Crowdfunding Dataset Report

From the data we can conclude several things regarding crowdfunding. For one, crowdfunding projects in the “theater” category were observed to be attempted the most of any other category. Furthermore, there are a greater number of campaigns with a goal less than $10,000 than greater than $10,000. When looking at campaigns with a goal below $10000, campaigns with a goal between $1000 to $4999 have a greater probability of success (and therefore a lower probability of failure) when compared to projects with a goal between $5000 and $9999 as well as below $1000. Finally, we can observe that, generally, the number of campaigns started throughout a year are consistent, however, campaigns started in June and July have a greater amount of success than throughout most of the rest of the year.

There are, however, limitations to the dataset. It is hard to draw fair comparisons between goals with lower monetary value and higher monetary value goals, this is due to the low number of goals above $10,000. Another limitation of the dataset is that there is no way to obtain distributions of donations based on amount or time. While average can be calculated by comparing the amount pledged and the number of backers, this fails to capture variability within campaigns.

Despite the limitations of the dataset, there is still many observations that can be made from graphing or filtering tables. A graph which compares distribution of percent funded across categories can give further insight into the categories which find more success in the dataset. We can also produce a bar chart which compares campaigns which have either been selected as a staff pick or been given a spotlight and compare the achievement of the goal of those campaigns to campaigns which have been given neither a spotlight nor a staff pick. The purpose of this would be to observe some effect of spotlighting and staff pick in the performance of a campaign. This effect of spotlight and staff pick can also be observed by comparing those values with the number of backers via another bar chart, this would give some baseline idea of the influence in these values in helping a campaign reach other people. Another possible plot could be plotting a line chart plotting the percent funded vs the campaign length (by subtracting deadline by start date) to observe the relationship between the two and if there is some threshold of time which provides the greatest opportunity of success for each category.

Justification of Median

Based on the central tendency calculations of successful and unsuccessful crowdfunding campaigns. The median number of backers (Successful = 201, Unsuccessful = 115) would be a better measure to summarize the data then the mean number of backers (Successful = 851, Unsuccessful = 586). This is because the dataset has a large amount of variance in both the successful and unsuccessful groups, which indicates that the actual value in each group varies greatly from the mean; this is trend is also noted by the standard deviation (successful = 1267, unsuccessful = 961), which indicates a similar amount of variance in the data. The mean does not account for the outliers, such as few campaigns with a huge number of backers which would misrepresent the data (which is likely the case as the dataset has a positive skew; mean > median). The median however, helps to somewhat mitigate the outliers by giving each data point a similar “weight” in the dataset (a value greater/less than the median will not shift the median regardless of magnitude.