Sunwoo Kim

UCI DA BootCamp

Homework 1

**Kickstarter Data Report**

Given this data set, one conclusion we can draw is that theater and music are the two most popular Kickstarter campaign categories. Additionally, both of these categories have rates of success comfortably over 50%. As such, our data suggests that individuals or organizations interested in funding a theater or music project have a good chance at success. A second conclusion can be drawn from the line chart that displays the percent of success, failure, and cancelation by month. According to this chart, campaigns that were started in May have the highest proportion of success and campaigns started in April have the lowest percent of failure. Additional testing and data could suggest that campaigns started in April and May have strong chances of success. A final conclusion we can draw from the data is from the table breakdown by sub-category. The most populous sub-category is plays, but the percentage of success for play is not exceptionally high. However, campaigns from our dataset in the documentary and rock music sub-categories have all succeeded. With a large sample size and more rigorous testing, these two sub-categories may prove to have high probabilities of success.

While we were able to draw some interesting conclusions from this data set, it is not without limitations. One limitation is the number of observations. As listed in the homework prompt, there have been over 300,000 Kickstarter campaigns. The 4,000 observations provided to us—which barely makes up 1% of the population—may not be representative of the entire population. To further support the possibility that the sample is not representative, our sample has approximately 50% successful campaigns while it is stated in the homework prompt that only 1/3 of campaigns found success. Another limitation is the range of dates. The campaigns in this data set range from 2009 to 2017. Since we have no data on campaigns from 2018 and later, our analyses may not apply to more recent campaigns—especially if there were any events that changed society’s willingness to donate or updated Kickstarter’s policies.

Another possible graphic would be a table and line chart exploring the relationship between categories and percent funded. Analyzing the data in this way could demonstrate which categories have high support—which would be measured by how much a campaign exceeded (or failed to reach) their goal. Additionally, it would be interesting to stratify the successful campaigns by goal amount and list how long these campaigns took from beginning to end. Plotting a scatterplot with goal amount on the x-axis and campaign duration in days on the y-axis could give an idea of how long campaigns will take based on their goal amount.

**Statistical Analysis Responses**

* For both the successful and failed outcomes, median summarizes the data more meaningfully than the mean does. Some of the successful campaigns have extremely high backer counts—for example, the largest backer count was 26457. These large values are few in number but pull the mean of the data to a larger number. In this scenario, the robust median of 62 is a better measure of central tendency than the mean of 194.43. Similarly, for the failed campaigns, there are a few observations with high backer counts. However, the data is most dense near zero; thus, the median of 4 is a more appropriate measure of central tendency.
* The subset of successful campaigns (713167.38) has a higher sample variance than the subset of failed campaigns (3775.69). This makes sense because in general, the unsuccessful campaigns likely failed due to insufficient backers. It follows that the number of backers for most failed projects will be low and near zero. On the contrary, the numbers of backers for successful campaigns vary greatly because each campaign has a unique goal amount. In addition, most campaigns exceed their goals by varying amounts. These factors contribute to the high variance of the successful campaigns.