Module 1 Challenge

1. Given the provided data, what are three conclusions that we can draw about crowdfunding campaigns?
   * Out of 1,000 sample projects provided in this module, 5.7% of the projects were canceled, 36.4% of the projects failed, 1.4% of the projects are live, and 56.5% of the projects were successful. Therefore, based on this sample, we can conclude that more than half of the projects were successful.
   * When 1,000 sample projects were categorized, top 3 categories in count were theater with 344, film & video with 178, and music with 175. Therefore, based on this sample, if one tries to receive funding from one of these three categories, there may be more competition.
   * Out of the 344 sample projects categorized under theater, 187 were categorized once again under the plays sub-category. Therefore, we could conclude that more than half of the projects under the category of theater are categorized under the plays sub-category. Therefore, if one tries to receive funding from theater/plays category, it has the highest number of competitors.
2. What are some limitations of this dataset?
   * The sample is limited to 1,000. I’m not sure of the source of these samples, but if it may be limited to Kickstarter and Indiegogo platforms. Since there are more than 2 crowdfunding platforms online, this dataset is limited.
3. What are some other possible tables and/or graphs that we could create, and what additional value would they provide?
   * There are two columns titled “staff\_pick” and “spotlight.” We could create a table/graph to check how “staff\_pick” and “spotlight” value relates to “outcome” and “Percent Funded” column. From this table/graph, we may be able to find out if being a staff picked project or spotlight project affects the outcome.
   * “country” and “category & sub-category” column could be useful for another table/graph because once created, we could check the preference of category for each country. From this table/graph, one could decide which country he/she needs to launch his/her project from.

Module 1 Challenge: Bonus Statistical Analysis

1. Use your data to determine whether the mean or the median better summarizes the data.
   * I think median would be a better choice to summarize the data because upon checking Mean + (3 \* Standard Deviation), it seems like there are outliers to skew the normal curve. For example, Successful Campaigns had Mean of 851.15 (Rounded to hundredths). I added 3 multiplied by Standard Deviation of 1267.37 (Rounded to hundredths). The answer I got from this calculation was 4653.26. However, the maximum backer\_count for the successful campaigns was 7295 which shows that it is an outlier. Since there is an outlier to this data, I would prefer to use median.
2. Use your data to determine if there is more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?
   * The sample variance of successful campaigns was 1,606,216.59, and the sample variance of failed campaigns was 924,113.46. Also, the standard deviation of successful campaigns was 1,267.37, and the standard deviation of failed campaigns was 961.31. From the value of sample variance and standard deviation, I can assume that there is more variability in successful campaigns. This makes sense because higher value in simple variance means lower consistency. Low consistency means that it is more difficult to predict.