Assignment1 Report

**Conclusion 1:**

Table 1 and Figure 1 present the number of canceled, failed, live, and successful crowdfunding campaigns for different parent categories, as well as the total number of campaigns. There are a total of 1000 campaigns, with theater having the highest count of 344, accounting for 34.40% of the total. On the other hand, journalism has the lowest count of 4, accounting for 0.40% of the total.

Among the 57 canceled campaigns, theater has the highest number of cancellations with 23, accounting for 40.35% of the total, while journalism has 0 cancellations, accounting for 0%. For the 364 failed campaigns, theater still has the highest number of failures with 132, accounting for 36.26% of the total, while journalism has 0 failures, accounting for 0%. In terms of 565 successful campaigns, theater remains at the top with 187, accounting for 33.10% of the total, while journalism has only 4 successful campaigns, accounting for 0.71%. Regarding the 14 live campaigns, film & video has the highest count with 5, accounting for 35.71% of the total, while food, journalism, and music have no live campaigns, accounting for 0%.

It is observed that some categories have a higher total number of campaigns, indicating both a larger number of campaigns and a higher number of failures or cancellations. Therefore, it is difficult to evaluate the performance of different categories solely based on these numbers. To gain a better understanding, I calculated the proportions. Table 2 presents the proportions of successful, failed, canceled, and live campaigns for each category.

***Table 1****. The outcome counts for different parent categories.*

A screenshot of a data

Description automatically generated

***Figure 1****. A stacked-column chart of the outcome counts for different parent categories.*

A graph of different colored bars

Description automatically generated

***Table 2****. The percentage of outcomes for different parent categories.*

A table with numbers and a few percentages

Description automatically generated

The overall success rate for all campaigns is the highest at 56.50%, followed by a failure rate of 36.40%, a cancellation rate of 5.70%, and a live rate of 1.40%. Among all the categories, journalism has the highest success rate of 100%, the lowest failure rate of 0%, and the lowest cancellation rate of 0%. On the other hand, games have the lowest success rate of 43.75% and the highest failure rate of 47.92%. Photography has the highest cancellation rate of 9.52%. We can say that journalism performs well while games perform poorly. However, it should be noted that the category of journalism has a relatively small total number of campaigns, so we cannot confidently conclude that this category will always be the best choice in the future. Also, we cannot definitively say that the future trend for games is unfavorable, as there are still 6.25% of campaigns currently live.

**Conclusion 2:**

From Table 3 and Figure 2, it can be observed that among the sub-categories under each parent category, without adjusting for the parent category, the plays sub-category has the highest number of campaigns with 344, accounting for 34.40% of the total, while world music has the lowest number with 3, accounting for 0.30%. According to the proportion information in Table 4, world music and audio have the highest success rates at 100%, with the lowest failure rates at 0%. They belong to the music and journalism groups, respectively. On the other hand, mobile games have the lowest success rate at 30.77% and belongs to the games group, while science fiction has the highest failure rate at 64.29% and belongs to the film & video group. Among them, television has the highest cancellation rate at 17.65% and belongs to the film & video group, while audio, world music, translations, wearables, metal, electric music, radio & podcasts, mobile games, science fiction all have a cancellation rate of 0%.

Combining with Conclusion 1, I filtered different performing parent categories to get different sub-categories and examine the results. Journalism only has one sub-category, audio, with a total of 4 campaigns and a success rate of 100%. Therefore, we can infer that the audio sub-category may show good performance and success rate. However, due to the small number of campaigns and the limited diversity in sub-category types, we cannot conclude that audio is a high-performing category. The small sample size may introduce sampling bias, lack of representativeness and statistical power, resulting in high uncertainty and instability of the results.

The games category is divided into mobile games and video games. The success rates of these two sub-categories are relatively low compared to other sub-categories, at 30.77% and 48.57% respectively, with high failure rates at 61.54% and 42.86% respectively, and low cancellation rates at 0% and 2.86% respectively. Overall, the games category performs poorly, and its sub-categories also show poor performance. However, the proportions of live campaigns for these two sub-categories are 7.69% and 5.71% respectively, indicating uncertain results. Furthermore, the games category has a total of 48 campaigns, with relatively small sample sizes for each sub-category, 13 and 35 respectively, especially for mobile games, resulting in a lack of statistical power and representativeness.

Furthermore, it is noted that within the music group, world music has a relatively high success rate. However, this does not necessarily indicate outstanding performance of the music or world music category, as the total number of campaigns for world music is only 3, all of which were successful. The small sample size and the performance of other sub-categories in the music group should also be considered.

Lastly, within the film & video group, although science fiction has the highest failure rate, it does not necessarily imply poor performance of the film & video or science fiction category, as the number of campaigns for science fiction is relatively small at 14. Additionally, the success rates and failure rates of other sub-categories under the film & video group are not too low and high compared to other sub-categories.

***Table 3****. The outcome counts for different sub-categories.*

A screenshot of a computer

Description automatically generated

***Figure 2****. A stacked-column chart of the outcome counts for different sub-categories.*

A graph with different colored bars

Description automatically generated

***Table 4****. The percentage of outcomes for different sub-categories.*

A screenshot of a computer

Description automatically generated

**Conclusion 3:**

From Figure 3, it can be observed that when considering the cumulative data for all years, the number of successful campaigns in each month is higher than the number of failed campaigns and the number of canceled campaigns. However, there is a notable decrease in the number of successful campaigns in August compared to other months, with a relatively higher failure rate and cancellation rate. Therefore, it can be inferred that there may be a need to reduce the number of crowdfunding campaigns in August.

***Figure 3****. The distribution of outcome counts for 12-month projects.*

A graph of different colored lines

Description automatically generated

**Limitation:**

1. Small sample sizes under certain conditions: It can lead to inaccuracies in statistical results, lack of representativeness, sampling biases, and an inability to capture true population characteristics or differences. Furthermore, the obtained data results may not be generalizable to the overall population and broader contexts. Additionally, due to the limited sample size, complex relationships or trends may not be adequately captured, resulting in erroneous interpretations or overinterpretations of the data. Small sample sizes can also lead to unstable results, with significant fluctuations observed across different samples, thereby rendering the interpretation and prediction of results unreliable.

2. The unequal sample sizes for different levels of the same variable in this dataset can lead to misleading comparison results. The large differences in sample sizes across different conditions may result in exaggerated effects in conditions with larger samples, while the true effects may not be properly represented in conditions with smaller samples. Additionally, the imbalance in sample sizes can introduce bias in inferring the relationships between different conditions. Conditions with larger samples may be more representative, while conditions with smaller samples may suffer from sampling bias, thus affecting accurate inferences about the overall population.

3. The data distribution of continuous variables in the dataset exhibits significant outliers and skewness, which may lead to inaccurate statistical inferences. Traditional statistical methods based on the assumption of a Gaussian distribution are not applicable. The presence of outliers and skewness can impact the interpretation of statistical indicators. The mean, as a measure of central tendency, can be heavily influenced by extreme values, leading to an inaccurate representation of the central position of the data set. Similarly, measures of dispersion, such as standard deviation, can lose accuracy due to the influence of extreme values. Skewed distributions and outliers also affect data analysis and visualization, making it challenging to observe and understand the underlying patterns and features of the data.

4. The dataset lacks clear explanations for its variables, and the requirements for different levels of categorization are not well-defined. For example, it is unclear how the "goal" variable is determined and what unit it represents. The meaning of "staff\_pick" and "spotlight" being either false or true is also not clearly explained. Moreover, the criteria used to classify "failed" and "successful" outcomes are not specified. This can lead to misinterpretation of the data, an inability to understand exactly what each variable represents and its relevance to the research question, which can lead to erroneous analysis and conclusions. If it is not clear how certain variables were obtained or their units are not known, the reliability of the data may be affected. Failure to understand the classification criteria for variables and the meaning of different levels can lead to misleading results for comparisons between different groups or conditions and difficulties in interpreting the results.

**Other possible tables/graphs:**

1. Descriptive statistics table and demographic table: Create a descriptive statistics table that includes measures such as mean, median, standard deviation, minimum, and maximum for each variable. This helps us quickly understand the central tendency, dispersion, and range of the data. Additionally, create a demographic table that includes background information of the sample, such as age groups and gender, to better understand the characteristics and composition of the sample.

2. Histograms for continuous variables: Create histograms to visualize the distribution of continuous variables, including observing peaks, dispersion, and outliers, to enhance the reliability of the statistical analysis.

3. Boxplots: Create boxplots for different levels or variables to display the minimum, first quartile, median, third quartile, and maximum values of the variables. This visualization helps us understand the distribution, outliers, and anomalies of the variables.

4. Line charts and bar charts: Observe and compare the frequencies or percentages of different levels of variables and their trends. These charts can help us identify relationships, trends, and patterns between variables, as well as facilitate effective comparison and visualization.

5. Scatter plots and correlation plots: If we want to explore the relationships between different variables, we can create scatter plots or conduct correlation analysis, calculate correlation coefficients, and create correlation plots. This helps us observe and understand the relationships, trends, and correlations between variables, supporting further analysis and interpretation.

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

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

1. Both for the successful and failed conditions, the median is more appropriate. The data distribution is skewed, uneven, and contains extreme values. The median is less sensitive to outliers and better represents the central tendency of the dataset. Based on the information of standard deviation and variance for both conditions, it can be observed that the data has a high degree of dispersion. The mean is influenced by outliers and may not accurately represent the overall data. Despite the relatively large sample size, the skewed distribution makes the mean less suitable, and the median should be used instead.

2. Comparing variances, standard deviations, and ranges reveals higher variability in the successful group. However, we can't make the general conclusion that "there is more variability with successful campaigns." Firstly, there is a significant difference in sample sizes between the successful group (565) and the unsuccessful group (364), with the successful group having a larger sample size. Having more data points may result in larger variances, but it does not imply a wider or more diverse range of backer variations. Secondly, both the successful and unsuccessful groups have non-Gaussian distributions with significant skewness and the presence of numerous outliers. This may be due to various reasons, such as sampling issues, which can lead to inaccurate comparison results for variances, standard deviations, and ranges between the two groups. Thirdly, besides standard deviation, variance, and range, additional statistical measures, such as quartiles, can be used for a more comprehensive evaluation. By analyzing these indicators, the variation between the successful and unsuccessful groups can be assessed more comprehensively. Lastly, successful and unsuccessful campaigns may have different characteristics and target audiences, leading to potentially different patterns of backer variations. Therefore, when comparing the two groups, it is not appropriate to simply conclude that the successful group exhibits a higher degree of backer variability. It is important to consider or control for other factors (like covariables), such as market trends, product features, and marketing strategies, to explain backer variations.