Reddit Stock Mentions, Price Volatility, and Trading Volume

Robert DeWitt, Katherine Hirth, & Katherine Yu

I. Introduction & Problem Statement

Subreddit r/WallStreetBets was created in 2012 by Jaime Rogozinski for those interested in high-risk, high-return trades but who desired a forum that was less serious than traditional stock market forums or communities.[1] More recently, it has become a popular way for retail investors to communicate with each other and even mobilize against big hedge funds and corporations. In January of 2021, r/WallStreetBets gained rapid national notoriety for the subreddit's pivotal role in the organized short squeeze of GameStop stock, an event which facilitated the rise of GameStop's stock price from below $20 per share to over $340 per share. Since GameStop's stock surge, speculation on the rising power of the retail investor has become a consistent theme in financial news. Intuitive and no-commission trading apps such as Robinhood have drastically increased the general public's access to the stock market, introducing retail investors as a market-moving force to be reckoned with. r/WallStreetBets has remained a surprising centerpiece for speculation on the rise of the retail investor's market influence, and the subreddit has swelled to over 10 million members since January. But to what extent is this speculation valid, i.e. does a relationship exist between r/WallStreetBets and the market for individual stocks?

II. Objective

Our goal is to scrape recent 'Daily Discussion' threads on subreddit r/WallStreetBets for stock ticker mentions. Using the frequency of stock mentions per day for each unique stock ticker, we will then explore if there is any correlation between mention frequency, stock price volatility, and stock trading volume to better understand what kind of relationship exists between stock prices and social media buzz.

III. Methodology & Data

Our first step in this project is to scrape all possible stock tickers listed on the New York Stock Exchange and the NASDAQ so that we are able to identify when they appear in reddit comments. Using python and Xpath, we pulled each NYSE and NASDAQ stock ticker from eodddata.com and saved this information as a list in our python file.[2]

Then, we scrape r/WallStreetBets for recent daily discussion thread links and modify those links to sort all comments by 'top' which displays the comments which have the most 'up votes' at the top of the page.[3]

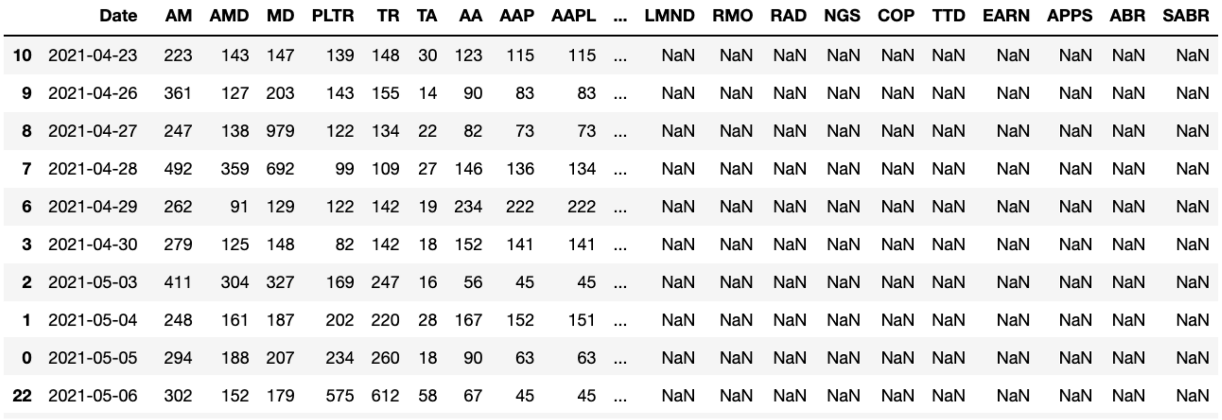
However, when scraping the comments from each thread, we run into challenges because of the way that reddit is structured. Instead of showing every single comment and reply (a comment on another comment), reddit hides almost all comments behind click-through buttons so the original thread page only shows a very small number of comments. Additionally, reddit does not load an entire page at once and, as one scrolls down on the page, more comments and click-through buttons are loaded. These two issues make it very difficult to scrape every single comment and reply from a thread.

Given this convoluted structure, we try three different methods of web scraping to get the highest number of comments possible. First, since comments are filed under ‘p’ tags on the site, we attempt a simple Xpath search of the html content for the text within each ‘p’ tag. However, this method returns less than 50 comments per thread because any comments which are not loaded on the original thread page or are hidden behind click-through buttons are not listed under a ‘p’ tag.

Second, we try an Xpath search of the entire html content for text within the general script of the page. This excludes any coding for side-bars, headers, and footers on the page. Looking at this information, we notice that any text comment, even hidden ones which did not appear with a ‘p’ tag, is preceded by the following pattern: ‘“media":{"richtextContent":{"document":[{"c":[{"e":"text","t":’. This enables us to turn the content into a string and split it every time ‘"e":"text","t":’ appears. Then, since the first string after each split corresponds to a comment, we can pull the comments on each page. This returns roughly 300 comments per thread which, though not close to the thousands which are posted daily, is a big improvement upon our original ‘p’ tag search. Searching through Xpath was therefore capable of returning some results, but for sites with scroll-down and click-through barriers, it may not be the best method of web scraping.

Our third approach uses an API in conjunction with the thread links scraped through Xpath to generate a much larger number of comment text results. Many companies running sites with a large amount of data, like Spotify, Reddit, and Twitter, have public API’s for more accessible data gathering. After using Xpath to find the relevant Daily Discussion threads, we call the Pushshift API to get a list of comment ID’s on each thread.[4] Then we call the API again to get a list of the text body for those comments ID’s for every 100 comments (due to API query limits). With this method we were able to get nearly all the comments on each thread. We chose to cap the results at 20,000 comments for each thread since the API occasionally crashes when dealing with larger amounts of data. This cap does not affect the results however, since most threads had less than 20,000 comments.

Then, for each comment pulled from the API, we search for occurrences of the stock tickers in our previously scraped NYSE ticker list. We count how often each ticker is mentioned in a Daily Discussion thread and repeat this process across Daily Discussion threads from April 23rd to June 1st. Figure 1 below shows a subset of our final ticker count results from April 23rd through May 6th.



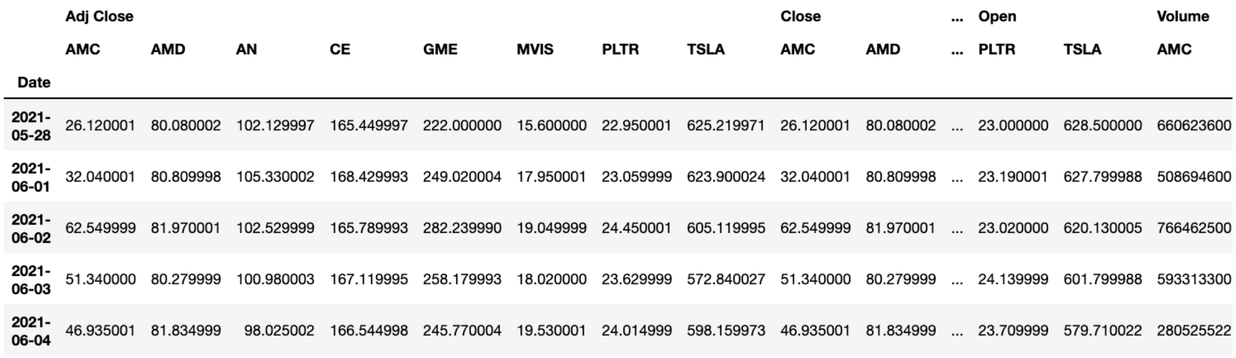
	Date	AM	AMD	MD	PLTR	TR	TA	AA	AAP	AAPL	...	LMND	RMO	RAD	NGS	COP	TTD	EARN	APPS	ABR	SABR
10	2021-04-23	223	143	147	139	148	30	123	115	115	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	2021-04-26	361	127	203	143	155	14	90	83	83	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	2021-04-27	247	138	979	122	134	22	82	73	73	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	2021-04-28	492	359	692	99	109	27	146	136	134	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	2021-04-29	262	91	129	122	142	19	234	222	222	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	2021-04-30	279	125	148	82	142	18	152	141	141	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2021-05-03	411	304	327	169	247	16	56	45	45	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2021-05-04	248	161	187	202	220	28	167	152	151	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0	2021-05-05	294	188	207	234	260	18	90	63	63	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
22	2021-05-06	302	152	179	575	612	58	67	45	45	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

(Above) Figure 1: Web Scraping Dataset

However, we encountered a few limitations to counting ticker occurrences. Due to the way we searched each comment, there were certain tickers symbols that were falsely counted. For example, some ticker symbols that are contained in other longer ticker symbols were double counted (i.e. AM and AMD). Additionally, ticker symbols that are also common words (i.e. AM or TIME) were falsely identified and therefore had biased counts. In order to ensure that our results are not influenced by these limitations, we are very selective in the stocks that we choose to analyze, making sure that each selected stock has the following two qualities:

1. We know that it is a 'meme' stock
2. It has a unique enough name that there will be minimal bias in the counts

We then use the Yahoo Finance library in Python to retrieve the stock prices, volume, etc. for the selected stocks. This is a snapshot of what our data looks like:



	Adj Close								Close		...	Open		Volume
	AMC	AMD	AN	CE	GME	MVIS	PLTR	TSLA	AMC	AMD	...	PLTR	TSLA	AMC
Date														
2021-05-28	26.120001	80.080002	102.129997	165.449997	222.000000	15.600000	22.950001	625.219971	26.120001	80.080002	...	23.000000	628.500000	660623600
2021-06-01	32.040001	80.809998	105.330002	168.429993	249.020004	17.950001	23.059999	623.900024	32.040001	80.809998	...	23.190001	627.799988	508694600
2021-06-02	62.549999	81.970001	102.529999	165.789993	282.239990	19.049999	24.450001	605.119995	62.549999	81.970001	...	23.020000	620.130005	766462500
2021-06-03	51.340000	80.279999	100.980003	167.119995	258.179993	18.020000	23.629999	572.840027	51.340000	80.279999	...	24.139999	601.799988	593313300
2021-06-04	46.935001	81.834999	98.025002	166.544998	245.770004	19.530001	24.014999	598.159973	46.935001	81.834999	...	23.709999	579.710022	280525522

(Above) Figure 2: Stock Data

IV. Analytic Methods & Results

For each of the eight stocks we chose, we focused our analysis on stock volatility and trading volume in order to evaluate whether frequency of mention on Reddit for a given stock is associated with larger trading volumes and increased price movements. For each stock we calculated annualized volatility and trading volume for each day that we scraped ticker count data for. Annualized volatility was calculated using a backward rolling window of 252 trading days for mean calculation. We then calculated the Pearson correlation coefficient and associated p-value between each stock's ticker scrape count and its daily trading volume and volatility. (See Appendix A-E for tables and graphs comparing selected stocks). Additionally, in order to investigate the relationship between our variables across time, we also calculated the correlation coefficients of lagged ticker count values (up to 3 period lag) with trading volume and volatility. Our general results can be seen in the table below, while plots and correlation matrices for each of the individual stocks can be found in Appendix F-M.

	MVIS	**AMC**	**AMD**	**PLTR**	**TSLA**	**GME**	**CE**	**AN**
Lag 0	0.98	0.96	0.85	0.83	0.64	0.95	-0.02	-0.33
Lag 1	0.81	0.79	0.19	0.67	0.44	0.76	0.18	-0.38
Lag 2	0.82	0.64	-0.061	0.58	0.33	0.39	0.38	-0.38
Lag 3	0.57	0.42	0.07	0.51	0.067	0.13	0.61	-0.37

(Above) Figure 3: Ticker Scrape Count and Trading Volume Correlation Results Table

	MVIS	**AMC**	**AMD**	**PLTR***	**TSLA**	**GME**	**CE**	**AN**
Lag 0	0.83	0.46	0.52	N/A	-0.027	0.57	-0.64	-0.69
Lag 1	0.89	0.57	0.45	N/A	-0.047	0.62	-0.69	-0.67
Lag 2	0.89	0.52	0.48	N/A	0.042	0.7	-0.66	-0.72
Lag 3	0.9	0.35	0.47	N/A	0.18	0.7	-0.64	-0.75

(Above) Figure 4: Ticker Scrape Count and Annualized Volatility Correlation Results Table

*Note: Palantir stock has not been public long enough for an accurate calculation of annualized volatility

For the majority of the stocks we analyzed, our results show heavy and statistically significant correlations between trading volume levels and the frequency of ticker mentions on r/WallStreetBets. Furthermore, the correlation often extends to Reddit ticker count levels 1-3 periods prior to a given day's trading volume. Perhaps unsurprisingly, many of the strongest correlations can be seen in stocks known as 'meme' stocks, such as AMC and GameStop (GME).

While our correlation results for volatility and stock ticker count levels are more subdued than those for trading volume, there remains a strong degree of correlation between volatility levels in certain stocks and frequency of ticker mention on Reddit, even at lagged intervals. Interestingly, the volatility correlations of both GME and MVIS increase with larger lags, implying that a large increase in GME or MVIS ticker mentions on r/WallStreetBets is associated with increasing levels of stock volatility in the following days.

Considered altogether, our results provide strong initial evidence that a significant relationship exists between reddit ticker mentions and stock volatility and trading volume. It is important to note, however, that high correlations in our results are not indicative of causality. While significant limitations exist in our analysis, such as our limited time window, our results suggest the area worthy of future exploration.

V. Economic Insights Derived

Should our results hold valid in a broader analytical study, they would have broad practical value in a variety of stock trading strategies. There are many strategies that rely on trading volume as a foundational component, and the ability to predict likely increases or decreases in trading volume could prove to be quite profitable in such circumstances. More importantly, the ability to predict a stock's future volatility is incredibly valuable in the options market. While options rely on implied volatility rather than historical volatility for pricing, the ability to gain some level of predictive ability over a stock's near-term future volatility might allow for a more accurate assessment of whether options are overpriced or underpriced at current price levels.

Our results may also be indicative of an underlying trend emerging in the market following the rise to prominence of r/WallStreetBets. In search of the next big 'meme' stock, many proprietary trading funds are also scraping reddit for possible insights. The possible consequence of this is that, in some form or another, trading firms are undertaking the same general analysis that we have started, but are ultimately trading on derived insights. The amplifying effect of this dynamic might reasonably explain why, after a few thousand mentions on r/WallStreetBets, we see large movements in a stock's trading volume and volatility. Of course, the causation assumption being made here requires far more investigation to be proven valid, but it is an interesting and relevant possible explanation nonetheless.

VI. Summary

Ultimately, through this project, we have learned that different methods of web scraping are useful for different types of websites. Though Xpath is a powerful tool, it cannot scrape information that is hidden behind scroll-downs or click-through buttons and is more suited to scraping sites where all of the target information is readily available on the original page.

For the majority of our sample of stocks, our results indicate strong levels of correlation between ticker mentions on r/WallStreetBets and stock volatility or trading volume. For many stocks, these results hold even at lagged intervals for scraped ticker counts, and were often strongest for those stocks commonly considered 'meme' stocks. While these results certainly require further investigation, such a strong current and lagged correlation might serve as a critical component in many types of trading strategies, and may also be indicative of the extent to which many financial firms are already trading based on scraped Reddit insights.

VII. Suggested Future Work

In terms of future research, there are a few areas that have room for improvement. First is the amount of data that we were able to scrape. Since it did take some time to call the API each time and collect data, due to our time limits on this project we were only able to collect a month's worth of data. If we had more data going further back in time we would be able to do a more in depth time series analysis.

Additionally, there were probably better ways to program our counter for ticker symbols. It is not computationally efficient due to multiple for loops, and there is probably a more efficient way to program the count. More importantly, further improvement on our string matching algorithm is needed. As we saw, our model picked up some erroneous counts for 2 letter ticker symbols as well as some common words like 'TIME', 'USA', 'IMO', etc. that are also ticker symbols as well. Although we could specify a list of words to be excluded, it would be excluding the ticker mentions at the same time.

With these 2 issues fixed, our data could become more usable and accurate.

VIII. References

[1] Stokel-Walker, Chris. "GameStop: the Oral History of r/WallStreetBets' Meme Stock Bubble." *British GQ,* 22 Mar. 2021, www.gq-magazine.co.uk/lifestyle/article/gamestop-stock-oral-history.

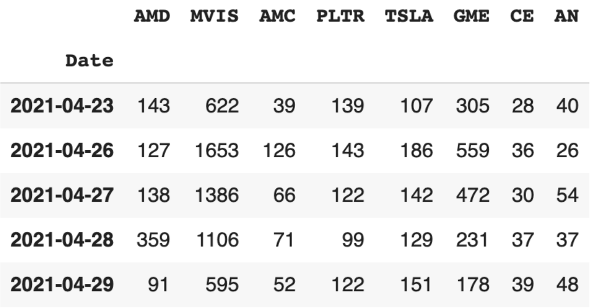
[2] "List of Symbols for New York Stock Exchange [NYSE] ." EOD Data, 1 June 2021, eoddata.com/stocklist/NYSE/A.htm. "List of Symbols for New York Stock Exchange [NYSE]." *EOD Data,* 1 June 2021, eoddata.com/stocklist/NYSE/A.htm.

[3] "r/Wallstreetbets Daily Discussions." *Reddit,* 4 June 2021, www.reddit.com/r/wallstreetbets/search?q=flair_name%3A%22Daily+Discussion%22&amp;restrict_sr=1&amp;sort=new.

[4] Baumgartner, Jason Michael. "Pushshift Reddit API Documentation." *GitHub,* GitHub, Inc. , 1 Oct. 2019, github.com/pushshift/api.

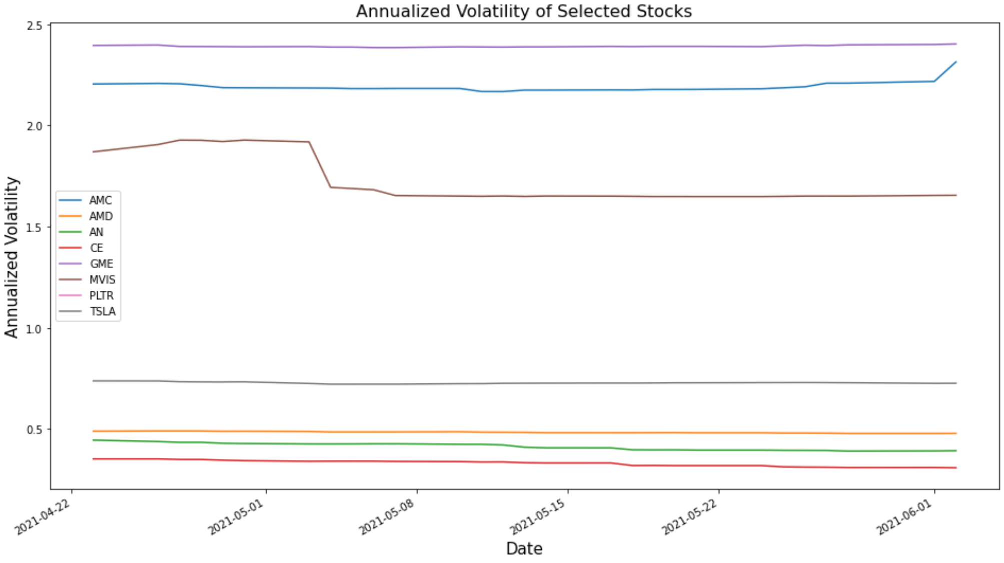
IX. Appendix

Appendix A: Mention Frequency of Selected Stocks between 4/23 and 4/29

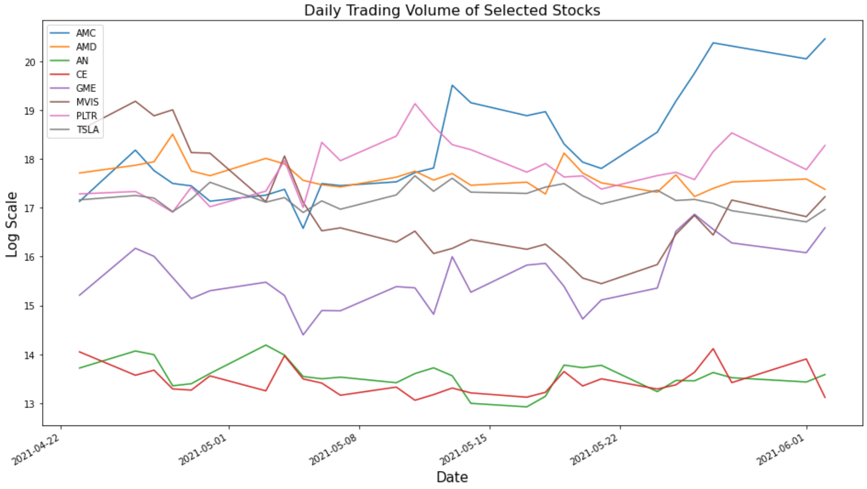


	AMD	MVIS	AMC	PLTR	TSLA	GME	CE	AN
Date								
2021-04-23	143	622	39	139	107	305	28	40
2021-04-26	127	1653	126	143	186	559	36	26
2021-04-27	138	1386	66	122	142	472	30	54
2021-04-28	359	1106	71	99	129	231	37	37
2021-04-29	91	595	52	122	151	178	39	48

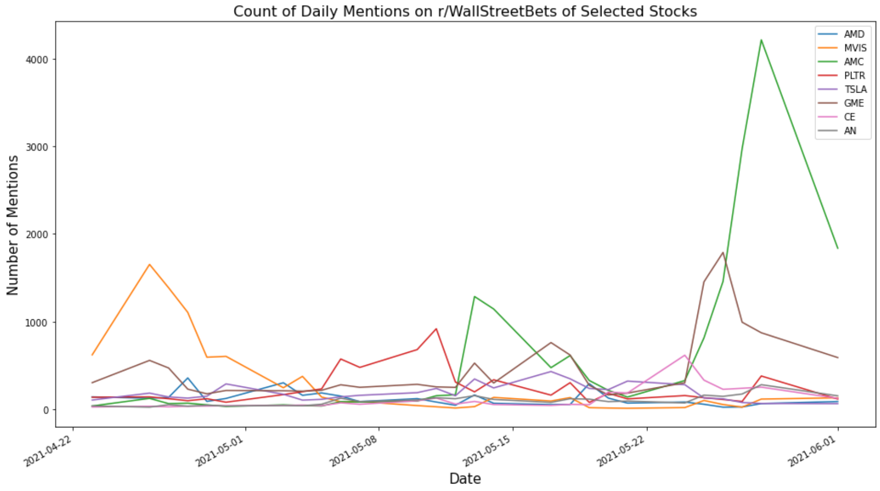
Appendix B: Annualized Volatility of Selected Stocks



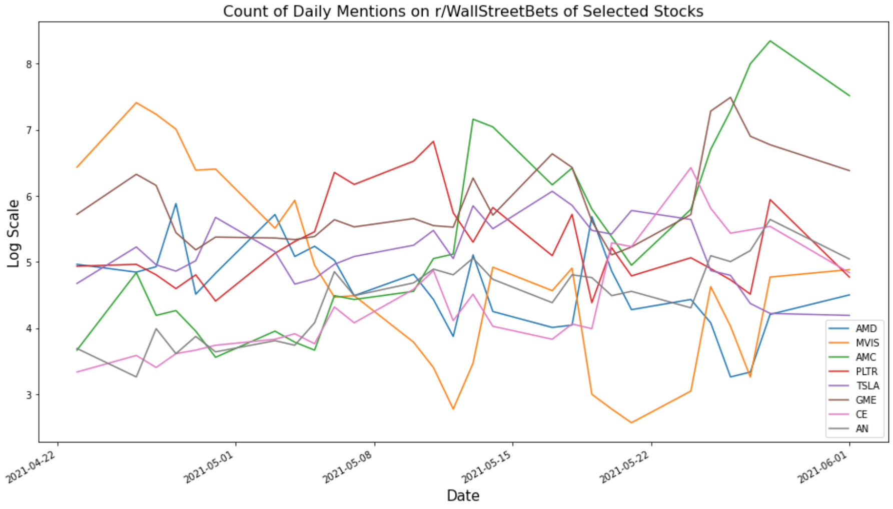
Appendix C: Daily Trading Volume of Selected Stocks



Appendix D: Count of Daily Mentions on r/WallStreetBets of Selected Stocks

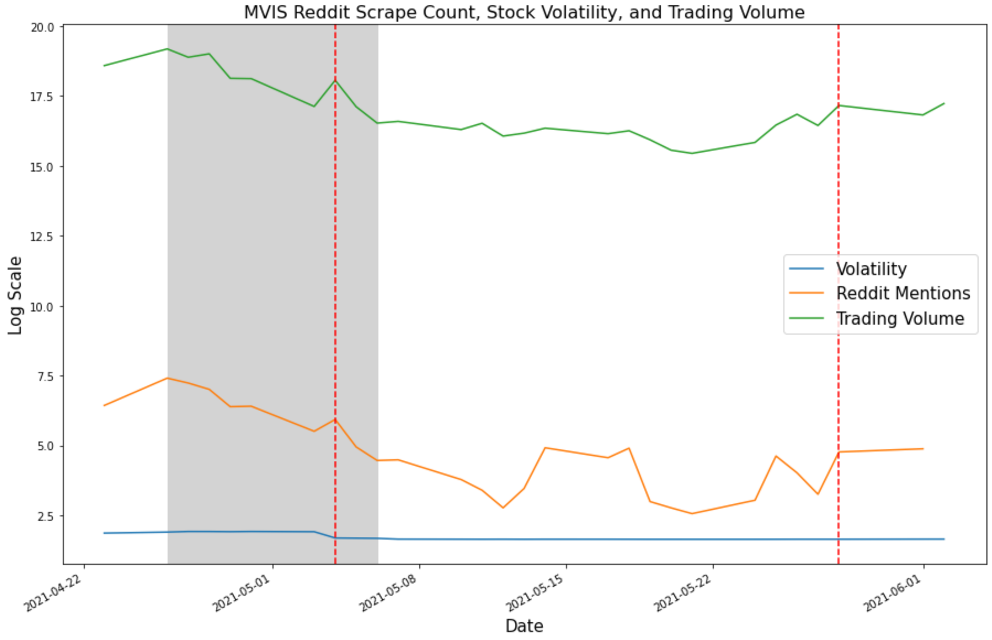


Appendix E: Log Count of Daily Mentions on r/WallStreetBets of Selected Stocks

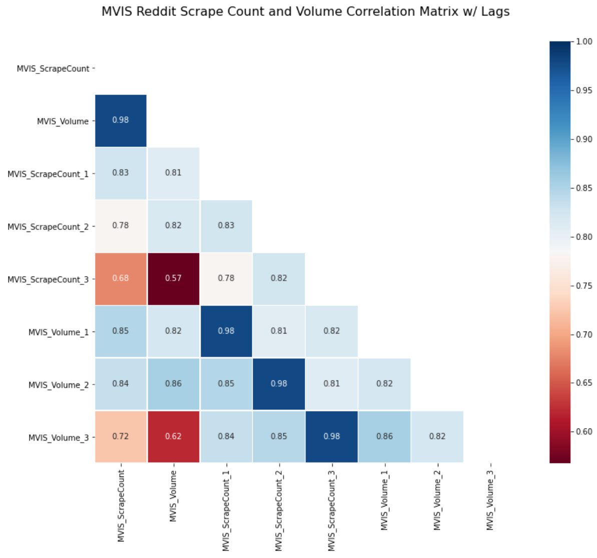


Appendix F: MVIS Results

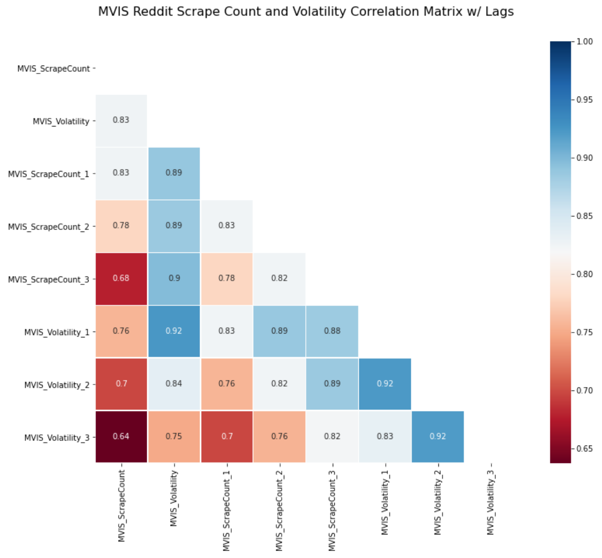
Scrape Count, Stock Volatility, and Trading Volume Over Time



Volume Correlation Matrix

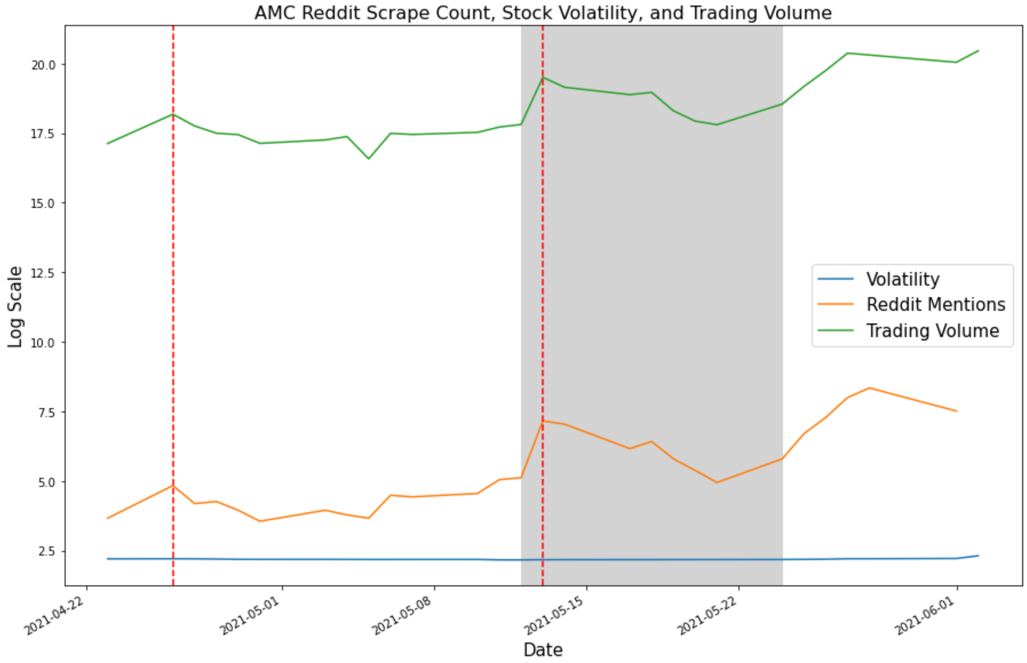


Volatility Correlation Matrix

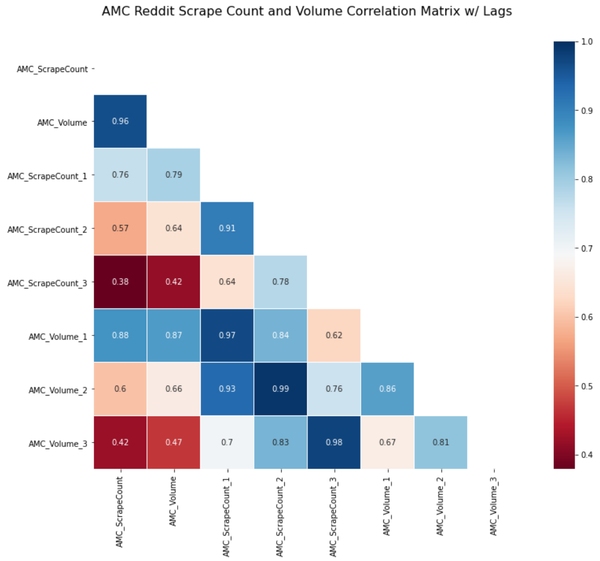


Appendix G: AMC Results

Scrape Count, Stock Volatility, and Trading Volume Over Time

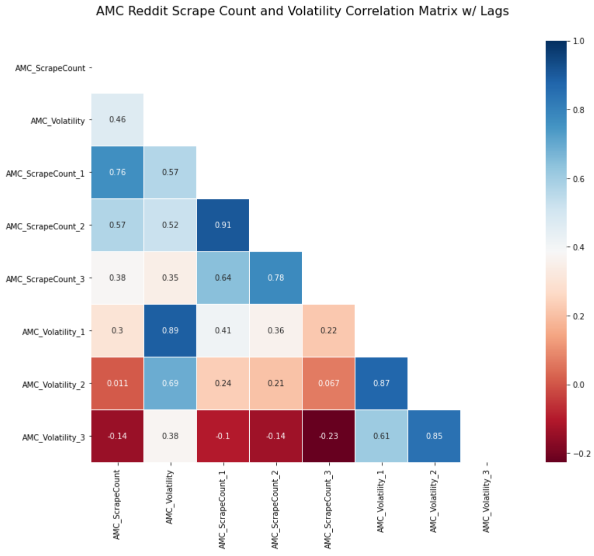


Volume Correlation Matrix



	AMC_ScrapeCount	AMC_Volume	AMC_ScrapeCount_1	AMC_ScrapeCount_2	AMC_ScrapeCount_3	AMC_Volume_1	AMC_Volume_2	AMC_Volume_3
AMC_ScrapeCount								
AMC_Volume	0.96							
AMC_ScrapeCount_1	0.76	0.79						
AMC_ScrapeCount_2	0.57	0.64	0.91					
AMC_ScrapeCount_3	0.38	0.42	0.64	0.78				
AMC_Volume_1	0.88	0.87	0.97	0.84	0.62			
AMC_Volume_2	0.6	0.66	0.93	0.99	0.76	0.86		
AMC_Volume_3	0.42	0.47	0.7	0.83	0.98	0.67	0.81	

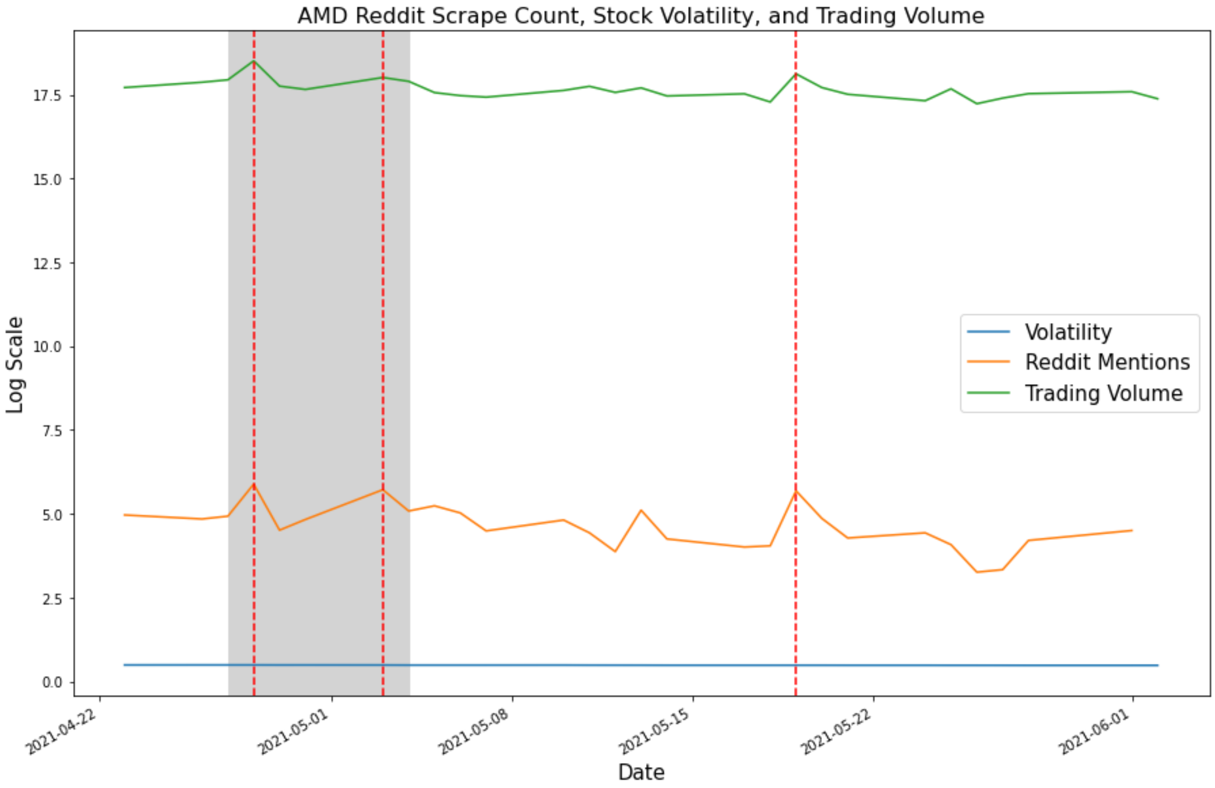
Volatility Correlation Matrix



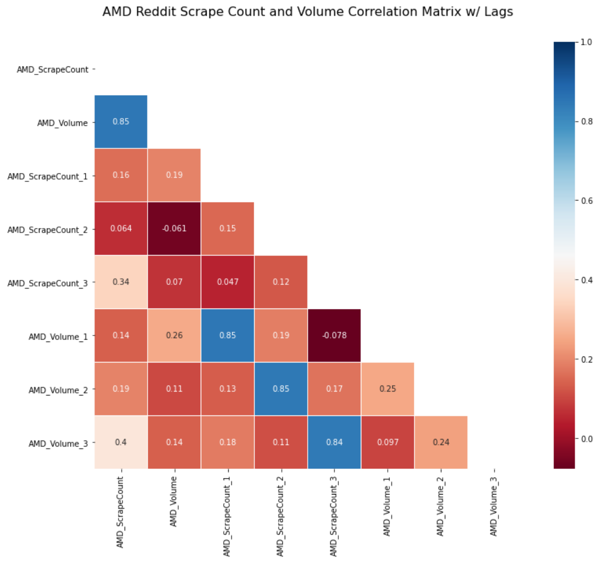
	AMC_ScrapeCount	AMC_Volatility	AMC_ScrapeCount_1	AMC_ScrapeCount_2	AMC_ScrapeCount_3	AMC_Volatility_1	AMC_Volatility_2	AMC_Volatility_3
AMC_ScrapeCount								
AMC_Volatility	0.46							
AMC_ScrapeCount_1	0.76	0.57						
AMC_ScrapeCount_2	0.57	0.52	0.91					
AMC_ScrapeCount_3	0.38	0.35	0.64	0.78				
AMC_Volatility_1	0.3	0.89	0.41	0.36	0.22			
AMC_Volatility_2	0.011	0.69	0.24	0.21	0.067	0.87		
AMC_Volatility_3	-0.14	0.38	-0.1	-0.14	-0.23	0.61	0.85	

Appendix H: AMD Results

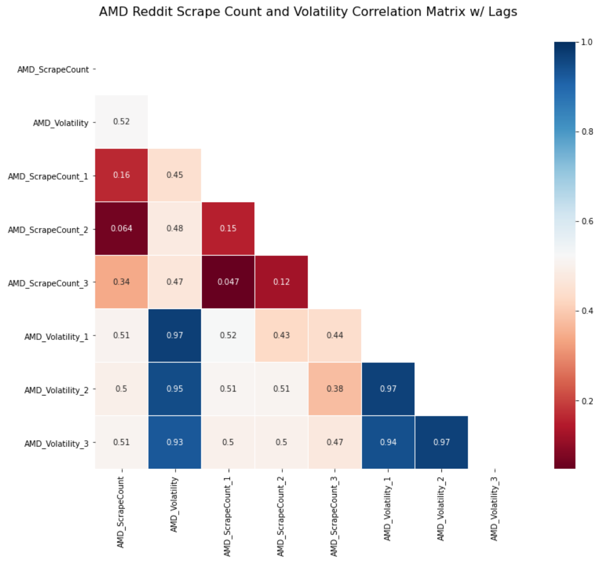
Scrape Count, Stock Volatility, and Trading Volume Over Time



Volume Correlation Matrix

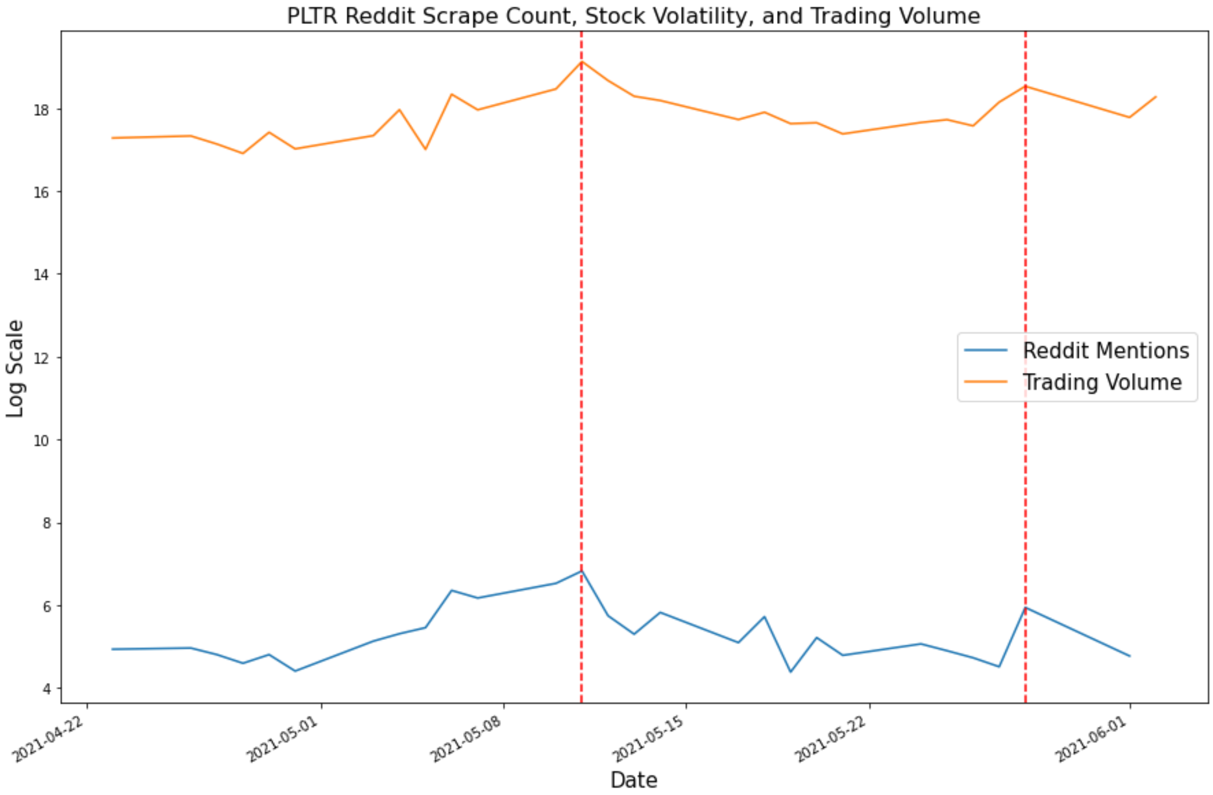


Volatility Correlation Matrix

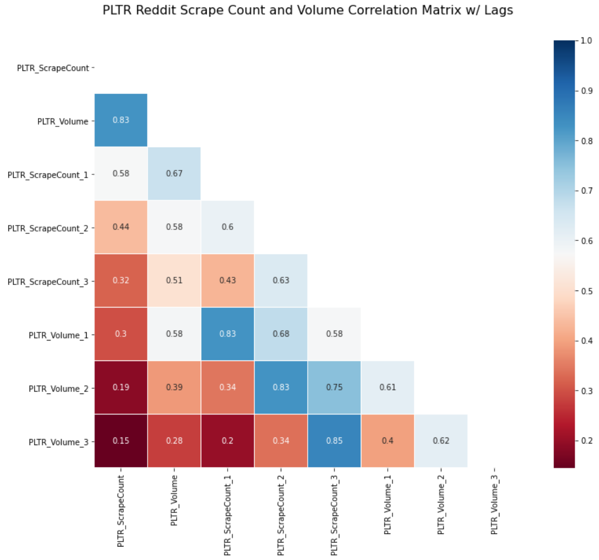


Appendix I: PLTR Results

Scrape Count, Stock Volatility, and Trading Volume Over Time



Volume Correlation Matrix

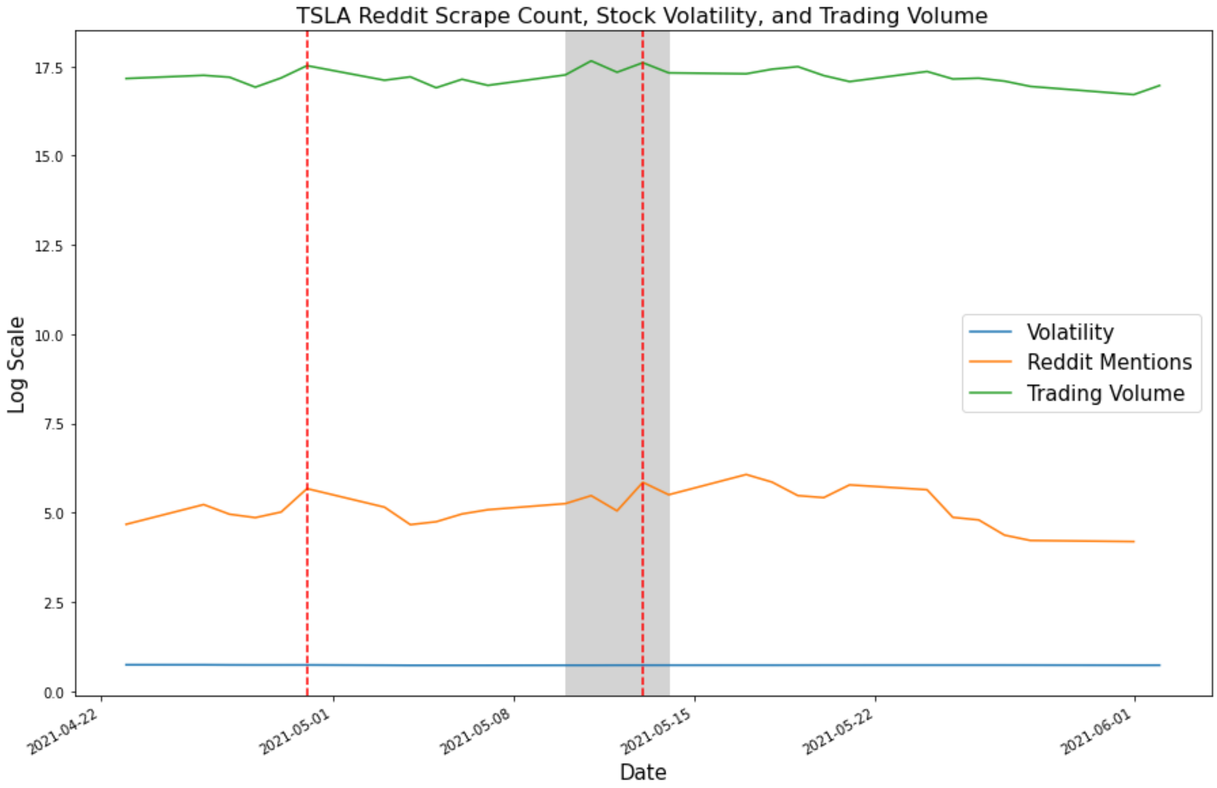


Volatility Correlation Matrix

Palantir stock has not been public long enough for an accurate calculation of annualized volatility

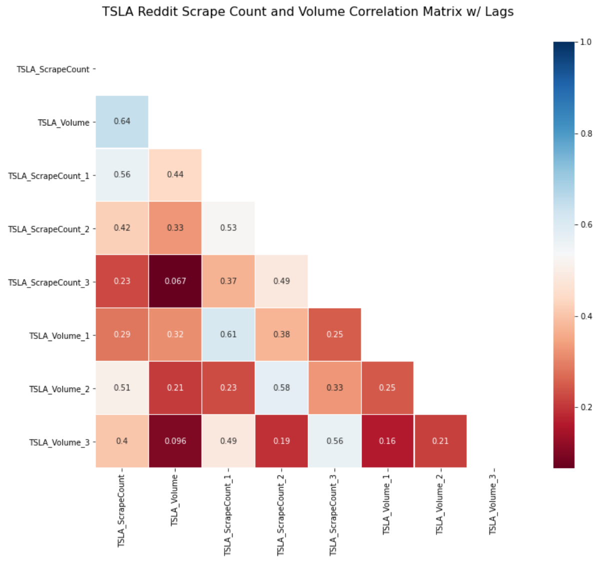
Appendix J: TSLA Results

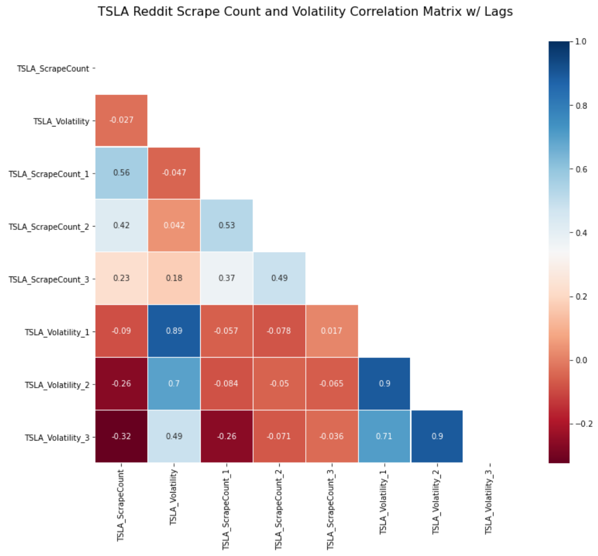
Scrape Count, Stock Volatility, and Trading Volume Over Time



Volume Correlation Matrix

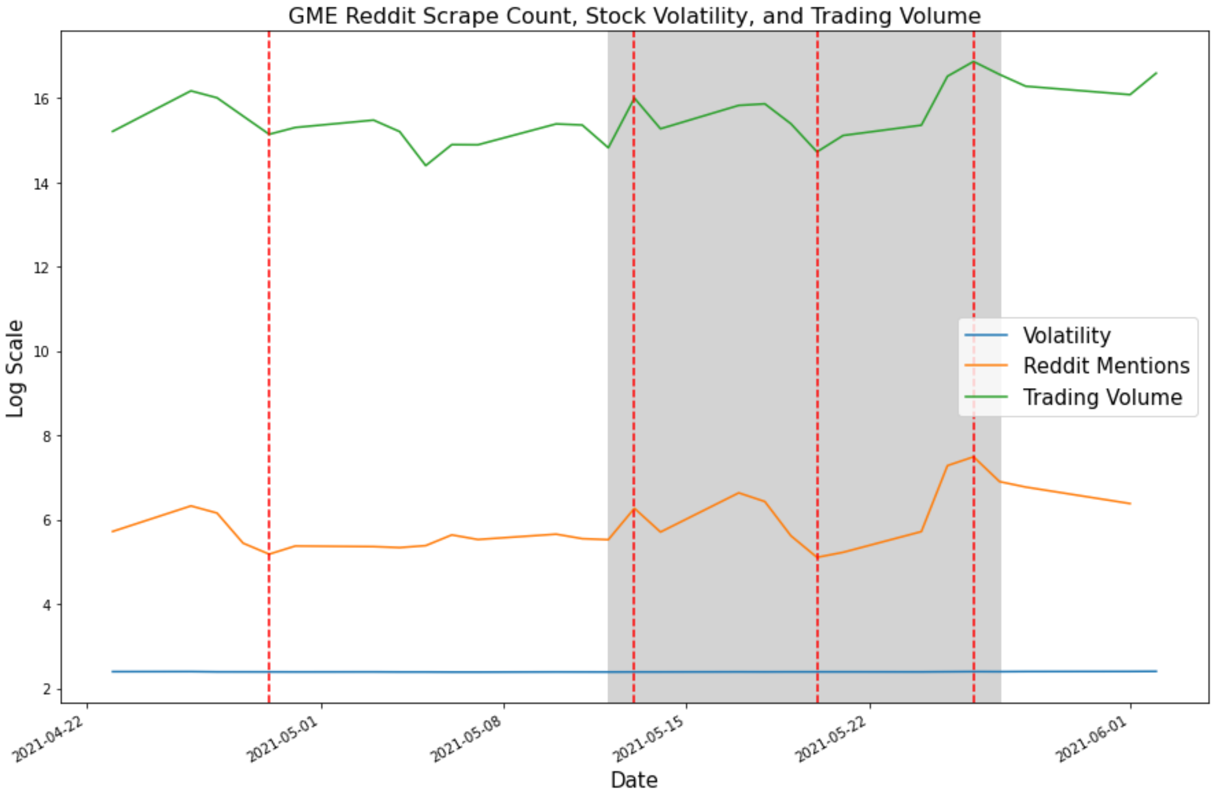
Volatility Correlation Matrix



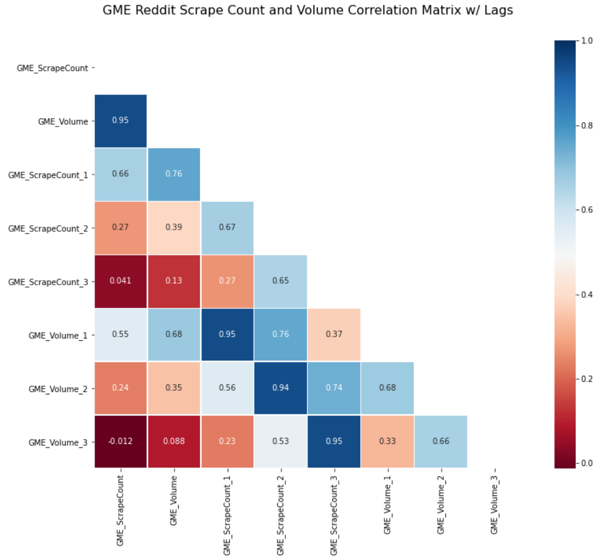


Appendix K: GME Results

Scrape Count, Stock Volatility, and Trading Volume Over Time

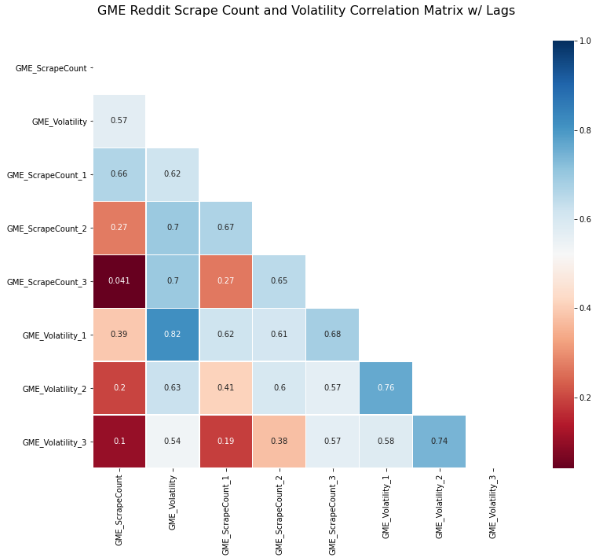


Volume Correlation Matrix



	GME_ScrapeCount	GME_Volume	GME_ScrapeCount_1	GME_ScrapeCount_2	GME_ScrapeCount_3	GME_Volume_1	GME_Volume_2	GME_Volume_3
GME_ScrapeCount								
GME_Volume	0.95							
GME_ScrapeCount_1	0.66	0.76						
GME_ScrapeCount_2	0.27	0.39	0.67					
GME_ScrapeCount_3	0.041	0.13	0.27	0.65				
GME_Volume_1	0.55	0.68	0.95	0.76	0.37			
GME_Volume_2	0.24	0.35	0.56	0.94	0.74	0.68		
GME_Volume_3	-0.012	0.088	0.23	0.53	0.95	0.33	0.66	

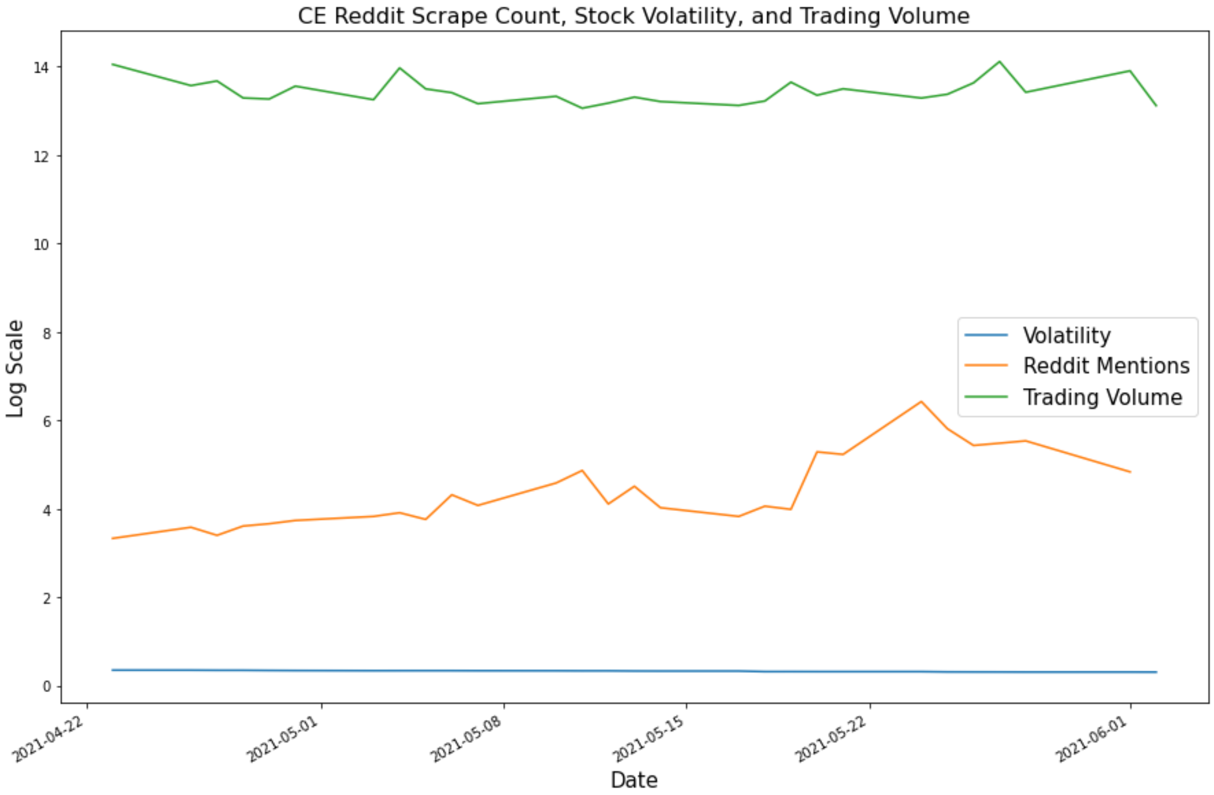
Volatility Correlation Matrix



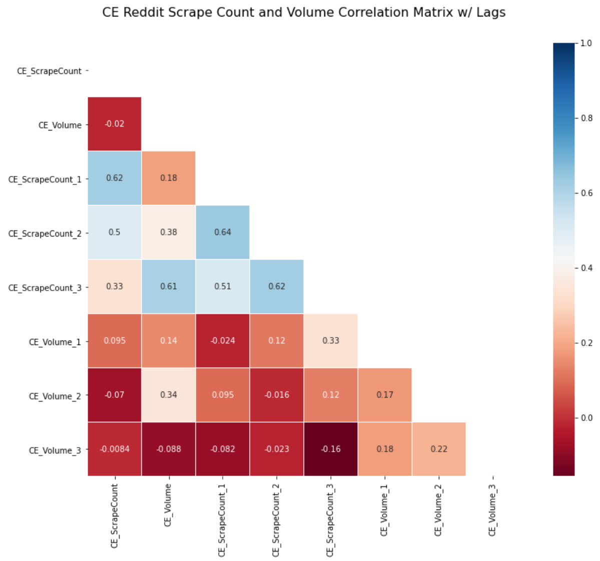
	GME_ScrapeCount	GME_Volatility	GME_ScrapeCount_1	GME_ScrapeCount_2	GME_ScrapeCount_3	GME_Volatility_1	GME_Volatility_2	GME_Volatility_3
GME_ScrapeCount								
GME_Volatility	0.57							
GME_ScrapeCount_1	0.66	0.62						
GME_ScrapeCount_2	0.27	0.7	0.67					
GME_ScrapeCount_3	0.041	0.7	0.27	0.65				
GME_Volatility_1	0.39	0.82	0.62	0.61	0.68			
GME_Volatility_2	0.2	0.63	0.41	0.6	0.57	0.76		
GME_Volatility_3	0.1	0.54	0.19	0.38	0.57	0.58	0.74	

Appendix L: CE Results

Scrape Count, Stock Volatility, and Trading Volume Over Time

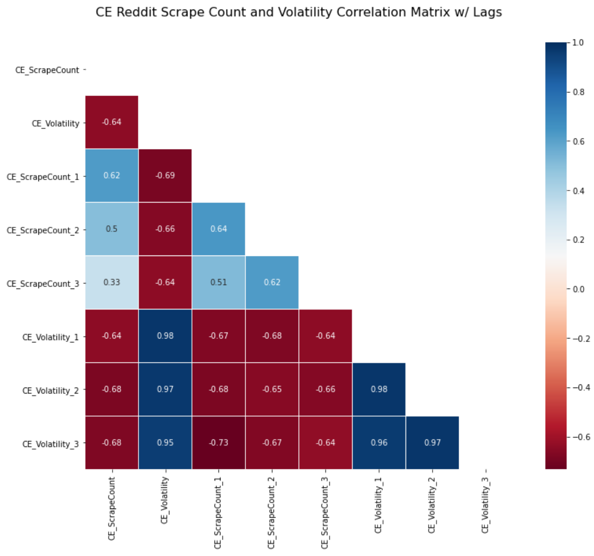


Volume Correlation Matrix



	CE_ScrapeCount	CE_Volume	CE_ScrapeCount_1	CE_ScrapeCount_2	CE_ScrapeCount_3	CE_Volume_1	CE_Volume_2	CE_Volume_3
CE_ScrapeCount								
CE_Volume	-0.02							
CE_ScrapeCount_1	0.62	0.18						
CE_ScrapeCount_2	0.5	0.38	0.64					
CE_ScrapeCount_3	0.33	0.61	0.51	0.62				
CE_Volume_1	0.095	0.14	-0.024	0.12	0.33			
CE_Volume_2	-0.07	0.34	0.095	-0.016	0.12	0.17		
CE_Volume_3	-0.0084	-0.088	-0.082	-0.023	-0.16	0.18	0.22	

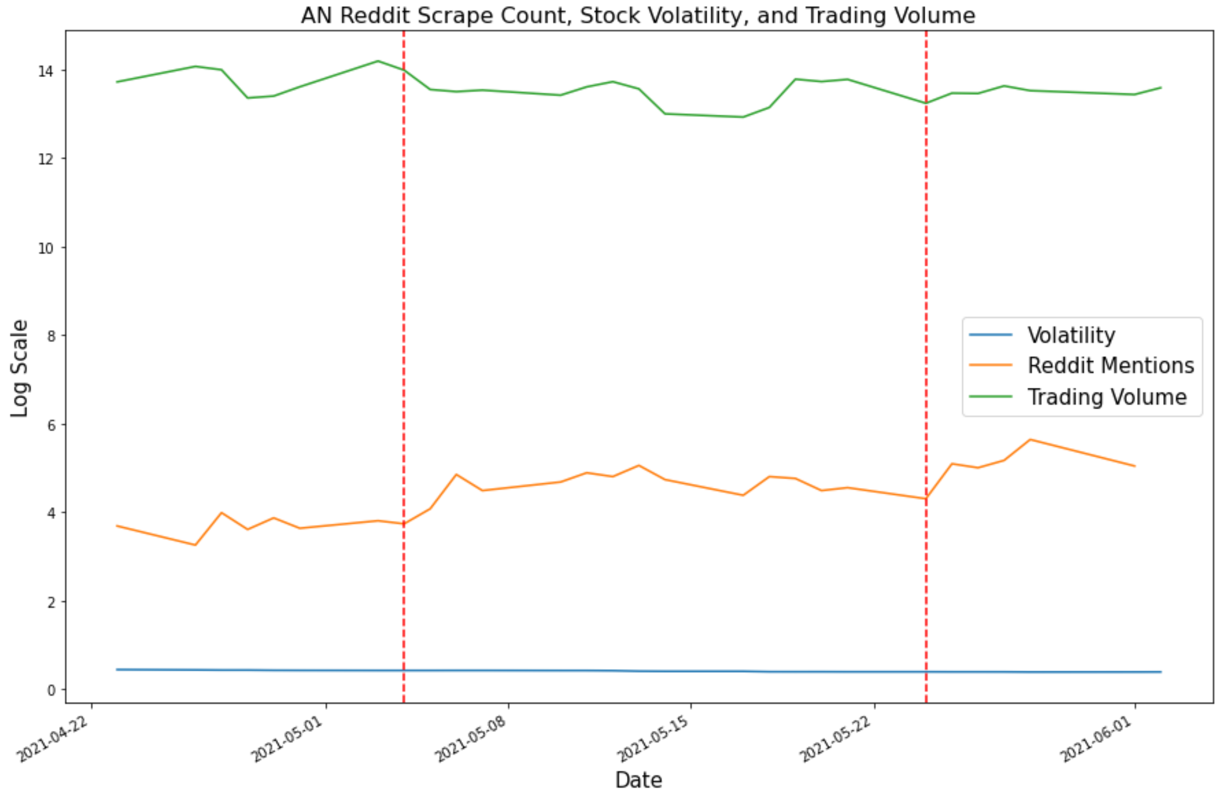
Volatility Correlation Matrix



	CE_ScrapeCount	CE_Volatility	CE_ScrapeCount_1	CE_ScrapeCount_2	CE_ScrapeCount_3	CE_Volatility_1	CE_Volatility_2	CE_Volatility_3
CE_ScrapeCount								
CE_Volatility	-0.64							
CE_ScrapeCount_1	0.62	-0.69						
CE_ScrapeCount_2	0.5	-0.66	0.64					
CE_ScrapeCount_3	0.33	-0.64	0.51	0.62				
CE_Volatility_1	-0.64	0.98	-0.67	-0.68	-0.64			
CE_Volatility_2	-0.68	0.97	-0.68	-0.65	-0.66	0.98		
CE_Volatility_3	-0.68	0.95	-0.73	-0.67	-0.64	0.96	0.97	

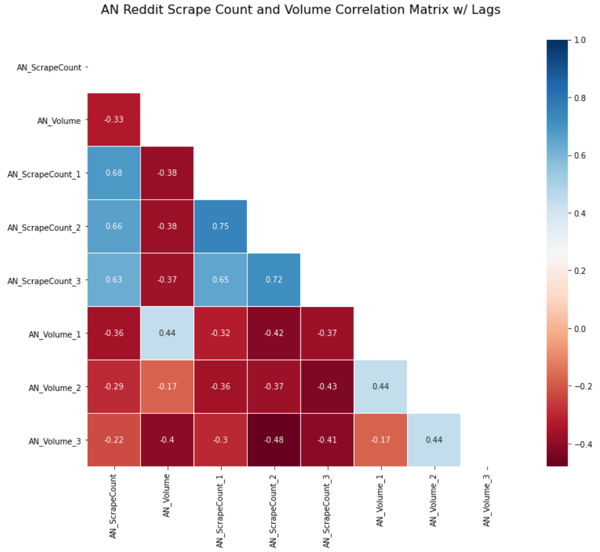
Appendix M: AN Results

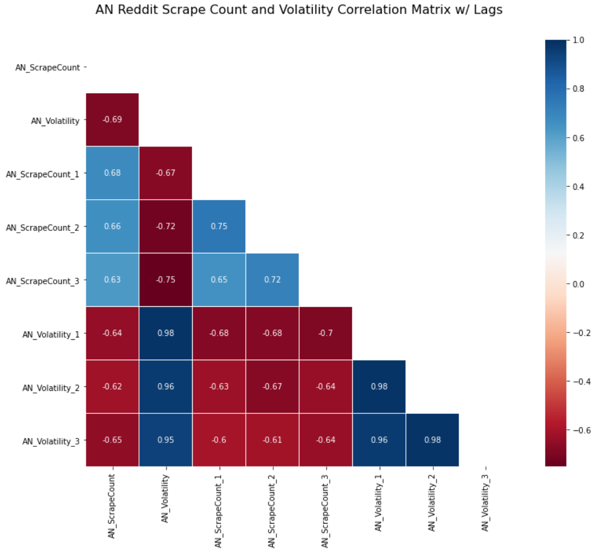
Scrape Count, Stock Volatility, and Trading Volume Over Time



Volume Correlation Matrix

Volatility Correlation Matrix


AN Reddit Scrape Count and Volume Correlation Matrix w/ Lags
AN_ScrapeCount
AN_Volume
AN_ScrapeCount_1
AN_ScrapeCount_2
AN_ScrapeCount_3
AN_Volume_1
AN_Volume_2
AN_Volume_3


AN Reddit Scrape Count and Volatility Correlation Matrix w/ Lags
AN_ScrapeCount
AN_Volatility
AN_ScrapeCount_1
AN_ScrapeCount_2
AN_ScrapeCount_3
AN_Volatility_1
AN_Volatility_2
AN_Volatility_3