Executive Summary

DS 710 Final Project by Moko Sharma [Mukund Raghav Sharma]

"Is there a correlation between the Sentiment of the Tweets of major Political Leaders and the Price Range of the SPY?" is the question I am trying to answer for my final project for DS 710. The three political leaders' tweets I will be considering are: **Donald Trump**, **Mike Pence** and **Paul Ryan**. For those who don't know, the **SPY** Exchange Traded Fund is the best-recognized and oldest ETF and is widely considered to be a great proxy variable of the health of the United State's Financial Markets. Hence, the over arching theme of my project is to deduce how much political announcements effect the financial markets.

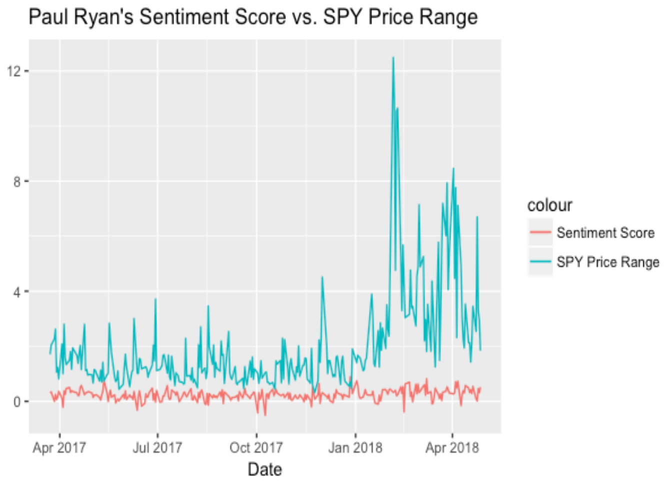
The reason I decided to answer the aforementioned question is the fact that I work at a Proprietary Trading Firm that primarily deals with Equity Options and have always been interested in studying the immediate effects of announcements political leaders on the financial markets. In the world of Equity Options, volatility of the underlying instrument is an important driving factor of the prices and hence, choosing the daily Price Range of the SPY was a natural choice to experiment with. The answer to the question I am addressing will be of interest to Equity and Index options traders and Quantitative Researchers who are constantly trying to decipher market signals and concocting up new strategies based on the changing political-economic environment.

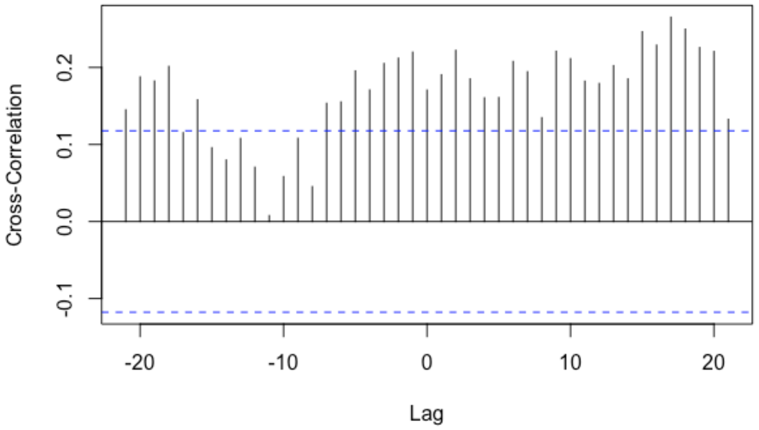
The two data sets that were acquired for the analysis were: the day-by-day Sentiment Score associated by the cleaned tweets and retweets of a political leader obtained by Tweepy's REST API and the SPY daily historical price data obtained by using the Pandas Data Reader. To generate the day-by-day Sentiment Score, **NLTK's Vader Sentiment Analyzer** was used and the average of the positive, neutral and negative sentiment score from this model was computed for each cleaned tweet. The cleaning process involved stripping off any hyperlinks, non-alphanumeric characters and Retweet identifying information. Next, the sentiment scores for tweets in one day were averaged up so that there was a grouping of the sentiment score by day and then converted into a dataframe that was merged with the **Historical SPY Price** dataframe on the Date. The resultant dataframe for each of the political leaders was saved as a CSV file to be used for analysis via R.

The CSV files were loaded into R as dataframes for analysis and the time series portions of the dataframe were extracted out. Since we are dealing with time series data and want to conduct a correlation analysis we have to handle the issues of autocorrelation, trends and seasonality that is sometimes known as non-stationarity. To discern whether the time series at hand was stationary, we make use of the **Augmented Dickey-Fuller** test; using the **Durbin-Watson** test checks the autocorrelation of a time series. For each of these tests, we use a significance level of 0.05. If a trend, seasonality and / or autocorrelation are detected, differencing a time series is used to nullify the effects. Once the effects of non-stationarity and autocorrelation are removed we proceed to plot the **Cross-Correlation** between the two time series in question and then conduct the **Pearson's product-moment correlation** test. The Cross-Correlation plot will give us an indication of how similar two series are as a function of the displacement of one relative to another.

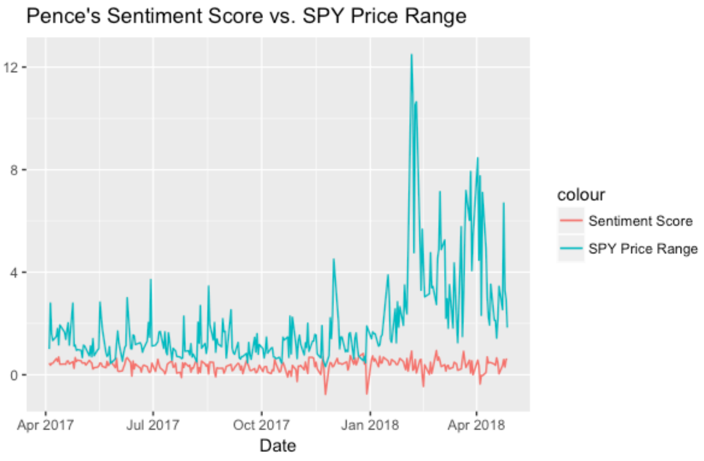
The Cross Correlation plots and Pearson's product-moment correlation test's p-value both highlighted the fact that neither Donald Trump nor Mike Pence's tweets' sentiment scores were correlated with the SPY price range as we failed to reject the null hypothesis. However, Paul Ryan's tweets' sentiment scores were significantly correlated with a p-value of **0.004** and significant cross-correlation across a majority of lags considered. The alternative hypothesis for the aforementioned correlation test is accepted for the case of Paul Ryan. From this conclusion, we can use this information to prospectively consider making volatility trades or at least consider **Paul Ryan** as an important figure whose tweets have historically, been correlated with the SPY price ranges.

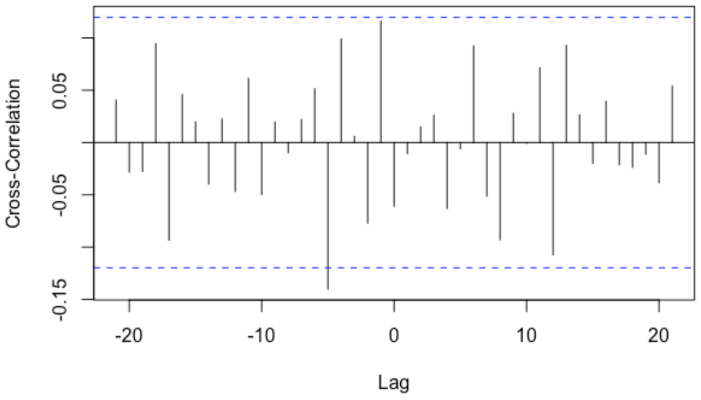
Paul Ryan’s Plots [p-value = 0.0043 for Pearson’s product-moment correlation]



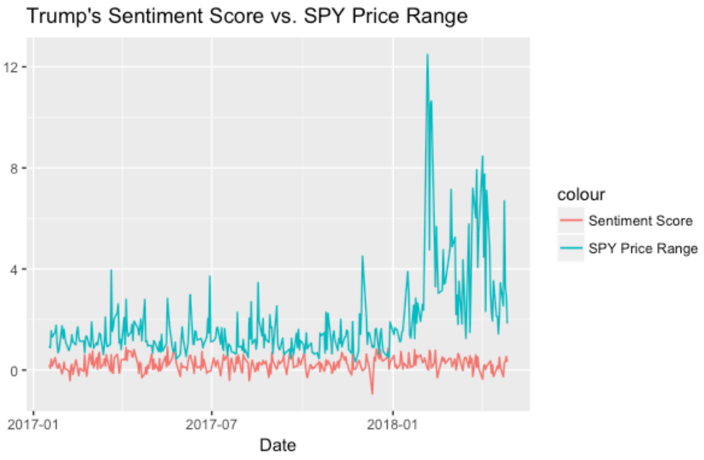


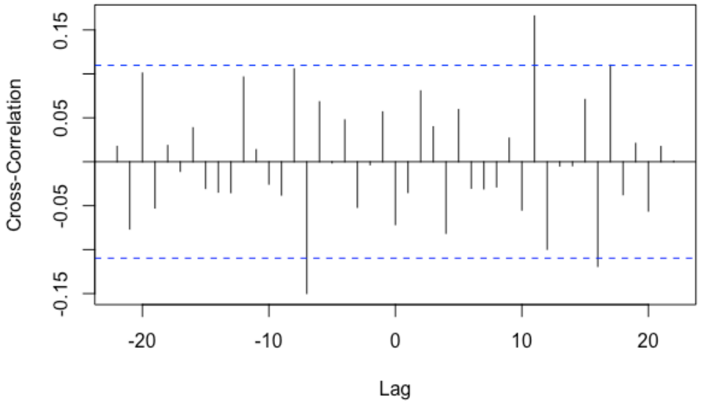
Mike Pence’s Plots [p-value = 0.3213 for Pearson’s product-moment correlation]





Donald Trump’s Plots [p-value = 0.1901 for Pearson’s product-moment correlation]


Trump's Sentiment Score vs. SPY Price Range
Date
colour
Sentiment Score
SPY Price Range


Cross-Correlation
Lag