HOFSTRA UNIVERSITY
FRANK G. ZARB SCHOOL OF BUSINESS

BAN204 – Final Project Cover Page

Student Name	**Last four digits of ID #**
Jingyi Wang	*8628*

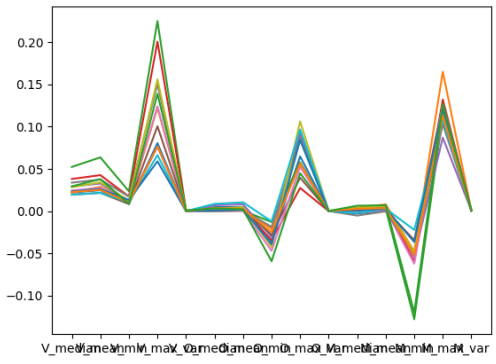
I hereby certify that I am the author of this document and all sources used in the preparation of this assignment have been cited in accordance with Hofstra's Code of Student Conduct (available at https://www.hofstra.edu/pdf/studentaffairs/deanofstudents/commstandards/commstandards_guidetopride.pdf*) directly or paraphrased in the document. Sources are properly credited according to accepted standards for professional publications. I also certify that this paper was prepared by me (all group members if it is a group paper) for this purpose.*

PAGE 2: Summary

Unlike many people, my interest in Ethereum was sparked by the concept of a "metaverse". As the most well-known application of Ethereum, NFTs can essentially function as a unique code connected to a digital file, allowing to distinguish the original file from any potential copies. It has thus become the cornerstone of many metaverse projects.

Ether 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 Ethereum, and about people's attitudes towards Ethereum. In this Python-based project, I have collected and organized historical data from S&P 500 companies and ETH-USD. I normalized the raw price data to remove as much price stickiness as possible using indicators such as Volatility Rate, Movement Rate, and Opening Performance. After that, by using Elbow Method and K-Means Clustering, all the observers were divided into 4 and 6 groups, and 22 stocks with the most similar performance to ETHUSD were successfully found.



In addition, I learned about people's attitudes towards the term Ethereum through Netlytic and found several twitter users who are highly followed, which allowed me to learn more quickly about Ethereum-related knowledge and follow their analysis of the Ethereum market.

PAGE 3: Problem Statement and Dataset Selection

Note: Discuss what dataset you selected and why. What real-world problem did you try to address? What hypothesis did you plan to test with the dataset and why?

For Clustering analysis, I selected historical trading data for 504 S&P 500 companies and ETH-USD, with attributes including daily adj Close, Close Price, High, Low, Open, and Volume. The data set was divided into a training set (9/1/2021 to 11/31/2021), and a test set (1/1/2021 to 2/28/2021).

I understand that both stock prices and ETH-USD price movements are very complex and they are affected by so many factors that it is difficult for even the most sophisticated prediction algorithms to accurately predict prices. The problem I want to solve is not to make price predictions, but to understand which public companies have close correlation with ETH, which allows me to get multiple perspectives on information related to Ethereum and those companies that are closely linked to the cryptocurrency.

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.

For Social Media Analysis, I filtered 10,000 twitter posts containing the words "ETH", "Ethereum".
My goal is to understand the public's attitude towards Ethereum, and to find the most popular twitter users through the network analysis, which will help me to speed up my learning process and keep up with the latest information.

PAGE 4: Previous work / Literature review

Note: Discuss/cite possible approaches to solving the problem discussed in the previous section with references to relevant literature (e.g., scholarly publications, white papers, industry reports, expert blogs, etc.). It is recommended to cite at least 5 external sources.

Was the selected dataset used by other researchers/organizations and for what purposes?

I started by learning about some stock analysis terminology, and related articles include：

1. Stock Trading Terms，

   https://www.marketbeat.com/financial-terms/；

   https://www.timothysykes.com/blog/trading-terms-you-need-to-know/

2. Forces That Move Stock Prices，

   https://www.investopedia.com/articles/basics/04/100804.asp；

I also learned about some of the issues that need to be considered when performing stock analysis, such as price stickiness and trends, and learned some common models used in stock analysis and their principles, including Moving Average, Linear Regression, K-Nearest Neighbors, Auto ARIMA，and etc.（Stock Prices Prediction Using Machine Learning and Deep Learning Techniques，https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/）

In addition, I also refer to many cases of stock clustering analysis, which are basically limited to clustering between stocks, and there is no case of clustering analysis between stocks and cryptocurrencies on blockchain. Due to the specificity of blockchain, I had to abandon attributes such as "Debt To Equity", "Debt To Assets", "Dividend Yield " and so on. Therefore, this cluster analysis is only based on the market performance of stock prices and does not reflect the fundamental dimension of these companies.

PAGE 5: Data Preprocessing

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

STEP 2: All historical transaction data on this list was collected on the Yahoo for two different time periods. I used nearly 60 days of data (2021-9-1 to 2021-11-30) as the training set and 40 days of data from 2021-1-1 to 2021-2-28 as the testing set. Although I use an unsupervised model, I want to verify the stability of the classification by using a test set.

STEP 3: Take the Training set as an example, the original data format is 63*3030, with two hierarchy levels of INDEX, in which there are two stocks lacking any data. I first removed the stocks with no trading data (the columns are all NA), and made sure there is no other missing data.

STEP 4: The original table is split into 6 tables according to different attributes. Take Open.csv for example, this table shows all Open Prices for all stocks (including ETHUSD) for the 60 trading days.

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

	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

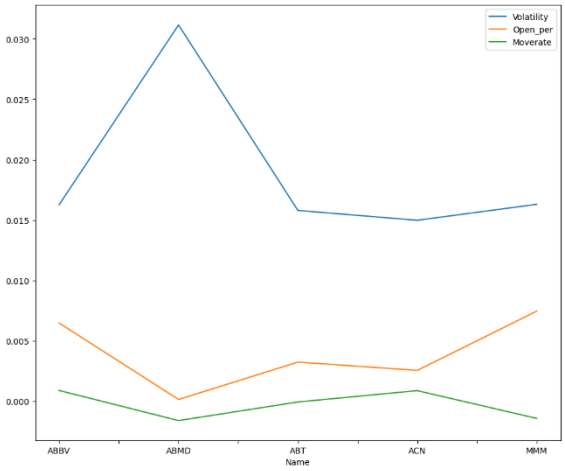
STEP 7: Prepare the testing set in the same way as above (step 3 to step 6).

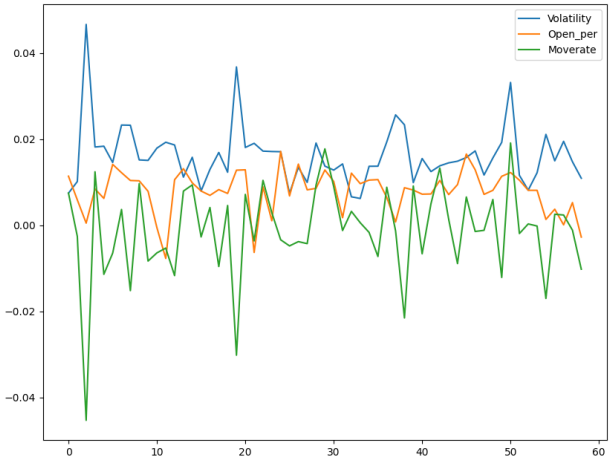
PAGE 6-8: Analysis

- What Data Mining methods did use and why? Did you run multiple experiments?

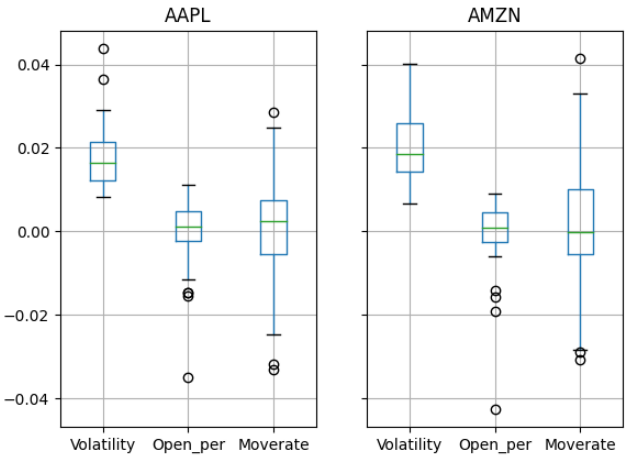
  Python:

  1. Descriptive Analysis: With visualisation, it is easier to compare the dynamics of multiple stocks visually or to identify unusual trading days. For example, the chart on the left shows that there is a large difference between the indicators of different stocks (any 5 stocks) on the same day, while the chart on the right shows the performance of the stock MMM over 60 trading days, and we can pay particular attention to these peaks.



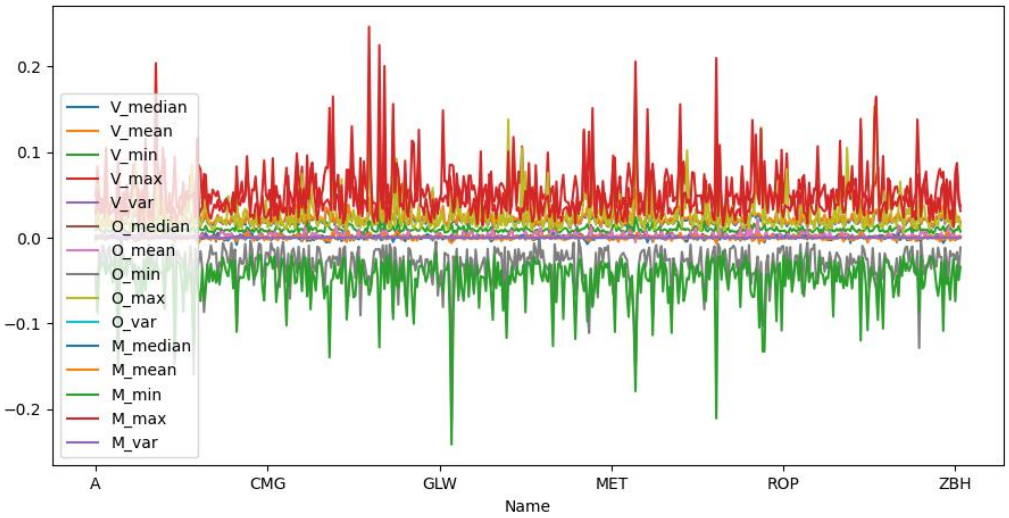


We can also use ANOVA, and boxplots to compare market performance between stocks: the graph on the right shows the distribution of AAPL and AMZN on the three indicators over a 60-day period.



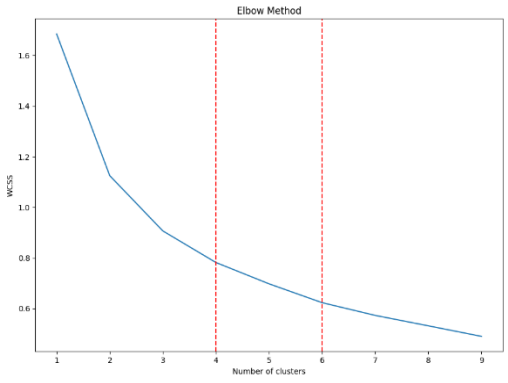
2. Clustering Analysis: K-means

- Flatten the dataset: Convert time series to: max, min, mean, median, var; and perform cluster analysis on this basis



- 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:

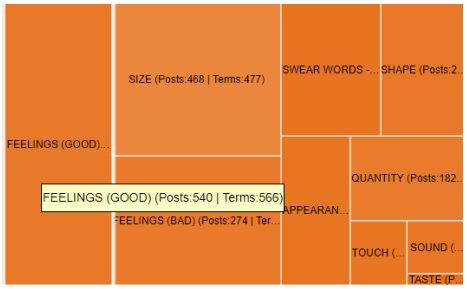
  Find the set of stocks with the same classification as ETH-USD and take the merged set to get a total of 22 stocks

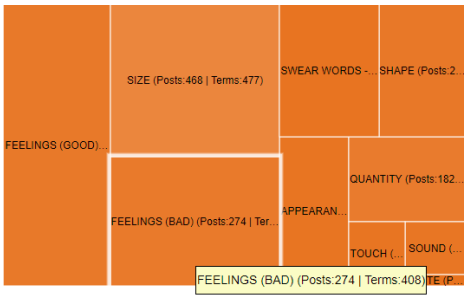
- Multiple experiments: Repeating the previous steps, a total of 40 days of data (K=4) between January and February of this year were used to obtain 20 stocks, 18 of which were the same as the original classification.

3. Social Media Analysis using Netlytic:

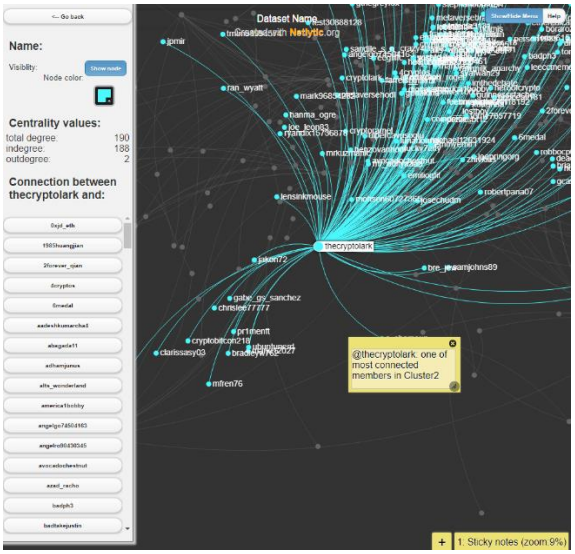
1. Text Analysis:

   There were more positive feelings being expressed in these posts:
   （540 Positive Posts vs 274 Negative Posts）

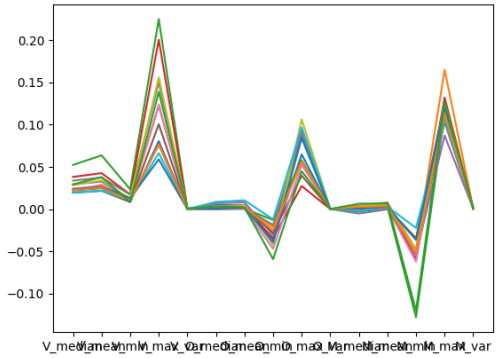




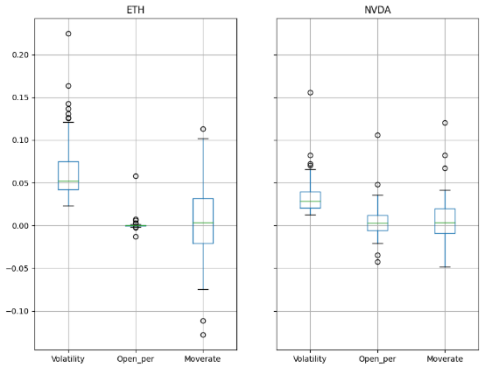
2. Network Analysis: I found a few interesting twitter users that I follow :)



- How did you evaluate your results?



The clustered stocks show a clear trend towards similar market performance relative to the initial haphazard charts. In addition, the results of the classification with the testing set confirm that the classification is stable and that the same stocks do not have significantly different clustering results depending on the time period in which the data was collected.



The most important point is that when I researched similar stocks one by one, I found a clear chain of industries related to blockchain computing resources.

For example, NVDA (NVIDIA): a fabless semiconductor company that designs and sells graphics processors;

AMD (Advanced Micro Devices, Inc.): that develops computer processors and related technologies;

Not surprisingly, there is also TSLA.

- What did you discovered?

In this list, there are also stocks that are not so much related to Ether, such as "DLTR" (Dollar Tree), which, through research, has seen unusual fluctuations in its stock price due to aggressive investors. It is possible that the converging market performance of these stocks with Ethereum is due to correlation, or 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']

PAGE 9-10: Conclusions

Note: Discuss how the results of your analysis can be used by decision makers?

Also reflect on what you learned when working on this assignment.

Through this project, I got a list of a total of 22 companies that are more convergent with the market movements of Ethereum. 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.

For this project I followed the SEMMA methodology, collecting data, building the training set and the test set, explore data, modify variables, setup model (unsupervised clustering analysis), using testing set to evaluate my model.

I learned a lot about Python packages, including: urllib.request, ssl, pandas_datareader, pandas, numpy, datetime, sklearn.cluster, and matplotlib.pyplot.

I learned how to deal with the problem of clustering and analyzing multivariate data in time series. In addition, while referring to various details of the code, I solved many problems caused by package updates, which all improved my experience with the Python language.

References

1. Cluster analysis on stock selection
https://towardsdatascience.com/clustering-analysis-on-stock-selection-2c2fd079b295
2. 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
3. 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
4. Stock Clustering with Time Series Clustering in R
https://medium.com/@panda061325/stock-clustering-with-time-series-clustering-in-r-63fe1fabe1b6
5. Step by Step: Twitter Sentiment Analysis in Python
https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58d
6. Ethereum USD (ETH-USD) Price, Value, News & History
https://finance.yahoo.com/quote/ETH-USD/

Appendices

1. python scripts