William Clewett

upenn-phi-data-pt-06-2020-u-c

Excel Challenge - Writeup

A quick note before my analysis: I decided to put the limitations of the dataset before the conclusions piece, since I thought the dataset had some glaring limitations that it would be difficult to draw conclusions without first addressing them. This writeup was done before the bonus, so none of my analysis references the bonus material. Thank You!

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

At the time the data was sampled we know there were approximately 300,000 projects launched through Kickstarter. When we look at the Outcome by Parent Category, we can see that there is a large variance in the number of projects sampled for each Parent Category. It would be beneficial to know that the sampling of the Parent Categories matched their true population distribution. For instance, Theater projects represent nearly 34% of the dataset, while Journalism projects represent just six-tenths of 1% of the dataset. Going deeper, projects that were specifically plays by Sub-Category represented 26% of the data. While this might not seem as impactful as the Parent Category variance, it is important to note that no other Sub-Category has a weight larger than 6.3%. This is the greatest issue with the dataset. We do not know if our data is reflective of the true population when we begin our analysis.

Some of the smaller issues that exist within the dataset include the inclusion of currently live projects and the inability to see the true median donations for projects. Since we are trying to analyze hidden trends within past Kickstarter projects, it would not be in our best interest to look at projects whose results could be subject to change. When we calculate the average donation, we take the total funds donated and divide it by the number of backers. With only one measure of central tendency and no ability to look at the actual donation dataset, our view of each project’s donation framework becomes limited. It could be possible that successful projects benefit from outlier donations, and that it is less likely to succeed from a homogenous set of donations.

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

The data seems to show a high variance in success rate based on which category the project is based in. When we look specifically at the parent category, we see that Music projects have a 79.4% success rate, without the inclusion of currently live projects. On the other hand, technology projects have a significantly high cancelation rate, 29.7% vs the sample average of 8.5%. If we set aside the dataset limitations and were to attempt to design successful Kickstarter project profiles, it would be crucial to pick projects that were based in either Music, Theater, or Film & Video.

Without filtering the Outcome by Sub-Category sheet, we notice that Theater/Plays is clearly the most popular Sub-Category. When we filter by Parent Category, we see that most Parent Categories show nearly binary results at the Sub-Category level. From these results it is easy to conclude that we should have sampled more data for each Parent Category, since most of the Sub-Categories simply do not have enough observations. However, if we tossed the limitations aside and believed that this data was representative of the whole, we could conclude that certain Sub-Categories are currently infeasible on Kickstarter.

Outcome by Date Created needs a little bit more work before we can understand its meaning. Before we make any conclusions on this sheet, we need to create a table that shows the percent of projects that were successful based on what month they were created.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Count** | | | | **% of Grand Total, by period** | | | |
| **Month** | **canceled** | **failed** | **successful** | **Grand Total** | **canceled** | **failed** | **successful** | **Grand Total** |
| Jan | 34 | 149 | 182 | 365 | 9.3% | 40.8% | 49.9% | 1 |
| Feb | 27 | 106 | 202 | 335 | 8.1% | 31.6% | 60.3% | 1 |
| Mar | 28 | 108 | 180 | 316 | 8.9% | 34.2% | 57.0% | 1 |
| Apr | 27 | 102 | 192 | 321 | 8.4% | 31.8% | 59.8% | 1 |
| May | 26 | 126 | 234 | 386 | 6.7% | 32.6% | 60.6% | 1 |
| Jun | 27 | 147 | 211 | 385 | 7.0% | 38.2% | 54.8% | 1 |
| Jul | 43 | 150 | 194 | 387 | 11.1% | 38.8% | 50.1% | 1 |
| Aug | 33 | 134 | 166 | 333 | 9.9% | 40.2% | 49.8% | 1 |
| Sep | 24 | 127 | 147 | 298 | 8.1% | 42.6% | 49.3% | 1 |
| Oct | 20 | 149 | 183 | 352 | 5.7% | 42.3% | 52.0% | 1 |
| Nov | 37 | 114 | 183 | 334 | 11.1% | 34.1% | 54.8% | 1 |
| Dec | 23 | 118 | 111 | 252 | 9.1% | 46.8% | 44.0% | 1 |

Using the above table, we can see how successful projects are within their respective time period. We can conclude that projects started in the second to fifth month of the year did better than the group. To drive this home, when we start a project in between February 1st to May 31st, we can expect a success rate of 59.5%, however when we start a project in between June 1st and January 31st our expected success rate drops nearly 9% to 50.9%. If we decided that this conclusion was worth more research, we could look at trends in consumer discretionary spending, and even what months Kickstarter sees the best cash inflows.

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

When thinking about what further analysis we could do on the Kickstarter data, it is best to start with what data columns were unused by the end of our analysis. What stuck out to me the most was the two Boolean variables, Staff Pick and Spotlight. I think it could be valuable to create a table grouping projects into the following four groups: False Staff Pick with False Spotlight, True Staff Pick with False Spotlight, False Staff Pick with True Spotlight, True Staff Pick with True Spotlight. We could then create a chart that displays the success rate of each group, which would most likely be best displayed in a bar chart. The only issue that could arise with these two variables is that projects become staff picks or are spotlighted only after they are appearing likely to succeed. This would ultimately nullify the legitimacy of the analysis.

Another interesting chart would be a Scatterplot comparing Backers Count and Average Donation, while coloring each point based on whether it was successful or not. With this chart we could visualize whether the success of a project was better determined by Average Donation or Backers Count. In other words, do successful Kickstarter projects more often rely on finding smaller groups of individuals that are willing to give a lot, or do they more often rely on the project becoming popular with a large group of individuals. I created an example below showing projects where donations were USD denominated.

There were outliers left out after >1,000 USD Average Donation and >2000 Total Backers, but it was relevant to note that all projects with greater than 1000 backers in USD denominated projects were successful. Of the projects with >1,000 USD Average Donation, three out of five had failed. Based on the above, I believe it is safe to assume that if your project generates more than 200 individual USD supporters, you can feel very safe that your Kickstarter will be deemed a success.