Alex Tsai

4/15/2023

**Excel Challenge Report**

* Some Data Conclusions:
  + It appears that many campaigns in this sample are of the theatrical nature with roughly 35% of them having a sub-category of “Plays” and a primary category of “Theater” – notably, the only subcategory that “Theater” has is “Plays”. Other prominent categories include music and film at 18% each, making the majority of the campaigns here related to the creative arts.
  + Surprisingly, over 55% of the campaigns in this sample have succeeded. I was expecting more of a 90-10 split between unsuccessful and successful campaigns because crowdfunding is another phenomenon of the internet and only an extreme minority of user-generated content ever gets any attention at all.
  + It seems that time doesn’t have a strong influence on campaign success rates. Given that campaigns can last for quite a long time, it’s reasonable to suspect that there’s no purpose in assessing how people’s generosity vary over the months. The signs of seasonality could be due to noise and chance. It is notable how success rates have a small spike around the middle of the year. If we had information about the kind of people making those donations, perhaps we could make guesses about why that minor increase in success rates exist – perhaps younger people are more susceptible to impulsive spending during the summer break, only to regret it when they review their balance sheets at the beginning of their school year.
* Data Limitations:
  + The sample size here is only 1,000 campaigns out of the over 500,000 projects Kickstarter alone ever had. The data we have here is not remotely representative of the crowdfunding industry and the sample data breaks down especially quickly when we break what little data we have into even smaller categories.
    - For example: When breaking down campaign outcomes by proposed campaign goals, we have categories such as the “$30,000 to $34,999” category that have a total population of 7 campaigns. This is not a good size for statistical analysis. It’s like a boxer bragging about being undefeated because he fought one person and retired after that one match – the record is worthless.
    - I would like to draw conclusions about how increased funding goals may be inversely proportional to campaign success rates, but I do not have enough information to elaborate upon that claim with the majority of the sample dataset being localized within the small intervals of <10,000 and >50,000 backers. Yes, we see that failures overtake successes at the end and generalize about how “expensive bad, backer leave”, but it’s not really that informative when the middle of the intervals is so sparsely populated that we don’t get to see how the curves converge.
  + Speaking of campaign goals, data inconsistency is also a problem. It is not clear based on the data whether the columns with financial data are even standardized under the same currency. Considering how there is a separate “Currency” column in the raw data, it is completely possible that the charts I have generated is not at all representative of the actual transactions that have been made on this platform.
  + Additionally, the small sample size here also calls into question the reliability of our category-based pivot tables: Are there really that many theater/play projects, or did we only capture that sample by chance?
    - GoFundMe, for example, is a platform greatly known for people begging other people for financial aid in order to pay for medical bills that should not have been that expensive in the first place. As different platforms have different focuses and niches, the sample data we collected may not be representative of the overall online crowdfunding space.
* Other data visualizations:
  + For exploratory data analysis, I would also generate a histogram of the goals themselves to see how much money people are asking for in the first place. I would also generate a pivot chart to see how much success each category of crowdfunding campaigns is getting. Perhaps tech gadgets and things involving cryptocurrency garnered more traction because pump and dump schemes were more popular a few years ago.
  + Other columns include the average time each campaign took – it seems there are campaigns that finished in the same day, so a distribution of how long campaigns last amongst successful ones and those that expire would be interesting.
  + It would also be interesting to see the distribution of funding goals between the different categories as there are successful campaigns with 20 times the funding they asked for. It would be interesting to see how successful the successful campaigns are on a histogram.
  + Similarly, I would break down other columns based on these categories. Simple bivariate analysis of some columns may bring new insight to the dataset. Perhaps arts and crafts projects have more success not because of high demand but because of lower overall expected budget, making their goals more achievable in the grand scheme of things.
* The Better Measure of Central Tendency:
  + Much like how Jeff Bezos can walk into a homeless shelter and turn the average person into a multi-billionaire and much like how almost all tweets online will never have more than 5 likes, the number of backers supporting campaigns is a statistic that is greatly right-skewed.
  + The means are greatly influenced by the outliers of the data set as roughly 70% of the distribution is under the mean for both successful and unsuccessful campaigns. Due to the existence of great outliers, the standard deviation is also huge, making the mean not a great measure of central tendency for the dataset at all. Due to the reasons listed above, it is better to go with the median to assess how many backers a typical campaign is getting.
    - Incidentally, one standard deviation left of the mean reaches negative numbers. While it would be hilarious to assume that some 15% of people are losing backers by setting up a crowdfunding campaign, this is an observation that in practice makes no sense.
  + It would be more informative to notice how 50% of all failed campaigns have less than 114 backers and compare that with the much higher median of successful campaigns at 201 backers. In general, it seems as if successful campaigns are twice as popular and well-supported as unsuccessful campaigns.
* Regarding Variance:
  + According to the data generated, successful campaigns have a much greater variance than unsuccessful ones, which makes sense because successful campaigns, in general, receive more money.
  + One may notice from the other descriptive statistics that the mean, median, minimum, and maximum backers received for successful campaigns are greater than unsuccessful ones. This makes sense because if unsuccessful campaigns received more money, they would not be unsuccessful in the first place. By definition, unsuccessful campaigns are those that did not receive enough money, thereby making the range of backers shorter amongst all campaign goal ranges as they were never popular enough to get enough backers to donate enough money.
  + Furthermore, the discrepancy in variance can be easily explained by going to the Percent Funded column on the spreadsheet and ordering the numbers from greatest to least. While unsuccessful campaigns are restricted between 0% to less than 100% funding, there is no upper bound for successful campaigns. There are 24 campaigns alone that received over 1000% of their target goals. Simply because of data definitions, the nonexistent upper bound of successful campaigns allow for greater variation than the restricted range of the unsuccessful campaigns.