Question for assignment:

* Given the provided data, what are three conclusions that we can draw about crowdfunding campaigns?
* What are some limitations of this dataset?
* What are some other possible tables and/or graphs that we could create, and what additional value would they provide?

Requirements for Grading:

* Presents a cohesive written analysis that:
  + Draws three conclusions from the data (10 points)
  + States limitations of the dataset and suggestions for additional tables of graph (10 points)

**Written Report:**

Based on the data provided, we can conclude the following:

1. Crowdfunding campaigns with “theater” as their parent category are most likely to have funds donated to them (34.4%), and those with “plays” as their sub-category are most likely to have funds donated to them (34.4%). Upon further review, those are the same campaigns, as the parent category has only one sub-category, which is “plays.”
2. While the “theater” parent category has the most (by count) of successful campaigns (187 of 344), the “journalism” parent category has the highest rate of successful campaigns (4 of 4); granted, this is too low of a sample size to be significant, but interesting, nonetheless.
3. Campaigns started in June or July are more likely to be successful (63.2%), with 11.3% of the total campaigns (113 campaigns). Looking at the number of successful campaigns for each month, September also has a high percentage of successful campaigns, with 61.6% of the campaigns being successful. This is because there are no live campaigns and a much lower number of failed campaigns for those started in September.
4. (Bonus): By looking at the number of backers, it appears that the more backers a campaign has, the more likely it is to be successful.

There are some limitations of this dataset.

1. With only being given the total amount donated, we can’t see which campaigns had large individual donations, therefore helping them become successful in a shorter amount of time. Also, having the individual donations would mean we can see where the money is coming from, both the person and the location to see if there are patterns there.
2. The dataset is rather small. Being only a thousand rows long, with 11 years of data, there is an average of ~7.5 campaigns per month each year, with some months having 4 or 5 campaigns. If there were more campaigns overall, then any analysis could help predict future campaigns.
3. I’m not confident, but since there is a column called “currency,” one might assume that the monetary columns are in that currency and not normed to US currency (or another currency). If that is the case, then our initial analysis and any analysis done without norming the currency would be inaccurate.

Here are some possible tables we could include:

1. I would want global economic data to be able to compare the donation rate to economic conditions at the same time. We could use that information to help predict donation rates in the future based on similar economic conditions in the past.
2. I would want the individual donations table to analyze what kinds of campaigns are getting large donations as well as to see if campaigns are more likely to receive donations near the beginning, middle, or end of the campaign duration. This could be useful to know when more marketing could be done by the campaign team to help increase donations.

Graphs I would include:

1. One graph that could help with an analysis would be the outcome vs. country or currency. Using this, we could determine if the country or currency influences the outcome of the campaign.
2. I would also investigate the staff\_pic and spotlight columns vs. the outcome to see if those had any type of relationship for the success of a campaign.

**Statistical Analysis of Failed vs. Outcome Summary Statistics:**

For both the successful and failed outcomes vs. backers\_count, the median is a better measure because the data is right skewed. As you can see in the graphs I created in the “Statistical Analysis” tab in the workbook, most of the data lives in the first bucket on the left with many outliers (~8% of the data in both cases are outliers). As these outliers are very large compared to the rest of the data, the mean is larger than the median, and therefore less representative of the data.

There is more variability within the successful campaign data, because the standard deviation, variance, range, and interquartile range are all higher than their counterparts within the failed campaign data. This makes sense as for a campaign to fail, the percentage funded must be somewhere between 0% and 100%, inclusive. However, a campaign that succeeds is one that is funded at 100% or more, and there are many highly funded. For example, there are 24 campaigns that were funded at 1000% percent or more within the data, with another 66 campaigns funded between 500% and 1000%.