Robert Bentz

UOM – Data Science

Crowdfunding Goal Analysis

Starting with the assumption that I correctly interpreted the, “Ask”, and performed the directions properly.

Given the provided data in the, “CrowdfundingBook.xlsx”, spreadsheet, there are several questions to address:

* Conclusions that may be drawn from the category sheet include:
  + The performing arts appear to be much more successful at crowdfunding campaigns (film & video, music and theater).
  + There also appear to be a lot more crowdfunding campaign attempts for the performing arts.
  + Appears the failure rate is pretty consistent between groups, 40-50% (not calculated, just visual observation).
  + Although journalism is by far the least popular category, it is 100% successful.
* Conclusions for the countries sheet include:
  + All countries had a preference to attempt performing arts crowdfunding campaigns.
  + Attempting publishing in Great Britain appears to be a bad gamble, with no success listed.
  + Italy has 100% success at technology crowdsourcing campaigns. Great place to build tech ideas.
  + Australia is also good for tech. There were some failed campaigns, but quite a few more were attempted than in Italy.
* Conclusions for the sub-category sheet:
  + Plays are by far the most popular campaign, across all countries.
  + Of the plays attempted in each country, the failure rate is 40-50% for all the countries.
* Conclusions for the date sheets:
  + After a rocky start, the success rate for the campaigns is slowly rising.
  + The failure rate seemed pretty consistent until 2020, when it dropped like a rock.
  + 2020 appears to be incomplete, based upon no values for several categories, and previous year campaigns hovering at 100 per year. 2020 has two campaigns for the entire year.
  + Seasonally it would seem June and July are the best times of year to attempt a crowdfunding campaign. Success rate is higher and the failure rate is lower.

Some limitations of the dataset appear to be:

* The blurb column does not appear to be correctly synced to the category and sub-category columns.
* The 2020 data appears incomplete, and possibly 2019.
  + 2020 has no cancelled, live or successful data.
  + 2020 has only two total campaigns, compared to ~100 for other years.
* At this time, the columns are not defined. Some column headings are not intuitive as to what relevance they hold. Staff pick and spotlight columns are true/false, but are not defined anywhere.
* The dataset is comparing apples and oranges. The native currency to the locations is being used. There should be a conversion to a common currency unit to give a baseline for the money comparisons.
* Calculating the averages may artificially skew the view of donations. A few very large donors would artificially inflate the averages of all smaller donations.

Additional tables or graphs that could provide more useful information from this dataset would be:

* Creating a column for US dollars, and use currency calculations to list all monetary values in the US dollar. This may affect how any graphs are used due to how the dollar is weighted against other currency (conversion rate).
* A pivot table and chart showing the US dollar as a value comparison against the category field. What category produces the most money (divide total dollars by number of projects).
* How does the length of campaign time influence money and success rate?

For the Statistical Analysis for the generated outcome table, I assumed (but did not see in the directions) from the listing of nine items that the reference was for the category. Hence the category column was created although not requested.

The median best describes the outcome table columns. There are some large numbers that really skew the results, biasing the column’s values to the high side. There are only two ‘low’ values, both for the journalism category that can contribute to lowering the median and averages. I don’t believe they are outliers though.

Variability is highest with the successful outcome group. The sample size is more than double. Standard deviation is more than doubled. Unusual because the sample size is so much bigger, the variability should be decreasing, not increasing. Even random error rates should decrease with the much larger population size. Standard deviation should be decreasing. The theater category is most likely shifting, or skewing the results high (there are no values even close to it).

I would have to say the theater category is a significant outlier, influenced by people’s enjoyment of the medium. This may have been influenced by the types of people that contribute to crowdfunding as opposed to a random population sample. More advertising and awareness of theater production crowdfunding compared to a specialty funding opportunity with journalism for example.