Assignment 4

BAN 250

Jingyi Wang

Dataset:

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.

Formula for calculating new variables:

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

Topics covered:

1. K-Means Clustering Analysis

2. Hoteling T^2 Test

Topic 1: Clustering Analysis:

Q1. How do you handle time series problems? What are the problems?

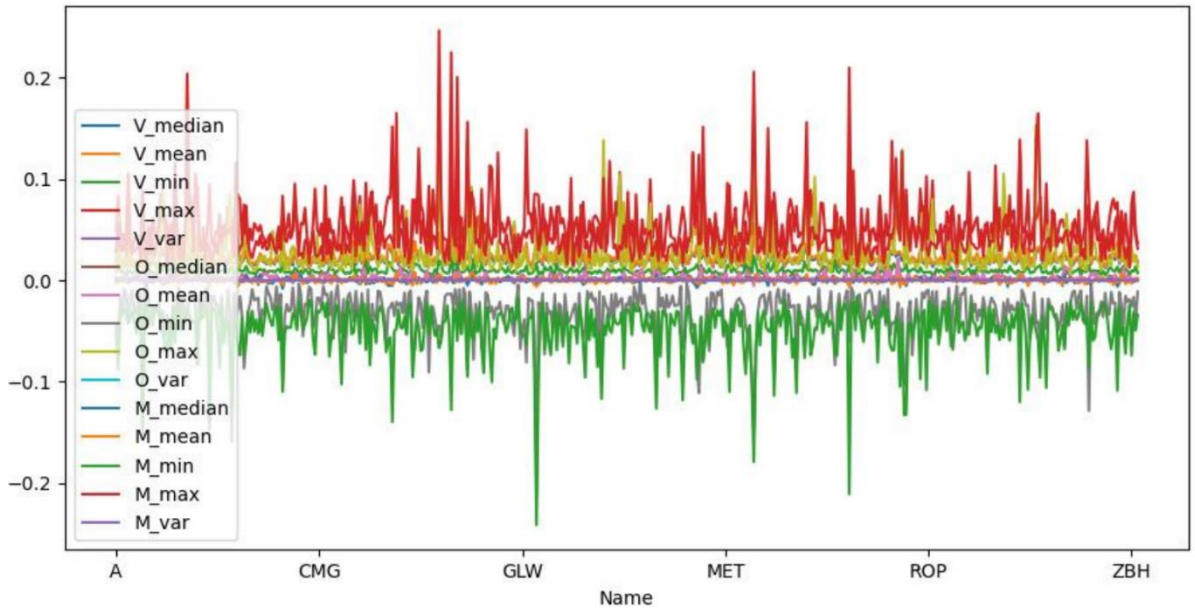
Since the dataset I used was already without missing data and had been standardized, the only thing I had to deal with in order to perform the clustering analysis was the time series problem.

As you can see from the screenshot on the top right, there are 60 sets of data (60 days) for each stock in the original dataset. I take the approach of aggregating the data set by stock name and expressing the distribution of a variable over the time period as "maximum", "minimum", "mean", "median", and "variance". For example, the "Volatility" column for each stock will become: v-median, v_mean, v_min, v_max, v_var.

In this way, the original data matrix for a particular stock will be represented by one row, which gives the structure shown in the screenshot below right.

```
> head(Stock)
  Name       Date  Volatility    Open_per     Moverate
1  MMM  9/2/2021 0.007471515 0.011421519  0.007495857
2  MMM  9/3/2021 0.010182020 0.005913867 -0.002565512
3  MMM  9/7/2021 0.046703347 0.000514454 -0.045321233
4  MMM  9/8/2021 0.018207227 0.008464030  0.012447455
5  MMM  9/9/2021 0.018396060 0.006260112 -0.011389651
6  MMM 9/10/2021 0.014609845 0.014162516 -0.006460266
```

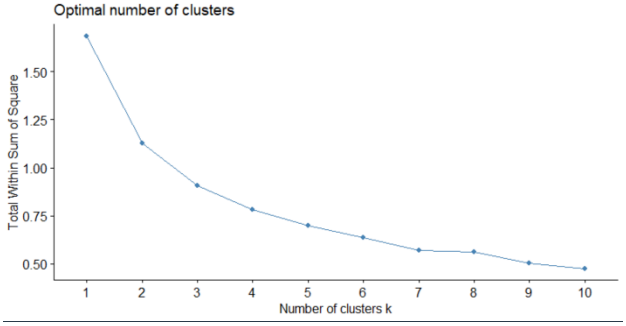
```
> head(Data)
  Name V_median   V_mean    V_min    V_max    V_var
1    A 0.017441 0.020904 0.008750 0.049057 0.000092
2  AAL 0.030922 0.033658 0.016268 0.083770 0.000191
3  AAP 0.018178 0.020576 0.010052 0.047290 0.000062
4 AAPL 0.016485 0.017979 0.008302 0.043742 0.000051
5 ABBV 0.015093 0.016274 0.008062 0.031225 0.000034
6  ABC 0.017292 0.019423 0.005200 0.045900 0.000076
  O_median     O_mean     O_min    O_max    O_var
1 0.002734  0.001141 -0.042327 0.010575 0.000063
2 0.001731  0.001567 -0.061151 0.062046 0.000212
3 0.002481  0.001728 -0.016926 0.013745 0.000032
4 0.001132 -0.000128 -0.035066 0.011143 0.000057
5 0.004659  0.006470 -0.008849 0.023663 0.000056
6 0.003794  0.003356 -0.014242 0.017403 0.000027
   M_median    M_mean     M_min    M_max    M_var
1  0.000605 -0.002258 -0.049146 0.028962 0.000205
2  0.000485 -0.001328 -0.087873 0.057683 0.000719
3  0.002083  0.002114 -0.036352 0.032911 0.000196
4  0.002441  0.000546 -0.033102 0.028536 0.000167
5 -0.000698  0.000886 -0.027999 0.045591 0.000137
6  0.000842 -0.000193 -0.031890 0.025383 0.000166
```



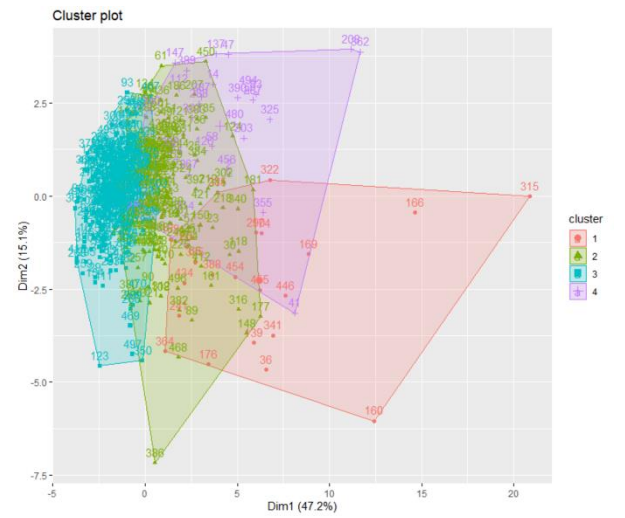
The biggest problem with this is that some relationships may not be expressed in terms of mean or variance, so this processing will result in the loss of temporal information. I wish I could find a better solution.

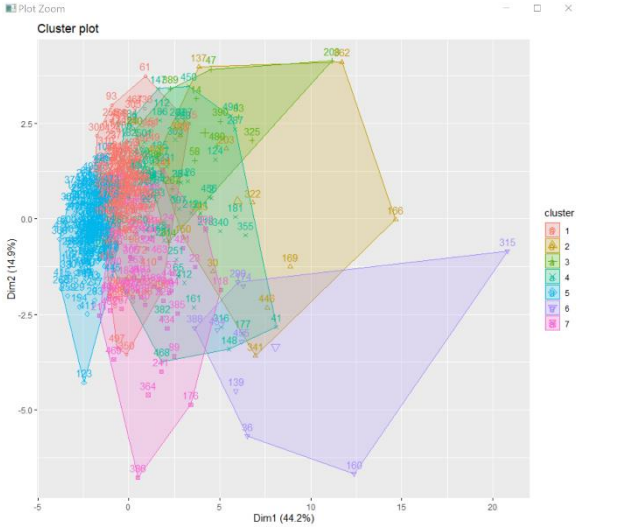
Q2. How to determine the value of K? Complete the cluster analysis.

With the Eblow Method, I decided to try both the k=4 and k=7 cases.




```r
#Elbow Method for finding the optimal number of clusters
set.seed(123)
fviz_nbclust(Data[-c(1,17,18)], kmeans, method = "wss")

#k=7
final <- kmeans(Data[-1], 7, nstart = 25)
print(final)
fviz_cluster(final, data = Data[-1])
Data$cluster_7 = final$cluster
#k=4
final_4 <- kmeans(Data[-1], 4, nstart = 25)
fviz_cluster(final_4, data = Data[-1])
print(final_4)
Data$cluster_4 = final_4$cluster
```





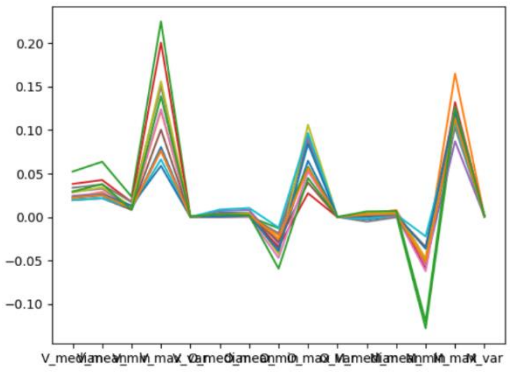
Q3. Interpret the result.

Through clustering analysis, I found 21 and 13 stocks in the same group as ETH respectively, and obtained their merges and intersections.

```
> Namelist_7
 [1] "AMD"  "DISH" "DXCM" "ETH"  "ETSY" "GNRC" "IPGP" "LUMN" "MTCH" "NVDA" "PENN" "PTC"  "TSLA" "WBA"
> Namelist_4
 [1] "ANET" "BEN"  "DLTR" "ENPH" "ETH"  "ETSY" "EXPE" "F"    "FMC"  "INTU" "LUMN" "LYV"  "MRNA" "MTCH"
[15] "NVDA" "PFE"  "PWR"  "QCOM" "TER"  "TSLA" "UA"   "UAA"
```

```
> same = newlist[duplicated(newlist)]
> same
[1] "ETH"  "ETSY" "LUMN" "MTCH" "NVDA" "TSLA"
```

```
> newlist = newlist[!duplicated(newlist)]
> newlist
 [1] "ANET" "BEN"  "DLTR" "ENPH" "ETH"  "ETSY" "EXPE" "F"    "FMC"  "INTU" "LUMN" "LYV"  "MRNA"
[14] "MTCH" "NVDA" "PFE"  "PWR"  "QCOM" "TER"  "TSLA" "UA"   "UAA"  "AMD"  "DISH" "DXCM" "GNRC"
[27] "IPGP" "PENN" "PTC"  "WBA"
```



The clustered stocks show a clear trend towards similar market performance relative to the initial haphazard charts.

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.

However, 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.

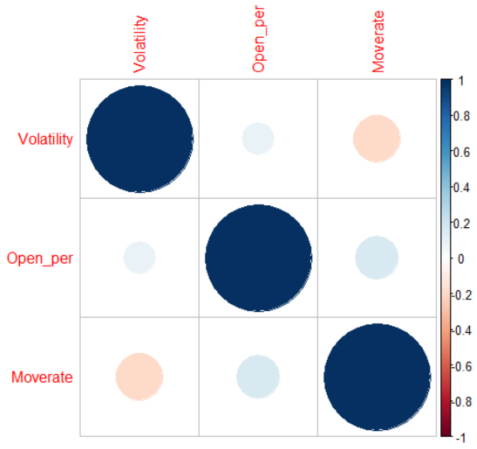
Topic 2. Hoteling T^2 Test

Q1. Why was the sample compared? Why is Hoteling used? Is the effect of time factor taken into account?

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 reason I performed the cluster analysis was to understand which public companies have close correlation with ETH.

I want to test whether it is reasonable to put blockchain currencies together with stocks for cluster analysis by comparing the sample of ETH with the sample of stocks two by two. Hoteling was used because it was a comparison between two samples, with three variables.

Since the principle of Hotelling T is to analyze the mean and variance of vectors and the distribution of variables, even if our data contains time series, the final result does not reflect the time factor, which means that we actually lose some information.



Q2. Is there a correlation between these three variables?

As you can see, Volatitity has a slight positive correlation with Moverate, but the opening performance is relatively independent.

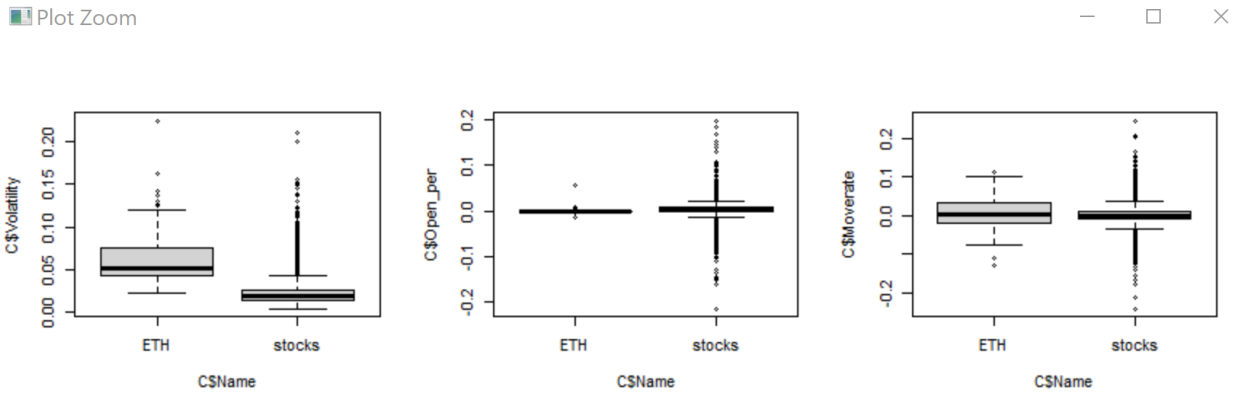
Q3. What is the result? Is this result consistent with the results of the cluster analysis, i.e. they are not different?

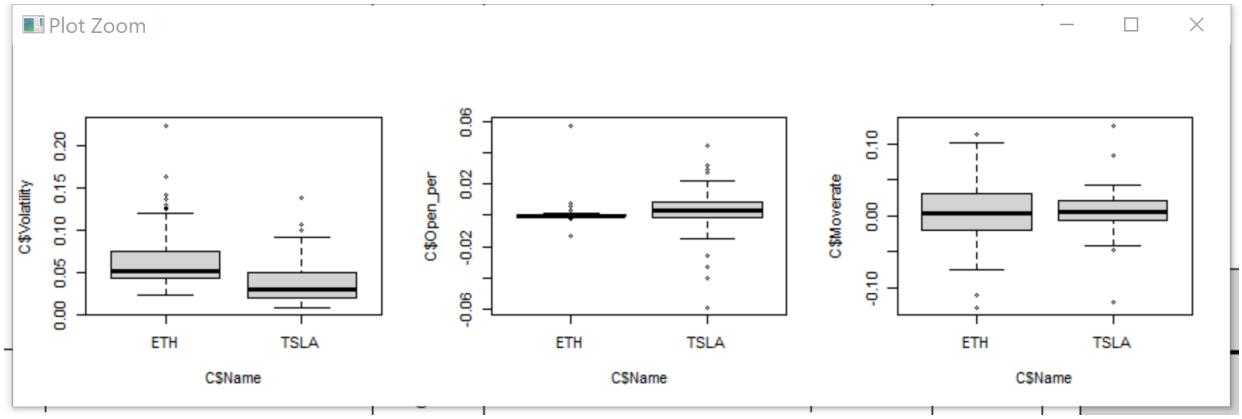
```
> m1 <- with(C, HotellingsT2(cbind(C$Volatility,C$Open_per,C$Moverate) ~ C$Name))
> m1

        Hotelling's two sample T2-test

data:  cbind(C$Volatility, C$Open_per, C$Moverate) by C$Name
T.2 = 400.14, df1 = 3, df2 = 30265, p-value < 2.2e-16
alternative hypothesis: true location difference is not equal to c(0,0,0)

> m2 <- with(D, HotellingsT2(cbind(D$Volatility,D$Open_per,D$Moverate) ~ D$Name))
> m2

        Hotelling's two sample T2-test

data:  cbind(D$Volatility, D$Open_per, D$Moverate) by D$Name
T.2 = 26.839, df1 = 3, df2 = 385, p-value = 8.882e-16
alternative hypothesis: true location difference is not equal to c(0,0,0)
```

M1 is a comparison of ETH with all stocks and we can see that p-value < 0.05, so we reject H0, in other words, we have evidence to support that the ETH is different form STOCKS.

M2 is comparing the ETH with the stocks data previously in the same cluster as it, and we get the same conclusion that they are different.

However, it is clear from the boxplot that when comparing ETH with all other stocks: the main difference is in Volatility, which is obviously much larger than the average volatility of all other stocks; and when we compare ETH with TSLA which is in the same cluster, the difference in volatility between the two is somewhat smaller.





In addition, because ETH is traded 24 hours a day, it will have a very stable opening performance, while stocks will have a significantly different opening performance relative to ETH because of the presence of over-the-counter trading.

I think this shows that the magnitude of stock price fluctuations largely determines the results of their cluster analysis.