Module 1 Challenge

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

In the given data, we will try to discover any hidden trends within the context of crowdfunding. Our analysis of the dataset aimed to uncover key patterns and insights that can inform strategic decisions to find success in crowdfunding. Here are the conclusions:

(A) The first stack bar graph shows the relationship between count of the campaigns that were successful, failed, cancelled, or are currently live per parent category. From this graph, we can see “film & video”, “music” and theatre had the most total backers (“grand total”), and all had a high successful campaign count (>100 backers). Therefore, we can conclude that when starting a new project, a category in “Theatre” has a high chance of receiving funds and being successful.

(B), The second stack bar graph shows the relationship between the count of the sub-categories that were successful, failed, cancelled, or are currently live. From this table, we can see that “plays” had high outcome count for “successful” (187). This graph is useful as it can help us narrow down what category did well under the parent category. We can conclude that when start a new project, we should start out in “plays”.

(C), In the line graph depicts the relationship between the successful, failed and cancelled over a 12-month period. Overall, the successful outcomes were greater compared to the failed and cancelled outcomes over a 12-month period, with failed coming in second and cancelled in last. If we look at the trend between the successful, failed and cancelled. Successful and failed had minor fluctuation between the months of January to May. However, there was a steeply increase from May all the way to July, which was the peak outcome for the whole year, then rapidly decrease from July to August. From August to Dec, it slowly increases then plateau off. Failed continued to have fluctuation until August where it suddenly decreased until September, then it gradually increased towards the end of the year. Cancelled on the other hand, remain constant all year with small fluctuation.

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

1. Not including the average donation to the graph – this information is valuable because it gives us an idea on the amount of donation we can received over a period. Gives a better insight into economic status of that year and month.

2. The data for currency is provided in many different currencies (AUD, USD etc). This data can be confusing and inefficient when conveying this information to someone else. Can possible convert all the value to 1 currency to better illustrate the information.

3. The number of countries selected in the data is not a good representation of the whole population. From our data, we have 7 countries compared to 195 countries in the whole world.

4. Selection criteria for the countries is unknown. As stated above, we don’t know why these 7 countries were selected, this could mislead our conclusion.

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

1. Horizontal bar graph – they list the ranking from highest to lowest. Useful when we want to compare the outcome performances of all the subcategories. We can also add a filter for the outcomes (i.e. successful, failed, cancelled and live) which we can use to compare the performance.

2. Scatter plot – shows the correlation between 2 variables. They can show us if the data has any trends (positive or negative), no correlation, clustering trends or spot out outliers. As our data contains large values, we can easily identify them. Also, in the line graph we made, we can plot this as scatter plot and helps us identify the correlation.

Bonus Statistical Analysis Questions:

1. Use your data to determine whether the mean or the median better summarises the data.

The mean gives us a good representation of the data because it uses every value in the dataset. However, it is highly sensitive to extreme values and outliers therefore affects our distribution negatively. Whereas the median is less sensitive to extreme values and/or has outliers as it considers the position of the value therefore can be better than the mean in showing distribution. In our dataset, both contains high values therefore I would use the median to summarise the data. However, I believe using both the mean and median values can give us a more comprehensive summary of the data.

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

To determine if there more variability with successful or unsuccessful campaigns we use these three metrics: variance, standard deviation and z-score. However, I will only explain the variance and standard deviation (as z score is not calculated).

Variance describes how far values in the data are from the mean and standard deviation describes how spread out the data is from the mean. In our dataset, the successful campaigns had a higher standard deviation compared to the unsuccessful campaign. A high standard deviation mean that the data is more spread from the mean therefore the results is more less predictable. Whereas a lower standard deviation mean that the data is closer to the mean therefore result is more predictable/reliable. Similarly with variance, a smaller value equals to less variance and a high value equal to more variance (i.e distance from the mean). Successful campaign had a high degree of variance compared to unsuccessful campaigns therefore we can say that successful is riskier option compared to unsuccessful campaigns.

In conclusion, unsuccessful campaigns are more variability compared to successful campaigns. This is because the standard deviation and variance value is lower compared to successful campaigns.