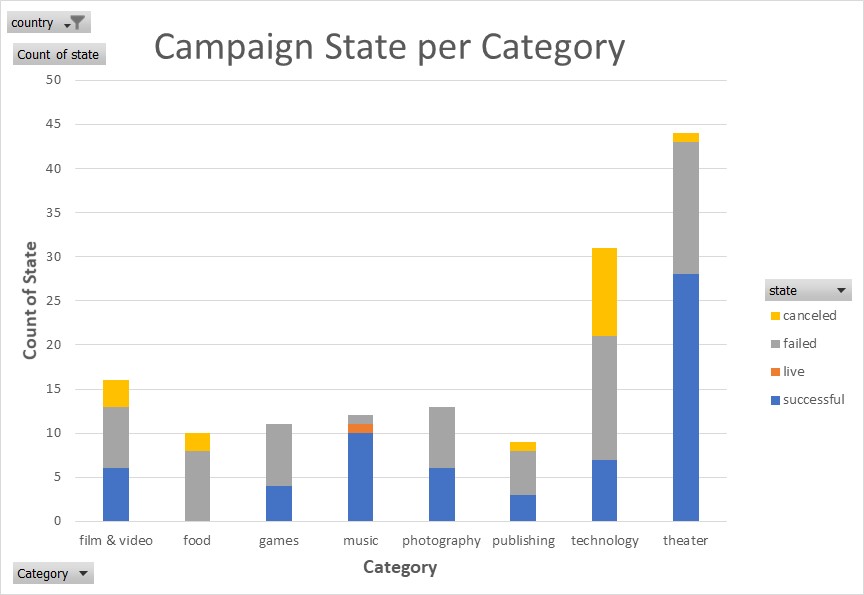
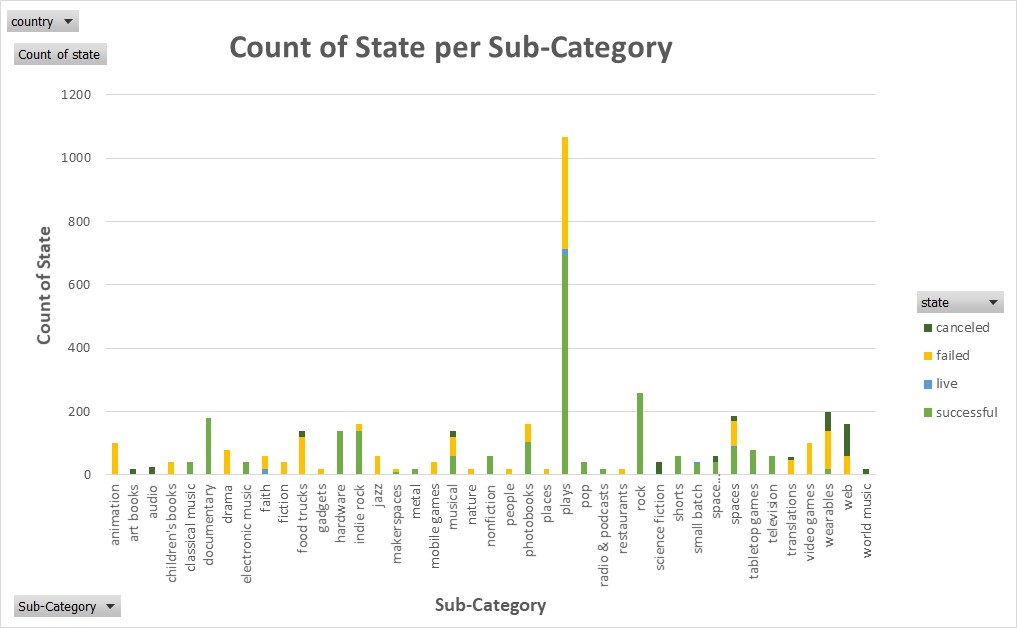
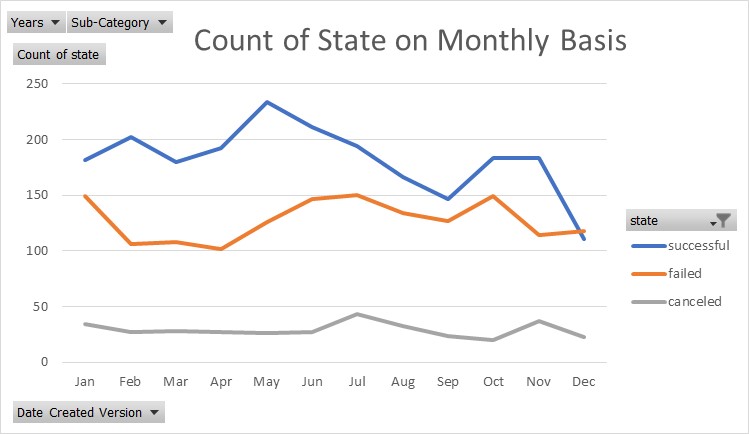
Name: Thao Hoang

Date Submitted: Sep 23, 2020

**Unit 1 – Excel Homework**

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





Insights from the graphs above:

- Figure 1:

* Out of all categories, **Theater** has the greatest number of total projects launched, then follows by technology, then film & video, and so on.
* The stacked bar graph helps to visualize how successful a category is and thus the investors can focus on the categories with highest successful rate, such as theater, music, photography, and film & video. They can also lessen the money spent on less successful projects in categories such as games, publishing, and technology. Since none of the launched projects were successful in food category, they should not spend money in this category.

- Figure 2:

* The sub-category that has the exceptionally high number of projects launched is Play, which is ~5times higher than many other sub-categories. Within the Play sub-category alone, about ~60-70% of the launched projects are successful.
* Similarly, the investors can start to narrow down which sub-categories are worth investing money in based on this graph, such as: Plays, Documentary, Classical Music, Hardware, Photobooks, Rock, Small Batch, Tablet Games, etc. They can then save money from spending on projects with higher failure rate such as: Animation, Drama, Fiction, Video Games, etc.

- Figure 3:

* This graph shows the trend in success and failure rate according to the month of the year. According to this graph, we see that the successful rate is highest in the first half of the year and lower in the second half of the year, the peak rate is during May.
* The failure rate is exactly opposite of the success rate. Thus, the investors should spend more money in launching projects during the first half of the year.

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

- The number of projects launched in each categories and sub-categories are not equal, few of them have less then 5-10 projects launched in the past. Thus, the investors must keep in mind that the results do not represent the population sample, they can provide inaccurate insights.

- The data only includes the targeting countries and not the entire the world. Thus, the data insights can be applicable to the countries listed but not other parts of the world.

- Each project was only launched for ~1 month before it is marked as successful, failed, or cancelled. This only tells us if the project was successful during that period but not the rest of the year. This can create fault conclusion about the true potentials of those projects that were marked as Failed during the period they were tested.

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

* World Map with highlighted colors that show the countries with higher successful rate.
* Time wise: besides monthly trend, we can also look at quarterly, seasonal, etc
* Instead of stacked bar chart, we can also create pie chart for the same purpose but different visualization
* Create a chart that shows Percent Funded and Average Donation per Category and Sub-category.
* Create a chart that shows number of Packers Count per Category and Sub-category to identify which ones have the most potential supporters if they are launched again.

**BONUS**

- See Excel File, “Bonus” Sheet

**STATISTICAL ANALYSIS**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Successful | Failed | Canceled |
| Mean | 194 | 18 | 27 |
| Median | 62 | 4 | 2 |
| Minimum | 1 | 0 | 0 |
| Maximum | 26457 | 1293 | 1501 |
| Variance | 712841 | 3773 | 11615 |
| Standard Deviation | 844 | 61 | 108 |
| 1st Quartile or 25% Percentile | 29 | 1 | 0 |
| 2nd Quartile or 50% Percentile | 62 | 4 | 2 |
| 3st Quartile or 75% Percentile | 141 | 12 | 10 |
| IQR = 3rd Q – 1st Q | 112 | 11 | 10 |
| Lower limit for outliers | -139 | -167 | -168 |
| Upper limit for outliers | 309 | 180 | 178 |

**1/ Use your data to determine whether the mean or the median summarizes the data more meaningfully.**

- The histograms (shown in the Excel file, “Statistical Analysis” Sheet) were created for each of the successful/ failed/ canceled data set. Neither of them follows a normal distribution curve. Thus, the Median would be a better prediction of the chance of success/ failed/ canceled. This is because some values are extremely large or extremely small compared to most other values. Thus, when we average out the data set with these outliers, the mean value is not meaningful.

**2/ Use your data to determine if there is more variability with successful or unsuccessful campaigns. Does this make sense? Why or why not?**

- Variability measures the amount of dispersion of the dataset, or how far away the data points tend to fall from the center. When a distribution has lower variability (narrower distribution curve), the values in a dataset are more consistent. Understanding variability helps you grasp the likelihood of unusual events.

* **Variance:** According to the table above, “Successful” has the largest Variance value, thus larger variability. As the result, “Successful” campaigns data set is less consistent, and the data are more dispersed from each other and from the Mean.
* **The interquartile range (IQR)** is also a robust measure of variability in a similar manner that the median is a robust measure of central tendency. IQR is a great alternative to Standard Deviation when the data does not follow normal distribution such as in this case. Quartiles divide a rank-ordered data set into four equal parts, where:
  + Q1 is the "middle" value in the first half of the rank-ordered data set.
  + Q2 is the median value in the set.
  + Q3 is the "middle" value in the second half of the rank-ordered data set.
  + The interquartile range is equal to Q3 minus Q1.

This case, our calculated IQR values show that “Successful” campaigns have much larger IQR values, thus higher variability of the data set. This makes sense as mentioned earlier, “Successful” campaigns have more extreme outliers compared to “Failed” and “Canceled”, which makes the data set more dispersed.

Source: <https://stattrek.com/statistics/dictionary.aspx?definition=interquartile%20range>

Standard deviation