Chris Mikus
Data Analytics & Visualization
George Washington University
April 4, 2020

Excel Case Assignment Analysis

Question 1: Given the provided data, what are three conclusions we can draw about Kickstarter campaigns?
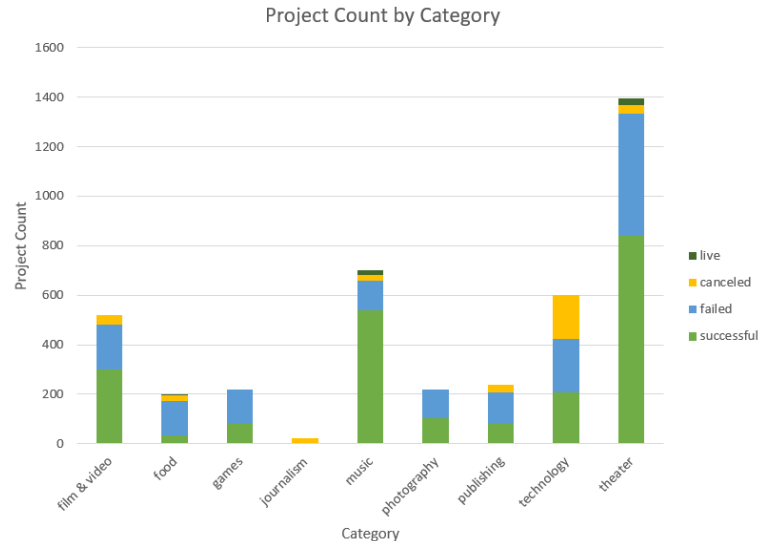
1. Conclusions from Category/Subcategory Bar Charts

Figure 1: Stacked Bar Chart of Project Counts for Each Category

a. Kickstarter campaigns in the music category are the most likely to be successful, with approximately 77% - 540 out of 700 - of all campaigns meeting or exceeding their contribution goals, but success in the music category is very dependent on subcategory. As shown in Figure 2, rock, classical, electronic, metal, and pop campaigns were always successful, while jazz and faith music campaigns were never successful (to date).
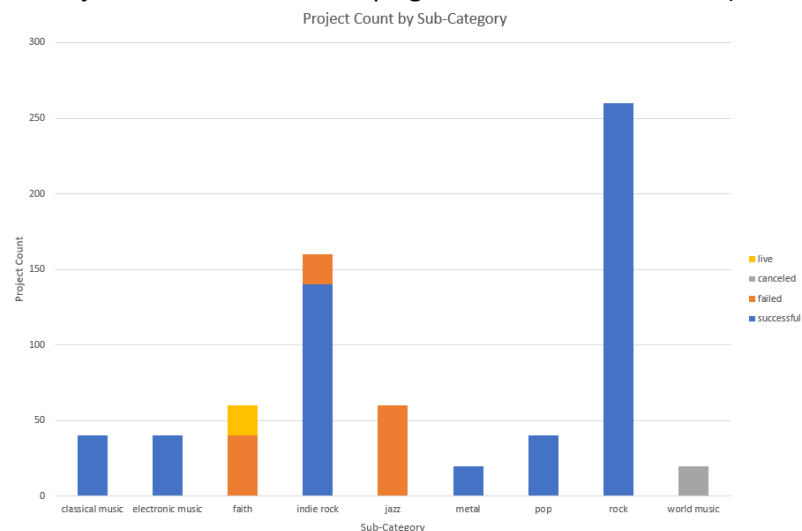

Figure 2: Stacked Bar Chart of Project Counts for Each Music Subcategory

b. Theater projects were the second most successful category, with 60% of campaigns (839 out of 1393) meeting or exceeding contribution goals. Interestingly, the success of music projects does not contribute to the success of musical theater projects, as there were an equal number of successful and failed musical theater campaigns (there were 60 successful and 60 failed musical theater campaigns, with the percentage of successful campaigns dropping below 50% when including canceled projects).
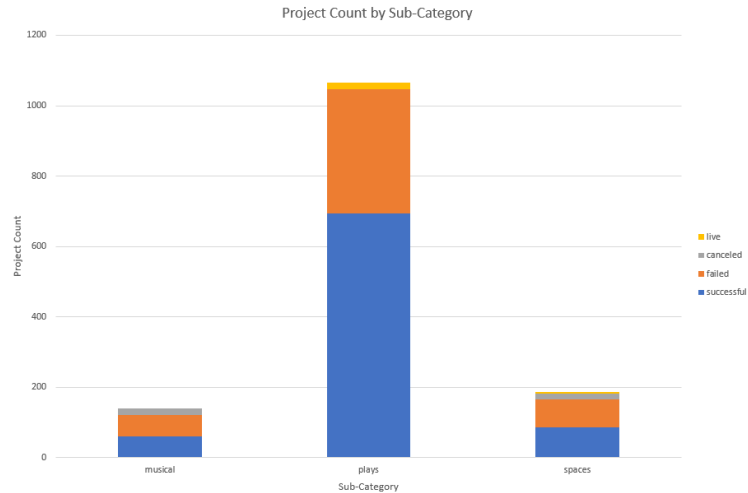

Figure 3: Stacked Bar Chart of Project Counts for Each Theater Subcategory

c. Excluding journalism (of which all 24 projects were canceled), campaigns in the food category have the lowest success rate, with only 17% (34 out of 200) of projects meeting or exceeding contribution goals. As with the music category, the success rate of food campaigns was very dependent on the subcategory. Shown in Figure 4, all Kickstarter campaigns for food trucks and restaurants did not meet contribution goals. However, all small batch food campaigns were either successful or are currently live, so small batch campaigns still have the potential for success on Kickstarter.
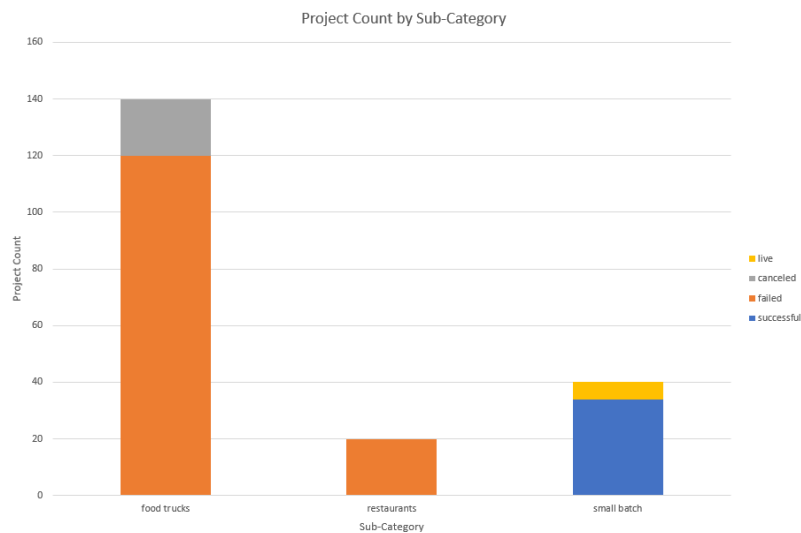

Figure 4: Stacked Bar Chart of Project Counts for Each Food Subcategory

2. Conclusions from Project States by Project Launch Month Line Chart
    a. As shown in Figure 5, from 5/17/2009 to 3/15/2017, projects launched in the month of May have accounted for the highest number of successful projects as well as the highest percentage of successful projects, with approximately 61% of all projects launched meeting or exceeding contribution goals. However, the ratio of successful to failed projects is highest among projects launched in February (1.91 successful campaigns for every failed campaign, as opposed to 1.86 for May launches, see Excel worksheet PivotTable_Line_Count). Kickstarter campaigns launched between February and May have the highest success rates (with successful projects constituting 60%, 57%, 60%, and 61% of total projects launched in February, March, April, and May, respectively).
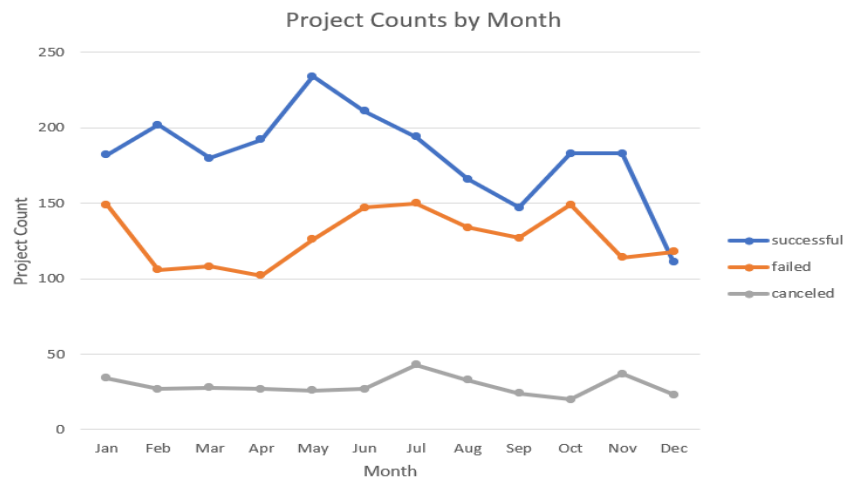

Figure 5: Line Graph of Project Counts by Launch Month

3. Conclusions from the Bonus Line Chart (Outcomes Based on Goal)
    a. Projects with a contribution goal under $10000 carry a success rate of over 50%, while campaigns with a contribution goal of $45000 or greater carry a failure rate of over 50%. Canceled projects typically increase as the project goal increases (from 4% for goals of less than $1000 to approximately 23% for projects with goals above $50000).
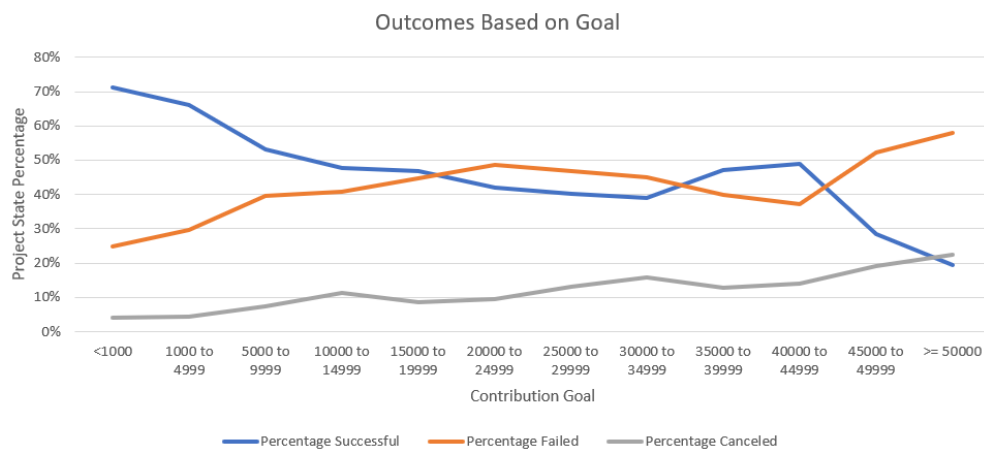

Figure 6: Line Graph of Project Outcomes Based on Contribution Goals

Question 2: What are some limitations of this dataset?

1. It is important to note that just because a campaign was successful on Kickstarter does not mean that the project was developed after funding was obtained. Likewise, projects that failed to achieve goal contributions on Kickstarter might still have been completed via alternative means. A principal limitation of the data is that it cannot be used as a reliable metric to predict the types of projects that will reach full development and implementation – analysis is restricted to the study of projects that are likely to obtain contribution goals on Kickstarter.
2. Both "Successful" and "Failed" projects have a concrete set of conditions that must be satisfied to qualify for placement in their respective categories (successful projects met contribution goals, and failed projects did not). "Canceled" projects do not have a clean set of conditions and analysis of these projects is thus limited in utility. Canceled projects may have been "cut short" by those proposing the campaign because it was apparent that the project would fall short of contribution goals. As such, there may be instances where canceled projects should be considered the same as "failed" projects, which could affect the percentages of project states.
3. Although the data contained in the "pledged" column is crucial in determining whether a project reached goal contributions, it is limited in the sense that individual contribution values are not provided. Since a normal distribution cannot be assumed for donation amounts, calculating an average donation could lead to inaccurate conclusions. Some projects may have received most of their funding from a few generous donors and may be indicative of the vested interest of a few people rather than more widespread interest among the general populace. Using individual contribution data to generate a median donation could help mitigate this situation.

Question 3: What are some other possible tables and/or graphs that we could create?

1. Contribution Goals by Category/Subcategory: it was observed that certain categories and subcategories were much more successful than others, and that cheaper projects were more likely to be successful. It could be worth investigating whether more successful categories enjoyed higher success rates because their contribution goals were generally lower.
2. Launch Month/Year vs. Project State Percentage: for the case homework, a line graph was generated showing the project counts by the month in which the project was launched, with a different line for each project state (successful, failed, canceled). However, this presentation is susceptible to misinterpretation because some months had higher project counts than others. A higher number of successful projects does not mean that the projects were more likely to be successful – it means that there were more launched that month. Line graphs with project state percentages would mitigate the differing project counts across launch months.
3. Project State Percentages by Country: it would be worth investigating the success rates of campaigns by the countries in which they were launched. There may be regions of the world where public awareness of Kickstarter may not be as prevalent as others, and it would be worth shifting marketing resources to increase awareness of the platform in these regions.
4. Launch-to-Deadline Duration vs. Project State Percentages: from the data set, campaigns are visible on Kickstarter for different amounts of time. It would be worth investigating how the amount of time a project is on Kickstarter can affect the likelihood of a project's success. For those trying to determine a project schedule and looking to allocate an appropriate amount of time to generate funds, this could also help answer the question "What is an ideal donation window that will optimize donation acquisition speed with project success rates?"

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

The median summarizes the data more meaningfully than the mean. From the statistics produced for the data, the mean and median for both the successful and failed projects are much closer to the minimum donor count than the maximum donor count, indicating that both data sets have a skewed distribution primarily composed of smaller donor counts with a handful of larger donor counts (note that "smaller" and "larger" are relative terms to each data set – successful projects are expected to have a higher number of backers than failed projects). The presence of these high donor count "outliers" increases the mean of the data sets because an average depends on the actual values in the data set. A median is outlier-resilient, as it considers the *positions* of the individual data points in the set (rather than the actual values), using a singular point in the middle of the data set as an approximation for the central value. The use of a median as a descriptive statistic is more useful for data that does not follow a normal distribution.

From the variance statistics, there is more variability in the data for the successful projects than in the failed projects. This makes sense because there are more methods to achieve a successful outcome than there are ways to achieve a failed outcome. Successful projects might have obtained their goal contributions from a small number of large contributions, many small contributions, or a combination of the two. Failed projects, on the other hand, will result from the combination of low backer count and low contribution. As a result, we see a much wider range in backer counts for successful projects (26456) than we do for failed projects (1293). Variance is determined by first squaring the difference between each individual data point and the mean of the data set, and then calculating the average of these squared differences. Data sets with individual values falling far from the mean will drastically affect the variance. The data set of successful projects contains more of these values, leading to the higher degree of variability in the data.