Kickstarter Project - Milestone Report

Problem Statement

When Creators are seeking for ways to bring their project to life, they have several means of achieving this. Some projects are easy to bring to life based on stuff that creators already have, while others require a bit of financial assistance. For ideas that need help, they can go through a financer such as a publisher, for writers, or a studio, for game developers. However, not all creators want to go that route. For those, there is the crowdfunding path where their project can come to life with the help of other people who want to give money to see their creations.

Kickstarter's platform allows crowdfunding to easily happen. A Creator can post a project proposal, describing the idea, the path to the idea, and how the community can help. The community of crowdfunders can give money to the Creator to see their project come to life.

However, the Creator will only receive funding from the community if the project is 'fully backed', i.e. the monetary goal amount set by the Creator was reached by the community. Kickstarter states that the success of projects is only 37.06%. While not all ideas will, or should, be successfully funded, there are ways for Creators to improve their proposal, or "campaign", to reach their funding goal.

A prediction model can help both Kickstarter and the Creator by providing a way to increase the success rate and help Creators optimize their campaign. Campaigns can be edited to follow patterns of success of previous successful campaigns. They can also be optimized to release on certain days to reach a wide amount of audience members.

Kickstarter Data

Kickstarter, unfortunately, does not provide a public API and the community has limited access to their data. There are organizations, such as Web Robots, who, as a side project, scrape sites and provide the data they gathered to the public. I will be using the Kickstarter dataset provided by Web Robots to analyze successful Kickstarter campaigns. Web Robots provides their files in JSON or CSV with updates every month.

The initial kickstarter dataset is provided by webrobots.io from their free web scraping data projects. The dataset is pulled from the 2019-05-16 set and contains 210k rows with 37 columns. The data is split into 56 csv files.

Initial Cleaning

For the most part, the data set is generally clean. However there still needs to be some cleaning done. We'll start with combining the csv files into one single file. I'll be using the glob

module to find all the file names for reading into a dataframe using pandas. From there, I'll be using pd.concat to attach all the DataFrames together into one large data frame.

The first data point to clean would be the timestamps and the format that they're stored in. According to the webrobots.io website, the time format is in unix, or epoch time. Using the time module, we're able to easily convert the epoch timestamps stored on the 'created_at', 'deadline', 'launched_at', and 'state_changed_at' columns. We store the value into two separate columns, "x_date" and "x_time" where x is one of the original columns. By storing time and date into separate columns, we'll be able to do easier analysis later on when we're exploring the data.

Our next points to clean is the columns stored in JSON formats since the web scrape project pulled the information in JSON. To do this, we simply use the json module to read the string in as a dict and pull information that way. Since there are several columns that store format in this manner, we'll be using a dict to store column:category information. There are also some columns that contain multiple categories that we want to pull so this way will also address that. We'll then use a for loop over key-value pairs to pull information in an easy manner.

Doing a bit of simple analysis on the data set, there are several columns that contain a large number of nulls compared to our 210k rows. From our exploration, the columns 'friends', 'is_backing', 'is_starred', and 'permissions' look to be related to a certain account. It provides information that is not relevant to the overall analysis.

Additionally, there are other columns that, while informational, provide little value to an analysis. For example, there is the currency_symbol which shows which symbol the currency uses. This would be helpful but we already have a currency column which provides more detail than the symbol. For the "$", this would relate to USD, CAD, and AUS. The symbol is not able to tell us this information. Other columns such as "photo" also fall under the unusable list and are therefore dropped.

Finally, there is an opportunity to pull more information from Kickstarter using the URLs provided by the dataset. To prepare for this, we'll be cleaning up the URLs once more by reading the JSON into a dict and parsing through the data that way.

Web Scraping

This is easily the most frustrating portion of the wrangling process. The script to scrape information itself was simple. Using the requests library and the BeautifulSoup library, I was able to pull the description from the HTML of the Kickstarter. However, when running this script on 210,088 rows, each iteration took about 2 seconds. This totalled to an expected time of 116 hours or almost 5 days. This quickly became unreasonable.

I knew right away that there could be optimizations made to speed this process up. I broke down my script into three parts: requesting the HTML, parsing the HTML, and creating the text from the HTML.

I ended up comparing between two libraries: urllib and request. The urllib library ended up being way faster on several tests when pulling the HTML of the page. It also ended up being more consistent with what it would pull. The requests library would sometimes be unable to pull information from a page. As for the cause, I'm still unsure but I do know that it was worth the switch. This ended up making the script run way faster and averaged less than 1 second per iteration.

When parsing the html, I compared BeautifulSoup and the selectolax libraries. From my testing, the selectolax library ended up being faster by a few ms but with the benefit of BeautifulSoup being able to parse multiple tags, the minimal speed increase of selectolax was not enough to overcome BeautifulSoup. Additionally, I also ended up parsing the image and video tags to count how many images and videos are in each description. Time was not saved in this section.

For creating the description text, there was an optimization issue since each p tag was separated in our HTML parsing. To overcome this, I used list comprehension when joining text.

Overall, the time taken for the script was reduced in half.

To run this script on a personal computer is taxing. There runs the risk of the computer turning off, the laptop going into sleep mode, and a multitude of other issues. To overcome this, I searched for ways to run this on a cloud server. Google Cloud Platform became the answer for this. This will allow the script to run without worry of any computer outages or errors.

Using GCP's free trial, I spun up a Linux based compute engine to simply run my script. The overall time ran for 2.5 days, with error checking for each row to ensure that the script does not crash.

Further Cleaning

Once the script finished, there were additional steps to clean since the data was not always pulled in the nicest way. To parse through html, the parser script stored information into a comma delimited list resulting in one additional column. The cleaning process addressed this issue by moving to individual columns. Further cleaning also included removing of rows that resulted in errors during the parsing process. Since there were only 184 rows, and those rows were consistently returning errors, it made sense to remove them. Additional processes included removing NaNs and creating new columns such as 'percent goal' or 'video usage'.

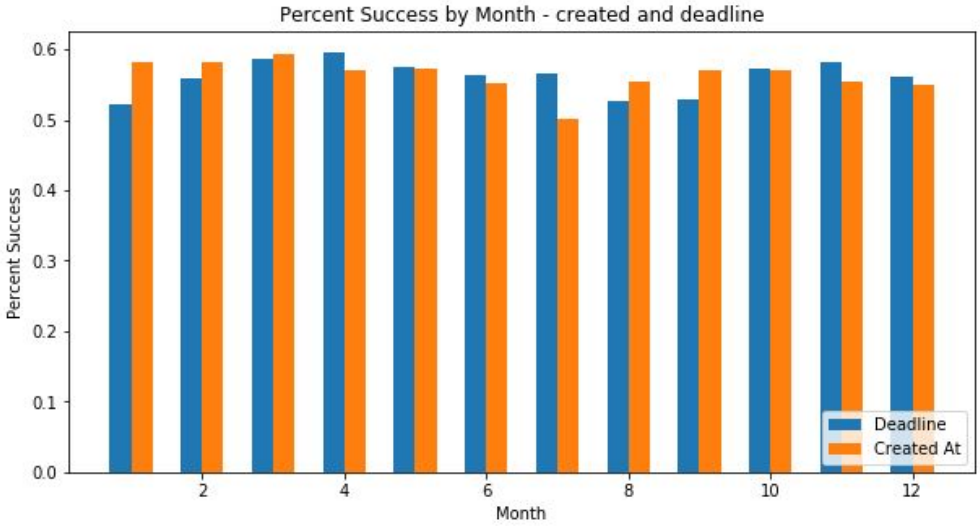
Past the cleaning, I realized that after certain procedures such as crafting the 'percent goal' or moving all monetary values to USD that other columns were not necessary for the analytics portion. This resulted in a new cleaning process where I dropped several of the tables that were not used for analytics such as the web urls or the original goal in JPY or GBP. I also made a final column to calculate the median of the rewards since that would be useful for analytics.

Exploratory Data Analysis

Diving deep into the data set, we can find several points that may lead to a solution for our problem.

Projects by Month

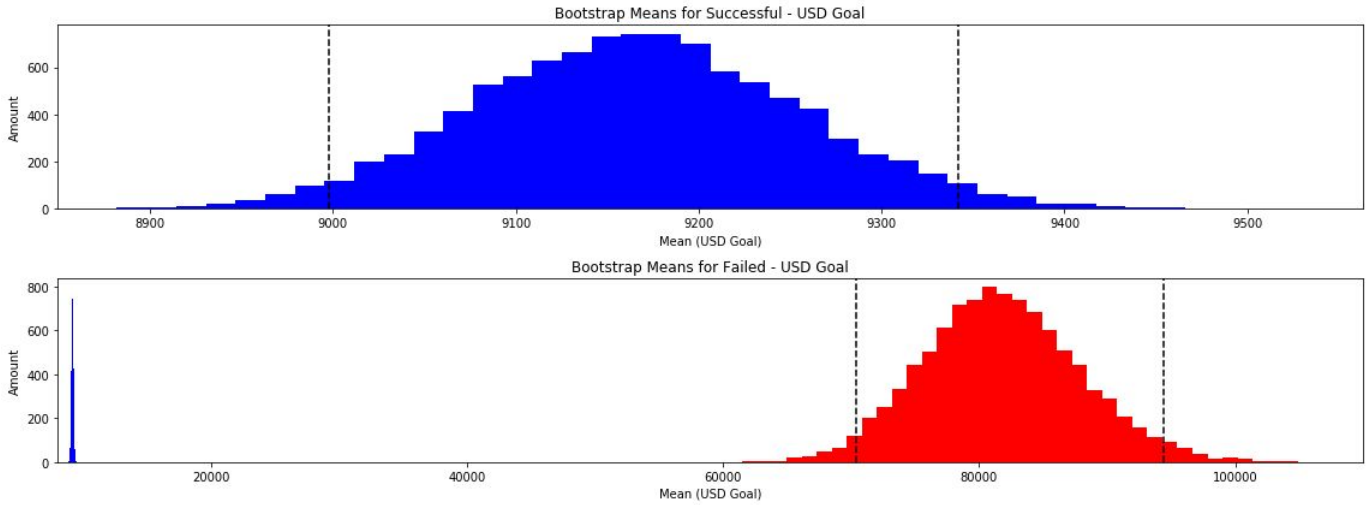
Kickstarter projects are being started no matter the time. Some months, namely January and December, have more projects stated during that time. Others have less projects starting at that time. However, frequency does not mean success.



Taking a look at the success rate by month shows that projects started during March are the most successful while projects starter in July are the least. Deadline date also plays a large part in success rate where projects ending on April are the most successful.

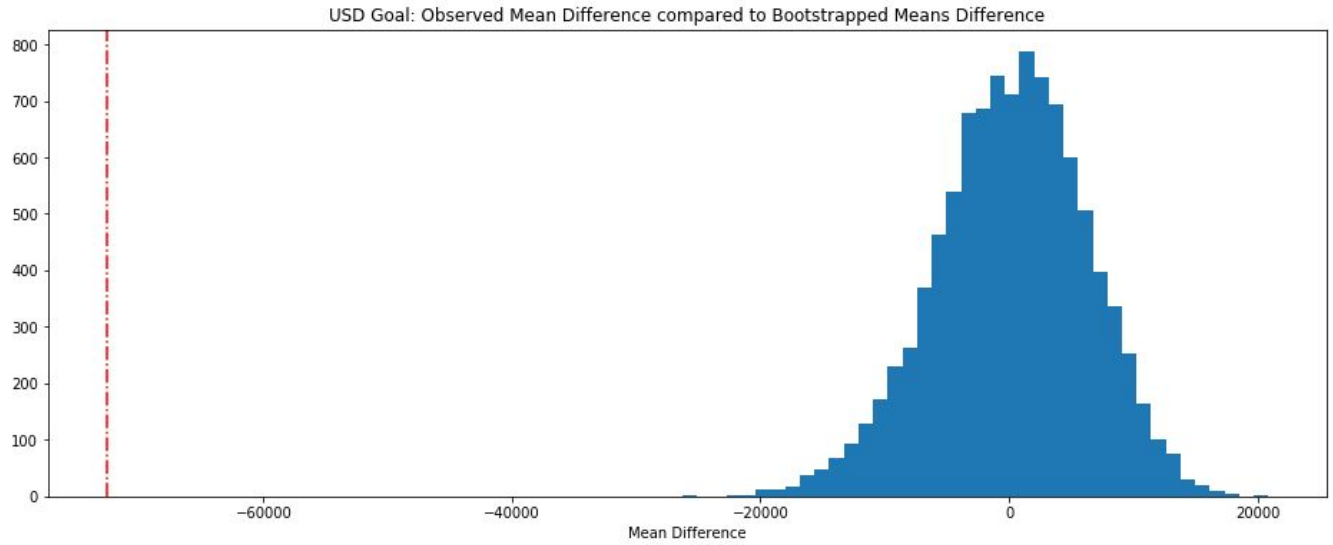
Projects by Goal (USD)

Each Kickstarter project has a goal that the Creators want to reach and is established clearly for the crowd. By the nature of it, projects with tiny goals such as $100 will usually be funded since it does not take many people to fund whereas projects with more lofty goals will take much longer and will need more work to be funded. Breaking down the data set into successes and failures, is there a difference in average goals and would that have an affect on projects?



The first graph shows the set of bootstrap replicates for the mean of goals (usd) for successful projects. The second one shows the failed projects compared to the successful ones. We can see a large difference in their means. If we perform a hypothesis test and assume that their means are equal and no different, how likely are we to get the observed difference of means between successful and failed projects.

While the above graphs were created using the complete data set, after limiting our bootstrapped replicates on data within 1000 (USD) and 1 million (USD), the difference between the groups was stark.

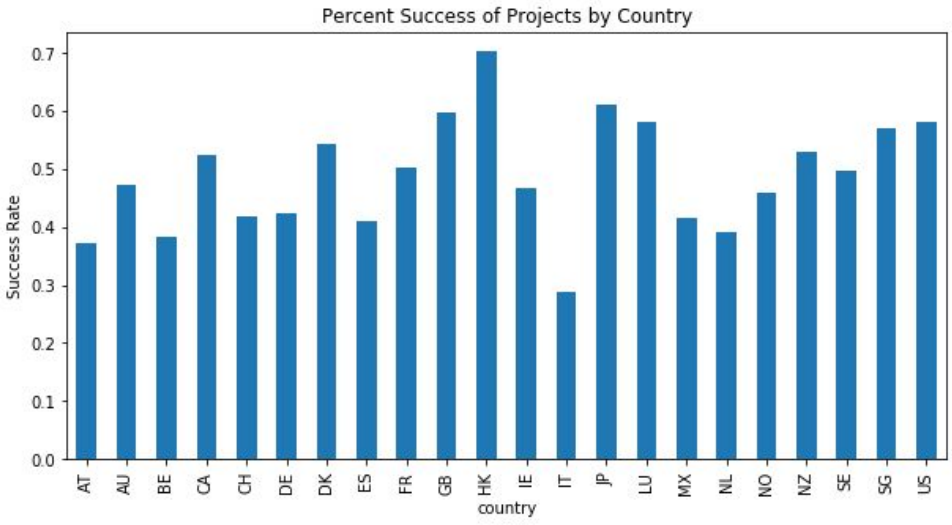


The blue histogram is the bootstrap replicates of the successes and failures if their means were equal. We can see that it still has a wide mean different spread but it centers around 0. However, the main thing to notice is that the observed mean difference, the vertical red line at -72,000. This shows that given the assumption that the mean between successful and failed projects are equal, the probability that their different reaches what we observed is very very slim. The calculated p-value reached 0.00 but it is essentially small or as close to 0 as

possible. This shows that we can reject the hypothesis that the means of successful projects and failed projects are equal.

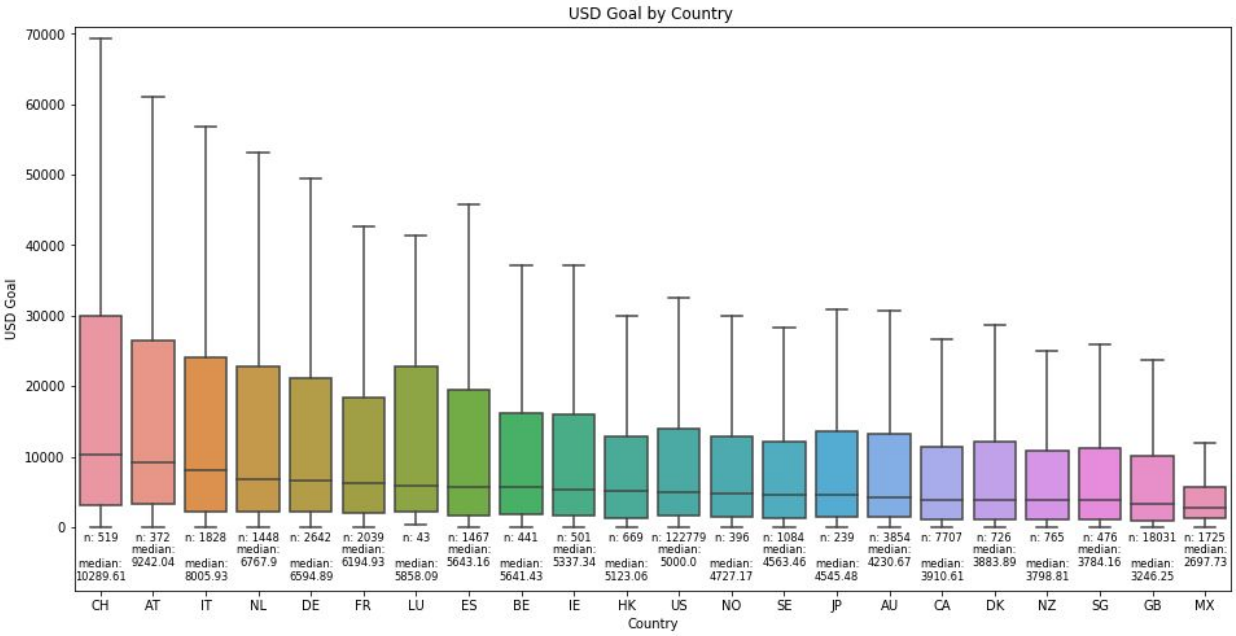
Projects by Country

Kickstarter is a global idea and so has projects being started all over the world. The main country where projects start out of is the United States where 72% of Kickstarter projects originated. However, does this necessarily mean that leads to a successful project?



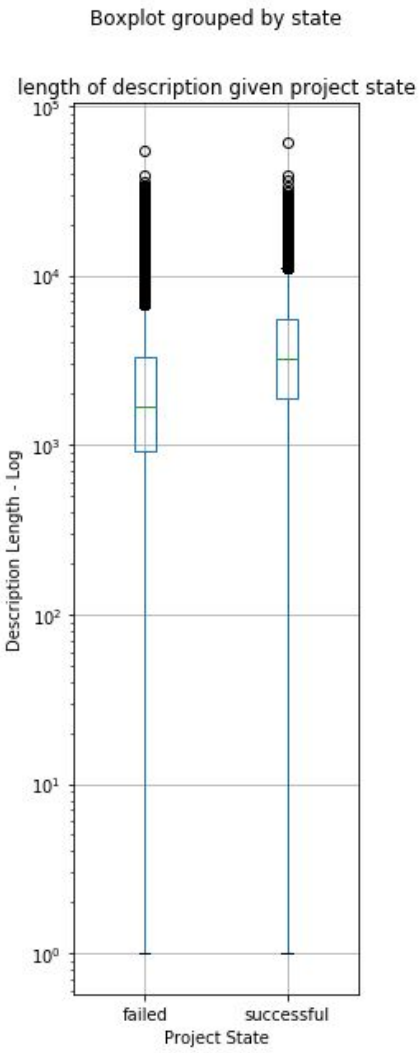
By taking a quick look at the success rate per Country, we can see that the US is actually at around a 58% success rate. The highest one is Hong Kong with a success rate of 74.7%. The country with the lowest success rate is Italy with a success rate of 32.8%.

We can also explore countries in relation to their USD Goal.



Here, we can see that

Projects Descriptions - Length



Descriptions of projects are the best way to entice potential backers to support a Kickstarter. We can take a look at the lengths of descriptions and see if they determine success.

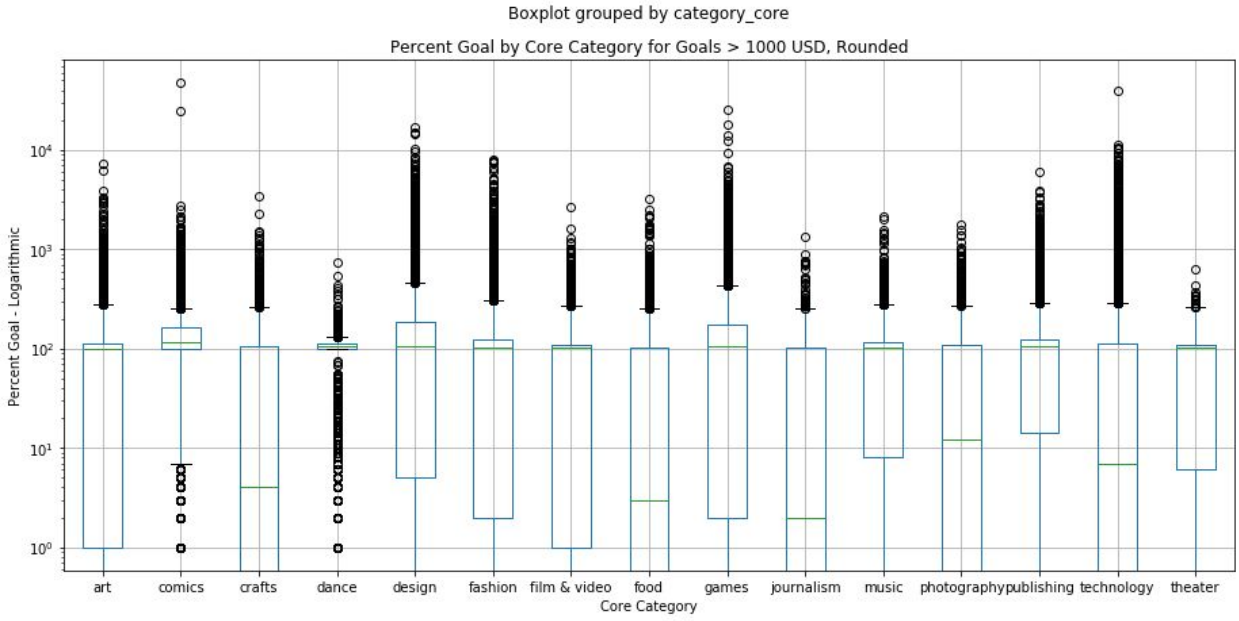
From a boxplot of description length broken by the success or failure of the project, we can see that successful projects, on average, use more words in their description than projects that fail. However, it's worth noting that both project states have descriptions with lengths of 0 (i.e. they could have zero text or use images as descriptions). There are also points outside of the boxplot range that we may want to pay attention to. These may occur due the description having a pareto distribution.

Projects by Category

Category - Percent Goal

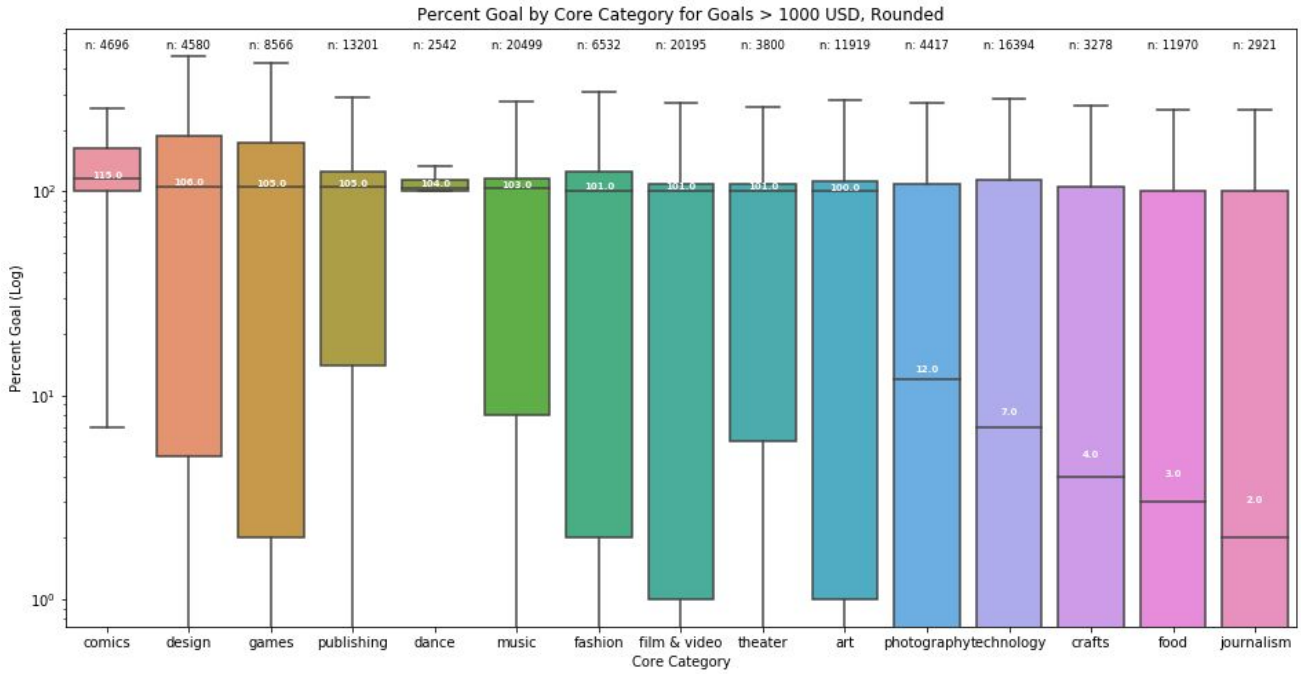
Each project on Kickstarter is broken down into one of 15 categories and further broken down into subcategories. These categories are selected by the creator and generally encompass what the project entails. These range from "art" to "food" to "technology". Some may overlap such as "film & video" "dance" project. It's up to the creators to distinguish where their project belongs.

Are certain categories more successful? Does the final product drive backers to the project? To start, we can look at the percent success of projects based on their category.



For most categories, it looks like their average percent to goal is 100, meaning they've hit their goal. For some, namely the crafts, food, journalism, photography, and technology categories, it looks like they're more affected by other factors and bring their success rate down. Another interesting point is the dance category looks to center around 100% success rate. The highest 'max' between the categories look to be between design and games.

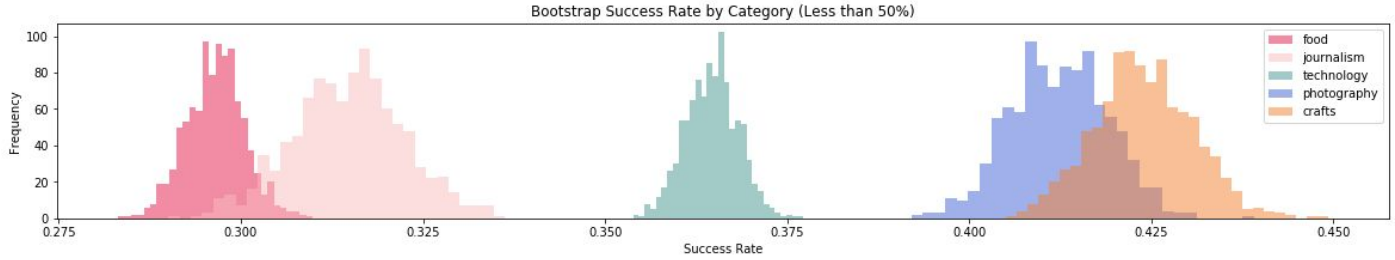
Let's dive deeper into the categories.

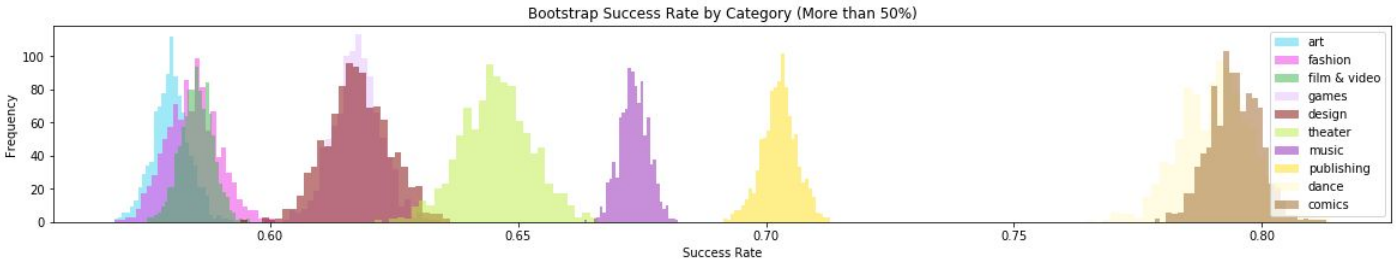


From this we can see that comics has the highest median success rate, followed by design and games. It's also interesting that these are some of the categories with the lower counts having smaller ranges compared to categories with higher counts with very large ranges.

Category - Success Rate

We can use the bootstrap method on the categories and their success rates.

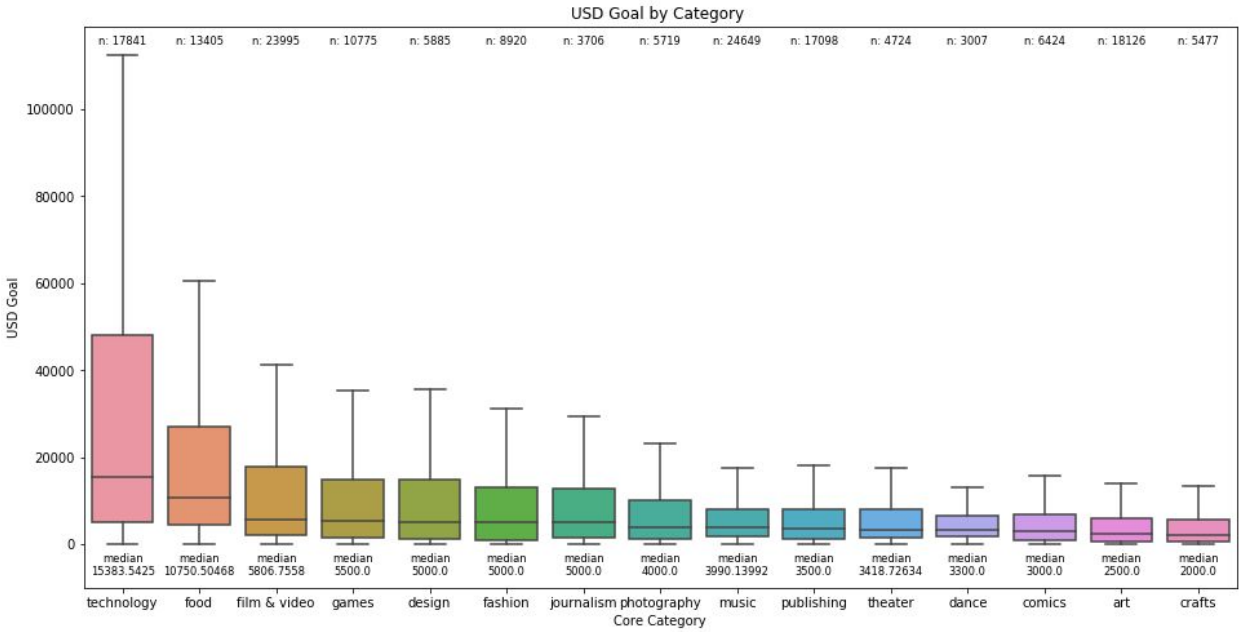




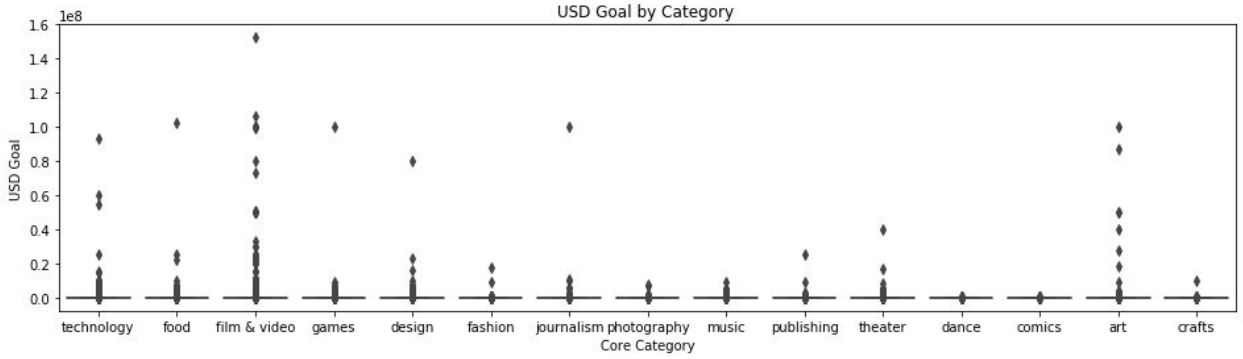
We can see that success rates range from ~28% to ~90%. The lowest category success rates belongs to food and the highest belong to dance and comics. With this, we can predict based on categories on whether a project will be successful or not.

Category - USD Goal

We'd also like to see how the categories reflect their USD Goals relative to other categories. We can do that with a boxplot.

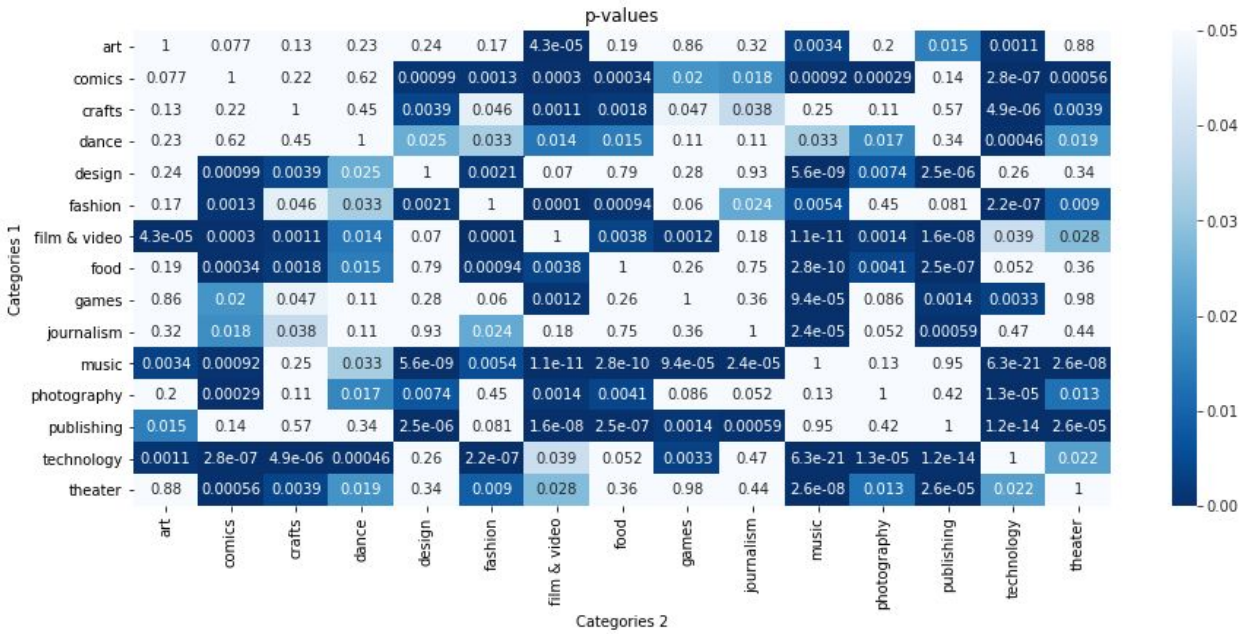


We can see that technology has the widest range and also the highest median goal. Other categories tend to be in the same range for their goals. We can also see that the size of each category fluctuates widely ranging from 24,000 to 3,000. A deep look into the categories would be to also look at their outliers and see how they affect each category.



It looks like Film & Video have the widest range of outliers. Technology also has a spread of outliers in the higher ranges. For a few categories such as food, games, and design, there look to be a cluster of data points with lower goals but have a few high reaching goals that may be affecting the mean. Art, similar to Technology, looks to also have a wide spread of data points in the higher goal range.

We'd also want to compare categories between each other to see whether or not the difference between them is significant in predicting success.



Using a 5% threshold, we can reject the hypothesis that the categories have similar average usd goals. This shows that simply comparing successful and failed projects as a whole is not enough to tell a complete story and that we would need to pay attention to the category of the project. If we performed a one way ANOVA test on our USD Goals with a null hypothesis that mean usd goals between categories are the same, we'd 2.185e-26, well below an alpha of 0.05. Since we're rejecting the null hypothesis, we can further explore to determine which categories differ the most by performing a Tukey test.

group1	group2	meandiff	lower	upper	reject
art	film & video	75098.7355	38120.565	112076.9061	True
comics	film & video	101276.2274	48490.4682	154061.9866	True
comics	technology	64089.8212	9415.2691	118764.3734	True
crafts	film & video	99138.275	42867.8098	155408.7403	True
crafts	technology	61951.8688	3905.8911	119997.8466	True
dance	film & video	101035.8861	28345.2743	173726.498	True
fashion	film & video	92658.3509	46060.9628	139255.7391	True
fashion	technology	55471.9448	6745.2902	104198.5993	True
film & video	food	-58887.1959	-99405.3324	-18369.0593	True
film & video	games	-72766.196	-116341.5353	-29190.8566	True
film & video	music	-97359.664	-131436.7193	-63282.6087	True
film & video	photography	-95079.9702	-150372.6429	-39787.2974	True
film & video	publishing	-97447.5457	-135053.6495	-59841.442	True
film & video	technology	-37186.4062	-74332.4488	-40.3636	True
film & video	theater	-72442.0367	-132252.4474	-12631.626	True
music	technology	60173.2578	23237.9634	97108.5523	True
photography	technology	57893.564	794.9656	114992.1623	True
publishing	technology	60261.1396	20046.8471	100475.432	True

Taking the results of the Tukey test and looking at those that it rejects based on an alpha of 0.5, we can see that film & video looks to be the most different to other categories in terms of mean usd goals. This is followed by technology which make up the other comparisons in this list (group1 and group2 include either tech or film/video). Several factors could be playing into this but it's worth noting that technology and film & video are one of the highest size categories that also have one of the higher ranges of usd goals.

Projects on Word Choice

We can also take a look at the categories and their most used description terms.

From this we can see some key terms that each category uses and some terms that are shared across categories. Immediately, we can see the term “one” being shared across multiple categories, regardless of relation. Taking a look, it seems as if the descriptions use “one” to describe their product in a way it stands above others (i.e. “one of a kind”, or “one perfect brew”). There are also other instances where one is just a passing term not meant to have any meaning related to the product (i.e. “one day”, or “one last thing”).

Some categories that require creativity use the word “design” seen in the design, fashion, and art categories. Some words like “app” or “device” live solely in the Technology category and words like “recipe” and “farm” belong to the food category.

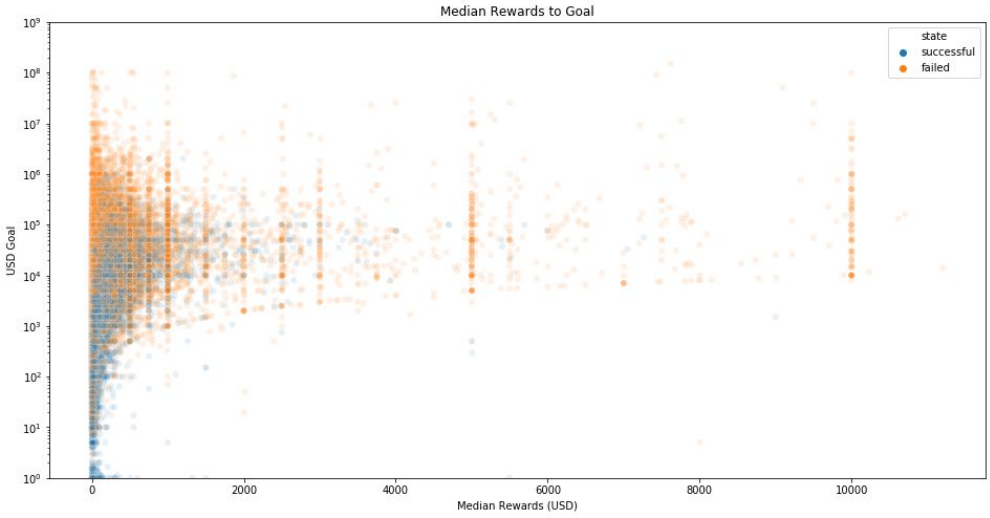
In addition to descriptions, Kickstarter projects also include blurbs as a way to entice people to click into their project.

Similar to the descriptions, “one” becomes a key term used in a lot of these blurbs. Again, it’s use varies but most use it to describe their project in the superlative (“one of the most”).

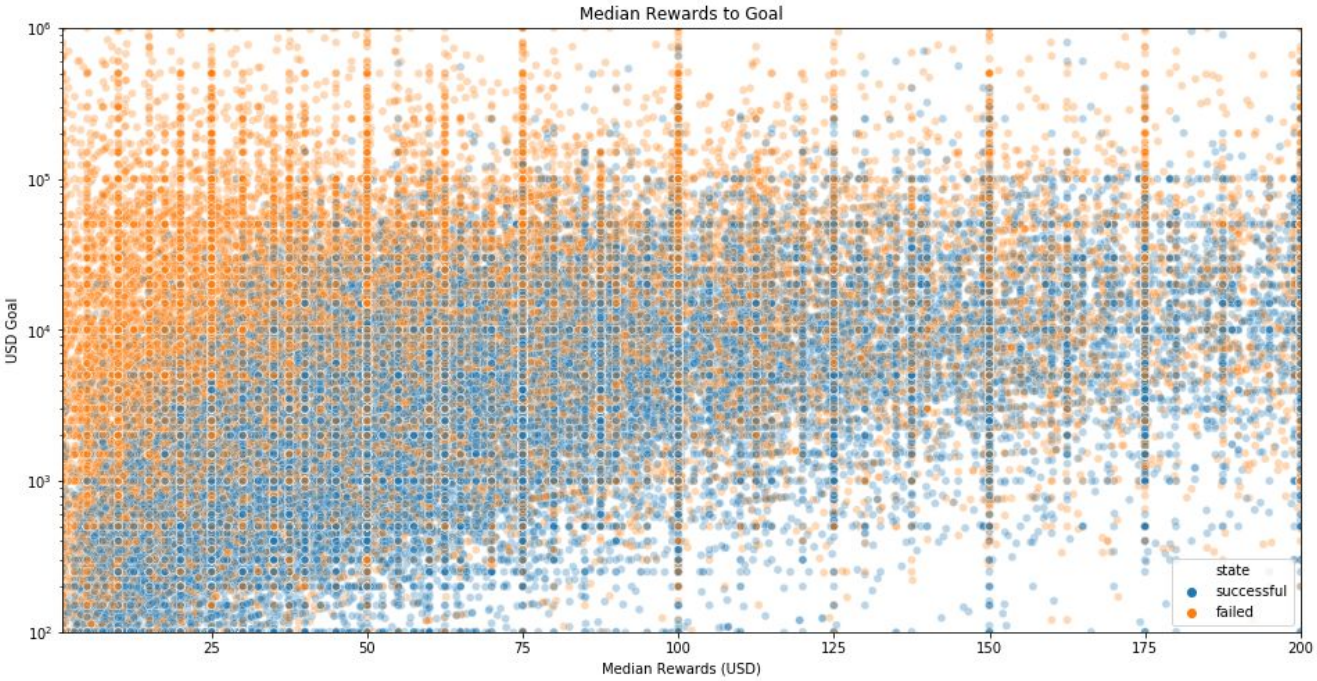
The blurbs also use the word “world” frequently. The usage generally falls into two categories, the world as a whole or the world of the product they want to enter. Things like “world premier” or “around the world” describes the first category while phrases like “word of wine”, “art world”, or “word of sugar” describe the second.

Projects by Rewards

Let's explore how rewards for Kickstarters can affect their progress.

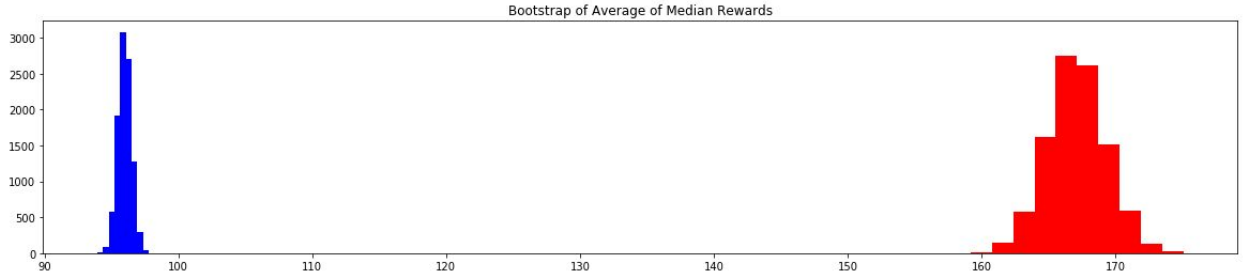


Here we've put the median reward price in relation to their goal. We can see that most Kickstarters that have a median reward of over 4000 usually fail. The median rewards over 200 actually encompass such a small amount of Kickstarters. Let's zoom in and see if we can focus on a core percent of projects.

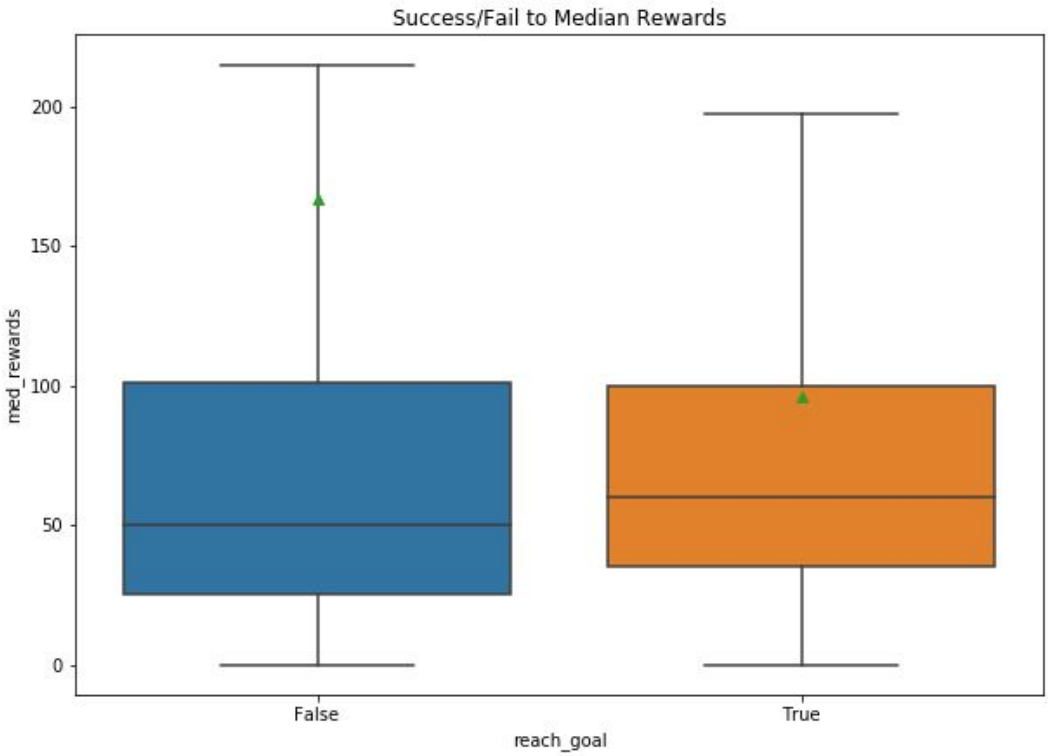


Here, we're only looking at projects with a median reward less than 200. Right away, we can see that median rewards less than 25 and goals more than 10,000 usually fail. There's a nice curve of success vs fail as we increase in median rewards. It makes sense that if you have small reward prizes, it'll be hard to reach bigger goals. However, it's not all successful since there's still patches of failures between the successes if you're in the higher median v goal area.

Let’s also explore the differences between whether successful projects have a different range of median rewards compared to failed projects. Taking a bootstrap sample of the average of the median rewards we get a stark difference.



We can see that the median rewards for successful projects do not resemble median rewards for failed projects. If we perform a t-test on the two groups, we get a p-value of 1.21e-215. That is incredibly small and well below an alpha of 0.05. From this, we can gleam that successful projects tend of have rewards that are more affordable for people. However, we can see that the mean isn’t always the best way when it comes to dealing with large values. If we take a look at a boxplot we can see right away that successful and failed projects have similar looking spreads. The median of failed projects are actually less than the median of successful projects. The means of each respective group is included as green dots and from this, we can see how impactful outliers are, namely those failed projects that had high goals and high reward costs.


Success/Fail to Median Rewards
200
150
100
50
0
med_rewards
False
True
reach_goal