HOMEWORK – 01 - EXCEL

Cindy WONG

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

* The data range are quite large, i.e., min = 4, max = 344. Distribution of the samples are unevenly distributed.
* The most popular crowdfunding category is PLAYS where they have the most successful amongst the other categories. Despite that, it also has the highest failed campaigns.
* The min & max count in category is smaller compared to sub-category. In the sub category data range, there is an extreme data point – “Play” whereas the rest of the sub category are pretty consistent.

1. What are some limitations of this dataset?

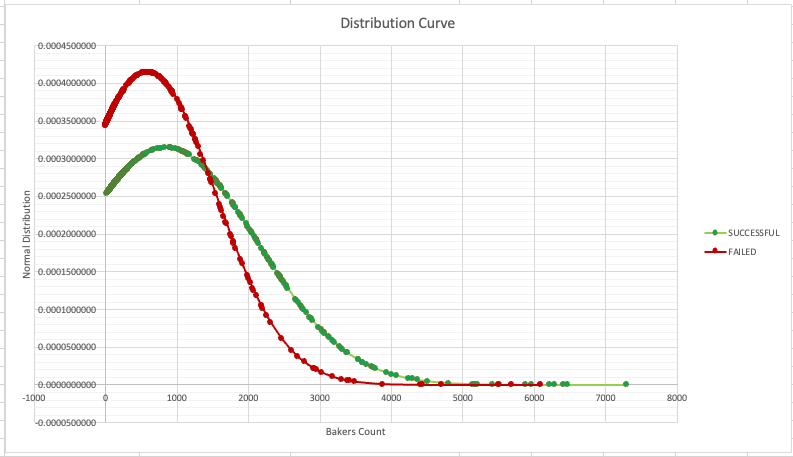
* **QUALITY OF THE DATA:** Data collected from different sources might vary in quality and format, might be inconsistent, i.e., people who might pledge on this project might not have access/ did not get a chance to pledge on another project.
* **CONFLICTS on biased views and personal preferences**
* **DESIGN LIMITATIONS - METHODS/ TECHNIQUES** **USED TO COLLECT THE DATA:** The data collection process, i.e., has it included all of the data, how was the data collected, did everyone answer the questionnaire truthfully etc.
* **INCOMPLETE DATASET** – Perhaps during the data collection process, if all of the value has been entered correctly.
* **TIME/COST CONSTRAINTS**

1. What are some other possible tables and/or graphs that we could create, and what additional value would they provide?

* We could create a normal distribution chart to see if the data is normally distributed, using the information of Mean, Median, Min, Max, Variance and Standard Deviation.
* Having the min & max value, we could see the range of a dataset, i.e., how spread out the data is.
* By plotting a normal distribution graph using scattered plot, we could get the information if the data is normally distribution, skewed right or left.
* When mean & median have values that are close, we could then use the mean to describe the overall dataset. However, if the value is significantly different, we could use median instead/ best represent the data. When mean is used (normally distributed graph), we could use standard deviation to describe the spread. However, if the graph is skewed left/right, we could use IQR to describe the spread.
* With variance information, we could see if a dataset has more variability. The smaller the number of the variance, the lower the variance.

**Bonus Challenge**

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



* From the dataset that we have, as the graph is right skewed and the value difference for mean and median is quite significant. Hence, it would be better to use a median to summarise/represent the data.

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

Successful campaigns has higher variability because the standard deviation of the number of the backers count is higher.

Yes, it makes sense as high variability represents a less consistent data, so it’s harder to make a prediction as the datasets are more dissimilar and might indicate that the extreme values doesn’t represent the overall datasets.