**Module 1: Excel Challenge**

Clare Sakauye

UC Davis Data Analytics Bootcamp

March 23rd, 2023

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

Given the provided data, three conclusions we can draw about crowdfunding campaigns are one, theatre is the most popular category for projects, two, the most successful sub-category can differ based on success metrics and three, crowdfunding platforms are a relatively lucrative way of generating funds for new projects.

Starting with the first pivot table and chart displaying the parent categories and number of corresponding outcomes it is clear that theatre has the highest number of total projects based on count. Theatre as a whole has 344 of the total 1000 campaigns with the second highest parent category, film & video, reaching 178, just over half the number for theatre.

Assuming that popularity is defined by the majority category, we could say that theatre is the most popular crowdfunding campaign category for creators because it has the highest number of projects via the outcome totals.

Knowing that theatre is the most popular category, it is beneficial to explore the sub-categories beneath it and their relative successes. From the second pivot table and chart, it is clear that only one sub-category falls under theatre. Plays is the most successful sub-category based on outcome totals. It has a total of 187 successes, more than any other sub-category. However, by this measure, it is also the least successful sub-category with the highest number of failures at 132.

Success could also be determined based on number of successes per sub-category total, therefore it could be said that audio is the most successful because out of the four projects, all four succeeded. Additionally, it can be said that audio is too small of a sub-category to provide meaningful insight and that web should be considered the most successful due to its relative size of 51 campaigns and a success rate of 36, putting it in the middle between the extremes of plays and audio and more aligned with the rest of the data set. The above examples outline that identifying the most successful sub-category is dependent upon the measure of success defined by either the data analyst or audience.

Lastly, it can be concluded that crowdfunding platforms are a relatively lucrative way of generating funds for new projects. The third pivot table and chart that measure outcomes over time, do not give much insight into whether the success of a project can be related to seasonality. It does however show that the number of successful projects far exceeds the number of canceled or failed projects. Out of the 986 projects, excluding live campaigns, 565 have been successful and 421 have either failed or been canceled. Therefore, the data implies that over fifty percent of projects on crowdfunding platforms succeed in generating funds.

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

Some limitations of this dataset may be data origin, data amount and information pulled. The data used in this challenge is for educational purposes only and isn’t clear as to whether it was created or pulled from an existing source. If analyzing the success of projects on crowdfunding platforms, it would be important to understand from where the data originated.

Since we are analyzing crowdfunding as whole and not looking at just Kickstarter or Indiegogo, it would be of greater benefit to pull data from both websites. It is possible that the data on Kickstarter or conversely Indiegogo could contain unique patterns or relationships. Having data from both websites could provide a clearer picture of the overall success of projects.

Additionally, this dataset is limited in number. We are only analyzing a thousand projects. It might be important to know how many projects exist across all crowdfunding platforms. That way we could better understand how significant this pool of a thousand is. Furthermore, creating pivot tables and charts is not dependent on data size. We’d create a pivot table the same way for one thousand samples as we would two thousand. Meaning, if we doubled the current sample, or even tripled it, it would not cost any more significant time or energy.

Lastly, it could be beneficial to have greater information pulled into the data set. For example, from the Background we know that celebrities and the general population both use crowdfunding platforms. Of the owners, apart from already knowing a celebrity by name, it is difficult to identify who is famous and who is not. There could be a relationship between the success of a project and the relative fame of the owner. Looking into the sub-categories, it is clear that mobile games has a majority of failures that far outweighs its successes. Of the thirteen total outcomes, eight are failures. Similar numbers are also seen in science-fiction. Television, conversely, has eleven successes out of seventeen total outcomes. It would be interesting to see if any television stars own projects in that sub-category versus those who may be lesser known in mobile games or science-fiction. Given time and the right resources, the project owners’ names could be cross-referenced for fame on the internet. Or there could be information given in addition to the blurb, allowing owners to self-identify their fame or lack thereof.

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

A couple other possible tables and/or graphs that we could create are staff pick versus outcome and spotlight versus outcome. Staff pick and spotlight compared to the outcome could provide us information on whether being staff picked or spotlighted has any impact on success. I would create pivot tables and bar charts for both. The data for the two labels would be broken into two clusters of false and true, with separate bars for the different outcomes. We could also compare these two graphs and determine if it is better to be spotlighted, staff picked, both or neither when it comes to project outcomes. Analyzing staff pick and spotlight could also highlight whether projects that receive one or both of these labels see a higher number of backers, meaning a possible better chance to achieve one’s goal.

**Statistical Analysis**

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

For the statistical analysis of the number of backers in relation to the successful outcome of a crowdfunded campaign, the median summarizes the data better than the mean. The mean and median are both used to determine the central tendency of a given data set. In this case, when trying to understand the relationship between the number of backers and the success of a project, it can be easy to be mislead by the mean.

Looking at the statistical summary of successful campaigns, the mean is 851 and the median is 201. The mean is significantly higher than the median because it is taking the average of all the successful projects. There are a number of projects with backers greater than 851 which is skewing the mean to show a higher number than the central tendency. The number of projects with fewer than 851 backers is then being masked by the average. The delta between the minimum and maximum of this data set is so great that the average is being influenced by extreme values. The median, however, tells us that 201 is the middle of the data set, meaning there are an equal number of projects on both sides of 201. This tells us that more of the data set has successful campaigns with closer to 200 backers than any other number. Meaning, that even though there are successful campaigns with backers greater than one thousand, most of the projects in the data set are significantly less than that.

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

Based on my summary statistics table, there is more variability with the successful campaigns. The successful campaigns have a variance of 1,603,374 while the unsuccessful campaigns have a variance of 921,575. This makes sense because the data set for successful campaigns is larger and has a greater delta between the maximum and minimum values, which we can use to indicate a greater spread/distribution of data.

Variance is used to determine how far different data points are from the average. With a larger data set, any given value, especially the minimum and maximum, will be of a greater distance to the center than that of a smaller data set. In this case specifically, the successful campaigns have a higher mean, standard deviation and count, therefore when applying the formula for variance the successful campaigns will be greater.