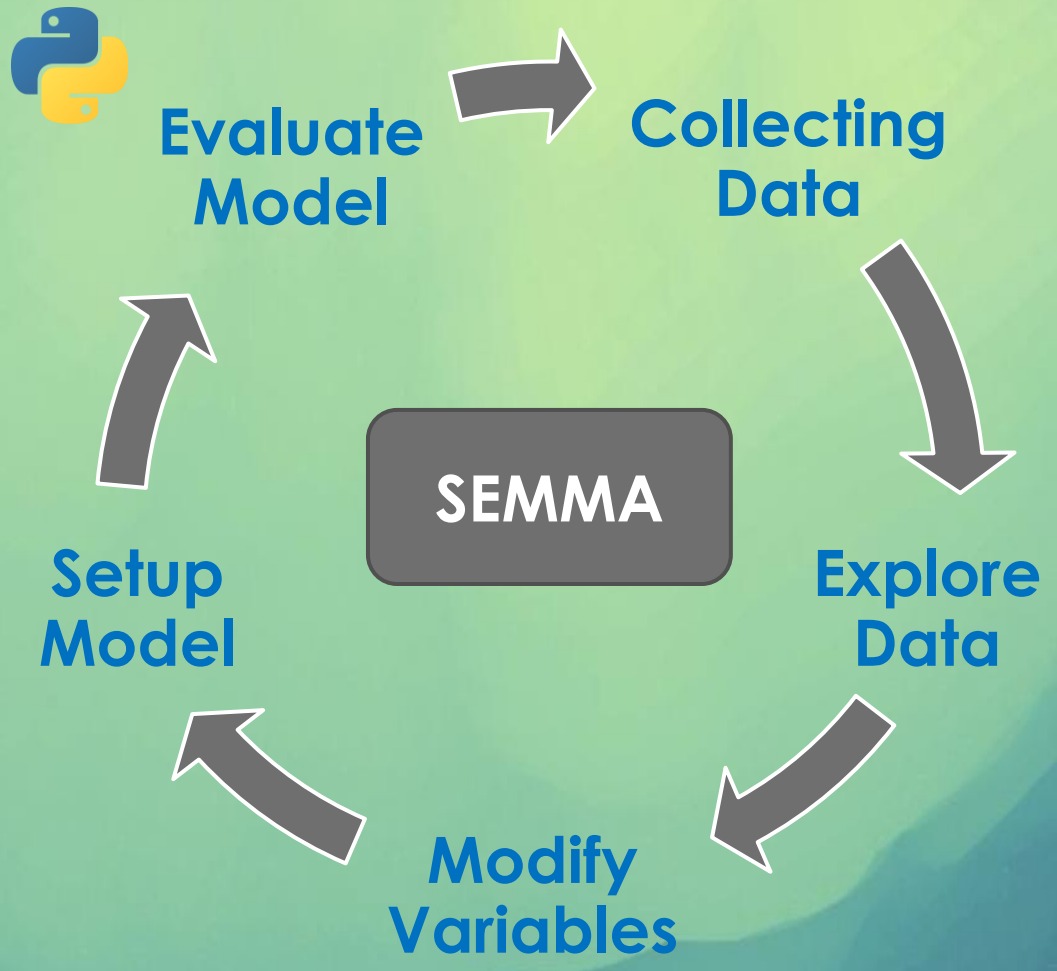
BAN204 – FINAL PROJECT

CLUSTERING ANALYSIS ETH AND S&P 500 COMPANIES

JINGYI WANG

Ethereum is up about 81% since Jan. 1st of this year, well above Bitcoin's 26%.
As someone who is new to the US stock market and the cryptocurrency world, I was curious about which public companies would have a strong correlation with the rise and fall of Ether, and about people's attitudes towards it.

SUMMARY



My hypothesis is that companies that are highly correlated will be affected by the same positive or negative news, and thus they should converge in market performance.

I hope to find these companies by organizing the collected data into meaningful indicators and performing unsupervised cluster analysis based on them.

Theoretically, we can find price pockets or investment opportunities by comparing the market performance of these companies in the short term, which can also help us diversify risk when choosing an asset portfolio.

SUMMARY

For Clustering analysis:

- I selected historical trading data for 504 S&P 500 companies and ETH-USD
- attributes including daily adj Close, Close Price, High, Low, Open, and Volume.
- Training set (9/1/2021 to 11/31/2021)
  Testing set (1/1/2021 to 2/28/2021)
- Only based on the market performance of stock prices and does not reflect the fundamental dimension of these companies.

For Social Media Analysis:

I filtered 10,000 twitter posts containing the words "ETH", "Ethereum".

DATASET SELECTION

STEP 1: I first found a list of all the S&P 500 stock symbols from the Wikipedia; (urllib.request)

STEP 2: Collect data on the Yahoo for two different time periods; (pandas_datareader)

STEP 3: Remove missing data; (pandas, numpy)

- 63*3030, two hierarchy levels of indexes
- two stocks lacking any data.

STEP 4: Split the raw data into 6 tables according to different attributes.
Take Open.csv for example:

	Attributes	Adj Close	Adj Close.1	Adj Close.2
0	Symbols	MMM	ABT	ABBV
1	Date	nan	nan	nan
2	9/1/2021	191.8784637	126.909996	110.9251938
3	9/2/2021	193.3167572	127.9958115	110.6781921
4	9/3/2021	192.8208008	128.4440765	110.2829895
5	9/7/2021	184.0819244	127.8961868	107.7240067
6	9/8/2021	186.3732758	128.5636139	109.0380783
7	9/9/2021	184.2505493	127.4877701	105.9850845

	MMM	ABT	ABBV	ABMD
9/1/2021	194.5500031	125.7200012	121	365.3900146
9/2/2021	194.0700073	127.5	113.5500031	366.1700134
9/3/2021	194.4600067	129	111.1800003	368.269989
9/7/2021	192.9199982	128.1199951	110.4499969	371.9200134
9/8/2021	185.6399994	128.5599976	108.0599976	375.269989
9/9/2021	187.5399933	129.1499939	109.4400024	374.480011
9/10/2021	186.8600006	128	107.4800034	368.3699951
9/13/2021	185.3000031	129.6699982	107.3899994	362.1700134
9/14/2021	185.6499939	127.3499985	107.9899979	354.519989

DATA PREPROCESSING

STEP 5: Calculate the desired indicators
Volatility Rate = (High – Low)/Open
Movement Rate = (Today's Close Price – Yesterday's Close Price)/Yesterday's Close Price
Opening Performance = (Today's Open Price – Yesterday's Close Price)/Yesterday's Close Price

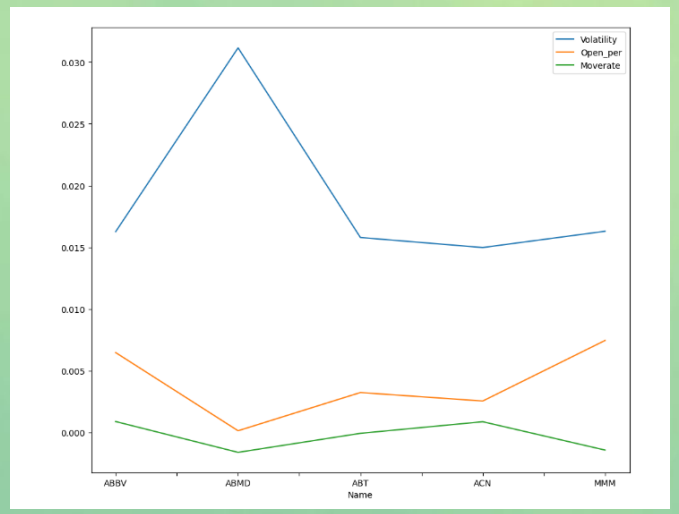
STEP 6: Combine the above three tables together and flatten the table
The final table was obtained in the following format: 30269 rows by 5 columns

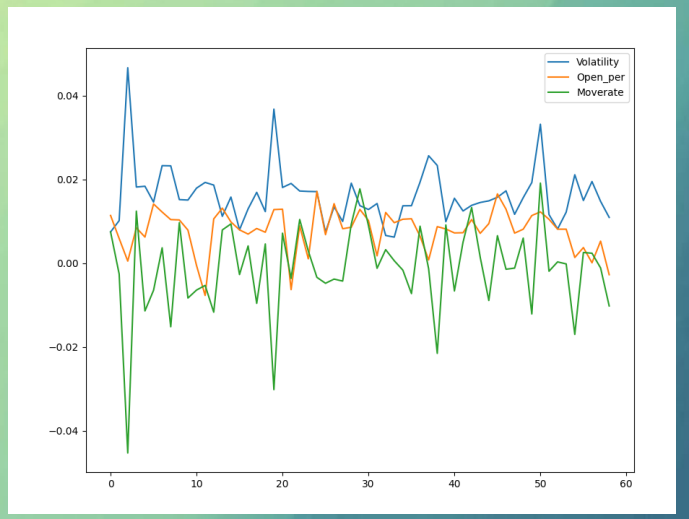
STEP 7: Prepare the testing set in the same way as above (step 3 to step 6).

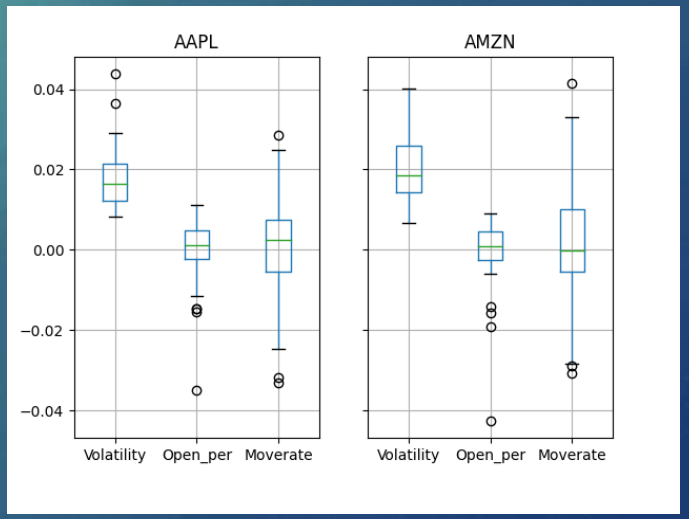
	Name	Date	Volatility	Open_per	Moverate
0	MMM	9/2/2021	0.00747	0.01142	0.00750
1	MMM	9/3/2021	0.01018	0.00591	-0.00257
2	MMM	9/7/2021	0.04670	0.00051	-0.04532
3	MMM	9/8/2021	0.01821	0.00846	0.01245
4	MMM	9/9/2021	0.01840	0.00626	-0.01139
5	MMM	9/10/2021	0.01461	0.01416	-0.00646
6	MMM	9/13/2021	0.02331	0.01224	0.00368
7	MMM	9/14/2021	0.02327	0.01042	-0.01517
8	MMM	9/15/2021	0.01521	0.01035	0.00970

DATA PREPROCESSING

1. Descriptive Analysis:







- First chart shows that there is a large difference between the indicators of different stocks (any 5 stocks) on the same day
- The second chart shows the performance of the stock MMM over 60 trading days, and we can pay particular attention to these peaks.
- Hoteling T/MANOVA：Using boxplots to compare market performance between stocks
  the distribution of AAPL and AMZN on the three indicators over a 60-day period.

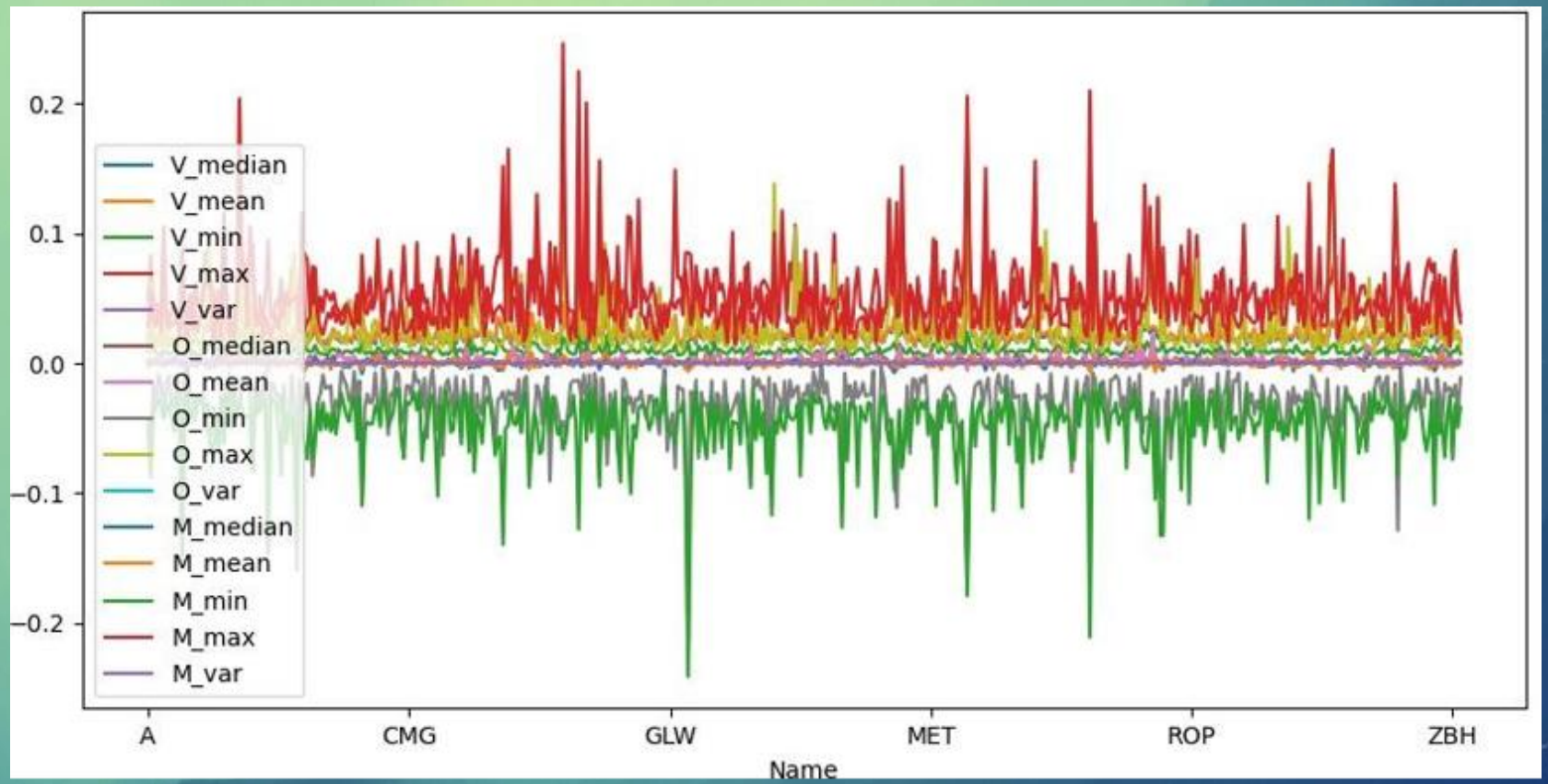
ANALYSIS

2. Clustering Analysis: K-means

- Flatten the dataset:

  Convert time series to: max, min, mean, median, var;

- perform cluster analysis on this basis



ANALYSIS

Elbow Method:

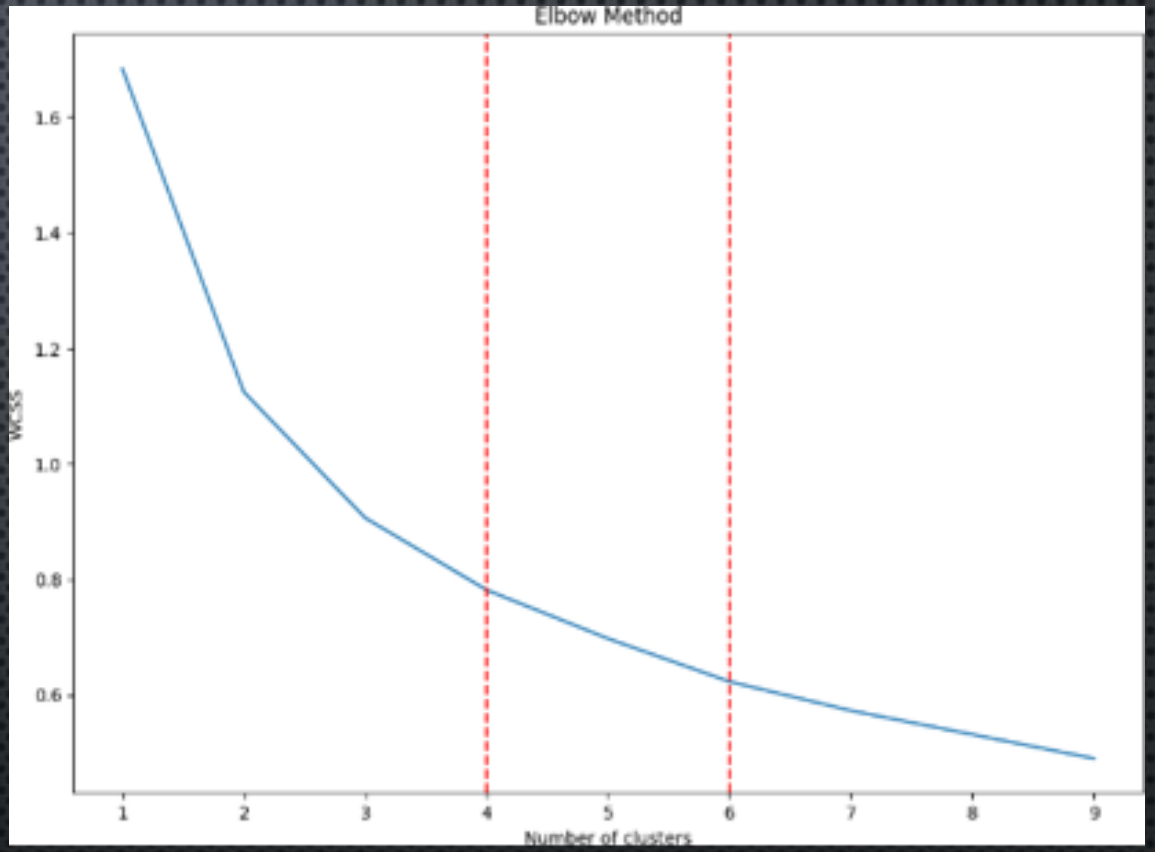
Based on the Elbow Method, I decided to try to divide the data into 4 groups (K=4) or 6 groups (K=6)

K-Means:

Find the set of stocks with the same classification as ETH-USD and take the merged set to get a total of 22 stocks

Multiple experiments:

Repeating the previous steps using testing set, obtain 20 stocks, 18 of which were the same as the original classification.

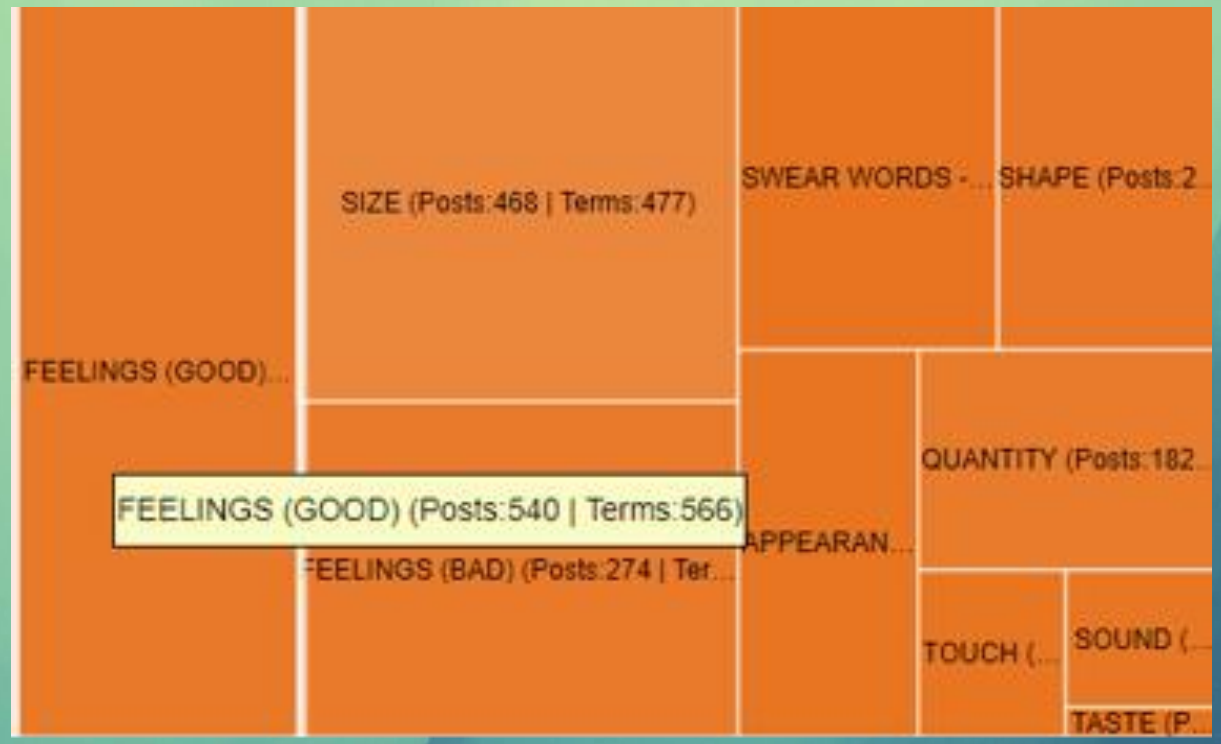


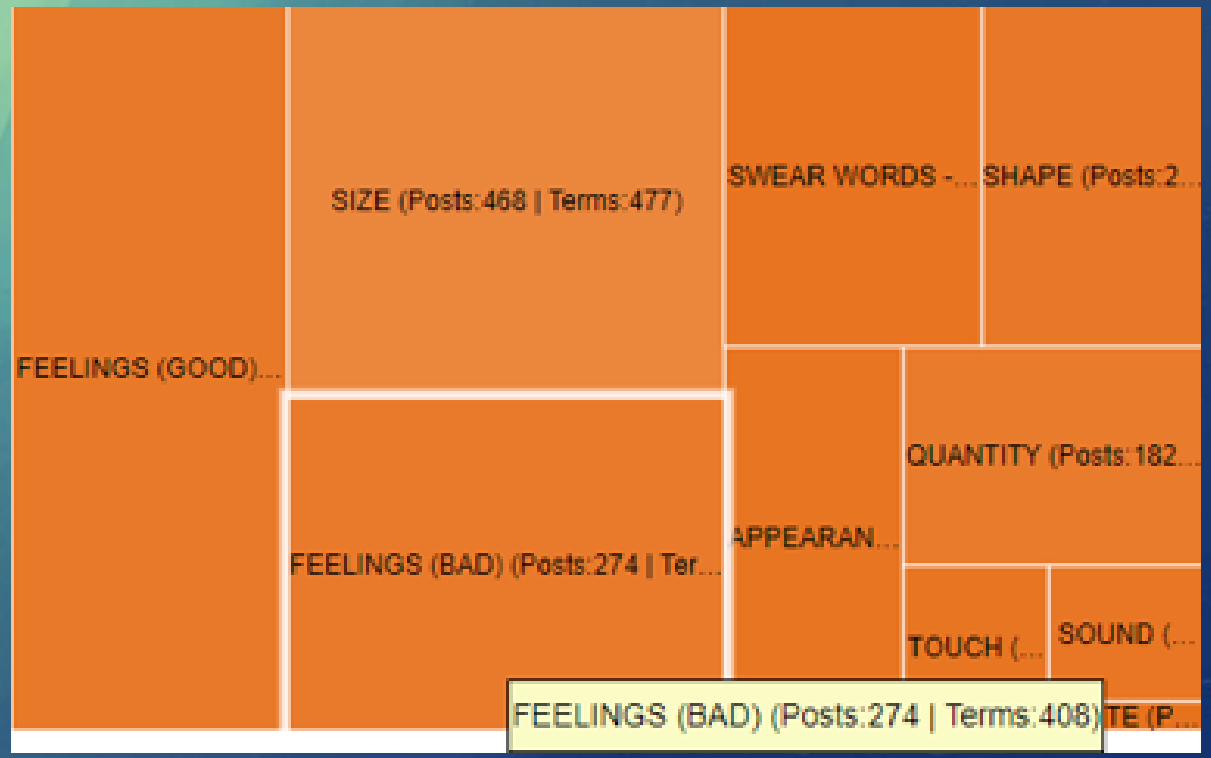
ANALYSIS

3. Social Media Analysis using Netlytic:

- Text Analysis:
  There were more positive feelings being expressed in these posts:

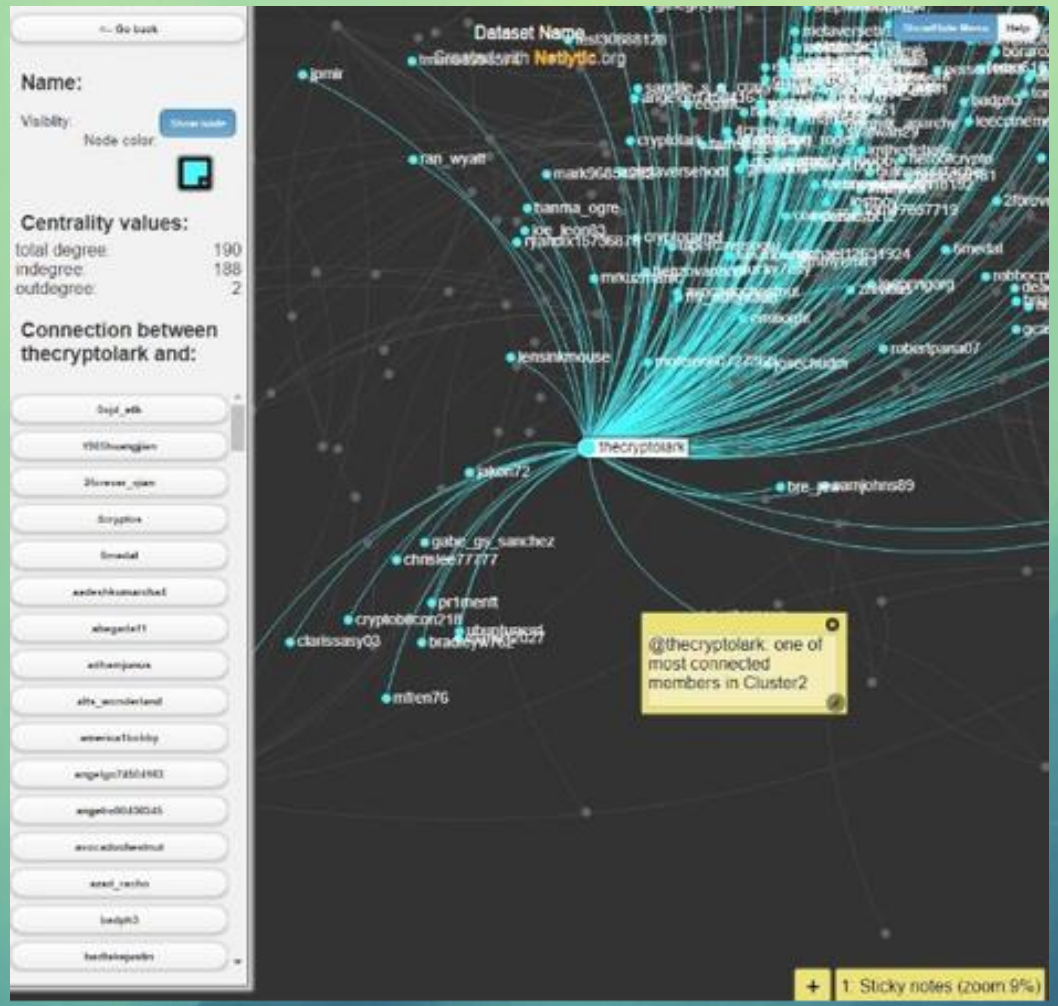
540 Positive Posts vs 274 Negative Posts





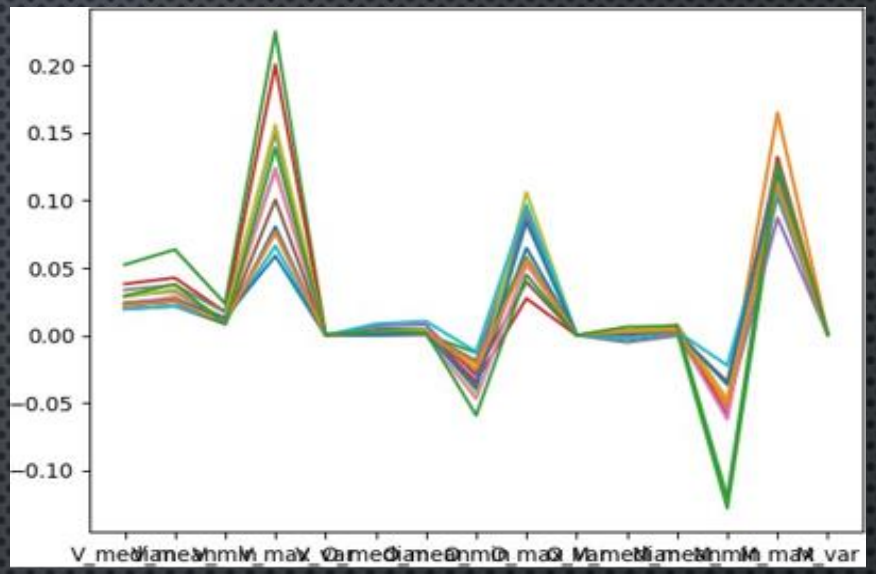
ANALYSIS

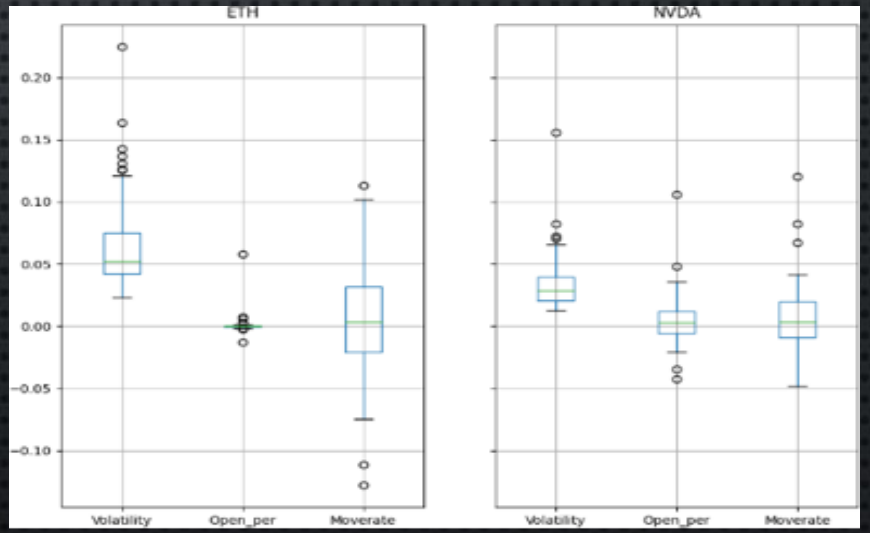
- Network Analysis: I found a few interesting twitter users :)



ANALYSIS

- The clustered stocks show a clear trend towards similar market performance relative to the initial haphazard charts.
- The classification is stable that the same stocks do not have significantly different clustering results depending on the time period in which the data was collected.
- I do find a clear chain of industries related to blockchain computing resources.
  - NVDA (NVIDIA): a fabless semiconductor company that designs and sells graphics processors;
  - AMD (Advanced Micro Devices, Inc.): that develops computer processors and related technologies;
  - TSLA is also in the list.





WHAT DID I FIND?

- In this list, there are also stocks that are not so much related to Ether, such as "DLTR" (Dollar Tree):

➢ It is said that there is an aggressive investor who causes its share price to fluctuate abnormally, which means it could be a coincidence due to other factors.

- We can conclude by saying that all these companies have had relatively large stock movements in recent times and have shown a clear upward trend overall.

- final_list: ['BEN', 'DLTR', 'ETH', 'ETSY', 'AMD', 'FMC', 'LUMN', 'MTCH', 'NVDA', 'PFE', 'PWR', 'TER', 'TSLA']

CONCLUSIONS

REFERENCES

Cluster analysis on stock selection
https://towardsdatascience.com/clustering-analysis-on-stock-selection-2c2fd079b295

Evolution of Financial Time Series Clusters.
Azzalini, D., Azzalini, F., Mazuran, M., & Tanca, L. (2019). In SEBD.
http://ceur-ws.org/Vol-2400/paper-12.pdf

Clustering stock price time series data to generate stock trading recommendations: An empirical study. Binoy B. Niar, P.K. Saravana Kumar, N.R. Sakthivel, U.Vipin. (2017)
https://doi.org/10.1016/j.eswa.2016.11.002

Stock Clustering with Time Series Clustering in R
https://medium.com/@panda061325/stock-clustering-with-time-series-clustering-in-r- 63fe1fabe1b6

Step by Step: Twitter Sentiment Analysis in Python
https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58d

Ethereum USD (ETH-USD) Price, Value, News & History
https://finance.yahoo.com/quote/ETH-USD/

Appendices
1. python scripts

Packages including:
- urllib.request,
- ssl,
- pandas_datareader,
- pandas,
- numpy,
- datetime,
- sklearn.cluster
- matplotlib.pyplot.