



Project 7

Trade&Ahead

Unsupervised learning

Date: 11/Mar/2023

Name: Ann Mariya Jomon

Content / Agenda

1. Executive Summary
2. Business Problem Overview and Solution Approach
3. EDA Results
4. Data Preprocessing
5. K-Means Clustering
6. Hierarchical Clustering
7. Appendix

The background is a light cream color with a large, faint, light-yellow circular shape in the center. Scattered around are various geometric elements: a green circle at the top center, a red circle at the bottom right, a yellow circle at the bottom left, and several dark blue circles and outlines of circles. One dark blue circle is at the top left, another at the top right, and a third at the bottom center. There are also outlines of circles: one at the top left, one at the top right, one at the bottom left, and one at the bottom right.

OI

Executive Summary

Actionable insights & Recommendations

- Trade&Ahead can use the clusters identified to segregate different companies based on its level of riskiness and returns.
- To provide a high-quality advice, it would be better if Trade&Ahead can analyze their client needs and risk appetites.
- Further, the company could try to identify the risk-return dynamics of different sectors.
- The clustering should be done at frequent intervals to ensure unexpected changes are taken into account.
- Trade&Ahead can also rank the clusters based on its level of risk or the returns it generate. This will help to easily identify suitable stocks for their clients.

The background is a light cream color with a large, faint yellow circular gradient in the center. Scattered around are several geometric elements: a green circle at the top center, a red circle at the bottom right, a yellow circle at the bottom left, and several dark blue circles and thin black outlines of circles. The number '02' is prominently displayed in a dark blue serif font inside a light yellow rounded rectangle on the left side.

02

Business problem overview and Solution approach

Business problem overview & solution approach

Problem overview

- Trade&Ahead is a financial consultancy firm who provide their customers with personalized investment strategies.
- They wish to analyze the data, group the stocks based on the attributes provided, and gain insights about the characteristics of each group based on data of few companies listed on the New York Stock Exchange.

Solution approach

1. EDA
2. Feature Scaling
3. K-means clustering
4. Cluster profiling based on k-means
5. Hierarchical clustering
6. Cluster profiling based on hierarchical clustering
7. Comparison between k-means and hierarchical clustering
8. Actionable insights and recommendations

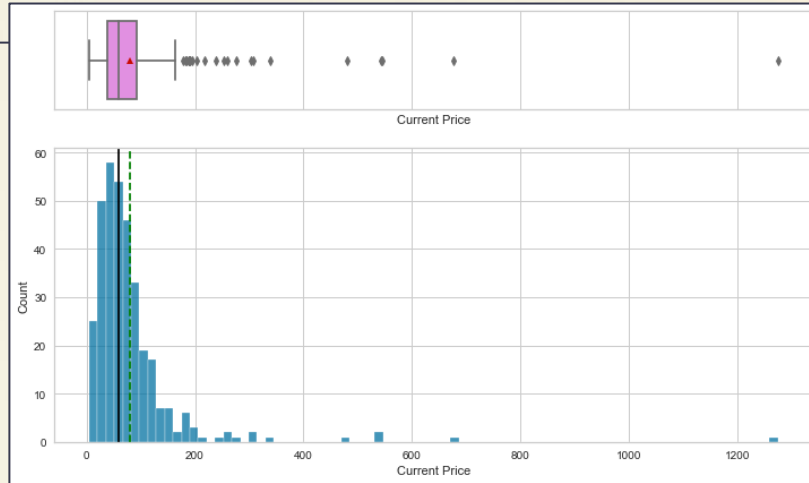
The background is a light cream color with a large, faint, light-yellow circular gradient in the center. Scattered around are various geometric elements: a green circle at the top center, a red circle at the bottom right, a yellow circle at the bottom left, and several dark blue circles of different sizes. Some of these dark blue circles are positioned on the outlines of larger, thin-lined circles. The number '03' is displayed in a dark blue serif font, enclosed within a white rounded rectangle with a thin dark blue border. A small dark blue circle is also located to the right of this rectangle.

03

EDA Results

Univariate Analysis

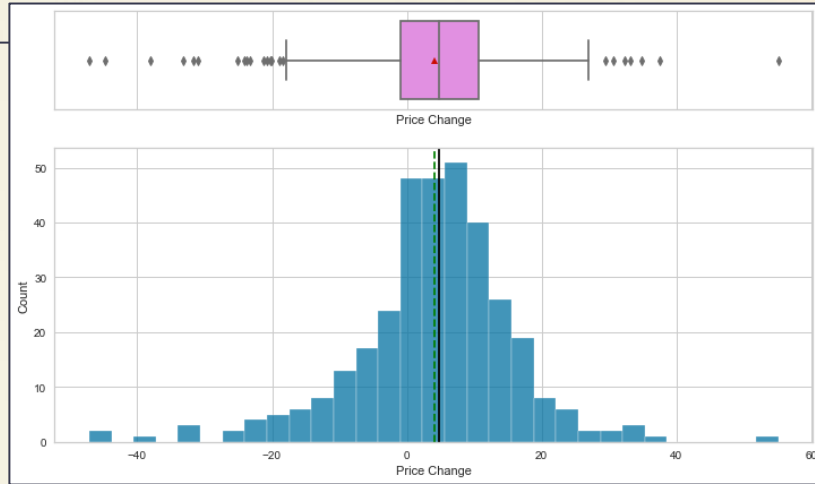
Current Price



- The current stock prices are heavily skewed to the right.
- There are outliers after the right whisker.
- The highest price is at \$1275.
- Most of the stock prices range between 0 to 200.
- The mean price is greater than the median.

Univariate Analysis

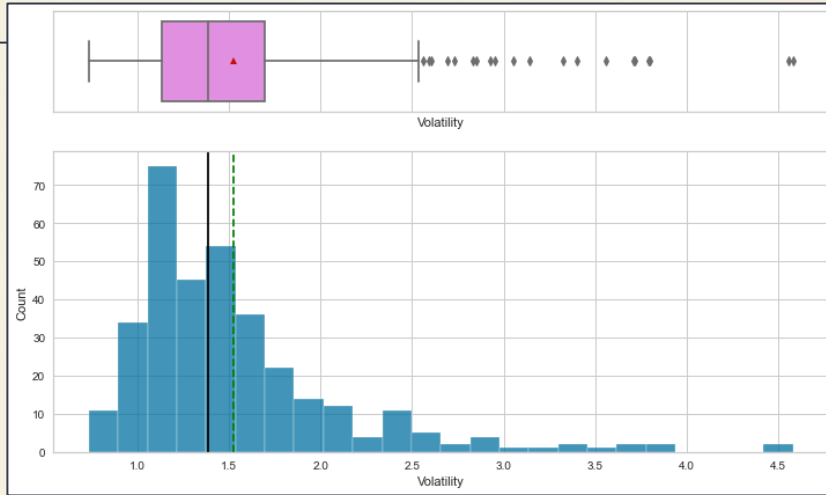
Price Change



- The distribution is normal.
- Most the price changes are an increase of 0% - 10%.
- There are outliers on both sides of the boxplot.
- The overall range is approximately between -20% to 40%.

Univariate Analysis

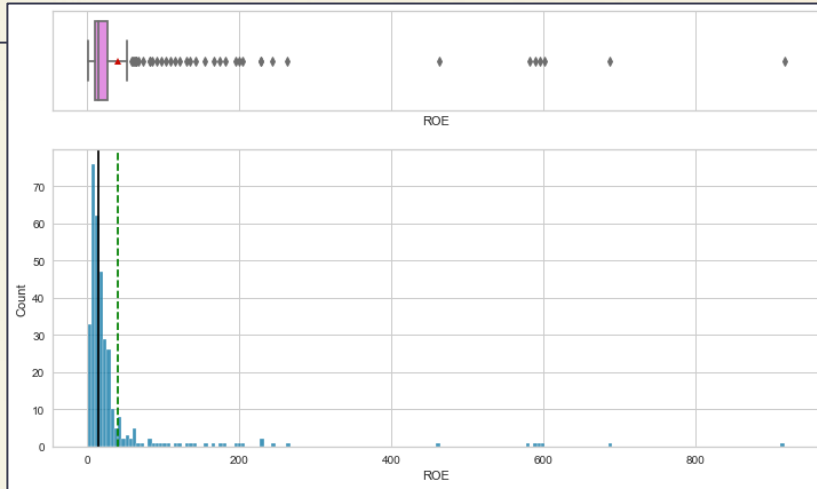
Volatility



- The distribution is right skewed.
- The mean standard deviation is at 1.5.
- Most of the stocks have a standard deviation of 1 to 1.5.
- There are outliers.
- The highest level of volatility is at a standard deviation of 4.5.

Univariate Analysis

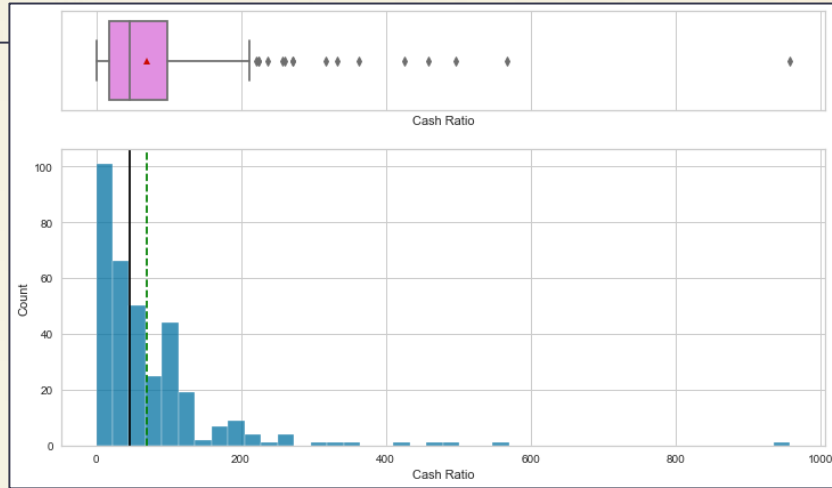
ROE



- The distribution is right skewed.
- The mean ROE is around 40%.
- There are extreme values of ROE such as 200%, 600%, and even around 900%.
- However, these are proper values since it indicates an extremely high level of performance of the company. This is due to extremely high net incomes for certain companies.

Univariate Analysis

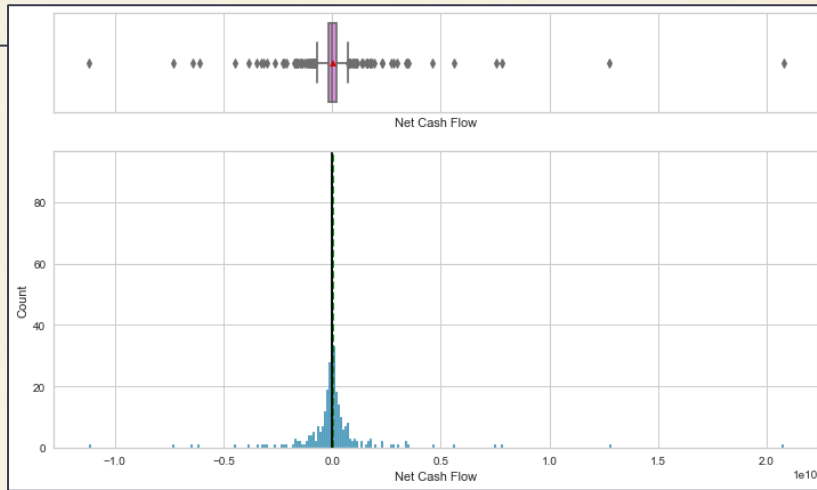
Cash ratio



- The distribution is right skewed.
- There are companies with a cash ratio of 0. These companies do not have sufficient cash to meet their liabilities.
- The mean cash ratio is at 70 and this is greater than the median.
- The highest cash ratio is at 958.
- 50% of the companies have a cash ratio of 47 or more.

Univariate Analysis

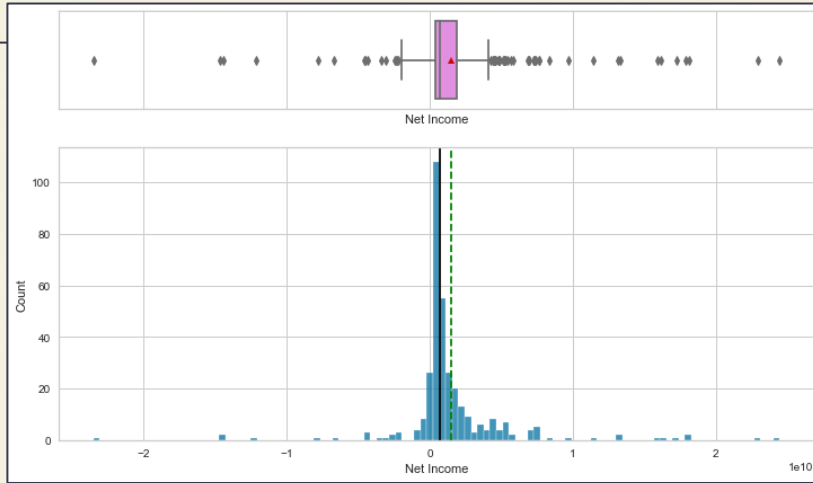
Net cashflow



- The distribution is normal.
- There are outliers.
- The mean and median values are coinciding.
- There are companies with negative cashflows which means the cash outflows exceed the cash inflows.

Univariate Analysis

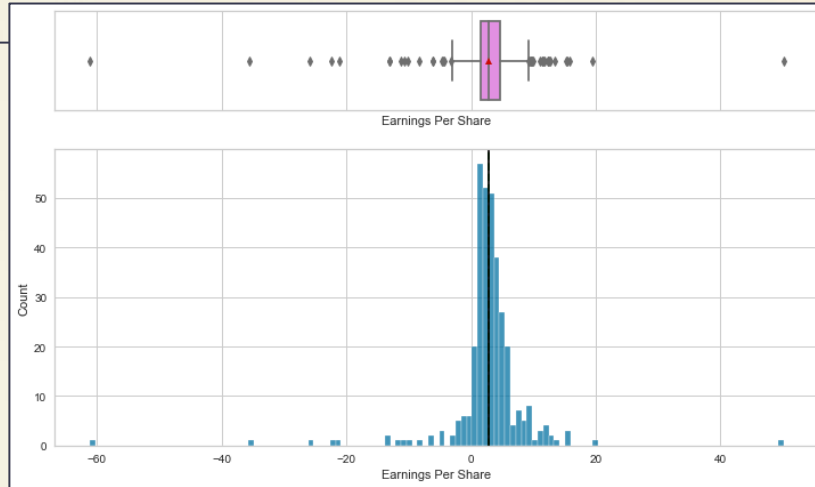
Net income



- The distribution is normal.
- There are outliers.
- The mean net income is greater than the median.
- There are companies with negative income and this is due to the expenses, interest, and tax exceeding the total revenue of the company.

Univariate Analysis

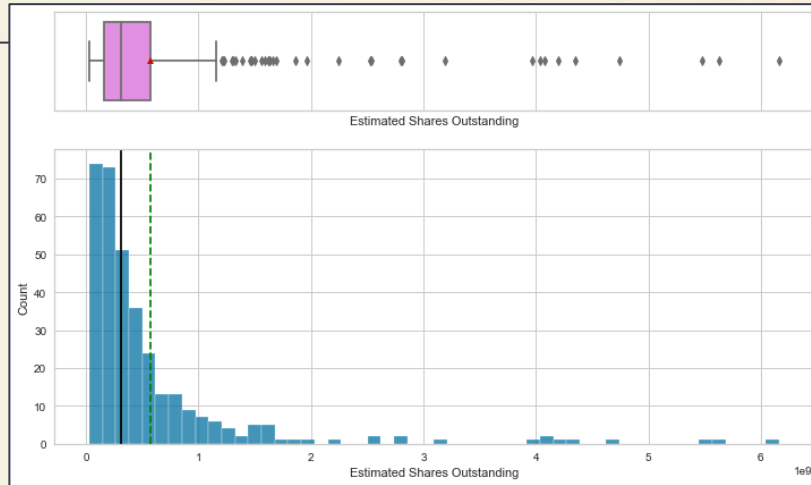
Earnings per share



- The distribution is normal.
- The mean and median EPS is the same.
- There are outliers.
- Some of the companies have a negative EPS. These are the companies which are having a negative income and hence do not have profits to distribute to their shareholders.

Univariate Analysis

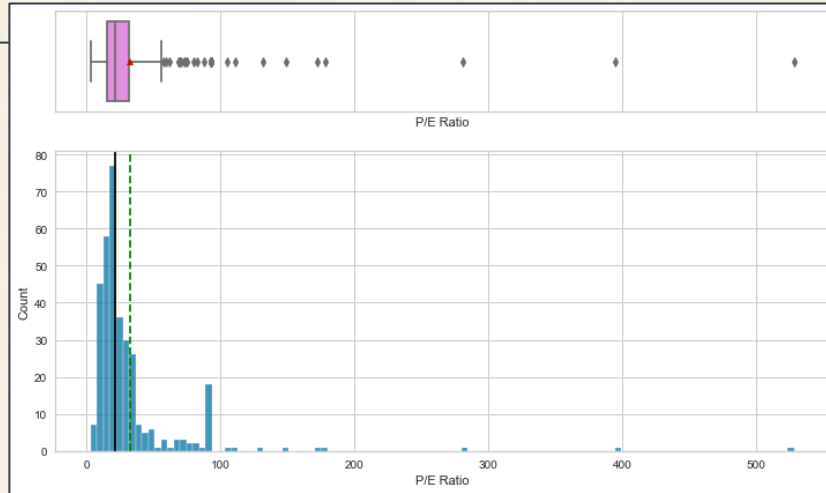
Estimated shares outstanding



- The distribution is right skewed.
- The mean shares outstanding is greater than the median.
- There are outliers.
- Companies with higher shares outstanding will have a lower price for their stocks.

Univariate Analysis

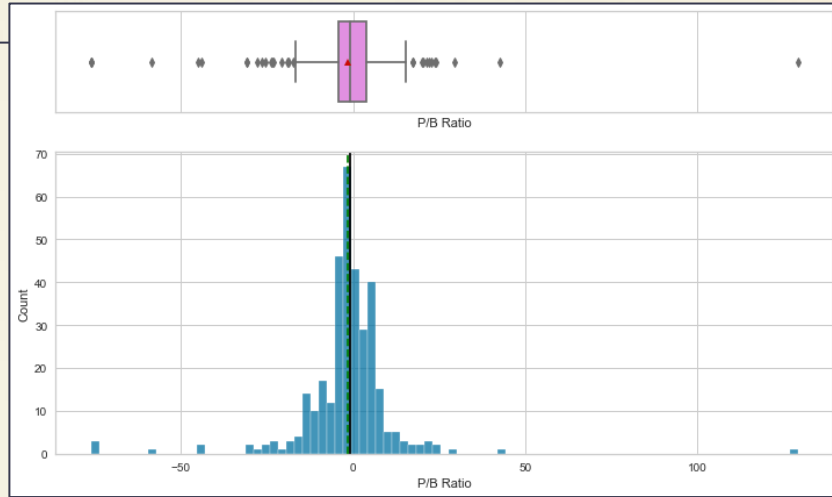
P/E Ratio



- The distribution is slightly right skewed.
- The mean P/E ratio is greater than the median.
- There are outliers.
- The P/E ratio ranges between 3 to 528.

Univariate Analysis

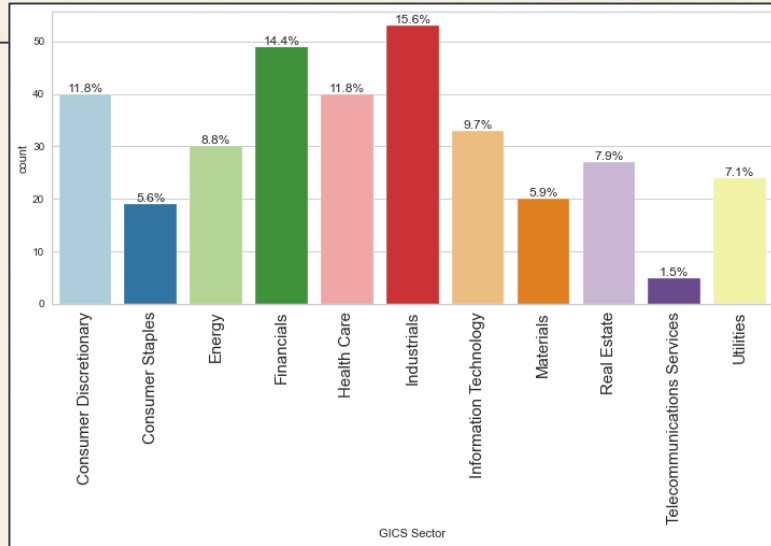
P/B Ratio



- The distribution is normal.
- The mean P/B ratio is slightly lower than the median.
- There are outliers.
- There are companies with a negative P/B ratio.
- The highest P/B ratio is at 129.

Univariate Analysis

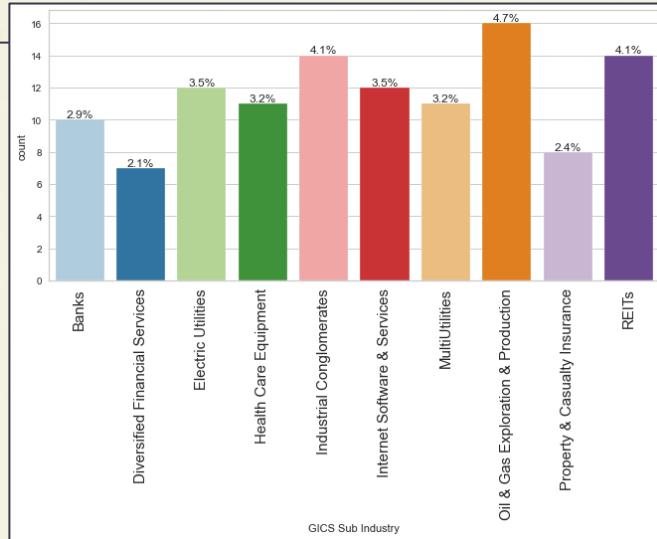
GICS Sector



- The barplot shows the top 10 GICS Sectors.
- Industrials and Financials are the top 2 sectors.
- Out of the top 10, Telecommunication services is the sector with lower number of companies.

Univariate Analysis

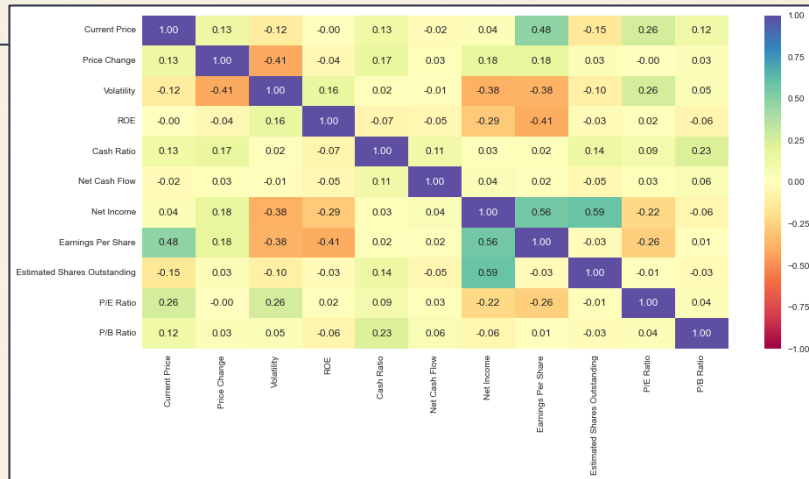
GICS Sub industry



- Oil & Gas exploration and production is the top sub industry, followed by Industrial conglomerates and REITs.
- Diversified financial services is the least common sub industry.

Bivariate Analysis

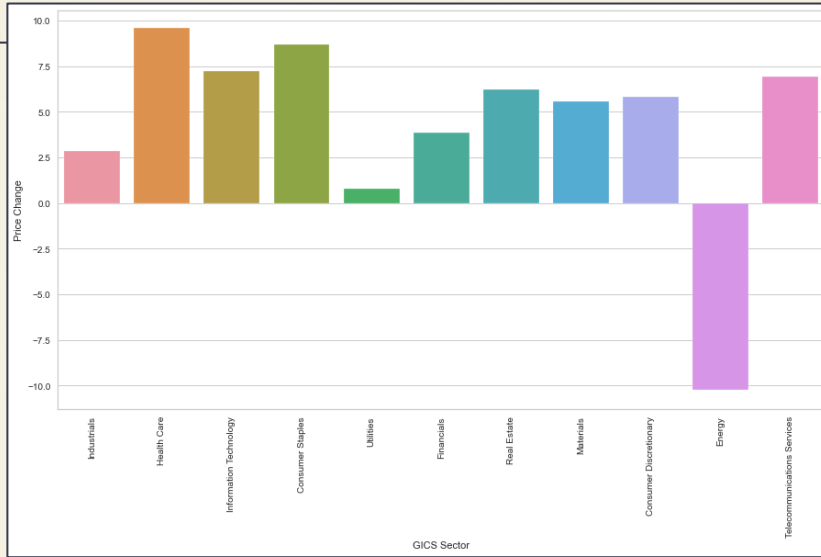
Correlation



- The highest positive correlation is between estimated shares outstanding and net income.
- EPS has a positive correlation with current price and net income, meaning an increase in current price or net income leads to an increase in EPS.
- The highest negative correlation is between Volatility and price change, and EPS and ROE.

Bivariate Analysis

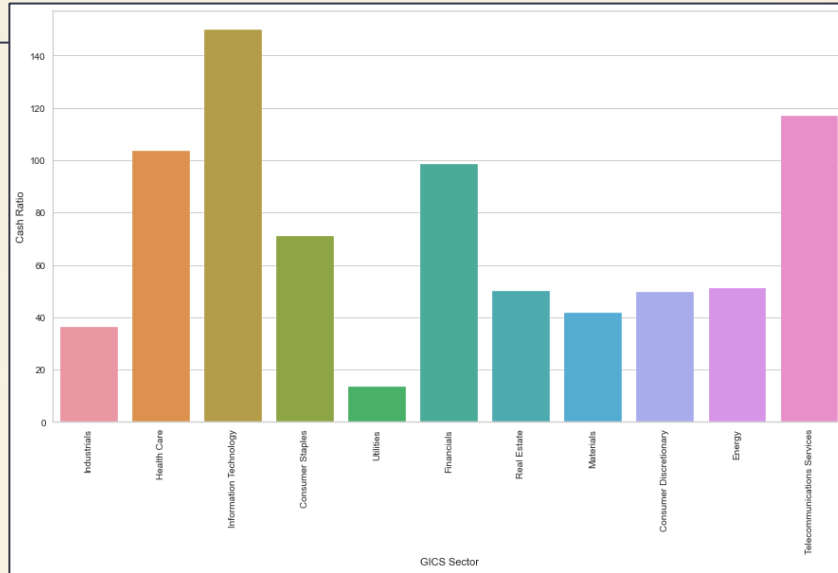
Price change over economic sectors



- The price change is highest for the Energy sector as it has a very high negative price change.
- The highest positive price change is for the health sector.
- The lowest price change is for the Utilities sector, meaning Utilities sector is comparatively more stable than the other sectors.

Bivariate Analysis

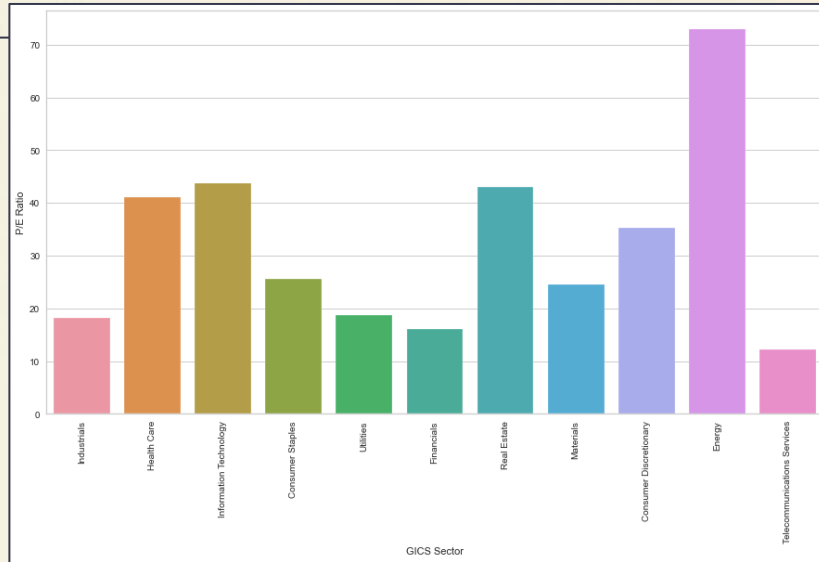
Cash ratio over economic sectors



- Cash ratio is highest for the IT sector and is followed by the Telecommunications services sector.
- Although Utilities sector was more stable than the rest of the sectors, it has the lowest cash ratio.

Bivariate Analysis

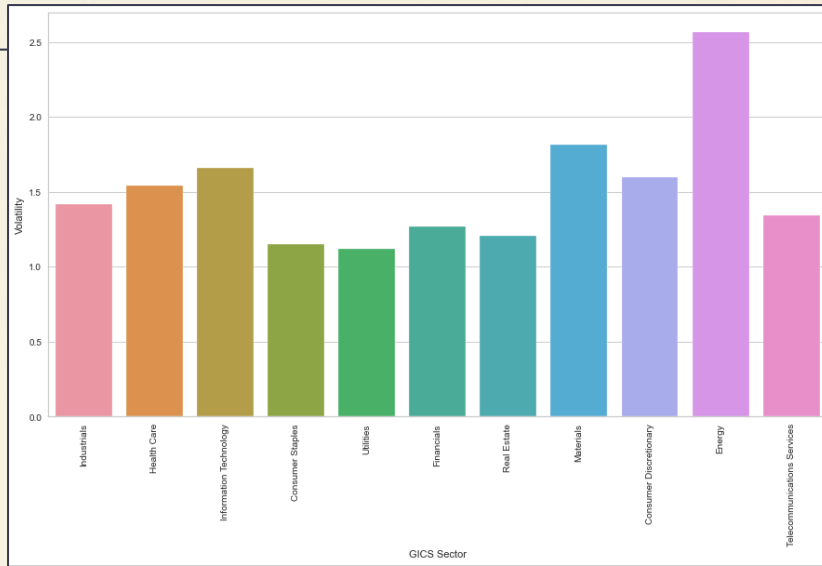
P/E ratio over economic sectors



- The P/E ratio is highest for the Energy sector, and this would suggest shareholders to invest more into the Energy sector.
- The lowest P/E ratio is for the Telecommunications Services sector and a lower P/E ratio would mean that the shareholders will be reluctant to invest here.

Bivariate Analysis

Volatility over economic sectors



- Though the Energy sector had a higher P/E ratio, it also has the highest volatility rate for its stock prices. Only shareholders who are willing to take higher risk, will invest in the Energy sector.
- The lowest volatility is for the Utilities sector which is in line with Utilities having the lowest price changes.

The background is a light cream color with a large, faint, light-yellow circular shape in the center. Scattered around are various geometric elements: a green circle at the top left, a red circle at the bottom right, a yellow circle at the bottom left, and several dark blue circles of different sizes. Some of these dark blue circles are positioned on the outlines of larger, thin-lined circles.

O4

Data Preprocessing

Missing values, Duplicates Check

No missing values

```
1 # checking for missing values in
2 df.isnull().sum() ## Complete t

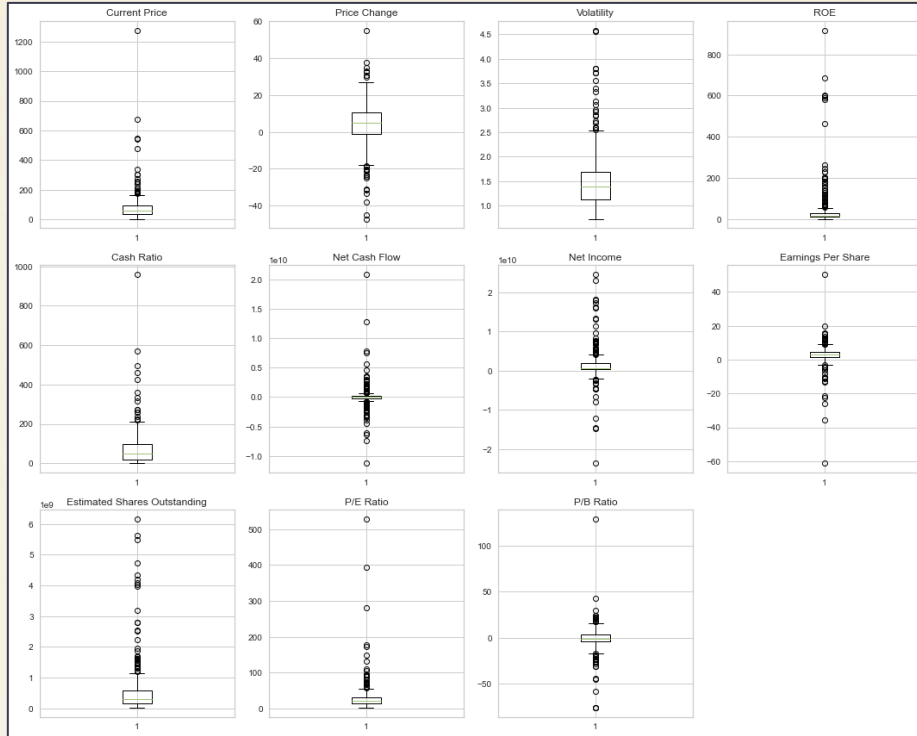
Ticker Symbol      0
Security           0
GICS Sector        0
GICS Sub Industry  0
Current Price      0
Price Change       0
Volatility          0
ROE                0
Cash Ratio         0
Net Cash Flow      0
Net Income         0
Earnings Per Share 0
Estimated Shares Outstanding 0
P/E Ratio          0
P/B Ratio          0
dtype: int64
```

No duplicates

```
1 # checking for duplicate
2 df.duplicated().sum()## C

0
```

Outlier Check





All the values are proper,
hence no need for outlier
treatment



Feature Scaling

```
1 # scaling the data before clustering
2 scaler = StandardScaler()
3 subset = df[numeric_columns].copy() ## Complete the code to scale the
4 subset_scaled = scaler.fit_transform(subset)
```

```
1 # creating a dataframe of the scaled data
2 subset_scaled_df = pd.DataFrame(subset_scaled, columns=subset.columns)
```

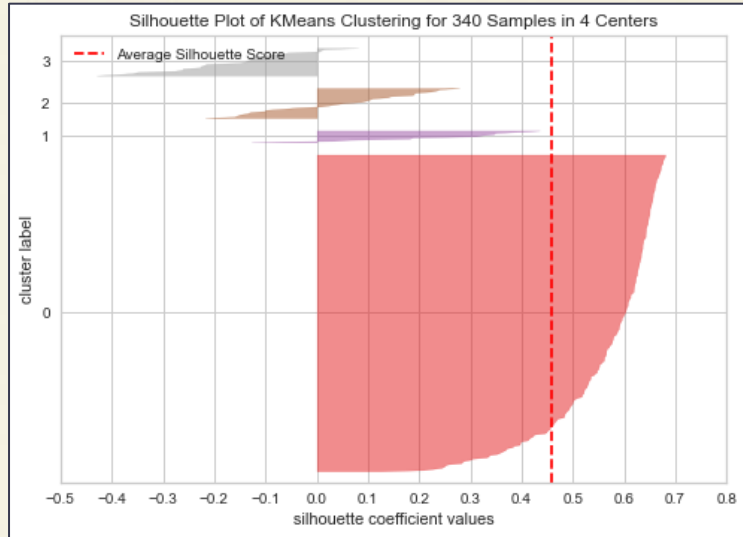
- The data is scaled to ensure that all the numerical variables have values in the same units.
 - `StandardScaler()` has been used to do this.
- 
- 

The background is a light cream color with a large, faint, light-yellow circular gradient in the center. Scattered around are various geometric elements: a green circle at the top center, a red circle at the bottom right, a yellow circle at the bottom left, and several dark blue circles and thin-lined circles of different sizes. The number '05' is prominently displayed in a dark blue serif font, enclosed within a white rounded rectangle with a thin dark blue border. A small dark blue circle is positioned to the right of this rectangle.

05

K-Means Clustering

Optimal number of clusters



- The data was split into 4 different clusters based on the elbow method and silhouette scores.
- **4 clusters** gave a silhouette score of 0.457.
- The highest silhouette score was for 3 clusters which was at 0.464. But 3 clusters seemed too less.
- Hence, 4 is appropriate and has a higher silhouette score as well.
- Majority of the data fall under Cluster 0 as seen from the given Silhouette plot.

Cluster profiling

KM_segments	0	1	2	3
Current Price	72.40	50.52	38.10	234.17
Price Change	5.07	5.75	-15.37	13.40
Volatility	1.39	1.13	2.91	1.73
ROE	34.62	31.09	107.07	25.60
Cash Ratio	53.00	75.91	50.04	277.64
Net Cash Flow	-14046223.83	-1072272727.27	-159428481.48	1554926560.00
Net Income	1482212389.89	14833090909.09	-3887457740.74	1572611680.00
Earnings Per Share	3.62	4.15	-9.47	6.05
Estimated Shares Outstanding	438533835.67	4298826628.73	480398572.85	578316318.95
P/E Ratio	23.84	14.80	90.62	74.96
P/B Ratio	-3.36	-4.55	1.34	14.40
count_in_each_segment	277	11	27	25

- Stocks belonging to cluster 3 have the highest mean current price, price change, cash ratio, net cashflow, EPS, and P/B ratio.
- Cluster 2 has the highest mean volatility, ROE, and P/E ratio.
- Cluster 1 has the highest mean net income and estimated shares outstanding.
- Approximately 81% of the companies fall under cluster 0.

Cluster Profiling

Cluster 0

Risk: Low

Returns: Low

- 82% of companies fall under cluster 0.
- Cluster 0 is dominated by companies belonging to Industrials and Financials sectors.
- Has the highest net income but ROE is very low.
- Has highest estimated shares outstanding which could possibly be the reason for low prices.
- Has a low price change and volatility suggesting lower risk.
- Cash ratio is low-moderate but only less than 50% of the companies have a positive net cash flow.
- Low-moderate level of EPS. All companies in this cluster have a positive EPS.
- P/E ratio and P/B ratios are low. Almost 50% of the companies have a P/B ratio.

Cluster Profiling

Cluster 1

Risk: Moderate

Returns: Low-Moderate

- 3% of companies fall under cluster 1.
- Cluster 1 has more companies belonging to Financials sectors.
- Has a low-moderate volatility which suggests why the ROE is low.
- Very poor net cashflow which suggests why the cash ratio is low.
- Has a moderate net income. Most of the companies (around 75%) have a positive net income.
- Has a low-moderate level of EPS. Less than 25% of companies have negative EPS, possibly due to few companies having negative net income.
- Has a low-moderate P/E ratio. More than 50% of companies have a negative P/B ratio.

Cluster Profiling

Cluster 2

Risk: High
Returns: High

- 8% of companies fall under cluster 2.
- Cluster 2 has more companies belonging to Energy sector.
- Highest volatility with very high price changes. But majority of the price changes are negative.
- However, gives the highest ROE.
- P/E ratio is the highest as well.
- Low price, net cash flow, and P/B ratio. Has lowest cash ratio as well.
- 75% of the companies have a negative net income and consequently negative EPS.

Cluster Profiling

Cluster 3

Risk: Low-Moderate

Returns: Moderate-High

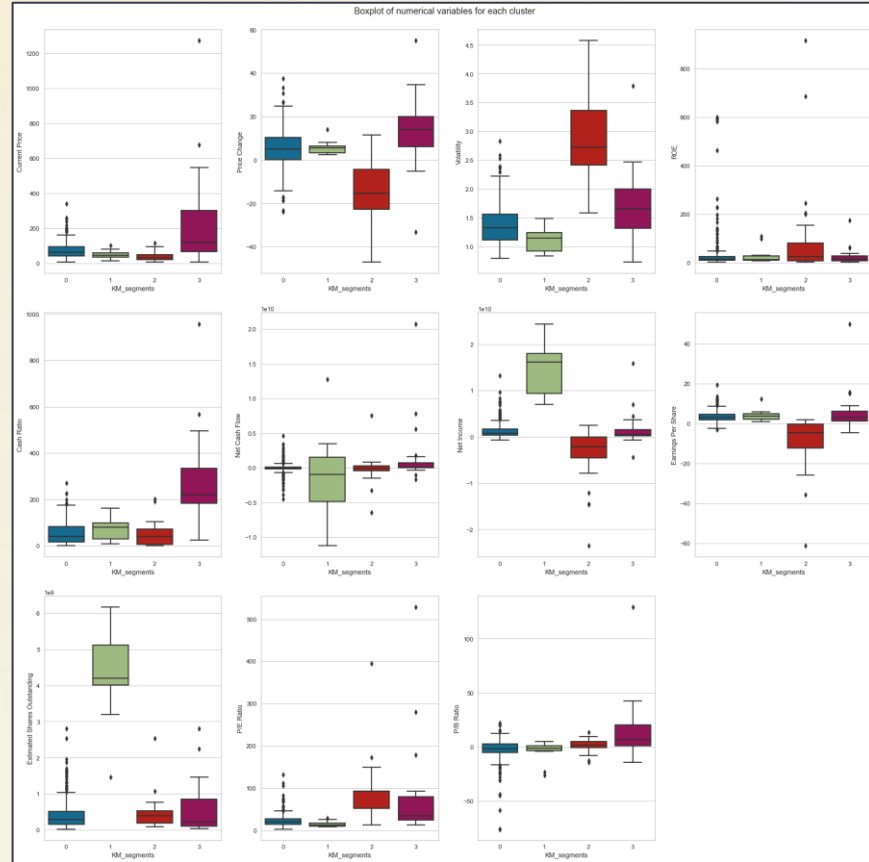
- 7% of companies fall under cluster 3.
- Cluster 3 has more companies belonging to Health Care sector.
- Has the highest prices, price change, cash ratio, net cashflow, EPS, and P/B ratio.
- Moderately volatile but has low ROE.
- 75% of companies have a positive net income.
- 25% of companies have a negative P/B ratio. Has a moderate level of P/E ratio as well.

Sectors under each cluster

- Cluster 0 majorly comprised of companies from the Industrials sector.
- Cluster 1 majorly comprised of companies from the Financials sector.
- Cluster 2 majorly comprised of companies from the Energy sector.
- Cluster 3 majorly comprised of companies from the Health Care sector.

KM_segments	GICS sectors	Security
0	Consumer Discretionary	33
	Consumer Staple.	17
	Energy	6
	Financials	45
	Health Care	29
	Industrials	52
	Information Technology	24
	Materials	19
	Real Estate	26
	Telecommunications Services	2
	Utilities	24
1	Consumer Discretionary	1
	Consumer Staples	1
	Energy	1
	Financials	3
	Health Care	2
	Information Technology	1
2	Telecommunications Services	2
	Energy	22
	Industrials	1
	Information Technology	3
3	Materials	1
	Consumer Discretionary	6
	Consumer Staples	1
	Energy	1
	Financials	1
	Health Care	9
	Information Technology	5
	Real Estate	1
	Telecommunications Services	1

Boxplot of numerical variables for each cluster

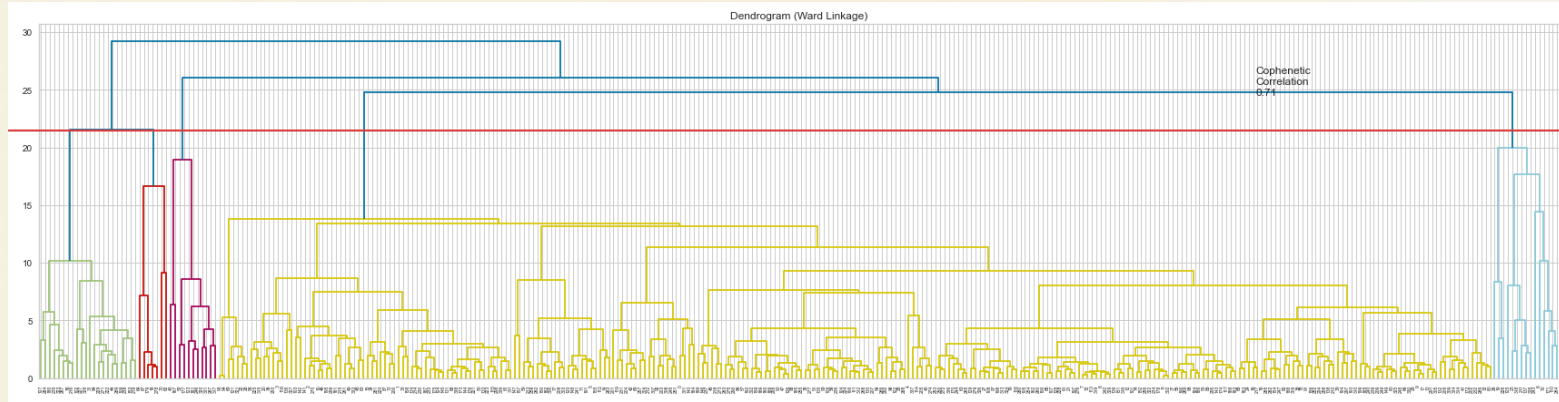


The background is a light cream color with a large, faint yellow circular gradient in the center. Scattered around are various geometric elements: a green circle at the top center, a red circle at the bottom right, a yellow circle at the bottom left, and several dark blue circles and outlines of circles. One dark blue circle is at the top left, another at the top right, and a third at the bottom center. There are also outlines of circles without dots: one at the top left, one at the top right, and one at the bottom right.

06

Hierarchical Clustering

Optimal number of clusters



- **5 clusters** were selected as the optimal number of clusters.
- This was based on Euclidean distance with ward linkage.
- Ward linkage gives a better clustering than the other linkages although average linkage gave the highest cophenetic correlation.

Cluster Profiling

HC_segments	0	1	2	3	4
Current price	326.20	72.43	46.67	84.36	36.44
price Change	10.56	5.26	5.17	3.85	-16.07
Volatility	1.64	1.43	1.08	1.83	2.83
ROE	14.40	25.51	25.00	633.57	57.50
Cash Ratio	309.47	60.88	58.33	33.57	42.41
Net Cash Flow	288850666.67	196157425.09	-3040666666.67	-568400000.00	-472834090.91
Net Income	864498533.33	1623022236.93	14848444444.44	-4968157142.86	-3161045227.27
Earnings Per Share	7.79	3.65	3.44	-10.84	-8.01
Estimated Shares Outstanding	544900261.30	462816085.05	4564959946.22	398169036.44	514367806.20
P/E Ratio	113.10	24.65	15.60	42.28	85.56
P/B Ratio	19.14	-2.62	-6.35	-11.59	0.84
count_in_each_segment	15	287	9	7	22

- Stocks belonging to cluster 0 have the highest mean current price, price change, cash ratio, net cashflow, EPS, P/E ratio, and P/B ratio.
- Cluster 2 has the highest mean net income and estimated shares outstanding.
- Cluster 3 has the highest mean ROE.
- Cluster 4 has highest mean volatility.
- Approximately 84% of the companies fall under cluster 1.

Cluster Profiling

Cluster 0

Risk: Low-Moderate

Returns: Low

- 4% of companies fall under cluster 0.
- Cluster 0 has more companies belonging to Health Care sector.
- Has the highest prices, price changes, cash ratio, net cash flow, EPS, P/E ratio and P/B ratio.
- Even though it has the highest price changes, it doesn't have the highest volatility.
- Has the lowest ROE.
- Has a very low net income but of the companies have a positive net income.
- 75% of companies have P/B ratio.

Cluster Profiling

Cluster 1

Risk: Low-Moderate

Returns: Low

- 84% of companies fall under cluster 1.
- Cluster 1 has more companies belonging to Industrials sector.
- Has low volatility and low ROE.
- 75% of companies have a positive price change.
- Less than 25% of companies have a negative net income and EPS.
- Has a very low net cash flow.
- More than 50% of companies have a negative P/B ratio.

Cluster Profiling

Cluster 2

Risk: Low

Returns: Low

- 3% of companies fall under cluster 2.
- Cluster 2 has more companies belonging to Financials sector.
- Has the highest net income but EPS is very low
- The current prices are really low as the shares outstanding are highest.
- Has the lowest volatility and low price changes.
- ROE is very low.
- P/E and P/B ratio are very low.
- Less than 25% of the companies have a positive net cash flow.

Cluster Profiling

Cluster 3

Risk: Low-Moderate

Returns: High

- 2% of companies fall under cluster 3.
- Cluster 3 has more companies belonging to Consumer staples and Energy sector.
- Has the highest ROE but does not have a very high volatility or price change.
- Low cash ratio and net cashflow.
- Around 75% of companies have a negative net income and EPS.
- Has a very low P/B ratio.

Cluster Profiling

Cluster 4

Risk: High

Returns: Low-Moderate

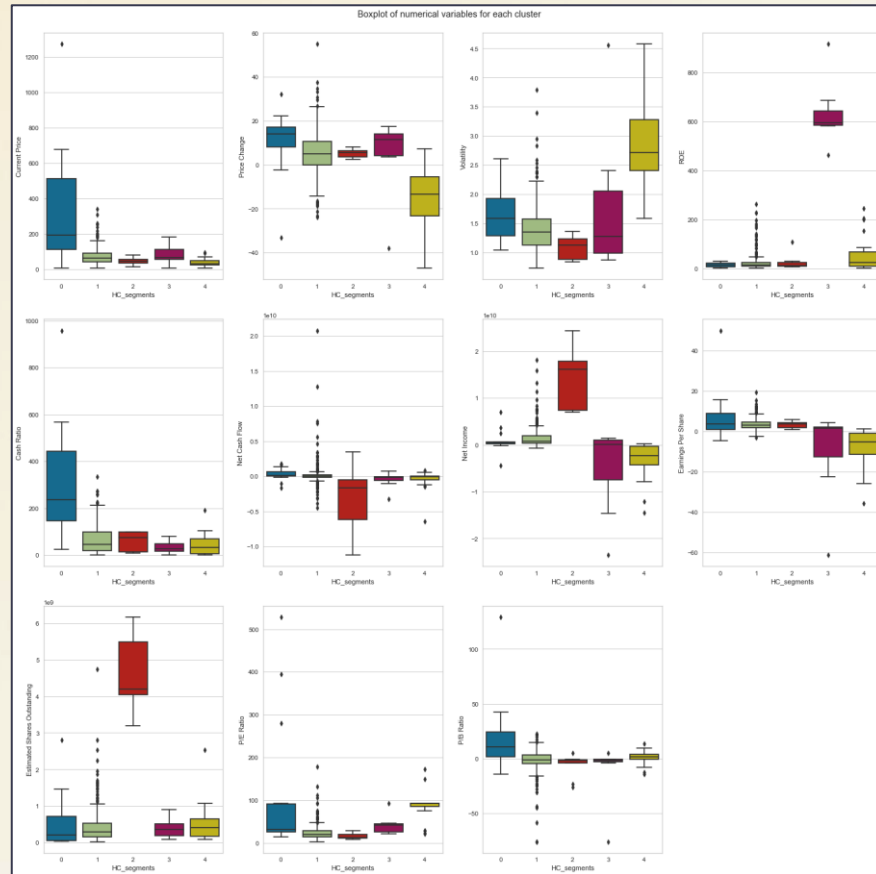
- 6% of companies fall under cluster 4.
- Cluster 4 has more companies belonging to Energy sector.
- Has the highest volatility a high negative price changes.
- Has decent ROE.
- Cash ratio and net cash income are low.
- 75% of the companies have negative net income and EPS.
- Low price but lower number of shares outstanding.
- Has a higher P/E ratio and low P/B ratio.

Sectors under each cluster

- Cluster 0 majorly comprised of companies from the Healthcare sector.
- Cluster 1 majorly comprised of companies from the Industrials sector.
- Cluster 2 majorly comprised of companies from the Financials sector.
- Cluster 3 majorly comprised of companies from the Consumer staples and Energy sectors.
- Cluster 4 majorly comprised of companies from the Energy sector.

HC_segments	GICS Sector	Security
0	Consumer Discretionary	3
	Consumer Staples	1
	Health Care	5
	Information Technology	4
	Real Estate	1
	Telecommunications Services	1
1	Consumer Discretionary	35
	Consumer Staples	15
	Energy	7
	Financials	45
	Health Care	34
	Industrials	52
	Information Technology	28
	Materials	19
	Real Estate	26
2	Telecommunications Services	2
	Utilities	24
	Consumer Discretionary	1
	Consumer Staples	1
	Energy	1
	Financials	3
3	Health Care	1
	Telecommunications Services	2
	Consumer Discretionary	1
	Consumer Staples	2
	Energy	2
	Financials	1
4	Industrials	1
	Energy	20
	Information Technology	1
	Materials	1

Boxplot of numerical variables for each cluster



The background is a light cream color with a large, faint, light-yellow circular shape in the center. Scattered around are various geometric elements: a green circle at the top center, a red circle at the bottom right, a yellow circle at the bottom left, and several dark blue circles of different sizes. Some of these dark blue circles are positioned on the outlines of larger, thin-lined circles.

07

Appendix

Data Background and contents

- **Ticker Symbol:** An abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market
- **Company:** Name of the company
- **GICS Sector:** The specific economic sector assigned to a company by the Global Industry Classification Standard (GICS) that best defines its business operations
- **GICS Sub Industry:** The specific sub-industry group assigned to a company by the Global Industry Classification Standard (GICS) that best defines its business operations
- **Current Price:** Current stock price in dollars
- **Price Change:** Percentage change in the stock price in 13 weeks
- **Volatility:** Standard deviation of the stock price over the past 13 weeks
- **ROE:** A measure of financial performance calculated by dividing net income by shareholders' equity (shareholders' equity is equal to a company's assets minus its debt)
- **Cash Ratio:** The ratio of a company's total reserves of cash and cash equivalents to its total current liabilities
- **Net Cash Flow:** The difference between a company's cash inflows and outflows (in dollars)
- **Net Income:** Revenues minus expenses, interest, and taxes (in dollars)
- **Earnings Per Share:** Company's net profit divided by the number of common shares it has outstanding (in dollars)
- **Estimated Shares Outstanding:** Company's stock is currently held by all its shareholders
- **P/E Ratio:** Ratio of the company's current stock price to the earnings per share
- **P/B Ratio:** Ratio of the company's stock price per share by its book value per share (book value of a company is the net difference between that company's total assets and total liabilities)

Data Background and contents

Statistical summary

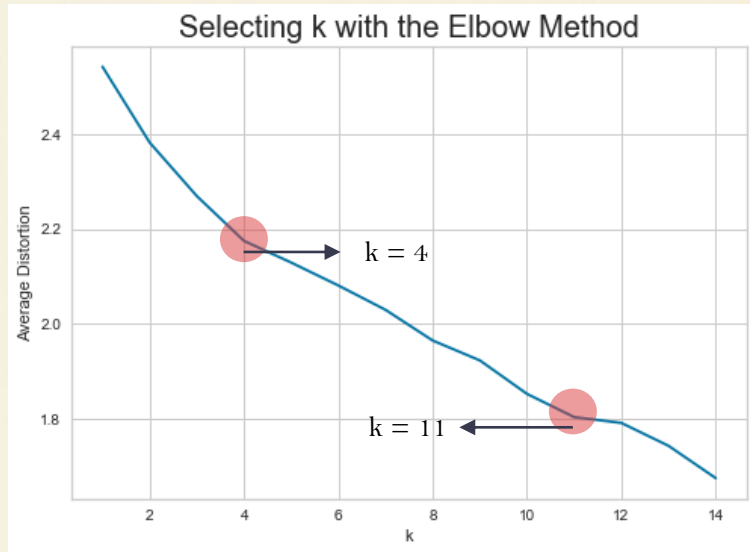
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Ticker Symbol	340	340	AAL	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Security	340	340	American Airlines Group	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GICS Sector	340	11	Industrials	53	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GICS Sub Industry	340	104	Oil & Gas Exploration & Production	16	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Current Price	340.0	NaN	NaN	NaN	80.862345	98.055086	4.5	38.555	59.705	92.880001	1274.949951
Price Change	340.0	NaN	NaN	NaN	4.078194	12.006338	-47.129693	-0.939484	4.819505	10.695493	55.051683
Volatility	340.0	NaN	NaN	NaN	1.525976	0.591798	0.733163	1.134878	1.385593	1.695549	4.580042
ROE	340.0	NaN	NaN	NaN	39.597059	96.547538	1.0	9.75	15.0	27.0	917.0
Cash Ratio	340.0	NaN	NaN	NaN	70.023529	90.421331	0.0	18.0	47.0	99.0	958.0
Net Cash Flow	340.0	NaN	NaN	NaN	55537620.588235	1946365312.175789	-11208000000.0	-193906500.0	2098000.0	169810750.0	20764000000.0
Net Income	340.0	NaN	NaN	NaN	1494384602.941176	3940150279.327937	-23528000000.0	352301250.0	707336000.0	1899000000.0	24442000000.0
Earnings Per Share	340.0	NaN	NaN	NaN	2.776662	6.587779	-61.2	1.5575	2.895	4.62	50.09
Estimated Shares Outstanding	340.0	NaN	NaN	NaN	577028337.754029	845849595.417695	27672156.86	158848216.1	309675137.8	573117457.325	6159292035.0
P/E Ratio	340.0	NaN	NaN	NaN	32.612563	44.348731	2.935451	15.044653	20.819876	31.764755	528.039074
P/B Ratio	340.0	NaN	NaN	NaN	-1.718249	13.966912	-76.119077	-4.352056	-1.06717	3.917066	129.064585

Rows: 340
Columns: 15

Integer: 4
Float: 7
Object: 4

K-Means Clustering Technique

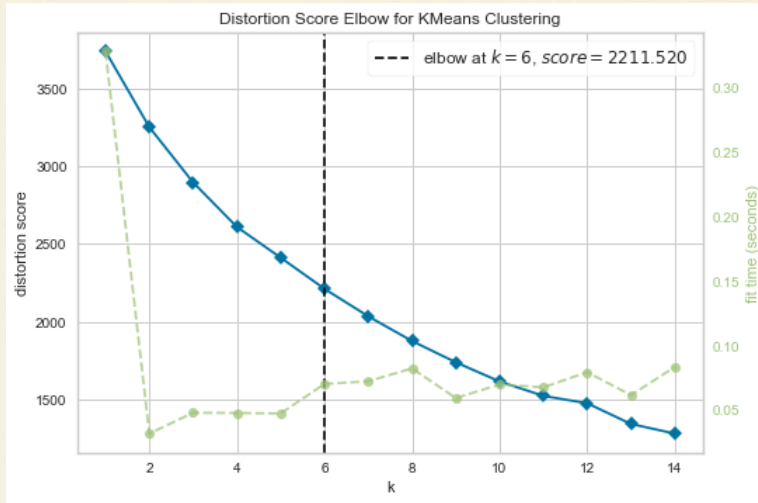
Elbow method



- There is a visible elbow at $k=4$ and $k=11$.
- 11 Clusters may be too much for the given business case.
- Hence, $k=4$ will be the number of clusters chosen as per elbow method.

K-Means Clustering Technique

Elbow method - Visualizer



- According to the obtained figure, elbow is identified at $k = 6$.
- So, now there are two options for selecting k , i.e., 4 and 6.
- Next, let's compare the silhouette scores of both.

K-Means Clustering Technique

Silhouette scores

- When comparing the silhouette scores, 4 clusters have a higher silhouette score than 6 clusters.
- Therefore, the optimal number of clusters will be 4.

Number of clusters	Silhouette Score
2	0.439696395
3	0.464440567
4	0.457722597
5	0.432283364
6	0.400542274
7	0.397633536
8	0.40278402
9	0.377858598
10	0.134589383
11	0.142183216
12	0.204466962
13	0.234248748
14	0.121025265

Hierarchical Clustering Technique

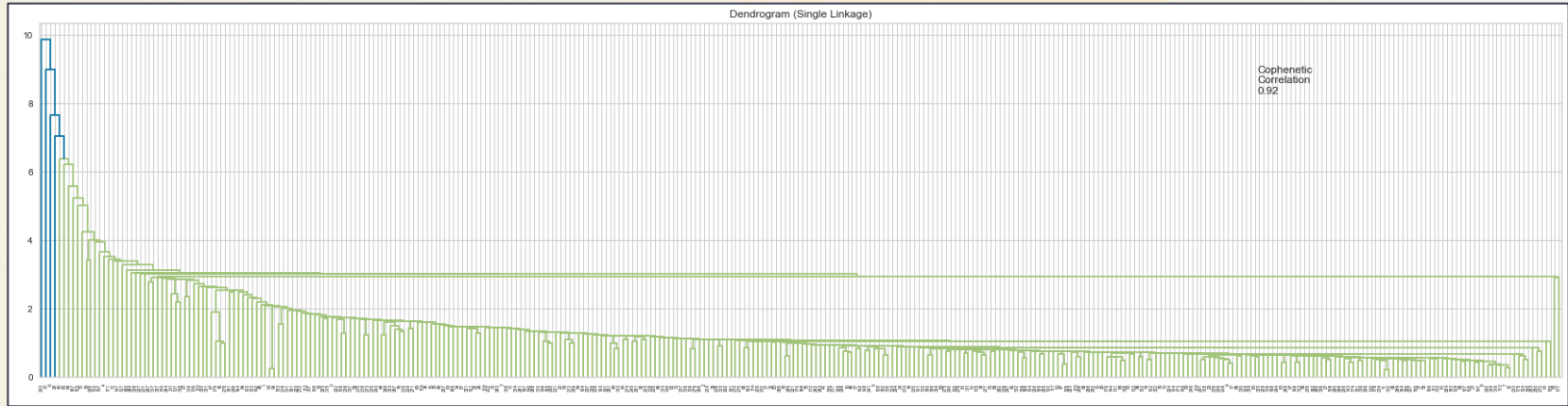
Cophenetic correlation

Distance	Linkage	Cophenetic correlation
Euclidean distance	single linkage	0.923990004
	complete linkage	0.825126969
	average linkage	0.943608456
	weighted linkage	0.871693048
Chebyshev distance	single linkage	0.909058001
	complete linkage	0.598605026
	average linkage	0.935348241
	weighted linkage	0.913481551
Mahalanobis distance	single linkage	0.908723158
	complete linkage	0.838711399
	average linkage	0.929841973
	weighted linkage	0.895591916
Cityblock distance	single linkage	0.926745358
	complete linkage	0.761734342
	average linkage	0.933500912
	weighted linkage	0.730850634

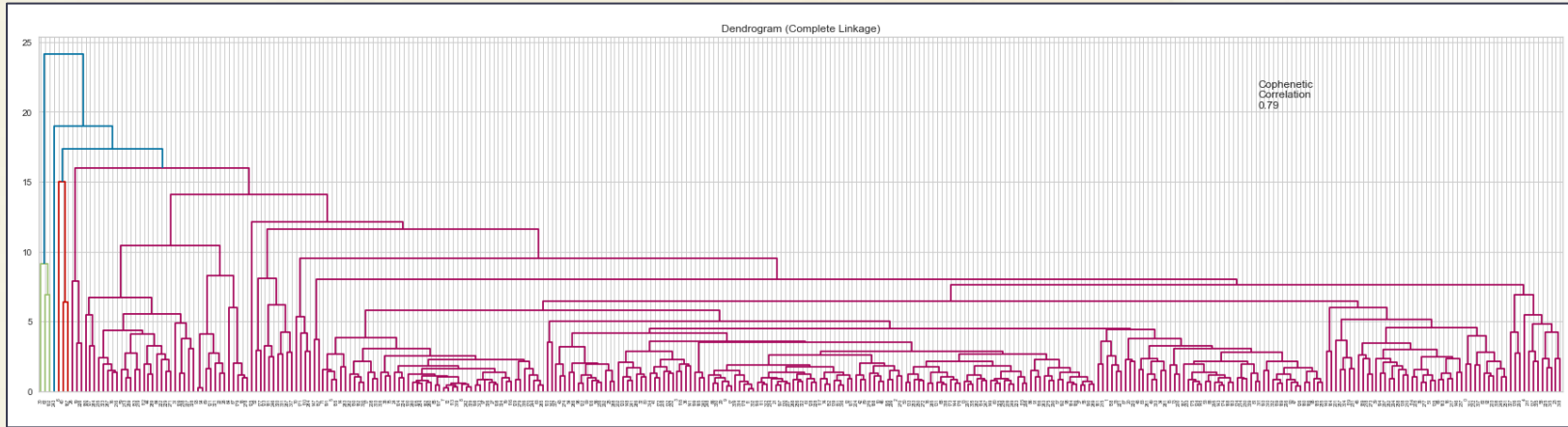
- Euclidean distance with average linkage gives the highest cophenetic correlation.

Distance	Linkage	Cophenetic correlation
Euclidean distance	single linkage	0.923990004
	complete linkage	0.825126969
	average linkage	0.943608456
	weighted linkage	0.871693048
	centroid linkage	0.9314012446828154
	ward linkage	0.7101180299865353

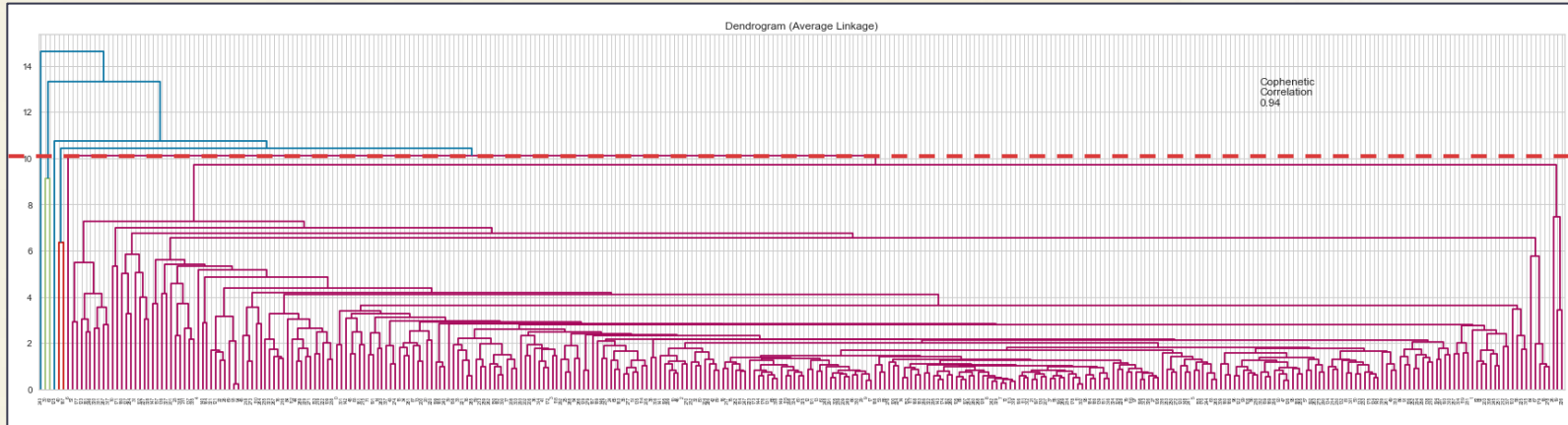
Dendrogram – Single linkage (Euclidean)



Dendrogram – Complete linkage (Euclidean)

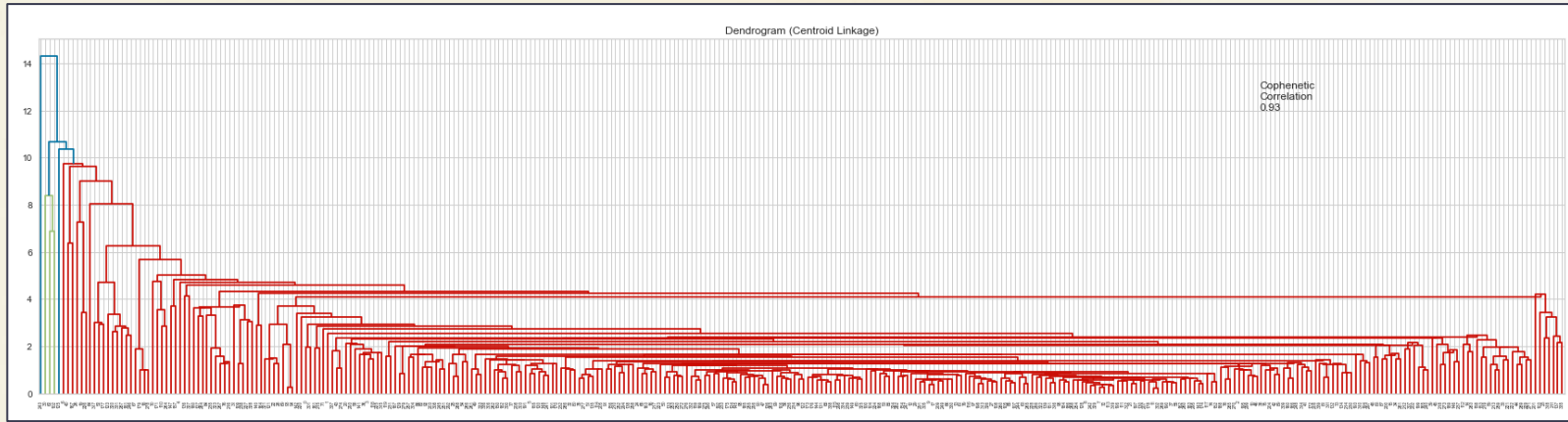


Dendrogram – Average linkage (Euclidean)



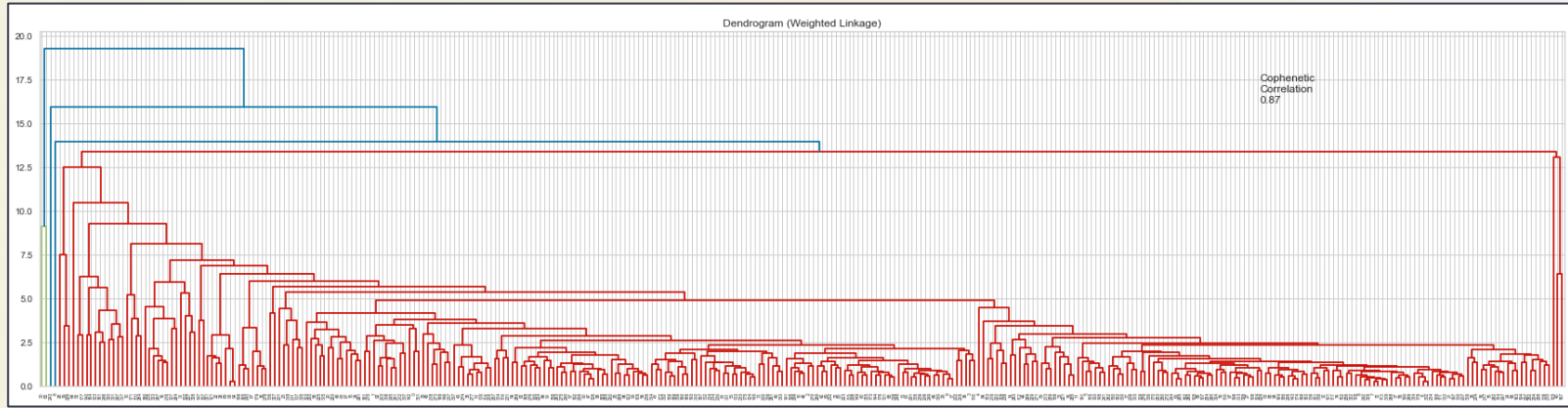
- Average linkage under Euclidean distance gives the highest cophenetic correlation.
- Based on the dendrogram 6 clusters seem appropriate. However, there is no clear distinction between the clusters.

Dendrogram – Centroid linkage (Euclidean)

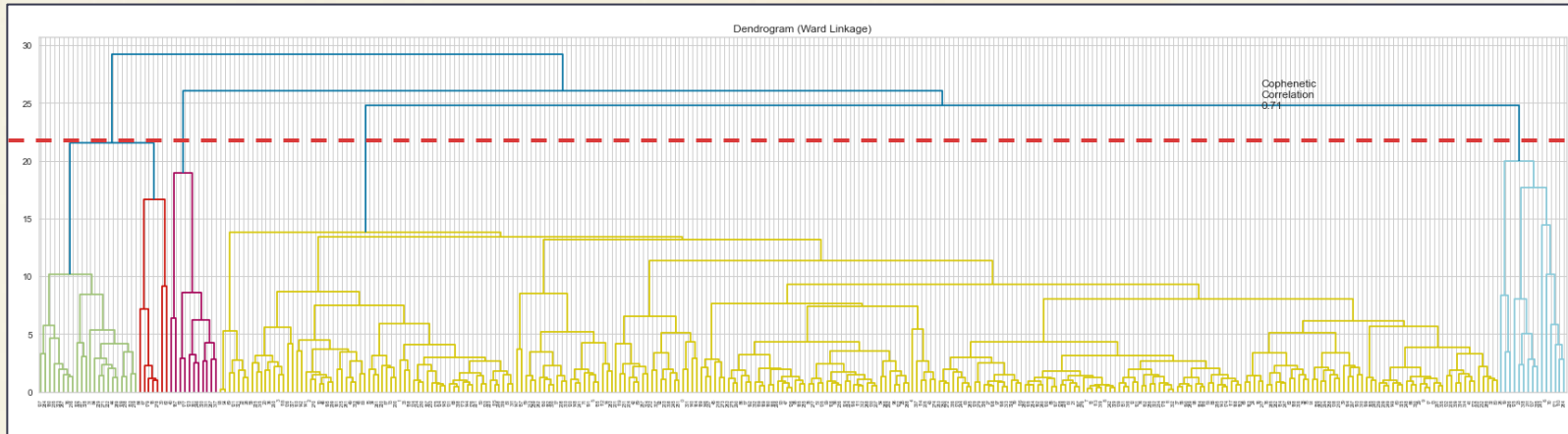


- Centroid linkage has a high cophenetic correlation. However, the dendrogram does not show clear distinction between the clusters.

Dendrogram – Weighted linkage (Euclidean)



Dendrogram – Ward linkage (Euclidean)



- Ward linkage has the lowest cophenetic correlation.
- However, the clusters have much clear distinction when compared to rest of the linkages.
- Hence, 5 clusters have been identified based on ward linkage.

K-means vs Hierarchical Clustering

- K-means had 4 clusters while hierarchical had 5 clusters.
- K-means clusters had a better split of the companies than hierarchical.
- Both had a cluster with majority of the companies in it (around 82%-84%). Both these clusters had majority of companies under Industrials sector.
- Both clusters didn't take much time to execute.
- However, more time was spent for hierarchical clustering on choosing the number of clusters as the initial dendrogram for average linkage was quite complex to interpret.
- The cluster profiles for clusters under k-means and hierarchical weren't very similar.

The background is a light cream color with a large, faint, yellowish circular gradient in the center. Scattered around the edges are several thin black circles and solid colored dots. In the top left, there is a single dark blue dot. In the top right, a thin black circle has a dark blue dot on its circumference. In the bottom left, a thin black arc contains three dots: red, yellow, and green. In the bottom right, there is a thin black circle without a dot.

Thank You!