Capstone 1 - Milestone Report - DW - 3/7/20

Problem Statement:

Kickstarter is a crowdfunding website with a stated purpose of helping to “bring creative projects to life.” According to Kickstarter’s own website, funding tends to be “all-or-nothing.” Projects that receive substantial funding (>20%) are more likely than not to hit their goals, except most projects never hit that 20% threshold. Exposure tends to snowball support, so how can projects get that initial exposure and build momentum?

Kickstarter is a good method of obtaining funding for some projects but is absolutely dreadful for others. Can we help those who are considering launching a project on Kickstarter decide whether to make the leap? Is there a way to know if a Kickstarter project will be successful before it is launched? What can be done to maximize the chance of success?

Data Wrangling:

The data was obtained from Kaggle after being scrapped from Kickstarter.com.

I noticed there was already a binary state column in the data which compares the goal target to the actual amount of funding received. I checked to verify the binary state column was correct by creating a new column that should be its equivalent by checking whether the funding received in USD is greater than or equal to the goal in USD, and I noted no difference.

I looked at the distribution of funding goals and determine that there were a large number of Kickstarters with targets under 1000. I decided to remove these data as they seem to be less serious campaigns. These data had a much higher probability of being successful (even when compared to the 1000 to 2000 dollar range) which is not an interesting result due to the ease of success and the lack of necessity for such a small amount of funding.

I discovered that there were duplicate project IDs. The duplicates had almost identical information except for the amount of funding received, so apparently, some information was taken from different points in time. I was not expecting duplicates, and if they were systematic, they could have a massive impact on modeling. I decided to remove the duplicates keeping the one with the larger funding received, indicating that it was newer information.

I decided to remove any projects where the deadline came after the last project submitted (approximately when the data was captured). These projects would not have had sufficient time to reach their goal and should not be included in the model.

I created a new column calculating the number of days between the goal date and the launch date. It appeared that 30 days was by far the most popular choice. I created a new column for

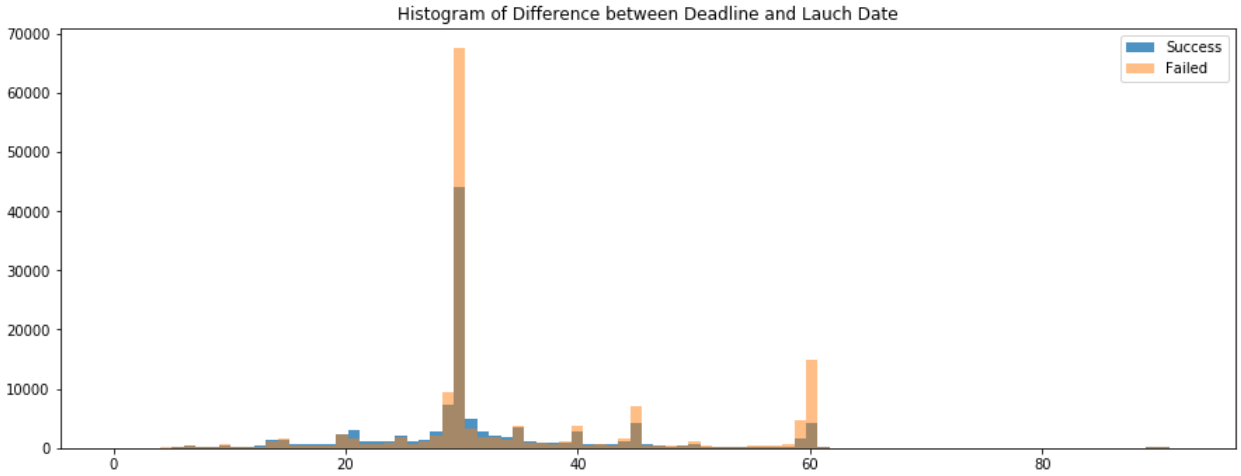
the day of the week and noted Tuesday was the most popular day of the week for launch. I created a new numerical column to count the length of the name.

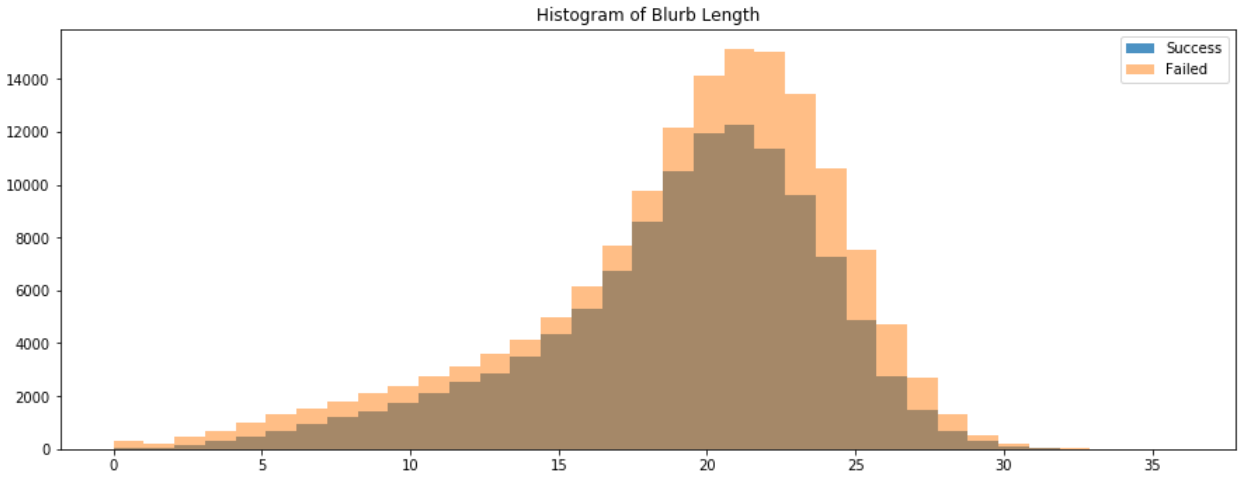
I examined both the name and the blurb for various keywords to find out if certain words were associated with more or less failure.

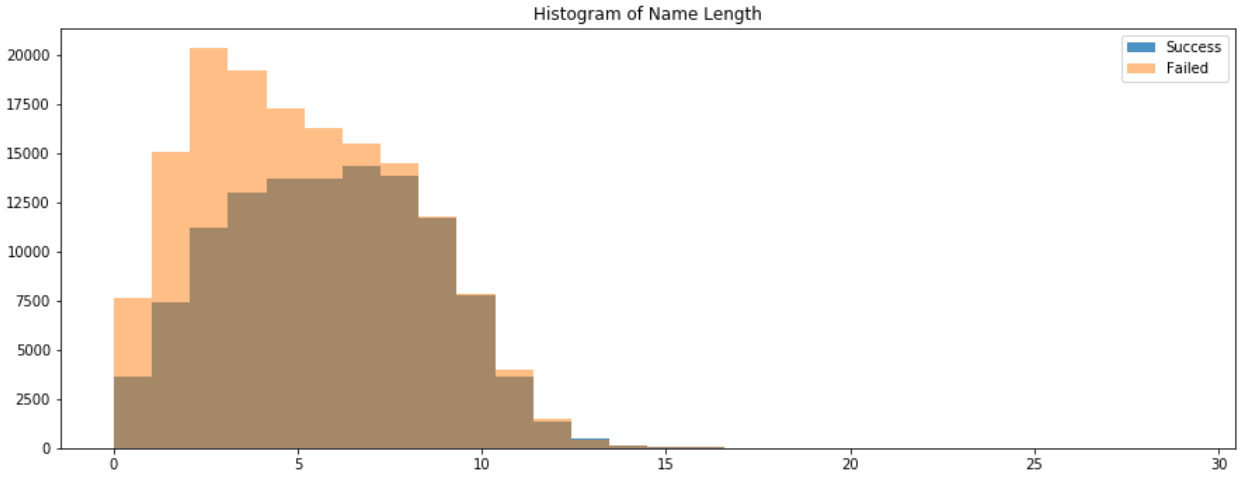
Many ambitious or successful Kickstarter campaigns were “outliers” in terms of the number of backers, the amount of support, or the goal in USD. However, as this data is crucial to understanding successful Kickstarter campaigns, I do not consider this data to have any sort of outlier or problem that needed to be handled. When visualizing this data it was important to use a log scale due to the wide variability between projects.

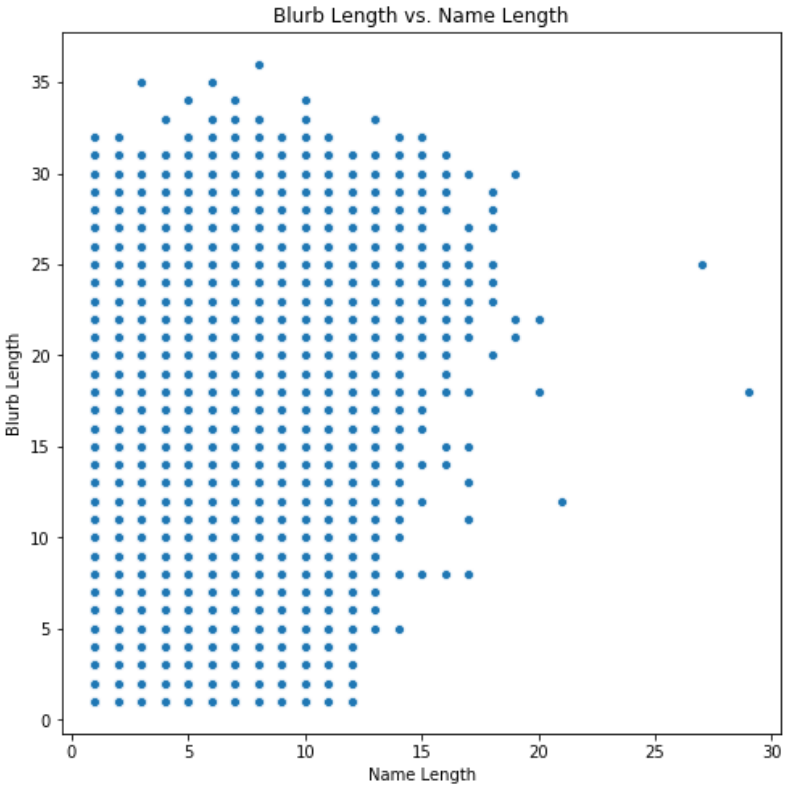
Data Story:

I focused the story on the differences between successful and failed Kickstarter projects. I looked at the histograms of the numerical data: days between the deadline and launch date, the blurb length, and the name length.





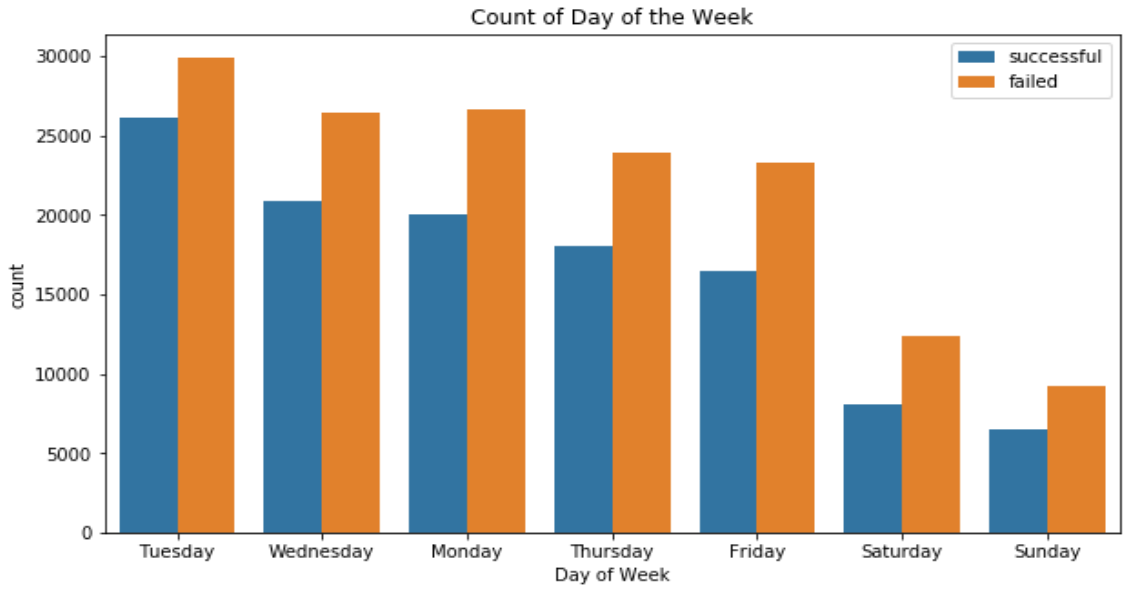


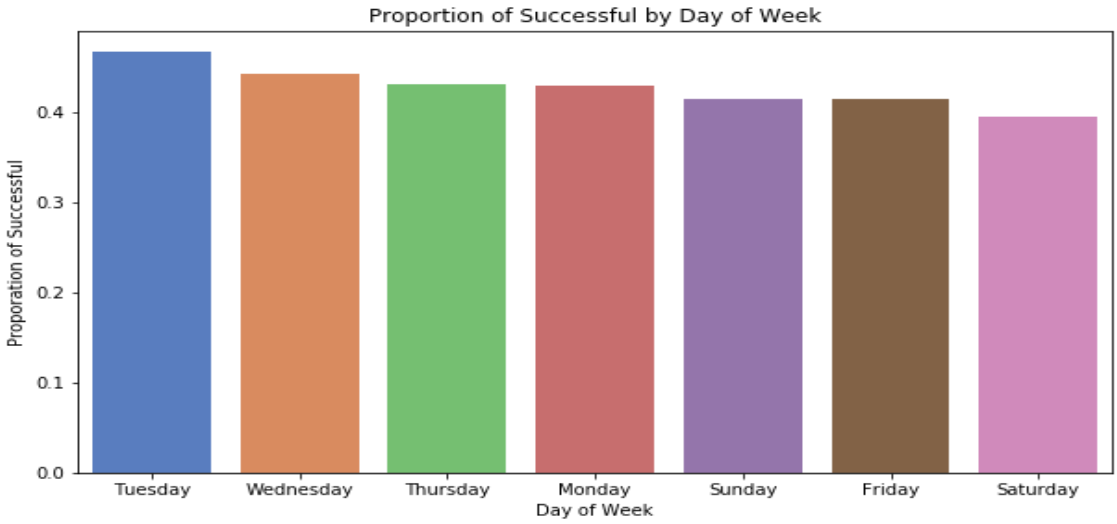


None of the distributions, separated out by success and failure, appeared to be incredibly different, with the name length distributions appearing the most different. However, I tested the differences in the means with bootstrapping and found the differences were statistically significant.

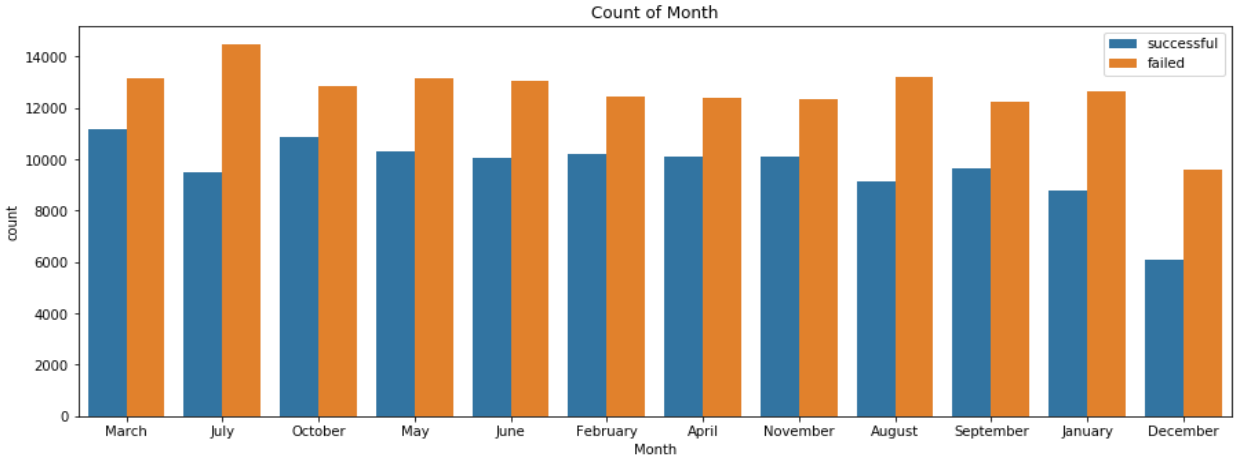
I decided to plot the two numerical variables I decided to keep against each other to check for correlation. Blurb length and name length were only very slightly positively correlated with a Pearson's r of less than 0.19%.

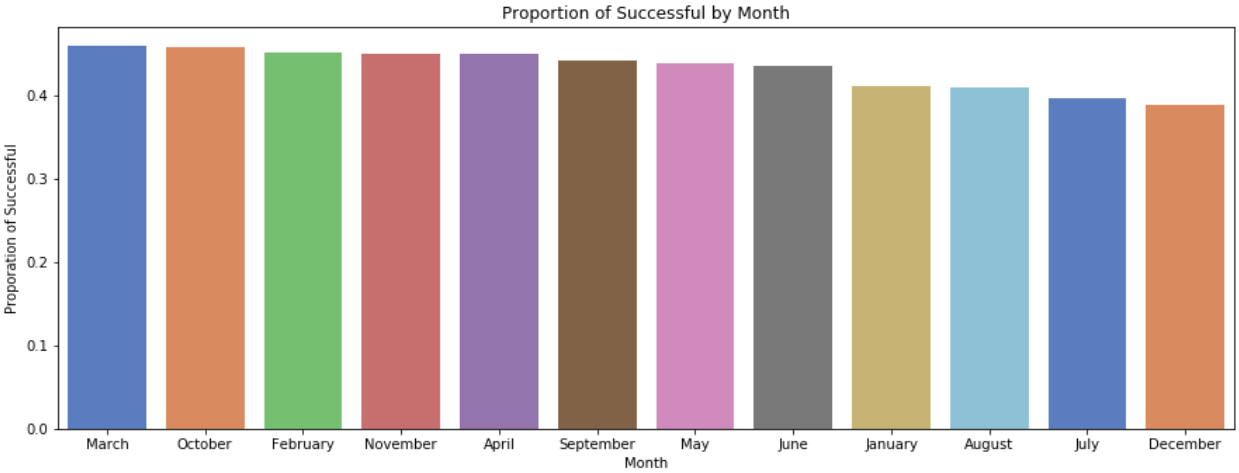
I then examine categorical variables looking for major differences between success and failure in each category. For the day of the week category, there was not major variability, but Friday, Saturday, and Sunday had lower activity and a slightly lower chance of success.



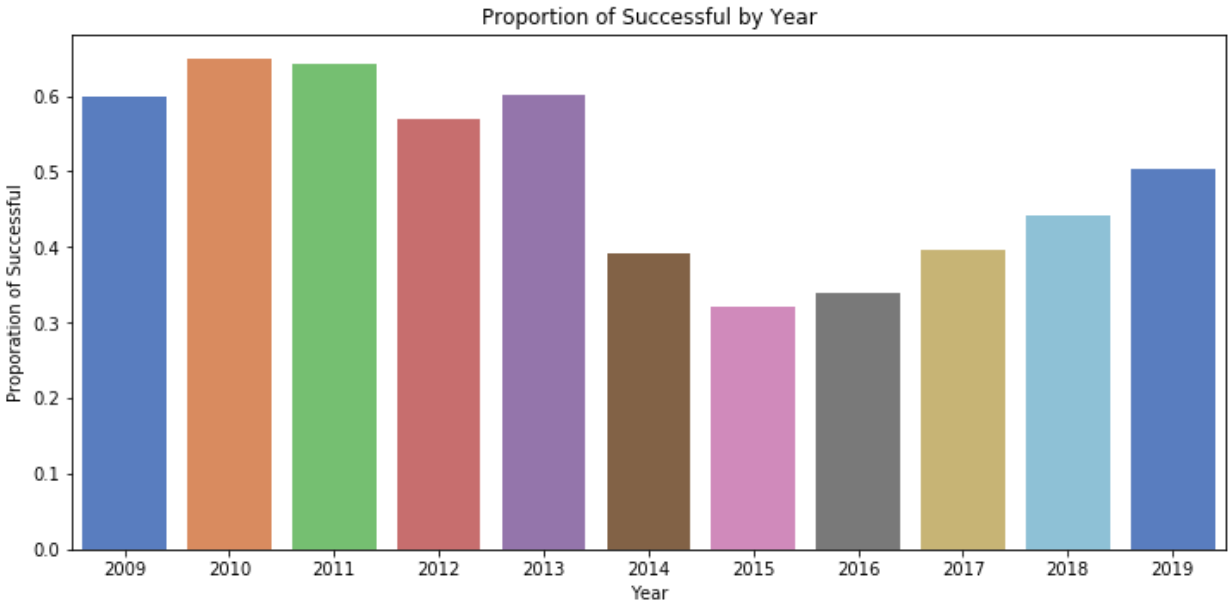


The month was the next variable I displayed similarly. There was not a high level of variability except for in December which was an overall down month.



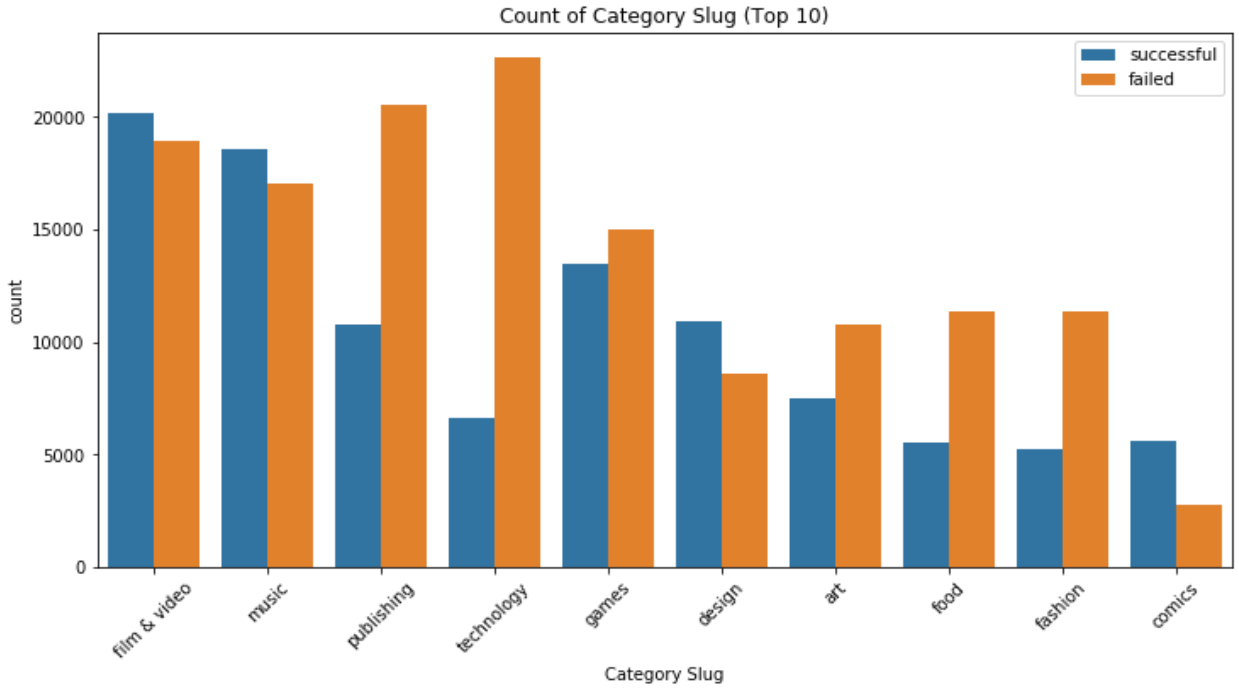


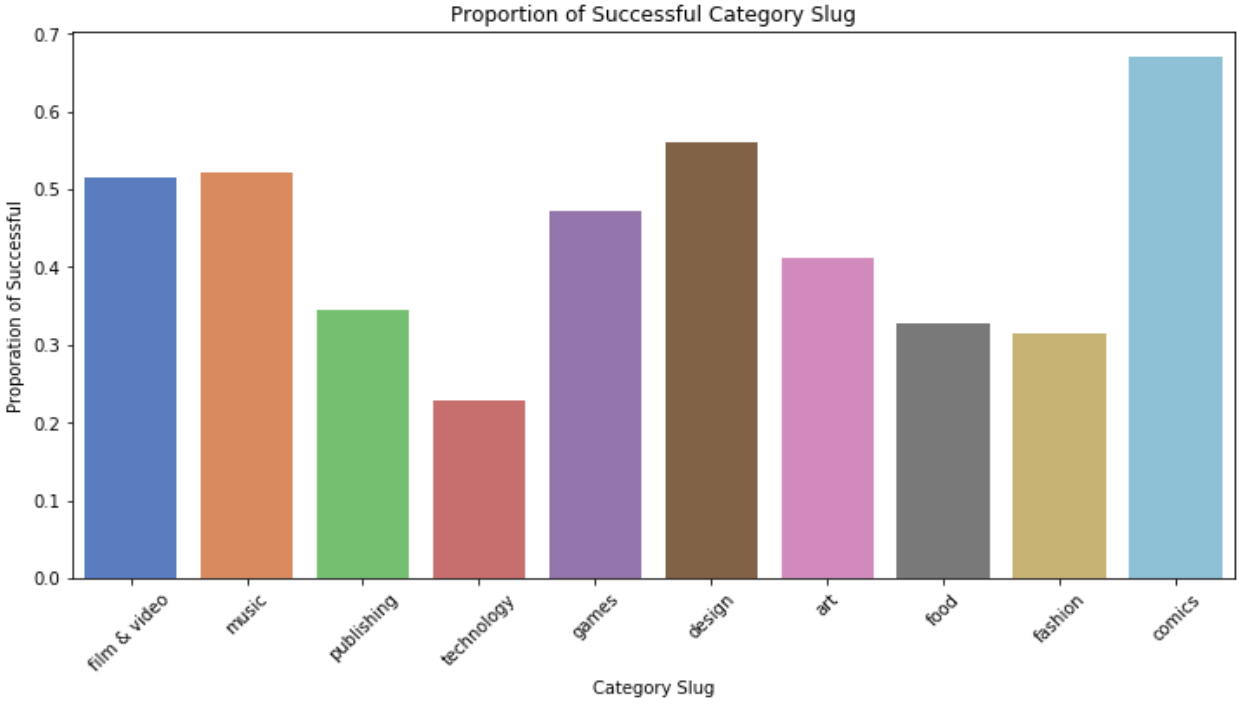
The most interesting discovery came while examining a variable that is not useful for modeling, the year of launch. There was a huge drop off in success from 2013 vs. 2014.



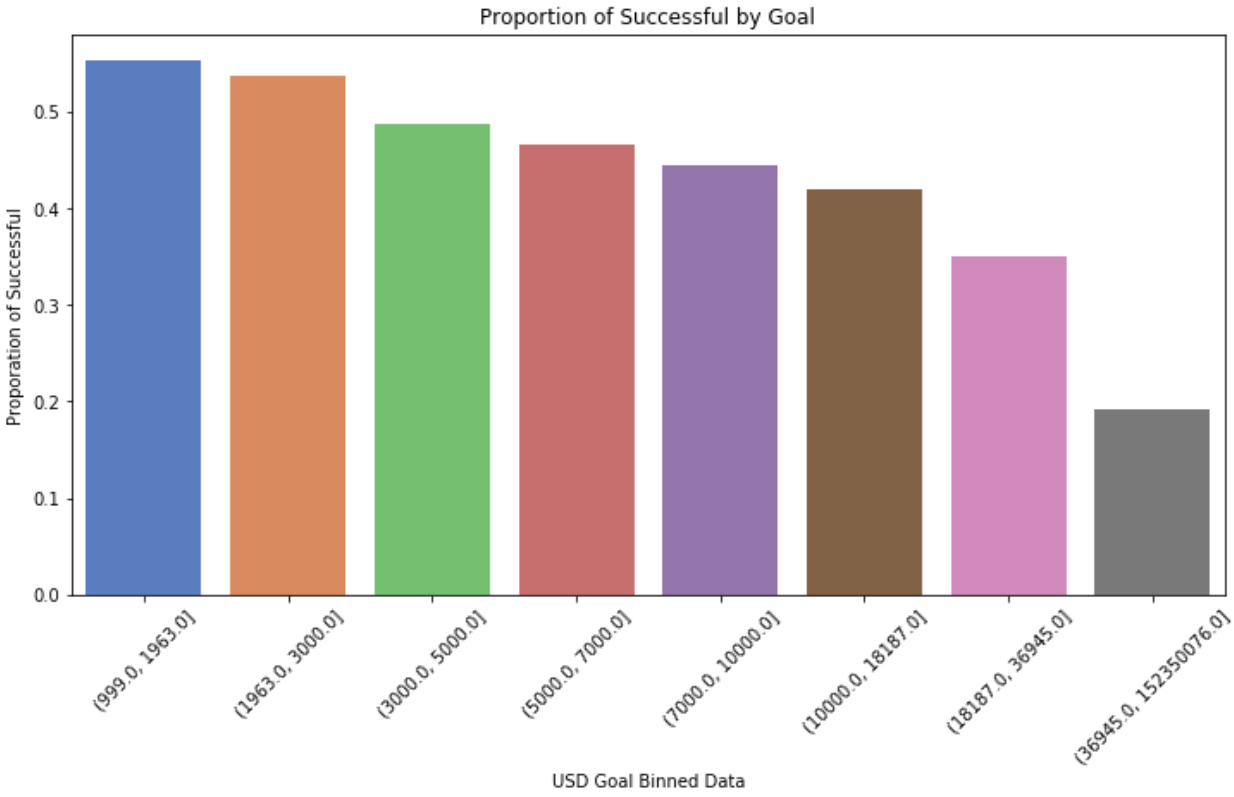
The success rate was 60% in 2013. In 2014, there were almost twice as many projects with almost a doubled average goal. With more ambition came more failure, and projects succeeded at under 40% in 2014. I have not yet been able to decipher what changed in 2014 through research of Kickstarter.

There was quite a bit of variability in the category slug column which could be useful for modeling purposes.





I found the most important variable, the goal amount, which I visualized using qcut. 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% chance of success. However, success is far from guaranteed for less ambitious projects.



Inferential Statistics:

I identified two numerical columns that were most useful: the name length and the blurb length of projects. I performed 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 Pearson's R statistic. The two variables have a very low correlation. Unfortunately, despite the means being significantly different, in practice, the amounts were very small and the variables will likely only have a small impact on the accuracy of my final model.

I used a chi-squared test on two categorical variables 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 modeling.