The Influence of Reddit Activity on GameStop's Stock Performance

By: Eduardo Cisneros & Diego Pretelt

Introduction

Data Selection

1. GME.csv
    a. Gamestop Stock Data from 2002 to Present Day
        i. Highs, Lows, Open, Close, Volume
2. Six-months-of-gme-on-reddit-comments.csv
    a. Reddit comments mentioning GameStop from May 2021 - October 2021
        i. Text, Date/Time Created, Subreddit
3. Six-months-of-gme-on-reddit-posts.csv
    a. Reddit posts mentioning GameStop from May 2021 - October 2021
        i. Text, Date/Time Created, Subreddit

Data Transformation

- Loaded csv files as data frames using pandas
- Created new data frames for posts and comments that counted the amount of posts and comments on a given date, then combined and called these mentions
- Finally, combined new Reddit mentions data frame with GameStop stock data by date to get a fully cleaned dataset

	Date	Open	High	Low	Close	Adj Close	Volume	Reddit Mention Count
0	2021-05-03	44.372501	44.372501	39.902500	40.549999	40.549999	21044000	730
1	2021-05-04	39.750000	40.372501	37.950001	40.182499	40.182499	16030000	804
2	2021-05-05	40.457500	41.375000	39.582500	39.869999	39.869999	11221600	991
3	2021-05-06	40.215000	41.180000	38.900002	40.252499	40.252499	11771200	794
4	2021-05-07	40.027500	41.852501	39.375000	40.277500	40.277500	11738400	645
...	...	...	...	...	...	...	...	...
122	2021-10-25	42.355000	43.700001	41.814999	43.492500	43.492500	5771200	2774
123	2021-10-26	43.340000	46.250000	43.125000	44.459999	44.459999	8706800	3744
124	2021-10-27	45.000000	45.772499	43.082500	43.377499	43.377499	4428000	4032
125	2021-10-28	43.790001	45.785000	43.750000	45.712502	45.712502	6784800	4113
126	2021-10-29	45.702499	46.437500	44.500000	45.877499	45.877499	9176000	7288

ETL Process

- Once data was fully cleaned, we created code that would allow us to load the data into MongoDB

ETL PIPELINE DIAGRAM

kaggle

1. GME.csv
2. six-months-of-gme-on-reddit-comments.csv
3. six-months-of-gme-on-reddit-posts.csv

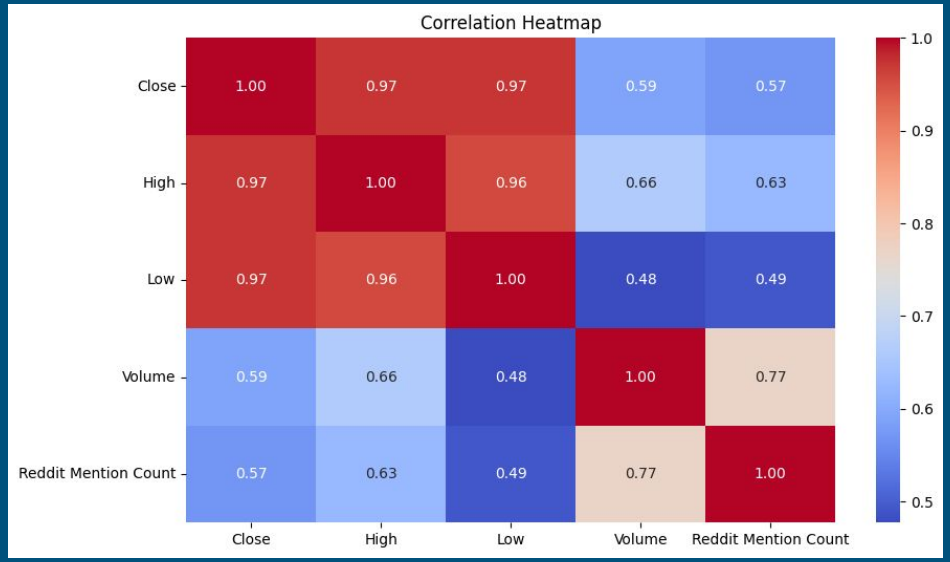
Transformation

- Cleaned using pandas to extract necessary data to complete statistical analysis
  - Dates
  - Stock prices
  - Posts/comments per trading day

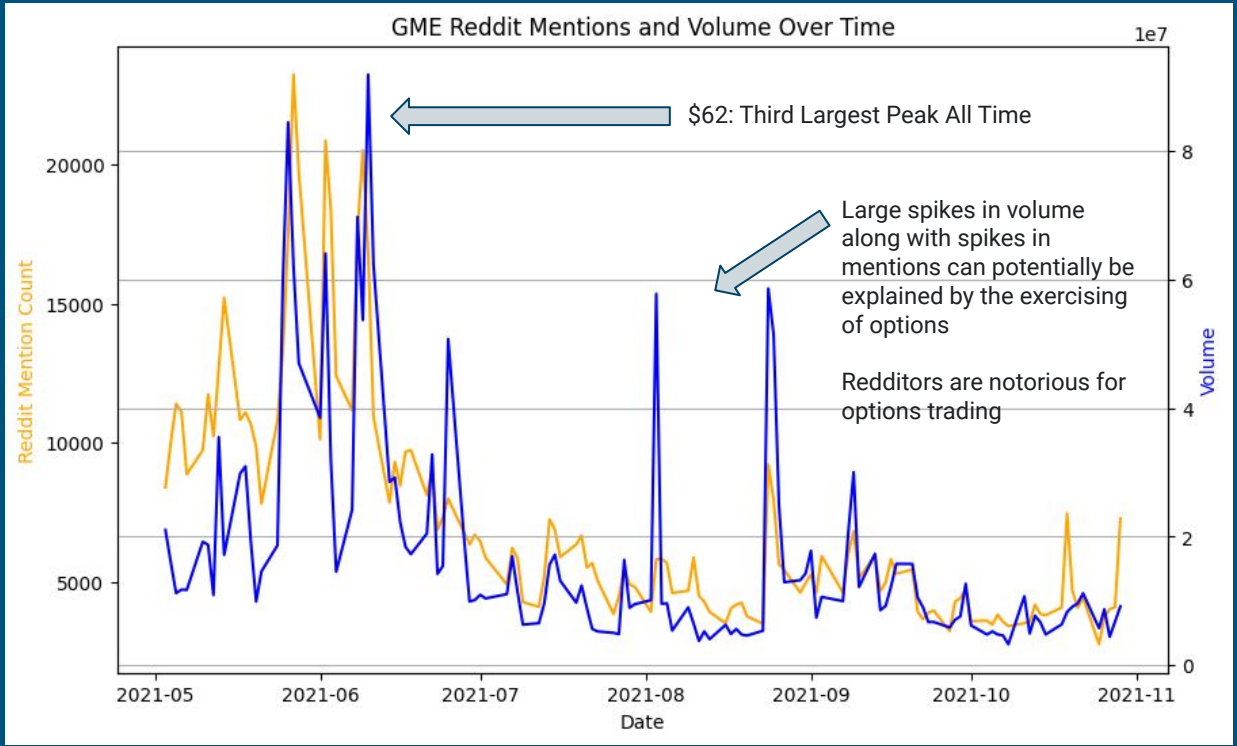
MongoDB®

Analysis

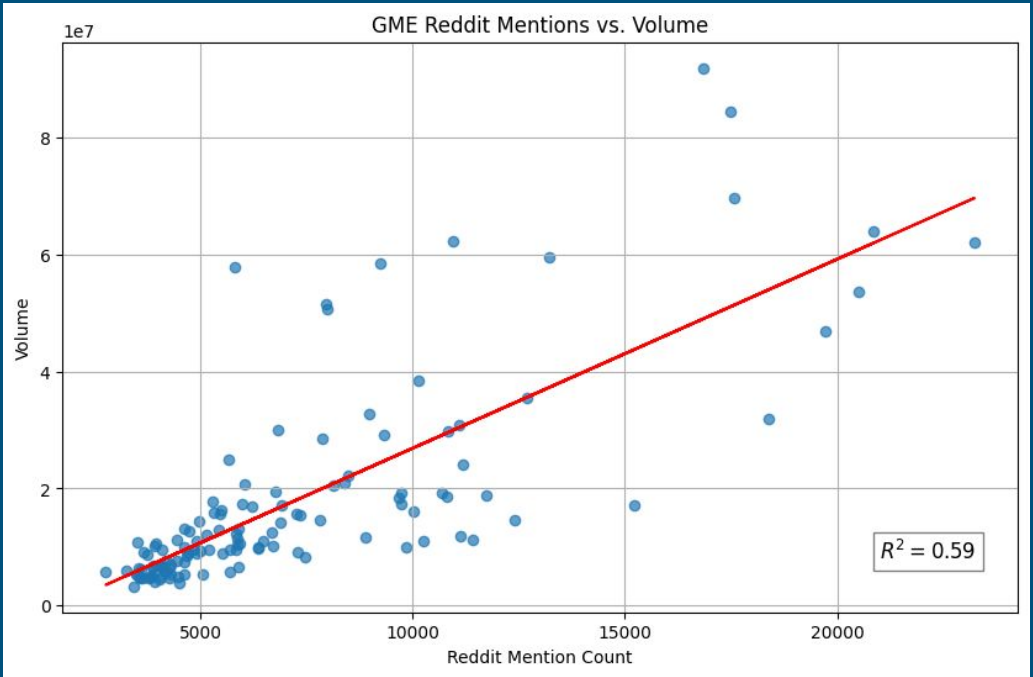
- First analysis we did was trying to find correlation between any of our GameStop stock metrics
  - Used seaborn
- We found Volume to have the highest correlation to the amount of Reddit mentions on a given trading day
  - .77
- Because of this, we decided to continue analyzing the relationship between Volume and Mentions



Time Series


GME Reddit Mentions and Volume Over Time
1e7
$62: Third Largest Peak All Time
Large spikes in volume along with spikes in mentions can potentially be explained by the exercising of options
Redditors are notorious for options trading
Reddit Mention Count
Volume
20000
15000
10000
5000
8
6
4
2
0
2021-05
2021-06
2021-07
2021-08
2021-09
2021-10
2021-11
Date

Linear Regression



R^2 value of .59 tells us that 59% of the change in Volume can be explained by the Reddit Mention Count

Moderate relationship

Cloud Storage

Google Cloud Storage:

- Stored in CSV/Text Format
- Can be accessed and transformed to df for data analysis
- Access Control
- Requires JSON key for access
- Safe Key Storage (IMPORTANT)

MongoDB Storage:

- Stored in Documents where each doc is a row of data (JSON)
- Requires url specific to database and credentials
- Can be accessed and transformed to DF for data analysis

Challenges

- Finding Adequate Datasets
- Data Manipulation
  - Combining three csv files into one data frame
- Data Standardization
  - Pandas DFs to Docs in MongoDB
  - Unix Timestamps to YYYY/MM/DD
- Data Visualization
  - Time series graph was difficult due to different Y axis values
- Google Cloud Storage

Conclusion

- There are is statistical evidence that proves there is a moderate/strong relationship between GameStop's reddit activity and its stock performance
- Increases in social media activity around a stock most likely leads to the stock is being traded heavily (higher volume)
- These trends could potentially be used to predict volume, then therefore predict potential movement in a stock

Thank You!