# Predicting Stock Market Movements through Social Media Sentiment Analysis: A Data-Driven Approach to Understanding Public Perception and Financial Markets

Aryan Bhardwaj & Tyler Riddick

July 1 2024

## 1 Data Summary

The data collection process for this project has consumed the majority of our efforts thus far. We began by setting up a PostgreSQL database on Railway, enabling easy access for both team members. Currently, our database comprises two main tables; however, we plan to further subdivide these tables as we gather more data (as discussed further in Section 2). Additionally, we are working on deploying virtual machines using Google Cloud's Compute Engine to handle the extensive task of scraping historical post data and automating the daily scraping of current post data. This process has proven to be more challenging than initially anticipated.

Our data comprises two main components: stock market information and social media content. Financial data is readily accessible and has been collected from sources such as Yahoo Finance and CoinMarketCap.

Obtaining social media data has been more challenging. Social APIs either require lengthy approval processes (as with Reddit) or are expensive (as with Twitter). Consequently, we opted to build our own web scrapers. We are using the Selenium package in Python to extract data from Twitter and Reddit, and utilizing the psycopg2 and SQLAlchemy packages to establish connections between our local machines and our PostgreSQL database hosted on Railway.

## 2 Data Design

Currently, our database consists of two main areas: social media data and stock market data. We are connecting these two areas via the date. We can collect stock market data for our selected companies for any date we wish, and we are also scraping the date and time each social media post was originally published. Connecting these two tables by the date is the most straightforward option when considering our final goal which is to use social media post sentiment to predict future stock market movement. An entity relationship diagram showing our database can be found in Figure 1.

Our social media data is going into our posts table, which consists of the following columns: "platform", which is the platform the post was published on (Twitter, Reddit, etc.); "topic", which is the name of the company; "post", which is the actual text of the post excluding any media or non-character content such as emojis; "likes", which is the number of likes, up-votes, or hearts given to the post; and "date_posted", which is the date the post was made and is also a foreign key referring to the stock movement tables. This table also contains a "post_id" column, a serial primary key column to distinguish each post.

Our stock market data is going into several tables named for the stock market title of the organization the data refers to, though all of them have the same structure. The table "gme" is GameStop, "wing" is Wingstop, "nvda" is Nvidia, and "dogecoin" is Dogecoin. These tables have the following columns: "date", which is just the date of the observation and is also the primary key of each table; "open", which is the stock price when the day began; "high" and "low", which are the highest and lowest stock prices seen for the company on that day, respectively; "close", which is the stock price when the day ended; "adj_close", which is the closing price after accounting for any corporate actions;


Figure 1: An ERD diagram displaying our current database structure.

and "volume", which is the number of shares traded that day. The "dogecoin" table has a different column called "market_cap" which is aggregate market value.

# 3 Sample EDAs

**Stock Prices Over Time:**

These visuals shows the progression of stock prices over a specific period. The data used here is purely financial, focusing on the Date and stock price columns. It helps to understand the overall trend and volatility of the stock over time, providing a foundational view of its performance. These line graphs can be found in Figures 2, 3, 4, & 5.

**Likes vs. Stock Price:**

These scatter plots illustrate the relationship between the number of likes on tweets mentioning the company and the corresponding stock prices. The data was joined on the Date column to match the financial data with the social media data. It helps to identify any potential correlation between public engagement on social media and stock price movements. Our social media data so far is limited to only Gamestop, though we will collect data for the other organizations as the capstone progresses. Figure 6 shows the closing price in US dollars vs the average number of likes per post, and Figure 7 shows average number of likes per post over time as a scatter plot, with each point colored and sized according to GameStop's closing price that day. The latter graph shows a clear correlation between closing cost and time, though there seems to be no clear correlation between average number of likes per post and closing cost.

**Heatmap of Open, Close, and Volume:**

These heatmaps visualize the mean trading volume across different price ranges for a stock. It uses financial data columns for Open, Close, and Volume to illustrate how trading volume correlates with price movements. This visual representation provides insights into market activity patterns and the relationship between trading volume and price levels. These heatmaps can be found in Figures 8 thorugh 11.

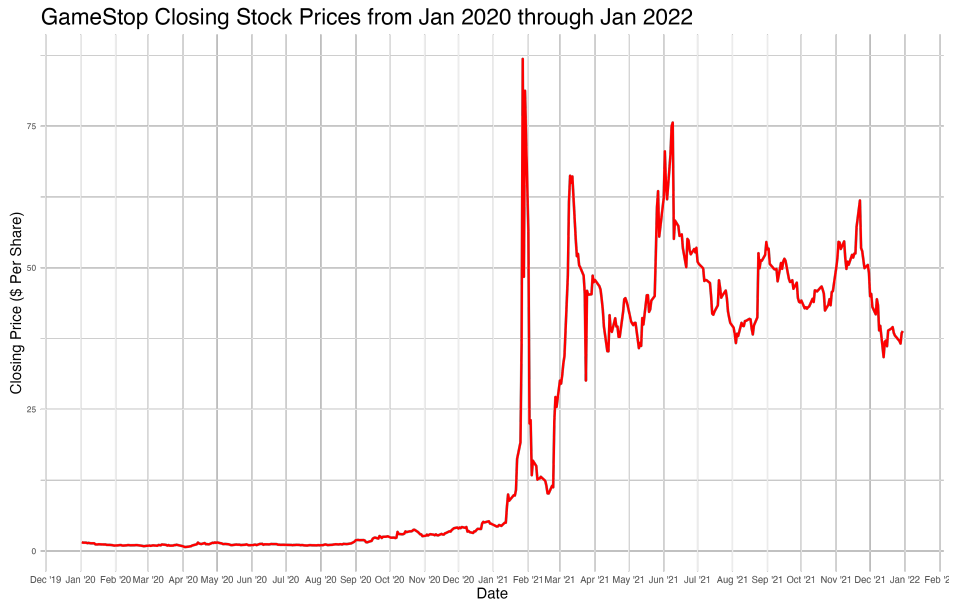


Figure 2: Closing price over time for GME (GameStop).

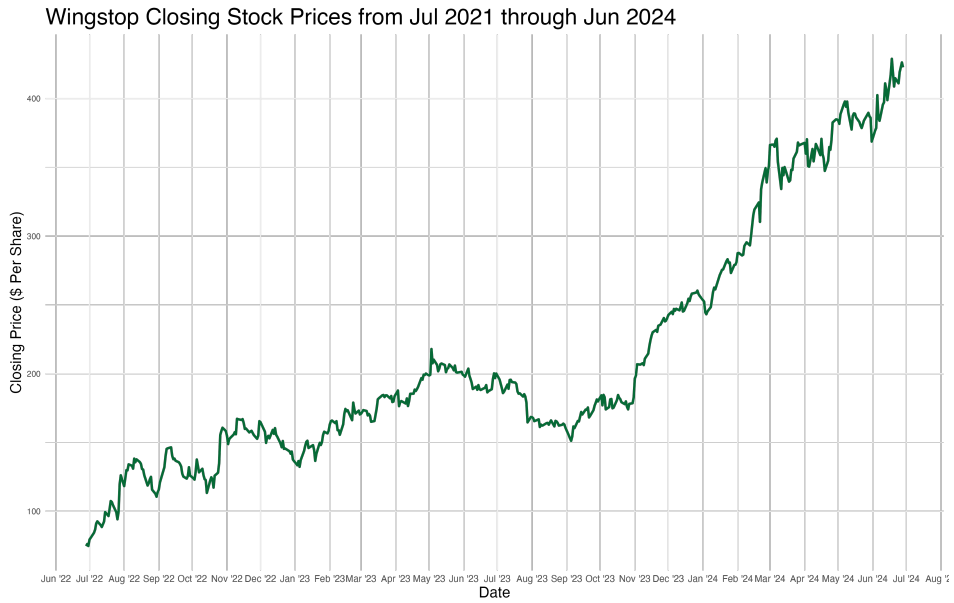


Figure 3: Closing price over time for WING (Wingstop).

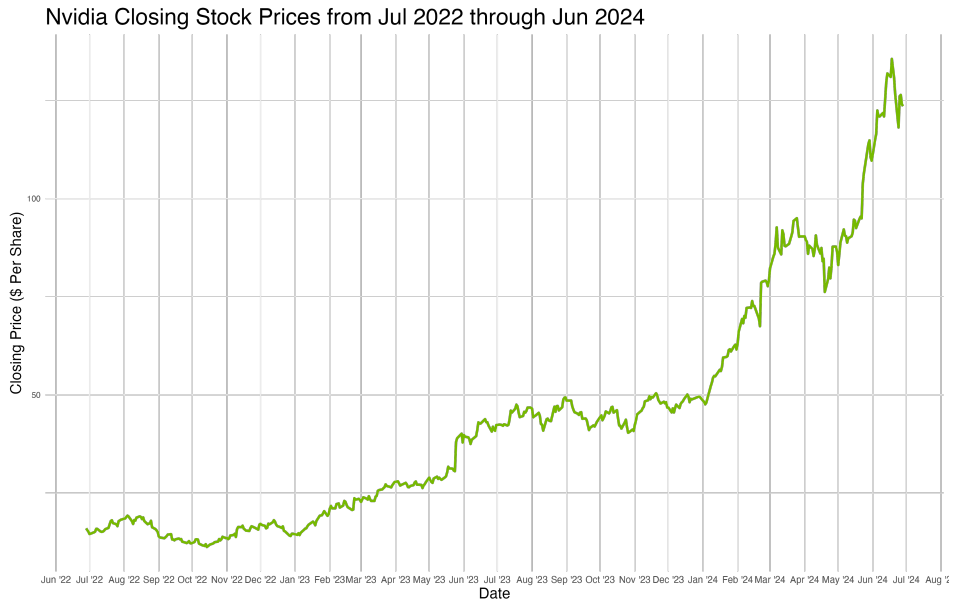


Figure 4: Closing price over time for NVDA (Nvidia).


Figure 5: Closing price over time for Dogecoin.

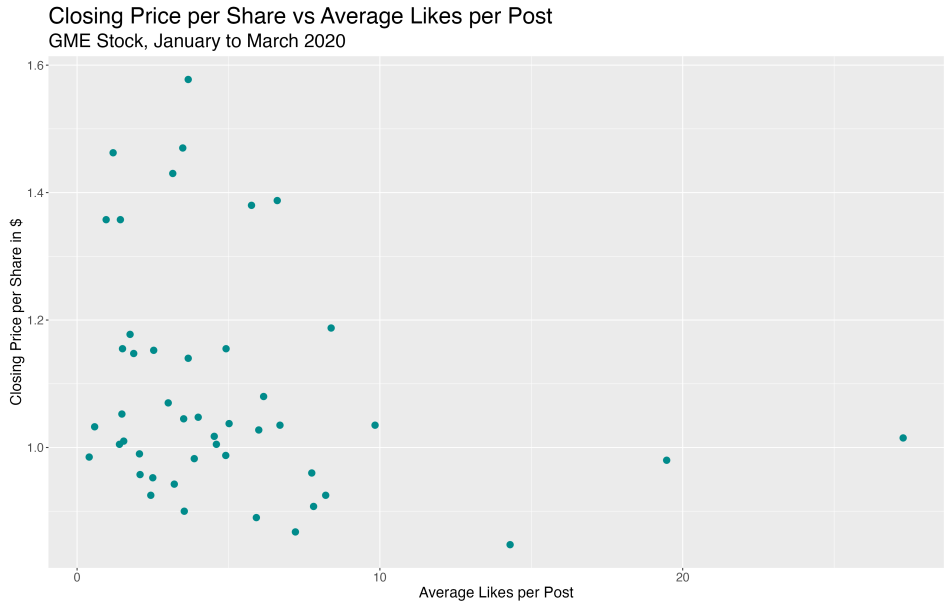


Figure 6: Closing price in USD vs average number of likes per post.

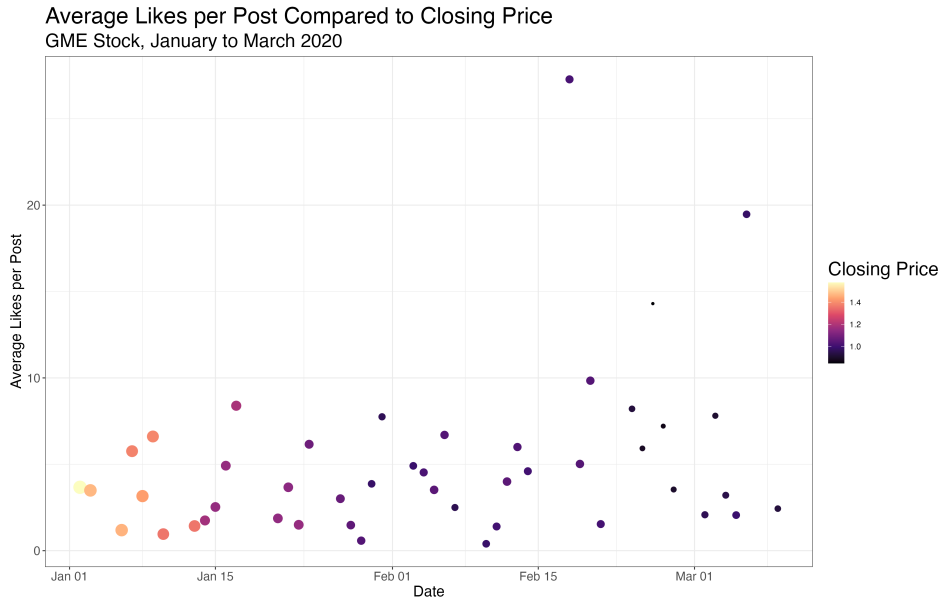


Figure 7: Average number of likes per post over time. Point color and size represent the closing cost in USD.

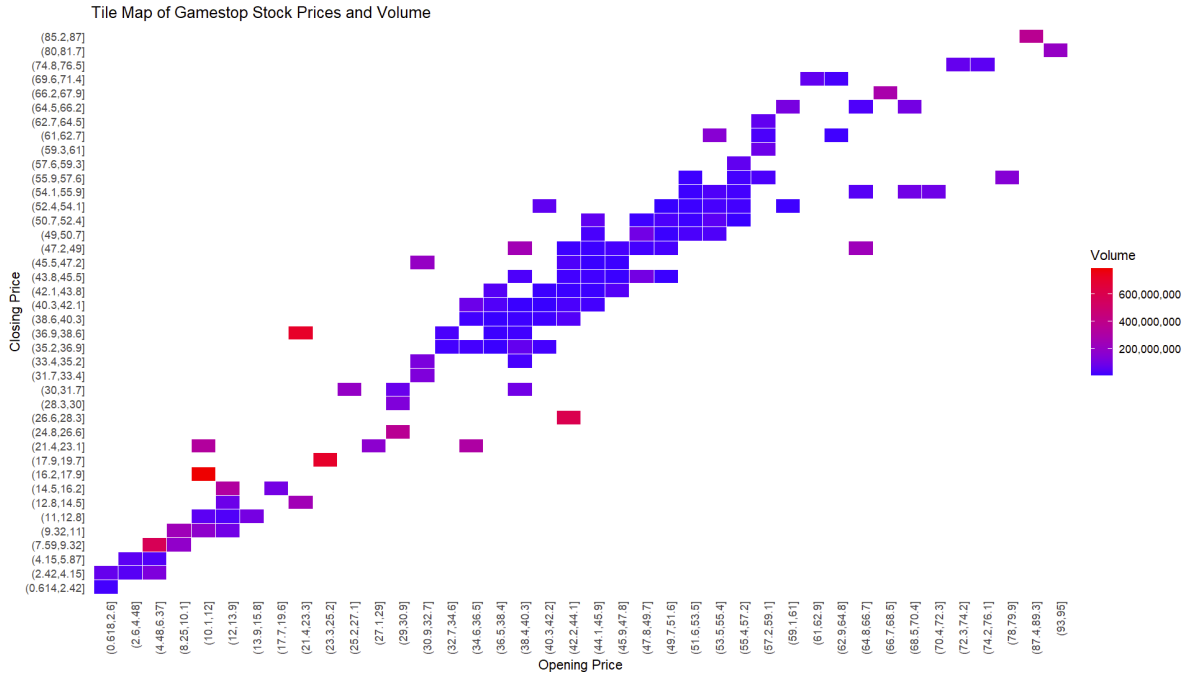


Figure 8: Analyzing Trading Volume Patterns Across GME (GameStop) Stock Prices

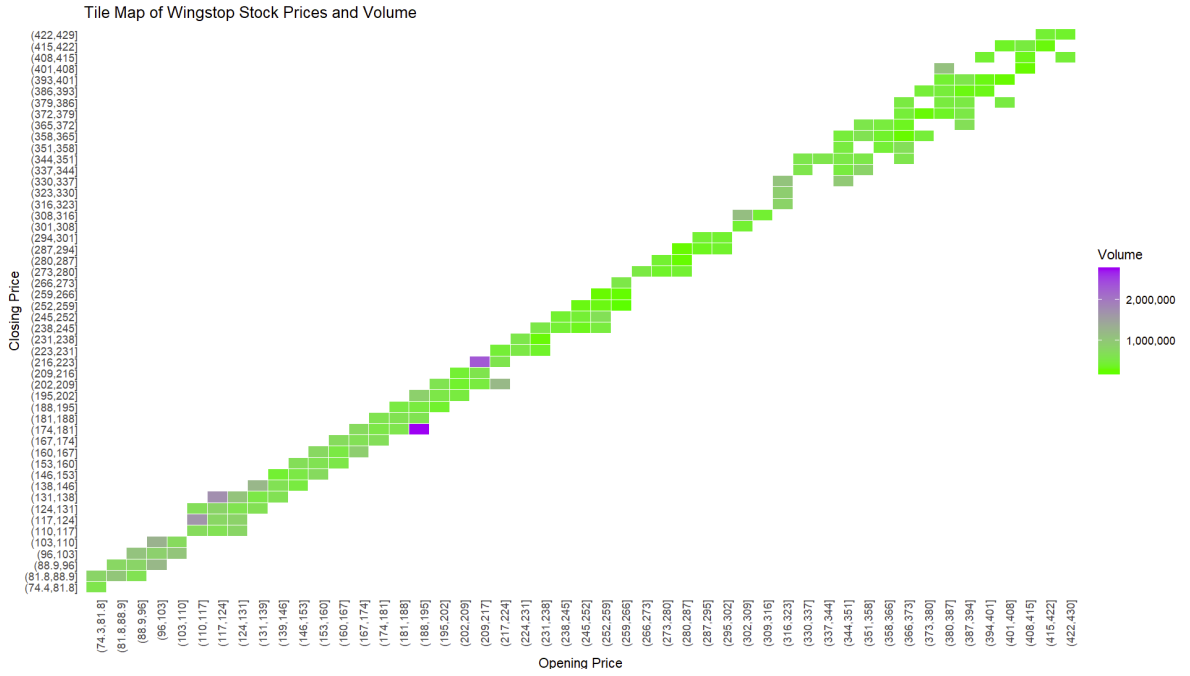


Figure 9: Analyzing Trading Volume Patterns Across WING (Wingstop) Stock Prices

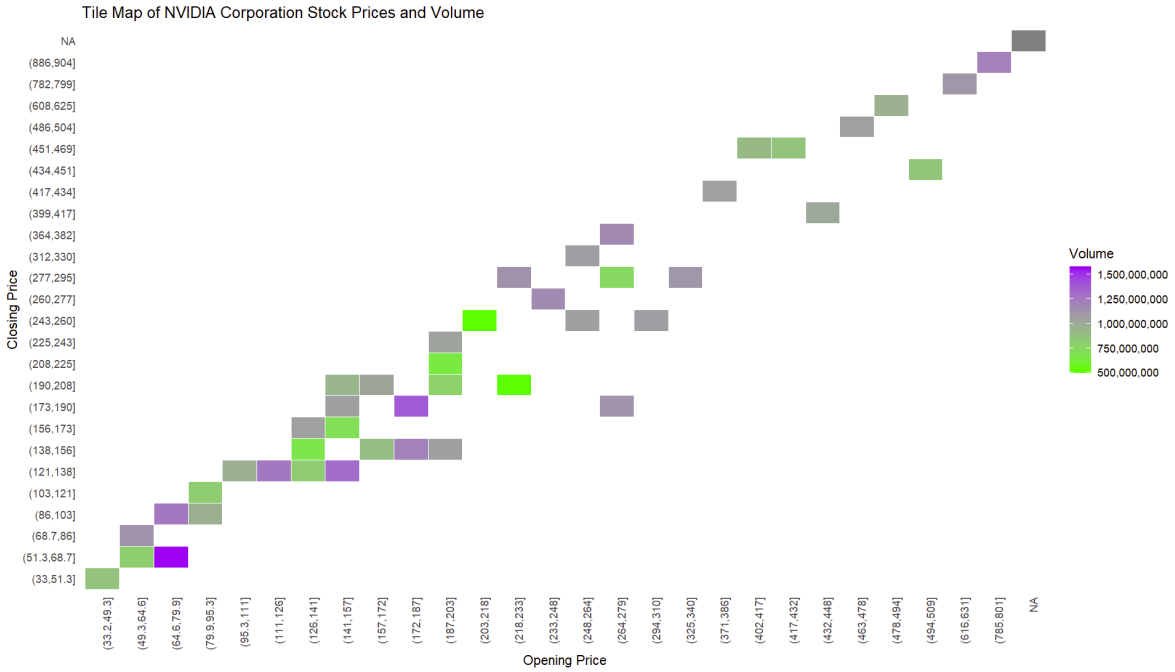


Figure 10: Analyzing Trading Volume Patterns Across NVDA (Nvidia) Stock Prices

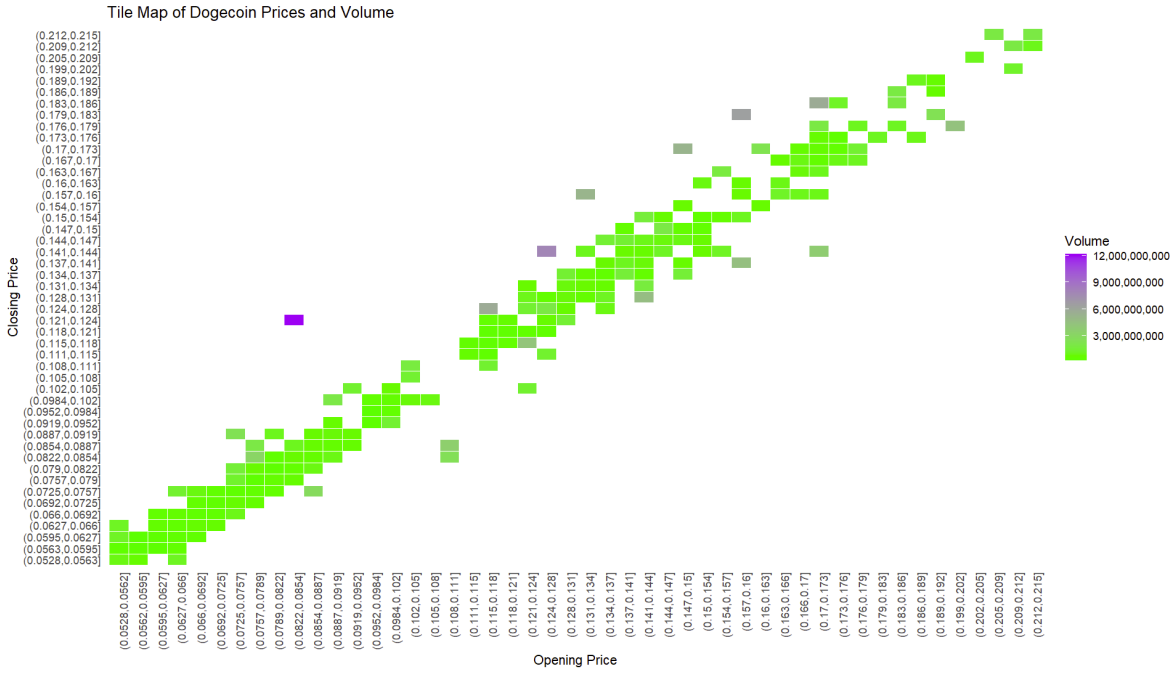


Figure 11: Analyzing Trading Volume Patterns Across Dogecoin Prices