Robert Austin Peron

Rutgers Data Bootcamp

Homework #1

June 13, 2020

1. Given the provided data, what are three conclusions we can draw about Kickstarter campaigns

After summarizing the data between subcategory and project state, as well as the project’s created dates, here are 3 conclusions we can draw about Kickstarter campaigns:

1. In the given data, I found that out of the 4,114 campaigns we analyzed, currently we show 53% of all campaigns analyzed ended meeting their funding goal and were “successful”, 37% of campaigns did not meet their funding goal and were “failed”, while 8% of campaigns were “cancelled” and 1% of campaigns are still “live” with no outcome. We can determine from this summary that 1 out of every 2 kickstarter campaigns have been successful in the aggregate.
2. From the subcategory data and column, I found the sub-category with the highest total count of related kickstarter campaigns was for “Plays” which had a total count of 1,066 campaigns. The sub-categories with the 2nd and 3rd highest kickstarter campaigns started are “Rock” and “Wearables”. We can conclude from this data, that the artists and communities involved in “Plays” and “Rock””, as well as independent tech designers of “Wearables” are the most frequent sub-categories to seek funding through Kickstarter campaigns.
3. From the “CreatedDate” data and line chart, I found that with the data provided based on the date the campaign was started, the highest count of projects created that were “successful” were created in May, and the month with the highest count of projects created that were “failed” or that were “cancelled” were created in July.
4. What are some limitations of this dataset?

There are a couple of limitations in the dataset that was provided regarding Kickstarter campaigns. The first of the limitations is that there are no data points that provide us with a location or associated address with the kickstarter campaigns that are listed in the data. With this information, we could analyze more deeply into the successful/failure campaigns by identifying the creation location and being able to see if there are any significant trends regarding any kickstarter campaigns geographical location.

An additional limitation I noticed is that the data we have on kickstarter campaigns ends in March 2017. With this limited data set of roughly 8 years, we are missing the remaining data in 2017, as well as missing outcomes campaigns that are classified as “live”. However, understanding that we only have a limit to what data we could access, sort of limitation is to be expected.

A final limitation I noticed is that there is no data point that determines who, or what entity created the campaign. There could be possible differences between projects that were created by a solo artist compared to a theater company, or local community organization. Campaigns outcomes can be influenced by the exposure that larger sized groups or organizations have with their outreach and campaign awareness, against solo artists or tech designers that are independent and have fewer connections. I think understanding the type of creator of a campaign could help us understand further on what influences the outcomes of these campaigns.

1. What are some other possible tables and/or graphs that we could create?

For the data we analyzed, there are a couple of other tables and graphs that we could have created to better understand the dataset. My first recommendation would be creating a new column in the data set titled “Campaign Duration”, where we find the length of time a campaign was live until it reached its outcome. From there, we could build a table that could breakdown the 3 main outcomes (cancelled, failed, successful) and analyze the average time of each of these outcomes to determine whether the length of time a campaign continues has any influence on its outcome.

Another graph I would create would involve the average donation analyzed by each parent category or sub-category. We could use the given information on average donations we calculated to analyze which categories/sub-categories communities have the highest average donation counts, as well as the average duration of the campaigns in those category/sub-categories. This could present us clear understanding of what communities are more likely to respond to a kickstarter campaign fund, and what you could expect the average donation could be given a certain duration of time. This could give us information to plan future campaigns with average donations, expected length of campaigns broken down by that projects cat/sub-cat.

## Bonus

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

Based on the analysis of the mean and median for the count of backers for successful and failed campaigns, the mean seems to be the figure that is more relevant for this data set. The mean includes the campaigns backer counts that had absolutely no support, which the median measure does not capture. I used the mode function to measure the mode for each status, and I found that for failed campaigns, their mode was 0. I believe using the mean for a comparative analysis is crucial to understanding what requirements a future campaign may need in terms of outreach to online audiences to meet their funding goals.

Use your data to determine if there is more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?

Based on the analysis of the variance and standard deviation of the count of backers for failed and successful kickstarter campaigns, the data shows there is greater variability in the count of backers for successful campaigns than failed campaigns. To start, the variance measure for successful campaigns was 712,841, compared to failed campaigns at 3,773. This number alone implies that successful campaigns have greater variability. The standard deviation for successful campaigns is 844, compared to failed campaigns with a standard deviation of 61. This means that successful campaigns have a larger distance away from the mean for each standard deviation in a normal distribution when compared to failed campaigns. Overall, both measures signal to us that the variability of backers for successful campaigns is greater than failed campaigns.