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 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 of 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 or undefined campaign states. Rows containing these campaign states are removed from the analysis.
For future statistical analysis, a new column - binary_state - is created. In order to create a binary column, failed campaigns absorb cancelled and suspended campaigns as well.

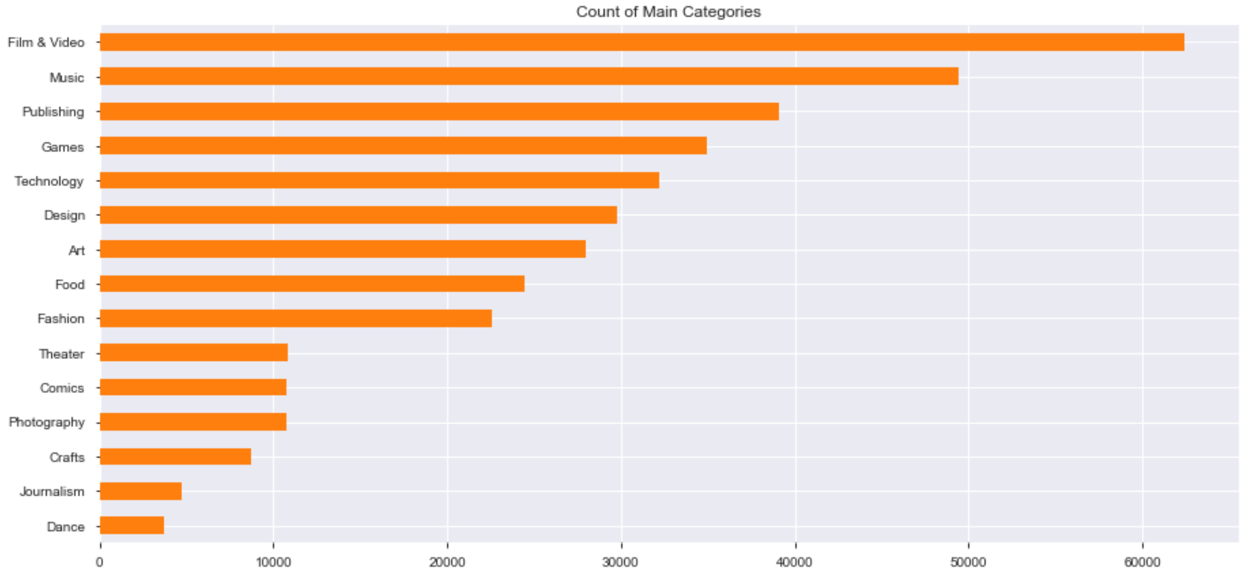
Outliers: Due to the nature of this dataset, it can be expected that some columns will contain outliers. For example, ambitious campaigns who set a very high campaign goal or campaigns that exceeded expectations and raised thousands of dollars more than expected. Most columns in the Kickstarter data 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

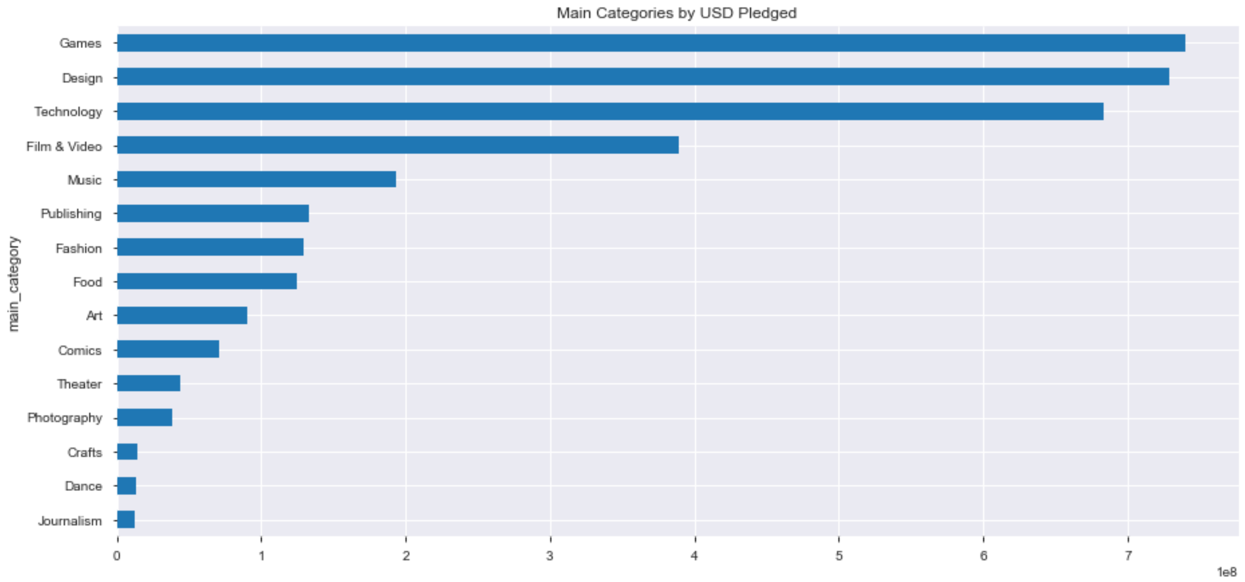
Campaign Categories

*Research Question 1***:** Is there a statistically significant relationship between campaign category and the success of a campaign?
Research Question 2: Is there a correlation between campaign categories and total USD pledged?
Hypothesis 1: There is a statistically significant relationship between campaign category and the success of a campaign.
Hypothesis 2: There is a correlation between campaign categories and total USD pledged.

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:

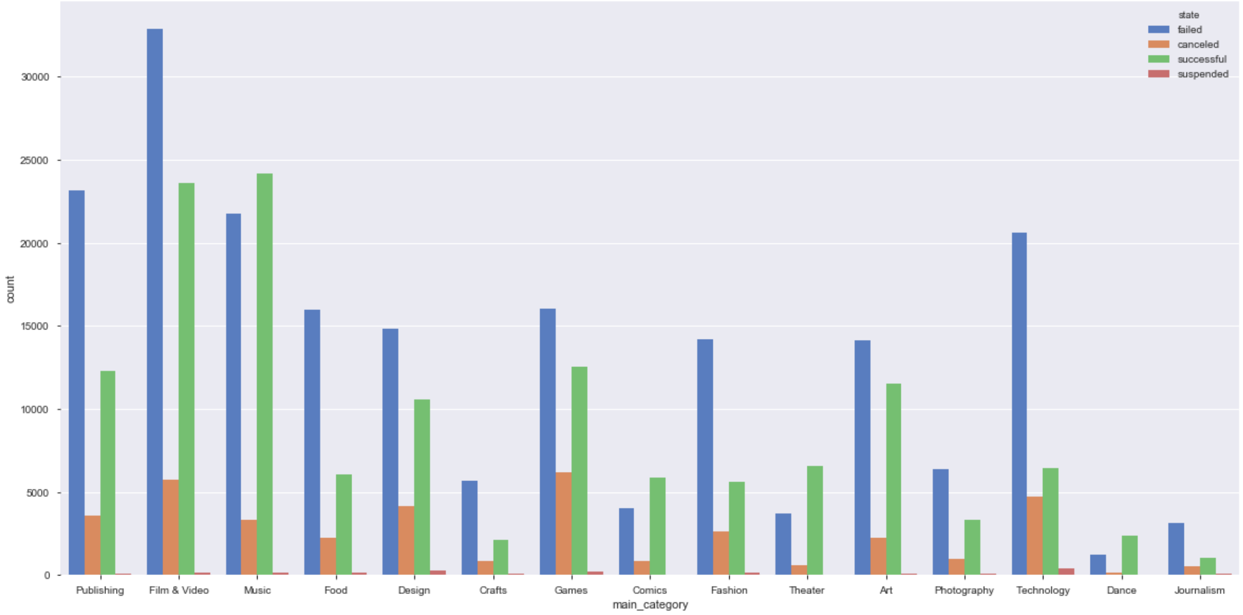


Next, I want to look at the campaigns based off of the total sum of usd_pledged_real per campaign category. When this is applied, we get the following:

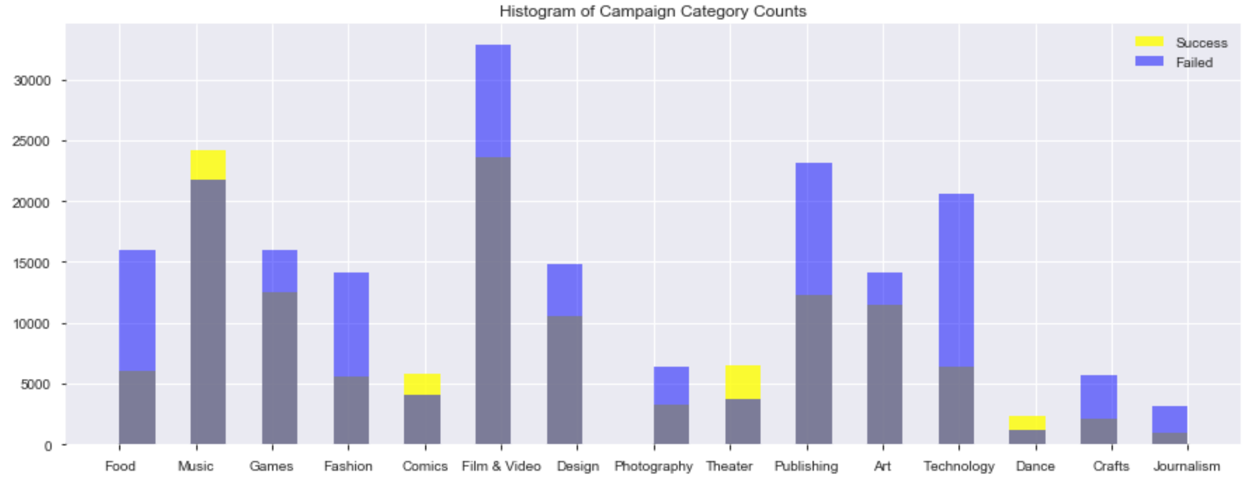


Interestingly, the main categories based on the sum of pledged USD does not line up with which campaign categories produce the most campaigns. Although games are the 4th ranked main category in terms of number of campaigns, it far outpaces Film & Video. Campaigns with the main category games brought 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!

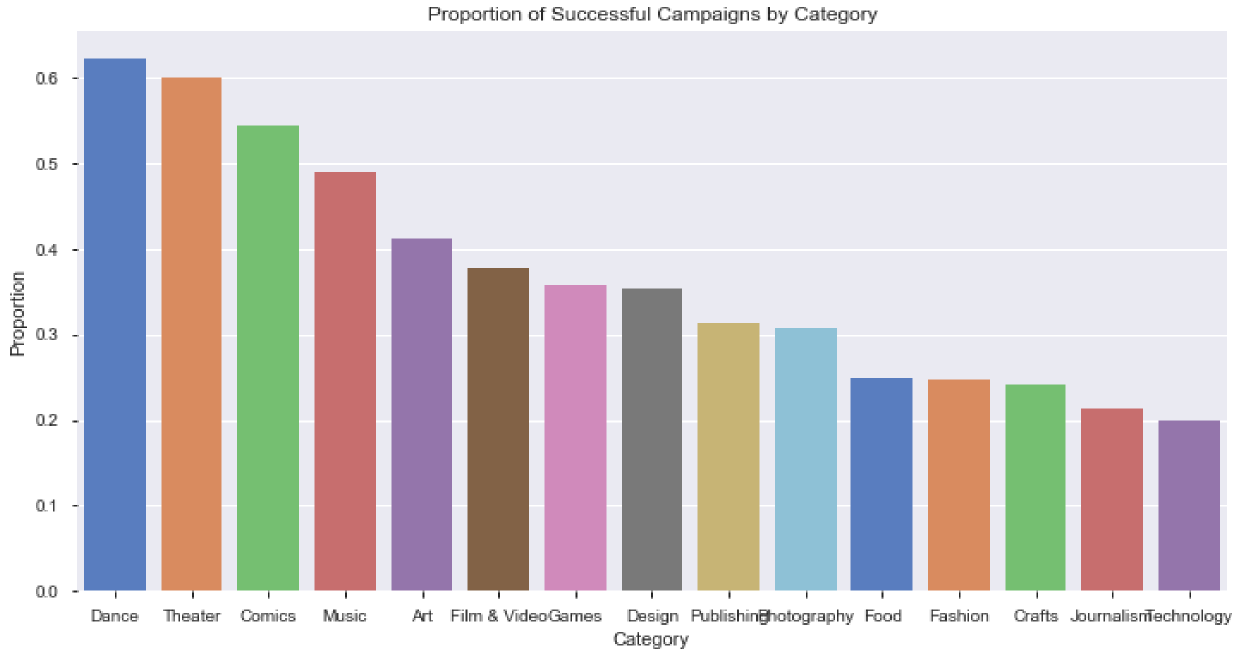
From here, I broke down each category by their campaign state to begin to look at the success rate within a campaign category. This chart is based off of the count of each campaign state within each campaign category:



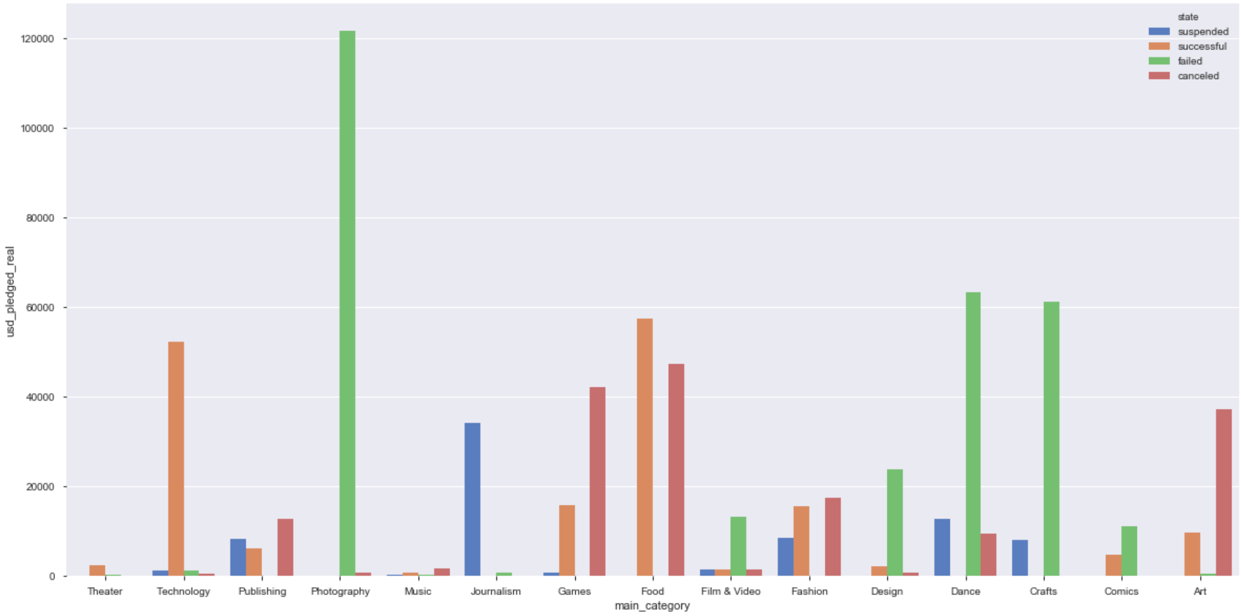
Another way to look at this more concisely is by breaking down the campaign categories by only success and failure. The other campaign statuses are minimal in comparison to success and failure, but will be taken into account when determining the overall campaign success rate.



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.

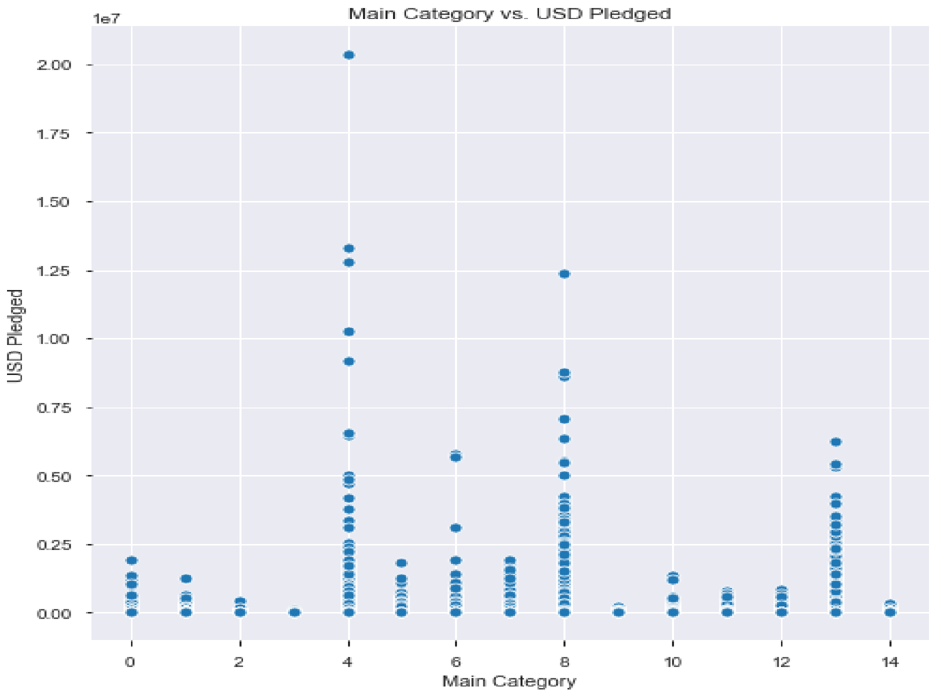


Next, I broke this down further by looking at the total sum of usd_pledged_real by the campaign state, per campaign category:



In order to test my first hypothesis, I conducted a bootstrap resampling and found the p-value of the main category in relation to the campaign state. The p-value was 0.0, which means that the null hypothesis is rejected. There is no relationship between the main category and the success of a campaign.

In order to test my second hypothesis, I created a scatter plot of the main category and USD pledged and then calculated Pearsons correlation:



The Pearsons correlation is very weak at 0.06. The null hypothesis is rejected.

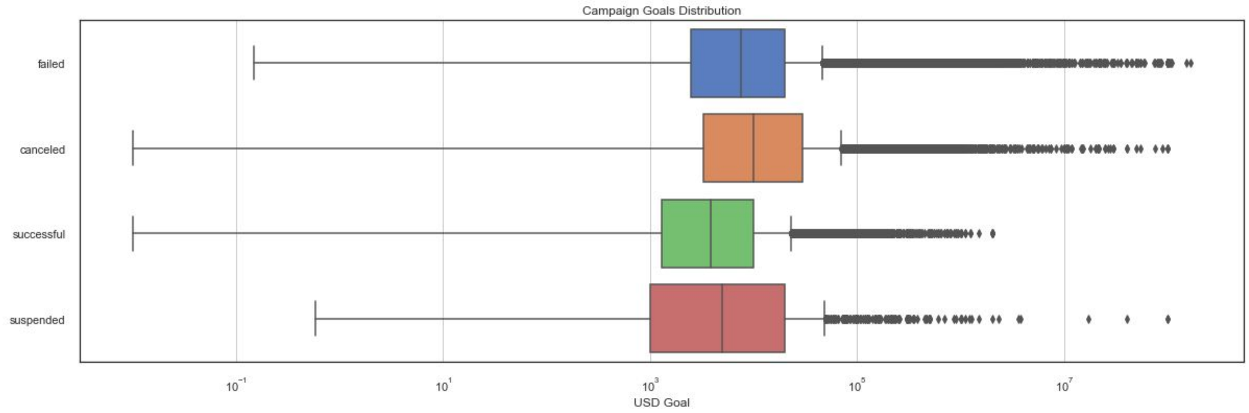
Campaign Goals

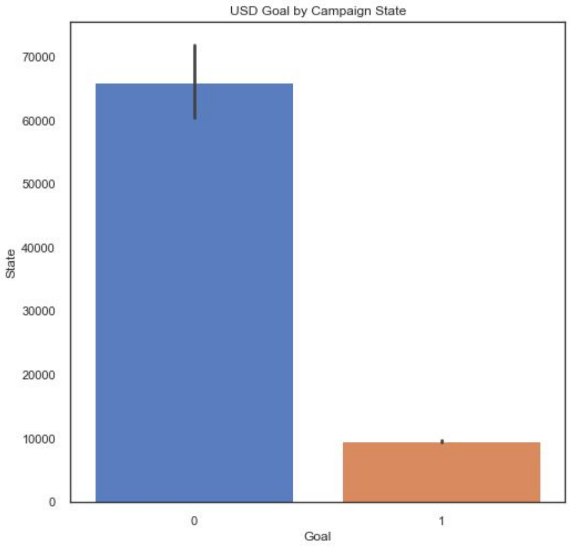
Research Question: Are goals of failed campaigns typically higher than those of successful campaigns? Are initial campaign goals too lofty for the campaign content? Leading to a higher rate of failure?

Hypothesis: Goals of failed Campaigns are higher on average than goals of successful campaigns.

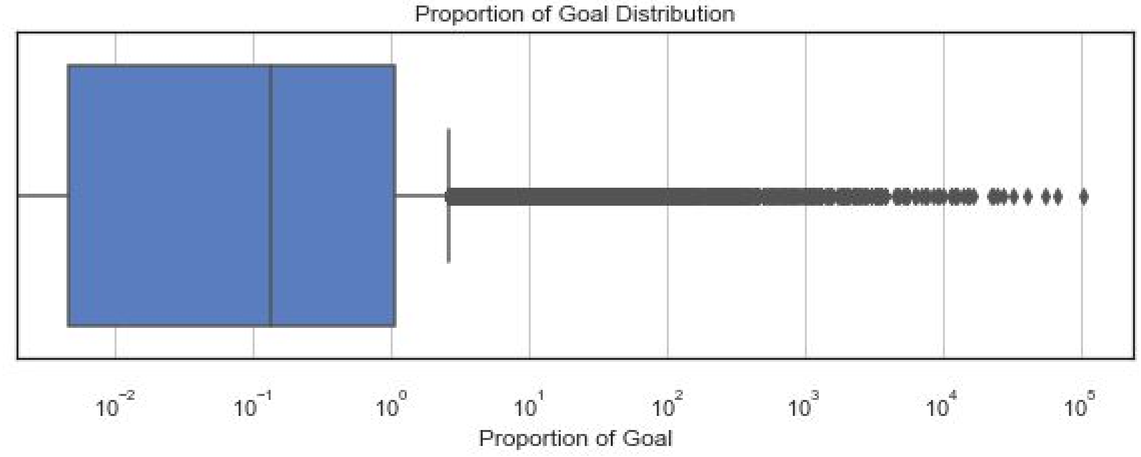
.

Kickstarter campaigns seek funding ranging 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 null hypothesis, I created a new column in the dataframe called goal_prop. This column is for the proportion of the goal to the USD pledged.



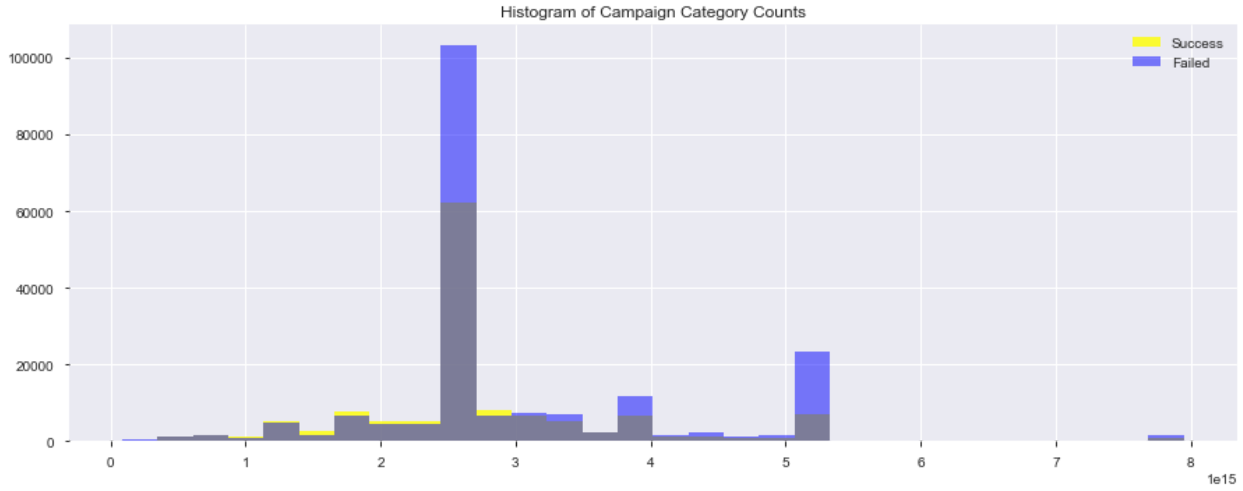
Campaign Duration

Research Question 1: Is there a statistically significant relationship between the duration of a campaign and total USD pledged?
Hypothesis 1: Longer campaigns generate a higher total of pledges.
Research Question 2: Is there a statistically significant relationship between the duration of a campaign and the number of backers?
Hypothesis 2: Longer campaigns have a high number of backers.

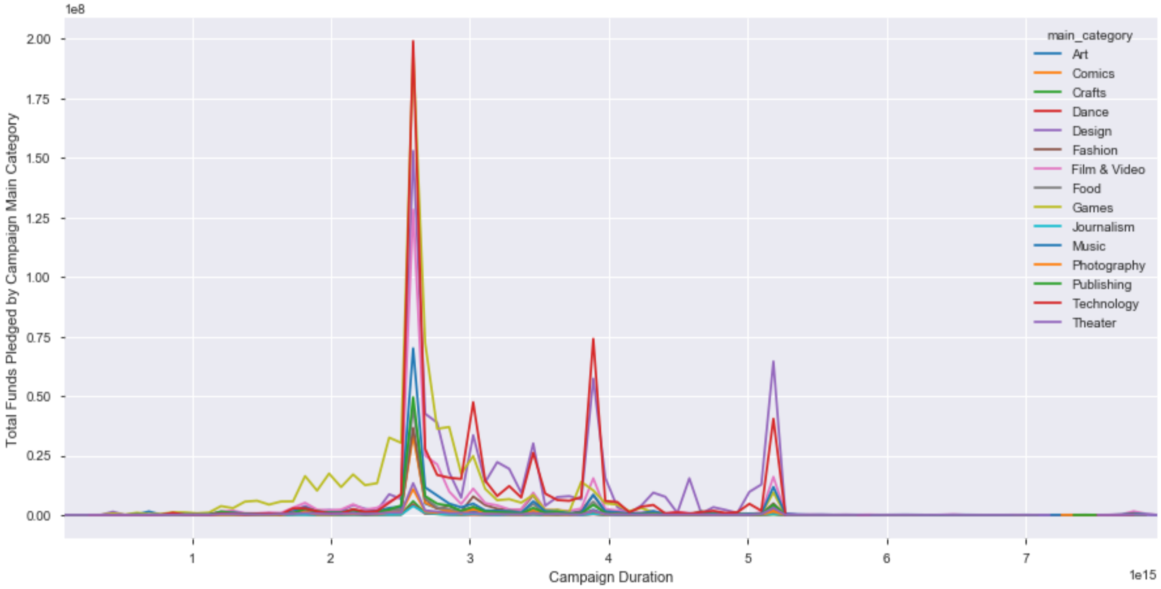
After exploring the campaign data, it is important to add a column that shows the duration of a campaign to contextualize how much time it has taken for successful campaigns to reach or surpass their funding goal, or for determining the average length of time of a failed campaign. In order to create a campaign duration column, both the launched and deadline columns are converted to datetime. From there, the campaign_duration column is 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.

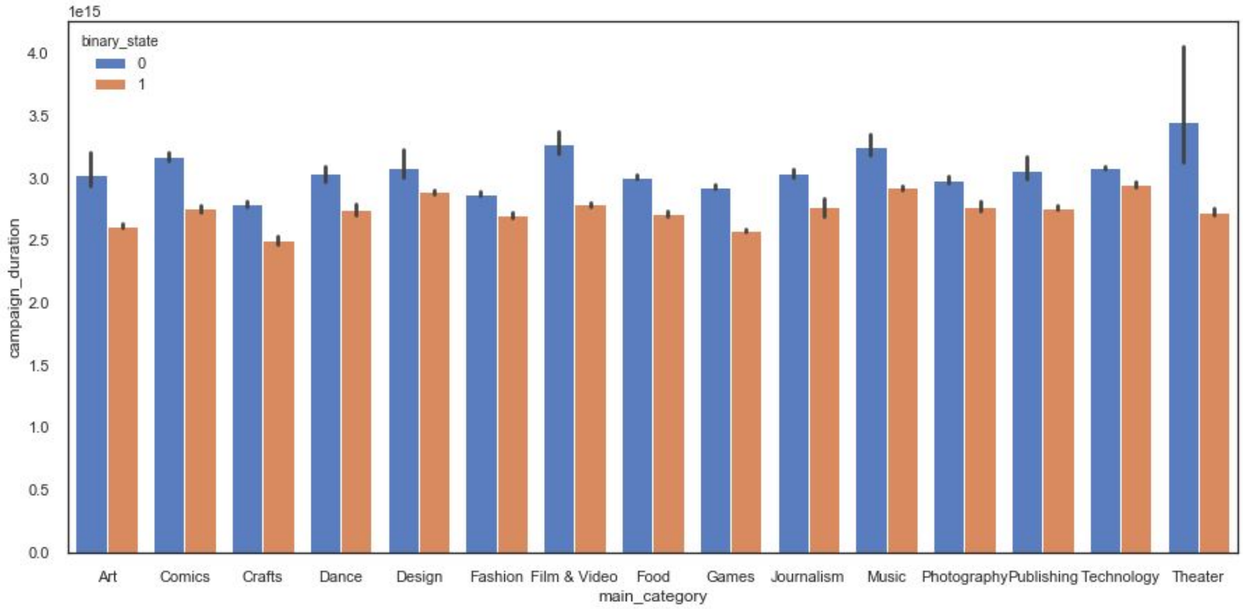
According to Kickstarter.com: “Projects on Kickstarter can last anywhere from 1 - 60 days. We've done some research, and found that projects lasting any longer are rarely successful. We recommend setting your campaign at 30 days or less. Campaigns with shorter durations have higher success rates, and 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:

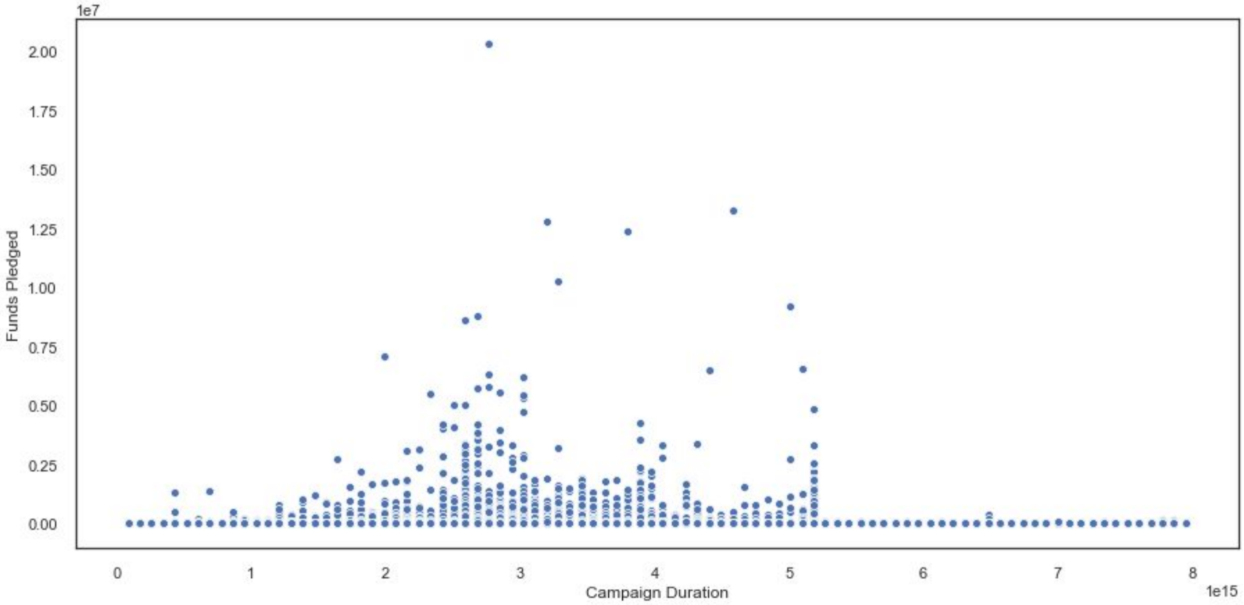


Does the duration have any correlation with USD pledged? Following the same patterns as our previous duration analysis, the campaigns with the highest USD pledged by category occur around the 30 day mark of a campaign:



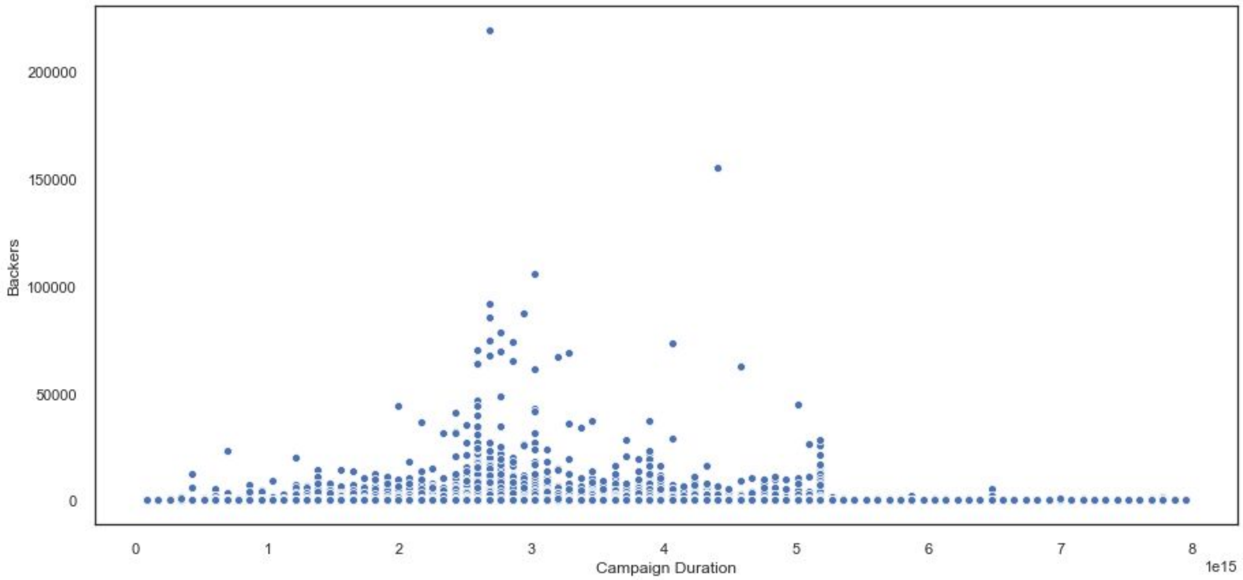


To test null hypothesis 1, a scatterplot was created and Pearsons correlation was calculated.



Pearsons correlation is 0.001 indicating no relationship between the campaign duration and the total USD pledges. The null hypothesis is rejected.

The same hypothesis test was conducted for null hypothesis 2:

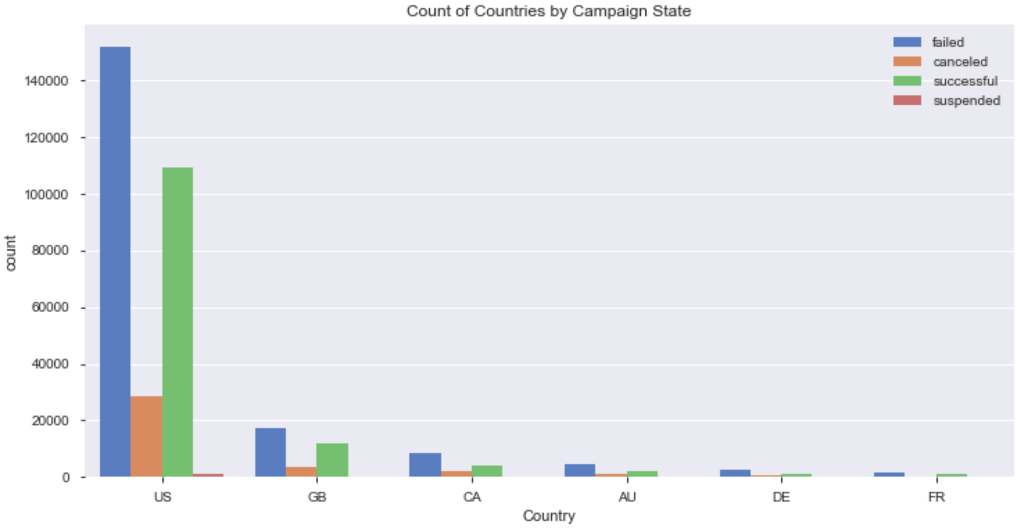


Pearsons correlation is also 0.001 for hypothesis 2, indicating no relationship between the campaign duration and number of campaign backers. The null hypothesis is rejected.

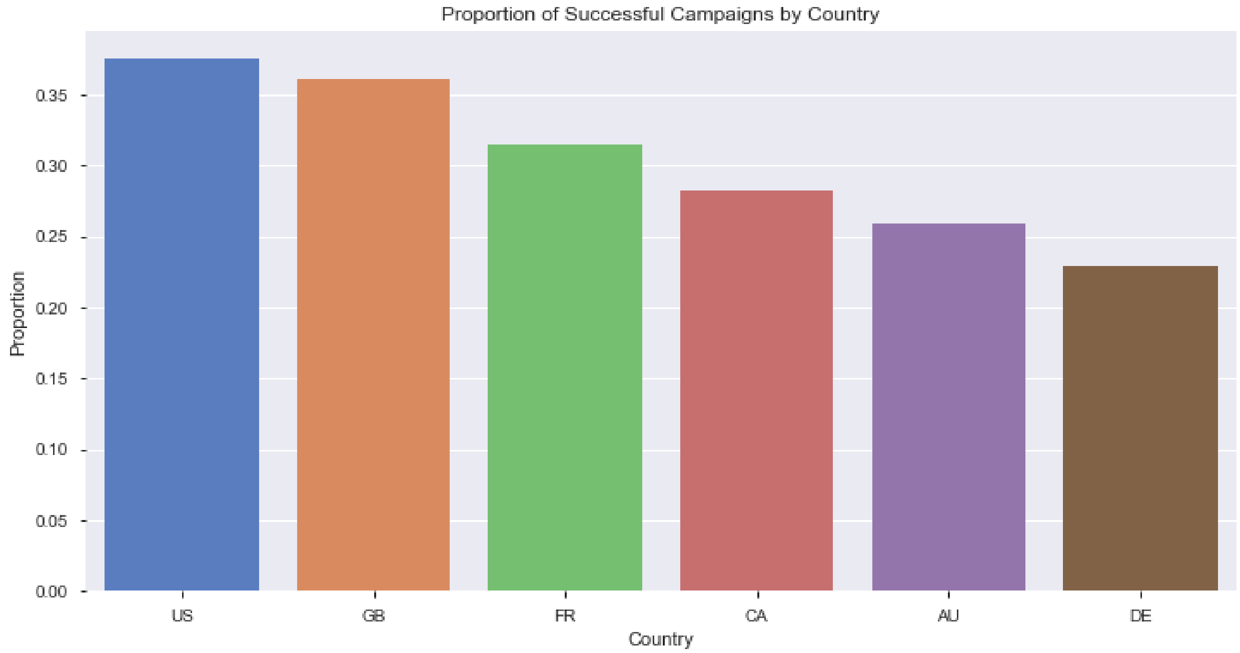
Campaign Country

Research Question: Does the country which a campaign is launched from determine the success of a campaign?
Hypothesis: If a campaign is launched from the US, it has a higher likelihood of succeeding than a campaign launched from any other 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:



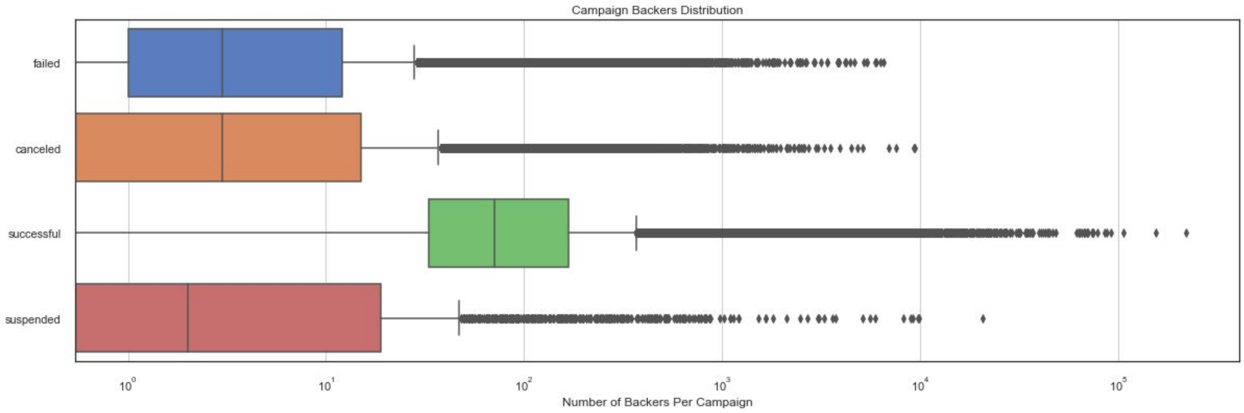
While the US has 109,299 successful campaigns, they have 152061 failures accounting for a 0.375744% success rate. Great Britain - while far behind in campaign volume - has 12067 successful campaigns and 17387 failed campaigns, accounting for a 0.361363% success rate. These two success rates are very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



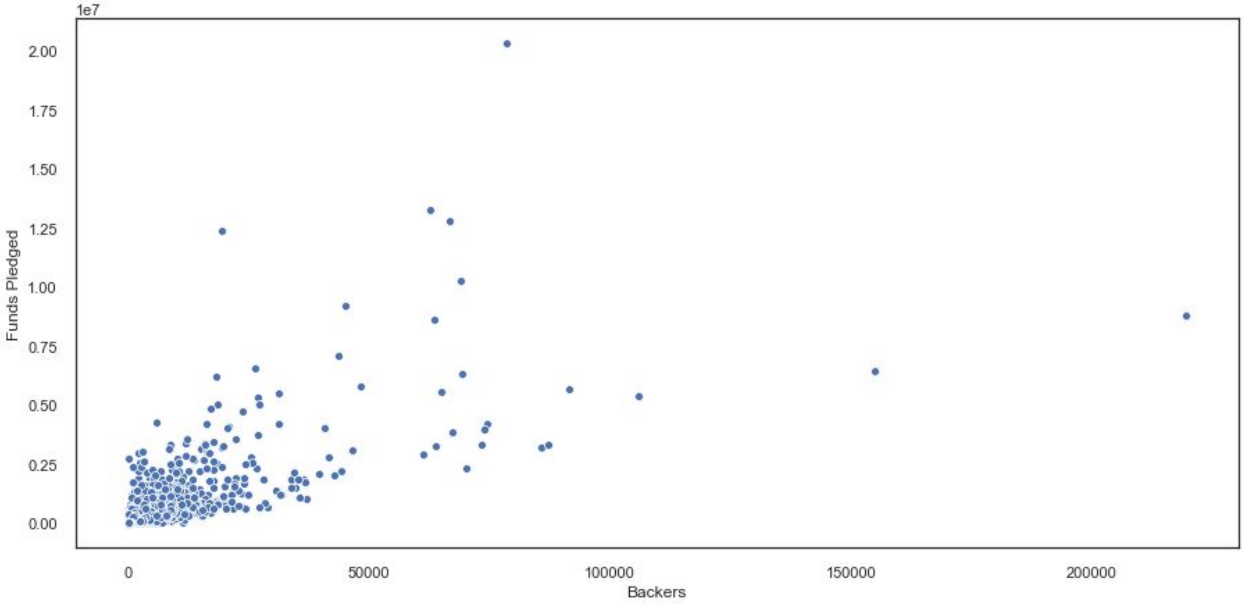
Campaign Backers

Research Question: Is there a statistically significant relationship between the number of backers per campaign and campaign success?
Hypothesis: There is a statistically significant relationship between the number of backers per campaign and campaign success.

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.



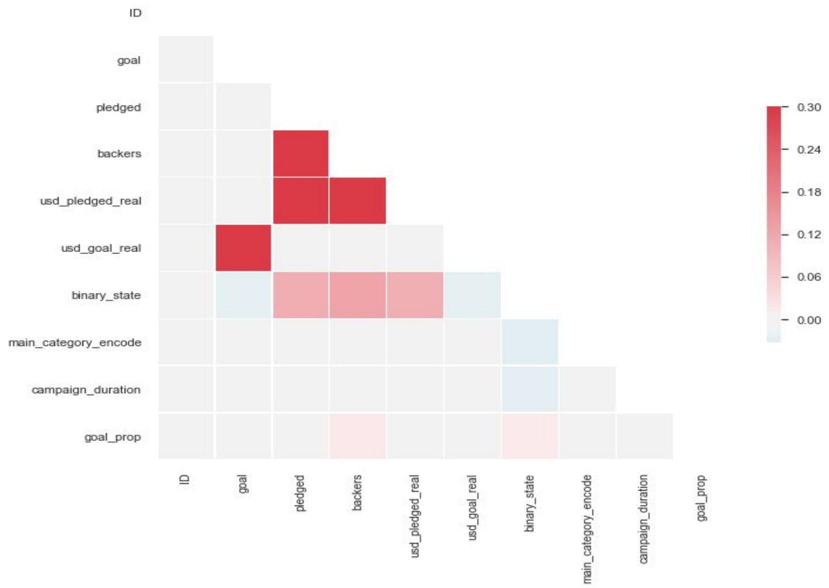
To test the null hypothesis, a scatterplot was created and the Persons correlation was calculated.



The Pearsons correlation is 0.752, making this a statistically significant relationship. The null hypothesis is accepted.

Correlations Matrix

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


ID
goal
pledged
backers
usd_pledged_real
usd_goal_real
binary_state
main_category_encode
campaign_duration
goal_prop