Final Project Report

Brent Alonzo

05/08/2025

Purpose

The purpose of this project is to determine whether the volume of positive sentiment expressed in tweets referencing major companies (Apple (AAPL) and Amazon (AMZN) specifically) has any measurable relationship with the companies' stock prices. The project applies sentiment analysis to a large dataset of tweets to classify them as positive, negative, or neutral.

After filtering and categorizing the tweets by company, the results are analyzed alongside historical stock price data. The goal is to determine whether increased positive sentiment corresponds with changes in stock price. Visualizations were created to show sentiment trends over time, stock price trends, and scatter plots comparing tweet volume to stock price. Also, statistical analysis was performed to check the significance of any observed correlations.

Data

The data for this analysis comes from StephanAkkerman/stock-market-tweets-data dataset (https://huggingface.co/datasets/StephanAkkerman/stock-market-tweets-data/viewer/default/train?p=9236&views%5B%5D=train). It consists of over 900,000 tweets mentioning a variety of companies' ticker symbols. The original dataset was divided into three columns: id, created_at, and text. We decided that a new dataset will be created that will only consider the companies Apple and Amazon with three columns: ticker_symbol, text, and created_at. The date of these tweets range from April 9th, 2020 to July 16th, 2020. Another source of data used was from Yahoo Finance historical data for the stock prices of Apple and Amazon from the same time period.

Techniques

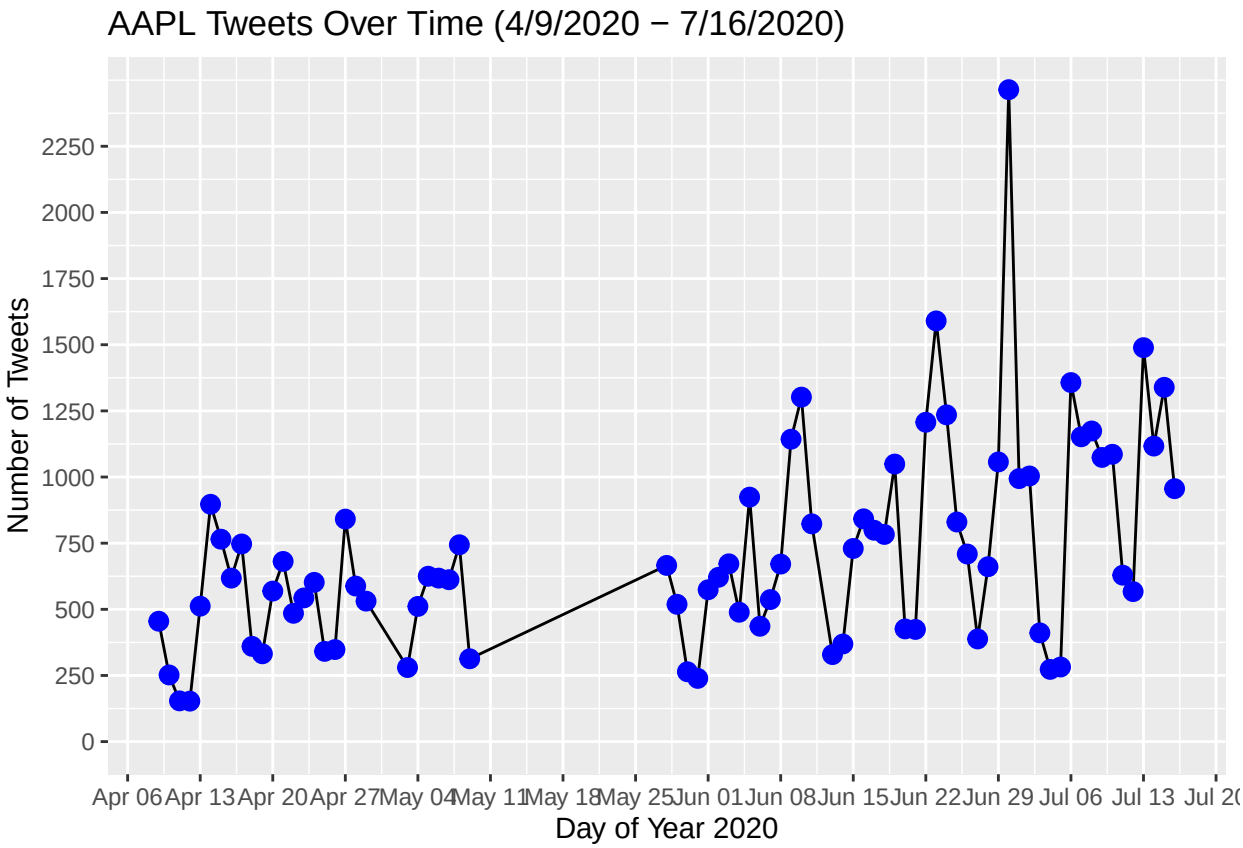
After filtering for Apple and Amazon, the dataset contained over 190,000 tweets. VADER was then applied to this subset and combined the sentiment scores with the original data, resulting in a final dataset with the columns: ticker_symbol, text, created_at, and compound.

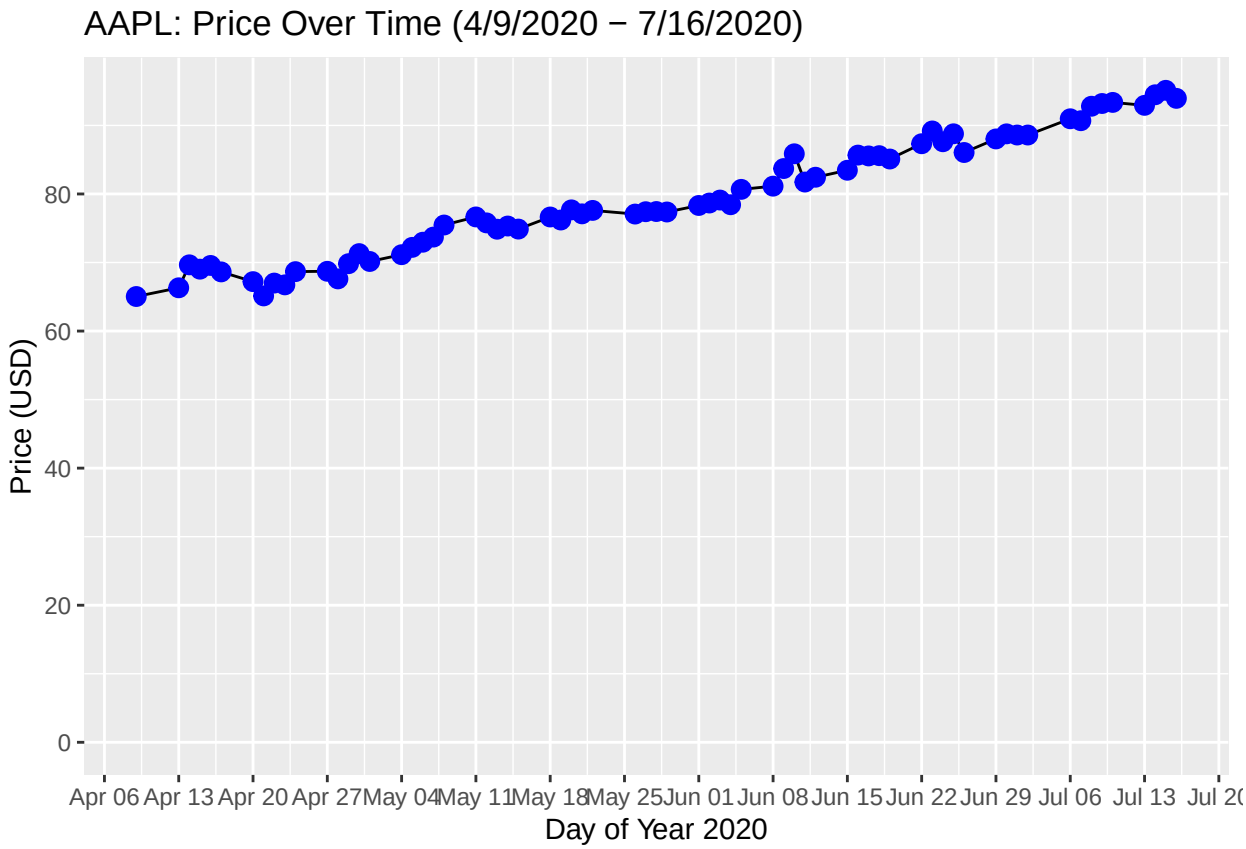
This dataset was then split into three sentiment categories: positive, negative, and neutral. Tweets with a VADER score above 0.05 were labeled positive, below -0.05 as negative, and between -0.05 and 0.05 as neutral. The threshold of 0.05/-0.05 was chosen for simplicity.

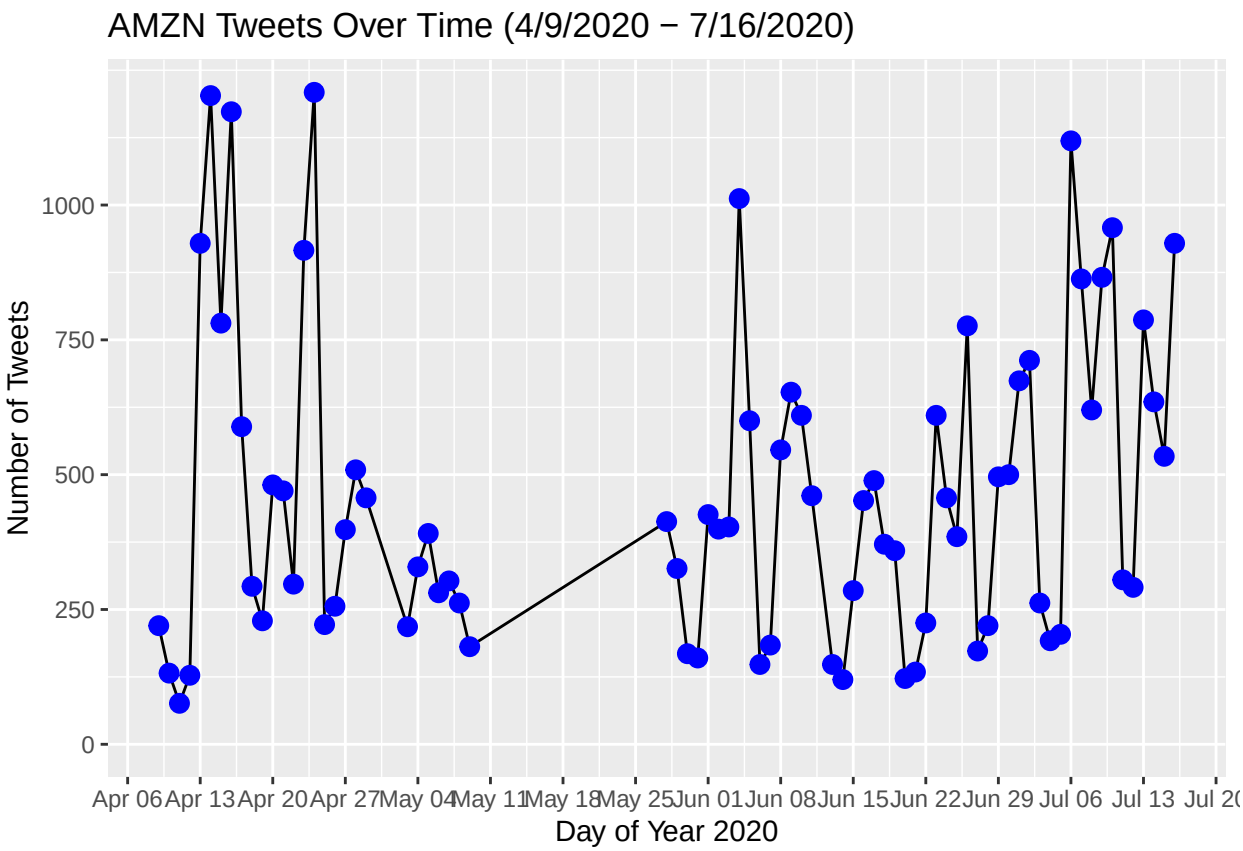
Positive sentiment data was combined with historical stock prices of Apple and Amazon to create visualizations using ggplot2. Daily tweet counts and corresponding stock closing prices were plotted. A scatterplot of tweet volume versus stock price was generated with a linear regression line to examine potential relationships. Each data point was labeled with the corresponding date, though not in chronological order.

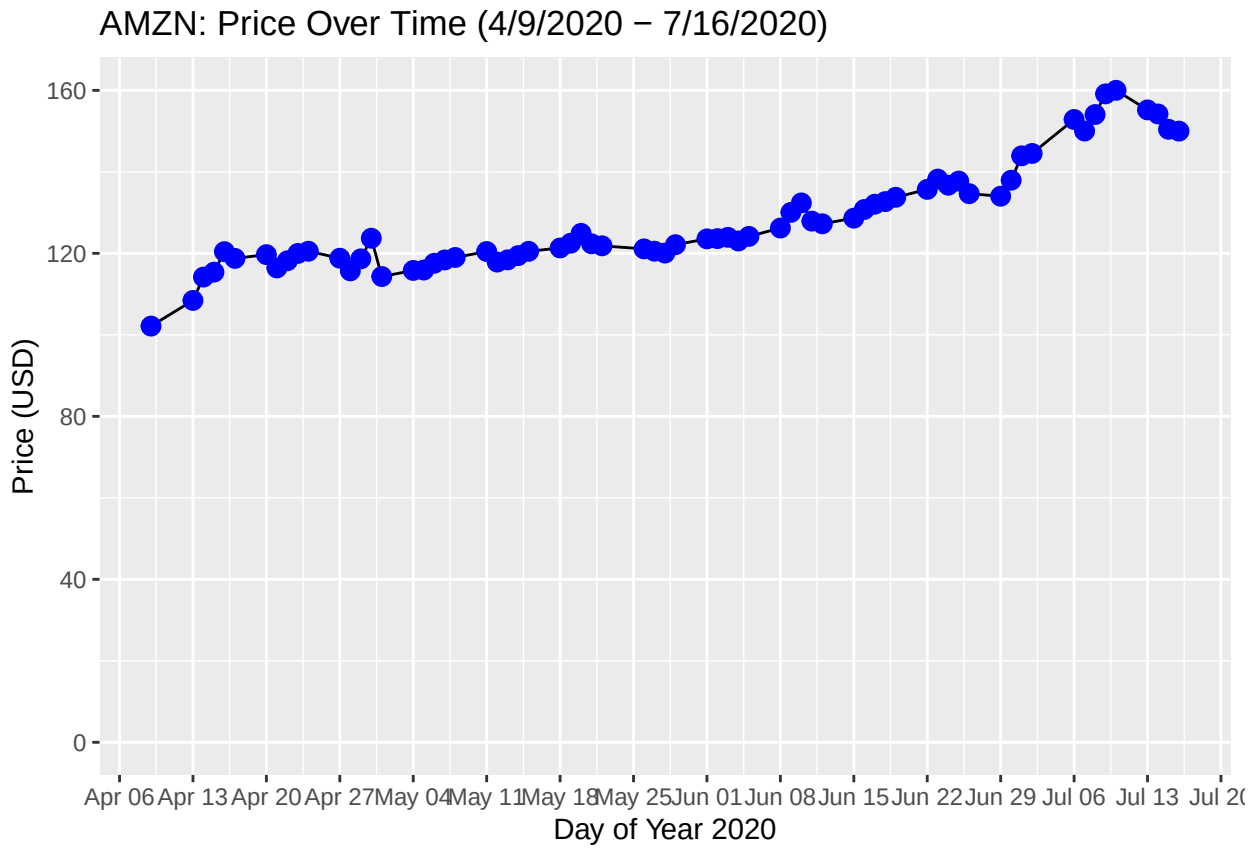
Positive, neutral, and negative sentiment data will be used to predict future Apple and Amazon stock prices using Machine Learning models.

Visualization

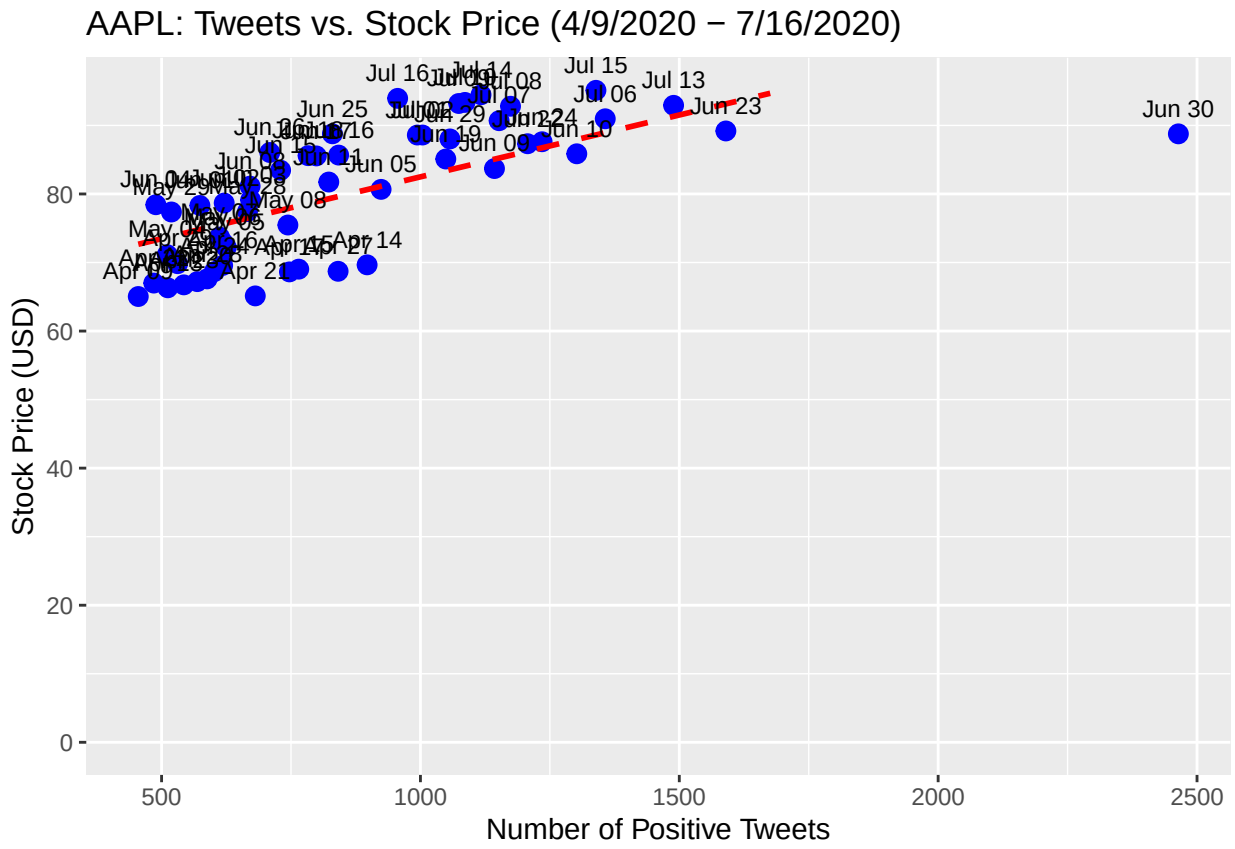



AAPL: Price Over Time (4/9/2020 − 7/16/2020)
Price (USD)
0
20
40
60
80
Apr 06
Apr 13
Apr 20
Apr 27
May 04
May 11
May 18
May 25
Jun 01
Jun 08
Jun 15
Jun 22
Jun 29
Jul 06
Jul 13
Day of Year 2020

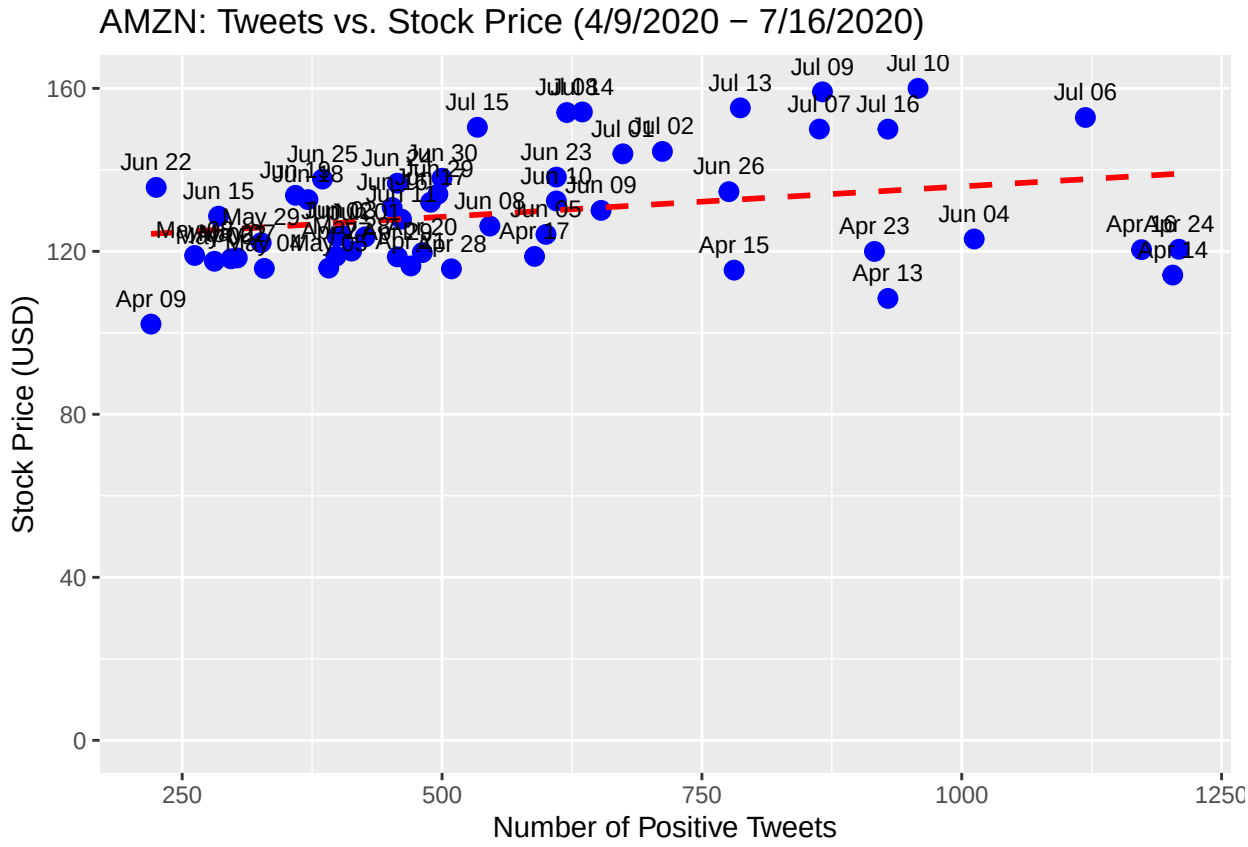

AMZN Tweets Over Time (4/9/2020 − 7/16/2020)
Number of Tweets
0
250
500
750
1000
Apr 06
Apr 13
Apr 20
Apr 27
May 04
May 11
May 18
May 25
Jun 01
Jun 08
Jun 15
Jun 22
Jun 29
Jul 06
Jul 13
Day of Year 2020




```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



Analysis

To perform descriptive statistical analysis on our data, we'll use the summary function for the Apple and Amazon scatterplots.

```
##
## Call:
## lm(formula = tweet_count ~ price, data = aapl_combined)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -338.94 -138.61  -28.19   76.38 1366.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1214.697    312.912  -3.882 0.000299 ***
## price          26.044      3.873   6.725 1.48e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 266.1 on 51 degrees of freedom
## Multiple R-squared:   0.47,  Adjusted R-squared:  0.4596
## F-statistic: 45.22 on 1 and 51 DF,  p-value: 1.481e-08

## [1] 0.6855549
```

The linear regression analysis gave an R-squared value of 0.47, meaning that about 47% of the variation in Apple's stock price can be explained by the number of positive tweets. The p-value for the tweet count coefficient was 1.48e-08, which is well below the typical significance level of 0.05. This shows that the relationship between tweet count and stock price is statistically significant. Additionally, the correlation between the two was 0.686, suggesting a moderately strong positive linear relationship.

```
##
## Call:
## lm(formula = tweet_count ~ price, data = amzn_combined)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -394.45 -164.56  -83.72   63.85  698.95
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -108.220    332.334  -0.326   0.7460
## price          5.363      2.546   2.106   0.0401 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 257.5 on 51 degrees of freedom
## Multiple R-squared:  0.08001,    Adjusted R-squared:  0.06197
## F-statistic: 4.435 on 1 and 51 DF,  p-value: 0.04014
```

```
## [1] 0.2828625
```

The linear regression analysis gave an R-squared value of 0.08, meaning that about 8% of the variation in Amazon's stock price can be explained by the number of positive tweets. The p-value for the tweet count coefficient was 0.0401, which is just below the typical 0.05 threshold. This means the relationship is statistically significant, but the effect is relatively weak. The correlation between tweet count and stock price was 0.283, suggesting a low positive linear relationship.

Conclusion

In conclusion, the impact of positive tweet volume on stock price seems to depend on the company. For Apple (AAPL), the low p-value and relatively high R-squared value of 0.47 suggest a statistically significant and moderately strong relationship. On the other hand, Amazon (AMZN) had a much lower R-squared value of 0.08 and a p-value just below 0.05, pointing to a weaker, but still statistically significant, connection. Overall, the results suggest that tweet volume might affect the stock price for some companies and that it's not consistent across different companies.