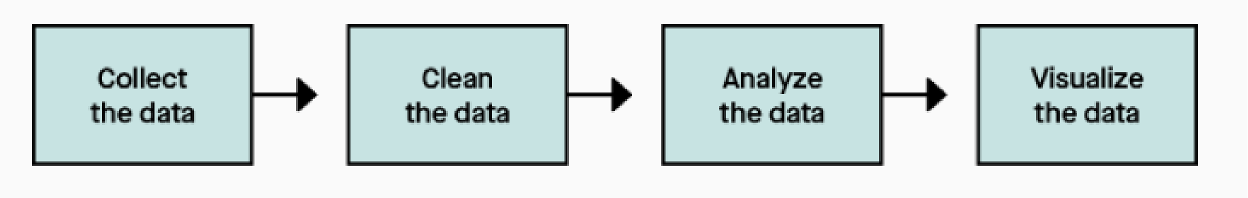
Cleaning and Formatting Data



This lesson demonstrates how to handle raw data that has formatting issues, inconsistencies, missing data, or duplicate data.

The first step is data collection, which is when raw data is initially collected or obtained. Raw data is data from the primary source that has had no processing or formatting done to it. Raw data is often messy and can't be used for analysis immediately. It can have a variety of issues, including the following:

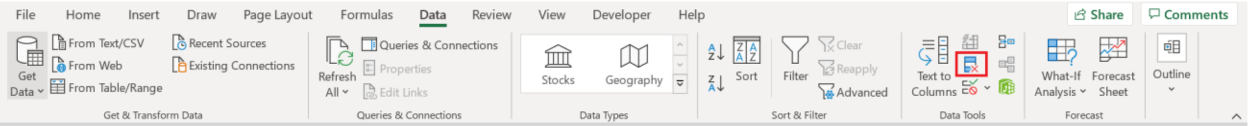
- Inaccurate data.
- Inconsistent formats throughout the data.
- Missing data due to optional fields or system outages.
- Unclear field definitions that are used inconsistently by users.
- Data that was entered wrong on purpose (fraud).

If raw data is used in analysis, it can lead to incorrect results; it's important to use clean data instead. So, in this lesson, we learn how to carry out the second step in the data analysis process: data cleaning.

One of the first things to check in a new dataset is the consistency and accuracy of data types. If a column has the header Temperature but the values are formatted as dates, the dataset won't make sense. Similarly, if a column has the header Number of students but contains some integers and some decimals, that may be confusing.

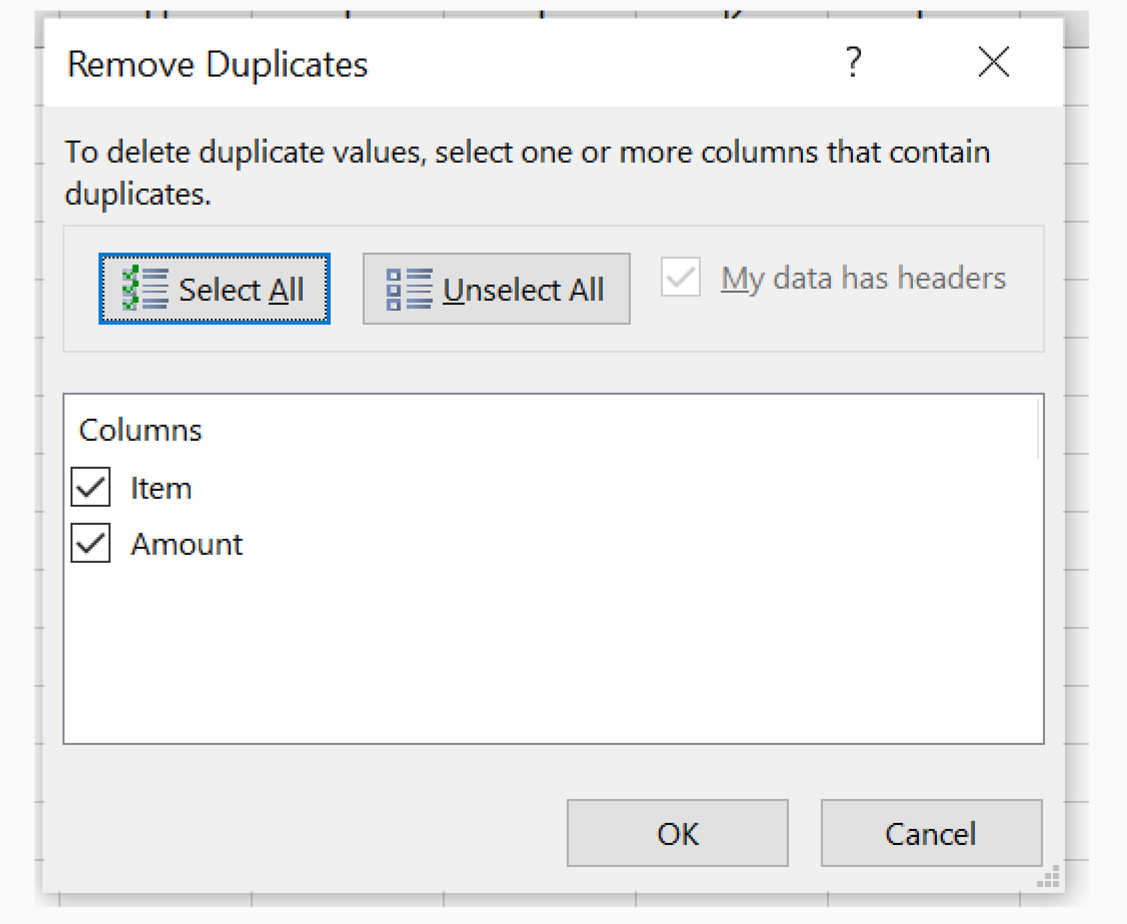
Removing duplicates

Fortunately, Excel has a tool that removes all the duplicates from your data at once. To use this tool, start by highlighting the data that you want to remove duplicates from. In this case, that's cells A2:B14. Then, in the Data tab of the ribbon, click the Remove Duplicates button.



That will bring up the Remove Duplicates dialog, which lets you choose one or more columns that have duplicates. If you select all fields in this dialog, Excel removes duplicate rows that have the same values across all columns. Or, if only one field is selected in this window, then Excel removes rows that have the same value for that particular column.

For this example, try selecting all fields so that only duplicates with the same item name and the same amount get removed. Assume that if the item name is the same but the amount is different, that represents two separate expenses.



Remove Duplicates

To delete duplicate values, select one or more columns that contain duplicates.

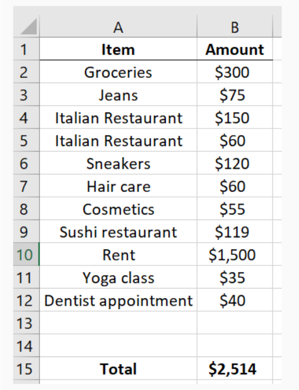
☑ My data has headers

Columns

☑ Item

☑ Amount

Click OK, and you'll see a dialog telling you that 2 duplicates were removed and 11 unique values remain.



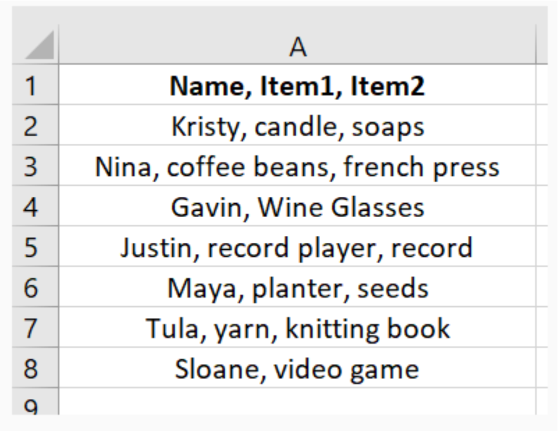
	A	B
1	**Item**	**Amount**
2	Groceries	$300
3	Jeans	$75
4	Italian Restaurant	$150
5	Italian Restaurant	$60
6	Sneakers	$120
7	Hair care	$60
8	Cosmetics	$55
9	Sushi restaurant	$119
10	Rent	$1,500
11	Yoga class	$35
12	Dentist appointment	$40
13		
14		
15	**Total**	**$2,514**

Notice that one of the Rent rows has been removed, as well as one of the Dentist appointment rows. Now the new total accurately reflects your expenditures for the last month: $2,514.

Splitting text into columns

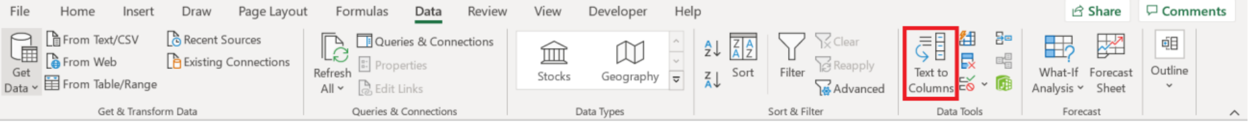
You may not always receive data as a tabular Excel file. There are lots of data file types, such as text files, XML files, CSV files, and Microsoft Access Database files, to name a few. When data is obtained in a different file type, you may need to do some work to make it compatible with Excel's data structure.

A common data file type is a comma-separated value file (CSV file). This file type contains data that looks like strings of text or numbers, and each piece of data is separated by commas , instead of separated into different cells. The example below shows a CSV file containing your friends' names and the holiday gifts that you've bought for each of them.

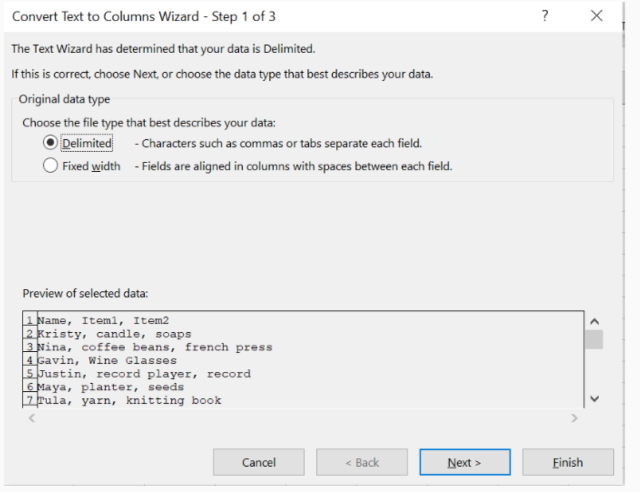


	A
1	**Name, Item1, Item2**
2	Kristy, candle, soaps
3	Nina, coffee beans, french press
4	Gavin, Wine Glasses
5	Justin, record player, record
6	Maya, planter, seeds
7	Tula, yarn, knitting book
8	Sloane, video game
9	

As you can see, the CSV data structure makes it hard to visually separate the names from the gift items. Excel has a useful tool called Text to Columns that splits data into separate cells, using a symbol of your choice or fixed length as the splitting location. To try it out, highlight the cells that need to be split up (A1:A8). Then click the Text to Columns button in the ribbon's Data tab.



You will see the following dialog:



Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

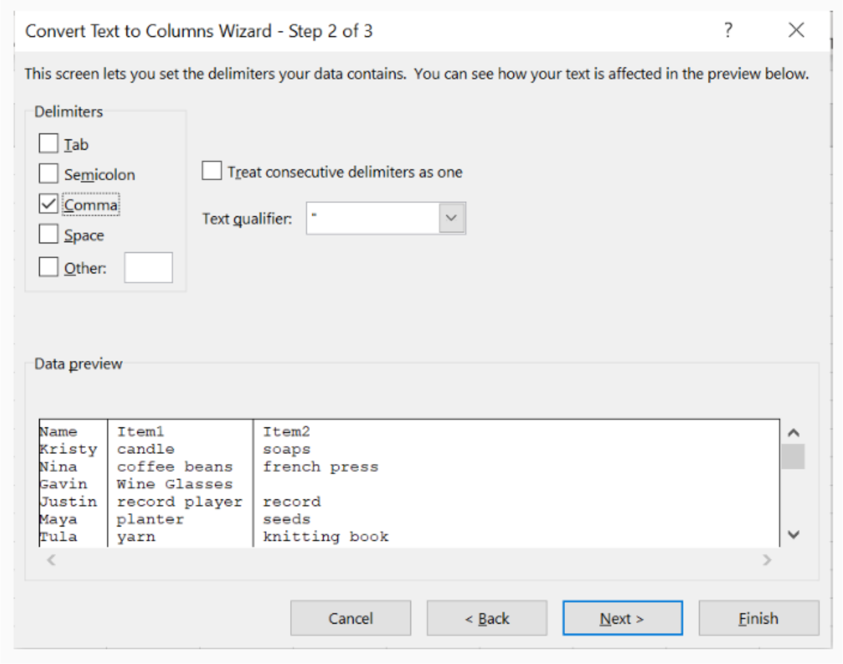
● Delimited - Characters such as commas or tabs separate each field.

○ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

```
1 Name, Item1, Item2
2 Kristy, candle, soaps
3 Nina, coffee beans, french press
4 Gavin, Wine Glasses
5 Justin, record player, record
6 Maya, planter, seeds
7 Tula, yarn, knitting book
```

Here, you can select if the separation happens at a particular character (Delimited) or after a certain number of characters (Fixed width). Because this is a CSV file, leave the Delimited option selected. Then click Next.



Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

- □ Tab
- □ Semicolon
- ☑ Comma
- □ Space
- □ Other:

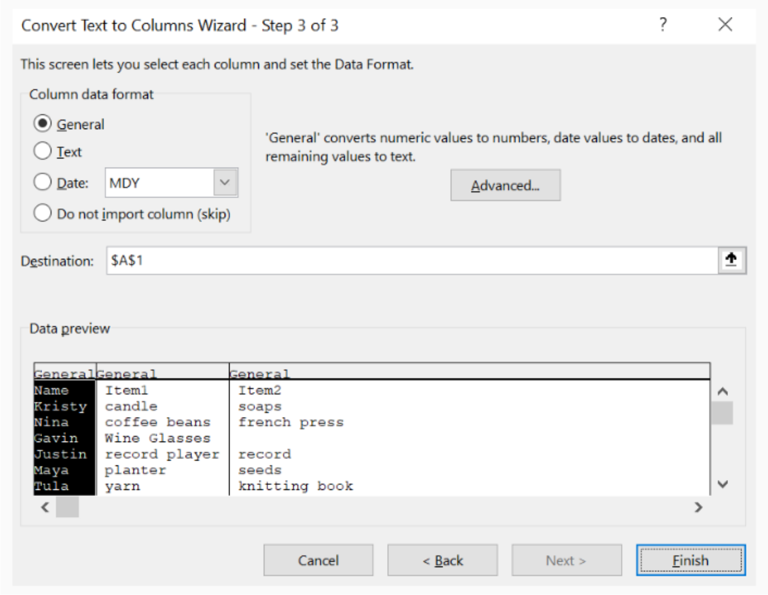
□ Treat consecutive delimiters as one

Text qualifier: "

Data preview

```
Name    Item1          Item2
Kristy  candle         soaps
Nina    coffee beans   french press
Gavin   Wine Glasses
Justin  record player  record
Maya    planter        seeds
Tula    yarn           knitting book
```

This second screen lets you choose which character triggers the separation. Select the Comma checkbox. The bottom of the dialog shows you a preview of how your data looks with your selection. Click Next.



Convert Text to Columns Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

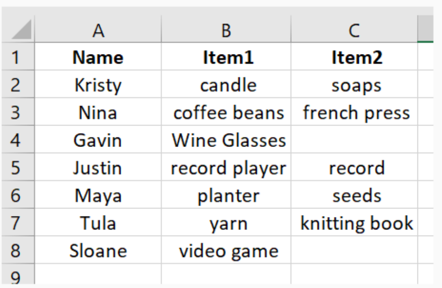
- ● General
- ○ Text
- ○ Date: MDY
- ○ Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Destination: A1

Data preview

```
GeneralGeneral        General
Name    Item1          Item2
Kristy  candle         soaps
Nina    coffee beans   french press
Gavin   Wine Glasses
Justin  record player  record
Maya    planter        seeds
Tula    yarn           knitting book
```



	A	B	C
1	**Name**	**Item1**	**Item2**
2	Kristy	candle	soaps
3	Nina	coffee beans	french press
4	Gavin	Wine Glasses	
5	Justin	record player	record
6	Maya	planter	seeds
7	Tula	yarn	knitting book
8	Sloane	video game	
9			

This final window lets you choose the data type for each column. Because you aren't changing any data types, you can press Finish. Now your data has been split out into separate cells and is much easier to look at!

Resolving null values

If data was collected using a survey and some of the questions were optional, the data may have some blanks or null values. A null value is a blank or empty cell in a set of data. The problem with null values is that they can impact the totals, averages, and other numeric calculations. If you take an average of 2, 4, and null, for example, you're potentially missing an important value that could change the average.

How are nulls dealt with? The favored way of dealing with nulls is imputing (filling in) the missing value. This is done by estimating a reasonable value for the null. Imputing can be done in several ways. First, check if there is any other record in the data that has similar values for the other columns. If there is, it can be used as a good estimator for the missing value. Other imputation methods include finding a relevant average, median, or mode. If none of these methods seem reasonable for your null value, then the last resort is deleting the record altogether.

Resolving null values requires a bit of exploration and creativity. It's best to find a reasonable estimate than to guess a value at random.

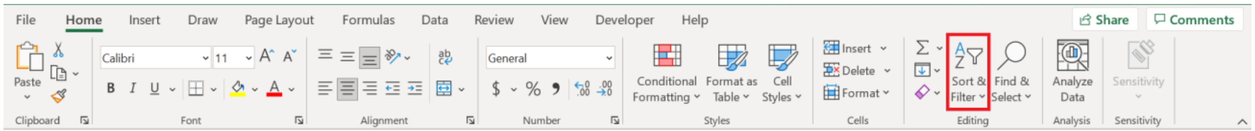
Sorting and filtering data

Sorting data

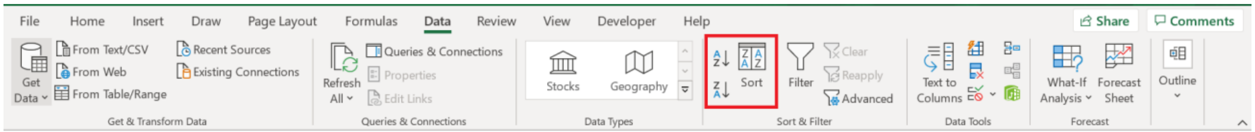
Now that we now how to clean data, we learn how to use Excel's sorting and filtering tools to systematically organize data and focus on relevant data.

Sorting data means to systematically organize it. You can organize your data alphabetically so that it's faster to find what you're looking for. Or you can organize your data in ascending or descending order for a particular column.

There are two places to find the Sort tool in the Excel ribbon. In the example above, you accessed this tool from the Home tab.



You can also access this tool from the Data tab, as shown below. Both Sort tools have the exact same functionalities, so you can use either one.

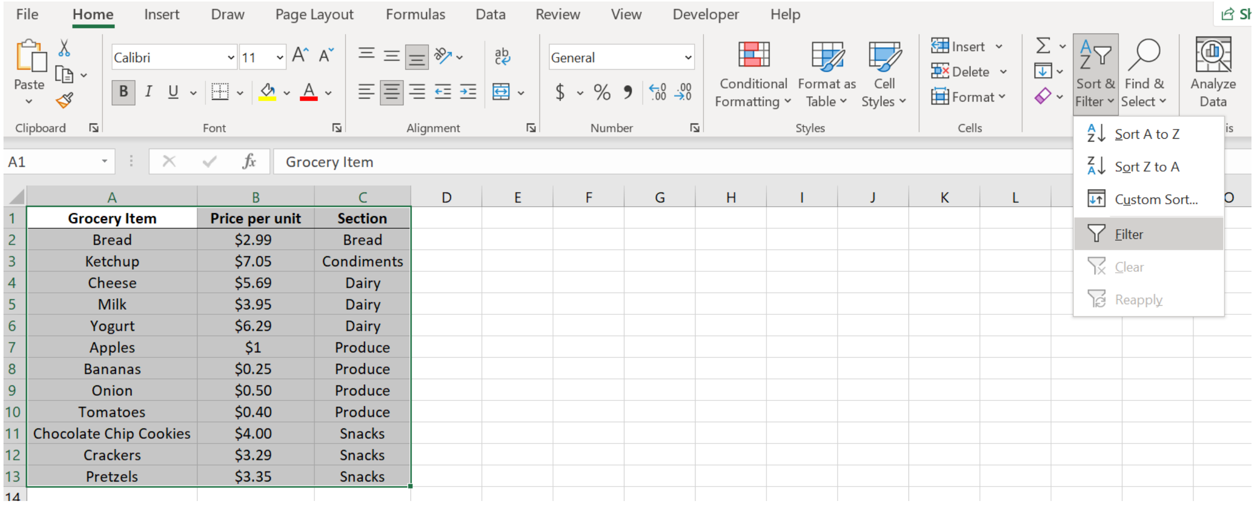


Filtering data

The Filter tool temporarily hides all data that you don't need, so that only relevant data is visible.

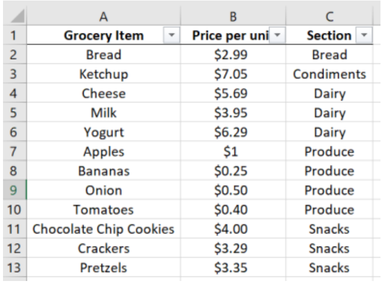
As with the Sort tool, there are two places where you can find the Filter tool in the ribbon. The first is in the Home tab, which is used in the example above.

The second location is in the Data tab, as shown below. Both Filter tools have the same functionalities, so you can use either one.



Grocery Item	Price per unit	Section
Bread	$2.99	Bread
Ketchup	$7.05	Condiments
Cheese	$5.69	Dairy
Milk	$3.95	Dairy
Yogurt	$6.29	Dairy
Apples	$1	Produce
Bananas	$0.25	Produce
Onion	$0.50	Produce
Tomatoes	$0.40	Produce
Chocolate Chip Cookies	$4.00	Snacks
Crackers	$3.29	Snacks
Pretzels	$3.35	Snacks

Select Filter. Your data table now has a small drop-down arrow next to each header or field name. Keep in mind that if your data doesn't have headers, you will see these arrows on the first row of the table.



Grocery Item	Price per uni	Section
Bread	$2.99	Bread
Ketchup	$7.05	Condiments
Cheese	$5.69	Dairy
Milk	$3.95	Dairy
Yogurt	$6.29	Dairy
Apples	$1	Produce
Bananas	$0.25	Produce
Onion	$0.50	Produce
Tomatoes	$0.40	Produce
Chocolate Chip Cookies	$4.00	Snacks
Crackers	$3.29	Snacks
Pretzels	$3.35	Snacks

You can filter items shown in the Section column by clicking the drop-down arrow next to Section. A filter menu will appear and show all the possible values in that column.

Remember that filtering only hides irrelevant data temporarily. To clear the filter, you can click the drop-down next to Section and select Clear Filter From Section. Or you can click the Sort & Filter button again and then Clear.

Formatting data

This lesson demonstrates ways to format data in Excel in order to make that data visually easier to work with.

When you're working with data in a table structure, patterns and key information can get lost in the sea of values. Formatting data in Excel gives you control over how the information is presented. It also helps you orient yourself. You can even tell Excel how to format data depending on if the data meets a condition that you specify.

<u>Conditional formatting</u>

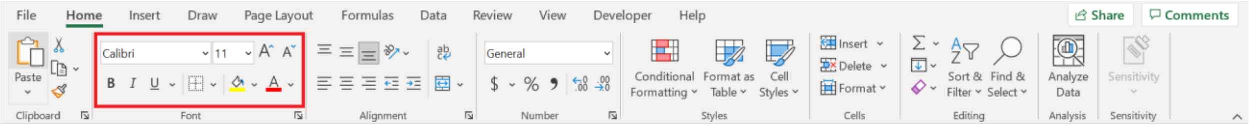
A feature that automatically applies formatting to cells that meet certain criteria that you've specified.

<u>Freeze Panes</u>

An Excel tool that lets you freeze one or more columns or rows, so that they remain fixed in place while the rest of the spreadsheet is free to scroll.

Adding borders and updating fonts

You can use borders and fonts to make data look more organized and help key information stand out. All the borders and font-formatting tools are located in the Home tab of the ribbon.



If you have a large dataset, it can be helpful to make the headers bold. This way, they stand out and are easier to find when you're scrolling through a dataset. Similarly, if you're looking at a few specific columns or rows in a large dataset, it can be helpful to highlight them with a color. That makes it easier to spot your columns or rows of interest as you're scrolling back and forth between them.

Borders organize data in a way that draws attention to certain information.

As you may have noticed, bolding the headers and totals makes it easier to differentiate them from the data. Borders also help separate out totals and headers from the data. Highlighting makes calculated or special columns stand out. Finally, centering the text and figures makes it easier for the eyes to scan.

Here are some other best practices for formatting data:

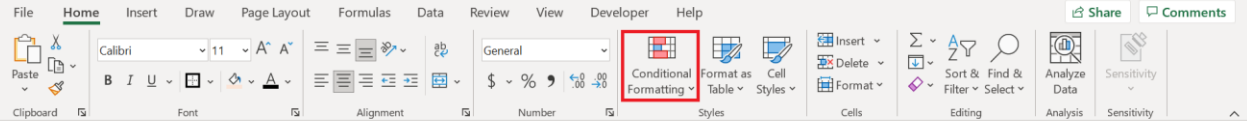
- Use a different font color for calculated values or for values where the user can change the input.
- Avoid having blank rows or columns in the middle of your data, because they can cause data to be left out of formulas (or out of view).
- Include headers for all columns, describing what's in the column. If column headers are missing or blank, create them based on what the values of the column represent.
- Avoid including trailing or leading whitespace in your data. Spaces count as characters, and they may throw off formulas.

In general, you want your data to be visually easy to follow; it shouldn't require the user to think too hard to figure out what's there.

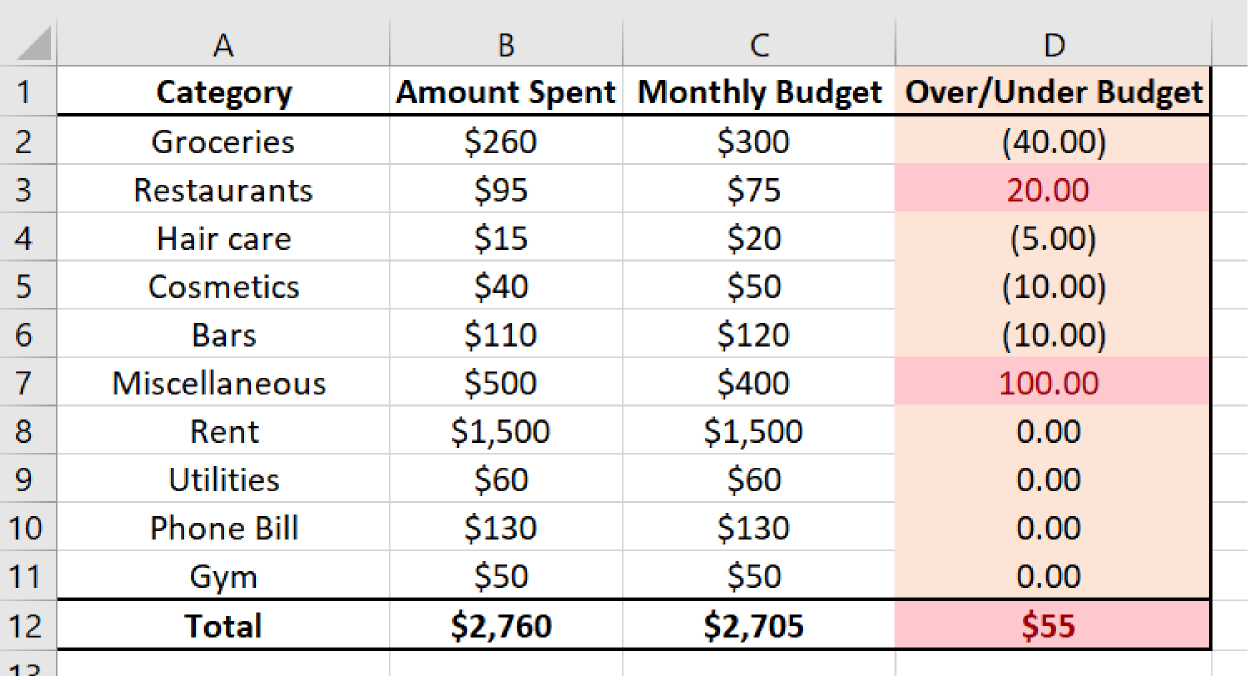
Adding Conditional Formatting

Excel's Conditional Formatting tool automatically formats data for you if the data meets a certain condition or set of conditions that you've specified.

The Conditional Formatting tool is in the Home tab of the ribbon.



For example, imagine that you want to highlight all positive values in the Over/Under Budget column (column D). To do so, you'd select cells D2:D12, then click Conditional Formatting > Highlight Cell Rules > Greater Than. Because you want all values greater than 0 highlighted red, enter 0 in the first box. The second box indicates that the cells will be highlighted in red, with dark red font, if they meet this condition. Click OK.



	A	B	C	D
1	**Category**	**Amount Spent**	**Monthly Budget**	**Over/Under Budget**
2	Groceries	$260	$300	(40.00)
3	Restaurants	$95	$75	20.00
4	Hair care	$15	$20	(5.00)
5	Cosmetics	$40	$50	(10.00)
6	Bars	$110	$120	(10.00)
7	Miscellaneous	$500	$400	100.00
8	Rent	$1,500	$1,500	0.00
9	Utilities	$60	$60	0.00
10	Phone Bill	$130	$130	0.00
11	Gym	$50	$50	0.00
12	**Total**	**$2,760**	**$2,705**	**$55**
13				

Now you can quickly see the categories that you went over budget on.