Data Analytics Bootcamp

Columbia University

Dan Cabrera

12/21/23

Module 1 Challenge

Part I Questions:

* Given the provided data, what are three conclusions that we can draw about crowdfunding campaigns?
* What are some limitations of this dataset?
* What are some other possible tables and/or graphs that we could create, and what additional value would they provide?

Crowdfunded projects—that is, creative projects (e.g., films, apps, games) funded by the public using an online donation platform—are a potentially risky endeavor, as creators often rely on complete strangers to provide necessary capital to launch a project. While many crowdfunded projects succeed in becoming fully funded (i.e., meeting a funding goal the creators set for themselves), many projects fail or are canceled (i.e., the projects do not meet the funding goal but may still get some funding). Creators who are considering crowdfunding can look at historical data of other such projects to help understand the best conditions to launch a successful project. Based on our analysis, it appears that a project dealing with plays, launched in June or July, and with a goal of $1,000 to $1,499 could have a high likelihood of success.

After reviewing data on 1,000 crowdfunded projects, three trends emerge:

1. Theater, film & video, and music are the most popular categories of project. Digging deeper, projects about plays are the most popular type of project by an overwhelming amount. Of the 1,000 projects analyzed, 344 were about plays; the next highest amount dealt with rock music, at 85 projects. Documentary projects were the third most popular, with 60 projects.
   1. While these categories and sub-categories are the most popular, the type of project does not necessarily indicate they will be more successful than projects of other types. At a cursory glance, the projects in the above-mentioned categories and sub-categories succeeded more than they failed, but only marginally. More analysis is required to determine the exact ratios of success and failure, but overall, there does not seem to be one category of project that overwhelmingly succeeds. In fact, of the three top categories, projects succeeded less than 60% of the time. Only one project type, journalism, had a 100% success rate, though there were only four projects in this category.
2. While we cannot make a firm deduction on which project type will be successful, the data indicate that the project’s launch month has some correlation to its success. Projects launched in June or July saw the most success. The failure rate of projects during the same months were not outside of the normal ranges of the number of failed projects. September had the fewest failed projects, but that did not correlate with an increase in successful projects, as the number of successful projects in September was near or slightly below the average. The reason(s) there are more successful projects that launched in June or July is unclear, however.
3. The data indicate that the funding goal may play a role in determining a project’s success or failure. Projects with funding goals of between $15,000 and $35,000 saw the most success, with some funding bands (of $5,000 increments) in the above range showing a 100% success rate. However, there were a very small number of projects with funding goals in the range, with only 38 projects (small relative to the 1,000 we analyzed). A more accurate picture may come from looking at projects with a funding goal of between $1,000 to $1,499. There were 231 projects in this range, and 83% were successful (this was the highest percentage of success apart from the small sample size that had 100% success). This indicates that projects with small goals are naturally more likely to meet or exceed their funding goals.

Further analysis can help us glean better insights. However, there are some limitations to the data, which may make getting insights difficult. A notable limitation is that the dataset only contains 1,000 entries, which may be a small fraction of the total number of projects. We would have to ensure that the dataset is representative of all projects and has statistical significance. Another limitation is the broad categories and sub-categories of the projects. For instance, we do not know if a music project is to put on a live show or record an album; if a play project is to write a drama or stage a musical. Drilling down further into a project’s details could allow for a more refined analysis.

Other possible analyses we can perform and additional tables and charts we can make include the below:

* We could potentially look for the optimal number of backers and the average donation size by comparing number of backers vs. project outcome, and average donation size vs. project outcome. For example, a comparative analysis could show if projects that attract fewer backers with higher donations would likely lead to more chances of success. Or, similarly, if projects that attract many backers with smaller donations could also have positive results.
* The data show the country where the project is located, which could help us see if certain countries prefer certain projects. However, the data does not specify a more exact location (city, state, county), so this analysis would be limited.
* We can see which projects exceeded their goal, and by how much. Finding these ultra-popular projects, and analyzing similarities among them, can help uncover possible strategies for launching a project. Projects with outsized funding can indicate there is a gap in the marketplace for those types of projects; however, there is the possibility that these projects have unique attributes that other projects cannot replicate (e.g., a celebrity launching the project). The dataset does not include information about the projects’ creators.
* Some projects are staff picks or spotlights, which presumably means the staff of the crowdfunding platform pick or highlight specific projects to be featured on the main page or in a list of top picks. Doing so likely raises the visibility and profile of certain projects, potentially leading to more funders and a higher success rate. The number of staff picks and spotlights may be statistically insignificant, but it would be an interesting analysis to see if or how they influence a project’s funding outcome.
* Another analysis could look at the length the project was live on the platform. Each project has a deadline for when it can reach its funding goal. It would be interesting to see if a project’s funding duration contributed to its success or failure. Similarly, we did not analyze a project category’s success or failure rate over time. That is, do some categories always have successful projects, or are some categories fads and backers lose interest in the category over time?

Part II Questions (Statistical Analysis):

* Use your data to determine whether the mean or the median better summarizes the data.
* Use your data to determine if there is more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?

There is an uneven distribution of backers per project. In the 565 successful projects, roughly 70% or projects have fewer than 1,000 backers. Each of the remaining ~30% of successful projects have more than 1,000 backers, and the number of backers for these projects varies widely, from just above 1,000 to almost 7,300.

<1,000 backers

>=1,000 backers

Given this skew, using the **median** number of backers would more realistically summarize the data, as there are more projects with fewer backers and having the lower number (the median being 201 and the mean being 851) is more representative of the majority of projects.

The same is true for failed projects, as the distribution of backers is similar to that of successful projects. The second chart below shows the number of backers per failed project, which is nearly identical to the chart showing successful projects.

Looking at the number of backers alone (especially when not including other data, such as the number of backers compared to average donation, amount pledged, etc.) does not indicate whether a project will be a success or failure. The variability of each dataset is similar. Intuitively, one could reason that the more backers a project has, the more likely it is to succeed. However, many projects have more than 1,000 backers each, yet they still failed. This suggests that success or failure depends on other factors.