# **DATA 542-FINAL REPORT**

INITIAL STATISTIC:

The part A of the project involved the initial statistic. This was the crucial phase which involved reading the two csv files and a lot of cleaning of the data. Setting up the code to walk through all the csv files and then cleaning them was a tedious task. The steps followed for the initial statistic were:

* The first part involved reading only the all detailed files and the newest review files independently and load it into the data frame.
* The next step involved setting up the right indexes for the data frames, getting rid of the duplicates, as well as removing the columns that had only NaN values.
* Further refining of data was done to remove the duplicates based only on the apptitle and the text as it was observed that few appTitles and text were being duplicated. Hence it was necessary to get rid of them in order to have clean data which could be further used for processing.
* The newest review file when loaded had approximately 2.7 million records and after cleaning and refining the data, about 1.1 million records were used for further processing.

TEXT PROCESSING:

The part B of the project was the most interesting part. This involved cleaning up the reviews which had been put forth by the users. Before this, in order to answer few questions, the content rating column from the all detailed files was merged with the newest review files in order to accommodate

Questions based on sub categories. For part B, the following procedure was followed:

* For removal of non-English characters, the NLTK library and its words corpus was used. Though the corpus of NLTK is not exhaustive, it was a good way to carry out the cleaning part of the text.
* Removal of punctuations and non ascii characters were carried out using regular expressions.
* However, the removal of reviews having just two words was something I felt was not needed, meaning we need to retain reviews even if they are just two words. The reason for this is that even just two words for app or just a word can give the sentiment or the feeling the user had towards the app and that information could be used forward for the purpose of sentiment analysis.

After considerable amount of cleaning and pre-processing the data, yes, we can say that there is difference in the reviews. The thorough analysis done using different plots and normalization is an evidence for that fact.

LEARNINGS:

Data wrangling is one of the most important skillset that Data scientists would posses. This project helped me gain a good insight of how it would be to deal with messy data. The project helped me improve my skills from in understanding and enhancing my data pre processing skills. Extensive usage of Pandas library coupled with other libraries like NLTK helped to understand the power these libraries posses which help in data wrangling. The most important aspect of data analysis understood is that DATA is always messy and it needs to beaten to get good meaning out of it.