Eric Llorente Alarcon

Columbia University

Data Analytics Boot Camp

Weekly Challenge 1

August 20, 2023

**Written Report**

After analyzing the stacked-column and outcomes by month pivot tables and charts, I was able to deduct the following three conclusions for the data:

* Plays can be considered an outlier in the data set.
* July is the month with the most successes for projects.
* Photography and Web projects had the highest percentage of success with around 70% and 67% respectively. This is considering that Audio, World Music, and Metal had a 100% success rate with very few backers.

A limitation that the data faces is that visualizations are not able to depict a better story of the data distribution. One suggestion that I would have, is to have a column that calculates the percentage of success or failure in Parent categories or sub-categories in relationship with backers to predict the probability of success. I also filtered out some of the outliers to get a better visualization of the data. By filtering out Theater and Tech, I was able get a clearer picture of the success in other categories like Photography, Audio, World Music, and Metal.

**Statistical Analysis**

In this case, I believe that the mean summarizes the data better because it can give you a better idea of the distribution. One thing that I did additionally to calculating the mean, median, min, max, variance and standard deviation, was to calculate the data range and count the number of projects that had a backer’s count above the mean. From this I was able to see that the data had a wide range and did not have a normal distribution because of the outliers. There were 162 projects with backer’s counts above the mean, which would account for only 29% of the data. If the data was normally distributed, the percentage of projects with a backer’s counts over the mean should reflect around 50% of the data under the standard deviation curve. Another clue that gave away the discrepancy in the normal distribution is that the minimums for the data sets were all within one standard deviation from the mean. Since you cannot have negative backers, I expected the minimum to be somewhere around three standard deviations below the mean.