Milestone 3

Submit a 2-3 page pdf with Revised Project Statement, EDA, and baseline model.

You should also include any jupyter notebooks you used to produce the EDA and baseline.

Your submission should include:

A description of the data: what type of data are you dealing with? What methods have you used to explore the data (initial explorations, data cleaning and reconciliation, etc)?
Visualizations and captions that summarize the noteworthy findings of the EDA.
A revised project question based on the insights you gained through EDA.

Trump Tweets- Milestone 3

Donald Trump, the 45th president of the United States is seen as a highly controversial figure and is very well known for use of Twitter using his handle @realdonaldtrump. As President, Trump has a unique source of power and even a slight social media presence may heavily impact changes on a national scale. Some see his Tweets as reckless and some others enjoy his content. His tweets may have a long lasting universal effect and we will observe that effect on a particular category.

The S&P 500 is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States and represents a market cap of over $25 trillion. It is one of the most commonly followed equity indices, and many consider it to be one of the best representations of the U.S. stock market. In a recent experience with someone from a big investing firm, it was clear the many of the giant trading companies use social media to aid in trading and those trades performed in large quantities may ~~cause~~ correlate with market changes. We will observe the effect of Trump's tweets and presence on twitter to the Stock Market data, particularly the S&P 500 index movement.

A couple connections to observe include (but not limited to):-

1. Impact of Trump Tweets (Quantity) on stock market
2. Impact of Trump Tweets (Sentiment) on S&P 500

We shall use the above connections, datapoints and given tweets in order to predict the stock market price change after the release of the tweets. In other words, we want to predict how people trade or how the markets react after Trump Tweets. A complication here is that the markets are open only on certain days and closed on other days such as weekends and holidays, so we shall attempt to spend some time understanding the impact of time interval on prediction. A potential method to do so would be to compare prediction accuracy within a day or more than a day based on the tweet release date.

Data Sources

Data Source 1: Twitter data from http://www.trumptwitterarchive.com/

Columns = source, text, created_at, id_str

Count: 28,505 tweets

Data Source 2: Stock Market (S&P 500 Index) Stock-mendation, Sharecast Data Source Account

Columns = Date, Open Price, Close Price, High Price, Low Price, Volume

Count: 1483 Dates

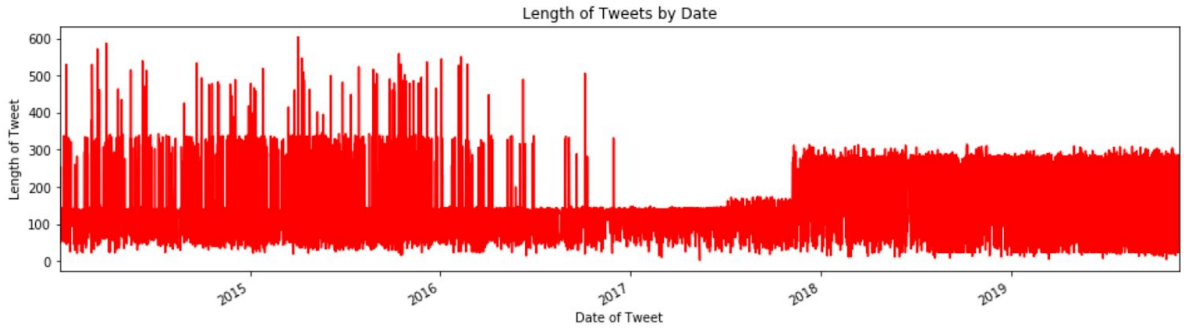
We have also found sources and libraries online which would enable us in our analysis, we shall mention these sources and library in our final paper.

Data Description- What data are we looking at?

The dataset for Trump tweets consists of all of Trump's Tweets back from 2014 and similarly, the dataset for stock market data grabs the S&P 500 ETF costs back from January 1st of 2014. Columns of interest for the trump dataset include the actual textual data and the date of tweet. Other data such as the source of tweet (iphone android) can also be explored.

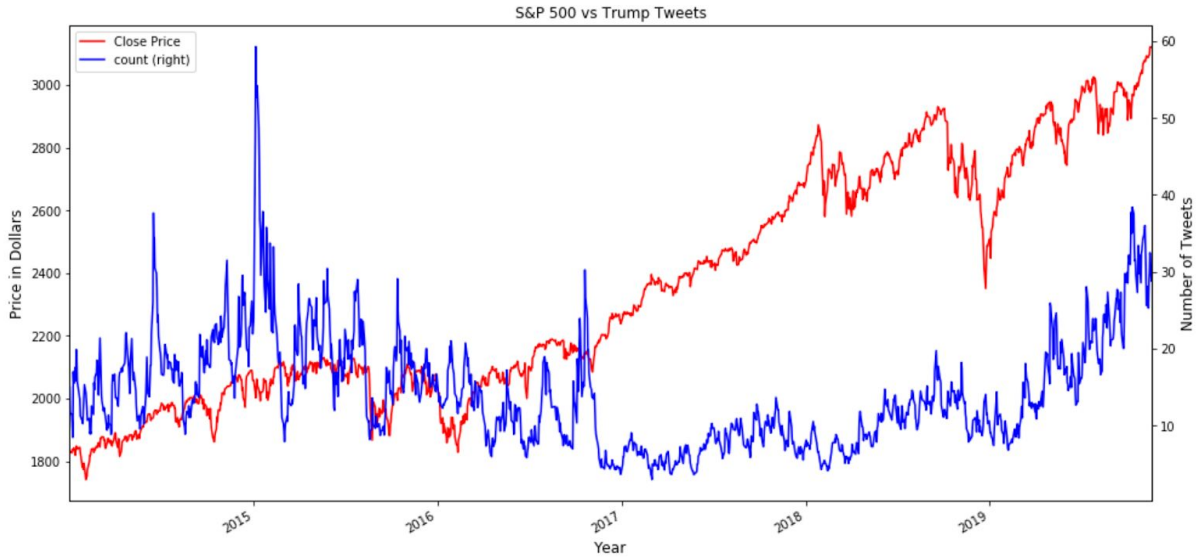
Methods

The dataset for Trump tweets provided some interesting results for his tweets back from 2014. The dataset consists of the created_at column, text (tweet content) as well as the tweet source (device). We can see that there are 28504 observations for these three columns. Also, the created_at columns has the correct format for datetime, so converting it to a datetime object would enable a time series analysis to be performed. The image below depicts this plot.



After cleaning the dataset which constituted to removing outliers of extreme lengths most likely due to errors while obtaining the dataset, we can see that the average length of a tweet was 144.87 characters. As for the source of Tweet, Twitter for iPhone was the most used followed by Twitter for android and then Twitter web client. There were several other sources used, but those were used so rarely that it's almost negligible. The next step is to clean the actual textual content of the tweets because many of the tweets contain emojis and other non-traditional characters that our code cannot recognize. Using regular expressions, the non-traditional characters and any

symbol distinct to an alphanumeric value will be remapped into a new one that will work with our code. After cleaning the text, we can perform our sentiment analysis to see whether the content of Tweet is positive, negative or neutral in nature using the TextBlob library in Python. Here we obtain: Positive tweets: 55.07%, Neutral tweets: 26.57%, Negative tweets: 18.35%. The accompanying Jupyter notebook has more observations on stock data. We have correlated stock data to the number of tweets and found a slightly negative correlation coefficient of -0.11. Below is a plot depicting the number of tweets and the closing price of stocks since 2014 until November 18th 2019 (Updated data).



The presidency of Donald Trump began at noon EST on January 20, 2017, which may be worth taking a closer look at.

So, to improve the initial question we had, we will explore whether the content/sentiment and number of Trump's tweets will cause an upward/downward shift in S&P 500 markets.
NB: The accompanying notebook will also have any preliminary model, but half of our group members dropped out of the group, so we'll try our best to get this model working. If not, we shall include it in our next milestone.