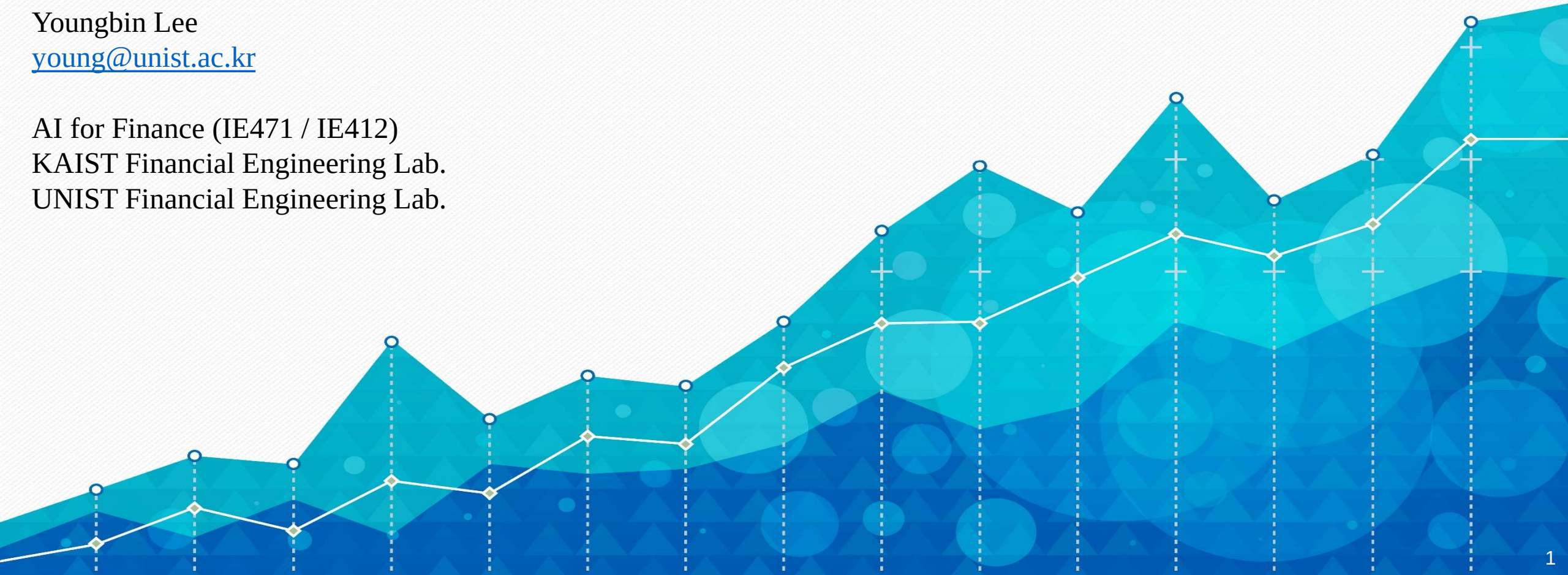
Hands-on Practice on Financial AI Session

Session 2

Within Stock Classification via Clustering

Youngbin Lee
young@unist.ac.kr

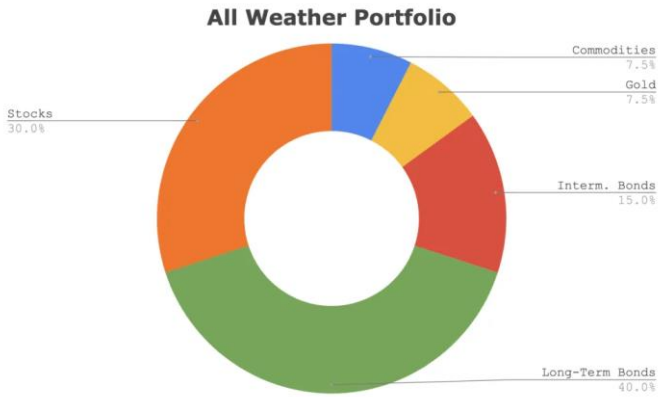
AI for Finance (IE471 / IE412)
KAIST Financial Engineering Lab.
UNIST Financial Engineering Lab.



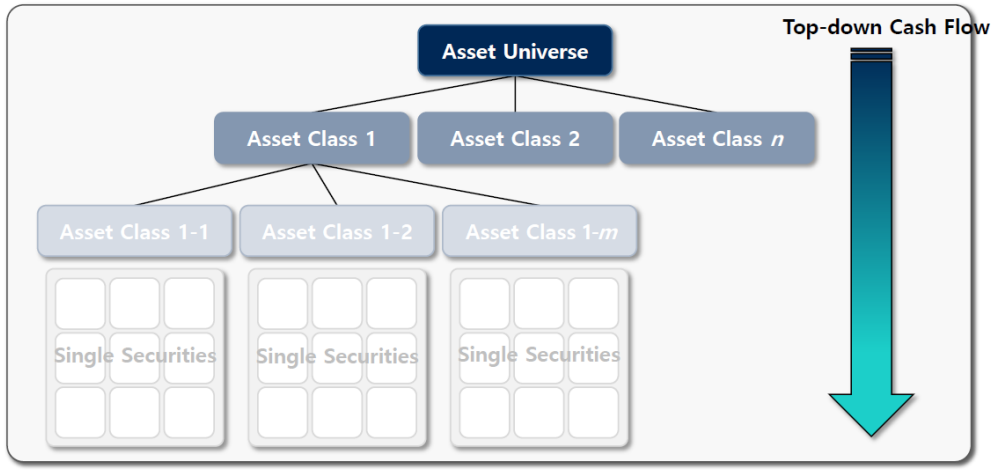
Contents

Introduction

A Problem statement



[Security selection] VS [Asset allocation]



- **Asset allocation matters**
  - Asset allocation decides how the investment should be distributed
    - Equities, bonds, cash, commodities, etc.
  - Asset allocation decisions explain a majority of investment return variabilities [1]

- **How to define asset classes?** (for single-stock universe)
  - Traditional methods
    - Style classification: large-cap, small-cap, growth, value, ...
    - Industry classification: Health care, Technology, Financials, …
    - Market classification: US stocks, developed market stocks, emerging market stocks, …
  - AI methods (we will do)
    - Clustering

[1] Brinson, G. P., Hood, L. R., & Beebower, G. L. (1986). Determinants of portfolio performance. *Financial Analysts Journal*, *42*(4), 39-44.

Introduction

B Impact of asset classification schemes

EXHIBIT 2
Sharpe Ratios of Security Selection and Four Classification Schemes

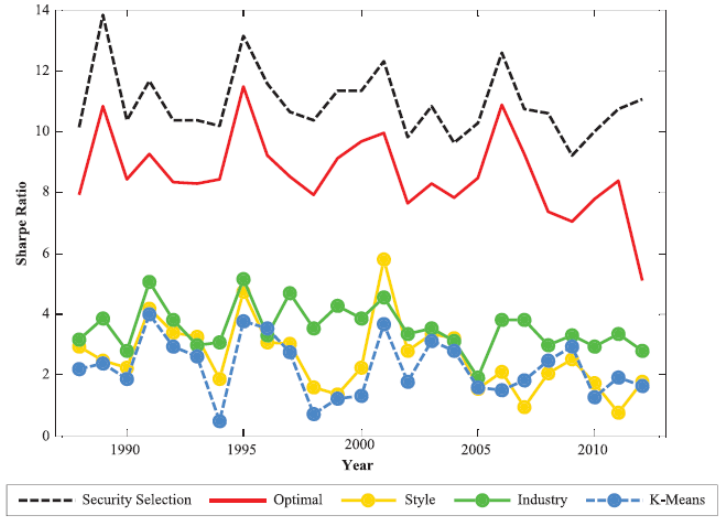
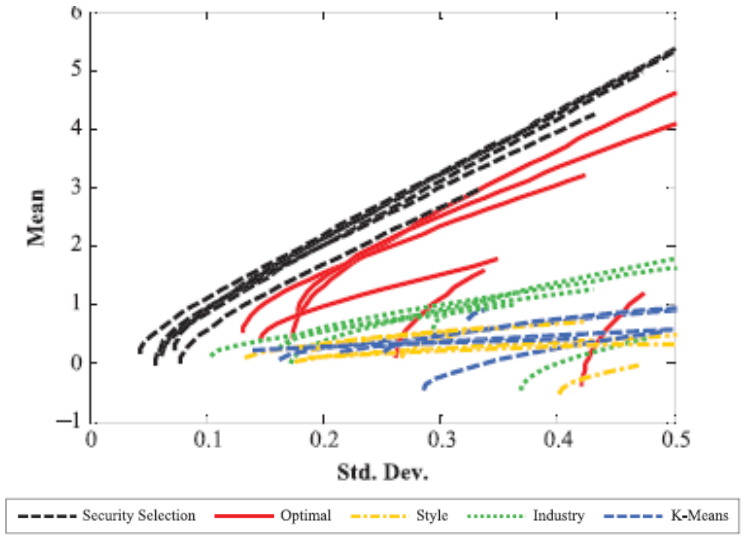


EXHIBIT 3
Efficient Frontiers of Security Selection and Four Classification Schemes

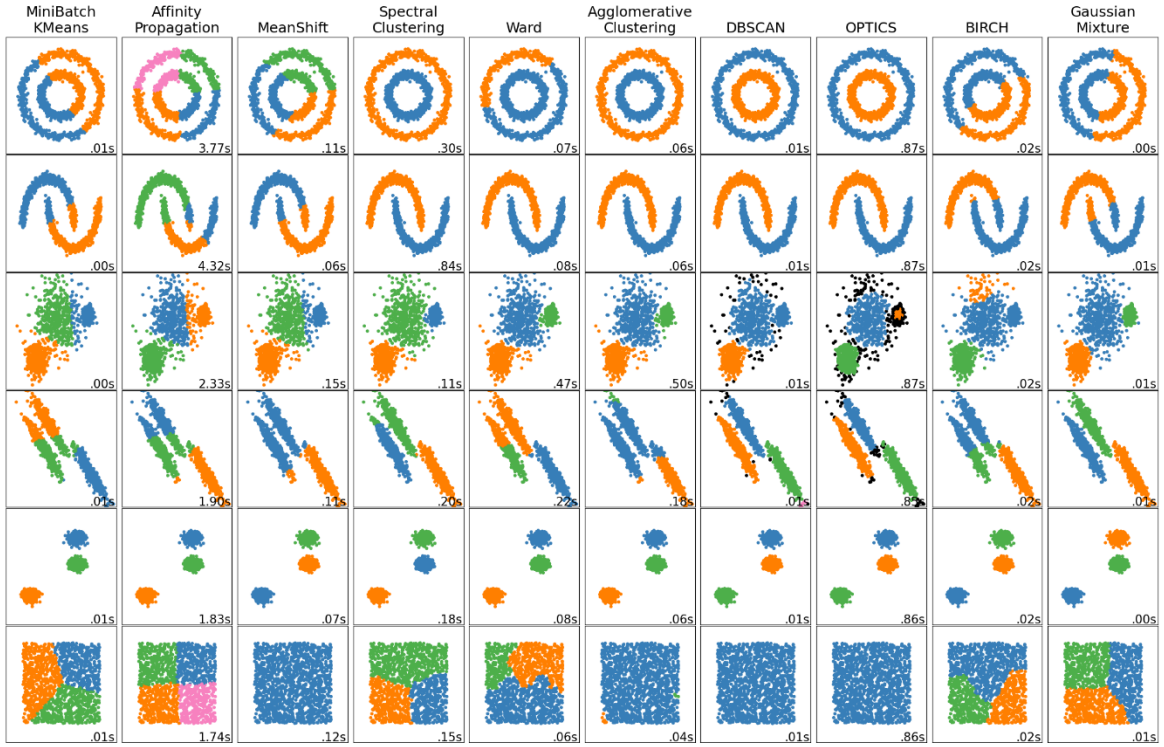


- **Asset classification and investment performance**
  - Investment performance is significantly affected by the employed asset classification scheme [2]
    - In exhibits, each line represents different classification scheme
    - A poor classification scheme choice can result in poor portfolio return
  - Assets in different classes are expected to have different characteristics in market (e.g., return, risk)

[2] Kim, W. C., Lee, Y., & Lee, Y. H. (2014). Cost of Asset Allocation in Equity Market: How Much Do Investors Lose Due to Bad Asset Class Design?. The Journal of Portfolio Management, 41(1), 34-44.

AI model

A Clustering



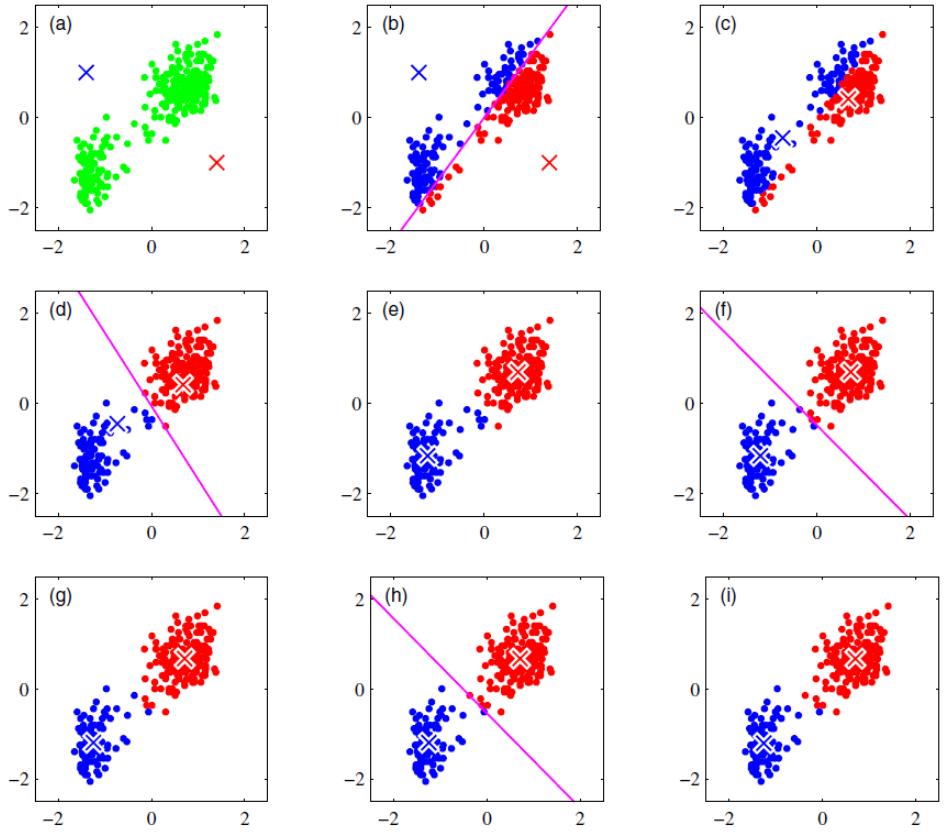
[A comparison of the clustering algorithms in scikit-learn]

- **Unsupervised learning**
  - The goal is to understand data itself by learning patterns from unlabeled data
    - E.g., Dimensionality reduction, Clustering
- **Clustering**
  - The organization of unlabeled data into similarity groups called clusters
    - A cluster is a collection of data items which are similar between them, and dissimilar to data items in other clusters
- **Methods**
  - Conventional algorithms
    - k-means (we will do), DBSCAN, Gaussian Mixtures, Spectral clustering, …
  - Deep clustering
    - Usually combines representation learning with deep neural networks

Source: http://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf

AI model

B k-means clustering



[Illustration of the k-means algorithm]

- **Algorithm**
  - Initialize k centroids
    - k is a user-defined parameter
  - Iterate until convergence
    - Step 1. Assign each data point to its closest centroid
    - Step 2. Move each centroid to the center of data points assigned to it

Source: https://www.cs.toronto.edu/~jlucas/teaching/csc411/lectures/tut8_handout.pdf

AI model

B k-means clustering

The goal is to find a set of vectors $\{\boldsymbol{\mu}_k\}$, the centers of the clusters $k = 1, \dots, K$, such that the cost function J is minimum.

J: the sum of squares of the distances of each data point to its closest vector $\boldsymbol{\mu}_k$

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left||x_n - \boldsymbol{\mu}_k\right||^2$$

where responsibilities $r_{nk} \in \{0,1\}$ are binary indicator variables describing which of the K clusters the data point x_n is assigned to.

Step 1. **Assign each data point to its closest centroid**
This can be done by computing responsibilities.

$$r_{nk} \begin{cases} 1 & if\ k = argmin_j \left||x_n - \boldsymbol{\mu}_j\right||^2 \\ 0 & otherwise \end{cases}$$

Step 2. **Move each centroid to the center of data points assigned to it**
This can be done by computing new sample means for every cluster.

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

- **The two steps are repeated in turn until convergence**
  - Because each step reduces the value of the cost function J, convergence of the algorithm is assured

Source: Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

AI model

C Performance evaluation

- **Silhouette coefficient**
  - A measure of cohesion compared to separation
  - Range: [-1, 1]
    - Higher scores indicate dense and well separated clusters

For data point i in cluster C_I, the **Silhouette coefficient**

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

where

a_i is the mean distance between a sample i and all other points in the same cluster (cohesion)

$$a_i = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$$

b_i is the mean distance between a sample i and all other points in the next nearest cluster (separation)

$$b_i = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

- **Davies-Bouldin index**
  - A measure of average similarity
  - Lowest possible score: 0
    - Lower scores indicate better separated clusters

For cluster C_i for $i = 1, \ldots, k$, the **Davies-Bouldin index**

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

where

R_{ij} is the similarity (the ratio of within-cluster distances to between-cluster distances)

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

s_i is the average distance between each point of cluster i and the centroid of that cluster

d_{ij} is the distance between cluster centroids i and j

Source: Silhouette (clustering) – Wikipedia, Davies–Bouldin index – Wikipedia

Implementation

A Data

WRDS hosts 350+TB of data, partnering with global vendors
Compustat Fundamentals provides standardized North American and global financial statement and market data

- **financial_statements.csv**
  - Quarterly financial statement of 2094 US stocks
  - Date
    - Dec 31 2020 or Jan 4 2021

	datadate_x	tic	name	gsector	market_cap	PBR	PER	EPS	ROE	net_income	net_cash_flow	volatility
0	20210104	PNW	PINNACLE WEST CAPITAL CORP	utilities	8.905297e+09	1.099	18.112	5.63	0.082	550.559	49.685	0.016509
1	20210104	ABT	ABBOTT LABORATORIES	health_care	1.933824e+11	6.183	96.174	1.89	0.107	4473.000	2907.000	0.013776
2	20210104	ALK	ALASKA AIR GROUP INC	industrial	6.089215e+09	1.627	10.233	-5.64	-0.145	-1324.000	1154.000	0.035188
3	20210104	MATX	MATSON INC	industrial	2.442149e+09	1.944	20.084	2.86	0.107	193.100	-8.700	0.030216
4	20210104	Y	ALLEGHANY CORP	financial	8.430382e+09	1.005	26.306	-1.75	-0.003	101.754	-411.621	0.021499
...	...	...	...	...	...	...	...	...	...	...	...	...
2089	20210104	NES	NUVERRA ENVIRONMENTAL SOLUTN	energy	3.280576e+07	0.243	-0.513	-4.73	-0.436	-44.143	9.990	0.062768
2090	20210104	MHH	MASTECH DIGITAL INC	industrial	1.752520e+08	3.142	32.530	0.91	0.206	9.861	4.883	0.036276
2091	20210104	ISDR	ISSUER DIRECT CORP	IT	6.473660e+07	2.384	58.476	0.50	0.070	2.106	3.785	0.031666
2092	20210104	AMPE	AMPIO PHARMACEUTICALS INC	health_care	2.738814e+08	27.435	-20.666	-0.07	-2.600	-15.894	10.814	0.100264
2093	20210104	ENSV	ENSERVCO CORP	energy	9.617410e+06	3.202	-1.591	-0.60	-3.031	-2.509	0.804	0.350034

2094 rows × 12 columns

- **security_daily.csv**
  - Daily close price of US stocks
  - Period
    - Jan 4 2021 ~ Jul 4 2021

	datadate	tic	conm	prccd
254	20210104	AIR	AAR CORP	34.3600
255	20210105	AIR	AAR CORP	36.0100
256	20210106	AIR	AAR CORP	38.4500
257	20210107	AIR	AAR CORP	38.6200
258	20210108	AIR	AAR CORP	37.9500
...	...	...	...	...
9692678	20211227	DTRUY	DAIMLER TRUCK HOLDING AG	18.2717
9692679	20211228	DTRUY	DAIMLER TRUCK HOLDING AG	18.1591
9692680	20211229	DTRUY	DAIMLER TRUCK HOLDING AG	18.4389
9692681	20211230	DTRUY	DAIMLER TRUCK HOLDING AG	18.2940
9692682	20211231	DTRUY	DAIMLER TRUCK HOLDING AG	18.3600

5078928 rows × 4 columns

Implementation

B-1 k-means clustering with financial statements

	datadate_x	tic	name	gsector	market_cap	PBR	PER	EPS	ROE	net_income	net_cash_flow	volatility
0	20210104	PNW	PINNACLE WEST CAPITAL CORP	utilities	8.905297e+09	1.099	18.112	5.63	0.082	550.559	49.685	0.016509
1	20210104	ABT	ABBOTT LABORATORIES	health_care	1.933824e+11	6.183	96.174	1.89	0.107	4473.000	2907.000	0.013776
2	20210104	ALK	ALASKA AIR GROUP INC	industrial	6.089215e+09	1.627	10.233	-5.64	-0.145	-1324.000	1154.000	0.035188
3	20210104	MATX	MATSON INC	industrial	2.442149e+09	1.944	20.084	2.86	0.107	193.100	-8.700	0.030216
4	20210104	Y	ALLEGHANY CORP	financial	8.430382e+09	1.005	26.306	-1.75	-0.003	101.754	-411.621	0.021499
...	...	...	...	...	...	...	...	...	...	...	...	...
2089	20210104	NES	NUVERRA ENVIRONMENTAL SOLUTN	energy	3.280576e+07	0.243	-0.513	-4.73	-0.436	-44.143	9.990	0.062768
2090	20210104	MHH	MASTECH DIGITAL INC	industrial	1.752520e+08	3.142	32.530	0.91	0.206	9.861	4.883	0.036276
2091	20210104	ISDR	ISSUER DIRECT CORP	IT	6.473660e+07	2.384	58.476	0.50	0.070	2.106	3.785	0.031666
2092	20210104	AMPE	AMPIO PHARMACEUTICALS INC	health_care	2.738814e+08	27.435	-20.666	-0.07	-2.600	-15.894	10.814	0.100264
2093	20210104	ENSV	ENSERVCO CORP	energy	9.617410e+06	3.202	-1.591	-0.60	-3.031	-2.509	0.804	0.350034

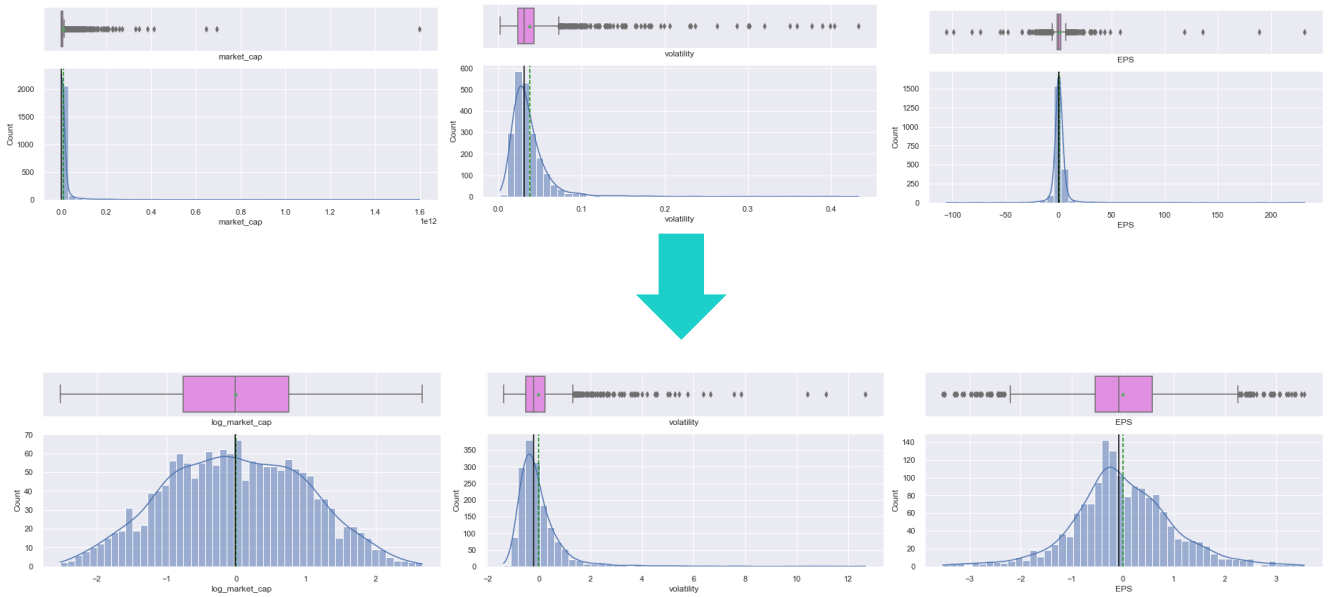
2094 rows × 12 columns

- **Identifying information**
  - Tic: Ticker symbol
  - Name: Company name
  - gsector: GIC sectors
    - Global Industry Classification Standard

- **Cluster features**
  - **Market cap**: Total dollar market value of a company's outstanding shares of stock (size)
  - **PER**: Price Earnings Ratio (value)
  - **EPS**: Earnings Per Share
  - **ROE**: Return on Equity (profitability)
  - **Net income**: Revenues minus expenses, interest, and taxes
  - **Net cash flow**: The sum of a company's cash inflows and outflows over a period of time
  - **Volatility**: Standard deviation of daily log returns over the last 3 months

Implementation

B-1 k-means clustering with financial statements



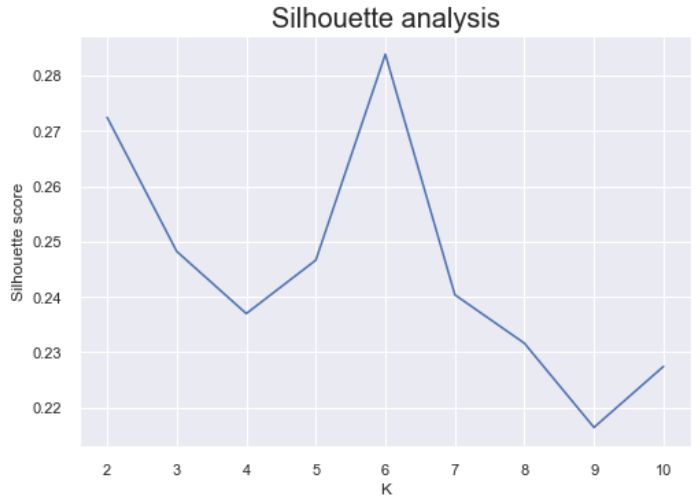
- **Removing outliers**
  - EPS, net income, PER, ROE, net cash flow
    - 3% from each of both tails are removed to reduce the effect of outliers
    - Because even a small amount of outliers can significantly distort the clustering results

- **Scaling features**
  - Market cap
    - A logarithmic transformation (base 10) is applied to make small values and large values closer
    - Because the distribution is highly skewed towards large values
  - All features
    - A standard scaling is applied to make every feature have the same scales
    - Because you will measure Euclidean distance for k-means clustering

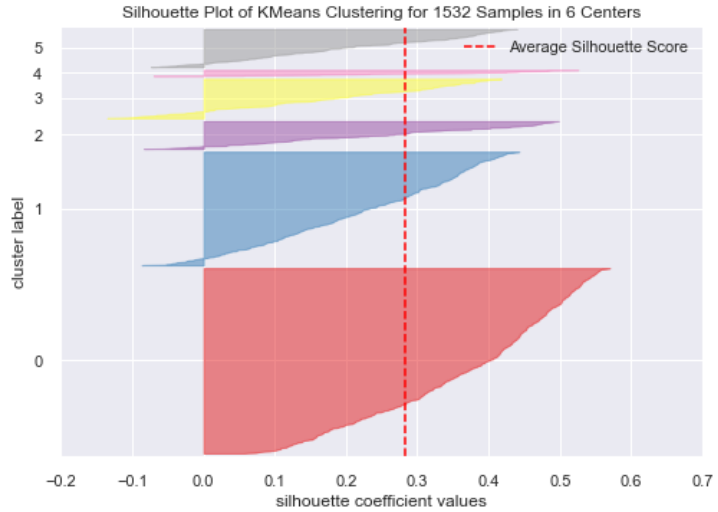
Source: https://github.com/rochitasundar/Stock-clustering-using-ML

Implementation

B-1 k-means clustering with financial statements



- **Silhouette analysis**
  - The optimal number of clusters is determined to be 6
  - Because the score is highest with k=6

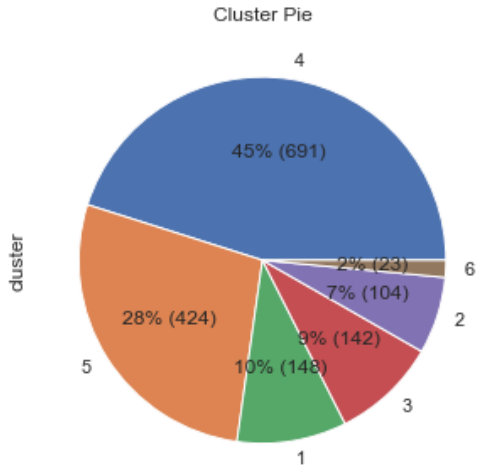


- **Silhouette plot**
  - Silhouette scores for each data point are visualized
  - Cluster size is visualized from the thickness of the plot

Implementation

B-1 k-means clustering with financial statements

```
Silhouette score:  0.284
Davies-Bouldin score:  1.203
```

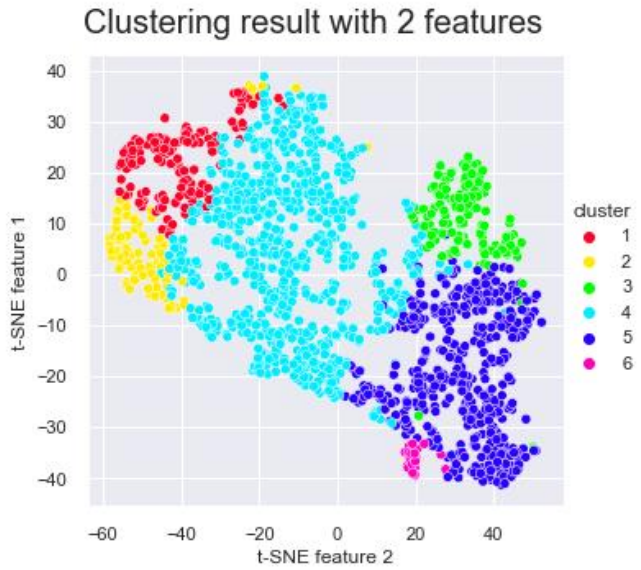


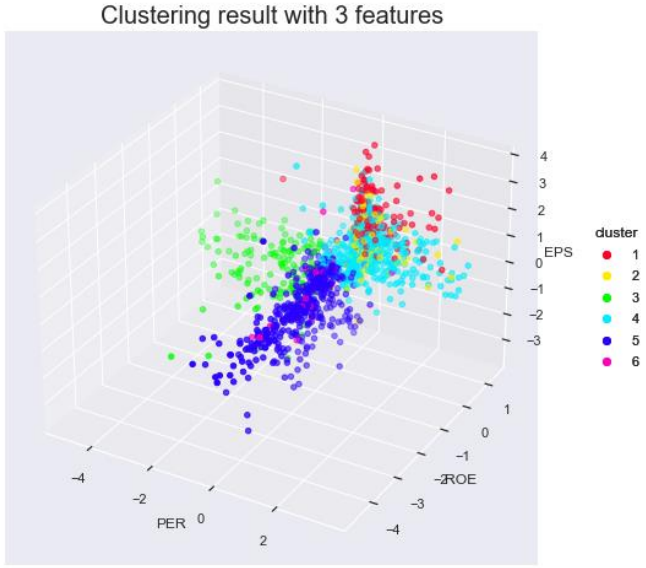

```
# fit kmeans
kmeans = KMeans(n_clusters=num_clusters,
                random_state=2022).fit(df)

# print out scores
print('Silhouette score: ',
    round(metrics.silhouette_score(df, kmeans.labels_, metric='euclidean'), 3))
print('Davies-Bouldin score: ',
  round(davies_bouldin_score(df, kmeans.labels_), 3))

# add cluster labels to dataset
df['cluster'] = kmeans.labels_
```

- **Run k-means**
  - Fitting scaled data into k-means model with k=6
    - The cluster size is visualized from pie chart
  - The scores indicate goodness of fit based on similarity
    - However, they do not necessarily represent the actual performance of clustering as we are dealing with unlabeled data
    - The distributions of each feature will be investigated in later slides

Implementation

B-1 k-means clustering with financial statements





- **Clustering results**
  - Clustering was done with 7 features, but data points in each cluster are visualized with 2 or 3 features in standardized scale
  - t-SNE (t-distributed Stochastic Neighbor Embedding) is used to visualize high-dimensional data
    - A dimensionality reduction technique that embeds the points from a higher dimension to a lower dimension with an objective of optimizing local distances (trying to preserve the neighborhood of that point)
      - › Unlike PCA, it is a non-linear method

Implementation

B-2 k-means clustering with style factors

	datadate_x	tic	name	gsector	market_cap	PBR	PER	EPS	ROE	net_income	net_cash_flow	volatility
0	20210104	PNW	PINNACLE WEST CAPITAL CORP	utilities	8.905297e+09	1.099	18.112	5.63	0.082	550.559	49.685	0.016509
1	20210104	ABT	ABBOTT LABORATORIES	health_care	1.933824e+11	6.183	96.174	1.89	0.107	4473.000	2907.000	0.013776
2	20210104	ALK	ALASKA AIR GROUP INC	industrial	6.089215e+09	1.627	10.233	-5.64	-0.145	-1324.000	1154.000	0.035188
3	20210104	MATX	MATSON INC	industrial	2.442149e+09	1.944	20.084	2.86	0.107	193.100	-8.700	0.030216
4	20210104	Y	ALLEGHANY CORP	financial	8.430382e+09	1.005	26.306	-1.75	-0.003	101.754	-411.621	0.021499
...	...	...	...	...	...	...	...	...	...	...	...	...
2089	20210104	NES	NUVERRA ENVIRONMENTAL SOLUTN	energy	3.280576e+07	0.243	-0.513	-4.73	-0.436	-44.143	9.990	0.062768
2090	20210104	MHH	MASTECH DIGITAL INC	industrial	1.752520e+08	3.142	32.530	0.91	0.206	9.861	4.883	0.036276
2091	20210104	ISDR	ISSUER DIRECT CORP	IT	6.473660e+07	2.384	58.476	0.50	0.070	2.106	3.785	0.031666
2092	20210104	AMPE	AMPIO PHARMACEUTICALS INC	health_care	2.738814e+08	27.435	-20.666	-0.07	-2.600	-15.894	10.814	0.100264
2093	20210104	ENSV	ENSERVCO CORP	energy	9.617410e+06	3.202	-1.591	-0.60	-3.031	-2.509	0.804	0.350034

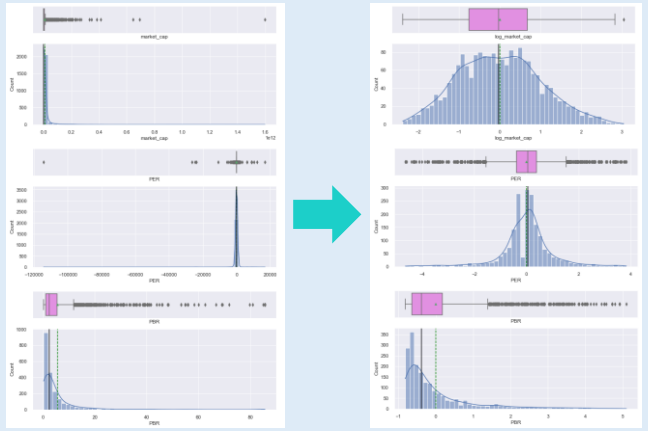
2094 rows × 12 columns

- **Identifying information**
  - Tic: Ticker symbol
  - Name: Company name
  - gsector: GIC sectors
    - Global Industry Classification Standard
- **Cluster features**
  - Market cap: Total dollar market value of a company's outstanding shares of stock
  - PBR: Price-to-Book Ratio
  - PER: Price Earnings Ratio

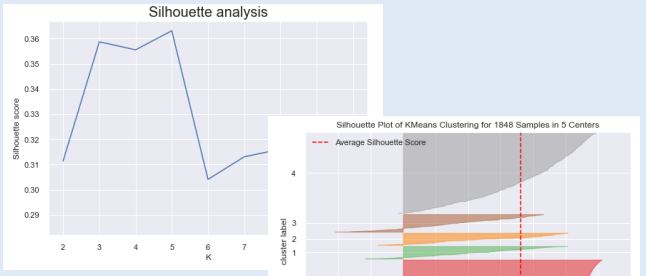
Implementation

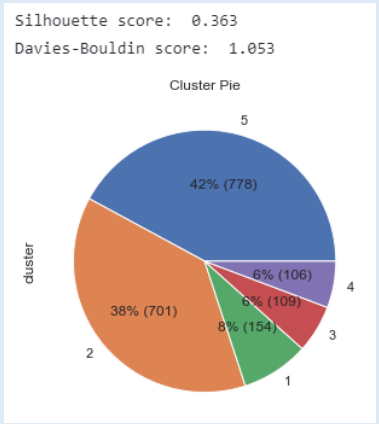
B-2 k-means clustering with style factors

(1) Feature scaling

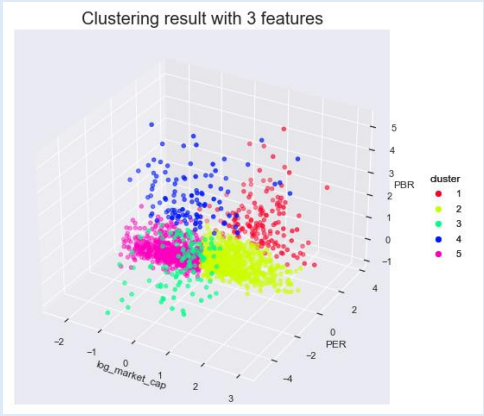


(2) k-means clustering





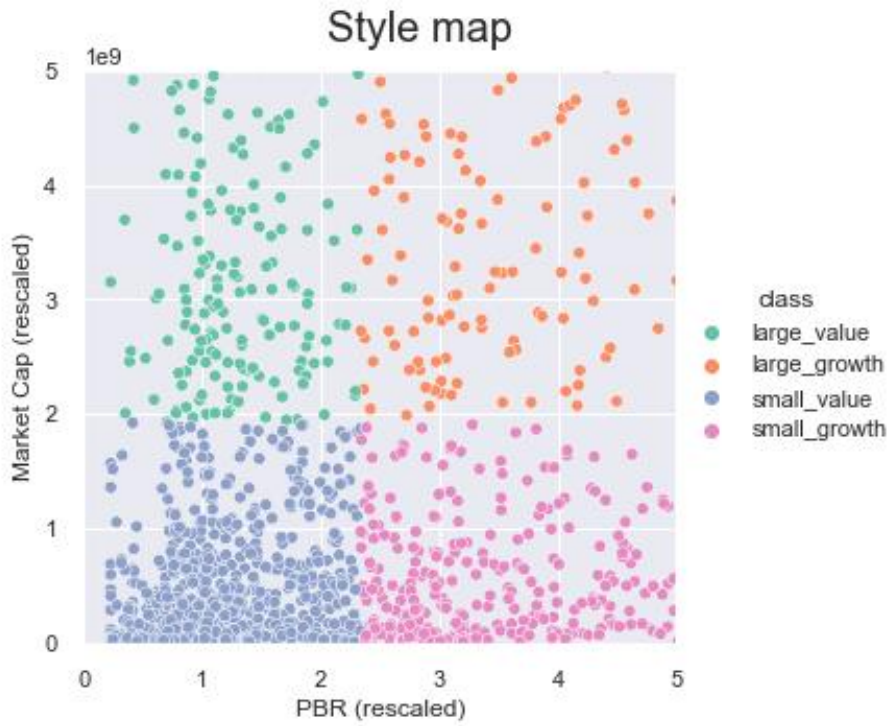
(3) Clustering result



- **Same process as *B-1*, with different number of clusters (k=5)**

Implementation

B-3 Traditional method (Style classification)



- **Style classification**
  - Size (Market cap)
    - Stocks in the top 40% were classified as large
    - Stocks in the bottom 60% were classified as small
  - Value (PBR)
    - Stocks in the upper half were classified as value
    - Stocks in the lower half were classified as growth
  - There are 4 classes
    - Large-Value, Large-Growth, Small-Value, Small-Growth

Implementation

C Analysis

B-1. k-means with financial statements

	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6
1	INTUITIVE SURGICAL INC	FIDELITY NATIONAL INFO SVCS	PINTEREST INC	DOMINION ENERGY INC	BRIDGEBIO PHARMA INC	ARCTURUS THERAPETCS HOLD INC
2	FISERV INC	KRAFT HEINZ CO	SEAGEN INC	ZIMMER BIOMET HOLDINGS INC	INVITAE CORP	ASSEMBLY BIOSCIENCES INC
3	DUKE ENERGY CORP	MARRIOTT INTL INC	ZENDESK INC	WILLIAMS COS INC	NOVAVAX INC	CHINA AUTOMOTIVE SYSTEMS INC
4	GLOBAL PAYMENTS INC	IQVIA HOLDINGS INC	AVALARA INC	ROLLINS INC	TG THERAPEUTICS INC	HEAT BIOLOGICS INC
5	NORFOLK SOUTHERN CORP	KINDER MORGAN INC	SAREPTA THERAPEUTICS INC	AES CORP (THE)	PTC THERAPEUTICS INC	WESTWATER RESOURCES INC

B-2. k-means with style factors

	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
1	META PLATFORMS INC	JPMORGAN CHASE & CO	EXACT SCIENCES CORP	LILLY (ELI) & CO	MATADOR RESOURCES CO
2	JOHNSON & JOHNSON	UNITEDHEALTH GROUP INC	BIOMARIN PHARMACEUTICAL INC	TEXAS INSTRUMENTS INC	SMILEDIRECTCLUB INC
3	PAYPAL HOLDINGS INC	BANK OF AMERICA CORP	LIVE NATION ENTERTAINMENT	LOCKHEED MARTIN CORP	MACROGENICS INC
4	COCA-COLA CO	VERIZON COMMUNICATIONS INC	GUARDANT HEALTH INC	ILLINOIS TOOL WORKS	COHERUS BIOSCIENCES INC
5	MERCK & CO	COMCAST CORP	HOWMET AEROSPACE INC	MODERNA INC	MAGNOLIA OIL & GAS CORP

B-3. Traditional method (Style classification)

	cluster_large_value	cluster_large_growth	cluster_small_value	cluster_small_growth
1	JPMORGAN CHASE & CO	AMAZON.COM INC	FULTON FINANCIAL CORP	SOUTHWESTERN ENERGY CO
2	BANK OF AMERICA CORP	TESLA INC	PDC ENERGY INC	MEDNAX INC
3	COMCAST CORP	META PLATFORMS INC	ALLEGHENY TECHNOLOGIES INC	PRECIGEN INC
4	AT&T INC	JOHNSON & JOHNSON	FIRST MERCHANTS CORP	NGM BIOPHARMACEUTICAL INC
5	EXXON MOBIL CORP	MASTERCARD INC	WESBANCO INC	TPI COMPOSITES INC

[Five most popular firms in cluster]

Checklist for Asset Classes

1. Do they have distinct characteristics?

- Distribution of financial features

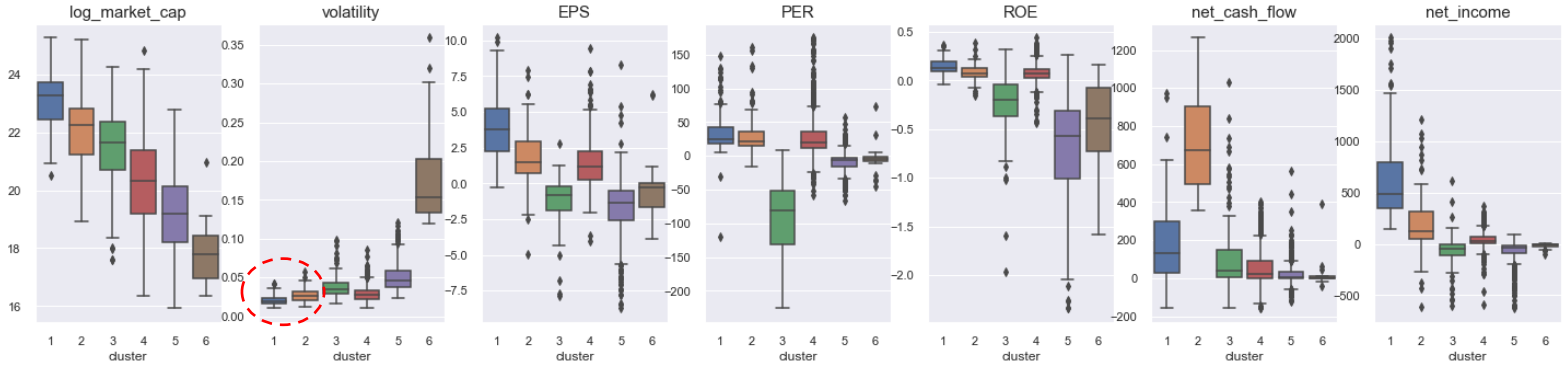
2. Do they move differently in the market?

- Return correlations
- Return and risk characteristics

Implementation

C Analysis

B-1. k-means with financial statements



[Boxplot]

cluster	log_market_cap	volatility	EPS	PER	ROE	net_cash_flow	net_income
1	23.143487	0.020301	4.041689	33.972872	0.149885	176.990230	655.284770
2	22.059823	0.026975	1.913433	31.048798	0.077894	717.509567	193.128269
3	21.408351	0.038003	-1.224718	-91.481338	-0.248669	113.574099	-71.282655
4	20.270595	0.029023	1.375094	28.247340	0.073624	57.531339	39.433779
5	19.140726	0.050010	-1.826705	-9.228061	-0.707828	21.716849	-79.047538
6	17.812397	0.182299	-0.408696	-4.603261	-0.472870	20.841652	-18.697435

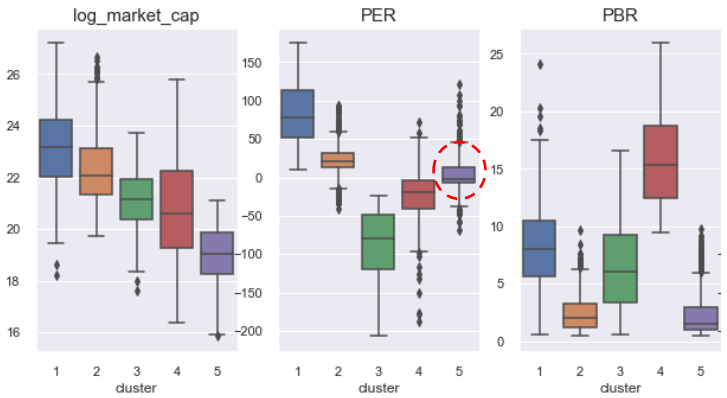
[Mean table]

- **Distribution of financial features**
  - Clusters with higher market cap tend to have lower volatility
  - Stocks in cluster 1 have highest mean value of net income
  - Financial properties are similar within clusters, and distinct across clusters

Implementation

C Analysis

B-2. k-means with style factors



[Boxplot]

cluster	log_market_cap	PER	PBR
1	23.124188	84.869597	8.337851
2	22.323108	23.294966	2.457762
3	21.105701	-85.737743	6.281294
4	20.724720	-28.471811	15.999170
5	18.978872	3.100269	2.243172

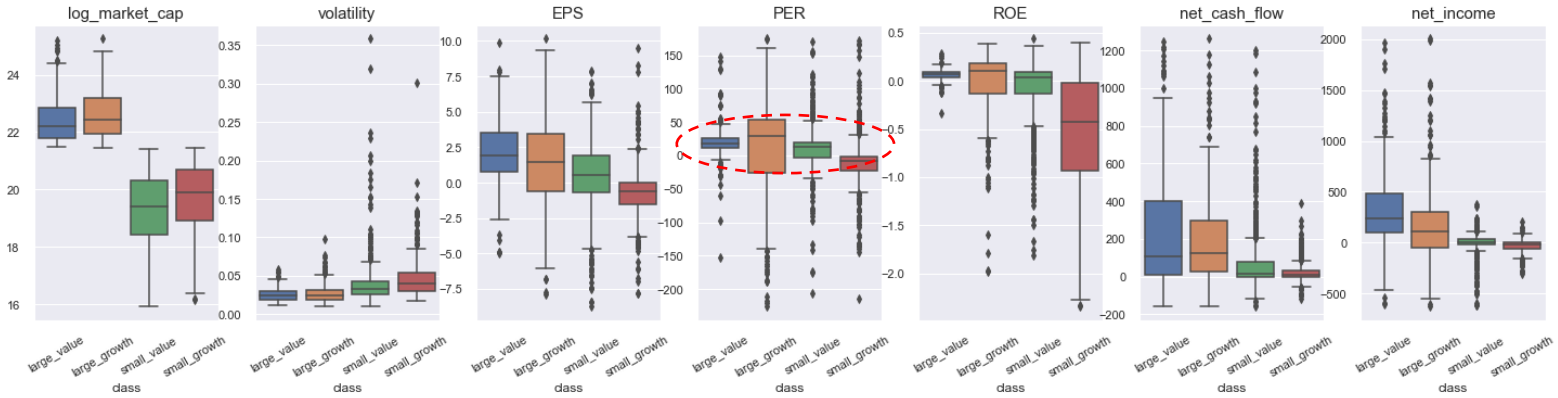
[Mean table]

- **Distribution of financial features**
  - Stocks in cluster 5 tend to have values of PER around zero
  - Stocks in cluster 4 have highest mean value of PBR
  - Financial properties are similar within clusters, and distinct across clusters

Implementation

C Analysis

B-3. Traditional method (Style classification)



[Boxplot]

class	log_market_cap	volatility	EPS	PER	ROE	net_cash_flow	net_income
large_growth	22.595455	0.026831	1.363923	11.311698	-0.022923	199.396682	177.956045
large_value	22.448989	0.025575	2.124186	20.730762	0.060035	259.638401	349.697320
small_growth	19.698659	0.046631	-0.642689	-7.868692	-0.542556	22.903694	-33.232794
small_value	19.297846	0.038938	0.394488	10.496966	-0.079792	77.808894	-3.053063

[Mean table]

- **Distribution of financial features**
  - Variations are higher
  - Large stocks exhibit similar values of features to each other, and so do small stocks
  - Financial properties are dissimilar within clusters, and indistinguishable across clusters

Implementation

C Analysis

Average correlations within and between clusters

	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6
Cluster_1	0.262514	0.295900	0.136110	0.219293	0.111795	0.070901
Cluster_2		0.417184	0.160402	0.306824	0.155933	0.092798
Cluster_3			0.249587	0.154893	0.190726	0.187356
Cluster_4				0.240692	0.142497	0.100783
Cluster_5					0.177150	0.183045
Cluster_6						0.223911

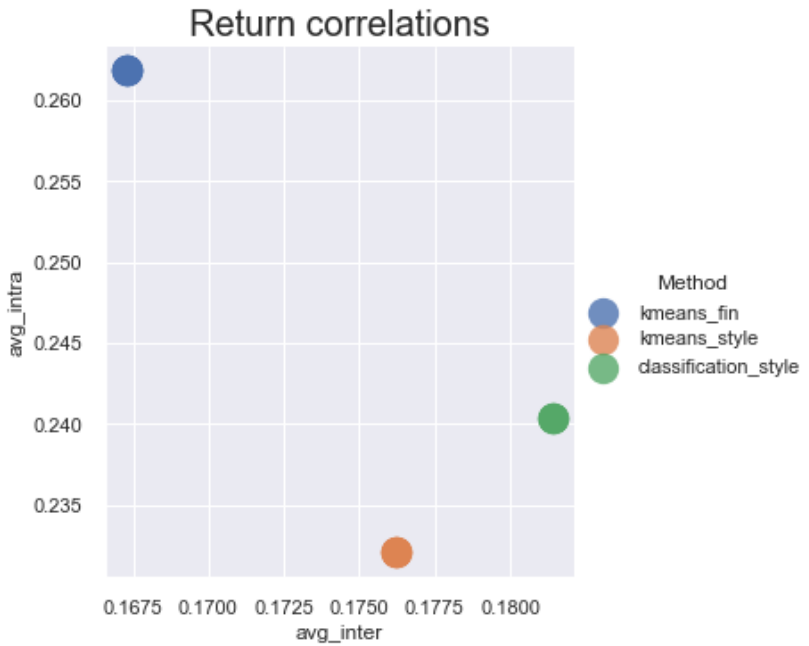
[B-1. k-means with financial statements]

	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5
Cluster_1	0.243923	0.212901	0.201449	0.173543	0.146562
Cluster_2		0.310553	0.166337	0.135527	0.186204
Cluster_3			0.233048	0.218457	0.166173
Cluster_4				0.211603	0.154999
Cluster_5					0.161168

[B-2. k-means with style factors]

	Cluster_1	Cluster_2	Cluster_3	Cluster_4
Cluster_1	0.225191	0.214156	0.171565	0.164185
Cluster_2		0.345019	0.146391	0.240435
Cluster_3			0.188593	0.151701
Cluster_4				0.202643

[B-3. Traditional method (Style classification)]

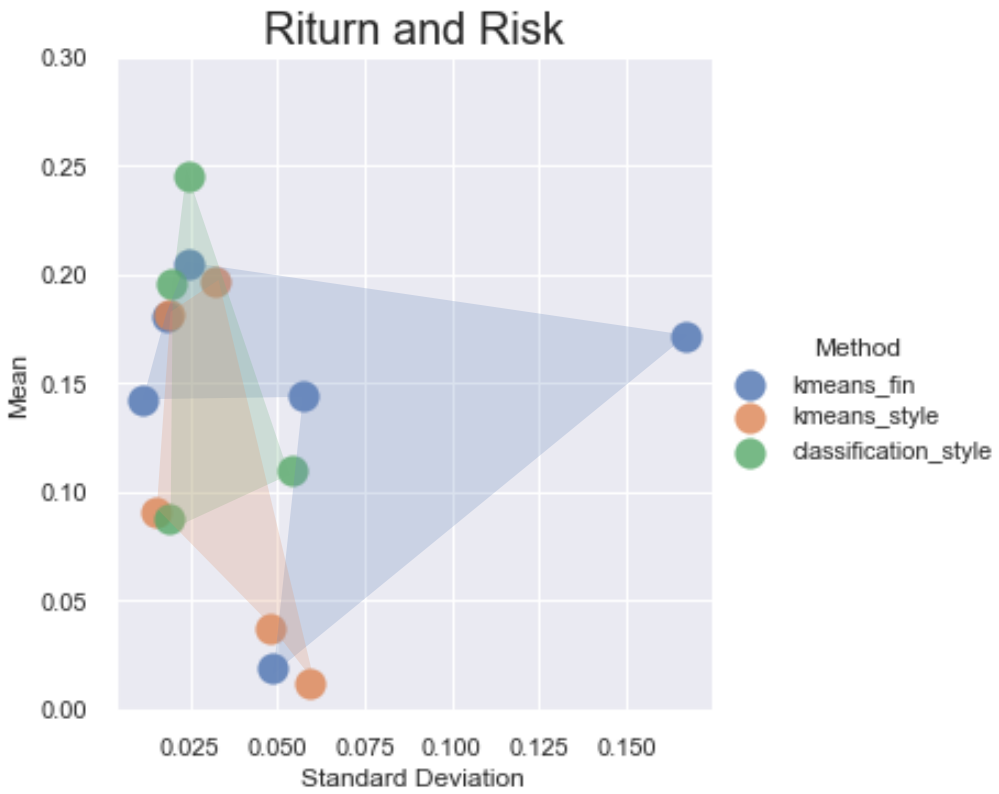


- avg_intra: Average correlations of daily log return within blusters
- avg_inter: Average correlations of daily log return between clusters

- **Return correlations**
  - We expect assets in the same cluster would move similarly, and assets in different clusters would move differently
  - K-means clustering with financial statements (blue) have the most desired correlations between stocks
    - Higher intra-correlations and lower inter-correlations

Implementation

C Analysis



- Each point stands for an asset class from three different methods
- Y-axis: Mean of average daily log returns in each class
- X-axis: Standard Deviation of average daily log returns in each class

- **Return and Risk**
  - We expect assets in different clusters would have distinct return and risk characteristics (for the purpose of diversification)
  - K-means clustering with financial statements (blue) have the most desired return and risk characteristics
    - Clusters have diverse mean and standard deviation of returns

Conclusion

A Summary

- A key determinant of portfolio returns is **asset allocation**, which requires proper asset classification
  - In this session, we focus on within-stock classification
- **AI models** allow you to classify assets in a different way than traditional methods
  - Existing asset classification methods include style-based and industry-based
- **K-means clustering** is one of the simplest and popular unsupervised machine learning algorithms
  - It is implemented by iterating two steps
    - Step 1. Assigning each data point to its closest centroid
    - Step 2. Updating centroids to the center of data points assigned to it
  - When clustering data without labels, we can evaluate performance with metrics such as Silhouette coefficient and Davies-Bouldin index
- As a result of **classifying stocks via clustering** with financial properties, stocks are grouped differently from the existing method
  - Financial features are distinct across classes
  - Stocks in different classes have different return and risk characteristics

Conclusion

B Homework

"Implement stock classification via clustering on your own"

- **Main**
  - K-means clustering
    - with your own number of clusters (k)
- **Extension**
  - Another clustering method
    - E.g., DBSCAN, Gaussian mixtures, Spectral clustering, …

Questions & Answers

- **E-mail: young@unist.ac.kr**

Thank You.