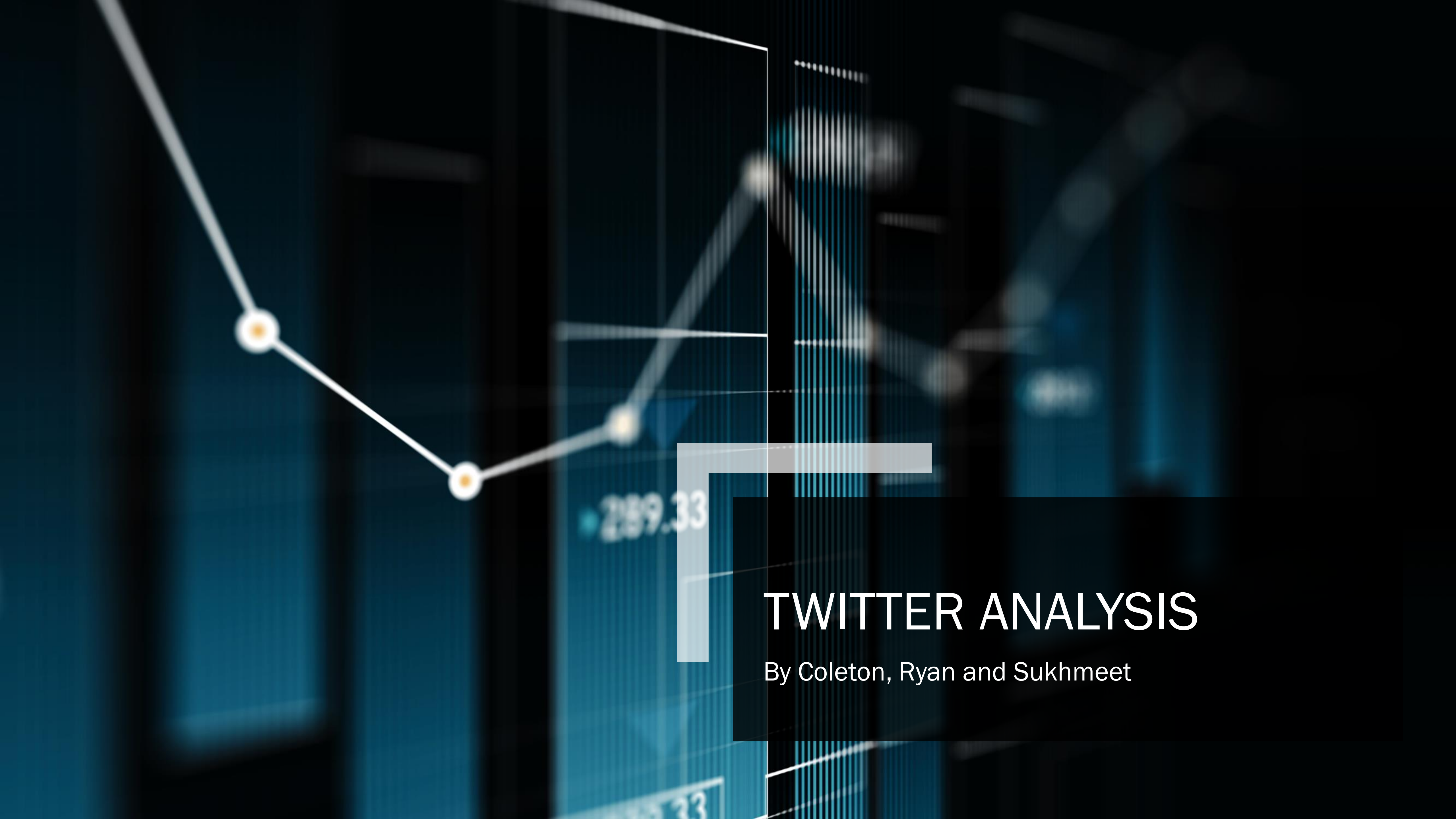


TWITTER ANALYSIS

By Coleton, Ryan and Sukhmeet

BUSINESS QUESTION

What is the impact of Twitter Activity on the Performance of Tesla Stock?

Factors Considered

Analysis of tweet content examining its association with stock performance

Focused on the post Covid-19 period to avoid extenuating circumstances

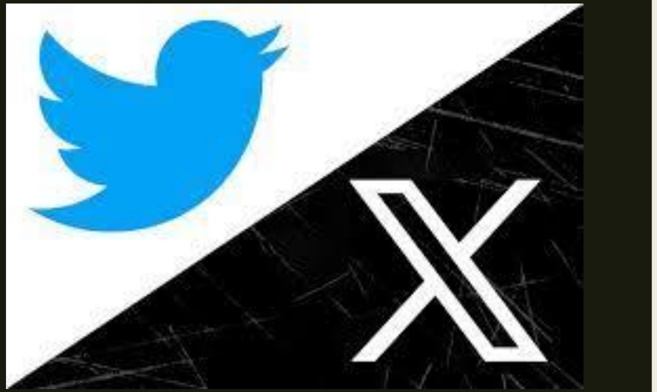
Utilization of various statistics for precise Tesla stock forecasts

Project Background

Stock market forecasting is becoming an increasingly popular field with the advent of AI and machine learning techniques

However, it is a very difficult task due to the endless variables that effect stock sentiment each day

We decided to use Twitter to explore whether the topics and sentiment of Tweets revolving around a highly volatile stock like Tesla could be used to predict future prices



LET’S DIVE INTO THE TWO DATASETS

Dataset 1 - Hashtag Tesla Tweets

- We have conducted our analysis using tweets that were posted between April 11th, 2022, and June 30th, 2022, harvested from a Kaggle Dataset
- The dataset contains five columns, each serving a distinct purpose:
  - Date & Time, Profile Picture Link, Twitter ID, Tweet Text, and Tweet Link
- The only three essential columns were Date & Time, Twitter ID, and Tweet Text
- The goal was to identify the overall sentiment of these Tweets for any given day so that we could pair this sentiment with the progression of Adj. close stock values overtime

Elon Musk
@elonmusk

the bird is freed

9:19 AM · Oct 28, 2022 · Twitter for iPhone

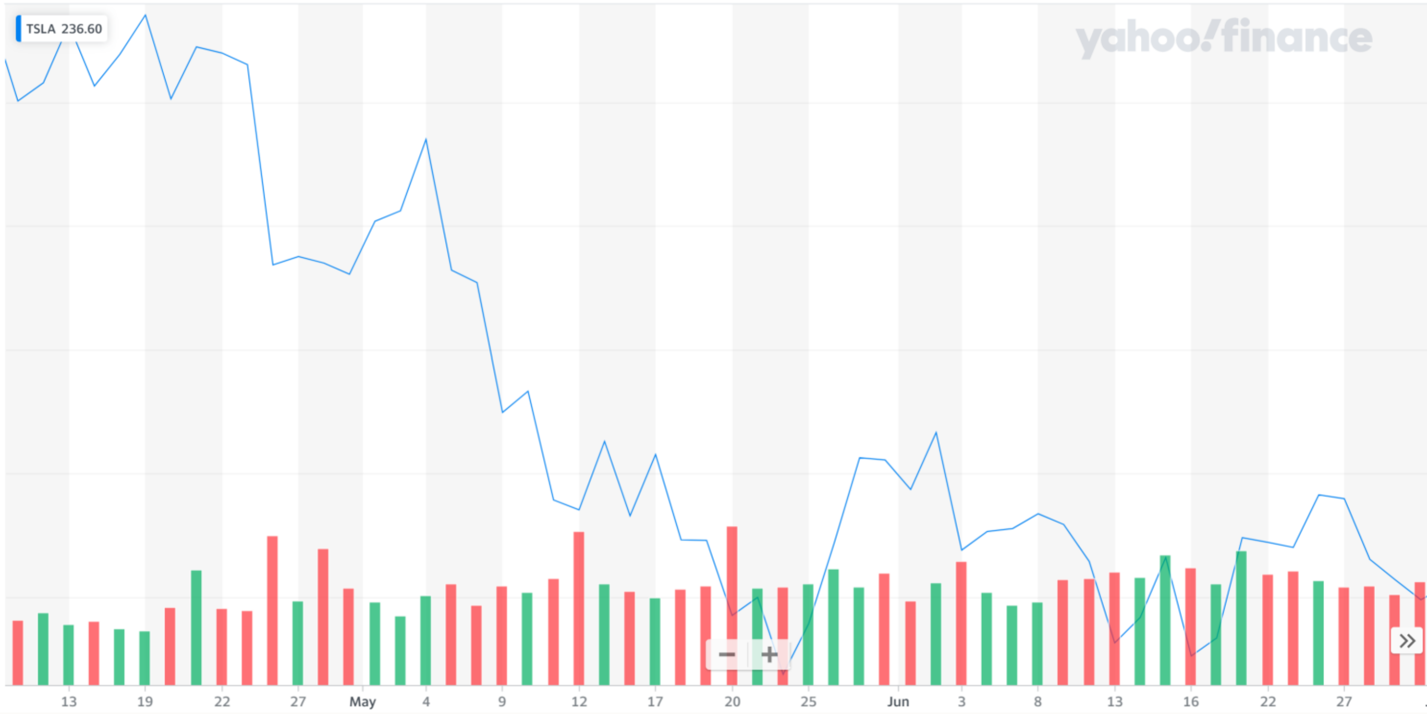
33.6K Retweets 5,301 Quote Tweets 154K Likes

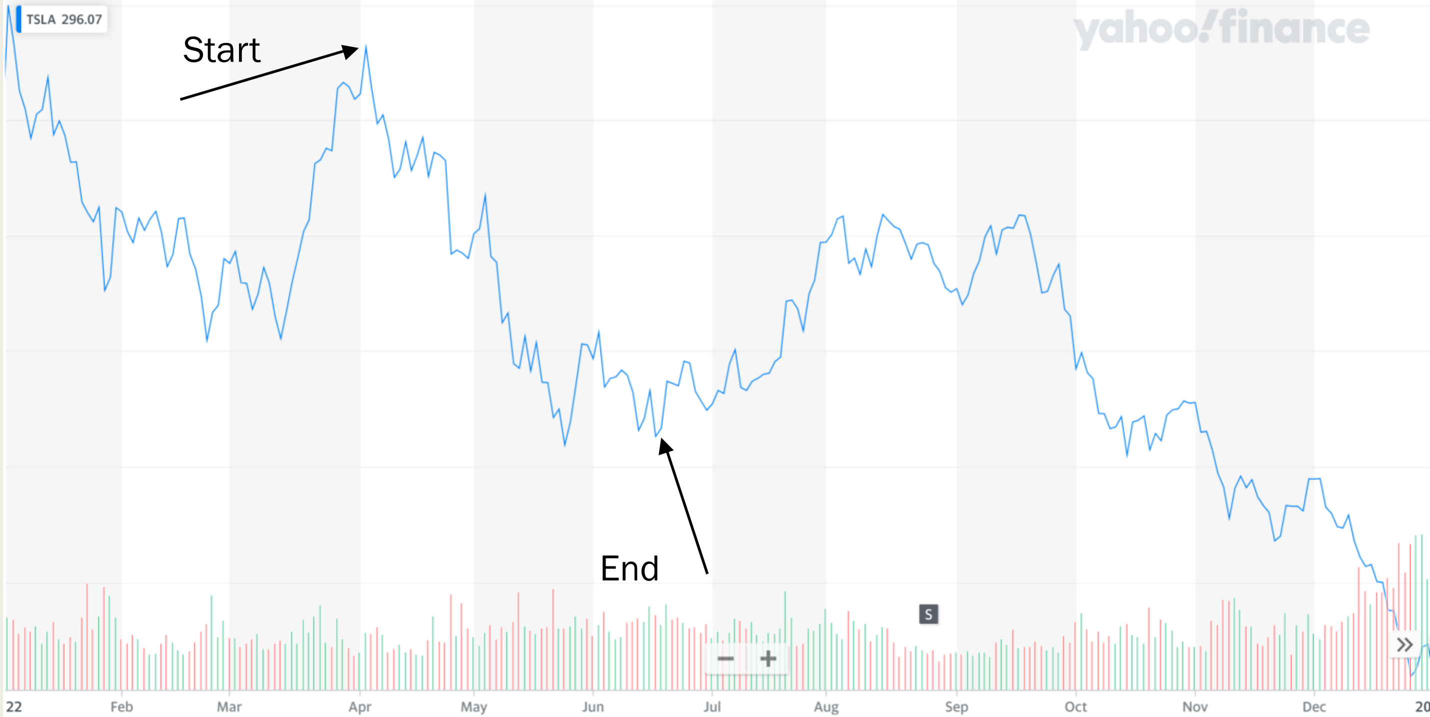
Dataset 2 - Tesla Stock (TSLA)

- We sourced our Tesla stock data from finance.yahoo.com where we obtained daily stock price information spanning from April 11th, 2022, to June 30th, 2022
- This dataset is comprised of seven columns that represent various stock statistics:
  - Date, Open, Close, Adj. Close, High, Low, and Volume
- The columns we thought were most valuable were Date, Adj. Close, and Volume
- Date is the primary/foreign key of both datasets that we used as a merging factor once all preprocessing measures were completed
- Adj. Close was used over Close since the stock market is only open from 9:30 AM to 4:00 PM while Twitter never sleeps
  - *Therefore, we needed to use Adj. Close to take after-market trading into consideration since Tweets are posted all day long… even after 4:00 PM*

Visualizing TSLA Stock Data

- Graph on top shows the timeframe of our data
  - *It can be seen that there is a general downward trend for TSLA during this time period*
- Graph on bottom shows TSLA's 2022 stock performance
  - *Our data starts at the second peak and finishes at the middle trough*
- **We anticipate there to be a negative sentiment analysis within our Twitter Data**





DATA CLEANING

Data Cleaning

- Removed the unnecessary columns from each dataset to leave us with:
  - *Hashtag Tesla Tweets: Date & Time, Twitter ID, and Tweet Text*
    - 152,000 rows/Tweets
  - *Tesla Stock: Date, Adj. Close, and Volume*
    - 150 rows/days
- Renamed column names
- Converted the “Date” columns of both datasets to datetime: (2022-04-11)
  - *Allowed to manipulate the data in a way so that the software was able to read the data as dates rather than strings*
- Removed all Twitter data after June 30th , 2022 to ensure that the dates of both datasets aligned from April 11th, 2022, to June 30th, 2022
  - *The Tweets dataframe was reduced to approximately 50,000 rows from 150,000 at this point*
- Omitted NAs from the Tweets dataframe

Data Cleaning

- Imported Textcat library to identify the languages of each tweet and appended these languages to the Tweets dataframe as a new column
- Removed all rows in the Tweets dataframe that consisted of languages other than English
  - *This reduced the number of rows to approximately 25,000 from 50,000*
- Calculated the volume of Tweets per day and appended these values as another column to the Tweets dataframe
  - *On average, there were 321 English tweets per day revolving around Tesla*
- Ran a groupby() function to group together each date's Tweets into one corpus
  - *This reduced the Tweets dataframe down to 80 observations*
  - *At this point, the Tweets and Tesla Stock dataframes had 80 and 54 rows respectively*

DATA ANALYSIS

Data Analysis

Determined the Sentiment Score and Sentiment Polarity

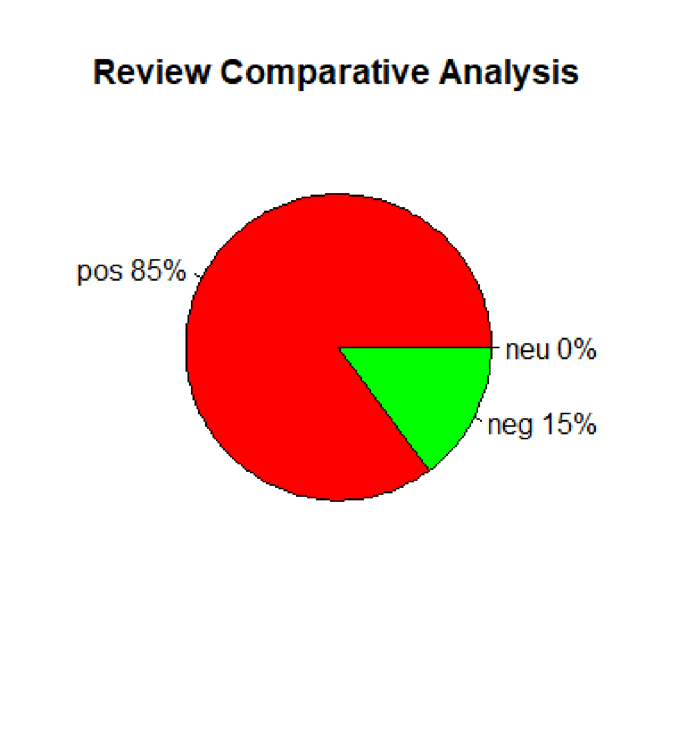
Data Processing

Performed Regression Analysis to create a model

Remember our goal: To develop a model that determines whether Twitter sentiment influences Tesla stock valuation

Sentiment Analysis

- Imported the positive and negative word libraries
- Created a function that applied a sentiment score to our newly condensed text corpuses
  - *A sentiment score is a value that reflects the magnitude of positive or negative opinions present throughout a body of text*
  - *We used this function to remove the words Tesla and TSLA from our corpuses as well since we believed they would skew our results*
  - *These values were appended to the Tweets dataframe next to their respective corpus*
- Assigned sentiment polarity to each sentiment score by appending another column:
  - *Any sentiment score less than zero was "negative*
  - *Any sentiment score equal to zero was "neutral"*
  - *Any sentiment score greater than zero was "positive"*



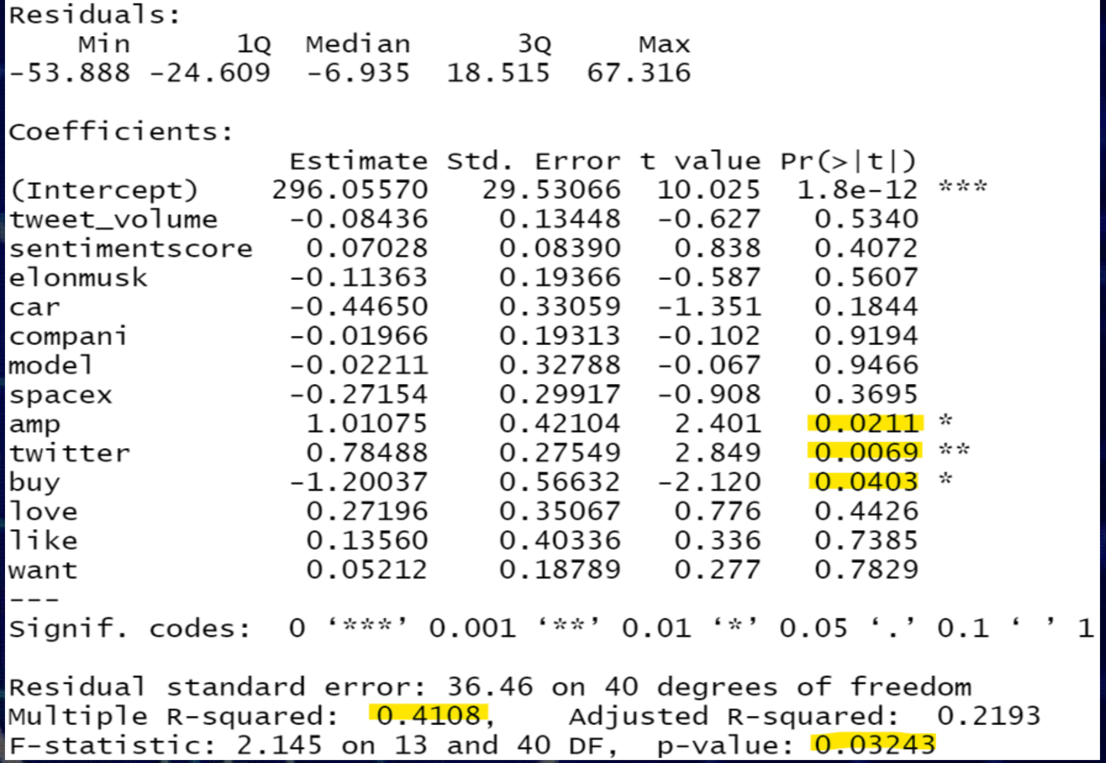
More Data Processing…

- Combined the Tweets and Tesla stock dataframes using the merge() function
  - *Needed to drop NAs again since the dataframes were of unequal length*
- This left us with one dataframe of 54 rows and various columns including: Date, Sentiment Score, Tweet Volume, Corpus, Sentiment Polarity, and Adj. Close
- Established a single, large corpus of all the daily corpuses of Tweets
- Created a Term by Document Matrix which processed the data further by removing stop words, performing stemming, removing punctuation, removing numbers, converting words to lowercase characters, etc.
- From this matrix, we were able to identify key words/topics by filtering for words that had a frequency of over 1,000 in the large corpus
- We then manually selected a list of words we wanted to test in our regression model to see if those mentions had a direct effect on stock valuation
- These key words were then merged to our overall dataframe as columns which provided the frequency of those words occurring in each daily Tweet corpus

Regression Model

Using our final dataframe, we ran an OLS regression using Adj. Close as our response variable and the following columns as our predictor variables: Tweet Volume, Sentiment Score, Elon Musk, Car, Compani, Model, SpaceX, Amp, Twitter, Buy, Love, Like, and Want.

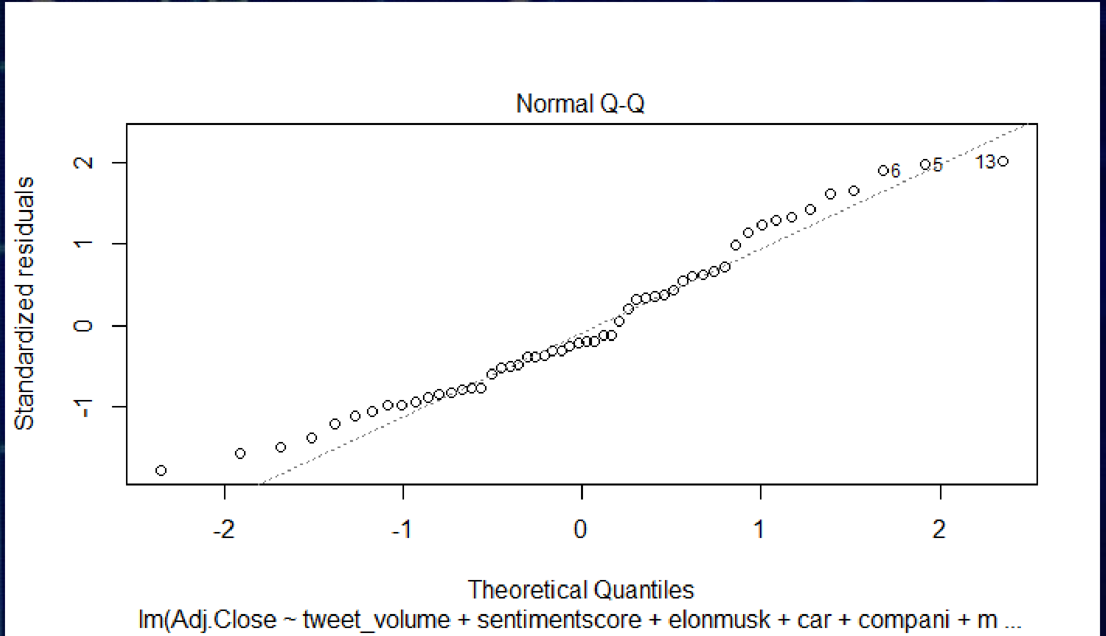
Results: As shown by the OLS summary, our model had a Multiple R-squared value of 0.4108 which means that the total variation in the predictor variables was not well explained by the response variable. Therefore, the model has a very poor goodness of fit. Furthermore, the parametric t-tests show that only three predictor variables were statistically significant predictors (SpaceX, Amp, and Twitter). This is surprising since Sentiment Score was not even a significant predictor.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-53.888 -24.609  -6.935  18.515  67.316 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   296.05570   29.53066  10.025  1.8e-12 ***
tweet_volume   -0.08436    0.13448  -0.627   0.5340    
sentimentscore  0.07028    0.08390   0.838   0.4072    
elonmusk       -0.11363    0.19366  -0.587   0.5607    
car            -0.44650    0.33059  -1.351   0.1844    
compani        -0.01966    0.19313  -0.102   0.9194    
model          -0.02211    0.32788  -0.067   0.9466    
spacex         -0.27154    0.29917  -0.908   0.3695    
amp             1.01075    0.42104   2.401   0.0211 *  
twitter         0.78488    0.27549   2.849   0.0069 ** 
buy            -1.20037    0.56632  -2.120   0.0403 *  
love            0.27196    0.35067   0.776   0.4426    
like            0.13560    0.40336   0.336   0.7385    
want            0.05212    0.18789   0.277   0.7829    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.46 on 40 degrees of freedom
Multiple R-squared:  0.4108,	Adjusted R-squared:  0.2193 
F-statistic: 2.145 on 13 and 40 DF,  p-value: 0.03243
```



Solutions and Recommendations

- Our overall conclusion after analyzing the results is that you cannot use Twitter sentiment to predict the Adj. Close values of Tesla stock
- However, our results may be misleading since we needed to trim down our data significantly since our computers could not handle the large volumes of our original datasets
- In the future, we would like to perform the regression again using the entire datasets while breaking them down into training and testing sets
- Furthermore, we would have liked to incorporate other variables such as average Tweet length, but were unable to do so due to time constraints and some hardships with data preprocessing

Summary

- Data preprocessing and sentiment analysis are two parts of our project that really went well
  - *We were able to merge, condense, and create new variables for two very large datasets*
  - *We implemented the sentiment score and sentiment polarity functions perfectly and were able to use that data in our OLS regression model*
- The biggest challenge of our group was that our computers had trouble processing such large amounts of data
  - *We had incredibly long run times and there were multiple occasions where our dplyr groupby() function did not execute properly when working perfectly fine just hours before*
  - *We were luckily able to save our working directories with our regression analysis and export smaller dataframe subsets to circumvent these hardships*
- We still believe our regression model could be greatly improved upon with more time
  - *Variables could be added, more data could be analyzed, and other types of regression could be used (ie. logistic, lasso, ridge, etc.)*
- We split our group project into two parts: Coleton and Ryan primarily worked on the code while Suk primarily worked on the presentation
  - *We communicated over Zoom and text messages*

Dataset References

Shafaghi, A. (2022, November 14). Hashtag tesla tweets (+150K tweets). Kaggle. https://www.kaggle.com/datasets/alishafaghi/hashtag-tesla-tweets

Thakur, V. (2022, July 12). Twitter dataset: Tesla. Kaggle. https://www.kaggle.com/datasets/vishesh1412/twitter-dataset-tesla

THANK YOU!