**Conclusions from the initial data set**

* Kickstarter campaigns are far more prevalent in the United States than in any other country included in the data sample.
* Theater, music, and film/video are the three most popular categories of campaigns. Theater campaigns are especially popular, with the total number of theater campaigns being almost double that of music or film/video campaigns. Furthermore, plays are by far the most popular sub-category in the data set, representing almost 35% of all campaigns. The second most popular sub-category, rock music, accounts for only 8.5% of all campaigns, demonstrating the disparity between the popularity of play campaigns and that of all other campaign sub-categories.
* Campaigns created in June and July are the most likely to be successful, while campaigns created in August are the least likely to be successful. Alternatively, campaigns started in January are the most likely to fail, while campaigns created in September are the least likely to fail.

**Limitations of this data set**

Considering that over 75% of the data points are from campaigns based in the United States of America, this data set does not necessarily provide an accurate reflection of global trends in crowdfunding campaigns. No country other than the United States of America has more than 5% representation in the data sample. Entire regions of the world, namely Central and South America, Africa, Eastern Europe, and Asia, have no representation in the data sample. The absence Asian data in the sample is particularly problematic because roughly 60% of the world population can be found on the Asian continent[[1]](#footnote-1). China and India alone comprise roughly 36% of the world population and are two of the largest emerging markets in the world, yet they are not represented in this study[[2]](#footnote-2). Furthermore, the relatively small data sample from outside of the United States is more likely to be skewed because such a small sample is less likely to reflect the broader non-US population. The non-US data may therefore skew the entire data set. For this reason, the data set may not be an accurate representation of the US crowdfunding activity.

Concerning outcome categorization, it is not clear how a “cancelled” campaign is defined. A cancelled campaign could often be no different than a failed campaign. In many cases, the only difference between the two would be that the campaign organizers cancelled an otherwise failed campaign before the deadline. The ambiguous nature of cancelled campaigns could skew data to make certain campaign categories appear more successful, or at least less unsuccessful, than they truly are.

The categorization of campaigns limits the usefulness of this dataset. For example, what are the criteria for differentiating between a documentary and a journalism project? Similarly, a non-fiction publishing project could just as easily be categorized as journalism, but, in the dataset, these are distinct parent categories/sub-categories. The theater parent category proves to be particularly problematic. Theater campaigns are almost twice as prevalent in the dataset than any other campaign parent-category. Despite comprising almost 35% of the dataset, theater campaigns have only one sub-category: plays. The play sub-category is also 3.5 times more prevalent than any other sub-category in the data set, causing Figure 3 (campaign outcomes broken down by sub-category) to look very skewed.

*The final limitation in this dataset concerns the dates. The campaigns in this dataset were all created between 2010 and 2020. However, in the analysis, only the month in which a project was created is considered. Although trends emerge from this analysis, as seen in Figure 3, such analysis operates under the assumption that campaign activity has followed the same monthly patterns over the course of the last decade.*

Additional Tables and Charts

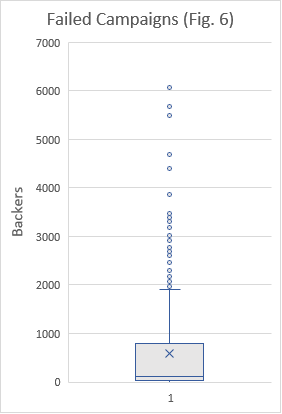
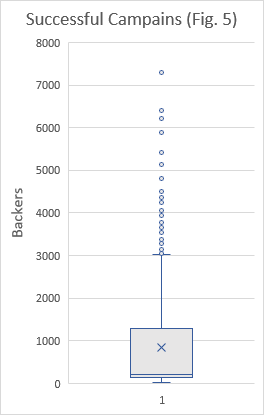
Some additional tables and charts could have been included in order to better analyze the dataset and to account for some of limitations mentioned above. The first addition would not involve adding another table of chart, but modifying an existing chart, namely Figure 1, Number of Campaigns Broken Down by Campaign Status (Successful, Live, Failed, or Cancelled) per Campaign Category. It would be useful to express each outcome (failed, cancelled, live, and successful) as a percentage of the total number of outcomes for each parent category. These percentages could then be added to each campaign category on the chart in order to make the chart more readable. This would also make it easier to compare campaign outcomes without having to consult the corresponding table.

A second addition could be a chart that tracks the number of campaign outcomes over time. Instead of the x axis consisting of only months, it would consist of every year and all the months in each year. Its basic structure would resemble a monthly stock chart, where the user can easily locate a value for any month within any year. This would make it possible to find long-term trends related to the time of the year. Furthermore, filters could be added to the table and chart so that the user can filter out all months except for one and see how campaign outcomes for a given month have changed over time.

As mentioned in the previous section, the ambiguity surrounding campaign outcome categorizations, particularly cancelled and unsuccessful campaigns, makes it difficult to accurately analyze campaigns based on their outcomes. If all cancelled campaigns were instead grouped with failed campaigns, Figure 3, Number of Campaigns Broken Down by Campaign Outcome per Month, would be clearer and more concise. Specifically, the y axis could be scaled in a way that would be easier to interpret data points for both successful and failed campaigns.

Statistical Analysis

Both the successful campaigns data set and the failed campaigns data sets have means that are significantly higher than the respective medians. Generally, a data set in which the mean is higher than the median is said to be right skewed. Looking at the box and whisker plots below (Fig. 5 and Fig.6), it is apparent that all of the outliers have values *greater* than 1.5 X IQR, confirming that both data sets are right skewed. In a right-skewed data set, the mean reflects skewed data, while the median is the better measure of central tendency. For that reason, the median value summarizes each data set more accurately.



The successful campaign data set has a significantly higher variability (variance = 1,630,374 and standard deviation = 1,266) than the failed campaign data set (variance = 921,575 and standard deviation = 960). This difference in variability makes sense because failed campaigns have an upper-bound on donations, defined as the campaign goal. By this same logic, the number of backers in a failed campaign is also subject to an upper bound. In other words, if a campaign exceeds its goal, then it is no longer categorized as unsuccessful. Unsuccessful campaigns can therefore be said to have a smaller possible range of backers. Successful campaigns, however, are not subject to these restraints. For successful campaigns, there is no limit to the amount of money that can be raised, allowing the amount of money raised, and by extension the number of backers, to vary more greatly on the right end of the dataset.

1. https://worldpopulationreview.com/continents/asia-population [↑](#footnote-ref-1)
2. https://statisticstimes.com/demographics/china-vs-india-population.php [↑](#footnote-ref-2)