

PREPARING DATA IN EXCEL

1	IMPORTING DATA INTO EXCEL SPREADSHEETS	2
1.1	INTRODUCTION	2
1.2	LEARNING ABOUT DELIMITERS.....	3
1.2.1	<i>Loading data into a spreadsheet</i>	<i>3</i>
1.2.2	<i>Delimiters.....</i>	<i>3</i>
1.3	LOADING TEXT FILES: METHOD 1	4
1.3.1	<i>Text files: open or import?</i>	<i>4</i>
1.3.2	<i>Your scenario</i>	<i>4</i>
1.3.3	<i>Method 1 for text files</i>	<i>5</i>
1.3.4	<i>Instructions</i>	<i>9</i>
1.4	LOADING TEXT FILES: METHOD 2	10
1.4.1	<i>Another way to load text file datasets.....</i>	<i>10</i>
1.4.2	<i>Method 2 for text files</i>	<i>10</i>
1.4.3	<i>Instructions</i>	<i>10</i>
1.5	CONVERT TEXT TO COLUMNS WIZARD	11
1.5.1	<i>Loading other file types</i>	<i>11</i>
1.5.2	<i>Excel files.....</i>	<i>11</i>
1.5.3	<i>Copy and paste</i>	<i>12</i>
1.5.4	<i>Instructions</i>	<i>14</i>
1.6	DRAG AND DROP	15
2	ORGANIZING DATA.....	16
2.1	INTRODUCTION	16
2.1.1	<i>Instructions</i>	<i>18</i>
2.2	CELLS AND RANGES	18
2.2.1	<i>Instructions</i>	<i>20</i>
2.3	TABLES.....	21
2.3.1	<i>Instructions</i>	<i>24</i>
2.4	WORKSHEETS.....	25
2.5	WORKBOOK.....	28
2.5.1	<i>Instructions</i>	<i>30</i>
2.6	EXCEL FUNCTIONS	31
2.6.1	<i>Instructions</i>	<i>32</i>
3	CLEANING DATA PART 1	33
3.1	INTRODUCTION	33
3.1.1	<i>Data cleaning processes</i>	<i>33</i>
3.2	REMOVING DUPLICATE DATA	34
3.2.1	<i>A scenario</i>	<i>35</i>
3.2.2	<i>Instructions</i>	<i>38</i>
3.3	REMOVING EXTRA SPACES	38
3.3.1	<i>Instructions</i>	<i>42</i>
3.4	SEPARATING DATA	43
3.4.1	<i>Instructions</i>	<i>45</i>
3.5	CORRECTING INACCURATE DATA	46
3.5.1	<i>Instructions</i>	<i>51</i>

3.6	UPDATING MISSING DATA	51
3.6.1	<i>Instructions</i>	53
3.7	UPDATING MISSING DATA CONTINUED	53
3.7.1	<i>Instructions</i>	55
3.8	OTHER CONSIDERATIONS.....	55
4	CLEANING DATA PART 2	55
4.1	INTRODUCTION	56
4.2	STANDARDIZING TEXT DATA.....	57
4.2.1	<i>Instructions</i>	59
4.3	FORMATTING DATES.....	59
4.3.1	<i>Instructions</i>	61
4.4	FORMATTING NUMBERS.....	61
4.4.1	<i>Instructions</i>	63
4.5	WORKING WITH OUTLIERS.....	63
4.5.1	<i>Methods for handling outliers</i>	66
4.6	VISUALIZING YOUR DATA.....	67
4.6.1	<i>Instructions</i>	72
5	CONSOLIDATING DATA.....	72
5.1	INTRODUCTION	72
5.2	MANIPULATING HEADERS.....	73
5.2.1	<i>Instructions</i>	75
5.3	THE CONSOLIDATE TOOL	75
5.3.1	<i>Instructions</i>	77
5.4	THE CONSOLIDATE TOOL CONTINUED	78
5.4.1	<i>Instructions</i>	78
5.5	COPY AND PASTE	79
5.5.1	<i>Instructions</i>	84
5.6	USING VLOOKUP	85
5.6.1	<i>Instructions</i>	87
5.7	USING INDEX/MATCH	88
5.7.1	<i>Instructions</i>	89

1 Importing Data into Excel Spreadsheets

1.1 Introduction

Preparing data involves cleaning and organizing. It also includes knowing how to fix bad and missing data, and consolidating data so that it's easier to analyze (this includes all types of data). Many people consider this step the "least exciting" of the data analytics process, but it's also the most important part. Think about it. If your data isn't in good shape to begin with, your analysis won't be trustworthy. There's an old saying: "Garbage in, garbage out." Your task in this step is to take out the garbage!

Throughout this course, you'll encounter the processes and skills necessary to clean and prepare your data with spreadsheets so that it's ready for analysis. In this first lesson, you'll learn how to load your data into Excel.

1.2 Learning About Delimiters

1.2.1 Loading data into a spreadsheet

The first step when preparing your data in Excel is to load it. Unless you're working with a very small amount of data that you type in directly, you'll almost always load your data from some external source. This means you'll need to become very familiar with some of the more common sources of data, including the following:

- Excel files
- text files
- CSV files
- copying & pasting data
- dragging & dropping data

As a data analyst, it's also possible that you may need to import data from a proprietary or customized system. While that's beyond the scope of this course, the general approach you'll learn here will help you understand the process.

1.2.2 Delimiters

Before we go any further, you need to understand delimiters. Older versions of Excel required that you go through a series of steps to load data. This was an attempt to give you a dataset with values located in the correct columns and rows so that it would save you some work. You'd have to specify special characters, such as commas, tabs, or spaces to separate individual values. We call these special characters that act as separators *delimiters*. And even though modern spreadsheet programs have improved the process of loading data, they aren't always perfect. So, you still need to know how to use delimiters.

Suppose you have the following text file that shows the color, quantity, state, and price for online items sold:

color,qty,state,price

blue,7,Vermont,14.50

red,5,Michigan,4.95

green,3,Colorado,12.25

You can tell by examining the data that commas are separating the values. Datasets separated by commas are `Comma Separated Variable` files, or CSV files, and they are perhaps the most common type of format you'll encounter. In most cases, the commas themselves are `not` part of the actual data — they're simply there to act as delimiters.

Watch for numerical data that uses commas to separate thousands (like 25,000), or text data, such as customer comments. In these cases, use commas and other punctuation in a way that doesn't distort your data. For example, the sentence `"I really liked the pizza, but the salad was terrible"` contains a comma. You need to be careful how you load this sentence so that you don't treat it as two separate sentences by inadvertently using the comma as a delimiter!

As a data analyst, you'll need to be proficient with every step of the data analytics pipeline. So, knowing how to use delimiters will also help you do a better job of collecting and storing data — because you'll know what's necessary during the data preparation step.

1.3 Loading Text files: Method 1

1.3.1 Text files: open or import?

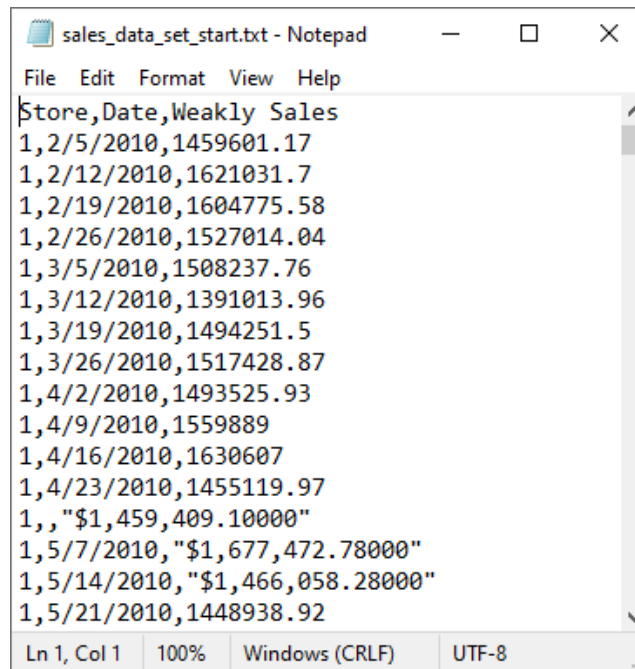
Until now, we've used the word "load" to generically describe the process of bringing your data into a spreadsheet. But to be more specific, you'll use `Open` or `Import` when you're pulling your data into Excel. While earlier versions of Excel were more rigid about using `Open` or `Import`, recent versions offer greater flexibility for users to load data. In most cases, you can use either `Open` or `Import` to load your data, but there are some exceptions. Let's look at a scenario to illustrate this!

1.3.2 Your scenario

Suppose that you work as a data analyst for a national retail store. You're part of a team of other analysts, and you've been asked to help prepare data for analysis. Your first step is to load the data into Excel.

One of your colleagues has sent you the dataset titled `sales_data_set_start.txt`. This dataset has several errors, and you'll use it throughout the course as a way to learn how to clean and organize data.

If you were to open it in a text editor like Notepad, you'd see something like this:

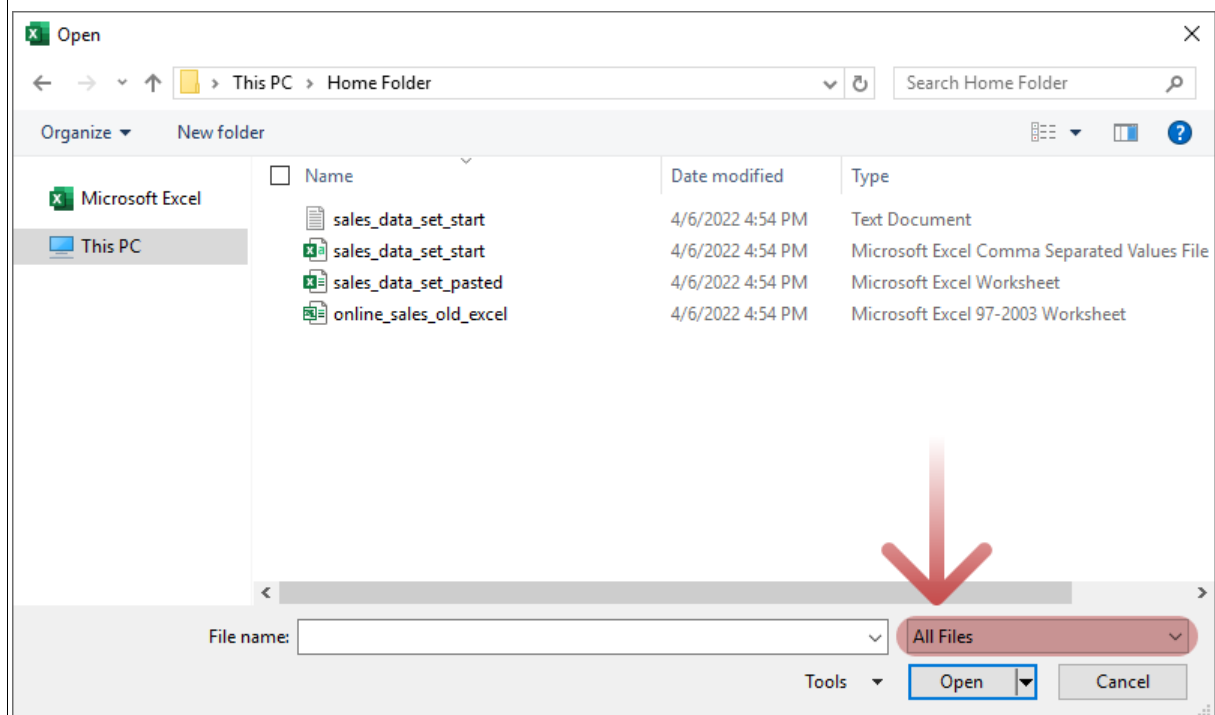


```
File Edit Format View Help
Store,Date,Weekly Sales
1,2/5/2010,1459601.17
1,2/12/2010,1621031.7
1,2/19/2010,1604775.58
1,2/26/2010,1527014.04
1,3/5/2010,1508237.76
1,3/12/2010,1391013.96
1,3/19/2010,1494251.5
1,3/26/2010,1517428.87
1,4/2/2010,1493525.93
1,4/9/2010,1559889
1,4/16/2010,1630607
1,4/23/2010,1455119.97
1,, "$1,459,409.10000"
1,5/7/2010, "$1,677,472.78000"
1,5/14/2010, "$1,466,058.28000"
1,5/21/2010,1448938.92
Ln 1, Col 1 100% Windows (CRLF) UTF-8
```

Take a moment to inspect the dataset above. Looking closely, you'll see that it contains commas that separate the data into columns. You'll also notice some obvious errors, such as "Weekly Sales" — and some inconsistent values. Your goal on this screen is to pull the text file data into Excel so that you can manipulate it and prepare it for analysis. There are multiple ways to do this; let's look at two of them.

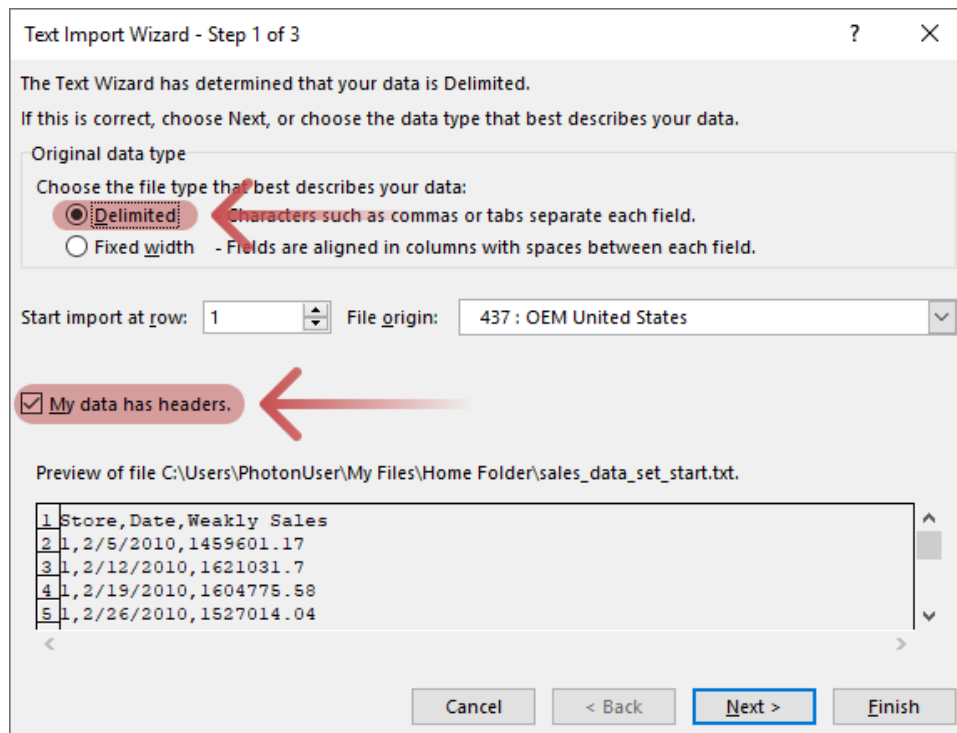
1.3.3 Method 1 for text files

In Excel, if you were to select **File, Open, Browse**, you'd see a window similar to this:



In the bottom-right (circled), you can choose the file type to open. For this exercise, you can set it to **All Files** or change it to **Text Files**.

If you were to open the dataset, you'd see a window for the **Text Import Wizard**. Notice that you started out choosing **Open**, but you were quickly moved over to using **Import**. Excel has blurred the lines between **Open** and **Import** to make it a seamless experience for the user.



Text Import 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.

Start import at row: 1 File origin: 437 : OEM United States

☒ **My data has headers.**

Preview of file C:\Users\PhotonUser\My Files\Home Folder\sales_data_set_start.txt.

1	Store, Date, Weekly Sales
2	1, 2/5/2010, 1459601.17
3	1, 2/12/2010, 1621031.7
4	1, 2/19/2010, 1604775.58
5	1, 2/26/2010, 1527014.04

Buttons: Cancel, < Back, **Next >**, Finish

The main thing you would need to focus on as shown above is to make sure that **Delimited** and **My data has headers** are both selected.

If you were to click on **Next**, you'd see the following screen, which is where you select the type of delimiter. It will probably default to **Tab**, which is incorrect. You want **Comma**, so you would deselect **Tab** and select **Comma** as shown below.

You can see the difference in the **Data preview** window at the bottom. If you were to toggle between the choices, you'd see how it affects the data in the preview window!

Text Import 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

Store	Date	Weekly Sales
1	2/5/2010	1459601.17
1	2/12/2010	1621031.7
1	2/19/2010	1604775.58
1	2/26/2010	1527014.04

Cancel < Back Next > Finish

Of course, if your data file used tabs, spaces, or semicolons as delimiters, you'd want to select those.

After you click **Next**, you'd see the final screen of the **Text Import Wizard**:

Text Import 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.

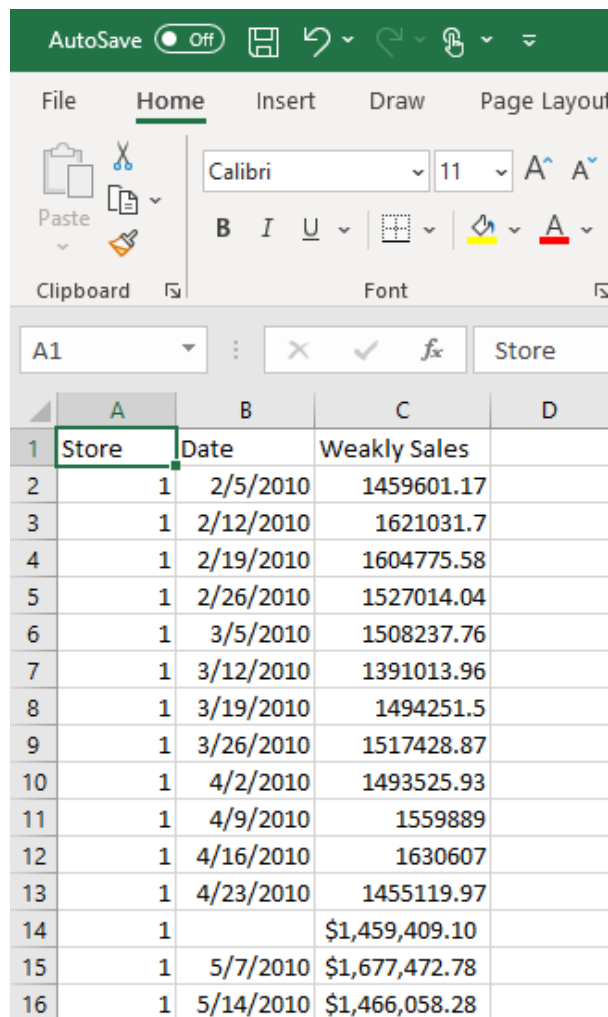
Advanced...

Data preview

Store	Date	Weekly Sales
1	2/5/2010	1459601.17
1	2/12/2010	1621031.7
1	2/19/2010	1604775.58
1	2/26/2010	1527014.04

Cancel < Back Next > Finish

Here you can set the **Data Format** for each column. For now, you'll use the defaults. When you're done, click **Finish**, and you'll see your data loaded in Excel:



	A	B	C	D
1	Store	Date	Weekly Sales	
2	1	2/5/2010	1459601.17	
3	1	2/12/2010	1621031.7	
4	1	2/19/2010	1604775.58	
5	1	2/26/2010	1527014.04	
6	1	3/5/2010	1508237.76	
7	1	3/12/2010	1391013.96	
8	1	3/19/2010	1494251.5	
9	1	3/26/2010	1517428.87	
10	1	4/2/2010	1493525.93	
11	1	4/9/2010	1559889	
12	1	4/16/2010	1630607	
13	1	4/23/2010	1455119.97	
14	1		\$1,459,409.10	
15	1	5/7/2010	\$1,677,472.78	
16	1	5/14/2010	\$1,466,058.28	

You may notice some cells contain ##### instead of displaying data. This happens when the column isn't wide enough to display the data. To resize a column, hover your mouse cursor between two column headers (e.g., B and C, just above the first row of data) until a double-headed arrow appears, and then double-click. The column will automatically adjust its width to display the data correctly.

1.3.4 Instructions

1. In Excel, select **File, Open, Browse**, and view the choices.
2. Locate the `sales_data_set_start.txt` dataset, and click **Open**.
 - To find the dataset: click **This PC**, and select the **Home Folder**.
 - Change the file type to **All Files** or **Text Files** to display the dataset.
3. On the first screen of the **Text Import Wizard** make sure that **Delimited** and **My data has headers** are both selected. Click **Next** to continue.
4. On the second screen of the **Text Import Wizard**, select the correct delimiter.
 - Deselect **Tab** and select **Comma**. Click **Next** again to continue.

5. On the final screen of the Wizard, you can leave everything as-is and select **Finish**.

1.4 Loading Text Files: Method 2

1.4.1 Another way to load text file datasets

1.4.2 Method 2 for text files

The next method we'll look at feels like a "hack," but it's sometimes necessary, especially if your dataset wasn't saved with the proper file extension in the first place. For this second method, you would go into your file explorer on your local machine and change the file extension from **.txt** to **.csv**.

To do this, you'd make a copy of the text file ending in "txt" so you still have the original, then change the file extension of the new file to end in "csv," so now it would look like **sales_data_set_start.csv**. Then, you'd go back to Excel, and select **Open** to load the file as you did in Method 1.

This time, you don't need to change the file type to see the file. Excel recognizes the CSV file just as it recognizes a regular Excel file (**.xlsx**), and it will open it in Excel without using the Wizard. CSV files have become so ubiquitous in the world of data that Excel finally made it straightforward to simply open them like any other workbook. There is one point of caution for this method, and it's that you *need to be sure* your dataset is comma-separated before you change the file extension!

CSV files are wonderful for passing datasets between different applications because they've become a standard among data file formats. However, it's often advantageous to save your CSV file as an Excel file (**.xlsx**). This is especially true if you plan to use Excel to manipulate the file going forward. If you're working with smaller files, like the one you're using in this lesson, the file size will **increase** after being converted to an Excel file because of the extra overhead that goes along with XLSX files. But, if your CSV file is quite large (larger than 100 MB), then saving it as an Excel file will often reduce the size of the file. And having it in XLSX format allows Excel to be more efficient with the file because it's optimized to work with Excel's native file format.

Finally, there is one more thing you need to know. If you simply go to **Data** > **Get Data** to import your data, you'll be using Excel's Power Query tool, which modifies the data format and uses advanced features that are beyond the scope of this course.

For the sake of convenience, we have created a copy of the file **sales_data_set_start.txt** and renamed this copy to **sales_data_set_start.csv** for the exercise below.

1.4.3 Instructions

1. In Excel, select **File**, **Open**, **Browse**, and view the choices.

- Notice that while you're browsing for files in Excel, the application displays information about each file, such as **Type** and **Size**.

2. Open the file `sales_data_set_start.csv` that we created for you.

3. Use **File > Save As** to save your CSV file as an XLSX file: `sales_data_set_start.xlsx`.

- Use the drop-down menu on the right side of your screen, next to the **Save** icon, to select the file type **Excel Workbook (*.xlsx)**.
- If you get a warning message saying the file already exists, select **Ok** to replace the existing file.

1.5 Convert Text to Columns Wizard

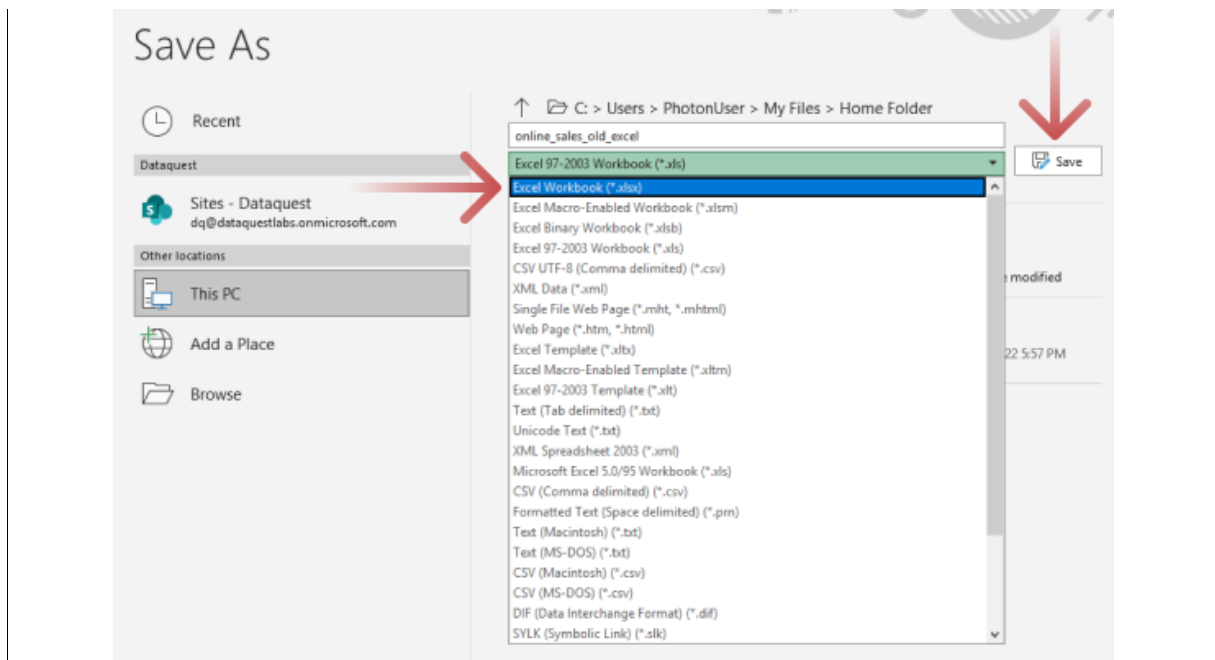
1.5.1 Loading other file types

1.5.2 Excel files

So far, you've seen how to load text files and CSV files. Opening an Excel file (`.xlsx`), like the one you just created on the previous screen, opens as expected — there isn't anything special you need to do because Excel just recognizes it and loads it.

It's quite possible that you'll encounter older versions of Excel files, such as those ending in `.xls`. The older XLS files are stored in binary format, whereas the latest XLSX files use a compressed XML file format, so they are always smaller. It's always a good idea to use **File > Save As** to save old XLS files in the newer XLSX format for speed and efficiency. But, if there's ever a reason you need to go back to the old format, you can always use **File > Save As** to convert to the old XLS format if necessary.

If you did so, you'd have two copies of the workbook, one with the old XLS extension and one with the new XLSX extension. With an XLS file already opened in Excel, selecting **File > Save As** opens the following window:



1.5.3 Copy and paste

There are more methods we'll look at for loading data into Excel. This next one may seem primitive, but copying and pasting is something you might spend more time doing as a data analyst than you may realize.

Let's go back to your scenario. Suppose you were to use a text editor like Notepad to bring up the original text file. You could select and copy all of the data in your text editor window, then paste it into an Excel worksheet:

is that the checkbox for **My data has headers** isn't available because you're simply pasting data — but Excel doesn't know what kind of data you're pasting. Now you would proceed as you did previously when you used the **Text Import Wizard** to choose delimiters.

You will then see that your data is now nicely divided into the appropriate columns:

	A	B	C	D
1	Store	Date	Weakly Sales	
2	1	2/5/2010	1459601.17	
3	1	2/12/2010	1621031.7	
4	1	2/19/2010	1604775.58	
5	1	2/26/2010	1527014.04	
6	1	3/5/2010	1508237.76	
7	1	3/12/2010	1391013.96	
8	1	3/19/2010	1494251.5	
9	1	3/26/2010	1517428.87	
10	1	4/2/2010	1493525.93	
11	1	4/9/2010	1559889	
12	1	4/16/2010	1630607	
13	1	4/23/2010	1455119.97	
14	1		\$1,459,409.10	
15	1	5/7/2010	\$1,677,472.78	
16	1	5/14/2010	\$1,466,058.28	
17	1	5/21/2010	1448938.92	
18	1	5/28/2010	1514259.78	

Using copy and paste may work equally well for you if you need to copy data from a website or an online document. If you do this, you'll need to pay attention to any unusual characters that may come along for the ride. Just remember that the **Text Import Wizard** has an **Other** category if you need to specify a custom delimiter!

For the sake of convenience, we have created the workbook `sales_data_set_pasted.xlsx` with the contents of `sales_data_set_start.txt` already copied to column **A** for the exercise below.

1.5.4 Instructions

Save XLS file as XLSX file

1. Open the file `online_sales_old_excel.xls`.

2. Use **File > Save As** to save the old Excel file type (XLS) to the newer **Excel Workbook (*.xlsx)** file type.

Convert Text to Columns Wizard

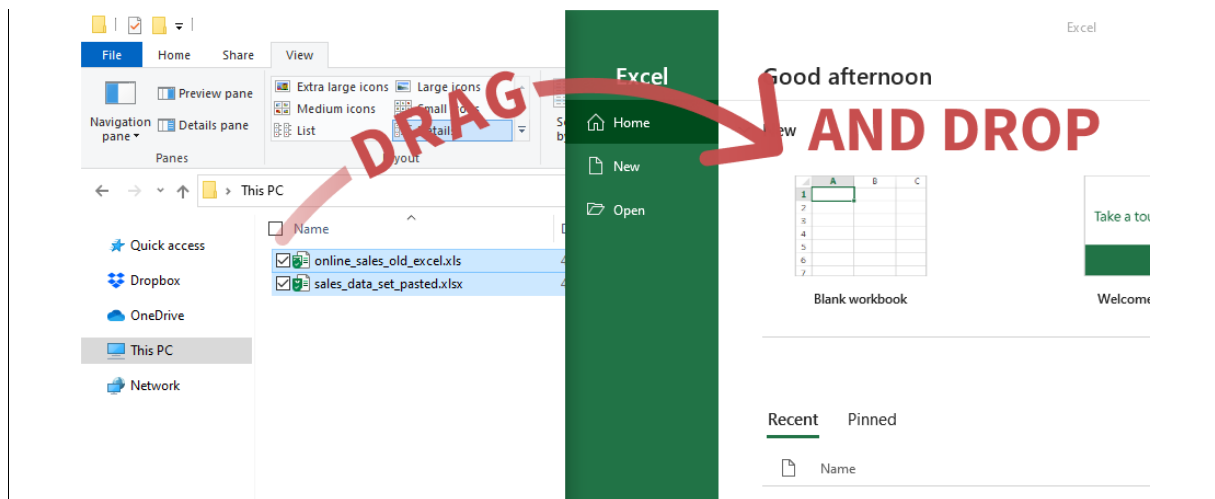
1. Open the file `sales_data_set_pasted.xlsx`.
2. Select column **A**, then select the **Data** tab, then **Text to Columns**.
3. Follow the **Wizard** as before, and choose the appropriate delimiters.
4. Click **Finish** to finish.

1.6 Drag and Drop

The last method for loading data files sometimes gets overlooked, but it's worth knowing. If you have a data file, and you want to load it into Excel, you can simply use drag and drop to load the file. The process is very similar to when you copy files from one folder to another on your computer using drag and drop. When you do this with XLSX or CSV files, the files will simply load as usual. When you do this with text files, it's a lot like what you just experienced with an Excel file with data pasted from a text file. You'll need to convert the file from one column to multiple columns using the **Wizard**.

So, why would you want to use drag and drop? If you have several data files that you need to load in a short period of time, then using drag and drop may be a little faster than navigating Excel's menu system. Plus, some people prefer the visual aspect provided by dragging and dropping icons. You can also load multiple data files *at the same time* using drag and drop (while you can also do this using the Excel menu system, you may find it faster and more intuitive using drag and drop). As we said at the beginning of this lesson, there are multiple ways to accomplish the same thing, and you may find some to be more convenient than others.

You could do this by opening Excel and your file explorer, then positioning them side by side, similar to the following. You would then select the two files `online sales old excel.xlsx` and `sales data set start.xlsx` in your file explorer and drag them at the same time on top of the **Blank workbook** icon in Excel as shown below:



After you did this, Excel would open two workbooks for you, each with their own dataset. If you were required to use a **Wizard**, Excel would take you through the usual prompts.

Since we're using an integrated version of Excel directly in the browser, it isn't possible to use the drag and drop technique we just learned. However, using this technique on your local computer would be fine.

That's it! Now you've taken a comprehensive tour of how to open files in Excel.

2 Organizing Data

2.1 Introduction

In the previous lesson, you learned how to load data into a spreadsheet from various data sources. Once you've loaded your data, you need to organize it for analysis.

For example, you ended the last lesson by learning how you could drag and drop two different files into an Excel spreadsheet simultaneously. If you did that, you would produce two different Excel workbooks, each with their respective datasets. If your goal was to analyze both of those datasets together, then it would be more convenient to have them in the same workbook, and possibly even on the same worksheet. This lesson is all about organizing your data to best prepare it for analysis.

Before we go any further, let's go through a quick refresher on the structure of Excel — and some common terminology.

It's not uncommon for data professionals to misuse terms when referring to data processes and tools. Although this is often harmless, it can create confusion when

specifying requirements, and that can lead to projects that don't meet requirements. We'll take a few minutes to clarify some terms that often get used loosely in the world of spreadsheets.

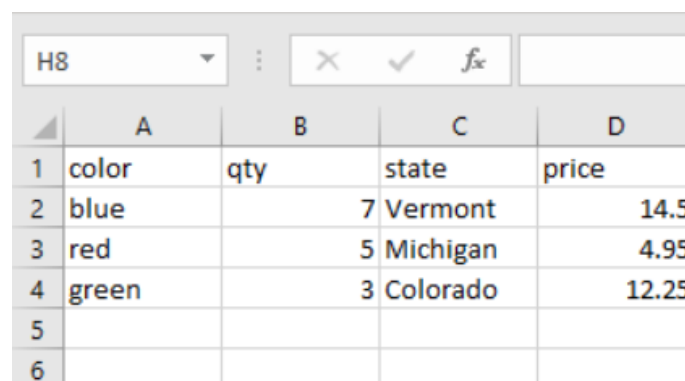
Let's start with the word "spreadsheet." This word really refers to the computer application that creates spreadsheets. Microsoft Excel is an example of a spreadsheet application. The file you produce by using a spreadsheet application is called a "workbook," often ending in ".XLSX." Having said that, it's not uncommon for people to use the words "spreadsheet" and "workbook" synonymously. You'll often hear people say things like "Let's look at the spreadsheet," when they really mean "Let's look at the workbook."

Using the words "spreadsheet" and "workbook" interchangeably may be a minor mistake, but things get even more complicated when people use the terms "workbook" and "worksheet" synonymously. A workbook contains one or more worksheets. And a worksheet in Excel is a single tab (so the words "worksheet" and "tab" are synonymous)!

Within a worksheet, across the top, you have columns that go from left to right, and those columns are assigned letters, starting with A. You also have rows that go from top to bottom, and those rows are assigned numbers starting at 1. We locate each of the individual cells within a worksheet by using a combination of the column and row together, such as A21 or D43.

This combination of column letter and row number is the cell's address. Whenever you want to access the contents of a particular cell, you use its address in a formula. We also call this referencing a cell. Now it's time to see this in action!

Open the workbook you saved from the previous lesson named online sales old excel.xlsx. You should see something like the following:



	A	B	C	D
1	color	qty	state	price
2	blue		7 Vermont	14.5
3	red		5 Michigan	4.95
4	green		3 Colorado	12.25
5				
6				

Suppose you want to access the contents of Cell B3 and place the value in Cell E1. To do so, select Cell E1, and then type in =B3, hit Enter, and you'll see the value of Cell B3. (Remember that any entries starting with the = symbol tell Excel you have a formula!)

To answer the following questions, you'll need to leave your workbook open. When completed, you do not need to save your workbook (or you can save it under a new name if you wish)!

2.1.1 Instructions

We've provided a solutions file ([sales_combined_solutions.xlsx](#)) in case you get stuck or would like to verify your answers while working on the exercises in this lesson.

1. Open the workbook [online_sales_old_excel.xlsx](#) if it isn't already open.

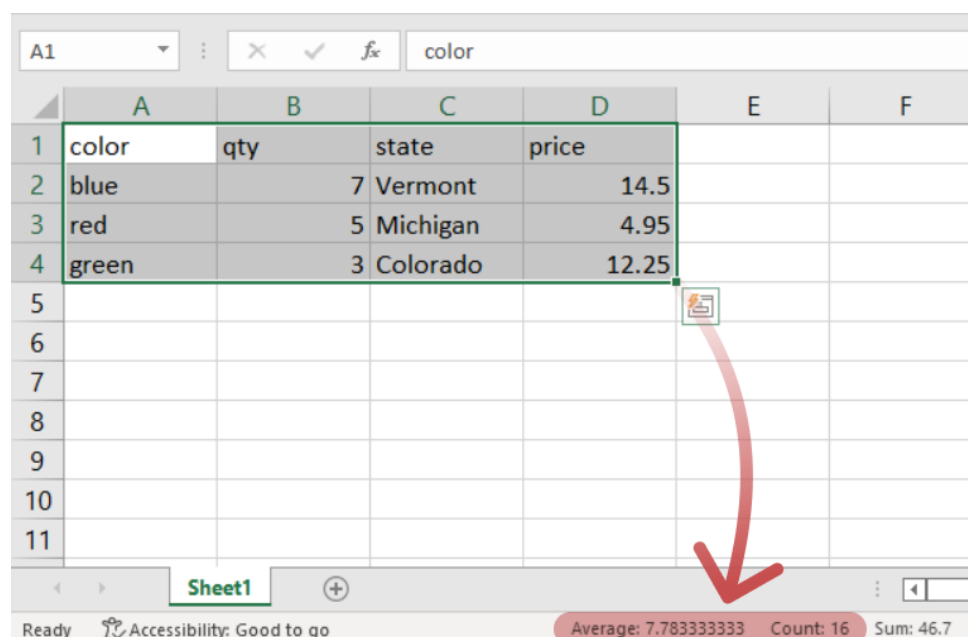
- You can find the file by selecting **File, Open, Browse** and navigating to the **Home Folder** under **This PC**.

2. Select **Cell F1**, and enter the following formula for rounding a number to 0 decimals: `=ROUND(D2,0)`.

2.2 Cells and Ranges

We just learned that we can access a single cell by referencing its column-row address. That's very helpful when we want to focus only on a single cell. But, what happens if we want to access more than a single cell? When we access more than a single cell at the same time, we're accessing a range of cells, or simply just a range. On this screen, you'll be looking at the workbook [online_sales_old_excel.xlsx](#) again.

Now, suppose that we select all the cells (the range of cells) from **A1** to **D4**. After doing so, you would see this summary in the lower right of your workbook in what's called the **status bar**:



That works fine if you just want some quick information about a range of cells. But, if you want to use a range inside a formula, you need to know how to access the range using the correct conventions. To specify a range in a formula, you use a colon symbol `:` to separate the upper left corner where the range starts from the lower right corner where it ends. If you wanted to work with the same range you just highlighted above, you'd use `A1:D4` to specify that range.

For example, let's say you want the count of all cells that aren't blank from **Cell A1** through **Cell D4**, and you want your result in **Cell E1**. To do this, you'd use the function `COUNTA` that returns the number of non-blank cells.

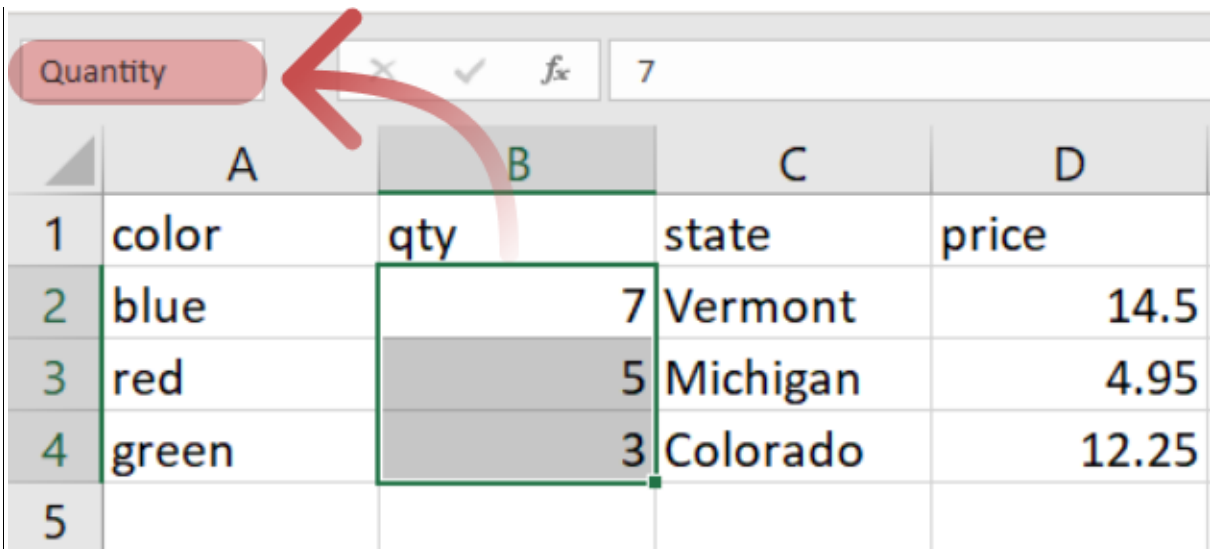
If you were to select cell **E1** and type in `=COUNTA(`, you could enter the range needed by first selecting cell **A1**, then while holding down the `SHIFT` key, selecting cell **D4**. After hitting `Enter`, you'd see the count of all the cells in the range you selected that contain data.

Alternatively, you can manually type in the range using the cell addresses, `=COUNTA(A1:D4)`. If you did this correctly, you would see `16` as your answer, because that's the number of cells that are not empty in your range (and this matches the status bar information you saw previously).

Ranges, by definition, are contiguous; they do not contain any gaps. However, we can access multiple ranges at the same time to create a "super range" made up of smaller sub-ranges. We can do this by separating each sub-range with a comma symbol `,`. For instance, suppose we wanted to get the count of the range `B2:B4` as well as the range `D2:D4`. What do you think your formula would look like? If you guessed `=COUNTA(B2:B4,D2:D4)`, you're correct!

Also, since a range is comprised of cells, a single cell is *technically* a range — although most people reserve the word *range* for those instances involving more than one cell. Imagine a scenario wherein you want the count of several non-contiguous individual cells, such as **A1**, **B2**, **C3**, **D3**, and **D4**. What do you suppose that formula would look like? If you're thinking `=COUNTA(A1,B2,C3,D3,D4)`, you're correct again!

Finally, you can name your ranges for ease of use later on in formulas. To do this, use your mouse to select the range `B2:B4`, and then in the upper left **Name Box** just above cell **A1**, you would type in the name `Quantity`.



	A	B	C	D
1	color	qty	state	price
2	blue	7	Vermont	14.5
3	red	5	Michigan	4.95
4	green	3	Colorado	12.25
5				

If you were to do the same thing for **Cells D2:D4**, and name that range **Price**, you could use those names in formulas instead of column/row addresses.

```
=COUNTA(Quantity,Price)
```

Isn't this formula easier to read and understand than the same formula using a range of cells?

The concept of ranges is important because ranges are a way to isolate specific cells containing data that interests you. On the next screen, you'll look at **tables**, which are special types of ranges.

2.2.1 Instructions

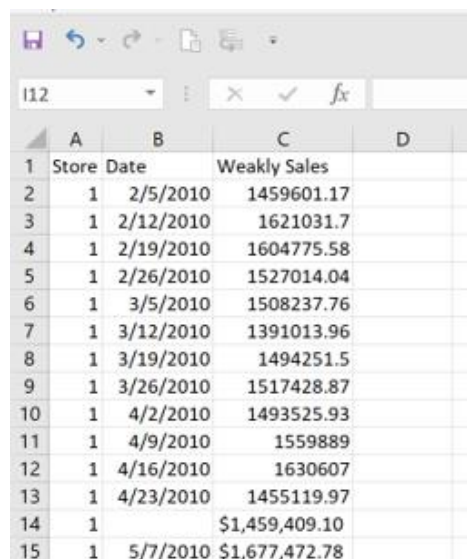
1. Open the workbook **online_sales_old_excel.xlsx** if it isn't already open.
2. Highlight all the range **A1:D4**. If you look in the status bar in the lower right corner of your window, you'll get some basic information about the range you just selected.
3. Select cell **E1**, and type **=COUNTA (**.
4. Select cell **A1**, then while holding down the **SHIFT** key, select cell **D4**. When you hit **Enter**, you'll see the count of all the numbers in the range you selected.
5. Type the formula **=COUNTA (B2:B4, D2:D4)** in cell **F1**. Overwrite the contents of the cell, if necessary.
6. Type the formula **=COUNTA (A1, B2, C3, D3, D4)** in cell **G1**.

7. Rename your range by using your mouse to select the range of cells **B2:B4**, and then in the upper left **Name Box** just above cell **A1**, type in the name **Quantity**.
8. Rename the range **D2:D4**, and name it **Price**.
9. To see how to use the named ranges in a formula, go to cell **H1** and enter **=COUNTA(Quantity,Price)**, then hit **Enter**.
- 10.

2.3 Tables

On the previous screen, you learned about ranges. On this screen, you'll learn about tables, which are a special type of range in Excel. A table is a range where the data fits nicely into a single collection of columns and rows. A table is also treated differently than a range in Excel because it is specially formatted with several features built-in. Unlike a range, a table needs to be contiguous. Also, every table comes with headers. Even if your data doesn't already have headers, Excel will assign some generic headers for you.

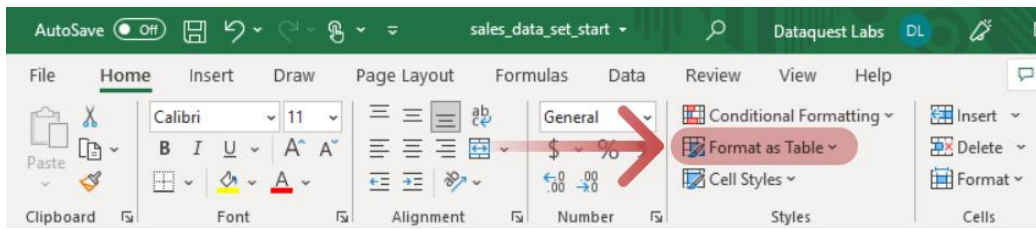
If you were to open **sales data set start.xlsx**, you'd see this:



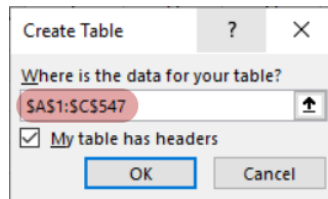
	A	B	C	D
1	Store	Date	Weekly Sales	
2	1	2/5/2010	1459601.17	
3	1	2/12/2010	1621031.7	
4	1	2/19/2010	1604775.58	
5	1	2/26/2010	1527014.04	
6	1	3/5/2010	1508237.76	
7	1	3/12/2010	1391013.96	
8	1	3/19/2010	1494251.5	
9	1	3/26/2010	1517428.87	
10	1	4/2/2010	1493525.93	
11	1	4/9/2010	1559889	
12	1	4/16/2010	1630607	
13	1	4/23/2010	1455119.97	
14	1		\$1,459,409.10	
15	1	5/7/2010	\$1,677,472.78	

Note: You may need to resize your columns to display the data properly. You'll also notice some obvious errors, such as "Weekly Sales" — we'll fix this and other errors in upcoming exercises.

If you were to select any cell within the range of data, and go to the **Home** tab under the **Styles** group, then select **Format as Table** and choose a style, you'd see this:



You would see a dialog box asking you to confirm the selected range, and then you would click **OK**.



So, you've just seen how to convert your range to a table — congratulations! *But now what?* Why would you want to convert a range to a table, and what does that do for you?

There are several advantages to using a table over a range. Here are five of those advantages:

1. You get nice visual formatting that separates the rows for easy viewing using colors and shades:

	A	B	C	D
1	Store	Date	Weekly Sales	
2	1	2/5/2010	1459601.17	
3	1	#####	1621031.7	
4	1	#####	1604775.58	
5	1	#####	1527014.04	
6	1	3/5/2010	1508237.76	
7	1	#####	1391013.96	
8	1	#####	1494251.5	
9	1	#####	1517428.87	
10	1	4/2/2010	1493525.93	
11	1	4/9/2010	1559889	
12	1	#####	1630607	
13	1	#####	1455119.97	
14	1		\$1,459,409.10	

2. You get automatic filters added to your data:

	A	B	C	D
1	Store	Date	Weekly Sales	
2	1	2/5/2010	1459601.17	
3	1	#####	1621031.7	
4	1	#####	1604775.58	
5	1	#####	1527014.04	
6	1	3/5/2010	1508237.76	
7	1	#####	1391013.96	
8	1	#####	1494251.5	
9	1	#####	1517428.87	
10	1	4/2/2010	1493525.93	
11	1	4/9/2010	1559889	
12	1	#####	1630607	
13	1	#####	1455119.97	
14	1		\$1,459,409.10	

3. Your headers remain visible even when you scroll down:

The screenshot shows the Excel ribbon with the 'View' tab selected. The 'Freeze Panes' button is highlighted with a red circle. A red arrow points from this button to the first row of the table in the spreadsheet below, indicating that the headers will be frozen.

	A	B	C	D	E	F	G	H	I	J
1	Store	Date	Weekly Sales							
2	1	2/5/2010	1459601.17							
3	1	#####	1621031.7							
4	1	#####	1604775.58							
5	1	#####	1527014.04							
6	1	3/5/2010	1508237.76							

Note: For this feature to work, you must select a cell within the table itself.

4. Your table expands as you add new columns or rows:

In your table, if you were to select **Cell D1** and type in "Forecast," then hit **Enter**, you would see that your table would know to treat the new column as part of the table and would just expand the table accordingly. Pretty cool!

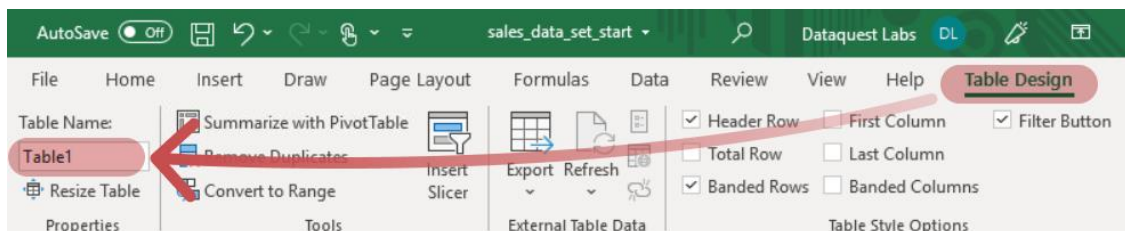
	A	B	C	D	E
1	Store	Date	Weekly Sales	Forecast	
2	1	2/5/2010	1459601.17		
3	1	#####	1621031.7		
4	1	#####	1604775.58		
5	1	#####	1527014.04		
6	1	3/5/2010	1508237.76		
7	1	#####	1391013.96		
8	1	#####	1494251.5		
9	1	#####	1517428.87		

5. Your table will automatically name your ranges (columns) for you.

Your table will use your header names as the names of the ranges for each of the respective columns. So, the range of **Column A** is automatically named "Store." As you saw on the previous screen, that can really come in handy! Be aware, to access these automatically named ranges, you must also reference the table name. We'll learn more about this on an upcoming screen.

After giving you all these great reasons why you'd want to convert a range to a table, you might be wondering if there are any reasons *not* to use a table. One reason might be that you may not need all the extra formatting for your data. Maybe you just want your data to be raw data, and that's what you need for your analysis. Another reason may be that your data is non-contiguous, and it may be more convenient for you to keep it that way. If for some reason you need to convert your table back to a range, simply right-click anywhere in the table, and choose **Table > Convert to Range**, and you'll be asked to confirm that you wish to do this. It's very convenient! As with any of the decisions you'll face in your career as a data analyst, you'll need to determine the best approaches for your projects, and those will vary depending on your data and your project goals.

Tables, like ranges, can and should be named. To name your table, you'd select any cell in the table, then click on the **Table Design** tab. On the far left, just below the **File** menu, you'd see a **Table Name** box. In that box, you'd see a default name that Excel provided, like "Table1." This is where you'd rename your table.



One final note about tables: you can have (and often will have) multiple tables in a single worksheet. The same is true of ranges; you'll often have multiple ranges existing within one worksheet. The next screen will describe worksheets and how they help with organizing your data.

2.3.1 Instructions

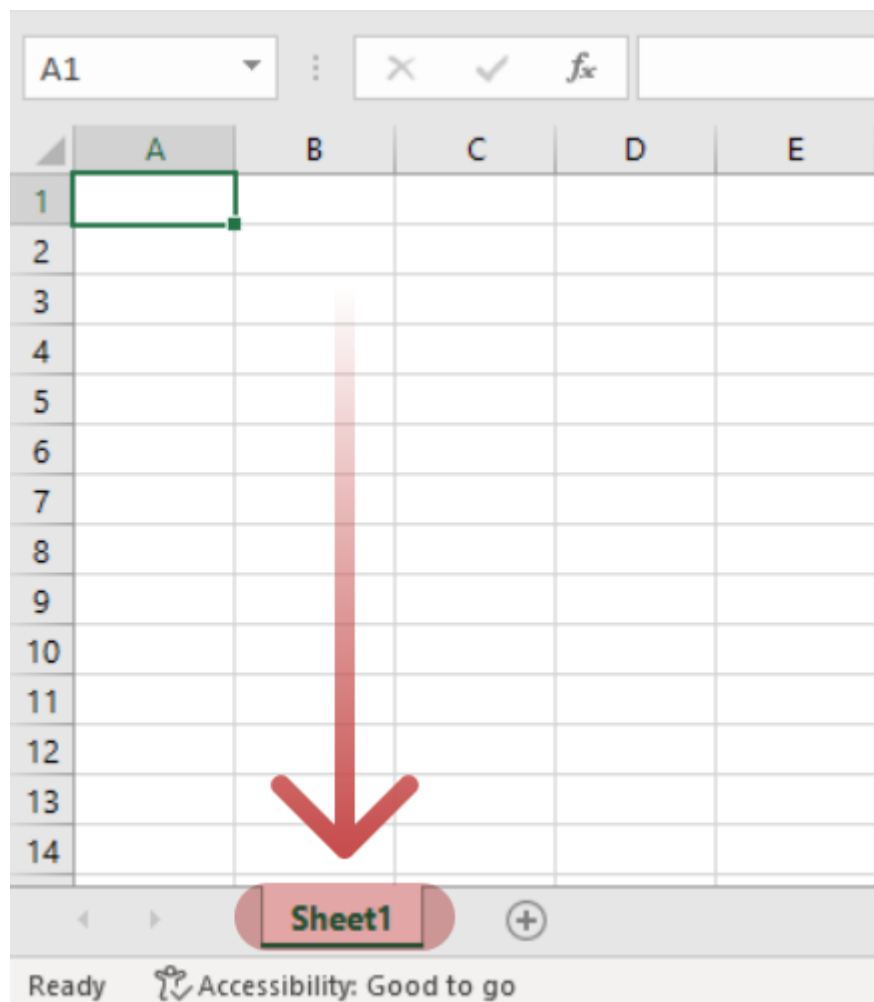
1. Open `sales_data_set_start.xlsx`.
2. Select any cell within the range of data, and in the **Home** tab under the **Styles** group, select **Format as Table** and choose a style (this example uses `Light Blue, Table Style Light 2` but you can choose any style you like).
3. When you see a dialog box asking you to confirm the selected range, click **OK**.

- If you get a message asking if you'd like to "Discover more about your data," click on "Not now" to proceed.

4. To name your table, select any cell in the table, then click on the **Table Design** tab.
5. On the far left, you'll see a **Table Name** box. In that box, you'll see a default name that Excel provided, like "Table1." Rename your table "Sales_Data", noting that you can't use a space in your name.
6. Sort the table by **Column C** using **Sort Smallest to Largest** by clicking on the down arrow next to column header name **Weakly Sales**.

2.4 Worksheets

As you learned earlier in this lesson, a worksheet is a single sheet in Excel. When you first start Excel and open a blank workbook, Excel automatically gives you a single worksheet and names it "Sheet1":



A worksheet is the same thing as a tab, and you will see these terms used interchangeably. One of the first things you should do when you're working with Excel is to make sure that your worksheets have meaningful names. For starters, "Sheet1" is not a good name! The great thing is that you have a lot of flexibility when renaming your worksheets. Here are some general guidelines:

1. You can't leave the name blank. If you try this, you'll get an error message.
2. You can't use duplicate names. Since Excel ignores upper and lower case, the names "DATA" and "data" are equivalent, so you can't use both!
3. You can't use more than 31 characters (and you probably don't want your names to be very long as this will take up too much space).
4. You can't use the word "history", because it's a reserved name by Excel to track changes.
5. You can't use special characters, such as : \ / ? * [or]

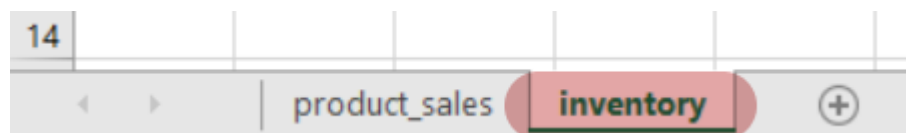
The fastest way to rename a tab is to double-click on the tab name. Alternatively, you can right-click on the tab name and select **Rename**, but that involves an extra step.



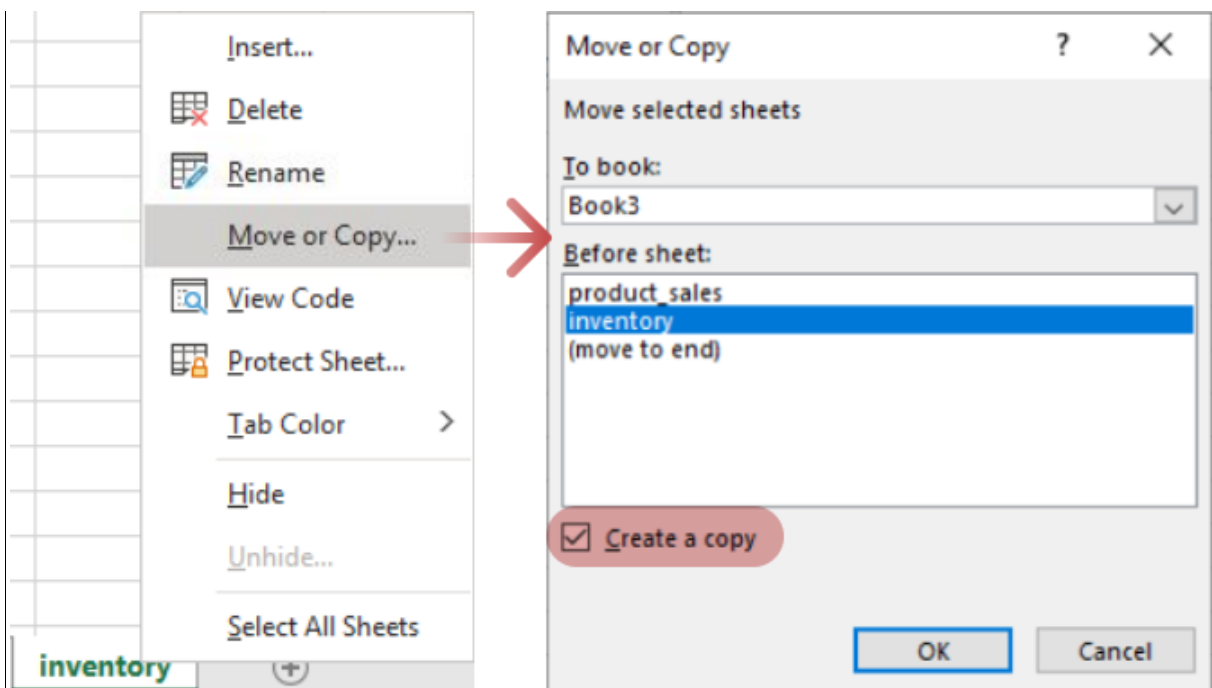
You can also insert a new tab by clicking on the plus sign "+" to the right of the final tab.



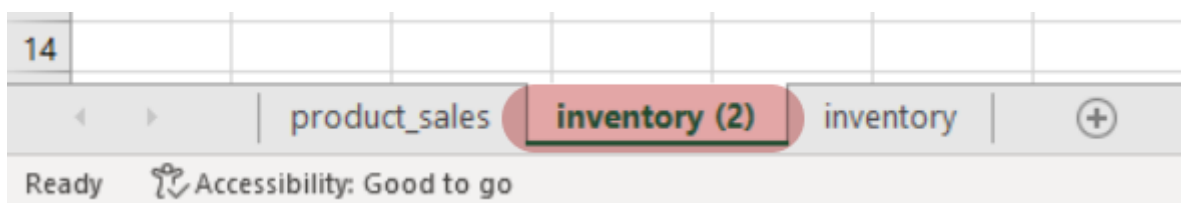
You could rename the new tab "inventory."



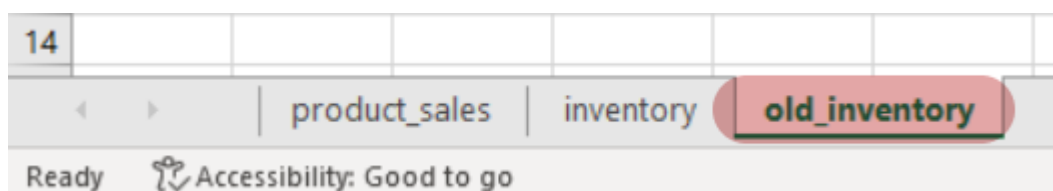
Let's say you have a worksheet and you want to make a copy. You can do this by holding down **CTRL** and then using your left mouse button to drag the tab name over one position. Alternatively, you can right-click and choose **Move or Copy...**, and you'd see this dialog box:



Make sure the box for **Create a copy** is checked, otherwise you'll only be *moving* the tab, not *copying* it. Either way you do it, you'll get a duplicate copy that has been renamed with a number after it. If you were to do this using the "inventory" tab, you would see "inventory (2)":

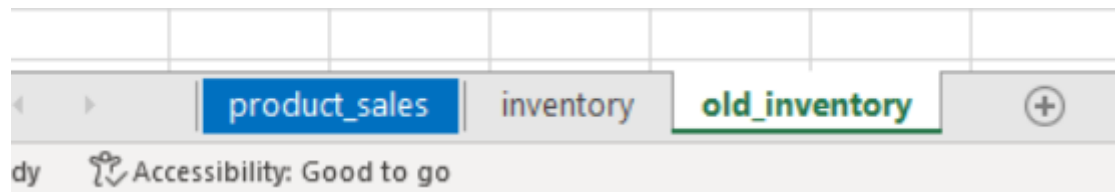


As you know, it's always a good idea to use meaningful names so you would change the name of the copied tab to "old_inventory". Also, you may not like where this tab is located, so you could move it using your mouse by clicking on the tab name and dragging it to the desired location.



If you want to delete a worksheet, right-click on the tab name and select **Delete**. If your worksheet has any content, you'll be asked to confirm that you want to delete it because you can't "undo" it!

Finally, another nice feature is the ability to set a tab color, which can help you locate tabs quickly if you have a workbook that contains multiple tabs. You can really see the color when you select another tab:



To change the color of a tab, right-click on the tab name, and select **Tab color**, and then select a color.

On the next screen, you'll see how worksheets interact with workbooks.

Instructions

1. Open a new instance of Excel by selecting **File > New** and selecting **Blank workbook**.
2. Rename the tab **Sheet1** to **product_sales**.
3. Insert a new tab by clicking on the plus sign **+** to the right of the last tab.
4. When you add the new tab, rename it to **inventory**.
5. Make a copy of the **inventory** tab using either method shown above.
6. Rename the **inventory (2)** tab to **old_inventory**.
7. Move the **old_inventory** tab so that it is to the right of your other two tabs.
8. Change the tab color for **product_sales**. The example above used blue, but you can choose whatever color you like.

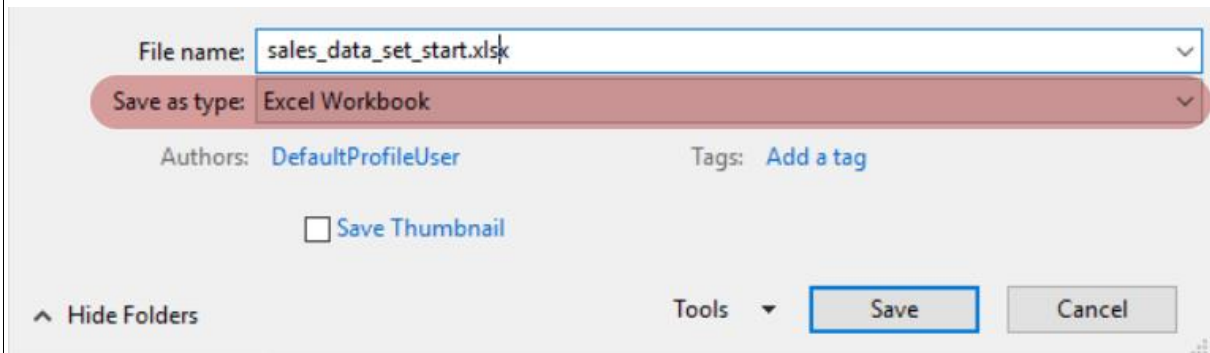
2.5 Workbook

An Excel workbook is a collection of one or more worksheets. A workbook can never have less than one worksheet, and the maximum number of worksheets per workbook is limited only by your computer's memory.

A workbook is the actual file you create when you save your work, and it ends in .XLSX. As we mentioned in the previous lesson, we always recommend that you save and/or convert your workbooks to the latest workbook version available to maximize the size and efficiency benefits offered by using the latest format.

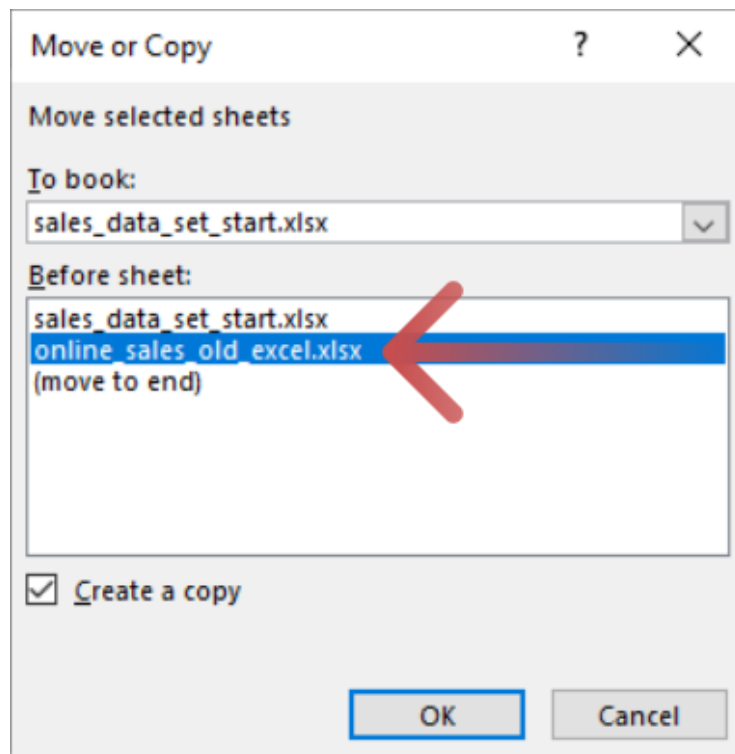
The exception to this is if you have a non-XLSX file that you'll be sharing with other non-Excel applications, such as databases. In those situations, it may be beneficial to keep the files in a common format that is friendly to all the applications that will use it. That type of

file format may be CSV or TXT, or something else. That's a decision you'll need to make when considering how you'll use the dataset. One way you can change the format of your workbook is to specify the type when you save it. You saw this in the previous lesson when you went to save a workbook by using **File > Save As**:



Because an Excel workbook is an actual computer file, many of the ways you manipulate it will involve using your computer's operating system, which is likely Windows or Mac OS. Using your operating system, you can rename, delete, move, and copy your Excel workbook just like you would do with any other file, so we won't cover those processes in this lesson.

You can also have multiple workbooks open at the same time, and Excel will treat this as having individual sessions of Excel running simultaneously. If you were to open the workbook `sales_data_set_start.xlsx` and also open the workbook `online_sales_old_excel.xlsx`, when you use **Move or Copy** on the `sales_data_set_start` tab of `sales_data_set_start.xlsx`, you'd select `online_sales_old_excel.xlsx` under **To book** and you'd see this dialog box:



You need to know that you can only copy a worksheet to another workbook when the other workbook is currently open — that's why you had both workbooks open at the same time! Now, you'll see a new worksheet in the workbook `online_sales_old_excel.xlsx` called "sales_data_set_start," just like the name of the original worksheet. You may also notice that whoever created the worksheet in the `online_sales_old_excel.xlsx` workbook never renamed the original sheet and just left it as "Sheet1." This isn't a good idea - we recommend you give worksheets meaningful names.

On the next screen, you'll see how to use some Excel functions to complement what you've learned thus far in this lesson.

2.5.1 Instructions

1. Open the workbook `sales_data_set_start.xlsx`.
2. Open the workbook `online_sales_old_excel.xlsx`.
3. Switch back to the workbook `sales_data_set_start`.

- You can switch between workbooks under **File** and making the appropriate selection under `Recent`.

4. Use **Move or Copy** to make a copy of your `sales_data_set_start` worksheet and send it to the workbook `online_sales_old_excel`.

- Be sure to enable **Create a copy** when using **Move or Copy**.
- If you get a message about "Keep an eye on it", click on "Got it" to proceed.

- After copying the worksheet from `sales_data_set_start` to `online_sales_old_excel`, Excel automatically switches to the destination workbook: `online_sales_old_excel`.

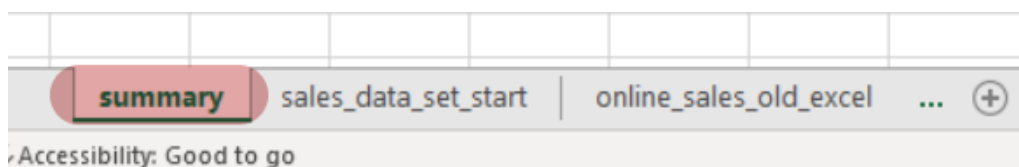
5. Rename the tab "Sheet1" to "online_sales".
6. Delete any data in cells `E1`, `F1`, `G1`, and `H1` so you are left with just the original data for `online_sales_old_excel`.
7. Save your workbook as `sales_combined.xlsx`.

2.6 Excel Functions

On this screen, you're going to complement the skills you just learned by using some Excel functions. More specifically, you're going to see how using functions along with named tables and ranges is extremely helpful when it comes to organizing your data in Excel.

Suppose, for example, that you have multiple worksheets in a workbook and you want to access data across those worksheets. What do you think is an easy way to do that so that it's understandable and easy to follow? You could do it using cell addresses, and that would work, but it wouldn't be easy to follow. The better way is to use named ranges and tables, because using a name will always make more sense than using a meaningless address made up of a column-row combination. Let's see this in action.

Suppose you were to open `sales_combined.xlsx` and insert a new tab called "summary" and move it toward the front as seen below:



Now suppose you went to tab "online_sales" and converted it to a table called "Online_Only":

You would now have three tabs:

- The "summary" tab is currently blank.
- The "sales_data_set_start" tab contains a table called "Sales_Data" that you created earlier in this lesson.
- The "online_sales" tab contains a table called "Online_Only."

Now going to the first tab, "summary," suppose you typed the following in cells `A1` and `B1`:

	A	B	C
1	Total weekly sales	Total online quantity	
2			
3			
4			
5			

Note: You may need to resize your columns for the data to display correctly.

Now you want to access the information in your tables. Here's the exciting part where you use named tables and ranges. Under "Total weekly sales," you want the sum of all sales in that column. Suppose that in cell **A2** you entered the formula `=SUM(Sales_Data[Weakly Sales])`.

Let's break this apart. You already know what `=SUM(` means. That's the start of the formula. The next item, `Sales_Data`, is the name of the table you created earlier in this lesson. The next item after the `[` is the name of the range that was automatically assigned when you created the table. So, your formula is getting the sum of the range `Weakly Sales` that exists in the table called `Sales_Data`. Make sense?

Congratulations! You've successfully worked with tables, ranges, worksheets, and workbooks! Save your workbook

2.6.1 Instructions

1. Open the workbook you created on the previous screen, `sales_combined.xlsx`. Insert a new tab called "summary," and move it toward the front as shown above.
2. Go to the tab "sales_data_set_start", and make sure it has been converted to a table called "Sales_Data" from an earlier screen in this lesson.
3. Go to the tab "online_sales", and convert it to a table, and rename it to "Online_Only."

- The range for the table should be `A1:D4`.

4. On the "summary" tab in cell **A1**, type "Total weekly sales" and in cell **B1**, type "Total online quantity".

- You may need to resize your columns.

5. In cell **A2**, enter the formula `=SUM(Sales_Data[Weakly Sales])`.
6. Similarly, in cell **B2**, enter a formula that uses the "Online_Only" table and the named range "qty" to get the sum of all items sold online.

3 Cleaning Data Part 1

3.1 Introduction

In the previous lesson, you learned how to organize your data into a spreadsheet using various spreadsheet structures. After you've organized your data, you need to clean it for analysis.

Cleaning your data involves things like removing duplicate data, removing extra spaces, and correcting inaccurate or missing data.

Cleaning your data is also called "manipulating" your data because it often requires you to make changes to your data (sometimes significant changes). This means that you need to "know" your data very well because your goal is to manipulate the data without damaging the integrity of the data.

Suppose you have an old family photograph of a favorite pet that you'd like to restore — a dog named Duncan. Imagine that the photograph is mostly intact, but it has some rips, stains, and some fading. Your best bet is to create a digital scan of the photograph and manipulate it at the pixel level. In this way, you can think of a digitized photograph as being a type of visual dataset. With your photograph, you have most of the data you need, but there are some problem areas you need to fix to complete the "whole picture."

You can restore the photo because you know what your dog should look like — you "know" your data! If the part of the picture with the dog's nose is missing, you wouldn't replace it with a strawberry! And if there is a rip through the picture where your dog's fur should be, you can take pixels from the surrounding fur of the same color and fill it in because that makes sense. The surrounding data combined with your knowledge of the overall data helps you fill in the missing pieces.

When you're working with datasets that have "rips, stains, and fading," you'll do the same thing. You'll use your knowledge of the dataset to restore it to a complete picture that you can analyze with confidence!

3.1.1 Data cleaning processes

Data cleaning can take quite a bit of time, and it often involves several processes. In fact, there are so many processes to review, we'll take the next two lessons to cover them all. Here are the high-level processes required for cleaning data:

- Finding and removing duplicate data
- Finding and removing extra spaces
- Separating data without delimiters

- Correcting inaccurate or irrelevant data
- Updating missing or incomplete data
- Standardizing the case for text
- Standardizing formats for each data type (dates, numbers, etc.)
- Working with outliers
- Visualizing to verify clean data

It's time to get started on this very important piece of organizing your data!

3.2 Removing Duplicate Data

On the previous screen, we went through a high-level overview of the processes involved in cleaning your data. One of the first things you should do when cleaning your data is to remove duplicate data. Removing duplicate data will simplify the rest of your tasks because it will trim the dataset by the amount of duplicate data you have, meaning that you won't spend time cleaning data that is redundant (i.e., data that you'll eventually remove anyway).

So, what exactly is **duplicate** data? To start, duplicate data **doesn't** mean **data values that are the same**. For example, you may have a sales dataset that shows multiple customers purchased the same item — let's say "earbuds." Your dataset will show "earbuds" multiple times, and that's perfectly fine because each of the earbuds was part of a unique transaction, which is represented as an individual row (also called a **record**) in a worksheet, like this:

Transaction ID	Customer #	Item
1470000	341	earbuds
1470001	242	earbuds
1470002	547	earbuds
1470003	751	earbuds
1470004	412	earbuds

So in the above example, "earbuds" is **not duplicate data** just because it appears more than once in the dataset.

However, things become problematic when the entire row is duplicated and there is nothing unique about it. Here you can see two identical rows where the **Transaction ID** is 1470001, so we should remove one of them:

Transaction ID	Customer #	Item
1470000	341	earbuds
1470001	242	earbuds
1470001	242	earbuds
1470002	547	earbuds
1470003	751	earbuds
1470004	412	earbuds

The above example is a classic case of duplicated data. The reasons for duplication are many, and duplicated data is a very real problem that you'll encounter frequently as a data analyst. Now let's look at some duplicated data in a dataset, and you'll learn ways to correct it!

3.2.1 A scenario

Imagine that you're a data analyst working for an online retail store. You're part of a team of analysts investigating sales information, and it's your role to clean up the dataset prior to analysis. You've been given the file `features_start.xlsx`, which is based on an actual retail dataset that you can find [here](#). The original dataset is much larger, and it's a CSV file as opposed to the XLSX file you will be using during the next few lessons. Fortunately, if you wanted to use the original dataset to explore and learn more on your own, you have experience working with CSV files, so you'll know how to work with it! Suppose you open the XLSX file, and you see something like this:

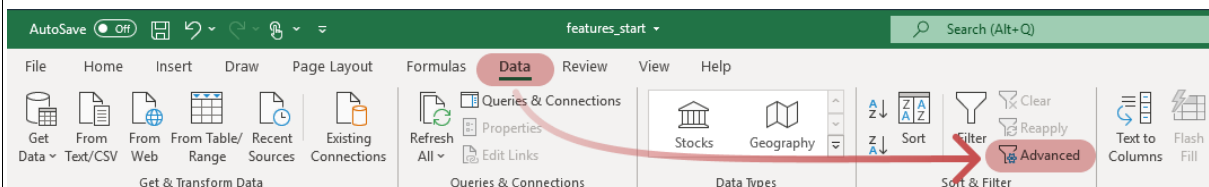
	A	B	C	D	E	F	G	H	I	J	K	L
1	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
2	1	2/5/2010	42.31	2.572	NA	NA	NA	NA	NA	211.0963582	8.106	FALSE
3	1	12-Feb-10	38.51	2.548	NA	NA	N/A	NA	NA	211.2421698	8.106	TRUE
4	1	2/19/2010	39.93	2.514	NA	NA	NA	NA	NA	211.2891429	8.106	FALSE
5	1	2/26/2010	46.63	2.561	NA	NA	NA	NA	NA	211.3196429	8.106	FALSE
6	1	Friday, March 5, 2010	46.5	2.625	NA	NA	NA	NA	NA	211.3501429	8.106	FALSE
7	1	3/12/2010	57.79	\$2.67	NA	NA	NA	NA	NA	211.3806429	8.106	FALSE
8	1	3/19/2010	54.58	2.72	NA	NA	NA	NA	NA	211.215635	8.106	FALSE
9	1	3/26/2010	51.45	2.732	NA	NA	NA	NA	NA	211.0180424	8.106	FALSE
10	1	4/2/2010	62.27	2.719	NA	NA	NA	NA	NA	210.8204499	7.808	FALSE
11	1	4/9/2010	65.86	2.77	NA	NA	NA	NA	NA	210.6228574	7.808	FALSE
12	1	4/16/2010	66.32	2.808	NA	NA	NA	NA	NA	210.4887	7.808	FALSE
13	1	Friday, April 23, 2010	64.84	2.795	NA	NA	NA	NA	NA	210.4391228	7.808 F	
14	1	Friday, April 23, 2010	64.84	2.795	NA	NA	NA	NA	NA	210.4391228	7.808 F	
15	1	Friday, April 23, 2010	64.84	2.795	NA	na	NA	NA	NA	210.4391228	7.808 F	
16	1	4/30/2010	67.41	\$2.78	NA	na	NA	NA	NA	210.3895456	7.808	FALSE
17	1	5/7/2010	72.55	2.835	NA	na	NA	NA	NA	210.3399684	7.808	FALSE
18	1	5/14/2010	74.78	2.854	NA	NA	NA	NA	NA	210.3374261	7.808	FALSE
19	1	5/21/2010	76.44	2.826	NA	na	NA	NA	NA	210.6170934	7.808	FALSE
20	1	5/28/2010	80.44	2.759	NA	NA	NA	NA	NA	210.8967606	7.808	FALSE
21	1	Friday, June 4, 2010	80.69	2.705	NA	NA	NA	NA	NA	211.1764278	7.808	FALSE
22	1	6/11/2010	80.43	2.668	NA	NA	NA	NA	NA	211.4560951	7.808	FALSE
23	1	6/18/2010	84.11	2.637	NA	NA	NA	NO	NA	211.4537719	7.808	FALSE
24	1	25-Jun-10	84.34	2.653	NA	NA	NA	NA	NA	211.3386526	7.808 F	
25	1	7/2/2010	80.91	2.669	NA	NA	NA	NA	NA	211.2235333	7.787	FALSE
26	1	7/9/2010	80.48	2.642	NA	NA	NA	NA	NA	211.108414	7.787	FALSE
27	1	Friday, July 16, 2010	83.15	2.623	NA	NA	NA	NA	NA	211.1003854	7.787	FALSE

It's always a good idea to visually scan your dataset for a few minutes to give yourself an idea of what you're working with. As you do so, you would ask yourself questions like these:

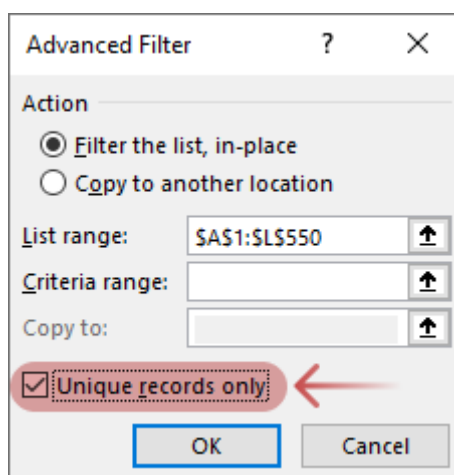
- How many columns and rows do you have?
- Does the dataset have headers, and do they make sense to you?

- Do you immediately see any problems, such as duplicate rows, inconsistent data, or missing data?

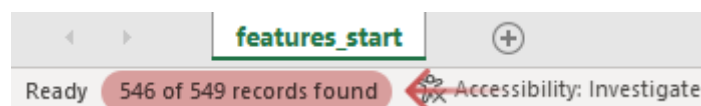
You discover that there are plenty of opportunities to clean up this dataset! As stated above, your first job is to look for duplicate rows. There are a couple of different ways you can do this. One quick way to identify duplicate rows is to first select the entire range of data (using **CTRL+A** works well for this). Then, you would go to **Data, Sort & Filter, Advanced**:



After doing the above, you'd see this dialog box confirming your range. Notice the check box in the lower left that says Unique records only. (Note that *record* and *row* are the same thing, and you will hear those terms used interchangeably often). This is the key to what you're trying to do.



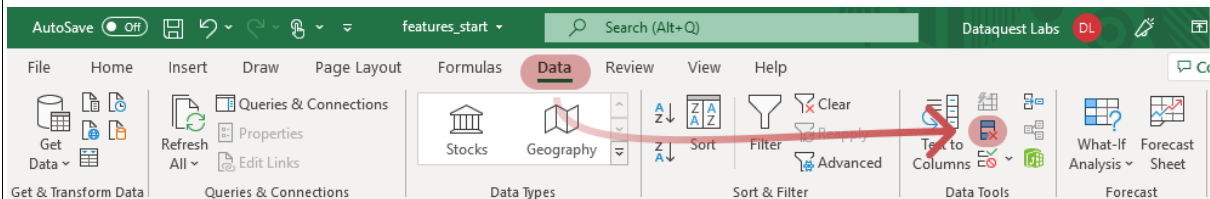
After you do this, if you look in the lower left corner of your status bar, you'll see that Excel found 546 unique records out of 549 (note that Excel did *not* count the header row)!



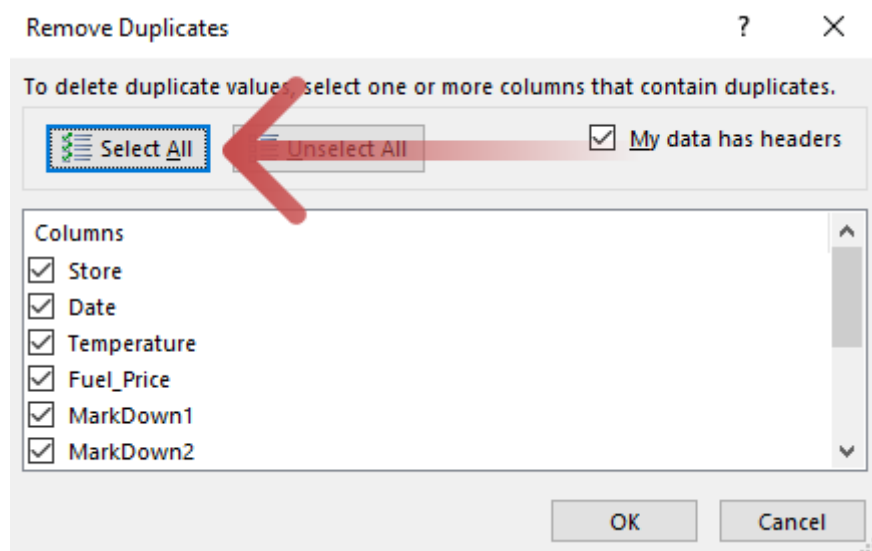
This tells you that you've got three duplicate records, but it doesn't actually do anything with them. Excel simply filtered the data so that only unique records are being displayed, but the duplicates are still in the workbook. You can see some rows aren't displaying when you look at row 13 because it immediately jumps to row 16. The rows that have been skipped are some of the duplicate rows.

11	1	4/9/2010	65.86
12	1	4/16/2010	66.32
13	1	Friday, April 23, 2010	64.84
16	1	4/30/2010	67.41
17	1	5/7/2010	72.55
18	1	5/14/2010	74.78

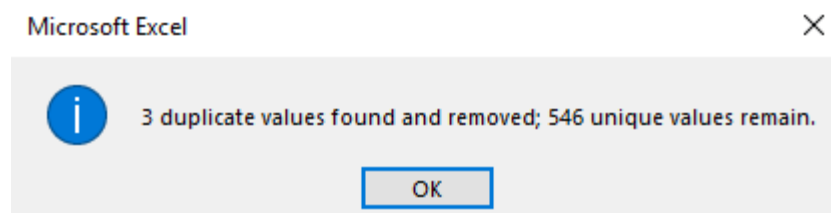
To actually remove the duplicate rows from the workbook, you'd again select the entire range (using **CTRL+A**), then go to **Data, Data Tools, Remove Duplicates**:



You would then see this dialog box, where you'd want to click **OK**:



After you clicked **OK**, you would see this confirmation dialog box showing you how many records were removed:



Now your dataset has been cleaned of all duplicate records!

3.2.2 Instructions

We've provided a solutions file (`features_level_1_solutions.xlsx`) in case you get stuck or would like to verify your answers while working on the exercises in this lesson.

1. Open the file `features_start.xlsx`, review it for a few minutes, and then save it as `features_level_1.xlsx`.

- You'll build on this file for the rest of the lesson, so please keep it open as you move through the lesson screens.

2. Select the entire range of data.
3. Use **Data, Sort & Filter, Advanced**, and then use the dialog box to confirm your range. Check the box for **Unique records only**, and click **OK**.
4. Use **CTRL+Z** to undo your last action and restore the full view of the dataset (i.e., no skipped rows).
5. To remove your duplicate rows, select the entire range, then go to **Data, Data Tools, Remove Duplicates**.
6. When the dialog box appears, click **OK** and confirm that the correct number of records were removed.
7. Save your file.

3.3 Removing Extra Spaces

On the previous screen, you learned how to remove duplicate rows. Now that you have a set of unique rows, you need to look at ways to clean the individual data values. Before we get started, let's take a quick quiz. What's the difference between the two data values below?

- `earbuds`
- `earbuds`

These values look identical. However, behind the scenes, the second value actually has three spaces before and three spaces after the word `earbuds`. It is being rendered on your screen to look the same as the first value, but it isn't! Here is how the two data values really look when we replace the spaces with dashes:

- `earbuds`
- `---earbuds---`

This phenomenon is *extremely* common in datasets, especially because spaces (and tabs) are unseen characters. They are there, but they may not be obvious to the viewer. The spaces before the value are called *leading* spaces, and the spaces after the value are called *trailing* spaces. Many applications (including this viewer) will strip off the spaces for you as a matter of convenience for display purposes. But, when the data is used for analysis, you can't be guaranteed that you'll get the same courtesy.

Thankfully, there's a straightforward way to remove leading and trailing spaces from your data using a very common Excel function called **TRIM**. As the name implies, the **TRIM** function will trim, or strip off, leading and trailing spaces from data values. This function takes one argument, and that is just the cell value itself. So, the formula will look like this **=TRIM(L2)**, where **L2** is the cell value you're trimming. Note that the **TRIM** function will *not* remove spaces from inside text that exists between other characters. For example, using **TRIM** on "Hello world," as in **=TRIM("Hello world")** would *not* affect the space between the two words.

As mentioned on the previous screen, you'll continue to work with the same dataset you worked on in the last screen called **features_level_1.xlsx**. Please go ahead and open that workbook if it isn't already opened. For the purposes of teaching this skill, you'll be focused on **Column L**, which is named **IsHoliday**. Take a few moments to scroll up and down and look at that column. As you do so, you'll see a variety of problems. In particular, if you select cell **L2**, you'll notice the word "FALSE" doesn't line up with the other words. If you place your cursor in the formula bar, you'll see that there are 12 leading spaces before the word "FALSE."



Cell **L42** has a similar problem, only it has *both* leading and trailing spaces. Before you use the **TRIM** function, you may be wondering if there's a simple way to know if your data contains any leading or trailing spaces. There are two Excel functions that can help you with this. Those functions are the **LEFT** and **RIGHT** functions, which both return characters from the left-most and right-most part of a cell, respectively. To start, you would need to create some "temporary" columns in your worksheet, which is a very common practice in data analytics.

Suppose you added three new headers at the top of Columns **M**, **N**, and **O** and named them **left space**, **right space**, and **trimmed**. Then suppose you used the **LEFT** function in cell M2 and entered **=LEFT(L2,1)**:

=LEFT(L2, 1)												
D	E	F	G	H	I	J	K	L	M	N	O	P
Price	MarkDown	MarkDown	MarkDown	MarkDown	MarkDown	CPI	Unemployment	IsHoliday	left space	right space	trimmed	
2.572	NA	NA	NA	NA	NA	211.0963582	8.106	FALSE	=LEFT(L2, 1)			
2.548	NA	NA	N/A	NA	NA	211.2421698	8.106	TRUE				
2.514	NA	NA	NA	NA	NA	211.2891429	8.106	FALSE				
2.561	NA	NA	NA	NA	NA	211.3196429	8.106	FALSE				
2.625	NA	NA	NA	NA	NA	211.3501429	8.106	FALSE				

The **LEFT** function has the format **LEFT(text to evaluate, number of characters to take)**. So, you just asked Excel to take the value in cell **L2** and take one character from the left-most position. This means that if the value in cell **L2** has at least one leading blank space, you'll see nothing returned, which is what you see because you already know that cell **L2** has at least one leading space.

L	M	N	O
IsHoliday	left space	right space	trimmed
FALSE			
TRUE	T		
FALSE	F		
FALSE	F		
FALSE	F		
FALSE	F		
FALSE	F		
FALSE	F		
FALSE	F		
FALSE	F		
F	F		
FALSE	F		

Now suppose you copied the formula all the way down to Row 547. If you did so, you'd see that cell **L42** is also blank. It's time to apply the same approach by using the **RIGHT** function to find the trailing spaces. The **RIGHT** function works the same way as the **LEFT** function, only it takes a certain number of characters from the right-most position.

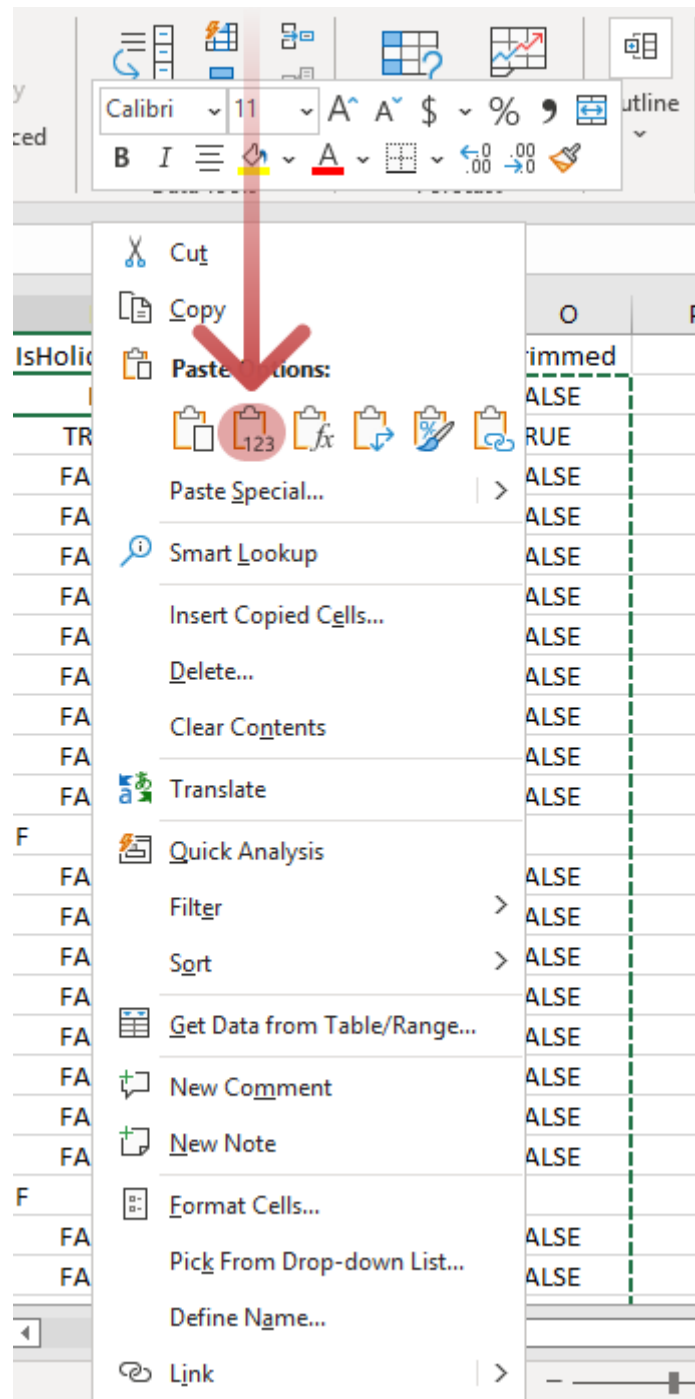
Now suppose in cell **N2**, under the header "right space," you entered the formula **=RIGHT(L2, 1)** and copied it all the way down the rows. What would you see? This time, you'd see that row 42 as well as rows 73 through 78 all have trailing spaces! This means you would need to do some trimming!

To do this, in cell **O2** you would type in the formula **=TRIM(L2)** and copy it all the way down the rows by double-clicking on the **fill handle** (the small square in the lower right of the selected cell). After double-clicking the fill handle, the column will **autofill** with the correct formula for each row.

L	M	N	O	P
IsHoliday	left space	right space	trimmed	
FALSE		E		
TRUE	T	E		
FALSE	F	E		
FALSE	F	E		
FALSE	F	E		
FALSE	F	E		
FALSE	F	E		
FALSE	F	E		
FALSE	F	E		
FALSE	F	E		
FALSE	F	E		
F	F	F		
FALSE	F	E		
FALSE	F	E		

You may be wondering why you'd want to apply the **TRIM** function to all the rows, even those that don't need trimming. The main reason is because it's possible that you may have visually missed some blank cells when you reviewed your results from columns **M** and **N**. The second reason is because it won't hurt anything, and it makes the formulas in your workbook more consistent.

As a final step for this scenario, you would copy all the cell values in the range **O2:O547** and then paste them "as values" over the original values in column **L**.



Now let's practice removing extra spaces!

3.3.1 Instructions

1. In your worksheet, create three new headers at the top of Columns **M**, **N**, and **O**, and name them **left space**, **right space**, and **trimmed**, respectively.
2. In cell **M2**, under the header **left space**, enter the formula **=LEFT(L2,1)** and hit **Enter**.

3. Copy the above formula all the way down to row 547 using autofill as is shown in the animation above.
4. In cell **N2**, under the header **right space**, enter the formula **=RIGHT(L2,1)**, and hit **Enter**. As before, copy this formula all the way down the rows using autofill.
5. In cell **O2**, under the header **trimmed**, enter the formula **=TRIM(L2)**, and hit **Enter**. As before, copy this formula all the way down the rows using autofill.
6. Copy all the cell values in the range **O2:O547**, and then paste them over the original values in column **L**, and be sure to use the option **Paste as Values**.
7. After you've pasted your values, delete columns **M**, **N**, and **O**.
8. Save your workbook.

3.4 Separating Data

On the previous screens, you learned how to remove unwanted rows and spaces. Another common data manipulation method is to separate data, usually text data, but you can also separate numerical data. You have experience separating data using delimiters, such as commas and other designated characters. However, there are often situations when you cannot use delimiters to separate data, or doing so may not be so straightforward.

Suppose that you have a list of ten-digit telephone numbers including the area codes, but they are given to you without any characters in-between the numbers. Suppose the first five rows look like this, including the header row:

phone_number

1115553434

1115558787

1115553425

1115559087

1115554410

A ten-digit phone number is comprised of the following:

- An **area code** (first three numbers)
- An **exchange number** (next three numbers)
- A **line number** (last four numbers)

For example, the customer service number for Microsoft support is 800-936-5700. The area code is **800**, the exchange number is **936**, and the line number is **5700**.

Now suppose that you need to separate the above list of phone numbers into area code, exchange, and line number so that you can analyze them as individual fields. If you were to copy and paste them into a new worksheet, you'd see something like this:

	A	B
1	phone_number	
2		
3	1115553434	
4		
5	1115558787	
6		
7	1115553425	
8		
9	1115559087	
10		
11	1115554410	
12		
13		
14		

Suppose that you removed the blank lines between the data and inserted headers in columns **B**, **C**, and **D** called **area_code**, **exchange**, and **line_number**. If so, you'd have something like this:

	A	B	C	D	E
1	phone_number	area_code	exchange	line_number	
2	1115553434				
3	1115558787				
4	1115553425				
5	1115559087				
6	1115554410				
7					
8					
9					

This is the part where you would separate the actual data. You would use some familiar Excel functions to do this. In particular, you'd use the `LEFT` and `RIGHT` functions, and you'll learn about a new function called `MID`. Before we jump into the formulas, let's take a few moments to plan out what we need to do.

Your goal is to separate each phone number into its three components. You also know that for every ten-digit phone number, you can be certain that the first three digits will be the area code, the next three digits will be the exchange, and the last four digits will be the line number. If you think about it in terms of individual characters, you can take out the three left-most characters, then take out three more characters starting at the fourth position, then take out the four right-most characters. Now that you have a plan, it's time to take your plan and implement it using your functions.

You already know that the `LEFT` function takes a certain number of characters from the left-most part of a value, and you know the `RIGHT` function takes a certain number of characters from the right-most part of a value. In this case, you'd enter your formulas for `LEFT` and `RIGHT`, which would be used for `area_code` and `line_number` in columns `B` and `D`, respectively. Then you'd copy your formulas down all the rows using autofill.

The new function you'll learn about, `MID`, takes a certain number of characters starting from a given location. It has the format `=MID(text to evaluate, starting position, number of characters to take)`. In your scenario, for `C2`, you would want it to look like `=MID(A2,4,3)`. This means you're asking Excel to take the value in cell `A2`, and starting with the fourth character, take the next three characters. If you were to enter that formula and copy it down all the rows using autofill, you'd now have:

	A	B	C	D	E
1	phone_number	area_code	exchange	line_number	
2	1115553434	111	555	3434	
3	1115558787	111	555	8787	
4	1115553425	111	555	3425	
5	1115559087	111	555	9087	
6	1115554410	111	555	4410	
7					
8					
9					

Now you know how to separate data for easier analysis!

3.4.1 Instructions

1. Open the workbook `phone_numbers.xlsx`.

2. Insert headers in columns **B**, **C**, and **D**, and name them **area_code**, **exchange**, and **line_number**, respectively.
3. Enter formulas for **=LEFT** and **=RIGHT** to extract the appropriate numbers for **area_code** and **line_number** in cells **B2** and **D2**, respectively. Copy your formulas down all the rows using autofill.
4. In cell **C2**, enter the formula **=MID (A2, 4, 3)**. Copy the formula down all the rows using autofill.

3.5 Correcting Inaccurate Data

On the previous screens, you learned how to remove unwanted rows and spaces, and how to separate data. On this screen, you'll spend some time correcting inaccurate data. Inaccurate data is data that doesn't make sense, or data that will cause errors or problems during analysis. The reasons for inaccurate data are many, and inaccurate data can manifest itself in many ways.

One of the best ways to determine if data is accurate or not is to *know your data*. You've heard this plenty of times before, and you'll hear it repeated throughout this course. Let's take a look at some examples of inaccurate data.

Suppose you're analyzing data from a survey that took place in the last six months of 2021, and one of the pieces of information you collected is the date, which shows the date that the survey was conducted. Your data looks like this:

date
11/21/21
11/22/21

date
11/21/29
11/23/21
11/24/21

At first glance, this may look OK. However, if you look more closely at the third row of data, **11/21/29**, you'll see that the day and year are switched. If the survey took place in 2021, then how could there be data for 2029 — especially if that date is in the future? More than likely, the data should look like **11/29/21**. A follow-up investigation may be necessary to confirm, but it appears this is a transposition error.

Now suppose that you have data that shows the number of cars sold at two local dealerships in the last three days. The data looks like this:

dealership A: 5, 7, 4

dealership B: 4, 5, 2.45

This data should make you suspicious about the number of cars sold at dealership B, because you can't sell 2.45 cars; this data should be discrete, not continuous!

Let's look at another type of inaccurate data. Suppose you've been given data on temperatures in a local pond taken over the last five days. Here is the data:

temperature
42.7
51.3
47.2
fifty
52.5

This one may be rather obvious because a temperature should be a number, not a word like "fifty." Not all data inaccuracies will be this obvious. Some are very subtle.

The following is a data problem that has been plaguing computer programmers since computers were invented. See if you can find the problem in this set of data that shows prices for online products:

price
11.29
15.9O
21.07
17.21
23.47

If you guessed the second row with the price **15.90**, you're correct! The problem is that the last digit in **90** is not the number zero, but the capital letter O! The correct entry should be **90** with a zero at the end. If you look at the next line with **21.07** you'll see that the zero looks different because it's a zero, as it should be. Often, the zero will have a diagonal line to distinguish it from the letter O. This remains a common problem today, and you may even see situations where the number one (1) is replaced with the lowercase letter l. Be on the lookout for situations like this where numbers and letters are used interchangeably, as this can wreak havoc on your analysis.

Now, time to pick up where you left off on a previous screen. Suppose you have the workbook **features_level_1.xlsx** open and you've examined it. You have gone column by column, slowly scrolling down the sheet. You've noticed problems in column **A**, **Store**. There are many inconsistencies with the data:

	A	B	C	D
32	1	9/3/2010	81.21	2.577
33	1	9/10/2010	78.69	2.565
34	1	9/17/2010	82.11	2.582
35	1	9/24/2010	80.94	2.624
36	1	10/1/2010	71.89	2.603
37	one	Friday, October 8, 2010	63.93	2.633
38	1	10/15/2010	67.18	2.72
39	1	10/22/2010	69.86	2.725
40	1	10/29/2010	69.64	2.716
41	1	11/5/2010	58.74	2.689

For column **B**, **Date**, you'll see many formatting inconsistencies, which you'll fix in the next lesson. And you'll even see a couple of missing dates, which you'll address on the next screen. Otherwise, there really isn't any inaccurate data, so we move on to the next column.

Looking at column **C**, **Temperature**, it mostly looks okay except for cells **C133** and **C339** that say "hot" and "cold," respectively.

Column **D**, **Fuel Price**, looks okay except that it has inconsistent formatting. Some values have dollar signs (\$), while most do not. Also, the number of decimal places doesn't match. As with column **B**, we'll address formatting in the next lesson.

The analysis team has just informed you that they **do not** need the data in columns **E** through **I**, the columns called **Markdown1-5**. Look at the data. You'll see inconsistent usage of "NA" throughout those columns. How would you go about cleaning that up? If you were to delete those columns, you'd see this:

	A	B	C	D	E	F	G	H
1	Store	Date	Temperature	Fuel_Price	CPI	Unemployment	IsHoliday	
2	1	2/5/2010	42.31	2.572	211.0963582	8.106	FALSE	
3	1	12-Feb-10	38.51	2.548	211.2421698	8.106	TRUE	
4	1	2/19/2010	39.93	2.514	211.2891429	8.106	FALSE	
5	1	2/26/2010	46.63	2.561	211.3196429	8.106	FALSE	
6	1	Friday, March 5, 2010	46.5	2.625	211.3501429	8.106	FALSE	
7	1	3/12/2010	57.79	\$2.67	211.3806429	8.106	FALSE	
8	1	3/19/2010	54.58	2.72	211.215635	8.106	FALSE	
9	1	3/26/2010	51.45	2.732	211.0180424	8.106	FALSE	
10	1	4/2/2010	62.27	2.719	210.8204499	7.808	FALSE	
11	1	4/9/2010	65.86	2.77	210.6228574	7.808	FALSE	

You have three more columns to look at for inaccuracies, which are now columns **E**, **F**, and **G**. The columns named **CPI** and **Unemployment** (columns **E** & **F**) both look pretty good

except for some missing values that we'll address on the next screen. We won't make any changes to these columns right now.

Time to move on to column **G**, **IsHoliday**. In this column, you'll see many inaccuracies. The word "False" might be "No" or "F", or it may be lowercase (which is a format inconsistency issue).

Follow the instructions below to clean this dataset using what you just learned.

3.5.1 Instructions

1. Return to the workbook **features_level_1.xlsx**.
2. Fix the values in column **A** for rows 37, 93, 191, 201, 376, 385, 418, 428, and 482 by replacing them with the proper numbers that surround them above or below (i.e., stores 1, 2, or 3).
 - In cell **A37**, you can replace "one" with 1.
 - In cell **A93**, you can also replace that value with 1.
 - Continue this for all the rows until they have their appropriate store numbers.
3. In cells **C133** and **C339** that say "hot" and "cold," delete those values for now.
4. Delete all the columns **E** through **I**, **Markdown1-5**, by selecting the column and then **Home, Cells, Delete**. You can also select the columns and right-click to select **Delete**.
5. Assuming that "T" and "yes" are **TRUE**, and "F" and "no" are **FALSE**, use your skills to correct the values in the **IsHoliday** column so that it contains only **TRUE** and **FALSE** values. Lowercase versions do not need to be changed.
6. Save your workbook.

3.6 Updating Missing Data

On the previous screen, you learned how to resolve inaccurate data. Along the way, you saw missing data that was also problematic. On this screen, you'll see how to manage missing data in your datasets.

Missing data is a big problem with datasets, and it's something you'll encounter all of the time in your career as a data analyst. Fixing missing data has a lot in common with fixing inaccurate data, in that you need to know something about your data before you just start entering new data in a blank cell.

The process of fixing missing data is also known as *imputation*, which is a term you may hear from other data specialists. To better understand imputation, let's go back to our original example of the photograph of your family dog, Duncan. Suppose that there's a rip in the photo that cuts through the middle of where the dog's tail is, and the tear is only 18 inch wide. You can safely assume that the missing 18-inch gap should have the same look as the remaining part of the tail. To fix this, you would copy and paste pixels from the surrounding areas to fill in the missing data.

Fixing missing data in your dataset is no different. You already learned from your work on the previous screens that you have missing data. You'll use what you know about the surrounding data to fill in the gaps, just like in your photo.

If you look at the workbook *features_level_1.xlsx*, you'll notice there are many places where the store number is missing. In every case, the store number is surrounded by other store numbers that are consistently the same.

	A	B	C	D	E	F	G	H
57	1	2/25/2011	62.9	3.065	213.535609	7.742	FALSE	
58	1	3/4/2011	59.58	3.288	213.8233327	7.742	FALSE	
59	1	3/11/2011	53.56	3.459	214.1110564	7.742	FALSE	
60		3/18/2011	62.76	3.488	214.3627114	7.742	FALSE	
61	1	3/25/2011	69.97	3.473	214.5999389	7.742	FALSE	
62	1	4/1/2011	59.17	3.524	214.8371664	7.682	FALSE	
63	1	4/8/2011	67.84	3.622	215.0743939	7.682	FALSE	

In this situation, you can insert the number of the store that matches the numbers around it.

Also, it appears that a couple of the dates are missing in the *Date* column. Take a few moments to inspect this column and try to determine how you might fill those in. As a data analyst, you'll often need to apply critical thinking skills that require you to investigate the data on a deeper level. This only comes with practice!

	A	B	C	D	E	F	G	H
191	two	3/26/2010	51.26	2.732	210.6766095	8.324	FALSE	
192	2	4/2/2010	63.27	2.719	210.4798874	8.2	FALSE	
193	2	4/9/2010	65.41	2.77	210.2831653	8.2	FALSE	
194	2	4/16/2010	68.07	2.808	210.1495463	8.2	FALSE	
195	2		65.11	2.795	210.1000648	8.2	FALSE	
196	2	4/30/2010	66.98	2.78	210.0505833	8.2	FALSE	
197	2	5/7/2010	71.28	2.835	210.0011018	8.2	FALSE	
198	2	5/14/2010	73.31	2.854	209.9984585	8.2	FALSE	
199	2	5/21/2010	74.83	2.826	210.2768443	8.2	FALSE	

After some review of the *Date* column, you may have noticed that the dates are one week apart from each other. Unless you see other concerns, the best thing for you to do is to fill in those cells with a date that fits with the other dates around it. The new date should be one week after the previous row and one week before the next row.

Another way to fix missing data is by using the fill handle like you used in an earlier screen. Previously, you used it to fill an entire column based on a formula you entered into one cell. In this case, you can use the values of surrounding cells to indicate a pattern and use it to fill in a missing value. In order for autofill to recognize the weekly pattern in the **Date** column, **you need to select at least two cells directly above a missing value** before it will work correctly. Here is an example of it in action:

	A	B	C	D	E	
189	2	3/12/2010	57.56	2.667	NA	NA
190	2	3/19/2010	54.52	2.72	NA	NA
191	2	3/26/2010	51.26	2.732	NA	NA
192	2	4/2/2010	63.27	2.719	NA	NA
193	2	4/9/2010	65.41	2.77	NA	NA
194	2	4/16/2010	68.07	2.808	NA	NA
195	2		65.11	2.795	NA	NA
196	2	4/30/2010	66.98	2.78	NA	NA
197	2	5/7/2010	71.28	2.835	NA	NA
198	2	5/14/2010	73.31	2.854	NA	NA
199	2	5/21/2010	74.83	2.826	NA	NA
200	2	28-May-10	81.13	2.759	NA	NA
201	2	6/4/2010	81.81	2.705	NA	NA
202	2	6/11/2010	83.4	2.668	NA	NA
203	2	6/18/2010	85.81	2.637	NA	NA

Now let's practice!

3.6.1 Instructions

1. Return to the workbook **features_level_1.xlsx**.
2. For all rows in the **Store** column, fill in the appropriate store number for **blank cells only** and *take note of the count of the cells that you had to fix* in order to answer a question below.
3. Proceed to column **B**, the **Date** column, and fill in the appropriate date for **blank cells only**. *Take note of the new dates you enter* in order to answer the question below.
4. Save your workbook.

3.7 Updating Missing Data Continued

You continue reviewing your worksheet, and now you look at column **C**, **Temperature**. You'll notice a few missing values in this column. By this time, you're pretty certain that the rows are ordered by date, and that will help you fix these missing values. Before we proceed, do you have any thoughts on how you might fill in a missing temperature when it's surrounded by other values? We can approach this type of problem in a variety of ways.

For example, you could look at the weather report for that particular date and store location, and that would probably be a reasonable estimate, although it may take you some time to do that, especially if you have a lot of missing data. Another approach to this situation is to calculate a number in the middle of the surrounding numbers. You'll learn about measures of central tendency in a future course, but for now, you'll use the average of the surrounding numbers.

The average is also called the *mean*, and you calculate it by taking the sum of all the numbers and dividing by the count of all the numbers. So, for the set of numbers **3, 1, 4, 1, 5, 9, 2, 6, 5**, you'd add up all the numbers and get **36**. Now you'd divide by the count of **9** to get an average (or mean) of **4**.

Now the question is *How many temperature values do you average?* That's a good question, and, as with many decisions you'll make as a data analyst, you'll need to apply some professional judgment. Taking the average of the two nearest temperatures seems reasonable, but it may be influenced (also called *biased*) by a temperature that is unseasonably high or low, so perhaps two is too few. Taking too many temperatures may be overdoing it, and it may also introduce unwanted bias from temperatures too far away. For the purposes of this exercise, taking five temperature values above and below the missing value seems reasonable. Using the **=AVERAGE** function to fill in the missing value for **C133** would look like this:

<div> <div>AVERAGE</div> <div>✕ ✓ fx</div> <div>=AVERAGE(C128:C132, C134:C138)</div> </div>					
	A	B	C	D	E
121	1	5/18/2012	70.33	3.63	221.742674
122	1	5/25/2012	77.22	3.561	221.744944
123	1	6/1/2012	77.95	3.501	221.7472139
124	1	6/8/2012	78.3	3.452	221.7494839
125	1	6/15/2012	79.35	3.393	221.7626421
126	1	6/22/2012	78.39	3.346	221.8030211
127	1	6/29/2012	84.88	3.286	221.8434
128	1	7/6/2012	81.57	3.227	221.8837789
129	1	7/13/2012	77.12	3.256	221.9241579
130	1	7/20/2012	80.42	3.311	221.9327267
131	1	7/27/2012	82.66	3.407	221.9412954
132	1	8/3/2012	86.11	3.417	221.9498642
133	1	8/10/2012	C134:C138)	3.494	221.9584329
134	1	8/17/2012	84.85	3.571	222.0384109
135	1	8/24/2012	77.66	3.62	222.1719457
136	1	8/31/2012	80.49	\$3.64	222.3054805
137	1	9/7/2012	83.96	3.73	222.4390153
138	1	9/14/2012	74.97	3.717	222.5820193
139	1	9/21/2012	69.87	3.721	222.7818386
140	1	9/28/2012	76.08	3.666	222.9816579
141	1	10/5/2012	68.55	3.617	223.1814772
142	1	10/12/2012	62.99	3.601	223.3812965
143	1	10/19/2012	67.97	3.594	223.4257233
144	1	10/26/2012	69.16	3.506	223.4442513

Upon further examination, you'd discover that the columns `CPI` and `Unemployment` both have several missing values that you would need to fix. You'd also note that several cells will have "NA," which is **not** a missing value, so you would leave those as-is. Your focus is on the blank cells only! The `IsHoliday` column has no blank cells, so you can also leave it as-is.

Now let's practice!

3.7.1 Instructions

1. Return to the workbook `features_level 1.xlsx`.
2. Take the average temperature for missing cells. Start in cell `C133`, type in the formula `=AVERAGE(C134:C138,C128:C132)`, and hit `Enter`. Do this for all remaining blank temperatures.
3. Look at the remaining columns and continue to fill in the missing values using the same approach you used for the temperature column. You'll discover that the columns `CPI` and `Unemployment` both have several missing values that you'll need to fix. Fix the missing values in those columns now.
4. Save your workbook.

3.8 Other Considerations

What do you do when you have **lots** of missing data?

If you have one or more records that are missing most of the data, or several records near each other that are mostly blank, you need to decide which is better, imputing the missing data or deleting the blank records altogether. Going back to the analogy of the dog photograph, if the only thing remaining in the photograph is a portion of the tail, do you really have enough information to recreate an entire photo, or should you abandon that effort and try to find other sources?

Deleting records that are mostly blank is common practice, especially if you have enough remaining data to support your analysis. Once again, this is a matter of professional judgment that you'll need to exercise as you consider your dataset and how to analyze it.

4 Cleaning Data Part 2

4.1 Introduction

In the previous lesson, you learned methods for cleaning your data. In this lesson, you'll pick up where you left off. In particular, you'll focus on these processes:

- Standardizing the case for text
- Standardizing formats for dates and numbers
- Working with outliers
- Visualizing to verify clean data

As a reminder, the reason why you take the time to clean your data is so that the final analysis will be trustworthy. While it may take quite a bit of time to clean the data, imagine the time it would take you to go back and clean the data if it wasn't done correctly the first time. Imagine the mistakes in the analysis, and the poor decisions made based on flawed information. Imagine how that may reflect on your credibility as a data professional.

Now, imagine if you take the time to do it correctly the first time, and how satisfied you'll feel knowing that the data is trustworthy. Imagine how great you'll feel when you know the data led to a successful analysis and good decisions! That's a much better scenario, and *that's* what you need to focus on!

One of the things you'll learn about cleaning data is that consistency is a good thing. As you saw in the last lesson, you worked with a column in your dataset that had multiple ways of expressing *false*. One of those inconsistencies was between the uppercase **FALSE** and lowercase **false**.

7.682	FALSE	
7.682	FALSE	
7.682	FALSE	
7.682	FALSE	
7.682	false	
7.682	false	
7.962	false	
7.962	false	

In most situations, there isn't a significant difference between those two instances of the word. However, computers are pickier than humans — and a lot less forgiving. As a data professional, you'll benefit from knowing how computers think.

Internally, computers use a series of zeros (0's) and ones (1's) to follow instructions — it's their internal language. We call these zeros and ones *machine code*. All computer data gets converted to machine code. The uppercase letter **F** uses a different machine code than the lowercase letter **f**. The integer number **5** uses a different machine code than the real number **5.00**, even though both numbers are mathematically equal. Computers use machine code to follow explicit instructions. Standardizing your data ensures that it will be consistent so that the machine code is also consistent when the computer analyzes it.

In the following screens, you'll see how to standardize your data so that it is consistent.

4.2 Standardizing Text Data

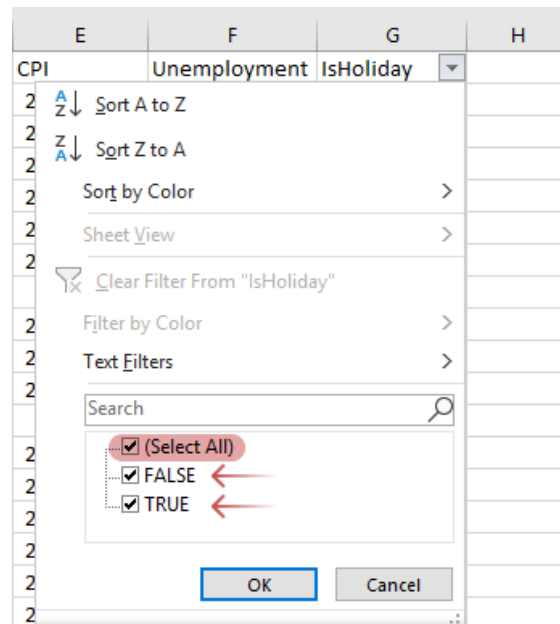
By this point, you've manipulated your data using a number of different methods. It's time to go back to your dataset and continue to clean it using some of methods outlined in this lesson.

To start, you'll open the last dataset from the previous lesson called `features_level_1.xlsx`.

Take a few moments to review the current state of the dataset. You'll quickly recall that column **G**, `IsHoliday`, contains text that is mixed format.

In particular, most of the values on this column are uppercase "FALSE," yet some of the values are lowercase "false." As we pointed out in the previous lesson, even if you use a filter on this column, the filter won't distinguish the uppercase from the lowercase.

If you use **Sort & Filter, Filter** on the `IsHoliday` column, you'll see that Excel has made things convenient for you by displaying "FALSE" only, and it doesn't include the lowercase version of false. Be careful, as this can be deceiving!



If you scroll down the **IsHoliday** column, you'll see a handful of values where the word "false" is lowercase. This is something that you'd need to update to make them all consistent.

Another thing you should do with your datasets is to standardize the header names. The naming conventions you use should make sense, but they should also be consistent in *how* they are named. If you look at the header names in your dataset, you'll see the following. Everything looks pretty consistent (the names all start with an uppercase letter). However, the column names **Fuel Price** and **IsHoliday** are inconsistent because both use two words, yet one uses an underscore symbol (**_**), and the other does not.

C	D	E	F	G
Temperature	Fuel Price	CPI	Unemployment	IsHoliday
42.31	2.572	211.0963582	8.106	FALSE
38.51	2.548	211.2421698	8.106	TRUE
39.93	2.514	211.2891429	8.106	FALSE
46.63	2.561	211.3196429	8.106	FALSE

If you change things so *both* columns use an underscore between the two words, you'd see the following:

C	D	E	F	G
Temperature	Fuel Price	CPI	Unemployment	Is Holiday
42.31	2.572	211.0963582	8.106	FALSE
38.51	2.548	211.2421698	8.106	TRUE
39.93	2.514	211.2891429	8.106	FALSE
46.63	2.561	211.3196429	8.106	FALSE

Follow the instructions below to implement the changes above.

4.2.1 Instructions

We've provided a solutions file (`features_level_2_solutions.xlsx`) in case you get stuck or would like to verify your answers while working on the exercises in this lesson.

1. To start, open the last dataset from the previous lesson called `features_level_1.xlsx`.
2. Use **Data, Sort & Filter, Filter** on the `IsHoliday` column to confirm that "false" is not one of the filtering options.
3. Change it so that *all* instances of "false" become the uppercase "FALSE." As you do this, double-check that all of the "TRUE" values are uppercase and convert them if necessary. Take note of the count of values that you update for both "TRUE" and "FALSE"!
4. Change the column `IsHoliday` to `Is_Holiday`.
5. Save it as `features_level_2.xlsx`.

4.3 Formatting Dates

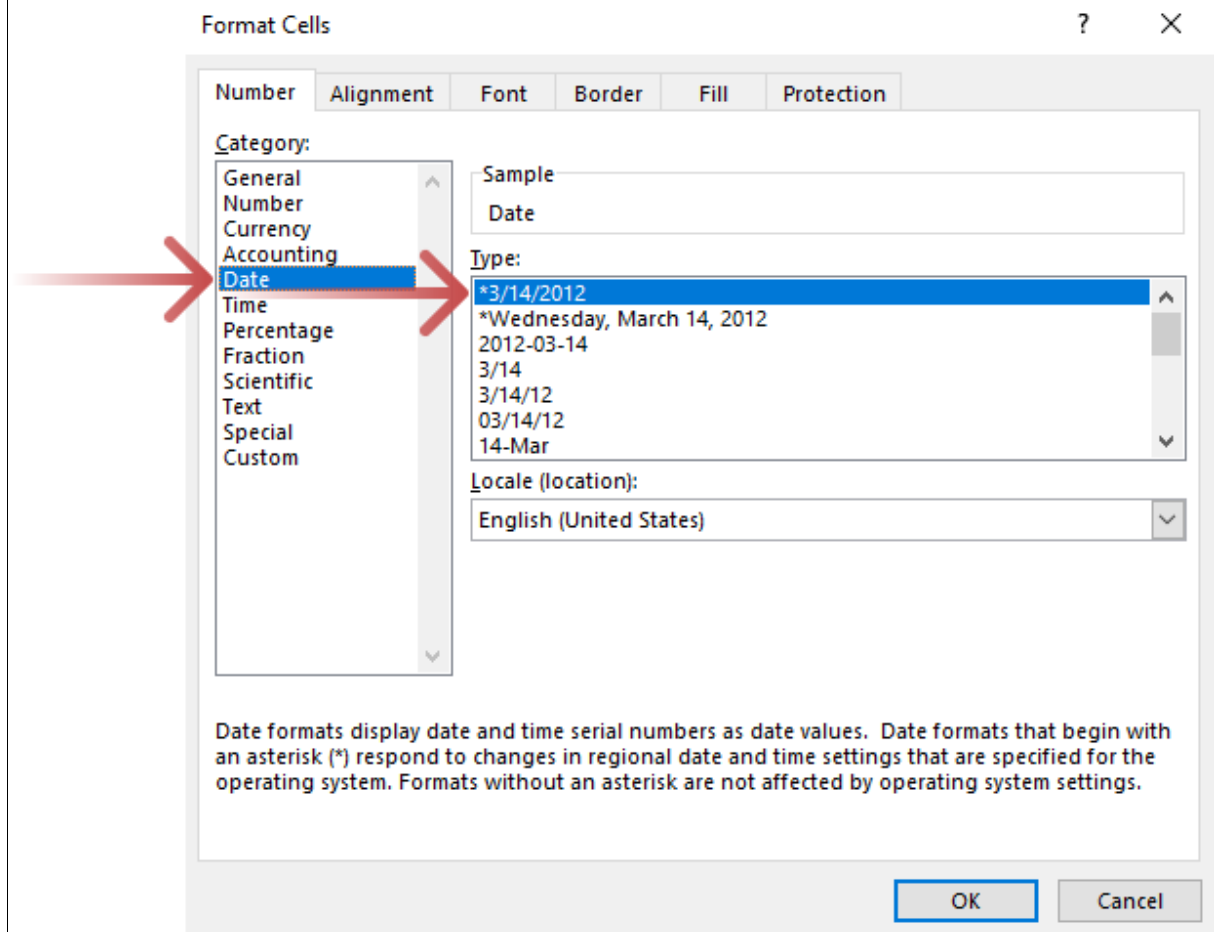
Dates are very common in datasets, so knowing how to work with them is very important. In most cases, you'll need to make sure the dates are in some type of order, either ascending or descending. In addition, dates can be expressed in various formats, so you'll need to make sure that the formatting is consistent throughout the dataset.

If you look at your most recent workbook `features_level_2.xlsx` and focus on column `B`, `Date`, you'll see several inconsistent date formats:

	A	B	C	D	E	F	G
1	Store	Date	Temperature	Fuel_Price	CPI	Unemployment	IsHoliday
2	1	2/5/2010	42.31	2.572	211.0963582	8.106	FALSE
3	1	12-Feb-10	38.51	2.548	211.2421698	8.106	TRUE
4	1	2/19/2010	39.93	2.514	211.2891429	8.106	FALSE
5	1	2/26/2010	46.63	2.561	211.3196429	8.106	FALSE
6	1	Friday, March 5, 2010	46.5	2.625	211.3501429	8.106	FALSE
7	1	3/12/2010	57.79	\$2.67	211.3806429	8.106	FALSE
8	1	3/19/2010	54.58	2.72	211.215635	8.106	FALSE
9	1	3/26/2010	51.45	2.732	211.0180424	8.106	FALSE
10	1	4/2/2010	62.27	2.719	210.8204499	7.808	FALSE

Your main task is to convert the dates so that they all have consistent formatting. You first need to decide on a date format. Since most of the dates in column `B` use the format `month/day/year`, like `3/26/2010`, you decide to keep that format. As with most things in Excel, there are often multiple ways to accomplish the same thing. For this task,

you would revisit an old friend by selecting the **Date** column and using **Format Cells** (CTRL+1). You would then select the first option under **Type** for the **Date** category as shown below:



If you did that, you'd see that cell **B6** and several other dates in column **B** didn't convert as you had hoped. The date there still looks like this: **Friday, March 5, 2010**. Excel can't always convert dates because they may be in an unknown format, so Excel doesn't know what to do. In this case, the date **Friday, March 5, 2010** really isn't a date; it's text that looks like a date. In these situations, you may need to do a manual conversion to match the other dates.

You can type in the dates directly, which is error-prone because it's manual entry. Or you can use the preferred method, which is to use the **Fill Handle** as you did previously with missing values. If you did that, you'd see the following:

	A	B	C	D	E	F	G
1	Store	Date	Temperature	Fuel_Price	CPI	Unemployment	IsHoliday
2	1	2/5/2010	42.31	2.572	211.0963582	8.106	FALSE
3	1	2/12/2010	38.51	2.548	211.2421698	8.106	TRUE
4	1	2/19/2010	39.93	2.514	211.2891429	8.106	FALSE
5	1	2/26/2010	46.63	2.561	211.3196429	8.106	FALSE
6	1	3/5/2010	46.5	2.625	211.3501429	8.106	FALSE
7	1	3/12/2010	57.79	\$2.67	211.3806429	8.106	FALSE
8	1	3/19/2010	54.58	2.72	211.215635	8.106	FALSE
9	1	3/26/2010	51.45	2.732	211.0180424	8.106	FALSE
10	1	4/2/2010	62.27	2.719	210.8204499	7.808	FALSE
11	1	4/9/2010	65.86	2.77	210.6228574	7.808	FALSE
12	1	4/16/2010	66.32	2.808	210.4887	7.808	FALSE
13	1	4/23/2010	64.84	2.795	210.4391228	7.808	FALSE
14	1	4/30/2010	67.41	\$2.78	210.3895456	7.808	FALSE
15	1	5/7/2010	72.55	2.835	210.3399684	7.808	FALSE
16	1	5/14/2010	74.78	2.854	210.3374261	7.808	FALSE
17	1	5/21/2010	76.44	2.826	210.6170934	7.808	FALSE
18	1	5/28/2010	80.44	2.759	210.8967606	7.808	FALSE
19	1	6/4/2010	80.69	2.705	211.1764278	7.808	FALSE
20	1	6/11/2010	80.43	2.668	211.4560951	7.808	FALSE

Excellent! Now your dates are clean and consistent! On the next screen, you'll learn how to format numbers.

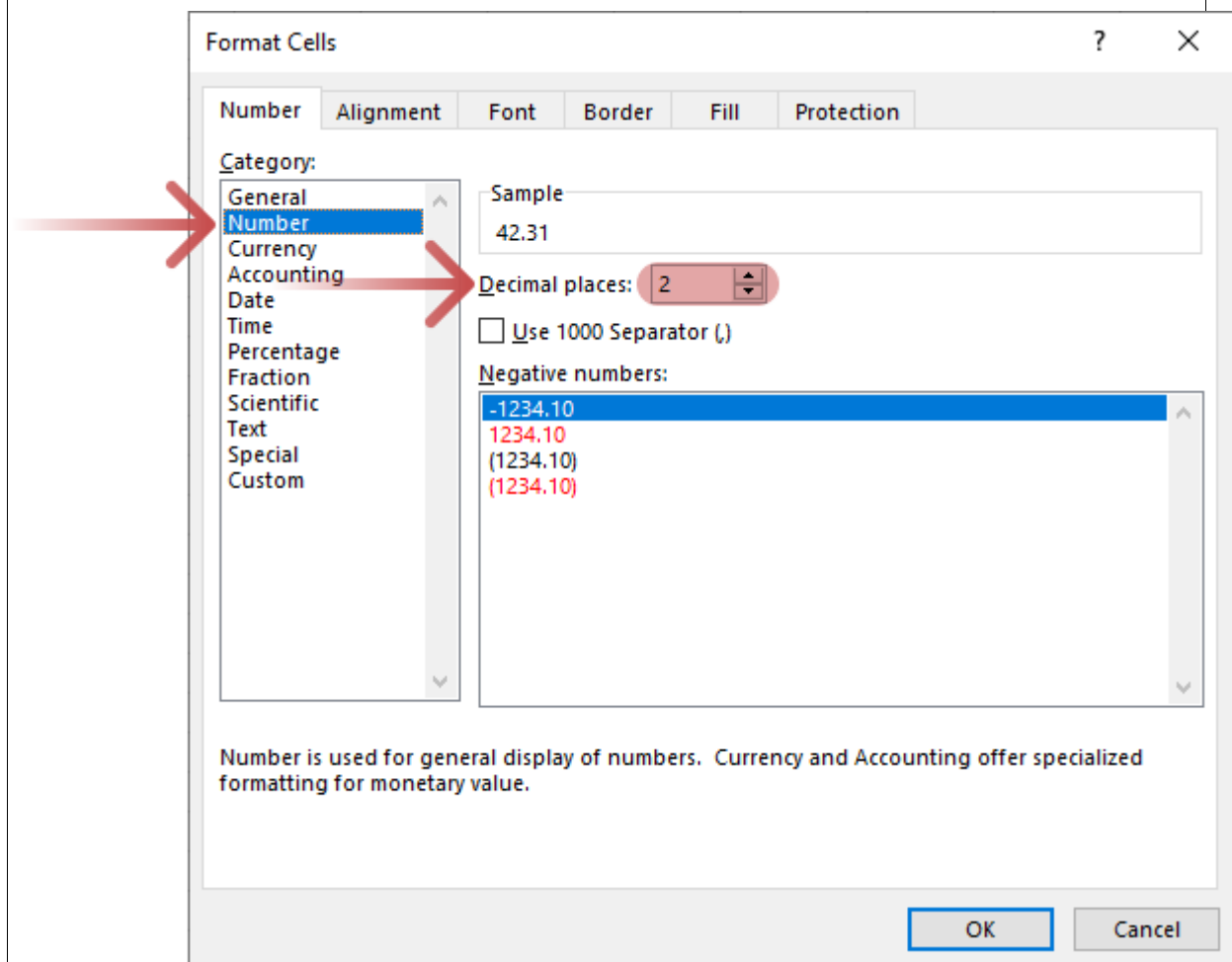
4.3.1 Instructions

1. Select all of the data in column **B**, and then click **CTRL+1** to bring up the **Format Cells** window.
2. Once there, select **Category: Date** and then choose the first format option under **Type**, as shown above.
3. Before you click **OK**, take a moment to scroll through the various date types offered in the window. As you can see, there are many ways to express a date.
4. When you're done, click **OK**.
5. Using the **Fill Handle**, convert the remaining (text) dates so that they all have consistent formatting.
6. Save your workbook.

4.4 Formatting Numbers

The majority of datasets you'll encounter will involve numbers. As with dates and text, you'll want to be sure that your numbers are consistent within their respective columns so you can analyze them properly. Going back to your dataset, please open **features_level_2.xlsx** if necessary. Your focus this time will start with

column **C**, **Temperature**. Most of the numbers have two decimal places, and your analysis team has indicated that is sufficient for their purposes. So, you'll need to convert all the numbers to conform to two decimal places using **Format Cells**:



Moving on to column **D**, **Fuel Price**, take a few moments to survey what you have. You'll quickly notice that in addition to an inconsistent number of decimal places, you've also got several values that use dollar signs "\$".

D
Fuel_Price
2.572
2.548
2.514
2.561
2.625
\$2.67
2.72
2.732
2.719
2.77
2.808
2.795
\$2.78
2.835
2.854

To fix both of these issues at the same time, you'll want to use **Format Cells** again as you saw above.

Finally, you have two more columns that contain numbers, **CPI** and **Unemployment**, that you also need to fix to have a consistent format. Again, this is where **Format Cells** will come in handy!

4.4.1 Instructions

1. Select column **C**, **Temperature**, then using **Format Cells**, choose **Category: Number**, then set your decimal places to 2. Click **OK**.
2. Format column **D**, **Fuel_Price**, in the same way; **Category: Number** with 2 decimal places.
3. Format column **E**, **CPI**, to have five decimal places
4. Format column **F**, **Unemployment**, to have three decimal places.
5. Save your workbook.

4.5 Working with Outliers

When working with numbers, there are times when the numbers are accurate but will still cause problems for you. These problems occur when the numbers you're working with contain extreme values that are outside the normal range. Extreme values are often called **outliers** because they exist outside the main group of values. We'll use an example to illustrate this point.

Suppose that you're a data analyst who works for a car dealership franchise that owns seven dealerships for both new and used cars. You're part of a team of data professionals who is working on a project that will ultimately identify and analyze the sticker prices of the cars at each of the dealerships. You've just been given a dataset for the used car lot at one of the dealerships, as follows:

make	model	year	price
Ford	Mustang	1995	8,500
Chevrolet	Silverado	2003	12,300
Dodge	Ram	2001	11,000
Toyota	Camry	2000	7,400
Honda	Accord	1999	7,900
Honda	Civic	2003	8,250

make	model	year	price
Ford	Taurus	2002	7,000
Chevrolet	Camaro	1998	9,750
Jeep	Wrangler	2002	8,900
Porsche	911	2020	97,000

Take a few moments and look at the list. As you can see, most of the prices aren't too far from each other. The exception is the price of the used Porsche, which is far greater than the price of the other vehicles listed. As a data analyst responsible for cleaning the data, you know that one of the things that will be analyzed is the "average" price of the cars (you also know this as the mean). Even before you do any math, you already know that the price of the Porsche will strongly influence the average price of all the vehicles. If you simply take the average price of all the vehicles, you'll get an answer that isn't really representative of the majority of vehicles.

In this case, the average price is \$17,800. Does this price really represent the average vehicle price? If the dealership advertised this as the average price of the cars on their lot, what would customers think when they visited the lot?

Imagine if a customer has a maximum budget of \$9,000. There is a good chance they wouldn't even visit the lot, even though there are several vehicles that fall within their budget! In that sense, the dealership is hurting itself by using a number that misrepresents the values across the vehicle population.

As you can see, the outlier has distorted the price. This distortion is also called *skewing*, and you need to know how to handle it before it causes problems with the analysis. Before we go any further, you need to know that there are statistical methods for identifying outliers. Those methods are beyond the scope of this course — you'll get a closer look at those methods in future courses. For now, you're assuming that the price of the Porsche is an outlier, and you need to determine the best way to deal with it.

4.5.1 Methods for handling outliers

There are a number of different ways to work with outliers so that your data is better prepared for analysis. Some of those methods are rather sophisticated, and some of them are very straightforward. Perhaps the easiest (and most obvious) thing to do is to remove the outliers.

Removing outliers is not something to take lightly. Outliers may be unusual, but they are *still part of the data*. As with all of the decisions you make as a data professional, you need to know your data and *how it will be used*. For the purposes of advertising an average price, the dealership probably shouldn't have included the Porsche in its inventory in the first place because its value would clearly distort the average value of the total population. So, removing it to give customers better insight into the average vehicle price is reasonable. However, if the dealership is comparing its average inventory price to another dealership, then keeping the outlier is recommended. **When it comes to representing data, context is everything!**

Method 1: You could remove the row for the Porsche and take the average price of all the remaining vehicles. Doing so would give you an average price of \$9,000. Does that price seem more representative of the vehicle prices? Yes, it certainly does!

Method 2: There is another quick method to minimize the effect of outliers. You can calculate the *median* price of the vehicles, which is the mid-point of a set of sorted numbers. In the case of your vehicle prices, you have the following:

7,000	7,400	7,900	8,250	8,500	8,900	9,750	11,000	12,300	97,000
-------	-------	-------	-------	-------	-------	-------	--------	--------	--------

Because there is an even number of values, you take the average of the middle two values: 8,500 and 8,900, giving you \$8,700 as your representative (median) price. You can also use the Excel function `MEDIAN()` to accomplish this and get the same result. You can see that including your outlier in the median calculation had virtually no effect!

Telling customers that your typical price for used vehicles is around 8 to 9 thousand dollars is much more representative than telling them it is \$17,800, and you'll probably get more customers to visit your lot that way!

Instructions

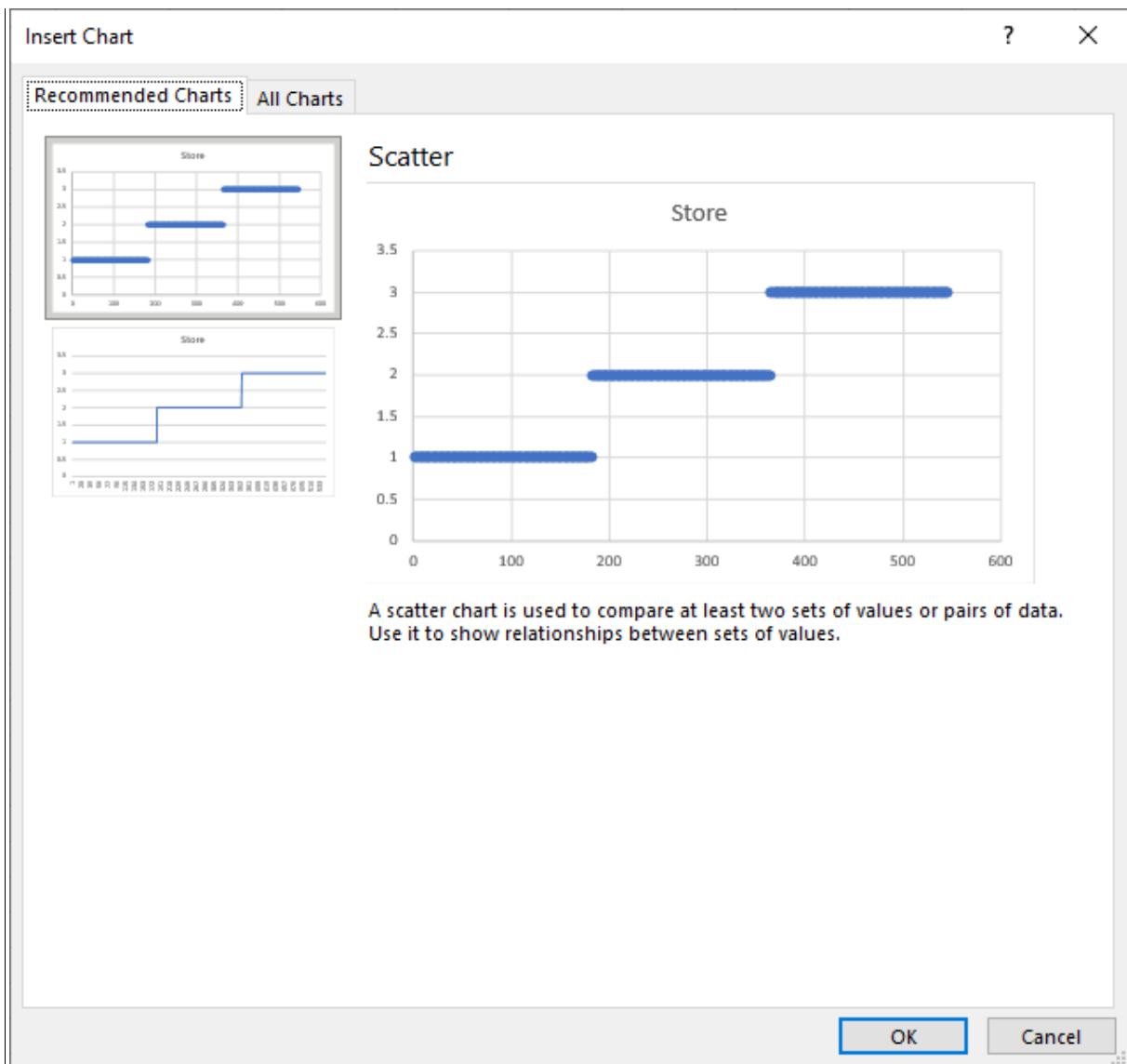
1. Use the plus sign at the bottom of your workbook to add a blank worksheet.
2. Starting in cell **A1**, fill column **A** with the price of the 10 vehicles shown above.
3. In cell **C1**, find the average price of all the vehicles using the **AVERAGE** function.
4. Remove the last row for the Porsche, and in cell **E1**, find the average price of all the remaining vehicles.
5. Reinsert the row for the Porsche you just removed, and in cell **G1**, use the **MEDIAN ()** function to get the median value of all vehicle prices.

4.6 Visualizing Your Data

Up until now, you've verified your data by looking at it line by line. You've also used filtering to verify that your data is what you expect it to be. Those are good things to do, but it's also a good idea to create some quick visualizations to make sure you didn't miss anything. On this screen, you'll see how to apply some visualizations to your data to verify that it's as clean as you intended it to be. In Excel, visualizations are **charts**. Visualizing your data is also a large part of the analysis process, and you'll spend quite a bit of time on that in the following course.

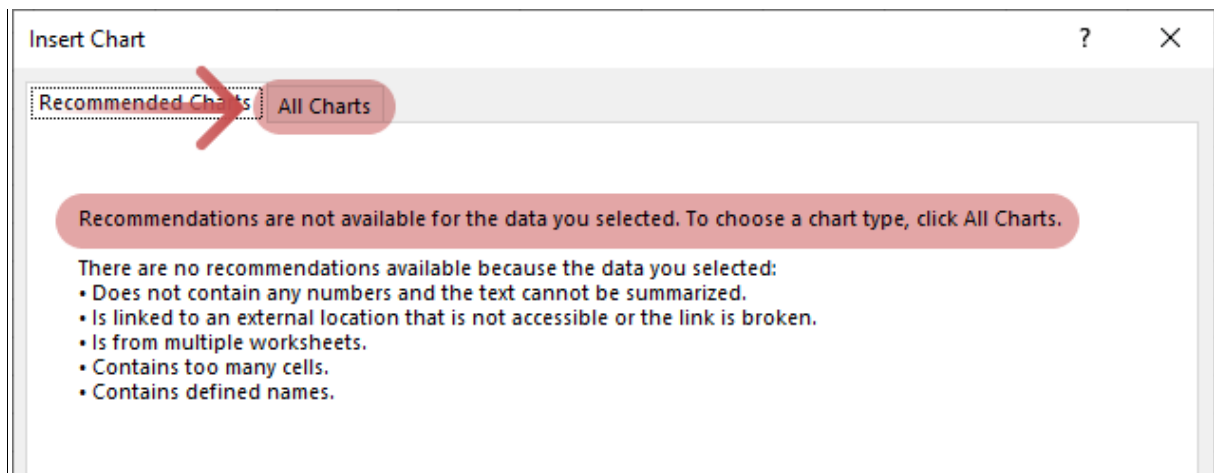
You'll continue working with **features_level_2.xlsx**, and you'll want to check your data. If you were to apply filtering and look for unusual values, you'd see that all looks good. But it's also good to apply visualizations column by column to make sure you don't miss anything.

After selecting the data, if you were to use **Insert, Recommended Charts**, you would get a preview of the chart. For example, the chart for column **A** would look something like this:

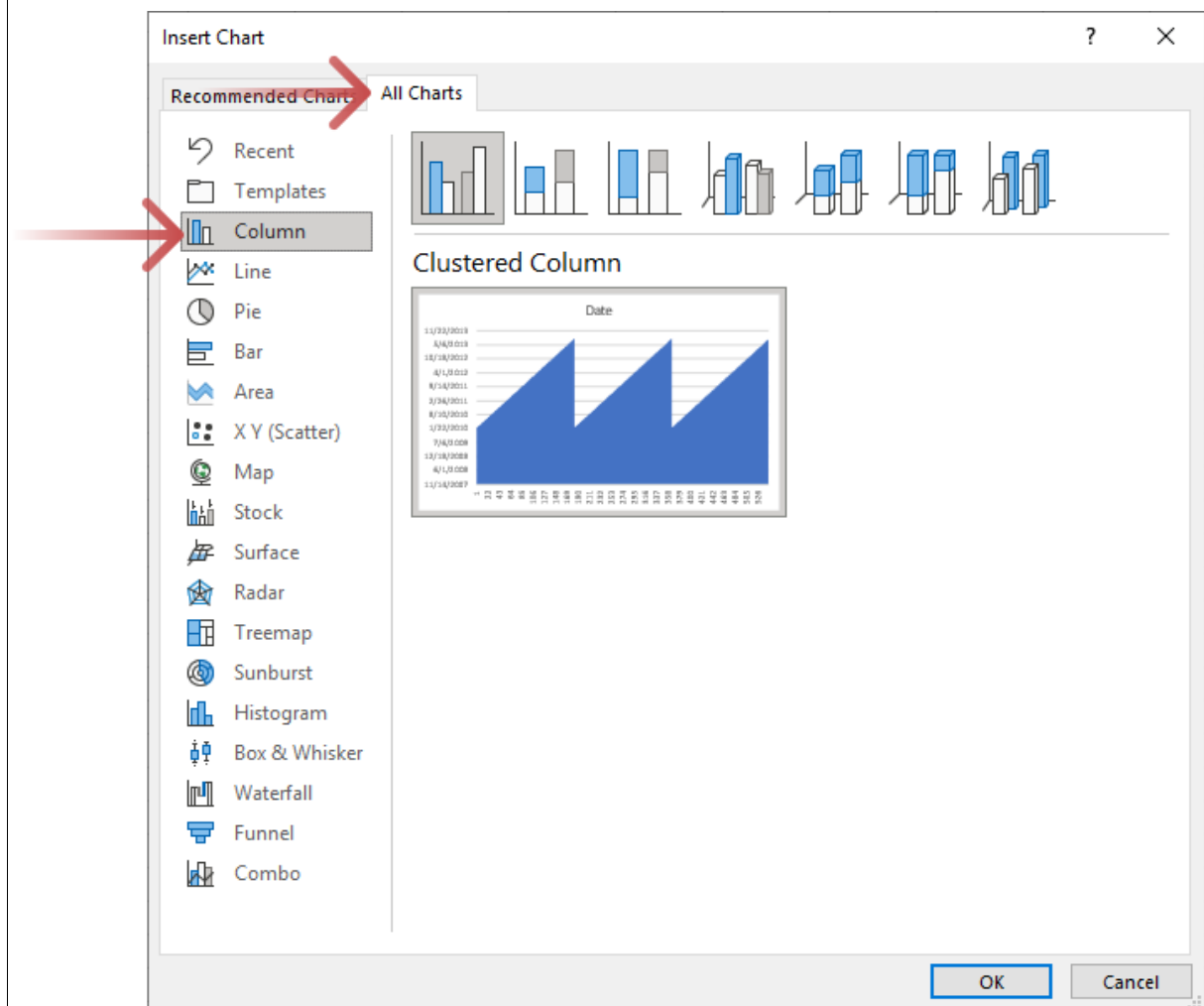


Excel does its best to recommend the most appropriate chart for the data you select. It may take some getting used to, but over time, you'll be able to quickly read and assess what the charts are telling you. In the previous case, you're looking at multiple instances of three stores, which is what you expect. Time to move on to the next column, **B**.

You'll notice for column **B**, **Date**, that there is no good recommended chart. Sometimes the data doesn't lend itself well to being visualized (we've provided those reasons in the dialog box shown below), and this is one of those situations.

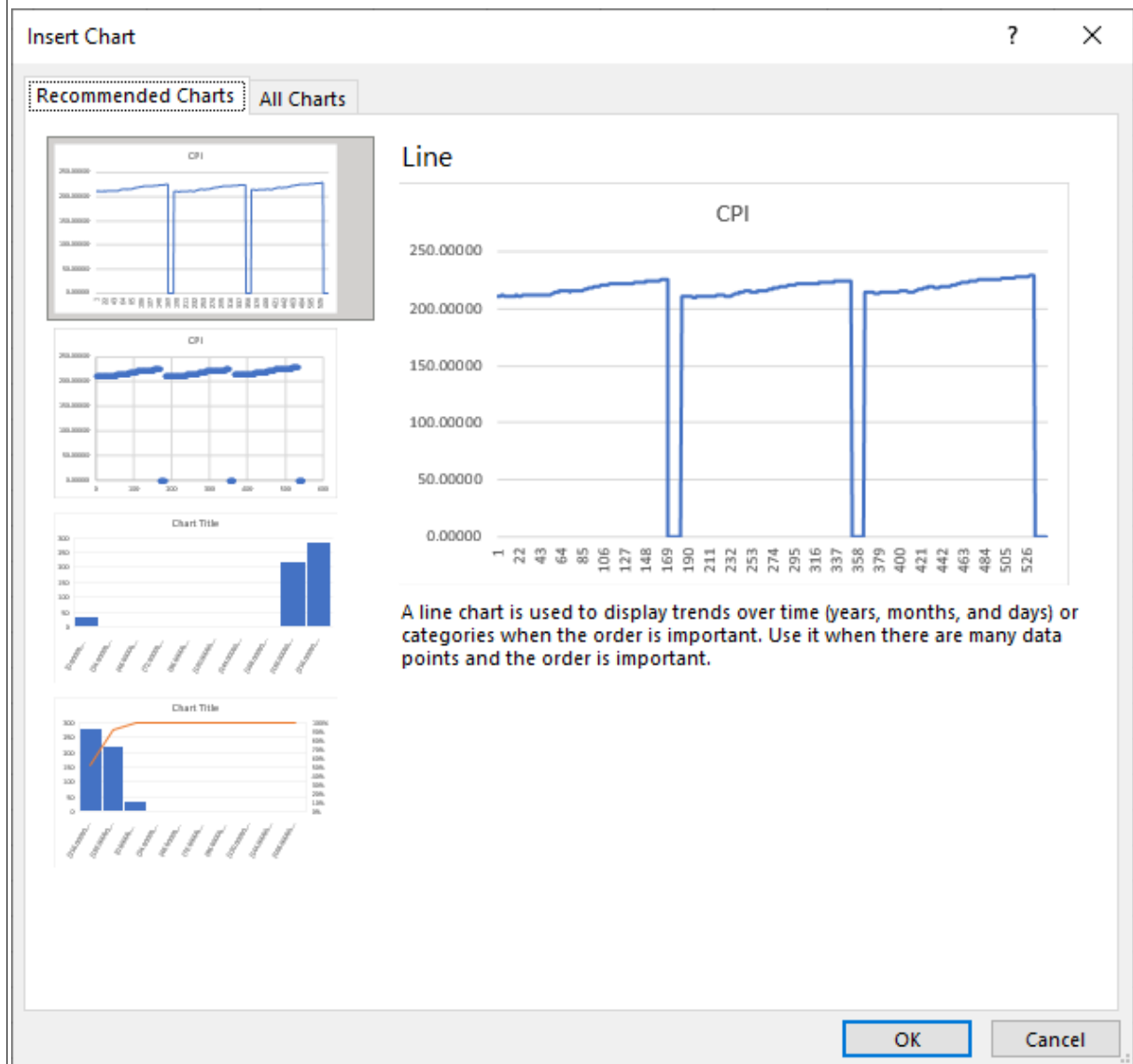


You could choose the **All Charts** tab and try to insert a chart manually. The **Clustered Column** chart (the first option under **Column**) seems to give you something that makes sense. It's the dates arranged by store number. It looks okay, so time to move on.



As you move through the next several columns, you'll start to notice a pattern in the charts. In particular, you'll see things grouped into three segments (just like you saw with

the dates), and that's because your data has nicely grouped things by each of the three stores. For example, the column **CPI** shows three distinct groupings:



Keep going through the columns to verify all is well. You'll continue to see this pattern until you get to the last column, **Is Holiday**. A preview chart would show you the following:

This is something you would need to fix before you analyze your data!

4.6.1 Instructions

1. Open `features_level_2.xlsx` to the `features_start` tab. Select all columns, and apply **Data, Sort & Filter, Filter**, then do a quick check to see if the selections you're given are those that you expect.
2. Select column `A`, then select **Insert, Recommended Charts**. Don't insert the chart, the preview is sufficient.
3. Continue to look at the recommended charts for each of the columns, and see if they look good to you.
4. Fix the formatting in the `Is_Holiday` column for each of the "TRUE" and "FALSE" values to match the rest.

- Use Copy & Paste to replace incorrectly formatted "TRUE" and "FALSE" values with correctly formatted versions.
- Keep track of the number of values you fix to answer a question below.

5. Rerun your chart test and verify that you only have two entries for "TRUE" and "FALSE" in the `Is_Holiday` column.
6. Save your workbook.

5 Consolidating Data

5.1 Introduction

Consolidating data, sometimes called *combining* or *merging* data, is one of the final steps in the data cleaning process. This is when you take the data and organize it into as few tables as you can to support analysis. Having multiple tables or datasets isn't necessarily a problem because there are methods to join and extract the data during an analysis. But, having your data in one table has its advantages. The biggest advantage of having your data in one table is that you can go to one location to get the information that you need without having to consolidate the data every time you want to review it.

Having your data in one location simplifies analysis. The only caveat to having your data in one table is that you need to make sure that you keep it updated from the individual data sources that feed into it. Depending on how you approach this, you may be able to keep it updated automatically, or you may need to do it manually. Either way, making sure the data is current is part of a data analyst's job.

Sometimes, the analysis team consolidates the data, but often, the data cleaning team does it as a final step before submitting the data for analysis (a courtesy that the analysis team will appreciate!).

Think back to the scenario from the previous lesson. Suppose that you work for a company that owns seven car dealerships. Now suppose that every car dealership maintains its own financial records that includes information like `weekly gross sales`, `weekly costs`, `weekly profits`, `inventory levels new`, and `inventory levels used`. A single franchise business owner owns your company. The owner wants to know not only how each dealership is performing but also how *all* dealerships are performing *together*. In order to analyze all of the dealerships together, it's optimal to consolidate the data so that you have one single source to analyze. That's what this lesson is all about. You're about to learn methods for consolidating data including the following:

- Manipulating headers
- Using Excel's Consolidate tool
- Using copy and paste as needed
- Using `VLOOKUP`
- Using `INDEX/MATCH`

You'll see some of these same concepts in other courses, such as databases and programming. So, learning about them now will prepare you for other areas of study within data analytics.

5.2 Manipulating Headers

By now, you know how to use headers. Headers are column names (or field names) that are a helpful way of telling the user something about the data in that column. If you see a column named `date` or `sales tax`, you immediately know the type of data in those columns. However, if you see a column named `xyz` or `1_2A`, those header names don't tell you much about the data.

As you learned in an earlier lesson, header names should be meaningful, but you should also strive to keep them brief for the sake of space. This is a challenge you will face frequently as you work with datasets.

Also in your career as a data analyst, there is a good chance you'll encounter datasets that have no headers whatsoever! Unlabeled data is, unfortunately, very common.

Suppose that you work for a company that collects sales data weekly, but it's only the raw data with no headers. It's wise to insert headers into the dataset as soon as it's practical.

That way, if another team uses the dataset, they will know the type of data they are working with. Data that is unlabeled is data that is just waiting for mistakes. So, always label your data with headers!

Now you're going to visit a dataset that you worked with toward the beginning of this course. Find and open the workbook named `sales_data_set_start.xlsx`. When you open it, you'll see something like this:

	A	B	C
1	Store	Date	Weakly Sales
2	1	2/5/2010	1459601.17
3	1	2/12/2010	1621031.7
4	1	2/19/2010	1604775.58
5	1	2/26/2010	1527014.04
6	1	3/5/2010	1508237.76
7	1	3/12/2010	1391013.96
8	1	3/19/2010	1494251.5
9	1	3/26/2010	1517428.87
10	1	4/2/2010	1493525.93
11	1	4/9/2010	1559889
12	1	4/16/2010	1630607
13	1	4/23/2010	1455119.97
14	1		\$1,459,409.10
15	1	5/7/2010	\$1,677,472.78

Take a few minutes to familiarize yourself with the data. You should see several things that we need to fix. To start, the `Date` column is missing some dates, as you saw with your earlier dataset. Also, the data in column `C` has inconsistent formatting. Finally, the header name for column `C`, `Weakly Sales` is misspelled, and you also need an underscore between the two words to match your other dataset.

Time to get to work!

5.2.1 Instructions

1. Using what you learned previously, fill in the missing dates in the workbook `sales_data_set_start.xlsx`. Take note of the number of dates you fill in.
2. Format column `C` to a standard number with two decimal places, `no` dollar sign, and `no` comma to separate the number by 000s using **Format Cells**.
3. Fix the header name for column `C` by changing `Weakly Sales` to `Weekly Sales`.
4. When you're finished cleaning the dataset, save your workbook as `sales_data_set_clean.xlsx`.

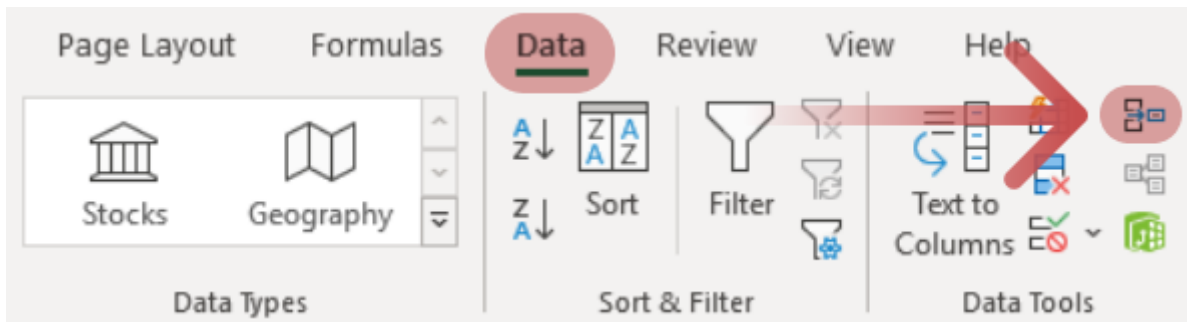
5.3 The Consolidate Tool

On the previous screen, when you counted the number of rows in the workbook `sales_data_set_clean.xlsx`, did you notice that it contained the same number of rows as the earlier dataset you worked with, `features_level_2.xlsx`? As it turns out, these two datasets are related to each other. In fact, you'll notice that the first two columns, `Store` and `Date`, are identical in both workbooks. That won't always be the case, but in this situation, that has made things convenient for our purposes.

For you to combine the data, what you really want to do is take the column `Weekly Sales` from `sales_data_set_clean.xlsx` and combine it with the larger worksheet in `features_level_2.xlsx` so that your data is in a single table in one worksheet.

There are multiple ways to do this, but on this screen, you'll learn how to consolidate the data into a single table using Excel's **Consolidate** tool.

The **Consolidate** tool is very helpful when you want to consolidate data from multiple workbooks or worksheets into a single worksheet. You'll find the **Consolidate** tool icon in the **Data** tab within the **Data Tools** group. Familiarize yourself with where this tool is located because you'll use it soon!



The **Consolidate** tool is very good at taking two or more columns and performing some mathematical operation on them to produce a third column. You'll see a general example of this first, and then you'll do it yourself step-by-step.

Suppose you have a column called **monthly_sales_east** with the following values:

300
200
350
450
150

Now suppose you have another column in a different workbook called **monthly_sales_west** with the following values:

100
550
200
300
250

If you want to see the **combined** sales for both east and west regions in a new column, you could use the **Consolidate** tool and ask for the sum. Your new column would produce the following, which is the sum of both columns row-by-row:

400
750
550
750
400

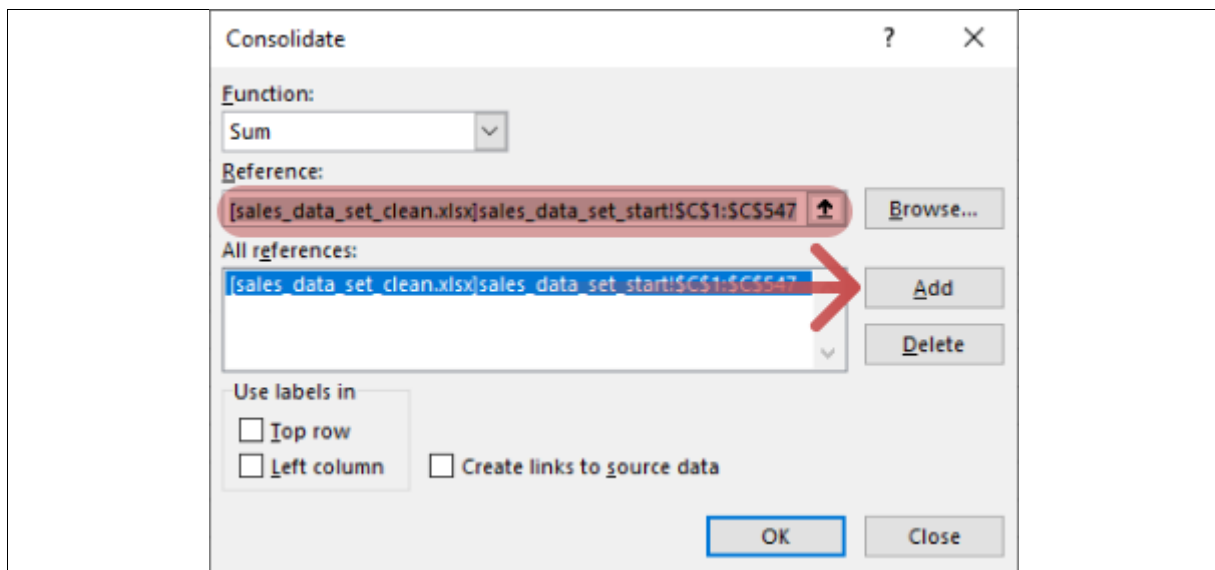
Now you're going to get hands-on experience with the tool and apply it to your datasets.

5.3.1 Instructions

We've provided a solutions file (`features_level_3_solutions.xlsx`) in case you get stuck or would like to verify your answers while working on the exercises in this lesson.

1. Open both workbooks, `sales_data_set_clean.xlsx` and `features_level_2.xlsx`, if they aren't already open.
2. Go to `features_level_2.xlsx`, and select the cell `H1` because column `H` is where your data will go. Once there, click on the **Consolidate** tool icon in the **Data** tab, and you'll see this dialog box:

- With your cursor inside the **Reference** section of **Consolidate**, head to the workbook `sales_data_set_clean.xlsx` by minimizing the `features_level_2.xlsx` workbook and select the range `C1:C547` inside `sales_data_set_clean.xlsx`, then click on the **Add** button in **Consolidate**. You'll then see this:



- You can leave the **Function** menu at the top to **Sum**, but you can use the drop-down arrow to view the other available options.
- Do **not** click **OK** yet! We will continue this exercise on the next screen.
-

5.4 The Consolidate Tool Continued

You've just added your first column to the **Consolidate** tool, and you'll see it listed in the **All references** section. Here's where you'll use a trick to get what you want. Your goal is to combine the **Weekly Sales** data in the main table. You don't need to sum the numbers, you just want the original values. The **Consolidate** tool doesn't offer a choice for value only, so you're going to take the sum of the values with blank values, yielding only the values. Any number plus zero is that number, and that's what you're going to do next.

5.4.1 Instructions

1. With the **Consolidate** tool dialog box still open, select a new range for the **Reference** section.

- This time, select the blank range **D2:D547** in **sales_data_set_clean.xlsx**. Remember that you're doing this to get the sum of column **C** and a blank column **D**, which will result in the values of column **C**.

2. As before, click the **Add** button afterwards, then click **OK**.

- If you go to `features_level_2.xlsx`, you'll see the data in column `H` as desired.
- The header name doesn't get transferred, so give it the name `Weekly Sales`.

3. To make things consistent, remove any filtering from columns `A` through `G` by clicking on `Clear` under `Data, Sort & Filter`, then save your workbook as `features_level_3.xlsx`.

4. Find the sum of the newly consolidated `Weekly Sales` data in column `H`.

Your worksheet should now look like the following:

	A	B	C	D	E	F	G	H	I
1	Store	Date	Temperature	Fuel_Price	CPI	Unemployment	Is_Holiday	Weekly_Sales	
2	1	2/5/2010	42.31	2.57	211.09636	8.106	FALSE	1459601.17	
3	1	2/12/2010	38.51	2.55	211.24217	8.106	TRUE	1621031.70	
4	1	2/19/2010	39.93	2.51	211.28914	8.106	FALSE	1604775.58	
5	1	2/26/2010	46.63	2.56	211.31964	8.106	FALSE	1527014.04	
6	1	3/5/2010	46.50	2.63	211.35014	8.106	FALSE	1508237.76	
7	1	3/12/2010	57.79	2.67	211.38064	8.106	FALSE	1391013.96	
8	1	3/19/2010	54.58	2.72	211.21564	8.106	FALSE	1494251.50	
9	1	3/26/2010	51.45	2.73	211.01804	8.106	FALSE	1517428.87	
10	1	4/2/2010	62.27	2.72	210.82045	7.808	FALSE	1493525.93	

5.5 Copy and Paste

Copy and paste is often overlooked as a useful tool, perhaps because it seems too simple. However, it can be very handy when you're working with data. You can use copy and paste to copy data from one worksheet or workbook and paste it into a single worksheet table.

Perhaps the most important thing you need to remember when you're performing a copy and paste is to focus on the `paste` part. When copying data in Excel, you have only one option, and that is `copy`. However, when pasting data, you'll often see multiple options, and those options can affect your data significantly.

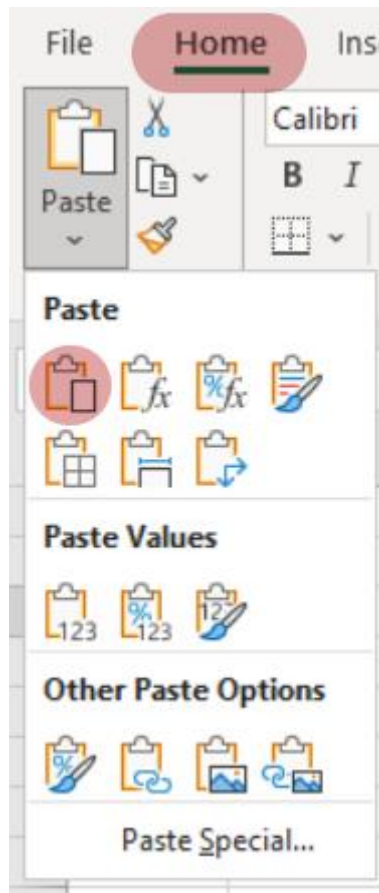
After you copy data in Excel, your `Paste` menu in the `Home` tab will activate. You'll see something like this:



As you can see, there are many paste icons that offer you various choices on how to paste your copied data. You can even select **Paste Special...** at the bottom to view additional options. It's beyond the scope of this lesson to review every paste option, but we'll cover some of the more important options that you're likely to use as a data analyst.

Perhaps one of the most common decisions you'll make when you paste data is whether to include things like formulas and formatting, or just pasting the values alone, or some combination of those.

The default paste option is **Paste All**, which will paste the formula, the number format, and the font format (called source formatting). This icon will give you **Paste All**, and so will the keyboard shortcut **CTRL+V**:



Using **Paste All** is the default for most users, and in many cases, it works well. However, if you want to specify how the data is pasted with more control, you have several options. The second icon gives you the option to paste with only the formula, but no formatting:



The third icon gives you the option to paste the formula with the number formatting:



The last icon gives you the option to paste the formula with the font (source) formatting:



The next two options we'll cover are equally important for data analysts. Often, all you want is the value of a cell. You don't want the formula or any of the formatting. Just the value, and that's it! In this case, you'll select **Values**:



In the event you want some formatting with your values, you can use one of the adjacent icons to the right to select from **number** or **font** formatting.

The last paste option we'll cover on this screen is incredibly valuable and not well-known to many users. It's called transpose. Transpose is a fancy term that means to rotate the axis of a table so that the columns and rows are switched. For example, suppose you've been given the following data on monthly sales from three stores:

January	200	250	200
February	300	200	350
March	250	200	300
April	350	350	400
May	200	300	350
June	150	200	300

Now suppose that your analysis requires that you list the months as column headers instead of at the start of each row. How would you do that? The easiest way to accomplish that is to transpose the table. You'll do that now.

5.5.1 Instructions

1. Open the **transpose.xlsx** workbook containing the monthly data shown above.
2. Select the entire table, which is the range **A1:D6**, and copy it.
3. Move your cursor to cell **F1** and, using the **Paste** tool, select the **Transpose** icon to paste your table in a transposed way.
4. The transposed table will look like this:

[illegible]

5.6 Using VLOOKUP

On this screen, you'll see some methods for combining datasets using Excel functions that are very useful. In particular, there is a function called `VLOOKUP()` that you'll want to know.

Before we get started, suppose that you have two tables in Excel, and at least one of the field names is common between the two tables. For example, let's say one of your tables contains an employee's social security number and name, and the other table contains social security number and annual salary.

The first table may look like this:

SSN	Name
111-22-1111	W. Lang
222-33-0101	T. Patel
000-11-0000	L. Kim
444-99-2222	B. Williams
333-22-2121	J. Smith

The second table may look like this:

SSN	Salary
111-22-1111	85,000
222-33-0101	79,000
000-11-0000	87,000
444-99-2222	56,000
333-22-2121	64,000

Suppose that you want to create a third table that lists `Name` along with `Salary`. Take a few moments, look at both tables, and think about how you'd do this.

If you guessed that the key is using the common values listed in **SSN**, *you're correct!* In fact, the actual term for a field that is common between tables is the word **key**. The key provides the information that links two or more tables, enabling you to combine the data. You'll hear quite a bit more about keys in other courses, especially those involving databases. In this example, if you want to know the salary of "W. Lang", all you'd need to do is to take the **SSN** of "111-22-1111" and match it to the **SSN** in the other table next to **Salary**, and you'd retrieve an annual salary of "85,000".

SSN	Name		SSN	Salary
111-22-1111	W. Lang		111-22-1111	85,000
222-33-0101	T. Patel	↑	222-33-0101	79,000
000-11-0000	L. Kim		000-11-0000	87,000
444-99-2222	B. Williams		444-99-2222	56,000
333-22-2121	J. Smith		333-22-2121	64,000

Answer

When you use a key value to find information, you're "looking up" the value in the table that matches your key. Excel has a function called **VLOOKUP()**, (which stands for **vertical lookup**), and we use it to find and extract data from a table. For **VLOOKUP()**, you supply it with four parameters in the parentheses:

1. The value you're trying to find
2. The range, starting with the column that includes the value you're trying to find
3. The index number of the range where your answer is located
4. An exact (FALSE) or approximate (TRUE) return value

As such, the parameters in **VLOOKUP()** will look like this: **VLOOKUP(value to find, starting column to find it, index for return value, false)**. You'll use it now to get some information.

Suppose that you want to get the value in **Weekly Sales** when the **Temperature** was 84.34, let's see how to do that now.

First, we open up the workbook **features_level_3.xlsx**. In cell **J1**, enter the header **temp**, and in **K1**, enter **sales**. In cell **J2**, enter the value you seek, which is 84.34. Finally, in cell **K2** enter the formula **=VLOOKUP(J2,C:H,6, FALSE)**.

SUM									
=VLOOKUP(J2,C:H,6, FALSE)									
	B	C	D	E	F	G	H	I	J
1	Date	Temperatur	Fuel_Pric	CPI	Unemployment	Is_Holiday	Weekly_Sales		temp
2	2/5/2010	42.31	2.57	211.09636	8.106	FALSE	1459601.17		84.34
3	2/12/2010	38.51	2.55	211.24217	8.106	TRUE	1621031.70		
4	2/19/2010	39.93	2.51	211.28914	8.106	FALSE	1604775.58		
5	2/26/2010	46.63	2.56	211.31964	8.106	FALSE	1527014.04		
6	3/5/2010	46.50	2.63	211.35014	8.106	FALSE	1508237.76		
7	3/12/2010	57.79	2.67	211.38064	8.106	FALSE	1391013.96		
8	3/19/2010	54.58	2.72	211.21564	8.106	FALSE	1494251.50		
9	3/26/2010	51.45	2.73	211.01804	8.106	FALSE	1517428.87		



When you hit **Enter**, you'll see your answer, which is 1597868.05:

	J	K
	temp	sales
	84.34	1597868.05

You can imagine building a separate table of information by consolidating data using **VLOOKUP()** and extracting individual values.

The **VLOOKUP()** function is very versatile, but it has some limitations, such as the following:

- The function cannot search columns to the left of the search range
- If there are duplicate matches, it only returns the first one
- Inserting a column in your table may break the formula

On the next screen, we'll introduce you to other functions that will help consolidate your data.

5.6.1 Instructions

Suppose that you want to know if it was a holiday when the CPI was 211.4951902. To do that, you'll use **VLOOKUP()** and apply it to the **CPI** and **Is_Holiday** columns.

1. Open up the **features_level_3.xlsx** workbook you saved earlier in screen 4.
2. In cell **J1**, enter the header **cpi_val**, and in **K1** enter **holiday**.
3. In cell **J2**, enter the value you seek, which is 211.4951902.

- Finally, in cell **K2** enter the formula `=VLOOKUP(J2,E:H, 3,FALSE)`.
- Save your workbook.

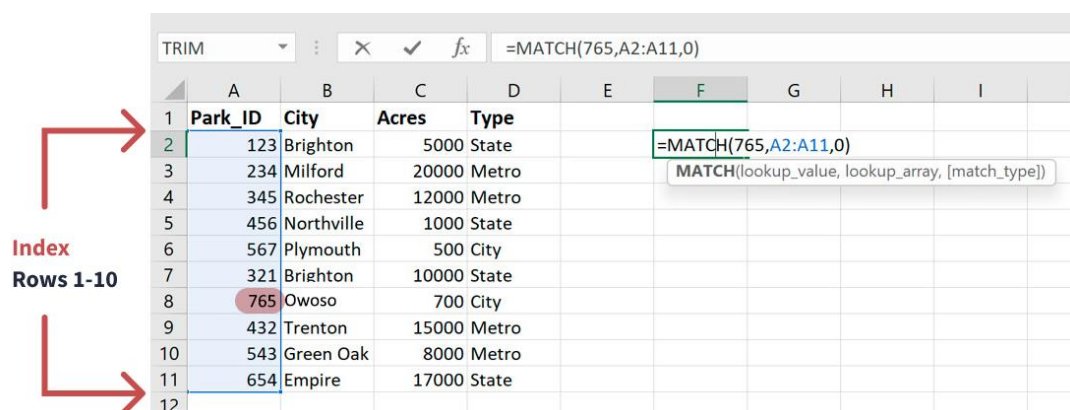
5.7 Using INDEX/MATCH

On the previous screen, we introduced you to `VLOOKUP()`, which many consider a more "advanced" Excel function because it takes several parameters, and the idea of using index numbers in a range can be a new concept to many users. On this screen, you'll learn about two more functions, `INDEX()` and `MATCH()`, that are often used together, so most users just refer to them as `INDEX/MATCH`.

Recall from using `VLOOKUP()` that you had to supply an index number in a range to locate your answer, and that concept will help you to understand how `INDEX/MATCH` works. This may not make sense now (it will shortly), but you'll use the answer from `MATCH()` as an input to `INDEX()`, so let's begin by looking at how `MATCH()` works.

The function `MATCH()` will return the index position of a value in a given range. The parameters in `MATCH()` will look like this: `MATCH(value to find, range to find it, 0)`, where the last parameter is similar to `VLOOKUP()`, in which you specify an exact or approximate match, with `0` being an exact match. Also, the range you use to find your index number can be either horizontal (across columns) or vertical (down rows).

As an example, suppose you're working with some local non-profits to get information on various parks in your area. You've been given a small table that lists information about parks, and you've been asked to consolidate some of the information. You've been asked to identify the `City`, `Acres`, and `Type` for the `Park ID` equal to "765." To begin, you need to get the index location of the `Park ID`. Here's where you use `MATCH()`:



	A	B	C	D	E	F	G	H	I	J
1	Park_ID	City	Acres	Type						
2	123	Brighton	5000	State		=MATCH(765,A2:A11,0)				
3	234	Milford	20000	Metro						
4	345	Rochester	12000	Metro						
5	456	Northville	1000	State						
6	567	Plymouth	500	City						
7	321	Brighton	10000	State						
8	765	Owoso	700	City						
9	432	Trenton	15000	Metro						
10	543	Green Oak	8000	Metro						
11	654	Empire	17000	State						
12										

As you can see, you'd use the formula `=MATCH(765,A2:A11,0)` to return the index position of "765" from the given range, which returns "7" because "765" is in

the seventh row of the range A2:A11 (even though it is in row 8 of the worksheet). This key piece of information will help you complete your task.

Now it's time to use INDEX(). The function INDEX() looks like the following: INDEX(range to search, index number). So, in this example, if you want to find the City where Park_ID equals "765", you'd use the following formula: =INDEX(B2:B11,7), because City is found in the range B2:B11, and you want the value in the seventh index position. Note that the range to search parameter needs to be a single column or row; it cannot span multiple columns and rows.

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J
1	Park_ID	City	Acres	Type						
2	123	Brighton	5000	State						
3	234	Milford	20000	Metro						
4	345	Rochester	12000	Metro						
5	456	Northville	1000	State						
6	567	Plymouth	500	City						
7	321	Brighton	10000	State						
8	765	Owosso	700	City						
9	432	Trenton	15000	Metro						
10	543	Green Oak	8000	Metro						
11	654	Empire	17000	State						
12										

The formula bar shows `=INDEX(B2:B11,7)`. A tooltip for the INDEX function is displayed, showing the syntax: `INDEX(array, row_num, [column_num])` and `INDEX(reference, row_num, [column_num], [area_num])`.

Consider how you'd combine both functions to get the same result. If you're thinking you can use the output of MATCH() to get the index number parameter you need inside the INDEX() function, you're correct! Doing so would give you this formula: =INDEX(B2:B11,MATCH(765,A2:A11,0)), which produces the city name "Owosso" because it's the value in the seventh position in the range B2:B11.

Now think about how you'd do this same thing to find information from the columns Acres and Type. What would you need to change in the current formula and why?

Now it's time for you to try this.

5.7.1 Instructions

1. Open the workbook features_level_3.xlsx.
2. In cells M1, N1, and O1, add the header names sales, store, and fuel.
3. You want to get information on the store that had weekly sales of 1448938.92. So, in cell M2 enter 1448938.92.

- You want to get the store number, so in cell **N2**, under **store**, enter this formula: `=INDEX (A2 : A547, MATCH (M2, H2 : H547, 0))`.
- You should see store number "1" after you hit **Enter**.

4. Now do the same thing for cell **02**, and get the fuel price from column **D**.

- You should only need to tweak your formula slightly, so now it looks like this: `=INDEX (D2:D547,MATCH (M2,H2:H547,0))`. You should now see "2.826" or "2.83" (depending on your formatting) in cell **02**.

5. Save your workbook.

Here's a shot of using the `INDEX` formula:

[illegible]