Hi! Good to see you!

My name is Sally, and I'm here to teach you all about processing data.

I'm a measurement and analytical lead at Google.

My job is to help advertising agencies and companies measure success and

analyze their data,

so I get to meet with lots of different people to show them how

data analysis helps with their advertising.

Speaking of analysis, you did great earlier learning how to gather and

organize data for analysis.

It's definitely an important step in the data analysis process, so well done!

Now let's talk about how to make sure that your organized data is complete and

accurate.

Clean data is the key to making sure your data has integrity before you analyze it.

We'll show you how to make sure your data is clean and tidy.

Cleaning and processing data is one part of the overall data analysis process.

As a quick reminder, that process is Ask, Prepare,

Process, Analyze, Share, and Act.

Which means it's time for us to explore the Process phase, and

I'm here to guide you the whole way.

I'm very familiar with where you are right now.

I'd never heard of data analytics until I went through a program similar to

this one.

Once I started making progress, I realized how much I enjoyed data analytics and

the doors it could open.

And now I'm excited to help you open those same doors!

One thing I realized as I worked for

different companies, is that clean data is important in every industry.

For example, I learned early in my career to be on the lookout for duplicate data,

a common problem that analysts come across when cleaning.

I used to work for a company that had different types of subscriptions.

In our data set, each user would have a new row for each subscription type

they bought, which meant users would show up more than once in my data.

So if I had counted the number of users in a table without accounting for

duplicates like this, I would have counted some users twice instead of once.

As a result, my analysis would have been wrong, which would have led to problems in

my reports and for the stakeholders relying on my analysis.

Imagine if I told the CEO that we had twice as many customers as we

actually did!?

That's why clean data is so important.

So the first step in processing data is learning about data integrity.

You will find out what data integrity is and

why it is important to maintain it throughout the data analysis process.

Sometimes you might not even have the data that you need,

so you'll have to create it yourself.

This will help you learn how sample size and random sampling can save you time and

effort.

Testing data is another important step to take when processing data.

We'll share some guidance on how to test data before your analysis

officially begins.

Just like you'd clean your clothes and your dishes in everyday life,

analysts clean their data all the time, too.

The importance of clean data will definitely be a focus here.

You'll learn data cleaning techniques for all scenarios,

along with some pitfalls to watch out for as you clean.

You'll explore data cleaning in both spreadsheets and databases,

building on what you've already learned about spreadsheets.

We'll talk more about SQL and how you can use it to clean data and

do other useful things, too.

When analysts clean their data, they do a lot more than a spot check to make sure

it was done correctly.

You'll learn ways to verify and report your cleaning results.

This includes documenting your cleaning process, which has lots of

benefits that we'll explore.

It's important to remember that processing data is just one of the tasks you'll

complete as a data analyst.

Actually, your skills with cleaning data might just end up being something you

highlight on your resume when you start job hunting.

Speaking of resumes, you'll be able to start thinking about how to build

your own from the perspective of a data analyst.

Once you're done here, you'll have a strong appreciation for clean data and

how important it is in the data analysis process.

So let's get started!

# Course 4 overview: Set your expectations

Welcome to the fourth course in the program! As you make your way through the certificate, this course and others that follow will begin to focus more on practical, skills-based assignments and projects.

In this course, you’ll learn to clean data by checking it for completeness and correctness. You’ll review a variety of approaches to clean data in spreadsheets and databases. Then, you’ll gain essential troubleshooting skills that will enable you to fix any errors. An important step in cleaning data is creating reports to communicate the changes you’ve made to others. You’ll understand how to do that in order to ensure the accuracy and reliability of data. Together, these skills will help ensure your data analysis is successful.

Course menu showing courses 1 to 8. Courses 1, 2, and 3 are complete. Course 4 is active, and courses 5, 6, 7, and 8 are not yet complete. Course menu showing: Course 1: Foundations: Data, data, everywhere; Course 2: Ask questions to make data-driven decisions; Course 3: Prepare data for exploration; Course 4: Process data from dirty to clean; Course 5: Analyze data to answer questions; Course 6: Share data through the art of visualization; Course 7: Data analysis with R programming; Course 8: Data analytics capstone project: Complete a case study.

## Course Overview

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](https://www.coursera.org/learn/data-preparation/home/welcome)
4. **Process Data from Dirty to Clean** (this course)
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/welcome)
7. [Data Analysis with R Programming](https://coursera.org/learn/data-analysis-r/home/welcome)
8. [Google Data Analytics Capstone: Complete a Case Study](https://coursera.org/learn/google-data-analytics-capstone/home/welcome)

## Course 4 content

Each course is broken into modules. Here’s a quick overview of the skills you’ll gain in each of the five Course 4 modules.

### **Module 1: The importance of integrity**

Data integrity is critical to successful analysis. In this part of the course, you’ll explore methods and steps that analysts take to check their data for integrity. This includes knowing what to do when you don’t have enough data. You’ll also learn about random samples and understand how to avoid sampling bias. All of these methods will also help you ensure your analysis is successful.

### **Module 2: Clean data for more accurate insights**

Every data analyst wants to analyze clean data. In this part of the course, you’ll learn the difference between clean and dirty data. Then, you’ll practice cleaning data in spreadsheets and other tools.

### **Module 3: Data cleaning with SQL**

Knowing a variety of ways to clean data can make a data analyst’s job much easier. In this part of the course, you’ll use SQL to clean data from databases. In particular, you’ll explore how SQL queries and functions can be used to clean and transform your data before an analysis.

### **Module 4: Verify and report cleaning results**

When you clean data, you make changes to the original dataset. It’s important to verify the changes you make are accurate and to let your teammates know about the changes. In this part of the course, you’ll learn to verify that data is clean and report your data cleaning results. With verified clean data, you’re ready to begin analyzing!

### **Module 5: Optional: Add data to your resume**

Creating an effective resume will help you in your data analytics career. In this part of the course, you’ll learn all about the job application process. Your focus will be on building a resume that highlights your strengths and relevant experience.

### Module 6: Course wrap-up

Review the course glossary and prepare for the next course in the Google Data Analytics Certificate program.

## What to expect

Each module includes a series of lessons with many types of learning opportunities. These include:

* **Videos** for instructors to teach new concepts and demonstrate the use of tools
* **In-video questions** that pop up from time to time to help you to check your understanding of key concepts and skills
* **Step-by-step guides** you can use to follow along with instructors as they demonstrate tools
* **Readings** to explore topics more in-depth and build on the concepts from the videos
* **Discussion forums** to share, explore, and reinforce lesson topics
* **Discussion prompts** to promote thinking and engagement in the discussion forums
* **Practice quizzes** to prepare you for graded quizzes
* **Graded quizzes** to measure your progress and give you valuable feedback

This program was designed to let you work at your own pace—your personalized deadlines are just a guide. There is no penalty for late assignments. To earn your certificate, you simply need to complete all of the work.

If you miss two assessment deadlines in a row, or if you miss an assessment deadline by two weeks, you'll see a **Reset deadlines** option on the **Grades** page. Click it to switch to a new course schedule with updated deadlines. You can use this option as many times as you need—it won’t remove any progress you’ve already made in the course, but you may find new course content if the instructor updated the course after you started. If you cancel a subscription and then reactivate it, your deadlines will automatically reset.

In this course, you'll be assessed with graded quizzes and activities. Both are based on the wide variety of learning materials and activities that reinforce the important skills you’ll develop. And both can be taken more than once.

## Tips for success

* It is strongly recommended that you go through the items in each lesson in the order they appear because new information and concepts build on previous knowledge.
* Participate in all learning opportunities to gain as much knowledge and experience as possible.
* If something is confusing, don’t hesitate to replay a video, review a reading, or repeat a self-review activity.
* Use the additional resources that are referenced in this course. They are designed to support your learning. You can find all of these resources in the [**Resources**](https://www.coursera.org/learn/ask-questions-make-decisions/resources/xW7lI) tab.
* When you encounter useful links in this course, bookmark them so you can refer to the information later for study or review.
* Understand and follow the [**Coursera Code of Conduct**](https://www.coursera.support/s/article/208280036-Coursera-Code-of-Conduct?) to ensure that the learning community remains a welcoming, friendly, and supportive place for all members.

**Updates to the course**

As you complete this course, you may notice updates to the content, like new practice materials and additional examples. These updates ensure the program provides up-to-date skills and guidance that will help you in your data analytics career. If you previously completed a graded activity, you may need to repeat the assessment in order to complete this course.  For more information, check out [the course discussion forum.](https://www.coursera.org/learn/process-data/discussions)

Welcome back. In this video,

we're going to discuss data integrity and

some risks you might run into as a data analyst.

A strong analysis depends on the integrity of the data.

If the data you're using is compromised in any way,

your analysis won't be as strong as it should be.

Data integrity is the accuracy, completeness,

consistency, and trustworthiness of

data throughout its lifecycle.

That might sound like a lot of

qualities for the data to live up to.

But trust me, it's worth it to check for them

all before proceeding with your analysis.

Otherwise, your analysis could be wrong.

Not because you did something wrong,

but because the data you were working

with was wrong to begin with.

When data integrity is low,

it can cause anything from the loss of a single pixel in

an image to an incorrect medical decision.

In some cases,

one missing piece can make all of your data useless.

Data integrity can be

compromised in lots of different ways.

There's a chance data can be

compromised every time it's replicated,

transferred, or manipulated in any way.

Data replication is the process of

storing data in multiple locations.

If you're replicating data at

different times in different places,

there's a chance your data will be out of sync.

This data lacks integrity

because different people might not be

using the same data for

their findings, which can cause inconsistencies.

There's also the issue of data transfer,

which is the process of copying

data from a storage device

to memory, or from one computer to another.

If your data transfer is interrupted,

you might end up with an incomplete data set,

which might not be useful for your needs.

The data manipulation process involves changing

the data to make it more organized and easier to read.

Data manipulation is meant to make

the data analysis process more efficient,

but an error during the process

can compromise the efficiency.

Finally, data can also be compromised

through human error, viruses,

malware, hacking, and system failures,

which can all lead to even more headaches.

I'll stop there.

That's enough potentially bad news to digest.

Let's move on to some potentially good news.

In a lot of companies, the data warehouse or

data engineering team takes

care of ensuring data integrity.

Coming up, we'll learn about checking

data integrity as a data analyst.

But rest assured, someone

else will usually have your back too.

After you've found out what data you're working with,

it's important to double-check that your data is

complete and valid before analysis.

This will help ensure that your analysis and

eventual conclusions are accurate.

Checking data integrity is a vital step in

processing your data to get it ready for analysis,

whether you or someone else at your company is doing it.

Coming up, you'll learn even more

about data integrity. See you soon!

**More about data integrity and compliance**

This reading illustrates the importance of data integrity using an example of a global company’s data. Definitions of terms that are relevant to data integrity will be provided at the end.

**Scenario: calendar dates for a global company**

Calendar dates are represented in a lot of different short forms. Depending on where you live, a different format might be used.

* In some countries, **12/10/20** (DD/MM/YY) stands for October 12, 2020.
* In other countries, the national standard is YYYY-MM-DD so October 12, 2020 becomes **2020-10-12**.
* In the United States, (MM/DD/YY) is the accepted format so October 12, 2020 is going to be **10/12/20**.

Now, think about what would happen if you were working as a data analyst for a global company and didn’t check date formats. Well, your data integrity would probably be questionable. Any analysis of the data would be inaccurate. Imagine ordering extra inventory for December when it was actually needed in October!

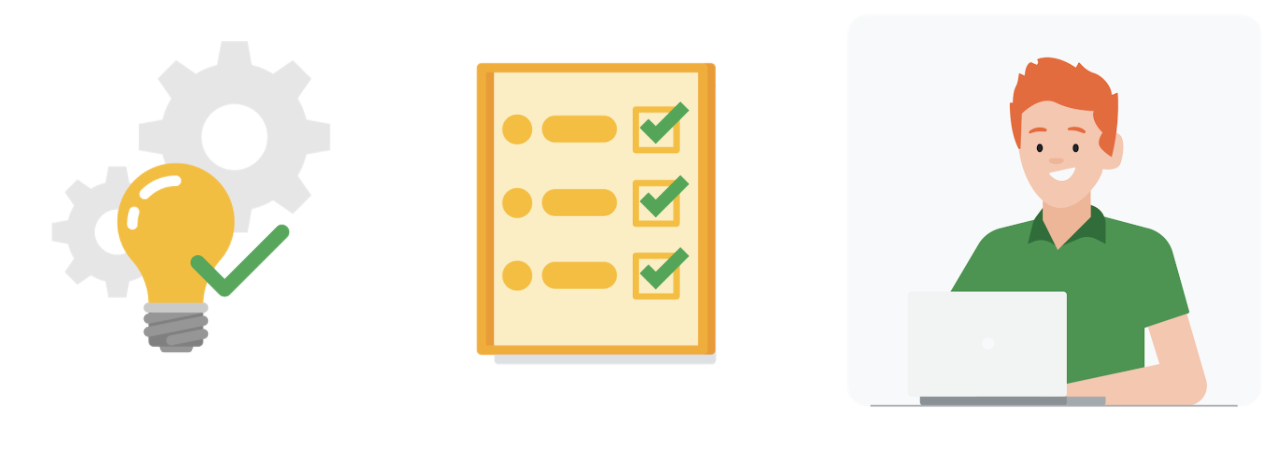
A good analysis depends on the integrity of the data, and data integrity usually depends on using a common format. So it is important to double-check how dates are formatted to make sure what you think is December 10, 2020 isn’t really October 12, 2020, and vice versa.

Here are some other things to watch out for:

* **Data replication compromising data integrity:** Continuing with the example, imagine you ask your international counterparts to verify dates and stick to one format. One analyst copies a large dataset to check the dates. But because of memory issues, only part of the dataset is actually copied. The analyst would be verifying and standardizing incomplete data. That partial dataset would be certified as compliant but the full dataset would still contain dates that weren't verified. Two versions of a dataset can introduce inconsistent results. A final audit of results would be essential to reveal what happened and correct all dates.
* **Data transfer compromising data integrity:** Another analyst checks the dates in a spreadsheet and chooses to import the validated and standardized data back to the database. But suppose the date field from the spreadsheet was incorrectly classified as a text field during the data import (transfer) process. Now some of the dates in the database are stored as text strings. At this point, the data needs to be cleaned to restore its integrity.
* **Data manipulation compromising data integrity:** When checking dates, another analyst notices what appears to be a duplicate record in the database and removes it. But it turns out that the analyst removed a unique record for a company’s subsidiary and not a duplicate record for the company. Your dataset is now missing data and the data must be restored for completeness.

**Conclusion**

Fortunately, with a standard date format and compliance by all people and systems that work with the data, data integrity can be maintained. But no matter where your data comes from, always be sure to check that it is valid, complete, and clean before you begin any analysis.



**Reference: Data constraints and examples**

As you progress in your data journey, you'll come across many types of data constraints (or criteria that determine validity). The table below offers definitions and examples of data constraint terms you might come across.

| **Data constraint** | **Definition** | **Examples** |
| --- | --- | --- |
| **Data type** | Values must be of a certain type: date, number, percentage, Boolean, etc. | If the data type is a date, a single number like 30 would fail the constraint and be invalid |
| **Data range** | Values must fall between predefined maximum and minimum values | If the data range is 10-20, a value of 30 would fail the constraint and be invalid |
| **Mandatory** | Values can’t be left blank or empty | If age is mandatory, that value must be filled in |
| **Unique** | Values can’t have a duplicate | Two people can’t have the same mobile phone number within the same service area |
| **Regular expression (regex) patterns** | Values must match a prescribed pattern | A phone number must match ###-###-#### (no other characters allowed) |
| **Cross-field validation** | Certain conditions for multiple fields must be satisfied | Values are percentages and values from multiple fields must add up to 100% |
| **Primary-key** | (Databases only) value must be unique per column | A database table can’t have two rows with the same primary key value. A primary key is an identifier in a database that references a column in which each value is unique. More information about primary and foreign keys is provided later in the program. |
| **Set-membership** | (Databases only) values for a column must come from a set of discrete values | Value for a column must be set to Yes, No, or Not Applicable |
| **Foreign-key** | (Databases only) values for a column must be unique values coming from a column in another table | In a U.S. taxpayer database, the State column must be a valid state or territory with the set of acceptable values defined in a separate States table |
| **Accuracy** | The degree to which the data conforms to the actual entity being measured or described | If values for zip codes are validated by street location, the accuracy of the data goes up. |
| **Completeness** | The degree to which the data contains all desired components or measures | If data for personal profiles required hair and eye color, and both are collected, the data is complete. |
| **Consistency** | The degree to which the data is repeatable from different points of entry or collection | If a customer has the same address in the sales and repair databases, the data is consistent. |

Hey there, it's good to remember to check for data integrity.

It's also important to check that the data you use aligns with the business

objective.

This adds another layer to the maintenance of data integrity

because the data you're using might have limitations that you'll need to deal with.

The process of matching data to business objectives can actually be pretty

straightforward.

Here's a quick example. Let's say you're an analyst for

a business that produces and sells auto parts.

Play video starting at ::29 and follow transcript0:29

If you need to address a question about the revenue generated by the sale of

a certain part, then you'd pull up the revenue table from the data set.

Play video starting at ::37 and follow transcript0:37

If the question is about customer reviews,

then you'd pull up the reviews table to analyze the average ratings.

But before digging into any analysis,

you need to consider a few limitations that might affect it.

If the data hasn't been cleaned properly, then you won't be able to use it yet.

You would need to wait until a thorough cleaning has been done.

Now, let's say you're trying to find how much an average customer spends.

You notice the same customer's data showing up in more than one row.

This is called duplicate data.

To fix this, you might need to change the format of the data, or

you might need to change the way you calculate the average.

Otherwise, it will seem like the data is for two different people, and

you'll be stuck with misleading calculations.

You might also realize there's not enough data to complete an accurate analysis.

Maybe you only have a couple of months' worth of sales data.

There's slim chance you could wait for more data, but

it's more likely that you'll have to change your process or

find alternate sources of data while still meeting your objective.

I like to think of a data set like a picture.

Take this picture. What are we looking at?

Play video starting at :1:48 and follow transcript1:48

Unless you're an expert traveler or know the area,

it may be hard to pick out from just these two images.

Play video starting at :1:55 and follow transcript1:55

Visually, it's very clear when we aren't seeing the whole picture.

When you get the complete picture, you realize... you're in London!

Play video starting at :2:4 and follow transcript2:04

With incomplete data,

it's hard to see the whole picture to get a real sense of what is going on.

We sometimes trust data because if it comes to us in rows and columns, it seems

like everything we need is there if we just query it. But that's just not true.

I remember a time when I found out I didn't have enough data and

had to find a solution.

Play video starting at :2:26 and follow transcript2:26

I was working for an online retail company and

was asked to figure out how to shorten customer purchase to delivery time.

Faster delivery times usually lead to happier customers.

When I checked the data set, I found very limited tracking information.

We were missing some pretty key details.

So the data engineers and I created new processes to track additional information,

like the number of stops in a journey.

Using this data, we reduced the time it took from purchase to delivery and

saw an improvement in customer satisfaction. That felt pretty great!

Learning how to deal with data issues while staying focused on your objective

will help set you up for success in your career as a data analyst. And

your path to success continues.

Next step, you'll learn more about aligning data to objectives. Keep it up!

# Well-aligned objectives and data

You can gain powerful insights and make accurate conclusions when data is well-aligned to business objectives. As a data analyst, alignment is something you will need to judge. Good alignment means that the data is relevant and can help you solve a business problem or determine a course of action to achieve a given business objective.

In this reading, you will review the business objectives associated with three scenarios. You will explore how clean data and well-aligned business objectives can help you come up with accurate conclusions. On top of that, you will learn how new variables discovered during data analysis can cause you to set up data constraints so you can keep the data aligned to a business objective.

## Clean data + alignment to business objective = accurate conclusions

### **Business objective**

Account managers at Impress Me, an online content subscription service, want to know how soon users view content after their subscriptions are activated.



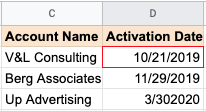
To start off, the data analyst verifies that the data exported to spreadsheets is clean and confirms that the data needed (when users access content) is available. Knowing this, the analyst decides there is good alignment of the data to the business objective. All that is missing is figuring out exactly how long it takes each user to view content after their subscription has been activated.

Here are the data processing steps the analyst takes for a user from an account called V&L Consulting. (These steps would be repeated for each subscribing account, and for each user associated with that account.)

### **Step 1**

| **Data-processing step** | **Source of data** |
| --- | --- |
| Look up the activation date for V&L Consulting | Account spreadsheet |

**Relevant data in spreadsheet:**

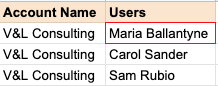


**Result**: October 21, 2019

### **Step 2**

| **Data-processing step** | **Source of data** |
| --- | --- |
| Look up the name of a user belonging to the V&L Consulting account | Account spreadsheet (users tab) |

**Relevant data in spreadsheet**:



**Result**: Maria Ballantyne

### **Step 3**

| **Data-processing step** | **Source of data** |
| --- | --- |
| Find the first content access date for Maria B. | Content usage spreadsheet |

**Relevant data in spreadsheet:**

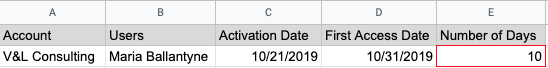


**Result**: October 31, 2019

### **Step 4**

| **Data-processing step** | **Source of data** |
| --- | --- |
| Calculate the time between activation and first content usage for Maria B. | New spreadsheet calculation |

**Relevant data in spreadsheet**:



**Result**: 10 days

### **Pro tip 1**

In the above process, the analyst could use **VLOOKUP** to look up the data in Steps 1, 2, and 3 to populate the values in the spreadsheet in Step 4. [VLOOKUP](https://support.microsoft.com/en-us/office/vlookup-function-0bbc8083-26fe-4963-8ab8-93a18ad188a1) is a spreadsheet function that searches for a certain value in a column to return a related piece of information. Using **VLOOKUP** can save a lot of time; without it, you have to look up dates and names manually.

Refer to the [VLOOKUP](https://support.google.com/docs/answer/3093318?hl=en) page in the Google Help Center for how to use the function in Google Sheets.

### **Pro tip 2**

In Step 4 of the above process, the analyst could usethe **DATEDIF** function to automatically calculate the difference between the dates in column C and column D. The function can calculate the number of days between two dates.

Refer to the Microsoft Support [DATEDIF](https://support.microsoft.com/en-us/office/datedif-function-25dba1a4-2812-480b-84dd-8b32a451b35c) page for how to use the function in Excel. The [DAYS360](https://support.microsoft.com/en-us/office/days360-function-b9a509fd-49ef-407e-94df-0cbda5718c2a) function does the same thing in accounting spreadsheets that use a 360-day year (twelve 30-day months).

Refer to the [DATEDIF](https://support.google.com/docs/answer/6055612?hl=en) page in the Google Help Center for how to use the function in Google Sheets.

## Alignment to business objective + additional data cleaning = accurate conclusions

### **Business objective**

Cloud Gate, a software company, recently hosted a series of public webinars as free product introductions. The data analyst and webinar program manager want to identify companies that had five or more people attend these sessions. They want to give this list of companies to sales managers who can follow up for potential sales.



The webinar attendance data includes the fields and data shown below.

| **Name** | **<First name> <Last name>** | **This was required information attendees had to submit** |
| --- | --- | --- |
| **Email Address** | xxxxx@company.com | This was required information attendees had to submit |
| **Company** | <Company name> | This was optional information attendees could provide |

### **Data cleaning**

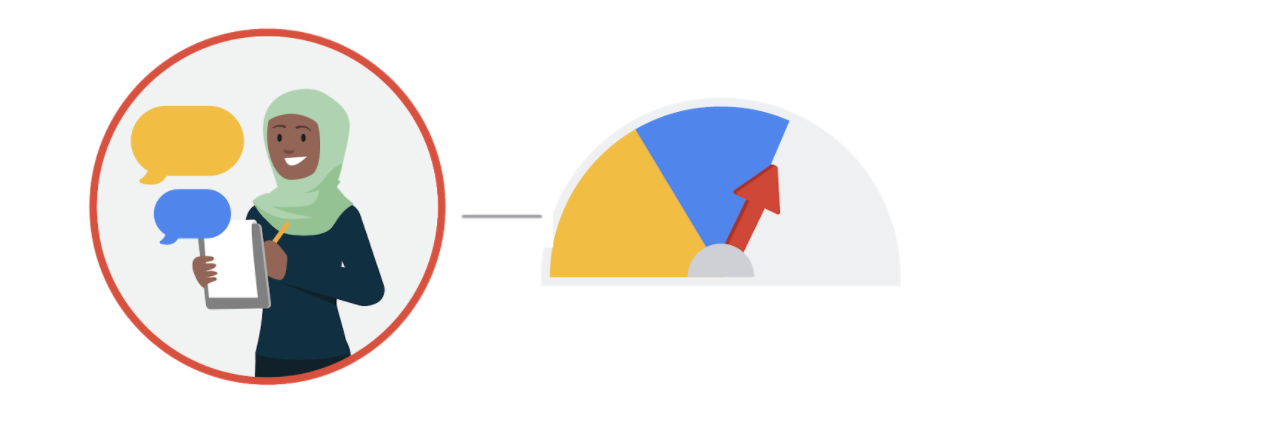
The webinar attendance data seems to align with the business objective. But the data analyst and program manager decide that some data cleaning is needed before the analysis. They think data cleaning is required because:

* The company name wasn’t a mandatory field. If the company name is blank, it might be found from the email address. For example, if the email address is username@google.com, the company field could be filled in with Google for the data analysis. This data cleaning step assumes that people with company-assigned email addresses attended a webinar for business purposes.
* Attendees could enter any name. Since attendance across a series of webinars is being looked at, they need to validate names against unique email addresses. For example, if Joe Cox attended two webinars but signed in as Joe Cox for one and Joseph Cox for the other, he would be counted as two different people. To prevent this, they need to check his unique email address to determine that he was the same person. After the validation, Joseph Cox could be changed to Joe Cox to match the other instance.

## Alignment to business objective + newly discovered variables + constraints = accurate conclusions

### **Business objective**

An after-school tutoring company, A+ Education,  wants to know if there is a minimum number of tutoring hours needed before students have at least a 10% improvement in their assessment scores.



The data analyst thinks there is good alignment between the data available and the business objective because:

* Students log in and out of a system for each tutoring session, and the number of hours is tracked
* Assessment scores are regularly recorded

### **Data constraints for new variables**

After looking at the data, the data analyst discovers that there are other variables to consider. Some students had consistent weekly sessions while other students had scheduled sessions more randomly even though their total number of tutoring hours was the same. The data doesn’t align as well with the original business objective as first thought, so the analyst adds a data constraint to focus only on the students with consistent weekly sessions. This modification helps to get a more accurate picture about the enrollment time needed to achieve a 10% improvement in assessment scores.

## Key takeaways

Hopefully these examples give you a sense of what to look for to know if your data aligns with your business objective.

* When there is clean data and good alignment, you can get accurate insights and make conclusions the data supports.
* If there is good alignment but the data needs to be cleaned, clean the data before you perform your analysis.
* If the data only partially aligns with an objective, think about how you could modify the objective, or use data constraints to make sure that the subset of data better aligns with the business objective.

### 1.

Question 1

Fill in the blank: Data \_\_\_\_\_ involves the accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle.

1 point

manipulation

replication

transferral

integrity

### 2.

Question 2

Which process do data analysts use to make data more organized and easier to read?

1 point

Data uniformity

Data transfer

Data replication

Data manipulation

### 3.

Question 3

Before analysis, a company collects data from countries that use different date formats. Which of the following actions would improve the data integrity?

1 point

Remove data in an unfamiliar date format

Organize the data by country

Change all of the dates to the same format

Leave the dates in their current formats

### 4.

Question 4

In this spreadsheet, what common data problem appears in rows 2 and 7 ?

|  |  |  |  |
| --- | --- | --- | --- |
|  | **First** | **Last** | **CustID** |
| **1** | Douglas | Pool | 10794 |
| **2** | Ronnie | Mazlan | 10351 |
| **3** | Tonya | Butcher | 10990 |
| **4** | Yanni | Morningside | 10184 |
| **5** | Eliza | Fe | 10212 |
| **6** | Travis | Tatien | 10746 |
| **7** | Ronnie | Mazlan | 10351 |

1 point

Data virus

System failure

Insufficient data

Duplicate data

Every analyst has been in a situation where there is insufficient data to help

with their business objective.

Considering how much data is generated every day, it may be hard to believe, but

it's true.

So let's discuss what you can do when you have insufficient data.

We'll cover how to set limits for the scope of your analysis and

what data you should include.

Play video starting at ::21 and follow transcript0:21

At one point, I was a data analyst at a support center.

Every day, we received customer questions, which were logged in as support tickets.

Play video starting at ::29 and follow transcript0:29

I was asked to forecast the number of support tickets coming in per month to

figure out how many additional people we needed to hire.

It was very important that we had sufficient data spanning back at least

a couple of years because I had to account for year-to-year and seasonal changes.

If I just had the current year's data available,

I wouldn't have known that a spike in January is common and

has to do with people asking for refunds after the holidays.

Because I had sufficient data,

I was able to suggest we hire more people in January to prepare.

Challenges are bound to come up,

but the good news is that once you know your business objective,

you'll be able to recognize whether you have enough data.

And if you don't, you'll be able to deal with it before you start your analysis.

Now, let's check out some of those limitations you might come across and

how you can handle different types of insufficient data.

Play video starting at :1:22 and follow transcript1:22

Say you're working in the tourism industry, and

you need to find out which travel plans are searched most often.

If you only use data from one booking site,

you're limiting yourself to data from just one source.

Other booking sites might show different trends that you would want to consider for

your analysis.

If a limitation like this impacts your analysis, you can stop and

go back to your stakeholders to figure out a plan.

If your data set keeps updating, that means the data is still incoming and

might not be complete.

So if there's a brand new tourist attraction that you're analyzing interest

and attendance for, there's probably not enough data for you to determine trends.

For example, you might want to wait a month to gather data.

Or you can check in with the stakeholders and ask about adjusting the objective.

For example, you might analyze trends from week to week instead of month to month.

You could also base your analysis on trends over the past three months and say,

"Here's what attendance at the attraction for month four could look like."

Play video starting at :2:22 and follow transcript2:22

You might not have enough data to know if this number is too low or too high.

But you would tell stakeholders that it's your best estimate based on the data that

you currently have.

On the other hand, your data could be older and no longer be relevant.

Outdated data about customer satisfaction won't include the most recent responses.

So you'll be relying on the ratings for hotels or

vacation rentals that might no longer be accurate.

In this case, your best bet might be to find a new data set to work with.

Data that's geographically-limited could also be unreliable.

If your company is global, you wouldn't want to use data limited to travel

in just one country. You would want a data set that includes all countries.

So that's just a few of the most common limitations you'll come across and

some ways you can address them.

You can identify trends with the available data or wait for more data if time allows;

you can talk with stakeholders and adjust your objective; or you can look for

a new data set.

Play video starting at :3:22 and follow transcript3:22

The need to take these steps will depend on your role in your company and

possibly the needs of the wider industry.

But learning how to deal with insufficient data is always a great way to set yourself

up for success.

Your data analyst powers are growing stronger. And just in time.

After you learn more about limitations and solutions,

you'll learn about statistical power, another fantastic tool for you to use.

See you soon!

# When you find an issue with your data

When you are getting ready for data analysis, you might realize you don’t have the data you need or you don’t have enough of it. In some cases, you can use what is known as proxy data in place of the real data. Think of it like substituting oil for butter in a recipe when you don’t have butter. In other cases, there is no reasonable substitute and your only option is to collect more data.

Consider the following data issues and suggestions on how to work around them.

## Data issue 1: no data

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the analysis after you have collected more data. | If you are surveying employees about what they think about a new performance and bonus plan, use a sample for a preliminary analysis. Then, ask for another 3 weeks to collect the data from all employees. |
| If there isn’t time to collect data, perform the analysis using proxy data from other datasets.  This is the most common workaround. | If you are analyzing peak travel times for commuters but don’t have the data for a particular city, use the data from another city with a similar size and demographic. |

## Data issue 2: too little data

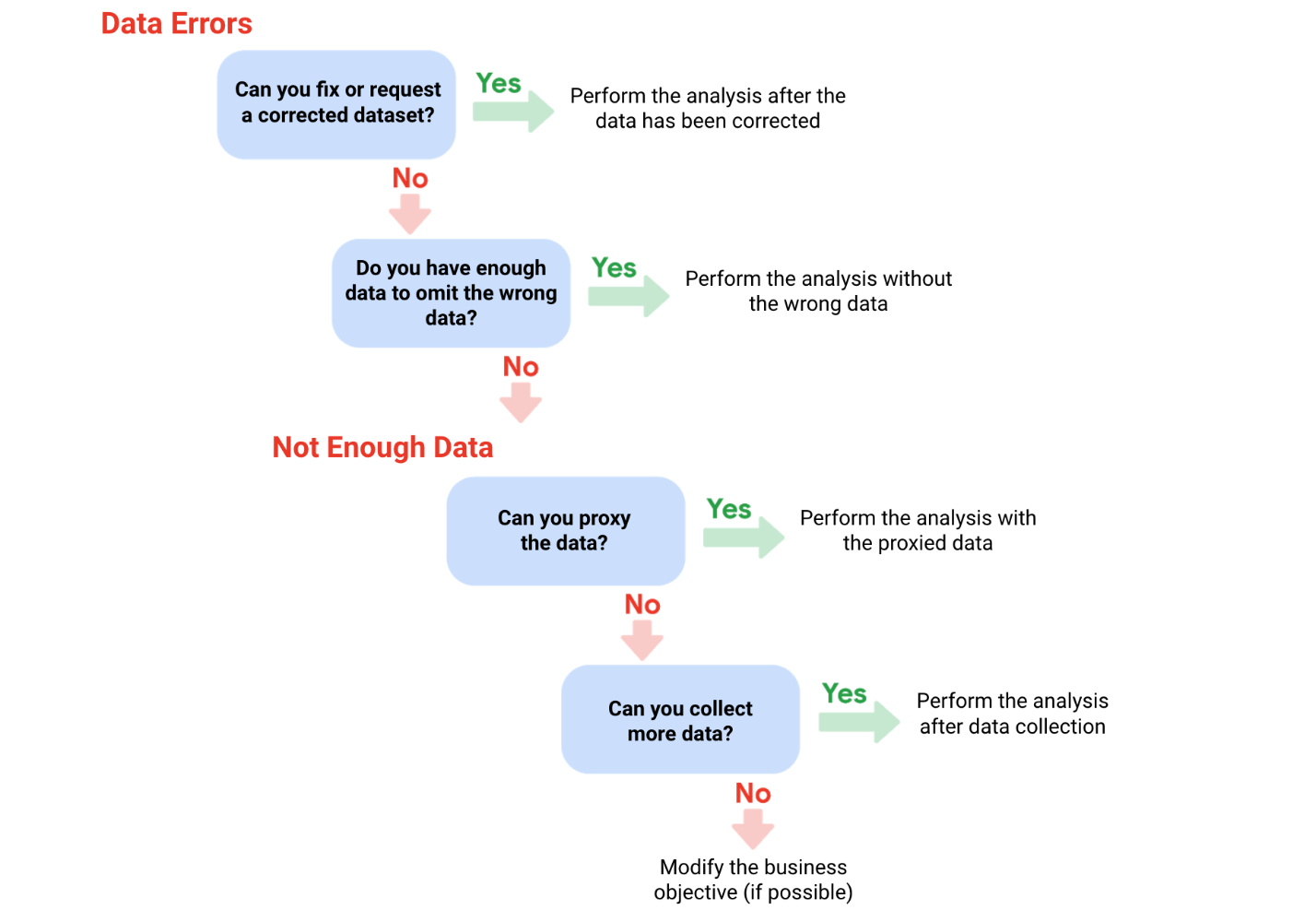
| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Do the analysis using proxy data along with actual data. | If you are analyzing trends for owners of golden retrievers, make your dataset larger by including the data from owners of labradors. |
| Adjust your analysis to align with the data you already have. | If you are missing data for 18- to 24-year-olds, do the analysis but note the following limitation in your report: this conclusion applies to adults 25 years and older only. |

## Data issue 3: wrong data, including data with errors\*

| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| If you have the wrong data because requirements were misunderstood, communicate the requirements again. | If you need the data for female voters and received the data for male voters, restate your needs. |
| Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors. | If your data is in a spreadsheet and there is a conditional statement or boolean causing calculations to be wrong, change the conditional statement instead of just fixing the calculated values. |
| If you can’t correct data errors yourself, you can ignore the wrong data and go ahead with the analysis if your sample size is still large enough and ignoring the data won’t cause systematic bias. | If your dataset was translated from a different language and some of the translations don’t make sense, ignore the data with bad translation and go ahead with the analysis of the other data. |

***\* Important note:*** Sometimes data with errors can be a warning sign that the data isn’t reliable. Use your best judgment.

### Use the following decision tree as a reminder of how to deal with data errors or not enough data:



Okay, earlier we talked about having

the right kind of data to meet your business objective

and the importance of having the right amount of data to

make sure your analysis is as accurate as possible.

You might remember that for data analysts,

a population is all possible data values

in a certain dataset.

If you're able to use 100 percent of

a population in your analysis, that's great.

But sometimes collecting information about

an entire population just isn't possible.

It's too time-consuming or expensive.

For example, let's say a global organization

wants to know more about pet owners who have cats.

You're tasked with finding out which kinds of

toys cat owners in Canada prefer.

But there's millions of cat owners in Canada,

so getting data from all of

them would be a huge challenge.

Fear not! Allow me to introduce you to... sample size!

When you use sample size or a sample,

you use a part of a population

that's representative of the population.

The goal is to get enough information

from a small group within

a population to make predictions

or conclusions about the whole population.

The sample size helps

ensure the degree to which you can be

confident that your conclusions

accurately represent the population.

For the data on cat owners,

a sample size might contain data about

hundreds or thousands of people rather than millions.

Using a sample for analysis is more

cost-effective and takes less time.

If done carefully and thoughtfully,

you can get the same results using

a sample size instead of trying to hunt

down every single cat owner

to find out their favorite cat toys.

There is a potential downside, though.

When you only use a small sample of a population,

it can lead to uncertainty.

You can't really be 100 percent

sure that your statistics are

a complete and accurate representation of the population.

This leads to sampling bias,

which we covered earlier in the program.

Sampling bias is when a sample

isn't representative of the population as a whole.

This means some members of the population are

being overrepresented or underrepresented.

For example, if the survey used to collect data from

cat owners only included people with smartphones,

then cat owners who don't have

a smartphone wouldn't be represented in the data.

Using random sampling can help

address some of those issues with sampling bias.

Random sampling is a way of

selecting a sample from a population so

that every possible type of

the sample has an equal chance of being chosen.

Going back to our cat owners again,

using a random sample of cat owners means cat owners

of every type have an equal chance of being chosen.

Cat owners who live in apartments in Ontario would have

the same chance of being represented as

those who live in houses in Alberta.

As a data analyst,

you'll find that creating sample sizes usually

takes place before you even get to the data.

But it's still good for you to know that

the data you are going to analyze is

representative of the population

and works with your objective.

It's also good to know what's

coming up in your data journey.

In the next video, you'll have an option to become even

more comfortable with sample sizes. See you there.

# Calculate sample size

Before you dig deeper into sample size, familiarize yourself with these terms and definitions:

| **Terminology** | **Definitions** |
| --- | --- |
| **Population** | The entire group that you are interested in for your study. For example, if you are surveying people in your company, the population would be all the employees in your company. |
| **Sample** | A subset of your population. Just like a food sample, it is called a sample because it is only a taste. So if your company is too large to survey every individual, you can survey a representative sample of your population. |
| **Margin of error** | Since a sample is used to represent a population, the sample’s results are expected to differ from what the result would have been if you had surveyed the entire population. This difference is called the margin of error. The smaller the margin of error, the closer the results of the sample are to what the result would have been if you had surveyed the entire population. |
| **Confidence level** | How confident you are in the survey results. For example, a 95% confidence level means that if you were to run the same survey 100 times, you would get similar results 95 of those 100 times. Confidence level is targeted before you start your study because it will affect how big your margin of error is at the end of your study. |
| **Confidence interval** | The range of possible values that the population’s result would be at the confidence level of the study. This range is the sample result +/- the margin of error. |
| **Statistical significance** | The determination of whether your result could be due to random chance or not. The greater the significance, the less due to chance. |

## Things to remember when determining the size of your sample

When figuring out a sample size, here are things to keep in mind:

* Don’t use a sample size less than 30. It has been statistically proven that 30 is the smallest sample size where an average result of a sample starts to represent the average result of a population.
* The confidence level most commonly used is 95%, but 90% can work in some cases.

Increase the sample size to meet specific needs of your project:

* For a **higher** confidence level, use a larger sample size
* To **decrease** the margin of error, use a larger sample size
* For **greater** statistical significance, use a larger sample size

**Note:** Sample size calculators use statistical formulas to determine a sample size. More about these are coming up in the course! Stay tuned.

### **Why a minimum sample of 30?**

This recommendation is based on the **Central Limit Theorem (CLT)** in the field of probability and statistics. As sample size increases, the results more closely resemble the normal (bell-shaped) distribution from a large number of samples. A sample of 30 is the smallest sample size for which the CLT is still valid. Researchers who rely on **regression analysis** –statistical methods to determine the relationships between controlled and dependent variables –also prefer a minimum sample of 30.

Still curious? Without getting too much into the math, check out these articles:

* [Central Limit Theorem (CLT)](https://www.investopedia.com/terms/c/central_limit_theorem.asp): This article by Investopedia explains the Central Limit Theorem and briefly describes how it can apply to an analysis of a stock index.
* [Sample Size Formula](https://www.statisticssolutions.com/dissertation-resources/sample-size-calculation-and-sample-size-justification/sample-size-formula/): This article by Statistics Solutions provides a little more detail about why some researchers use 30 as a minimum sample size.

## Sample sizes vary by business problem

Sample size will vary based on the type of business problem you are trying to solve.

For example, if you live in a city with a population of 200,000 and get 180,000 people to respond to a survey, that is a large sample size. But without actually doing that, what would an acceptable, smaller sample size look like?

Would 200 be alright if the people surveyed represented every district in the city?

**Answer**: It depends on the stakes.

* A sample size of 200 might be large enough if your business problem is to find out how residents felt about the new library
* A sample size of 200 might not be large enough if your business problem is to determine how residents would vote to fund the library

You could probably accept a larger margin of error surveying how residents feel about the new library versus surveying residents about how they would vote to fund it. For that reason, you would most likely use a larger sample size for the voter survey.



## Larger sample sizes have a higher cost

You also have to weigh the cost against the benefits of more accurate results with a larger sample size. Someone who is trying to understand consumer preferences for a new line of products wouldn’t need as large a sample size as someone who is trying to understand the effects of a new drug. For drug safety, the benefits outweigh the cost of using a larger sample size. But for consumer preferences, a smaller sample size at a lower cost could provide good enough results.



## Knowing the basics is helpful

Knowing the basics will help you make the right choices when it comes to sample size. You can always raise concerns if you come across a sample size that is too small. A sample size calculator is also a great tool for this. Sample size calculators let you enter a desired confidence level and margin of error for a given population size. They then calculate the sample size needed to statistically achieve those results.

Refer to the [Determine the Best Sample Size](https://www.coursera.org/learn/process-data/lecture/mSj5A/determine-the-best-sample-size) video for a demonstration of a sample size calculator, or refer to the [Sample Size Calculator](https://www.coursera.org/learn/process-data/supplement/ZqcDw/sample-size-calculator) reading for additional information.



## Key takeaways

As you continue on your data analytics journey, be sure to familiarize yourself with key terms including population, sample, margin of error, confidence level, and confidence interval before calculating sample size. Remember that a minimum sample size of 30 is recommended and that sample size varies depending on the specific business problem. Also consider the trade-off between accuracy and cost when determining sample size, as larger sample sizes provide more accurate results but at a higher cost. Finally, use sample size calculators to determine the appropriate sample size for your study.

### 1.

Question 1

What are some strategies data professionals can use when they do not have enough data to meet a business objective? Select all that apply.

1 point

Consider whether it is possible to adjust the objective.

Locate another relevant dataset to work with.

Use smaller-scale data until they can find more complete data.

Use hypothetical data that aligns with their own predictions.

### 2.

Question 2

Which of the following are limitations that might lead to insufficient data? Select all that apply.

1 point

Duplicate data

Data that updates continually

Data from a single source

Outdated data

### 3.

Question 3

A data analyst wants to find out how many middle school students in Helsinki have laptops. It is unlikely that they can survey every middle schooler in the city. Instead, they survey enough students to represent all middle schoolers. This describes what data analytics concept?

1 point

Using a sample

Confidence level

Statistical significance

Margin of error

### 4.

Question 4

Fill in the blank: Sampling \_\_\_\_\_ occurs when some members of a population are overrepresented or underrepresented in the data.

1 point

integrity

duplication

randomization

bias

Hey, there.

We've all probably dreamed of having a superpower at least once in our lives.

I know I have. I'd love to be able to fly.

But there's another superpower you might not have heard of: statistical power.

Play video starting at ::14 and follow transcript0:14

Statistical power is the probability of getting meaningful results from a test.

I'm guessing that's not a superpower any of you have dreamed about.

Still, it's a pretty great data superpower.

For data analysts, your projects might begin with the test or study.

Hypothesis testing is a way to see if a survey or

experiment has meaningful results.

Here's an example.

Let's say you work for a restaurant chain that's planning a marketing campaign for

their new milkshakes.

You need to test the ad on a group of customers before turning it into a

nationwide ad campaign.

Play video starting at ::50 and follow transcript0:50

In the test, you want to check whether customers like or dislike the campaign.

You also want to rule out any factors outside of the ad that might lead

them to say they don't like it.

Play video starting at :1:1 and follow transcript1:01

Using all your customers would be too time consuming and expensive.

So, you'll need to figure out how many customers you'll need to show that the ad

is effective.

Fifty probably wouldn't be enough.

Even if you randomly chose 50 customers,

you might end up with customers who don't like milkshakes at all.

And if that happens, you won't be able to measure the effectiveness of your ad in

getting more milkshake orders since no one in the sample size would order them.

That's why you need a larger sample size:

so you can make sure you get a good number of all types of people for your test.

Usually, the larger the sample size, the greater the chance you'll have

statistically significant results with your test.

And that's statistical power.

Play video starting at :1:46 and follow transcript1:46

In this case, using as many customers as possible will show the actual differences

between the groups who like or

dislike the ad versus people whose decision wasn't based on the ad at all.

Play video starting at :1:58 and follow transcript1:58

There are ways to accurately calculate statistical power, but

we won't go into them here.

You might need to calculate it on your own as a data analyst.

Play video starting at :2:8 and follow transcript2:08

For now, you should know that statistical power is usually shown as a value

out of one.

So if your statistical power is 0.6, that's the same thing as saying 60%.

In the milk shake ad test,

if you found a statistical power of 60%, that means there's a 60% chance

of you getting a statistically significant result on the ad's effectiveness.

Play video starting at :2:32 and follow transcript2:32

"Statistically significant" is a term that is used in statistics.

If you want to learn more about the technical meaning, you can search online.

But in basic terms, if a test is statistically significant,

it means the results of the test are real and not an error caused by random chance.

Play video starting at :2:52 and follow transcript2:52

So there's a 60% chance that the results of the milkshake ad

test are reliable and real

and a 40% chance that the result of the test is wrong.

Play video starting at :3:2 and follow transcript3:02

Usually, you need a statistical power of at least 0.8 or

80% to consider your results statistically significant.

Play video starting at :3:11 and follow transcript3:11

Let's check out one more scenario.

We'll stick with milkshakes because, well, because I like milkshakes.

Imagine you work for a restaurant chain that wants to launch a brand-new birthday

cake flavored milkshake.

Play video starting at :3:24 and follow transcript3:24

This milkshake will be more expensive to produce than your other milkshakes.

Your company hopes that the buzz around the new flavor will bring in more

customers and money to offset this cost.

They want to test this out in a few restaurant locations first.

So let's figure out how many locations you'd have to use to be confident in your

results.

Play video starting at :3:44 and follow transcript3:44

First, you'd have to think about what might prevent you from getting

statistically significant results.

Are there restaurants running any other promotions that might bring

in new customers?

Do some restaurants have customers that always buy the newest item,

no matter what it is?

Do some location have construction that recently started,

that would prevent customers from even going to the restaurant?

Play video starting at :4:7 and follow transcript4:07

To get a higher statistical power, you'd have to consider all of these factors

before you decide how many locations to include in your sample size for

your study.

Play video starting at :4:16 and follow transcript4:16

You want to make sure any effect is most likely due to the new milkshake flavor,

not another factor.

Play video starting at :4:23 and follow transcript4:23

The measurable effects would be an increase in sales or

the number of customers at the locations in your sample size.

That's it for now.

Coming up, we'll explore sample sizes in more detail, so

you can get a better idea of how they impact your tests and studies.

Play video starting at :4:39 and follow transcript4:39

In the meantime, you've gotten to know a little bit more about milkshakes and

superpowers.

And of course, statistical power.

Sadly, only statistical power can truly be useful for data analysts.

Though putting on my cape and

flying to grab a milkshake right now does sound pretty good.

# When data isn't readily available

Earlier, you learned how you can still do an analysis using proxy data if you have no data. You might have some questions about proxy data, so this reading will give you a few more examples of the types of datasets that can serve as alternate data sources.

## Proxy data examples

Sometimes the data to support a business objective isn’t readily available. This is when proxy data is useful. Take a look at the following scenarios and where proxy data comes in for each example:

| **Business scenario** | **How proxy data can be used** |
| --- | --- |
| A new car model was just launched a few days ago and the auto dealership can’t wait until the end of the month for sales data to come in. They want sales projections now. | The analyst proxies the number of clicks to the car specifications on the dealership’s website as an estimate of potential sales at the dealership. |
| A brand new plant-based meat product was only recently stocked in grocery stores and the supplier needs to estimate the demand over the next four years. | The analyst proxies the sales data for a turkey substitute made out of tofu that has been on the market for several years. |
| The Chamber of Commerce wants to know how a tourism campaign is going to impact travel to their city, but the results from the campaign aren’t publicly available yet. | The analyst proxies the historical data for airline bookings to the city one to three months after a similar campaign was run six months earlier. |

## Open (public) datasets

If you are part of a large organization, you might have access to lots of sources of data. But if you are looking for something specific or a little outside your line of business, you can also make use of open or public datasets. (You can refer to this [Medium article](https://medium.com/thinkdata/is-there-a-difference-between-open-data-and-public-data-2b8d5608b2f1) for a brief explanation of the difference between open and public data.)

Here's an example. A nasal version of a vaccine was recently made available. A clinic wants to know what to expect for contraindications, but just started collecting first-party data from its patients. A **contraindication** is a condition that may cause a patient not to take a vaccine due to the harm it would cause them if taken. To estimate the number of possible contraindications, a data analyst proxies an open dataset from a trial of the injection version of the vaccine. The analyst selects a subset of the data with patient profiles most closely matching the makeup of the patients at the clinic.

There are plenty of ways to share and collaborate on data within a community. Kaggle ([kaggle.com](https://www.kaggle.com/)) which we previously introduced, has datasets in a variety of formats including the most basic type, Comma Separated Values (CSV) files.



### **CSV, JSON, SQLite, and BigQuery datasets**

* CSV: Check out this [Credit card customers](https://www.kaggle.com/sakshigoyal7/credit-card-customers) dataset, which has information from 10,000 customers including age, salary, marital status, credit card limit, credit card category, etc. (CC0: Public Domain, Sakshi Goyal).
* JSON: Check out this JSON dataset for [trending YouTube videos](https://www.kaggle.com/datasnaek/youtube-new) (CC0: Public Domain, Mitchell J).
* SQLite: Check out this SQLite dataset for 24 years worth of [U.S. wildfire data](https://www.kaggle.com/rtatman/188-million-us-wildfires) (CC0: Public Domain, Rachael Tatman).
* BigQuery: Check out this [Google Analytics 360](https://www.kaggle.com/bigquery/google-analytics-sample) sample dataset from the Google Merchandise Store (CC0 Public Domain, Google BigQuery).

Refer to the Kaggle [documentation for datasets](https://www.kaggle.com/docs/datasets) for more information and search for and explore datasets on your own at [kaggle.com/datasets](https://www.kaggle.com/datasets).

As with all other kinds of datasets, be on the lookout for duplicate data and ‘Null’ in open datasets. Null most often means that a data field was unassigned (left empty), but sometimes Null can be interpreted as the value, 0. It is important to understand how Null was used before you start analyzing a dataset with Null data.

## ****Key takeaways****

As you work on data analysis projects, proxy data can often be used to estimate or predict outcomes when actual data is not available. Open or public datasets can be used as proxy data sources, and there are many available online repositories for finding relevant datasets. But be cautious when using proxy data and ensure that it is well-suited for the intended purpose. Finally, check for duplicate data and null values in open datasets before using them for analysis.

Great to see you again.

In this video, we'll go into more detail

about sample sizes and data integrity.

If you've ever been to a store that hands out samples,

you know it's one of life's little pleasures.

For me, anyway!

those small samples are also a very smart way for

businesses to learn more about their products from

customers without having to give everyone a free sample.

A lot of organizations use sample size in a similar way.

They take one part of something larger.

In this case, a sample of a population.

Sometimes they'll perform complex tests on

their data to see if it meets their business objectives.

We won't go into all the calculations

needed to do this effectively.

Instead, we'll focus on a "big picture"

look at the process and what it involves.

As a quick reminder,

sample size is a part of

a population that is representative of the population.

For businesses, it's a very important tool.

It can be both expensive and

time-consuming to analyze an entire population of data.

Using sample size usually makes

the most sense and can still

lead to valid and useful findings.

There are handy calculators

online that can help you find sample size.

You need to input the confidence level,

population size, and margin of error.

We've talked about population size before.

To build on that, we'll learn about

confidence level and margin of error.

Knowing about these concepts will help you

understand why you need them to calculate sample size.

The confidence level is the probability that

your sample accurately reflects the greater population.

You can think of it the same

way as confidence in anything else.

It's how strongly you feel that you

can rely on something or someone.

Having a 99 percent confidence level is ideal.

But most industries hope for at least a

90 or 95 percent confidence level.

Industries like pharmaceuticals

usually want a confidence level

that's as high as possible when

they are using a sample size.

This makes sense because

they're testing medicines and need

to be sure they work and are safe for everyone to use.

For other studies,

organizations might just need to know that

the test or survey results

have them heading in the right direction.

For example, if a paint company

is testing out new colors,

a lower confidence level is okay.

You also want to consider

the margin of error for your study.

You'll learn more about this soon,

but it basically tells you how

close your sample size results are to what

your results would be if you use

the entire population that your sample size represents.

Think of it like this.

Let's say that the principal of

a middle school approaches you

with a study about students' candy preferences.

They need to know an appropriate

sample size, and they need it now.

The school has a student population of 500,

and they're asking for a confidence level of

95 percent and a margin of error of 5 percent.

We've set up a calculator in a spreadsheet,

but you can also easily find this type of calculator by

searching "sample size calculator" on the internet.

Just like those calculators,

our spreadsheet calculator doesn't show any of

the more complex calculations

for figuring out sample size.

All we need to do is input

the numbers for our population,

confidence level, and margin of error.

And when we type 500 for our population size,

95 for our confidence level percentage,

5 for our margin of error percentage,

the result is about 218.

That means for this study,

an appropriate sample size would be 218.

If we surveyed 218 students and

found that 55 percent of them preferred chocolate,

then we could be pretty confident that

would be true of all 500 students.

218 is the minimum number

of people we need to survey based on

our criteria of a 95 percent confidence level

and a 5 percent margin of error.

In case you're wondering,

the confidence level and margin of error

don't have to add up to 100 percent.

They're independent of each other.

So let's say we change our margin of error

from 5 percent to 3 percent.

Then we find that our sample size

would need to be larger,

about 341 instead of 218,

to make the results of the study

more representative of the population.

Feel free to practice with an online calculator.

Knowing sample size and how to

find it will help you when you work with data.

We've got more useful knowledge coming your way,

including learning about margin of error. See you soon!

# Sample size calculator

In this reading, you will learn the basics of sample size calculators, how to use them, and how to understand the results. A **sample size calculator** tells you how many people you need to interview (or things you need to test) to get results that represent the target population. Let’s review some terms you will come across when using a sample size calculator:

* **Confidence level**: The probability that your sample size accurately reflects the greater population.
* **Margin of error**: The maximum amount that the sample results are expected to differ from those of the actual population.
* **Population**: This is the total number you hope to pull your sample from.
* **Sample**: A part of a population that is representative of the population.
* **Estimated response rate**: If you are running a survey of individuals, this is the percentage of people you expect will complete your survey out of those who received the survey.

## How to use a sample size calculator

In order to use a sample size calculator, you need to have the population size, confidence level, and the acceptable margin of error already decided so you can input them into the tool. If this information is ready to go, check out these sample size calculators below:

* [Sample size calculator by surveymonkey.com](https://www.surveymonkey.com/mp/sample-size-calculator/)
* [Sample size calculator by raosoft.com](http://www.raosoft.com/samplesize.html)

## What to do with the results

After you have plugged your information into one of these calculators, it will give you a recommended sample size. Keep in mind, the calculated sample size is the **minimum** number to achieve what you input for confidence level and margin of error. If you are working with a survey, you will also need to think about the estimated response rate to figure out how many surveys you will need to send out. For example, if you need a sample size of 100 individuals and your estimated response rate is 10%, you will need to send your survey to 1,000 individuals to get the 100 responses you need for your analysis.

Now that you have the basics, try some calculations using the sample size calculators and refer back to this reading if you need a refresher on the definitions.

### 1.

Question 1

Fill in the blank: Hypothesis testing is a way to see if a survey or experiment has \_\_\_\_\_ results.

1 point

meaningful

insufficient

random

probable

### 2.

Question 2

A research team conducts an experiment to determine if a new cybersecurity tool is more effective than the previous version. What type of results are required for the experiment to be statistically significant?

1 point

Results that are real and not caused by random chance

Results that are inaccurate and should be ignored

Results that are hypothetical and in need of more testing

Results that are unlikely to occur again

### 3.

Question 3

In order to have a high confidence level in a customer survey, what should the sample size accurately reflect?

1 point

The trends from previous customer surveys

The predictions of stakeholders

The entire population

The most valuable members of the population

### 4.

Question 4

Fill in the blank: Typically, a data professional aims to achieve a statistical power of at least \_\_\_\_\_ to consider their results statistically significant.

1 point

0.8, or 80%

0.1, or 10%

1, or 100%

0.6, or 60%

Hey there! Earlier, we touched on margin of error without explaining it completely.

Well, we're going to right that wrong in this video by explaining margin of

error more.

We'll even include an example of how to calculate it.

Play video starting at ::14 and follow transcript0:14

As a data analyst, it's important for you to figure out sample size and variables

like confidence level and margin of error before running any kind of test or survey.

It's the best way to make sure your results are objective, and

it gives you a better chance of getting statistically significant results.

But if you already know the sample size, like when you're given survey results

to analyze, you can calculate the margin of error yourself.

Then you'll have a better idea of how much of a difference there is between your sample and

your population.

We'll start at the beginning with a more complete definition.

Margin of error is the maximum that the sample results are expected

to differ from those of the actual population.

Play video starting at :1: and follow transcript1:00

Let's think about an example of margin of error.

Play video starting at :1:3 and follow transcript1:03

It would be great to survey or test an entire population,

but it's usually impossible or impractical to do this.

So instead, we take a sample of the larger population.

Play video starting at :1:15 and follow transcript1:15

Based on the sample size,

the resulting margin of error will tell us how different the results might be

compared to the results if we had surveyed the entire population.

Play video starting at :1:27 and follow transcript1:27

Margin of error helps you understand how reliable the data from your hypothesis

testing is.

Play video starting at :1:33 and follow transcript1:33

The closer to zero the margin of error, the closer your results from your

sample would match results from the overall population.

Play video starting at :1:43 and follow transcript1:43

For example,

let's say you completed a nationwide survey using a sample of the population.

You asked people who work five-day workweeks whether they like the idea of

a four-day workweek.

So your survey tells you that 60% prefer a four-day workweek.

The margin of error was 10%,

which tells us that between 50 and 70% like the idea.

So if we were to survey all five-day workers nationwide,

between 50 and 70% would agree with our results.

Play video starting at :2:18 and follow transcript2:18

Keep in mind that our range is between 50 and 70%.

That's because the margin of error is counted in both directions

from the survey results of 60%.

If you set up a 95% confidence level for your survey,

there'll be a 95% chance that the entire population's responses will

fall between 50 and 70% saying, yes, they want a four-day workweek.

Play video starting at :2:46 and follow transcript2:46

Since your margin of error overlaps with that 50% mark, you can't say for

sure that the public likes the idea of a four-day workweek.

In that case, you'd have to say your survey was inconclusive.

Play video starting at :3: and follow transcript3:00

Now, if you wanted a lower margin of error, say 5%,

with a range between 55 and 65%, you could increase the sample size.

But if you've already been given the sample size,

you can calculate the margin of error yourself.

Play video starting at :3:18 and follow transcript3:18

Then you can decide yourself how much of a chance your results have

of being statistically significant based on your margin of error.

In general, the more people you include in your survey,

the more likely your sample is representative of the entire population.

Play video starting at :3:34 and follow transcript3:34

Decreasing the confidence level would also have the same effect,

but that would also make it less likely that your survey is accurate.

Play video starting at :3:43 and follow transcript3:43

So to calculate margin of error, you need three things:

population size, sample size, and confidence level.

Play video starting at :3:52 and follow transcript3:52

And just like with sample size,

you can find lots of calculators online by searching "margin of error calculator."

Play video starting at :3:59 and follow transcript3:59

But we'll show you in a spreadsheet,

just like we did when we calculated sample size.

Play video starting at :4:5 and follow transcript4:05

Lets say you're running a study on the effectiveness of a new drug.

You have a sample size of 500 participants

whose condition affects 1% of the world's population.

That's about 80 million people, which is the population for your study.

Play video starting at :4:23 and follow transcript4:23

Since it's a drug study, you need to have a confidence level of 99%.

You also need a low margin of error.

Let's calculate it.

We'll put the numbers for population,

confidence level, and

sample size, in the appropriate spreadsheet cells.

And our result is a margin of error of close to 6%, plus or minus.

When the drug study is complete, you'd apply the margin of error to your results

to determine how reliable your results might be.

Play video starting at :5:4 and follow transcript5:04

Calculators like this one in the spreadsheet are just one of the many tools

you can use to ensure data integrity.

Play video starting at :5:11 and follow transcript5:11

And it's also good to remember that checking for data integrity and aligning

the data with your objectives will put you in good shape to complete your analysis.

Play video starting at :5:21 and follow transcript5:21

Knowing about sample size, statistical power, margin of error, and

other topics we've covered will help your analysis run smoothly.

That's a lot of new concepts to take in.

If you'd like to review them at any time,

you can find them all in the glossary, or feel free to rewatch the video!

Soon you'll explore the ins and outs of clean data.

The data adventure keeps moving!

I'm so glad you're moving along with it.

You got this!

# All about margin of error

**Margin of error** is the maximum amount that the sample results are expected to differ from those of the actual population. More technically, the margin of error defines a range of values below and above the average result for the sample. The average result for the entire population is expected to be within that range. We can better understand margin of error by using some examples below.

## Margin of error in baseball



Imagine you are playing baseball and that you are up at bat. The crowd is roaring, and you are getting ready to try to hit the ball. The pitcher delivers a fastball traveling about 90-95mph, which takes about 400 milliseconds (ms) to reach the catcher’s glove. You swing and miss the first pitch because your timing was a little off. You wonder if you should have swung slightly earlier or slightly later to hit a home run. That time difference can be considered the margin of error, and it tells us how close or far your timing was from the average home run swing.

## Margin of error in marketing

The margin of error is also important in marketing. Let’s use A/B testing as an example. **A/B testing** (or split testing) tests two variations of the same web page to determine which page is more successful in attracting user traffic and generating revenue. User traffic that gets monetized is known as the **conversion rate**. A/B testing allows marketers to test emails, ads, and landing pages to find the data behind what is working and what isn’t working. Marketers use the **confidence interval** (determined by the conversion rate and the margin of error) to understand the results.

For example, suppose you are conducting an A/B test to compare the effectiveness of two different email subject lines to entice people to open the email. You find that subject line A: “Special offer just for you” resulted in a 5% open rate compared to subject line B: “Don’t miss this opportunity” at 3%.

Does that mean subject line A is better than subject line B? It depends on your margin of error. If the margin of error was 2%, then subject line A’s actual open rate or confidence interval is somewhere between 3% and 7%. Since the lower end of the interval overlaps with subject line B’s results at 3%, you can’t conclude that there is a statistically significant difference between subject line A and B. Examining the margin of error is important when making conclusions based on your test results.

## Want to calculate your margin of error?

All you need is population size, confidence level, and sample size. In order to better understand this calculator, review these terms:

* **Confidence level**: A percentage indicating how likely your sample accurately reflects the greater population
* **Population**: The total number you pull your sample from
* **Sample**: A part of a population that is representative of the population
* **Margin of error**: The maximum amount that the sample results are expected to differ from those of the actual population

In most cases, a 90% or 95% confidence level is used. But, depending on your industry, you might want to set a stricter confidence level. A 99% confidence level is reasonable in some industries, such as the pharmaceutical industry.

After you have settled on your population size, sample size, and confidence level, plug the information into a margin of error calculator like the ones below:

* [Margin of error calculator by Good Calculators (free online calculators)](https://goodcalculators.com/margin-of-error-calculator/)
* [Margin of error calculator by CheckMarket](https://www.checkmarket.com/sample-size-calculator/#sample-size-margin-of-error-calculator)

## Key takeaways

Margin of error is used to determine how close your sample’s result is to what the result would likely have been if you could have surveyed or tested the entire population. Margin of error helps you understand and interpret survey or test results in real-life.  Calculating the margin of error is particularly helpful when you are given the data to analyze. After using a calculator to calculate the margin of error, you will know how much the sample results might differ from the results of the entire population.

### .

Question 1

Fill in the blank: Margin of error is the \_\_\_\_\_ amount that the sample results are expected to differ from those of the actual population.

1 point

maximum

average

median

minimum

### 2.

Question 2

What elements are required to calculate margin of error? Select all that apply.

1 point

Population size

Sample size

Mean calculation

Confidence level

### 3.

Question 3

In a survey about a new gardening product, 80% of respondents report they would buy the product again. The margin of error for the survey is 5%. Based on that margin of error, what range reflects the population's true response?

1 point

75-85%

73-78%

70-80%

75-80%

### 4.

Question 4

In an employee satisfaction survey, 60% of respondents report that they prefer commuting to work via train. The margin of error for the survey is 4%. Based on that margin of error, what range reflects the population's true response?

1 point

64-68%

50-60%

56-64%

46-54%

# Glossary terms from module 1

## ****Terms and definitions for Course 4, Module 1****

**Accuracy:** The degree to which the data conforms to the actual entity being measured or described

**Completeness:** The degree to which the data contains all desired components or measures

**Confidence interval:**  A range of values that conveys how likely a statistical estimate reflects the population

**Confidence level:** The probability that a sample size accurately reflects the greater population

**Consistency:** The degree to which data is repeatable from different points of entry or collection

**Cross-field validation:** A process that ensures certain conditions for multiple data fields are satisfied

**Data constraints:** The criteria that determine whether a piece of a data is clean and valid

**Data integrity:** The accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle

**Data manipulation:** The process of changing data to make it more organized and easier to read

**Data range:** Numerical values that fall between predefined maximum and minimum values

**Data replication:** The process of storing data in multiple locations

**DATEDIF:** A spreadsheet function that calculates the number of days, months, or years between two dates

**Estimated response rate**: The average number of people who typically complete a survey

**Hypothesis testing:** A process to determine if a survey or experiment has meaningful results

**Mandatory:** A data value that cannot be left blank or empty

**Margin of error**: The maximum amount that the sample results are expected to differ from those of the actual population

**Random sampling:** A way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen

**Regular expression (RegEx):** A rule that says the values in a table must match a prescribed pattern

### 1.

Question 1

Fill in the blank: If a test is statistically significant, the results are less likely to be due to \_\_\_\_\_ and more likely to be due to a real difference between the groups being compared.

1 point

insufficient data

bias

random chance

causation

### 2.

Question 2

A delivery company stores customer data in a delivery database and a customer relationship database. When a customer changes their shipping address, the change is not reflected in the delivery database. This leads to the shipments being delivered to the wrong address. What data integrity problem does this scenario describe?

1 point

Replication

Manipulation

Transfer

Gathering

### 3.

Question 3

In a survey about a new type of sustainable material, 55% of respondents report they would be willing to pay more for clothes made from this material. The margin of error for the survey is 3%. Based on that margin of error, what range reflects the population's true response?

1 point

50-53%

52-58%

50-52%

58-60%

### 4.

Question 4

A candy manufacturer conducts a survey to learn more about its customer base. Although young people are known to purchase a large percentage of its candy, due to age requirements, the survey is only sent to customers who are 18 years or older. What is likely to result?

1 point

Sampling bias

Random sampling

Unbiased sampling

Geographically limited sampling

### 5.

Question 5

Fill in the blank: To determine whether a survey or experiment has meaningful \_\_\_\_\_, a data team uses hypothesis testing.

1 point

results

significance

action items

process steps

### 6.

Question 6

A data professional in the logistics industry wants to calculate the margin of error for a study about transportation route efficiency. They know the sample size and confidence level. What must they also know in order to accurately calculate margin of error?

1 point

Distribution

Correlation

Population size

Testing methodology

### 7.

Question 7

A data professional copies a dataset from a USB drive to their computer. They accidentally unplug the USB before the process is complete, which causes the dataset on their computer to be incomplete. What data integrity problem does this scenario describe?

1 point

Transfer

Replication

Cleaning

Manipulation

### 8.

Question 8

Which of the following statements accurately describe sample size, population, and confidence level? Select all that apply.

1 point

For effective outcomes, a data professional aims for a high confidence level in their sample.

A confidence level of 75% is considered ideal by most industries.

The goal of random sampling is to ensure every possible type of the sample has an equal chance of being chosen.

Using sample size makes it possible to get enough information from a small group within a population to draw conclusions about the whole population.

Module 2

Can you guess what inaccurate or bad data costs businesses every year?

Thousands of dollars? Millions? Billions?

Well, according to IBM,

the yearly cost of poor-quality data is $3.1 trillion in the US alone.

That's a lot of zeros.

Play video starting at ::20 and follow transcript0:20

Now, can you guess the number one cause of poor-quality data?

Play video starting at ::25 and follow transcript0:25

It's not a new system implementation or a computer technical glitch.

The most common factor is actually human error.

Here's a spreadsheet from a law office.

It shows customers, the legal services they bought,

the service order number, how much they paid, and the payment method.

Play video starting at ::49 and follow transcript0:49

Dirty data can be the result of someone typing in a piece of data incorrectly;

Play video starting at ::56 and follow transcript0:56

inconsistent formatting;

Play video starting at :1: and follow transcript1:00

blank fields; or the same piece of data being entered more than once,

which creates duplicates.

Play video starting at :1:12 and follow transcript1:12

Dirty data is data that's incomplete, incorrect, or

irrelevant to the problem you're trying to solve.

Play video starting at :1:19 and follow transcript1:19

When you work with dirty data, you can't be sure that your results are correct.

In fact, you can pretty much bet they won't be.

Earlier, you learned that data integrity is critical to reliable data

analytics results, and clean data helps you achieve data integrity.

Clean data is data that's complete, correct, and

relevant to the problem you're trying to solve.

When you work with clean data, you'll find that your projects go much more smoothly.

Play video starting at :1:47 and follow transcript1:47

I remember the first time I witnessed firsthand how important clean data

really is.

I had just started using SQL, and I thought it worked like magic.

I could have the computer sum up millions of numbers, saving me tons of time and

effort.

But I quickly discovered that only works when the data is clean.

If there was even one accidental letter in a column that should only have numbers,

the computer wouldn't know what to do.

So, it would throw an error, and suddenly, I'm stuck.

And there's no way I could add up millions of numbers by myself.

So, I have to clean out that data to make it work.

The good news is that there's plenty of effective processes and

tools to help you do that.

Coming up, you'll gain the skills and

knowledge you need to make sure the data you work with is always clean.

Play video starting at :2:34 and follow transcript2:34

Along the way, we'll dig deeper into the difference between clean and dirty data,

and why clean data is so important.

We'll also talk about different ways to clean your data and

common problems to look for during the process.

Play video starting at :2:47 and follow transcript2:47

Ready to start?

Let's do it.

Clean data is incredibly

important for effective analysis.

If a piece of data is entered into

a spreadsheet or database incorrectly,

or if it's repeated,

or if a field is left blank,

or if data formats are inconsistent,

the result is dirty data.

Small mistakes can lead to

big consequences in the long run.

I'll be completely honest with you,

data cleaning is like brushing your teeth.

It's something you should do and do

properly because otherwise it can cause serious problems.

For teeth, that might be cavities or gum disease.

For data, that might be costing

your company money, or an angry boss.

But here's the good news.

If you keep brushing twice a day,

every day, it becomes a habit.

Soon, you don't even have to think about it.

It's the same with data.

Trust me, it will make you look great

when you take the time to clean up that dirty data.

As a quick refresher,

dirty data is incomplete,

incorrect, or irrelevant to

the problem you're trying to solve.

It can't be used in a meaningful way, which

makes analysis very difficult, if not impossible.

On the other hand, clean data is complete,

correct, and relevant to

the problem you're trying to solve.

This allows you to understand and analyze

information and identify important patterns,

connect related information, and draw useful conclusions.

Then you can apply what you learn

to make effective decisions.

In some cases, you won't have

to do a lot of work to clean data.

For example, when you use

internal data that's been verified

and cared for by

your company's data engineers and data warehouse team,

it's more likely to be clean.

Let's talk about some people you'll

work with as a data analyst.

Data engineers transform data into

a useful format for

analysis and give it a reliable infrastructure.

This means they develop, maintain,

and test databases, data processors and related systems.

Data warehousing specialists develop processes

and procedures to effectively store and organize data.

They make sure that data is available,

secure, and backed up to prevent loss.

When you become a data analyst,

you can learn a lot by

working with the person who maintains

your databases to learn about their systems.

If data passes through the hands of

a data engineer or a data warehousing specialist first,

you know you're off to a good start on your project.

There's a lot of great career opportunities as

a data engineer or a data warehousing specialist.

If this kind of work sounds interesting to you,

maybe your career path will involve

helping organizations save lots of time,

effort, and money by making

sure their data is sparkling clean.

But even if you go in

a different direction with your data analytics career

and have the advantage of working with

data engineers and warehousing specialists,

you're still likely to have to clean your own data.

It's important to remember:

no dataset is perfect.

It's always a good idea to examine and

clean data before beginning analysis.

Here's an example. Let's say

you're working on a project where you

need to figure out how many people

use your company's software program.

You have a spreadsheet that was created internally and

verified by a data engineer

and a data warehousing specialist.

Check out the column labeled "Username."

It might seem logical that you can just scroll down and

count the rows to figure out how many users you have.

Play video starting at :3:40 and follow transcript3:40

But that won't work because one person

sometimes has more than one username.

Play video starting at :3:49 and follow transcript3:49

Maybe they registered from different email addresses,

or maybe they have a work and personal account.

In situations like this,

you would need to clean the data by

eliminating any rows that are duplicates.

Play video starting at :4:6 and follow transcript4:06

Once you've done that, there won't

be any more duplicate entries.

Then your spreadsheet is ready to be put to work.

So far we've discussed working with internal data.

But data cleaning becomes even

more important when working with

external data, especially if

it comes from multiple sources.

Let's say the software company from our example surveyed

its customers to learn how

satisfied they are with its software product.

But when you review the survey data,

you find that you have several nulls.

Play video starting at :4:40 and follow transcript4:40

A null is an indication that

a value does not exist in a data set.

Note that it's not the same as a zero.

In the case of a survey,

a null would mean the customers skipped that question.

A zero would mean they provided zero as their response.

To do your analysis,

you would first need to clean this data.

Step one would be to decide what to do with those nulls.

You could either filter them out and

communicate that you now have a smaller sample size,

or you can keep them in and learn from the fact

that the customers did not provide responses.

There's lots of reasons why this could have happened.

Maybe your survey questions

weren't written as well as they could be.

Maybe they were confusing or biased,

something we learned about earlier.

We've touched on the basics of

cleaning internal and external data,

but there's lots more to come.

Soon, we'll learn about the common errors to be

aware of to ensure your data is complete,

correct, and relevant. See you soon!

I am Angie. I'm a program manager

of engineering at Google.

I truly believe that cleaning Data

is the heart and soul of data.

It's how you get to know your data:

its quirks, its flaws, it's mysteries.

I love a good mystery.

I remember one time I found somebody had purchased,

I think it was one million dollars

worth of chicken sandwiches in one transaction.

This mystery drove me nuts.

I had all these questions.

Could this have really happened?

Or maybe it was a really big birthday party?

How did they make a million dollars

worth of chicken sandwiches?

I was cleaning my data

and trying to figure out where did it go wrong.

We ended up finding out that we'd been squaring and

multiplying all of our transactions

for a very specific case.

It took us about three days to figure this out.

I will never forget the moment when it was like,

aha! We got to the bottom of it.

The result is our data was cleaned, and we

had this great dataset that we could use for analysis.

But what I loved was just the mystery of it

and getting to know

all these weird intricacies about my dataset.

It felt like a superpower almost,

like I was a detective,

and I had gone in there and I had really

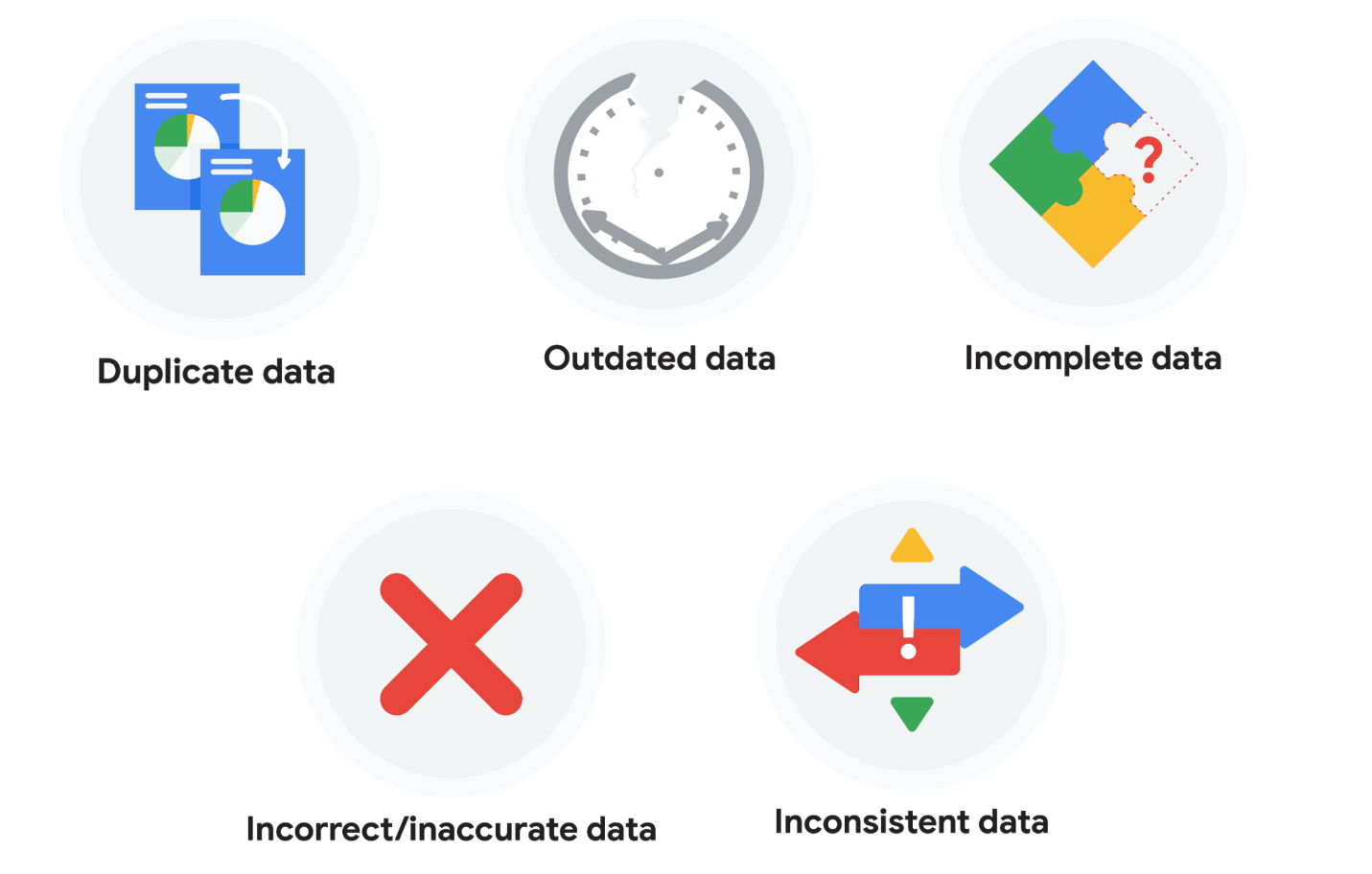
solved something. I love cleaning data!

# What is dirty data?

Earlier, we discussed that **dirty data** is data that is incomplete, incorrect, or irrelevant to the problem you are trying to solve.  This reading summarizes:

* Types of dirty data you may encounter
* What may have caused the data to become dirty
* How dirty data is harmful to businesses

## Types of dirty data



### **Duplicate data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data record that shows up more than once | Manual data entry, batch data imports, or data migration | Skewed metrics or analyses, inflated or inaccurate counts or predictions, or confusion during data retrieval |

### **Outdated data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that is old which should be replaced with newer and more accurate information | People changing roles or companies, or software and systems becoming obsolete | Inaccurate insights, decision-making, and analytics |

### **Incomplete data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that is missing important fields | Improper data collection or incorrect data entry | Decreased productivity, inaccurate insights, or inability to complete essential services |

### **Incorrect/inaccurate data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that is complete but inaccurate | Human error inserted during data input, fake information, or mock data | Inaccurate insights or decision-making based on bad information resulting in revenue loss |

### **Inconsistent data**

| **Description** | **Possible causes** | **Potential harm to businesses** |
| --- | --- | --- |
| Any data that uses different formats to represent the same thing | Data stored incorrectly or errors inserted during data transfer | Contradictory data points leading to confusion or inability to classify or segment customers |

### **Business impact of dirty data**

For further reading on the business impact of dirty data, enter the term “dirty data” into your preferred browser’s search bar to bring up numerous articles on the topic. Here are a few impacts cited for certain industries from a previous search:

* **Banking**: Inaccuracies cost companies between 15% and 25% of revenue ([source](https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/)).
* **Digital commerce:** Up to 25% of B2B database contacts contain inaccuracies ([source](https://www.demandgen.com/dirty-data-what-is-it-costing-you/)).
* **Marketing and sales**: 99% of companies are actively tackling data quality in some way ([source](https://www.dqglobal.com/blog/why-bad-data-is-wasting-your-marketing-efforts/)).
* **Healthcare**: Duplicate records can be 10% and even up to 20% of a hospital’s electronic health records ([source](https://searchhealthit.techtarget.com/feature/Hospitals-battle-duplicate-medical-records-with-technology)).

## Key takeaways

Dirty data includes duplicate data, outdated data, incomplete data, incorrect or inaccurate data, and inconsistent data. Each type of dirty data can have a significant impact on analyses, leading to inaccurate insights, poor decision-making, and revenue loss. There are a number of causes of dirty data, including manual data entry errors, batch data imports, data migration, software obsolescence, improper data collection, and human errors during data input. As a data professional, you can take steps to mitigate the impact of dirty data by implementing effective data quality processes.

Hey, there. In this video,

we'll focus on common issues associated with dirty data.

These includes spelling and other texts errors,

inconsistent labels, formats and field lane,

missing data and duplicates.

This will help you recognize problems

quicker and give you the information you

need to fix them when you

encounter something similar during your own analysis.

This is incredibly important in data analytics.

Let's go back to our law office spreadsheet.

As a quick refresher,

we'll start by checking out

the different types of dirty data it shows.

Sometimes, someone might

key in a piece of data incorrectly.

Other times, they might not

keep data formats consistent.

Play video starting at ::46 and follow transcript0:46

It's also common to leave a field blank.

Play video starting at ::51 and follow transcript0:51

That's also called a null,

which we learned about earlier.

If someone adds the same piece of data more than once,

that creates a duplicate.

Play video starting at :1:4 and follow transcript1:04

Let's break that down.

Then we'll learn about a few other types of

dirty data and strategies for cleaning it.

Misspellings, spelling variations, mixed up letters,

inconsistent punctuation, and typos in general,

happen when someone types in a piece of data incorrectly.

As a data analyst,

you'll also deal with different currencies.

For example, one dataset could be in

US dollars and another in euros,

and you don't want to get them mixed up.

We want to find these types of

errors and fix them like this.

Play video starting at :1:41 and follow transcript1:41

You'll learn more about this soon.

Clean data depends largely on

the data integrity rules that an organization follows,

such as spelling and punctuation guidelines.

For example, a beverage company

might ask everyone working in

its database to enter data about

volume in fluid ounces instead of cups.

It's great when an organization

has rules like this in place.

It really helps minimize

the amount of data cleaning required,

but it can't eliminate it completely.

Like we discussed earlier,

there's always the possibility of human error.

The next type of dirty data

our spreadsheet shows is inconsistent formatting.

In this example,

something that should be formatted as

currency is shown as a percentage.

Until this error is fixed,

like this, the law office

will have no idea how much money

this customer paid for its services.

We'll learn about different ways to solve

this and many other problems soon.

We discussed nulls previously,

but as a reminder,

nulls are empty fields.

This kind of dirty data requires

a little more work than just fixing

a spelling error or changing a format.

In this example,

the data analysts would need

to research which customer had

a consultation on July 4th, 2020.

Then when they find the correct information,

they'd have to add it to the spreadsheet.

Play video starting at :3:19 and follow transcript3:19

Another common type of dirty data is duplicated.

Play video starting at :3:27 and follow transcript3:27

Maybe two different people added

this appointment on August 13th,

not realizing that someone else had already done it or

maybe the person entering the data

hit copy and paste by accident.

Whatever the reason, it's the data analyst job to

identify this error and correct

it by deleting one of the duplicates.

Play video starting at :3:51 and follow transcript3:51

Now, let's continue on to

some other types of dirty data.

The first has to do with labeling.

To understand labeling,

imagine trying to get a computer to correctly identify

panda bears among images

of all different kinds of animals.

You need to show the computer thousands

of images of panda bears.

They're all labeled as panda bears.

Any incorrectly labeled picture,

like the one here that's just bear,

will cause a problem.

The next type of dirty data is having

an inconsistent field length.

You learned earlier that a field is a single piece of

information from a row or column of a spreadsheet.

Field length is a tool for determining

how many characters can be keyed into a field.

Assigning a certain length to the fields in

your spreadsheet is a great way to avoid errors.

For instance, if you have

a column for someone's birth year,

you know the field length is four

because all years are four digits long.

Some spreadsheet applications have

a simple way to specify field lengths

and make sure users can only enter

a certain number of characters into a field.

This is part of data validation.

Data validation is a tool for checking

the accuracy and quality of

data before adding or importing it.

Data validation is a form of data cleansing,

which you'll learn more about soon.

But first, you'll get familiar

with more techniques for cleaning data.

This is a very important part of the data analyst job.

I look forward to sharing

these data cleaning strategies with you.

### 1.

Question 1

Which data professionals are most often responsible for ensuring that data is available, secure, and backed up to prevent loss?

1 point

Data analysts

Data engineers

Data warehousing specialists

Data scientists

### 2.

Question 2

Fill in the blank: If a dataset contains \_\_\_\_\_, this is an indication that those values do not exist in the dataset.

1 point

anomalies

outliers

zeroes

nulls

### 3.

Question 3

A data professional works in a spreadsheet column that can only contain six-digit customer ID numbers. They ensure the data points in the column are always exactly six-digits long using which data analytics tool?

1 point

Labels

Filtering

Data formatting

Field length

### 4.

Question 4

Fill in the blank: Data validation is a tool for checking the \_\_\_\_\_ of data before adding or importing it.

1 point

accessibility and openness

origin and age

size and structure

accuracy and quality

Hi. Now that you're familiar with some of the most common types of dirty data,

it's time to clean them up.

As you've learned, clean data is essential to data integrity and

reliable solutions and decisions.

The good news is that spreadsheets have all kinds of tools you can use to get your

data ready for analysis.

The techniques for data cleaning will be different depending on the specific data

set you're working with.

So we won't cover everything you might run into, but this will give you a great

starting point for fixing the types of dirty data analysts find most often.

Think of everything that's coming up as a teaser trailer of data cleaning tools.

I'm going to give you a basic overview of some common tools and techniques, and

then we'll practice them again later on.

Here, we'll discuss how to remove unwanted data, clean up text to remove

extra spaces and blanks, fix typos, and make formatting consistent.

However, before removing unwanted data,

it's always a good practice to make a copy of the data set.

That way, if you remove something that you end up needing in the future,

you can easily access it and put it back in the data set.

Once that's done, then you can move on to getting rid of the duplicates or

data that isn't relevant to the problem you're trying to solve.

Typically, duplicates appear when you're combining data sets from more than

one source or using data from multiple departments within the same business.

You've already learned a bit about duplicates, but

let's practice removing them once more now using this spreadsheet,

which lists members of a professional logistics association.

Duplicates can be a big problem for data analysts.

So it's really important that you can find and

remove them before any analysis starts.

Here's an example of what I'm talking about.

Play video starting at :2: and follow transcript2:00

Let's say this association has duplicates of one person's $500

membership in its database.

Play video starting at :2:8 and follow transcript2:08

When the data is summarized,

the analyst would think there was $1,000 being paid by this member

and would make decisions based on that incorrect data.

But in reality, this member only paid $500.

These problems can be fixed manually, but most spreadsheet applications also offer

lots of tools to help you find and remove duplicates.

Play video starting at :2:35 and follow transcript2:35

Now, irrelevant data, which is data that doesn't fit the specific problem that

you're trying to solve, also needs to be removed.

Going back to our association membership list example, let's say a data analyst

was working on a project that focused only on current members.

They wouldn't want to include information on people who are no longer members,

Play video starting at :3:2 and follow transcript3:02

or who never joined in the first place.

Play video starting at :3:11 and follow transcript3:11

Removing irrelevant data takes a little more time and effort because you have

to figure out the difference between the data you need and the data you don't.

But believe me,

making those decisions will save you a ton of effort down the road.

Play video starting at :3:25 and follow transcript3:25

The next step is removing extra spaces and blanks.

Extra spaces can cause unexpected results when you sort, filter, or

search through your data.

And because these characters are easy to miss, they can lead to unexpected and

confusing results.

For example, if there's an extra space and in a member ID number,

when you sort the column from lowest to highest, this row will be out of place.

Play video starting at :3:57 and follow transcript3:57

To remove these unwanted spaces or blank cells, you can delete them yourself.

Play video starting at :4:7 and follow transcript4:07

Or again, you can rely on your spreadsheets,

which offer lots of great functions for removing spaces or blanks automatically.

The next data cleaning step involves fixing misspellings,

inconsistent capitalization, incorrect punctuation, and other typos.

These types of errors can lead to some big problems.

Let's say you have a database of emails that you use to keep in touch with your

customers.

If some emails have misspellings, a period in the wrong place, or

any other kind of typo, not only do you run the risk of sending an email to

the wrong people, you also run the risk of spamming random people.

Think about our association membership example again.

Misspelling might cause the data analyst to miscount the number of

professional members if they sorted this membership type

Play video starting at :5: and follow transcript5:00

and then counted the number of rows.

Play video starting at :5:6 and follow transcript5:06

Like the other problems you've come across,

you can also fix these problems manually.

Play video starting at :5:16 and follow transcript5:16

Or you can use spreadsheet tools, such as spellcheck,

autocorrect, and conditional formatting to make your life easier.

There's also easy ways to convert text to lowercase, uppercase, or proper case,

which is one of the things we'll check out again later.

All right, we're getting there.

The next step is removing formatting.

This is particularly important when you get data from lots of different sources.

Every database has its own formatting,

which can cause the data to seem inconsistent.

Creating a clean and consistent visual appearance for your spreadsheets will help

make it a valuable tool for you and your team when making key decisions.

Most spreadsheet applications also have a "clear formats" tool,

which is a great time saver.

Cleaning data is an essential step in increasing the quality of your data.

Now you know lots of different ways to do that.

In the next video, you'll take that knowledge even further and

learn how to clean up data that's come from more than one source.

Welcome back. So far you've learned a lot about

dirty data and how to clean up

the most common errors in a dataset.

Now we're going to take that a step further and

talk about cleaning up multiple datasets.

Cleaning data that comes from

two or more sources is very common for data analysts,

but it does come with some interesting challenges.

A good example is a merger,

which is an agreement that unites

two organizations into a single new one.

In the logistics field,

there's been lots of big changes recently,

mostly because of the e-commerce boom.

With so many people shopping online,

it makes sense that the companies responsible

for delivering those products to their homes

are in the middle of a big shake-up.

When big things happen in an industry,

it's common for two organizations to

team up and become stronger through a merger.

Let's talk about how that will

affect our logistics association.

As a quick reminder,

this spreadsheet lists association member ID numbers,

first and last names, addresses,

how much each member pays in dues,

when the membership expires,

and the membership types.

Now, let's think about what would happen if

the International Logistics Association decided to get

together with the Global Logistics Association in

order to help their members handle

the incredible demands of e-commerce.

First, all the data from

each organization would need to be

combined using data merging.

Data merging is the process of combining

two or more datasets into a single dataset.

This presents a unique challenge because

when two totally different datasets are combined,

the information is almost

guaranteed to be inconsistent and misaligned.

For example,

the Global Logistics Association's spreadsheet

has a separate column for a person's suite, apartment,

or unit number, but

the International Logistics Association

combines that information with their street address.

This needs to be corrected to make

the number of address columns consistent.

Next, check out how

the Global Logistics Association

uses people's email addresses as their member ID,

while the International Logistics Association

uses numbers.

This is a big problem

because people in a certain industry,

such as logistics, typically

join multiple professional associations.

There's a very good chance that these datasets include

membership information on the exact same person,

just in different ways.

It's super important to remove those duplicates.

Also, the Global Logistics Association

has many more member types than the other organization.

Play video starting at :2:59 and follow transcript2:59

On top of that, it uses a term,

"Young Professional" instead of "Student Associate."

Play video starting at :3:8 and follow transcript3:08

But both describe members who are

still in school or just starting their careers.

If you were merging these two datasets,

you'd need to work with your team to fix the fact that

the two associations

describe memberships very differently.

Now you understand why the merging of

organizations also requires the merging of data,

and that can be tricky.

But there's lots of other reasons why

data analysts merge datasets.

For example, in one of my past jobs,

I merged a lot of data from multiple sources

to get insights about our customers' purchases.

The kinds of insights I gained helped me

identify customer buying patterns.

When merging datasets, I always

begin by asking myself some key questions to

help me avoid redundancy

and to confirm that the datasets are compatible.

In data analytics, compatibility describes

how well two or more datasets are able to work together.

The first question I would ask is,

do I have all the data I need?

To gather customer purchase insights,

I wanted to make sure I had data on customers,

their purchases, and where they shopped.

Next I would ask,

does the data I need exist within these datasets?

As you learned earlier in this program,

this involves considering

the entire dataset analytically.

Looking through the data before I start using

it lets me get a feel for what it's all about,

what the schema looks like,

if it's relevant to my customer purchase insights,

and if it's clean data.

That brings me to the next question.

Do the datasets need to be

cleaned, or are they ready for me to use?

Because I'm working with more than one source,

I will also ask myself,

are the datasets cleaned to the same standard?

For example, what fields are regularly repeated?

How are missing values handled?

How recently was the data updated?

Finding the answers to

these questions and understanding

if I need to fix any problems

at the start of a project is

a very important step in data merging.

In both of the examples we explored here,

data analysts could use

either the spreadsheet tools or SQL queries to clean up,

merge, and prepare the datasets for analysis.

Depending on the tool you decide to use,

the cleanup process can be simple or very complex.

Soon, you'll learn how to make

the best choice for your situation.

As a final note, programming languages like R

are also very useful for cleaning data.

You'll learn more about how to use

R and other concepts we covered soon.

# Common data-cleaning pitfalls

In this reading, you will learn the importance of data cleaning and how to identify common mistakes. Some of the errors you might come across while cleaning your data could include:



## Common mistakes to avoid

* **Not checking for spelling errors**: Misspellings can be as simple as typing or input errors. Most of the time the wrong spelling or common grammatical errors can be detected, but it gets harder with things like names or addresses. For example, if you are working with a spreadsheet table of customer data, you might come across a customer named “John” whose name has been input incorrectly as “Jon” in some places. The spreadsheet’s spellcheck probably won’t flag this, so if you don’t double-check for spelling errors and catch this, your analysis will have mistakes in it.
* **Forgetting to document errors**: Documenting your errors can be a big time saver, as it helps you avoid those errors in the future by showing you how you resolved them. For example, you might find an error in a formula in your spreadsheet. You discover that some of the dates in one of your columns haven’t been formatted correctly. If you make a note of this fix, you can reference it the next time your formula is broken, and get a head start on troubleshooting. Documenting your errors also helps you keep track of changes in your work, so that you can backtrack if a fix didn’t work.
* **Not checking for misfielded values**: A misfielded value happens when the values are entered into the wrong field. These values might still be formatted correctly, which makes them harder to catch if you aren’t careful. For example, you might have a dataset with columns for cities and countries. These are the same type of data, so they are easy to mix up. But if you were trying to find all of the instances of Spain in the country column, and Spain had mistakenly been entered into the city column, you would miss key data points. Making sure your data has been entered correctly is key to accurate, complete analysis.
* **Overlooking missing values**: Missing values in your dataset can create errors and give you inaccurate conclusions. For example, if you were trying to get the total number of sales from the last three months, but a week of transactions were missing, your calculations would be inaccurate.  As a best practice, try to keep your data as clean as possible by maintaining completeness and consistency.
* **Only looking at a subset of the data**: It is important to think about all of the relevant data when you are cleaning. This helps make sure you understand the whole story the data is telling, and that you are paying attention to all possible errors. For example, if you are working with data about bird migration patterns from different sources, but you only clean one source, you might not realize that some of the data is being repeated. This will cause problems in your analysis later on. If you want to avoid common errors like duplicates, each field of your data requires equal attention.
* **Losing track of business objectives**: When you are cleaning data, you might make new and interesting discoveries about your dataset-- but you don’t want those discoveries to distract you from the task at hand. For example, if you were working with weather data to find the average number of rainy days in your city, you might notice some interesting patterns about snowfall, too. That is really interesting, but it isn’t related to the question you are trying to answer right now. Being curious is great! But try not to let it distract you from the task at hand.
* **Not fixing the source of the error:** Fixing the error itself is important. But if that error is actually part of a bigger problem, you need to find the source of the issue. Otherwise, you will have to keep fixing that same error over and over again. For example, imagine you have a team spreadsheet that tracks everyone’s progress. The table keeps breaking because different people are entering different values. You can keep fixing all of these problems one by one, or you can set up your table to streamline data entry so everyone is on the same page. Addressing the source of the errors in your data will save you a lot of time in the long run.
* **Not analyzing the system prior to data cleaning:** If we want to clean our data and avoid future errors, we need to understand the root cause of your dirty data. Imagine you are an auto mechanic. You would find the cause of the problem before you started fixing the car, right? The same goes for data. First, you figure out where the errors come from. Maybe it is from a data entry error, not setting up a spell check, lack of formats, or from duplicates. Then, once you understand where bad data comes from, you can control it and keep your data clean.
* **Not backing up your data prior to data cleaning**: It is always good to be proactive and create your data backup before you start your data clean-up. If your program crashes, or if your changes cause a problem in your dataset, you can always go back to the saved version and restore it. The simple procedure of backing up your data can save you hours of work-- and most importantly, a headache.
* **Not accounting for data cleaning in your deadlines/process**: All good things take time, and that includes data cleaning. It is important to keep that in mind when going through your process and looking at your deadlines. When you set aside time for data cleaning, it helps you get a more accurate estimate for ETAs for stakeholders, and can help you know when to request an adjusted ETA.

## Key takeaways

Data cleaning is essential for accurate analysis and decision-making. Common mistakes to avoid when cleaning data include spelling errors, misfielded values, missing values, only looking at a subset of the data, losing track of business objectives, not fixing the source of the error, not analyzing the system prior to data cleaning, not backing up your data prior to data cleaning, and not accounting for data cleaning in your deadlines/process. By avoiding these mistakes, you can ensure that your data is clean and accurate, leading to better outcomes for your business.

## Additional resources

Refer to these "top ten" lists for data cleaning in Microsoft Excel and Google Sheets to help you avoid the most common mistakes:

* [Top ten ways to clean your data](https://support.microsoft.com/en-us/office/top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19): Review an orderly guide to data cleaning in Microsoft Excel.
* [10 Google Workspace tips to clean up data](https://support.google.com/a/users/answer/9604139?hl=en#zippy=): Learn best practices for data cleaning in Google Sheets.

### 1.

Question 1

To create a clean and consistent visual appearance for a spreadsheet, which tool ensures all font types, sizes, and colors are uniform?

1 point

Autocorrect

Clear formats

Spellcheck

Conditional formatting

### 2.

Question 2

What is the process of combining two or more datasets into a single dataset?

1 point

Data transferring

Data composition

Data validation

Data merging

### 3.

Question 3

Fill in the blank: In data analytics, \_\_\_\_\_ describes how well two or more datasets are able to work together.

1 point

alignment

agreement

compatibility

suitability

### 4.

Question 4

What are some benefits of documenting any errors you find while data cleaning? Select all that apply.

1 point

Keep track of changes

Having a backup of your dataset

Save time by not repeating errors in the future

More efficient troubleshooting

# Step-by-Step guide: Data-cleaning features in spreadsheets

This reading outlines the steps the instructor performs in the next video, [Data-cleaning features in spreadsheets](https://www.coursera.org/learn/process-data/lecture/Ez3u5/data-cleaning-features-in-spreadsheets). In the video, the instructor explains how to use menu options in spreadsheets to fix errors.

Keep this step-by-step guide open as you watch the video. It can serve as a helpful reference if you need additional context or clarification while following the video steps. This is not a graded activity, but you can complete these steps to practice the skills demonstrated in the video.

**What you’ll need**

If you’d like to follow along with the examples in this video, choose a spreadsheet tool. Google Sheets or Excel are recommended.

To access the spreadsheet the instructor uses in this video, click the link to the template to create a copy of the dataset. If you don’t have a Google account, download the data directly from the attachments below.

Link to logistics data: [International Logistics Association Memberships - Data for Cleaning](https://docs.google.com/spreadsheets/d/1jmxXS6ZJEMtaoli5__qApb9LE_nXkU2ysf5c8N1tiQA/template/preview)

Link to cosmetics data: [Cosmetics Inc. - Data for Cleaning](https://docs.google.com/spreadsheets/d/12U9Y4IVAGwml7XWBBgC4j9l0cCjqIZlqJc9vu3jr6Ig/template/preview?resourcekey=0-ds9iuh8tsuB7PwGd2dHMDA#gid=0)

OR

[International Logistics Association Memberships - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/5zlIemyvQtKsj4BGZQEdxA_7f286fde512b4f14b3246a6e68b333e1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1706832000&Signature=f8HqNPfYuHiudRIerbd74td1DApFzMzDMa2JYlidTLd1x6QG-FeA0sg8s7NuPSNlJ4RPyHBVMyko6mp1GQgP2dP--FxGqghOH~SjmBhZXR~pcXwjmeoARmZNXLzmihHjO1kKh6dM0TMAp6Wq~ivq6-VAr4ZdhBrZsLYuHOKpXuY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/5zlIemyvQtKsj4BGZQEdxA_7f286fde512b4f14b3246a6e68b333e1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1706832000&Signature=f8HqNPfYuHiudRIerbd74td1DApFzMzDMa2JYlidTLd1x6QG-FeA0sg8s7NuPSNlJ4RPyHBVMyko6mp1GQgP2dP--FxGqghOH~SjmBhZXR~pcXwjmeoARmZNXLzmihHjO1kKh6dM0TMAp6Wq~ivq6-VAr4ZdhBrZsLYuHOKpXuY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[Cosmetics Inc. - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/t3_O7KI_Rp6KVfCyGvTxGA_26e91b3aa1d44284956a4ae860d114e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1706832000&Signature=QSEaS20VZSCLtZgeMvON8n0IKWomvG0LcVWjY99uEL7HtuBFbBqqjZVTK-r53S1zkImsFvZHhHd5BtmrLqmFkq-LKhV~U1dKT80648cFrLwqGp15nXDWDhXKHpvQpFCOLEmypyAKyqSbv1KLaHA591e5gD4bHdfZB3NUYIZnZ5c_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/t3_O7KI_Rp6KVfCyGvTxGA_26e91b3aa1d44284956a4ae860d114e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1706832000&Signature=QSEaS20VZSCLtZgeMvON8n0IKWomvG0LcVWjY99uEL7HtuBFbBqqjZVTK-r53S1zkImsFvZHhHd5BtmrLqmFkq-LKhV~U1dKT80648cFrLwqGp15nXDWDhXKHpvQpFCOLEmypyAKyqSbv1KLaHA591e5gD4bHdfZB3NUYIZnZ5c_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



## Example 1: Use conditional formatting to highlight blank cells

Conditional formatting is a spreadsheet tool that changes how cells appear when values meet specific conditions.

1. Open the spreadsheet [International Logistics Association Memberships - Data for Cleaning](https://docs.google.com/spreadsheets/d/1jmxXS6ZJEMtaoli5__qApb9LE_nXkU2ysf5c8N1tiQA/template/preview).
2. Select the range of cells to which you’ll apply conditional formatting. In this example, you’ll select **columns A** through **L**, except for **columns F** and **H**. To select all columns except for **F** and **H**: a. Select cell **A** to highlight **column A**. b. Hold down the SHIFT key and at the same time use your mouse to select cell **E**. This will highlight all the columns between **A** and **E**. c. To select the remainder of the columns, hold down the CONTROL (Windows) or COMMAND (Mac) key while you select cells **G**, **I**, **J**, **K**, and **L**. d. **Columns A** through **L** in your spreadsheet should be highlighted except **Column F** and **Column H**.
3. From the menu, select **Format**, then **Conditional formatting**. The columns you’ve selected should turn a light shade of green, and a new **Conditional format rules** tool will appear. Additionally, the **Apply to range** field should indicate the cells you’ve selected.
4. Next, apply a condition to these cells to change the cell color if the cell is empty. In the **Format cells if** drop-down, select **Cell is empty**.
5. Select the **Formatting style** field. Select a bright color from the drop-down to make the blank cells stand out.
6. Select **Done**.

## Example 2: Remove duplicates

Remove duplicates is a spreadsheet tool that automatically searches for and eliminates duplicate entries from a spreadsheet. This is faster and easier than searching the data by scrolling through it.

1. Create a copy of your dataset by right clicking the **Association ABC membership** tab and selecting **Duplicate**. This is a good practice, as it protects against accidentally deleting important data. Continue working in the new sheet, **Copy of Association ABC memberships**.
2. In the menu, select **Data**, then **Data cleanup**, then **Remove duplicates**.
3. Check the box next to **Data has header row**.
4. Check the box next to **Select All** to inspect the entire spreadsheet.
5. Select **Remove duplicates**.

## Example 3: Format dates consistently

Format dates to make all of the data in your spreadsheet consistent. This makes items easier to find and manipulate.

1. Select **column J** (Membership valid through), which contains dates.
2. From the menu, select **Format**, then **Number**, then **Date**.

## Example 4: Use split to separate data into columns

The split menu option is helpful when you have more than one piece of data in a cell and you want to separate those pieces of data into different cells.

1. Select **column L** (Certification).
2. In the menu, select **Data**, then **Split text to columns**.
3. The delimiter (for example, a comma) will be automatically detected.
4. If needed, specify the separator manually in the dropdown that appears in your spreadsheet.

## Example 5: Use split to fix numbers stored as text

**SPLIT** is a spreadsheet function that divides text around a specified character and puts each fragment into a new, separate cell.

1. Open the spreadsheet [Cosmetics Inc. - Data for Cleaning](https://docs.google.com/spreadsheets/d/12U9Y4IVAGwml7XWBBgC4j9l0cCjqIZlqJc9vu3jr6Ig/template/preview?resourcekey=0-ds9iuh8tsuB7PwGd2dHMDA#gid=0).
2. Notice that cell **F12** contains an error.
3. Select **column E** (Total).
4. In the menu select **Data**, then select **Split text to columns**.
5. This removes the quotation marks from cell **E12**, so the spreadsheet recognizes the data in the cell as a number, resolving the error in cell **F12**.

Hi again. As you learned earlier,

there's a lot of different ways to clean up data.

I've shown you some examples of

how you can clean data manually,

such as searching for and fixing

misspellings or removing empty spaces and duplicates.

We also learned that lots

of spreadsheet applications have

tools that help simplify and

speed up the data cleaning process.

There's a lot of great efficiency tools

that data analysts use all the time,

such as conditional formatting,

removing duplicates, formatting dates,

fixing text strings and substrings,

and splitting text to columns.

We'll explore those in more detail now.

The first is something called conditional formatting.

Conditional formatting is a spreadsheet tool

that changes how cells

appear when values meet specific conditions.

Likewise, it can let you know when a cell

does not meet the conditions you've set.

Visual cues like this are very useful for data analysts,

especially when we're working in

a large spreadsheet with lots of data.

Making certain data points standout

makes the information easier to understand and analyze.

For cleaning data, knowing when

the data doesn't follow the condition is very helpful.

Let's return to the logistics association spreadsheet

to check out conditional formatting in action.

We'll use conditional formatting

to highlight blank cells.

That way, we know where there's

missing information so we can add it to the spreadsheet.

To do this, we'll start by

selecting the range we want to search.

For this example we're not focused on address

3 and address 5.

The fields will include

all the columns in our spreadsheets, except

for F and H. Next,

we'll go to Format and choose Conditional formatting.

Play video starting at :2:2 and follow transcript2:02

Great. Our range is automatically indicated in the field.

The format rule will be to

format cells if the cell is empty.

Play video starting at :2:15 and follow transcript2:15

Finally, we'll choose the formatting style.

I'm going to pick a shade of bright pink,

so my blanks really stand out.

Play video starting at :2:26 and follow transcript2:26

Then click "Done," and

the blank cells are instantly highlighted.

The next spreadsheet tool removes duplicates.

As you've learned before, it's

always smart to make a copy of the data

set before removing anything. Let's do that now.

Play video starting at :2:48 and follow transcript2:48

Great, now we can continue.

You might remember that our example spreadsheet

has one association member listed twice.

Play video starting at :3:1 and follow transcript3:01

To fix that, go to Data and select "Remove duplicates."

"Remove duplicates" is a tool that automatically searches

for and eliminates duplicate entries from a spreadsheet.

Choose "Data has header row" because our spreadsheet

has a row at the very top that

describes the contents of each column.

Next, select "All" because

we want to inspect our entire spreadsheet.

Finally, "Remove duplicates."

Play video starting at :3:37 and follow transcript3:37

You'll notice the duplicate row was

found and immediately removed.

Play video starting at :3:45 and follow transcript3:45

Another useful spreadsheet tool

enables you to make formats consistent.

For example, some of the dates in

this spreadsheet are in a standard date format.

Play video starting at :3:59 and follow transcript3:59

This could be confusing if you wanted to

analyze when association members joined,

how often they renewed their memberships,

or how long they've been with the association.

To make all of our dates consistent,

first select column J,

then go to "Format,"

select "Number," then "Date."

Now all of our dates have a consistent format.

Before we go over the next tool,

I want to explain what a text string is.

In data analytics, a text string

is a group of characters within a cell,

most often composed of letters.

An important characteristic of

a text string is its length,

which is the number of characters in it.

You'll learn more about that soon.

For now, it's also useful to know that

a substring is a smaller subset of a text string.

Now let's talk about Split.

Split is a tool that divides a text string around

the specified character and puts

each fragment into a new and separate cell.

Split is helpful when you have more than one piece of

data in a cell and you want to separate them out.

This might be a person's first and last name

listed together,

or it could be a cell that

contains someone's city, state,

country, and zip code,

but you actually want each of those in its own column.

Let's say this association wanted to analyze all of

the different professional certifications

its members have earned.

To do this, you want

each certification separated out into its own column.

Right now, the certifications are separated by a comma.

That's the specified text separating each item,

also called the delimiter.

Let's get them separated.

Highlight the column, then select "Data,"

and "Split text to columns."

Play video starting at :6:2 and follow transcript6:02

This spreadsheet application automatically

knew that the comma was a delimiter

and separated each certification.

But sometimes you might need to

specify what the delimiter should be.

You can do that here.

Play video starting at :6:18 and follow transcript6:18

Split text to columns is also helpful for

fixing instances of numbers stored as text.

Sometimes values in your spreadsheet

will seem like numbers,

but they're formatted as text.

This can happen when copying and pasting from

one place to another or if the formatting's wrong.

For this example, let's check out

our new spreadsheet from a cosmetics maker.

If a data analyst wanted to determine total profits,

they could add up everything in column

F. But there's a problem;

one of the cells has an error.

If you check into it,

you learn that the "707" in this cell

is text and can't be changed into a number.

When the spreadsheet tries to multiply the cost of

the product by the number of units sold,

it's unable to make the calculation.

But if we select

the orders column and choose "Split text to columns,"

Play video starting at :7:26 and follow transcript7:26

the error is resolved because

now it can be treated as a number.

Coming up, you'll learn about

a tool that does just the opposite.

CONCATENATE is a function that joins

multiple text strings into a single string.

Spreadsheets are a very important part of data analytics.

They save data analysts time and effort

and help us eliminate errors each and every day.

Here, you've learned about some of

the most common tools that we use.

But there's a lot more to come.

Next, we'll learn even more about data cleaning

with spreadsheet tools. Bye for now!

# Step-by-Step: Optimize the data-cleaning process

This reading outlines steps the instructor performs in the following video, [Optimize the data-cleaning process](https://www.coursera.org/learn/process-data/lecture/ohiCl/optimize-the-data-cleaning-process). The video teaches some useful spreadsheet functions, which can make your data-cleaning even more successful.

Keep this step-by-step guide open as you watch the video. It can serve as a helpful reference if you need additional context or clarification while following the video steps. This is not a graded activity, but you can complete these steps to practice the skills demonstrated in the video.

### **What you’ll need**

If you would like to access the spreadsheet the instructor uses in this video, click the link to the dataset to create a copy. If you don’t have a Google account, you may download the data directly from the attachments below.

Link to logistics data: [International Logistics Association Memberships - Data for Cleaning](https://docs.google.com/spreadsheets/d/1jmxXS6ZJEMtaoli5__qApb9LE_nXkU2ysf5c8N1tiQA/template/preview)

Link to cosmetics data: [Cosmetics Inc. - Data for Cleaning](https://docs.google.com/spreadsheets/d/12U9Y4IVAGwml7XWBBgC4j9l0cCjqIZlqJc9vu3jr6Ig/template/preview?resourcekey=0-ds9iuh8tsuB7PwGd2dHMDA#gid=0)

OR

[International Logistics Association Memberships - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/p1JSr3UeTOOi_2-3Yr67dg_2e42a0117ecc42978ff39c15b0fbd3f1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1706832000&Signature=KYYX6pxlZm8nfUhNLSxggKGVEFYqTpZ3pqdV-SU6VaTESVsAG~EJTvM-JQbv~kKvyokTGhiGkJUvtnjPWVDkdcTP7OY8tVHUXcX1Go80siaAc68HPVUmCGuWtg0cSOK-e9s2Thl0UQGQoF4wEOEalFMQ6xuaF6h10KPsi1mUZWg_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/p1JSr3UeTOOi_2-3Yr67dg_2e42a0117ecc42978ff39c15b0fbd3f1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1706832000&Signature=KYYX6pxlZm8nfUhNLSxggKGVEFYqTpZ3pqdV-SU6VaTESVsAG~EJTvM-JQbv~kKvyokTGhiGkJUvtnjPWVDkdcTP7OY8tVHUXcX1Go80siaAc68HPVUmCGuWtg0cSOK-e9s2Thl0UQGQoF4wEOEalFMQ6xuaF6h10KPsi1mUZWg_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[Cosmetics Inc. - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/htYeVDpXRHCUSNSclnWcqQ_06ea60f3250b4ba2a59e4fc7d2e2a5e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1706832000&Signature=IMSaeBa7QyNUlWI~AExDhKjdQSWU0jotfnPIWi1vDEhz17tx727ZfoeyG9Jd~9y-A2dYE2ojfuxotvCrVGoqNgpC6i1CTDy0aB4vRNRQstG30-QkdEr1FBtnIBcuHrSMUtuc-4MnHpB1obp8KzSgg8y3f1N5uVBuPDqLzUiPw5s_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/htYeVDpXRHCUSNSclnWcqQ_06ea60f3250b4ba2a59e4fc7d2e2a5e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1706832000&Signature=IMSaeBa7QyNUlWI~AExDhKjdQSWU0jotfnPIWi1vDEhz17tx727ZfoeyG9Jd~9y-A2dYE2ojfuxotvCrVGoqNgpC6i1CTDy0aB4vRNRQstG30-QkdEr1FBtnIBcuHrSMUtuc-4MnHpB1obp8KzSgg8y3f1N5uVBuPDqLzUiPw5s_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



## Example 1: The COUNTIF function

**COUNTIF** is a spreadsheet function that returns the number of cells within a range that match a specified value.

### **Use** COUNTIF **to find numbers lower than 100**

1. Open the International [Logistics Association Memberships - Data for Cleaning](https://docs.google.com/spreadsheets/d/1jmxXS6ZJEMtaoli5__qApb9LE_nXkU2ysf5c8N1tiQA/template/preview) dataset, and scroll down to row 74.
   1. **Note:** The dataset has 72 rows, and row 73 is left blank for separation.
2. In cell **H74**, enter **Member Dues < 100** to label the calculation.
3. In cell **I74**, enter the formula **=COUNTIF(I2:I72,"<100")** to count how many members in the cell range **I2:I72** pay dues of less than $100. This formula returns a value of 1, indicating one value is below $100.
4. In cell **I55**, change -$200 to $200. Cell **I74** should now display the value 0.

### **Use** COUNTIF **to find numbers higher than 500**

1. In cell **H75**, enter **Member Dues > 500**.
2. In cell **I75**, enter the formula **=COUNTIF(I2:I72,">500")** to count how many members in cell range **I2:I72** pay dues of greater than 500. This formula returns a value of 1, indicating one value is above 500.
3. In cell **I44**, change $1,000 to $100. Cell **I75** should now display the value 0.

## Example 2: The LEN function

The **LEN** function is useful if you have a certain piece of information in your spreadsheet that you know must contain a certain length.

1. Right click cell **A**.
2. Select **+ Insert one column right** to create a new, empty column.
3. Select cell **B1** and enter **LEN** to name the new column.
4. In cell **B2**, enter **=LEN(A2)**. This function references the value of cell **A2** and returns its length, 6.
5. Double-click on the lower right corner of cell **B2**. This will copy the function through the rest of the column. Each cell will show the length of the Member ID in that row.

## Example 3: Use conditional formatting

Conditional formatting is a spreadsheet tool that changes how cells appear when values meet specific conditions.

1. To highlight all of column **B** except for the header, select cell **B**. Then press CONTROL (Windows) or COMMAND (MAC) and select cell **B1**.
2. Navigate to the **Format** menu, and choose **Conditional Formatting**.
3. Set the **Format rules** field to **Is not equal to** and enter **6** as the value.
4. Select **Done**.
5. Notice cell **B36** is highlighted because its value is 7.

## Example 4: The LEFT and RIGHT functions

**LEFT** is a function that returns a set number of characters from the left side of a text string. **RIGHT** is a function that returns a set number of characters from the right side of a text string.

### **The** LEFT **function**

1. Use the [Cosmetics Inc. - Data for Cleaning](https://docs.google.com/spreadsheets/d/12U9Y4IVAGwml7XWBBgC4j9l0cCjqIZlqJc9vu3jr6Ig/template/preview?resourcekey=0-ds9iuh8tsuB7PwGd2dHMDA#gid=0) dataset.
2. Select cell **H1**, and enter **Left**.
3. In cell **H2**, enter **=LEFT(A2, 5)** to extract the first five characters from cell **A2**. This function will show the substring 51993.
4. Select cell **H2**.
5. Select and hold the fill handle, the small circle in the corner of a selected cell, then drag this formula down to populate the rest of this column.

### **The** RIGHT **function**

1. Select cell **I1**, and enter **Right**.
2. In cell **I2**, enter **=RIGHT(A2, 4)** to extract the last four characters from cell **A2**. This function will show the substring Masc.
3. Select cell **I2**.
4. Select and hold the fill handle and drag this formula down to populate the rest of this column.

## Example 5: The MID function

MID is a function that returns a segment from the middle of a text string.

1. Select cell **J1**, and enter **Mid**.
2. In cell **J2**, enter **=MID(D2, 4, 2)** to extract the two-letter state code that starts at character four in cell **D2**.
3. Double-click the fill handle and to automatically populate the rest of this column.

## Example 6: The CONCATENATE function

**CONCATENATE** is a spreadsheet function that joins together two or more text strings.

1. Select cell **K1**, and enter **Concatenate**.
2. In cell **K2**, enter **=CONCATENATE(H2, I2)** to combine the values from columnsH and I.
3. Double-click the fill handle and to automatically populate the rest of this column.

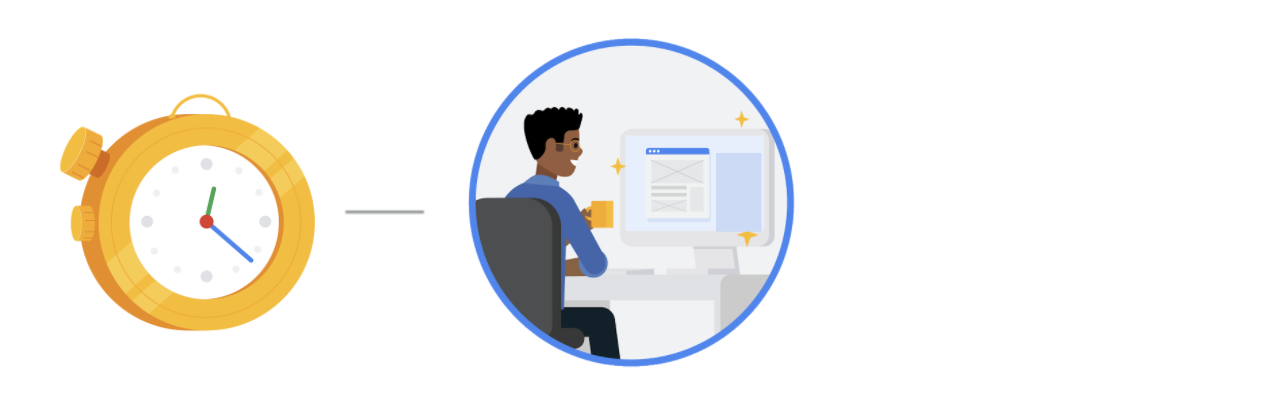
## Example 7: TRIM function

**TRIM** is a function that removes leading, trailing, and repeated spaces in data.

1. Select cell **L1**, and enter **Trim**.
2. In cell **L2**, enter **=TRIM(C2)** to remove any leading, trailing, or repeated spaces.
3. Double-click the fill handle and to automatically populate the rest of this column.
4. Welcome back. You've learned about
5. some very useful data- cleaning tools that
6. are built right into spreadsheet applications.
7. Now we'll explore how functions can
8. optimize your efforts to ensure data integrity.
9. As a reminder, a function is a set of instructions that
10. performs a specific calculation
11. using the data in a spreadsheet.
12. The first function we'll discuss is called COUNTIF.
13. COUNTIF is a function that returns
14. the number of cells that match a specified value.
15. Basically, it counts the number of
16. times a value appears in a range of cells.
17. Let's go back to
18. our professional association spreadsheet.
19. In this example,
20. we want to make sure
21. the association membership prices are listed accurately.
22. We'll use COUNTIF to check for some common problems,
23. like negative numbers or a value that's
24. much less or much greater than expected.
25. To start, let's find the least expensive membership:
26. $100 for student associates.
27. That'll be the lowest number that exists in this column.
28. If any cell has a value that's
29. less than 100, COUNTIF will alert us.
30. We'll add a few more rows at
31. the bottom of our spreadsheet,
32. Play video starting at :1:22 and follow transcript1:22
33. then beneath column H, type "member dueS
34. less than $100." Next,
35. type the function in the cell next to it.
36. Every function has a certain syntax
37. that needs to be followed for it to work.
38. Syntax is a predetermined structure that
39. includes all required information
40. and its proper placement.
41. The syntax of a COUNTIF function should be
42. like this: Equals COUNTIF,
43. open parenthesis, range, comma,
44. the specified value in
45. quotation marks and a closed parenthesis.
46. It will show up like this.
47. Play video starting at :2:5 and follow transcript2:05
48. Where I2 through I72 is the range,
49. and the value is less than 100.
50. This tells the function to go through column I,
51. and return a count of all cells that
52. contain a number less than 100.
53. Turns out there is one!
54. Scrolling through our data,
55. we find that one piece of data was
56. mistakenly keyed in as a negative number.
57. Let's fix that now. Now we'll
58. use COUNTIF to search for
59. any values that are more than we would expect.
60. The most expensive membership type
61. is $500 for corporate members.
62. Type the function in the cell.
63. Play video starting at :3:5 and follow transcript3:05
64. This time it will appear like this:
65. I2 through I72 is still the range,
66. but the value is greater than 500.
67. Play video starting at :3:21 and follow transcript3:21
68. There's one here too. Check it out.
69. Play video starting at :3:28 and follow transcript3:28
70. This entry has an extra zero.
71. It should be $100.
72. Play video starting at :3:36 and follow transcript3:36
73. The next function we'll discuss is called LEN.
74. LEN is a function that tells you the length of
75. the text string by counting
76. the number of characters it contains.
77. This is useful when cleaning data
78. if you have a certain piece of information in
79. your spreadsheet that you
80. know must contain a certain length.
81. For example, this association
82. uses six-digit member identification codes.
83. If we'd just imported this data
84. and wanted to be sure our codes are
85. all the correct number of digits, we'd use LEN.
86. The syntax of LEN is equals LEN,
87. open parenthesis, the range,
88. and the close parenthesis.
89. We'll insert a new column after Member ID.
90. Play video starting at :4:25 and follow transcript4:25
91. Then type an equals sign and LEN.
92. Add an open parenthesis.
93. The range is the first Member ID number in A2.
94. Finish the function by closing the parenthesis.
95. It tells us that there are six characters in cell A2.
96. Let's continue the function through
97. the entire column and
98. find out if any results are not six.
99. But instead of manually
100. going through our spreadsheet to search
101. for these instances, we'll use conditional formatting.
102. We talked about conditional formatting earlier.
103. It's a spreadsheet tool that
104. changes how cells appear when
105. values meet specific conditions. Let's practice that now.
106. Select all of column B except for the header.
107. Then go to Format and choose Conditional formatting.
108. The format rule is to format cells
109. if not equal to six.
110. Play video starting at :5:38 and follow transcript5:38
111. Click "Done." The cell
112. with the seven inside is highlighted.
113. Play video starting at :5:47 and follow transcript5:47
114. Now we're going to talk about LEFT and RIGHT.
115. LEFT is a function that gives you
116. a set number of characters from
117. the left side of a text string.
118. RIGHT is a function that gives you
119. a set number of characters from
120. the right side of a text string.
121. As a quick reminder, a text string
122. is a group of characters within a cell,
123. commonly composed of letters, numbers, or both.
124. To see these functions in action,
125. let's go back to the spreadsheet from
126. the cosmetics maker from earlier.
127. This spreadsheet contains product codes.
128. Each has a five-digit numeric code and then
129. a four-character text identifier.
130. Play video starting at :6:35 and follow transcript6:35
131. But let's say we only want to work with
132. one side or the other.
133. You can use LEFT or RIGHT to
134. give you the specific set of
135. characters or numbers you need.
136. We'll practice cleaning up our data
137. using the LEFT function first.
138. The syntax of LEFT is equals LEFT, open parenthesis,
139. the range, a comma,
140. and a number of characters from
141. the left side of the text string we want.
142. Then, we finish it with a closed parenthesis.
143. Here, our project requires
144. just the five-digit numeric codes.
145. In a separate column,
146. Play video starting at :7:16 and follow transcript7:16
147. type equals LEFT, open parenthesis, then the range.
148. Our range is A2. Then,
149. add a comma, and then
150. number 5 for our five- digit product code.
151. Finally, finish the function with a closed parenthesis.
152. Our function should show up like this.
153. Press "Enter." And now,
154. we have a substring,
155. which is the number part of the product code only.
156. Play video starting at :7:50 and follow transcript7:50
157. Click and drag this function through the entire column to
158. separate out the rest of
159. the product codes by number only.
160. Play video starting at :8:2 and follow transcript8:02
161. Now, let's say our project only needs the
162. four-character text identifier.
163. Play video starting at :8:9 and follow transcript8:09
164. For that, we'll use the RIGHT function,
165. and the next column will begin the function.
166. The syntax is equals RIGHT,
167. open parenthesis, the range,
168. a comma and the number of characters we want.
169. Then, we finish with a closed parenthesis.
170. Let's key that in now.
171. Equals right, open parenthesis, and the range is still
172. A2. Add a comma.
173. This time, we'll tell it that we want
174. the first four characters from the right.
175. Close up the parenthesis and press "Enter."
176. Then, drag the function throughout the entire column.
177. Play video starting at :8:58 and follow transcript8:58
178. Now, we can analyze the product in
179. our spreadsheet based on either substring.
180. The five-digit numeric code
181. or the four character text identifier.
182. Hopefully, that makes it clear how you can use LEFT and
183. RIGHT to extract substrings
184. from the left and right sides of a string.
185. Now, let's learn how you can extract
186. something in between.
187. Here's where we'll use something called MID.
188. MID is a function that gives you
189. a segment from the middle of a text string.
190. This cosmetics company lists
191. all of its clients using a client code.
192. It's composed of the first three letters
193. of the city where the client is located,
194. its state abbreviation,
195. and then a three- digit identifier.
196. But let's say a data analyst needs to
197. work with just the states in the middle.
198. The syntax for MID is equals MID,
199. open parenthesis, the range, then a comma.
200. When using MID, you
201. always need to supply a reference point.
202. In other words, you need to set
203. where the function should start.
204. After that, place another comma,
205. and how many middle characters you want.
206. In this case, our range is D2.
207. Let's start the function in a new column.
208. Play video starting at :10:21 and follow transcript10:21
209. Type equals MID, open parenthesis, D2.
210. Then the first three characters represent a city name,
211. so that means the starting point is the fourth.
212. Add a comma and four.
213. We also need to tell the function
214. how many middle characters we want.
215. Add one more comma,
216. and two, because the state abbreviations
217. are two characters long.
218. Press "Enter" and bam,
219. we just get the state abbreviation.
220. Continue the MID function through the rest of the column.
221. Play video starting at :11:2 and follow transcript11:02
222. We've learned about a few functions that help
223. separate out specific text strings.
224. But what if we want to combine them instead?
225. For that, we'll use CONCATENATE,
226. which is a function that joins
227. together two or more text strings.
228. The syntax is equals
229. CONCATENATE, then an open parenthesis
230. inside indicates each text
231. string you want to join, separated by commas.
232. Then finish the function with a closed parenthesis.
233. Just for practice, let's say we needed to rejoin
234. the left and right text strings back
235. into complete product codes.
236. In a new column, let's begin our function.
237. Play video starting at :11:47 and follow transcript11:47
238. Type equals CONCATENATE, then an open parenthesis.
239. The first text string we want to join is
240. in H2. Then add a comma.
241. The second part is in I2.
242. Add a closed parenthesis and press "Enter".
243. Drag it down through the entire column,
244. Play video starting at :12:16 and follow transcript12:16
245. and just like that, all of
246. our product codes are back together.
247. Play video starting at :12:24 and follow transcript12:24
248. The last function we'll learn about here is TRIM.
249. TRIM is a function that removes leading,
250. trailing, and repeated spaces in data.
251. Sometimes when you import data,
252. your cells have extra spaces,
253. which can get in the way of your analysis.
254. Play video starting at :12:44 and follow transcript12:44
255. For example, if this cosmetics maker
256. wanted to look up a specific client name,
257. it won't show up in the search if it has extra spaces.
258. You can use TRIM to fix that problem.
259. The syntax for TRIM is equals TRIM,
260. open parenthesis, your range, and closed parenthesis.
261. In a separate column,
262. Play video starting at :13:11 and follow transcript13:11
263. type equals TRIM and an open parenthesis.
264. The range is C2,
265. as you want to check out the client names.
266. Close the parenthesis and press "Enter".
267. Finally, continue the function down the column.
268. Play video starting at :13:32 and follow transcript13:32
269. TRIM fixed the extra spaces.
270. Play video starting at :13:42 and follow transcript13:42
271. Now we know some very useful functions that can
272. make your data cleaning even more successful.
273. This was a lot of information.
274. As always, feel free to go back and
275. review the video and then practice on your own.
276. We'll continue building on these tools soon,
277. and you'll also have a chance to practice.
278. Pretty soon, these data cleaning steps will
279. become second nature, like brushing your teeth.

**Workflow automation**

In this reading, you will learn about workflow automation and how it can help you work faster and more efficiently. Basically, workflow automation is the process of automating parts of your work. That could mean creating an event trigger that sends a notification when a system is updated. Or it could mean automating parts of the data cleaning process. As you can probably imagine, automating different parts of your work can save you tons of time, increase productivity, and give you more bandwidth to focus on other important aspects of the job.



**What can be automated?**

Automation sounds amazing, doesn’t it? But as convenient as it is, there are still some parts of the job that can’t be automated. Let's take a look at some things we can automate and some things that we can’t.

| **Task** | **Can it be automated?** | **Why?** |
| --- | --- | --- |
| Communicating with your team and stakeholders | No | Communication is key to understanding the needs of your team and stakeholders as you complete the tasks you are working on. There is no replacement for person-to-person communications. |
| Presenting your findings | No | Presenting your data is a big part of your job as a data analyst. Making data accessible and understandable to stakeholders and creating data visualizations can’t be automated for the same reasons that communications can’t be automated. |
| Preparing and cleaning data | Partially | Some tasks in data preparation and cleaning can be automated by setting up specific processes, like using a programming script to automatically detect missing values. |
| Data exploration | Partially | Sometimes the best way to understand data is to see it. Luckily, there are plenty of tools available that can help automate the process of visualizing data. These tools can speed up the process of visualizing and understanding the data, but the exploration itself still needs to be done by a data analyst. |
| Modeling the data | Yes | Data modeling is a difficult process that involves lots of different factors; luckily there are tools that can completely automate the different stages. |

**More about automating data cleaning**

One of the most important ways you can streamline your data cleaning is to clean data where it lives. This will benefit your whole team, and it also means you don’t have to repeat the process over and over. For example, you could create a programming script that counted the number of words in each spreadsheet file stored in a specific folder. Using tools that can be used where your data is stored means that you don’t have to repeat your cleaning steps, saving you and your team time and energy.

**More resources**

There are a lot of tools out there that can help automate your processes, and those tools are improving all the time. Here are a few articles or blogs you can check out if you want to learn more about workflow automation and the different tools out there for you to use:

* Towards Data Science’s [**Automating Scientific Data Analysis**](https://towardsdatascience.com/automating-scientific-data-analysis-part-1-c9979cd0817e)
* MIT News’ [**Automating Big-Data Analysis**](https://news.mit.edu/2016/automating-big-data-analysis-1021)
* TechnologyAdvice’s [**10 of the Best Options for Workflow Automation Software**](https://technologyadvice.com/blog/information-technology/top-10-workflow-automation-software/)

**Key takeaways**

As a data analyst, automation can save you a lot of time and energy, and free you up to focus more on other parts of your project. The more analysis you do, the more ways you will find to make your processes simpler and more streamlined.

# Step-by-Step: Different data perspectives

This reading outlines the steps the instructor performs in the next video, [Different data perspectives](https://www.coursera.org/learn/process-data/lecture/BcY0L/different-data-perspectives). The video teaches you different methods data analysts use to view data differently and how looking at different views leads to more efficient and effective data cleaning.

Keep this step-by-step guide open as you watch the video. It can serve as a helpful reference if you need additional context or clarification while following the video steps. This is not a graded activity, but you can complete these steps to practice the skills demonstrated in the video.

### **What you’ll need**

If you’d like to follow along with the examples in this video, choose a spreadsheet tool. Google Sheets or Excel are recommended.

To access the spreadsheet the instructor uses in this video, click the link to the template to create a copy of the dataset. If you don’t have a Google account, download the data directly from the attachments below.

Link to template: [Cosmetics, Inc](https://docs.google.com/spreadsheets/d/1J8wiEi7R9Jt3kNOjV1Bp-w1Zw1GvIbXgd78EeoXT9Mg/template/preview).

OR

[Cosmetics Inc. - Data for Pivot Table and VLOOKUP](https://d3c33hcgiwev3.cloudfront.net/mjJpNpPfRiC7blak2QdsSA_f5269cdfb2da42ddb06577759b173be1_Cosmetics-Inc.---Data-for-Pivot-Table-and-VLOOKUP.xlsx?Expires=1706832000&Signature=JziKY9aOHYDxU-DgzSf-q-3~Hlh-76RPvks2jo-1L3BQsYju0IzrOFSd-fue7VvpmoWf1ARgWDIBS5IjLlJFoL-opV4ElQc1kL0HOHXoijljRNOJpy25taqXs6GOreFRfvrQs0fH-OjViRFXTaqY7KrnfRNC1Qviq2X5Zjcv7dw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/mjJpNpPfRiC7blak2QdsSA_f5269cdfb2da42ddb06577759b173be1_Cosmetics-Inc.---Data-for-Pivot-Table-and-VLOOKUP.xlsx?Expires=1706832000&Signature=JziKY9aOHYDxU-DgzSf-q-3~Hlh-76RPvks2jo-1L3BQsYju0IzrOFSd-fue7VvpmoWf1ARgWDIBS5IjLlJFoL-opV4ElQc1kL0HOHXoijljRNOJpy25taqXs6GOreFRfvrQs0fH-OjViRFXTaqY7KrnfRNC1Qviq2X5Zjcv7dw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



## Example 1: Pivot tables

A pivot table is a data summarization tool. It can be used in data processing and in data cleaning, for which pivot tables offer a quick, clutter-free view of your data. Pivot tables help sort, reorganize, group, count, total, or average data in a dataset.

1. In the Cosmetics Inc. spreadsheet, select the data you need to include. In this case, select the entire spreadsheet by dragging from the top-left cell to the bottom-right cell that contains data.
2. Select **Insert**, then **Pivot Table**. Choose **New sheet** and **Create**. Google Sheets creates a new sheet where you can define the pivot table.
3. To add specific data to your pivot table, refer to the **Pivot table editor** on the right side of the window. For example, in the video, the instructor says they want to view only the most profitable products—the ones for which Cosmetics, Inc. has at least $10,000 in orders.

a. In the **Pivot table editor** panel, next to **Rows**, select **Add**.

b. From the columns list, select **Total**.

c. Below **Rows**, from the **Order** dropdown list, select **Descending** to put the most profitable items at the top.

d. To add another row with the product codes, next to **Rows**, select **Add**.

e. From the column list, select **Products**.

f. Notice that the top two most ordered products are **15143Exfo** and **3279Masc**. The rest of the orders total less than $10,000.

## Example 2: VLOOKUP

**VLOOKUP** is a spreadsheet function that vertically searches for a certain value in a column to return a corresponding piece of information. It's rare for all of the data an analyst will need to be in the same place. Usually, you'll have to search across multiple sheets or even different databases. **VLOOKUP** helps bring the information together.

1. In the Cosmetics Inc. spreadsheet on the **Sheet 1** tab, select a cell in the first empty column adjacent to the top row of your data, such as H2.
2. In the selected cell, enter **=VLOOKUP(A2, 'Sheet 2'!A1:B31, 2, false)** a. **Note:** This references information in another sheet. Make sure you have Sheet 2 in your workbook. b. This formula will take the value in cell **A2** of **Sheet 1** and check for that value in **Sheet 2** among the cells from **A1:B31** in the 2nd column (which corresponds with the 2 in the formula). Because the formula includes “**false**,” it will search only for an exact match. It will then output the value of column **B** in Sheet 2 as the result.
3. Press **Enter** to input the formula. The result is LashX Mascara.
4. Next, **select the cell** and **drag** the fill handle in the lower-right corner down to populate the other cells in the sheet with the formula.
5. To identify the products mentioned, select **Edit > Find and Replace**. In the **Find** text box, enter the product codes, then press **Enter**.

## Example 3: Plotting

The plotting tool allows analysts to quickly create a graph, chart, table, or other visual from data. Plotting is useful for identifying skewed data or outliers.

1. In **Sheet 1** of the Cosmetics, Inc. spreadsheet, **select column B**, which contains the prices.
2. **Select Insert > Chart**. a. By default, Google Sheets creates a bar chart. b. Drag the chart to the side so you can view the data in the sheet.
3. Check for obvious outliers and fix them in the spreadsheet. For example, you might notice that an item in the middle of the chart has an extremely low price of $0.73. The decimal point is in the wrong place. In cell **B15**, correct this price to $7.30, and notice that Google Sheets automatically updates the chart.

Hi, let's get into it.

Motivational speaker Wayne Dyer once said,

"If you change the way you look at things, the things you look at change."

This is so true in data analytics.

No two analytics projects are ever exactly the same.

So it only makes sense that different projects require us

to focus on different information differently.

Play video starting at ::26 and follow transcript0:26

In this video,

we'll explore different methods that data analysts use to look at data differently

and how that leads to more efficient and effective data cleaning.

Play video starting at ::36 and follow transcript0:36

Some of these methods include sorting and filtering, pivot tables,

a function called VLOOKUP, and plotting to find outliers.

Play video starting at ::47 and follow transcript0:47

Let's start with sorting and filtering.

As you learned earlier, sorting and

filtering data helps data analysts customize and

organize the information the way they need for a particular project.

But these tools are also very useful for data cleaning.

Play video starting at :1:3 and follow transcript1:03

You might remember that sorting involves arranging data into a meaningful

order to make it easier to understand, analyze, and visualize.

Play video starting at :1:13 and follow transcript1:13

For data cleaning,

you can use sorting to put things in alphabetical or numerical order,

so you can easily find a piece of data.

Play video starting at :1:22 and follow transcript1:22

Sorting can also bring duplicate entries closer together for faster identification.

Play video starting at :1:29 and follow transcript1:29

Filters, on the other hand, are very useful in data cleaning when you want to

find a particular piece of information.

Play video starting at :1:36 and follow transcript1:36

You learned earlier that filtering means showing only the data that meets

a specific criteria while hiding the rest.

Play video starting at :1:44 and follow transcript1:44

This lets you view only the information you need.

Play video starting at :1:48 and follow transcript1:48

When cleaning data, you might use a filter to only find values

above a certain number, or just even or odd values.

Again, this helps you find what you need quickly and

separates out the information you want from the rest.

Play video starting at :2:5 and follow transcript2:05

That way you can be more efficient when cleaning your data.

Play video starting at :2:10 and follow transcript2:10

Another way to change the way you view data is by using pivot tables.

Play video starting at :2:15 and follow transcript2:15

You've learned that a pivot table is a data summarization tool that is used in

data processing.

Play video starting at :2:21 and follow transcript2:21

Pivot tables sort, reorganize, group, count, total or average data stored in the database.

In data cleaning, pivot tables are used to give you a quick, clutter-

free view of your data. You can choose to look at the specific parts

of the data set that you need to get a visual in the form of a pivot table.

Play video starting at :2:44 and follow transcript2:44

Let's create one now using our cosmetic makers spreadsheet again.

Play video starting at :2:49 and follow transcript2:49

To start, select the data we want to use.

Here, we'll choose the entire spreadsheet. Select "Data" and then "Pivot table."

Play video starting at :3:5 and follow transcript3:05

Choose "New sheet" and "Create."

Play video starting at :3:11 and follow transcript3:11

Let's say we're working on a project that requires us

to look at only the most profitable products.

Items that earn the cosmetics maker at least $10,000 in orders.

So the row we'll include is "Total" for total profits.

Play video starting at :3:31 and follow transcript3:31

We'll sort in descending order to put the most profitable items at the top.

Play video starting at :3:38 and follow transcript3:38

And we'll show totals.

Play video starting at :3:41 and follow transcript3:41

Next, we'll add another row for products

Play video starting at :3:49 and follow transcript3:49

so that we know what those numbers are about.

We can clearly determine tha the most profitable products

have the product codes 15143 E-X-F-O and 32729

M-A-S-C.

Play video starting at :4:6 and follow transcript4:06

We can ignore the rest for

this particular project because they fall below $10,000 in orders.

Play video starting at :4:17 and follow transcript4:17

Now, we might be able to use context clues to assume we're talking

about exfoliants and mascaras.

But we don't know which ones, or if that assumption is even correct.

Play video starting at :4:30 and follow transcript4:30

So we need to confirm what the product codes correspond to.

Play video starting at :4:34 and follow transcript4:34

And this brings us to the next tool.

It's called VLOOKUP.

Play video starting at :4:39 and follow transcript4:39

VLOOKUP stands for vertical lookup.

It's a function that searches for

a certain value in a column to return a corresponding piece of information.

When data analysts look up information for a project, it's rare for

all of the data they need to be in the same place.

Usually, you'll have to search across multiple sheets or

even different databases.

Play video starting at :5:5 and follow transcript5:05

The syntax of the VLOOKUP is equals VLOOKUP, open parenthesis,

then the data you want to look up.

Next is a comma and where you want to look for that data.

Play video starting at :5:19 and follow transcript5:19

In our example, this will be the name of a spreadsheet followed by an exclamation point.

Play video starting at :5:25 and follow transcript5:25

The exclamation point indicates that we're referencing a cell in a different

sheet from the one we're currently working in.

Play video starting at :5:32 and follow transcript5:32

Again, that's very common in data analytics.

Play video starting at :5:35 and follow transcript5:35

Okay, next is the range in the place where you're looking for

data, indicated using the first and last cell separated by a colon.

After one more comma is the column in the range containing the value to return.

Play video starting at :5:50 and follow transcript5:50

Next, another comma and the word "false,"

which means that an exact match is what we're looking for.

Play video starting at :5:59 and follow transcript5:59

Finally, complete your function by closing the parentheses.

To put it simply, VLOOKUP searches for the value in the first argument in

the leftmost column of the specified location.

Play video starting at :6:12 and follow transcript6:12

Then the value of the third argument tells VLOOKUP to

return the value in the same row from the specified column.

Play video starting at :6:20 and follow transcript6:20

The "false" tells VLOOKUP that we want an exact match.

Play video starting at :6:25 and follow transcript6:25

Soon you'll learn the difference between exact and approximate matches.

But for now, just know that V lookup takes the value in one cell and

searches for a match in another place.

Play video starting at :6:38 and follow transcript6:38

Let's begin.

Play video starting at :6:41 and follow transcript6:41

We'll type equals VLOOKUP.

Play video starting at :6:49 and follow transcript6:49

Then add the data we are looking for, which is the product data.

Play video starting at :6:56 and follow transcript6:56

The dollar sign makes sure that the corresponding part of the reference

remains unchanged.

Play video starting at :7:3 and follow transcript7:03

You can lock just the column, just

the row, or both at the same time.

Play video starting at :7:19 and follow transcript7:19

Next, we'll tell it to look at Sheet 2, in both columns

Play video starting at :7:33 and follow transcript7:33

We added 2 to represent the second column.

Play video starting at :7:37 and follow transcript7:37

The last term, "false," says we wanted an exact match.

Play video starting at :7:47 and follow transcript7:47

With this information, we can now analyze the data for

only the most profitable products.

Play video starting at :8:1 and follow transcript8:01

Going back to the two most profitable products, we can search for

15143 E-X-F-O And 32729 M-A-S-C.

Go to Edit and then Find. Type

in the product codes and

search for them.

Play video starting at :8:31 and follow transcript8:31

Now we can learn which products we'll be using for this particular project.

Play video starting at :8:36 and follow transcript8:36

The final tool we'll talk about is something called plotting.

When you plot data, you put it in a graph chart, table, or

other visual to help you quickly find what it looks like.

Play video starting at :8:49 and follow transcript8:49

Plotting is very useful when trying to identify any skewed data or outliers.

For example, if we want to make sure the price of each product is correct,

we could create a chart. This would give us a visual aid that helps

us quickly figure out if anything looks like an error.

Play video starting at :9:10 and follow transcript9:10

So let's select the column with our prices.

Play video starting at :9:15 and follow transcript9:15

Then we'll go to Insert and choose Chart.

Play video starting at :9:24 and follow transcript9:24

Pick a column chart as the type. One of these prices looks extremely low.

Play video starting at :9:31 and follow transcript9:31

If we look into it,

we discover that this item has a decimal point in the wrong place.

Play video starting at :9:41 and follow transcript9:41

It should be $7.30, not 73 cents.

Play video starting at :9:51 and follow transcript9:51

That would have a big impact on our total profits.

So it's a good thing we caught that during data cleaning.

Play video starting at :9:58 and follow transcript9:58

Looking at data in new and

creative ways helps data analysts identify all kinds of dirty data.

Play video starting at :10:5 and follow transcript10:05

Coming up, you'll continue practicing these new concepts so

you can get more comfortable with them.

You'll also learn additional strategies for ensuring your data is clean,

and we'll provide you with effective insights.

Great work so far.

# Step-by-Step: Even more data-cleaning techniques

This reading outlines the steps the instructor performs in the next video, [Even more data-cleaning techniques](https://www.coursera.org/learn/process-data/lecture/Ei2IH/even-more-data-cleaning-techniques). This video teaches you different methods data analysts use in data mapping. Data mapping is the process of matching fields from one database to another. It’s critical to the success of data migration, data integration, and many other data-management activities. This video contains one activity for you to practice.

Keep this step-by-step guide open as you watch the video. It can serve as a helpful reference if you need additional context or clarification while following the video steps. This is not a graded activity, but you can complete these steps to practice the skill demonstrated in the video.

### **What you’ll need**

If you’d like to follow along with the example in this video, choose a spreadsheet tool, such as Google Sheets or Excel.

To access the spreadsheet the instructor uses in this video, click the link to the template to create a copy of the dataset. If you don’t have a Google account, download the data directly from the attachments below.

Link to templates:

[International Logistics Association memberships](https://docs.google.com/spreadsheets/d/1ST8zMU2NCVgWCfwOge-8iWJguZ-NwW8C8s5p3kfC2K0/template/preview)

[Global Logistics Association](https://docs.google.com/spreadsheets/d/186Yx3S-ejZr1cJunsal2tUV9J1abixhBUDjXbUtQT7I/template/preview)

[Logistics Association Merger](https://docs.google.com/spreadsheets/d/1rWFRcunKBh2t6zy5zxOxcBrDyokyJZL9OJnZ5XH5tFw/template/preview)

Downloads:

[International Logistics Association memberships](https://d3c33hcgiwev3.cloudfront.net/xSEEsOv6TRa77FfljFKFyQ_c0b53b24e0b54f6fa5d83c5c8683a1e1_International-Logistics-Association-memberships.xlsx?Expires=1706832000&Signature=a1cbg2yM097~vBpSVDC27M3odwPPCStBPe~0-MvHx~zzpunBZz8tiaaSQyoeMEcWk6C4BoGDH64mrX8t-DOCCZHW1UV27x95ThMA801VllcPagIHxua9CTHq4rcZAhcPVw~LbKXl8xKLODwNpvvb7p1QHeJdUerSqScQY18LsE4_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/xSEEsOv6TRa77FfljFKFyQ_c0b53b24e0b54f6fa5d83c5c8683a1e1_International-Logistics-Association-memberships.xlsx?Expires=1706832000&Signature=a1cbg2yM097~vBpSVDC27M3odwPPCStBPe~0-MvHx~zzpunBZz8tiaaSQyoeMEcWk6C4BoGDH64mrX8t-DOCCZHW1UV27x95ThMA801VllcPagIHxua9CTHq4rcZAhcPVw~LbKXl8xKLODwNpvvb7p1QHeJdUerSqScQY18LsE4_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[Global Logistics Association](https://d3c33hcgiwev3.cloudfront.net/tMtjHQztTiesjo8vTKb8UA_a5b6cb5cf2404d9abe4b7c3797da94e1_Global-Logistics-Association.xlsx?Expires=1706832000&Signature=J6b36y-rBA~jQuxUe5Phq4cDfi2EHuRWwkHiiwURdB-xPcY7K7q5jsVf-P69ttO4wQkA4FIGPeI0H~4Lsq5dNlbZpyKR21A-UcruGzXljB8H2g8C2OK7Ccd7Cx5vxkQREe10j~EPXDN2AZKL8Erkleq4VPk6pC2hxjjLSuxUVd8_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/tMtjHQztTiesjo8vTKb8UA_a5b6cb5cf2404d9abe4b7c3797da94e1_Global-Logistics-Association.xlsx?Expires=1706832000&Signature=J6b36y-rBA~jQuxUe5Phq4cDfi2EHuRWwkHiiwURdB-xPcY7K7q5jsVf-P69ttO4wQkA4FIGPeI0H~4Lsq5dNlbZpyKR21A-UcruGzXljB8H2g8C2OK7Ccd7Cx5vxkQREe10j~EPXDN2AZKL8Erkleq4VPk6pC2hxjjLSuxUVd8_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[Logistics Association Merger](https://d3c33hcgiwev3.cloudfront.net/rCJ95nZHTpCqGu0S0vglvg_a244e9e881604d688bef52d5388febe1_Logistics-Association-Merger.xlsx?Expires=1706832000&Signature=LdeEmF9HCWeMQ~ep~nUVMbZ9sAVDzf9O9odHuM1Knj88Jt0iStlsCGW4D5jofKYyO-tS55QbPlh7HnziZjQb6W91Kvuun03QBR76OiE~HZ1stRtqW3yr6mytWwlrnvP-hgZ-e-0qSxIO1y0gNVhJtrqQw8Mli1KLvdsX3w6pFDI_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/rCJ95nZHTpCqGu0S0vglvg_a244e9e881604d688bef52d5388febe1_Logistics-Association-Merger.xlsx?Expires=1706832000&Signature=LdeEmF9HCWeMQ~ep~nUVMbZ9sAVDzf9O9odHuM1Knj88Jt0iStlsCGW4D5jofKYyO-tS55QbPlh7HnziZjQb6W91Kvuun03QBR76OiE~HZ1stRtqW3yr6mytWwlrnvP-hgZ-e-0qSxIO1y0gNVhJtrqQw8Mli1KLvdsX3w6pFDI_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



## Example: CONCATENATE

**CONCATENATE** is a function that joins together two or more text strings. In the video, you’ll learn how to use **CONCATENATE** to clean data after two datasets have been combined.

1. Open the dataset spreadsheet titled **Global Logistics Association**. When prompted, select **USE TEMPLATE**.
2. **Insert** a new column to the right of column **E**. Label it **New Address** in cell **F1**.
3. In the second row of the new column (cell **F2**), enter **=CONCATENATE (D2,E2)** and press **Enter**.
   1. You will notice that some results need a space between the street address and the unit or suite number, such as: 25 Dyas RdSte. 101.
   2. You could manually clean the data later to add a space between Rd and Ste., but **CONCATENATE** can actually do it for you.
   3. The **CONCATENATE** formula can help you format the data as it is merged by entering an additional string to insert a space between Rd and Ste.
   4. Enter **=CONCATENATE(D2, " ", E2)** and you will have an address that is formatted like this: 25 Dyas Rd Ste. 101. Much better!
4. Ensure the new data in the cell accurately reflects the merging of the two previous columns.
5. **Select cell F2** and **drag down** to apply the formula to all rows in the column.

Hello. So far you've learned about a lot of

different tools and functions that

analysts use to clean up data for analysis.

Now we'll take a step back and talk about some of

the really big picture aspects of clean data.

Knowing how to fix specific problems,

either manually with spreadsheet tools,

or with functions, is extremely valuable.

But it's also important to think about

how your data has moved between

systems and how it's evolved

along it's journey to your data analysis project.

To do this, data analysts

use something called data mapping.

Data mapping is the process of

matching fields from one database to another.

This is very important to the success of data migration,

data integration,

and lots of other data management activities.

As you learned earlier,

different systems store data in different ways.

For example, the state field in

one spreadsheet might show Maryland spelled out.

But another spreadsheet might store it as MD.

Play video starting at :1:8 and follow transcript1:08

Data mapping helps us note these kinds of differences

so we know when data is moved and

combined it will be compatible.

As a quick reminder,

compatibility describes how well

two or more data sets are able to work together.

The first step to data mapping is

identifying what data needs to be moved.

This includes the tables and the fields within them.

We also need to define the desired format for

the data once it reaches its destination.

To figure out how this works let's go back to

the merger between our two logistics associations.

Play video starting at :1:48 and follow transcript1:48

Starting with the first data field,

we'll identified that we need to

move both sets of member IDs.

To define the desired format,

we'll choose whether to use

numbers like this spreadsheet,

or email addresses like the other spreadsheet.

Next comes mapping the data.

Depending on the schema and number of

primary and foreign keys in a data source,

data mapping can be simple or very complex.

As a reminder, a schema is

a way of describing how something is organized.

A primary key references a column in which each value is

unique and a foreign key is

a field within a table that is

a primary key in another table.

For more challenging projects there's

all kinds of data mapping software programs you can use.

These data mapping tools will analyze

field by field how to move data

from one place to another then they automatically clean,

match, inspect, and validate the data.

They also create consistent naming conventions,

ensuring compatibility when the data

is transferred from one source to another.

When selecting a software program to map your data,

you want to be sure that it

supports the file types you're working with,

such as Excel, SQL, Tableau, and others.

Later on, you'll learn more about

selecting the right tool for a particular task.

For now, let's practice mapping data manually.

First, we need to determine the content of

each section to make sure

the data ends up in the right place.

For example, the data on when memberships

expire would be consolidated into a single column.

This step makes sure that each piece of information

ends up in the most appropriate place

in the merged data source.

Now, you might remember that some of the data

was inconsistent between the two organizations,

like the fact that one uses

a separate column for suite apartment or

unit number but the other doesn't.

Play video starting at :4:11 and follow transcript4:11

This brings us to the next step,

transforming the data into a consistent format.

This is a great time to use concatenate.

As you learned before,

concatenate is a function that joins

together two or more text strings,

which is what we did earlier with

our cosmetics company example.

We'll insert a new column

Play video starting at :4:43 and follow transcript4:43

and then type equals concatenate,

Play video starting at :4:50 and follow transcript4:50

then the two text strings we want to make one.

Play video starting at :4:58 and follow transcript4:58

Drag that through the entire column.

Play video starting at :5:8 and follow transcript5:08

Now we have the consistency in

the new merged association lists of member addresses.

Play video starting at :5:20 and follow transcript5:20

Now that everything's compatible,

it's time to transfer the data to its destination.

There's a lot of different ways to

move data from one place to another,

including querying, import wizards,

and even simple drag and drop.

Here's our merged spreadsheet.

Play video starting at :5:41 and follow transcript5:41

It looks good, but we still want to make

sure everything was transferred properly.

We'll go into the testing phase of data mapping.

For this, you inspect a sample piece of data

to confirm that it's clean and properly formatted.

Play video starting at :6: and follow transcript6:00

It's also a smart practice to do

spot checks on things such as the number of nulls.

For the test, you can use a lot

of the data cleaning tools we discussed previously,

such as data validation,

conditional formatting, COUNTIF, sorting, and filtering.

Finally, once you've determined

that the data is clean and compatible,

you can start using it for analysis.

Data mapping is so important because even one mistake

when merging data can ripple throughout an organization,

causing the same error to appear again and again.

This leads to poor results.

On the other hand, data mapping

can save the day by giving you

a clear road map you can follow to make sure

your data arrives safely at it's destination.

That's why you learn how to do it.

# Develop your approach to cleaning data

As you continue on your data journey, you’re likely discovering that data is often messy—and you can expect raw, primary data to be imperfect. In this reading, you’ll consider how to develop your personal approach to cleaning data. You will explore the idea of a cleaning checklist, which you can use to guide your cleaning process. Then, you’ll define your preferred methods for cleaning data. By the time you complete this reading, you’ll have a better understanding of how to methodically approach the data cleaning process. This will save you time when cleaning data and help you ensure that your data is clean and usable.

## Consider your approach to cleaning data

Data cleaning usually requires a lot of time, energy, and attention. But there are two steps you can take before you begin to help streamline your process: creating a cleaning checklist and deciding on your preferred methods. This will help ensure that you know exactly how you want to approach data cleaning and what you need to do to be confident in the integrity of your data.

### **Your cleaning checklist**

Start developing your personal approach to cleaning data by creating a checklist to help you identify problems in your data efficiently and identify the scale and scope of your dataset. Think of this checklist as your default “what to search for” list.

Here are some examples of common data cleaning tasks you could include in your checklist:

* **Determine the size of the dataset:** Large datasets may have more data quality issues and take longer to process. This may impact your choice of data cleaning techniques and how much time to allocate to the project.
* **Determine the number of categories or labels:** By understanding the number and nature of categories and labels in a dataset, you can better understand the diversity of the dataset. This understanding also helps inform data merging and migration strategies.
* **Identify missing data:** Recognizing missing data helps you understand data quality so you can take appropriate steps to remediate the problem. Data integrity is important for accurate and unbiased analysis.
* **Identify unformatted data:** Identifying improperly or inconsistently formatted data helps analysts ensure data uniformity. This is essential for accurate analysis and visualization.
* **Explore the different data types:** Understanding the types of data in your dataset (for instance, numerical, categorical, text) helps you select appropriate cleaning methods and apply relevant data analysis techniques.

There might be other data cleaning tasks you’ve been learning about that you also want to prioritize in your checklist. Your checklist is an opportunity for you to define exactly what you want to remember about cleaning your data; feel free to make it your own.

### **Your preferred cleaning methods**

In addition to creating a checklist, identify which actions or tools you prefer using when cleaning data. You’ll use these tools and techniques with each new dataset—or whenever you encounter issues in a dataset—so this list should be compatible with your checklist.

For example, suppose you have a large dataset with missing data. You’ll want to know how to check for missing data in larger datasets, and how you plan to handle any missing data, before you start cleaning. Outlining your preferred methods can save you lots of time and energy.

## Key takeaways

The data you encounter as an analyst won’t always conform to your checklist or your preferred actions and tools. But having these things can make common data cleaning tasks much easier to complete. As is so often the case, thoughtful planning sets up any project for success!

### 1.

Question 1

What is the relationship between a text string and a substring?

1 point

A text string is a row of data within a table. A substring is one cell within that row.

A text string is a column of data within a table. A substring is one cell within that column.

A text string is a group of characters within a cell. A substring is a smaller subset of that text string.

A text string is the list of attributes at the top of columns within a table. A substring is a single attribute within that list.

### 2.

Question 2

A data analyst uses the **COUNTIF** function to count the number of times a value less than 50 occurs between spreadsheet cells D2 through F100. What is the correct syntax?

1 point

**=COUNTIF(D2:F100,">50")**

**=COUNTIF(D2:F100,"<50")**

**=COUNTIF(D2:F100,<50)**

**=COUNTIF(D2:F100,>50)**

### 3.

Question 3

Fill in the blank: To remove leading, trailing, and repeated spaces in data, analysts use the \_\_\_\_ function.

1 point

**MID**

**LEFT**

**TRIM**

**RIGHT**

### 4.

Question 4

Which spreadsheet tool searches for matches to a specified value in one column, returning a corresponding piece of information from another location?

1 point

**CONCATENATE**

**LEN**

**VLOOKUP**

SPLIT

# Glossary terms from module 2

## ****Terms and definitions for Course 4, Module 2****

**Clean data:** Data that is complete, correct, and relevant to the problem being solved

**Compatibility:** How well two or more datasets are able to work together

**CONCATENATE:** A spreadsheet function that joins together two or more text strings

**Conditional formatting:** A spreadsheet tool that changes how cells appear when values meet specific conditions

**Data engineer:** A professional who transforms data into a useful format for analysis and gives it a reliable infrastructure

**Data mapping:** The process of matching fields from one data source to another

**Data merging:** The process of combining two or more datasets into a single dataset

**Data validation:** A tool for checking the accuracy and quality of data

**Data warehousing specialist:** A professional who develops processes and procedures to effectively store and organize data

**Delimiter:** A character that indicates the beginning or end of a data item

**Dirty data:** Data that is incomplete, incorrect, or irrelevant to the problem to be solved

**Duplicate data:** Any record that inadvertently shares data with another record

**Field length:** A tool for determining how many characters can be keyed into a spreadsheet field

**Incomplete data:** Data that is missing important fields

**Inconsistent data:** Data that uses different formats to represent the same thing

**Incorrect/inaccurate data:** Data that is complete but inaccurate

**LEFT:** A function that returns a set number of characters from the left side of a text string

**LEN:** A function that returns the length of a text string by counting the number of characters it contains

**Length:** The number of characters in a text string

**Merger:** An agreement that unites two organizations into a single new one

**MID:** A function that returns a segment from the middle of a text string

**Null:** An indication that a value does not exist in a dataset

**Outdated data:** Any data that has been superseded by newer and more accurate information

**Remove duplicates:** A spreadsheet tool that automatically searches for and eliminates duplicate entries from a spreadsheet

**Split:** A function that divides text around a specified character and puts each fragment into a new, separate cell

**Substring:** A smaller subset of a text string

**Text string:** A group of characters within a cell, most often composed of letters

**TRIM:** A function that removes leading, trailing, and repeated spaces in data

**Unique:** A value that can’t have a duplicate

### 1.

Question 1

To identify ways to improve the shipping process, a data analyst merges a dataset of client order data with a dataset of shipping data. What should the data team do to ensure the compatibility of the two datasets?

1 point

Map the data

Spotcheck for null values

Use a visualization

Apply a data structure

### 2.

Question 2

Fill in the blank: When typing a **MID** function, the correct \_\_\_\_\_ to follow is **=MID(range, reference starting point, number of middle characters)**.

1 point

algorithm

system

validation

syntax

### 3.

Question 3

In this spreadsheet, which function will extract Clay Casey’s four-digit postcode?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| (n/a) | **A** | **B** | **C** | **D** |
| **1** | **Customer** | **Street** | **Territory** | **Country** |
| **2** | Terry Stewart | 707 Collins St, Docklands | VIC 3008 | Australia |
| **3** | Che Price | 177 William St, Tennant Creek | NT 0860 | Australia |
| **4** | Stanley Moss | 430 Little Collins St, | CXI 6798 | Australia |
| **5** | Clay Casey | 250 Queen St, Sydney | NSW 3000 | Australia |

1 point

**=RIGHT(4,C5)**

**=LEFT(4,C5)**

**=LEFT(C5,4)**

**=RIGHT(C5,4)**

### 4.

Question 4

Fill in the blank: In a **VLOOKUP** function, the word false tells **VLOOKUP** that an \_\_\_\_\_ match is desired.

1 point

inexact

exact

uncertain

approximate

### 5.

Question 5

In the following spreadsheet, a data professional wants to create product IDs in column C. The IDs should include the item name plus its version number. Which function will create the ID Life\_vest04?

|  |  |  |  |
| --- | --- | --- | --- |
| (n/a) | **A** | **B** | **C** |
| **1** | **Product** | **Version** | **ID** |
| **2** | Raft | 05 |  |
| **3** | Oar | 01 |  |
| **4** | Life\_vest | 04 |  |
| **5** | Tether\_rope | 02 |  |

1 point

**=CONCATENATE(A4\*B4)**

**=CONCATENATE(A4+B4)**

**=CONCATENATE(A4\_B4)**

**=CONCATENATE(A4,B4)**

### 6.

Question 6

A data analyst wants to make sure their dataset contains only numbers below 100 from cell A2 through cell A50. Which of the following **COUNTIF** statements could they use?

1 point

**=COUNTIF(A2:A50,"<100")**

**=COUNTIF(A2:A50,">=100")**

**=COUNTIF(A2:A50, >100)**

**=COUNTIF(A2:A50, <=100)**

### 7.

Question 7

A data analyst uses the **SPLIT** tool to place each grain and dairy product into new, separate cells. What is the semicolon’s function in this scenario?

|  |  |  |
| --- | --- | --- |
| (n/a) | **A** | **B** |
| **1** | **Grains** | **Dairy** |
| **2** | Ricel Corn | Milk; Cheese |
| **3** | Oats; Quinoa | Yogurt; Butter |
| **4** | Wheat; Barley; Farro | Buttermilk; Ice Cream |

1 point

Substring

Duplicate

String

Delimiter

### 8.

Question 8

A junior data analyst needs to search their spreadsheet for a particular client ID. In order to identify all cells containing the ID, they use a spreadsheet tool that changes how cells appear when values meet specific conditions. What tool do they use?

1 point

Conditional formatting

Field length

Cell filtering

Data merging

0:01

Welcome back and great job on that last weekly challenge.

Now that we know the difference between cleaning dirty data and

some general data cleaning techniques,

let's focus on data cleaning using SQL.

Coming up we'll learn about the different data cleaning functions in

spreadsheets and SQL

and how SQL can be used to clean large data sets.

I'll also show you how to develop some basic search queries for databases

and how to apply basic SQL functions for transforming data and cleaning strings.

Cleaning your data is the last step in the data analysis process

before you can move on to the actual analysis, and

SQL has a lot of great tools that can help you do that.

Play video starting at ::39 and follow transcript0:39

But before we start cleaning databases, we'll take a closer look at SQL and

when to use it.

I'll see you there.

Advertising agencies get money from their clients

to advertise their brand.

These agencies use our products,

use certain Google platforms, advertising platforms,

and I help them with how to best use those platforms,

different strategies they can use to be best in class.

A lot of the folks at the advertising agencies

have reports that they have to send out to their clients.

These reports take a lot of time to create and visualize,

and so what I do is I help the practitioners

and the analytics teams use

a particular product that enables them to create

those reports much faster and much easier.

If you're going to start off as a data analyst,

it opens tons of doors

because everybody is tracking data,

is using data, needs to use data, regardless of industry.

Anywhere from health care,

to advertising, to e-commerce,

to entertainment,

anything and everything, everybody uses data,

so everybody needs you as a data analyst.

SQL makes our lives easier

when we're analyzing lots of different data.

It's only somewhat recently

that the SQL programs that we use now

can give us instant results for

analyzing millions or billions of data.

Years ago, maybe about five years ago or so,

even though we could still

analyze those millions of rows,

we would end up having to wait fifteen minutes,

thirty minutes for the queries to run.

But now it's instantaneous,

and so that's really exciting,

and we can do so much more with that power.

SQL has helped a lot in my career because it's one

of those fundamental things you have

to know as a data analyst.

Back in the day, not everyone knew SQL,

so knowing SQL was definitely a competitive advantage.

Nowadays, I would say more people,

maybe most people know it.

It's a core skill and

highly sought after by everybody.

So, knowing SQL, becoming a data analyst

makes you quite popular from recruiters,

so I think that's really fun.

I taught myself SQL,

so my knowledge about SQL is something I

hold near and dear, close to my heart

since it's something that almost I've made for

myself, and I feel so much satisfaction from it.

So that's why I really like SQL.

One of the fun things about SQL and

another reason why I really enjoy using

it is because when you type something in that query,

and you just hit Control, Shift, Enter,

or once you've run the query,

you get the results almost

instantly, depending on the platform you use.

But it's fascinating to see if you think

conceptually how much analysis

the computer is doing for you based on

that little bit of

command code or a little bit of code you wrote,

and it's just so powerful

if you think about what's happening behind the scenes.

So I think that's fun to look at.

We live in a world of big data,

and it keeps getting bigger.

The computing power is also increasing exponentially.

With all the data that we can track,

the more and more we can track that data,

the more and more we need data analysts.

Our career prospects are basically skyrocketing.

I'm Sally, I'm

a measurement and analytical lead at Google.

Hello, again.

So before we go over all the ways data analysts use SQL to clean data,

I want to formally introduce you to SQL.

We've talked about SQL a lot already.

You've seen some databases and some basic functions in SQL,

and you've even seen how SQL can be used to process data.

But now let's actually define SQL.

SQL is a structured query language that analysts use to work with databases.

Data analysts usually use SQL to deal with large datasets

because it can handle huge amounts of data.

And I mean trillions of rows.

That's a lot of rows to wrap your head around.

So let me give you an idea about how much data that really is.

Play video starting at ::41 and follow transcript0:41

Imagine a data set that contains the names of all 8 billion people in the world.

It would take the average person 101 years to read all 8 billion names.

SQL can process this in seconds.

Personally, I think that's pretty cool.

Other tools like spreadsheets might take a really long time to process that much

data, which is one of the main reasons data analysts choose to use SQL,

when dealing with big datasets.

Let me give you a short history on SQL.

Development on SQL actually began in the early 70s.

Play video starting at :1:15 and follow transcript1:15

In 1970, Edgar F.Codd developed the theory about relational databases.

You might remember learning about relational databases a while back.

This is a database that contains a series of tables that can be connected to form

relationships.

At the time IBM was using a relational database management system called

System R.

Well, IBM computer scientists were trying to figure out a way to manipulate and

retrieve data from IBM System R.

Their first query language was hard to use.

So they quickly moved on to the next version, SQL.

In 1979, after extensive testing SQL, now just spelled S-Q-L, was released publicly.

By 1986, SQL had become the standard language for

relational database communication, and it still is.

This is another reason why data analysts choose SQL.

It's a well-known standard within the community.

The first time I used SQL to pull data from a real database was for

my first job as a data analyst.

I didn't have any background knowledge about SQL before that.

I only found out about it because it was a requirement for that job.

The recruiter for that position gave me a week to learn it.

So I went online and researched it and ended up teaching myself SQL.

They actually gave me a written test as part of the job application process.

I had to write SQL queries and functions on a whiteboard.

But I've been using SQL ever since.

And I really like it.

And just like I learned SQL on my own,

I wanted to remind you that you can figure things out yourself too.

There's tons of great online resources for learning.

So don't let one job requirement stand in your way without doing

some research first.

Now that we know a little more about why analysts choose to work with SQL when

they're handling a lot of data and a little bit about the history of SQL,

we'll move on and learn some practical applications for it.

Coming up next, we'll check out some of the tools we learned in spreadsheets and

figure out if any of those apply to working in SQL.

Spoiler alert, they do.

See you soon.

**How a junior data analyst uses SQL**

In this reading, you will learn more about how to decide when to use SQL, or Structured Query Language. As a data analyst, you will be tasked with handling a lot of data, and SQL is one of the tools that can help make your work a lot easier. SQL is the primary way data analysts extract data from databases. As a data analyst, you will work with databases all the time, which is why SQL is such a key skill. Let’s follow along as a junior data analyst uses SQL to solve a business task.

**The business task and context**

The junior data analyst in this example works for a social media company. A new business model was implemented on February 15, 2020 and the company wants to understand how their user-growth compares to the previous year. Specifically, the data analyst was asked to find out how many users have joined since February 15, 2020.



**Spreadsheets functions and formulas or SQL queries?**

Before they can address this question, this data analyst needs to choose what tool to use. First, they have to think about where the data lives. If it is stored in a database, then SQL is the best tool for the job. But if it is stored in a spreadsheet, then they will have to perform their analysis in that spreadsheet. In that scenario, they could create a pivot table of the data and then apply specific formulas and filters to their data until they were given the number of users that joined after February 15th. It isn’t a really complicated process, but it would involve a lot of steps.

In this case, the data is stored in a database, so they will have to work with SQL. And this data analyst knows they could get the same results with a single SQL query:

1

2

3

4

5

6

SELECT

    COUNT(DISTINCT user\_id) AS count\_of\_unique\_users

FROM

    table

WHERE

    join\_date >= '2020-02-15'

Spreadsheets and SQL both have their advantages and disadvantages:

| **Features of Spreadsheets** | **Features of SQL Databases** |
| --- | --- |
| Smaller data sets | Larger datasets |
| Enter data manually | Access tables across a database |
| Create graphs and visualizations in the same program | Prepare data for further analysis in another software |
| Built-in spell check and other useful functions | Fast and powerful functionality |
| Best when working solo on a project | Great for collaborative work and tracking queries run by all users |

**Key takeaways**

When it comes down to it, where the data lives will decide which tool you use. If you are working with data that is already in a spreadsheet, that is most likely where you will perform your analysis. And if you are working with data stored in a database, SQL will be the best tool for you to use for your analysis. You will learn more about SQL coming up, so that you will be ready to tackle any business problem with the best tool possible.

# SQL dialects and their uses

In this reading, you will learn more about SQL dialects and some of their different uses. As a quick refresher, **Structured Query Language**, or SQL, is a language used to talk to databases. Learning SQL can be a lot like learning a new language—including the fact that languages usually have different dialects within them. Some database products have their own variant of SQL, and these different varieties of SQL dialects are what help you communicate with each database product.

These dialects will be different from company to company and might change over time if the company moves to another database system. So, a lot of analysts start with Standard SQL and then adjust the dialect they use based on what database they are working with. Standard SQL works with a majority of databases and requires a small number of syntax changes to adapt to other dialects.

As a junior data analyst, it is important to know that there are slight differences between dialects. But by mastering Standard SQL, which is the dialect you will be working with in this program, you will be prepared to use SQL in any database.

## More information

You may not need to know every SQL dialect, but it is useful to know that these different dialects exist. If you are interested in learning more about SQL dialects and when they are used, you can check out these resources for more information:

* LearnSQL’s blog, [**What Is a SQL Dialect, and Which One Should You Learn?**](https://learnsql.com/blog/what-sql-dialect-to-learn/)
* Software Testing Help’s article, [**Differences Between SQL Vs MySQL vs SQL Server**](https://www.softwaretestinghelp.com/sql-vs-mysql-vs-sql-server/)
* Datacamp’s blog, [**SQL Server, PostgreSQL, MySQL... what's the difference? Where do I start?**](https://www.datacamp.com/community/blog/sql-differences)Note that there is an error in this blog article. The comparison table incorrectly states that SQlite uses subqueries instead of window functions. Refer to the [**SQLite Window Functions**](https://sqlite.org/windowfunctions.html)documentation for proper clarification.
* SQL Tutorial’s tutorial, [**What is SQL**](https://www.sqltutorial.org/what-is-sql/)

# Review: Set up your BigQuery account

**Note:** This reading is also in Courses 3 and 5 of this program. If you’re taking the courses in order, you may either review it or move on to the next new course item, [Hands-On Activity: Processing time with SQL](https://www.coursera.org/learn/process-data/quiz/9fIb9/hands-on-activity-processing-time-with-sql). If you haven’t taken Courses 3 or 5 you should complete this reading before proceeding to the next course item.

As you’ve been learning, BigQuery is a database you can use to access, explore, and analyze data from many sources. Now, you’ll begin using BigQuery, which will help you gain SQL knowledge by typing out commands and troubleshooting errors. This reading will guide you through the process of setting up your very own BigQuery account.

**Note:** Working with BigQuery is not a requirement of this program. Additional resources for other SQL database platforms are also provided at the end of this reading if you choose to use them instead.

## BigQuery account options

BigQuery offers a variety of account tiers to cater to various user needs and has two free-of-charge entry points, a sandbox account and a free-of-charge trial account. These options allow you to explore the program before selecting the best choice to suit your needs. A sandbox account allows you to practice writing queries and to explore public datasets free of charge, but it has [quotas and limits](https://cloud.google.com/bigquery/quotas), as well as some additional [restrictions](https://cloud.google.com/bigquery/docs/sandbox#limits). If you prefer to use BigQuery with the standard limits, you can set up a free-of-charge trial account instead. The free-of-charge trial is a trial period prior to paying for a subscription. In this instance, there is no automatic charge, but you will be asked for payment information when you create the account.

This reading provides instructions for setting up either account type. An effective first step is to begin with a sandbox account and switch to a free-of-charge trial account when needed to run the SQL presented upcoming courses.

### **Sandbox account**

The sandbox account is available at no cost, and anyone with a Google account can use it. However, it does have some limitations. For instance, you are limited to a maximum of 12 projects at a time. This means that, to create a 13th project, you'll need to delete one of your existing 12 projects. Additionally, the sandbox account doesn't support all operations you’ll do in this program. For example, there are limits on the amount of data you can process and you can’t insert new records into a database or update the values of existing records. However, a sandbox account is perfect for most program activities, including all of the activities in this course. Additionally, you can convert your sandbox account into a free-of-charge trial account at any time.

**Set up your sandbox account**

To set up a sandbox account:

1. Visit the [BigQuery sandbox documentation](https://cloud.google.com/bigquery/docs/sandbox#limits) page.
2. Log in to your preferred Google account by selecting the profile icon in the BigQuery menu bar.
3. Select the **Go to BigQuery** button on the documentation page.
4. You'll be prompted to select your country and read the terms of service agreement.
5. This will bring you to the **SQL Workspace**, where you'll be conducting upcoming activities. By default, BigQuery creates a project for you.

After you set up your account, the name of the project will be in the banner in your BigQuery console.

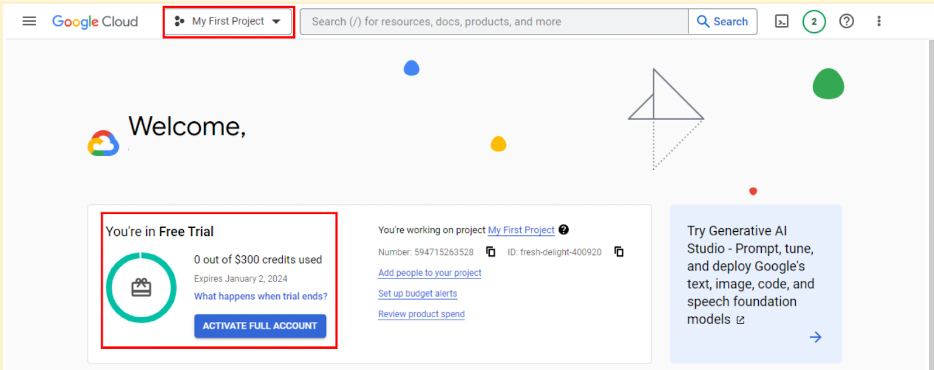
### **Free-of-charge trial**

If you wish to explore more of BigQuery's capabilities with fewer limitations, consider the Google Cloud Free Trial. It provides you with $300 in credit for Google Cloud usage during the first 90 days. If you're primarily using BigQuery for SQL queries, you're unlikely to come close to this spending limit. After you've used up the $300 credit or after 90 days, your free trial will expire, and you will only be able to use this account if you pay to do so. Google won't automatically charge your payment method when the trial ends. However, you'll need to set up a payment option with Google Cloud. This means that you’ll need to enter your financial information. Rest assured, it won't charge you unless you consciously opt to upgrade to a paid account. If you're uncomfortable providing payment information, don't worry; you can use the BigQuery sandbox account instead.

**Set up your free-of-charge trial**

1. Go to the [BigQuery](https://cloud.google.com/bigquery) page.
2. Select **Try BigQuery free**.
3. Log in using your Google email, or create an account free of charge if you don't have one. [Click here](https://cloud.google.com/bigquery?utm_source=google&utm_medium=cpc&utm_campaign=na-US-all-en-dr-bkws-all-all-trial-e-dr-1605212&utm_content=text-ad-none-any-DEV_c-CRE_665665924750-ADGP_Hybrid+%7C+BKWS+-+MIX+%7C+Txt_BigQuery-KWID_43700077225652770-kwd-274188433361&utm_term=KW_bigquery%20account-ST_bigquery+account&gclid=CjwKCAjwkNOpBhBEEiwAb3MvvYQXjIQ4TRnkITJoSXz7DFez4T-XKPG5IpfKmxUg2iHPEmiJBNQByhoCLVgQAvD_BwE&gclsrc=aw.ds) to create an account.
4. Select your country, a description of your organization or needs, and the checkbox to accept the terms of service, Then select **CONTINUE**.
5. Enter your billing information and select **START MY FREE TRIAL**.

After you set up your account, your first project, titled **My First Project** will be in the banner.



### **Transferring between BigQuery accounts**

With either a sandbox or free-of-charge trial account, you have the flexibility to upgrade to a paid account at any time. If you upgrade, all your existing projects will be retained and transferred to your new account. If you started with a free-of-charge trial, but choose not to upgrade when it ends, you can switch to a sandbox account. However, note that projects from your trial won't transfer to your sandbox. Essentially, creating a sandbox is like starting from scratch.

## Get started with other databases (if not using BigQuery)

It’s easiest to follow along with the course activities if you use BigQuery, but you may use other SQL platforms, if you prefer. If you decide to practice SQL queries on other database platforms, here are some resources to get started:

* [Getting Started with MySQL](https://dev.mysql.com/doc/mysql-getting-started/en/)
* [Getting 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)
* [Getting Started with PostgreSQL](https://www.postgresql.org/docs/10/tutorial-start.html)
* [Getting Started with SQLite](https://www.sqlite.org/quickstart.html)

## Key takeaways

BigQuery offers multiple account options. Keep the following in mind when you choose an account type:

* **Account tiers:** BigQuery provides various account tiers to cater to a wide range of user requirements. Whether you're starting with a sandbox account or exploring a paid account with the free-of-charge trial option, BigQuery offers flexibility to choose the option that aligns best with your needs and budget.
* **Sandbox limitations:** While a sandbox account is a great starting point, it comes with some limitations, such as a cap on the number of projects and restrictions on data manipulation operations like inserting or updating records, which you will encounter later in this program. Be aware of these limitations if you choose to work through this course using a sandbox account.
* **Easy setup and upgrades:** Getting started with any BigQuery account type is quick and easy. And if your needs evolve, you have the flexibility to modify your account status at any time. Additionally, projects can be retained even when transitioning between account types.

Choose the right BigQuery account type to match your specific needs and adapt as your requirements change!

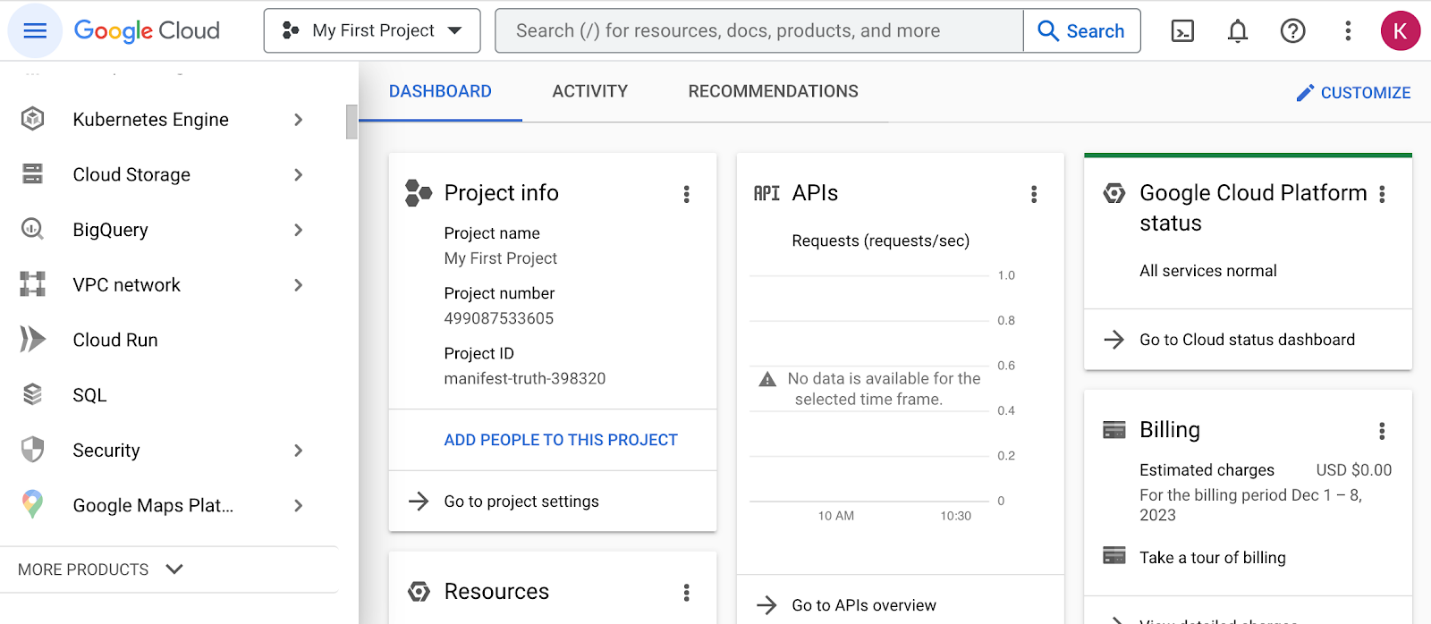
# Review: Get started with BigQuery

**Note:** This reading is also in Courses 3 and 5 of this program. If you’re taking the courses in order, you may either review it or move on to the next course item, [Hands-On Activity: Processing time with SQL](https://www.coursera.org/learn/process-data/quiz/9fIb9/hands-on-activity-processing-time-with-sql). If you haven’t taken Courses 3 or 5 you should complete this reading before proceeding to the next course item.

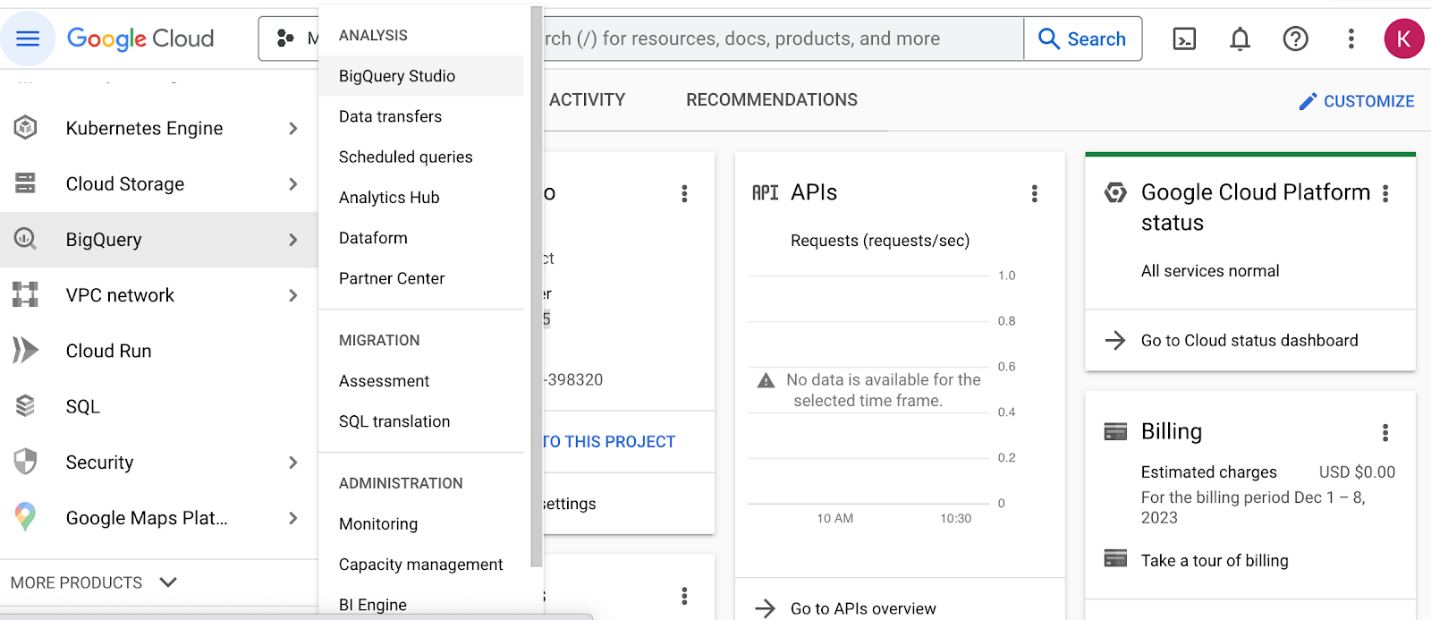
BigQuery is a data warehouse on the Google Cloud Platform used to query and filter large datasets, aggregate results, and perform complex operations. Throughout this program, you’re going to use BigQuery to practice your SQL skills and collect, prepare, and analyze data. At this point, you have set up your own account. Now, explore some of the important elements of the SQL workspace. This will prepare you for the upcoming activities in which you will use BigQuery. Note that BigQuery updates its interface frequently, so your console might be slightly different from what is described in this reading. That’s okay; use your troubleshooting skills to find what you need!

## Log in to BigQuery

When you log in to BigQuery using the landing page, you will automatically open your project space. This is a high-level overview of your project, including the project information and the current resources being used. From here, you can check your recent activity.

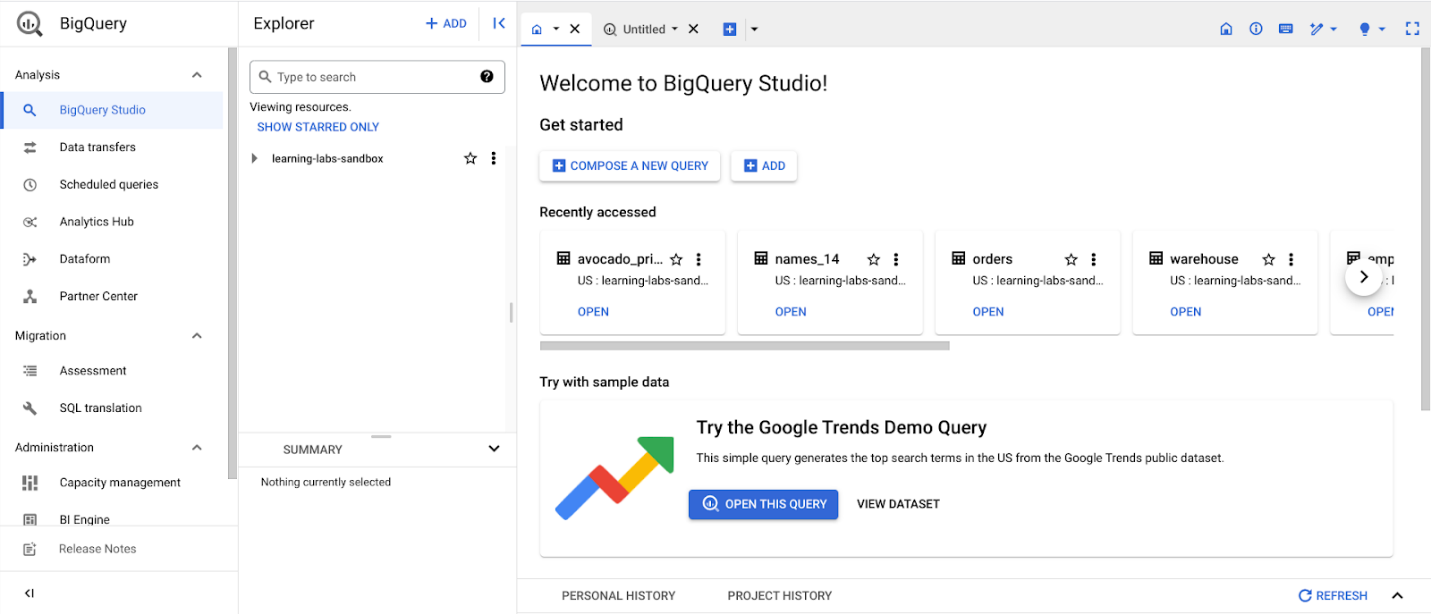


Navigate to your project’s BigQuery Studio by selecting BigQuery from the navigation menu and BigQuery Studio from the dropdown menu.



## BiqQuery Studio components

Once you have navigated to BigQuery from the project space, most of the major components of the BigQuery console will be present: the **Navigation** pane, the **Explorer** pane, and the **SQL Workspace**.

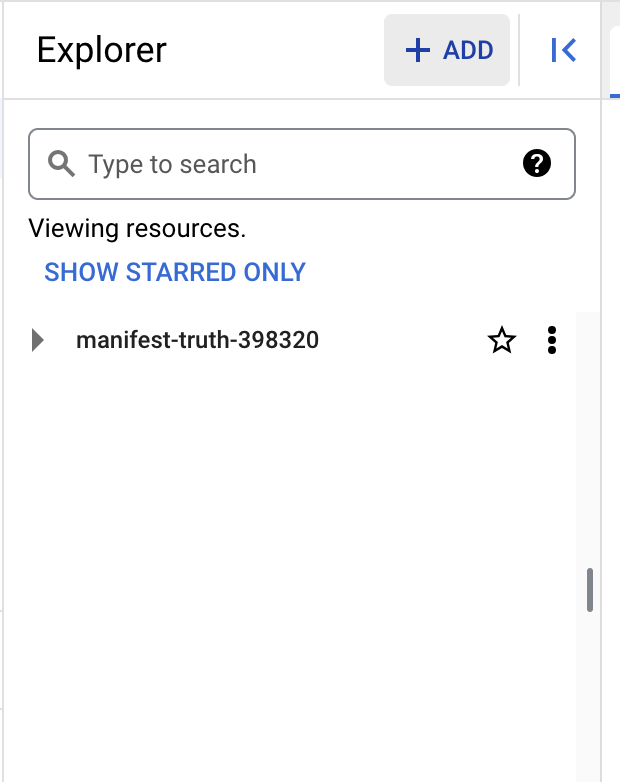


### The Navigation pane

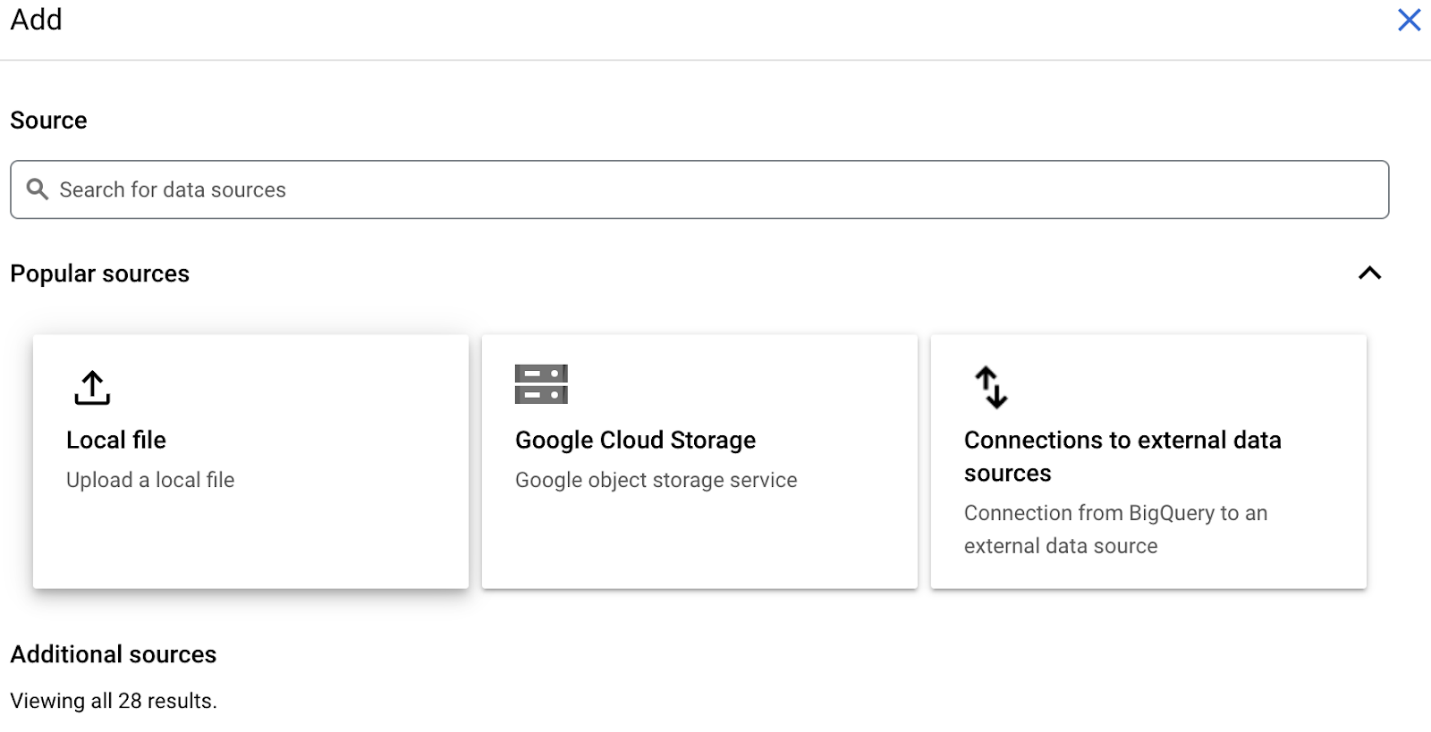
On the console page, find the **Navigation** pane. This is how you navigate from the project space to the BigQuery tool. This menu also contains a list of other Google Cloud Project (GCP) data tools. During this program, you will focus on BigQuery, but it’s useful to understand that the GCP has a collection of connected tools data professionals use every day.

### The Explorer pane

The **Explorer** pane lists your current projects and any starred projects you have added to your console. It’s also where you’ll find the **+ ADD** button, which you can use to add datasets.

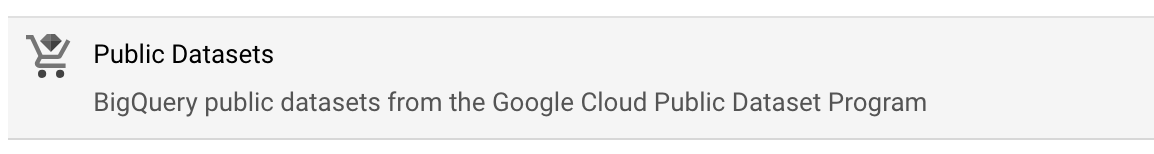


This button opens the **Add** dialog that allows you to open or import a variety of datasets.

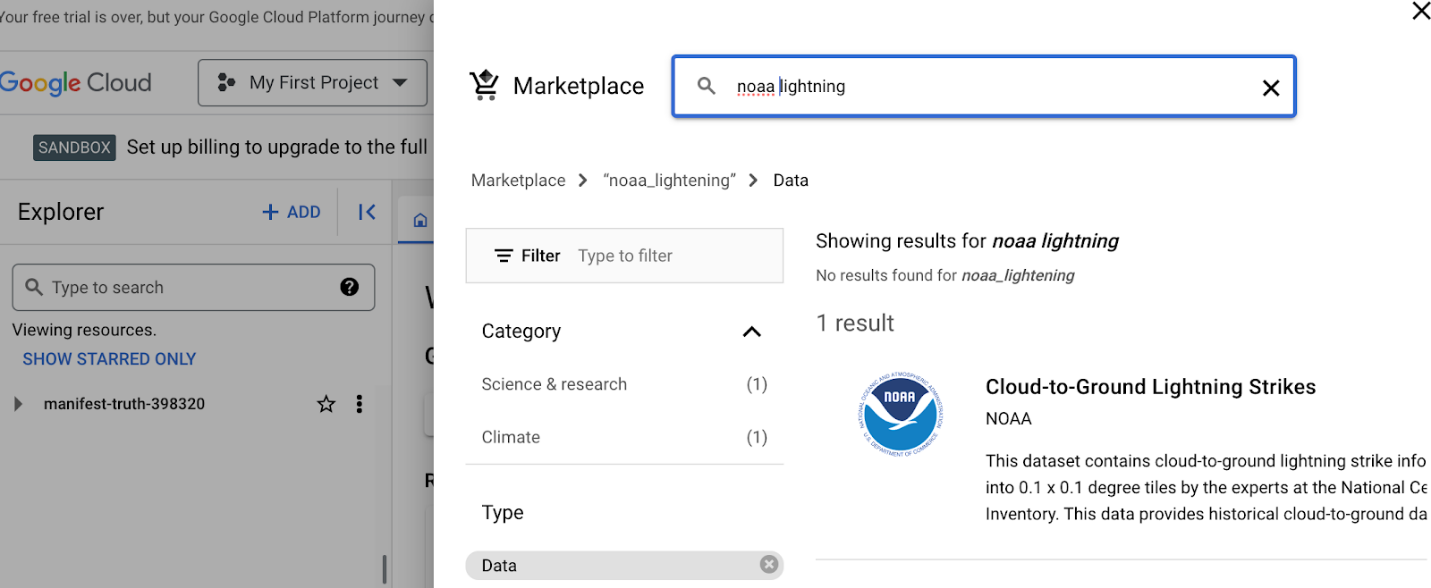


### Add Public Datasets

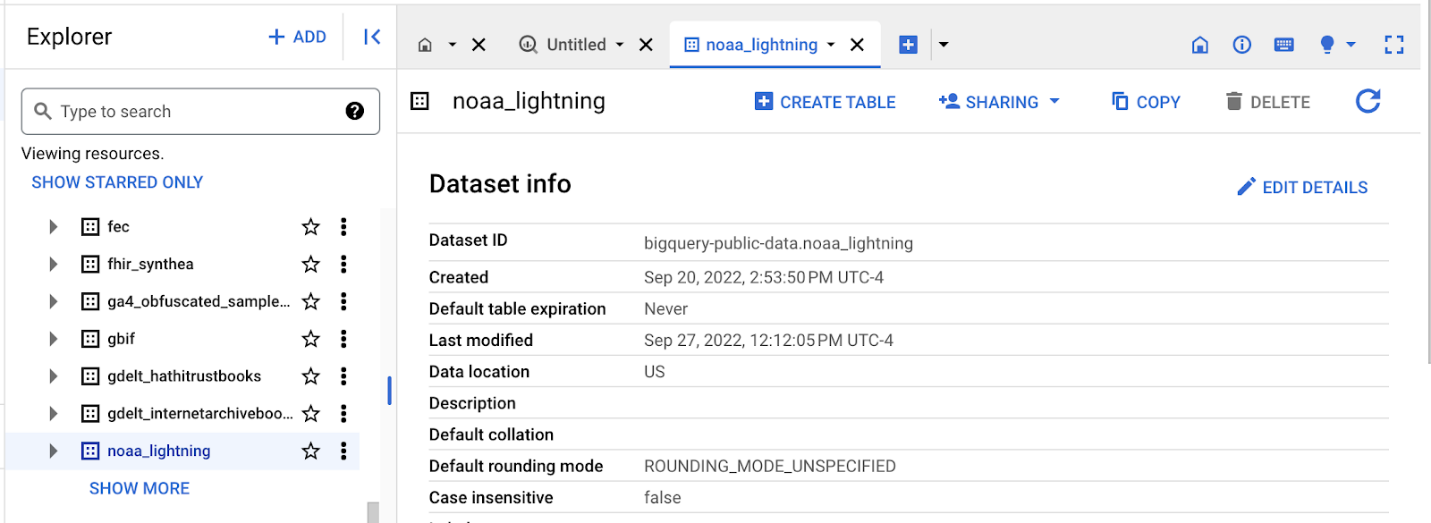
BigQuery offers a variety of public datasets from the Google Cloud Public Dataset Program. Scroll down the **Add** dialog to the **Public Datasets** option.



Select **Public Datasets**. This takes you to the **Public Datasets Marketplace**, where you can search for and select public datasets to add to your BigQuery console. For example, search for the "noaa lightning" dataset in the Marketplace search bar. When you search for this dataset, you will find NOAA’s Cloud-to-Ground Lightning Strikes data.

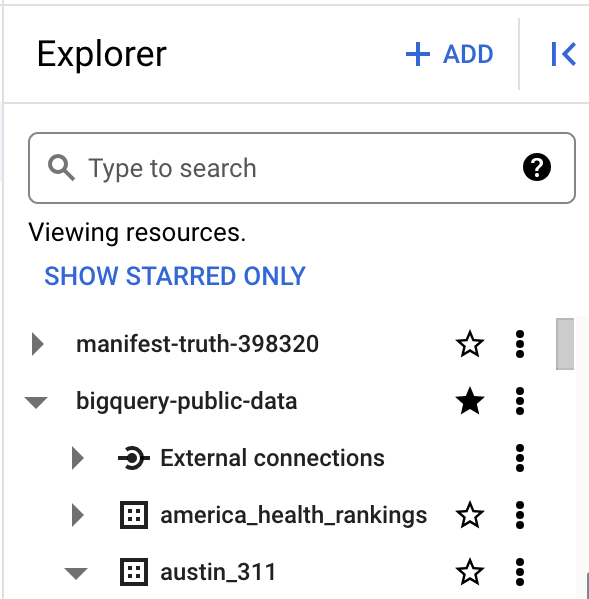


Select the dataset to read its description. Select **View dataset** to create a tab of the dataset’s information within the SQL workspace.

The Explorer Pane lists the noaa\_lightning and other public datasets.

### Star and examine Public Datasets

You added the public noaa\_lightning dataset to your BigQuery Workspace, so the **Explorer** pane displays the noaa\_lightning dataset, along with the list of other public datasets. These datasets are nested under bigquery-public-data. Star bigquery-public-data by navigating to the top of the **Explorer** pane and selecting the star next to bigquery-public-data.



Starring bigquery-public-data will enable you to search for and add public datasets by scrolling in the **Explorer** pane or by searching for them in the **Explorer** search bar.

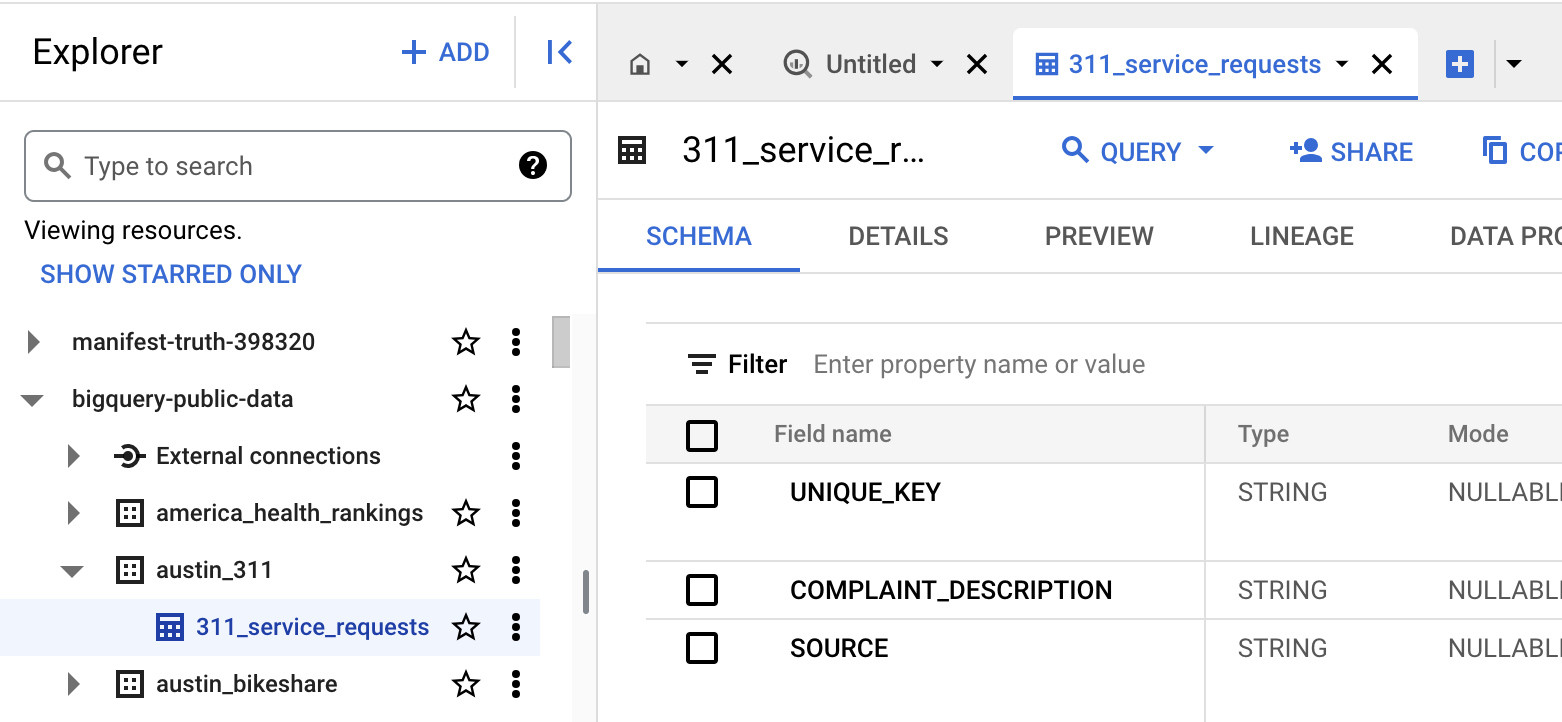
For example, you might want to select a different public dataset. If you select the second dataset, "austin\_311," it will expand to list the table stored in it, “311\_service\_requests.”

A screenshot of a search engine

Description automatically generatedThe Explorer pane with the “bigquery-public data” and “austin\_311” datasets expanded, revealing the “311\_service\_requests” table

When you select a table, its information is displayed in the SQL Workspace. Select the 311\_service\_requests table to examine several tabs that describe it, including:

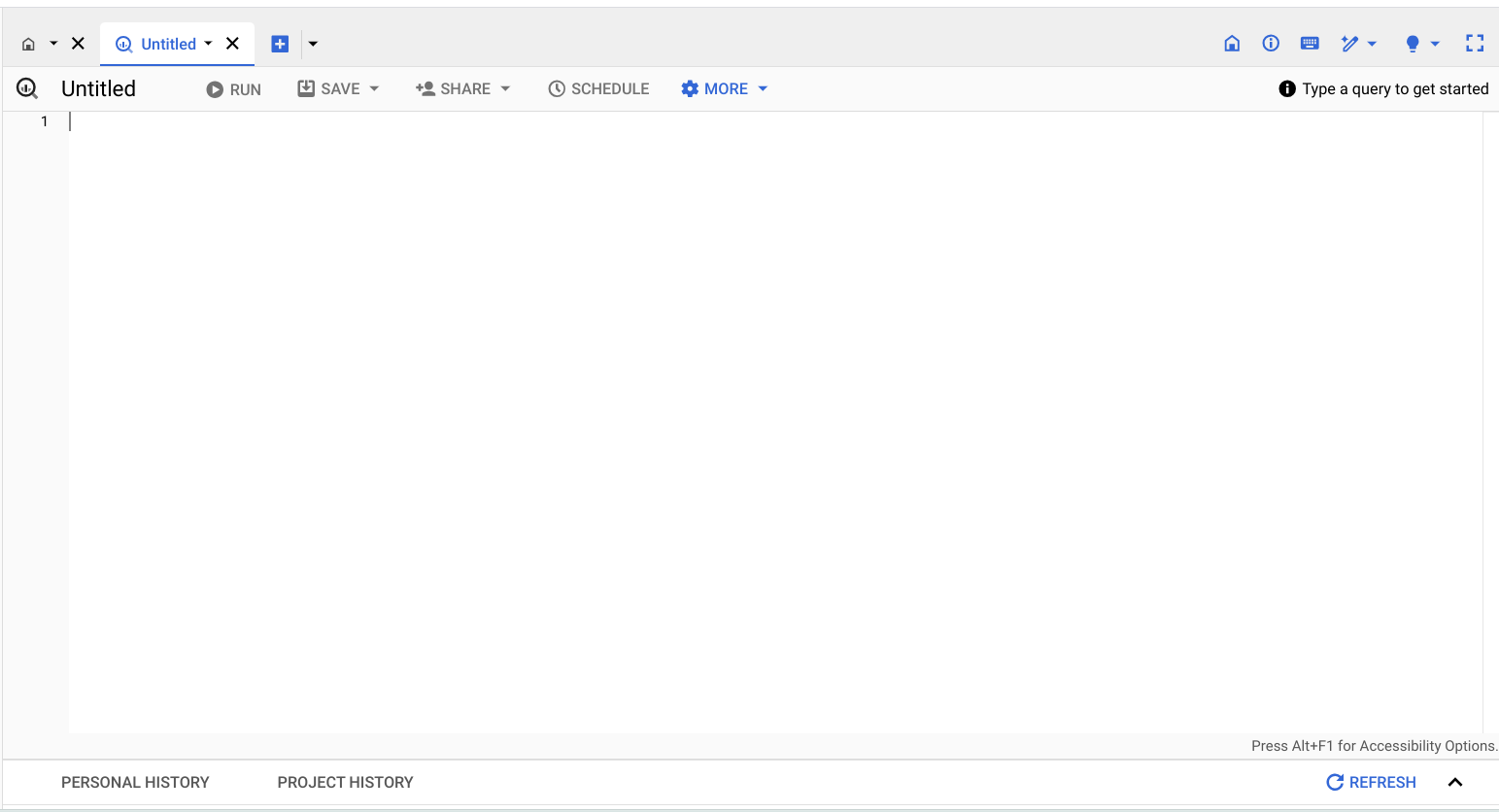
* **Schema**, which displays the column names in the dataset
* **Details**, which contains additional metadata, such as the creation date of the dataset
* **Preview**, which shows the first rows from the dataset



Additionally, you can select the **Query** button from the menu bar in the SQL Workspace to query this table.

### The SQL Workspace

The final menu pane in your console is the SQL Workspace. This is where you will actually write and execute queries in BigQuery.



The SQL Workspace also gives you access to your personal and project history, which stores a record of the queries you’ve run. This can be useful if you want to return to a query to run it again or use part of it in another query.

## Upload your data

In addition to offering access to public datasets, BigQuery also gives you the ability to upload your own data directly into your workspace. Access this feature by opening the **+ ADD** menu again or by clicking the three vertical dots next to your project’s name in the Explorer pane. This will give you the option to create your own dataset and upload your own tables. You will have the opportunity to upload your own data in an upcoming activity to practice using this feature!

## Key takeaways

BigQuery's SQL workspace allows you to search for public datasets, run SQL queries, and even upload your own data for analysis. Whether you're working with public datasets, running SQL queries, or uploading your own data, BigQuery’s SQL workspace offers a range of features to support all kinds of data analysis tasks. Throughout this program, you will be using BigQuery to practice your SQL skills, so being familiar with the major components of your BigQuery console will help you navigate it effectively in the future!

# Optional: Upload the customer dataset to BigQuery

In the next video, the instructor uses a specific dataset. The instructions in this reading are provided for you to upload the same dataset in your BigQuery console.

You must have a BigQuery account to follow along.

## Prepare for the next video

* First, download the .csv file from the attachment below.

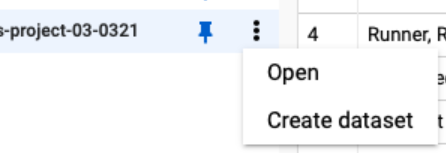
[Customer Table - Sheet1](https://d3c33hcgiwev3.cloudfront.net/F0iSyYcLT9iIksmHCw_Y-Q_191f150a80d74fda96f9df2aa2e3b533_Customer-Table---Sheet1.csv?Expires=1706832000&Signature=eozzUEo7~SrJJBCSxPf73BMSEsSUrfAY9lfqKhk47yOxLMqf0ktAjNpZUXeeZBzCFWwp6aeYSLcKCiBiF6Ohu1cnRpAJk-NMTHjQkg6sIbUdwON1WbIZ1FlHt5QmFBW0sV2308nOh2Z39lmQHdsI9BU2kQZ~2VBpfS3-soOTjls_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[CSV File](https://d3c33hcgiwev3.cloudfront.net/F0iSyYcLT9iIksmHCw_Y-Q_191f150a80d74fda96f9df2aa2e3b533_Customer-Table---Sheet1.csv?Expires=1706832000&Signature=eozzUEo7~SrJJBCSxPf73BMSEsSUrfAY9lfqKhk47yOxLMqf0ktAjNpZUXeeZBzCFWwp6aeYSLcKCiBiF6Ohu1cnRpAJk-NMTHjQkg6sIbUdwON1WbIZ1FlHt5QmFBW0sV2308nOh2Z39lmQHdsI9BU2kQZ~2VBpfS3-soOTjls_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

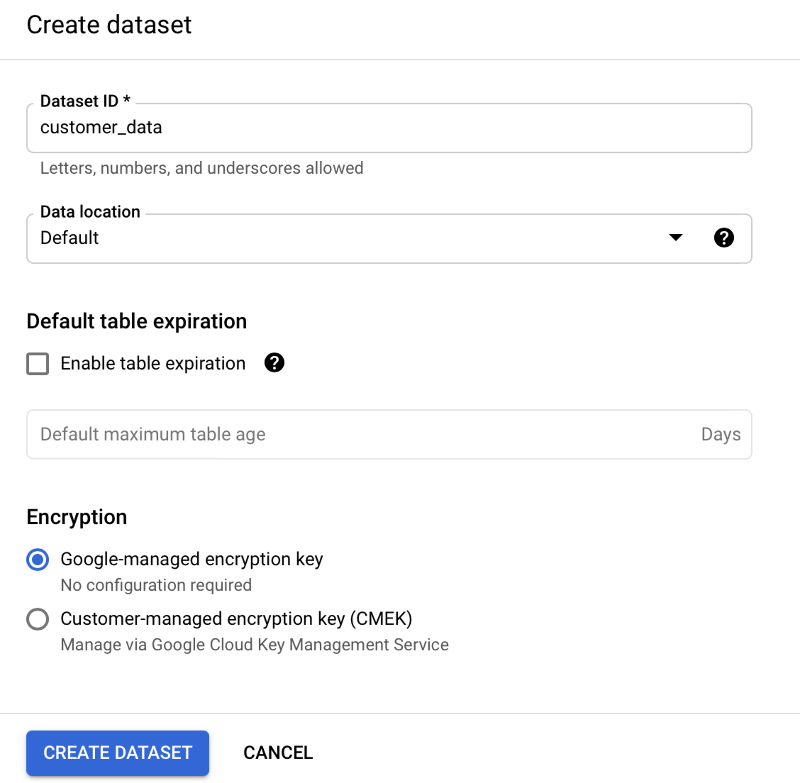
* Next, complete the following steps in your BigQuery console to upload the Customer Table dataset.

**Step 1**: Open your BigQuery console and click on the project you want to upload the data to.

**Step 2:** In the **Explorer** on the left, click the **Actions** icon (three vertical dots) next to your project name and select **Create dataset**.



**Step 3:** In the upcoming video, the name "customer\_data" will be used for the dataset. If you plan to follow along with the video, enter **customer\_data** for the Dataset ID.



**Step 4:** Click **CREATE DATASET** (blue button) to add the dataset to your project.

**Step 5:** In the **Explorer** on the left, click to expand your project, and then click the **customer\_data** dataset you just created.

**Step 6:** Click the **Actions** icon (three vertical dots) next to **customer\_data** and select **Open**.

**Step 7:** Click the blue **+** icon at the middle to open the **Create table** window.

This image shows the Create Table button.

**Step 8:** Under Source, for the **Create table from** selection, choose where the data will be coming from.

* Select **Upload**.
* Click **Browse** to select the Customer Table .csv file you downloaded.
* Choose **CSV** from the file format drop-down.

**Step 9:** For **Table name,** enter **customer\_address** if you plan to follow along with the video.

**Step 10:** For **Schema,** click the **Auto detect** check box.

**Step 11:** Click **Create** **table** (blue button). You will now see the **customer\_address** table under your **customer\_data** dataset in your project.

**Step 12:** Click **customer\_address** and then select the **Preview** tab. Confirm that you see the data shown below.



And now you have everything you need to follow along with the next video. This is also a great table to use to practice querying data on your own. Plus, you can use these steps to upload any other data you want to work with.

0:00

Hey, welcome back.

So far we've learned that SQL has

some of the same tools as spreadsheets,

but on a much larger scale.

In this video, we'll learn some of

the most widely used SQL queries that you can

start using for your own data cleaning

and eventual analysis. Let's get started.

We've talked about queries as requests you put into

the database to ask it to do things for you.

Queries are a big part of using SQL.

It's Structured Query Language, after all.

Queries can help you do a lot of things,

but there are some common ones that data analysts

use all the time. So let's start there.

First, I'll show you how to use the SELECT query.

I've called this one out before,

but now I'll add some new things for us to try out.

Right now, the table viewer is blank

because we haven't pulled anything from the database yet.

For this example, the store we're working

with is hosting a giveaway

for customers in certain cities.

We have a database containing

customer information that we can use to

narrow down which customers are eligible for

the giveaway. Let's do that now.

We can use SELECT to specify

exactly what data we want to interact with in a table.

If we combine SELECT with FROM,

we can pull data from any table in

this database as long as they

know what the columns and rows are named.

We might want to pull the data about

customer names and cities from one of the tables.

To do that, we can input SELECT name, comma,

city FROM

customer underscore data dot customer underscore address.

To get this information from

the customer underscore address table,

which lives in the customer underscore data, data set.

SELECT and FROM help specify

what data we want to extract from the database and use.

We can also insert new data into

a database or update existing data.

For example, maybe we have

a new customer that we want to insert into this table.

We can use the INSERT INTO

query to put that information in.

Let's start with where we're trying to insert this data,

the customer underscore address table.

Play video starting at :2:26 and follow transcript2:26

We also want to specify which columns we're adding

this data to by typing their names in the parentheses.

Play video starting at :2:38 and follow transcript2:38

That way, SQL can tell the database

exactly where we were inputting new information.

Then we'll tell it what values we're putting in.

Play video starting at :3:8 and follow transcript3:08

Run the query, and just like that,

it added it to our table for us.

Now, let's say we just need to

change the address of a customer.

Well, we can tell the database to update it for us.

To do that, we need to tell it we're trying

to update the customer underscore address table.

Play video starting at :3:32 and follow transcript3:32

Then we need to let it know

what value we're trying to change.

Play video starting at :3:45 and follow transcript3:45

But we also need to tell

it where we're making that change

specifically so that it doesn't

change every address in the table.

Play video starting at :3:57 and follow transcript3:57

There. Now this one customer's address has been updated.

If we want to create a new table for this database,

we can use the CREATE TABLE IF NOT EXISTS statement.

Keep in mind, just running a SQL query

doesn't actually create a table for the data we extract.

It just stores it in our local memory.

To save it, we'll need to download it as

a spreadsheet or save the result into a new table.

As a data analyst,

there are a few situations

where you might need to do just that.

It really depends on

what kind of data you're pulling and how often.

If you're only using a total number of customers,

you probably don't need a CSV file

or a new table in your database.

If you're using the total number of customers per day

to do something like track

a weekend promotion in a store,

you might download that data as

a CSV file so you can visualize it in a spreadsheet.

But if you're being asked to pull

this trend on a regular basis,

you can create a table that will

automatically refresh with the query you've written.

That way, you can directly download

the results whenever you need them for a report.

Another good thing to keep in mind,

if you're creating lots of tables within a database,

you'll want to use

the DROP TABLE IF EXISTS statement

to clean up after yourself.

It's good housekeeping.

You probably won't be deleting

existing tables very often.

After all, that's the company's data, and

you don't want to delete

important data from their database.

But you can make sure you're cleaning up

the tables you've personally made so that

there aren't old or unused tables with

redundant information cluttering the database.

There. Now you've seen some of

the most widely used SQL queries in action.

There's definitely more query keywords for you to

learn and unique combinations

that'll help you work within databases.

But this is a great place to start.

Coming up, we'll learn even more about queries in

SQL and how to use them to

clean our data. See you next time.

0:03

Hi, I'm Evan. I'm

a learning portfolio manager here at Google.

I don't think I'm a computer science

or super engineering type,

but I really, really like working with numbers, so actually,

I went into accounting. And

about after two years of accounting I said,

"Wow, I really don't want to do all this by hand,"

so I took my first information systems class,

where they taught me the language SQL or S-Q-L,

and it completely opened up my mind.

Between a working knowledge of spreadsheets where

you change one cell and

the whole spreadsheet changes because

those amazing calculated fields and

SQL where I can query

billions of rows of data in a matter of second,

I was completely sold on my love for data.

I've dedicated my life and

my career to just communicating

that passion and getting folks excited about

the things that they can do with their data.

Why is SQL such an amazing first language to pick up?

Well, there's so many things that you can do with it.

I will first caveat and say,

I am not a computer science major.

I don't know deep down Java and Python,

and I was a little bit apprehensive

of learning a computer language.

It's like a pseudo-programming language, but in reality,

you can write your first SQL statement as you're going to

find out here in just five minutes or less.

SQL, honestly, it's one of those languages

that's easy to learn and even more fun to master.

I've been learning SQL for 15 years.

I've been teaching it for 10.

As you're going to see in some of these

hands-on labs you'll be working through,

it's very easy to return

data from within a database or a data set.

Just select whatever columns from

whichever database that you're

pulling from, and immediately you get the data back.

Now, the really fun part

is actually teasing apart and saying,

I wonder if I change my query,

add these more columns,

filter this data set a different way,

share with my colleagues.

It's meant to be an interactive querying language,

and "query" means "asking a question."

If I can challenge you one thing,

it's that the syntax for picking up SQL,

much like the rules of a chess game,

are very easy to pick up.

But the hard part is actually not

the syntax writing, much

like with any programming language,

but the actual what question

do you want to ask of your data?

What I would encourage you to do is be

super curious about whatever data set that you're given.

Spend a lot of time, even before you touch your keyboard,

in thinking about what data set or what insights you

can get from your data. And then start having fun.

There's many different ways to write

the same correct SQL statement,

so try one out, share it with your friends and then

start returning that data back for insights. Good luck.

0:00

It's so great to have you back.

Now that we know some basic SQL queries

and spent some time working in a database,

let's apply that knowledge to

something else we've been talking about;

preparing and cleaning data.

You already know that cleaning and completing

your data before you analyze it is an important step.

In this video, I'll show you

some ways SQL can help you do just that.

Including how to remove duplicates as well as

four functions to help you clean string variables.

Earlier we covered how to remove duplicates and

spreadsheets using the Remove duplicates tool.

In SQL, we can do the same thing

by including Distinct in our Select statement.

For example, let's say the company we work

for has a special promotion for customers in Ohio.

We want to get the customer IDs

of customers who live in Ohio,

but some customer information

has been entered multiple times.

We can get these customer IDs by writing,

select customer\_id from customer\_data.customer\_address.

This query will give us

duplicates if they exist in the table.

If customer ID 9080 shows up three times in our table,

our results will have three of

that customer ID, but we don't want that.

We want a list of all unique customer IDs.

To do that, we add

Distinct to our Select statement by writing,

Select Distinct customer\_

id from customer\_data.customer\_address.

Play video starting at :1:51 and follow transcript1:51

Now the customer ID

9080 will show up only once in our results.

You might remember we talked before about

text strings as a group of characters within a cell,

commonly composed of letters, numbers or both.

These text strings need to be clean sometimes.

Maybe they've been entered differently in

different places across your database,

and now they don't match.

In those cases, you'll need to

clean them before you can analyze them.

Here are some functions you can use in

SQL to handle string variables.

You might recognize some of these functions

from when we talked about spreadsheets.

Now it's time to see them work in a new way.

Pull up the data set we shared right before this video,

and you can follow along step by

step with me during the rest of this video.

The first function I want to show you is length,

which we've encountered before.

If we already know the length

our string variables are supposed to be,

we can use length to double

check that our string variables are consistent.

For some databases, this query is written as len,

but it does the same thing.

Let's say we're working with

the customer\_address table from our earlier example,

we can make sure that all country codes have

the same length by using Length on each of these strings.

To write our SQL query,

let's first start with Select and From.

We know our data comes from

the customer\_address table

within the customer\_data dataset.

We add customer\_ data.customer

\_address after the From clause.

Then under Select,

we'll write Length,

and then the column we want to check; country.

To remind ourselves what this is,

we can label this column in

our results as letter\_in \_country.

We add as letters\_in

\_country after Length parentheses country.

The result we get is a list of the number of

letters in each country listed for each of our customers.

It seems that almost all of them are twos,

which means the country field contains only two letters,

but we notice one that has three.

That's not good. We want our data to be consistent.

Let's check out which countries

were incorrectly listed in our table.

We can do that by putting the Length parenthesis,

country parentheses function that

we created into the Where clause,

because we're telling SQL to filter the data to show

only customers whose country

contains more than two letters.

Now we'll write, Select country

from customer\_data.customer\_address,

where Length parentheses, country

parentheses greater than 2.

When we run this query,

we now get the two countries where the number

of letters is greater than the 2 we expect to find.

The incorrectly listed countries show

up as USA instead of US.

If we created this table,

then we could update our table so that this entries

shows up as US instead of USA.

But in this case, we didn't create this table,

so we shouldn't update it.

We still need to fix this problem,

so we can pull a list of all the customers in the US,

including the two that have USA instead of US.

The good news, is that,

we can account for this error in our results by

using the sub-string function in our SQL query.

To write our SQL query,

let's start by writing the basic structure.

Select, From, Where.

We know our data is coming from

the customer\_address table

from the customer\_data dataset.

We type in customer\_data.customer \_ address.

After from. Next,

we tell SQL what data we want it to give us.

We want all the customers in the US by their IDs,

so we type into customer\_id after select.

Finally, we want SQL to filter

out only American customers,

so we use the substring function after the where clause.

We're going to use the substring function

to pull the first two letters of

each country so that all of them are

consistent and only contain two letters.

To use the substring function,

we first need to tell SQL,

the column where we found this error country.

Then we specify which letter to start with.

We want SQL to pull the first two letters,

so we're starting with the first letter,

so we type in one.

Then we need to tell SQL how many letters,

including this first letter to pull.

Since we want the first two letters,

we need SQL to pull

two total letters, so we type in two.

This will give us the first two letters of each country.

We want US only,

so we'll set this function to equals US.

When we run this query,

we get a list of all customer IDs

of customers whose country is the US,

including the customers that had USA instead of US.

Going through our results,

it seems like we have a couple duplicates where

the customer ID is shown multiple times.

Remember how we get rid of duplicates.

We add distinct before customer underscore ID.

Play video starting at :8:4 and follow transcript8:04

Now when we run this query,

we have our final list of

customer IDs of the customers who live in the US.

Finally, let's check out

the trim function which you've come across before.

This is really useful if you find entries with

extra spaces and need to

eliminate those extra spaces for consistency.

For example, let's check out

the state column in our customer\_address table.

Just like we did for the country column,

we want to make sure the state column

has the consistent number of letters.

Let's use the length function again to learn if

we have any state that has more than two letters,

which is what we would expect to find in our data table.

We start writing our SQL query by

typing the basic SQL structure of

, select from where.

We're working with the customer\_address table

in the customer\_data dataset,

so we type in customer\_data,

dot customer\_address after from.

Next, we tell SQL what we want it to pull.

We want it to give us

any state that has more than two letters,

so we type in state after select.

Finally, we want SQL

to filter for states that have more than two letters.

This condition is written in the where clause.

We type in length,

parentheses state, parentheses,

and that it must be greater than

two because we want

the states that have more than two letters.

We want to figure out what

the incorrectly listed states look like, if we have any.

When we run this query,

we get one result.

We have one state that has

more than two letters. But hold on.

How can this state that seems like it has two letters,

O and H for Ohio have more than two letters?

We know that there are more than two characters

because we use the length parentheses state,

parentheses greater than two statement

in the where clause when filtering our results.

That means the extra characters that SQL

is counting must then be a space.

There must be a space after the age.

This is where we would use the trim function.

The trim function removes any spaces.

Let's write a SQL query that accounts for this error.

Let's say we want a list of

all customer IDs of the customers who live in,

OH, for Ohio.

We start with the basic SQL structure.

Select from where.

We know the data comes from,

the customer\_address table,

and the customer\_data dataset.

We type in customer\_data.customer\_address after from.

Next, we tell SQL what data we want.

We want SQL to give us

the customer IDs of customers who live in Ohio.

We type in customer\_id after select.

Since we know we have some duplicate customer entries,

we'll go ahead and type in distinct before

customer ID to remove

any duplicate customer IDs from appearing in our results.

Finally, we want SQL to give us

the customer IDs of the customers who live in Ohio.

We're asking SQL to filter the data.

This belongs in the where clause.

Here's where we'll use the trim function.

To use the trim function,

we tell SQL the column we want to remove spaces from,

which is state in our case,

and we want only Ohio customers,

so we type in equals OH.

That's it. We have

all customer IDs of the customers who live in Ohio,

including that customer with the extra space after the

H. Making sure that

your string variables are

complete and consistent will save you

a lot of time later by

avoiding errors or miscalculations.

That's why we clean data in the first place.

Hopefully, functions like length, substring,

and trim will give you the tools you need to

start working with string variables in your own datasets.

Next up, we'll check out some other ways you can work

with strings and more advanced cleaning functions.

Then you'll be ready to start working in

SQL on your own. See you soon.

## Cleaning your data



Your new dataset contains historical sales data, including details such as car features and prices. You can use this data to find the top 10 most popular cars and trims. But before you can perform your analysis, you’ll need to make sure your data is clean. If you analyze dirty data, you could end up presenting the wrong list of cars to the investors. That may cause them to lose money on their car inventory investment.

### Step 1: Inspect the fuel\_type column

The first thing you want to do is inspect the data in your table so you can find out if there is any specific cleaning that needs to be done. Get an initial understanding of the data table by clicking on the **PREVIEW** tab that sits below the **car\_info** toolbar.

A screenshot of a computer

Description automatically generated

According to the [data’s description](https://archive.ics.uci.edu/ml/datasets/Automobile), the **fuel\_type** columnshould only have two unique string values: **diesel** and **gas**. To check and make sure that’s true, run the following query. You can generate the default query setup by clicking on the **QUERY** button and selecting the **In split tab**. This will give you a dual view of the info window and the query.

A screenshot of a computer

Description automatically generated

Next, we can generate the first query in the workspace:

1

2

3

4

5

SELECT

  DISTINCT fuel\_type

FROM

  your project name.cars.car\_info

LIMIT 1000

**NOTE**: Within the **FROM** clause of the syntax above, you will need to begin the **Table ID** line with your personalized project name, period, the dataset name, period, and end with the table name. It's important to understand that the personal project name will be unique to each learner. You can also locate and copy the full **Table ID** filename by clicking on the **DETAIL** option tab in your **car\_info** **Table info** window. Once copied, paste it after the **FROM** clause and run the above query.

This returns the following results:

A screenshot of a computer

Description automatically generated

This confirms that the fuel\_type column doesn’t have any unexpected values. Also note that the default **LIMIT 1000** is added to your query, but in this case, BigQuery is only returning two distinct fuel types.

### Step 2: Inspect the length column

Next, you will inspect a column with numerical data. The length column should contain numeric measurements of the cars. So you will check that the minimum and maximum lengths in the dataset align with the [data description](https://archive.ics.uci.edu/ml/datasets/Automobile), which states that the lengths in this column should range from 141.1 to 208.1. Run this query to confirm:

1

2

3

4

5

SELECT

  MIN(length) AS min\_length,

  MAX(length) AS max\_length

FROM

  you project name.cars.car\_info;

Your results should confirm that 141.1 and 208.1 are the minimum and maximum values respectively in this column.

A screenshot of a cell phone

Description automatically generated

### Step 3: Fill in missing data

Missing values can create errors or skew your results during analysis. You’re going to want to check your data for null or missing values. These values might appear as a blank cell or the word ***null*** in BigQuery.

You can check to see if the **num\_of\_doors** column contains null values using this query:

1

2

3

4

5

6

SELECT

  \*

FROM

  your project name.cars.car\_info

WHERE

  num\_of\_doors IS NULL;

This will select any rows with missing data for the **num\_of\_doors** column and return them in your results table. You should get two results, one Mazda and one Dodge:

A screenshot of a computer

Description automatically generated

In order to fill in these missing values, you check with the sales manager, who states that all Dodge gas sedans and all Mazda diesel sedans sold had four doors. If you are using the BigQuery free trial, you can use this query to update your table so that all Dodge gas sedans have four doors:

1

2

3

4

5

6

7

8

UPDATE

  your project name.cars.car\_info

SET

  num\_of\_doors = "four"

WHERE

  make = "dodge"

  AND fuel\_type = "gas"

  AND body\_style = "sedan";

You should get a message telling you that three rows were modified in this table. To make sure, you can run the previous query again:

1

2

3

4

5

6

SELECT

  \*

FROM

  your project name.cars.car\_info

WHERE

  num\_of\_doors IS NULL;

Now, you only have one row with a **NULL** value for num\_of\_doors.Repeat this process to replace the null value for the Mazda.

If you are using the BigQuery Sandbox, you can skip these **UPDATE** queries; they will not affect your ability to complete this activity.

### Step 4: Identify potential errors

Once you have finished ensuring that there aren’t any missing values in your data, you’ll want to check for other potential errors. You can use **SELECT DISTINCT** to check what values exist in a column. You can run this query to check the num\_of\_cylinders column:

1

2

3

4

SELECT

  DISTINCT num\_of\_cylinders

FROM

  your project name.cars.car\_info;

After running this, you notice that there are one too many rows.There are two entries for two cylinders: rows 6 and 7. But the two in row 7 is misspelled.

A screenshot of a number

Description automatically generated

To correct the misspelling for all rows, you can run this query if you have the BigQuery free trial:

1

2

3

4

5

6

UPDATE

  your project name.cars.car\_info

SET

  num\_of\_cylinders = "two"

WHERE

  num\_of\_cylinders = "tow";

You will get a message alerting you that one row was modified after running this statement. Tocheck that it worked, you can run the previous query again:

1

2

3

4

SELECT

  DISTINCT num\_of\_cylinders

FROM

  your project name.cars.car\_info;

Next, you can check the compression\_ratio column. According to the [data description](https://archive.ics.uci.edu/ml/datasets/Automobile),the compression\_ratio column values should range from 7 to 23. Just like when you checked the length values , you can use **MIN** and **MAX** to check if that’s correct:

5

2

3

4

1

  your project name.cars.car\_info;

  MIN(compression\_ratio) AS min\_compression\_ratio,

  MAX(compression\_ratio) AS max\_compression\_ratio

FROM

SELECT

Notice that **this returns a maximum of 70**. But you know this is an error because the maximum value in this column should be 23, not 70. So the 70 is most likely a 7.0. Run the above query again without the row with 70 to make sure that the rest of the values fall within the expected range of 7 to 23.

1

2

3

4

5

6

7

SELECT

  MIN(compression\_ratio) AS min\_compression\_ratio,

  MAX(compression\_ratio) AS max\_compression\_ratio

FROM

  your project name.cars.car\_info

WHERE

  compression\_ratio <> 70;

Now the highest value is 23, which aligns with the data description. So you’ll want to correct the 70 value. You check with the sales manager again, who says that this row was made in error and should be removed. Before you delete anything, you should check to see how many rows contain this erroneous value as a precaution so that you don’t end up deleting 50% of your data. If there are too many (for instance, 20% of your rows have the incorrect 70 value), then you would want to check back in with the sales manager to inquire if these should be deleted or if the 70 should be updated to another value. Use the query below to count how many rows you would be deleting:

1

2

3

4

5

6

SELECT

   COUNT(\*) AS num\_of\_rows\_to\_delete

FROM

   your project name.cars.car\_info

WHERE

   compression\_ratio = 70;

Turns out there is only one row with the erroneous 70 value. So you can delete that row using this query:

1

2

DELETE your project name.cars.car\_info

WHERE compression\_ratio = 70;

If you are using the BigQuery sandbox, you can replace **DELETE** with **SELECT** to see which row would be deleted.

### Step 5: Ensure consistency

Finally, you want to check your data for any inconsistencies that might cause errors. These inconsistencies can be tricky to spot—sometimes even something as simple as an extra space can cause a problem.

Check the **drive\_wheels** column for inconsistencies by running a query with a **SELECT DISTINC**T statement:

1

2

3

4

5

 SELECT

  DISTINCT drive\_wheels

FROM

  your project name.cars.car\_info;

It appears that 4wd appears twice in results. However, because you used a **SELECT DISTINCT** statement to return unique values, this probably means there’s an extra space in one of the 4wd entries that makes it different from the other 4wd.

A screenshot of a phone

Description automatically generated

To check if this is the case, you can use a **LENGTH** statementto determine the length of how long each of these string variables:

1

2

3

4

5

SELECT

  DISTINCT drive\_wheels,

  LENGTH(drive\_wheels) AS string\_length

FROM

  your project name.cars.car\_info;

According to these results, some instances of the 4wd string have four characters instead of the expected three (4wd has 3 characters). In that case, you can use the **TRIM** function to remove all extra spaces in the drive\_wheels column if you are using the BigQuery free trial:

1

2

3

4

5

UPDATE

  your project name.cars.car\_info

SET

  drive\_wheels = TRIM(drive\_wheels)

WHERE TRUE;

Then, you run the **SELECT DISTINCT** statement againto ensure that there are only three distinct values in the drive\_wheels column:

1

2

3

4

 SELECT

  DISTINCT drive\_wheels

FROM

  your project name.cars.car\_info;

And now there should only be three unique values in this column! Which means your data is clean,  consistent, and ready for analysis!

### 1.

Question 1

What are some key benefits of using SQL for data analytics projects? Select all that apply.

1 point

Manage huge amounts of data

Insertable images and text formatting

Adaptable for multiple database programs

Powerful data cleaning tools

### 2.

Question 2

Which SQL function cleans string variables by extracting a substring from a string variable?

1 point

**COUNTIF**

**SUBSTR**

**LENGTH**

**LEN**

### 3.

Question 3

You are working with a database table that contains data about playlists, and you discover there are duplicate entries. What SQL clause will remove the duplicates from the playlist\_id column?

1 point

**SELECT DUPLICATE playlist\_id**

**SELECT ONLY playlist\_id**

**SELECT DISTINCT playlist\_id**

**SELECT REMOVE playlist\_id**

### 4.

Question 4

You are working with a database table that contains data about turtles. What SQL clause will return any turtle ages that are less than three digits long from the turtle\_age column?

1 point

**LENGTH(turtle\_age) > 3**

**LENGTH turtle\_age > 3**

**LENGTH turtle\_age < 3**

**LENGTH(turtle\_age) < 3**

### 5.

Question 5

You are working with a database table that contains data about cookbooks. What SQL clause will retrieve the first eight letters of each data point in the recipe\_name column, then store the result in a new column called recipe\_listing?

1 point

**SUBSTR(recipe\_name, 1, 8) new\_recipe\_listing**

**SUBSTR(recipe\_name, 8) to recipe\_listing**

**SUBSTR(recipe\_name, 1, 8) AS recipe\_listing**

**SUBSTR(recipe\_name, 8) AS recipe\_listing**

# Optional: Upload the store transactions dataset to BigQuery

In the next video, the instructor uses a specific dataset. The instructions in this reading are provided for you to upload the same dataset in your BigQuery console so you can follow along.

You must have a BigQuery account to follow along.

## Prepare for the next video

* First, download the .csv file from the attachment below.

[Lauren's Furniture Store Transaction Table](https://d3c33hcgiwev3.cloudfront.net/0cvJS5ocTSu-77CVZvn6kw_55078d160a924e49aca0837f03d995f1_Lauren-s-Furniture-Store-Transaction-Table.csv?Expires=1706832000&Signature=FBQePOTmdfmSjsQFqhd3Ct986Q6VbeIvclDWYWB9hEJyLIe4tnWx8VVVtXjfeX6O7PabaMSD3iV7D85jSY0JwqU2t8wXKh93hsPS7TtzTeM4x9mT9XNUkrtxQkcz6ozLMLVeiINQVgv08CsO31OU4h10WpFS~eUBZkdkyHPsUPA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

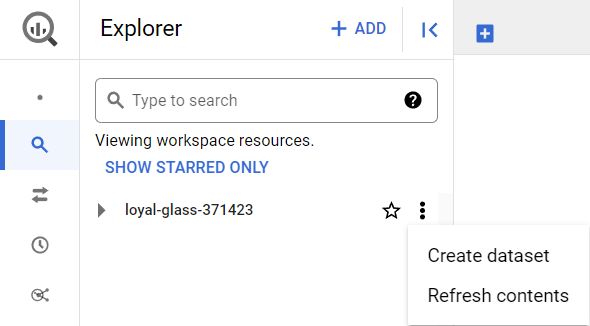
[CSV File](https://d3c33hcgiwev3.cloudfront.net/0cvJS5ocTSu-77CVZvn6kw_55078d160a924e49aca0837f03d995f1_Lauren-s-Furniture-Store-Transaction-Table.csv?Expires=1706832000&Signature=FBQePOTmdfmSjsQFqhd3Ct986Q6VbeIvclDWYWB9hEJyLIe4tnWx8VVVtXjfeX6O7PabaMSD3iV7D85jSY0JwqU2t8wXKh93hsPS7TtzTeM4x9mT9XNUkrtxQkcz6ozLMLVeiINQVgv08CsO31OU4h10WpFS~eUBZkdkyHPsUPA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

* Next, complete the steps below in your BigQuery console to upload the Store Transaction dataset.

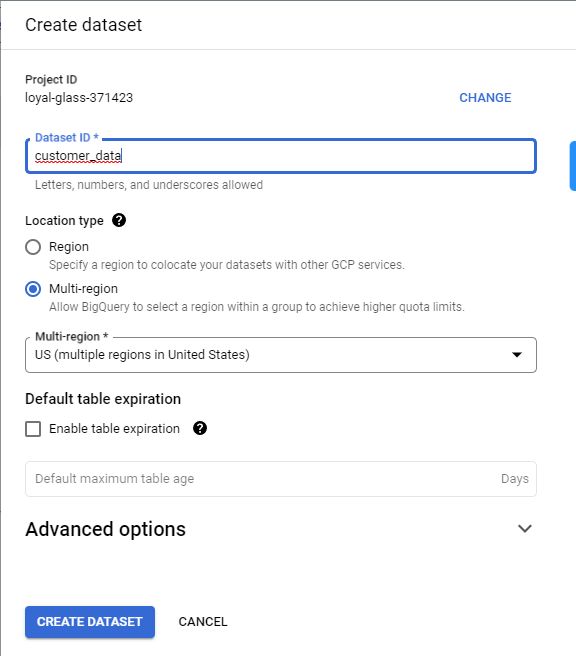
**Note:** These steps will be different from what you performed before. In previous instances, you selected the **Auto detect** check box to allow BigQuery to auto-detect the schema. This time, you will choose to create the schema by editing it as text. This method can be used when BigQuery doesn't automatically set the desired type for a particular field. In this case, you will specify **STRING** instead of **FLOAT** as the type for the purchase\_price field.

**Step 1**: Open your BigQuery console and click on the project you want to upload the data to. If you already created a **customer\_data** dataset for your project, jump to step 5; otherwise, continue with step 2.

**Step 2:** In the **Explorer** on the left, click the **Actions** icon (three vertical dots) next to your project name and select **Create dataset**.

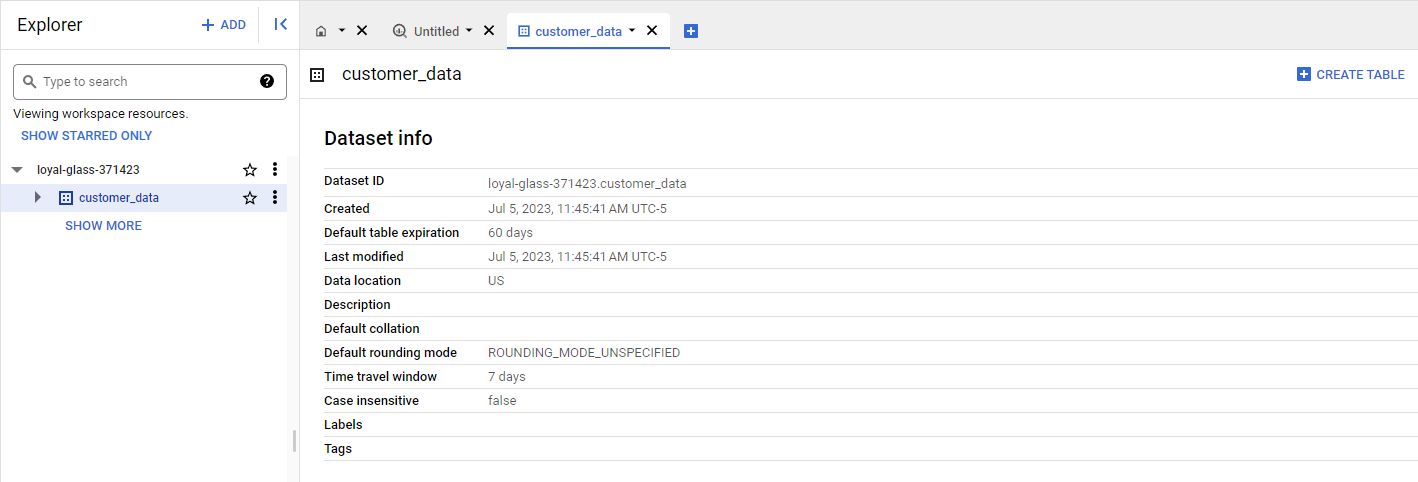


**Step 3:** In the **Create dataset** window, enter **customer\_data** for the Dataset ID. Make sure the Location type is set to **Multi-region, US (Multiple regions in United States),** and all the default Advanced options remain set to the **Google-managed encryption key** option.



**Step 4:** Click **CREATE DATASET** (blue button) to add the dataset to your project.

**Step 5:** In the **Explorer** pane, click on the expansion arrow under your project name, and then click the **customer\_data** dataset.



**Step 6:** On the far right hand side of the new **Dataset info** page, click the blue **+ CREATE TABLE** button to open the **Create table** window. Use the visual settings in the next image to complete the steps below.



**Step 7:** Under **Source**, for the **Create table from** selection, choose where the data will be coming from.

* Select **Upload**.
* Click **Browse** to select the Store Transaction Table .csv file you downloaded.
* Choose **CSV** from the file format drop-down.

**Step 8:** For Table name, enter **customer\_purchase** if you plan to follow along with the video.

**Step 9:** For Schema, click the toggle switch for **Edit as text**. This opens up a box for the text.

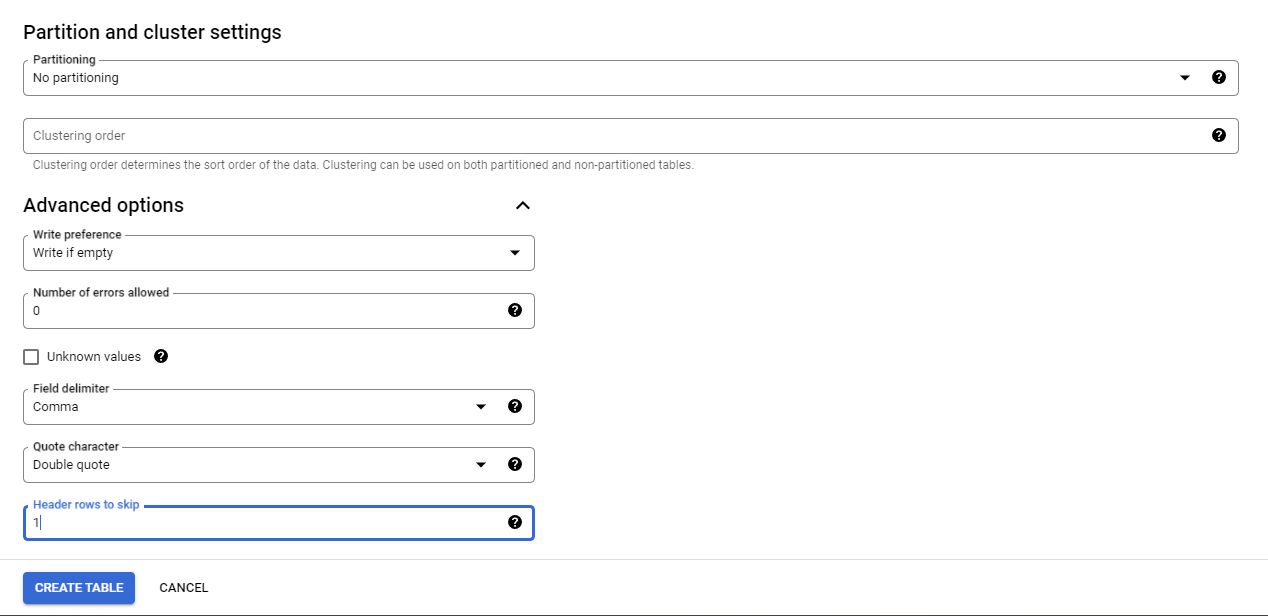
**Step 10:** Copy and paste the following text into the box. Be sure to include the opening and closing brackets. They are required.

[ { "description": "date", "mode": "NULLABLE", "name": "date", "type": "DATETIME" }, { "description": "transaction id", "mode": "NULLABLE", "name": "transaction\_id", "type": "INTEGER" }, { "description": "customer id", "mode": "NULLABLE", "name": "customer\_id", "type": "INTEGER" }, { "description": "product name", "mode": "NULLABLE", "name": "product", "type": "STRING" }, { "description": "product\_code", "mode": "NULLABLE", "name": "product\_code", "type": "STRING" }, { "description": "product color", "mode": "NULLABLE", "name": "product\_color", "type": "STRING" }, { "description": "product price", "mode": "NULLABLE", "name": "product\_price", "type": "FLOAT" }, { "description": "quantity purchased", "mode": "NULLABLE", "name": "purchase\_size", "type": "INTEGER" }, { "description": "purchase price", "mode": "NULLABLE", "name": "purchase\_price", "type": "STRING" }, { "description": "revenue", "mode": "NULLABLE", "name": "revenue", "type": "FLOAT" } ]

**Step 11:** Scroll down and expand the **Advanced options** section.

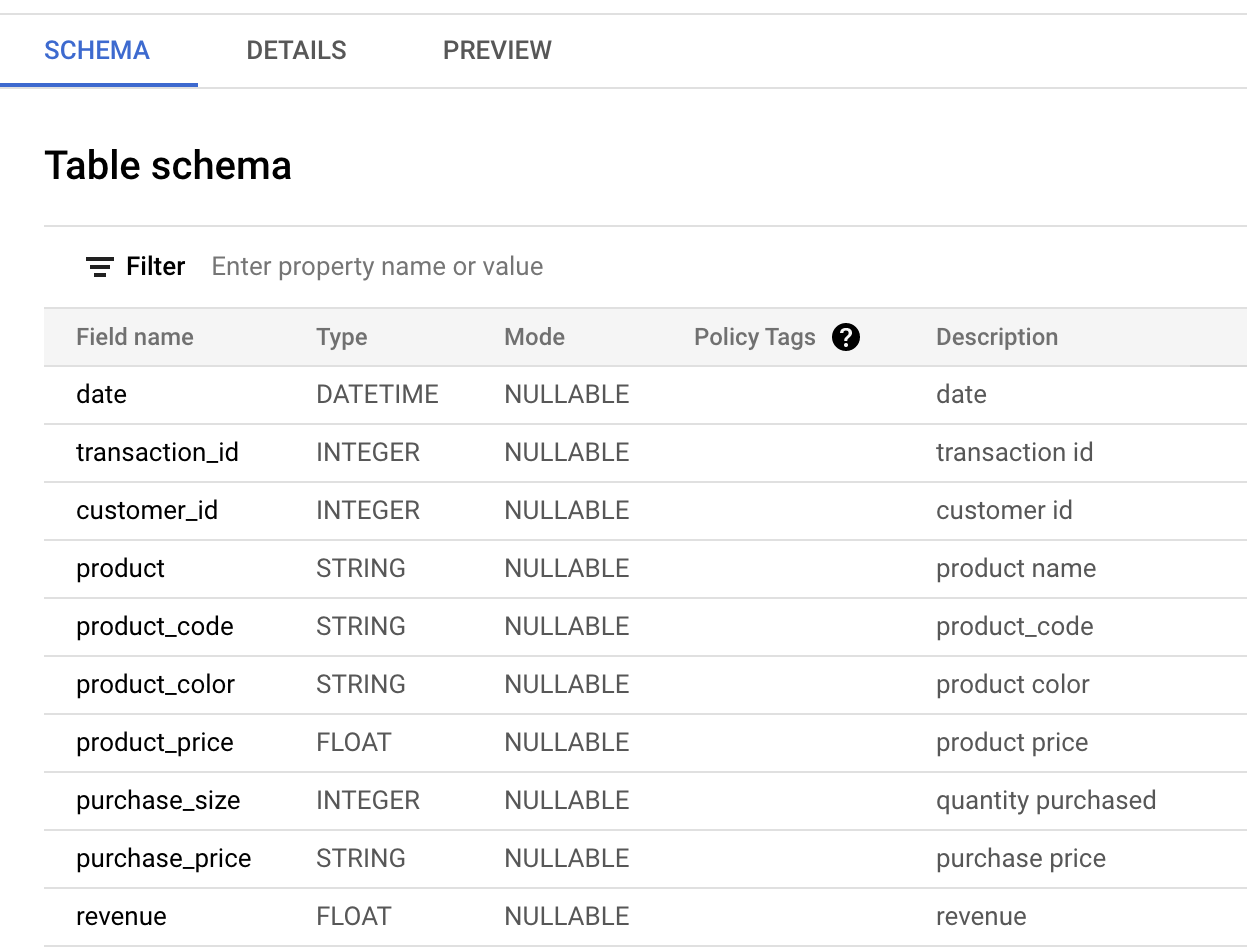
**Step 12:** For the **Header rows to skip** field, enter **1**.

**NOTE**: It is very important that you don't skip the last step, or you will receive 'parsing' errors, as BigQuery will try to apply the schema editing functions to the title row.

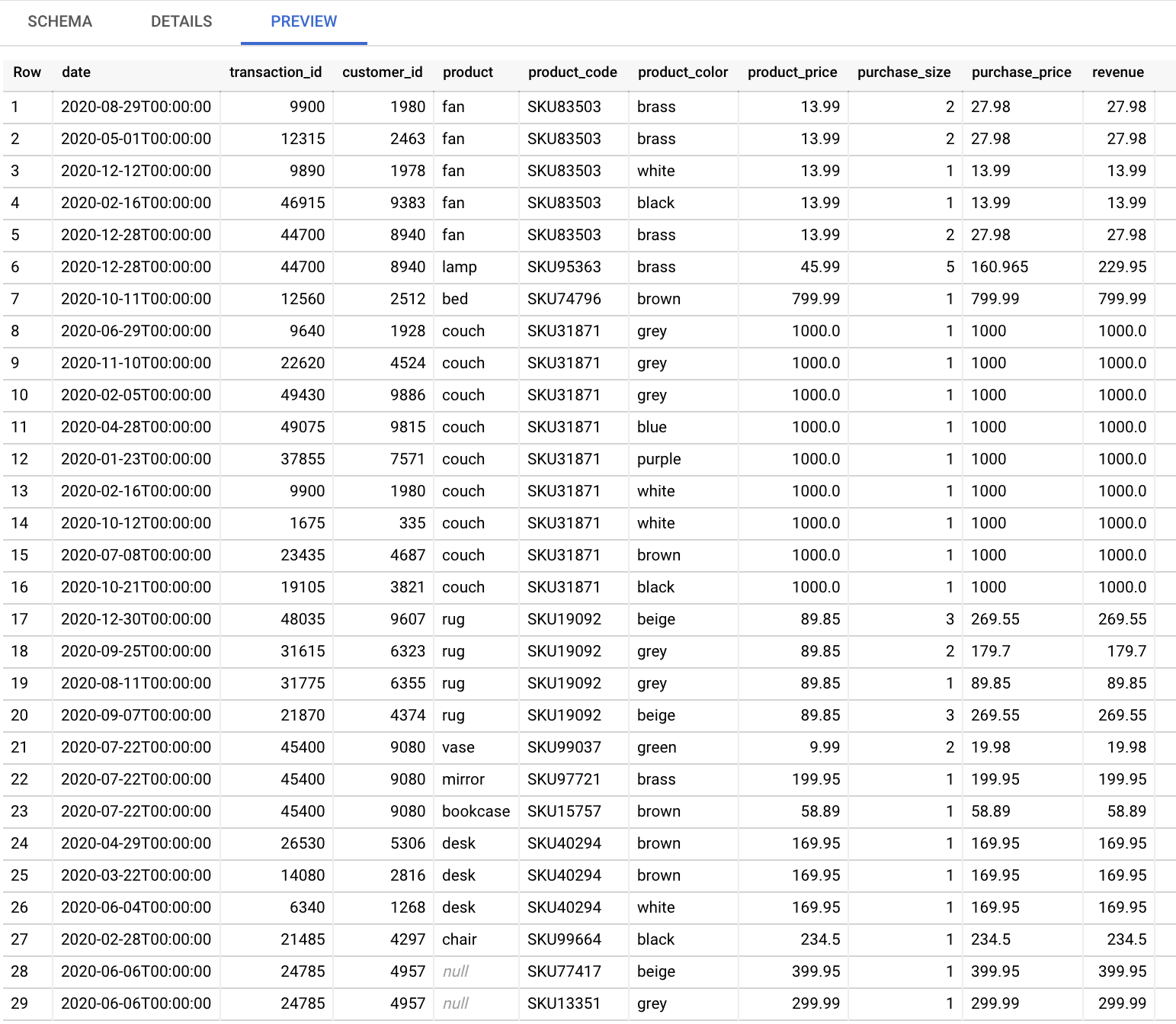


**Step 13:** Click **Create** **table** (blue button). You will now see the **customer\_purchase** table under your **customer\_data** dataset in your **Explorer** pane.

**Step 14:** Click the **customer\_purchase** table and in the **Schema** tab, confirm that the schema matches the schema shown below.



**Step 15:** Click the **Preview** tab and confirm that your data matches the data shown below.



Congratulations, you are now ready to follow along with the video!

0:01

Hi there and welcome back.

So far we've gone over some basic SQL queries and

functions that can help you clean your data.

We've also checked out some ways you can deal with string variables in SQL to make

your job easier.

Get ready to learn more functions for dealing with strings in SQL.

Trust me, these functions will be really helpful in your work as a data analyst.

In this video, we'll check out strings again and

learn how to use the cast function to correctly format data.

When you import data that doesn't already exist in your SQL tables,

the data types from the new dataset might not have been imported correctly.

This is where the CAST function comes in handy.

Basically, CAST can be used to convert anything from one data type to another.

Let's check out an example.

Imagine we're working with Lauren's Furniture Store.

The owner has been collecting transaction data for the past year, but

she just discovered that they can't actually organize their data because it

hadn't been formatted correctly.

So we'll help her by converting her data to make it useful again.

For example, let's say we want to sort all purchases by

purchase\_price in descending order.

That means we want the most expensive purchase to show up first in our results.

To write the SQL query, we start with the basic SQL structure.

SELECT, FROM, WHERE, we know the data is stored in

the customer\_purchase table in the customer\_dataset.

So we write customer\_data.customer\_purchase

after FROM.

Next, we tell SQL what data to give us in the select clause.

We want to see the purchase\_price data,

so we type purchase\_price after SELECT.

Next is the where clause.

We are not filtering out any data since we want all purchase

prices shown, so we can take out the where clause.

Finally, to sort the purchase\_price

in descending order, we type ORDER BY

purchase\_price DESC at the end of our query.

Let's run this query.

Play video starting at :2:28 and follow transcript2:28

We see that 89.85 shows up at the top

with 799.99 below it, but

we know that 799.99 is a bigger number than 89.85.

The database doesn't recognize that these are numbers, so

it didn't sort them that way.

If we go back to the customer\_purchase table and take a look at its schema,

we can see what data type the database thinks purchase\_price is.

It says here the database thinks purchase\_price is a string,

when in fact it is a float, which is a number that contains a decimal.

That is why 89.85 shows up before 799.99.

When we sort letters,

we start from the first letter before moving on to the second letter.

So if we want to sort the words apple and orange in descending order,

we start with the first letters a and o.

Since o comes after a, orange will show up first, then apple.

The database did the same with 89.85 and 799.99.

It started with the first letter, which in this case was 8 and 7 respectively.

Since 8 is bigger than 7, the database sorted 89.85 first and

then 799.99 because the database treated these as text strings.

The database doesn't recognize these strings as floats because they haven't

been typecast to match that data type yet.

Typecasting means converting data from one type to another,

which is what we'll do with the CAST function.

We use the CAST function to replace purchase\_price with a new

purchase\_price that the database recognizes as float instead of string.

We start by replacing purchase\_price with CAST.

Then we tell SQL the field we want to change, which is the purchase\_price field.

Next is a data type we want to change purchase\_price to,

which is the FLOAT data type.

BigQuery stores numbers in a 64 bit system,

so the FLOAT data type is referenced as float 64 in our query.

This might be slightly different in other SQL platforms, but basically the 64 and

float 64 just indicates that we're casting numbers in the 64 bit system as FLOATs.

We also need to sort this new field so we change purchase\_price

after ORDER BY to CAST purchase\_price as FLOAT64.

This is how we use the cast function to allow SQL to recognize

the purchase\_price column as FLOATs instead of text strings.

Now we can sort our purchases by purchase\_price.

And just like that,

Lauren's Furniture Store has data that can actually be used for analysis.

As a data analyst, you'll be asked to locate and organize data a lot,

which is why you want to make sure you convert between data types early on.

Businesses like our Furniture Store are interested in timely sales data,

and you need to be able to account for that in your analysis.

The CAST function can be used to change strings into other data types too,

like date and time.

As a data analyst, you might find yourself using data from various sources.

Part of your job is making sure the data from those sources is recognizable and

usable in your database so that you won't run into any issues with your analysis.

And now you know how to do that.

The CAST function is one great tool you can use when you're cleaning data.

And coming up,

we'll cover some other advanced functions that you can add to your toolbox.

See you soon.

0:00

Hey there. Great to see you again.

So far we've seen some SQL functions in action.

In this video, we'll go over more uses for CAST and

then learn about CONCAT and COALESCE. Let's get started.

Earlier we talked about the CAST function,

which lets us type cast text strings into floats.

I called out that the CAST function can

be used to change into other data types too.

Let's check out another example of how

you can use CAST in your own data work.

We've got the transaction data we were working

with from our Lauren's furniture store example,

but now we'll check out the purchase date field.

The furniture store owner has asked us to look at

purchases that occurred during

their sales promotion period in December.

Let's write a SQL query that will pull date and

purchase\_price for all purchases

that occurred between December 1st,

2020 and December 31st, 2020.

We start by writing the basic SQL structure;

SELECT, FROM, WHERE.

We know the data comes from

the customer\_purchase table in

the customer\_data data set,

so we write customer\_data.customer\_purchase after FROM.

Next we tell SQL what data to pull.

Since we want date and purchase\_price,

we add them into the SELECT statement.

Play video starting at :1:38 and follow transcript1:38

Finally, we want SQL to filter for

purchases that occurred in December only,

so we type date BETWEEN

2020-12-01

Play video starting at :1:52 and follow transcript1:52

and 2020-12-31 in the WHERE clause.

Let's run the query.

Four purchases occurred in December,

but the date field looks odd.

That's because the database recognizes

the date field as date time,

which consists of the date and time.

Our SQL query still works correctly even if

the date field is date time instead of date.

But we can tell SQL to convert the date field into

the date data type so

we see just the date and not the time.

To do that, we use the CAST function again.

We'll use the CAST function to replace the date field in

our select statement with

the new date field that will

show the date and not the time.

We can do that by typing CAST and adding

the date as the field we want to change,

then we tell SQL the data type we want instead,

which is the date data type.

Play video starting at :2:57 and follow transcript2:57

There. Now we can have

cleaner results for purchases that

occurred during the December sales period.

CAST is a super useful function

for cleaning and sorting data,

which is why I wanted you to see

it in action one more time.

Next up, let's check out the CONCAT function.

CONCAT lets you add strings together to create

new text strings that can be used as unique keys.

Going back to our customer\_purchase table,

we see that the furniture store

sells different colors of the same product.

The owner wants to know if customers prefer

certain colors so the owner can

manage store inventory accordingly.

The problem is the product\_code

is the same regardless of the product color.

We need to find another way to separate products by color

so we can tell if customers

prefer one color over the others.

We'll use CONCAT to produce a unique key

that'll help us tell the products

apart by color and count them more easily.

Let's write our SQL query by

starting with the basic structure.

SELECT, FROM, WHERE.

We know our data comes from

the customer\_purchase table and

the customer\_data data set,

so we type customer\_data.customer\_purchase after FROM.

Next we tell SQL what data to pull.

We use the CONCAT function here to

get that unique key of product and color.

We type CONCAT,

the first column we want,

product\_code, and the other column

we want, product\_color.

Play video starting at :4:49 and follow transcript4:49

Finally, let's say we want to look at couches,

so we filter for couches by

typing product='couch' in the WHERE clause.

Now we can count how many times each couch was purchased

and figure out if customers

preferred one color over the others.

Play video starting at :5:14 and follow transcript5:14

With CONCAT, the furniture store can find out

which color couches are the most popular and order more.

I've got one last advanced function

to show you, COALESCE.

COALESCE can be used to return non-null values in a list.

Null values are missing values.

If you have a field that's optional in your table,

it'll have null in that field for

rows that don't have appropriate values to put there.

Let's open the customer\_purchase table

so I can show you what I mean.

In the customer\_purchase table,

we can see a couple of rows where

product information is missing.

That is why we see nulls there.

But for the rows where product name is null,

we see that there is product\_code data

that we can use instead.

We'd prefer SQL to show us the product name,

like bed or couch because it's easier for us to read.

But if the product name doesn't exist,

we can tell SQL to give us the product\_code instead.

That is where the COALESCE function comes into play.

Let's say we wanted a list

of all products that were sold.

We want to use the product name column

to understand what product was sold.

So we write our SQL query with the basic SQL structure,

SELECT, FROM, WHERE.

We know our data comes from

customer\_purchase table and the customer\_data data set,

so we type customer\_data.customer\_purchase after FROM.

Next, we tell SQL the data we want.

We want a list of product names.

But if names aren't available,

then give us the product code.

Here is where we type COALESCE,

then we tell which column to check first, product,

and which column to check second,

if the first column is null, product\_code.

We'll name this new field as product\_info.

Play video starting at :7:20 and follow transcript7:20

Finally, we are not filtering out

any data so we can take out the WHERE clause.

This gives us product information for each purchase.

Now we have a list of all products

that were sold for the owner to review.

COALESCE can save you

time when you're making calculations too,

by skipping any null values

and keeping your math correct.

Those were just some of the

advanced functions you can use to clean

your data and get it ready for

the next step in the analysis process.

You'll discover more as you continue working in SQL.

But that's the end of this video and this module.

Great work. We've covered a lot of ground.

You learned the different data cleaning functions

and spreadsheets and SQL,

and the benefits of using

SQL to deal with large data sets.

We also added some SQL formulas

and functions to your toolkit,

and most importantly, we got to experience some of

the ways that SQL can help you

get data ready for your analysis.

After this, you'll get to spend

some time learning how to verify and report

your cleaning results so that your data is

squeaky clean and your stakeholders know it.

But before that, you've got

another weekly challenge to tackle. You've got this.

Some of these concepts might seem challenging at first,

but they'll become second nature

to you as you progress in your career.

It just takes time and practice.

Speaking of practice, feel free to go back to any of

these videos and re-watch or

even try some of these commands on your own.

Good luck and I'll see you again when you're ready.

**Your SQL experience (so far)**



So far, you have been introduced to many different tools available in SQL. As a brief review, you learned how to complete tasks such as:

* Getting data from a table using **SELECT** statements.
* De-duplicating data using commands like **DISTINCT** and **COUNT +** **WHERE**.
* Manipulating string data with **TRIM()** and **SUBSTR**.
* Creating/dropping tables with **CREATE TABLE** and **DROP TABLE**.
* Changing data types with **CAST**.

Some of these tasks are more challenging than others, and learning all the various SQL functions takes work. But, when you practice different functions, you can master the skills needed to make SQL work the way you need it to. Take a moment to think about the parts of SQL that you’ve found most challenging.

# Glossary terms from module 3

## ****Terms and definitions for Course 4, Module 3****

**CAST:** A SQL function that converts data from one datatype to another

**COALESCE:** A SQL function that returns non-null values in a list

**CONCAT:** A SQL function that adds strings together to create new text strings that can be used as unique keys

**DISTINCT:** A keyword that is added to a SQL SELECT statement to retrieve only non-duplicate entries

**Float:** A number that contains a decimal

**Substring:** A subset of a text string

**Typecasting:** Converting data from one type to another

### 1.

Question 1

After a company merger, a data analyst receives a dataset with millions of rows of data. They need to use this data to identify insights for an important project. What tool would be most efficient for the analyst to use?

1 point

Word processor

SQL

CSV

Spreadsheet

### 2.

Question 2

A junior data professional notices their Boolean column is incorrectly storing True/False values as strings. What SQL function can the analyst use to convert the data type from a string to Boolean?

1 point

**CAST**

**TRIM**

**LENGTH**

**SUBSTR**

### 3.

Question 3

Fill in the blank: A junior data analyst in the engineering department uses the SQL command \_\_\_\_\_ to add a new row of product specification data to their organization’s database.

1 point

**INSERT INTO**

**UPDATE**

**CREATE TABLE IF NOT EXISTS**

**DROP TABLE IF EXISTS**

### 4.

Question 4

You are working with a database table that has columns about ice cream, such as ice\_cream\_flavor. Which **SUBSTR** function and **AS** command will retrieve the first 4 characters of each flavor and store the result in a new column called flavor\_ID?

1 point

**SUBSTR(ice\_cream\_flavor, 1, 4) AS flavor\_ID**

**SUBSTR(1, 4) ice\_cream\_flavor, AS flavor\_ID**

**SUBSTR(ice\_cream\_flavor, 4) AS flavor\_ID**

**SUBSTR(ice\_cream\_flavor 1, 4, AS) flavor\_ID**

### 5.

Question 5

In SQL, what function can be used to remove leading spaces from a piece of data?

1 point

**TRIM**

**AVG**

**SUBSTR**

**FORMAT**

### 6.

Question 6

While working with a database table that contains the column invoice\_number, you notice that there are some duplicate entries. Which SQL clause would you use in a query to return the invoice\_number data without these duplicates?

1 point

**DROP invoice\_number**

**DUPLICATE invoice\_number**

**DELETE invoice\_number**

**DISTINCT invoice\_number**

### 7.

Question 7

Fill in the blank: Data teams use the SQL command \_\_\_\_\_ to tidy up a database that is currently cluttered with many irrelevant tables.

1 point

**INSERT INTO**

**CREATE TABLE IF NOT EXISTS**

**DROP TABLE IF EXISTS**

**UPDATE**

### 8.

Question 8

You are using a database table that includes the column credit\_card\_numbers, and you want to check for any fraudulent activity. Which SQL clause will help you identify any credit card numbers that are more than 16 characters long?

1 point

**WHERE(credit\_card\_numbers) < 16**

**LENGTH(credit\_card\_numbers) > 16**

**COUNT(credit\_card\_numbers) > 16**

**IDENTIFY(credit\_card\_numbers) < 16**

### 9.

Question 9

After joining multiple tables containing data about patient visits to a hospital, you find a significant number of null values in the patient\_intake column. What SQL function can you use to replace these null values with a value in a different column?

1 point

**CONCAT**

**TRIM**

**CAST**

**COALESCE**

Hi there, great to have you back.

You've been learning a lot about

the importance of clean data and

explored some tools and strategies to

help you throughout the cleaning process.

In these videos, we'll be

covering the next step in the process:

verifying and reporting on

the integrity of your clean data.

Verification is a process to confirm that

a data cleaning effort was well-

executed and the resulting data is accurate and reliable.

It involves rechecking your clean dataset,

doing some manual clean ups if needed,

and taking a moment to sit back and really

think about the original purpose of the project.

That way, you can be confident that the data you

collected is credible and appropriate for your purposes.

Making sure your data is properly

verified is so important because it

allows you to double-check that the work you did to

clean up your data was thorough and accurate.

For example, you might have referenced

an incorrect cellphone number

or accidentally keyed in a typo.

Verification lets you catch

mistakes before you begin analysis.

Without it, any insights you gain from

analysis can't be trusted for decision-making.

You might even risk misrepresenting populations or

damaging the outcome of a product that

you're actually trying to improve.

I remember working on a project where I

thought the data I had was sparkling

clean because I'd use all the right tools and processes,

but when I went through the steps

to verify the data's integrity,

I discovered a semicolon that I had forgotten to remove.

Sounds like a really tiny error, I know,

but if I hadn't caught

the semicolon during verification and removed it,

it would have led to some big changes in my results.

That, of course, could have led

to different business decisions.

There's an example of why verification is so crucial.

But that's not all.

The other big part of

the verification process is reporting on your efforts.

Open communication is a lifeline

for any data analytics project.

Reports are a super effective way

to show your team that you're

being 100 percent transparent about your data cleaning.

Reporting is also a great opportunity to

show stakeholders that you're accountable,

build trust with your team,

and make sure you're all on the same page

of important project details.

Coming up, you'll learn

different strategies for reporting,

like creating data- cleaning reports,

documenting your cleaning process,

and using something called the changelog.

A changelog is a file containing

a chronologically ordered list of

modifications made to a project.

It's usually organized by version and

includes the date followed by a list of added,

improved, and removed features.

Changelogs are very useful for keeping track of how

a dataset evolved over the course of a project.

They're also another great way to

communicate and report on data to others.

Along the way, you'll also see

some examples of how verification and

reporting can help you avoid repeating

mistakes and save you and your team time.

Ready to get started? Let's go!

In this video,

we'll discuss how to begin the process of verifying your data-cleaning efforts.

Play video starting at ::7 and follow transcript0:07

Verification is a critical part of any analysis project.

Without it you have no way of knowing that your insights can be relied on for

data-driven decision-making.

Think of verification as a stamp of approval.

Play video starting at ::20 and follow transcript0:20

To refresh your memory, verification is a process to confirm that a data-cleaning

effort was well-executed and the resulting data is accurate and reliable.

It also involves manually cleaning data to compare your expectations with what's

actually present.

The first step in the verification process is going back to your

original unclean data set and comparing it to what you have now.

Review the dirty data and try to identify any common problems.

For example, maybe you had a lot of nulls.

In that case, you check your clean data to ensure no nulls are present.

To do that, you could search through the data manually or

use tools like conditional formatting or filters.

Play video starting at :1:6 and follow transcript1:06

Or maybe there was a common misspelling like someone keying in the name of

a product incorrectly over and over again.

In that case, you'd run a FIND in your clean data to make sure no instances of

the misspelled word occur.

Play video starting at :1:21 and follow transcript1:21

Another key part of verification involves taking a big-picture view of your project.

This is an opportunity to confirm you're actually focusing on the business problem

that you need to solve and the overall project goals

and to make sure that your data is actually capable of solving that problem

and achieving those goals.

Play video starting at :1:41 and follow transcript1:41

It's important to take the time to reset and focus on the big picture

because projects can sometimes evolve or

transform over time without us even realizing it.

Maybe an e-commerce company decides to survey 1000 customers to

get information that would be used to improve a product.

But as responses begin coming in, the analysts notice a lot of comments about

how unhappy customers are with the e-commerce website platform altogether.

So the analysts start to focus on that.

While the customer buying experience is of course important for

any e-commerce business, it wasn't the original objective of the project.

The analysts in this case need to take a moment to pause, refocus,

and get back to solving the original problem.

Play video starting at :2:32 and follow transcript2:32

Taking a big picture view of your project involves doing three things.

First, consider the business problem you're trying to solve with the data.

Play video starting at :2:41 and follow transcript2:41

If you've lost sight of the problem,

you have no way of knowing what data belongs in your analysis.

Taking a problem-first approach to analytics is essential at all stages of

any project.

You need to be certain that your data will actually make it possible to solve your

business problem.

Second, you need to consider the goal of the project.

It's not enough just to know that your company wants to analyze customer feedback

about a product.

What you really need to know is that the goal of getting this feedback is to make

improvements to that product.

On top of that, you also need to know whether the data you've collected and

cleaned will actually help your company achieve that goal.

And third, you need to consider whether your data is capable of solving

the problem and meeting the project objectives.

That means thinking about where the data came from and

testing your data collection and cleaning processes.

Play video starting at :3:35 and follow transcript3:35

Sometimes data analysts can be too familiar with their own data,

which makes it easier to miss something or make assumptions.

Play video starting at :3:43 and follow transcript3:43

Asking a teammate to review your data from a fresh perspective and

getting feedback from others is very valuable in this stage.

Play video starting at :3:51 and follow transcript3:51

This is also the time to notice if anything sticks out to

you as suspicious or potentially problematic in your data.

Again, step back, take a big picture view, and

ask yourself, do the numbers make sense?

Play video starting at :4:7 and follow transcript4:07

Let's go back to our e-commerce company example.

Imagine an analyst is reviewing the cleaned up data

from the customer satisfaction survey.

The survey was originally sent to 1,000 customers, but what if

the analyst discovers that there is more than a thousand responses in the data?

This could mean that one customer figured out a way to take the survey more

than once.

Or it could also mean that something went wrong in the data cleaning process, and

a field was duplicated.

Either way, this is a signal that it's time to go back to the data-cleaning

process and correct the problem.

Play video starting at :4:42 and follow transcript4:42

Verifying your data ensures that the insights you gain from analysis can be

trusted.

It's an essential part of data-cleaning that helps companies avoid big mistakes.

This is another place where data analysts can save the day.

Play video starting at :4:55 and follow transcript4:55

Coming up,

we'll go through the next steps in the data-cleaning process. See you there.

# Step-by-Step: Verification of data cleaning

This reading outlines the steps the instructor performs in the following video, [Verification of data cleaning](https://www.coursera.org/learn/process-data/lecture/Hx69i/the-final-step-in-data-cleaning). The video demonstrates how to verify cleaned data in both spreadsheets and SQL.

Keep this step-by-step guide open as you watch the video. It can serve as a helpful reference if you need additional context or clarification while following the video steps. This is not a graded activity, but you can complete these steps to practice the skills demonstrated in the video.

### **What you’ll need**

If you’d like to follow along with the examples in this video, choose a spreadsheet tool. Google Sheets or Excel are recommended.

To access the spreadsheet the instructor uses in this video, click the link to the template to create a copy of the dataset. If you don’t have a Google account, download the data directly from the attachments below.

Link to dataset: [Jeff’s Party Planet - Data for Cleaning](NULL)

OR

[Jeff's Party Planet - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/ve64rJeHTTylDeO1R6V-YQ_3611b683e1844cb5a165b96ea68f71e1_Jeff-s-Party-Planet---Data-for-Cleaning.xlsx?Expires=1706832000&Signature=knEZjWLhsdvZiX9OeDxDg2V7mmWwyL7gYBiMm2PMa4I4DvjfEJa9A78CJxHCBF6AMpb4PWOBphBOmVtwqD9co9RAhpJIGR0Ktun-g~FdAfRGpXZQwEWDPSOoK74m~BuMWfTjFGukS2bTeFESSgkfFt1Qn1bwZDnR2ZBR8u5QN3E_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/ve64rJeHTTylDeO1R6V-YQ_3611b683e1844cb5a165b96ea68f71e1_Jeff-s-Party-Planet---Data-for-Cleaning.xlsx?Expires=1706832000&Signature=knEZjWLhsdvZiX9OeDxDg2V7mmWwyL7gYBiMm2PMa4I4DvjfEJa9A78CJxHCBF6AMpb4PWOBphBOmVtwqD9co9RAhpJIGR0Ktun-g~FdAfRGpXZQwEWDPSOoK74m~BuMWfTjFGukS2bTeFESSgkfFt1Qn1bwZDnR2ZBR8u5QN3E_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**Note:** The SQL table used in this example is not available for this activity.

## Example 1: Verify data with spreadsheets

Use spreadsheet tools such as Find and Replace and pivot tables to find, understand, and fix errors in your spreadsheet.

### **Use Find and Replace to replace all instances of a mistake**

1. Use the [Jeff’s Party Planet - Data for Cleaning](https://docs.google.com/spreadsheets/d/1RaDdSEp2V6D09FE6LOFkiJGv9CMT83GIV_U9YnY2rvI/template/preview?resourcekey=0-IU2-k90CX0mrt0ebwrvwDw#gid=0) dataset.
2. From the **Edit** menu, choose **Find and Replace** to open the **Find and replace** dialog box.
3. In the **Find** field, enter the misspelled word in the supplier name, **Plos**.
4. In the **Replace with** field, enter **Plus**.
5. Click **Replace all** to replace all instances of "Plos" with "Plus". Click **Done** to close the **Find and replace** dialog box.
6. Select the **Undo** button to use a different method to correct this misspelling. This can also be done with **Ctrl** (Windows) or **Command** (Mac) **Z**.

### **Use a pivot table to understand errors in a spreadsheet**

1. Select the **Suppliers** column.
2. Select **Insert > Pivot Table**. In the **Create pivot table** dialog box, choose **New Sheet** then **Create**.
3. This creates a new tab that is mostly blank.
4. Additionally, the **Pivot table editor** pane is in the window.
5. Next to **Rows**. Select **Add**, then the **Suppliers** column.
6. Next to **Values**, select **Add** then select **Suppliers**. This adds a value for the **Suppliers** column.
7. By default, Google Sheets sets the value to summarize by **COUNTA** (the total number of values in a range). This will show how many times each supplier name comes up. It’s a great way to check for misspellings and other anomalies. **Note:** Don’t use **COUNT**, because **COUNT** counts only numerical values.
8. When there is only one instance of the misspelled name, manually change it to the correct spelling.
9. To return to the original sheet, select the **Sheet1** tab.

## Example 2: Use a CASE statement to verify data in SQL

Use **CASE** statements to correct misspellings in SQL.

1. The SQL table used in this example is not available for download, but if you were performing a similar query you’d first make sure to load the data in BigQuery.

2. Start your SQL query with the basic structure:

**SELECT**

**FROM**

**WHERE**

3. In the **FROM** clause, specify the table you're pulling data from after **FROM**. For example, **project-id.customer\_data.customer\_name**

4. In the **SELEC**T clause, specify the columns you want to return. In this example, you want **customer\_id** and **first\_name**.

5. However, there is a misspelling in a customer’s first name.

i. To correct the misspelled name "Tnoy" to "Tony", use a **CASE** statement.

ii. Enter **CASE**. On the next line, enter **WHEN first\_name = 'Tnoy'THEN 'Tony'**. This tells SQL to replace any instances of **Tnoy** in the **first\_name** column with **Tony**.

iii. On the next line, add the **statement ELSE first\_name** to keep other names as they are.

iv. End the statement with **END AS cleaned\_name**.This creates a new column called **cleaned\_name** that will contain the data cleaned with the **CASE** statement.

6. Delete the **WHERE** clause because you don’t want to filter the query.

7. The final statement should be:

1

2

3

4

5

6

7

8

SELECT

    Customer\_id,

    CASE

    WHEN first\_name = 'Tnoy' THEN 'Tony'

    ELSE first\_name

    END AS cleaned\_name

FROM

   project-id.customer\_data.customer\_name

8. This SQL query will correct the misspelled name and leave other names unchanged in a new column called **cleaned\_name**. Note that this query corrects only the display of the name; it does not update the table’s data.

Hey there. In this video,

we'll continue building on the verification process.

As a quick reminder,

the goal is to ensure that

our data-cleaning work was done

properly and the results can be counted on.

You want your data to be verified so you know

it's 100 percent ready to go.

It's like car companies running tons of tests to

make sure a car is safe before it hits the road.

You learned that the first step in

verification is returning to

your original, unclean dataset

and comparing it to what you have now.

This is an opportunity to search for common problems.

After that, you clean up

the problems manually. For example,

by eliminating extra spaces

or removing an unwanted quotation mark.

But there's also some great tools for

fixing common errors automatically,

such as TRIM and remove duplicates.

Earlier, you learned that TRIM

is a function that removes leading,

trailing, and repeated spaces and data.

Remove duplicates is a tool that automatically searches

for and eliminates duplicate entries from a spreadsheet.

Now sometimes you had an error that

shows up repeatedly, and it can't be

resolved with a quick manual edit or

a tool that fixes the problem automatically.

In these cases, it's helpful to create a pivot table.

A pivot table is

a data summarization tool

that is used in data processing.

Pivot tables sort, reorganize, group,

count, total or average data stored in a database.

We'll practice that now using

the spreadsheet from a party supply store.

Let's say this company was

interested in learning which of

its four suppliers is most cost-effective.

An analyst pulled this data on

the products the business sells,

how many were purchased,

which supplier provides them,

the cost of the products, and the ultimate revenue.

The data has been cleaned.

But during verification, we noticed that one of

the suppliers' names was keyed in incorrectly.

Play video starting at :2:16 and follow transcript2:16

We could just correct the word as "plus,"

but this might not solve the problem

because we don't know if this was

a one-time occurrence or if

the problem's repeated throughout the spreadsheet.

There are two ways to answer that question.

The first is using Find and replace.

Find and replace is a tool that looks for

a specified search term in

a spreadsheet and allows

you to replace it with something else.

We'll choose Edit. Then Find and replace.

We're trying to find P-L-O-S,

the misspelling of "plus" in the supplier's name.

In some cases you might not want to replace the data.

You just want to find something. No problem.

Just type the search term,

leave the rest of the options as

default and click "Done."

But right now we do want to replace it with

P-L-U-S. We'll type that in here.

Then click "Replace all" and "Done."

Play video starting at :3:20 and follow transcript3:20

There we go. Our misspelling has been corrected.

That was of course the goal.

But for now let's undo our Find and

replace so we can

practice another way to determine if

errors are repeated throughout a dataset,

like with the pivot table.

We'll begin by selecting the data we want to use.

Choose column C. Select "Data." Then "Pivot Table."

Choose "New Sheet" and "Create."

Play video starting at :3:59 and follow transcript3:59

We know this company has four suppliers.

If we count the suppliers and

the number doesn't equal four,

we know there's a problem.

First, add a row for suppliers.

Play video starting at :4:13 and follow transcript4:13

Next, we'll add a value for

our suppliers and summarize by COUNTA.

COUNTA counts the total number of values within a specified range.

Here we're counting the number of times

a supplier's name appears in

column C. Note that there's also function called COUNT,

which only counts the numerical values

within a specified range.

If we use it here,

the result would be zero.

Not what we have in mind.

But in other special applications,

COUNT would give us information

we want for our current example.

As you continue learning more

about formulas and functions,

you'll discover more interesting options.

If you want to keep learning,

search online for spreadsheet formulas and functions.

There's a lot of great information out there.

Our pivot table has counted the number of misspellings,

and it clearly shows that the error occurs just once.

Otherwise our four suppliers

are accurately accounted for in our data.

Now we can correct the spelling, and we

verify that the rest of the supplier data is clean.

This is also useful practice when querying a database.

If you're working in SQL,

you can address misspellings using a CASE statement.

The CASE statement goes through

one or more conditions and

returns a value as soon as a condition is met.

Let's discuss how this works in real life

using our customer\_name table.

Check out how our customer,

Tony Magnolia, shows up as Tony and Tnoy.

Tony's name was misspelled.

Let's say we want a list of our customer IDs and

the customer's first names so we can write

personalized notes thanking

each customer for their purchase.

We don't want Tony's note to be

addressed incorrectly to "Tnoy."

Here's where we can use: the CASE statement.

We'll start our query with the basic SQL structure.

SELECT, FROM, and WHERE.

We know that data comes from

the customer\_name table

in the customer\_data dataset,

so we can add customer underscore data

dot customer underscore name after FROM.

Next, we tell SQL what data to pull in the SELECT clause.

We want customer\_id and first\_name.

We can go ahead and add customer

underscore ID after SELECT.

But for our customer's first names,

we know that Tony was misspelled,

so we'll correct that using CASE. We'll

add CASE and then

WHEN and type first underscore name equal "Tnoy."

Next we'll use the THEN command and type "Tony,"

followed by the ELSE command.

Here we will type first underscore name,

followed by End As

and then we'll type cleaned underscore name.

Finally, we're not filtering our data,

so we can eliminate the WHERE clause.

As I mentioned, a CASE statement

can cover multiple cases.

If we wanted to search for a few more misspelled names,

our statement would look similar to the original,

with some additional names like this.

Play video starting at :8:6 and follow transcript8:06

There you go. Now that you've learned how you can use

spreadsheets and SQL to fix errors automatically,

we'll explore how to keep track of our changes next.

**Data-cleaning verification checklist**

This reading will give you a checklist of common problems you can refer to when doing your data cleaning verification, no matter what tool you are using. When it comes to data cleaning verification, there is no one-size-fits-all approach or a single checklist that can be universally applied to all projects. Each project has its own organization and data requirements that lead to a unique list of things to run through for verification.



Keep in mind, as you receive more data or a better understanding of the project goal(s), you might want to revisit some or all of these steps.

**Correct the most common problems**

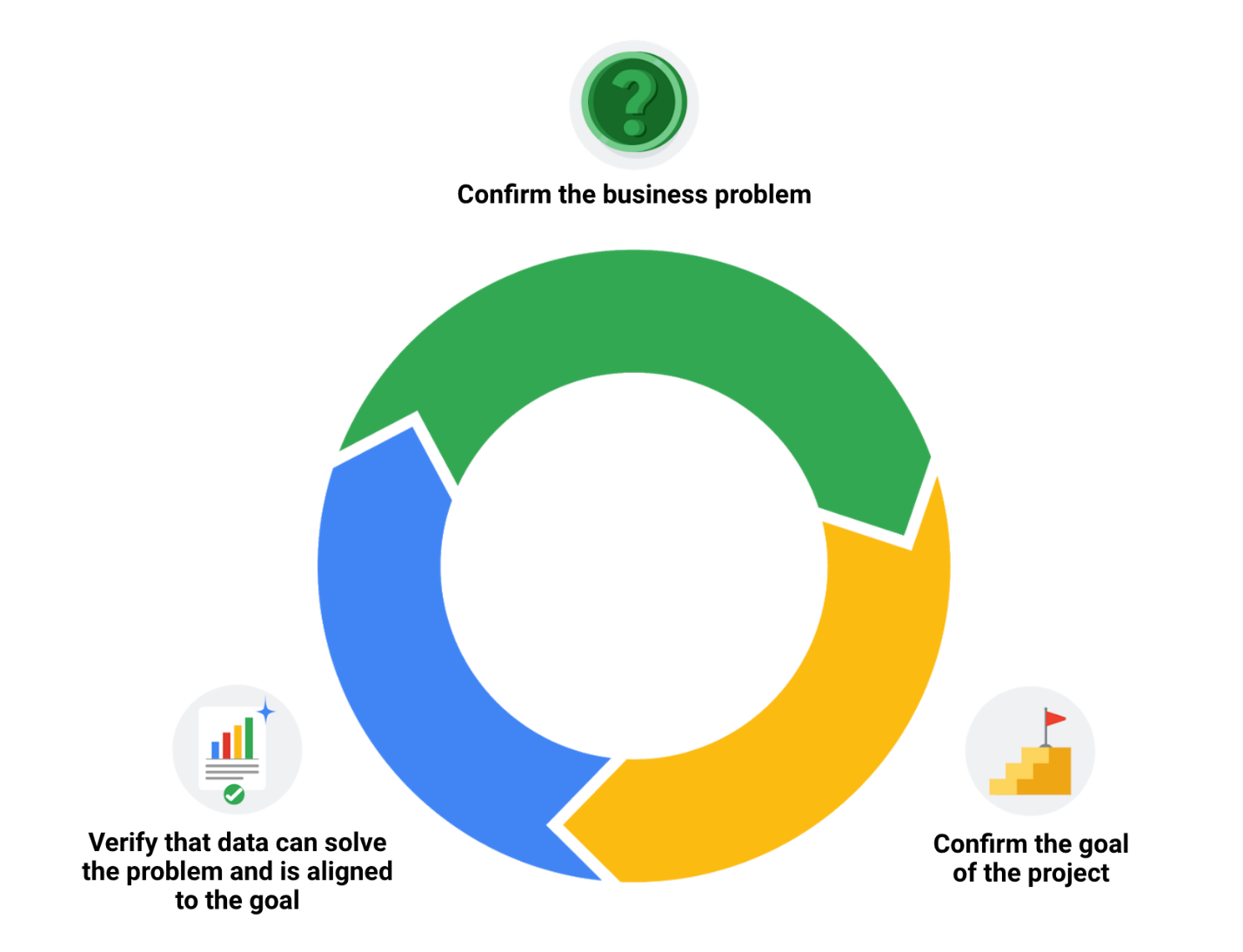
Make sure you identified the most common problems and corrected them, including:

* **Sources of errors**: Did you use the right tools and functions to find the source of the errors in your dataset?
* **Null data**: Did you search for NULLs using conditional formatting and filters?
* **Misspelled words**: Did you locate all misspellings?
* **Mistyped numbers**: Did you double-check that your numeric data has been entered correctly?
* **Extra spaces and characters**: Did you remove any extra spaces or characters using the **TRIM** function?
* **Duplicates**: Did you remove duplicates in spreadsheets using the **Remove Duplicates** function or **DISTINCT** in SQL?
* **Mismatched data types**: Did you check that numeric, date, and string data are typecast correctly?
* **Messy (inconsistent) strings**: Did you make sure that all of your strings are consistent and meaningful?
* **Messy (inconsistent) date formats**: Did you format the dates consistently throughout your dataset?
* **Misleading variable labels (columns)**: Did you name your columns meaningfully?
* **Truncated data:** Did you check for truncated or missing data that needs correction?
* **Business Logic**: Did you check that the data makes sense given your knowledge of the business?

**Review the goal of your project**

Once you have finished these data cleaning tasks, it is a good idea to review the goal of your project and confirm that your data is still aligned with that goal. This is a continuous process that you will do throughout your project-- but here are three steps you can keep in mind while thinking about this:

* Confirm the business problem
* Confirm the goal of the project
* Verify that data can solve the problem and is aligned to the goal



### 1.

Question 1

Which of the following tasks are involved in the verification process? Select all that apply.

1 point

Rechecking the data-cleaning effort

Asking stakeholders to check and confirm the data is clean

Considering whether the data is credible and appropriate for the project

Manually fixing errors found in the data

### 2.

Question 2

Which function enables a data professional to count the total number of spreadsheet values within a specified range?

1 point

**COUNTA**

**TOTAL**

**SUM**

**WHOLE**

### 3.

Question 3

Fill in the blank: Changelogs are files containing \_\_\_\_\_ ordered lists of modifications made to a project.

1 point

chronologically

alphabetically

randomly

directionally

### 4.

Question 4

A data professional discovers that SUV is spelled SWV in the database column car\_types. Which **CASE** clause will enable them to correct the misspellings?

1 point

**WHEN CASE car\_types = 'SWV' THEN 'SUV'**

**CASE WHEN car\_types = 'SWV' THEN 'SUV'**

**WHEN car\_types CASE = 'SWV' THEN 'SUV'**

**CASE car\_types = 'SWV' THEN 'SUV'**

Hi again. Now that you've

learned how to make your data squeaky clean,

it's time to address all the dirt you've left behind.

When you clean your data,

all the incorrect or outdated information is

gone, leaving you with the highest-quality content.

But all those changes you made to

the data are valuable too.

In this video, we'll discuss

why keeping track of changes is

important to every data project and how to

document all your cleaning changes to

make sure everyone stays informed.

This involves documentation which

is the process of tracking changes,

additions, deletions and errors

involved in your data cleaning effort.

You can think of it like a crime TV show.

Crime evidence is found at

the scene and passed on to the forensics team.

They analyze every inch of

the scene and document every step,

so they can tell a story with the evidence.

A lot of times,

the forensic scientist is

called to court to testify about

that evidence, and they have

a detailed report to refer to.

The same thing applies to data cleaning.

Data errors are the crime,

data cleaning is gathering evidence,

and documentation is detailing

exactly what happened for peer review or court.

Having a record of how a data set evolved

does three very important things.

First, it lets us recover data-cleaning errors.

Instead of scratching our heads,

trying to remember what we

might have done three months ago,

we have a cheat sheet to rely on if we

come across the same errors again later.

It's also a good idea to create

a clean table rather

than overriding your existing table.

This way, you still have the original data

in case you need to redo the cleaning.

Second, documentation gives you a way to

inform other users of changes you've made.

If you ever go on vacation or get promoted,

the analyst who takes over for you

will have a reference sheet to check in with.

Third, documentation helps you to determine the

quality of the data to be used in analysis.

The first two benefits assume the errors aren't fixable.

But if they are, a record gives

the data engineer more information to refer to.

It's also a great warning for ourselves that the data set

is full of errors and should be avoided in the future.

If the errors were time-consuming to fix,

it might be better to check out

alternative data sets that we can use instead.

Data analysts usually use

a changelog to access this information.

As a reminder, a changelog is a file containing

a chronologically ordered list of

modifications made to a project.

You can use and view a changelog in

spreadsheets and SQL to achieve similar results.

Let's start with the spreadsheet.

We can use Sheet's version history,

which provides a real-time tracker of

all the changes and who made them

from individual cells to the entire worksheet.

To find this feature,

click the File tab,

and then select Version history.

Play video starting at :3:19 and follow transcript3:19

In the right panel,

choose an earlier version.

Play video starting at :3:26 and follow transcript3:26

We can find who edited the file and

the changes they made in the column next to their name.

Play video starting at :3:34 and follow transcript3:34

To return to the current version,

go to the top left and click "Back."

If you want to check out changes in a specific cell,

we can right-click and select Show Edit History.

Play video starting at :3:58 and follow transcript3:58

Also, if you want others to be

able to browse a sheet's version history,

you'll need to assign permission.

Play video starting at :4:9 and follow transcript4:09

Now let's switch gears and talk about SQL.

The way you create and view a changelog with

SQL depends on the software program you're using.

Some companies even have their own separate software

that keeps track of changelogs and important SQL queries.

This gets pretty advanced.

Essentially, all you have to do is

specify exactly what you did and

why when you commit a query to

the repository as a new and improved query.

This allows the company to revert

back to a previous version

if something you've done crashes the system,

which has happened to me before.

Another option is to just add

comments as you go while you're cleaning data in SQL.

This will help you construct

your changelog after the fact.

For now, we'll check out

query history, which tracks all the queries you've run.

Play video starting at :5:7 and follow transcript5:07

You can click on any of them to revert back to

a previous version of your query or to

bring up an older version to find what you've

changed. Here's what we've got.

I'm in the Query history tab.

Listed on the bottom right are

all the queries that run by date and time.

You can click on this icon to the right of

each individual query to bring it up to the Query editor.

Changelogs like these are

a great way to keep yourself on track.

It also lets your team get

real-time updates when they want them.

But there's another way to keep the communication

flowing, and that's reporting.

Stick around, and you'll learn some easy ways to share

your documentation and maybe

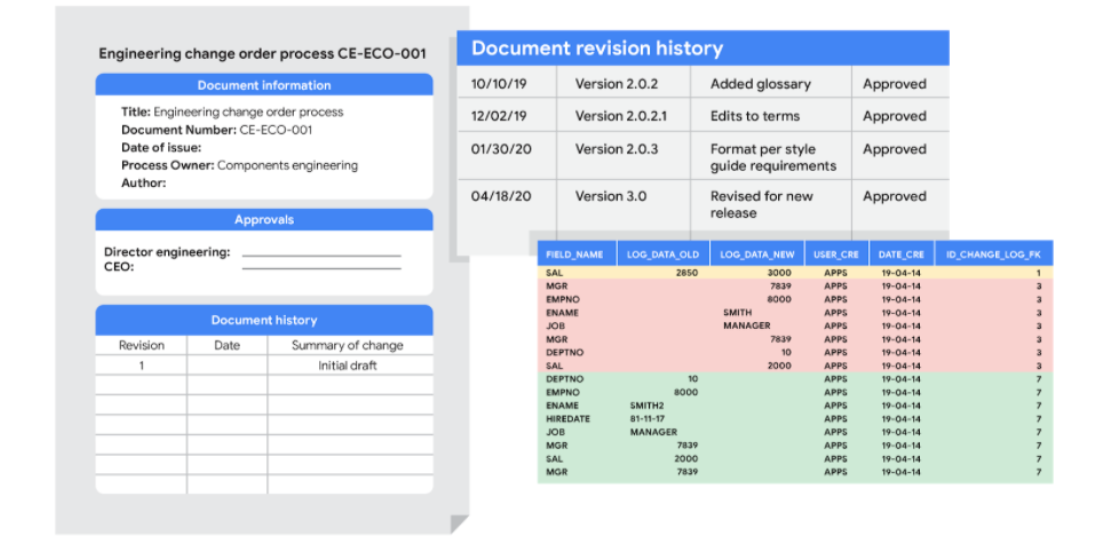
impress your stakeholders in the process.

See you in the next video.

**Embrace changelogs**

What do engineers, writers, and data analysts have in common? Change.

Engineers use **engineering change orders** (ECOs) to keep track of new product design details and proposed changes to existing products. Writers use **document revision histories** to keep track of changes to document flow and edits. And data analysts use **changelogs** to keep track of data transformation and cleaning. Here are some examples of these:



**Automated version control takes you most of the way**

Most software applications have a kind of history tracking built in. For example, in Google sheets, you can check the version history of an entire sheet or an individual cell and go back to an earlier version. In Microsoft Excel, you can use a feature called **Track Changes**. And in BigQuery, you can view the history to check what has changed.

Here’s how it works:

|  |  |
| --- | --- |
| Google Sheets | 1. Right-click the cell and select **Show edit history**. 2. Click the left-arrow < or right arrow > to move backward and forward in the history as needed. |
| Microsoft Excel | 1. If Track Changes has been enabled for the spreadsheet: click **Review**.2. Under **Track Changes**, click the **Accept/Reject Changes** option to accept or reject any change made. |
| BigQuery | Bring up a previous version (without reverting to it) and figure out what changed by comparing it to the current version. |

**Changelogs take you down the last mile**

A **changelog** can build on your automated version history by giving you an even more detailed record of your work. This is where data analysts record all the changes they make to the data. Here is another way of looking at it. Version histories record *what* was done in a data change for a project, but don't tell us *why*. Changelogs are super useful for helping us understand the reasons changes have been made. Changelogs have no set format and you can even make your entries in a blank document. But if you are using a shared changelog, it is best to agree with other data analysts on the format of all your log entries.

Typically, a changelog records:

* Data, file, formula, query, or any other component that changed
* Description of what changed
* Date of the change
* Person who made the change
* Person who approved the change
* Version number
* Reason for the change

Let’s say you made a change to a formula in a spreadsheet because you observed it in another report and you wanted your data to match and be consistent. If you found out later that the report was actually using the wrong formula, an automated version history would help you *undo* the change. But if you also recorded the reason for the change in a changelog, you could go back to the creators of the report and let them know about the incorrect formula. If the change happened a while ago, you might not remember who to follow up with. Fortunately, your changelog would have that information ready for you! By following up, you would ensure data integrity outside your project. You would also be showing personal integrity as someone who can be trusted with data. That is the power of a changelog!

Finally, a changelog is important for when lots of changes to a spreadsheet or query have been made. Imagine an analyst made four changes and the change they want to revert to is change #2. Instead of clicking the undo feature three times to undo change #2 (and losing changes #3 and #4), the analyst can undo just change #2 and keep all the other changes. Now, our example was for just 4 changes, but try to think about how important that changelog would be if there were hundreds of changes to keep track of.

**Bonus tip**



If an analyst is making changes to an existing SQL query that is shared across the company, the company most likely uses what is called a **version control system**. An example might be a query that pulls daily revenue to build a dashboard for senior management.

Here's how a version control system affects a change to a query:

1. A company has official versions of important queries in their **version control system**.
2. An analyst makes sure the most up-to-date version of the query is the one they will change. This is called **syncing**
3. The analyst makes a change to the query.
4. The analyst might ask someone to review this change. This is called a **code review** and can be informally or formally done. An informal review could be as simple as asking a senior analyst to take a look at the change.
5. After a reviewer approves the change, the analyst submits the updated version of the query to a repository in the company's version control system. This is called a **code commit**. A best practice is to document exactly what the change was and why it was made in a comments area. Going back to our example of a query that pulls daily revenue, a comment might be: *Updated revenue to include revenue coming from the new product, Calypso*.
6. After the change is **submitted**, everyone else in the company will be able to access and use this new query when they **sync** to the most up-to-date queries stored in the version control system.
7. If the query has a problem or business needs change, the analyst can ***undo*** the change to the query using the version control system. The analyst can look at a chronological list of all changes made to the query and who made each change. Then, after finding their own change, the analyst can **revert** to the previous version.
8. The query is back to what it was before the analyst made the change. And everyone at the company sees this reverted, original query, too.

**Key takeaways**

Engineers, writers, and data analysts use different methods to keep track of changes they make to their work. Automated version control, changelogs, and version control systems are all common tools used to track changes. Changelogs are particularly useful, as they can be used to record the reasons for changes made to data. This can help to ensure data integrity and consistency. Version control systems are most commonly used when making changes to shared queries. They enable analysts to track any changes made and revert to previous versions if necessary.

Great, you're back.

Let's set the stage.

The crime is dirty data.

We've gathered the evidence.

It's been cleaned, verified, and cleaned again.

Now it's time to present our evidence.

We'll retrace the steps and

present our case to our peers.

As we discussed earlier, data cleaning,

verifying, and reporting is a lot like crime drama.

Now it's our day in court.

Just like a forensic scientist

testifies on the stand about the evidence,

data analysts are counted on to present

their findings after a data cleaning effort.

Earlier, we learned how to document and

track every step of the data cleaning process,

which means we have solid information to pull from.

As a quick refresher,

documentation is the process

of tracking changes, additions,

deletions, and errors involved in a data cleaning effort,

changelogs are good example of this.

Since it's staged chronologically,

it provides a real-time account of every modification.

Documenting will be a huge time saver

for you as a future data analyst.

It's basically a cheatsheet you

can refer to if you're working with

the similar data set or need to address similar errors.

While your team can view changelogs directly,

stakeholders can't and have to

rely on your report to know what you did.

Lets check out how we might document

our data cleaning process

using example we worked with earlier.

In that example, we found that this association had

two instances of the same membership

for $500 in its database.

Play video starting at :1:48 and follow transcript1:48

We decided to fix this

manually by deleting the duplicate info.

Play video starting at :1:57 and follow transcript1:57

There're plenty of ways we could go

about documenting what we did.

One common way is to just create a doc

listing out the steps we took and the impact they had.

For example, first on

your list would be that you

remove the duplicate instance,

Play video starting at :2:17 and follow transcript2:17

which decreased the number of rows from 33 to 32,

Play video starting at :2:29 and follow transcript2:29

and lowered the membership total by $500.

Play video starting at :2:39 and follow transcript2:39

If we were working with SQL,

we could include a comment in

the statement describing the reason for

a change without affecting

the execution of the statement.

That's something a bit more advanced,

which we'll talk about later.

Regardless of how we capture and share our changelogs,

we're setting ourselves up for success by

being 100 percent transparent about our data cleaning.

This keeps everyone on the same page and shows

project stakeholders that we are

accountable for effective processes.

In other words, this helps

build our credibility as witnesses who can

be trusted to present

all the evidence accurately during testimony.

For dirty data, it's an open and shut case.

Welcome back.

By now it's safe to say that verifying, documenting and

reporting are valuable steps in the data-cleaning process.

You have proof to give stakeholders that your data is accurate and reliable.

And the effort to attain it was well-executed and documented.

The next step is getting feedback about the evidence and using it for good,

which we'll cover in this video.

Play video starting at ::25 and follow transcript0:25

Clean data is important to the task at hand.

But the data-cleaning process itself can reveal insights that are helpful to

a business.

The feedback we get when we report on our cleaning can transform data collection

processes, and ultimately business development.

For example,

one of the biggest challenges of working with data is dealing with errors.

Some of the most common errors involve human mistakes like mistyping or

misspelling,

flawed processes like poor design of a survey form,

and system issues where older systems integrate data incorrectly.

Whatever the reason, data-cleaning can shine a light on the nature and

severity of error-generating processes.

Play video starting at :1:7 and follow transcript1:07

With consistent documentation and reporting,

we can uncover error patterns in data collection and entry procedures

and use the feedback we get to make sure common errors aren't repeated.

Maybe we need to reprogram the way the data is collected

or change specific questions on the survey form.

Play video starting at :1:26 and follow transcript1:26

In more extreme cases,

the feedback we get can even send us back to the drawing board

to rethink expectations and possibly update quality control procedures.

For example, sometimes it's useful to schedule a meeting with a data engineer

or data owner to make sure the data is brought in properly

and doesn't require constant cleaning.

Play video starting at :1:48 and follow transcript1:48

Once errors have been identified and addressed,

stakeholders have data they can trust for decision-making.

And by reducing errors and inefficiencies in data collection,

the company just might discover big increases to its bottom line.

Congratulations!

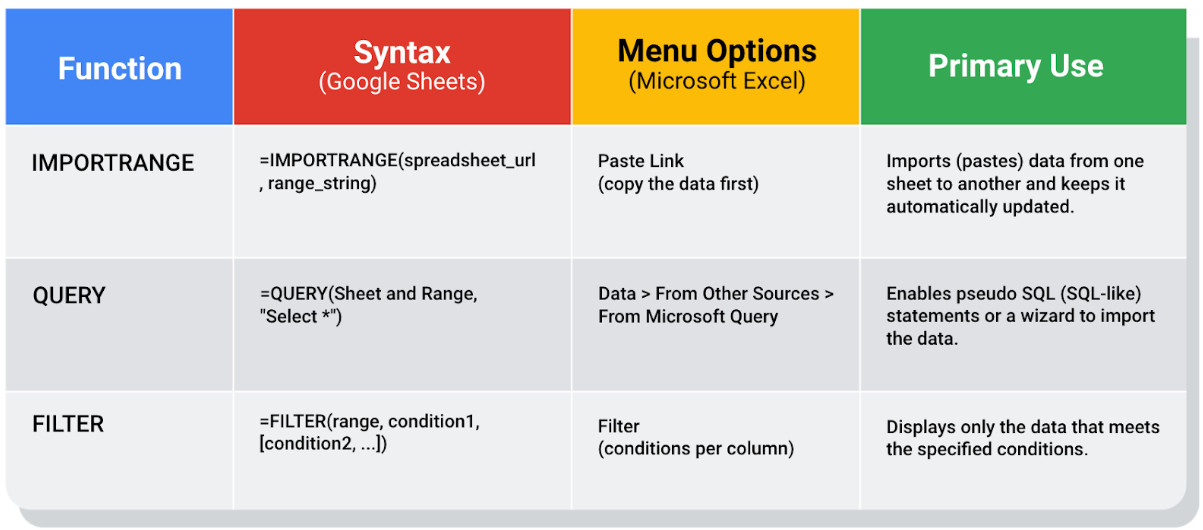
You now have the foundation you need to successfully verify a report on your

cleaning results.

Stay tuned to keep building on your new skills.

# Advanced functions for speedy data cleaning

In this reading, you will learn about some advanced functions that can help you speed up the data cleaning process in spreadsheets. Below is a table summarizing three functions and what they do:

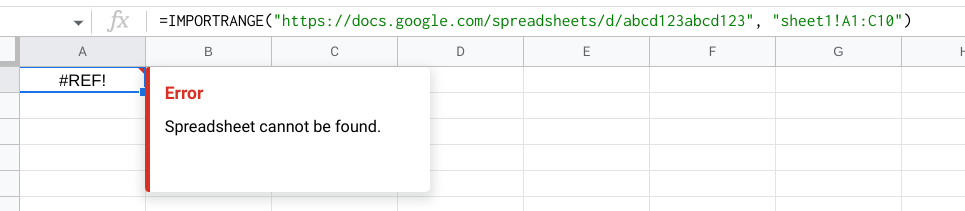


## Keeping data clean and in sync with a source

The [**IMPORTRANGE**](https://support.google.com/docs/answer/3093340?hl=en) function in Google Sheets and the[**Paste Link**](https://professor-excel.com/how-to-paste-cell-links/) feature (a Paste Special option in Microsoft Excel) both allow you to insert data from one sheet to another. Using these on a large amount of data is more efficient than manual copying and pasting. They also reduce the chance of errors being introduced by copying and pasting the wrong data. They are also helpful for data cleaning because you can “cherry pick” the data you want to analyze and leave behind the data that isn’t relevant to your project. Basically, it is like canceling noise from your data so you can focus on what is most important to solve your problem. This functionality is also useful for day-to-day data monitoring; with it, you can build a tracking spreadsheet to share the relevant data with others. The data is synced with the data source so when the data is updated in the source file, the tracked data is also refreshed.

In Google Sheets, you can use the **IMPORTRANGE** function. It enables you to specify a range of cells in the other spreadsheet to duplicate in the spreadsheet you are working in. You must allow access to the spreadsheet containing the data the first time you import the data.

**The URL shown below is for syntax purposes only. Don't enter it in your own spreadsheet. Replace it with a URL to a spreadsheet you have created so you can control access to it by clicking the Allow access button.**



Refer to the [Google support page for IMPORTRANGE](https://support.google.com/docs/answer/3093340?hl=en) for the sample usage and syntax.

### **Example of using** IMPORTRANGE

An analyst monitoring a fundraiser needs to track and ensure that matching funds are distributed. They use **IMPORTRANGE** to pull all the matching transactions into a spreadsheet containing all of the individual donations. This enables them to determine which donations eligible for matching funds still need to be processed. Because the total number of matching transactions increases daily, they simply need to change the range used by the function to import the most up-to-date data.

On Tuesday, they use the following to import the donor names and matched amounts:

**=IMPORTRANGE("https://docs.google.com/spreadsheets/d/abcd123abcd123", "sheet1!A1:C10", "Matched Funds!A1:B4001")**

On Wednesday, another 500 transactions were processed. They increase the range used by 500 to easily include the latest transactions when importing the data to the individual donor spreadsheet:

**=IMPORTRANGE("https://docs.google.com/spreadsheets/d/abcd123abcd123", "Matched Funds!A1:B4501")**

**Note: The above examples are for illustrative purposes only. Don't copy and paste them into your spreadsheet. To try it out yourself, you will need to substitute your own URL (and sheet name if you have multiple tabs) along with the range of cells in the spreadsheet that you have populated with data.**

## Pulling data from other data sources

The [**QUERY**](https://support.google.com/docs/answer/3093343?hl=en) function is also useful when you want to pull data from another spreadsheet. The **QUERY** function's SQL-like ability can extract specific data within a spreadsheet. For a large amount of data, using the **QUERY** function is faster than filtering data manually. This is especially true when repeated filtering is required. For example, you could generate a list of all customers who bought your company’s products in a particular month using manual filtering. But if you also want to figure out customer growth month over month, you have to copy the filtered data to a new spreadsheet, filter the data for sales during the following month, and then copy those results for the analysis. With the **QUERY** function, you can get all the data for both months without a need to change your original dataset or copy results.

The **QUERY** function syntax is similar to **IMPORTRANGE**. You enter the sheet by name and the range of data that you want to query from, and then use the SQL **SELECT** command to select the specific columns. You can also add specific criteria after the **SELECT** statement by including a **WHERE** statement. But remember, all of the SQL code you use has to be placed between the quotes!

Google Sheets run the Google Visualization API Query Language across the data. Excel spreadsheets use a query wizard to guide you through the steps to connect to a data source and select the tables. In either case, you are able to be sure that the data imported is verified and clean based on the criteria in the query.

### **Examples of using QUERY**

Check out the [Google support page for the QUERY function](https://support.google.com/docs/answer/3093343?hl=en) with sample usage, syntax, and examples you can download in a Google sheet.

Link to make a copy of the sheet: [QUERY examples](https://docs.google.com/spreadsheets/d/1815H5TCe91LLT6tD6FmxMHmeJAAkr4o5Q6rNpV6xiFk/copy)

### **The solution**

Analysts can use SQL to pull a specific dataset into a spreadsheet. They can then use the **QUERY** function to create multiple tabs (views) of that dataset. For example, one tab could contain all the sales data for a particular month and another tab could contain all the sales data from a specific region. This solution illustrates how SQL and spreadsheets are used well together.

## Filtering data to get what you want

The [**FILTER**](https://support.google.com/docs/answer/3093197?hl=en) function is fully internal to a spreadsheet and doesn’t require the use of a query language. The **FILTER** function lets you view only the rows (or columns) in the source data that meet your specified conditions. It makes it possible to pre-filter data before you analyze it.

The **FILTER** function might run faster than the **QUERY** function. But keep in mind, the **QUERY** function can be combined with other functions for more complex calculations. For example, the **QUERY** function can be used with other functions like **SUM** and **COUNT** to summarize data, but the **FILTER** function can't.

### **Example of using FILTER**

Check out the [Google support page for the FILTER function](https://support.google.com/docs/answer/3093197?hl=en) with sample usage, syntax, and examples you can download in a Google sheet.

Link to make a copy of the sheet: [FILTER examples](https://docs.google.com/spreadsheets/d/1caULJLQvQuzBnCN7rO9utg0xSKrYms7wM0Ph7A2JXY4/copy)

### 1.

Question 1

What objectives can be achieved by documenting the evolution of a dataset? Select all that apply.

1 point

Inform other users of changes

Determine the quality of the data

Recover data-cleaning errors

Communicate data insights to stakeholders

### 2.

Question 2

Fill in the blank: After a change to a query is submitted, all team members will be able to access the new query once they \_\_\_\_\_ the most up-to-date version control system.

1 point

copy

sync to

revert from

upload

### 3.

Question 3

What information is typically included in a changelog? Select all that apply.

1 point

Frequently asked questions about the change

The component that changed and the reason why

The date of the change

A description of the change

### 4.

Question 4

A data professional makes a change to a file. Then, they ask a colleague to evaluate the change to identify any potential issues. What does this scenario describe?

1 point

Code commit

Sync

Revert

Code review

# Glossary terms from module 4

## ****Terms and definitions for Course 4, Module 4****

**CASE:** A SQL statement that returns records that meet conditions by including an if/then statement in a query

**Changelog:** A file containing a chronologically ordered list of modifications made to a project

**COUNTA:** A spreadsheet function that counts the total number of values within a specified range

**Find and replace:** A tool that finds a specified search term and replaces it with something else

**Verification:** A process to confirm that a data-cleaning effort was well executed and the resulting data is accurate and reliable

### 1.

Question 1

Fill in the blank: A data scientist keeps code for data analysis pipelines in a \_\_\_\_\_, which enables them to track the evolution of the pipelines over time.

1 point

dataset

changelog

version control system

dashboard

### 2.

Question 2

A data analyst performs a customer segmentation project. During the verification process, they focus on the big picture view of attracting the company’s ideal customers and meeting customer needs. What actions should they take to achieve this goal? Select all that apply.

1 point

Consider the business problem

Consider the goal

Consider the reporting

Consider the data

### 3.

Question 3

Which SQL clause will consider the condition ‘Florida’ and return the value ‘East’ when that condition is met?

1 point

|  |
| --- |
| **CASE**  **WHEN us\_states = 'Florida' THEN 'East' END** |
| **CASE us\_states = 'Florida' THEN 'East' END** |

|  |
| --- |
| **WHEN**  **CASE us\_states = 'Florida' THEN 'East' END** |
| **WHEN us\_states = 'Florida' CASE 'East' END** |

### 4.

Question 4

What is the process of tracking changes, additions, deletions, and errors during data cleaning?

1 point

Documentation

Observation

Cataloging

Recording

### 5.

Question 5

During verification, you wonder if one of your data modifications was an effective update. What can you reference to revisit the modification and your reasoning behind it?

1 point

Data table

Text editor

Notepad

Changelog

### 6.

Question 6

A team is tasked with determining how many customers have made a purchase in the past month. Using a pivot table in Google Sheets, what function will total the number of non-empty sales in the purchase\_date column of its customer database?

1 point

**CONCAT**

**CHECK**

**CASE**

**COUNTA**

### 7.

Question 7

Which of the following statements accurately describe code review and code commit? Select all that apply.

1 point

Code review must involve numerous formal approvals.

Code review occurs prior to code commit.

Code commit might involve submitting changes to a data pipeline in a version control system.

An example of code review is a web developer submitting their code for evaluation by a more experienced developer.

### 8.

Question 8

Fill in the blank: To correct a misspelling in their spreadsheet, a data professional uses \_\_\_\_\_ to search for any instance of “compurer” and change it to “computer.”

1 point

find and replace

Remove duplicates

formatting

**TRIM**

Module 5

Hey there, thanks for stopping by once again.

So earlier we checked out some potential career paths that might open up for

you once you complete the program.

You might also have explored the advantages of networking and

building an online presence.

And I want to tell you just by being here now, you've shown you're committed.

You're taking a big step in your future career.

Coming up, we'll spend some time building your resume.

You might already have a resume that you've used or been saving and

that's great.

There's a good chance you'll still be able to use it even if you're planning to

switch careers.

Together we'll find out what kinds of changes to your resume you might want

to make.

But before that, we'll figure out what the whole application process is like.

Then we'll explore the best way to write or adjust your resume to make it

as professional-looking as possible and ready for your role as a data analyst.

We'll also take a peek at some examples of other resumes.

After that, we'll have you do a little self-analysis as we review the different

types of data analyst jobs out there,

so you can think about which ones might be best for you.

While I'm definitely not a career counselor, we can still think of this as

a kind of career counseling session. You'll get a better idea of how to build

your resume while thinking about your bigger career picture at the same time.

So let's get started!

Hi again.

Right now, it seems like the perfect time to take a step back from learning about

data analytics,

so you can get excited about what comes after you're done here.

The road to finding a job can be challenging,

but you're building up your skillset and

learning what it takes to be a data analyst.

In this video, we'll cover what you can expect from your job search,

plus some tips for using your newfound skills and

knowledge to make your search easier.

I remember when I first started out.

I reached out to as many people as I could to learn about their career paths,

their companies, and their roles.

I wanted to get a good idea of what to expect.

And that's what we're doing now:

giving you an idea of what to expect during your own job search.

It's important to remember that everyone's search will be different.

It might depend on where you live, what your interests are within the field, and

personal preferences,

like the type of work environment you feel comfortable in.

This is all part of making this journey your own as you hunt for

a job that's perfect for you.

The most common way to start is by checking out available jobs.

There's a lot of job sites that are built specifically for

people seeking employment.

You can also go to company websites

where they usually post job listings too.

Play video starting at :1:13 and follow transcript1:13

These sites might even have an option to send you an alert when a role matching

your search becomes available.

Once you find a few that you like,

do some research to learn more about the companies and

the details about the specific positions you'll be applying for.

Play video starting at :1:28 and follow transcript1:28

Then you can update your resume or create a new one.

You'll want it to be specific and reflect what each company is looking for.

But you can definitely have a master resume that you tweak for each position.

Play video starting at :1:40 and follow transcript1:40

It can also help to create a spreadsheet with all of your experiences and

accomplishments to help you decide what to include in your resume for each.

Play video starting at :1:49 and follow transcript1:49

If you're using a professional networking site like LinkedIn,

you might already have connections who can help you with your job search.

Maybe you know someone who can write a referral for you

or knows of a job within their company that would suit you.

And even if you don't have any luck with your connections,

you can also reach out to employees of the companies you're interested in.

They might be able to give you some insight on the best ways to highlight your

skills and experience when applying.

And, it's okay if they don't write back. Keep trying!

This is probably a good time to tell you of the most challenging part of

a job search:

hearing the word "no."

You will probably hear it a lot, and that's 100% okay.

It's part of everyone's experience,

especially when switching career paths. People you reach out to might not be able

to help you.

Companies you would love to work for might not have any openings.

Jobs you applied for might be filled by someone else, and

that's all part of the process.

The key is to stay focused.

Don't get discouraged, and above all else believe in yourself.

Okay, speech over, but don't forget it, or I'll be forced to give more speeches.

So, back to your search.

If the company you're applying to is interested,

your first point of contact might be a recruiter.

A recruiter might also reach out to you based on their own research.

They may find your professional profile online and think you're a good match for

a position.

Speaking of which, that's another reason to keep building and

refreshing your online profile.

Recruiters are there to make sure you're a legitimate candidate for

the job posted in the description.

So when you talk with the recruiter, whether on the phone, online or

in person, be professional and personable.

It's natural to feel nervous here.

So, it can help to refer back to your resume to wow them with your knowledge

of the data analytics industry.

And remember, recruiters are also looking for someone and

they're hoping it'll be you.

Here's another tip.

Using technical terms like "SQL" and

"clean data" will show recruiters that you know what you're doing.

Play video starting at :3:55 and follow transcript3:55

Recruiters probably won't go into too much detail about the ins and outs.

But they want to see that you know what you're talking about.

They might also give you prep materials or other recommendations.

Play video starting at :4:7 and follow transcript4:07

Take advantage of these because recruiters want you to do well.

Next up is usually the hiring manager.

This is the most important step.

The hiring manager's job is to evaluate whether you have the ability to do

the work and whether you'd be a good fit for their team.

Your job is to convince them that yes, you do,

and yes, you would be.

A good thing you can do here is use LinkedIn or

other professional sites to research the hiring managers or

even other analysts who have a similar role to the one you're applying for.

The more information you have about the job,

the better your chances of actually getting it.

You should also use this opportunity to ask lots of questions to

help you figure out if the company's a good fit for you.

You can do this when you talk to recruiters too.

Play video starting at :4:55 and follow transcript4:55

Now if the hiring manager sees you as a fit,

it's very possible you'll have at least one more interview.

The point of these interviews is to give your future stakeholders and

teammates a chance to decide if you're the best candidate for the position.

The next step is the best step.

If all goes well, you'll get an official offer.

Usually by phone first and maybe followed by an official letter.

At this point, feel free to celebrate.

Call everyone and celebrate some more.

But even if it's your dream job,

make sure it's a competitive offer before you sign.

Remember, if they reach out to you with an offer,

that means they want you as much as you want them.

If you're interviewing at other places, you can leverage this to figure

out if negotiating for a more competitive offer is possible.

You should also research salaries, benefits, vacation time, and

any other factors that are important to you for similar jobs.

If you can show specific research like company x gives y amount more for

the same role,

there's usually some room to negotiate your salary, vacation days, or

something else.

Keep in mind, you'll need to find a balance between what you want,

what they want to give you, and what's fair.

So know your own worth but also understand that the company hiring you has

already placed a certain value on your role.

Okay, let's say that everything works out, and

you're happy with a negotiated deal and excited to join your new team.

Even then, hit pause and

give yourself at least two weeks before you officially start.

Why?

Well, if you're already employed somewhere else during your job search,

it's customary and polite to give at least a two-week notice at your old job

before starting at the new one.

Plus, it's good to give yourself a break before starting your exciting new

adventure.

You've earned it.

By now you should have a pretty good idea of what to expect

when you start your data analyst job search.

Coming up we'll talk more about building your resume.

See you in the next video.

Great, you're back.

When you take a picture,

you usually try to capture lots

of different things in one image.

Maybe you're taking a picture of

the sunset and want to capture the clouds,

the tree line and the mountains.

Basically, you want a snapshot of that entire moment.

You can think of building a resume in the same way.

You want your resume to be a snapshot of all that

you've done both in school and professionally.

In this video, we'll go through

the process of building a resume,

which you'll be able to add your own details too.

Keep in mind this is a snapshot.

When managers and recruiters

look at what you've included in your resume,

they should be able to tell right

away what you can offer their company.

The key here is to be brief.

Try to keep everything in one page and

each description to just a few bullet points.

Two to four bullet points is enough but

remember to keep your bullet points concise.

Sticking to one page will help you stay

focused on the details that best

reflect who you are or who you want to be professionally.

One page might also be all that

hiring managers and recruiters have time to look at.

They're busy people, so you want to get

their attention with your resume as quickly as possible.

Now let's talk about actually building your resume.

This is where templates come in.

They're a great way to build

a brand new resume or reformat one you already have.

Programs like Microsoft Word or Google Docs and

even some job search websites

all have templates you can use.

A template has placeholders

for the information you'll need to

enter and its own design elements

to make your resume look inviting.

You'll have a chance to explore

this option a little later.

For now, we'll go through the steps you can

take to make your resume professional,

easy to read and error-free.

If you already have a resume document,

you can use these steps to tweak it.

Now, there's more than one way to build a resume,

but most have contact information

at the top of the document.

This includes your name,

address, phone number, and email address.

If you have multiple email addresses or phone numbers,

use the ones that are most

reliable and sound professional.

It's also great if you can use

your first and last name in your email address,

like janedoe17@email.com.

You should also make sure that your contact information

matches the details that you've

included on professional websites.

While most resumes have

contact information in the same place,

it's up to you how you organize that info.

A format that focuses more on skills and qualifications

and less on work history is great for

people who have gaps in their work history.

It's also good for those who

are just starting out their career

or making a career change, and that might be you.

If you do want to highlight your work history,

feel free to include details of

your work experience starting with your most recent job.

If you've had lots of jobs that are

related to a new position you're applying for,

this format make sense.

If you're editing a resume you already have,

you can keep it in the same format

and adjust the details.

If you're starting a new one or

building a resume for the first time,

choose the format that makes the most sense for you.

There's lots of resume resources online.

You should browse through

a bunch of different resumes to get

an idea of the formats you think work best for you.

Once you've decided on your format,

you can start adding your details.

Some resumes begin with the summary,

but this is optional.

A summary can be helpful if you

have experience that is not traditional

for a data analyst

or if you're making a career transition.

If you decide to include a summary,

keep it to one or two sentences that highlight

your strengths and how

you can help the company you're applying to.

You'll also want to make sure your summary

includes positive words about yourself,

like dedicated and proactive.

You can support those words with data,

like the number of years you've worked or the tools

you're experienced in like SQL and spreadsheets.

A summary might start off with something like

hardworking customer service representative

with over five years of experience.

Once you've completed this program

and have your certificate,

you'll be able to include that too,

which could sound like this,

"entry-level data analytics professional recently

completed the Google

Data Analytics Professional Certificate."

Sounds pretty good, doesn't it?

Another option is leaving

a placeholder for your summary while you build

the rest of your resume and then

writing it after you finish the other sections.

This way, you can review the skills and experience you've

mentioned and grab two or

three of the highlights to use in your summary.

It's also good to note that the summary might change a

little as you apply for different jobs.

If you're including a work experience section,

there's lots of different types

of experience you could add.

Outside of jobs with other companies,

you could also include volunteer positions you've

had and any freelance or side work you've done.

The key here is the way in

which you describe these experiences.

Try to describe the work you did in a way

that relates to the position you're applying for.

Most job descriptions have

minimum qualifications or requirements listed.

These are the experiences, skills,

and education you'll need to be considered for the job.

It's important to clearly state them in your resume.

If you're a good match,

the next step is checking out preferred qualifications,

which lots of job descriptions also include.

These aren't required,

but every additional qualification you

match makes you a more

competitive candidate for the role.

Including any part of

your skills and experience that matches

a job description will help

your resume rise above the competition.

If a job listing describes

a job responsibility as

"effectively managing data resources,"

you'll want to have your own description

that reflects that responsibility.

For example, if you volunteered or

worked at a local school or community center,

you might say that you "effectively managed

resources for after-school activities."

Later on, you'll learn

more ways to make your work history work for you.

It's helpful to describe your skills

and qualifications in the same way.

For example, if a listing talks

about organization and partnering with others,

try to think about relevant experiences you've had.

Maybe you've helped organize the food drive

or partnered with someone to start an online business.

In your descriptions,

you want to highlight the impact you've

had in your role, as

well as the impact the role had on you.

If you helped a business get

started or reach new heights,

talk about that experience

and how you played a part in it.

Or if you worked at a store when it first opened,

you can say that you helped launch

the successful business by

ensuring quality customer service.

If you used data analytics in any of your jobs,

you'll definitely want to include that as well.

We'll cover how to add

specific data analysis skills a little bit later.

One way to do this is to

follow a formula in your descriptions:

Accomplished X as measured by Y, by doing Z.

Here's an example of how this might read on a resume:

Selected as one of 275 participants nationwide for

this 12-month professional development program for high-

achieving talent based on

leadership potential and academic success.

If you've gained new skills in one of your experiences,

be sure to highlight them all and how they helped.

This is probably as good a spot as

any to bring up data analytics.

Even if this program is

the first time you really thought about data analytics,

now that you're equipped with some knowledge,

you'll want to use that to your benefit.

If you've ever managed money,

maybe that means you helped

the business analyze future earnings.

Or maybe you created a budget based

on your analysis of previous spending.

Even if it was for your own or a friend's small business,

it's still data that you've analyzed.

Now you can reflect on when

and how and use it in your resume.

After you've added work experience and skills,

you should include a section for

any education you've completed.

Yes, this course absolutely counts.

You can add this course as part of your education,

and you can also refer to it in

your summary and skill sections.

Depending on the format of your resume,

you might want to add a section

for technical skills you've

acquired both in this course and elsewhere.

Besides technical skills like SQL,

you could also include

language proficiencies in this section.

Having some ability in a language other than

English can only help your job search.

Now you have an idea of how to make

your resume look professional and appealing.

As you move forward,

you'll learn even more about

how to make your resume shine.

By the end, you'll have a resume you can be proud of.

Next up, we'll talk about how to make your resume

truly unique. See you soon.

Great to see you again.

Building a strong resume is

a great way to find success in your job hunt.

You've had the chance to start building your resume,

and now we'll take the next step by showing

you how to refine your resume for

data analytics jobs. Let's get started.

For data analytics, one of the most important things

your resume should do is

show that you are a clear communicator.

Companies looking for analysts want to know that

the people they hire can do the analysis,

but also can explain it to

any audience in a clear and direct way.

Your first audience as a data analyst will most

likely be hiring managers and recruiters.

Being direct and coherent in

your resume will go a long way with them as well.

Let's start with the summary section.

While you won't go into too much detail in

this section about any of your work experiences,

it's a good spot to point out if you're

transitioning into a new career role.

You might add something like,

"transitioning from a career in

the auto industry and seeking

a full-time role in the field of data analytics."

One strategy you can use in your summary and throughout

your resume is P-A-R, or PAR statements.

PAR stands for Problem, Action, Result.

This is a great way to help you

write clearly and concisely.

Instead of saying something like,

"was responsible for writing two blogs a month,"

you'd say, "earned little-known website

over 2,000 new clicks through strategic blogging."

The website being little-known is the problem.

The strategic action is

the strategic blogging. And

the result is the 2,000 new clicks.

Adding PAR statements to your job descriptions or

skill section can help with

the organization and consistency in your resume.

They definitely helped me when I changed jobs.

Speaking of the skill section,

make sure you include any skills and qualifications

you've acquired through this course and on your own.

You don't need to be super technical.

But talking about your experience with spreadsheets,

SQL, Tableau, and R,

which is a programming language that we'll get to later,

will enhance your resume

and your chances of getting a job.

If you're listing qualifications or skills,

you might include a spot for

programming languages and then list SQL and R,

which are both a part of

the Google Data Analytics certificate.

You might even add in the top functions,

packages or formulas that

you're comfortable with in each.

It also makes sense to include skills you've

acquired in spreadsheets like pivot tables.

Pivot tables, SQL, R,

and lots of other terms we covered here might

get you noticed by hiring managers and recruiters.

But you definitely want your resume to

accurately represent your skills and abilities.

Only add these skills

after you've completed this certificate.

Once you start applying the ideas we

talked about here to your resume,

you'll be well on your way to setting

yourself apart from other candidates.

After you've completed your final course,

you'll have the opportunity to complete

a case study and link it on your resume.

This'll be a great opportunity

to show recruiters and hiring

managers the skills you've

learned while earning your certificate.

Before you know it, you'll have

a pretty great resume that you can

update quickly whenever you're

searching for a data analyst job.

Nothing wrong with that. Up next,

we'll talk more about adding

experienced to your resume. Bye for now.

Hello, my name is Joseph.

I'm a people analyst at Google.

As a people analyst, my job is to

work with executives and

HR business partners to

use data to make informed people decisions.

Inclusion is very essential to the work that we do.

As you know, sometimes you can start with

data and have your own bias in it.

For us in this field that is very sensitive,

it requires that we have a diverse set of people who

have different backgrounds to

have this lens of data to work.

Being a black professional,

I can better tell a story about people of color

that is a lot more personal to me.

Being an analyst requires me

to take data and tell a story with it.

On a personal standpoint,

I'm very passionate about this space of

increasing representation in the tech industry.

For example, outside of work,

I run a nonprofit called Sankofa Tech.

Our whole goal is essentially to

help develop the next generation of

black engineers who can essentially be in

this field and represent our experience

using data as a foundation and offer technology as

the powering moving factor going forward.

It's critical that we have

more black people in the technology sector.

As you all know, in the next 10-20 years, AI,

machine learning, will be like just speaking

English in this country or even the entire world.

So the more we can have more black people in this field,

the more we can represent it in

the products that are being built,

and the more that our experiences are being influenced

in every single product that these companies do build.

It's definitely critical that

we have more black engineers,

we have more black data scientists to do the analysis,

and also just black data analysts to help tell

the story that's more

inclusive of our experience as well.

It's definitely essential that we do have people from

different backgrounds, colors, creeds

to really understand data,

and have the alliance to it,

and tell the story,

and make it very personal to our audience.

**The importance of diversity on a data analytics team**

As you’ve been learning, there are aspects of data analysis in which bias can play a role. This is why diversity and inclusion are important in data teams and can often be key to taking fair and accurate insights from the data. In this reading, you’re going to explore some of the ways bias can influence data analysis, the ethical considerations of working with data, and how you can mitigate bias.

A group of people standing in a room

Description automatically generated

**Data collection and bias**

Even before data is collected, bias might have been involved in determining what data is collected and how. Biased data can lead to inaccurate insights, or it can reinforce existing inequity. An example of this might be a bank that focuses its data collection efforts on predominantly affluent areas. Perhaps residents there have a history of being more profitable bank clients. However, this would exclude lower-income individuals from the bank’s  marketing strategies. A more diverse data team is more likely to notice gaps in data collection and understand the impact on new loan initiatives. Plus, extending loans to a broader base can help banks attract more customers, opening up new markets and revenue streams.

**Data interpretation and bias**

Data analysts interpret data to gain insights and guide decision-making. For example, an analyst in a healthcare organization might review data about patient outcomes. If the analyst doesn’t understand or account for disparities in healthcare among various populations, they may present an incomplete report on what the data means. With a diverse group of data analysts, it’s more likely that someone will be familiar with these disparities and factor in additional data to create a more complete picture. For example, a diverse team analyzing healthcare data may notice that a particular study on heart disease focuses primarily on male participants. Team members may recognize this gender bias and recommend conducting a similar study with a more balanced representation of all participants to understand sex-specific risk factors.

**Ethical considerations**

Data analysts often have access to a great deal of data and must consider what's right and fair when collecting and interpreting it. It’s important to establish and follow ethical guidelines for data use, some of which are included below:

* **Informed consent:** Obtain informed and voluntary consent from individuals before collecting their data. This ensures that people are aware of how their data will be used and gives them the opportunity to make an informed decision about data sharing.
* **Anonymity and confidentiality:** Protect the privacy of individuals through de-identification and confidentiality measures. This guideline helps prevent unauthorized access to sensitive data and maintains trust in data handling.
* **Data security:** Implement robust data security measures to safeguard data from breaches and unauthorized access. Data security is foundational for protecting individuals' information.
* **Transparency:** Transparency in data use practices is critical for building trust. Clearly communicate how data is collected, processed, and shared.
* **Data ownership and control:**  Make sure to give people control over their data, including the ability to access, correct, or delete it.

If your team is small or lacks diversity, there are steps you can take to identify and mitigate bias in data collection and analysis:

* Educate yourself on unconscious bias and its impact on data analysis.
* Review data collection methods to find and fix bias in tools and survey questions.
* Be transparent in documenting all aspects of your data collection and analysis methods to allow others to review and identify potential biases in your methodology.
* Consult with diverse data experts and seek peer reviews from colleagues with diverse perspectives.
* Engage in continuous learning and in discussions with other data analysts to stay up to date on the latest research and techniques.

**Key takeaways**

Bias in data collection and interpretation can have wide-ranging impacts. Data analysts can make a positive impact by practicing ethical data collection e and striving to engage diverse voices in their work.

# CareerCon resources on YouTube

The data analytics industry is always changing and constantly aiming to improve its diversity. Google is proud to support a well-rounded education and a more inclusive environment for all of our learners.

Kaggle's CareerCon resources are for anyone interested in a data analyst career.

## What is CareerCon?

Have you ever wanted to get into the mind of a data analyst? Kaggle’s CareerCon is an annual and free digital event whose aim is to help new data analysts land their first job in the field. Recorded sessions from CareerCon offer tons of firsthand knowledge and expert advice from top data analysts and hiring managers through seminars, coding workshops, and resume advice.

Although the resources offered are aimed at data scientists, the principles and guidelines are still similar to what data analysts can expect on their career journey.

### **Important note about CareerCon**

Most likely due to COVID-19, CareerCon 2019 was the last event held. At the time of this writing, there were no publicly available plans for future CareerCon events.

## CareerCon 2019 resources

Browse the [full sessions for CareerCon 2019](https://www.youtube.com/playlist?list=PLqFaTIg4myu-npFrYu6cO7h7AI6bkcOlL).

Be sure to check out [Portfolio and resume analysis with data science hiring managers](https://www.youtube.com/watch?v=cBbYhhH399c&list=PLqFaTIg4myu-npFrYu6cO7h7AI6bkcOlL&index=8): A panel of hiring managers discusses what they are seeking in candidates and how they examine different resumes submitted by job seekers like you. Learn from the mistakes of others and get ahead of the curve by adapting your resume/portfolio to avoid the noted mistakes and capitalize on what others have done well in their resumes.

## Highlights from CareerCon 2018

[How to build a compelling data science portfolio and resume](https://www.youtube.com/watch?v=xrhPjE7wHas&list=PLqFaTIg4myu-dNobDHQZPrD2wH27PthCG): A hiring manager from Quora reviews actual resumes from data science candidates and gives candid feedback on areas of improvement. Learn what to include and omit from your resume and  portfolio as well as formatting tips. This offers a great firsthand look into what hiring managers are seeking when reviewing your resume and portfolio.

[Overview of the Data Science Interview Process](https://www.youtube.com/watch?v=X6orAXDIrds&list=PLqFaTIg4myu-dNobDHQZPrD2wH27PthCG&index=5): Hiring managers at Google discuss typical data science interviews, including the soft and hard skills you will want to prioritize. You will get a better sense of the interview process from both sides, and better prepare yourself for what to expect when interviewing for a data science role.

[Live Breakdown of Common Data Science Interview Questions](https://www.youtube.com/watch?v=aXUsrKPTBvY&list=PLqFaTIg4myu-dNobDHQZPrD2wH27PthCG&index=6): Watch a mock interview to see how a Kaggle data scientist answers questions during a data science interview. The video also includes live coding! This video is great preparation for some of the most commonly asked data science interview questions.

[Am I a Good Fit? Identifying Your Best Data Science Job Opportunities](https://www.youtube.com/watch?v=0W0Zrc-m5r8&list=PLqFaTIg4myu-dNobDHQZPrD2wH27PthCG&index=2): Ever wonder where you will fit in for your future career? This chat with Jessica Kirkpatrick, an intelligence manager, gives you a great breakdown of the different types of categories within the data science job market, the different types of job opportunities you may notice, and how you can frame previous work and skills from another career to fit into the data science job market.

[Real Stories from a Panel of Successful Career Switchers](https://www.youtube.com/watch?v=iP0Fxg4oqUQ&list=PLqFaTIg4myu-dNobDHQZPrD2wH27PthCG&index=8): Are you switching careers? Awesome! Learn from people who were in the same position as you and successfully switched their careers into data science. This panel discusses the different experiences in their careers and life that shifted them into the data science field.

Welcome back. Everyone out

there has their own personal work history.

We all started somewhere,

whether part-time or full-time.

What matters for your resume is

how you present the work you've done.

In this video, we'll hone in on work history,

and how you can translate yours

effectively for your data analyst resume.

If you don't have a specific section for

work history in your resume, that's okay.

You can use the same basic ideas to

adjust your skills and qualifications section.

The good news is that you

already have a lot of the skills that

recruiters and hiring agents look

for when they hire data analysts.

You've probably used lots of them in previous jobs.

We call these "transferable skills."

Transferable skills are skills and qualities that can

transfer from one job or industry to another.

Think about all the positions

you've held, associate, owner,

team member, manager, and

how they might be used as a data analyst.

Let's start with the big one that we

talked about before: communication.

When job descriptions say they want

strong communication skills for a data analyst,

it usually means they want

someone who can speak about what they

do to people who aren't as technical or analytical.

If someone who's not familiar with

the analytics can understand what you're talking about

when you try to explain it to them,

your communication skills are usually good-to-go.

You've probably had to communicate in other jobs you had,

whether with employees, customers or clients,

team members, or managers.

You might have had to give presentations too,

whether formal or informal.

In your work history section,

you can highlight how your effective

communication skills have helped you.

You can also refer to specific presentations

you've made and the outcomes of those presentations,

and you can even include

the audience for your presentations,

especially if you present it to

large groups or people in senior positions.

After listing job details,

like the place and length of employment,

you might add something like,

"effectively implemented and communicated

daily workflow to fellow team members,

resulting in an increase in productivity."

Here you'd change the details based on the work you did.

Since you'll be working in the world of data,

including any quantitative data would be ideal.

For example, the increase in

productivity might have been a 15 percent increase.

As long as you have a way to back up your data,

hopefully with more data,

then you can put it in your resume.

This example brings us to the next transferable skill.

Data analysts are problem-solvers.

When problems arise in a database or lines of code,

data analysts need to be able to

find and troubleshoot the problem.

If you have no prior experience working with data,

you can still talk about your problem-solving skills.

That last example we shared does

a great job of showing an ability to problem-solve.

It's actually written as a PAR,

or problem, action, results statement,

which we talked about earlier.

The problem is that the daily workflow procedures

were not in place.

The action is that you put the procedures

into effect and communicated them to your team,

and the result is that

productivity increased by 15 percent.

This makes it clear that there was

a problem, and you solved it.

We can also use a statement to point out teamwork

as an important quality to

bring to the data analyst world.

While you might have plenty of work to do on your own,

it'll always be for the benefit of the team.

Team means not only the data team you're part of,

but the whole company as well.

That's a few skills you can add to

your work experience and

skills and qualifications sections.

All of these are known as soft skills.

Soft skills are non-technical traits

and behaviors that relate to how you work.

Being detail-oriented and demonstrating perseverance are

two more examples of soft skills that

anyone hiring a data analyst will look for.

Companies want to know that you will do

your analysis carefully and to completion,

no matter what setbacks you might face along the way.

If you worked at a retail job,

you can talk about how your attention to detail helps you

find discrepancies while handling a high volume of money,

and you could add how you continue to

practice customer service at a high level,

despite a high turnover rate at the management level.

These are just some examples to think

about and apply to your work details.

Take a moment and think back to your last job,

or maybe it's your current job.

What soft skills do you use to find success?

Are you starting to understand how those are

transferable to the world of data analytics?

Using PAR statements and focusing on

your transferable soft skills can

really add to the power of your resume.

Now you can keep powering on to the next step to continue

learning about the data analytics field

and your future job in it.

See you in the next video.

Hi, I'm Kate. I'm a senior product analyst at Google.

I have always been perhaps an annoyingly curious person.

Even as a child, I remember I would

take things apart just to see how they worked.

I just love seeing how things work together,

and I love asking new questions.

I love having more information.

I think that makes me a more

well-rounded person and definitely a better analyst.

Every step in my career,

including the first step in the army,

I always picked at what I could in terms of trying to

self teach on things like databases and things like data.

One of my first forays into data was,

I had been deployed and when I came back,

I worked with the personnel office

and we had to do things like track where

everybody was and what their pay was

and their rank and if they were getting awards,

and there wasn't a single system to work through in that,

so I use an Access database.

It took me forever to learn what a foreign key

was and what a primary key was.

I'll be totally honest,

I did really poorly.

I ended up going back to Excel,

but it was a really good learning experience.

After my time in the army,

I didn't have a sense of what I wanted to do.

I had been doing personnel,

but I really did still enjoy the technology piece.

I somehow spun my army career into logistics and

got a job doing

logistics for what they call the roundhouse.

It is where they work on the locomotive engines.

I did a lot of database maintenance.

When I left the railroad,

I went to a welding company where I started

out as a logistics person working on trucks.

I mean truck parts. But then I was able to

transition into a more database data-focused role.

After my time at the welding company,

I was ready to try something a lot more technical.

I actually ended up working

for a small consulting firm that was very

boutiquey and did a lot of work with Tableau,

where we started to work with

companies and taught them how to do data visualization.

I did Tableau training for awhile.

But really I was there for over six years,

and through my time there,

I did database engineering,

I did data engineering.

I got to run a team of analysts,

I got to teach people how to do consulting.

There was a lot of growth for

me in that six-year time period.

After that, I decided to come to Google.

I get to work with stakeholders on

translations throughout the Google world.

If anybody wants to translate

something from one language to another,

I get to work on the analytics of that.

That means that if you

take 500 different languages

or 40 different languages, what does it cost?

How many words do we translate?

What does that translation quality look like?

If I look back on my career,

I would have told myself five,

ten years ago to focus on something.

Don't try to feel too overwhelmed.

The important thing to be able to do is to

be able to ask the right question

and know how to answer it.

I have confidence.

Confidence is really important

because people are coming to me for answers.

That's my job, is to think

really hard about the questions and give

them answers that make

them better and make the company better.

The fact that I know that I can do this now,

now that I've put that time and effort into it,

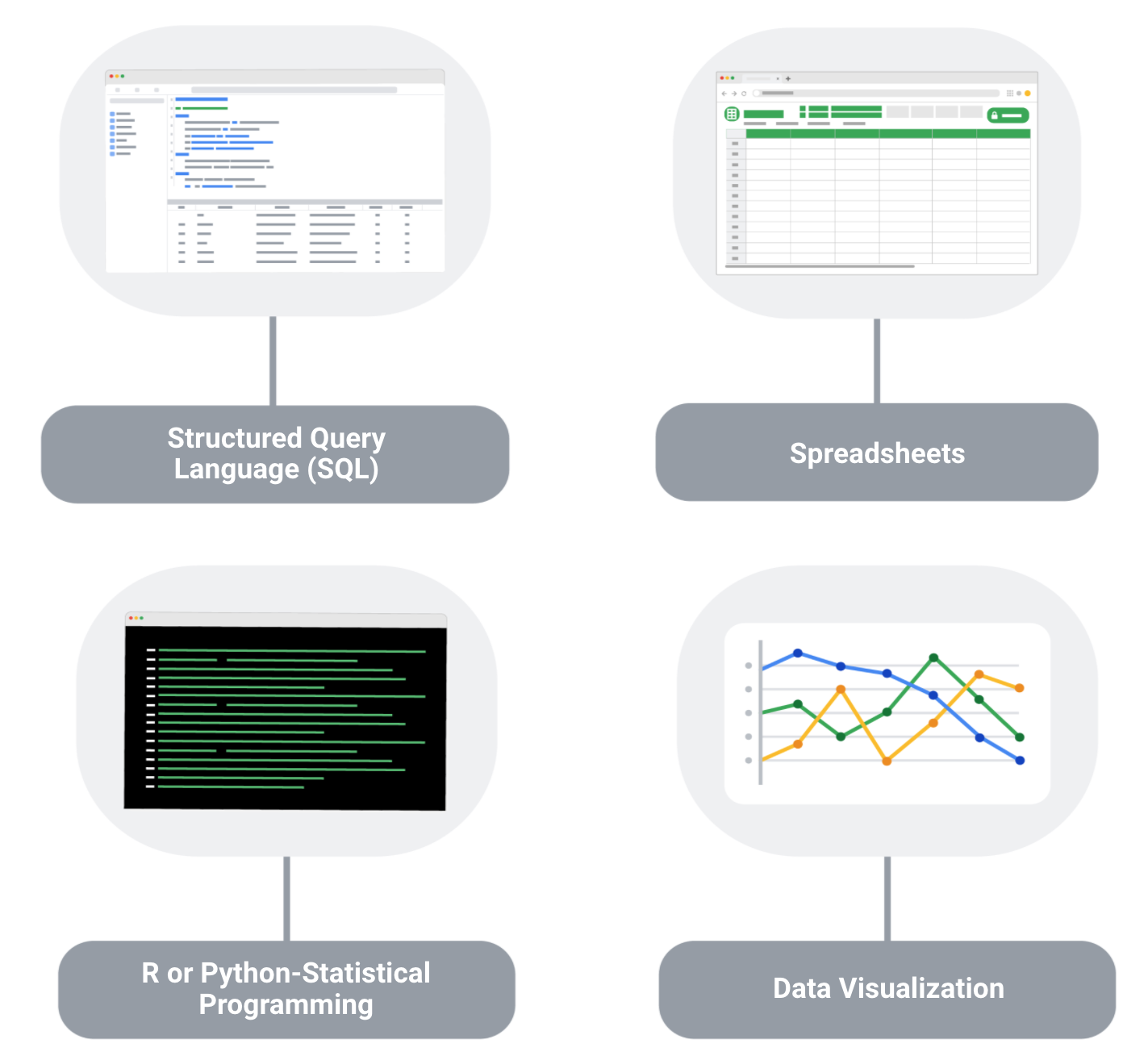
it's really, really rewarding.

**Add technical skills to your resume**

Technical skills are crucial when building a solid resume. They demonstrate to employers that you have the skills necessary to successfully perform the job. Think of technical skills as your toolbox: How you list each skill on your resume is how you demonstrate to employers that you are capable of using those tools.

**Common technical skills for entry-level data analysts**

It takes lots of skills to be a successful data analyst, and these are some common ones that employers seek out when hiring for data analyst jobs:



**1. Structured Query Language (SQL):** SQL is considered a basic skill that is pivotal to any entry-level data analyst position. SQL helps you communicate with databases, and more specifically, it is designed to help you retrieve information from databases. Every month, thousands of data analyst jobs posted require SQL, and knowing how to use SQL remains one of the most common job functions of a data analyst.

**2. Spreadsheets:** Although SQL is popular, 62% of companies still prefer to use spreadsheets for their data insights. When getting your first job as a data analyst, the first version of your database might be in spreadsheet form, which is still a powerful tool for reporting or even presenting data sets. So, it is important for you to be familiar with using spreadsheets for your data insights.

**3. Data visualization tools:** Data visualization tools help to simplify complex data and enable the data to be visually understood. After gathering and analyzing data, data analysts are tasked with presenting their findings and making that information simple to grasp. Common tools that are used in data analysis include Tableau, Microstrategy, Data Studio, Looker, Datarama, Microsoft Power BI, and many more. Among these, Tableau is best known for its ease of use, so it is a must-have for beginner data analysts. Also, studies show that data analysis jobs requiring Tableau are expected to grow about 34.9% over the next decade.

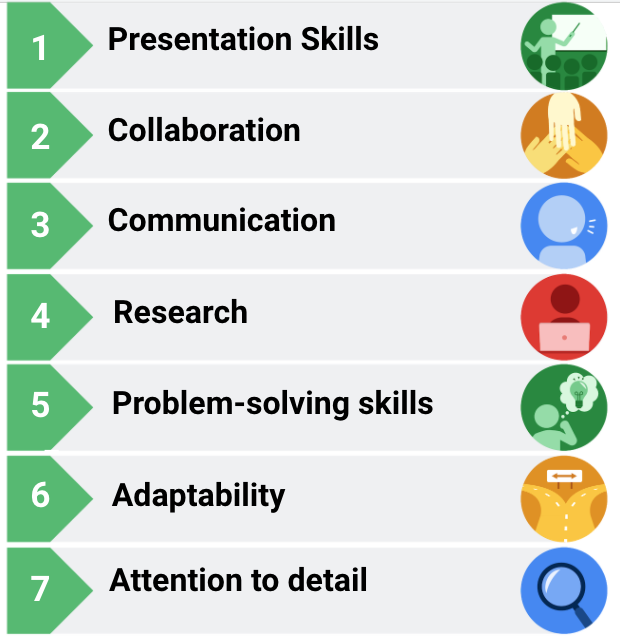
**4**. **R or Python programming:** Since only less than a third of entry-level data analyst positions require knowledge of Python or R, you don’t need to be proficient in programming languages as an entry-level data analyst. But, R or Python are great additions to have as you become more advanced in your career.

**Key takeaways**

As a data analyst, you are often asked to collect and analyze data with a specific purpose in mind. Knowing which platform and language to use helps you analyze the data to decipher which information is important, to probe for any anomalies, prepare questions, assess risks, and so much more.

# Add professional skills to your resume

There is more than just data when it comes to being a data analyst—there are plenty of professional skills that can set you apart from other candidates so that potential employers will notice you and know that you have the ability to succeed in this role. Here are some of the most common professional skills you will find in an entry-level data analyst resume.

List of data analytics skills/characteristics next to icons. Presentation skills, collaboration, communication, research, problem-solving skills, adaptability, attention to detail.

1. **Presentation skills**

Although gathering and analyzing data is a big part of the job, presenting your findings in a clear and simple way is just as important. You will want to structure your findings in a way that allows your audience to know exactly what conclusions they are supposed to draw.

2. **Collaboration**

As a data analyst, you will be asked to work with lots of teams and stakeholders—sometimes internal or external—and your ability to share ideas, insights, and criticisms will be crucial. It is important that you and your team—which might consist of engineers and researchers—do your best to get the job done.

3. **Communication**

Data analysts must communicate effectively to obtain the data that they need. It is also important that you are able to work and clearly communicate with teams and business leaders in a language that they understand.

4. **Research**

As a data analyst, even if you have all of the data at your disposal, you still need to analyze it and draw crucial insights from it. To analyze the data and draw conclusions, you will need to conduct research to stay in-line with industry trends.

5. **Problem-solving skills**

Problem-solving is a big part of a data analyst’s job, and you will encounter times when there are errors in databases, code, or even the capturing of data. You will have to adapt and think outside the box to find alternative solutions to these problems.

6. **Adaptability**

In the ever-changing world of data, you have to be adaptable and flexible. As a data analyst, you will be working across multiple teams with different levels of needs and knowledge, which requires you to adjust to different teams, knowledge levels, and stakeholders.

7. **Attention to detail**

A single line of incorrect code can throw everything off, so paying attention to detail is critical for a data analyst. When it comes to understanding and reporting findings, it helps if you focus on the details that matter to your audience.

### Adding professional skills to your resume

Here are a few ways that you can add professional skills to your resume:

1. Analyze your previous work experience and find opportunities to insert a professional skill. For example, if you worked in a restaurant, you could emphasize your communication and adaptability skills that you utilized to effectively function during peak hours.
2. Call attention to your problem-solving, presentation, research, and communication skills in previous projects or relevant coursework.
3. Add a mix of professional and technical skills in the skills or summary section of your resume.

Glossary

Data Analytics

Terms and definitions from Course 4

A

A/B testing: The process of testing two variations of the same web page to determine which

page is more successful at attracting user traffic and generating revenue

Accuracy: The degree to which data conforms to the actual entity being measured or

described

B

C

CASE: A SQL statement that returns records that meet conditions by including an if/then

statement in a query

CAST: A SQL function that converts data from one datatype to another

Changelog: A file containing a chronologically ordered list of modifications made to a project

Clean data: Data that is complete, correct, and relevant to the problem being solved

COALESCE: A SQL function that returns non-null values in a list

Compatibility: How well two or more datasets are able to work together

Completeness: The degree to which data contains all desired components or measures

CONCAT: A SQL function that adds strings together to create new text strings that can be

used as unique keys

CONCATENATE: A spreadsheet function that joins together two or more text strings

Conditional formatting: A spreadsheet tool that changes how cells appear when values meet

specific conditions

Confidence interval: A range of values that conveys how likely a statistical estimate reflects

the population

Confidence level: The probability that a sample size accurately reflects the greater population

Consistency: The degree to which data is repeatable from different points of entry or

collection

COUNTA: A spreadsheet function that counts the total number of values within a specified

range

COUNTIF: A spreadsheet function that returns the number of cells in a range that match a

specified value

Cross-field validation: A process that ensures certain conditions for multiple data fields are

satisfied

D

Data constraints: The criteria that determine whether a piece of a data is clean and valid

Data engineer: A professional who transforms data into a useful format for analysis and gives

it a reliable infrastructure

Data ethics: Well-founded standards of right and wrong that dictate how data is collected,

shared, and used

Data governance: A process for ensuring the formal management of a company’s data assets

Data integrity: The accuracy, completeness, consistency, and trustworthiness of data

throughout its life cycle

Data manipulation: The process of changing data to make it more organized and easier to

read

Data mapping: The process of matching fields from one data source to another

Data merging: The process of combining two or more datasets into a single dataset

Data range: Numerical values that fall between predefined maximum and minimum values

Data replication: The process of storing data in multiple locations

Data transfer: The process of copying data from a storage device to computer memory or

from one computer to another

Data validation: A tool for checking the accuracy and quality of data

Data warehousing specialist: A professional who develops processes and procedures to

effectively store and organize data

DATEDIF: A spreadsheet function that calculates the number of days, months, or years

between two dates

Delimiter: A character that indicates the beginning or end of a data item

Dirty data: Data that is incomplete, incorrect, or irrelevant to the problem to be solved

DISTINCT: A keyword that is added to a SQL SELECT statement to retrieve only non-duplicate

entries

Duplicate data: Any record that inadvertently shares data with another record

E

Estimated response rate: The average number of people who typically complete a survey

F

Field length: A tool for determining how many characters can be keyed into a spreadsheet field

Find and replace: A tool that finds a specified search term and replaces it with something else

Float: A number that contains a decimal

G

H

Hypothesis testing: A process to determine if a survey or experiment has meaningful results

I

Incomplete data: Data that is missing important fields

Inconsistent data: Data that uses different formats to represent the same thing

Incorrect/inaccurate data: Data that is complete but inaccurate

J

K

L

LEFT: A function that returns a set number of characters from the left side of a text string

LEN: A function that returns the length of a text string by counting the number of characters it

contains

Length: The number of characters in a text string

M

Mandatory: A data value that cannot be left blank or empty

Margin of error: The maximum amount that the sample results are expected to differ from

those of the actual population

Merger: An agreement that unites two organizations into a single new one

MID: A function that returns a segment from the middle of a text string

N

Null: An indication that a value does not exist in a dataset

O

Outdated data: Any data that has been superseded by newer and more accurate information

P

Q

R

Random sampling: A way of selecting a sample from a population so that every possible type

of the sample has an equal chance of being chosen

Regular expression (RegEx): A rule that says the values in a table must match a prescribed

pattern

Remove duplicates: A spreadsheet tool that automatically searches for and eliminates

duplicate entries from a spreadsheet

RIGHT: A function that returns a set number of characters from the right side of a text string

S

Soft skills: Nontechnical traits and behaviors that relate to how people work

Split: A spreadsheet function that divides text around a specified character and puts each

fragment into a new, separate cell

Statistical power: The probability that a test of significance will recognize an effect that is

present

Statistical significance: The probability that sample results are not due to random chance

SUBSTR: A SQL function that extracts a substring from a string variable

Substring: A subset of a text string

Syntax: The predetermined structure of a language that includes all required words, symbols,

and punctuation, as well as their proper placement

T

Text string: A group of characters within a cell, most often composed of letters

Transferable skills: Skills and qualities that can transfer from one job or industry to another

TRIM: A function that removes leading, trailing, and repeated spaces in data

Typecasting: Converting data from one type to another

U

Unique: A value that can’t have a duplicate

V

Validity: The degree to which data conforms to constraints when it is input, collected, or

created

Verification: A process to confirm that a data-cleaning effort was well executed and the

resulting data is accurate and reliable

VLOOKUP: A spreadsheet function that vertically searches for a certain value in a column to

return a corresponding piece of information

W

X

Y

Z

Great job on finishing this course.

You've covered a lot of ground

already and learned so much.

There's only one thing left to do: keep going.

But if you ever need to check

back on what you've learned,

these videos will still be available to you.

Right now, I'm excited for you to meet

your instructor for the next course, Aiyana.

She's ready to guide you

through the next part of the program

as you continue in your journey

to becoming a data analyst.

You've learned how to prepare and process data,

and in the next course,

Aiyana will show you how to analyze it.

We'll explore how to make sense of

all the data you've collected and cleaned.

And you'll learn how to ask the right questions

and use data to find answers.

We'll also show you how to

organize and format the data once again,

so it's completely ready for analysis.

We'll talk about aggregation and joins,

two processes that allow you to gather

all the data you need and

summarize it for your stakeholders.

Both spreadsheets and SQL will make an appearance too.

We'll give you some more practice with

using them for calculations.

We'll explain how temporary tables in SQL work.

Awesome job so far and best of luck on what's next!