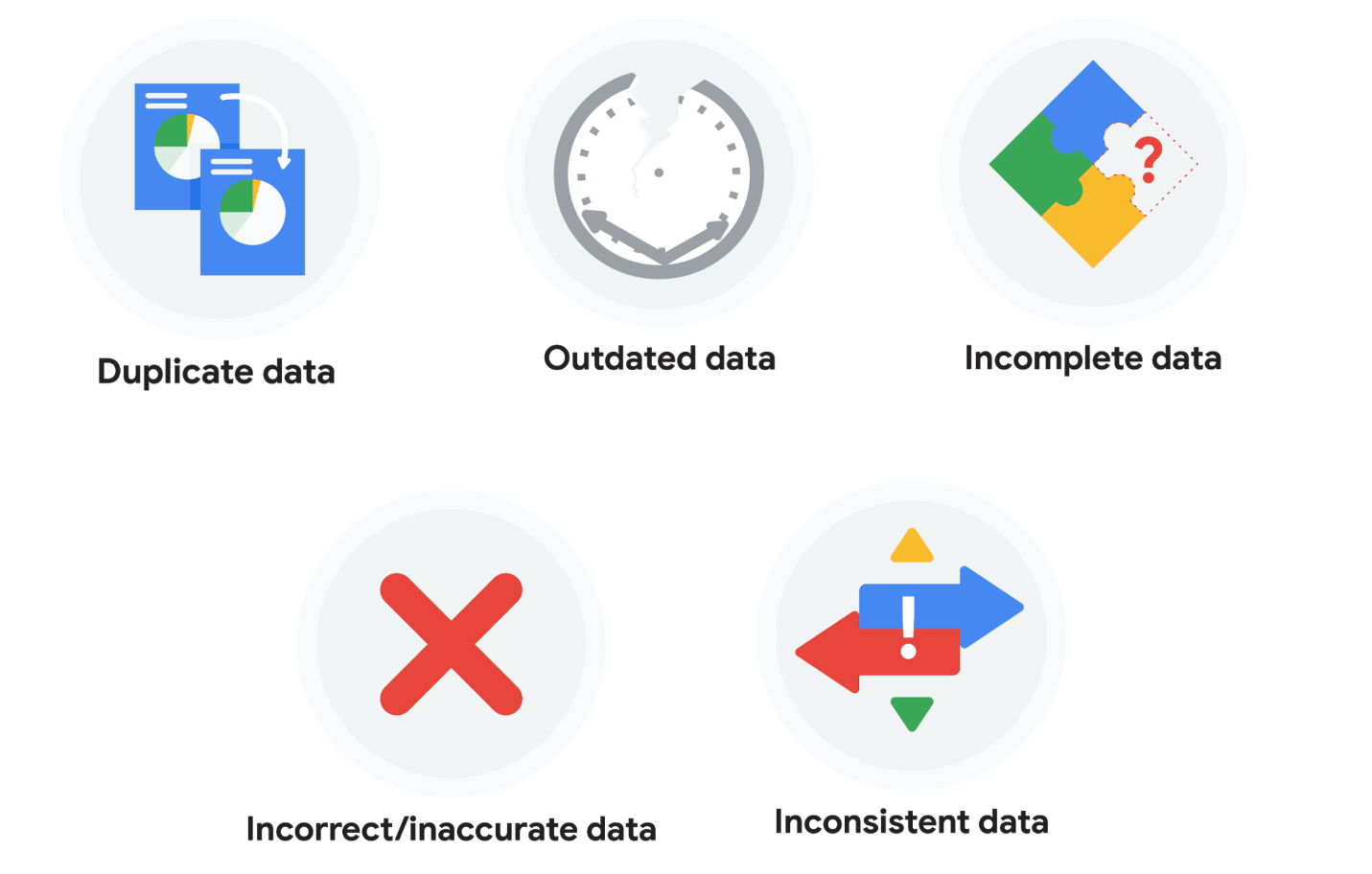
**What is dirty data?**

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**Outdated data**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

Validity

Definition

The concept of using data integrity principles to ensure measures conform to defined business rules or constraints

Example

Data collected five years ago used technology that is not approved or supported by the business

Accuracy

Definition

The degree of conformity of a measure to a standard or a true value

Example

Addresses in the business database are identified as incorrect when compared to the public postal service database

Completeness

Definition

The degree to which all required measures are known

Example

NULL/missing value for the item “Number of employees per store”

Consistency

Definition

The degree to which a set of measures is equivalent across systems

Example

Date of store opening stored in both MM/DD/YYYY and MM/YY formats

**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.

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

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

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

**What you’ll need**

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

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

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

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

OR

[International Logistics Association Memberships - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/5zlIemyvQtKsj4BGZQEdxA_7f286fde512b4f14b3246a6e68b333e1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1721952000&Signature=OLlGiefaxKXlV5-KT0DQexhEbV3E1NeMGDjj6jYACTop2-MOr-JpBV2gm1815bU62ob3-kjOFK4cf7Io-y1QNLiqZINZeKDYcEewVaRmq4eLw0sPW54szKfxk37dlmJgs2gnTjFpGswxSRikmyXZa3jjbXuzVZd96888RrCR8zU_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/5zlIemyvQtKsj4BGZQEdxA_7f286fde512b4f14b3246a6e68b333e1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1721952000&Signature=OLlGiefaxKXlV5-KT0DQexhEbV3E1NeMGDjj6jYACTop2-MOr-JpBV2gm1815bU62ob3-kjOFK4cf7Io-y1QNLiqZINZeKDYcEewVaRmq4eLw0sPW54szKfxk37dlmJgs2gnTjFpGswxSRikmyXZa3jjbXuzVZd96888RrCR8zU_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[Cosmetics Inc. - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/t3_O7KI_Rp6KVfCyGvTxGA_26e91b3aa1d44284956a4ae860d114e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1721952000&Signature=NeqyFquH0VXLYVx7RxcG8b8oUV-rESIzYMC~HEDlEaVOlklafYyIIyGNkmOx7Lw9C0D7ublujuTjd739KqINtwQs7CG8U4kau~kKDTQuTa7RI8Ijd0jJgMJXPiD7b1XwkMw9MdqyZb93XiGPYirXf5ANkaahGqoUiNbpkfNBSjA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/t3_O7KI_Rp6KVfCyGvTxGA_26e91b3aa1d44284956a4ae860d114e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1721952000&Signature=NeqyFquH0VXLYVx7RxcG8b8oUV-rESIzYMC~HEDlEaVOlklafYyIIyGNkmOx7Lw9C0D7ublujuTjd739KqINtwQs7CG8U4kau~kKDTQuTa7RI8Ijd0jJgMJXPiD7b1XwkMw9MdqyZb93XiGPYirXf5ANkaahGqoUiNbpkfNBSjA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



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

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

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

## Example 2: Remove duplicates

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

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

## Example 3: Format dates consistently

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

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

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

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

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

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

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

1. Open the spreadsheet [Cosmetics Inc. - Data for Cleaning](https://docs.google.com/spreadsheets/d/12U9Y4IVAGwml7XWBBgC4j9l0cCjqIZlqJc9vu3jr6Ig/template/preview?resourcekey=0-ds9iuh8tsuB7PwGd2dHMDA#gid=0).
2. Notice that cell **F12** contains an error.
3. Select **column E** (Orders).
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. This resolves the error in cell **F12**.

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

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

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

### **What you’ll need**

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

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

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

OR

[International Logistics Association Memberships - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/p1JSr3UeTOOi_2-3Yr67dg_2e42a0117ecc42978ff39c15b0fbd3f1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1721952000&Signature=Iw1kpkLr3bqYrtFZdQsZm26L8meCdeKsr8e0k7L7gMfQMbxacqpIM21rGCgGchuTePZIhy~vwaTqvKheLTtNsm2anN2jQ52ZkLZcmY5~jahm48R9vzht~ubb7zWSwd8NqomuunAHGRqW3SRSR~XioHBv1oWXR0VO7vUtkJqeNZw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/p1JSr3UeTOOi_2-3Yr67dg_2e42a0117ecc42978ff39c15b0fbd3f1_International-Logistics-Association-Memberships---Data-for-Cleaning.xlsx?Expires=1721952000&Signature=Iw1kpkLr3bqYrtFZdQsZm26L8meCdeKsr8e0k7L7gMfQMbxacqpIM21rGCgGchuTePZIhy~vwaTqvKheLTtNsm2anN2jQ52ZkLZcmY5~jahm48R9vzht~ubb7zWSwd8NqomuunAHGRqW3SRSR~XioHBv1oWXR0VO7vUtkJqeNZw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[Cosmetics Inc. - Data for Cleaning](https://d3c33hcgiwev3.cloudfront.net/htYeVDpXRHCUSNSclnWcqQ_06ea60f3250b4ba2a59e4fc7d2e2a5e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1721952000&Signature=kkVclzga5GyIuNM67p3iJZgXHr8ZcDitZpRdO00edRO7bf~5f21ueYTKp0n4Q0kBLCmLn5owgT9iwiuOOvti3l3HJ3GZpeWuGPRlbkqzCrsUlZSiszxBYxTvDaxBShN4pFQlK6NOw9shNxGdjfTFYLtbI2a3k3dAsRZOQXxrUBw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[XLSX File](https://d3c33hcgiwev3.cloudfront.net/htYeVDpXRHCUSNSclnWcqQ_06ea60f3250b4ba2a59e4fc7d2e2a5e1_Cosmetics-Inc.---Data-for-Cleaning.xlsx?Expires=1721952000&Signature=kkVclzga5GyIuNM67p3iJZgXHr8ZcDitZpRdO00edRO7bf~5f21ueYTKp0n4Q0kBLCmLn5owgT9iwiuOOvti3l3HJ3GZpeWuGPRlbkqzCrsUlZSiszxBYxTvDaxBShN4pFQlK6NOw9shNxGdjfTFYLtbI2a3k3dAsRZOQXxrUBw_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)



## Example 1: The COUNTIF function

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

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

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

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

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

## Example 2: The LEN function

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

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

## Example 3: Use conditional formatting

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

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

## Example 4: The LEFT and RIGHT functions

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

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

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

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

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

## Example 5: The MID function

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

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

## Example 6: The CONCATENATE function

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

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

## Example 7: TRIM function

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

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

# 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.

# Glossary terms from module 2

## **Terms and definitions for Course 4, Module 2**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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