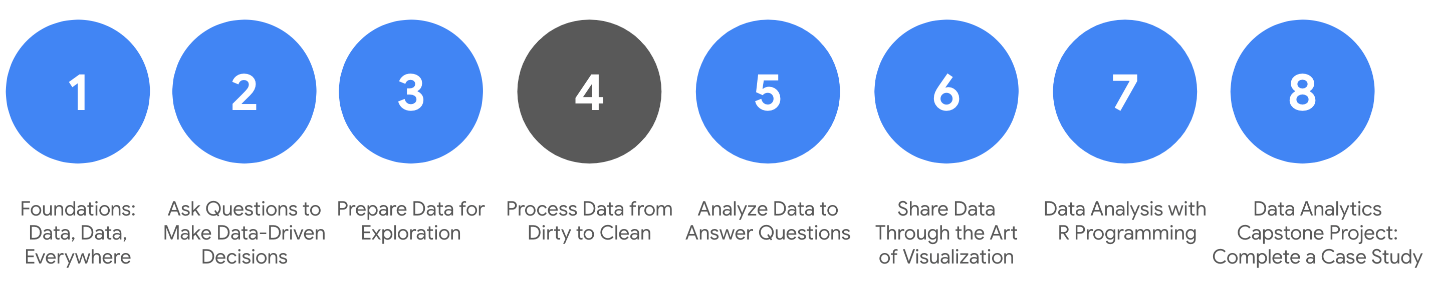
Course syllabus



1. [Foundations: Data, Data, Everywhere](https://www.coursera.org/learn/foundations-data/home/welcome)
2. [Ask Questions to Make Data-Driven Decisions](https://www.coursera.org/learn/ask-questions-make-decisions/home/welcome)
3. [Prepare Data for Exploration](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 D​ata Analytics Capstone: Complete a Case Study](https://coursera.org/learn/google-data-analytics-capstone/home/welcome)

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 attention on practical, hands-on assignments and projects. This progressively increases the amount of time you have to develop important job skills.

In the last course, you worked on some basic skills you will need as an entry-level data analyst. You learned about data structures, and discovered how to obtain, apply, organize, and protect data.

In this course, you will learn to make sure your data is clean by checking it for completeness and correctness. You will review a variety of approaches to clean data in spreadsheets and databases. You will also learn how to verify that your data is clean and how to create reports to communicate that information to others. Ensuring the accuracy and reliability of data is a critical part of a data analyst’s job.

**C​ourse content**

C​ourse 4 – Process Data from Dirty to Clean

1. **E​nsuring data integrity.** Data integrity is necessary to ensure a successful analysis. In this part of the course, you will explore methods and steps that analysts take to check data for integrity. This includes knowing what to do when you have an insufficient amount of data. You will also learn about sample size, avoiding sample bias, and using random samples. All of these measures also help to ensure a successful data analysis.
2. **U​nderstanding clean data.** Every data analyst wants clean data to work with when performing an analysis. In this part of the course, you will learn the difference between clean and dirty data. You will practice data cleaning techniques in spreadsheets and other tools.
3. **C​leaning data using SQL.** Knowing a variety of ways to clean data can make an analyst’s job much easier. In this part of the course, you will use SQL to clean data from databases. You will explore how SQL queries and functions can be used to clean and transform your data before an analysis.
4. **V​erifying and reporting cleaning results.** Cleaning data is an important step in the data analysis process. In this part of the course, you will verify that data is clean and report data cleaning results. With verified clean data, you will be ready for the next step in the data analysis process.
5. **(​Optional) Adding data to your resume.** Creating an effective resume will help you in your data analytics career. In this part of the course, you will learn all about the job application process. Your focus will be on building a resume that highlights your strengths and relevant experience.
6. **C​ompleting the Course Challenge.** At the end of this course, you will be able to apply what you have learned in the Course Challenge. The Course Challenge will ask you questions about the key concepts and then will give you an opportunity to put them into practice as you go through prepared scenarios.

**What to expect**

You can plan to finish this program in about four to five weeks. You will earn course credit after completing all the prescribed activities which include:

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

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

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

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

This course also contains practical information to prepare you for the job market as a data analyst. Use the recommendations to add what you learned about cleaning data to your resume.

**Tips**

* Try to complete all items in order. All new information builds on earlier learnings.
* Treat every task as if it is real-world experience. Have a mindset that you are working at a company or in an organization as a data analyst. This will help you apply what you learn in this program to the real world.
* Repeat demonstrated tasks on your own for extra practice and speed.
* Even though they aren’t graded, participate in and complete all practice items. They will help you build a strong foundation as a data analyst and better prepare you for the graded assessments.
* Take advantage of all additional resources provided, including discussion forums and links to learning content.
* W​hen you encounter useful links in the course, remember to bookmark them so you can refer to the information for study or review.
* Additional resources are free, but some sites place limits on how many articles you can access for free each month. Sometimes you can register on the site for full access, but you can always bookmark a resource and come back to view it later.

Now that you know how to proceed, you can take your first steps towards working with all kinds of data, and learn how to keep data integrity a priority in all of your projects. Stay the course (pun intended)!

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!

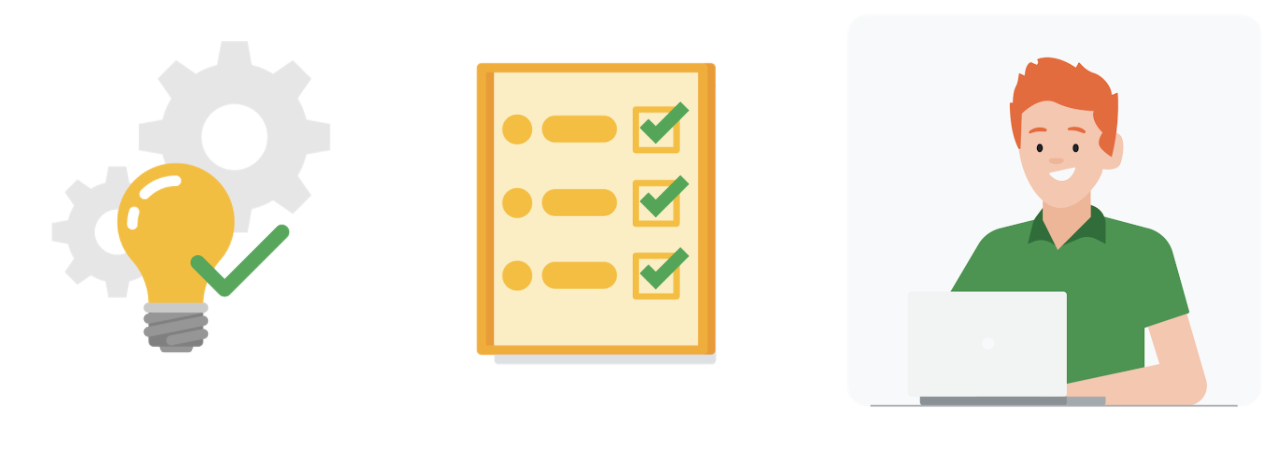
A good analysis depends on the integrity of the data, and data integrity usually depends on using a common format. So it is important to double-check how dates are formatted to make sure what you think is December 10, 2020 isn’t really October 12, 2020, and vice versa.

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. |

# 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.



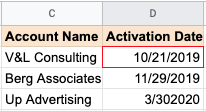
To start off, the data analyst verifies that the data exported to spreadsheets is clean and confirms that the data needed (when users access content) is available. Knowing this, the analyst decides there is good alignment of the data to the business objective. All that is missing is figuring out exactly 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:**

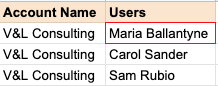


**Result**: October 21, 2019

### **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**:

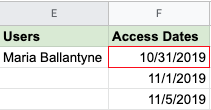


**Result**: Maria Ballantyne

### **Step 3**

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

**Relevant data in spreadsheet:**

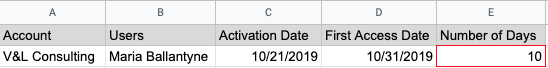


**Result**: October 31, 2019

### **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.

R​efer 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).

R​efer 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 |

### **D​ata cleaning**

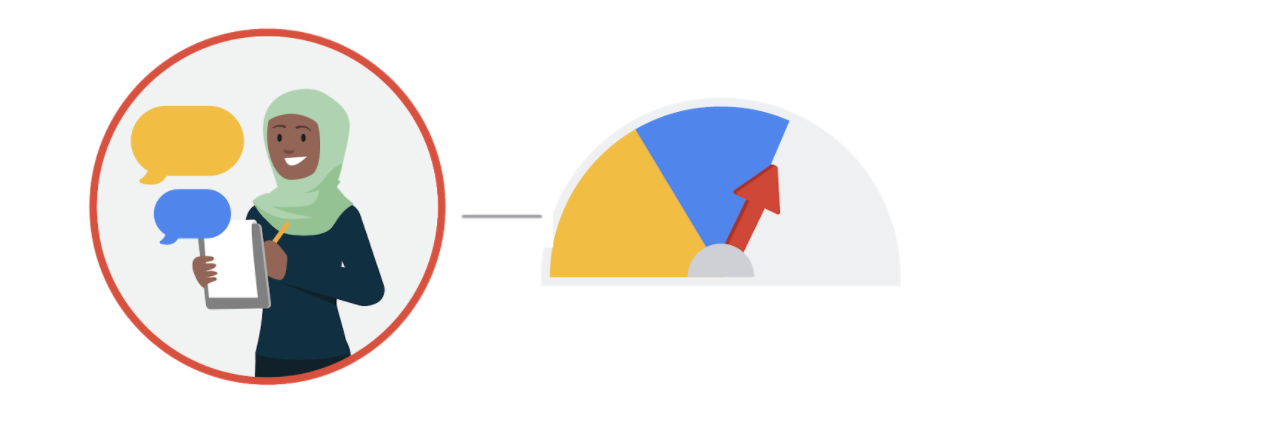
The webinar attendance data seems to align with the business objective. But the data analyst and program manager decide that some data cleaning is needed before the analysis. They think data cleaning is required because:

* 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

### **D​ata 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.

# What to do 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

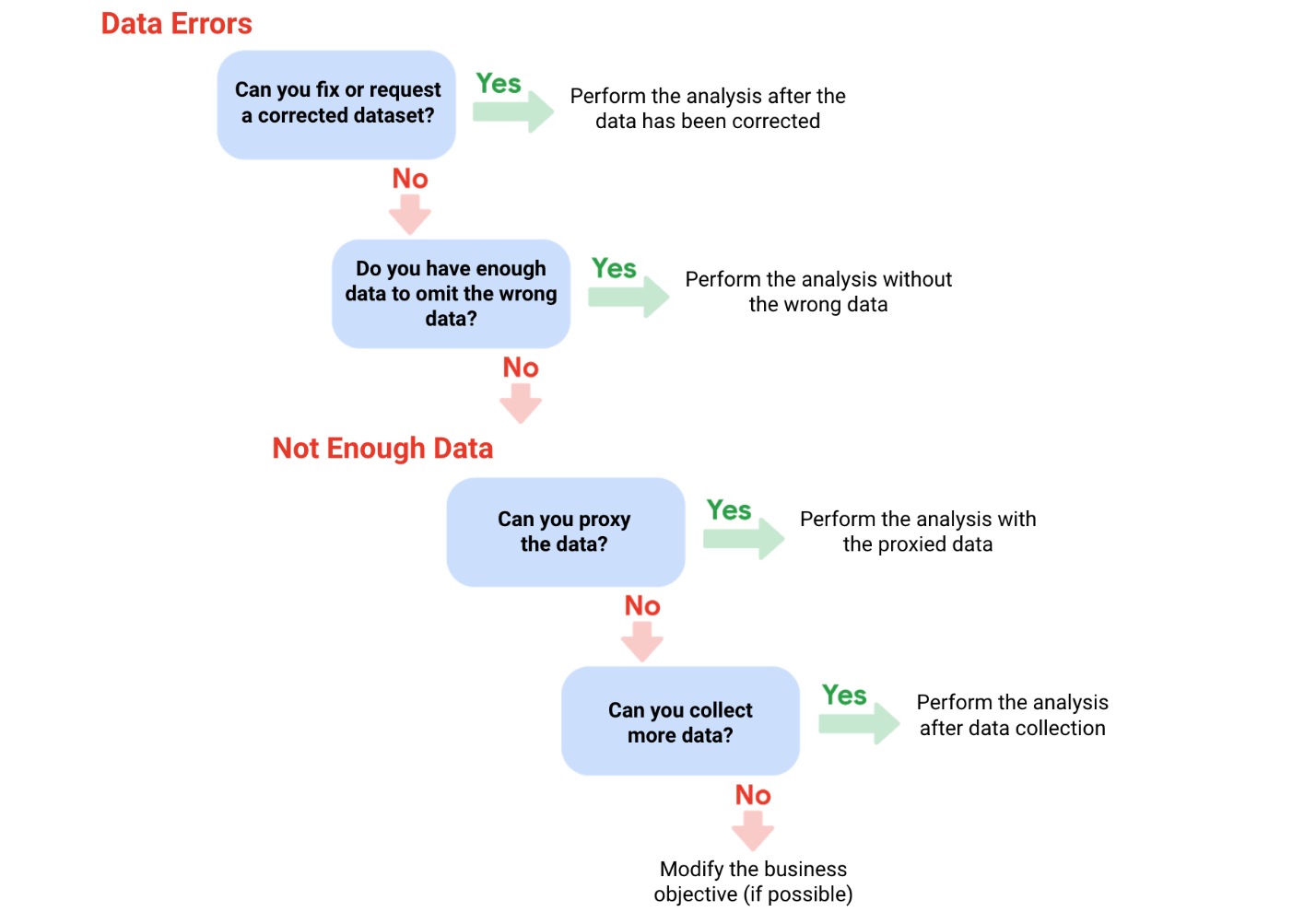
| **Possible Solutions** | **Examples of solutions in real life** |
| --- | --- |
| Do the analysis using proxy data along with actual data. | If you are analyzing trends for owners of golden retrievers, make your dataset larger by including the data from owners of labradors. |
| Adjust your analysis 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:



# Calculating 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.



# What to do when there is no data

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 [Towards Data Science article](https://towardsdatascience.com/is-there-a-difference-between-open-data-and-public-data-6261cd7b5389) 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, J​SON, 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.

# 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.

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 G​ood 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 takeaway**

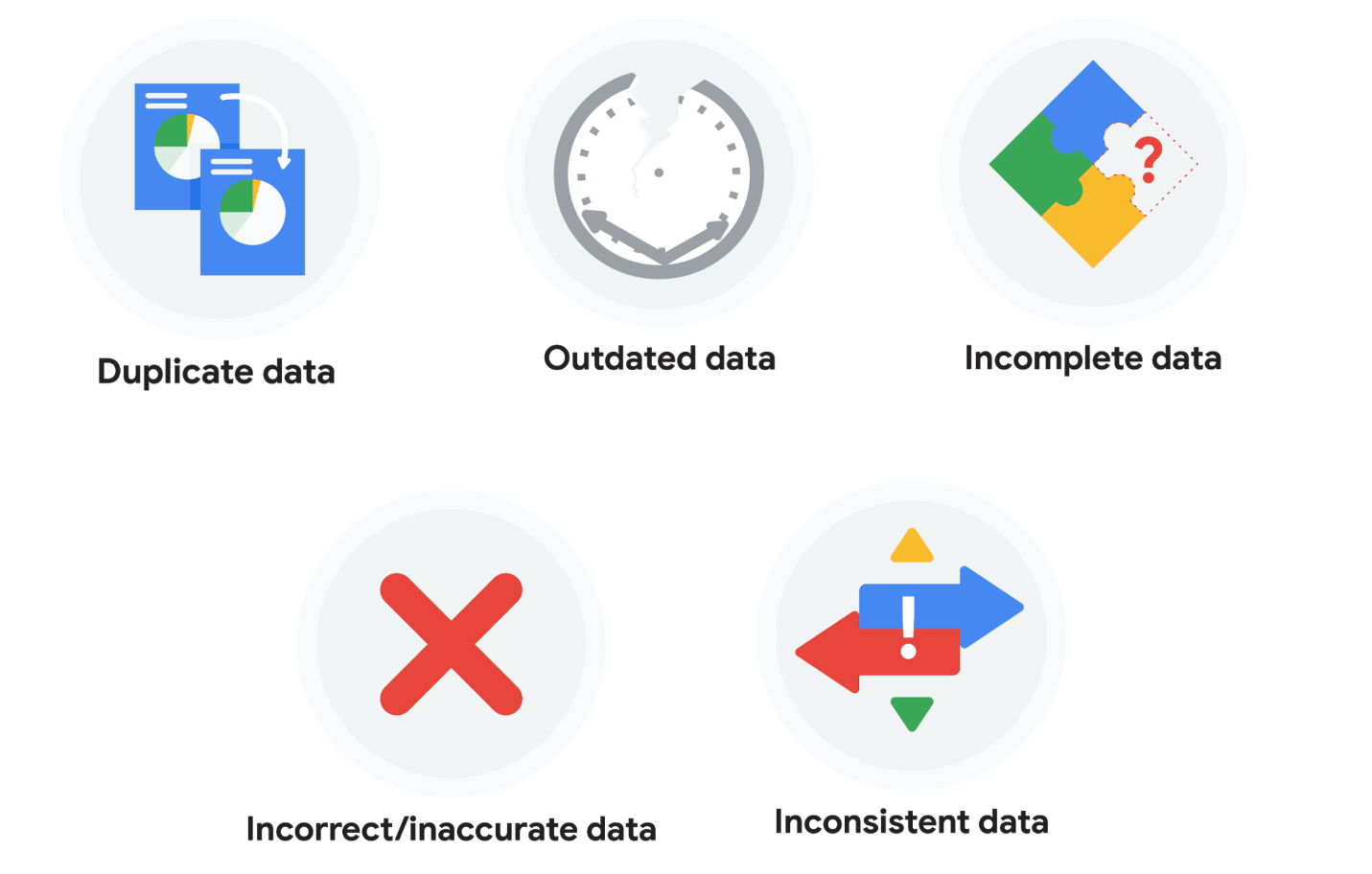
Margin of error is used to determine how close your sample’s result is to what the result would likely have been if you could have surveyed or tested the entire population. Margin of error helps you understand and interpret survey or test results in real-life.  Calculating the margin of error is particularly 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.

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**: 8 out of 10 companies have said that dirty data hinders sales campaigns ([source](https://www.dqglobal.com/2011/05/04/obsolete-or-dirty-data/)).
* **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)).

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.

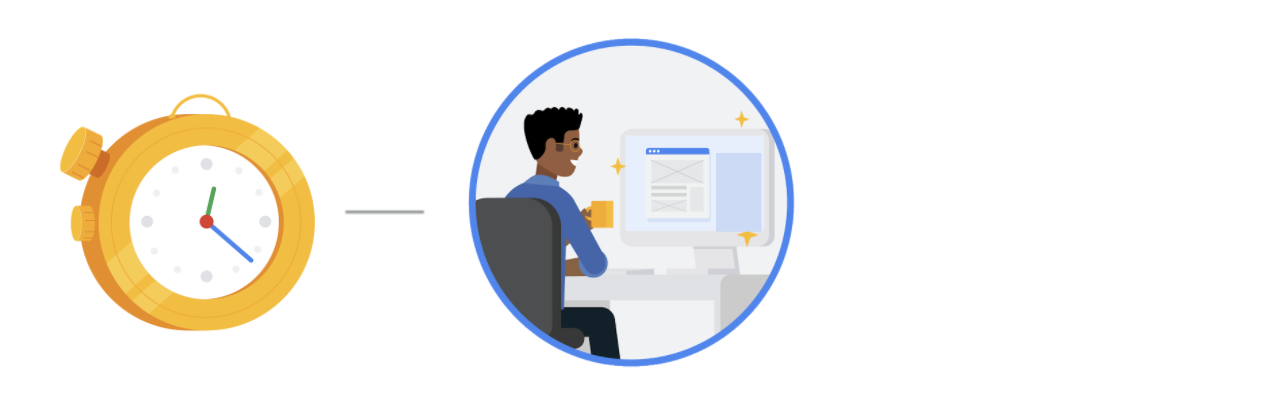
**Additional resources**

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.

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/)

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.

Learning Log: Develop your approach to cleaning data



**Overview**



By this point, you have started working with real data. And you may have noticed that data is often messy-- you can expect raw, primary data to be imperfect. In this learning log, you will develop an approach to cleaning data by creating a cleaning checklist, considering your preferred methods for data cleaning, and deciding on a data cleaning motto. By the time you complete this entry, you will have a stronger understanding of how to approach the data cleaning process methodically. This will help you save time cleaning data in the future and ensure that your data is clean and usable.

**Fill out the Data Cleaning Approach Table**



The problem with data cleaning is that it usually requires a lot of  time, energy, and attention from a junior data analyst. One of the best ways to lessen the negative impacts of data cleaning is to have a plan of action or a specific approach to cleaning the data.

In order to help you develop your own approach, you’ll use the instructions from this learning log to fill out a Data Cleaning Approach Table in your [learning log template](https://docs.google.com/document/d/1W_onDb60axr-Zur7KyL_5dwrW5eZmJF49zM7_FwkHyQ/template/preview). The table will appear like this in the template:

Table

Description automatically generated with medium confidence

Once you have completed your Data Cleaning Approach Table, you will spend some time reflecting on the data cleaning process and your own approach.



**Access your learning log**

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

Link to learning log template: [Develop your approach to data cleaning](https://docs.google.com/document/d/1W_onDb60axr-Zur7KyL_5dwrW5eZmJF49zM7_FwkHyQ/template/preview)

OR

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

**Learning Log Template\_ Develop your approach to cleaning data**DOCX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/26Lfz62pSRai38-tqXkWmQ_bbcdfe84d8b1423b8db5e155d46367c4_Learning-Log-Template_-Develop-your-approach-to-cleaning-data.docx?Expires=1648080000&Signature=Mlhu1ETzv9c5KLlplwhS28TMio3zyQ4A6nAsnwOe-NtYcQz9xcXPPTVibW1xB97tMlispdk-f2IhHgazfZAKxLKHxmAtAfLKhXROEec7mNeLIM3x97ePIYKdGJ-gJ2lL3-URpWPnMooHwC8EqvUzEoY~2y-6Q4e2rnxMU6gYEq4_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



**Step 1: Create your checklist**



You can start developing your personal approach to cleaning data by creating a standard checklist to use before your data cleaning process. Think of this checklist as your default "what to search for" list.

With a good checklist, you can efficiently and, hopefully, swiftly identify all the problem spots without getting sidetracked. You can also use the checklist to identify the scale and scope of the dataset itself.

Some things you might include in your checklist:

* Size of the data set
* Number of categories or labels
* Missing data
* Unformatted data
* The different data types

You can use your own experiences so far to help you decide what else you want to include in your checklist!

**Step 2: List your preferred cleaning methods**



After you have compiled your personal checklist, you can create a list of activities you like to perform when cleaning data. This list is a collection of procedures that you will implement when you encounter specific issues present in the data related to your checklist or every time you clean a new dataset.

For example, suppose that you have a dataset with missing data, how would you handle it? Moreover, if the data set is very large, what would you do to check for missing data? Outlining some of your preferred methods for cleaning data can help save you time and energy.

**Step 3: Choose a data cleaning motto**



Now that you have a personal checklist and your preferred data cleaning methods, you can create a data cleaning motto to help guide and explain your process. The motto is a short one or two sentence summary of your philosophy towards cleaning data. For example, here are a few data cleaning mottos from other data analysts:

1. "Not all data is the same, so don't treat it all the same."
2. "Be prepared for things to not go as planned. Have a backup plan.”
3. "Avoid applying complicated solutions to simple problems."

The data you encounter as an analyst won’t always conform to your checklist or activities list regardless of how comprehensive they are. Data cleaning can be an involved and complicated process, but surprisingly most data has similar problems. A solid personal motto and explanation can make the more common data cleaning tasks easy to understand and complete.

**Reflection**



Now that you have completed your Data Cleaning Approach Table, take a moment to reflect on the decisions you made about your data cleaning approach. Write 1-2 sentences (20-40 words) answering each of the following questions:

* What items did you add to your data cleaning checklist? Why did you decide these were important to check for?
* How have your own experiences with data cleaning affected your preferred cleaning methods? Can you think of an example where you needed to perform one of these cleaning tasks?
* How did you decide on your data cleaning motto?

Using SQL as a junior data analyst

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:

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 |

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/)

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. If you have hopped around courses, [Using BigQuery](https://www.coursera.org/learn/data-preparation/supplement/DYOQK/using-bigquery) in the **Prepare Data for Exploration** course covers how to set up a BigQuery account.

**Prepare for the next video**

* First, download the CSV file from the attachment below.

**Customer Table - Sheet1**CSV File

[Download file](https://d3c33hcgiwev3.cloudfront.net/F0iSyYcLT9iIksmHCw_Y-Q_191f150a80d74fda96f9df2aa2e3b533_Customer-Table---Sheet1.csv?Expires=1648080000&Signature=FujiSOTJ4r9VkMIhQ~xtMzMU36uxFXyt98wrA9R16rRxqg0u7jaQs5hpJNrIrBrDr7nitrirQ8~vBRkmtiJwXQy2NhCUHdHtW8-TI~2phCSzjBHNQ-QVZ76CAOo6c3SLHj3ulN~tEyNcySGYaO~244X8gOEoP~vNDUaU8XbKPkE_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

* Next, complete the following steps in your BigQuery console to upload the Customer Table dataset.

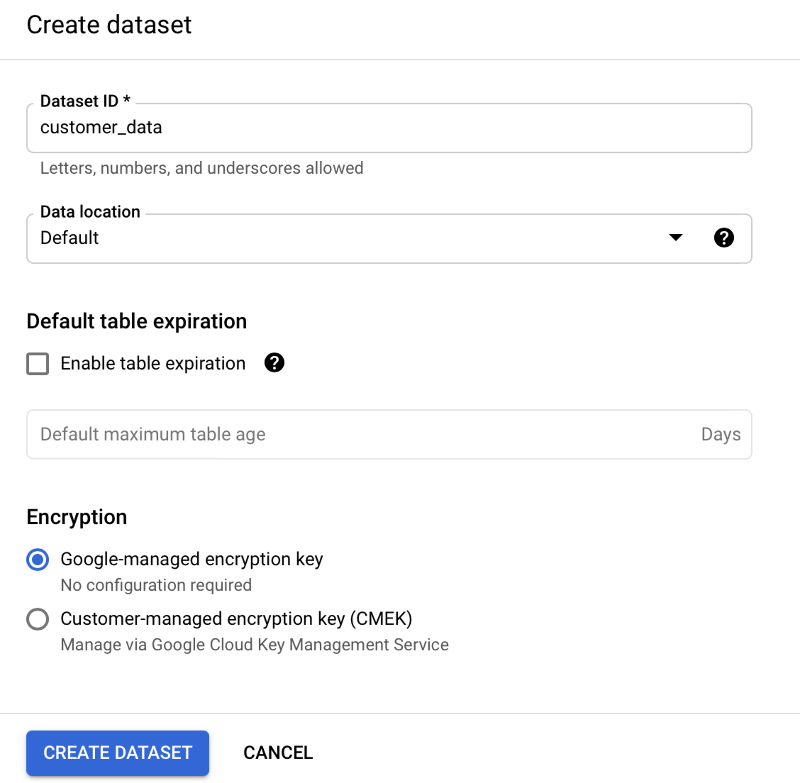
**Step 1**: Open your BigQuery console and click on the project you want to upload the data to.

**Step 2:** In the Explorer on the left, click the Actions icon (three vertical dots) next to your project name and select **Create dataset**.

Graphical user interface, application

Description automatically generated with medium confidence

**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.

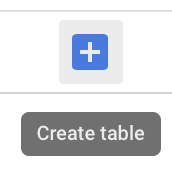


**Step 4:** Click **CREATE DATASET** (blue button) to add the dataset to your project.

**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 top right to open the Create table window.



**Step 8:** Under Source, for the Create table from selection, choose where the data will be coming from.

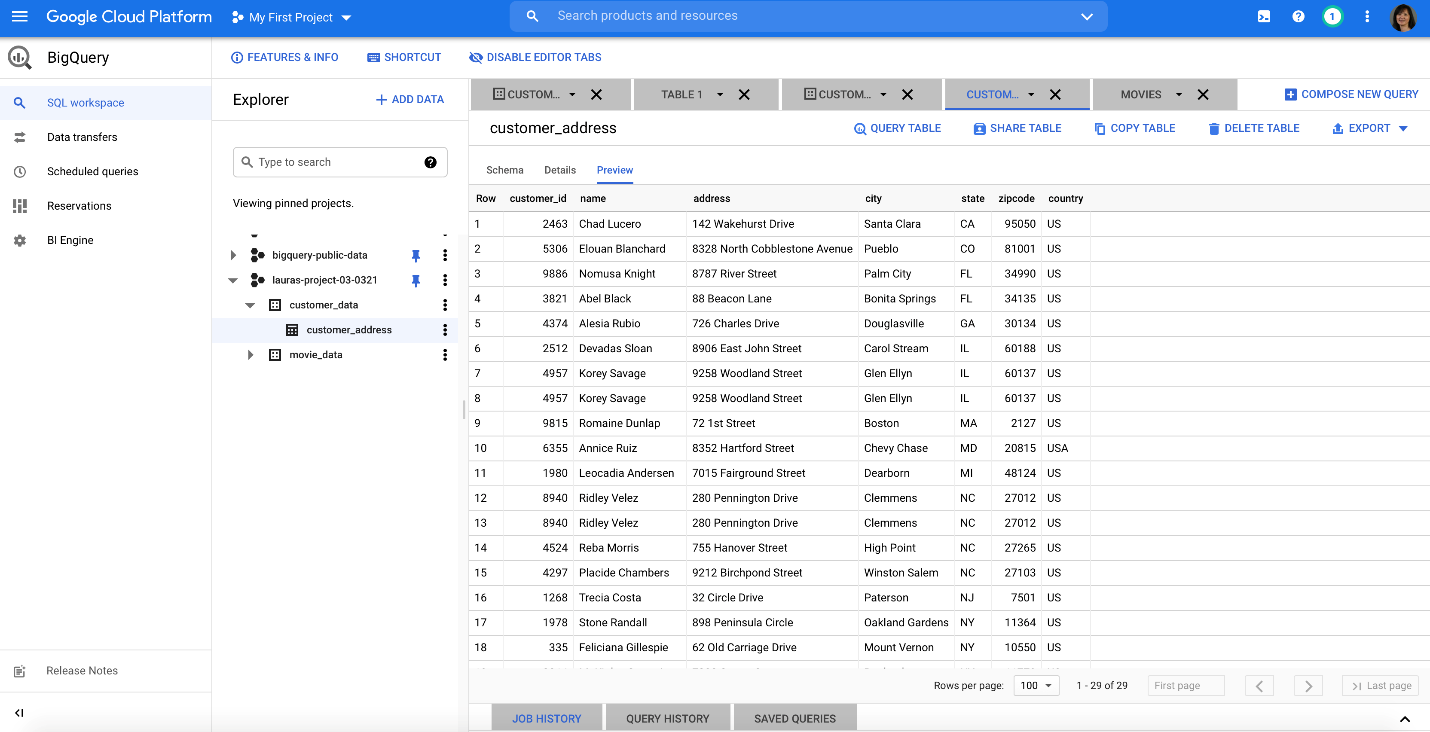
* Select **Upload**.
* Click **Browse** to select the Customer Table CSV file you downloaded.
* Choose **CSV** from the file format drop-down.

**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.

## Hands-On Activity: Clean data using SQL

**Total points**2

### 1.

**Question 1**



## Activity overview



In previous lessons, you learned about the importance of being able to clean your data where it lives. When it comes to data stored in databases, that means using SQL queries. In this activity, you will create a custom dataset and table, import a CSV file, and use SQL queries to clean automobile data.

In this scenario, you are a data analyst working with a used car dealership startup venture. The investors want you to find out which cars are most popular with customers so they can make sure to stock accordingly.

By the time you complete this activity, you will be able to clean data using SQL. This will enable you to process and analyze data in databases, which is a common task for data analysts.



### **What you will need**

To get started, download the automobile\_data CSV file. This is data from an external source that contains historical sales data on car prices and their features.

Click the link to the automobile\_data file to download it. Or you may download the CSV file directly from the attachments below.

Link to data: [automobile\_data](https://drive.google.com/u/0/uc?id=1cJtuw-6mxZk7BNkcsLYEvfjW0l_PdKxA&export=download)

OR

Download download data:

**automobile\_data**CSV File

[Download file](https://d3c33hcgiwev3.cloudfront.net/CHOsbSb3RYGzrG0m98WBiQ_f6e1579446f9464699fab3ea55ecb6f1_automobile_data.csv?Expires=1648080000&Signature=DCNKGurBzEusm6TseiaRsL3XCf5gtXBzhvmdNxBBSRpguYk9upHUHjf~gQ-3HDZIek6dlkvgmxAa-eqEqK2uZNgPbrAaC8PO~AEmEpVgUzYAPm8AriZ~-MMcI0ZKkvVzncOi4EK5k843xqv-H9EAmUyR97bO3if8wUfrM9w9lOc_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



## Upload your data



Similarly to a previous BigQuery activity, you will need to create a dataset and a custom table to house your data. Then, you’ll be able to use SQL queries to explore and clean it. Once you’ve downloaded the automobile\_data file, you can create your dataset.

### **Step 1: Create a dataset**

Go to the **Explorer pane** in your workspace and **click the three dots next to your pinned project** to open the menu. From here, **select** **Create dataset.**

Graphical user interface, application

Description automatically generated

From the Create dataset menu, fill out some information about the dataset.  **Input the Dataset ID as cars;** you can leave the Data location as Default. Then **click CREATE DATASET**.

Graphical user interface, text, application

Description automatically generated

The cars dataset should appear under your project in the Explorer pane as shown below. **Click on the three dots next to the cars dataset** to open it.

Graphical user interface, application

Description automatically generated

### **Step 2: Create table**

After you open your newly created dataset, you will be able to add a custom table for your data.

From the cars dataset, **click CREATE TABLE**.

Graphical user interface, text, email

Description automatically generated

**Under Source, upload the automobile\_data CSV**. Under Destination, make sure you are uploading into your cars dataset and **name your table car\_info.** You can **set the schema to Auto-detect**. **Then, click Create table.**

Graphical user interface, application

Description automatically generated

After creating your table, it will appear in your Explorer pane. You can **click on the table to explore the schema and preview your data.** Once you have gotten familiar with your data, you can start querying it.

## 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. According to the [data’s description](https://archive.ics.uci.edu/ml/datasets/Automobile), the **fuel\_type column** should only have **two unique string values: diesel and gas**. To check and make sure that’s true, **run the following query:**

**SELECT   DISTINCT fuel\_type FROM   cars.car\_info;**

This returns the following results:

Graphical user interface, text, application, email

Description automatically generated

This confirms that the fuel\_type column doesn’t have any unexpected values.

### **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**

**SELECT   MIN(length) AS min\_length,   MAX(length) AS max\_length FROM   cars.car\_info;**

Your results should confirm that 141.1 and 208.1 are the minimum and maximum values respectively in this column.

Table

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:**

**SELECT   \* FROM   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:

Table

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:**

**UPDATE   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:**

**SELECT   \* FROM   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:**

**SELECT   DISTINCT num\_of\_cylinders FROM   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.**

Table

Description automatically generated

To **correct the misspelling for all rows, you can run this query if you have the BigQuery free trial:**

**UPDATE   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. To **check that it worked, you can run the previous query again:** **SELECT   DISTINCT num\_of\_cylinders FROM   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:**

**SELECT   MIN(compression\_ratio) AS min\_compression\_ratio,   MAX(compression\_ratio) AS max\_compression\_ratio FROM   cars.car\_info;**

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.

**SELECT   MIN(compression\_ratio) AS min\_compression\_ratio,   MAX(compression\_ratio) AS max\_compression\_ratio FROM   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:

**SELECT**

**COUNT(\*) AS num\_of\_rows\_to\_delete**

**FROM**

**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:**

**DELETE 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 DISTINCT statement:**

**SELECT   DISTINCT drive\_wheels FROM   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.

Table

Description automatically generated

To check if this is the case, you can **use a LENGTH statement** to determine the length of how long each of these string variables:

**SELECT   DISTINCT drive\_wheels,   LENGTH(drive\_wheels) AS string\_length FROM   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**:

**UPDATE**

**cars.car\_info**

**SET**

**drive\_wheels = TRIM(drive\_wheels)**

**WHERE TRUE;**

Then, you **run the SELECT DISTINCT statement again** to ensure that there are only three distinct values in the drive\_wheels column:

**SELECT   DISTINCT drive\_wheels FROM   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!

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. If you have hopped around courses, [Using BigQuery](https://www.coursera.org/learn/data-preparation/supplement/DYOQK/using-bigquery) in the **Prepare Data for Exploration** course covers how to set up a BigQuery account.

**Prepare for the next video**

* First, download the CSV file from the attachment below.

**Lauren's Furniture Store Transaction Table**CSV File

[Download file](https://d3c33hcgiwev3.cloudfront.net/Hhbf0PXYQVmW39D12OFZ8A_ef465a1628a943b98ba9d0e7696877f1_Lauren-s-Furniture-Store-Transaction-Table.csv?Expires=1648080000&Signature=MNgBtfZ7OMTRl8KeMHqyBokNKlrk-3y5SeZpzc6RIolumz2dt1gaRbRiIU6pdAR9u-etCFndSOi6nVG9phHbJ~GsD-pIvi4H2IkLB2CvagQAlwTVne3oKPbEj-G9~4PeimBlXsnS8qggK~obORb9TeHlZ6sRgW1cwYcRqsM-rkc_&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.

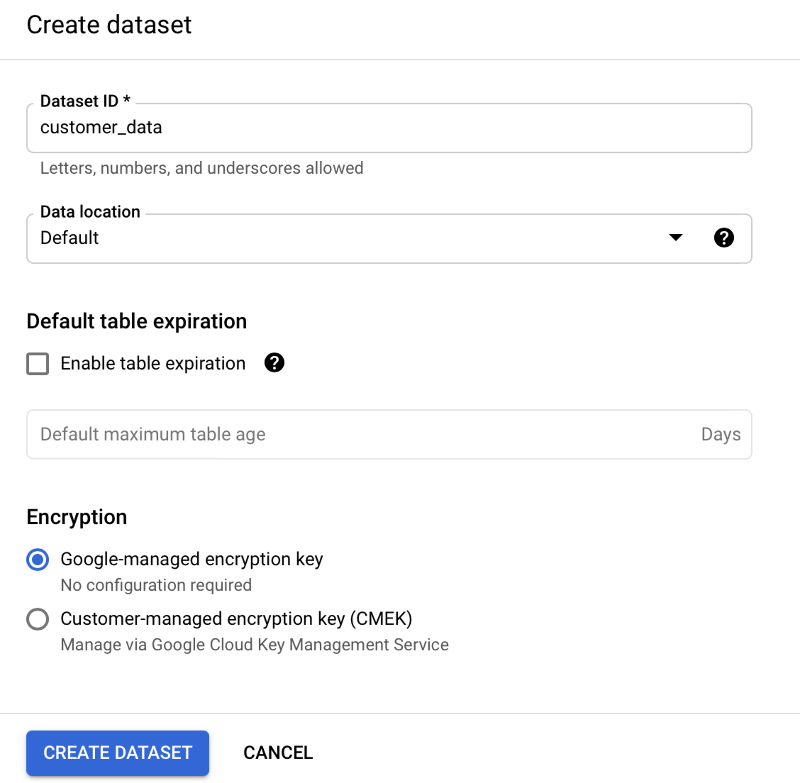
**Step 1**: Open your BigQuery console and click on the project you want to upload the data to. If you already created a **customer\_data** dataset for your project, jump to step 5; otherwise, continue with step 2.

**Step 2:** In the Explorer on the left, click the Actions icon (three vertical dots) next to your project name and select **Create dataset**.

Graphical user interface, application

Description automatically generated with medium confidence

**Step 3:** Enter **customer\_data** for the Dataset ID.

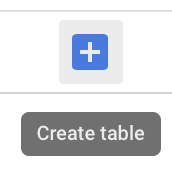


**Step 4:** Click **CREATE DATASET** (blue button) to add the dataset to your project.

**Step 5:** In the Explorer, click to expand your project, and then click the **customer\_data** dataset.

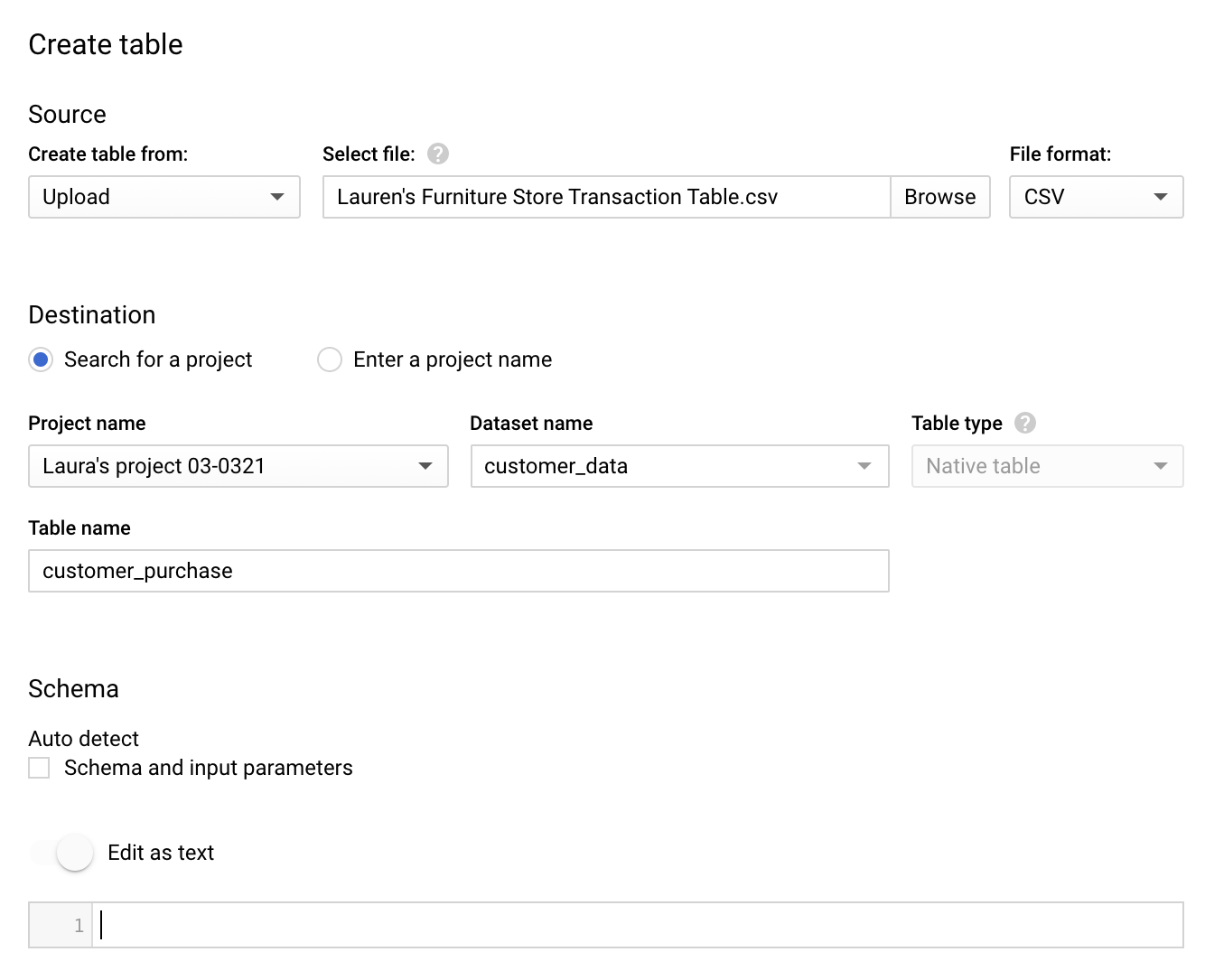
**Step 6:** Click the Actions icon (three vertical dots) next to customer\_data and select **Open**.

**Step 7:** Click the blue **+** icon at the top right to open the Create table window.



**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 Store Transaction Table CSV file you downloaded.
* Choose **CSV** from the file format drop-down.



**Step 9:** For Table name, enter **customer\_purchase** if you plan to follow along with the video.

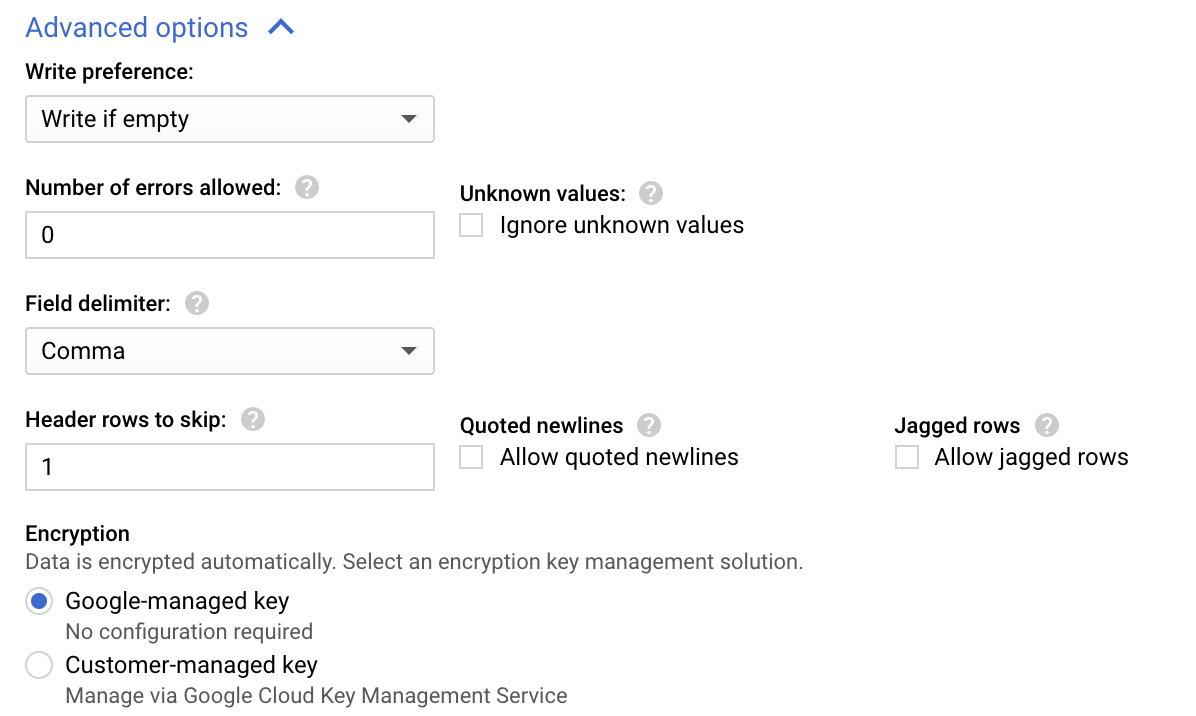
**Step 10:** For Schema, click the toggle switch for **Edit as text**. This opens up a box for the text.

**Step 11:** 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 12:** Scroll down and expand the **Advanced options** section.

**Step 13:** For the **Header rows to skip** field, enter **1**.



**Step 14:** Click **Create** **table** (blue button). You will now see the **customer\_purchase** table under your **customer\_data** dataset in your project.

**Step 15:** Click the **customer\_purchase** table and in the **Schema** tab, confirm that the schema matches the schema shown below.

Table

Description automatically generated

**Step 16:** Click the **Preview** tab and confirm that your data matches the data shown below.

Table

Description automatically generated

Congratulations, you are now ready to follow along with the video!

## Self-Reflection: Challenges with SQL

**Total points**1

### 1.

**Question 1**



## Overview



Now that you have practiced writing SQL functions, you can pause for a moment and think about what you are learning. In this self-reflection, you will consider your thoughts about your experience with learning SQL and respond to brief questions.

This self-reflection will help you develop insights into your own learning and prepare you to identify your successes and difficulties with learning SQL so you can understand how to develop your skills further. As you answer questions—and come up with questions of your own—you will consider concepts, practices, and principles to help refine your understanding and reinforce your learning. You’ve done the hard work, so make sure to get the most out of it: This reflection will help your knowledge stick!

## 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.

Data-cleaning verification: A 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**

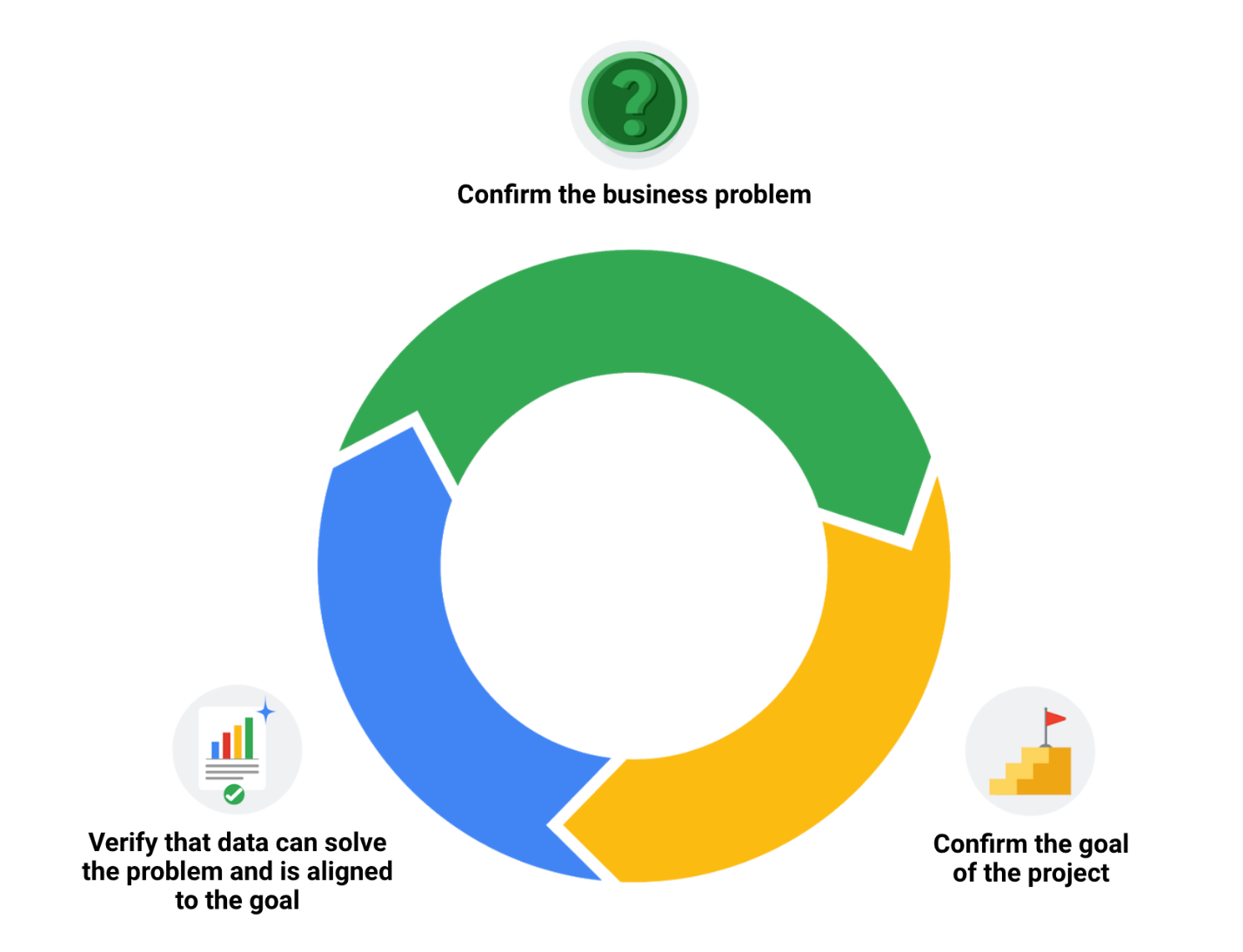
Make sure you identified the most common problems and corrected them, including:

* **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:

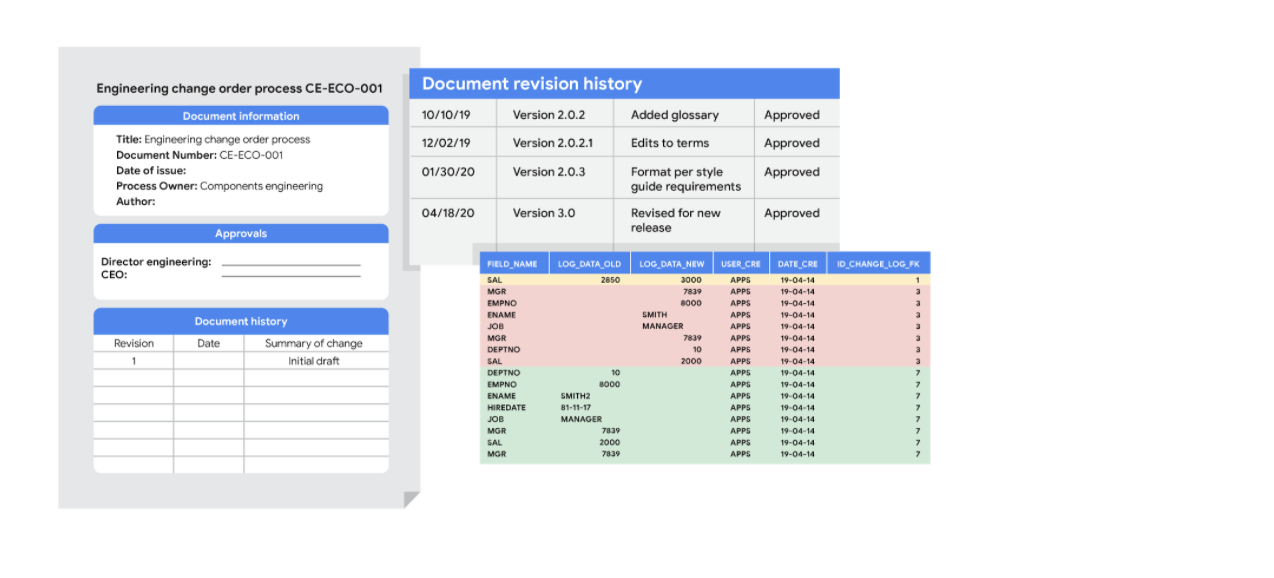
* Confirm the business problem
* Confirm the goal of the project
* Verify that data can solve the problem and is aligned to the goal



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 this type of information:

* 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.

**What also happens IRL (in real life)**



A junior analyst probably only needs to know the above with one exception. 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 is 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.

## Self-Reflection: Creating a changelog

**Total points**1

### 1.

**Question 1**



## Overview



Now that you have learned about the importance of keeping track of changes in your data analysis, you can pause for a moment and track what you are learning. In this self-reflection, you will consider your thoughts about changelogs and respond to brief questions.

This self-reflection will help you develop insights into your own learning and prepare you to incorporate changelogs into your data cleanings procedures. As you answer questions—and come up with questions of your own—you will consider concepts, practices, and principles to help refine your understanding and reinforce your learning. You’ve done the hard work, so make sure to get the most out of it: This reflection will help your knowledge stick!

## The importance of changelogs



In previous activities, you’ve reviewed the different types of questions to ask before exploring data, the importance of pre-cleaning data, the basic functions of SQL, how to clean data with spreadsheets, and more. As a junior data analyst, most of your projects will consist of these activities. As you have experienced, each of these tasks follows a complicated process. Therefore, consistent and accurate record-keeping is essential to keeping you on track.

A **changelog** is a document used to record the notable changes made to a project over its lifetime across all of its tasks. It is typically curated so that the changes it records are listed chronologically across all versions of the project.

The major benefit to using changelogs is that contributors and users connected with the project get a specific list of what important alterations have been made, when they were made, and sometimes, what version they were released for. It is an invaluable tool for communicating how the project has evolved over time to coworkers, management, and stakeholders.

### **Best practices for changelogs**



A changelog for a personal project may take any form desired. However, in a professional setting and while collaborating with others, readability is important. These guiding principles help to make a changelog accessible to others:

* Changelogs are for humans, not machines, so write legibly.
* Every version should have its own entry.
* Each change should have its own line.
* Group the same types of changes. For example, Fixed should be grouped separately from Added.
* Versions should be ordered chronologically starting with the latest.
* The release date of each version should be noted.

All the changes for each category should be grouped together. Types of changes usually fall into one of the following categories:

* Added: new features introduced
* Changed: changes in existing functionality
* Deprecated: features about to be removed
* Removed: features that have been removed
* Fixed: bug fixes
* Security: lowering vulnerabilities

## Examine a sample changelog



Examine the figure below for an example of a changelog. Note that the following example is written in [Markdown](https://docs.github.com/en/free-pro-team@latest/github/writing-on-github/basic-writing-and-formatting-syntax), as it is common to keep changelogs as a readme file in a code repository.



1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

# Changelog

This file contains the notable changes to the project

Version 1.0.0 (02-23-2019)

## New

    - Added column classifiers (Date, Time, PerUnitCost, TotalCost, etc. )

    - Added Column “AveCost” to track average item cost

## Changes

    - Changed date format to MM-DD-YYYY

    - Removal of whitespace (cosmetic)

## Fixes

    - Fixed misalignment in Column "TotalCost" where some rows did not match with correct dates

    - Fixed SUM to run over entire column instead of partial







## What to record in a changelog



Now that you're familiar with the example, consider what changes you need to record in a changelog. To start, you record the various changes, additions, and fixes that were discussed above. Arrange them using bullets or numbering with one change per line. Group similar changes together with a label describing the change immediately above them.

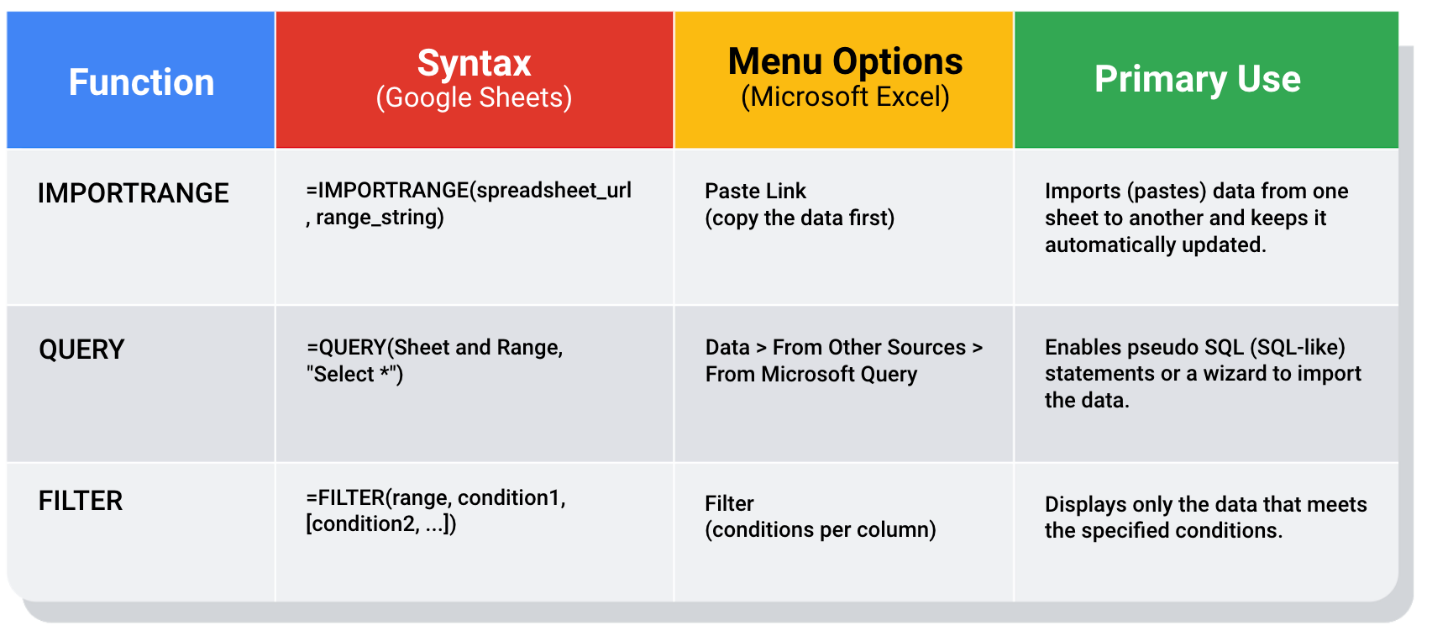
Use different version numbers for each milestone reached in your project. Within each version, place the logged changes that were made since the previous version (milestone). Dates are not generally necessary for each change, but they are recommended for each version.

In an upcoming course, you will have the opportunity to complete a capstone project. This will be a great chance to demonstrate your ability to organize a project like a professional data analyst by keeping your own changelog.

You can do this using a simple text file or spreadsheet and include your changelog with the project write-up. It will help you stay organized and collaborate with others. Keep this in mind when you reach the capstone project in an upcoming course, and don’t be afraid to revisit this lesson if you have questions.

# Advanced functions for speedy data cleaning

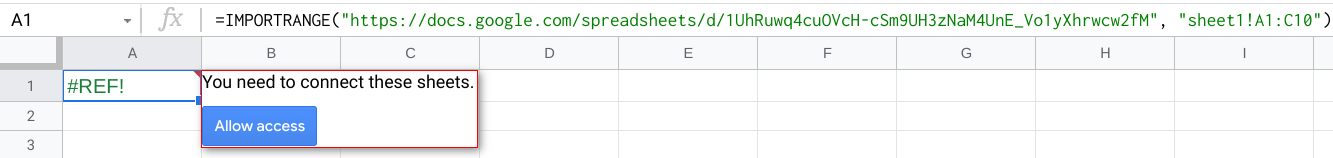
In this reading, you will learn about some advanced functions that can help you speed up the data cleaning process in spreadsheets. Below is a table summarizing three functions and what they do:

IMPORTRANGE: Syntax: =IMPORTRANGE(spreadsheet\_url, range\_string) Menu Options: Paste Link (copy the data first) Primary Use: Imports (pastes) data from one sheet to another and keeps it automatically updated QUERY: Syntax: =QUERY(Sheet and Range, "Select \*") Menu Options: Data > From Other Sources > From Microsoft Query Primary Use: Enables pseudo SQL (SQL-like) statements or a wizard to import the data. FILTER: Syntax: =FILTER(range, condition1, [condition2, ...]) Menu Options: Filter(conditions per column) Primary Use: Displays only the data that meets the specified conditions.

## 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.

If you are using IMPORTRANGE in Google sheets, data can be pulled from another spreadsheet, but you must allow access to the spreadsheet the first time it pulls 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/1cOsHnBDzm9tBb8Hk\_aLYfq3-o5FZ6DguPYRJ57992\_Y”, “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/1cOsHnBDzm9tBb8Hk\_aLYfq3-o5FZ6DguPYRJ57992\_Y”, “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)

### **Real life 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)

**Hands-On Activity: Build a resume**

**Total points**2

**1.**

**Question 1**



Activity Overview



Earlier, you learned about what makes an effective resume. In this activity, you’ll begin building your resume or work on your existing one.

By the time you complete this activity, you’ll have a stronger understanding of common resume formats and decide on a template for your data analytics resume that you’ll complete later. This is an important part of the job application process: A strong resume is essential to moving forward as a data analytics professional.



What you will need

To use the templates for this course item, click the links below and select “Use Template.”

Link to template 1: [Template Example 1](https://docs.google.com/document/d/1qn_zOg-0E7pca6bEk6BGIEBNBuIZdiPnwOTKm76Q-jA/template/preview)

Link to template 2: [Template Example 2](https://docs.google.com/document/d/1l-aMPMNRxZ0zSOQcNGg4jMqQO5FW1coZiV2m7jz6CFw/template/preview)

OR

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

**Resume Template 1**DOCX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/ufNix5roSVGzYsea6JlR4A_68e010ef3f8f4d8d92978d8c1cbf7184_Resume-Template-1.docx?Expires=1648080000&Signature=ibvnVSBZSu-qh6ynQWAHaikUPzg5-Ymu7mMB8rbzBpogNIm7OfnAc7RLF-kgKOB0Kr9GwoQMfuF4i~auLTSi7j8brlk4x4vz~YQQz2ht35bOo7km08JeL2ECZby2d2aNKvZIEfuyZLLUR6wNFyFsWztKW5Xzo80f2VocQWZr67Y_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**Resume Template 2**DOCX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/S0mDPfhiSoaJgz34YvqGxA_48ad285cfea74858a876be2ef64c4a49_Resume-Template-2.docx?Expires=1648080000&Signature=Vtlrwcv-da48VIy2xP8Tqs47GzAdHy~0At8gWm8Ks6vVToTVGeGnJV9bl3ntlvD850sfvpZ9kkPZOdZcgRzvkY7zL3wT0UlgXtH5ruFktvjxcT2ps6kYzglYPj6lGZkMQw19XXEnSglU9LwlIDTpGc7Ealo-xlpJ7EEMt3mb4Vk_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



Format your resume



First, you’ll make some decisions about the structure and layout of your resume. You’ll start by examining the two templates and decide which format you like best. Spend some time browsing the templates, as well as templates available to you on the web. Take note of the things you like or dislike about the various formats. It’s okay to take some time and be selective here. You’ll be spending a lot of time on this document, so picking a format you’re happy with will help you get off to a strong start.

Formats and templates



Before creating your resume, you need to make some design decisions. While you may make some small tweaks and changes to tailor the content of your resume to specific roles you are applying for, the structure and format of the resume likely won’t change. This means it is important to spend time thinking about how you’ll want to structure your resume.

Keep your resume format concise



There’s no “best” format for a resume. Instead, think about what you want to highlight about yourself to potential employers.

For instance, if you have relevant work experience, then pick a format to highlight that.

If you are transitioning from a different career and don’t yet have relevant work experience, then you may want to pick a format that highlights your technical skills and portfolio projects. Some resume formats include a **Summary** or **Goals** section at the top to help candidates add context to their application, while other resume formats avoid these sections completely and save that space for sections such as **Skills** and **Experience**.

Whatever format you pick, make sure to follow the one-page rule and keep the completed version on just a single page. If the one-page rule seems limiting, think about the purpose resumes serve in the hiring process overall. Resumes are short documents designed to communicate the most pertinent information about yourself to recruiters and hiring managers at a glance. These are different from longer, multi-page Curriculum Vitaes (CVs) that exhaustively list every relevant thing the candidate has ever done.

If an employer wants a detailed history of your past work experiences and accolades, they might specifically request a CV (curriculum vitae) instead. If they don’t, always assume they prefer a resume. While it is generally considered acceptable for resumes of applicants with extensive work history applying for senior technical roles to have two-page resumes, these are the exception rather than the rule. When applying for a data analyst position, keep it to one page!

Select a format



Once you decide on a template, resist the urge to begin filling it out. The next lessons will focus on best practices for communicating your skills and experience in meaningful, impactful ways. Keep this resume template handy, as you’ll be working on it further.

# 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.

### **I​mportant 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.

## Hands-On Activity: Adding skills to a resume

**Total points**2

### 1.

**Question 1**



## Activity Overview



In the last activity, you chose a format for your resume. Now, you will work on your resume by adding information about you, as well as the job-ready skills you’ve developed in this program!

By the time you complete this activity, you will be able to build a document that describes your skills, experience, and achievements. This is very important for applying for jobs as a data analyst.



### What you will need

To get started, open the resume template that you chose in the previous activity. If you can’t find the template you chose, you can access the resume templates below.

To use the templates for this course item, click the links below and select “Use Template.”

Link to resume template 1: [Resume Template 1](https://docs.google.com/document/d/1qn_zOg-0E7pca6bEk6BGIEBNBuIZdiPnwOTKm76Q-jA/template/preview)

Link to resume template 2: [Resume Template 2](https://docs.google.com/document/d/1l-aMPMNRxZ0zSOQcNGg4jMqQO5FW1coZiV2m7jz6CFw/template/preview)

OR

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

**Resume Template 1**DOCX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/CPOyY0iTSByzsmNIk7gccQ_8f9d6fe0c2a346a0b655b42479ae3a32_Resume-Template-1.docx?Expires=1648080000&Signature=LlqqtLSNasbsley4sKqcsr1I2fPDUAkbcuTqTIh7fWAvrHgddb1AGpE~dQA7W21N67X7irI1GqiTcryiiZCBaUL1pNuhzZwVdLM0ytA576LGQWepbNyx-Z7m0-QoSw5yoA5f0RceNS2RZJh8PGLkmlMaWoGn~P4HeG2f1tUQK7o_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**Resume Template 2**DOCX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/y_9YlhGCQEq_WJYRglBKQQ_2295c95c90554ac8be26ed0fa0e55a5c_Resume-Template-2.docx?Expires=1648080000&Signature=kgsNWMolOFvy31Kx9LIkgpydPnEyQCgXwBZsLqDn5TQALdQQ6RlL~cqkuZDXbJ4e4otFW2~Rr9~-bhroKdosM~BMZS8Noi~2QyR4SVB7uROoKx3NMiCpDPHotPh4s-MGdqFqe5WKnsBpEhDGMCwAi-Yr2cVhDTf6d3Jfh-5Mzso_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



## Add skills to your resume



Data analysts are expected to have strong technical skills and abilities, so effectively highlighting those skills is a crucial part of crafting your resume. Focus on your skills for this activity. Don’t worry about adding your work experience just yet; you’ll complete that in a future activity.

### Get help from the real world



Reviewing real-world resumes is always a great idea. It can help you get a feel for how others in the industry are representing their experience and skills. You can find resumes on job sites and LinkedIn or even just by searching for “data analyst resume.”  There are many ways to represent your technical skills, and taking a moment to understand how other data analysts do this may give you some great ideas!

### What skills to add



The skills section on your resume likely only has room for 2-4 bullet points, so be sure to use this space effectively. You might want to prioritize technical skills over soft skills. This is a great chance for you to highlight some of the skills you’ve picked up in these courses, such as:

* Strong analytical skills
* Pattern recognition
* Relational databases and SQL
* Strong data visualization skills
* Proficiency with spreadsheets, SQL, R, and Tableau

Notice how the skills listed above communicate a well-rounded data analyst’s skill set without being wordy. The skills section summarizes what you’re capable of doing while listing the technology and tools you are proficient in.

Many companies use algorithms to screen and filter resumes for keywords. If your resume does not contain the keywords they are searching for, a human may never even read your resume. Reserving at least one bullet point to list specific programs you are familiar with is a great way to make sure your resume makes it past automated keyword screenings and onto the desk of a recruiter or hiring manager.

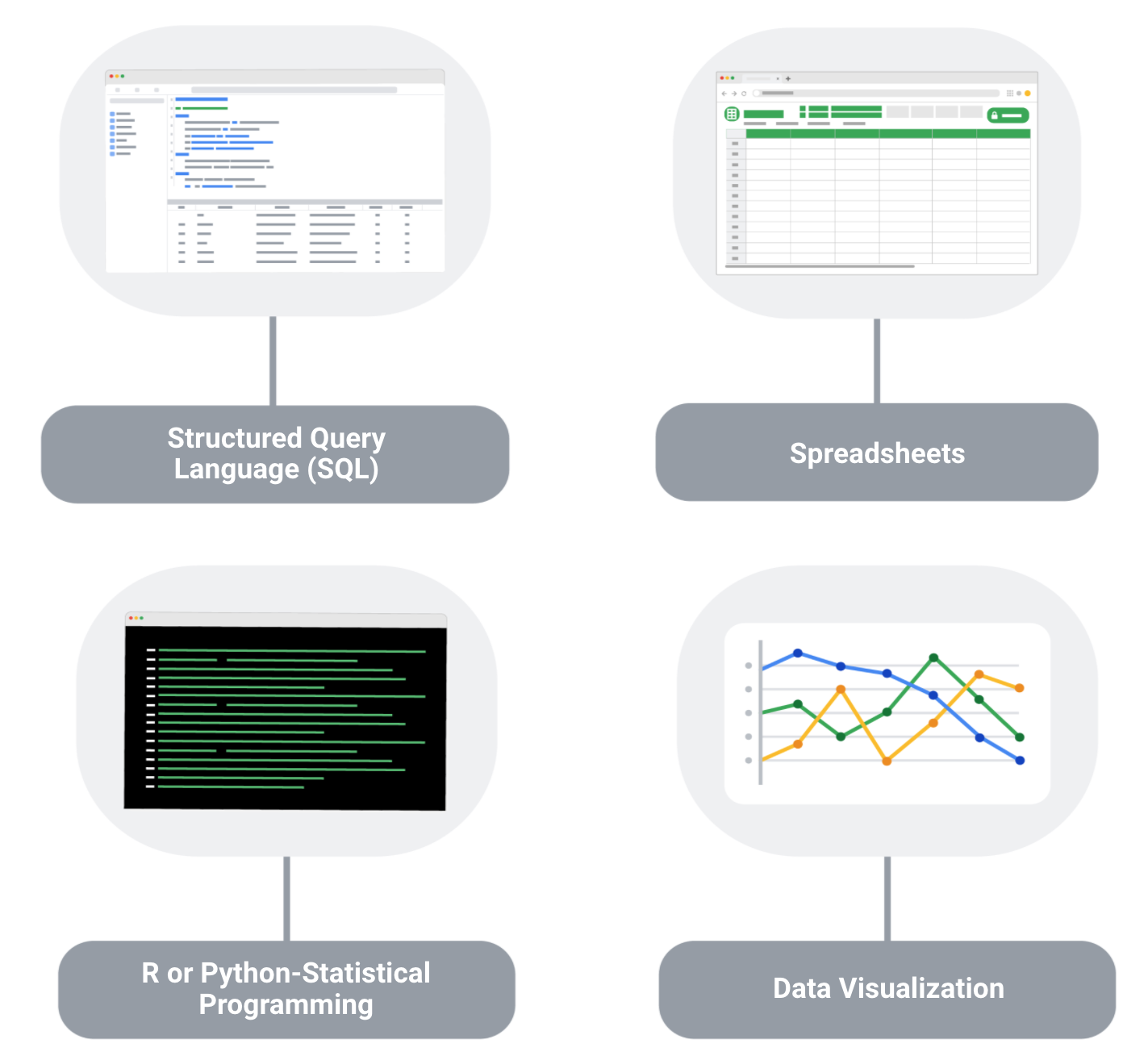
Take a moment to complete the skills section of your resume.

# Adding professional skills to your resume

Technical skills are crucial when building a solid resume. They demonstrate to employers that you have the professional skills necessary to successfully perform the job. Think of professional 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 professional 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.

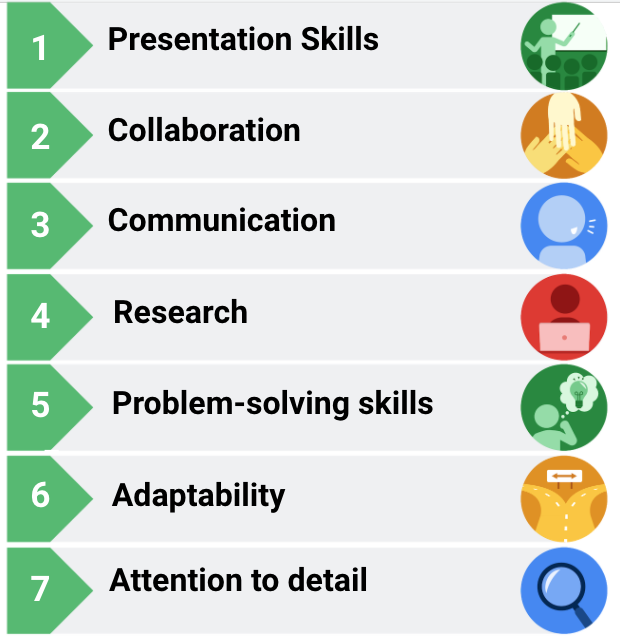
**4**. **R or Python programming:** Since only less than a third of entry-level data analyst positions require knowledge of Python or R, you don’t need to be proficient in programming languages as an entry-level data analyst. But, R or Python are great additions to have as you become more advanced in your career.

## Key takeaway

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.

# Adding soft skills to your resume

There is more than just data when it comes to being a data analyst—there are plenty of soft 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 soft 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 soft skills to your resume

Here are a few ways that you can add soft skills to your resume:

1. Analyze your previous work experience and find opportunities to insert a soft 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 soft and professional skills in the skills or summary section of your resume.

**Hands-On Activity: Adding experience to a resume**

**Total points**2

**1.**

**Question 1**



Activity overview



In the last activity, you added skills to your resume. In this activity, you will work on your resume by adding your experience.

By the time you complete this activity, you will understand how to frame your work history and experience to fit your resume. This will enable you to build a document that effectively describes your skills, experience, and achievements. This is important for applying for jobs as a data analyst.



What you will need

To get started, open up the resume template that you chose in the previous activity. If you can’t find the template you chose, you can access the resume templates below.

First, access the templates (if needed) as well as the Work Experience Bullet Example.

To use the templates for this course item, click the links below and select “Use Template.”

Link to resume template 1: [Resume Template 1](https://docs.google.com/document/d/1qn_zOg-0E7pca6bEk6BGIEBNBuIZdiPnwOTKm76Q-jA/template/preview)

Link to resume template 2: [Resume Template 2](https://docs.google.com/document/d/1l-aMPMNRxZ0zSOQcNGg4jMqQO5FW1coZiV2m7jz6CFw/template/preview)

Link to example document: [Work Experience Bullet Example](https://docs.google.com/document/d/1fea0Yzldlr1eMf5eZ9aWRTeeVP0-ZS5h7DMqXt__6Mg/template/preview)

OR

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

**Resume Template 1**DOCX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/hPHYeJ8qQcWx2HifKrHFyA_aec69fce611c44a6be37229b3a6d1267_Resume-Template-1.docx?Expires=1648080000&Signature=N24QXo3aiIrcJIN2Q2SZJo5TBC2dTYJDIdLIqQbiKiknMaga6-Q4qtvOWdty~v8XOtMBltvjAuzew94z6eP4bTjB47vN7Da7Afvu6nQIhu6DSMLoo2dufMJmx9yJAsbWJVYmdJxMo~9aUgGEq8KZCslRxfgTaWV1LsbUe74Y8EQ_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**Resume Template 2**DOCX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/4kkrF6aORiqJKxemjkYqPg_ec81a42c2c654a16900a100866a4a4f3_Resume-Template-2.docx?Expires=1648080000&Signature=D9j5HlEw~2-TqodrAlbeCvTVok0zK9Poa3FmalVgWupfBUAE2rNz3TgrTUUhxhPxWMGj0qSSgNuFeaq0WNI~BrHpAPR98I8WxO0-CzD~Q8VPYui3kLIs7gBqCl~CcreAa2XjxI8CdJFG2Nw8Ls2KfOaYkvM9vJLCxTIzybXAL30_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**Resume Template Example with Experience**DOCX File

[Download file](https://d3c33hcgiwev3.cloudfront.net/4eUfT3k5SFKlH095OahSPA_0a16d65999de4b55b84f7862ed46e7ef_Resume-Template-Example-with-Experience.docx?Expires=1648080000&Signature=Tvlf4XmOTkN5hWOW28dx~uJ5u2KpP-nPOWfqkZsecCSlab-qh5vTcxirMIqwWEUHUMdWWiGvlVL4OyH0ga9PLKeheTOTJOt0UPLNC-vcuz9ZIfF-3PxBNb6eDCL~7L4s8gdadncKq0tjZytPexwkc7GSYEFTTReY52z4zTLlf-4_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



Best practices for writing about experience



One of the most important functions of a resume is communicating your prior work experience in a favorable light. This can often be challenging, as the one-page format forces job seekers to summarize all of their work experience into a few bullet points.

Resume best practices will help you select the most relevant parts of your work experience and communicate them in the shortest, most impactful way possible.

As you think about how to represent your work experience on your resume effectively, it might be helpful to refer to these best practices:

Focus on your accomplishments first, and explain them using the formula **“Accomplished X, as measured by Y, by doing Z.”**

* These statements help you communicate the most important things a recruiter or hiring manager is searching for—the impact of your work.
* Whenever possible, use numbers to explain your accomplishments. For example, “Increased manufacturing productivity by 15% by improving shop floor employee engagement,” is better than “Increased manufacturing productivity.”

Phrase your work experience and duties using **Problem-Action-Result** (PAR) statements.

* For example, instead of saying “was responsible for two blogs a month,” phrase it as “earned little-known website over 2,000 new clicks through strategic blogging.”

Describe jobs that highlight **transferable skills** (those skills that can transfer from one job or industry to another).

* This is especially important if you are transitioning from another industry into data analytics.
* For example, communication is a skill often used in job descriptions for data analysts, so highlight examples from your work experience that demonstrate your ability to communicate effectively.

Describe jobs that highlight your **soft skills**.

* These are non-technical traits and behaviors that relate to how you work.
* Are you detail-oriented? Do you have grit and perseverance? Are you a strong critical thinker? Do you have leadership skills?
* For instance, you could give an example of when you demonstrated leadership on the job.
* Showing is always more effective than telling.

This is almost always the hardest part of crafting a resume, especially if you are transitioning from a different career field. However, if you take a moment to think deeply about your previous work experience, you’ll likely discover that you can find ways to represent your work experiences in a way that highlights your ability to do things important to data analyst roles, such as thinking critically or making data-driven decisions.

Get the interview first



Remember that the goal of a resume is to get you an interview. You may find that you need to brainstorm and carefully edit your resume to effectively summarize your background. In the end, you will have all of your various responsibilities and accomplishments from previous jobs synthesized into a few bullet points. That way, your resume will highlight what potential employers like to know about applicants.

Effective resumes communicate that you are a candidate who understands the needs of the role and you have the skills and experience to warrant an interview. During interviews, you can expect questions about your experience and that’s when you can go into more detail.

Add your work experience



Now that you have had some time to think about your work experience, add it to your resume. Keep in mind the best practices we shared above and that creating a resume is a process. You’ll likely come back to work on it multiple times and change phrasing or formatting. With effort and time, you’ll eventually get your resume to a place where you’re satisfied with the final result.

Sample experience description



Earlier, you downloaded a template of bullet points describing work experience. Refer to this as you write about your own work experience. Notice how it demonstrates factual, measurable successes and job experience that is applicable to the role of a data analyst in a short, concise manner.

If you need inspiration or want to see how other data analysts have structured their resumes, take a moment to search for resumes of real data analysts. You can find these easily on sites such as LinkedIn. Seeing how other data analysts have structured and worded their own resumes may give you valuable insight about more effective ways to highlight your own experience.