Final Project - MATH 2820 (L)

Contributors:
Grant Bowlds: Wrote the code for the data mining, sanitized the tweets and completed the analysis between volume and effect on Dow Jones.
Immanual John Milton: Wrote much of the final proposal and final project. Looked into completing analysis on other world leaders, prevented by Twitter policies
James Tang: Wrote code to find the common words from President Trump. Analyzed the impact of tweets on the market to the market and on international countries

Introduction:

The main goal of this project is to quantify to the extent possible the relationship between the number of President Trump's tweets and the Dow Jones Industrial Average during the time from Trump's inauguration until 3/31/2020. The data used includes all of Trump's tweets from his inauguration until 3/31/20, as well as the Dow Jones Industrial average from that timeframe; data was collected and compared per day to grant the largest sample size to approach the population size. We also analyzed whether the content of the tweets was relevant to the performance of the market. Our hypothesis is that tweets about the market would have a positive effect on the market, while tweets about foreign powers would have a negative effect on the market. We do not believe that the number of tweets will directly affect the market in a measurable, quantifiable way.

Our R code and data sources are at the bottom of this document for reference. The other code is spread throughout this project.

Model:

Data Gathering

We began by gathering the relevant data. This began with the retrieval of the data that we need to use to test our hypothesis. We first gathered the Tweets that President Trump tweeted using Kaggle user Austin Reese's account that had scrapped all of President Trump's tweets from April 15th, 2019 until January 20th, 2020 and placed the contents of each of the tweets with the associated timestamp into a CSV file. A representative subset of that Excel file is included in the repository below.

```r
library(ggplot2)
library(dplyr)
library(readr)
library(lubridate)
library(tidyr)
library(wesanderson)
library(ggrepel)


tweets <- read.csv(file = "C:/Users/grant/OneDrive/Desktop/MATH 2820L/Twitter-DJI/Data/trumptweets.csv")
dji <- read.csv(file = "C:/Users/grant/OneDrive/Desktop/MATH 2820L/Twitter-DJI/Data/^DJI.csv")
```

Our next step was to gather the other piece of data that was essential for this project. We needed the stock information from during this same time period. We decided to use the Dow Jones Industrial Average to best analyze the status of the stock market on that particular day. We once again used an outside resource: Yahoo Finance. Yahoo Finance included a CSV file of the market information of every day including every day of Trump's presidency, which included the timeframe that we were looking at analyzing. A representative subset of that Excel file is included in the repository below.

```r
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(ggplot2)
library(foreign)
trumptweets <-read.csv(url("https://raw.githubusercontent.com/grantbowlds/Twitter-DJI/master/Data/trumptweets_processed.csv"))
DJI <- read.csv(url("https://raw.githubusercontent.com/grantbowlds/Twitter-DJI/master/Data/%5EDJI.csv"))
```
```

Data Processing

The next step in this process was to process the data that we received from the two sources. We had gigabytes worth of data that we needed to pare down to be able to provide meaningful analysis. We first did this by only selecting the data from Trump's inauguration date (1/29/2017) to 3/31/2020. Then, using the Tidyverse package features, the list of dates were completed to integrate the dates that President Trump did not tweet. Weekend dates were also removed as the market, thereby Dow Jones, obviously is not open on weekends. A new column was added to the Dow Jones Industrial average CSV file to calculate the daily percent change from open to close of the market. These two data sets were then merged into a single data frame that could be used for visualization and analysis. Tidyverse packages were also used to run analysis and mine text from the tweets themselves. To test some of our hypotheses, dates, where the tweets did not meet our testing criteria, were not used.

```
dji <- dji %>%
  mutate(perchange = (Close-Open)/Open *100)

dji <- dji %>%
  mutate(perchange = (Close-Open)/Open *100)

tweets <- tweets %>%
  mutate(date = ymd_hms(date))
tweets <- tweets %>%
  mutate(date = round_date(date, "day"))

dji <- dji %>%
  mutate(date = ymd(Date))


overlapDates <- interval(ymd(20170120), ymd(20200331)
```

```
tweets <- tweets %>%
  filter(date %within% overlapDates)
dji <- dji %>%
  filter(date %within% overlapDates)

tweetBD <- tweets %>% group_by(date) %>% tally()

tweetBD <- tweetBD %>%
  mutate(date = ymd(date))

tweetBD %>%
  mutate(date = as.Date(date))

tweetBD$n[tweetBD$n>5] <- 5
tweetBD$n <- as.integer(tweetBD$n)

dplr <- left_join(dji, tweetBD, by=c("date"))
```

```
right = replicate(804, 0)
right <- as.integer(right)

dplr$n <- coalesce(dplr$n, right)

nZero <- dplr %>%
  filter(n == 0)
nOne <- dplr %>%
  filter(n == 1)
nTwo <- dplr %>%
  filter(n == 2)
nThree <- dplr %>%
  filter(n == 3)
nFour <- dplr %>%
  filter(n == 4)
fiveUp <- dplr %>%
  filter(n >= 5)

suppressMessages(library("dplyr"))
```

We also used the tidytext to identify key features from the content of Trump's tweets. Using a sentiment analysis package, we were able to create a dummy variable, flagging each tweet as either positive or negative. We then mined our tweets to create a list of Trump's most commonly used words and bigrams, which we later utilized to flag tweets containing economic or international relations related terms. Grouping by day, we merged this dataset with our Dow Jones Industrial Average CSV file. We added columns containing the frequency of market mentions per day, the frequency of international mentions per day, and the sentiment ratio (positive tweets divided by total tweets) per day.

```r
# Additional Processing
names(trumptweets)[10] <- "sentiment"
names(trumptweets)[11] <- "market_check"
names(trumptweets)[12] <- "intl_check"

text_results <- trumptweets %>%
  mutate(Date = as.Date(date)) %>%
  group_by(Date) %>%
  summarise(market_mentions = sum(market_check),
            intl_mentions = sum(intl_check),
            sentiment_ratio =
              length(sentiment[sentiment=="positive"])/ length(sentiment))
DJI <- DJI %>%
  mutate(Date = as.Date(Date))


DJI_text <- right_join(DJI, text_results)
DJI_text <- na.omit(DJI_text)

DJI_text <- DJI_text %>%
  mutate(net_market = Close - Open) %>%
  mutate(market_groups = ifelse(market_mentions == 0, 0,
                                ifelse(market_mentions >= 1 & market_mentions < 5, 1, 2)),
         intl_groups = ifelse(intl_mentions == 0, 0,
                              ifelse(intl_mentions >= 1 & intl_mentions < 5, 1, 2)))
```

Data Plots

In order to figure out Trump's specific impact on the market based on his comments on other countries and the market itself, we mined the most common words that Trump said. The relevant code is below with the results of the code.

```r
# ------------------------------------------------------------------------
# Exploratory Analysis
# ------------------------------------------------------------------------

# Top Words
replace_reg <- "https?://[^\\s]+|&amp;|&lt;|&gt;|\bRT\\b"
# Top Single Words
words <- trumptweets %>%
  mutate(text = str_replace_all(content, replace_reg, "")) %>%
  unnest_tokens(word, content, token = "tweets")
words <- words %>%
  anti_join(stop_words, by = "word") %>%
  filter(!str_detect(word, "realdonaldtrump"))

words_count <- words %>%
  group_by(word) %>%
  count()
trump_words <- words_count %>%
  arrange(-n)

# ------------------------------------------------------------------------
```

```r
# Top Word Pairs
bigrams <- trumptweets %>%
  mutate(text = str_replace_all(content, replace_reg, "")) %>%
  unnest_tokens(bigram, content, token = "ngrams", n = 2)
bigrams <- bigrams %>%
  separate(bigram, into = c("first","second"), sep = " ", remove = FALSE) %>%
  anti_join(stop_words, by = c("first" = "word")) %>%
  anti_join(stop_words, by = c("second" = "word")) %>%
  filter(str_detect(first, "[a-z]") &
           str_detect(second, "[a-z]")) %>%
  filter(!str_detect(first, "[.]") &
           !str_detect(second, "[.]")) %>%
  filter(!str_detect(first, "http") & !str_detect(first, "donaldtrump") &
           !str_detect(second, "http") & !str_detect(first, "donaldtrump"))

bigrams_count <- bigrams %>%
  group_by(bigram) %>%
  count()
trump_bigrams <- bigrams_count %>%
  arrange(-n)
```

word <chr>	n <int>
trump	5059
president	2765
people	2444
donald	1819
country	1765
america	1621
time	1553
obama	1444
dont	1431
run	1144

1-10 of 10 rows

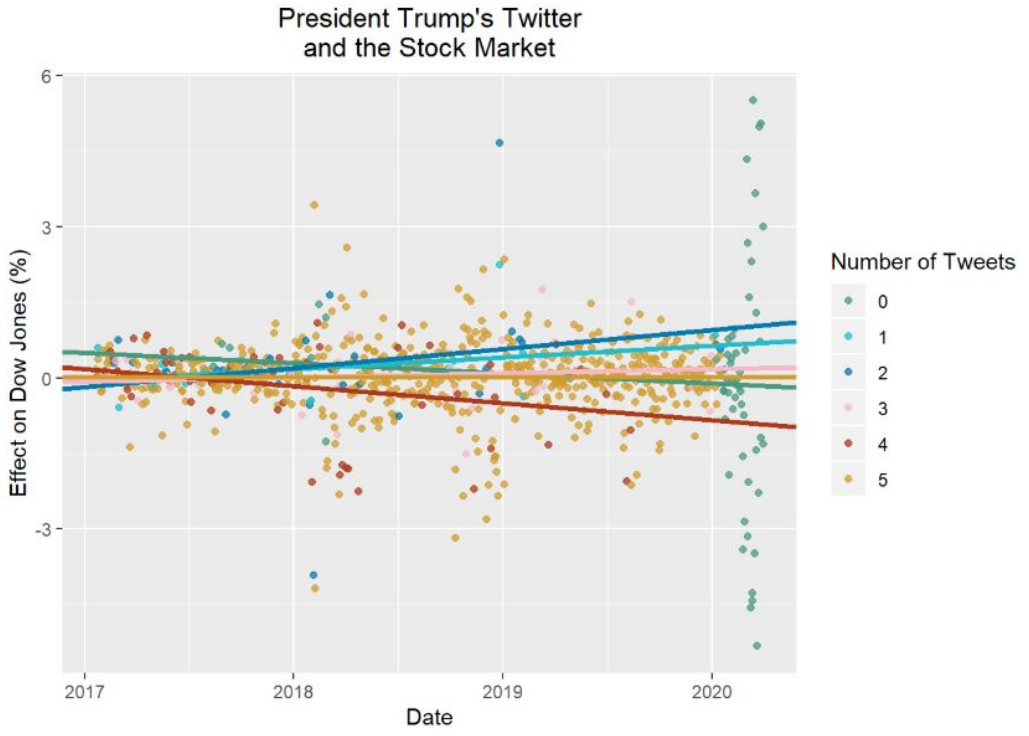
bigram <chr>	n <int>
donald trump	1526
fake news	582
white house	337
crooked hillary	329
witch hunt	319
hillary clinton	312
celebrity apprentice	258
president obama	221
trump tower	199
north korea	197

1-10 of 10 rows

Next, we also created the linear models that combined the Dow Jones Industrial Average and Trump's tweets after the data had been sanitized. The relevant R code and the subsequent plot follows.

```
mZero=lm(perchange~date,data=nZero)
summary(mZero)
confint(mZero)
mOne=lm(perchange~date,data=nOne)
summary(mOne)
confint(mOne)
mTwo=lm(perchange~date,data=nTwo)
summary(mTwo)
confint(mTwo)
mThree=lm(perchange~date,data=nThree)
summary(mThree)
confint(mThree)
mFour=lm(perchange~date,data=nFour)
summary(mFour)
confint(mFour)
mFive=lm(perchange~date,data=fiveUp)
summary(mFive)
confint(mFive)
```

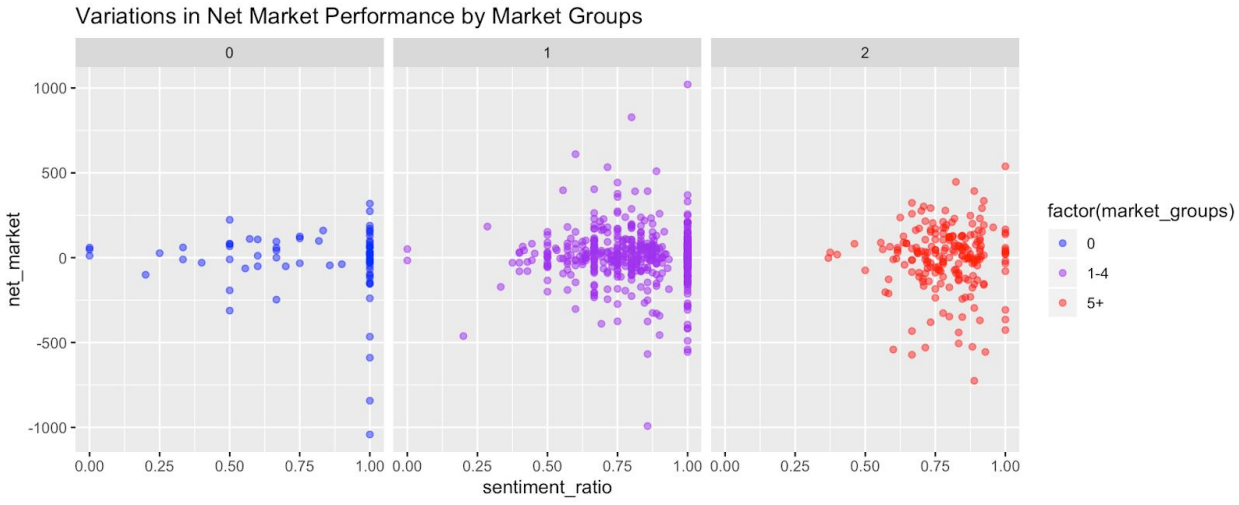
```
p <- ggplot(dplr, aes(y = perchange, x = date, color = factor(n))) +
    theme(plot.title = element_text(hjust = 0.5))+
    geom_point(alpha = 0.75)+
    geom_abline(intercept = 10.0730611, slope = -0.0005575, size =1.2, color = "#49997c")+
    geom_abline(intercept = -1.111e+01, slope = 6.437e-04, size =1.2, color = "#1ebecd")+
    geom_abline(intercept = -1.784e+01, slope = 1.029e-03, size =1.2, color = "#027ab0")+
    geom_abline(intercept = -3.9308183, slope = 0.0002251, size =1.2, color = "#F4BAC8")+
    geom_abline(intercept = 16.0250693, slope = -0.0009236, size =1.2, color = "#ae3918")+
    geom_abline(intercept = 1.803e-01, slope = -9.026e-06, size =1.2, color = "#d19c2f")+
    labs(title= "President Trump's Twitter\nand the Stock Market", y="Effect on Dow Jones (%)", x = "Date",
         color = "Number of Tweets")
```



Next, we tried to determine if plots would show a relationship between the market performance or trading volume and President Trump's tweet mentions. We get similar results for international mentions groups vs market performance and market/international mentions groups vs. volume

```r
# Market Groups --> Net Market
ggplot(DJI_text, aes(y = net_market, x = sentiment_ratio, color = factor(market_groups))) +
  geom_point(alpha = 0.5) +
  labs(title = "Variations in Net Market Performance by Market Groups") +
  scale_color_manual(labels = c("0", "1-4", "5+"), values = c("blue","purple","red")) +
  facet_wrap( ~ factor(market_groups))
# Market Groups --> Volume
ggplot(DJI_text, aes(y = Volume, x = sentiment_ratio, color = factor(market_groups))) +
  geom_point(alpha = 0.5) +
  scale_color_manual(labels = c("0", "1-4", "5+"), values = c("blue","purple","red")) +
  facet_wrap( ~ factor(market_groups))
```

```r
# Intl Groups --> Net Market
ggplot(DJI_text, aes(y = net_market, x = sentiment_ratio, color = factor(intl_groups))) +
  geom_point(alpha = 0.5) +
  scale_color_manual(values = c("blue","purple","red")) +
  facet_wrap( ~ factor(intl_groups))
# Intl Groups --> Volume
ggplot(DJI_text, aes(y = Volume, x = sentiment_ratio, color = factor(intl_groups))) +
  geom_point(alpha = 0.5) +
  scale_color_manual(values = c("blue","purple","red")) +
  facet_wrap( ~ factor(intl_groups))

# These plots all seem to suggest that there is no relationship between tweet mentions
# and market performance or trading volume
```



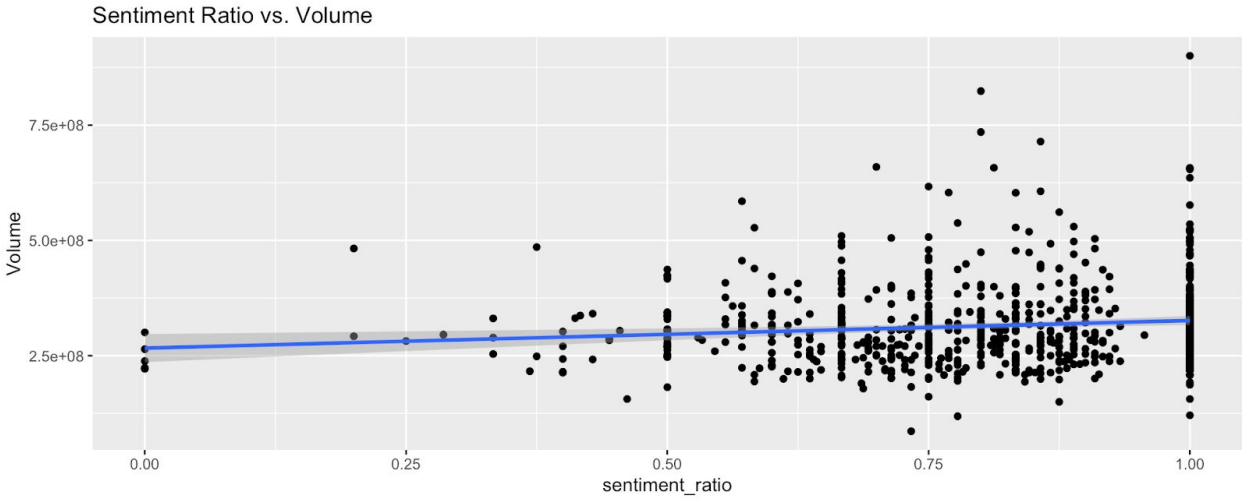
Next, we will build linear models and check their significance between the net market and the sentiment ratio with the market and international mentions from President Trump.

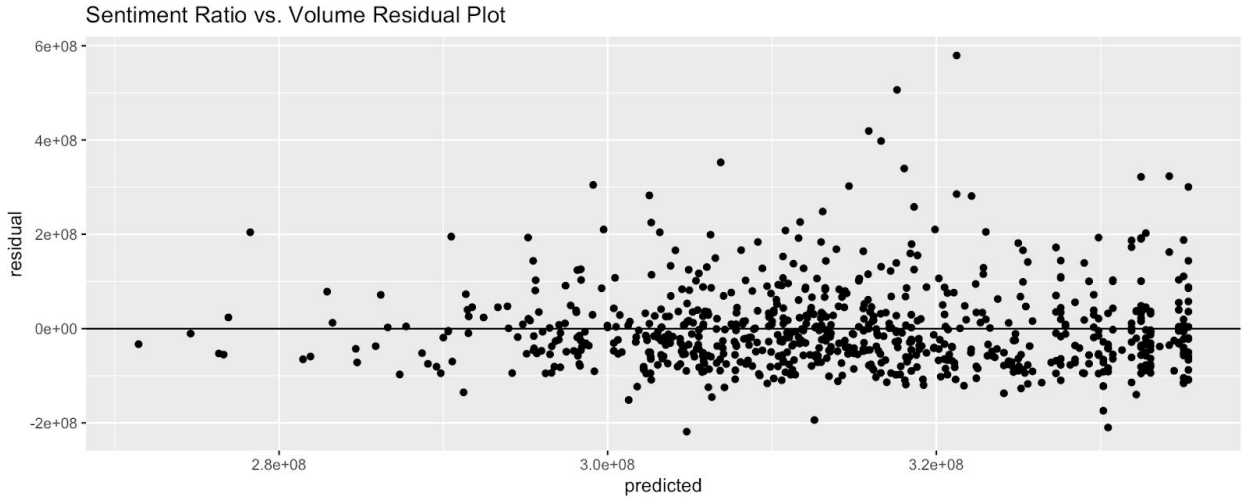
```r
# Regression
# Net Market
model1 <- lm(net_market ~ sentiment_ratio + market_mentions + intl_mentions, data = DJI_text)
summary(model1)
# None are significant at 5% level

# Volume
# Remove Outlier
DJI_text <- DJI_text[-652,]
model2 <- lm(Volume ~ sentiment_ratio + market_mentions + intl_mentions, data = DJI_text)
# market mentions, intl mentions not significant
# sentiment is significant on volume at the 5% level, let's examine this further
```

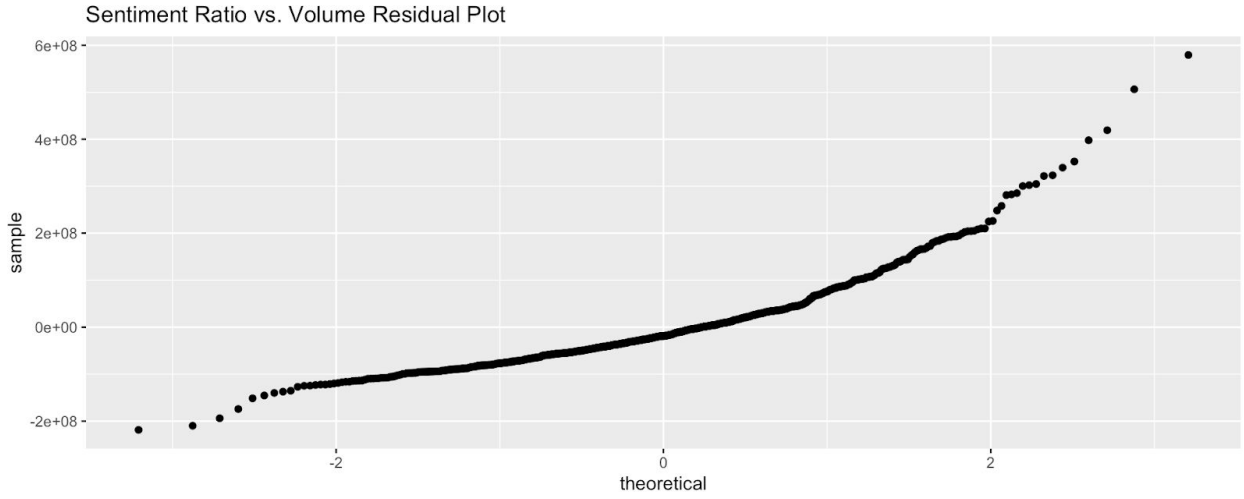
We will now build those plots using the sentiment ratio and the tweeting volume to find any potential impact of the numbers of the tweets using the following R code, and the subsequent plot.

```r
# Tweet sentiment is significant on volume at the 5% significance level
ggplot(DJI_text, aes(x = sentiment_ratio, y = Volume)) +
  geom_point() +
  labs(title = "Sentiment Ratio vs. Volume") +
  geom_smooth(method="lm")

mod2_results <- data.frame(sentiment_ratio = DJI_text$sentiment_ratio,
                           observed = DJI_text$Volume,
                           predicted = model2$fitted.values,
                           residual = model2$residuals)
model_volume <- lm(observed ~ sentiment_ratio, data = mod2_results)
summary(model_volume)
# Now signficant at the 1% level
```

```r
# residual plot
ggplot(mod2_results, aes(y = residual, x = predicted)) +
  geom_point() +
  labs(title = "Sentiment Ratio vs. Volume Residual Plot") +
  geom_hline(yintercept = 0)
# qq plot
ggplot(mod2_results, aes(sample = residual)) +
  geom_qq() +
  labs(title = "Sentiment Ratio vs. Volume Residual Plot")
```


Sentiment Ratio vs. Volume
Volume
7.5e+08
5.0e+08
2.5e+08
0.00
0.25
0.50
0.75
1.00
sentiment_ratio


Sentiment Ratio vs. Volume Residual Plot
residual
6e+08
4e+08
2e+08
0e+00
-2e+08
2.8e+08
3.0e+08
3.2e+08
predicted


Sentiment Ratio vs. Volume Residual Plot
sample
6e+08
4e+08
2e+08
0e+00
-2e+08
-2
0
2
theoretical

Data Analysis

From looking at the regression lines for the number of tweets less than the average per day (4.37), it is apparent that fewer tweets indicate a more positive slope for percent change of the Dow Jones. Regression and QQ plots were made for each trend line, and it appears all regression models, with the exception of N equal to zero, were normal. For an N equal to one and two, the 97.5% confidence interval was positive on both ends, indicating that we can say with 97.5% confidence that the market will go up when President Trump tweets one or two times. None of the other models had such significance, and R^2 increased with N. The code for the plots are below. The plot itself is above.

```
ggplot(mod_results, aes(y = residual, x = predicted)) +
    geom_point() +
    geom_hline(yintercept = 0)
ggplot(mod_results, aes(sample = residual)) +
    geom_qq()
```

```
ggplot(mod_results1, aes(y = residual, x = predicted)) +
    geom_point() +
    geom_hline(yintercept = 0)
ggplot(mod_results1, aes(sample = residual)) +
    geom_qq()
```

```
ggplot(mod_results2, aes(y = residual, x = predicted)) +
    geom_point() +
    geom_hline(yintercept = 0)
ggplot(mod_results2, aes(sample = residual)) +
    geom_qq()
```

```
ggplot(mod_results3, aes(y = residual, x = predicted)) +
    geom_point() +
    geom_hline(yintercept = 0)
ggplot(mod_results3, aes(sample = residual)) +
    geom_qq()
```

```
ggplot(mod_results4, aes(y = residual, x = predicted)) +
    geom_point() +
    geom_hline(yintercept = 0)
ggplot(mod_results4, aes(sample = residual)) +
    geom_qq()
```

```
ggplot(mod_results5, aes(y = residual, x = predicted)) +
    geom_point() +
    geom_hline(yintercept = 0)
ggplot(mod_results5, aes(sample = residual)) +
    geom_qq()
```

Our text analysis results indicate that market mention frequency and international mention frequency do not impact market performance or volume. The values for all three categories appear to be similar. However, it is important to note that sentiment ratio appears to increase variation in market performance. While sentiment ratio is not significant on market performance, sentiment ratio does appear to be significant on volume. A simple regression of sentiment ratio on observed volume shows that sentiment ratio is significant at the 1% level. The slope of this regression line is around 5.9e7, indicating that a one unit increase in sentiment ratio will increase volume by around 59 million units. Additionally, we see that the residual and qq-plot represent an approximately normal distribution, satisfying our assumptions.

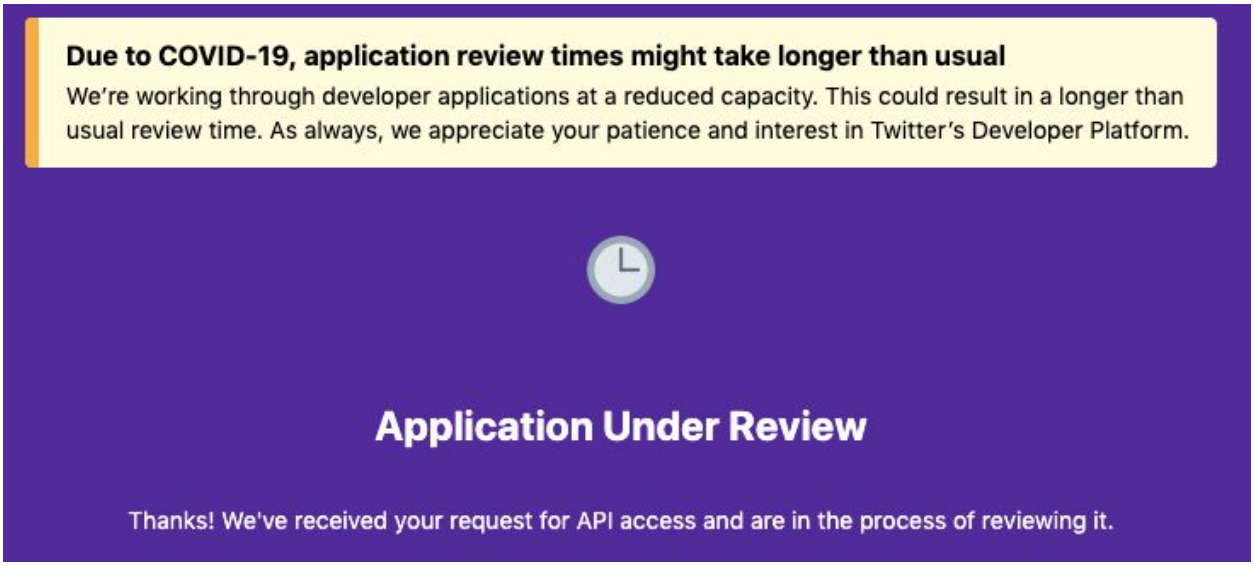
Conclusion:

When Donald Trump only tweets one to two times a day, there is a positive impact on the Dow Jones Industrial Average. When he tweets more than this, the R^2 value becomes so low that we cannot determine a trend. Since the market's value naturally increases with time due to inflation, and we can conclude that Donald Trump's tweets increase market variability, there is a negative correlation between the number of tweets and the success of the stock market on that given day.

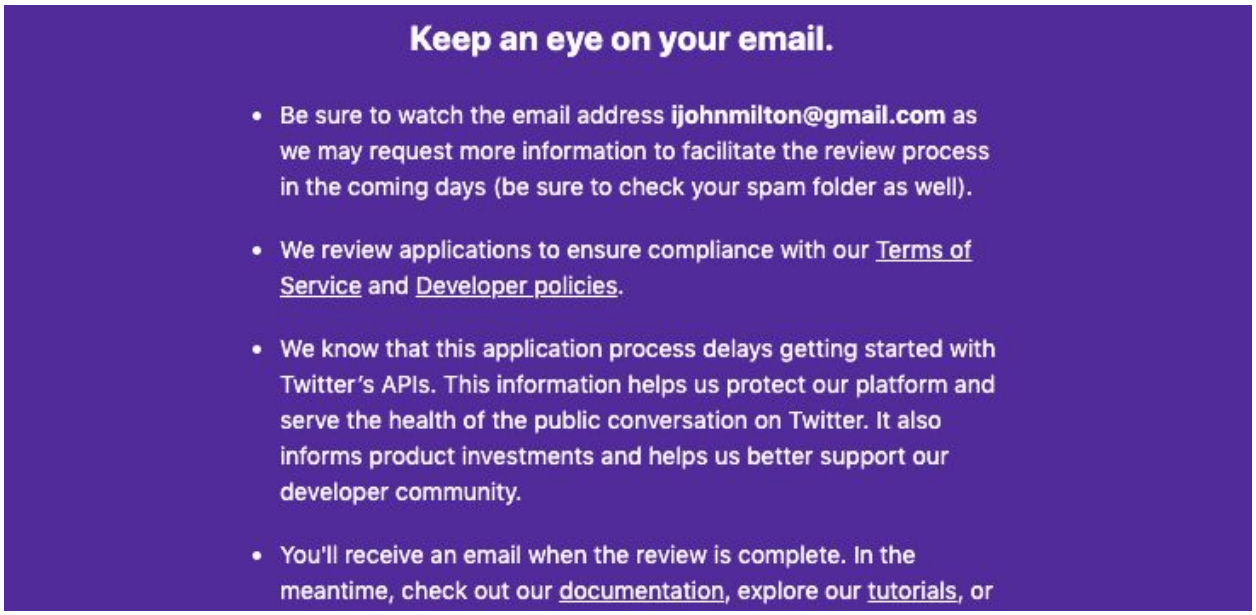
Additionally, the content of Trump's tweets appears to impact the market but only slightly. Trump's mention of the economy or international issues does not impact the stock market. However, the more

positive his tweets are in a given day, the more volume the market receives and the greater the variability in market performance.

In terms of our information, we were limited in scope by the lack of data sources on other world leaders. We began looking into former presidents like President Obama and other world leaders but were unable to find a good source of information that included all their tweets with the appropriate time stamps. We attempted to utilize a tweet scraper to solve this problem using Python, however, Twitter requires passwords that only developers are able to access in order to utilize a script that grants access to Tweets. We applied to receive developer access but received the following image from Twitter. (Image underneath next paragraph) We would also have preferred to be able to conduct this research over a longer period of time that didn't include COVID-19 because that is an unprecedented variable that could confound much of this information, however, we found that the most complete reading of the data would include this information; in conclusion on this point, we would have preferred if the COVID-19 pandemic didn't occur, but since it did, we had to use that information.



What we would have liked to do in terms of this information would be to aggregate other world leaders and compare them to both the United States stock exchanges and their own countries. This would allow us to be able to tell if there is an exaggerated effect from President Trump's tweets from others in a similar position. This required the tweet scrapper, but that was contingent on Twitter policies but that is a direction this project could go in the future. Another thing we would like to investigate would be whether President Trump only tweets on days that the market is doing well. This is very similar to our initial question of whether his tweets influence the market, but this would require much more data about stock prices at specific points of the day, but bears future investigation.



Data Sources:

R and HTML files: https://github.com/grantbowlds/Twitter-DJI
Data: https://github.com/grantbowlds/Twitter-DJI/tree/master/Data