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

“Theater” is the largest category of crowdfunding campaigns, specifically the sub-category “plays”, which is four times as large as the next sub-category. The “music” and “film & video” categories come in roughly tied for second-largest, with about half as many total as “theater”.

Summer, especially June and July, is the best time to have a successful campaign, but August is bad. Maybe there’s an end-of-summer effect that causes people to give less to crowdfunding campaigns?

Set a crowdfunding campaign in the range of $15,000-$35,000. Campaigns in the $15,000-$35,000 range tend to be more successful, verging on 100% success rates, but even campaigns in the $35,000-$50,000 range have a ~75% success rate. Campaigns in the $5,000-$15,000 range fare 50-50, a roughly even chance of success or failure.

Bonus fourth conclusion! Most campaigns tend to have smaller numbers of backers, in the low hundreds range, so a crowdfunding campaign doesn’t have to reach thousands of donors to be successful. A few hundred is probably sufficient.

* What are some limitations of this dataset?

No demographic data for donors, in case you are looking to target your fundraising dollars most effectively. Maybe people aged 45-60 are obsessed with funding documentary or music campaigns, but the 60-80’s are into animation and science fiction. Who knows?

It hasn’t been filtered. There are clearly some outliers in the data. Removing them would be helpful. Data cleaning is something I need to learn how to do.

Is it just me, or are the blurbs nonsensical? I guess it’s irrelevant.

It would be nice to know if there is more data on how the campaigns were marketed, beyond the spotlight and staff pick columns. Not sure what that data would look like. Maybe campaigns are helped by last-minute advertising pushes? What if there was a donor match campaign, where every dollar gets matched by Nonprofit.org up to $10,000? We don’t know this.

Limited country set. A cursory glance shows it focused to Europe, Australia, Canada, and the US, with the US representing over 75% of the data. This data won’t tell you anything about a crowdfunding campaign in India, South Africa, or Peru, and barely tells you anything about a country outside of the US.

Data barely enters 2020. Three-year-old data may be all that’s available for analysis, but recent events, like an unforeseen global covid pandemic, may alter crowdfunding behavior.

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

A visualization that shows the number of backers of failed and successful campaigns, that also shows the mean and the median. Maybe a scatter plot? I don’t know yet. But something that shows how a few outliers are skewing the data. The additional value would be that it tells the data analyst (me) and anybody else looking at the data that they need to ignore the outliers (clean the data).

Compare the success and failure rates based on “spotlight” (which, I presume, is if the website spotlights the campaign). Same with “staff pick”. It would show how useful “spotlight” and “staff pick” are to possible campaigns (so you can sell them for advertising dollars, which makes me feel uncomfortably Machiavellian).

* Does the mean or the median better summarize the data? Why?

I think the median better summarizes the data for backers, although they both provide useful information. The data is right-skewed, so the mean (average) has a significantly higher value (successful 851, failed 586) than the median (successful 201, failed 115). For passing curiosity, I checked the modes as well (successful 85, failed 1). Even the standard deviation is greater than the mean (std dev: 1266 successful, 960 failed, vs. mean 851 successful, 586 failed), which makes no sense at all (you’d have negative values for the number of backers of successful campaigns within one StdDev). Given all of that, I think you can pretty safely say that some massive outliers (maxs of 7295 and 6080) are pulling the data right-wards.

In sum, most campaigns tend to have smaller numbers of backers, with success and failure around 200 and 100, respectively. The mean is skewed by outliers.

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

There’s more variation with the successful campaigns, which makes sense. The variance for failed campaigns is roughly 45% less than the value for successful campaigns. You would expect failed campaigns to have less support, and therefore less variability. More importantly, there are more successful campaigns than failed campaigns, so you’d expect more total variability, as there are simply more data points with successful campaigns (about a third fewer failed campaigns compared to successful ones).