**Background**

Crowdfunding platforms like Kickstarter and Indiegogo have been growing in success and popularity since the late 2000s. From independent content creators to famous celebrities, more and more people are using crowdfunding to launch new products and generate buzz, but not every project has found success.

To receive funding, the project must meet or exceed an initial goal, so many organizations dedicate considerable resources looking through old projects in an attempt **to discover “the trick” to finding success.** For this week's Challenge, you will organize and analyze a database of 1,000 sample projects to uncover any hidden trends.

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

The initial analysis was to look at the Parent and Sub-Category of the projects previously presented. Based on the chart that were created, below bullet points are conclusions we can draw:

1. Not a lot of attempts to try “Journalism”, only 4 projects were presented and all of the projects were successful.
2. Out of all the projects, “Theater” has the highest attempts. Handful projects are still “live”, and the number of “successful” is fairly higher than the “failed” numbers. Even though the attempts are higher, you still have higher chance to be successful with “Theater” category project.
3. Looking at “Music” and “Film & Video” parent category, the total number of projects are almost the same. Their “failed”, “canceled” and “successful” are almost the same, but “Film & Video” draws a difference with 3 counts of “live” outcome.
4. The next lowest attempt is “Photography”. There are not many attempts with this category, out of 42 total projects, the “successful” outcome is more than half of the total and “failed” outcome is half of the “successful” outcome. There is a pretty high chance to be successful.

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

The data set does not provide enough information as to why a project would be “live”, “canceled”. It would have been nice to also see the population based on ppl’s interests in different categories. Also, I noticed some categories have limited data provided for instance “Food” is only “Food Truck”, but does not provide much info on what type of food they are. Similarly, “Technology” also has limitation, it says “Web” and “Wearable”, but does not provide much info on what they type of wearables are, also what category of web is. Same with “Theater”, it is obvious that theater will be around “Plays”, but does not really break into parts to indicate what type of plays (drama, comedy etc.) were successful, failed, canceled or even currently live.

* **What are some other possible tables and/or graphs that we could create, and what additional value would they provide?**
* I noticed anything below 100% funded were considered failed. I would like to know the percent against the outcomes to understand at what point they were canceled, failed and successful to understand the impact of the donations.
* I believe it would be also valuable to look at the difference between “Date Created” and “Date Ended” and understand how the duration impacts the outcome of success, fail and canceled.
* Another one would be also looking at the bakers count against the outcome to analyze the count of bakers a project might need to be successful based on their category.

**STATISTICAL ANALYSIS**

|  |  |  |
| --- | --- | --- |
| **Statistical Measures** | **Successful** | **Failed** |
| The mean number of backers | 851.15 | 585.62 |
| The median number of backers | 201 | 114.5 |
| The minimum number of backers | 16 | 0 |
| The maximum number of backers | 7295 | 6080 |
| The variance of the number of backers | 1603373.73 | 921574.68 |
| The standard deviation of the number of backers | 1266.24 | 959.99 |

* **A brief and compelling justification of whether the mean or median better summarizes the data**

Mean is the average value for a given set of numbers while median is the value that falls in the middle of the dataset. After running the calculation, median summarizes this data set better. When we look at the max and min numbers, we can see the significant difference between two values. This indicates the outliers which are not the true representation of this data set. Hence median will give more representative value than mean in this situation.