Impact of Reddit WallStreetBets Forum on GME and AMC Stocks in Early 2021

Authors: Hsiao Chen Yeh, Yanying Jiang, Julien Szarata

Project Description and Motivation

The Coronavirus pandemic shook the lives of many around the world and brought many changes to what the world considered "normal". While it was a very frightening time with the constant news of each country's number of cases or number of deaths increasing constantly, there were also plenty of new learning opportunities that arose out of this dire situation. From the change of corporate work culture, to the sudden wave of new retail investors that began entering the stock market, the world was changing in ways that may not have occurred without the pandemic. As Data Science students, we wanted to apply the coursework and skills that we have cultivated during our time in this program to explore one of the many interesting things that occurred during this time.

Our group wanted to take a look at the memestock[1] trading frenzy that took place at the beginning of 2021 with some names you likely heard from news sources such as GameStop and AMC Theaters.These memestocks grew in popularity on the subreddit r/WallStreetBets where small retail investors joined forces to trigger a short squeeze[2] on these stocks to get back at the hedge funds that were shorting the stocks. Before long even news sources were talking about this subreddit and the memestocks themselves

For this project we wanted to explore the relationship between these memstocks' prices and the activity or chatter on r/WallStreetBets. We wanted to know if there were any correlations between stock price and r/WallStreetBets activity. Our goal is to tell the overall story of the short squeezes that happened in the first half of 2021. Specifically we would like to answer the following questions:

- How did subreddit r/WallStreetBets impact the stock market?
- What were the popular memes stocks that took off in early 2021?
- When was the first major spike in the stock market that reflected the meme stock frenzy?
- How does chatter on the subreddit r/WallStreetBets react with the actual movement of the stock?
- Does high activity on reddit only happen when stock is rising or could it also happen when stock is dropping?

The following assumptions were made based on the news before we conducted our analysis.

- We assumed GameStop trader Keith Gill, known as u/DeepFuckingValue on the subreddit r/WallStreetBets was the key player who attracted a flood of retail cash into GameStop since his reddit posts on September, 2019. We chose this time as our starting point of investigation.
- We assumed that GameStop (GME) and AMC Theaters (AMC) would be the biggest stocks in terms of chatter or activity on the r/WallStreetBets subreddit during the memestock craze.

---

[1] A meme stock refers to the shares of a company that have gained a cult-like following online and through social media platforms. (Investopedia)

[2] A short squeeze is an unusual condition that triggers rapidly rising prices in a stock or other tradable security. It occurs when a security has a significant amount of short sellers, meaning lots of investors are betting on its price falling. The short squeeze begins when the price jumps higher unexpectedly and gains momentum as a significant measure of the short sellers decide to cut losses and exit their positions. (Investopedia)

Data Sources

Reddit Data Download through API

Initial data import, RRD is Reddit Data for all historical posts and comments in the subreddit group r/wallstreetbets from the beginning to present. To download Reddit data, we used the API praw and psaw , the Python Reddit API Wrapper to find the IDs of the historical posts then search for the posts together with comments according to the IDs and download all historical data in the subreddit group r/wallstreetbets. The returned data file is a csv file containing around 4 million rows of records, with each record representing a post or its comment. Major variables of the dataset include reddit post id, created timestamp, title, body, identifier for post or comment. The time period covered in this dataset is from the beginning of r/wallstreetbets in 2012 to the end of Aug. 2022. For reference, we are including the github link of our code to download the RRD data.

Stock Data From Yahoo Finance

GME and AMC are historical daily stock data from August 31, 2019 until September 30, 2022 for those respective symbols, downloaded from Yahoo Finance. You can acquire this data by adjusting the time frame and download it as a csv file. Yahoo Finance presents their data based on the stock exchange. In our case NYSE therefore our data is in EST time zone. There are 777 Rows x 7 Attributes for each respective stock. Among those attributes are- Date, Open, High, Low, Close, Adjusted Close, and Volume for the stock.

News Sources that Help Us Identify Significant Events

- Robinhood temporarily suspended trading on Jan 28th 2021: A halt on trader's ability to buy stocks played an important role in the movement of the stock. This information helps us better understand our analysis findings and interpret our results more accurately.
- Top 10 popular memestocks: To corroborate our initial speculation that GME and AMC were the top two meme stocks among the meme stock frenzy in 2021, we decided to narrow down a list of popular meme stocks and investigated if GME and AMC in fact ranked the highest.
- Who is GameStop trader Deep F—ing Value: This Wikipedia page gave us background on how the memestock craze began to unfold on r/wallstreetbets. Knowing that user u/DeepFuckingValue started his posts on reddit and GME value began to rise in September, 2019, we decided to to make this our starting point of investigation before we narrow down further on the time period of interests that highlight the actual spike in GME's stock price.

Data Cleaning and Manipulation

Manipulation Technology Choices
Data manipulation was done using Python in a Deepnote notebook in order to create a shared environment that could yield reproducible results with an audit trail of our data transformation steps.

Timestamp Conversion
Reddit data was provided in PST and Yahoo Finance data was tied to the NYSE in EST. Pandas.to_datetime() function was used in order to change the timestamps into datetime objects, which were then converted by using a pandas.Timedelta.

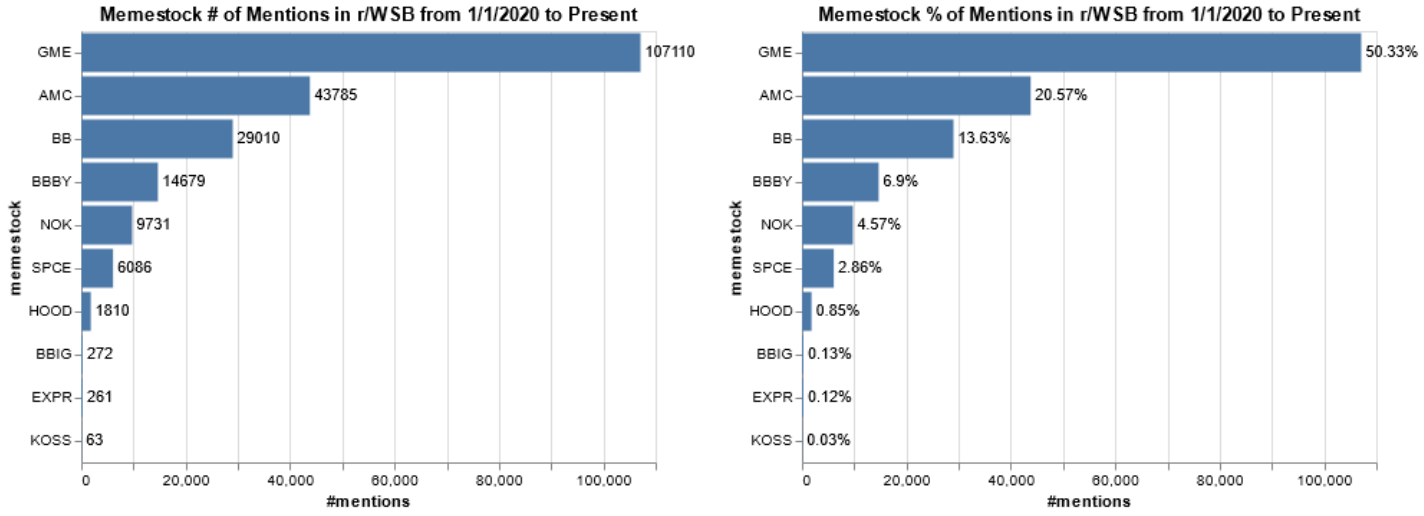
Missing Data
Throughout several visualizations we noticed that our Yahoo Finance data had gaps that our Reddit data did not have gaps for, this was due to non-trading days. We decided to keep the fragmented bar chart because we think it better shows the flow of what happened during this time period.

Aggregation and analysis
Data for both datasets was aggregated into both hourly and daily time windows using pandas.groupby() function to do things like count mentions of GMC and AMC in the reddit data. This allows for comparison against stock price over corresponding time periods.

Analysis and Visualization

Deciding which Memestocks to focus on

In order to decide which memestocks we would be focusing on for the project we first came up with a list of the top 10 memestocks from 2021. From these 10 tickers, we needed to establish a ranking system that would put these in order from one to ten to analyze how big a part of the subreddit they were. The method that we settled on was to count how many times each stock ticker was mentioned on r/WallStreetBets, add those counts up, and that would be the score we used to rank them. Initially we created a function that would loop through the list of 10 memestocks and then count the number of times the ticker was counted in both the title and body fields from our Reddit dataset. However, the initial string count method we used resulted in over counting some tickers such as “BB” where it would match in other tickers like “BBBY” and “BBIG”, and in other random strings. To resolve this issue, we used regex to only count exact matches for each ticker and get more accurate.

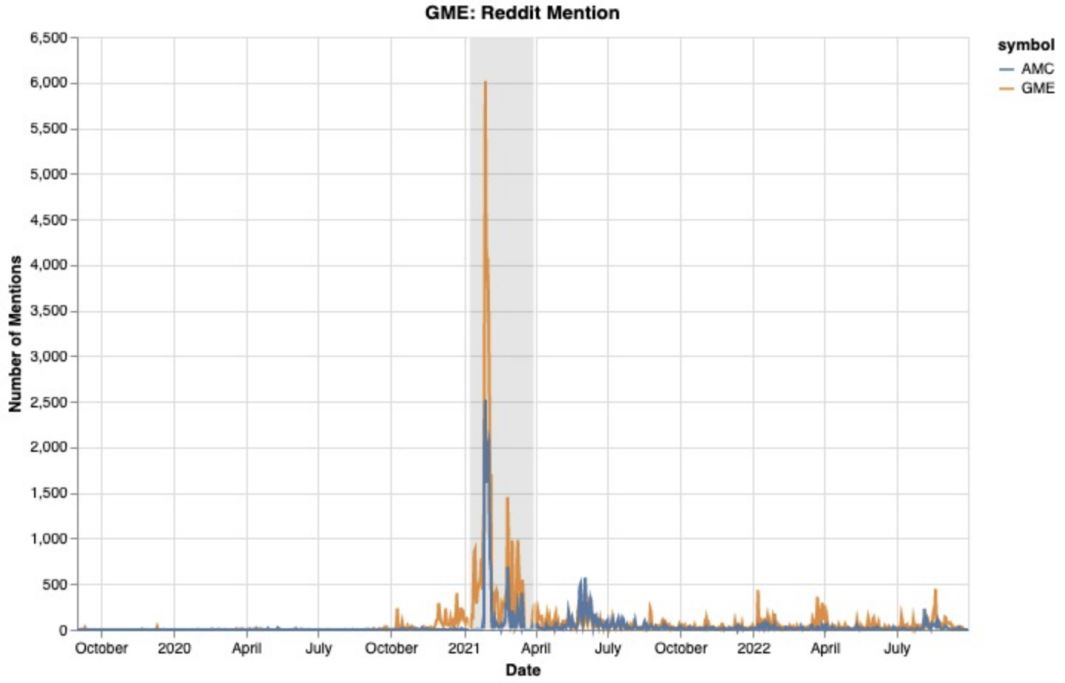


We now had our 10 memestocks ranked by number of counts in the subreddit between our specified dates, and also decided to check each tickers proportion to the total count of those 10 stocks. After doing this we saw that our top two, GME and AMC, accounted for 70% of the total count and felt it was reasonable to focus on these two for our analysis and also confirmed our initial assumption about GME and AMC in terms of activity and chatter on r/WallStreetBets.

When did the first spike of meme stock occur?

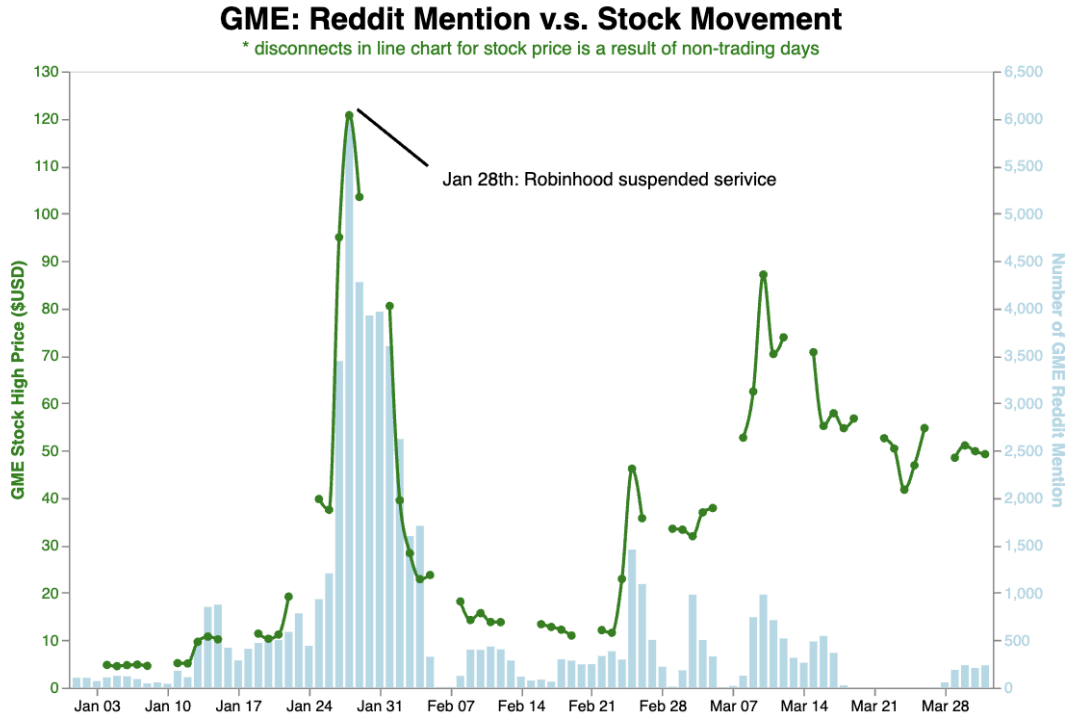
In the following graph, we have a layered line chart that shows GME and AMC’s reddit mentions on r/WallStreetBets. Our time period of investigation was set from 2019-08-31 - 2022-10-01. Because based on the news, September, 2019

was when reddit user u/DeepFuckingValue started to gain traction on subreddit r/WallStreetBets and the value of GameStop began to rise. Once we had this visualization, we were able to identify the initial major spike that occurred in early 2021 and further narrow down our time frame to three months from 2021-01-01 - 2021-04-01 for a more in depth analysis.

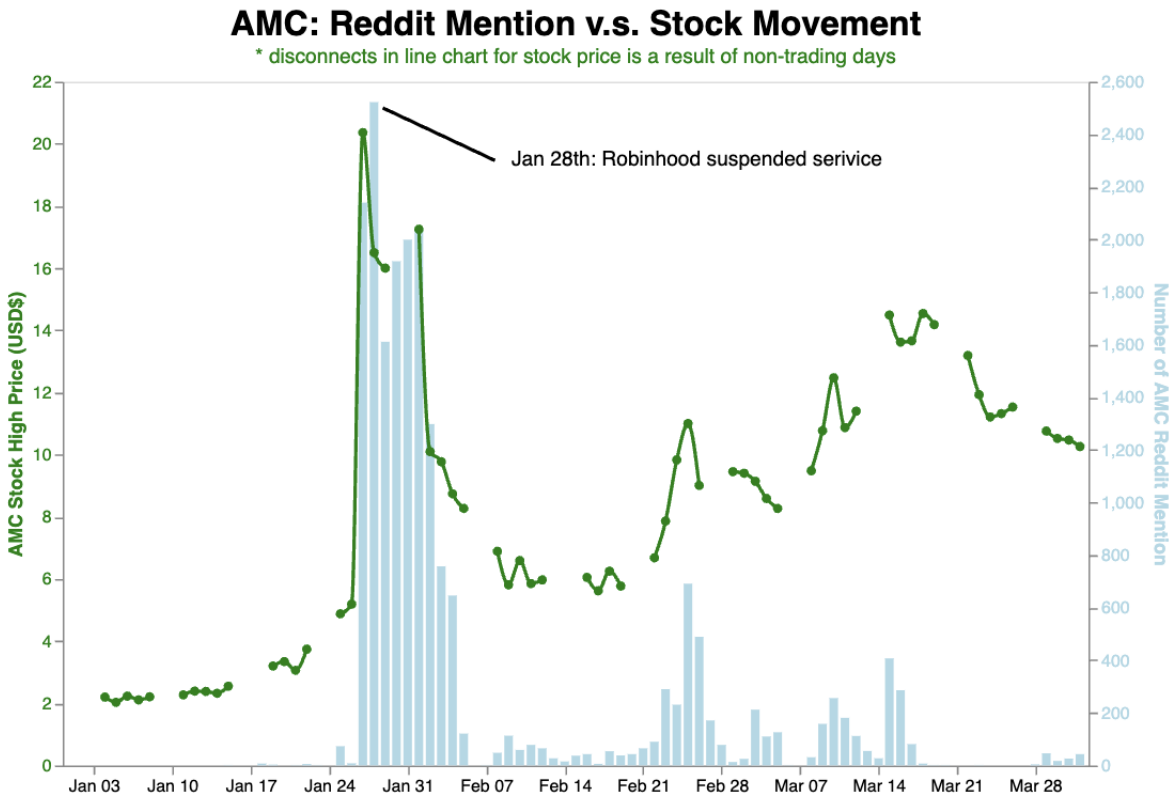


What happened during the first initial spike?

The following visualization gives a three month view from 2021-01-01 - 2021-04-01 with the number of Reddit mention data as bar charts corresponding stock high price as line charts.



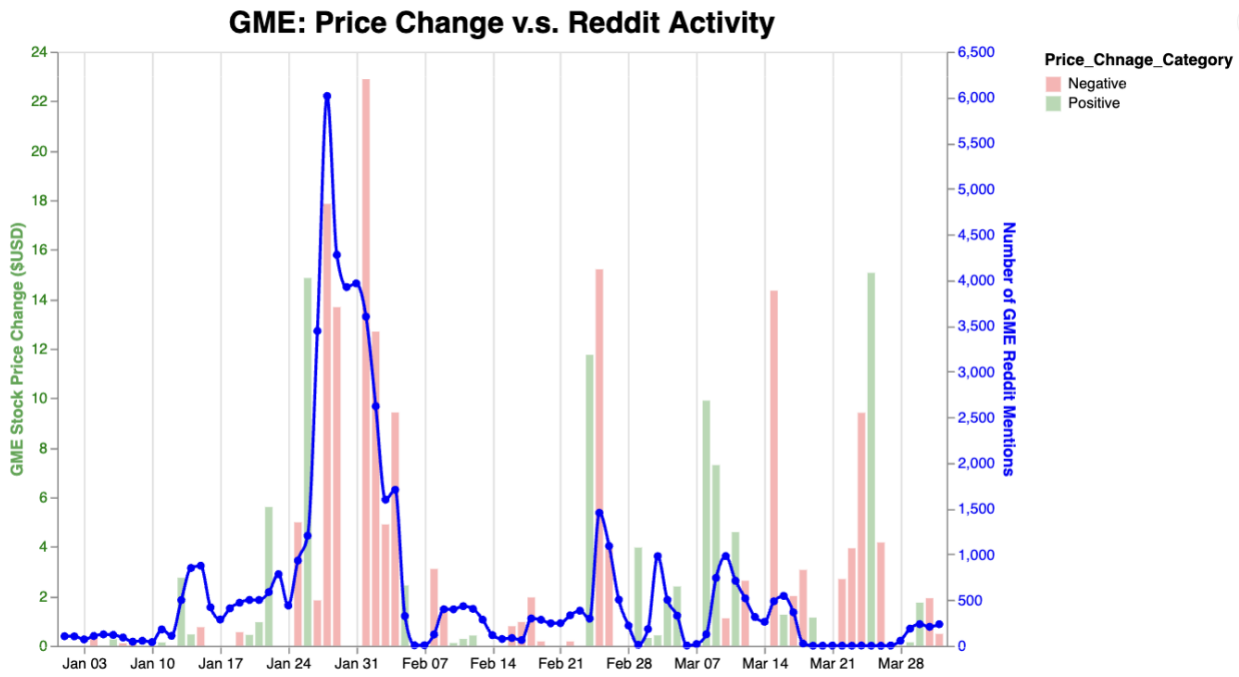
In order to further investigate what actually happened during the first initial spike of the meme stock frenzy, we decided to zoom into the time period of 2021-01-01 - 2021-04-01 that encompassed the first spike and the waves that followed within this three month period. From the graph, we can see that **GME stock value** had the highest **2429% growth from $4.775 in the beginning of January to its highest price of $120.75 on Jan 28, 2021.** Overall, the volume in reddit chatter seemed to align with the movement in the stock market with an exception **on Jan 28th where stock price ceased to rise as a response to chatter volume.** The dip in stock price was due to the fact that trading platform Robinhood temporarily suspended its service on meme stocks such as GME and AMC.



AMC showed a similar pattern within the same time period. Although comparatively AMC didn't rise in value as dramatically as GME. It still shows **an 825% increase in value from $2.2 in the beginning of January to $20.36 on Jan 28th.** Similarly, AMC's growth was impacted by the event of Robinhood suspending its service.

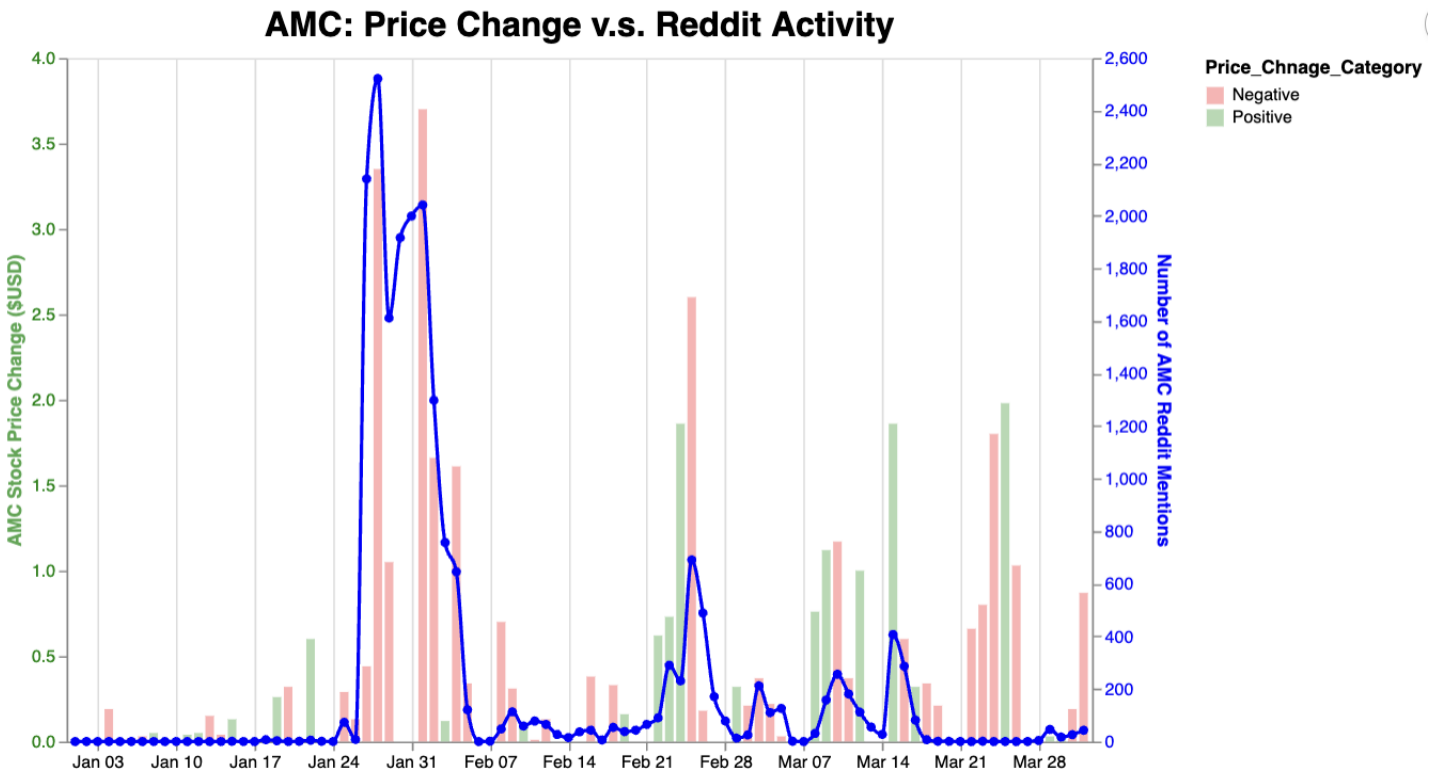
Compare Stock Price Change to Reddit Activity

We were curious whether high activity on reddit impacts solely on a rise in the stock market or could it also impact a drop in the stock value. The following graphs show the dollar amount price change of the stock represented by a bar chart. The color of the bar reflects whether or not the change was in a positive direction or negative. Layered into the visualization is a line chart showing the daily reddit mentions.



We can see that for most major spikes in positive stock movement we see a corresponding spike in reddit chatter on wallstreetbets, when the stock price goes down, we see that interest is lost and reddit chatter ceases. However, within this chart we can also see something that would suggest otherwise - the large price change bars around Mar 25th occurred both in a negative than positive direction and no additional chatter happened at that time period.

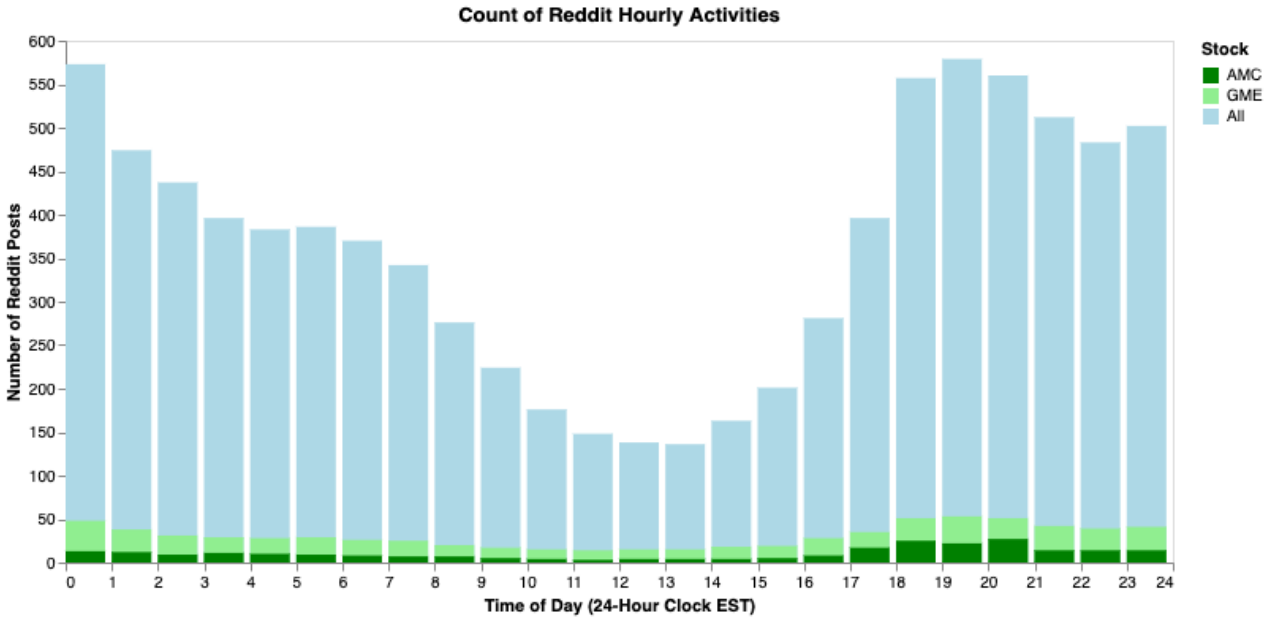
This data suggests that there is a relationship between the jump in stock price and the wallstreetbets chatter, but not one strong enough to suggest that stock price was heavily influenced by chatter. For example, the large spike on Jan 26th was heavily influenced by the chatter leading up to it, but after that point the cat was out of the bag and everyone heard about short squeezing GME. We would go even further to suggest that after that spike that positive stock price changes seems to be the one influencing wallstreetbets chatter.



In the AMC chart we can see similar patterns regarding positive stock movement and reddit mentions, as well as the outlier in March. Positive stock movement followed by reddit chatter can be seen around Feb 25th. This led us to wonder when exactly was the majority of reddit chatter - was it during trading hours while there was high movement in the market or after the market was closed?

When is reddit chatter most active?

We would like to explore at what time of the day is Reddit the most active. So we calculated the average count number of posts and their comments per hour for all reddit r/wallstreetbets posts and for each of the top 2 mostly mentioned stocks, for the period of 01/01/2021 to 04/01/2021.



In the above visualization, we use a stacked bar highlighting activities for GME and AMC in light green and dark green respectively. We can see that during the evenings, Reddit is much more active than during the daytime. Since Reddit is a casual discussion forum that people are more likely to use after work, this pattern is aligned with most people's social media usage behavior.

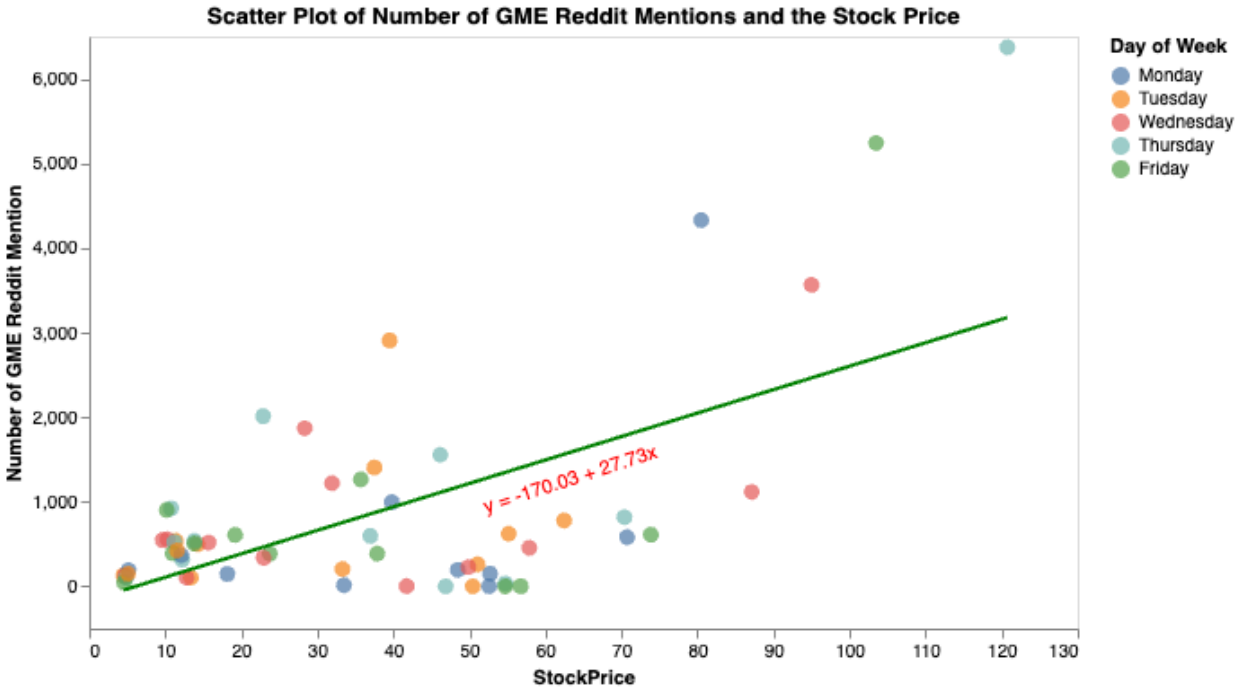
We can also infer from the bar chart that the discussions on the stock GME and AMC are consistent with the overall pattern.

Correlation Analysis and Visualization

GME - Reddit Mention and Stock Price Correlation Analysis

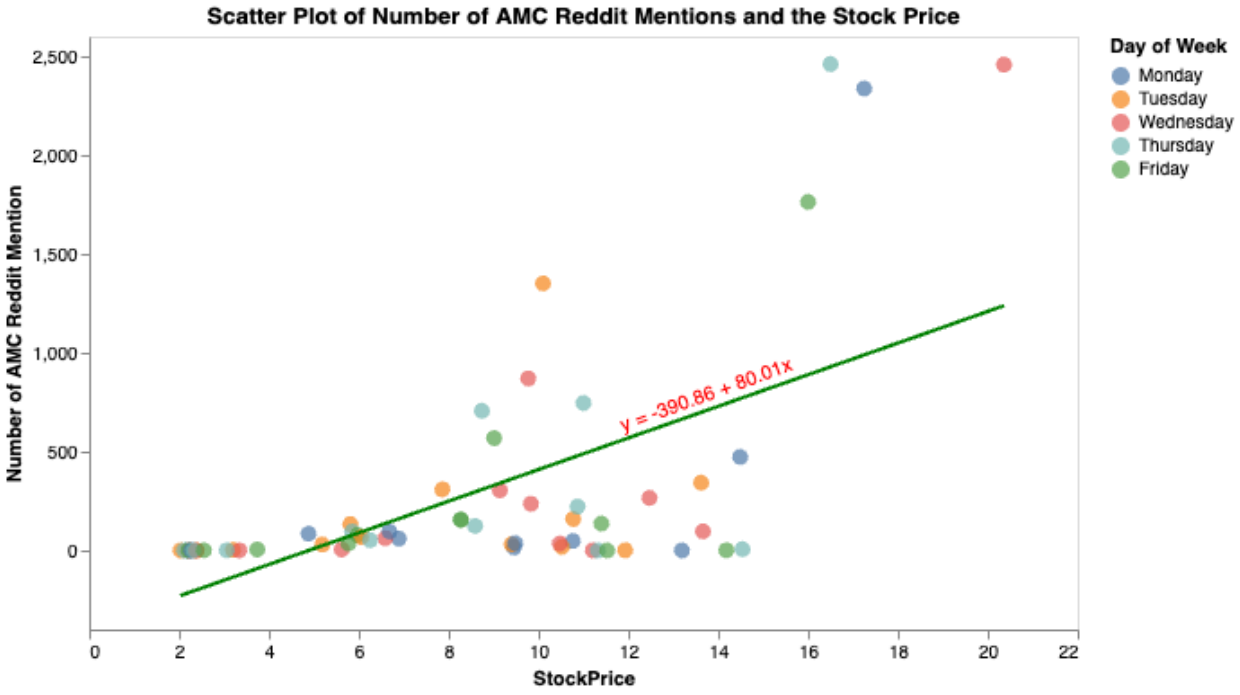
We conduct a correlation analysis to see if there is a relationship between reddit posts and the actual movement of the stock and how strong that relationship may be for GME. We estimate the Pearson correlation coefficient to quantify the direction and strength of the linear association between the Yahoo finance and Reddit wsb datasets.

The correlation coefficient of stock price and count of reddit mentions for GME is 0.61, indicating a positive correlation between the two variables.



In addition, we wanted to explore if there is any pattern embedded in the day of week. So we add color coding for each day of week in the above scatter plot. The color pattern is pretty random according to the above plot so we don't have any conclusion for this.

AMC - Reddit Mention and Stock Price Correlation Analysis



We also do the same for AMC. The correlation coefficient of stock price and count of reddit mentions for AMC is 0.61, indicating a positive correlation between the two variables.

Again, we are trying to find the pattern for the day of week, but we don't see any clear relationship between the day of week and mentions of the stock.

Summary and Future Research Opportunities

- In this project, we have conducted various analyses to explore and visualize the relationship between stock price change and the activities on reddit r/wallstreetbets forum. We concluded that there is positive correlation between the two datasets, for the top 2 meme stocks GME and AMC, during the period of 01/01/2021 - 04/01/2021, when meme stocks took off.

- During the project we had encountered a couple of challenges. The first one we faced was that the Kaggle data we originally chose for reddit wsb mentions only has data through Aug. 2021. So we were missing a big chunk of data for the most recent year. We ended up figuring out how to use reddit api praw and psaw to download the full dataset for the whole subgroup r/wallstreetbets so we are able to conduct the analysis for the period we chose.

- Another challenge was that we wanted to see the correlation between yahoo finance and reddit mentions per hour. However, Yahoo finance doesn't provide us with hourly historical stock prices. We ended up calculating the difference between the open and close price to support our analysis in the above section: Compare GME/AMC Price Change to Reddit Activity.

- We also wanted to explore the impact on the stock price of an important person: Keith Gill, who started the GameStop madness back in September 2019. According to his wikipedia, his posts under the username "DeepFuckingValue" on Reddit were cited as a driving factor in the GameStop short squeeze of Jan. 2021 and the subsequent trading craziness in meme stocks. However our dataset has all user info anonymized so we weren't able to conduct the exploration on his behaviors in an efficient way. This also creates another future research opportunity for us after we figure out how to obtain the historical data for a specific user on Reddit.

- In the future, we would like to apply more advanced techniques using machine learning methods to conduct predictions for meme stock spike from the reddit wsb posting activities. And we are also thinking about looking at stock trade volumes and correlating that to mentions of reddit.

Statement of Work

Our team collaborated using Slack for instant messaging, Google Meet for scheduled calls, and Deepnote for working on the notebook together. We generally met at least once or twice a week to discuss the project status, trouble areas, next steps, bounce ideas off of each other, and to solve problems together. We kept a work-list with assignments for each team member to work on and updated the list after each meeting. Outside of our scheduled calls we would mostly communicate via Slack for quick updates. We also would use our meetings to discuss schedules for when we were free to meet next, what we should try to have finished before the next meeting, as well as decide on when we felt would be a good time to schedule a check-in with our assigned faculty member.

In total, we had three scheduled video meetings with our project advisor, Oleg, where we asked for feedback on our project direction and ran ideas by him on what we were thinking about adding or removing from the project. The check-ins with Oleg were extremely valuable as he helped us realize things that we had overlooked and assisted us when we would feel stuck.

Project Task	**Yanying**	**Hsiao**	**Julien**
Project Proposal Brainstorming	x	x	x
Project Proposal Writing	x	x	x
Coordinating Team Meetings & Advisor Meetings		x	x
r/WallStreetBets Data set	x		
Yahoo! Finance Data set		x	
Relevant Sources	x	x	x
Deepnote Coding Collaboration	x	x	x
Data Manipulation & Cleaning	x	x	x
Analysis: Memestock Selection			x
Analysis: Memestock Price vs Reddit Activity		x	
Analysis: Correlation	x		
Visualizations	x	x	x
Final Report	x	x	x