The Official
CompTIA
Data+
Student Guide
(Exam DA0-001)

Course Edition: 1.0

Acknowledgments

Robin E. Hunt, ThinkData Solutions, Inc., Author
Thomas Reilly, Senior Vice President, Learning
Katie Hoenicke, Vice President, Product Management
Becky Mann, Director, Product Development
Evan Burns, Senior Manager, Learning Technology Operations and Implementation
James Chesterfield, Senior Manager, User Experience Design

Notices

Disclaimer

While CompTIA, Inc. takes care to ensure the accuracy and quality of these materials, we cannot guarantee their accuracy, and all materials are provided without any warranty whatsoever, including, but not limited to, the implied warranties of merchantability or fitness for a particular purpose. The use of screenshots, photographs of another entity's products, or another entity's product name or service in this book is for editorial purposes only. No such use should be construed to imply sponsorship or endorsement of the book by, nor any affiliation of such entity with CompTIA. This courseware may contain links to sites on the Internet that are owned and operated by third parties (the "External Sites"). CompTIA is not responsible for the availability of, or the content located on or through, any External Sites. Please contact CompTIA if you have any concerns regarding such links or External Sites.

Trademark Notice

Copyright Notice

Table of Contents

About This Course

CompTIA is a not-for-profit trade association with the purpose of advancing the interests of information technology (IT) professionals and IT channel organizations; its industry-leading IT certifications are an important part of that mission. CompTIA's Data+ Certification is an early-career data analytics certification for professionals tasked with developing and promoting data-driven business decision-making.

> This exam will certify the successful candidate has the knowledge and skills required to transform business requirements in support of data-driven decisions by:
>
> - Mining data
> - Manipulating data
> - Applying basic statistical methods
> - Analyzing complex datasets while adhering to governance and quality standards throughout the entire data life cycle
>
> *CompTIA Data+ Exam Objectives*

Course Description

Course Objectives

This course can benefit you in two ways. If you intend to pass the CompTIA Data+ (Exam DA0-001) certification examination, this course can be a significant part of your preparation. But certification is not the only key to professional success in the field of data analysis. Today's job market demands individuals with demonstrable skills, and the information and activities in this course can help you build your data skill set so that you can confidently perform your duties in any entry-level data analysis role.

On course completion, you will be able to do the following:

- Identify basic concepts of data schemas
- Understand different data systems
- Understand types and characteristics of data
- Compare and contrast different data structures, formats, and markup languages
- Explain data integration and collection methods
- Identify common reasons for cleansing and profiling data
- Execute different data manipulation techniques
- Explain common techniques for data manipulation and optimization
- Apply descriptive statistical methods
- Describe key analysis techniques
- Understand the use of different statistical methods
- Use the appropriate type of visualization
- Express business requirements in a report format
- Design components for reports and dashboards

- Distinguish different report types
- Summarize the importance of data governance
- Apply quality control to data
- Explain master data management concepts
- Identify common data analytics tools

Target Student

The Official CompTIAData+ (Exam DA0-001) Guide is the primary course you will need to take if your job responsibilities include gathering and collecting data for use in reporting, research of visualization, and analysis within your organization. You can take this course to prepare for the CompTIA Data+ (Exam DA0-001) certification examination.

Prerequisites

To ensure your success in this course, you should have 18–24 months of hands-on experience working in a business intelligence, report/data analyst job role.

You should have a working knowledge of Microsoft Excel or a spreadsheet program. You should understand how to build basic math calculations, like add, subtract, divide, and multiply (basic arithmetic). You should know how to build basic functions like Sums, Average, and Count. You should understand the basics of sorting and filtering data sets in Excel or a similar spreadsheet program. You should have a working knowledge of how to build very basic pivot tables. You should have some understanding of databases and all knowledge toward understanding how databases designed will be helpful. You should have a basic understanding of how to build simple charts in using data.

The prerequisites for this course might differ significantly from the prerequisites for the CompTIA certification exams. For the most up-to-date information about the exam prerequisites, complete the form on this page: www.comptia.org/training/resources/exam-objectives.

How to Use the Study Notes

The following notes will help you understand how the course structure and components are designed to support mastery of the competencies and tasks associated with the target job roles and will help you to prepare to take the certification exam.

As You Learn

At the top level, this course is divided into **lessons,** each representing an area of competency within the target job roles. Each lesson is composed of a number of topics. A **topic** contains subjects that are related to a discrete job task, mapped to objectives and content examples in the CompTIA exam objectives document. Rather than follow the exam domains and objectives sequence, lessons and topics are arranged in order of increasing proficiency. Each topic is intended to be studied within a short period (typically 30 minutes at most). Each topic is concluded by one or more activities designed to help you to apply your understanding of the study notes to practical scenarios and tasks.

In addition to the study content in the lessons, there is a glossary of the terms and concepts used throughout the course. There is also an index to assist in locating particular terminology, concepts, technologies, and tasks within the lesson and topic content.

Watch throughout the material for the following visual cues.

Student Icon	Student Icon Descriptive Text
❗	A **Note** provides additional information, guidance, or hints about a topic or task.
⚠	A **Caution** note makes you aware of places where you need to be particularly careful with your actions, settings, or decisions so that you can be sure to get the desired results of an activity or task.

As You Review

Any method of instruction is only as effective as the time and effort you, the student, are willing to invest in it. In addition, some of the information that you learn in class may not be important to you immediately, but it may become important later. For this reason, we encourage you to spend some time reviewing the content of the course after your time in the classroom.

Following the lesson content, you will find a table mapping the lessons and topics to the exam domains, objectives, and content examples. You can use this as a checklist as you prepare to take the exam and as you review any content that you are uncertain about.

As a Reference

The organization and layout of this book make it an easy-to-use resource for future reference. Guidelines can be used during class and as after-class references when you're back on the job and need to refresh your understanding. Taking advantage of the glossary, index, and table of contents, you can use this book as a first source of definitions, background information, and summaries.

How to Use the CompTIA Learning Center

The CompTIA Learning Center is an intuitive online platform that provides access to the eBook and all accompanying resources to support the Official CompTIA curriculum. The CompTIA Learning Center can be accessed at learn.comptia.org. An access key to the CompTIA Learning Center is delivered upon purchase of the eBook.

Use the CompTIA Learning Center to access the following resources:

- **Online Reader**—The interactive online reader provides the ability to search, highlight, take notes, and bookmark passages in the eBook. You can also access the eBook through the CompTIA Learning Center eReader mobile app.
- **Videos**—Videos complement the topic presentations in this study guide by providing short, engaging discussions and demonstrations of key technologies referenced in the course.
- **Assessments**—Practice questions help to verify your understanding of the material for each lesson. Answers and feedback can be reviewed after each question or at the end of the assessment. A timed final assessment provides a practice-test-like experience to help you to determine how prepared you feel to attempt the CompTIA certification exam. You can review correct answers and full feedback after attempting the final assessment.
- **Strengths and Weaknesses Dashboard**—The Strengths and Weaknesses Dashboard provides you with a snapshot of your performance. Data flows into the dashboard from your practice questions, final assessment scores, and your indicated confidence levels throughout the course.

Lesson 1

Identifying Basic Concepts of Data Schemas

LESSON INTRODUCTION

Data analysts work with data, and often that data is stored in a database. This is why it is of the utmost importance that you understand the basic foundations of databases. For example, the type of database you're working with can alter your decision-making as a data analyst, so it is crucial that you have a solid understanding of what differentiates relational from non-relational databases. Further, understanding how the tables in a relational database are designed and how the fields interact with each other will help you build necessary sets for analysis.

Lesson Objectives

In this lesson, you will do the following:

- Identify relational and non-relational databases.
- Understand the way we use tables, primary keys, and normalization.

Topic 1A

Identify Relational and Non-Relational Databases

EXAM OBJECTIVES COVERED
1.1 Identify basic concepts of data schemas and dimensions.

Data analysts are usually tasked with working with databases, and not building them. However, a little time spent understanding the design theory will help you to understand and interpret the designs of databases. Here we will cover the key characteristics of the two main types of databases: relational and non-relational.

Relational Databases

If you are already a data analyst or data worker in some area of the world, then you have likely been exposed to relational databases and the software used to maintain them. If you haven't, we'll walk you through the basics. A **relational database** uses tables to store data that is captured. You might also note that this type of format is often referred to as tabular schema, due to the rows and columns it employs. Software that maintains relational databases is often referred to as a **relational database management system (RDBMS)**.

Tables are created with fields, often called field names, and each row is a record of data that represents the data of that table. For example, if you wanted to create a spreadsheet of all the employees in a company, you would create field names like first name, last name, and employee number. Then you would list on each row the person's first name, last name, and their employee number, under designated columns. Tables in a relational database have two main design components: the field name and the data type, which we will explore deeper later in this course.

There is often a great amount of time spent designing a relational database. Database architects must determine what data needs to be stored and how to relate the tables of data to each other through relationships.

When you are using a relational database, you need to be able to retrieve the data that's stored within. **Structured query language (SQL)** is the language used to query and manage data in a relational database. SQL is a programming language for data; it's not a single software, which is important to remember. Just like any spoken language, there are different forms of SQL. For example, when using SQL in Microsoft, you are likely using T-SQL, which has additional coding elements that are specific to their software.

When data is retrieved from relational databases, there is usually some version of SQL involved in the process, whether you are coding it directly through SQL statements or using an interface that is coding it in the background for you.

Non-Relational Databases

A **non-relational database**, often referred to as NoSQL, is any alternative to a relational SQL database. This type of database addresses the need for web-based databases to handle large amounts of traffic and data, and are easier to scale for web applications.

A non-relational database does not use the same tabular schema (rows and columns) that you find in relational databases. Whereas a relational database contains tables with defined relationships, non-relational databases store their data in several different ways depending on the type of data required by a specific development. A non-relational database is more flexible than a relational database, and it is ideal for data that doesn't fit into the more rigid structure of a relational database.

The way we retrieve data from non-relational databases also differs from the way we do this for relational databases. While relational databases require traditional SQL-based queries, you will often use different programming languages to retrieve the data you need from a non-relational database.

There are four basic categories of NoSQL databases.

- **Document-oriented databases** store data in XML documents or JSON. They are flexible in that they allow developers to reshape the data to meet the format needed for the application.
- **Key-value stores** store each value with a key value. This is similar to a table with just two columns: a field (key) and value.
- **Column-oriented databases** store data in columns rather than rows. This design can make for easier analysis in some cases, such as when counting the total number of orders.
- **Graph stores** store individual elements as nodes. This type of database is more complex and focuses on the relationships between data elements. In a graph database, the connections are first-class elements of the database and stored directly. You will often find that graph databases run side by side with a relational database.

Identify Key Differences between Relational and Non-Relational Databases

It is important to have a firm understanding of the foundational concepts of databases. As a data analyst, how data is designed and stored determines how you will access and work with data. There are a few key differences between the two types of databases that bear repeating to compare them directly.

Relational Databases:

- Relational databases have tables that store fields in columns, and rows of records that hold data in relational database management systems (RDBMS).
- Relational databases use primary and foreign keys to establish relationships that control how certain data relates to data in other tables.
- Relational databases leverage normalization techniques for optimal design.
- Relational databases use traditional SQL language to query and handle transactions within the database.
- Relational databases expect data to fit within the tables that are designed for the database.

Non-Relational Databases:

- Non-relational databases are the alternative to relational databases and do not follow the same structural requirements of a SQL-based database.
- Non-relational databases are much easier to scale and build for web-based applications and do not require the same level of detailed planning and structure that relational databases require.
- Non-relational databases use various programming languages to retrieve and handle transactions within the stored data.
- There are four categories of non-relational databases: document-oriented (the most prominent), key-value stores, column-oriented, and graph stores.
- Non-relational databases can store structured and non-structured data with more flexibility.

Relational Database	**Non-Relational Database**
Contains tables that store fields in columns, and rows of records that hold data	Does not follow the same structural requirements as an SQL-based database
Uses primary and foreign keys to establish relationships that control how certain data relates to data in other tables	
Leverages normalization techniques for optimal design	Does not require detailed planning and structure to build
Uses traditional SQL language to query and handle transactions within the database	Uses various programming languages to retrieve and handle transactions within the stored data
Expects data to fit within the tables that are designed for the database	Can store structured and non-structured data with more flexibility

Review Activity:

Relational and Non-Relational Databases

Answer the following questions:

1. **Which type of database should you use when you need a defined structure with relationships within the data?**

2. **What are the two main components of a table design?**

3. **Which type of database stores data in document format and uses XML or JSON?**

4. **What is the language that is used for relational databases?**

5. **Which type of database provides the most scalable and flexible option for web-based applications?**

Topic 1B

Understand the Way We Use Tables, Primary Keys, and Normalization

EXAM OBJECTIVES COVERED

1.1 Identify basic concepts of data schemas and dimensions.
2.3 Given a scenario, execute data manipulation techniques.
5.1 Summarize important data governance concepts.

When you're working with tables in a database, it is important to know up front that the way the data is stored differs from the way we query it for reporting. Here we will dive into the way databases are designed in theory. We will discover the process of normalization, types of keys, types of relationships, and how these relationships impact the data in our systems.

Normalization

When a database is designed, the data is structured for optimal storage and use within the program. That means the data is **normalized data**. Have you heard the saying "everything has a place, and everything in its place"? That is, in a nutshell, what the process of normalizing data means. Normalizing data is a form of organization, and it supports the design by optimizing storage. Normalization of data also adds flexibility in working with the data to support the design of front-end interfaces.

We have tables that hold customer information, and we don't store all the customer and salesperson data in the same table. Customer information goes to customer table, and salesperson data goes to a salesperson table.

As a data analyst, you hope to work with databases that have followed proper design theory and the principles of the normal forms. Most database designers will at least attempt to get data to the third normal form.

- *The First Normal Form (1NF)* eliminates redundant information in individual tables. Each set of related data will be stored in a dedicated table. Each table of related information will have a primary key assigned.
- In the *Second Normal Form (2NF)*, related information that is applicable to multiple tables will have its own table and will be associated through the use of a foreign key.
- The *Third Normal Form (3NF)* eliminates fields that do not depend on a key. You will likely find that while some designs go to the 3NF, most do not.
- The less practical fourth and fifth normal forms are rarely used, so we won't go into them here, but you should know they exist.

The tables below provide some examples of different forms.

StudentID	FirstName	LastName	Tuition Amt1	Tuition Amt2	Tuition Amt3
1000	Susie	Baker	1,500.00	1,500.00	1,500.00
1001	Daniel	Bishop	1,575.00	1,575.00	1,575.00
1002	John	Brown	1,475.00	1,475.00	1,475.00

Not Normalized

This table shows student data that is not normalized. The table has a column for every tuition payment, and it denotes each with a tuition amount and number. This would mean every tuition payment that ever needs to be created must have a field name. This is not a functional design or method.

StudentID	FirstName	LastName
1000	Susie	Baker
1001	Daniel	Bishop
1002	John	Brown

First Normal Form (1NF)

This table shows how we would normalize student data to the 1NF by creating a single student record in a table. We eliminate redundant data (the tuition payments) and store that related data in a separate table.

PayID	StudentID	TuitionAmt
252	1000	1,500.00
253	1001	1,575.00
254	1002	1,475.00
255	1000	1,500.00
256	1001	1,575.00
257	1002	1,475.00
258	1000	1,500.00
259	1001	1,575.00
260	1002	1,475.00

Second Normal Form (2NF)

This table shows how we would normalize tuition payment data to the 1NF. Note that we now use a dedicated PayID, rather than repeating the tuition payment amounts. Further, the StudentID is how we identify the student in this second table, making it the primary key. Once data has been normalized, relationships between the data can then be established.

Relationships in Data

A relational database is "relational" because relationships are formed between keys in the tables. This design allows a unique identifier to be assigned to a record of data in the table, which distinguishes that record from every other record in that table. A key used to identify a record is referred to as a **primary key**. When a primary key is used in another table to refer to your record, it's known as a **foreign key**.

Suppose you've been accepted to attend a university. On your first day, you are assigned a student number that is unique to you. That number will identify you in your school's other systems. If we are following the rules of normal forms, your student number will be associated with all records pertaining to you. For example, your student number, name, and address may be stored in one table, while your student number and tuition payment information will be stored in another table. There are several benefits of this design.

- The university can simply associate your student number, or key, to any necessary information.
- A change to student data will automatically carry over where it belongs.

If the school's database was not designed using this normalization method, to refer to you in another table, they would have to retype all information about you each time. We call that redundant data. When there is redundant data, and a change needs to be made, that information would need to be manually changed in many different places.

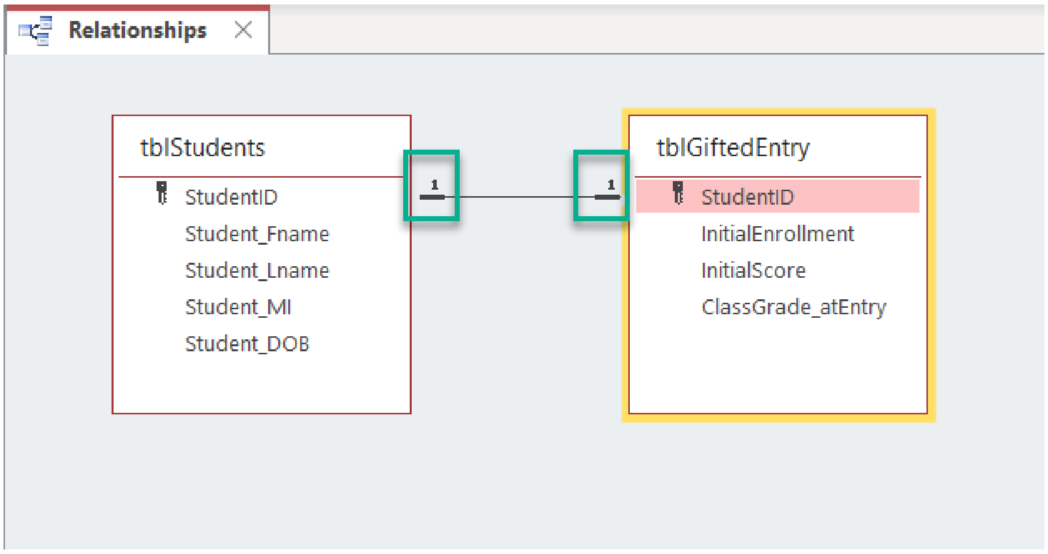
Imagine how big payroll records would get over a five-year period if each employee's information had to be repeatedly typed in every time checks were issued. Not only would it take more time to do payroll, but it would also create more storage, and data quality would be questionable (because we all know how easy it is to mis-key information).

Types of Relationships

Database tables that use primary and foreign keys to create relationships have varying levels of **cardinality**, which describes how many possible occurrences of one entity (record in a table) can be associated with the number of occurrences in another (records in another table).

One-to-One Relationship

A **one-to-one relationship** means that one record in a table will be associated with only one record in the other table. In a one-to-one relationship, the primary key in the first table is often the primary key in the other table as well.



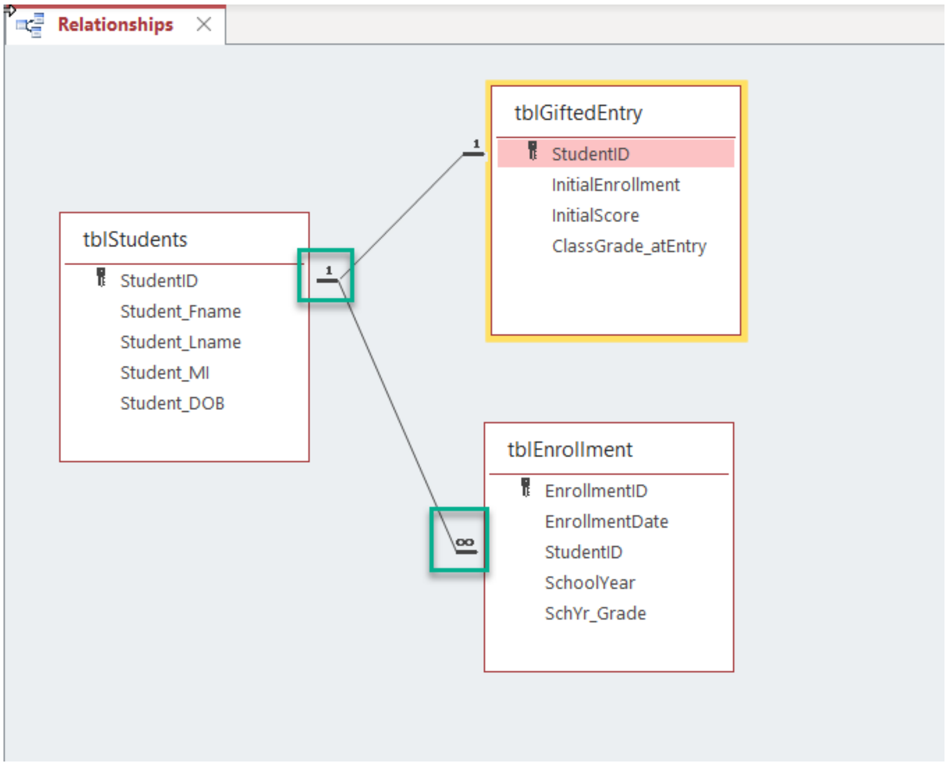
One-to-One Relationship in Microsoft Access (Used with permission from Microsoft.)

Let's walk through an example. Suppose when a student enrolls in school they are entered into the tblStudents table. The student record contains basic student information, like StudentID, date of birth, and other identifying information. If a student meets the initial score requirements needed for gifted services, a record will also be entered into the gifted entry table. The Student ID is the primary key for both tblStudents and tblGiftedEntry.

The symbols on the relationship line indicate the type of relationship that exists between these two tables. There can be only one record per student in tblStudents and only one record per student for GiftedEntry. Thus, the cardinality of the two tables is a one-to-one relationship.

One-to-Many Relationship

In a **one-to-many relationship**, a primary key is joined to a foreign key, meaning there is one record (in the table in which the key is primary) associated with multiple records in other tables (in which the key is foreign)).

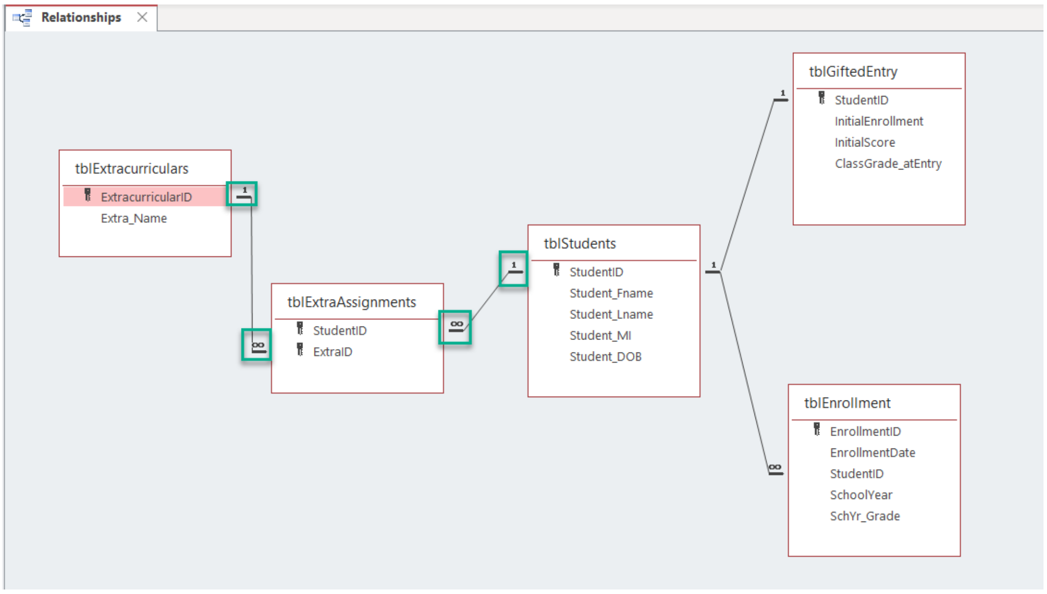


One-to-Many Relationship in Microsoft Access (Used with permission from Microsoft.)

Suppose each time our student enters into a new school year, we use their StudentID from the student information table to associate them to their new enrollment record in the tblEnrollment table. StudentID is a primary key in tblStudents but a foreign key in tblEnrollment. That means we have one record for each student in tblStudents, and we have as many enrollment records for that student as years they are enrolled. (If this student went to school for five years, we would have five records to represent each year they were enrolled.) On the diagram, you see a "1" next to tblStudent and an infinity symbol near tblEnrollment. These symbols represent the cardinality. There can only be one record in tblStudents for a student, but an infinite number can exist for that same student in tblEnrollment, and thus this is a one-to-many relationship.

Many-to-Many Relationship

A **many-to-many relationship** means that you have many records associated with many other records. Data tools can't resolve a many-to-many relationship without the use of another table that uses the associated keys from each table to serve as a bridge.



Resolving a Many-to-Many Relationship with a Junction Table in Microsoft Access (Used with permission from Microsoft.)

Suppose our student records management system also tracks participation in extracurricular activities. Many students are involved in more than one activity; for example, a student who is in band and on the debate team. There are also many other students in band and on the debate team with this one student. The tblExtracurriculars table uses the ExtracurricularID field to identify the activity along with the Extra_Name field, which allows us to identify which student is participating in that activity. We also have each student's basic information in tblStudents. One record represents the activity (tblExtracurriculars), and one represents the student (tblStudents). In order to allow many students to do many extracurricular activities, we have a bridge table, tblExtraAssignments, that contains the two fields from our separate records (ExtracurricularID and StudentID). StudentID from the main student information table. On the diagram, the many-to-many cardinality is represented by a "1" next to our two single records and an infinity symbol next to our bridge table, tblExtraAssignments.

In the bridge table, the Extracurricular ID is called ExtraID. This is the same data, but has a different name. While you would hope that the names would be consistent, this isn't always the case and something you should be aware of as a data analyst.

Referential Integrity

A database design that uses primary and foreign keys to set relationships between records must also establish referential integrity. **Referential integrity** ensures that a foreign key definitively has a primary key in the related table. Let's revisit the one-to-one relationship in our student information example. We have one student record that represents a student in the main student table, and one student record for that same student in the gifted table. Referential integrity guarantees that the StudentID must exist in one table (such as the main student table) before a record can be created in the other table (the gifted table).

Referential integrity helps to prevent the occurrence of bad or missing data in any of the tables in the design. In short, it ensures that database users are not allowed to add records when a related record doesn't exist. Let's return to our student example. If the database did not have referential integrity, someone could add a gifted services record for a student who does not exist. With referential integrity, you would not be allowed to add a gifted services record for an unknown student.

There are additional settings that can be established when referential integrity is set. These options will affect how data updates and/or will delete data across the related tables.

- **Cascade update**: When a primary key is changed and cascade update is enforced, the primary key will change in all other related tables. The benefit to this design is that you do not have to identify all the records in the various tables that use the key and manually update them.
- **Cascade delete**: When a primary key record is deleted and cascade delete is enforced, all records in various tables that are related to the record through that key will be automatically deleted.

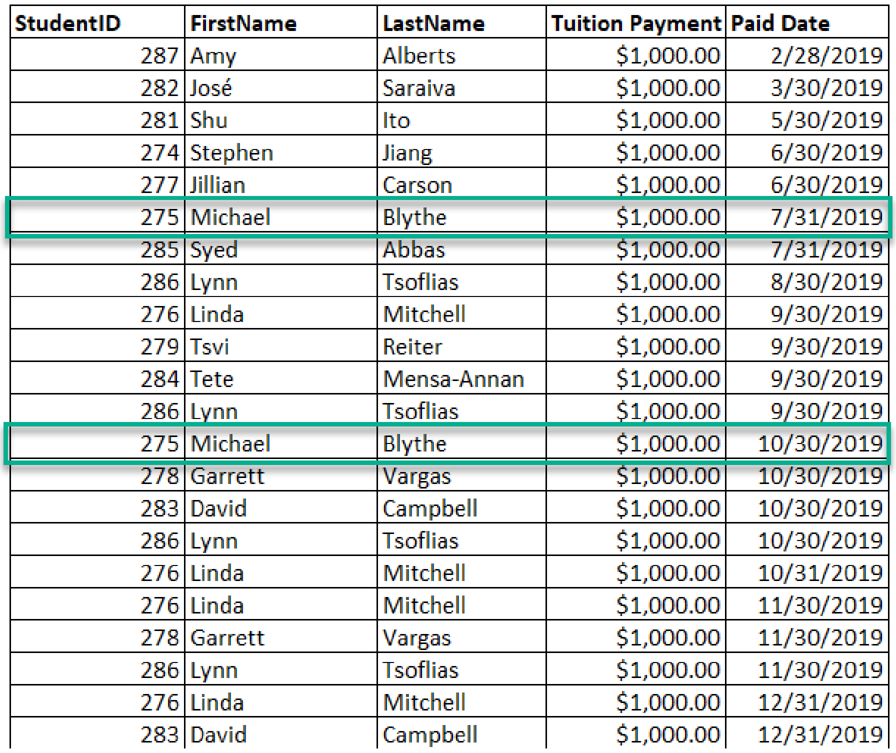
Think about how student information might change over a four-year period. They may move to a new apartment, change their name, or change their phone number. Now suppose a student's primary key is a random number followed by the first three letters of their last name. When that person's last name changes, the key must also change. When referential integrity has a cascade update setting, that key only needs to be updated with the first three letters of the new last name in one place; the update will cascade to all other records. Similarly, if a student drops out of the university, the cascade delete setting ensures that all records relating to the student are deleted when the key is deleted. Due to the permanent nature of cascade delete, you will likely find that it is not set on many relationships. In our example, the enforcement of cascade delete could mean we lose all the historical records of a student's tenure. So while this option may be set on some relationships, it is used less frequently than the more convenient cascade update setting.

As a data analyst, you are not likely to be designing databases. You merely need to know what referential integrity means to a design. If a database is designed without these integrity rules, you might find gaps in the data due to people making additions without following the order of data entry, or deleting information while leaving related information behind.

Denormalization

Data that is not properly designed using normalization would be referred to as **denormalized data**. When data is not structured into tables using normalization, you will find lots of redundant and repetitive data.

Consider the example below. For every tuition payment, student information—such as student ID, first name, and last name—is entered into the table. This is an example of denormalized data because the student information is manually keyed in and repeated each time a tuition payment is made.



StudentID	FirstName	LastName	Tuition Payment	Paid Date
287	Amy	Alberts	$1,000.00	2/28/2019
282	José	Saraiva	$1,000.00	3/30/2019
281	Shu	Ito	$1,000.00	5/30/2019
274	Stephen	Jiang	$1,000.00	6/30/2019
277	Jillian	Carson	$1,000.00	6/30/2019
275	Michael	Blythe	$1,000.00	7/31/2019
285	Syed	Abbas	$1,000.00	7/31/2019
286	Lynn	Tsoflias	$1,000.00	8/30/2019
276	Linda	Mitchell	$1,000.00	9/30/2019
279	Tsvi	Reiter	$1,000.00	9/30/2019
284	Tete	Mensa-Annan	$1,000.00	9/30/2019
286	Lynn	Tsoflias	$1,000.00	9/30/2019
275	Michael	Blythe	$1,000.00	10/30/2019
278	Garrett	Vargas	$1,000.00	10/30/2019
283	David	Campbell	$1,000.00	10/30/2019
286	Lynn	Tsoflias	$1,000.00	10/30/2019
276	Linda	Mitchell	$1,000.00	10/31/2019
276	Linda	Mitchell	$1,000.00	11/30/2019
278	Garrett	Vargas	$1,000.00	11/30/2019
286	Lynn	Tsoflias	$1,000.00	11/30/2019
276	Linda	Mitchell	$1,000.00	12/31/2019
283	David	Campbell	$1,000.00	12/31/2019

Example of Denormalized Data in Microsoft Excel (Used with permission from Microsoft.)

It's important to understand there are other reasons for denormalized data to exist that are not related to design flaws in the data entry process. Data analysts actually denormalize data for the purposes of analysis quite regularly through querying. The process of querying involves gathering data from all the different single source tables and joining that information together for the purposes of data warehousing, data mining, analysis, and visualization. The end result in the data set appears denormalized. Normalizing data breaks it down, but denormalization brings it back together.

Let's revisit our student data example. If we were asked to produce a list of meaningful data that shows each student and all their tuition payment records, we'd see a lot of repetitive data. We don't want to store the data that way, for obvious reasons, but for reporting purposes we need that redundancy so we can create appropriate reports with all the necessary information. As you go through this course and begin to understand how we join data together, you will be performing denormalization.

In the work that we do as data analysts, we don't always use the formal terms of "normalizing" or "denormalizing" data. In general conversation, we say that we are joining data, or querying data with joins.

Review Activity:

Tables, Primary Keys, and Normalization

Answer the following questions:

1. **Which elements of a database are necessary in order to establish relationships between tables?**

2. **To ensure data updates effectively across tables, what can be set on the relationships between tables?**

3. **When the design of the data forces a person to repetitively enter the same information over and over, this data would be considered _____.**

4. **A database architect creates a design in which a table has a primary key and associated fields. This is an example of what type of design theory?**

Lesson 1

Summary

After reading this lesson, you should be able to identify the key differences between relational and non-relational databases, such as the structure and scalability of designs. You should also be able to recognize whether data is normalized or denormalized, and understand the benefits of each.

As a data analyst, you will not find yourself designing databases for enterprise organizations but rather using the data from these systems as designed. However, the more you discover about design theory, the easier it becomes to investigate the databases you work with and recognize what you may need to do with the data they store.

Guidelines in Understanding Data Schema

Consider these best practices and guidelines when familiarizing yourself with the databases you will be using in your analysis.

1. Identify what types of databases you're working with (relational or non-relational, or both).
2. Identify the different technologies the organization has adopted.
3. Familiarize yourself with the tables, views, and relationships that exist within the databases you have permission to view.
4. Understand the cardinality or types of relationships (one-to-one, one-to-many, many-to-many) that exist in the tables.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 2

Understanding Different Data Systems

LESSON INTRODUCTION

Learning about common types of database processing and storage systems will help you understand the different types of structures and schemas you will encounter in your work as a data analyst. You don't have to know everything about all of these data systems, as they are controlled by other data roles in an organization. However, you will be accessing data stored and/or processed with these technologies.

Lesson Objectives

In this lesson, you will do the following:

- Describe types of data processing and storage systems.
- Explain how data changes.

Topic 2A

Describe Types of Data Processing and Storage Systems

EXAM OBJECTIVES COVERED

1.1 Identify basic concepts of data schemas and dimensions.

As a data analyst, you will work with data that is stored in different types of data management systems. Data analysts will also sometimes have access to the source systems that provide data to these management systems as part of their role to identify and validate data. All these technologies work with data at different points. Further, having a firm understanding of the basic data schemas will help you identify how the data has been joined together, ultimately informing your work as an analyst.

Types of Data Processing

There are two types of processing in the world of databases:

- **Online transactional processing (OLTP)** is a class of software that allows large numbers of database transactions in real time, typically over the internet. A **database transaction** is any change to data in a system, whether it's an insertion, deletion, or query.
- **Online analytical processing (OLAP)** is a class of software that allows complex analysis to be conducted on large databases without negatively affecting transactional systems.

Every day, people perform transactions that are processed with OLTP. One familiar example is using an automated teller machine (ATM). When you make a balance request of the ATM, you are querying the database that holds the information of your account. When you make a deposit or withdrawal, you are creating a record of that transaction. A key benefit of using the ATM is that you can log into your banking account online and see that this transaction has occurred.

The group or organization that owns the ATM needs to analyze the data for all of the transactions that occur. Even doing the most basic analysis (e.g., sums or counts) of billions of transactions is tough for an Excel spreadsheet. You wouldn't even consider trying to analyze data this way, and your machine would likely not be able to handle it. This is where OLAP comes into play. This technology allows a high volume of data to be analyzed more effectively because it is meant to handle massive transactions from OLTP.

Source Systems

A **source system** is the system of record (i.e., information storage system) for any given data element or piece of information. It's dedicated to a particular type of information, and the term "system of record" implies that the system holds the absolute truth of that data; that is you can trust this system to be correct.

Organizations have a wide variety of source systems. Examples include systems for accounting information, payroll, human resources, and more. When an organization has a high volume of data, they will implement master data management software and policies to create a single source of truth for their dimensional data. A source system is often the first entry of any given information. For example, accounting systems help support what an organization will need to meet IRS regulations. Human resource systems are dedicated to the people that work at an organization and are meant to support the different processes centered around people. A Customer Relationship Management (CRM) system is a system of record and source system for all the customer relationship aspects. These different source systems are dedicated to specific functions and processes within an organization. If we didn't have source systems dedicated to processes, someone would have to build a single data system that covers all processes within an organization. That would be a major undertaking and honestly financially impossible to maintain for most organizations.

Let's consider a company that sells a product. This company will likely have a dedicated system or systems that support that effort. All information regarding the sale of the product is interrelated, even if it is located in entirely different systems. For instance, some of the employees in the HR system are salespeople in the product source system, while others who support the production, shipping, or receiving of products are in the production source system. Because the organization has multiple sources of information, and each source system is likely not connected to the other source systems, there is a disconnect in the reporting and data needs of the organization that can be solved with the use of a data warehouse.

Data Warehouses and Data Marts

Data Warehouses

A **data warehouse** is a technology that is dedicated to the storage of company data from a wide range of sources for reporting and decision-making purposes. Data warehouses unify and hold the data from multiple source systems. Different data from different systems can be queried together without impacting the performance of the separate source systems. The advent of the data warehousing model gives us a reporting model that overcomes the gap in reporting while still allowing each source system to do what it is defined to do.

To learn more about the valuable concepts of data warehousing, you can explore Dimensional Modeling Techniques from The Data Warehouse Toolkit, Third Edition by Ralph Kimball. https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/

Data warehouses provide a single source of truth, meaning the data in the warehouse is trustworthy, as there's a lot of rigor involved in confirming what data should go into the warehouse, determining where that data comes from, and effectively relating the data for the use of reporting.

Data warehouses can also leverage OLAP technology, making data processing more effective. For example, OLAP can summarize thousands of records far faster than traditional tools like Excel. The "heavy lifting" of the data work, such as grouping data by month and then totaling that data, can be done by OLAP technology through the use of OLAP cubes.

- These cubes are three-dimensional and provide data that is already grouped by different dimensions. Thus, when analysts access data, they can further analyze it without having to group and total it first.
- When an organization's data warehousing professionals know the types of reports that are routinely run, they can use this knowledge to design cubes that support business intelligence requirements within the organization.

Note that some skills used by the professionals who build data warehouses are also shared and leveraged by data analysts.

- *Data analysts join data together through queries, gather data from different systems, and combine them for analysis.*
- *Data analysts will work with source systems and warehouses in their daily work.*

We will cover these skills more deeply in the coming lessons.

Data Marts

A **data mart** is a subset of the data warehouse that is dedicated to a specific department or group. Data marts ensure that the data needed by a defined group of users is available to them. Data marts also ensure that data is secure and only available to the users who should have access to that data.

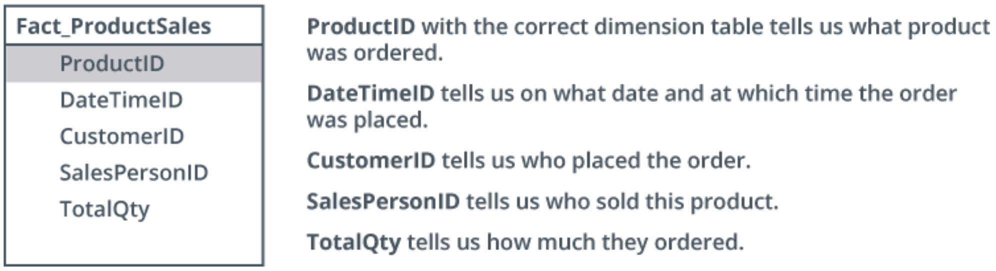
Imagine it this way: The data warehouse is like a shopping mall of data, and the data mart is a particular store in the mall. Unlike the stores in the mall, however, you can't just walk into any data mart. Your access is controlled based on what data you need.

Data warehouses and data marts are structured, and that structure allows us to report on or build visualizations from data without requiring access to every single source system. Without a data warehouse, we would have to integrate all of the data from different source systems ourselves.

Schemas Used in Data Warehousing

Next, let's delve more deeply into the structure of a data warehouse. In data warehousing, two common schemas are used to relate data tables to one another: snowflake and star schemas. These schemas get their names based on how they look when all the tables are structured. A key similarity between these two schemas is the use of fact tables and dimension tables. A **fact table** holds the "facts" about a particular business process or event (in most cases numbers, metrics, or measurements) and contains keys to relate to the other tables. A **dimension table** holds attributes or the categorical information that support the fact tables. The information contained within a dimension table gives us more information related to the fact table, thus providing us with a full set of data that will be more meaningful for the reporting of these facts. Both the snowflake and star schema use joins to relate these two types of tables, and both schemas also use normalization rules.

Here is an example of a fact table named Fact_ProductSales. It contains the facts related to a sale: the product sold, the date and time the product was ordered, the customer who ordered the product, the salesperson who sold the product, and how much of the product was ordered.



Fact_ProductSales
ProductID
DateTimeID
CustomerID
SalesPersonID
TotalQty

ProductID with the correct dimension table tells us what product was ordered.

DateTimeID tells us on what date and at which time the order was placed.

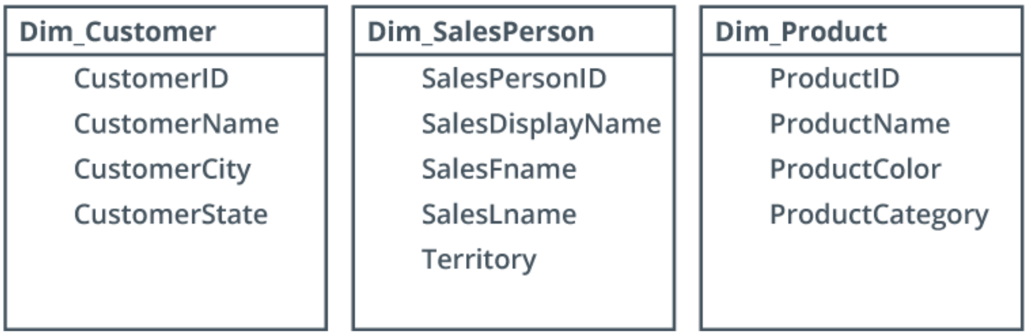
CustomerID tells us who placed the order.

SalesPersonID tells us who sold this product.

TotalQty tells us how much they ordered.

Sample Fact Table Showing Keys and Facts

Next, let's look at some examples of dimension tables that are related to the Fact_ProductSales table. Note how these tables are used to meaningfully categorize data by customer, salesperson, and product.



Dim_Customer	Dim_SalesPerson	Dim_Product
CustomerID	SalesPersonID	ProductID
CustomerName	SalesDisplayName	ProductName
CustomerCity	SalesFname	ProductColor
CustomerState	SalesLname	ProductCategory
	Territory	

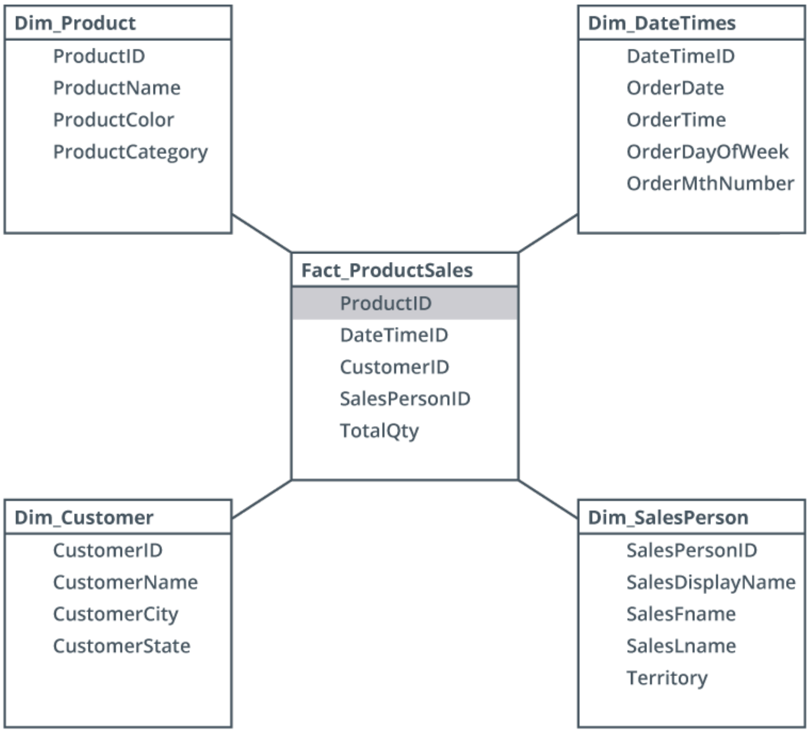
Dimension Tables with Details to Suppport the Fact Table

If you are currently an analyst who works with source data, the terms for fact tables and dimension tables are likely unfamiliar to you because you are used to working with queries on the actual tables by name. You likely do not call them fact or dimension tables.

Now that we have covered the common elements of both schemas, let's uncover the differences between them.

Star Schema

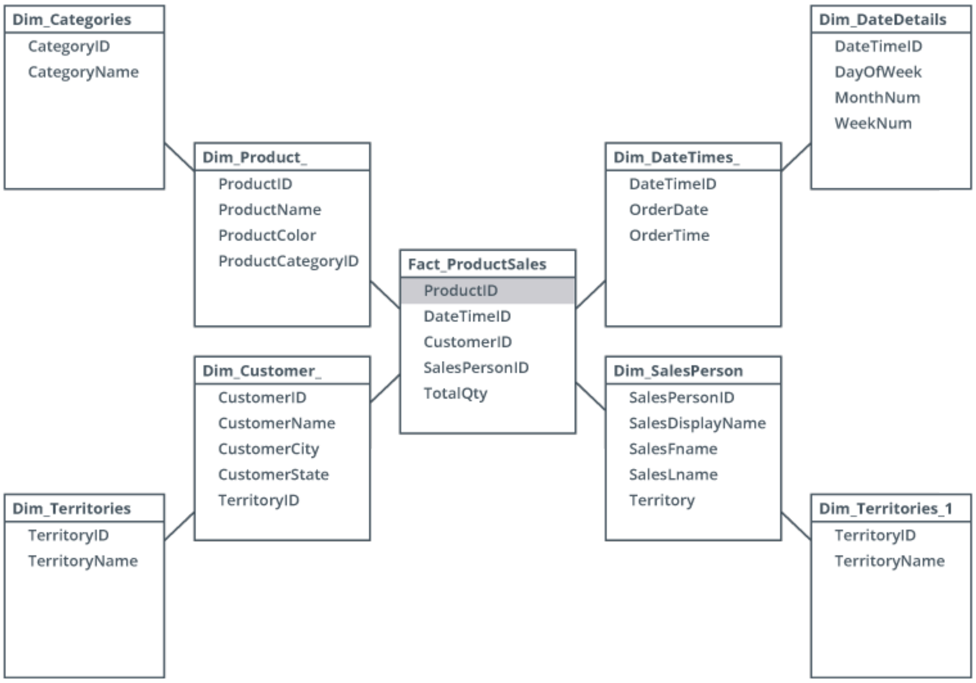
In a **star schema**, any related dimension tables are joined to a single fact table in a visual layout that looks like a star. It is one of the simplest schemas to use.



Example of Star Schema

Snowflake Schema

A **snowflake schema** also features a single fact table joined to related dimension tables. However, unlike in the star schema, the dimension tables in a snowflake schema might also be tied to other dimension tables, creating the appearance of a snowflake.



Example of Snowflake Schema

Data Lakes and Lakehouses

Data Lake

In the world of data and within an organization, we really only analyze a percentage of all the data that we collect. Organizations make a significant investment in software, storage, and reporting capabilities. However, they need a place to hold data of all types when it doesn't follow the same rigid structure that is designed by databases and data warehouses. When data has been collected but is not yet ready for cleaning or analysis, it can be stored in what is called a **data lake**. Data lakes hold both structured and unstructured types of information, allowing an organization to store large amounts of data of all types in its original format. Data lakes serve as a kind of "catch-all" for data.

The biggest drawback of data lakes is that, unlike a data warehouse, there is less rigorous oversight of and control over how data is entered. That's not to say that there is no data governance, but a data lake isn't acknowledged as the single source of truth like a data warehouse.

Data Lakehouse

All technologies have pros and cons. Data warehouses, although extremely powerful, require diligence in determining what goes in and where it comes from, and preparing that information for analysis. Data lakes capture large amounts data in holding, while waiting for those processes to be conducted.

Because the world of data is constantly evolving, new innovations can change the way we store and access data. The **data lakehouse** is a data management system that combines the best of both data warehousing and data lakes. Data lakehouses provide flexibility, like a data lake, and yet are often more cost effective than a data warehouse. Data lakehouses serve up information not only for data analysts, but also data scientists and data engineers. Data lakehouses support business intelligence and analytics projects in addition to machine learning and data science.

While some companies will make use of these newer data technologies, other companies might still use databases or exports from their source systems. Your ability to access an organization's data will likely be controlled by the information technology (IT) department, and you will be granted permission to the data that you need to access through some form of data governance plan. How you connect to any data source will be controlled by the technologies and software that your organization has adopted. Regardless, only after you determine where the data lives and how you access it can you begin to use it for further analysis.

Review Activity:

Types of Data Processing and Storage Systems

Answer the following questions:

1. **Which subset of a data warehouse holds data that is relevant to a specific department?**

2. **What is the most flexible storage system for holding large amounts of both structured and unstructured data?**

3. **In an organization, which technology holds a system of record, or is a software system dedicated to a specific task?**

4. **What choices do organizations have when they want to combine data from different source systems into one unified data management system?**

5. **What type of table stores information that is categorical for the use of reporting?**

6. **What are two common schemas that are used to relate data in a data warehouse?**

Topic 2B

Explain How Data Changes

EXAM OBJECTIVES COVERED

1.1 Identify basic concepts of data schemas and dimensions.

Now that we have discussed the different ways we process and store data, it's time to understand how data changes in the systems from which you will be gathering it. Let's dive into the concept of slowly changing dimensions and how it impacts the data you will be reporting.

Overview of Slowly Changing Dimensions

As you have now learned, data warehouses capture data from multiple source systems. These source systems are active every day, so the data in the warehouse is consistently updated. When a data warehouse is designed, the data in the dimension tables will change based on the decisions made by the architects who design databases, and the fact tables will update as new facts are created (such as a new sale occurring). Data analysts are not typically responsible for making design decisions, but because of the way we analyze data, it's important to understand how data changes.

Slowly changing dimensions are a way of updating dimension data. Dimension data is categorical data, and it doesn't change as often as fact data does. To demonstrate, let's consider two pieces of data for a business. A company sells a variety of products, which are named in a field called Product Name. Each time a new order is created, that order is given a purchase order (PO)number. The product name is dimension data because it doesn't change often, but data related to the specific orders (such as quantity ordered) is fact data because it is continually changing.

It may be hard to distinguish between fact data and dimension data at the beginning of your data journey. We tend to think of facts simply as being any true, verifiable pieces of information, so all data may seem like facts. When working with data warehouses, however, the general definition of facts in everyday life isn't applicable. Dimension data does not change very often, but it is used alongside fact data that changes regularly.

There are three primary types of slowly changing dimensions:

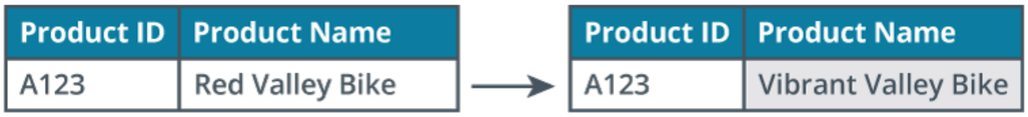
- **Slowly Changing Dimension Type 1** overwrites the existing value in a field of data with a new value and does not retain the history of that field.
- **Slowly Changing Dimension Type 2** adds a new row for the new value and also maintains the historical record.
- **Slowly Changing Dimension Type 3** adds a new column for the current value but also retains the original column with the original value.

The Impact of Slowly Changing Dimensions

Data warehouses often do not update in real time, but rather on a scheduled basis (such as once every 24 hours). Data warehouses often hold the entire history of data for reporting and business intelligence projects. How the data changes plays a major role in how the data is stored in the data warehouse. While dimension data doesn’t change as often as fact data, it does still change. How the data warehouse is designed to handle these changes is based on the slowly changing dimensions.

Let’s use an example to walk through how updates are made to dimension data for each of the three types of slowly changing dimensions. Our scenario is a company that sold a popular bike with the product name “Red Valley Bike” from January 2000 to 2010. In 2011, the company decided to rename the product “Vibrant Valley Bike.” Here’s how that product name would be updated according to the different types of slowly changing dimensions.

1. ***Type 1 Slowly Changing Dimensions*:** This method overwrites the existing value (Red Valley Bike) with the new value (Vibrant Valley Bike). Every data set we generate from any point in time will now have the most current name applied—the previous product name will not be retained. Type 1 may be preferred if only the most current product name is needed, and the former name of the bike is no longer relevant.



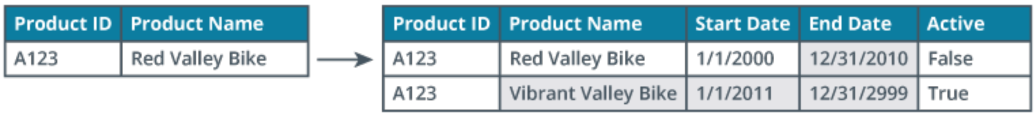
Product ID	Product Name
A123	Red Valley Bike

Product ID	Product Name
A123	Vibrant Valley Bike

Slowly Changing Dimensions Type 1

This change type means that every time we use the product ID A123, the associated product name will be “Vibrant Valley Bike,” even for data that was created prior to the name change (when the product was named “Red Valley Bike”). When we report on this table, we will establish a join from the product ID as a primary key to the product ID as a foreign key in the sales table. The key did not change, but the value of the product name did change. New sales records will only reflect the current information for that product name.

2. ***Type 2 Slowly Changing Dimensions*:** This method allows us to create reports showing that the product name was formerly “Red Valley Bike” for all historical reports between January of 2000 to December of 2010, even though the bike is now named “Vibrant Valley Bike.” Type 2 is preferred if it's important to have a record of historical data for the product name. As depicted in the next image, this method retains the record with the original product name and creates a record with the new product name, providing the most complete history of the product name. The records include start and end dates for each name of the bike.



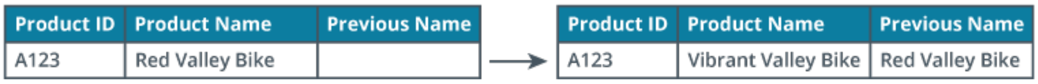
Product ID	Product Name
A123	Red Valley Bike

Product ID	Product Name	Start Date	End Date	Active
A123	Red Valley Bike	1/1/2000	12/31/2010	False
A123	Vibrant Valley Bike	1/1/2011	12/31/2999	True

Slowly Changing Dimensions Type 2

This means the data warehouse holds both the historical information as well as the current information. This type of slowly changing dimension provides you with the most detailed history for reporting and allows you to leverage the dates of the change to provide accurate information at that time in any of your reporting.

3. ***Type 3 Slowly Changing Dimensions:*** In this method, a current value column would be created with the current product name (Vibrant Valley Bike), while the original column is retained and shows the old product name (Red Valley Bike). Unlike with Type 2, there is only a single record for the product using Type 3. It is important to note that Type 3 only maintains the last two changes made to the data. For example, using this method, if the product name changed a third time (e.g., to Vibrant Red Valley Bike), then the current name of the product and the last product name would be retained, but not the original product name.



Product ID	Product Name	Previous Name
A123	Red Valley Bike	

Product ID	Product Name	Previous Name
A123	Vibrant Valley Bike	Red Valley Bike

Slowly Changing Dimensions Type 3

Using this method, a new column is added showing the previous name. If the bike is renamed again, this update would only capture the new product name (Red Vibrant Valley Bike) and the previous name (Vibrant Valley Bike). The original product name (Red Valley Bike) would be lost, meaning we lose some history for data in this field when it changes.

As a data analyst, you should learn to view data through the lens of Type 1, Type 2, or Type 3 slowly changing dimensions, because this understanding will help you to communicate what you can report historically. As you grow in your function-writing capability, you will learn to write logical functions that can help you return the appropriate information for whatever report you are writing.

Review Activity:

Explain How Data Changes

Answer the following questions:

1. **What process does slowly changing dimensions describe?**

2. **Which type of slowly changing dimension is in use when a new record is added to account for a change to dimension data?**

3. **If the name of a product changes using Type 1 slowly changing dimensions, what impact would this method have on reporting prior to the change?**

4. **Which slowly changing dimension type provides the most complete history of dimension data, and why?**

Lesson 2

Summary

After reading this lesson, you should be able to identify transactional systems that leverage OLTP processing and analytical systems that leverage OLAP for complex analysis. You should also be able to identify the source systems of data at an organization. You should understand the differences between data warehouses, data marts, data lakes, and data lakehouses. You should also be familiar with dimension and fact tables, and two common schemas used for data warehousing: star and snowflake. You should be able to relate how data updates in a system, including the way that slowly changing dimensions impacts dimensional data.

Guidelines in Understanding Data Systems

Consider these best practices and guidelines when familiarizing yourself with the data systems you will be working with.

1. Identify the source systems in use at the organization.
2. Identify what technologies have been implemented for the purposes of storing and reporting on data. For example, do they use data warehouses, data lakes, data marts, or data lakehouses?
3. Recognize when data is updated at the organization, and what slowly changing dimension is set when dimensional data is changed.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 3

Understanding Types and Characteristics of Data

LESSON INTRODUCTION

As data analysts, we must be able to identify the types of data we are working with, as this controls what we can do with the data. At the start of a project, our first goal is to determine the type of data at a high level. Then we approach the data at the most granular level, which is the field data type. We also must be able to control the data type to meet the needs of our analysis. The better you understand the different types and characteristics of data, the better you can approach each data set you work with.

Lesson Objectives

In this lesson, you will do the following:

- Understand types of data.
- Break down the field data types.

Topic 3A

Understand Types of Data

EXAM OBJECTIVES COVERED

1.2 Compare and contrast different data types.

The two major high-level types of data are qualitative and quantitative. A core understanding of each data type is crucial, because the type of data you're working with controls what you can do with it and how you approach it. Here we will dive into what differentiates the two types of data and explore their characteristics.

Quantitative Data

Quantitative data is defined by the fact that it is portrayed through numbers, meaning it can have a numerical value (like price) or be measured (like weight, distance, or height). There are two main characteristics of quantitative data: discrete and continuous. The general rule is that **discrete data** can be counted and can only take on a certain number of values, while **continuous data** can be measured and can use any value.

For example, consider a survey that asks how many people live in a household. The number that respondents supply is discrete; the person responding can definitively count the number of people living in their house. Another point to consider is that discrete data is generally a whole number—a household can't contain 2.5 people. When considering quantitative data, ask yourself if the data be counted. If it can, then it is likely discrete. Discrete data can only take on a certain number of values like the number of people in the household, where continuous data can take on any value.

Now suppose that same survey asks for the age, height, and weight of each person in the home. This quantitative data is continuous because the values can be measured. Continuous data can also be divided to be even more precise; for example, you could describe a person's age right down to the millisecond.

When working with quantitative data, you can identify its characteristic by determining whether the data is something you can count (discrete) or something you can measure (continuous).

Qualitative Data

Qualitative data is defined by qualities of the data. Qualitative data is also known as categorical data because it can be arranged into groups/categories based on these qualities, but these qualities in and of themselves do not have a number value. Let's go back to our survey example from earlier. We are asked for the number of people who live in the home; this is quantitative data. Suppose, however, that the survey next asks us to provide each person's favorite color. This data is qualitative because we are only able to count the person who answered that particular color, but not the color itself. The color red, for example, isn't greater than or less than green. We can, however, add counts or averages to qualitative data. For example, we could count how many respondents said they liked a specific color and then apply averages to the counts for each category. Qualitative data consists of any text-based quality that does not inherently have a numerical value.

To dive deeper into the discussion on qualitative data, we need to understand the characteristics of this type of data: nominal and ordinal. **Nominal data** is information that has no natural order. In our survey example, the colors red, blue, and green are examples of nominal data. The data is categorical in nature (the answers are all colors) but does not have a natural order (one color is not greater than or less than another). Even though there is no natural order, the answers are different from each other: red isn't blue, and blue isn't green. **Ordinal data** is information that follows a natural order. Ordinal data is very similar to nominal data, but one key difference is that it has a naturally occurring order. Let's go back to our survey example. Suppose after being asked for the number of people in the household, we are asked to provide the grade level in school for anyone under the age of 18. These responses would be ordinal because there is a natural order in which we obtain schooling. Children might skip a grade, but it doesn't change the naturally occurring order.

The easiest way to distinguish if a category is nominal or ordinal is determining whether there is a natural order.

Why the Data Type Matters

So why does any of this matter to the analyst? The type of data you're working with determines how you work with it. When you see categories of information arranged in a list or displayed in a chart, this data is categorical, or qualitative. To make qualitative data suitable for analysis, you will need to add counts, averages, or sums next to each category of information.

Categorical data (qualitative) will also be used to group data into variables for evaluation and statistical analysis to find insights on that category. If we are studying groups of people and their performance on a test. We may want to group them into categories like their gender, race, or other groupings that are needed for the analysis.

It's important to remember that we frequently mix qualitative data with quantitative data for the purposes of analysis and visualization. We measure the color red (qualitative), but we can add a count of people who prefer that color (quantitative).

As analysts, we work with both qualitative data and quantifiable data to highlight areas of focus and uncover possible issues in the data. Understanding the high-level data types, and being able to use both, helps drive decision-making.

Review Activity:

Types of Data

Answer the following questions:

1. **What two types of data are at the highest level?**

2. **Suppose a survey asks respondents for their level of education. This would be considered qualitative data with which characteristic?**

3. **If you ask people to tell you their favorite tv show, what type of data are you collecting?**

4. **When determining how much product was ordered over time, the quantity ordered is which high-level type of data?**

5. **When data can be measured, which characteristic of data is it?**

Topic 3B

Break Down the Field Data Types

EXAM OBJECTIVES COVERED

1.2 Compare and contrast different data types.

Each field of information in a table has a field name (which may or may not be meaningful) and a data type. No matter how you acquire your data, whether through spreadsheets, databases, or data warehouses, at some point you will be met with the data types for an individual field. Understanding which data types exist at a field level, and what they will or will not do, is critical for the analyst.

Introduction to Field Data Types

The field data type controls what type of information can be stored in that field. There are several different core data types: numbers, text, and dates represent most of the data that we work with. At the point you begin working with any given data set, you will want to first identify the field data types. These data types also have different attributes or characteristics, which we will explain next.

A field in a table can only be one of several data types, but the different software programs data analysts use can break these types down and surface even more attributes of the data.

Text/Alphanumeric Field Data Types

The most versatile of all the field data types for design purposes is alphanumeric. You may also see this referred to as “char,” “text,” or “string”; however, these all reference text data types. When this data type is set as a text data type it is alphanumeric which allows the field to store any type of letter and/or number within the field.

Text field data types hold qualitative data (categorical) which makes them great for grouping information. When we count these text-based fields they take on a fixed number of values. As a data analyst, you can have half of your results select a favorite color of red, but a single person can’t “half” select a value.

Its also important to note that what we see from a human perspective and what is defined by the database will not always be in sync.

As an example, consider the postal code. When you see 35007-0023, your first thought would likely be that this data type is a number. However, the dash that lies between the first five digits and the last four digits requires this field to be set as an alphanumeric data type. When conducting data analysis, you will spend quite a bit of time identifying the actual, defined data types controlled by the database and confirming what needs to be done with each field in order to use it in reporting. Unfortunately, what makes the alphanumeric data type versatile for design purposes makes it harder to work with as an analyst. Because the alphanumeric data type is capable of catching everything, you may find that a lot of fields you expect to be a different data type are actually stored as text.

Alphanumeric data types are character data types. They may also be described as text or string in most systems. The data type name itself is dependent on how that particular software names that data type, but they all refer to the same text strings.

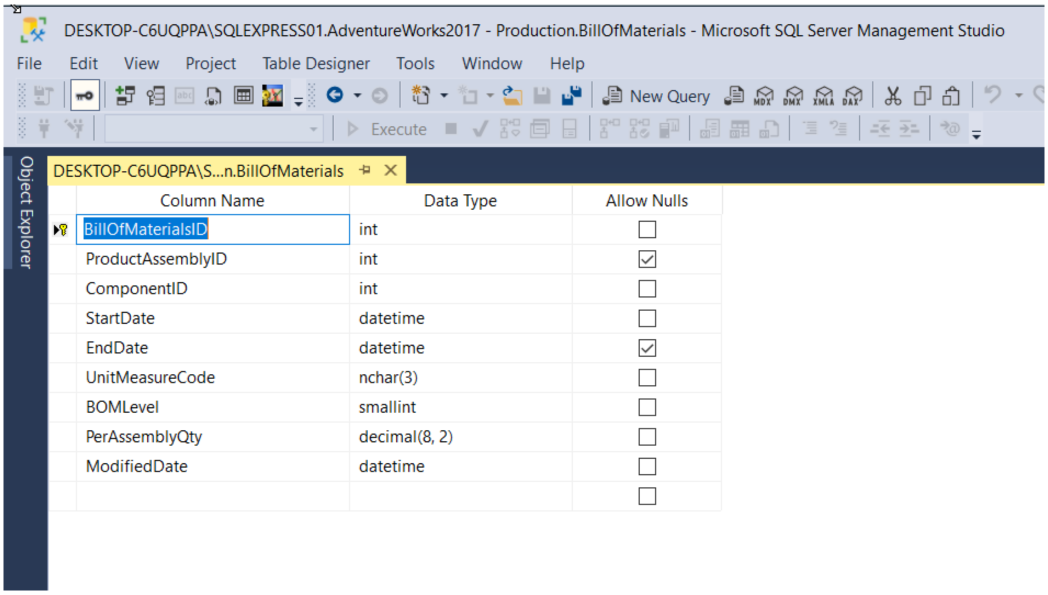
Date Data Type

The date field data type is exactly what it sounds like; it captures dates. However, this data type also serves another purpose. Fields that are defined with a date data type can be used to not only capture a date and/or time in a field, but also to calculate other dates or other information from a date. For example, we can use the field that has a date field data type to pull information such as the month number, week number, or even day of the week. Suppose a company wants a report notification to be sent to the sales manager as soon as each order is placed, but the product ordering system was not designed to do this. If the order date is set as a date field data type, it can be used to calculate when that notification should be sent. This is why it's so important that a database is set up with the proper field data types; there are different steps that we need to perform when reporting that are just not possible if the correct data types have not been used for the fields.

Number Data Types

At its most basic level, the number data type will not allow any text, as it only stores numbers (as you might expect). Use of the number data type allows for calculations. It is very important to use the number data type if you need to do any basic arithmetic in your reporting. Addition, subtraction, multiplication, and division all require the number data type.

The number data type has a variety of different attributes that can be applied depending on the type of number that is being stored. Each database system or data software will provide a variety of number data types, but they will all represent numbers.


DESKTOP-C6UQPPA\SQLEXPRESS01.AdventureWorks2017 - Production.BillOfMaterials - Microsoft SQL Server Management Studio
File
Edit
View
Project
Table Designer
Tools
Window
Help
New Query
MDX
DMX
XMLA
DAX
Execute
Object Explorer
DESKTOP-C6UQPPA\S...n.BillOfMaterials

Column Name	Data Type	Allow Nulls
BillOfMaterialsID	int	□
ProductAssemblyID	int	☑
ComponentID	int	□
StartDate	datetime	□
EndDate	datetime	☑
UnitMeasureCode	nchar(3)	□
BOMLevel	smallint	□
PerAssemblyQty	decimal(8, 2)	□
ModifiedDate	datetime	□
		□

Design View of the Table Production.BillOfMaterials from AdventureWorks2017 in Microsoft SQL Server Management Studio (Used with permission from Microsoft.)

In this image, we see the design view of the BillOfMaterials table in AdventureWorks2017. If you review the data type column for the ID fields, you will see “int,” meaning integer. The field called BOMLevel is assigned “smallint,” meaning small integer. Its important to remember that database design considers storage and size. The difference between int and smallint for TSQL is the difference in storing four bytes of data in the field (int) versus two bytes of data in the field (smallint).

If your number data is stored as a text data type, then you will need to convert it to a number in order to use it within mathematical calculations.

Currency Data Type

The currency field data type is, of course, a number. It can be a number data type that is simply formatted to be displayed as currency with a currency symbol and decimal places, or it could be a field that is assigned a currency data type. When you are dealing with currency, it may be preferable to use the currency data type over a basic number data type formatted to look like currency.

Defining numbers as currency, with field data types like “money” or “smallmoney,” enables the use of currency symbols for a particular geographical area. This is especially important for organizations that serve a worldwide audience and thus conduct international sales across countries that utilize different currencies.

Boolean Data Type

The Boolean data type is commonly used for Yes/No, True/False, and On/Off responses. It is represented by a number that is either 1 or 0. Each system will have its own defined way of working with the Boolean data type.

You may also see data that uses the Boolean data type expressed as 0 or 1, and in some data software may appear as a 0 or −1.

Data Type Conversion

Field data types are controlled by the database’s design, and thus their usefulness hinges on the database designer recognizing the varying data types and understanding the importance of being intentional with the choice of each field data type. However, sometimes the database designer does not set the most suitable field data types. When this occurs, as an analyst you may need to convert field data to different types. For example, if a database designer stores everything as text, you will likely need to spend time converting the data types as needed for the various reporting requirements. Suppose you need to perform mathematical functions with this data (such as addition or subtraction). You can’t calculate alphanumeric data mathematically, so you would need to use a conversion function to change the field data type to a number.

It’s important to remember that database designers and data analysts have different objectives. The designer’s goal is to optimize storage and design, whereas the analyst’s goal is to be able to easily and accurately report data. With that said, when a data type isn’t as you need it to be, you will need to convert it to the data type that’s appropriate to meet your requirements. For example, if a field containing a date is not set as the date data type, and you need various date calculations, then you will convert the field to the data type. We will address how you actually accomplish this task later in the course.

Review Activity:

Field Data Types

Answer the following questions:

1. **Why are field data types important to the data analyst?**

2. **Which field data type is the most versatile of all?**

3. **What feature controls what type of information is stored in a field?**

4. **Who defines the field level data type in the source system, and what can an analyst do if it is incorrect for the type of analysis required?**

Lesson 3

Summary

After this lesson, you should be able to recognize the type of data you are working with and understand the field data type. Data types matter to all data professionals, regardless of the role you play. You should recognize data that can be counted, summed, or averaged as quantitative data. You should also understand that categorical data can be grouped for further analysis and visualization. When assessing data types at the field level, it's important that you are able to identify when the field data type is not suitable for your reporting needs, or when it fails to meet the standards that are set for your organization, so that you may transform the data type accordingly.

Guidelines in Understanding Types of Data

Consider these best practices and guidelines when familiarizing yourself with the different types of high-level and field data you will be working with.

1. Spend time discovering the data types that are assigned to the field data.
2. Document potential conversions of data that are needed (e.g., dates stored as text need to be converted to dates).
3. Remember that field data types might be identified differently based on the system that is reading the data.
4. Recognize that, as a data analyst, you typically can't change the source system, but you can control the data types within the software you use to interact.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 4

Comparing and Contrasting Different Data Structures, Formats, and Markup Languages

LESSON INTRODUCTION

Data can either be structured or unstructured. How we interact with that data varies: it might be through direct connections to the data or through exports of varying file formats. We will also likely encounter and work with data that has been marked up with different languages and standards, like XML and JSON. It is important to understand that the way we work with data is typically dictated by the way the data was designed.

Lesson Objectives

In this lesson, you will do the following:

- Differentiate between structured data and unstructured data.
- Recognize different file formats.
- Understand the different code languages used to read data.

Topic 4A

Differentiate Between Structured Data and Unstructured Data

EXAM OBJECTIVES COVERED

1.3 Compare and contrast common data structures and file formats.

It's important to understand that not all data is equal. The employee information captured by an HR system is vastly different from data in a social media platform (captured by a social media app). For example, the HR system is structured to account for all employees and their hire data information. However, while a social media post might have some structured information, like the date of the post and the user, if a video or image is added, it becomes unstructured data. We have thus far focused mostly on structured data—data that is structured into tables in a column and row format. But the world is also full of unstructured data. With the explosion of mobile devices and other tools that allow us to create videos and images with a click or a swipe, the world of unstructured data has fully integrated itself into our everyday lives. As a data analyst, you will likely work with both structured and unstructured data to some degree.

Structured Data

When our data fits nicely into tables, we know we are working with **structured data**. Just as the name implies, structured data has structure. Consider fields of information that are easy to distinguish, fit neatly in a schema, and can be restructured to fit into other structured data systems (like data warehouses). When you think about databases and spreadsheets, you can immediately recognize structured data. We previously discussed different types of field data, such as numeric, alphanumeric, and date data types. These are all types of structured data.

Most people work with structured data at some point in their lives. When you view an Excel file or work in a SharePoint list, the information you're seeing has a structure and data type. That is structured data.

For the data analyst, working with structured data (or putting data into a structure) is an everyday part of the job. In an earlier lesson, we learned about relational databases. A relational database is the most traditional store of structured data, but it is not the only type—think of all the software you might use today that allows you to manually key in information. Even the information you provide when setting up a free email account is an example of structured data.

Unstructured Data

Unstructured data is data that is not organized in a predefined manner to meet the standards for structured data. Unstructured data does not fit neatly into tables, but instead has **undefined fields**. It is important to consider that there is far more unstructured data in the world than structured data. It cannot be stored in a traditional relational database and instead is stored in a non-relational database. If we want to work with unstructured data, we may use a data lake or blob storage. It is important to remember that there is still some structured data around unstructured data.

BLOB Storage is a framework designed to store large amounts of unstructured data, like video, audio, and text, and more effective for storage versus a database.

Let's start with the most basic unstructured data type: image data. Image data is a prime example of unstructured data because while we can structure the data around an image, to truly know what's in the image we must view it. Consider the CAPTCHA security measure, which requires us to identify what's in a picture in order to log it into a website. The security measure knows we are not a machine when we respond accurately, because while a machine can read some data around an image, it can't look at the image like a human does.

Videos are another type of unstructured data. While there is structured data surrounding the video—such as the title, running length, time of upload, and who uploaded it—we must actually watch the video to know what it is about. The same goes for audio;it must be listened to in order to gain full context. Other types of unstructured data include documents and even PDFs. These data types, while not the type of data that you commonly find and that is easily structured into databases, are still data.

There are also some significant challenges associated with the storage of unstructured data. The size of an image file is based on the quality of that image. An audio file's size is based on both its quality and its length, and the same goes for video files. As you can imagine, finding space to store long, high-quality videos may be difficult for some organizations.

There are many examples of unstructured data in our everyday lives. For example, consider the reactions people have to a social media post. A company may post an image, video, or text to social media and want to know what the public thinks of their post. This data is important because the company can gauge how well a product or service might sell depending on how people react to the post. What people think of a post is unstructured data. To capture this data in its most basic form, a company would have to hire people to read every post and report on the information. Therefore, many social media sites have a reaction or "like" feature that allows companies to more succinctly determine what people think of the post. While what people think about a post is unstructured data, the number of "likes" a post gets is structured data. If you really think about it, you are likely to realize that there is far more unstructured data in the world than structured.

Unstructured data and structured data may have crucial differences, but both are useful in their own way. Consider the following scenario, in which we see both types of data being used. Suppose an organization that is hiring new employees receives handwritten applications and paper copies of the applicants' résumés. Someone who works at the company scans that documentation into a system that converts the paper documents to PDF files and stores them. These two documents are unstructured data. The staff member of the company then creates a record in the Application Interview system, filling out the applicant's name and contact information. The applicant's PDF resume can even be attached to the record. The information entered into fields in the Application Interview system, such as first name and last name, is structured data. Thus, we are able to take unstructured data, in the form of a PDF resume, and turn some of it into structured data.

There will always be unstructured data in the world. We as people produce a huge volume of unstructured data daily. Consider videos; many years ago, the only way to store video was on VHS tapes, and the only way to read VHS tapes was with a VCR player. Now almost anyone can record a video on their phone and upload it to the internet, where it can be watched on numerous devices. Unstructured data is here to stay, which is why we are always seeking ways to make it structured.

Semi-Structured Data

Semi-structured data is a mix of both structured and non-structured data. Emails, JSON files, XML files, zipped files, and even web pages (HTML) are examples of semi-structured data. Semi-structured data does not meet the rigid standards of structured data, so it cannot be stored in a relational database. However, semi-structured data does contain some structure through the use of tags and attributes that group the data and describe its storage methods. These tags allow us to search for and view the data, and also help us to transform unstructured data into more structured data sets.

Review Activity:

Structured and Unstructured Data

Answer the following questions:

1. **Would data that is stored in a relational database be considered structured or unstructured data, and why?**

2. **Video, audio, and images are all examples of what type of data?**

3. **Is most data in the world structured or unstructured, and why?**

Topic 4B

Recognize Different File Formats

EXAM OBJECTIVES COVERED

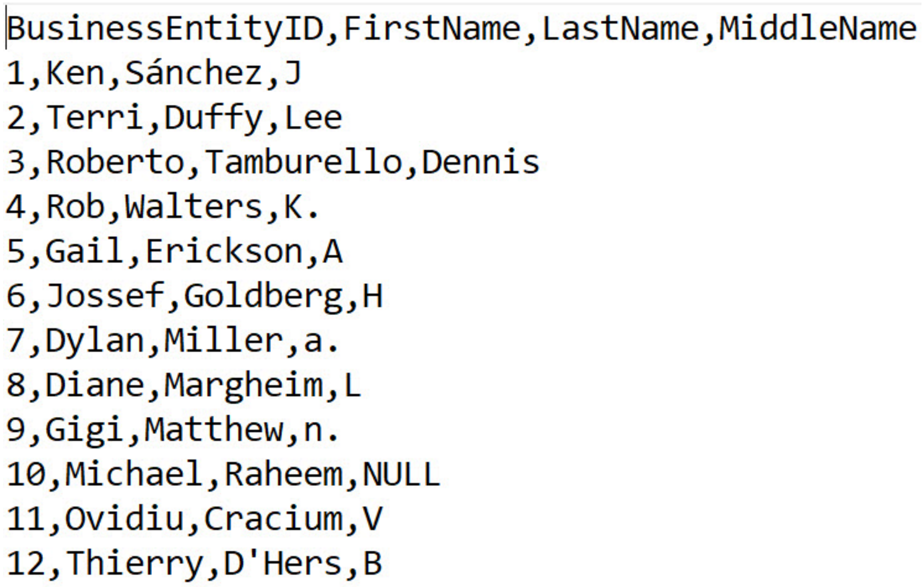
1.3 Compare and contrast common data structures and file formats.

Based on what we've covered so far, or maybe from your own experience, you've likely realized that not all data comes to us in the form of a database. One of the key skills a data analyst must have is the ability to work with data in any format. It is extremely important to have a sound understanding of the file formats typically used for most data, as these formats will control how you will work with the data, what you will need to do to maintain the data, and what tools you can use with the data. When you are working with data in various file formats, you will need to import and export data. Importing data means you are bringing data into a system, and exporting means you are sending it out. Most software systems will allow you to export information, and the resulting file is typically formatted as a delimited file.

Delimited Files

Delimited files are files in which some form of character separates each field of data from the other data fields. If you have ever opened a text file and found it was full of commas, or even weird-looking spaces, this file was likely downloaded from a data system. The commas and spaces are characters intended to support the conversion of the data into a structured file format.

The most common type of delimited file is comma separated values (CSV). In this file format, each field of data is separated by a comma. If you were to open up a CSV file in Notepad or a text editor, you might see something that looks like this image.



```
BusinessEntityID,FirstName,LastName,MiddleName
1,Ken,Sánchez,J
2,Terri,Duffy,Lee
3,Roberto,Tamburello,Dennis
4,Rob,Walters,K.
5,Gail,Erickson,A
6,Jossef,Goldberg,H
7,Dylan,Miller,a.
8,Diane,Margheim,L
9,Gigi,Matthew,n.
10,Michael,Raheem,NULL
11,Ovidiu,Cracium,V
12,Thierry,D'Hers,B
```

Comma Delimited File Opened in Notepad (Used with permission from Microsoft.)

In this image, the first line represents the column headings from the original CSV file.

It's important to remember that the data will not always come with headings.

When you look at all the lines in the image, you can see that each field is separated by a comma. The page break forms the notation for the record. For example, you will notice there is no comma after "MiddleName" but the next line begins with 1. Then there is a comma between each field of information before it moves on to the next line, and so on. This is what is considered a delimited file, and in this case, the delimiter is a comma.

There are multiple types of delimited files and formats. If you have a pipe delimited file, that means the | symbol separates fields of information. If you have a TAB delimited file, then the delimiter is a TAB. These are the most common delimited files, although you could see other types of characters.

Why We Use Delimited Files

Working with CSV or Tab delimited files has grown easier over the years. While it wasn't always the case, tools like Excel will now automatically understand and interpret the data in a delimited file. If you have Excel installed on your machine and you open a CSV file, the file will automatically open in Excel, as opposed to a text editor (like Notepad or Wordpad). When you attempt to open a file that's in a format Excel doesn't recognize, the software will present you with an import option that lets you define the different information that is applicable to that file.

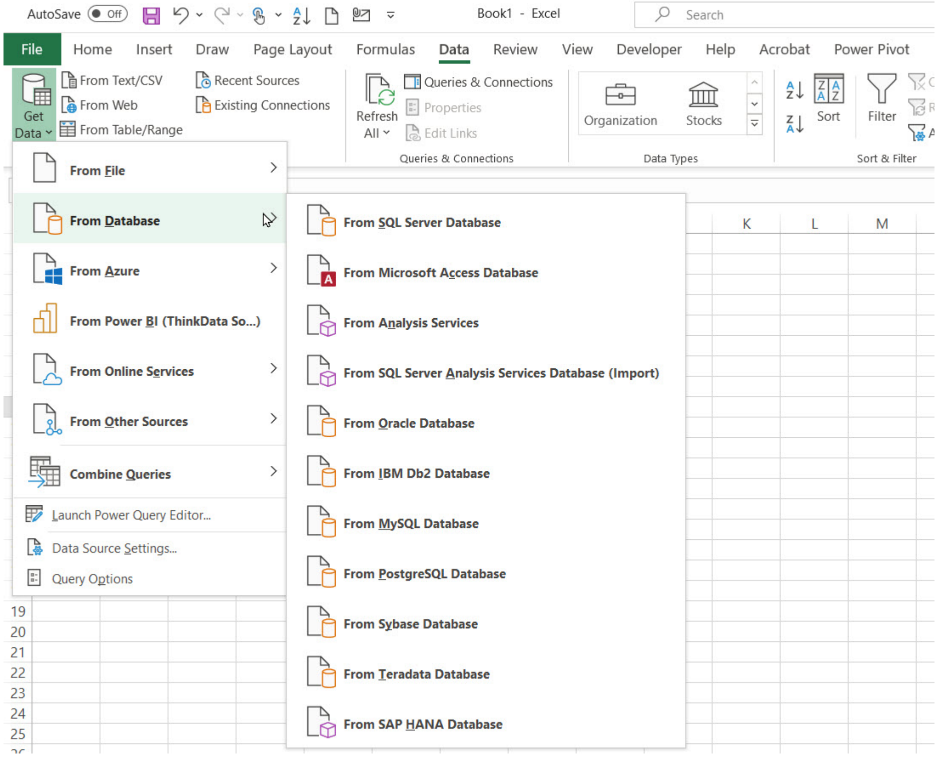
You can import the data in delimited files into Excel, and you can also export data out of Excel to save as other formats. When you are working with an Excel file, you would click the "Save As" option and choose the file type that you want to save the data to.

When you are exporting data out of Excel into a delimited format, with the plan to transfer the data to another system, the system you are transferring it to will define the requirements for you.

Flat Files

Delimited files that are exported out of a system are **flat files**. Flat files are not connected to the database; when the database updates, the flat file does not update. The flat file will continue to display the data as it was when you exported it. In order to see the updated data, you will need to export the file again.

These days we have more capability to be connected to live data in real time. Many of the tools we use to connect to the data systems that store data provide real-time information, removing constraints of working with flat files. You will notice that different tools have different connections. For example, Excel can connect to an SQL database that you have permission to access, or you can connect to a SharePoint list to gather data.



Microsoft Excel O365 Get Data Command (Used with permission from Microsoft.)

This image shows the various database connections that are natively available to Excel.

There are also flat file databases where data can be structured into records that follow a uniform structure or design, but no relationships exist within the database; you need to manually look at the data to determine how to relate it. There are plenty of flat file databases in the world, the most popular of which are mainframe databases. Most of the flat file databases you hear about today were implemented decades ago. These flat file databases don't have the same flexibility as a relational database, but for some organizations, modernizing these databases is a major undertaking. Changing out these older databases for relational (or even non-relational) databases is extremely costly and runs the risk of interrupting processes that rely on the databases.

File Extensions

There are several different types of flat or delimited files. The table shown provides the file extensions and a brief description for the most common of them. The system that generates the flat file will either supply the file with the extension or give you an option to export in a certain format. When you plan to import a delimited file into a particular system or software, it's important that you first identify which file type that system can read, and then save your file appropriately.

File Extension	Information
.CSV (comma separated value)	Comma separates the values
.CSV (comma separated value) with UTF – Encoding	Comma separates the values; this extension is preferred when dealing with web-based coding
.TAB (tab delimited)	Tab separates the values
.TXT (tab delimited)	When saving from Excel as a tab delimited file, Excel will generate the delimiter and line breaks as needed but use an extension of .TXT
.TSV (tab delimited)	Tab separates the values

Common Delimited File Extensions

This is certainly not an exhaustive list, but the key takeaway is that different formats exist, and different programs read or require certain file types. If your file is not already in the acceptable format, you will need to convert it.

Review Activity:

File Formats

Answer the following questions:

1. **What are some common delimiters you will likely encounter in text-based files?**

2. **What is the difference between importing and exporting?**

3. **What is a major drawback to working with CSV files?**

Topic 4C

Understand the Different Code Languages Used for Data

EXAM OBJECTIVES COVERED

1.3 Compare and contrast common data structures and file formats.

As a data analyst, you will encounter different languages that will impact the way you work with data. Some languages are dedicated to structured data, such as SQL. There are also markup languages that work with unstructured data, but these first need structure in order to be read by websites, web applications, and even other data systems.

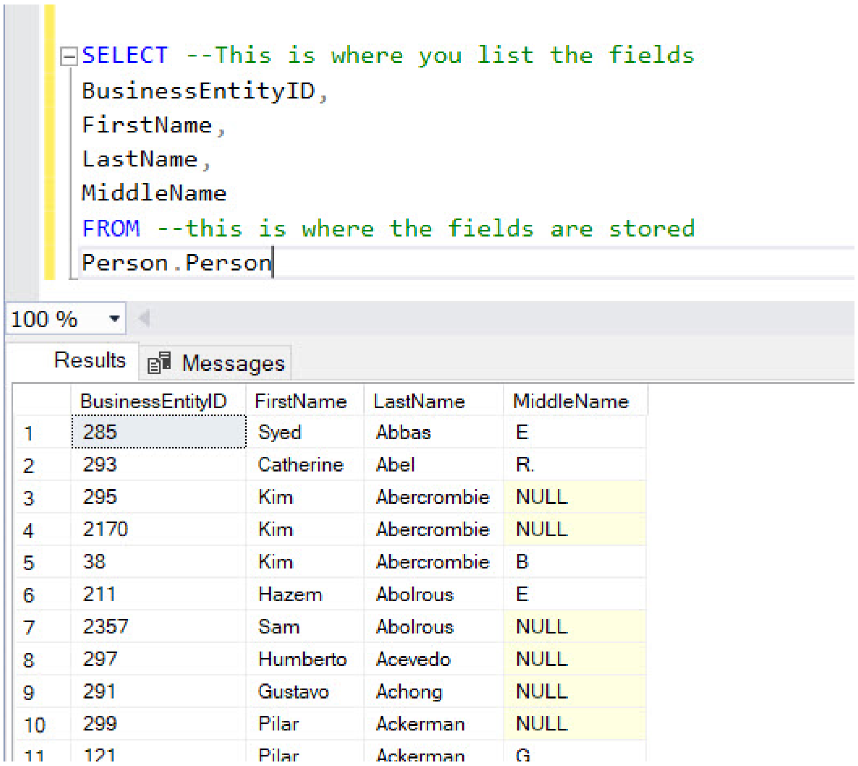
Structured Query Language (SQL)

The most comprehensive language is SQL, as this language provides syntax that is dedicated to working with data. One of the primary skills of a data analyst is the ability to query data, and just like the name implies, that is what SQL will allow you to do.

SQL statements provide information to databases (SQL databases, or even data warehouses) that allow data analysts to retrieve data from tables stored in the database. SQL statements not only query data, but also allow you to perform actions on the data. How extensively you use SQL as an analyst depends on your particular job role within the organization where you work. For most analysts, knowing how to use the basic SQL statement is an important first step.

SQL has a required syntax, but it is also important to understand that different versions of SQL do exist depending on the system you are using. For example, T-SQL, which stands for Transact – SQL, is used in software developed by Microsoft. Oracle also uses SQL, but it uses PL/SQL, or Procedural Language/SQL. When searching for syntax examples, it is helpful to look for the version of SQL that your particular software uses.

Let's look at the universal SELECT statement, which has two required key words to query data: SELECT, which should list the fields you are interested in accessing, and FROM, the name of the table where the fields are stored. This particular statement selects information from a table called Person.



```
SELECT --This is where you list the fields
BusinessEntityID,
FirstName,
LastName,
MiddleName
FROM --this is where the fields are stored
Person.Person
```

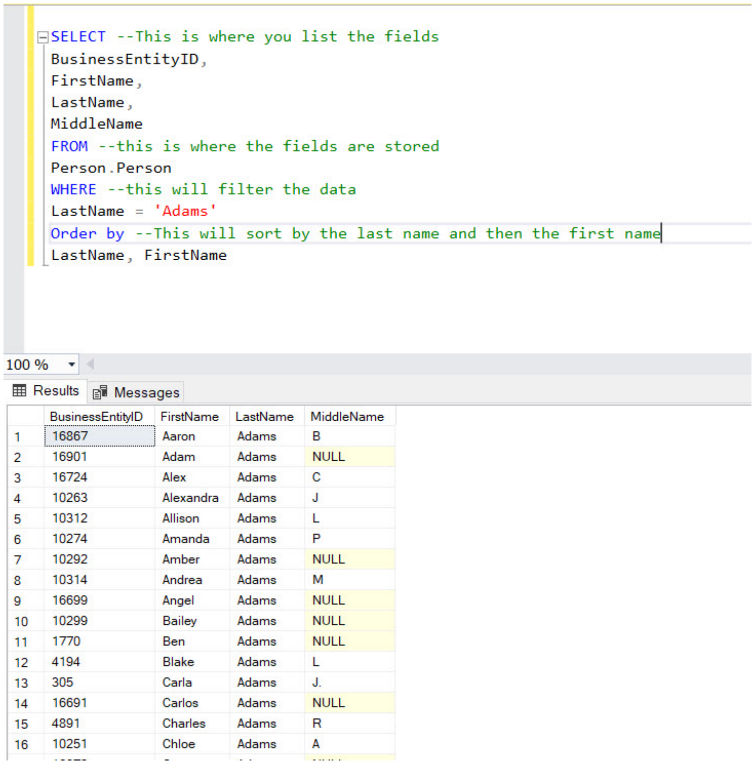
	BusinessEntityID	FirstName	LastName	MiddleName
1	285	Syed	Abbas	E
2	293	Catherine	Abel	R.
3	295	Kim	Abercrombie	NULL
4	2170	Kim	Abercrombie	NULL
5	38	Kim	Abercrombie	B
6	211	Hazem	Abolrous	E
7	2357	Sam	Abolrous	NULL
8	297	Humberto	Acevedo	NULL
9	291	Gustavo	Achong	NULL
10	299	Pilar	Ackerman	NULL
11	121	Pilar	Ackerman	G

Basic SQL Select Statement in Microsoft SQL Server Management Studio (Used with permission from Microsoft.)

Additional statements can be added to the SELECT statement that will allow you to filter the data: the WHERE key word is first used to filter data, and the ORDER BY statement then sorts the results in ascending order by default. (To sort in descending order instead, you must specify that by including the desc after the ORDER BY and FIELD key words.)

The statements used to filter data should always occur in the same order; the WHERE key word cannot be placed after the ORDER BY key word. Key word order is important because the data system needs to filter the requested data before it sorts the data it's displaying.

This next SQL statement selects fields from the Person table, but only if the last name is Adams, as we have designated. Data is sorted by last name and then by first name.



```
SELECT --This is where you list the fields
BusinessEntityID,
FirstName,
LastName,
MiddleName
FROM --this is where the fields are stored
Person.Person
WHERE --this will filter the data
LastName = 'Adams'
Order by --This will sort by the last name and then the first name
LastName, FirstName
```

	BusinessEntityID	FirstName	LastName	MiddleName
1	16867	Aaron	Adams	B
2	16901	Adam	Adams	NULL
3	16724	Alex	Adams	C
4	10263	Alexandra	Adams	J
5	10312	Allison	Adams	L
6	10274	Amanda	Adams	P
7	10292	Amber	Adams	NULL
8	10314	Andrea	Adams	M
9	16699	Angel	Adams	NULL
10	10299	Bailey	Adams	NULL
11	1770	Ben	Adams	NULL
12	4194	Blake	Adams	L
13	305	Carla	Adams	J.
14	16691	Carlos	Adams	NULL
15	4891	Charles	Adams	R
16	10251	Chloe	Adams	A

Basic Select with Filter and Sort Statements in Microsoft SQL Server Management Studio (Used with permission from Microsoft.)

The order in which you lay out your fields does not have to match the sort order. The SELECT statement will control the layout order of your results, but the ORDER BY key word controls how the fields are sorted. These functions are independent of each other.

SQL also contains the syntax and key words to perform actions on data. You may need to use these techniques over time as you progress in your career, or they may be completed by the database administrators in your organization. Such actions might create tables, update fields of data, or append (i.e., copy) data from one location to another.

As you grow in your role as a data analyst, you will learn even more about the various features of SQL. In the early stages of your career, however, it's important that you simply understand how to read the statements, even if you aren't familiar with how to key the code. Like everything else, learning SQL takes practice and exposure.

HyperText Markup Language (HTML)

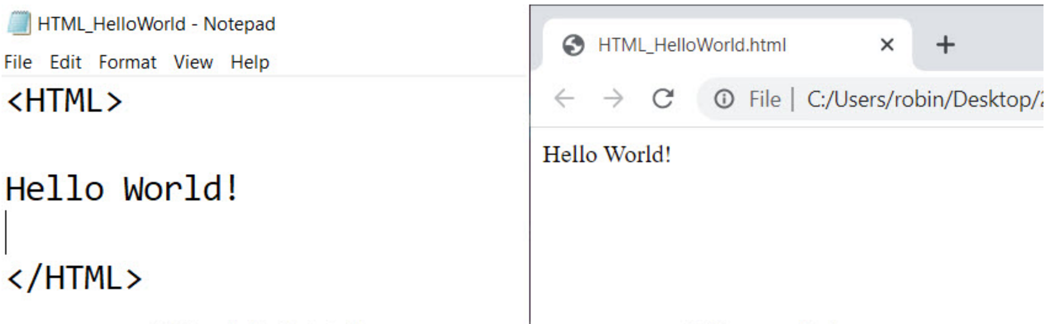
Created by Tim Berners Lee in the early 1990s, **HyperText Markup Language (HTML)** is a language dedicated to presenting data in a browser-based environment. We all experience the benefits of HTML every time we look at a web page on the Internet. HTML is derived from **Standard Generalized Markup Language (SGML)**, which is considered to be the parent of all markup languages. SGML provides the standard that defines all markup languages and is also widely used for data structures.

HTML is considered a markup and tag language. It allows you to mark up any type of text for presentation on a web browser, which makes HTML suitable as the language of the web. The browser reads the HTML and produces a web page or website following the structure that is laid out in the code.

HTML has a set of predefined tags based on the HTML language (e.g., HTML, TITLE) so that when the browser sees those lines of code, it knows how to present that information and where. For example, you can create a text file with HTML tags and then open that text document in the browser to see the results.

It might be hard to imagine that as recently as 20 years ago, we developed web pages with very limited tools dedicated to designing the pages of the World Wide Web—we would literally code the HTML. Now there are many tools that use a graphical user interface (GUI, pronounced gooey) to design web pages, transforming the points and clicks you make into HTML for use in the browser. The original GUIs for HTML were FrontPage and Dreamweaver; WordPress is a well-known example that's used today.

The syntax of HTML involves the use of tags in the form of angle brackets to mark up a document intended to be displayed on a web browser. These tags typically consist of an opening tag (< >) indicating the start of an element and a closing tag (/ < >) indicating the end of the element, with the text content of the element enclosed between the two tags. As an example, the screenshot below shows a pair of these tags, both in the text document and how they appear in the browser. As you can see, the text between the tags is presented in the browser.


HTML_HelloWorld - Notepad
File Edit Format View Help
HTML_HelloWorld.html
File | C:/Users/robin/Desktop/

```
<HTML>

Hello World!

</HTML>
```

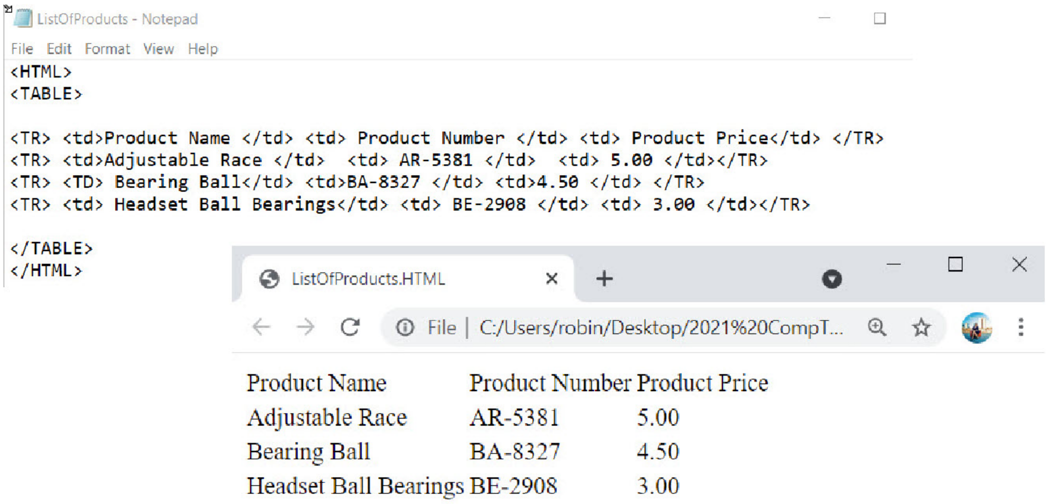
Hello World!

HTML coded in NotePad

HTML opened in browser

Text Editor and Browser Side by Side

For a more involved example, let's create a product-pricing table in HTML for display on a company's website. To create and organize an HTML table, the table must be declared with a <Table> tag. Each row of the table is designated with a <tr> tag, and each cell of data with a <td> tag. HTML presents the data using these simple markup tags, and the browser knows how to interpret and work with these tags, so we end up seeing a structured table of product-pricing information on the web browser.


ListOfProducts - Notepad
File Edit Format View Help
ListOfProducts.HTML
File | C:/Users/robin/Desktop/2021%20CompT...

```
<HTML>
<TABLE>

<TR> <td>Product Name </td> <td> Product Number </td> <td> Product Price</td> </TR>
<TR> <td>Adjustable Race </td>  <td> AR-5381 </td>  <td> 5.00 </td></TR>
<TR> <TD> Bearing Ball</td> <td>BA-8327 </td> <td>4.50 </td> </TR>
<TR> <td> Headset Ball Bearings</td> <td> BE-2908 </td> <td> 3.00 </td></TR>

</TABLE>
</HTML>
```

Product Name	Product Number	Product Price
Adjustable Race	AR-5381	5.00
Bearing Ball	BA-8327	4.50
Headset Ball Bearings	BE-2908	3.00

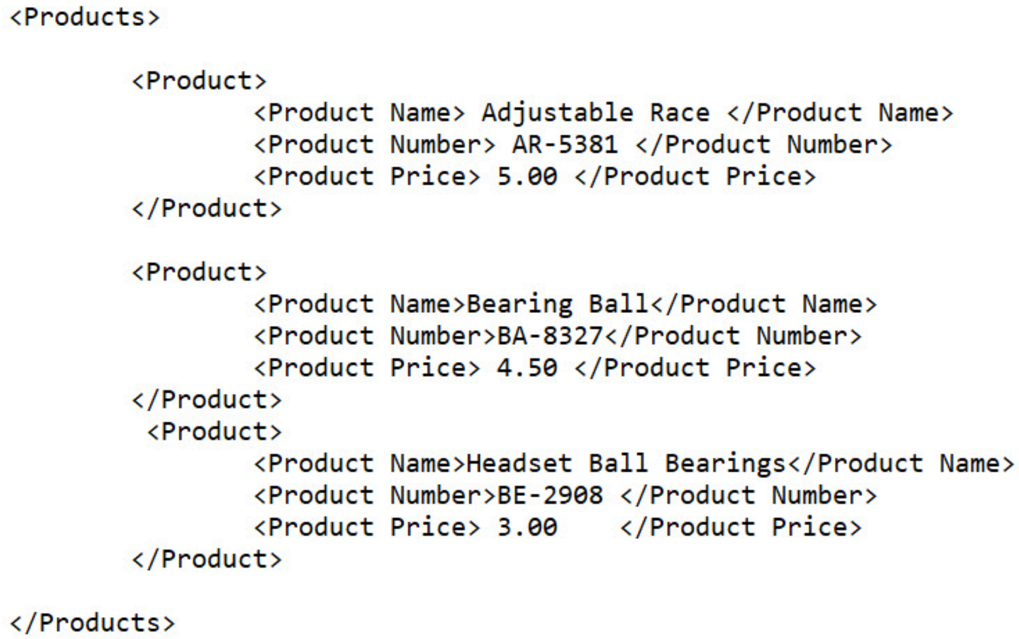
Using Simple Table Tags to Create a Product List

Note that the HTML code does not distinguish between specific elements of the table, such as the first row of cells functioning as the table header, or that the first <td> tag in each row represents a product name. The logical organization of the table becomes more apparent when we read it on the browser.

Extensible Markup Language (XML)

Like HTML, **Extensible Markup Language (XML)** is also a text-based markup language that uses tags derived from SGML. Unlike HTML, the primary purpose of XML is to transfer data, not display it. XML also uses opening and closing tags, but one of the key differences of XML is that the "tags" are not defined by the language itself. Rather, the author can invent the tags and the structure. This ability allows the developer to tag the fields of data with something more meaningful.

As an example, consider the example of the product-pricing table from our HTML discussion. Rather than being limited to just using <TR> and <TD> tags, we can use more relevant tags with XML, as shown in the following example.



```
<Products>

        <Product>
                <Product Name> Adjustable Race </Product Name>
                <Product Number> AR-5381 </Product Number>
                <Product Price> 5.00 </Product Price>
        </Product>

        <Product>
                <Product Name>Bearing Ball</Product Name>
                <Product Number>BA-8327</Product Number>
                <Product Price> 4.50 </Product Price>
        </Product>
         <Product>
                <Product Name>Headset Ball Bearings</Product Name>
                <Product Number>BE-2908 </Product Number>
                <Product Price> 3.00    </Product Price>
        </Product>

</Products>
```

XML Markup of a List of Products

The XML tag now indicates that this is a table containing “products.” Each product line has a product tag with an open and a close, and each field of data is tagged according to the data that it holds. We don’t need to see this content laid out in a table; we can easily read this code and recognize which fields are the product names and which are the product numbers, because each piece of data is tagged appropriately for what it represents.

Developing XML markup may not be a typical task for most data analysts. There are tools dedicated to generating XML.

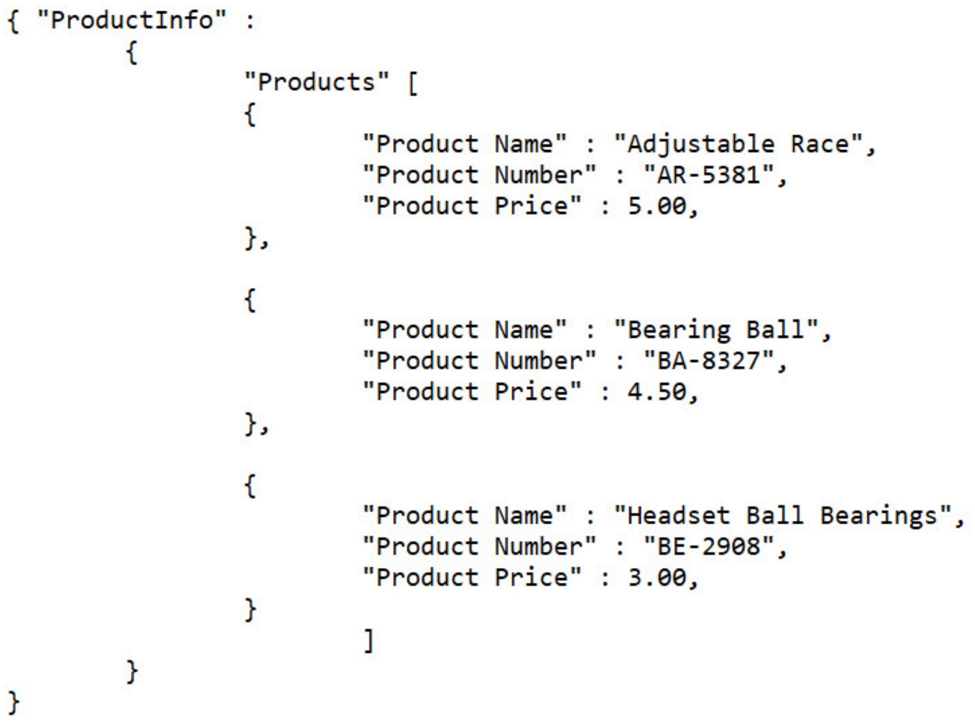
JavaScript Object Notation (JSON)

JavaScript Object Notation (JSON) is an object-oriented, event-driven programming language that allows us to interact with websites (as opposed to HTML, which simply displays information). JSON is a child of JavaScript and is based on the JavaScript language. With the advent of web applications, JavaScript has become one of the core languages of the World Wide Web. JSON is used to transfer information and interact with programming languages like JavaScript.

JSON and XML are alike in several ways. Both utilize tags or keys that are created by the developer and thus more relevant to the data. JSON and XML can both be read, parsed, and used by many programming languages. However, JSON has a more simplified coding syntax than XML, with fewer requirements, and it is not considered a markup language because it is object oriented.

These language concepts are extremely important for software developers, but as a data analyst, you just need a basic understanding of the code languages that you might encounter in your work. This could include XML, JSON, or both, depending on the decisions made by the development team for that software.

Like JavaScript, JSON uses curly brackets for its syntax, and it does not require the opening and closing tags seen in XML and HTML. JSON is written using **key value pairs**. Let's return to our product example. When we write the product list in JSON, the Product Name is an example of a key, and the associated value is "Adjustable Race." The key value is always a text string. The value in the "value" section of the key value pair must be one of the following data types: a string, an object, a number, an array, Boolean, or null.



```
{ "ProductInfo" :
        {
                "Products" [
                {
                        "Product Name" : "Adjustable Race",
                        "Product Number" : "AR-5381",
                        "Product Price" : 5.00,
                },

                {
                        "Product Name" : "Bearing Ball",
                        "Product Number" : "BA-8327",
                        "Product Price" : 4.50,
                },

                {
                        "Product Name" : "Headset Ball Bearings",
                        "Product Number" : "BE-2908",
                        "Product Price" : 3.00,
                }
                        ]
        }
}
```

Basic JSON for Product List

One element that makes JSON coding simpler is the ability to use an array (list of elements). In our JSON product list, the array is the list of the products. An array is noted using square brackets and contains comma-separated elements. The ability to use an array helps make JSON more streamlined than XML, which requires that each product have its own line of code.

Review Activity:

Code Languages Used for Data

Answer the following questions:

1. **Which markup language is dedicated to the presentation of information on the web?**

2. **Which markup language supports the ability to use an array?**

3. **What language is used to write queries in a structured database environment?**

4. **Which markup language supports the transfer of data between systems and is a child of SGML?**

Lesson 4

Summary

After this lesson, you should be able to distinguish between structured and unstructured data. You should be able to recognize different file formats, like comma separated values and tab delimited files. The type of data and file format will affect the analyst's work strategies, so it's important to be very familiar with these concepts. You should also have a basic understanding of the various code languages that interact with or mark up data to allow work within various systems.

Guidelines in Understanding Data Structures, File formats, and Languages

Consider these best practices and guidelines when familiarizing yourself with the different data structures, file formats, and code languages you will encounter when working with data.

1. Recall that structured data is data that is organized and stored in tables, in rows and columns.
2. Recall that unstructured data is data that is not organized in a predefined manner, such as audio, images, and videos.
3. Recognize delimited files, or files in which some form of character separates each field of data from the other data fields.
4. Understand that the most common type of delimited file is CSV, which stands for comma separated values. In this file format, each field of data is separated by a comma.
5. Understand that flat files do not have a live connection to the database and require additional steps when the data is to be refreshed with a new file.
6. Recognize SQL, a key language that allows analysts to query data (Select, From, Where, Order By).
7. Identify the use of markup languages like HTML, a language dedicated to browser-based environment markup and tag language.
8. Identify the use of XML, a text-based markup language with the primary purpose of transferring data (not displaying it).

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 5

Explaining Data Integration and Collection Methods

LESSON INTRODUCTION

When data is created through the various systems of an organization, there is a need for processes that help transform data to make it more meaningful and ensure the data meets any requirements. Extract, Transform, and Load (ETL) and Extract, Load, Transform (ELT) processes make our data consistent. Not all data changes occur in real time, so understanding how data goes from its natural state to the state where it can be loaded into a database or data warehouse is valuable for the analyst. The creation of application programming interface (API) and web scraping is traditionally in the domain knowledge of software developers, but knowing how different data iscollected and passed through systems, and when this occurs, can provide valuable insight to the analyst for their own reporting. Further, the analyst should understand how public data and survey data are collected and used.

Lesson Objectives

In this lesson, you will do the following:

- Make data usable for data analysis and reporting.
- Explain API/web scraping and other collection methods.
- Collect and use public data.
- Use and collect survey data.

Topic 5A

Understand the Processes of Extracting, Transforming, and Loading Data

EXAM OBJECTIVES COVERED
2.1 Explain data acquisition concepts.

Extracting, transforming, and loading data are critical actions in making data usable for reporting and visualization. These processes make data available to users within an organization who just need to use the data for their reports, without needing to know the complexity of the data and how it's connected.

The most common method used to prepare data for a data warehouse is **ETL (Extract, Transform, Load)**. ETL is the process that occurs when moving data from source systems to data warehouses by extracting data from the source, transforming the data, and then loading it to the warehouse. **ELT (Extract, Load, Transform)** is a more modern method that is used when preparing data for data lakes, and this process holds data in preparation for transformation. In both cases, these methods are often performed by the data professionals who are tasked with ensuring data is made available for reporting and are responsible for making data available to the data warehouse or the data lake. However, data analysts perform a similar ETL process by also extracting and transforming data for meeting business intelligence and reporting requirements.

Extracting Data

Extracting data (the "E" in ETL) is the process of accessing the source data from the system and then converting that data into a format that can be transformed and loaded into the data warehouse. How you extract data from a system is controlled by that system itself.

The software (source system) that stores the data has built-in tools to allow you to extract the data from that system. For example, in Microsoft SQL Server, data architects and data engineers are likely using the OLAP technology provided by SQL Server Integration Service (SSIS) to create packages that allow the extraction of data for the purposes of then transforming and loading the data into a data warehouse.

Data analysts use a similar process when extracting data, but they are likely using different tools, such as Power BI, a popular data transformation tool and dashboarding software. As data analysts, we are connecting to the SQL database or data warehouse with the credentials that allow us to gain access to the table(s) needed for our report. We do not create packages like the data warehousing professional, but we are creating a connection to that data for the purposes of transforming it. We would use the transformation tool that is built inside the software (in Power BI, this tool is called Power Query) and then load it to our dashboard.

In most cases, you only need read permissions to extract a data set.

For data analysts using the ETL process, there are a multitude of connections that are built into any of the popular reporting tools. How you connect to "extract" the data you need varies by the tool. Almost all tools will allow you to connect to the common file format CSV (comma separated values).

In earlier years, unless you were the software developer or the database admin in the IT department, you hardly ever had back-end access to the database. You would either export the data you needed, or you would be given the data by someone who could export it for you. The tools and technologies we are given today provide greater access to data than ever before.

Here's a specific example of how an organization might need to extract data as part of a job task or an ETL process for the warehouse. In a database, a company's sales staff are stored in a table with their first and last name, address, and other employee-related information, with one record per person. The organization's customers are stored in a table with one record per customer. When a salesperson sells to a customer, a record of the sale is created and stored in an orders table. To create a monthly sales report, an ETL process must extract the salespeople, the customers, and the orders from the database. These three objects combined provide the necessary data for the report.

Once you have connected to your data, you can begin the next stage of work: transformation of the data to meet business requirements.

Transforming Data

Transforming data (the "T" in ETL) is the act of making the data more meaningful for the purposes of reporting and decision-making. Data transformations will also occur when the data must meet the data quality standards for an organization. Transformation of data can involve many different techniques. The goal is to transition data from its natural state to the format needed for reporting or requirements. Here are some examples of the need to transform data in a business context:

- For customer order data, if the data type of the order date field is actually text, then it is transformed to a date data type.
- If the company needs all the states to be two-letter abbreviations and uppercased text to make that data consistent for end users, then the transformation steps will transform it to meet that requirement.

Transformation in Data Warehousing

In data warehousing, the transformation step allows business users to access data that has already been cleaned and transformed. By implementing a data warehouse and ETL processes, the business removes the need for every user to complete the same redundant steps to transform the data. The data warehouse provides users with a clean data set ready to meet their reporting needs, whatever they may be.

Remember that in the case of data lakes, transformation occurs after the data is extracted and loaded.

Transformation for the Data Analyst

Although data transformation is integrated into the ETL process for data warehousing, data analysts will also perform these same types of transformations as needed when working with data . It's important to remember that not all data we access is neatly stored in a data warehouse, and that some data may not have made it to the warehouse yet. This means that data workers will be actively transforming data as a part of their job.

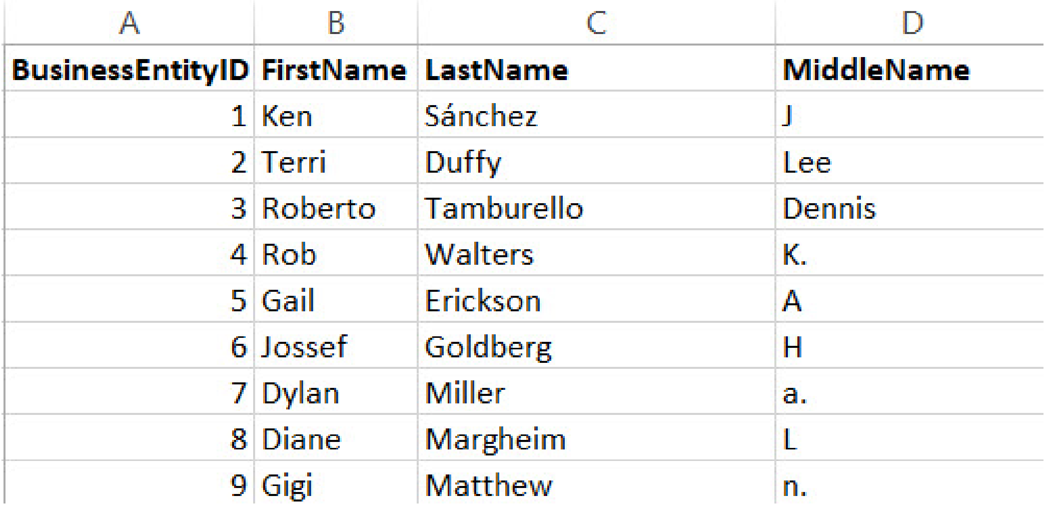
Recall that data is stored under a different set of rules and requirements, such as optimizing the data for storage using the normal forms. However, the way that data is stored is not always optimal for data reporting and analysis, which is why data transformation is needed. When we transform data, we are making it more readable and more meaningful.

Common Examples of Data Transformation

To explore some specific reasons for transforming data, let's consider the example of an HR database that holds information for all the employees. Before employees make it into the HR system, they will have completed their necessary new hire paperwork. As a data analyst, you are tasked with writing a data set that will report on the new hires based on their digital paperwork.

In the data, the first names and last names of the new hires are currently stored in two individual fields. This is ideal in one sense, because it ultimately allows us to sort the information by last name and first name. However, when we want to show this data on a report or dashboard, it is preferred that these names are combined, so when we see the full names they don't include what we call a river of white space. A river of white space is inconsistent, extra spacing between the first and last names of employees that occurs when the names are left as individual fields. This white space happens because the last names all begin at the same starting point, regardless of the character length of the first names.

Data warehouse professionals and data analysts will commonly transform name data to make it more readable, combining the first and last name to display together.



A	B	C	D
BusinessEntityID	**FirstName**	**LastName**	**MiddleName**
1	Ken	Sánchez	J
2	Terri	Duffy	Lee
3	Roberto	Tamburello	Dennis
4	Rob	Walters	K.
5	Gail	Erickson	A
6	Jossef	Goldberg	H
7	Dylan	Miller	a.
8	Diane	Margheim	L
9	Gigi	Matthew	n.

Microsoft Excel with Names Stored in Individual Fields (Used with permission from Microsoft.)

You can see in this screenshot how extra spacing would happen when some first names are only three characters and others are seven characters.

In a data warehouse, first and last names are often stored individually alongside an additional "display name" fieldcreated after data transformation to merge the first and last names. This design allows employees to use this data without needing to know and use the function to combine the names every time.

Continuing with the example of employee names, we should also note the inconsistent format of the employees' middle names in the data shown. Some people have entered their entire middle name, some a capitalized middle initial, and others a lowercase initial. Some middle initials are followed by a period, whereas others are not.

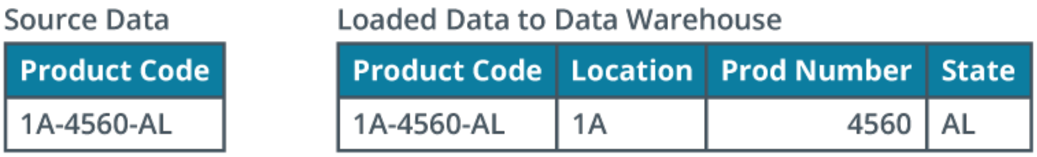
If we want consistency in this data, we will transform the field to only include the middle initial, so that the data (even though entered differently) will all appear in the same format. We would likely use a function to pull only the first character from the left-most side of the field MiddleName, a function to capitalize the initial, and yet another function to create the "." at the end of the initial.

These two examples help demonstrate what it means to transform data in a way that is more consistent and readable. As we get into later lessons on data mining and cleaning, we will spend more time learning more about the transformation process.

Loading Data

Loading data is the process of moving the data into it's target destination, such as a data warehouse. A database architect or data warehouse professional uses tools dedicated to the ETL process based on the software that has been selected by their organization.

Data analysts should understand how data is loaded into a data warehouse using ETL and how this process might change the data. For example, if you are a data analyst working with a source system, and the data from that system is transformed and loaded into the warehouse, the warehouse data may not have a direct one-to-one correlation with the data fields in the source system. As an example, let's consider product codes. The source system stores and formats a product code one way, but after being transformed and loaded into the warehouse, the product code data appears in a different format. This has occurred because the product code in the source system denotes the location, product number, and state, all within the one code. However, once the data is loaded into the data warehouse, these elements of the product code are broken out separately.



Source Data

Product Code
1A-4560-AL

Loaded Data to Data Warehouse

Product Code	Location	Prod Number	State
1A-4560-AL	1A	4560	AL

Product Code Data in a Source System and After Load to Warehouse

The ETL process thereby allows users of the data to see the information in a form that is compartmentalized for reporting purposes, and ultimately more meaningful. Next, we'll explore the different types of data loads used to move data into the warehouse.

Full Load and Delta Load

Different methods can be used to load data into a data warehouse, and the type of load used determines how the data is read from the source systems and in turn placed into the data warehouse. The data warehouse professional makes decisions about what type of load to use, so as a data analyst, you simply need to know what decision was made and understand the impact it has on your data. The two types of loads are as follows:

- **Full load** means you are loading all data into a data system (database, data warehouse, or source software) for the very first time. For example, when a new data system is designed, all historical data from the previous system is loaded into the new system.
- **Delta load** means you are loading new data into a data system and updating any existing data that has changed since the last load. A delta load is how data in the data warehouse gets updated. Decisions about when these loads occur are made by the database administrators (DBAs), the architects of the systems.

The timing of a delta load is important information for the data analyst to know. For example, if the delta load occurs once daily at midnight, then the analyst would not see updates to the data until that time. In this scenario, If a sales order is placed on Monday at 4 p.m., then any sales data accessed from the warehouse will not show the order until Tuesday at 12 a.m.

As this example illustrates, knowing about the types of data loads can help you develop a basic understanding of how data makes it into your data warehouse and when certain data will be in the system. This information can provide important background knowledge when accessing and analyzing data.

Extract, Load, Transform (ELT)

The steps of the extract, load, transform (ELT) process follow a different order than in ETL, with transformations to the data happening after the data is loaded, not before. The ELT process allows data to be moved into data storage systems faster, because the transformations take place after the data is loaded.

Data lakes often use the ELT process. Remember that data lakes are meant to hold both unstructured and structured data. This means the data isn’t ready for transformation, but it still needs to be extracted and loaded into the data storage.

The ELT process is ideal for data lakes because they typically hold more real-time data that is updated minute by minute. Data warehouses will still hold data that is updated regularly, but not as frequently as the data in the data lake. ELT is meant to increase data availability and improve processing times.

Review Activity:

The Processes of Extracting, Transforming, and Loading Data

Answer the following questions:

1. **Which method is most commonly used to move data from a source system into a data warehouse?**

2. **Which method is most commonly used to move data from a source system into a data lake?**

3. **What is the biggest difference between the processes of ETL and ELT?**

4. **Which load type is being used when new data is loaded and existing data that has changed since the last load is updated?**

Topic 5B

Explain API/Web Scraping and Other Collection Methods

EXAM OBJECTIVES COVERED

2.1 Explain data acquisition concepts.
1.3 Compare and contrast common data structures and file formats.

Not all data is neatly kept within the servers of an organization. Companies will often leverage third-party tools, typically hosted in the cloud, to serve a particular function for their organization as opposed to building their own internal systems. Reaching the back end of these tools is not as simple as just getting permissions and connecting to them. Common third-party tools include accounting software, customer relationship management software, and even marketing-related software. All of these software applications capture important business data. When purchasing third-party tools, technology departments need to determine if the software is able to connect and interface with other technology systems used by the organization. The ability to share data across systems is important for the collection and analysis of data and to ensure that different business functions are integrated.

A data analyst's job is to gather data from systems. For this reason, a solid understanding of API and other methods of sharing and gaining data, such as machine data, web services, or web scraping, can be immensely helpful in your career.

Application Programming Interface (API)

An **application programming interface (API)** is a set of protocols within a computer system that allows two unrelated systems to communicate. APIs have been around for a long time, but they have gained more visibility in the data world in the last decade. Defined simply, an API is the ability to share data across systems.

We can use a restaurant as an example for understanding conceptually how an API works. When you want to place an order from the menu, you give the order to your server (the interface), and the server goes to the kitchen and provides the order to the cooks. You don't necessarily see all the ins and outs of your order being made in the kitchen, but the food that comes back to you via your server (the same interface) matches your specific menu selection. Just as the restaurant server places your order with the kitchen staff, an API places a "call" or request for specific data from another system. The API then returns this data to the requesting application, similar to how your server returns to your table with your desired meal.

The biggest benefit of an API is the ability to access data from a dedicated system. As an example, imagine that you need up-to-date weather information because your company analyzes the weather forecasts for construction projects when bidding. However, you don't want all the overhead costs of building your own data system for weather data. In this case, you can use an API to call the specific data that you need. This process is a two-way street because the weather app itself must also have an API that allows you to request the data. Some APIs might also utilize a pull/push method for updating data. When pulling data, it is updated continuously regardless of whether a change has occurred. Consider our weather data; with pull, the data will update every five minutes, no matter what. When pushing data, updates are sent out only when a change has occurred. In our weather data example, the API will push an update to the data in the system whenever the weather data itself is updated. The decision to use pull or push is specific to the needs of the program.

Understanding the concept of an API is important in case you need to acquire data that may not be stored in your company's internal systems. As a data analyst, you should become familiar with the APIs that are applicable to the organization where you work. You must learn how to use a particular API to get the data you need from any given system.

Web Services

A **web service** is a type of API that allows a hosted computer on a network to share data back and forth with a computer in the same hosted environment. It's important to remember that while web services are a type of API, not all APIs are web services. The key difference between a web service and most other APIs is the use of a hosted network. Web services require this network and use different protocols to transfer the data between machines.

Web services use XML because this markup language is generic in nature, allowing the data to be transferred to a multitude of systems that are designed in many different languages. Web services are typically dedicated and designed to a particular process.

As an example, imagine that your organization has built two internal data systems to handle customer data and orders. The system that handles orders needs to have the data for the customers that are being added by the sales team. The two systems can communicate via a dedicated web service that passes customer information details from the customers system to the orders system, so that the data doesn't have to be manually entered into both systems.

As a data analyst, you likely will not be responsible for programming web services, but you will rely on web services to provide data to the systems that you work with in your role. Developers who design web services need to make complicated decisions on when and how data should be passed. In your role as a data analyst, it is enough for you to know that data starts out in one system and then is passed into other systems using the web service. It's important for the analyst to know what data the web service is passing from system to system, along with the timing of data transfer. If the service execution is *synchronous*, then the system that calls on the web service waits for a response to the request. It the web service is *asynchronous*, other functions can continue while awaiting a response. If data has not been transferred from one system into another as expected, it could be that the web service needs to work through its process before the data arrives. It's like any other workflow we encounter in any system: the rules have to be met before the data appears where it belongs.

Web Scraping

Web scraping (sometimes called data scraping) is the act of pulling information from a website. Many variations of web scraping are used by data analysts. You may use a tool, or you may web scrape by hand.

To understand web scraping, let's start with an example. Imagine that you are a data analyst working at a local firm that offers a service, and you want to compare your pricing to the pricing of your competitors. Your firm's competitors provide their pricing on their website. Using the hand scraping method, you would navigate to the competitor's website and then proceed to copy/paste or manually key the pricing values into a spreadsheet.A web-scraping tool would allow you to collect the same information from competitor websites using an automated process that produces the data in a format like CSV or even Excel.

Note that not all websites allow web scraping. When working with any data or with any type of programming or technology, read the terms of use first to ensure you are following the legal terms that you must agree to when using the data.

Machine Data

When we talk about the role computer software plays in data collection, we must address machine data. Machine data refers to data that is produced by a machine rather than a human. This doesn't mean that humans are completely uninvolved in the formation of machine data; just that a machine collects the data. For example, when you log into a network, you are physically typing the credentials needed to log yourself in, but the computer time stamps where you logged in from, what time you logged in, and how long you were logged in. You didn't directly give the computer this information, but your interaction with the machine allows it to capture this data.

In the manufacturing industry, workers who maintain the machines use machine data to determine whether the machine is functioning properly. Let's consider a paper mill as an example. A paper mill has a roll line. Paper moves through this line as its being produced, and when it gets to the end of the line it is in the final stages of the process. The machine reports the amount of papers that rolls through that line, because there is a certain amount of production the plant must hit each day to meet demand. This data, produced by the machine, keys workers in to whether they are on track to make their goals. The machine also produces data on its temperature and different parts, which helps the workers maintain the machine.

One of the most interesting business startups in recent years, with truly big data, is one that works to conserve artwork in museums and buildings. This type of machine uses proprietary sensors to detect the room's temperature and humidity throughout the day. This is important because these values need to be consistently maintained to prevent damage to the art by temperature drops and increases in humidity. What makes this truly machine data is the fact that the information the sensors collect is transactional and automatically recorded in real time. Before these sensors were available, museums hired humans to measure and record this information periodically throughout the day.

The biggest value of machine data is that we don't have to enter any of it by hand; it's built to generate data in various formats for analysis.

Review Activity:

API/Web Scraping and Other Collection Methods

Answer the following questions:

1. **What is a benefit of using an API?**

2. **What is the key difference between a web service and most other APIs?**

3. **What language do web services use?**

4. **What is web scraping?**

5. **Sensors that detect the temperature of a person on a job site would be collecting what type of data?**

Topic 5C

Collect and Use Public and Publicly Available Data

EXAM OBJECTIVES COVERED
2.1 Explain data acquisition concepts.

Data analysts will not always have firsthand access to the data that they need within the data systems of their organization. When internal data systems are lacking, you may find that public and publicly available data can help support your analysis. Let's explore what public data is, and how to access it.

Overview of Public and Publicly Available Data

Public data is data that has been made available to the public through various legal requirements. Some popular sources of public data include the U.S. Census Bureau and the U.S. Department of Education. Certain organizations are required by law to provide public data, such as some federal departments and state agencies. Other organizations may make data available to the public even though they are not legally required to do so; this is **publicly available data**. Some organizations (especially nonprofits) make their data publicly available for use in research, while some data is made publicly available for learning purposes, and may merely represent sample data. An example of a site that provides data for learning purposes is Kaggle, an online community and public data platform that provides access to sample data.

Note that for some types of public data, the entirety of the data collected may not be provided to the public. An example of this is education data, where averages of how a student population performed on a standardized test are publicly available, but the actual student records with individual test scores and identifying information are not made available to the public.

In many cases public data is aggregated, but that doesn't make it any less powerful for analysis. **Aggregated data** is data that has already been compiled and summarized (e.g., summed, counted, or averaged by group) for the purposes of analysis and reporting. For example, United States Census data is aggregated by race and ethnicity, health, housing, and other categories. Individual household responses to surveys aren't available for download.

Finding Public and Publicly Available Data

There are thousands of useable data sets online. A simple Google search for the types of public data sets you require can yield a wide variety of results for you to review and consider.

Sites That Provide Public Data

There are numerous websites that give us access to public data. Here are just a few that you may encounter frequently:

- U.S. Government Open Data - https://www.data.gov/
- U.S. Census Bureau - https://data.census.gov/cedsci/
- U.S. Department of Commerce - https://data.commerce.gov/

The U.S. Government Open Data site hosts data from federal agencies under the OPEN Government Data Act, which requires government data be made public. Data can be searched by key word and filtered by location, topic, topic category, data set type, tags, format, organization type, organization, publisher, and bureau.

The U.S. Census Bureau offers aggregated U.S. Census data in the form of geo-profiles and pre-tabulated data tables with fields that can be customized and filtered. Microdata access allows for the creation of custom tables that are not available in the premade tables on data.census.gov. Download formats are limited to Excel and CSV, but multiple APIs are available for varying categorical options.

The U.S. Commerce Department collects, stores, and analyzes data on the nation's economy, population, and environment. Most data sets can be exported as Excel and CSV, with some being available as text/HTML and/or allowing API access. Search results can be filtered by view type, tags, or federated domain.

Sites That Provide Publicly Available Data

There are literally thousands of websites that have made their data publicly available. A few websites that you may find extremely useful are listed here:

- Pew Research Center - https://www.pewresearch.org/
- Kaggle - https://www.kaggle.com/
- GeoPostcodes - https://www.geopostcodes.com/

The Pew Research Center conducts public opinion polling, demographic research, content analysis, and other data-driven social science research. With an account, the published results can be exported as zipped files that will include PDFs of the methodology, questionnaire, and topline results, as well as a 'Read Me' text document and a .SAV file that can be converted into a CSV or Excel file.

Kaggle provides community-driven data sets that can be explored and analyzed, as well as Notebooks for machine learning and a public API. The available data sets will vary depending on licensing and source information and on the user who uploaded the data. Most data sets will be available for download as CSV, though they require an account for export.

GeoPostcodes offers a comprehensive database of post codes with details on countries, administrative division, postal/zip codes, cities and other localities, statistical codes, geocodes, time zones, languages, postal boundaries, and street names. Sample data can be exported with an account, or their proprietary data can be purchased.

Considerations for Using Public Data

It is critical that you first read the terms of use for any public data set before interacting with it, as these terms describe any requirements for using the data. It's extremely important that data analysts understand the terms and methods related to the data and cite them appropriately for any research and presentations.

In addition to the terms that you agree to follow when using public data, you should also read about the methods of data collection and any key definitions that are available related to the data. These methods will explain how the data was collected and any techniques that apply to the data.

Review Activity:

Public and Publicly Available Data

Answer the following questions:

1. **When an organization is required to provide data by law, what would it be called? And what are some sites that provide this type of data?**

2. **What is a consideration of public data that must be addressed before use?**

3. **Public data sources often do not provide the individual data. They will typically provide the data in what format?**

Topic 5D

Use and Collect Survey Data

EXAM OBJECTIVES COVERED
2.1 Explain data acquisition concepts.

Large-scale surveys that collect data from large groups of people, like those run by the U.S. Census Bureau, post-operative surveys, and customer voice surveys, can be a valuable data collection method for research. As with anything, data quality is imperative for accurate reporting. Each survey type and collection method requires special care to ensure data quality. From market research to customer satisfaction, surveys can be used to gain information from people about a topic, a process, or a service. Data analysts will sometimes be involved in the process of developing a survey, and they will more than likely be involved in working with the data that is collected from surveys.

Considerations for Using Surveys

When designed properly and shared effectively, surveys provide a valuable source of data for research, insight, and analysis. A great deal of thought is required in the development of surveys, including the answer types used and how the results are presented. The questions should be accurate for what we are trying to determine through the survey and should be free of bias.

One way you might see bias in a survey is the use of **leading questions**, or questions that prompt certain responses or sway the survey results based on word choices, tone, and how the question is framed. An example of a leading question would be, “How awesome was your customer service today?”, with the answer options “extremely awesome,” “awesome,” and “somewhat awesome.” Considering the language and overall tone of this question, you can see how it leads someone toward a positive experience. It would be difficult to accurately gauge the customer experience by these answers, particularly if it was negative.

Another consideration for an effective survey is whether the provided answer options are detailed enough to elicit a useful response. For example, if only yes or no responses are provided for the question “Are you satisfied with our customer service experience?”, valuable details may be missed about the customer service experience (particularly if the answer is no). Further, the answer you get may not be entirely accurate because you are forcing respondents to default to a yes or a no, when they may have been in the middle. For example, consider a customer who is typically happy but a little frustrated with the service at the moment; neither option really fits. In this case, the answers cannot help inform the process of creating a better customer service experience.

Types of Survey Answers

Surveys often contain a variety of answer types, such as single option, multiple choice, Likert scale, and text-based responses. In addition, some surveys will have built-in logic that will generate questions based on answer responses. Answer types should be carefully chosen to capture the information that is needed by an organization for a particular question.

Imagine that an organization wants to survey customers to see if they would purchase a new service. Let's ask the same question in different ways according to different answer types, and see how the answer type might impact the data.

- **Single choice:** Customers can only provide a single answer. We might ask, "Would you purchase the new service?" In response, customers can reply yes or no.
- **Multiple choice:** Customers can select between multiple options. We might ask, "Would you purchase the new service?" In response, customers could choose one of a few options, such as "in the next six months," "in the next 12 months," or "I would not be interested in this service."
- **Likert scale:** A scale is used to gain details about customers' agreement with or attitude toward certain topics. We might ask, "How likely are you to purchase the following services?" Customers can then select the applicable option from a scale that ranges from "not very likely" to "extremely likely."
- **Text based:** Text-based answers allow for more detailed, written feedback. We might ask, "Tell us your thoughts about this new service." Customers would then provide in their own words an answer in a free-form text box.

Review Activity:

Survey Data

Answer the following questions:

1. **What should a survey be free of?**

2. **Which survey question type uses a scale to gain details about customers' agreement with or attitude toward certain topics?**

3. **What are some common answer types you will find on most surveys?**

4. **Why is it important that a survey provides well-written questions and appropriate answer choices?**

Lesson 5

Summary

After this lesson, you should have a firm understanding of the processes used to move data from a storage system to a data warehouse (Extract, Transform, Load) or data lake (Extract, Load, Transform). You should also have a general understanding of how APIs (Application Programming Interfaces) and web scraping allow you to collect data from systems, and the role that web services play to pass data between systems. While you can collect data yourself, you should know that some data is collected entirely by machines. You should also understand how to find and use public data sets that can be used to support your data projects. Further, when you develop surveys to collect data, you must take particular care not to draft leading and biased questions, and put thought into the answer types.

Guidelines for Extracting, Transforming, and Loading Data from a Source System

Consider these best practices and guidelines when familiarizing yourself with the processes used to extract, transform, and load data.

1. Understand that extracting data is the process of accessing the source data from the system and then converting that data into a format that will then be transformed and loaded into the data warehouse.
2. Understand that transforming data is the act of making the data more meaningful for the purposes of reporting and decision-making.
3. Understand that loading data is the process of moving the data into a tool or its target destination, such as a data warehouse.
4. Differentiate between a full load (loading all the data into the data system for the very first time) and a delta load (loading new data and updating any existing data that has changed since the last load).
5. Recognize the key difference between ELT and ETL: in the ETL process, data is extracted, transformed, and then loaded, whereas in the ELT process, the data is extracted, loaded, and then transformed.

Guidelines for Collecting Data via API, Web Scraping, and Other Methods

Consider these best practices and guidelines when familiarizing yourself with the other methods of data collection you will encounter when working with data.

1. Understand that an application programming interface (API) is a set of protocols within a computer system that allow two unrelated systems to communicate.
2. Recognize how web services (a type of API) allow a hosted computer on a network to share data back and forth with a computer in the same hosted environment.
3. Recognize that some data is produced by a machine rather than a human, which saves the effort of having to enter it by hand.

4. When you need to collect data that doesn't exist in the source system, understand that you can use web scraping to collect data from outside.
5. Know that certain organizations are required by law to provide public data, such as some federal departments and state agencies.
6. Ensure you read the terms and agreements related to any public data you plan to use.
7. When using surveys to collect data, consider whether the provided answer options can elicit an accurate and useful response, and make sure questions are free from bias.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 6

Identifying Common Reasons for Cleansing and Profiling Data

LESSON INTRODUCTION

Rapid changes in business practices and business requirements are the main reasons we encounter less-than-perfect data sets or data structures in our work, leading to the need for cleansing and profiling data. Data sets that have a poor design or are attempting to retrofit a process into an off-the-shelf software also often need to be cleaned or profiled. When a company designs a system or process for handling their data, they base it on what they know in that moment. If you have ever heard the saying, “flying the plane while building it,” that is a fair description of how many organizations approach building data-centric systems. Once you know the types of imperfections data systems may have, you can start handling these common issues.

Lesson Objectives

In this lesson, you will do the following:

- Learn to profile data.
- Address redundant, duplicated, and unnecessary data.
- Work with missing values.
- Address invalid data.
- Convert data to meet specifications.

Topic 6A

Learn to Profile Data

EXAM OBJECTIVES COVERED
2.2 Identify common reasons for cleansing and profiling datasets.

In its simplest terms, **data profiling** means to learn the basics about the data that you are working with and discern information from that data. Profiling data will help you to identify any data quality issues so that you can correct them. It will also help inform what data cleansing needs to occur in order to improve the data for reporting and analysis.

The data analyst should always start by profiling any data set that they plan to work with while performing their role.

Steps of Data Profiling

Profiling data involves forming an understanding of the data by learning about aspects such as the following:

- The source of the data
- Keys of the data
- Relationships within the data
- Record counts (e.g., how many product records are in the set, or how many employees are represented)

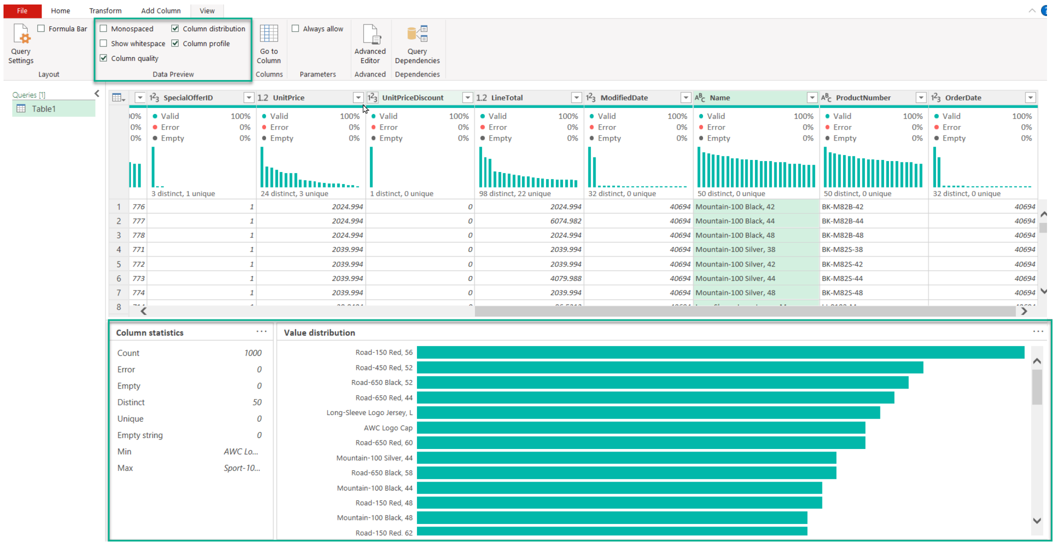
Here are a few basic steps that data analysts should follow to profile data:

1. Identify and document the source of the data and its integrity. For example, was this data collected in a system with a defined process, or was it manually keyed from a paper document?
2. Identify the fieldnames and data types, and determine whether they are appropriate for what you are reporting. For example, are the date fields set as dates, or are they text? Depending on what you discover, you may need to convert the data types for some fields.
3. Determine the main fields identified for reports and what their column profile entails. For example, if you are reporting on products, what products are represented in the data?
4. Check whether a key field (primary, natural, or foreign) is represented where expected. For example, if working with product sale data, do all the sales have an associated product code?
5. Recognize the total of all the data in the data set. For example, if the total of every record combined is $500,000, and $1 million suddenly displays upon reporting, is there an obvious issue here?

Data Profiling Tools and Techniques

Data analysts can use either manual techniques or advanced software to profile data. However, when working with large amounts of data, the use of basic manual techniques can be a challenge. Most of the popular data tools (such as Power Query in Excel, PowerBI, and Tableau) contain built-in tools to help you profile data. In Power Query, the process is as simple as checking the column profile checkbox. In Tableau, profiling data requires just a right click and the selection of the describe option.

Column profiling gives us basic insight into that column of data. For data types like numbers, you can get insight on the max, min, and averages of that column. For text data types, you can obtain distinct or unique counts of the values in that column.



☐ Formula Bar
☐ Monospaced ☑ Column distribution
☐ Show whitespace ☑ Column profile
☑ Column quality
Data Preview
☐ Always allow

	SpecialOfferID	UnitPrice	UnitPriceDiscount	LineTotal	ModifiedDate	Name	ProductNumber	OrderDate
00%	● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%
0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%
0%	● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%
	3 distinct, 1 unique	24 distinct, 3 unique	1 distinct, 0 unique	98 distinct, 22 unique	32 distinct, 0 unique	50 distinct, 0 unique	50 distinct, 0 unique	32 distinct, 0 unique
1 776	1	2024.994	0	2024.994	40694	Mountain-100 Black, 42	BK-M82B-42	40694
2 777	1	2024.994	0	6074.982	40694	Mountain-100 Black, 44	BK-M82B-44	40694
3 778	1	2024.994	0	2024.994	40694	Mountain-100 Black, 48	BK-M82B-48	40694
4 771	1	2039.994	0	2039.994	40694	Mountain-100 Silver, 38	BK-M82S-38	40694
5 772	1	2039.994	0	2039.994	40694	Mountain-100 Silver, 42	BK-M82S-42	40694
6 773	1	2039.994	0	4079.988	40694	Mountain-100 Silver, 44	BK-M82S-44	40694
7 774	1	2039.994	0	2039.994	40694	Mountain-100 Silver, 48	BK-M82S-48	40694

Column statistics	
Count	1000
Error	0
Empty	0
Distinct	50
Unique	0
Empty string	0
Min	AWC Lo...
Max	Sport-10...

Power Query Column Profile (Used with permission from Microsoft.)

This image shows the column profile, distribution, and quality based on the first 1,000 rows of data for a sample set. At the bottom of the window, the column profile provides us with the column statistics and the value distribution. At the top of the column, the column distribution checkbox gives us a histogram view of the column values. At the top of the window, the column quality provides valid, error, and empty percentages.

Review Activity:

Learn to Profile Data

1. **Why is it important for the data analyst to profile data?**

2. **What are some elements of data that are assessed when profiling?**

3. **Name a few popular tools that can be used to profile data.**

Topic 6B

Address Redundant, Duplicated, and Unnecessary Data

EXAM OBJECTIVES COVERED

2.2 Identify common reasons for cleansing and profiling datasets.

As you fulfill your role as a data analyst, you are likely to encounter redundant data and duplicated data. Learning how to identify redundant and duplicated data early on in a project will help you to provide more accurate data for your analysis. You will also sometimes find that you may have more fields than you need, and you will want to either remove them or exclude them when writing queries.

Redundant Data

Redundant data is identical data that is stored in multiple places. For example, if an organization stores a list of products in its customer relationship software for its salesforce to quote, and the same list of products is also stored in its accounting software for invoicing, that data would be considered redundant. You may also encounter redundant data when there is a design flaw in a system, or when the system is intentionally designed to capture data everywhere. The challenge you will have when working with redundant data is determining which record represents the absolute truth and is the most accurate.

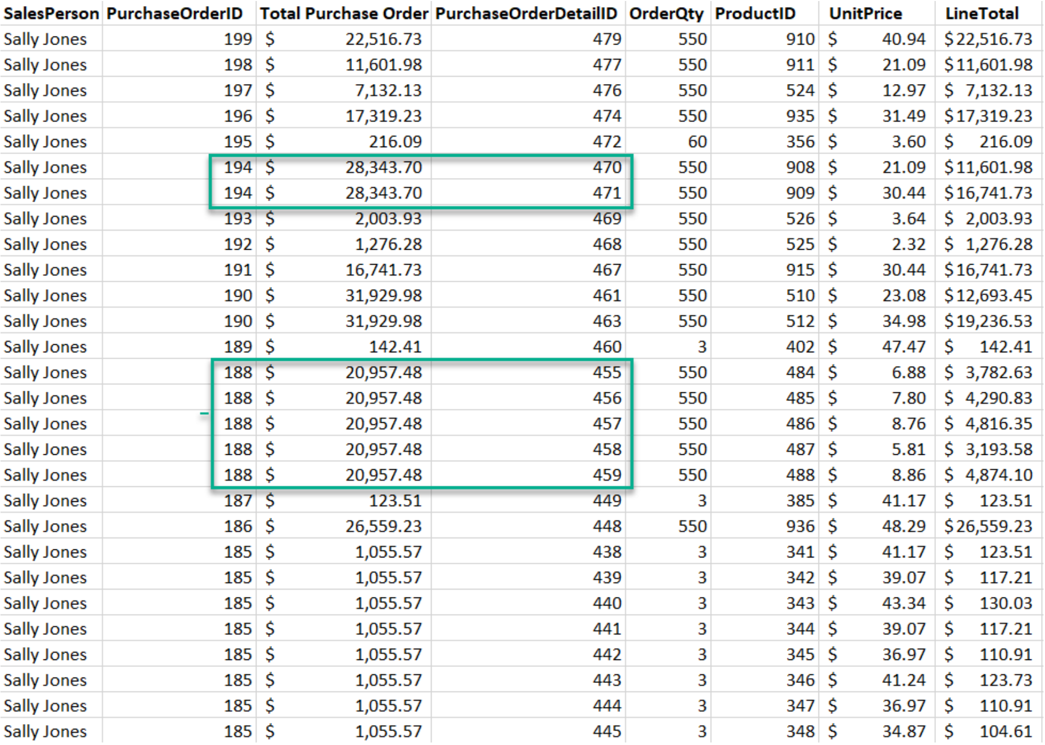
With any data project, you should determine if any data might be redundant up front, to ensure you they are working with the source that represents the most accurate set of information. In our example of a product list located in both the customer relationship management (CRM) software and in the accounting software, you would need to determine which source represents the most accurate information for the products. Decision makers and data owners typically know which source represents the most accurate data for the business objective, as they have a more intimate knowledge of not only the data itself but also the processes of the data. For example, the sales manager will know which clients the sales team has been working with, and a general manager will understand what's happened on a project. They have context around the data that you as an analyst may not be able to determine from just looking at the data.

Duplicated Data

Duplicated data is data that is repeated within the same data set. It is important to note that there's no inherent problem with duplicated data, as long as you expect it. However, if you don't know that duplicated records exist when you're running a report, it can cause your report to be invalid. Consider Sally Jones, a top salesperson, who is responsible for creating purchase orders for each sale she makes. Sally's name appears alongside these purchase orders, which are each given an ID number. The purchase order contains all the items that are being purchased, with each item being listed as a separate line item. This means that purchase orders involving more than one item will result in duplicated data (e.g., the purchase order ID will appear three times if three different items are purchased). When looking quickly at the data for Sally's sales, the analyst might assume that each record represents a separate purchase order, but in reality, there could be multiple line items for one purchase order. As this example demonstrates, if you are not critically looking at the entire data set, the data could be misinterpreted, causing you to inadvertently inflate the numbers. Thus, one of the very first actions a data analyst should take when working with data is to look through the data sets for any duplicated records or fields.

There are many reasons why duplicated records might exist. It could be due to a design flaw in the data design, or maybe the data was inappropriately joined with another data set. Or, it could just be that you received a data set that is full of duplicated data by design.

Going back to our previous example, consider the data for Sally Jones's purchase orders, as shown in the screenshot.



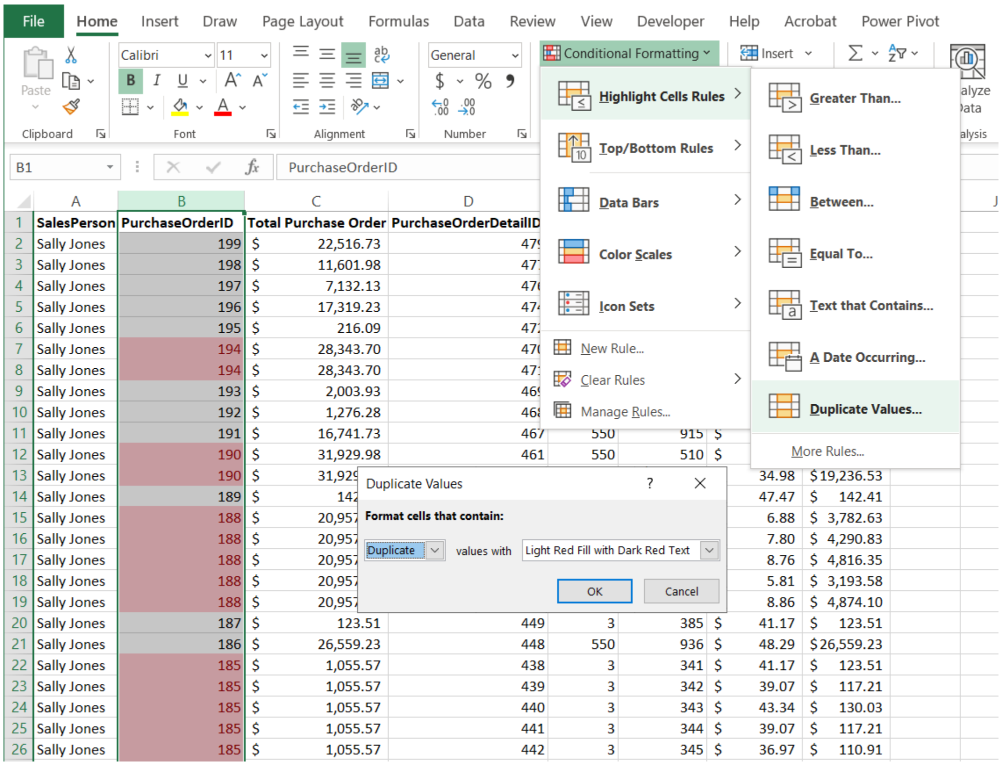
SalesPerson	PurchaseOrderID	Total Purchase Order	PurchaseOrderDetailID	OrderQty	ProductID	UnitPrice	LineTotal
Sally Jones	199	$ 22,516.73	479	550	910	$ 40.94	$22,516.73
Sally Jones	198	$ 11,601.98	477	550	911	$ 21.09	$11,601.98
Sally Jones	197	$ 7,132.13	476	550	524	$ 12.97	$ 7,132.13
Sally Jones	196	$ 17,319.23	474	550	935	$ 31.49	$17,319.23
Sally Jones	195	$ 216.09	472	60	356	$ 3.60	$ 216.09
Sally Jones	194	$ 28,343.70	470	550	908	$ 21.09	$11,601.98
Sally Jones	194	$ 28,343.70	471	550	909	$ 30.44	$16,741.73
Sally Jones	193	$ 2,003.93	469	550	526	$ 3.64	$ 2,003.93
Sally Jones	192	$ 1,276.28	468	550	525	$ 2.32	$ 1,276.28
Sally Jones	191	$ 16,741.73	467	550	915	$ 30.44	$16,741.73
Sally Jones	190	$ 31,929.98	461	550	510	$ 23.08	$12,693.45
Sally Jones	190	$ 31,929.98	463	550	512	$ 34.98	$19,236.53
Sally Jones	189	$ 142.41	460	3	402	$ 47.47	$ 142.41
Sally Jones	188	$ 20,957.48	455	550	484	$ 6.88	$ 3,782.63
Sally Jones	188	$ 20,957.48	456	550	485	$ 7.80	$ 4,290.83
Sally Jones	188	$ 20,957.48	457	550	486	$ 8.76	$ 4,816.35
Sally Jones	188	$ 20,957.48	458	550	487	$ 5.81	$ 3,193.58
Sally Jones	188	$ 20,957.48	459	550	488	$ 8.86	$ 4,874.10
Sally Jones	187	$ 123.51	449	3	385	$ 41.17	$ 123.51
Sally Jones	186	$ 26,559.23	448	550	936	$ 48.29	$26,559.23
Sally Jones	185	$ 1,055.57	438	3	341	$ 41.17	$ 123.51
Sally Jones	185	$ 1,055.57	439	3	342	$ 39.07	$ 117.21
Sally Jones	185	$ 1,055.57	440	3	343	$ 43.34	$ 130.03
Sally Jones	185	$ 1,055.57	441	3	344	$ 39.07	$ 117.21
Sally Jones	185	$ 1,055.57	442	3	345	$ 36.97	$ 110.91
Sally Jones	185	$ 1,055.57	443	3	346	$ 41.24	$ 123.73
Sally Jones	185	$ 1,055.57	444	3	347	$ 36.97	$ 110.91
Sally Jones	185	$ 1,055.57	445	3	348	$ 34.87	$ 104.61

Sally Jones's Purchase Orders and Details in Microsoft Excel (Used with permission from Microsoft.)

When we look at the PurchaseOrderID list for Sally's sales, we see that in some cases one PO covers multiple line items (indicated by the PurchaseOrderDetailID field). Along with the PurchaseOrderID, the Total Purchase Order (the amount of the sale for the entire order) is also duplicated for each line item. If we totaled all the data in the Total Purchase Order column, as is, we would produce an inflated amount.

Taking steps to identify your data set for duplicates and redundancy up front helps prevent you from using inaccurate information and inflated numbers.

Programs like Excel have built-in tools to help you identify duplicate records. For example, you can use simple conditional formatting to highlight any column and format the duplicates so those fields stand out. In Excel, there is also a Remove Duplicates command that will allow you to select the fields that form a duplicated record and then remove them. When you write queries as an analyst, you can also set a property or use the key word "Distinct" in your "Select" statement to provide a list with only the non-duplicated records.



B1 | PurchaseOrderID

	A	B	C	D				
1	SalesPerson	PurchaseOrderID	Total Purchase Order	PurchaseOrderDetailID				
2	Sally Jones	199	$ 22,516.73	47				
3	Sally Jones	198	$ 11,601.98	47				
4	Sally Jones	197	$ 7,132.13	47				
5	Sally Jones	196	$ 17,319.23	47				
6	Sally Jones	195	$ 216.09	47				
7	Sally Jones	194	$ 28,343.70	47				
8	Sally Jones	194	$ 28,343.70	47				
9	Sally Jones	193	$ 2,003.93	46				
10	Sally Jones	192	$ 1,276.28	46				
11	Sally Jones	191	$ 16,741.73	467	550	915	$	
12	Sally Jones	190	$ 31,929.98	461	550	510	$	
13	Sally Jones	190	$ 31,929				34.98	$19,236.53
14	Sally Jones	189	$ 142				47.47	$ 142.41
15	Sally Jones	188	$ 20,957				6.88	$ 3,782.63
16	Sally Jones	188	$ 20,957				7.80	$ 4,290.83
17	Sally Jones	188	$ 20,957				8.76	$ 4,816.35
18	Sally Jones	188	$ 20,957				5.81	$ 3,193.58
19	Sally Jones	188	$ 20,957				8.86	$ 4,874.10
20	Sally Jones	187	$ 123.51	449	3	385	$ 41.17	$ 123.51
21	Sally Jones	186	$ 26,559.23	448	550	936	$ 48.29	$26,559.23
22	Sally Jones	185	$ 1,055.57	438	3	341	$ 41.17	$ 123.51
23	Sally Jones	185	$ 1,055.57	439	3	342	$ 39.07	$ 117.21
24	Sally Jones	185	$ 1,055.57	440	3	343	$ 43.34	$ 130.03
25	Sally Jones	185	$ 1,055.57	441	3	344	$ 39.07	$ 117.21
26	Sally Jones	185	$ 1,055.57	442	3	345	$ 36.97	$ 110.91

Duplicate Values

Format cells that contain:

Duplicate values with Light Red Fill with Dark Red Text

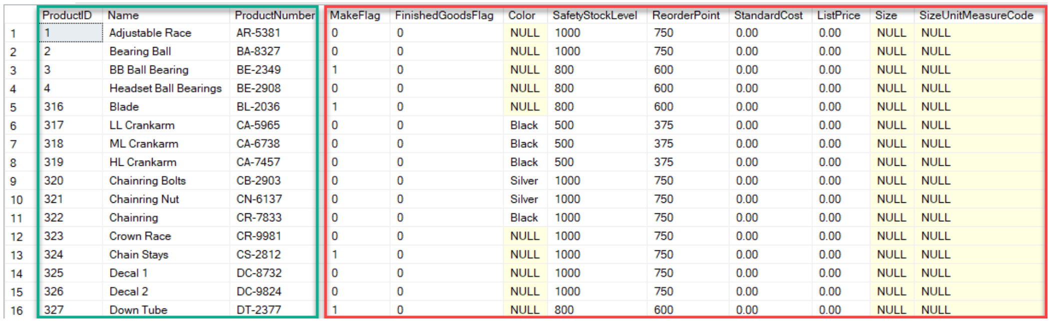
OK Cancel

Highlight Duplicates in Microsoft Excel (Used with permission from Microsoft.)

When working with a large data set, it is virtually impossible to spot every duplicate by eye. You can use commands like conditional formatting to support your duplicate detection.

Unnecessary Fields

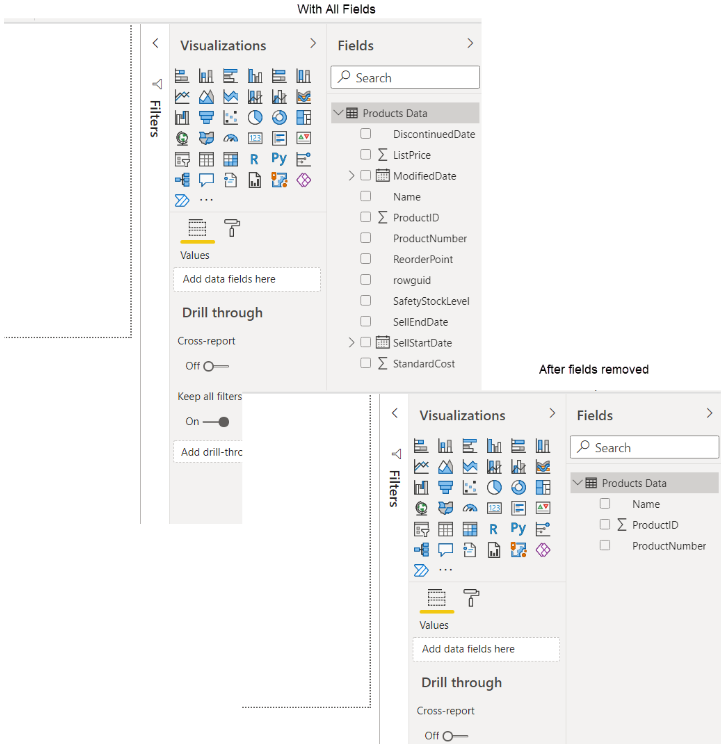
You will commonly have access to more data than you need, and a routine part of your work as an analyst is to remove that unnecessary data. When working with a data set that contains everything, the unnecessary fields represent what we call "noise" in the data. These extra fields have no real meaning for the analysis. For example, if we have been provided an export of data from the accounting software that represents all the products stored in the system, and the export includes every field in that table, then it's likely some of the fields should be removed before analysis. You can see in the screenshot of the accounting product list, with every field included, that some of these fields are not needed for analytical purposes.



	ProductID	Name	ProductNumber	MakeFlag	FinishedGoodsFlag	Color	SafetyStockLevel	ReorderPoint	StandardCost	ListPrice	Size	SizeUnitMeasureCode
1	1	Adjustable Race	AR-5381	0	0	NULL	1000	750	0.00	0.00	NULL	NULL
2	2	Bearing Ball	BA-8327	0	0	NULL	1000	750	0.00	0.00	NULL	NULL
3	3	BB Ball Bearing	BE-2349	1	0	NULL	800	600	0.00	0.00	NULL	NULL
4	4	Headset Ball Bearings	BE-2908	0	0	NULL	800	600	0.00	0.00	NULL	NULL
5	316	Blade	BL-2036	1	0	NULL	800	600	0.00	0.00	NULL	NULL
6	317	LL Crankarm	CA-5965	0	0	Black	500	375	0.00	0.00	NULL	NULL
7	318	ML Crankarm	CA-6738	0	0	Black	500	375	0.00	0.00	NULL	NULL
8	319	HL Crankarm	CA-7457	0	0	Black	500	375	0.00	0.00	NULL	NULL
9	320	Chainring Bolts	CB-2903	0	0	Silver	1000	750	0.00	0.00	NULL	NULL
10	321	Chainring Nut	CN-6137	0	0	Silver	1000	750	0.00	0.00	NULL	NULL
11	322	Chainring	CR-7833	0	0	Black	1000	750	0.00	0.00	NULL	NULL
12	323	Crown Race	CR-9981	0	0	NULL	1000	750	0.00	0.00	NULL	NULL
13	324	Chain Stays	CS-2812	1	0	NULL	1000	750	0.00	0.00	NULL	NULL
14	325	Decal 1	DC-8732	0	0	NULL	1000	750	0.00	0.00	NULL	NULL
15	326	Decal 2	DC-9824	0	0	NULL	1000	750	0.00	0.00	NULL	NULL
16	327	Down Tube	DT-2377	1	0	NULL	800	600	0.00	0.00	NULL	NULL

Product List Showing All Fields of Data, Displayed in Query Grid of Microsoft SQL Server Management Studio (Used with permission from Microsoft.)

Removing unnecessary data makes for more efficient data analysis. Having columns of data you don't use increases the size of the file and requires the software to process all of the data fields that are there, even if they are never going to be used. Using the example of the accounting product list, suppose that you need only the product-specific fields for your analysis, such as ProductID, Name, and ProductNumber. When profiling the data, you discover that there are additional fields (e.g., MakeFlag, FinishedGoodsFlag) that include data captured on the back end by the accounting software. These unnecessary columns of data can be removed so that the data set only includes the needed product-related fields. Data programs such as Power BI, Tableau, and Excel allow you to remove columns of data that you don't need. In the screenshot of Power BI, you can see the difference in the amount of fields before the removal and after. The list is cleaner and more manageable.



Before and After List of Data Fields in Microsoft Power BI (Used with permission from Microsoft.)

When you use data software like Power BI or Tableau to connect to a data set, the software will display all the fields available for you in the data set. If you keep the unnecessary fields, they will show up as an available field to use in your analysis, even if you never use it.

To remove fields in Excel, we highlight the columns we do not need and remove them. If you are building a query, then you would only include the fields needed in that data set. These approaches allow you to eliminate noise in your data.

Review Activity:

Redundant, Duplicated, and Unnecessary Data

Answer the following questions:

1. **What does it mean for data to be redundant?**

2. **When data is duplicated and then totaled, what problem will you encounter?**

3. **Spotting duplicates can be a challenge in a large data set. What type of command could you use to highlight them?**

4. **Your data set contains every field of that table or query, and only a few fields are relevant for your study. What should you do with the unnecessary fields?**

5. **What are some of the drawbacks of leaving unnecessary fields in your data?**

Topic 6C

Work with Missing Values

EXAM OBJECTIVES COVERED

2.2 Identify common reasons for cleansing and profiling datasets.

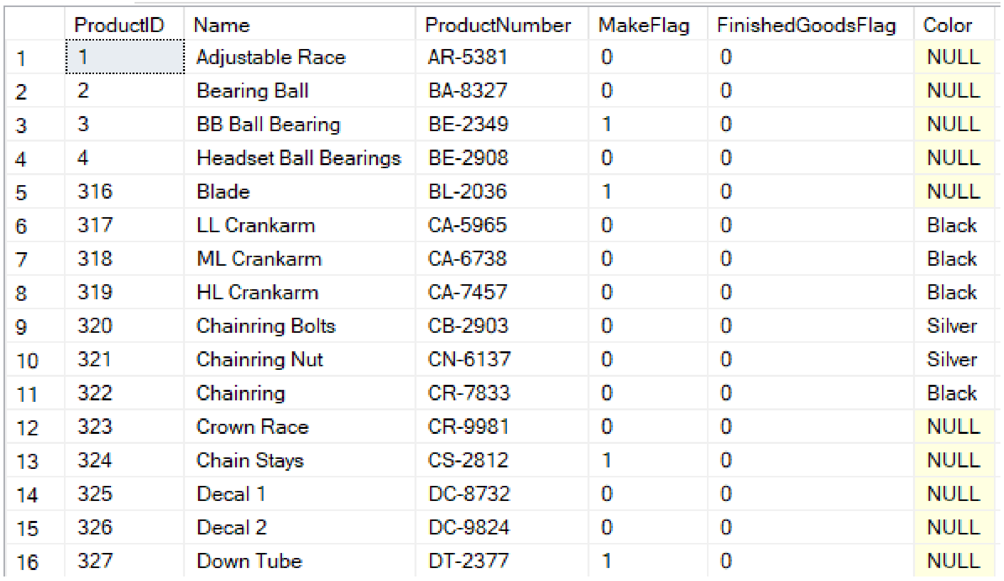
When you work with data, you will at some point be faced with missing data. Missing data can occur for a number of reasons. It could be missing due to survey respondents failing to fill in an answer, or because the data has not yet been entered into a system. It could also be the result of joining data. A null value will often exist to represent that missing data. The data analyst must be able to identify the cause of the null data in order to determine the best way to address it for a particular data project. Nulls do not just represent a bad or missing data entry—they can provide valuable insight and meaningful information.

Causes of Null Values

In any given data set, the analyst may encounter null values. **NULL** means that there is no value in a field. The field might appear to be blank, may contain the word "NULL," or, in the case of Excel, might return an "N/A."

Value Is Not Applicable to the Field

Null values can occur when the value isn't applicable to that field. For example, consider the SQL table shown here. In this list of products, color is a valid field, but when a product doesn't have that characteristic it is null.

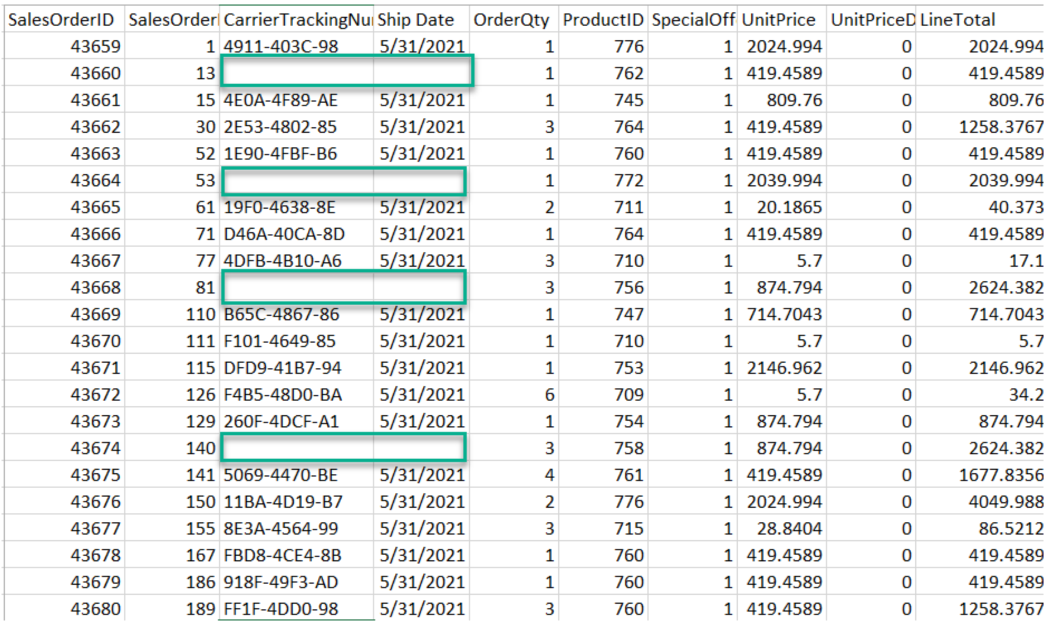


	ProductID	Name	ProductNumber	MakeFlag	FinishedGoodsFlag	Color
1	1	Adjustable Race	AR-5381	0	0	NULL
2	2	Bearing Ball	BA-8327	0	0	NULL
3	3	BB Ball Bearing	BE-2349	1	0	NULL
4	4	Headset Ball Bearings	BE-2908	0	0	NULL
5	316	Blade	BL-2036	1	0	NULL
6	317	LL Crankarm	CA-5965	0	0	Black
7	318	ML Crankarm	CA-6738	0	0	Black
8	319	HL Crankarm	CA-7457	0	0	Black
9	320	Chainring Bolts	CB-2903	0	0	Silver
10	321	Chainring Nut	CN-6137	0	0	Silver
11	322	Chainring	CR-7833	0	0	Black
12	323	Crown Race	CR-9981	0	0	NULL
13	324	Chain Stays	CS-2812	1	0	NULL
14	325	Decal 1	DC-8732	0	0	NULL
15	326	Decal 2	DC-9824	0	0	NULL
16	327	Down Tube	DT-2377	1	0	NULL

NULL Value in SQL Table in Microsoft SQL Server Management Studio (Used with permission from Microsoft.)

Data Set Does Not Yet Have the Information

Another reason you might encounter null values is when the data set does not have all the information it needs at that time. For example, if a data set contains records that represent the process of shipping orders, and the order hasn't been shipped yet, the Ship Date field will be null.



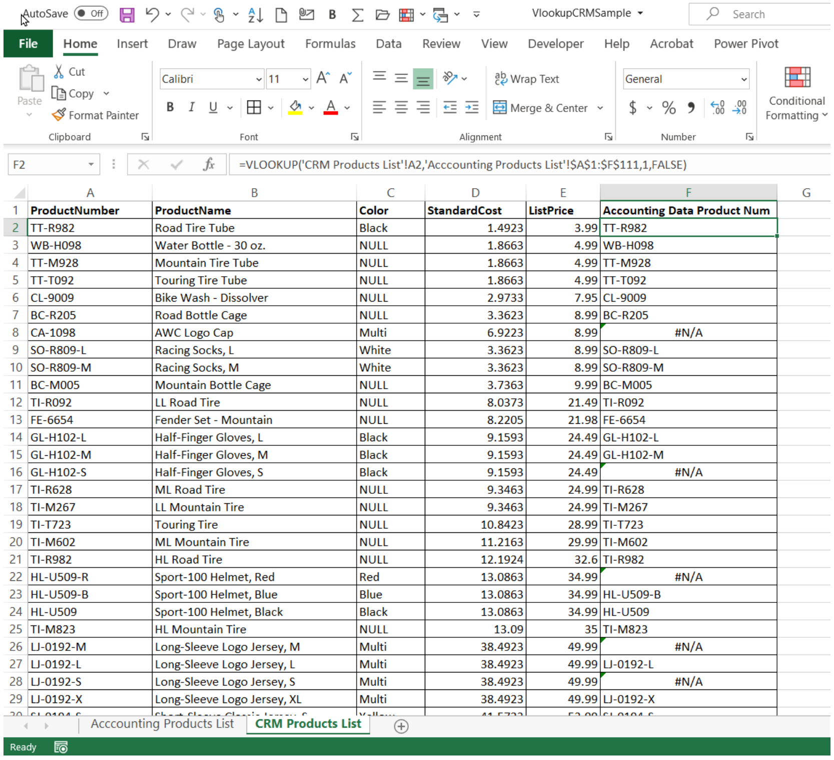
SalesOrderID	SalesOrder	CarrierTrackingNu	Ship Date	OrderQty	ProductID	SpecialOff	UnitPrice	UnitPriceD	LineTotal
43659	1	4911-403C-98	5/31/2021	1	776	1	2024.994	0	2024.994
43660	13			1	762	1	419.4589	0	419.4589
43661	15	4E0A-4F89-AE	5/31/2021	1	745	1	809.76	0	809.76
43662	30	2E53-4802-85	5/31/2021	3	764	1	419.4589	0	1258.3767
43663	52	1E90-4FBF-B6	5/31/2021	1	760	1	419.4589	0	419.4589
43664	53			1	772	1	2039.994	0	2039.994
43665	61	19F0-4638-8E	5/31/2021	2	711	1	20.1865	0	40.373
43666	71	D46A-40CA-8D	5/31/2021	1	764	1	419.4589	0	419.4589
43667	77	4DFB-4B10-A6	5/31/2021	3	710	1	5.7	0	17.1
43668	81			3	756	1	874.794	0	2624.382
43669	110	B65C-4867-86	5/31/2021	1	747	1	714.7043	0	714.7043
43670	111	F101-4649-85	5/31/2021	1	710	1	5.7	0	5.7
43671	115	DFD9-41B7-94	5/31/2021	1	753	1	2146.962	0	2146.962
43672	126	F4B5-48D0-BA	5/31/2021	6	709	1	5.7	0	34.2
43673	129	260F-4DCF-A1	5/31/2021	1	754	1	874.794	0	874.794
43674	140			3	758	1	874.794	0	2624.382
43675	141	5069-4470-BE	5/31/2021	4	761	1	419.4589	0	1677.8356
43676	150	11BA-4D19-B7	5/31/2021	2	776	1	2024.994	0	4049.988
43677	155	8E3A-4564-99	5/31/2021	3	715	1	28.8404	0	86.5212
43678	167	FBD8-4CE4-8B	5/31/2021	1	760	1	419.4589	0	419.4589
43679	186	918F-49F3-AD	5/31/2021	1	760	1	419.4589	0	419.4589
43680	189	FF1F-4DD0-98	5/31/2021	3	760	1	419.4589	0	1258.3767

Ship Date and Carrier Tracking Number Both NULL in Microsoft Excel (Used with permission from Microsoft.)

The values shown here are null because that data is unavailable and has not yet been created. In this scenario, the fact that the ship date is null provides valuable information; it lets us know that the order has not been shipped.

Data Set Has No Match for Expected Information

Null values can also occur when there is no match for the expected information. Recall our example of a company that stores its product list in both the customer relationship management (CRM) system and the accounting system. Suppose we want to ensure that all products are listed in both systems. We know that the CRM system is the true source of the data, so we would likely use the VLOOKUP function in Excel (a function that searches for data values within another data set) to compare the list of product codes in that system to the accounting system product list. Anytime VLOOKUP encounters a product that is in the CRM list, but not the accounting list, it will place an N/A for the null value. In this scenario, the NULL values are useful in that they tell us which products need to be added to the accounting system.



F2 =VLOOKUP('CRM Products List'!A2,'Acccounting Products List'!$A$1:$F$111,1,FALSE)

	A	B	C	D	E	F
1	ProductNumber	ProductName	Color	StandardCost	ListPrice	Accounting Data Product Num
2	TT-R982	Road Tire Tube	Black	1.4923	3.99	TT-R982
3	WB-H098	Water Bottle - 30 oz.	NULL	1.8663	4.99	WB-H098
4	TT-M928	Mountain Tire Tube	NULL	1.8663	4.99	TT-M928
5	TT-T092	Touring Tire Tube	NULL	1.8663	4.99	TT-T092
6	CL-9009	Bike Wash - Dissolver	NULL	2.9733	7.95	CL-9009
7	BC-R205	Road Bottle Cage	NULL	3.3623	8.99	BC-R205
8	CA-1098	AWC Logo Cap	Multi	6.9223	8.99	#N/A
9	SO-R809-L	Racing Socks, L	White	3.3623	8.99	SO-R809-L
10	SO-R809-M	Racing Socks, M	White	3.3623	8.99	SO-R809-M
11	BC-M005	Mountain Bottle Cage	NULL	3.7363	9.99	BC-M005
12	TI-R092	LL Road Tire	NULL	8.0373	21.49	TI-R092
13	FE-6654	Fender Set - Mountain	NULL	8.2205	21.98	FE-6654
14	GL-H102-L	Half-Finger Gloves, L	Black	9.1593	24.49	GL-H102-L
15	GL-H102-M	Half-Finger Gloves, M	Black	9.1593	24.49	GL-H102-M
16	GL-H102-S	Half-Finger Gloves, S	Black	9.1593	24.49	#N/A
17	TI-R628	ML Road Tire	NULL	9.3463	24.99	TI-R628
18	TI-M267	LL Mountain Tire	NULL	9.3463	24.99	TI-M267
19	TI-T723	Touring Tire	NULL	10.8423	28.99	TI-T723
20	TI-M602	ML Mountain Tire	NULL	11.2163	29.99	TI-M602
21	TI-R982	HL Road Tire	NULL	12.1924	32.6	TI-R982
22	HL-U509-R	Sport-100 Helmet, Red	Red	13.0863	34.99	#N/A
23	HL-U509-B	Sport-100 Helmet, Blue	Blue	13.0863	34.99	HL-U509-B
24	HL-U509	Sport-100 Helmet, Black	Black	13.0863	34.99	HL-U509
25	TI-M823	HL Mountain Tire	NULL	13.09	35	TI-M823
26	LJ-0192-M	Long-Sleeve Logo Jersey, M	Multi	38.4923	49.99	#N/A
27	LJ-0192-L	Long-Sleeve Logo Jersey, L	Multi	38.4923	49.99	LJ-0192-L
28	LJ-0192-S	Long-Sleeve Logo Jersey, S	Multi	38.4923	49.99	#N/A
29	LJ-0192-X	Long-Sleeve Logo Jersey, XL	Multi	38.4923	49.99	LJ-0192-X

Microsoft Excel Vlookup for Data Match (Used with permission from Microsoft.)

You will note that there are several "#N/As" present in column F (the Accounting Data Product Number), which signifies that those products were not identified in the accounting system. They may not exist on that sheet, or they may just have issues that can be addressed with cleaning. Either way, the null value gives us valuable insight on where we should begin our investigation.

When the analyst queries data (discussed later in the course) and joins tables together, this process can also produce a NULL value, similar to the VLOOKUP function. If a matching record does not exist in both tables, a NULL is produced.

Survey Data Is Incomplete

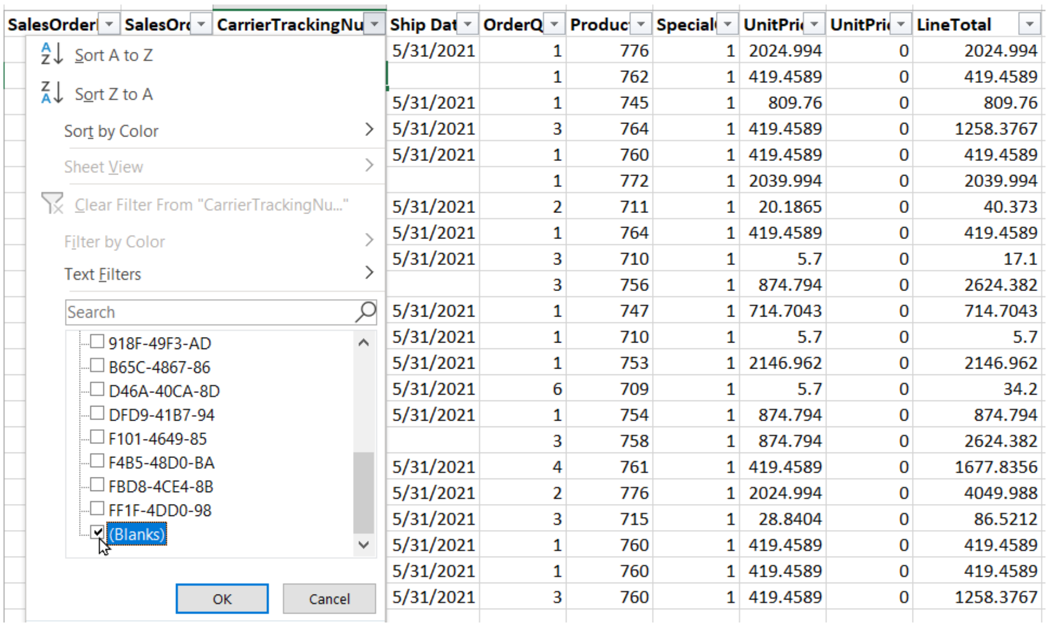
Another scenario in which the analyst may encounter NULL values is when working with survey data. For example, customers answering a product satisfaction survey may not answer every question. When you extract the data from that survey, NULL values will appear anywhere an answer was not provided.

There is no one-size-fits-all approach to addressing NULL values within data. Determining why the null value exists is important because it will drive what you should do when working with NULL values.

Filtering Null Values

As discussed, NULL values can provide valuable insights when we are analyzing data. The way we work with null values depends on the reason they exist. One way we can utilize the information provided by null values is through filtering.

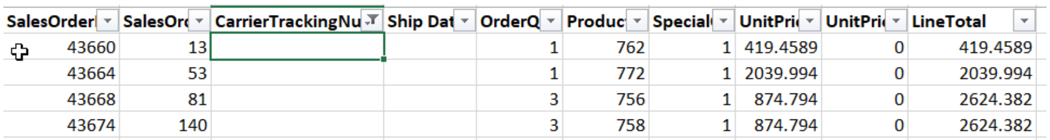
Let's return to our example of data on the shipping process, where NULL values for the ship date show for any orders that have not yet shipped. This data can be filtered to show only NULL values, or only the values that are not null, which gives us the ability to view shipped and not shipped orders in two lists.



SalesOrder	SalesOrd	CarrierTrackingNu	Ship Dat	OrderQ	Produc	Special	UnitPri	UnitPri	LineTotal
			5/31/2021	1	776	1	2024.994	0	2024.994
				1	762	1	419.4589	0	419.4589
			5/31/2021	1	745	1	809.76	0	809.76
			5/31/2021	3	764	1	419.4589	0	1258.3767
			5/31/2021	1	760	1	419.4589	0	419.4589
				1	772	1	2039.994	0	2039.994
			5/31/2021	2	711	1	20.1865	0	40.373
			5/31/2021	1	764	1	419.4589	0	419.4589
			5/31/2021	3	710	1	5.7	0	17.1
				3	756	1	874.794	0	2624.382
			5/31/2021	1	747	1	714.7043	0	714.7043
			5/31/2021	1	710	1	5.7	0	5.7
			5/31/2021	1	753	1	2146.962	0	2146.962
			5/31/2021	6	709	1	5.7	0	34.2
			5/31/2021	1	754	1	874.794	0	874.794
				3	758	1	874.794	0	2624.382
			5/31/2021	4	761	1	419.4589	0	1677.8356
			5/31/2021	2	776	1	2024.994	0	4049.988
			5/31/2021	3	715	1	28.8404	0	86.5212
			5/31/2021	1	760	1	419.4589	0	419.4589
			5/31/2021	1	760	1	419.4589	0	419.4589
			5/31/2021	3	760	1	419.4589	0	1258.3767

When Filtering in Excel, NULL Values Are Referred to as Blanks (Used with permission from Microsoft.)

As shown here, filtering the carrier tracking numbers and ship date by NULL values gives us a list of only the orders that have not yet been shipped.



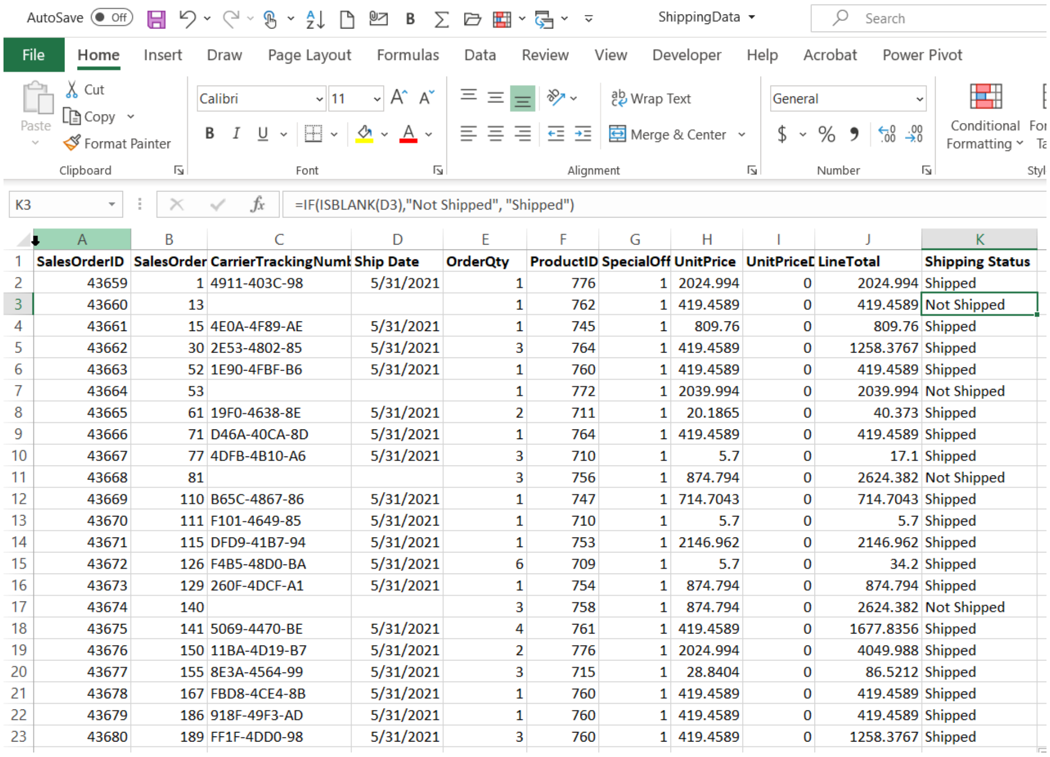
SalesOrder	SalesOrd	CarrierTrackingNu	Ship Dat	OrderQ	Produc	Special	UnitPri	UnitPri	LineTotal
43660	13			1	762	1	419.4589	0	419.4589
43664	53			1	772	1	2039.994	0	2039.994
43668	81			3	756	1	874.794	0	2624.382
43674	140			3	758	1	874.794	0	2624.382

Shipping Data After Filtering by NULL Values (i.e., Blanks) in Excel (Used with permission from Microsoft.)

Replacing Missing Values

When we want to replace the null value with a more meaningful result, we can use functions. Recall the product list in which NULL values appeared in the color field if color wasn't applicable for that product. A function could be used to change the NULL to read "No Color," "N/A," or whichever other answer appropriately describes the missing value in the field.

Let's return to this lesson's example of shipping data. Recall that null values exist for the ship date when the order hasn't yet shipped. We could write a function (e.g., an IF function) that creates a new field based on the null value, which indicates whether the order has shipped or has not shipped.



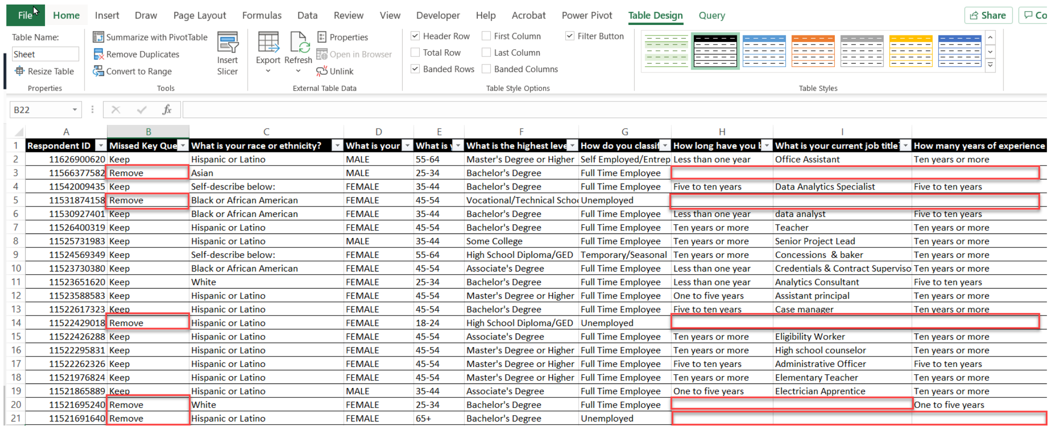
K3 | =IF(ISBLANK(D3),"Not Shipped", "Shipped")

	A	B	C	D	E	F	G	H	I	J	K
1	SalesOrderID	SalesOrder	CarrierTrackingNuml	Ship Date	OrderQty	ProductID	SpecialOff	UnitPrice	UnitPriceD	LineTotal	Shipping Status
2	43659	1	4911-403C-98	5/31/2021	1	776	1	2024.994	0	2024.994	Shipped
3	43660	13			1	762	1	419.4589	0	419.4589	Not Shipped
4	43661	15	4E0A-4F89-AE	5/31/2021	1	745	1	809.76	0	809.76	Shipped
5	43662	30	2E53-4802-85	5/31/2021	3	764	1	419.4589	0	1258.3767	Shipped
6	43663	52	1E90-4FBF-B6	5/31/2021	1	760	1	419.4589	0	419.4589	Shipped
7	43664	53			1	772	1	2039.994	0	2039.994	Not Shipped
8	43665	61	19F0-4638-8E	5/31/2021	2	711	1	20.1865	0	40.373	Shipped
9	43666	71	D46A-40CA-8D	5/31/2021	1	764	1	419.4589	0	419.4589	Shipped
10	43667	77	4DFB-4B10-A6	5/31/2021	3	710	1	5.7	0	17.1	Shipped
11	43668	81			3	756	1	874.794	0	2624.382	Not Shipped
12	43669	110	B65C-4867-86	5/31/2021	1	747	1	714.7043	0	714.7043	Shipped
13	43670	111	F101-4649-85	5/31/2021	1	710	1	5.7	0	5.7	Shipped
14	43671	115	DFD9-41B7-94	5/31/2021	1	753	1	2146.962	0	2146.962	Shipped
15	43672	126	F4B5-48D0-BA	5/31/2021	6	709	1	5.7	0	34.2	Shipped
16	43673	129	260F-4DCF-A1	5/31/2021	1	754	1	874.794	0	874.794	Shipped
17	43674	140			3	758	1	874.794	0	2624.382	Not Shipped
18	43675	141	5069-4470-BE	5/31/2021	4	761	1	419.4589	0	1677.8356	Shipped
19	43676	150	11BA-4D19-B7	5/31/2021	2	776	1	2024.994	0	4049.988	Shipped
20	43677	155	8E3A-4564-99	5/31/2021	3	715	1	28.8404	0	86.5212	Shipped
21	43678	167	FBD8-4CE4-8B	5/31/2021	1	760	1	419.4589	0	419.4589	Shipped
22	43679	186	918F-49F3-AD	5/31/2021	1	760	1	419.4589	0	419.4589	Shipped
23	43680	189	FF1F-4DD0-98	5/31/2021	3	760	1	419.4589	0	1258.3767	Shipped

Microsoft Excel IF Function for Shipping Status (Used with permission from Microsoft.)

The logical function we create can determine whether or not the date field is blank and can surface a shipped or not shipped status field accordingly.

As another example, consider the NULL values that you will see when there are unanswered survey questions. If respondents have failed to respond to several key questions, we could write an IF function that returns information stating these critical questions were not answered. This is a way to replace a missing value with an appropriate result.



	Respondent ID	Missed Key Que	What is your race or ethnicity?	What is your	What is	What is the highest leve	How do you classif	How long have you k	What is your current job title	How many years of experience
2	11626900620	Keep	Hispanic or Latino	MALE	55-64	Master's Degree or Higher	Self Employed/Entrep	Less than one year	Office Assistant	Ten years or more
3	11566377582	Remove	Asian	MALE	25-34	Bachelor's Degree	Full Time Employee			
4	11542009435	Keep	Self-describe below:	FEMALE	35-44	Bachelor's Degree	Full Time Employee	Five to ten years	Data Analytics Specialist	Five to ten years
5	11531874158	Remove	Black or African American	FEMALE	45-54	Vocational/Technical Schoo	Unemployed			
6	11530927401	Keep	Black or African American	FEMALE	35-44	Bachelor's Degree	Full Time Employee	Less than one year	data analyst	Five to ten years
7	11526400319	Keep	Hispanic or Latino	FEMALE	45-54	Bachelor's Degree	Full Time Employee	Ten years or more	Teacher	Ten years or more
8	11525731983	Keep	Hispanic or Latino	MALE	35-44	Some College	Full Time Employee	Ten years or more	Senior Project Lead	Ten years or more
9	11524569349	Keep	Self-describe below:	FEMALE	55-64	High School Diploma/GED	Temporary/Seasonal	Ten years or more	Concessions & baker	Ten years or more
10	11523730380	Keep	Black or African American	FEMALE	45-54	Associate's Degree	Full Time Employee	Less than one year	Credentials & Contract Superviso	Ten years or more
11	11523651620	Keep	White	FEMALE	25-34	Bachelor's Degree	Full Time Employee	Less than one year	Analytics Consultant	Five to ten years
12	11523588583	Keep	Hispanic or Latino	FEMALE	45-54	Master's Degree or Higher	Full Time Employee	One to five years	Assistant principal	Ten years or more
13	11522617323	Keep	Hispanic or Latino	FEMALE	45-54	Bachelor's Degree	Full Time Employee	Five to ten years	Case manager	Ten years or more
14	11522429018	Remove	Hispanic or Latino	FEMALE	18-24	High School Diploma/GED	Unemployed			
15	11522426288	Keep	Hispanic or Latino	FEMALE	45-54	Associate's Degree	Full Time Employee	Ten years or more	Eligibility Worker	Ten years or more
16	11522295831	Keep	Hispanic or Latino	FEMALE	45-54	Master's Degree or Higher	Full Time Employee	Ten years or more	High school counselor	Ten years or more
17	11522262326	Keep	Hispanic or Latino	FEMALE	45-54	Master's Degree or Higher	Full Time Employee	Five to ten years	Administrative Officer	Five to ten years
18	11521976824	Keep	Hispanic or Latino	FEMALE	45-54	Master's Degree or Higher	Full Time Employee	Ten years or more	Elementary Teacher	Ten years or more
19	11521865889	Keep	Hispanic or Latino	MALE	35-44	Associate's Degree	Full Time Employee	One to five years	Electrician Apprentice	Ten years or more
20	11521695240	Remove	White	FEMALE	25-34	Bachelor's Degree	Full Time Employee			One to five years
21	11521691640	Remove	Hispanic or Latino	FEMALE	65+	Bachelor's Degree	Unemployed			

Survey Logical Function in Microsoft Excel (Used with permission from Microsoft.)

Assume we have decided that surveys which were not fully completed should be removed from our analysis, because we require a complete record for each respondent (including responses to those key questions) to gain the most meaning out of the data set. The logical function can be used to remove the records that do not have complete answers from our sample. We can create a field that allows us to easily determine what records can be kept and what are to be removed.

It should be noted that in many cases, the decision will be made to include all responses regardless of their completeness, because there are still insights to be found in the answers to the questions that the respondent did complete. Regardless of whatever decision is made for data handling, the action should be documented and noted somewhere when the analysis findings are published.

It is important to know that null values should not be replaced with one of the answer choices from a list just because the field wasn't filled out. Consider a survey question with yes or no answer options. Inexperienced analysts might choose the default "no" option to replace the null, but this would be a mistake.

Review Activity:

Missing Values

Answer the following questions:

1. **What appears in a data set when a field has no value?**

2. **Describe two examples of how NULL values can be useful when working with data?**

3. **When a survey is completed and an answer is skipped, what can you do with that information?**

Topic 6D

Address Invalid Data

EXAM OBJECTIVES COVERED

2.2 Identify common reasons for cleansing and profiling datasets.

When data is not correct, you will need to address the issue by replacing it with valid data or removing it from your data set entirely. Invalid data is often hard to spot, but if you know how to look and test for it, then you will end up with better, more accurate data. There are many techniques we can use to correct invalid data, including removing or replacing the data. The goal is to first verify whether the data is valid. After that, we can determine the best approach to addressing any invalid data depending on business requirements and needs.

Identifying Invalid Data

Invalid data is data that is incorrect. Data could be invalid for many reasons, and there are a number of situations in which you may encounter invalid data in your work as a data analyst. Here are a few reasons that invalid data occurs.

- *Data can become invalid over time*. Suppose that product data has a tax percentage that is hard coded into the system at 9%. At a future date, that tax percentage by law is changed to 10%. All the values that were generated from that 9% are now invalid, because the tax percentage should be 10%. As this scenario shows, understanding the math behind the data can help you detect issues.

- *Survey data may be invalid if a question is deemed invalid.* Suppose when working with survey data, you find that a survey used leading questions, or the answer options had inadvertent bias. There is no way to go backward and reframe the questions (and answers), so the invalid data must be removed from the analysis.

- *Outliers or extreme values can sometimes indicate invalid data*. These values that lie far outside the normal range could signal a technical issue, or a problem with the collection or extraction of the data.

- *Data that cannot possibly be correct is invalid*. Data that is impossible needs to be investigated. Examples of impossible data include invalid zip codes, exceptional values for a transaction amount (like 999999), and even invalid dates.

Issues with Invisible Characters

Data can also become invalid when something as simple as a mistake in manual entry occurs. Here are a few character-based issues that can lead to invalid data.

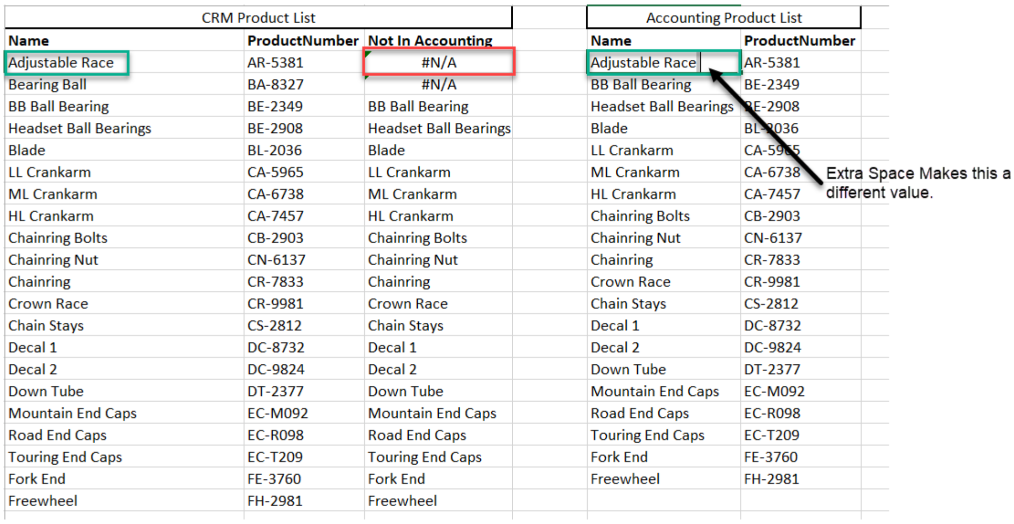
- **Non-printable characters** are characters that do not produce a written symbol, such as spaces and tabs. These types of characters will present challenges when working with data and need to be removed.

- **Leading spaces** are spaces at the front of a field of information.

- **Trailing spaces** are spaces at the end of a field of information.

Non-printable characters can cause problems when you are working with data. For example, " world of data " (with spaces) and "world of data" (no spaces) are not a match, because a data program will read the spaces as actual valid characters. The analyst can write different functions and commands to address this type of issue, as discussed in a later lesson.

To demonstrate how non-printable characters can cause invalid data, let's revisit the example in which we used a VLOOKUP function in Excel to find out whether products were contained in both the CRM and accounting systems. Expanding on this example, suppose we focus on product names to determine if the products are in both systems. As the following image shows, the function tells us that the product "Adjustable Race" doesn't exist in the accounting system, but it is clearly visible to us in the accounting product list. It's hard to see, but there's actually a trailing space in this product name in the accounting product list, which causes a NULL value to be returned. The trailing space makes these products technically different. The system doesn't know that the space is invisible; it sees it as an actual valid character and cannot find a match for that product name.



CRM Product List

Name	ProductNumber	Not In Accounting
Adjustable Race	AR-5381	#N/A
Bearing Ball	BA-8327	#N/A
BB Ball Bearing	BE-2349	BB Ball Bearing
Headset Ball Bearings	BE-2908	Headset Ball Bearings
Blade	BL-2036	Blade
LL Crankarm	CA-5965	LL Crankarm
ML Crankarm	CA-6738	ML Crankarm
HL Crankarm	CA-7457	HL Crankarm
Chainring Bolts	CB-2903	Chainring Bolts
Chainring Nut	CN-6137	Chainring Nut
Chainring	CR-7833	Chainring
Crown Race	CR-9981	Crown Race
Chain Stays	CS-2812	Chain Stays
Decal 1	DC-8732	Decal 1
Decal 2	DC-9824	Decal 2
Down Tube	DT-2377	Down Tube
Mountain End Caps	EC-M092	Mountain End Caps
Road End Caps	EC-R098	Road End Caps
Touring End Caps	EC-T209	Touring End Caps
Fork End	FE-3760	Fork End
Freewheel	FH-2981	Freewheel

Accounting Product List

Name	ProductNumber
Adjustable Race	AR-5381
BB Ball Bearing	BE-2349
Headset Ball Bearings	BE-2908
Blade	BL-2036
LL Crankarm	CA-5965
ML Crankarm	CA-6738
HL Crankarm	CA-7457
Chainring Bolts	CB-2903
Chainring Nut	CN-6137
Chainring	CR-7833
Crown Race	CR-9981
Chain Stays	CS-2812
Decal 1	DC-8732
Decal 2	DC-9824
Down Tube	DT-2377
Mountain End Caps	EC-M092
Road End Caps	EC-R098
Touring End Caps	EC-T209
Fork End	FE-3760
Freewheel	FH-2981

Using the VLOOKUP Function in Microsoft Excel to Compare Product Names in the CRM and Accounting Systems (Used with permission from Microsoft.)

In this scenario, you would want to correct the product name by either removing the trailing space or using a function that ignores any extra spaces. Your organization's business rules will dictate how you handle the fix.

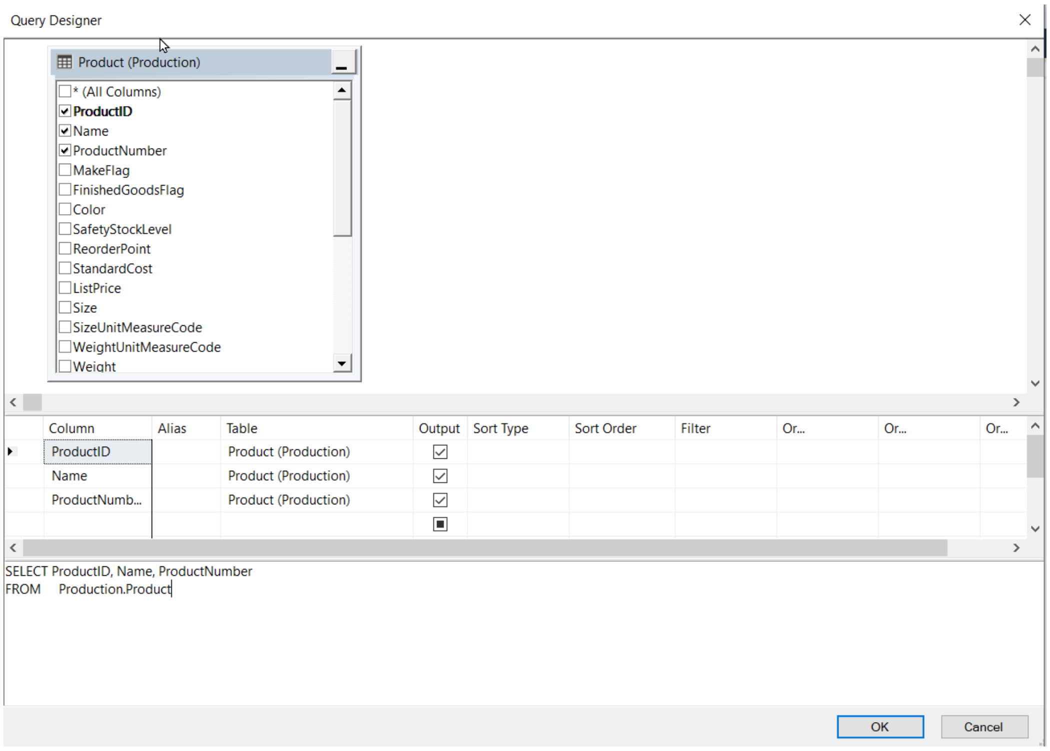
While it may seem like the function didn't work properly, a valid method was used to determine if there was a match between the product lists. When you match information between data systems, it is a very literal process. What a computer sees and what a human may see can trick you, unless you know to look for issues that can lead to invalid data.

Removing Invalid Data

After determining that data is invalid, one way of addressing this issue is to remove (or not include) the data in question. Removing invalid data could be as simple as just removing that field from the data set. It could also mean not including that data in a query.

The decision to remove or not include data is simple. Discovering that data may not be valid is the bigger, more critical challenge for the analyst.

Note the visual example shown in Query Designer involving a product table. All of the available fields from the table are shown, but you could simply check or write in the needed fields, and not include the others. You can remove columns by writing a SELECT statement that doesn't include them. In the screenshot below, only the ProductID, Name, and ProductNumber fields are included in the query.



☐ * (All Columns)
☑ ProductID
☑ Name
☑ ProductNumber
☐ MakeFlag
☐ FinishedGoodsFlag
☐ Color
☐ SafetyStockLevel
☐ ReorderPoint
☐ StandardCost
☐ ListPrice
☐ Size
☐ SizeUnitMeasureCode
☐ WeightUnitMeasureCode
☐ Weight

Column	Alias	Table	Output	Sort Type	Sort Order	Filter	Or...	Or...	Or...
ProductID		Product (Production)	☑						
Name		Product (Production)	☑						
ProductNumb...		Product (Production)	☑						
			■						

```
SELECT ProductID, Name, ProductNumber
FROM    Production.Product
```

Query to Include ProductID, Name, and ProductNumber in AdventureWorks Database in Microsoft SQL Server Management Studio (Used with permission from Microsoft.)

Replacing Invalid Data with Valid Data

Invalid data can also sometimes be addressed by changing it to the correct value. For example, suppose that a customer satisfaction survey asks customers to type in the product that they ordered, and what the customer types in might not always match the official name of the product. An analyst could change the invalid entry (what the customer typed) to valid data (the official product name), but only if they are able to verify that the customer is indeed referring to that particular product.

At times, it may not be possible to correct invalid data to make it useable. In this case, it must be either removed or replaced with a value indicating that the information can't be validated. When correcting or replacing invalid data, the goal is to determine the best way to address the issue so as to meet business requirements.

Review Activity:

Invalid Data

Answer the following questions:

1. **What is invalid data?**

2. **When writing a query to create a data set, what should you do with the fields that have invalid data?**

3. **How can non-printable characters, leading spaces, and trailing spaces cause invalid data?**

4. **What should you do when you encounter invalid values that can definitively be defined?**

Topic 6E

Convert Data to Meet Specifications

EXAM OBJECTIVES COVERED

2.2 Identify common reasons for cleansing and profiling datasets.

As an analyst, you will not only use data for analysis and visuals, but will also move data from system to system. It could be that you are moving historical data from an old system to a new system, or you could just be moving data that is redundant between systems. Databases are designed so that fields have specific data types. When you are manipulating data and preparing it to be stored somewhere else, that data must meet the system's expectations or the data will be mismatched and fail to load properly. Data also must meet the specifications for analysis. If it fails to do so, it may need to be converted to a different data type.

Data That Does Not Meet Specifications

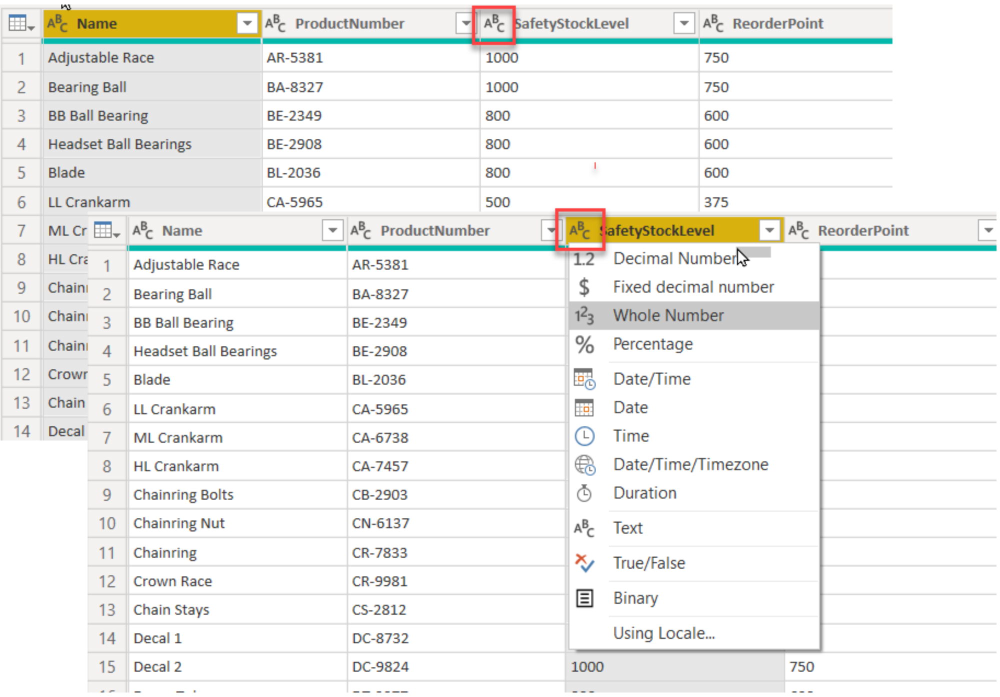
People often discover that data doesn't meet specifications when it fails to load into the system, or when the system doesn't load all of the data. When software gives you the ability to bulk load data to its underlying database or data storage, then it will often provide the specifications to load that data. Be sure to explore the software help menu to find this information. Software vendors will provide a template to use for loading data.

Suppose you have been tasked with transferring missing products to the accounting system from the CRM system. However, the list of products fails to transfer because of a mismatch between the data types. In this situation, you would hope that the system provides meaningful messages that let you know exactly what failed, but that is not always the case. Often, a failed transfer often results from a field or two not having the expected data type, so you must identify the fields causing the issue and then convert them to meet specifications.

Converting Data Types

When data fails to meet the specifications for analysis or visuals, you may need to convert data types. For example, if you need to calculate numbers, then your data type must be a number. The same goes for dates—if you need date-specific calculations, then you must work with data that is set to a date data type.

No single data software is identical in its methods for using functions to convert data types, but all data software will have this capability. Tools such as Power Query and Tableau allow you to adjust the data types without having to write a separate calculation, as shown in the next set of screenshots.



	Name	ProductNumber	SafetyStockLevel	ReorderPoint
1	Adjustable Race	AR-5381	1000	750
2	Bearing Ball	BA-8327	1000	750
3	BB Ball Bearing	BE-2349	800	600
4	Headset Ball Bearings	BE-2908	800	600
5	Blade	BL-2036	800	600
6	LL Crankarm	CA-5965	500	375
7	ML Cr			
8	HL Cra			
9	Chaini			
10	Chaini			
11	Chaini			
12	Crowr			
13	Chain			
14	Decal			

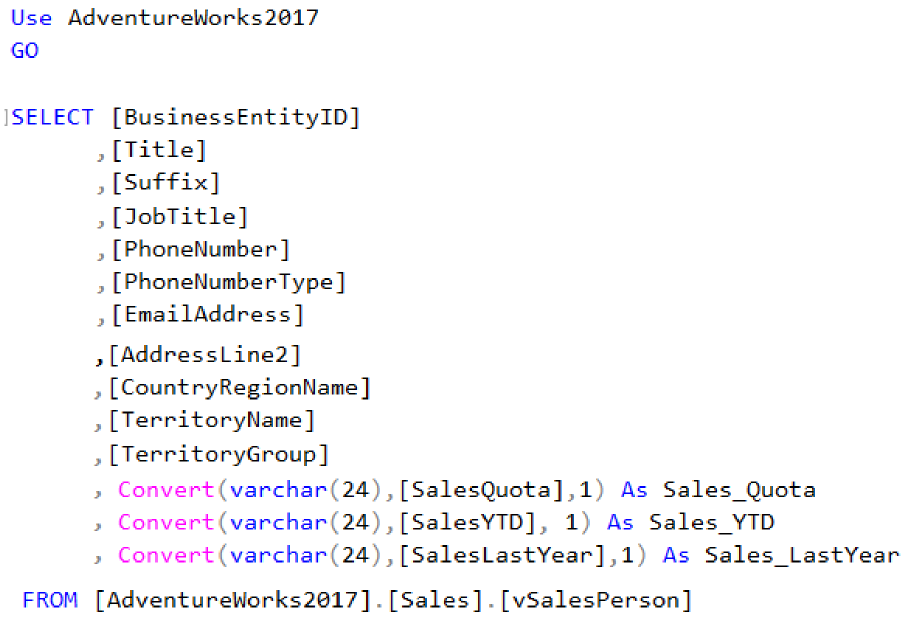
	Name	ProductNumber	SafetyStockLevel	ReorderPoint
1	Adjustable Race	AR-5381		
2	Bearing Ball	BA-8327		
3	BB Ball Bearing	BE-2349		
4	Headset Ball Bearings	BE-2908		
5	Blade	BL-2036		
6	LL Crankarm	CA-5965		
7	ML Crankarm	CA-6738		
8	HL Crankarm	CA-7457		
9	Chainring Bolts	CB-2903		
10	Chainring Nut	CN-6137		
11	Chainring	CR-7833		
12	Crown Race	CR-9981		
13	Chain Stays	CS-2812		
14	Decal 1	DC-8732		
15	Decal 2	DC-9824	1000	750

Data Type Conversion in Microsoft Power Query (Used with permission from Microsoft.)

In Microsoft Power Query, the data types are read from the first row of data. You will always want to validate that the data type is as intended. It can be converted by selecting the appropriate data type from the button to the left of the field name, as you see in this screenshot.

It is important as a data analyst that you are aware of the data types that are associated with the data you are working with. You will want to add data type validation, which means validating the data types are correct, to your approach on any data project.

In some software or scripts, you can write functions that will allow you to perform any type of conversion needed to meet the requirements. This example script is converting the SalesQuota field to a varchar data type with a 24-character length, and renaming this conversion with a label called "Sales _Quota."



```
Use AdventureWorks2017
GO

SELECT [BusinessEntityID]
      ,[Title]
      ,[Suffix]
      ,[JobTitle]
      ,[PhoneNumber]
      ,[PhoneNumberType]
      ,[EmailAddress]
      ,[AddressLine2]
      ,[CountryRegionName]
      ,[TerritoryName]
      ,[TerritoryGroup]
      , Convert(varchar(24),[SalesQuota],1) As Sales_Quota
      , Convert(varchar(24),[SalesYTD], 1) As Sales_YTD
      , Convert(varchar(24),[SalesLastYear],1) As Sales_LastYear
 FROM [AdventureWorks2017].[Sales].[vSalesPerson]
```

SQL Statement That Converts Sales Quota, Sales YTD, and Sales Last Year (Used with permission from Microsoft.)

Why do databases use field names with no spaces? Spaces are converted to a "%20" when processed internally by servers. Using no spaces is one way to optimize processing. Underscores are a single character where a %20 is 3 additional characters.

Review Activity:

Convert Data to Meet Specifications

Answer the following questions:

1. **You need to create calculations that are date specific, and you discover that the date field is set as text. What should you do?**

2. **What are some of the issues of data not having the correct data type?**

Lesson 6

Summary

After this lesson, you should have a better understanding of the need to profile data to identify the source data, the key data, needed transformations, and the records counts. You should have learned about the impacts of redundant and duplicated data, and how to handle them. You should have also gained strategies for working with missing values and invalid data, and understand how to convert data when it doesn't meet specifications.

Guidelines in Cleansing and Profiling Data

Consider these best practices and guidelines when profiling data and addressing redundant, duplicated, unnecessary, and missing data.

1. Always profile your data sets to understand what is in the data, and the volume of the data.
2. Redundant data is the same data that is stored in multiple places, and the record representing the absolute truth must be determined.
3. Duplicated data is data that is repeated within the same data set and can lead to invalid calculations by inflating results.
4. NULL means that there is no value in a field. The field might appear to be blank or contain the word "null." Determining why the null value exists is important because it will drive what you do when working with this type of value.
5. Functions can be used to replace NULL values in the data with a more meaningful result or to create a field that makes it easier to eliminate data from the set when warranted.
6. Keep in mind the following issues that can cause invalid data: non-printable characters, leading spaces, trailing spaces, data becoming invalid over time, and survey data being invalid.
7. Removing invalid data could be as simple as just removing that field of data from the data set. It could also mean not including that data in a query.
8. Invalid data sometimes can be addressed by changing it to the correct value.
9. Data that does not meet specifications when it fails to load into the system, or when the system does not load all the data, must be converted to the correct data type.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 7

Executing Different Data Manipulation Techniques

LESSON INTRODUCTION

When we encounter a data set that is not meaningful, we must replace the values with more readable information. Data analysts regularly recode data for better analysis and more meaningful reports. We use strategies like deriving variables and imputing values when we encounter a need for data that is not in our data set. These different techniques allow us to create new values from data that is available. On the other hand, sometimes we have more data than we need for our analysis and can leverage aggregation or sampling to focus only on the values we need for further analysis. We will likely discover a need to join data sets from multiple sources and must understand the impact of joins when we query data.

Lesson Objectives

In this lesson, you will do the following:

- Manipulate field data and create variables.
- Transpose and append data.
- Query data.

Topic 7A

Manipulate Field Data and Create Variables

EXAM OBJECTIVES COVERED
2.3 Given a scenario, execute data manipulation techniques.

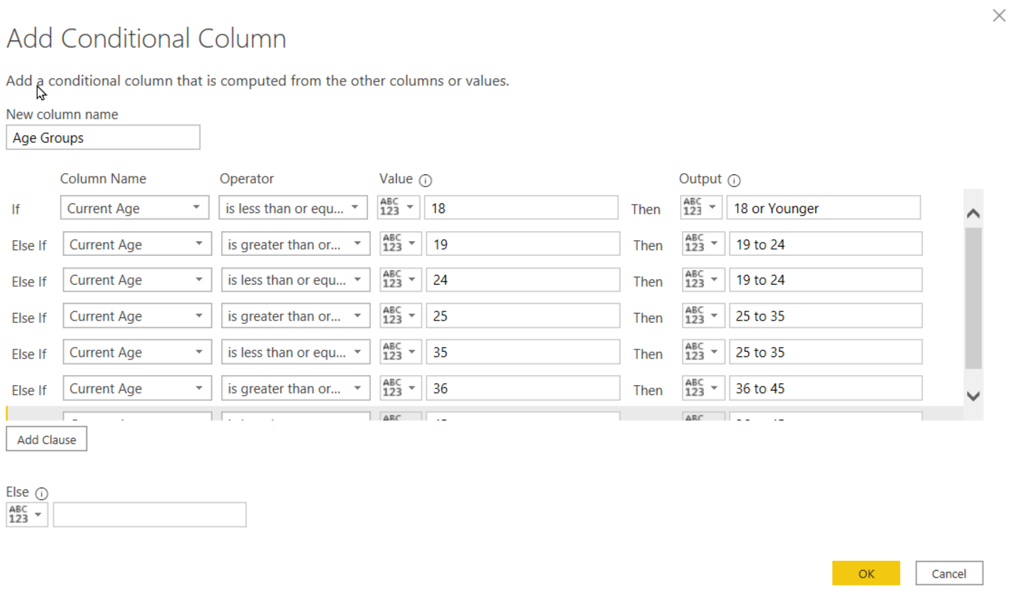
In a perfect world, all data would exist at our fingertips just as we need it for reporting. But the world is certainly not perfect, and thus the data analyst's everyday routine involves recoding data and using functions to derive variables. You will find that values can often be created by imputing them from historical data as needed. You might also need to mask data when reporting, especially when that data is sensitive.

Recoding Numerical and Categorical Data

When we talk about data in the context of recoding, we are referring to a single data field or variable. The act of **recoding** data involves changing the current value of that variable to a different value. We recode data for a few reasons: to make it more meaningful, to group it more effectively for analysis, and to correct the data.

Numerical Data

When recoding numerical data, you can use math to create groups of data. Consider a survey that asks respondents to provide their current age as a whole number. The answers vary based on the respondent, but each response is a number. Now suppose that for the purposes of our report, we want to group the respondents into age groups: 18 or younger, 19 to 24, 25 to 35, 36 to 45, 46 to 55, and 55 or older. The actions you take to perform this grouping will vary depending on the software used. The screenshot below shows the creation of conditional columns in Power BI. This is just one example, but all programs intended for data analysis use will have some function or command that allows you to group numerical data together.



Add Conditional Column

Add a conditional column that is computed from the other columns or values.

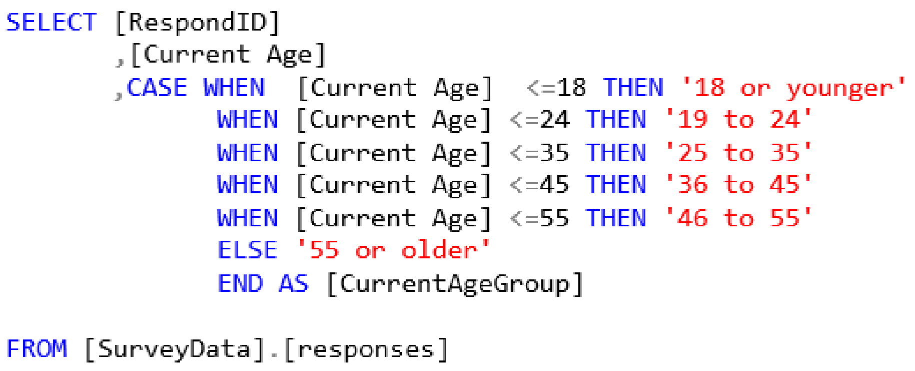
New column name

Age Groups

	Column Name	Operator	Value		Output
If	Current Age	is less than or equ...	18	Then	18 or Younger
Else If	Current Age	is greater than or...	19	Then	19 to 24
Else If	Current Age	is less than or equ...	24	Then	19 to 24
Else If	Current Age	is greater than or...	25	Then	25 to 35
Else If	Current Age	is less than or equ...	35	Then	25 to 35
Else If	Current Age	is greater than or...	36	Then	36 to 45

Current Age Groups Created via Conditional Columns in Microsoft Power BI (Used with permission from Microsoft.)

The case statement below is an example of how you might use an SQL statement to recode ages to age groups. This case statement creates a field called CurrentAgeGroup and groups each of the individual ages based on the current age.



```
SELECT [RespondID]
       ,[Current Age]
       ,CASE WHEN  [Current Age]  <=18 THEN '18 or younger'
             WHEN [Current Age] <=24 THEN '19 to 24'
             WHEN [Current Age] <=35 THEN '25 to 35'
             WHEN [Current Age] <=45 THEN '36 to 45'
             WHEN [Current Age] <=55 THEN '46 to 55'
             ELSE '55 or older'
             END AS [CurrentAgeGroup]

FROM [SurveyData].[responses]
```

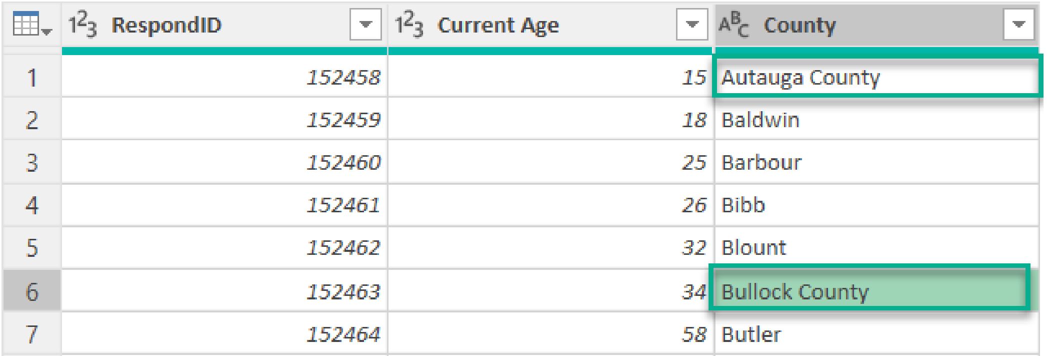
Case Statement in SQL to Recode Ages to Age Groups (Used with permission from Microsoft.)

There are other ways to record numerical data, but the most important takeaway is understanding what recoding is and why it's done.

It is more important to understand the concept of recoding and know when to apply it. Each software you might encounter might use a different approach. It might have a command for recoding, or you may have to create a function to recode your data.

Categorical Data

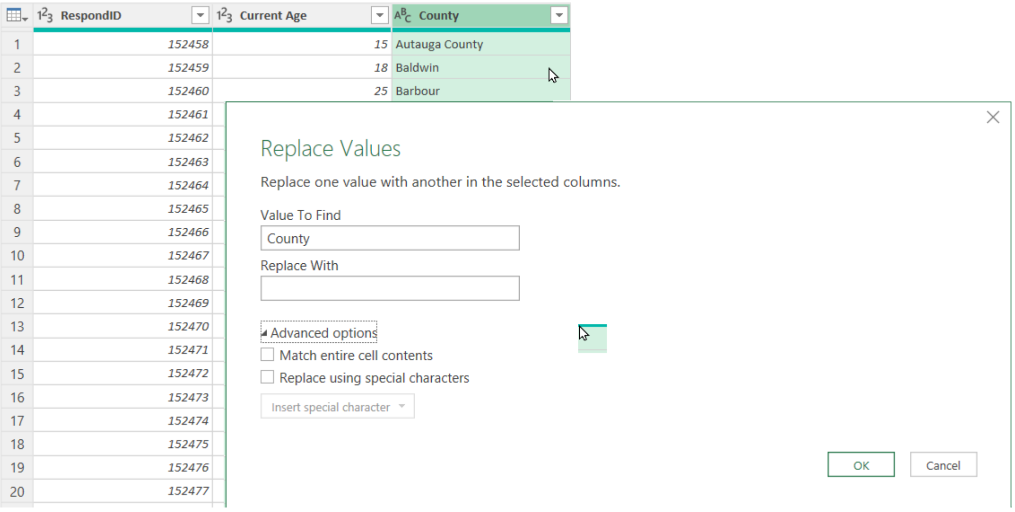
When we are working with categorical data, we sometimes need to recode data in order to correct it. Suppose our survey asked respondents to provide the county they lived in at the time the survey was taken. A free-form text-response object allows respondents to type in the name of the county, but it also opens up the possibility for error.



	RespondID	Current Age	County
1	152458	15	Autauga County
2	152459	18	Baldwin
3	152460	25	Barbour
4	152461	26	Bibb
5	152462	32	Blount
6	152463	34	Bullock County
7	152464	58	Butler

Data Inconsistency Requiring Recoding in Power Query (Used with permission from Microsoft.)

For example, if you look at the screenshot above, you will see that some respondents included the word "county" in their response. We only want the name of the county itself in this data, so we will want to recode this variable to remove the extra word. The screenshot below shows one way this can be done, by using the Replace Values command in Power Query.



	RespondID	Current Age	County
1	152458	15	Autauga County
2	152459	18	Baldwin
3	152460	25	Barbour
4	152461		
5	152462		
6	152463		
7	152464		
8	152465		
9	152466		
10	152467		
11	152468		
12	152469		
13	152470		
14	152471		
15	152472		
16	152473		
17	152474		
18	152475		
19	152476		
20	152477		

Replace Values

Replace one value with another in the selected columns.

Value To Find: County

Replace With:

Advanced options

□ Match entire cell contents

□ Replace using special characters

Using Replace Values Command in Power Query (Used with permission from Microsoft.)

Here, we are removing the word "county" from the column so that all the counties are listed by just their name without including "county" in the field.

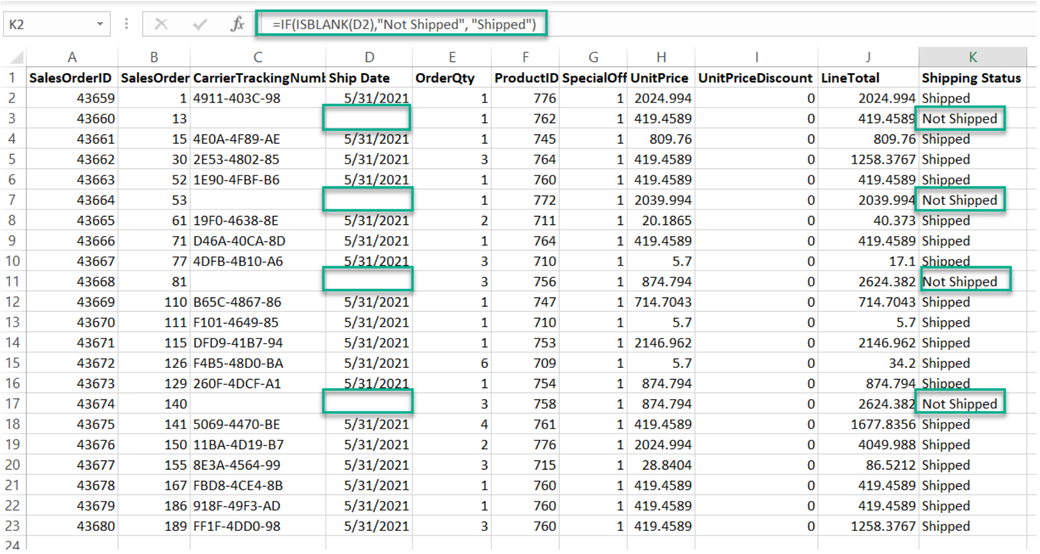
When we recode the data in this way, we don't have to create a new field. However, if we wanted to, we could have duplicated the column, recoded the data in the new column, and named the new column Corrected County. Again, different software programs have different methods of recoding data to remove errors, but what's most important is understanding why we might need to recode data for this purpose.

Derived Variables

Sometimes our analysis requires us to do more than just manipulate data that already exists; there are times when we must create new data from the existing data. A **derived variable** is a data point that is "derived" or created from existing data.

Consider the following example. Suppose we are working with a company that wants to improve its shipping process. In order to do that, we need to calculate the number of days between the time an order is placed and when it has shipped. Even if there is no existing data for this metric, we can create it by deriving the variable "number of days" from the existing data values for the order date and the shipped date. This information can then be further analyzed to determine why some products may take longer than others.

Another example of a derived variable, as shown in the screenshot below, is a newly created field labeled Shipping Status that provides a value of "shipped" or "not shipped" when the Ship Date has a Null value.

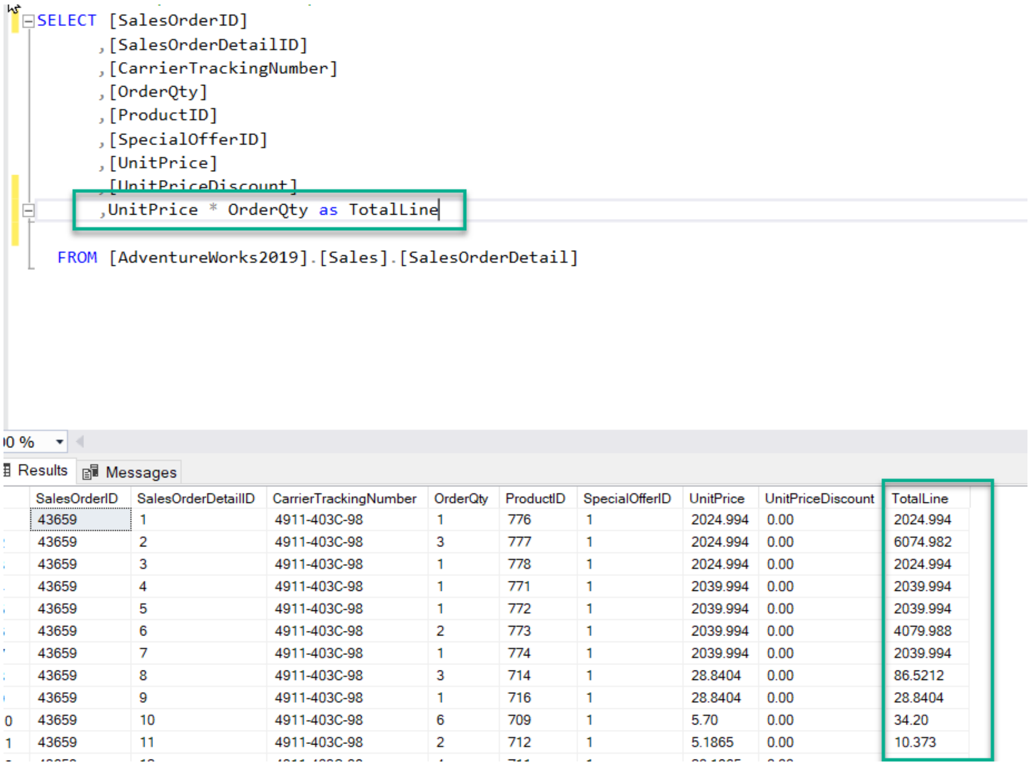


K2 =IF(ISBLANK(D2),"Not Shipped", "Shipped")

	A	B	C	D	E	F	G	H	I	J	K
1	SalesOrderID	SalesOrder	CarrierTrackingNumb	Ship Date	OrderQty	ProductID	SpecialOff	UnitPrice	UnitPriceDiscount	LineTotal	Shipping Status
2	43659	1	4911-403C-98	5/31/2021	1	776	1	2024.994	0	2024.994	Shipped
3	43660	13			1	762	1	419.4589	0	419.4589	Not Shipped
4	43661	15	4E0A-4F89-AE	5/31/2021	1	745	1	809.76	0	809.76	Shipped
5	43662	30	2E53-4802-85	5/31/2021	3	764	1	419.4589	0	1258.3767	Shipped
6	43663	52	1E90-4FBF-B6	5/31/2021	1	760	1	419.4589	0	419.4589	Shipped
7	43664	53			1	772	1	2039.994	0	2039.994	Not Shipped
8	43665	61	19F0-4638-8E	5/31/2021	2	711	1	20.1865	0	40.373	Shipped
9	43666	71	D46A-40CA-8D	5/31/2021	1	764	1	419.4589	0	419.4589	Shipped
10	43667	77	4DFB-4B10-A6	5/31/2021	3	710	1	5.7	0	17.1	Shipped
11	43668	81			3	756	1	874.794	0	2624.382	Not Shipped
12	43669	110	B65C-4867-86	5/31/2021	1	747	1	714.7043	0	714.7043	Shipped
13	43670	111	F101-4649-85	5/31/2021	1	710	1	5.7	0	5.7	Shipped
14	43671	115	DFD9-41B7-94	5/31/2021	1	753	1	2146.962	0	2146.962	Shipped
15	43672	126	F4B5-48D0-BA	5/31/2021	6	709	1	5.7	0	34.2	Shipped
16	43673	129	260F-4DCF-A1	5/31/2021	1	754	1	874.794	0	874.794	Shipped
17	43674	140			3	758	1	874.794	0	2624.382	Not Shipped
18	43675	141	5069-4470-BE	5/31/2021	4	761	1	419.4589	0	1677.8356	Shipped
19	43676	150	11BA-4D19-B7	5/31/2021	2	776	1	2024.994	0	4049.988	Shipped
20	43677	155	8E3A-4564-99	5/31/2021	3	715	1	28.8404	0	86.5212	Shipped
21	43678	167	FBD8-4CE4-8B	5/31/2021	1	760	1	419.4589	0	419.4589	Shipped
22	43679	186	918F-49F3-AD	5/31/2021	1	760	1	419.4589	0	419.4589	Shipped
23	43680	189	FF1F-4DD0-98	5/31/2021	3	760	1	419.4589	0	1258.3767	Shipped
24											

IF Function in Microsoft Excel Showing Shipped or Not Shipped Status (Used with permission from Microsoft.)

It's also important to note that in some software systems, you will see values (such as an order total) on your screen, but that total is not actually stored in the data. Database designers and software developers are trying to optimize storage and processing, and so they will often not store data that can be calculated. We must create it for our reporting purposes, and that is a derived variable because we are using multiple fields to determine the answer. In this screenshot, we are using the UnitPrice and multiplying it by the OrderQty to create the TotalLine field. This provides us the total of that line item.



```
SELECT [SalesOrderID]
      ,[SalesOrderDetailID]
      ,[CarrierTrackingNumber]
      ,[OrderQty]
      ,[ProductID]
      ,[SpecialOfferID]
      ,[UnitPrice]
      ,[UnitPriceDiscount]
      ,UnitPrice * OrderQty as TotalLine

  FROM [AdventureWorks2019].[Sales].[SalesOrderDetail]
```

	SalesOrderID	SalesOrderDetailID	CarrierTrackingNumber	OrderQty	ProductID	SpecialOfferID	UnitPrice	UnitPriceDiscount	TotalLine
	43659	1	4911-403C-98	1	776	1	2024.994	0.00	2024.994
	43659	2	4911-403C-98	3	777	1	2024.994	0.00	6074.982
	43659	3	4911-403C-98	1	778	1	2024.994	0.00	2024.994
	43659	4	4911-403C-98	1	771	1	2039.994	0.00	2039.994
	43659	5	4911-403C-98	1	772	1	2039.994	0.00	2039.994
	43659	6	4911-403C-98	2	773	1	2039.994	0.00	4079.988
	43659	7	4911-403C-98	1	774	1	2039.994	0.00	2039.994
	43659	8	4911-403C-98	3	714	1	28.8404	0.00	86.5212
	43659	9	4911-403C-98	1	716	1	28.8404	0.00	28.8404
0	43659	10	4911-403C-98	6	709	1	5.70	0.00	34.20
1	43659	11	4911-403C-98	2	712	1	5.1865	0.00	10.373

SQL Statement in Microsoft SQL Server Management Studio That Calculates the Total (Used with permission from Microsoft.)

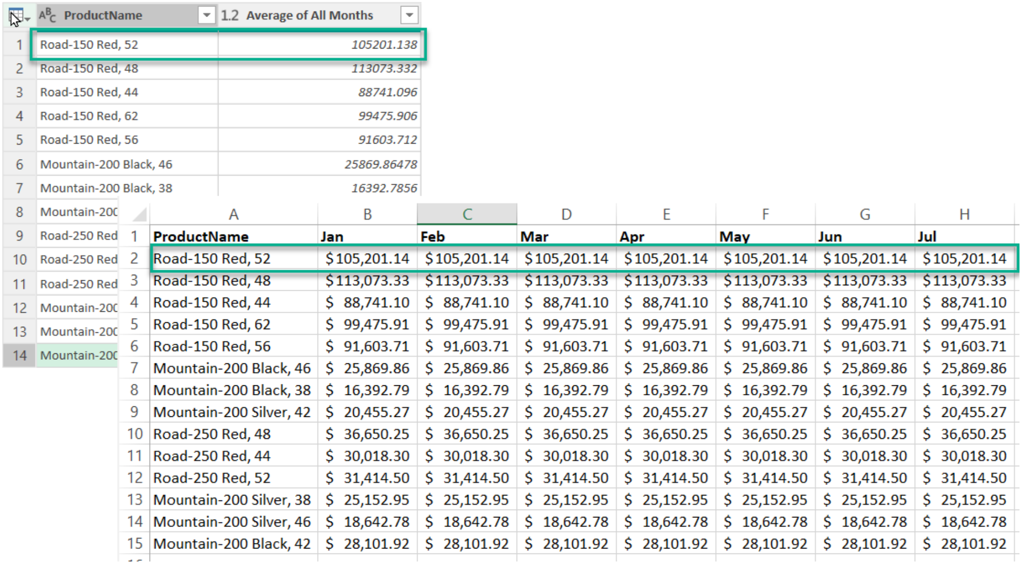
Imputing Values

Imputing values means to replace data with an estimated value. Missing data or null values can cause issues in analysis, and in data software they may even be entirely excluded from your data set.

As you have already discovered, when working with null values there are several ways to review the data set and determine information that could be the replaced value.

There are multiple methods that can be used to impute values. You can choose the average of records and use those to impute values for a record, or you might use a predicted value.

If we are forecasting next year's sales for an organization and we want to visualize the data, there must be a value specified for each month and each product. We will impute the value to be the average of all months combined for each product, and place that value in each month for the next year. This means we are using the same value for all 12 months for each product.



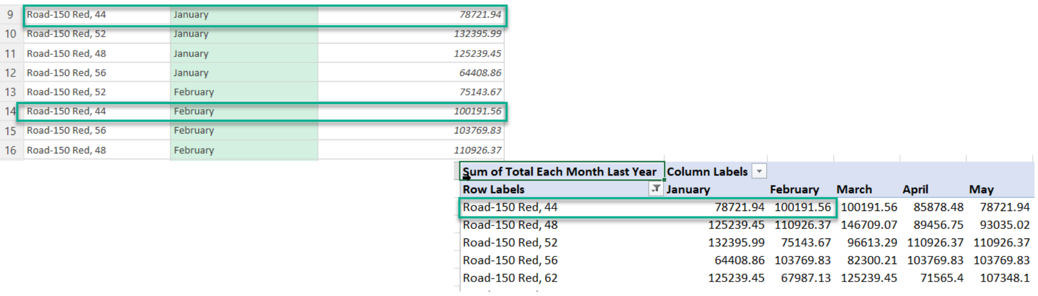
	ProductName	1.2 Average of All Months
1	Road-150 Red, 52	105201.138
2	Road-150 Red, 48	113073.332
3	Road-150 Red, 44	88741.096
4	Road-150 Red, 62	99475.906
5	Road-150 Red, 56	91603.712
6	Mountain-200 Black, 46	25869.86478
7	Mountain-200 Black, 38	16392.7856
8	Mountain-20C	
9	Road-250 Red	
10	Road-250 Red	
11	Road-250 Red	
12	Mountain-20C	
13	Mountain-20C	
14	Mountain-20C	

	A	B	C	D	E	F	G	H
1	ProductName	Jan	Feb	Mar	Apr	May	Jun	Jul
2	Road-150 Red, 52	$105,201.14	$105,201.14	$105,201.14	$105,201.14	$105,201.14	$105,201.14	$105,201.14
3	Road-150 Red, 48	$113,073.33	$113,073.33	$113,073.33	$113,073.33	$113,073.33	$113,073.33	$113,073.33
4	Road-150 Red, 44	$ 88,741.10	$ 88,741.10	$ 88,741.10	$ 88,741.10	$ 88,741.10	$ 88,741.10	$ 88,741.10
5	Road-150 Red, 62	$ 99,475.91	$ 99,475.91	$ 99,475.91	$ 99,475.91	$ 99,475.91	$ 99,475.91	$ 99,475.91
6	Road-150 Red, 56	$ 91,603.71	$ 91,603.71	$ 91,603.71	$ 91,603.71	$ 91,603.71	$ 91,603.71	$ 91,603.71
7	Mountain-200 Black, 46	$ 25,869.86	$ 25,869.86	$ 25,869.86	$ 25,869.86	$ 25,869.86	$ 25,869.86	$ 25,869.86
8	Mountain-200 Black, 38	$ 16,392.79	$ 16,392.79	$ 16,392.79	$ 16,392.79	$ 16,392.79	$ 16,392.79	$ 16,392.79
9	Mountain-200 Silver, 42	$ 20,455.27	$ 20,455.27	$ 20,455.27	$ 20,455.27	$ 20,455.27	$ 20,455.27	$ 20,455.27
10	Road-250 Red, 48	$ 36,650.25	$ 36,650.25	$ 36,650.25	$ 36,650.25	$ 36,650.25	$ 36,650.25	$ 36,650.25
11	Road-250 Red, 44	$ 30,018.30	$ 30,018.30	$ 30,018.30	$ 30,018.30	$ 30,018.30	$ 30,018.30	$ 30,018.30
12	Road-250 Red, 52	$ 31,414.50	$ 31,414.50	$ 31,414.50	$ 31,414.50	$ 31,414.50	$ 31,414.50	$ 31,414.50
13	Mountain-200 Silver, 38	$ 25,152.95	$ 25,152.95	$ 25,152.95	$ 25,152.95	$ 25,152.95	$ 25,152.95	$ 25,152.95
14	Mountain-200 Silver, 46	$ 18,642.78	$ 18,642.78	$ 18,642.78	$ 18,642.78	$ 18,642.78	$ 18,642.78	$ 18,642.78
15	Mountain-200 Black, 42	$ 28,101.92	$ 28,101.92	$ 28,101.92	$ 28,101.92	$ 28,101.92	$ 28,101.92	$ 28,101.92

Imputed Value Using the Total Average of All Months per Product in Microsoft Excel (Used with permission from Microsoft.)

In this example, we are using the "Group By" function in Microsoft Power Query to create the total for each month, and then creating an average of all months. In Microsoft Excel, we display that average for each month and each product. We are using the same annualized average.

We could also predict that each month might look identical to that same month from last year, and place that value into our data set. With this method, we assume that each month of the following year will be exactly as it was in the last year.



9	Road-150 Red, 44	January	78721.94
10	Road-150 Red, 52	January	132395.99
11	Road-150 Red, 48	January	125239.45
12	Road-150 Red, 56	January	64408.86
13	Road-150 Red, 52	February	75143.67
14	Road-150 Red, 44	February	100191.56
15	Road-150 Red, 56	February	103769.83
16	Road-150 Red, 48	February	110926.37

Sum of Total Each Month Last Year	Column Labels				
Row Labels	January	February	March	April	May
Road-150 Red, 44	78721.94	100191.56	100191.56	85878.48	78721.94
Road-150 Red, 48	125239.45	110926.37	146709.07	89456.75	93035.02
Road-150 Red, 52	132395.99	75143.67	96613.29	110926.37	110926.37
Road-150 Red, 56	64408.86	103769.83	82300.21	103769.83	103769.83
Road-150 Red, 62	125239.45	67987.13	125239.45	71565.4	107348.1

Imputing a Value from the Same Month Last Year Using Microsoft Excel and Power Query (Used with permission from Microsoft.)

Here, you can see that we calculated the values for this year based on the information from last year's sales.

Imputation decisions or techniques should be documented in the analysis report. Not all missing values should be imputed; a more accurate analysis may be supported if the data is simply left as missing and noted as such in the analysis report.

Reduction in Data Sets

Reduction in terms of data mining means to reduce the volume of data. However, you must be careful not to invalidate the data when you do this. In simpler terms, you can't just decide to remove or delete records to reduce the volume. Large volumes of data slow down the processing of information, and in some cases, you can use methods like aggregating the data and sampling to reduce the volume without jeopardizing the analysis.

Aggregate Data as a Method of Reduction

A great example of a data set that has gone through a reduction process is the United States Census. Instead of providing us every single record of each census entry, they provide us the data aggregated by different subgroups, like geography, gender, and race/ethnicity, along with their respective aggregated totals. This allows us to have the total number we need without having to create that summary ourselves.

When you are aggregating data, make sure the method you choose doesn't involve cherry-picking, or choosing only the records you want in the analysis. You must be able to communicate how the data was reduced with a valid approach.

Sampling Data

For larger data sets, you will find that sampling the data can be an effective way to reduce data while still producing an accurate outcome. There are multiple methods of sampling data, and the approach you choose will likely be decided by a group of stakeholders on the project or other statisticians that are a part of a larger team. Two common types of sampling are:

- **Simple random sampling**, where each record of data has an equal chance of being selected into the data set used for analysis.
- **Stratified sampling**, where you break your data into subgroups, like gender, and then randomly sample from each of the groups.

There are, of course, other sampling methods that are appropriate for different types of analysis. What's most important to understand is that when you are sampling data, you are not using the entire population of records but only a sample of the records.

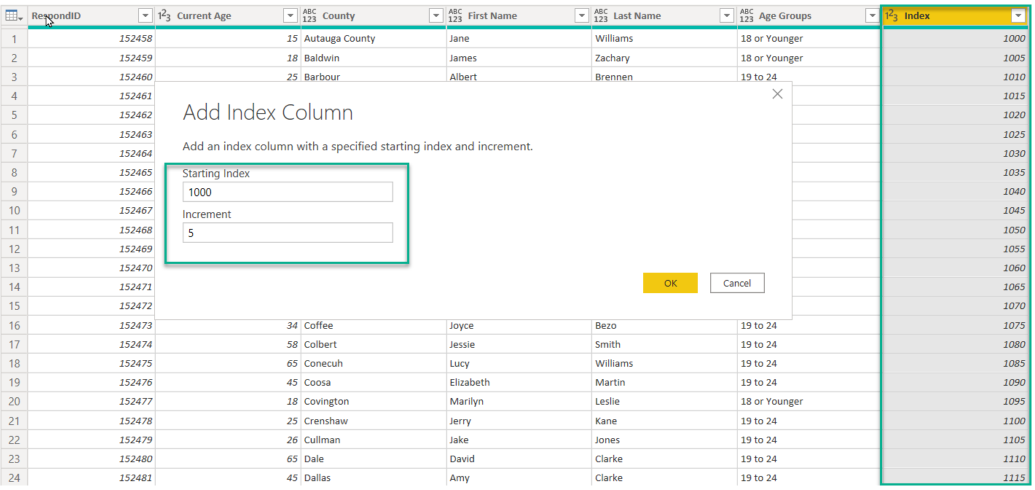
Masking Values

Masking values means to hide or change the original value. For example, a person will enter their social security number into a system, but when we display that data in our reports we will likely mask all but the last four characters, or may even mask the field entirely.

This masking should be applied to any data that is sensitive but still needs to appear in the report. If you do not need to display sensitive data, you could just exclude it from the data set used for your reporting.

Sensitive data is typically known as **personally identifiable information (PII)** . When PII is shared outside of the organization that represents the people who the data is about, that data is required to be masked in order to protect their identities. An example of this is test scores shared outside of the school's records.

When you want to maintain a unique record for a person without disclosing their PII, you can easily create an index field using an index generation tool. An **index field** applies a unique number to a record. The index shown in the screenshot below starts at 1,000 and has a five in the increment. You could also easily just select “start at 0” or “start at 1” to create a unique index field.



	RespondID	Current Age	County	First Name	Last Name	Age Groups	Index
1	152458	15	Autauga County	Jane	Williams	18 or Younger	1000
2	152459	18	Baldwin	James	Zachary	18 or Younger	1005
3	152460	25	Barbour	Albert	Brennen	19 to 24	1010
4	152461						1015
5	152462						1020
6	152463						1025
7	152464						1030
8	152465						1035
9	152466						1040
10	152467						1045
11	152468						1050
12	152469						1055
13	152470						1060
14	152471						1065
15	152472						1070
16	152473	34	Coffee	Joyce	Bezo	19 to 24	1075
17	152474	58	Colbert	Jessie	Smith	19 to 24	1080
18	152475	65	Conecuh	Lucy	Williams	19 to 24	1085
19	152476	45	Coosa	Elizabeth	Martin	19 to 24	1090
20	152477	18	Covington	Marilyn	Leslie	18 or Younger	1095
21	152478	25	Crenshaw	Jerry	Kane	19 to 24	1100
22	152479	26	Cullman	Jake	Jones	19 to 24	1105
23	152480	65	Dale	David	Clarke	19 to 24	1110
24	152481	45	Dallas	Amy	Clarke	19 to 24	1115

Add Index Column

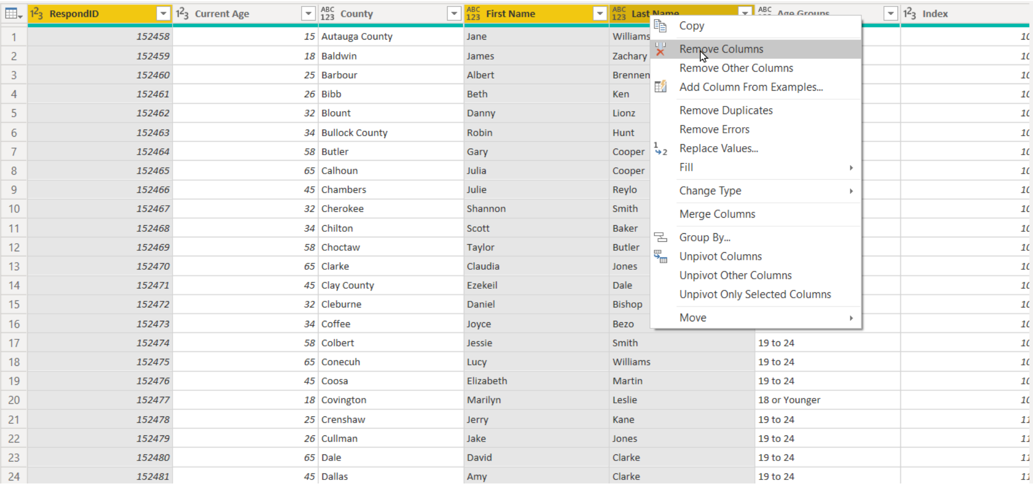
Add an index column with a specified starting index and increment.

Starting Index: 1000

Increment: 5

Index Field in Power BI Power Query (Used with permission from Microsoft.)

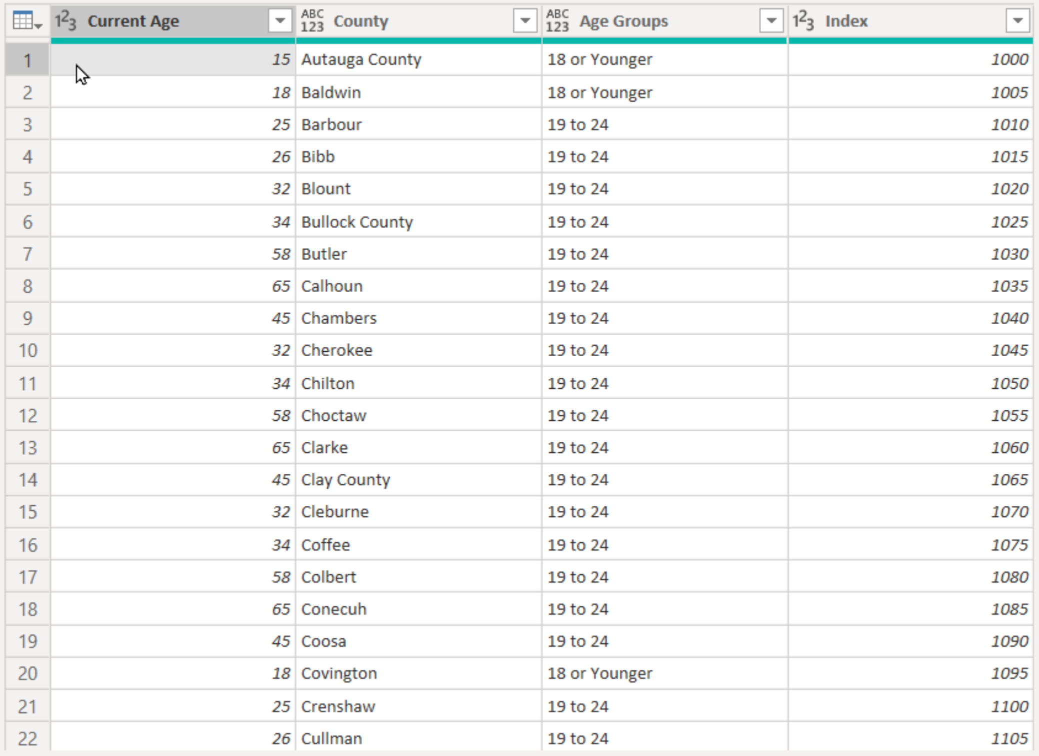
The next step after creating an index would be to remove the data from the query. Please note this does not remove it from the underlying data set.



	RespondID	Current Age	County	First Name	Last Name	Age Groups	Index
1	152458	15	Autauga County	Jane	Williams		10
2	152459	18	Baldwin	James	Zachary		10
3	152460	25	Barbour	Albert	Brennen		10
4	152461	26	Bibb	Beth	Ken		10
5	152462	32	Blount	Danny	Lionz		10
6	152463	34	Bullock County	Robin	Hunt		10
7	152464	58	Butler	Gary	Cooper		10
8	152465	65	Calhoun	Julia	Cooper		10
9	152466	45	Chambers	Julie	Reylo		10
10	152467	32	Cherokee	Shannon	Smith		10
11	152468	34	Chilton	Scott	Baker		10
12	152469	58	Choctaw	Taylor	Butler		10
13	152470	65	Clarke	Claudia	Jones		10
14	152471	45	Clay County	Ezekeil	Dale		10
15	152472	32	Cleburne	Daniel	Bishop		10
16	152473	34	Coffee	Joyce	Bezo		10
17	152474	58	Colbert	Jessie	Smith	19 to 24	10
18	152475	65	Conecuh	Lucy	Williams	19 to 24	10
19	152476	45	Coosa	Elizabeth	Martin	19 to 24	10
20	152477	18	Covington	Marilyn	Leslie	18 or Younger	10
21	152478	25	Crenshaw	Jerry	Kane	19 to 24	11
22	152479	26	Cullman	Jake	Jones	19 to 24	11
23	152480	65	Dale	David	Clarke	19 to 24	11
24	152481	45	Dallas	Amy	Clarke	19 to 24	11

Remove PII from the Query Data (Used with permission from Microsoft.)

Your data set now has a unique identifier for each record that replaces the original key field, which was the respondent ID. The first name and last name fields have also been removed for privacy reasons.



	Current Age	County	Age Groups	Index
1	15	Autauga County	18 or Younger	1000
2	18	Baldwin	18 or Younger	1005
3	25	Barbour	19 to 24	1010
4	26	Bibb	19 to 24	1015
5	32	Blount	19 to 24	1020
6	34	Bullock County	19 to 24	1025
7	58	Butler	19 to 24	1030
8	65	Calhoun	19 to 24	1035
9	45	Chambers	19 to 24	1040
10	32	Cherokee	19 to 24	1045
11	34	Chilton	19 to 24	1050
12	58	Choctaw	19 to 24	1055
13	65	Clarke	19 to 24	1060
14	45	Clay County	19 to 24	1065
15	32	Cleburne	19 to 24	1070
16	34	Coffee	19 to 24	1075
17	58	Colbert	19 to 24	1080
18	65	Conecuh	19 to 24	1085
19	45	Coosa	19 to 24	1090
20	18	Covington	18 or Younger	1095
21	25	Crenshaw	19 to 24	1100
22	26	Cullman	19 to 24	1105

Final Data in Query with All the PII Removed (Used with permission from Microsoft.)

Review Activity:

Manipulate Field Data and Create Variables

Answer the following questions:

1. **What is one reason why data may need to be recoded?**

2. **Imagine a data set with a column for Start Time and a column for End Time, where a column for Total Time has been added to the data set. The Total Time column would be an example of what?**

3. **What does it mean to impute values?**

4. **Why would a value be imputed?**

5. **Reduction can be achieved through what two methods?**

Topic 7B

Transpose and Append Data

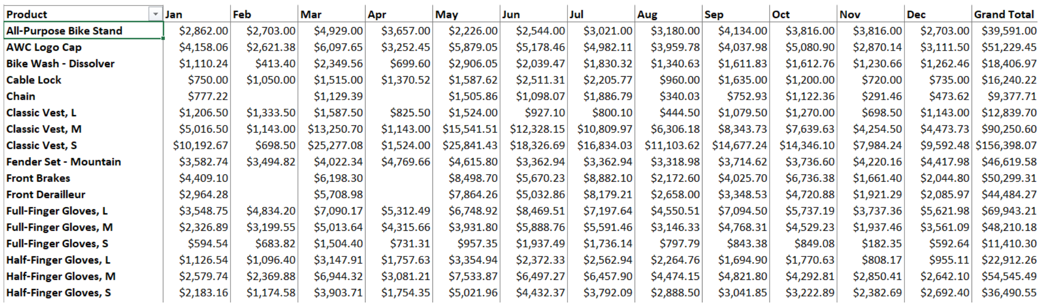
EXAM OBJECTIVES COVERED

2.3 Given a scenario, execute data manipulation techniques.

As data professionals, we may be given data that has already been aggregated and put in pivot format, meaning it has a row, column, and value. This data can be difficult to work with unless we transform it into a traditional data set that is vertical and in tabular or record format. Knowing how to change data from a wide format to a vertical record set can be helpful when all we have are totals in a cross tab format. We may also find that we need to append data from one location to the next, or append data from multiple sets to create a new comprehensive data set. The ability to transpose and append data is crucial, as you will need to manipulate data.

Transposing Data

There are times when you will need to work with data that is given to you in a cross tab or pivot format, meaning it is not in the style of a record set. Imagine you have a data set that shows a company's product sales performance for each month of the year. This data set contains the name of each product sold in the first column. The months (January through December) are set as the first row of the data set, and each value of that product's performance or sales for that month will appear under the appropriate month column and the appropriate product row. You do not have access to the individual line data that was used to create the pivot.

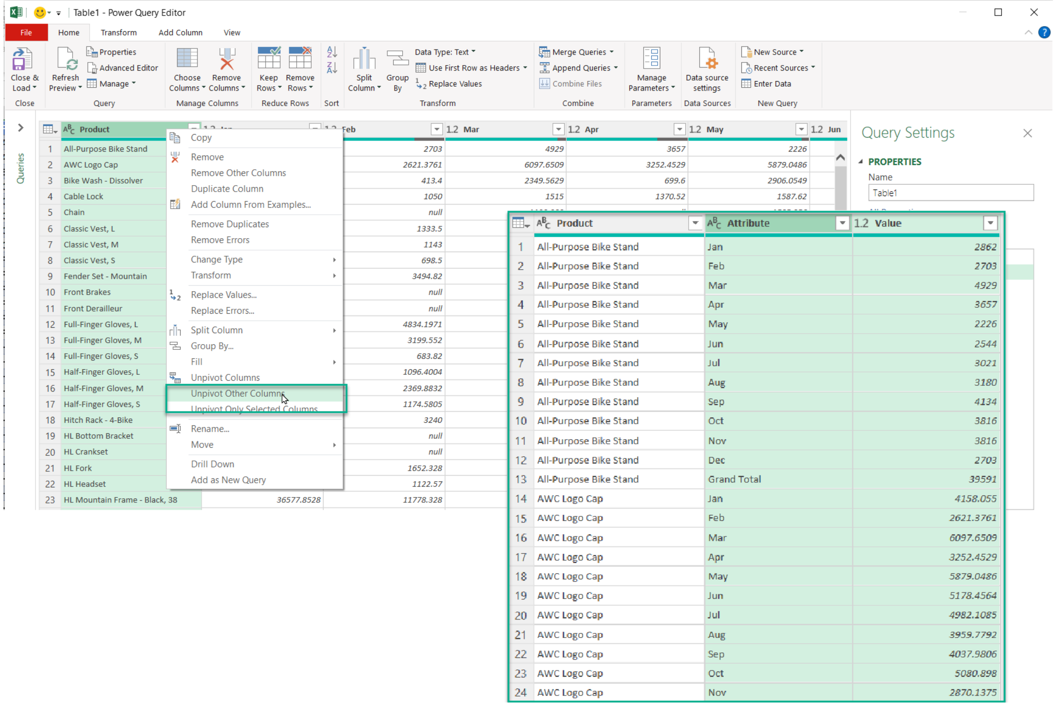


Product	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Grand Total
All-Purpose Bike Stand	$2,862.00	$2,703.00	$4,929.00	$3,657.00	$2,226.00	$2,544.00	$3,021.00	$3,180.00	$4,134.00	$3,816.00	$3,816.00	$2,703.00	$39,591.00
AWC Logo Cap	$4,158.06	$2,621.38	$6,097.65	$3,252.45	$5,879.05	$5,178.46	$4,982.11	$3,959.78	$4,037.98	$5,080.90	$2,870.14	$3,111.50	$51,229.45
Bike Wash - Dissolver	$1,110.24	$413.40	$2,349.56	$699.60	$2,906.05	$2,039.47	$1,830.32	$1,340.63	$1,611.83	$1,612.76	$1,230.66	$1,262.46	$18,406.97
Cable Lock	$750.00	$1,050.00	$1,515.00	$1,370.52	$1,587.62	$2,511.31	$2,205.77	$960.00	$1,635.00	$1,200.00	$720.00	$735.00	$16,240.22
Chain	$777.22		$1,129.39		$1,505.86	$1,098.07	$1,886.79	$340.03	$752.93	$1,122.36	$291.46	$473.62	$9,377.71
Classic Vest, L	$1,206.50	$1,333.50	$1,587.50	$825.50	$1,524.00	$927.10	$800.10	$444.50	$1,079.50	$1,270.00	$698.50	$1,143.00	$12,839.70
Classic Vest, M	$5,016.50	$1,143.00	$13,250.70	$1,143.00	$15,541.51	$12,328.15	$10,809.97	$6,306.18	$8,343.73	$7,639.63	$4,254.50	$4,473.73	$90,250.60
Classic Vest, S	$10,192.67	$698.50	$25,277.08	$1,524.00	$25,841.43	$18,326.69	$16,834.03	$11,103.62	$14,677.24	$14,346.10	$7,984.24	$9,592.48	$156,398.07
Fender Set - Mountain	$3,582.74	$3,494.82	$4,022.34	$4,769.66	$4,615.80	$3,362.94	$3,362.94	$3,318.98	$3,714.62	$3,736.60	$4,220.16	$4,417.98	$46,619.58
Front Brakes	$4,409.10		$6,198.30		$8,498.70	$5,670.23	$8,882.10	$2,172.60	$4,025.70	$6,736.38	$1,661.40	$2,044.80	$50,299.31
Front Derailleur	$2,964.28		$5,708.98		$7,864.26	$5,032.86	$8,179.21	$2,658.00	$3,348.53	$4,720.88	$1,921.29	$2,085.97	$44,484.27
Full-Finger Gloves, L	$3,548.75	$4,834.20	$7,090.17	$5,312.49	$6,748.92	$8,469.51	$7,197.64	$4,550.51	$7,094.50	$5,737.19	$3,737.36	$5,621.98	$69,943.21
Full-Finger Gloves, M	$2,326.89	$3,199.55	$5,013.64	$4,315.66	$3,931.80	$5,888.76	$5,591.46	$3,146.33	$4,768.31	$4,529.23	$1,937.46	$3,561.09	$48,210.18
Full-Finger Gloves, S	$594.54	$683.82	$1,504.40	$731.31	$957.35	$1,937.49	$1,736.14	$797.79	$843.38	$849.08	$182.35	$592.64	$11,410.30
Half-Finger Gloves, L	$1,126.54	$1,096.40	$3,147.91	$1,757.63	$3,354.94	$2,372.33	$2,562.94	$2,264.76	$1,694.90	$1,770.63	$808.17	$955.11	$22,912.26
Half-Finger Gloves, M	$2,579.74	$2,369.88	$6,944.32	$3,081.21	$7,533.87	$6,497.27	$6,457.90	$4,474.15	$4,821.80	$4,292.81	$2,850.41	$2,642.10	$54,545.49
Half-Finger Gloves, S	$2,183.16	$1,174.58	$3,903.71	$1,754.35	$5,021.96	$4,432.37	$3,792.09	$2,888.50	$3,041.85	$3,222.89	$2,382.69	$2,692.40	$36,490.55

Data Presented in Wide Format in Excel (Used with permission from Microsoft.)

Viewing data in a pivot format can be extremely impactful for reporting purposes on large data sets. There will be times, however, when you will need this data in a record format rather than pivot format, as this will allow you to create different styles of visuals and perform additional analysis. The data in the pivot format must be transposed in order to structure it in record format.

When you **transpose** data, you are reversing the direction of the data. In our product sales example, you would use the Unpivot command to create a record set that lists the product name, the month, and that month's value on every row.



Name: Table1

	Product	Jan	Feb	Mar	Apr	May	Jun
1	All-Purpose Bike Stand		2703	4929	3657	2226	
2	AWC Logo Cap		2621.3761	6097.6509	3252.4529	5879.0486	
3	Bike Wash - Dissolver		413.4	2349.5629	699.6	2906.0549	
4	Cable Lock		1050	1515	1370.52	1587.62	
5	Chain		null				
6	Classic Vest, L		1333.5				
7	Classic Vest, M		1143				
8	Classic Vest, S		698.5				
9	Fender Set - Mountain		3494.82				
10	Front Brakes		null				
11	Front Derailleur		null				
12	Full-Finger Gloves, L		4834.1971				
13	Full-Finger Gloves, M		3199.552				
14	Full-Finger Gloves, S		683.82				
15	Half-Finger Gloves, L		1096.4004				
16	Half-Finger Gloves, M		2369.8832				
17	Half-Finger Gloves, S		1174.5805				
18	Hitch Rack - 4-Bike		3240				
19	HL Bottom Bracket		null				
20	HL Crankset		null				
21	HL Fork		1652.328				
22	HL Headset		1122.57				
23	HL Mountain Frame - Black, 38	36577.8528	11778.328				

	Product	Attribute	Value
1	All-Purpose Bike Stand	Jan	2862
2	All-Purpose Bike Stand	Feb	2703
3	All-Purpose Bike Stand	Mar	4929
4	All-Purpose Bike Stand	Apr	3657
5	All-Purpose Bike Stand	May	2226
6	All-Purpose Bike Stand	Jun	2544
7	All-Purpose Bike Stand	Jul	3021
8	All-Purpose Bike Stand	Aug	3180
9	All-Purpose Bike Stand	Sep	4134
10	All-Purpose Bike Stand	Oct	3816
11	All-Purpose Bike Stand	Nov	3816
12	All-Purpose Bike Stand	Dec	2703
13	All-Purpose Bike Stand	Grand Total	39591
14	AWC Logo Cap	Jan	4158.055
15	AWC Logo Cap	Feb	2621.3761
16	AWC Logo Cap	Mar	6097.6509
17	AWC Logo Cap	Apr	3252.4529
18	AWC Logo Cap	May	5879.0486
19	AWC Logo Cap	Jun	5178.4564
20	AWC Logo Cap	Jul	4982.1085
21	AWC Logo Cap	Aug	3959.7792
22	AWC Logo Cap	Sep	4037.9806
23	AWC Logo Cap	Oct	5080.898
24	AWC Logo Cap	Nov	2870.1375

Unpivot Command in Power Query (Used with permission from Microsoft.)

The Unpivot command transposes the values into a vertical record set by converting the month field to an attribute and the actual value is then listed in a field labeled "value." We will often change the field name attribute to the appropriate field name after the data has been unpivoted. In the case of this example, we would rename that Attribute field to Month and the Value field to Revenue.

Unpivot commands can be used in many programs. While you may be more familiar with using the paste command in Excel to transpose data, you will likely find that using unpivot commands makes this process even simple.

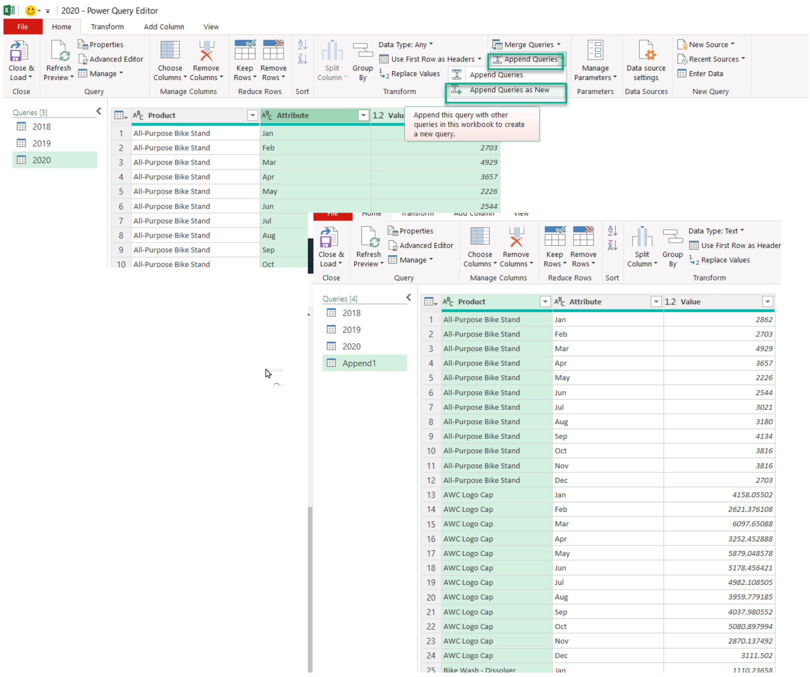
Appending Data

To **append** data means to combine data from one data set with another data set. It's a type of query that we use to perform this action on the data. We can use this type of command to create an inline append or intermediate append.

- An **inline append** will combine data sets until you achieve your final result; all the sets are combined, leaving just the combined set.
- An **intermediate append** will retain the separate data sets and also create a new data set with all the combined data.

Let's return to our earlier example of forecasting the next year's sales numbers. Suppose we have different years with the data in individual pivot format. We can transpose that data to a vertical format and append the data sets into a new data set.

When you are working with appends in tools like Power Query, the functions are called Append Queries. You can append data from one table to another, or you can append data into a new table. In this case, you are not only copying data but also creating an entirely new data set. The Append Query function allows us to take different tables of information and combine them into one.



Append this query with other queries in this workbook to create a new query.

	Product	Attribute	1.2 Valu
1	All-Purpose Bike Stand	Jan	
2	All-Purpose Bike Stand	Feb	2703
3	All-Purpose Bike Stand	Mar	4929
4	All-Purpose Bike Stand	Apr	3657
5	All-Purpose Bike Stand	May	2226
6	All-Purpose Bike Stand	Jun	2544
7	All-Purpose Bike Stand	Jul	
8	All-Purpose Bike Stand	Aug	
9	All-Purpose Bike Stand	Sep	
10	All-Purpose Bike Stand	Oct	

	Product	Attribute	1.2 Value
1	All-Purpose Bike Stand	Jan	2862
2	All-Purpose Bike Stand	Feb	2703
3	All-Purpose Bike Stand	Mar	4929
4	All-Purpose Bike Stand	Apr	3657
5	All-Purpose Bike Stand	May	2226
6	All-Purpose Bike Stand	Jun	2544
7	All-Purpose Bike Stand	Jul	3021
8	All-Purpose Bike Stand	Aug	3180
9	All-Purpose Bike Stand	Sep	4134
10	All-Purpose Bike Stand	Oct	3816
11	All-Purpose Bike Stand	Nov	3816
12	All-Purpose Bike Stand	Dec	2703
13	AWC Logo Cap	Jan	4158.05502
14	AWC Logo Cap	Feb	2621.376108
15	AWC Logo Cap	Mar	6097.65088
16	AWC Logo Cap	Apr	3252.452888
17	AWC Logo Cap	May	5879.048578
18	AWC Logo Cap	Jun	5178.456421
19	AWC Logo Cap	Jul	4982.108505
20	AWC Logo Cap	Aug	3959.779185
21	AWC Logo Cap	Sep	4037.980552
22	AWC Logo Cap	Oct	5080.897994
23	AWC Logo Cap	Nov	2870.137492
24	AWC Logo Cap	Dec	3111.502
25	Bike Wash - Dissolver	Jan	1110.23658

Power BI/Power Query Append to New Query Statement (Used with permission from Microsoft.)

The screenshot shows an intermediate append in Power Query where we are combining the data sets from 2018 through 2020 into a single new data set for us to use.

Review Activity:

Transpose and Append Data

Answer the following questions:

1. **Which action reverses the direction of the data?**

2. **What does it mean to append data?**

3. **What command can you use to reverse the direction of the data?**

Topic 7C

Query Data

EXAM OBJECTIVES COVERED

2.3 Given a scenario, execute data manipulation techniques.

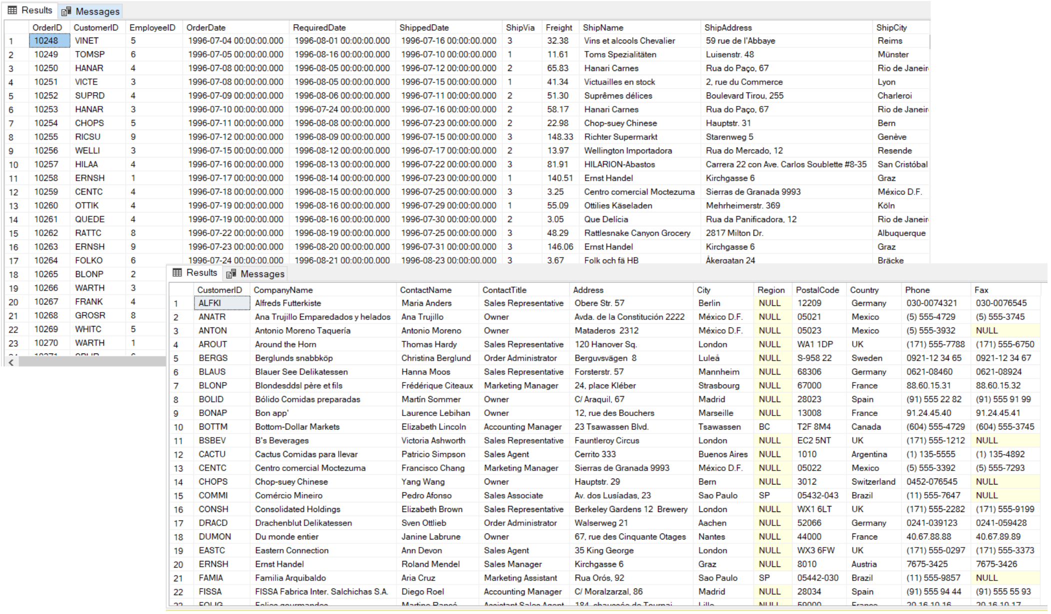
One of our key goals as data analysts is to merge multiple data sets into a single data set. The way we do this is by querying our data (also referred to as merging or blending data). At the end of the day, if you understand the basic concepts behind writing a query, then you can master any tool that is used to arrange that data into a single data set.

Know that most data programs have some form of querying capability. Conceptually, they all work the same, although each program may refer to the process differently (merge, blend, or query).

Querying Data

You may find that the first step you perform as an analyst is **querying** the data. You will create your data set and then you will begin the data transformations, like combining names or creating functions. You will use **joins** when working with any two or more data sets that need to be combined into a single data set for reporting.

Let's start with a list of invoices that we have in our database. This invoice table contains very valuable invoice details, but it's missing customer details. It only contains the CustomerID, which is a foreign key (meaning CustomerID is the primary key in the customer table). For the data set to be more meaningful, it will need to include the customer's name and their geography details, so that we cannot only report on these elements but also map that information in later visualizations. Because the information exists in two different tables, we will use a query to join them.



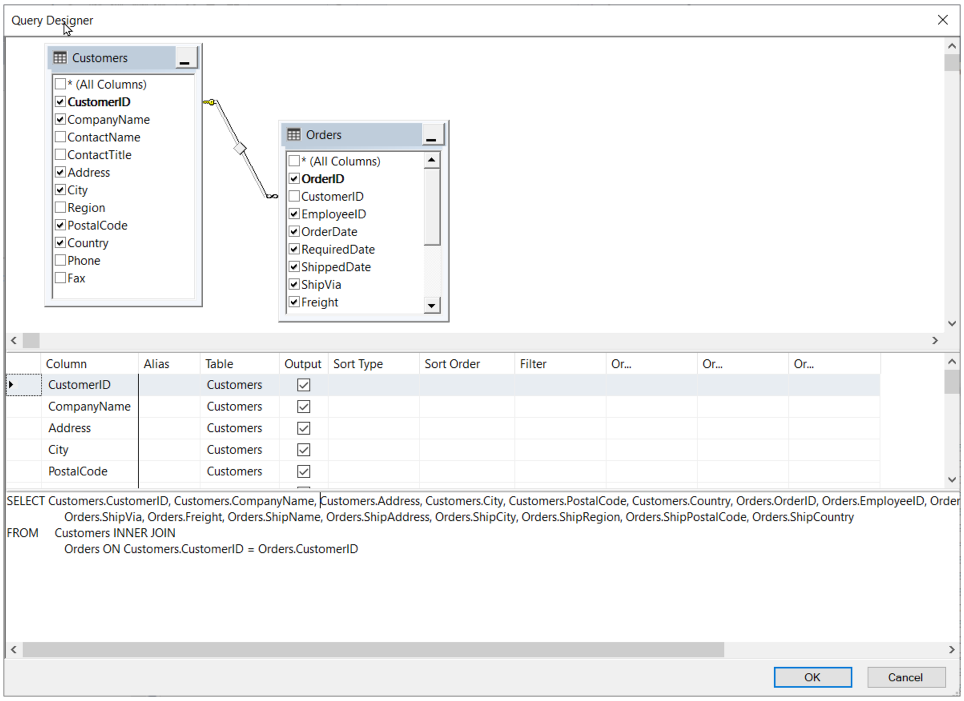
	OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate	ShippedDate	ShipVia	Freight	ShipName	ShipAddress	ShipCity
1	10248	VINET	5	1996-07-04 00:00:00.000	1996-08-01 00:00:00.000	1996-07-16 00:00:00.000	3	32.38	Vins et alcools Chevalier	59 rue de l'Abbaye	Reims
2	10249	TOMSP	6	1996-07-05 00:00:00.000	1996-08-16 00:00:00.000	1996-07-10 00:00:00.000	1	11.61	Toms Spezialitäten	Luisenstr. 48	Münster
3	10250	HANAR	4	1996-07-08 00:00:00.000	1996-08-05 00:00:00.000	1996-07-12 00:00:00.000	2	65.83	Hanari Carnes	Rua do Paço, 67	Rio de Janeir
4	10251	VICTE	3	1996-07-08 00:00:00.000	1996-08-05 00:00:00.000	1996-07-15 00:00:00.000	1	41.34	Victuailles en stock	2, rue du Commerce	Lyon
5	10252	SUPRD	4	1996-07-09 00:00:00.000	1996-08-06 00:00:00.000	1996-07-11 00:00:00.000	2	51.30	Suprêmes délices	Boulevard Tirou, 255	Charleroi
6	10253	HANAR	3	1996-07-10 00:00:00.000	1996-07-24 00:00:00.000	1996-07-16 00:00:00.000	2	58.17	Hanari Carnes	Rua do Paço, 67	Rio de Janeir
7	10254	CHOPS	5	1996-07-11 00:00:00.000	1996-08-08 00:00:00.000	1996-07-23 00:00:00.000	2	22.98	Chop-suey Chinese	Hauptstr. 31	Bern
8	10255	RICSU	9	1996-07-12 00:00:00.000	1996-08-09 00:00:00.000	1996-07-15 00:00:00.000	3	148.33	Richter Supermarkt	Starenweg 5	Genève
9	10256	WELLI	3	1996-07-15 00:00:00.000	1996-08-12 00:00:00.000	1996-07-17 00:00:00.000	2	13.97	Wellington Importadora	Rua do Mercado, 12	Resende
10	10257	HILAA	4	1996-07-16 00:00:00.000	1996-08-13 00:00:00.000	1996-07-22 00:00:00.000	3	81.91	HILARION-Abastos	Carrera 22 con Ave. Carlos Soublette #8-35	San Cristóbal
11	10258	ERNSH	1	1996-07-17 00:00:00.000	1996-08-14 00:00:00.000	1996-07-23 00:00:00.000	1	140.51	Ernst Handel	Kirchgasse 6	Graz
12	10259	CENTC	4	1996-07-18 00:00:00.000	1996-08-15 00:00:00.000	1996-07-25 00:00:00.000	3	3.25	Centro comercial Moctezuma	Sierras de Granada 9993	México D.F.
13	10260	OTTIK	4	1996-07-19 00:00:00.000	1996-08-16 00:00:00.000	1996-07-29 00:00:00.000	1	55.09	Ottilies Käseladen	Mehrheimerstr. 369	Köln
14	10261	QUEDE	4	1996-07-19 00:00:00.000	1996-08-16 00:00:00.000	1996-07-30 00:00:00.000	2	3.05	Que Delícia	Rua da Panificadora, 12	Rio de Janeir
15	10262	RATTC	8	1996-07-22 00:00:00.000	1996-08-19 00:00:00.000	1996-07-25 00:00:00.000	3	48.29	Rattlesnake Canyon Grocery	2817 Milton Dr.	Albuquerque
16	10263	ERNSH	9	1996-07-23 00:00:00.000	1996-08-20 00:00:00.000	1996-07-31 00:00:00.000	3	146.06	Ernst Handel	Kirchgasse 6	Graz
17	10264	FOLKO	6	1996-07-24 00:00:00.000	1996-08-21 00:00:00.000	1996-08-23 00:00:00.000	3	3.67	Folk och fä HB	Åkergatan 24	Bräcke
18	10265	BLONP	2								
19	10266	WARTH	3								
20	10267	FRANK	4								
21	10268	GROSR	8								
22	10269	WHITC	5								
23	10270	WARTH	1								

	CustomerID	CompanyName	ContactName	ContactTitle	Address	City	Region	PostalCode	Country	Phone	Fax
1	ALFKI	Alfreds Futterkiste	Maria Anders	Sales Representative	Obere Str. 57	Berlin	NULL	12209	Germany	030-0074321	030-0076545
2	ANATR	Ana Trujillo Emparedados y helados	Ana Trujillo	Owner	Avda. de la Constitución 2222	México D.F.	NULL	05021	Mexico	(5) 555-4729	(5) 555-3745
3	ANTON	Antonio Moreno Taquería	Antonio Moreno	Owner	Mataderos 2312	México D.F.	NULL	05023	Mexico	(5) 555-3932	NULL
4	AROUT	Around the Horn	Thomas Hardy	Sales Representative	120 Hanover Sq.	London	NULL	WA1 1DP	UK	(171) 555-7788	(171) 555-6750
5	BERGS	Berglunds snabbköp	Christina Berglund	Order Administrator	Berguvsvägen 8	Luleå	NULL	S-958 22	Sweden	0921-12 34 65	0921-12 34 67
6	BLAUS	Blauer See Delikatessen	Hanna Moos	Sales Representative	Forsterstr. 57	Mannheim	NULL	68306	Germany	0621-08460	0621-08924
7	BLONP	Blondesddsl père et fils	Frédérique Citeaux	Marketing Manager	24, place Kléber	Strasbourg	NULL	67000	France	88.60.15.31	88.60.15.32
8	BOLID	Bólido Comidas preparadas	Martín Sommer	Owner	C/ Araquil, 67	Madrid	NULL	28023	Spain	(91) 555 22 82	(91) 555 91 99
9	BONAP	Bon app'	Laurence Lebihan	Owner	12, rue des Bouchers	Marseille	NULL	13008	France	91.24.45.40	91.24.45.41
10	BOTTM	Bottom-Dollar Markets	Elizabeth Lincoln	Accounting Manager	23 Tsawassen Blvd.	Tsawassen	BC	T2F 8M4	Canada	(604) 555-4729	(604) 555-3745
11	BSBEV	B's Beverages	Victoria Ashworth	Sales Representative	Fauntleroy Circus	London	NULL	EC2 5NT	UK	(171) 555-1212	NULL
12	CACTU	Cactus Comidas para llevar	Patricio Simpson	Sales Agent	Cerrito 333	Buenos Aires	NULL	1010	Argentina	(1) 135-5555	(1) 135-4892
13	CENTC	Centro comercial Moctezuma	Francisco Chang	Marketing Manager	Sierras de Granada 9993	México D.F.	NULL	05022	Mexico	(5) 555-3392	(5) 555-7293
14	CHOPS	Chop-suey Chinese	Yang Wang	Owner	Hauptstr. 29	Bern	NULL	3012	Switzerland	0452-076545	NULL
15	COMMI	Comércio Mineiro	Pedro Afonso	Sales Associate	Av. dos Lusíadas, 23	Sao Paulo	SP	05432-043	Brazil	(11) 555-7647	NULL
16	CONSH	Consolidated Holdings	Elizabeth Brown	Sales Representative	Berkeley Gardens 12 Brewery	London	NULL	WX1 6LT	UK	(171) 555-2282	(171) 555-9199
17	DRACD	Drachenblut Delikatessen	Sven Ottlieb	Order Administrator	Walserweg 21	Aachen	NULL	52066	Germany	0241-039123	0241-059428
18	DUMON	Du monde entier	Janine Labrune	Owner	67, rue des Cinquante Otages	Nantes	NULL	44000	France	40.67.88.88	40.67.89.89
19	EASTC	Eastern Connection	Ann Devon	Sales Agent	35 King George	London	NULL	WX3 6FW	UK	(171) 555-0297	(171) 555-3373
20	ERNSH	Ernst Handel	Roland Mendel	Sales Manager	Kirchgasse 6	Graz	NULL	8010	Austria	7675-3425	7675-3426
21	FAMIA	Familia Arquibaldo	Aria Cruz	Marketing Assistant	Rua Orós, 92	Sao Paulo	SP	05442-030	Brazil	(11) 555-9857	NULL
22	FISSA	FISSA Fabrica Inter. Salchichas S.A.	Diego Roel	Accounting Manager	C/ Moralzarzal, 86	Madrid	NULL	28034	Spain	(91) 555 94 44	(91) 555 55 93

Orders Table and Customer Table in Northwind Database (Used with permission from Microsoft.)

Consider these two data sets. Each table contains only the data it is designed to store for efficiency's sake. To combine this data into a single, more meaningful set, it must include the necessary key fields from each table. In our example, we can use the CustomerID field in each of the data sets to join the customers to the order information.

We bring this data into a query either through an SQL statement or a query designer, such as what's used in tools like SQL Server Management Studio, Microsoft Access, or Tableau. You will leverage the key fields, which in this example is the CustomerID, to bring all the records together. Below you will see a screenshot of one such query designer in the process of merging the two data sets.



Customers

- □ * (All Columns)
- ☑ **CustomerID**
- ☑ CompanyName
- □ ContactName
- □ ContactTitle
- ☑ Address
- ☑ City
- □ Region
- ☑ PostalCode
- ☑ Country
- □ Phone
- □ Fax

Orders

- □ * (All Columns)
- ☑ **OrderID**
- □ CustomerID
- ☑ EmployeeID
- ☑ OrderDate
- ☑ RequiredDate
- ☑ ShippedDate
- ☑ ShipVia
- ☑ Freight

Column	Alias	Table	Output	Sort Type	Sort Order	Filter	Or...	Or...	Or...
CustomerID		Customers	☑						
CompanyName		Customers	☑						
Address		Customers	☑						
City		Customers	☑						
PostalCode		Customers	☑						

```
SELECT Customers.CustomerID, Customers.CompanyName, Customers.Address, Customers.City, Customers.PostalCode, Customers.Country, Orders.OrderID, Orders.EmployeeID, Order
          Orders.ShipVia, Orders.Freight, Orders.ShipName, Orders.ShipAddress, Orders.ShipCity, Orders.ShipRegion, Orders.ShipPostalCode, Orders.ShipCountry
FROM    Customers INNER JOIN
          Orders ON Customers.CustomerID = Orders.CustomerID
```

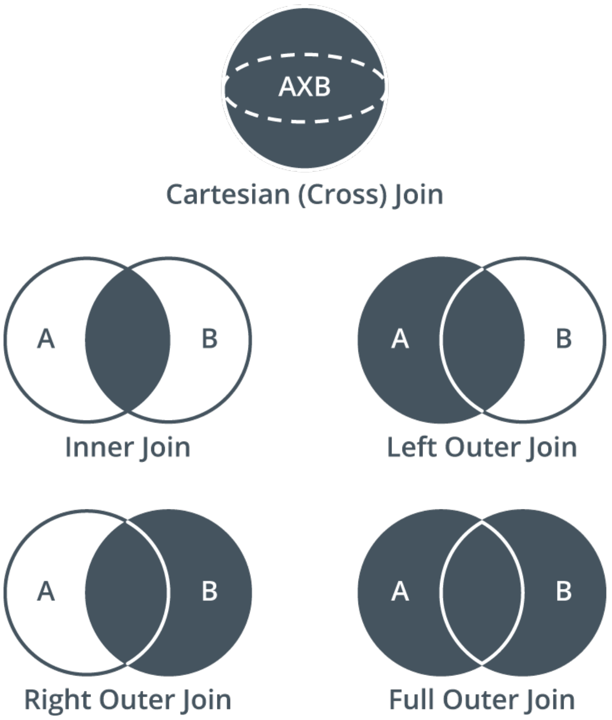
View of Basic Query in SQL Server Management Studio (Used with permission from Microsoft.)

When a relationship in the database exists between the tables you are using in your query, it will automatically show the join line between the two tables on the key field. When a relationship doesn't exist, the analyst will need to create the join line by dragging and dropping the key field from one table to the key field in the other table. In the case of our screenshot, we would drag and drop CustomerID from the Customers table to CustomerID in the Orders table. The initial join type will be an inner join. This join type defaults to showing records that exist in both tables, which means it will only show customers that have made orders (and will not show customers that have not made an order).

Types of Joins

There are multiple join types that you can use when querying data, and it can be hard to understand which join would be best in any given situation. The decision about which join type to use should be based on what results you need from your data set.

- **Cross join/Cartesian join**: Data doesn't have a direct join on a key field.
- **Inner join**: Data is joined so that only records that exist in both tables appear in the result.
- **Left outer join**: The left table displays all results of the left table, while only matching records in the other (right) table appear in the result.
- **Right outer join**: The right table displays all results of the right table, while only matching records in the other (left) table appear in the result.
- **Full outer join**: Data is joined so that all records, matched or unmatched, show in the results.

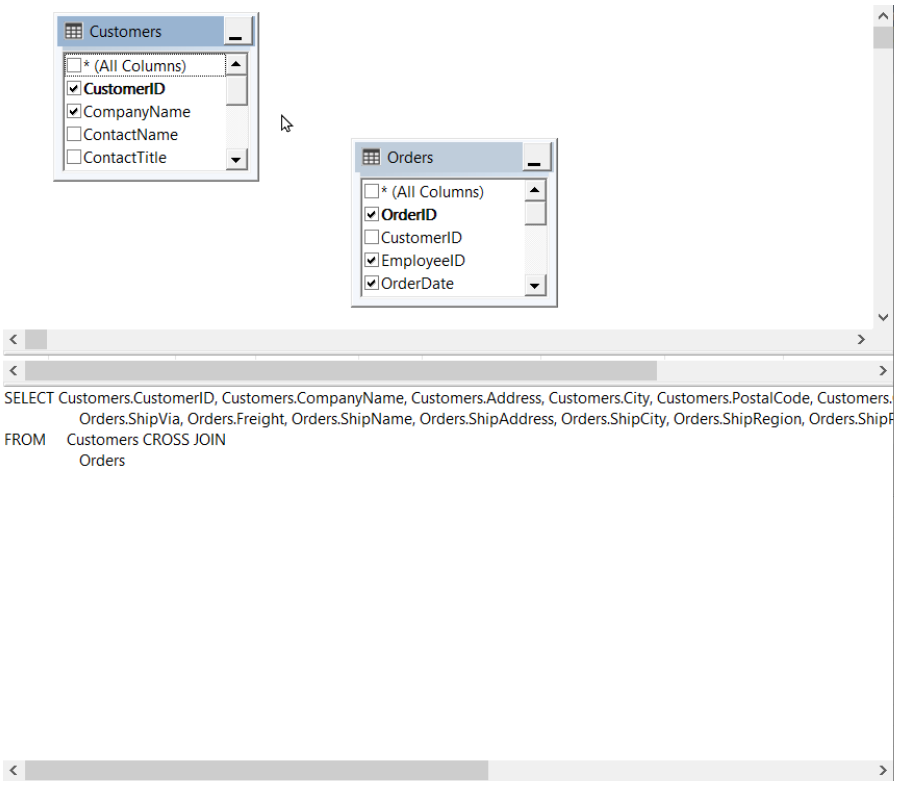


Join Type Diagram

The Impact of Each Join to Data

Because joins ultimately control what data shows in the results, it is important to understand the impact of each type. Let's return to our example, in which we want to query data from the customers data set and the invoices data set, and break down each outcome using different join types.

Cross Join


Customers

□ * (All Columns)
☑ **CustomerID**
☑ CompanyName
□ ContactName
□ ContactTitle

Orders

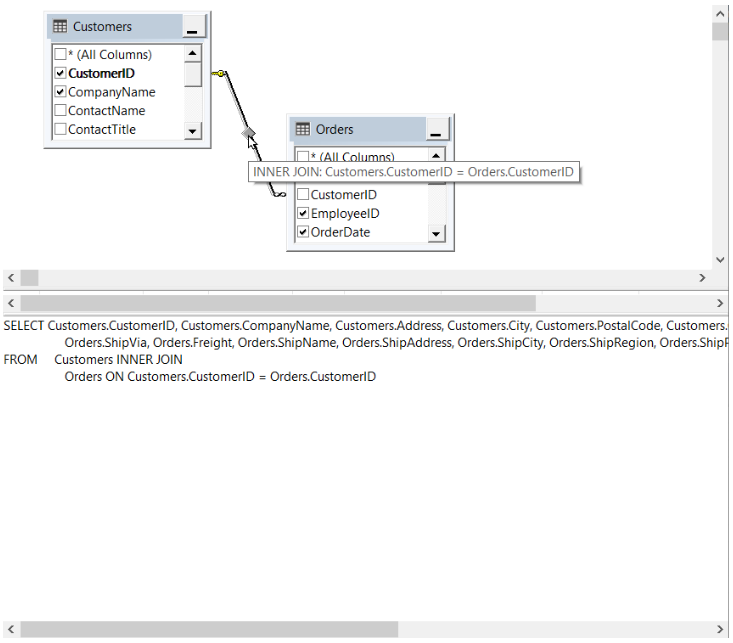
□ * (All Columns)
☑ **OrderID**
□ CustomerID
☑ EmployeeID
☑ OrderDate

```
SELECT Customers.CustomerID, Customers.CompanyName, Customers.Address, Customers.City, Customers.PostalCode, Customers.
          Orders.ShipVia, Orders.Freight, Orders.ShipName, Orders.ShipAddress, Orders.ShipCity, Orders.ShipRegion, Orders.ShipF
FROM    Customers CROSS JOIN
          Orders
```

Cross Join on Customers and Orders in Microsoft SQL Management Studio (Used with permission from Microsoft.)

When tables are added to a query and no join is specified, the end result will be that each record in each table is tied together. In the case of our query above, this would associate each customer to every order in the orders table, meaning every order that has been placed will be associated to every customer regardless of whether its their order. When it does not know how to join the records through the key, it will default to joining every record to every record. Because of this, cross joins are rarely used.

Inner Join

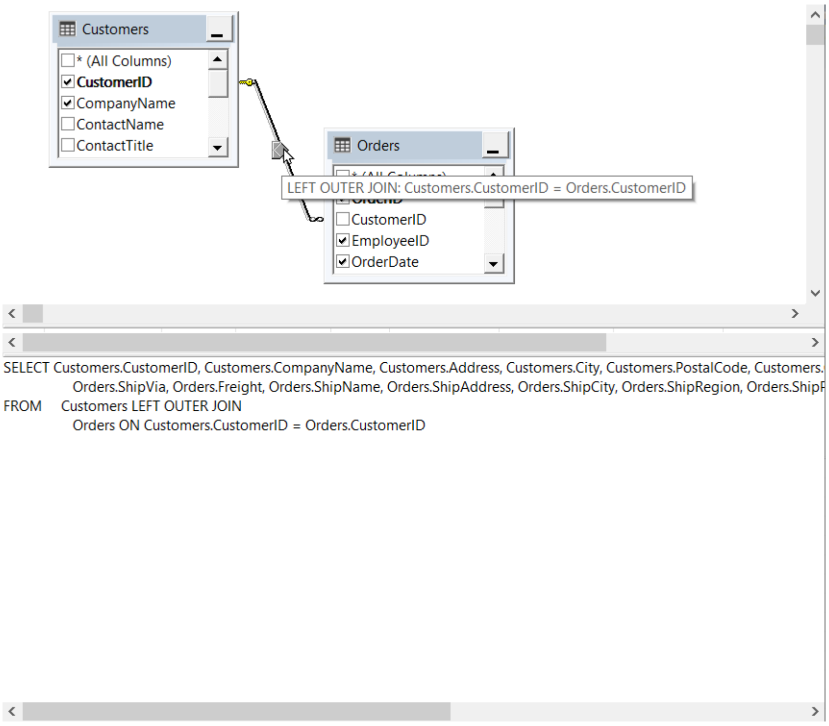


```
SELECT Customers.CustomerID, Customers.CompanyName, Customers.Address, Customers.City, Customers.PostalCode, Customers.
         Orders.ShipVia, Orders.Freight, Orders.ShipName, Orders.ShipAddress, Orders.ShipCity, Orders.ShipRegion, Orders.ShipF
FROM     Customers INNER JOIN
         Orders ON Customers.CustomerID = Orders.CustomerID
```

Inner Join on Customers and Orders in Microsoft SQL Management Studio (Used with permission from Microsoft.)

Inner joins are typically the default join type for most programs that query. The result of an inner join for this data set is that for every customer who has placed an order, there will be a record in the results. This means that by using an inner join, you will automatically exclude any customer who has not placed an order.

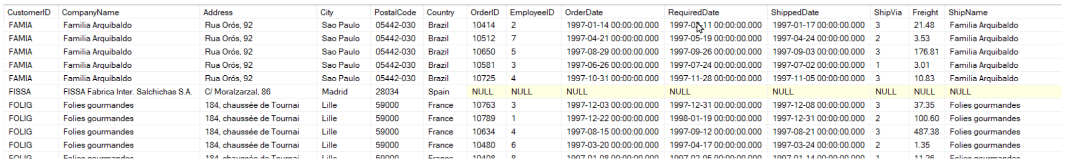
Left Outer Join



```
SELECT Customers.CustomerID, Customers.CompanyName, Customers.Address, Customers.City, Customers.PostalCode, Customers.
         Orders.ShipVia, Orders.Freight, Orders.ShipName, Orders.ShipAddress, Orders.ShipCity, Orders.ShipRegion, Orders.ShipF
FROM     Customers LEFT OUTER JOIN
         Orders ON Customers.CustomerID = Orders.CustomerID
```

Left Outer Join on Customers and Orders in Microsoft SQL Management Studio (Used with permission from Microsoft.)

Left outer joins specify that all records from the table on the left side and only matching records from the right table show in the results. In this case, the left table is the Customer table, which means all customers will show in the results whether they have no orders, one order, or multiple orders.

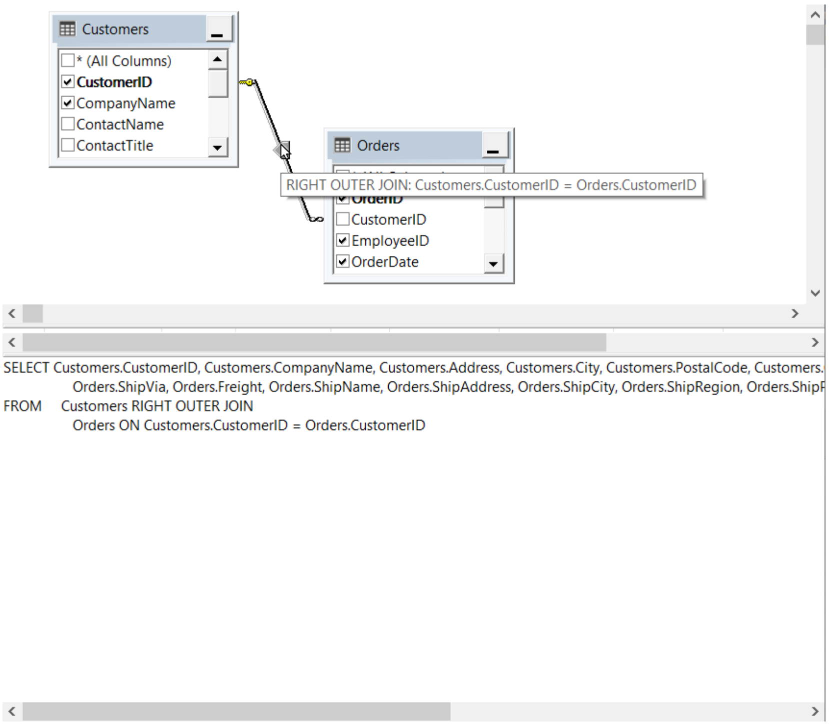


CustomerID	CompanyName	Address	City	PostalCode	Country	OrderID	EmployeeID	OrderDate	RequiredDate	ShippedDate	ShipVia	Freight	ShipName
FAMIA	Familia Arquibaldo	Rua Orós, 92	Sao Paulo	05442-030	Brazil	10414	2	1997-01-14 00:00:00.000	1997-0[illegible]-11 00:00:00.000	1997-01-17 00:00:00.000	3	21.48	Familia Arquibaldo
FAMIA	Familia Arquibaldo	Rua Orós, 92	Sao Paulo	05442-030	Brazil	10512	7	1997-04-21 00:00:00.000	1997-05-19 00:00:00.000	1997-04-24 00:00:00.000	2	3.53	Familia Arquibaldo
FAMIA	Familia Arquibaldo	Rua Orós, 92	Sao Paulo	05442-030	Brazil	10650	5	1997-08-29 00:00:00.000	1997-09-26 00:00:00.000	1997-09-03 00:00:00.000	3	176.81	Familia Arquibaldo
FAMIA	Familia Arquibaldo	Rua Orós, 92	Sao Paulo	05442-030	Brazil	10581	3	1997-06-26 00:00:00.000	1997-07-24 00:00:00.000	1997-07-02 00:00:00.000	1	3.01	Familia Arquibaldo
FAMIA	Familia Arquibaldo	Rua Orós, 92	Sao Paulo	05442-030	Brazil	10725	4	1997-10-31 00:00:00.000	1997-11-28 00:00:00.000	1997-11-05 00:00:00.000	3	10.83	Familia Arquibaldo
FISSA	FISSA Fabrica Inter. Salchichas S.A.	C/ Moralzarzal, 86	Madrid	28034	Spain	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
FOLIG	Folies gourmandes	184, chaussée de Tournai	Lille	59000	France	10763	3	1997-12-03 00:00:00.000	1997-12-31 00:00:00.000	1997-12-08 00:00:00.000	3	37.35	Folies gourmandes
FOLIG	Folies gourmandes	184, chaussée de Tournai	Lille	59000	France	10789	1	1997-12-22 00:00:00.000	1998-01-19 00:00:00.000	1997-12-31 00:00:00.000	2	100.60	Folies gourmandes
FOLIG	Folies gourmandes	184, chaussée de Tournai	Lille	59000	France	10634	4	1997-08-15 00:00:00.000	1997-09-12 00:00:00.000	1997-08-21 00:00:00.000	3	487.38	Folies gourmandes
FOLIG	Folies gourmandes	184, chaussée de Tournai	Lille	59000	France	10480	6	1997-03-20 00:00:00.000	1997-04-17 00:00:00.000	1997-03-24 00:00:00.000	2	1.35	Folies gourmandes
FOLIG	Folies gourmandes	184, chaussée de Tournai	Lille	59000	France	10408	8	1997-01-08 00:00:00.000	1997-02-05 00:00:00.000	1997-01-14 00:00:00.000	1	11.26	Folies gourmandes

Left Outer for Customers and Orders Results (Used with permission from Microsoft.)

When a customer who has no order shows up in the results, it will produce a null for all the order fields (because that order doesn't exist).

Right Outer Join

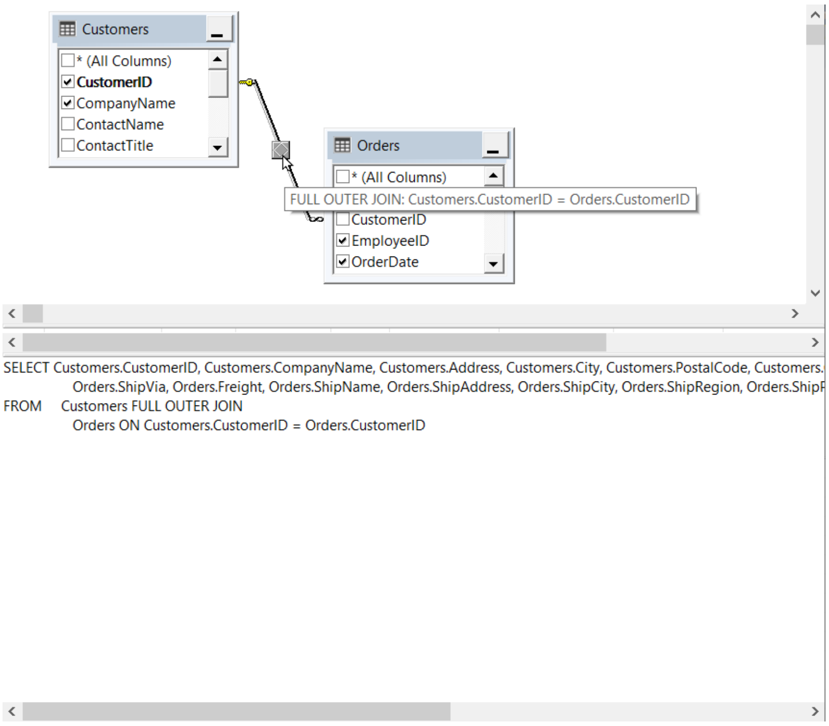


```
SELECT Customers.CustomerID, Customers.CompanyName, Customers.Address, Customers.City, Customers.PostalCode, Customers.
            Orders.ShipVia, Orders.Freight, Orders.ShipName, Orders.ShipAddress, Orders.ShipCity, Orders.ShipRegion, Orders.ShipP
FROM     Customers RIGHT OUTER JOIN
            Orders ON Customers.CustomerID = Orders.CustomerID
```

Right Outer Join on Customers and Orders in Microsoft SQL Management Studio (Used with permission from Microsoft.)

Right outer joins function the same way as left outer joins, except in this type all records from the right table and only matching records in the left are shown in the results. This join type will also show Nulls when a match doesn't exist. With this join type, the results would make clear any instances where we have orders that did not have a customer record.

Full Outer Join



```
SELECT Customers.CustomerID, Customers.CompanyName, Customers.Address, Customers.City, Customers.PostalCode, Customers.
            Orders.ShipVia, Orders.Freight, Orders.ShipName, Orders.ShipAddress, Orders.ShipCity, Orders.ShipRegion, Orders.ShipF
FROM    Customers FULL OUTER JOIN
            Orders ON Customers.CustomerID = Orders.CustomerID
```

Full Outer Join on Customers and Orders in Microsoft SQL Management Studio (Used with permission from Microsoft.)

A full outer join is like creating a left outer and right outer at the same time, meaning it displays the same results you would see if you ran both a left and right join query, but in one single set. This is a great way to troubleshoot potential bad records, or records due to be corrected. The result here would show all customers regardless of their orders, and also all orders.

Review Activity:

Query Data

Answer the following questions:

1. **Which action involves merging multiple data sets into a single data set?**

2. **Which join type results in a display that only contains records that exist in both tables?**

3. **Which join type is used the least, and why?**

4. **Which join type provides results that help us troubleshoot potential bad records or records due to be corrected?**

Lesson 7

Summary

After this lesson, you should understand that recoding is used to change the current value of a single data field or variable to a different value to make it more meaningful, group it more effectively for analysis, or correct the data. You've learned that imputing values means to replace data with an estimated value, such as average records or a predicted value. You've also discovered that reduction in data mining means reducing the volume of data. You should have learned about the different types of joins that affect the results shown by queries. You should also be able to recognize the reasons we transpose (reverse) data and append (combine) data.

Guidelines in Manipulating Data Sets and Joins

Consider these best practices and guidelines when working with data sets and joins.

1. When performing reduction of data, use caution in order to avoid invalidating the data.
2. Remember to mask values when working with sensitive data, or personally identifiable data (PII).
3. A data set sometimes needs to be transposed from the format we've received in order to be workable for other data formats.
4. When working with any two or more data sets that need to be combined into a single data set for reporting, querying may happen before the transformation process.
5. The different types of joins do not produce the same data sets. You must choose the type that best suits your needs.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 8

Explaining Common Techniques for Data Manipulation and Optimization

LESSON INTRODUCTION

A common misconception that data professionals have is that they will be working with clean data of the appropriate data type, and that all of the data they need will exist in the data set. However, having data that is clean and ready for analysis rarely happens, especially when dealing with legacy systems (older technology). However, even when using modern software, there will always be an opportunity to clean and manipulate data to better suit your needs.

Lesson Objectives

In this lesson, you will do the following:

- Use functions to manipulate data.
- Use common techniques for query optimization.

Topic 8A

Use Functions to Manipulate Data

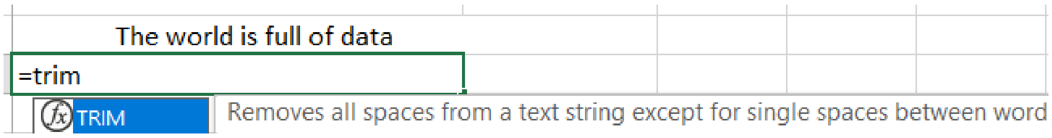
EXAM OBJECTIVES COVERED

2.3 Given a scenario, execute data manipulation techniques.
2.4 Explain common techniques for data manipulation and query optimization.

The most common reason for using functions is the need to create data that doesn't exist in your database, or at least doesn't exist in the way you might require it for reporting. Databases are already large, so storing every single calculation that you would ever use (or maybe never use) does not make sense. The data that is stored can provide you with valuable information that can be used to create data. You can also use functions to manipulate data, such as conversion-related or format-related functions.

Text Functions

Several **text functions** can assist a data analyst in working with text-based fields. The type of text-based function you may need will be based on the data you are working with and the challenges it may present. When working with a software system and its fields, you will learn the nuances of that data. It could be that the data has non-printable characters (e.g., spaces) in the field, which might require you to use functions like TRIM and/or CLEAN. The TRIM function removes spaces that are not legitimate spaces between words from data fields, while CLEAN will remove all non-printable characters.

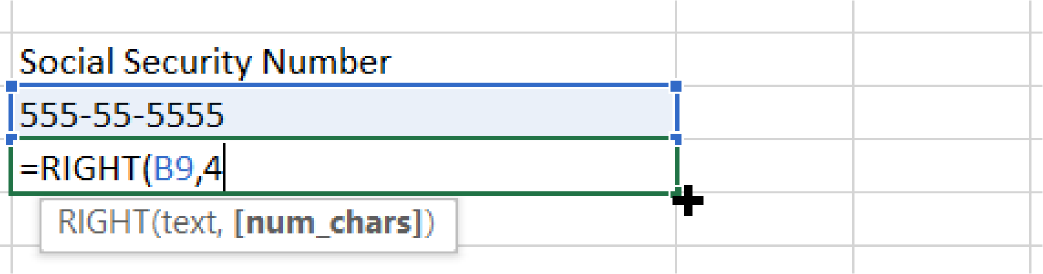

The world is full of data
=trim

Removes all spaces from a text string except for single spaces between word

Microsoft Excel Trim Function to Remove Leading and Trailing Spaces (Used with permission from Microsoft.)

At times, non-printable characters are present because the type of system design has a defined number of characters for a field, whether or not they are used. The outcome means that when the data is exported, it includes non-printable characters and/or spaces. As shown in this screenshot, the TRIM function can be used to remove these spaces.

To isolate characters from a field, you could use functions like LEFT(), RIGHT(), and MID() to create fields that just contain those characters. As an example, suppose that a system stores a person's entire social security number, but you only want to display the last four characters (numbers in this case). To accomplish this, you could build a function that would isolate the four right-most characters of that field (see Excel screenshot).



Social Security Number

555-55-5555

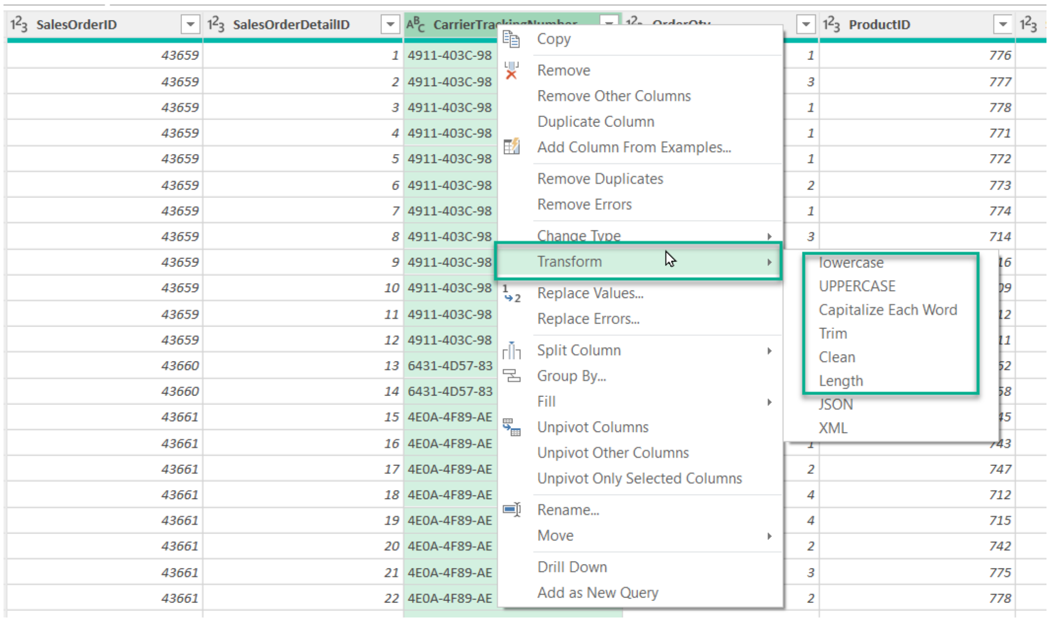
=RIGHT(B9,4

Microsoft Excel Right Function to Extract the Last Four Characters (i.e., Digits) of the Social Security Number (Used with permission from Microsoft.)

Here is a list of common functions in Excel that you may find helpful when working with text. Knowing how these functions work is beneficial for your work not only in Excel, but also because you will see these same types of functions in other programs that do data prep, such as Power Query and Tableau Data Prep.

- **=TRIM()** – Trims leading and trailing spaces.
- **=CLEAN()** – Removes all non-printable characters.
- **=LEFT()** – Removes a number of characters from the left side of the field.
- **=RIGHT()** – Removes a number of characters from the right side of the field.
- **=MID()** – Removes a number of characters from the middle of the field.
- **=UPPER()** – Converts the field to uppercase.
- **=LOWER()** – Converts the field to lowercase.
- **=PROPER()** – Capitalizes each word.
- **=LEN()** – Produces a number that tells you how many countable characters are in that field.

These types of text functions already exist in many data prepping tools, so you will not need to create a function to accomplish this type of task—you can just execute that function for a particular column. For example, in Power Query, you can execute a text function by selecting Transform and choosing from a menu of options, as shown in the following screenshot.



SalesOrderID	SalesOrderDetailID	CarrierTrackingNumber	OrderQty	ProductID
43659	1	4911-403C-98	1	776
43659	2	4911-403C-98	3	777
43659	3	4911-403C-98	1	778
43659	4	4911-403C-98	1	771
43659	5	4911-403C-98	1	772
43659	6	4911-403C-98	2	773
43659	7	4911-403C-98	1	774
43659	8	4911-403C-98	3	714
43659	9	4911-403C-98		16
43659	10	4911-403C-98		09
43659	11	4911-403C-98		12
43659	12	4911-403C-98		11
43660	13	6431-4D57-83		52
43660	14	6431-4D57-83		58
43661	15	4E0A-4F89-AE		45
43661	16	4E0A-4F89-AE	1	743
43661	17	4E0A-4F89-AE	2	747
43661	18	4E0A-4F89-AE	4	712
43661	19	4E0A-4F89-AE	4	715
43661	20	4E0A-4F89-AE	2	742
43661	21	4E0A-4F89-AE	3	775
43661	22	4E0A-4F89-AE	2	778

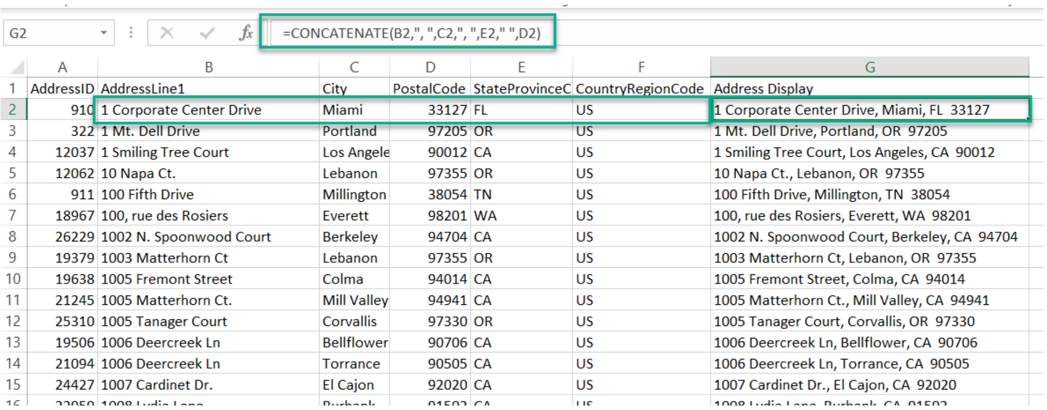
Transform Option in Microsoft Power Query (Used with permission from Microsoft.)

While these sorts of commands used to be completed by manual function building, data prep tools now have basic transformation commands built in.

Combining Data Fields

When working with data, you might need to combine multiple fields into a single field using a **merge fields function**, or CONCATENATE function. As a common example, first and last name are both commonly used in a concatenation formula to create a display name. Another reason for combining fields of information is to form a unique key for that record. For instance, you might want to produce a unique index that includes not only a number, but other information, (e.g., "Customer 1" versus just "1").

Let's walk through how to perform the basic CONCATENATE function in Excel, which is commonly used to combine different address fields to create and display a single consolidated address.



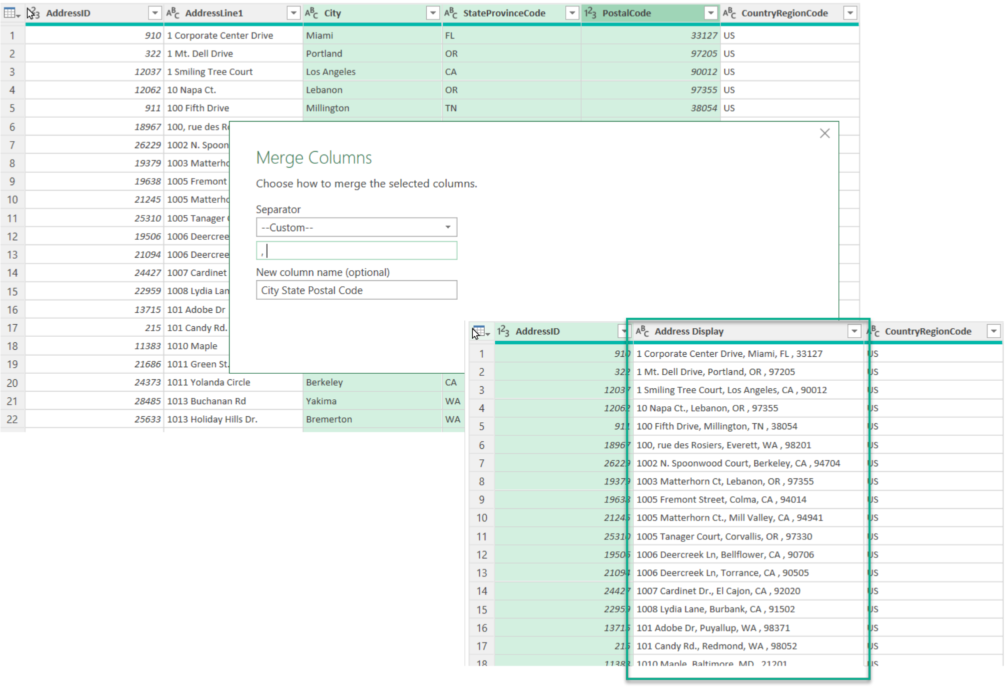
```
=CONCATENATE(B2,", ",C2,", ",E2," ",D2)
```

	A	B	C	D	E	F	G
1	AddressID	AddressLine1	City	PostalCode	StateProvinceC	CountryRegionCode	Address Display
2	910	1 Corporate Center Drive	Miami	33127	FL	US	1 Corporate Center Drive, Miami, FL 33127
3	322	1 Mt. Dell Drive	Portland	97205	OR	US	1 Mt. Dell Drive, Portland, OR 97205
4	12037	1 Smiling Tree Court	Los Angele	90012	CA	US	1 Smiling Tree Court, Los Angeles, CA 90012
5	12062	10 Napa Ct.	Lebanon	97355	OR	US	10 Napa Ct., Lebanon, OR 97355
6	911	100 Fifth Drive	Millington	38054	TN	US	100 Fifth Drive, Millington, TN 38054
7	18967	100, rue des Rosiers	Everett	98201	WA	US	100, rue des Rosiers, Everett, WA 98201
8	26229	1002 N. Spoonwood Court	Berkeley	94704	CA	US	1002 N. Spoonwood Court, Berkeley, CA 94704
9	19379	1003 Matterhorn Ct	Lebanon	97355	OR	US	1003 Matterhorn Ct, Lebanon, OR 97355
10	19638	1005 Fremont Street	Colma	94014	CA	US	1005 Fremont Street, Colma, CA 94014
11	21245	1005 Matterhorn Ct.	Mill Valley	94941	CA	US	1005 Matterhorn Ct., Mill Valley, CA 94941
12	25310	1005 Tanager Court	Corvallis	97330	OR	US	1005 Tanager Court, Corvallis, OR 97330
13	19506	1006 Deercreek Ln	Bellflower	90706	CA	US	1006 Deercreek Ln, Bellflower, CA 90706
14	21094	1006 Deercreek Ln	Torrance	90505	CA	US	1006 Deercreek Ln, Torrance, CA 90505
15	24427	1007 Cardinet Dr.	El Cajon	92020	CA	US	1007 Cardinet Dr., El Cajon, CA 92020
16	22050	1008 Lydia Lane	Burbank	91502	CA	US	1008 Lydia Lane, Burbank, CA 91502

CONCATENATE Function in Microsoft Excel (Used with permission from Microsoft.)

In this function, the syntax allows us to add each field and necessary character to create a proper display of the newly created address. Now our address is in one field that is ideal for display purposes on reports.

When you use data prep tools, this type of data transformation may be referred to as Merge Fields (or something similar).



	AddressID	AddressLine1	City	StateProvinceCode	PostalCode	CountryRegionCode
1	910	1 Corporate Center Drive	Miami	FL	33127	US
2	322	1 Mt. Dell Drive	Portland	OR	97205	US
3	12037	1 Smiling Tree Court	Los Angeles	CA	90012	US
4	12062	10 Napa Ct.	Lebanon	OR	97355	US
5	911	100 Fifth Drive	Millington	TN	38054	US
6	18967	100, rue des R				
7	26229	1002 N. Spoon				
8	19379	1003 Matterhc				
9	19638	1005 Fremont				
10	21245	1005 Matterhc				
11	25310	1005 Tanager (				
12	19506	1006 Deercree				
13	21094	1006 Deercree				
14	24427	1007 Cardinet				
15	22959	1008 Lydia Lan				
16	13715	101 Adobe Dr				
17	215	101 Candy Rd.				
18	11383	1010 Maple				
19	21686	1011 Green St.				
20	24373	1011 Yolanda Circle	Berkeley	CA		
21	28485	1013 Buchanan Rd	Yakima	WA		
22	25633	1013 Holiday Hills Dr.	Bremerton	WA		

Merge Columns

Choose how to merge the selected columns.

Separator

--Custom--

,

New column name (optional)

City State Postal Code

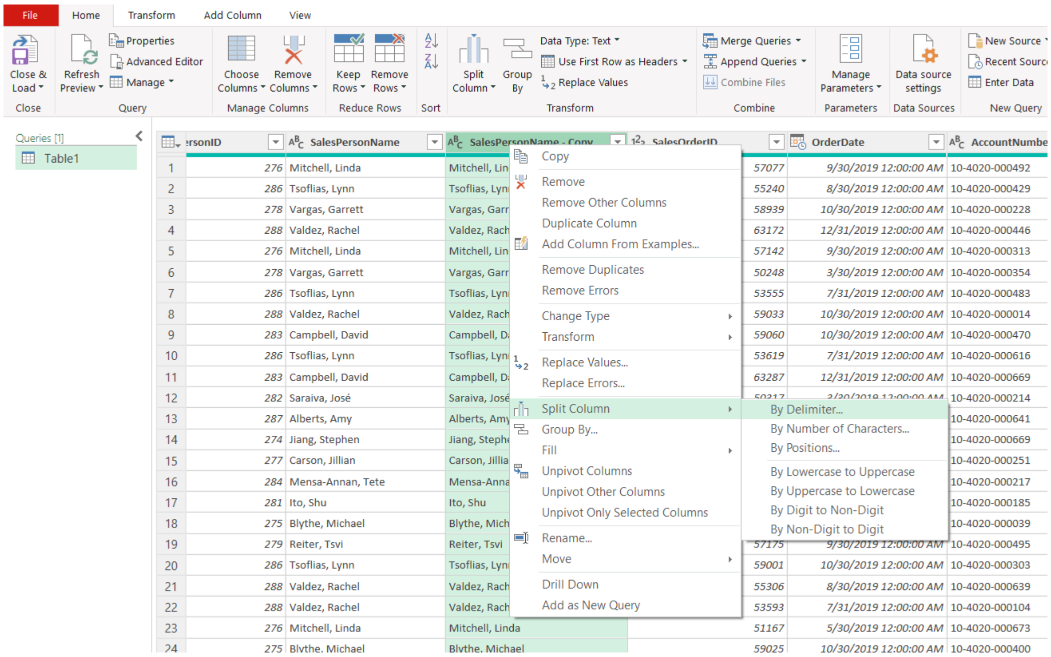
	AddressID	Address Display	CountryRegionCode
1	910	1 Corporate Center Drive, Miami, FL , 33127	US
2	322	1 Mt. Dell Drive, Portland, OR , 97205	US
3	12037	1 Smiling Tree Court, Los Angeles, CA , 90012	US
4	12062	10 Napa Ct., Lebanon, OR , 97355	US
5	911	100 Fifth Drive, Millington, TN , 38054	US
6	18967	100, rue des Rosiers, Everett, WA , 98201	US
7	26229	1002 N. Spoonwood Court, Berkeley, CA , 94704	US
8	19379	1003 Matterhorn Ct, Lebanon, OR , 97355	US
9	19638	1005 Fremont Street, Colma, CA , 94014	US
10	21245	1005 Matterhorn Ct., Mill Valley, CA , 94941	US
11	25310	1005 Tanager Court, Corvallis, OR , 97330	US
12	19506	1006 Deercreek Ln, Bellflower, CA , 90706	US
13	21094	1006 Deercreek Ln, Torrance, CA , 90505	US
14	24427	1007 Cardinet Dr., El Cajon, CA , 92020	US
15	22959	1008 Lydia Lane, Burbank, CA , 91502	US
16	13715	101 Adobe Dr, Puyallup, WA , 98371	US
17	215	101 Candy Rd., Redmond, WA , 98052	US
18	11383	1010 Maple, Baltimore, MD , 21201	US

Merge Columns in Excel Power Query (Used with permission from Microsoft.)

In Power Query, the Merge Columns command will produce the same end result as CONCATENATE.

Parsing Strings for Information

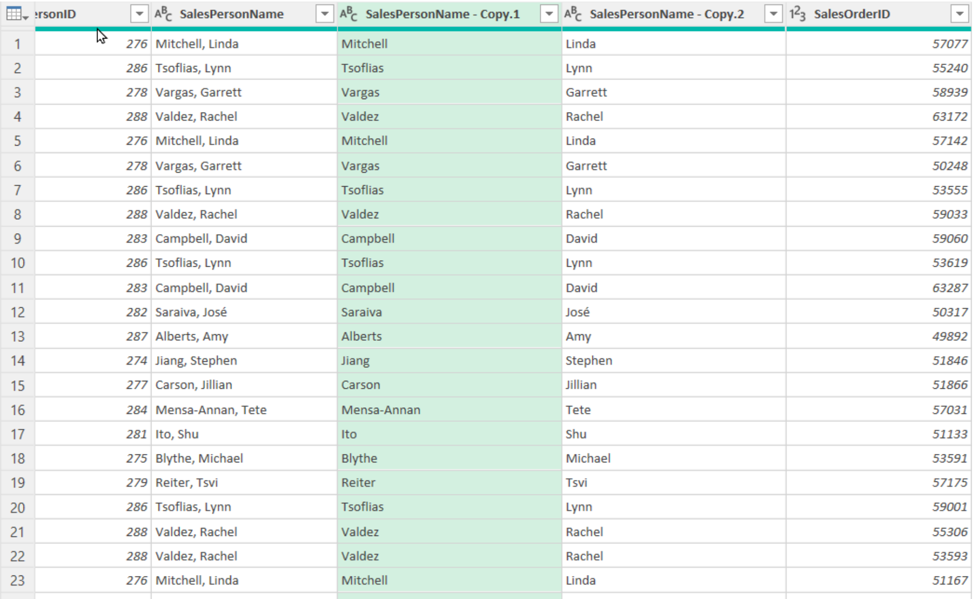
Parsing means to break data into parts. When we parse strings, we are extracting data out of a field for use. For example, if you want to sort data by last name and first name, but these names are combined in the same field, you will need to break the data into individual fields.



	rsonID	SalesPersonName	SalesPersonName - Copy	SalesOrderID	OrderDate	AccountNumbe
1	276	Mitchell, Linda	Mitchell, Lin	57077	9/30/2019 12:00:00 AM	10-4020-000492
2	286	Tsoflias, Lynn	Tsoflias, Lyn	55240	8/30/2019 12:00:00 AM	10-4020-000429
3	278	Vargas, Garrett	Vargas, Garr	58939	10/30/2019 12:00:00 AM	10-4020-000228
4	288	Valdez, Rachel	Valdez, Rach	63172	12/31/2019 12:00:00 AM	10-4020-000446
5	276	Mitchell, Linda	Mitchell, Lin	57142	9/30/2019 12:00:00 AM	10-4020-000313
6	278	Vargas, Garrett	Vargas, Garr	50248	3/30/2019 12:00:00 AM	10-4020-000354
7	286	Tsoflias, Lynn	Tsoflias, Lyn	53555	7/31/2019 12:00:00 AM	10-4020-000483
8	288	Valdez, Rachel	Valdez, Rach	59033	10/30/2019 12:00:00 AM	10-4020-000014
9	283	Campbell, David	Campbell, D	59060	10/30/2019 12:00:00 AM	10-4020-000470
10	286	Tsoflias, Lynn	Tsoflias, Lyn	53619	7/31/2019 12:00:00 AM	10-4020-000616
11	283	Campbell, David	Campbell, D	63287	12/31/2019 12:00:00 AM	10-4020-000669
12	282	Saraiva, José	Saraiva, José	50317	3/30/2019 12:00:00 AM	10-4020-000214
13	287	Alberts, Amy	Alberts, Amy			10-4020-000641
14	274	Jiang, Stephen	Jiang, Stephe			10-4020-000669
15	277	Carson, Jillian	Carson, Jillia			10-4020-000251
16	284	Mensa-Annan, Tete	Mensa-Anna			10-4020-000217
17	281	Ito, Shu	Ito, Shu			10-4020-000185
18	275	Blythe, Michael	Blythe, Mich			10-4020-000039
19	279	Reiter, Tsvi	Reiter, Tsvi	57175	9/30/2019 12:00:00 AM	10-4020-000495
20	286	Tsoflias, Lynn	Tsoflias, Lyn	59001	10/30/2019 12:00:00 AM	10-4020-000303
21	288	Valdez, Rachel	Valdez, Rach	55306	8/30/2019 12:00:00 AM	10-4020-000639
22	288	Valdez, Rachel	Valdez, Rach	53593	7/31/2019 12:00:00 AM	10-4020-000104
23	276	Mitchell, Linda	Mitchell, Linda	51167	5/30/2019 12:00:00 AM	10-4020-000673
24	275	Blythe, Michael	Blythe, Michael	59025	10/30/2019 12:00:00 AM	10-4020-000400

Using Split Columns in Microsoft Power Query to Create Individual Names (Used with permission from Microsoft.)

Using the comma, we can break the SalesPersonName field into the last name and first name. In our screenshot, we have duplicated the SalesPersonName field because we want to keep the original for display purposes. We then navigated to the Split Column command, and chose to split by delimiter (in this case, a comma) to create new fields as seen in the next screenshot.

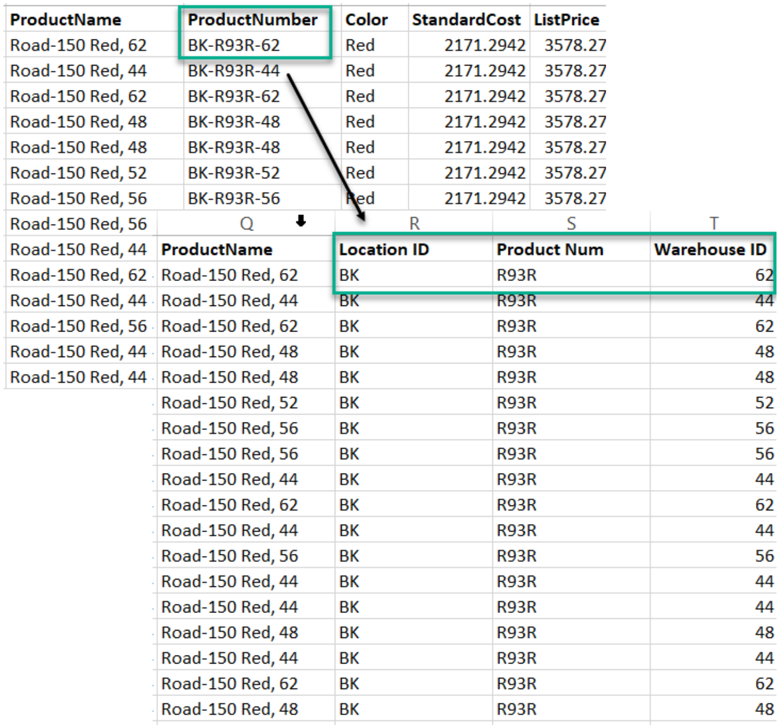


	rsonID	SalesPersonName	SalesPersonName - Copy.1	SalesPersonName - Copy.2	SalesOrderID
1	276	Mitchell, Linda	Mitchell	Linda	57077
2	286	Tsoflias, Lynn	Tsoflias	Lynn	55240
3	278	Vargas, Garrett	Vargas	Garrett	58939
4	288	Valdez, Rachel	Valdez	Rachel	63172
5	276	Mitchell, Linda	Mitchell	Linda	57142
6	278	Vargas, Garrett	Vargas	Garrett	50248
7	286	Tsoflias, Lynn	Tsoflias	Lynn	53555
8	288	Valdez, Rachel	Valdez	Rachel	59033
9	283	Campbell, David	Campbell	David	59060
10	286	Tsoflias, Lynn	Tsoflias	Lynn	53619
11	283	Campbell, David	Campbell	David	63287
12	282	Saraiva, José	Saraiva	José	50317
13	287	Alberts, Amy	Alberts	Amy	49892
14	274	Jiang, Stephen	Jiang	Stephen	51846
15	277	Carson, Jillian	Carson	Jillian	51866
16	284	Mensa-Annan, Tete	Mensa-Annan	Tete	57031
17	281	Ito, Shu	Ito	Shu	51133
18	275	Blythe, Michael	Blythe	Michael	53591
19	279	Reiter, Tsvi	Reiter	Tsvi	57175
20	286	Tsoflias, Lynn	Tsoflias	Lynn	59001
21	288	Valdez, Rachel	Valdez	Rachel	55306
22	288	Valdez, Rachel	Valdez	Rachel	53593
23	276	Mitchell, Linda	Mitchell	Linda	51167

Final Result of Split Column in Microsoft Power Query (Used with permission from Microsoft.)

Now, we have two additional fields: SalesPersonName - Copy.1 and SalesPersonName - Copy.2. We can rename these to be more meaningful (such as “LastName” and “FirstName”) and can use them to sort the data in the column by last name and first name.

Another scenario in which you might parse data strings would be when a company uses a group of fields to form a nomenclature or customer ID that represents different pieces of information. The presence of a delimiter in a field (such as dashes) can signal that the field contains different pieces of information, and that we may need to break it out into separate fields to work with its individual parts. For example, a company could have created a product number which combines the location ID, product number, and warehouse ID into a single field, using a dash to separate each individual piece of the information. We can then use this field, and the dash, as a method to break information out into their respective columns.



ProductName	ProductNumber	Color	StandardCost	ListPrice
Road-150 Red, 62	BK-R93R-62	Red	2171.2942	3578.27
Road-150 Red, 44	BK-R93R-44	Red	2171.2942	3578.27
Road-150 Red, 62	BK-R93R-62	Red	2171.2942	3578.27
Road-150 Red, 48	BK-R93R-48	Red	2171.2942	3578.27
Road-150 Red, 48	BK-R93R-48	Red	2171.2942	3578.27
Road-150 Red, 52	BK-R93R-52	Red	2171.2942	3578.27
Road-150 Red, 56	BK-R93R-56	Red	2171.2942	3578.27
Road-150 Red, 56				
Road-150 Red, 44				
Road-150 Red, 62				
Road-150 Red, 44				
Road-150 Red, 56				
Road-150 Red, 44				
Road-150 Red, 44				

Q	R	S	T
ProductName	Location ID	Product Num	Warehouse ID
Road-150 Red, 62	BK	R93R	62
Road-150 Red, 44	BK	R93R	44
Road-150 Red, 62	BK	R93R	62
Road-150 Red, 48	BK	R93R	48
Road-150 Red, 48	BK	R93R	48
Road-150 Red, 52	BK	R93R	52
Road-150 Red, 56	BK	R93R	56
Road-150 Red, 56	BK	R93R	56
Road-150 Red, 44	BK	R93R	44
Road-150 Red, 62	BK	R93R	62
Road-150 Red, 44	BK	R93R	44
Road-150 Red, 56	BK	R93R	56
Road-150 Red, 44	BK	R93R	44
Road-150 Red, 44	BK	R93R	44
Road-150 Red, 48	BK	R93R	48
Road-150 Red, 44	BK	R93R	44
Road-150 Red, 62	BK	R93R	62
Road-150 Red, 48	BK	R93R	48

Breaking Apart Product Number into Fields by Using Text to Columns and Delimiter in Microsoft Excel (Used with permission from Microsoft.)

As you can see in our screenshot, the original ProductNumber field has been broken into three different columns to show the Location ID, the Product Num, and the Warehouse ID.

Date Functions

We live in a world that operates around time, so you are likely to encounter date-related data in your organization. Date fields reveal important information, such as when a product is ordered, shipped, and delivered. **Date functions** can help us derive valuable attributes from date fields, such as determining the day of the week, the week number, or the month or year from a single date. We can also use date functions for further analysis, such as determining how long it takes to accomplish a task (e.g., the amount of time from when an order is created to when it is delivered). You can also use dates to signal the start of a process. There are several software automation tools (such as Microsoft SharePoint, SalesForce, and PowerAutomate) that support workflows that are triggered by different calculations and dates.

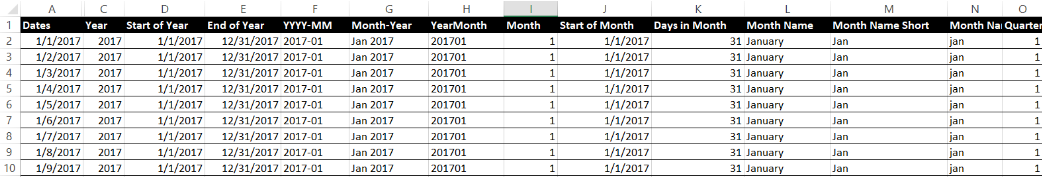
As a data analyst, you will encounter many different reasons to use date functions. Here are some of the most common functions and their purposes:

- **NOW()**—This function uses the computer system time to tell the current date and time of a calculation. When you use NOW() and another date like ShipDate in a function together, it can tell you how many days have elapsed from the time it was shipped, or how many days are left until it should be shipped.

- **TODAY ()**—This function gives you the date but not the time.
- **DATEDIFF()**—This function can be used to calculate the amount of days between two dates: a start date and an end point. NOW() or TODAY() is that start or end point for the DATEDIFF().
- **NETWORKDAYS()**—This function calculates how many business days exist between a start and end date. Business days are Monday through Friday, excluding Saturdays and Sundays.
- **WEEKDAY([StartDate])**—This Excel function will return a number 1 through 7 to designate the day of the week.
- **WEEKNUM([StartDate])**—This Excel function will return the number of the week of that year as 1 through 52.
- **MONTH([StartDate])**—This function returns a 1–12 to designate the month of the year.

Most software will have some variation of these date functions that you can use for your analysis.

Data analysts will commonly use date tables that are already prefilled with date information for every date of the year. In this type of table, you would see items like Month Name, Month Number, Quarter, Year, and Day of the Week. The use of a prefilled date table saves you from having to write these calculations over and over. Some common data tools will allow you to use queries to generate a date table for the purposes of your analysis.



	A	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Dates	Year	Start of Year	End of Year	YYYY-MM	Month-Year	YearMonth	Month	Start of Month	Days in Month	Month Name	Month Name Short	Month Na	Quarter
2	1/1/2017	2017	1/1/2017	12/31/2017	2017-01	Jan 2017	201701	1	1/1/2017	31	January	Jan	jan	1
3	1/2/2017	2017	1/1/2017	12/31/2017	2017-01	Jan 2017	201701	1	1/1/2017	31	January	Jan	jan	1
4	1/3/2017	2017	1/1/2017	12/31/2017	2017-01	Jan 2017	201701	1	1/1/2017	31	January	Jan	jan	1
5	1/4/2017	2017	1/1/2017	12/31/2017	2017-01	Jan 2017	201701	1	1/1/2017	31	January	Jan	jan	1
6	1/5/2017	2017	1/1/2017	12/31/2017	2017-01	Jan 2017	201701	1	1/1/2017	31	January	Jan	jan	1
7	1/6/2017	2017	1/1/2017	12/31/2017	2017-01	Jan 2017	201701	1	1/1/2017	31	January	Jan	jan	1
8	1/7/2017	2017	1/1/2017	12/31/2017	2017-01	Jan 2017	201701	1	1/1/2017	31	January	Jan	jan	1
9	1/8/2017	2017	1/1/2017	12/31/2017	2017-01	Jan 2017	201701	1	1/1/2017	31	January	Jan	jan	1
10	1/9/2017	2017	1/1/2017	12/31/2017	2017-01	Jan 2017	201701	1	1/1/2017	31	January	Jan	jan	1

Sample Date Table Displayed in Microsoft Excel (Used with permission from Microsoft.)

You can then join this date table within your data model, so that you do not have to calculate these types of standard date functions.

Logical Functions and Conditional Data

Logical functions will check if a condition is met and return a result based on whether or not the condition is met. Logical functions work with data that is either true or false. There are several types of logical functions; a couple of examples are IF, ISNULL, AND, and OR.

An IF function is a logical function that uses a logical test to validate whether a condition is true or false. The logical test will return a specified value if the condition is true and a different value if the condition is false.

To demonstrate one possible use of the IF function, recall the scenario introduced in Lesson 6 in which NULL values for the ship date field indicated that a product had not yet shipped.

Instead of leaving the blank values in this field, you could write an IF function to create a logical test that verifies whether that field is null. If the ship date field is empty, the IF function would return the words “Not Shipped,” and if the field has a date, the function would return the date the product was shipped. The function would look similar to this: IF(ISNULL(ShipDate),”Not Shipped”, “ShipDate”). The IF function could thereby provide a more readable field.

Although the syntax of functions will vary between programs, a logical function can be used to test whether a condition of the data is true or false and then supply the appropriate value.

Aggregation and the Basic Types of Aggregate Functions

There are functions that we write on a row-by-row basis, like an IF function or date conversion. In contrast, **aggregate functions** are written for all or a group of records, not just for a single record. Aggregate functions work with a column of data. For example, to calculate the total of all orders you would use the SUM function. To obtain the total number of orders placed by a region, you could group the records by region and then use an aggregate function like COUNT to count them.

Many people are exposed to aggregate functions through their use of Excel. For example, the Auto Sum feature is a familiar aggregate function used to total a column.

Here are some common examples of aggregate functions:

- **SUM** will add all the records together to produce a total. You will see this function used for amounts and quantities.
- **COUNT** will count all the records as individual lines to produce a record count.
- **DISTINCT COUNT** will count all the records in that column, but will only count the field one time, even if it appears multiple times.
- **AVERAGE** will total all the values in that column and then divide them by the count of values.
- **MAX** will give the largest value in that column.
- **MIN** will give the smallest value in that column.

System Functions

Most reporting tools come packaged with **system functions** that track report-related information (such as page numbers, refresh dates, report names, and more), removing the need for you to manually add this information. These functions provide the people who will read your reports with valuable insight about them. You may find that you commonly use system functions in the headers and footers of your reports. Here are some examples of helpful system functions:

- When running a weekly report, you may want to provide readers with its run date or refresh date/time. This information lets everyone know that the data is only up-to-date to that specific date or time. For instance, if a report is dated Thursday at 5 p.m., people viewing the report on Friday at 8 a.m. will automatically know that this data does not include events that occurred later Thursday night.
- Another helpful system function provides page numbers for paginated reports. This function is convenient when a report is multiple pages long, as the software will automatically generate page numbers.

Review Activity:

Functions to Manipulate Data

Answer the following questions:

1. **Which function removes all non-printable characters? Which function trims leading and trailing spaces? What type of function would we use to classify these as?**

2. **Which function is used to combine data fields?**

3. **If you had a data set with a column for Full Name, and you needed columns for First Name and Last Name, what action would you need to perform?**

4. **The TODAY function will provide and not provide what, respectively?**

5. **Which function is used to test whether a condition is true or false?**

6. **SUM, COUNT, DISTINCT COUNT, AVERAGE, MAX, and MIN are all what type of function?**

7. **Which function type is program or tool dependent?**

Topic 8B

Use Common Techniques for Query Optimization

EXAM OBJECTIVES COVERED

2.4 Explain common techniques for data manipulation and query optimization.

One of the outcomes of our work as data analysts is to hand data via dashboards or paginated reports to decision makers. In some cases, we also want these decision makers to be able to run these reports themselves. As analysts, one of our goals is to ensure that we write not only accurate queries with valid information, but also that we write them so that they are as efficient as possible.

In some organizations, there are specific job roles dedicated to optimizing query performance. The data analyst also uses techniques for optimizing queries. It could be that we leverage indexing or temporary tables to improve query performance, if our permissions allow us to do this. We can also write nested queries, verify the query execution plan, and increase performance where available for us to make changes.

Filtering Data

Databases will store all the data collected over time, but as a data analyst you will have the need to report on a specific date period or specific product. When this is the case, you will likely want to filter your data on the query level to reduce the amount of information that appears in the result.

If you are filtering your data in an SQL statement, you will be filtering it through the Where statement, as shown here.

```
SELECT ORDERID, ORDERDATE, ORD_AMOUNT

FROM Orders

WHERE ORDERDATE > 1/1/2020
```

When the query is executed, this code will only show the data that meets the specific requirements you set in the Where statement. In this example, the query will only show Order dates that are greater than Jan. 1, 2020. Every order before Jan. 1, 2020 will be excluded, giving you a much smaller data set. We will dig deeper into the variations of filters and types of filters in a later lesson.

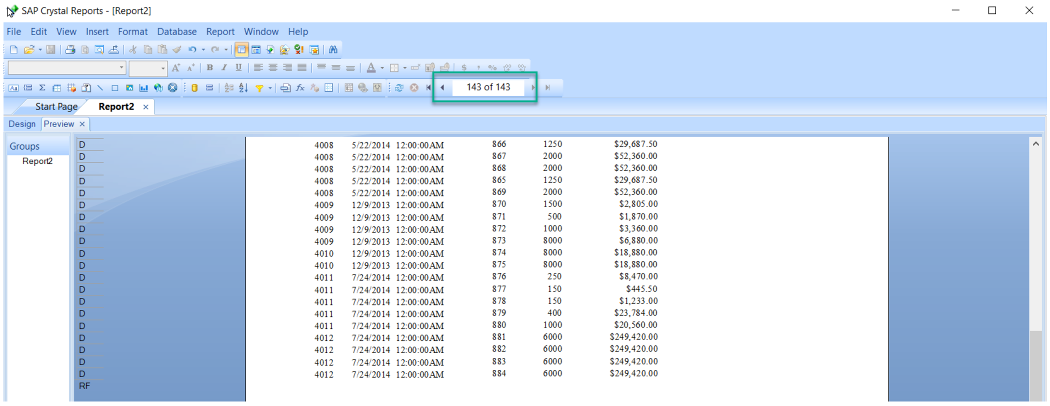
When you hard code a filter into a Where statement, it will always show only the data indicated with the filter. If you need to see different data every time you run this query, you would want to deploy another strategy instead, like the use of parameters.

Parameterization

A **parameter** is a method of adding a criteria to a query that can be used to filter and reduce the result set. Queries that pull a large amount of data over time can take a long time to process. One way to optimize data is by specifying a parameter in the query that filters the data before the result set is run, showing only what you need in the results.

The user who runs the query can either manually add values to the parameter, or these values can be pulled from another place in the software, leveraging the back-end design. For example, if a user executes a query on a particular product, the software itself can tell the query what value should be used to filter the data, through a parameter that's already been designed.

The benefit of parameters is that you start your work with just the data that you need versus bringing all of it into a tool like Tableau and filtering the data there.

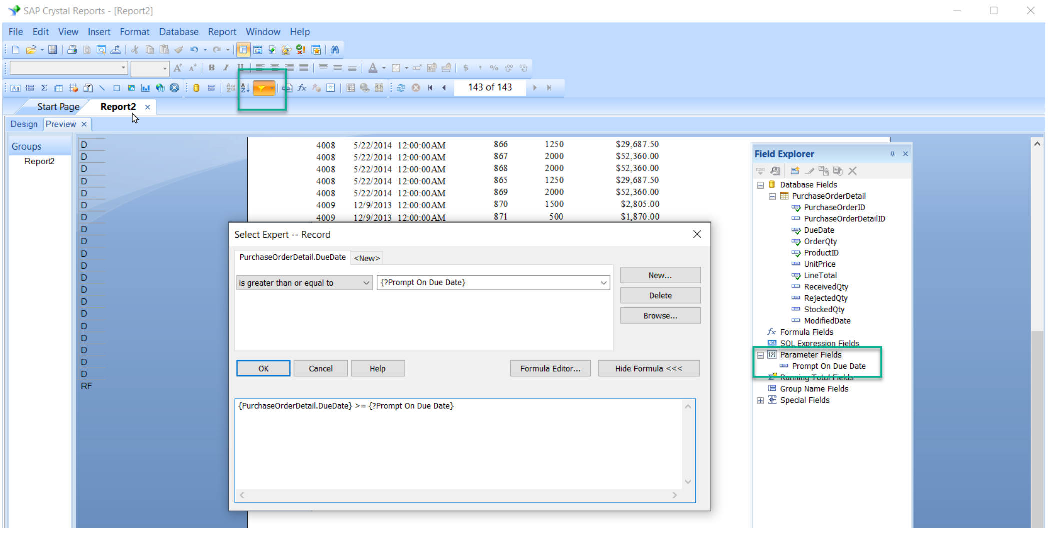


```
4008   5/22/2014  12:00:00AM   866   1250    $29,687.50
4008   5/22/2014  12:00:00AM   867   2000    $52,360.00
4008   5/22/2014  12:00:00AM   868   2000    $52,360.00
4008   5/22/2014  12:00:00AM   865   1250    $29,687.50
4008   5/22/2014  12:00:00AM   869   2000    $52,360.00
4009   12/9/2013  12:00:00AM   870   1500     $2,805.00
4009   12/9/2013  12:00:00AM   871    500     $1,870.00
4009   12/9/2013  12:00:00AM   872   1000     $3,360.00
4009   12/9/2013  12:00:00AM   873   8000     $6,880.00
4010   12/9/2013  12:00:00AM   874   8000    $18,880.00
4010   12/9/2013  12:00:00AM   875   8000    $18,880.00
4011   7/24/2014  12:00:00AM   876    250     $8,470.00
4011   7/24/2014  12:00:00AM   877    150       $445.50
4011   7/24/2014  12:00:00AM   878    150     $1,233.00
4011   7/24/2014  12:00:00AM   879    400    $23,784.00
4011   7/24/2014  12:00:00AM   880   1000    $20,560.00
4012   7/24/2014  12:00:00AM   881   6000   $249,420.00
4012   7/24/2014  12:00:00AM   882   6000   $249,420.00
4012   7/24/2014  12:00:00AM   883   6000   $249,420.00
4012   7/24/2014  12:00:00AM   884   6000   $249,420.00
```

AdventureWorks 2019 Order Details Report in Crystal Reports 2016 (Copyright © 2021 SAP SE or an SAP affiliate company. All rights reserved.)

The report displayed in the screenshot contains all data from Jan. 1, 2011 up until the end of 2014, which is when the data contained in this system ends. When this report is run without parameters and filters in place, we are provided with all the data in the table, producing 143 pages of reporting information. If we only want to look at certain periods of time, we can use parameters to filter out only what we need.

You can also develop parameters into reports that prompt you to enter information. These parameters are then used to pass data back to the command that filters the data, allowing you to run different data sets from a single report by changing the parameter.



4008	5/22/2014 12:00:00AM	866	1250	$29,687.50
4008	5/22/2014 12:00:00AM	867	2000	$52,360.00
4008	5/22/2014 12:00:00AM	868	2000	$52,360.00
4008	5/22/2014 12:00:00AM	865	1250	$29,687.50
4008	5/22/2014 12:00:00AM	869	2000	$52,360.00
4009	12/9/2013 12:00:00AM	870	1500	$2,805.00
4009	12/9/2013 12:00:00AM	871	500	$1,870.00

Select Expert -- Record

PurchaseOrderDetail.DueDate <New>

is greater than or equal to {?Prompt On Due Date}

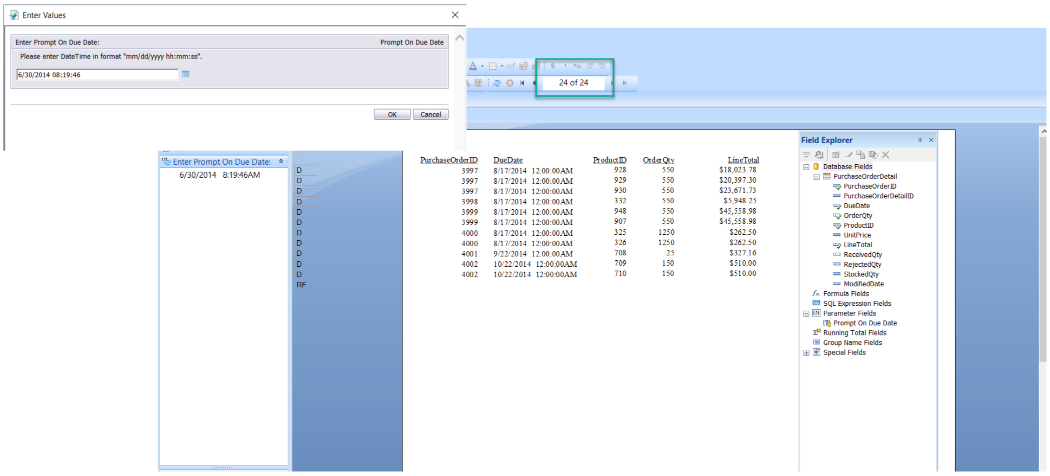
New... Delete Browse...

OK Cancel Help Formula Editor... Hide Formula <<<

```
{PurchaseOrderDetail.DueDate} >= {?Prompt On Due Date}
```

Screenshot of Parameters in Crystal Reports 2016
(Copyright © 2021 SAP SE or an SAP affiliate company. All rights reserved.)

As you can see in the screenshot, we have developed a parameter for the DueDate of the report. We then place that parameter into the filter of the report, which in Crystal Reports is called the Select Expert.



Enter Values

Enter Prompt On Due Date: Prompt On Due Date

Please enter DateTime in format "mm/dd/yyyy hh:mm:ss".

6/30/2014 08:19:46

OK Cancel

Enter Prompt On Due Date:
6/30/2014 8:19:46AM

PurchaseOrderID	DueDate	ProductID	OrderQty	LineTotal
3997	8/17/2014 12:00:00AM	928	550	$18,023.78
3997	8/17/2014 12:00:00AM	929	550	$20,397.30
3997	8/17/2014 12:00:00AM	930	550	$23,671.73
3998	8/17/2014 12:00:00AM	332	550	$5,948.25
3999	8/17/2014 12:00:00AM	948	550	$45,558.98
3999	8/17/2014 12:00:00AM	907	550	$45,558.98
4000	8/17/2014 12:00:00AM	325	1250	$262.50
4000	8/17/2014 12:00:00AM	326	1250	$262.50
4001	9/22/2014 12:00:00AM	708	25	$327.16
4002	10/22/2014 12:00:00AM	709	150	$510.00
4002	10/22/2014 12:00:00AM	710	150	$510.00

Parameter Prompt in Crystal Reports 2016
(Copyright © 2021 SAP SE or an SAP affiliate company. All rights reserved.)

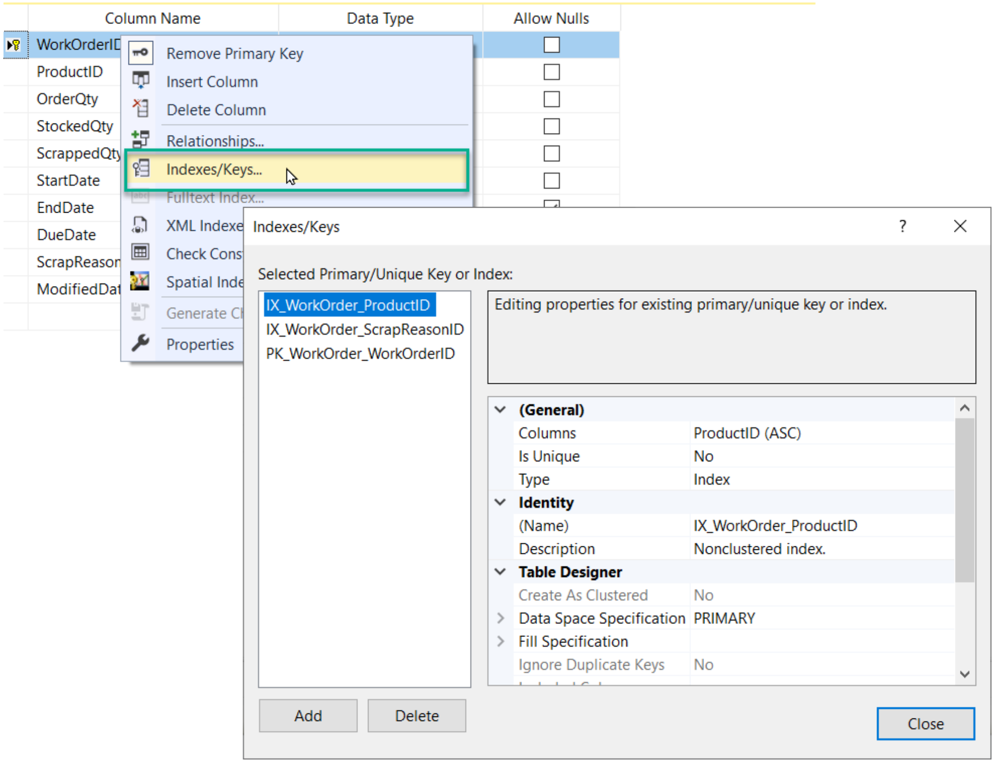
Now that we have developed the parameter, when we execute the report it will prompt us to enter our date, at which point it will filter the data accordingly. In the screenshot, we can see the filter is set to only include data from June 30, 2014 and on, resulting in a 24-page result. This is a drastically smaller and more manageable report than the original.

As with other techniques we have learned, each software we might use will handle the creation and use of parameters differently. For example, parameters can be directly coded in SQL statements to read data. When a query is run, the data is passed to the SQL statement and filtered to show only information related to the specified parameters. This process happens entirely in the back end of the system, but the end user sees the results.

Indexing Data

Fields that will be commonly sorted, queried, or filtered will likely be indexed. **Indexing** is a field property setting that tells the database that a field needs to be indexed. It's an internal processing that just controls how much data must be looked at when processing the data for queries. When fields are indexed, it makes data processing faster and ultimately speeds up the performance of the query. There are some fields that are automatically indexed when they are created, like key fields.

If you have access to the back-end database, you can confirm whether a field is indexed or not. However, you will need edit permissions for the database to change the properties. See the following screenshot for how to access the index in SQL Server Management Studio.



Column Name	Data Type	Allow Nulls
WorkOrderID		□
ProductID		□
OrderQty		□
StockedQty		□
ScrappedQty		□
StartDate		□
EndDate		
DueDate		
ScrapReason		
ModifiedDat		

Indexes/Keys

Selected Primary/Unique Key or Index:

IX_WorkOrder_ProductID
IX_WorkOrder_ScrapReasonID
PK_WorkOrder_WorkOrderID

Editing properties for existing primary/unique key or index.

(General)	
Columns	ProductID (ASC)
Is Unique	No
Type	Index
Identity	
(Name)	IX_WorkOrder_ProductID
Description	Nonclustered index.
Table Designer	
Create As Clustered	No
Data Space Specification	PRIMARY
Fill Specification	
Ignore Duplicate Keys	No

Indexes/Keys Screen via SQL Server Management Studio (Used with permission from Microsoft.)

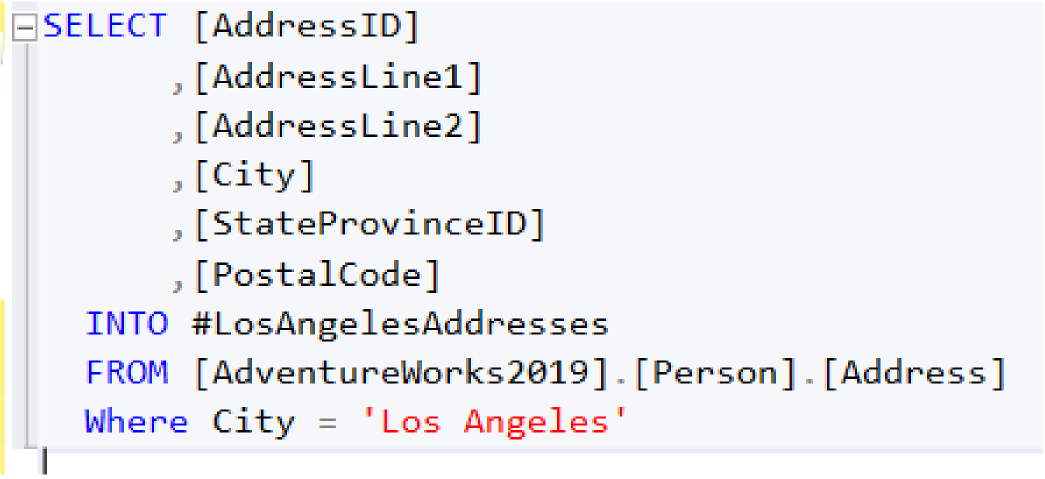
You can right-click any field in the table to access the index and keys for that table.

Temporary Tables

A **temporary table** is a table that is stored on the database server until a user disconnects from the server. Similar to a permanent table, temporary tables provide records, but they are for temporary use only. These temporary tables can improve processing speeds for queries simply because they likely contain a smaller number of records than the permanent table.

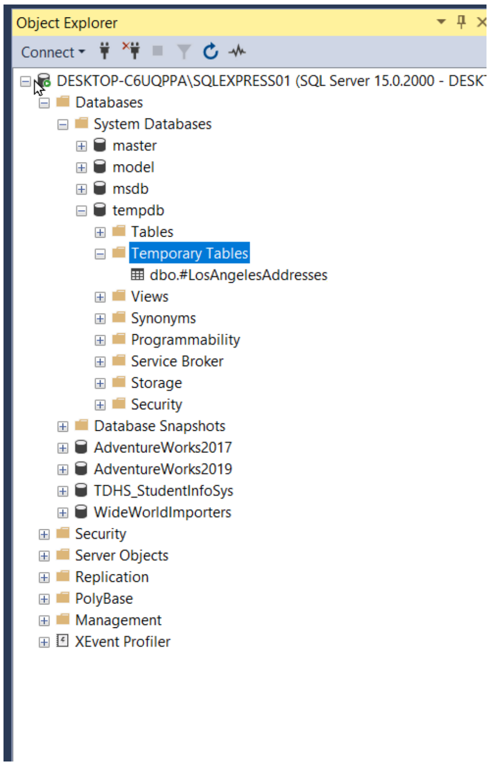
Most of this type of querying occurs in the backend of databases.As a data analyst, you may never see these processes if you don't have access to the design.

Temporary tables can be created through SQL statements. Create the table using the Create Table syntax, and then add records to the temporarily created table using Insert syntax. You would create the table when querying the permanent table.



```
SELECT [AddressID]
      ,[AddressLine1]
      ,[AddressLine2]
      ,[City]
      ,[StateProvinceID]
      ,[PostalCode]
  INTO #LosAngelesAddresses
  FROM [AdventureWorks2019].[Person].[Address]
  Where City = 'Los Angeles'
```

Create Table Syntax for a Temporary Table in Microsoft SQL Server Management Studio (Used with permission from Microsoft.)



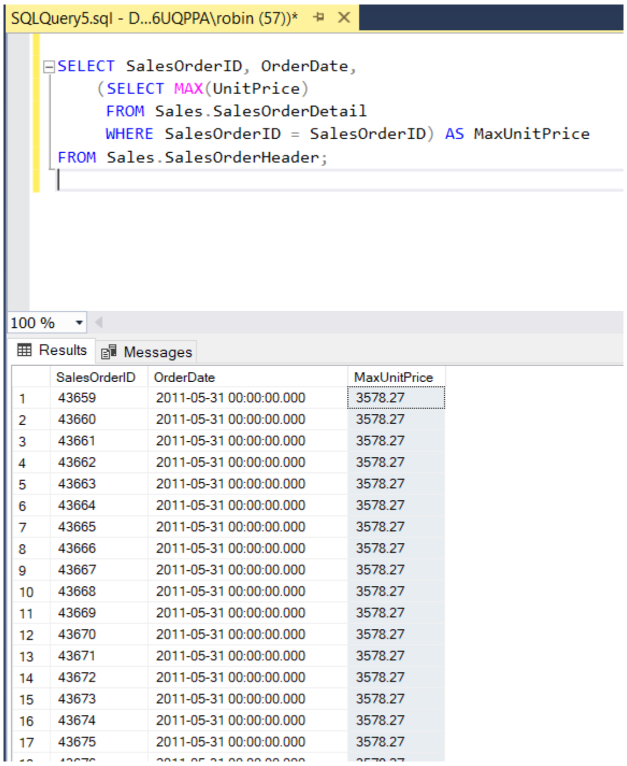
Temporary Table Found in Tempdb in Microsoft SQL Server Management Studio (Used with permission from Microsoft.)

Subquerying and Subsets of Information

In cases where you might not have the ability to create temporary tables within the design of a database, you may need to subquery information to get a subset of information.

A **subquery**, also commonly referred to as a nested query, is a query that is nested inside another query statement. Subquerying a subset of information means accessing a smaller set of data rather than querying the whole table. Reducing the data to only what you need will improve the performance of the data.

To demonstrate the use of a subquery, suppose that you need sales data, but only for the highest-priced item on that day. Without a subquery, you would have to run a query to get all of the sales data and then run another query to get the Max. With a subquery, you can basically run them together. In the screenshot, you can see the syntax used for the subquery.



```
SELECT SalesOrderID, OrderDate,
    (SELECT MAX(UnitPrice)
     FROM Sales.SalesOrderDetail
     WHERE SalesOrderID = SalesOrderID) AS MaxUnitPrice
FROM Sales.SalesOrderHeader;
```

	SalesOrderID	OrderDate	MaxUnitPrice
1	43659	2011-05-31 00:00:00.000	3578.27
2	43660	2011-05-31 00:00:00.000	3578.27
3	43661	2011-05-31 00:00:00.000	3578.27
4	43662	2011-05-31 00:00:00.000	3578.27
5	43663	2011-05-31 00:00:00.000	3578.27
6	43664	2011-05-31 00:00:00.000	3578.27
7	43665	2011-05-31 00:00:00.000	3578.27
8	43666	2011-05-31 00:00:00.000	3578.27
9	43667	2011-05-31 00:00:00.000	3578.27
10	43668	2011-05-31 00:00:00.000	3578.27
11	43669	2011-05-31 00:00:00.000	3578.27
12	43670	2011-05-31 00:00:00.000	3578.27
13	43671	2011-05-31 00:00:00.000	3578.27
14	43672	2011-05-31 00:00:00.000	3578.27
15	43673	2011-05-31 00:00:00.000	3578.27
16	43674	2011-05-31 00:00:00.000	3578.27
17	43675	2011-05-31 00:00:00.000	3578.27
18	43676	2011-05-31 00:00:00.000	3578.27

SQL View of a Subquery That Pulls the MaxUnitPrice in Microsoft SQL Server Management Studio (Used with permission from Microsoft.)

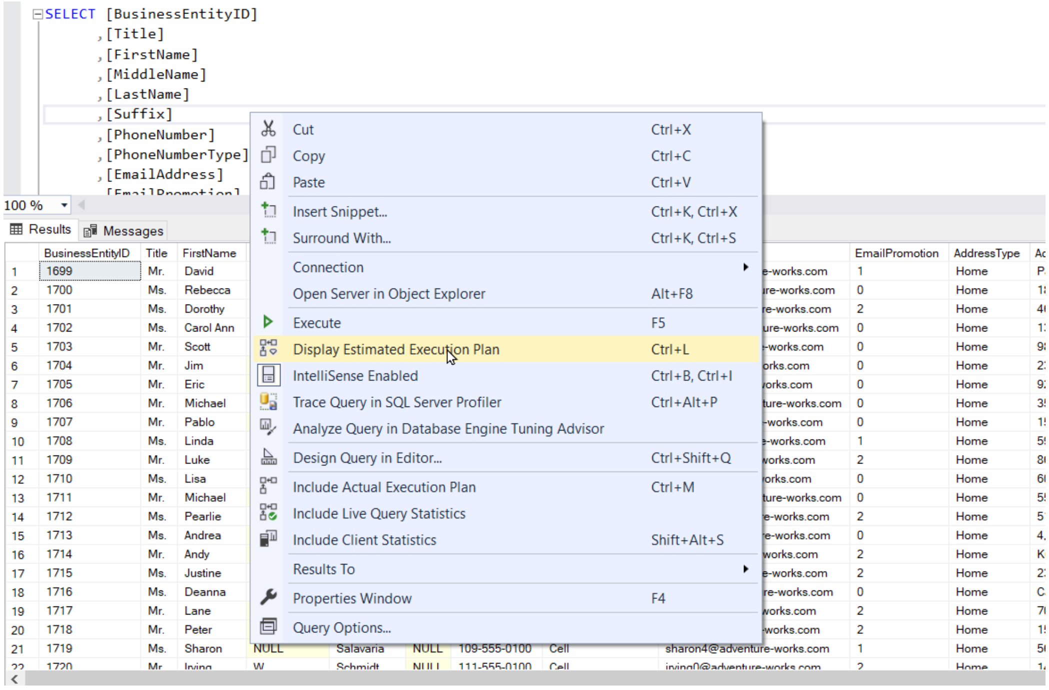
A subquery is placed within the statement of another query, allowing you to narrow the results of the query to only the data that you need. Anytime you query another query, you are creating a subquery. The original query must execute first. The nested query returns a specified subset of that retrieved data and provides the end result of the two queries.

Query Execution Plan

A **query execution plan** is the order of steps in which a query is processed. It is a visual representation that provides details about how the query executes. The query execution plan includes an **estimated execution plan** of possible requirements for executing the query and the **actual execution plan** that is known once the query has been executed.

The estimated execution plan can reveal areas of high server usage to process the information.

Your organization will likely have employee database administrators who spend time fine-tuning queries for optimal performance of the queries and also the server.



```
SELECT [BusinessEntityID]
      ,[Title]
      ,[FirstName]
      ,[MiddleName]
      ,[LastName]
      ,[Suffix]
      ,[PhoneNumber]
      ,[PhoneNumberType]
      ,[EmailAddress]
      ,[EmailPromotion]
```

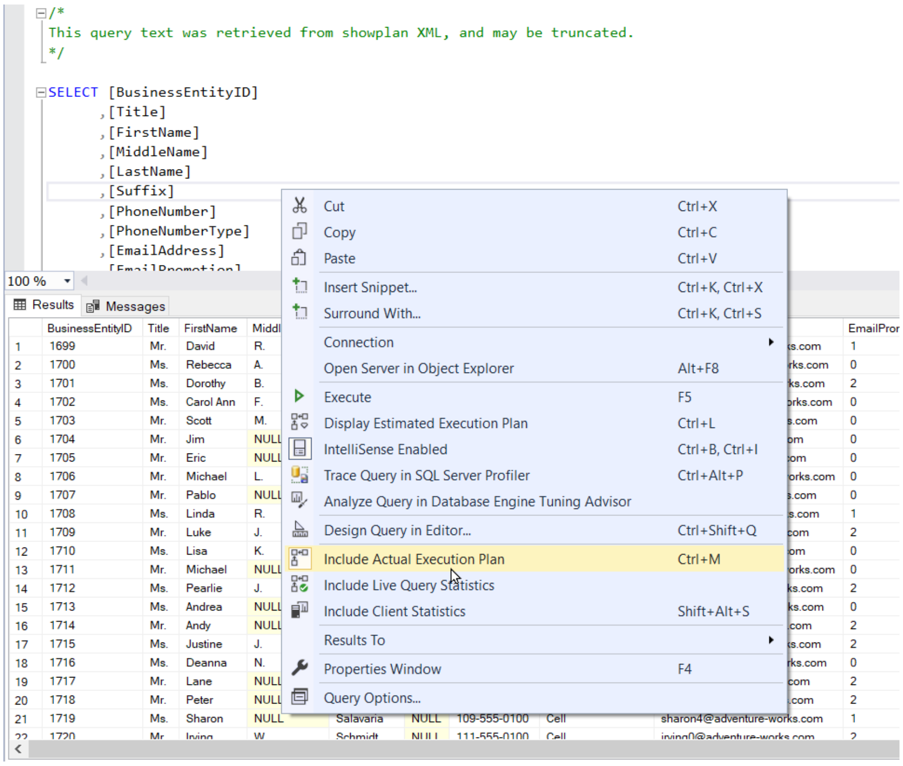
100 %
Results
Messages
Cut Ctrl+X
Copy Ctrl+C
Paste Ctrl+V
Insert Snippet... Ctrl+K, Ctrl+X
Surround With... Ctrl+K, Ctrl+S
Connection
Open Server in Object Explorer Alt+F8
Execute F5
Display Estimated Execution Plan Ctrl+L
IntelliSense Enabled Ctrl+B, Ctrl+I
Trace Query in SQL Server Profiler Ctrl+Alt+P
Analyze Query in Database Engine Tuning Advisor
Design Query in Editor... Ctrl+Shift+Q
Include Actual Execution Plan Ctrl+M
Include Live Query Statistics
Include Client Statistics Shift+Alt+S
Results To
Properties Window F4
Query Options...

	BusinessEntityID	Title	FirstName	EmailPromotion	AddressType
1	1699	Mr.	David	1	Home
2	1700	Ms.	Rebecca	0	Home
3	1701	Ms.	Dorothy	2	Home
4	1702	Ms.	Carol Ann	0	Home
5	1703	Mr.	Scott	0	Home
6	1704	Mr.	Jim	0	Home
7	1705	Mr.	Eric	0	Home
8	1706	Mr.	Michael	0	Home
9	1707	Mr.	Pablo	0	Home
10	1708	Ms.	Linda	1	Home
11	1709	Mr.	Luke	2	Home
12	1710	Ms.	Lisa	0	Home
13	1711	Mr.	Michael	0	Home
14	1712	Ms.	Pearlie	2	Home
15	1713	Ms.	Andrea	0	Home
16	1714	Mr.	Andy	2	Home
17	1715	Ms.	Justine	2	Home
18	1716	Ms.	Deanna	0	Home
19	1717	Mr.	Lane	2	Home
20	1718	Mr.	Peter	2	Home
21	1719	Ms.	Sharon	1	Home

Display the Estimated Plan Generated in Microsoft SQL Server Studio (Used with permission from Microsoft.)

In the case of queries that take a long time to execute, you can view the estimated execution plan beforehand, as shown in this screenshot.

When running the query, including an actual plan gives you a tab that lets you review the execution plan.



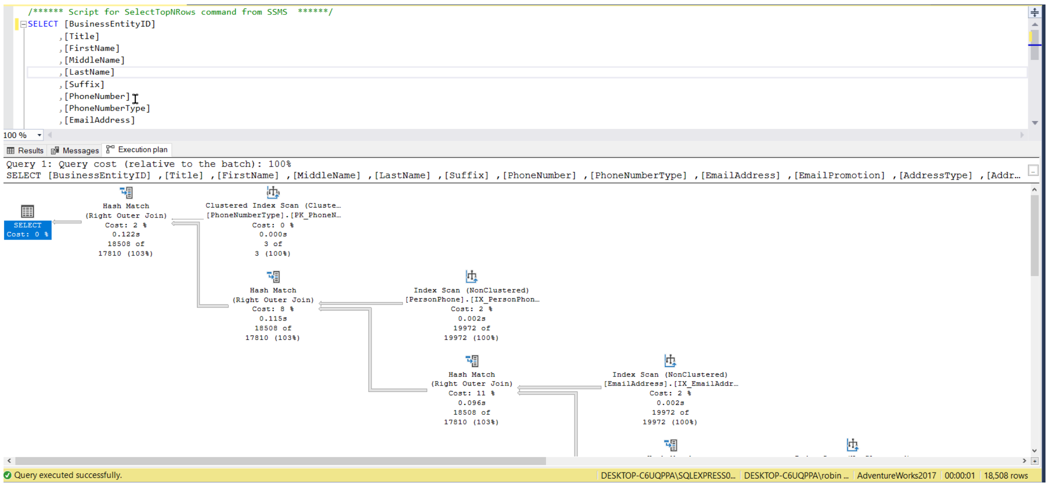
```
/*
This query text was retrieved from showplan XML, and may be truncated.
*/

SELECT [BusinessEntityID]
      ,[Title]
      ,[FirstName]
      ,[MiddleName]
      ,[LastName]
      ,[Suffix]
      ,[PhoneNumber]
      ,[PhoneNumberType]
      ,[EmailAddress]
      ,[EmailPromotion]
```

100 %
Results
Messages
Cut Ctrl+X
Copy Ctrl+C
Paste Ctrl+V
Insert Snippet... Ctrl+K, Ctrl+X
Surround With... Ctrl+K, Ctrl+S
Connection
Open Server in Object Explorer Alt+F8
Execute F5
Display Estimated Execution Plan Ctrl+L
IntelliSense Enabled Ctrl+B, Ctrl+I
Trace Query in SQL Server Profiler Ctrl+Alt+P
Analyze Query in Database Engine Tuning Advisor
Design Query in Editor... Ctrl+Shift+Q
Include Actual Execution Plan Ctrl+M
Include Live Query Statistics
Include Client Statistics Shift+Alt+S
Results To
Properties Window F4
Query Options...

	BusinessEntityID	Title	FirstName	Middl	EmailPror
1	1699	Mr.	David	R.	1
2	1700	Ms.	Rebecca	A.	0
3	1701	Ms.	Dorothy	B.	2
4	1702	Ms.	Carol Ann	F.	0
5	1703	Mr.	Scott	M.	0
6	1704	Mr.	Jim	NULL	0
7	1705	Mr.	Eric	NULL	0
8	1706	Mr.	Michael	L.	0
9	1707	Mr.	Pablo	NULL	0
10	1708	Ms.	Linda	R.	1
11	1709	Mr.	Luke	J.	2
12	1710	Ms.	Lisa	K.	0
13	1711	Mr.	Michael	NULL	0
14	1712	Ms.	Pearlie	J.	2
15	1713	Ms.	Andrea	NULL	0
16	1714	Mr.	Andy	NULL	2
17	1715	Ms.	Justine	J.	2
18	1716	Ms.	Deanna	N.	0
19	1717	Mr.	Lane	NULL	2
20	1718	Mr.	Peter	NULL	2
21	1719	Ms.	Sharon	NULL	1

Including the Actual Execution Plan (Used with permission from Microsoft.)

When viewing the query execution plan, each part of the graphical representation can be hovered over for a more detailed view of each step and its impacts. When you build queries for large data sets, you will want to deploy strategies like this to execute your queries faster.



```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT [BusinessEntityID]
      ,[Title]
      ,[FirstName]
      ,[MiddleName]
      ,[LastName]
      ,[Suffix]
      ,[PhoneNumber]
      ,[PhoneNumberType]
      ,[EmailAddress]
```

```
Query 1: Query cost (relative to the batch): 100%
SELECT [BusinessEntityID] ,[Title] ,[FirstName] ,[MiddleName] ,[LastName] ,[Suffix] ,[PhoneNumber] ,[PhoneNumberType] ,[EmailAddress] ,[EmailPromotion] ,[AddressType] ,[Addr…
```

Query executed successfully. | DESKTOP-C6UQPPA\SQLEXPRESS0… | DESKTOP-C6UQPPA\robin … | AdventureWorks2017 | 00:00:01 | 18,508 rows

Visual View of the Query Execution Plan (Used with permission from Microsoft.)

Review Activity:

Common Techniques for Query Optimization

Answer the following questions:

1. **What can you create to optimize data load time and filter data to certain criteria?**

2. **What fields are automatically indexed when they are created?**

3. **When does a temporary table stop being stored on the database server?**

4. **What is being created when querying another query?**

5. **What are the two types of query execution plans?**

Lesson 8

Summary

After this lesson, you should be able to describe some of the common types of functions that a data analyst will use to manipulate data: text functions, merge functions, parsing, date functions, logical functions, aggregate functions, and system functions. You should also have gained some additional strategies for optimizing your queries and be familiar with back-end options, like temporary tables and parameters.

Guidelines for Data Manipulation and Optimization

Consider these best practices and guidelines when manipulating data and optimizing queries.

1. Text functions can help you clean up leading and trailing spaces, delete non-printable characters, and remove a set number of characters.
2. You can also use text functions to transform data, like changing the case, counting the length, or joining characters (CONCATENATE).
3. You will use date functions and date tables in your day-to-day life, so make sure you are very familiar with them.
4. Logical functions are used when you need to test whether a condition of the data is true or false, and then to supply the appropriate value.
5. If you need to perform a calculation for all or a group of records, rather than for a single record, you will want to use an aggregate function.
6. System functions, packaged with most reporting tools, eliminate the need for you to manually track information about your report.
7. Filters can be applied to any data set to reduce the volume of data in the results.
8. Parameters are a way to specify filters to data and can be used through the backend of the data sets or through users answering prompts.
9. A subquery allows you to run two queries together and improves the performance of your data.
10. A query execution plan helps you estimate how the query will process. There are often people dedicated to this role when maintaining large-scale systems.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 9

Applying Descriptive Statistical Methods

LESSON INTRODUCTION

The work of a data analyst involves summarizing overall findings and insights after analyzing the data. Data analysts describe data in conversations and presentations, and thus learning how to describe the data through various methods (e.g., averages of data, range of data) is a necessary part of the analyst's role. Data analysis also draws conclusions about the distribution of data and may show these findings visually in many workplace scenarios.

At the beginning of a project, descriptive statistical methods are invaluable for understanding your data and finding issues that need to be addressed. Understanding how to describe your data will help you to either be confident in the accuracy and quality of the data, or speak to a lack of confidence in the data, depending on what your analysis reveals.

Lesson Objectives

In this lesson, you will do the following:

- Use measures of central tendency.
- Use measures of dispersion.
- Use frequency and percentages.

Topic 9A

Use Measures of Central Tendency

EXAM OBJECTIVES COVERED
3.1 Given a scenario, apply the appropriate descriptive statistical methods.

The **measures of central tendency** are mathematical functions used to find the center of a data set. Although these mathematical functions are easy enough to calculate by hand, the software we use today makes them even easier to perform. When approaching a new data set, you will be expected to determine the measures of central tendency at a minimum. The answers to these calculations are the beginning of your discovery of what's in the data and will help you to describe the data during your analysis.

Overview of the Measures of Central Tendency

When working with data, finding the central or middle values in the data can help you to describe the overall data set. Whether you are analyzing sales orders, student scores on an achievement test, or any other data set, the calculation of these functions will provide insights into your data. The most common measures of central tendency are the mean, the median, and the mode.

- **Mean** is the average of a set of numbers, calculated by adding all the values and then dividing that sum by the total number of values.
- **Median** is the middle number within a group of sorted numbers. When two middle numbers exist, they are then averaged to calculate the median.
- **Mode** is the number that shows up most often in a data set. It may also be referred to as the modal value.

Mean

The mean, or average of a group of numbers, has many uses in data analysis. To find the mean, add all the values together, count the number of values, and divide the sum by that number. Most programs will have a mathematical function for the mean, so you do not have to manually perform the calculation. For example, in Excel, this function is called the average.

Here are just a few examples of when it might be helpful to determine the mean within a workplace:

- A company is analyzing the annual sales goals for next year. The organization wants to look at sales performance for the last three years to determine what they will forecast for the next year. As opposed to just relying on guesswork, calculating the mean of total sales for those three years would give an average value based on sales history that the company can use with more confidence to develop their future sales goals.

- A company is setting goals for its employees in sales, as well as gauging performance of a new salesperson. The mean can help the company decide what performance should be expected. For example, the average sales achieved by every new salesperson in the first year could serve as the initial first-year target goal for all new salespeople. The mean provides a valuable and realistic goal for each new salesperson hired.

Calculating the Mean

To demonstrate the steps of calculating the mean, let's use an example of test scores. The mean of a group of scores can provide insight into the performance of a group of students. For example, suppose a school has two eighth-grade classrooms. In this scenario, you can use the mean to calculate the average test score for each classroom as well as the school's entire eighth-grade average (i.e., the mean of scores for all eighth-grade students).

Student test scores are grouped by classroom. The mean score can be calculated for classroom 105, classroom 106, and both classrooms, as shown in the accompanying screenshot. These averages can be compared with one another to obtain insight into that school year's eighth-grade test scores, as shown in the next two screenshots.

Grade Level	Classroom Number	Student Score
8th	105	75
8th	105	80
8th	105	95
8th	105	55
8th	105	80
8th	105	88
8th	105	99
8th	105	75
8th	105	70
8th	105	90
8th	105	85
8th	105	70
8th	105	95
8th	105	80

$$Total\ Scores\ for\ Classroom\ 105 = 1,137$$

$$Test\ Score\ Records\ for\ Classroom\ 105 = 14$$

$$Mean\ Score\ for\ Classroom\ 105 = 1,137 \div 14 = 81.21429$$

Grade Level	Classroom Number	Student Score
8th	106	75
8th	106	80
8th	106	45
8th	106	55
8th	106	80
8th	106	88
8th	106	85
8th	106	75
8th	106	70
8th	106	85
8th	106	85
8th	106	70
8th	106	75

$$\begin{aligned}
&Total\ Scores\ for\ Classroom\ 106 = 968\\
&Test\ Score\ Records\ for\ Classroom\ 106 = 13\\
&Mean\ Score\ for\ Classroom\ 106 = 968 \div 13 = 74.46154
\end{aligned}$$

$$\begin{aligned}
&Total\ Scores\ for\ Eighth\ Grade = 2,105\\
&Test\ Score\ Records\ for\ Eighth\ Grade = 27\\
&Mean\ Score\ for\ Eighth\ Grade = 2,105 \div 27 = 77.96296
\end{aligned}$$

The mean score for classroom 105 is 81.21 and is three points higher than the collective average of both classrooms. The mean score for classroom 106 is 74.46, which is three points lower than the collective average of both classrooms.

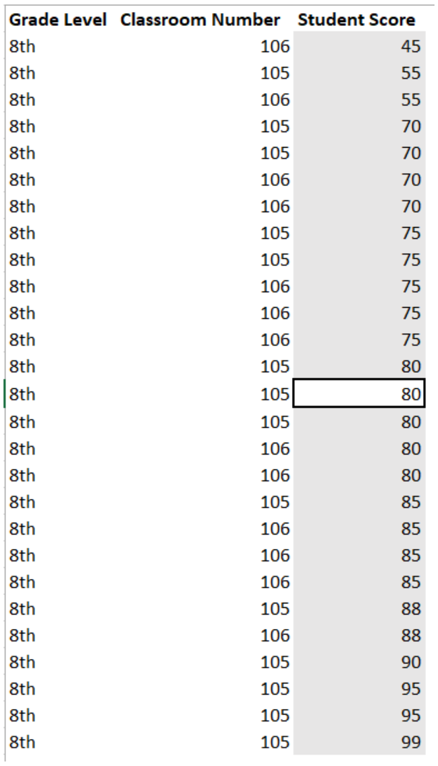
The mean is a very versatile calculation that returns a single aggregated value, but keep in mind that the mean is just one value that can be determined for a group of numbers. When combined with other descriptive statistics, the mean can create a valuable set of information for a business to use.

Median

The median, the middle value within an ordered set of numbers, allows you to establish a midpoint value for the data and see how many values are above or below it. Sometimes, the median can be more meaningful than the mean when describing data, as we can demonstrate using our example of student test scores. Recall that the average score is a 77.96 for both eighth-grade classrooms. The average is determined based on the total of all the numbers, so a few high or low scores can skew the mean, making this measure less informative in assessing student performance. To address this flaw, we'll find the middle score (median) for both classrooms

1. First, sort the values in ascending or descending order.
2. Next, determine the score that is in the middle. Because the total number of scores is an odd number, we can pinpoint the exact middle value, which in this case is 80.

A median score of 80 means that at least half of the students scored above 80, even though the average was only 77.96. These measures indicate that a few students scored really low, thereby bringing down the mean/average.

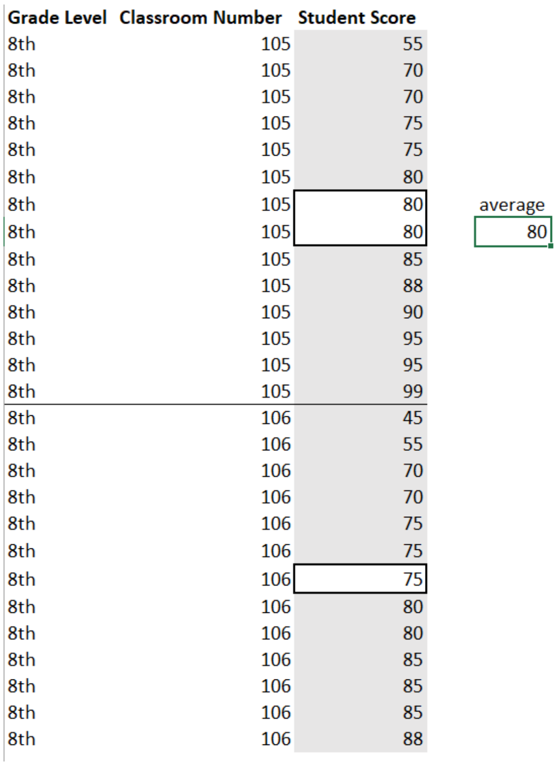


Grade Level	Classroom Number	Student Score
8th	106	45
8th	105	55
8th	106	55
8th	105	70
8th	105	70
8th	106	70
8th	106	70
8th	105	75
8th	105	75
8th	106	75
8th	106	75
8th	106	75
8th	105	80
8th	105	80
8th	105	80
8th	106	80
8th	106	80
8th	105	85
8th	106	85
8th	106	85
8th	106	85
8th	105	88
8th	106	88
8th	105	90
8th	105	95
8th	105	95
8th	105	99

Determining the Median for All Eighth-Grade Test Scores

Continuing with this example, let's next determine the median for each individual classroom. Classroom 105 has 14 students, an even number, and classroom 106 has 13 students, an odd number.

1. First, sort each classroom by scores in ascending order.
2. For classroom 106, with an odd number of students, simply find the number directly in the middle.
3. For classroom 105, with an even number of students, average the two middle scores in the list to determine the median.



Grade Level	Classroom Number	Student Score
8th	105	55
8th	105	70
8th	105	70
8th	105	75
8th	105	75
8th	105	80
8th	105	80
8th	105	80
8th	105	85
8th	105	88
8th	105	90
8th	105	95
8th	105	95
8th	105	99
8th	106	45
8th	106	55
8th	106	70
8th	106	70
8th	106	75
8th	106	75
8th	106	75
8th	106	80
8th	106	80
8th	106	85
8th	106	85
8th	106	85
8th	106	88

average
80

Determining the Median for the Test Scores of Each Classroom, One with an Odd Number of Values and One With an Even Number of Values

The data set in this screenshot reveals that students scored lower overall in classroom 106 based on both the mean score (74.46) and the median score (75), as opposed to the mean and median scores for classroom 105 (81.21 and 80, respectively). We can also draw possible conclusions about student performance within each classroom based on the median scores. For example, in classroom 106, the median score of 75 is higher than the classroom average of 74.46, suggesting that only one or two low scores may be driving the class average down.

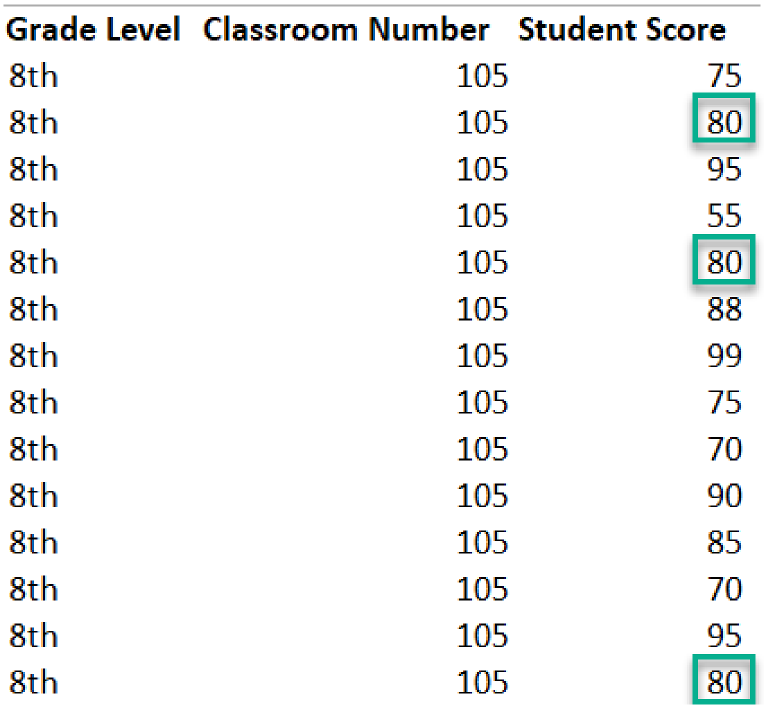
As this example demonstrates, looking at both the mean and median may help you to spot outliers in the data more effectively. **Outliers** are values in the data set that don't seem to be within the norm of all the other data. For instance, in the analysis of eighth-grade student test scores, a few very high or low scores were outliers that impacted classroom averages and overall measures of eighth-grade student performance.

Mode

The mode is determined by counting which values show up most frequently in a set of numbers. Simply count how many times each single number shows up in the data, and the number that shows up the most is the mode. Continuing with the example of student test scores, you can determine the mode either by using a calculation or by manually counting. You may end up with either a single mode or a multiple mode.

- A single mode is one single number that appears most often in the data set.
- A multiple mode exists when two or more numbers appear an equal number of times.

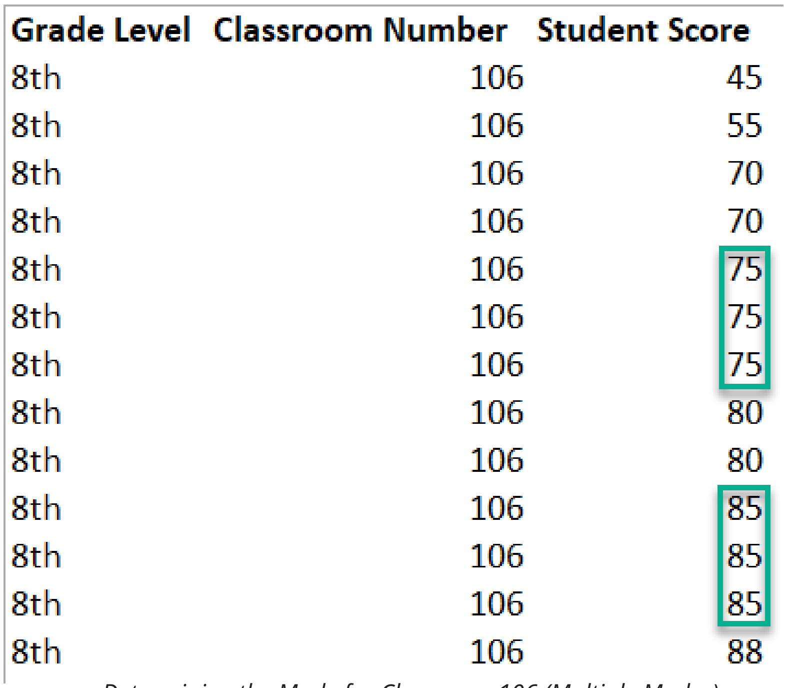
In Excel, you can use the calculations =MODE.SNGL or =MODE.MULT to return the mode on a group of numbers. The following screenshots show how to determine the mode for Classrooms 105 and 106.



Grade Level	Classroom Number	Student Score
8th	105	75
8th	105	80
8th	105	95
8th	105	55
8th	105	80
8th	105	88
8th	105	99
8th	105	75
8th	105	70
8th	105	90
8th	105	85
8th	105	70
8th	105	95
8th	105	80

Determining the Mode for Classroom 105 (Single Mode)

For classroom 105, the score of 80 is the mode because it has the highest frequency, appearing three times in the data set.



Grade Level	Classroom Number	Student Score
8th	106	45
8th	106	55
8th	106	70
8th	106	70
8th	106	75
8th	106	75
8th	106	75
8th	106	80
8th	106	80
8th	106	85
8th	106	85
8th	106	85
8th	106	88

Determining the Mode for Classroom 106 (Multiple Modes)

For classroom 106, there are two modes: 75 and 85 are the most frequent scores, each showing up three times in the data set.

Review Activity:

Measures of Central Tendency

Answer the following questions:

1. **Which measure of central tendency is calculated by adding all the values together, counting the number of values, and dividing the sum by that number?**

2. **What are values in the data set that don't seem to be within the norm of all the other data?**

3. **Which measure of central tendency is the middle value within an ordered set of numbers?**

4. **When you count which value shows up most frequently in a set of numbers, what will you end up with? What if two or more values show up an equal amount of times?**

Topic 9B

Use Measures of Dispersion

EXAM OBJECTIVES COVERED

3.1 Given a scenario, apply the appropriate descriptive statistical methods.

Along with measures of central tendency, you will also use **measures of dispersion** to determine how spread out the data is from the center of the data set. You will learn how to calculate range, variance, standard deviation, and z-scores to better describe your data set and the distribution of your data points. Knowing the variability and distribution of your data will help you determine accurate tests to use in further analysis.

Overview of the Measures of Dispersion

Several measures commonly used to determine the distribution ("spread") of a data set are described below.

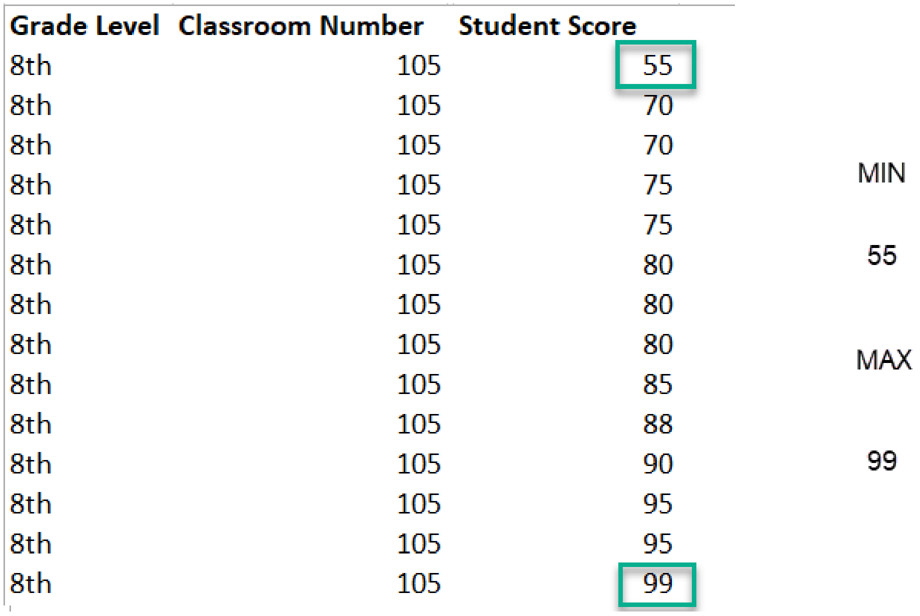
- **Min**, which stands for minimum, is the smallest number in the data set.
- **Max**, or maximum, is the largest number in the data set.
- **Range** is the difference between the highest and lowest values.
- **Variance** is the average squared distance from the mean of the data for a single data point.
- **Standard deviation** shows how dispersed the data is in relation to the mean of all of the data.
- A **z-score** shows how many standard deviations a data point is from the mean.

These measures can be manually calculated fairly easily if you know how to look for their values. However, practically speaking you will likely use the functions that are built into programs like Excel, Tableau, or SPSS (a popular statistical analysis tool) to perform these types of calculations.

The Range of Data

To find the range of a data set, you must start by finding the highest (maximum) value and the lowest (minimum) value. Then subtract the lowest value from the highest value to determine the range of the data.

Let's revisit the example of student test scores to calculate the range. As shown in the following screenshot, test scores are sorted lowest to highest. The first value is the minimum, and the last value is the maximum. These are the two key values you need to determine the range.



Grade Level	Classroom Number	Student Score
8th	105	55
8th	105	70
8th	105	70
8th	105	75
8th	105	75
8th	105	80
8th	105	80
8th	105	80
8th	105	85
8th	105	88
8th	105	90
8th	105	95
8th	105	95
8th	105	99

MIN

55

MAX

99

Identifying the Min and Max in a Data Set

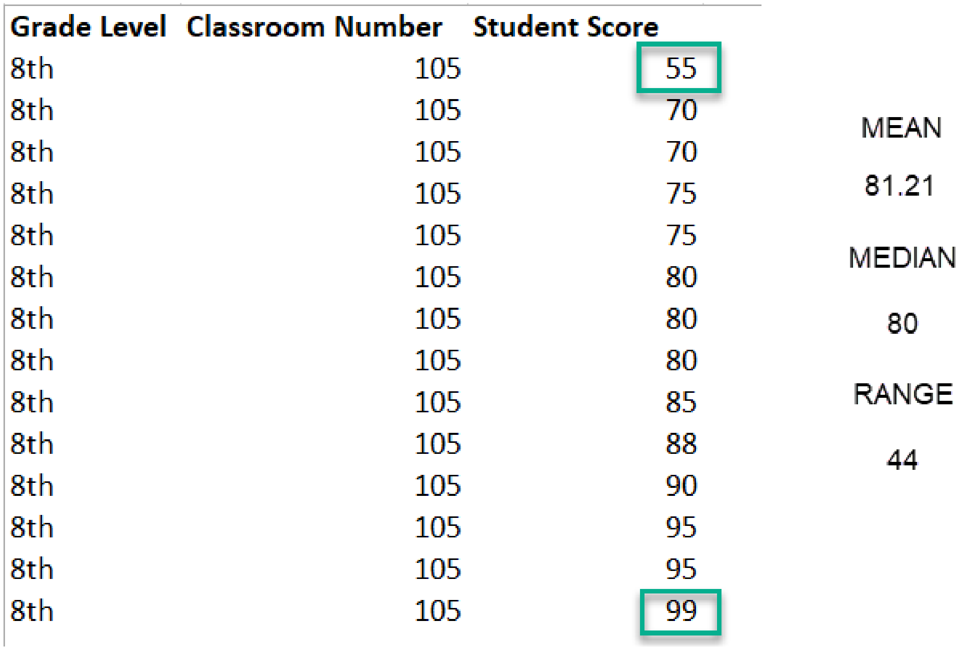
The first value is the minimum, and the last value is the maximum. These are the two key values you need to determine the range.

$$RANGE = MAX - MIN$$

$$RANGE = 99 - 55 = 44$$

We subtract the minimum number (55) from the maximum number (99) to calculate the range. In this example, the range is 44.

The range of 44 shows us that there is a lot of variability in this data set, and it's likely due to an outlier. Remember that an outlier is a value that lies far outside the norm of the other data points. In this case, the outlier could be a very low score or a very high score. Let's look at the mean, median, and range of the data together for classroom 105.



Grade Level	Classroom Number	Student Score
8th	105	55
8th	105	70
8th	105	70
8th	105	75
8th	105	75
8th	105	80
8th	105	80
8th	105	80
8th	105	85
8th	105	88
8th	105	90
8th	105	95
8th	105	95
8th	105	99

MEAN

81.21

MEDIAN

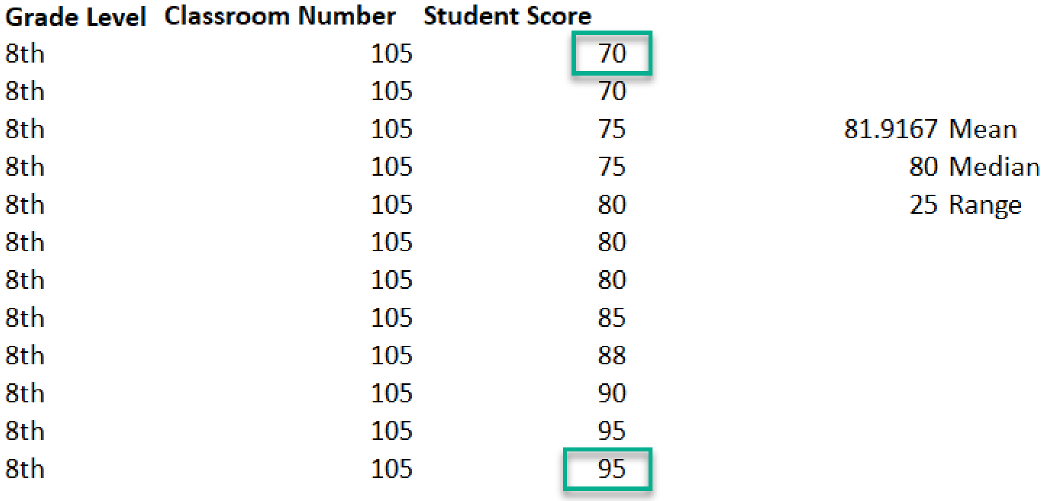
80

RANGE

44

Range of the Entire Classroom 105

If the highest and lowest scores were closer in value, the range would be smaller, indicating less variation in the data set. For example, removing the highest score of 99 and the lowest score of 55 would drive the range value down, as shown in the next screenshot.



Grade Level	Classroom Number	Student Score		
8th	105	70		
8th	105	70		
8th	105	75	81.9167	Mean
8th	105	75	80	Median
8th	105	80	25	Range
8th	105	80		
8th	105	80		
8th	105	85		
8th	105	88		
8th	105	90		
8th	105	95		
8th	105	95		

Range After Removing the Highest and Lowest Scores

When the highest and lowest scores are removed, there is now only a 25-point difference between the highest and lowest values, as opposed to the original range of 44 points. As this example demonstrates, the range is smaller when the values are closer together.

Note that by itself, the range does not provide enough information about the distribution of data when the data set contains an outlier. Range will often be used as part of another function to help determine the distribution.

Standard Deviation

Standard deviation is a statistical measure of variability that uses all the data points in the data set and shows how dispersed the data is in relation to the mean. Standard deviation is represented by the formula shown.

$$\sqrt{\frac{\sum(x-\overline{x})^2}{n-1}}$$

Formula for Standard Deviation Calculation

In the formula, *x* represents each of the values of the data, *x-bar* is the mean, *n* represents the number of data points, and the sigma symbol is used to denote a sum.

The standard deviation formula is calculated using the following steps:

1. For each value in the data set, subtract the mean and square that number. Then add together the results of these calculations and divide that sum by the total number of values in the data set minus one. The result of this calculation represents the variance—the average squared distance from the mean of the data for a single data point.

2. Determine the standard deviation by taking the square root of the variance.

Calculating Standard Deviation

Let's break this formula down to make it more understandable. This easiest variable to determine is *n*, which represents our sample size (number of values). Using the example of student test scores for classroom 105, *n* is the total number of student scores for the class. You can use a count function or manually count the number of data points to determine *n*.

Grade Level	Classroom Number	Student Score
8th	105	55
8th	105	70
8th	105	70
8th	105	75
8th	105	75
8th	105	80
8th	105	80
8th	105	80
8th	105	85
8th	105	88
8th	105	90
8th	105	95
8th	105	95
8th	105	99

There are 14 scores. For this part of the formula, subtract 1 from 14 to get 13, and use this number as the denominator in the standard deviation calculation.

$$\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$$\sqrt{\frac{\sum(x-\bar{x})^2}{14-1}}$$

$$\sqrt{\frac{\sum(x-\bar{x})^2}{13}}$$

Next, we need to calculate the value for *x-bar*, a statistical symbol depicted as an *x* with a line above it. The *x-bar* represents the mean of the scores, which in this case is 81.

$$\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$$\sqrt{\frac{\sum(x-81)^2}{13}}$$

The next step is to calculate (x – mean)^2 for each student score. The *x* represents each value of the data (or each individual student score). Thus, we subtract the mean from each student score and then square that value.

Student Score	–	Mean	= (Score – 81)
55	–	81	–26
70	–	81	–11
70	–	81	–11
75	–	81	–6
75	–	81	–6
80	–	81	–1
80	–	81	–1
80	–	81	–1
85	–	81	4
88	–	81	7
90	–	81	9
95	–	81	14
95	–	81	14
99	–	81	18

Deducting the Mean from the Score

For example, for the first score, the calculation is 55 – 81 = –26. Perform this same calculation for all other individual scores in classroom 105.

Then we square each resulting value by multiplying the number by itself.

Student Score	–	Mean	= (Score – 81)	(Score – Mean)^2
55	–	81	–26	676
70	–	81	–11	121
70	–	81	–11	121
75	–	81	–6	36
75	–	81	–6	36
80	–	81	–1	1
80	–	81	–1	1
80	–	81	–1	1
85	–	81	4	16
88	–	81	7	49
90	–	81	9	81
95	–	81	14	196
95	–	81	14	196
99	–	81	18	324

Squaring the Values

For the first score, we multiply -26 x 26 to get a value of 676. This same calculation is performed for every score in the data set.

Note that, mathematically, when you square a negative value it returns a positive value.

Now that we've calculated the (x – mean)^2 value for each student score, we will total (or sum) the results, which gives us our numerator: 1,855. We divide our numerator by our denominator: 1,855 / 13 = 142.69.

$$\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$$\sqrt{\frac{SUM(676,121,121,36,36,1,1,1,16,49,81,196,196,324)}{13}}$$

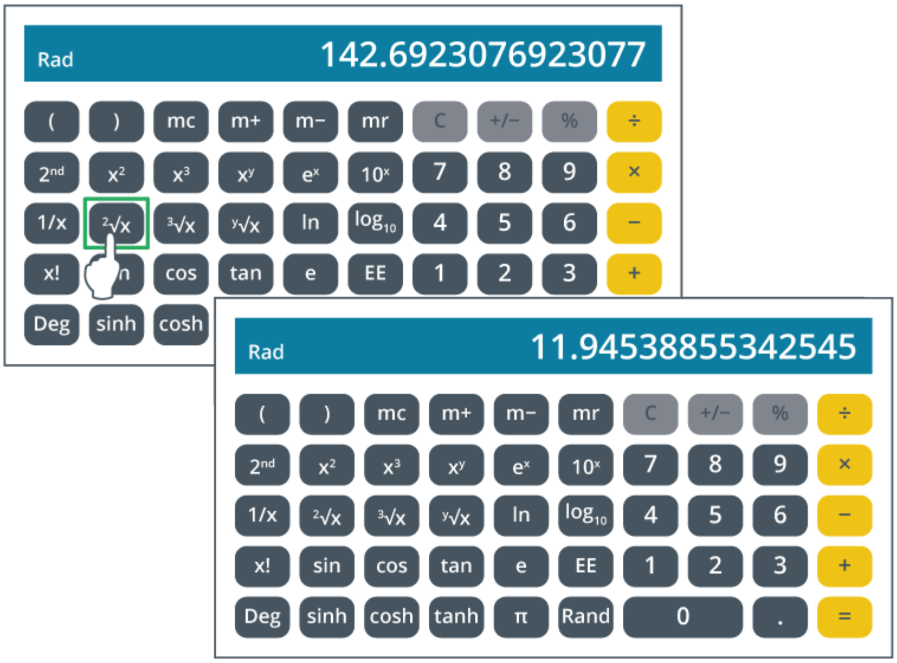
$$\sqrt{\frac{1855}{13}}$$

$$\sqrt{142.69} = 11.945$$

The final step is to determine the square root of the resulting value. In mathematics, the square root of a number equals the number that, when multiplied by itself, yields that given quantity.

The square root of 142.69 is 11.945, or 11.95 after rounding (11.95 x 11.95 = 142.69). **11.95** is our standard deviation for the entire group of scores in classroom 105.

There are multiple ways to achieve the square root of a value, but the most common is to use the square root button on a calculator.



Rad 142.6923076923077

Rad 11.94538855342545

( ) mc m+ m− mr C +/− % ÷
2nd x² x³ xʸ eˣ 10ˣ 7 8 9 ×
1/x ²√x ³√x ʸ√x ln log10 4 5 6 −
x! sin cos tan e EE 1 2 3 +
Deg sinh cosh tanh π Rand 0 . =

Using a Scientific Calculator to Determine the Square Root

Meaning of Standard Deviation

Standard deviation gives a more accurate depiction of the average distance of the values from the mean, providing a better understanding of the variability within a data set. A low standard deviation indicates that the values are closer to the mean, whereas a high standard deviation tells you the values are more spread out, and thus further from the mean.

In our example of student scores, the mean score was 81, and on average, the data values were about 11.95 points either way plus or minus from that mean.

Z-Scores

To determine how data is distributed, we can calculate the z-score of each individual data point and then visualize it. A z-score is used to pinpoint how many standard deviations away from the mean a value actually is. Calculating the z-score allows us to easily see if our data is following the empirical rule and clearly shows how much deviation is occurring in the data through a visual. The z-score is calculated using the following formula

$$z = \frac{x - \bar{x}}{S}$$

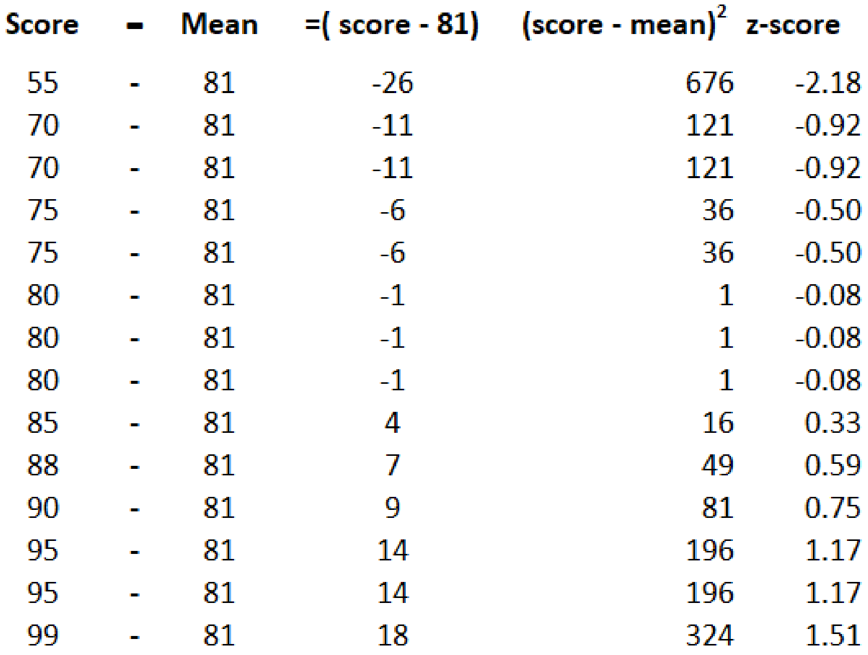
Formula for Z-Score Calculation

In the formula, *s* is the standard deviation of a sample, *x* represents each value in the data set, *x-bar* is the mean of all values in the data set, and *z* is the standard score.

Continuing with the data set for classroom 105, let's calculate the z-score for the student with a test score of 55. Using our formula, *x* = 55; the *x-bar* (mean) is 81; and *S* is the standard deviation, which we previously calculated as 11.95. The calculation is performed as follows: (55 – 81) / 11.95 or –26 / 11.95 = –2.18. This z-score means that the student's test score is a little over 2 standard deviations away from the middle of the data. (On a distribution curve, this data point would be placed on the left side of the bell curve. We'll show visuals of the distribution curve in the next section.)

Next, let's calculate the z-score for our highest score, which is 99. Plugging in our values, we calculate (99 – 81) / 11.95 or 18 / 11.95 = 1.51. The highest test score for classroom 105 is a little over 1.5 standard deviations away from the center of the data.

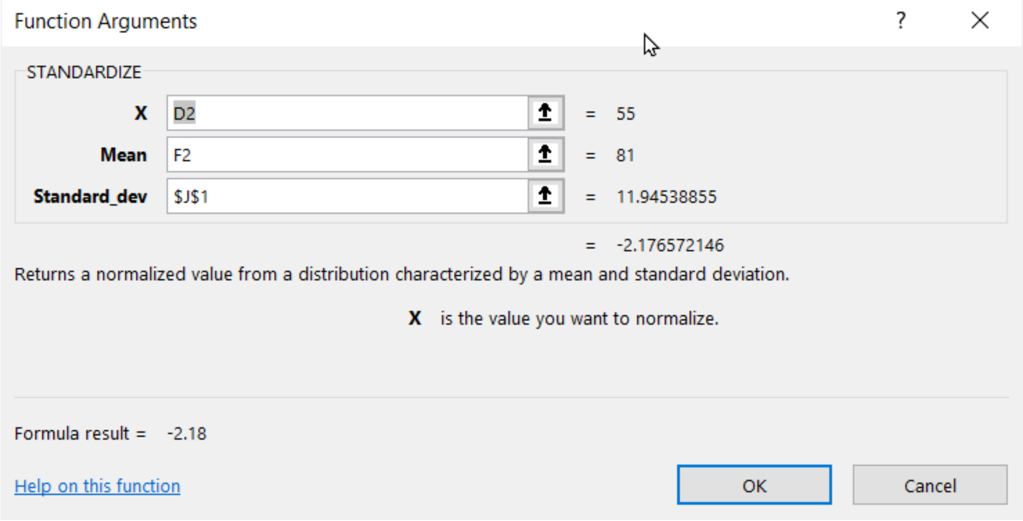
The next image shows the z-scores for all of the test scores, indicating their overall distance from the center of the data.



Score	-	Mean	=(score - 81)	$(\text{score - mean})^2$	z-score
55	-	81	-26	676	-2.18
70	-	81	-11	121	-0.92
70	-	81	-11	121	-0.92
75	-	81	-6	36	-0.50
75	-	81	-6	36	-0.50
80	-	81	-1	1	-0.08
80	-	81	-1	1	-0.08
80	-	81	-1	1	-0.08
85	-	81	4	16	0.33
88	-	81	7	49	0.59
90	-	81	9	81	0.75
95	-	81	14	196	1.17
95	-	81	14	196	1.17
99	-	81	18	324	1.51

Z-Score Values for All of Our Student Test Scores

In Excel, we can use the STANDARDIZE () function to calculate z-scores.



STANDARDIZE

X	D2	= 55
Mean	F2	= 81
Standard_dev	$J$1	= 11.94538855

= -2.176572146

Returns a normalized value from a distribution characterized by a mean and standard deviation.

X is the value you want to normalize.

Formula result = -2.18

STANDARDIZE Function in Excel (Used with permission from Microsoft.)

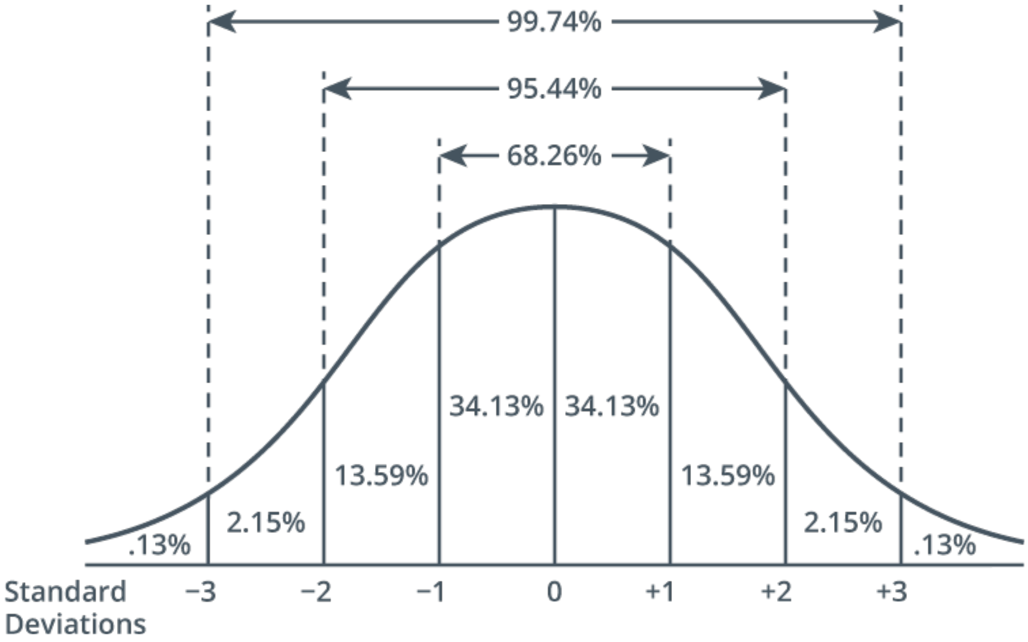
Distribution of a Data Set

It's important for the analyst to understand that, mathematically, you can have two vastly different data sets that have the same middle point. Therefore, knowing how all of the data is distributed makes a difference in your ability to describe the data set. Let's review some of the statistical measures discussed so far and consider how each can affect the distribution of data points:

- The mean, median, and mode involve analyzing the center of the data set.
- The range provides information about the outermost values.

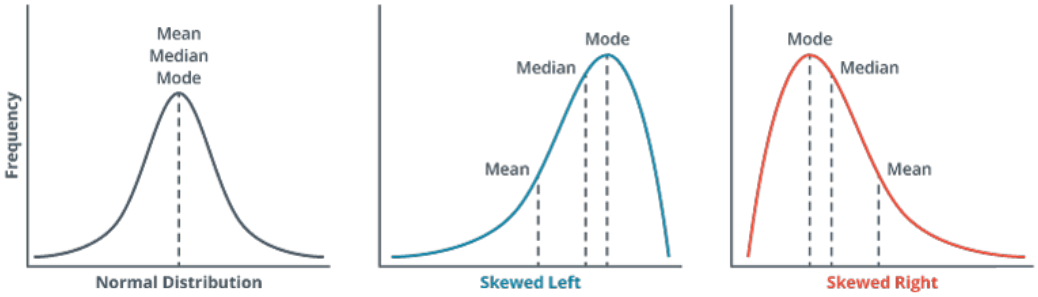
- Standard deviation shows how much of the data deviates from the middle point.
- The z-score is used to pinpoint how many standard deviations away a value lies from the mean.

Using these statistical measures, the analyst can visually depict the type of data distribution for a data set. Data with **normal distribution** follows a bell shape curve, with the mean being the middle and all other data following three points to the left or three points to the right of the mean. In normal distribution, the **empirical rule** refers to this tendency of most data points to fall within three points of the mean either on the positive side or the negative side of the curve.



Overview of the Normal Distribution Bell Curve

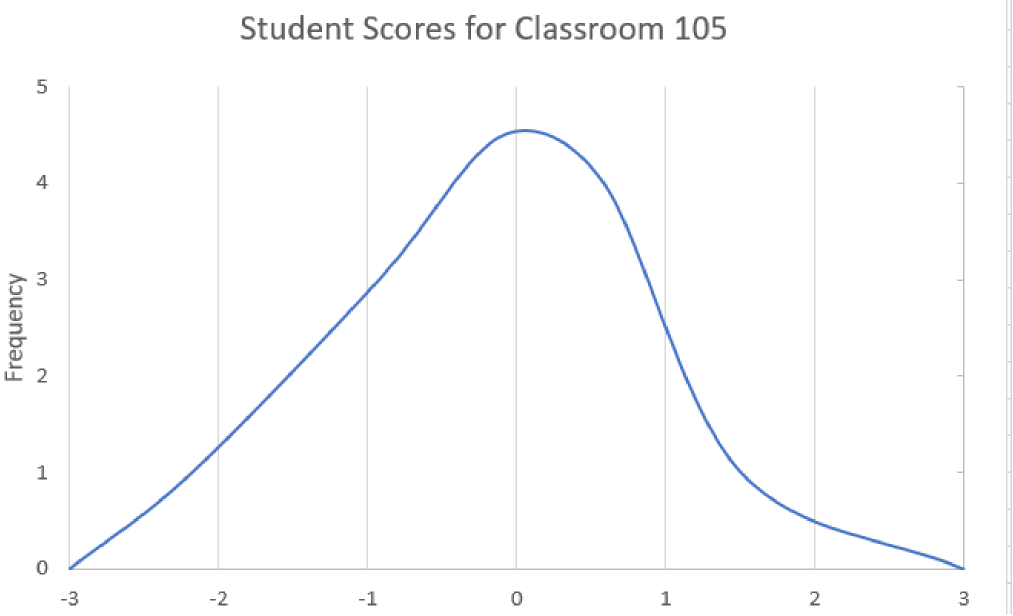
When the population of our data is within the normal distribution, meaning that 99.74% of our data falls within three standard deviations of the middle (mean), then our data can be further analyzed using tests that are designed for parametric data. **Parametric data** exists when the data set is within the rules of normal distribution. **Non-parametric data** exists when the data is not within the rules of normal distribution, with values that frequently deviate from the mean.



Example of Normal and Skewed Distribution

Looking at the images here, you should notice that the curve representing a normal distribution appears balanced, with both sides of the curve falling away from the middle of the data set equally. However, when looking at skewed data, the curve changes; more of the data falls on one side than the other.

When you're ready to further analyze your data, keep in mind that these tests are designed to work either with parametric or non-parametric data, not both. You must identify whether your data is within the rules of normal distribution, and use the appropriate test, or your results could be invalid.



Student Scores Mapped in a Simple Bell Curve Line Graph in Excel (Used with permission from Microsoft.)

Using the z-scores previously calculated for the students in classroom 105, we can plot the values to confirm whether the data is normally distributed. Looking at our distribution curve, we can see that all of the student scores appear within three points on both sides of the middle of the data. This data is not skewed left or right; it is normally distributed.

Review Activity:

Measures of Dispersion

Answer the following questions:

1. **What calculation involves finding the highest (maximum) value and the lowest (minimum) value?**

2. **What is the x-bar representative of in the calculation for standard deviation?**

3. **What is defined as following a bell shape curve, with the mean being the middle and all other data following three points to the left or three points to the right of the mean?**

4. **What is the tendency of most data points to fall within three points of the mean either on the positive side or the negative side of the curve?**

5. **How do you test whether data is following the empirical rule through visualization?**

Topic 9C

Use Frequency and Percentages

EXAM OBJECTIVES COVERED

3.1 Given a scenario, apply the appropriate descriptive statistical methods.

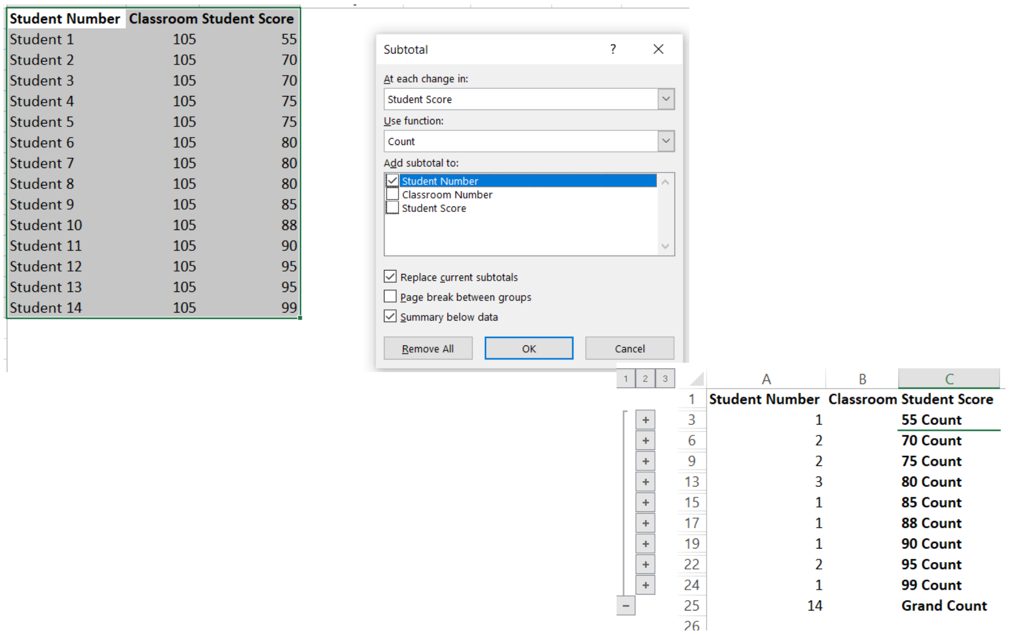
Frequency, or the number of times a data point occurs, is helpful when we need to describe large volumes of data. Percentage calculations, such as percentage difference and percentage change, can give insight into a process, a person's performance, or almost anything measured with quantitative data. Data can be expressed in several ways using percentages:

- We can show frequency as a percentage.
- Differences between two data sets can be expressed using percentage difference.
- The degree of change over time can be measured by calculating percentage change.

Here, we will learn how to calculate frequency and percentages, as well as how these measures can be useful in describing and analyzing data.

Frequency

Frequency is the number of times that a data point occurs within a data set. Calculating the frequency of any data point is as simple as grouping the data and counting it up. You can do this in several ways. One method is to use the SUBTOTAL command in Excel. Using the example of student data, this command will calculate the number of students in the classroom who achieved a certain score, as shown in the screenshot.



Student Number	Classroom	Student Score
Student 1	105	55
Student 2	105	70
Student 3	105	70
Student 4	105	75
Student 5	105	75
Student 6	105	80
Student 7	105	80
Student 8	105	80
Student 9	105	85
Student 10	105	88
Student 11	105	90
Student 12	105	95
Student 13	105	95
Student 14	105	99

Subtotal

At each change in: Student Score

Use function: Count

Add subtotal to:

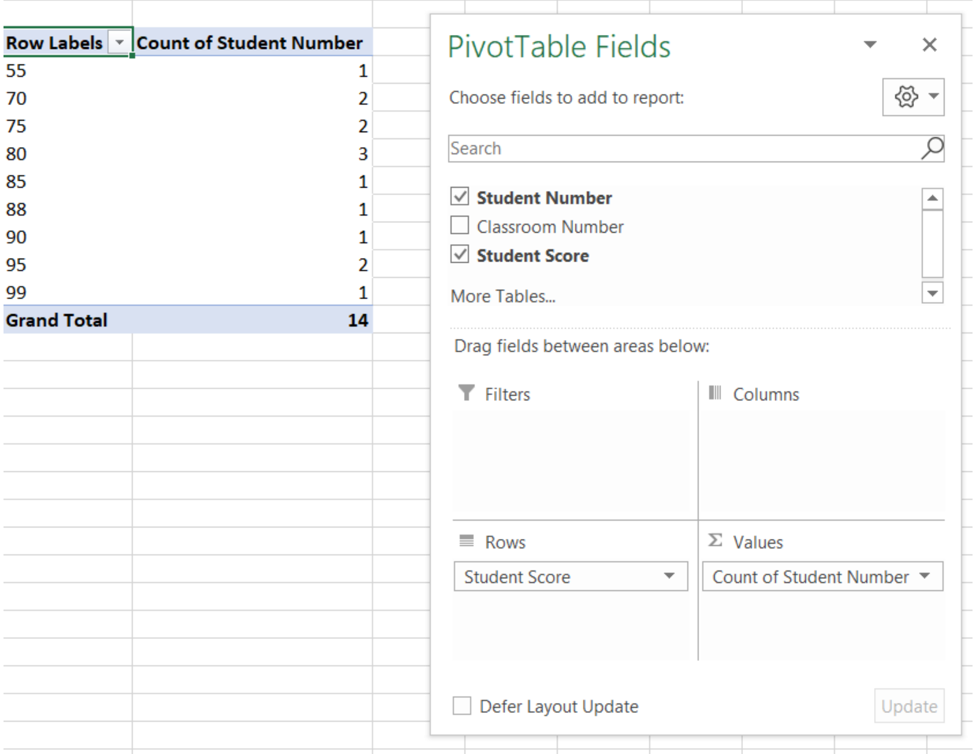
☑ Student Number
☐ Classroom Number
☐ Student Score

☑ Replace current subtotals
☐ Page break between groups
☑ Summary below data

	A	B	C
1	Student Number	Classroom	Student Score
3	1		**55 Count**
6	2		**70 Count**
9	2		**75 Count**
13	3		**80 Count**
15	1		**85 Count**
17	1		**88 Count**
19	1		**90 Count**
22	2		**95 Count**
24	1		**99 Count**
25	14		**Grand Count**
26			

Using the Microsoft Excel Subtotal Command (Used with permission from Microsoft.)

Pivot tables can be applied to data to create a data point and the cross section of *x* and *y*, meaning a row and column. You can also quickly use a pivot table in Excel to provide the frequency of any data in a group by leveraging just the row and value of the Pivot Table Fields and using the count. For example, to count the number of students who achieved each test score, you could design a simple pivot table with the following fields: student number and student score. As shown in the next screenshot, we create a row with the student score and then count the number of students with that score (value field).



Row Labels	Count of Student Number
55	1
70	2
75	2
80	3
85	1
88	1
90	1
95	2
99	1
Grand Total	**14**

PivotTable Fields

Choose fields to add to report:

Search

☑ **Student Number**
□ Classroom Number
☑ **Student Score**

More Tables...

Drag fields between areas below:

Filters

Columns

Rows: Student Score

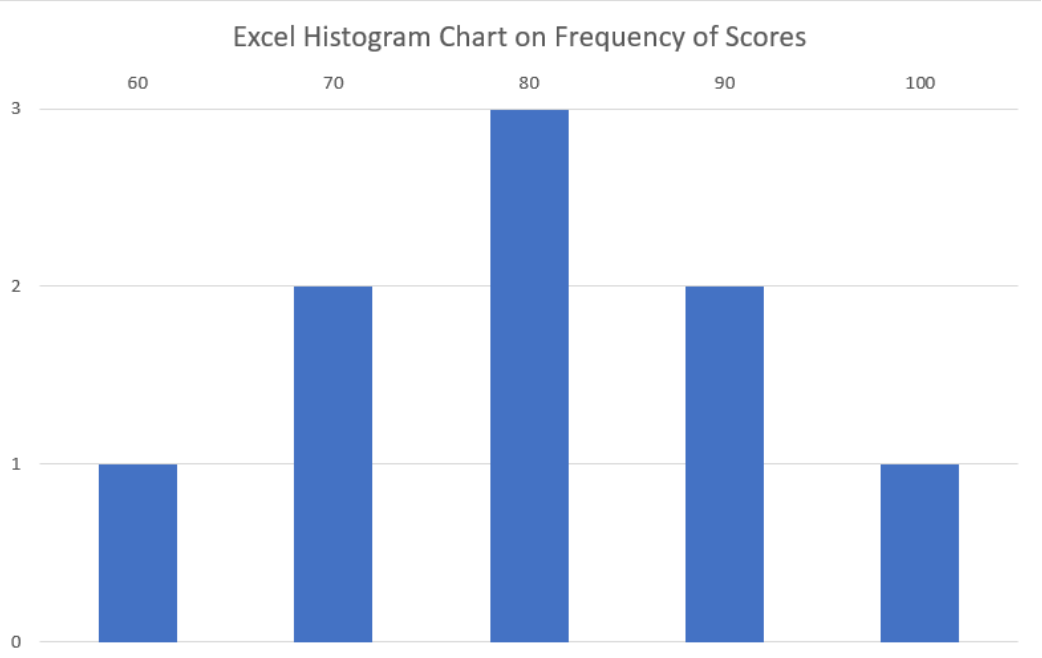
Values: Count of Student Number

□ Defer Layout Update

Creating a Pivot Table In Microsoft Excel to Show Frequency Data (Used with permission from Microsoft.)

The pivot table creates a simple frequency table that can easily be used for additional analysis, such as percentages and visuals, which makes the data easier to describe and understand than just a list of values.

Let's look at a visualization of the frequency table in a histogram, a column chart that shows an approximate distribution of data values. The histogram divides the range of values in a data set into a set of intervals (called bins). The visual shows the frequency of data points in each bin to provide a more meaningful picture of frequency data. The following screenshot shows a histogram to display how students scored on the test in this classroom.



Histogram in Excel Providing a Visual of Student Test Scores (Used with permission from Microsoft.)

Based on the frequency table of scores, we can display the data using the Excel chart for creating a histogram. In this example, Excel divided our data into 10-point bins to create the column visual. The shape of the histogram reveals information about the distribution of our data. Note that this histogram has a bell shape, meaning this data appears to be normally distributed.

Frequency can also be visualized as percentage values. For example, we can determine what percentage of children in our classroom received each score on the test.

Score	Frequency	Frequency %
55	1	7%
70	2	14%
75	2	14%
80	3	21%
85	1	7%
88	1	7%
90	1	7%
95	2	14%
99	1	7%

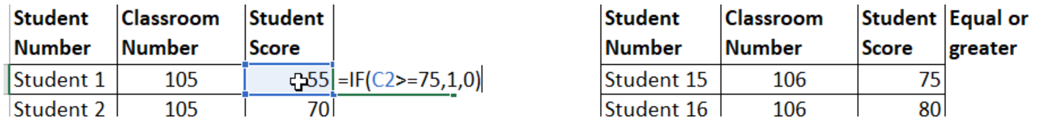
Displaying Frequency Data as Percentages

To calculate the percentages, you can either use the software functions provided by the tool you are using or do it manually. To manually calculate the percentages, divide the frequency for each score by the the total of all values in the frequency column (which is 14 in our example). For example, for score 55 the calculation would be 1/14 = 0.07, or 7 percent.

Percentage Difference

Percentages allow us to quantify and compare data in meaningful ways during analysis. The **percentage difference** is a calculation performed by determining the absolute value of the difference between two numbers, dividing by the average of the two values, and then multiplying by 100%. This displays the product as a percentage. The percentage difference can be a valuable way to determine the difference in percentage between two data points.

In our example of test scores, percentage differences can be helpful in comparing student performance in each classroom. For instance, suppose you want to compare the percentages of students who scored 75 or greater on the test in each classroom. You could calculate the percentage difference to complete this analysis. As a first step, determine the number of students who scored above 75 in both classrooms, as shown in the following screenshot.



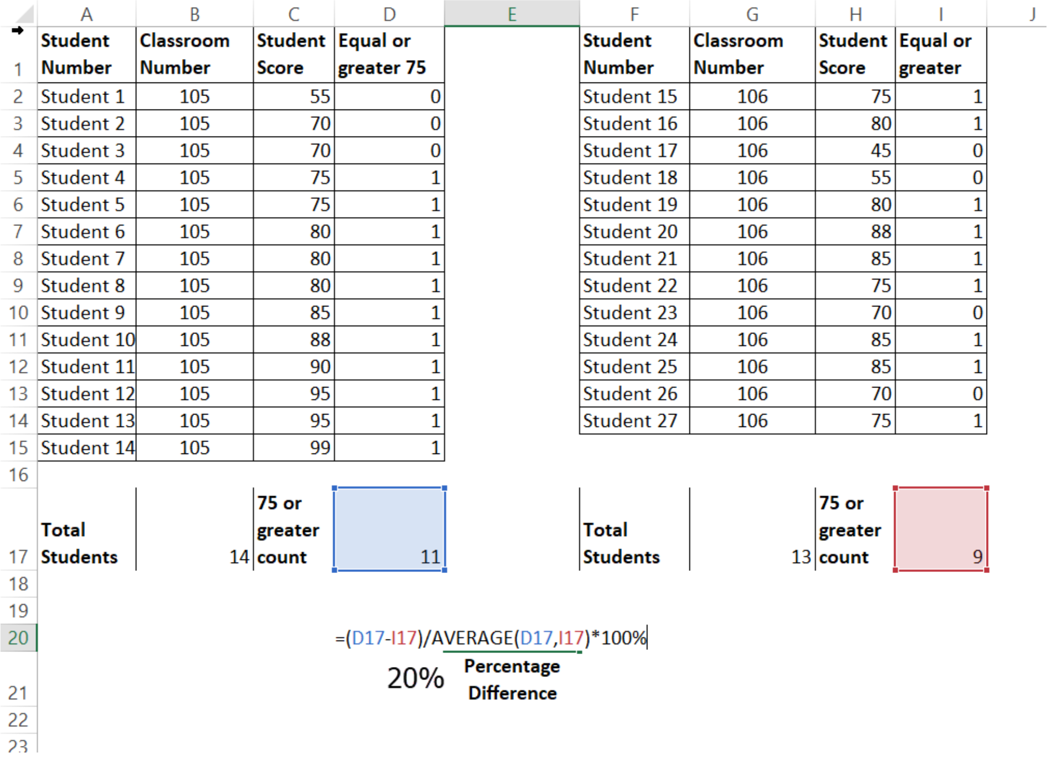
Student Number	Classroom Number	Student Score	
Student 1	105	55	=IF(C2>=75,1,0)
Student 2	105	70	

Student Number	Classroom Number	Student Score	Equal or greater
Student 15	106	75	
Student 16	106	80	

Logical Function in Excel to Determine Students with Scores of 75 or Greater (Used with permission from Microsoft.)

You can write a logical function to easily calculate which students scored 75 or greater. In Excel, an IF function will return a 1 if the score is 75 or greater and a 0 if the score is below 75: =IF(Student>=75, 1, 0).

After executing the IF function, use an aggregate function, in this case Count, to total the number of students who scored 75 or greater in each classroom. You now have all the information needed to determine the percentage difference.



	A	B	C	D	E	F	G	H	I	J
1	Student Number	Classroom Number	Student Score	Equal or greater 75		Student Number	Classroom Number	Student Score	Equal or greater	
2	Student 1	105	55	0		Student 15	106	75	1	
3	Student 2	105	70	0		Student 16	106	80	1	
4	Student 3	105	70	0		Student 17	106	45	0	
5	Student 4	105	75	1		Student 18	106	55	0	
6	Student 5	105	75	1		Student 19	106	80	1	
7	Student 6	105	80	1		Student 20	106	88	1	
8	Student 7	105	80	1		Student 21	106	85	1	
9	Student 8	105	80	1		Student 22	106	75	1	
10	Student 9	105	85	1		Student 23	106	70	0	
11	Student 10	105	88	1		Student 24	106	85	1	
12	Student 11	105	90	1		Student 25	106	85	1	
13	Student 12	105	95	1		Student 26	106	70	0	
14	Student 13	105	95	1		Student 27	106	75	1	
15	Student 14	105	99	1						
16										
17	Total Students	14	75 or greater count	11		Total Students	13	75 or greater count	9	
18										
19										
20				=(D17-I17)/AVERAGE(D17,I17)*100%						
21				20%	Percentage Difference					
22										
23										

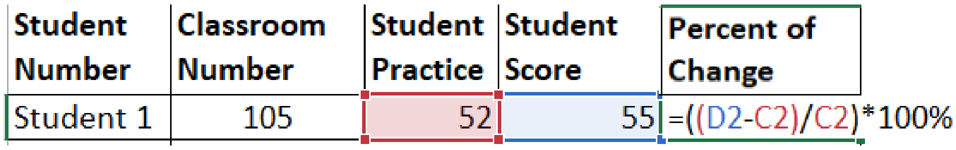
Excel Calculation of the Percentage Difference Between Classrooms for Students Who Scored 75 or Greater (Used with permission from Microsoft.)

The formula for percentage difference determines the absolute value of the difference between the students who scored 75 or greater in each classroom (11 – 9 = 2), divides by the average of the two values (10), and then multiplies by 100%. In other words, 2/10 x 100% = 20%. There is a 20% difference between classroom 105 and classroom 106 for scores that are 75 or greater.

Percentage Change

Percentage change represents the difference between a new value and an original value (either the last value or an older value), and it can be an increase or a decrease. The percentage change calculation helps us determine how values change over time. The percentage change calculation can be used in a number of scenarios, such as analyzing price increases or decreases, showing a salesperson's performance month by month, or following a child's development by capturing changes in height and weight each year. The formula for percentage change involves subtracting the newest value from the last value, dividing that number by the last value, and then multiplying by 100%.

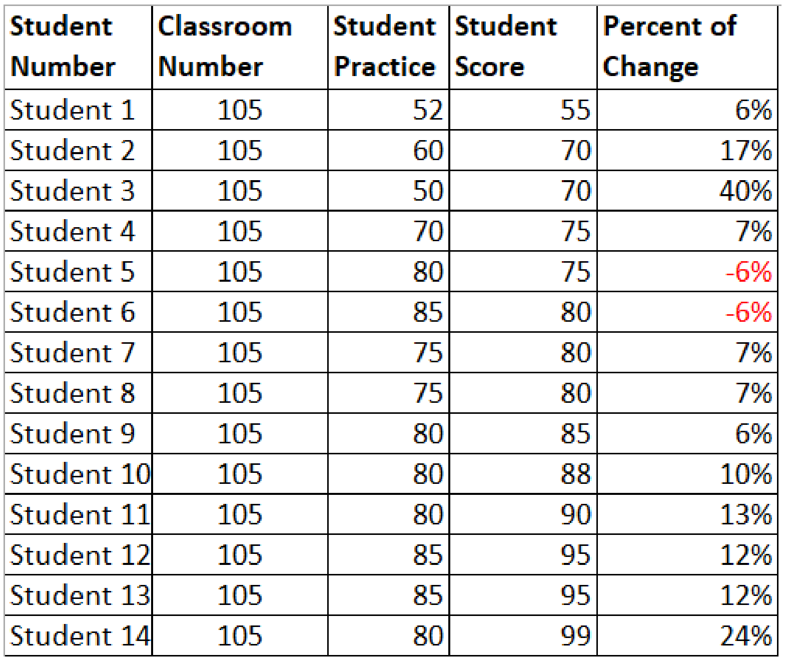
To demonstrate, let's calculate the percentage change of student test scores in classroom 105. Suppose that the students took a practice test at the beginning of the semester, and we'd like to calculate the percentage change between practice test scores and actual test scores on a student-by-student basis.



Student Number	Classroom Number	Student Practice	Student Score	Percent of Change
Student 1	105	52	55	=((D2-C2)/C2)*100%

Excel Calculation for Percentage Change (Used with permission from Microsoft.)

This calculation can be performed for each student in the class to determine whether there was a positive increase or negative decrease between the practice test and actual test scores.



Student Number	Classroom Number	Student Practice	Student Score	Percent of Change
Student 1	105	52	55	6%
Student 2	105	60	70	17%
Student 3	105	50	70	40%
Student 4	105	70	75	7%
Student 5	105	80	75	-6%
Student 6	105	85	80	-6%
Student 7	105	75	80	7%
Student 8	105	75	80	7%
Student 9	105	80	85	6%
Student 10	105	80	88	10%
Student 11	105	80	90	13%
Student 12	105	85	95	12%
Student 13	105	85	95	12%
Student 14	105	80	99	24%

Percentage Change for Each Student

As shown by the percentage change calculations, all but two students saw their performance increase from the practice test to the actual test.

Review Activity:

Frequency and Percentages

Answer the following questions:

1. **Differences between two data sets could be expressed using what calculation?**

2. **What are two aspects of calculating frequency?**

3. **A pivot table, histogram, and percentages are three ways of displaying what?**

4. **Which calculation involves subtracting the newest value from the last value, dividing that number by the last value, and then multiplying by 100%?**

Lesson 9

Summary

After this lesson, you should be able to explain the measures of central tendency used to find the center of the data and the measures of dispersion used to determine how much variation exists in the data. You should understand the major differences between mean, mode, and median: the mean gives you the average of a set of numbers, the median is the middle number within a group of sorted numbers, and the mode is the number that shows up most often in the data set. You should be familiar with aggregate functions that can help you identify the range of data, such as Max and Min, and should be able to use standard deviation to determine how close or spread out the data is. You should also be able to calculate frequency and percentages to describe large volumes of data.

Guidelines in Using Frequency and Percentages

Consider these best practices and guidelines when familiarizing yourself with the calculations for frequency, percentage difference, and percentage change.

1. While mean, median, and mode are all measures of central tendency, each measures something different. They are not interchangeable, as each gives you different information, especially when outliers exist. You will use each measure at some point in your work.

2. Standard deviation gives a more accurate depiction of the average distance of the values from the mean, providing a better understanding of the variability within a data set. A low standard deviation indicates that the values are closer to the mean, whereas a high standard deviation tells you the values are more spread out, and thus farther from the mean.

3. When plotting distribution, the shape of the curve will tell us whether the data is normally distributed or not. Normal distribution consists of a bell-shaped curve that follows the empirical rule (most data points fall within three points of the mean either on the positive side or the negative side of the curve).

4. If your distribution curve indicates normal distribution, you know your data is parametric. If the curve is skewed right or left, your data is non-parametric.

5. Calculating the frequency of any data point is as simple as grouping the data and counting it up. To gain further insight, we can use percentages to show what the frequency numbers mean.

6. Percentage difference is a valuable way to determine the difference in percentage between two data points (or data sets), while percentage change represents the difference between a new value and an original value (change over time).

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 10

Describing Key Analysis Techniques

LESSON INTRODUCTION

When you are preparing to work with data, you must first determine what you are researching, where the data comes from, and what types of analysis you wish to perform. You will likely approach all data sets in this same manner, regardless of which type of analysis you end up conducting. Analysis can range from exploratory, which is performed at a high level and applicable to all data sets, to other types that are more specific to the goal and outcome of the research. Thus, it is important that you understand how to create a research question and identify the source of data before moving on to the actual analysis.

Lesson Objectives

In this lesson, you will do the following:

- Get started with analysis.
- Recognize types of analysis.

Topic 10A

Get Started with Analysis

EXAM OBJECTIVES COVERED
3.3 Summarize types of analysis and key analysis techniques

All research begins with a question. As data analysts, our goal is to answer these questions with the data that is available. In order to do this, we must first understand how to develop appropriate research questions and assess whether our data set is of high quality with enough observations to research.

Research Questions

When you are preparing to research a topic, one of the first steps you must take is to determine your **research questions**. If you are working on a student project, you may be able to collect your own thoughts and form your own questions. If you are working for an organization, or within a team, these questions may be created by others for you, which you must then work to analyze.

Research questions that are too broad can be challenging to answer. You should attempt to focus them down to a true or false statement. These true or false statements make up a key component of all analysis—the **hypothesis statement**. It's what you believe to be true, and what you will test with analysis techniques to show that it is true (or maybe false). You will learn more about hypotheses as you discover different types of analysis techniques.

Let's walk through an example. Consider student test scores as a subject for our research. We believe (or hypothesize) that students would perform better on the test if they were given extra study hours. Students were broken naturally into two groups due to the fact that they were in different classrooms. We want to know whether the students in the classroom with extra study hours performed better than the students who did not have those extra study hours. Our research question would be, "Did the students in the classroom with extra study hours perform better than students without?" The answer to this question is either yes, they did, or no, they did not.

Data Sources and Collection Methods

After you have narrowed down your research questions, you will need to dive into the data and the methods of collection used. Ask yourself the following:

- What is the source of the data?
- Did this data come from a source system?
- Was this data collected through a polling system, or did someone ask questions in a one-on-one interview?

This step is important because different collection methods have different considerations and can potentially result in vastly different qualities of data.

Let's return to our student testing example. Suppose that you are given the option to choose the source of your data between a dedicated software for student testing or from a teacher's memory, two years later. The student scores stored in the software were automatically recorded after the tests were taken; this data was then exported for each student in each classroom into a spreadsheet. The data from the teacher's memory is collected via interview; the teacher recalls the list of students and their test scores, and you record that data. Which data set do you believe would have the highest quality data?

This likely seems like a silly exercise, with an obvious answer, as it's not hard to imagine that the software is more reliable than memory. However, this easily points out why the source and collection method of your data is important when considering your analysis. You can't control all of the data when it is captured, and you can't change the process of collection after it is collected. The best you can do is note how that data was collected, so that people will know the source and collection method.

Observations

After you have identified your research question and data source, you must next determine whether the amount of data in hand is enough to adequately answer the research question. While you can work with small sample sizes and still create very legitimate research results, you must be careful to note the amount of data you are using in your research. This is what's known as the **“n” count**, and it's always disclosed in reputable research. You may see it written in this way: “In a recent study of 27 students in the eighth grade . . .”

Let's return to our example of studying whether or not children benefited from extra study hours. In this scenario, each student is an observation.

Review Activity:

Get Started with Analysis

Answer the following questions:

1. **In which manner should a research question be posed, and why?**

2. **Why is it important to determine the sources and collection methods for the data you're working with?**

Topic 10B

Recognize Types of Analysis

EXAM OBJECTIVES COVERED
3.3 Summarize types of analysis and key analysis techniques

There is no single answer for the type of analysis you might require when performing your job as a data analyst. However, there are some types you are more likely to encounter than others. You will likely always perform some level of exploratory analysis, and at other times perform a mix of others, like performance, trend, and link analysis. The type of analysis you perform is based on the data you have on hand and the overall goals of the analysis. In order to know which type of analysis is best suited for your needs, you must first understand what the different types of analysis are and why we use them.

Exploratory Analysis

As a data analyst, **exploratory analysis** is something you should be prepared to perform on literally every data set you encounter. The point of exploratory analysis is to determine the main characteristic of the data set. For example, exploratory analysis can help you determine how much of your data is qualitative (categorical) versus quantitative (countable/measurable). This identification helps inform what type of data cleaning and/or transformation is required. Exploratory analysis provides you with up-front information about the data and is also the easiest type of analysis to visualize.

There are many other key benefits of exploratory analysis.

- Informs what type of statistical testing is appropriate.
- Helps you to investigate potential relationships between data early on.
- Allows you to identify correlation between data points.
- Informs whether your research questions can be potentially answered with the data.

Let's go back to our earlier example of student test scores and study hours. In order to answer your research question, you will need to know how every student performed, and how each classroom performed in comparison. At the very beginning of your research, you would need to identify whether any students scored very high or very low, as these scores will impact your analysis. Exploratory analysis can provide you with this information.

All data should be explored through exploratory analysis before it is analyzed any further. This is not only a necessary process of the analyst to learn the data, but also a mathematical process that has been in use since the 1970s, when John Tukey encouraged statisticians to use the method to help determine better methods of data collection and discover potential research questions. Tukey recognized that the process helped researchers formulate hypotheses and ultimately led to a better understanding of how to deal with missing data and transformations needed for the variables being studied.

Performance Analysis

Performance analysis is analysis that is done to measure the performance of a particular product, outcome, or scenario against a defined objective. Performance analysis uses qualitative and quantitative data. Performance analysis data can also be considered non-parametric, meaning that it doesn't follow the normal distribution curve and thus is not able to be analyzed by all types of testing.

Performance analysis is frequently used to test how a particular software program performs to inform optimization of the software. In business, performance analysis is used in combination with metrics like **key performance indicators (KPIs)**. KPIs are measurements/goals that are established to help identify whether a business is achieving its objectives. KPIs can show the overall health of products, work processes, or even sales goals for new salespeople.

If we were to conduct a performance analysis on our student scores, we might start by setting a KPI, such as an expectation that students achieve at least three points higher than the average when they have had the extra study time. Then we would look at their actual scores and measure how many are over the three-point average.

KPIs vary because organizations and their goals vary. Understanding how to develop them for the use of these measures in your analysis means that you understand the business, you understand their goals, and you have a good understanding of the data. The overall goal is to determine if the data for that scenario is meeting the defined targets.

Gap Analysis

Gap analysis is the study of a present state, desired state, and the gaps between the two. After identifying these gaps, you can then develop a project (or projects) that help you get to the desired state. When you define how you will accomplish the desired state through the development of a project, you are creating a **scope**, which includes measurable tasks that are needed to meet the desired end state.

For example, suppose a company has a goal to achieve $100,000 in total revenue for a product. The company must determine the present or current state and then compare it to the goal to identify the gap between those values. The product is a bike that has brought in $75,000 since its debut one year ago. The difference between the current state, $75,000, and the desired state, $100,000, is a gap of $25,000. To earn the extra $25,000, the company decides to redesign the bike pedals and give the bike a new color scheme: this is the scope.

Gap analysis is also commonly used for project management efforts on software development projects. Suppose a company wants to design a new piece of software to replace the paperwork that is currently completed during deliveries. The current state is that the staff delivery team writes out a paper delivery receipt and requires customers to sign the receipt upon delivery of their order. The desired state is that the delivery staff will use a tablet and request digital signatures. The software development team will study the current state and their knowledge of the desired state, and focus on the gaps to deliver the proper estimate and scope. In the world of software development, or really any development project, **scope creep** can occur when the scope changes from the original plan and incurs adjustments. These adjustments can cause issues in meeting the various projects, causing a gap in reaching the desired state on time and within budget.

Trend Analysis

Trend analysis is defined by measuring a trend on historical data to predict a future outcome. You will typically find trend analysis is used in all industries to determine how something is performing over time or predicting issues with a subject. A subject is a process, person, or product.

- We can use trend analysis for product or market research, strategic initiatives, future outcomes, and financial trends.
- Geographic trends are based on geographic locations.
- Short-term, or temporal, trends occur for a period of time, but not in the long term. For example, if it was predicted that there was going to be a shortage of medical supplies, like gloves and face masks, we might see an increase in purchases of those products for a short period of time.
- Time-related trends occur over a specified period of time. For example, an increase in toy purchases in the months leading up to Christmas is a time-related trend.

Google Trends is a tool that tracks key words used in internet searches to show how these key words are trending in the Google Search engine. The tool allows you to key in a subject or text phrase to search over various points of time, or in a particular geography. It then shows you a visual of how that word or phrase is performing in searches.

Trend analysis is also commonly used to forecast future values. If you manage data for an investment firm, you will certainly need to perform trend analysis for a specified period of time for businesses in the portfolio. If you see that a company is trending up or down, that information can inform the investment organization on what they might provide to the company (whether help to recover from a down trend or support to amplify an upward trend).

In the case of our student scores, when trying to determine whether extra study hours have boosted scores, we may want to look at the score performance over the last three years. This information can help us determine whether the two classrooms studied were on track with where they were expected to be in the future, or if they were already trending up before the extra study hours were implemented. This is important because if we discover student scores were already improving on a three-year trend, then any upward momentum might have nothing to do with the extra study hours. It may sound counterintuitive to want to disprove a hypothesis, but when we do, we can come up with new questions. For example, we could next pose the idea that the scores are up because the teacher is spending more time in the classroom.

Link Analysis

Link analysis is used to determine how a single data point links to other data points, focusing on the relationships and connections in a database. When you determine how data is linked from a primary key to a foreign key, you are performing a type of link analysis in the database.

Link analysis is often used in advertising, particularly for social media networks. For example, suppose you have a relationship with a social media group, and members of that group start to purchase an item from a particular online retailer. Your association with the group of purchasers indicates to the retailer that you may also be interested in their products. The end result is that you could see advertising in your social feeds to purchase similar products from the retailer.

When a network of items exists, link analysis can occur. Consider six degrees of separation, which posits that all people are six or fewer acquaintances away from each other. This idea that you know someone who knows someone, who knows someone else, is essentially the concept behind link analysis.

Link analysis has three main components: a network, a node, and a link. The network is a set of nodes and links. The node is a single point (person, account, product). The link itself is the relationship between the different nodes.

Let’s relate this back to our student test scores example. Imagine that the study hours are spent in a software with a social component (network). Each student is a point (node). Suppose that some of our students performed well on the test, yet others in the extra-hours classroom did not. We might want to explore that data further by viewing the connections between the students (links). Did the students leverage the social components of the software to discuss difficult topics amongst each other? If we found a relationship between student engagement within the network and higher score performance, we would want to explore that relationship further.

Link analysis is all about relationships among the data points and can provide some very interesting insights when a network component exists within the data.

Review Activity:

Types of Analysis

Answer the following questions:

1. **Profiling data is similar to what type of analysis?**

2. **Which type of analysis is conducted on non-parametric data, or data that does not follow the normal distribution curve?**

3. **Performance analysis measures the performance of a particular product, outcome, or scenario against what?**

4. **Which metric establishes goals to help identify whether a business is achieving its objectives?**

5. **Which type of analysis studies the difference between the desired state and current state?**

6. **Trend analysis is commonly used for what purpose?**

7. **What are the three main components of link analysis?**

Lesson 10

Summary

After this lesson, you should have a better understanding of what needs to be done before you begin data analysis, and what types of analysis you will use throughout your work. You should understand the importance of identifying a research question, data source, and the amount of data you'll use in your research.

You should also have a firm understanding of the varying types of analysis. Major types of analysis you will encounter and should be prepared to use include exploratory analysis, performance analysis, scope and gap analysis, trend analysis, and link analysis.

Guidelines in Describing Key Analysis Techniques

Consider these best practices and guidelines when creating research questions and recognizing the varying techniques that are used to analyze data.

1. The first step when preparing to analyze data should always be to develop a research question. Research questions should be posed as a "yes or no" statement, so they can be answered during hypothesis testing.
2. You must also determine the source and collection method of data, as this can affect the quality of your data.
3. You must then determine whether the amount of data in hand is enough to adequately answer the research question. The amount of data you use in your research is the "n" count, and it's always disclosed in reputable research.
4. You must be familiar with the purposes of different types of analysis, so you can choose the appropriate method for your research.
    - The point of exploratory analysis is to determine the main characteristics of the data set.
    - Performance analysis measures the performance of a particular product, outcome, or scenario against a defined objective.
    - Gap analysis measures the difference between the desired state and present state and involves determining the scope, or tasks needed to reach the desired state.
    - Trend analysis measures a trend on historical data to determine performance over time and predict a future outcome.
    - Link analysis helps us determine how a single data point links to other data points, and it can be used to focus on relationships and connections in a database.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 11

Understanding the Use of Different Statistical Methods

LESSON INTRODUCTION

Inferential statistics involves reaching conclusions based on evidence and reasoning using data, and includes methods such as confidence intervals, t-tests, and hypothesis testing. Using inferential statistics on both large and small sample sizes allows us to create findings for just about anything that we can gather data on and measure. With inferential statistics, we can analyze data gathered about programs, people, things, and even interventions. This differs from descriptive analysis, which just aims to describe the data we are using to draw conclusions. Inferential statistics also allow us to utilize smaller samples to represent a larger population. We use different statistical tests to determine the statistical significance of our findings, leveraging p-values.

Lesson Objectives

In this lesson, you will do the following:

- Understand the importance of statistical tests.
- Break down the hypothesis test.
- Understand tests and methods to determine relationships between variables.

Topic 11A

Understand the Importance of Statistical Tests

EXAM OBJECTIVES COVERED

3.2 Explain the purpose of inferential statistical methods.

When we are working with a sample of data (with or without the entire population), we use statistical methods and processes to make assessments, or judgments, on different outcomes. **Population** doesn't mean the population of the United States; it refers to the population of a group of records that meet a certain criterion. Consider our example of student test scores. We have a population of records for test scores in two different classrooms of students. We can compare these classrooms to each other and use the data to interpret the scores. Our findings can then inform policy or performance measures. When we are interpreting data, we must choose the appropriate statistical test for the data that we have to ensure accurate results. You may not be performing these tests yourself; this might be the role of a more senior statistician. Regardless, it is good to have a basic understanding of these tests and concepts.

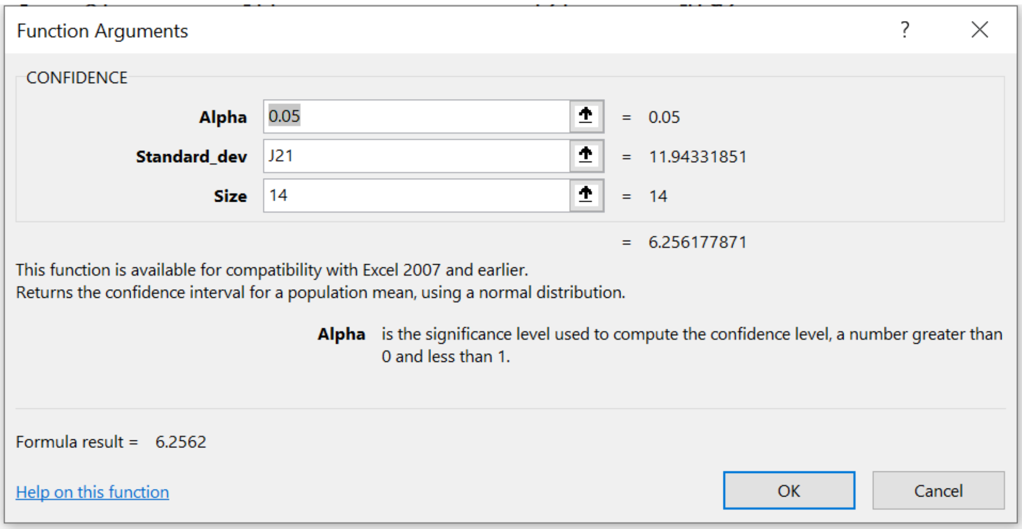
Confidence Intervals

A **confidence interval** is a calculation of values that describes the certainty or uncertainty of an estimate made on the analysis. It allows us to identify how confident we are that our estimate represents the sample or population on which our analysis focuses.

Confidence intervals can be subjective, and there is no set percentage that indicates a good level of confidence. However, most analysts strive for a 95% confidence interval, which means we are 95% confident in the results of our analysis.

How to Calculate a Confidence Interval

We can use a function in Excel to determine the confidence interval for our sample. First, we will decide which confidence level we want to use—in most cases, this will be 95%. Then, we must identify the number of records, known as the n value, in our sample; in this example, that number is 14. Next, calculate and input the standard deviation.



Function Arguments

CONFIDENCE

Alpha 0.05 = 0.05

Standard_dev J21 = 11.94331851

Size 14 = 14

= 6.256177871

This function is available for compatibility with Excel 2007 and earlier.
Returns the confidence interval for a population mean, using a normal distribution.

Alpha is the significance level used to compute the confidence level, a number greater than 0 and less than 1.

Formula result = 6.2562

CONFIDENCE Function in Excel (Used with permission from Microsoft.)

The result of our CONFIDENCE function is 6.2562, and our mean for this set of data is 81. To create an upper bound, we add the 6.2562 to the mean of 81. To get the lower bound, we deduct the 6.2562 from the mean. Thus, in our small sample, we would have an upper bound of 87.25 and a lower bound of 74.74. If this was written into our paper or presentation, we would say: "if this sample was true of the entire population of that school in that classroom, we believe that 95% of all the students would fall between a test score of 74.74 and 87.25."

T-Tests and P-Values

T-Tests

A **t-test** is used when comparing two groups to determine if there is a significant difference between the means of both groups. We know from our descriptive analysis that the dependent variable is normally distributed.

There are two important variables when conducting a t-test: the **dependent variable**, which is what we are measuring, and the **independent variable** that is different between the groups. The dependent variable is our main data point, and we use it to determine the mean, median, mode, and standard deviation. In our student test score example, the test scores are our dependent variables. The classroom is our independent variable. Students in classroom 105 received five extra study hours, and they have a higher mean than the students in classroom 106. When we want to test whether those five extra study hours made a significant difference in the scores over the classroom that did not get the extra study hours, we use a t-test.

Statistical Significance and the P-Value

In our student test scores example, the higher mean value in classroom 105 shows us that the students in that classroom outperform their peers in the other classroom. But while it's a true statement to say that they had higher scores, that doesn't make it statistically significant. For **statistical significance** to exist, it cannot have happened by chance and must be a significant difference.

When the data is normally distributed, we can use a t-test with the two different means in the sample to determine the probability that the means are different due to chance. The value of significance is referenced by the **p-value**. The "p" in p-value stands for probability, as the p-value tells us the probability of the occurrence of an outcome. A high p-value indicates that the outcome is likely not repeatable and thus not significant, whereas a low p-value tells us the event is likely repeatable, meaning it did not occur due to chance and is thus statistically significant.

Most studies desire the p-value to be less than .05 ($p > .05$), as this signals that the difference between the two groups is significant and not just due to chance. For most studies, if your p-value is greater than 5%, then it is not a statistically significant difference. While a p-value that is equal to .05 is a typical level of significance, you will often see values ranging from .01 to .10.

We mentioned previously that statistical tests are important because they can determine changes in policy. Returning to our student test score example, the p-value is important because if the data is statistically significant, the school will likely want to ensure that all classrooms get that five extra hours of study time in order to improve outcomes for all students.

Review Activity:

The Importance of Statistical Tests

Answer the following questions:

1. **The calculation of values that describes the certainty or uncertainty of an estimate made on the analysis is known as what?**

2. **What is the percentage of a confidence interval that is most commonly strived for in analysis?**

3. **What are the two variables we use when conducting a t-test?**

4. **Which two conditions must be met for data to be considered statistically significant?**

5. **What does the "p" in p-value stand for?**

Topic 11B

Break Down the Hypothesis Test

EXAM OBJECTIVES COVERED
3.2 Explain the purpose of inferential statistical methods.

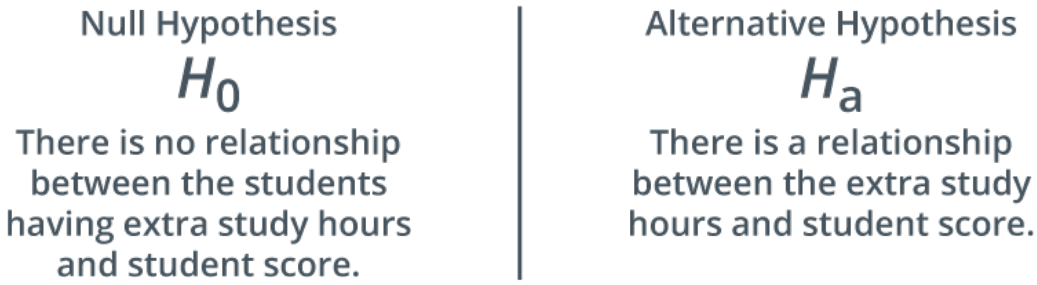
As we've previously discussed, when analyzing data our research question is converted into a hypothesis, which is a true or false statement that states we either believe a relationship exists or we do not believe a relationship exists. In order to test that hypothesis, we use hypothesis testing to either accept or reject the statement. This means we accept either the null hypothesis or the alternative hypothesis. We will also discuss the types of errors, type I and type II, that occur with hypothesis testing and the impact of rejecting the wrong hypothesis.

The Null Hypothesis

When you convert your research question to a positive statement, you create an alternative hypothesis. The **alternative hypothesis** assumes that a relationship between two variables *does exist*. When you convert your question to a false or negative statement, you create a null hypothesis. The **null hypothesis** assumes that a relationship between two variables *does not exist*.

In our student test scores example, examples of these hypotheses would be as follows:

- Alternative hypothesis: There is a relationship between the five extra study hours given to students in classroom 105 and their higher scores. (In essence, the extra study hours contributed to the higher scores.)
- Null hypothesis: There is no relationship between the five extra study hours and the higher scores in classroom 105. (In essence, the extra study hours made no difference at all, for either classroom.)



Null Hypothesis

H_0

There is no relationship between the students having extra study hours and student score.

Alternative Hypothesis

H_a

There is a relationship between the extra study hours and student score.

Image of the Null and Alternative Hypothesis

Understanding the Results of Hypothesis Testing

In our student test scores example, we are hoping that the investment of five extra study hours made a difference statistically between the two groups. Thus, we want to disprove the null hypothesis and prove the alternative hypothesis to be true. However, we need to be careful when attempting to reject the null hypothesis so that we don't get ahead of ourselves and reject it in error. There are two types of errors that can occur when conducting hypothesis testing.

- A **type I error** occurs when we reject the correct hypothesis and accept the incorrect hypothesis. In essence, we inadvertently create a false positive. Suppose we made a mistake in our data analysis and found that the five hours of extra study was statistically significant, but it actually was not. In this instance, we have committed a type I error.
- A **type II error** occurs when we accept the incorrect hypothesis and reject the correct hypothesis. In essence, we create a false negative. Suppose we found that the five hours of extra study made a statistically significant difference, and we should have thus accepted the alternative hypothesis. But if we incorrectly reject the alternative hypothesis, and instead accept the null hypothesis and claim the extra study made no difference, we have committed a type II error.

There can be massive consequences for creating false positives and false negatives. For example, if your organization does medical testing on critical diseases, and your test returns false positives, this means people who are not sick will receive unnecessary (and potentially harmful) treatment. The reverse could be even worse: the creation of a false negative leads to people who are actually sick not receiving treatment, due to the test returning an incorrect conclusion that they are not sick.

Review Activity:

The Hypothesis Test

Answer the following questions:

1. **Which type of hypothesis assumes that a relationship between two variables does exist?**

2. **Which type of hypothesis assumes that a relationship between two variables does not exist?**

3. **Which type of error creates a false negative?**

4. **What are some of the impacts of type I and type II errors on hypothesis testing? Use an example.**

Topic 11C

Understand Tests and Methods to Determine Relationships Between Variables

EXAM OBJECTIVES COVERED

3.2 Explain the purpose of inferential statistical methods.

When working with data, it is often important to find the relationships between the data points. We can use various statistical methods or tests to determine whether there is or isn't a relationship between the values. Often, in the search of these relationships we find ways to improve processes or outcomes with minimal changes needed. Using our data and the results of these statistical tests, we can produce findings that prove or disprove the stated hypothesis. It's not enough to just produce data and counts; we need to use that information to look for correlation between variables.

Chi-Square

A **chi-square statistic** compares the size of the difference between the expected result and the actual result. It is used to measure how the model compares to the actual data.

A **chi-square test**, which produces the chi-square statistic, can be used to determine if a difference exists between groups. The test can be used to compare all sorts of variables: scores on a test, responses to a medical test, answers to a survey question, and more.

Chi-square is pronounced as \ 'kī-' skwer \, and it rhymes with the word "sky."

There are a few reasons why we would choose to conduct chi-square analysis.

- A chi-square test is used to compare actual results to what we expected the results would be.
- A chi-square test also allows us to rule out that the observations happened by chance.
- Chi-square testing identifies how confident we are that the results are (or are not) different from what we expected and that there is a relationship between the variables.
- Chi-square testing is useful when we are analyzing data from a random sample and working with a categorical variable, like education, race, or gender.

Let's return to our student test score example. Suppose the vendor that is administering the extra study hours has created a survey for students to complete after they have finished the extra study, asking them to identify how prepared they are for the test. The resulting data set of responses is matched with the students' actual test scores.

Student Preparedness	Fail	Pass	Total
Very Prepared	9	17	**26**
Somewhat Prepared	11	40	**51**
Not Prepared	12	11	**23**
Total	**32**	**68**	**100**

Data Prepared for Chi-Square Test of Independence

In our example, we categorize student scores by Fail and Pass. Then we aggregate the student data on their indication of preparedness and score category. We can use the chi-square test to determine whether their test scores appear to be related to the level of preparedness they had previously identified.

Chi-Square Tests

Two common types of chi-square tests are the **test of independence** and **goodness of fit**. Both tests compare actual results against an expectation, but goodness of fit is used for the distribution of the data and the test of independence is used for the relationship between the variables.

- Test of independence: Tests against multiple variables.
- Goodness of fit: Tests against a single variable.

Test of Independence

Continuing with our test scores example, let's say we want to determine whether our test takers scored a pass or fail on the test, accounting for how prepared they felt before taking the test. We assume that if they felt more prepared, they were less likely to fail the test. Suppose we have 100 students in total. Of those 100 students, 32 of them had a failing grade and 68 earned a passing grade. The hypothesis is that the more prepared the student felt, the more likely they were to pass the test. The chi-square test of independence could be used to see if there is a significant relationship between preparation and passing scores.

There are several ways to perform a chi-square test. You can use manual functions in Excel, statistical analysis software (like Stata and SPSS), or even online chi-square calculators.

Student Preparedness	Fail	Pass	Total
Very Prepared	9 (8.32) [0.06]	17 (17.68) [0.03]	**26**
Somewhat Prepared	11 (16.32) [1.73]	40 (34.68) [0.82]	**51**
Not Prepared	12 (7.36) [2.93]	11 (15.64) [1.38]	**23**
Total	**32**	**68**	**100**

The chi-square statistic is	6.9338
The p-value is	0.031213
This result is significant p < .05	

Output of a Chi-Square Test

Let's break down the results to more easily see what we're looking at. The first cell of information under the "Fail" column shows us the nine students who felt very prepared for the test but still failed. The (8.32) is the calculated expected result we would find in this sample, and the [0.06] is the chi-square point for this data. The cell to the right shows us the 17 students who felt very prepared and passed the test. You can follow those numbers down the rest of the column to see that most students who indicated they were somewhat or very prepared did pass the test, while those who were not prepared split pretty evenly between failing and passing.

The established p-value is set at a threshold of .05 for this example, which means we can feel more confident in our conclusions.

Goodness of Fit

If we wanted to test whether the students performed as expected overall, we would use goodness of fit. This test allows us to determine whether the pass or fail counts followed the hypothesized distribution. We would look at the number of actual observations (test takers) who had either passed or failed the test (this data would include the 100 students in our sample). The chi-square goodness of fit tests the hypothesized distribution and then provides the significance value to make a conclusion about whether or not it met what we expected. If we were to believe that 15% of our students will fail the test and 85% were to pass, then we can compare these values to the actual scores to determine if our data met that expectation.

Simple Linear Regression

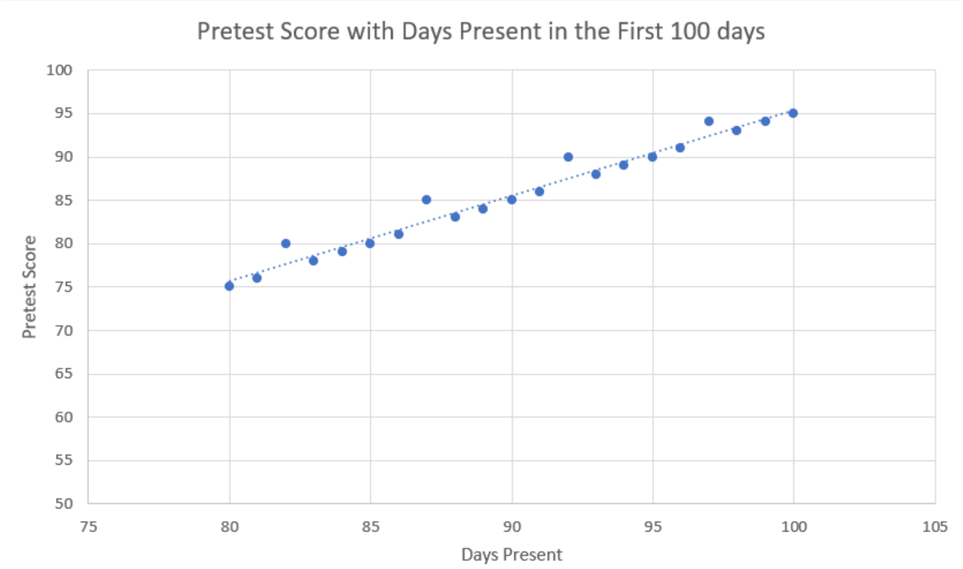
Regression analysis is a type of statistical method used to estimate relationships between a dependent variable and one or more independent variables. The outcome is a calculation that we will typically see visualized as a line on a scatter plot.

Simple linear regression is called simple because it is used to study the relationship between one dependent variable and one predictor, or independent variable. *Linear* refers to the straight-line relationship between the two quantitative values. The analysis tells us which predictor may have the largest impact.

Let's return to our test scores example. Suppose after the first 100 days of school, the eighth-grade students take a pretest. While plotting those pretest scores, along with attendance records, we notice that there appears to be a relationship between the two. "Appears" is the key word here. What this means is that it seems that the more present a student was, the more likely the score they achieved was higher.

We have one independent variable, days present, which is what we use as the predictor. We want to determine if a relationship exists between this variable and the score a student earns on their eighth-grade pretest (our dependent variable).

Let's look at the scatter plot of pretest scores in Microsoft Excel. In this visual, we have added a linear trend line, which provides the best fit line through the data points. Our question in the pretest analysis is, does days present have any relationship with the pretest score? Our research will attempt to determine if a student being more or less present in the first 100 days of school has any impact on their pretest score. This type of analysis will give us limited insight into the potential relationship between the two variables.

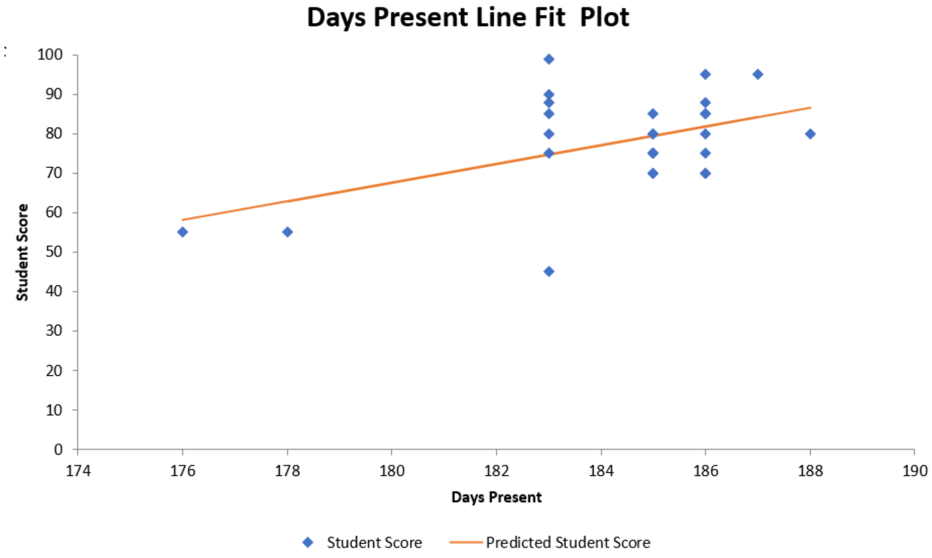


XY Scatter Plot with Linear Trendline in Microsoft Excel (Used with permission from Microsoft.)

In the visual, we see that there is a positive trend, meaning the line is going up in scores as the days present increases, and the plotted scores seem to fit the line pretty closely. This data does appear to have some relationship, but further analysis is always required.

Suppose the pretest analysis was interesting enough to school administrators that they've given the go-ahead for an exploration of whether days present has an actual impact on the actual scores. This is now the newest project for the analyst. We can use simple linear regression to assess whether or not the student absence variable is a strong predictor of the eighth-grade test score.

Using the Excel Statistical Analysis ToolPak, we use the actual test score with the days present through the whole school year. What we notice is that the markers on the line fit plot shown here do not fit the line as closely as the pretest example. So although it does appear that there is some relationship between attendance and scores, we can't suggest that days present is closely related to the scores. We will need to study further, and include more factors, to determine what truly impacts student test scores.



Simple Linear Regression Line Fit Plot from Regression in Excel Data Analysis ToolPak (Used with permission from Microsoft.)

The use of statistical software will create more statistical outputs than just the linear trend line, like the p-value and other necessary statistics for confidence. These tools all help us to determine how strong the relationship is between two variables, so that we can further study whether the values are correlated and have a causal relationship.

When there appears to be a relationship between variables, you will always want to further investigate. This is why you hear the phrase "correlation doesn't mean causation." Although there appears to be a relationship between present days and scores, it doesn't necessarily mean the score—either low or high—-can be explained by days present.

Correlation

Correlation is the statistical association between two (or more) equal variables. It does not tell us that one influences the other, but it does indicate that if one variable changes, the other variable changes as well. When researching correlation, we are simply attempting to determine what relationship might exist between two variables; we are not attempting to prove that one variable caused the other. A relationship in which one variable is proven to have an effect on another would be considered a **causal relationship**. Correlation only describes the relatedness of the relationship; in order to investigate whether a causal relationship exists, a correlation coefficient must be calculated.

Let's consider our student test scores example. Suppose we learn that a popular movie came out the weekend before our eighth-grade students took their test. Almost all our students went to the local theater and watched the movie at some point over the weekend. We gather data on which students watched the movie on Saturday night, which students watched the movie on Sunday night, and which students didn't watch the movie at all. We then take their test scores and test for correlation with the variable of watching the movie.

When we look for correlation, we discover that the students who watched the movie over the weekend all scored higher than 75 on their tests. At this time, we can confidently say that students who watched the movie scored higher on their test. But we cannot say that students scored higher on their test *because* they watched the movie. Nor can we say that any student who watched the movie is more likely to have scored above 75. We can only identify the strength of the relationship.

Pearson's correlation coefficient is a calculation used to measure a linear relationship between the data points, returning a value that is plus or minus 1 to determine the strength of the relationship. The correlation coefficient value is expressed as an **r value**. An r value that is close to 1 tells us that there is a strong correlation between the values, while an r value of or close to 0 means there is no correlation. R values between 0.4 and 0.7 represent a moderate correlation. There is also the coefficient of determination, which is expressed as R^2 or the square of the correlation coefficient value. This value is used to interpret the determination, and it is easier to understand if you multiply it by 100%. Let's see this in action by using the student scores and absenteeism example from our simple linear regression output.

Regression Statistics	
Multiple R	0.48512063
R Square	0.23534203
Adjusted R Square	0.20475571
Standard Error	11.0981969
Observations	27

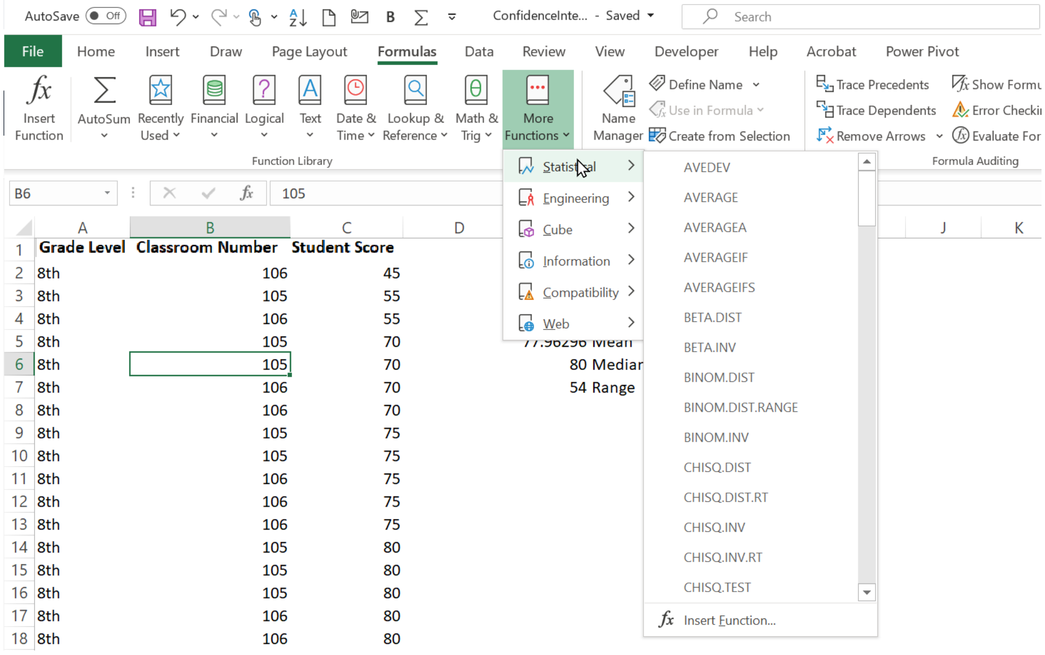
Regression Statistics Output using Excel Data Analysis ToolPak Regression

The Multiple R is Pearson's Correlation coefficient (r) and is currently showing 0.48, which means that there is a moderate correlation between the test score and the days absent. If we look at the R Square value, we can theorize that the 23% of variability in the score is explained by the variability in the days absent. This of course means that 77% of the variability is explained by other data not expressed in this sample.

This tells us that even though there appears to be a relationship between the number of days absent and the test score, the strength of that relationship is not high enough to explain or consider absences as the only factor affecting test scores.

Use Excel to Apply Statistical Methods

Excel has a multitude of functions for statistical use. You can find many of these functions in the Formulas tab of the Excel ribbon. If you navigate to the Statistical category, you will see all the available functions that have been defined for statistical analysis.



	A	B	C	D
1	Grade Level	Classroom Number	Student Score	
2	8th	106	45	
3	8th	105	55	
4	8th	106	55	
5	8th	105	70	77.96296 Mean
6	8th	105	70	80 Median
7	8th	106	70	54 Range
8	8th	106	70	
9	8th	105	75	
10	8th	105	75	
11	8th	106	75	
12	8th	106	75	
13	8th	106	75	
14	8th	105	80	
15	8th	105	80	
16	8th	105	80	
17	8th	106	80	
18	8th	106	80	

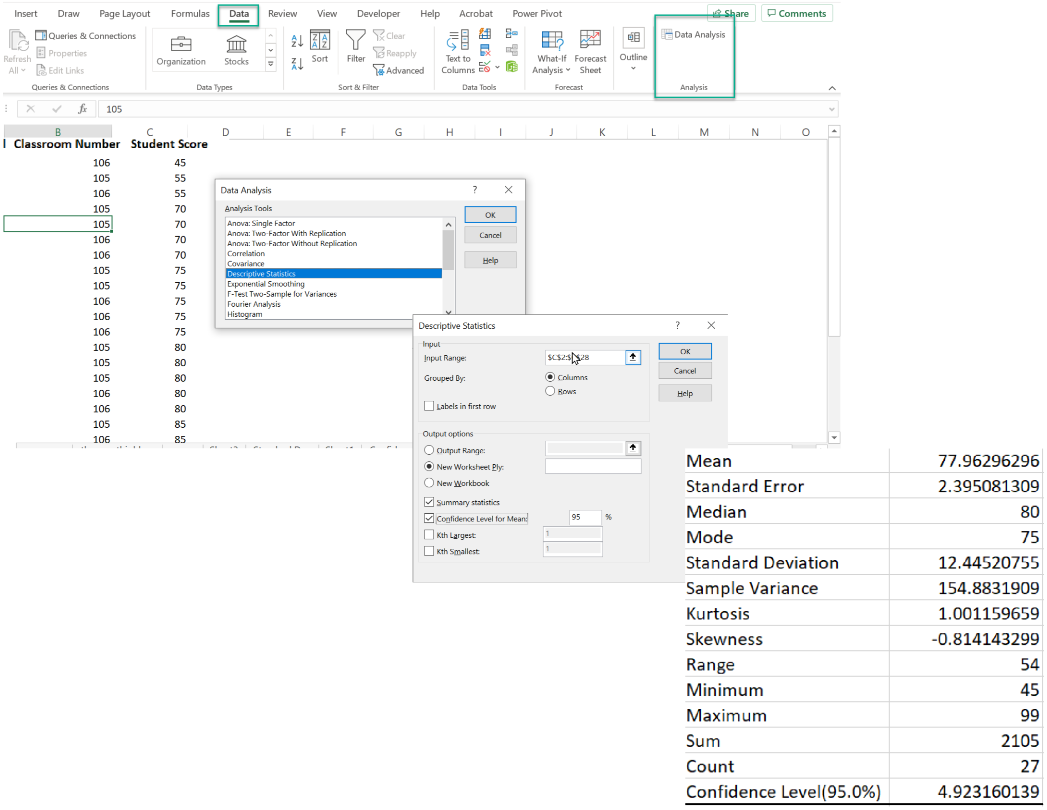
Excel Formulas Filtered for Statistical Functions (Used with permission from Microsoft.)

You can hover over each function to view basic information on that function, as well as access Help options that will provide further information and examples.

Some of the functions available in Excel include the following:

- AVERAGE (mean)
- MEDIAN
- MODE (single or multiple)
- MIN
- MAX
- STDEV (standard deviation)
- STANDARDIZE (z-score)
- CONFIDENCE

To run statistical tests directly in Excel, you will want to add the Excel Add-In to the Analysis ToolPak. The Analysis ToolPak will allow you to run and view outcomes for a variety of tests.



Classroom Number	Student Score
106	45
105	55
106	55
105	70
105	70
106	70
106	70
105	75
105	75
106	75
106	75
106	75
105	80
105	80
105	80
106	80
106	80
105	85
106	85

Data Analysis

Analysis Tools
- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram

Descriptive Statistics

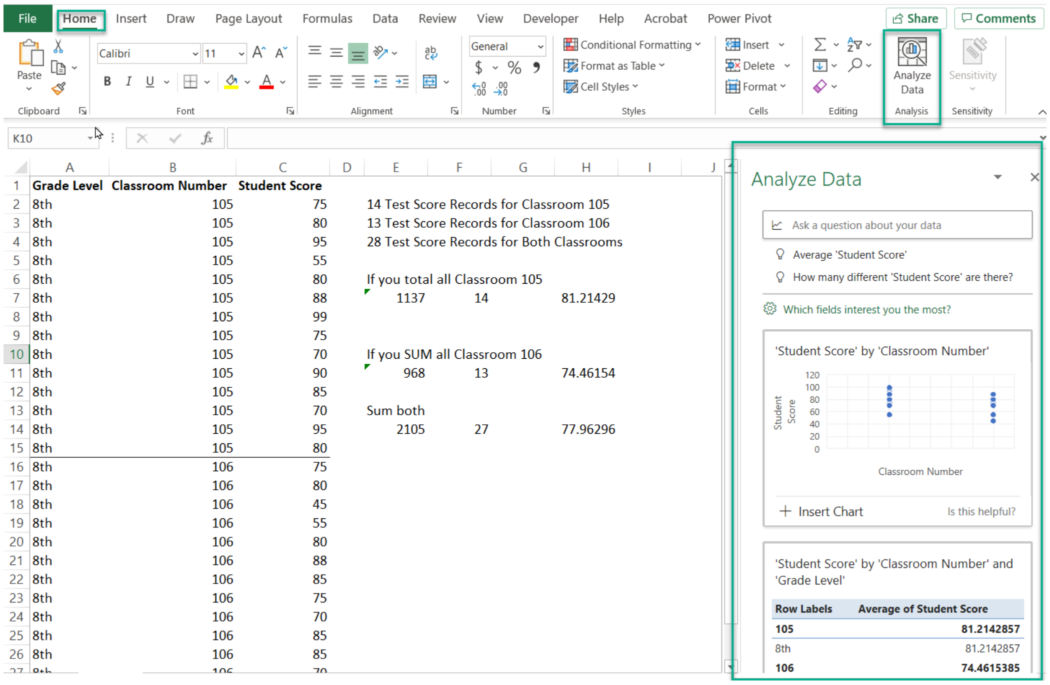
Input
Input Range: $C$2:$C$28
Grouped By: ● Columns ○ Rows
□ Labels in first row

Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook
☑ Summary statistics
☑ Confidence Level for Mean: 95 %
□ Kth Largest: 1
□ Kth Smallest: 1

Mean	77.96296296
Standard Error	2.395081309
Median	80
Mode	75
Standard Deviation	12.44520755
Sample Variance	154.8831909
Kurtosis	1.001159659
Skewness	-0.814143299
Range	54
Minimum	45
Maximum	99
Sum	2105
Count	27
Confidence Level(95.0%)	4.923160139

Excel Statistical Analysis ToolPak (Used with permission from Microsoft.)

While Excel is easy to use and widely available, other programs that are dedicated to statistical analysis will have a more comprehensive set of statistics tests. Despite this, later versions of Excel do contain new functionalities intended to support the data analyst. Explore the Home and Data tabs of the Ribbon in your version of Excel for all the possibilities.



	A	B	C	D	E	F	G	H
1	Grade Level	Classroom Number	Student Score					
2	8th	105	75		14 Test Score Records for Classroom 105			
3	8th	105	80		13 Test Score Records for Classroom 106			
4	8th	105	95		28 Test Score Records for Both Classrooms			
5	8th	105	55					
6	8th	105	80		If you total all Classroom 105			
7	8th	105	88		1137	14		81.21429
8	8th	105	99					
9	8th	105	75					
10	8th	105	70		If you SUM all Classroom 106			
11	8th	105	90		968	13		74.46154
12	8th	105	85					
13	8th	105	70		Sum both			
14	8th	105	95		2105	27		77.96296
15	8th	105	80					
16	8th	106	75					
17	8th	106	80					
18	8th	106	45					
19	8th	106	55					
20	8th	106	80					
21	8th	106	88					
22	8th	106	85					
23	8th	106	75					
24	8th	106	70					
25	8th	106	85					
26	8th	106	85					
27	8th	106	70					

Row Labels	Average of Student Score
105	**81.2142857**
8th	81.2142857
106	**74.4615385**

Working with the Selected Values in Columns A, B, and C and Using the Analyze Data Option in Later Versions of Excel (Used with permission from Microsoft.)

One newer functionality in Excel is Analyze Data, as shown above, which exists in Excel O365 and gives the data analyst suggested options to review, as well as the ability to ask questions about your data.

Review Activity:

Tests and Methods to Determine Relationships Between Variables

Answer the following questions:

1. **What type of analysis will typically involve an *x* and *y* scatter plot with a line?**

2. **What are two commonly used types of chi-square tests?**

3. **Correlation is used to measure what?**

4. **A relationship in which one variable is proven to have an effect on another would be considered what?**

5. **Regarding simple linear regression, how do we refer to the variables and the outcome?**

Lesson 11

Summary

After this lesson, you should be able to explain the use of inferential statistics to reach conclusions on the data and the relationships between data points. You should be able to describe relationships between a dependent variable and independent variable(s). You should also have discovered how to understand confidence intervals, to know how confident we are in the data that we are working with, and how to create the upper and lower bounds. You should have a practical understanding of the p-value and its importance in giving us statistical significance in our data with different statistical test methods. You should be able to distinguish when to use a t-test when comparing two different groups. You should also be familiar with the various methods used to test a hypothesis and the relationship between variables. We use chi-square and simple linear regression as we look for correlation.

Guidelines in Understanding Methods of Statistical Analysis

Consider these best practices and guidelines when familiarizing yourself with the statistical methods you will be working with.

1. Confidence intervals can be subjective, and there is no set percentage that indicates a good level of confidence.
2. A t-test is used when comparing two groups to determine if there is a significant difference between the means of both groups.
3. The dependent variable is our main data point, and we use it to determine the mean, median, mode, and standard deviation.
4. For statistical significance to exist, the difference between two values cannot have happened by chance and must be significant.
5. Most studies desire the p-value to be less than .05 ($p > .05$), as this signals that the difference between the two groups is significant and not just due to chance.
6. A hypothesis statement is a true or false statement. An alternative hypothesis assumes that a relationship exists between two variables, while a null hypothesis assumes that a relationship does not exist between two variables.
7. A type I error occurs when we reject the correct hypothesis and accept the incorrect hypothesis. A type II error occurs when we accept the incorrect hypothesis and reject the correct hypothesis.

8. A chi-square test can be used to determine if a difference exists between groups.

9. Simple linear regression is used to study the relationship between one independent variable and one dependent variable, and it is simple because there are only two variables involved.

10. Correlation is the statistical association between two (or more) equal variables that tells us if one variable changes, the other(s) will too.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 12

Using the Appropriate Type of Visualization

LESSON INTRODUCTION

After you analyze your data, you may realize that the information can be more easily digested if you can present it visually. When data lends itself to visualization, you must first figure out which type of visualization is best suited for your data. Learning about data visualization is really centered on learning how to effectively share insight with visuals. It could be that basic visuals are all you need to make a point. In other cases, you might require advanced visuals in order to bring certain issues to light. You might even need to map data. When deciding what visual you'll use, it's in your best interest to consider how a certain type of visualization is intended to be used alongside the story your data needs to tell.

Lesson Objectives

In this lesson, you will do the following:

- Use basic visuals.
- Build advanced visuals.
- Build maps with geographical data.
- Use visuals to tell a story.

Topic 12A

Use Basic Visuals

EXAM OBJECTIVES COVERED

4.4 Given a scenario, apply the appropriate type of visualization

Visuals can be extremely impactful to learning and processing. Consider that when children are first learning how to read, they are often presented with a picture alongside a letter. For example, to introduce the short vowel sound for the letter "A" a child may be shown a picture of an apple. This helps children remember that "A" sounds like "ah" as in "apple." This same line of thinking works for data as well. When lines of data are represented by a visual, we can easily see and digest the information without having to work with all the records. Basic visuals can help support the understanding of thousands or even millions of records of data and are commonly used to effectively represent information in a meeting, report, or dashboard.

Pie Charts and Treemaps

There are two types of visualization that are commonly used to display information in a way that shows the proportion of one (or more) data points: pie charts and treemaps.

Pie Charts

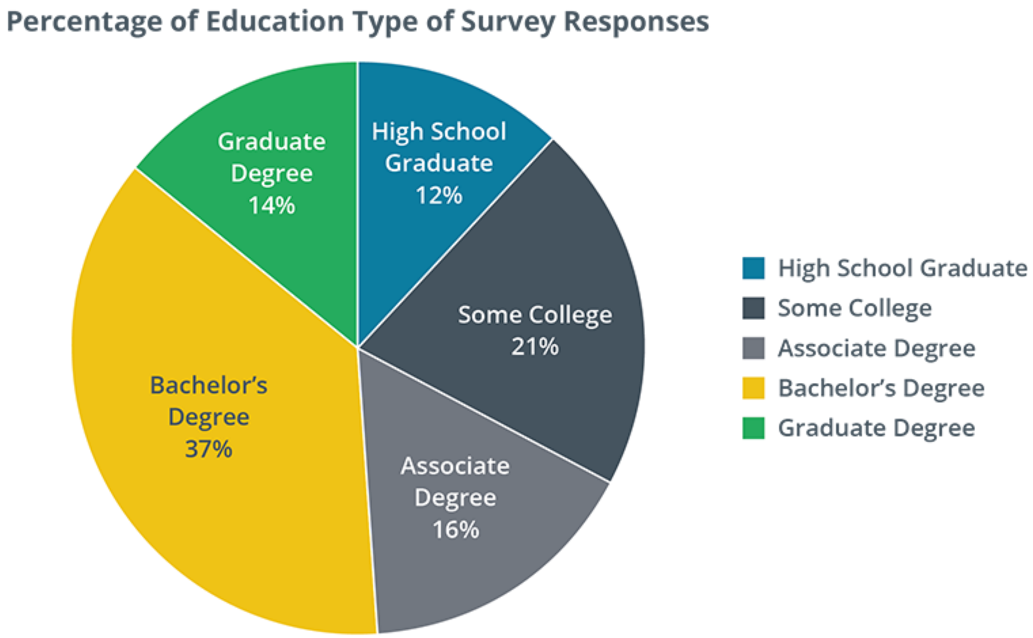
The most basic of all visuals is the **pie chart**, a circle broken into slices to represent percentages of information. The pie chart is easy to understand and thus is commonly used. Although there are other visuals that can show percentages, it is important to remember that a pie chart gives us a proportional visual for a whole group of records.

Let's see a pie chart in action. Suppose we have a data set from a survey where we asked respondents to specify their highest level of education. We have 610 responses, and the answer options were high school graduate, some college, associate degree, bachelor's degree, and graduate degree. First, let's aggregate this data into a table to count the total responses for each answer option.

Education Type	Count
High School Graduate	75
Some College	125
Associate Degree	100
Bachelor's Degree	225
Graduate Degree	85

Education Type with Count of Respondents

While the table is certainly more effective than reading all the records, it still requires the viewer to assess each value individually and then determine which education level received the most responses. But when we visualize this data in a pie chart, a simple glance easily shows us which answer has the most responses—and the highest proportion of the pie.

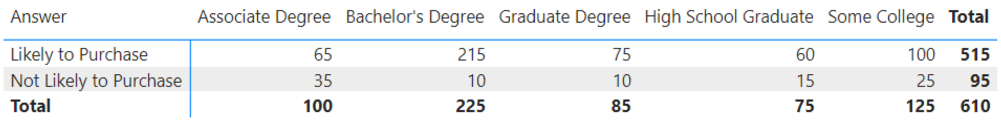


Basic Pie Chart Created in Excel (Used with permission from Microsoft.)

We can easily see that 37% of the responses to our survey responded that their highest education is a bachelor's degree.

Treemaps

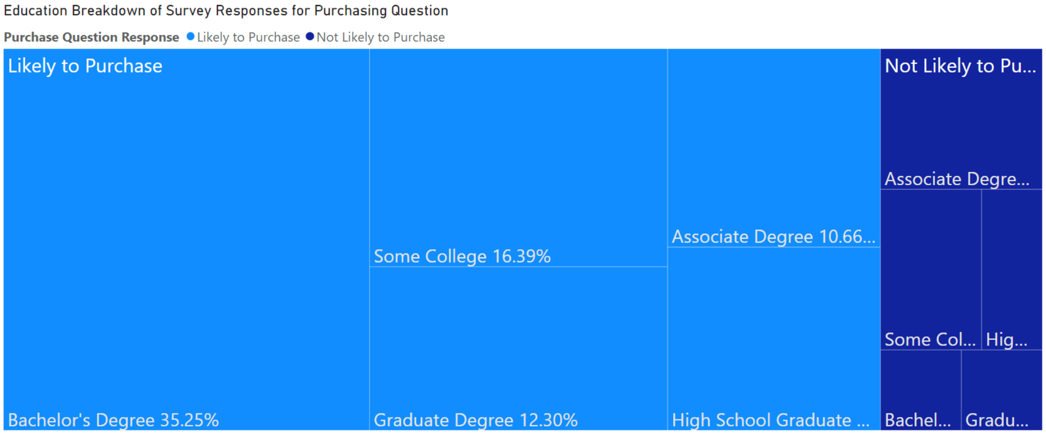
A **treemap** is a rectangle that shows the proportion of values using smaller rectangles within the larger one. Treemaps also can display a hierarchy when visualizing two data points, with a breakdown of both data points. For example, suppose we additionally surveyed our respondents about their likelihood of purchasing a product.



Answer	Associate Degree	Bachelor's Degree	Graduate Degree	High School Graduate	Some College	**Total**
Likely to Purchase	65	215	75	60	100	**515**
Not Likely to Purchase	35	10	10	15	25	**95**
Total	**100**	**225**	**85**	**75**	**125**	**610**

Survey Responses Displayed as a Matrix in Power BI (Used with permission from Microsoft.)

There were only two responses: likely to purchase or not likely to purchase. We can show the breakdown of this answer, as well as a breakdown of the education levels gathered previously, inside a treemap.



Treemap in Power BI Shows Purchase Responses Plus Level of Education Percentages with Tool Tip (Used with permission from Microsoft.)

This visualization allows us to easily see two things:

1. The majority of the respondents indicated they were likely to purchase the product.
2. Within people who were likely to purchase the product, the majority had bachelor's degrees.

They can be simple to read, but always remember that when you are working as an analyst you will want to be sure to have descriptions that include what information can be gleaned from each visual.

Column and Bar Charts

When you want to display the distribution of data, you'll want to choose a column chart or bar chart. These types of visualization show the total values of categories of data. Both bar charts and column charts are set on an *x*-axis (horizontal axis) and *y*-axis (vertical axis). The major difference between the two is the direction of the display: a bar chart reads horizontally while a column chart reads vertically. Either type of chart can show the distribution of basic information, so the choice of which to use depends on your preference.

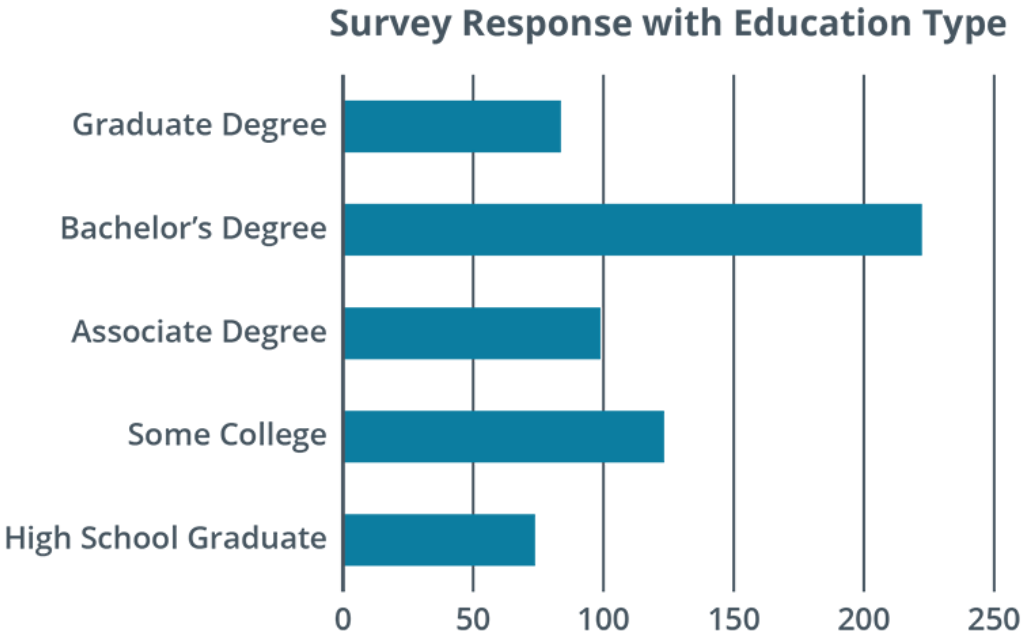
Bar Charts

In a simple **bar chart**, the *y*-axis (vertical axis) lists the categories of information and the *x*-axis (horizontal axis) is labeled with a set of (discrete number) values that are a set distance from each other.

Let's return to our survey example to create a bar chart. We list the categories, or the different answer options, on the *y*-axis. Then we draw the rectangle for each category across the *x*-axis to the total count for that answer. This bar chart allows the viewer to clearly see at a glance that most respondents indicated they have a bachelor's degree.

Education Type	Count
High School Graduate	75
Some College	125
Associate Degree	100
Bachelor's Degree	225
Graduate Degree	85

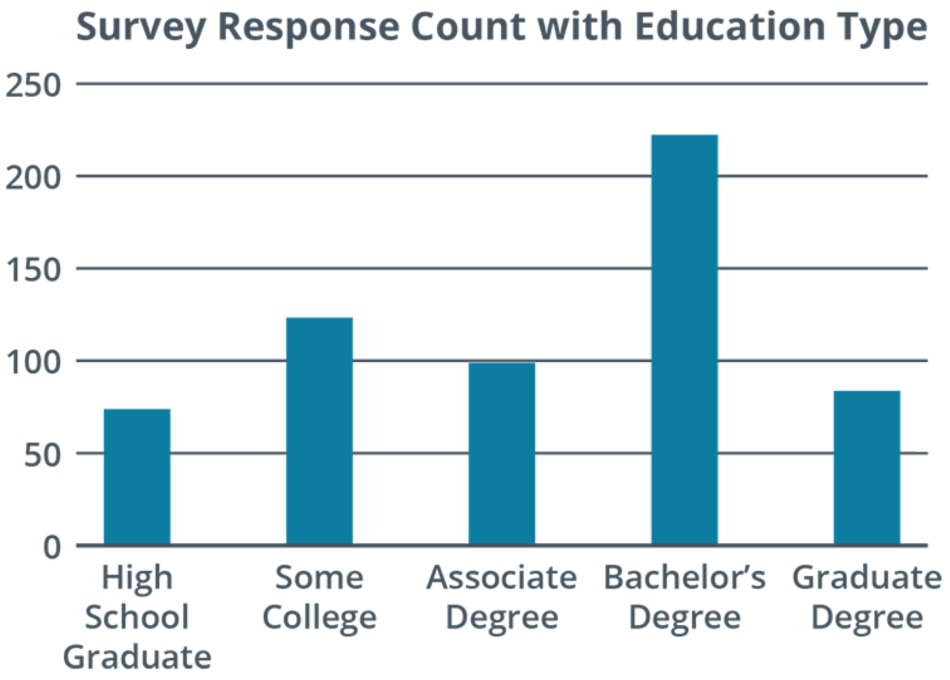
Table with Education Type with Count of Respondents



Count of Survey Responses by Type of Education Displayed in a Bar Chart in Microsoft Excel (Used with permission from Microsoft.)

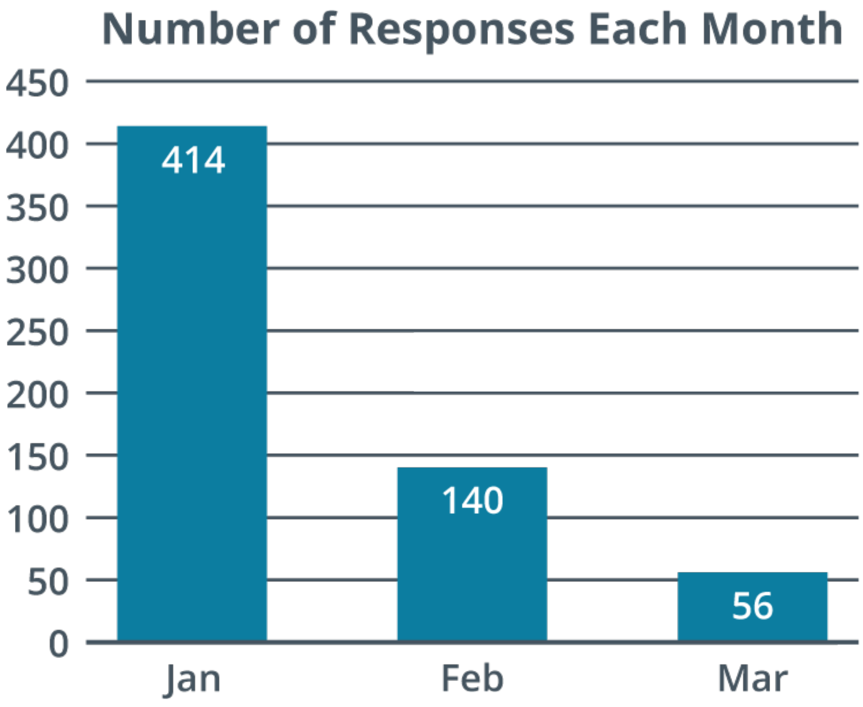
Column Charts

A **column chart** displays the same information as a bar chart but swaps the axes, meaning the *x*-axis lists the categories of information and the *y*-axis is labeled with the numerical values. Using the survey example again, we list the answer options on the *x*-axis and draw rectangles up to the appropriate value for each.



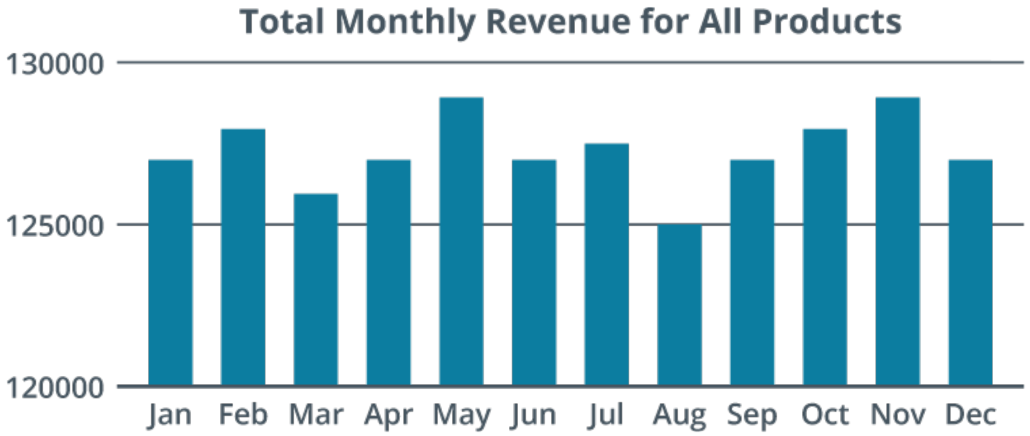
Level of Education Displayed in a Column Chart in Microsoft Excel (Used with permission from Microsoft.)

Column charts are also useful when we need to display ordinal information. For example, suppose our survey responses came in over a period of three months, and we want to see how many responses came in each month. Displaying this information in a column chart allows us to quickly see that we received the most responses in January, followed by much fewer responses in February, and even fewer in March.



Responses Collected Each Month in a Column Chart in Microsoft Excel (Used with permission from Microsoft.)

Column charts are also often used to show revenues, sales, or other information consisting of dollar amounts, as this type of chart allows us to isolate and quickly compare each time period.



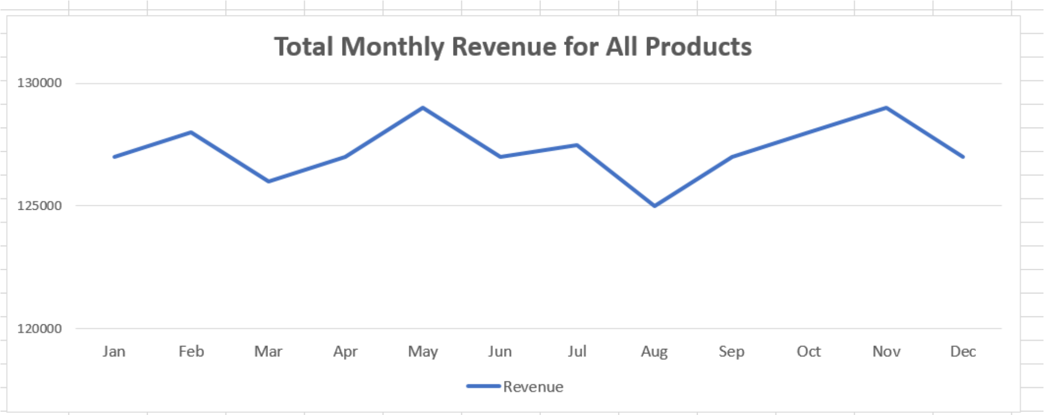
Monthly Revenue in a Column Chart in Microsoft Excel (Used with permission from Microsoft.)

Column charts are generally easy to read, but some people might prefer line graphs for this type of information. Thus, you will need to work with the consumer or decision maker when determing the appropriate visualization type.

Line Graphs

A **line graph**, or run chart, consists of either a single horizontal line or a group of multiple lines that represent different data points at different times. This is typically the visual of choice when we want to look at time series data, or data over intervals of time, as the connected line makes it easier to see exactly how that data changes over time.

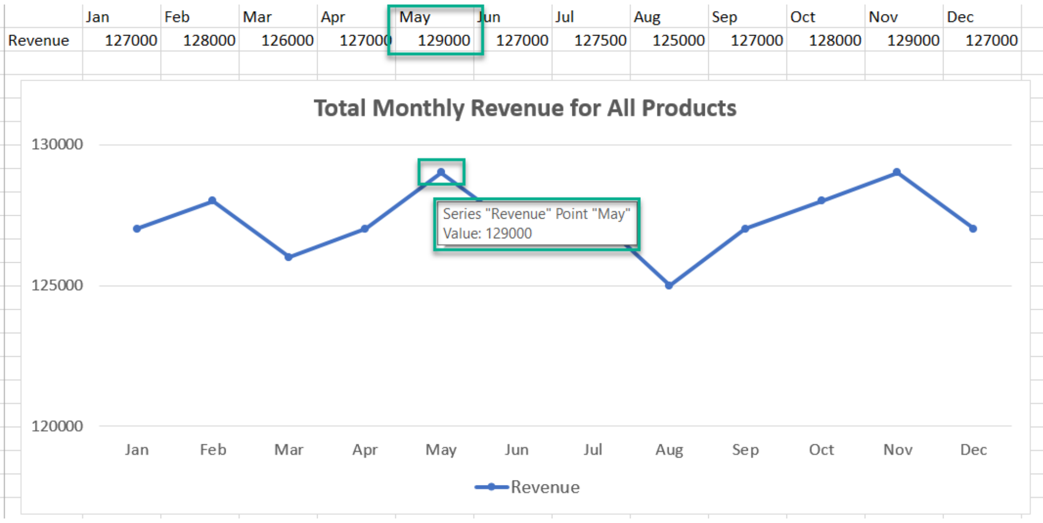
Let's return to the same series of monthly revenue data that we used in the column chart and now visualize it using a basic line graph. When we view the data this way, it can be easier to spot the trend of the data over a period of time. For example, we see that revenue peaks in May and declines to a low point in August, at which point it starts rising again to reach a high point in November.



Monthly Revenue in a Line Graph in Microsoft Excel (Used with permission from Microsoft.)

Adding Markers

To further emphasize certain points in a line graph, we can add a marker, which is often a dot or square placed on the line that makes it a little easier to distinguish the point. In a dashboard design, we can control what information shows up when the mouse hovers over the dot, allowing us to provide even more insight into the visual. In most popular visualization tools, the marker can also be used to display more information (in a pop-up) about the data point.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Revenue	127000	128000	126000	127000	129000	127000	127500	125000	127000	128000	129000	127000

Series "Revenue" Point "May"
Value: 129000

Revenue Each Month Presented in a Line Graph with Markers in Excel (Used with permission from Microsoft.)

Tools like Power BI and Tableau allow you to build more information into your pop-up display than Excel, like our example here that just shows the series name and the value.

Review Activity:

Basic Visuals

Answer the following questions:

1. **What type of visual appears as a circle broken into slices to represent percentages of information?**

2. **A treemap shows what information alongside values?**

3. **What is the difference between a column chart and a bar chart?**

4. **What do the lines in a line graph represent?**

Topic 12B

Build Advanced Visuals

EXAM OBJECTIVES COVERED

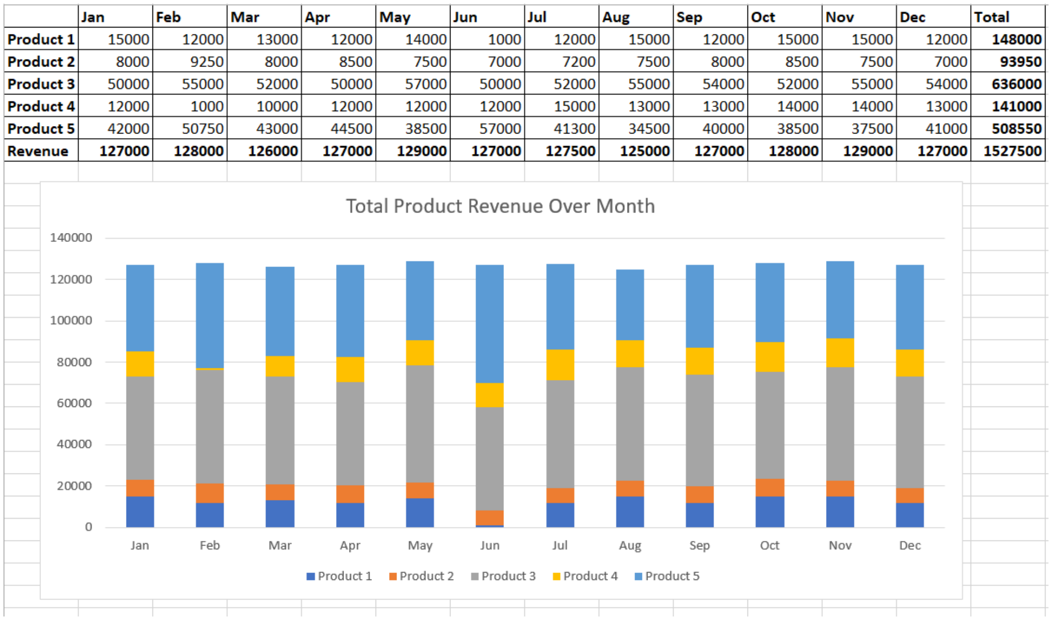
4.4 Given a scenario, apply the appropriate type of visualization

Basic visuals will show one or two data points in an easy-to-interpret manner, but advanced visuals are needed to represent more than just a single data point. As an analyst, you will create not only the visuals themselves, but will also create additional information to lie inside the visual, as done with stacked columns or bar charts. It's important to choose the correct type of visual for the data you are working with because you don't want to overcomplicate data visualization as much as you don't want to oversimplify it.

It can be challenging for people who do not build visuals for themselves to read them. As data analysts, we know what we are trying to show in a visual and must always keep in mind that the more advanced the visual appears, the more explanation it might require, until everyone who consumes the information understands how to read it.

Stacked Column/Bar Charts

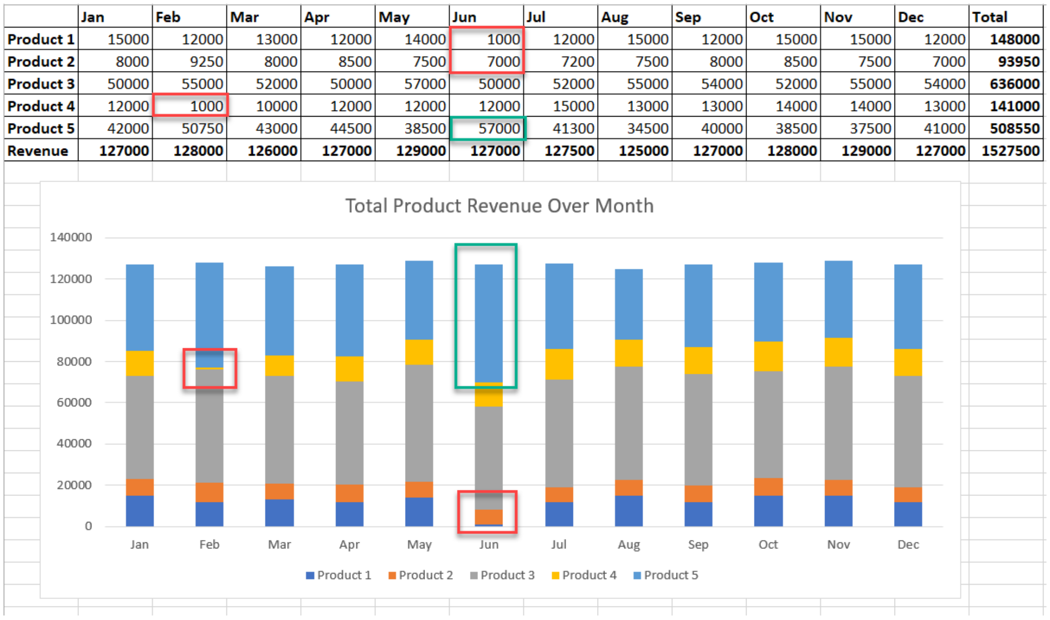
When we build basic column and bar charts, we set a single data point on each axis, which gives us a high-level look at the values. However, if we want to provide a deeper insight into those values, we can add additional data points to create a **stacked chart**, either bar or column. The stacked column/bar chart breaks the bar or column into separate portions, each representing an additional data point. For example, suppose we need to visualize monthly revenues for a suite of products. We could create a set of basic column charts, with each chart representing the monthly revenue of a single product. Or, we could use a stacked column chart to portray all product revenues in a single visual. To build this type of chart, we start off the same way: the *x*-axis contains the categories (months), and the *y*-axis contains the numerical values. However, we now have an additional data point of the products. The stacked column chart breaks the category columns up so that each of the five products makes up a portion of each column.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Product 1	15000	12000	13000	12000	14000	1000	12000	15000	12000	15000	15000	12000	**148000**
Product 2	8000	9250	8000	8500	7500	7000	7200	7500	8000	8500	7500	7000	**93950**
Product 3	50000	55000	52000	50000	57000	50000	52000	55000	54000	52000	55000	54000	**636000**
Product 4	12000	1000	10000	12000	12000	12000	15000	13000	13000	14000	14000	13000	**141000**
Product 5	42000	50750	43000	44500	38500	57000	41300	34500	40000	38500	37500	41000	**508550**
Revenue	**127000**	**128000**	**126000**	**127000**	**129000**	**127000**	**127500**	**125000**	**127000**	**128000**	**129000**	**127000**	**1527500**

Stacked Column Chart in Microsoft Excel for Product Performance (Used with permission from Microsoft.)

For example, if we look at the column for March, we can see that product 5 (light blue) and product 3 (gray) made up a large portion of total revenue for the month, with products 1, 2, and 4 (dark blue, orange, and yellow, respectively) making up a lesser amount. The stacked column/bar chart allows us to compare something within a category; in this case, the visual allows us to see that some products are performing better than others.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Product 1	15000	12000	13000	12000	14000	1000	12000	15000	12000	15000	15000	12000	**148000**
Product 2	8000	9250	8000	8500	7500	7000	7200	7500	8000	8500	7500	7000	**93950**
Product 3	50000	55000	52000	50000	57000	50000	52000	55000	54000	52000	55000	54000	**636000**
Product 4	12000	1000	10000	12000	12000	12000	15000	13000	13000	14000	14000	13000	**141000**
Product 5	42000	50750	43000	44500	38500	57000	41300	34500	40000	38500	37500	41000	**508550**
Revenue	**127000**	**128000**	**126000**	**127000**	**129000**	**127000**	**127500**	**125000**	**127000**	**128000**	**129000**	**127000**	**1527500**

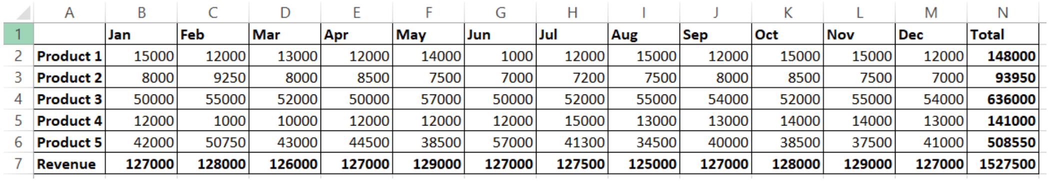
Highlighting Key Visual Information in a Stacked Column Chart in Microsoft Excel (Used with permission from Microsoft.)

This type of visualization also helps us more easily compare values for a data point across categories. For example, consider our product revenue example. When looking at the table, it might be hard to spot the low months for any given product. However, when you visualize this data in a stacked column chart, you can immediately see the low points for any given product. For example, we can see that product 4 (in yellow) performed much worse in February than it did in other months. And in June, we notice that both product 1 (dark blue) and product 2 (orange) had lower-than-normal revenues, while product 5 (light blue) had an above-average month. While total revenue in June appears to be fairly consistent with the year, the chart allows us to clearly see that the higher sales for product 5 that month made up the difference for the other two products having a low-revenue month.

Line Graphs with Multiple Lines

We previously covered how to visualize data with a basic line graph. A line graph that has multiple lines to represent each data point over the time series is no different than a line graph with a single line, except of course for the fact that it has multiple lines (and thus can portray multiple data points).

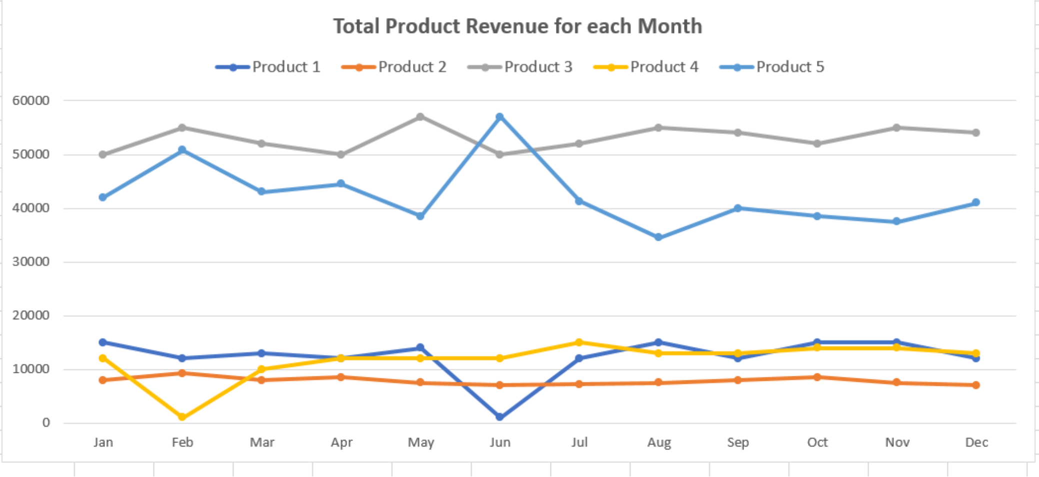
Let's return to our monthly revenue example. Suppose we want to visualize our data in a way that really highlights the top-producing products among all the other products.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Total**
2	**Product 1**	15000	12000	13000	12000	14000	1000	12000	15000	12000	15000	15000	12000	**148000**
3	**Product 2**	8000	9250	8000	8500	7500	7000	7200	7500	8000	8500	7500	7000	**93950**
4	**Product 3**	50000	55000	52000	50000	57000	50000	52000	55000	54000	52000	55000	54000	**636000**
5	**Product 4**	12000	1000	10000	12000	12000	12000	15000	13000	13000	14000	14000	13000	**141000**
6	**Product 5**	42000	50750	43000	44500	38500	57000	41300	34500	40000	38500	37500	41000	**508550**
7	**Revenue**	**127000**	**128000**	**126000**	**127000**	**129000**	**127000**	**127500**	**125000**	**127000**	**128000**	**129000**	**127000**	**1527500**

Data Used for Multiline Chart in Excel (Used with permission from Microsoft.)

We also want to display each product's high and low points. With this goal, a line graph with multiple lines and markers would be the best bet.

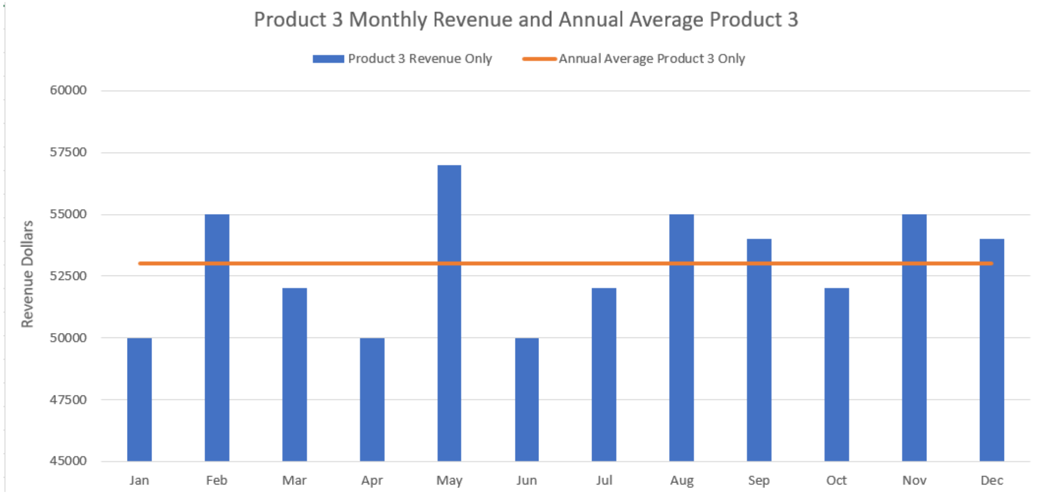


Lines with Markers for Each Product in Microsoft Excel (Used with permission from Microsoft.)

On the *x*-axis we place our categories (months), and the *y*-axis lists the monthly revenue values. Each product is represented by a different colored line. This visual allows us to see that products 3 and 5 are by far our top performers, with product 3 being the largest and most consistent source of revenue. Product 5 hits a high point in June and bottoms out in August, but it appears to level out again in the later months of the year. This graph allows us to very easily see the trends in data over time across multiple products, so that we may make comparisons.

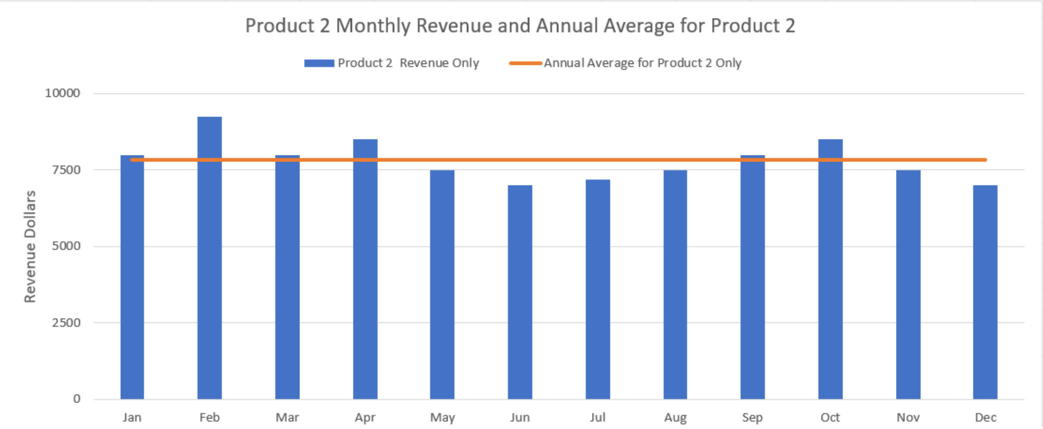
Combination Charts

There is also value in building visuals that combine columns and lines, particularly when you want to compare one or more data points (columns) against a trend (a line). You can use either single columns or stacked columns in a **combination chart**. Let's return to our product revenue example, assuming this time that we want to compare monthly revenues for product 3 against the average product sale.



Combo Chart in Excel Showing Monthly Revenue and Annual Average for Product 3 Only (Used with permission from Microsoft.)

This simple combination chart allows us to compare the performance of product 3 against its annual average. The contrast between the average and monthly values shows us that there are a few high months and a few low months for this product, rather than a consistent month-by-month performance. It's worth noting that the gap between the lowest value and highest value is $7,000 in revenue.

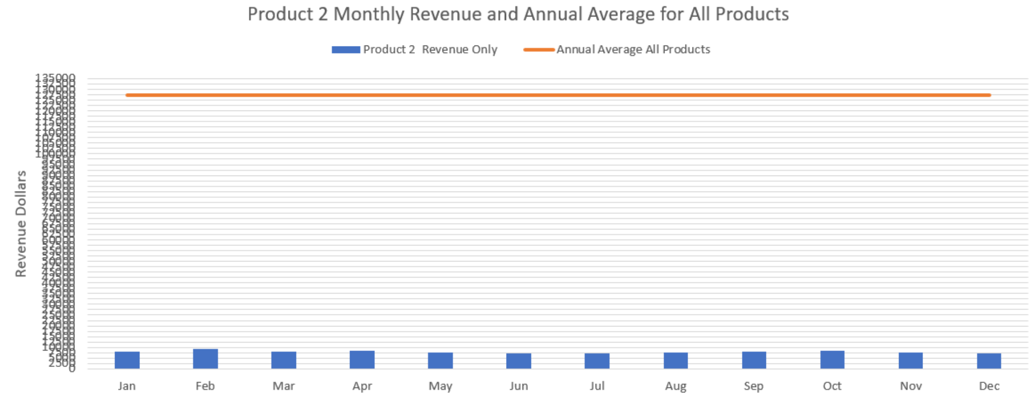


Combo Chart in Excel Showing Monthly Revenue and Annual Average for Product 2 Only (Used with permission from Microsoft.)

Let's do this again, but this time using the data for only product 2, which is one of the less revenue-producing products. By comparing this product's monthly revenue to the product's annual average, we can see that despite having revenues lower than the average, this product consistently performs pretty consistently across the year.

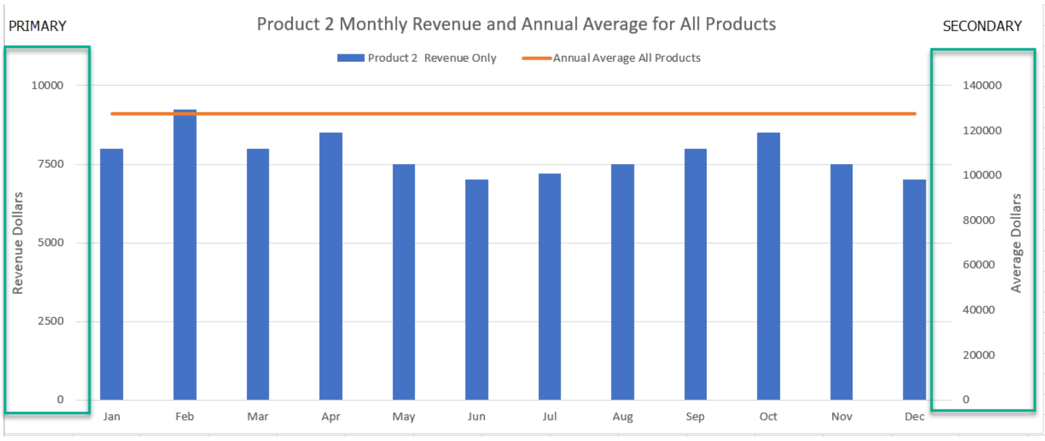
To make your visuals easier to read, it is crucial to give them proper titles and data labels, like what you see for the monthly revenues and annualized average related to specific products. For any visual you build, you will want to consider what additional elements should be included to make it more digestible for the audience/consumers.

Combination charts also allow us to utilize a secondary axis. The secondary axis can be used when the gaps between two sets of numbers are so large that the distance makes it hard to read the chart purely from a visual standpoint. Suppose we want to create a chart that compares the monthly revenue for product 2 against the annual average for all products. As you can seein the screenshot, the annual average for every product of $127,290 is far greater than the revenues per month for product 2.



Product 2 Monthly Revenue and Annual Average for All Products Displayed in a Combo Chart in Excel Without the Secondary Axis for the Average (Used with permission from Microsoft.)

This certainly is an accurate depiction of the product performance, but because the difference between values causes the *y*-axis to greatly lengthen, its hard to even see where the product has high or low months. To visualize this information better, we can leverage the secondary axis.



Product 2 Monthly Revenue and Annual Average for All Products Displayed in a Combo Chart in Excel Using the Secondary Axis for the Average (Used with permission from Microsoft.)

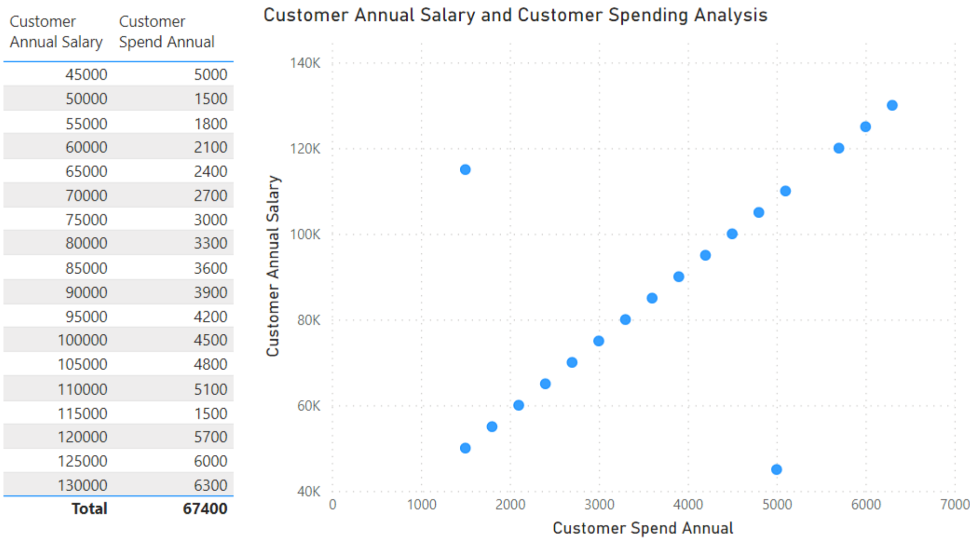
We now present both sets of data in a more readable format. We can see not only that this product's revenue lies way below the average of all products, but also the higher and lower revenue months on the primary axis.

Scatter Plots and Bubble Charts

Scatter Plots

A **scatter plot** consists of two variables plotted on the *x*-axis and *y*-axis, with a dot placed on the graph where the two data points converge on both of the axes. Scatter plots help us determine whether there is a relationship between the two variables placed on the axes, and are especially useful when we want to spot outliers.

Let's use an example to create a scatter plot. Suppose we are conducting an internal market analysis and want to determine whether an individual's salary has any bearing on the amount of money they spend on our products annually. To do this, we need to compare two variables: their annual salary and their annual spending on our products. We can place these two values on the axes of a scatter plot, and then add a dot to the graph at every point where the annual spend matches the annual salary indicated.



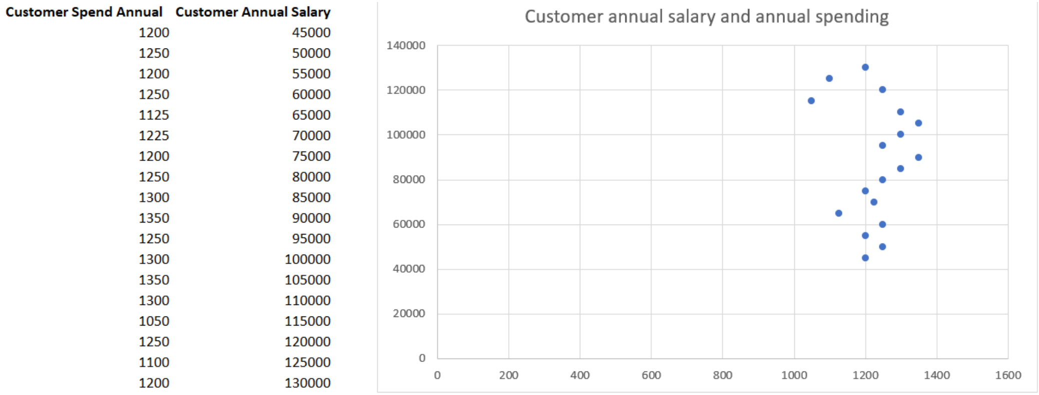
Customer Annual Salary	Customer Spend Annual
45000	5000
50000	1500
55000	1800
60000	2100
65000	2400
70000	2700
75000	3000
80000	3300
85000	3600
90000	3900
95000	4200
100000	4500
105000	4800
110000	5100
115000	1500
120000	5700
125000	6000
130000	6300
Total	**67400**

Basic Scatter Plot in Microsoft Power BI (Used with permission from Microsoft.)

We can analyze the layout of the dots to determine whether there is any correlation between these two variables. In this example, the dots are steadily headed in an upward direction and in a straight line. This trend shows us that the more money an individual makes, the more they spend on our products. If the trend of the dots had instead gone steadily downward, we would make the opposite conclusion—that the more money an individual makes, the less they spend on our products.

Returning to this scatter plot, we should note that there are two outliers in our data. The data point at the top left indicates this person has a high annual income, but spends less on our products. The next outlier on the bottom right indicates an individual who spends a fairly high amount of money on our products despite having a low annual salary. Despite these two outliers, however, our conclusion still stands. The outliers are an exception to the trend, but the trend still exists.

If there truly is no trend at all, that's when we can determine that the two variables are not related. For example, let's plot a separate set of data containing annual salaries and annual spending for a different company.



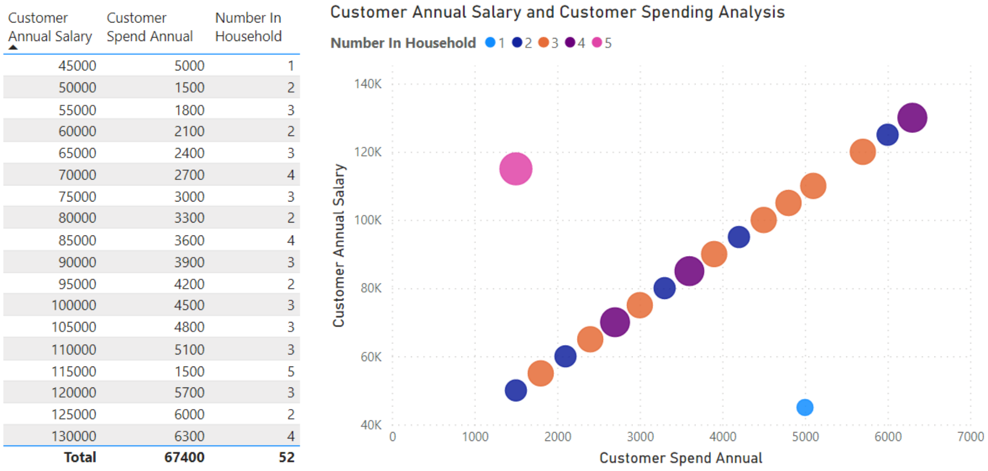
Customer Spend Annual	Customer Annual Salary
1200	45000
1250	50000
1200	55000
1250	60000
1125	65000
1225	70000
1200	75000
1250	80000
1300	85000
1350	90000
1250	95000
1300	100000
1350	105000
1300	110000
1050	115000
1250	120000
1100	125000
1200	130000

Scatterplot in Microsoft Power BI Showing No Relationship between Annual Salary and Annual Spend (Used with permission from Microsoft.)

This time, you will notice that the dots are generally placed in the same area. All the individuals in this data set spend between $1,000 and $1,400 regardless of their annual salary. A scatter plot helps us visualize the connection between two variables so we can easily see whether there is (or isn't) a strong relationship.

Bubble Charts

Unlike scatter plots, a **bubble chart** allows us to work with more than just two data points/variables on the *x*-axis and *y*-axis. This visual will allow you to add a third variable or dimension of the data as represented by the size of the dot (or bubble). When working with the scatter plot and wanting additional data to convey additional meaning, you can use the bubble chart to show the third value via dot size. Let's look again at our annual salary and spending data for the first company we analyzed, but this time let's also visualize the number of people in the household, as we suspect this variable might play a role in spending.



Customer Annual Salary	Customer Spend Annual	Number In Household
45000	5000	1
50000	1500	2
55000	1800	3
60000	2100	2
65000	2400	3
70000	2700	4
75000	3000	3
80000	3300	2
85000	3600	4
90000	3900	3
95000	4200	2
100000	4500	3
105000	4800	3
110000	5100	3
115000	1500	5
120000	5700	3
125000	6000	2
130000	6300	4
Total	67400	52

Bubble Chart in Microsoft Power BI (Used with permission from Microsoft.)

This additional data point gives some context to the two outliers we saw previously. For example, we can now see that the high salary, low spending outlier has five people in the household, whereas the low salary, high spending outlier only has one person in the household. We can draw the conclusion that the one-person household can afford to spend more on our products despite having a lower salary, as opposed to the household with more people (and thus more overall spending needs).

Scatter plots and bubble charts are useful visualizations when you want to assess for an existing relationship between two variables, and bubble charts specifically can provide additional insight into what additional factors may contribute to the presence of outliers.

Histograms

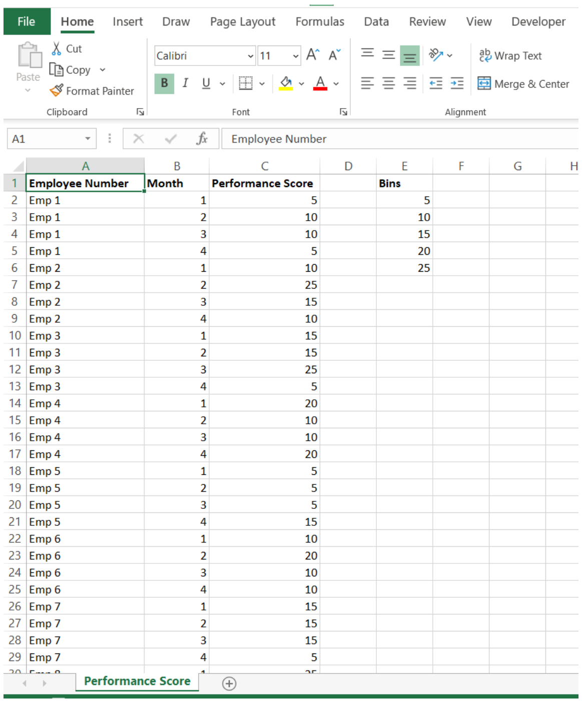
A **histogram** is similar to a column chart, but the key differentiator is the ability to show a frequency of values that are grouped by bins, or class intervals. Each bin is aligned on the *x*-axis, while the metric we want to assess against is listed on the *y*-axis. The size or height of each column portrays the volume of the numbers in each bin. Aesthetically, while a column chart has spaces between the columns, there is no space between the columns of a histogram. When we are formatting the data for histogram creation, there are four rules to keep in mind.

1. Each bin should be the same size. Your bins can contain groups of 5, 19, 100, or even 1,000 records. The size you need ultimately depends on the number of records you're analyzing. No matter how many records end up in your bin, make sure the group size is consistent. For example, suppose our data set includes values ranging from 1 to 100. We could create 20 bins in groups of 5 or 10 bins in groups of 10. Outliers must be included in your bins. You will want to consider the lowest number in your data set and the highest number, and then choose the size of your group accordingly. For example, if most of your data set includes values from 80 to 100, but you have one value that is 20, your bins need to include values from 20 to 100.

2. To make the chart easier to read, whole numbers should be used for both the start and end of the bin, when possible.

3. The number of bins should not be exceedingly large. While the number of bins ultimately depends on your volume of data, a good rule of thumb is to use 20 or fewer bins.

While you can build a histogram manually, a tool such as Excel makes this task easier, particularly when you're working with large data sets. Let's walk through an example. Suppose that at the end of each month, a company asks its employees to complete a performance survey consisting of Likert scale questions, with choices from 1 to 5. The values chosen are then summarized to create a performance score for each employee, for each month. This goes on for a period of 4 months.

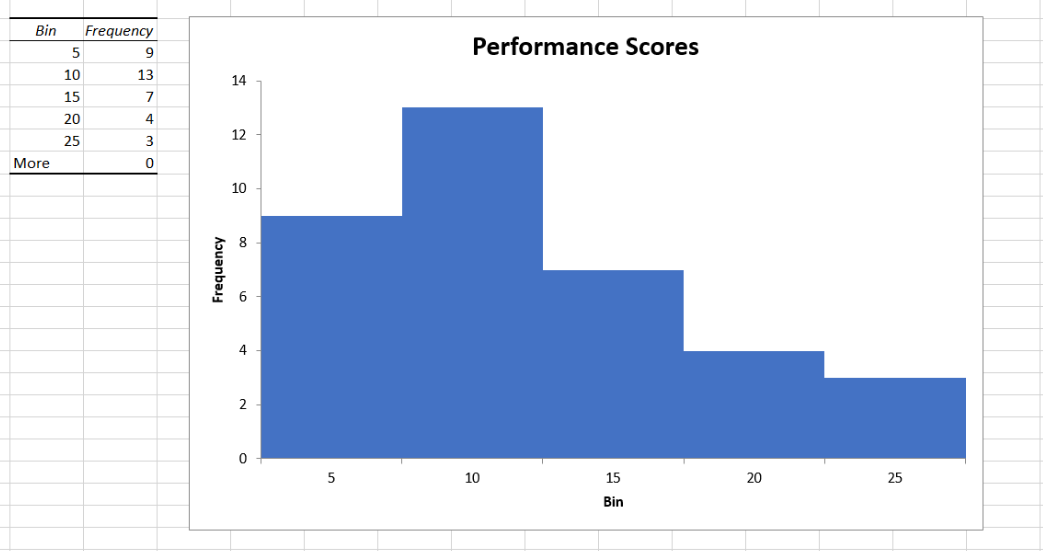


	A	B	C	D	E
1	Employee Number	Month	Performance Score		Bins
2	Emp 1	1	5		5
3	Emp 1	2	10		10
4	Emp 1	3	10		15
5	Emp 1	4	5		20
6	Emp 2	1	10		25
7	Emp 2	2	25		
8	Emp 2	3	15		
9	Emp 2	4	10		
10	Emp 3	1	15		
11	Emp 3	2	15		
12	Emp 3	3	25		
13	Emp 3	4	5		
14	Emp 4	1	20		
15	Emp 4	2	10		
16	Emp 4	3	10		
17	Emp 4	4	20		
18	Emp 5	1	5		
19	Emp 5	2	5		
20	Emp 5	3	5		
21	Emp 5	4	15		
22	Emp 6	1	10		
23	Emp 6	2	20		
24	Emp 6	3	10		
25	Emp 6	4	10		
26	Emp 7	1	15		
27	Emp 7	2	15		
28	Emp 7	3	15		
29	Emp 7	4	5		

Performance Score Data in Excel (Used with permission from Microsoft.)

If we want to see the frequency of the performance scores across the four months for which the survey was performed, we should use a software visualization tool to build a histogram from our data.

The highest score in our data set is 25, so we will break our data out into five bins of five. You can see the bins identified in our table to the right of the actual values. We will use the Excel data analysis ToolPakto build our frequency output, which is how many times a score showed up and in what bin, and then display that information in our chart.



Bin	Frequency
5	9
10	13
15	7
20	4
25	3
More	0

Final Output of Histogram Using Excel Data Analysis ToolPak (Used with permission from Microsoft.)

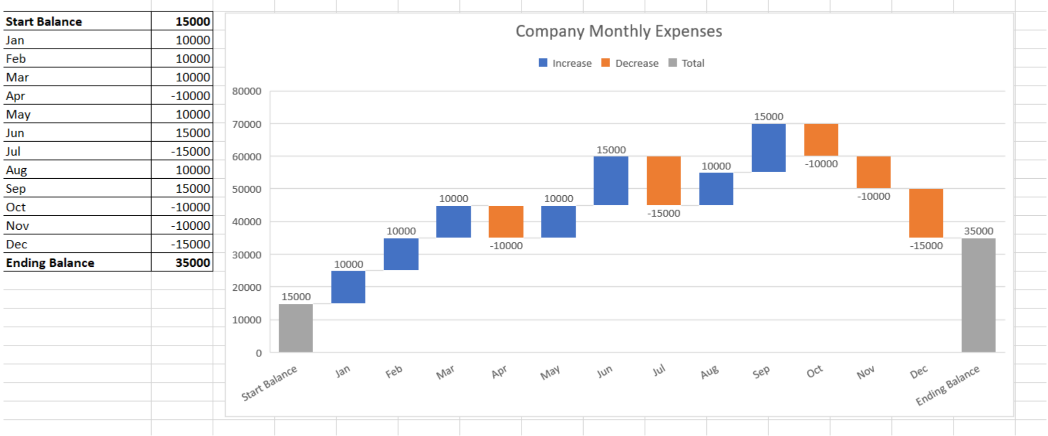
Looking at the histogram, we see that the most selected score is 10; that's the mode of these values. We can also see that most of our data landed in the leftmost area of the histogram. In the performance survey, low values meant low performance ratings. Thus, the histogram makes it clear that the employees overall denoted lower performance scores over the four-month period than they did higher scores.

When you create a histogram in Excel, the default chart contains gaps between the bars. In our visual, we can decrease the gap width to remove the space between the columns in the chart-formatting options.

Waterfall Charts

A **waterfall chart** is used to show performance over time, such as when tracking operating expenses, cash flow, or even growth in customers or investments. A waterfall chart visualizes how the money flows through from starting balance to ending balance. The time period is set on the *x*-axis while the money value is set on the *y*-axis.

For example, let's suppose we want to visualize the cash flow for our company for an entire year. Our starting balance is $15,000, and each month our columns either move up the *y*-axis (when cash increases) or move down the *y*-axis (when cash decreases). The columns continue through each month, until the end of the year, at which point we are presented with our ending balance.



Start Balance	**15000**
Jan	10000
Feb	10000
Mar	10000
Apr	-10000
May	10000
Jun	15000
Jul	-15000
Aug	10000
Sep	15000
Oct	-10000
Nov	-10000
Dec	-15000
Ending Balance	**35000**

Waterfall Chart in Microsoft Excel Displaying Company Monthly Expenses with Starting Bank Balance and Ending Balance (Used with permission from Microsoft.)

The size of the blocks is based on the amount of cash earned or lost each month. The axis in this scenario sets major lines for every $10,000. If we focus on the columns for the starting point and the first quarter of the year, we will see that we are steadily moving up to $45,000. (We start with $15,000 and add $10,000 in January, February, and March.) Every time we add or take away money the waterfall chart shows not only the change in cash flow for that month, but also demonstrates the performance in relation to every previous month. At the end of the year, our ending balance gives us our total balance for the entire year, reflecting all the ups and downs. In this example, we end the year with an additional $20,000.

When you use Excel to build waterfalls, don't forget to set the starting balance and the ending balance as totals. You do this by right-clicking on those columns and selecting "set as total."

Review Activity:

Advanced Visuals

Answer the following questions:

1. **Which type of visual displays additional data points as separate proportions?**

2. **Which type of visual would work best for visualizing the run time (in years) of multiple television series?**

3. **Combination charts typically use what type of visual alongside columns (either single or stacked)?**

4. **What differs between how a scatter plot and a bubble chart plot points?**

5. **In which way does a histogram display values that a column chart does not?**

6. **Which type of visual is best to display performance over time?**

Topic 12C

Build Maps with Geographical Data

EXAM OBJECTIVES COVERED

4.4 Given a scenario, apply the appropriate type of visualization

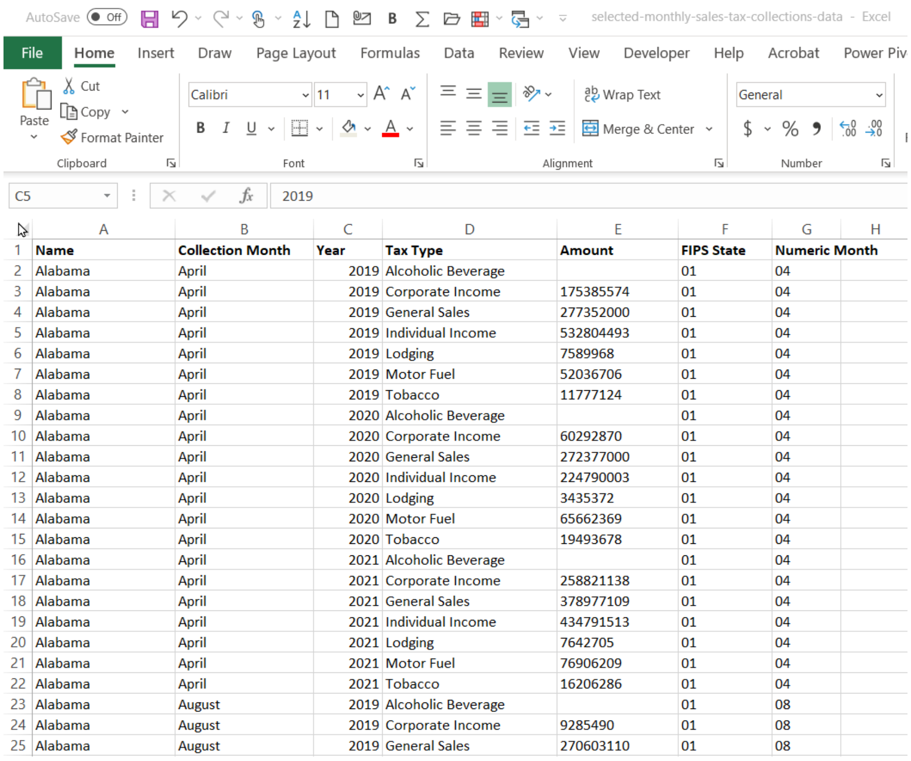
These days, many organizations spread beyond typical borders, crossing over into many different geographic areas. Even smaller companies are likely to have data from different geographical areas, particularly when it comes to surveys and polls. When working with data that has geographic fields, you may find there is a desire to map the data, to show how it appears across the country or even the world. In order to do this, you first will need to confirm that your geographic fields are recognized by the visualization tool.

Preparing Geo Fields for Mapping

Any tool that provides mapping, like Power BI or even Tableau, will also have built-in mapping visuals. There are also software programs specifically designed for mapping, like ArcGIS, which is dedicated to mapping data and offers comprehensive mapping components.

These visuals do have an expectation that your data is recognized as a geographic field. If the program you are using does not automatically detect your data as a geographic field, you can easily convert it. It is also important to ensure that geographic data is complete before you start the mapping process. For example, someone in a survey may have filled in their address but forgotten the zip code. It could also be that you have only partial geographic data—you might have city and state data, but lack county data. You may also find that the software expects very specific data, like latitudinal and longitudinal coordinates. We wouldn't ask a survey taker to provide this data, but we will use data sets to merge the data we have with this required data.

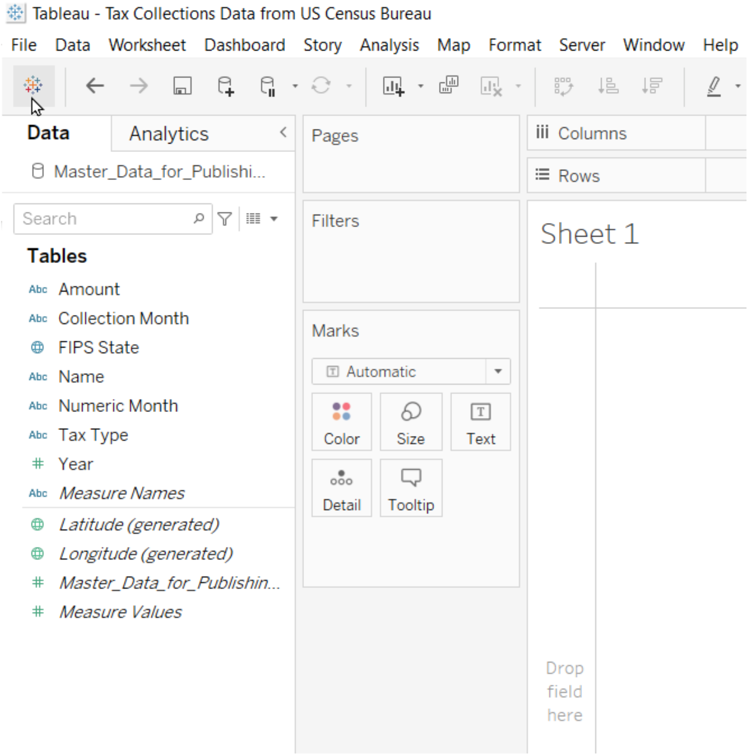
Let's walk through preparing geographic data for mapping using a sample data set. The Excel list shown here contains data for the monthly amount of tax collected by each tax type. The name column contains the name of the state—that's our geographic field. Other columns contain the collection month, year of collection, tax type, amount, FIPS state codes, and numeric month.



	A	B	C	D	E	F	G	H
1	Name	Collection Month	Year	Tax Type	Amount	FIPS State	Numeric Month	
2	Alabama	April	2019	Alcoholic Beverage		01	04	
3	Alabama	April	2019	Corporate Income	175385574	01	04	
4	Alabama	April	2019	General Sales	277352000	01	04	
5	Alabama	April	2019	Individual Income	532804493	01	04	
6	Alabama	April	2019	Lodging	7589968	01	04	
7	Alabama	April	2019	Motor Fuel	52036706	01	04	
8	Alabama	April	2019	Tobacco	11777124	01	04	
9	Alabama	April	2020	Alcoholic Beverage		01	04	
10	Alabama	April	2020	Corporate Income	60292870	01	04	
11	Alabama	April	2020	General Sales	272377000	01	04	
12	Alabama	April	2020	Individual Income	224790003	01	04	
13	Alabama	April	2020	Lodging	3435372	01	04	
14	Alabama	April	2020	Motor Fuel	65662369	01	04	
15	Alabama	April	2020	Tobacco	19493678	01	04	
16	Alabama	April	2021	Alcoholic Beverage		01	04	
17	Alabama	April	2021	Corporate Income	258821138	01	04	
18	Alabama	April	2021	General Sales	378977109	01	04	
19	Alabama	April	2021	Individual Income	434791513	01	04	
20	Alabama	April	2021	Lodging	7642705	01	04	
21	Alabama	April	2021	Motor Fuel	76906209	01	04	
22	Alabama	April	2021	Tobacco	16206286	01	04	
23	Alabama	August	2019	Alcoholic Beverage		01	08	
24	Alabama	August	2019	Corporate Income	9285490	01	08	
25	Alabama	August	2019	General Sales	270603110	01	08	

Data from the US Census Burea Data Tools Site for Tax Collection (Used with permission from Microsoft.)

However, when we put the data into Tableau to perform mapping, we can see that some of our data types are incorrectly identified. The FIPS state data is identified as a geographic field, when this is a number. On the other hand, the Name data, which should be identified as a geographic field, is marked as text.



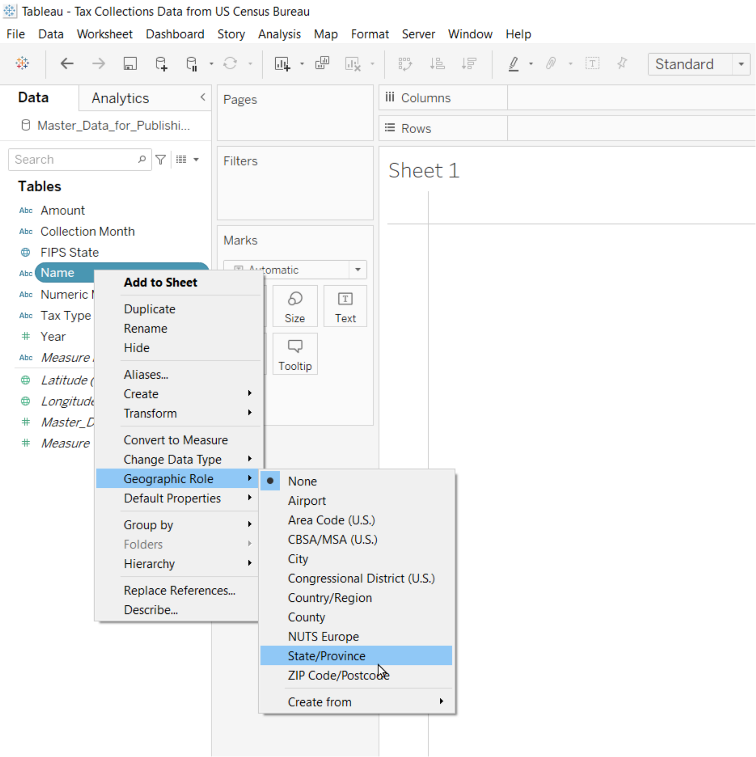
Master_Data_for_Publishi...

Tables

- Abc Amount
- Abc Collection Month
- ⊕ FIPS State
- Abc Name
- Abc Numeric Month
- Abc Tax Type
- # Year
- Abc *Measure Names*
- ⊕ *Latitude (generated)*
- ⊕ *Longitude (generated)*
- # *Master_Data_for_Publishin...*
- # *Measure Values*

Data Transferred to Tableau for Mapping (© 2021 Salesforce.com, Inc. All rights reserved. Used here with permission.)

In order to effectively map, we will need to ensure that these fields are properly recognized. The process used to convert data to a geographic field depends on the software you are using. To do this in Tableau Desktop, simply right-click on the field you need to change (in this case, Name), go to "Geographic Role," and choose the proper type.



Converting to Geographic Role Field in Tableau Desktop (© 2021 Salesforce.com, Inc. All rights reserved. Used here with permission.)

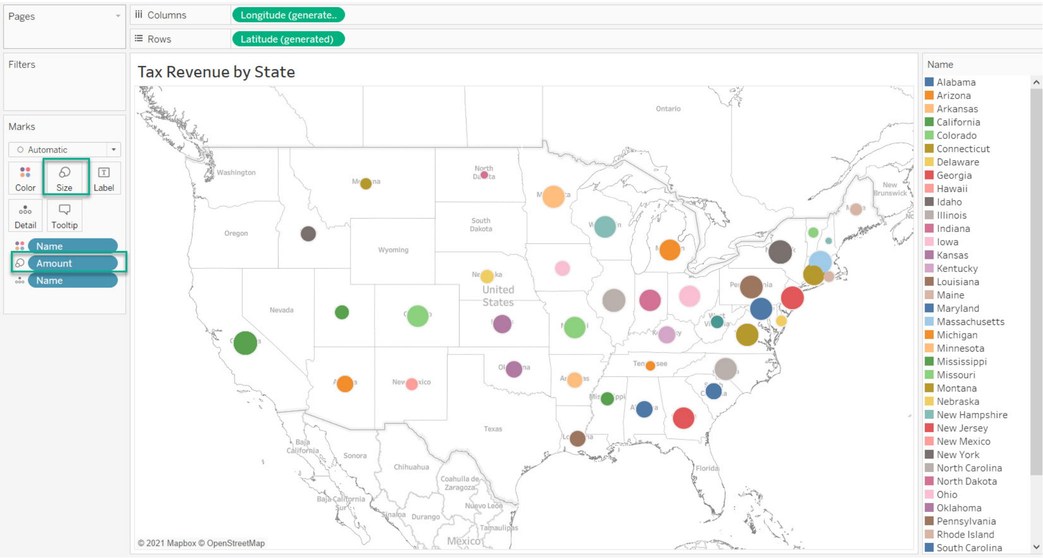
There is also software intended to be used specifically for mapping, like ArcGIS, which is dedicated to mapping data and thus offers comprehensive mapping components.

Geographic Maps

There are two styles of maps you will commonly use to display geographic data: the **dot map**, which uses markers to note specific spots on the map, and the **filled map**, which fills in the borders of a location. You may also see **layered maps**, which blend the two types of maps together.

Dot Maps

Dot maps are typically used when we want to pinpoint a location, like a postal code area, or highlight an amount of data occurring within a particular point, such as within a particular state. For example, suppose we want to identify which states reported individual tax type incomes to the US Census Bureau in 2019, and how much was reported. Using a dot map, we can easily identify not only the states that reported the tax types (via the dot placed in the center of each state), but also how much was reported—the larger the dot, the greater the amount that was reported. We can see that South Dakota did not report these taxes, while North Dakota did. However, North Dakota reported a small amount compared to a state like New York.



Columns: Longitude (generate..

Rows: Latitude (generated)

Marks: Automatic

Name

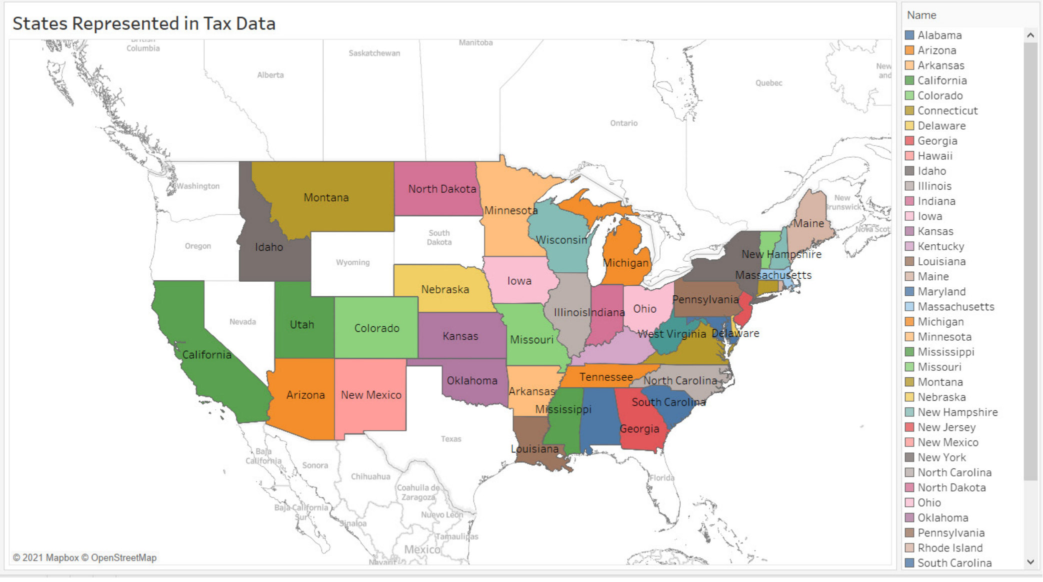
Amount

Name

Dot Map in Tableau Desktop (© 2021 Salesforce.com, Inc. All rights reserved. Used here with permission.)

Filled Maps

Filled maps fill the border of a geographic location. In our scenario since we are referring to the entire state versus a point in the state, we may want to use a filled map in some mapping scenarios, but you do lose the sizing component of the dot in a filled map.



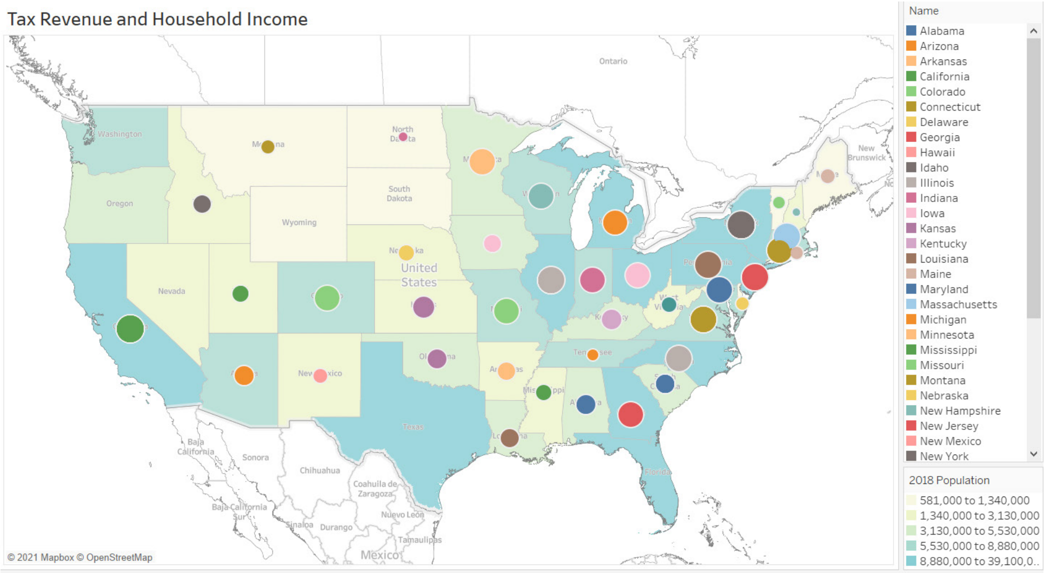
Filled Map in Tableau Desktop (© 2021 Salesforce.com, Inc. All rights reserved. Used here with permission.)

This map highlights each state that is in the data by filling in the whole state boundary with a color. We can easily detect which states are in this data set, and which states are not.

Layered Maps

When you want to visualize multiple data points, a layered map may be your best choice. Layered maps contain different layers for various data points within the data set, using both dot and filled approaches, and can give you the most comprehensive experience.

Let's return to our tax reporting example. Suppose that we also want to visualize average household income of each state. We could present two different maps to convey this information. Or, we could use a layered map to display household income and tax reporting data in the same visual.



Data Layer Map with Dots to Show Tax Revenue with Dots, and Household Income for Each State with a Filled Background in Tableau (© 2021 Salesforce.com, Inc. All rights reserved. Used here with permission.)

In this visual, the filled map represents the average household income for each state, while the dot map represents the state tax data from the US Census Bureau.

Review Activity:

Maps with Geographical Data

Answer the following questions:

1. **Which software is dedicated to mapping and offers comprehensive mapping components?**

2. **Which data transformation will likely be needed before visualizing geographic fields?**

3. **Which type of map is used to highlight the amount of data occurring within a particular point?**

4. **Which type of map is used to visualize multiple data points?**

5. **A filled map highlights what portions of a visual with a color?**

Topic 12D

Use Visuals to Tell a Story

EXAM OBJECTIVES COVERED

4.4 Given a scenario, apply the appropriate type of visualization.

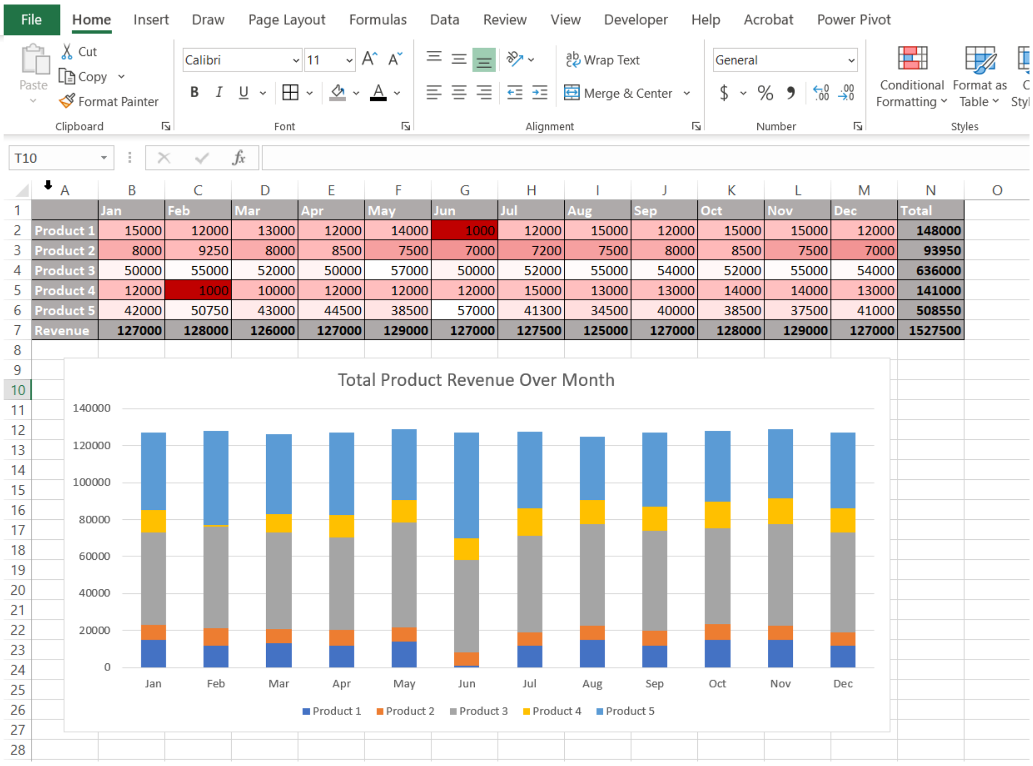
Visuals always tell a story, but how well they tell that story depends on the storyteller. In this case, the storyteller is our choice of visualization. When building visuals, you will want to choose visualization types that best support the overall story your data is telling. Heat maps, infographics, and simple word clouds can be key components of visual storytelling, especially when combined with other charts, graphs, and maps as appropriate.

Heat Maps

A **heat map** is any visual that uses color to draw attention to a "hot" spot, or a part of the visual that needs pointing out. A heat map can use color scales to point out a high or low number, or it can use color to draw attention to points on a geographic map.

Color Scales

Suppose we have product sales numbers for the year, and we want to draw attention to the months with extremely low sales. We can use Color Scales with Conditional Formatting in Excel to create a heat map for these values. There are multiple colors to choose from, so you can use whatever best suits your needs. For example, here we are using red, since we are highlighting low sales. But we could use green, or any other color, if we instead wanted to draw attention to months of high sales.



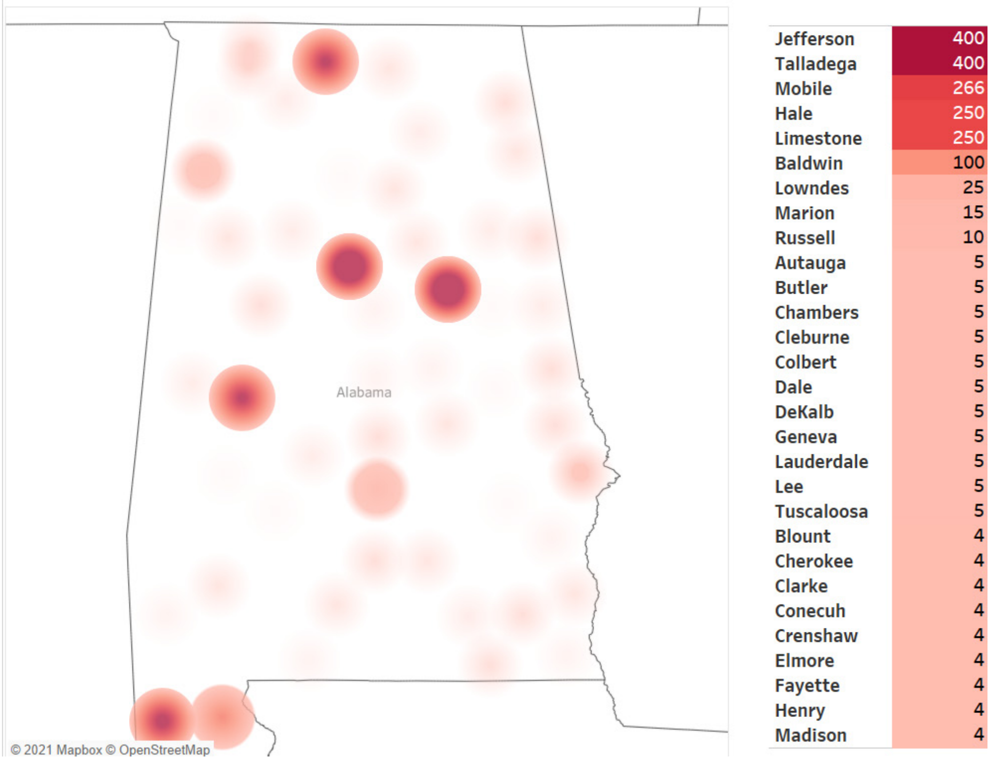
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
Product 1	15000	12000	13000	12000	14000	1000	12000	15000	12000	15000	15000	12000	148000
Product 2	8000	9250	8000	8500	7500	7000	7200	7500	8000	8500	7500	7000	93950
Product 3	50000	55000	52000	50000	57000	50000	52000	55000	54000	52000	55000	54000	636000
Product 4	12000	1000	10000	12000	12000	12000	15000	13000	13000	14000	14000	13000	141000
Product 5	42000	50750	43000	44500	38500	57000	41300	34500	40000	38500	37500	41000	508550
Revenue	127000	128000	126000	127000	129000	127000	127500	125000	127000	128000	129000	127000	1527500

Color Scale Conditional Formatting in Excel (Used with permission from Microsoft.)

As you can see, the red color applied onto these values allows us to easily see which months are most impacted, and gives us an idea of which values we should be looking at in the stacked chart underneath the heat map.

Geographical Heat Maps

Heat maps can also be used when you want to draw attention to points on a geographic map. Let's look at an example. Suppose we want to identify how many locations for a certain business exist in each county in the state of Alabama. On the right-hand side, we have our list of the number of locations per county, organized from most to least. To visually show which counties have the highest number of locations, we use a heat map.



Jefferson	400
Talladega	400
Mobile	266
Hale	250
Limestone	250
Baldwin	100
Lowndes	25
Marion	15
Russell	10
Autauga	5
Butler	5
Chambers	5
Cleburne	5
Colbert	5
Dale	5
DeKalb	5
Geneva	5
Lauderdale	5
Lee	5
Tuscaloosa	5
Blount	4
Cherokee	4
Clarke	4
Conecuh	4
Crenshaw	4
Elmore	4
Fayette	4
Henry	4
Madison	4

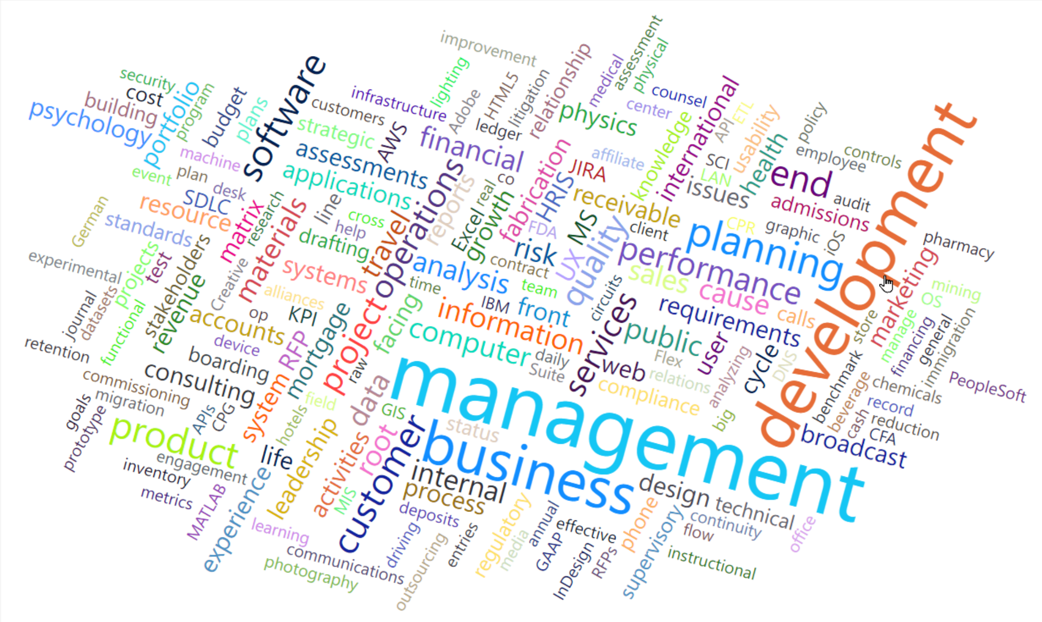
Density Map and Color Scales in Tableau Dashboard (© 2021 Salesforce.com, Inc. All rights reserved. Used here with permission.)

The list of counties uses a color scale; more dense/darker colors indicate a higher number of locations. On the map, we place dots on every county, which vary in color density corresponding to the number of locations. This allows us to see the areas with the most locations in list format and geographically. This information is useful when it is important to understand the geographic spread of a variable, such as when a business is considering adding more locations.

Word Clouds

A **word cloud** is a visual representation of the words used in a particular body of text. A word cloud analyzes the text that you're working with and visually portrays the words in varying sizes to identify how often a particular word is used in the text. The largest words are those that are used most frequently, while the smallest words are those that appear less frequently. You can tweak the settings to suit your needs, such as choosing to exclude certain words from the analysis that you may not want to skew results. You can also remove **stop words**, like "the," "and," "or," and "to." Stop words are words that appear frequently in your data but do not need to be counted; removing them leaves the more relevant words to be counted and visualized.

For example, suppose an organization is planning internal training sessions for the coming year and decides to ask its employees to identify topics that they would like to see training on. Each employee provides the organization with one to five topics, written into a free-form survey box. If we want to visualize this data, we will likely choose to use a word cloud.



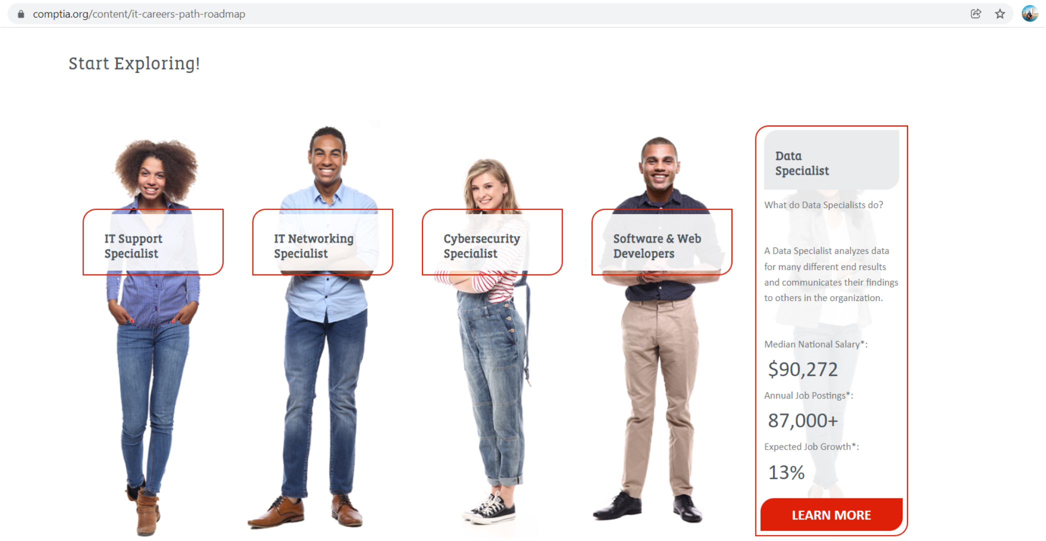
Sample Word Cloud Developed with Word Cloud Visual in Power BI (Used with permission from Microsoft.)

The largest words—management, development, and business—are the words that were used most often by the employees. This tells us that management training, for example, is something that many of the employees would be interested in, which can help the organization plan. A word cloud is not only visually appealing, but also more easily analyzed than a count of each word.

Infographics

An **infographic** is any combination of visuals, artwork, photos, and language that tells the story of your data in a compelling and graphically appealing way. Infographics allow people to easily digest information and are often used on websites, handouts, social media posts, and magazines. It is unlikely that a company would use an infographic to deliver operational data to help run a production line, but that same company might provide potential clients a beautifully designed infographic to explain why they should partner together.

There are tools that are dedicated to the building and development of infographics. People who work in graphic design might use Adobe Illustrator, while those who are less graphically savvy may use a tool that provides a little more hand-holding, like Canva or another online infographic tool. No matter which tool you're using, it's important to remember that you must have the data and the numbers needed to design a visually compelling infographic in the first place.



What do Data Specialists do?

A Data Specialist analyzes data for many different end results and communicates their findings to others in the organization.

Median National Salary*: $90,272

Annual Job Postings*: 87,000+

Expected Job Growth*: 13%

CompTIA's Website Displays an Infographic About Different Job Roles in IT

This screenshot displays an infographic that presents information for each of the technical pathways provided by CompTIA. When you hover over the infographic on the website, it is interactive, allowing you to not only visually see information but also to navigate and discover more information. You can visit the site and see the infographic for yourself, here: https://www.comptia.org/content/it-careers-path-roadmap.

Review Activity:

Visuals to Tell a Story

Answer the following questions:

1. **A heat map can also be referred to as what?**

2. **What can be used in Excel to create a color scale?**

3. **The data points of a geographic heat map are usually colored based on what?**

4. **Which type of visual would be appropriate for visualizing open text responses from a survey?**

5. **What are some tools that might be used to create an infographic?**

Lesson 12

Summary

After this lesson, you should have a better understanding of how to use and read basic visuals, like pie charts, tree maps, bar and column charts, and line graphs. You should also have discovered how to leverage more advanced charts, like stacked charts, graphs with multiple lines, combination charts, scatter plots, bubble charts, histograms, and waterfall charts. You should have learned about mapping geographic data and highlighting information with color scales, and discovered the practical visual storytelling that is provided by using word clouds and infographics.

Guidelines in Using the Varying Types of Visualizations

Consider these best practices and guidelines when familiarizing yourself with the data visualizations you will be working with.

1. Basic visuals are the best choice when you want to share information in a way that is easy to understand, using a visual type that is common and thus recognizable by most people.
2. Pie charts and tree maps are used to show the proportions of data as a whole.
3. Bar charts and column charts can be used to clearly display counts and percentages for data.
4. Line charts (either with a single line or multiple lines) effectively show value changes over time periods.
5. Stacked charts are useful when you want to add another dimension of data to your visual.
6. Combination charts are used to compare one or more data points against a trend.
7. Scatter plots and bubble charts are both used to determine whether a relationship exists between two variables, but only the bubble chart allows for the addition of a third dimension, which adjusts the size of the bubble.
8. Histograms use bins to visualize the frequency of your data, and are used quite often in statistics analysis.
9. Waterfall charts show the ebb and flow of values and are extremely useful anytime you have a starting and ending position, especially for financial analysis.
10. Before you can map data, you must make sure you've prepared the geographic fields so that the software you use recognizes your data as geographic.
11. A dot map is used to draw attention to a particular point on a map, while a filled map is used to highlight the entire boundary of a country, state, city, or county.

12. Heat maps use color to draw visual attention to areas of interest.

13. Word clouds are the visualization of choice when you have a lot of text-based information.

14. Infographics use a mix of data, charts, and imagery to generate a creative view of the data that is easily digestible and visually appealing.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 13

Expressing Business Requirements in a Report Format

LESSON INTRODUCTION

A key skill for the data analyst is the ability to translate information into business requirements. Developing business requirements in the data world means that, as an analyst, you understand the types of requirements needed to perform any business request that comes your way. You must then work through those requirements to deliver the appropriate style of report or dashboard, with the correct type of filters, and at the right time to the right people. From the outset, a data analyst must develop a high-level understanding of the audience for the data, where the data came from, and how it will be delivered. This information will guide you in designing reports and dashboards that meet requirements and have the appropriate view filters and navigation. Understanding and outlining the requirements for data helps to ensure that reporting not only meets specification, but that it is also usable by the people who need the information.

Lesson Objectives

In this lesson, you will do the following:

- Consider audience needs when developing a report.
- Describe data source considerations for reporting.
- Describe considerations for delivering reports and dashboards.
- Develop reports or dashboards.
- Sort and filter data.

Topic 13A

Consider Audience Needs When Developing a Report

EXAM OBJECTIVES COVERED

4.1 Given a scenario, translate business requirements to form a report.

The development of business requirements requires an analysis of the audience who will be using a report or visualization. It's important to understand the different needs and desired insights across different types of consumers within an organization, including both internal and external stakeholders, because this information will guide what data you will include in your reports. It's also important to consider the level of access to data that should be provided to different types of consumers.

The Audience

When developing any type of reporting, you need to determine your **audience**—the people who will be using the data within your reports and dashboards. An understanding of your audience is important because this knowledge will help you to identify what type of data they need and what access they may require to support the reporting process. The audience may be a person within a specific role, a department, or even the entire organization.

The people who are in your audience would also be considered the **distribution list**, or those who receive the report or dashboard. The distribution list can sometimes be noted on the actual report or dashboard. This allows the audience to know who else is receiving the report, thereby avoiding questions like, "Did John see this report?"

It's not only important to understand who the users of your report are, but also which individuals will have access to specific types of information. Consider the following example involving human resources (HR) data and salary information. Although these functions each have a dedicated departmental team, only HR leadership and the managers of each department in the organization should be able to access certain information related to performance reviews. Because salary information is confidential, the analyst should have a list of managers who need access to the information, because only these individuals should ultimately be the audience for this data.

Consumer Types

Information presented in visual format is used by different types of consumers. In an organization, both internal and external parties consume data. It's important to remember that each type of consumer has different requirements and/or needs. When writing your initial business requirements, be sure to determine who your audience will be and what type of consumer they are. This knowledge allows you to determine what type of view would be ideal and what access they will need to the information, which in turn helps you design better visualizations and provide more insight. Types of consumers include the following:

- C-level executives
- Management
- External vendors/stakeholders
- General public
- Technical experts

Think about the type of consumer that will use your information when writing business requirements and designing visuals. When designing visualizations for a particular consumer type, consider how to give people access to the data and what type of viewing requirements they may have. For example, you can't give the general public access to your internal data systems for security purposes.

It's also important to consider what insights they should gain from a visualization. The data content should be geared toward the overall project. This can include the standards of the data, like what fields are to be named, and also what data is included. Remember when developing a report or data project that your consumers do not need all of the data at all times. Clarification on what data content is needed for the objective falls right in line with the consumer type. A person who is in a management position might need details or insight into a particular set of business processes, or even department-level insight. In contrast, the C-level executives (Chief Executive Officer, Chief Operating Officer, and Chief Marketing Officer) likely need to view information from an organizational perspective.

Likewise, external vendors or technical experts will use the information differently than the other stakeholders, even if they see the same type of visualization. As an example, if your organization works with a vendor to provide bike parts, that vendor doesn't need to know your annual revenues, but it will need to know the expectation you have for bike parts. This information helps the vendor ensure they have all the necessary parts ordered for fulfillment.

Review Activity:

Audience Needs When Developing a Report

Answer the following questions:

1. **The people who are in your audience would also be considered what?**

2. **C-level executives, management, external vendors/stakeholders, the general public, and technical experts are all types of what?**

3. **Why is it important to consider who will be the users of a report?**

Topic 13B

Describe Data Source Considerations for Reporting

EXAM OBJECTIVES COVERED

4.3 Given a scenario, use appropriate methods for dashboard development.

When you are performing a data-related project, you will need to determine the source of the data needed for reporting, as well as plan for how the data will be viewed, described, and communicated through reporting. These considerations are key to forming a project plan, establishing and meeting business requirements, and developing reports and visuals that will be shared with key stakeholders in your organization.

Documenting the Source Data

When you are visualizing and reporting on data, you will want to make sure you first identify the source of your data. Remember that not everyone will have the same level of data source access that you do as an analyst. At least some of the people in your audience will lack access to the direct source of the data used in a visualization. Because these people do not have equal access to the data, they cannot freely run the report.

In these cases, you will have to develop a new source for your data project from the original source data. This way, others can be given access to the data they need. When you do this, you will need to note the source data for your project. This can be documentation that you add onto the report itself or include in an appendix. It could even be as simple as a "read me" file, which is just a notepad file located in the same folder as your project. As an example, suppose you are writing a report that shows new hires from the HR system and their corresponding payroll information from the payroll system. As an analyst, you have access to both source software systems, and you create a query that combines the data from both systems to create the report. You should document that the source of the report is the query you created and the source data for the query are the two software systems (HR and payroll).

Not all data that we use for reporting and visualization is actually stored in the database, so we write many functions and formulas to accomplish our projects. When you write any form of calculation on your data, you want to be sure to document that not only for yourself but for every consumer of the report. For example, if you have created a field to calculate a commission amount based on a change to the rates, the documentation that supports the project should include the calculation, why it was created, and what it does. This is especially important when you write logical functions, so that consumers know what logic is being applied and why.

Why We Document Data Sources

Here are just a few of many reasons why we document our data sources.

- *To account for changes in role and prepare others who might take over the work.* For example, if you are promoted, you will want the data analyst who steps into your role to have adequate information to support the reports and projects you have already developed.
- *To expedite the process of resolving issues.* When issues arise in the data that must be addressed with proper documentation, the process to identify the issue can begin immediately when we have documented the source. Without documentation, we are required to reverse engineer the project to determine where the source data came from.
- *To facilitate a return to the project.* For example, we may want to add an enhancement to a report after a large amount of time has passed, and the source documentation will streamline that process.
- *To answer questions about where the data came from.* When your report is presented in a meeting, people will undoubtedly ask where this data came from. You don't want to be put on the spot, searching for an answer; the source documentation makes it easy to provide this information when asked.

Determining Access to Data

As a data analyst, you will often have greater access to data than a typical staff member of the organization. You will need to provide users with just the data that they need, rather than all the data available, which some individuals may not have permission to access.

Gaining Approval/Approval Granted

Before you even begin reporting on data, it is important to first gain approval to design the dashboard or report. You must also gain approval to give that dashboard or report to others to use. The approval process varies between organizations, and in some cases even between departments. You might need to go through multiple levels of approval depending on your organization, but you will likely need to speak with your manager, and the managers of those who will need access, to start this process. The Information Technology (IT) department may specifically grant the approval to the data that feeds your dashboard or set of reports.

Reporting with Varied Permissions and Access

You will provide data to other users through reporting in several ways depending on their permissions and access to the data. It may be that they are given permissions to the database, or you might need to develop a way to give people data for their reporting without them having full permissions. For example, if your report has direct access to an SQL database, when users access your information they must also have access to the SQL database. As an analyst, when you are describing the data source considerations you will want to be familiar with your organization's process of giving users access to data, as these types of permissions can vary depending on the data/system/technology at any given organization.

However, there are also ways to give users access to your data without requiring direct access through the organization's systems. To demonstrate, let's continue with our example of confidential payroll data. Suppose that a department manager needs to see salary information for each of their employees, but managers are not allowed access to the source data that holds the information needed. To meet these requirements, you might create a view of the data that gives them just the data that they need. Or, you might automate a way to create a data set and export the report (e.g., using a simple CSV format).

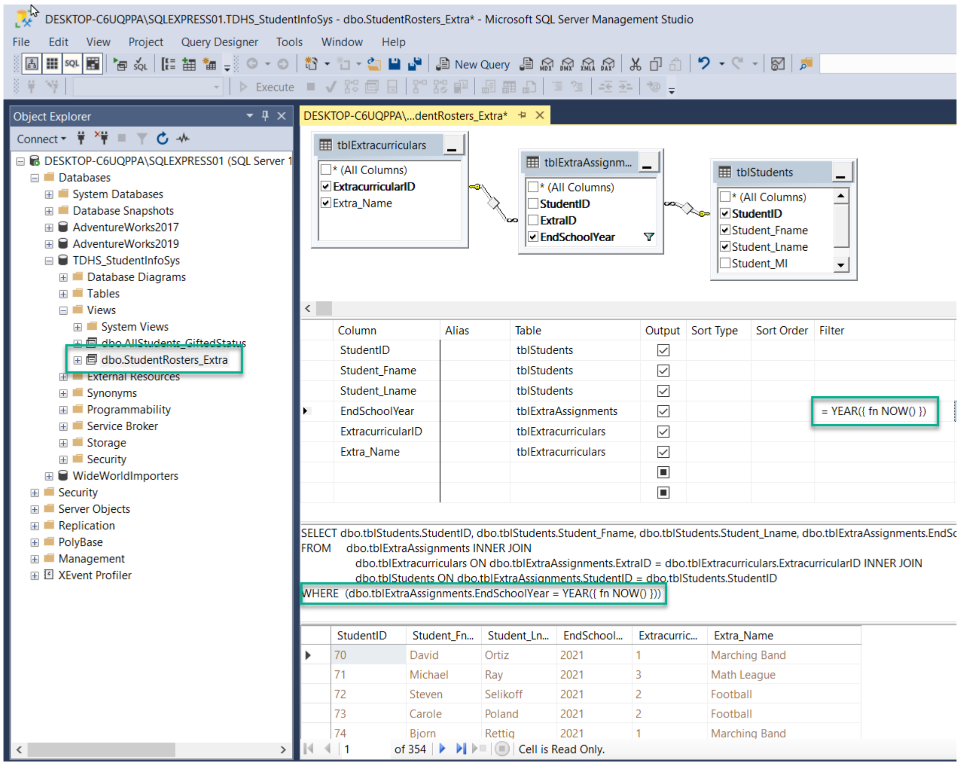
Determining what the audience can access (and more importantly, what they cannot access) is an important step in developing overall business requirements for data, as it will affect how you present data and reporting to the organization.

Developing Views of the Data

The analyst often builds queries or views of the data that provide information to reports or dashboards. These views allow you to join the data and effectively pull in just the information that you need from each table. These views then serve as the source of the reports and/or visualizations.

Any time we organize tables and/or query and join them, we are building a data model. A **data model** organizes the data and the relationships of data elements so that the data is ready to use and meaningful for every user who needs a report. The person who accesses the report will not need to organize the data themselves.

Where you can build views will be determined by the permissions you have within your organization's data system. If you have access to either the database or the data warehouse, where you can create views that are available to all people who have access, then you might build your query (data model) and save it as View. You can then leverage the View when you use tools like Power BI or Tableau for reporting. Building this view eliminates the need to use the data prep tools in these applications, because you can model and filter the data beforehand.



Column	Alias	Table	Output	Sort Type	Sort Order	Filter
StudentID		tblStudents	☑			
Student_Fname		tblStudents	☑			
Student_Lname		tblStudents	☑			
EndSchoolYear		tblExtraAssignments	☑			= YEAR({ fn NOW() })
ExtracurricularID		tblExtracurriculars	☑			
Extra_Name		tblExtracurriculars	☑			
			■			
			■			

```
SELECT dbo.tblStudents.StudentID, dbo.tblStudents.Student_Fname, dbo.tblStudents.Student_Lname, dbo.tblExtraAssignments.EndSc
FROM    dbo.tblExtraAssignments INNER JOIN
            dbo.tblExtracurriculars ON dbo.tblExtraAssignments.ExtraID = dbo.tblExtracurriculars.ExtracurricularID INNER JOIN
            dbo.tblStudents ON dbo.tblExtraAssignments.StudentID = dbo.tblStudents.StudentID
WHERE  (dbo.tblExtraAssignments.EndSchoolYear = YEAR({ fn NOW() }))
```

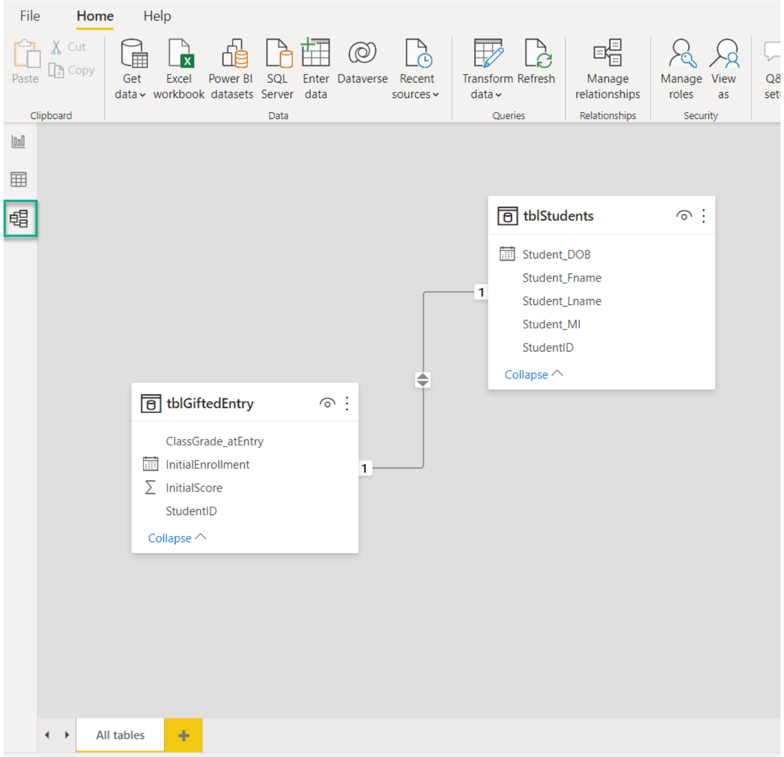
StudentID	Student_Fn...	Student_Ln...	EndSchool...	Extracurric...	Extra_Name
70	David	Ortiz	2021	1	Marching Band
71	Michael	Ray	2021	3	Math League
72	Steven	Selikoff	2021	2	Football
73	Carole	Poland	2021	2	Football
74	Bjorn	Rettig	2021	1	Marching Band

1 of 354 Cell is Read Only.

StudentRosters_Extra View in TDHS_StudentInfoSys SQL Server Management Studio (Used with permission from Microsoft.)

Assuming that you have permissions to create views in the database, you can hard code filters in your views. For example, you may want to only show data for the current year. You would hard code the appropriate date field, in this case end school year, indicating the current year for use as a filter. Then you would add the appropriate date type filter. We have used a function of Year (Now()) so that it looks at the current date and returns the current year to use for the filter. This ensures the data in your view is only current-year data.

You may find that you have access to all the tables you need, but no permissions to create a view inside the database. In this case, you will create a data model using a tool like Power BI. This model is built to bring in, join, and clean the data within Power BI.



Screenshot of Data Model of Students and Gifted Status from the TDHS_StudentInfoSys Database in Power BI Data Model View (Used with permission from Microsoft.)

Using the model view, which is the button on the left vertical bar that has tables related as an icon, we can see the data model that is built for this Power BI file. It displays the relationship between tblGiftedEntry and tblStudents, and it will also allow you to bring in other data sets as needed and add their relationships.

Regardless of the method for filtering the data with views, your goal is to narrow down the tables into just the information you need for the report. Using a hard-coded filter in the query allows the software you use to only process what data is needed up front.

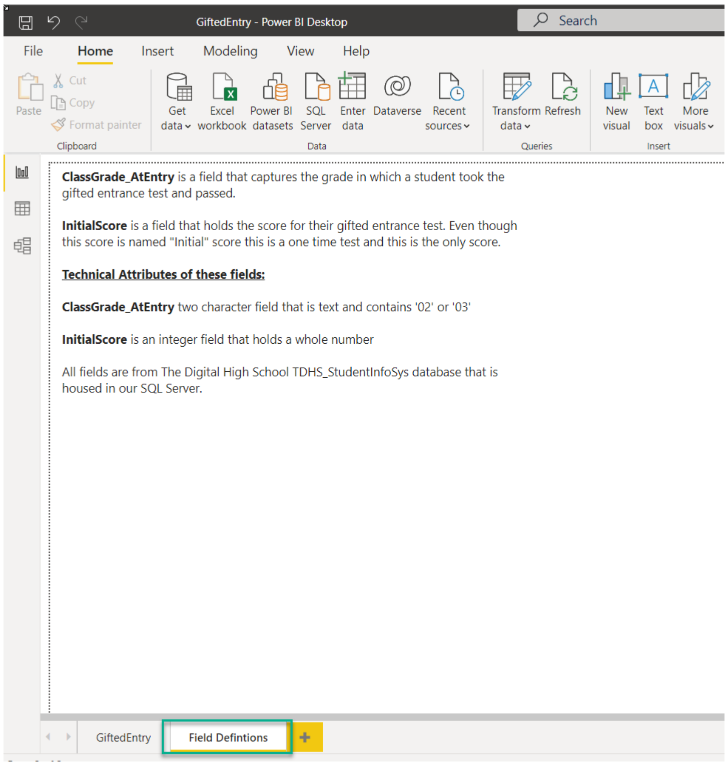
Data Fields and Attributes

As an analyst, when considering the source of the data you're using for reporting, you should also ensure that the data field names and attributes are set accordingly.

Field Definitions

You likely know a lot about the fields of data used in reporting, but others may not have the same experience working with that data. For example, while most people might recognize FNAME and LNAME as the first and last name fields, they probably don't know that the field containing Location, Warehouse, and Bin number is called LWB. For each data project, you should provide **field definitions** to clarify what information each field contains, as the names of the fields may not be readily apparent to everyone.

Providing proper field definitions for a report or dashboard is key to everyone productively working with the information and being able to answer their own questions about each data point they see without having to confirm their understanding of what the field means (or worse, operate on the wrong understanding). Including field definitions in the dashboard is one way that we can ensure the field information is available to the user without requiring them to open up any further emails or documents. When the dashboard is published, the page of information will be included.



ClassGrade_AtEntry is a field that captures the grade in which a student took the gifted entrance test and passed.

InitialScore is a field that holds the score for their gifted entrance test. Even though this score is named "Initial" score this is a one time test and this is the only score.

Technical Attributes of these fields:

ClassGrade_AtEntry two character field that is text and contains '02' or '03'

InitialScore is an integer field that holds a whole number

All fields are from The Digital High School TDHS_StudentInfoSys database that is housed in our SQL Server.

Example of Field Definitions as Page in Power BI
(Used with permission from Microsoft.)

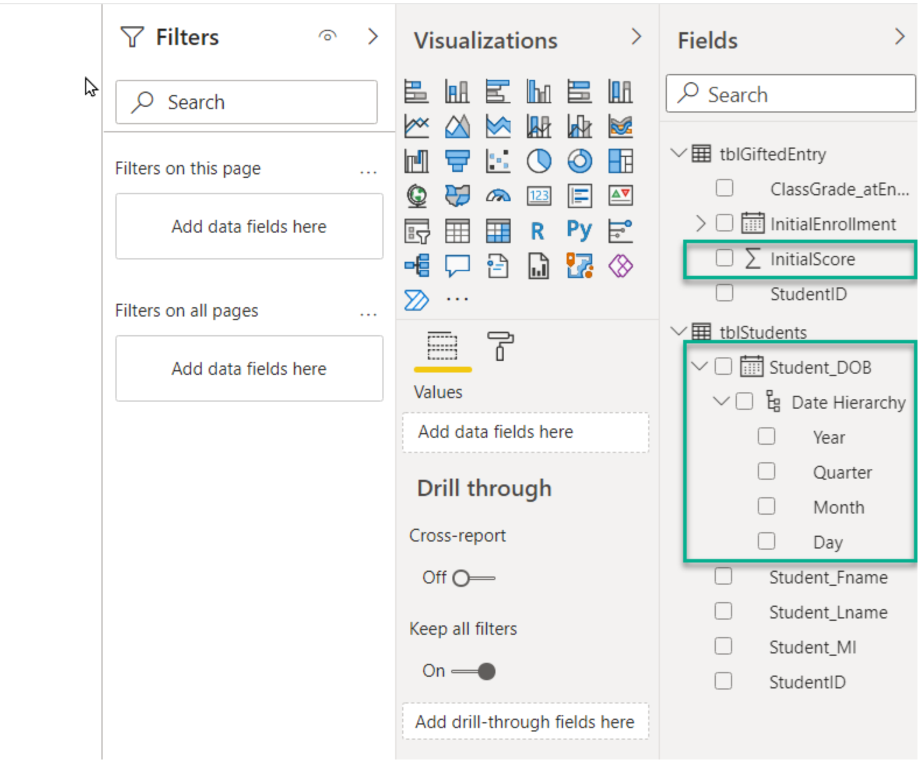
Let's look at an example. In our screenshot, the naming of the field "InitialScore" might imply that multiple scores exist, and this field is just the initial score. However, this is just poor naming of the actual field stored in the database, as there is only one field for scores. The data analyst can clarify this possible misconception in the field definitions. Each data analyst or organization may approach field definitions in a different way, but the goal is always to ensure that the user understands where the data comes from and what the field really contains.

Reading Field Attributes

In addition to providing proper field definitions, it's important to understand how visualization software will read fields based on their data types. *Dimensions* and *measures* are terms commonly used to describe attributes of data visualization software.

- Dimensions are attributes for categorical data that are used to label and provide meaningful insight about the data. Dimensions are typically the text-based values in our data sets, such as color, product name, or employee name.
- Measures are attributes for number-related values. These are the fields likely to be averaged, summed, or otherwise aggregated.
- Date fields may have an associated date hierarchy depending on the software you are working with, such as Power BI.

In the screenshot below, we have connected to The Digital High School Database, also referred to as TDHS_StudentInfoSys database, and also connected tblGiftedEntry and tblStudents into the data model.



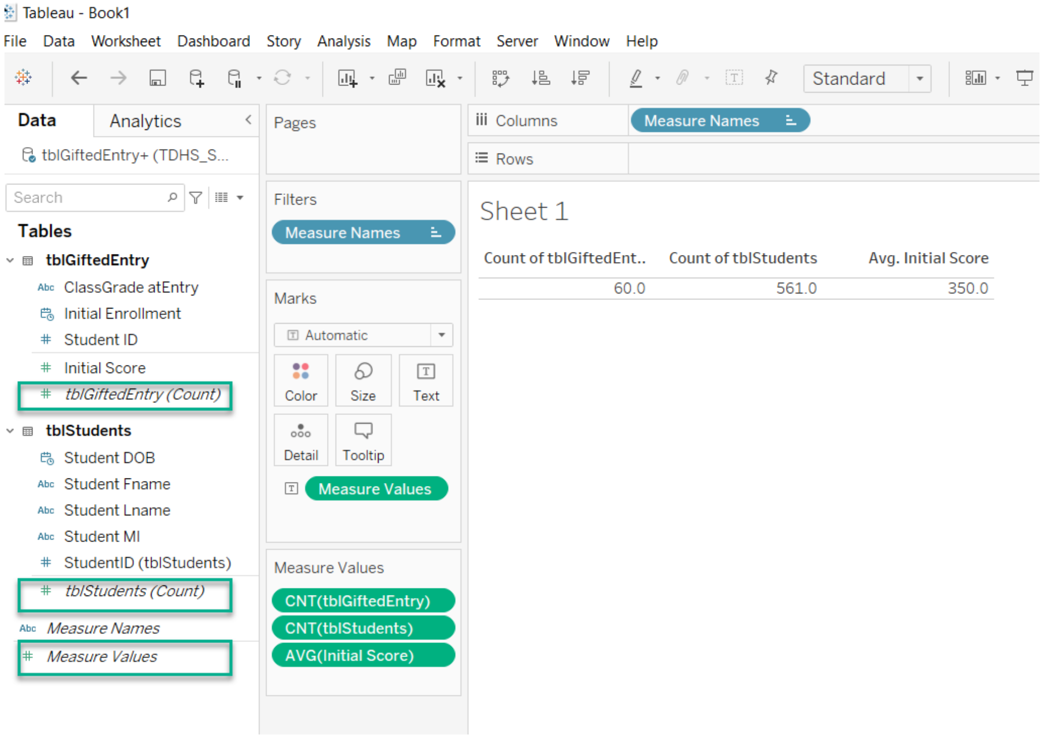
Cross-report: Off

Keep all filters: On

Power BI Connected to TDHS_StudentInfoSys Tables for Gifted Entry and Student Information (Used with permission from Microsoft.)

Power BI reads the InitialScore field as a value that can be aggregated and provides a date hierarchy for all the date fields. The other fields are read as dimension data. However, if needed, they can be converted to other types of aggregated information when appropriate.

Tableau will also read field information but will interpret it differently than Power BI, as you can see in the image below. This is the same database, and the same tables, tblGiftedEntry and tblStudents.



Count of tblGiftedEnt..	Count of tblStudents	Avg. Initial Score
60.0	561.0	350.0

Fields Displayed in Tableau Desktop from TDHS_StudentInfoSys (© 2021 Salesforce.com, Inc. All rights reserved. Used here with permission.)

Note that Tableau provides some initial aggregated fields for you to select. It also displays the data type for each of the fields in the data. The Tableau canvas displays Measure Names and Measure Values, which will take the fields that it sees as numbers and attempt to give you an aggregated view of these fields, if you choose to utilize them. They will default to sum or count, so be careful to adjust as appropriate. For our example, we applied the measure values but adjusted the IntialScore from Sum to Average.

Every system reads fields slightly differently and offers different features meant to enhance productivity.

Review Activity:

Data Source Considerations for Reporting

Answer the following questions:

1. **What organizes the data and the relationships of data elements so that the data is ready to use and meaningful for every user who needs a report?**

2. **If only part of a table from a database should be shared for the audience of a report, what can be created to only provide the permitted data for access?**

3. **The creation of what only allows the software in use to process the data that is needed up front?**

4. **What should be provided to help clarify the names of data fields?**

Topic 13C

Describe Considerations for Delivering Reports and Dashboards

EXAM OBJECTIVES COVERED

4.1 Given a scenario, translate business requirements to form a report.
4.3 Given a scenario, use appropriate methods for dashboard development.

When preparing a report or dashboard, you should understand how the data and visuals will be viewed or consumed, as this can affect your choice of software. You will also need to know how often the data will be delivered and how frequently people need updates to their reports. Reports that provide crucial data to an organization are commonly run on a recurring schedule.

Determining How Visuals Will Be Viewed

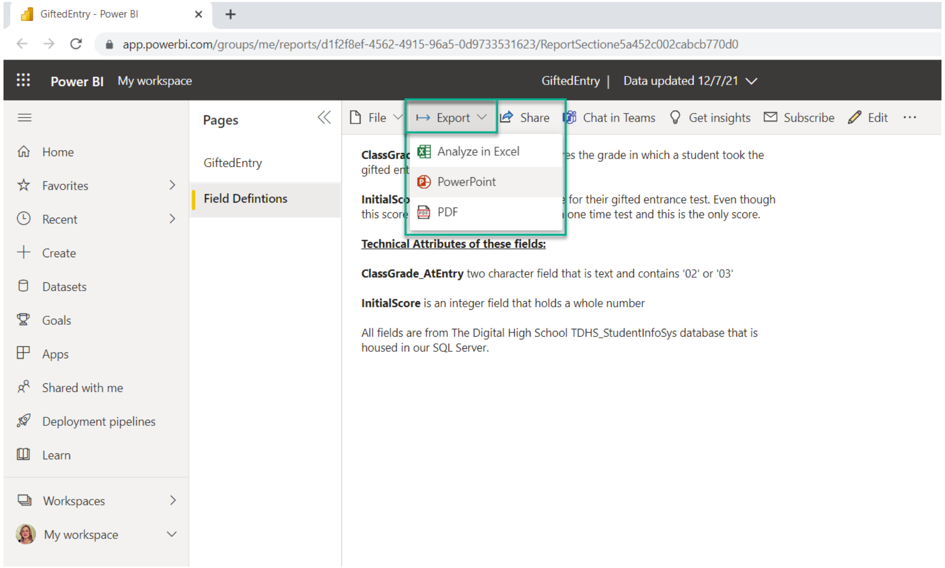
When building a report or dashboard, it's important to identify how the visuals will be viewed. There are many options for displaying reports and dashboards, including the following methods:

- Viewed via a web interface, like a website, or through an internal intranet site
- Converted to other software, such as presentation software like PowerPoint or Prezi
- Converted to a file format, such as a PDF
- Printed on paper

Determining the appropriate format for viewing reports and dashboards is an important consideration related to business requirements, because you will need to not only build the report or dashboard, but also export it to the proper format. For example, a visual may need to be included in an annual report or quarterly update that is drafted in Microsoft Word and then exported to a PDF.

If your report or dashboard is to be made available to a web interface, either the corporate intranet or the organization's public-facing website, it is important to gather the additional requirements needed. For example, you should know who is responsible for making the report or dashboard available on the company's website and what specifically they need from you, as the data analyst, to complete the task. It may require you to use a different software or gain additional permissions.

Most dashboard and reporting tools, such as Power BI, Crystal Reports, and others, can print and/or export to various formats. For example, when you publish a dashboard in Power BI, it immediately enables you to export to PowerPoint or PDF format. You should perform a test run of these exports as you refine your dashboard or report to ensure that the visuals translate as expected and are in a helpful format for people reviewing the information.



ClassGrac es the grade in which a student took the gifted ent

InitialSco for their gifted entrance test. Even though this score one time test and this is the only score.

Technical Attributes of these fields:

ClassGrade_AtEntry two character field that is text and contains '02' or '03'

InitialScore is an integer field that holds a whole number

All fields are from The Digital High School TDHS_StudentInfoSys database that is housed in our SQL Server.

Published Power BI File in Online Power BI Workspace with Export Commands (Used with permission from Microsoft.)

Power BI in our example allows you to publish to a web interface called My Workspace, and even to others if you have permissions. This allows you to have both a web version of your dashboard and also the available export commands to PowerPoint and PDF.

Determining How Data Will Be Delivered

In a business setting, data is delivered in a number of ways.

- A report, dashboard, or visual is presented in a meeting.
- Information is delivered as a part of a data section on the company website.
- Data is shared in the “report sharing” option of software applications like Power BI or Tableau, which have different environments companies can leverage to share reports across an organization.
- Data can be printed as a paper report and given at a monthly meeting so that consumers like managers, C-level executives, or technical experts can review and mark up information for their needs.
- Data can be subscribed to through the subscription options made available in a software.

The available options for delivering data to your audience depend on the tools and software used when developing your reports and dashboards. Many popular visualization tools allow you to build interactive reports through web-based environments. With these tools, consumers will often subscribe to the reports that they want, and can receive updates based on times they define. As an example, Salesforce, a popular CRM tool, has its own internal report builder and dashboard options. These features allow for interactive experiences directly inside the software environment. You can easily subscribe to any report that is available to you when you have access to the Salesforce Reports.

Some reports should be delivered on a set schedule (e.g., daily or weekly) as defined by business requirements. For example, a weekly reportmight be delivered as an Excel or PDF file via email. This type of delivery is a form of static information that people can access according to their schedule. Most enterprise report and dashboard software programs offer some form of scheduled delivery service.

The consideration of whether your data is a continuous/live feed or static data might drive some of your decisions for how that data will be delivered to users in your organization. If the data is only updated once a week by a data analyst who simply needs to refresh it, then the overhead of creating a subscription service or interactive dashboard experience might be a waste. However, if the data is live, active, and always up-to-date, then delivering it routinely or through a dashboard might be more effective for the audience than waiting each week to receive an update.

You must determine the delivery method(s) that will be used for your developed reports up front. The method needed can impact what tools you use, so you will definitely want to identify it as part of the requirements.

Frequency of Reporting

A key part of developing business requirements for data and reporting is determining how often reports should be updated and how often they should be sent. For example, some businesses send month-end reporting two days after the end of the month to ensure that all data for that month is represented. If you are developing a reporting schedule, be sure to consider the desired timing. For instance, if you want to send out reports as soon as new data is loaded, then write a reporting schedule according to that timing.

Even if it's difficult to control the timing of when data is updated, it is important for the audience to understand that their data is up-to-date as of a certain time, so that there is no misunderstanding about what data is contained within the reports. As we will cover in a later lesson, you can use the system functions to provide the print date and refresh date of your report or visualization: these would be considered key elements. To ensure that everyone in the distribution list knows what to expect, the analyst should document requirements that explain the frequency of reporting and also how often the data is updated.

Recurring Reports

Just like how many work teams have regular meetings that occur weekly, monthly, or even annually, reports can also be created on a schedule. A report that is set to repeatedly run on certain dates or at specific times is a **recurring report**. This schedule will include the following:

- The report to be sent
- The output of the report
- Who will receive the report

Reports that provide crucial information to the organization are likely to be run on a recurring basis. Recurring reports are delivered in different ways based on organizational policy and audience preferences. If you are reporting to key executives, for example, you may deliver the report in a print or PDF format (the process for which can also be recurring and sometimes automated).

You may find in the company policy that there are set procedures for exactly where reports can and cannot be stored. As organizations strive for efficiency, they may publish reports on web-based platforms and offer scheduled refreshes so that consumers can have updated data regularly. They may leverage shared drives, in platforms like SharePoint, so that users can access the information from multiple devices whenever it is needed. Along with the institution of these platforms comes IT and security policies intended to protect the information while still allowing the users to securely access data when needed.

Review Activity:

Considerations for Delivering Reports and Dashboards

Answer the following questions:

1. **What key point must be considered when preparing to create the visuals of a report or dashboard?**

2. **What are two methods we might use to publish Power BI reports?**

3. **What limits the delivery options of a report or dashboard?**

4. **When planning the delivery of a report, we must consider how the dashboard will be delivered, and what else?**

5. **Recurring reports can be delivered in multiple ways depending on the organization's policies. What are a few ways these reports are delivered?**

Topic 13D

Develop Reports or Dashboards

EXAM OBJECTIVES COVERED

4.1 Given a scenario, translate business requirements to form a report.
4.3 Given a scenario, use appropriate methods for dashboard development.

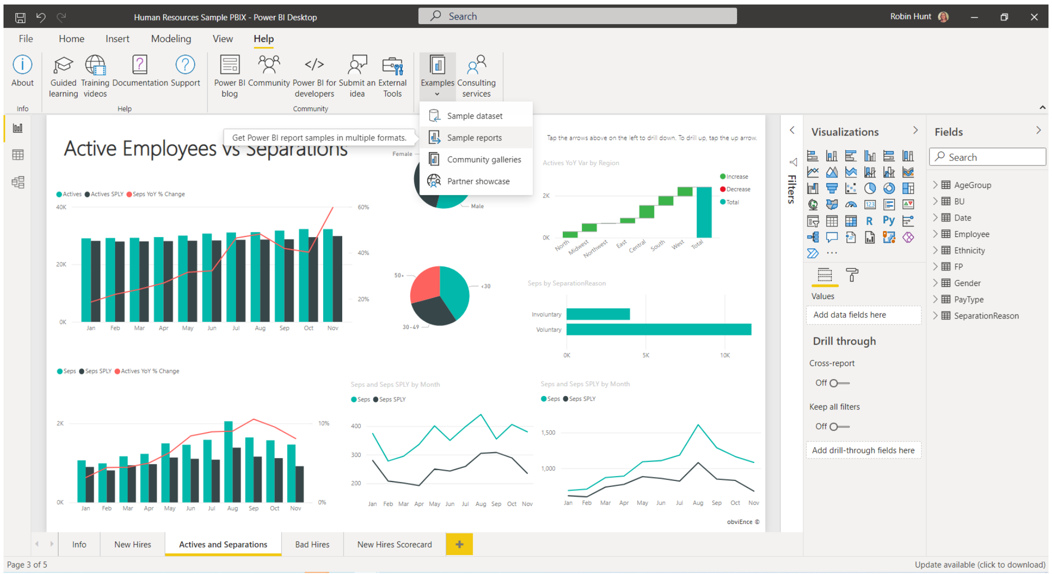
There are three main reporting types that you will work with as an analyst: dashboards, paginated reports, and spreadsheets. Each of these styles requires a slightly different approach during the design process. In this lesson, you will learn about these three main reporting types and how to approach design using mockup and wireframing. You will also discover the different types of visuals and suggestions for navigation.

Visualization Layouts

Data analysts work with three main types of reporting: dashboards, paginated reports, and spreadsheets.

Dashboards

A **dashboard** is an interactive, visual display of information. Dashboards are built for screen viewing and can be designed for mobile devices, tablets, and monitors of all sizes. Dashboards should be designed with visuals that make sense when viewed together on the screen. These visuals are meant to give high-level insight about the data and can then be used to further investigate information. The screenshot that follows shows a popular free sample of a data dashboard in Power BI. You can download any of their samples from the Examples option in the Help tab.

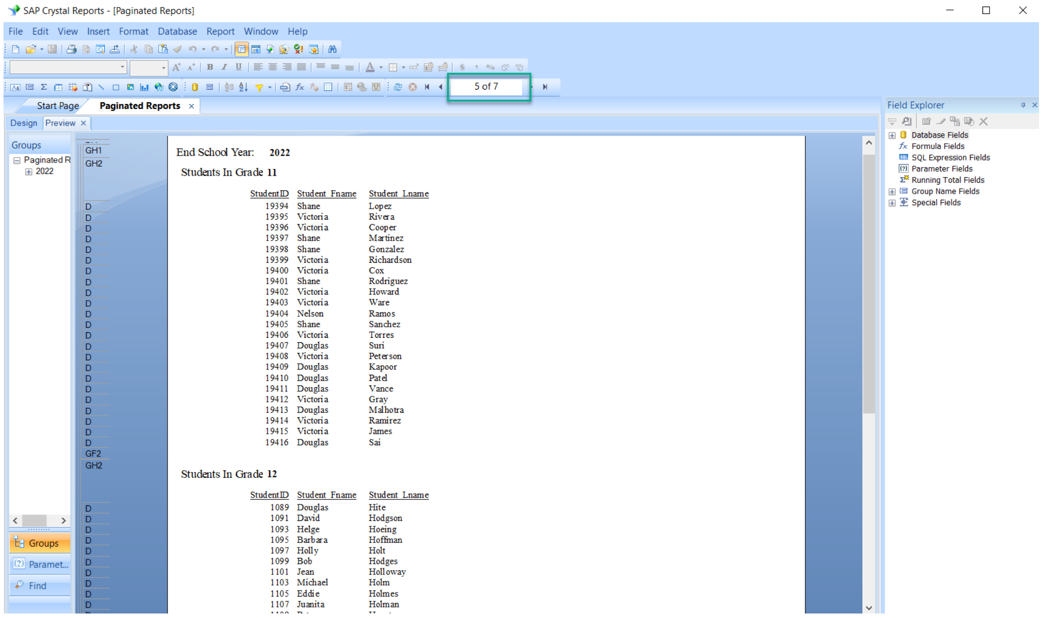


Screenshot of Sample HR Dashboard in Power BI (Used with permission from Microsoft.)

Dashboards can be designed for a specific process or department. They can be very impactful to an organization as they can not only present information, but also provide an ability to interact with the data through visuals, filters, and slicers.

Paginated Reports

A **paginated report** is a multipage report that is not suitable for display on a dashboard. Anytime you need more than a single page of information, tools like Power BI Report Builder or Crystal Reports can be helpful in building paginated reports. These are reports that have lines of information and carry over pages, like in our example below where we have seven pages of student roster information.



End School Year: **2022**

Students In Grade **11**

StudentID	Student_Fname	Student_Lname
19394	Shane	Lopez
19395	Victoria	Rivera
19396	Victoria	Cooper
19397	Shane	Martinez
19398	Shane	Gonzalez
19399	Victoria	Richardson
19400	Victoria	Cox
19401	Shane	Rodriguez
19402	Victoria	Howard
19403	Victoria	Ware
19404	Nelson	Ramos
19405	Shane	Sanchez
19406	Victoria	Torres
19407	Douglas	Suri
19408	Victoria	Peterson
19409	Douglas	Kapoor
19410	Douglas	Patel
19411	Douglas	Vance
19412	Victoria	Gray
19413	Douglas	Malhotra
19414	Victoria	Ramirez
19415	Victoria	James
19416	Douglas	Sai

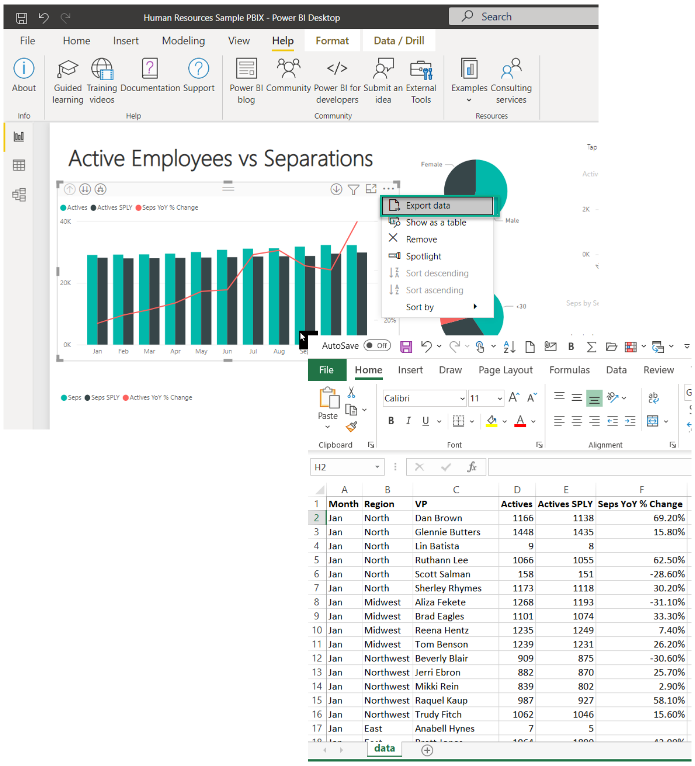
Students In Grade **12**

StudentID	Student_Fname	Student_Lname
1089	Douglas	Hite
1091	David	Hodgson
1093	Helge	Hoeing
1095	Barbara	Hoffman
1097	Holly	Holt
1099	Bob	Hodges
1101	Jean	Holloway
1103	Michael	Holm
1105	Eddie	Holmes
1107	Juanita	Holman

Screenshot of Student Rosters in Crystal Report over Multiple Pages (Copyright © 2021 SAP SE or an SAP affiliate company. All rights reserved.)

Spreadsheets

A **spreadsheet** is a worksheet of data in tabular form. Spreadsheets are an ideal tool for people in your organization who need to export and work with data as part of their roles. As an analyst, you can provide insight on how to take data from a dashboard and export it to a spreadsheet. This gives a person the flexibility to work with the data, such as by creating pivot tables and various other reports. Spreadsheets are often helpful when data and reports are needed for a specific job task or scenario, but they are limiting when used for organizational reporting.



	A	B	C	D	E	F
1	Month	Region	VP	Actives	Actives SPLY	Seps YoY % Change
2	Jan	North	Dan Brown	1166	1138	69.20%
3	Jan	North	Glennie Butters	1448	1435	15.80%
4	Jan	North	Lin Batista	9	8	
5	Jan	North	Ruthann Lee	1066	1055	62.50%
6	Jan	North	Scott Salman	158	151	-28.60%
7	Jan	North	Sherley Rhymes	1173	1118	30.20%
8	Jan	Midwest	Aliza Fekete	1268	1193	-31.10%
9	Jan	Midwest	Brad Eagles	1101	1074	33.30%
10	Jan	Midwest	Reena Hentz	1235	1249	7.40%
11	Jan	Midwest	Tom Benson	1239	1231	26.20%
12	Jan	Northwest	Beverly Blair	909	875	-30.60%
13	Jan	Northwest	Jerri Ebron	882	870	25.70%
14	Jan	Northwest	Mikki Rein	839	802	2.90%
15	Jan	Northwest	Raquel Kaup	987	927	58.10%
16	Jan	Northwest	Trudy Fitch	1062	1046	15.60%
17	Jan	East	Anabell Hynes	7	5	

Screenshot of Export Data from Power BI HR Sample Dashboard (Used with permission from Microsoft.)

In this image, we actually generated the .csv file from the dashboard to produce the spreadsheet. Whether the spreadsheet comes from a visual or not, as a data analyst you will find many spreadsheets are being used for everyday reporting.

Mockup and Wireframing for Design

It is never too early in the design process to create a mockup or wireframe. **Mockup** simply means to draw out a potential layout. This can happen on a piece of paper or inside a software, like Snagit, where you can draw different screens and add objects. **Wireframing** is creating mockups of multiple screens that are likely connected. You can do this with paper and pen, or you can use tools like Figma that allow you to draw the screens and add interactions.

These processes are commonly used in user experience (UX) design and user interface (UI) design. The purpose is twofold, in that you are forced to consider how the business user will be able to interact with the information and the data story overall.

One of the benefits of mockups and wireframing is that they provide valuable information about the potential user experience of the reporting package to be delivered. They help people understand how the data may look, allowing them to provide you with feedback and better information on what they may want to see in the visuals and reports. This process also aids in building proper navigation across multiple dashboards. Mockups and wireframing may feel like a lot of up-front work at the beginning of any project, but it will be a massive time saver in the long run.

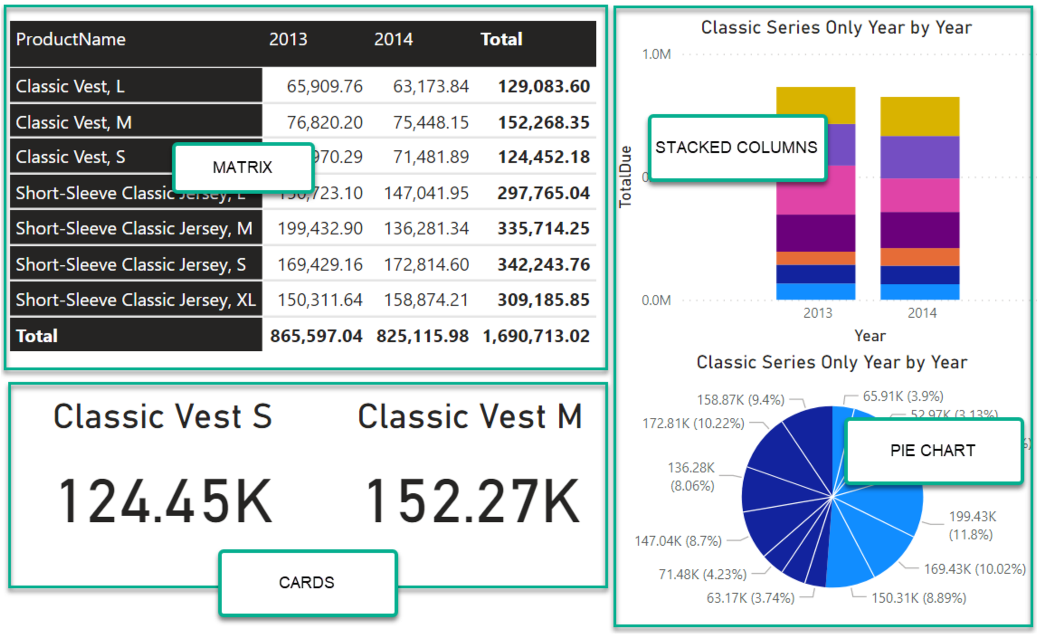
Types of Visuals

Many different types of visuals can be used to display data and express the underlying table. As an analyst, you will want to work with your decision makers to define what they need from the data sets and determine the best visuals to display the outcomes for the report or dashboard.

Just because you can create a type of visual from data doesn't make it the correct visual for that report or dashboard. When you begin the design process, it is important to discuss various ways that data can be displayed and to understand how people want to see the information laid out visually.

Several iterations of design typically will occur in a data project to visualize data. It's good to know what types of visuals commonly make it onto dashboards.

- **Table**—Displays columns and rows of information, and may also include aggregate information. A table is great for information that simply needs to be subtotaled, or for the purpose of showing the potential values of any dimension in a list format.
- **Matrix**—This display is common in data analytics and can also be referred to as a pivot or cross tab format. A matrix shows values at the cross section of a row and column. This display is great for showing multiple dimensions and aggregated values.
- **Graphical visuals (pies, bars, stacked)**—Varied graphs and charts are used to visualize, break down, and otherwise summarize data, and are very common in dashboards.
- **Cards**—Typically used for totals or aggregated values. Cards are an easy way to display critical values on the dashboard in just a text-based form, without the visual aspect of a graph or chart.



ProductName	2013	2014	Total
Classic Vest, L	65,909.76	63,173.84	**129,083.60**
Classic Vest, M	76,820.20	75,448.15	**152,268.35**
Classic Vest, S	[illegible]970.29	71,481.89	**124,452.18**
Short-Sleeve Classic [illegible]	[illegible],723.10	147,041.95	**297,765.04**
Short-Sleeve Classic Jersey, M	199,432.90	136,281.34	**335,714.25**
Short-Sleeve Classic Jersey, S	169,429.16	172,814.60	**342,243.76**
Short-Sleeve Classic Jersey, XL	150,311.64	158,874.21	**309,185.85**
Total	**865,597.04**	**825,115.98**	**1,690,713.02**

Screenshot Mockup of Table, Matrix, Graphics, and Cards in Power BI (Used with permission from Microsoft.)

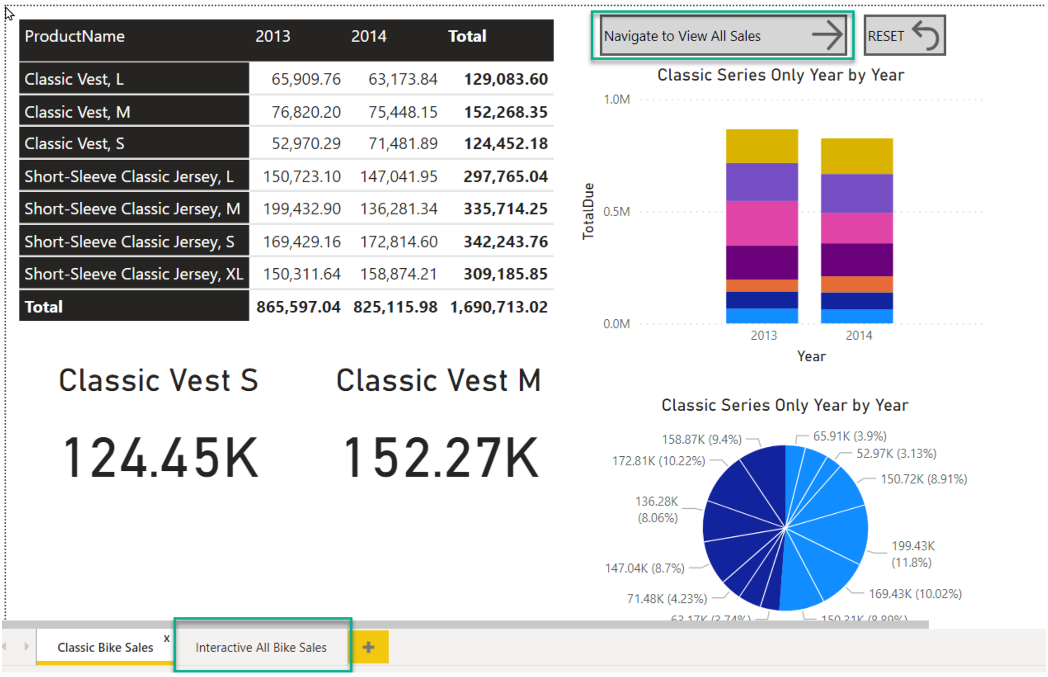
In this screenshot, we have a matrix that gives us product values from our AdventureWorks sample data. We also have standard visuals, like stacked column and pie charts, and then cards. These cards provide just the high-level totals for the information we have specified.

Types of Dashboard Navigation

Dashboards should make the proper navigation tools available for users. For example, users may need to navigate between dashboards, or they may need to refresh or even reset a dashboard.

Navigating to Other Dashboards

When providing a series of dashboards, it's important that consumers have an easy way to navigate from one dashboard to another. Each software handles this navigation differently. When developing business requirements, you might want to determine what dashboards make sense to share together, so the consumer can explore them without having to think about where to access them. This is where the roles of the data analyst and UX designer merge. You must always think about how the consumer will be using the dashboard in order to make it as simple as possible for them to get to the information they need without having to do a lot of digging and searching.



ProductName	2013	2014	Total
Classic Vest, L	65,909.76	63,173.84	**129,083.60**
Classic Vest, M	76,820.20	75,448.15	**152,268.35**
Classic Vest, S	52,970.29	71,481.89	**124,452.18**
Short-Sleeve Classic Jersey, L	150,723.10	147,041.95	**297,765.04**
Short-Sleeve Classic Jersey, M	199,432.90	136,281.34	**335,714.25**
Short-Sleeve Classic Jersey, S	169,429.16	172,814.60	**342,243.76**
Short-Sleeve Classic Jersey, XL	150,311.64	158,874.21	**309,185.85**
Total	**865,597.04**	**825,115.98**	**1,690,713.02**

Classic Vest S

124.45K

Classic Vest M

152.27K

Dashboard in Power BI with Bookmarks for Navigation and Filter Resets (Used with permission from Microsoft.)

Power BI allows you to establish bookmarks to clear your filters and navigate to other areas of the dashboard. You will see on the screenshot the option to navigate to view all sales, which will navigate the users to the interactive "all bike sales" page in the dashboard.

Refreshing Data

You can build your dashboard to automatically refresh. This is a great method when you have real-time access to data. In addition, you might provide a visible button that allows consumers who are on the dashboard to refresh on demand.

Resetting Filters

If you look at our earlier screenshot from Power BI, you should notice there is also an option to reset. In the software, you can use filters and create bookmarks that allow the user to select the bookmark button and apply the filter. This reset button allows them to then reset any filter that has been applied.

Review Activity:

Develop Reports or Dashboards

Answer the following questions:

1. **What are the three main types of reporting?**

2. **What visual is also known as a pivot or cross tab format?**

3. **What is defined as creating mockups of multiple screens that are likely connected?**

4. **When should a report be set to automatically refresh?**

Topic 13E

Understand Ways to Sort and Filter Data

EXAM OBJECTIVES COVERED

4.1 Given a scenario, translate business requirements to form a report.
4.3 Given a scenario, use appropriate methods for dashboard development.

As you have discovered, when working with databases and tables of information, every single field is captured, whether or not all the data is needed for reporting. Most reports or dashboards are filtered in some way to restrict access to just the needed fields, allowing users to interact with the dashboard and select only the filters they need at that time. As a natural part of data analyst work, you will deal with data that needs to be sorted using various sort commands and methods.

When developing business requirements, you must consider what types of data might be filtered and what sorts might be applicable up front. You'll need to determine where it makes sense to use hard-coded filters, interactive filters, and visuals as filters, and what the appropriate sorts are.

Another common way of filtering data is by date. Businesses often rely on dates in reports when making decisions, so the ability to filter by date range is key. The analyst may also need to create date tables to link data with dates. Be sure to ask what dates are applicable for the reporting project, so you can figure out how to sort them for the best user experience.

Sorting Data

Sorting data is a technique that is commonly deployed by the data analyst. Data is typically in **natural order**, meaning the order in which the data is entered. This is often not effective for the displays of the report. The most common sort method is **ascending and descending order**, meaning A–Z or Z–A for text-based fields. For numerical data, ascending is the minimum number sorted to the top and the maximum number at the bottom, with descending order being the reverse of that. While this is the most common method, you will also encounter more advanced sorting methods.

Advanced Sorting Methods

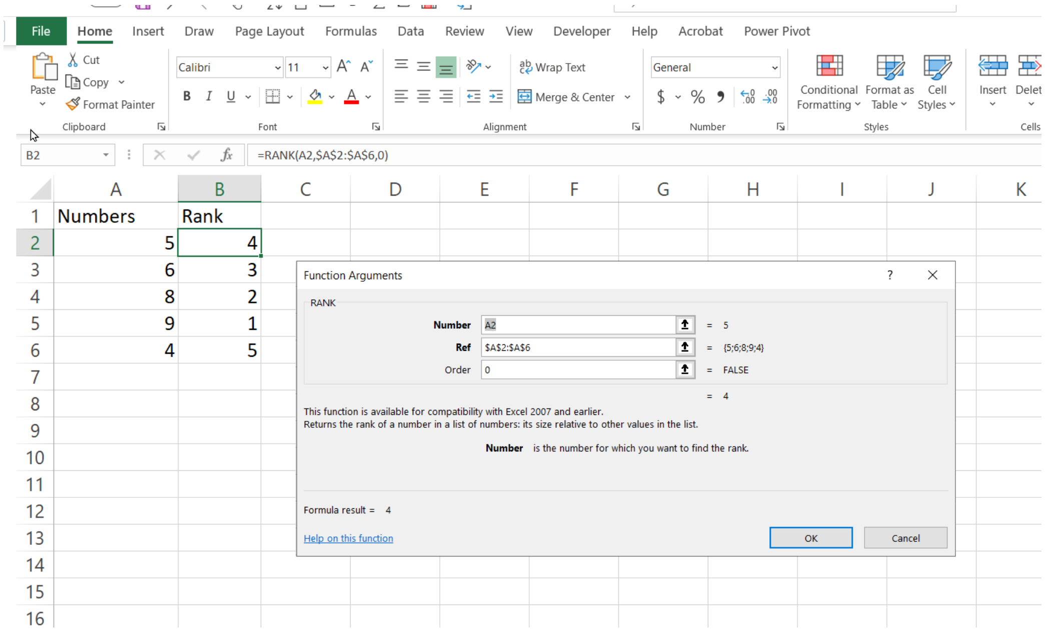
Multi-sorting is a method of sorting where you sort within a sort. Imagine that you are reviewing the phone list. It has each last name sorted, then within the last name it sorts the first name. This helps to display all the last names, as example Smith together, and then sorts the first names A to Z. Not all data is appropriate for multiple sorting, but when you have groups of data you can easily sort them using the multi-sort options that are available in all data tools.

Top N and Bottom N sorts sort the data to display the top or bottom portions of data in a set based on how many values exist in the set and the number you specify as the portion, or N. For example, if you want to view the top 25 most-ordered products, you would use the function built into the data software to create the Top N sort, which will filter your data to the specified top 25 values.

Custom sorts are where you create the data set to include the value and the sort order you need for your visualization. This is great for nominal data that does not have a sort, but you need to create one. A great example is education from high school, some college, bachelor's degree, and graduate degrees. If you sort this data by the first letter of the education, then it would be "out of order" from how you obtain the education. Traditionally you'd attend high school first before any other type of education. You can easily create the category list and apply the appropriate sort order. Then you would use that table to sort the information into your data model.

There are also available sort functions, like Rank and Order. These functions are sort related, but they both supply different information to the set, and then you could sort by them if needed.

In our image below, we see the numbers in column A are in natural order and we have used the RANK function in Excel, and the order is 0 because we want the largest number to be ranked the highest for this example.



B2 =RANK(A2,$A$2:$A$6,0)

	A	B
1	Numbers	Rank
2	5	4
3	6	3
4	8	2
5	9	1
6	4	5

Function Arguments

RANK

Number	A2	= 5
Ref	$A$2:$A$6	= {5;6;8;9;4}
Order	0	= FALSE

= 4

This function is available for compatibility with Excel 2007 and earlier.
Returns the rank of a number in a list of numbers: its size relative to other values in the list.

Number is the number for which you want to find the rank.

Formula result = 4

Help on this function

OK Cancel

Rank Function in Microsoft Excel (Used with permission from Microsoft.)

In our example, you see that of the five available numbers the number 9 which is the highest value is also ranked as number 1. Number 4 which is our lowest value is ranked number 5.

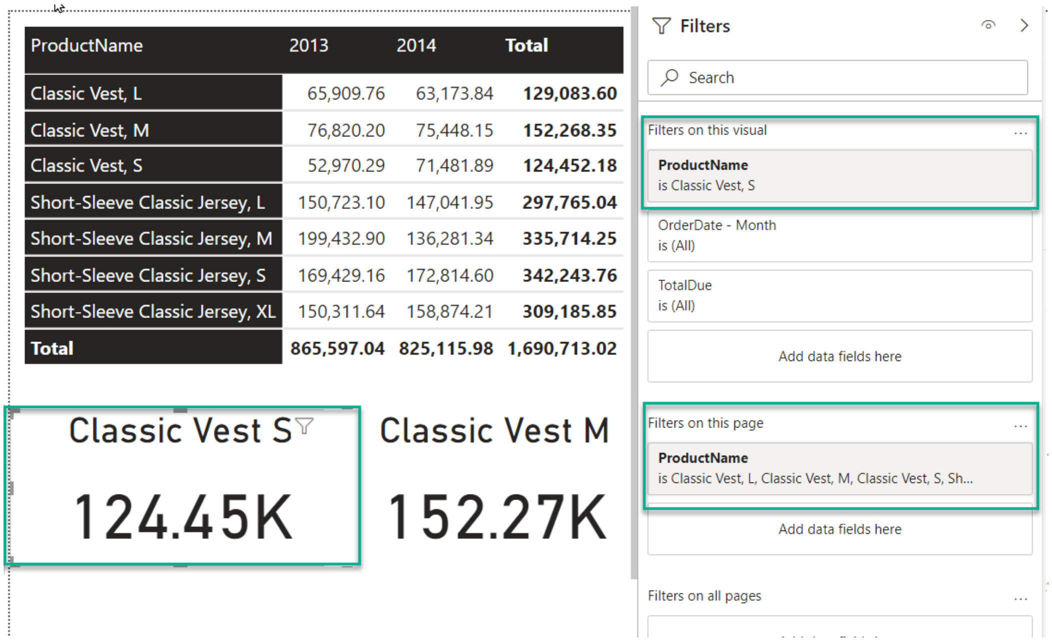
As a data analyst there are benefits to leveraging the various sorts for the data you are presenting. Carefully consider what sort you need and be sure to note this information so that others will know what types of sort method has been supplied. Remember that Top N sorts also filter so you will want to be sure that you don't exclude data when you need the entire set.

Filter Methods for Visuals

As part of developing the business requirements for data, it is important to determine where filters will occur and if these filters are hard coded or interactive.

Applying Hard-Coded Filters on Visuals

Hard-coded filters are coded into the view or the visual. The filters are automatically applied to the visual; the user does not adjust them. Most visualization software programs have filters that can hard code a visual, page, or the entire report by whatever field and condition you specify. The following screenshot shows page filters in Power BI.



ProductName	2013	2014	Total
Classic Vest, L	65,909.76	63,173.84	**129,083.60**
Classic Vest, M	76,820.20	75,448.15	**152,268.35**
Classic Vest, S	52,970.29	71,481.89	**124,452.18**
Short-Sleeve Classic Jersey, L	150,723.10	147,041.95	**297,765.04**
Short-Sleeve Classic Jersey, M	199,432.90	136,281.34	**335,714.25**
Short-Sleeve Classic Jersey, S	169,429.16	172,814.60	**342,243.76**
Short-Sleeve Classic Jersey, XL	150,311.64	158,874.21	**309,185.85**
Total	**865,597.04**	**825,115.98**	**1,690,713.02**

Classic Vest S

124.45K

Classic Vest M

152.27K

Filters

Search

Filters on this visual

ProductName
is Classic Vest, S

OrderDate - Month
is (All)

TotalDue
is (All)

Add data fields here

Filters on this page

ProductName
is Classic Vest, L, Classic Vest, M, Classic Vest, S, Sh...

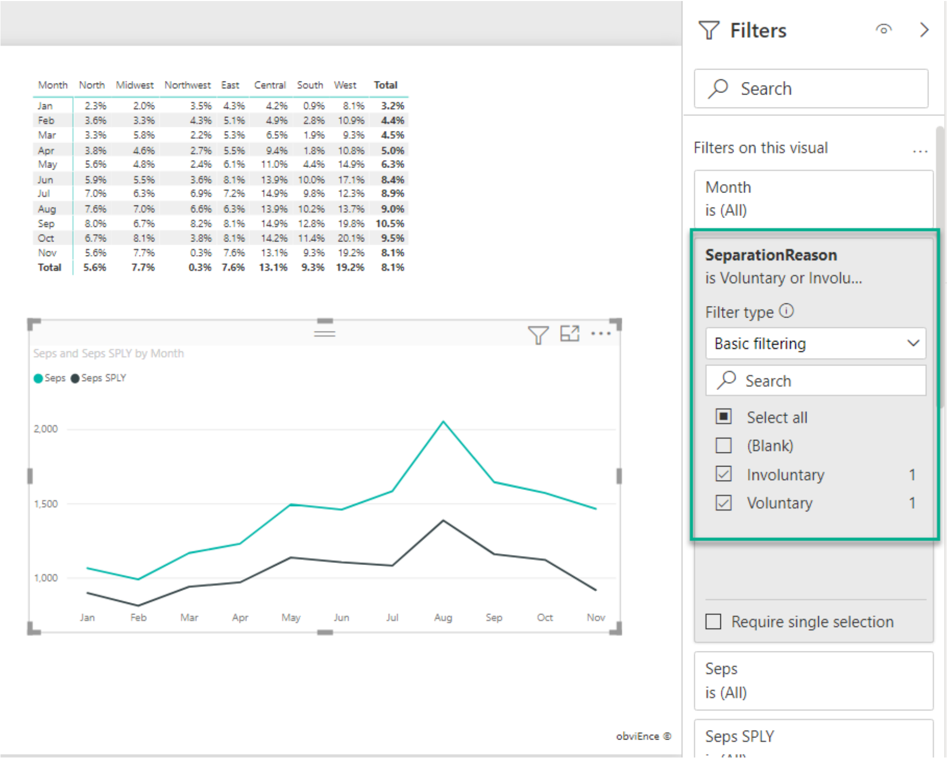
Add data fields here

Filters on all pages

Power BI Visual Filters and Page Filters (Used with permission from Microsoft.)

In our image, you will note that the Classic Vest S card is set with a product name filter on the visual to ensure that it stays as Classic Vest S and only shows that total. Note that there are also page filters that apply to the entire dashboard so it only shows these specific products, and all other products are excluded.

Hard-coded filters can be useful in many business scenarios. For example, you could hard code a manager's report so that it only shows the employees listed as being under that manager's supervision. You might create a hard-coded page filter so that all the visuals on that dashboard are automatically filtered by that manager. In another example, you might use a hard-coded filter that does not show null or blank values in the report. A variety of hard-coded filters can be applied to meet reporting and business needs within your organization.



Month	North	Midwest	Northwest	East	Central	South	West	**Total**
Jan	2.3%	2.0%	3.5%	4.3%	4.2%	0.9%	8.1%	**3.2%**
Feb	3.6%	3.3%	4.3%	5.1%	4.9%	2.8%	10.9%	**4.4%**
Mar	3.3%	5.8%	2.2%	5.3%	6.5%	1.9%	9.3%	**4.5%**
Apr	3.8%	4.6%	2.7%	5.5%	9.4%	1.8%	10.8%	**5.0%**
May	5.6%	4.8%	2.4%	6.1%	11.0%	4.4%	14.9%	**6.3%**
Jun	5.9%	5.5%	3.6%	8.1%	13.9%	10.0%	17.1%	**8.4%**
Jul	7.0%	6.3%	6.9%	7.2%	14.9%	9.8%	12.3%	**8.9%**
Aug	7.6%	7.0%	6.6%	6.3%	13.9%	10.2%	13.7%	**9.0%**
Sep	8.0%	6.7%	8.2%	8.1%	14.9%	12.8%	19.8%	**10.5%**
Oct	6.7%	8.1%	3.8%	8.1%	14.2%	11.4%	20.1%	**9.5%**
Nov	5.6%	7.7%	0.3%	7.6%	13.1%	9.3%	19.2%	**8.1%**
Total	**5.6%**	**7.7%**	**0.3%**	**7.6%**	**13.1%**	**9.3%**	**19.2%**	**8.1%**

Seps and Seps SPLY by Month

Filters on this visual

Month
is (All)

SeparationReason
is Voluntary or Involu...

Filter type ⓘ

Basic filtering

Search

▣ Select all
□ (Blank)
☑ Involuntary 1
☑ Voluntary 1

□ Require single selection

Seps
is (All)

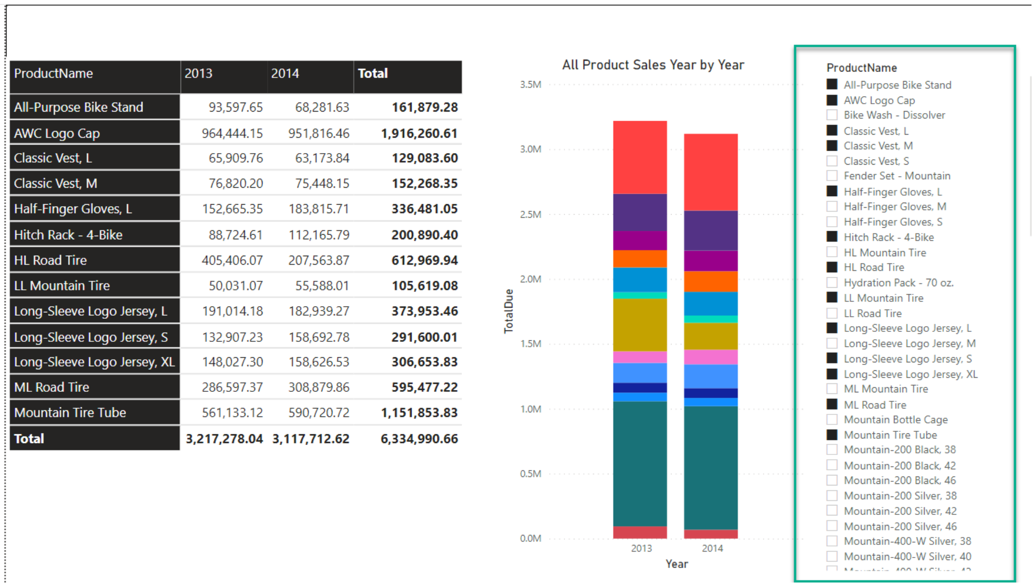
Seps SPLY

Visual Filter that Does Not Show Blanks Using Power BI Sample HR Report
(Used with permission from Microsoft.)

In this example, the visual is hard coded with a filter to not show any blank or null values. We do this by selecting that visual and checking only the items we want.

Applying Interactive Filters to Visuals

Interactive filters are filters that allow the consumer to adjust a slicer or filter option on a dashboard to narrow down the data they want to see. As an example of when interactive filters can be useful in business, consider a sales organization that has multiple products, and each salesperson has access to view the sales of every product. The company wants the salespeople to be able to choose the products that they are interested in viewing. Each salesperson at the company needs to be able to see different products, either one at a time or many at one time, so they can investigate issues or highlight key information about that product. In this case, a slicer or filter option would allow the salesperson to filter the data to their specifications.



ProductName	2013	2014	Total
All-Purpose Bike Stand	93,597.65	68,281.63	**161,879.28**
AWC Logo Cap	964,444.15	951,816.46	**1,916,260.61**
Classic Vest, L	65,909.76	63,173.84	**129,083.60**
Classic Vest, M	76,820.20	75,448.15	**152,268.35**
Half-Finger Gloves, L	152,665.35	183,815.71	**336,481.05**
Hitch Rack - 4-Bike	88,724.61	112,165.79	**200,890.40**
HL Road Tire	405,406.07	207,563.87	**612,969.94**
LL Mountain Tire	50,031.07	55,588.01	**105,619.08**
Long-Sleeve Logo Jersey, L	191,014.18	182,939.27	**373,953.46**
Long-Sleeve Logo Jersey, S	132,907.23	158,692.78	**291,600.01**
Long-Sleeve Logo Jersey, XL	148,027.30	158,626.53	**306,653.83**
ML Road Tire	286,597.37	308,879.86	**595,477.22**
Mountain Tire Tube	561,133.12	590,720.72	**1,151,853.83**
Total	**3,217,278.04**	**3,117,712.62**	**6,334,990.66**

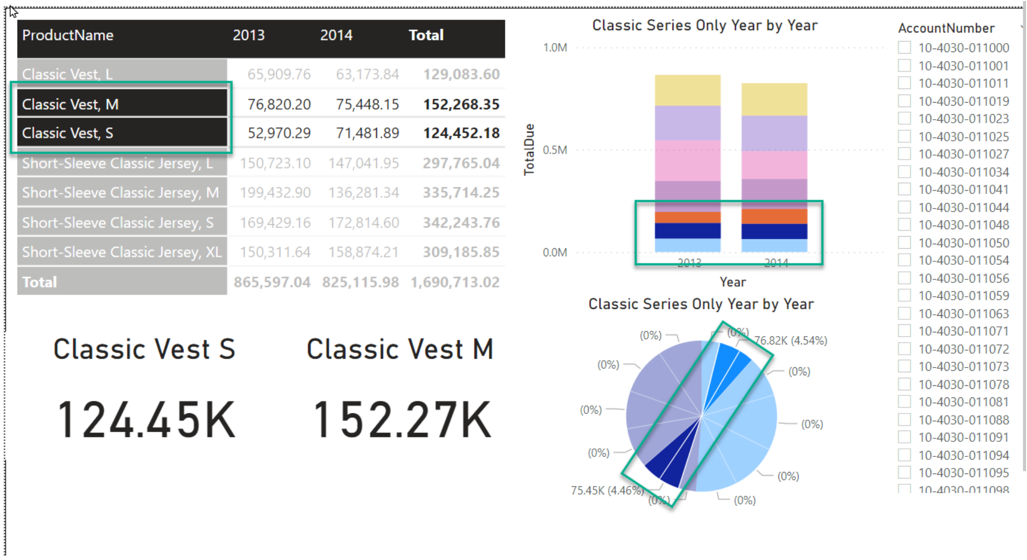
ProductName
■ All-Purpose Bike Stand
■ AWC Logo Cap
□ Bike Wash - Dissolver
■ Classic Vest, L
■ Classic Vest, M
□ Classic Vest, S
□ Fender Set - Mountain
■ Half-Finger Gloves, L
□ Half-Finger Gloves, M
□ Half-Finger Gloves, S
■ Hitch Rack - 4-Bike
□ HL Mountain Tire
■ HL Road Tire
□ Hydration Pack - 70 oz.
■ LL Mountain Tire
□ LL Road Tire
■ Long-Sleeve Logo Jersey, L
□ Long-Sleeve Logo Jersey, M
■ Long-Sleeve Logo Jersey, S
■ Long-Sleeve Logo Jersey, XL
□ ML Mountain Tire
■ ML Road Tire
□ Mountain Bottle Cage
■ Mountain Tire Tube
□ Mountain-200 Black, 38
□ Mountain-200 Black, 42
□ Mountain-200 Black, 46
□ Mountain-200 Silver, 38
□ Mountain-200 Silver, 42
□ Mountain-200 Silver, 46
□ Mountain-400-W Silver, 38
□ Mountain-400-W Silver, 40

Slicer in Power BI Set to Only Show Product Names (Used with permission from Microsoft.)

In our dashboard here, we have added a slicer that will allow us to select the items we want to review, and it will automatically filter all visuals on the dashboard.

Visuals That Filter Other Visuals

In the world of dashboarding, visuals can also serve as filters. This means that when a consumer clicks on any column or bar, all the data is filtered based on that selection. This type of filter can be especially meaningful when people are seeking information for actionable insights. For example, in a Classic Vest research study, suppose you want to see how both small and medium Classic Vest products are performing. Having a visual filter, like a matrix, will allow us to select and see that specific data.



ProductName	2013	2014	Total
Classic Vest, L	65,909.76	63,173.84	129,083.60
Classic Vest, M	76,820.20	75,448.15	**152,268.35**
Classic Vest, S	52,970.29	71,481.89	**124,452.18**
Short-Sleeve Classic Jersey, L	150,723.10	147,041.95	297,765.04
Short-Sleeve Classic Jersey, M	199,432.90	136,281.34	335,714.25
Short-Sleeve Classic Jersey, S	169,429.16	172,814.60	342,243.76
Short-Sleeve Classic Jersey, XL	150,311.64	158,874.21	309,185.85
Total	865,597.04	825,115.98	1,690,713.02

Classic Vest S
124.45K

Classic Vest M
152.27K

AccountNumber
□ 10-4030-011000
□ 10-4030-011001
□ 10-4030-011011
□ 10-4030-011019
□ 10-4030-011023
□ 10-4030-011025
□ 10-4030-011027
□ 10-4030-011034
□ 10-4030-011041
□ 10-4030-011044
□ 10-4030-011048
□ 10-4030-011050
□ 10-4030-011054
□ 10-4030-011056
□ 10-4030-011059
□ 10-4030-011063
□ 10-4030-011071
□ 10-4030-011072
□ 10-4030-011073
□ 10-4030-011078
□ 10-4030-011081
□ 10-4030-011088
□ 10-4030-011091
□ 10-4030-011094
□ 10-4030-011095
□ 10-4030-011098

Survey in Power BI by Selecting the Two Products of the Matrix Classic Vest, M and Classic Vest, S (Used with permission from Microsoft.)

In our sample, we selected the Classic Vest, M and Classic Vest, S. The impact to the dashboard is that the parts not applicable to those selections have a different opacity than the parts of the dashboard that are impacted. Looking at the highlights, you can see the parts of the stack and pieces of the pie that represent these products.

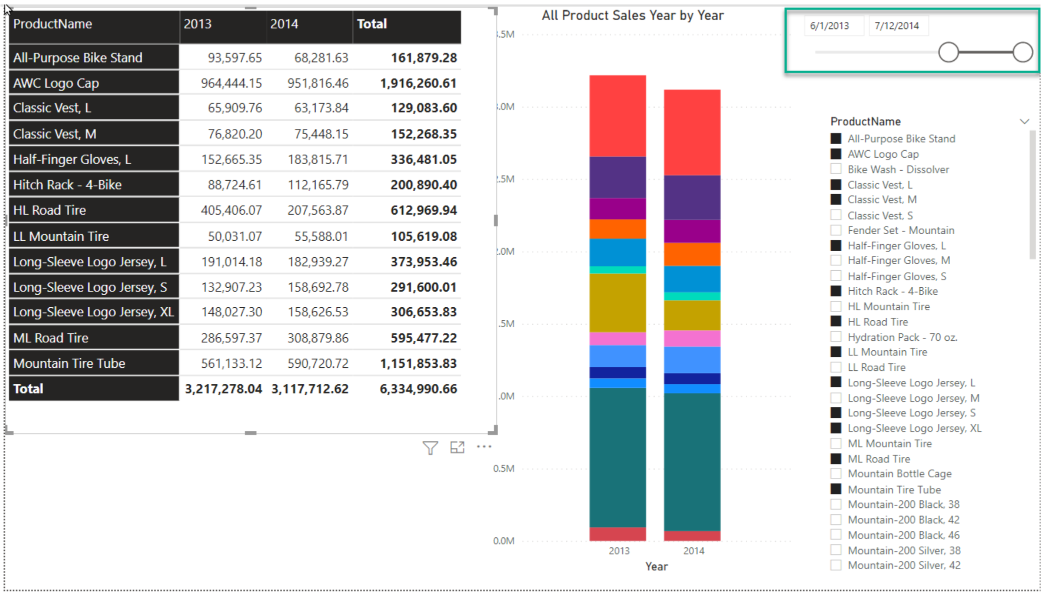
Depending on the tool used, some visuals on the dashboard are automatically meant to serve as filters by default, whereas in other tools, the filters have to be manually set for a dashboard item/visual. Each software, like Power BI or Tableau, handles this feature differently.

Filtering by Date Ranges

In almost any organization across all industries, dates are one of the most common fields used as a filter for data. Businesses often like to analyze and organize data by time. For example, an organization might want to run a daily report that shows what occurred each day. To do so, a filter can be applied that only shows that day in time.

Methods for Filtering by Date Range

A company also might want to filter data by a period of time, such as weekly, monthly, or annually. This is done through the use of a **date filter**, which filters a data field by a starting and ending point. The most common statement used for this filter would be a BETWEEN AND statement that would allow you to look at a data field between a start date and an end date. This statement can also be written as greater than or equal to a start date, and less than or equal to an end date.



ProductName	2013	2014	Total
All-Purpose Bike Stand	93,597.65	68,281.63	**161,879.28**
AWC Logo Cap	964,444.15	951,816.46	**1,916,260.61**
Classic Vest, L	65,909.76	63,173.84	**129,083.60**
Classic Vest, M	76,820.20	75,448.15	**152,268.35**
Half-Finger Gloves, L	152,665.35	183,815.71	**336,481.05**
Hitch Rack - 4-Bike	88,724.61	112,165.79	**200,890.40**
HL Road Tire	405,406.07	207,563.87	**612,969.94**
LL Mountain Tire	50,031.07	55,588.01	**105,619.08**
Long-Sleeve Logo Jersey, L	191,014.18	182,939.27	**373,953.46**
Long-Sleeve Logo Jersey, S	132,907.23	158,692.78	**291,600.01**
Long-Sleeve Logo Jersey, XL	148,027.30	158,626.53	**306,653.83**
ML Road Tire	286,597.37	308,879.86	**595,477.22**
Mountain Tire Tube	561,133.12	590,720.72	**1,151,853.83**
Total	**3,217,278.04**	**3,117,712.62**	**6,334,990.66**

6/1/2013 7/12/2014

ProductName

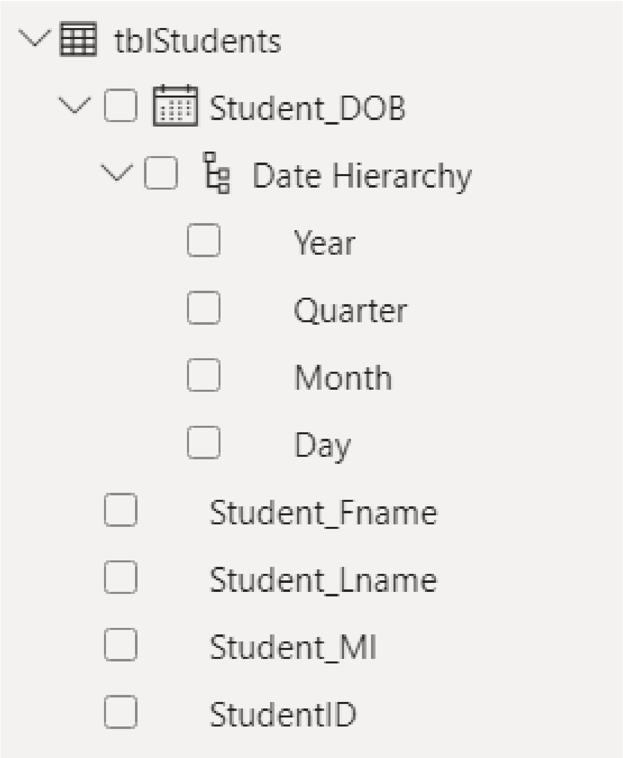
■ All-Purpose Bike Stand
■ AWC Logo Cap
□ Bike Wash - Dissolver
■ Classic Vest, L
■ Classic Vest, M
□ Classic Vest, S
□ Fender Set - Mountain
■ Half-Finger Gloves, L
□ Half-Finger Gloves, M
□ Half-Finger Gloves, S
■ Hitch Rack - 4-Bike
□ HL Mountain Tire
■ HL Road Tire
□ Hydration Pack - 70 oz.
■ LL Mountain Tire
□ LL Road Tire
■ Long-Sleeve Logo Jersey, L
□ Long-Sleeve Logo Jersey, M
■ Long-Sleeve Logo Jersey, S
■ Long-Sleeve Logo Jersey, XL
□ ML Mountain Tire
■ ML Road Tire
□ Mountain Bottle Cage
■ Mountain Tire Tube
□ Mountain-200 Black, 38
□ Mountain-200 Black, 42
□ Mountain-200 Black, 46
□ Mountain-200 Silver, 38
□ Mountain-200 Silver, 42

Date Filters in Microsoft Power BI (Used with permission from Microsoft.)

In this image, we can see a date range option that filters all the data on the dashboard when the specified date is set to the range inside the slicer. This allows the user to interact not only by the other slicer, but also to specify particular dates.

When writing the proper business requirements for any reporting project, you will want to identify which fields are needed for filtering, so you can identify which dates may be needed for any given report. For example, when reporting on the onboarding of new employees, we might want to look at their actual hire date, which may differ from their start date. We should identify up front which dates are needed for the report.

Most reporting tools will have some form of **date hierarchy** and syntax that allow you to filter by date.



- tblStudents
  - □ Student_DOB
    - □ Date Hierarchy
      - □ Year
      - □ Quarter
      - □ Month
      - □ Day
  - □ Student_Fname
  - □ Student_Lname
  - □ Student_MI
  - □ StudentID

Date Hierarchy in Microsoft Power BI (Used with permission from Microsoft.)

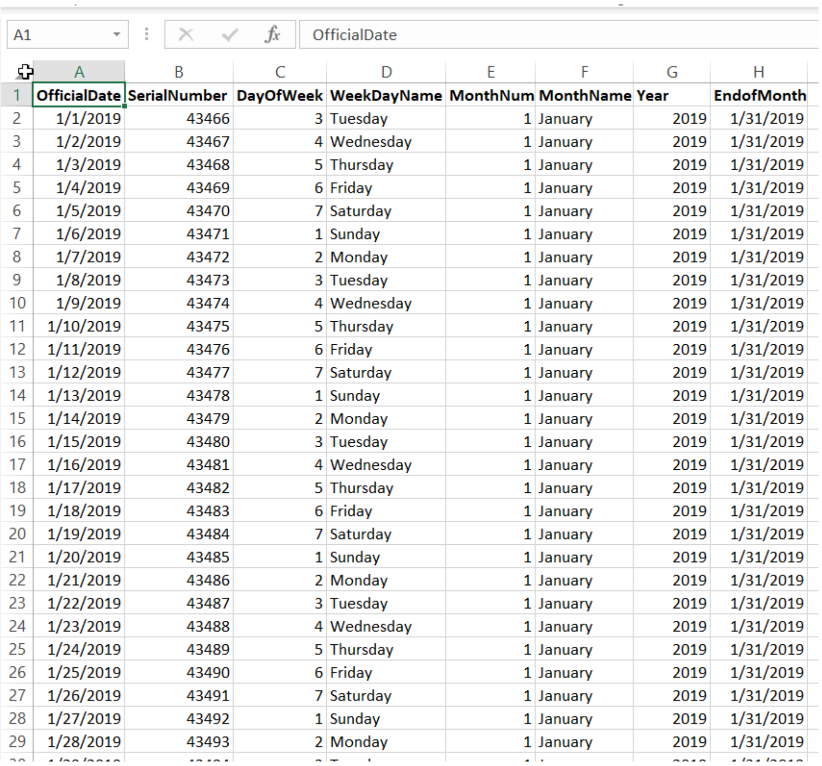
Power BI will recognize any date field, applying a date hierarchy that allows you to see that date represented by month, quarter, and year without having to write functions to create this information for your report.

Date Tables

Data analysts often implement a **date table** when the data only represents the date in which an event or transaction occurred, and not all the dates within a specified period. For example, there are 365 days in the calendar year, but a company won't have sales on literally every day. If you want to see how the organization performed on all days of the year, you'd first need a data set that contained all the days of the year, and then you could establish a report that would show each day of the year and whether or not there were sales on that day.

When important dates are not included in the data, the analyst will establish a date table. Methods of creating the date table include using a script that creates dates in Power BI or creating an Excel spreadsheet with all the dates listed in each row that are being imported into the visualization of the data. A data table may contain more than just the dates, including variations of the date.

Imagine that we want to see what business day in the year holds the most sales. We need to have access to data that has every day of the year over multiple years, and we would want to see the day of the week. We can then combine that date data with the transactions, leveraging a join on the date fields. This data combined will let us know on which days, like Tuesday, the company sees the most of any given transaction. For example, is there a higher sales volume on any one day of the week versus any other day? In order to see this type of information, we must know every day, and every day a sale came in, plus what day of the week it came in.



	A	B	C	D	E	F	G	H
1	OfficialDate	SerialNumber	DayOfWeek	WeekDayName	MonthNum	MonthName	Year	EndofMonth
2	1/1/2019	43466	3	Tuesday	1	January	2019	1/31/2019
3	1/2/2019	43467	4	Wednesday	1	January	2019	1/31/2019
4	1/3/2019	43468	5	Thursday	1	January	2019	1/31/2019
5	1/4/2019	43469	6	Friday	1	January	2019	1/31/2019
6	1/5/2019	43470	7	Saturday	1	January	2019	1/31/2019
7	1/6/2019	43471	1	Sunday	1	January	2019	1/31/2019
8	1/7/2019	43472	2	Monday	1	January	2019	1/31/2019
9	1/8/2019	43473	3	Tuesday	1	January	2019	1/31/2019
10	1/9/2019	43474	4	Wednesday	1	January	2019	1/31/2019
11	1/10/2019	43475	5	Thursday	1	January	2019	1/31/2019
12	1/11/2019	43476	6	Friday	1	January	2019	1/31/2019
13	1/12/2019	43477	7	Saturday	1	January	2019	1/31/2019
14	1/13/2019	43478	1	Sunday	1	January	2019	1/31/2019
15	1/14/2019	43479	2	Monday	1	January	2019	1/31/2019
16	1/15/2019	43480	3	Tuesday	1	January	2019	1/31/2019
17	1/16/2019	43481	4	Wednesday	1	January	2019	1/31/2019
18	1/17/2019	43482	5	Thursday	1	January	2019	1/31/2019
19	1/18/2019	43483	6	Friday	1	January	2019	1/31/2019
20	1/19/2019	43484	7	Saturday	1	January	2019	1/31/2019
21	1/20/2019	43485	1	Sunday	1	January	2019	1/31/2019
22	1/21/2019	43486	2	Monday	1	January	2019	1/31/2019
23	1/22/2019	43487	3	Tuesday	1	January	2019	1/31/2019
24	1/23/2019	43488	4	Wednesday	1	January	2019	1/31/2019
25	1/24/2019	43489	5	Thursday	1	January	2019	1/31/2019
26	1/25/2019	43490	6	Friday	1	January	2019	1/31/2019
27	1/26/2019	43491	7	Saturday	1	January	2019	1/31/2019
28	1/27/2019	43492	1	Sunday	1	January	2019	1/31/2019
29	1/28/2019	43493	2	Monday	1	January	2019	1/31/2019

Date Table in Microsoft Excel (Used with permission from Microsoft.)

In this sample, we have all the relevant date fields we need, with additional calculations for DayOfWeek, WeekDayName, MonthNum, and MonthName. To create a date table in Excel, start by adding the first date you need and extend to the last date needed. Include all the date variations you may need or want for your table.

Review Activity:

Ways to Sort and Filter Data

Answer the following questions:

1. **Which type of filter does not allow the user to adjust it?**

2. **Which type of filter does give the user the ability to adjust it?**

3. **What else can serve as a filter?**

4. **What filter would be a consideration for payroll data that is done on a monthly basis?**

5. **What type of table is typically used with the payroll data when working with a report that uses a date filter?**

6. **When data is not yet sorted, what is the order of the data?**

7. **What are two sort methods you can apply to your data sets?**

8. **What type of sort will also filter data?**

Lesson 13

Summary

After this lesson, you should have gained an understanding of the types of business requirements you must gather for any of your data projects and visualizations. You should consider the consumer/audience and the types of data reporting they may need, the source of and access to data, and how the report will be viewed and delivered. You have discovered the types of visuals that you may find on a dashboard and should have a basic understanding of the design process, using tools to create mockups and wireframing to help design your reporting projects. You should know what types of filters can be used and have a good understanding of the different sort types that are available. You should have also discovered that every decision, data set, and source should be documented, from the field description to listing relevant information on tabs in yourreports.

Guidelines in Expressing Business Requirements for Reporting

Consider these best practices and guidelines when familiarizing yourself with the necessary considerations when translating information into business requirements.

1. Before you do anything, remember to always document critical information for both yourself and others.
2. Identify your audience and the types of data they will need access to, so you can adjust how the report will be presented.
3. Determining the source of the data helps you identify whether additional access to data is needed for visualizations and reporting.
4. If additional permissions are required, seek out and follow the approval processes and policies at your organization.
5. When field names are likely to be unclear to users, or even possibly misleading, providing field definitions can help clarify what information these fields contain.
6. Recall that there are different ways the visuals will be viewed—such as via a web interface, as a file format, in other software, or on paper—and the type of display matters because it affects how you export your data.
7. The way the report is delivered—such as in a meeting, through "report sharing" in software, printed, or through a subscription—must be identified up front, since it can affect what software you should use.
8. The type of data you're working with, and the delivery method, will affect how frequently reports should be updated and how often they should be sent to the audience.
9. Recurring reports regularly occur on a schedule and are delivered in different ways based on organizational policy and audience preferences.

10. Mockups and wireframing can help your audience see what to expect before you publish a report, which can lead to feedback that could change your direction.

11. Ask up front what types of filters and sorts are preferred for dashboards; some companies may want to provide interactive filters so the audience can choose what they want to see, while in other instances hard-coded filters are best.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 14

Designing Components for Reports and Dashboards

LESSON INTRODUCTION

When developing reports and dashboards as a data analyst, you must consider not only how to make the data meaningful for what you are reporting on for the organization, but also in accordance with the organization's style. Your audience will desire a great user experience, so reports and dashboards must be designed with this in mind. Using proper colors, knowing the needs of your audience, and following the company style guide are all important parts of the process. Everything must be considered right down to the font style and size used for the data. You also mustn't forget to include key elements, like the refresh date, and the narrative and key talking points, so you can be sure your report conveys what the data is truly saying. You should also provide answers for the questions your audience may have and include critical supporting information that not only helps them, but also helps you move on to the next task at hand.

Lesson Objectives

In this lesson, you will do the following:

- Design elements for reports/dashboards.
- Utilize standard elements.
- Create a narrative and other written elements.
- Understand deployment considerations.

Topic 14A

Choose Design Elements for Reports/Dashboards

EXAM OBJECTIVES COVERED

4.2 Given a scenario, use appropriate design components for reports and dashboards.

When designing reports and/or dashboards, you must consider the organization and the consumers at all times. You should seek guidance for how to follow your organization's brand strategy and remain true to the organizational brand when making design decisions. Effective design components in reporting involve the use of appropriate colors, fonts, and layouts to provide the best user experience.

Branding Guidelines

The age and size of the organization you support will influence the overall branding guidelines of the organization. When a company adopts a brand identity, they will want that brand reflected in all the artifacts that represent the company. Data analysts often create reports or visualizations that will go into other presentations and documents, or onto websites. Some companies establish branding procedures and **style guides**. As a data analyst, you should always ask if these resources exist.

Style guides commonly contain different variations of an organization's logo and guidelines for how it can be used. They may include recommendations of logo placement, and provide different logo styles or placements dependent of the type of medium you are working with. Printed pages, presentations, and websites might all have different logo placements that are acceptable.

When a report is intended for different consumer types, the company may have additional guidelines for serving different consumers. For example, a specific PowerPoint design template may be required for C-level executives. Guidelines for public-facing reports that are delivered to customers or the general public may include requirements about what types of disclaimers are to be placed and where.

If your organization has a marketing department, that's a good first resource for learning about branding guidelines, so you can ensure your reports reflect the brand.

When no organization style guide exists, many designers and developers use APA guidelines for business.

Appropriate Color Schemes

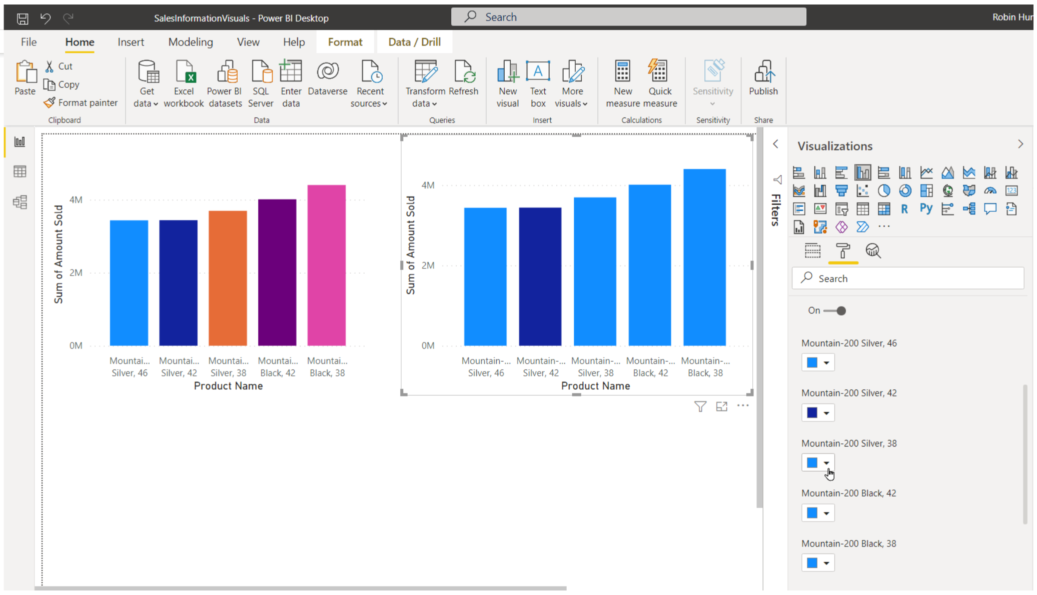
A style guide that tells you exactly what colors you can and can't use makes it easier to make design and color choices for a report or dashboard. However, if a style guide for color doesn't exist, you must decide on an appropriate color scheme for the audience who will consume the information. Here are some important considerations related to color:

- Rather than choosing random colors, select specific colors with the user experience in mind.
- Avoid distracting color schemes (e.g., the use of too many different colors) that detract from the information presented.
- Use a consistent color scheme throughout the dashboarding experience so that users don't have to acclimate to new colors as they move through different dashboards.

Some software programs, like Tableau, recommend themes that have been designed using high-contrast colors to ensure accessibility for colorblind readers. If not using a default theme, it may be worthwhile to test your visual for accessibility with a tool like "ColorBlindly" or another similar tool.

Customizing Colors

The colors for most visuals are customizable. It is important to remember that visuals will sort to different values by default, and that this will shift the colors to different categories. For example, if you start with Mountain - 200 Silver, 46 as a blue color, you will want to ensure that same blue is used for that product throughout the dashboard, regardless of the default options of the visual. You can do this by adjusting the colors manually where required.



On

Mountain-200 Silver, 46 🟦

Mountain-200 Silver, 42 🟦

Mountain-200 Silver, 38 🟦

Mountain-200 Black, 42 🟦

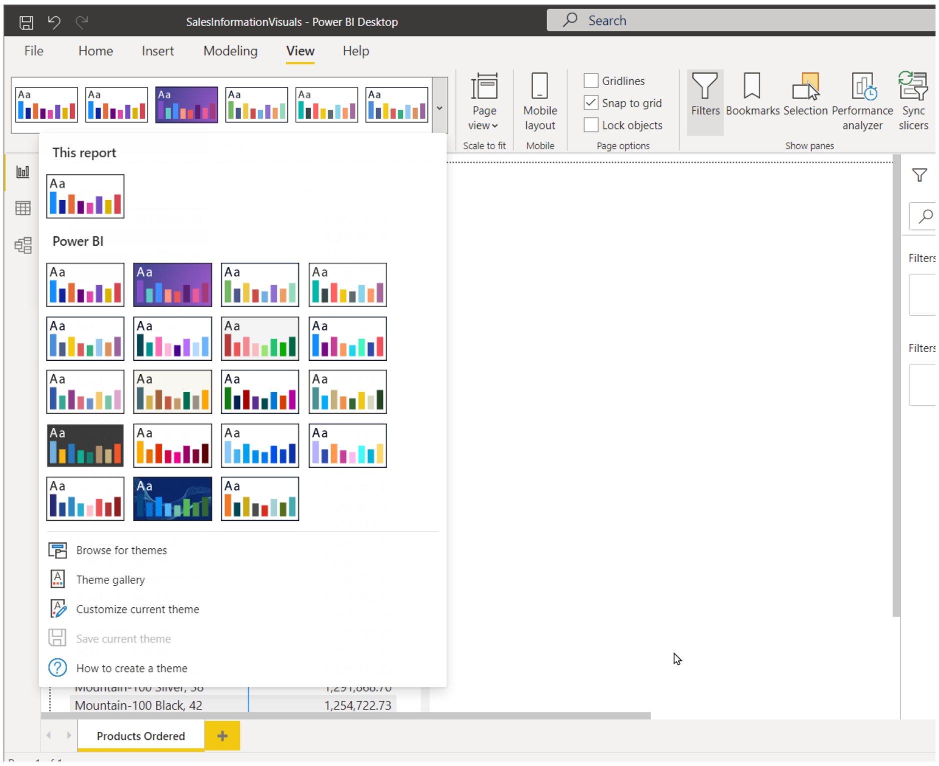
Mountain-200 Black, 38 🟦

Visual Color Editing in Microsoft Power BI Desktop (Used with permission from Microsoft.)

We would use the format options in the Power BI format menu for data colors to edit the colors of the visual on the right to match the color of each product in the visual on the left. Although this visual will ultimately be placed on another page in the file, it's important to ensure consistent use of color throughout the report.

Default Color Schemes

Most software products have a default color scheme applied. Consider discussing several of those palettes with decision makers in your organization—like the marketing team, other data workers in your department, or the leadership that is responsible for the work you are doing—to determine what is most suitable for the brand. When there is no policy or style guide to follow, it's important to determine when you should use certain default color schemes, and which palettes should never be used.

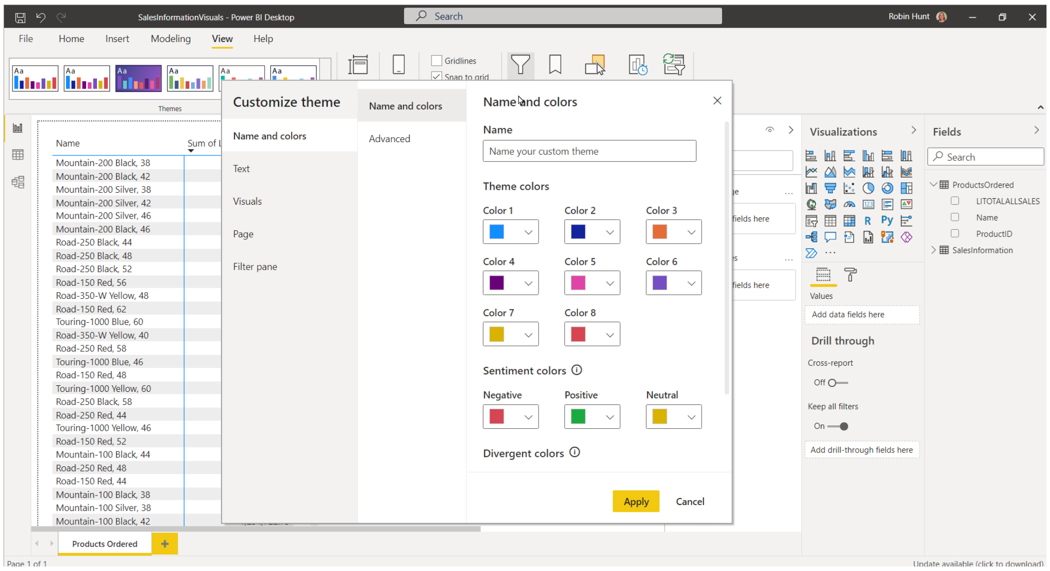


Color Schemes in Microsoft Power BI (Used with permission from Microsoft.)

In Microsoft Excel, the built-in themes are on the Page Layout tab of the ribbon.

When a style guide exists, you may want to create a theme or template that incorporates the colors you need by default. This step is a massive time-saver, requiring you to only tweak color as needed in the future. For example, in Microsoft Excel, you can create several variations of chart templates and select to use those templates by default, or you can apply an existing template to any chart you create.

The following screenshots show ways to customize themes and their colors using Microsoft Power BI.



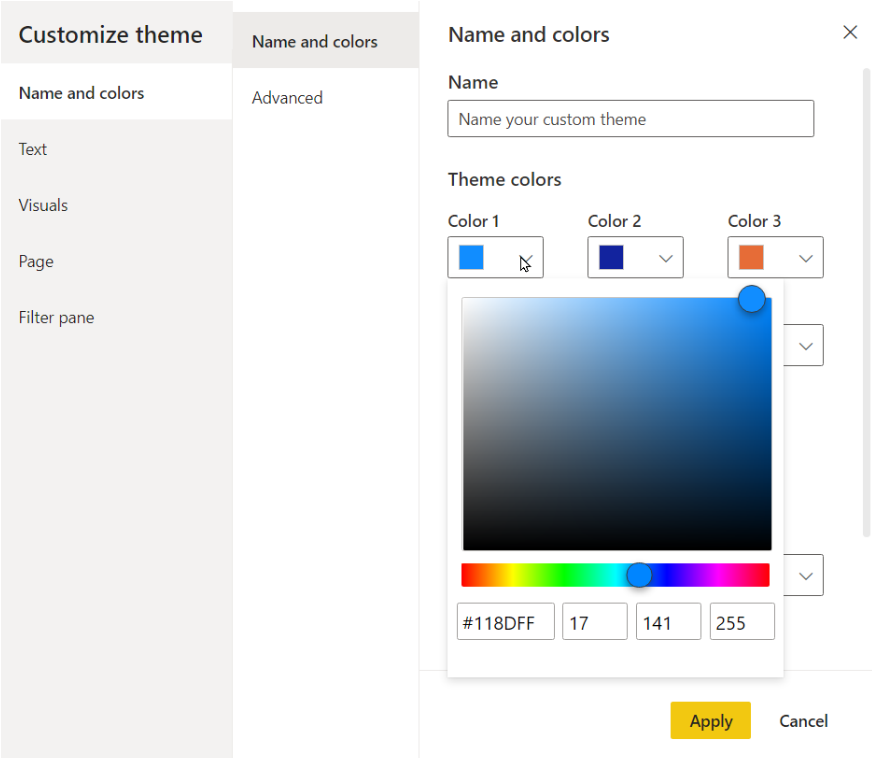
Name and colors

Name: Name your custom theme

Name	Sum of I
Mountain-200 Black, 38	
Mountain-200 Black, 42	
Mountain-200 Silver, 38	
Mountain-200 Silver, 42	
Mountain-200 Silver, 46	
Mountain-200 Black, 46	
Road-250 Black, 44	
Road-250 Black, 48	
Road-250 Black, 52	
Road-150 Red, 56	
Road-350-W Yellow, 48	
Road-150 Red, 62	
Touring-1000 Blue, 60	
Road-350-W Yellow, 40	
Road-250 Red, 58	
Touring-1000 Blue, 46	
Road-150 Red, 48	
Touring-1000 Yellow, 60	
Road-250 Black, 58	
Road-250 Red, 44	
Touring-1000 Yellow, 46	
Road-150 Red, 52	
Mountain-100 Black, 44	
Road-250 Red, 48	
Road-150 Red, 44	
Mountain-100 Black, 38	
Mountain-100 Silver, 38	
Mountain-100 Black, 42	

Customize Theme in Microsoft Power BI (Used with permission from Microsoft.)

You can customize almost every object in an existing theme, or create an entirely new theme to meet the standards and requirements for your organization.



Name and colors

Name: Name your custom theme

#118DFF	17	141	255

Custom Color Selection in Microsoft Power BI Customize Theme (Used with permission from Microsoft.)

When a style guide requires specific colors, you should never be approximating which color matches the requirement. Use the hexadecimal code or the RGB values to ensure the exact color is selected. The hexadecimal code and RGB values are a mathematical way to ensure color accuracy versus using your eyes on various screens. Style guides will often provide you these values.

Appropriate Fonts and Layout

It is important to use the appropriate fonts, font sizes, and layouts when creating a report or dashboard. Let's go through some key considerations.

Font and Layout Considerations

When selecting a font, one of the first steps is to consider the delivery format of the dashboard or report. If a report or dashboard is simply printed on a local printer, then your choice of a font is less challenging and involves fewer considerations. However, if you are sending a report to a production-level printer for mass printing on a high-quality scale, then you may have to send fonts to the printshop in advance. For reports and dashboards presented on the web, the browser will translate font choices, which has implications for people viewing the report onscreen. You can't control the output of the report on every machine, so it's important to keep in mind how certain fonts might be displayed depending on the technology or device being used.

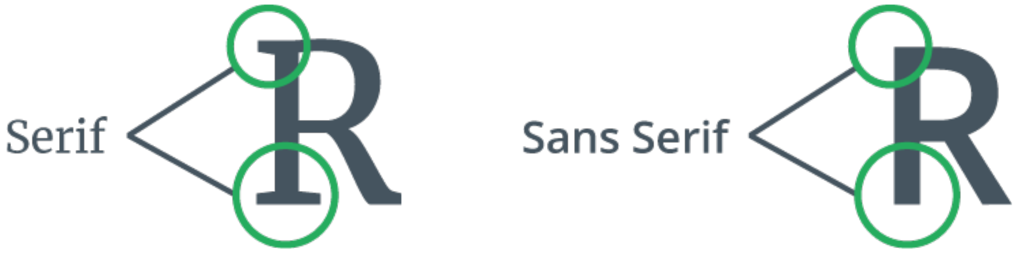
When designing a dashboard that includes text information, think carefully through the layout. Will the dashboard be delivered in a mobile format, meaning it will be read on a phone or tablet? Or will it be displayed on a basic desktop PC or Mac screen? Readability is key, so whatever fonts you decide to use should be versatile for both print and web.

It is important to identify how much space your audience will have from left to right and top to bottom at the very beginning of any project, as this will affect your layout. An iPhone screen is significantly smaller than a laptop screen, so a dashboard that will only be available on a mobile device would require a different layout than a report viewed on the web interface of a computer.

⚠ *When working on any project, data related or otherwise, it is important to follow ADA compliance standards that ensure information can be viewed, read, and interpreted by a more diverse audience. Different software exists to read and interpret the items on a screen, so be sure to consider your audience and know how various software translates reports and dashboards for people with visual and hearing impairments. There are several tools that can be leveraged to determine if your project meets compliance standards. For more information on ADA compliance, visit https://www.ada.gov/index.html.*

Serif vs. Sans Serif Fonts

Serif and sans serif are the two basic types of fonts. In a serif font, the letter edges have lines that make the font more readable when there are paragraphs of text at smaller font sizes, such as the written narrative of a report. Sans serif fonts do not have these lines and can be useful for headings, labels, and other elements surrounding visuals.

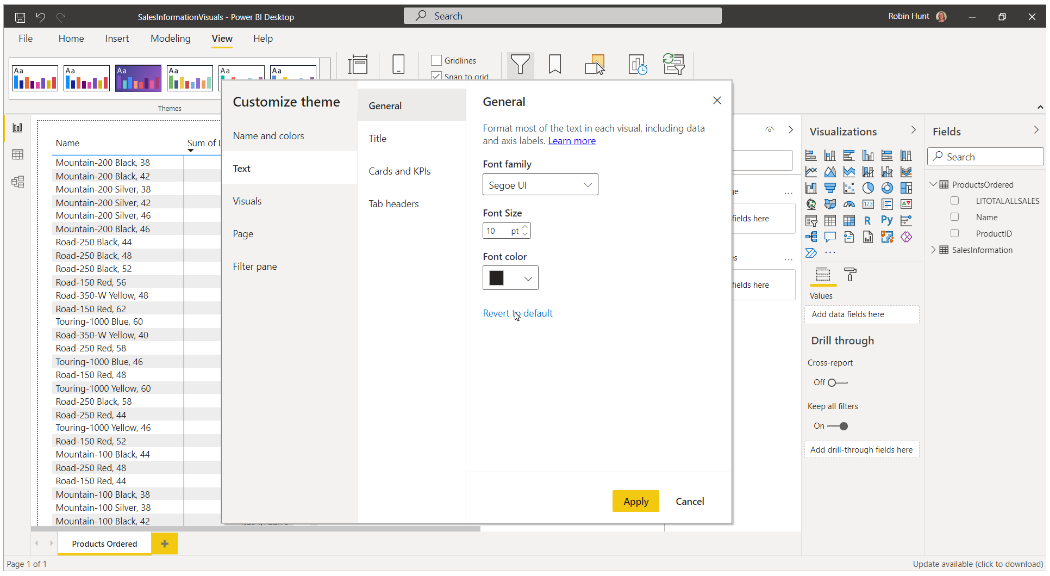


Comparison of a Serif Font and Sans Serif Font

This visual shows a side-by-side comparison of a serif font (Times New Roman) and sans serif font (Ariel) at the same point size. The serif font has little lines on the edges of the letter, whereas the sans serif font does not. It's important to note that two different fonts may not appear to be the same physical size, even if set in the same point size, as every font presents differently.

Customizing Fonts

By customizing a theme, you can apply a desired font throughout your report or dashboard to save time. As shown in the next screenshot, in Microsoft Power BI, you can change the default font of a theme to a customized font family, font size, and font color.



Customize theme

- Name and colors
- Text
- Visuals
- Page
- Filter pane

- General
- Title
- Cards and KPIs
- Tab headers

General

Format most of the text in each visual, including data and axis labels. Learn more

Font family: Segoe UI

Font Size: 10 pt

Font color: ■

Revert to default

Apply Cancel

Name	Sum of L
Mountain-200 Black, 38	
Mountain-200 Black, 42	
Mountain-200 Silver, 38	
Mountain-200 Silver, 42	
Mountain-200 Silver, 46	
Mountain-200 Black, 46	
Road-250 Black, 44	
Road-250 Black, 48	
Road-250 Black, 52	
Road-150 Red, 56	
Road-350-W Yellow, 48	
Road-150 Red, 62	
Touring-1000 Blue, 60	
Road-350-W Yellow, 40	
Road-250 Red, 58	
Touring-1000 Blue, 46	
Road-150 Red, 48	
Touring-1000 Yellow, 60	
Road-250 Black, 58	
Road-250 Red, 44	
Touring-1000 Yellow, 46	
Road-150 Red, 52	
Mountain-100 Black, 44	
Road-250 Red, 48	
Road-150 Red, 44	
Mountain-100 Black, 38	
Mountain-100 Silver, 38	
Mountain-100 Black, 42	

Customize Theme for Fonts in Microsoft Power BI (Used with permission from Microsoft.)

Naming Conventions

As data analysts, we will see a lot of different naming conventions when working with data. We will see what a field is technically named in the database, and we can use various methods to make sure that the name the audience will see is meaningful. You will want to carefully label elements to convey what they mean in regard to the data, not just what the database architect happened to name them.

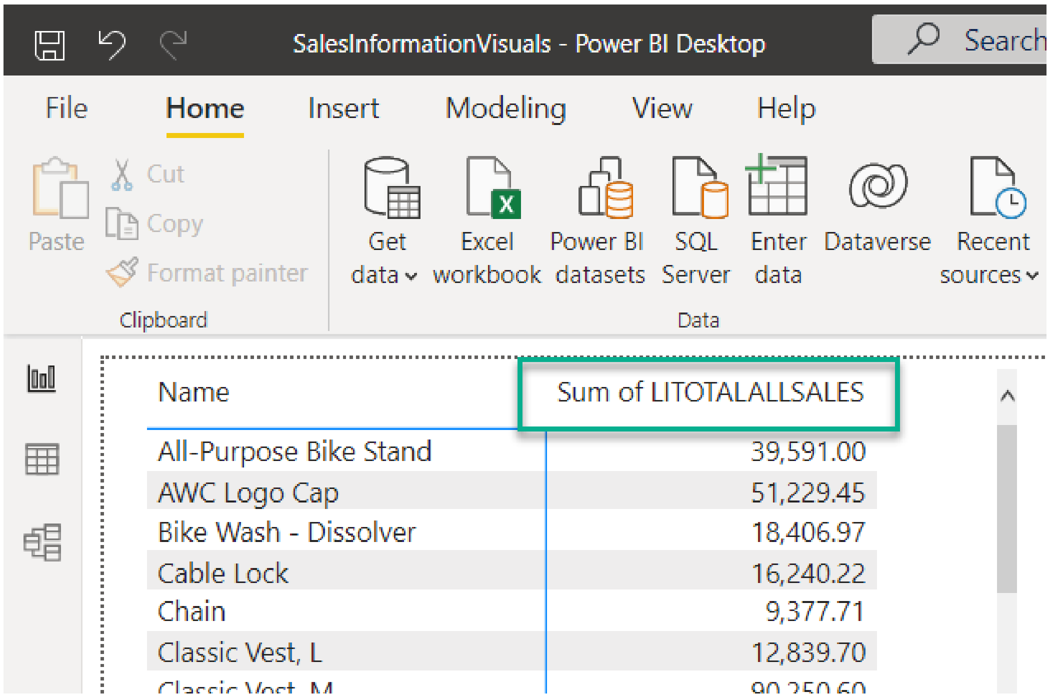
Captioning Fields

As discussed previously, the database or data set is designed with certain field names. These field names may not always make sense to people without knowledge of the database design. **Captioning** allows you to designate more meaningful names for fields in a report or dashboard. For example, if a field in a data set is named "FirstN" for First Name, you could use captioning to display the field as "First Name." Captioning does not change the underlying field name in the database or data set, but instead displays a custom name that clarifies the information being presented.

A few different methods are available for changing captions:

- If you change the caption in the query that provides data to the visual, then every object that you build using that query will already have the appropriate caption.
- If you change the caption in the transformation step of Power Query, then all the visuals that use that query will have the appropriate caption.
- If you build the caption into a visual, then only that visual will have that caption. In this case, every time you use that field for a new visual, you will need to caption it again.

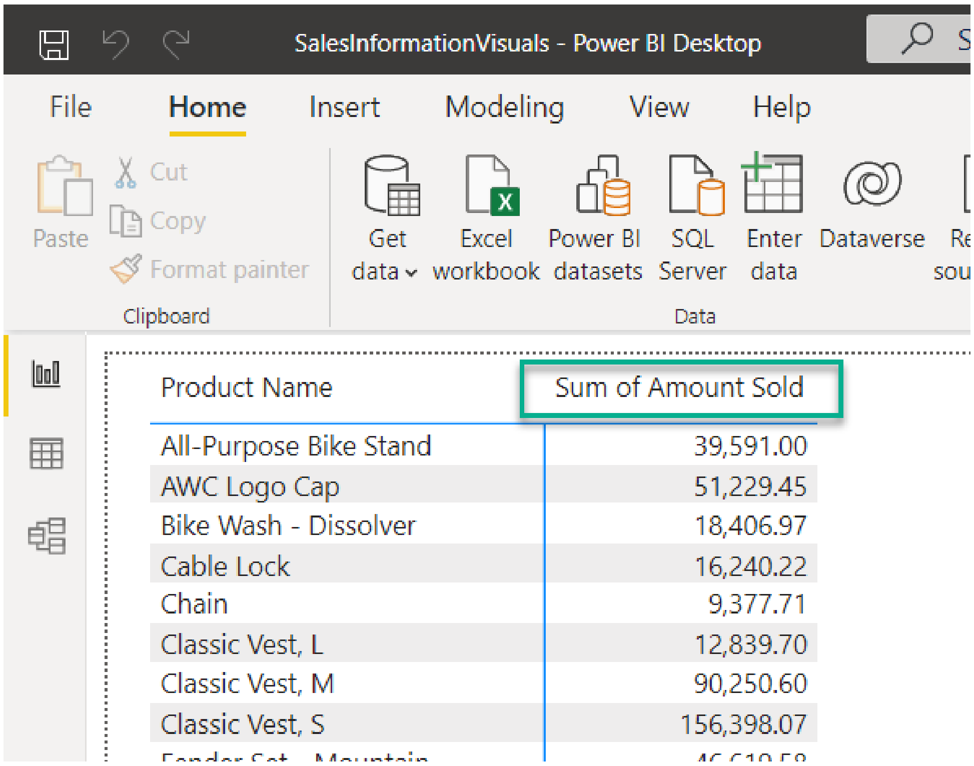
In our sample below, we see "Sum of LITOTALALLSALES" as the label of this field of information. This name is created automatically by Power BI when we sum the field named LITOTALALLSALES. This is a prime example of something you will want to address as a data analyst.



Name	Sum of LITOTALALLSALES
All-Purpose Bike Stand	39,591.00
AWC Logo Cap	51,229.45
Bike Wash - Dissolver	18,406.97
Cable Lock	16,240.22
Chain	9,377.71
Classic Vest, L	12,839.70
Classic Vest, M	90,250.60

Power BI Showing Automated Label (Used with permission from Microsoft)

You can make this adjustment in several different places depending on the software you are using and how much access you have to the backend source. In our next screenshot, let's look at a more appropriate label.



Product Name	Sum of Amount Sold
All-Purpose Bike Stand	39,591.00
AWC Logo Cap	51,229.45
Bike Wash - Dissolver	18,406.97
Cable Lock	16,240.22
Chain	9,377.71
Classic Vest, L	12,839.70
Classic Vest, M	90,250.60
Classic Vest, S	156,398.07
Fender Set - Mountain	46,619.58

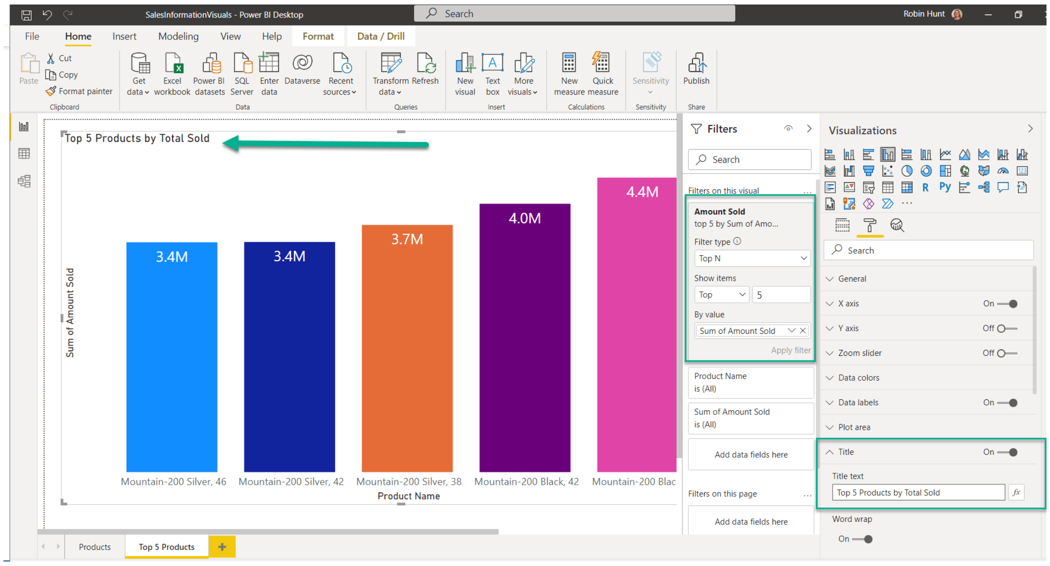
Power BI Caption Adjusted to Show More Meaningful Information (Used with permission from Microsoft)

Because this value represents the total amount sold ofthat product, it makes sense to caption it as what it is (Sum of Amount Sold) rather than leave the label that was automatically generated. We could have made it say "Total Amount Sold" as well—what we decide just depends on the preference of the people who request the reports.

Visual and Page Titles

When developing visuals and other elements of a report or dashboard, the software will often create default titles based on the fields built into the object (e.g., matrix or column chart). Because these default titles are rarely meaningful outside of designing visuals, it's important to change them so that they present intentional information to consumers.

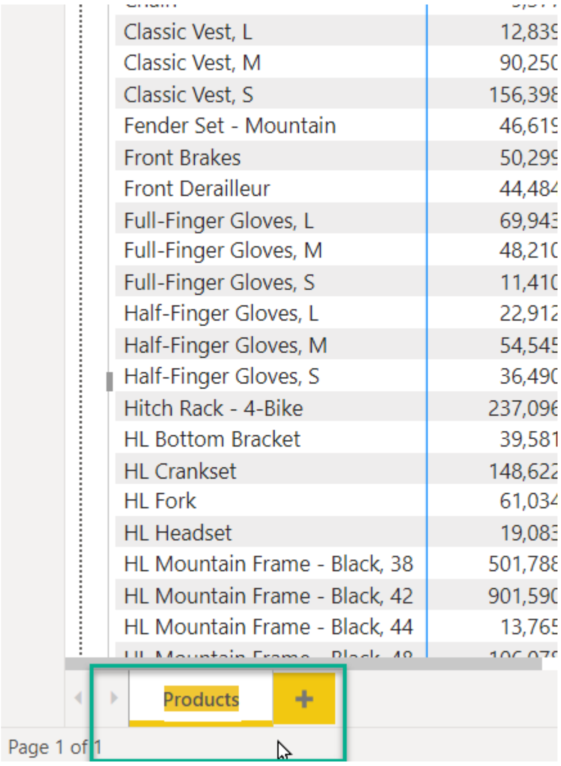
Check the title of the visual to ensure that it is appropriate and update as needed. As an example, consider a visual depicting the top five products sold. If the visual is built using data from a full product list, then a default title based on the fields used, such as "Products and Total Sold," would be inaccurate. The title of the visual can be appropriately updated as "Top 5 Products by Total Sold" to accurately describe what is presented, as shown in the next screenshot.



Amount Sold
top 5 by Sum of Amo...
Filter type: Top N
Show items: Top 5
By value: Sum of Amount Sold
Product Name is (All)
Sum of Amount Sold is (All)
X axis: On
Y axis: Off
Zoom slider: Off
Data labels: On
Title: On
Title text: Top 5 Products by Total Sold

Top 5 Products Visual in Microsoft Power BI (Used with permission from Microsoft.)

Pages within a dashboard product should also be titled meaningfully. The default "Page 1" doesn't describe what viewers see on the page. For example, in the next screenshot we could more appropriately title the page "Products" since this is a visual showing total products sold. When you publish visuals, the titles you've set will be displayed.



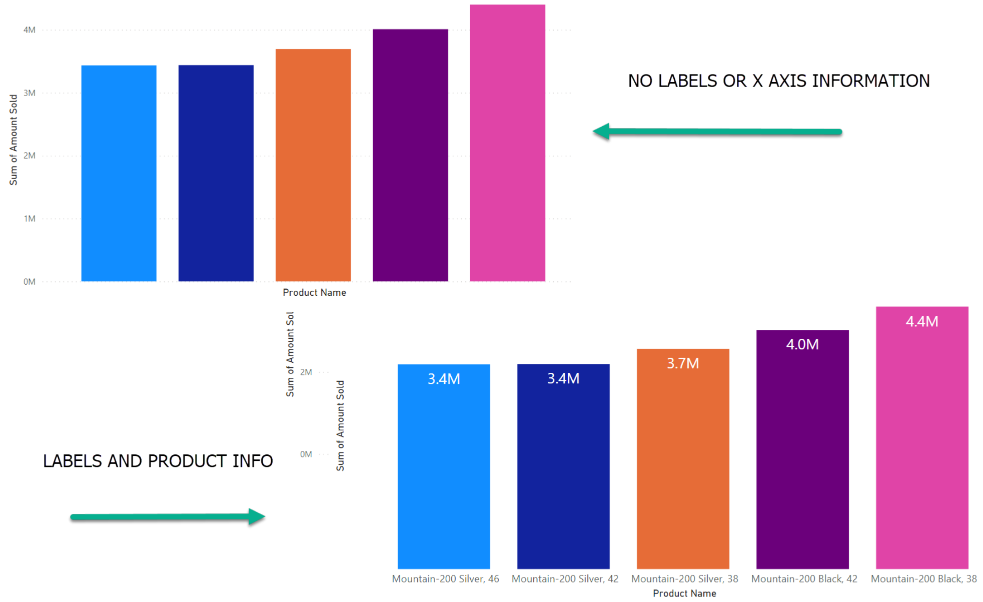
Classic Vest, L	12,839
Classic Vest, M	90,250
Classic Vest, S	156,398
Fender Set - Mountain	46,619
Front Brakes	50,299
Front Derailleur	44,484
Full-Finger Gloves, L	69,943
Full-Finger Gloves, M	48,210
Full-Finger Gloves, S	11,410
Half-Finger Gloves, L	22,912
Half-Finger Gloves, M	54,545
Half-Finger Gloves, S	36,490
Hitch Rack - 4-Bike	237,096
HL Bottom Bracket	39,581
HL Crankset	148,622
HL Fork	61,034
HL Headset	19,083
HL Mountain Frame - Black, 38	501,788
HL Mountain Frame - Black, 42	901,590
HL Mountain Frame - Black, 44	13,765

Changing Page Titles in Microsoft Power BI (Used with permission from Microsoft)

You can create meaningful and descriptive page titles for your dashboard pages by changing the name of the tab (like what you see highlighted in the screenshot above).

Labels and Legends

Legends and labels help consumers easily distinguish what they are viewing and understand the visual faster. People immediately seek these elements when viewing a visual to understand the purpose of the visual and interpret the information presented. The following screenshot shows a visual with and without labels. As you can plainly see, adding labels to visuals makes a major readability difference.

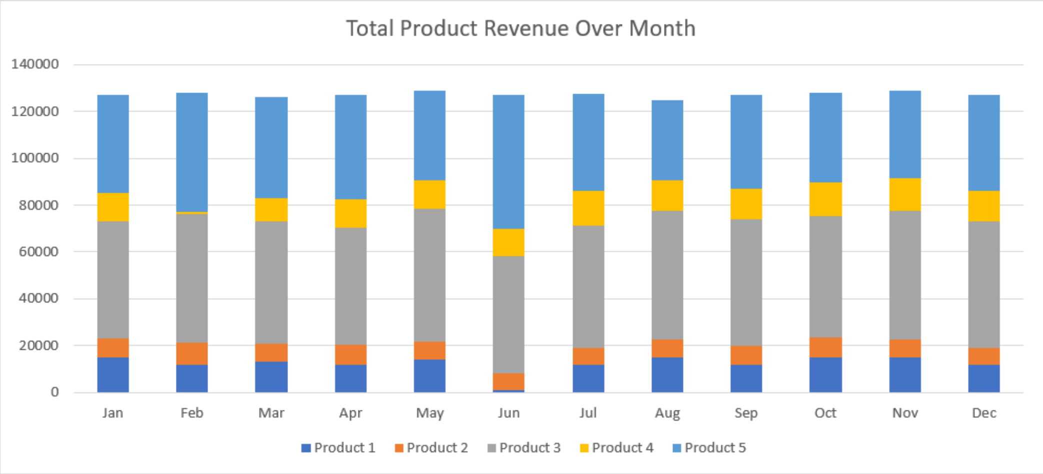


Comparison of a Visual in Microsoft Power BI with and Without Labels (Used with Permission from Microsoft.)

Labels can be easily turned on and off in Microsoft Power BI. In the image without specific labels for the x-axis, users would have to hover over each column in a dashboard environment to know which specific product is being visualized. On the y-axis, when labels are added for the sum of amount sold, users immediately know the value for each product, without having to look back to the left or hover over each product.

A **legend** is a labeling element that lets you know which color represents which value in a visual. Legends can be turned on or off, and they can be formatted like any other object in a visual. They can be placed to the left, right, top, or bottom of a visual.

You do not need a legend for every visual that you create. When designing a visual, you will need to determine if the legend provides value, and if so, where it should be placed. For example, if you have visible labels associated with the bar, pie, or block, then you likely don't need a legend. However, when working with stacked bars or stacked columns, the legend greatly aids in understanding without cluttering the visual with labels on every item, as shown in the following screenshot.



Stacked Column Chart with Legend Displayed (Used with Permission from Microsoft)

In this stacked column chart, which represents product sales each month, implementing a legend is the most appropriate choice. Because of the size of the columns and the stack, trying to place product labels on the columns would clutter the visual. The legend explains the corresponding color for each product and presents the information in an easy-to-interpret format.

Review Activity:

Design Elements for Reports/Dashboards

Answer the following questions:

1. **Style guides typically contain variations of what dependent on the output?**

2. **What are some considerations when choosing appropriate colors?**

3. **What are the two basic types of fonts?**

4. **What three categories are typically offered for customizing fonts?**

5. **What does not change the underlying field name in the database or data set, but displays a custom name that clarifies the information being presented?**

6. **What should be considered regarding naming conventions when reporting?**

Topic 14B

Utilize Standard Elements for Reports/Dashboards

EXAM OBJECTIVES COVERED
4.2 Given a scenario, use appropriate design components for reports and dashboards.

As a data analyst, part of your role is to not only provide information through reports and dashboards, but also to provide critical information *about* the reports and dashboards. You will find that version numbers, refresh dates, and other special fields will help you provide this information by simply adding them to your work. Page numbers, file paths, and print dates are common and important standard elements to consider for every reporting or dashboard project.

Standard Information and Formatting Elements

Several standard elements provide important information and help users navigate through reports. Many of these components, such as references, are part of the expected report format and are best practices for communicating data in a business setting.

Version Numbers

Version numbers are recommended for reports that are a part of a documented process, or are in production but still in development. When there are multiple iterations of a report, version numbers serve as a valuable reference so that people can immediately identify whether they are looking at the correct version of a given report. Version numbers can be placed in the different sections of a report or on a certain portion of a screen. As a data analyst, the version number will also allow you to confirm that people are using the correct version of the report. It allows you to quickly check without having to look for the various changes that have occurred since the last version.

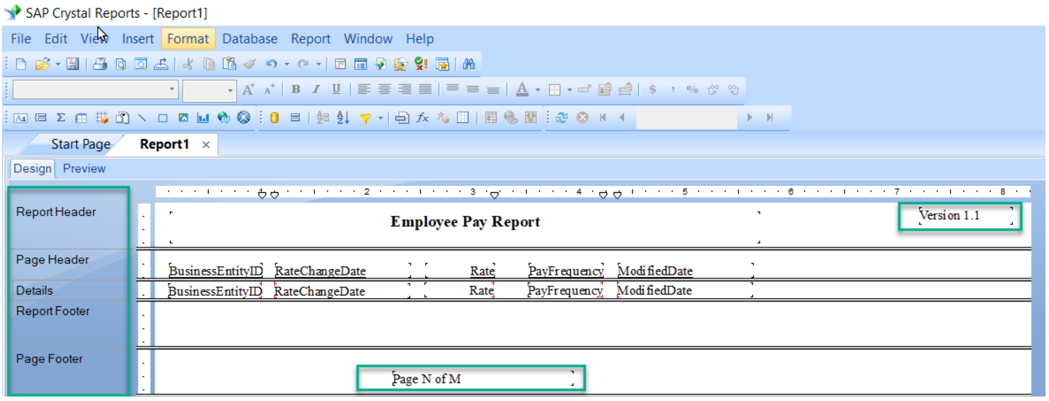
Changes that do happen between versions should be documented. For example, version 1.2 should have a “what’s in this version” list to detail all the changes that are new to this version. This not only provides critical updates, but also makes the update more meaningful, giving us the ability to see how much has changed over time from the original version. Notes on versioning can also help us identify items that might not be complete and slate them for later versions. For example, suppose you have been asked for a new field, but that field isn’t in the underlying data set. You can note that the field will be added in the next version (or whichever version it will potentially be in).

Elements of Paginated Reports

Page numbers are recommended for paginated reports to maintain order when they are printed and used in meetings. Paginated reports will have multiple sections that help users to better navigate and read the report.

- The **report header** appears once at the top of the first page of a report. The report header can be used to title the report, and the version number can be placed on the top right.
- The **page header** is located at the top of each page of a report. This element is a good place to include field headings and information that needs to be on every page, such as a page number.
- The **page footer** appears at the bottom of each page of a report. The page footer is a typical location for references to the data, page numbers, and version numbers.
- The **report footer** appears at the end of the reported data. Sometimes it shows at the bottom of the page before the page footer. This section would not be appropriate for page numbers, since it only appears once when the report concludes.

Crystal Reports is a program that develops paginated reports, and it includes report sections in the Design and Preview tabs of the tool. You can very clearly distinguish the sections of the paginated report in the design view, as shown in this next screenshot. The details section will include the individual line items of the data that repeat over each page.



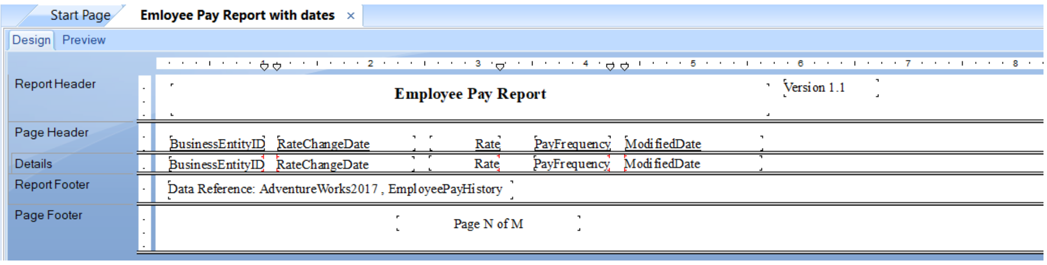
Design View in Crystal Reports Showing Report Sections (Copyright © 2021 SAP SE or an SAP affiliate company. All rights reserved.)

In this example of an employee pay report, the report header contains the title of the report, with the version number in the top right corner. The page header contains different field labels for the information contained in the report's details section. It appears that the detail section and the page header are showing the same fields. However, when in the report view, the page header simply displays the field name label, while the details section will repeat to show the individual field data. The page footer contains a special field in Crystal Reports called "Page N of M." This field will show the current page number of any given page viewed and also the total number of pages for the report.

Reference Data Sources

It is a recommended practice to include **references** to the data sources when designing any given report or dashboard. Including reference information in the report will also save you time and energy as an analyst. You might be asked for reference information if it isn't readily displayed, and this will require you to investigate the design of the report to determine the source of the data.

Data references can be included as additional information in the report footer, as shown in this screenshot of Crystal Reports. You can also use special fields within your reporting program to point to the location of the file and the actual file name of the data source.



Employee Pay Report

Version 1.1

BusinessEntityID RateChangeDate Rate PayFrequency ModifiedDate

BusinessEntityID RateChangeDate Rate PayFrequency ModifiedDate

Data Reference: AdventureWorks2017 , EmployeePayHistory

Page N of M

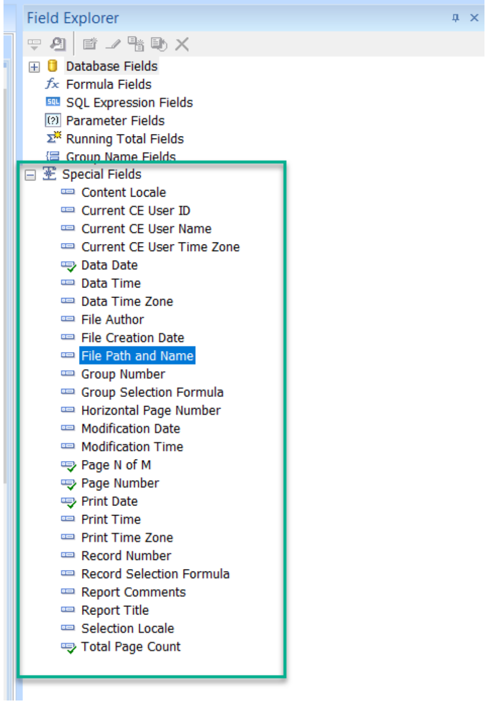
Data Reference in Report Footer in Crystal Reports (Copyright © 2021 SAP SE or an SAP affiliate company. All rights reserved.)

Other Special Fields

Different programs (e.g., Crystal Reports, Excel) contain common special fields that will help maintain information about a report without requiring a lot of time or manual updates. The data analyst and business decision makers should decide what special fields to include in the report or dashboard. Just know that there is not a one-size-fits-all approach for every report or dashboard. Some common special fields might include the following:

- Page numbers
- File paths
- Time zones
- Location
- Specific date types

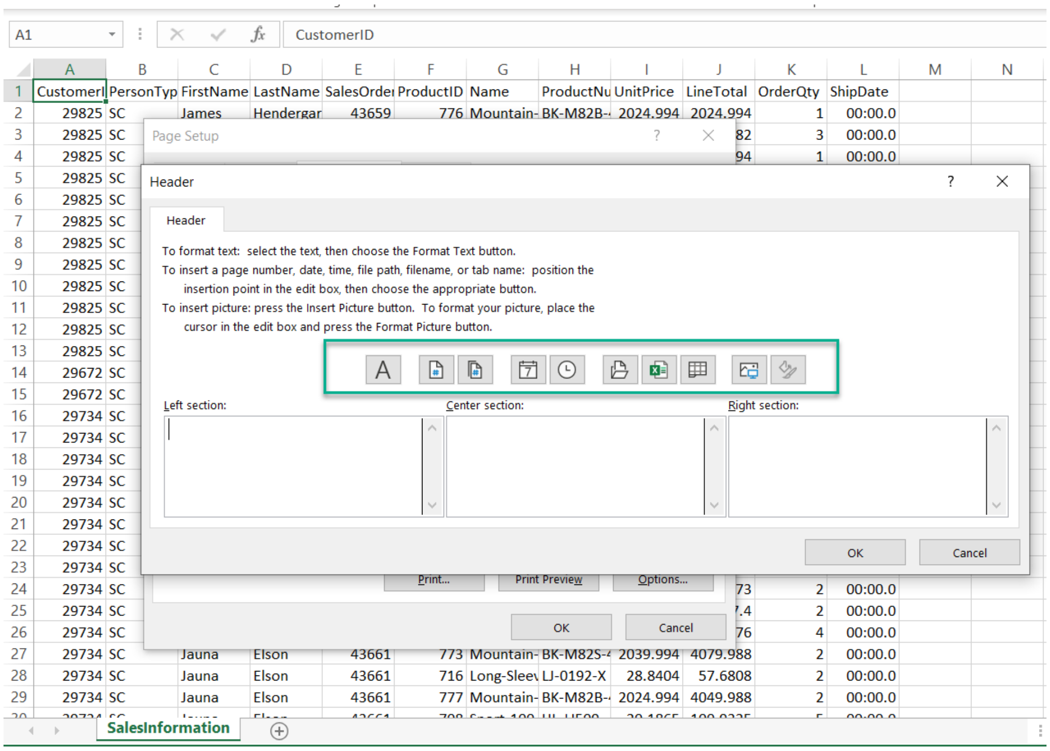
When you add special fields to your report design in Crystal Reports, these values will update automatically as you work with the report.



Database Fields
Formula Fields
SQL Expression Fields
Parameter Fields
Running Total Fields
Group Name Fields
Special Fields
Content Locale
Current CE User ID
Current CE User Name
Current CE User Time Zone
Data Date
Data Time
Data Time Zone
File Author
File Creation Date
File Path and Name
Group Number
Group Selection Formula
Horizontal Page Number
Modification Date
Modification Time
Page N of M
Page Number
Print Date
Print Time
Print Time Zone
Record Number
Record Selection Formula
Report Comments
Report Title
Selection Locale
Total Page Count

Examples of Special Fields in Crystal Reports (Copyright © 2021 SAP SE or an SAP affiliate company. All rights reserved.)

This screenshot shows the special fields that are made available to data analysts who use Crystal Reports for paginated reports. You can include these in any section of the report where it makes sense; the most common locations are the report and page headers and footers.



A1 CustomerID

	A	B	C	D	E	F	G	H	I	J	K	L
1	CustomerI	PersonTyp	FirstName	LastName	SalesOrder	ProductID	Name	ProductNu	UnitPrice	LineTotal	OrderQty	ShipDate
2	29825	SC	James	Hendergar	43659	776	Mountain-	BK-M82B-	2024.994	2024.994	1	00:00.0
3	29825	SC								82	3	00:00.0
4	29825	SC								94	1	00:00.0
5	29825	SC										
6	29825	SC										
7	29825	SC										
8	29825	SC										
9	29825	SC										
10	29825	SC										
11	29825	SC										
12	29825	SC										
13	29825	SC										
14	29672	SC										
15	29672	SC										
16	29734	SC										
17	29734	SC										
18	29734	SC										
19	29734	SC										
20	29734	SC										
21	29734	SC										
22	29734	SC										
23	29734	SC										
24	29734	SC								73	2	00:00.0
25	29734	SC								.4	2	00:00.0
26	29734	SC								76	4	00:00.0
27	29734	SC	Jauna	Elson	43661	773	Mountain-	BK-M82S-	2039.994	4079.988	2	00:00.0
28	29734	SC	Jauna	Elson	43661	716	Long-Sleev	LJ-0192-X	28.8404	57.6808	2	00:00.0
29	29734	SC	Jauna	Elson	43661	777	Mountain-	BK-M82B-	2024.994	4049.988	2	00:00.0

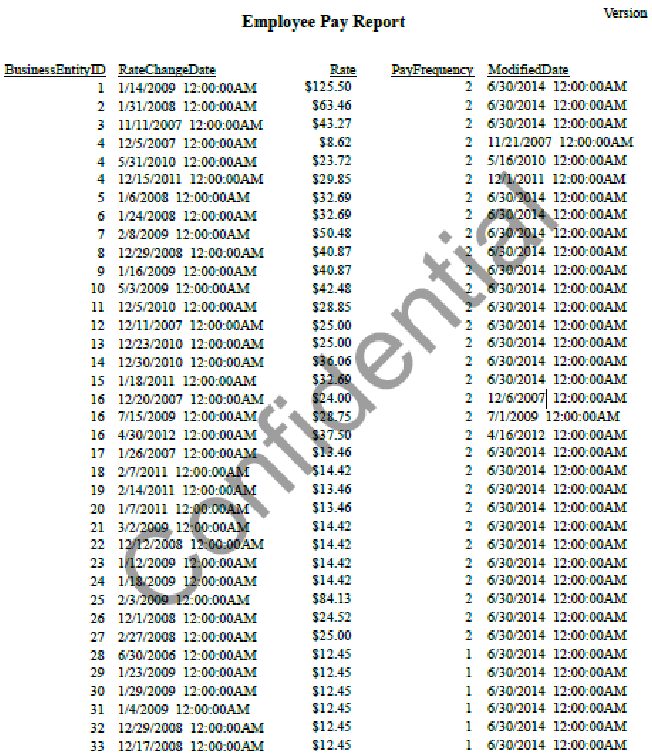
To format text: select the text, then choose the Format Text button.
To insert a page number, date, time, file path, filename, or tab name: position the insertion point in the edit box, then choose the appropriate button.
To insert picture: press the Insert Picture button. To format your picture, place the cursor in the edit box and press the Format Picture button.

Header and Footer in Excel Displaying Special Fields (Used with permission from Microsoft)

In Excel, if you want to include information in the header and footer of the printable page, you can use special fields to add text, page numbers, file names, and other fields without having to manually type these items on the pages.

Watermarks

A **watermark** marks information as required for any given report. It provides a high-level reminder about the information contained in a report. For example, "Confidential" or "Do Not Distribute" watermarks can be added to reports with confidential information that are not to be distributed. A "Not Final" watermark can be used to convey that the information is a draft and is not the final version. Watermarks can also be used as part of a brand design strategy for reports that are made available to the public for use. When the report is viewed or printed, it will contain the appropriate brand marks.


Confidential

Employee Pay Report

Version 1.1

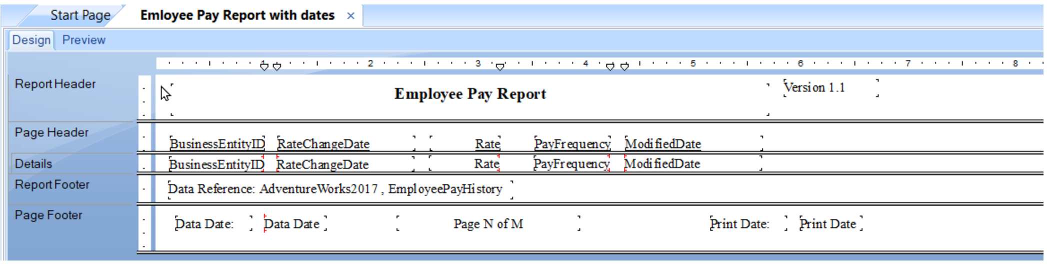
BusinessEntityID	RateChangeDate	Rate	PayFrequency	ModifiedDate
1	1/14/2009 12:00:00AM	$125.50	2	6/30/2014 12:00:00AM
2	1/31/2008 12:00:00AM	$63.46	2	6/30/2014 12:00:00AM
3	11/11/2007 12:00:00AM	$43.27	2	6/30/2014 12:00:00AM
4	12/5/2007 12:00:00AM	$8.62	2	11/21/2007 12:00:00AM
4	5/31/2010 12:00:00AM	$23.72	2	5/16/2010 12:00:00AM
4	12/15/2011 12:00:00AM	$29.85	2	12/1/2011 12:00:00AM
5	1/6/2008 12:00:00AM	$32.69	2	6/30/2014 12:00:00AM
6	1/24/2008 12:00:00AM	$32.69	2	6/30/2014 12:00:00AM
7	2/8/2009 12:00:00AM	$50.48	2	6/30/2014 12:00:00AM
8	12/29/2008 12:00:00AM	$40.87	2	6/30/2014 12:00:00AM
9	1/16/2009 12:00:00AM	$40.87	2	6/30/2014 12:00:00AM
10	5/3/2009 12:00:00AM	$42.48	2	6/30/2014 12:00:00AM
11	12/5/2010 12:00:00AM	$28.85	2	6/30/2014 12:00:00AM
12	12/11/2007 12:00:00AM	$25.00	2	6/30/2014 12:00:00AM
13	12/23/2010 12:00:00AM	$25.00	2	6/30/2014 12:00:00AM
14	12/30/2010 12:00:00AM	$36.06	2	6/30/2014 12:00:00AM
15	1/18/2011 12:00:00AM	$32.69	2	6/30/2014 12:00:00AM
16	12/20/2007 12:00:00AM	$24.00	2	12/6/2007 12:00:00AM
16	7/15/2009 12:00:00AM	$28.75	2	7/1/2009 12:00:00AM
16	4/30/2012 12:00:00AM	$37.50	2	4/16/2012 12:00:00AM
17	1/26/2007 12:00:00AM	$13.46	2	6/30/2014 12:00:00AM
18	2/7/2011 12:00:00AM	$14.42	2	6/30/2014 12:00:00AM
19	2/14/2011 12:00:00AM	$13.46	2	6/30/2014 12:00:00AM
20	1/7/2011 12:00:00AM	$13.46	2	6/30/2014 12:00:00AM
21	3/2/2009 12:00:00AM	$14.42	2	6/30/2014 12:00:00AM
22	12/12/2008 12:00:00AM	$14.42	2	6/30/2014 12:00:00AM
23	1/12/2009 12:00:00AM	$14.42	2	6/30/2014 12:00:00AM
24	1/18/2009 12:00:00AM	$14.42	2	6/30/2014 12:00:00AM
25	2/3/2009 12:00:00AM	$84.13	2	6/30/2014 12:00:00AM
26	12/1/2008 12:00:00AM	$24.52	2	6/30/2014 12:00:00AM
27	2/27/2008 12:00:00AM	$25.00	2	6/30/2014 12:00:00AM
28	6/30/2006 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
29	1/23/2009 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
30	1/29/2009 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
31	1/4/2009 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
32	12/29/2008 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
33	12/17/2008 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM

Confidential Watermark in a PDF of the Report

Important Dates

There are two important dates that should be included with reports and dashboards: when a report is refreshed and when it is printed. The knowledge of when these actions occurred can provide value to the report audience.

The **refresh date** is the date and time that the data was last updated (if applicable). This date lets the audience know that they are only seeing data that was available up to that day and time, and not anything since then. The **print date** of the report tells us, as you might expect, when the report was printed. It's important to remember that the refresh date and print date for a report are not always the same. For example, data might be refreshed on Friday before a meeting, but the report is printed on Monday. Including these two separate dates in the report lets the audience know that the data they are viewing is up to date as of Friday, but the printed format or PDF was distributed on Monday. Programs such as Crystal Reports provide several options and formats for dates and times. Dates can be built into the report design as special fields, as shown in the next screenshot.



Report Header: Employee Pay Report — Version 1.1

Page Header: BusinessEntityID RateChangeDate Rate PayFrequency ModifiedDate

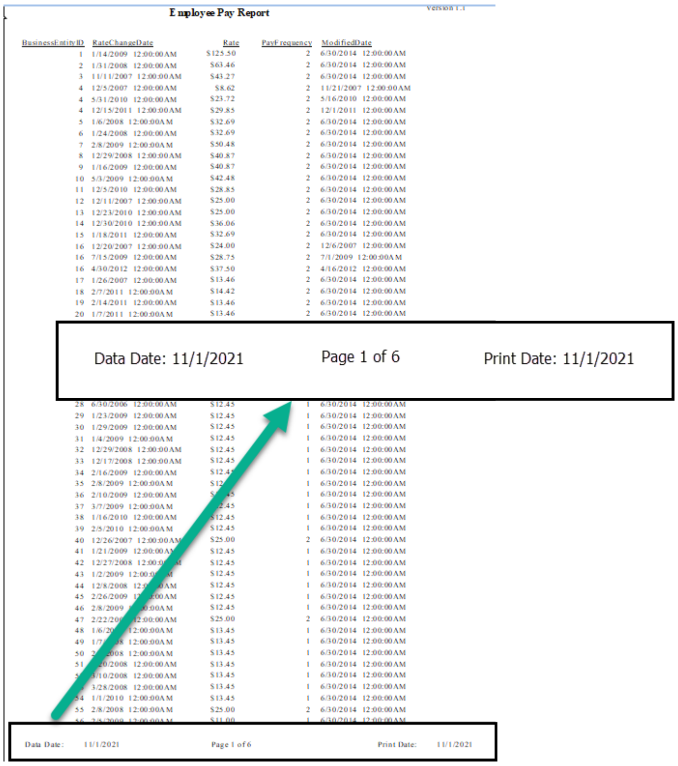
Details: BusinessEntityID RateChangeDate Rate PayFrequency ModifiedDate

Report Footer: Data Reference: AdventureWorks2017 , EmployeePayHistory

Page Footer: Data Date: Data Date — Page N of M — Print Date: Print Date

Data Date and Print Date Shown in Page Footer in Crystal Reports (Copyright © 2021 SAP SE or an SAP affiliate company. All rights reserved.)

The refresh date and print date special fields have been added as a page footer for this report. Varying programs will name their special fields slightly differently, so you should refer to the particular program in use's help detail to confirm what the special field may mean if you are unsure.



Employee Pay Report

Version 1.1

BusinessEntityID	RateChangeDate	Rate	PayFrequency	ModifiedDate
1	1/14/2009 12:00:00AM	$125.50	2	6/30/2014 12:00:00AM
2	1/31/2008 12:00:00AM	$63.46	2	6/30/2014 12:00:00AM
3	11/11/2007 12:00:00AM	$43.27	2	6/30/2014 12:00:00AM
4	12/5/2007 12:00:00AM	$8.62	2	11/21/2007 12:00:00AM
4	5/31/2010 12:00:00AM	$23.72	2	5/16/2010 12:00:00AM
4	12/15/2011 12:00:00AM	$29.85	2	12/1/2011 12:00:00AM
5	1/6/2008 12:00:00AM	$32.69	2	6/30/2014 12:00:00AM
6	1/24/2008 12:00:00AM	$32.69	2	6/30/2014 12:00:00AM
7	2/8/2009 12:00:00AM	$50.48	2	6/30/2014 12:00:00AM
8	12/29/2008 12:00:00AM	$40.87	2	6/30/2014 12:00:00AM
9	1/16/2009 12:00:00AM	$40.87	2	6/30/2014 12:00:00AM
10	5/3/2009 12:00:00AM	$42.48	2	6/30/2014 12:00:00AM
11	12/5/2010 12:00:00AM	$28.85	2	6/30/2014 12:00:00AM
12	12/11/2007 12:00:00AM	$25.00	2	6/30/2014 12:00:00AM
13	12/23/2010 12:00:00AM	$25.00	2	6/30/2014 12:00:00AM
14	12/30/2010 12:00:00AM	$36.06	2	6/30/2014 12:00:00AM
15	1/18/2011 12:00:00AM	$32.69	2	6/30/2014 12:00:00AM
16	12/20/2007 12:00:00AM	$24.00	2	12/6/2007 12:00:00AM
16	7/15/2009 12:00:00AM	$28.75	2	7/1/2009 12:00:00AM
16	4/30/2012 12:00:00AM	$37.50	2	4/16/2012 12:00:00AM
17	1/26/2007 12:00:00AM	$13.46	2	6/30/2014 12:00:00AM
18	2/7/2011 12:00:00AM	$14.42	2	6/30/2014 12:00:00AM
19	2/14/2011 12:00:00AM	$13.46	2	6/30/2014 12:00:00AM
20	1/7/2011 12:00:00AM	$13.46	2	6/30/2014 12:00:00AM

Data Date: 11/1/2021   Page 1 of 6   Print Date: 11/1/2021

BusinessEntityID	RateChangeDate	Rate	PayFrequency	ModifiedDate
28	6/30/2006 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
29	1/23/2009 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
30	1/29/2009 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
31	1/4/2009 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
32	12/29/2008 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
33	12/17/2008 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
34	2/16/2009 12:00:00AM	$12.4[illegible]	1	6/30/2014 12:00:00AM
35	2/8/2009 12:00:00AM	$12[illegible]	1	6/30/2014 12:00:00AM
36	2/10/2009 12:00:00AM	$[illegible]45	1	6/30/2014 12:00:00AM
37	3/7/2009 12:00:00AM	[illegible]2.45	1	6/30/2014 12:00:00AM
38	1/16/2010 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
39	2/5/2010 12:00:00AM	$12.45	1	6/30/2014 12:00:00AM
40	12/26/2007 12:00:00AM	$25.00	2	6/30/2014 12:00:00AM
41	1/21/2009 12:00:00A[illegible]	$12.45	1	6/30/2014 12:00:00AM
42	12/27/2008 12:00:[illegible]M	$12.45	1	6/30/2014 12:00:00AM
43	1/2/2009 12:00:[illegible]M	$12.45	1	6/30/2014 12:00:00AM
44	12/8/2008 12:[illegible]0AM	$12.45	1	6/30/2014 12:00:00AM
45	2/26/2009 1[illegible]:00AM	$12.45	1	6/30/2014 12:00:00AM
46	2/8/2009 [illegible]0:00AM	$12.45	1	6/30/2014 12:00:00AM
47	2/22/200[illegible] 12:00:00AM	$25.00	2	6/30/2014 12:00:00AM
48	1/6/20[illegible] 12:00:00AM	$13.45	1	6/30/2014 12:00:00AM
49	1/7/[illegible]08 12:00:00AM	$13.45	1	6/30/2014 12:00:00AM
50	2/[illegible]2008 12:00:00AM	$13.45	1	6/30/2014 12:00:00AM
51	[illegible]20/2008 12:00:00AM	$13.45	1	6/30/2014 12:00:00AM
5[illegible]	[illegible]3/10/2008 12:00:00AM	$13.45	1	6/30/2014 12:00:00AM
[illegible]	3/28/2008 12:00:00AM	$13.45	1	6/30/2014 12:00:00AM
54	1/1/2010 12:00:00AM	$13.45	1	6/30/2014 12:00:00AM
55	2/8/2008 12:00:00AM	$25.00	2	6/30/2014 12:00:00AM
56	2/5/2009 12:00:00AM	$11.00	1	6/30/2014 12:00:00AM

Data Date: 11/1/2021   Page 1 of 6   Print Date: 11/1/2021

Preview of the Employee Pay Report in Crystal Reports (Copyright © 2021 SAP SE or an SAP affiliate company. All rights reserved.)

The special fields included in our Employee Pay report automatically show the refresh date of the data (currently labeled as "data date" in our report), the print date, and page numbers in preview mode.

Review Activity:

Standard Elements for Reports/Dashboards

Answer the following questions:

1. **What element should you include when reports are a part of a documented process?**

2. **What is the difference between a report header and a page header?**

3. **What can be added to confidential reports which are not to be distributed?**

4. **What two dates are important to include in a report?**

Topic 14C

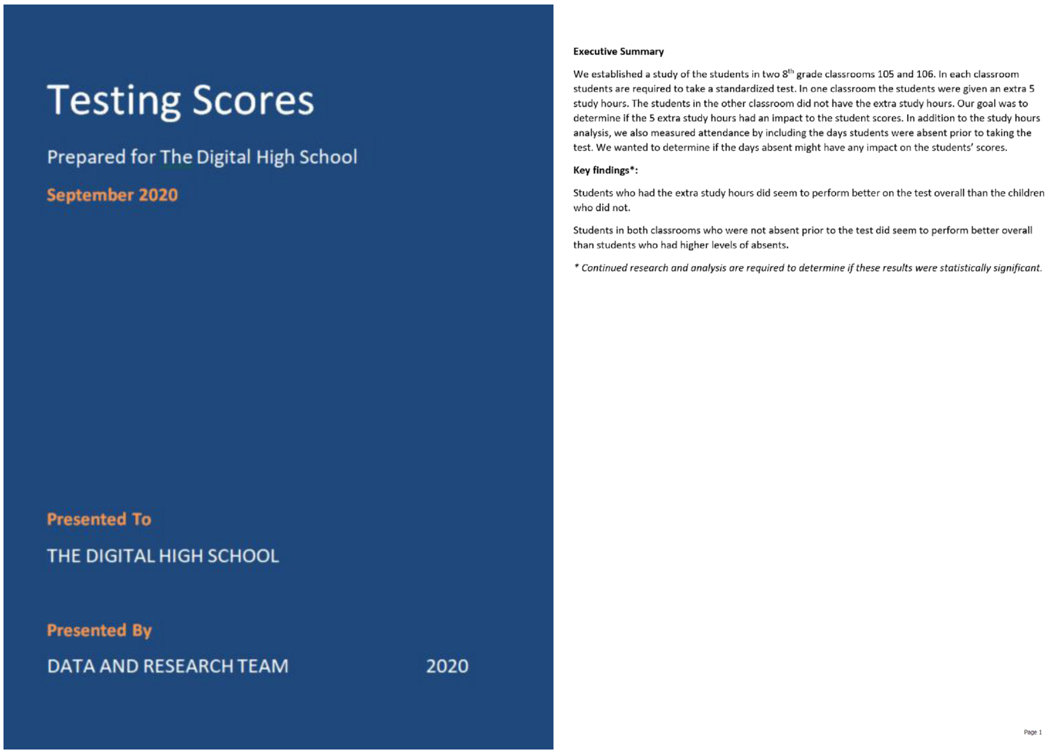
Create a Narrative and Other Written Elements

EXAM OBJECTIVES COVERED
4.2 Given a scenario, use appropriate design components for reports and dashboards.

Writing a narrative for a report or dashboard can provide valuable insights about what the audience will discover in the report. A narrative provides high-level details about the information contained in the report and can include a summary and key findings. The details, style, and elements of the narrative can change depending on the intended audience and how the data will be used. For example, a good deal of narrative is required for scientific studies that include data that will be published by an institution. The studies will include narrative elements such as an abstract, methods, reference citations, and other requirements for written information, outside of just the data outcomes.

Narrative

In any data project, the basic **narrative** for a report typically includes a report cover page and a summary of the contents of the report or dashboard.



Testing Scores

Prepared for The Digital High School

September 2020

Presented To

THE DIGITAL HIGH SCHOOL

Presented By

DATA AND RESEARCH TEAM 2020

Executive Summary

We established a study of the students in two 8th grade classrooms 105 and 106. In each classroom students are required to take a standardized test. In one classroom the students were given an extra 5 study hours. The students in the other classroom did not have the extra study hours. Our goal was to determine if the 5 extra study hours had an impact to the student scores. In addition to the study hours analysis, we also measured attendance by including the days students were absent prior to taking the test. We wanted to determine if the days absent might have any impact on the students' scores.

Key findings*:

Students who had the extra study hours did seem to perform better on the test overall than the children who did not.

Students in both classrooms who were not absent prior to the test did seem to perform better overall than students who had higher levels of absents.

** Continued research and analysis are required to determine if these results were statistically significant.*

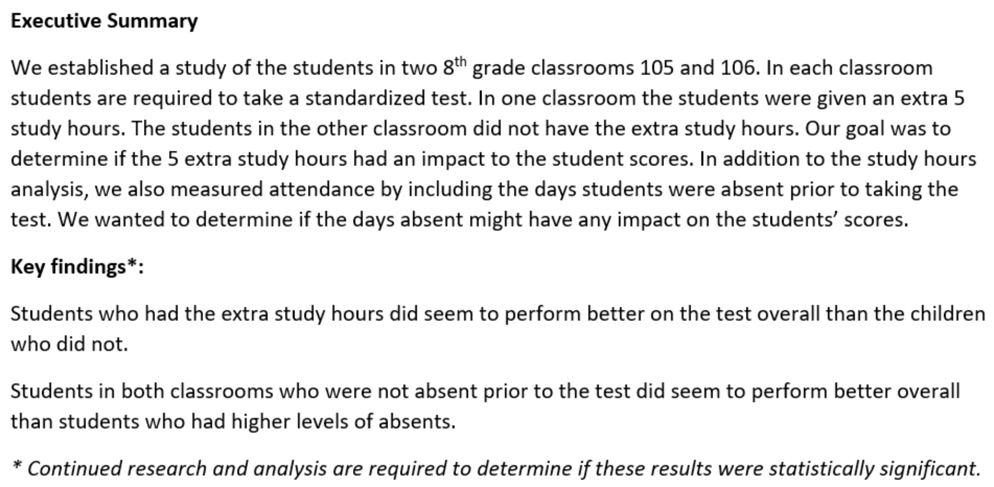
Page 1

Cover Page with Executive Summary using Microsoft Word (Used with permission from Microsoft.)

If you are printing your work, you will most certainly want to have a cover page that can provide the name of the presentation or document. In some cases, the name of the presentation for the cover page might be “Annual Report.” For our example, we are creating a “Testing Scores” document to present to the administration of our school, The Digital High School.

Key Findings

You will also find that when a narrative is included within the report, it will contain key findings that are meant to identify problems or areas in need of improvement, or highlight areas that are going positively.



Executive Summary

We established a study of the students in two 8th grade classrooms 105 and 106. In each classroom students are required to take a standardized test. In one classroom the students were given an extra 5 study hours. The students in the other classroom did not have the extra study hours. Our goal was to determine if the 5 extra study hours had an impact to the student scores. In addition to the study hours analysis, we also measured attendance by including the days students were absent prior to taking the test. We wanted to determine if the days absent might have any impact on the students’ scores.

Key findings*:

Students who had the extra study hours did seem to perform better on the test overall than the children who did not.

Students in both classrooms who were not absent prior to the test did seem to perform better overall than students who had higher levels of absents.

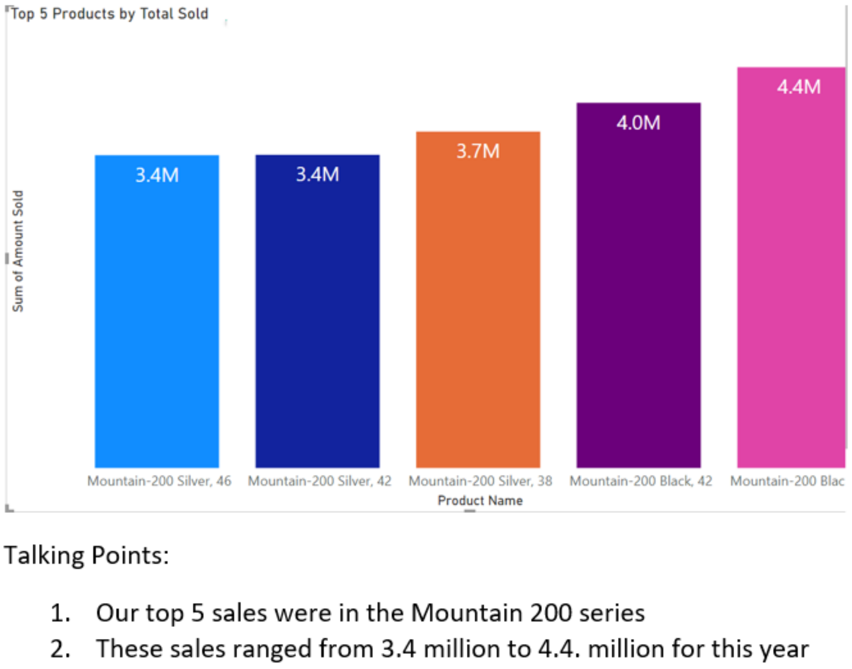
** Continued research and analysis are required to determine if these results were statistically significant.*

Summary and Key Finding Draft for Report Cover Page in Microsoft Word (Used with permission from Microsoft.)

For example, this summary highlights what we studied, providing a high-level detail of the studies performed. You also can see the key findings, with careful notes of what we did find and what is needed to further determine statistically significant results.

Talking Points

Talking points are another useful narrative element to include within a report, especially to assist in meetings and presentations. For example, if your supervisor will be presenting visuals of a report during a meeting, you could include key takeaways about the visual to provide talking points that will support the presenter in explaining the information to the audience.



Talking Points:

1. Our top 5 sales were in the Mountain 200 series
2. These sales ranged from 3.4 million to 4.4. million for this year

Visual Developed in Power BI and Microsoft Word with Talking Points (Used with permission from Microsoft.)

Imagine that a manager is tasked with presenting the information in this chart. Rather than trying to read the chart during the presentation, the talking points will provide the manager with some of the high-level findings to help keep them on track and direct the presentation.

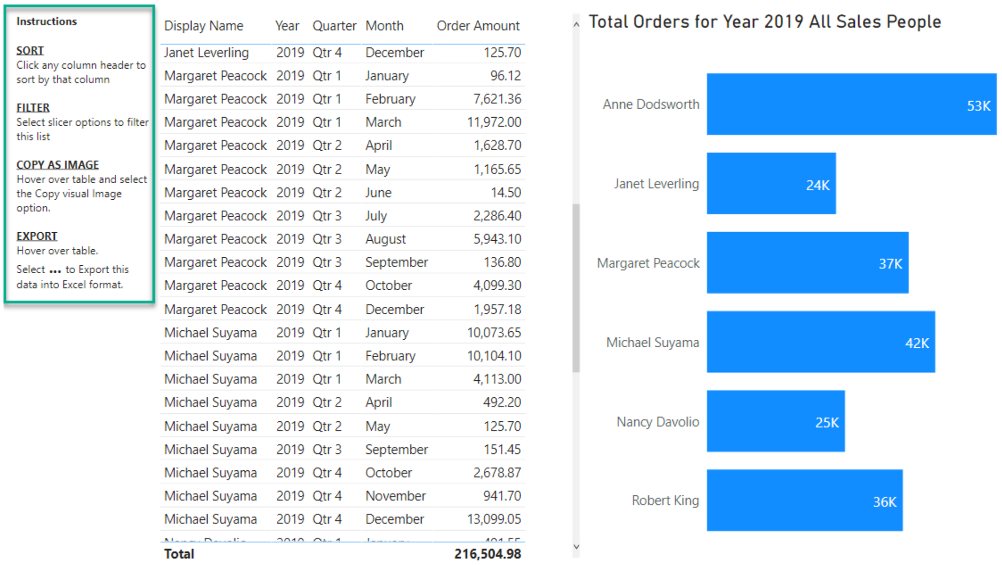
Key Definitions and Calculations

When writing the narrative, you should also include key definitions and insight into calculations where necessary. For example, recall the earlier example in the course involving test scores for a cohort of students. In the report studying student performance, it was important to describe the cohort of students represented by the data, such as the fact that the students were in the eighth grade and in classrooms 105 and 106. When the audience reads this narrative information, they will know that the report data does not include all students in the school, but only eighth graders in these two specific classrooms. You could also consider making note of any calculations that were developed in addition to the data provided. For example, suppose we created a variable for study hours that returned a yes or no in our data set, so we always knew whether a student had those hours or not. In our narrative, we might address this calculation as follows:

> "You will see a field called study hours. This field indicates the students were offered five extra study hours. Only students in classroom 105 were given an option of study hours. If they are in classroom 105, they are marked with a Yes for study hours, and if not in 105, study hours is set to No."

Instructions for Using the Report/Dashboard

Providing instructions for how to use and interact with a dashboard is an area where you can set yourself apart from all other analysts. When we build dashboards, it is easy to forget that we know exactly how they work and exactly what to touch on the dashboard. People are uneasy when they move around on a computer and things change unexpectedly. When a user clicks on an item in a dashboard, it might automatically filter other visuals based on what they selected. This is one of the key features of a dashboard, but if the user is unfamiliar with the technology, they may immediately feel as if they have broken it. Simple screenshots and some written instructions can help support people who work with your dashboard and improve the user experience. This can be done through the use of a training document, or it can also be noted on the dashboard (if space allows).



Instructions

SORT
Click any column header to sort by that column

FILTER
Select slicer options to filter this list

COPY AS IMAGE
Hover over table and select the Copy visual Image option.

EXPORT
Hover over table.
Select ... to Export this data into Excel format.

Display Name	Year	Quarter	Month	Order Amount
Janet Leverling	2019	Qtr 4	December	125.70
Margaret Peacock	2019	Qtr 1	January	96.12
Margaret Peacock	2019	Qtr 1	February	7,621.36
Margaret Peacock	2019	Qtr 1	March	11,972.00
Margaret Peacock	2019	Qtr 2	April	1,628.70
Margaret Peacock	2019	Qtr 2	May	1,165.65
Margaret Peacock	2019	Qtr 2	June	14.50
Margaret Peacock	2019	Qtr 3	July	2,286.40
Margaret Peacock	2019	Qtr 3	August	5,943.10
Margaret Peacock	2019	Qtr 3	September	136.80
Margaret Peacock	2019	Qtr 4	October	4,099.30
Margaret Peacock	2019	Qtr 4	December	1,957.18
Michael Suyama	2019	Qtr 1	January	10,073.65
Michael Suyama	2019	Qtr 1	February	10,104.10
Michael Suyama	2019	Qtr 1	March	4,113.00
Michael Suyama	2019	Qtr 2	April	492.20
Michael Suyama	2019	Qtr 2	May	125.70
Michael Suyama	2019	Qtr 3	September	151.45
Michael Suyama	2019	Qtr 4	October	2,678.87
Michael Suyama	2019	Qtr 4	November	941.70
Michael Suyama	2019	Qtr 4	December	13,099.05
Total				**216,504.98**

Power BI Dashboard with Table and Bar Graph with Instructions (Used with permission from Microsoft.)

In the screenshot, we are providing simple instructions to let users know how to sort and filter their information, as well as how to copy an image and export the data.

Other Supporting Materials

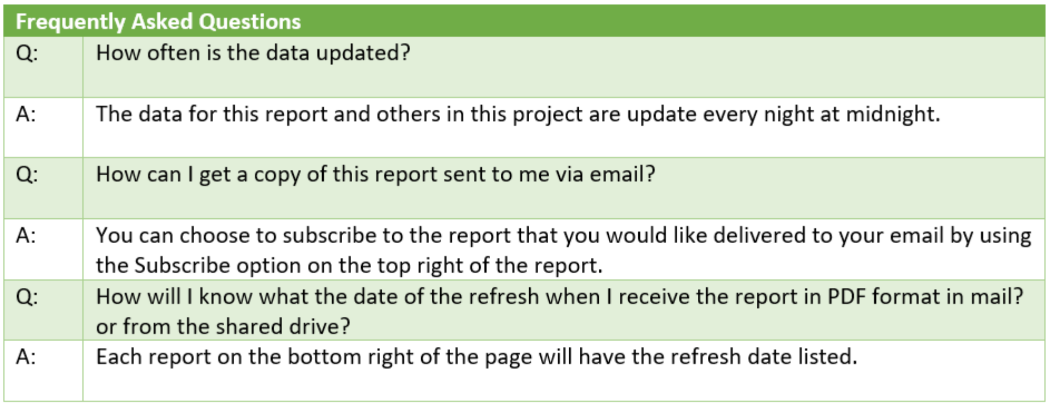
Other supporting materials included as narrative content in the report include Frequently Asked Questions (FAQs) and appendixes.

Frequently Asked Questions (FAQs)

Frequently asked questions, or FAQs, address anticipated, common questions related to the report and its data. Topics of FAQs may include the following:

- When refreshes take place
- The source of the data
- How often people will be able to receive information

FAQs are typically written in a question-and-answer format. They can be displayed in several ways, such as on a printed page or a web page.



Frequently Asked Questions	
Q:	How often is the data updated?
A:	The data for this report and others in this project are update every night at midnight.
Q:	How can I get a copy of this report sent to me via email?
A:	You can choose to subscribe to the report that you would like delivered to your email by using the Subscribe option on the top right of the report.
Q:	How will I know what the date of the refresh when I receive the report in PDF format in mail? or from the shared drive?
A:	Each report on the bottom right of the page will have the refresh date listed.

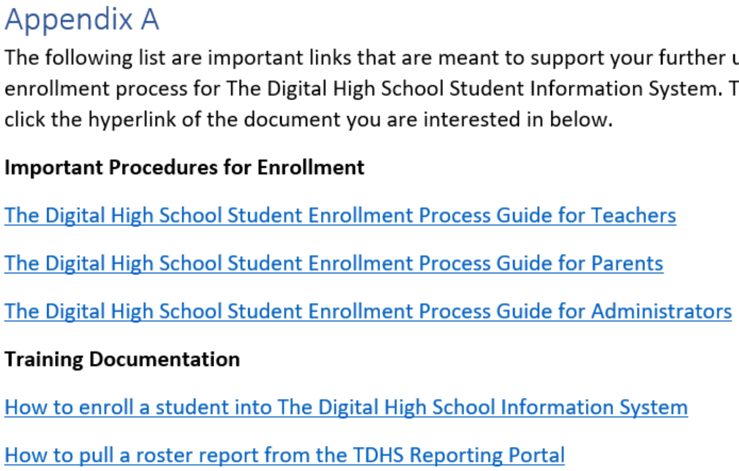
Frequently Asked Questions Displayed in Microsoft Excel (Used with permission from Microsoft.)

This example displays a few questions that may be included, if relevant, for a report or project. How often the data is updated is a very common question; anything that you expect would be asked should go in the FAQs.

Appendixes

An **appendix** is used to provide additional details and information related to the report or process. It is an ideal place to reference materials that support or supplement what consumers are reading or viewing, but that are not essential to the main content of the report. An appendix allows users to access more extensive information as needed without cluttering the report or visuals with too much detail.

If the report involves a process or procedure, that procedure document might be included in an appendix. If multiple procedures are involved, the appendix could include an appendix of links to those procedures, as shown in the screenshot.



Appendix A

The following list are important links that are meant to support your further understanding of the enrollment process for The Digital High School Student Information System. To read the information just click the hyperlink of the document you are interested in below.

Important Procedures for Enrollment

The Digital High School Student Enrollment Process Guide for Teachers

The Digital High School Student Enrollment Process Guide for Parents

The Digital High School Student Enrollment Process Guide for Administrators

Training Documentation

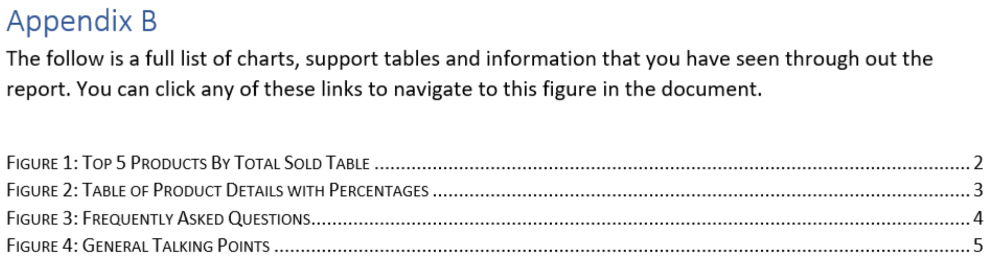
How to enroll a student into The Digital High School Information System

How to pull a roster report from the TDHS Reporting Portal

Screenshot of Appendix A with Important Links in Microsoft Word (Used with permission from Microsoft.)

The sample appendix listed here just provides a list of relevant links. This type of documentation can vary in detail; this is just one sample or style.

Some appendixes may include more detailed tables and figures to support the visuals throughout the report. This type of appendix can be especially useful in long documents that are designed with lots of data outcomes, narrative, charts, and visuals.



Appendix B

The follow is a full list of charts, support tables and information that you have seen through out the report. You can click any of these links to navigate to this figure in the document.

Screenshot of Appendix B with Detailed List of Figures in Microsoft Word (Used with permission from Microsoft.)

The sample here is created by using Insert Caption in Microsoft Word and then using Insert Table of Figures. This is a valuable command for data analysts who are integrating their visuals into reporting documents that have many charts, graphs, and tables throughout the document.

Review Activity:

Narrative and Other Written Elements

Answer the following questions:

1. **Narrative provides what type of detail about the report?**

2. **If you are preparing a dashboard for an audience who is not very familiar with how to work the dashboard, you might supply instructions for its use through which method?**

3. **What is typically written in a question-and-answer format?**

4. **What is used to provide additional details and information that is related, but not crucial, to a report?**

Topic 14D

Understand Deployment Considerations

EXAM OBJECTIVES COVERED

4.3 Given a scenario, use appropriate methods for dashboard development.

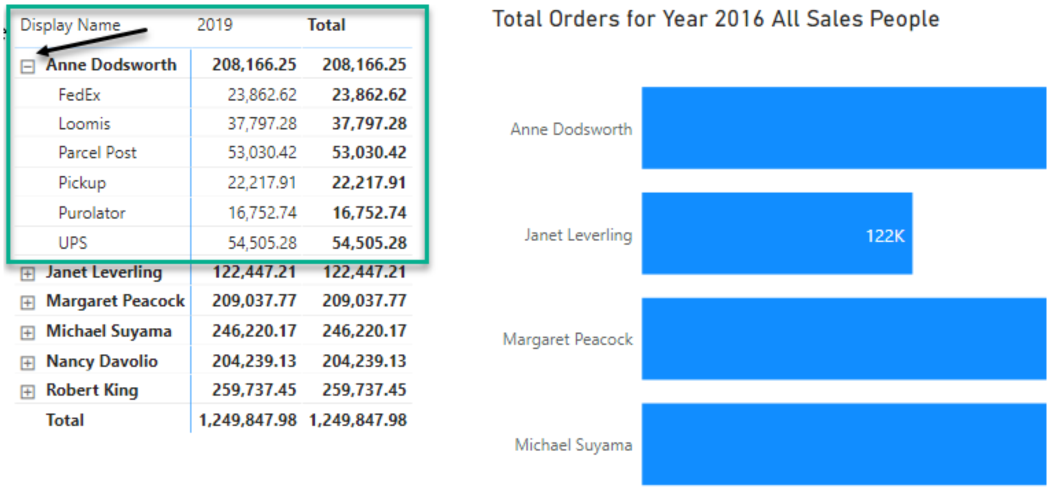
When developing dashboards, we have lots of different technical items to consider. Does the dashboard update efficiently? Does it filter effectively for the user? Do the users have the correct permissions and licensing to be able to use and share the dashboard? These are all questions you should ask yourself as you prepare to deploy to production.

Techniques for Dashboard Optimization

Dashboards and filters have numerous capabilities and options for consumers to read through data and determine information. When creating a dashboard, it's important to provide consumers with access to the data they need, but it's also not feasible to build a million pages in a dashboard to account for every scenario or piece of information. Techniques to optimize the dashboard experience and presentation of data include visual filters, drill-through capabilities, and tooltips.

Visual Filters

Visual filters allow us to provide additional row values, which can be expanded or collapsed to provide more detail without having to create more pages or more visuals. Every tool has some version of this capability; you will need to explore the actual tool to determine how it can be done within that tool.



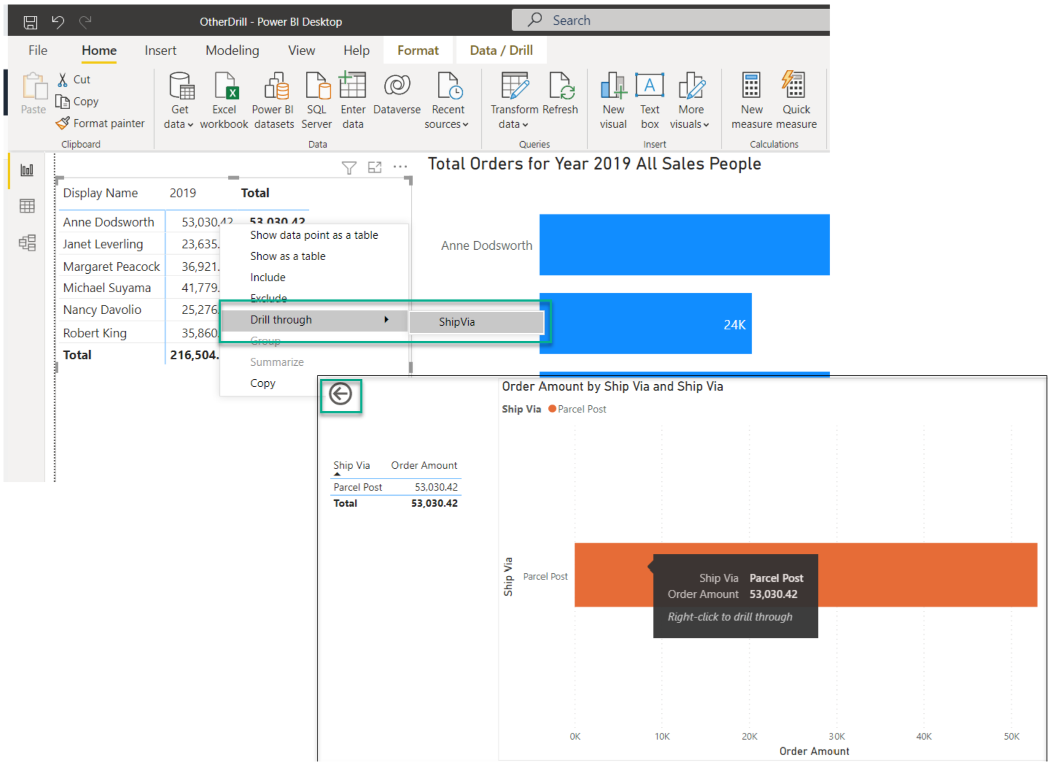
Display Name	2019	Total
⊟ Anne Dodsworth	208,166.25	208,166.25
FedEx	23,862.62	23,862.62
Loomis	37,797.28	37,797.28
Parcel Post	53,030.42	53,030.42
Pickup	22,217.91	22,217.91
Purolator	16,752.74	16,752.74
UPS	54,505.28	54,505.28
⊞ Janet Leverling	122,447.21	122,447.21
⊞ Margaret Peacock	209,037.77	209,037.77
⊞ Michael Suyama	246,220.17	246,220.17
⊞ Nancy Davolio	204,239.13	204,239.13
⊞ Robert King	259,737.45	259,737.45
Total	1,249,847.98	1,249,847.98

Expand Option in Matrix with Bar Graph in Microsoft Power BI (Used with permission from Microsoft.)

In the screenshot, not only can we see Anne's totals, but we can also click the expand option and see each total separated out by each shipper (like FedEx, Loomis, and Parcel Post). This gives us more insight into the total itself.

Drill-Through Pages

Tools like Power BI also have **drill-through capability** that allows you to select a value and drill down to a deeper visualization of the information you selected. This prevents the user from having to jump to multiple pages and filter the visual. They are able to instead view another visual, filtered with the drill-through capability.



Display Name	2019	Total
Anne Dodsworth	53,030.42	53,030.42
Janet Leverling	23,635.	
Margaret Peacock	36,921.	
Michael Suyama	41,779.	
Nancy Davolio	25,276.	
Robert King	35,860.	
Total	**216,504.**	

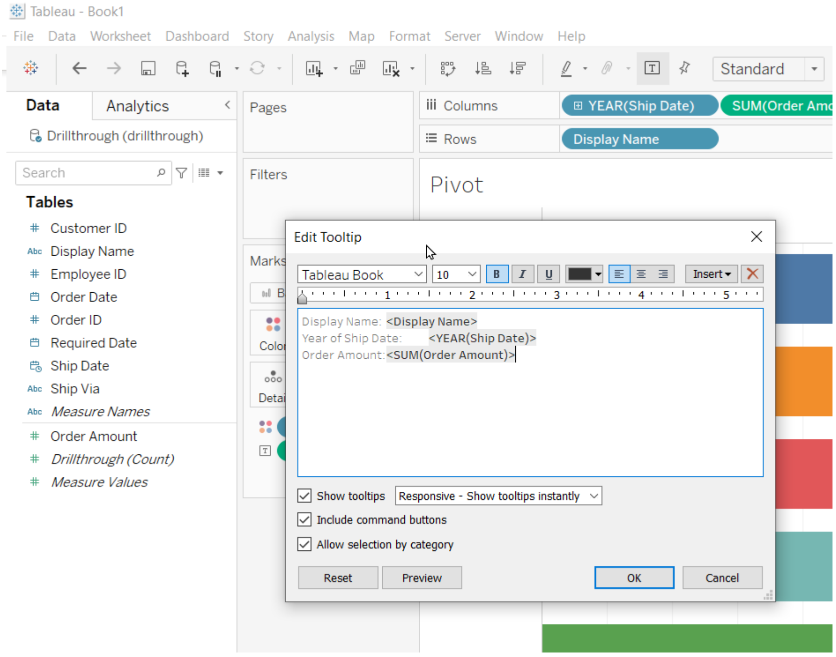
Ship Via	Order Amount
Parcel Post	53,030.42
Total	**53,030.42**

Drill-Through Values in Microsoft Power BI (Used with permission from Microsoft.)

In our drill-through example, we see Anne's value in the matrix. We can drill through this value to see other information about what makes up that particular number, by leveraging drill through to another visual.

Effective Tooltips

When you build a visual, you will notice that it contains **tooltips** by default. Tooltips can provide additional information to the user who simply hovers over a value. You can build a tooltip for almost any visual you create to provide additional information that might be valuable for the user.



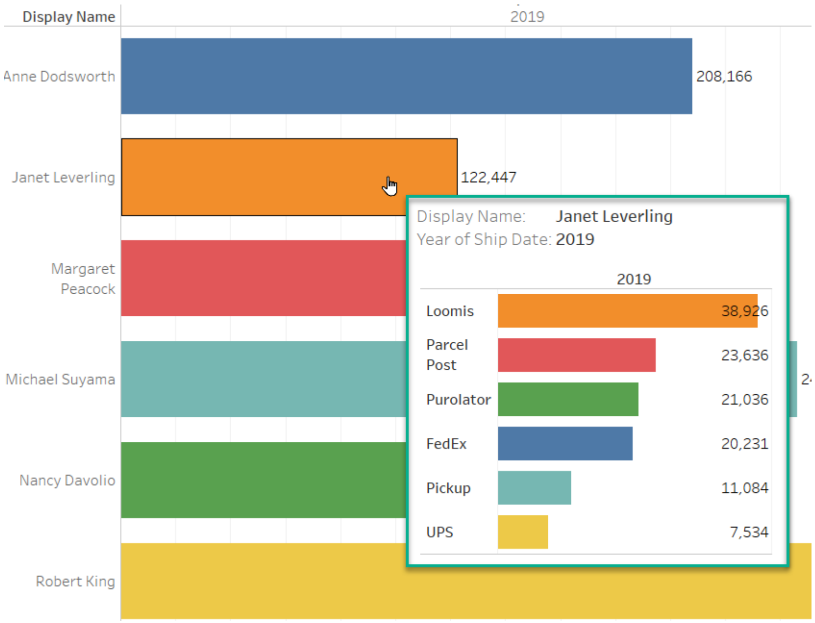
Display Name: <Display Name>
Year of Ship Date: <YEAR(Ship Date)>
Order Amount: <SUM(Order Amount)>

☑ Show tooltips Responsive - Show tooltips instantly
☑ Include command buttons
☑ Allow selection by category

Default Tooltip in Tableau Desktop

The default tooltip in this example shows Display Name, or the name of the customer; YEAR(Ship Date), or the year in which the order was shipped; and SUM(Order Amount), or the total order amount. This information is based on the values that are used to create the visual.

There are times when providing other visuals as the tooltip makes more sense than just the default tooltip. In tools like Power BI and Tableau, you can set the tooltip to show another visual or table that is in the file. These tabs are hidden from view when the dashboard is published, but always available on a hover. In this next screenshot, when we hover over Janet's bar we see another visual that shows the breakdown of what Janet's shipping choices have been. Loomis immediately stands out at the top of her values.



Custom Visual Tooltip in Tableau Desktop

Other Considerations

When discussing optimization with dashboards, just like with databases and queries, we make decisions that can impact the user experience. For example, data refreshes can take some time to process depending on how much data you need to update. If your data doesn't change every second, then you should consider how often it needs to refresh. Another consideration is how you connect to your data—either through a live connection or import connection—as this can impact performance. Row-level security can conveniently provide filters to different components of the report, but it can have a negative impact on performance. As a data analyst, you will learn the nuances of each tool that you use as you attempt to design meaningful experiences that run smoothly for the users.

Deploy to Production

When our work on our reports and dashboards has been completed, and they're ready for our audience, it is time to deploy our work to production. This can mean a lot of different things depending on your organization. It is best defined as going from a proof of concept that has been tested and validated to actually providing a useable product that has been deployed for others to use. There are some key action items to perform when you are ready to deploy to production. Before deployment, you should ensure that:

1. Your documentation and instructions have been prepared.
2. Everything is named and labeled correctly.
3. The items function as intended and have been tested for use by someone other than you. (This is important because you know how you intended it to work, meaning you might miss something that another person might do with your dashboard to create an unintended result.)
4. You have confirmed through the organization's processes that approval has been granted.
5. You have confirmed that permissions have been adequately defined and are set.
6. Required licenses, when applicable, have been assigned to the appropriate users.

Let's dive deeper into that last point.

Licensing Requirements

You will always want to learn up front what licenses are required for reporting at your organization. There are multiple types of licenses and features that are applicable at each level. Let's say, for example, that you have created a dashboard and it's ready to go. You don't want to find out at this point that you can't deliver the final product because no one has licenses--or worse, there are no licenses that can be given to them. This can happen due to cost, or when the purchased licenses are already assigned to other users.

As an example, you can purchase Tableau Desktop and a Creator account, but to share your reports with other creators, they need to also have the licenses from Tableau that allow them access the reports you share. To share and give access, you must remember that =the consumers who read reports have different license requirements, but they will still likely require some type of license to view the work.

You can download PowerBI Desktop for free, but if you want to share reports freely, this comes with different levels of licensing. There are also different licenses required to share the reports within SharePoint. When you are ready for production, this is not the time to discover that licenses are needed. You will want to know ahead of time, so that you can design your reports and seamlessly deploy to production.

Review Activity:

Deployment Considerations

Answer the following questions:

1. **What are three techniques we use to optimize the dashboard experience and presentation of data?**

2. **What action should you take before deployment to production to ensure you can share your report with others?**

Lesson 14

Summary

After this lesson, you should have a better understanding of the design elements that should be considered for every data project. You should know how to follow a brand or style guide, and should be familiar with some simple guidelines to follow when an organization does not have a style guide. You should also be able to identify standard elements to be included in any report or dashboard, like version numbers, print dates, and refresh dates. You should have a better understanding of what types of written elements might support your overall project, and should have learned why providing frequently asked questions can save you time. You should also be able to identify items that will be on your checklist for deployment to production, like testing, permissions, and licenses.

Guidelines in Designing Components for Reports and Dashboards

Consider these best practices and guidelines when familiarizing yourself with the various considerations you should have when preparing a report or dashboard.

1. When a company adopts a brand identity, they will want that brand reflected in all the artifacts that represent the company.
2. Style guides commonly contain different variations of an organization's logo and guidelines for how it can be used.
3. Avoid distracting color schemes (e.g., the use of too many different colors) that detract from the information presented.
4. When adhering to color requirements in a style guide, use the hexadecimal code or the RGB values to ensure the exact color is selected.
5. In a serif font, the letters have lines that make the font more readable when there are paragraphs of text at smaller font sizes, such as the written narrative of a report.
6. In a sans serif font, the letters do not have these lines and can be useful for headings, labels, and other elements surrounding visuals.
7. Captioning allows you to designate more meaningful names for fields in a report or dashboard without changing the underlying field.
8. Labels help consumers easily distinguish what they are viewing and understand the visual faster; legends should show users which color represents which value in a visual.
9. Version numbers are recommended for reports that are a part of a documented process or that are in production, but still in development.
10. The refresh date and print date are not always the same.

11. FAQs address common questions that you would anticipate to receive related to the report and its data.

12. Techniques to optimize the dashboard experience and presentation of data include visual filters, drill-through capabilities, and tooltips.

13. You will want to identify up front what licenses are required for reporting at your organization. There are multiple types of licenses and features that are applicable at each level.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 15

Distinguishing Different Report Types

LESSON INTRODUCTION

As an analyst, you will discover that there are differences in the types of reports you may create. Some will be static, and some will be dynamic. This designation refers to the way that the data refreshes, and what this means for the data analyst is either the audience will serve themselves from the dashboard you created, or you will update your reports and provide them a static copy. You will also receive one-time requests, which is referred to as ad-hoc reporting. You may find that these requests can sometimes lead to regular reporting for the organization. Whether fulfilling a one-time request or running a routine report, you will also need to consider the timing of the reporting, whether it covers a specified period of time or just a point in time.

Lesson Objectives

In this lesson, you will do the following:

- Differentiate between static and dynamic report types.
- Determine the uses for ad-hoc and self-service reporting.
- Explain the importance of timing in reporting.

Topic 15A

Understand How Updates and Timing Affect Reporting

EXAM OBJECTIVES COVERED

4.5 Compare and contrast types of reports.

A report that has to be manually updated by the analyst and then provided to a consumer is a perfect example of a static report. However, when you have access to the data through your permissions, and you can directly update your report in real time, this is dynamic reporting. As an organization looks at their past and present to predict future outcomes, the data analyst is tasked with creating reports that cover a point in time. These points in time can be a week, a month, or a specific month this year compared to that same month last year.

Static Versus Dynamic Reports

A **static report** is a report that does not update automatically. A **dynamic report**, also known as a real-time report, is connected to the data and can be refreshed on demand or regularly updated automatically. There are many similarities between the two types of reports.

- The data we capture and aim to report on comes from the same database, data set, or source.
- We develop the same types of the data joins for both types of reports.
- The way we visualize the data will be the same.

The key difference between the two types of reports is that a static report doesn't update, whereas a dynamic report can be refreshed to visualize new data. In order to see new data in a static report, the data analyst must run a new report.

The use of static versus dynamic in terms of reporting is also sometimes used to refer to the consumer's ability to interact with the report. A dynamic report may be dynamic in the sense that a consumer can refresh and filter data, thus creating various scenarios from the data. Meanwhile, a static report is set to only show a specific base of information, with no additional filtering and refresh capabilities.

A standalone spreadsheet that is not connected to a database would be an example of a static report. A Power BI dashboard that is directly connected to an SQL server would be dynamic. You may find that you, as a data analyst, have the ability to create dynamic reports, whereas others need a static version of your report or dashboard. You will want to create dynamic reports as much as possible, even if its just you that ends up using them.

Point-In-Time Reporting

A **point-in-time report** reflects on a specific point in time, whether looking at data from the most recent month or data from one year prior. It is important to understand that this type of report can be either static or dynamic. When an analyst runs a point-in-time report each day, prints that report, or sends a PDF via email, it is static. When an analyst instead delivers a report that allows the consumer to refresh and gain new data every day, this report is dynamic (but still point in time, as it is still providing daily data).

Point-in-time reports are frequently used for presentations and meetings, and they are typically scheduled to be run on a set schedule. An annual report is a prime example of a point-in-time report, as it is a look back at the previous year and the performance measures that have been set. Another example would be a report of last month's sales shown at a monthly sales meeting to help forecast the next month's sales volume. We might develop and use multiple point-in-time reports when decisions need to be made for the future based on past information.

Real-Time Reporting

Data is always changing, and a **real-time report** will automatically provide the most up-to-date data. For example, some data warehouses load the updated data every night so that the next morning all the reports contain the most up-to-date data from the day before. This would be an example of a real-time report. Depending on the rules of the organization, the variations of "real time" will need to be defined, and the analyst will need to understand how often data is actually updated. The consistent known is that real-time data changes and is updated regularly.

Real-time data is particularly useful for dashboarding. When we can refresh our dashboards and see what's happening in "real time," we have a greater opportunity to get in front of issues and make better decisions (particularly when it comes to financials). For example, suppose your organization's technology department has a help desk call center. Displaying real-time data in a dashboard could be incredibly helpful for the department. Consider that the team sees on the dashboard that there's been a sudden spike in call volume. Being able to see this in real time would operationally help the organization identify the presence of an issue pretty quickly.

Many modernized cloud-based software programs, like CRM systems, provide real-time information. As soon as a new contact has been engaged, that information is made readily available to all reports. It's important to remember that not all data is suitable for dashboards. While the aggregate of transactional data might be relevant for a dashboard, the line-by-line nature of the individual data may lend itself for other styles of reporting.

Review Activity:

How Updates and Timing Affect Reporting

Answer the following questions:

1. **What is the key difference between static reports and dynamic reports?**

2. **A report that covers a specified period of time would be referred to as what type of report?**

3. **A Power BI dashboard directly connected to a SQL server would be an example of which type of report?**

Topic 15B

Differentiate Between Types of Reports

EXAM OBJECTIVES COVERED
4.5 Compare and contrast types of reports.

There are many different types of reports you will work with as an analyst. Compliance and operational reports are extremely common, and typically run on a recurring basis. When more detailed information is needed about operational details, especially for strategic business initiatives that cover long periods of time, you may use tactical reporting. When you want to use research and data to inform and even change business practices, you would create research-driven reports. Sometimes, you'll receive one-time requests for a report that doesn't already exist. These ad-hoc reports can turn into regularly used reports if they truly provide meaningful information. In some cases, you won't even be the one running reports; self-service reports allow users to run their own reports, and it removes the need for a middleman.

Operational and Compliance Reports

Reports that provide crucial data to an organization are often run on a recurring basis. Two types of reports that fit this criteria are operational reports and compliance reports.

An **operational report** is used to inform on the health and status of a project, product, or organization. For example, a company that has identified key performance measures and key performance indicators (KPIs) will likely want to see this data regularly. There are many metrics that an organization will generate about its business. These metrics can become KPIs, but not all metrics *are* KPIs. A metric is designated as a KPI when that metric can be used with other KPIs to improve the overall health of an organization.

A **compliance report** is a report that must be run for compliance or regulatory reasons and includes financial reports, health reports, and safety reports. Consider that organizations in the United States must remain compliant with Occupational Safety and Health Administration (OSHA) standards. As part of these standards, businesses are required to report various type of injuries and near misses, and must hold safety meetings. It is not enough for an organization to "say" that the safety meetings were conducted. They must have data that proves the meetings were held and that people attended. This means the organization is being compliant to the standards. Not unlike OSHA compliance, public corporations also have compliance and oversight agencies, like the U.S. Securities and Exchange Commission (SEC). There are additional standards and reporting requirements that must be met to be considered compliant with these regulations.

In general, any report that helps support a company's operations or covers regulations will be run regularly and routinely.

Tactical and Research-Driven Reporting

When we work with an organization to build reports, we will encounter all types of requests. We will create different styles of reports, like static or dynamic. Some of the reports will be tactical, when we are trying to achieve operational success by monitoring our process. However, they may also be research-based, when we are trying to analyze the "why" behind the process or make improvements.

Tactical Dashboards

A **tactical dashboard** is focused on the operational details of a process or operation. Strategic business initiatives that cover long periods of time and are focused on overall improvement of the business must be monitored and measured to be effective. For example, if a company wants to trim approximately three days off the lead time between sales to delivery over the next year, it will benefit from implementation of a tactical dashboard, or reports that show how the team is doing on this specific metric.

In this scenario, the tactical dashboard is created for the employees who manage the day-to-day processes that the company is trying to improve. Tactical dashboards can also be used by the leaders of a department to watch and track overall performance, so they can identify when performance is below expectations. The reports in tactical dashboards are needed on a regular basis, and real-time data is crucial when progress toward a long-term goal is being monitored.

Research-Driven Reports

A **research-driven report** relies on research to inform and even change business practices. Research-driven reports go beyond simply reporting on the overall health and status of an organization, aiming instead to drive business and move the organization to new goalposts. These reports may happen as part of a specific project, but they do not follow a pattern of reporting, such as we see with day-to-day tactical reports.

Let's revisit our strategic initiative of reducing the lead time between sales and delivery. In trying to figure out how to accomplish this goal, we might study the current lead time in an attempt to determine why it may be longer than we desire. Research studies, and their associated reports, can steer us toward changes we might be able to make to help reach that initiative. Research-driven reports are also useful when we need to figure out why a process has failed to work as intended or has other problems. This type of research analysis is called **root cause analysis**, as it attempts to find the root cause of the problem that occurred.

Ad-Hoc Reporting

A report generated in response to a one-time request is known as an **ad-hoc report**. These reports are generally time sensitive and typically require a quick turnaround. Ad-hoc reports don't already exist in a system, but instead need to be created by someone with the right skills. Data analysts create a lot of ad-hoc reports because they typically have greater access to data.

An ad-hoc report may be requested when someone needs information on a certain subject, topic, or product that is not represented in the reporting that already exists. This type of report is used once and never revisited.

Let's consider an example. Suppose an operations manager wants to determine if the production time of a particular product has increased. This is important because an increase in production time means that this product costs more to produce, and thus lowers the overall margins made when the company sells the product. This report doesn't exist—some of the information needed might be contained in other reports, but the operations manager requests a report that provides all the necessary information in one spot. This is an example of a one-time request that would lead to an ad-hoc report.

In some cases, ad-hoc reports can become standard reports. If a one-time report turns out to be useful on an ongoing basis, and people begin to depend on the information presented, what started as an ad-hoc report can become an "all the time" report. Returning to our example, suppose after receiving the ad-hoc report, the operations manager realizes the report is extremely helpful when looking at margins. If the operations manager asks for this to be an ongoing report that is run on all products with lower margins, the ad-hoc report has become a standard report.

Ad-hoc reports are created in response to a quick request for information that doesn't already exist. Besides that differentiator, ad-hoc reports have all the features of any other type of organizational report. They can be static or dynamic and point in time or real time.

Self-Service Reporting

A **self-service report**, or on-demand report, is one that is run directly by the consumer. When business users can leverage dashboards, or run their own reports from the systems the organization has purchased, they are performing what we call self-service.

In the early days of business, a request for a report would always come into the technology department, which would require someone to create and then email the report (or save it in a shared folder). The technology department was the middleman between the user and the data. Data analysts with access to the data are certainly able to benefit from the self-service model, as well as other business users who are familiar with working with data. Self-service reporting breaks down barriers to accessing data and limits delays due to all requests for data going to one person in the technology department.

Dashboarding has become a phenomenon in the data industry, because once a dashboard is built, other users can serve themselves with the information from that dashboard. Dashboards allow people without a data analysis background to interact with the data and create their own reports whenever they would like, with premade designs created by the data analyst that provide plenty of insight into the data.

Review Activity:

Types of Reports

Answer the following questions:

1. **Compliance reporting is performed to adhere to rules and standards required by certain regulatory agencies and organizations. Provide some examples of regulations that would require reporting for compliance.**

2. **Which type of report uses metrics to provide a picture of the overall health of the organization?**

3. **Tactical dashboards focus on what part of a process or operation?**

4. **Which type of report relies on research to inform and even change business practices?**

5. **A report generated in response to a one-time request is known as what?**

Lesson 15

Summary

You should now be able to determine when a report is static versus dynamic, and have a better understanding of point-in-time reporting. You should have more insight into how often a data analyst will use dynamic report capabilities to create static reports for others. You should also understand that real-time reporting may have various meanings in an organization, and that data is sometimes available the minute its entered, like in popular cloud-based programs. You should be familiar with operational and compliance reports, which are commonly run on a recurring basis, and tactical and research-based reports, which use common tools but have distinct types of reporting requirements and are meant to serve different areas of focus. You have discovered that ad-hoc requests are usually focused asks from different consumers, typically with a short turnaround time, for one-time reports. However, recently organizations have begun deploying data and access to their staff directly, so that people can serve themselves with data.

Guidelines in Distinguishing Different Reports Types and Recognizing Requests for Reports

Consider these best practices and guidelines when familiarizing yourself with the various reports you will be working with.

1. The key difference between static and dynamic reports is that a static report doesn't update, whereas a dynamic report can be refreshed to visualize new data.

2. A standalone spreadsheet that is not connected to a database would be an example of a static report.

3. When an analyst delivers a report that allows the consumer to refresh and gain new data every day, this report is dynamic.

4. When an analyst runs a point-in-time report each day, prints that report, or sends a PDF via email, it is static.

5. An organization will have a mix of true real-time data and data stored in warehouses that updates in near real time, or at different intervals during the day.

6. Real-time data is great for dashboarding, as this allows us operationally to see items as they are occurring in the data.

7. Key metrics can become key performance indicators when they are extremely useful in indicating the health of the organization, group, or process.

8. Tactical dashboards are typically focused on operational information for overall monitoring and improvement for an operation.

9. Research-driven reports usually will seek to inform or improve upon a particular program or process.

10. Root cause analysis is a research analysis that seeks to determine the real cause of an issue, such as a problem in a process or procedure.

11. Data analysts will commonly conduct ad-hoc reporting as a result of one-time requests for specific information that is not already represented in a report or dashboard.

12. Ad-hoc reports can become regularly accessed reports if the information proves to be extremely useful to the organization.

13. Self-service reporting allows users to pull and work with the data themselves, without requiring another person to gather and send it.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 16

Summarizing the Importance of Data Governance

LESSON INTRODUCTION

When a company adopts a data governance plan, it involves the people, processes, and technology needed to control data. Data governance aims to ensure that data maintains its quality and integrity by establishing definitions, rules, and standardization. Data governance also allows for the organization to adhere to regulations and helps maintain compliance. Data governance impacts the quality of data in a very positive way for the data analyst, as it sets the rules of the data from the top level to all levels of the organization.

Lesson Objectives

In this lesson, you will do the following:

- Define data governance.
- Understand access requirements and policies.
- Understand security requirements.
- Understand entity relationship requirements.

Topic 16A

Define Data Governance

EXAM OBJECTIVES COVERED

5.1 Summarize important data governance concepts.

Data governance involves the controls, processes, and people that ensure quality and secure data. It can also be said that data governance encompasses the processes, procedures, and protocol involved in managing the availability, usability, and integrity of data. Data governance includes roles that cover different areas of the data, from ownership to stewardship. These roles play a major part in ensuring data quality as well as data classifications. Data governance provides processes and accountability, ensuring compliance and also awareness of the jurisdictions in which the organization is legally responsible.

Data Governance

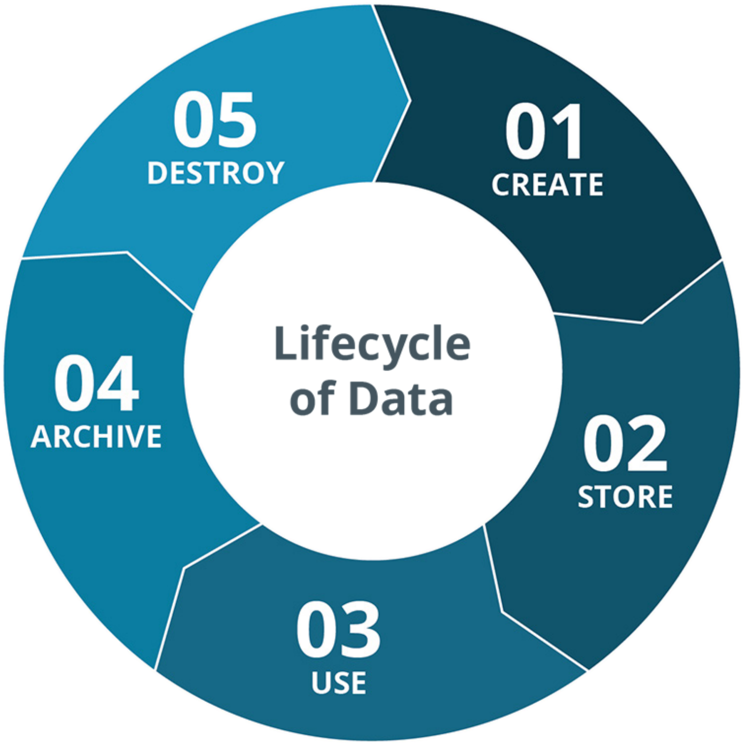
Data governance is a large umbrella term for a framework used to govern data in an organization. Data governance covers the technology, people, and processes that help control data within an organization during its different lifecycles. For a data governance plan to be effective, it must involve every department in an organization. While all organizations implement some form of data governance, it is typically more mature in industries that are regulated or required to meet specific standards. There are measures of maturity that help indicate where an organization may be in its overall path to full data governance. This model, like all things, varies depending on the provider of the model. However, the typical range goes from "completely unaware" to "effective," effective meaning that the organization has reached its goal of information and data management.

Transparency, accountability, and standardization are key elements of a strong data governance plan.

- **Transparency** means everyone in the organization has access to the data governance policies and understands why they are in place.
- **Accountability** means these policies are being followed, and there are accountability measures in place to ensure the policies are followed.
- **Standardization** means data is consistently labeled, categorized, and described for use within the organization.

The Lifecycle of Data

Data has a lifecycle. It's created, stored, used, archived, and deleted. Each stage in the **lifecycle of data** has different rules and requirements for the data an organization will work with related to the regulations and compliance requirements for the industry. There are departments within each organization that have various needs and requirements that are set based on this lifecycle.



The Lifecycle of Data

- **Stage 1 Create**: Data is created through manual entry, software interfaces, external feeds, automated capture, databases, file systems, and many other methods.
- **Stage 2 Store**: The storage of data refers to the locations used to house data, which inform the way we protect data, give permissions, and set up data backup and recovery plans.
- **Stage 3 Use**: Data is used to support operational needs and objectives. Data is viewed, manipulated, processed, and saved. Data will also be overwritten, or values deleted, as needed. Data may also be shared outside the organization.
- **Stage 4 Archive**: When data is not regularly needed, it is archived. It can still be accessed, if necessary, but it is no longer in everyday use.
- **Stage 5 Destroy**: Data that is no longer needed or required to be kept is destroyed. Data must be authorized for destruction in this stage, and destruction methods must be legally compliant.

There are certain types of data that all businesses need to destroy or archive. For example, consider business tax records. In some cases, an organization is required to hold these documents for three or seven years, depending on the type of document. However, while the organization is required to store that data for this period of time, it doesn't need to be available for direct access, so we might archive that data. When the time is up for that documentation, we can destroy it and eliminate the cost of keeping it archived. There will be dedicated processes for performing these actions in the data governance plan. Data governance plans ultimately include planning, process, and policy for each stage of the data lifecycle.

When working as an analyst, you will most likely find that in regulated industries organizations have a dedicated team or department that handles their data governance policies, procedures, and measures. In nonregulated industries, organizations may have a more disjointed data governance plan, lack universal procedures, and silo governance within specific departments.

Roles Within a Data Governance Team

Everyone in an organization tends to participate in data governance in some way, whether or not they have a defined role within a data governance team. For example, database administrators and security professionals perform very specific tasks that ultimately ensure different elements of data are addressed and accounted for. There are also attorneys and compliance teams that work to ensure organizations follow the appropriate legal requirements for the data they work with. These functions could be performed by a single person, or entire teams can be dedicated to these roles. It's important to remember that how data governance is handled differs across every organization.

There are a few roles that are commonly utilized in data governance teams across organizations: data owner, data steward, and data custodian.

- A **data owner** is a senior (executive) role. The data owner holds the ultimate responsibility for maintaining the confidentiality, integrity, and availability of the information asset. The owner also typically selects a steward and custodian, directs their actions, sets the budget, and allocates resources for sufficient controls.
- A **data steward** is primarily responsible for data quality. The data steward ensures data is labeled, identified with the appropriate metadata, and collected and stored in a format that complies with applicable laws and regulations.
- A **data custodian** manages the system on which the data assets are stored. This includes the responsibilities of enforcing access control, encryption, and backup/recovery measures.

Where does the data analyst fit within these data governance roles? An analyst must adhere to the standards and requirements that align the organization's data governance plan. The analyst's role is to adhere to data governance policies but not necessarily to create them. As an analyst, you must keep data governance policies in mind when gaining access to and working with data.

There is no universal data governance plan for every organization; you will need to refer to the internal policies of each organization for plans, processes, and approaches to data governance.

Jurisdiction Requirements

Data is constantly being created, stored, and analyzed worldwide. Thus, as we work with data for different organizations, it is important that we know which jurisdictions we are accountable to (locally, nationally, and even internationally) and what regulations we must follow.

Data sovereignty is the idea that the country in which data is stored has control over that data. It describes the legal dynamics of the collection and usage of data in a global economy. Laws vary widely from country to country. Some impose restrictions on how data can be used, how it can be moved from one country to another, and what type of encryption can be used to protect it.

Legal Jurisdiction

Jurisdiction is the official power to make legal decisions and judgments. While data analysts are not responsible for jurisdiction, you should know that your organization, and especially its data governance and compliance teams, will monitor any applicable laws. Jurisdiction includes the following:

- Federal laws
- Federal regulations
- State laws
- International law
- Laws in other countries

Regulations and Compliance

Regulations are rules that are implemented by an authority and have the backing of law. Regulations are implemented for organizations that fall under an authority's control. These regulations describe their legal authority and include details regarding compliance requirements. An example is the Federal Energy Regulatory Commission (FERC), which regulates interstate transmission of electricity, natural gas, and oil. An organization that works in the energy industry is most likely regulated by FERC. If any regulations apply to you and your organization, based on the data that you work with, you will want to be aware of these requirements. You will likely discover that many organizations have a compliance officer and department to help navigate this.

There are a few major regulations that you should be familiar with.

- The General Data Protection Regulation (GDPR) enforces rules on organizations that work with the collection and/or analysis of data for people that reside in or offer services to entities in the European Union (EU). More information on the GDPR can be found here: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en.
- The Privacy Shield is a set of data protection requirements that applies to data transferred from the EU to the United States. Businesses who wish to work with international data can self-certify to prove that the protections they offer are adequate under the Privacy Shield framework. For more information on the Privacy Shield, visit: https://www.privacyshield.gov/US-Businesses.
- Children's Online Privacy Protection Act (COPPA) is a US federal law designed to protect the privacy of children (inside and outside of the Unites States) under the age of 13. Entities who operate online and collect data on children are required to follow the regulations set by COPPA. More information on this law can be found here: https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule.
- Family Educational Rights and Privacy Act (FERPA) is a US federal law that protects the privacy of student education records. More information on FERPA can be found here: https://www2.ed.gov/policy/gen/reg/ferpa/index.html.
- The Health Insurance Portability and Accountability Act (HIPAA) is a US federal law designed to protect the privacy of healthcare-related information, or personal health information (PHI). The website for HIPAA can be found here: https://www.hhs.gov/hipaa/index.html.

- The Payment Card Industry Data Security Standard (PCI DSS) is a global data protection standard established and maintained by a consortium of payment card companies, designed to reduce fraud by increasing the security of cardholder data. For more information, visit: https://www.pcisecuritystandards.org/document_library.

Data Classifications

Data classifications are a way to categorize or classify data. There are typically three levels of classification within an organization, although the specific levels used can vary from organization to organization. Three of the most common classifications are public, sensitive, and confidential.

- **Public**—Disclosure would not cause a negative impact to the organization.
- **Sensitive**—Disclosure would cause harm to the organization.
- **Confidential**—Disclosure would cause considerable harm to the organization.

Each level of data classification aligns with its own set of controls and procedures intended to protect the data. The data owner on a data governance team is responsible for applying these classifications to data, but as an analyst you must know the classification level of any data that you plan to work with.

Data Privacy and Protection

When you are working as an analyst, you will likely encounter data that identifies a person in some way. This data should be handled with care due to the numerous privacy and protection laws organizations must follow. Some of the most common types of personal data are described below.

- Personally identifiable information (PII) describes data that can be used to directly or indirectly identify an individual. It is important to note that PII exists in every organization in every industry, and it is included in other regulations along with the other types of personal data listed here.
- **Protected health information (PHI)** describes health-related data that can be used to identify an individual. PHI includes information about a person's past, present, or future health, as well as payments and data used in the operation of a healthcare business.
- **Personally identifiable financial information (PIFI)** describes information about a consumer provided to a financial institution. PIFI includes information such as account number, credit/debit card number, personal information (such as name and contact information), and social security number.
- **Intellectual property (IP)** describes intangible products of human thought and ingenuity. Intellectual property is protected by various laws, such as copyrights, patents, trademarks, and trade secrets.

Review Activity:

Data Governance

Answer the following questions:

1. **What are the three key elements of a strong data governance plan?**

2. **What are the five steps of the data lifecycle, in order from start to finish?**

3. **Which data governance role is primarily responsible for data quality?**

4. **How is data sovereignty defined?**

5. **Which regulation is a US federal law designed to protect the privacy of healthcare-related information?**

6. **There are typically how many levels of data classification in an organization?**

7. **What are three common classifications?**

Topic 16B

Understand Access Requirements and Policies

EXAM OBJECTIVES COVERED
5.1 Summarize important data governance concepts.

When sharing information, there are important agreements that will need to be in place to outline what data can be shared, who it can be shared with, and how it can be used. There are approval processes that grant the use of data and provide the permission to move forward. Data also has a life span, at the end of which it should be archived and sometimes destroyed. From use agreements to approvals, these are critical details for an analyst to be aware of as you start your data projects.

Data Use Agreements

A **data use agreement** is an agreement between two parties about the exchange of data that specifies what data will be shared and how that data can be used. Typically, a data use agreement is a legal agreement. Some variations include contracts, non-disclosure agreements, memorandums of understanding, and other legal instruments. Any document that addresses the use and exchange or sharing of information is a form of data use agreement.

Non-Disclosure Agreements

A **non-disclosure agreement (NDA)** defines the conditions under which an entity (such as a person or supplier) cannot disclose information to outside parties. An NDA includes specific descriptions of the legal ramifications for breaking the agreement, which provides a legal basis for protecting information assets and serves as a deterrent to the sharing of information.

Acceptable Use Agreements

An **acceptable use agreement** describes not only how data can be used, but also for what purpose. For example, suppose a company is giving a vendor access to its customer data for the purpose of researching the customers. However, the company would not want the vendor to use that data to create its own target customer list. In this case, an acceptable use agreement can be created to define the acceptable ways the vendor can use that data.

Acceptable use agreements also establish requirements for deidentification, or the removal of personal data, especially when privacy regulations like GDPR or HIPAA apply to the data. The goal is to reduce the risk that data can be reidentified.

Memorandums of Understanding

A **memorandum of understanding (MOU)** is an acceptable use agreement that establishes the rules of engagement between two parties and defines roles and expectations. They are used for describing and enforcing the data usage agreements between two parties. An MOU is a nonbinding agreement, meaning it is difficult to enforce in a court setting because it is not a formal contract, but it is still an agreement that lays out how the parties will adhere to the data policies surrounding the data that is to be used.

Release Approvals

When working with private information and data that has sensitive and confidential classifications, you will likely be involved in the process of release approval. Before an organization can release data to you for analysis, and before you can release it to others, these actions must be approved.

When personal and sensitive or confidential data is released without proper approval, there are serious consequences (even if the release was accidental). Always ere on the side of caution when preparing to work with personal data. When in doubt, reach out to your manager to inquire about the appropriate process at your organization.

Data Retention and Destruction Policies

When there are no outside regulations dictating what data an organization must retain or delete, the organization should consider the cost to store and maintain unneeded data when deciding how to handle it. But when there are regulations that require an organization to hold or destroy data within a certain time frame, the organization must follow jurisdiction laws and/or industry regulations. There are two key factors in the holding of data that are often specified in laws and regulations: data retention and data deletion/destruction.

Data retention defines the time span for which data must be kept. This includes not only the minimum amount of time data must be kept, but also the maximum (or “no longer than”) amount of time data can be held before it must be destroyed. Data retention requirements are often identified according to data classification levels, but the specific details should be defined in separate data retention policies and procedures. Effective data retention mitigates the potential issues surrounding data loss but also, and more frequently, for ongoing and future litigation. If you remove documents that you are required to hold, there can be ramifications to the organization through the legal system. In most organizations, you will find not only a compliance officer but also a well-versed attorney to help support the organization in adhering to the rules and regulations that must be followed.

Data destruction describes the legally compliant means through which data must be removed and made inaccessible. The required level of destruction is closely linked to the data’s classification. There are multiple steps needed to delete data in a way that ensures it is never recovered.

How long data is to be retained and how it is to be destroyed should be specified in an organization's data governance policy. As always, situational awareness of the data you are working with and any related internal policies is paramount.

Review Activity:

Access Requirements and Policies

Answer the following questions:

1. **What is an acceptable use agreement?**

2. **What are the two other most common types of data use agreements?**

3. **What process will typically occur when you need to work with private information and data that has sensitive and confidential classifications?**

4. **How do we refer to the time span for which data must be kept?**

5. **What do we call the legally compliant means through which data must be removed and made inaccessible?**

Topic 16C

Understand Security Requirements

EXAM OBJECTIVES COVERED
5.1 Summarize important data governance concepts.

In order to adhere to regulations and standards, organizations that collect and use data must also follow security protocols and policies as defined by the data governance plan. Each organization may have further security measures, processes, and software in place inside the information technology group responsible for security. The type of security measures in use depends on the type of data that is stored and collected and how it is classified. These security protocols, measures, and methods are applicable to the transmission of data and detail how to address data breaches and handle data access. From where we store data to how we send data, we should always have security of the data in mind.

Data Processing

As we discuss data governance and data security, it is important to note the types of data processing that can occur at an organization. These are choices that are not made by the analyst, but typically made by the IT and Executive teams who make decisions on the hardware and software to support the organization. Any or all of these may be occurring at any time in an organization. There are five high- level data processing types:

- Transaction processing
- Distributed processing
- Real-time processing
- Batch processing
- Multiprocessing

Transaction Processing

Transaction processing is used for transactional data that is mission critical to an organization. For example, an organization that sells large volumes online is likely using some form of transaction processing and leveraging systems that are meant for this type of data. This allows for transactional data to be captured and processed effectively in real time. The hardware and software components are designed so that data can move quickly and accurately. Consider an online retailer that participates in Cyber Monday sales; a system is needed that can handle a high volume of transactions placed from all over the country. With this type of processing, exact data—such as what was ordered and where it will be shipped—must be complete to continue with the transaction.

Distributed Processing

Distributed processing takes large-volume data sets and distributes them across multiple servers. A distributed data system, meant for these types of processes, is designed so that if one server fails, another server can take on the tasks and continue the data processes. In distributed processing, the servers are independent of each other and can be located in different geographical areas.

Real-time Processing

Real-time processing is like transaction processing, wherein the output is needed in real time, but it handles the data a little differently. Imagine that you are navigating to a destination using GPS navigation, and you miss your turn. The GPS navigation doesn't stop or turn off. It leverages real-time processing, acknowledging that you missed the original direction and carrying on by routing you to the next turn that will take you to your destination. Real-time processing is preferred over transaction processing when the answer is not expected to be exact, and approximate results are acceptable.

Batch Processing

Batch processing is used when processing a large amount of data. Imagine you are wanting to analyze 10 years of transactional history that shows every single invoice line item within that period. Accuracy is more important than processing speed, and lots of transactions will take time to process. Batch processing will process the data in batches, saving on the resources costs that are allocated for processing. Batch processing can also impact costs by leveraging the computational resources that are being used for the analysis of these large sets.

Multiprocessing

Multiprocessing involves the use of two or more processors working on a single data set. Unlike distributed processing, where the data is distributed across multiple servers in different places, multiprocessing occurs within the same server leveraging multiple processors. This allows faster processing for very large data sets (think millions of records), although it is a more expensive option, simply accounting for the machine expense and the amount of memory needed to process data.

Data Transmission

Data transmission is the process of transferring (sending or receiving) data. Data that is actively being transferred is **data in transit**. Data that is being stored is **data at rest**, and data that has been transmitted and is now present in memory or being queried is **data in use**, or data in processing. Data transmission can occur across a network, between computers, from a mobile device to a networked device, and more—even across a copper wire. Unfortunately, the transfer of data can be interrupted by malicious actors. Therefore, all data should be encrypted during or before transmission. Security professionals are responsible for ensuring that data is encrypted, although data analysts are often provided a means to encrypt data (typically in the form of a software that follows the necessary standards).

Data transmission is subject to oversight by the legal jurisdictions we mentioned earlier in this lesson because data is captured about all human, business, and natural activities on Earth. Thus, it makes sense that there is jurisdictional interest in data transmission, because these transmissions can affect commerce, financial markets, and societies.

Data Encryption

Data encryption is the process of using algorithms that will "scramble" data from its original plaintext into another form, known as **cyphertext**, so that it can't be read. This process ensures that sensitive data is not exposed while in transit. Encryption is reversible with the decryption key that unscrambles the cyphertext back to its unencrypted form, the original value. When data is encrypted, no human or machine can read it. It must be decrypted to be read.

Data can also be encrypted at rest, which means that the data is encrypted while it's not moving in transit.

Encryption Standards and Laws

The **Advanced Encryption Standard (AES)** specifies a Federal Information Processing Standards (FIPS)-approved cryptographic algorithm that can be used to protect electronic data. The AES is supported in many current software systems.

The legal standing of encryption varies widely across the world. Some countries do not constrain the use of encryption, whereas others impose very strict limitations. An excellent resource that highlights the legal stance of varying countries around the world can be accessed via: https://www.gp-digital.org/world-map-of-encryption/.

De-Identification and Masking of Data

As a data analyst, there will be times when you need to de-identify or mask certain elements of data before sharing or distributing it, particularly personal data. **De-identification** is the process of removing fields that can be used to identify an individual or information that must remain anonymous. As you learned earlier, masking involves hiding that type of field by showing something else in its place, like an asterisk, and is critical when working with PII.

For example, suppose we are sharing student data for a research project. We can share the number of children that attend a school within a year, how many students are in each grade level, and the number of children categorized by race or poverty status, but we cannot share any information that would allow someone to tie this data back to a single individual. Aggregating our data into counts is one form of de-identification.

Banding, or providing the aggregated data by group, is another technique used to de-identify data. For example, if you are working with data for children from the ages of five to 18, you might create "bands" of ages in specific ranges using a calculation or condition. This condition would provide a field that places every record within that age range within the band.

Many databases already have data masking systems in place. For example, you are likely familiar with seeing asterisks in place of characters in passwords to keep this information from being visibly exposed. There are several different ways to approach field masking for data, whether through aggregation, banding, or creating functions like an index field.

Before you share data with anyone outside your organization (or inside with a different level of access), always double-check that your data adheres to the privacy rules that are set forth.

Data Breaches

Even with a thriving data governance plan, protection policy, and security software that works to discover vulnerabilities, organizations can still experience data breaches. Again, as with all things related to data governance and security, you will need to review your company's internal organizational policies on what to do in the event of a breach.

Data Breach

A **data breach** occurs when information is read, modified, or deleted without authorization. "Read" in this sense can mean either seen by a person or transferred to a network or storage media. A data breach can involve the access of any type of data, although it is most commonly used in reference to corporate information and intellectual property. A privacy breach refers specifically to the loss or disclosure of personal and sensitive data.

There are varying levels of severity for data breaches. A breach can be as simple as an unauthorized figure logging into your unlocked machine while you are away or as complex as an unknown actor hacking a network to gain access to all the files within. It's important to note that breaches can occur at almost any level of an organization, and even in the most accidental of circumstances. For example, mistakenly leaving personally identifying information in a data set that is viewed by an outside company via a shared drive is a form of data breach.

A breach of any type can have severe consequences for an organization and the individuals involved—both the people at the organization who are responsible for the breach and also the individuals whose information was breached. Data breaches cause negative publicity for the affected organizations, ruining their reputation and causing people to lose trust in them. Organizations that have experienced a data breach typically have to pay heavy fines and can experience significant damage due to the loss of intellectual property. Further, a data breach can cause considerable problems for the people whose data was read in the breach, including but not limited to the misuse of personal information and identity theft.

Notification and Escalation of Breach

The way an organization is required to respond to a breach depends on the laws and/or regulations specific to the type and classification of the information that was breached.

The data governance and security teams at any given organization should be familiar with these requirements, which indicate who must be notified when a breach occurs.

When a data breach occurs, it is critical that the appropriate person on the team be notified so that they can take charge of the situation and formulate the appropriate response. Any breach of personal data and most breaches of intellectual property should be escalated to senior decision makers so they can properly consider any impacts from legislation and regulation. Even if a data breach is considered to be minor, if you attempt to correct it without alerting the proper person, you could place the company in legal jeopardy.

After a data breach is escalated, it may be found that public disclosure is needed. Although it is unlikely you will be responsible for the public notification of data breaches, it is important that you know the reporting requirements for your organization. Regulations outline who is to be notified and at what point. If public disclosures are not made when required, the organization will suffer fines and penalties.

Data Access

Permissions on data is the data governance concept that impacts data analysts the most on a practical level. There are several types of permissions on data that serve different functions.

Read and Read/Write Permissions

Two types of permissions you will commonly work with are read and read/write.

- **Read permissions** simply give you the ability to read the data.
- **Read/write permissions** give you the ability to read and also change the data.

When we need to enter and change information within a system, we need read/write permissions. When we are connecting to data to create a report, we only need read permissions.

When you find you don't have access to the data you need, you will need to use the appropriate channels to request permissions.

Role-Based Permissions

Role-based permissions are those that you gain simply because you serve a certain role at the company. Role-based permissions restrict or grant access to certain company information based on the role that person holds within an organization. Different employees are given different levels of access.

For example, let's consider a company's marketing department. Any employee in a management position has access to HR-related data for all department employees. However, nonmanagement employees in the marketing department only have access to their own HR data.

The ability to assign role-based permissions reduces administrative burden, as it eliminates the need to assign permissions one person at a time. It also makes managing permissions an easier task for a company overall.

User Group Permissions

There are times when an organization might also need **user group permissions**, or permissions that are specific to users in a group regardless of their roles. These permissions are not suitable to everyone in a particular role, and may even be needed by people who serve in different roles.

For example, suppose a company's marketing and product development departments share a process for contracting with vendors. If the company adopts a software that contains information specific to that process, only the users who need access to that system (in both departments) will be assigned to the user group with the associated permissions.

As a data analyst, you will certainly be assigned to role-based and user group permissions that are suitable for the tasks you will be given.

Saving Data Files and Storage Types

Most companies have policies that dictate where you should and should not save your work. We'll discuss a few of the ways information is typically stored.

Shared Drive

Shared drives are common in any server-based environment. These drives can be group or user based and have permissions associated with them that control access. For example, a drive that everyone who works for a particular company can access is a shared drive. These drives are likely backed up every night through server technologies and the IT department.

In most cases, you will work with your manager and IT department to create secure locations on shared drives where files can be saved.

! *When you need to save sensitive and confidential information, it is important to first identify who has access to the folders in which you plan to save. You do not want to be responsible for inadvertently causing a breach of information due to saving sensitive information in the wrong shared drive.*

Cloud Drives

Cloud drives are not unlike shared drives in that they allow access to a folder system that is shared in a server environment. One of the key differences between the two is where that server is stored. While shared drives use a traditional server sitting inside your company's IT office, the server for cloud drives resides on the "cloud." Files on a cloud drive are being consistently backed up and always available when needed.

One of the key benefits of a cloud drive over a traditional server is the presence of version history, which allows end users to see multiple versions of a document within a single document file. In a traditional shared drive, we must save files with different dates and titles, such as draft or final, and can only see those different versions in different documents.

Because a cloud drive is constantly updating and able to show multiple versions, you will want to be careful about who has access to the cloud folder you save your work to.

! *The data analyst always has greater access to information than most others, which means we have an additional responsibility to protect information.*

Local Storage

No matter what type of machine you're working on (whether laptop or desktop, Mac or PC), that machine has **local drives**. We access local drives similarly to how we access shared drives, but local drives are local to the specific equipment, whereas a shared drive is part of a larger server infrastructure. Most people recognize the most common local drive, the C: drive. People are also familiar with saving files into the Documents folder or their Desktop; these locations are actually a part of the local machine, in particular within the C: drive.

Unlike with shared drives, others who work at your company can't be given permission to access your C: drive. If you need to share files that are saved on your local drive, you must either re-save the files to a shared or cloud-based drive or transfer the files (such as via email).

It is important that you never send sensitive or confidential information via email. No matter where you save or upload your files, you must always be sensitive to the information you are sharing and who has access to any given location.

A downside of local drives is that they are difficult to back up through the company network, unlike shared drives and cloud drives, which make use of software and other technologies to protect your information in the event your machine crashes or otherwise becomes unusable. For example, a local drive can become corrupted, causing you to lose your data, or your laptop could be lost or stolen. This is why you will commonly find that organizations implement policies against using local drives and instead encourage the use of shared and cloud drives for the work performed at the organization.

Review Activity:

Security Requirements

Answer the following questions:

1. **What is data that is actively being transferred?**

2. **In what state should data be before transmission?**

3. **What is the approved cryptographic algorithm that can be used to protect electronic data?**

4. **What is the difference between de-identification and masking?**

5. **What might be needed if a data breach is escalated?**

6. **What are the two types of data access we commonly work with?**

7. **What are the three most common storage options for data?**

Topic 16D

Understand Entity Relationship Requirements

EXAM OBJECTIVES COVERED

5.1 Summarize important data governance concepts.

As you can imagine, an organization typically utilizes many different databases across an organization. As a part of the organization's overall data governance strategy, there is likely an ongoing effort to capture the model of these databases. In the same way that a data architect uses models when designing a database, the organization uses entity relationship models to identify relationships between the data held in the various databases. These models, when available, help the data analyst know what data is related to each other, which ultimately assists you in creating queries and joints.

Entity Relationship Models

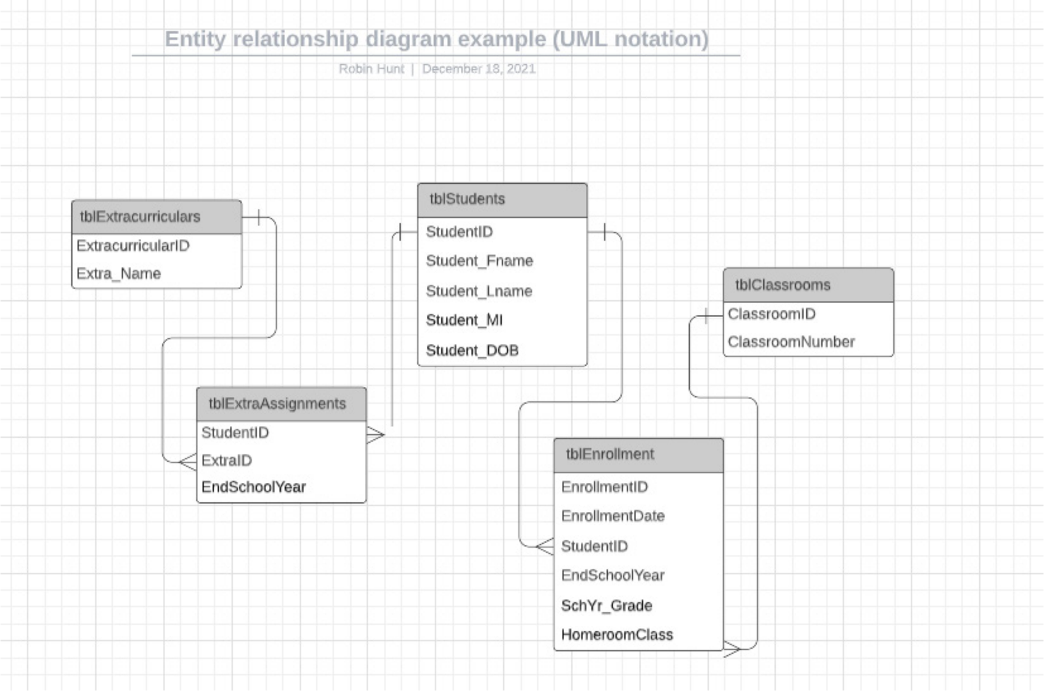
Entity relationship models give the data governance team a bird's eye view into the systems with data so they can more easily apply data classifications and provide valuable information to the teams that are required to protect and store this data. There are three basic model types: conceptual, logical, and physical.

- A **conceptual data model** is the conceptual view of what should exist in a data system and how it could be related.
- A **logical data model** is a more detailed view of the conceptual model that includes data fields and the relationships between them.
- A **physical data model** is the actual data system with tables, relationships, fields, and attributes.

Specific data governance teams are responsible for obtaining the physical data models, typically software, that the organization will use.

An **entity relationship diagram** is the pictorial representation of a database model that shows how entities (like people or objects) relate to each other through the data.

We can use Lucidchart, a diagramming software that allows us to create entity relationship diagrams, to look at an example. This is a basic diagram of The Digital High School Student Information System.



Entity Relationship Diagram (UML Notation) Created in Lucidchart (www.lucidchart.com)

The Unified Modeling Language (UML) is a general development and modeling language in software and database engineering. This would be a physical model, as it describes the exact layout of this database.

Record Linkage Restrictions

Record linkage, also known as data linkage, is the process of identifying, matching, and merging records that correspond to a matching record, whether between several data sets or within one data set. We link these records from each data set based on the identifiers they share. Let's imagine we have a student record for Sally Jones. Sally's record exists in multiple systems; there's information related to her enrollment, her test scores, and other student-related information. We use the process of data linkage to link all her records, when necessary, into a single data set that will then provide us with a full picture of her student information.

There are some instances in which our data systems cannot ever be related to each other due to organizational policies. These cases are known as **record link restrictions**, and they are set due to regulations intended to protect data sets with sensitive information. For example, suppose a health technology organization has a data system that handles customer notifications for new products and services that are designed to help customers stay on top of their health. The organization also has a system that contains PHI and medical records for each customer. To reduce the risk of anyone being able to associate the individuals in both systems, a record link restriction has likely been set to ensure these two systems are never able to communicate.

Data Constraints

Data integrity is the existence of accurate and consistent data in the database, and it can be maintained by data constraints. **Data constraints** are integrity rules that limit the types of data that can go into a column or table within a database system.

There are four types of data integrity rules.

- **Domain integrity** is the acceptable values for a field. For example, the appropriate data type is a data constraint. Suppose the data should be numbers and not text. If the data type is defined as numbers, the field will not allow text.
- **Entity integrity** is the unique identifier of a record as defined using a primary key field. For example, if the primary key field cannot allow a null value, this is a constraint that ensures all records can be uniquely identified.
- Referential integrity refers to the integrity of data between two tables, as we covered earlier in this course. A constraint would ensure the data follows the proper rules for adding, updating, and deleting records. For example, this type of integrity prevents the creation of a payroll record for a person who's not an employee. It uses the key (primary and/or foreign) to establish the relationship that requires the person to exist before being entered into the payroll system.
- **User-defined integrity** is based on business rules that are not covered by the other data integrity settings. These types of integrity use logic that goes beyond just a setting to constrain to a data type or format. For example, suppose an organization has products in their system that do not have a specified color because they are multicolor. When the data entry occurs, logic can be defined that when these products are entered, the field for color will default to multicolor. The data entry person doesn't have to attempt to guess what color option to apply, nor do they even have to spend time keying that information in. This ensures that we have exactly the right information expected in that field.

The goal of constraints overall is to provide high-quality data, which means it is correct and entered in accurately. While constraint sounds like a negative, in most cases it is a must for data integrity.

Review Activity:

Entity Relationship Requirements

Answer the following questions:

1. **What are the three basic types of entity relationship models?**

2. **What is the definition of an entity relationship diagram?**

3. **What is set to ensure two systems are never able to communicate?**

4. **What are data constraints?**

5. **Data constraints can occur at what level?**

Lesson 16

Summary

In this lesson, you learned that data governance is a large umbrella term for a framework used to govern data in an organization, and it covers the technology, people, and processes involved. You should now be able to summarize the various roles and jurisdictions that are part of the data governance process, and have gained a deeper understanding of the types of regulations that organizations must comply with. You should be familiar with the three common data classifications: public, sensitive, and confidential. You should be able to summarize the most common data use agreements and recognize that you must always be granted approval before the release of any data. You should be able to describe the need and use for data retention and destruction. You should have a better understanding of data security protocols, like data encryption, and be able to summarize the high levels of data processing. You should be able to describe what data masking and de-identification means and understand why it's especially important when working with PII. You should be familiar with the varying types of permissions on data: read, read/write, role based, and user group. You should also be able to differentiate between shared drives, cloud drives, and local drives. You should be able to explain the use of entity relationship diagrams, as well as the concepts of record linkage and data constraints.

Guidelines in Understanding the Importance of Data Governance

Consider these best practices and guidelines when familiarizing yourself with data governance plans.

1. A strong data governance plan has transparency, accountability, and standardization.
2. Data has a lifecycle. It's created, stored, used, archived, and deleted.
3. Regulations apply to different organizations depending on the type or organization and what they produce.
4. The disclosure of public data would not affect the organization, but the disclosure of sensitive data would cause harm, and the disclosure of confidential data would cause considerable harm.
5. A data use agreement is an agreement between two parties about the exchange of data that specifies what data will be shared and how that data can be used.
6. Data retention requirements and the required level of destruction is closely linked to its classification.
7. Encryption ensures that sensitive data is not exposed while in transit.
8. De-identification is the process of removing fields that can be used to identify an individual or information that must remain anonymous, while masking is showing something else in its place.

9. A data breach occurs when information is read, modified, or deleted without authorization, and it can have severe consequences for all parties involved.

10. Local drives are often discouraged, as they require backup in case your machine becomes unusable and do not allow you to give others permission to view your work.

11. An entity relationship diagram is the pictorial representation of a database model that shows how "entities" (like people or objects) relate to each other through the data.

12. Some data systems cannot ever be related to each other; this is a case of record link restrictions.

13. Data constraints are limitations on what can be done with the data, and they can occur at a system-to-system level, record level, or field level.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 17

Applying Quality Control to Data

LESSON INTRODUCTION

Have you ever heard the phrase “garbage in, garbage out”? In the world of data analysis, the use of proper data validation methods can turn some of that “garbage in” into “one person’s trash is another’s person’s treasure.” Quality data leads to quality reporting. A report would be problematic if the data was inaccurate, incomplete, or inconsistent. In this lesson, we will describe the quality assurance process and explain the reasons we check for data quality. We will learn how to understand data quality metrics and walk through different methods to verify and validate the data that we provide for reporting.

Lesson Objectives

In this lesson, you will do the following:

- Describe characteristics, rules, and metrics of data quality.
- Identify reasons to quality check data and methods of validation.

Topic 17A

Describe Characteristics, Rules, and Metrics of Data Quality

EXAM OBJECTIVES COVERED

5.2 Given a scenario, apply data quality control concepts.

As an analyst, you will most certainly encounter data that doesn't have the level of quality that gives you comfort in presenting results. This is an unfortunate fact of life for a data analyst. In this lesson, we will discover quality assurance, quality data, and methods to validate and verify the information so that all can be confident in the results.

Reasons to Check Data Quality

There are times where data is just bad. It could be bad due to human error, a bad database design, or just something that went awry in the process of getting the data from the system to the point of analysis. All data that you work with on a project needs validation and verification throughout the entire project.

- **Data validation** is the process of confirming the type, structure, and accurate representation of the data. An email address is a great example to use for validation: an email address always has an "@" symbol and usually ends with a ".com," ".net," or another variation, all of which can be easily validated with a few logical tests to confirm that the data is valid. If the email address is in the wrong format, then that data is not valid.

- **Data verification** is the process of confirming that the data is accurate or true. For example, suppose we have a value in the product price field that is listed at $50, but all of the other products are listed at $10. The fact that this product is five times higher than the others signals that we should verify the $50 product price is correct, because human error could've occurred when it was entered into the system.

Imagine that spreadsheets of data are extracts from various systems that need to be consolidated and transferred into a single data set. The spreadsheets are maintained by various people and departments, with different standards on how the data is entered and maintained. In this scenario, the analyst should check for data completeness, accuracy, and consistency to ensure the quality of the acquired data. Data quality checks are critical when data is acquired from other sources.

In the business setting, several situations necessitate the need for data verification and validation checks:

- **Mergers and acquisitions**—When companies merge, or one company acquires another, data from multiple organizations is combined. The involved organizations may have some of the same systems, but their data will eventually all need to be migrated to the same organizational data system. It's important to note that when two large organizations merge, they might have different standards developed for their data prior to the merge of the data sets. You will want to ensure that the data meets the correct structure and type (validation) and is correct (verification).

- **Data manipulation**—When data is manually keyed into a system, or manually changed by hand, there is the opportunity for human error. This data needs verification and possibly also validation, depending on the fields used. For example, you may need to confirm that records for email addresses and postal codes meet the requirements for this data (validation).
- **Data transformations**—Any transformations that are performed should always include a process for quality checks to ensure the data transformations worked as intended. As an example, in Power Query, the replace text command is case sensitive. If the data analyst doesn't write the transformation to include the correct case, then even if the text is the same, the program will not know to transform it.
- **Human error**—Any situation in which human error may have occurred necessitates quality checks for a multitude of reasons. People may key in the wrong information or build an invalid calculation using incorrect numbers. They could also have joined data improperly—a mistake that is difficult to detect in a large data set without diligence. These errors would be difficult to catch without checking for data quality.
- **Data transfers**—The transfer of data between systems is another common reason for quality checks. The data analyst will want to ensure that data transfers correctly from one system into the other system. When data flows from one system to the next, it's important to verify that the data actually made it to the intended destination. For example, suppose you are using Microsoft Power Automate to flow data from a SharePoint list into the database. To create these flows, different conditions are set that determine when to transfer the data. If a condition is not set properly, or if someone breaks the rules of a condition, then the information may not flow through as intended, leading to missing information that should be in the system. The opposite could also occur, where information that should not be in the system actually did flow through because a condition was not properly specified.

Understanding Quality

Quality assurance (QA) is the process of ensuring that the data used in analysis is of a high enough quality that it gives decision makers confidence in the findings. QA is often referred to as "data cleaning." As a data analyst, you will develop a data quality plan when approaching a data project that includes how to handle impossible and missing values, validation, and verification. **Quality data** can be defined as data that has integrity and is accurate, complete, and consistent.

- **Data completeness** means that the data is complete and that all expected and required fields are entered. For example, if the data set contains first name, last name, and email addresses, does every record have each one of these fields completed? If not, this issue should be further investigated to determine if a quality source of this missing information exists and can be used to correct the missing data.
- **Data accuracy** means that the data in the field is correct and accurate. Imagine that we combined two data sets of products for a company. Although the products are the same, the product names may not exactly match because they were entered in differently. Recall from an earlier lesson our example of an organization renaming the Red Valley Bike to Vibrant Valley Bike. For the most accuracy, every occurrence of Red Valley Bike should be updated to reflect the

correct name. There may be a few reports that would be an exception, where the original name should still remain, but overall the existence of Red Valley Bike could indicate that not all systems were updated with the new bike name. If there are no quality checks, we can get bad groupings on our data (i.e., two distinct products with sales for each, even though they are the same product just renamed) or just invalid data altogether.

- **Data consistency** means that the data is consistently entered the same way and as intended based on business rules. As an example, the staffing data of a healthcare facility contains the credentials of nurses and doctors. These credentials could be entered inconsistently (e.g., MD vs. M.D.; PhD vs. PHD; Lpn vs. LPN). If they have not been entered consistently, the data analyst should apply data cleaning and transformation techniques to correct them and make them consistent with the required standard for that organization.

These are simple examples, but rest assured that there are many varying opportunities for the data analyst to perform quality assurance. Human beings completing forms may not be thinking about the impacts to the data and reporting; they are most likely focused on getting to the next field, or completing the entry of the information so they can move on to whatever they have coming next in their day.

It is important to understand that data can have multiple data quality issues. It can be inaccurate, inconsistent, and incomplete all at once.

Rules and Metrics for Data Quality

If a data governance team exists at your organization, they will likely determine the list of rules and metrics for quality standards. These rules and metrics serve as key performance indicators (KPI) to evaluate data quality and monitor progress. Regardless of the existence of data governance and quality standards, as a data analyst you will want to always ensure your data is high quality. You want to report with confidence, and the people you report to will look to you for that confidence in the data.

As a data analyst, you don't always have control over the systems that provide data. Thus, achieving a 100% quality rating for any data requires internal support and measures within the organization. You can identify metrics that can be used to set measurable goals for data quality based on the needs of an organization. For example, suppose an organization wants to achieve a high level of data quality for sales contacts established for accounts, as well as their emails, a critical means of communication. The company sets a 95% quality rating as a metric for these fields. Other crucial data captured about sales contacts may include physical location, but this field is of lesser importance, so a quality rating of 80% is set as the goal. These fields are available to be analyzed, and you can count them and provide percentages.

Creating rules about the data is also a key part of ensuring quality; this can be internal to the system design or through a process. Continuing with our example of data involving sales contacts, suppose the organization establishes a business rule that the names of sales contacts must contain a first name and last name, must be complete (not left blank), and must contain a single space between them. A quality check can be performed to test that these rules have been followed by verifying that the first and last name fields contain a value (e.g., are not null) and that they contain a single space when combined. Both tests would fail if data was missing.

Emails could also be tested for null information and to confirm that they contain a "." followed by an appropriate value for an email address (e.g., .com or .net). For physical location, data can be tested for completeness and whether it conforms to rules such as two-letter abbreviations for states. Failure to meet these quality standards triggers a process to complete or correct these fields, with the level of effort directly related to the metrics (e.g., quality rating percentages) set.

In addition to assessing the quality of the data itself, evaluating data processes is also important for an organization. These processes may be incomplete or not followed, leading to quality issues. As a data analyst, consider asking two questions about data that does not pass quality checks:

1. Was the entry bad?
2. Did the data itself not conform to the rules?

As an example, a person may type out the entire state name for the location of sales contacts rather than use the state's abbreviation if the data system does not have constraints to restrict the entry. Alternatively, the person may not have followed the proper procedures for entering the state's name into the system.

Conformity and nonconformity to the required rules and procedures impact the quality of our data. As an analyst, you may not be able to control these procedures, but you can support the process. For example, your job role may involve letting your managers know how many rows passed through all the quality checks and how many rows failed to meet the standard.

Review Activity:

Characteristics, Rules, and Metrics of Data Quality

Answer the following questions:

1. **What is the difference between data validation and data verification?**

2. **What are a few reasons why we might check the quality of data?**

3. **What do we call data that is accurate, complete, and consistent?**

4. **A system with a state field that always lists states as an abbreviated value and in proper case has what?**

5. **What are used to set measurable goals for data quality based on the needs of an organization?**

Topic 17B

Identify Reasons to Quality Check Data and Methods of Data Validation

EXAM OBJECTIVES COVERED

5.2 Given a scenario, apply data quality control concepts.

Data validation and verification are important parts of the quality assurance process that help us ensure data meets quality standards, meets the business requirements, and can be trusted for decision-making. Whether you use tools to support the validation of data or manual methods, the data quality (or lack thereof) will impact all data analysis. Data analysts should not only validate but also verify the quality of data that they work with, because whenever data is manipulated or manually entered there is the potential for human error. Data analysts can benefit from developing a checklist for assessing data quality. Tasks such as data profiling, data audits, checking calculations, and checking visuals are key components of quality control that the data analyst regularly performs.

Data Validation Methods

Validating data is important throughout all stages of the process or project. When performing quality checks on all the data and visuals you produce and work with as an analyst, a checklist can help you keep track of what's been done (and what still needs to be done). Here, we cover some of the tasks commonly performed while validating data.

Data Profiling

Data profiling (introduced in an earlier lesson) prior to working with a data set provides reasonable expectations about the data and valuable information for checking your work. Some critical information to know includes total record counts and the total amounts of all summed data. For example, after data profiling, if you expect five million dollars in revenue, then a result of 10 million should immediately indicate a problem, either with your math or in the joining of the data with other sets. Knowing how many records you have when loading them from one system to another is another simple yet imperative method for validating your data.

Cross Validation

In research where the findings can have a profound impact on a process or program, there are several methods you might use to conduct **cross validation**. As a data analyst, you will want to look at the data collection methods. Was this data collected from a survey and/or a live interviewer via the phone? Did the questions in either method contain any bias, and were the questions consistent across both collection methods? Validating the collection of the data is one way to ensure you have quality data from both sets. It can be as simple as confirming that what you see in all systems regarding that data and information is the same.

Peer Review

Peer review can be critical for data projects. This is a practical way to gain feedback on the data product before it is published or goes live. Peer review can be as simple as having someone walk through your presentation and provide suggestions for improvement or as complex as running a deep review to test if your results are repeatable and accurate.

Data Audits

When an organization has a strong data governance plan in place, you may find they leverage data audits. A **data audit** is the process of not only assessing data quality, but also whether the data can achieve a specific purpose or objective. For example, if the goal is to report on invoices that have been generated over the last five years, the data audit would include confirming that you have five years of data and that this data meets the requirements that have been set for the report. Although data audits are a great way to ensure quality, this is not a method you will find in every organization.

Automated Validation

Automated validation is when we utilize the power of software to ensure we achieve a validated result. Let's return to our previous example of email addresses to see how automated validation works. When a person attempts to input a bad email address into a field, automated validation can prevent that email address from ever being saved if it doesn't meet the standards of the email field. Automated validation can also ensure that phone numbers or dates are entered in the right format. By ensuring that the data in these fields was entered correctly from the very beginning, this will reduce the amount of verification you must conduct on these fields at a later time. In essence, if you know that there's no possibility a date of birth might use an invalid date format, thanks to automated validation, you will have more confidence that you are working with validated field data. It is important to note that automated validation doesn't entirely prevent verification issues. For example, someone could key in a birth date in the correct format, but with a bad date, like 1/1/1900 instead of 1/1/1990.

There are software tools dedicated to helping organizations validate their data. These tools use processes that look at groups of records and confirm the completeness by creating a master record. (You'll learn more about master data management in the next lesson.) Data validation software tools can be set up to provide automated reporting about data processes. For example, reports can be generated to confirm what data was transferred from the source system to the data warehouse. Data validation tools will let you know how many records passed through successfully and also how many failed to go through. These reports will also often provide reasons why records may have failed to transfer that can give meaningful insight into how to solve the failure. Can you imagine the process of loading millions of records without automated validation? It would be very time consuming and prone to errors.

Let's walk through an example of how data validation tools help ensure data quality. To support an older company that has recently adopted a new customer relationship management (CRM) software, you are tasked with ensuring that all the historical invoices are loaded into the CRM. The system requires data on all accounts, contacts, products, and invoices. You will not manually enter these records, but will use an API that is built to work with the CRM. You create various exports of all the necessary data sets and load them into the system. Because the older CRM system did not have any controls on the account names, you anticipate some duplicated accounts. The automated validation in the new CRM system provides a list of potential duplicates and allows you to merge these fields with no issues, giving a higher quality data set. Automated validation also would allow you to check for data loss. When you load information into the new CRM system, if you expected to see an account and failed to find it, the system could provide a list of the invoices that could not load because of missing values.

Data Verification Methods

As you now know, verification is the process of confirming that data is accurate. The lines between verification and validation can be blurred, and in some cases both are in order for the data you might be working with. There are some verification steps you should perform on all data sets.

Verifying Field Level Data

Field level data can be incorrect for many reasons. It sounds simple, but verifying field data can become complicated very quickly when you have a large amount of data. Looking for data that seems to be outside the norm, or having a discussion with the team that is responsible for that data, can add clarity and help you verify the data. When field data is incorrect, we can attempt to provide a verified value if needed. This process is strongly dependent on whether you have information that can be used to verify the data and thus provide the correct information.

Verifying Record Counts

Data loss is the intentional or accidental loss of information through human error or an ineffective process. Data loss occurs when records are lost, are incomplete, are poorly named, or may have accidentally dropped out.

A discrepancy between data types in the source system and the system where data is transferred can also result in missing data. If the date field in the source system is a text data type, but the system where the data is headed contains an actual date field, data loss will occur if you do not convert the date field to a recognizable format—or all the records might fail when you attempt to load them.

When you know your data, the people in your organization, and the quality standards for data, you can build quality check queries that check for null values or information that is missing.

Verifying Calculations

The mathematical equations within the programs we use as analysts are nothing short of amazing. They perform complicated computations that require little more from us than correctly inputting the right values into a formula. However, we can introduce errors by making mathematical mistakes or using the wrong equations. It is always important to verify calculations, such as spot-checking for the intended outcome and that the correct information was supplied to formulas and equations.

Verifying Visuals

You must always confirm that visuals are labeled correctly and meet business requirements, and that data is sorted effectively for the visual. You should also confirm that your visuals make sense to users. As an example, suppose your visual is titled "Top 5 Products." You should make sure that this visual does indeed show the top five products, and not the bottom five. Also, you should always look for spelling errors in your visual titles, labels, and text. You will also want to confirm that colors are consistent across visuals, meaning if you used purple to refer to "graduate degree," you must ensure that purple references graduate degree throughout the project.

Review Activity:

Reasons to Quality Check Data and Methods of Data Validation

Answer the following questions:

1. **What process involves validating that the data can achieve the specific purpose of an objective?**

2. **What are at least two examples of how automated validation helps us achieve a validated result?**

3. **If data fails to transfer from the source system to another, what would we say has occurred?**

Lesson 17

Summary

After this lesson, you should understand that quality data leads to quality reporting. As a data analyst, you will spend large amounts of time validating and verifying the data that you work with throughout a project, and should be familiar with the reasons why we perform these actions. You should understand how data completeness, data accuracy, and data consistency contribute to the quality of data. You should be familiar with common data validation methods, such as data profiling, peer review, and cross validation. You should also be familiar with the steps taken to verify data, including verifying field level data, record counts, calculations, and visuals.

Guidelines in Applying Quality Control to Data

Consider these best practices and guidelines when familiarizing yourself with what makes quality data and how we can validate and verify data.

1. Data validation is the process of validating that the data meets standards, made it to the systems in which it belongs, and overall meets the business requirements for what it's intended to accomplish.

2. Data verification is the process of confirming the type, structure, and accurate representation of the data.

3. Quality data can be defined as data that has integrity, or data that is accurate, complete, and consistent.

4. Data quality checks are a routine part of the data analyst's job, and should be performed on all data.

5. When data is manipulated by individuals, there is always an opportunity for human error.

6. Any data transformations that are performed should include a process for quality checks, to ensure the data transformations worked as intended.

7. The data governance team of your organization will determine the list of rules and metrics for quality standards. These rules and metrics serve as KPIs to evaluate data quality and monitor progress.

8. Data profiling prior to working with a data set provides reasonable expectations about the data and valuable information for checking your work.

9. Some organizations have a data audit process in place to ensure data achieves its purpose, but regardless of whether a separate data audit is implemented, you must still always check for quality.

10. The use of software to ensure we achieve a validated result allows us to load millions of records quickly and accurately, something that if done manually would be extremely time consuming and prone to errors.

11. It is important to verify records because data loss occurs when records are lost, are incomplete, are poorly named, or may have accidentally dropped out.

12. It is always important to double-check calculations, such as spot-checking the intended outcome and that the correct information was supplied to formulas and equations.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Lesson 18

Explaining Master Data Management Concepts

LESSON INTRODUCTION

Master data management is key to data quality, as it enables the existence of a single source of truth at an organization. Master data management at large organizations with compliance regulations is traditionally supported through data governance and the use of dedicated software.

Lesson Objectives

In this lesson, you will do the following:

- Explain the basics of master data management.
- Describe master data management processes.

Topic 18A

Explain the Basics of Master Data Management

EXAM OBJECTIVES COVERED

5.3 Explain master data management (MDM) concepts.

Regardless of the size of an organization, there are multiple reasons for master data management. Data quality and integrity are something that all organizations want for their data and their reporting. New systems and legacy systems support the organization with the same type of data, and having a known source of master data can ensure that we are working with accurate data.

Master Data Management

Master data is important data about general business operations that is used frequently in analysis. It is shared by people across the organization to make critical decisions. Master data changes less frequently and is intended for dimension data, not transactional data. Although master data may vary by organization, here are some common types:

- Customers
- Products
- Suppliers
- Accounts
- Employees
- Locations
- Assets
- Reference data (countries, states, status types)

Master data is likely a part of an organization's data governance initiative. **Master data management** includes tools and processes that are used to create the single source of truth, or the "golden record" for the data that is considered critical at the organization. Master data management is really the management of the "master file." Whether or not a company has adopted a large-scale software, like Informatica or Relatio, for master data management, they likely have a source that everyone uses as the "master" source of information.

Having a single source of truth is important for all general business operations and for effectively analyzing and communicating information. Consider something as simple as project names. Suppose marketing, HR, and sales all use different naming conventions for a project. When the analyst creates a report that combines information from all three departments, the same project could have three different names. Master data management would aim to solve this issue by creating a golden record of project names, helping to avoid confusion about which project is being referenced and ensuring consistent data.

Organizations that have implemented master data management will have policies to help maintain master data so that it is accurate and up to date. Consider the data that an organization may have regarding the sales pipeline. Suppose that the company maintains a list of prospective customers that the sales team hopes to convert to paying customers. The policies and procedures of master data management will define when a customer in the sales pipeline should be added to the master data. These policies also outline how any changes to the master data should be handled. For example, if a customer's address changes, master data management policies help ensure that the master data is updated with the new address so that the entire organization has access to the correct address for all forms and documents related to that customer.

! *Organizations vary in their master data management efforts. Some organizations may not have any initiatives in place related to master data. Other companies will have policies and procedures related to master data management, but the process is handled manually without the the use of tools and software. Large-scale organizations are more likely to use some form of software for master data management, because there is just too much data to effectively manage using manual approaches.*

Benefits of Master Data Management

One of the biggest benefits of master data management is that it improves data quality and data integrity. Master data management and data governance initiatives help ensure that data is managed by a set of rules and procedures. Having a single source of truth and a process for managing data helps safeguard against inconsistencies or inaccuracies in the data. For example, if master data management is in place, the data analyst can be more confident that a customer name in the system in fact represents an actual customer, because it has met the rules of the master data management policy. Master data management ensures everyone in the organization is working with data that is accurate and vetted.

When master data management is a part of the organization's operations, it supports all organizational functions by streamlining access to data and shortening the time to data-driven actions. The master data management process allows people to access high-quality data that is most likely in a suitable format for reporting and visualization. In addition, when new systems are introduced to an organization, they can automatically be populated with high-quality lists of master data.

Reasons for Master Data Management

Master data management is part of an organization's data governance initiatives. As a refresher, data governance lays out the rules of handling data and how entities should ensure those rules are followed. Data governance includes policies and procedures for how data is managed and maintained, including procedures to ensure data security. Compliance with these policies and regulations requires rules that are applied at the macro level (e.g., internet communications, interstate and international digital financial transactions) and at the micro level (e.g., within an organization, between organizations).

Data governance initiatives encompass the people of an organization and the technology needed to support their efforts. Successful data governance relies on documenting data handling policies and evolving these policies as needed according to influences from technology or regulations. Systems that involve people and technology are in place to ensure the policies are understood and executed to maintain data and data security.

Data governance and master data management combined can help organizations in a number of circumstances. Master data management helps support the transfer of data from older systems (often referred to as legacy systems) to newer, more advanced systems. Strong master data management is also beneficial during mergers and acquisitions. When a business merges with or acquires another company, the master data of each organization will also need to be merged. Although businesses in the same industry (e.g., finance) may use some common terms and practices, each company will have a unique approach for handling and referencing their data. Just a few possible differences include the term *customers* versus *clients* or using two-letter abbreviations for states versus spelling the whole state name out in a field. Master data management plays a key role in addressing these discrepancies and integrating data during mergers and acquisitions.

Master data management also supports compliance with regulatory mandates like HIPAA, GDPR, and SOX by helping to identify the personal data that is collected and letting us know where this data is being used, which is critical for maintaining proper security of PII. Companies that are required to house personal data must be able to adequately identify that all personal data is maintained and secured. In some cases, they must also be able to destroy personal data upon request. Master data management supports the effort to maintain this data effectively, and on a large scale.

Organizations vary in their master data management efforts. Some organizations may not have any initiatives in place related to master data. Other companies will have policies and procedures related to master data management, but the process is handled manually without the use of tools and software. Large-scale organizations are more likely to use some form of software for master data management, because there is just too much data to effectively manage using manual approaches.

Master Data Management vs. Data Warehouse

Master data management software may seem similar in nature to a data warehouse, because both are meant to give us a single source of truth. However, although they are meant to work together, they are in fact two different tools.

Master data management software and the data warehouse both hold data that can be used by the entire organization. They both work with multiple sources of the data and increase data quality. However, their goals are entirely different. Master data management establishes a master record of data related to business operations. Master data maintains dimension data, whereas a data warehouse will hold the dimension data as controlled by the master data and the transactional data from the source systems.

In simpler terms, master data management software will work with the data warehouse to ensure the data in the data warehouse is updated with the master data. The software may pull data from the data warehouse as a part of master data management, and then send the master data back to the warehouse. Master data management also helps create more accurate data in the data warehouse by ensuring that consolidated data is a more complete and accurate record.

Review Activity:

The Basics of Master Data Management

Answer the following questions:

1. **Master data management focuses on what type of data?**

2. **What does master data management aim to create?**

3. **What are some benefits of and reasons why organizations have master data management?**

Topic 18B

Describe Master Data Management Processes

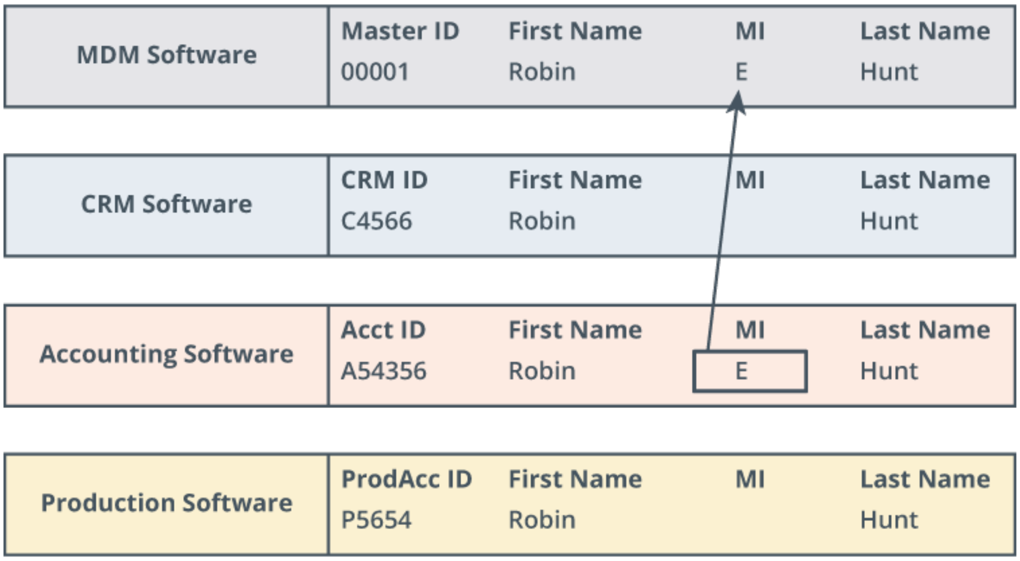
EXAM OBJECTIVES COVERED

5.3 Explain master data management (MDM) concepts.

The processes involved in master data management are dictated by the data governance team and often managed using the required settings in the software that is selected for master data management (if in use). There are some common scenarios to expect when trying to achieve the "golden record" for any data set. Consolidation of data and standardization of fields are two key processes to help create a single source of truth. Data dictionaries are another important tool related to master data management, serving as useful repositories that capture data within the organization.

Consolidation of Multiple Data Fields

Master data management requires the consolidation of multiple fields of information. This is one way to establish a single, reliable source of truth. Master data management software can help enable the consolidation of data from multiple systems to create one record of trust. Consider customer data from accounting, CRM, and production software systems, as shown in the next screenshot. The master data management software will combine the data from all three systems to create a single record.



MDM Software	Master ID	First Name	MI	Last Name
	00001	Robin	E	Hunt

CRM Software	CRM ID	First Name	MI	Last Name
	C4566	Robin		Hunt

Accounting Software	Acct ID	First Name	MI	Last Name
	A54356	Robin	E	Hunt

Production Software	ProdAcc ID	First Name	MI	Last Name
	P5654	Robin		Hunt

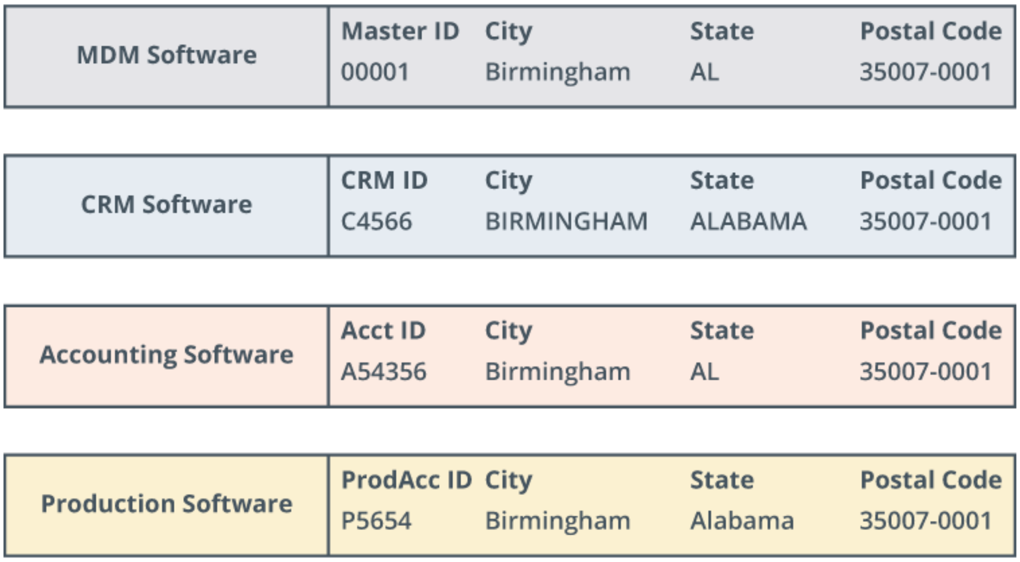
Consolidation of the Records in Three Systems to Create a Single Master Record

The master data management software uses the data from all three records—the CRM software, accounting software, and production software—to form a master record that represents the data for this one customer.

Field Standardization

Standardizing fields of data can be advantageous for many reasons, particularly when working with different systems that house the same information, or in the case of an acquisition or merger. When you report on data that has not been standardized, a fair amount of transformation and cleaning will be dedicated to that process. Field standardization can be a very efficient and effective way to reduce these efforts.

As shown in the screenshot, the standardization process involves making changes to the data, such as transforming all data for states to two-letter, uppercase abbreviations and using a consistent format for postal codes.



MDM Software	Master ID	City	State	Postal Code
	00001	Birmingham	AL	35007-0001

CRM Software	CRM ID	City	State	Postal Code
	C4566	BIRMINGHAM	ALABAMA	35007-0001

Accounting Software	Acct ID	City	State	Postal Code
	A54356	Birmingham	AL	35007-0001

Production Software	ProdAcc ID	City	State	Postal Code
	P5654	Birmingham	Alabama	35007-0001

Standardizing Fields of Data Using Master Data Management

The city, state, and postal code fields are formatted differently in these three different software systems. The master data management software standardizes these fields.

Data Dictionary

The master data dictionary is a critical part of master data management and a valuable resource for the data analyst. This document can serve as the authority on all definitions that have been agreed upon for the organization, as well as key metrics. In general, any type of **data dictionary** is a valuable resource document for the data analyst. The data dictionary may contain all the information related to an organization's data, or it can be specific to a single database or software. It will likely contain data elements and their definitions and field attributes, along with the relationships and structure of the data.

The data dictionary is a living tool—as definitions change or new data systems are added to the organization, they become a part of the data dictionary. When no data dictionary exists, the analyst will need to manually identify this information to verify that their work is correct. Systems that are large and complex can be deceiving to a data analyst when they are discovering what's available to them through the back end of a system. Having a complete data dictionary helps to increase their ability to find the appropriate fields and tables and to understand the relationships needed to build their required data sets.

Data dictionaries can be created by hand or generated by tools that work with databases. The simplest form of a data dictionary can still be a powerful asset in the hands of the data analyst.

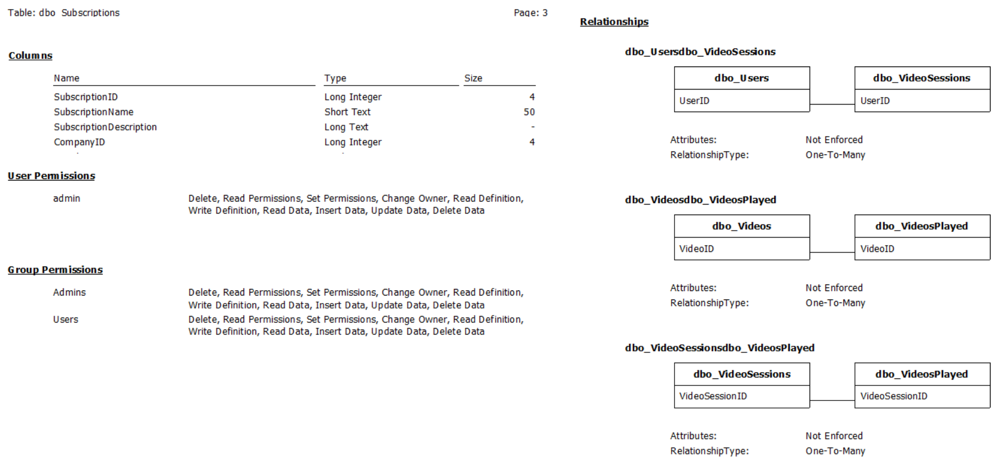


Table: dbo Subscriptions

Page: 3

Columns

Name	Type	Size
SubscriptionID	Long Integer	4
SubscriptionName	Short Text	50
SubscriptionDescription	Long Text	-
CompanyID	Long Integer	4

User Permissions

admin	Delete, Read Permissions, Set Permissions, Change Owner, Read Definition, Write Definition, Read Data, Insert Data, Update Data, Delete Data

Group Permissions

Admins	Delete, Read Permissions, Set Permissions, Change Owner, Read Definition, Write Definition, Read Data, Insert Data, Update Data, Delete Data
Users	Delete, Read Permissions, Set Permissions, Change Owner, Read Definition, Write Definition, Read Data, Insert Data, Update Data, Delete Data

Relationships

dbo_Usersdbo_VideoSessions

dbo_Users	dbo_VideoSessions
UserID	UserID

Attributes:	Not Enforced
RelationshipType:	One-To-Many

dbo_Videosdbo_VideosPlayed

dbo_Videos	dbo_VideosPlayed
VideoID	VideoID

Attributes:	Not Enforced
RelationshipType:	One-To-Many

dbo_VideoSessionsdbo_VideosPlayed

dbo_VideoSessions	dbo_VideosPlayed
VideoSessionID	VideoSessionID

Attributes:	Not Enforced
RelationshipType:	One-To-Many

Microsoft Access Documenter Tool (Used with permission from Microsoft.)

Even tools intended for lightweight databases, like Microsoft Access, often have the ability to generate basic data dictionary-type information. This screenshot provides information on the columns of a particular table, shows a layout of the tables, and defines relationships. This is information that you might find in a data dictionary.

Review Activity:

Master Data Management Processes

Answer the following questions:

1. **In master data management, what process is used to match the records in different systems to create one record of trust?**

2. **What two processes will likely need to be performed on data for a report that is not standardized through master data management?**

3. **What resource contains all the master definitions of data within the organization?**

Lesson 18

Summary

After this lesson, you will have an understanding of why master data management is key to data quality and how it provides a single source of truth at an organization. You will be able to describe how master data management at large organizations is supported through data governance and software. You should be able to describe why the consolidation and standardization of data can be advantageous when working with different systems that house the same information or in the case of an acquisition or merger. You should also understand that a data dictionary is a valuable resource document that contains all the definitions and information related to an organization's data.

Guidelines in Understanding Master Data Management

Consider these best practices and guidelines when familiarizing yourself with the basics and processes of master data management.

1. Master data changes infrequently and is dimension data, not transactional.
2. Master data management uses software to create a single source of truth in the form of a complete record for data that is critical to an organization.
3. Master data management is a part of the overall data governance initiatives at an organization.
4. Master data management software works with the data warehouse to provide data that is complete and accurate.
5. There will be varying levels of master data management at an organization based on the compliance and regulation requirements.
6. Master data management involves consolidating data to create a truly complete record from multiple sources.
7. Field standardization can be a very efficient and effective way to reduce transformation and cleaning efforts.
8. The master data dictionary is an authoritative document that defines all master data and key metrics at an organization.
9. The data dictionary is an essential resource for the data analyst and typically contains data elements and their definitions and field attributes, along with the relationships and structure of the data.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Appendix A

Identifying Common Data Analytics Tools

EXAM OBJECTIVES COVERED

3.4 Identify common data analytics tools.

Appendix Introduction

Data analysts encounter all types of data sets, from databases to flat files. In order to analyze these varying types of data, we need a toolbox. A data analyst can apply data concepts to any data set, and most tools can be used to accomplish many different goals. However, the way each tool operates depends on how that tool was developed.

There are all sorts of tools. Some are a language used to work with data, others are dedicated to statistical analysis, and still others are part of a platform intended to accomplish a multitude of tasks.

As a data analyst, you may work with any number of tools. It's not necessary for you to learn how to execute every function within every tool, but it is important to understand what type of function you need and how that tool performs it. What's paramount is that you recognize which tools are commonly used in the analyst's world, and why.

Proprietary and Open Source Languages

Each language is different, but they are all uniquely used to deal with data. There are different ways to interact with these languages; you may directly code these languages, or you may work with software that leverages them. Before we dig into more specifics, it's important to understand the difference between open source and proprietary. **Open source** is freely open for people to use, and **proprietary** is custom to the vendor.

Both proprietary and open source languages allow us to work with large data sets, apply transformations, code graphical outcomes, and perform statistical analysis. Just like any language, there are variations in the open source languages used in data analysis. For example, when working within a Microsoft product, you're actually using Transact-SQL (TSQL), which is a Microsoft and Sybase proprietary extension of SQL.

Open source languages allow a program to communicate with the use of **middleware** from the front to the back. Some of the most commonly used open source languages are as follows:

- SQL
- R
- Python

Data Transformation Tools

Tools that are dedicated to data transformation allow us to easily transform our data in order to meet business requirements. It's important to know which tools you should reach for when you need to transform data. Tools like Microsoft Excel, and many others, have made major expansions to support data analysts through routine feature development and expansions. Simple improvements, like row counts in Excel from 65,000 rows to one million, and additions, like Power Query serving both Power BI and Excel for the analysts' work, allow us to effectively manage and work with data that exceeds a million rows. The addition of new features in Tableau, like their data prep tools, also enable us to perform more data preparation than with the original product.

Data transformation tools are used to mine data, explore data, expand data, link to other data sets, and build models. There are several data transformation tools that are very commonly used by data analysts.

- Microsoft Excel
- Tableau Prep
- Microsoft Power BI
- Rapid Miner

Visualization Tools

Tools that are dedicated to building dashboard representations of data allow us to visualize our data. Any of the tools covered here can be used for this purpose; it doesn't matter which tool you select.

Tools that allow us to visually represent data all share common visual elements, such as charts, graphs, tables, and filters. While each tool has its own strengths, it ultimately doesn't matter what tool is selected. Data analysts are typically required to use the software packages that the organization has purchased. Some of the most popular tools for data visualization include the following:

- Tableau
- Microsoft Power BI
- Qlik
- ArcGIS
- AWS QuickSight

Statistical Tools

Although most data-related software involves some form of statistical analysis, like the Data Analysis ToolPak in Excel, using tools dedicated to performing statistical tests will offer more test capability and features for data preparation.

These tools offer more than just the basic statistics, but also different types of statistical analysis examples that include tests for parametric and nonparametric data, prediction and regression models, and other types of statistical analysis like fixed effects and mixed effects. They not only perform the statistical analysis outcomes but visualization. These tools are designed with statistical analysis in mind, and they are accepted by researchers and scholars all over for use of their analysis.

When there is a heavy focus on statistical analysis you will likely find tools that are dedicated to not only working with data but have a focus on comprehensive statistical tests and ease of use. Some of the most popular tools for statistical analysis include the following:

- SAS
- IBM SPSS
- IBM SPSS Modeler
- Stata
- Minitab

Paginated Reporting Tools

Before the advent of dashboards, we would find pages of reports being printed with lines of information for use. These reports still provide critical information and are not a thing of the past. However, there are tools that are dedicated to performing the types of reports that are to be printed over multiple pages. Reports that have lines of data and are not suitable for dashboarding often make their way to paginated reports.

These are some common tools used to work with data that is to be reported over multiple pages.

- SQL Reporting Services (SSRS)
- Crystal Reports
- Power BI Report Builder

Platform Tools

Over the years, major software vendors have determined that platforms, or suites of tools combined, can be extremely helpful for an organization. These offerings provide the ability to perform many different tasks within one suite of programs.

Platform tools allow companies to share data, create reports, build dashboards with organizational data, and more. It is also not uncommon for a data professional to work within multiple cloud platforms at their organization.

There are many different types of platforms, and the list continues to grow. Some of the most popular are as follows:

- Business Objects
- MicroStrategy
- Apex
- IBM Cognos
- Dataroma
- Cloudera
- Alteryx
- Oracle Analytics
- Domo
- Microsoft Power Platform
- RapidMiner

Review Activity:

Common Data Analytics Tools

Answer the following questions:

1. **An organization that is very involved in deep research and statistical analysis would need the flexibility and tools provided by what type of software?**

2. **An organization has decided to start using dashboards to deliver real-time information from the call center to operations. What type of software might the company consider using?**

3. **When a vendor owns the code and does not share it out for public use, what is this software called? What about software that is freely available for public use?**

4. **What type of software is a whole suite of (typically cloud-based) tools that provides many different options for an organization?**

Appendix A

Summary

While each tool that we encounter will likely specialize in a particular function, they will still share features with other, similar software. As a data analyst, your main responsibility is to understand the underlying concepts of the work you're performing, regardless of the tool that will be used. When you do know what tool you will need to use, you will need to learn how that tool specifically handles the tasks you're performing. Typically, the software, platforms, and features you will be required to use are those purchased by your organization. You may work with several types of software within the same organization.

Guidelines in Identifying Common Data Analytics Tools

Consider these best practices and guidelines when familiarizing yourself with the various tools you may use while fulfilling the data analyst role.

1. When using a tool for the first time, you should identify the tool's specialty, so you can then know what scenarios it would be most appropriate for.

2. When you have multiple tools at your disposal, use the best one for the project at hand.

3. Remember that licensing plays a large role in the features that are available to any user in a software or platform, and thus you must be able to identify whether a software is proprietary or open source.

4. While there are dedicated data transformation tools, data transformation happens in almost every type of data-related software.

5. When needing to show data in the forms of graphs, charts, and maps, visualization tools might be your best choice for these capabilities.

6. When you need statistical tests that do more than you can accomplish with the Excel Data Analysis ToolPak, you should explore dedicated statistical tools which have a larger breadth of tests.

7. Paginated reporting tools are designed to work with data that is in list format with line items.

8. Organizations implement platform tools as a way to cover many of their various needs for data, like the ability to build visuals, store data, and share information.

Practice Questions: Additional practice questions are available on the CompTIA Learning Center.

Appendix B

Mapping Course Content to CompTIA Data+ Certification (Exam DA0-001)

Achieving CompTIA Data+ certification requires candidates to pass Exam DA0-001. This table describes where the exam objectives for Exam DA0-001 are covered in this course.

1.0 Data Concepts and Environments	Covered in
1.1 Identify basic concepts of data schemas and dimensions. Databases Relational Non-relational Data mart/data warehousing/data lake Online transactional processing (OLTP) Online analytical processing (OLAP) Schema concepts Snowflake Star Slowly changing dimensions Keep current information Keep historical and current information	Lesson 1, Topic A Lesson 2, Topic A Lesson 2, Topic B
1.2 Compare and contrast different data types. Date Numeric Alphanumeric Currency Text Discrete vs. continuous Categorical/dimension Images Audio Video	Lesson 1, Topic B Lesson 3, Topic A Lesson 3, Topic B

1.0 Data Concepts and Environments	Covered in
1.3 Compare and contrast common data structures and file formats. Structures Structured Defined rows/columns Key value pairs Unstructured Undefined fields Machine data Data file formats Text/Flat file Tab delimited Comma delimited JavaScript Object Notation (JSON) Extensible Markup Language (XML) HyperText Markup Language (HTML)	Lesson 4, Topic A Lesson 4, Topic B Lesson 4, Topic C

2.0 Data Mining	Covered in
2.1 Explain data acquisition concepts. Integration Extract, transform, load (ETL) Extract, load, transform (ELT) Delta load Application programming interfaces (APIs) Data collection methods Web scraping Public databases Application programming interface (API)/web services Survey Sampling Observation	Lesson 5, Topic A Lesson 5, Topic B Lesson 5, Topic C Lesson 5, Topic D
2.2 Identify common reasons for cleansing and profiling data sets. Duplicate data Redundant data Missing values Invalid data Non-parametric data Data outliers Specification mismatch Data type validation	Lesson 6, Topic A Lesson 6, Topic B Lesson 6, Topic C Lesson 6, Topic D Lesson 6, Topic E

2.0 Data Mining	Covered in
2.3 Given a scenario, execute data manipulation	Lesson 1, Topic B
techniques.	Lesson 7, Topic A
Recoding data	Lesson 7, Topic B
Numeric	Lesson 7, Topic C
Categorical	Lesson 8, Topic A
Derived variables	
Data merge	
Data blending	
Concatenation	
Data append	
Imputation	
Reduction/aggregation	
Transpose	
Normalize data	
Parsing/string manipulation	
2.4 Explain common techniques for data manipulation	Lesson 8, Topic A
and query optimization.	Lesson 8, Topic B
Data manipulation	
Filtering	
Sorting	
Date functions	
Logical functions	
Aggregate functions	
Query optimization	
Parametrization	
Indexing	
Temporary table in the query set	
Subset of records	
Execution plan	

3.0 Data Analysis	Covered in
3.1 Given a scenario, apply the appropriate descriptive	Lesson 9, Topic A
statistical methods.	Lesson 9, Topic B
Measures of central tendency	Lesson 9, Topic C
Mean	
Median	
Mode	
Measures of dispersion	
Range	
Max	
Min	
Distribution	
Variance	
Standard deviation	
Frequencies/percentages	
Percent change	
Percent difference	
Confidence intervals	

3.0 Data Analysis	Covered in
3.2 Explain the purpose of inferential statistical methods. t-tests Z-score p-values Chi-squared Hypothesis testing Type I error Type II error Simple linear regression Correlation	Lesson 11, Topic A Lesson 11, Topic B Lesson 11, Topic C
3.3 Summarize types of analysis and key analysis techniques. Process to determine type of analysis Review/refine business questions Determine data needs and sources to perform analysis Scoping/gap analysis Type of analysis Trend analysis Comparison of data over time Performance analysis Tracking measurements against defined goals Basic projections to achieve goals Exploratory data analysis Use of descriptive statistics to determine observations Link analysis Connection of data points or pathway	Lesson 10, Topic A Lesson 10, Topic B
3.4 Identify common data analytics tools. Structured Query Language (SQL) Python Microsoft Excel R Rapid mining IBM Cognos IBM SPSS Modeler IBM SPSS SAS Tableau Power BI Qlik MicroStrategy BusinessObjects Dataroma Domo AWS QuickSight Stata Minitab	Appendix A

4.0 Visualization	Covered in
4.1 Given a scenario, translate business requirements to form a report. Data content Filtering Views Date range Frequency Audience for report Distribution list	Lesson 13, Topic A Lesson 13, Topic C Lesson 13, Topic D Lesson 13, Topic E
4.2 Given a scenario, use appropriate design components for reports and dashboards. Report cover page Instructions Summary Observations and insights Design elements Color schemes Layout Font size and style Key chart elements Titles Labels Legends Corporate reporting standards/style guide Branding Color codes Logos/trademarks Watermark Documentation elements Version number Reference data sources Reference dates Report run date Data refresh date Frequently asked questions (FAQs) Appendix	Lesson 14, Topic A Lesson 14, Topic B Lesson 14, Topic C

4.0 Visualization	Covered in
4.3 Given a scenario, use appropriate methods for dashboard development. Dashboard considerations Data sources and attributes Field definitions Dimensions Measures Continuous/live data feed vs. static data Consumer types C-level executives Management External vendors/stakeholders General public Technical experts Development process Mockup/wireframe Layout/presentation Flow/navigation Data story planning Approval granted Develop dashboard Deploy to production Delivery considerations Subscription Scheduled delivery Interactive (drill down/roll up) Saved searches Filtering Static Web interface Dashboard optimization Access permissions	Lesson 13, Topic B Lesson 13, Topic C Lesson 13, Topic D Lesson 13, Topic E Lesson 14, Topic D
4.4 Given a scenario, apply the appropriate type of visualization. Line chart Pie chart Bubble chart Scatter plot Bar chart Histogram Waterfall Heat map Geographic map Tree map Stacked chart Infographic Word cloud	Lesson 12, Topic A Lesson 12, Topic B Lesson 12, Topic C Lesson 12, Topic D

4.0 Visualization	**Covered in**
4.5 Compare and contrast types of reports.	Lesson 15, Topic A
Static vs. dynamic reports	Lesson 15, Topic B
Point-in-time	
Real time	
Ad-hoc/one-time report	
Self-service/on demand	
Recurring reports	
Compliance reports (e.g., financial, health, and safety)	
Risk and regulatory reports	
Operational reports [e.g., performance, key performance indicators (KPIs)]	
Tactical/research report	

5.0 Data Governance, Quality, and Controls	**Covered in**
5.1 Summarize important data governance concepts.	Lesson 16, Topic A
Access requirements	Lesson 16, Topic B
Role-based	Lesson 16, Topic C
User group-based	Lesson 16, Topic D
Data use agreements	
Release approvals	
Security requirements	
Data encryption	
Data transmission	
De-identify data/data masking	
Storage environment requirements	
Shared drive vs. cloud based vs. local storage	
Use requirements	
Acceptable use policy	
Data processing	
Data deletion	
Data retention	
Entity relationship requirements	
Record link restrictions	
Data constraints	
Cardinality	
Data classification	
Personally identifiable information (PII)	
Personal health information (PHI)	
Payment card industry (PCI)	
Jurisdiction requirements	
Impact of industry and governmental regulations	
Data breach reporting	
Escalate to appropriate authority	

5.0 Data Governance, Quality, and Controls	Covered in
5.2 Given a scenario, apply data quality	Lesson 17, Topic A
control concepts.	Lesson 17, Topic B
Circumstances to check for quality	
Data acquisition/data source	
Data transformation/data flow	
Pass through	
Conversion	
Data manipulation	
Final product (report/dashboard, etc.)	
Automated validation	
Data field to data type validation	
Number of data points	
Data quality dimensions	
Data consistency	
Data accuracy	
Data completeness	
Data integrity	
Data attribute limitations	
Data quality rule and metrics	
Conformity	
Non-conformity	
Rows passed	
Rows failed	
Methods to validate quality	
Cross-validation	
Sample/spot check	
Reasonable expectations	
Data profiling	
Data audits	
5.3 Explain master data management (MDM) concepts.	Lesson 18, Topic A
Processes	Lesson 18, Topic B
Consolidation of multiple data fields	
Standardization of data field names	
Data dictionary	
Circumstances for MDM	
Mergers and acquisitions	
Compliance with policies and regulations	
Streamline data access	

Glossary

acceptable use agreement An agreement that describes not only how data can be used, but also for what purpose

accountability When data governance plans are being followed and there are accountability measures in place

actual execution plan The actual process used to execute a query

ad-hoc report A report that is generated in response to a one-time request

Advanced Encryption Standard (AES) A Federal Information Standards (FIPS)-approved cryptographic algorithm that can be used to protect electronic data

aggregate functions Functions that are written for all or a group of records, as opposed to a single record

aggregated data Data that has already been compiled and summarized for the purposes of analysis and reporting

alternative hypothesis The assumption that a relationship exists between two variables

append To combine data from one data set with another data set

appendix A part of the narrative that provides additional details related to the report or process that is not essential to the main content

application programming interface (API) A library of programming utilities used, for example, to enable software developers to access functions of the TCP/IP network stack under a particular operating system

ascending and descending order A method of sorting in which fields are sorted with the minimum on top (for ascending) or maximum on top (for descending)

audience The people who will be using the data within your reports and dashboards

automated validation Using the power of software to ensure data achieves a validated result

bar chart A chart that displays information, listing the categories on the y-axis and the discrete values on the x-axis

batch processing Processing a large amount of data

bubble chart A visual that plots points on an x-axis and y-axis similar to a scatter plot, but with the addition of the size of the dot representing a third variable

captioning Designating more meaningful names for fields in a report or dashboard

cardinality How many possible occurrences of one entity can be associated with the number of occurrences in another

cascade delete Referential integrity setting that deletes all related records when the primary key is deleted

cascade update Referential integrity setting that updated all related records when the primary key is changed

causal relationship A relationship in which one variable is proven to have an effect on another

chi-square statistic A value that compares the size of the difference between the expected result and the actual result

chi-square test A test used to determine if a difference exists between groups; produces the chi-square statistic

cloud drives Drives that are stored on the “cloud”

column chart A chart that displays information, listing the categories on the x-axis and the discrete values on the y-axis

combination chart A chart that combines columns and lines to compare one or more data points (columns) against a trend (line)

compliance report A report that must be run for compliance or regulatory reasons

conceptual data model The conceptual view of what should exist in a data system and how it could be related

confidence interval A calculation of values that describes the certainty or uncertainty of an estimate made on the analysis

continuous data A characteristic of quantitative data that identifies data that can be measured and can use any value

correlation The statistical association between two (or more) equal variables that tells us if one variable changes, the other(s) will too

cross validation Determining whether data collected across different methods is consistent and accurate

custom sorts Sorting when you create the data set to include the value and the sort order you need for your visualization

cyphertext Data that has been scrambled to be unreadable through encryption

dashboard An interactive, visual display of information

data accuracy The state of data being correct and accurate

data at rest Data that is being stored

data audit The process of assessing that the data achieves a specific objective or purpose

data breach When information is read, modified, or deleted without authorization

data classifications A way to categorize or classify data

data completeness The state of data being complete with all expected and required fields entered

data consistency The state of data having been entered consistently and as intended based on business rules

data constraints Integrity rules that limit the types of data that can go into a column or table within a database system

data custodian The person who manages the system on which the data assets are stored

data destruction The legally compliant means through which data must be removed and made inaccessible

data dictionary A document that serves as the authority on all definitions that have been agreed upon for the organization, as well as key metrics

data encryption The process of using algorithms that will “scramble” data from its original plaintext into another form so that it can’t be read

data governance A large umbrella term for a framework used to govern data in an organization

data in transit Data that is actively being transferred

data in use Data that has been transmitted and is now present in memory or being queried

data lake A technology for storing large amounts of structured and unstructured types of information in their original format

data lakehouse A data management system that combines the best of both data warehousing and data lakes

data loss The intentional or accidental loss of information through human error or an ineffective process

data mart A subset of the data warehouse that is dedicated to a specific department or group

data model A model that organizes data and the relationships of data elements so that the data is ready to use and meaningful for every user who needs a report

data owner The person who holds the ultimate responsibility for maintaining the confidentiality, integrity, and availability of the information asset

data profiling Determining the volume of data, the types of data and quality of the data

data retention The time span for which data must be kept

data sovereignty The idea that the country in which data is stored has control over that data

data steward The person who is primarily responsible for data quality

data transmission The process of transferring data

data use agreement An agreement between two parties about the exchange of data that specifies what data will be shared and how that data can be used

data validation The process of confirming the type, structure, and accurate representation of the data

data verification The process of confirming that the data is accurate or true

data warehouse A technology that is dedicated to the store of company data from a wide range of sources for reporting and decision making purposes

database transaction Any change to data in a system, whether its an insertion, deletion, or query

date filter A filter that narrows down data by a starting and ending point

date functions Functions that manipulate fields set to date data type and return date-related information

date hierarchy A hierarchy using a date field that provides high-level information, like year, quarter, month, and day

date table A table that is full of each date and various information, like month, day of the week, week number, year, and other date-related information

de-identification The process of removing fields that can be used to identify an individual or information that must remain anonymous

delimited files Files in which some form of character separates each field of data from the other data fields

delta load The method of loading new data into a data system and updating any existing data that has changed since the last load

denormalized data Data that has not gone through a normalization process and contains repetitive data

dependent variable The variable we are measuring when comparing two groups

derived variable A data point that is derived or created from existing data

dimension table A table that holds attributes or the categorial information that supports the fact tables

discrete data A characteristic of quantitative value that identifies data that can be counted and can only take on a certain number of values

distributed processing The process of distributing large-volume data sets across multiple servers

distribution list The people who are in your audience, who will receive a report or dashboard

domain integrity The acceptable values for a field

dot map A map that displays geographic data using markers to note specific spots on the map

drill-through capability The ability to select a value and drill down to a deeper visualization of that information

duplicated data Data that is repeated within the same data set

dynamic report A report that is connected to the data and can be refreshed on demand or regularly updated automatically; also known as real-time report

ELT (Extract, Load, Transform) The process that occurs when moving data from source systems to data lakes, which holds data in preparation for transformation

empirical rule The tendency of most data points in normal distribution to fall within three points of the mean either on the positive or negative side of the curve

entity integrity The unique identifier of a record as defined using a primary key field

entity relationship diagram The pictorial representation of a database model; also known as entity relationship model

estimated execution plan The estimated requirements for executing a query

ETL (Extract, Transform, Load) The process that occurs when moving data from source systems to data warehouses by extracting data from the source, transforming the data, and then loading it to the warehouse

exploratory analysis Analysis that determines the main characteristics of a data set

Extensible Markup Language (XML) A system for structuring documents so that they are human- and machine-readable; information within the document is placed within tags, which describe how information within the document is structured

fact table A table that holds the "facts" about a particular business process or event and contains keys to relate to the other tables

field definitions Descriptive information about what each field contains, intended to clarify field names that may be ambiguous

filled map A map that displays geographic data by filling in the borders of a location

flat files Delimited files that are exported out of a system

foreign key A field or fields that are primary in another table

frequency The number of times that a data point occurs within a data set

frequently asked questions Information that addresses anticipated, common questions related to the report and its data, often included as part of the narrative; also known as FAQs

full load The method of loading all data into a data system for the very first time

gap analysis The study of a present state, desired state, and the gaps that exist between the two

goodness of fit A chi-square test that tests against a single variable to analyze the relationship between variables

hard-coded filters Filters that are coded into the view or the visual

heat map A visual that uses color to draw attention to a "hot" spot, or a part of the visual that needs pointing out; also known as color scale

histogram A chart that groups values into bins, or class intervals, on the x-axis and lists the metric we want to assess against on the y-axis

HyperText Markup Language (HTML) A system of coded tags that identify the structure of the document files used for web pages

hypothesis statement What you believe to be true, and what you will test with analysis techniques to show that it is true (or maybe false)

imputing Replacing data with an estimated value

independent variable The variable that is different between two groups that we are comparing

index field A field that applies a unique number to a record

indexing A field property setting that tells the database that a field needs to be indexed

infographic Any combination of visuals, artwork, photos, and language that tells the story of your data in a compelling and graphically appealing way

inline append An append query that combines data sets until all are combined.

intellectual property (IP) Intangible products of human thought and ingenuity

interactive filters Filters that allow the consumer to adjust a slicer or filter option on a dashboard to narrow down the data they want to see

intermediate append An append query that creates a combined data set but also retains the separate data sets

invalid data Data that is incorrect

JavaScript Object Notation (JSON) An object-oriented, event-driven programming language that allows us to interact with websites

joins A join line between fields in a query

jurisdiction The official power to make legal decisions and judgments

key performance indicators (KPIs) Measurements/goals that are established to help identify whether a business is achieving its objectives

Key Value Pair A type of non relational structure that establishes a unique identifier or key field and maps it to a value

layered maps Maps that display geographic data by both using markers and filling in borders

leading questions Questions that are written in a way that produces the desired answer

leading spaces Spaces at the front of a field of information

legend A labeling element that lets you know which color represents which value in a visual

lifecycle of data The five stages of the life of data: create, store, use, archive, and destroy

line graph A graph that consists of either a single horizontal line or a group of multiple lines that represent different data points at different times; also known as run chart

link analysis Analysis that helps us determine how a single data point links to other data points

local drives Drives that are local to individual equipment

logical data model A more detailed view of the conceptual model that includes data fields and the relationships between them

logical functions Functions that check if a condition is met and return a result based on whether or not the condition is met

machine data Data that is produced by a machine rather than a human

many-to-many relationship Many records in a table are associated to many other records in other tables

masking The act of hiding the original value of data by showing something else in its place; also known as anonymization

master data Important data about general business operations that is used frequently in analysis

master data management Tools and processes that are used to create the single source of truth or the "golden record" for the data that is considered critical at the organization

max The largest number in a data set

mean The average of a set of numbers, calculated by adding all the values and then dividing that sum by the total number of values

measures of central tendency Mathematical functions used to find the center of a data set

measures of dispersion Mathematical functions used to determine the distribution of a data set; also known as measures of variability

median The middle number within a group of sorted numbers

memorandum of understanding (MOU) An acceptable use agreement that establishes the rules of engagement between two parties and defines roles and expectations

merge fields function A function used to combine different fields to create and display a single consolidated field; also known as CONCATENATE function

middleware Software that is the middle between an operating system and the applications that use a language

min The smallest number in a data set

mockup To draw out a potential layout

mode The number that shows up most often in a data set

multiprocessing The process of two or more processors working on a single data set

multi-sorting A method of sorting where you sort within a sort

"n" count The amount of data being used in research

narrative A description of a report or dashboard that typically includes a report cover page and a summary of the contents

natural order The order in which the data is entered

nominal data A characteristic of qualitative data that identifies information that has no natural order

non-printable characters Characters that do not produce a written symbol

non-relational database A database system that stores data without the use of relational database models; also known as NoSQL

non-disclosure agreement (NDA) An agreement that defines the conditions under which an entity cannot disclose information to outside parties

non-parametric data Data that is not within the rules of normal distribution, with values that frequently deviate from the mean

normal distribution A bell-shaped curve that depicts the data distribution for a data set

normalized data Data that is structured for optimal storage and use within a program

NULL There is no value in a data field

null hypothesis The assumption that a relationship does not exist between two variables

one-to-many relationship One record in a table is associated with multiple records in another table or tables

one-to-one relationship One record in a table is associated only with one record in another table

online analytical processing (OLAP) A class of software that allows complex analysis to be conducted on large databases without negatively affecting transactional systems

online transactional processing (OLTP) A class of software that allows large numbers of database transactions in real time, typically over the internet

open source Freely available for public use

operational report A report that informs on the health and status of a project, product, or organization

ordinal data A characteristic of qualitative data that identifies information that follows a natural order

outliers Values in the data set that don't seem to be within the norm of all the other data

page footer A place for information at the bottom of each page of a paginated report

page header A place for information at the top of each page of a paginated report

paginated report A multipage report that is not suitable for display on a dashboard

parameter A method of adding a criteria to a query that can be used to filter and reduce the result set

parametric data Data that is within the rules of normal distribution

parsing Breaking data into parts

Pearson's correlation coefficient A calculation used to measure a linear relationship between data points, returning a value that is plus or minus one to determine the strength of the relationship

peer review A method of validating data that involves gaining feedback from others before publication

percentage change A calculation that represents the difference between a new value and an original value (either the last value or an older value)

percentage difference A calculation performed by determining the absolute value of the difference between two numbers, dividing by the average of the two values, and then multiplying by 100%

performance analysis Analysis that measures the performance of a particular product, outcome, or scenario against a defined objective

personally identifiable financial information (PIFI) Information about a consumer provided to a financial institution

personally identifiable information (PII) Data that can be used to directly or indirectly identify an individual; also known as personally identifiable data (PID)

physical data model The actual data system with tables, relationships, fields, and attributes

pie chart A circle broken into slices to represent percentages of information

point-in-time report A report that reflects on a specific point in time

population The population of a group of records that meet a certain criterion

primary key A unique identifier for a record that cannot contain duplicates and is used to reference a record

print date The date when the report was printed

proprietary Owned by a vendor

protected health information (PHI) Health-related data that can be used to identify an individual

public data Data that has been made available to the public through various legal requirements

publicly available data Data that has been made available to the public without any legal requirement

p-value A value in statistical tests that tells the probability that an observed difference occurred by chance

qualitative data Data that can be arranged into groups or categories based on its qualities; also known as categorical data

quality assurance (QA) The process of ensuring that the data used in analysis is of a high enough quality that it gives decision makers confidence in the findings; also known as data cleaning

quality data Data that has integrity and is accurate, complete, and consistent

quantitative data Data that is portrayed through numbers, meaning it can have a numerical value or be measured

query execution plan The order of steps in which a query is processed

querying Gathering fields from one or more tables; also known as merging or blending

r value A value that when closer to 1 shows a strong correlation between values and when closer to 0 means no correlation; also known as correlation coefficient value

range The difference between the highest and lowest values in a data set

read permissions Permission to read data

read/write permissions Permission to read and also change data

real-time processing The processing of data needed in real time

real-time report A report that automatically updates to provide the most current data

recoding Changing the current value of a variable to a different value

record link restrictions Instances in which data systems cannot ever be related to each other through organizational policies

record linkage The process of identifying, matching, and merging records that correspond to a matching record; also known as data linkage

recurring report A report that is set to repeatedly run on certain dates or at specific times

reduction Reducing the volume of data

redundant data Identical data that is stored in multiple places

references Information about the data sources used for a report or dashboard

referential integrity Established to maintain that records are not orphaned by ensuring the proper table has the key field

refresh date The date and time that the data was last updated

regression analysis A type of statistical method used to estimate the relationship between a dependent variable and one or more independent variables

regulations Rules that are implemented by an authority for organizations that fall under that authority's control

relational database Structured database in which information is stored in tables where columns represent typed data fields and rows represent records. Tables can have relationships, established by linking a unique primary key field in one table with the same value in a foreign key field in another table. The overall structure of a particular database and its relations is called a schema; a relational database is also known as a tabular schema.

relational database management system (RDBMS) Software that maintains relational databases

report footer A place for information on the last page of a paginated report

report header A place for information at the top of the first page of a paginated report

research questions Questions that are formed to determine the focus of analysis

research-driven report A report that relies on research to inform and even change business practices

role-based permissions Permissions granted according to the role served at a company

root cause analysis Analysis conducted for a research-driven report that attempts to find the root cause of a problem

scatter plot A visual that consists of two variables plotted on the x-axis and y-axis, with a dot placed on the graph where the two data points converge on both of the axes

scope The overall outline of the project, with measurable tasks that are needed to meet the desired end state

scope creep What occurs when the scope changes from the original plan and incurs adjustments

self-service report A report that is run directly by the consumer; also known as on-demand report

semi-structured data Data that is a mix of both structured and non-structured data

shared drives Drives in a server-based environment that utilize permissions to control access

simple linear regression A statistical method used to study the relationship between one independent variable and one dependent variable

simple random sampling A type of sampling for which each record of data has an equal chance of being selected into the dataset used for analysis

slowly changing dimensions A way of updating dimension data

snowflake schema The relational model of fact and dimension tables that looks like a snowflake, in which dimension tables are joined to a single fact table and also other dimension tables

sorting A method of ordering data by ascending or descending by letter, number, or date

source system The system of record for any given data element or piece of information

spreadsheet A worksheet of data in tabular form

stacked chart A chart that breaks the bar or column into separate portions, each representing an additional data point

standard deviation A value that shows how dispersed the data is in relationship to the mean of all of the data

standardization When data is consistently labeled, categorized, and described for use within the organization

Standardized Generalized Markup Language (SGML) A language that provides the standard for all markup languages and is widely used for data structures

star schema The relational model of fact and dimension tables that looks like a star, in which all dimension tables are joined to a single fact table

static report A report that does not update automatically

statistical significance The designation of the difference between two values as being significant and not by chance

stop words Common or high-frequency words that are used in the business, but not appropriate for counting in word clouds

stratified sampling A type of sampling in which you break your data into subgroups and then randomly sample from each of the groups

structured data Data that is organized and stored in tables, in rows and columns

structured query language (SQL) Programming and query language common to many relational database management systems.

style guides Guidelines that detail how reports must be presented for a specific organization

subquery A query that is nested inside another query statement; also known as nested query

system functions Functions that track information about a report

tactical dashboard A dashboard that is focused on the operational details of a process or operation

talking points A narrative element in reports that supports a presentation in explaining the information

temporary table A table that is stored on the database server until a user disconnects from the server

test of independence A chi-square test that tests against multiple variables to analyze the distribution of the data

text functions Functions that manipulate text-based data, such as TRIM or CLEAN

tooltips A popup display that allows you to show additional information about a data point when you hover over it in the visual

Top N and Bottom N sorts Sorts that display the top or bottom of the data in a set based on a number you specify, or N

trailing spaces Spaces at the end of a field of information

transaction processing The processing of transactional data that is mission critical to an organization

transparency When everyone in the organization has access to the data governance policies and understands why they are in place

transpose To reverse the direction of data

treemap A rectangle that shows the proportion of values using smaller rectangles within the larger one

trend analysis Analysis that measures a trend on historical data to determine performance over time and predict a future outcome

t-test A test used to compare two groups when determining whether there is a significant difference between the means of both groups

type I error An error that occurs when the correct hypothesis is rejected and the incorrect hypothesis is accepted; also known as false positive

type II error An error that occurs when the incorrect hypothesis is accepted and the correct hypothesis is rejected; also known as false negative

undefined fields Fields that are not defined in a database table

unstructured data Data that is not organized in a predefined manner to meet the standards for structured data

user group permissions Permissions that are specific to users in a group regardless of their roles

user-defined integrity Integrity based on business rules that are covered by the other data integrity settings

variance The average squared distance from the mean of the data for a single data point

version numbers A reference that identifies which iteration of a report is being viewed

visual filters The ability to collapse or expand rows to see more information or details

waterfall chart A chart that visualizes performance over time with a series of columns

watermark Information displayed as an overlay on a report indicating content is confidential or should not be printed or distributed

web scraping The act of pulling information from a website; also known as data scraping

web service Any software that provides network communication between devices

wireframing Creating mockups of multiple screens that are likely connected

word cloud A visual representation of the words used in a particular body of text

z-score A value that shows how many standard deviations a data point is from the mean

Solutions

Activity 1-1: Relational and Non-Relational Databases

Answer the following questions:

1. **Which type of database should you use when you need a defined structure with relationships within the data?**

Relational databases are designed with tables where a structure is defined.

2. **What are the two main components of a table design?**

Field names and data types

3. **Which type of database stores data in document format and uses XML or JSON?**

Non-relational or NoSQL database

4. **What is the language that is used for relational databases?**

SQL, or structured query language

5. **Which type of database provides the most scalable and flexible option for web-based applications?**

NoSQL or non-relational databases provide the most scalable and flexible options for web-based design.

Activity 1-2: Tables, Primary Keys, and Normalization

Answer the following questions:

1. **Which elements of a database are necessary in order to establish relationships between tables?**

Primary keys and foreign keys, which are used to uniquely identify information across tables

2. **To ensure data updates effectively across tables, what can be set on the relationships between tables?**

Cascade updates, a referential integrity setting

3. **When the design of the data forces a person to repetitively enter the same information over and over, this data would be considered ______.**

Denormalized

4. **A database architect creates a design in which a table has a primary key and associated fields. This is an example of what type of design theory?**

Designing in the normal forms

Activity 2-1: Types of Data Processing and Storage Systems

Answer the following questions:

1. **Which subset of a data warehouse holds data that is relevant to a specific department?**

Datamart

2. **What is the most flexible storage system for holding large amounts of both structured and unstructured data?**

Data lake

3. **In an organization, which technology holds a system of record, or is a software system dedicated to a specific task?**

Source system

4. **What choices do organizations have when they want to combine data from different source systems into one unified data management system?**

Data warehouses, data lakes, and data lakehouses

5. **What type of table stores information that is categorical for the use of reporting?**

Dimension tables

6. **What are two common schemas that are used to relate data in a data warehouse?**

Star schema and snowflake schema

Activity 2-2: Explain How Data Changes

Answer the following questions:

1. **What process does slowly changing dimensions describe?**

Slowly changing dimensions describes three ways to change dimension data in a system. There are three types: Type 1, Type 2, and Type 3.

2. **Which type of slowly changing dimension is in use when a new record is added to account for a change to dimension data?**

Type 2

3. **If the name of a product changes using Type 1 slowly changing dimensions, what impact would this method have on reporting prior to the change?**

This method will only show the most current name of the product in reports dated prior to the name change, even if the product name was originally something different.

4. Which slowly changing dimension type provides the most complete history of dimension data, and why?

Type 2 has the most complete history of change, as every record of the change is maintained, with time stamps, in the table dedicated to that dimension.

Activity 3-1: Types of Data

Answer the following questions:

1. What two types of data are at the highest level?

Quantitative and qualitative

2. Suppose a survey asks respondents for their level of education. This would be considered qualitative data with which characteristic?

This is ordinal data, because it has a natural occurring order.

3. If you ask people to tell you their favorite tv show, what type of data are you collecting?

People's favorite TV shows would be qualitative data because it is not numerical. We can count the number of people who gave a specific response, but the name of the TV show itself is not a number.

4. When determining how much product was ordered over time, the quantity ordered is which high-level type of data?

Because quantity is a number, this would be considered quantitative data.

5. When data can be measured, which characteristic of data is it?

Continuous

Activity 3-2: Field Data Types

Answer the following questions:

1. Why are field data types important to the data analyst?

Because they determine how you work with a field. If the field is already set up with the correct data type, then you can just work with it in calculations. But if it is not set as the data type that suits your needs, you must convert it.

2. Which field data type is the most versatile of all?

Alphanumeric, also known as "text" or "string"

3. What feature controls what type of information is stored in a field?

The field data type designed within a system

4. Who defines the field level data type in the source system, and what can an analyst do if it is incorrect for the type of analysis required?

The designer of a database defines the field data type, but an analyst can convert that data type to meet their need.

Activity 4-1: Structured and Unstructured Data

Answer the following questions:

1. **Would data that is stored in a relational database be considered structured or unstructured data, and why?**

A relational database stores structured data; we know this because structured data fits into columns and rows.

2. **Video, audio, and images are all examples of what type of data?**

Unstructured data, because these items do not fit neatly into structured data tables with field names and require additional action, like watching a video, to gain context.

3. **Is most data in the world structured or unstructured, and why?**

The majority of data in the world is unstructured. A process must be performed to structure data into tables and fields.

Activity 4-2: File Formats

Answer the following questions:

1. **What are some common delimiters you will likely encounter in text-based files?**

Comma, pipe, and tab are common delimiters.

2. **What is the difference between importing and exporting?**

Importing is the act of bringing data into a system, while exporting is the act of taking it out.

3. **What is a major drawback to working with CSV files?**

CSV files are not connected to the live data, so the file will not update when the data updates. A new file must be exported to see the changes.

Activity 4-3: Code Languages Used for Data

Answer the following questions:

1. **Which markup language is dedicated to the presentation of information on the web?**

HTML (HyperText Markup Language)

2. **Which markup language supports the ability to use an array?**

JSON (JavaScript Object Notation)

3. What language is used to write queries in a structured database environment?

SQL (Structured Query Language)

4. Which markup language supports the transfer of data between systems and is a child of SGML?

XML (Extensible Markup Language)

Activity 5-1: The Processes of Extracting, Transforming, and Loading Data

Answer the following questions:

1. Which method is most commonly used to move data from a source system into a data warehouse?

ETL is the most commonly used method for data warehouses, because data warehouses require that data has thought and structure prior to loading.

2. Which method is most commonly used to move data from a source system into a data lake?

ELT is the most commonly used method for data lakes, because having the transformations take place after the data is loaded allows the data to be moved into the data storage system faster.

3. What is the biggest difference between the processes of ETL and ELT?

In the ETL process, the data must be transformed before it is loaded. There is a fair amount of planning that has to occur to know the transformations and structure before it is loaded. ELT allows the data to be loaded first, and then transformed when it becomes necessary.

4. Which load type is being used when new data is loaded and existing data that has changed since the last load is updated?

Delta load

Activity 5-2: API/Web Scraping and Other Collection Methods

Answer the following questions:

1. What is a benefit of using an API?

Answers may vary. Potential answers might include the following:

- Allows two unrelated systems to communicate.
- Accesses data from a dedicated system.
- Removes the need for you to build the data system yourself.
- Accesses data that is not stored in the company's internal system.

2. What is the key difference between a web service and most other APIs?

The use of a hosted network

3. What language do web services use?

XML

4. What is web scraping?

Web scraping is the act of pulling information from a website.

5. Sensors that detect the temperature of a person on a job site would be collecting what type of data?

This is machine data, because it is collected by a machine and doesn't require a human to manually track and key it into a system.

Activity 5-3: Public and Publicly Available Data

Answer the following questions:

1. When an organization is required to provide data by law, what would it be called? And what are some sites that provide this type of data?

Public data; U.S. Census Bureau, Data.Gov, state or federal agencies

2. What is a consideration of public data that must be addressed before use?

Answers may vary. Potential answers might include the following:

- Reading the terms of use
- Reading the methods of data collection
- Reading any key definitions that are available

3. Public data sources often do not provide the individual data. They will typically provide the data in what format?

It is typically aggregated—data that has already been compiled and summarized (e.g., summed, counted, or averaged by group) for the purposes of analysis and reporting.

Activity 5-4: Survey Data

Answer the following questions:

1. What should a survey be free of?

Bias

2. Which survey question type uses a scale to gain details about customers' agreement with or attitude toward certain topics?

A Likert scale

3. **What are some common answer types you will find on most surveys?**

Answers may vary. Potential answers might include the following:

- Single option
- Multiple choice
- Likert scale
- Text-based response

4. **Why is it important that a survey provides well-written questions and appropriate answer choices?**

When researching for information to improve processes or gauge effectiveness, appropriate questions and answer types help us get to the truth of how well something is performing. Insufficient or inappropriate questions and answer types can skew the results and thus lead to ineffective research.

Activity 6-1: Learn to Profile Data

1. **Why is it important for the data analyst to profile data?**

Answers may vary. Potential answers might include the following:

- Learn the basics of the data
- Discern information about the data
- Detect any data quality issues

2. **What are some elements of data that are assessed when profiling?**

Answers may vary. Potential answers might include the following:

- The source of the data
- Keys of the data
- Relationships within the data
- Record counts

3. **Name a few popular tools that can be used to profile data.**

Answers may vary. Potential answers might include the following:

- Power Query
- Power BI
- Tableau

Activity 6-2: Redundant, Duplicated, and Unnecessary Data

Answer the following questions:

1. **What does it mean for data to be redundant?**

The same data is represented in multiple places.

2. **When data is duplicated and then totaled, what problem will you encounter?**

The duplicates will artificially inflate the totals.

3. **Spotting duplicates can be a challenge in a large data set. What type of command could you use to highlight them?**

Conditional formatting in tools such as Excel highlights duplicates.

4. **Your data set contains every field of that table or query, and only a few fields are relevant for your study. What should you do with the unnecessary fields?**

They should be removed.

5. **What are some of the drawbacks of leaving unnecessary fields in your data?**

Those fields increase the size of your data. The fields will show as potential fields for use in data software. They create clutter or "noise" in your data.

Activity 6-3: Missing Values

Answer the following questions:

1. **What appears in a data set when a field has no value?**

Null value

2. **Describe two examples of how NULL values can be useful when working with data?**

Answers may vary. Potential answers might include the following:

- A NULL value can be used as a filter.
- A NULL value can be used to indicate that something has not yet occurred (e.g., a NULL value for a ship date indicates that a product hasn't been shipped).
- Null values can indicate that there is no match for the data, which can be helpful when trying to determine if all the data in one system is present in another system.

3. **When a survey is completed and an answer is skipped, what can you do with that information?**

You must handle skipped responses with care. You may find that you must exclude that record from your results.

Activity 6-4: Invalid Data

Answer the following questions:

1. **What is invalid data?**

Data that is not correct

2. **When writing a query to create a data set, what should you do with the fields that have invalid data?**

Do not include them in the statement.

3. **How can non-printable characters, leading spaces, and trailing spaces cause invalid data?**

A data program will read the spaces as actual valid characters.

4. **What should you do when you encounter invalid values that can definitively be defined?**

Replace the values with the correct ones.

Activity 6-5: Convert Data to Meet Specifications

Answer the following questions:

1. **You need to create calculations that are date specific, and you discover that the date field is set as text. What should you do?**

Convert that field to a date data type.

2. **What are some of the issues of data not having the correct data type?**

You can't use text to calculate a number. Invalid data types might cause data to fail to transfer between systems.

Activity 7-1: Manipulate Field Data and Create Variables

Answer the following questions:

1. **What is one reason why data may need to be recoded?**

Answers may vary. Potential answers might include the following:

- Make the data more meaningful
- Group it more effectively for analysis
- Correct the data

2. Imagine a data set with a column for Start Time and a column for End Time, where a column for Total Time has been added to the data set. The Total Time column would be an example of what?

A derived variable

3. What does it mean to impute values?

Replace data with an estimated value

4. Why would a value be imputed?

Missing data or null values can cause issues in analysis.

5. Reduction can be achieved through what two methods?

Aggregation and sampling

Activity 7-2: Transpose and Append Data

Answer the following questions:

1. Which action reverses the direction of the data?

Transpose

2. What does it mean to append data?

To append can be described as creating a combined data set from multiple sets. You can also append data by copying it from one location to another. An intermediate append creates an entirely new set, and an inline append combines data into an existing set.

3. What command can you use to reverse the direction of the data?

Answers may vary. Potential answers might include the following:

- Transpose in paste commands in Excel
- Unpivot command in Power Query and other data tools

Activity 7-3: Query Data

Answer the following questions:

1. Which action involves merging multiple data sets into a single data set?

Querying

2. Which join type results in a display that only contains records that exist in both tables?

Inner join

3. Which join type is used the least, and why?

Cross join, because it associates every record from one table to every record from another table

4. Which join type provides results that help us troubleshoot potential bad records or records due to be corrected?

Full outer join

Activity 8-1: Functions to Manipulate Data

Answer the following questions:

1. Which function removes all non-printable characters? Which function trims leading and trailing spaces? What type of function would we use to classify these as?

CLEAN and TRIM. They are text functions.

2. Which function is used to combine data fields?

Merge Fields or CONCATENATE

3. If you had a data set with a column for Full Name, and you needed columns for First Name and Last Name, what action would you need to perform?

Parsing

4. The TODAY function will provide and not provide what, respectively?

It will provide the date but not the time.

5. Which function is used to test whether a condition is true or false?

IF

6. SUM, COUNT, DISTINCT COUNT, AVERAGE, MAX, and MIN are all what type of function?

Aggregate functions

7. Which function type is program or tool dependent?

System functions

Activity 8-2: Common Techniques for Query Optimization

Answer the following questions:

1. What can you create to optimize data load time and filter data to certain criteria?

A parameter

2. What fields are automatically indexed when they are created?

Key fields

3. When does a temporary table stop being stored on the database server?

When the user disconnects from the server

4. What is being created when querying another query?

A subquery

5. What are the two types of query execution plans?

Estimated execution plan and actual execution plan

Activity 9-1: Measures of Central Tendency

Answer the following questions:

1. Which measure of central tendency is calculated by adding all the values together, counting the number of values, and dividing the sum by that number?

Mean

2. What are values in the data set that don't seem to be within the norm of all the other data?

Outliers

3. Which measure of central tendency is the middle value within an ordered set of numbers?

Median

4. When you count which value shows up most frequently in a set of numbers, what will you end up with? What if two or more values show up an equal amount of times?

Single mode (if there's one number that appears most frequently) and multiple mode (if two or more numbers appear most frequently)

Activity 9-2: Measures of Dispersion

Answer the following questions:

1. What calculation involves finding the highest (maximum) value and the lowest (minimum) value?

Range

2. What is the *x*-bar representative of in the calculation for standard deviation?

Mean

3. What is defined as following a bell shape curve, with the mean being the middle and all other data following three points to the left or three points to the right of the mean?

Normal distribution

4. **What is the tendency of most data points to fall within three points of the mean either on the positive side or the negative side of the curve?**

Empirical rule

5. **How do you test whether data is following the empirical rule through visualization?**

Calculate the z-score and plot the values on the distribution curve to visualize how much deviation is occurring

Activity 9-3: Frequency and Percentages

Answer the following questions:

1. **Differences between two data sets could be expressed using what calculation?**

Percentage difference

2. **What are two aspects of calculating frequency?**

Grouping and counting

3. **A pivot table, histogram, and percentages are three ways of displaying what?**

Frequency

4. **Which calculation involves subtracting the newest value from the last value, dividing that number by the last value, and then multiplying by 100%?**

Percentage change

Activity 10-1: Get Started with Analysis

Answer the following questions:

1. **In which manner should a research question be posed, and why?**

A research question should be posed as a true or false statement so that it can be answered as part of hypothesis testing.

2. **Why is it important to determine the sources and collection methods for the data you're working with?**

Different collection methods have different considerations and can potentially result in vastly different qualities of data. You need to understand the source and quality before you start working with the data.

Activity 10-2: Types of Analysis

Answer the following questions:

1. **Profiling data is similar to what type of analysis?**

Exploratory analysis

2. **Which type of analysis is conducted on non-parametric data, or data that does not follow the normal distribution curve?**

Performance analysis

3. **Performance analysis measures the performance of a particular product, outcome, or scenario against what?**

A defined objective

4. **Which metric establishes goals to help identify whether a business is achieving its objectives?**

Key performance indicators (KPIs)

5. **Which type of analysis studies the difference between the desired state and current state?**

Gap analysis

6. **Trend analysis is commonly used for what purpose?**

Forecasting future values

7. **What are the three main components of link analysis?**

A network, a node, and a link

Activity 11-1: The Importance of Statistical Tests

Answer the following questions:

1. **The calculation of values that describes the certainty or uncertainty of an estimate made on the analysis is known as what?**

Confidence interval

2. **What is the percentage of a confidence interval that is most commonly strived for in analysis?**

95%

3. **What are the two variables we use when conducting a t-test?**

Dependent variable and independent variable

4. **Which two conditions must be met for data to be considered statistically significant?**

It cannot have happened by chance and must be a significant difference.

5. **What does the “p” in p-value stand for?**

Probability

Activity 11-2: The Hypothesis Test

Answer the following questions:

1. **Which type of hypothesis assumes that a relationship between two variables does exist?**

An alternative hypothesis

2. **Which type of hypothesis assumes that a relationship between two variables does not exist?**

A null hypothesis

3. **Which type of error creates a false negative?**

Type II error

4. **What are some of the impacts of type I and type II errors on hypothesis testing? Use an example.**

Answers may vary. Potential answers might include the following:

Students are not given critical study hours when this study time could have helped improve their test scores, because the test error said the study time would not help. This is a type I error.

Patients who are not sick get the treatment unnecessarily (type I error), and patients who are sick get no treatment (type II error).

Activity 11-3: Tests and Methods to Determine Relationships Between Variables

Answer the following questions:

1. **What type of analysis will typically involve an *x* and *y* scatter plot with a line?**

Simple linear regression

2. **What are two commonly used types of chi-square tests?**

Goodness of fit and test of independence

3. **Correlation is used to measure what?**

The relationship between two (or more) equal variables

4. **A relationship in which one variable is proven to have an effect on another would be considered what?**

A causal relationship

5. **Regarding simple linear regression, how do we refer to the variables and the outcome?**

Predictors and criterion

Activity 12-1: Basic Visuals

Answer the following questions:

1. What type of visual appears as a circle broken into slices to represent percentages of information?

A pie chart

2. A treemap shows what information alongside values?

The proportions

3. What is the difference between a column chart and a bar chart?

The axes are swapped.

4. What do the lines in a line graph represent?

Different data points at different times

Activity 12-2: Advanced Visuals

Answer the following questions:

1. Which type of visual displays additional data points as separate proportions?

Stacked chart (bar or column)

2. Which type of visual would work best for visualizing the run time (in years) of multiple television series?

A line graph with multiple lines

3. Combination charts typically use what type of visual alongside columns (either single or stacked)?

Line graph

4. What differs between how a scatter plot and a bubble chart plot points?

The size of the data point

5. In which way does a histogram display values that a column chart does not?

The ability to show a frequency of values grouped by bins

6. Which type of visual is best to display performance over time?

A waterfall chart

Activity 12-3: Maps with Geographical Data

Answer the following questions:

1. **Which software is dedicated to mapping and offers comprehensive mapping components?**

ArcGIS

2. **Which data transformation will likely be needed before visualizing geographic fields?**

Converting data types

3. **Which type of map is used to highlight the amount of data occurring within a particular point?**

Dot map

4. **Which type of map is used to visualize multiple data points?**

Layered map

5. **A filled map highlights what portions of a visual with a color?**

Border/boundary

Activity 12-4: Visuals to Tell a Story

Answer the following questions:

1. **A heat map can also be referred to as what?**

Color scales

2. **What can be used in Excel to create a color scale?**

Conditional formatting

3. **The data points of a geographic heat map are usually colored based on what?**

Density

4. **Which type of visual would be appropriate for visualizing open text responses from a survey?**

A word cloud

5. **What are some tools that might be used to create an infographic?**

Answers may vary. Potential answers might include the following:

- Adobe Illustrator
- Adobe InDesign
- Microsoft Word
- Canva

Activity 13-1: Audience Needs When Developing a Report

Answer the following questions:

1. **The people who are in your audience would also be considered what?**

Distribution list

2. **C-level executives, management, external vendors/stakeholders, the general public, and technical experts are all types of what?**

Consumers

3. **Why is it important to consider who will be the users of a report?**

This information can affect what type of data you need to present and also what access and viewing requirements should be set.

Activity 13-2: Data Source Considerations for Reporting

Answer the following questions:

1. **What organizes the data and the relationships of data elements so that the data is ready to use and meaningful for every user who needs a report?**

A data model

2. **If only part of a table from a database should be shared for the audience of a report, what can be created to only provide the permitted data for access?**

A view

3. **The creation of what only allows the software in use to process the data that is needed up front?**

A hard-coded filter

4. **What should be provided to help clarify the names of data fields?**

Field definitions

Activity 13-3: Considerations for Delivering Reports and Dashboards

Answer the following questions:

1. **What key point must be considered when preparing to create the visuals of a report or dashboard?**

How they will be viewed/displayed

2. **What are two methods we might use to publish Power BI reports?**

PDF and PowerPoint are the most common. We could also use Excel exports and printed pages.

3. **What limits the delivery options of a report or dashboard?**

The tool/software/platform used to create it

4. **When planning the delivery of a report, we must consider how the dashboard will be delivered, and what else?**

The frequency, or how often the report will be delivered

5. **Recurring reports can be delivered in multiple ways depending on the organization's policies. What are a few ways these reports are delivered?**

Answers may vary. Potential answers might include the following:

Printed, shared folders on a server; email; dashboards

Activity 13-4: Develop Reports or Dashboards

Answer the following questions:

1. **What are the three main types of reporting?**

Dashboards, paginated reports, and spreadsheets

2. **What visual is also known as a pivot or cross tab format?**

Matrix

3. **What is defined as creating mockups of multiple screens that are likely connected?**

Wireframing

4. **When should a report be set to automatically refresh?**

When there is real-time access to the data

Activity 13-5: Ways to Sort and Filter Data

Answer the following questions:

1. **Which type of filter does not allow the user to adjust it?**

A hard-coded filter

2. **Which type of filter does give the user the ability to adjust it?**

An interactive filter

3. **What else can serve as a filter?**

A visual

4. **What filter would be a consideration for payroll data that is done on a monthly basis?**

A date filter

5. **What type of table is typically used with the payroll data when working with a report that uses a date filter?**

A date table

6. **When data is not yet sorted, what is the order of the data?**

Natural order, which is the way that it was entered in the system

7. **What are two sort methods you can apply to your data sets?**

Answers may vary. Potential answers might include the following:

- Ascending and descending order
- Multi-sort
- Top N and Bottom N sorts
- Custom sorts

8. **What type of sort will also filter data?**

Top N and Bottom N sorts will also filter data.

Activity 14-1: Design Elements for Reports and Dashboards

Answer the following questions:

1. **Style guides typically contain variations of what dependent on the output?**

The organization's logo and logo placement

2. What are some considerations when choosing appropriate colors?

Answers may vary. Potential answers might include the following:

- Tailoring to the user
- Avoiding distracting colors
- Keeping the color scheme consistent

3. What are the two basic types of fonts?

Serif and sans serif

4. What three categories are typically offered for customizing fonts?

Family, size, and color

5. What does not change the underlying field name in the database or data set, but displays a custom name that clarifies the information being presented?

Captioning

6. What should be considered regarding naming conventions when reporting?

Field names, visual and page titles, labels, and legends

Activity 14-2: Standard Elements for Reports and Dashboards

Answer the following questions:

1. What element should you include when reports are a part of a documented process?

Version numbers

2. What is the difference between a report header and a page header?

A report header is only shown once on the top of the first page of the report, while a page header is shown at the top of every page of the report.

3. What can be added to confidential reports which are not to be distributed?

Watermarks

4. What two dates are important to include in a report?

The refresh date and print date

Activity 14-3: Narrative and Other Written Elements

Answer the following questions:

1. **Narrative provides what type of detail about the report?**

High-level

2. **If you are preparing a dashboard for an audience who is not very familiar with how to work the dashboard, you might supply instructions for its use through which method?**

Simple screenshots and some written instructions

3. **What is typically written in a question-and-answer format?**

FAQs/Frequently Asked Questions

4. **What is used to provide additional details and information that is related, but not crucial, to a report?**

An appendix

Activity 14-4: Deployment Considerations

Answer the following questions:

1. **What are three techniques we use to optimize the dashboard experience and presentation of data?**

Visual filters, drill-through capabilities, and tooltips

2. **What action should you take before deployment to production to ensure you can share your report with others?**

Ensure that licensing requirements have been assigned to the appropriate users

Activity 15-1: How Updates and Timing Affect Reporting

Answer the following questions:

1. **What is the key difference between static reports and dynamic reports?**

Dynamic reports can be refreshed on demand or regularly updated automatically, and static reports cannot.

2. **A report that covers a specified period of time would be referred to as what type of report?**

Point-in-time report

3. **A Power BI dashboard directly connected to a SQL server would be an example of which type of report?**

Dynamic report

Activity 15-2: Types of Reports

Answer the following questions:

1. **Compliance reporting is performed to adhere to rules and standards required by certain regulatory agencies and organizations. Provide some examples of regulations that would require reporting for compliance.**

Answers may vary. Potential answers might include the following:

OSHA, SEC, Sarbanes-Oxley Act, HIPPA

2. **Which type of report uses metrics to provide a picture of the overall health of the organization?**

Operational report

3. **Tactical dashboards focus on what part of a process or operation?**

Operational details

4. **Which type of report relies on research to inform and even change business practices?**

Research-driven report

5. **A report generated in response to a one-time request is known as what?**

Ad-hoc report

Activity 16-1: Data Governance

Answer the following questions:

1. **What are the three key elements of a strong data governance plan?**

Transparency, accountability, and standardization

2. **What are the five steps of the data lifecycle, in order from start to finish?**

Create, store, use, archive, destroy

3. **Which data governance role is primarily responsible for data quality?**

A data steward

4. **How is data sovereignty defined?**

It is the idea that the laws governing the country in which data is stored have control over that data. It describes the legal dynamics of the collection and usage of data in a global economy.

5. **Which regulation is a US federal law designed to protect the privacy of healthcare-related information?**

The Health Insurance Portability and Accountability Act (HIPPA)

6. There are typically how many levels of data classification in an organization?

Three

7. What are three common classifications?

Public, sensitive, and confidential

Activity 16-2: Access Requirements and Policies

Answer the following questions:

1. What is an acceptable use agreement?

It is an agreement that describes not only how the data can be used, but also for what purpose.

2. What are the two other most common types of data use agreements?

The non-disclosure agreement (NDA) and memorandum of understanding (MOU)

3. What process will typically occur when you need to work with private information and data that has sensitive and confidential classifications?

Release approvals

4. How do we refer to the time span for which data must be kept?

Data retention

5. What do we call the legally compliant means through which data must be removed and made inaccessible?

Data destruction

Activity 16-3: Security Requirements

Answer the following questions:

1. What is data that is actively being transferred?

Data in transit

2. In what state should data be before transmission?

Encrypted

3. What is the approved cryptographic algorithm that can be used to protect electronic data?

The Advanced Encryption Standard (AES)

4. What is the difference between de-identification and masking?

De-identification is the process of removing fields that can be used to identify an individual or information that must remain anonymous. Masking involves hiding that type of field by showing something else in its place, like an asterisk.

5. **What might be needed if a data breach is escalated?**

A public disclosure

6. **What are the two types of data access we commonly work with?**

Read and read/write

7. **What are the three most common storage options for data?**

Shared drives, cloud drives, and local drives

Activity 16-4: Entity Relationship Requirements

Answer the following questions:

1. **What are the three basic types of entity relationship models?**

Conceptual data model, logical data model, and physical data model

2. **What is the definition of an entity relationship diagram?**

The pictorial representation of a database model that shows how entities (like people or objects) relate to each other through the data

3. **What is set to ensure two systems are never able to communicate?**

Record link restrictions

4. **What are data constraints?**

Rules that are set on records or fields of data

5. **Data constraints can occur at what level?**

A system-to-system level, record level, or field level

Activity 17-1: Characteristics, Rules, and Metrics of Data Quality

Answer the following questions:

1. **What is the difference between data validation and data verification?**

Data validation is the process of confirming the type, structure, and accurate representation of the data, while data verification is the process of confirming that the data is accurate or true.

2. What are a few reasons why we might check the quality of data?

Answers may vary. Potential answers might include the following:

- A merger has occurred.
- Data has been manipulated.
- Data has been transferred.
- Data has been transformed.
- Human error may have occurred.

3. What do we call data that is accurate, complete, and consistent?

Quality data

4. A system with a state field that always lists states as an abbreviated value and in proper case has what?

Data consistency

5. What are used to set measurable goals for data quality based on the needs of an organization?

Metrics

Activity 17-2: Reasons to Quality Check Data and Methods of Data Validation

Answer the following questions:

1. What process involves validating that the data can achieve the specific purpose of an objective?

A data audit

2. What are at least two examples of how automated validation helps us achieve a validated result?

Answers may vary. Potential answers might include the following:

- It tells us how many records succeeded or failed during a data transfer.
- It can provide a list or count of duplicated records.
- It can perform data quality checks on data loss.
- It can give us a list of failures for missing data.
- It can determine discrepancies between regular data load size or data record count.
- It can reduce and prevent human error for data entry.

3. If data fails to transfer from the source system to another, what would we say has occurred?

Data loss

Activity 18-1: The Basics of Master Data Management

Answer the following questions:

1. **Master data management focuses on what type of data?**

Answers may vary. Potential answers might include the following:

- Dimension data
- Customer Data
- Product Data
- Supplier Data
- Location Data

2. **What does master data management aim to create?**

A single source of truth

3. **What are some benefits of and reasons why organizations have master data management?**

Answers may vary. Potential answers might include the following:

- Compliance with regulations and mandates
- Data integrity
- Data quality
- Streamlining access to data
- Automatically populating data sets

Activity 18-2: Master Data Management Processes

Answer the following questions:

1. **In master data management, what process is used to match the records in different systems to create one record of trust?**

Consolidation

2. **What two processes will likely need to be performed on data for a report that is not standardized through master data management?**

Cleaning and transformation

3. **What resource contains all the master definitions of data within the organization?**

A master data dictionary

Activity App-1: Common Data Analytics Tools

Answer the following questions:

1. **An organization that is very involved in deep research and statistical analysis would need the flexibility and tools provided by what type of software?**

This organization would need software that is dedicated to statistical analysis and has many types of test modules available. Examples would be Stata, Minitab, and SPSS.

2. **An organization has decided to start using dashboards to deliver real-time information from the call center to operations. What type of software might the company consider using?**

This organization would need software that has visualization tools dedicated to dashboarding. Examples would be Tableau, Power BI, or Qlik.

3. **When a vendor owns the code and does not share it out for public use, what is this software called? What about software that is freely available for public use?**

Proprietary implies that the vendor owns the code, software, and tool that is used. When it is available for public use, it is open source.

4. **What type of software is a whole suite of (typically cloud-based) tools that provides many different options for an organization?**

This is a platform tool.