Statistical Data Analysis – Capstone 1 – DW

Refer to the Jupyer notebook for visualization and progress.

Based on an extensive examination and visualization of the data, the most important variable in explaining success of Kickstart project is the goal of the projects. The more ambitious the project, the more likely it is to fail. I used the qcut function in pandas to display the categorical result vs. the numerical goal, and the probability of success across the 8 cuts steadily drops from 55% to a 19% change of success. However, success is far from guaranteed for less ambitious projects, making prediction difficult.

The most interesting discovery in exploring the data was noticing that the success rate plummeted between 2013 and 2014. The success rate was over 60% in 2013. In 2014, there were almost twice as many projects with almost a doubled average goal. With more ambition came more failure, and projects succeeded at under 40% in 2014. I have not yet been able to decipher what changed in 2014 through research of Kickstarter. This does make me consider throwing out data prior to 2014 as it may not be as relevant to prediction as the newer data.

I identified two numerical columns that were most useful to prediction: the name length and the blurb length of projects. I performed use bootstrapping to compare the means of the name length of successful projects and the name length of unsuccessful projects and found there was a statistically significant difference. I had similar results looking at the blurb length. The plotted the two numerical variables against each other and found the Pearson's R statistic. The two variables clearly had a very low correlation. Unfortunately, despite the means being significantly different, in practice the amounts were very small and the variables only have small predictive power.

I used a chi-squared test on two categorical columns that I was unsure whether to keep for modeling purposes. I performed a chi-squared test on the day of the week column (and the result column) and the month column (and the result column). Both results were statistically significant, so I decided to keep the variables around for prediction.