U of T SCS Data Analytics Boot Camp

Module Challenge 1 Written Report

Thursday June 14, 2023

Gurpal Singh Gill, X435733

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

From Tab “Pivot Table 1”, we can see that the dataset is derived from campaigns mostly in the theater category, followed by film & video and music. These appear to be the most popular categories. However, in proportion to total outcomes of each category, the most successful campaigns were from the technology category (66.67% success rate). It is important to note that the success rate of technology is not far off from other runner ups such as photography, and publishing. On the contrary, the games category had the highest failure rate of 47.91%.

As per Tab “Pivot Table 2”, the plays sub-category generated the greatest number of campaigns. This is no surprise as plays was the only sub-category of theater, and the theater category also had the greatest number of campaigns as well. Furthermore, there is more data in the plays sub-category than all other categories, and that too by a landslide.

Lastly, Tab “Pivot Table 3” helps break down the outcome of campaigns based on months. From this, we can determine that the months of June and July appear to be the peak months. These months present the highest number of successful campaign months of the entire year. Furthermore, the month of August seems to be the least attractive month, as it had the least number of successful campaigns and the greatest number of failed campaigns. Fund seekers will want to avoid this month.

**What are some limitations of this dataset?**

The first noticeable limitation of the dataset is its parent categories. These categories appear very similar in nature as they relate mostly to arts and entertainment (except for food and technology. As such, the categories raise the question of potential bias by ruling out other categories like sports and learning, and the data may not show a true reflection of successful crowdfunding campaigns.

Furthermore, the data appears to be very saturated. Majority of the data points fall within the theater category (at least 34%), and there is a large gap between this category and the category with the second most data points. Similarly, the plays sub-category also makes up for most of the data. We can see from the graph in Tab “Pivot Table 3” that there is an imbalance in the data between the plays sub-category and the rest of the categories. This, as well, appears to be a red flag which can alter the decision that is made behind successful campaigns.

Lastly, there is no data past year 2020. Given the assumption that we are analyzing data in the present year (2023), it is difficult to present a solution if we do not have access to more recent information. If the data included years 2021 and 2022, at least, then we would have more clarity on trends of recent times. This is important as many things such as technology and trends likely change over a duration of 2 or more years.

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

For Pivot Tables 1 and 2, I would recommend a pie chart for each category and sub-category over a stacked columned graph. Since there are multiple categories, as well as outcomes, a pie chart would provide a better visual representation of the data. It would also make it easier for the viewer to determine which campaign categories were successful and failed against the entire pie chart.

Along with that, a clustered column graph with x-axes as years and y-axes as outcomes would be beneficial for data analysis. The analyzer would be able to compare how the 4 outcomes scored between each year (2010-2020.

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

Based on the statistical analysis, the mean is a better representation to summarize the data. This is because, for both outcomes, the mean is closer to the standard deviation than the median, and by a large amount.

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

As per the variance values for each outcome, it can be concluded that there is more variability with successful campaigns than unsuccessful campaigns. This, in fact, makes sense because the higher the variance, the higher the variability in the data.