Daniel Palombo

February 4, 2021

Rutgers Data Science Bootcamp

Homework 1

1. Given the provided data, what are three conclusions we can draw about Kickstarter campaigns?
   1. One conclusion is that when looking at the totals is that a little over half of the Kickstarter campaigns succeed while the other half are either fail or are cancelled. When looking at the data by categories you can conclude that the major of success in Kickstarter campaigns are in the categories Film & Video, Music, and Theater. Those three categories are the only ones that succeed more than half of the time while every other category underperforms in comparison to the total averages.
   2. Another conclusion when looking at subcategories is that within the three categories that have the most success, that success comes from a few subcategories that have 100% success while the other subcategories have 0% success. For example, within the category of Film & Video, Documentary films, Short Films, and Television all have 100% success while the other subcategories have 0% success. Also, within the category of Music, Classical, Electronic, Indie Rock, Metal, Pop, and Rock all have at least 87% success while the other subcategories have 0% success. Lastly, within the category of Theater, Plays have 65% success while the other subcategories have below a 46% success. Although the theater category is more evenly spread in comparison it still shows a somewhat similar trend to the other two categories being analyzed.
   3. My last conclusion I take from this dataset is that there is a total of between 250 to 400 Kickstart campaigns each month with the majority happening within the first 6 months. However, while the number of failed and canceled Kickstart campaigns stays pretty consistent month to month, the number of successes drops dramatically in the month of December.
2. What are some limitations of this dataset?
   1. Some limitations are that it does not show any percentages in the current data set which makes it difficult to compare two categories or subcategories that have a large difference in the number of kickstart campaigns they launched. Since the number of Kickstart campaigns launched varies dramatically between categories it could result in inaccurate conclusions. For example, if the Theater category may seem like it is the most successful because it has the largest number of successes, however, it also has the largest number of failures. This is why in some cases it could be useful to look at percentage success rate instead of looking at the total number. Another limitation is that is difficult to compare month to month success and failure because the total number of Kickstart campaigns fluctuates month to month. So, naturally if there are a larger number of Kickstart campaigns in a month there will be more successes, but there will likely also be more failures. The one benefit to this is to see which months these metrics go in opposite directions and to be able to tell a story with that. A good metric to make could be showing the change in the difference between success and failure month to month. This would show what months successes increase/decrease more than failures or that failures increase/decrease more than successes.
3. What are some other possible tables and/or graphs that we could create?
   1. Some other possible tables and/or graphs that we could create are the ones I mentioned above. For example, a table that shows percentages of successes, failures, cancelled, and live Kickstarts for different categories and sub categories. Along with this table could go pie charts that show visualizations of how different categories and subcategories look in terms of success rate.
   2. Another possible table could be one that shows the month-to-month differences between success and failures for each category. Then you could make a line chart that shows positive months to launch a kickstart and negative months to launch a kickstart based on historical data. This could allow for Kickstart campaigns to be more connected to the timing that they will be successful in.
4. Bonus:
   1. Use your data to determine whether the mean or the median summarizes the data more meaningfully.
      1. In this dataset the median seems to summarize the data more accurately because there are some outliers on the high end of the dataset that skew the data. Therefore, the median is more meaningful in this dataset.
   2. Use your data to determine if there is more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?
      1. There is more variability with successful campaigns. This makes sense because unsuccessful campaigns are likely to be consistent with having a low number of backers with fewer outliers while success campaigns may have a larger variety of possibilities in how they raised money for their campaigns. For example, a successful campaign could have had less backers who all put larger amounts of money or may have had many backers who all put a little amount of money. Because they raised more money, the possibilities of number of backers that provided that money increases.