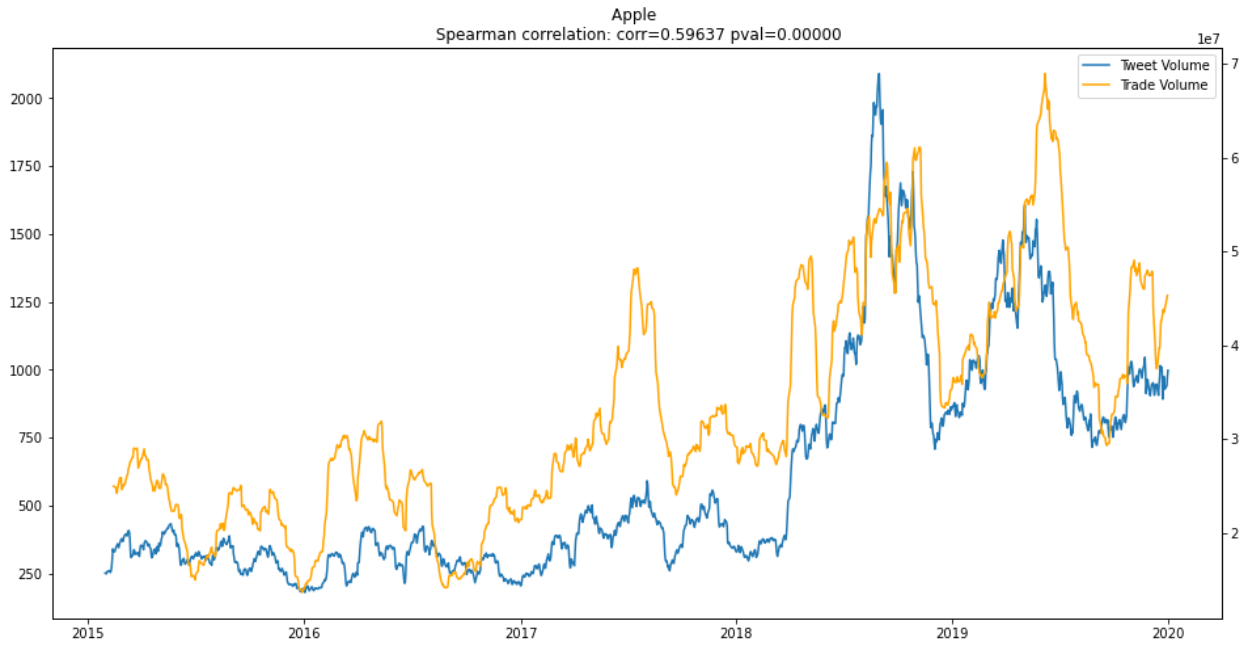
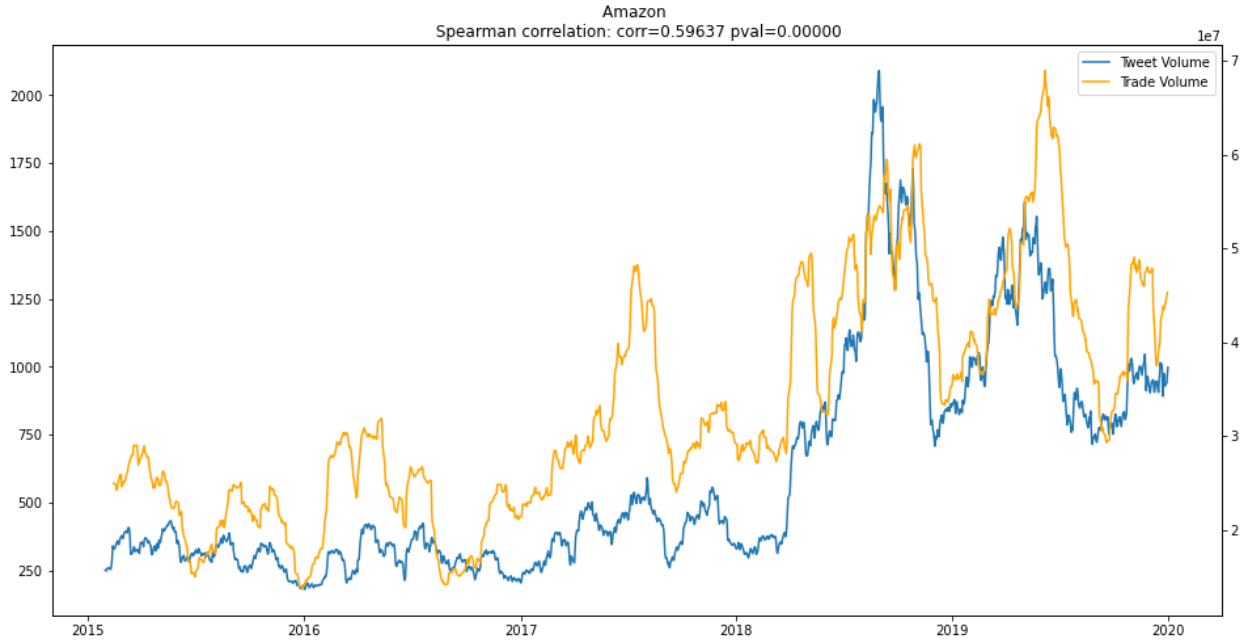
Phase I:

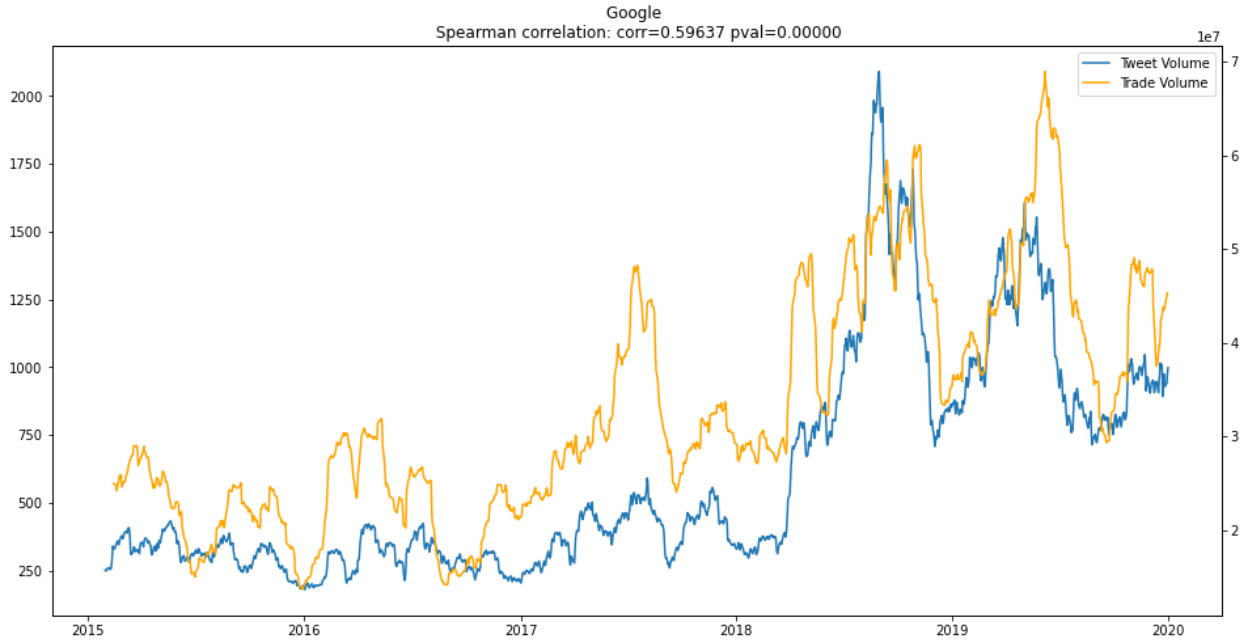
I wanted to work on a dataset that has to do with the stock market and its relation to twitter. I want to see how direct of a relationship there is between a company's stock price and tweets about that company to see if there is some sort of correlation. I have selected two separate datasets that I may combine into one dataset. The first describes several tweets about companies and the second describes stock prices over the same time period. I thought of this idea after the recent hysteria caused with Gamestop and Reddit users. I thought it would be interesting to see if there was an influence on a stock by these public sites. My Primary Data source is from Kaggle. It is the tweet dataset and I have provided a link at the end of this document to the original dataset that I was provided. The dataset is full of tweets about the top companies in the stock market. It is super cool because it has hundreds to thousands of separate tweets about big companies in the stock market. It mainly focuses on Apple, Facebook, Amazon, Tesla, and Microsoft. But this dataset ranges from 2015 to 2020 and has every piece of metadata about the tweet. This ranges from the time to the user who posted them. There is a lot of possibility with this dataset because there is a lot of correlation with my other dataset.

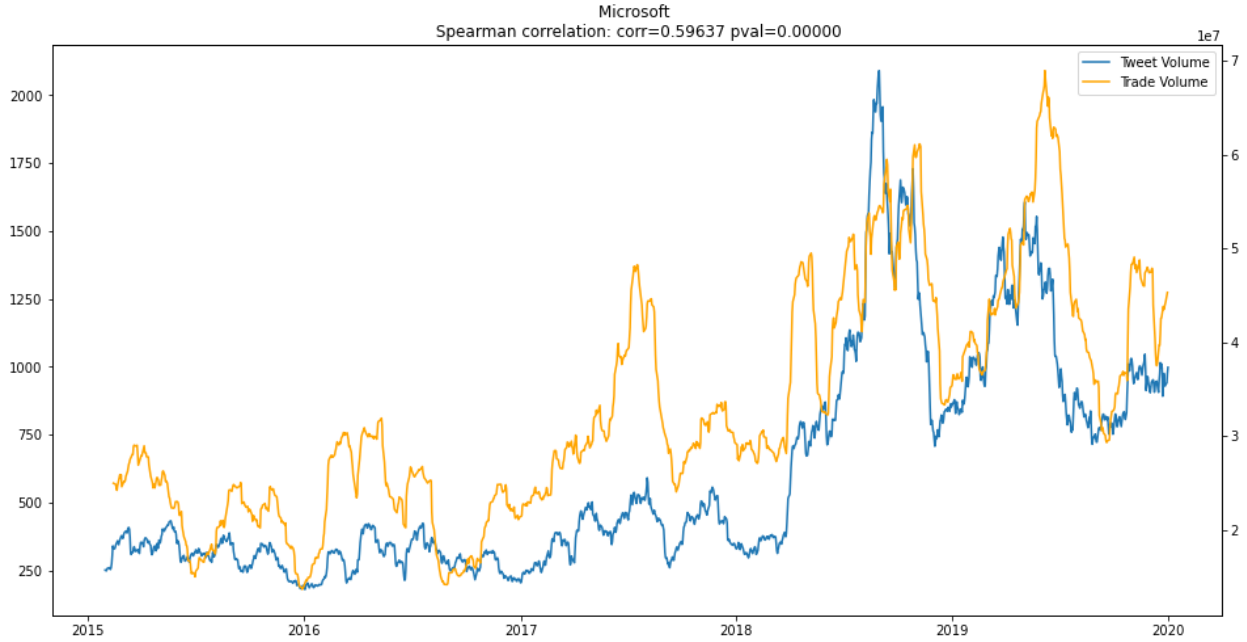
Phase II

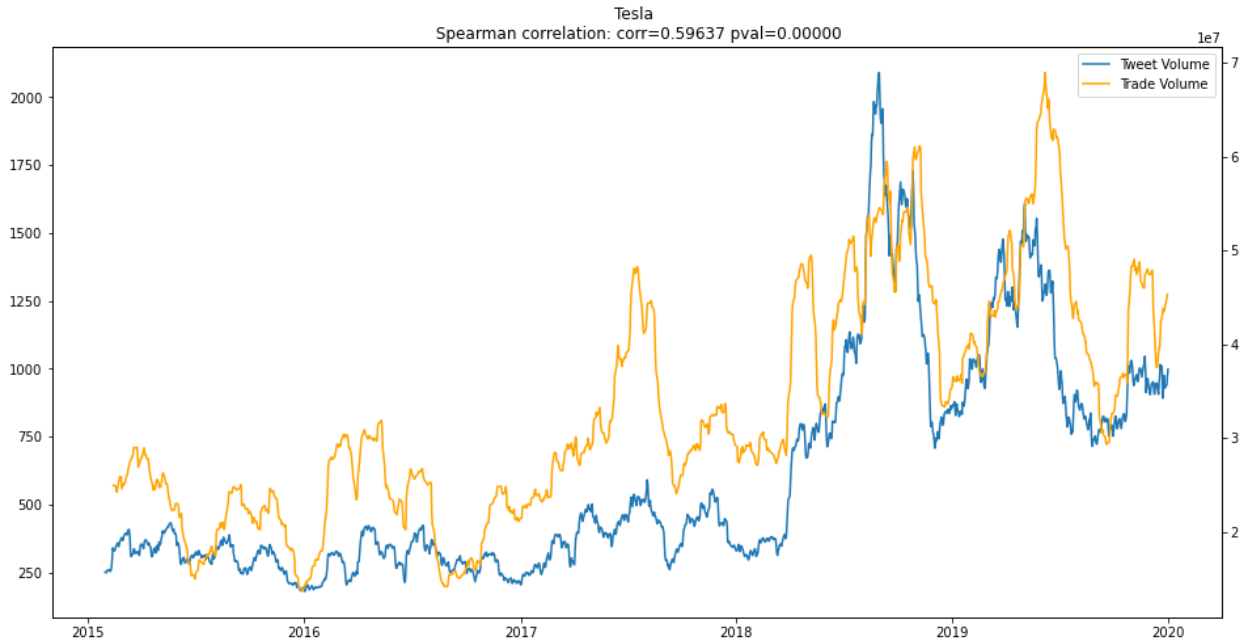
In my experiments I decided to look at what the relationship between tweet data and stock price was. To go about this process I decided to run experiments to find the relationship between 2 main variables. I started by loading the files into a csv so that I would be able to manipulate the data within them. Then I removed any rows where I had any null values for the tweet. I did that because there was only fifty thousand rows which is very small when compared to the millions of rows we already have. I decided it was best to remove them because it was incomplete data and I didn't think that it would be good to include data if we couldn't say who it came from. Then I looked at an average of 15 days on either side of the date in question to get a much smoother graph. I did that so the graph would be easier to read and so that it would be easier to tell if there was a correlation. After running the code here are my results:



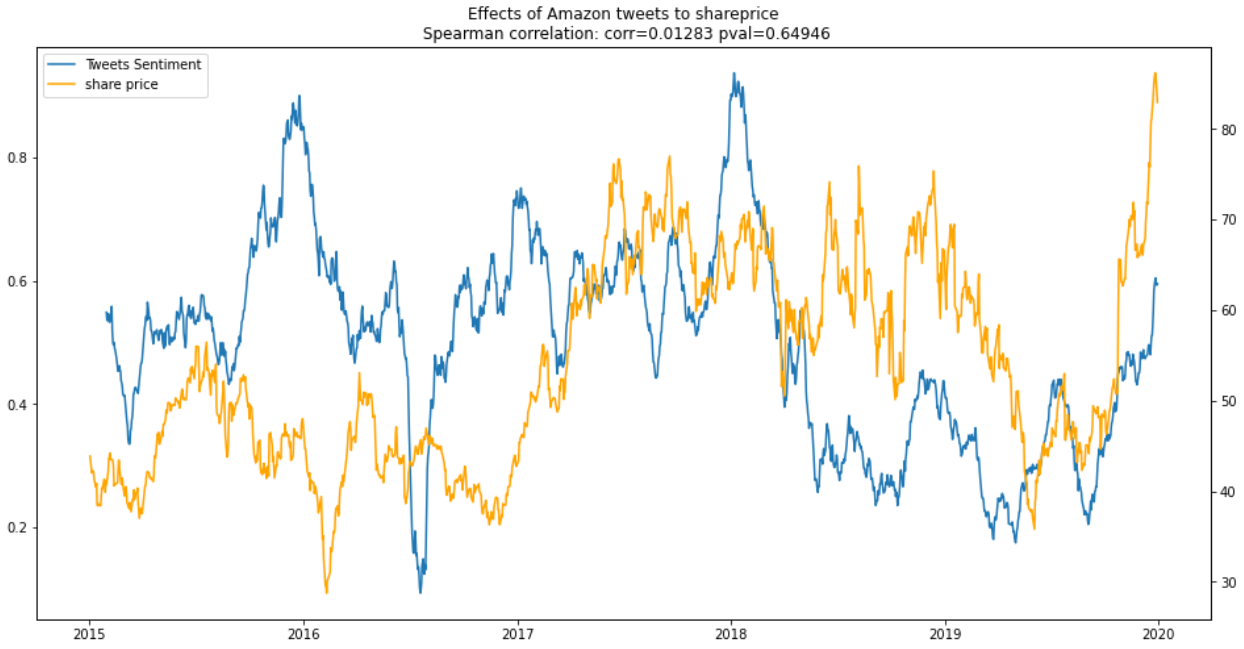

Amazon
Spearman correlation: corr=0.59637 pval=0.00000
1e7
Tweet Volume
Trade Volume
2000
1750
1500
1250
1000
750
500
250
7
6
5
4
3
2
2015
2016
2017
2018
2019
2020

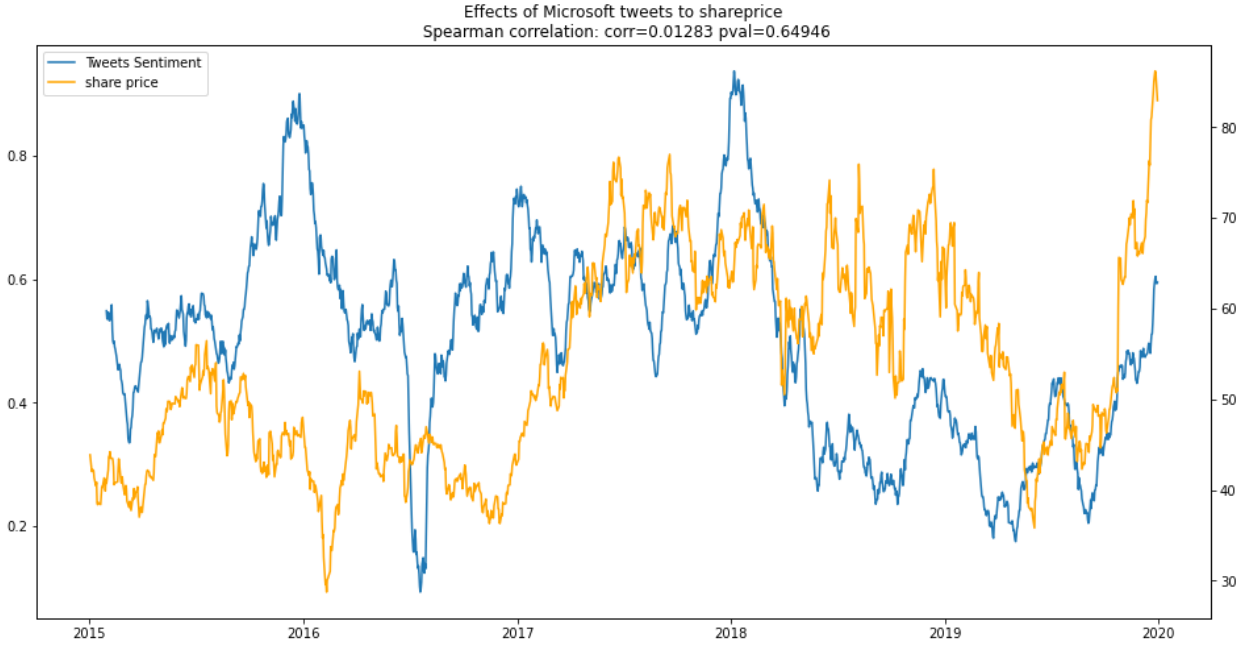

Google
Spearman correlation: corr=0.59637 pval=0.00000
1e7
Tweet Volume
Trade Volume
2000
1750
1500
1250
1000
750
500
250
7
6
5
4
3
2
2015
2016
2017
2018
2019
2020

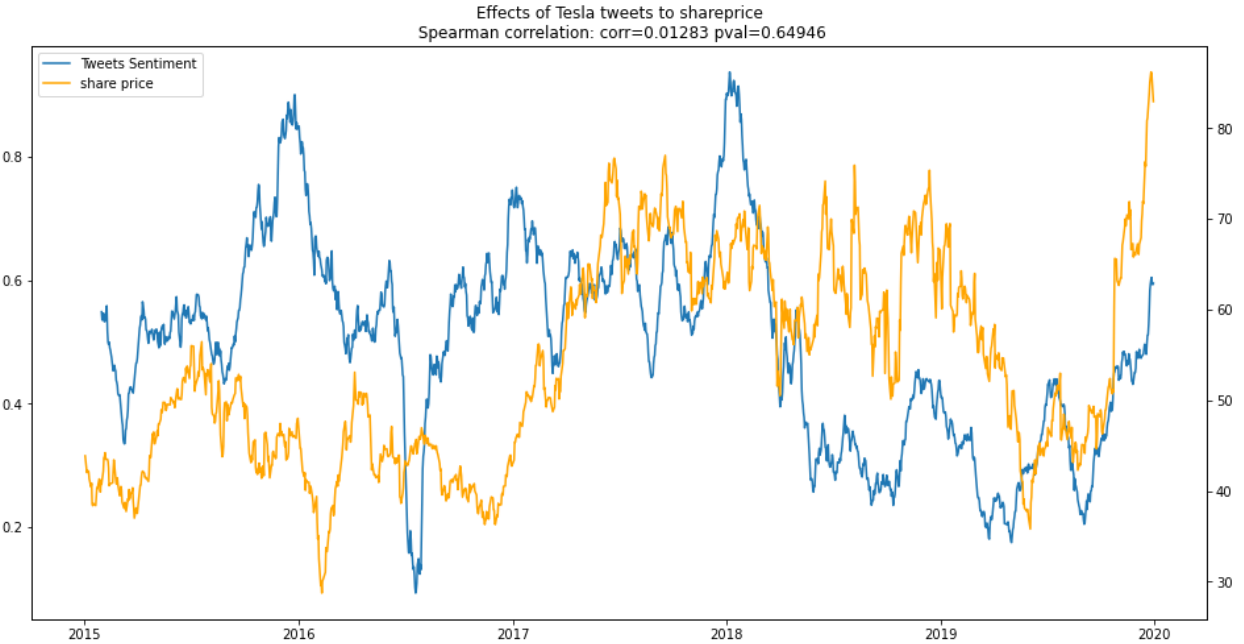

Microsoft
Spearman correlation: corr=0.59637 pval=0.00000
1e7
Tweet Volume
Trade Volume
2000
1750
1500
1250
1000
750
500
250
7
6
5
4
3
2
2015
2016
2017
2018
2019
2020

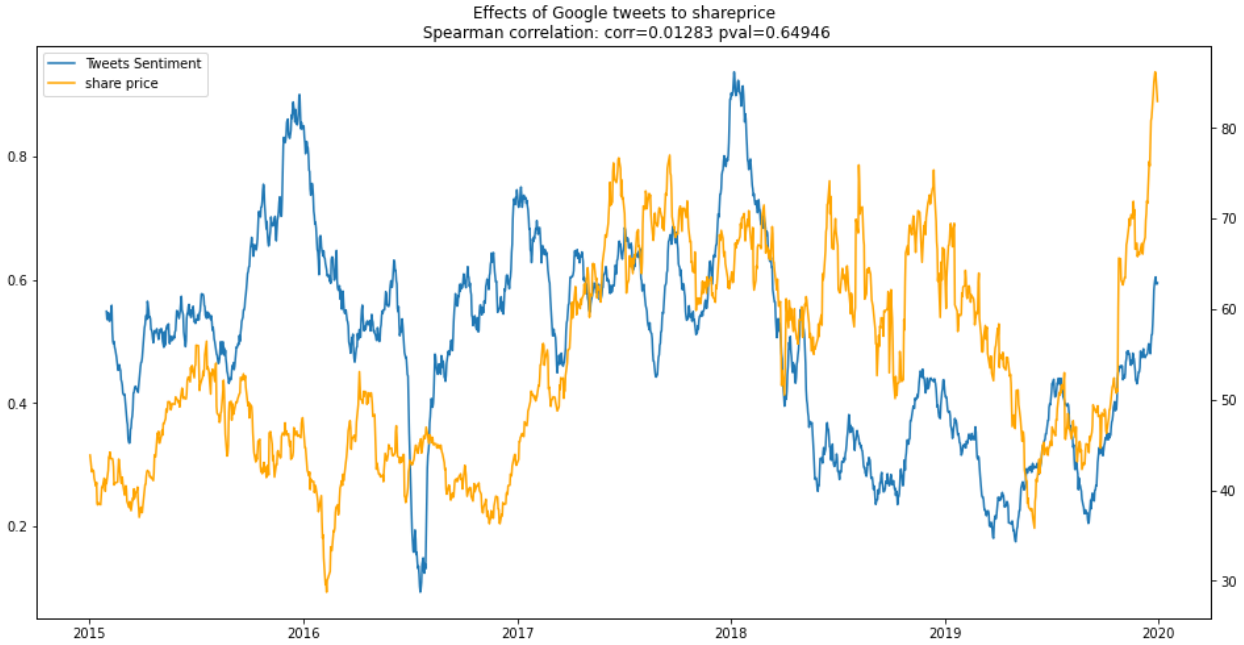

Tesla
Spearman correlation: corr=0.59637 pval=0.00000
1e7
Tweet Volume
Trade Volume
2000
1750
1500
1250
1000
750
500
250
7
6
5
4
3
2
2015
2016
2017
2018
2019
2020

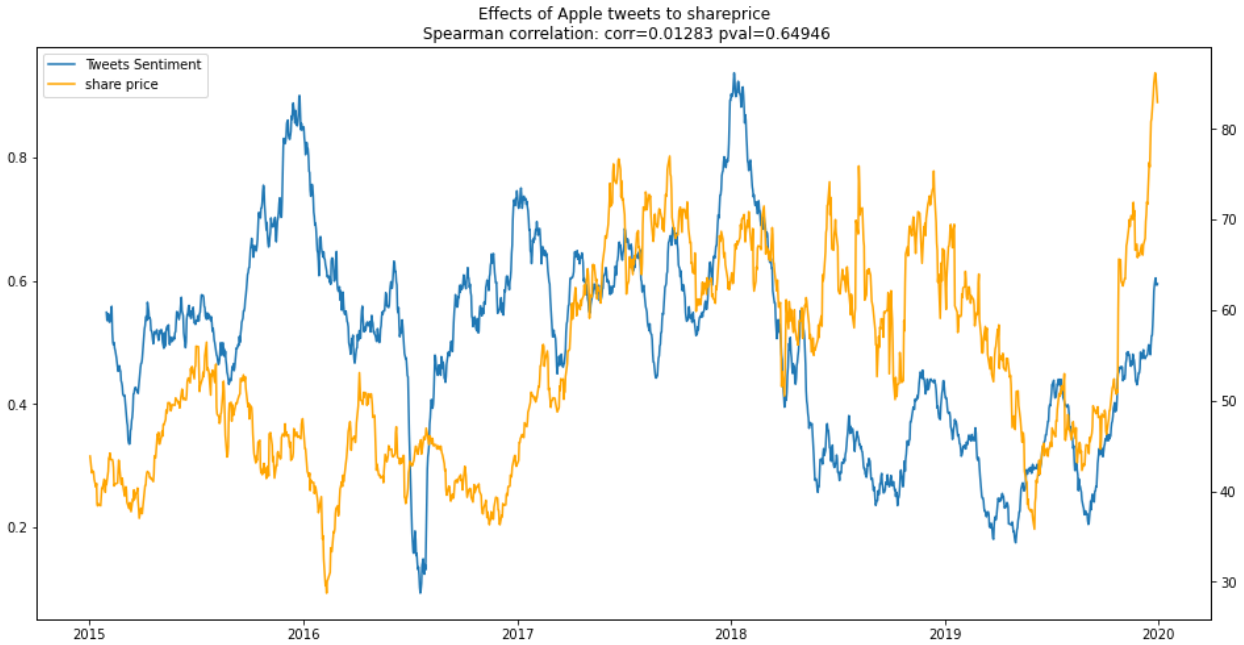
I thought this was interesting. There does seem to be some sort of relationship between the two traits volume of tweets and volume traded. I wanted to go a step further and see if what the tweet said had any effect on the price of the stock. This seemed really daunting at first because I had no idea on how to go about determining if a tweet was positive or not. I came across this tool called Afinn that actually does this for me. Afinn assigns a score to any text that is passed through it. The more negative the sentiment of the text, the more negative of a score it will receive. Likewise, the more positive the sentiment is of the text, the more positive of a score the text will receive. After I assigned a score to each tweet, I plotted the average score for a 15 day period with the average price for a 15 day period. I looked to see if there was any correlation and this is what I saw:




Effects of Microsoft tweets to shareprice
Spearman correlation: corr=0.01283 pval=0.64946
Tweets Sentiment
share price
0.8
0.6
0.4
0.2
80
70
60
50
40
30
2015
2016
2017
2018
2019
2020


Effects of Tesla tweets to shareprice
Spearman correlation: corr=0.01283 pval=0.64946
Tweets Sentiment
share price
0.8
0.6
0.4
0.2
80
70
60
50
40
30
2015
2016
2017
2018
2019
2020





I was pleased to see that there seemed to be another correlation between these categories as well. There were some problems with my dataset though. For starters, the dataset was very large. The dataset contained over four million rows of data that each had to be examined to see what the sentiment was for this tweet. This took a very long time to execute. Another issue with the dataset being so large is that it was very hard to actually upload to GitHub. I ended up using to the dataset in GitHub to link the csv files. As I already mentioned, another issue with my dataset was that there were some null values. This was easily fixed by deleting the rows that had these null values. Lastly, There was an issue with the tweet data and specifically referring to when the tweet was posted. This column was formatted in seconds since the tweet was posted and not in the form of a date. I ended up finding a

function within pandas that changes this to the date that it was posted. This was a very useful and easy solution to the problem.

Results:

For both of these categories the results seem to be that there is a correlation. My hypothesis at the beginning of this was that there would be a correlation because the price of a stock is the value that the public believes that stock to be worth. Hence, if people are saying a lot of good things about the company, the price will go up and the opposite if they are saying negative things. I believe this would be true for all categories of the stock market, however because I did not test data that covered all areas of the stock market, I cannot make that assumption. All that I can say form these findings is that for very large tech companies, their price does seem to follow this trend.

References:

Afinn:

https://pypi.org/project/afinn/

Yahoo Finance:

https://pypi.org/project/yfinance/

Tweet Dataset:

https://www.kaggle.com/omermetinn/tweets-about-the-top-companies-from-2015-to-2020