Homework Questions

1. **Given the provided data, what are three conclusions we can draw about Kickstarter campaigns?**

    - Theatre is the parent category with the highest number of projects created by people.
    - Among all the sub categories, play is the most popular one in terms of number of projects.
    - In terms of the created date of projects, May is the month where most successful projects were created in.

2. **What are some limitations of this dataset?**

    - The dataset doesn't have any information on the authors of projects so we don't know whether some successful projects come from same authors/creators.
    - Another limitation could be the data set only provides the number of backers and the total amount of fund that each project pledged but it doesn't show the amount of money that each backer offers. For example, some projects might have smaller number of backers but all backers provide a large amount of money compare to those have a larger number of backers but with smaller amount of money.

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

    - According to the graphs and charts that we have done, we have some ideas about the most popular category of projects, but we didn't include any information on which category has the highest percentage of successful projects and what's the difference in percentages among different categories. I think we can create a table which enables us to make the comparison among the percentage of successful, failed, cancelled and live projects in each category.
    - We can also create a graph on the relationship between the length of each project and their successfulness.

Bonus Questions

1. **Use your data to determine whether the mean or the median summarizes the data more meaningfully.**
   Based the data analysis that I have done for both successful and failed campaigns. I believe the median summarizes the data more meaningful as the median represents the middle value of the set of data. However, the mean value includes the animalities and outliers which will drag the mean value higher that it would be. In this case, both the maximum values in both successful and failed campaigns are dramatically higher and it indicates that what I mentioned above has a even higher chance to happen.

2. **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 higher variability with successful campaigns as it has higher variance and standard deviation. Basically, variance represents the difference between number of x and average value, higher variance usually means the data sets are more spread out. I think it does make sense to me as you can tell successful campaigns have a wider range between its maximum value and minimum value (26457&1). However, for unsuccessful campaigns, the range between its maximum and minimum is narrower (1293&0). Therefore, successful campaigns will be more variable. You can also compare their standard deviation (844 V.s 61.43) which has a better indication since standard deviation are the standardizes values. Furthermore, we can see some connections and correlations between number of backers and the successfulness of the campaigns. The number of backers could indicate the probability of success in some extend.