Final Report [Info W18]

Data Analysis on the Success of Kickstarter Campaigns

Authors: Neha Nagabothu, Ananya Gupta, Rimika Banerjee

Context:

While brainstorming project ideas with our newly formed group, we found that one thing we were all curious about was crowdfunding campaigns on sites such as Kickstarter, Indiegogo, etc. We decided that it would be interesting to look at what makes fundraising campaigns on such sites successful, and if there are certain factors that successful campaigns all exhibit. Thus, our group idea was to look at historical data on various Kickstarter campaigns, and attempt to make sense of what factors influence a campaign being funded or not. We acquired our datasets from Kaggle, an online community of open data sets. Over the past two weeks, we have worked together to conduct preliminary data analysis and even train a machine learning model using Scikit-learn to predict whether or not a Kickstarter campaign will be funded based on various factors.

Questions:

We are focusing on various factors for each campaign that we are analyzing, in order to see what makes a campaign "successful", which we defined as the campaign being fully backed and meeting its financial goal. By exploring the data, we are hoping to get insights into:

1) What should campaign creators focus on in order to achieve their financial goals, based on the data?
2) How can existing founders predict campaign success utilizing the metrics referenced to in the data set?
3) How long should a Kickstart fundraising campaign run for optimal performance?
4) When should campaigns be launched? Does seasonality influence campaign success?

Data Source:

After much digging, we ultimately decided to use Kaggle, a community of open data sets, for this project. This data was originally sourced and compiled from the Kickstarter platform. This dataset has campaign data for both 2016 and 2018.

- https://www.kaggle.com/kemical/kickstarter-projects

Dataset Structure:

This Kaggle dataset has over 300,000 entries of various fundraising campaign projects (2016 - 2018) and has the following columns of extensive data: *'ID', 'name', 'category', 'main_category', 'currency', 'deadline', 'goal', 'launched', 'pledged', 'state', 'backers', 'country', 'usd pledged'*

These are the columns of our interest:

- 'name': Name of project
- 'category'/ 'main_category': Category of project
- 'currency': Currency used for funds
- 'deadline' :Deadline for crowdfunding
- 'goal': The amount of money the creator seeks for the project
- 'launched': Date the campaign was launched
- 'pledged': Amount received by the crowd
- 'state': State of project (ie. Failed, Successful, Other)
- 'backers': Number of backers
- 'country': Country pledged from - a percentage breakdown
- 'usd pledged': USD pledged - amount

Additional Dataset:

https://www.kaggle.com/socathie/kickstarter-project-statistics#most_backed.csv

- This dataset focuses on the most backed projects, and provides insight into factors influencing highly successful campaigns

Initial Read of the Main Dataset - What is Relevant?:

Some of the important factors from the dataset are category, main category, campaign deadline, funding goal, date launched, amount pledged, current state the project is in, number of backers, and country pledged from.

The campaign deadline and date launched are also important because it considers:

1. The exact time in the year it was launched (seasonality, etc)
2. The actual duration of the campaign

Apart from seasonality, we would like to determine the ideal campaign length, and understand how deviating from that ideal length would influence a Kickstarter campaign.

Amount pledged is also important - maybe certain factors lead to campaigns being only 50% or even significantly overfunded! It also gives an indicator of whether the funding goal was reached or not, and is a strong metric of success for the campaign.

Data Cleaning and Performing Sanity Checks:

Before starting our analysis, we had to conduct a sanity check and clean up our data to ensure we are doing our due diligence.

Understanding the Data:

We used the info() function to see how pandas interpreted the data type of each column in our dataset. Most of its interpretation was correct, except it interpreted the ID column as an integer rather than a string, and so we set the datatype of the ID as a string to avoid misinterpretation.

Dropping Irrelevant Columns:

We dropped the USD pledged column and the goal column because the dataset author included columns that were not converted to USD or improperly converted to USD. We also dropped the ID column since it is not necessary for our data analysis.

Parsing and adding relevant columns:

We added the 'name_length' column by getting the length of the string values in 'name.' We also obtained the 'campaign_length' column by subtracting the 'launched' date from the 'deadline.' We parsed these columns and converted them to string values, so we could utilize the Python datetime module, and we also added 'campaign_start_month' and 'campaign_start_szn'.

Handling Null Values:

One of the columns with a significant number of null values was dropped entirely (it had 3797 null values). The other column with null values was "name", and we utilized the fillna() function to repopulate blank values with "NA". After these operations, we ran isnull().sum again and found that there were no null values in the dataset.

Machine Learning Data Transformations:

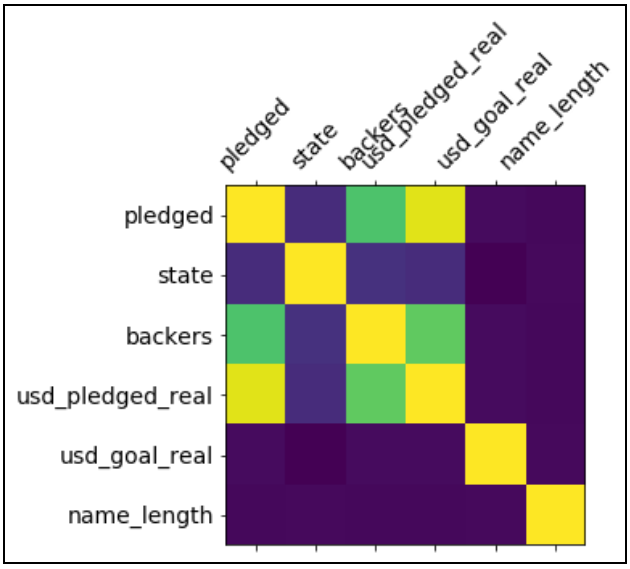
Transforming Numerical Data:

For the numerical data, after replacing all of the null values, we utilized the MinMaxScaler. This is valuable because if a deal amount is in the millions, but other values are binary or smaller (e.g: number of backers definitely will not exceed one million), we will not weight one column more than the other. The MinMaxScaler takes the highest value of the column and the lowest value of the column and scales it down to a range between 0 and 1.

Transforming Categorical Data:

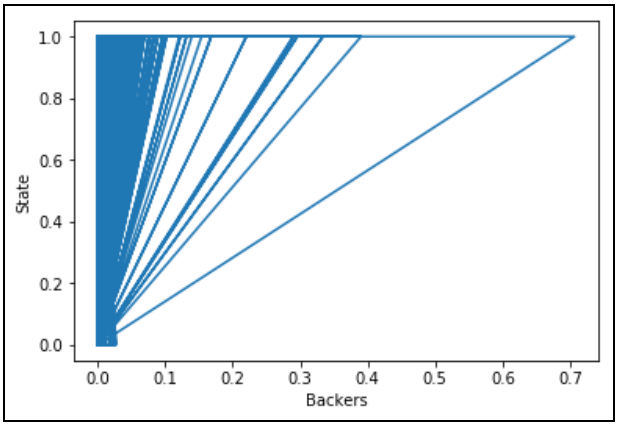
For categorical data, we needed to transform everything to numerical data. Originally, we thought that the best methodology would be to take numbers to represent each category (for example, Technology = 1, Film = 2, etc). However, after further research and testing, we realized that this would cause the data to be interpreted as ordinal (e.g. a certain number is better than another number) rather than nominal. Therefore, we decided to use the technique known as one-hot encoding. One-hot encoding takes a single categorical column, like "main_category", and converts it into many columns. For example, if "main_category" had the values Technology, Film, and Games, it would convert it into three columns known as "main_category_Technology", "main_category_Film", and "main_category_Games", with true or false values representing if the row has that attribute or not.

Question: What should campaign creators focus on in order to achieve their financial goals, based on the data?



We started out by conducting some exploratory analysis. To make sense of all the factors presented, we wanted to judge feature importance. The first method we utilized to determine which factors were important, and as a result, which factors campaign creators should focus on, was creating a Correlation Matrix. In the matrix above, you can easily view how different factors affect one another, illustrating correlations between variables. Lighter colors represent high correlation, indicating more linear association. Darker colors represent low correlation between the variables. These findings are based on the data, and not a larger population.

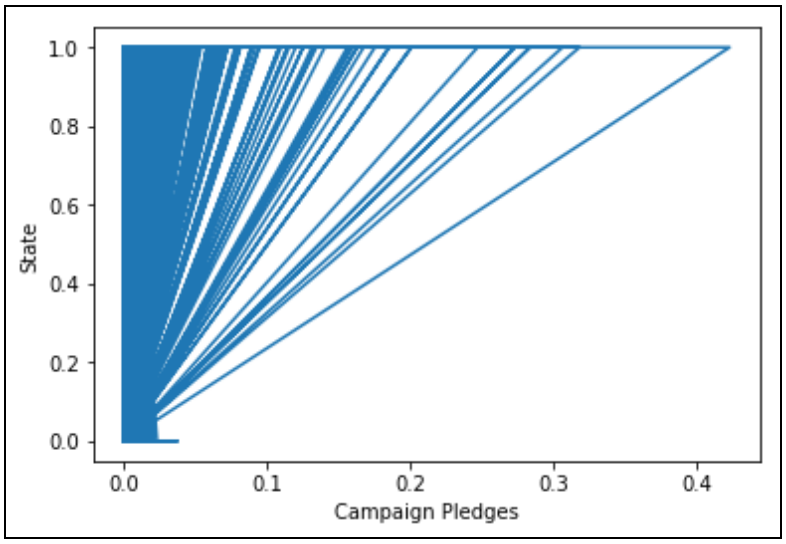
Number of Backers vs. State of Campaign:



The first analysis we performed was examining how the number of backers influences the state of the campaign ("*state*" meaning current condition the project is in - failed, successful, canceled). Going into this analysis, we thought about how the number of backers would actually affect whether a campaign is successful or not. We considered that larger numbers of backers

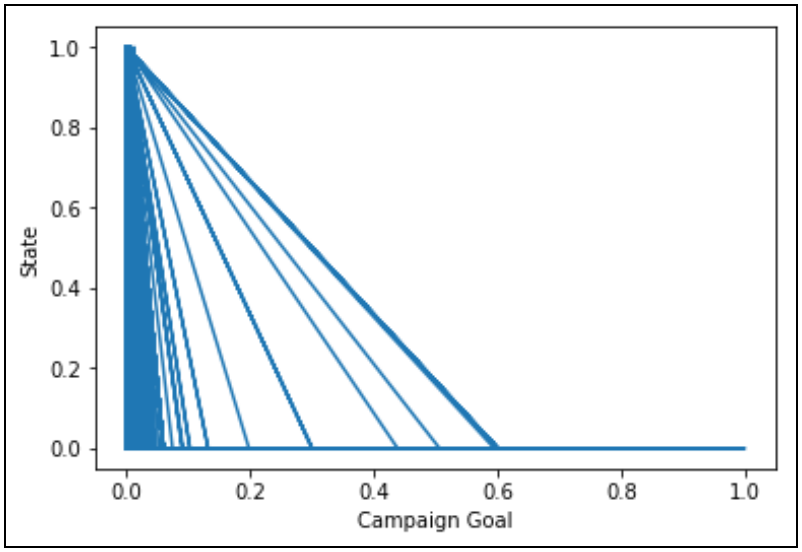
would mean a higher chance of success in most cases. However, we also thought about other influencing factors. For example, campaigns such as technology with higher investments would succeed with a smaller number of backers. Additionally, the campaign goal would significantly influence the number of backers needed to be successful. A creator with a campaign goal of only $1000 will not need a significant number of backers, while a creator with a campaign goal of $60,000 will definitely need several backers. Interestingly enough, the outlier for the number of backers (the rightmost line in the chart) shows that to reach success, that campaign needed a very large number of backers to be successful. Looking at the slope of the graph, we can hypothesize that not many campaigns had what it took to reach success at a presumably high goal amount.

Amount Pledged vs. State of Campaign



The second analysis we performed was examining how the amount pledged influences the state of the campaign ("*state*" meaning current condition the project is in - failed, successful, canceled). Similar to the graph above, this analysis shows one strong outlier (the rightmost line) which illustrates that to reach success, that campaign needed very high amounts pledged to reach success. The fact that this is an outlier seems counterintuitive - we had initially thought that it makes sense for a successful campaign to require relatively consistent high donations. However, this graph could also be indicating that a large portion of the Kickstarter campaigns didn't have such high funding goals, and thus, were able to reach success with relatively lower amounts pledged.

Campaign Goal vs. State of Campaign



The third analysis we performed was examining how the campaign goal value influences the state of the campaign ("state" meaning current condition the project is in - failed, successful, canceled). Intuitively, we believed that a higher campaign goal would probably make it less likely that the campaign was successful since a larger amount of money would be required. While this is of course not the only factor that would determine success, we believed this could definitely be a strong feature to take into consideration when predicting success. As we can see in the graph, our hypothesis seems to be true as the outliers on the far right and the general downward-sloping nature of the plot indicate that as campaign goals got higher, the states of campaigns tended towards zero (failure).

Question: How can existing founders predict campaign success utilizing the metrics referenced to in the data set?

Machine Learning Model:

We utilized the scikit learn package and the models XGBoost to create a predictive model of the potential success of Kickstarter campaigns.

After appropriately cleaning and transforming the data, we changed the state attribute into a binary attribute (successful or not successful). Next, we separated the dataset into Y (the target variable) and X. We randomly split the dataset into training and testing, using a test size of 33%. After running the predictions, we had an accuracy of 98.61% within the 2018 dataset for Kickstarter. The dataset is 378661 rows and has 221 columns, so there is a large amount of data to draw from to train the model.

Utilizing XGBoost, we were able to get an accuracy of 98.61% in predicting whether campaigns in the testing set would succeed or not. However, although this was very accurate in making

predictions, we wanted to get some more insight into how the model was making these decisions and apply them to understanding what ideal values that founders could focus on were.

The first value in each box of the tree is the attribute. The machine learning model is making decisions based on this attribute. If the attribute is true, it will move to the left, and if the attribute is false, it will move towards the right. In general, decision tree algorithms look at the amount of information gained when creating the tree. The algorithm hopes to maximize the amount of information obtained, and as a result, the first attribute it will look at is the one with the highest "information gain". From this, we can understand that the most significant measure of campaign success is the number of backers, followed by the goal amount and pledges. The second value in each box of the tree is the entropy value. The entropy value tells us about the amount of impurity or uncertainty in each group. In the first node of the tree, the entropy value is very high, and as the tree gains information and is able to group the dataset into different subsets, the entropy value decreases. The next value in each box of the tree is "samples". This value tells us how many samples are in each "subset" of the data, node by node. Adding up the boxes in line in each node will give you the original total number of "samples", or number of values in the data. The next piece of information is "values", which gives you the shape of the dataset or subset. The final piece of information is class, a binary attribute of either a 0 or 1. When this attribute is 1, it indicates a successful campaign, and when this attribute is 0, it indicates a failed campaign.

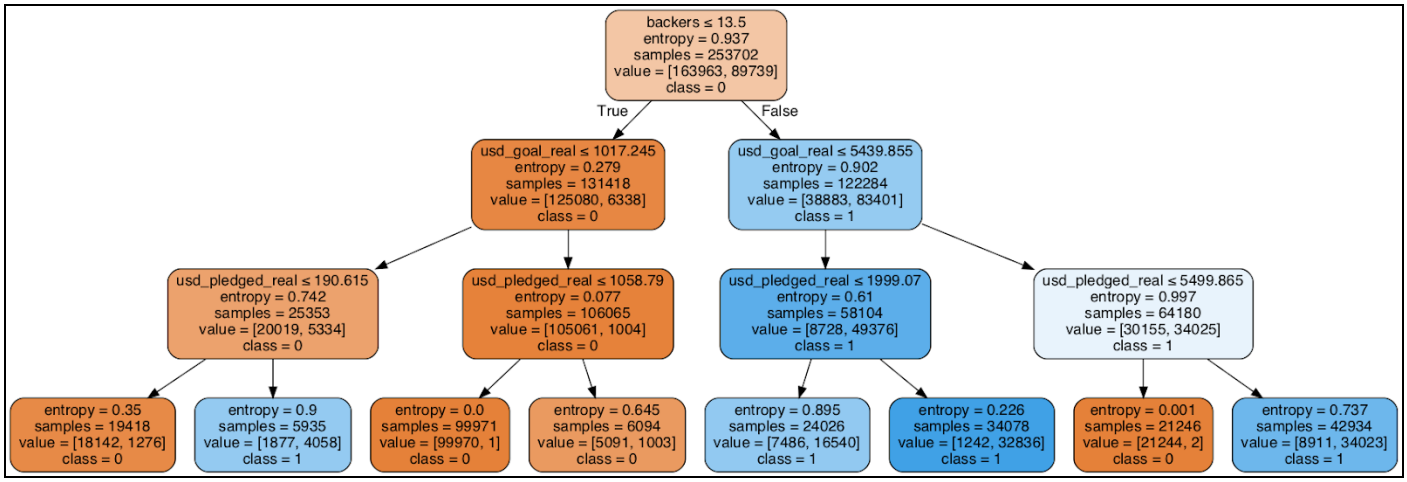
Based on the information provided by the two trees below, a campaign creator's #1 focus should be on getting backers. The creator's #2 focus should be setting a goal amount that makes sense in the context of their product, category, audience, and other factors that may not have been considered in this dataset. The creator's #3 focus should be on getting high pledge amounts.

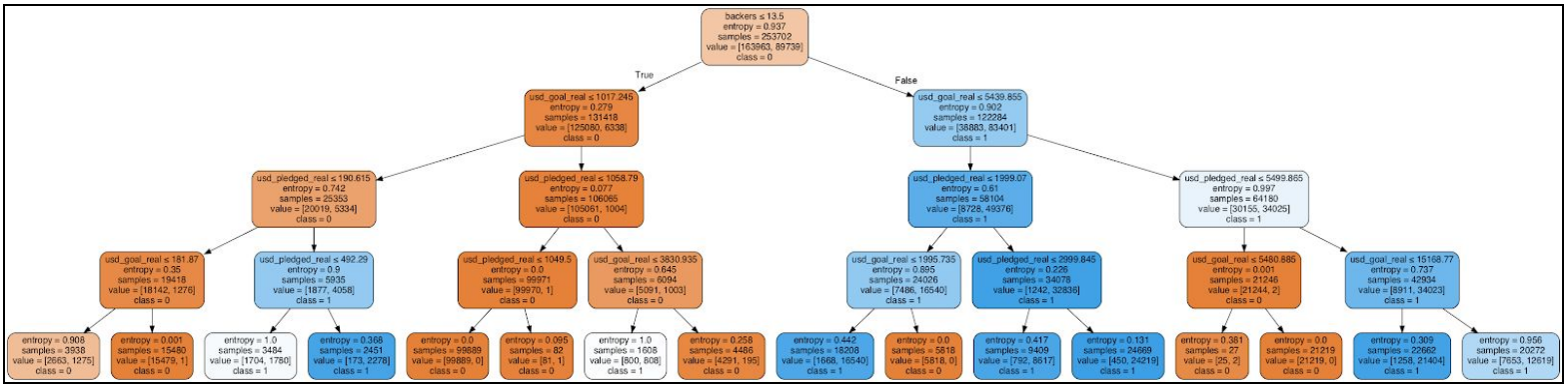
The first focus makes sense - backers are what support the creator's campaign. No matter what your campaign is or what your attributes are, backers are a clear sign of external validation to demonstrate the success of your campaign. Getting more backers is correlated with the number #3 focus of increasing pledge amounts, and it makes sense because each individual backer brings a creator closer to funding their final campaign goal.

The second focus also makes sense. Backers can affect how close you are to your goal amount, but what you set your goal amount at really influences whether you will reach your goal or not. Setting your goal extremely high can lead to an unsuccessful campaign, and setting a lower goal can make it easier to reach your campaign.

The third focus is also very important. Studies have shown that the amount already pledged can influence speculative backers, and the pledge amount also is a measure of how easy it is for a creator to reach their goal "in steps". It also influences the number of backers a creator would need to reach their individual campaign goal.

Limitations of the following decision tree - for readability, we pruned the tree to depths of 3 and 4 respectively, and there is a lot more data that could be analyzed. However, we are able to obtain a lot of information about feature significance and optimal values from the following decision tree.





Confusion Matrix

	Negative	Positive
True	79212 (tn)	44009 (tp)
False	208 (fn)	1529 (fp)

For context, here is a confusion matrix for the XGBoost machine learning model. 79212 values were “true negatives”, meaning that they were campaigns predicted to fail by the model and campaigns that DID fail. 44009 values were “true positives”, meaning that they were campaigns predicted to succeed and campaigns that DID succeed. There are 208 false negatives, meaning that they were campaigns predicted to fail but actually succeeded. There are 1429 false

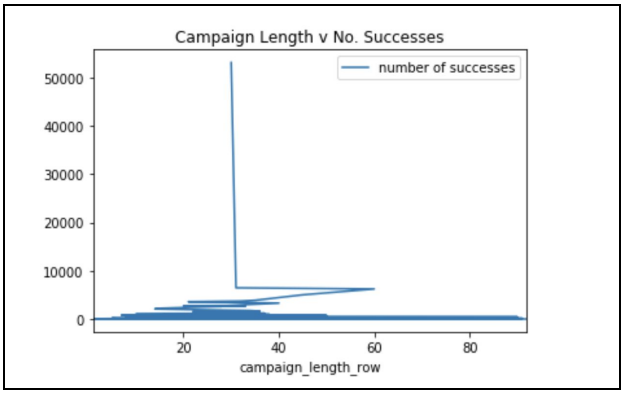
positives, meaning that they were campaigns predicted to succeed but actually failed. Overall, the accuracy is very high in predicting whether campaigns will succeed or not. Interestingly, in such a large dataset, there are only 208 values that were predicted to fail but actually succeeded, making it seem like the failure conditions are pretty set in stone. From this, we can hypothesize that these three numbers - backer count, goal amount, and amount pledged - are really what predict a campaign's success.

Question: How long should a Kickstart fundraising campaign run?

Campaign length is a difficult attribute for creators to decide on. In fact, most creators think that the longer their campaign is, the more successful it will be. However, this is not the case. Kickstarters are often funded because of the urgency of the campaign, and having a longer campaign that might not be getting funded removes that sense of urgency for backers. In fact, the most backing and pledging happens at the beginning of the campaign and the end of the campaign. We will hypothesize that a shorter to medium length will be most successful.

When researching our project, we looked at ideal fundraising campaign lengths. Kickstarter recommends that campaign lengths are about 30 days, so our hypothesis going into this analysis is that the average or even "ideal" length will be 30 days. This could be due to the fact that Kickstarter explicitly recommends this number, or that based on other confounding factors, 30 is the ideal length for any campaign. Based on both of these sources of information, we think that the ideal campaign length will be somewhere from 15 to 30 days, most likely exactly 30 days.

campaign_length	number of successes
30	53081
31	6458
60	6223
45	5009
35	3875



The above table to the left confirms our hypothesis that 30 days is the ideal campaign length. This table was generated after filtering out for only successful campaigns. We can conclude that 30 day campaigns are 8 times more likely to be successful than the amount of campaigns that are even a day longer. The above line plot to the right demonstrates an intense peak when campaign length is 30. This number might be skewed high because Kickstarter recommends 30 days as a goal which influences people to set their deadline sooner. 30 days can be noted as ideal because a shorter campaign length allows one to capitalize on momentum and heighten the sense of urgency for contributors.

	campaign_length
mean	32.156447
25%	30.000000
50%	30.000000
75%	34.000000

Length	Goal
30	9335.518821
31	12963.605748
60	10237.238801
45	13953.813951
35	14166.404183

The above table to the left tells us that the median is 30 days and at least 25% or more of the lengths of successful campaigns are 30 days. It is important to note what funding goal is in the ball-park for 30-day campaigns in the table above to the right. The average goal for successful 30 day campaigns is just under $10,000. The difference between the average goal for successful 60 day campaigns and successful 30 day campaigns is only around $900. From these results we can conclude that doubling the length of a campaign does not necessarily yield larger amounts or increase the success rate. This points to the general trend that Kickstart is attempting to promote: shorter is better. Kickstart has recently capped the maximum length of campaigns from 60 to 90 days.

Question: When should campaigns be launched? Does seasonality influence campaign success?

campaign_start_szn	number of successes
Spring	37006
Fall	35351
Summer	34274
Winter	27325

category	number of successes
Tabletop Games	2110
Product Design	1940
Shorts	1797
Music	1791
Documentary	1726
Theater	1436
Art	1017
Food	991
Rock	986
Indie Rock	983

The above table to the right demonstrates the number of successes divided per season. Spring is the most popular of launch season of successful campaigns. Spring campaigns are 35.4% more likely to be successful than campaigns launched in winter. We hypothesize that people are generally preoccupied during the winter with the holidays. Spring is characterized by new beginnings and fresh energy which leads to more ideas being launched, more excited

contributors, and more enthusiasm overall. Furthermore, if we glance at the most popular categories for successful campaigns in spring, we see that categories suited to the season perform really well. Shorts are probably popular because of the rising temperatures. Music/Rock/Indie Rock and Food might be especially popular in spring because of the rise of music festivals, concerts, and outdoor events.

Project Challenges:

We admittedly had a number of challenges throughout the duration of this project. First, it was tough to find data regarding crowdfunding campaign outcomes, that included very detailed information. We were looking for specifics such as duration of campaign, deadline, category of campaign, date launched, amount received in the end, etc. After a lot of digging, we finally found a few very extensive datasets on Kaggle.

Another challenge we faced was the dataset itself. It was fairly large, containing very detailed data on thousands of different crowdfunding campaigns, ranging from a wide variety of "categories". Thus, it was difficult to clean the dataset, and then determine what key insights to pull and analyze through graphs. Furthermore, there were several null values, making it even tougher to analyze and extract key insights.

Limitations:

When we began this study of Kickstarter campaigns, we hoped to learn more about factors influencing campaign success. We realize, however, that our analysis is confined to the sample of data that we have. Thus, we cannot make statements about the population that the data comes from. Throughout this paper, our analysis and inferences are based solely on the data.

Conclusions:

We have concluded that the top three factors influencing a campaign's state are number of backers, goal amount, and pledged amount. The ideal campaign length is around thirty days, and results for campaign creators will vary by category.

Another important conclusion we drew was that Spring campaigns have a higher success rate than Winter campaigns, based on the data. Perhaps this is a reflection of the spending habits of people - spending tends to go up in the Winter (due to reasons such as Black Friday and/or the Holidays), and down at the start of the new year.

Taking these learnings into consideration, campaign creators can prioritize/shift their efforts towards emulating the trends we drew from campaigns that reached their funding goals.

Potential Further Research Areas:

For this project, we defined a Kickstarter campaign's "success" as reaching its funding goal within the given time period. It is important to note that "success" can be defined differently. Perhaps another way to look at a campaign's success, is by seeing if the project actually launched or became a tangible product. Having data and research on this would allow campaign contributors to first see the probability of the project to launch, before donating to the Kickstarter campaign.

Moreover, there are other factors that could impact the probability of a campaign reaching its funding goals. Thus, it could be useful to analyze factors even beyond the ones we used in our project. One glaring factor that can potentially hold a lot of influence on the success of a campaign is the background of the campaign creator - specifically, his/her network and reputation. We can safely hypothesize that if the creator of the Kickstarter project has a more affluent network and more connections, he/she has a higher chance of receiving larger donations, thereby helping the creator reach the project's funding goal.