Summary of EDA

Tweet Dataset:

Dataset Characteristics:	Multivariate, Text	Number of Instances:	50000	Area:	Text
Attribute Characteristics:	Real	Number of Attributes:	11	Date Created:	1st of March, 2020
Associated Task:	Classification, Clustering	Missing Values?	None	Type of Learning:	Unsupervised

Dataset Information: The dataset was collected using the Twitter API called Tweepy. Tweepy API provides access to the entire RESTful API methods. Methods accept various parameters and returns responses. To collect the records, the tweets were streamed and stored in a json format first and then converted to the csv format. It has 11 attributes and more than 50000 records collected over three days' period.

Attribute Information: The attributes collected here are the information about the tweets like, from where it was tweeted (location), who tweeted it (id), when this was tweeted (created_at), how many followers does the person doing tweet have (followers_count) etc. These attributes are considered to be parameters that will help in determining the sentiments of the tweets.

Time Period Covered: As the Tweepy API helps in extracting the tweets of the time when it is running, we have collected the recent tweets over the three days' period, that was since 28th Feb, 2020.

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:

There were two different stages where data cleaning was performed:

1) During Streaming: This was the time when with the use of Tweepy API we listened to the tweets and started collecting the tweet's information. Every tweet had over 25 attributes and not all of them were needed for the data analysis. We captured those features and attributes that were necessary to evaluate the sentiments of the tweets and discarded rest of them. This is one of the dimensionality reduction technique where unnecessary attributes were removed. To tackle redundancy, we removed the retweets from consideration and so set the retweeted status to False.

2) After Streaming: The first thing to be taken care was the missing values. Sometimes there remains some missing value that can result in bad analysis, so removal of those missing values was a good step. We could have replaced the value with the mean or median of the rest of the tuples but it doesn't seem right because it is hard to judge something like sentiments by just some random numbers. The next step of cleaning was to remove all the stop words from the tweets. Stopwords are useless in such analysis and will only add to the computation and complexity of the task. This will help us further in just focusing on the words that actually add meaning to our analysis. This way we could focus more on the adjectives that was used which is considered to be an important factor of analyzing the sentiments of any tweet.

Different Types of Plots for Visualization:

Histogram Plot: This plot could be used to find the number of tweets that were made between different groups.

Bar Charts: It can be used to categorize the tweets based on the locations of the tweets. Like number of tweets that came from New York.

Pie Chart: Percentage of tweets that came from certain zone. Suppose if we categorize the locations around US in 5 zones. Then the percentage of tweets from each zone could be projected in the pie chart.

Scatter Plot: Scatter plot can be used to visualize the relation between two attributes to decide whether there is any correlation between any attribute or not in order to counter any multi-linearity problem.

Here are some of the plots that were visualized and analyzed by us with each briefly described below:

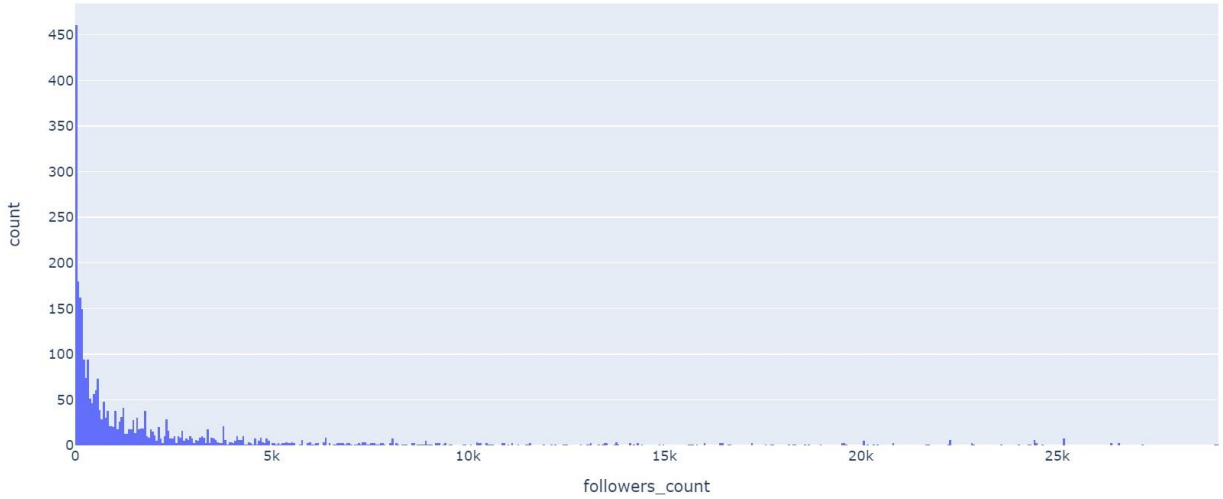

Fig1. This is the histogram plot for the followers_count

This is the plot for the number of tweets that were tweeted by the people that falls under a certain range of followers. From this we concluded that individuals having followers count between 0-49 made highest number of tweets which was 460.

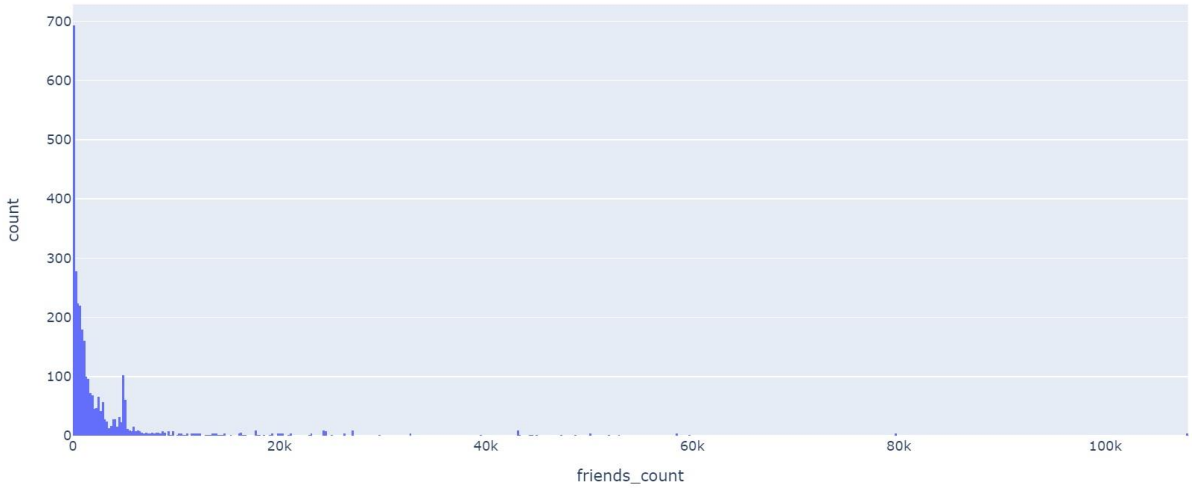

Fig2. This is the histogram plot for the friends_count

This is the plot for the number of tweets that were tweeted by the people that falls under a certain range of followers. From this we concluded that individuals having followers count between 0-49 made highest number of tweets which was 460.

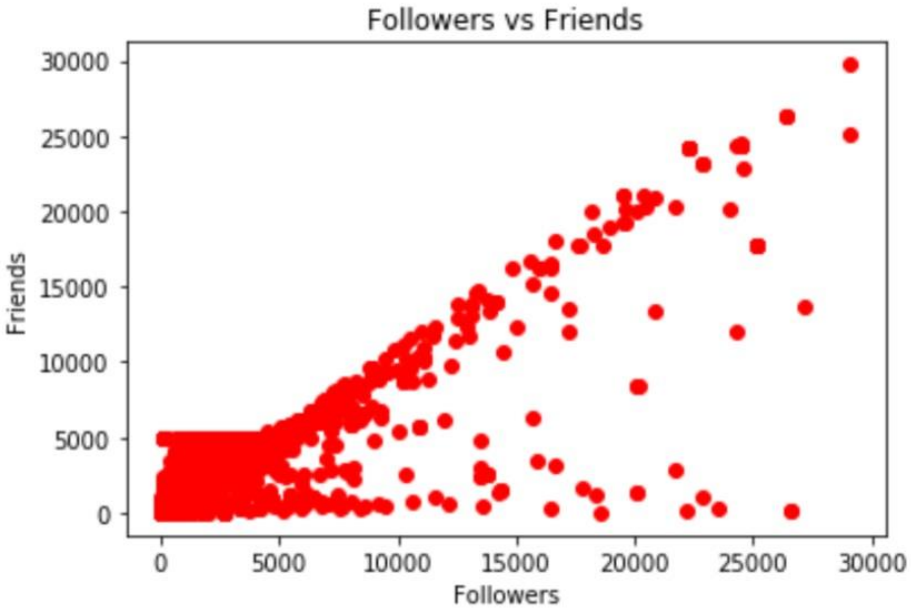

Fig3. Scatter plot of Followers vs. Friends

This is the plot between the number of followers and number of friends of the individual tweeting. First of all, looking at it there does seems to have some relation between the two, but cannot say it for sure as looking at the lower left section we can says something else might be going and to understand that we need to do a plotting for those having followers count less than 5000, and the same goes for the few data points that are on the lower right section of the graph. This needs a further analysis.

Note: *These are just the few plots and few of the visualizations. There could be a lot more things to be discovered*