ECOM193 - Statistical Machine Learning in Finance

Similarity of US technology stocks

Queen Mary
University of London

Hoang Long Luu

ID: 220932091

Table of Contents

Similarity of US technology stocks

I. Introduction

Clustering analysis, which is an unsupervised machine learning technique, aims to find homogeneous groups in the dataset based on some features. The ultimate goal of the method is to maximize the similarity level of observation within the group while maximizing the dissimilarity level between each group (James et al., 2023). In investment finance, the clustering method can help investors identify stock behavior, the correlation among firms, and the performance of the industry. The stocks are often analyzed by historical data such as liquidity, return, volatility, and other financial indicators. Combining the information provided by clustering analysis with other fundamental or technical methods, the investor can perform portfolio diversification and decide trading strategy more effectively (Kolanovic and T. Krishnamachari, 2017). There are two popular types of clustering analysis including hierarchical clustering and non-hierarchical clustering analysis.

II. Hierarchical and non-hierarchical clustering

Definition and type

Hierarchical clustering is the tree-like structure method that attempts to build a hierarchy of clusters. Agglomerative or bottom-up, and divisive hierarchical clustering or top-down clustering are two well-known types of this method (James et al., 2023). In agglomerative hierarchical clustering, the algorithm initially considers each data point as a single point and then they eventually merge with the closest pair into a bigger group. The process reoccurs until all data sets become a single cluster that contains all the data points. On the other hand, divisive hierarchical clustering is the method in which all data sets start with uniquely groups and then eventually break down into smaller clusters. The process stops when each observation becomes a single cluster (James et al., 2023).

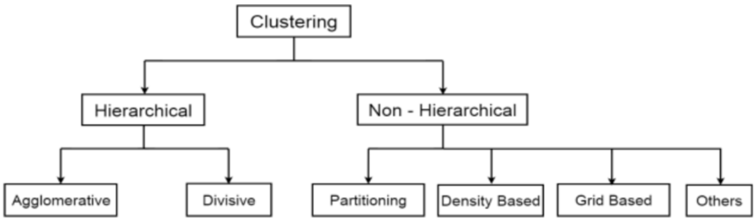


Figure 1: Categorization of clustering algorithms (Gülagiz and Sahin, 2017)

In contrast with hierarchical clustering, non-hierarchical clustering attempts to distribute the observations into distinct groups based only on similarity but without relying on the multilevel hierarchy structure. The non-hierarchical clustering divides into 4 different classes including Partitioning, Density Based, Grid Based, and other types (Kolanovic and T. Krishnamachari, 2017). The Density-based approach identifies the clusters based on the density of observations in an area. The method separates high-density regions from low-density regions. Some of the most popular algorithms of this method are DBSCAN and OPTICS. The Grill approach divides the data set into cells and then performs a clustering task on the grid structure. STING, DENCLUE, MAFIA, and CLIQUE are several notable examples of this method. Partitioning clustering methods classify the observations by a predefined number of centroids. In particular, the approach attempts to relocate the centroids in a way that minimizes the sum of squared distances from each point to the centroid (Gülagiz and Sahin, 2017). Several famous algorithms of Partitioning clustering are K-Means, K-Medoids, and Farthest First.

Distance measurement and similarity calculation

The similarity between clusters is normally calculated by several approaches such as Euclidean Distance, Manhattan Distance, Jaccard Distance, and Cosine Similarity in both types of clustering methods (James et al., 2023). In the context of stock analysis, Euclidean Distance is the most used. The Euclidean distance is the ordinal straight-line distance between two points in the space. The smaller the distance, the higher the similarity between the two groups. For instance, the formula Euclidean distance of two points $P_2(x_{1,}x_{2,}x_{3,}, \dots x_n)$ and $P_1(y_{1,}y_{2,}y_{3,}, \dots y_n)$ is displayed as:

$$d(P_1, P_2) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_n - x_n)^2}$$

Basically, minimize the distance of observation within a cluster while wider the distance between clusters.

Nonetheless, in hierarchical clustering methods, the groups with more than observations need the rule to calculate which is so-called linkages. In other words, linkage refers to the rule used to determine the similarity between two clusters of data points. There are four most common methods of linkage including complete, single, average, and centroid (James et al., 2023). The complete linkage is the method that chooses the pairwise of observations in a different group that has the longest distance, while the single linkage considers the shortest distance as the distance between groups. The average linkage method calculates all the dissimilarity between two groups and then takes the average of their distance. The centroid linkage estimates the dissimilarity by the distance between the centroid of a group to its counterpart in another group. The different methods led to differences in the dissimilarity between groups, thereby the choice of linkage rule influences how the hierarchical structure is formed and how clusters are merged at each step (James et al., 2023). Nonetheless, the ward's method is one of the most recommended and stable. The method attempts to minimize the sum of squared deviation within clusters (Vijaya, Sharma and Batra, 2019). In other words, the level increase of the sum of squared deviation is considered by the algorithm to decide merge or not. The formula of the ward is:

$$d(A,B) = \frac{|A||B|}{|A| + |B|} \times \ \|C(A) - C(B)\|^2$$

The $|A|\ and|B|$ is the size of the groups A and B while $\|C(A) - \ C(B)\|^2$ is the quare of the Euclidean distance between the centroids of the two groups.

Number of clusters

The number of clusters is a significant index that heavily impacts analysis effectiveness. The result is not so informative if the number of clusters is not suitable (James et al., 2023). Too few clusters may not capture the underlying structure of the data adequately while too many clusters can model noise in the data. In addition, the wrong-chosen number of clusters can lead to the reduction of model robustness and reproducibility.

In hierarchical clustering, the number of clusters is decided by reviewing the tree-based dendrogram (James et al., 2023). The high fusion or joining of two clusters, which estimates the vertical axis, decides the similarity of observations. In detail, the lower the vertical distance between fusion and dendrogram bottom is, the higher the level of similarity between observations is. The chosen threshold divides the number of clusters. The decision on where to cut the dendrogram depends on the problem domain and the criteria for specific circumstances but it still needs to obey the main clustering analysis (James et al., 2023).

In terms of non-hierarchical clustering, the Elbow method is commonly utilized to consider the number of groups (James et al., 2023). The method interprets the number of clusters again as the within-clusters sum of squares (WCSS). Normally, the optimal number of clusters is the one where the drop of WCSS is slowed down dramatically.

III. The case of 21 technology firms

Data processing

The data set contains the daily closing price of 21 technology companies from January 2019 to June 2021. In general, the data is sufficient and clean. The only problem is that the data includes the closing price on Saturdays and Sundays as well as holidays, which are the days that the Nasdaq 100 index does not operate. Therefore, the data on that day needs to be removed before calculating several indices like log return and volatility. After removing all weekends, the number of days moves from 900 days to 620 days.

Exploratory Analysis

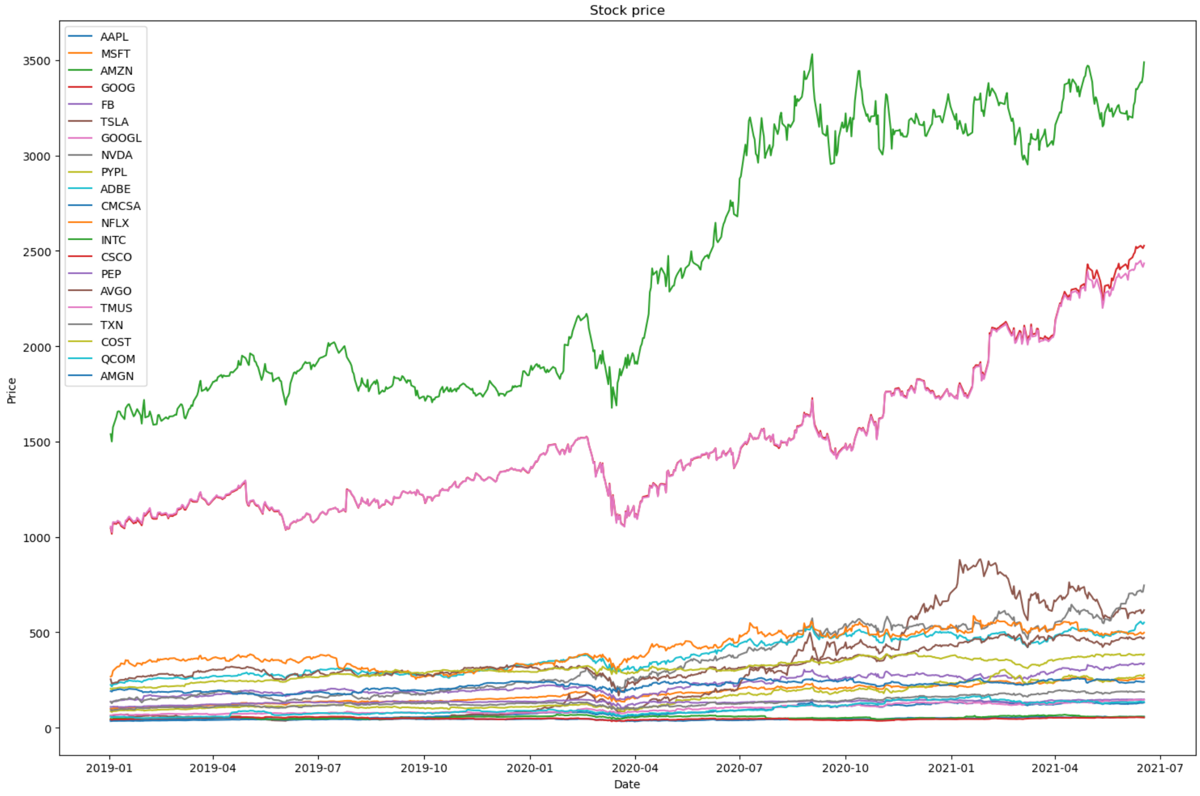


Figure 2

According to the chart, several general conclusions can be made. The prices of Amazon and Alphabet (the company formerly known as Google) were significantly higher than the rest. In addition, almost most companies endured a dramatic decrease in price from March 2020 to April 2020. The propel reason for this phenomenon is economic shock influenced by the Covid 19. The only outlier was Costco Wholesale Corporation (COST), in which the company price remains stable.

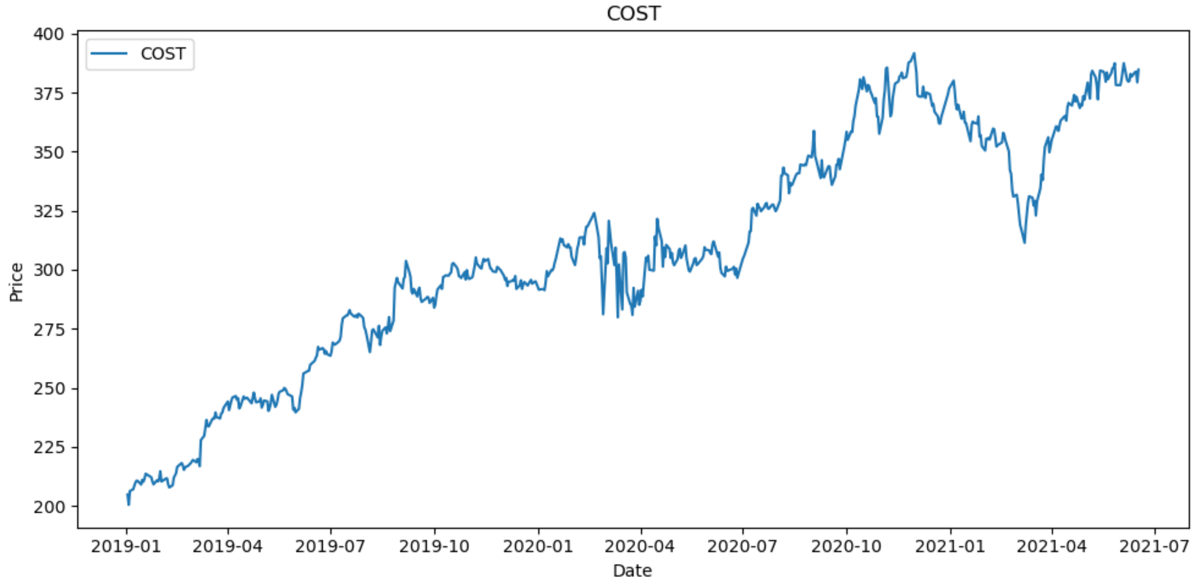


Figure 3

After the crisis, the general trend of stocks is to increase compared to before the crisis but with strong price fluctuations.

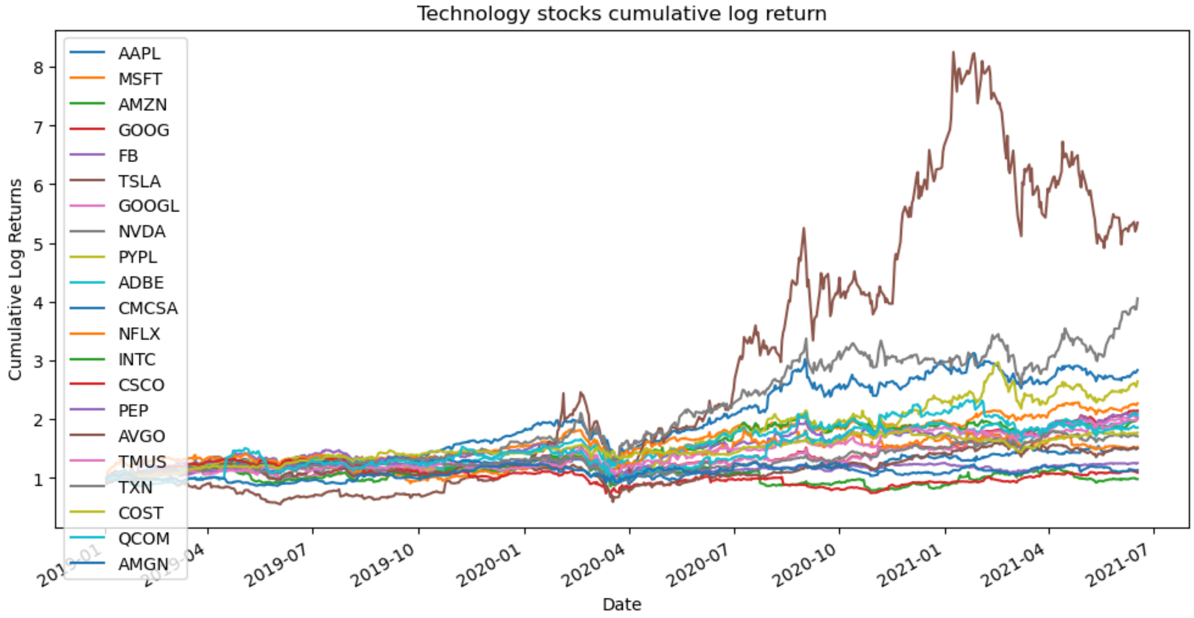
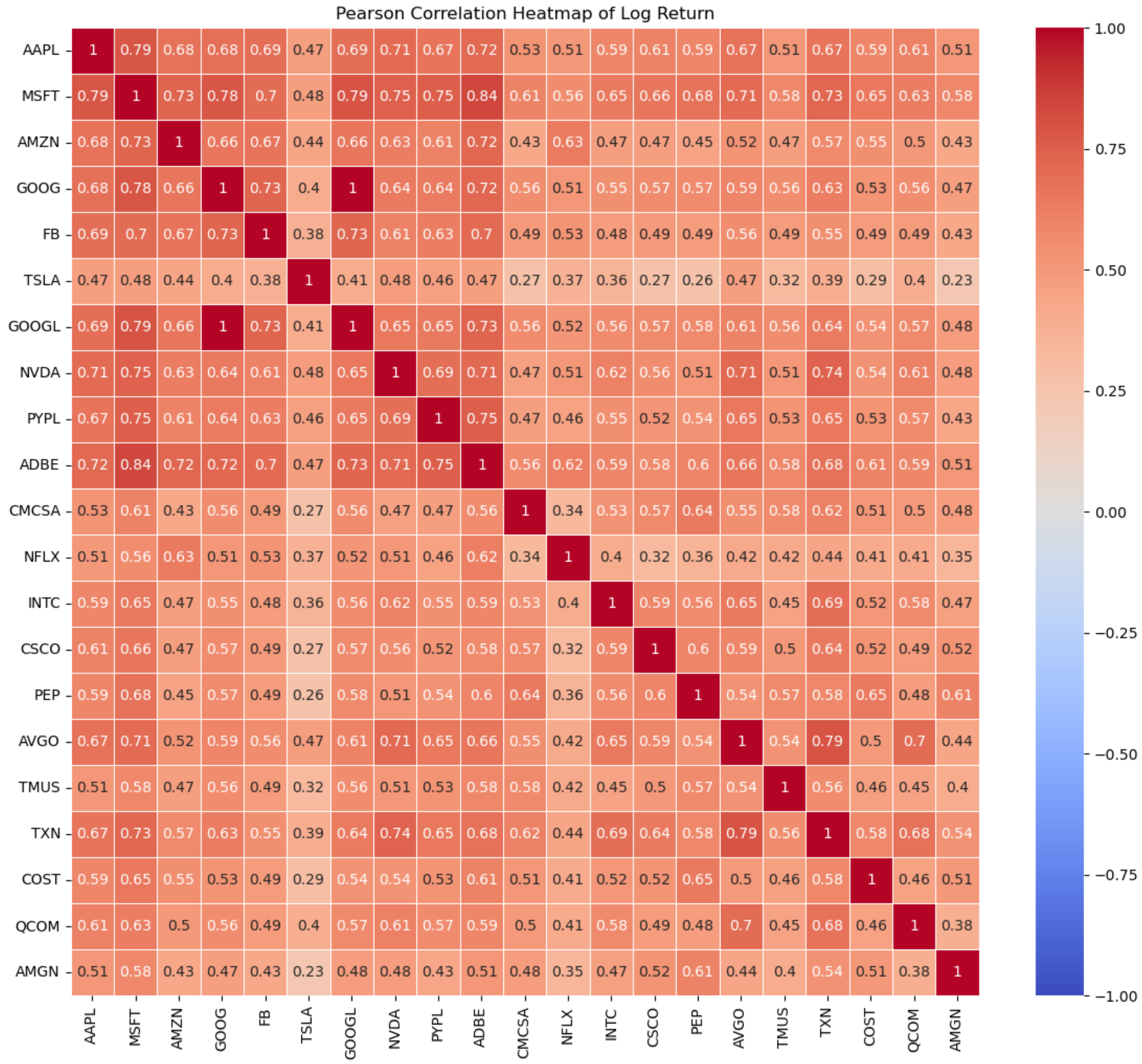


Figure 4

Based on the cumulative log return, TSLA and NVDA were two stocks that have good performance in the given period.



	AAPL	MSFT	AMZN	GOOG	FB	TSLA	GOOGL	NVDA	PYPL	ADBE	CMCSA	NFLX	INTC	CSCO	PEP	AVGO	TMUS	TXN	COST	QCOM	AMGN
AAPL	1	0.79	0.68	0.68	0.69	0.47	0.69	0.71	0.67	0.72	0.53	0.51	0.59	0.61	0.59	0.67	0.51	0.67	0.59	0.61	0.51
MSFT	0.79	1	0.73	0.78	0.7	0.48	0.79	0.75	0.75	0.84	0.61	0.56	0.65	0.66	0.68	0.71	0.58	0.73	0.65	0.63	0.58
AMZN	0.68	0.73	1	0.66	0.67	0.44	0.66	0.63	0.61	0.72	0.43	0.63	0.47	0.47	0.45	0.52	0.47	0.57	0.55	0.5	0.43
GOOG	0.68	0.78	0.66	1	0.73	0.4	1	0.64	0.64	0.72	0.56	0.51	0.55	0.57	0.57	0.59	0.56	0.63	0.53	0.56	0.47
FB	0.69	0.7	0.67	0.73	1	0.38	0.73	0.61	0.63	0.7	0.49	0.53	0.48	0.49	0.49	0.56	0.49	0.55	0.49	0.49	0.43
TSLA	0.47	0.48	0.44	0.4	0.38	1	0.41	0.48	0.46	0.47	0.27	0.37	0.36	0.27	0.26	0.47	0.32	0.39	0.29	0.4	0.23
GOOGL	0.69	0.79	0.66	1	0.73	0.41	1	0.65	0.65	0.73	0.56	0.52	0.56	0.57	0.58	0.61	0.56	0.64	0.54	0.57	0.48
NVDA	0.71	0.75	0.63	0.64	0.61	0.48	0.65	1	0.69	0.71	0.47	0.51	0.62	0.56	0.51	0.71	0.51	0.74	0.54	0.61	0.48
PYPL	0.67	0.75	0.61	0.64	0.63	0.46	0.65	0.69	1	0.75	0.47	0.46	0.55	0.52	0.54	0.65	0.53	0.65	0.53	0.57	0.43
ADBE	0.72	0.84	0.72	0.72	0.7	0.47	0.73	0.71	0.75	1	0.56	0.62	0.59	0.58	0.6	0.66	0.58	0.68	0.61	0.59	0.51
CMCSA	0.53	0.61	0.43	0.56	0.49	0.27	0.56	0.47	0.47	0.56	1	0.34	0.53	0.57	0.64	0.55	0.58	0.62	0.51	0.5	0.48
NFLX	0.51	0.56	0.63	0.51	0.53	0.37	0.52	0.51	0.46	0.62	0.34	1	0.4	0.32	0.36	0.42	0.42	0.44	0.41	0.41	0.35
INTC	0.59	0.65	0.47	0.55	0.48	0.36	0.56	0.62	0.55	0.59	0.53	0.4	1	0.59	0.56	0.65	0.45	0.69	0.52	0.58	0.47
CSCO	0.61	0.66	0.47	0.57	0.49	0.27	0.57	0.56	0.52	0.58	0.57	0.32	0.59	1	0.6	0.59	0.5	0.64	0.52	0.49	0.52
PEP	0.59	0.68	0.45	0.57	0.49	0.26	0.58	0.51	0.54	0.6	0.64	0.36	0.56	0.6	1	0.54	0.57	0.58	0.65	0.48	0.61
AVGO	0.67	0.71	0.52	0.59	0.56	0.47	0.61	0.71	0.65	0.66	0.55	0.42	0.65	0.59	0.54	1	0.54	0.79	0.5	0.7	0.44
TMUS	0.51	0.58	0.47	0.56	0.49	0.32	0.56	0.51	0.53	0.58	0.58	0.42	0.45	0.5	0.57	0.54	1	0.56	0.46	0.45	0.4
TXN	0.67	0.73	0.57	0.63	0.55	0.39	0.64	0.74	0.65	0.68	0.62	0.44	0.69	0.64	0.58	0.79	0.56	1	0.58	0.68	0.54
COST	0.59	0.65	0.55	0.53	0.49	0.29	0.54	0.54	0.53	0.61	0.51	0.41	0.52	0.52	0.65	0.5	0.46	0.58	1	0.46	0.51
QCOM	0.61	0.63	0.5	0.56	0.49	0.4	0.57	0.61	0.57	0.59	0.5	0.41	0.58	0.49	0.48	0.7	0.45	0.68	0.46	1	0.38
AMGN	0.51	0.58	0.43	0.47	0.43	0.23	0.48	0.48	0.43	0.51	0.48	0.35	0.47	0.52	0.61	0.44	0.4	0.54	0.51	0.38	1

Figure 5

The table shows the correlation of log return. In general, all the technology firms are positively correlated. The unprecedented phenomenon is seen in the case of GOOGL and GOOG. In detail, the two stock tickets are absolutely correlated. This result was normal since both GOOGLE and GOOG are the stocks of Alphabet. The main difference is that GOOGLE has the authority to vote while GOOG is not. Other remarkable positive relationships are the pair of ADBE and MSFT, TNX and AVGO, GOOGL and MSFT.

	ADBE and MSFT	TXN and AVGO	GOOGL and MSFT
p-value	6.355332e-168	4.966377e-136	2.800644e-134

Figure 6

The p-value show that all relationships are significantly positive.

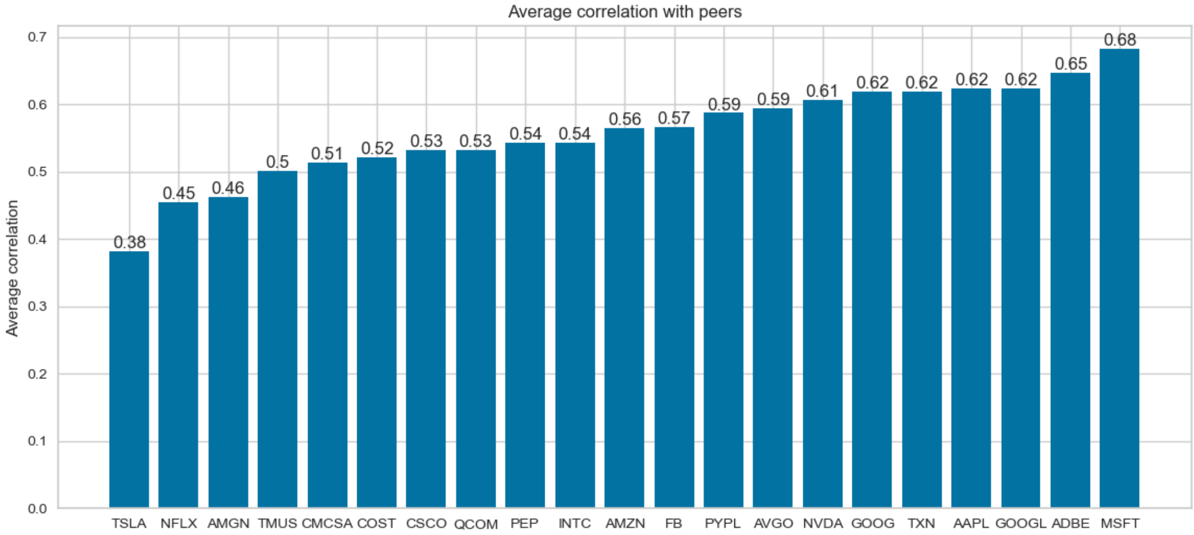


Figure 7

MSFT was the stock ticket that had the highest average correlation (0.68) with other firms in the data set. Followed MSFT are ADBE, GOOGL, AAPL, TXN, and NVDA. They had high correlation with their peer in the technology sections.

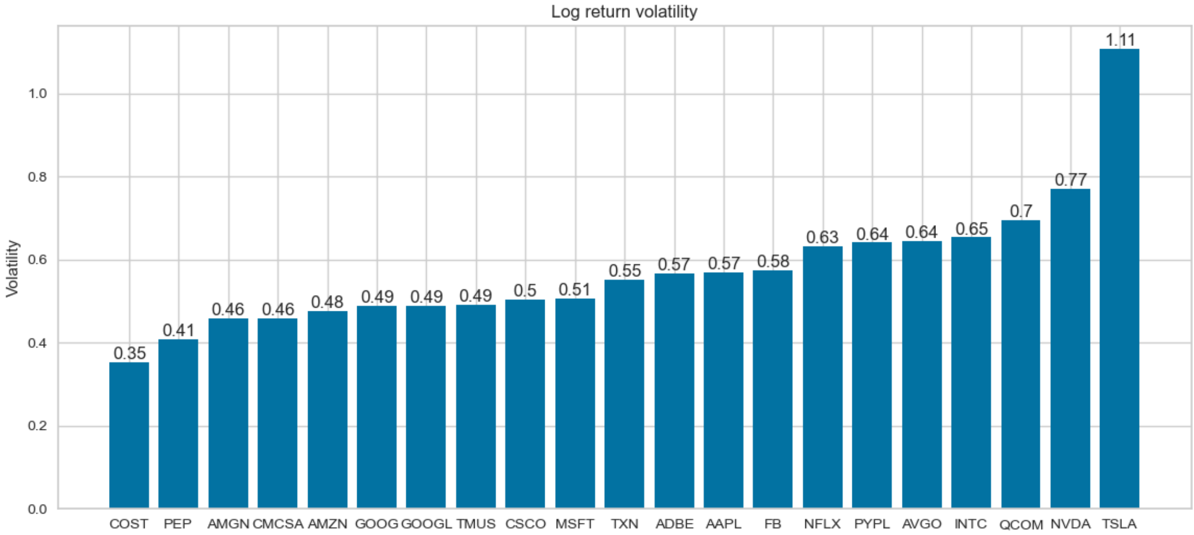


Figure 8

The bar chart shows the volatility of each stock. Tesla Corporation had an outstanding volatility level (1.11) that was nearly double the second one. The high volatility is often synonymous with higher risk and unpredictability, especially in short period. In contrast, COST had the lowest volatility with around 0.35. Therefore, the COST return is very stable at this stage.

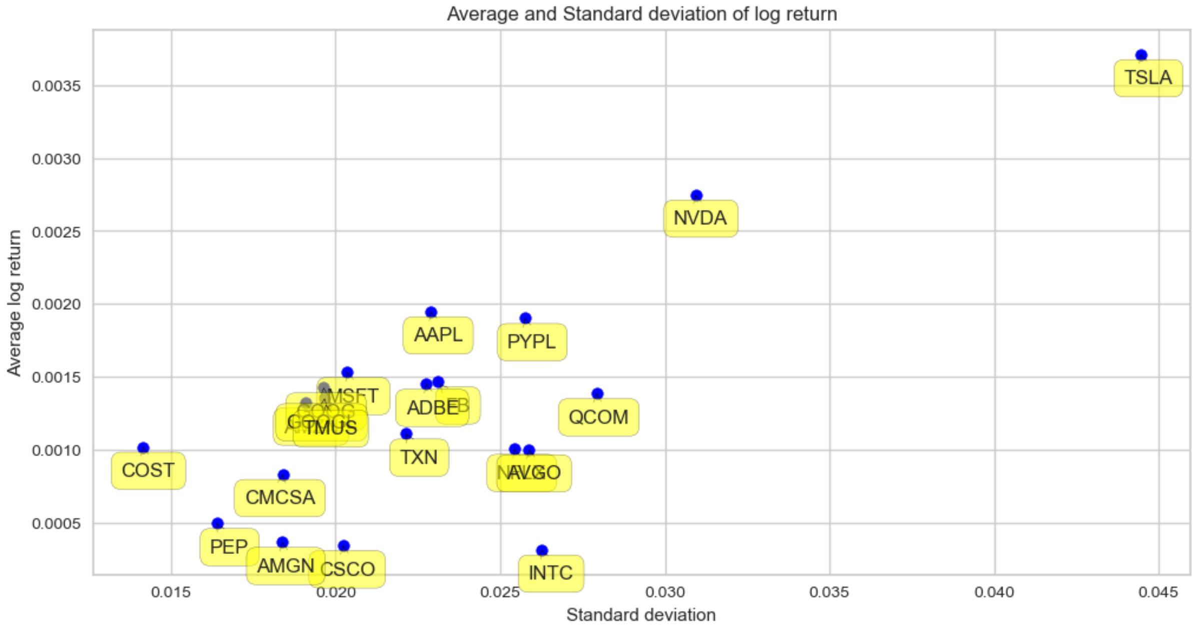


Figure 9

The scatter plot shows the relationship between average log return and standard deviation in given period. Based on the two metrics several estimations about the risk level of stocks can be pointed out. The clustering is a suitable approach to investigate the stocks and draw meaningful conclusions. Based on visual inspection, NVDA and especially TSLA are outliers in this data set. Nevertheless, the outsiders are not removed due to the small number of observations.

Clustering analysis

In this section, based on the closing price of stocks, average log return and stand deviation of log return are calculated and used for classification purpose. Several methods which is suitable for small number of stocks including Hierarchical clustering, and K-mean are used.

Before clustering, the data is standardized since stocks can have returns and risks on different scales. This ensures that both risk and return are given equal weight in the analyse process.

To comparing the performance of clustering algorithms, three metric is utilized. Silhouette coefficient is estimated by mean nearest-cluster distance and mean intra-cluster distance to determines how close each point in one cluster is to the points in the neighbouring clusters. The value range is between -1 and 1, where a higher value is good (Rousseeuw, 1987). Davies-Bouldin index indicate the average similarity measure of each cluster with its most similar cluster. The closer value to 0 is, the better the algorithm is (Davies and Bouldin, 1979). Finally, Calinski- Harabasz metric shows the dense and well-separated clusters by the ration of between group dispersion and within group dispersion. The higher value is, the better the algorithm is (Calinski and Harabasz, 1974).

Hierarchical clustering method (Agglomerative)

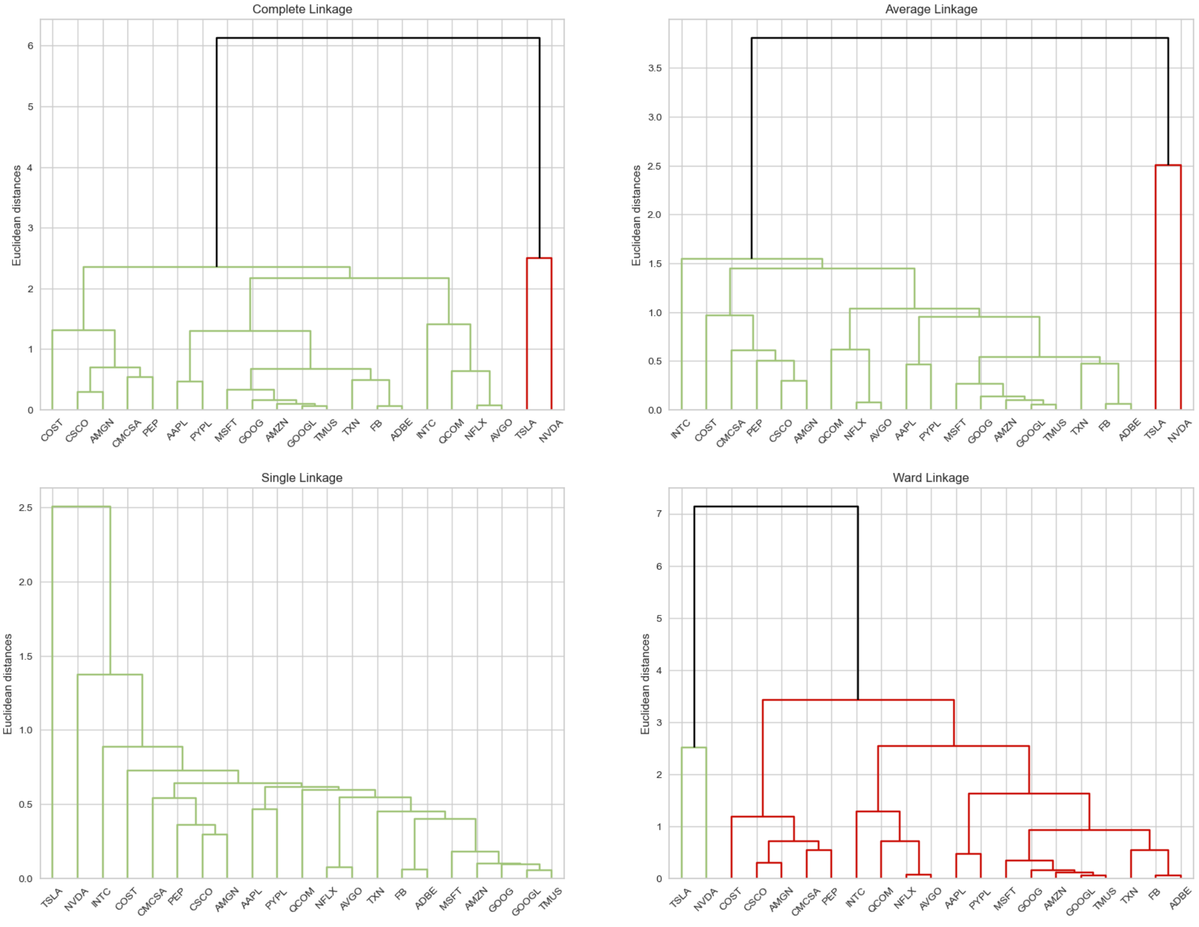


Figure 10

In the analysis, four types of linkage including complete, average, single, and ward are used simultaneously. In addition, Euclidean is chosen to calculate the distances or similarity. In this analysis, volatilities were the chosen index for classifying stocks. The chart dendrogram shows that the method should have 2 or 5 clusters.

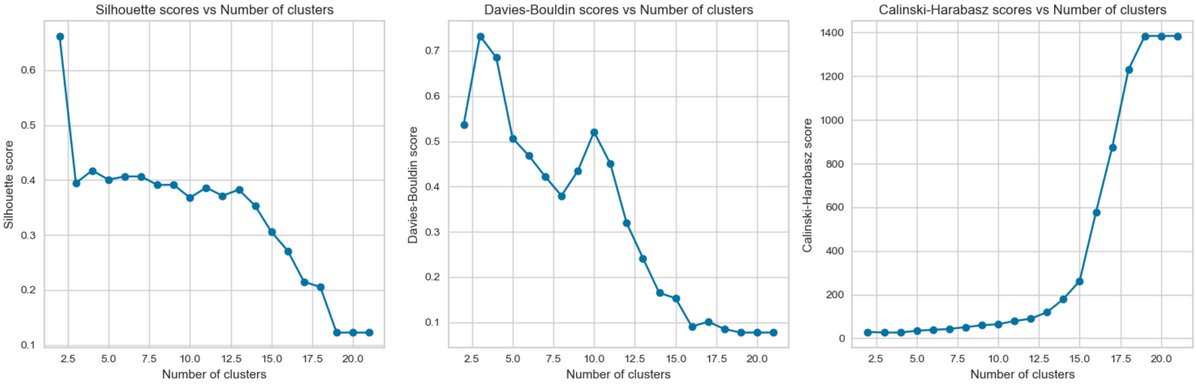


Figure 11

After double check by Silhouette coefficient and Davies-Bouldin index, each index suggests a different number of clusters. Davies-Bouldin and Calinski-Harabasz metrics show that 19 is the optimal number of clusters, while Silhouette scores proves that 2 is the most suitable. Nonetheless, 19 clusters are extremely high for a small number of observations in this research.

	Complete	Average	Single	Ward	Standard deviation	Average return
COST	1	1	0	0	0.014152	0.001019
PEP	1	1	0	0	0.016389	0.000496
AMGN	1	1	0	0	0.018387	0.000366
CMCSA	1	1	0	0	0.018398	0.000832
AMZN	1	1	0	0	0.019104	0.001322
GOOG	1	1	0	0	0.019637	0.001425
GOOGL	1	1	0	0	0.019668	0.001352
TMUS	1	1	0	0	0.019706	0.001309
CSCO	1	1	0	0	0.020222	0.000344
MSFT	1	1	0	0	0.020362	0.001531
TXN	1	1	0	0	0.022140	0.001109
ADBE	1	1	0	0	0.022750	0.001451
AAPL	1	1	0	0	0.022886	0.001947
FB	1	1	0	0	0.023103	0.001467
NFLX	1	1	0	0	0.025422	0.001004
PYPL	1	1	0	0	0.025740	0.001901
AVGO	1	1	0	0	0.025873	0.001001
INTC	1	1	0	0	0.026251	0.000314

QCOM	1	1	0	0	0.027928	0.001386
NVDA	0	0	0	1	0.030931	0.002748
TSLA	0	0	1	1	0.044440	0.003710

Figure 12

The results of all linkage but single are simultaneous. In detail, TSLA and NVDA belong to one group, and other stocks is a group, which is not surprised. It is worth noting that group numbers like 0 and 1 are just for identity purposes, not for ranking.

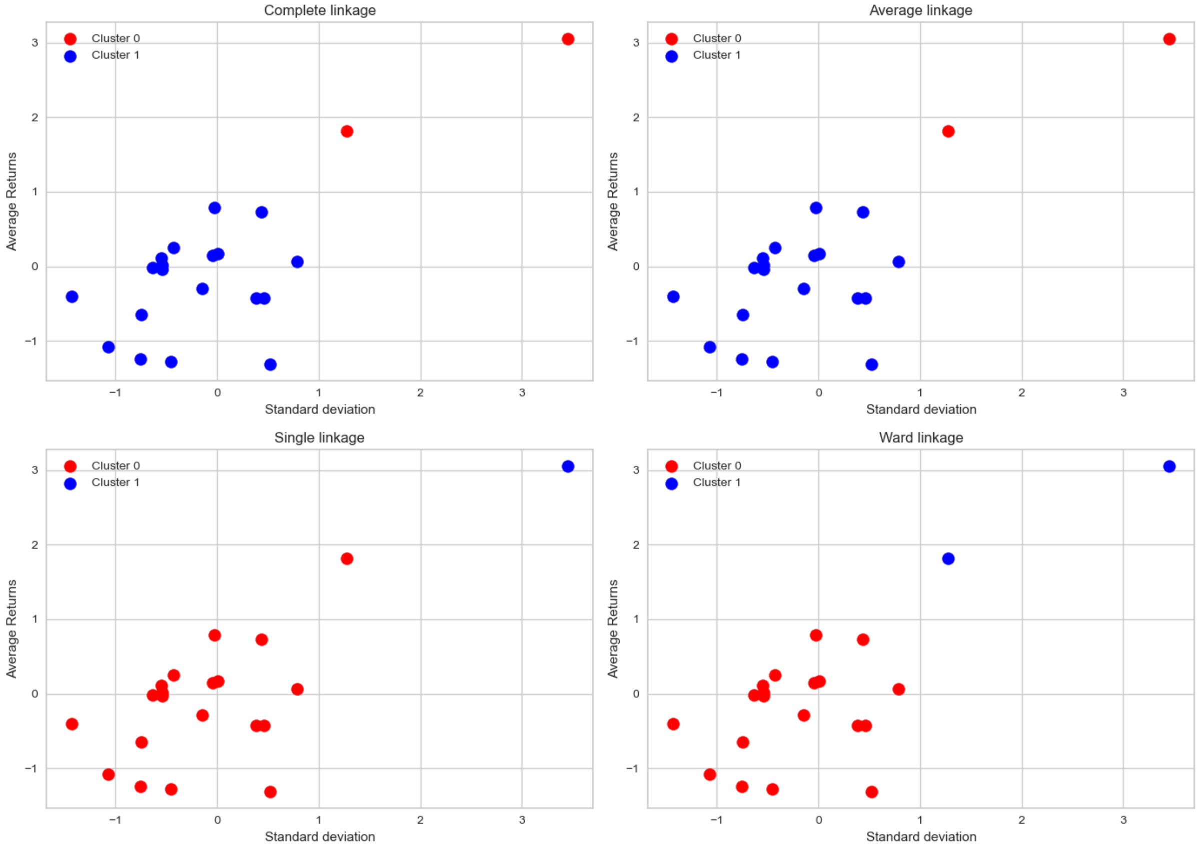


Figure 13

It is clearly that the result is impacted by the outliers, and the clarity of this result is not so good. Hence, it is hard to draw meaningful conclusion with this clustering decision.

Based on the standard deviation and average return, several characteristics of each group can be pointed out. The one has high return and risk while other has lower risk and return.

K-mean clustering

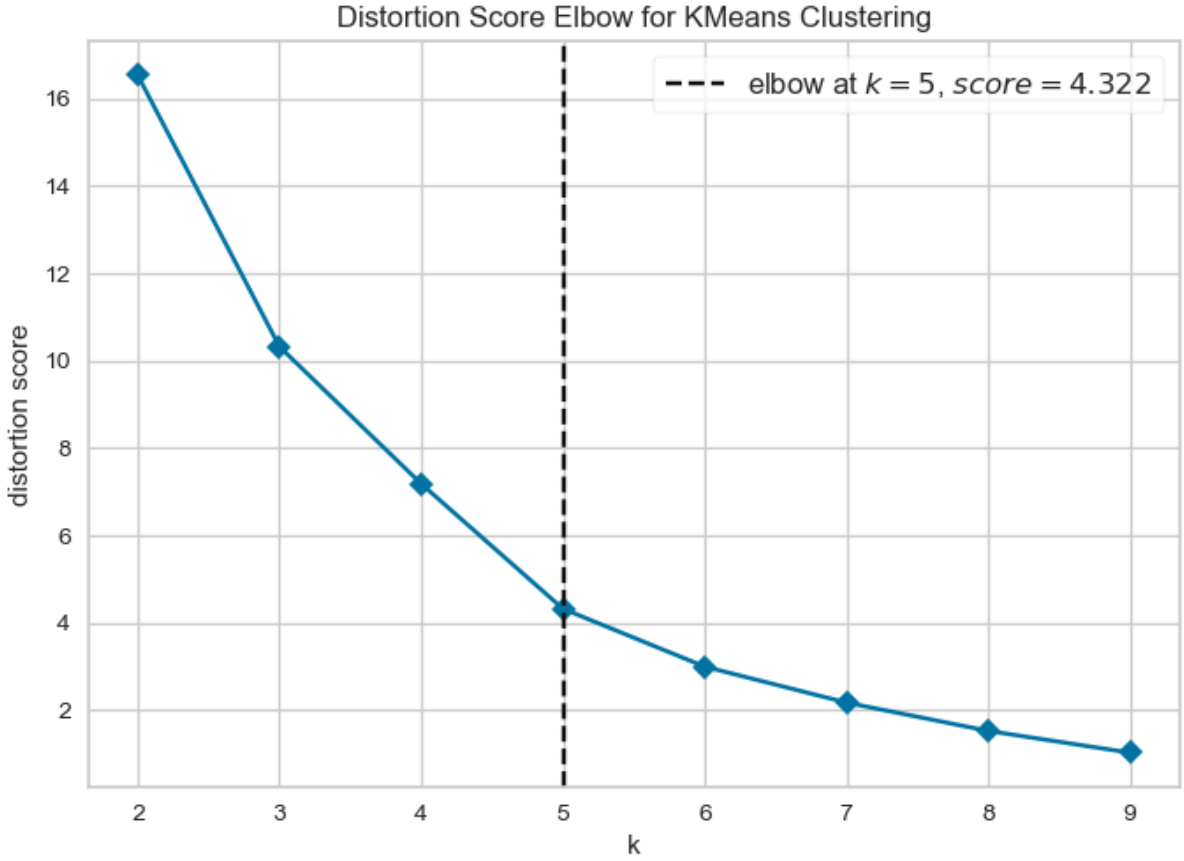
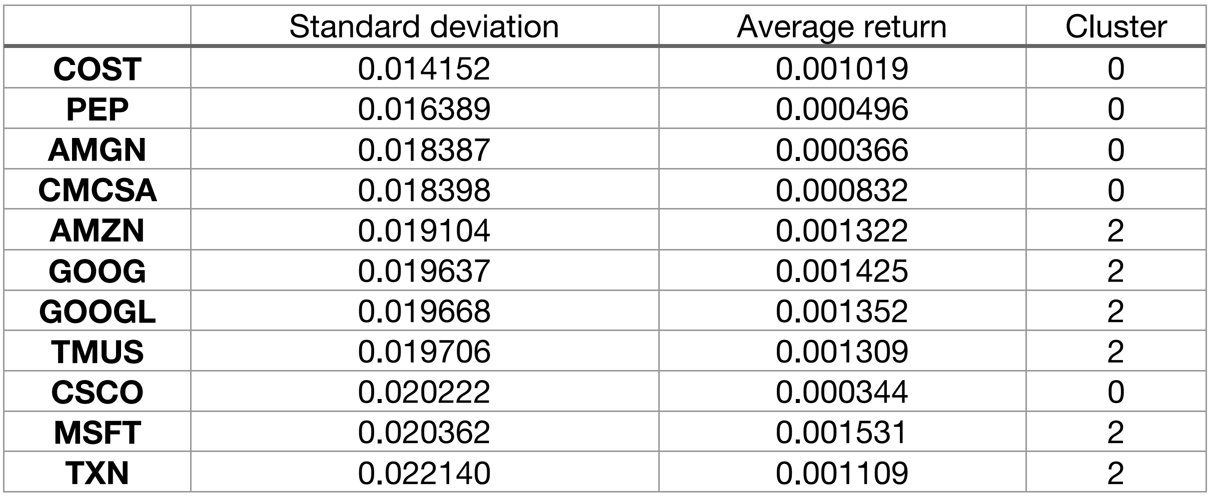


Figure 14

The Elbow method is used to identify the optimal number of clusters. Based on the figure, the optimal number of clusters in mean clustering is 5. In particular, the stock should be divided into five group based on standard deviation and average log return. The initial centroid is choised by K-mean++ algorithm to reduced Sensitivity to Initialization.



	Standard deviation	Average return	Cluster
COST	0.014152	0.001019	0
PEP	0.016389	0.000496	0
AMGN	0.018387	0.000366	0
CMCSA	0.018398	0.000832	0
AMZN	0.019104	0.001322	2
GOOG	0.019637	0.001425	2
GOOGL	0.019668	0.001352	2
TMUS	0.019706	0.001309	2
CSCO	0.020222	0.000344	0
MSFT	0.020362	0.001531	2
TXN	0.022140	0.001109	2

ADBE	0.022750	0.001451	2
AAPL	0.022886	0.001947	2
FB	0.023103	0.001467	2
NFLX	0.025422	0.001004	4
PYPL	0.025740	0.001901	2
AVGO	0.025873	0.001001	4
INTC	0.026251	0.000314	4
QCOM	0.027928	0.001386	4
NVDA	0.030931	0.002748	1
TSLA	0.044440	0.003710	3

Figure 15

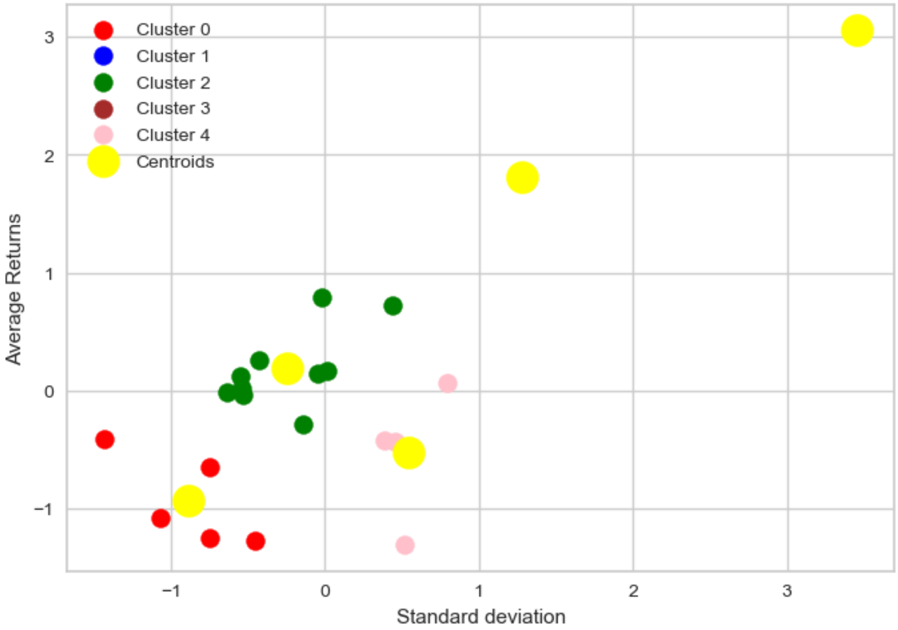
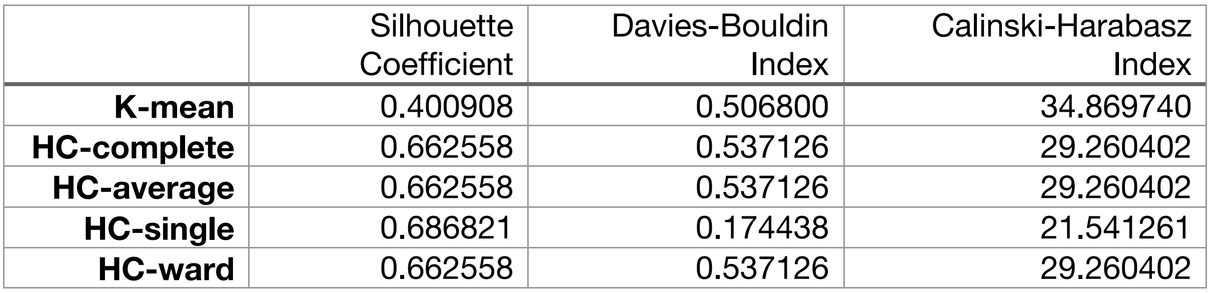


Figure 16

The figures show that there are five groups of stocks. Each outlier represents for one group while the rest is divided into 5 clear groups. Based on visual inspection, this result is more interpretative than the one in hierarchical clustering.

Evaluation

Silhouette coefficient, Davies-bouldin, and Calinski-Harabasz are used to evaluate the performance of algorithm. It is worth highlighting that the algorithm with better index is normally but not always the most appropriate with data set (Vendramin, Campello and Hruschka, 2010).



	Silhouette Coefficient	Davies-Bouldin Index	Calinski-Harabasz Index
K-mean	0.400908	0.506800	34.869740
HC-complete	0.662558	0.537126	29.260402
HC-average	0.662558	0.537126	29.260402
HC-single	0.686821	0.174438	21.541261
HC-ward	0.662558	0.537126	29.260402

Figure 17

The SC and DBI suppose that hierarchical clustering, especially single linkage, outperform K-mean, achieves better cluster cohesion and separation than K-means. Nevertheless, K-means outperforms in the Calinski-Harabasz Index (CHI), indicating better-defined groups. Considering the overall cluster representation and interpretability, K-means is chosen for its more generalized and holistic grouping of the data (Figure 16).

Conclusion

In general, the technology stack is divided into five groups (Figure 15, 16). Group 3, represented by only TSLA ticket, is the high-risk, high-return stocks. The same conclusion is made with the case of group 1, represented by NVDA, but in a lower level of risk and return. Next, group 2 is the group that has medium risk and medium return. The stock belonging to group 0 has low risk and low return while the members of group 4 have medium risk and low return.

From the given data, clusters 3 and 1 are suitable for risk-seeking investors due to their high returns and risk. Cluster 0 is more suited for risk-averse individuals because of its stability and lower returns. On the other hand, Cluster 2 strikes a balance between risk and return, making it ideal for risk-neutral people. Interestingly, stocks in Cluster 4 underperform within the tech sector, exhibiting moderate risk yet delivering low returns.

It is worth noting that the conclusion is created just based on the data itself without consulting any external data.

IV. Clustering analysis limitations and some solutions

Even though the clustering algorithm is an impressive method, it still contains some drawbacks.

The first disadvantage usually is no benchmark or ground truth to validate the correctness of the clustering results, which is the characteristic of unsupervised learning in general (James et al., 2023). Without this benchmark, analysts are often in the dark about the optimal configuration. The problem also occurs in this research, there are no true labels for each cluster if the researchers do not consult the benchmark on the external sources. To address this, several solutions can be pointed out. The specialist can utilize internal evaluation metrics such as the Silhouette coefficient, Davies-Bouldin index, and Calinski-Harabasz index to evaluate the performance of the algorithm (Han, Kamber, and Pei, 2012). For small databases like 21 stock ticket, visual inspection can be an effective technique. Scatter plots, pair plots, dendrograms, and heatmaps can help experiment analysts identify trends, outliers, or patterns. Nonetheless, this method contains vast disadvantages due to the limitations of human perspective when dealing with large and complex data.

Second, the algorithm is sensitive to the parameter's choice (Chiang and Mirkin, 2010). Methods such as hierarchical clustering, and K-mean required parameter tuning like the number of clusters, thresholds, type of linkage, and distance type, which play a pivotal role in the output. For instance, the result in hierarchical clustering is different based on the linkage. Several tools such as Score Elbow and Internal metrics can break down which parameters are the most possible choices for the algorithm.

Third, some clustering method like K-mean has susceptibility to local optimal rather than global optimal (James et al., 2023). If the solution of the algorithm is not a global optimum, it means that there are other ways where the sum of squared distances could be lower. Hence, the clusters might not be as tight and well-separated as possible. One of the reasonable explanations for why the K-mean is stuck with local optimal is the initial choice of centroid. Depending on the first position of centroids, the algorithm may find it difficult to converge to global optimal and highly trap in local optimal. In addition, the nature of the WCSS function, using in K-mean, is not convex when the number of clusters is bigger than one, therefore, further exacerbating the problem of local optima. To solve this problem, using a method that can run K-mean multiple times with different centroids and then choose the one that has the lowest WCSS is an effective strategy. A function like K-mean++ can ensure that process, but not guarantee the final decision is the most optimal (James et al., 2023). If the dataset is small and the number of dimensions is limited, visual inspection can indeed be a helpful tool to get a quick sense of the data.

Finally, clustering methods can be highly sensitive to outliers (Gagolewski, Bartoszuk and Cena, 2016). Algorithms such as K-mean are dramatically influenced by extreme data due to their underlying computations. For instance, an outlier can significantly move the centroid. Several algorithms are more robust to noise and outliers like k-medoids because they use actual data points to represent the cluster center, rather than computing a mean. In exchange, the method requires higher computational complexity and reduces scalability (Gülagiz and Sahin, 2017).

V. References

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1), pp.1–27. doi:https://doi.org/10.1080/03610927408827101.

Chiang, M.M.-T. and Mirkin, B. (2010). Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. *Journal of Classification*, 27(1), pp.3–40. doi:https://doi.org/10.1007/s00357-010-9049-5.

Davies, D.L. and Bouldin, D.W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, [online] PAMI-1(2), pp.224–227. doi:https://doi.org/10.1109/TPAMI.1979.4766909.

Gagolewski, M., Bartoszuk, M. and Cena, A. (2016). Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm. *Information Sciences*, 363, pp.8–23. doi:https://doi.org/10.1016/j.ins.2016.05.003.

Gülagiz, F.K. and Sahin, S. (2017). Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms. *ProQuest*, [online] 9(1), pp.6–14. Available at: https://www.proquest.com/docview/1873974618?pq-origsite=gscholar&fromopenview=true [Accessed 5 Apr. 2022].

Han, J., Kamber, M. and Pei, J. (2012). Cluster Analysis. *Data Mining*, pp.443–495. doi:https://doi.org/10.1016/b978-0-12-381479-1.00010-1.

James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. (2023). *An Introduction to Statistical Learning*. Springer Nature.

Kolanovic, M. and T. Krishnamachari, R. (2017). *Big Data and AI Strategies Machine Learning and Alternative Data Approach to Investing*. J.P.Morgan.

Mannor, S., Jin, X., Han, J., Jin, X., Han, J., Jin, X., Han, J. and Zhang, X. (2011). K-Medoids Clustering. *Encyclopedia of Machine Learning*, [online] pp.564–565. doi:https://doi.org/10.1007/978-0-387-30164-8_426.

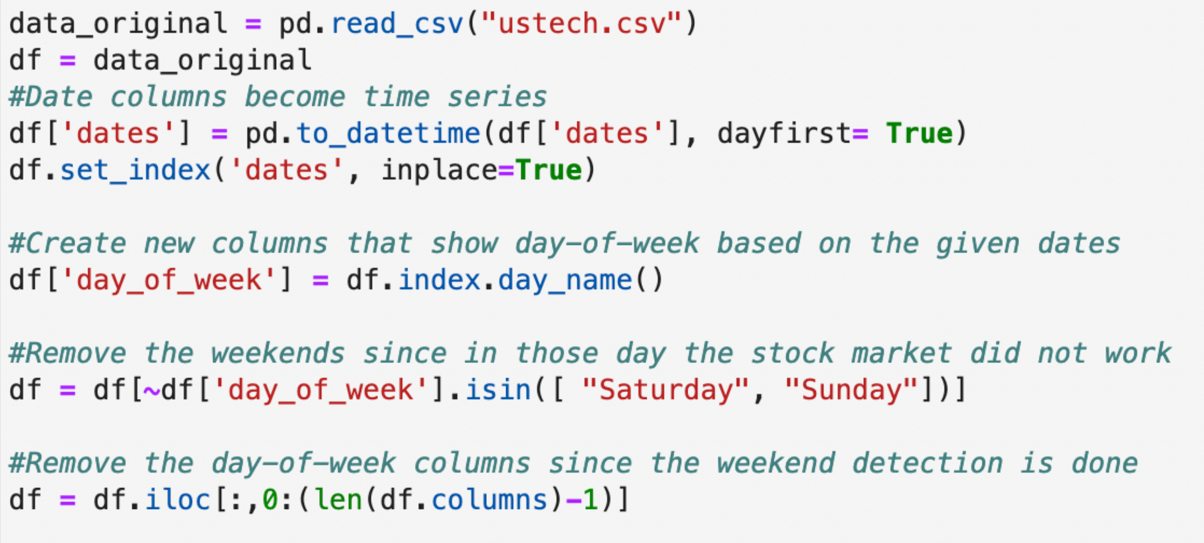
Rousseeuw, P.J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, pp.53–65. doi:https://doi.org/10.1016/0377-0427(87)90125-7.

Vendramin, L., Campello, R.J.G.B. and Hruschka, E.R. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4), p.n/a-n/a. doi:https://doi.org/10.1002/sam.10080.

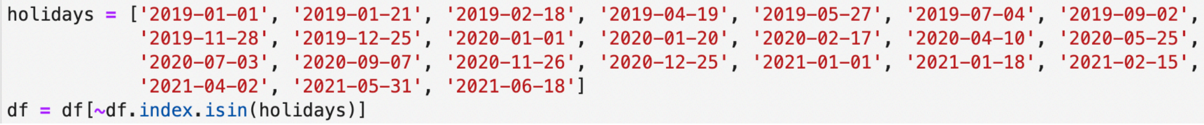
Vijaya, Sharma, S. and Batra, N. (2019). *Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering*. [online] IEEE Xplore. doi:https://doi.org/10.1109/COMITCon.2019.8862232.

VI. Appendixes

Appendix 1: Remove the weekend and holidays.

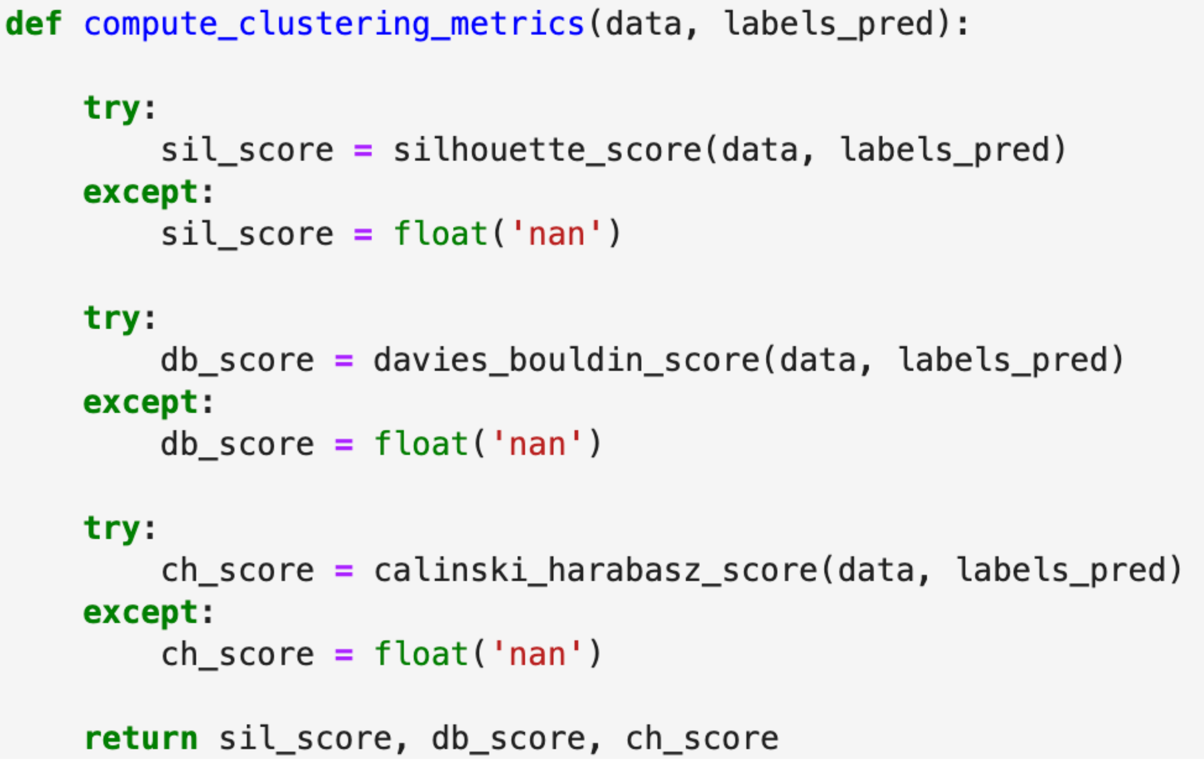


```
data_original = pd.read_csv("ustech.csv")
df = data_original
#Date columns become time series
df['dates'] = pd.to_datetime(df['dates'], dayfirst= True)
df.set_index('dates', inplace=True)

#Create new columns that show day-of-week based on the given dates
df['day_of_week'] = df.index.day_name()

#Remove the weekends since in those day the stock market did not work
df = df[~df['day_of_week'].isin([ "Saturday", "Sunday"])]

#Remove the day-of-week columns since the weekend detection is done
df = df.iloc[:,0:(len(df.columns)-1)]
```



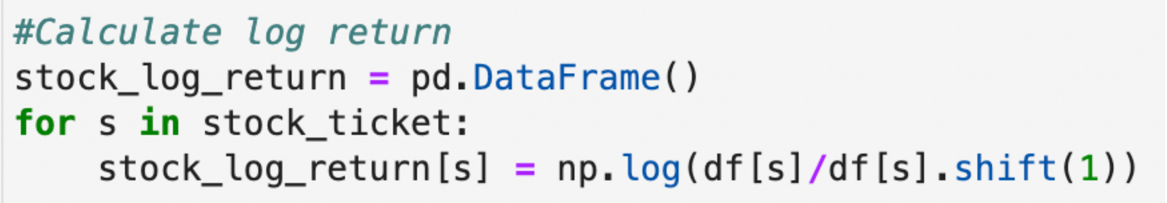
```
holidays = ['2019-01-01', '2019-01-21', '2019-02-18', '2019-04-19', '2019-05-27', '2019-07-04', '2019-09-02',
            '2019-11-28', '2019-12-25', '2020-01-01', '2020-01-20', '2020-02-17', '2020-04-10', '2020-05-25',
            '2020-07-03', '2020-09-07', '2020-11-26', '2020-12-25', '2021-01-01', '2021-01-18', '2021-02-15',
            '2021-04-02', '2021-05-31', '2021-06-18']
df = df[~df.index.isin(holidays)]
```

Appendix 2: Computing clustering metrics



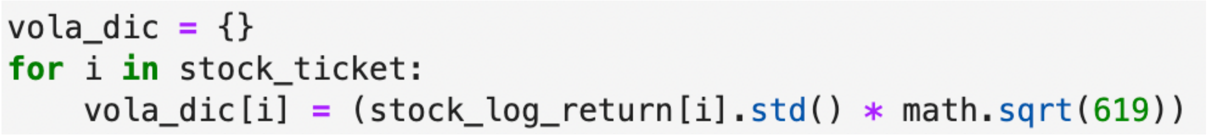
```
def compute_clustering_metrics(data, labels_pred):

    try:
        sil_score = silhouette_score(data, labels_pred)
    except:
        sil_score = float('nan')

    try:
        db_score = davies_bouldin_score(data, labels_pred)
    except:
        db_score = float('nan')

    try:
        ch_score = calinski_harabasz_score(data, labels_pred)
    except:
        ch_score = float('nan')

    return sil_score, db_score, ch_score
```

Appendix 4: Calculate daily log return.



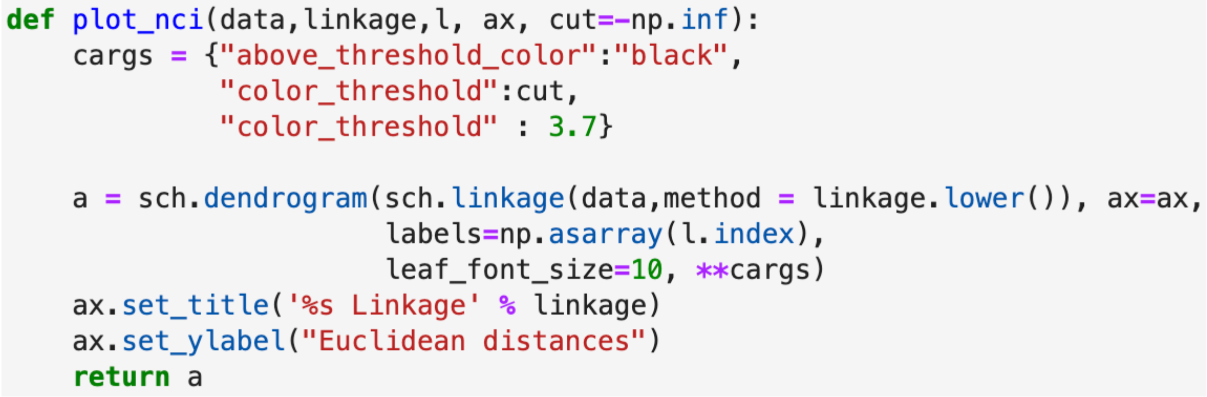
```
#Calculate log return
stock_log_return = pd.DataFrame()
for s in stock_ticket:
    stock_log_return[s] = np.log(df[s]/df[s].shift(1))
```

Appendix 3: Calculate volatility of stock log return.



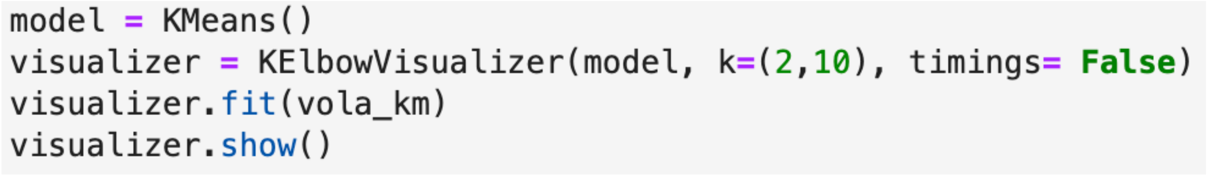
```
vola_dic = {}
for i in stock_ticket:
    vola_dic[i] = (stock_log_return[i].std() * math.sqrt(619))
```

Appendix 4: Drawing the dendrogram function.



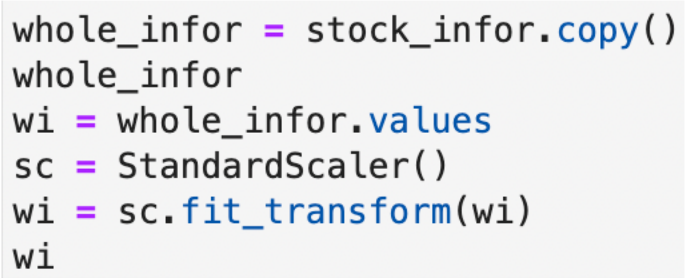
```
def plot_nci(data,linkage,l, ax, cut=-np.inf):
    cargs = {"above_threshold_color":"black",
             "color_threshold":cut,
             "color_threshold" : 3.7}

    a = sch.dendrogram(sch.linkage(data,method = linkage.lower()), ax=ax,
                       labels=np.asarray(l.index),
                       leaf_font_size=10, **cargs)
    ax.set_title('%s Linkage' % linkage)
    ax.set_ylabel("Euclidean distances")
    return a
```

Appendix 4: Elbow for K-Means.



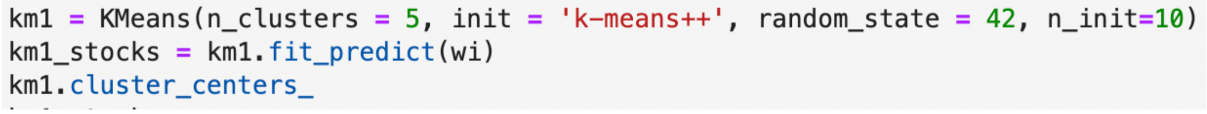
```
model = KMeans()
visualizer = KElbowVisualizer(model, k=(2,10), timings= False)
visualizer.fit(vola_km)
visualizer.show()
```

Appendix 5: Standard Scaler



```
whole_infor = stock_infor.copy()
whole_infor
wi = whole_infor.values
sc = StandardScaler()
wi = sc.fit_transform(wi)
wi
```

Appendix 6: One of the K-mean clustering



```
km1 = KMeans(n_clusters = 5, init = 'k-means++', random_state = 42, n_init=10)
km1_stocks = km1.fit_predict(wi)
km1.cluster_centers_
```