Shanara Hawkins

Data Science – 2023 Module 1 Report

Professor Booth

**DATA SCIENCE REPORT: A STATISTICAL ANALYSIS OF CROWDFUNDING CAMPAIGNS**

For over twenty years, crowdfunding campaigns have been steadily on the rise as a unique and innovative tool connecting investors with creators in various industries who are looking to raise funds to support a business or special project. To date, billions of dollars have been raised for creators using crowdfunding platforms, which is quite impressive, but at the same time, it also begs a few questions. For instance, what is the secret to increasing the chances of success? What factors result in instances of failed funding for what appears to be a great business project or venture idea? How can we leverage data science to gain insights on the actions steps that lead to successful funding? In statistical analysis, the most common measures of central tendency are mean, median, mode, and range. Depending on the perceptions you want to gain, these measures can be helpful in pinpointing the unique characteristics that offer a more precise and concise understanding of the values in a data set. Therefore, the initial step in providing the most accurate answers to the questions mentioned above begins by exploring the most relevant and appropriate data available on the topic.

To begin the process, a total of 1,000 sample projects were gathered and organized to see what statistical characteristics could be useful to create a quick assessment summary of the data points. After analyzing the mean and median, the results of the data indicate that the median values would be the better of the two to summarize the data. In the datasets from both the successful and failed campaigns, the average (mean) levels of backers were higher than their respective median value. In successful campaigns, the mean value was 851 and the median value was 201, and in failed campaigns, the mean value was 585, and the median value was 114. An analysis of several statistical summary charts generated from the Excel sheets of this sample project revealed a right-skewed distribution pattern, and most of the data points were positioned on the left side of the chart where the numerical values were smaller. In contrast to this, there were significantly less of the larger values positioned on the right side of the chart. Based on this information, the median value best represents the typical value to use to make more accurate judgements and recommendations. This point is visually illustrated by the chart’s data points positioning and its skewed distribution pattern.

Another measure of central tendency that provided more clarity upon analysis is the statistical range. The range represents the minimum and maximum values present in a dataset, and it’s useful when analyzing the variance of values amongst all data points presented. Based on the data from the campaign, there was more variability with both successful and unsuccessful campaigns. Given the minimum and maximum values calculated from the successful campaigns (16 and 7295 respectively) and failed campaigns (0 and 6080 respectively), it does make sense that there would be more variability within the campaigns because the data follows the fundamental logic that if the statistical range of values has a large number of variance between them, then the data points will be more scattered and spread across a wider range of diverse sets of values. Moreover, in the crowdfunding campaign Excel sheets and charts, the values inputted in the Goal, Pledged, and Percent Funded Columns shows a broad scope of variability and outliers within the numerical data of each section.

Considering the information presented, three conclusions that can be drawn about crowdfunding campaigns is that the levels of support for a particular project are best analyzed using the median value as the typical level of support received from backers, and not the mean value since it overestimates the average funding amount. Another conclusion we can draw is that backers who invest the largest amounts of funds generally represent the smaller group of backers in the population set. In other words, backers who usually pledge the highest amounts should be considered the exception and not the rule since there are fewer of them in comparison. On the other hand, most backers in the population set typically fund smaller amounts that statistically fall within or below the median values. Finally based on the statistical summaries, we can conclude that the variability found within a wider scope of values may pose a challenge when creating data visualization charts that are concise, precise, and easy to read and understand. This is because a diverse range of clusters and outliers in a data set may require alternative methods of statistical analysis that are more intricate and complex in nature.

On the negative side, the limitations that exist in this dataset could be the lack of historical data, which can be used to track various trends, identify user pain-points, substantiate investor interests and monitor behavioral trends. This information could be beneficial in management and planning for organizations who are dedicating extensive time and resources for a chance to obtain funding. In addition, cultural, economic, and geographical limitations may create additional limiting factors for both investors and creators. In fact, cultural norms and customs may play a huge role in a creator’s project or business. Therefore, misinterpreting customs and preferences, or the inability to communicate due to language barriers can all likely contribute to unintentional bias when reviewing campaign messages and proposals. Likewise, these limitations can subsequently result in inadequate campaign exposure, and fewer opportunities to interact with investors, adversely impacting connections, partnerships, support, and venture funding prospects.

In reference to the campaign Excel Data Sheets created for this project, some examples of charts that were used and are embedded in the Sheets tab include histograms, box and whiskers chart, and a bar chart. Alternatively, the Scatter Plot, Pareto Chart, a generic Campaign Success Rate Table, and a Filled Map could also be appropriate tables and charts used for this task. Scatter Plots would show the relationship between two variables charted as two separate data points. The Pareto Chart would be useful to pinpoint the usual causes or categories that contribute to campaign failure. A generic Campaign Success Rate Table could be used to calculate the most successful campaign categories. Then, sort the data into two categories—one that has the highest success rate and the other that has the lowest success rate. This information can then be used to improve, update, cancel, and/or replace the lowest performing category with one that may yield better performance outcomes. Finally, a Filled Map could be used to pinpoint the geographical regions with the highest and lowest success rates by category, interest, and several other types of classification.