UT Data Analysis Bootcamp

Tuesday/Thursday Section

Brian Fitzpatrick

02/02/2020

**Excel Homework Analysis (Due 02/08/2020)**

1. Given the provided data, what are three conclusions that we can draw about Kickstarter Campaigns?
   1. That in general Theater projects are the largest category in our data, and one of the only categories that possesses examples of projects in each state, the others being music and food. Further that within it’s subcategories, plays are the most successful with about 40% failing. The other two subcategories fail at closer to a 50% rate.
   2. In comparing successful and failed examples (worksheet ‘Analysis\_PT’) we can observe several interesting differences between two.
      1. That a successful project has a much higher average number of backers than an unsuccessful project (194.4 backers to 17.7 backers)
      2. That a successful project has a much lower average goal than an unsuccessful project ($9,866.97 to $60,556.39)
      3. The difference between average goal and average pledge is significantly larger for an unsuccessful project than a successful project (Unsuccessful project average pledged – average goal = -$58700.85 compared to a successful project’s difference of +$8712.30
   3. I would conclude from the below that the most important deciding factor in if a Kickstarter is successful or not comes from how achievable of a goal it sets for itself. If you turn your attention to the worksheet Bonus 1\_chart, you will see demonstrated a couple things.
      1. With a goal<$1000.00. Over 70 percent of projects are successfully funded through Kickstarter.
      2. As the goal amount increases, the trendline for successful projects is negative (since the lines are correlated, a see a positive trendline for unsuccessful projects, as well as cancelled projects)
      3. Beyond $40,000 dollars, we see a much sharper drop off in successful projects. Suggesting there may be something that requires more exploration of what might be happening at that level of goal.
      4. Similarly, we see the direction of the line of successful project goals reverse from its downward trend between $30,000 and $40,000. This movement suggests there is anomaly happening within that range and requires further investigation.
2. What are some limitations of this dataset?
   1. My short answer to that would be this is not a representative sample of Kickstarter projects; let’s work through that in two parts. The first is are we are getting enough data. In our set, we have a total of 4114 projects across various categories going back to 2009. However, if the internet is to be believed, a total of over 450,000 projects have been attempted through the platform. If the 4114 projects was a representative sample, this might be alright, but there are also problems within the data set, leading to my second point. This is almost certainly not a representative sample given that we have a lot of nulls we can observe when we start feeding the data into a pivot table. Just as an example, in my “PT Stacked Column,” if you expand the category ‘film & video’ you will observe that each of the six sub categories has a null value for one of the other possible state flags (being successful, failed, live, or cancelled.) So for this chart, it would suggest that 100% of animation projects have failed, whereas 100% of documentaries have succeeded in being funded. These null values occur across the data set.
3. What are some other possible tables and/or graphs we could create?
   1. Histograms for number of backers, average goal, average pledged amount, and average donation, separated by state flags.
   2. Cat and Whisker plots for the statistical analysis bonus section
   3. I would probably generate a pie chart of the number of projects in each parent category just for completionist sake.

**Bonus 2 Analysis**

1. Does the mean or the median summarize the data more meaningfully?
   1. Given that in both examples there is a very strong right skew, I believe that median gives the better picture of the information. In a successful project, the median falls solidly between the 1st and 3rd quartile, while the mean is above the 3rd quartile, suggesting that the outliers within the upper bound are pulling it further away from the center of gravity of the data. This also occurs with the failed projects, though not to the same degree.
2. Is there more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?
   1. There is more variability with successful projects than unsuccessful campaigns based on the number of backers. I think this is partly due to two things happening within the data. For a successful project, it can succeed with just one backer, but in general a successful project has on average more backers than an unsuccessful project. With Unsuccessful projects, the mode of the data is the value 0. Since an unexpected has fewer successful backers on average, you would expect to see a lower variability within the data.