Data Science Bootcamp Northwestern University

Data fundamentals in Excel

Student: Jorge Daniel Atuesta

February 24, 2021

Abstract

Kickstarter, one if not the most successful massive crowdfunding service, has been helping people bring creative projects to life since 2009 (Kickstarter, 2021). Based on the source Statista (2020), it is safe to say that with over $4.9 billion raised through the years, not all launched projects are successful. As of November 2020, the success rate of fully funding a project on the crowdfunding website was 38.21% (Statista, 2020). What happens to the rest of the projects? Well, some fail, others get canceled, and some are currently live. Using Microsoft Excel to organize and analyze a 4,000+ past Kickstarter projects database, we have identified the following trends: 1.If you are looking to launch your project using the Kickstarter platform, you have a higher chance of reaching your financial goal within the given time if your idea falls in the "music" category. 2. To further increase the likelihood of success in the Kickstarter platform, you can direct your project to be within the subcategory "rock." 3. According to the results, the ideal month of the year to launch your Kickstarter camping is in May. After identifying the trends, we identified the following limitations of the data collected: the data set does not represent to the full extend the entire population of Kickstarter projects. The data that is being analyzed could be updated to newer data from recent years. This data will not be able to make accurate forecasting modeling.

Keywords: Kickstarter, Database, Microsoft Excel, Data Analysis, Data trends.

Contents

Data Fundamentals in Excel

With over $4.9 billion raised in the past 11 years, Kickstarter has launched itself as one of the biggest crowdfunding platforms for new and upcoming businesses (Kickstarter, 2021). But not all projects that are launched on the platform find success. A study conducted by Statista (2020) publishes that only 38.21% of the launched projects reach their initial funding goal within the given time limit. This leaves a 61.79% unsuccessful rate. This rate is divided into the following categories "Failed" did not reach the intended financial goal within the time given, "Cancelled" the project is canceled due to lack of funding or time. "Unsuccessful" the project does not meet the funding goal within the given timeframe (Yeshitla, 2017).

Entrepreneurs are constantly on the hunt for ideas and the financial support to launch their ideas. Kickstarter serves as that bridge between a creative mind and the capital to start a successful business. With the crowdfunding option as a means of income, Kickstarter helps thousands of inventors, creators and entrepreneurs launch their business ideas every year. As mentioned previously, the success rate is only one-third, so what is it that one business idea has that the other doesn't allow it to become successful. We organized and analyzed a database with the information previous 4,000+ Kickstarter projects.

Results

As a starting point, we deconstructed the following question: what does it take to succeed when utilizing the Kickstart platform? To achieve this, we defined the possible states that a project could be in. There were four states 1. Successful 2. Failed 3. Canceled and 4. Live. With this, we proceeded to continue to organize, filter, and analyze the data. According to this data, the results indicate that the project has a higher possibility of success if it's under the category music and subcategories rock. It also shows that the campaign will be more successful if it launches in May. Here are the analyzed results of the dataset:

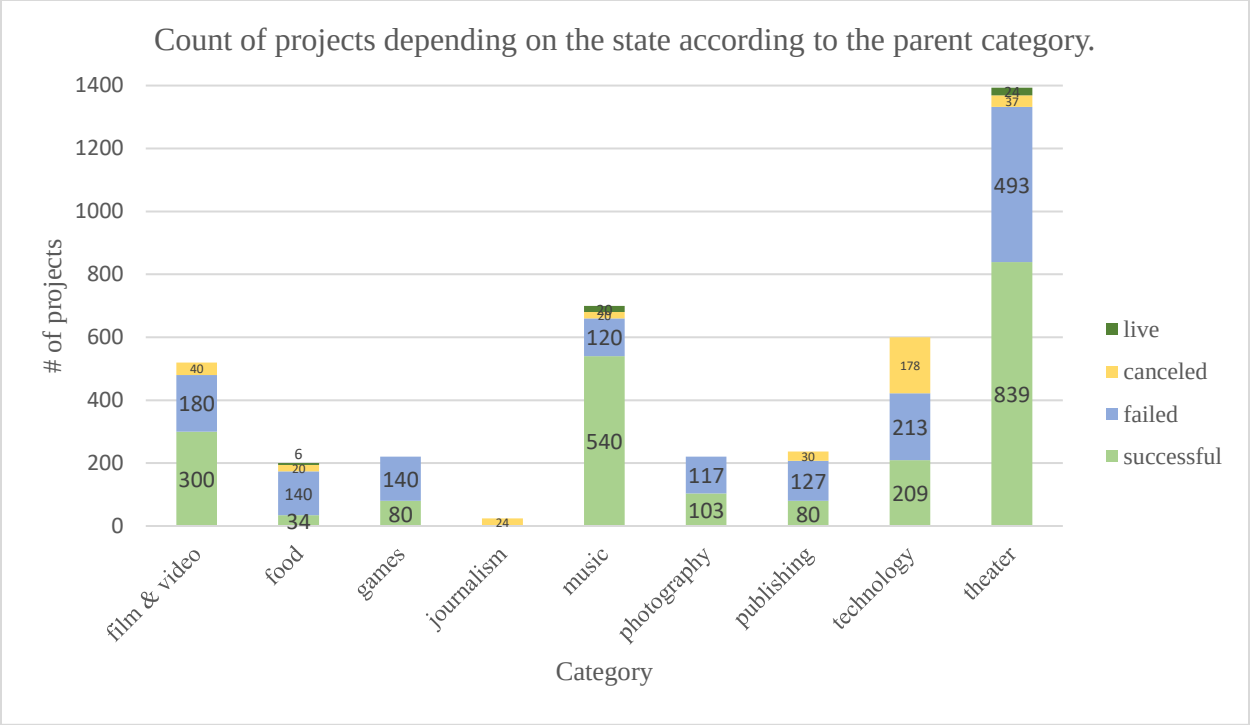

Figure 1. Count of projects depending on the state according to the parent category. This figure represents the count of projects within each category that has "Succeed," "Failed," "Canceled," or "Live."

According to *Figure1.* Count of projects depending on the state according to the parent category. It is safe to state that the category with the most successful projects is "music." With 540 cases of success and only 140 cases of failed or canceled, this category has a higher success rate than the other eight. On another note, the category "theater" has a total of 1393 projects launched on the platform, over 693 more projects than the following category "music," pointing out that theater is the most popular category when it comes to the number of projects launched. The technology field shows an increment in future projects and attraction for creators. Being the category with most "live" projects currently and having the third-highest count of projects places this category as the "up and coming." Finally, the data points the category of "journalism" to have raised some interest having 24 live projects. On the other hand, photography is the only category with cero live projects potentially showing that it didn't increase interest when the data was taken.

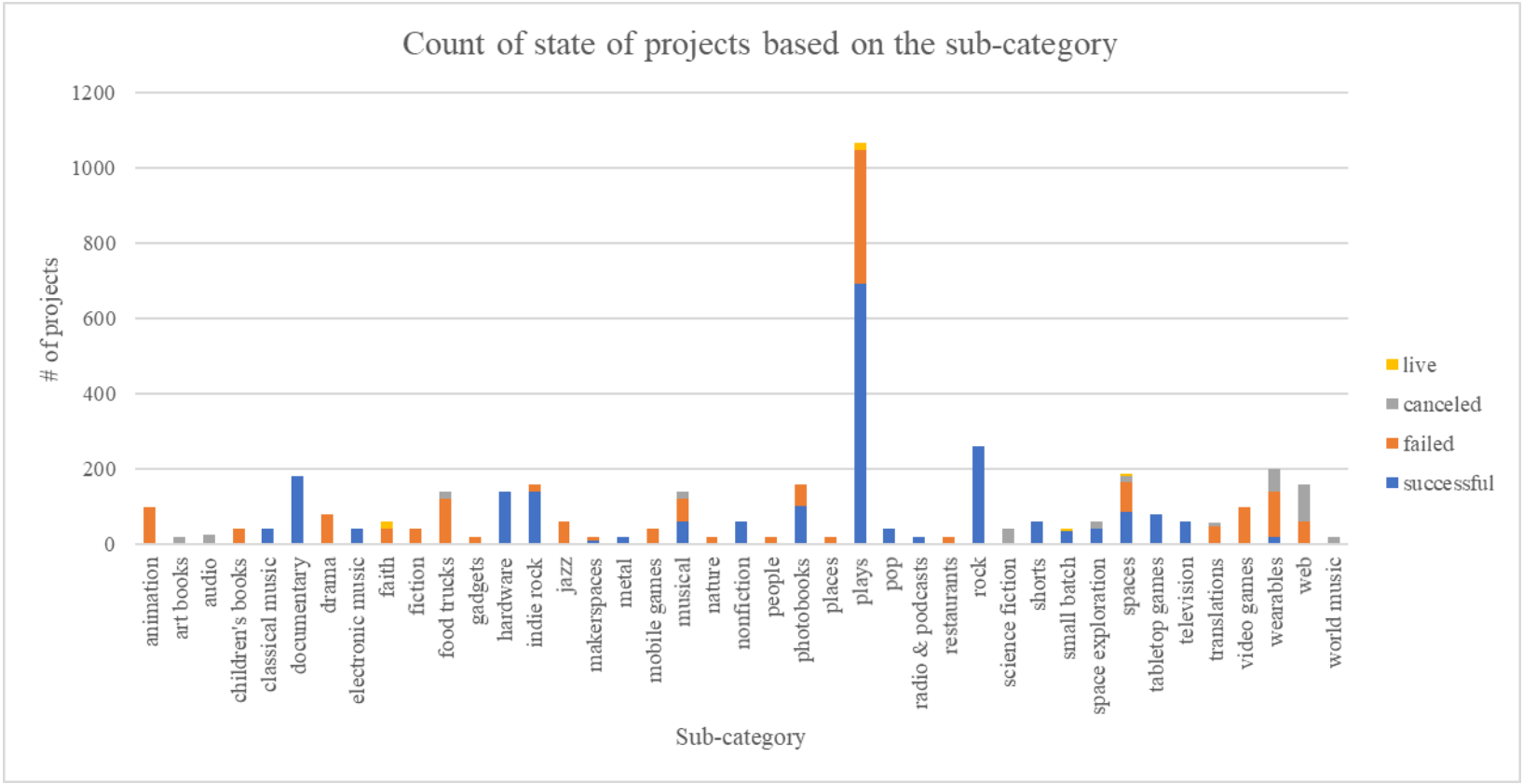

Figure 2. Count of state of projects based on the subcategory. This figure represents the number of schemes launched according to the subcategory and the current state.

The subcategory "plays" from the parent category "theater" has the overall most projects launched based on the data. The subcategory "rock" from the parent category "music" has the higher success rate of all the projects undertaken with 260 successful cases, followed by the subcategory "documentary" from the parent category "film" with 180 successful campaigns. The subcategory with the most canceled campaigns is "web," and the subcategory with the most "failed" campaigns is "plays."

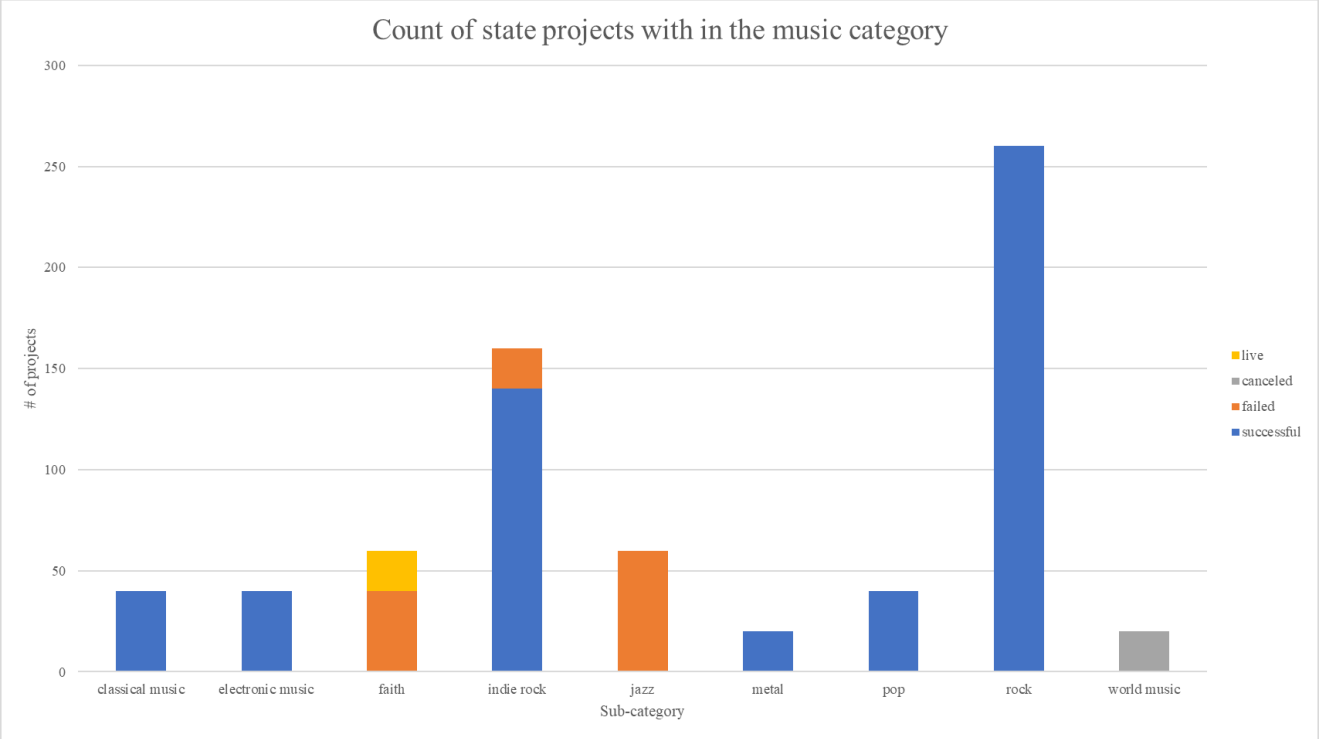

Figure 3. Count of state projects within the music category.

The subcategory rock has the highest success rate from the sample, whereas the "jazz" subcategory has the highest failed score. "world music" is the only subcategory that has been canceled, and although "faith" has some failed projects, it's the only subcategory that has up and coming campaigns illustrated on the figure by "live."

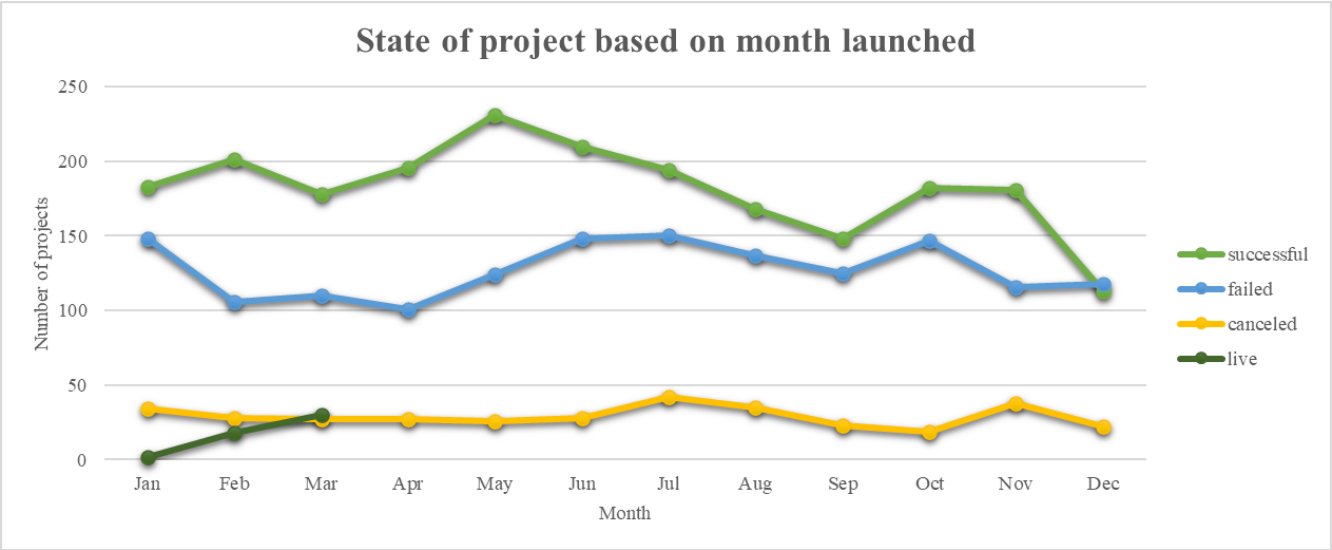

Figure 4. State of project based on month launched.

According to the data, the most effective month to launch a campaign on Kickstarter is May. This is independent of the year, category, or subcategory. On the other hand, the less effective month to launch the campaign, according to the data, was April.

Conclusions

After organizing and analyzing the data provided of over 4,000 Kickstarter campaign, the following conclusions were made:

1. If you are looking to launch your project using the Kickstarter platform, you have a higher chance of reaching your financial goal within the given time if your idea falls in the "music" category.
2. To further increase the likelihood of success in the Kickstarter platform, you can direct your project to be within the subcategory "rock," as shown in *Figure 3.* Count of state projects within the music category. With this said, you also have a high, more substantial chance of making your campaign successful if it falls under the category "film & video," specifically under the category "documentaries," as shown in *Figure 2.* Although the "theater" category and its subcategories have the most projects launched under their classes, it does not mean that it is the most successful or most comfortable to raise crowdfunding, as shown in *Figures 1 & 2.*
3. According to the results, the ideal month of the year to launch your Kickstarter camping is May, as shown in *figure 4.* With higher successful projects this month, it's recommended to utilize this knowledge to pick a month to execute a new project.

Limitations

The following limitations to the dataset were found:

1. Not an accurate representation of the entire population of Kickstarter projects.
2. The data that is being analyzed could be updated to newer data from recent years.
3. There is a significant variance in the data of the set.
4. This data will not be able to make accurate forecasting modeling.
5. The precision of boundaries

Diving more in-depth with the data

To make more data-driven decisions, here are some additional tables and graphs that could be included.

1. A graph showcasing the difference between countries and the state of the project.

2. Create a graph to see if the relationship between the Kickstarter feature "spotlight" and the project's state.
3. Graphs that filter out the data based on each category and subcategory to get more in-depth analysis.
4. Graph year to year to see if the year influences the month the campaign is being launched.
5. A graph showing the number of backers and the state of the project. To analyze whether or not the number of bakers can determine the project's outcome.

References

Kickstarter. (2021, February 17). *About Kickstarter*. Retrieved from Kickstarter: https://www.kickstarter.com/about?ref=global-footer

Statista. (2020, November 10). *Percentage of successfully funded Kickstarted projects as of November 2020*. Retrieved from Statista: https://www.statista.com/statistics/235405/kickstarter-project-funding-success-rate/#:~:text=This%20statistic%20shows%20the%20share,gone%20into%20successfully%20launched%20projects.

Yeshitla, L. (2017, June 5). *Analyzing success rates, failure rates, & trends of Kickstarter and Indiegogo products*. Retrieved from medium: https://medium.com/@lydiayeshitla/marketing-research-project-894c0a46c959

Bonus statistical analysis

Use your data to determine whether the mean or the median summarizes the data more meaningfully.

The critical factor in identifying when using the mean or the mode is the distribution of the sample. If the sample's distribution is "normal," then both median and mean will be similar. In this case, you could use either one. Usually, you would use the mean. If the data is skewed, then the mean and median will be drastically different in some cases. If this happens, the median, which is the value located in the middle of the sample, will be used to analyze the sample. The reason behind this is that the median is robust against outliers while the mean is sensitive to outliers generating an impact on future statistical analysis (Schremmer, 2017).

Looking at figures 5 & 6, both samples are positively skewed with a long tail to the data's right (Glen, 2014). It is essential to distinguish that if we were to use the mean to statistically analyze this data, we would be making a costly mistake as it would not be the correct amount. Therefore, we need to use the median result, as mentioned above, because it will be more robust to the outliers.

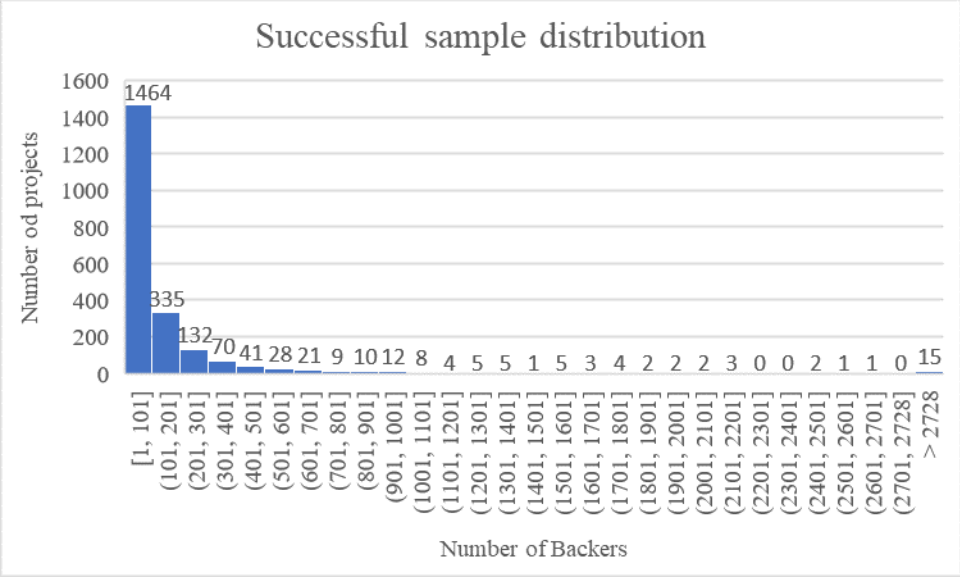

Figure 5. Successful sample distribution. This figure represents the distribution of the successful cases in our dataset. The data represents a positive skewed distribution for statistical analysis. It's essential to use the median rather than the mean. It is important to note that the bin width and overflow bin have been modified to visually represent the sample.

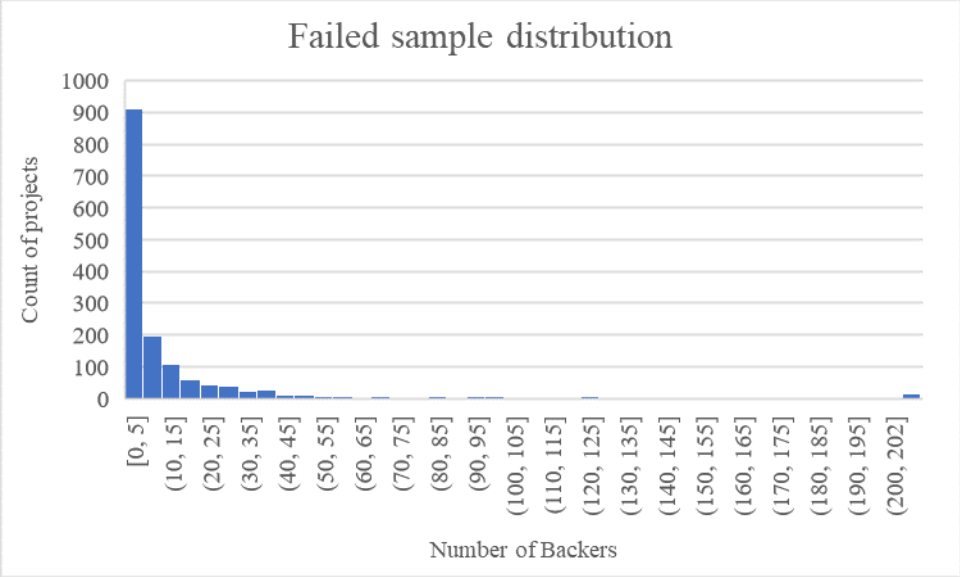

Figure 6. Failed sample distribution. This figure represents the distribution of the failed cases in our dataset. The data represents a positive skewed distribution for statistical analysis. It's essential to use the median rather than the mean. It is important to note that the bin width and overflow bin have been modified to visually represent the sample.

Use your data to determine if there is more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?

From the data gathered from the dataset, it is essential to note that both samples' variance (successful and failed) is high. This refers to the spread of the data gathered. Table.1 F-Test Two-Sample for Variances allows us to conclude that there is more variability in the successful data than unsuccessful. This could also be because of the amount of outliers collected on data used for each calculation. If we want to further investigate the variances, we can see the ratio is too high, 188.89. Additionally, the F value is higher than the F Critical one-tail. Furthermore, the test could be used if we had a null hypothesis and hypothesis as it will tell us if we reject or accept the null hypothesis.

Table 1.

F-Test Two-Sample for Variances

F-Test Two-Sample for Variances

	backers_count	*backers_count*
Mean	194.4251716	17.70980392
Variance	713167.3791	3775.689439

Observations	2185	1530
df	2184	1529
F	188.8840146	
P(F<=f) one-tail	0	
F Critical one-tail	1.081009586	