Kurt Shiple

02/08/19

Homework 1 Report

Homework 1 Report

**What are three conclusions we can make about Kickstarter campaigns given the provided data?**

Given the provided data, the category of music generates the highest proportion of success to canceled and failure. As a matter of fact, the average category has a success ratio of 1.16 with n=4064 while music yields a success ratio of 3.85 with n=680. To the contrary, journalism has the worst results in regard to success with a ratio of 0 and n=24. Theater had the most net success results with a success ratio of 1.58 with n=1369.

An additional insight could made about the provided data by looking at the outcomes by month. Again, averaging the proportion of success to failure and cancelation across the months yields a success ratio of 1.16 with n=4064. There is a four-month period where the success ratios are markedly higher than the rest. February, March, April, and May showed success ratios of 1.53, 1.32,1.48, and 1.53 with n= 334, n=315, n= 315, and n=323. The success ratio falls below average starting in July and the remaining months that follow. The worst success occurs in December with a ratio of 0.78 and n=253.

Other conclusions can be derived by looking at the outcome percentage based on goal requirement. The projects with the highest likelihood of success occur when the goal requirement is less than 1000 USD. The success percentage declines steadily as goal requirement increases with the lowest marks of success culminating with projects requiring more than 5000 USD.

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

We should be careful when making definitive conclusions about what correlates we may observe in regard to success, failure, cancelation and the like. What determines the state of a particular project only has to do with whether or not the goal in regard to donation was met. This is one way to measure success but for fresh eyes just now observing this data set, this could be misleading. It should be stated that measurement of success within this data set does not have anything to do with the accomplishments of an individual project after achieving its desired goal in regard to donation. Such a determination would require different data entirely.

It should also be mentioned that when comes to categorical variables, you are relying on the methodology of categorization of the host website and where certain Kickstarter projects are almost definitively film & video, for example, it is not uncommon for some projects to be interdisciplinary. This is relevant if one is trying to establish causal relationships between variables. Ultimately, one should recognize that categorical variables have a methodology to their organization that should be understood.

There does happen to be data missing as well as non-numeric data where there should be values. The non-numeric data comes as a result of using formulas and various columns of data to create new insights like percent funded. While the formulas used for this insight generate valid data for the most part, occasionally there are cells in which a figure is divided by 0 and thus becomes useless for analytical purposes. One way to combat this is to replace these non-numeric error messages with the average of the rest of the figures.

**What are some other possible tables/graphs that we could create?**

Some other tables, graphs, and methods of analysis could be created to demonstrate other insights. One could measure the count of state in a column graph that is separated by whether or not a particular project was picked by a staff member and this would allow us to see what type of effect that such social proof and marketing may have on the prospects of a given Kickstarter. Of course, one might suggest that something becomes selected by a staff member because it was inherently higher quality anyway. In any case, observing the relationship could be useful. Additionally, one could measure if there is a relationship between average donation and sub-category as well as backer count and sub-category. This could illuminate what the most popular genres are and thus how easy it is to generate interest in a given project. Lastly, I believe a multivariate regression could be done using percent funded as the dependent funded, state separated into dummy variables, and average donation as another dependent variable. This would demonstrate how much of the variance in in percent funded was determined by the selected dependent variables.