Brian Fitzpatrick

Module 1 Challenge

***Written Report***

**“Given the provided data, what are three conclusions that we can draw about crowdfunding campaigns?”**

The data allow you to draw several conclusions, the first one being most projects were successful. 565 out of 1000 projects were funded, or 565 out of 986 if the projects that are still “live” are excluded given we do not know their outcome. This also tells you that although most projects are successful, success is not guaranteed. 364 projects failed to meet their goal and 57 were cancelled, thus it is clearly necessary to have a project and a pitch that is appealing to potential donors, people are not just giving money to every project they see.

Examining the line graph showing outcome by month, success peaks in July, with a high percentage of projects being successful in both June and July. While projects are the least likely to be successful in August. Therefore, if you are looking to launch a project peak Summer is the ideal time to have your project out for funding, but you want to make sure you avoid August.

For the categories of projects that are most successful, journalism projects were the most successful, 100% were funded, although there were only four journalism projects so this conclusion is tentative at best. After journalism projects, technology projects were the most successful at receiving funding, while games projects were the least successful. Therefore, if you are choosing between pursuing a technology project and gaming project should go with the gaming project if you want a higher chance of funding.

**“What are some limitations of this dataset?”**

The dataset has a column for type of currency, but it is not clear from just looking at the dataset if the “goal” and “pledged” columns are all using the same currency or not. Were funds raised in CAD for example but then for the “goal” and “pledged” columns everything was converted to USD? Not knowing this makes comparison between countries difficult.

The dataset only has a N of 1000, large enough to draw broad conclusions, but drawing conclusions based on category and especially subcategory becomes less reliable particularly for the categories that don’t have that many projects, with the Ns for the sub-categories becoming even smaller. A small N means any conclusions drawn from, particularly comparing sub-categories within a category, will not lead to much certainty in your conclusions.

I would want more granular data than what even the sub-category provides. For example, for the “radio & podcast” subcategory, is there a specific type of podcase that tends to receive funding, ex. true crime? News? Etc. Having a dataset that included this additional detail would also be useful, assuming the dataset was large enough to include meaningful amounts of cases in these categories.

**“What are some other possible tables and/or graphs that we could create, and what additional value would they provide?”**

Providing a pivot table showing success by country and category of project, and then adding a filter for outcome, could be useful. Maybe certain types of projects do better in certain countries, ex. maybe food projects do well in the US, while music projects do better in Australia.

From the date created and ended columns a new variable for length of time a project is open could be created. Then could create a pivot table by outcome and length of time a project is open. A column pivot chart would then make a nice visual to compare how long projects are open by outcome. Maybe successful projects are simply open longer, or maybe projects that are successful receive their funding in a relatively short period of time compared to failed projects.

There is the “spotlight” variable in the dataset. Assuming this is whether or not a project received a spotlight on the funding site, a pivot table with whether or not a project was spotlighted and outcome. Can use this to answer the question of whether or not a spotlight has an impact on the outcome of a project.

Finally, could do a text analysis of the blurbs in the “blurb” category. Could code for example for more versus less technical language. Then create a pivot table comparing type of language to outcome. Maybe more technical language repels some potential donors?

**Statistical Analysis**

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

In this case, for the count of backers, median is better for summarizing the data. Looking at the standard deviations for both categories and then directly at the data, for both the successful and failed campaigns there are outliers. These outliers are skewing the data set and leading to a mean number of backers for both the successful and failed campaigns that is well above where 50% of the cases are. Thus making the median better for summarizing the data.

Also, by comparing the median and mean, given that the mean is much higher for both the successful and the failed campaigns, you can tell there is a skew to the dataset. This makes the mean less useful for summarizing the data. Assuming what you want to know is approximately how many backers does a campaign receive. The median is much more useful because the outliers have much less influence on it giving you a much more accurate picture of the central tendency of the data.

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

The standard deviation for the number of backers of successful campaigns is about 1.49 times the size of the mean, while the standard deviation for the number of backers of failed campaigns is about 1.64. While overall the standard deviation for successful campaigns is larger than for failed campaigns, thus if you are comparing mean to standard deviation there is more variability for the failed campaigns. This does not seem like a huge amount of variation between the number of backers for successful and unsuccessful campaigns, this said, more variation for the failed campaigns is what I would expect. For the successful campaigns all campaigns must have at least one backer and likely several, while for failed campaigns some campaigns likely have zero backers while there are some with several backers but they just did not reach their goal. It could also be that on average successful campaigns had backers that gave more money per backer, while failed campaigns had backers that gave less per backer requiring more backers to reach their goal. This combined with failed campaigns that received only a few or no backers would create greater variation than the successful campaigns.

All this said the range for the number of backers for successful campaigns is over a thousand larger than the range for the failed campaigns. This could be the results of outliers, but it could also be arguably expected because there are some campaigns asking for a relatively large amount of money that were successful, while there were also campaigns asking for relatively little money that were successful. If you assume all backers give the same amount of money this would mean you would get a greater range for successful versus failed campaigns, as failed campaigns with high goals would not reach their goal.