Bertrand Louis  
Rutgers BootCamp: Data Science  
March 2021

Kickstarter Homework – Data Analysis

1. Given the provided data, what are three conclusions we can draw about Kickstarter campaigns?
   1. Looking at the projects launched in Kickstarter, the data in the pivot table in sheet labeled “Categories” confirmed that none of the Journalism projects launched on the platform were successful. All of them got canceled. Therefore, Kickstarter is not an ideal platform to crow fund journalism projects.
   2. An important conclusion that I can deduct from this data is that all Kickstarter campaigns under the category of “Films & Video” with a Sub-category labeled “Documentary” were all successfully funded.
   3. Music and Theater as categories had the highest success rate of being funded in all project analyzed. They were had a 77% and 60% success rate. With a 17% success rate, Food as a category had the lowest rate of being successfully funded in all categories.
2. What are some limitations of this dataset?
   1. The dataset does not give a lot of information on the motivation of the donors and their incentive to fund a project. Therefore, it is hard to assess why donors support a project since the data on donor satisfaction is not available.
   2. Looking at the number of projects launched on a monthly basis, about half of them failed every month. We do not have any insights about the causes of their failures.
3. What are some other possible tables and/or graphs that we could create?
4. There some many other tables and/or graphs that could be created to get more insights on the data. For instance, I would create a pie chart to visualize which project category got the biggest slice or was the most successful among all of the projects on the Kickstarter platform.
5. I could also use a scatter plot to visually illustrate the relationships between two variables. For instance, I could use the scatter plot to see subcategory projects that were canceled on the Kickstarted platform.

**Bonus Questions**

***Use your data to determine whether the mean or the median summarizes the data more meaningfully.***

For this dataset I would rate use the median instead of the mean (shown in the excel sheet labeled Bonus 2) because the outlier 20,242 and 26,457. The outliers are skewing or pulling the data to right and causing the mean to be much higher that the average relevant number of backers support a project. The median gives the more typical value in this dataset.

***Use your data to determine if there is more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?***The range is the simplest and the most straightforward measure of variability to calculate. With that being said, there is more variability in the successful campaigns compare to the unsuccessful campaigns. If I look at the range of the successful campaigns, it much wider with 26,456 than the unsuccessful campaigns with a range of 1289.   
  
Also when I calculate the coefficient of variations (Standard deviation divide by mean) of the successful vs the unsuccessful campaign, the unsuccessful campaign has a coefficient of variation of 3.4 whereas the successful campaign has a higher coefficient of variation of 4.3. The coefficient of variation measures the dispersion of data around the mean. Therefore, the successful campaigns have more variations. One of the reasons why there is more variability in the successful campaigns is because of the outliers.