**MODULE 2**

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

### **LEARNING OBJECTIVES**

* Differentiate between clean and dirty data
* Explain the characteristics of dirty data
* Describe data cleaning techniques with reference to identifying errors, redundancy, compatibility and continuous monitoring
* Identify common pitfalls when cleaning data
* Demonstrate an understanding of the use of spreadsheets to clean data

**DATA CLEANING IS A MUST**

[**CLEAN IT UP!**](https://www.coursera.org/learn/process-data/lecture/nPwkO/clean-it-up)

Can you guess what inaccurate or bad data costs businesses every year? Thousands of dollars? Millions? Billions? Well, according to IBM, the yearly cost of poor-quality data is $3.1 trillion in the US alone. That's a lot of zeros.

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

It's not a new system implementation or a computer technical glitch. The most common factor is actually **human error**. Here's a spreadsheet from a law office. It shows customers, the legal services they bought, the service order number, how much they paid, and the payment method.



Dirty data can be the result of :

1. Someone typing in a piece of data incorrectly;
2. Inconsistent formatting;
3. Blank fields;
4. The same piece of data is entered more than once, which creates duplicates.

Dirty data is data that's incomplete, incorrect, or irrelevant to the problem you're trying to solve.

When you work with dirty data, you can't be sure that your results are correct. In fact, you can pretty much bet they won't be. Earlier, you learned that **data integrity is critical to reliable data analytics results, and clean data helps you achieve data integrity**. **Clean data is data that's complete, correct, and relevant to the problem you're trying to solve**. When you work with clean data, you'll find that your projects go much more smoothly.

I remember the first time I witnessed firsthand how important clean data really is. I had just started using SQL, and I thought it worked like magic. I could have the computer sum up millions of numbers, saving me tons of time and effort. But I quickly discovered that it only works when the data is clean. If there was even one accidental letter in a column that should only have numbers, the computer wouldn't know what to do. So, it would throw an error, and suddenly, I'm stuck. And there's no way I could add up millions of numbers by myself. So, I have to clean out that data to make it work. The good news is that there's plenty of effective processes and tools to help you do that. Coming up, you'll gain the skills and knowledge you need to make sure the data you work with is always clean.

Along the way, we'll dig deeper into the difference between clean and dirty data, and why clean data is so important. We'll also talk about different ways to clean your data and common problems to look for during the process.

[**WHY DATA CLEANING IS CRITICAL**](https://www.coursera.org/learn/process-data/lecture/ybXpU/why-data-cleaning-is-critical)

Clean data is incredibly important for effective analysis. If a piece of data is entered into a spreadsheet or database incorrectly, or if it's repeated, or if a field is left blank, or if data formats are inconsistent, the result is dirty data. Small mistakes can lead to big consequences in the long run.

I'll be completely honest with you, **data cleaning is like brushing your teeth**. **It's something you should do and do properly because otherwise it can cause serious problems**. For teeth, that might be cavities or gum disease.

For data, that might be costing your company money, or an angry boss. But here's the good news. If you keep brushing twice a day, every day, it becomes a habit. Soon, you don't even have to think about it. It's the same with data. Trust me, it will make you look great when you take the time to clean up that dirty data.

As a quick refresher, **dirty data is incomplete, incorrect, or irrelevant to the problem you're trying to solve**. It can't be used in a meaningful way, which makes analysis very difficult, if not impossible.

On the other hand, **clean data is complete, correct, and relevant to the problem you're trying to solve**. This allows you to understand and analyze information and identify important patterns, connect related information, and draw useful conclusions. Then you can apply what you learn to make effective decisions.

In some cases, you won't have to do a lot of work to clean data. For example, when you use internal data that's been verified and cared for by your company's data engineers and data warehouse team, it's more likely to be clean.

Let's talk about some people you'll work with as a data analyst. **Data engineers transform data into a useful format for analysis and give it a reliable infrastructure**. This means they develop, maintain, and test databases, data processors and related systems.

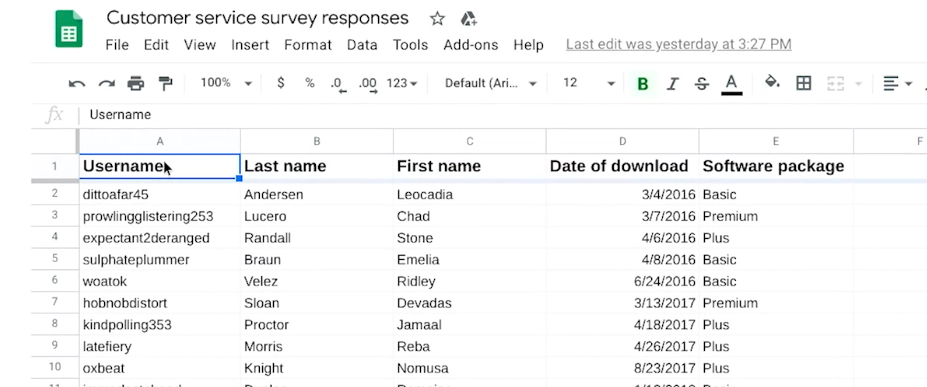
**Data warehousing specialists develop processes and procedures to effectively store and organize data. They make sure that data is available, secure, and backed up to prevent loss**.

When you become a data analyst, you can learn a lot by working with the person who maintains your databases to learn about their systems. If data passes through the hands of a data engineer or a data warehousing specialist first, you know you're off to a good start on your project.

**There's a lot of great career opportunities as a data engineer or a data warehousing specialist**. If this kind of work sounds interesting to you, maybe your career path will involve helping organizations save lots of time, effort, and money by making sure their data is sparkling clean.

But even if you go in a different direction with your data analytics career and have the advantage of working with data engineers and warehousing specialists, you're still likely to have to clean your own data. **It's important to remember: no dataset is perfect**. **It's always a good idea to examine and clean data before beginning analysis**.

Here's an example. Let's say you're working on a project where you need to figure out how many people use your company's software program. You have a spreadsheet that was created internally and verified by a data engineer and a data warehousing specialist. Check out the column labeled "Username." It might seem logical that you can just scroll down and count the rows to figure out how many users you have.



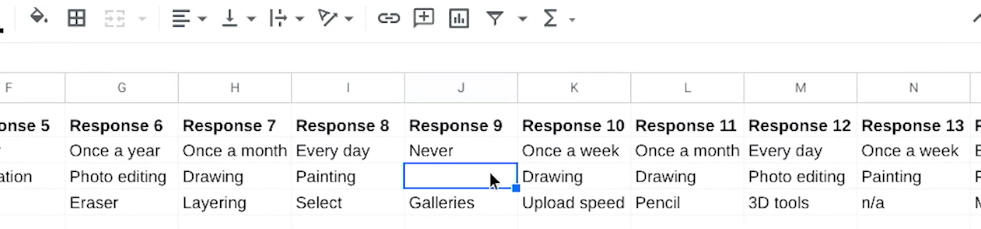
But that won't work because one person sometimes has more than one username.

Maybe they registered from different email addresses, or maybe they have a work and personal account. In situations like this, you would need to clean the data by eliminating any rows that are duplicates.

Once you've done that, there won't be any more duplicate entries. Then your spreadsheet is ready to be put to work. So far we've discussed working with internal data.

But data cleaning becomes even more important when working with external data, especially if it comes from multiple sources.

Let's say the software company from our example surveyed its customers to learn how satisfied they are with its software product. But when you review the survey data, you find that you have several nulls.



**A null is an indication that a value does not exist in a data set. Note that it's not the same as a zero.** In the case of a survey, a null would mean the customers skipped that question. A zero would mean they provided zero as their response.

To do your analysis, you would first need to clean this data. **Step one** would be to decide what to do with those nulls. You could either **filter them out and communicate that you now have a smaller sample size**, or **you can keep them in and learn from the fact that the customers did not provide responses**.

There's lots of **reasons why this could have happened**. Maybe your survey questions weren't written as well as they could be. Maybe they were confusing or biased, something we learned about earlier.

We've touched on the basics of cleaning internal and external data, but there's lots more to come. Soon, we'll learn about the common errors to be aware of to ensure your data is complete, correct, and relevant. See you soon!

[**ANGIE: I LOVE CLEANING DATA**](https://www.coursera.org/learn/process-data/lecture/hGbve/angie-i-love-cleaning-data)

I am Angie. I'm a program manager of engineering at Google. I truly believe that cleaning Data is the heart and soul of data. It's how you get to know your data: its quirks, its flaws, it's mysteries. I love a good mystery.

I remember one time I found somebody had purchased, I think it was one million dollars worth of chicken sandwiches in one transaction. This mystery drove me nuts. I had all these questions. Could this have really happened? Or maybe it was a really big birthday party? How did they make a million dollars worth of chicken sandwiches? I was cleaning my data and trying to figure out where it went wrong.

We ended up finding out that we'd been squaring and multiplying all of our transactions for a very specific case. It took us about three days to figure this out. I will never forget the moment when it was like, aha! We got to the bottom of it.

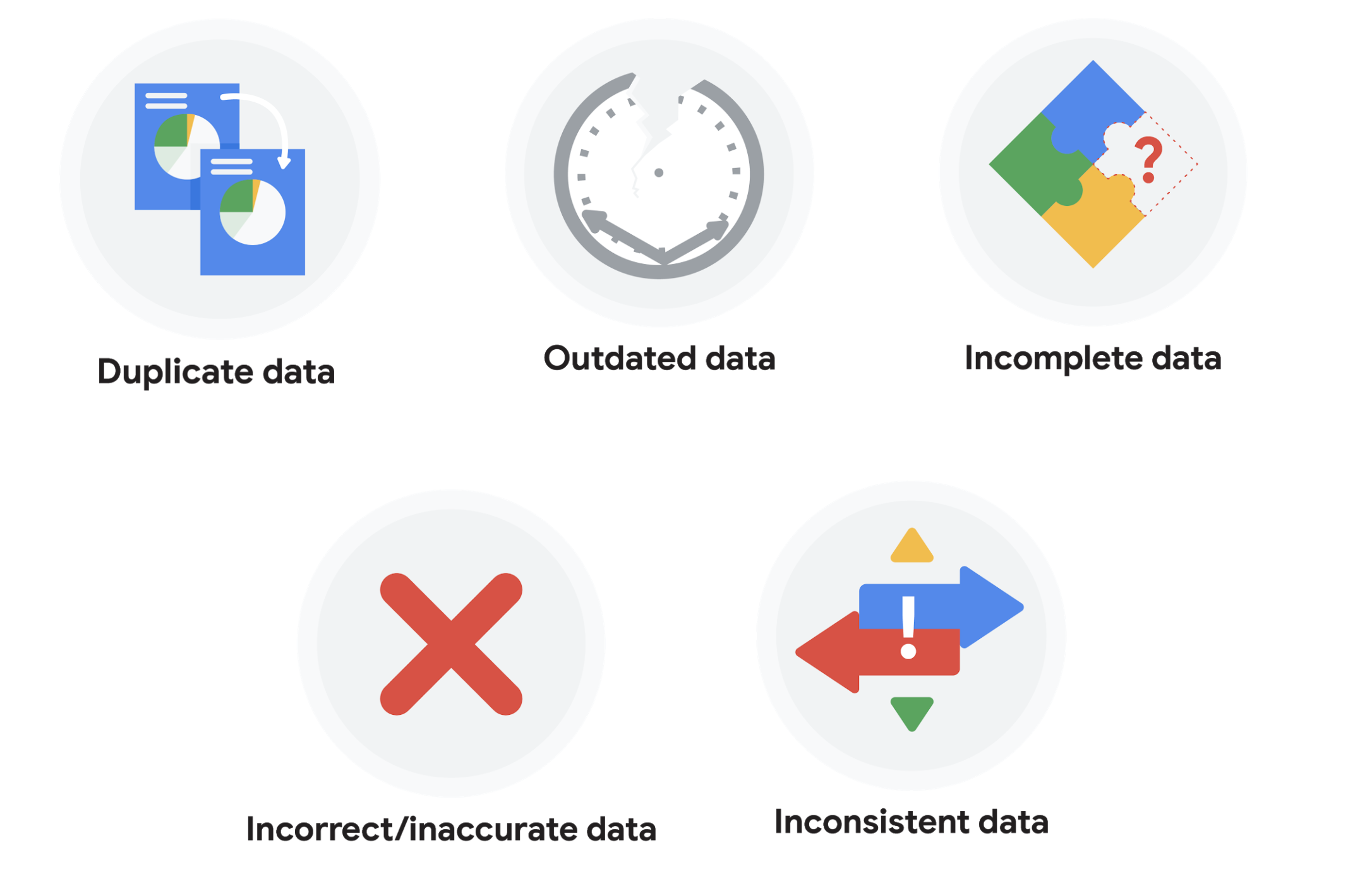
The result is our data was cleaned, and we had this great dataset that we could use for analysis. But what I loved was just the mystery of it and getting to know all these weird intricacies about my dataset. It felt like a superpower almost, like I was a detective, and I had gone in there and I had really solved something. I love cleaning data!

[**WHAT IS DIRTY DATA?**](https://www.coursera.org/learn/process-data/supplement/b90nY/what-is-dirty-data)

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

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

## **Types of dirty data**

****

### 

### **Duplicate data**

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

### **Outdated data**

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

### **Incomplete data**

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

### **Incorrect/inaccurate data**

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

### **Inconsistent data**

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

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

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

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

## **Key takeaways**

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

[**RECOGNIZE AND REMEDY DIRTY DATA**](https://www.coursera.org/learn/process-data/lecture/JQy3r/recognize-and-remedy-dirty-data)

Hey, there. In this video, we'll focus on common issues associated with dirty data. These include spelling and other text errors, inconsistent labels, formats and field lanes, missing data and duplicates.

This will help you recognize problems quicker and give you the information you need to fix them when you encounter something similar during your own analysis. This is incredibly important in data analytics.

Let's go back to our law office spreadsheet. As a quick refresher, we'll start by checking out the different types of dirty data it shows.

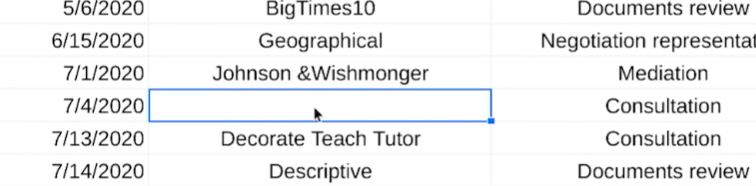
Sometimes, someone might key in a piece of data incorrectly.



Other times, they might not keep data formats consistent.



It's also common to leave a field blank.



That's also called **a null**, which we learned about earlier. **If someone adds the same piece of data more than once, that creates a duplicate.**

Let's break that down. Then we'll learn about a few other types of dirty data and strategies for cleaning it.

Misspellings, spelling variations, mixed up letters, inconsistent punctuation, and typos in general, happen when someone types in a piece of data incorrectly. As a data analyst, you'll also deal with different currencies. For example, one dataset could be in US dollars and another in euros, and you don't want to get them mixed up. We want to find these types of errors and fix them like this.

Clean data depends largely on the data integrity rules that an organization follows, such as spelling and punctuation guidelines.

For example, a beverage company might ask everyone working in its database to enter data about volume in fluid ounces instead of cups. It's great when an organization has rules like this in place. It really helps minimize the amount of data cleaning required, but it can't eliminate it completely. Like we discussed earlier, there's always the possibility of human error. The next type of dirty data our spreadsheet shows is inconsistent formatting. In this example, something that should be formatted as currency is shown as a percentage. Until this error is fixed, like this, the law office will have no idea how much money this customer paid for its services. We'll learn about different ways to solve this and many other problems soon.

We discussed nulls previously, but as a reminder, **nulls are empty fields**. This kind of dirty data requires a little more work than just fixing a spelling error or changing a format. In this example, the data analysts would need to research which customer had a consultation on July 4th, 2020. Then when they find the correct information, they'd have to add it to the spreadsheet.

Another **common type of dirty data** is **duplicated**.

Maybe two different people added this appointment on August 13th, not realizing that someone else had already done it or maybe the person entering the data hit copy and paste by accident. Whatever the reason, it's the data analyst job to identify this error and correct it by deleting one of the duplicates.

Now, let's continue on to some other types of dirty data. The first has to do with labeling. To understand labeling, imagine trying to get a computer to correctly identify panda bears among images of all different kinds of animals. 

You need to show the computer thousands of images of panda bears. They're all labeled as panda bears. Any incorrectly labeled picture, like the one here that's just “bear”, will cause a problem.

The next type of **dirty data is having an inconsistent field length**. You learned earlier that a field is a single piece of information from a row or column of a spreadsheet. Field length is a tool for determining how many characters can be keyed into a field.

**Assigning a certain length to the fields in your spreadsheet is a great way to avoid errors**.

For instance, if you have a column for someone's birth year, you know the field length is four because all years are four digits long. Some spreadsheet applications have a simple way to specify field lengths and make sure users can only enter a certain number of characters into a field. **This is part of data validation**.

Data validation is a tool for checking the accuracy and quality of data before adding or importing it. Data validation is a form of data cleansing, which you'll learn more about soon. But first, you'll get familiar with more techniques for cleaning data. This is a very important part of the data analyst job. I look forward to sharing these data cleaning strategies with you.

**FIRST STEPS TOWARD CLEAN DATA**

[**DATA-CLEANING TOOLS AND TECHNIQUES**](https://www.coursera.org/learn/process-data/lecture/epVyj/data-cleaning-tools-and-techniques)

Hi. Now that you're familiar with some of the most common types of dirty data, it's time to clean them up. As you've learned, **clean data is essential to data integrity and reliable solutions and decisions.** The good news is that spreadsheets have all kinds of tools you can use to get your data ready for analysis.

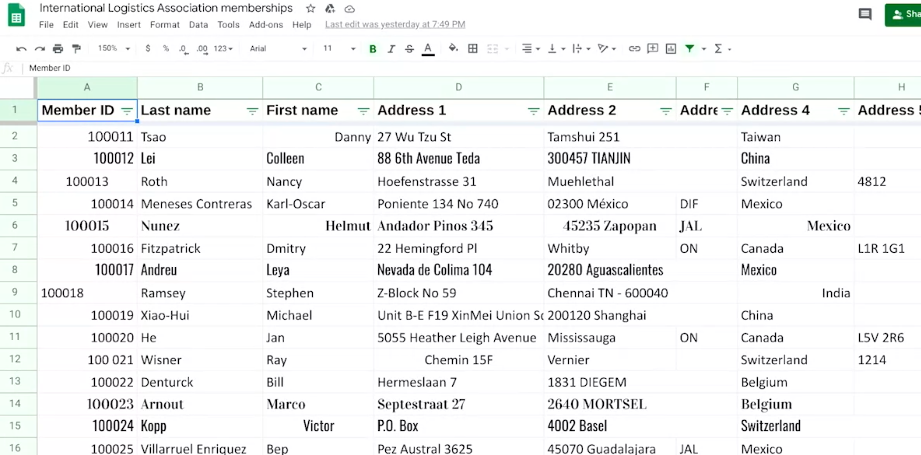
The techniques for data cleaning will be different depending on the specific data set you're working with.

This will give you a great starting point for fixing the types of dirty data analysts find most often. Think of everything that's coming up as a teaser trailer of data cleaning tools.

Here, we'll discuss how to remove unwanted data, clean up text to remove extra spaces and blanks, fix typos, and make formatting consistent. However, **before removing unwanted data, it's always a good practice to make a copy of the data set**. That way, if you remove something that you end up needing in the future, you can easily access it and put it back in the data set.

Once that's done, then you can move on to getting rid of the duplicates or data that isn't relevant to the problem you're trying to solve. Typically, duplicates appear when you're combining data sets from more than one source or using data from multiple departments within the same business.

You've already learned a bit about duplicates, but let's practice removing them once more now using this spreadsheet, which lists members of a professional logistics association.



Duplicates can be a big problem for data analysts. So it's really important that you can find and remove them before any analysis starts. Here's an example of what I'm talking about.

Let's say this association has duplicates of one person's $500 membership in its database.

When the data is summarized, the analyst would think there was $1,000 being paid by this member and would make decisions based on that incorrect data. But in reality, this member only paid $500. These problems can be fixed manually, but most spreadsheet applications also offer lots of tools to help you find and remove duplicates.

Now, irrelevant data, which is data that doesn't fit the specific problem that you're trying to solve, also needs to be removed. Going back to our association membership list example, let's say a data analyst was working on a project that focused only on current members. They wouldn't want to include information on people who are no longer members, or who never joined in the first place.

Removing irrelevant data takes a little more time and effort because you have to figure out the difference between the data you need and the data you don't. But believe me, making those decisions will save you a ton of effort down the road.

The next step is removing extra spaces and blanks. Extra spaces can cause unexpected results when you sort, filter, or search through your data. And because these characters are easy to miss, they can lead to unexpected and confusing results. For example, if there's extra space and a member ID number, when you sort the column from lowest to highest, this row will be out of place.

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

Or again, you can rely on your spreadsheets, which offer lots of great functions for removing spaces or blanks automatically. The next data cleaning step involves fixing misspellings, inconsistent capitalization, incorrect punctuation, and other typos. These types of errors can lead to some big problems. Let's say you have a database of emails that you use to keep in touch with your customers. If some emails have misspellings, a period in the wrong place, or any other kind of typo, not only do you run the risk of sending an email to the wrong people, you also run the risk of spamming random people. Think about our association membership example again. Misspelling might cause the data analyst to miscount the number of professional members if they sorted this membership type and then counted the number of rows.

Like the other problems you've come across, you can also fix these problems manually.

Or you can use spreadsheet tools, such as spellcheck, autocorrect, and conditional formatting to make your life easier. There's also easy ways to convert text to lowercase, uppercase, or proper case, which is one of the things we'll check out again later.

All right, we're getting there. The next step is removing formatting. This is particularly important when you get data from lots of different sources. Every database has its own formatting, which can cause the data to seem inconsistent.

Creating a clean and consistent visual appearance for your spreadsheets will help make it a valuable tool for you and your team when making key decisions. Most spreadsheet applications also have a "clear formats" tool, which is a great time saver.

Cleaning data is an essential step in increasing the quality of your data. Now you know lots of different ways to do that. In the next video, you'll take that knowledge even further and learn how to clean up data that's come from more than one source.

[**CLEAN DATA FROM MULTIPLE SOURCES**](https://www.coursera.org/learn/process-data/lecture/ZOoSh/clean-data-from-multiple-sources)

So far you've learned a lot about dirty data and how to clean up the most common errors in a dataset. Now we're going to take that a step further and talk about cleaning up multiple datasets.

**Cleaning data that comes from two or more sources is very common for data analysts, but it does come with some interesting challenges**.

A good example is a **merger**, which is an agreement that **unites two organizations into a single new one**.

In the logistics field, there's been lots of big changes recently, mostly because of the e-commerce boom. With so many people shopping online, it makes sense that the companies responsible for delivering those products to their homes are in the middle of a big shake-up. When big things happen in an industry, it's common for two organizations to team up and become stronger through a merger. Let's talk about how that will affect our logistics association.

As a quick reminder, this spreadsheet lists association member ID numbers, first and last names, addresses, how much each member pays in dues, when the membership expires, and the membership types.

Now, let's think about what would happen if the International Logistics Association decided to get together with the Global Logistics Association in order to help their members handle the incredible demands of e-commerce. First, all the data from each organization would need to be combined using data merging. Data merging is the process of combining two or more datasets into a single dataset. This presents a unique challenge because when two totally different datasets are combined, the information is almost guaranteed to be inconsistent and misaligned.

For example, the Global Logistics Association's spreadsheet has a separate column for a person's suite, apartment, or unit number, but the International Logistics Association combines that information with their street address. This needs to be corrected to make the number of address columns consistent. Next, check out how the Global Logistics Association uses people's email addresses as their member ID, while the International Logistics Association uses numbers. This is a big problem because people in a certain industry, such as logistics, typically join multiple professional associations. There's a very good chance that these datasets include membership information on the exact same person, just in different ways. It's super important to remove those duplicates. Also, the Global Logistics Association has many more member types than the other organization.

On top of that, it uses the term, "Young Professional" instead of "Student Associate."

But both describe members who are still in school or just starting their careers. If you were merging these two datasets, you'd need to work with your team to fix the fact that the two associations describe memberships very differently.

Now you understand why the merging of organizations also requires the merging of data, and that can be tricky.

But there's lots of other reasons why data analysts merge datasets.

For example, in one of my past jobs, I merged a lot of data from multiple sources to get insights about our customers' purchases. The kinds of insights I gained helped me identify customer buying patterns. When merging datasets, I always begin by asking myself some key questions to help me avoid redundancy and to confirm that the datasets are compatible.

In data analytics, compatibility describes how well two or more datasets are able to work together.

The first **question** I would ask is, **do I have all the data I need?** To gather customer purchase insights, I wanted to make sure I had data on customers, their purchases, and where they shopped.

Next I would ask, **does the data I need exist within these datasets?** As you learned earlier in this program, this involves considering the entire dataset analytically. Looking through the data before I start using it lets me get a feel for what it's all about, what the schema looks like, if it's relevant to my customer purchase insights, and if it's clean data. That brings me to the next question.

**Do the datasets need to be cleaned, or are they ready for me to use?**

Because I'm working with more than one source, I will also ask myself, **are the datasets cleaned to the same standard?**

**For example, what fields are regularly repeated?**

**How are missing values handled?**

**How recently was the data updated?**

Finding the answers to these questions and understanding if I need to fix any problems at the start of a project is a very important step in data merging. In both of the examples we explored here, data analysts could use either the spreadsheet tools or SQL queries to clean up, merge, and prepare the datasets for analysis. Depending on the tool you decide to use, the cleanup process can be simple or very complex. Soon, you'll learn how to make the best choice for your situation. As a final note, programming languages like R are also very useful for cleaning data. You'll learn more about how to use R and other concepts we covered soon.

[**COMMON DATA-CLEANING PITFALLS**](https://www.coursera.org/learn/process-data/supplement/m3iWu/common-data-cleaning-pitfalls)



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

## **COMMON MISTAKES TO AVOID**

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

## **Key takeaways**

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

## **Additional resources**

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

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

**CONTINUE CLEANING DATA IN SPREADSHEETS**

[**STEP-BY-STEP GUIDE: DATA-CLEANING FEATURES IN SPREADSHEETS**](https://www.coursera.org/learn/process-data/supplement/jIF1e/step-by-step-guide-data-cleaning-features-in-spreadsheets)



## **Activity overview**

****

### You’ve learned about cleaning data and its importance in meeting good data science standards. In this activity, you’ll do some data cleaning with spreadsheets, then transpose the data.

### By the time you complete this activity, you will be able to perform some basic cleaning methods in spreadsheets. This will enable you to clean and transpose data, which is important for making data more specific and accurate in your career as a data analyst.

### Step-By-Step Instructions

### Step 1: Access the template

To get started, first access the data spreadsheet.

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

Link to data spreadsheet: [Cleaning with spreadsheets](https://docs.google.com/spreadsheets/d/1PkAbgXC7C1g2dKzCCpaHBcAyPw-s1z7iUxIEJ0cCYWQ/template/preview)

OR

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

[Data Spreadsheet for Cleaning with Spreadsheets](https://d3c33hcgiwev3.cloudfront.net/9Ss-5kXlRP6rPuZF5VT-qA_d5c83c4e13d643cd87cb22632fcaa9c6_Data-Spreadsheet-for-Cleaning-with-Spreadsheets.xlsx?Expires=1712188800&Signature=XI4RZs-66RH~Q1ftYKElDEm9ryKc0xF8YKru5j1ZnVb8CI6z5P6VJzOB9hyF8udHhQxY1XszbPjHXemODG1HC2lZPEsLfEmcP20pC9v-FkzQUIAD~CH3u-Rxqgh4-0ISi0~tXYWUPsB6zd84Q2wwc64lTQ6xAYDfmVK63vOABww_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

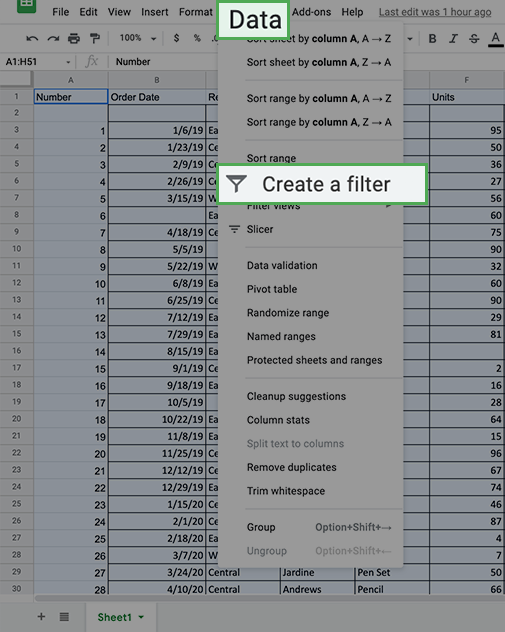
[XLSX File](https://d3c33hcgiwev3.cloudfront.net/9Ss-5kXlRP6rPuZF5VT-qA_d5c83c4e13d643cd87cb22632fcaa9c6_Data-Spreadsheet-for-Cleaning-with-Spreadsheets.xlsx?Expires=1712188800&Signature=XI4RZs-66RH~Q1ftYKElDEm9ryKc0xF8YKru5j1ZnVb8CI6z5P6VJzOB9hyF8udHhQxY1XszbPjHXemODG1HC2lZPEsLfEmcP20pC9v-FkzQUIAD~CH3u-Rxqgh4-0ISi0~tXYWUPsB6zd84Q2wwc64lTQ6xAYDfmVK63vOABww_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

### Step 2: Select and remove blank ells

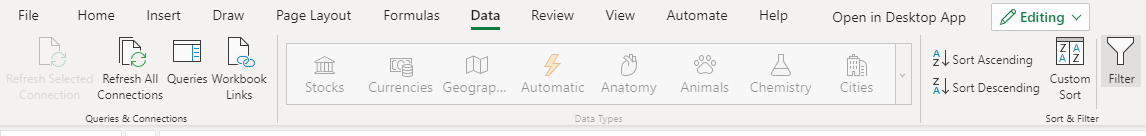
The first technique we’ll use is to select and eliminate rows containing blank cells by using filters. To eliminate rows with blank cells:

1. Highlight all cells in the spreadsheet. You can highlight Columns A-H by clicking on the header of Column A, holding Shift, and clicking on the header of Column H.

2. Click on the Data tab and pick the Create a filter option. In Microsoft Excel, this is called Filter.

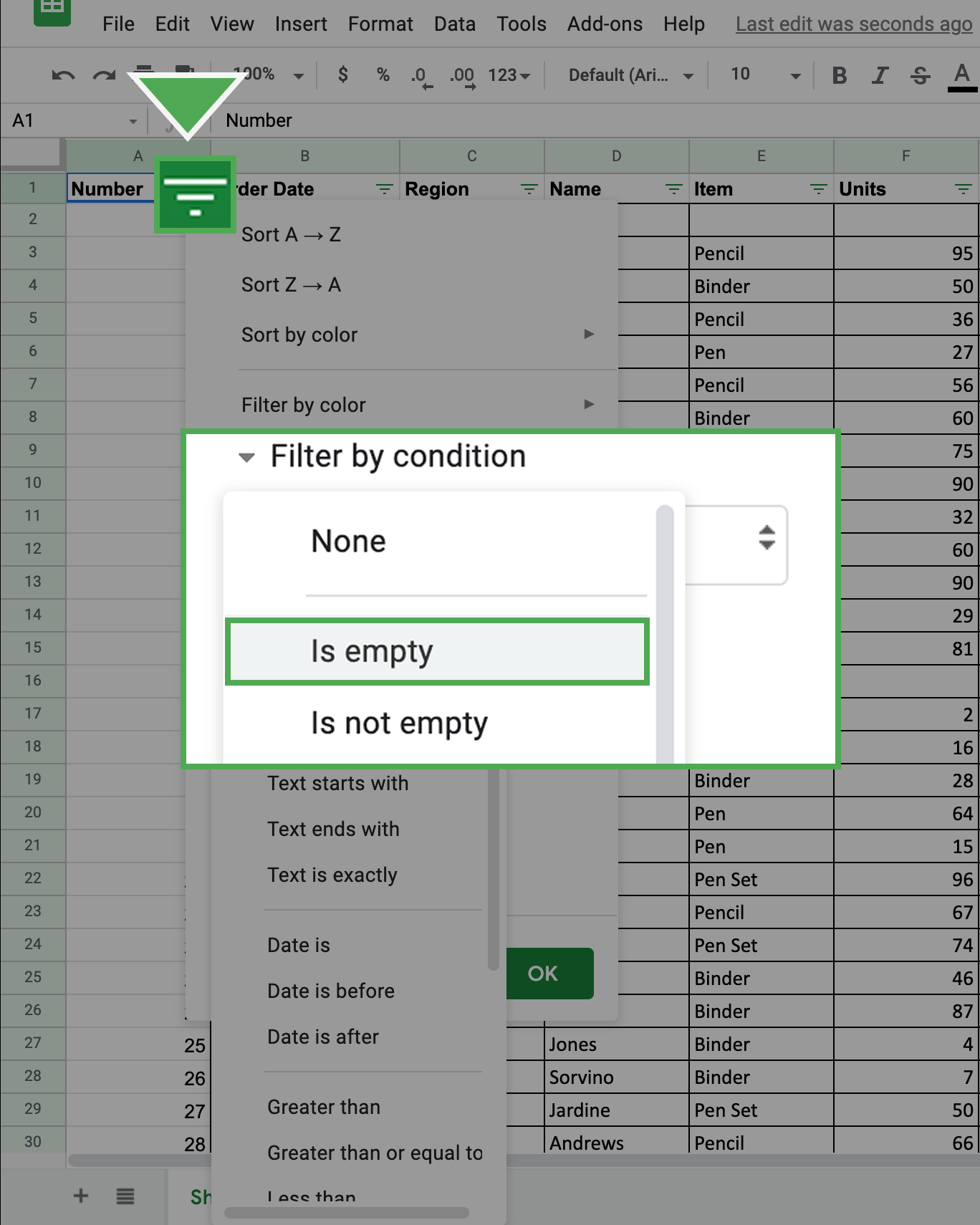


Excel:



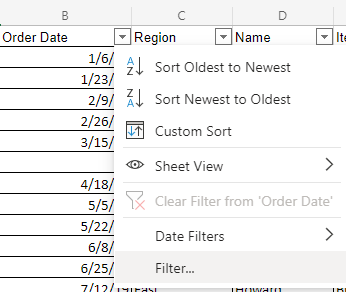
3. Every column now shows a green triangle in the first row next to the column title. Click the green triangle in Column B to access a new menu.

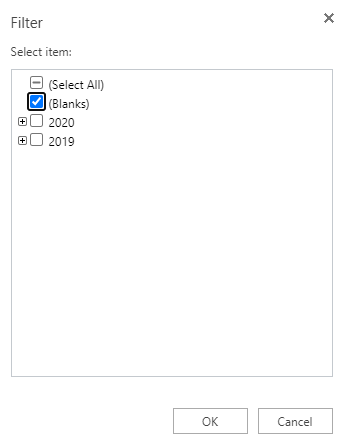
4. On that new menu, click Filter by condition and open the dropdown menu to select Is empty. Click OK.



In Excel, click the dropdown, then Filter... then make sure only (Blanks) is checked. Click OK.

Excel:





You can then review a list of all the rows with blank cells in that column.

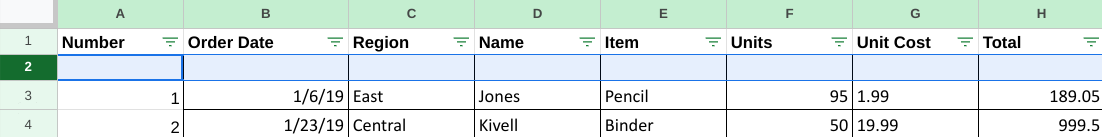
5. Select all these cells and delete the rows except the row of column headers.

6. Return to the Filter by condition and return it to None. In Excel, click Clear Filter from ‘Column’.

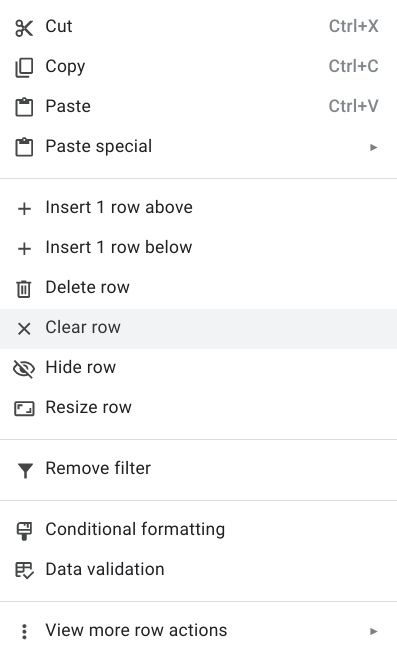
* Note: You will now notice that any row that had an empty cell in Column A will be removed (including the extra empty rows after the data).

7. Repeat this for Columns B-H.

8. Note: If you simply deleted the data from the row by tapping the backspace button, you will need to go a step further and *delete the empty row entirely* by left-clicking the row number located on the furthest left side of the screen.



9. Next, right click on the highlighted row to call up the drop down window, and select the Delete row option.



10. Continue to do this same operation for the remaining empty rows in the data set.

All the rows that had blank cells are now removed from the spreadsheet.

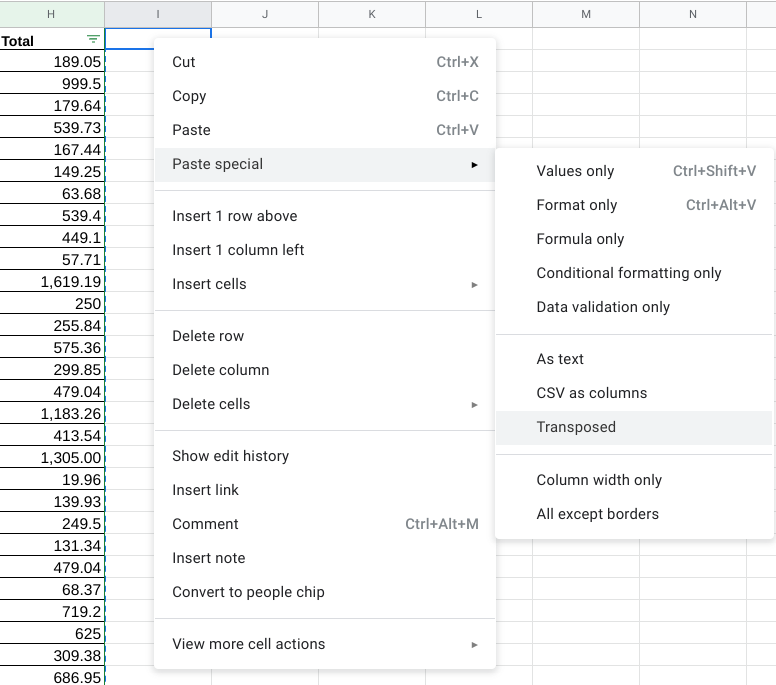
### Step 3: Transpose the data

The second technique you will practice will help you convert the data from the current long format (more rows than columns) to the wide format (more columns than rows). This action is called transposing. To transpose your data:

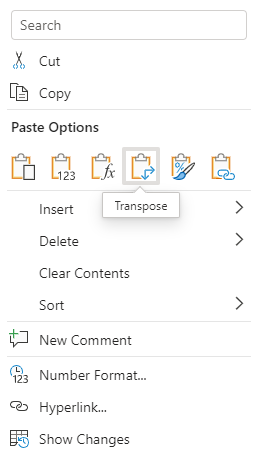
1. Highlight and copy the data that you want to transpose including the column labels. You can do this by highlighting Columns A-H. In Excel, highlight only the relevant cells (A1-H45) instead of the headers.

2. Right-click on cell I1. This is where you want the transposed data to start.

3. Hover over Paste Special from the right-click menu. Select the Transposed option. In Excel, select the Transpose icon under the paste options.

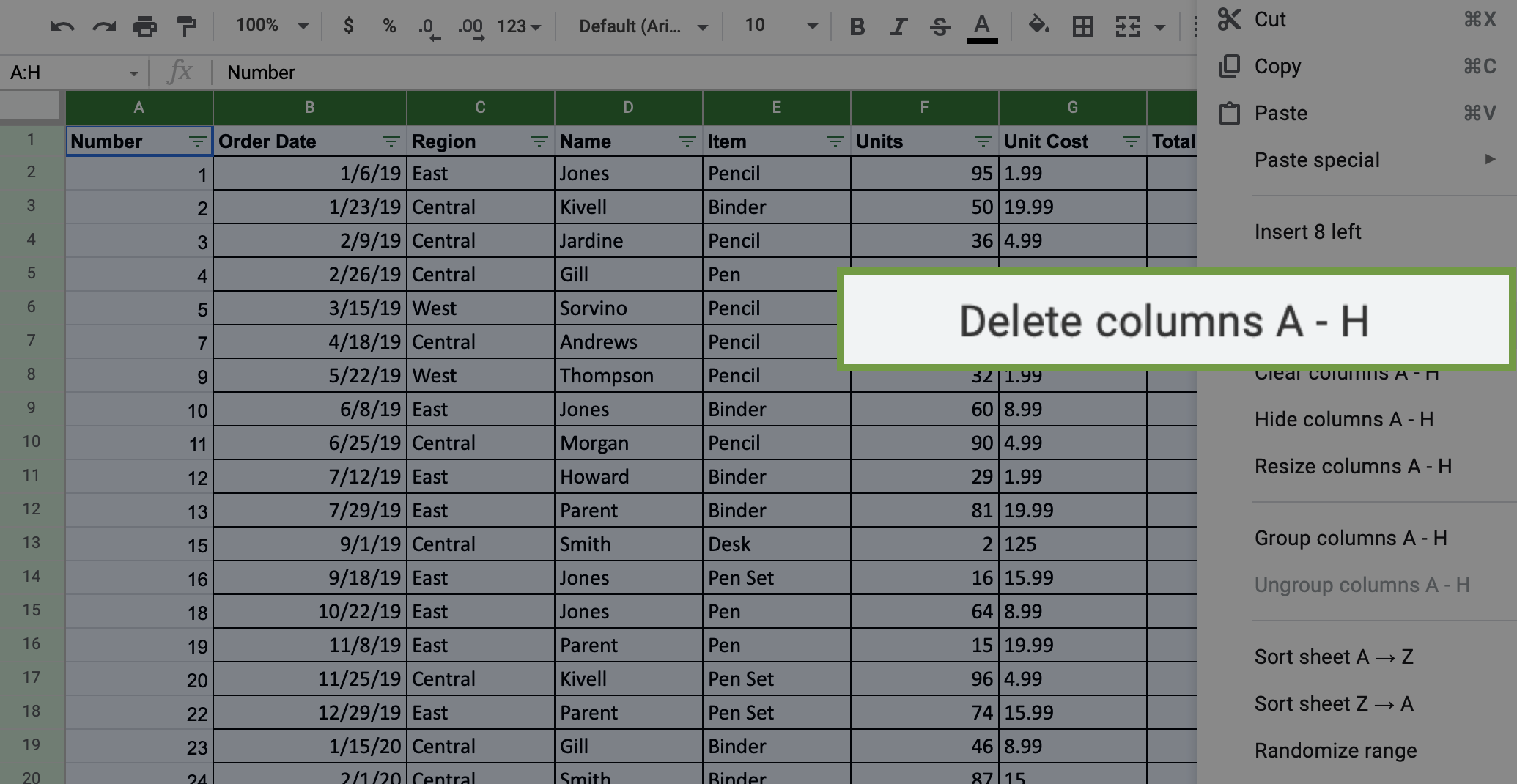


Excel:

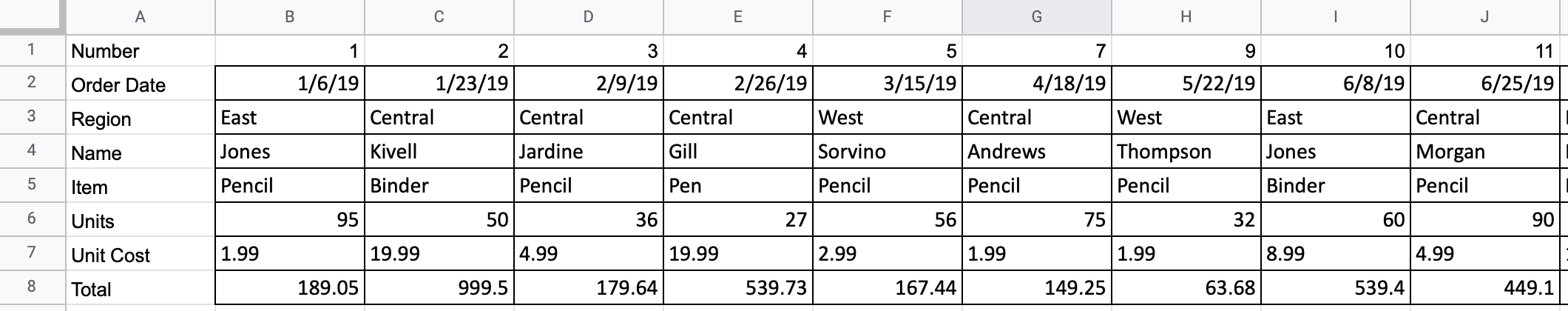


You should now find the data transformed into the new wide format. At this point, you should remove the original long data from the spreadsheet.

4. Delete the previous long data. The easiest way to do this is to click on Column A, so the entire column is highlighted. Then, hold down the Shift key and click on Column H. You should find these columns highlighted. Right-click on the highlighted area and select Delete Columns A - H.



Your screen should now appear like this:

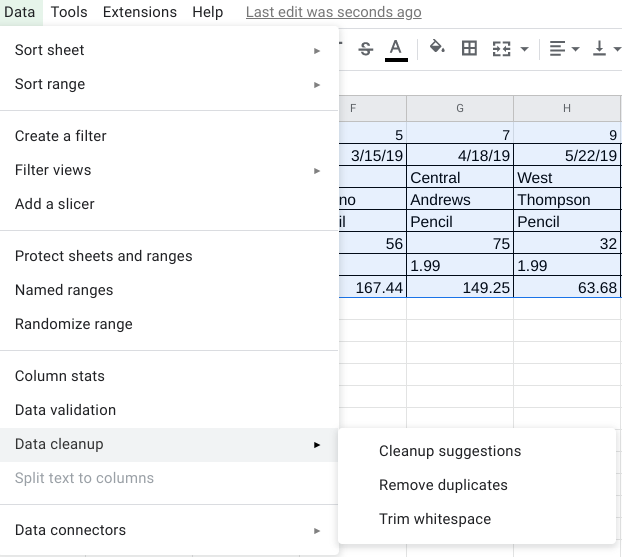


### Step 4: Get rid of extra spaces in cells with string data

Now that you have transposed the data, eliminate the extra spaces in the values of the cells.

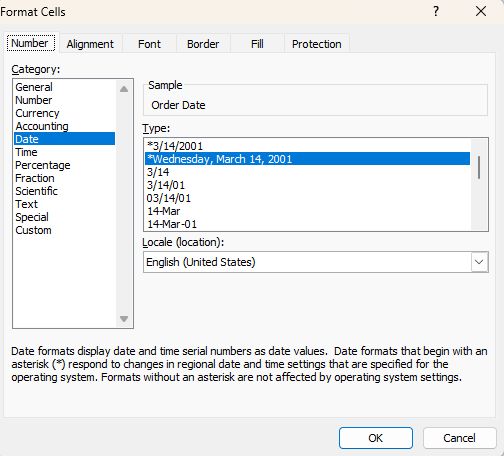
1. Highlight the data in the spreadsheet.

2. Click on the Data tab, then hover over Data cleanup and select Trim whitespace.



In Excel, you can use the TRIM command to get rid of white spaces. In any space beneath your data (such as cell A10), enter **=TRIM(A1)**. Then, drag the auto-fill square handle at the bottom right corner of the cell and pull it downward to call the rest of the data in the column without the white spaces.

Note: It's important to carry out the **TRIM** formula for each column individually, as it may affect certain types of cell values (like dates) undesirably. One option is right-clicking the "Order Date" column at the top, and selecting the Format Cells option to change the dates to string values before applying the **TRIM** formula.



For additional understanding of using the TRIM formula in Excel, please explore the [Microsoft Excel guide](https://support.microsoft.com/en-au/office/trim-function-50b6f593-4801-4f06-81c5-3345e7f80ddd) on using the formula correctly.

Now all the extra spaces in the cells have been removed.

### Step 5: Change lower/uppercase/proper case text

Next, you’ll process string data. The easiest way to clean up string data will depend on the spreadsheet program you are using. If you are using Excel, you’ll use a simple formula. If you are using Google Sheets, you can use an Add-on to do this with a few clicks. Follow the steps in the relevant section below.

### **Microsoft Excel**

****

If you are using Microsoft Excel, [this documentation](https://support.microsoft.com/en-us/office/change-the-case-of-text-in-excel-adc65f5b-958f-46a2-4d23-ab4d5faf48a8) explains how to use a formula to change the case of a text string. Follow these instructions to clean the string text and then move on to the confirmation and reflection section of this activity.

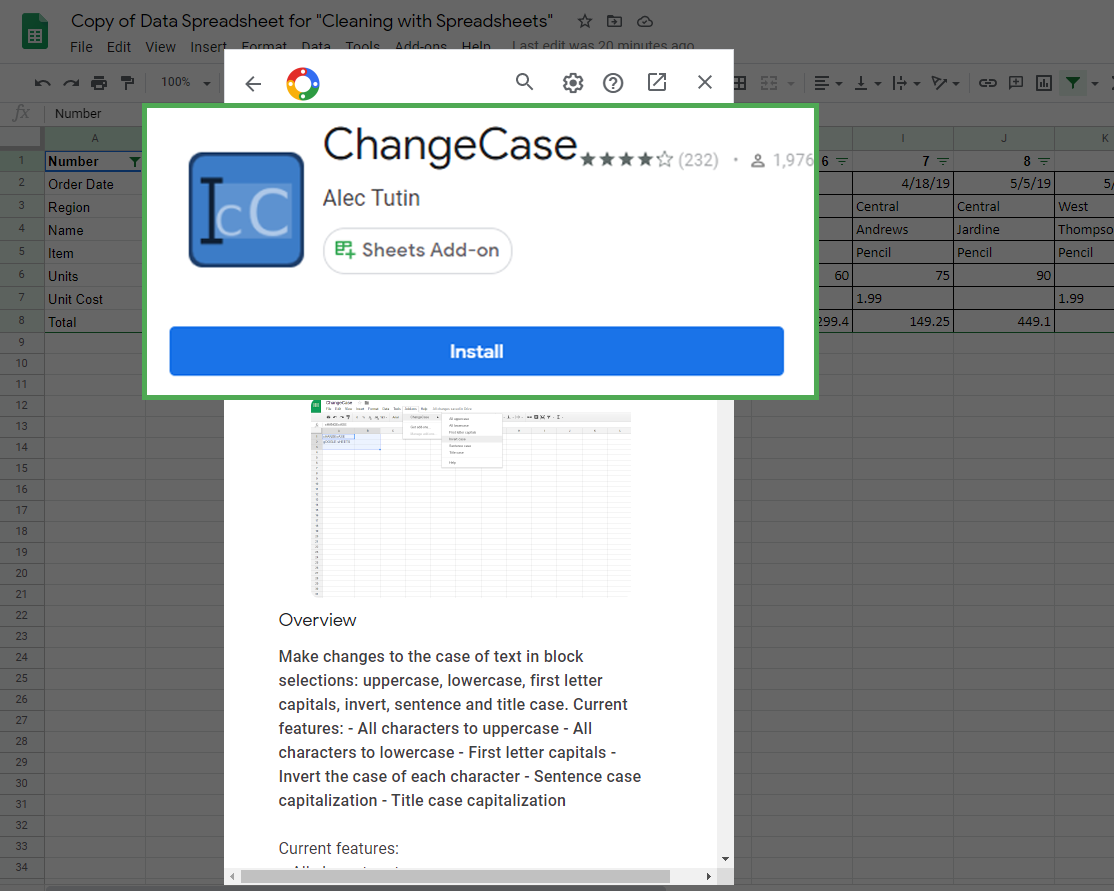
### **Google sheets**

****

If you’re completing this exercise using Google Sheets, you’ll need to install an add-in that will give you the functionality needed to easily clean string data and change cases.

*Google Sheets Add-on Instructions:*

1. Click on the Add-Ons option at the top of Google Sheets.
2. Click on Get add-ons.
3. Search for ChangeCase. It should appear like this:



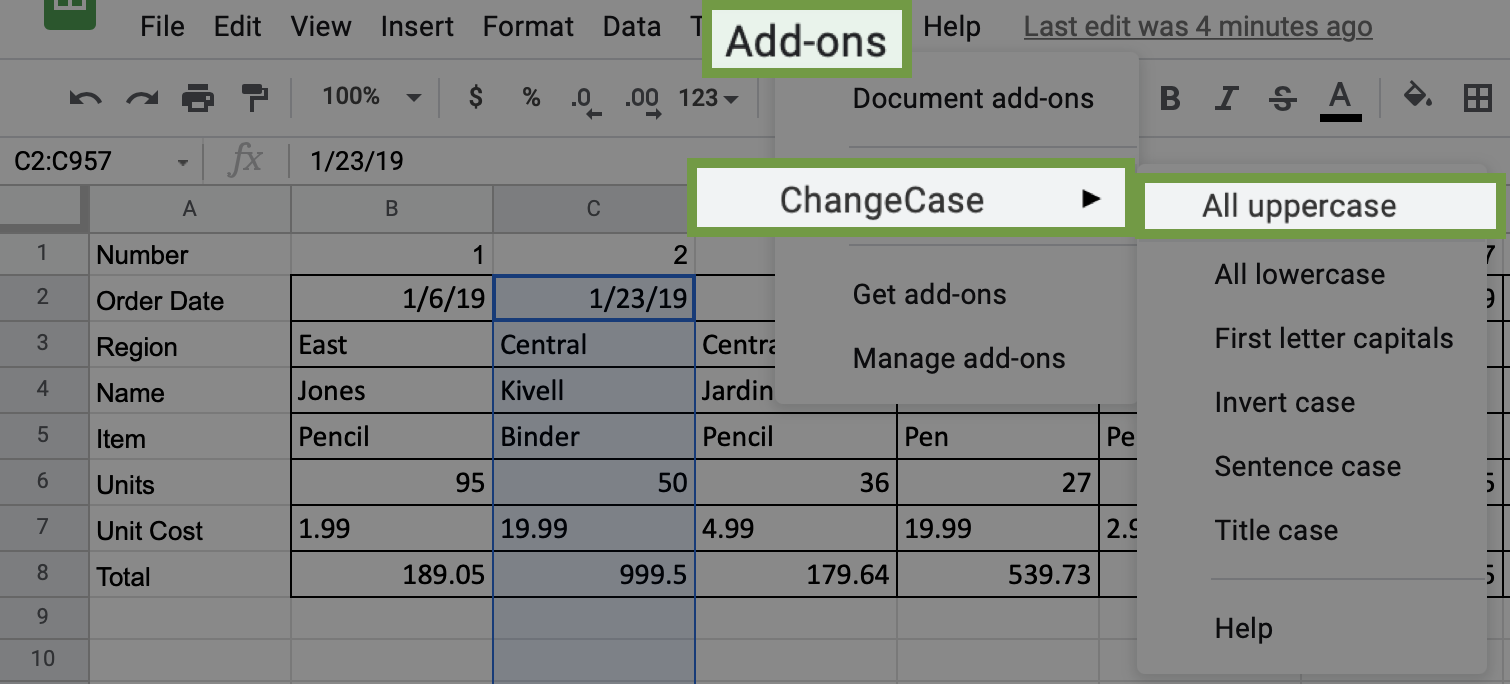
4. Click on Install to install the dd-on. It may ask you to login or verify the installation permissions.

Once you have installed the add-on successfully, you can access it by clicking on the Add-ons menu again.

Now, you can change the case of text data that shows up. To change the text in Column C to all uppercase:

1. Click on Column C. Be sure to deselect the column header, unless you want to change the case of that as well (which you don't).

2. Click on the Add-Ons tab and select ChangeCase. Select the option All uppercase. Notice the other options that you could have chosen if needed.



### Step 6: Delete all formatting

If you want to clear the formatting for any or all cells, you can find the command in the Format tab. To clear formatting:

1. Select the data for which you want to delete the formatting. In this case, highlight all the data in the spreadsheet by clicking and dragging over Rows 1-8.

2. Click the Format tab and select the Clear Formatting option.

In Excel, go to the Home tab, then hover over Clear and select Clear Formats.

You will notice that all the cells have had their formatting removed.

### Pro Tip: Save the activity template

Be sure to save a copy of the spreadsheet template you used to complete this activity. You can use it for further practice or to help you work through your thought processes for similar tasks in a future data analyst role.

[**DATA-CLEANING FEATURES IN SPREADSHEETS**](https://www.coursera.org/learn/process-data/lecture/Ez3u5/data-cleaning-features-in-spreadsheets)

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

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

**What you’ll need**

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

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

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

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

OR

[International Logistics Association Memberships - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/5zlIemyvQtKsj4BGZQEdxA_7f286fde512b4f14b3246a6e68b333e1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1712188800&Signature=Cex3IYuJQ~wlSBWoTHAoqv-RyekkBOkLKz8Tg4xH8~nwBsVTJlsmcRQZW3iODMh8aFYM3KNH5~osZmB9A8UVpQ3ZkoyuD3ZlawRbA6ZwgK717XSUMafr0NUKJzM2p-qvB~dcXVVQPdUn9L441EEZwENiPHPm-AfITCchYeBZRUY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/5zlIemyvQtKsj4BGZQEdxA_7f286fde512b4f14b3246a6e68b333e1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1712188800&Signature=Cex3IYuJQ~wlSBWoTHAoqv-RyekkBOkLKz8Tg4xH8~nwBsVTJlsmcRQZW3iODMh8aFYM3KNH5~osZmB9A8UVpQ3ZkoyuD3ZlawRbA6ZwgK717XSUMafr0NUKJzM2p-qvB~dcXVVQPdUn9L441EEZwENiPHPm-AfITCchYeBZRUY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[Cosmetics Inc. - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/t3_O7KI_Rp6KVfCyGvTxGA_26e91b3aa1d44284956a4ae860d114e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1712188800&Signature=DxAlgRfdDg-MvK8wDqbWdnMKjP9AOEcf7KrSUlo9Y7IYFjIILAEf-Vmz29zC7t6V3JlmwJuGGiwV2C4k7BqnoUkVcYZYMO-Gq5PmzaBe6xYAYBE5EtN0ddb~XxGbebZPCXrJ0RpvsYf4756oZsOwzcrkvFBi5ovRf4OWdeyoLVE_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/t3_O7KI_Rp6KVfCyGvTxGA_26e91b3aa1d44284956a4ae860d114e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1712188800&Signature=DxAlgRfdDg-MvK8wDqbWdnMKjP9AOEcf7KrSUlo9Y7IYFjIILAEf-Vmz29zC7t6V3JlmwJuGGiwV2C4k7BqnoUkVcYZYMO-Gq5PmzaBe6xYAYBE5EtN0ddb~XxGbebZPCXrJ0RpvsYf4756oZsOwzcrkvFBi5ovRf4OWdeyoLVE_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[](https://d3c33hcgiwev3.cloudfront.net/t3_O7KI_Rp6KVfCyGvTxGA_26e91b3aa1d44284956a4ae860d114e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1712188800&Signature=DxAlgRfdDg-MvK8wDqbWdnMKjP9AOEcf7KrSUlo9Y7IYFjIILAEf-Vmz29zC7t6V3JlmwJuGGiwV2C4k7BqnoUkVcYZYMO-Gq5PmzaBe6xYAYBE5EtN0ddb~XxGbebZPCXrJ0RpvsYf4756oZsOwzcrkvFBi5ovRf4OWdeyoLVE_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

## 

## 

## 

## 

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

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

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

## **Example 2: Remove duplicates**

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

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

## **Example 3: Format dates consistently**

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

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

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

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

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

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

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

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

[**STEP-BY-STEP: OPTIMIZE THE DATA-CLEANING PROCESS**](https://www.coursera.org/learn/process-data/supplement/2eSmP/step-by-step-optimize-the-data-cleaning-process)

Hi again. As you learned earlier, there's a lot of different ways to clean up data. I've shown you some examples of how you can clean data manually, such as searching for and fixing misspellings or removing empty spaces and duplicates. We also learned that lots of spreadsheet applications have tools that help simplify and speed up the data cleaning process.

There's a lot of great efficiency tools that data analysts use all the time, such as conditional formatting, removing duplicates, formatting dates, fixing text strings and substrings, and splitting text to columns. We'll explore those in more detail now.

The first is something called **conditional formatting**. Conditional formatting is a spreadsheet tool that changes how cells appear when values meet specific conditions. Likewise, it can let you know when a cell does not meet the conditions you've set. Visual cues like this are very useful for data analysts, especially when we're working in a large spreadsheet with lots of data. Making certain data points standout makes the information easier to understand and analyze. For cleaning data, knowing when the data doesn't follow the condition is very helpful.

Let's return to the logistics association spreadsheet to check out **conditional formatting in action**. We'll use conditional formatting to highlight blank cells. That way, we know where there's missing information so we can add it to the spreadsheet. To do this, we'll start by selecting the range we want to search. For this example we're not focused on address 3 and address 5. The fields will include all the columns in our spreadsheets, except for F and H. Next, we'll go to Format and choose Conditional formatting.

Great. Our range is automatically indicated in the field. The format rule will be to format cells if the cell is empty.

Finally, we'll choose the formatting style. I'm going to pick a shade of bright pink, so my blanks really stand out.

Then click "Done," and the blank cells are instantly highlighted. The next spreadsheet tool removes duplicates. As you've learned before, it's always smart to make a copy of the data set before removing anything. Let's do that now.

Great, now we can continue. You might remember that our example spreadsheet has one association member listed twice.

To fix that, go to Data and select "Remove duplicates." "Remove duplicates" is a tool that automatically searches for and eliminates duplicate entries from a spreadsheet. Choose "Data has header row" because our spreadsheet has a row at the very top that describes the contents of each column. Next, select "All" because we want to inspect our entire spreadsheet. Finally, "Remove duplicates."

You'll notice the duplicate row was found and immediately removed.

Another useful spreadsheet tool enables you to make formats consistent. For example, some of the dates in this spreadsheet are in a standard date format.

This could be confusing if you wanted to analyze when association members joined, how often they renewed their memberships, or how long they've been with the association. To make all of our dates consistent, first select column J, then go to "Format," select "Number," then "Date." Now all of our dates have a consistent format. Before we go over the next tool, I want to explain what a text string is. In data analytics, a text string is a group of characters within a cell, most often composed of letters. An important characteristic of a text string is its length, which is the number of characters in it. You'll learn more about that soon. For now, it's also useful to know that a substring is a smaller subset of a text string. Now let's talk about Split. Split is a tool that divides a text string around the specified character and puts each fragment into a new and separate cell. Split is helpful when you have more than one piece of data in a cell and you want to separate them out. This might be a person's first and last name listed together, or it could be a cell that contains someone's city, state, country, and zip code, but you actually want each of those in its own column. Let's say this association wanted to analyze all of the different professional certifications its members have earned. To do this, you want each certification separated out into its own column. Right now, the certifications are separated by a comma. That's the specified text separating each item, also called the delimiter. Let's get them separated. Highlight the column, then select "Data," and "Split text to columns."

This spreadsheet application automatically knew that the comma was a delimiter and separated each certification. But sometimes you might need to specify what the delimiter should be. You can do that here.

Split text to columns is also helpful for fixing instances of numbers stored as text. Sometimes values in your spreadsheet will seem like numbers, but they're formatted as text. This can happen when copying and pasting from one place to another or if the formatting's wrong. For this example, let's check out our new spreadsheet from a cosmetics maker. If a data analyst wanted to determine total profits, they could add up everything in column F. But there's a problem; one of the cells has an error. If you check into it, you learn that the "707" in this cell is text and can't be changed into a number. When the spreadsheet tries to multiply the cost of the product by the number of units sold, it's unable to make the calculation. But if we select the orders column and choose "Split text to columns,"

The error is resolved because now it can be treated as a number. Coming up, you'll learn about a tool that does just the opposite. CONCATENATE is a function that joins multiple text strings into a single string. Spreadsheets are a very important part of data analytics. They save data analysts time and effort and help us eliminate errors each and every day. Here, you've learned about some of the most common tools that we use. But there's a lot more to come. Next, we'll learn even more about data cleaning with spreadsheet tools. Bye for now!

[**OPTIMIZE THE DATA-CLEANING PROCESS**](https://www.coursera.org/learn/process-data/lecture/ohiCl/optimize-the-data-cleaning-process)

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

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

### **What you’ll need**

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

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

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

OR

[International Logistics Association Memberships - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/p1JSr3UeTOOi_2-3Yr67dg_2e42a0117ecc42978ff39c15b0fbd3f1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1712188800&Signature=GvOGRX9dfZWsL0WEd2bD9yUBW2XxZ0GvHT8nbmmKSezQN-MfE4Dx7-B1cdiPqhMNw09tFX6F0AK-Ko~-FbhEeT-ds-emrY6EWNlvrpTQyy-GCMIe-xpFKPKQYGHyWLyTIbGrvm2syjqUI8CFAhlI8OEncjQR2TZiuuXw1zYtBgY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/p1JSr3UeTOOi_2-3Yr67dg_2e42a0117ecc42978ff39c15b0fbd3f1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1712188800&Signature=GvOGRX9dfZWsL0WEd2bD9yUBW2XxZ0GvHT8nbmmKSezQN-MfE4Dx7-B1cdiPqhMNw09tFX6F0AK-Ko~-FbhEeT-ds-emrY6EWNlvrpTQyy-GCMIe-xpFKPKQYGHyWLyTIbGrvm2syjqUI8CFAhlI8OEncjQR2TZiuuXw1zYtBgY_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[Cosmetics Inc. - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/htYeVDpXRHCUSNSclnWcqQ_06ea60f3250b4ba2a59e4fc7d2e2a5e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1712188800&Signature=HpjUdP4P2BxLAyZtxVLHeWXUQGZSltBgFLxwTFCK0pOQeEJMZJfcC152SODpd2eK0fUa8orOyPm4RYV5oZQd7trtGO1-8lyb5sdbTz2Jp36i3XXtU8XbeR-IMvFXpFK6~02yY0vTgTNYcezk5jjzZTgscLIVRNLFbFnMFdA1nFg_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/htYeVDpXRHCUSNSclnWcqQ_06ea60f3250b4ba2a59e4fc7d2e2a5e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1712188800&Signature=HpjUdP4P2BxLAyZtxVLHeWXUQGZSltBgFLxwTFCK0pOQeEJMZJfcC152SODpd2eK0fUa8orOyPm4RYV5oZQd7trtGO1-8lyb5sdbTz2Jp36i3XXtU8XbeR-IMvFXpFK6~02yY0vTgTNYcezk5jjzZTgscLIVRNLFbFnMFdA1nFg_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[](https://d3c33hcgiwev3.cloudfront.net/htYeVDpXRHCUSNSclnWcqQ_06ea60f3250b4ba2a59e4fc7d2e2a5e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1712188800&Signature=HpjUdP4P2BxLAyZtxVLHeWXUQGZSltBgFLxwTFCK0pOQeEJMZJfcC152SODpd2eK0fUa8orOyPm4RYV5oZQd7trtGO1-8lyb5sdbTz2Jp36i3XXtU8XbeR-IMvFXpFK6~02yY0vTgTNYcezk5jjzZTgscLIVRNLFbFnMFdA1nFg_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

## 

## **Example 1: The COUNTIF function**

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

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

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

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

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

## **Example 2: The LEN function**

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

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

## **Example 3: Use conditional formatting**

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

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

## **Example 4: The LEFT and RIGHT functions**

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

### **The LEFT function**

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

### **The RIGHT function**

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

## **Example 5: The MID function**

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

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

## **Example 6: The CONCATENATE function**

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

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

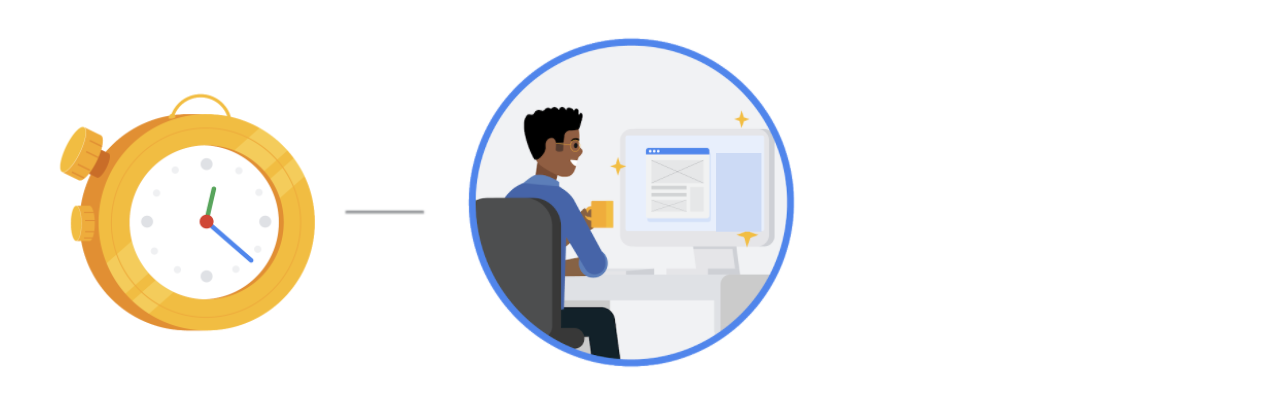
## **Example 7: TRIM function**

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

1. Select cell **L1**, and enter **Trim**.
2. In cell **L2**, enter **=TRIM(C2)** to remove any leading, trailing, or repeated spaces.
3. Double-click the fill handle and to automatically populate the rest of this column.

[**WORKFLOW AUTOMATION**](https://www.coursera.org/learn/process-data/supplement/QRQsR/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)
* Technology Advice’s [**10 of the Best Options for Workflow Automation Software**](https://technologyadvice.com/blog/information-technology/top-10-workflow-automation-software/)

## **Key takeaways**

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

[**STEP-BY-STEP: DIFFERENT DATA PERSPECTIVES**](https://www.coursera.org/learn/process-data/supplement/OYAXi/step-by-step-different-data-perspectives)

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

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

### **WHAT YOU’LL NEED**

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

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

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

OR

[Cosmetics Inc. - Data for Pivot Table and VLOOKUP](https://d3c33hcgiwev3.cloudfront.net/mjJpNpPfRiC7blak2QdsSA_f5269cdfb2da42ddb06577759b173be1_Cosmetics-Inc.---Data-for-Pivot-Table-and-VLOOKUP.xlsx?Expires=1712188800&Signature=g2quGyXB4H7YeaAfhLyBOVLb-2gmvh0~CkvzwvfvaPU~OrDDLr7qR1bRoSwXb609yr2kJWni6Ez-oPAbbiNLNTblbuUO8j2cLrCoOVoVgbeSokE3ryQ5KfM0iDX8KFb9LC1z830uEkmpTV3mMq8MjxgtBkR0gyZD4x8R4aFqCG8_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/mjJpNpPfRiC7blak2QdsSA_f5269cdfb2da42ddb06577759b173be1_Cosmetics-Inc.---Data-for-Pivot-Table-and-VLOOKUP.xlsx?Expires=1712188800&Signature=g2quGyXB4H7YeaAfhLyBOVLb-2gmvh0~CkvzwvfvaPU~OrDDLr7qR1bRoSwXb609yr2kJWni6Ez-oPAbbiNLNTblbuUO8j2cLrCoOVoVgbeSokE3ryQ5KfM0iDX8KFb9LC1z830uEkmpTV3mMq8MjxgtBkR0gyZD4x8R4aFqCG8_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[](https://d3c33hcgiwev3.cloudfront.net/mjJpNpPfRiC7blak2QdsSA_f5269cdfb2da42ddb06577759b173be1_Cosmetics-Inc.---Data-for-Pivot-Table-and-VLOOKUP.xlsx?Expires=1712188800&Signature=g2quGyXB4H7YeaAfhLyBOVLb-2gmvh0~CkvzwvfvaPU~OrDDLr7qR1bRoSwXb609yr2kJWni6Ez-oPAbbiNLNTblbuUO8j2cLrCoOVoVgbeSokE3ryQ5KfM0iDX8KFb9LC1z830uEkmpTV3mMq8MjxgtBkR0gyZD4x8R4aFqCG8_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

## 

## **Example 1: Pivot tables**

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

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

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

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

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

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

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

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

## **Example 2: VLOOKUP**

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

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

## 

## **Example 3: Plotting**

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

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

[**STEP-BY-STEP: EVEN MORE DATA-CLEANING TECHNIQUES**](https://www.coursera.org/learn/process-data/supplement/DFobu/step-by-step-even-more-data-cleaning-techniques)

This reading outlines the steps the instructor performs in the next video, [Even more data-cleaning techniques](https://www.coursera.org/learn/process-data/lecture/Ei2IH/even-more-data-cleaning-techniques). This video teaches you different methods data analysts use in data mapping. Data mapping is the process of matching fields from one database to another. It’s critical to the success of data migration, data integration, and many other data-management activities. This video contains one activity for you to practice.

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

### **What you’ll need**

If you’d like to follow along with the example in this video, choose a spreadsheet tool, such as Google Sheets or Excel.

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

Link to templates:

[International Logistics Association memberships](https://docs.google.com/spreadsheets/d/1th1mvhFtcg7hW_pDfWv1GaOxFCoIQLOyH7l3ab_W8a4/template/preview#gid=1764281342)

[Global Logistics Association](https://docs.google.com/spreadsheets/d/186Yx3S-ejZr1cJunsal2tUV9J1abixhBUDjXbUtQT7I/template/preview)

[Logistics Association Merger](https://docs.google.com/spreadsheets/d/16tkN0wpaSC3PlxQTGnOGyHWZy9Ui0C2NOIWxBY_gAv8/template/preview#gid=1764281342)

Downloads:

[International Logistics Association memberships](https://d3c33hcgiwev3.cloudfront.net/xSEEsOv6TRa77FfljFKFyQ_c0b53b24e0b54f6fa5d83c5c8683a1e1_International-Logistics-Association-memberships.xlsx?Expires=1712188800&Signature=YNVU-TShNQzpV4ZISi7JWVIezS2FFN2T0sACZ~wseVBoi5osKkl9AZ9YWDqEp329-4vp-bUJXeeXlOQfoAnSlYAKupIdozXU5qBb28PLY0FaQfGBvzWKouXdpsnpVSBkSngeVjVMxnvGWq4eEN2ZqmNQvSBoO5KHHqtPnf0m4XM_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/xSEEsOv6TRa77FfljFKFyQ_c0b53b24e0b54f6fa5d83c5c8683a1e1_International-Logistics-Association-memberships.xlsx?Expires=1712188800&Signature=YNVU-TShNQzpV4ZISi7JWVIezS2FFN2T0sACZ~wseVBoi5osKkl9AZ9YWDqEp329-4vp-bUJXeeXlOQfoAnSlYAKupIdozXU5qBb28PLY0FaQfGBvzWKouXdpsnpVSBkSngeVjVMxnvGWq4eEN2ZqmNQvSBoO5KHHqtPnf0m4XM_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[Global Logistics Association](https://d3c33hcgiwev3.cloudfront.net/tMtjHQztTiesjo8vTKb8UA_a5b6cb5cf2404d9abe4b7c3797da94e1_Global-Logistics-Association.xlsx?Expires=1712188800&Signature=jG~oeclQERHTAPlhtU1p3znhGEeH1t6aG6ngbTuhckriAqPLSARCZu97zOIZhVBCIeOVADMPEWTo-YBs-wZKko4~hHSuYzeuEIsGavPMiGILYY41FvHvUuLcSxpRHzXyUUA5w44zuhN-363QsqTaJwH9DYfB5pduQ2skwZOlVHA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/tMtjHQztTiesjo8vTKb8UA_a5b6cb5cf2404d9abe4b7c3797da94e1_Global-Logistics-Association.xlsx?Expires=1712188800&Signature=jG~oeclQERHTAPlhtU1p3znhGEeH1t6aG6ngbTuhckriAqPLSARCZu97zOIZhVBCIeOVADMPEWTo-YBs-wZKko4~hHSuYzeuEIsGavPMiGILYY41FvHvUuLcSxpRHzXyUUA5w44zuhN-363QsqTaJwH9DYfB5pduQ2skwZOlVHA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[Logistics Association Merger](https://d3c33hcgiwev3.cloudfront.net/rCJ95nZHTpCqGu0S0vglvg_a244e9e881604d688bef52d5388febe1_Logistics-Association-Merger.xlsx?Expires=1712188800&Signature=I9mshRPsIHSUmsVDWw8FLfAEinEjwVs86nk01Blp~tISsk5tQt3vs~gc54LhY7UM-Kr4AeIlw~onckgeV1dTFBarFBU5meCLWWdtLfsjFMSIx3L5wn-bn-hu8GUASnEZJ2shVI0HTqvIlJN95J0hbYCOoP~dCSiZQfwmbO30uqs_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/rCJ95nZHTpCqGu0S0vglvg_a244e9e881604d688bef52d5388febe1_Logistics-Association-Merger.xlsx?Expires=1712188800&Signature=I9mshRPsIHSUmsVDWw8FLfAEinEjwVs86nk01Blp~tISsk5tQt3vs~gc54LhY7UM-Kr4AeIlw~onckgeV1dTFBarFBU5meCLWWdtLfsjFMSIx3L5wn-bn-hu8GUASnEZJ2shVI0HTqvIlJN95J0hbYCOoP~dCSiZQfwmbO30uqs_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[](https://d3c33hcgiwev3.cloudfront.net/rCJ95nZHTpCqGu0S0vglvg_a244e9e881604d688bef52d5388febe1_Logistics-Association-Merger.xlsx?Expires=1712188800&Signature=I9mshRPsIHSUmsVDWw8FLfAEinEjwVs86nk01Blp~tISsk5tQt3vs~gc54LhY7UM-Kr4AeIlw~onckgeV1dTFBarFBU5meCLWWdtLfsjFMSIx3L5wn-bn-hu8GUASnEZJ2shVI0HTqvIlJN95J0hbYCOoP~dCSiZQfwmbO30uqs_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

## 

## **Example: CONCATENATE**

**CONCATENATE** is a function that joins together two or more text strings. In the video, you’ll learn how to use **CONCATENATE** to clean data after two datasets have been combined.

1. Open the dataset spreadsheet titled **Global Logistics Association**. When prompted, select **USE TEMPLATE**.
2. **Insert** a new column to the right of column **E**. Label it **New Address** in cell **F1**.
3. In the second row of the new column (cell **F2**), enter **=CONCATENATE (D2,E2)** and press **Enter**.
   1. You will notice that some results need a space between the street address and the unit or suite number, such as: 25 Dyas RdSte. 101.
   2. You could manually clean the data later to add a space between Rd and Ste., but **CONCATENATE** can actually do it for you.
   3. The **CONCATENATE** formula can help you format the data as it is merged by entering an additional string to insert a space between Rd and Ste.
   4. Enter **=CONCATENATE(D2, " ", E2)** and you will have an address that is formatted like this: 25 Dyas Rd Ste. 101. Much better!
4. Ensure the new data in the cell accurately reflects the merging of the two previous columns.
5. **Select cell F2** and **drag down** to apply the formula to all rows in the column.

Mark as completed

Like

Dislike

Report an issue

[**EVEN MORE DATA-CLEANING TECHNIQUES**](https://www.coursera.org/learn/process-data/lecture/Ei2IH/even-more-data-cleaning-techniques)

So far you've learned about a lot of different tools and functions that analysts use to clean up data for analysis. Now we'll take a step back and talk about some of the really big picture aspects of clean data. Knowing how to fix specific problems, either manually with spreadsheet tools, or with functions, is extremely valuable. But it's also important to think about how your data has moved between systems and how it's evolved along it's journey to your data analysis project. To do this, data analysts use something called data mapping. Data mapping is the process of matching fields from one database to another. This is very important to the success of data migration, data integration, and lots of other data management activities. As you learned earlier, different systems store data in different ways. For example, the state field in one spreadsheet might show Maryland spelled out. But another spreadsheet might store it as MD.

Data mapping helps us note these kinds of differences so we know when data is moved and combined it will be compatible. As a quick reminder, compatibility describes how well two or more data sets are able to work together. The first step to data mapping is identifying what data needs to be moved. This includes the tables and the fields within them. We also need to define the desired format for the data once it reaches its destination. To figure out how this works let's go back to the merger between our two logistics associations.

Starting with the first data field, we'll identify that we need to move both sets of member IDs. To define the desired format, we'll choose whether to use numbers like this spreadsheet, or email addresses like the other spreadsheet. Next comes mapping the data. Depending on the schema and number of primary and foreign keys in a data source, data mapping can be simple or very complex.

As a reminder, a schema is a way of describing how something is organized. A primary key references a column in which each value is unique and a foreign key is a field within a table that is a primary key in another table. For more challenging projects there's all kinds of data mapping software programs you can use. These data mapping tools will analyze field by field how to move data from one place to another then they automatically clean, match, inspect, and validate the data. They also create consistent naming conventions, ensuring compatibility when the data is transferred from one source to another. When selecting a software program to map your data, you want to be sure that it supports the file types you're working with, such as Excel, SQL, Tableau, and others.

Later on, you'll learn more about selecting the right tool for a particular task. For now, let's practice mapping data manually. First, we need to determine the content of each section to make sure the data ends up in the right place. For example, the data on when memberships expire would be consolidated into a single column. This step makes sure that each piece of information ends up in the most appropriate place in the merged data source. Now, you might remember that some of the data was inconsistent between the two organizations, like the fact that one uses a separate column for suite apartment or unit number but the other doesn't.

This brings us to the next step, transforming the data into a consistent format. This is a great time to use concatenate. As you learned before, concatenate is a function that joins together two or more text strings, which is what we did earlier with our cosmetics company example.

We'll insert a new column and then type equals concatenate, then the two text strings we want to make one.

Drag that through the entire column.

Now we have the consistency in the new merged association lists of member addresses.

Now that everything's compatible, it's time to transfer the data to its destination.

There's a lot of different ways to move data from one place to another, including querying, importing wizards, and even simple drag and drop. Here's our merged spreadsheet.

It looks good, but we still want to make sure everything was transferred properly. We'll go into the testing phase of data mapping. For this, you inspect a sample piece of data to confirm that it's clean and properly formatted.

It's also a smart practice to do spot checks on things such as the number of nulls. For the test, you can use a lot of the data cleaning tools we discussed previously, such as data validation, conditional formatting, COUNTIF, sorting, and filtering.

Finally, once you've determined that the data is clean and compatible, you can start using it for analysis. Data mapping is so important because even one mistake when merging data can ripple throughout an organization, causing the same error to appear again and again. This leads to poor results. On the other hand, data mapping can save the day by giving you a clear road map you can follow to make sure your data arrives safely at its destination. That's why you learn how to do it.

[**HANDS-ON ACTIVITY: CLEAN DATA WITH SPREADSHEET FUNCTIONS**](https://www.coursera.org/learn/process-data/quiz/YzKhK/hands-on-activity-clean-data-with-spreadsheet-functions)

**Activity Overview**

****

By now, you’ve been introduced to some useful techniques for cleaning spreadsheet data, such as sorting and filtering. In this activity, you'll continue to develop your data-cleaning skills by using spreadsheet functions.

### Scenario

Review the following scenario. Then complete the step-by-step instructions.

Imagine you are a data analyst working for a marketing agency based in San Francisco. The marketing agency wants to contact local boba tea shops to inquire about a potential collaboration for a new marketing campaign. The agency plans to visit the top-rated shops within a 10-mile radius of the center of their target area. To assist with planning, the agency asks your team to review external data related to ratings and locations of boba tea shops in San Francisco. One of your teammates has created a spreadsheet from an online source. However, the data is not in the greatest shape.

Your assignment is to identify the dirty elements in the dataset and clean them up.

By the time you complete this activity, you will be able to identify dirty elements in a dataset, remove duplicate data, and use the **COUNTIF** and **SPLIT** functions to help clean data.

### Step-By-Step Instructions

Follow the instructions to complete each step of the activity. Then answer the questions at the end of the activity before going to the next course item.

### **Step 1: Access the template**

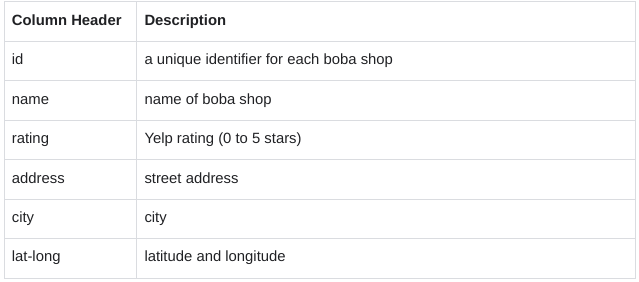
To get started, access the spreadsheet that contains the data. Click the link and make a copy of the [spreadsheet](https://docs.google.com/spreadsheets/d/1ETb45bbtIn-q3Z-eps9cw66GgS2Gye8cOfKIoFi65DU/template/preview).

Or, if you don’t have a Google account, you may download the dataset directly from the attachment below:

[San Francisco Boba Tea Shop Location Info](https://d3c33hcgiwev3.cloudfront.net/1G5STeL7RCeuUk3i-0QnEg_026c7c0c9f3b47efbd599d534b0937f1_San-Francisco-Boba-Tea-Shop-Location-Info.csv?Expires=1712188800&Signature=ZYuIVQIwR3UbeWLSjrZFbdipKfKesudKFSpTsKme3wvYiFv-fe5b2hVAfOa8cHlFxIiovlnIR5O9HMb1qzMHonL14DvqWlext5tVsWhPu-XjutgTLV2dvlCozBVLiZs-92zZC~ilUVKk1lNk9sBAH0bjlmqovQlQFa5t0yCjYEM_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

[CSV File](https://d3c33hcgiwev3.cloudfront.net/1G5STeL7RCeuUk3i-0QnEg_026c7c0c9f3b47efbd599d534b0937f1_San-Francisco-Boba-Tea-Shop-Location-Info.csv?Expires=1712188800&Signature=ZYuIVQIwR3UbeWLSjrZFbdipKfKesudKFSpTsKme3wvYiFv-fe5b2hVAfOa8cHlFxIiovlnIR5O9HMb1qzMHonL14DvqWlext5tVsWhPu-XjutgTLV2dvlCozBVLiZs-92zZC~ilUVKk1lNk9sBAH0bjlmqovQlQFa5t0yCjYEM_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A)

The dataset includes the following column headers:



### 

### **Step 2: Identify the dirty elements in your data**

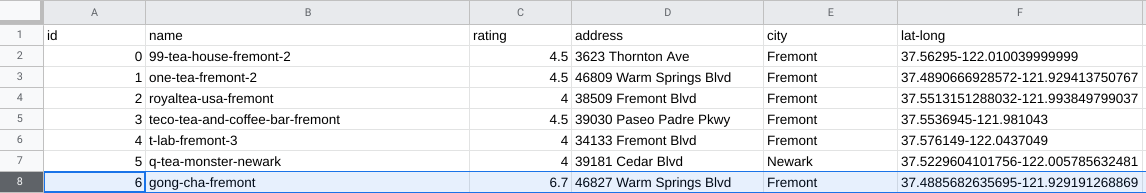
As a data analyst, your job is to present data that is readable, accurate, and visually appealing. Cleaning your data helps you achieve this goal. The first step is to identify the dirty elements in your data.

1. Rename your spreadsheet. Click Untitled Spreadsheet and enter a new name. You can use the name sf\_boba\_tea\_shop\_data or a similar name that describes the data your spreadsheet contains.
2. If you want to get a better view of your data, you can make the columns wider by dragging the right boundary of the column heading. This may apply to the name (B), address (D), and lat-long (F) columns.
3. Now, review your data and consider any problems you may need to address. The following are examples of errors that you can quickly identify and fix. This is not a comprehensive list of every potential problem, but is a great starting point for data cleaning.

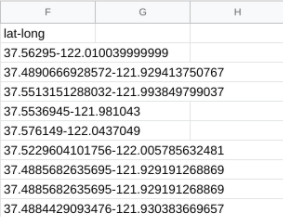
* First, there is at least one duplicate line (rows 20 and 21) in your dataset.



* Second, all Yelp ratings should fall between 0 and 5. However, at least one rating (in cell C8) falls outside of that range.



* Finally, the data for latitude and longitude is contained in a single column (F). In order for someone to be able to use this data for analysis, the two values should be in separate columns.



Now you know what issues to focus your attention on during the cleaning process.

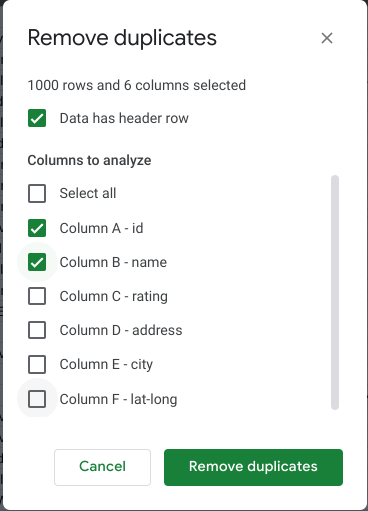
Your goal is to fix these errors and help create a clean dataset for analysis. You can address each issue in turn.

### 

### **Step 3: Remove duplicates**

The first step is to eliminate any duplicate entries from your dataset. As a best practice, duplicates should be removed even if they are not readily apparent.

1. To start, select columns A through F.
2. Then, in the menu bar, choose Data, then Data Cleanup, and select Remove duplicates.
3. In the pop-up window, click Data has header row. You want to remove duplicate boba shop id's and boba shop names. In the Columns to analyze section, make sure the relevant columns (id, name) are selected.



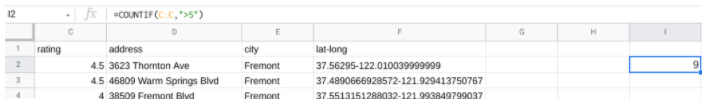
1. Once everything has been selected, click Remove duplicates.
2. If done correctly, 3 duplicate rows will be found and removed and 604 rows will remain.

### 

### **Step 4: Correct the ratings data**

Next, clean up any data that does not make sense. Yelp ratings should be less than 5 and greater than 0. Now, you will determine how many entries are inaccurate and correct them. You can use the **COUNTIF** function to perform this task.

1. The **COUNTIF** function quickly counts how many items in a range of cells meet a given criterion. In cell I2, enter **=COUNTIF(C:C,">5")**. The first entry (C:C) refers to the range where you are counting the data. In this case, the range is the entire rating column (C), which contains the Yelp ratings. The second entry refers to the criterion (>5), and tells the function to count all the values greater than 5.
2. Press Enter. You’ll notice that the function returns a value of 9. This tells you that your dataset contains 9 entries that have a rating greater than 5.

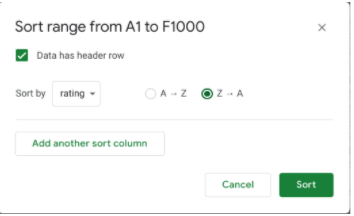


As a data analyst, it's your job to decide what to do with incorrect values or to ask the dataset owner for advice if you’re unsure. In this case, one effective approach would be to search on Yelp for the actual ratings. For this activity, you can just replace the incorrect ratings with the number 5. An efficient way to replace the ratings is to sort the data numerically from largest to smallest rating.

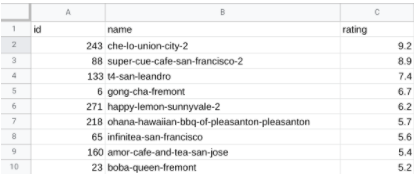
3. Select columns A through F.

4. Then, from the menu bar, choose Data, then Sort range, and select Advanced range sorting options.

5. In the pop-up window, check the box next to Data has header row. Sort by rating from Z →A. This way, the highest ratings will be listed first.



6. Click Sort. Check out your spreadsheet. At the start of the rating column, you should now find the 9 rows that have incorrect values (rating > 5).



7. Next, select the range of cells C2:C10. Press delete to delete the values that are greater than 5.

8. Replace all the values with the number 5. In cell C2, enter 5. Then, drag the fill handle down to cell C10 to fill the remaining cells with 5.

9. After replacing the incorrect ratings with the number 5, you may notice that the new value in cell I2 is 0. The output of the **COUNTIF** function now reflects the changes in your dataset. This confirms that the rating column no longer contains any values greater than 5.

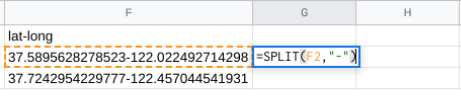
10. FInally, delete the formula from cell I2 since you don’t need this information anymore.

### 

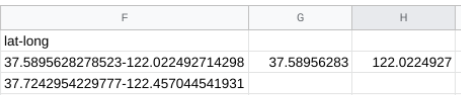
### **Step 5: Clean up the latitude and longitude data**

Next, clean up the latitude and longitude data by placing each value in a separate column. You can use the **SPLIT** function to accomplish this task.

1. The **SPLIT** function divides text around a specified character or string, and puts each fragment of text into a separate cell in the row. The **SPLIT** function will split the single lat-long column into two separate columns, one for latitude and the other for longitude. In cell G2, enter **=SPLIT(F2,"-")**. The first entry (F2) refers to the cell where the text is located. The second entry (“-”) refers to the fact that you are dividing the text based on the minus sign.



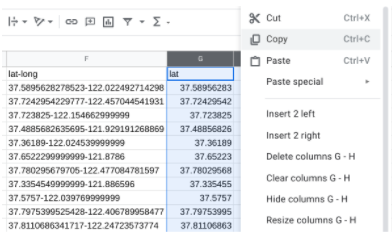
2. Press Enter. The result shows each fragment of text in a different cell.



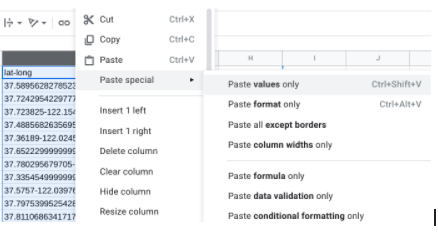
3. Select cell G2 again. In cell G2, double-click on the fill handle to split all the remaining lat-long entries.

4. Now add column headers to the two new columns (G and H). In cell G1, enter lat. In cell H1, enter long.

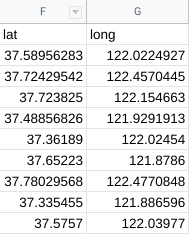
5. Next, replace the original lat-long data in column F with the new split entries in columns G and H. Select columns G and H, right-click, and choose Copy.



6. Then, select Column F, right-click, and choose Paste special and Paste values only.



7. Now the new lat column is column F, and the new long column is column G. Adjust the width of the lat column (F) to fit the data by dragging the right boundary of the column heading.



8. Next, select column H, right-click, and choose the Delete column.

9. Finally, the longitude values should be negative so that they are accurate coordinates for mapping. To make the values in the long column negative, multiply them by -1. In cell H2, enter **=G2\*-1**. The asterisk is the operator for multiplication. Press Enter.

10. Still in cell H2, double-click on the fill handle to fill in the rest of the values.

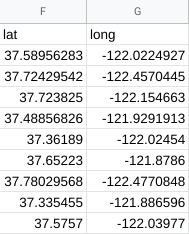
11. Next, add a column header. In cell H1, enter: long.

12. Now, replace the longitude data in column G with the new data in column H. Select column H, right-click, and choose Copy.

13. Select Column G, right-click, and choose Paste special and Paste values only.

14. Then, select column H, right-click, and choose the Delete column.

Columns F and G should look like this:



Now your data is cleaner, clearer, and easier to use.

### **Pro Tip: Save the activity template**

Be sure to save a copy of the spreadsheet template you used to complete this activity. You can use it for further practice or to help you work through your thought processes for similar tasks in a future data analyst role.

[**DEVELOP YOUR APPROACH TO CLEANING DATA**](https://www.coursera.org/learn/process-data/supplement/rKHUX/develop-your-approach-to-cleaning-data)

As you continue on your data journey, you’re likely discovering that data is often messy—and you can expect raw, primary data to be imperfect. In this reading, you’ll consider how to develop your personal approach to cleaning data. You will explore the idea of a cleaning checklist, which you can use to guide your cleaning process. Then, you’ll define your preferred methods for cleaning data. By the time you complete this reading, you’ll have a better understanding of how to methodically approach the data cleaning process. This will save you time when cleaning data and help you ensure that your data is clean and usable.

## 

## **Consider your approach to cleaning data**

Data cleaning usually requires a lot of time, energy, and attention. But there are two steps you can take before you begin to help streamline your process: **creating a cleaning checklist and deciding on your preferred methods**. This will help ensure that you know exactly how you want to approach data cleaning and what you need to do to be confident in the integrity of your data.

### 

### **Your cleaning checklist**

Start developing your personal approach to cleaning data by creating a checklist to help you identify problems in your data efficiently and identify the scale and scope of your dataset. Think of this checklist as your default “what to search for” list.

Here are some examples of common data cleaning tasks you could include in your checklist:

* **Determine the size of the dataset:** Large datasets may have more data quality issues and take longer to process. This may impact your choice of data cleaning techniques and how much time to allocate to the project.
* **Determine the number of categories or labels:** By understanding the number and nature of categories and labels in a dataset, you can better understand the diversity of the dataset. This understanding also helps inform data merging and migration strategies.
* **Identify missing data:** Recognizing missing data helps you understand data quality so you can take appropriate steps to remediate the problem. Data integrity is important for accurate and unbiased analysis.
* **Identify unformatted data:** Identifying improperly or inconsistently formatted data helps analysts ensure data uniformity. This is essential for accurate analysis and visualization.
* **Explore the different data types:** Understanding the types of data in your dataset (for instance, numerical, categorical, text) helps you select appropriate cleaning methods and apply relevant data analysis techniques.

There might be other data cleaning tasks you’ve been learning about that you also want to prioritize in your checklist. Your checklist is an opportunity for you to define exactly what you want to remember about cleaning your data; feel free to make it your own.

### **Your preferred cleaning methods**

In addition to creating a checklist, identify which actions or tools you prefer using when cleaning data. You’ll use these tools and techniques with each new dataset—or whenever you encounter issues in a dataset—so this list should be compatible with your checklist.

For example, suppose you have a large dataset with missing data. You’ll want to know how to check for missing data in larger datasets, and how you plan to handle any missing data, before you start cleaning. Outlining your preferred methods can save you lots of time and energy.

## **Key takeaways**

The data you encounter as an analyst won’t always conform to your checklist or your preferred actions and tools. But having these things can make common data cleaning tasks much easier to complete. As is so often the case, thoughtful planning sets up any project for success!

**MODULE 2 CHALLENGE**

[**GLOSSARY TERMS FROM COURSE 4 MODULE 2**](https://www.coursera.org/learn/process-data/supplement/itzXI/glossary-terms-from-module-2)

**Clean data:** Data that is complete, correct, and relevant to the problem being solved

**Compatibility:** How well two or more datasets are able to work together

**CONCATENATE:** A spreadsheet function that joins together two or more text strings

**Conditional formatting:** A spreadsheet tool that changes how cells appear when values meet specific conditions

**Data engineer:** A professional who transforms data into a useful format for analysis and gives it a reliable infrastructure

**Data mapping:** The process of matching fields from one data source to another

**Data merging:** The process of combining two or more datasets into a single dataset

**Data validation:** A tool for checking the accuracy and quality of data

**Data warehousing specialist:** A professional who develops processes and procedures to effectively store and organize data

**Delimiter:** A character that indicates the beginning or end of a data item

**Dirty data:** Data that is incomplete, incorrect, or irrelevant to the problem to be solved

**Duplicate data:** Any record that inadvertently shares data with another record

**Field length:** A tool for determining how many characters can be keyed into a spreadsheet field

**Incomplete data:** Data that is missing important fields

**Inconsistent data:** Data that uses different formats to represent the same thing

**Incorrect/inaccurate data:** Data that is complete but inaccurate

**LEFT:** A function that returns a set number of characters from the left side of a text string

**LEN:** A function that returns the length of a text string by counting the number of characters it contains

**Length:** The number of characters in a text string

**Merger:** An agreement that unites two organizations into a single new one

**MID:** A function that returns a segment from the middle of a text string

**Null:** An indication that a value does not exist in a dataset

**Outdated data:** Any data that has been superseded by newer and more accurate information

**Remove duplicates:** A spreadsheet tool that automatically searches for and eliminates duplicate entries from a spreadsheet

**Split:** A function that divides text around a specified character and puts each fragment into a new, separate cell

**Substring:** A smaller subset of a text string

**Text string:** A group of characters within a cell, most often composed of letters

**TRIM:** A function that removes leading, trailing, and repeated spaces in data

**Unique:** A value that can’t have a duplicate