Kickstarter Campaign Analysis

STAT 605: R for Data Science — Spring 2019

Jessica Chen, Jake Flores, Andrew Mike, Matthew Mutammara

April 22, 2019

1 Introduction

1.1 What is Kickstarter?

Kickstarter is a web-based platform that began in 2009 with the goal of promoting public innovation and entrepreneurship. The platform allows the public to create campaigns for their ideas or product for others in the community to fund and purchase. Kickstarter is known as one the first crowdfunding websites to be created, with several other companies creating their own versions in the years to come such as IndieGoGo and RocketHub. Campaigns have evolved to include over 15 different categories, including technology, artwork, books, comics and videogames. The campaign creator can decide their project funding goals and timeline for the public to fulfill those goals. Additionally, Kickstarter has opened their platform across the globe to include several high GDP producing countries such as United Kingdom, Germany, Japan, Canada, Denmark, Spain, Italy Singapore and Hong Kong.

Kickstarter makes a profit by charging a 5% fee for the total amount of funds raised but does not claim ownership rights over any of the products promoted on their site. Kickstarter maintains an 'all or nothing' rule for all campaigns, meaning in order for a campaign to earn any of the funds raised, they must reach the goal within the set time frame. If they fail to fundraise to their goal amount within the time frame, then there is no change of currency between the public backers and the campaign creator. To date Kickstarter has reported over $4 billion in total pledges, with a success rate of 40% in 2015.

1.2 Our Data

We gathered our primary data source form www.WebRobots.io/kickstarter-datasets with a timespan from January 2018 to December 2018. This .csv file that we utilized contained 2,556,392 rows and 37 columns of data. There are four specific columns that were quite helpful in our analysis: Category, State, Blurb/Description, and URLs. Our secondary data set came from a GitHub project by user **nalamidi** and included a vital column containing reward level data for our killer plot design.

1.3 Problem

Our research was motivated by a simple question: **What factors contribute to Kickstarter campaign success?** This report will discuss the four segments of analysis that we conducted throughout the semester: Exploratory Data Analysis, Geographic Distribution, Language Processing, and Community Interest.

2 Exploratory Data Analysis

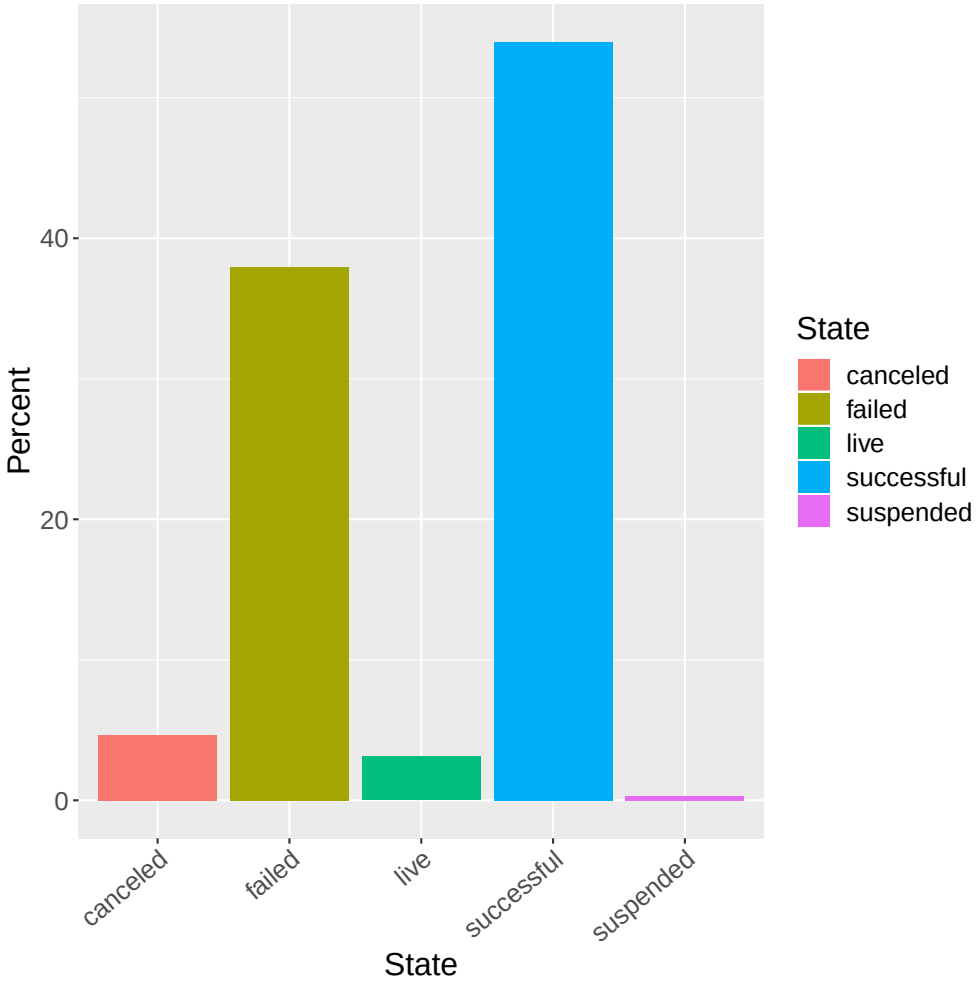
In our fist section we will set out to discover more about the data set. We wanted to understand basic patterns in the data, and determine where it would be best to utilize our resources to analyze the 2.5 million rows of data.

2.1 A Look At Campaign State

Our first graphic displays the five states of campaigns: canceled, failed, live, successful, and suspended.

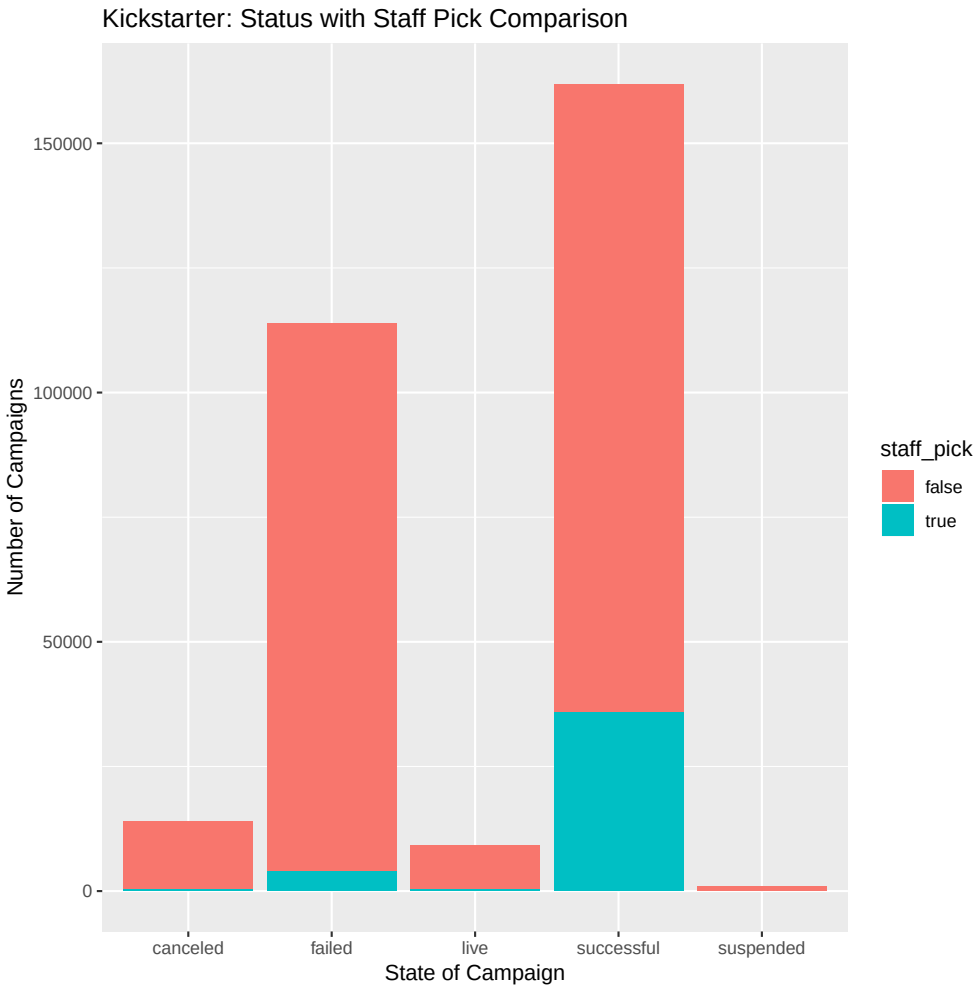
- Canceled campaigns were deleted by their creator and are no longer active.
- Failed campaigns are campaigns that did not meet their funding goal within the allotted time.
- Successful campaigns were able to meet or exceed their funding goals within their defined time frame.
- Live campaigns are still ongoing.
- Suspended campaigns have violated the terms of agreement with Kickstarter, which includes things like misrepresentation of funding or failure to disclose relevant details of the campaign.

From this graphic we determined that roughly 56% of campaigns were successful, while 36% of campaigns were failures, and the other three states constituted less than 8% of the campaigns. This graphic is important because it displays that a majority of campaigns either fail or succeed, making our data set valuable to better answer our main question.



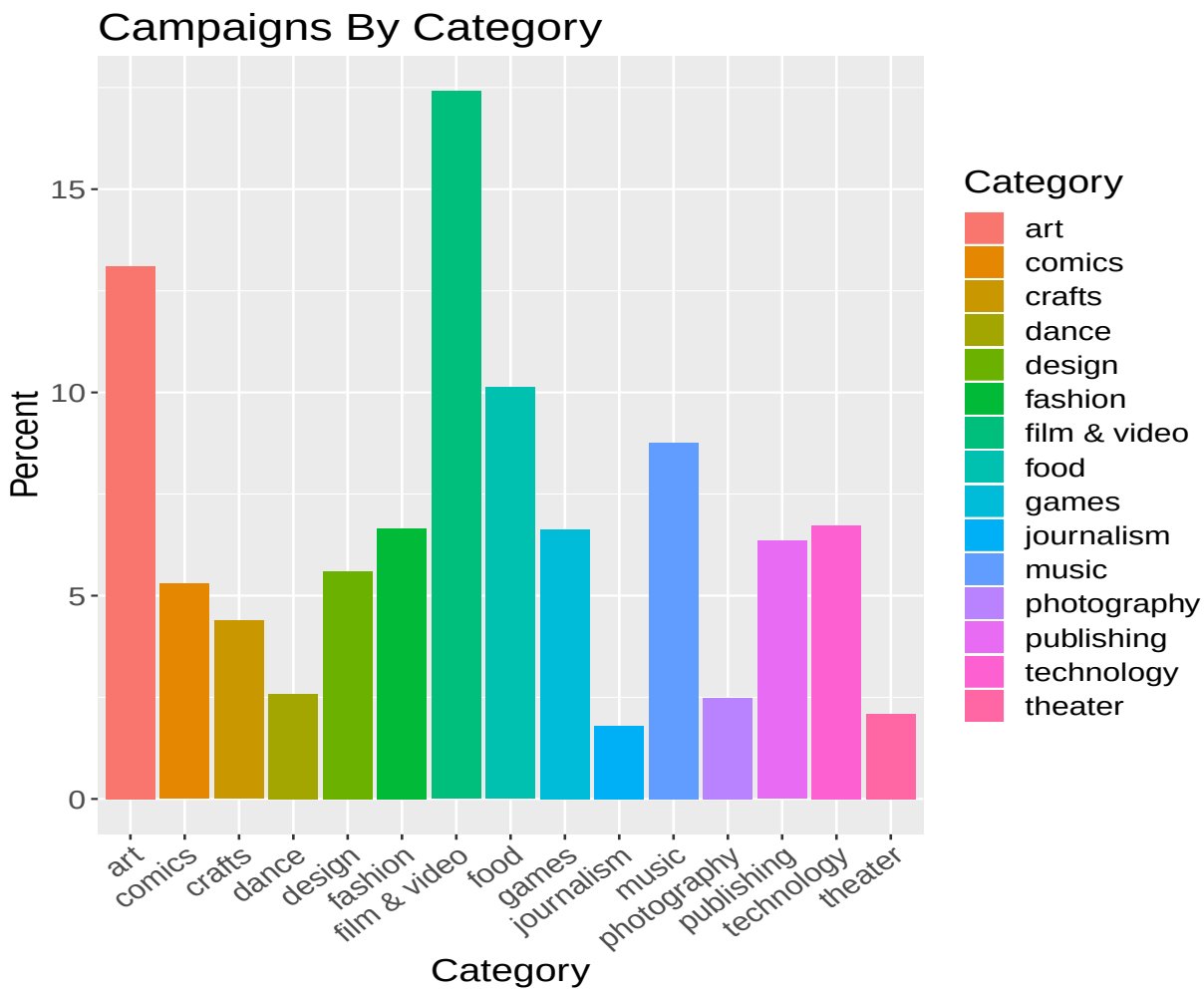
2.2 Staff Picks

To further the analysis of the state of campaigns, we looked into a specialized category that Kickstarter provides on their site: staff picks. According to Kickstarter's CEO, "When something sticks out as particularly compelling, whether it's a really fun video, creative and well-priced rewards, a great story, or an exciting idea (ideally all of the above!), we make the project a Staff Pick". There are a few perks to being selected as a staff pick, most importantly, improving exposure rates for your campaign. Kickstarter utilizes a "magical" algorithm that will display staff picks across the websites pages. Additionally, a staff pick has the opportunity to be published into the Kickstarter newsletter which is sent out weekly via email, furthering the exposure to the public. This bar plot displays the same plot of state of campaigns, but now includes an overlay to define what proportion of each category were staff picks. It is easy to see that only failed and successful campaigns were included as staff picks, but it also displays that the increased exposure may related to a successful campaign. (Source: CrowdCrux.com)



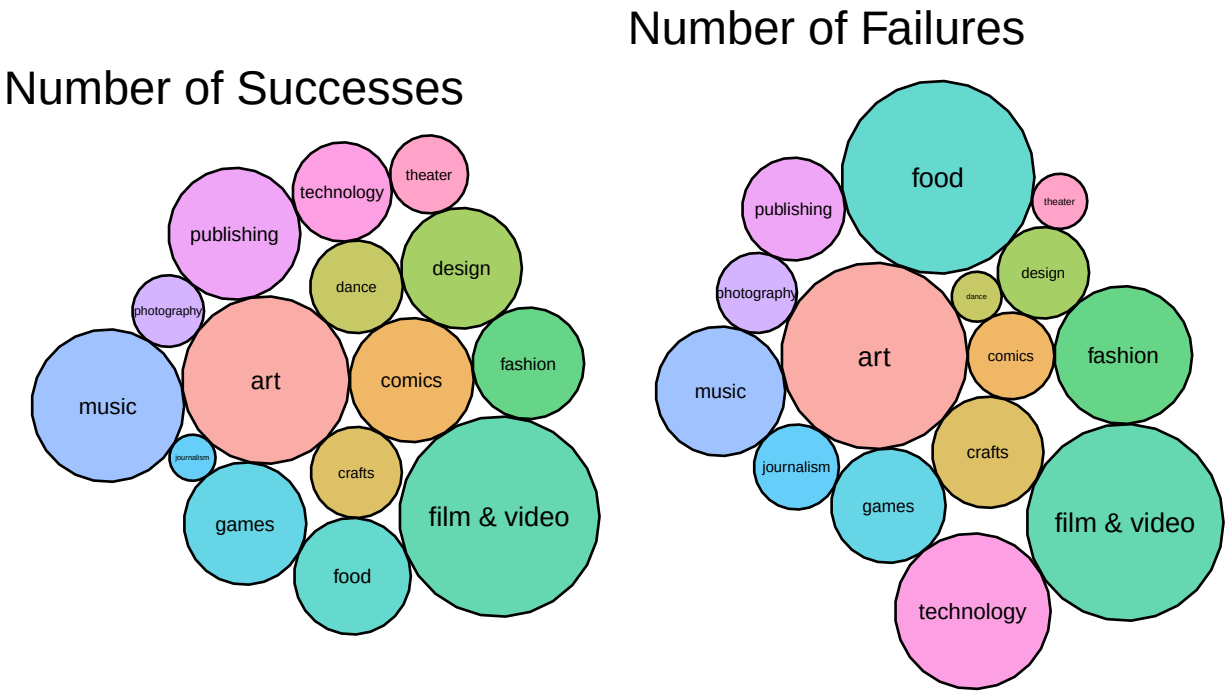
2.3 Campaigns By Category

This bar plot displays the top 15 categories on kickstarter by percentage. Of the top 15 there are 5 categories that make up a majority of the campaigns: film & video (19%), music (16%), technology (15%), art (10%) and publishing (8%).



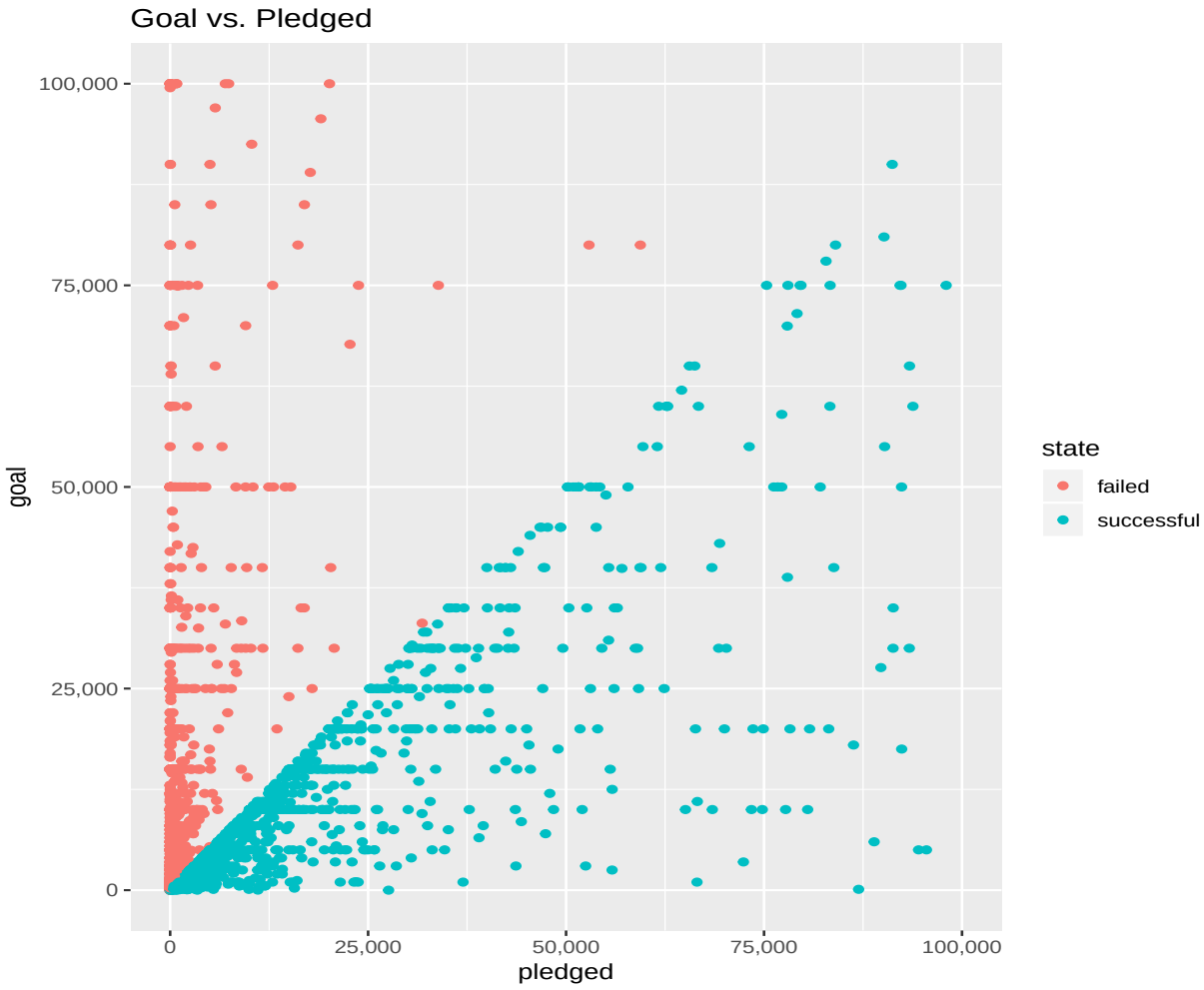
2.4 Comparing Success By Category

In order to visualize the success and failures for each category we developed a bubble plot of the top 15 categories. Each bubble and color corresponds to a category, while the size of each bubble corresponds to the number of successes or failures. Here we are able to visualize the ratio of success to failure for each category. For instance, the size of "film & video" in both successes and failures is roughly the same, displaying that there is close to a 50/50 chance for success or failure for that category. On the other hand there are several categories that have a higher chance for success than failure. If you look at comics, publishing and design, the size of their bubble is much larger for success, than failures, displaying that their success ratio is much greater than failures. Categories like food, technology, and fashion perform particularly poorly: their failure bubbles are even larger than their success bubbles, indicating a success rate below 50%.



2.5 Pledged Amount vs Goal Amount

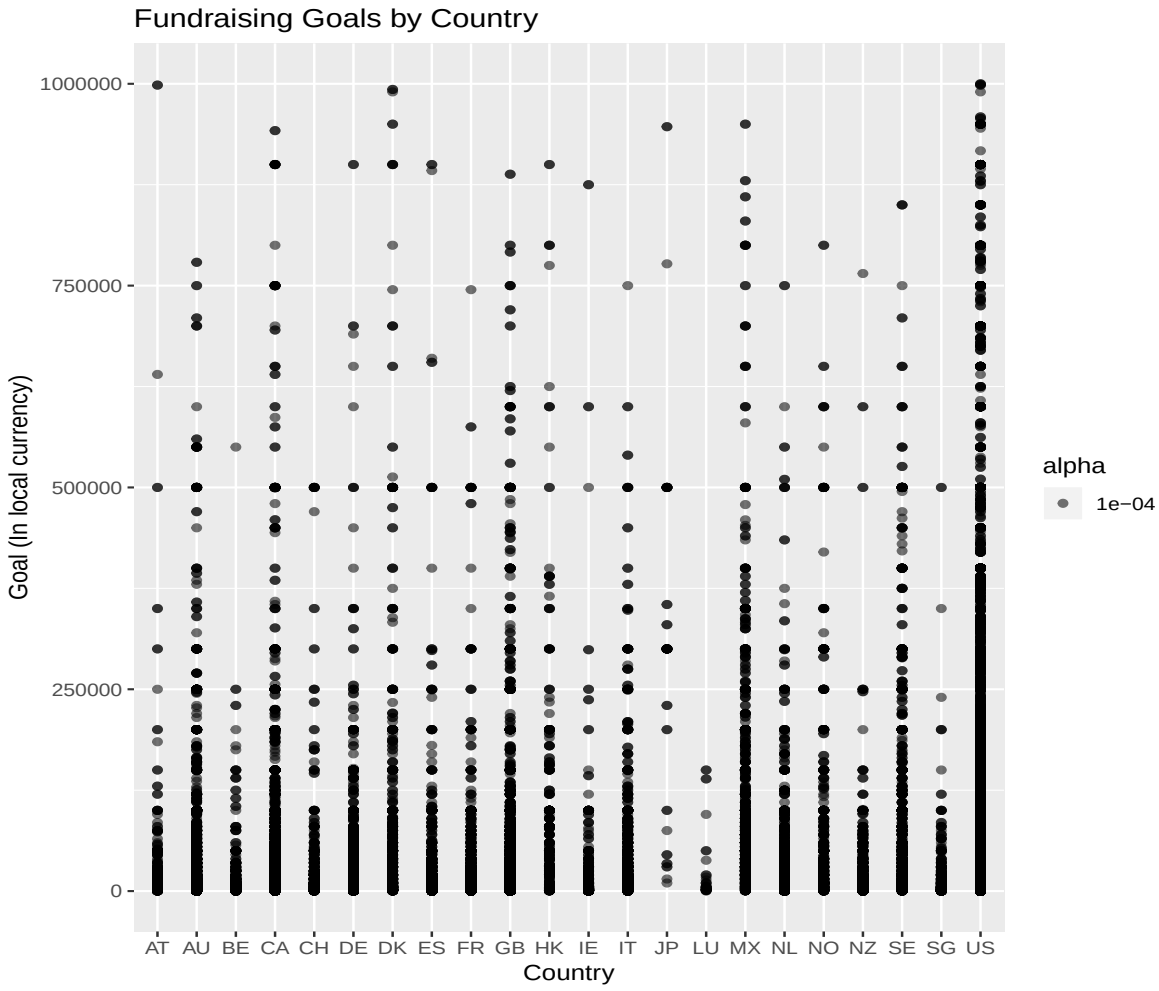
This scatter plot displays the pledged amount on the x-axis compared to the goal amount on the y-axis. The teal colored dots are representative of successful campaigns and the red defines failed campaigns. This plot originally included campaigns that asked for $1,000,000 but were considered outliers, therefore we zoomed in to look at the data below $100,000. There is a clear middle ground for campaigns displayed by this graphic, if a campaign asks for too much then they are at risk for failure, but if they ask for an amount that the public deems reasonable, then they succeed. This scatter plot is not great for predictions, because obviously those who succeed have met their fundraising goals (and colored teal) and vice versa, those who failed did not meet their fundraising goals.



3 Geographic Distribution

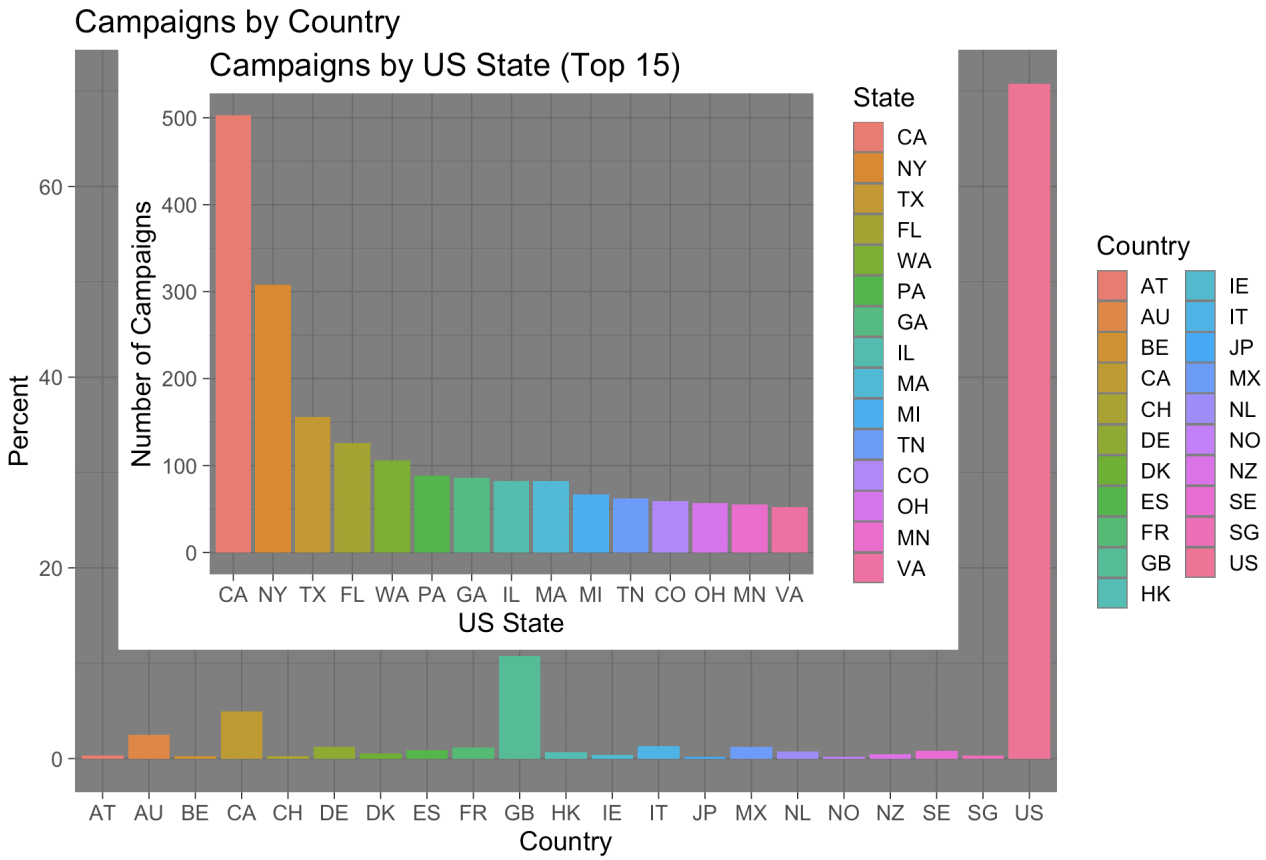
3.1 Fundraising Goals By Country

First we compare the fundraising goals by country. The x-axis represents the country name and the y-axis represents the fundraising goals in US dollars. As shown on the graph, the United States has many more ambitious fundraising goals, but that is likely a result of an overall higher number of campaigns. Still, we do notice that smaller countries like Japan, Italy, and Luxemborg do not have many campaigns with very high fundraising goals.



3.2 Geographic Distribution of Campaigns

The graphic below combines two bar plots of "Campaigns by Countries" and "Campaigns by U.S. State". We utilized grid to superimpose the second plot into the blank space of the first plot. The background plot displays that roughly 68% of campaigns originate from the United States, while Great Britain is far behind in second place with less than 20% of the global campaigns. We decided to investigate the United States' distribution of campaigns since they accounted for the majority of campaigns. We found that California, New York, and Texas accounted for the most campaigns by state, respectfully. This is interesting because Kickstarter was founded in New York, but California out produces New York by nearly 60% We believe this is due to influx of innovation that Silicon Valley brings to California but it could also be a factor of population size. California has 40 million residents, the state of New York has 18 million residents.



3.3 What is a Reward Level?

Another important Kickstarter topic that we discovered to have a fairly strong connection with success is the concept of reward levels. When setting up a Kickstarter campaign you can develop different reward levels, or backing levels that provide the public with options to support your product. For instance, a campaign for a clothing brand may offer a base reward level that costs $60 and rewards the backer with a hoodie from the company, the second tier of rewards may cost $80 and reward the backer with a hoodie and three pairs of socks from the company, and the third and final level may cost $150 but provide the backer with two hoodies, three socks and a signed letter from the CEO, thanking the consumer for their support. Every category has different methods for developing reward levels. A book campaign may offer a signed copy of the book as their highest reward or a cellphone case campaign may offer a lifetime warranty for their highest level reward. There is no limit on the number of reward levels that a campaign may have, but successful campaigns seemed to include more levels than failed campaigns. An example of one such reward level from the Kickstarter website can be seen below:

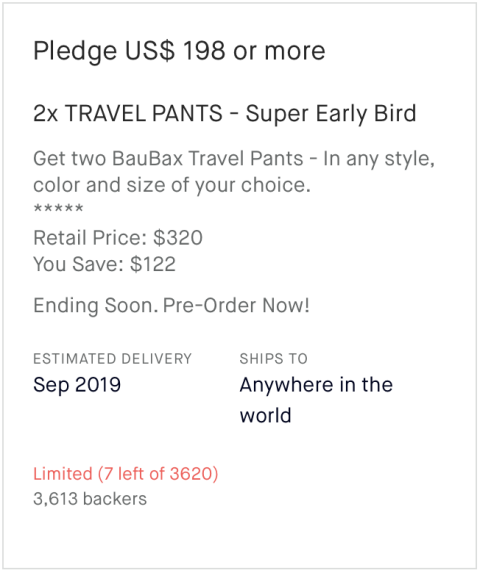


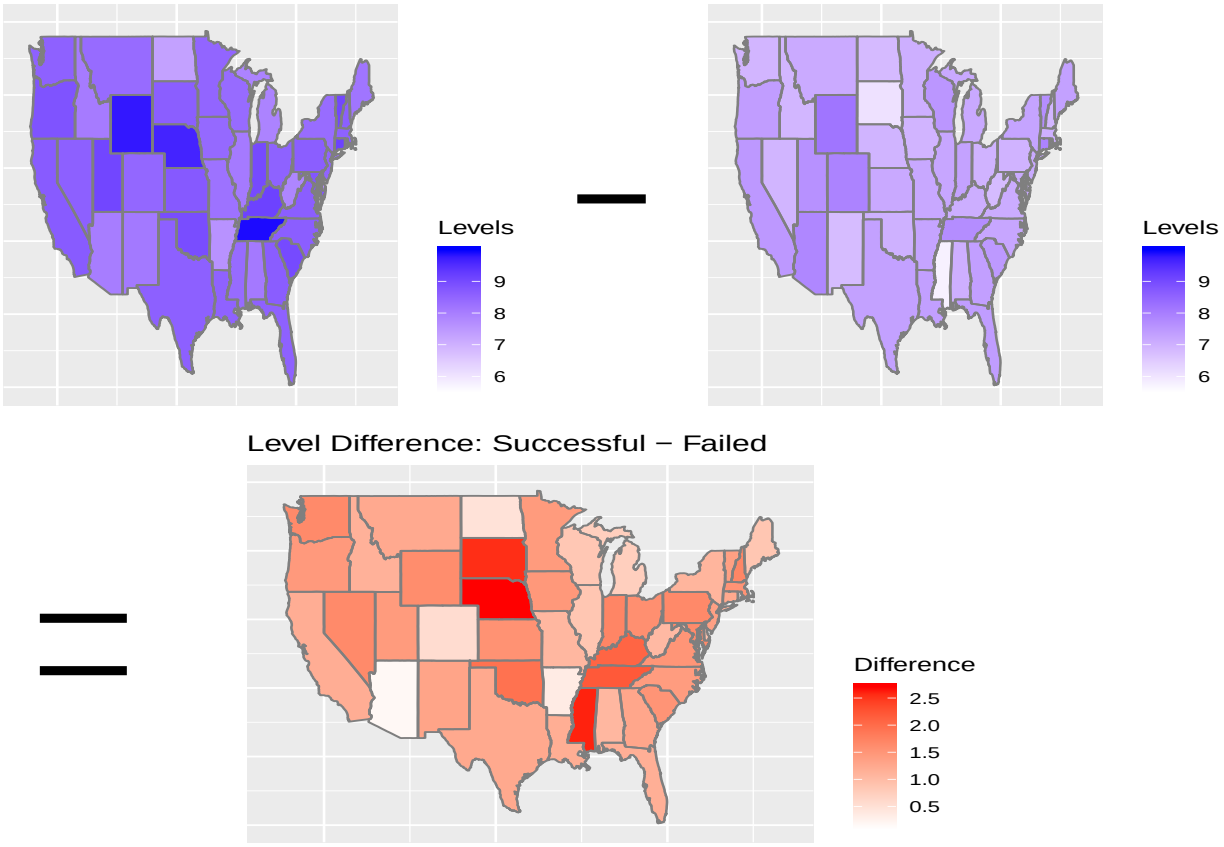
Figure 1: Example Reward Level

3.4 How Reward Levels Impact Success

We built the following plot by loading latitudes and longitudes and tracing them with grid to develop the map of the United States. We believe this is unique because we subtract one map from another to create a "difference map", seen at the bottom in red. The top left map of the United States is average number of reward levels in successful campaigns, while the map on the upper right is the average number of reward levels in failed campaigns. By subtracting the failed campaigns from the successful campaigns, we arrive at the bottom plot: The average difference in reward levels between successful campaigns and failed campaigns. Overall, there are two key takeaways from this difference map:

1. The difference in levels between successful and failed campaigns does not follow any notable geographic pattern.

2. More importantly, we see that the entire map shows positive values. This means that, *across every state*, the successful campaigns tend to offer more levels than the failed campaigns. So overall, we can say that offering more reward levels tends to be associated with campaign success,



4 Language Processing

In this section of the report we explored the relationship between keywords in the blurb and the success of a campaign. The blurb is a sentence or two that states something general about the kickstarter campaign to convince people that their campaign is worth their funding. We looked for specific words or symbols in the blurb and checked if the corresponding campaigns were more successful on average. The list of words and symbols includes "first", "!", and "best".
We analyzed if a keyword indicated success by comparing the percentage of successful campaigns the word appeared in to percentage of unsuccessful campaigns that the word appeared in.

- "first" appeared in:
  - 5.6% of successful campaigns
  - 3.4% of failed campaigns
- Exclamation marks appeared in:
  - 26.7% of successful campaigns
  - 21.8% of failed campaigns
- "best" appeared in:
  - 1.8% of successful campaigns
  - 1.45% of failed campaigns

While many of these percentages are very small, some of them do indicate notable changes in success. Successful campaign descriptions are 65% more likely to contain the word "first" and 22% more likely to contain an exclamation mark. When considering all possible words that can be used in a description, numbers like 5.6% for a single word are not as small as they sound. However, these small numbers should be kept in mind before drawing conclusions about the impact of a single keyword on campaign success.

4.1 Keyword Analysis With Shiny Application

Additionally, we created a shiny application which allows one to type in any word or combination of words and get these results in the format of a bar graph. The bar plot always has two bars. The first bar represents the percentage of failed campaigns that include one of queries, and second bar represents the percentage of successful campaigns. On the shiny application, there are five textboxes to enter queries into. Entering "first" into one of the textboxes while leaving the remaining textboxes empty will create a bar plot where the first bar is 3.4% and the second bar is 5.6%. The y-axis scale is always in percent but is rescaled to have the higher percentage fill up the bar plot. This allows for an easy comparison between successful and failed campaigns. Adding words or symbols to other queries allows the shiny application to check for multiple words at once. More specifically, the image below shows the failed and successful campaigns that included "first", "!", OR "new".

Figure 2: Shiny Application Used For Keyword Analysis

5 Community Interest: User Comments

Our final section of the report, we decided to evaluate community interest in the campaigns, and hypothesized that more successful campaigns usually have more comments than failed campaigns. We randomly selected 1000 campaigns from our existing data source to test this hypothesis. Next we extracted the URL for the

each campaign and used those URLs to extract the number of comments on each campaign's webpage. Using this data, we constructed a new table comment_counts that includes the URL, number of comments, and state (success vs failure) for each of the 1000 campaigns. This scraped web data table became the basis for exploration of our final section.

5.1 Comments and Campaigns

By comparing the means, quantiles, and percent of zeroes across successful and failed campaigns, we were able to see a strong correlation between the number of comments and campaign success.

Mean comments in successful campaigns: 26.94
Mean comments in failed campaigns: 0.94

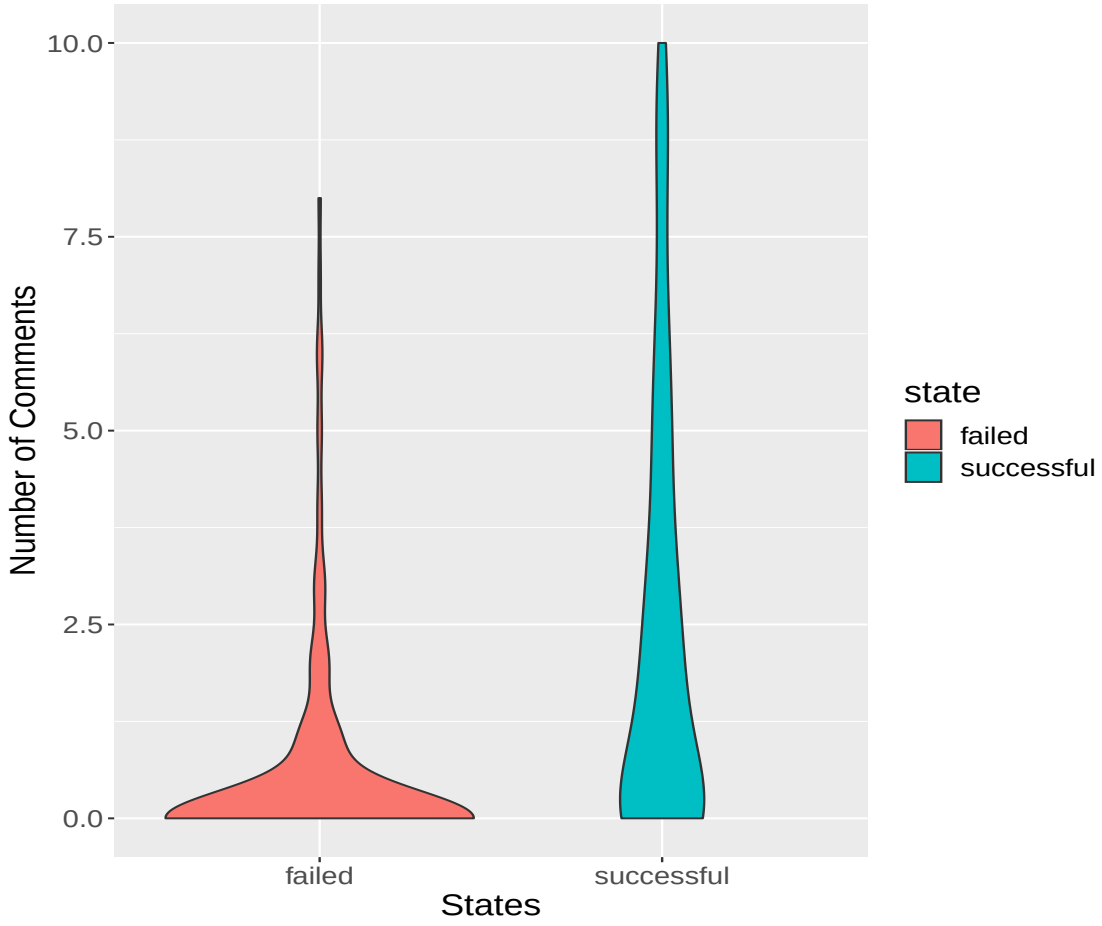
We calculated the mean number of comments for successful campaigns and the mean number of comments for failed campaigns. There is a clear difference between the two. For a successful campaign to occur it seems like it needs to have an average of at least 26 comments, while on the other hand unsuccessful campaigns will acquire only 1 comment. This suggests that a good campaign should induce the kickstarter community to ask questions and foster discussion about the product.

Percent of successful campaigns with 0 comments: 27.3%
We calculated the percent of campaigns that were successful and only had one comment to be 27%. On the other hand 73% of campaigns that we successful had more than one comment. Furthering the idea that inducing comments for campaigns can related to a successful campaign.

Percent of failed campaigns with 0 comments: 76.2%
Here we calculated the percent of campaigns that were successful and only had one comment to be 76%. If you do not get any comments, then your chance of success drops. Again, this furthers the idea that comments are an important factor to determine the success of a campaign.

5.2 Visualizing Comment Counts

We utilized the violin plot function of R to visualize the campaigns that failed and succeeded based off of the number of comments. The y-axis ranges from 0 comments to 10 comments, but many successful campaigns had well over 10 comments. We narrowed the y-axis here to ten to better visualize the data since most failed campaigns had less than one comment on average. The elongated base of the red, failed campaigns proves our hypothesis that community engagement in the for of comments related to success. The long spire like structure of the successful campaigns displays that they are more likely to obtain comments from the public.



6 Conclusions

Our project revealed valuable information about the nature of Kickstarter campaigns and what makes them successful. We found that categories like comics perform better than average while categories like food and fashion perform particularly poorly. We also found that, while geographic location does not impact success too much, including more reward levels improves campaign success rates across the country. When writing campaign descriptions there are some keywords that can help entice backers, but it should be noted that the numbers backing that hypothesis are fairly small. We believe that future research performing sentiment analysis on campaign descriptions may yield more robust results. And finally, we see a strong relationship between user comments and campaign success. This indicates that Kickstarter campaigns should work to engage the community and foster discussion about their project in order to improve their chances of success.

7 References

1. Primary dataset: *https://webrobots.io/kickstarter-datasets/*
2. Secondary dataset: *https://github.com/nalamidi/Kickstarter-Predictive-Analysis/tree/master/Kickstarter*
3. Kickstarter website: *https://kickstarter.com*
4. Kickstarter staff pick background information: *CrowdCrux.com*