Week 1 Homework: Crowdfunding Analysis

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

Three conclusions that can be reached based on the data provided are that:

* + The overall chance that there will be a positive result when crowd-funding is 0.565, with failure at 0.364. This suggests that should a crowd-funder be completed (barring cancels and lives), there is a 60.8% chance of a favourable outcome.
  + When referring to the categories of crowd-funder, those that pose the highest risk of failure are food (52.4%) and games (47.7%) when seen to completion. The category that has the highest success rate is photography at 70%, however this excludes the outlier of journalism.
  + When referring to the success rate per country, it can be seen that there is a reasonably low spread rate of success, ranging from 0.5 (CA) to 0.58 (GB).

\*These figures are inclusive of lives and cancels.

1. What are some limitations of this dataset?

Three limitations of this dataset are as follows:

* + 1. At a country level, there are not enough pieces of data to accurately ascertain the success rate of a crowd-funder. With the exception of the US, where the data seems to be localised, all countries are below 50 items of data.
    2. The dataset does not indicate the process each crowd-funder underwent, and why each was successful/unsuccessful. Potentially adding in information on advertisement budget and reachable audience numbers could increase ability to analyse crowd-funder success.
    3. The staff\_pick and spotlight columns are not based on formulaic ‘If’ statements. Without these, it is unclear what purpose they hold towards the success of each crowd-funder without combing through the data set for hidden trends.

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

Additional tables and graphs that could be added are:

**Histograms:** Could be used to categorise the ‘percent funded’ columns as a means of considering measures of central tendency. This could potentially be done with amount funded, but that would interfere with the limitation of each amount being in a different currency

**Bivariate scatterplots:** Comparing ‘backers\_count’ with ‘pledged’ could give an indication of backer patterns. By using the linear equation from the line of best fit by genre/category, you could get a rough indication of how many backers would be needed for future crowd-funders

**Boxplots:** Boxplots could be used for ‘pledged’, ‘percent funded’ and ‘backer-count’. This would be the least useful graph as the limitations of the ‘percent funded’ and ‘pledged’ are outlined in ‘Question 2’, while the bivariate scatterplots would be a more effective version to analyse the ‘backer-count’ section.

Bonus Section:

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

For this set of data, the median value is likely a better way of summarising the data. Looking at the ‘successful’ outcome, of over 500 pieces of data, there are 163 pieces of data over the mean. This implies that there is a high amount of outliers skewing the data-set upwards, especially when comparing 2-digit numbers with 4-digit numbers. Using the median reduces the amount of impact these high figures have on the rest of the data series, and is therefore the more appropriate measure of summarising the data set.

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

From the data, it can be seen that ‘successful’ campaigns had a higher variance than ‘failed’ campaigns. This conclusion is logical when considering the potential for each state of outcome. A ‘successful’ campaign can be of any size, ranging from low scale to significantly large. This means that the backer count can be anywhere above 1 with no defining limit.

The minimum value of ‘failed’ provides a slightly higher chance of variance with a potential value of 0, but where it falls short in variance comparatively is in it’s upper limit. A ‘failed’ outcome outlines that the goal, and thus the amount of backers, was lower than intended. Overall, it is likely, but not guaranteed, that the ‘failed’ outcome will have a lower maximum value than a successful outcome, resulting in a lower variability.