Original Research Question

We originally sought to investigate if fluctuations in the US stock market or presidential approval polls could predict President Trump's tweeting frequency.

Thoughts About Data and Sources

Market Data

We initially explored the course-provided market data. We quickly realized that these data were perhaps too niche for our use. Given that we were working from the assumption that President Trump would be tweeting as a result of headline news, we sought a way to capture overall market movement and market sentiment. We thus selected the SPDR S&P 500 Trust ETF (**SPY**) and CBOE Volatility Index (**VIX**) for our dataset.

We searched through several sources for data, including Quandl, Tiingo and IEX, but ultimately settled on Yahoo Finance for its ease of use and limited restrictions on redistributing the data.

Sources:
VIX - https://finance.yahoo.com/quote/%5EVIX/history?p=%5EVIX
SPY - https://finance.yahoo.com/quote/SPY/history?p=SPY

Polling Data

Polling data was provided by FiveThirtyEight's presidential approval tracker. They track individual polls being published, and publish a weighted aggregate version of all polls combined. We chose to use the output of their model, which aggregates the individual polls (weighted by reliability, sample size, and more).

Sources:
MODEL DESCRIPTION - https://fivethirtyeight.com/features/how-were-tracking-donald-trumps-approval-ratings/
TOTAL - https://projects.fivethirtyeight.com/trump-approval-data/approval_polllist.csv
TREND - https://projects.fivethirtyeight.com/trump-approval-data/approval_topline.csv

Tweet Archive

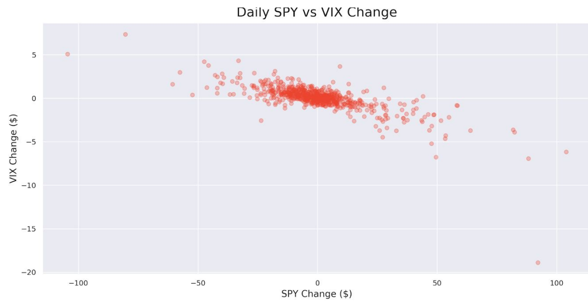
We scraped the president's tweets from the Trump Twitter Archive. Our research question is focused on the relationship between approval rate, stock market movement and tweet frequency, thus we decided to use a subset of the available date starting from when he became president. We also noticed that the days with the highest tweeting frequency were days where news came out about impeachment.

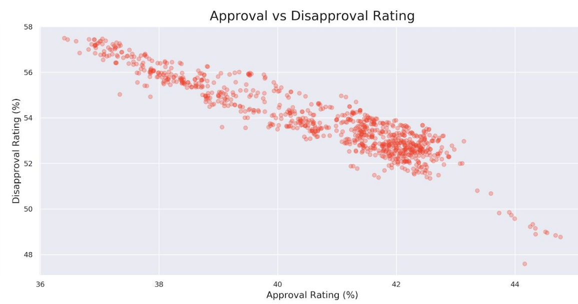
Sources:
ARCHIVE - http://www.trumptwitterarchive.com/archive

Our Path Down Exploratory Data Analysis

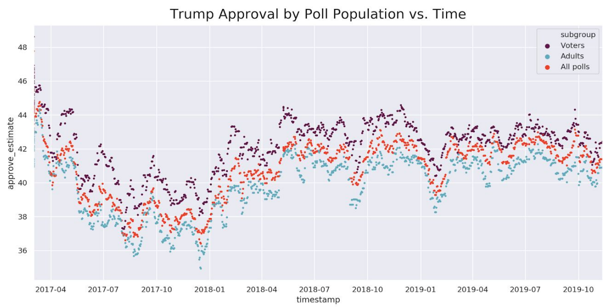
Visualizing the Data

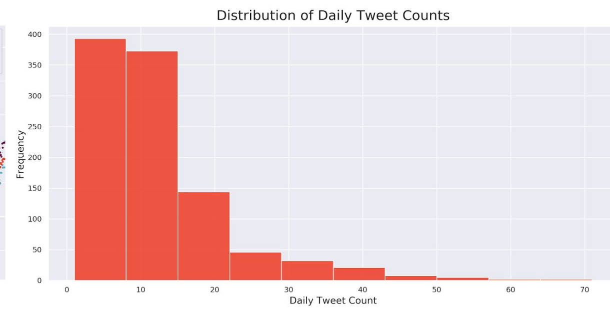
Upon first review of the separate sources of data, we decided to explore the different distributions and seek out any multicollinearity issues. From a logical standpoint, it seemed we should expect to find a connection between approval vs disapproval ratings and SPY vs VIX. As expected we discovered this upon plotting one against the other.





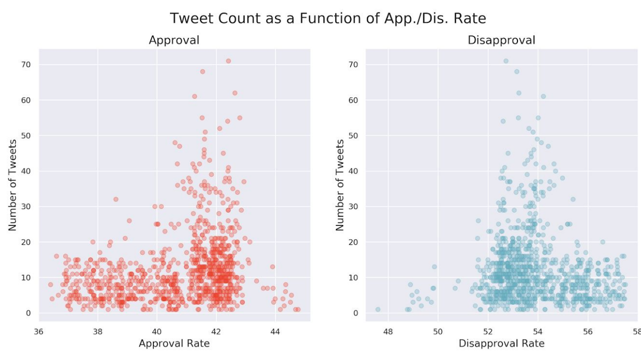
We also knew that the polling data we had collected consisted of multiple subpopulations based on likely voters, registered voters and a combined all. Plotting this (bottom left) we saw that the subgroups move almost exactly with one another, so we chose to use only the "all polls" observations (rather than potentially introduce bias from limiting the population subgroup).

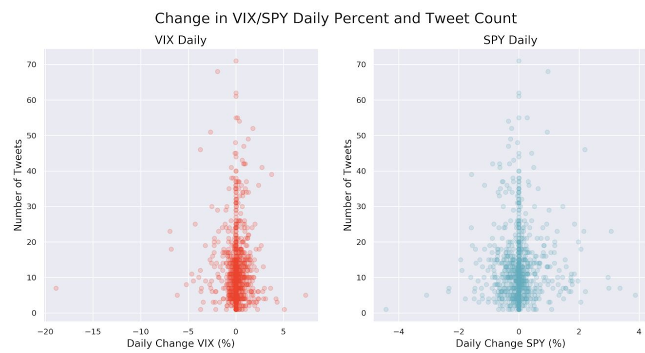




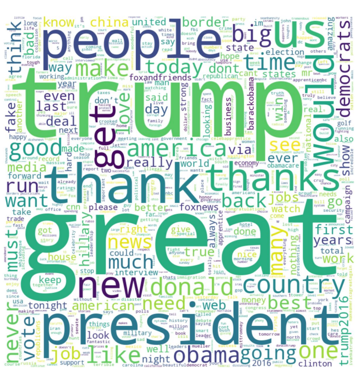
Another issue that we expected might come up, was a problem with the distribution of our outcome variable. As we graphed this (top right), we saw that it was indeed highly skewed. We noted that we might need to log-transform the counts for models that assume normally distributed outcomes.

Upon merging the data to a single dataframe, we next decided to start plotting the relationships between the different predictors and our outcome variable of daily tweet counts. However, no clear relationships stood out.





We then considered that we might be able to get more from these graphs if we were to assign each tweet to a specific topic. Our hope was that if we broke down the words used in the individual tweets, we would be able to find a way to categorize each of them. However, Trump seems to frequently use the same words that have no value for assigning a topic.



To the left is a wordcloud representing the frequency of particular words in all of the tweets we reviewed, after having cleaned out stopwords and other noise. With no luck here, we attempted to learn data-driven topics via the LDA algorithm in the gensim library. However, the learned topics did not clearly correspond to real-world topics. Many of them only consisted of words like “new”, “great”, “big”, “president” and “America”. At this point in our EDA process, we decided to no longer pursue this approach. Sometimes EDA can help you find paths not to go down.

Handling Missing Data

One issue we encountered with our data was the missing data that resulted from merging all the different data sources. Markets are closed on weekends and holidays, but Trump tweets during this period. Approval ratings were also not published daily. We carefully considered data imputation methods:

- **Approval rating -** We used linear interpolation, assuming that approval and disapproval likely did not swing wildly between surrounding estimates.
- **Volume -** For volume we decided to use the overall mean, as zeros would be outliers, and there is little day-to-day correlation in volume.
- **Daily Change Variables** - Part of our hypothesis for the project was that dramatic change would increase Trump's tweet frequency. Thus, it seemed imputing this as zero would capture the truest behavior during this time period.
- **Market Position** - Here we carried the Friday close value across the weekend. If the market was up, it might change his mood and behavior, and it created the most realistic version of the truth.

Revisiting the Research Question

There were a few instances during this process where we considered revising our research question. The most prominent of them was while exploring Trump's tweets and trying to assign topics to each one. If were were to do this, we thought it might be possible to develop a model that both predicted the frequency in which the President tweets, but also could predict the topic.

However, after being unable to assign tweets to all of the topics, we decided it would be best to stick with the original question at hand.

The Base Model

We explored many different models, documented in our notebook, but settled on random forests. The lack of any visible linear relationships in our EDA, and the inherent nature of what we are trying to predict guided us toward something based on a decision tree. The problem with a simple decision tree, or ones with boosting or bagging, is that it doesn't help us get at the answer to our research question. They provide tweet frequency estimates, but they don't tell us if the market or polling data are more important.

With random forests, the subsampling of features for each node will help us escape the shortcomings of a greedy algorithm. This design will prevent the same features from always being selected, and it will give us a better understanding of what type of predictors are most important.