Trade & Ahead Project

Unsupervised ML Problem

Yair Brama – December 2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- K-Means Clustering
- Hierarchical Clustering
- Appendix

Executive Summary - Business Context

As it well known, it is important to maintain a diversified portfolio when investing in stocks to maximize earnings under any market condition. Having a diversified portfolio tends to yield higher returns and face lower risk by tempering potential losses when the market is down. It is often easy to get lost in a sea of financial metrics to analyze while determining the worth of a stock.

By doing a cluster analysis, one can identify stocks that exhibit similar characteristics and stocks which exhibit minimum correlation. This will help investors better analyze stocks across different market segments and help protect against risks that could make the portfolio vulnerable to losses.

Trade & Ahead is a financial consultancy firm that provide their customers with personalized investment strategies. In this project we are provided with data comprising stock price and some financial indicators for a few companies listed under the New York Stock Exchange. The goal is to analyze the data, group the stocks based on the attributes provided, and share insights about the characteristics of each group.

Executive Summary – Final Recommendations

Based on the results of our models, we recommend to use the **K-means clustering** model. This model gives us **6 clusters** of companies, that we can profile as follows:

Cluster 0 *– 111 companies. These are companies that are generating profits on paper (earning per share is positive) but struggling to manage cash effectively (net cash flow is negative). It can be a red flag for liquidity issues, but it can also be an accounting planning that does not mean a long-term problem.*

Cluster 1 *– 153 companies. These companies are very stable in their share prices, with a correlation between their earnings and their shares prices.*

Cluster 2 *– 3 companies. These are the 3 leading companies in their current share price and in their P/E ratio, which shows that investors expect even more growth over time.*

Cluster 3 *- 34 companies. These companies have, by average, the best performances, although not the highest share price. That indicates stability and confidence by the market. This group includes leading companies in telecommunication (AT&T, Verizon), pharmaceutical (Pfizer), tech (Intel) and banking (JPMorgan Chase)*

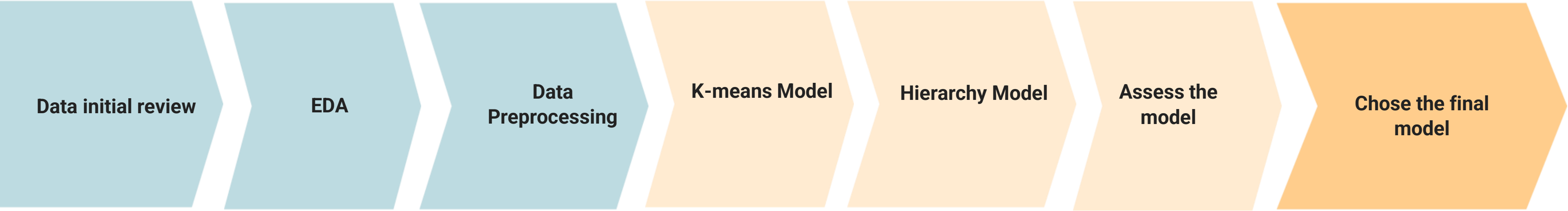
Executive Summary – Final Recommendations – Cont.

***Cluster 4** – 28 companies. These are companies that seem to struggle. Their price fell in the past year, and it was very volatile. They are losing cash and their earning per share is negative, which means the market is not confident about their long-term success. Note that there are 22 energy companies in this group (out of 28)*

***Cluster 5** – 11 companies. These companies show an anomaly that needs to be investigated further by investors. The ROE, on average, is very high, and the shares prices have gone up, but the cash flow is negative and the earning per share is negative. This could indicate one or more of the following:*

- *Temporary loses – Which impact the share price for the short term, but the ROE is still high*
- *High-debt – Can impact the shares but not the ROE at this point.*
- *Shares buy-back – The companies buy back their shares, which impact the cash and the earning, but the ROE is high, and the company is overall healthy.*

Solution Approach – Unsupervised Models



Data initial review

Check for null values

Check for duplicates

Check data types

EDA

Univariate Analysis – Box plots, Histograms

Data Preprocessing

Check outliers

Scale the numerical values

K-means Model

- Check the Elbow Plot
- Check Silhouette Scores
- Determine the number of clusters and fit the model

- Perform cluster profiling

Hierarchy Model

- Check the Cophenetic Correlation
- Generate Dendrogram

Fit the AgglomerativeClustering model

Perform cluster profiling

Assess the model

Adjust the no. of clusters and other parameters if needed

Perform cluster profiling for each iteration

Chose the final model

Decide on the best model with the best parameters (no. of clusters, metric, linkage)

Summarize the insights

Data Description

- **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

Data Description – Cont.

- **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 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)

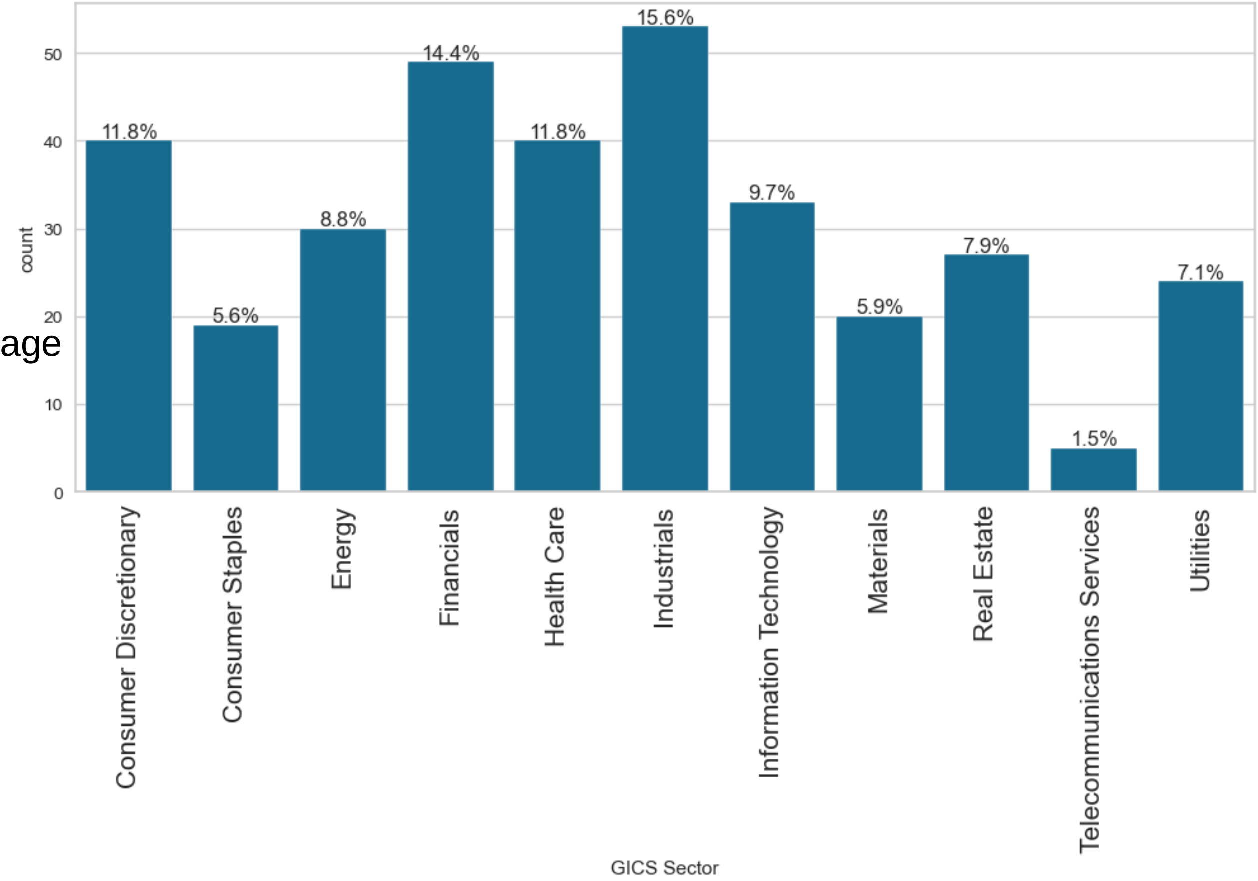
EDA Results - Data Overview

The data includes 340 rows and 15 columns. There are no missing values and no duplicates. The companies belong to 11 sectors, and 104 sub-sectors.

	mean	std	min	25%	50%	75%	max
Current Price	81	98	4	39	60	93	1,275
Price Change	4	12	-47	-1	5	11	55
Volatility	2	1	1	1	1	2	5
ROE	40	97	1	10	15	27	917
Cash Ratio	70	90	0	18	47	99	958
Net Cash Flow	55,537,621	1,946,365,312	-11,208,000,000	-193,906,500	2,098,000	169,810,750	20,764,000,000
Net Income	1,494,384,603	3,940,150,279	-23,528,000,000	352,301,250	707,336,000	1,899,000,000	24,442,000,000
Earnings Per Share	3	7	-61	2	3	5	50
Estimated Shares Outstanding	577,028,338	845,849,595	27,672,157	158,848,216	309,675,138	573,117,457	6,159,292,035
P/E Ratio	33	44	3	15	21	32	528
P/B Ratio	-2	14	-76	-4	-1	4	129

EDA Results - Univariate Analysis

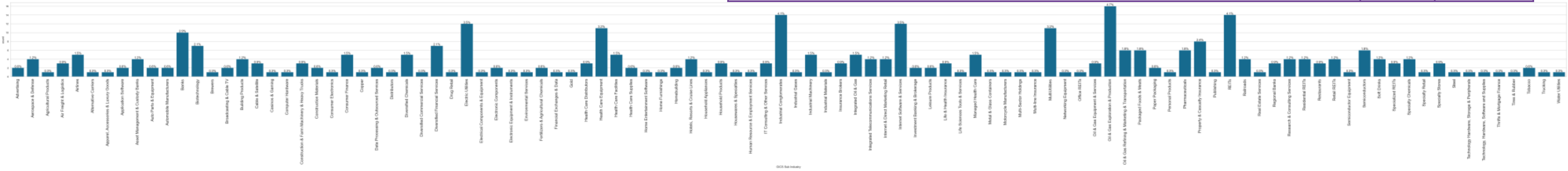
Observation – We have a balanced representation of companies from various sectors. The highest percentage is for industrial companies, and the lowest is for telco companies



EDA Results - Univariate Analysis

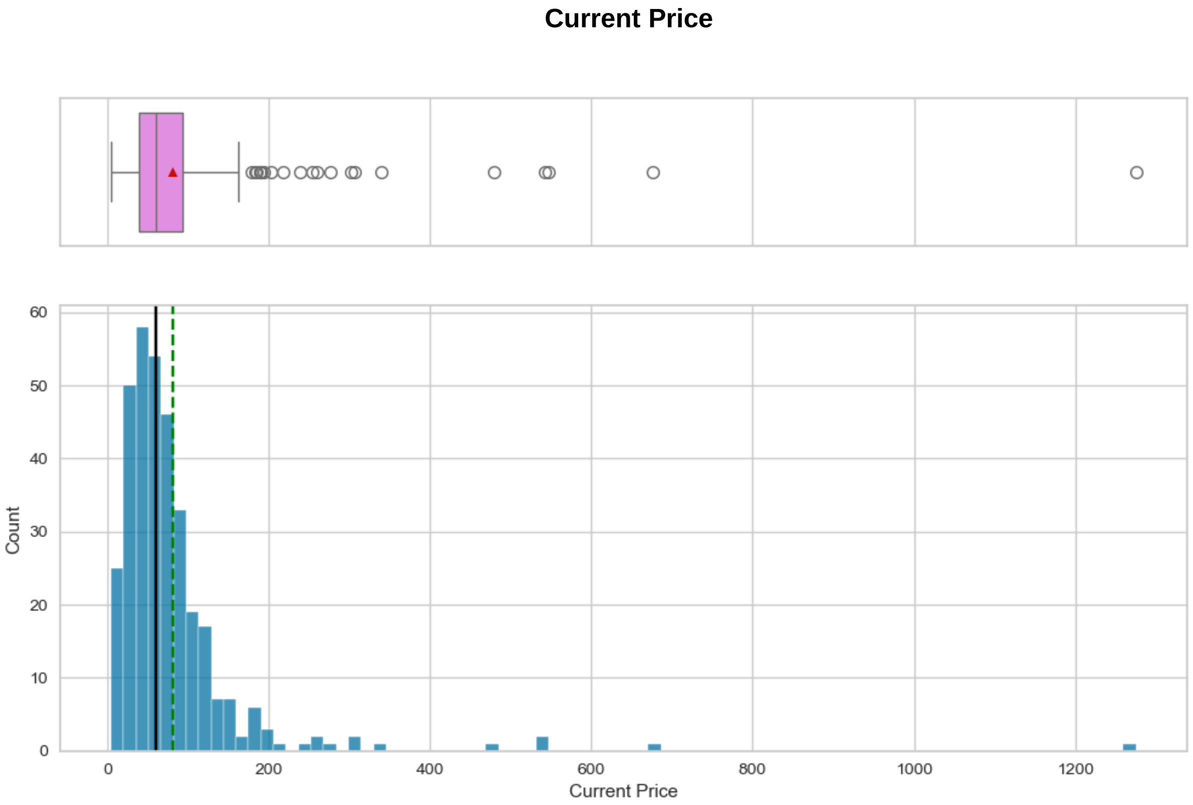
Observation – The sub-industries are very spread. 194 sub-industries have one company each, and the biggest sub-industry is oil and gas.

GICS Sub Industry	#	%
Oil & Gas Exploration & Production	16	5
REITs	14	4
Industrial Conglomerates	14	4
Electric Utilities	12	4
Internet Software & Services	12	4
Health Care Equipment	11	3
MultiUtilities	11	3
Banks	10	3
Property & Casualty Insurance	8	2
Diversified Financial Services	7	2
Biotechnology	7	2
Pharmaceuticals	6	2
Packaged Foods & Meats	6	2
Oil & Gas Refining & Marketing & Transportation	6	2
Semiconductors	6	2

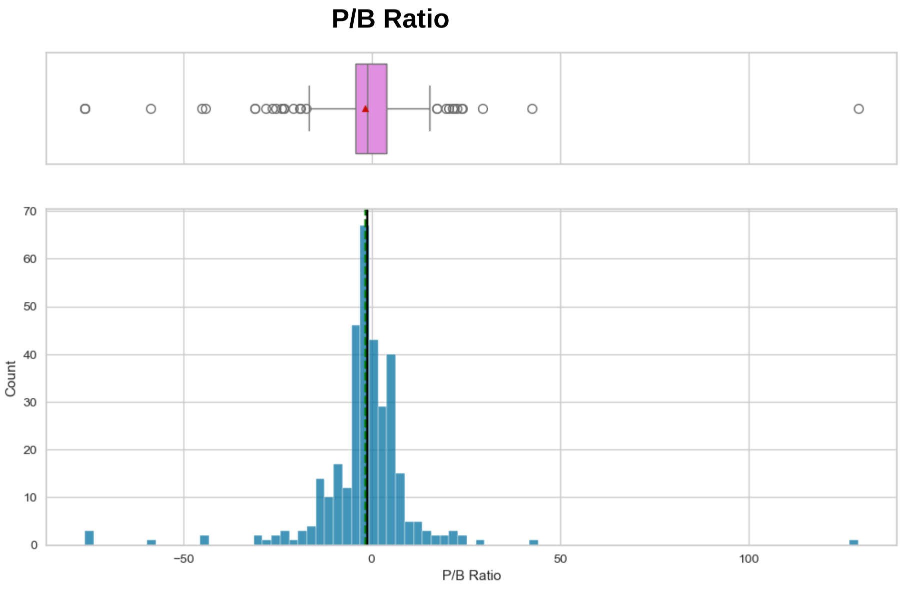


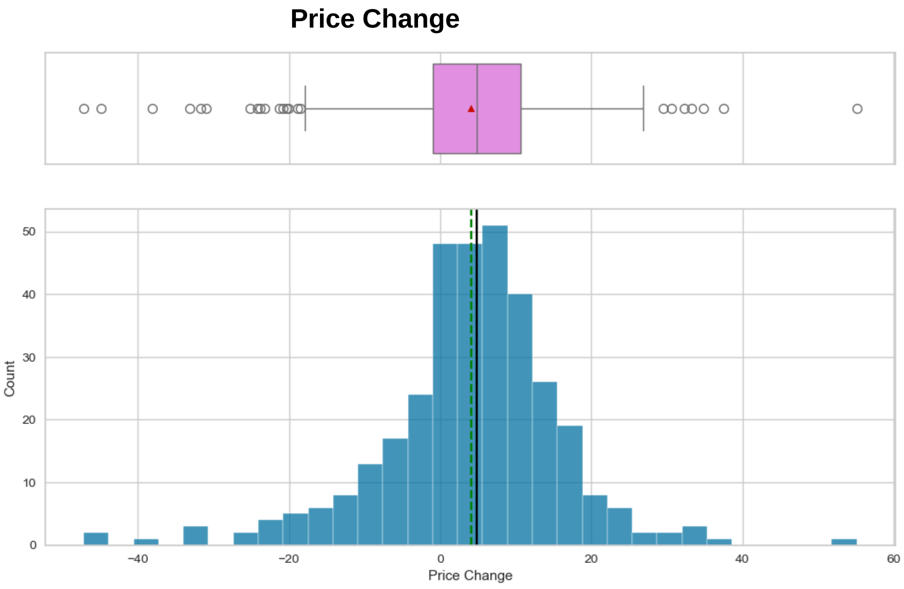
EDA Results - Univariate Analysis

Observation – The current share price is right skewed, with few companies that stand out and have a much higher price than the others



EDA Results - Univariate Analysis



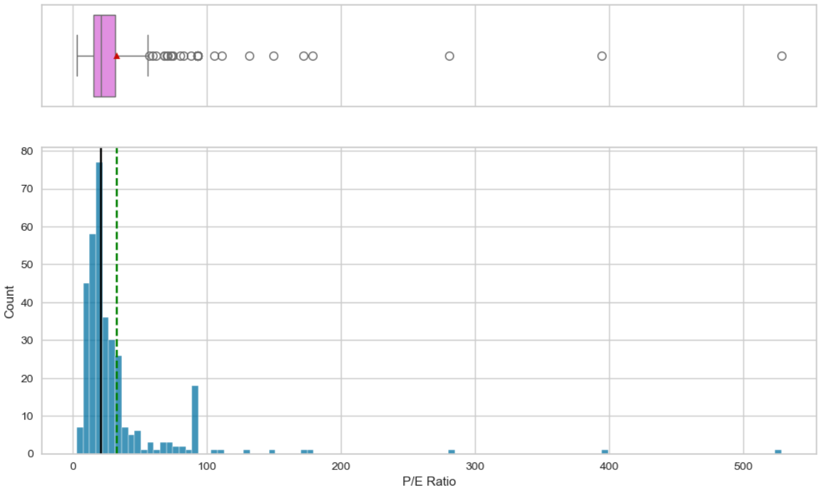


Observation – The P/B ratio and the price change are normally distributed, around $0

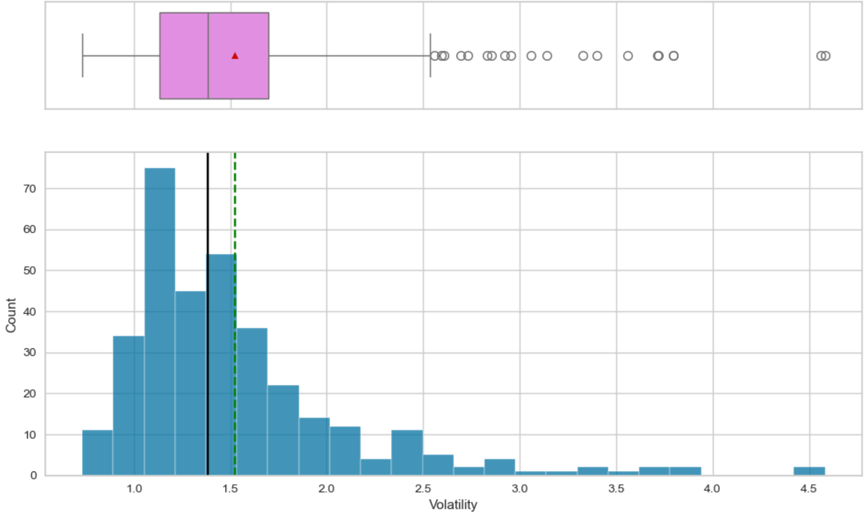
EDA Results - Univariate Analysis

Observation – The P/E ratio and the volatility are right skewed, with few outliers that have higher numbers

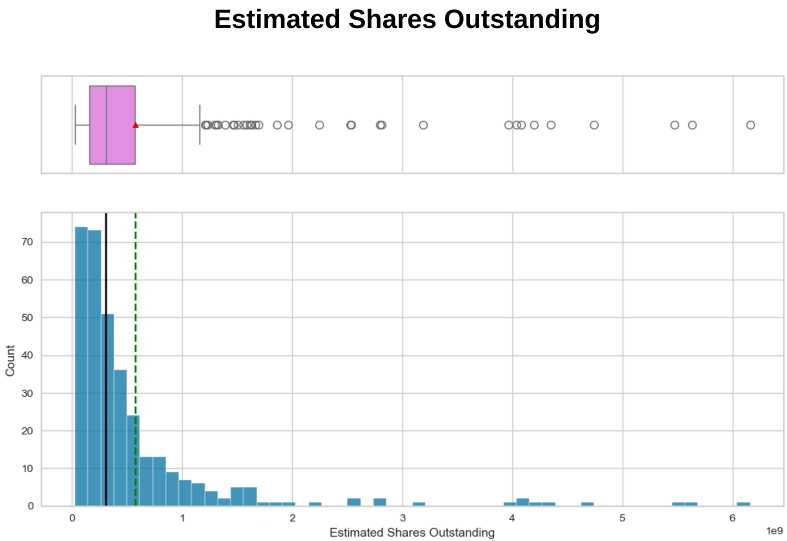
P/E Ratio

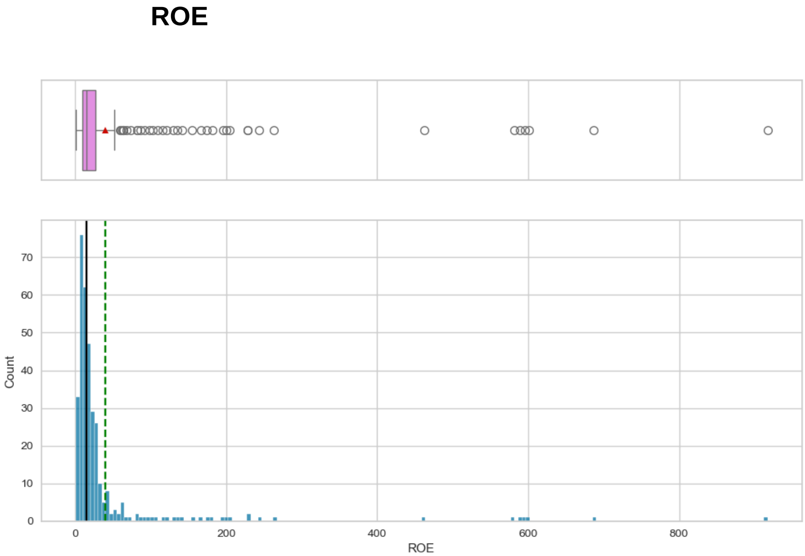


Volatility



EDA Results - Univariate Analysis

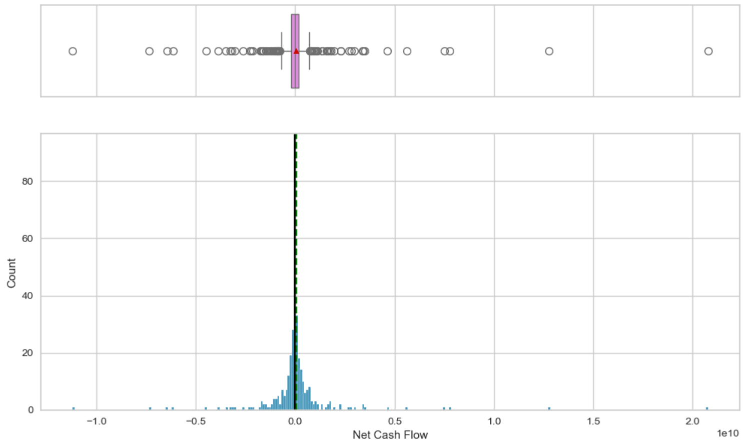




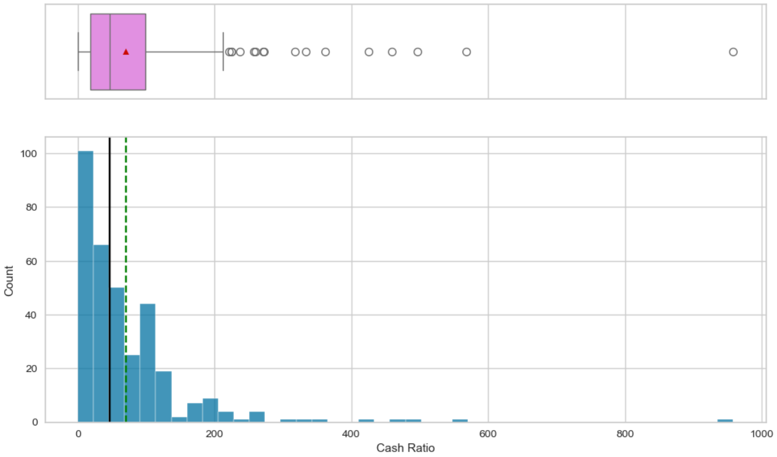
Observation – In accordance with the share prices, P/B ratio and volatility, the EPS and ROE are right skewed. This shows potential correlation between these attributes.

EDA Results - Univariate Analysis

Net Cash Flow



Cash Ratio

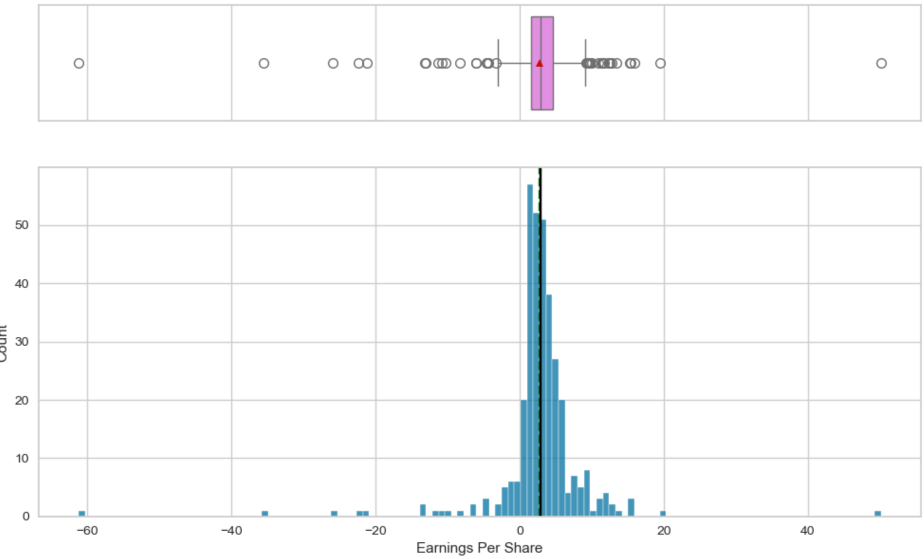


Observation – The net cash flow is normally distributed. The cash ratio that includes liabilities in addition to cash, is rightly skewed, like most of the other attributes.

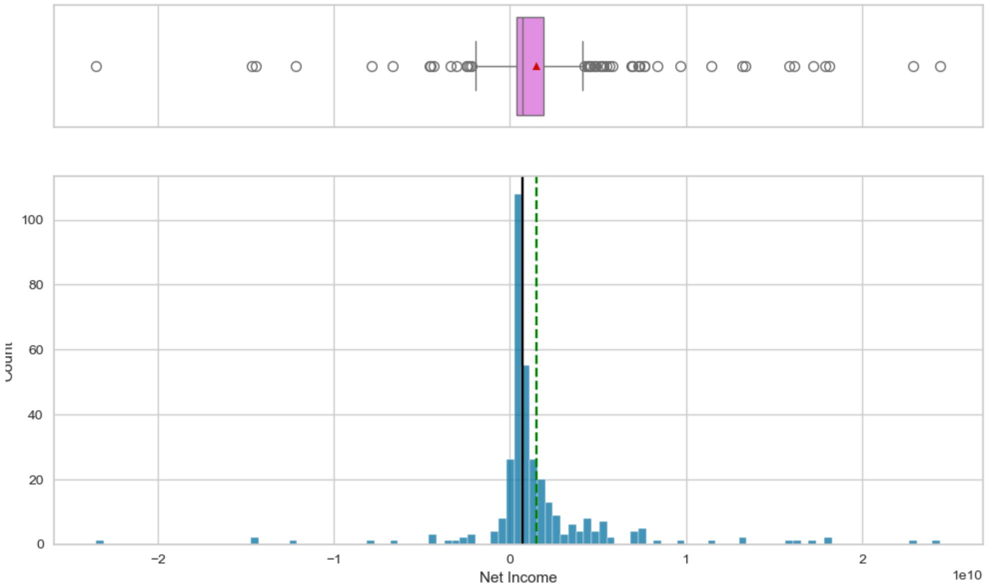
EDA Results - Univariate Analysis

Observation – The earning per share and the net income are normally distributed.

Earning Per Share



Net Income



Multivariate Analysis – Heatmap/Correlation

Observation –

Positive correlation –

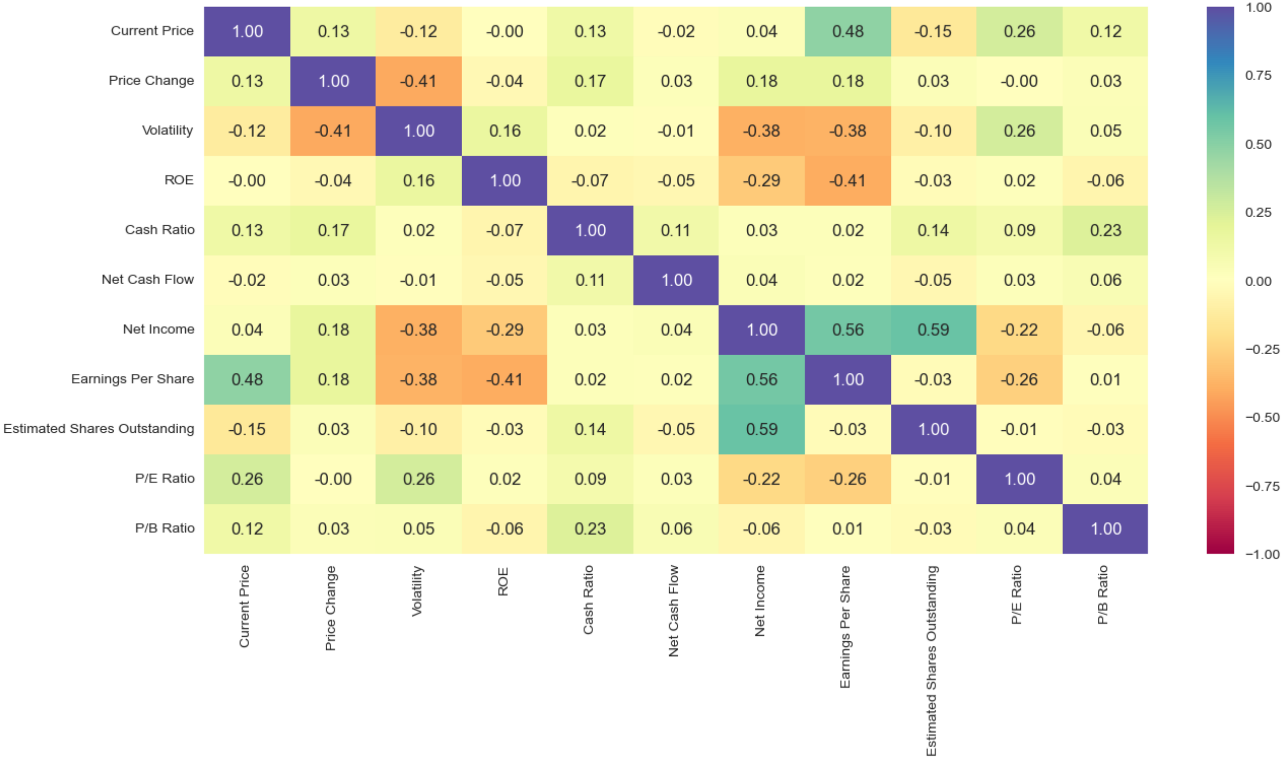
- EPS and current price
- EPS and net income
- EPS and estimated shares outstanding
- Net income and estimated shares outstanding

Negative correlation –

- Volatility and net income
- Volatility and EPS
- Volatility and price change

The price, income, earning per share and estimated shares outstanding seem to move in the same direction.

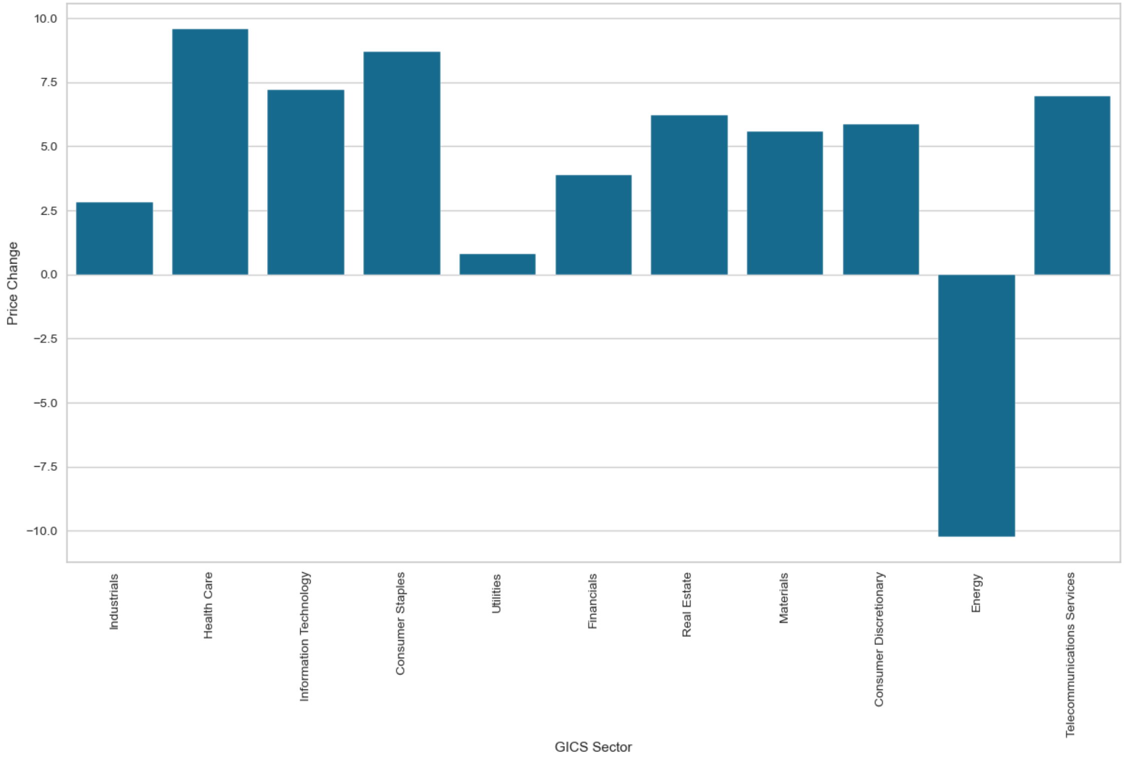
Volatility goes up as the prices and income go down



	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio
Current Price	1.00	0.13	-0.12	-0.00	0.13	-0.02	0.04	0.48	-0.15	0.26	0.12
Price Change	0.13	1.00	-0.41	-0.04	0.17	0.03	0.18	0.18	0.03	-0.00	0.03
Volatility	-0.12	-0.41	1.00	0.16	0.02	-0.01	-0.38	-0.38	-0.10	0.26	0.05
ROE	-0.00	-0.04	0.16	1.00	-0.07	-0.05	-0.29	-0.41	-0.03	0.02	-0.06
Cash Ratio	0.13	0.17	0.02	-0.07	1.00	0.11	0.03	0.02	0.14	0.09	0.23
Net Cash Flow	-0.02	0.03	-0.01	-0.05	0.11	1.00	0.04	0.02	-0.05	0.03	0.06
Net Income	0.04	0.18	-0.38	-0.29	0.03	0.04	1.00	0.56	0.59	-0.22	-0.06
Earnings Per Share	0.48	0.18	-0.38	-0.41	0.02	0.02	0.56	1.00	-0.03	-0.26	0.01
Estimated Shares Outstanding	-0.15	0.03	-0.10	-0.03	0.14	-0.05	0.59	-0.03	1.00	-0.01	-0.03
P/E Ratio	0.26	-0.00	0.26	0.02	0.09	0.03	-0.22	-0.26	-0.01	1.00	0.04
P/B Ratio	0.12	0.03	0.05	-0.06	0.23	0.06	-0.06	0.01	-0.03	0.04	1.00

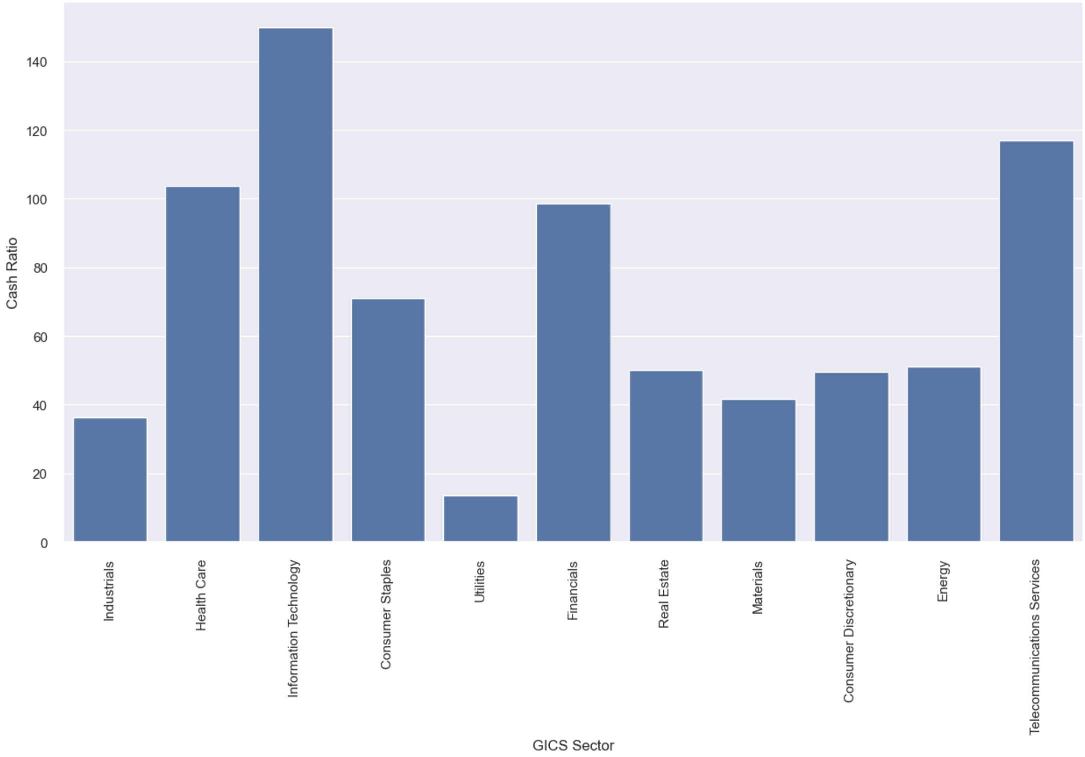
Multivariate Analysis – Price Change per Sector

Observation – The **Energy** sector stands out compares to the other sector, when it comes to the price changes. All the other sectors went up. Note that **utilities** didn’t change much



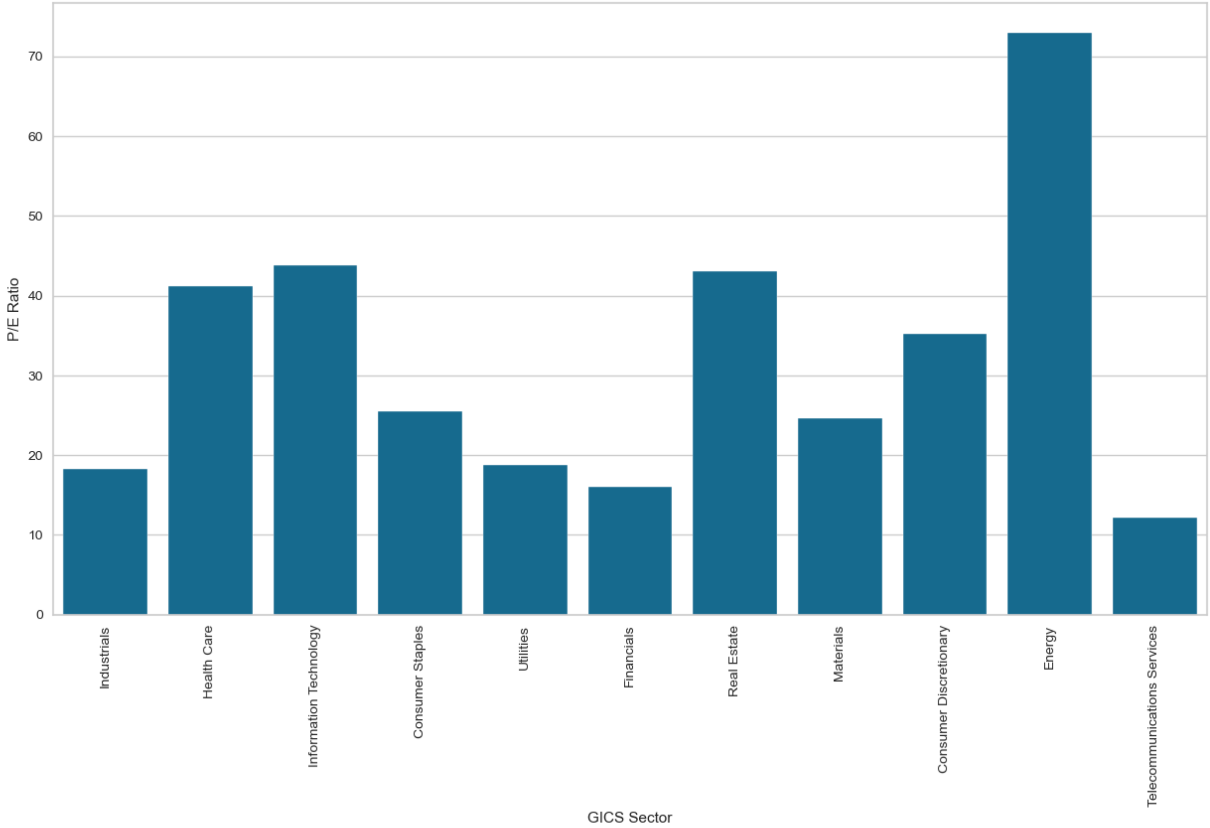
Multivariate Analysis – Cash Ratio per Sector

Observation – Cash ratio provides a measure of a company's ability to cover its short-term obligations using only cash and cash equivalents. **Utilities** companies have the lowest ratio, while **IT**, **Telco** and **finance** companies lead this measurement compared to others.



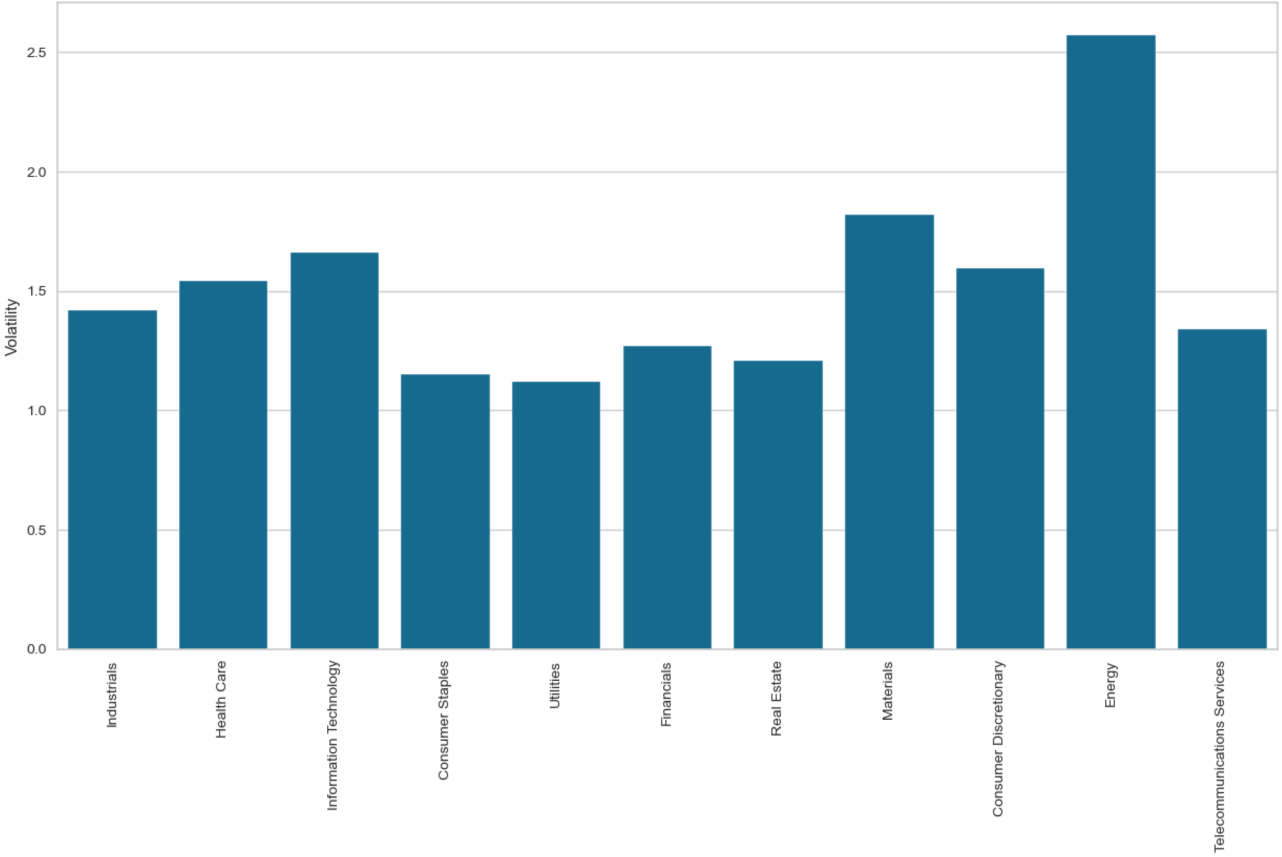
Multivariate Analysis – P/E Ratio per Sector

Observation – P/E ratios can help determine the relative value of a company's shares as they signify the amount of money an investor is willing to invest in a single share of a company per dollar of its earnings. **Energy** companies have the highest ratio, as they are in an industry that rely heavily on equipment and capital and built around long-term investments



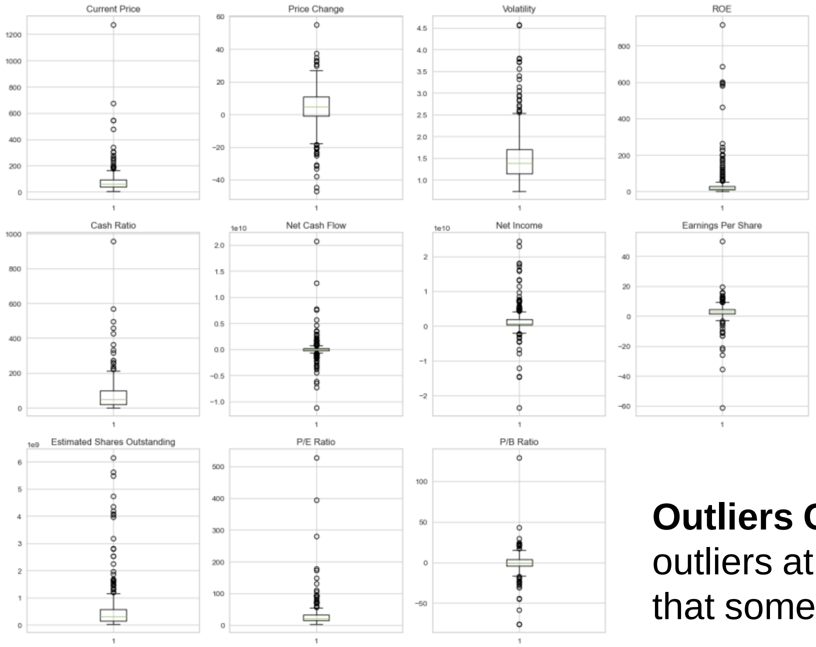
Multivariate Analysis – Volatility per Sector

Observation – Energy companies show the highest volatility, maybe due to their sensitivity to geo-political events in the past few years



Data Pre-processing

Scaling – Since this is unsupervised ML, we must normalize all the numeric values before running the model

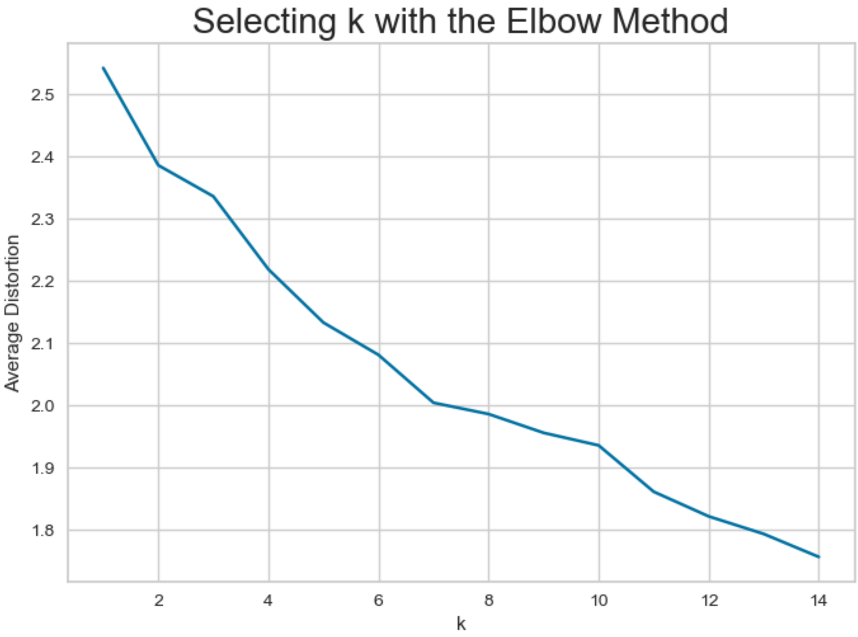


Outliers Check – We will not remove or normalize the outliers at this point, since the nature of stocks data is such that some stocks perform very different from others

K-mean Clustering

K-means Clustering– Checking Elbow Plot

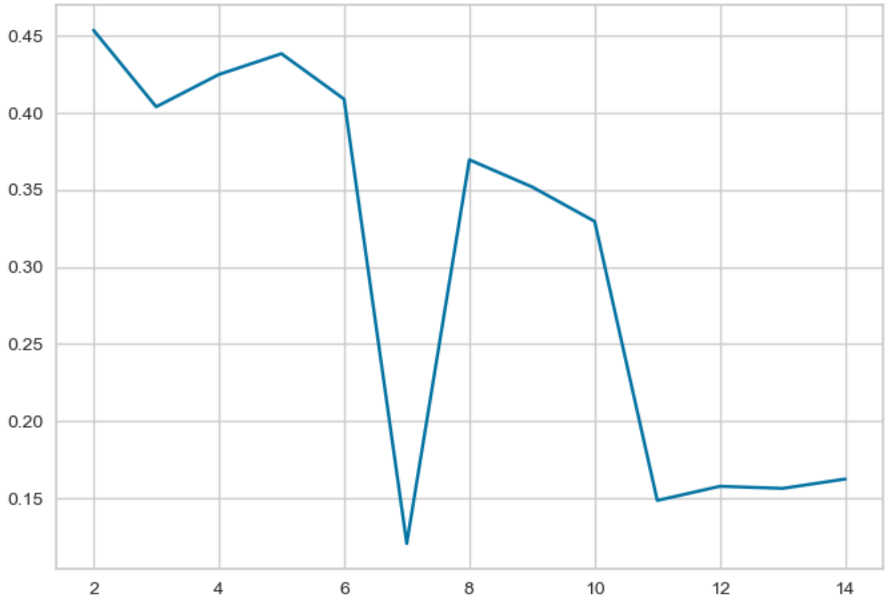
```
Number of Clusters: 1          Average Distortion: 2.5425069919221697
Number of Clusters: 2          Average Distortion: 2.3862098789299604
Number of Clusters: 3          Average Distortion: 2.33620927590848
Number of Clusters: 4          Average Distortion: 2.2190505638334423
Number of Clusters: 5          Average Distortion: 2.133404401901685
Number of Clusters: 6          Average Distortion: 2.0815036860937144
Number of Clusters: 7          Average Distortion: 2.0045413402786814
Number of Clusters: 8          Average Distortion: 1.986423782487441
Number of Clusters: 9          Average Distortion: 1.956222103389025
Number of Clusters: 10         Average Distortion: 1.9360473996664198
Number of Clusters: 11         Average Distortion: 1.8615942883461607
Number of Clusters: 12         Average Distortion: 1.8219574388532505
Number of Clusters: 13         Average Distortion: 1.7936924742607907
Number of Clusters: 14         Average Distortion: 1.7567842179093438
```



Results – The appropriate number of clusters seems to be 6 or 7.

K-means Clustering – Silhouette Scores

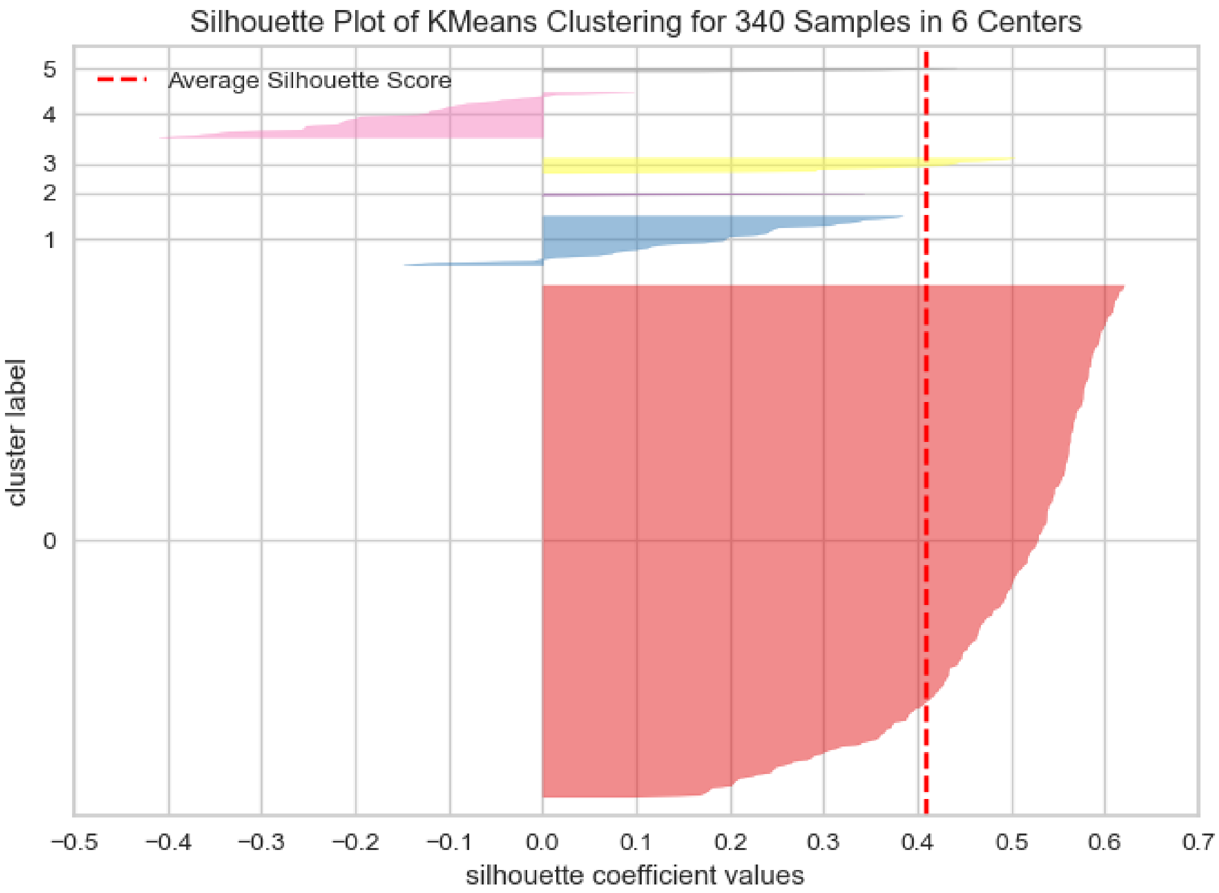
```
For n_clusters = 2, the silhouette score is 0.45335782729503565)
For n_clusters = 3, the silhouette score is 0.40374060030338865)
For n_clusters = 4, the silhouette score is 0.4246430808437099)
For n_clusters = 5, the silhouette score is 0.4381539778147092)
For n_clusters = 6, the silhouette score is 0.40869599703024256)
For n_clusters = 7, the silhouette score is 0.1207450219233897)
For n_clusters = 8, the silhouette score is 0.3693991650696542)
For n_clusters = 9, the silhouette score is 0.35185096182499204)
For n_clusters = 10, the silhouette score is 0.32950073703610283)
For n_clusters = 11, the silhouette score is 0.1486586842527321)
For n_clusters = 12, the silhouette score is 0.15784241071085106)
For n_clusters = 13, the silhouette score is 0.15646997458716602)
For n_clusters = 14, the silhouette score is 0.16253506827999134)
```



Results – The silhouette score goes down dramatically for 7 clusters, so we will continue with 6 clusters.

K-means Clustering – Silhouette Plot

Observation – The silhouette plot shows coefficiency cover of 0.4 on average for most clusters, which is the best we see, comparing to other options (see in the appendix)



K-mean Performance – Cluster Profiling

K-means Clustering – Cluster Profiling

segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	Count in segment
0	71	13	1.4	26	62.0	-193689576.6	1651605486.5	3.6	549560277.1	25.7	-2.0	111
1	73	-0.7	1.3	24	48.0	128553856.2	1425306261.4	3.8	365251101.3	22.0	-2.2	153
2	327	22	2	4	106.0	698240666.7	287547000.	0.8	366763235.3	401.0	-5.3	3
3	163	10	1.5	29	221.3	755269529.4	5933508588.2	5.7	1796106211.3	36.1	9.4	34
4	33	-16	2.9	70	46.8	-241306821.4	-2881485714.3	-6.9	508448395.6	73.9	1.5	28
5	90	6.63	1.5	347	36.5	-27609181.8	-1384512272.7	-3.5	263688145.9	33.6	-34.5	11

Results – Clusters 0 has a negative cash flow, with positive earning per share

Cluster 1 has the smallest price change over 13 months, and positive cash flow, which indicates stability

Cluster 2 has the highest price, price change and P/E ratio.

Cluster 3 has the leading results in many parameters (see above)

Cluster 4 shows the lowest performances, and “leads” only in volatility

Cluster 5 shows uniquely high ROE and less than average results in other parameters

K-means Clustering – Cluster Profiling – Group by Sectors

Cluster 0	Consumer Discretionary	11
	Consumer Staples	9
	Energy	5
	Financials	13
	Health Care	18
	Industrials	18
	Information Technology	14
	Materials	10
	Real Estate	12
	Telecommunications Services	1
Cluster 1	Consumer Discretionary	17
	Consumer Staples	6
	Financials	30
	Health Care	11
	Industrials	32
	Information Technology	10
	Materials	8
	Real Estate	14
	Telecommunications Services	1
	Utilities	24

Cluster 2	Consumer Discretionary	1
	Health Care	1
	Information Technology	1
Cluster 3	Consumer Discretionary	6
	Consumer Staples	2
	Energy	2
	Financials	4
	Health Care	10
	Information Technology	6
	Real Estate	1
	Telecommunications Services	3
Cluster 4	Energy	22
	Industrials	2
	Information Technology	2
	Materials	2
Cluster 5	Consumer Discretionary	5
	Consumer Staples	2
	Energy	1
	Financials	2
	Industrials	1

Observation – Clusters 0 and 1 have higher number of Health care, Industrials, IT and real estate companies
Cluster 3 has the leading results in many parameters, and a big part of it is health care companies
Cluster 4 shows the lowest performances, and “leads” only in volatility. Most of the energy companies are in this group

K-means Clustering – Cluster Profiling - Insights

Cluster 0:

- This cluster contains 111 companies from various sectors
- Average net cash flow is negative ($-194M)
- Price change is high and positive (13) but not the highest and the current price is $71

Cluster 1:

- This cluster contains 153 companies from various sectors
- Average net cash flow is positive ($130M), but not the best
- Price change is very close to 0, and volatility is the lowest, which indicates stability
- In other parameters, this cluster is not very different than cluster 0

Cluster 2:

- This cluster has only 3 companies from 3 sectors, 'Alexion Pharmaceuticals’, 'Amazon.com Inc’ and 'Netflix Inc.'
- The current price in this cluster is the highest by far from the other clusters ($327) and so is the P/E ratio (401) which indicate a very strong performance, creating a league of its own.

K-means Clustering – Cluster Profiling – Insights – Cont.

Cluster 3:

- These 34 companies have, by average, the best performances, with the second highest share price. That indicates stability and confidence by the market. This group includes leading companies in telecommunication (AT&T, Verizon), pharmaceutical (Pfizer), tech (Intel) and banking (JPMorgan Chase)

Cluster 4:

- These 28 companies seem to struggle. Their price fell in the past year, and it was very volatile. They are losing cash and their earning per share is negative, which means the market is not confident about their long-term success. There are 22 energy companies in this group.

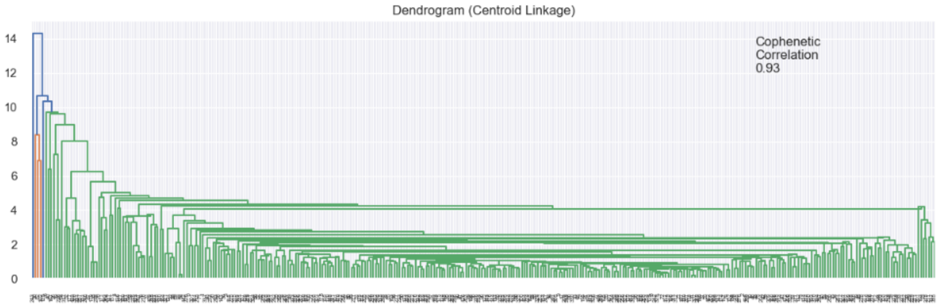
Cluster 5:

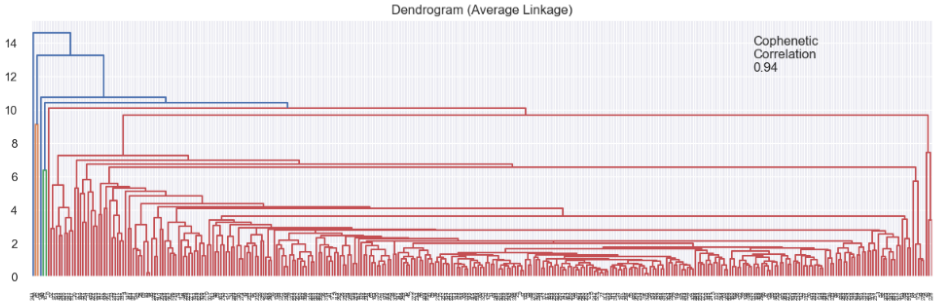
- These 11 companies show an anomaly that needs to be investigated further by investors. The ROE, on average, is very high, and the shares price has gone up, but the cash flow is negative and the earning per share is negative.

Hierarchical Clustering

Hierarchical Clustering– Cophenetic Correlation and Dendrogram

Highest cophenetic correlation is 0.942254, which is obtained with **Euclidean** distance and **average** linkage. The second highest is with **centroid** linkage (0.931401)





Observation – The cophenetic correlation is highest for average and centroid linkage methods.

- We will move ahead with average linkage.
- Since we found 6 clusters to be the appropriate number of clusters after running the K-mean clustering, we will start with assuming this is the right number of clusters

AgglomerativeClustering Model Performance – Using sklearn

AgglomerativeClustering Model – Cluster Profiling – 6 clusters

HC segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	Count in segment
0	77	4	2	35	67	-33197321	1538074667	3	560505037	32	-2	333
1	26	11	1	13	131	16755500000	13654000000	3	2791829362	14	2	2
2	24	-13	3	802	51	-1292500000	-19106500000	-42	519573983	61	2	2
3	105	16	1	8	958	592000000	3669000000	1	2800763359	80	6	1
4	1275	3	1	29	184	-1671386000	2551360000	50	50935516	25	-1	1
5	277	6	1	30	25	90885000	596541000	9	66951852	31	129	1

Results – Cluster 0 contains almost the entire set (333 records out of 340). This show a very poor clustering, and it only singles the most extreme performers in each category
Cluster 1 has the highest net cash flow and net income.
Cluster 2 has the highest ROE and the highest volatility
Cluster 3 has the highest cash ratio, price change and P/E ratio
Cluster 4 shows the highest price and the estimated shares outstanding
Cluster 5 shows the highest P/B ratio

AgglomerativeClustering Model – Cluster Profiling – 10 clusters

HC segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	Count in segment
0	75	4	2	34	67	-37431142	1546386296	3	562148126	31	-2	331
1	1275	3	1	29	184	-1671386000	2551360000	50	50935516	25	-1	1
2	5	-38	5	687	22	-3283000000	-14685000000	-22	654703522	28	-2	1
3	34	14	1	19	162	12747000000	11420000000	2	4738589212	14	4	1
4	105	16	1	8	958	592000000	3669000000	1	2800763359	80	6	1
5	44	11	2	917	80	698000000	-23528000000	-61	384444444	93	5	1
6	17	8	1	6	99	20764000000	15888000000	4	845069512	13	-1	1
7	676	32	1	4	58	1333000000	596000000	1	465625000	528	4	1
8	183	4	2	589	0	2000000	-271000000	-2	111522634	21	-76	1
9	277	6	1	30	25	90885000	596541000	9	66951852	31	129	1

Results – Again, one cluster, cluster 0 contains almost the entire set (331 records out of 340). This shows that for this data set, hierarchal clustering fails to perform well, and each cluster is organized around one leading parameter.

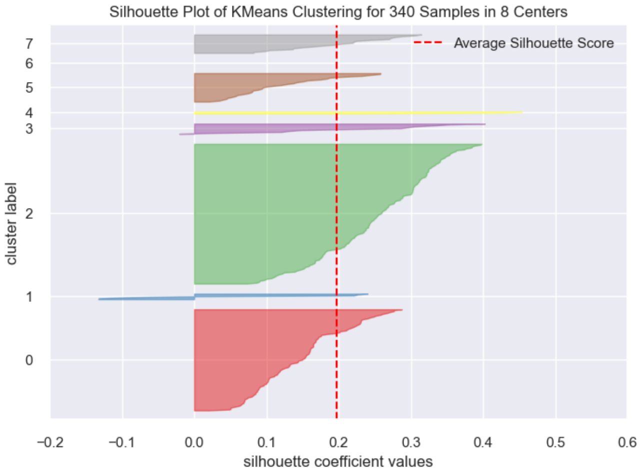
K-means vs Hierarchical Clustering – Comparison table

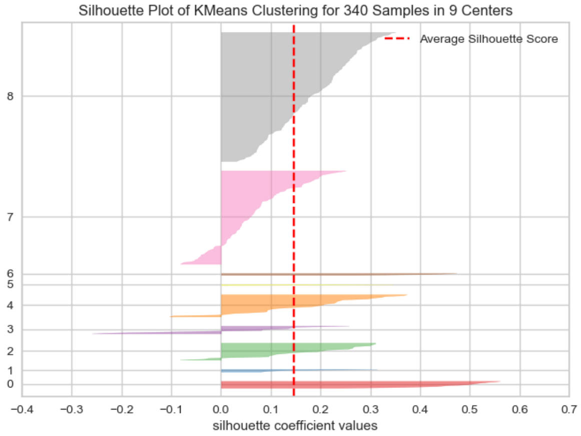
Parameter	**K-means model**	**Hierarchical clustering**
Execution Times (for this data set)	Less than a second for each step	Generating dendrograms took 15 seconds
Distinct clustering and their sizes	6 distinct clusters (sizes – 111, 153, 3, 34, 22, 11)	5 distinct clusters of 1 and all the rest in a 'default' cluster (cluster sizes – 335, 1,1,1,1,1)
No. of clusters	6	Attempted to create 6, 9 and 10
No. of observations	2 (6 and 8 clusters)	Multiple (different metrics, different # of clusters)

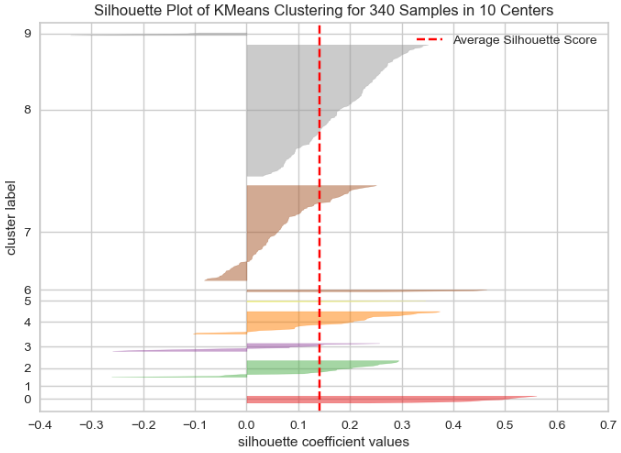
Summary – For this data set, K-means clustering resulted in a logical clustering that can be explained and used. Hierarchical clustering failed to create meaningful clustering, and resulted in multiple clusters of 1, while all the rest are piled in one cluster

APPENDIX

K-means Clustering – Silhouette Plot Comparison

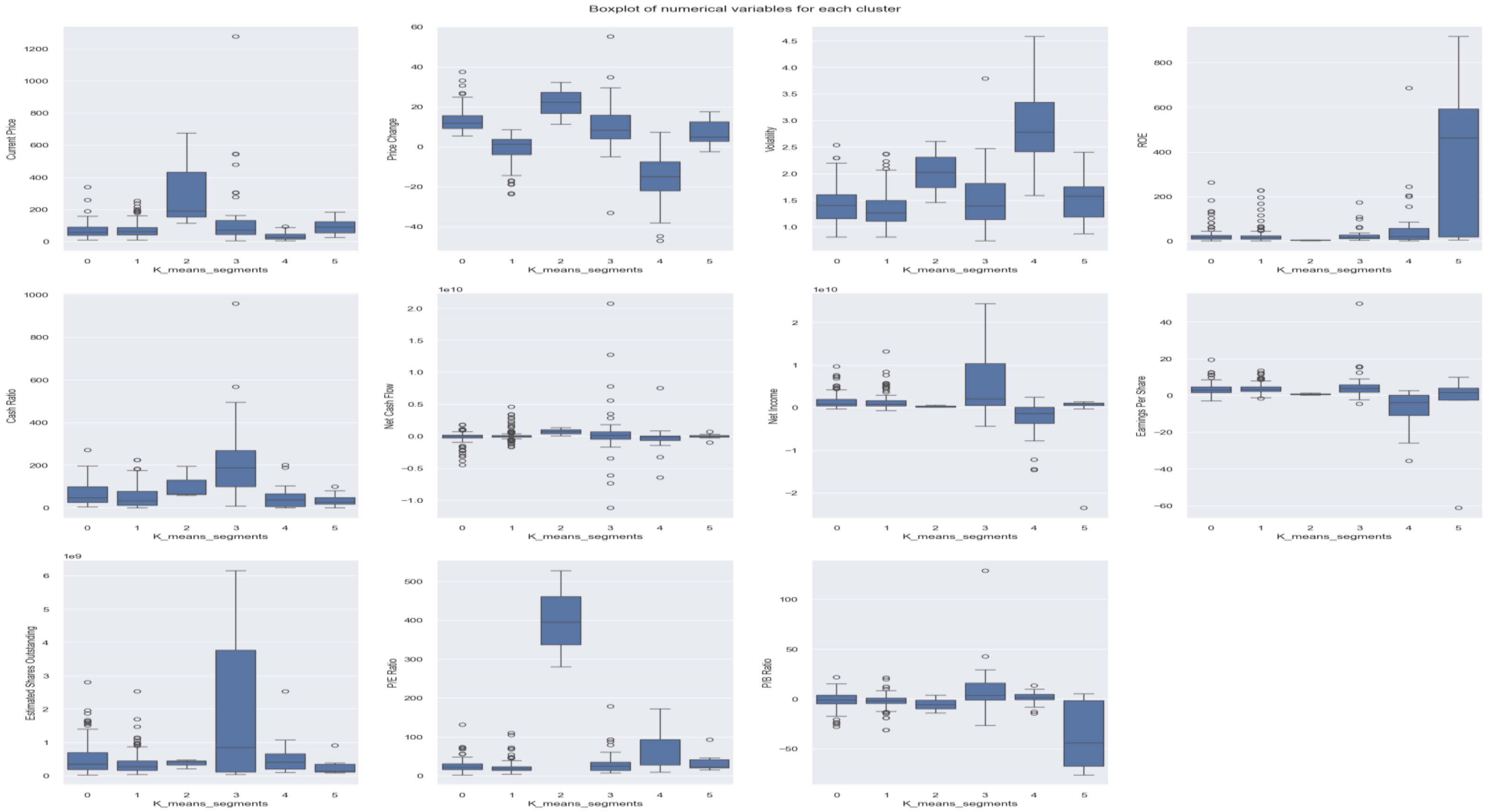






Observation – The silhouette plot coefficiency for 8, 9 and 10 clusters covers <0.2 in average, which is significantly less than 6 clusters (average of 0.4)

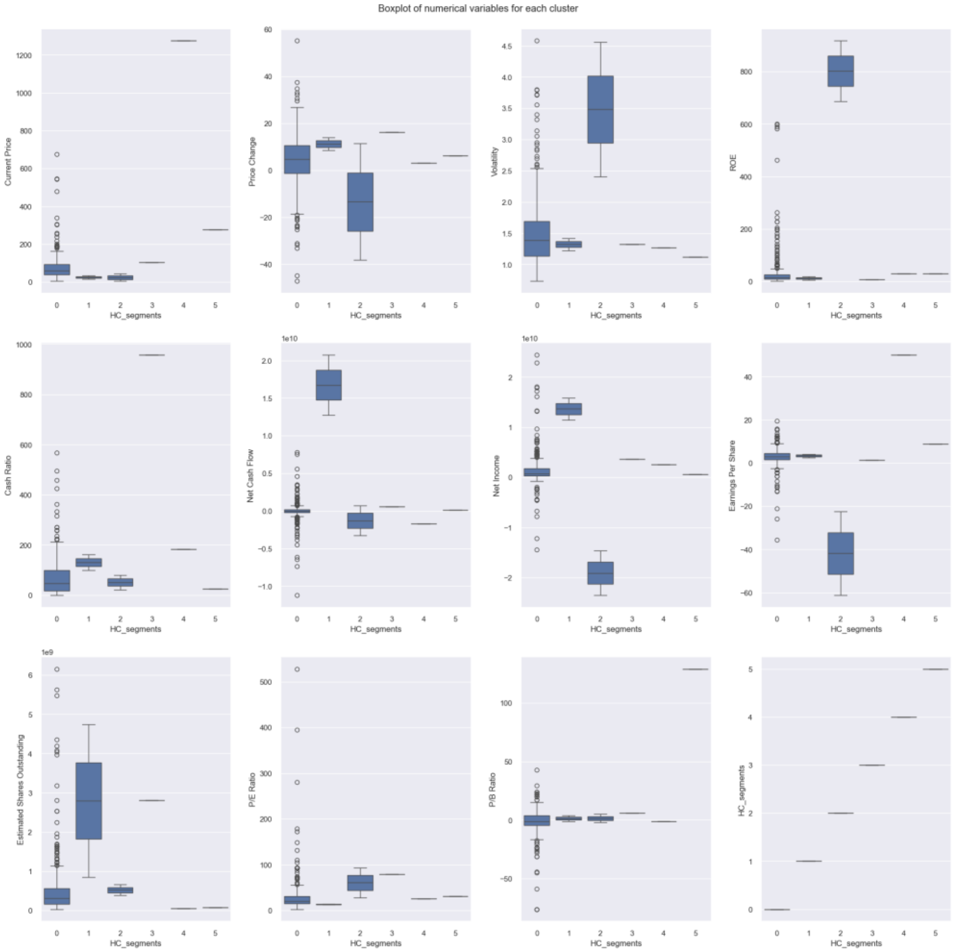
K-means Clustering – Cluster Profiling – Boxplot



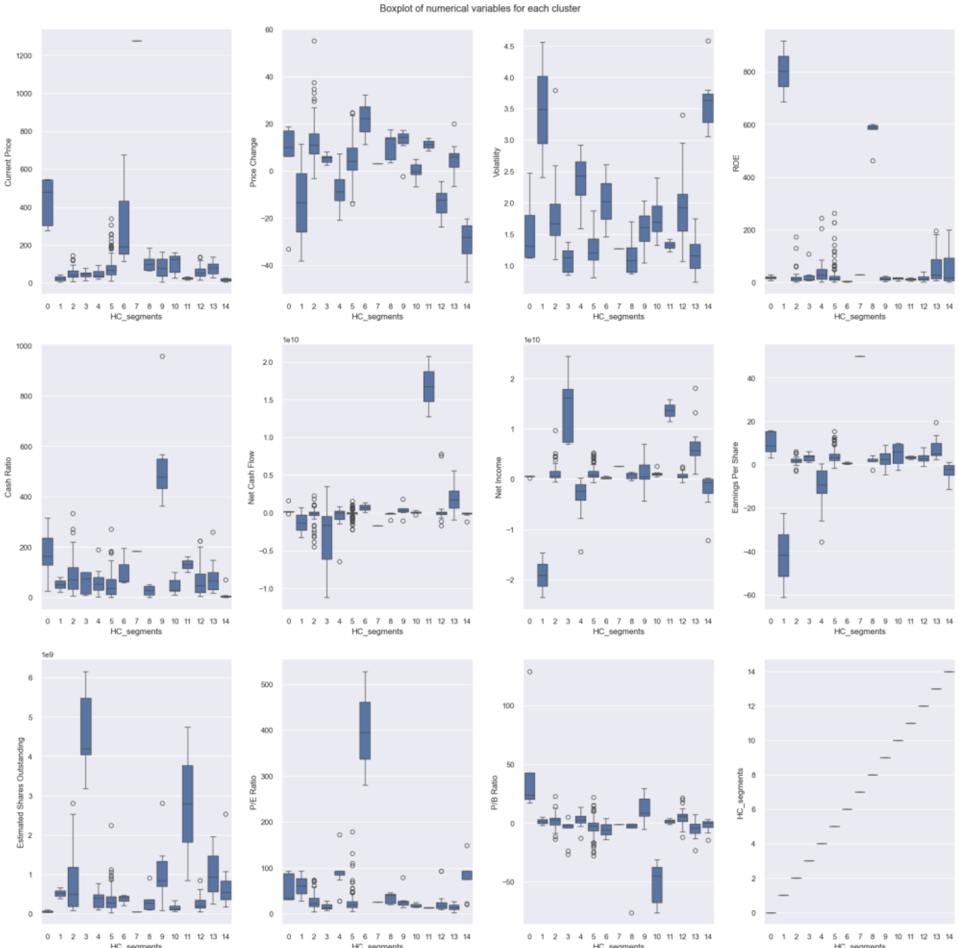
Hierarchical Clustering– Computing Cophenetic Correlation

```
Cophenetic correlation for Euclidean distance and single linkage is 0.9232271494002922.
Cophenetic correlation for Euclidean distance and complete linkage is 0.7873280186580672.
Cophenetic correlation for Euclidean distance and average linkage is 0.9422540609560814.
Cophenetic correlation for Euclidean distance and weighted linkage is 0.8693784298129404.
Cophenetic correlation for Chebyshev distance and single linkage is 0.9062538164750717.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.598891419111242.
Cophenetic correlation for Chebyshev distance and average linkage is 0.9338265528030499.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.9127355892367.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.925919553052459.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.792530720285.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.9247324030159737.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.8708317490180426.
Cophenetic correlation for Cityblock distance and single linkage is 0.9334186366528574.
Cophenetic correlation for Cityblock distance and complete linkage is 0.7375328863205818.
Cophenetic correlation for Cityblock distance and average linkage is 0.9302145048594667.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.731045513520281.
******************************************************************************************
Highest cophenetic correlation is 0.9422540609560814, which is obtained with Euclidean distance and average linkage.
```

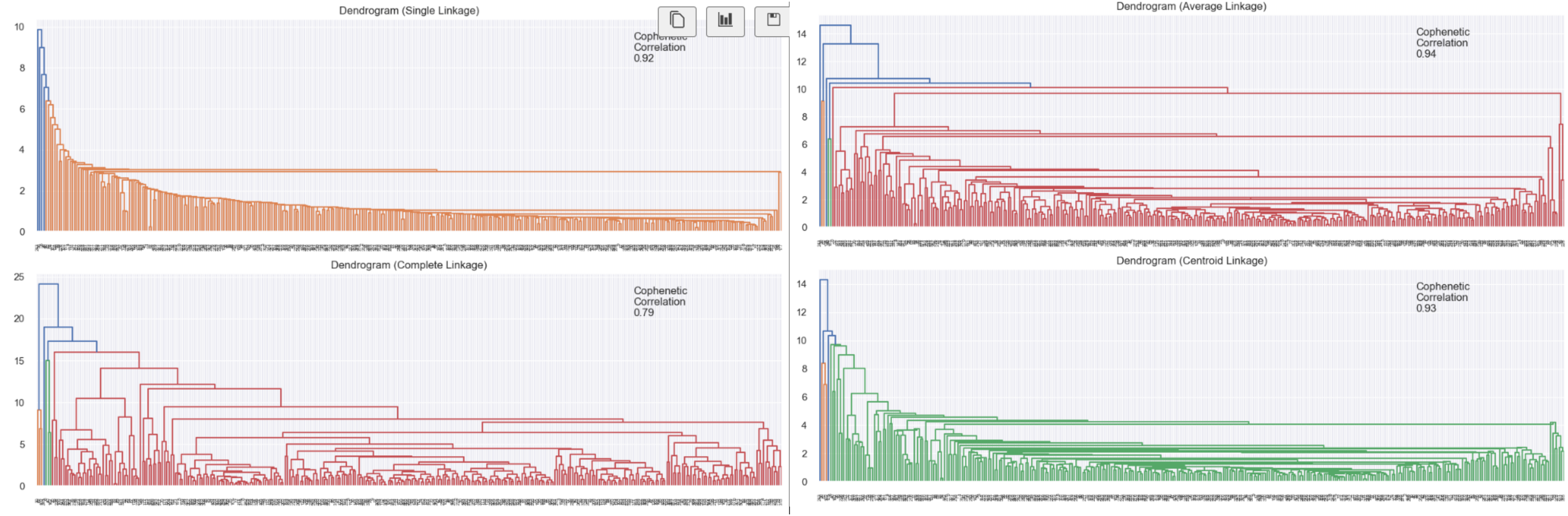
AgglomerativeClustering – 6 Clusters Profiling – Boxplot



AgglomerativeClustering – 10 Clusters Profiling – Boxplot



Hierarchical Clustering – Dendrograms plots



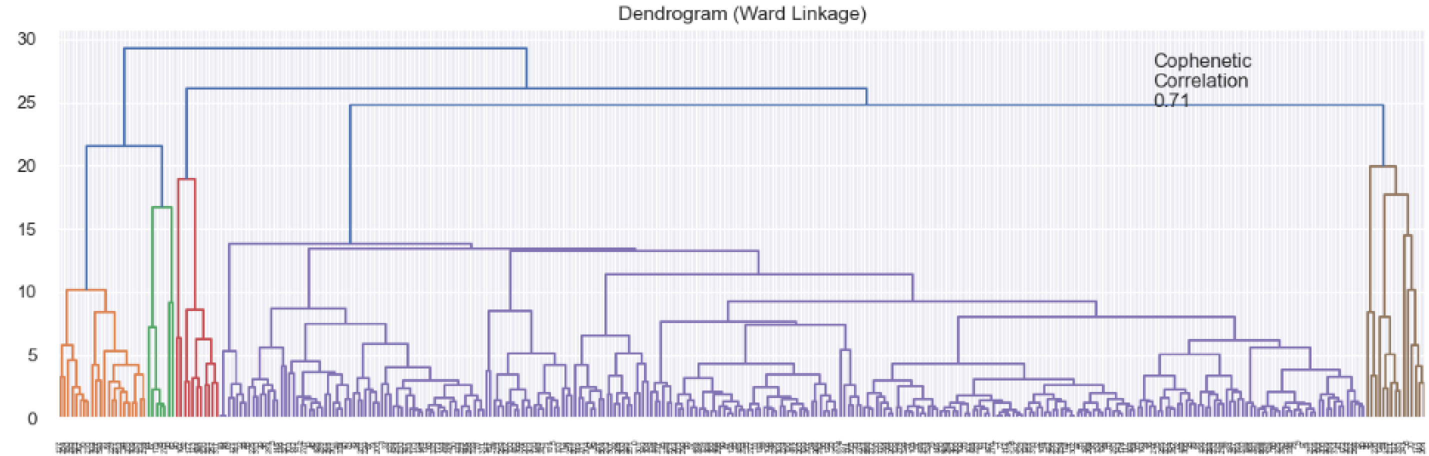
Dendrogram (Single Linkage)

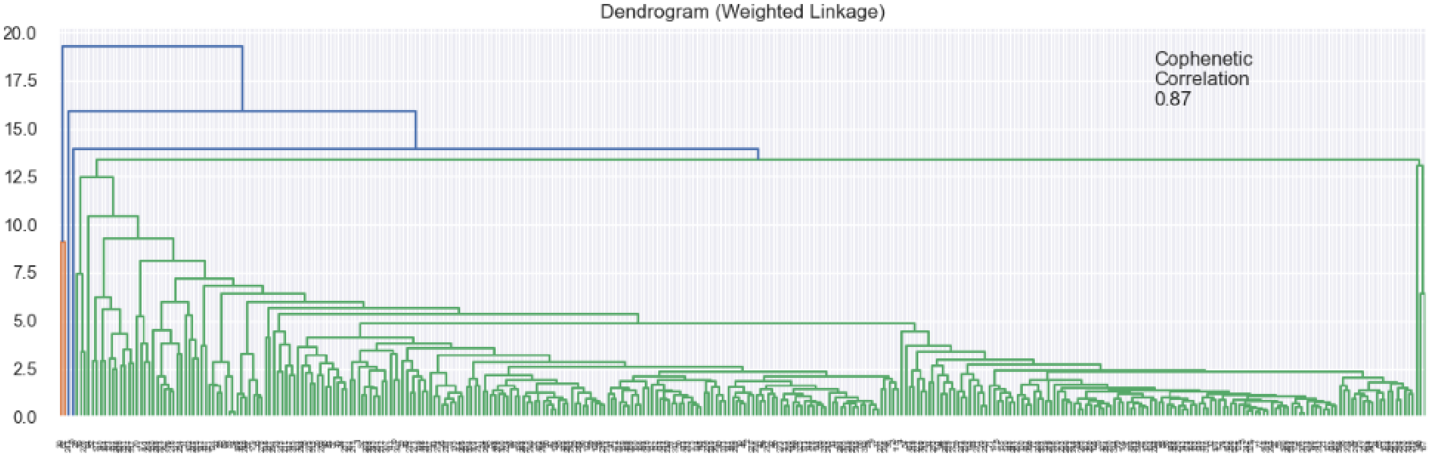
Dendrogram (Average Linkage)

Dendrogram (Complete Linkage)

Dendrogram (Centroid Linkage)

Hierarchical Clustering – Dendrograms plots





Great Learning

Happy Learning !