Kickstarter Projects Analysis

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

* This dataset is likely not representative of the larger population. While we may draw some broad conclusions from this data, there are some disproportionate trends for certain subcategories that should tell us our data collection scheme may warrant some revision.
* The US is the largest market for Kickstarter projects. There are more projects based in the US than any other country sampled—over 70%. That does *not* necessarily guarantee financial success for US-backed projects, merely that there are more projects there.
* The data here seems to indicate the percentage of success (over 40%) is roughly the same or lower than the percentage of failures, especially if combined with cancellations. This is interesting because our background knowledge claims roughly one-third of the outcomes end in success, not the 45% our dataset shows.
* Based on the Goal Analysis, it does seem to be a trend that smaller monetary goals are more conducive to success and larger monetary goals more conducive to failure/cancellation. As monetary goals increase, so do failure and cancellation percentages, while success percentages trend downward. A Kickstarter with a modest goal of less than $5000 has a much higher chance of success.
* Given this dataset, it would *seem* as though projects in the “Plays” category are an emerging trend, having a much higher quantity in total, which yields a higher quantity of success and failure. However, as explained below, if we calculate the standard error of some combined fields and draft charts that contain error bars, we will see some problems with our data’s variability, and we should take that into account when making assertions about Kickstarters as a whole.

*What are some limitations of this dataset?*

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

The main problem with this dataset is that it is very limited in size—roughly 4,100 out of over 300,000 possible projects. There are a couple trends that seem to buck common knowledge of the domain. An immediate trend that seemed odd was the lack of any successful video game projects. Moreover, our background of this topic explained that only one-third of all projects were successful, despite our dataset explaining a success-rate of over 45%. Domain knowledge would most likely indicate this trend may not reflect actual outcomes. With a larger dataset, the conclusion may change. That Kickstarters for plays seemed to have a *much* quantity in general, and, in return, a higher amount of success and failure, seems representative of the fact that this dataset is limited—one may venture to guess that given a larger sample size, the disparity in quantity of plays to the other categories may even out. A Regression Table shows a very large Standard Error, which indicates the dataset may not be very representative of the total population.

Moreover, there are virtually no data regarding backers. This seems to be an important oversight. It would be highly useful for this dataset to contain information regarding backers, if that data exists and obtainable in an ethical manner. The average age of a backer for any given category, for example, would perhaps help campaigns target audiences.

There are a few projects with very large goal amounts—one is $100 million—so that may skew averages taken of that particular field. Is that $100 million goal Kickstarter *really* representative of the success/failure rates in which we are interested? When we look at the number of backers, there are some large outliers, which informs us that using the *median* number of backers may describe the data more reasonably than the mean. For example, there are two successful projects with a large number of backers and *many* failed projects with 0 backers. The median—middle value—will help us understand the data better in this case.

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.795604193 |
| R Square | 0.632986031 |
| Adjusted R Square | 0.632896755 |
| Standard Error | 35615.79546 |
| Observations | 4113 |

We can run a Regression table of amount pledged to number of backers and this will give some answers regarding the dataset’s relationship to the population. With an R Square value of .63, we can say with a fair amount of certainty that the variation of amount pledged *(y)* is described by the variation in backers *(x)*. However, the large Standard Error indicates the dataset contains irregularities and may need expanding. As noted above, the data contains a small sample—roughly 1.37%—of total projects.

Because of the small sample size, error bars are necessary. And, when added to graphs, our data does not look so convincing. Creating a graph with error bars reflecting the standard deviation does seem to confirm the original suspicion that theater projects are not representative of the population because of all the overlap in the bars. It is unlikely we could *conclusively* state any of these values are larger than others in any given category, even the theater projects, which looked to be the largest category by a sizable amount.

In conclusion, while we can construct graphs, charts, and tables regarding this data, if our standard deviation and standard error values are high, it becomes very difficult to use the data in a meaningful way to construct a narrative regarding trends in the population. This dataset should be significantly expanded. Moreover, selection processes may need to be revised to include sample(s) more representative of the entire population.