Grace Palmer

Module 1

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

Three conclusions we can draw are: theatre and plays are the most represented project type by parent category and subcategory respectively; July boasts the most project launch dates but projects launched in June have a higher rate of success; finally, campaign goals and rate of success have a slight inverse correlation.

The pivot charts from sheets ‘Category PivotChart’ and ‘Subcategory PivotChart’, copied below, illustrate the first conclusion. From the data, we can presume that if a campaign is called at random, it is more likely to fall in the theatre & plays subcategory than any other.

The second conclusion can be observed by looking at sheets ‘Launch PivotTable’ and ‘Launch PivotChart’, chart copied below. July has seen a total of 94 launches, 58 of which were successful, whereas June returns 55 successful launches out of a total 87. We can conclude June is statistically the best month for successful launches.

The final conclusion can be reached by looking at sheet ‘Crowdfunding Goal Analysis’, chart below. I have included a trendline to demonstrate how the rate of success slightly decreases overall as the campaign goal increases.

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

One limitation of this dataset is that it is comprised of categories and subcategories that are represented at a significantly unbalanced rate. For example, journalism and theatre are both treated as categories, despite journalism describing 4 launches and theatre describing 344. Such a difference in representation can create issues when looking to make predictive or summative analyses using categories.

Another limitation is the date range of samples. All projects in the dataset have a launch date between 2010 and 2020, meaning no analysis can be performed on recent data.

Also, this dataset is limited by the ‘live’ outcome. It can be inferred from the data that the ‘live’ outcome is a user-reported status for the campaign or an error in the dataset, as multiple live projects have passed the launch deadline. The ‘live’ outcome introduces an aspect of ambiguity to the dataset that must be considered when making analyses.

Finally, the goal for each campaign is not given in a standardized currency. Variations in the value for each currency are not accounted for in the data, meaning that the goal given cannot be considered an empirical standard by which comparisons between campaigns originating in different countries can be made.

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

I’d like to look at how the rate of launches and rate of successful launches have changed overall. With a dataset spanning 10 years, we could draw conclusions about the usage and popularity of the crowdfunding platform on a macroscale. We could also look at the relationship between outcome and status as a staff pick or spotlight. If there is a correlation, we can make conclusions about the value of both functions as it relates to the success of any given campaign. I’d like to create a table describing the relationship between goal amount, % of goal reached, number of backers, and category and subcategory. This analysis could indicate which categories receive the most attention and draw the most enthusiasm. Finally, it would be interesting to see if the average donation correlates to the category and outcome of a launch. If correlation is shown, we can make predictions about donation expectations based on category and outcome expectations based on donation amounts.

**Statistical Analysis**

The mean is a better summary of this dataset because of the distribution of the values in the dataset and the information we are aiming to impart. The median indicates that the majority of values skew close to the minimum of the population, but as a single summary for this dataset it is misleading. Based off the median, we could expect to see successful projects count backers in the low hundreds. The mean, conversely, indicates an overall average in the upper hundreds, which better represents the distribution of backers across successful campaigns.

We find greater variance amongst the successful population than the failed population, which is to be expected given the differences in size between the datasets. As variance represents the difference between each number from each other and the mean, with more data points we would expect to see higher variation.