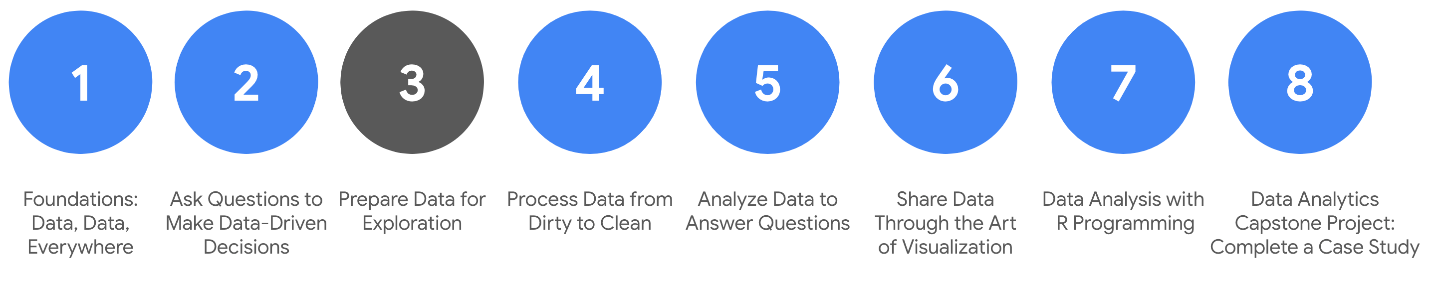
Course syllabus



1. [Foundations: Data, Data, Everywhere](https://www.coursera.org/learn/foundations-data/home/welcome)
2. [Ask Questions to Make Data-Driven Decisions](https://www.coursera.org/learn/ask-questions-make-decisions/home/welcome)
3. **Prepare Data for Exploration** *(this course)*
4. [Process Data from Dirty to Clean](https://www.coursera.org/learn/process-data/home/welcome)
5. [Analyze Data to Answer Questions](https://www.coursera.org/learn/analyze-data/home/welcome)
6. [Share Data Through the Art of Visualization](https://www.coursera.org/learn/visualize-data/home/week/1)
7. [Data Analysis with R Programming](https://coursera.org/learn/data-analysis-r/home/welcome)
8. [Google D​ata Analytics Capstone: Complete a Case Study](https://coursera.org/learn/google-data-analytics-capstone/home/welcome)

Welcome to the third course in the Google Data Analytics Certificate! So far, you have been introduced to the field of data analytics and discovered how data analysts can use their skills to answer business questions.

As a data analyst, you need to be an expert at structuring, extracting, and making sure the data you are working with is reliable. To do this, it is always best to develop a general idea of how all data is generated and collected, since every organization structures data differently. Then, no matter what data structure you are faced with in your new role, you will feel confident working with it.

You will soon discover that when data is extracted, it isn’t perfect. It might be biased instead of credible, or dirty instead of clean. Your goal is to learn how to analyze data for bias and credibility and to understand what clean data means. You will also get up close and personal with databases and even get to extract your own data from a database using spreadsheets and SQL. The last topics covered are the basics of data organization and the process of protecting your data.

And you will learn how to identify different types of data that can be used to understand and respond to a business problem. In this part of the program, you will explore different types of data and data structures. And best of all, you will keep adding to your data analyst tool box! From extracting and using data, to organizing and protecting it, these key skills will come in handy no matter what you are doing in your career as a data analyst.

**Course content**

Course 3 – Prepare Data for Exploration

1. **Understanding data types and structures:** We all generate lots of data in our daily lives. In this part of the course, you will check out how we generate data and how analysts decide which data to collect for analysis. You’ll also learn about structured and unstructured data, data types, and data formats as you start thinking about how to prepare your data for exploration.
2. **Understanding bias, credibility, privacy, ethics, and access:** When data analysts work with data, they always check that the data is unbiased and credible. In this part of the course, you will learn how to identify different types of bias in data and how to ensure credibility in your data. You will also explore open data and the relationship between and importance of data ethics and data privacy.
3. **Databases: Where data lives:** When you are analyzing data, you will access much of the data from a database. It’s where data lives. In this part of the course, you will learn all about databases, including how to access them and extract, filter, and sort the data they contain. You will also check out metadata to discover the different types and how analysts use them.
4. **Organizing and protecting your data:** Good organization skills are a big part of most types of work, and data analytics is no different. In this part of the course, you will learn the best practices for organizing data and keeping it secure. You will also learn how analysts use file naming conventions to help them keep their work organized.
5. **Engaging in the data community (optional):** Having a strong online presence can be a big help for job seekers of all kinds. In this part of the course, you will explore how to manage your online presence. You will also discover the benefits of networking with other data analytics professionals.
6. **Completing the C​ourse Challenge**: At the end of this course, you will be able to apply what you have learned in the Course Challenge. The Course Challenge will ask you questions about the key concepts and then will give you an opportunity to put them into practice as you go through two scenarios.

**What to expect**

This part of the program is designed to get you familiar with different data structures and show you how to collect, apply, organize, and protect data. All of these skills will be part of your daily tasks as an entry-level data analyst. You will work on a wide range of activities that are similar to real-life tasks that data analysts come across on a daily basis.

This course has five modules or weeks, and each has several lessons included. Within each lesson, you will find content such as:

* **V​ideos** of instructors teaching new concepts and demonstrating the use of tools
* **In-video questions** that pop up during or at the end of a video to check your learning
* **Readings** to introduce new ideas and build on the concepts from the videos
* [**Discussion forums**](https://www.coursera.org/learn/data-preparation/discussions) to discuss, explore, and reinforce new ideas for better learning
* **D​iscussion prompts** to promote thinking and engagement in the discussion forums
* **Hands-on activities** to introduce real-world, on-the-job situations, and the tools and tasks to complete assignments
* **Practice quizzes** to prepare you for graded quizzes
* **Graded quizzes** to measure your progress and give you valuable feedback

Hands-on activities promote additional opportunities to build your skills. Try to get as much out of them as possible. Assessments are based on the approach taken by the course to offer a wide variety of learning materials and activities that reinforce important skills. Graded and ungraded quizzes will help the content sink in. Ungraded practice quizzes are a chance for you to prepare for the graded quizzes. Both types of quizzes can be taken more than one time.

As a quick reminder, this course is designed for all types of learners, with no degree or prior experience required. Everyone learns differently, so the Google Data Analytics Certificate has been designed with that in mind. Personalized deadlines are just a guide, so feel free to work at your own pace. There is no penalty for late assignments. If you prefer, you can extend your deadlines by returning to **Overview** in the navigation pane and clicking **Switch Sessions**. If you already missed previous deadlines, click **Reset my deadlines** instead.

If you would like to review previous content or get a sneak peek of upcoming content, you can use the navigation links at the top of this page to go to another course in the program. When you pass all required assignments, you will be on track to earn your certificate.

**Optional speed track for those experienced in data analytics**

The Google Data Analytics Certificate provides instruction and feedback for learners hoping to earn a position as an entry-level data analyst. While many learners will be brand new to the world of data analytics, others may be familiar with the field and simply wanting to brush up on certain skills.

If you believe this course will be primarily a refresher for you, we recommend taking the practice diagnostic quiz offered this week. It will enable you to determine if you should follow the speed track, which is an opportunity to proceed to Course 4 after having taken each of the Course 3 Weekly Challenges and the overall Course Challenge. Learners who earn 100% on the diagnostic quiz can treat Course 3 videos, readings, and activities as optional. Learners following the speed track are still able to earn the certificate.

**Tips**

* Do your best to complete all items in order. All new information builds on earlier learning.
* Treat every task as if it is real-world experience. Have a mindset that you are working at a company or in an organization as a data analyst. This will help you apply what you learn in this program to the real world.
* Even though they aren’t graded, it is important to complete all practice items. They will help you build a strong foundation as a data analyst and better prepare you for the graded assessments.
* Take advantage of all additional resources provided.
* W​hen you encounter useful links in the course, remember to bookmark them so you can refer to the information later for study or review.

# Deciding if you should take the speed track

**This reading provides an overview of a speed track we offer to those familiar with data analytics.**

**If you are brand new to data analytics, you can skip the diagnostic quiz after this reading, and move directly to the next activity:** [**Data collection in our world**](https://www.coursera.org/learn/data-preparation/lecture/QCPVt/data-collection-in-our-world).

The Google Data Analytics Certificate is a program for anyone. A background in data analysis isn’t required. But you might be someone who has some experience already. If you are this type of learner, we have designed a **speed track** for this course. Learners who opt for the speed track can refresh on the basic topics and take each of the weekly challenges and the Course Challenge at a faster pace.

To help you decide if you’re a good match for the speed trackfor this course:

1. Take the optional [diagnostic quiz](https://www.coursera.org/learn/data-preparation/quiz/lV0eb/optional-familiar-with-data-analytics-take-our-diagnostic-quiz).
2. Refer to the [scoring guide](https://www.coursera.org/learn/data-preparation/supplement/EfmvJ/your-diagnostic-quiz-score-and-what-it-means) to determine if you’re a good fit for the speed track. A score of 90% or higher is the target goal for someone on the speed track.
3. Based on your individual score, follow the recommendations in the scoring guide for your next steps.

**Important reminder:** If you’re eligible for the speed track, you’re still responsible to complete all graded activities. In order to earn your certificate, you will need an overall score of 80% or higher on all graded materials in the program.

# Optional: Your diagnostic quiz score and what it means

Use your score to help you determine whether you should take the speed track. The speed track allows you to skip over the lesson material and go straight to the weekly challenges and the course challenge, which lead to your final course score. In order to earn your certificate, you will need an overall score of 80% or higher on all graded materials in this program. Read on to figure out your next steps based on your quiz score:

**If you scored 100% on the diagnostic quiz:**

* You’re probably very familiar with types of data and data structures and can take the speed track to move on to Course 4.
* You must take each of the weekly challenges and the course challenge, which will count toward the 80% overall score needed to earn the certificate. To help you find these items more quickly, we’ve identified them with asterisks in the course materials (for example: \*course challenge\*).
* After you complete the weekly challenges and course challenge, proceed to Course 4.
* You’re welcome to review videos, readings, and activities throughout the course based on your interests.

**If you scored between 90% and 99% on the diagnostic quiz:**

* You’re probably familiar with the types of data and data structures and might consider taking the speed track to move on to Course 4.
* However, we still recommend that you go through the Course 3 lesson materials to review areas where you might have some gaps before proceeding to Course 4.
* You must take each of the weekly challenges and the course challenge, which will count toward the 80% overall score needed to earn the certificate. To help you find these items more quickly, we’ve identified them with asterisks in the course materials (for example: \*course challenge\*).
* After you complete the weekly challenges and course challenge, proceed to Course 4.
* You’re welcome to review videos, readings, and activities throughout the course based on your interests.

**If you scored between 80% and 89% on the diagnostic quiz:**

* You likely have some background knowledge on types of data and data structures.
* However, we recommend that you go through the Course 3 lesson materials to review areas where you might have some gaps before proceeding to Course 4.
* You must take the weekly challenges and the course challenge, which will count toward the 80% overall score needed to earn the certificate. To help you find these items more quickly, we’ve identified them with asterisks in the course materials (for example: \*course challenge\*).

**If you scored less than 80% on the diagnostic quiz:**

* No problem — this course was made for you!
* We strongly recommend that you go through all of the Course 3 videos, readings, and activities, as the concepts taught are building blocks that will set you up for success on your learning path.
* You must take the weekly challenges and the course challenge, which will count toward the 80% overall score needed to earn the certificate.

Regardless of your score, the course material can help you supplement or identify gaps in your knowledge. Whether you take the speed track or complete the certificate at the recommended pace, good luck on your data endeavors!

# Selecting the right data

Following are some data-collection considerations to keep in mind for your analysis:

## How the data will be collected

Decide if you will collect the data using your own resources or receive (and possibly purchase it) from another party. Data that you collect yourself is called first-party data.

## Data sources

If you don’t collect the data using your own resources, you might get data from second-party or third-party data providers. **Second-party data** is collected directly by another group and then sold. **Third-party data** is sold by a provider that didn’t collect the data themselves. Third-party data might come from a number of different sources.

## Solving your business problem

Datasets can show a lot of interesting information. But be sure to choose data that can actually help solve your problem question. For example, if you are analyzing trends over time, make sure you use time series data — in other words, data that includes dates.

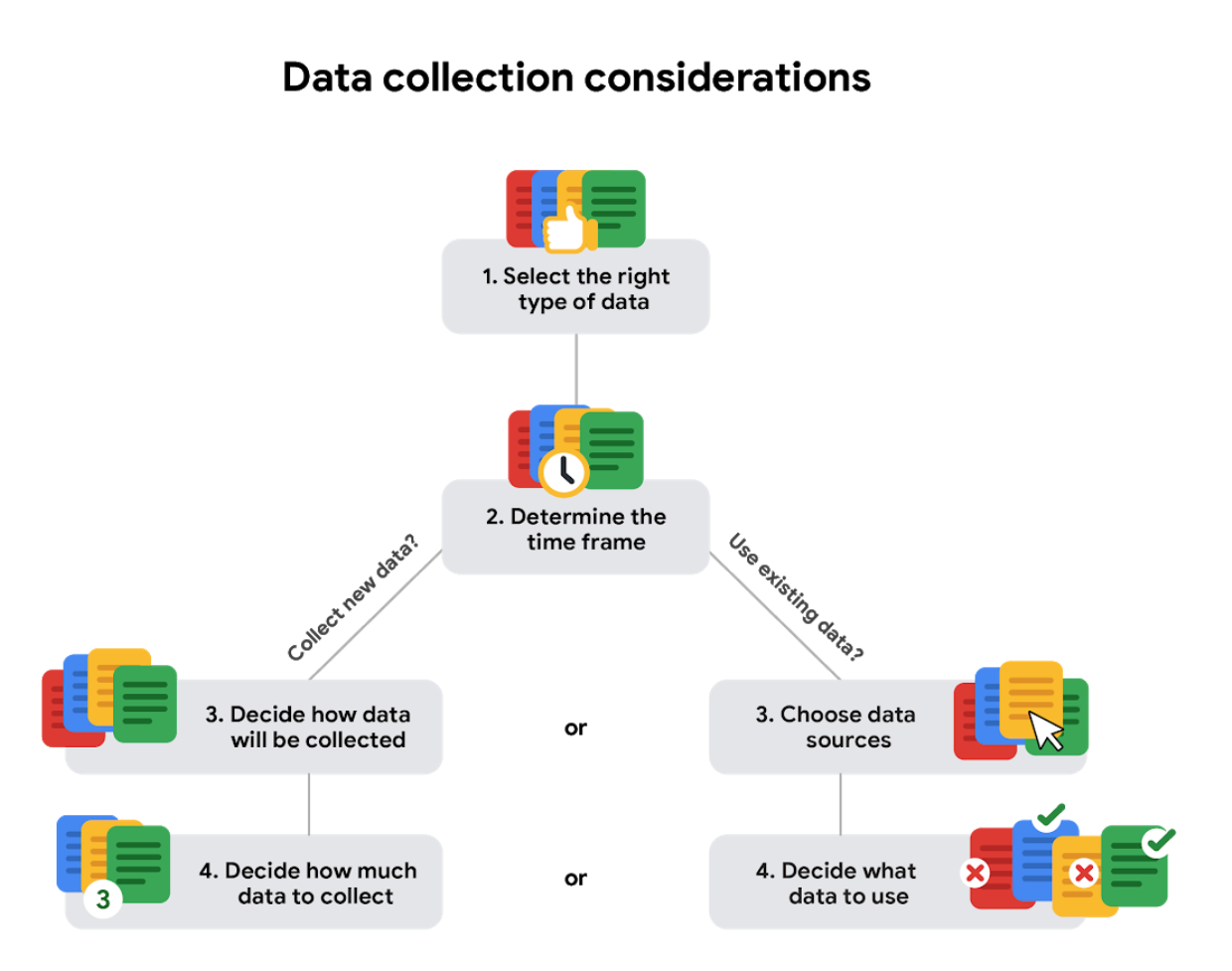
## How much data to collect

If you are collecting your own data, make reasonable decisions about sample size. A random sample from existing data might be fine for some projects. Other projects might need more strategic data collection to focus on certain criteria. Each project has its own needs.

## Time frame

If you are collecting your own data, decide how long you will need to collect it, especially if you are tracking trends over a long period of time. If you need an immediate answer, you might not have time to collect new data. In this case, you would need to use historical data that already exists.

Use the flowchart below if data collection relies heavily on how much time you have:



# Data formats in practice

When you think about the word "format," a lot of things might come to mind. Think of an advertisement for your favorite store. You might find it in the form of a print ad, a billboard, or even a commercial. The information is presented in the format that works best for you to take it in. The format of a dataset is a lot like that, and choosing the right format will help you manage and use your data in the best way possible.

Schematic

Description automatically generated with medium confidence

## Data format examples

As with most things, it is easier for definitions to click when we can pair them with real life examples. Review each definition first and then use the examples to lock in your understanding of each data format.

the following table highlights the differences between primary and secondary data and examples of each

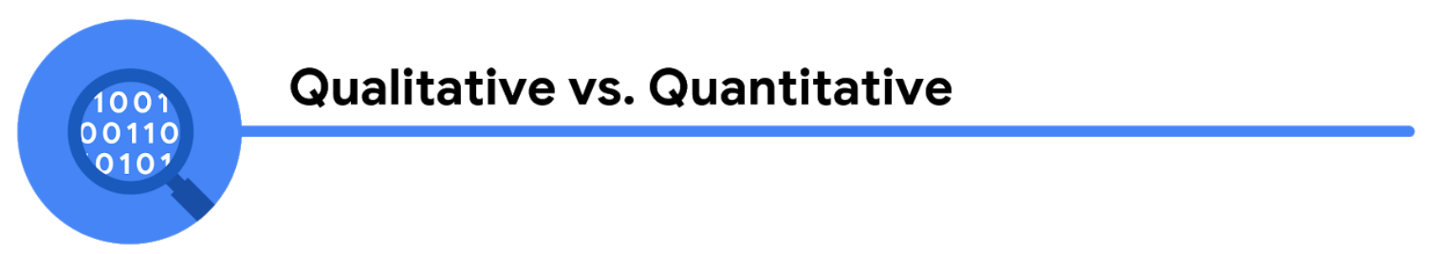
| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Primary data | Collected by a researcher from first-hand sources | - Data from an interview you conducted - Data from a survey returned from 20 participants - Data from questionnaires you got back from a group of workers |
| Secondary data | Gathered by other people or from other research | - Data you bought from a local data analytics firm’s customer profiles - Demographic data collected by a university - Census data gathered by the federal government |

the following table highlights the differences between internal and external data and examples of each

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Internal data | Data that lives inside a company’s own systems | - Wages of employees across different business units tracked by HR - Sales data by store location - Product inventory levels across distribution centers |
| External data | Data that lives outside of a company or organization | - National average wages for the various positions throughout your organization - Credit reports for customers of an auto dealership |

the following table highlights the differences between continuous and discrete data and examples of each

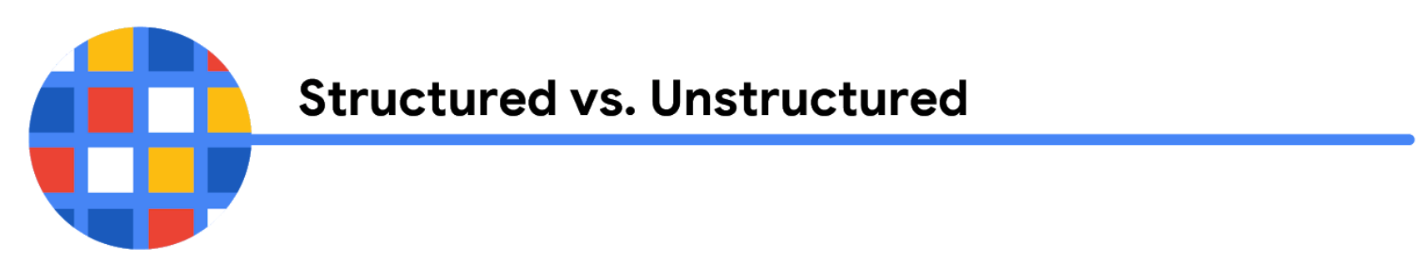
| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Continuous data | Data that is measured and can have almost any numeric value | - Height of kids in third grade classes (52.5 inches, 65.7 inches) - Runtime markers in a video - Temperature |
| Discrete data | Data that is counted and has a limited number of values | - Number of people who visit a hospital on a daily basis (10, 20, 200) - Room’s maximum capacity allowed - Tickets sold in the current month |

the following table highlights the differences between qualitative and quantitative data and examples of each

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Qualitative | Subjective and explanatory measures of qualities and characteristics | - Exercise activity most enjoyed - Favorite brands of most loyal customers - Fashion preferences of young adults |
| Quantitative | Specific and objective measures of numerical facts | - Percentage of board certified doctors who are women - Population of elephants in Africa - Distance from Earth to Mars |

the following table highlights the differences between nominal and ordinal data and examples of each

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Nominal | A type of qualitative data that isn’t categorized with a set order | - First time customer, returning customer, regular customer - New job applicant, existing applicant, internal applicant - New listing, reduced price listing, foreclosure |
| Ordinal | A type of qualitative data with a set order or scale | - Movie ratings (number of stars: 1 star, 2 stars, 3 stars) - Ranked-choice voting selections (1st, 2nd, 3rd) - Income level (low income, middle income, high income) |

the following table highlights the differences between structured and unstructured data and examples of each

| **Data Format Classification** | **Definition** | **Examples** |
| --- | --- | --- |
| Structured data | Data organized in a certain format, like rows and columns | - Expense reports - Tax returns - Store inventory |
| Unstructured data | Data that isn’t organized in any easily identifiable manner | - Social media posts - Emails - Videos |

## Self-Reflection: Unstructured data

**Total points**2

### 1.

**Question 1**



## Overview



Now that you have learned about unstructured data, you can pause for a moment and apply what you are learning. In this self-reflection, you will complete tasks with a neural network, consider your thoughts about data structuring, and respond to brief questions.

This self-reflection will help you develop insights into your own learning and prepare you to apply your knowledge of data structures to your interactions with unstructured data. As you complete tasks with a neural network website, you will explore concepts, practices, and principles to help refine your understanding and reinforce your learning. You’ve done the hard work, so make sure to get the most out of it: This reflection will help your knowledge stick!

## Data structuring with Quick, Draw!



In this self-reflection, you will explore the nature of unstructured data through a crowd-sourced dataset.

**Quick, Draw!** is a neural network dataset that has millions of pictures drawn by people separated into categories like plants, animals, or vehicles. On the Quick, Draw! website, you can view a large dataset of hundreds of thousands of real doodles made by people on the internet. You can also draw your own. Through this process, you can train a neural network to recognize objects and learn more about the importance of structured data.

1. Visit the [Quick, Draw! website](https://quickdraw.withgoogle.com/data/cloud).

2.In the upper left-hand corner, you will notice a drop-down menu like this:

Graphical user interface, application

Description automatically generated

Select a type of doodle to begin.

3. Click on different pictures to see details about the images on your screen. For example, there are more than one hundred thousand different drawings of elephants. Scroll through the list and see if there are any that don’t belong. If you find one that doesn’t match the intended object, click on it and select **Flag as inappropriate**.

4. Explore other categories of drawings. Select three categories that interest you and check out their doodles.

5. Optional: Explore further. Click **Get the data** to visit the GitHub page containing the entire dataset. As you become more familiar with data projects and start creating your own, you can return to this dataset and analyze it yourself. Click **Play the game** to draw your own doodles and contribute to Quick, Draw!’s dataset.

6. When you’re done, answer the reflection questions below.

Timeline

Description automatically generated with low confidence

## Reflection



Consider the doodles you found in the Quick, Draw! Dataset:

* What do you notice as you explored drawings in different categories? Are there consistent themes among the pictures in a category?
* If you didn’t know the category labels, how would you distinguish the pictures from each other? What would you look for?

Now, write 2-3 sentences (40-60 words) in response to each of these questions. Type your response in the text box below.

**1 point**



Your answer cannot be more than 10000 characters.

### 2.

**Question 2**

Consider what you know about structured and unstructured data and how it connects to the Quick, Draw! website:

* How would you describe the Quick, Draw! doodles you explored from a data point of view?
* How are these doodles different from or similar to other types of data that you have previously encountered?
* What about this data makes it unstructured?

Now, write 2-3 sentences (40-60 words) in response to each of these questions. Type your response in the text box below.

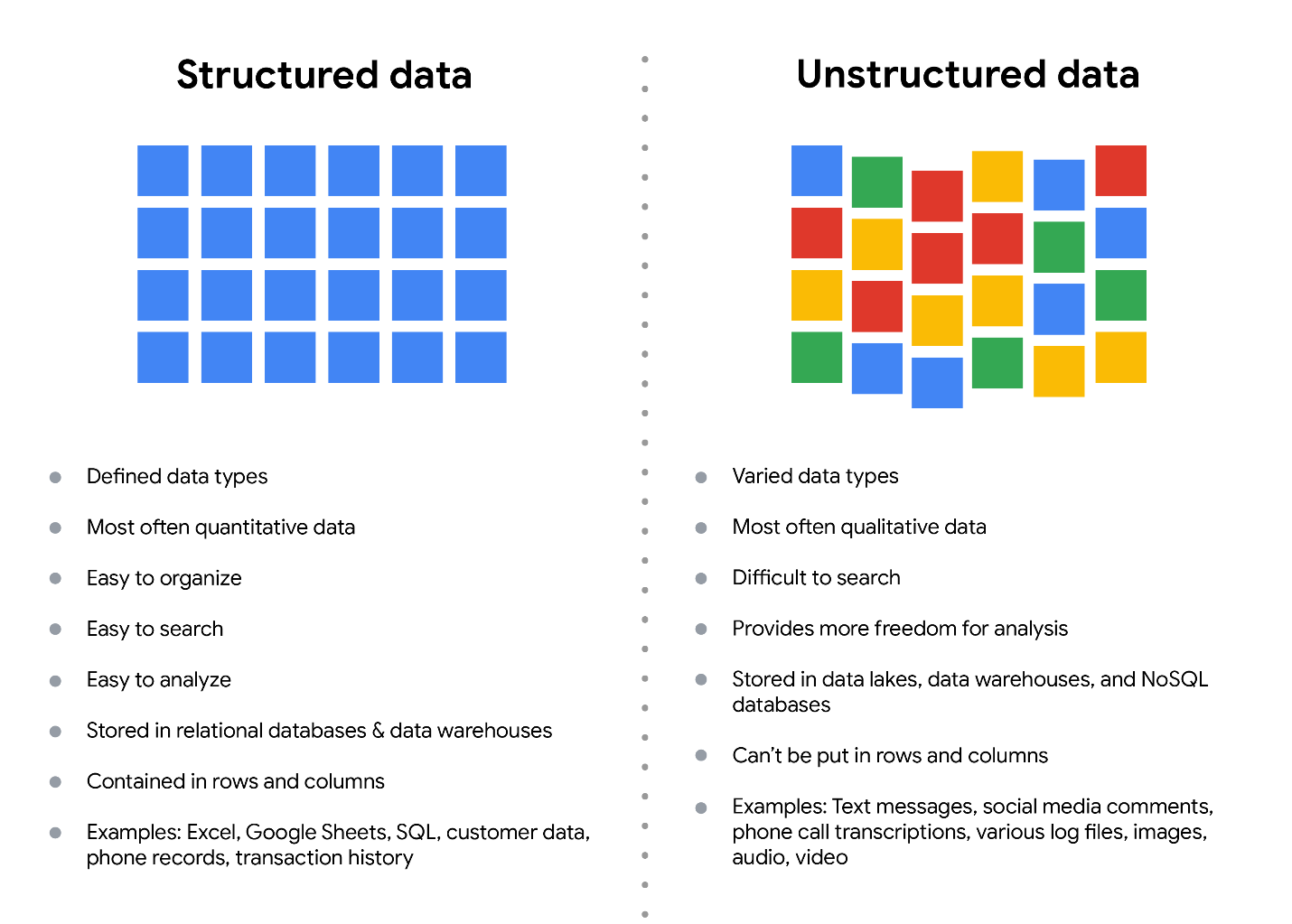
The structure of data

Data is everywhere and it can be stored in lots of ways. Two general categories of data are:

* **Structured data:** Organized in a certain format, such as rows and columns.
* **Unstructured data:** Not organized in any easy-to-identify way.

For example, when you rate your favorite restaurant online, you're creating structured data. But when you use Google Earth to check out a satellite image of a restaurant location, you're using unstructured data.

Here's a refresher on the characteristics of structured and unstructured data:

Structured data: - Defined data types - Most often quantitative data - Easy to organize - Easy to search - Easy to analyze - Stored in relational databases - Contained in rows and columns - Examples: Excel, Google Sheets, SQL, customer data, phone records, transaction history Unstructured data: - Varied data types - Most often qualitative data - Difficult to search - Provides more freedom for analysis - Stored in data lakes and NoSQL databases - Can't be put in rows and columns - Examples: Text messages, social media comments, phone call transcriptions, various log files, images, audio, video

**Structured data**

As we described earlier, **structured data** is organized in a certain format. This makes it easier to store and query for business needs. If the data is exported, the structure goes along with the data.

**Unstructured data**

**Unstructured data** can’t be organized in any easily identifiable manner. And there is much more unstructured than structured data in the world. Video and audio files, text files, social media content, satellite imagery, presentations, PDF files, open-ended survey responses, and websites all qualify as types of unstructured data.

**The fairness issue**

The lack of structure makes unstructured data difficult to search, manage, and analyze. But recent advancements in artificial intelligence and machine learning algorithms are beginning to change that. Now, the new challenge facing data scientists is making sure these tools are inclusive and unbiased. Otherwise, certain elements of a dataset will be more heavily weighted and/or represented than others. And as you're learning, an unfair dataset does not accurately represent the population, causing skewed outcomes, low accuracy levels, and unreliable analysis.

Data modeling levels and techniques

This reading introduces you to data modeling and different types of data models. Data models help keep data consistent and enable people to map out how data is organized. A basic understanding makes it easier for analysts and other stakeholders to make sense of their data and use it in the right ways.

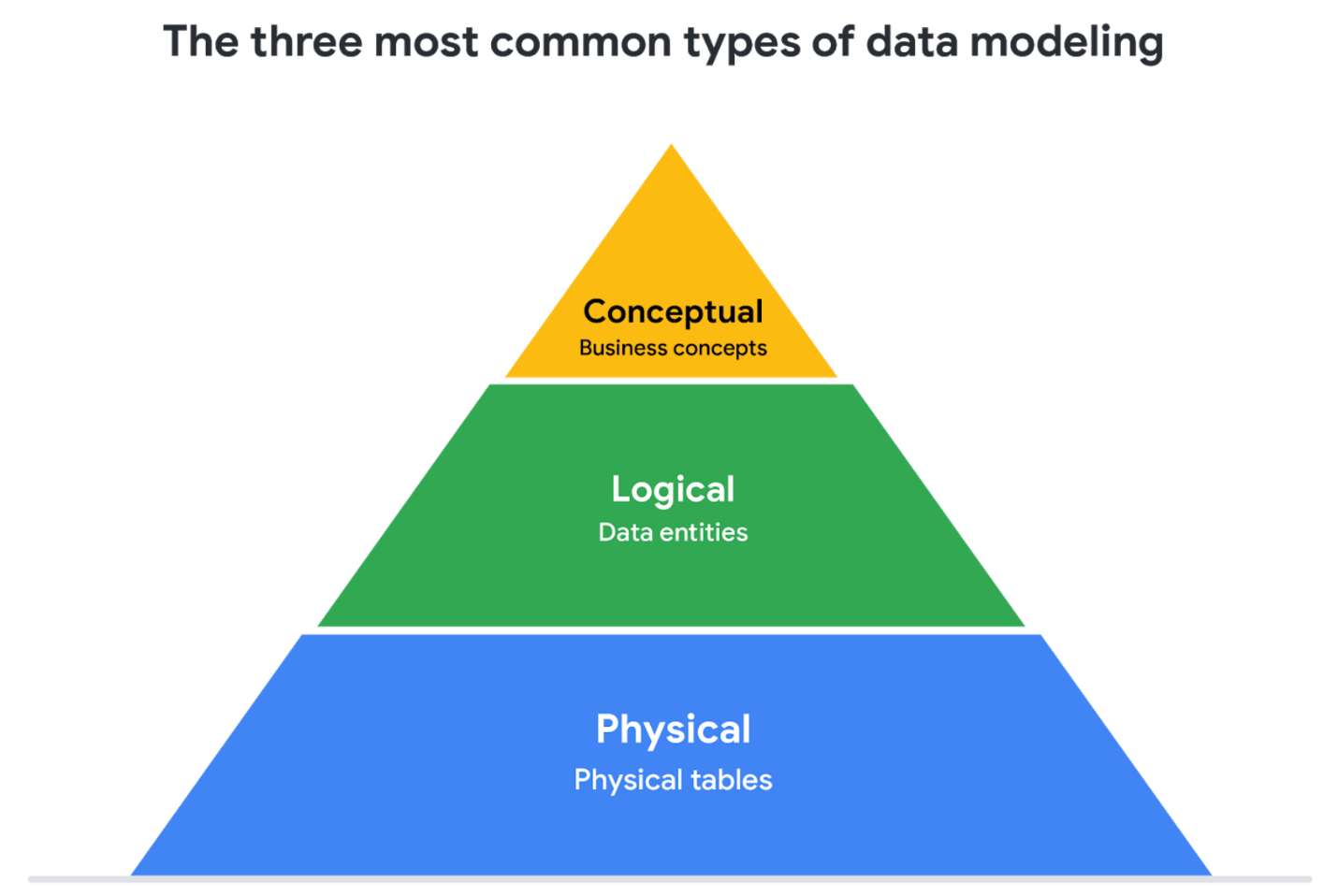
**Important note:** As a junior data analyst, you won't be asked to design a data model. But you might come across existing data models your organization already has in place.

**What is data modeling?**

**Data modeling** is the process of creating diagrams that visually represent how data is organized and structured.  These visual representations are called **data models**. You can think of data modeling as a blueprint of a house. At any point, there might be electricians, carpenters, and plumbers using that blueprint. Each one of these builders has a different relationship to the blueprint, but they all need it to understand the overall structure of the house. Data models are similar; different users might have different data needs, but the data model gives them an understanding of the structure as a whole.

**Levels of data modeling**

Each level of data modeling has a different level of detail.



1. **Conceptual data modeling** gives a high-level view of the data structure, such as how data interacts across an organization. For example, a conceptual data model may be used to define the business requirements for a new database. A conceptual data model doesn't contain technical details.
2. **Logical data modeling** focuses on the technical details of a database such as relationships, attributes, and entities. For example, a logical data model defines how individual records are uniquely identified in a database. But it doesn't spell out actual names of database tables. That's the job of a physical data model.
3. **Physical data modeling** depicts how a database operates. A physical data model defines all entities and attributes used; for example, it includes table names, column names, and data types for the database.

More information can be found in this [comparison of data models​.](https://www.1keydata.com/datawarehousing/data-modeling-levels.html)

**Data-modeling techniques**

There are a lot of approaches when it comes to developing data models, but two common methods are the **Entity Relationship Diagram (ERD)** and the **Unified Modeling Language (UML)** diagram. ERDs are a visual way to understand the relationship between entities in the data model. UML diagrams are very detailed diagrams that describe the structure of a system by showing the system's entities, attributes, operations, and their relationships. As a junior data analyst, you will need to understand that there are different data modeling techniques, but in practice, you will probably be using your organization’s existing technique.

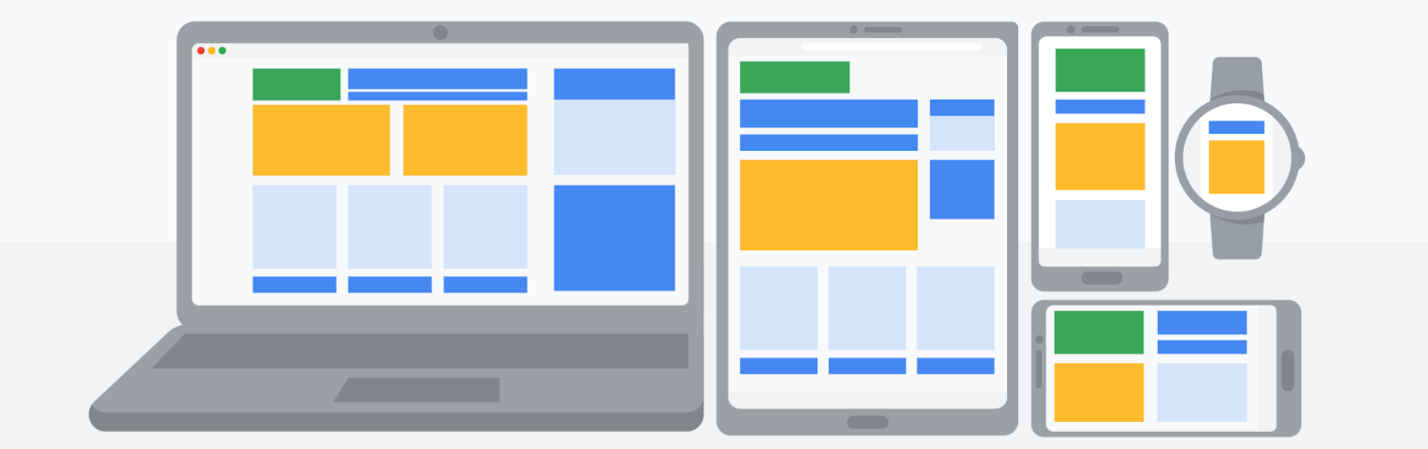
You can read more about ERD, UML, and data dictionaries in this [data modeling techniques article](https://dataedo.com/blog/basic-data-modeling-techniques).

**Data analysis and data modeling**

Data modeling can help you explore the high-level details of your data and how it is related across the organization’s information systems. Data modeling sometimes requires data analysis to understand how the data is put together; that way, you know how to map the data. And finally, data models make it easier for everyone in your organization to understand and collaborate with you on your data. This is important for you and everyone on your team!

# Understanding Boolean logic

In this reading, you will explore the basics of Boolean logic and learn how to use multiple conditions in a Boolean statement. These conditions are created with Boolean operators, including AND, OR, and NOT. These operators are similar to mathematical operators and can be used to create logical statements that filter your results. Data analysts use Boolean statements to do a wide range of data analysis tasks, such as creating queries for searches and checking for conditions when writing programming code.

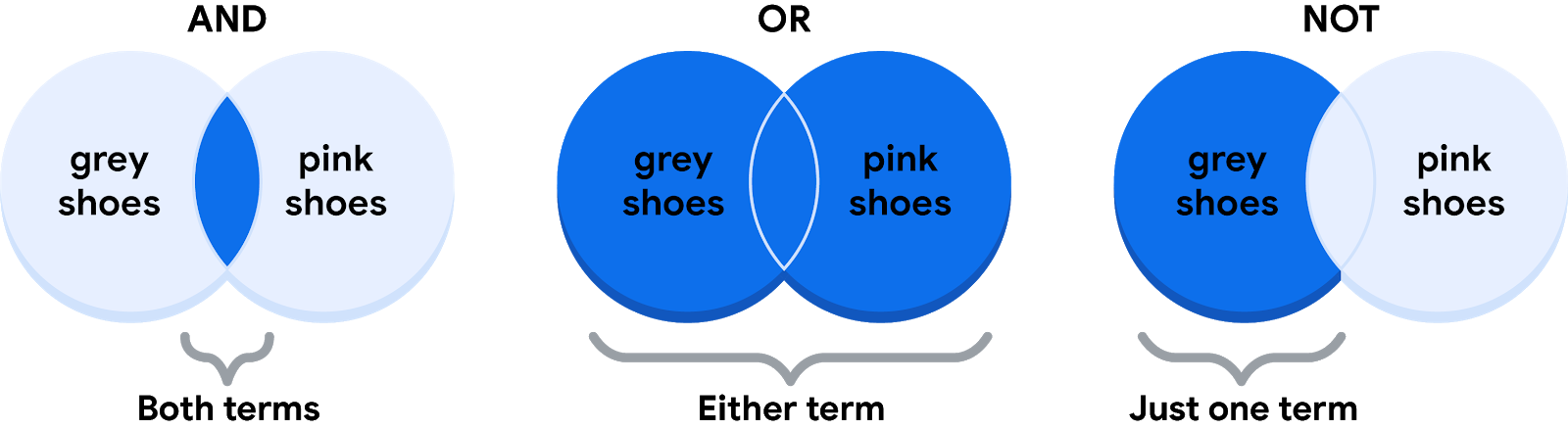


## B​oolean logic example

Imagine you are shopping for shoes, and are considering certain preferences:

* You will buy the shoes only if they are pink and grey
* You will buy the shoes if they are entirely pink or entirely grey, or if they are pink and grey
* You will buy the shoes if they are grey, but not if they have any pink

Below are Venn diagrams that illustrate these preferences. AND is the center of the Venn diagram, where two conditions overlap. OR includes either condition. NOT includes only the part of the Venn diagram that doesn't contain the exception.



### **The AND operator**

Your condition is “If the color of the shoe has any combination of grey and pink, you will buy them.” The Boolean statement would break down the logic of that statement to filter your results by both colors. It would say “IF (Color=”Grey”) AND (Color=”Pink”) then buy them.” The AND operator lets you stack multiple conditions.

Below is a simple truth table that outlines the Boolean logic at work in this statement. In the **Color is Grey** column, there are two pairs of shoes that meet the color condition. And in the **Color is Pink** column, there are two pairs that meet that condition. But in the **If Grey AND Pink** column, there is only one pair of shoes that meets both conditions. So, according to the Boolean logic of the statement, there is only one pair marked true. In other words, there is one pair of shoes that you can buy.

| **Color is Grey** | **Color is Pink** | **If Grey AND Pink, then Buy** | **Boolean Logic** |
| --- | --- | --- | --- |
| Grey/True | Pink/True | True/Buy | True AND True = True |
| Grey/True | Black/False | False/Don't buy | True AND False = False |
| Red/False | Pink/True | False/Don't buy | False AND True = False |
| Red/False | Green/False | False/Don't buy | False AND False = False |

### **The OR operator**

The OR operator lets you move forward if either one of your two conditions is met. Your condition is “If the shoes are grey or pink, you will buy them.” The Boolean statement would be “IF (Color=”Grey”) OR (Color=”Pink”) then buy them.” Notice that any shoe that meets either the **Color is Grey** or the **Color is Pink** condition is marked as true by the Boolean logic. According to the truth table below, there are three pairs of shoes that you can buy.

| **Color is Grey** | **Color is Pink** | **If Grey OR Pink, then Buy** | **Boolean Logic** |
| --- | --- | --- | --- |
| Red/False | Black/False | False/Don't buy | False OR False = False |
| Black/False | Pink/True | True/Buy | False OR True = True |
| Grey/True | Green/False | True/Buy | True OR False = True |
| Grey/True | Pink/True | True/Buy | True OR True = True |

### **The NOT operator**

Finally, the NOT operator lets you filter by subtracting specific conditions from the results. Your condition is "You will buy any grey shoe except for those with any traces of pink in them." Your Boolean statement would be “IF (Color="Grey") AND (Color=NOT “Pink”) then buy them.” Now, all of the grey shoes that aren't pink are marked true by the Boolean logic for the **NOT Pink** condition. The pink shoes are marked false by the Boolean logic for the **NOT Pink** condition. Only one pair of shoes is excluded in the truth table below.

| **Color is Grey** | **Color is Pink** | **Boolean Logic for NOT Pink** | **If Grey AND (NOT Pink), then Buy** | **Boolean Logic** |
| --- | --- | --- | --- | --- |
| Grey/True | Red/False | Not False = True | True/Buy | True AND True = True |
| Grey/True | Black/False | Not False = True | True/Buy | True AND True = True |
| Grey/True | Green/False | Not False = True | True/Buy | True AND True = True |
| Grey/True | Pink/True | Not True = False | False/Don't buy | True AND False = False |

## The power of multiple conditions

For data analysts, the real power of Boolean logic comes from being able to combine multiple conditions in a single statement. For example, if you wanted to filter for shoes that were grey or pink, and waterproof, you could construct a Boolean statement such as: “IF ((Color = “Grey”) OR (Color = “Pink”)) AND (Waterproof=“True”).”  Notice that you can use parentheses to group your conditions together.

Whether you are doing a search for new shoes or applying this logic to your database queries, Boolean logic lets you create multiple conditions to filter your results. And now that you know a little more about how Boolean logic is used, you can start using it!

## Additional Reading/Resources

* Learn about who pioneered Boolean logic in this historical article: [Origins of Boolean Algebra in the Logic of Classes](https://www.maa.org/press/periodicals/convergence/origins-of-boolean-algebra-in-the-logic-of-classes-george-boole-john-venn-and-c-s-peirce).
* F​ind more information about using AND, OR, and NOT from these [tips for searching with Boolean operators](https://libguides.mit.edu/c.php?g=175963&p=1158594).

## Hands-On Activity: Applying a function

**Total points**2

### 1.

**Question 1**



## Activity overview



In previous lessons, you got familiar with spreadsheets and data structures. In this activity, you will write functions in spreadsheets.

As a reminder, a **function** is a preset command that automatically performs a specified process or task using the data in a spreadsheet.

By the time you complete this activity, you will be able to apply the SUM function in spreadsheet software such as Google Sheets and Microsoft Excel. This will enable you to create dynamic spreadsheets, which are important for organizing and understanding data in your career as a data analyst.



### **What you will need**

To get started, first access the example spreadsheet of someone tracking their entertainment expenses.

To use the spreadsheet for this course item, click the link below and select Use Template.

Link to example spreadsheet: [Entertainment Expenses](https://docs.google.com/spreadsheets/d/1IJbiEEA2XIhkPctssmblodZkPWLzUgxzmjh0s1P6pv8/template/preview)

OR

If you don’t have a Google account, you can download the spreadsheet directly from the attachment below.

**Example Spreadsheet - Entertainment Expenses - Sheet1**CSV File

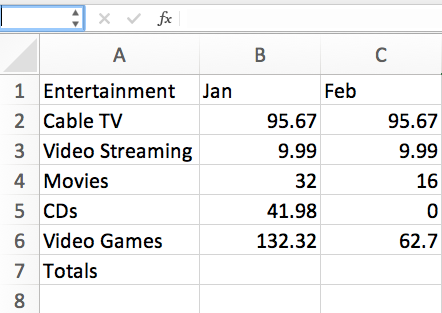
[Download file](https://d3c33hcgiwev3.cloudfront.net/xA1VTx3eQrqNVU8d3kK6gA_d43b9c83af434f3d849ce2c1cef3e995_Example-Spreadsheet---Entertainment-Expenses---Sheet1.csv?Expires=1648080000&Signature=j1O6s4Gv-86mPuEbcq-IDW895fP~TmpIop44goU6YDG9UV--1PdxJ75MufDtSQ3YJjtvq7A6A3XxPgX6AScqUKg3KKKvDApCLGDYD5K4PPDdXV5IRP7KdZUUvtCmOX19A5FOW6eOrychjVaxMxqeRp8hFs3SBg35RHrTjaqswkU_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



## Apply the SUM function



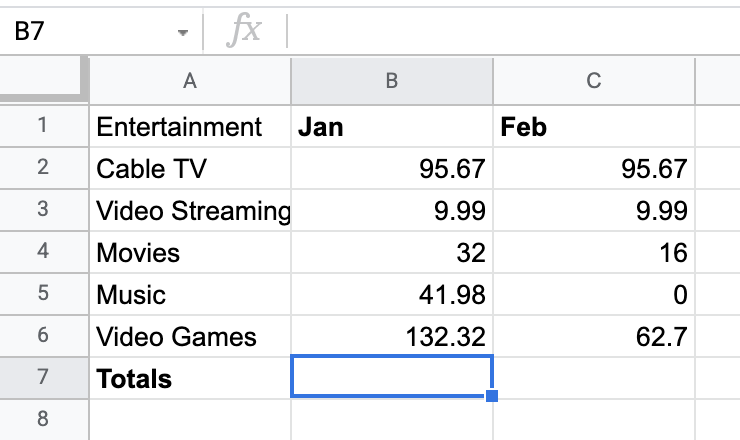
First, open the example spreadsheet. You will find the table below that contains data on the monthly entertainment expenses for January and February:

Column A: A1 - Entertainment A2 - Cable TV A3 - Video streaming A4 - Movies A5 - Music A6 - video games A7 - Totals Column B: B1 - Jan B2 - 95.67 B3 - 9.99 B4 - 32 B5 - 41.98 B6 - 132.32 Column C: C1 - Feb C2 - 95.67 C3 - 9.99 C4 - 16 C5 - 0 C6 - 62.7

Now, working in the example spreadsheet, you will write functions in cells to carry out certain tasks.

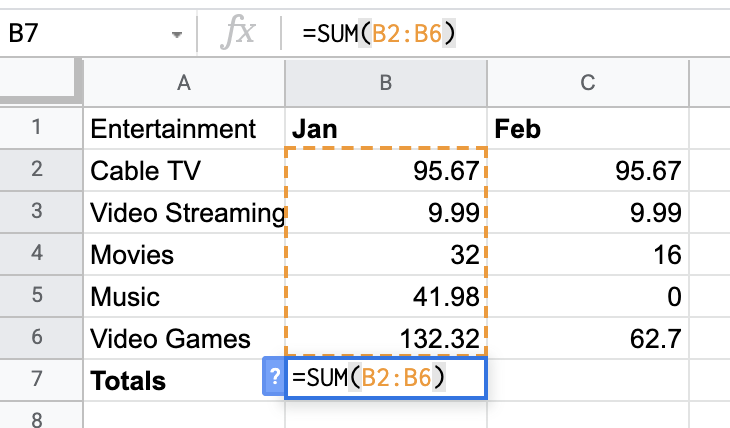
Your first goal is to fill in the cells B7 and C7. Each of these cells is supposed to be the sum of the numbers in the cells above it. For example, B7 should be the sum total of the numbers in the cells B2 to B6. To achieve this result:

1. Click on cell **B7**. The cell should have its border highlighted.

Column A: A1 - Entertainment A2 - Cable TV A3 - Video streaming A4 - Movies A5 - Music A6 - video games A7 - Totals Column B: B1 - Jan B2 - 95.67 B3 - 9.99 B4 - 32 B5 - 41.98 B6 - 132.32 Column C: C1 - Feb C2 - 95.67 C3 - 9.99 C4 - 16 C5 - 0 C6 - 62.7

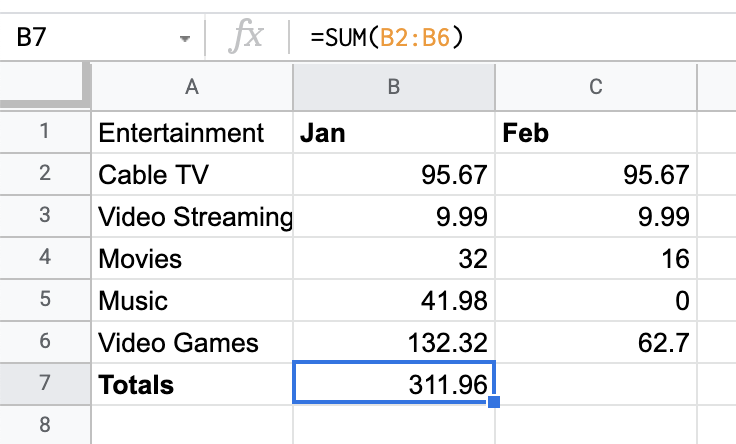
2. With that cell selected, type **=SUM(B2:B6)** like in the figure below.

Notice that this function both shows up in the cell and the field above the table. This field is called the **formula bar**. Once you’ve clicked on a cell, typing in the formula bar is the same thing as typing directly into the cell.

Column A: A1 - Entertainment A2 - Cable TV A3 - Video streaming A4 - Movies A5 - Music A6 - video games A7 - Totals Column B: B1 - Jan B2 - 95.67 B3 - 9.99 B4 - 32 B5 - 41.98 B6 - 132.32 B7: =SUM(B2:B6) Column C: C1 - Feb C2 - 95.67 C3 - 9.99 C4 - 16 C5 - 0 C6 - 62.7

The argument of the SUM function is the expression B2:B6. This expression represents a range of values starting from the first cell in the range (B2) to the last cell in the range (B6).  The word SUM instructs the spreadsheet to add up the values in that range of cells. This works similarly if you wish to add across the rows instead.

3. Press **Enter (Windows) or Return (Mac OS)**. The result below is what you should get.

Column A: A1 - Entertainment A2 - Cable TV A3 - Video streaming A4 - Movies A5 - Music A6 - video games A7 - Totals Column B: B1 - Jan B2 - 95.67 B3 - 9.99 B4 - 32 B5 - 41.98 B6 - 132.32 B7 - 311.96 Column C: C1 - Feb C2 - 95.67 C3 - 9.99 C4 - 16 C5 - 0 C6 - 62.7

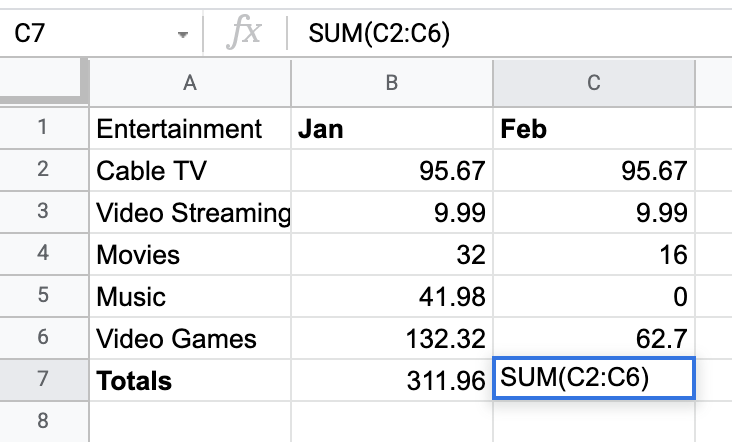
You will find that the SUM function in B7 is replaced by the numerical value (311.96) that is the sum of the numbers in cells B2 through B6. If the value in cell B7 is not equal to 311.96, check the function to ensure you have the correct range. The formula bar, however, still contains the SUM function. This is to inform people reading the spreadsheet how the value in cell B7 was determined.

## Find errors in functions



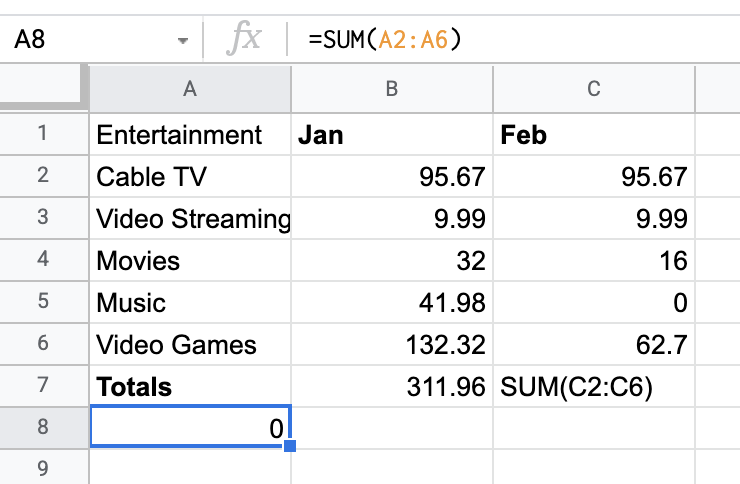
Syntax is very important for making proper functions in spreadsheets. Next, you can explore what happens when you leave out a character or make an error.

1. Click on cell **C7**. Enter the SUM function **SUM(C2:C6)** either in the cell itself or the formula bar. Do NOT include the **=**. Press **Enter or Return**. It should display the following:

Column A: A1 - Entertainment A2 - Cable TV A3 - Video streaming A4 - Movies A5 - Music A6 - video games A7 - Totals Column B: B1 - Jan B2 - 95.67 B3 - 9.99 B4 - 32 B5 - 41.98 B6 - 132.32 Column C: C1 - Feb C2 - 95.67 C3 - 9.99 C4 - 16 C5 - 0 C6 - 62.7 C7 - SUM(C2:C6)

The equal sign in the SUM command is not optional. Without it, the spreadsheet will interpret the input as a string. A string is text data. For the formula to work, it needs numeric data. This is why the command is uncalculated in C7. When the equal sign is included, the spreadsheet knows to carry out the sum calculation and return the result in the cell.

2. Spreadsheets handle string data quite differently than numerical data. Column A of this table is populated entirely by string data—the labels for each row. Try to input the SUM function on this column. In cell **A8**, type **=SUM(A2:A6)** and press **Enter or Return**.

Column A: A1 - Entertainment A2 - Cable TV A3 - Video streaming A4 - Movies A5 - Music A6 - video games A7 - Totals A8 - 0 Column B: B1 - Jan B2 - 95.67 B3 - 9.99 B4 - 32 B5 - 41.98 B6 - 132.32 B7 - 311.96 Column C: C1 - Feb C2 - 95.67 C3 - 9.99 C4 - 16 C5 - 0 C6 - 62.7 C7 - SUM(C2:C6)

You will find the spreadsheet calculated zero for the sum. This is because the program was asked to sum strings. When a given cell contains a string, the program considers the numerical value of the cell as zero.

That's how the SUM function in Excel works. There are many other functions available to you beyond SUM. If you know them, you can enter them just like how you entered the SUM function. There are many different spreadsheet programs, and they all have functionality similar to, if not exactly, like this.

## Confirmation and reflection



How would you write a function to calculate February’s entertainment expenses for Cable TV, Video Streaming, and Movies in the example spreadsheet?

**1 point**



SUM(C2:C6)



SUM(B2:C6)



=SUM(B2:C4)



=SUM(C2:C4)

### 2.

**Question 2**

During this activity, you explored spreadsheet functions and practiced writing them. In the text box below, write a 2-3 sentence (40-60 words) response to each of the following questions:

* When you wrote incorrect functions, what did you learn about spreadsheet data?
* How do you think this knowledge of the SUM spreadsheet function will help you write other kinds of functions?

Transforming data

**What is data transformation?**



A woman presenting data, a hand holding a medal, two people chatting, a ship's wheel being steered, two people high-fiving each other

In this reading, you will explore how data is transformed and the differences between wide and long data. **Data transformation** is the process of changing the data’s format, structure, or values. As a data analyst, there is a good chance you will need to transform data at some point to make it easier for you to analyze it.

Data transformation usually involves:

* Adding, copying, or replicating data
* Deleting fields or records
* Standardizing the names of variables
* Renaming, moving, or combining columns in a database
* Joining one set of data with another
* Saving a file in a different format. For example, saving a spreadsheet as a comma separated values (CSV) file.

**Why transform data?**

Goals for data transformation might be:

* Data **organization**: better organized data is easier to use
* Data **compatibility**: different applications or systems can then use the same data
* Data **migration**: data with matching formats can be moved from one system to another
* Data **merging**: data with the same organization can be merged together
* Data **enhancement**: data can be displayed with more detailed fields
* Data **comparison**: apples-to-apples comparisons of the data can then be made

**Data transformation example: data merging**

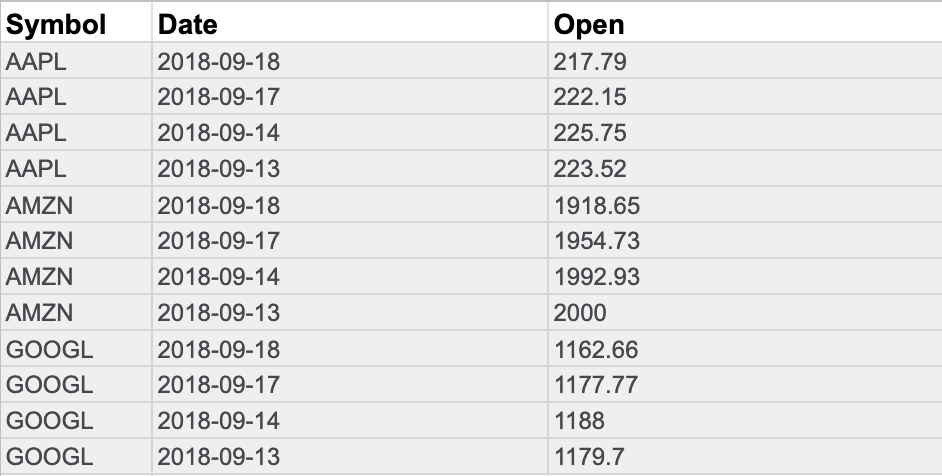
Mario is a plumber who owns a plumbing company. After years in the business, he buys another plumbing company. Mario wants to merge the customer information from his newly acquired company with his own, but the other company uses a different database. So, Mario needs to make the data compatible. To do this, he has to transform the format of the acquired company’s data. Then, he must remove duplicate rows for customers they had in common. When the data is compatible and together, Mario’s plumbing company will have a complete and merged customer database.

**Data transformation example: data organization (long to wide)**

To make it easier to create charts, you may also need to transform long data to wide data. Consider the following example of transforming stock prices (collected as long data) to wide data.

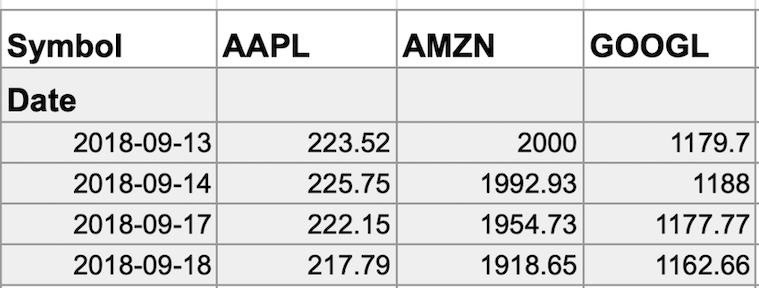
**Long data** is data where **each row contains a single data point** for a particular item. In the long data example below, individual stock prices (data points) have been collected for Apple (AAPL), Amazon (AMZN), and Google (GOOGL) (particular items) on the given dates.

**Long data example: Stock prices**



**Wide data** is data where **each row contains multiple data points** for the particular items identified in the columns.

**Wide data example: Stock prices**



With data transformed to wide data, you can create a chart comparing how each company's stock changed over the same period of time.

| **Wide data is preferred when** | **Long data is preferred when** |
| --- | --- |
| Creating tables and charts with a few variables about each subject | Storing a lot of variables about each subject. For example, 60 years worth of interest rates for each bank |
| Comparing straightforward line graphs | Performing advanced statistical analysis or graphing |

You might notice that all the data included in the long format is also in the wide format. But wide data is easier to read and understand. That is why data analysts typically transform long data to wide data more often than they transform wide data to long data. The following table summarizes when each format is preferred:

Data anonymization

**What is data anonymization?**

You have been learning about the importance of privacy in data analytics. Now, it is time to talk about **data anonymization** and what types of data should be anonymized. **Personally identifiable information**, or **PII**, is information that can be used by itself or with other data to track down a person's identity.

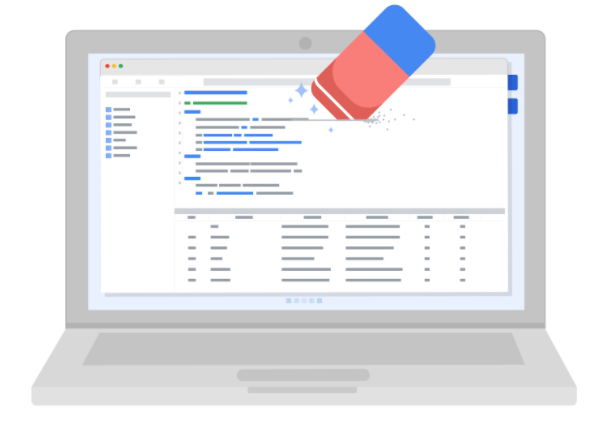
Data anonymization is the process of protecting people's private or sensitive data by eliminating that kind of information. Typically, data anonymization involves blanking, hashing, or masking personal information, often by using fixed-length codes to represent data columns, or hiding data with altered values.

**Your role in data anonymization**

Organizations have a responsibility to protect their data and the personal information that data might contain. As a data analyst, you might be expected to understand what data needs to be anonymized, but you generally wouldn't be responsible for the data anonymization itself. A rare exception might be if you work with a copy of the data for testing or development purposes. In this case, you could be required to anonymize the data before you work with it.

**What types of data should be anonymized?**

Healthcare and financial data are two of the most sensitive types of data. These industries rely a lot on data anonymization techniques. After all, the stakes are very high. That’s why data in these two industries usually goes through **de-identification**, which is **a process used to wipe data clean of all personally identifying information**.



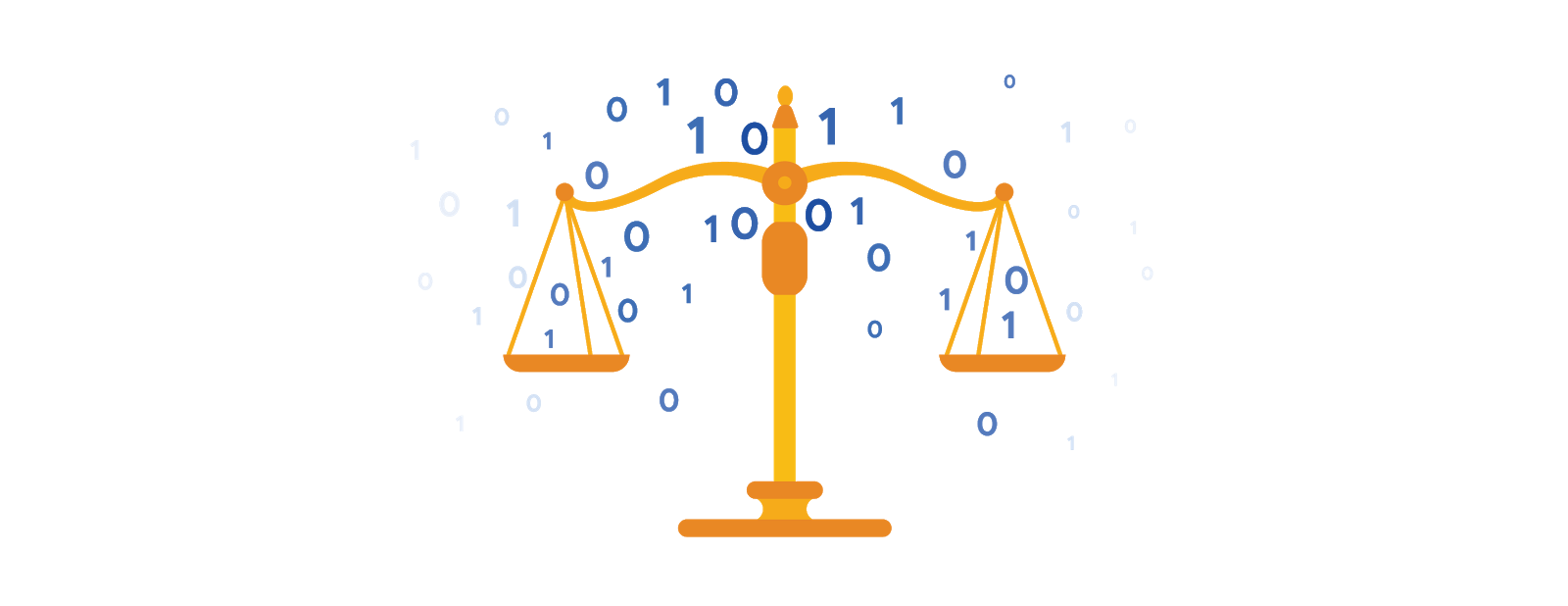
Data anonymization is used in just about every industry. That is why it is so important for data analysts to understand the basics. Here is a list of data that is often anonymized:

* Telephone numbers
* Names
* License plates and license numbers
* Social security numbers
* IP addresses
* Medical records
* Email addresses
* Photographs
* Account numbers

For some people, it just makes sense that this type of data should be anonymized. For others, we have to be very specific about what needs to be anonymized. Imagine a world where we all had access to each other’s addresses, account numbers, and other identifiable information. That would invade a lot of people’s privacy and make the world less safe. Data anonymization is one of the ways we can keep data private and secure!

The open-data debate

Just like data privacy, open data is a widely debated topic in today’s world. Data analysts think a lot about open data, and as a future data analyst, you need to understand the basics to be successful in your new role.



**What is open data?**

In data analytics, **open data** is part of **data ethics,** which has to do with using data ethically. **Openness** refers to free access, usage, and sharing of data. But for data to be considered open, it has to:

* Be available and accessible to the public as a complete dataset
* Be provided under terms that allow it to be reused and redistributed
* Allow universal participation so that anyone can use, reuse, and redistribute the data

Data can only be considered open when it meets all three of these standards.

**The open data debate: What data should be publicly available?**

One of the biggest benefits of open data is that credible databases can be used more widely. Basically, this means that all of that good data can be leveraged, shared, and combined with other data. This could have a huge impact on scientific collaboration, research advances, analytical capacity, and decision-making. But it is important to think about the individuals being represented by the public, open data, too.

**Third-party data** is collected by an entity that doesn’t have a direct relationship with the data. You might remember learning about this type of data earlier. For example, third parties might collect information about visitors to a certain website. Doing this lets these third parties create audience profiles, which helps them better understand user behavior and target them with more effective advertising.

**Personal identifiable information (PII)** is data that is reasonably likely to identify a person and make information known about them. It is important to keep this data safe***.*** PII can include a person’s address, credit card information, social security number, medical records, and more.

Everyone wants to keep personal information about themselves private. Because third-party data is readily available, it is important to balance the openness of data with the privacy of individuals.

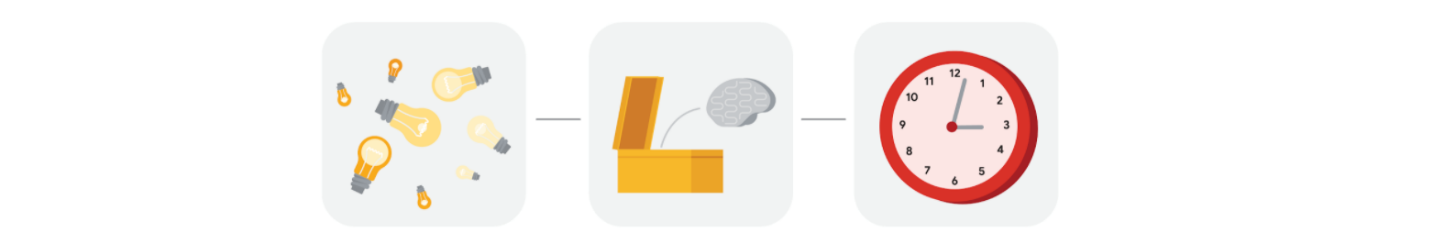
# Sites and resources for open data

Luckily for data analysts, there are lots of trustworthy sites and resources available for open data. It is important to remember that even reputable data needs to be constantly evaluated, but these websites are a useful starting point:

1. [**U.S. government data site**](https://www.data.gov/): Data.gov is one of the most comprehensive data sources in the US. This resource gives users the data and tools that they need to do research, and even helps them develop web and mobile applications and design data visualizations.
2. [**U.S. Census Bureau**](https://www.census.gov/data.html): This open data source offers demographic information from federal, state, and local governments, and commercial entities in the U.S. too.
3. [**Open Data Network**](https://www.opendatanetwork.com/): This data source has a really powerful search engine and advanced filters. Here, you can find data on topics like finance, public safety, infrastructure, and housing and development.
4. [**Google Cloud Public Datasets**](https://cloud.google.com/public-datasets): There are a selection of public datasets available through the Google Cloud Public Dataset Program that you can find already loaded into BigQuery.
5. [**Dataset Search**](https://datasetsearch.research.google.com/)**:** The Dataset Search is a search engine designed specifically for data sets; you can use this to search for specific data sets.

Databases in data analytics

Databases enable analysts to manipulate, store, and process data. This helps them search through data a lot more efficiently to get the best insights.



**Relational databases**

A **relational database** is a database that contains a series of tables that can be connected to show relationships. Basically, they allow data analysts to organize and link data based on what the data has in common.

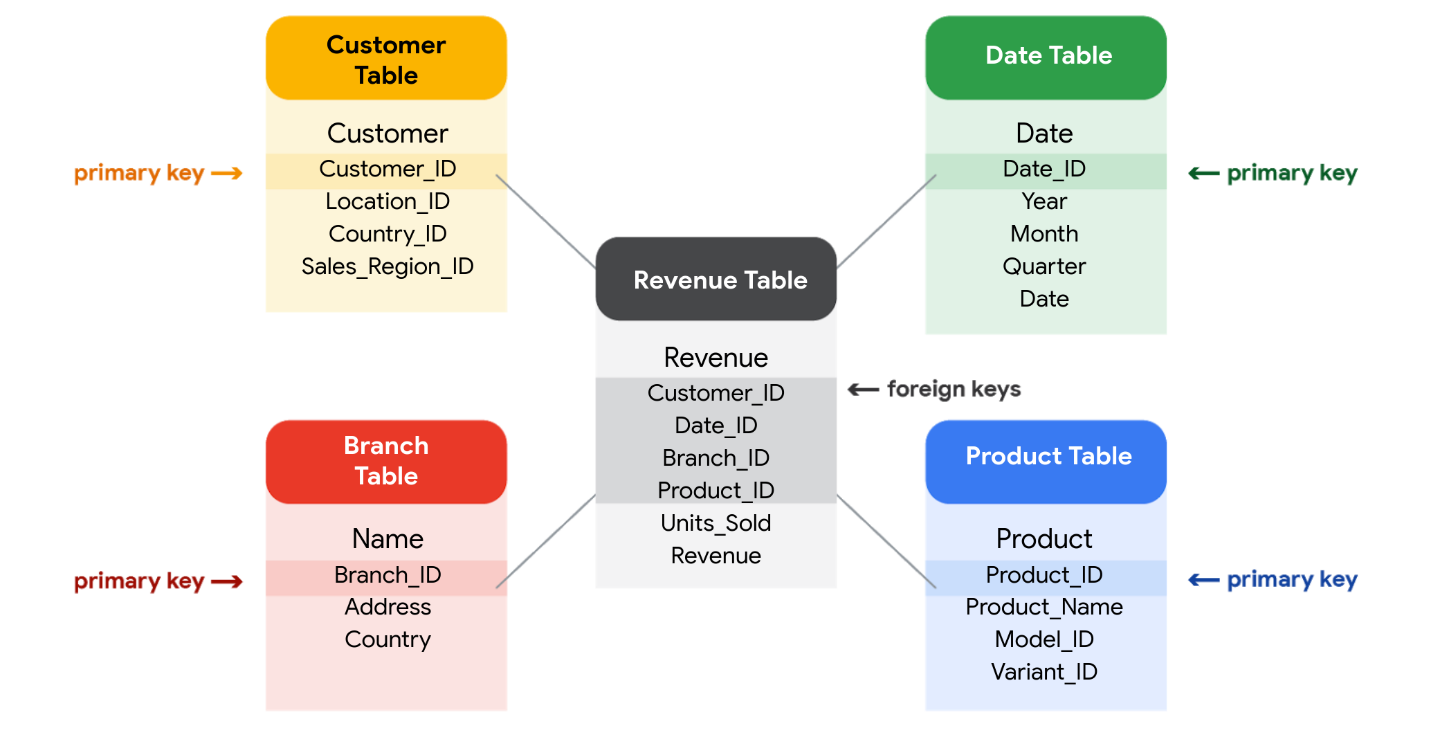
In a non-relational table, you will find all of the possible variables you might be interested in analyzing all grouped together. This can make it really hard to sort through. This is one reason why relational databases are so common in data analysis: they simplify a lot of analysis processes and make data easier to find and use across an entire database.

**The key to relational databases**

Tables in a relational database are connected by the fields they have in common. You might remember learning about primary and foreign keys before. As a quick refresher, a **primary key** is an identifier that references a column in which each value is unique. In other words, it's a column of a table that is used to uniquely identify each record within that table. The value assigned to the primary key in a particular row must be unique within the entire table. For example, if customer\_id is the primary key for the customer table, no two customers will ever have the same customer\_id.

By contrast, a **foreign key** is a field within a table that is a primary key in another table. A table can have only one primary key, but it can have multiple foreign keys. These keys are what create the relationships between tables in a relational database, which helps organize and connect data across multiple tables in the database.

Some tables don't require a primary key. For example, a revenue table can have multiple foreign keys and not have a primary key. A primary key may also be constructed using multiple columns of a table. This type of primary key is called a **composite key**. For example, if customer\_id and location\_id are two columns of a composite key for a customer table, the values assigned to those fields in any given row must be unique within the entire table.



**SQL? You’re speaking my language**

Databases use a special language to communicate called a query language. **Structured Query Language** (SQL) is a type of query language that lets data analysts communicate with a database. So, a data analyst will use SQL to create a query to view the specific data that they want from within the larger set. In a relational database, data analysts can write queries to get data from the related tables. SQL is a powerful tool for working with databases — which is why you are going to learn more about it coming up!

# Inspecting a dataset: A guided, hands-on tour

As a data analyst, you'll use data to answer questions and solve problems. When you analyze data and draw conclusions, you are generating insights that can influence business decisions, drive positive change, and help your stakeholders meet their goals.

Before you begin an analysis, it’s important to inspect your data to determine if it contains the specific information you need to answer your stakeholders’ questions. In any given dataset, it may be the case that:

* The data is not there (you have sandwich data, but you need pizza data)
* The data is insufficient (you have pizza data for June 1-7, but you need data for the entire month of June)
* The data is incorrect (your pizza data lists the cost of a slice as $250, which makes you question the validity of the dataset)

Inspecting your dataset will help you pinpoint what questions are answerable and what data is still missing. You may be able to recover this data from an external source or at least recommend to your stakeholders that another data source be used.

In this reading, imagine you’re a data analyst inspecting spreadsheet data to determine if it’s possible to answer your stakeholders’ questions.

## The scenario

You are a data analyst working for an ice cream company. Management is interested in improving the company's ice cream sales.

The company has been collecting data about its sales—but not a lot. The available data is from an internal data source and is based on sales for 2019. You’ve been asked to review the data and provide some insight into the company’s ice cream sales. Ideally, management would like answers to the following questions:

1. What is the most popular flavor of ice cream?
2. How does temperature affect sales?
3. How do weekends and holidays affect sales?
4. How does profitability differ for new versus returning customers?

## Download the data

You can download the data to follow along with this reading. To use the template for the sales data, click the link below and select “Use Template.”

Link to template: [Ice Cream Sales](https://docs.google.com/spreadsheets/d/1NgiKb8wCnJbUTuUkDUiNRpx9NhwncEmoKuPvgfYfOIY/template/preview?resourcekey=0-X3e7NzehG2Y74MIBhOaqeQ#gid=653912415)

OR

If you don’t have a Google account, you can download the spreadsheets directly from the attachments below:

**SalesByTemp**XLSX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/jmigEulNR7yooBLpTYe8Cw_9ecaf818f1a74b7987fe6a7d9af3c1f1_SalesByTemp.xlsx?Expires=1648080000&Signature=FVL35o2eM42gwoyW2ZLqCwcKWXrlLEK0VxS0fVZUZHsWGUWNp3dka4zxIe7vpokrrBOw133c9Yqk050Nc-Bl82gQEJDdrEbk04q~vu2-c1yXgrXS3eLBa92gNJYnUkIcxVc13oeeM87MTu57ff2XLET3yQfq-uaPBxBQbr-2Xz4_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**SalesByDay**XLSX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/B3ofmLtERPq6H5i7RFT6Pg_1ca5eec9c08941518e2c16034a2e65f1_SalesByDay.xlsx?Expires=1648080000&Signature=O4bRojnn8GFSfsrEkwFUK2CF-HiwolTCaTCltnTwAaYspyhbUDf5YbqrPH43pj3p3rSdMb~R4wcU~SDpm5Id3xOAjJ9QHQC2jiQqxgcMYLj99lnywQVAncwH4fbnMtxFpBVAc5bbn~QX7mRj~ATAPWMjyX8jpHfJKmBslBHDSk0_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**SalesByFlavor**XLSX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/DHN9hYWCSDCzfYWFgvgwgg_b0e0d35f6a4f4bde9c84ecd0dd69c0f1_SalesByFlavor.xlsx?Expires=1648080000&Signature=GmmsQqEVSX6VgRzT8whmjQJnsrN3BIrvl33O93OV8VDLayaveQaaUtTOEEOTN8DmgwdnfDTXxIVYz1WEEqv1nSge8KPA-THCnP5xmoG5Y1ZS~F3U5HJgG7G4jm8lsPsWwVL~57mIRfFsxAxL~X5qNnWXQluQXScSGlnCeJLzgcE_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



## Inspect the data

### Question 1: What is the most popular flavor of ice cream?

To discover the most popular flavor, you first need to define what is meant by "popular." Is the most popular flavor the one that generated the most revenue in 2019? Or is it the flavor that had the largest number of units sold in 2019? Sometimes your measurement choices are limited by what data you have—you can review your spreadsheet to find out if either of these definitions of “popular” make sense based on the available data.

Click the **flavors** tab on your spreadsheet to view the relevant data. The **flavors** sheet has three columns and 209 rows of data. The column headers are **week**, **units sold**, and **flavor**. This dataset did not come with a data description, so you have to figure out the significance of the columns on your own. Based on the data, you deduce that these columns provide information about the number of units sold for each ice cream flavor, by week, in 2019

Graphical user interface, application, table, Excel

Description automatically generated

In this case, you can discover what the most popular flavor is by using units sold as your measure. In particular, you can use the **units sold** column to calculate the total number of units sold during the year for each flavor. Unfortunately, the dataset does not provide the annual sales amount by flavor. In this case, your next step would be to ask your stakeholders if the annual sales per flavor data is available from another source. If not, you can add a statement about the current data’s limitations to your analysis.

### Question 2: How does temperature affect sales?

To explore your second question, you click the **temperatures** tab and check out the data. The **temperatures** sheet has two columns and 366 rows of data. The column headers are **temperature** and **sales**. The data may show total 2019 sales per temperature (for instance, the first entry might sum up $39.69 in sales for three separate days that each had a high of 60 degrees). Or, the data may show  a snapshot of sales and temperature for each day in 2019 (for instance, the first entry might refer to a single day with a high of 60 degrees and $39.69 in sales).

Table

Description automatically generated

So, which is it? It’s probably a daily snapshot because there are 365 entries for temperature, and multiple rows with the same temperature and different sales values. This implies that each entry is for a single day and not a summary of multiple days. However, without more information, you can’t be certain. Plus, you don’t know if the current data is listed in consecutive order by date or in a different order. Your next step would be to contact the owner of the dataset for clarification.

If it turns out that temperature does affect sales, you’ll be able to offer your stakeholders an insight such as the following: “When daily highs are above X degrees, average ice cream sales increase by Y amount. So the business should plan on increasing inventory during these times to maximize sales.”

### Question 3: How do weekends and holidays affect sales?

Next, you click on the **sales** tab to view the data about dates of sale. The **sales** sheet has two columns and 366 rows of data. The column headers are **date** and **sales**. This data is most likely total daily sales in 2019, as sales are recorded for each date in 2019.

Graphical user interface, application, table, Excel

Description automatically generated

You can use this data to determine whether a specific date falls on a weekend or holiday and add a column to your sheet that reflects this information. Then, you can find out whether sales on the weekends and holidays are greater than sales on other days. This will be useful to know for inventory planning and marketing purposes.

### Question 4: How does profitability differ for new customers versus returning customers?

Your dataset does not contain sales data related to new customers. Without this data, you won’t be able to answer your final question. However, it may be the case that the company collects customer data and stores it in a different data table.

If so, your next step would be to find out how to access the company’s customer data. You can then join the revenue sales data to the customer data table to categorize each sale as from a new or returning customer and analyze the difference in profitability between the two sets of customers. This information will help your stakeholders develop marketing campaigns for specific types of customers to increase brand loyalty and overall profitability.

## Conclusion

When working on analytics projects, you won’t always have all the necessary or relevant data at your disposal.  In many of these cases, you can turn to other data sources to fill in the gaps.

 Despite the limitations of your dataset, it’s still possible to offer your stakeholders some valuable insights. For next steps, your best plan of action will be to take the initiative to ask questions, identify other relevant datasets, or do some research on your own.  No matter what data you’re working with, carefully inspecting your data makes a big impact on the overall quality of your analysis.

# Metadata is as important as the data itself

Data analytics, by design, is a field that thrives on collecting and organizing data. In this reading, you are going to learn about how to analyze and thoroughly understand every aspect of your data.

A picture containing text, toy

Description automatically generated

Take a look at any data you find. What is it? Where did it come from? Is it useful? How do you know? This is where metadata comes in to provide a deeper understanding of the data. To put it simply, **metadata** is data about data. In database management, it provides information about other data and helps data analysts interpret the contents of the data within a database.

Regardless of whether you are working with a large or small quantity of data, metadata is the mark of a knowledgeable analytics team, helping to communicate about data across the business and making it easier to reuse data. In essence, metadata tells the who, what, when, where, which, how, and why of data.

## Elements of metadata

Before looking at metadata examples, it is important to understand what type of information metadata typically provides.

### Title and description

What is the name of the file or website you are examining? What type of content does it contain?

### Tags and categories

What is the general overview of the data that you have? Is the data indexed or described in a specific way?

### Who created it and when

Where did the data come from, and when was it created? Is it recent, or has it existed for a long time?

### Who last modified it and when

Were any changes made to the data?  If yes, were the modifications recent?

### Who can access or update it

Is this dataset public? Are special permissions needed to customize or modify the dataset?

## Examples of metadata

In today’s digital world, metadata is everywhere, and it is becoming a more common practice to provide metadata on a lot of media and information you interact with. Here are some real-world examples of where to find metadata:

### **Photos**

Whenever a photo is captured with a camera, metadata such as camera filename, date, time, and geolocation are gathered and saved with it.

### **Emails**

When an email is sent or received, there is lots of visible metadata such as subject line, the sender, the recipient and date and time sent. There is also hidden metadata that includes server names, IP addresses, HTML format, and software details.

### **Spreadsheets and documents**

Spreadsheets and documents are already filled with a considerable amount of data so it is no surprise that metadata would also accompany them. Titles, author, creation date, number of pages, user comments as well as names of tabs, tables, and columns are all metadata that one can find in spreadsheets and documents.

### **Websites**

Every web page has a number of standard metadata fields, such as tags and categories, site creator’s name, web page title and description, time of creation and any iconography.

### **Digital files**

Usually, if you right click on any computer file, you will see its metadata. This could consist of file name, file size, date of creation and modification, and type of file.

### **Books**

Metadata is not only digital. Every book has a number of standard metadata on the covers and inside that will inform you of its title, author’s name, a table of contents, publisher information, copyright description, index, and a brief description of the book’s contents.

## Data as you know it

Knowing the content and context of your data, as well as how it is structured, is very valuable in your career as a data analyst. When analyzing data, it is important to always understand the full picture. It is not just about the data you are viewing, but how that data comes together. Metadata ensures that you are able to find, use, preserve, and reuse data in the future. Remember, it will be your responsibility to manage and make use of data in its entirety; metadata is as important as the data itself.

# Exploring public datasets

**Open data** helps create a lot of **public datasets** that you can access to make data-driven decisions. Here are some resources you can use to start searching for public datasets on your own:

* The [Google Cloud Public Datasets](https://cloud.google.com/public-datasets) allow data analysts access to high-demand public datasets, and make it easy to uncover insights in the cloud.
* The [Dataset Search](https://datasetsearch.research.google.com/) can help you find available datasets online with keyword searches.
* [Kaggle](https://www.kaggle.com/datasets?utm_medium=paid&utm_source=google.com+search&utm_campaign=datasets&gclid=CjwKCAiAt9z-BRBCEiwA_bWv-L6PpACh6RzmrJjQjmNGCCE7kky1FCtc6Jf1qld-4NwDMYL0WsUyxBoCdwAQAvD_BwE) has an Open Data search function that can help you find datasets to practice with.
* Finally, [BigQuery](https://cloud.google.com/bigquery/public-data) hosts 150+ public datasets you can access and use.

### **Public health datasets**

1. [Global Health Observatory data](https://www.who.int/data/collections): You can search for datasets from this page or explore featured data collections from the World Health Organization.
2. [The Cancer Imaging Archive (TCIA) dataset](https://cloud.google.com/healthcare/docs/resources/public-datasets/tcia): Just like the earlier dataset, this data is hosted by the Google Cloud Public Datasets and can be uploaded to BigQuery.
3. [1000 Genomes](https://cloud.google.com/life-sciences/docs/resources/public-datasets/1000-genomes): This is another dataset from the Google Cloud Public resources that can be uploaded to BigQuery.

### **Public climate datasets**

1. [National Climatic Data Center](https://www.ncdc.noaa.gov/data-access/quick-links): The NCDC Quick Links page has a selection of datasets you can explore.
2. [NOAA Public Dataset Gallery](https://www.climate.gov/maps-data/datasets): The NOAA Public Dataset Gallery contains a searchable collection of public datasets.

### **Public social-political datasets**

1. [UNICEF State of the World’s Children](https://data.unicef.org/resources/dataset/sowc-2019-statistical-tables/): This dataset from UNICEF includes a collection of tables that can be downloaded.
2. [CPS Labor Force Statistics](https://www.bls.gov/cps/tables.htm): This page contains links to several available datasets that you can explore.
3. [The Stanford Open Policing Project](https://openpolicing.stanford.edu/): This dataset can be downloaded as a .CSV file for your own use.

Using BigQuery

[BigQuery](https://cloud.google.com/bigquery/docs) is a data warehouse on Google Cloud that data analysts can use to query, filter large datasets, aggregate results, and perform complex operations.

An upcoming activity is performed in BigQuery. This reading provides instructions to create your own BigQuery account, select public datasets, and upload CSV files. At the end of this reading, you can confirm your access to the BigQuery console before you move on to the activity,

**Note:** Additional getting started resources for a few other SQL database platforms are also provided at the end of this reading if you choose to work with them instead of BigQuery.

**Types of BigQuery accounts**

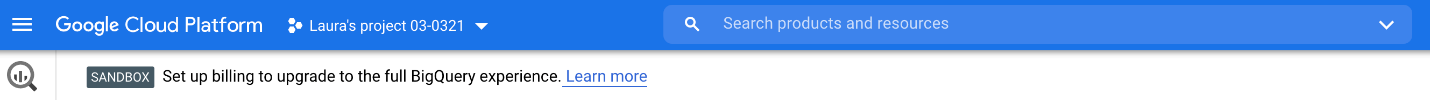
There are two different types of accounts: sandbox and free trial. A sandbox account allows you to practice queries and explore public datasets for free, but has additional [restrictions](https://cloud.google.com/bigquery/docs/sandbox#limits) on top of the [standard quotas and limits](https://cloud.google.com/bigquery/quotas). If you prefer to use BigQuery with the standard limits, you can set up a free trial account instead. More details:

* A **free** **sandbox account** doesn’t ask for a method of payment. It does, however, limit you to 12 projects. It also doesn't allow you to insert new records to a database or update the field values of existing records. These data manipulation language (DML) operations aren't supported in the sandbox.
* A **free trial account** requires a method of payment to establish a billable account, but offers full functionality during the trial period.

With either type of account, you can upgrade to a paid account at any time and retain all of your existing projects. If you set up a free trial account but choose not to upgrade to a paid account when your trial period ends, you can still set up a free sandbox account at that time. However, projects from your trial account won't transfer to your sandbox account. It would be like starting from scratch again.

**Set up a free sandbox account for use in this program**

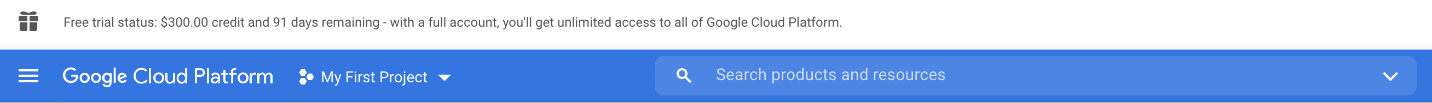
* Follow these [step-by-step instructions](https://cursive.io/shared/2da0e63f3-9de7-476f-997b-93fff70d7cb6) or watch the video, [Setting up BigQuery, including sandbox and billing options](https://www.coursera.org/learn/data-preparation/lecture/YCkys/setting-up-bigquery-including-sandbox-and-billing-options).
* For more detailed information about using the sandbox, start with the documentation, [Using the BigQuery sandbox](https://cloud.google.com/bigquery/docs/sandbox?hl=en_US).
* After you set up your account, you will see the project name you created for the account in the banner and **SANDBOX** at the top of your BigQuery console.



**Set up a free trial account instead (if you prefer)**

If you prefer not to have the sandbox limitations in BigQuery, you can set up a free trial account for use in this program.

* Follow these [step-by-step instructions](https://cursive.io/shared/2e98bf922-42d6-48c2-998f-6057389d0ccb) or watch the video, [Setting up BigQuery, including sandbox and billing options](https://www.coursera.org/learn/data-preparation/lecture/YCkys/setting-up-bigquery-including-sandbox-and-billing-options). The free trial offers $300 in credit over the next 90 days. You won’t get anywhere near that spending limit if you just use the BigQuery console to practice SQL queries. After you spend the $300 credit (or after 90 days) your free trial will expire and you will need to personally select to upgrade to a paid account to keep using Google Cloud Platform services, including BigQuery. **Your method of payment will never be automatically charged after your free trial ends.** If you select to upgrade your account, you will begin to be billed for charges.
* After you set up your account, you will see **My First Project** in the banner and the status of your account above the banner **–** your credit balance and the number of days remaining in your trial period.



**How to get to the BigQuery console**

In your browser, go to [console.cloud.google.com/bigquery](https://console.cloud.google.com/bigquery).

**Note:** Going to [console.cloud.google.com](https://console.cloud.google.com/) in your browser takes you to the main dashboard for the Google Cloud Platform. To navigate to BigQuery from the dashboard, do the following:

* Click the Navigation menu icon (Hamburger icon) in the banner.
* Scroll down to the **BIG DATA** section.
* Click **BigQuery** and select **SQL workspace**.

Watch the [How to use BigQuery](https://www.coursera.org/learn/data-preparation/lecture/YWn81/how-to-use-bigquery) video for an introduction to each part of the BigQuery SQL workspace.

**(Optional) Explore a BigQuery public dataset**

You will be exploring a public dataset in an upcoming activity, so you can perform these steps later if you prefer.

* Refer to these [step-by-step instructions](https://cursive.io/shared/242bde9a6-5415-4ce0-bbae-7e875d14d927).

**(Optional) Upload a CSV file to BigQuery**

These steps are provided so you can work with a dataset on your own at this time. You will upload CSV files to BigQuery later in the program.

* Refer to these [step-by-step instructions](https://cursive.io/shared/2dea0d610-ef6b-4ba8-8e44-d40dfeb0454b).

**Getting started with other databases (if not using BigQuery)**

It is easier to follow along with the course activities if you use BigQuery, but i​f you are connecting to and practicing SQL queries on other database platforms instead of BigQuery, here are similar getting started resources:

* [G​etting started with MySQL](https://dev.mysql.com/doc/mysql-getting-started/en/): This is a guide to setting up and using MySQL.
* [G​etting started with Microsoft SQL Server](https://docs.microsoft.com/en-us/sql/relational-databases/tutorial-getting-started-with-the-database-engine?view=sql-server-ver15): This is a tutorial to get started using SQL Server.
* [G​etting started with PostgreSQL](https://www.postgresql.org/docs/10/tutorial-start.html): This is a tutorial to get started using PostgreSQL.
* [Getting started with SQLite](https://www.sqlite.org/quickstart.html): This is a quick start guide for using SQLite.

In-depth guide: SQL best practices

*You can save this reading for future reference. Feel free to download a PDF version of this reading below:*

**DAC3 In-depth guide\_ SQL best practices.pdf**PDF File

[Open file](https://d3c33hcgiwev3.cloudfront.net/5vVDkB5qT1y1Q5Aeau9c_Q_6d0e31160e2e43479d172390d19853f1_DAC3-In-depth-guide_-SQL-best-practices.pdf?Expires=1648080000&Signature=djsbqs-v8Uxcneh5pIXczji2mp1UnH4SP3l1XgYGkY~ki~IzKZcEPb40JbSDJFb8RmbJyUJwZlWgMQO4nBq8C5f4w0lln1JSR82oZSTcBXMR9Uqw7cysquTeiAvzoNgZtZb2CBKEACOariL44cJXJ86HBdr-C5q7wkFjw7OMqu0_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

These best practices include guidelines for writing SQL queries, developing documentation, and examples that demonstrate these practices. This is a great resource to have handy when you are using SQL yourself; you can just go straight to the relevant section to review these practices. Think of it like a SQL field guide!

**Capitalization and case sensitivity**

With SQL, capitalization usually doesn’t matter. You could write SELECT or select or SeLeCT. They all work! But if you use capitalization as part of a consistent style your queries will  look more professional.

To write SQL queries like a pro, it is always a good idea to use all caps for clause starters (e.g., SELECT, FROM, WHERE, etc.). Functions should also be in all caps (e.g., SUM()). Column names should be all lowercase (refer to the section on snake\_case later in this guide). Table names should be in CamelCase (refer to the section on CamelCase later in this guide). This helps keep your queries consistent and easier to read while not impacting the data that will be pulled when you run them. The only time that capitalization does matter is when it is inside quotes (more on quotes below).

Vendors of SQL databases may use slightly different variations of SQL. These variations are called **SQL dialects**. Some SQL dialects are case sensitive. BigQuery is one of them. Vertica is another. But most, like MySQL, PostgreSQL, and SQL Server, aren’t case sensitive. This means if you searched for country\_code = ‘us’, it will return all entries that have 'us', 'uS', 'Us', and 'US'. This isn’t the case with BigQuery. BigQuery is case sensitive, so that same search would only return entries where the country\_code is exactly 'us'. If the country\_code is 'US', BigQuery wouldn’t return those entries as part of your result.

**Single or double quotes: '' or " "**

For the most part, it also doesn’t matter if you use single quotes ' ' or double quotes " " when referring to strings. For example, SELECT is a clause starter. If you put SELECT in quotes like 'SELECT' or "SELECT", then SQL will treat it as a text string. Your query will return an error because your query needs a SELECT clause.

But there are two situations where it does matter what kind of quotes you use:

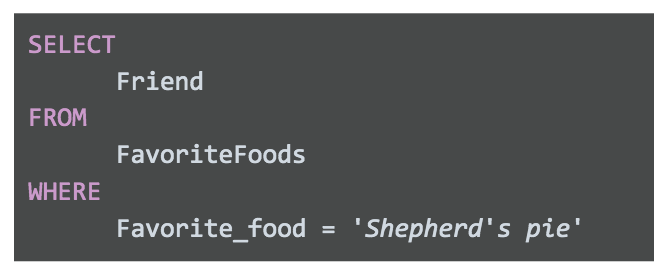
1. When you want strings to be identifiable in *any* SQL dialect
2. When your string contains an apostrophe or quotation marks

Within each SQL dialect there are rules for what is accepted and what isn’t. But a general rule across almost all SQL dialects is to use single quotes for strings. This helps get rid of a lot of confusion. So if we want to reference the country US in a WHERE clause (e.g., country\_code = 'US'), then use single quotes around the string 'US'.

The second situation is when your string has quotes inside it. Suppose you have a column of favorite foods in a table called FavoriteFoods and the other column corresponds to each friend.

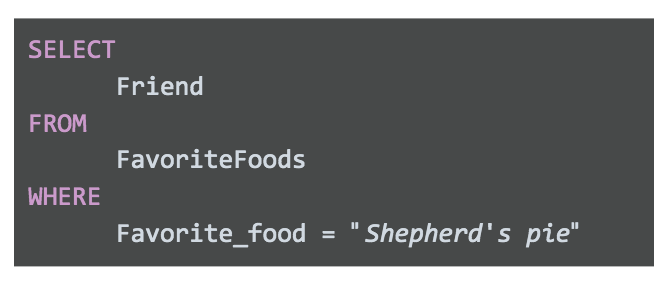
| **Friend** | **Favorite\_food** |
| --- | --- |
| Rachel DeSantos | Shepherd’s pie |
| Sujin Lee | Tacos |
| Najil Okoro | Spanish paella |

You might notice how Rachel’s favorite food contains an apostrophe. If you were to use single quotes in a WHERE clause to find the friend who has this favorite food, it would look like this:



**This won’t work.** If you run this query, you will get an error in return. This is because SQL recognizes a text string as something that starts with a quote 'and ends with another quote '. So in the bad query above,  SQL thinks that the Favorite\_food you are looking for is 'Shepherd'. Just 'Shepherd' because the apostrophe in Shepherd**'**s ends the string.

Generally speaking, this should be the only time you would use double quotes instead of single quotes. So your query would look like this instead:



SQL understands text strings as either starting with a single quote ' or double quote". Since this string starts with double quotes, SQL will expect another double quote to signal the end of the string. This keeps the apostrophe safe, so it will return "Shepherd's pie" and not 'Shepherd'.

**Comments as reminders**

As you get more comfortable with SQL, you will be able to read and understand queries at a glance. But it never hurts to have comments in the query to remind yourself of what you are trying to do. And if you share your query, it also helps others understand it.

For example:

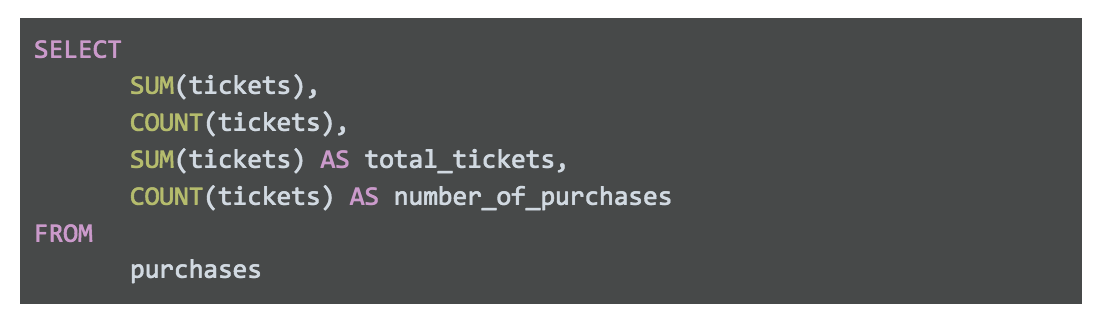
Text

Description automatically generated

You can use # in place of the two dashes, --, in the above query but keep in mind that # isn’t recognized in all SQL dialects (MySQL doesn’t recognize #). So it is best to use -- and be consistent with it. When you add a comment to a query using --, the database query engine will ignore everything in the same line after --. It will continue to process the query starting on the next line.

**Snake\_case names for columns**

It is important to always make sure that the output of your query has easy-to-understand names. If you create a new column (say from a calculation or from concatenating new fields), the new column will receive a generic default name (e.g., f0). For example:

The following table features the results of this query: f0: 8 f1: 4 total\_tickets: 8 Number\_of\_purchases: 4

Results are:

| **f0** | **f1** | **total\_tickets** | **number\_of\_purchases** |
| --- | --- | --- | --- |
| 8 | 4 | 8 | 4 |

The first two columns are named f0 and f1 because they weren’t named in the above query. SQL defaults to f0, f1, f2, f3, and so on. We named the last two columns total\_tickets and number\_of\_purchases so these column names show up in the query results. This is why it is always good to give your columns useful names, especially when using functions. After running your query, you want to be able to quickly understand your results, like the last two columns we described in the example.

On top of that, you might notice how the column names have an underscore between the words. Names should never have spaces in them. If 'total\_tickets' had a space and looked like 'total tickets' then SQL would rename SUM(tickets) as just 'total'. Because of the space, SQL will use 'total' as the name and won’t understand what you mean by 'tickets'. So, spaces are bad in SQL names. Never use spaces.

The best practice is to use snake\_case. This means that 'total tickets', which has a space between the two words, should be written as 'total\_tickets' with an underscore instead of a space.

**CamelCase names for tables**

You can also use CamelCase capitalization when naming your table. CamelCase capitalization means that you capitalize the start of each word, like a two-humped (Bactrian) camel. So the table TicketsByOccasion uses CamelCase capitalization. Please note that the capitalization of the first word in CamelCase is *optional;* camelCase is also used. Some people differentiate between the two styles by calling CamelCase,PascalCase, and reserving camelCase for when the first word isn't capitalized, like a one-humped (Dromedary) camel; for example, ticketsByOccasion.

At the end of the day, CamelCase is a style choice. There are other ways you can name your tables, including:

* All lower or upper case, like ticketsbyoccasion or TICKETSBYOCCASION
* With snake\_case,  like tickets\_by\_occasion

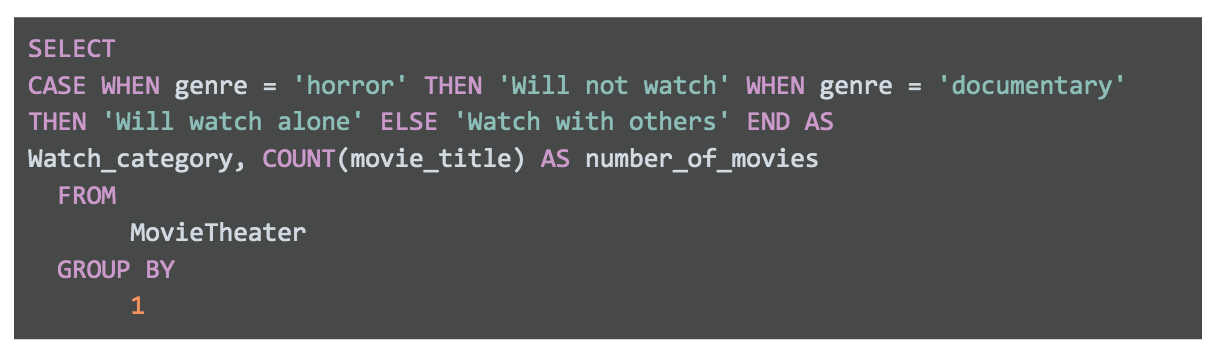
Keep in mind, the option with all lowercase or uppercase letters can make it difficult to read your table name, so it isn’t recommended for professional use.

The second option, snake\_case, is technically okay. With words separated by underscores, your table name is easy to read, but it can get very long because you are adding the underscores. It also takes more time to write. If you use this table a lot, it can become a chore.

In summary, it is up to you to use snake\_case or CamelCase when creating table names. Just make sure your table name is easy to read and consistent. Also be sure to find out if your company has a preferred way of naming their tables. If they do, always go with their naming convention for consistency.

**Indentation**

As a general rule, you want to keep the length of each line in a query <= 100 characters. This makes your queries easy to read. For example, check out this query with a line with >100 characters:

SELECT CASE WHEN genre = 'horror' THEN 'Will not watch' WHEN genre = 'documentary' THEN 'Will watch alone' ELSE 'Watch with others' END AS Watch\_category, COUNT(

This query is hard to read and just as hard to troubleshoot or edit. Now, here is a query where we stick to the <= 100 character rule:

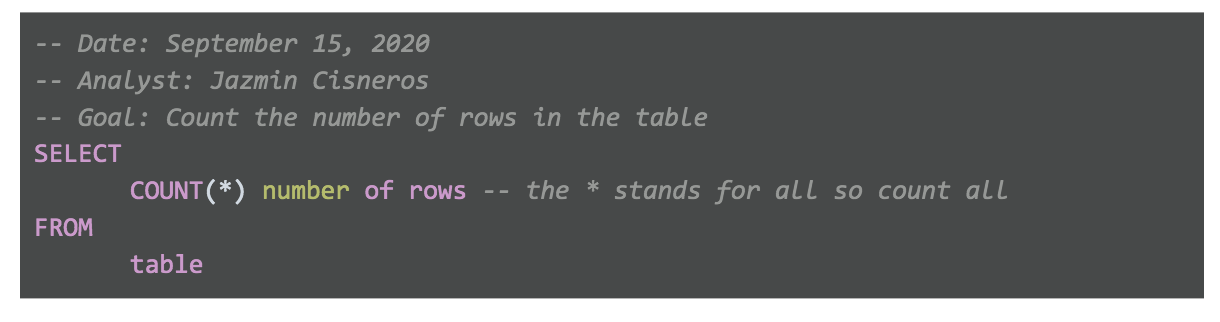
Text

Description automatically generated

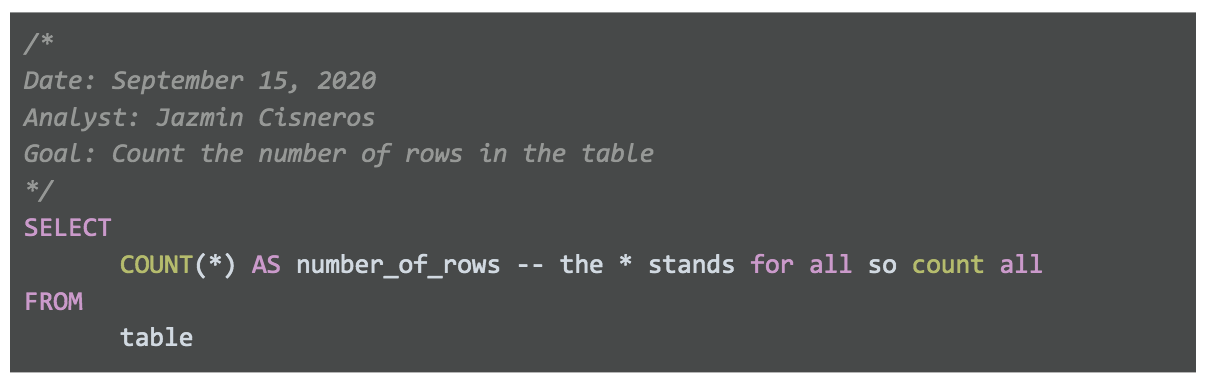
Now it is much easier to understand what you are trying to do in the SELECT clause. Sure, both queries will run without a problem because indentation doesn’t matter in SQL. But proper indentation is still important to keep lines short. And it will be valued by anyone reading your query, including yourself!

**Multi-line comments**

If you make comments that take up multiple lines, you can use -- for each line. Or, if you have more than two lines of comments, it might be cleaner and easier is to use /\* to start the comment and \*/ to close the comment. For example, you can use the -- method like below:

-- Date: September 15, 2020 -- Analyst: Jazmin Cisneros -- Goal: Count the number of rows in the table SELECT COUNT(\*) number of rows -- the \* stands for all so count all FROM table

Or, you can use the /\* \*/ method like below:

/\* Date: September 15, 2020 Analyst: Jazmin Cisneros Goal: Count the number of rows in the table \*/ SELECT COUNT(\*) AS number\_of\_rows -- the \* stands for all so count all FROM table

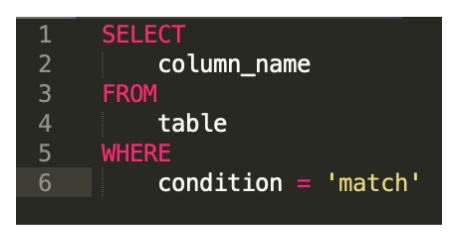
In SQL, it doesn’t matter which method you use. SQL ignores comments regardless of what you use: #, --, or /\* and \*/. So it is up to you and your personal preference. The /\* and  \*/ method for multi-line comments usually looks cleaner and helps separate the comments from the query. But there isn’t one right or wrong method.

**SQL text editors**

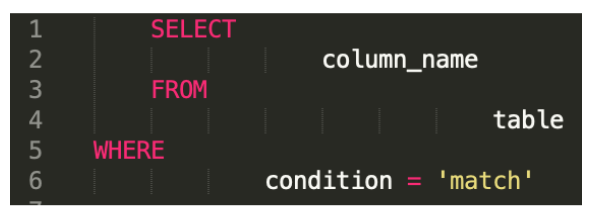
When you join a company, you can expect each company to use their own SQL platform and SQL dialect. The SQL platform they use (e.g., BigQuery, MySQL, or SQL Server) is where you will write and run your SQL queries. But keep in mind that not all SQL platforms provide native script editors to write SQL code. SQL text editors give you an interface where you can write your SQL queries in an easier and color-coded way. In fact, all of the code we have been working with so far was written with an SQL text editor!

**Examples with Sublime Text**

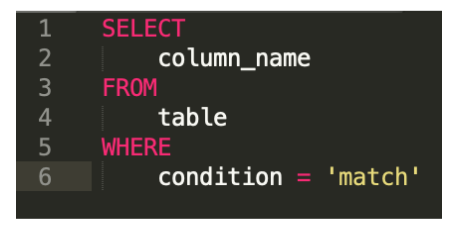
If your SQL platform doesn’t have color coding, you might want to think about using a text editor like [Sublime Text](https://www.sublimetext.com/) or [Atom](https://atom.io/). This section shows how SQL is displayed in Sublime Text. Here is a query in Sublime Text:



With Sublime Text, you can also do advanced editing like deleting indents across multiple lines at the same time. For example, suppose your query somehow had indents in the wrong places and looked like this:



This is really hard to read, so you will want to eliminate those indents and start over. In a regular SQL platform, you would have to go into each line and press BACKSPACE to delete each indent per line. But in Sublime, you can get rid of all the indents at the same time by selecting all lines and pressing Command (or CTRL in Windows) + [. This eliminates indents from every line. Then you can select the lines that you want to indent (i.e., lines 2, 4, and 6) by pressing the Command key (or the CTRL key in Windows) and selecting those lines. Then while still holding down the Command key (or the CTRL key in Windows), press  ] to indent lines 2, 4, and 6 at the same time. This will clean up your query and make it look like this instead:



Sublime Text also supports regular expressions. **Regular expressions** (or **regex**) can be used to search for and replace string patterns in queries. We won’t cover regular expressions here, but you might want to learn more about them on your own because they are a very powerful tool.

You can begin with these resources:

* [Search and replace in Sublime Text](https://sublime-text-unofficial-documentation.readthedocs.io/en/latest/search_and_replace/search_and_replace_overview.html)
* [Regex tutorial](https://www.regular-expressions.info/tutorialcnt.html) (if you don’t know what regular expressions are)
* [Regex cheat sheet](https://jdhao.github.io/2019/02/28/sublime_text_regex_cheat_sheet/)

Organization guidelines

This reading summarizes best practices for file naming, organization, and storage.



**Best practices for file naming conventions**

Review the following file naming recommendations:

* Work out and agree on file naming conventions early on in a project to avoid renaming files again and again.
* Align your file naming with your team's or company's existing file-naming conventions.
* Ensure that your file names are meaningful; consider including information like project name and anything else that will help you quickly identify (and use) the file for the right purpose.
* Include the date and version number in file names; common formats are YYYYMMDD for dates and v## for versions (or revisions).
* Create a text file as a sample file with content that describes (breaks down) the file naming convention and a file name that applies it.
* Avoid spaces and special characters in file names. Instead, use dashes, underscores, or capital letters. Spaces and special characters can cause errors in some applications.

**Best practices for keeping files organized**

Remember these tips for staying organized as you work with files:

* Create folders and subfolders in a logical hierarchy so related files are stored together.
* Separate ongoing from completed work so your current project files are easier to find. Archive older files in a separate folder, or in an external storage location.
* If your files aren't automatically backed up, manually back them up often to avoid losing important work.

# Learning Log: Review file structure and naming conventions



## Overview



In the previous lesson, you were introduced to file structuring and naming conventions. Now, you’ll complete an entry in your learning log reviewing these concepts and reflecting on why they are so important. By the time you complete this entry, you will have a stronger understanding of how and why data analysts use file structuring and naming conventions on the job. This will help you think critically about file structuring and naming for your own projects in the future and keep your work more organized.

## Review best practices



Before you begin thinking about what sort of naming conventions and patterns you would use in your own projects, take a moment to review the best practices for file structure and naming conventions.

When creating a file structure and naming convention pattern for a project, you should always:

* Work out your conventions early in your project. The earlier you start, the more organized you’ll be.
* Align file naming conventions with your team. Conventions are most useful when everyone follows them.
* Make sure filenames are meaningful. Stick to a consistent pattern that contains the most useful information needed.
* Keep file names short and to the point.

This includes understanding the expected structure of folders and files in a project. Where does your data live? Your spreadsheets? Your data visualizations? Being able to navigate your folders easily makes for a well-structured project.

Remember, there are some stylistic choices you’ll need to make when it comes to filename conventions. However, there are still best practices you should follow here, too:

| **Formatting Convention** | **Example** |
| --- | --- |
| Format Dates as yyyymmdd | SalesReport20201125 |
| Lead revision numbers with 0 | SalesReport20201125v02 |
| Use hyphens, underscores, or capitalized letters | SalesReport\_2020\_11\_25\_v02 |

You will reflect on the importance of these conventions and how you would approach file structuring and naming for your own projects in the learning log template linked below.



### Access your learning log

To use the learning log for this course item, click the link below and select “Use Template.”

Link to learning log template: [Review file structure and naming conventions](https://docs.google.com/document/d/1-l0JS6BNeggTsDXUbQhpX_rxRgukeHn8Zk6K-kbMzck/template/preview)

OR

If you don’t have a Google account, you can download the template directly from the attachment below.

**Learning Log Template\_ Review file structure and naming conventions**DOCX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/xyHfpcTnQB2h36XE57AdZg_540ee9a609e340bfa9346151c4b8ac2f_Learning-Log-Template_-Review-file-structure-and-naming-conventions.docx?Expires=1648080000&Signature=XnjpDH2TjpElTP5cRxXFVJNysvyOXmt~p3UeTt8EGEHrm36VzAQ1zse~pe96VRcNcTKs9Hjhnw4Sq-gv-0rCNDsnCnWf~26xKHWJ8qK7sKyK90c60s2b9T6QDN1LSZJSeSaGni~aUpZVJTRZNI9PJkzIkRwQFfJV3IS4gziTo5g_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



## Reflection



In your learning log template, write 2-3 sentences (40-60 words) responding to each of the questions about file structuring and naming conventions below:

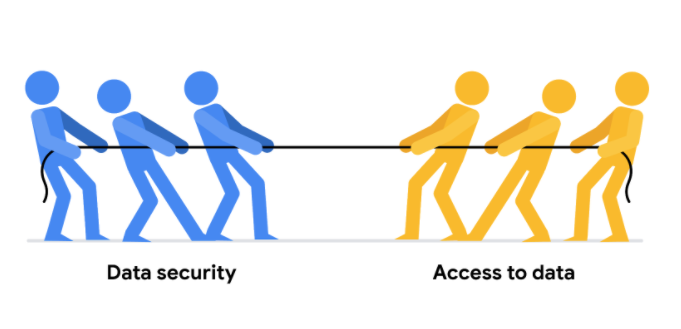
* Why are file structure and naming conventions so important? What are the consequences of poor organization for data analysts at work?
* How would you structure folders and files? What naming conventions would you use?
* What appeals to you about these choices?

# Balancing security and analytics

## The battle between security and data analytics

**Data security** means protecting data from unauthorized access or corruption by putting safety measures in place. Usually the purpose of data security is to keep unauthorized users from accessing or viewing sensitive data. Data analysts have to find a way to balance data security with their actual analysis needs. This can be tricky-- we want to keep our data safe and secure, but we also want to use it as soon as possible so that we can make meaningful and timely observations.

In order to do this, companies need to find ways to balance their data security measures with their data access needs.



Luckily, there are a few security measures that can help companies do just that. The two we will talk about here are encryption and tokenization.

**Encryption** uses a unique algorithm to alter data and make it unusable by users and applications that don’t know the algorithm. This algorithm is saved as a “key” which can be used to reverse the encryption; so if you have the key, you can still use the data in its original form.

**Tokenization** replaces the data elements you want to protect with randomly generated data referred to as a “token.” The original data is stored in a separate location and mapped to the tokens. To access the complete original data, the user or application needs to have permission to use the tokenized data and the token mapping. This means that even if the tokenized data is hacked, the original data is still safe and secure in a separate location.

Encryption and tokenization are just some of the data security options out there. There are a lot of others, like using authentication devices for AI technology.

As a junior data analyst, you probably won’t be responsible for building out these systems. A lot of companies have entire teams dedicated to data security or hire third party companies that specialize in data security to create these systems. But it is important to know that all companies have a responsibility to keep their data secure, and to understand some of the potential systems your future employer might use.

Developing a network

In this reading, you will be introduced to online and in-person opportunities to connect with other data analysts. This is part of how you develop professional relationships, which is very important when you are just starting out in your career.



**Important note for this reading**

* In-person gatherings may be restricted in your area due to the global COVID-19 pandemic. Follow the guidelines of your local health officials if you decide to attend an in-person meeting. We hope in-person meetings can soon resume safely.
* Links shared in this reading are provided as examples only and are subject to change. Google doesn’t sponsor or endorse them specifically.

**Online connections**

If you spend a few hours on social media every day you might be totally comfortable connecting with other data analysts online. But, where should you look if you don’t know any data analysts?

Even if you aren’t on social media and just created your LinkedIn profile yesterday, you can still use your online presence to find and network with other data analysts.

Knowing where to look is key. Here are some suggestions on where to start online:

* **Subscriptions** to newsletters like [Data Elixir](https://dataelixir.com/). Not only will this give you a treasure trove of useful information on a regular basis, but you will also learn the names of data science experts who you can follow, or possibly even connect with if you have good reason to.
* **Hackathons** (competitions) like those sponsored by [Kaggle](https://www.kaggle.com/), one of the largest data science and machine learning communities in the world. Participating in a hackathon might not be for everyone. But after joining a community, you typically have access to forums where you can chat and connect with other data analysts.
* **Meetups**, or online meetings that are usually local to your geography. Enter a search for ‘data science meetups near me’ to see what results you get. There is usually a posted schedule for upcoming meetings so you can attend virtually to meet other data analysts. Find out more information about [meetups happening around the world](https://www.meetup.com/topics/data-analytics/).
* **Platforms** like LinkedIn and Twitter. Use a search on either platform to find data science or data analysis hashtags to follow. You can also post your own questions or articles to generate responses and build connections that way. At the time of this writing, the LinkedIn #dataanalyst hashtag had 11,842 followers, the #dataanalytics hashtag had 98,412 followers, and the #datascience hashtag had 746,945 followers. Many of the same hashtags work on Twitter and even on Instagram.
* **Webinars** may showcase a panel of speakers and are usually recorded for convenient access and playback. You can see who is on a webinar panel and follow them too. Plus, a lot of webinars are free. One interesting pick is the [Tableau on Tableau webinar series](https://www.tableau.com/learn/series/how-we-do-data). Find out how Tableau has used Tableau in its internal departments.

**In-person (offline) gatherings**



In-person gatherings are super valuable in a digitized world. They are a great way to meet people. A lot of online relationships start from in-person gatherings and are carried on after people return home. Many organizations that sponsor annual gatherings also offer virtual meetings and resources during the rest of the year.

Here are a few suggestions to find in-person gatherings in your area:

* **Conferences** usually present innovative ideas and topics. The cost of conferences vary, and some are pricey. But lots of conferences offer discounts to students and some conferences like [Women in Analytics](https://womeninanalytics.com/about/) aim to increase the number of under-represented groups in the field. Leading research and advisory companies such as [Gartner](https://emtemp.gcom.cloud/ngw/eventassets/common/conference-calendar/gartner-conference-calendar.pdf) also sponsor conferences for data and analytics. The [KDNuggets list of meetings and online events](https://www.kdnuggets.com/meetings/index.html) for AI, analytics, big data, data science, and machine learning is useful.
* **Associations** or **societies** gather members to promote a field like data science. Many memberships are free. The [Digital Analytics Association](https://www.digitalanalyticsassociation.org/) is one example. The [KDNuggets list of societies and groups](https://www.kdnuggets.com/websites/societies.html) for analytics, data mining, data science, and knowledge discovery is useful.
* **User communities** and **summits** offer events for users of data analysis tools; this is a chance to learn from the best. Have you seen the [Tableau community](https://community.tableau.com/s/)?
* **Non-profit organizations** that promote the ethical use of data science and might offer events for the professional advancement of their members. The [Data Science Association](https://www.datascienceassn.org/) is one example.

**Key takeaways**

Your connections will help you increase your knowledge and skills. Making and keeping connections is also important to those already working in the field of data analytics. So look for online communities that promote data analysis tools or advance data science. And if available where you live, look for meetups to connect with more people face-to-face. Take advantage of both routes for the best of both worlds!  It is easier to have a conversation and exchange information in-person, but the key advantage of online connections is that they aren’t limited to where you live. Online communities might even connect you to an international crowd.