Data Analytics on Kickstart Campaign

**Plots**

Figure 1 Campaign Outcome per Sub-Category

Figure 2 Campaign Outcome per Parent-Category

Figure 3 Campaign Outcome per Month

**Question**

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

* Successful rate is higher at lower target goal. The Higher the target goal the successful rate is dropping while the opposite is observed on the failed rate. The higher the goal the failed rate is higher
* Sub-category “plays” is the most project kickstart among other sub-category
* The successful rate every month is higher than failed or cancelled. The cancelled project outcome rate is similar for every month.
* In average the total campaign per month is 342.

1. What are some limitations of this dataset?

* Data is limited up to 2017
* The duration of the campaign is various which make it not apple-to-apple to compare the project campaign
* There is “Spotlight” column seems to categorize if campaign is successful = TRUE otherwise it will mark as FALSE. Which kind of double dip with the “state” column
* The currency of the data is various. On the instruction it is not clear that we should convert them into the same unit before proceeding.

1. What are some other possible tables and/or graphs that we could create?

* Histogram of number successful/failed campaign vs no. of backers
* Whisker and box plot to see the spread of the no. of backers data on successful/failed campaign

**Bonus**

Figure 4 Percentage Rate Outcome vs Target Goal

**Bonus Statistical**

Statistical Parameter of Number of Backers vs Successful/Failed Kickstart

|  |  |  |
| --- | --- | --- |
| **Statistical Parameter** | **Successful** | **Failed** |
| Mean | 194 | 18 |
| Median | 62 | 4 |
| Mode | 27 | 0 |
| Max | 26457 | 1293 |
| Min | 1 | 0 |
| Variance | 713167 | 3776 |
| Std. Dev | 844 | 61 |
| Q1 | 29 | 1 |
| Q3 | 141 | 12 |
| IQT | 112 | 11 |
| Lower Bound | -139 | -16 |
| Upper Bound | 309 | 29 |

\* Use your data to determine whether the mean or the median summarizes the data more meaningfully.

I would think that median is more useful on this current dataset. As the data spread is very wide, one can see from the median where the dataset gravitate toward. Further investigation with histogram can be seen that on the higher range, the frequency is lesser and lesser. There is significant difference between the first two ranges to the next, This show that most of the data is laying in the first two range. The remaining ranges show the number of backers is not significant.

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

Figure 5 Histogram of Successful Backers

The variance/data spread is larger on the successful campaign which is make sense. Some Kickstarter projects are more interesting than others. The most attractive project will indeed attract more backers. If people are interested, it does not really matter whether target goal has been achieved. People will keep signing up to be involved in the project. Hence why some of project have backers that is offchart.

Figure 5 Histogram of Failed Backers

Generally, it is harder to convince people to buy ideas. Even if it is brilliant idea, but the marketing is not good, less people will attract or back the idea. This will reflect on the less variance of the failed project.