Kathryn Panger

Spring 2020

HOMEWORK QUESTIONS

1. **Given the provided data, what are three conclusions we can draw about Kickstarter campaigns?**
   1. Campaigns that ask for less than 15,000 and campaigns which ask for more than 45,000 are most likely to succeed.
   2. Campaigns which ask for amounts of money between these two values have comparable probabilities of success and failure.
   3. The greater the amount of money associated with the campaign goal, the more likely the campaign is to be cancelled.
2. **What are some limitations of this dataset?**
   1. There are too few variables to build a robust model explaining the likelihood of Kickstarter success.
   2. Sampling methods are unknown, which means that findings may be biased.
   3. Small range of years limits any analysis of change over time.
3. **What are some other possible tables and/or graphs that we could create?**
   1. Boxplots summarizing data from successful, failed, and cancelled Kickstarters.
   2. Bar chart showing the amount of money donated on average per donation category (i.e. film and video, etc.).
   3. I don’t think excel can do regression analysis, but that would be a good thing to do with **state** as the dependent variable and any or all of the remaining measured variables as the independent variables.

BONUS QUESTIONS

1. **Use your data to determine whether the mean or the median summarizes the data more meaningfully.**
   1. The median summarizes the data more effectively. This is because of the high levels of variance in backers for both successful and failed campaigns with a skew toward their minimum values.
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 variance among successful campaigns than unsuccessful ones. This makes sense for two reasons: 1) Because the conceptual bounds for successful and unsuccessful campaigns differ, and 2) because of the high proportion of campaigns that earned exactly zero dollars. I will explain both of these reasons below.

Reason #1:  
Unsuccessful campaigns by definition always earn less money than their goal, whereas successful campaigns will by definition have always meet or exceed their goal. This means that the values of failed campaigns conceptually have a lower bound of 0 and an upper bound of their goal amount, however successful campaigns will always have a lower bound of their goal amount with no upper limit. Because there is no upper limit for successful campaigns but there IS an upper limit for failed campaigns, we can expect greater variance in successful campaigns. There are some conditions under which this reasoning would not apply (e.g. if there is a strong relationship between campaign success and goal amount) however I would say that this line of reasoning, while not guaranteeing this pattern, is a sensible way of thinking about the pattern we have already observed.

Reason #2:  
There are probably a large amount of failed campaigns which earned exactly zero dollars. When many data points share the same value, this shrinks the variance of a sample. Dealing with many observations that share a value of zero is a common problem when analyzing data because it shrinks the variance and makes the mean less representative of the center of the data.