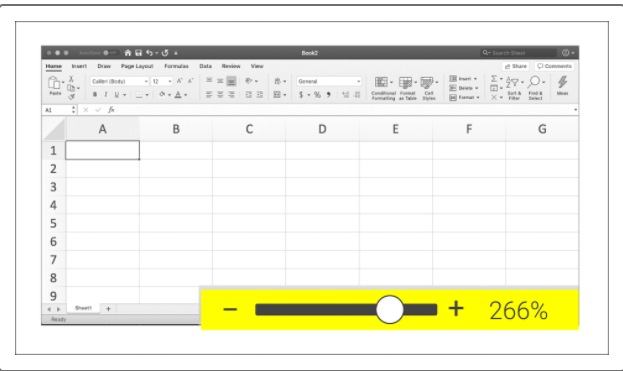
**Resize the Data**

When you open the data file, one thing you'll probably notice right away is the small text size. Because there are so many columns, Excel automatically resizes the sheet to fit everything on one screen, but that often makes the text too small to read. Let's adjust the zoom to make the text legible.

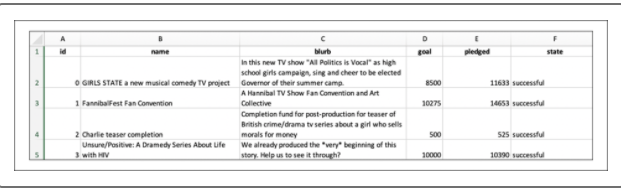
To do this, look at the bottom right of the screen, where you'll see a bar with a slider---a negative (minus) sign on the left and a positive (plus) sign on the right. To the right of the plus sign is a percentage.



This percentage indicates the magnification of the text. The text can be enlarged by dragging the small white circle to the right, or by clicking the "+" sign until the text is a comfortable size (in some versions of Excel, this may appear as a gray verticle bar instead).

## Tabular Data

When you open the data file in Excel, what you're looking at is a **worksheet**, also referred to as a sheet. At a glance, we can see that the data is arranged in rows and columns. Data in this format is called **tabular data**.



**Tabular data** is data that is displayed in a column and row format. This format isn't limited to Excel spreadsheets; any data displayed as a table is considered to be tabular. This includes digital tables on a website and printed tables in a textbook.

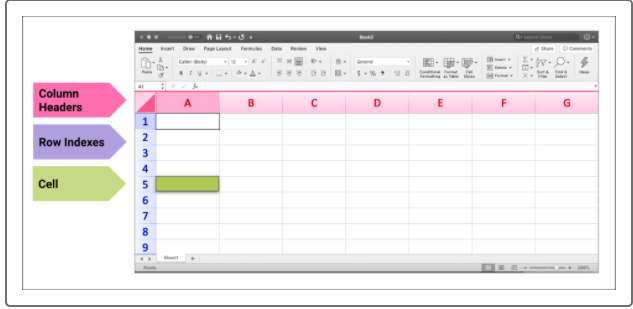
## Anatomy of a Worksheet

You have a worksheet of tabular data that you have resized to make it readable. Now let's break down the worksheet a bit more to understand what it's comprised of.

### Headers and Indexes

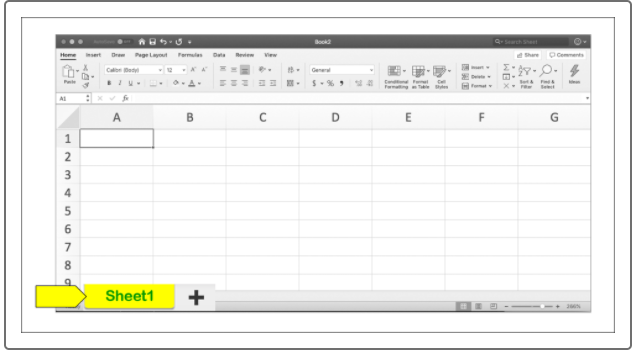
The letters along the top of the columns represent the **column headers,** and the numbers to the left of the rows are the **row indexes**. The headers and indexes help us identify where each data point is located.

In the following image, the cell is located at the intersection of column A, row 5, so we refer to it as A5.



### Worksheet Names

At the bottom of the sheet is a tab labeled "Sheet1." This is our current, or active, worksheet.



To create a new worksheet, click the plus sign (+). When multiple sheets are being used, the left and right arrows allow us to navigate between them. To help us more easily identify our data, let's rename "Sheet1" to "Kickstarter" by following these steps:

1. Right-click the sheet name (currently Sheet1).
2. Select Rename from the pop-up menu.
3. With the current name highlighted, type the new name and press Enter.

## How Many Columns and Rows Are There?

Using a mouse or laptop touchpad, scroll left and right and up and down to get a feel for how large the worksheet is. That's a lot of scrolling, right? Because of its size, manually scrolling through the worksheet will take some time; so instead we'll use a few shortcuts to get an idea of how many rows and columns we're working with.

To quickly view the last column containing data, press the Command and right arrow keys (Mac), or the CTRL and right arrow keys (Windows). This will bring the cursor to the final column. Each column's header is a letter of the alphabet, which keeps track of how many columns are in place.

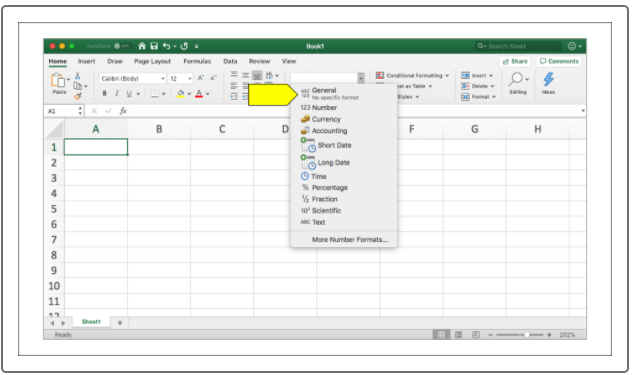
The length of a spreadsheet can be viewed using a similar shortcut: press the Command and down arrow keys (Mac), or the CTRL and down arrow keys (Windows), to bring the cursor to the very last row of data. The number to the left tells us how many rows of data are in the worksheet. Note that there are many keyboard shortcuts like these that can be used in Excel as well as other programs.

Smaller datasets (for example, a dataset with 5 columns and 10 rows) can be manipulated manually, which makes manual calculations within the worksheet a more doable task. However, since we are working with a substantial dataset, we'll need to use built-in formulas and formatting to perform our calculations, which we'll get to later on.

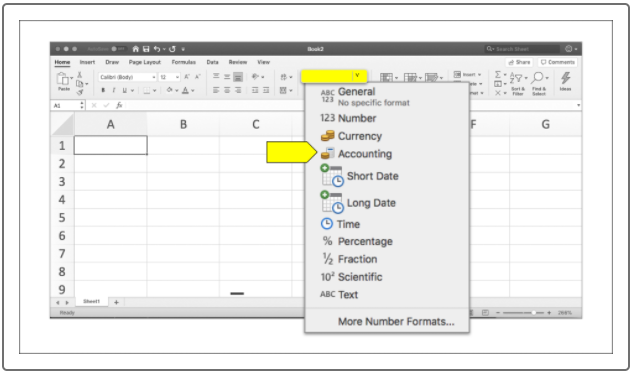
## What Types of Data Are Present?

Another reason to take an initial look at your data is to familiarize yourself with the types of data you'll be working with. Looking at the dataset, you'll notice that some of the columns contain text while other columns contain numbers. And not all numbers are in the same format; there are monetary units present as well as dates.

You can determine the data types in a column by clicking on a cell. For example, first click on the Home tab, then select cell C2. You'll see that it's a description of the Kickstarter campaign, and that Excel has assigned it a specific data type. Look at the top toolbar and note the dropdown menu that has General selected. This is the default format of the column's data.



Now let's check the next column over, **D2**. The value has changed to Accounting because this cell contains monetary data.



**Is the Data Readable?**

Turn your attention back to the top of the dataset and scroll to view columns I and J. You may notice that the data in these columns look like dates, but they're not in a readable format. This is a common issue in data analysis; you'll often encounter data that needs to be converted to make it readable. Right now, the Deadline and Launched\_at columns contain Unix timestamps rather than dates in a standard format.

How do we know the data is Unix timestamps and not random numbers? We already know that the data is supposed to represent a date, and timestamps like these are common. But to be sure that the data are timestamps, we can use a [timestamp converter (Links to an external site.)](https://www.epochconverter.com/) tool.

**NOTE**

Learn more about the use of [Unix timestamps (Links to an external site.)](https://websiteseochecker.com/blog/what-is-timestamp/).

Select the text in one of the cells and copy it. Then, paste it into the online converter tool to see how it's interpreted. After clicking the "Timestamp to Human date" button, the tool will provide the date and time assigned to that particular string. This confirms that we're working with Unix timestamps, so we know that we'll need to convert those timestamps into a format that's readable. This task will require an advanced formula, which we'll cover as we take a more in-depth look at the Kickstarter dataset.

Now we can see that our data is stored in columns A through N and rows 1 through 4115—that's a lot of data! Let's break down some of the information we're working with:

* The Goal column tells us how much money each campaign will need to succeed.
* The Pledged column tells us how much each campaign actually made.
* The Outcomes column tells us if the campaign met its goal.
* The Country column lists the country in which the campaign was started.

But even with this data breakdown, the current worksheet is still a little overwhelming, right? Let's make it more user-friendly by applying filters to the columns. **Filters** allow us to display only the specific data that we want to focus on.

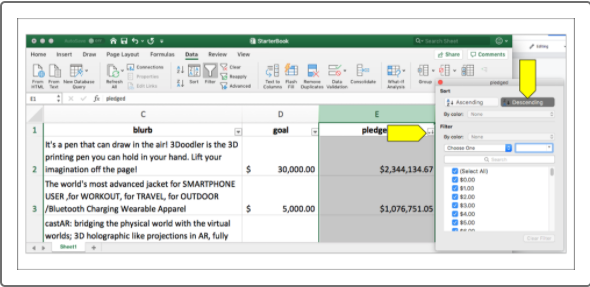
Let's focus first on the money raised by various campaigns. Louise estimates that her play will cost $12,000, so we can use data from the Pledged column to research projects with a similar monetary goal. We'll use filters to apply this customization.

## Filter the Data

Filters are commonly used in Excel because they make the data far more user-friendly. For instance, a dataset can be filtered to omit data we aren't interested in, or to single out specific data we want to learn more about. It helps negate some of the extra "noise"—information that is unnecessary to a project.

Follow these steps to add filters to your Excel worksheet. Let's add a filter to the Goal column first.

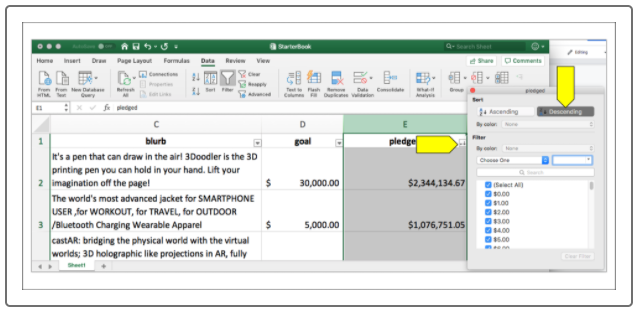
1. Begin by clicking the letter **D** above the goal column to select the entire column.
2. While the column is still selected, click the Data tab at the top of the screen.
3. Find the Filter button on the ribbon and then click it.



Once these steps are completed, a down arrow will appear next to the Goal header on the workbook; this will reveal the sorting and filtering menu.

Now we easily sort or filter the data as we see fit. Let's test this tool by sorting the pledged amount from highest to lowest.

1. Click the down arrow on the Pledged column (column E) to reveal the Filter menu.
2. Click "Descending," and then close the window. **Note:** "Ascending" sorts values from lowest to highest; "Descending" sorts values from highest to lowest, which is what we want. In some versions of Excel, you will see "Sort Largest to Smallest" instead of "Descending" and "Sort Smallest to Largest" instead of "Ascending."



**NOTE**

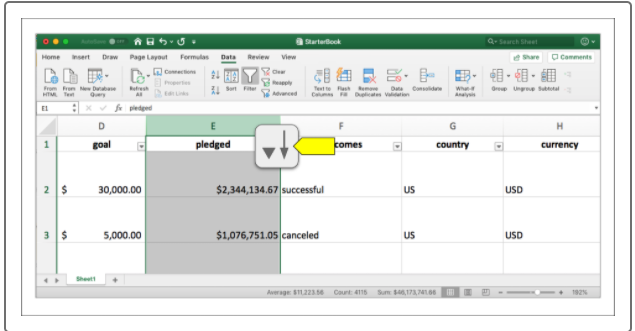
Your version of Excel may produce a dialogue window with the following message:

"Microsoft Excel found data next to your selection. Since you have not selected this data, it will not be sorted. What do you want to do?" If you receive this message, select "Expand the selection" before continuing.

This will sort the Pledged column as well as all of the data tied to it. We can verify this by looking at the data contained in column A: the ID numbers are no longer in order.

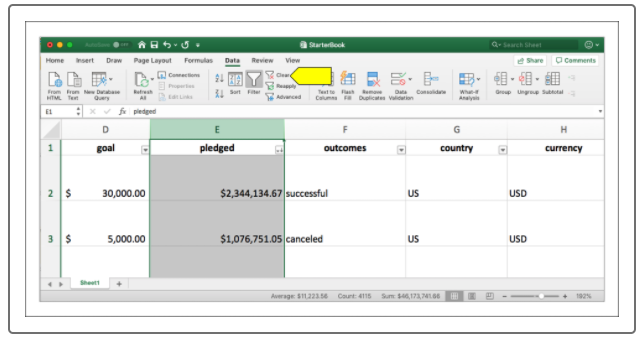
Now spend some time practicing how to format and sort different columns in your worksheet. Have fun with this! For example, try to find the highest goal, or combine filters by looking for the highest successful goal.

Remember, when you sort and filter multiple columns, it's easy to lose track of what's been sorted and filtered. Excel's subtle (but helpful) way of telling us which columns are sorted is to change the down arrow icon to a filter or arrow icon, depending on how the data was adjusted.



If you get lost in the data, that's okay! Clearing the filters and starting over can be done in two quick steps. To reset all filters, do the following:

1. Click the "Sort & Filter" button at the top right of the Excel window.
2. Click Clear.

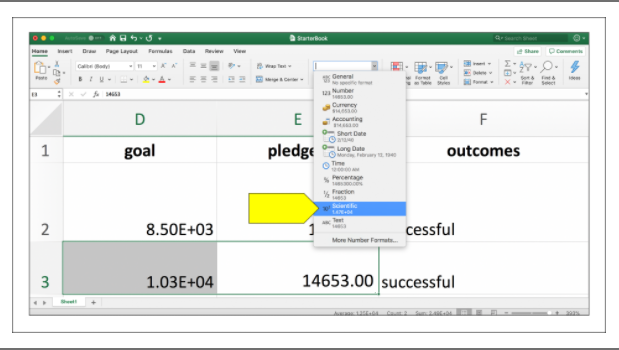


**NOTE**

An alternate way to revert your data to its original setting is to use the **Undo** feature. The hotkeys to quickly Undo a recent change are command + z (macOS) or ctrl + z (Windows).

**Format the Spreadsheet**

Let's backtrack a bit. The Goal column was automatically sorted, but now some of the numbers in that column look odd. This is because the column width doesn't capture the entire number, so it has been shortened using **scientific notation**.

****

**IMPORTANT**

**Scientific notation** is a method for displaying very large or very small numbers in a more compact way. Excel automatically adjusts those large or small numbers to fit within a column.

For example, the number **3E+07** can also be displayed as **3x10^7^**. This is scientific notation for 30,000,000. This means that the number starts with a 3 and has an exponent of 7 (seven zeros).

Note that if you are using a larger monitor, then the column may not be presented as scientific notation and is already in a normal numeric format.

You can automatically widen the column to hold all of the data by following these steps:

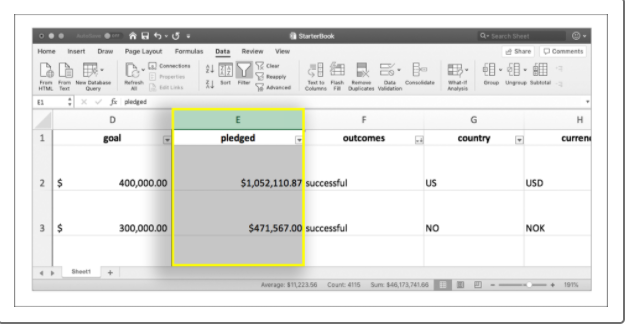
1. Place your cursor over the small line between columns D and E. The cursor should change in appearance to a vertical bar with an arrow pointing either left or right from the center.
2. Double-click to expand the column to fit the value with the most physical width.

You should now see the goal amounts arranged in order from highest to lowest. However,  you might notice that when you scroll further down the sheet, it's difficult to determine what's in each column. Some of the data may not make much sense without a heading attached to it. The solution? Freeze the header.

## Freeze and Unfreeze Rows, Columns, and Panes

**Freezing** is a feature of Excel that allows portions of a spreadsheet, such as a row or column, to be locked in place so that it is always displayed, no matter what part of the sheet we're looking at. This means that you don't need to commit column headers to memory; simply freeze them so that they are always displayed as you scroll through the sheet.

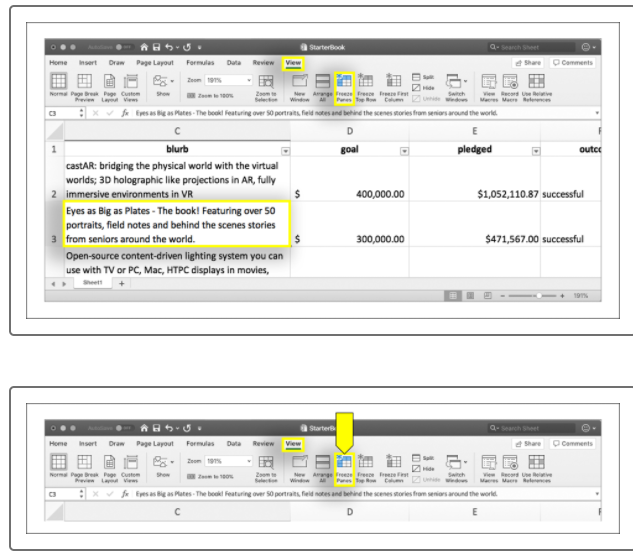
When a column and row are frozen together, they become a **pane**. Think of this as a window: the window can be opened and moved up and down or left and right, but the pane stays in place. The same principle applies here, as shown in the following image.

****

Now let's reset our view by unfreezing our panes. Click the Unfreeze Panes selection in the Freeze Panes drop down list while in the View tab, as shown in the video.

Say we want to set the name column and title row as a pane rather than freezing the first column and the top row. This is where the Freeze Panes button comes into play. This button freezes a column and row simultaneously.

Let's try it. In the worksheet, select cell C3 and then click Freeze Panes in the View tab. The new pane is created based on the location of the selected cell, in this case, C3.

****

This tells Excel to freeze the column to the left of C3 and the row above C3. Now scroll through the data, horizontally and vertically. See how both the **ID** and **name** columns stay in place? To complete the pane, the header row is also frozen. Pretty useful, right?

**IMPORTANT**

Remember to save your work periodically so that you don't lose it if the program crashes. Command+S (Mac) or CTRL+S (Windows) is a quick shortcut to save any work in progress.

**What's Conditional Formatting?**

**Conditional formatting** in Excel refers to customizing the appearance of a cell based on its value. This is a great way to provide viewers with information at a glance. And, much like customizing an essay or report with Microsoft Word, there are several ways to format a spreadsheet: each cell can be filled with a different color, various fonts and text sizes can be used, and so on. For our project, we'll apply conditional formatting to some of the columns in the worksheet in order to more easily view the outcomes of Kickstarter campaigns.

**IMPORTANT**

**Conditionals in Code** If you're an experienced Excel user who's looking forward to the coding portions of this course, that's great! Conditional formatting is a concept that you'll see in action often.

Conditionals make frequent appearances in all programming languages. When coding, they are referred to as conditional statements because they are paired with keywords such as *if*, *and*, and *or*.

When using conditionals in Excel, a cell can be highlighted a certain color based on its value. *If* the value of a cell is greater than 50, *then* color it green.

These same logical rules apply when conditional statements are included in coding script.

Let's put this into action by applying conditional formatting to our data.

**Use Conditional Formatting**

There are four main categories in the Outcome column---Successful, Failed, Canceled, and Live---so let's color-code by category. We can use conditional formatting to automatically apply a preselected color to the cell.

Watch the following video to learn about conditional formatting in Excel. Choose the video that corresponds to your operating system.

## Create a New Column

In your Kickstarter worksheet, name the next empty column (cell O1) Percentage Funded. In this column, we'll use Excel's **ROUND** formula to determine how much of the campaign's goal was met. Data from the Pledged and Goal columns will be used to find the percentage funded.

## Adding New Data to the Column

In the first empty row of our new Percentage Funded column, enter =ROUND(E2/D2\*100,0) into the cell. This is our first row of data, which begins on row 2. The very first row, the header row, is row 1.

Let's break down this formula to see how it works:

1. =ROUND tells Excel that we'll be using the ROUND formula.
2. (E2/D2 specifies which data is being utilized in the formula: column E, row 2 (pledged) divided by column D, row 2 (goal). The result is the percentage of completed donations, shown in decimal form.
3. \*100, multiplies that result by 100, giving us the percentage as a whole number.
4. 0) specifies how many numbers after the decimal will be viewed. In our case, that's zero.

Because there are so many rows of data, it would be time-consuming to apply the formula to each individual cell. Thankfully, there's a shortcut we can use to apply the formula to the entire column.

After inputting the formula and running it, the result should be a whole number of the percentage. Now, to apply the formula to the rest of the column, click the cell containing the new data. See how it's outlined in green? This tells us that it's the active cell.

Next, place your cursor over the bottom right corner of the cell. The crosshairs of the cursor change from white to black; this is how you know you're in the right spot. When the crosshairs are black, double-click that spot; Excel will automatically apply that formula to the entire column.

## Value Shading

To summarize what we just did: we created a new column of data using a new formula to find the percentage of a campaign's funding. Now we can quickly determine how close a campaign came to reaching---and in some cases, exceeding---their funding goal.

We'll now take it a step further and apply conditional formatting to the new column. This time, instead of assigning a color to a column based on a single word or phrase, we'll apply a range of colors, based on the percentage funded. Like we did for the Outcomes column, we'll create a formatting rule to let the colors do the talking for us.

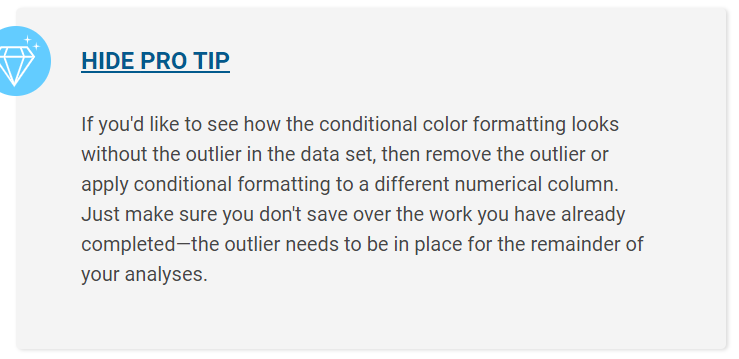
Highlight the Percentage Funded column in the worksheet, then make sure the Home tab is selected before clicking the Conditional Formatting button on the ribbon. Now we'll create a new rule by following these steps:

1. Select "Color Scales" followed by "More Rules." At this point, we can select colors based on the lowest and highest values. The colors will automatically change based on the cell's value.
2. Select the color red as the minimum color and blue for the maximum color.
3. Click OK.

The result is a color graded reference for Percentage Funded column. In a perfectly curated dataset, you would easily see which projects were fully funded by looking at the color grading alone. Our current dataset still has the color scheme, but if you begin to scroll down the sheet, you may notice that there is quite a lot of a single color. Why is this?

Even though you may have to scroll through the data a bit before you find a color transition, the data is providing a message: there is an outlier in the data set. An outlier is a data point that is abnormally high or low when compared to the rest of the data.

What does this mean for the data you're working with? Nothing yet, because we'll dive more deeply into outliers and how to handle them later. It's a good idea to keep it in mind as you work through your analysis, though.



## Average Donations

Kickstarter, like other crowdfunding platforms, allows project creators to add incentives for different pledge amounts. We can help Louise set up her incentives by first determining how much money people have pledged to campaigns historically. Is it $3? $10? Even $50? Excel will help us perform this type of calculation.

The first step is to create a column that will provide a quick look at the average donation. For this step we'll follow a process like the one for creating and filling the Percentage Funded column. When we created the Percentage Funded column, we:

1. Selected an empty column and added a header
2. Input the round function into the cell below the header
3. Applied the formula to the rest of the column

Let's do this again with a new column (column P) that we'll name "Average Donation."

Next, we'll use the ROUND formula again, but with a few modifications, shown below.

=ROUND(E2/L2,2)

We last used this formula to create a percentage as a whole number. This time we're finding an average, so we want to view the next two digits after the decimal. By modifying the formula to use a 2 instead of a 0, we're telling Excel to include those extra two digits.

To find the average, we'll use data from the Pledged column (column E) again. But instead of dividing by the goals (found in column D, as we did last time), we will divide by the backers (column L).

First, make sure the cell's data is set to General, then use this ROUND formula on the first cell of the Average Donation (P2) column. Then, in the next empty cell (P3), make sure the data type is set to General again, then use the same formula with a 4 instead of a 2 to specify how many digits will follow the decimal: =ROUND(E2/L2,4). See the difference? Since we were testing the formula in cell P3, let's delete it now that we've seen the difference.

Now apply the ROUND formula for the first cell to the entire column. Do this by selecting the top cell (P2) and hovering your mouse cursor over the bottom right corner of it until a black + sign appears. Double click the + to apply the formula to the entire column. Make sure there are only two digits after the decimal point.

When working with data, you're bound to encounter errors. Some errors occur due to a mistake, while others pop up because Excel doesn't recognize a formula's output format, even if it's valid.

Techniques used to fix errors vary. Sometimes errors are obscure and require research, while other errors have simpler fixes, such as correcting a typo. The process of researching an error and incorporating its solution is called **debugging**.

Recall the #DIV/0! error we encountered earlier. Why did it occur? Let's investigate.

Kickstarter requires every campaign to have a fundraising goal. However, not every campaign has backers, which means, in some cases, there is no number to divide by in the formula. Our formula, =ROUND(E2/L2,2) uses data from the Pledged and Backers columns. Let's look at row 124 and plug in the numbers ourselves. Now, our formula becomes =ROUND(0/0,2). If we were to take out a calculator and try to divide 0 by 0, we'd get an error there, too. No wonder it's not working correctly. The #DIV/0! error occurs because numbers are not divisible by zero.

While this error doesn't hinder our research, we can and should clean it up.

## Using IFERROR()

To improve the look of our formula output, we'll need to integrate the =IFERROR(value,value\_if\_error) formula. This formula catches errors and replaces them with a user-defined input. In addition, we'll add a bit of a twist by nesting this formula and the ROUND formula.

**IMPORTANT**

Nesting formulas occurs when one formula is nested, or lies inside, another formula. This is powerful because it allows both formulas to run simultaneously.

The **IFERROR** formula is designed specifically to hold another formula within it, while the ROUND formula is designed only to perform mathematical calculations. If the order was reversed, neither formula would be correct, resulting in additional errors. Therefore, the order of nesting is important.

To address the division error, we'll use this formula:

=IFERROR(ROUND(E2/L2,2),0)

Let's break down what's happening here.

* IFERROR is now the beginning of the entire formula, so the equals sign is in front of it instead of ROUND.
* The value is our intact ROUND formula; nothing has been changed.
* ,0) tells Excel that we want a zero-value input when the formula attempts to divide by 0. The formula is completed by closing the parenthesis.

We specify a zero-value input instead of text (such as "no backers") because we want a numerical data type throughout the column. This way, if we perform analysis on that column's data, we won't encounter additional errors.

Let's first edit the existing formula. While the first cell is selected, the dedicated formula bar (visible just above the active sheet) displays the current formula. Click inside this bar to edit the formula to use the IFERROR formula described earlier. Then, apply the updated formula to the entire column (do this by selecting the top cell and hovering over the bottom right corner and double clicking the black + when it appears). Next, scroll down to row 124. You should see that the error has been replaced by a zero.

Because Louise is interested in starting a campaign for theater, let's filter the worksheet to show only theater campaigns in the Category/Subcategory column. Take some time to familiarize yourself with this smaller dataset, also known as a subset of the data. Notice any differences? Similarities? This will allow a more focused view of our category of choice by trimming down the data and eliminating what we don't need.

## Create Subcategories

Start by putting the subcategories into their own column. In the worksheet, clear all the active filters (if there are any), and then follow these steps:

1. Select the "Category and Subcategory" column.
2. Copy the column using the keyboard shortcut Command+C (Mac) or CTRL+C (Windows).
3. Paste the data into the next empty column using the keyboard shortcut Command+V (Mac) or CTRL+V (Windows).
4. Click the Data tab.
5. Click the "Text to Columns" button.
6. The "Convert Text to Columns Wizard" appears.
7. Inside the "Convert Text to Columns Wizard:"
   * Select "Delimited" and click "Next."
   * Uncheck the "Tab" box and check "Other."
   * Place a backslash ( / ) in the box, then click "Next."
   * Select "Text" from the "Column data format."
   * Click "Finish."

Good work---you've broken down the data into more categories. Remember to name the new columns "Parent category" and "Subcategory" respectively. Let's use this new data to create a pivot table.

## Pivot Tables

**Pivot tables** are a powerful Excel tool that condenses data into a summary that delivers information based on our questions.

This is a lot to take in, so let's build a pivot table together, step by step.

**IMPORTANT**

Pivot tables and pivot charts (which we'll learn about next) are extremely versatile and powerful tools. They allow us to pick and choose the data we want to analyze and then tweak it with visual customizations. Pivot tables also let us continue to tweak the view by filtering our chosen data after it's been set to a graph.

Before we can create a pivot table, we need to know what questions we want the table to answer. Ask yourself the following:

* Which data do we want to see summarized?
* How do we want the data to be presented?

Now we have a starting point. Let's dive in! Watch and follow along with the following video to learn more about pivot tables. Choose the video that corresponds to your operating system.

## Pivot Charts

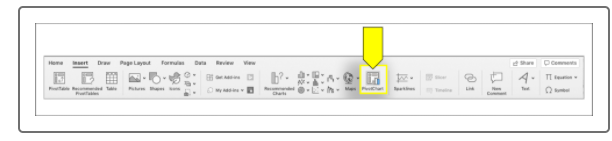
**Pivot charts** complement pivot tables by using table data to create visualizations such as bar charts and line graphs. Visualizations aid in uncovering a link or trend, especially when a table isn't able to.

Return to your pivot table and name the sheet it's currently in "Category Statistics." Let's add a visual representation of the data next by following these steps:

Click anywhere on the pivot table to bring up the PivotTable Analyze tab, and then click it.

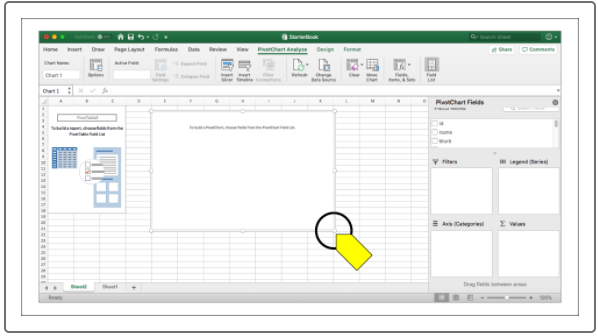


Click the PivotChart button, located in the Tools group on the ribbon.



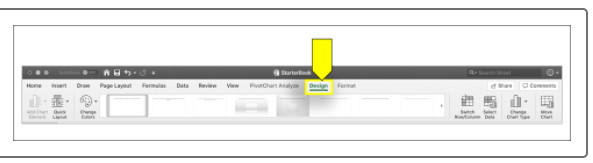
When this button is clicked, Excel will automatically choose a chart based on the table data, which in our case is a clustered column. Depending on your version of Excel, your chart may be automatically added. If it is not automatically added, click "OK" to add the chart to the sheet.

Resize the chart as needed by dragging the corner out from the center of the image.

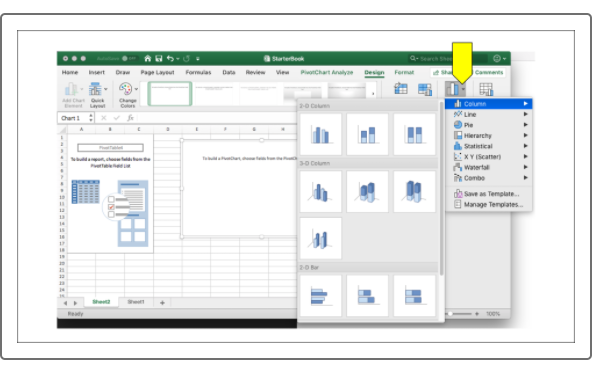


This bar chart is great, but it can be improved even further by changing it to a stacked bar chart.

Click on the chart, and then click the Design tab on the toolbar.



Click the Change Chart Type button, and then select Stacked Column Chart from the pop-up menu.

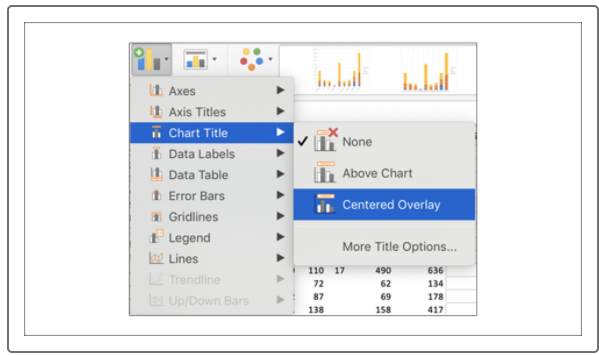


This is a great chart. Each category has a single bar that tells us the same information as the table. Even better, this chart is just as interactive as the table it's linked to. If we change a filter, the table will automatically update itself to reflect the new data. Additionally, information is displayed in a small box when we hover our mouse over a column.

We're not quite done yet, though. It's a good practice to give visualizations a title to link the chart to specific data. This way, if we revisit the image after not working directly with the data for any length of time, the title will remind us of what the data represents.

Add a title to your bar chart by completing the following steps:

1. Return to the Design tab and click Add Chart Element on the left side of the toolbar.
2. From the pop-up menu, select Chart Title followed by Centered Overlay.

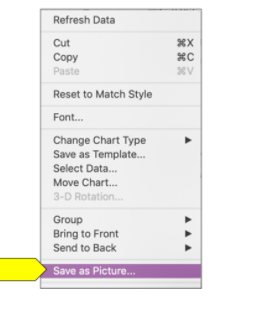


A text box with Chart Title will appear at the top of your chart. Enter a new name by double-clicking the placeholder text and typing in your own title, such as Parent Category Outcomes.

Now that we have a chart with a title, let's save it as an image to use in our report later. We'll be giving Louise a report of our findings when we're done, complete with images of our analysis. It's good to save images as we go; this way, we won't need to retrace our steps to save them later.

#### macOS

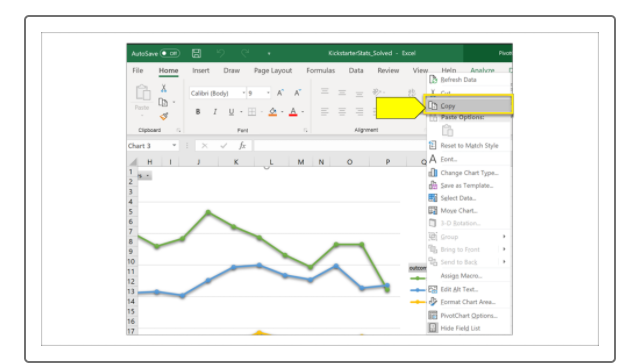
To save an image of your chart, select the chart by clicking anywhere on it. Then right-click and select "Save as Picture."



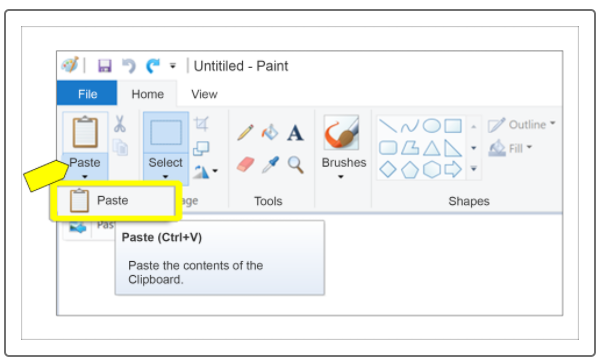
#### Windows

The steps below outline how to save an image of a chart (the chart in the image may not match the chart you have created or are trying to save).

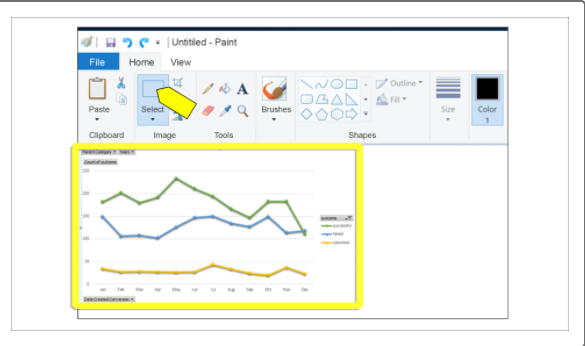
1. Right-click the image and select "Copy" (or use the CTRL+C keyboard shortcut).



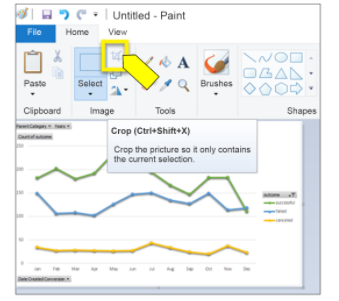
Open a graphics editor, such as Microsoft Paint, and paste the image using the Paste function from the ribbon or by pressing CTRL+V on your keyboard.



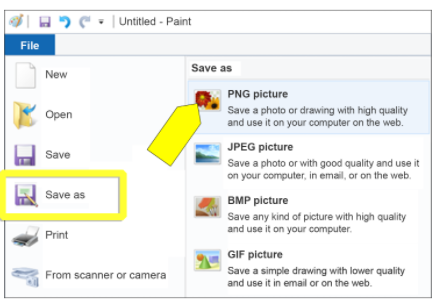
There may be excess white space around your chart. To trim it, use the Select tool to select only the chart.



Click the "Crop" button.



Save your image by clicking "File" followed by "Save as," and then select "PNG picture." Save this new image in the same directory as the data file. We'll need to access it later to add it to our report.



## Chart the Subcategories

Let's make another chart, this time using subcategories. We're using subcategories in order to focus our analysis on an area that is more relevant for Louise: theatrical productions. We'll follow a process similar to the one used to create the original chart using the parent categories, but this time we'll create a chart and pivot table at the same time. Return to the original Kickstarter spreadsheet to get started, and then follow these steps:

1. Click the Insert tab.
2. Click the PivotChart button.
3. Keep the default range selection and place the PivotChart in a new worksheet. Click OK.

**Tip:** The default range includes the entire worksheet, unless there is already a highlighted section. Make sure there are no sections highlighted before completing this step.

1. Name the new sheet Subcategory Statistics and relocate it to the right-most sheet tab (if it isn't there already).
2. Drag and drop the following:
   * Country and Parent Category to Filters
   * Outcomes to Columns
   * Subcategory to Rows
   * Outcomes to Values

**Note:** In some later versions of Excel, Outcomes should be dragged to "Legend (Series)" and Subcategory should be dragged to "Axis (Categories)."

1. Click the PivotTable Analyze tab and then click PivotChart.
2. Click on the graphic to make sure it is selected, then click the Design tab at the top of the screen and select Change Chart Type. Choose Column followed by Stacked Column.
3. (Optional) Customize the chart colors and grid.

The grid is made up of horizontal (and sometimes vertical) lines that help us quickly measure the data being displayed. This can be customized in Excel to change the line style or even remove them completely.

What do you notice? This new pivot chart looks almost exactly like the first one we created and has the same functionality. Let's now filter this chart by country to see if the origin of the campaign has an impact. First, select the U.S. and take a look at the statistics.

Now we know from our initial look at the data that the total number of Kickstarter campaigns is just over 4,000. You can double check this number by looking at the Grand Total on your pivot chart. By filtering for just the U.S., the number of campaigns is adjusted to 3,038. We can see this number in the Grand Total row of the table. Next, update the filter to show Great Britain's statistics.

## Convert Unix Timestamps to Readable Format

Look at the Deadline and Launched\_at columns in your worksheet. The data in these columns is not exactly easy to read. They contain Unix timestamps, which measure time as the number of seconds since midnight of January 1, 1970. While interesting, these timestamps are not exactly something we can use easily.

Thankfully, Excel helps us convert these timestamps into a day-month-year format that we can interpret. For this conversion, we'll need to use another formula.

In a new column (S), add the heading "Date Created Conversion." Then, enter the following formula, making sure that J2 is replaced with the first data cell of the Launched\_at column:

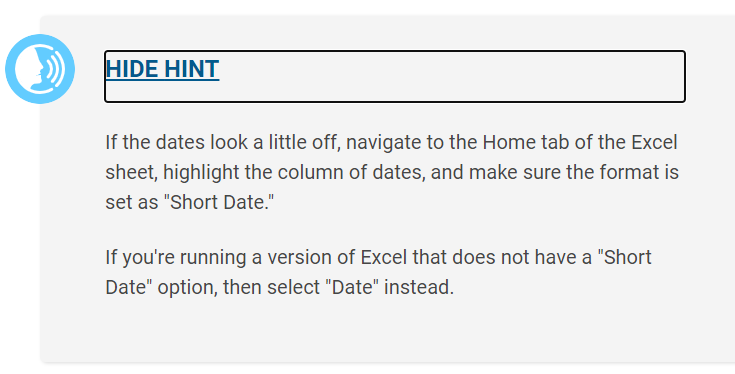
=(((J2/60)/60)/24)+DATE(1970,1,1)

Let's break this down.

* = tells Excel that we're using a formula.
* (((J2/60)/60)/24) iterates the following: cell J2 (the first cell in the Launched\_at column) divided by 60 (seconds), then divide that by 60 (minutes), then divide that by 24 (hours).
* +DATE tells Excel that we're using the DATE formula as well.
* (1970,1,1) is the date that the Unix timestamps began counting from, also known as the **epoch**.

After applying the formula to the column, change the format of the column from General to Date. Human-readable dates should fill the column cells.

Essentially, we're using the formula to figure out how many days, minutes, and seconds the timestamp translates to, and then we're adding it to the January 1, 1970 date. It's a little confusing, but it adds up to the actual time of the campaign launch.

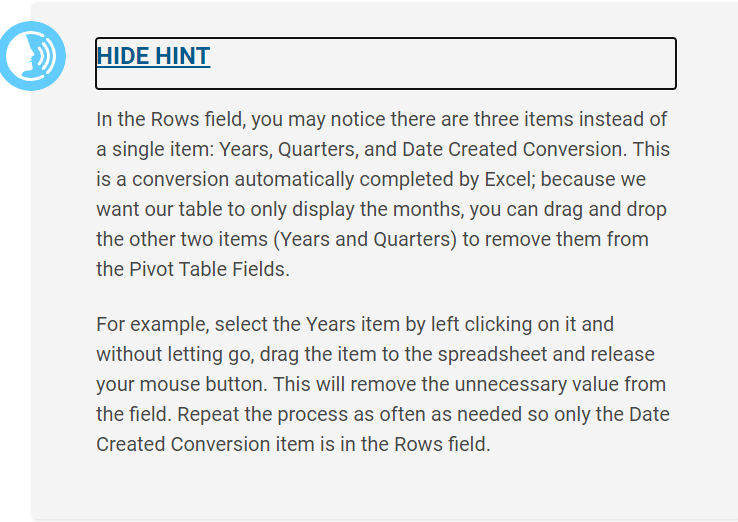


Follow these same steps to create a Date Ended Conversion column (T).

Input the same formula, but change J2 to the first cell of the Deadline column (I2). Test the formula on the first cell to ensure the conversion goes smoothly, and then apply it to the rest of the column.

## Create a New Pivot Chart

Now that we have a date range for each project, let's create another pivot chart. This time, instead of a stacked bar graph, we'll use a line chart to view the data. Filtering by parent category and years, we will want the Columns value to be "outcome," Rows value to be "Date Created Conversion," and Values to be "outcome." Note that when "outcome" is placed in the Values box, it will be renamed "Count of outcome."



**NOTE**

In some versions of Excel, Outcome should be dragged to "Legend (Series)" and Date Created should be dragged to "Axis (Categories)."

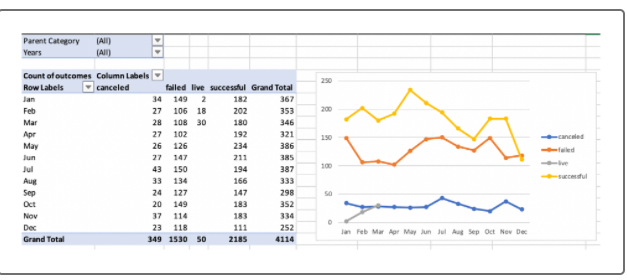
The steps for setting up a pivot line chart are similar to those followed to create stacked bar charts:

1. In the Kickstarter sheet, create a new PivotChart.
2. Rename the new sheet "Outcomes Based on Launch Date."
3. Right-click the chart image and hover over "Change Chart Type."
4. Select "Line" from the pop-up menu.
5. Select the "Line with Markers" chart type from the next pop-up menu.

**NOTE**

Some versions of Excel automatically parse the dates when pivot tables are created. This means that additional columns may be created without additional input from the user.

Line charts are helpful when trying to determine trends. We can see by looking at our new chart that the months of May and June both have a greater success rate.



A bar chart wouldn't be able to convey this information in the same manner, so it's important to keep in mind the type of data story we want to tell.

Save this image in your project folder so we can add it to our report later.

## Filter Charts

Go back to the Category Statistics worksheet and filter the chart to display only theater. Follow these steps:

1. In the Row Labels section on the right side of the page, click the arrow to activate the dropdown menu.
2. Click the "Select All" button to deselect everything. Scroll through the list to select only "theater."

We'll repeat these steps for each chart, but with a few minor differences:

* In the "Subcategory Statistics" sheet, we'll select Plays.
* In the "Outcomes Based on Launch Date" sheet, we'll select "theater" for the parent category.

As you continue to filter your data and refine your charts, keep an eye on how the data is displayed. Explore it using your filters. Do you notice trends between all the categories and subcategories?

Just by glancing at the data, we can determine that theater is a popular and successful type of campaign overall. By using filters, we can see that theater follows the overall trend: there is a spike of successful campaigns that began in June, but that tapers off by the end of the year. By comparison, the data around technology campaigns reveals a different story. Instead of one large spike, their trend lines are a bit all over the place and less predictable.

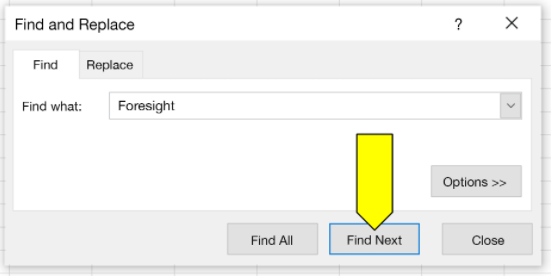
## Use the Search Function

So far, we've visualized all of the campaigns in our Category Statistics sheet. We've discovered trends in the theater category with our Outcomes sheets. (June seems to be a good month to launch a campaign!) All of this has provided us---and Louise--- with information to help plan her campaign. Now let's return to the Kickstarter sheet and focus on a single play: Foresight. This is a play from Great Britain that Louise enjoyed. She's curious about the market in Great Britain and would like to learn more about this particular play. We'll filter the dataset to view this campaign as well as other pertinent information.

Filter the Subcategory column to show only plays, and then use the Find function to search for "Foresight.” The keyboard shortcut to initiate the Find function is (Mac) Command+F or (Windows) CTRL+F. Alternatively, type “Foresight” into Excel’s search bar in the upper-right corner.

**NOTE**

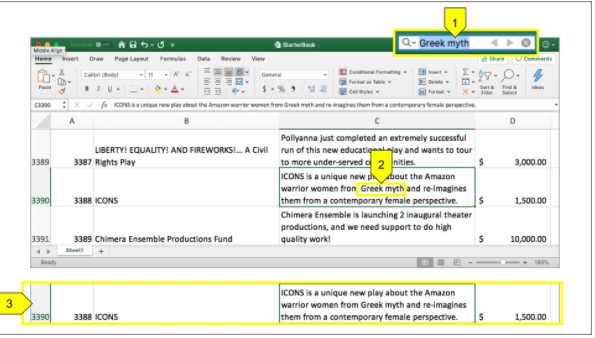
Newer versions of Excel don’t have a search bar in the upper right, but the same CTRL+F/Command+F shortcut will bring up a search box instead. This box will function in the exact same manner as the search bar in older versions of the software. Click “Find Next” to locate the next instance of the term, or “Find All” to view a list of all instances.



Looking at the color of the outcome allows us to quickly determine that the Foresight campaign was successful. It reached 100% of its goal -- it even went over by four dollars!

What else can we learn about this campaign from this data? The average donation is surprisingly high, considering there are only 17 backers. Scrolling further, we can also see that the campaign wasn't active for very long---just under a month.

**VLOOKUP** lets us pull specific columns from our main dataset into a new sheet without having to search for each column and then copy and paste the data. This way, we can pull only the data points we're interested in. For example, if we only want to see the blurb of the play, we can tell Excel to pull only the data in the blurb cell for that play.



To use VLOOKUP, first get all of the titles in a new sheet, which we'll name "Edinburgh Research." In our new sheet, let's add headers: column A will be named "Name" and column B will be named "blurb". Next, write the list of plays in column A. Be sure to spell the titles correctly and use proper capitalization. Here's our list:

* Be Prepared
* Checkpoint 22
* Cutting Off Kate Bush
* Jestia and Raedon
* The Hitchhiker's Guide to the Family

We'll start using the VLOOKUP formula in column B. In cell B2, type the following formula, then press Enter to run it:

=VLOOKUP(A2, Kickstarter!B:C, 2, FALSE)

You should now see the blurb for *Be Prepared* in cell B2. But what is VLOOKUP actually doing?

1. We're starting the formula with =VLOOKUP.
2. Excel uses the value in A2 ("Be Prepared") to look for data for that campaign.
3. We tell Excel to look for this data in the Kickstarter worksheet.
4. It begins looking in column B; then, it searches the next column, column C, for the data in the same row.
5. Excel will then reference the data from the second column, "blurb" (B is the starting column, so it's counted as the first). Then, Excel will retrieve that data and display it in the new table where we executed VLOOKUP.
6. FALSE tells Excel that we're interested in only exact matches.

Copy the formula into B3 through B6 to get the blurbs for all five plays. Let's also add what each play's goal was and the amount pledged. In the new table in the Edinburgh Research worksheet,  first update our headers by adding "Goal" and "Pledged" to columns C and D. To add additional values to the table, paste the same formula into cell C2, and then update its range to B:E. Originally, we told Excel to search for matching data in columns B through C (B:C), but now we want to search further down the same row, through column E instead.

It can be a little confusing, so let's work through this one together. In C2, use VLOOKUP to search for the play we typed into cell A2. The VLOOKUP formula starts in the left-most column of B:E in the Kickstarter worksheet, and then returns the value from the third column. This means that the third parameter will need to be revised as well, from 2 to 3. The updated formula will look like this:

=VLOOKUP(A2, Kickstarter!B:E, 3, FALSE)

Don't forget to specify exact matches.

Now, copy the formula into cells C2 through C5. We can use the same formula to get the amount pledged into column D by adjusting the VLOOKUP formula to take the value from the fourth column: =VLOOKUP(A2, Kickstarter!B:E, 4, FALSE).

## Mean, Median, and Mode

The **mean** is the sum of the data divided by the number of data points. You can think of the mean as answering the question "If every data point contributed the same amount, what would that amount be?" For example, if you and two friends all chipped in to buy a pizza, and you put in $12, one friend put in $7, and the other friend put in $5, the mean cost would be calculated this way:

(12 + 7 + 5) / 3 = 24 / 3 = $8.

The **median** answers the question "Where is the midpoint of the data?" Also known as the 50th percentile, the median is the value that splits the data into two equal halves: 50% of the data is lower than the median, and 50% of the data is higher. To calculate the median, sort the data points in order, and then locate the point in the middle. For example, if the grades on a quiz are 82, 79, 79, 77, 70, 90, 71, 86, 83, first we would put them in order:

70, 71, 77, 79, 79, 82, 83, 86, 90

Since there are 9 scores, the midpoint will be the 5th score after sorting. There are 4 scores above it and 4 scores below it:

70, 71, 77, 79, **79**, 82, 83, 86, 90

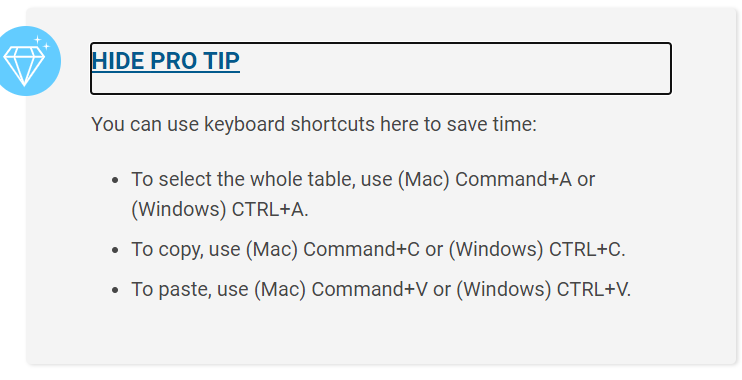
The median score on the quizzes is **79**.

The **mode** answers the question "What value shows up the most?" This question can be trickier than it seems. Take our quiz scores example. In this example, 79 shows up more than any other score, so 79 is the mode. Let's say there is a student who was absent that day and then takes a makeup quiz, scoring 82. Then there would be two students who scored 82, and two students who scored 79. In this case, the data has two modes: 79 and 82. A dataset can have any number of modes---or even no mode!

## Use Measures of Central Tendency with Crowdfunding Data

Let's see how measures of central tendency work in practice. We'll consider Kickstarter campaigns for plays in the U.S. and compare the statistics for the campaigns that succeeded versus those that failed. Follow these steps:

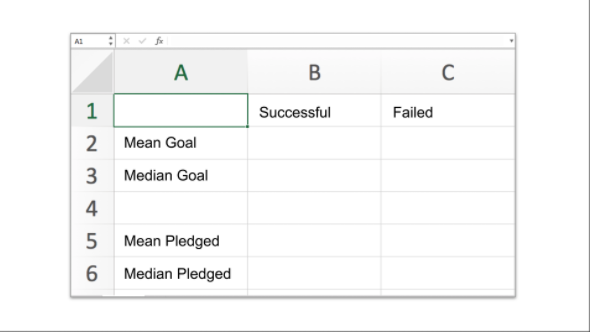
1. Clear any filters on the dataset.
2. For some versions of Excel, the entire header row must be selected first so multiple filters can be applied. Depending on your version, you may need to select the entire row, or only the header you want to filter. Then, apply the following filters:
   * Filter on subcategory for "plays."
   * Filter on country for "US"
   * Filter on outcome for "successful."
3. Copy the filtered dataset and paste it into a new worksheet named "Successful US Kickstarters."

****

1. Return to the Kickstarter worksheet and change the filter on outcome to "failed."
2. Copy and paste this dataset into a new worksheet and name it "Failed US Kickstarters."
3. Create another worksheet and name it "Descriptive Statistics."

We'll be pulling data from each of these new worksheets and performing a few measures of central tendency on them---which is just a fancy way of saying that we'll be finding the mean and median for each dataset's (the failed and successful U.S. campaigns) goal and pledged columns.

In the new worksheet, create a table like the one below to hold our results. This way, we'll be able to easily compare the goals and pledges for failed and successful campaigns alike. By comparing the two, we'll be able to determine whether there are any trends between the goals and pledges in successful or failed campaigns.



### Failed Campaigns

In B2, enter the formula used to find the mean of a dataset:

=AVERAGE('Successful US Kickstarters'!D:D)

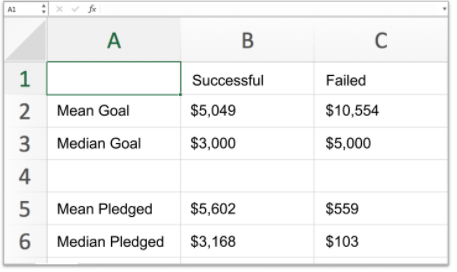
**IMPORTANT**

Excel doesn't have a MEAN function; it uses the less precise AVERAGE function to calculate the mean. Statisticians use the term "average" in many contexts, but prefer to be precise in their calculations. You can tell when a statistician is using Excel by the grumbling noise they make when they have to type AVERAGE instead of MEAN.

This formula tells Excel that we're looking for the average number in a dataset, but we're only looking for the average amount of "Successful U.S. Kickstarters." By adding (D:D), we're also pinpointing which column we're applying the formula to the Goals column. The colon indicates a range of data, so by adding D:D to the formula, we're specifying the entire column.

Let's add a few more to our new worksheet.

* In C2, enter the formula =AVERAGE('Failed US Kickstarters'!D:D).
* In B3, enter the formula =MEDIAN('Successful US Kickstarters'!D:D).
* In C3, enter the formula =MEDIAN('Failed US Kickstarters'!D:D).



This simple table allows us to determine a few things. For one, failed Kickstarter campaigns have much higher fundraising goals than successful Kickstarter campaigns. Louise is asking for more than twice the average successful Kickstarter goal, so this isn't great news for her campaign. In addition, the mean and median pledged amounts are much lower than the successful pledges, which indicates that failed Kickstarter campaigns are unsuccessful for reasons other than asking for too much money. In other words, if the failed projects were also getting a median pledge amount of around $3,000, it's possible that those that failed just asked for too high of a price. Since the median is much lower, there must be another factor keeping people from pledging to those unsuccessful projects.

Measures of central tendency distill a lot of information about the distribution of a dataset down to one number. However, two datasets can have the same mean or median but still look very different---that is, the spread of data between the two datasets can vary quite a bit. When considering the distribution of a dataset, we also want to have measures of its spread. **Measures of spread** include range, variance, standard deviation, and quartiles.

## Range

The simplest measure of spread is the range of a dataset. The **range** is the difference between the maximum value of the dataset and the minimum value of the dataset. For our purposes, the range does not capture as much information as we'd like. What we would really like to know is roughly how far each data point is from the center, or mean, or how much of the data is near the center.

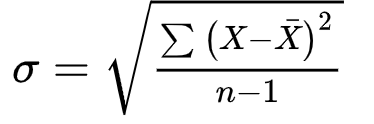
## Variance

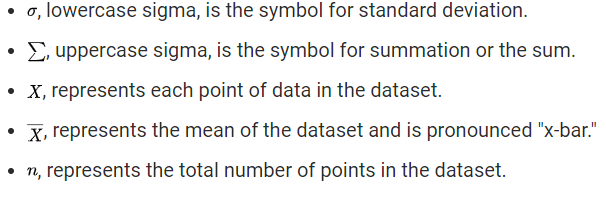
**Variance** is a measure of how far data points are from the center, or mean. To calculate the variance, do the following:

1. Subtract the mean from each data point.
2. Square the difference so that it's positive.
3. Take the average of those squared differences.

Because we've taken the average of the squared differences, the unit of variance doesn't quite match our dataset. To get the unit to match, we take the square root of the variance to standardize it, or get the **standard deviation.** Standard deviation is often represented with a lowercase sigma (σ). You'll also see variance represented as the standard deviation squared (σ^2^). Let's look at an example.

Imagine we have a dataset of five backers. We'll signify that this is a dataset by placing the data within brackets: [1, 3, 6, 7, 8], which then makes it into a "set" of numbers. How do we find the standard deviation? Let's begin working through the standard deviation equation:





This equation looks fairly complicated, so let's talk through what's happening.

1. Find the mean.
2. For each number in the dataset, subtract the mean and square the result.
3. Find the mean of these new numbers.
4. Take the square root of the mean.

Let's apply the equation to our set:

1. Find the mean: (1 + 3 + 6 + 7 + 8) / 5 (the sum the values divided by the number of values), or 25/5 = 5.
2. Next, we find the deviations. The deviations from the sample mean are
   * (1 -- 5) = --4
   * (3 -- 5) = --2
   * (6 -- 5) = 1
   * (7 -- 5) = 2
   * (8 -- 5) = 3

Ideally, we'd like to know the deviations from the actual population mean, but because we don't know the actual population mean, these deviations have a subtle and slight bias to them. We'll correct that bias in the next step.

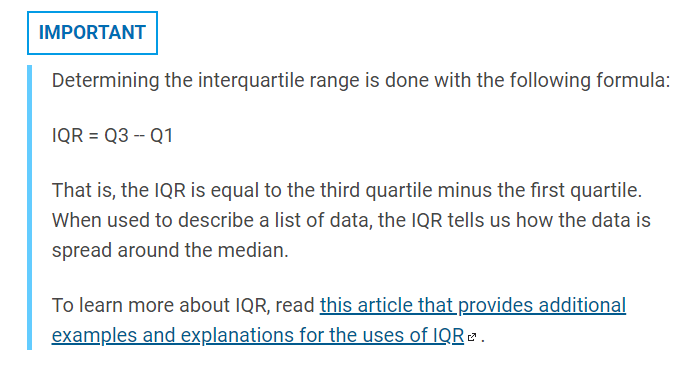
1. Find the variance. First, square all of those deviations; this way, we will always be working with positive numbers. This results in 16, 4, 1, 4, and 9. If we take the average of these values, we'll get a slightly smaller variance than the actual population. To correct for this bias, we instead divide by the number of samples minus 1 (if you're curious, this is known as Bessel's correction). Thus, the unbiased average of these values is (16 + 4 + 1 + 4 + 9) / (5-1)) = 34 / 4 =8.5.
2. The square root of 8.5 is about 2.92, which is the standard deviation.

Another method for measuring the spread is to calculate the interquartile range. After organizing the dataset from lowest value to highest value, we can break it into four separate parts known as quartiles.

## Quartiles

Like medians, **quartiles** are percentiles. The lower quartile is the 25th percentile, that is, 25% of the data is less than the lower quartile. Similarly, the upper quartile is the 75th percentile, so 75% of the data is less than the upper quartile. You may also see these referred to as the 1st and 3rd quartiles. (The 2nd quartile is the median, so that one already has its own fancy name).

The difference between the upper and lower quartiles is known as the **interquartile range** (IQR). The IQR gives us a sense of how far out you can go from the mean to get 50% of the data.



To see these concepts in action, let's add measures of spread to our "Descriptive Statistics" worksheet. Add new rows for the standard deviation, upper and lower quartiles, and IQR.

Here's how your updated table should look:



The function to calculate the standard deviation of a population in Excel is **STDEV.P**. (The other option is **STDEV.S**, which calculates the standard deviation based on a sample of the whole population. There's a subtle difference between these formulas (one is for the entire population of a dataset while the other is for a sample of the whole) that statisticians care about, but we're going to ignore it. Don't tell any of your statistician friends.

We'll be using the same range and worksheet data as we did with the AVERAGE formula, so the STDEV.P formula we enter into B4 is =STDEV.P('Successful US Kickstarters'!D:D).

To calculate the upper and lower quartiles, use the **QUARTILE.EXC** function. QUARTILE.EXC takes two arguments: the first argument is the data array, and the second is the quartile to be calculated. For the upper quartile, put 3:

=QUARTILE.EXC('Successful US Kickstarters'!D:D, 3)

For the lower quartile, put 1:

=QUARTILE.EXC('Successful US Kickstarters'!D:D, 1)

The IQR cell will be the difference between the upper and lower quartiles, so the two cells would be subtracted. In B7, type =B5-B6.

In datasets, **outliers** are extreme points of data; they can be much larger than the rest of the data or much smaller. But how do we define "extreme"? We can use the tools we've just learned along with some guidelines generally accepted by statisticians. There are two main techniques for determining outliers, and each technique uses a measure of central tendency and a measure of spread. We can use either the mean and standard deviation together, or the median and interquartile range (IQR) together.

**NOTE**

Why don't we use variance to determine outliers? We use standard deviation because taking the square root of the variance standardizes the "units" of the variance to match the "units" of the dataset. (This is also why it's called "standard" deviation.)

If we decide to use the first method---mean and standard deviation---the guideline is that any value that is more than 3 standard deviations higher or lower than the mean is considered an outlier. If we decide to use the second method---median and IQR---two guidelines need to be followed:

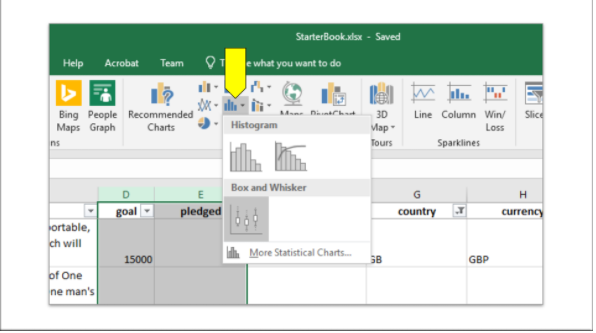
* Any value greater than the upper quartile plus 1.5 x IQR is considered an outlier.
* Any value less than the lower quartile minus 1.5 x IQR is considered an outlier.
* Which method do we use? In almost all cases, the IQR rule is preferred. Medians and quartiles are **robust** **statistics,** which means that they are less sensitive to outliers.
* Consider a county with a small population of people making a modest living. Now, imagine a billionaire moves into the county. The median income would barely change, if at all, but the mean would catapult to a much higher value. In fact, if the county is small enough, everyone but the billionaire could end up being "below average" based on the mean.
* So if the IQR rule is preferred, why is there a method that uses mean and standard deviation to determine outliers? For one thing, mean and standard deviation can be calculated more quickly. Finding percentiles requires sorting the data, which can be time-consuming with large datasets. The mean and standard deviation can be calculated without sorting data, which means that our computers won't need to work as hard to perform the calculations.
* Now that we can identify outliers, what do we do with them? This is a tricky question. Changing or removing data points changes the story you're trying to tell with your data. If the identified outliers are a mistake (say, the data was entered with a typo), ideally, we would just want to correct the mistake and leave the data point in the dataset; if that's not possible, we would have to throw out the data point. However, if an outlier is a legitimate member of the dataset, it's better to leave it in and tell the full story of the data.
* Taken together, measures of central tendency, measures of spread, and identification of outliers tell us about the distribution of our data, but showing the distribution makes the data story easier to understand---which has an even greater impact. **Box plots** are an effective way to show large amounts of information about a distribution in a small amount of space.

## Box Plots

We'll use **box plots**, also called **box and whisker plots**, to compare the distribution of campaign goals and the distribution of total amounts pledged for plays in Great Britain. Remember that Louise estimates she'll need to raise £4,000 for her future project.

First, clear all the filters from your data. Set a filter on "country" to only show rows with GB as the country. Then, set a filter on "Subcategory" to only show rows with "musical" as the subcategory.

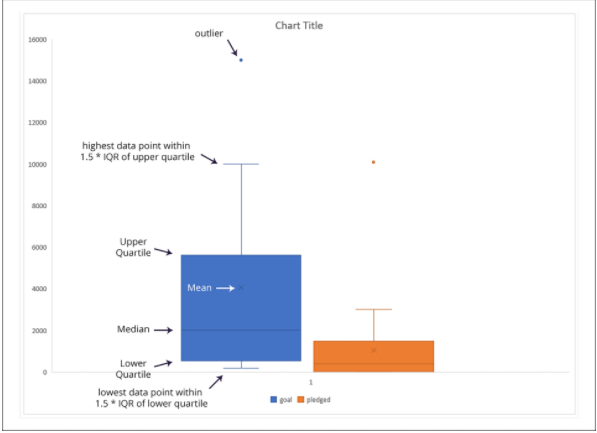
Select the Goal and Pledged columns and then go to the Insert tab. Click the Statistical Charts button and select "Box and Whisker."



Now move the chart to its own sheet by going to the Design tab and selecting the Move Chart button.



How do we read a box and whisker plot? The box shows the interquartile range with a line for the median and an "X" to indicate the mean. The whiskers show the extreme values within 1.5 times the interquartile range. Outliers are represented by labeled dots.



From these plots, we can see that the mean campaign goal is around £4,000. This is outside of the range of outliers for amount pledged, so Louise should probably try to get her play produced for less than £4,000. Half of the campaign goals are less than £2,000, which is just over the 3rd quartile for amounts pledged.