Summary of EDA

Summary of the data set that, at a minimum, answers the following questions: What is the unit of analysis? How many observations in total are in the data set? How many unique observations are in the data set? What time period is covered?

- This dataset has features that will help in analyzing the sentiments of the tweets. So, we will divide the sentiments into three different stage which is Neutral, Positive and Negative. So this will be the unit of our analysis. For now, we have in total 50000 tweets from where we have extracted features that will be used to calculate a score which will be used further classify the tweets into different sentiments. The whole datasets are the unique observations as we have avoided the retweets from others. The streaming of the Tweepy API considers the current period of the time during which it is run and collects the tweet for that particular time.

Brief summary of any data cleaning steps you have performed. For example, are there any particular observations / time periods / groups / etc. you have excluded?

- Data Cleaning: Different data cleaning was used to check the missing values, converting categorical data into numeric for the model to evaluate. Reducing the dataset by removing the redundant values in the beginning while streaming the tweets.

Description of outcome with an appropriate visualization technique.

- The visualization technique we will use is the scatter plot for different features to understand the relation between each features with every other. The following will be the graph of all the plots that we have found:

Description of key predictors with appropriate visualization techniques that compare predictors to the response. You should investigate all predictors in your data as part of your project. For the purpose of this assignment, pick the one or two predictors that you think are going to be most important in explaining the outcome. Your selection of predictors can either be guided by your domain knowledge or be the result of your EDA on all predictors.