Crowdfunding Using Kickstarter

# Problem

Crowdfunding is a method of raising capital for an idea without using traditional sources such as a bank, venture capital, record label, or any other well funded corporation. Since such corporations have a responsibility to their investors and shareholders, they tend to allocate their funds to safer investments and definition of safe can be highly subjective. Due to a corporation’s hierarchical nature, a select few such as the C-suite, executives and senior managers, decide whether an idea gets funded based on their perception of the market and their pre-existing biases. Kickstarter is a website that provides a platform for entrepreneurs to pitch their ideas to a broader population. As a result, Kickstarter provides an alternative to the existing financial institutions by democratizing the decision-making process to a wider audience who are potential consumers and/or people who share the same ideals as the owner of the intellectual property.

[Since Kickstarter’s inception](https://www.kickstarter.com/help/stats)[[1]](#footnote-1), approximately 9.5 million unique users have pledged more than 3.5 billion dollars to close to 392,100 projects. However, only 139,500 campaigns have successfully achieved their target which translates to a success rate of 35.95%. We can look into what makes a successful Kickstarter campaign.

# Data Source

The Kickstarter dataset has been collected from early 2009 to end of January 2018 and made available from [Kaggle](https://www.kaggle.com/kemical/kickstarter-projects)[[2]](#footnote-2). For each campaign, the dataset contains its name, general and sub category, launch date, deadline, number of backers, state, the goal and the pledged amount for the campaign in its native currency and their USD values. The state of a campaign can be failed, success, canceled, suspended, live and undefined.

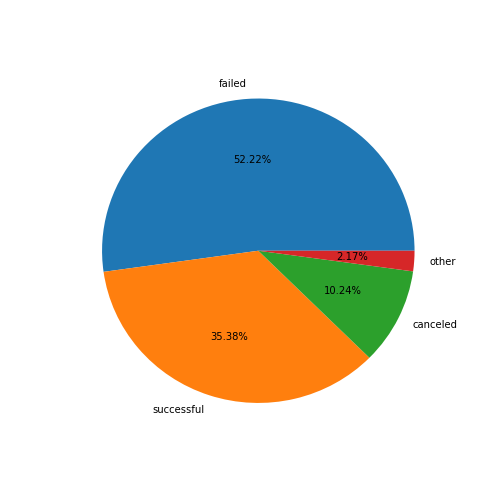
I have done the following preprocessing on the dataset:

* When loading the data, I was not able to use pandas’ default encoding to read the file. Hence, I had to use the chardet library to identify its encoding.
* The states live, undefined and suspended rarely occur hence, I have grouped them as other.
* I have calculated the campaign duration for each entry.
* I have created a new column that records the month project was launched to aggregate the data monthly.
* I have converted a subset of the data to JSON to be used by D3.

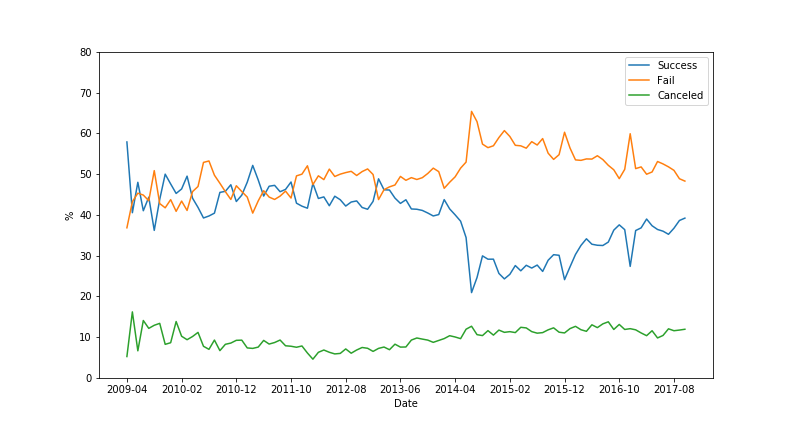
# Technologies

To process the data, I have used Python’s Pandas and NumPy libraries. To create the visualizations, I have used Python’s Matplotlib library along with the JavaScript’s D3 package. To use the D3 visualization, I had to convert the data into a specific JSON hierarchy. While attempting this task, I learned how to use OrderedDict from the collections library.

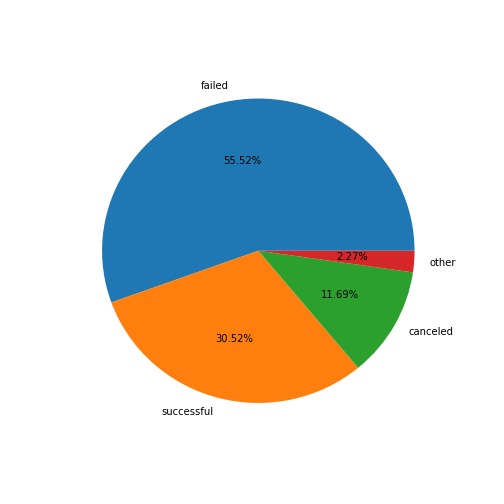
# Visualizations



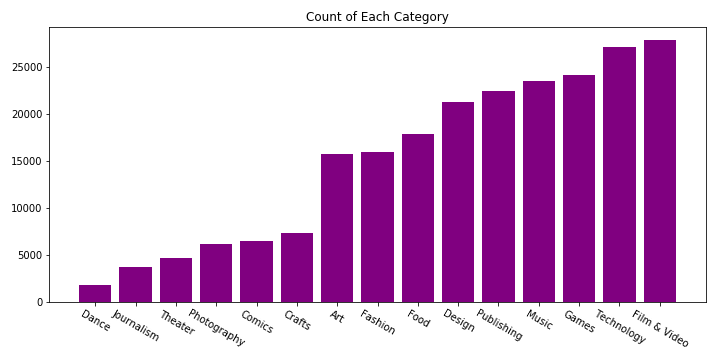
The goal of this project to understand what kind of projects succeed in Kickstarter. The pie chart provides a quick sanity check of the dataset, and it has a similar percentage of success to what is reported in the Kickstarter page. But, this raises an important question: Is this value a true representation of the current market? Change in customer and campaigner profile, site policy and other factors can easily change the success rate.



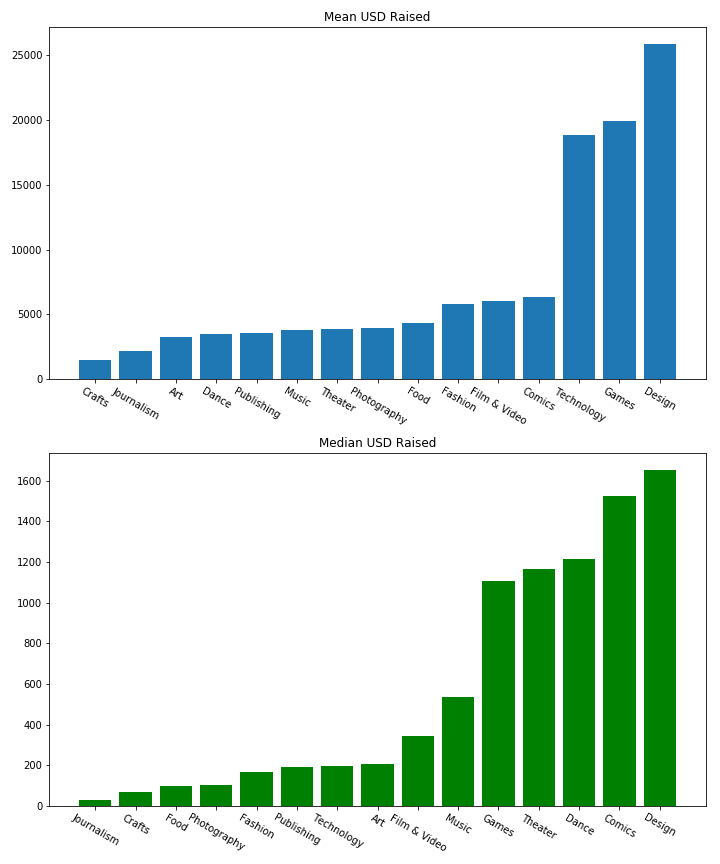
This visualization confirms my suspicions and shows that the pie chart is biased from the earlier stages of Kickstarter where more campaigns succeeded. The average success rate of campaigns hovered around 40% until mid 2014. Then we can observe a sharp decline in number of successful campaigns. The above visualization omits the last three months (2017 November to 2018 January) of the dataset because a good proportion of the campaigns are not finalized in that period.



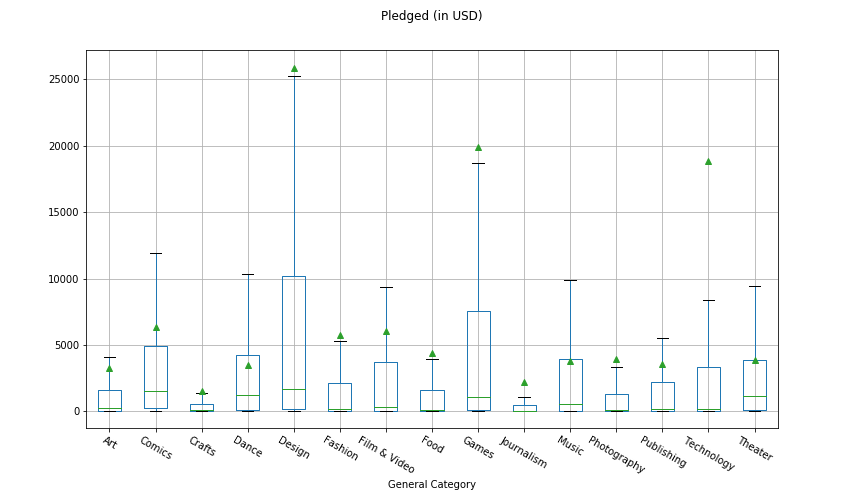
If we recreate the pie chart using only the data from *June 2014 to November 2017*, then we get a lower success rate. This rate is a more representative value of the current state of Kickstarter then using the whole dataset. *The rest of the visualizations will use this time range as well*. Now that we know the general success rate, we can look into more specific cases.



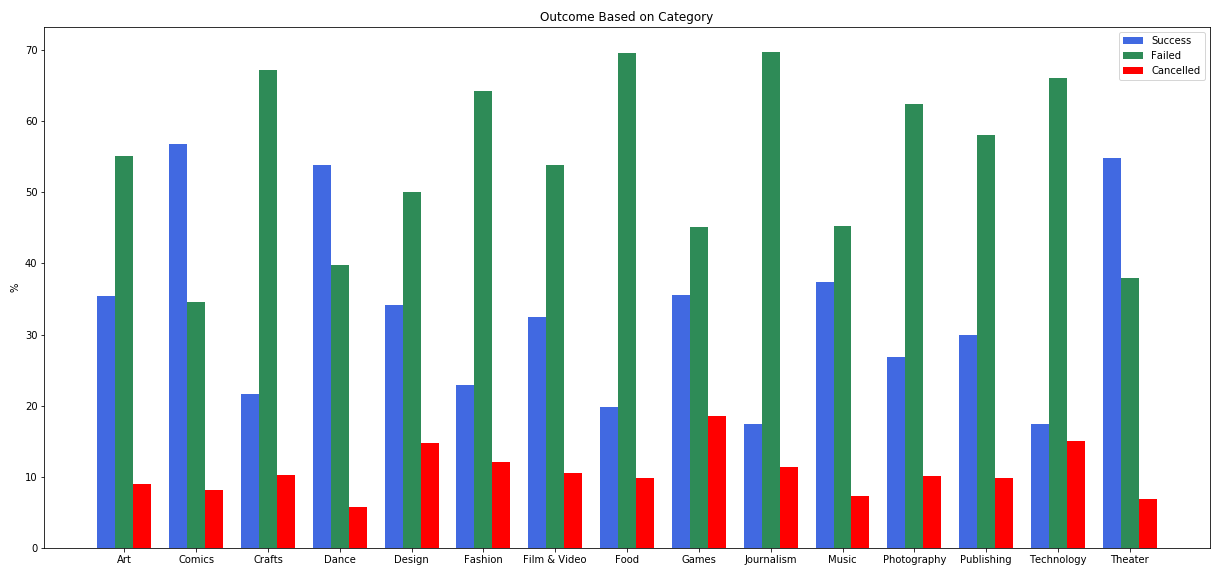
Before proceeding with the success rate of each category, it is important to understand the distribution of campaigns over categories. It seems Games, Technology and Film & Video are the most popular projects.



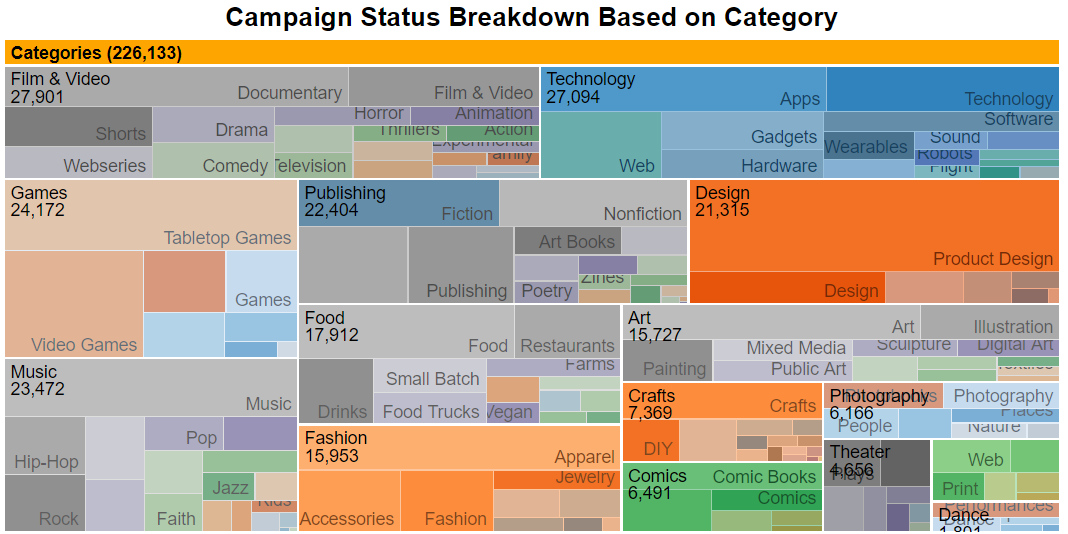
The median and mean USD raised by each category reveal interesting insights. Even though Technology is a popular category, it is highly skewed which can indicate Technology campaigns have a lower barrier to entry and/or more low-quality campaigns are proposed in this category resulting in less funds being raised. However, Technology also has a high mean, which indicates most of the funds for Technology are captured by few campaigns. The Design category exhibits an interesting pattern where they always raise more funds then any other category. Finally, even though there are fewer Dance campaigns, they tend to raise more funds then the others.



In the boxplot the green lines represent the median and the triangles represent the mean. The outliers are omitted from the visualization to retain the scale of all the categories. As a quick reminder, the whiskers of each boxplot capture 99.65% of all the underlying data. The previous conclusions regarding Technology is further confirmed in this visualization. The Design and Game categories have the highest variance which can indicate the budget range required for each category. For example, a single game developer would require less funds then a group of game developers. The means of Dance, Music and Theater fall in the interquartile range. This can indicate, these campaigns tend to set realistic goals because they know their target audience and know how much funds the supporters will allocate to their project.

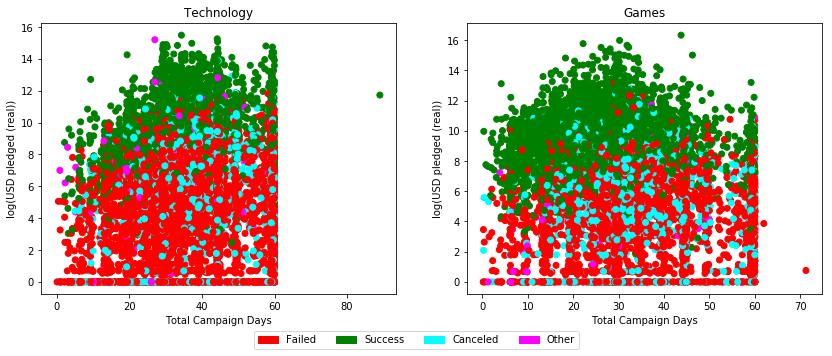


It seems Technology, Journalism, Food, Crafts and Fashion are the most difficult categories to succeed in. The reason Crafts and Fashion have a high failure rate is due to the subjectivity of the offered product. If the product is mainstream, most likely that a major corporation will produce it anyways. The reason Journalisms and Food campaigns tend to fail more often is they tend to have a limited audience. For example, a food campaign will be more likely to be perishable and a campaign located in Texas, might not be able to deliver to Toronto. Hence, that campaign will have a more local reach. Finally, the information Technology campaigns strengthens my conclusion that it has a lower barrier to entry. On the other hand, Theater, Comics, and Dance seem to have the most loyal audience. I believe these campaigns already have a pre-exiting audience and they simply use Kickstarter raise funds from them. Most likely, Music and Games have a similar pattern as well but there are simply too many campaigns in these categories causing a higher failure rate.



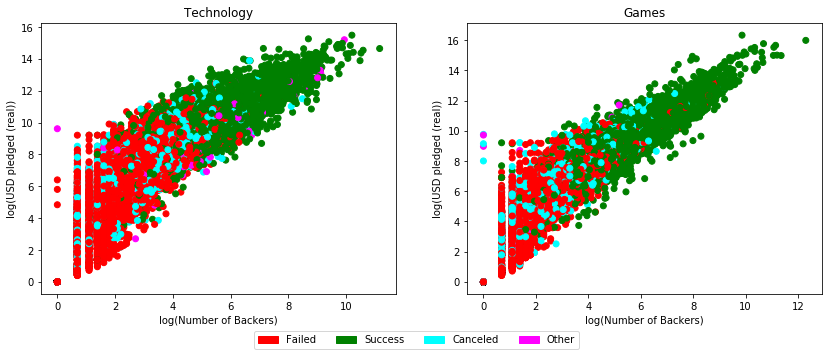
The above visualization further breaks down each category based on their sub categories. The user can access the state of a campaign by clicking each category. The source of the code is included in the zoomable.html file.

Let’s analyze whether Technology and Games are influenced by other factors. These categories are picked based because of the number of campaigns in each category and the high difference between their success rates.



I have applied a log transformation to the amount raised to prevent campaigns that raise more capital to skew the scale. In both categories, successful campaigns exhibit a reverse ‘v’ pattern. As the number of days increase, the amount they raise also increase until they reach a maturity and the amount raised starts to decrease. The peak for Technology happens between the 30 – 40 days range and the peak for Games happen at 30 days.

The difference between the categories can be caused by the nature of their offering. The market can be more efficient in identifying the utility of a technological device but once realized, the campaigns tend to raise more money. On the contrary Game campaigns most likely have an established audience, resulting in a high success rate in shorter campaigns. Also, campaigns that last longer than 30 days might be novel products resulting in a lower success rate and less funds being raised.



We can observe the positive relation between the pledged amount and the number of backers. Successful campaigns for the Technology category are clustered around a higher number of backers, however, there are higher number of campaigns that succeed with fewer backers for Games category. This can result from Games category offering more niche products. This assumption is strengthened when we compare the variance of the pledged amount. Technology campaigns attract a broader audience, hence it has a higher variance when we keep the number of backers fixed.

# Lessons Learned

* When aggregating results, one has to be careful about the changes in the trends. Otherwise, the aggregated result will not be a representative of the current state of the origin of the data.
* Technology, Film and Games are the most popular categories making them a more competitive space.
* Amount of funds raised do not always translate to success. For example, the Design category always raise more funds per campaign, but it has approximately have a success rate of 32% which is not too different from the overall 30.5%.
* Amount raised and the number of days a campaign lasts are not always positively related. I suspect brand recognition is at play.
* As expected, a campaign raises more funds if they have more backers. However, as the product is more niche, the target audience tends to give similar amount of funds.
* The analysis can be expanded by collecting additional data for the campaigns. For example, I suspect some of the campaigns have established audiences. Natural Language Processing can be used to expand the analysis to differentiate the campaigns that pitch new ideas from the campaigns that have a brand recognition. Additionally, if I can collect more data about the backers, some of the assumptions made in this report can be tested.

1. https://www.kickstarter.com/help/stats [↑](#footnote-ref-1)
2. https://www.kaggle.com/kemical/kickstarter-projects [↑](#footnote-ref-2)