Capstone 1: Predicting Kickstarter Success

The Problem: Kickstarter is a funding platform for creative projects. When creative groups, companies, or individuals have an idea, a clear plan and a final funding goal, they can submit their projects to the Kickstarter platform in order to garner public support and funding. The Kickstarter platform provides a space where campaigns can both ask for funding donations, and provide incentives and rewards to those who pledge funds to the project.

The Kickstarter platform is funded by fees collected from each donation, and from the overall funding amount when a campaign is successful. Kickstarter applies a 5% fee to any successful campaigns, and collects a 3-5% payment processing fee per donation, depending on the donation amount. If a campaign is not successful and does not reach their funding goal, Kickstarter does not collect the standard 5% fee. Therefore, in order for Kickstarter to continue their success, and increase their profits, they must host successful campaigns that reach or exceed their funding goal. Currently, successful Kickstarter campaigns are estimated at 35% of total campaigns, while failed campaigns are closer to 52%.

The Client: The ability to predict a successful Kickstarter campaign will be of great benefit to both Kickstarter as a company, and to companies and creators who launch campaigns on their website. Kickstarter has an inherent interest in running successful campaigns because of their fee structure, and their overall profits as a company. Additionally, competition from other crowdfunding platforms are gaining popularity and Kickstarter will need to remain competitive in offering services and exposure to clients that will lead to successful campaign outcomes.

The Approach: An analysis of successful Kickstarter campaigns will address metrics for campaigns that reach and exceed their funding goals. This includes the category of campaign, rewards/incentives offered, funding goal, funding time frame, and campaign description. Companies and creators who are launching campaigns also have an inherent interest in understanding the factors that create a successful campaign. Having a campaign or project reach or exceed funding status could alter the trajectory of a product or idea. Alternatively, campaigns and projects that end up failing to meet their funding goals could end up on life support.

By analyzing trends in successful campaigns, Kickstarter will be able to determine which campaigns are more likely to reach or exceed funded status. Armed with this data insight, Kickstarter will be able to make data driven, impactful decisions in regards to:

- Services offered to clients
- Fees that are collected from clients and from contributors,
- Campaign guidelines and recommendations

Data Wrangling

Overview:

The dataset that was used for analysis was provided in one .csv file, obtained from Kaggle. At first glance, the data is fairly clean containing 15 columns with 378,661 rows of data. Each Kickstarter campaign is represented by one row of data including the campaign name, the main category that the campaign falls under, the currency type that pledges are converted to, the campaign deadline, funding goal, the state of the campaign, how many backers supported the campaign, what country the campaign originated from, and then two columns that are conversions of the pledged amount column converted to USD.

Duplicate Data: I began the cleaning process by determining whether any data was duplicated. Each Kickstarter campaign is assigned a campaign ID, and I proceeded to work on deduplication based off of this column. In order to check for duplicate rows, I created a new data frame that would contain any potential duplicates. I created this data frame using df.duplicated() and then printing the shape of the new data frame. There were no duplicate rows that needed to be removed in the original data frame.

Null Values: Next, I determined whether there were any null values that needed to be addressed. To get a broad overview of all of the column names, I printed the column values, and examined whether there were any null values in each column. There were 4 null values in the **name** column, and 3797 null values in the **usd_pledged** column. The 4 null names are for campaigns that were cancelled or potentially created in error without a campaign name.

Column Adjustments:
Upon further inspection and research, the **usd_pledged** column and **usd_pledged_real** column are similar in that their existence had a common goal. The two columns were meant to convert the pledged entirely to USD, as some campaign pledges were in other countries' currency. The first column, **usd_pledged**, was created by Kickstarter, and looks as though it did not completely convert all pledges successfully. Alternatively, the **usd_pledged_real** column contains all correctly converted values. Because of this, I decided to remove the usd pledged column from the data frame, and create a new data frame called **clean_df,** using df.drop on the usd pledged column. To double check that everything went correctly, the column names and data frame shape are reprinted confirming that **usd_pledged** has been removed.
The campaign state column was examined to look at the total campaigns for each campaign state. There are 6 different campaign state categories: successful, failed, live, suspended, cancelled and undefined. We cannot possibly determine the campaign state of live, undefined, suspended, or canceled campaigns. Rows containing these campaign states are removed from the analysis.
For future statistical analysis, a new column - binary_state - is created that includes failure: 0, success: 1. Campaign outcomes are defined as a campaign that reaches its campaign duration and either met or exceeds its campaign goal amount, or failed to meet its campaign goal amount.

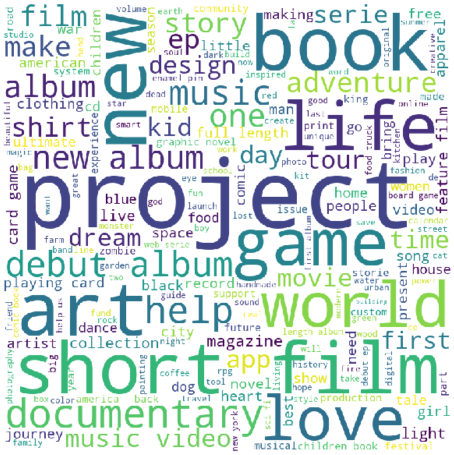
Outliers: Due to the nature of this dataset, it can be expected that some columns will contain outliers. For example, ambitious campaigns who set a very high campaign goal or campaigns that exceeded expectations and raised thousands of dollars more than expected. Most columns in the Kickstarter data set are objects, and would not have an outlier associated with them because they are categorical. In order to identify outliers in the appropriate columns (goal, pledged, usd_pledged_real, usd_goal_real) the datatypes are re-examined in order to remove the object columns. After object columns are removed there are only 6 columns left. From these 6 columns, a zscore over 3 is calculated, and any outliers identified are rejected. After the outliers are rejected, the data frame is left with 375,784 rows in comparison to the original 378,661. This will be helpful to take into account when statistical analysis is completed. It is important to identify outliers in order to account for possible statistical errors in the future. Outliers can skew statistical measures such as means and medians, and will need to be further considered when designing the predictive model. For data exploration purposes, the outliers continue to remain in the dataset at this time.

Data Storytelling & Statistics

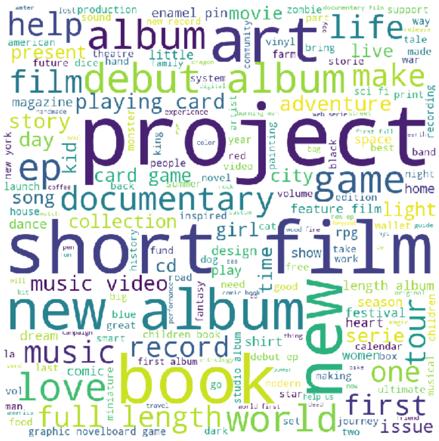
Campaign Names

Exploring the names of campaigns and whether they had any impact on the success or failure of a campaign was something I found very interesting. In order to analyze campaign names, I created three word clouds to get a visual understanding of what words were coming up in failed and successful campaigns.

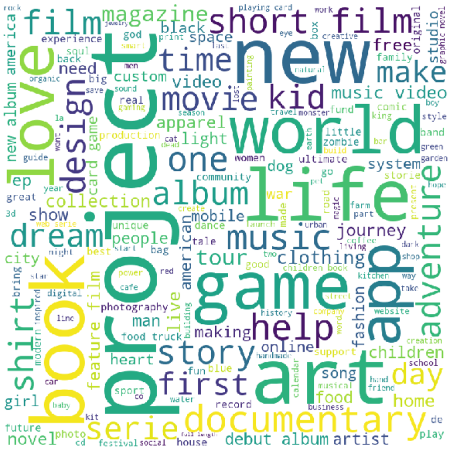
All Campaigns:



Successful Campaigns:

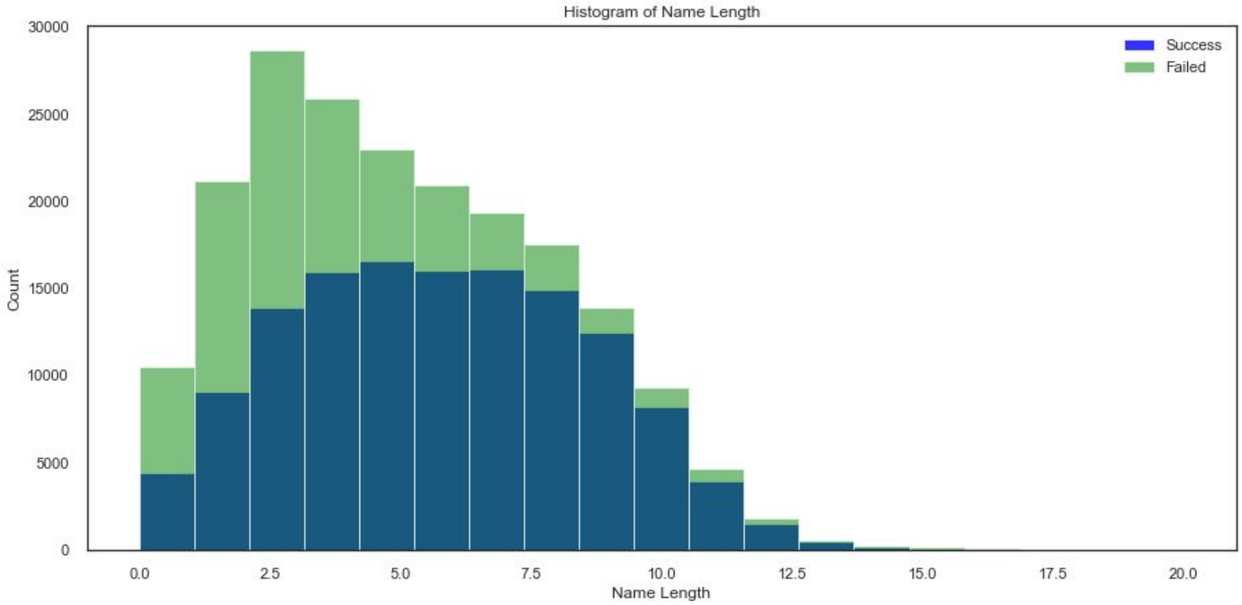


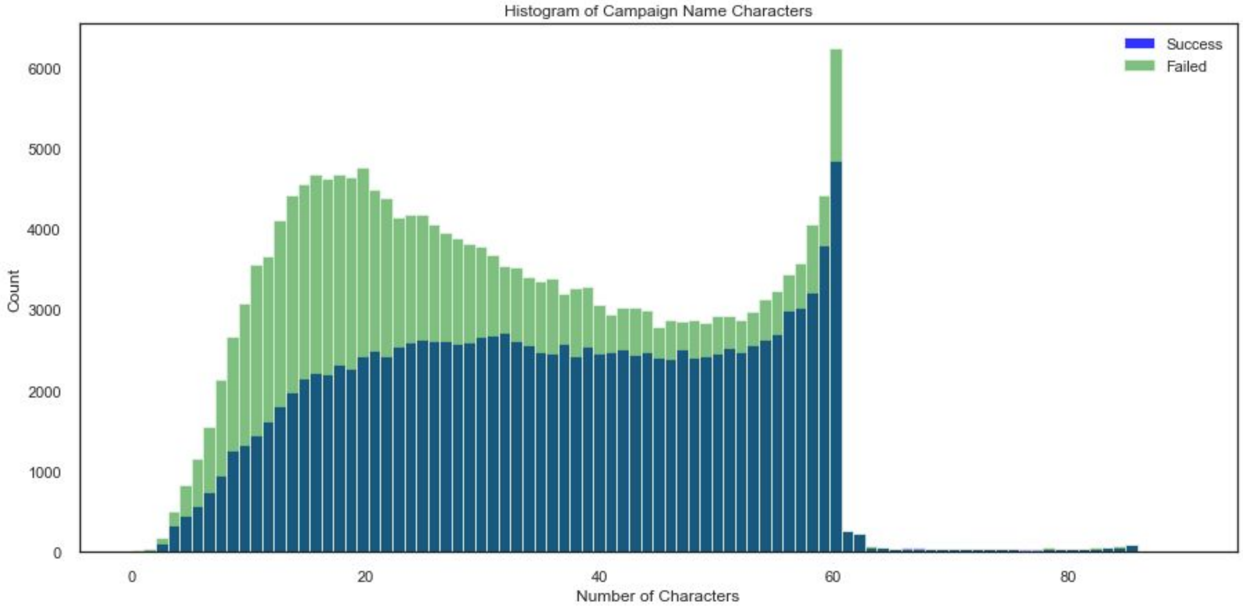
Failed Campaigns:



As you can see from the word clouds, many campaigns share similar words regardless of their success or failure. To dig into this a bit deeper I created two new columns: one for the number of characters in a campaign name (name_cl) and one for the number of words in a campaign name (name_len).

When plotted in a histogram, it appears as though campaign names with more characters are more likely to fail than campaigns with less characters in their name. Similarly, campaign names that have less words in them are more likely to succeed than campaigns with more words in them. Even though the campaign name itself doesn't seem to have an impact on success or failure, the number of characters and the number of words in a campaign name seems it may have an effect on the campaign outcome.

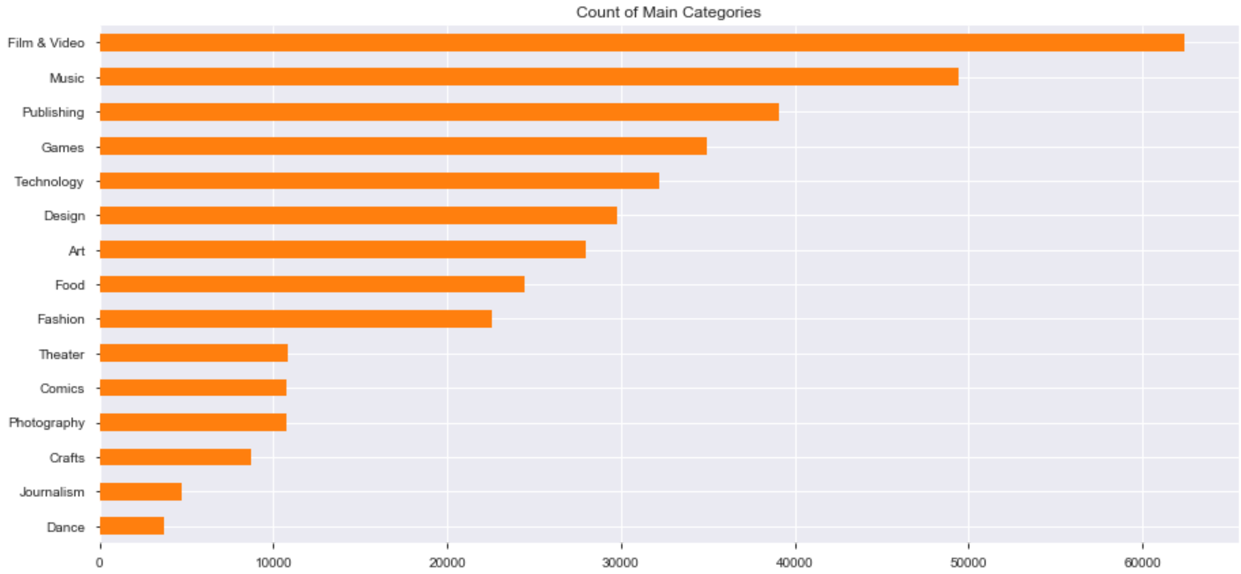




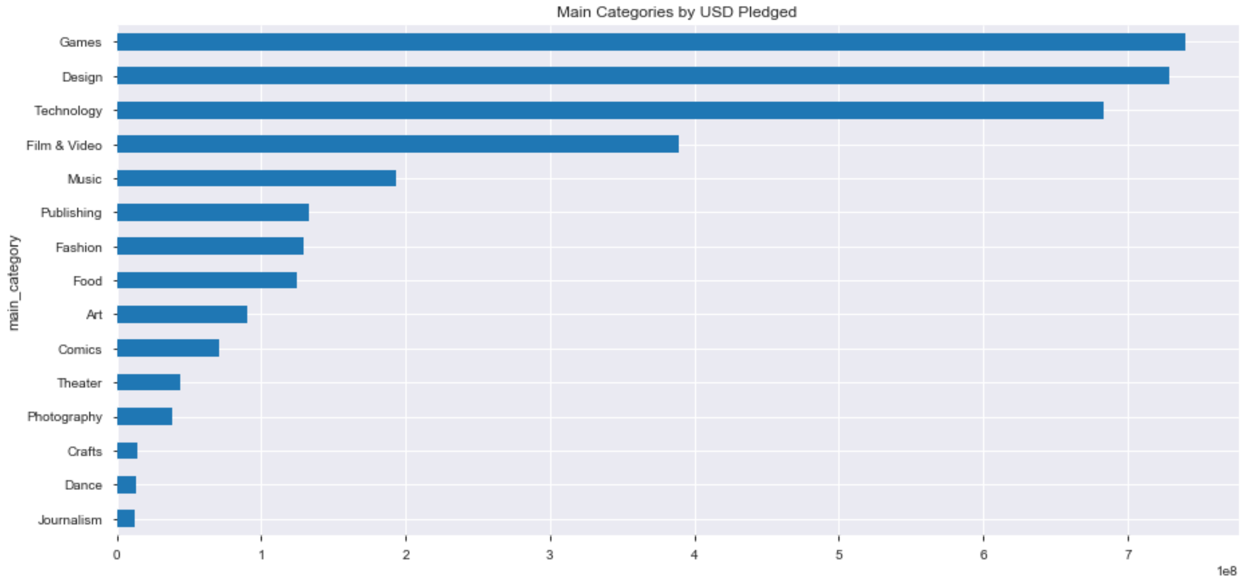
To test whether the length of the campaign name or the number of characters in a campaign name had an impact on the campaign outcome, I first tested the normality of the distributions using the Shapiro-Wilks test. I followed the normality test with the Kruskal-Wallis H-Test to determine whether the medians of the two groups were different. The Kruskal-Wallis test determined that the population medians were unequal. A bootstrap analysis was completed to compare the means of the two groups and found that there was a statistically significant difference. My analysis for both the number of characters in a campaign name and the length of the campaign name were very similar, finding that there was a statistically significant difference between the two outcome groups.

Campaign Categories

It's important to begin this analysis by looking at any potential relationships between the campaign category and it’s rate of success. Are there campaign categories that are simply more popular than others? Or are there categories that are overflowing with campaigns, not all of which are worth investing into? This chart represents a count of Kickstarter campaigns by category:

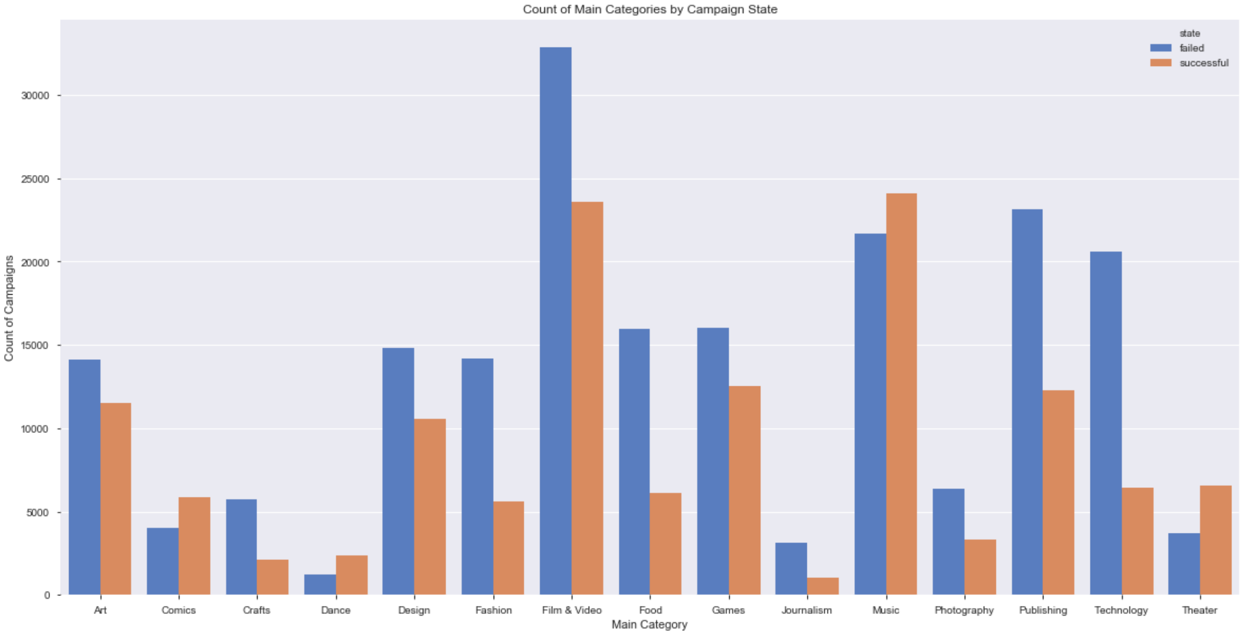


Next, I want to look at the campaigns based off of the total sum of usd_pledged_real per campaign category. When this is applied, we get the following:

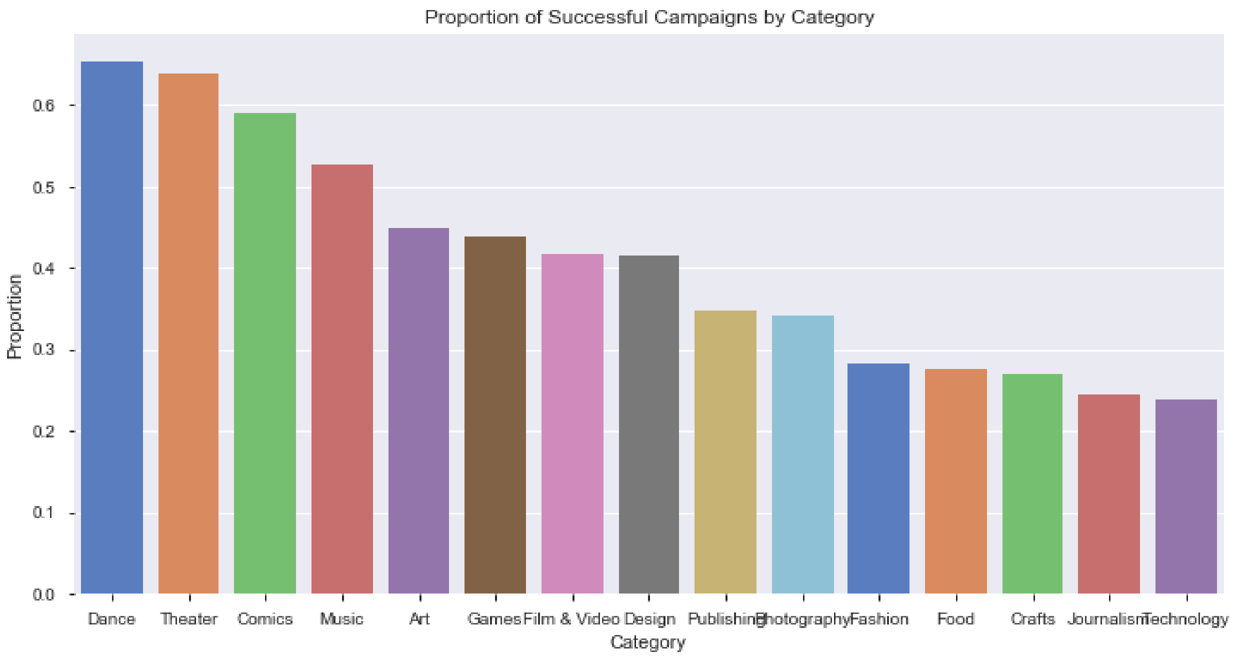


Interestingly, the main categories based on the sum of pledged USD does not line up with which campaign categories produce the most campaigns. Although games are the 4th ranked main category in terms of number of campaigns, it far outpaces Film & Video. Campaigns with the main category games brought in $739,853,563 in pledges alone from 34,943 campaigns. For Kickstarter, this accounts to a $678,832,833 profit from the successful Game campaigns, not accounting for additional fees collected at the time of each pledge - that's a lot of money!

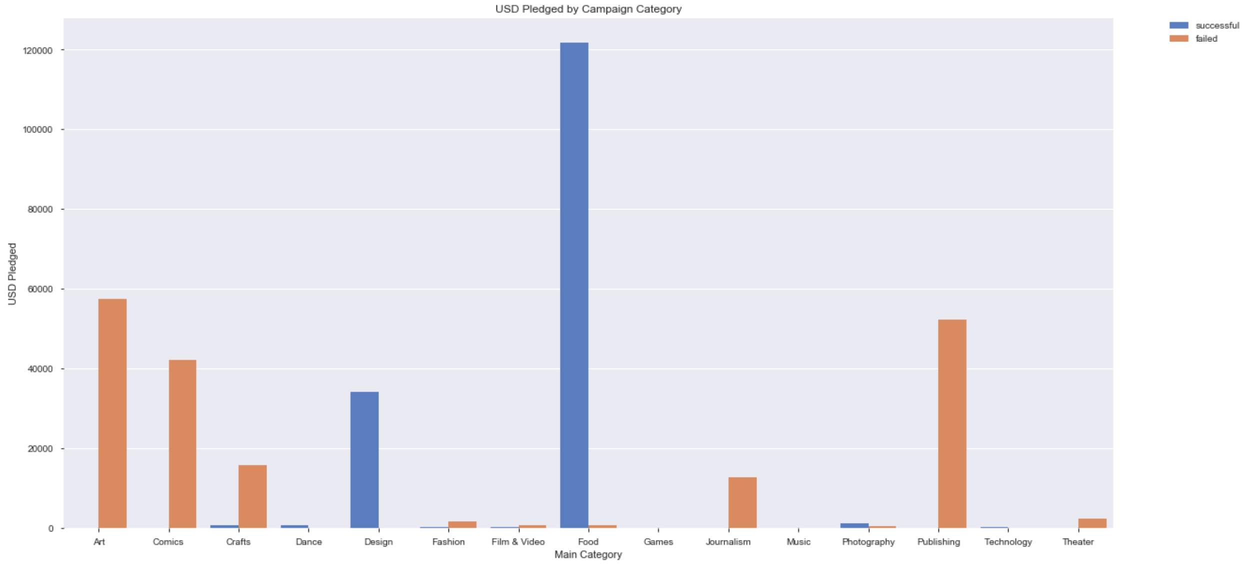
From here, I broke down each category by campaign outcome to begin to look at the success rate within a campaign category. This chart is based off of the count of each campaign state within each campaign category:



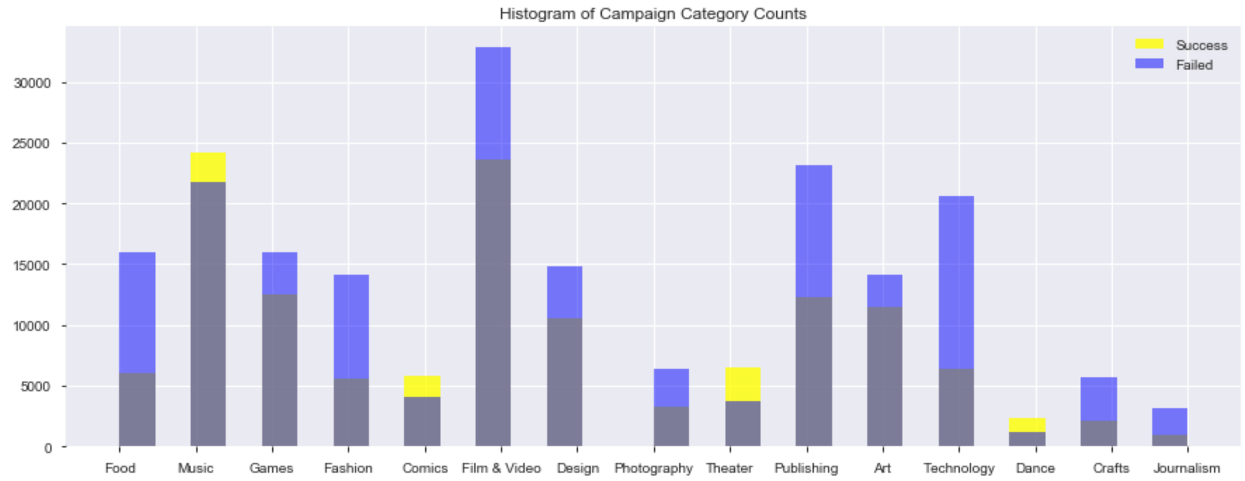
What proportion of these categories were successful? Proportion of success was determined for each main category and displayed in the chart below. Dance has the highest proportionality of success, while technology has the lowest proportionality of success. Based on what we know about the campaign categories, Dance has a total of 3,749 campaigns, pulling in a total of $112,997,480 across all campaign states. Technology has a total of 32,189 campaigns, pulling in a total of $683,918,915 across campaign states. This graphic is helpful in showing the proportionality of campaign success, but it should also be noted that some campaign categories have many more campaigns than others, leading to a higher chance of failure.



Next, I broke this down further by looking at the total sum of usd_pledged_real by the campaign state, per campaign category:



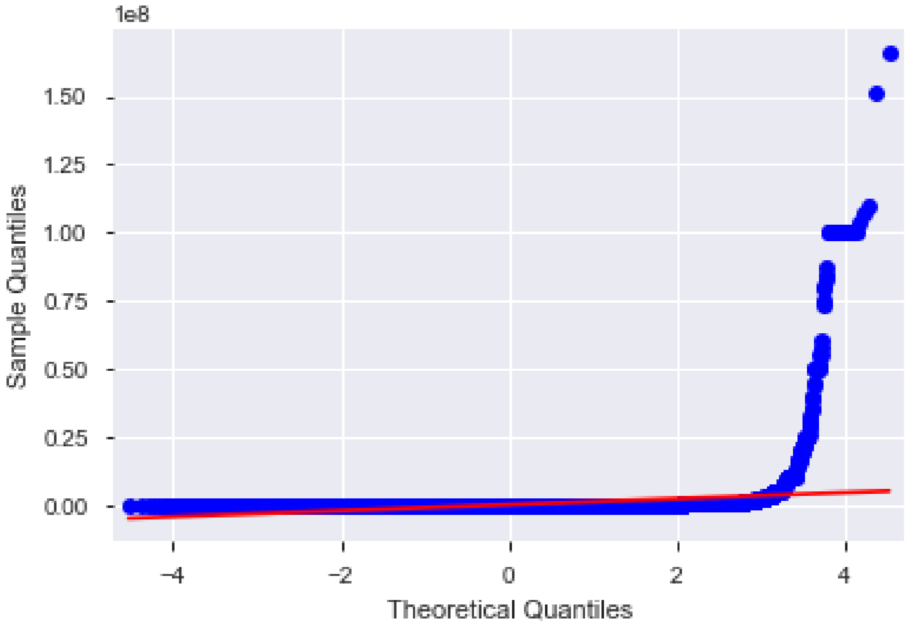
Because this is categorical data, I created a cross tab and histogram to visualize any relationship between campaign category and campaign outcome:

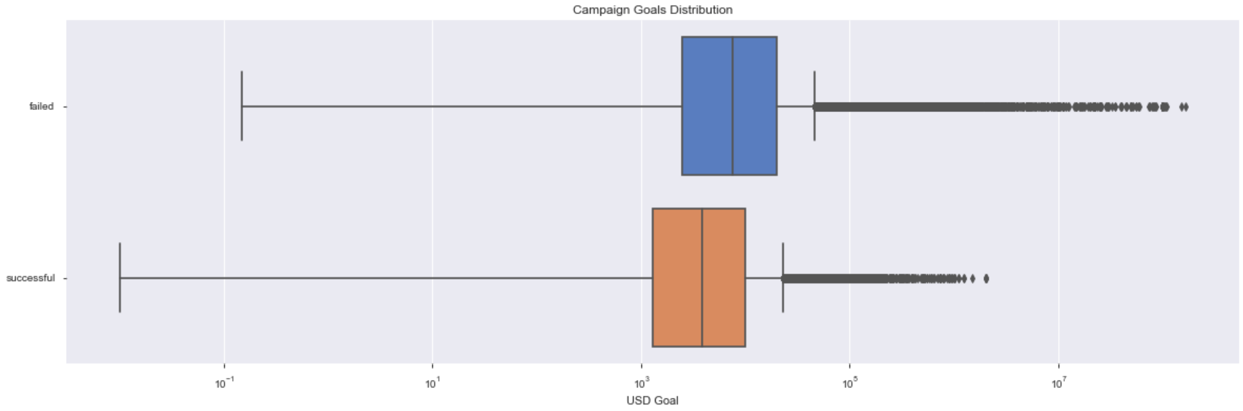


Campaign Goals

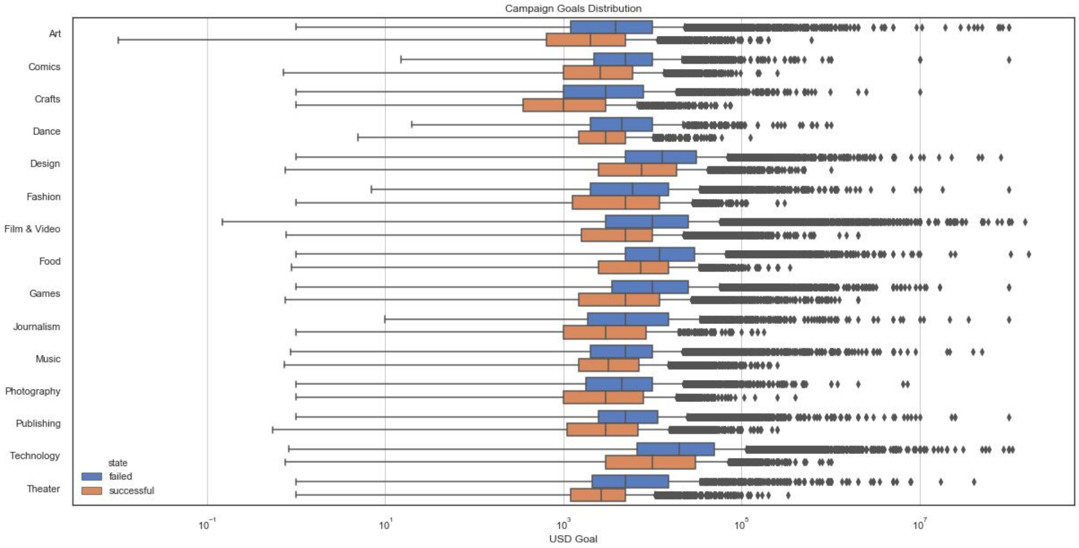
Kickstarter campaigns seek funding that ranges anywhere from $0 to $ 100,000,000, however the average campaign goal is $49300. With conversions for different currencies already taken into account, total pledges for campaigns range anywhere from $0 to a high of $20,338,986. On average, campaigns end up raising $9148 total funding regardless of their eventual success or failure.

In order to test the hypothesis, I first had to check whether or not the campaign goal data follows a normal distribution. I used three different tests to determine the normality of the distribution: Shapiro-Wilkes Test, qqplot, and a box-whisker plot. The Shapiro-Wilkes test loses p value reliability when there are over 5k data points, so the additional visualizations are helpful to validate the Shapiro-Wilkes finding of non-gaussian:





Not only does this distribution appear to be abnormal, it looks as though the campaign goal alone would not determine a campaign outcome. However, for some campaign categories, this may not be the case:



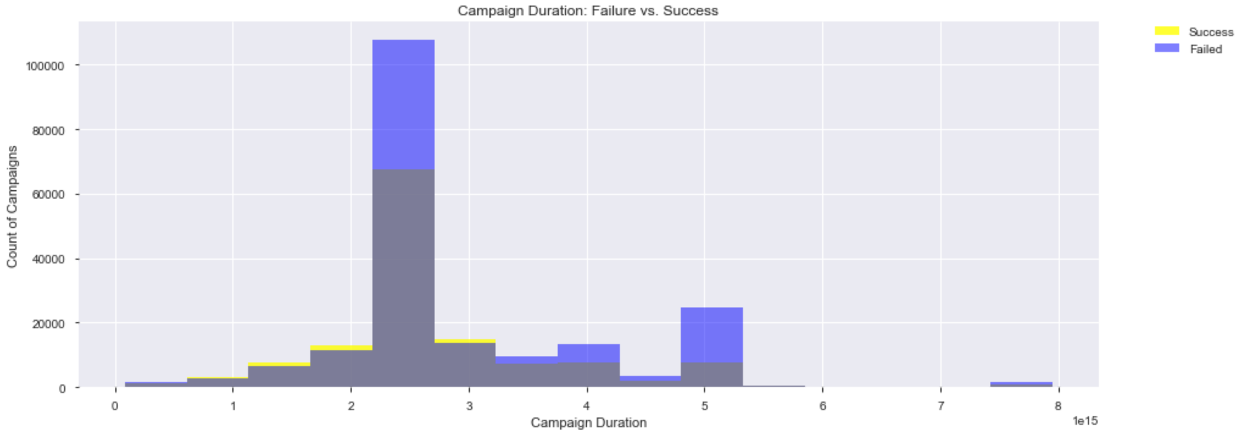
Campaign Duration

Research Question: Is there a statistically significant relationship between the duration of a campaign and campaign outcome?
Hypothesis: There is no statistically significant relationship between the duration of a campaign and the campaign outcome.

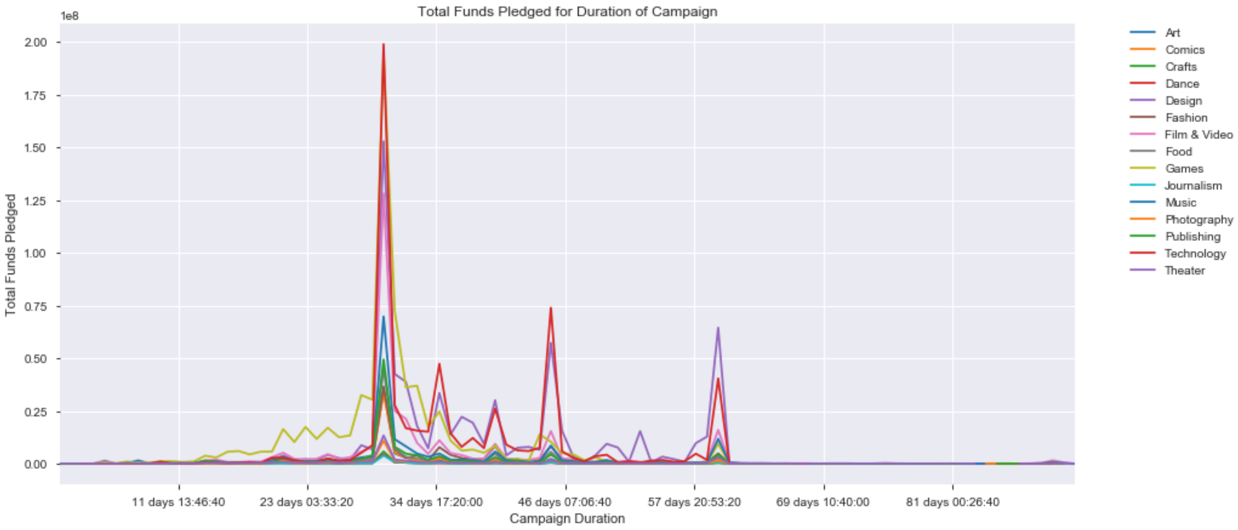
After exploring the campaign data, it is important to add a column that shows the duration of a campaign to contextualize how much time it has taken for successful campaigns to reach or surpass their funding goal, or for determining the average length of time of a failed campaign. In order to create a campaign duration column, both the launched and deadline columns are converted to datetime. From there, the campaign_duration and camp_days columns are created and added to clean_df by calculating the difference between launched and deadline. Exploration of this new column shows that the minimum campaign duration is 1 day, while the longest campaign duration was 16739 days (cancelled or suspended campaign). Successful campaigns typically run for an average length of time of 32 days, while failed campaigns typically run for an average length of 35 days.

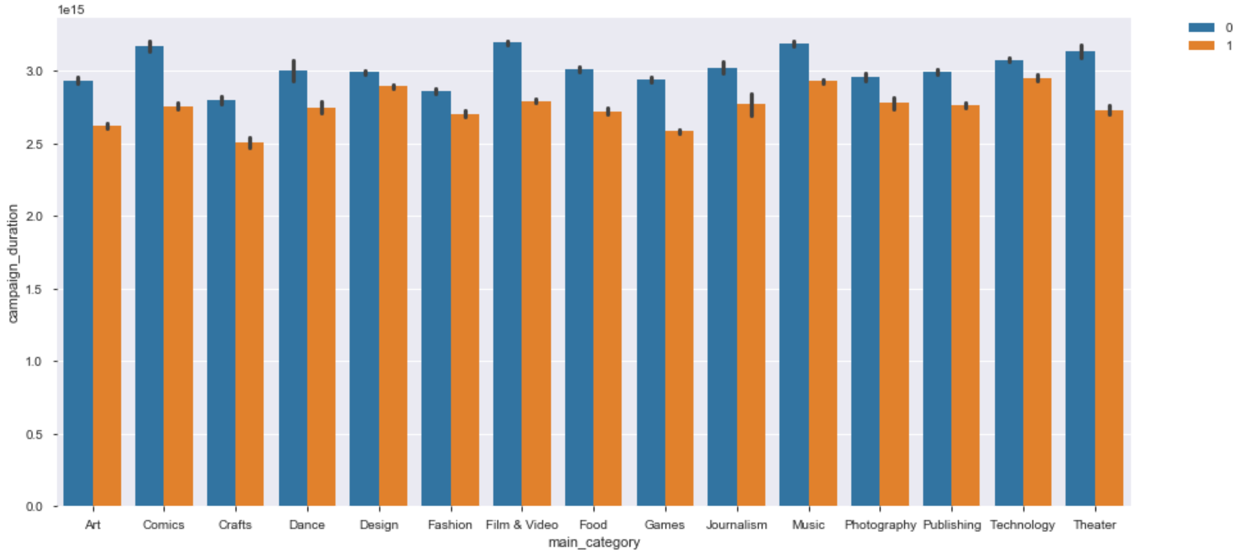
According to Kickstarter.com: “Projects on Kickstarter can last anywhere from 1 - 60 days. We've done some research, and found that projects lasting any longer are rarely successful. We recommend setting your campaign at 30 days or less. Campaigns with shorter durations have higher success rates, and create a helpful sense of urgency around your project.” We do have some outliers in both the failed data frame and successful data frame (duration of 92 days). These outliers could be potentially explained by previous Kickstarter policy.

Around the 30 day mark on this chart, we see the highest number of campaigns in their successful or failed state. This supports the analysis that the majority of campaigns end around the 30 day mark, and that more campaigns are failing around this campaign duration than succeeding. Here we see the count of campaign success and failure by campaign duration:

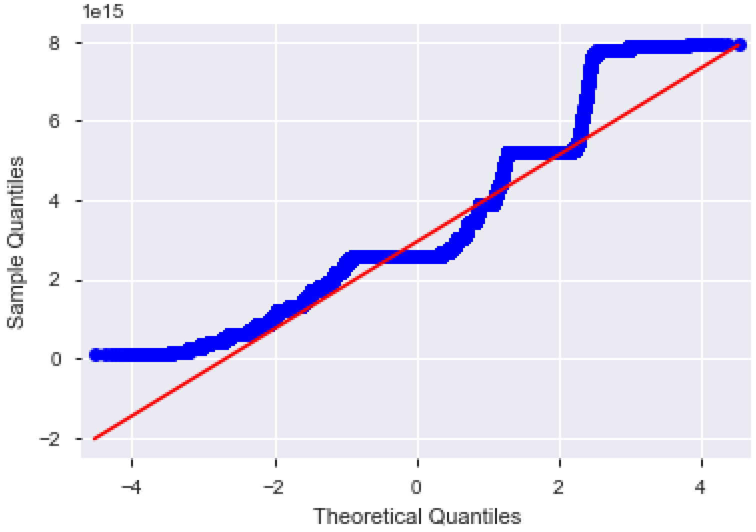


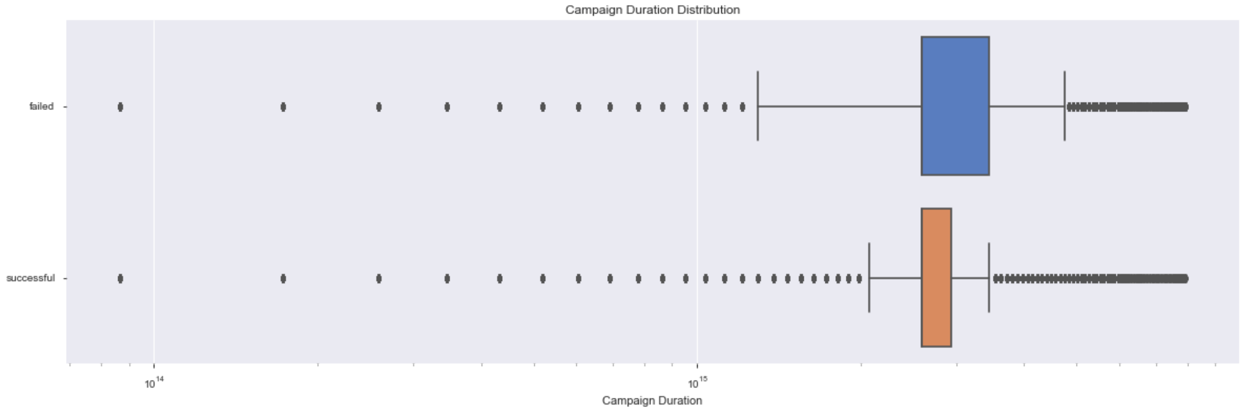
Does the duration have any correlation with USD pledged? Following the same patterns as our previous duration analysis, the campaigns with the highest USD pledged by category occur around the 30 day mark of a campaign:


Total Funds Pledged for Duration of Campaign
1e8
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Total Funds Pledged
11 days 13:46:40
23 days 03:33:20
34 days 17:20:00
46 days 07:06:40
57 days 20:53:20
69 days 10:40:00
81 days 00:26:40
Campaign Duration
Art
Comics
Crafts
Dance
Design
Fashion
Film & Video
Food
Games
Journalism
Music
Photography
Publishing
Technology
Theater



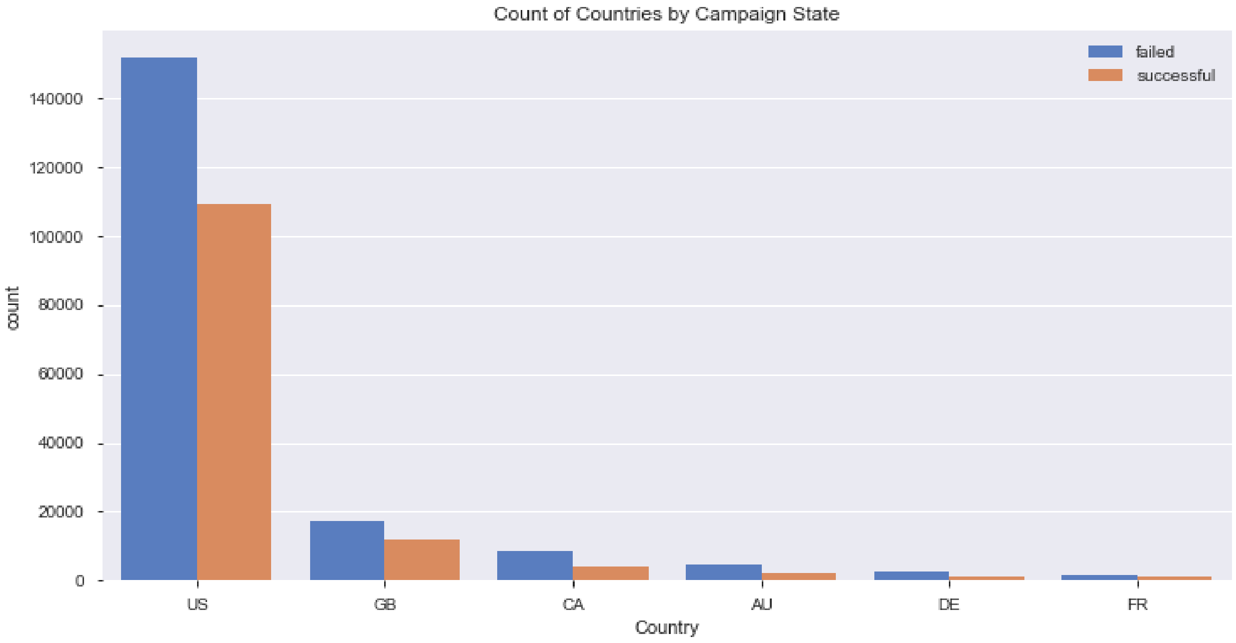
To test null hypothesis, I conducted the Shapiro-Wilkes test, qqplot and box-whisker plot to test the normality of the distribution of campaign durations:



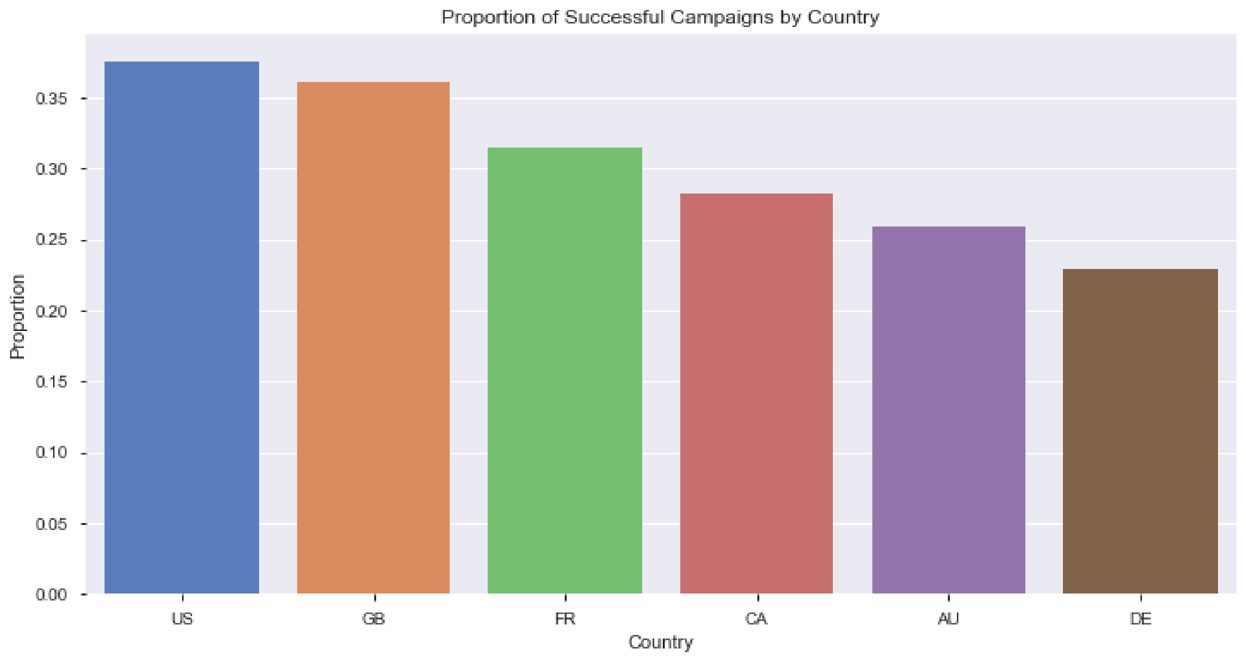


Campaign Country

Kickstarter was launched in the United States and is head-quartered in Brooklyn, NY. It seems unsurprising that campaigns in the US are more successful than campaigns outside of the US, by exposure alone. The US far exceeds any other country with their overall count of campaigns:



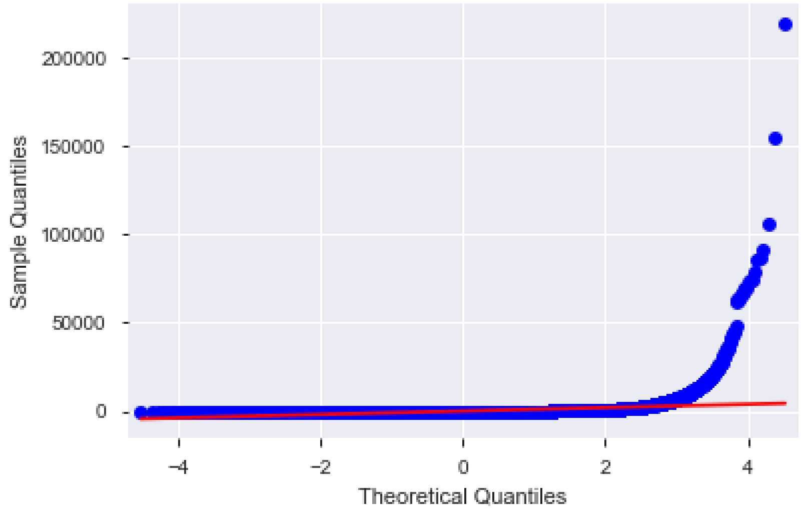
While the US has 109,299 successful campaigns, they have 152061 failures accounting for a 0.375744% success rate. Great Britain - while far behind in campaign volume - has 12067 successful campaigns and 17387 failed campaigns, accounting for a 0.361363% success rate. These two success rates are very similar to one another:

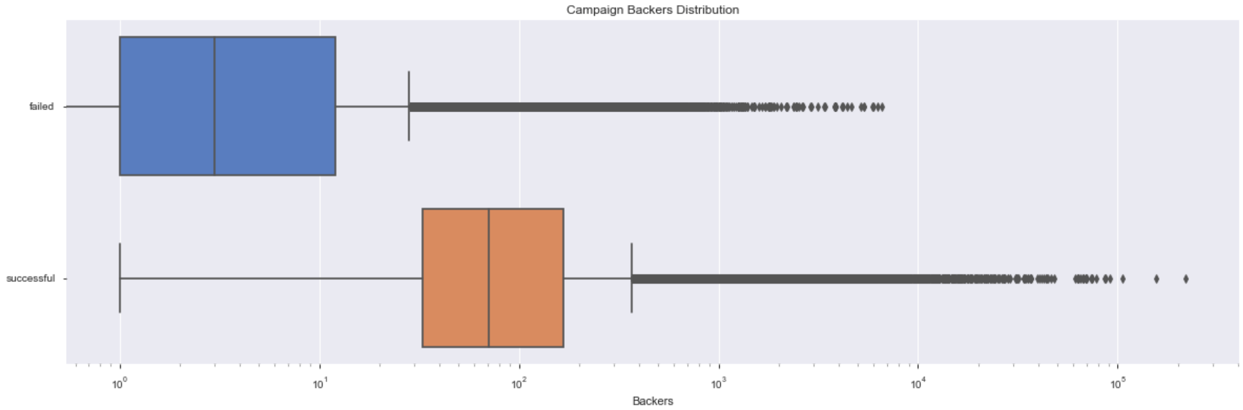


Campaign Backers

Research Question: Is there a statistically significant relationship between the number of backers per campaign and campaign outcome?
Hypothesis: There is no statistically significant relationship between the number of backers per campaign and campaign outcome.

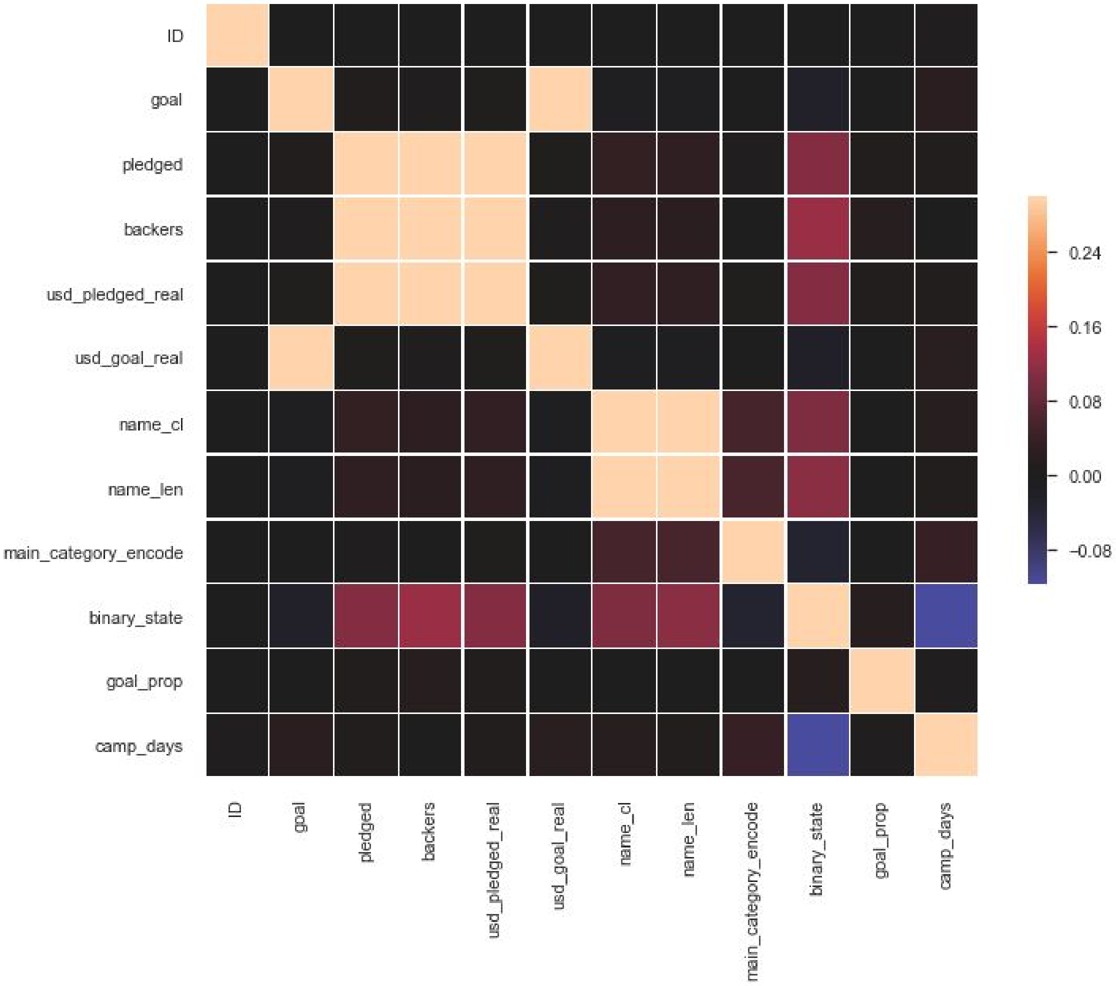
All campaigns have an average of 106 backers per campaign. Successful campaigns have an average of 263 backers per campaign. Failed campaigns have an average of 16 backers per campaign.





Correlations Matrix

A correlation matrix was created to provide an overview of possible relationships between variables from the Kickstarter dataset:



Observations from Data Storytelling and Statistics

From visualizing the dataset and exploring the relationships between features and campaign outcomes, there are several observations that I feel are worth exploring further in the machine learning models.
A challenge with this data set is that all feature distributions appear to be abnormal in some way. Using inferential statistics I examined features separated by the campaign outcomes. In doing so, I believe that features such as campaign_category, usd_pledged_real, campaign_duration, and backers will be particularly important to test in creating a classification model.

Modeling

Data Preparation

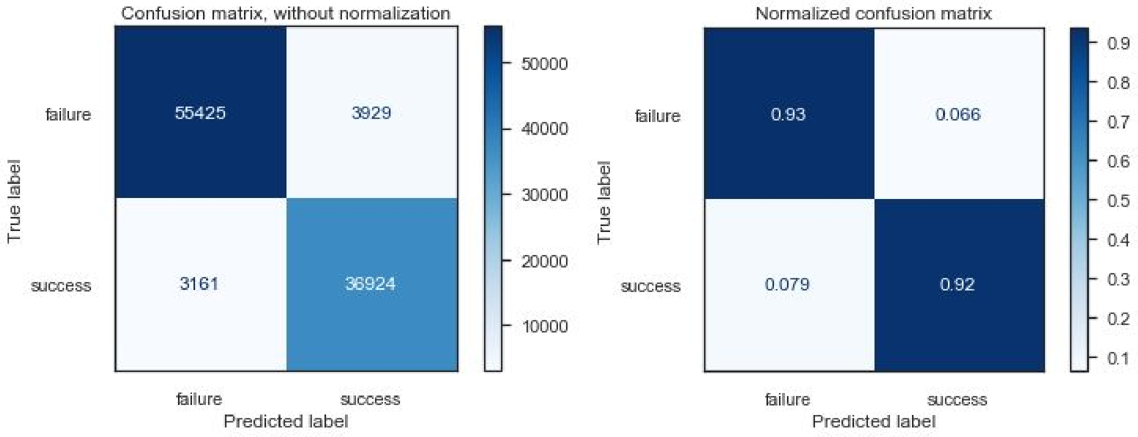
Because I had been prepping data as I went, not much needed to be adjusted with my data set in order to get it cleaned up for testing classification models. The main category's had already been encoded, a binary_state column was created for campaign outcomes, the name_len and name_cl columns had previously been created as well. The remainder of features i'd like to test out were numeric from the original data set.
To begin testing out classification models I created a new dataframe - model. The target label for this model is 0 - failure, 1 - success from the binary_state column.

Random Forest

The first classification model that I chose to explore is the Random Forest Classifier. I chose this model due to the abnormality of most of the Kickstarter data. Random forest reduces the chance of overfitting by analyzing random sub samples of data. This felt like a good model to try due to the abnormality of the Kickstarter dataset.
To test the Random Forest model, I created a baseline model using dummy classifiers to show what the baseline performance of the model would be if someone was simply guessing. Using the dummy classifier predicts the majority class to give better insight into parameters of the model.

The Kickstarter data is a total of 99,439 campaigns. 59,421(59%) of the campaign outcomes are 0 - failure, 40,018(41%) are 1-success. Using dummy classifiers the baseline model predicted all 99,439 campaign outcomes as 0-failure, because it is the majority class. The baseline classification accuracy of the baseline model is 59.6%.
Next, I ran the model with default parameters to compare accuracy to the baseline model:

Random Forest Classifier with Default Parameters:

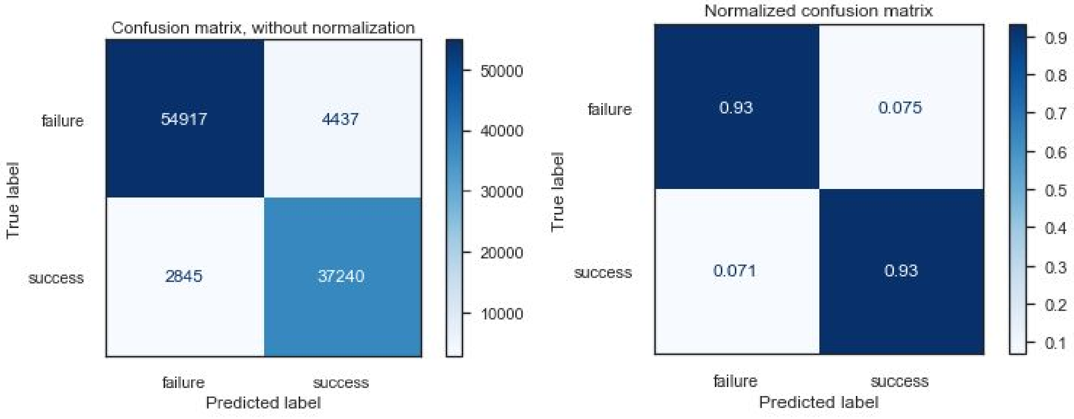


To optimize the Random Forest Model, I began tuning parameters with RandomSearchCV. I chose RandomSearchCV for hyperparameter tuning because RandomSearchCV because it is efficient, reliable and quick. RandomSearchCV was a good fit because I already had an understanding of which hyperparameters in particular we could look at tuning.
I used a 5 fold cross validation, and roc_auc scoring. RandomSearchCV determined the best parameters would be:

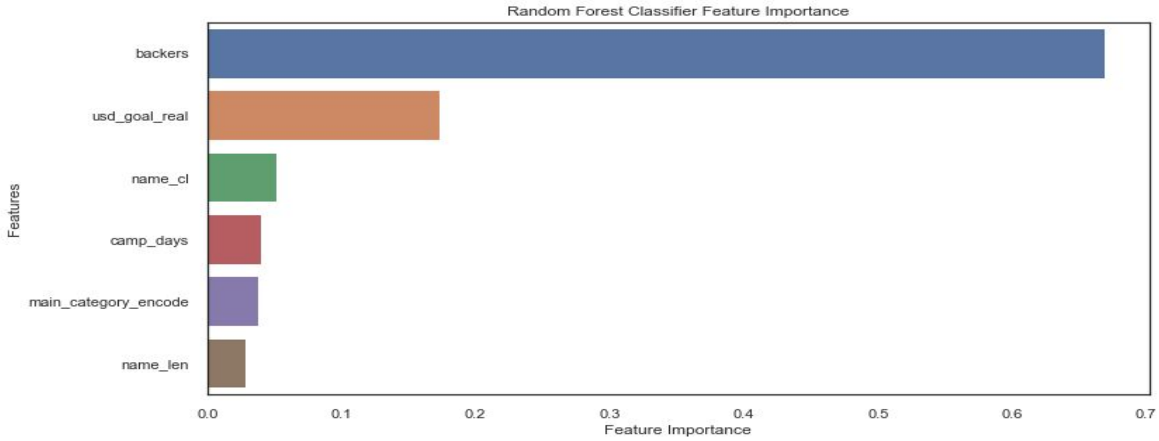
{'n_estimators': 196, 'min_samples_split': 10, 'max_leaf_nodes': 49, 'max_features': 0.7, 'max_depth': 17, 'bootstrap': True}

The most notable change from the Random Forest Classifier that was built using default parameters, and the best parameters identified by RandomSearch CV was the average number of nodes. With the default parameters, the average number of nodes was 35000 while the average maximum depth was 35. The best parameters estimate an average of 97 nodes and an average maximum depth of 8.

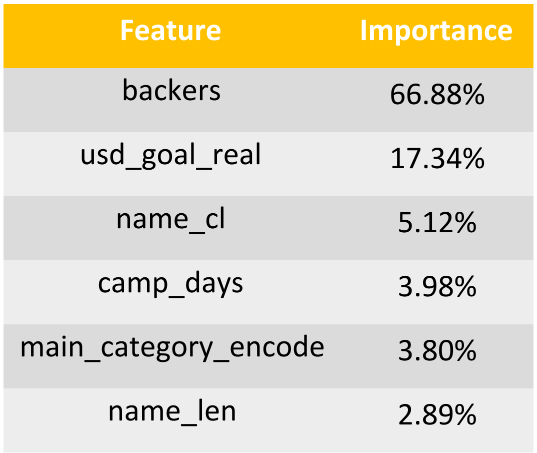
Random Forest Classifier with Optimal Parameters:



While using the best parameters set forth from RandomizedSearchCV, the Random Forest Classifier's F1 score rose from 90.49% to 90.8%. While the boost is relatively small, it accounts for a decrease in both false positives and false negatives. I chose to focus primarily on ROC_AUC scores, as they are better indicators of the models ability to distinguish between features. When the model was run with optimal parameters the ROC_AUC score dipped by 0.04% while the recall score rose from 92.1% to 92.9%. It appears that the efforts to improve the recall value effectively lowered the accuracy of other metrics.



I was surprised to find that the most important feature to the model was overwhelmingly backers at 66.88%.



Feature	Importance
backers	66.88%
usd_goal_real	17.34%
name_cl	5.12%
camp_days	3.98%
main_category_encode	3.80%
name_len	2.89%

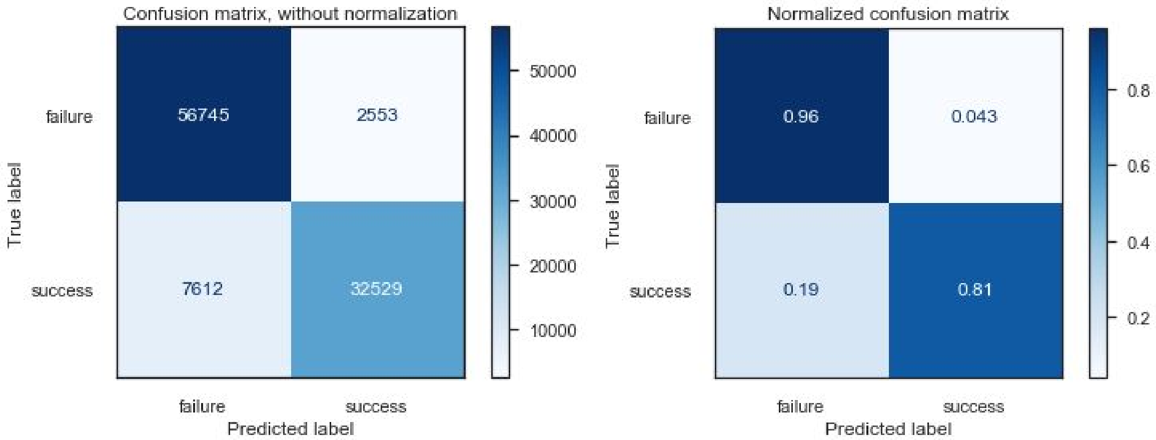
Logistic Regression

I chose to use Logistic Regression for an additional classifier model, because it can provide additional insight into the relevance of predictive features and as well as their direction of association.

In order to build my Logistic Regression model I took the same steps to set up a baseline accuracy score as I did with the Random Forest classifier. As with the Random Forest Classifier, the Logistic Regression baseline predicted all 0-failures, the majority class, accounting for a 59.6% accuracy score.

After determining the baseline classification accuracy, I ran the Logistic Regression model with default parameters:

Logistic Regression Model with Default Parameters:



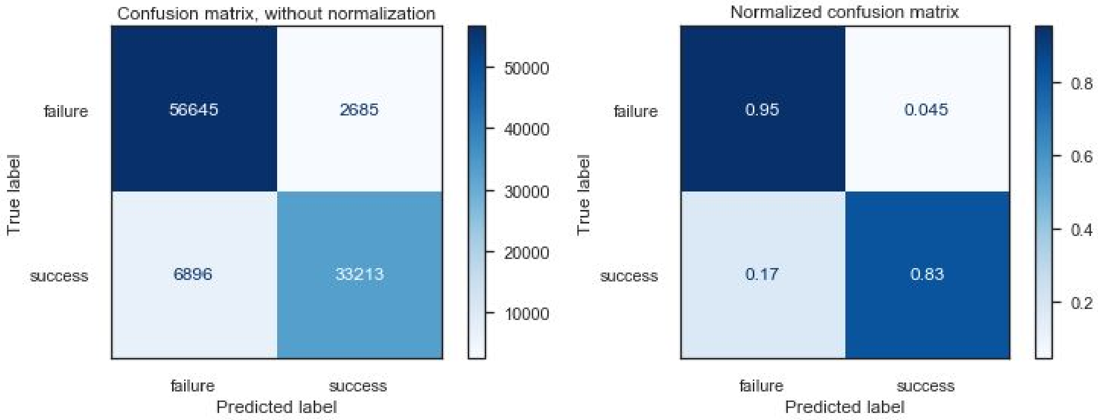
To best improve it’s accuracy, I chose GridSearchCV to tune the hyperparameters for the Logistic Regression model. I chose GridSearchCV because it is an exhaustive search option when determining the optimal hyperparameters.

GridSearchCV found that the optimal parameters would be:

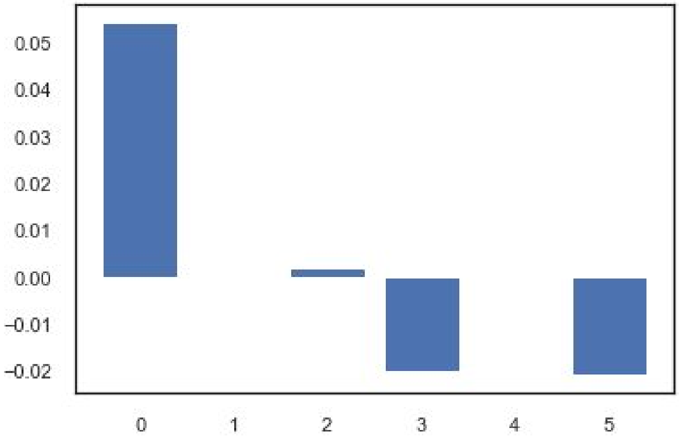
Best Penalty: l2

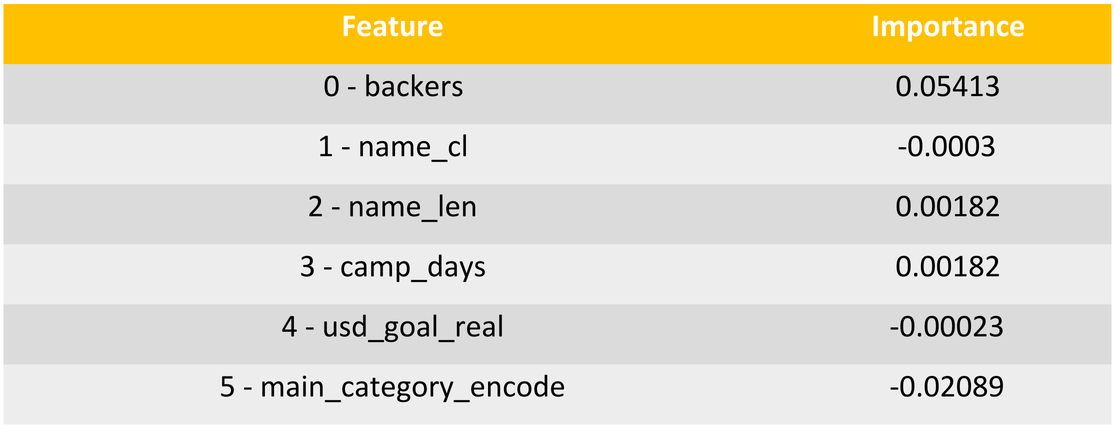
Best C: 21.544346900318832

Logistic Regression with Optimal Parameters



When the Logistic Regression model was run with optimal parameters, it came to the same accuracy as using default parameters. This will require further exploration to determine whether any features could be fine tuned in order to improve model accuracy, or if any additional parameters could be adjusted.





Feature	Importance
0 - backers	0.05413
1 - name_cl	-0.0003
2 - name_len	0.00182
3 - camp_days	0.00182
4 - usd_goal_real	-0.00023
5 - main_category_encode	-0.02089

Feature importance followed a similar pattern as it did with the Random Forest classifier, with Backers making the biggest impact.

Conclusion & Recommendations

After spending time exploring, manipulating and visualizing the Kickstarter dataset, I believe I can draw several conclusions that could help Kickstarter improve their business model for themselves and for their clients.

First, campaign categories do matter. The majority of campaign categories had higher failure rates than success rates, from examining the data this could possibly be attributed to their high goals, making it more difficult for these campaigns to reach or exceed their funding. Campaign categories such as comics, music, theater and dance tend to have lower goals on average in comparison to categories like film & video, publishing, and technology.

That leads me to the second conclusion - the goal of the campaign matters! If a campaign starts out with an incredibly ambitious goal, it will be much more difficult to reach and exceed said goal. The initial goal amount, coupled with the number of campaign backers, makes a big impact on the likelihood that a campaign will succeed.

While the length of the name and the number of characters in a name did not make a particularly large impact in the Random Forest classifier, it does appear to correlate some with campaign outcomes. The fewer number of characters in a campaign name and the fewer words in a campaign name lead to higher successful outcomes than more characters in a campaign name or more words in a campaign name. Campaign names often make the first impression of a campaign, and lengthier names appear to turn donors off.

The Kickstarter dataset that I have been working with did not include some data points that are generated from campaigns such as the campaign description, and further exploration of these additional features would be very interesting. Moving forward, i’m interested in exploring additional features and the possibilities of Natural Language Processing for the campaign names and the campaign descriptions.