**Crowdfunding Excel Challenge Assignment Report – Mark Speers**

**Primary Tasks Report**

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

* Though *Theater* had the most projects (344), if one was looking to invest, they may want to examine the chances of success instead of the number of projects. ***Technology* projects had the highest rate of success (67%), and thus, investors may find more success if they invest in Tech projects.**
* **The chances of success of crowdfunded campaigns vary widely across goal amounts.** The average success rate across all goal ranges was 53%, but variation between certain goal ranges was dramatic. Projects with a goals within 15000-24999 had a 100% success rate while projects with goals of 10000-14999 had a 44% chance of success. Projects with a goal greater than or equal to 50000 had the lowest success rate (37%). The variation in sample size between groups may have contributed to these dramatic disparities in success rates. All goal ranges with a success rate of 100% had at most 10 observations (projects), while those goal ranges with more modest success rates often had larger samples. However, there are some exceptions. For instance, projects with goal within 1000-4999 had a success rate of 83% across 231 projects. Additionally, there are several goal ranges where the sample size was small and the success rate was modest (see ranges: 10000-14999, 25000-29999, 35000-39999, 40000-44999, 45000-49999). Nonetheless, it is important to note that the statistics associated with ranges that have fewer than 10 cases are more likely to be biased than those with more than 100 cases, so interpretation of said ranges should take this caveat into account.
  + After further analysis it was revealed that goal ranges with less than 50 observations have a much higher average success rate (80%) than those with more than 50 observations (58%). This disparity may be due to bias sample problems discussed in the parent bullet.
* According to the Success by Launch Date chart, it seems that the number of successful projects is somewhat consistent throughout the year. However, if you look at individual years, the patterns vary pretty widely. **Nonetheless, there does seem to be a frequent spike in successful projects sometime in the summer, particularly in June or July.** Sometimes this may be due to an overall increase in crowdfunding projects, but in some years the increase in successful campaigns is **not** accompanied by an increase in failed, live, or canceled campaigns. Thus, one might prefer to begin your campaign in June or July to increase your odds of success.

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

* It is only a sample of crowdfunded projects, and we don’t know the sampling method.
* There isn’t a good description of the projects. It’s difficult to tell what the project was specifically about. For example, one project’s blurb says “Monitored multi-state encryption” but this project is classified as a “theater” project. More detailed information on the product/service being crowdfunded may be useful in predicting success.
* It didn’t come with a code book so it’s a little difficult to learn about the different measures in the dataset.

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

* Visualizing the difference in percentage successful is accomplished by the pivot charts showing the projects’ success status based on category, but this comparison could be quantified if we calculated the percentage of projects in each category for each outcome. This analysis would add a small amount of relevant information because the variation in percentage successful and percentage failure across categories has a seemingly small amount of variation, but we cannot easily quantify this variation based on the visual analysis of the stacked bar charts.
* If we visualized the interval between the start and end dates, we might be able to visualize how the length of the campaign relates to the chances of success. This could help someone pick the appropriate length for their own project.

**Bonus Statistical Analysis Report**

***Does the mean or median summarize the data more meaningfully?***

* In this case, the median summarizes the distribution of the data more accurately. Both successful and failed campaigns have a severe positive skew (right skew) and so the outliers at the far end of the skew inflate the mean artificially. To illustrate, for successful campaigns, it is easy to see that the vast majority of projects have a backer count between 16 and 416, but the mean/average backer count for successful campaigns is 851. In contrast, the median number of backers for successful campaigns is 201.

***Do the data indicate more variability in backer count of successful or failed campaigns? Does this make sense? Why or why not?***

* The data indicate more variability in the *backer count* of successful campaigns. I think, but am not sure, that this makes sense. If I’m thinking about this correctly, the variance measures how far away data points are from the mean, and this would result in the successful campaigns having a higher variance because of two features of the data set. First, the mean backer count of successful campaigns is higher than the mean backer count of failed campaigns, likely due to successful campaigns having higher, and more, outliers than failed campaigns. This elevated mean, combined with the fact that the backer count of both failed and successful campaigns are very positively skewed, results in more data points being further from the mean for successful campaigns. As the mean is pulled right on the x axis (backer count), the data points clustered at the left of the graph – because of the right/positive skew – become further from the mean, thus inflating the variance.