Vinayak Shankar

UC Berkeley Data Analytics Boot Camp

Crowdfunding Report

Given the provided data, what are three conclusions that we can draw about crowdfunding campaigns?

Based on the given dataset:

* Most crowdfunding campaigns are concentrated in three main categories: 1) film & video, 2) music, and 3) theater; most crowdfunding campaigns are also concentrated in one main sub-category: plays.
* Most crowdfunding campaigns are based in the United States.
* There is evidence to support a seasonal effect on the success of crowdfunding campaigns; there is an increase in the number of successful crowdfunding campaigns between May and July; there is also a decline in the number of successful crowdfunding campaigns between July and August. The total number of crowdfunding campaigns from year to year appears the same, around 100 per year.

What are some limitations of this dataset?

* This dataset only has results for 1000 crowdfunding campaigns and only from seven countries.
* In comparison, Kickstarter (one of the most well-known crowdfunding platforms) published data showing that they have hosted the launch of 585,379 crowdfunding campaigns to date (source: <https://www.kickstarter.com/help/stats>).
* Another popular crowdfunding website, Indiegogo, claims that their platform launches “thousands of campaigns every week” (source: <https://www.indiegogo.com/about/what-we-do>).
* Based on this, one major limitation of this dataset is that the sample size is incredibly small when compared to the entire population of crowdfunding campaigns and therefore might not be representative of the whole.
* Another limitation of this dataset is that it does not include any correlated marketing data from outside the crowdfunding platform itself. According to one published study, social media sites such as Twitter, YouTube, Facebook, etc. have profound impacts on the outcome of crowdfunding campaigns (source: <https://journals.sagepub.com/doi/10.1177/10949968221074726>).
* Another limitation of this dataset is that it does not include useful metadata about the people or entities that launched each campaign. For example, there is nothing in this dataset about how much money each project might have raised prior to launching its crowdfunding campaign; if a project already raised 90% of its capital prior to launching its crowdfunding campaign, that information would not be present in this dataset and could therefore be skewing the analysis (one hypothesis might be that if a project only needs to raise 10% through crowdfunding, it is more likely to succeed; such a hypothesis cannot be tested using this dataset).

What are some other possible tables and/or graphs that we could create, and what additional value would they provide?

* We could create histograms of the data distribution and they would add value by helping visualize outliers and showing if our dataset is skewed. I have created these histograms and included them in the Statistical Analysis sheet of the Excel file. These charts show us the distribution of the number of backers for successful campaigns and for failed campaigns.
* These charts help visualize the statistics rather than just displaying the values for mean, median, variance, and standard deviation in a table. The charts immediately show us that the dataset is not normally distributed and is skewed to the right with fat tails.
* This skewness might exist because the dataset is too small or because there are other strong variables for which this dataset does not account. For example, the histogram of the number of backers for successful campaigns shows a large spike at the far right end of the tail.
* One hypothesis might be that some outside factor such as a concurrent viral social media campaign or perhaps a celebrity endorsement could be greatly increasing both the number of backers and the success rate at the far end of the distribution.

Statistical Analysis

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

* Based on the histograms, the median appears to better summarize the data than the mean. For both successful and failed campaigns, the mean is far to the right of the median, which is consistent with the observed right skewed distribution. For both successful and failed campaigns, the median is much closer to the area where most of the backer count data is centered.

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

* The variance of the number of backers of successful campaigns is about 1.6 million compared to 0.9 million for failed campaigns. The standard deviation of the number of backers of successful campaigns is 1267 compared to 961 for failed campaigns. Based on these metrics, it appears that there is much more variability with successful campaigns than with failed campaigns.
* The salient questions here are: 1) Do successful campaigns truly have more variability in backer count than failed campaigns? 2) If yes, what factors might be correlated with campaign backer count variability?
* One reason that the answer to the first question might be “no” is that there was some flaw in the data collection method or in the dataset itself (for example: sample size is too small).
* Assuming that the answer to the first question is “yes,” one explanation for the observed increase in variability in the backer count of successful campaigns might be that there is no upper limit to how much money a crowdfunding campaign can raise. In contrast, a failed campaign cannot go below “zero” for money raised.
* Another reason for the observed increase in variability in the backer count of successful campaigns might be a confounding variable such as viral marketing campaigns on social media, celebrity endorsements, or the retention of crowdfunding campaign consultants. One published study hypothesizes that campaigns that hire consultants with proven experience in launching successful campaigns are more likely to succeed. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7203064/>).