Crowdfunding Data Analysis

**Given provided data provide three conclusions relevant to crowdfunding campaigns.**

From our graphs we can draw some summary observations. First, “Theater” is by far the most popular category in our dataset with 344 observations, nearly doubling the observations (total number of outcomes) of the next two largest categories: “Film & Video” (n = 178) and “Music” (n = 175). This trend continues with the next category by size, “Technology”, falling by nearly a factor of 2 (n = 96). The smallest category is “Journalism” (n = 4). This variation in sample sizes should give us appropriate pause when making comparisons between categories, and we should be particularly skeptical of saying anything definitive about “Journalism” as a category. A robust statistical test would be required to test, for example, the hypothesis that there is a difference in success rate between categories. That said, from our graph we might conclude that it would be more difficult to gain visibility with a “Theater” project as opposed to a “Journalism” or even “Photography” one. The client would want to consider this in estimating marketing costs and market saturation.

When we drilldown to subcategories the sample size disparity is exacerbated as there is only one subcategory for “Theater” but multiple for many others. Looking at all the categories does suggest some reason for optimism as there is not a single subcategory with only failures, but there are a couple with only successes, namely “World Music” and “Audio”. Though these are particularly small sample size subcategories, they are indicative of the overall trend: success is more common than failure at a ratio of 565 to 364 in the complete dataset.

This success is spread relatively evenly throughout the year. If we look at the number of successes, failures, and cancelations as a function of the month the campaign was launched, we see a nearly flat line. Again we are stymied; without a robust statistical test to test the hypothesis that rates of success or failure are different across months we cannot say anything definitive. However, there is the suggestion of an early summertime effect, as June and July appear to coincide with a spike in success that is not mirrored in any appreciable decrease in raw number of failures.

Drilling down this stability disappears. Looking at graphs of singular years, any number of different relationships can be found. Variability is greatly increased. In part this is likely due to smaller sample sizes. To decrease this variability, and provide a look at more contemporaneous data, we’ll examine the last five years for which data is provided (2016-2020). Looking at this graph we see suggestive parallelism in the trajectory of data points such that if successes increases from one month to the next, the is a roughly equal increase in failures. This is just an eyeball test, but it might make us conclude that the odds remain the same regardless of which month we start in. Of course, a robust statistical test is necessary to say for sure.

**What are the dataset’s limitations?**

As we’ve discussed, the data has some inherent limitations beyond analytical methodology. A major one for the variables we looked at is sample size. To compare between groups we need appropriate sample sizes or our conclusions may be erroneous. Another one is the temporality of the data. The data span what seems to me to be a decent amount of time, however there is the danger that market trends changed over the course of that time such that a lot of this data may actually be misleading. More broadly there might be limitations related to the data’s procurement: did the data come directly from the crowdfunding website? Are they a sample of campaigns, or are they all the campaigns? Are variables like “category” drawn from the crowdfunding website’s terminology, and are they consistent across websites in the case that these data are amalgamated from multiple sources?

**What other tables or graphs would provide additional value? How?**

A bar plot of percentage succeeded per category and subcategories would complement bar plots we made. It would provide a more direct visual comparison of the success to failure ratio that gives us one measure of the chance of success in a given category.

Another visualization could be done with a new variable “Duration of campaign” that measures the number of days between “Date created conversion” and “Date ended conversion”. We could then make a scatterplot of “Pledged” as a function of “Duration of campaign”. If we color code the data points according to “Outcome” then we could get a visual intuition for the impact of Duration on the outcome and the amount raised. We would see if failed campaigns are clustered, for example, at the low end of campaign duration. Simultaneously we could assess if longer campaigns always accrue more money, which we might intuitively think, or if there is a saturation point. A downside of this plot could be that “duration of campaign” might be confounded with “goal” such that campaigns with a low goal also end up with lower pledged amounts regardless of the duration of campaign.

A third graph could be a boxplot of “goal” broken up by category. For this graph I would suggest subsetting the data to only successful campaigns. Then we would have an idea of what a reasonable goal for a successful campaign in X category is. We could see if there is small or large variance in. If it is small then we would want to think seriously about setting a similar goal. If variance is large we might conclude the goal is either not important or campaign specific. To help distinguish between the two possibilities we could then take the categories with large variance and plot boxplots for their subcategories.

**Statistical Analysis**

The median serves as a better summary of this data due to the large right skew in both successful and failed backers. The variability is larger for successful campaigns which intuitively makes sense. Failed campaigns are by definition going to have fewer backers to begin with. We see this, for example, in the minimum value of each dataset, where failed campaigns have a 0 minimum of backers compared to 16 for successful campaigns. Failed campaigns are likely to be zero inflated and have total number of backers clustered towards this lower end.