Instructions:

**Conditional Formatting**

Highlight failed outcomes in red

canceled outcomes in yellow.

Live blue

Successful green

**Percentage Funded**

=ROUND(E2/D2\*100,0) – I would do differently.

**Value Shading**

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.

Highlight the column again and click Conditional Formatting followed by Manage Rules. Edit the current rule and adjust the Maximum Type to Percentile, then manually enter 90 as the value.

**Average Donations**

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.

**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.

**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."

 name the new columns "Parent category" and "Subcategory" respectively. Let's use this new data to create a pivot table.

**Pivot Tables and Charts**

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

Pivot Table Analysis

Insert/Pivot Table

Outcomes to colums values

categories

Country to filter

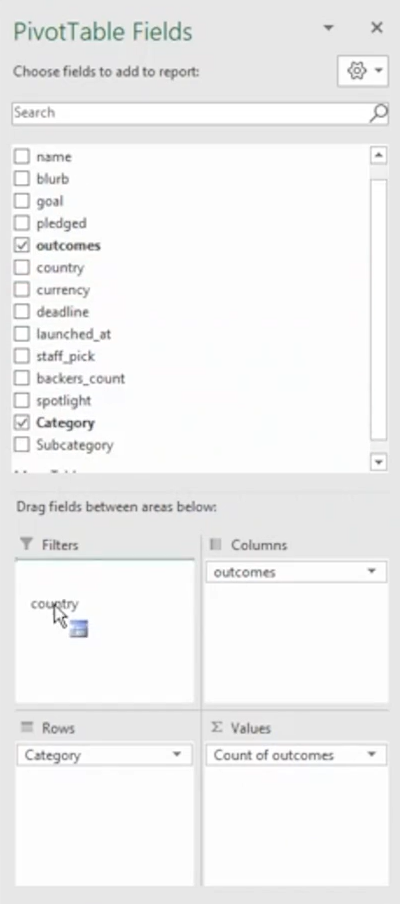
Experiment with the pivot table you just created by switching the **outcomes** to rows and the **parent category** to columns.

Switch back

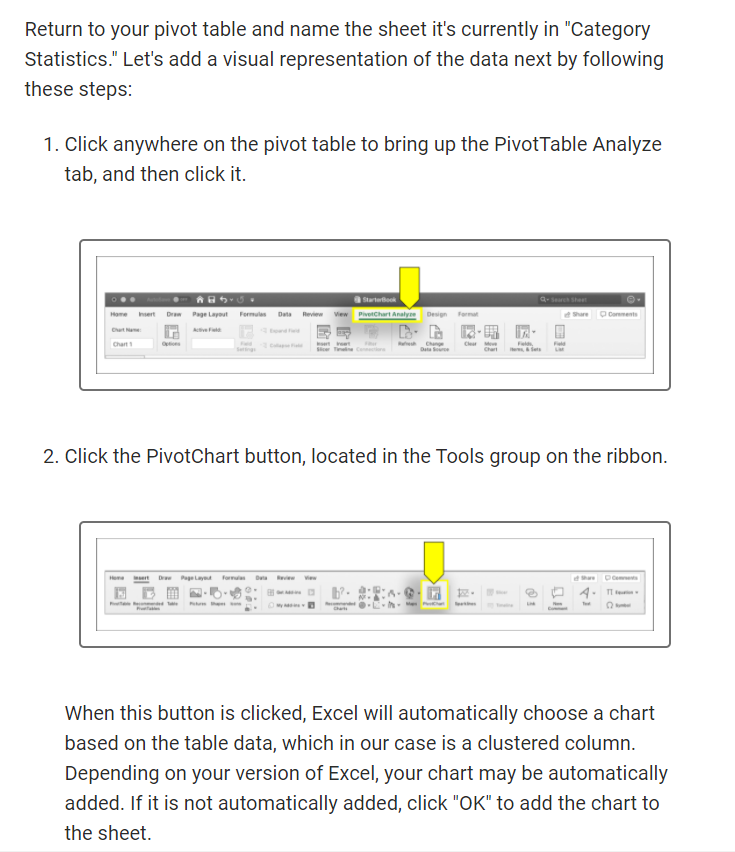
If we filter for only the United States campaigns, we will find that there were 525 successful theater Kickstarters.

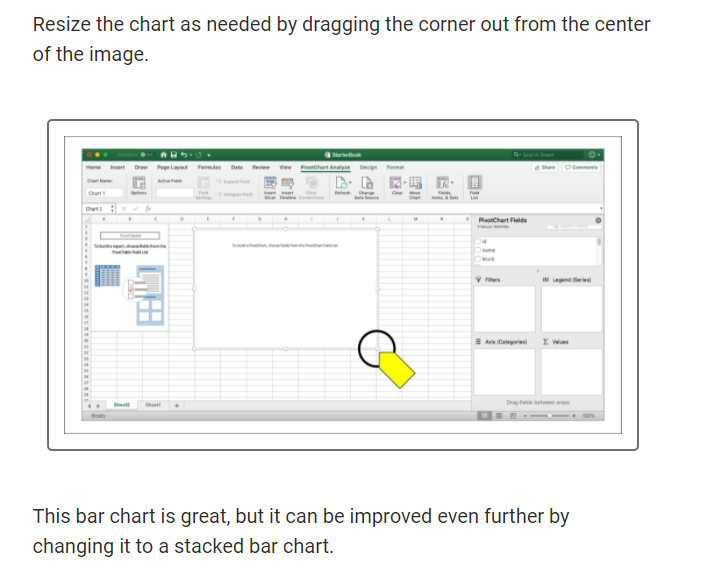
Name this sheet **Category Statistics**

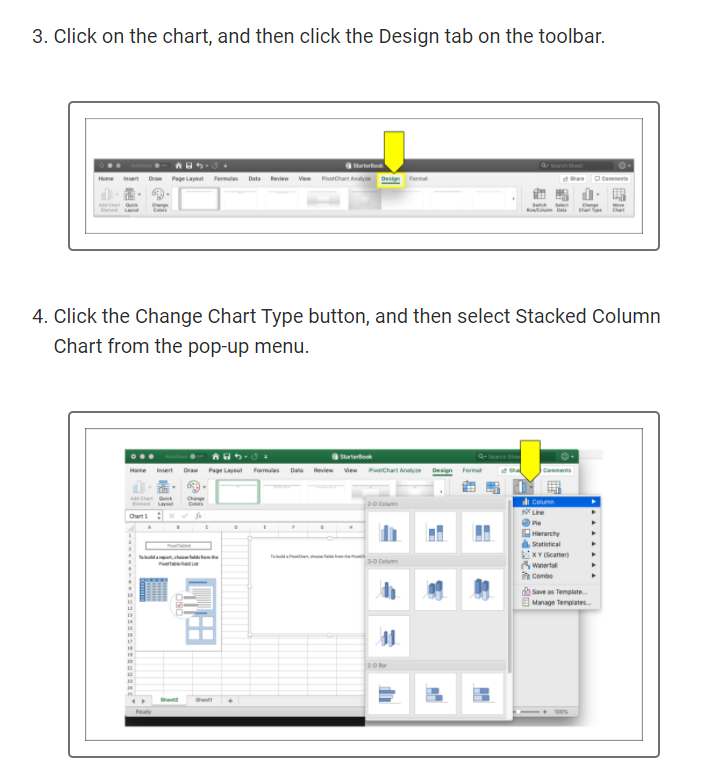
**Resulting PivotTable Fields**

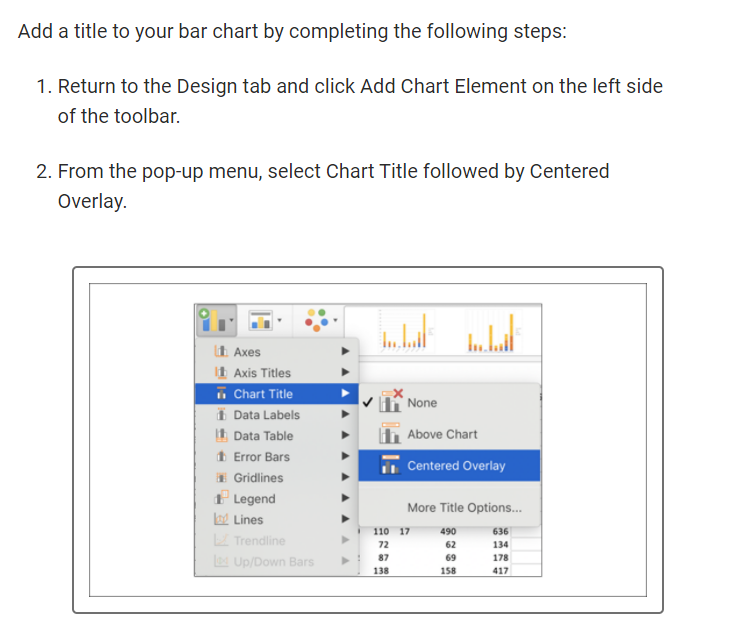
****

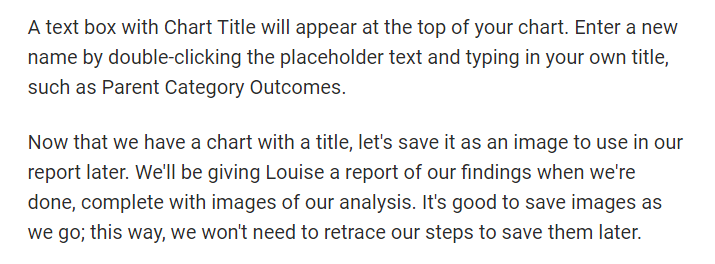
**Pivot charts** complement pivot tables by using table data to create visualizations such as bar charts and line graphs. Visualizations aid in uncovering a link or trend, especially when a table isn't able to.

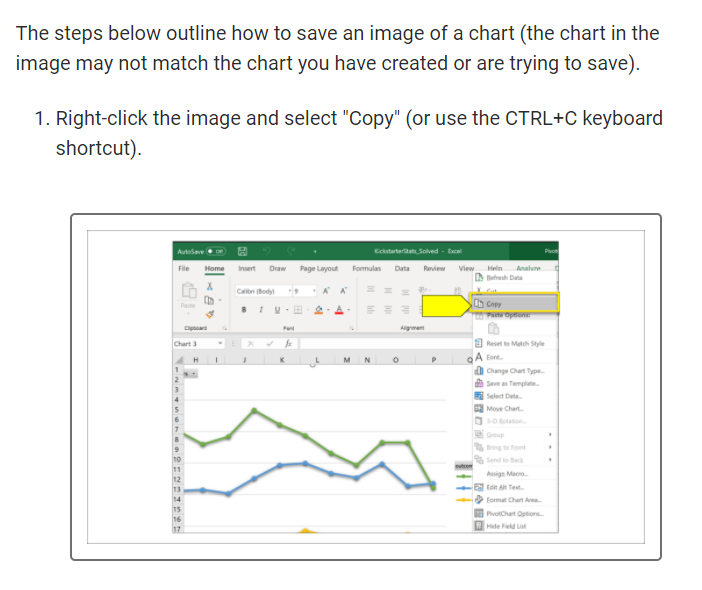


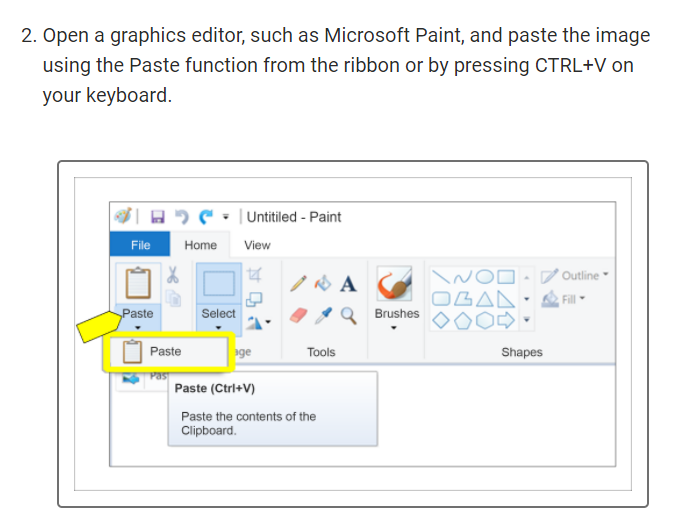


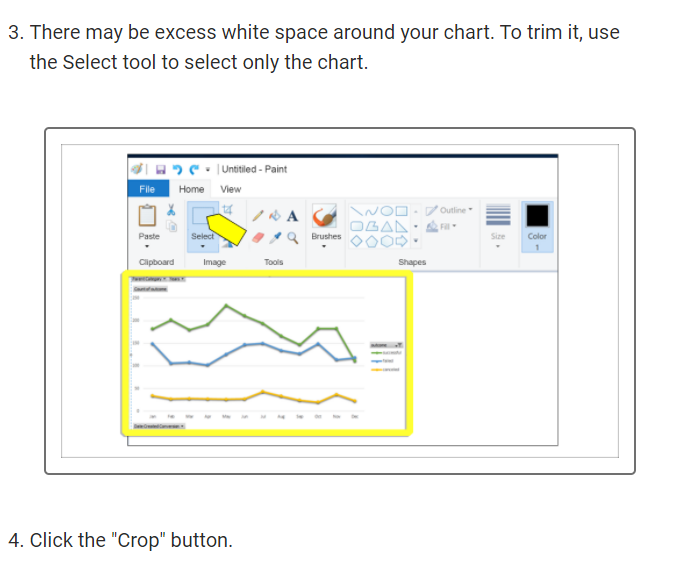


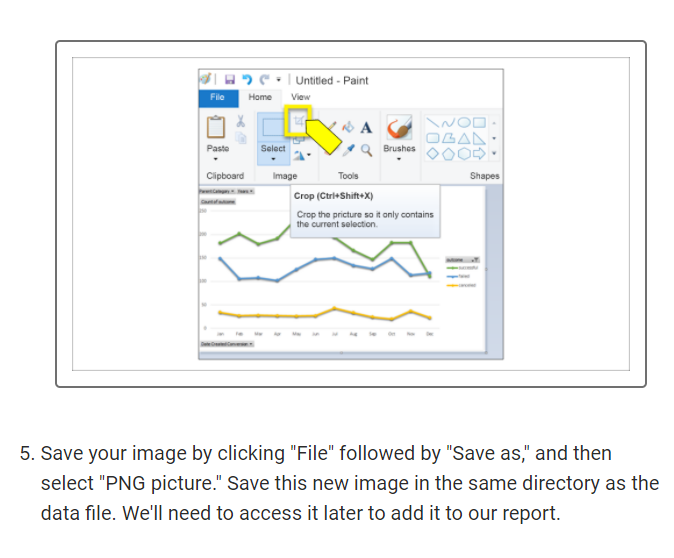


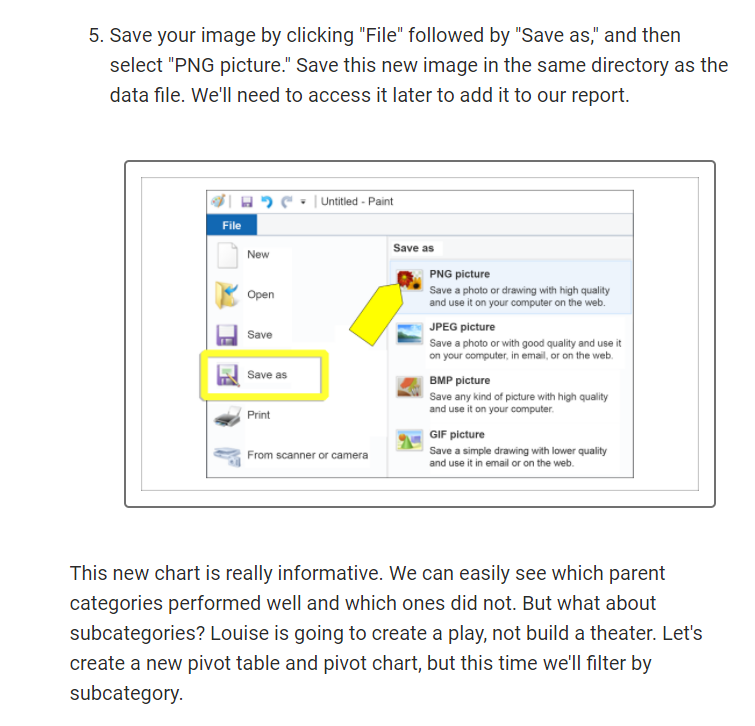












## Chart the Subcategories

Return to the original Kickstarter spreadsheet to get started, and then follow these steps:

1. Click the Insert tab.
2. Click the PivotTable 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 "US" 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 "US", 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

They contain Unix timestamps, which measure time as the number of seconds since midnight of January 1, 1970.

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)

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.

**HINT**

If the dates look a little off, navigate to the Home tab of the Excel sheet, highlight the column of dates, and make sure the format is set as "Short Date."

If you're running a version of Excel that does not have a "Short Date" option, then select "Date" instead.

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.

**Creating Line Chart**

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."

**HINT**

In the Rows field, you may notice there are three items instead of a single item: Years, Quarters, and Date Created Conversion. This is a conversion automatically completed by Excel; because we want our table to only display the months, you can drag and drop the other two items (Years and Quarters) to remove them from the Pivot Table Fields.

For example, select the Years item by left clicking on it and without letting go, drag the item to the spreadsheet and release your mouse button. This will remove the unnecessary value from the field. Repeat the process as often as needed so only the Date Created Conversion item is in the Rows field.

**NOTE**

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

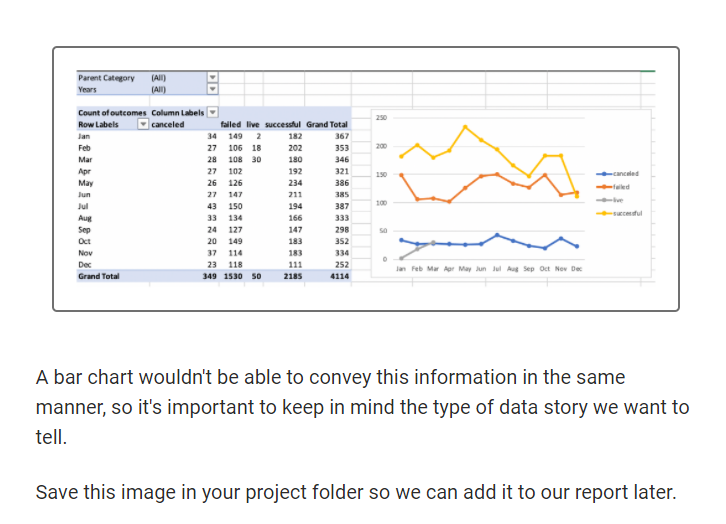
The steps for setting up a pivot line chart are similar to those followed to create stacked bar charts:

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.



**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.

**Use the Search Function**

Search for the play Walken on Sunshine and note the status of the campaign, average donation, duration, and backers.

**Vlookup**

To use VLOOKUP, first get all of the titles in a new sheet, which we'll name "Edinburgh Research."

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)

Copy the formula into B3 through B6 to get the blurbs for all five plays.

play's goal was and the amount pledged

**Mean, Median, and Mode**

Make a small dataset (fewer than 10 data points) with a median of 50 and a mean of 55.

**Use Measures of Central Tendency with Crowdfunding Data**

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

### Failed Campaigns

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

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

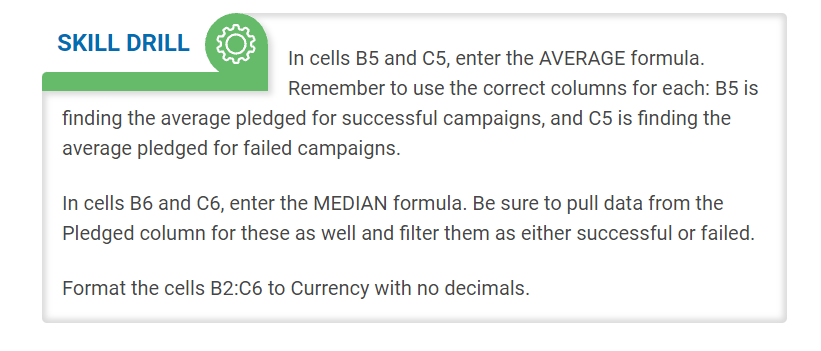
**IMPORTANT**

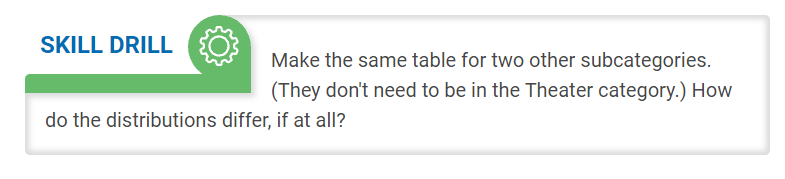
Excel doesn't have a MEAN function; it uses the less precise AVERAGE function to calculate the mean. Statisticians use the term "average" in many contexts, but prefer to be precise in their calculations. You can tell when a statistician is using Excel by the grumbling noise they make when they have to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

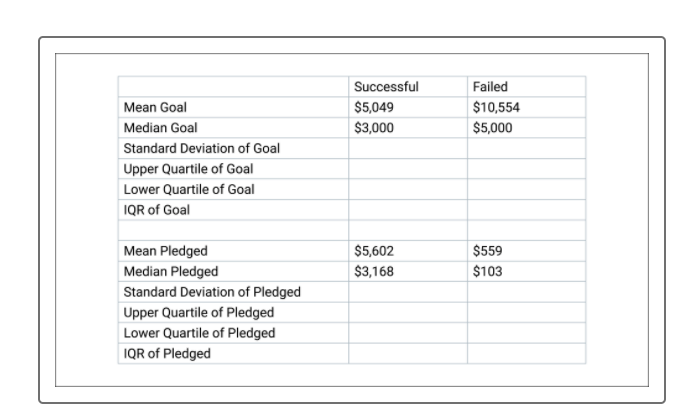
* In C2, enter the formula =AVERAGE('Failed US Kickstarters'!D:D).
* In B3, enter the formula =MEDIAN('Successful US Kickstarters'!D:D).
* In C3, enter the formula =MEDIAN('Failed US Kickstarters'!D:D).





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)

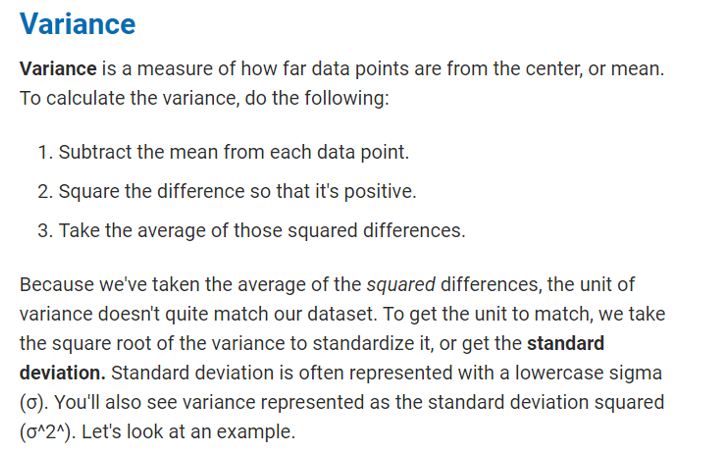
The IQR cell will be the difference between the upper and lower quartiles, so the two cells would be subtracted. In B7, type =B5-B6.

**Outliers**

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.

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.)

**Definition of VARIANCE is : how far data points are from the mean (average)**

****

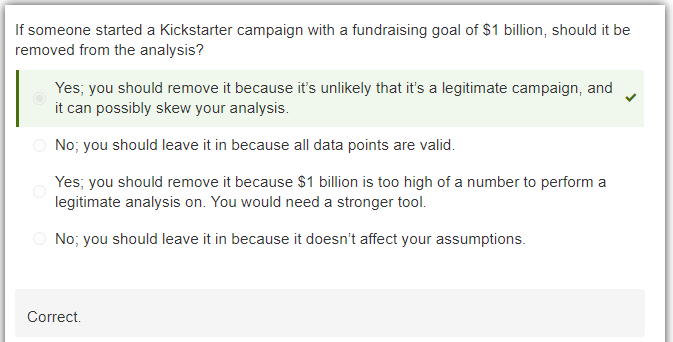
If we decide to use the first method---mean and standard deviation---the guideline is that any value that is more than 3 standard deviations higher or lower than the mean is considered an outlier. If we decide to use the second method---median and IQR---two guidelines need to be followed:

* 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.

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.



**Box Plots**

