Trade and Ahead

Unsupervised Learning

September 13, 2023

Contents / Agenda

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

Executive Summary

Summary of observations and conclusions:

- Net income becomes the highest mean, min and max value, surpassed all other parameters.
- Both clustering technique able to fit the dataset and clustering profiling.
- Hierarchical clustering able to give more distinct value with good observations in each parameter.
- All of the clusters are easily identify, grouped and able to present to their customers for stocks portfolio.

Business Problem Overview and Solution Approach

- Business problem overview:
    - Trade and Ahead could use these clusters as an starting point for financial statement analysis, especially for sector who do not fit/suitable in the cluster.
    - Cluster 3 was seen perform worst even though they are an important sector. More detailed analysis is needed to find the real root cause why their performance was worst.
    - Trade and Ahead could use this analysis for more observation and use this as forecasting analysis and create a supervised model for more solid financial analysis.

Business Problem Overview and Solution Approach

- Solution approach/business improvement/recommendation
  - Trade and Ahead should further identify the financial goals, risk tolerance, and investment behaviors for their clients, before recommend them a cluster as stocks portfolio.
  - Trade and Ahead could use this analysis to become financial advisor to focus on cluster 3 portfolio, where their business looks uncertain and challenging.
  - Trade and Ahead may need to proceed for modeling the good vs worst cluster to find the difference and prediction on the good sector to attract more investor.

EDA Results

- Data shape: 340 rows, 15 columns
- First 5 data head on the data as below:

	Ticker Symbol	Security	GICS Sector	GICS Sub Industry	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
0	AAL	American Airlines Group	Industrials	Airlines	42.349998	9.999995	1.687151	135	51	-604000000	7610000000	11.39	6.681299e+08	3.718174	-8.784219
1	ABBV	AbbVie	Health Care	Pharmaceuticals	59.240002	8.339433	2.197887	130	77	51000000	5144000000	3.15	1.633016e+09	18.806350	-8.750068
2	ABT	Abbott Laboratories	Health Care	Health Care Equipment	44.910000	11.301121	1.273646	21	67	938000000	4423000000	2.94	1.504422e+09	15.275510	-0.394171
3	ADBE	Adobe Systems Inc	Information Technology	Application Software	93.940002	13.977195	1.357679	9	180	-240840000	629551000	1.26	4.996437e+08	74.555557	4.199651
4	ADI	Analog Devices, Inc.	Information Technology	Semiconductors	55.320000	-1.827858	1.701169	14	272	315120000	696878000	0.31	2.247994e+09	178.451613	1.059810

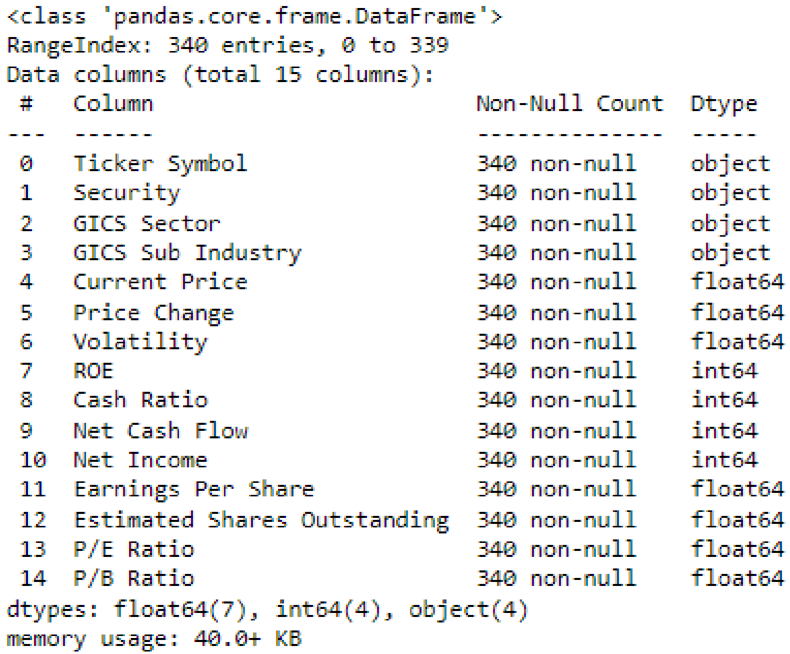
EDA Results

- Displaying a few rows of the dataset, as below:

	Ticker Symbol	Security	GICS Sector	GICS Sub Industry	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
102	DVN	Devon Energy Corp.	Energy	Oil & Gas Exploration & Production	32.000000	-15.478079	2.923698	205	70	830000000	-14454000000	-35.55	4.065823e+08	93.089287	1.785616
125	FB	Facebook	Information Technology	Internet Software & Services	104.660004	16.224320	1.320606	8	958	592000000	3669000000	1.31	2.800763e+09	79.893133	5.884467
11	AIV	Apartment Investment & Mgmt	Real Estate	REITs	40.029999	7.578608	1.163334	15	47	21818000	248710000	1.52	1.636250e+08	26.335526	-1.269332
248	PG	Procter & Gamble	Consumer Staples	Personal Products	79.410004	10.660538	0.806056	17	129	160383000	636056000	3.28	4.913916e+08	24.070121	-2.256747
238	OXY	Occidental Petroleum	Energy	Oil & Gas Exploration & Production	67.610001	0.865287	1.589520	32	64	-588000000	-7829000000	-10.23	7.652981e+08	93.089287	3.345102
336	YUM	Yum! Brands Inc	Consumer Discretionary	Restaurants	52.516175	-8.698917	1.478877	142	27	159000000	1293000000	2.97	4.353535e+08	17.682214	-3.838260
112	EQT	EQT Corporation	Energy	Oil & Gas Exploration & Production	52.130001	-21.253771	2.364883	2	201	523803000	85171000	0.56	1.520911e+08	93.089287	9.567952
147	HAL	Halliburton Co.	Energy	Oil & Gas Equipment & Services	34.040001	-5.101751	1.966062	4	189	7786000000	-671000000	-0.79	8.493671e+08	93.089287	17.345857
89	DFS	Discover Financial Services	Financials	Consumer Finance	53.619999	3.653584	1.159897	20	99	2288000000	2297000000	5.14	4.468872e+08	10.431906	-0.375934
173	IVZ	Invesco Ltd.	Financials	Asset Management & Custody Banks	33.480000	7.067477	1.580839	12	67	412000000	968100000	2.26	4.283628e+08	14.814159	4.218620

EDA Results

- No duplicated data was found.
- No missing data detected prior for analysis
- Data info as below:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 340 entries, 0 to 339
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Ticker Symbol                 340 non-null    object 
 1   Security                      340 non-null    object 
 2   GICS Sector                   340 non-null    object 
 3   GICS Sub Industry             340 non-null    object 
 4   Current Price                 340 non-null    float64
 5   Price Change                  340 non-null    float64
 6   Volatility                    340 non-null    float64
 7   ROE                           340 non-null    int64  
 8   Cash Ratio                    340 non-null    int64  
 9   Net Cash Flow                 340 non-null    int64  
 10  Net Income                    340 non-null    int64  
 11  Earnings Per Share            340 non-null    float64
 12  Estimated Shares Outstanding  340 non-null    float64
 13  P/E Ratio                     340 non-null    float64
 14  P/B Ratio                     340 non-null    float64
dtypes: float64(7), int64(4), object(4)
memory usage: 40.0+ KB
```

EDA Results

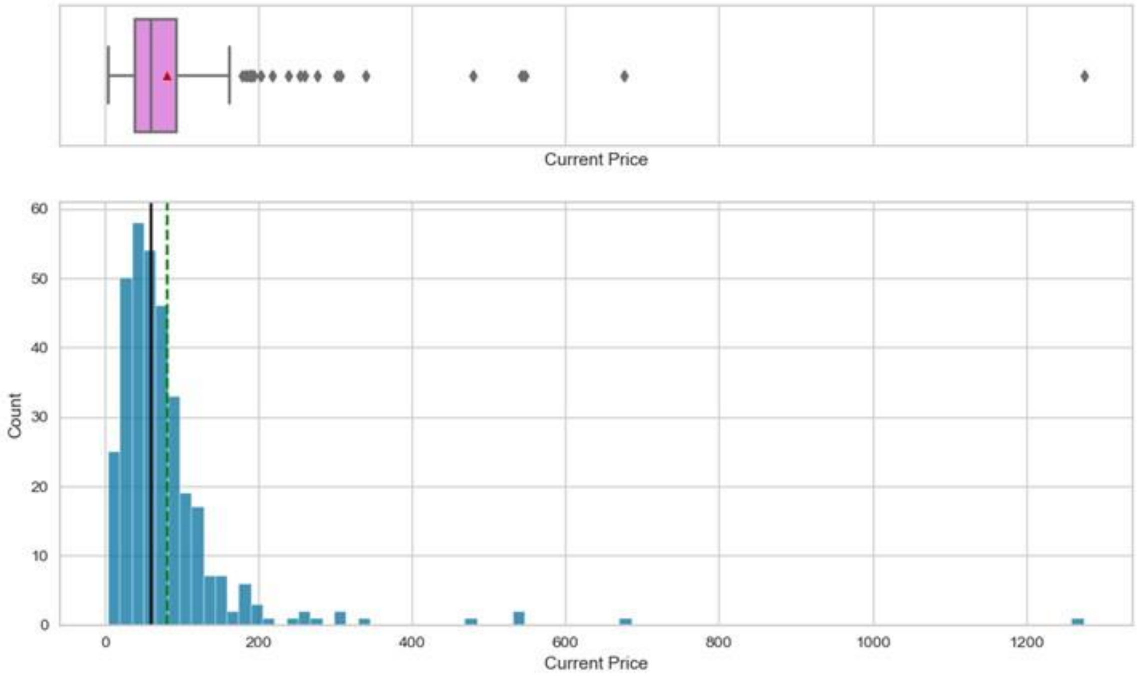
- Statistical summary for the dataset as below:

	count	mean	std	min	25%	50%	75%	max
Current Price	340.0	8.086234e+01	9.805509e+01	4.500000e+00	3.855500e+01	5.970500e+01	9.288000e+01	1.274950e+03
Price Change	340.0	4.078194e+00	1.200634e+01	-4.712969e+01	-9.394838e-01	4.819505e+00	1.069549e+01	5.505168e+01
Volatility	340.0	1.525976e+00	5.917984e-01	7.331632e-01	1.134878e+00	1.385593e+00	1.695549e+00	4.580042e+00
ROE	340.0	3.959706e+01	9.654754e+01	1.000000e+00	9.750000e+00	1.500000e+01	2.700000e+01	9.170000e+02
Cash Ratio	340.0	7.002353e+01	9.042133e+01	0.000000e+00	1.800000e+01	4.700000e+01	9.900000e+01	9.580000e+02
Net Cash Flow	340.0	5.553762e+07	1.946365e+09	-1.120800e+10	-1.939065e+08	2.098000e+06	1.698108e+08	2.076400e+10
Net Income	340.0	1.494385e+09	3.940150e+09	-2.352800e+10	3.523012e+08	7.073360e+08	1.899000e+09	2.444200e+10
Earnings Per Share	340.0	2.776662e+00	6.587779e+00	-6.120000e+01	1.557500e+00	2.895000e+00	4.620000e+00	5.009000e+01
Estimated Shares Outstanding	340.0	5.770283e+08	8.458496e+08	2.767216e+07	1.588482e+08	3.096751e+08	5.731175e+08	6.159292e+09
P/E Ratio	340.0	3.261256e+01	4.434873e+01	2.935451e+00	1.504465e+01	2.081988e+01	3.176476e+01	5.280391e+02
P/B Ratio	340.0	-1.718249e+00	1.396691e+01	-7.611908e+01	-4.352056e+00	-1.067170e+00	3.917066e+00	1.290646e+02

EDA Results

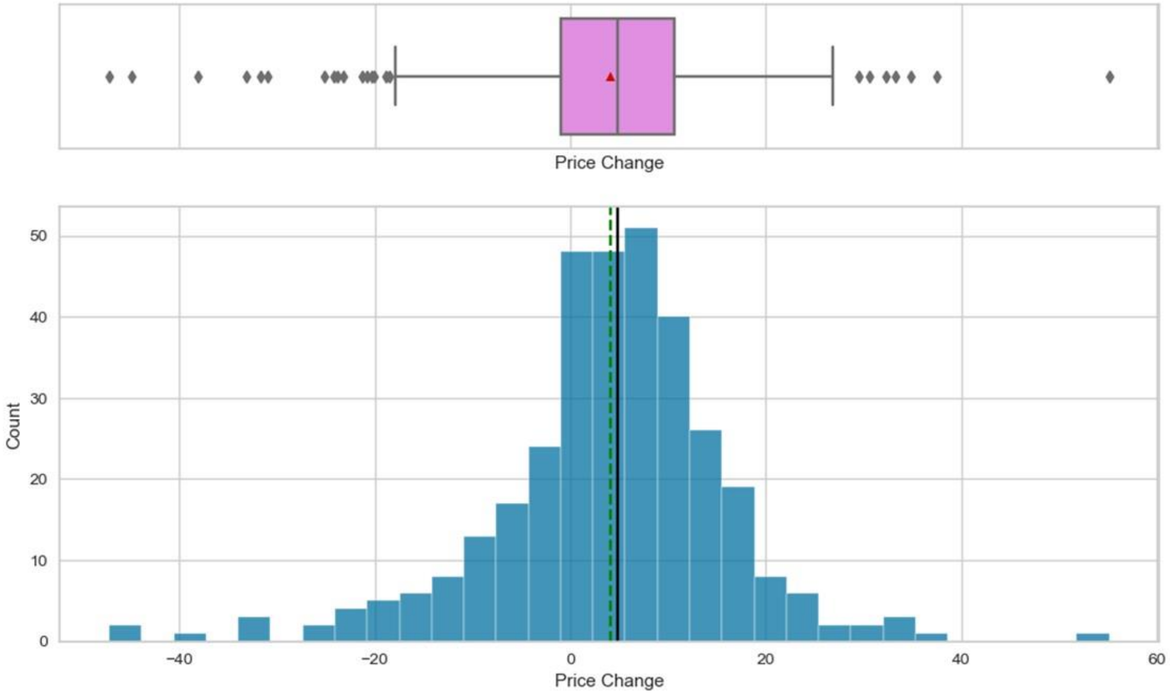
- Statistical summary shows:
  - All data collected was based on 340 data.
  - Net income conquered all the parameters with
    - highest mean: 1.494385 x 10^{9}
    - Lowest min: 2.352800 x 10^{10}
    - Highest max: 2.442 x 10^{10}

EDA Results – Current Price



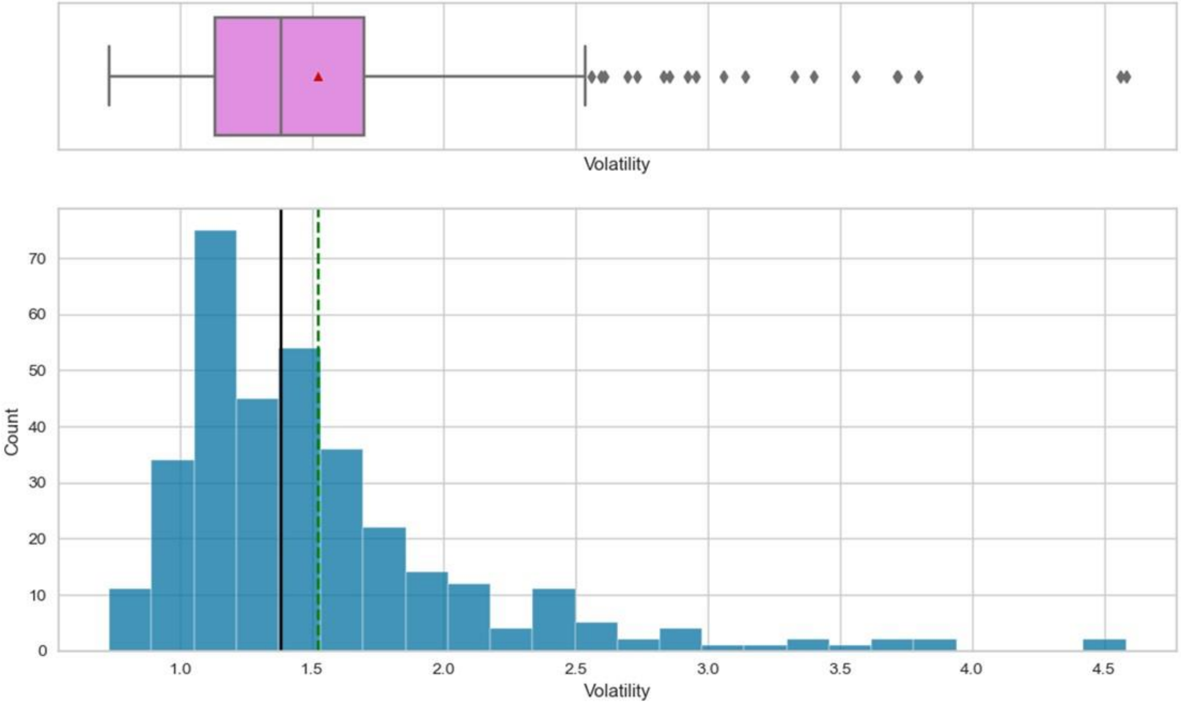
- Current Price observed was skewed to right with the highet amount estimated about 1300.
- The range current price can be seen vary from 10-200.
- Higher price than 200 was observed having few and vary around 300, 500 and 700.
- No stock is listed less than 0.

EDA Results – Price Change



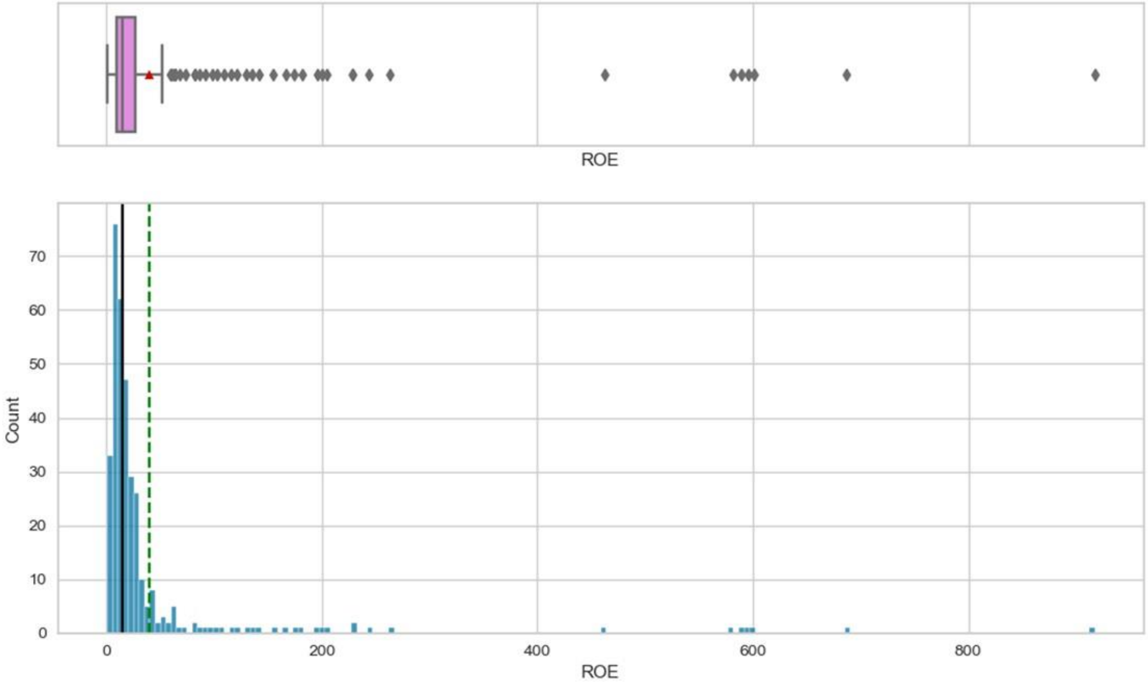
- Price change was observed having a good distribution of bell curve shaped, even though we can see some outlier from left and right.
- The average of price change estimated about 5, while some data on the left is at negative side, with -50 at the lowest and highest data at the right estimated at 55.

EDA Results - Volatility



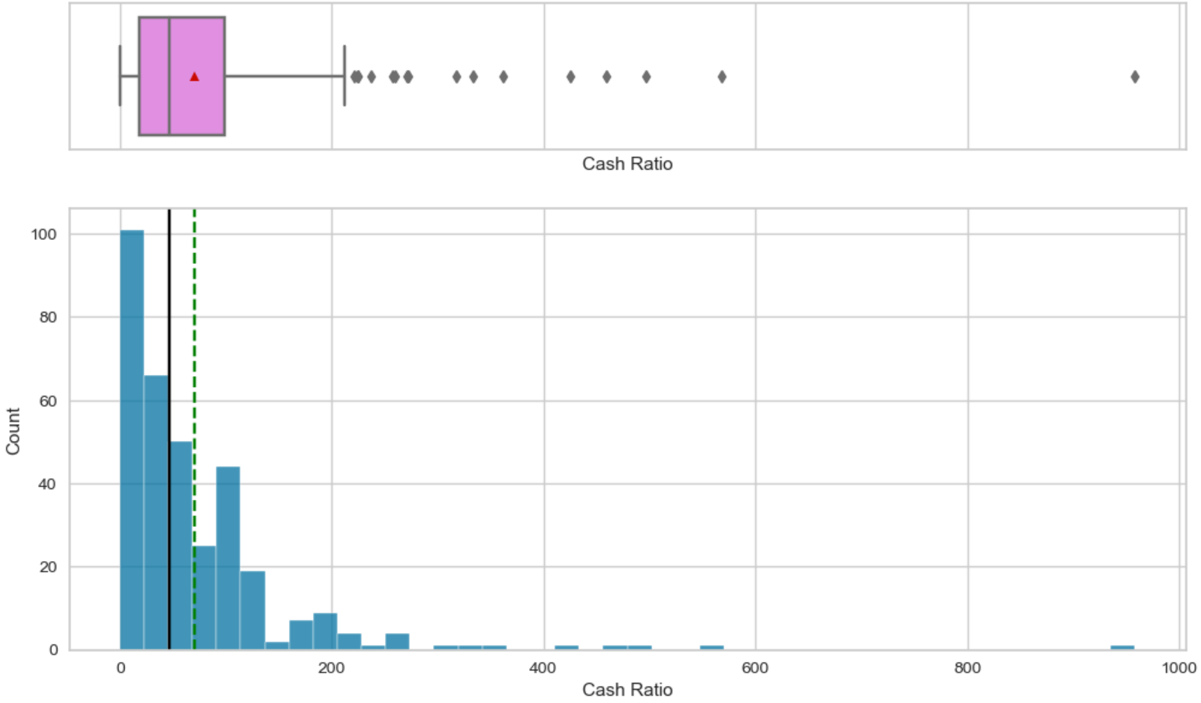
- It is observed the data is right skewed and few outlier at the right side.
- The evarage amount is estimated around 1.4.
- The highest volatility was observed at 4.5

EDA Results – ROE



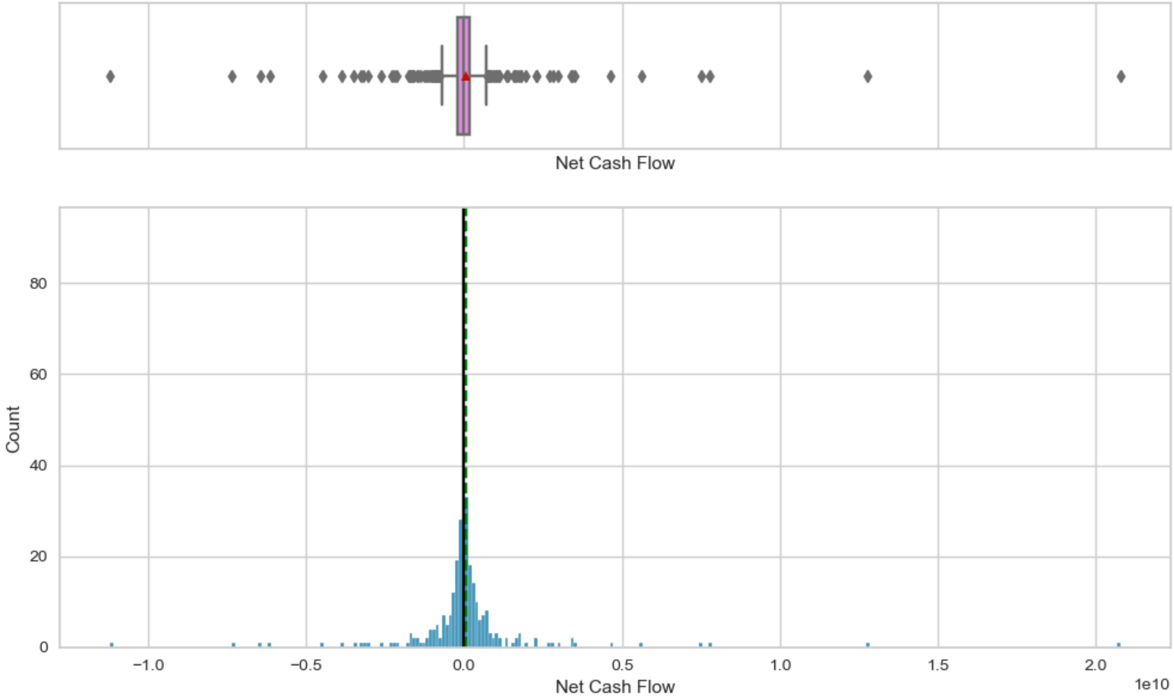
- The distribution is heavily skewed to the right side.
- No stock is listed less than 0.
- The highest ROE was observed at 950.

EDA Results – Cash Ratio



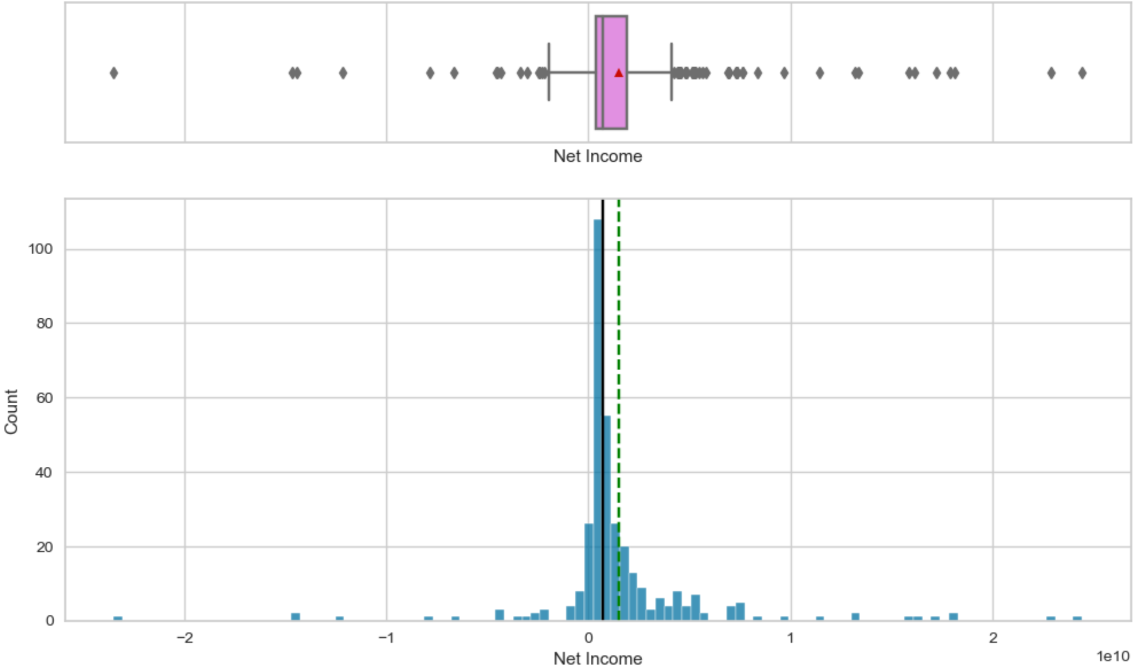
- The distribution is heavily skewed to the right side.
- No stock is listed less than 0.
- Average cash ratio is about 80, and the highest outlier of cash ratio estimated at 900.

EDA Results – Net Cash Flow



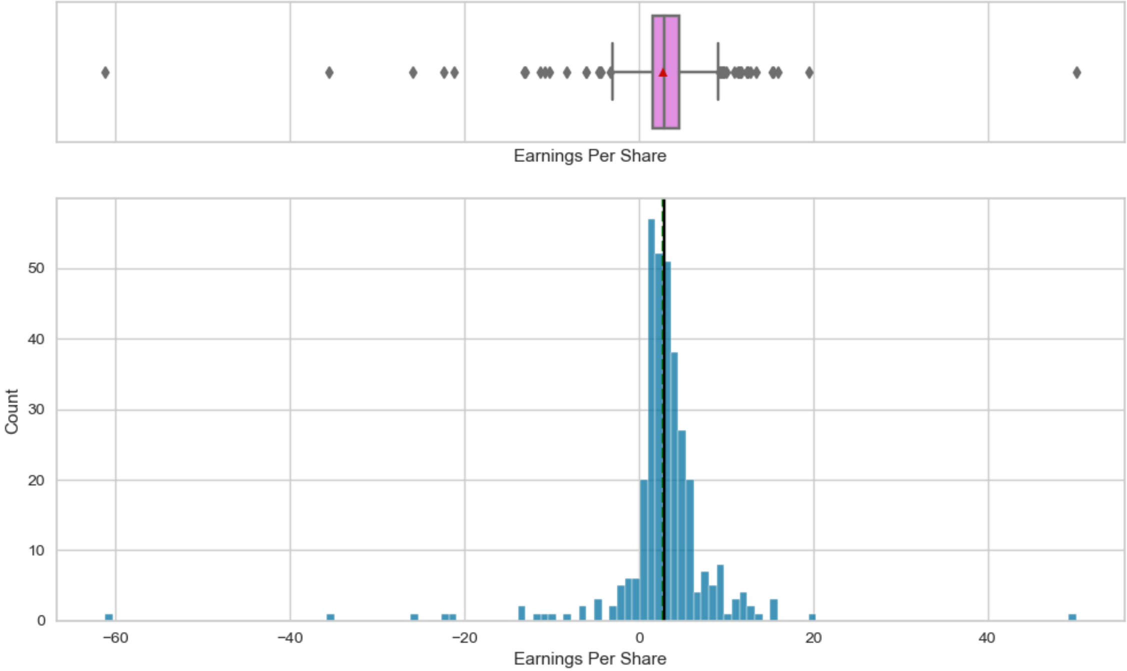
- Net cash flow observed having good bell curve of distribution even though there is an outlier from left and right side.
- The lowest net cash flow is at -1.5 (e10) and the highest net cash flow at 2.2 (e10).

EDA Results – Net Income



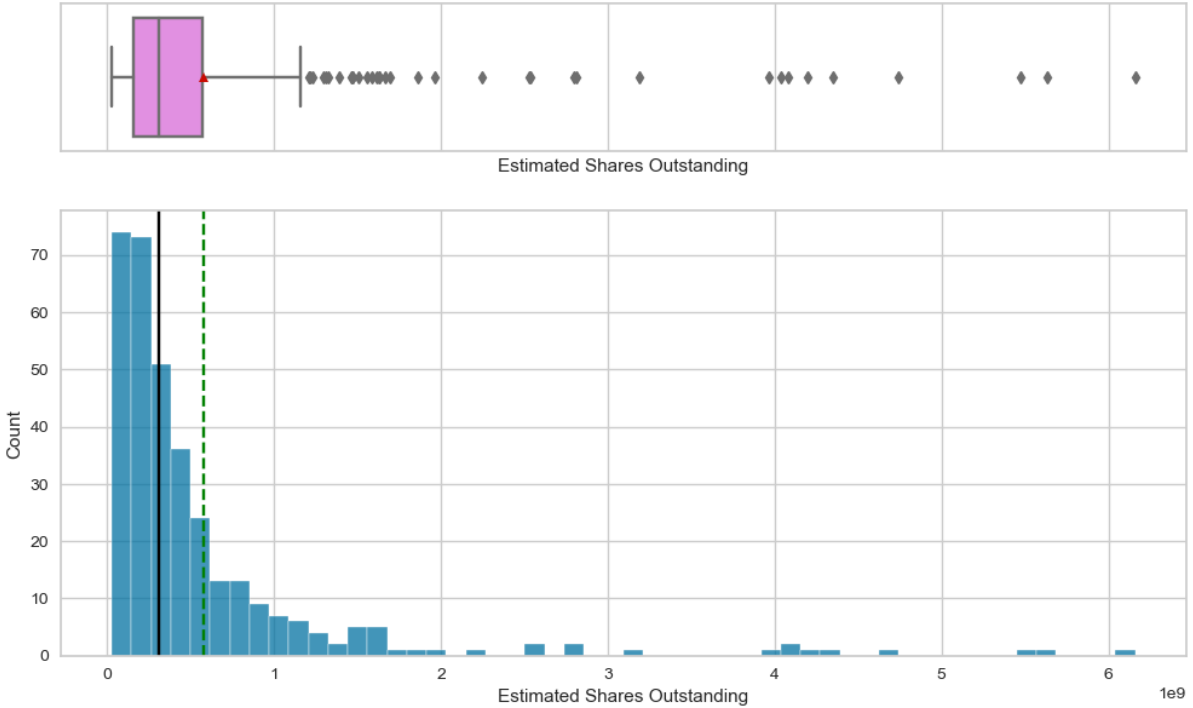
- Net income observed having the almost good of bell curve shape with an outlier from left and quite heavily at the right side.
- The lowest net income is at -2.5 (e10) and the highest is at the 2.6 (e10).
- It is observed some company is doing extremely good and some company loss a lot of money.

EDA Results – Earnings Per Share



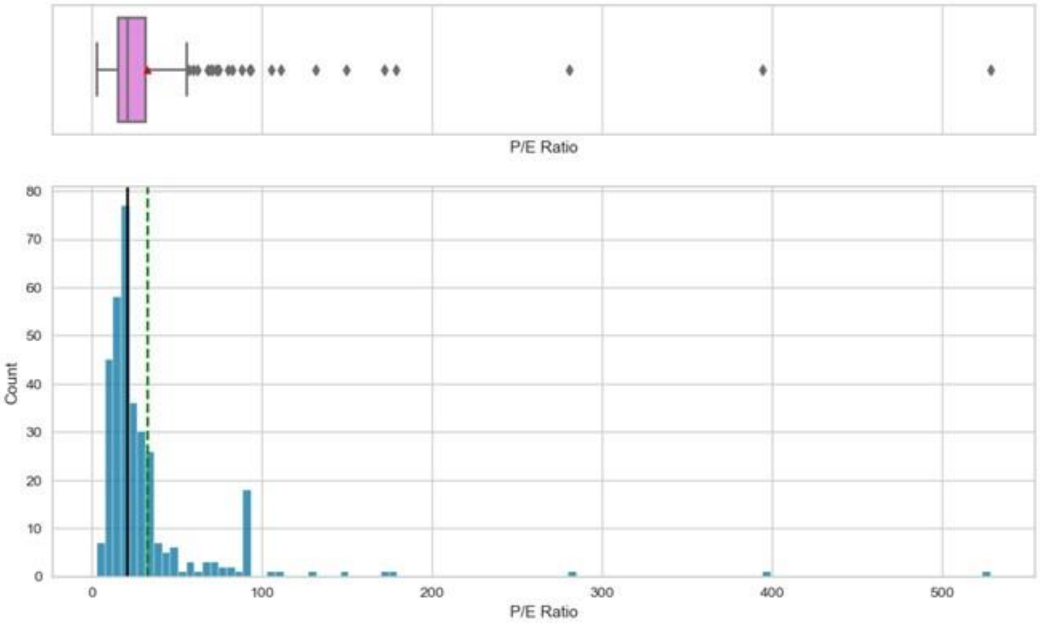
- Data shows some pattern of the bell shape curve but observed few heavy outliers from left and some outliers from right.
- The lowest earnings per share was recorded at -61 while the highest was recorded at 51.
- The average for earnings per share about 3 dollars.

EDA Results – Estimated Shares Outstanding



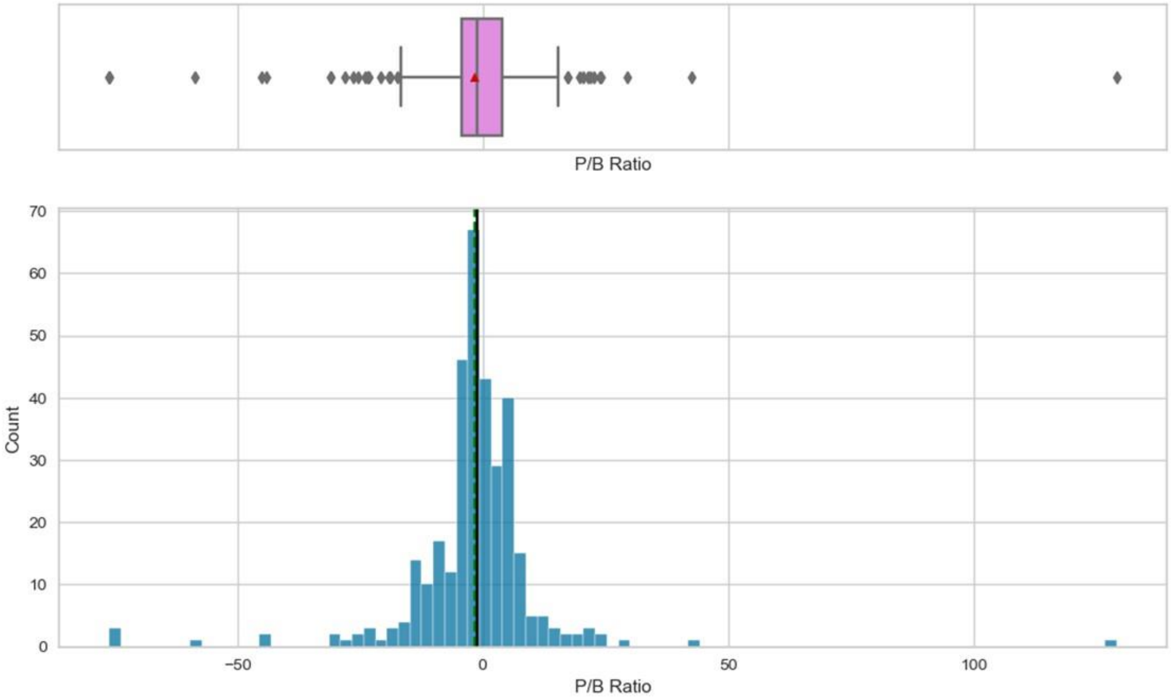
- The data observed was heavily skewed to the right with the highest outstanding recorded at 6.4-6.5 (e9).
- The average outstanding is recorded at 0.3 (e9).
- No outstanding was observed below 0.

EDA Results – P/E Ratio



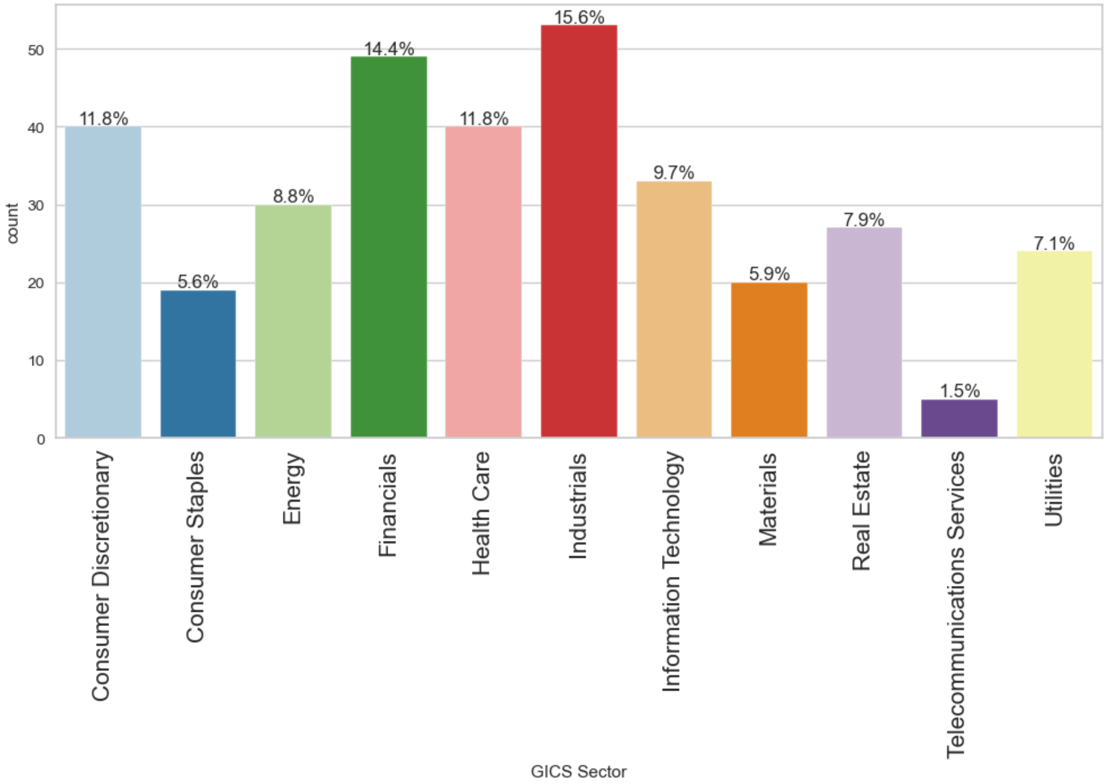
- The data is heavily skewed to the right side.
- No ratio was observed on the negative side and no ratio was observed below 0.

EDA Results – P/B Ratio



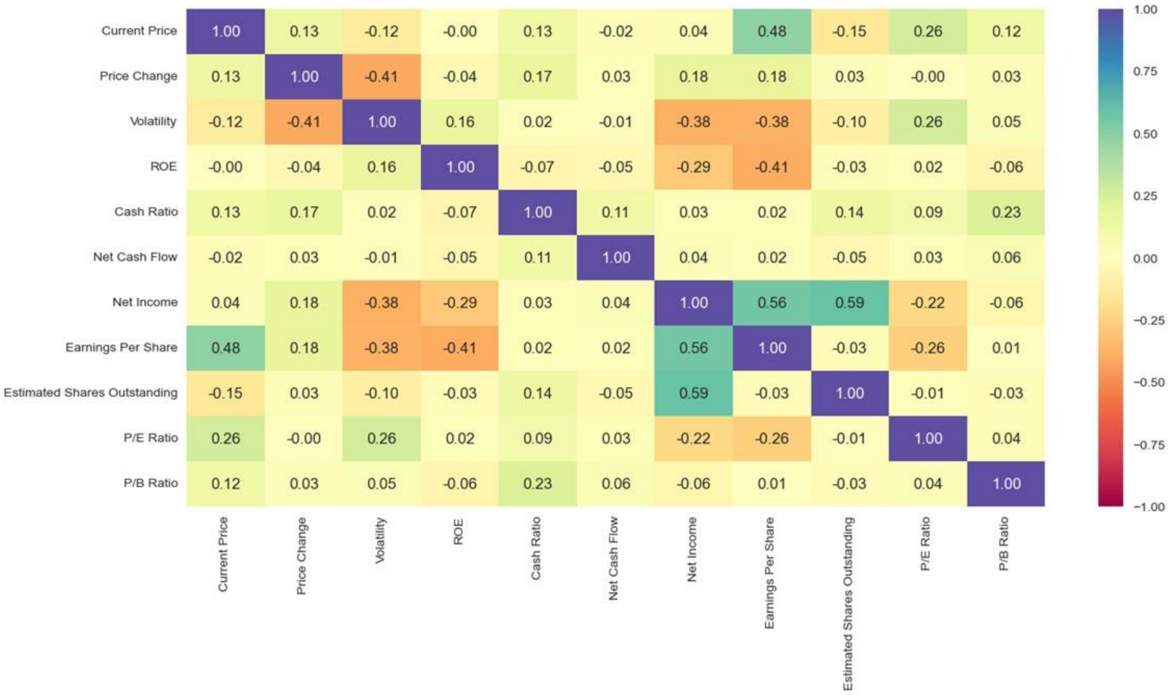
- Data shows some bell curve shape with few outliers on the left and right side.
- Data observed having negative value with the lowest around P/B ratio 150, and the highest is 160.

EDA Results – GICS Sector



- Top 3 contributor from GICS sector is:
  - Industrials with 15.6%
  - Financials with 14.4%
  - Health care and Consumer discretionary with the same percentage of 11.8%

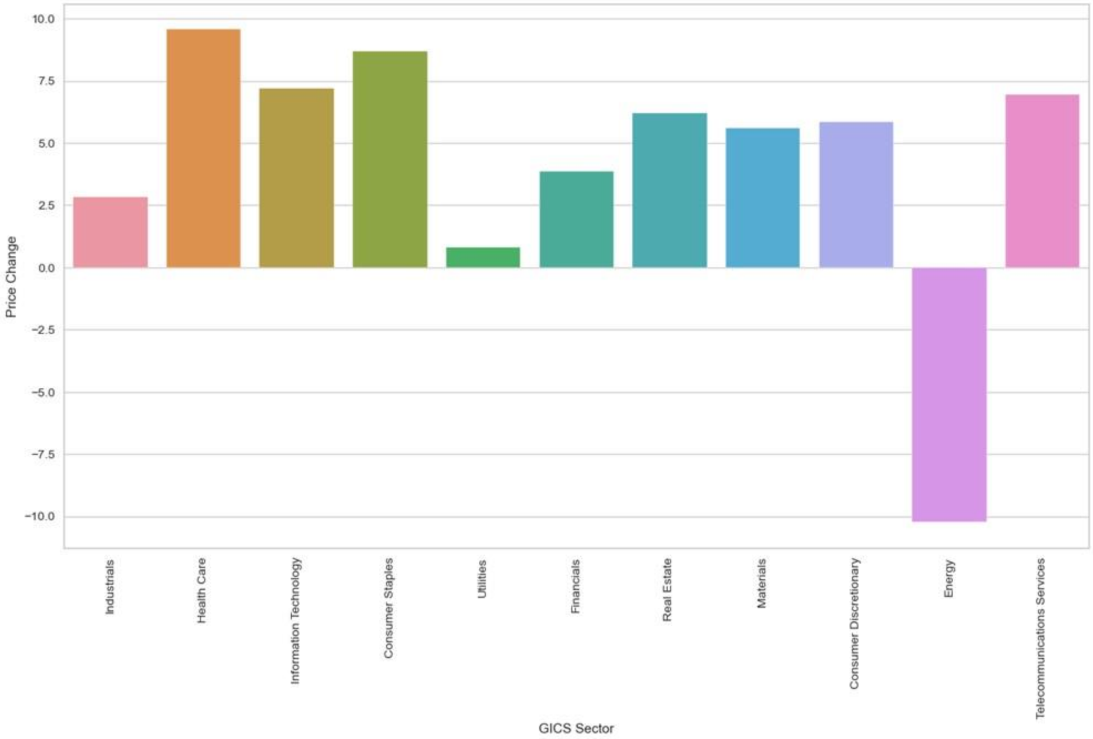
EDA Results – Variables Correlations



	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

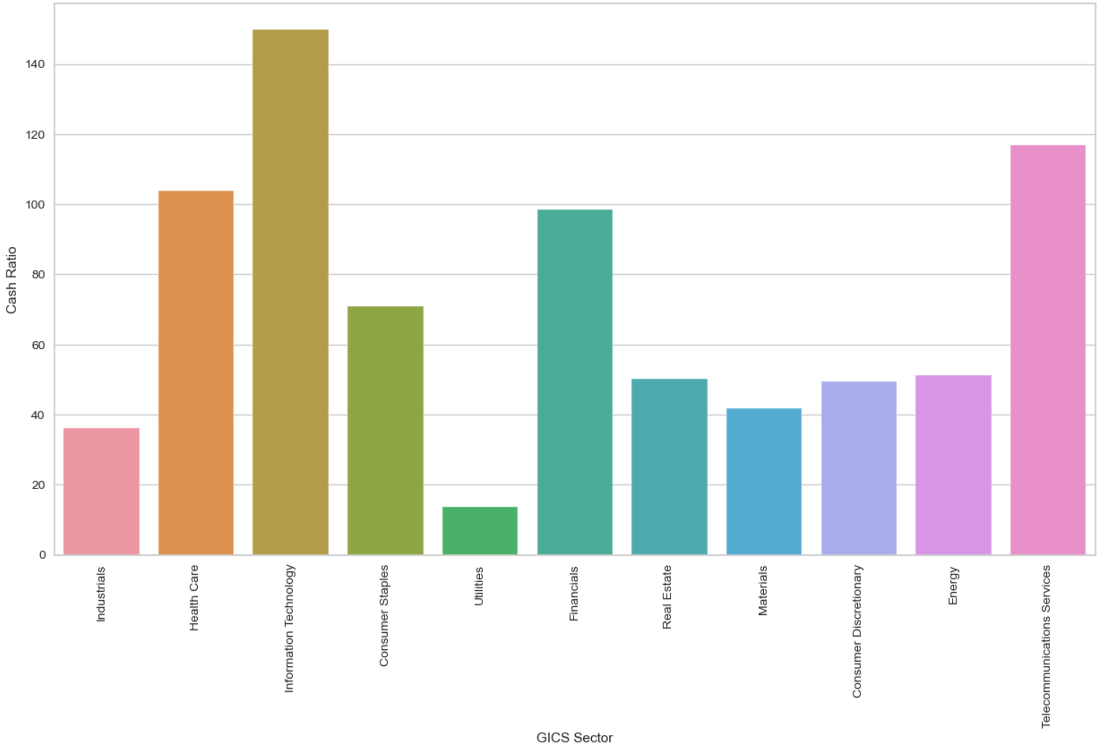
- Most parameter is observed having medium to low correlation, with few of them is having negative correlation.
- The medium correlation observed is net income, earnings per share and estimated shares outstanding.
- Among these three, net income was observed having negative correlation with volatility, where we could assume as the company generates higher income, its price most likely will be less volatile.

EDA Results – GISC Sector and Price Change



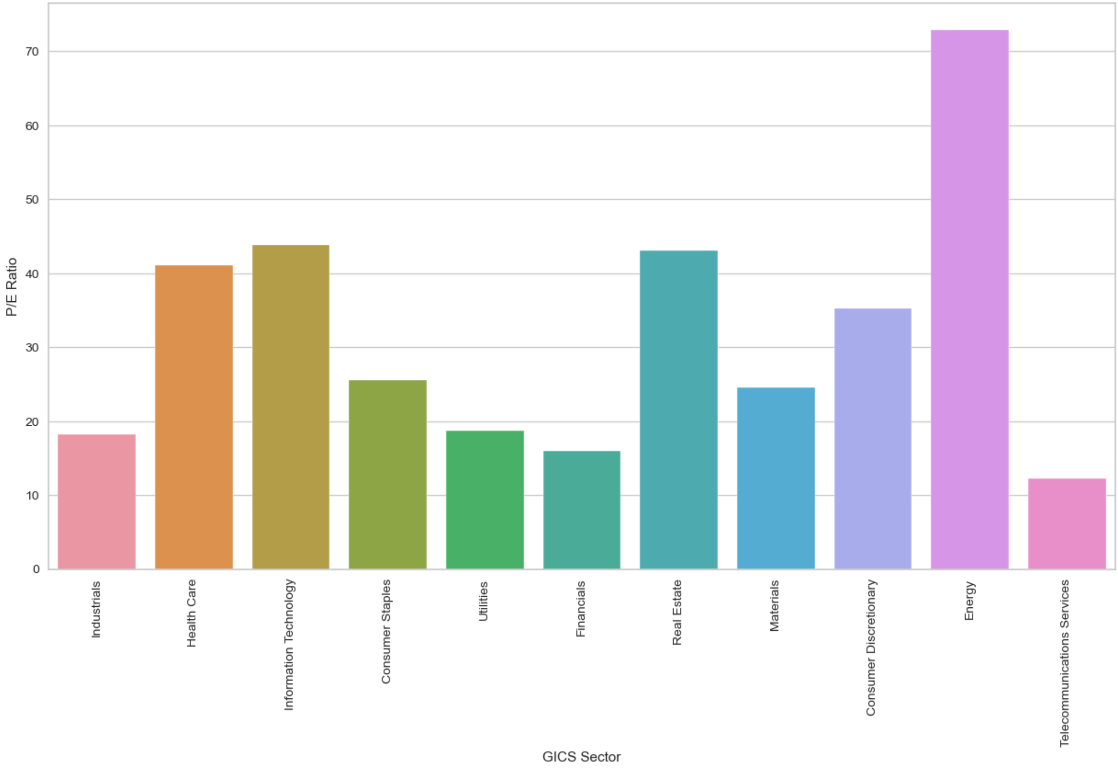
- Data shows Energy sector is having a negative of price change and it is the only sector shows a negative.
- The top 3 contributor for GISC sector is Health Care, Consumer Staples and information Technology.
- The bottom 3 contributor on the lowest price change is Utilities, Industrials and Financials.

EDA Results – GICS Sector and Cash Ratio



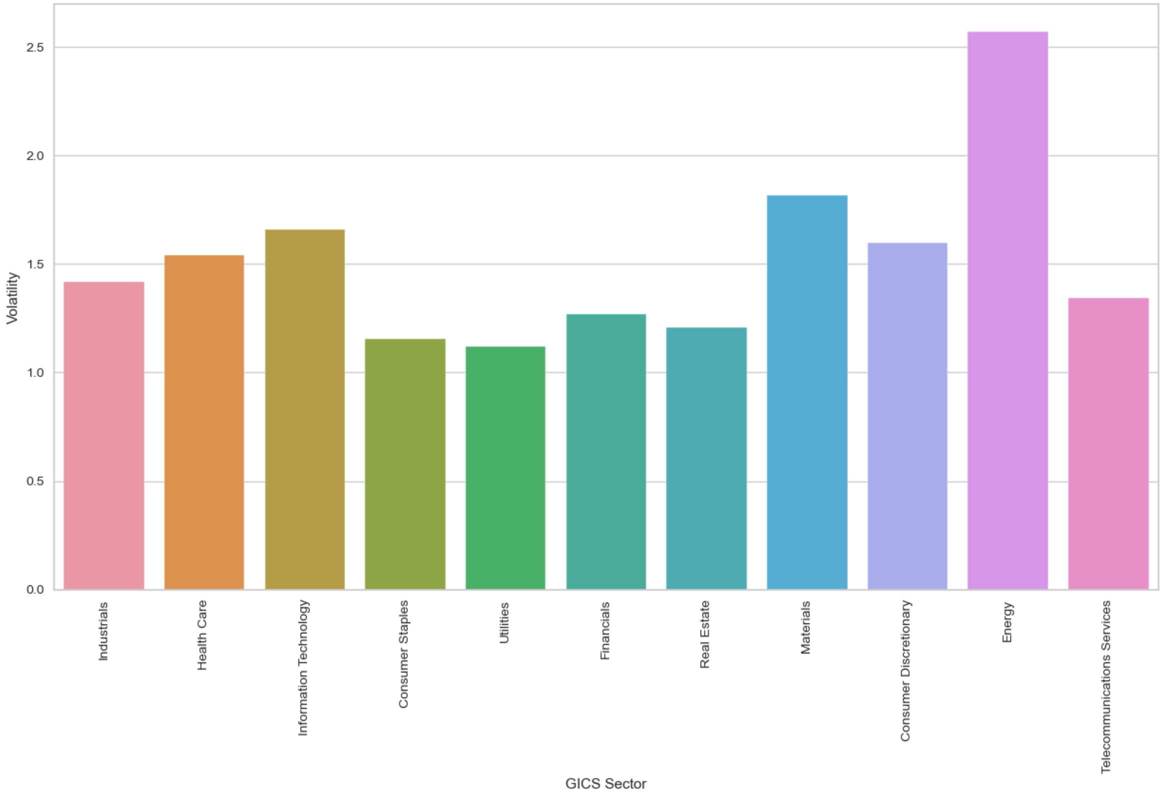
- The top 3 contributor for cash ratio is from Information Technology (150), Telecommunications Services (118) and Heltah Care (110).
- Utilities and Industrials is among the bottom contributor, which contribute about 80 and 58.

EDA Results – GICS Sector and P/E Ratio



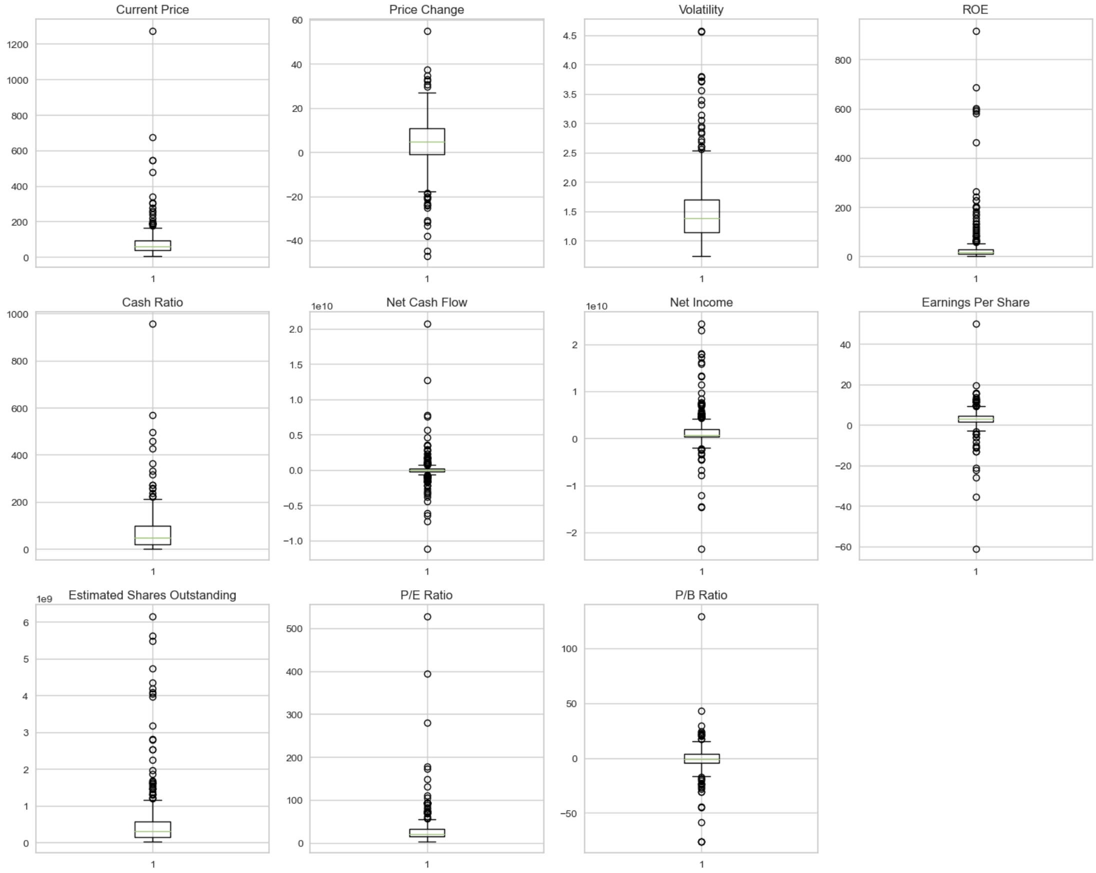
- Energy sector dominates strongly on P/E ratio, about 75 respectively, while Telecommunication Services contributes the final bottom, estimated about 11.
- Health Care, Information Technology and Real Estate is contribute the average on GICS sector.

EDA Results – GICS Sector and Volatility



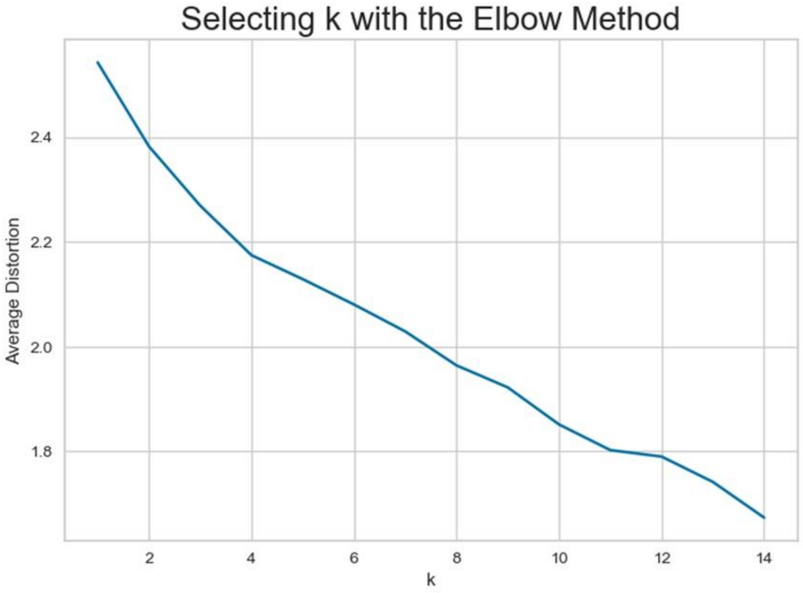
- Energy sector shows the most volatile sector, which contributes about 2.6
- Other sector shows almost similar and comparable volatility

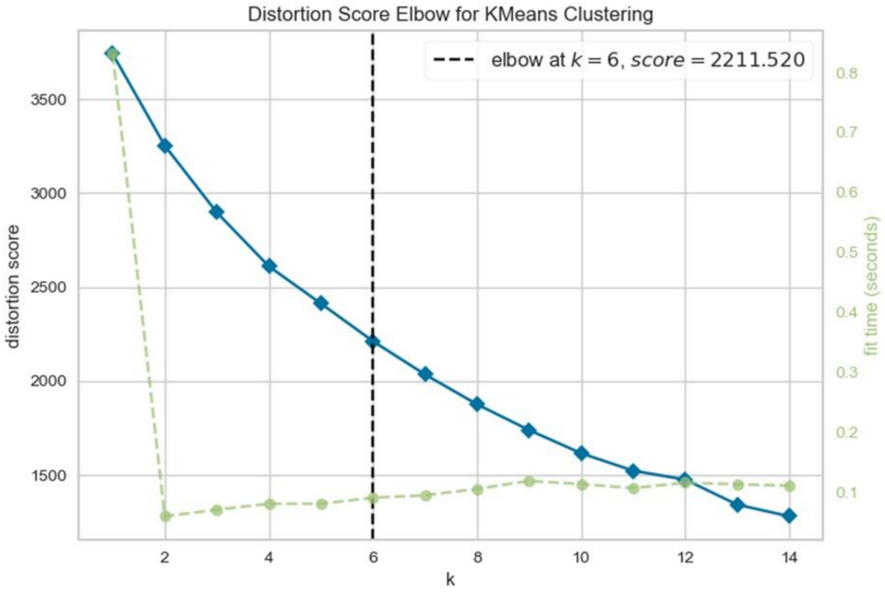
Data Preprocessing – Outlier Checking



- All data was found to have an outlier both left and right side.
- However, we will not impute this data as all of them is real data.
- We will proceed to perform cluster profiling.

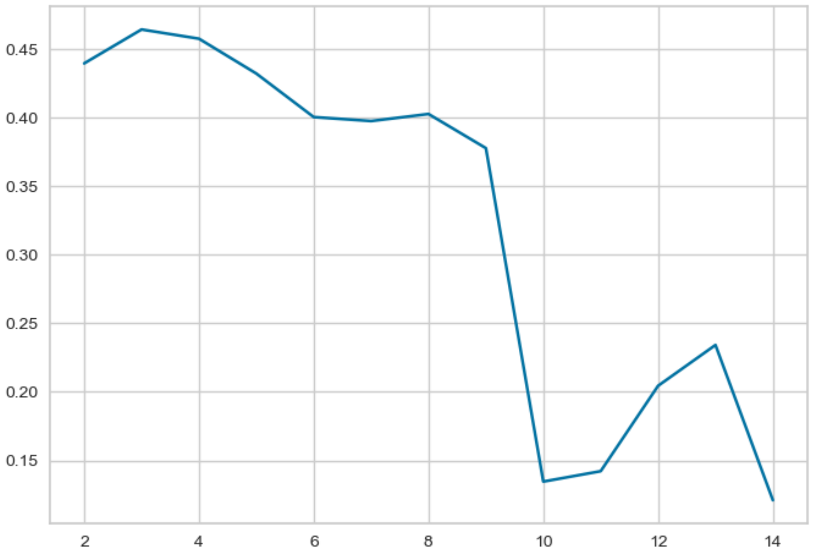
K-Means Clustering – Checking Elbow Plot

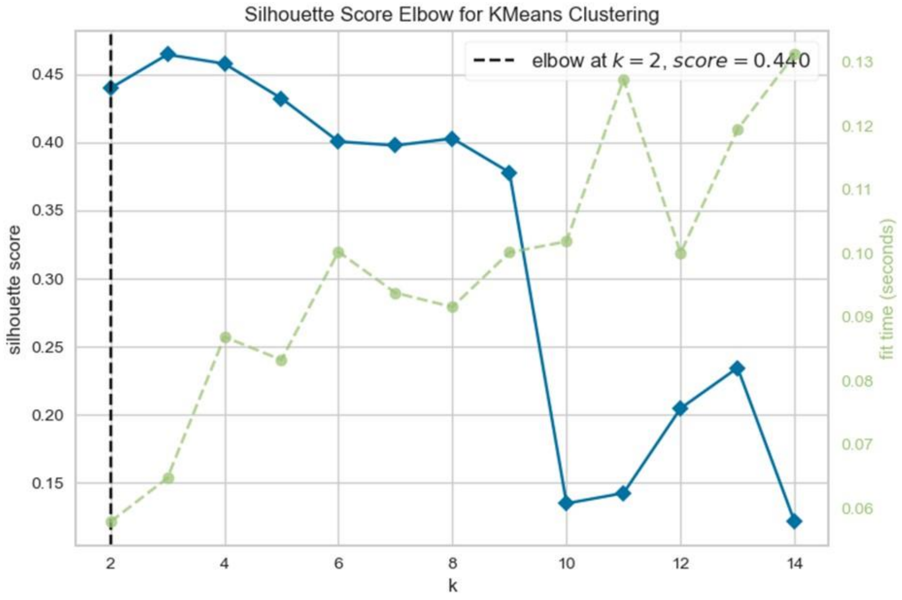




- From elbow plot, it is quite hard to find the k value as its plot keeps decreasing until max setting value.
- However, on the graph, k=4 may be a good value for us to start. We need to look for silhouette score for more clarification.

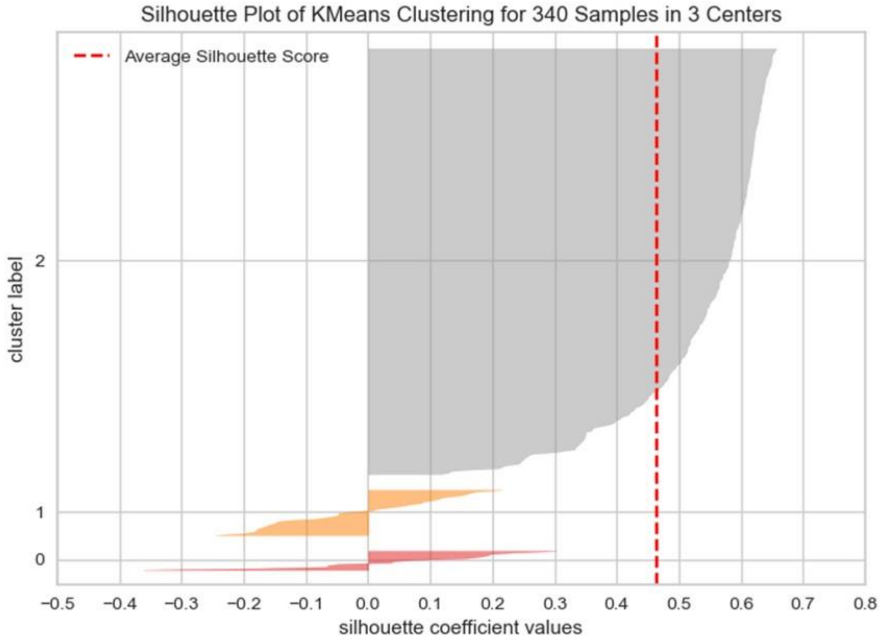
K-Means Clustering – Checking Silhouette Score





- Silhouette score provides much clear possible k value compared with elbow method.
- Based on the plot, k=3 is a good value for us to start a cluster profiling.

K-Means Clustering – Silhouette Plot of KMeans Clustering

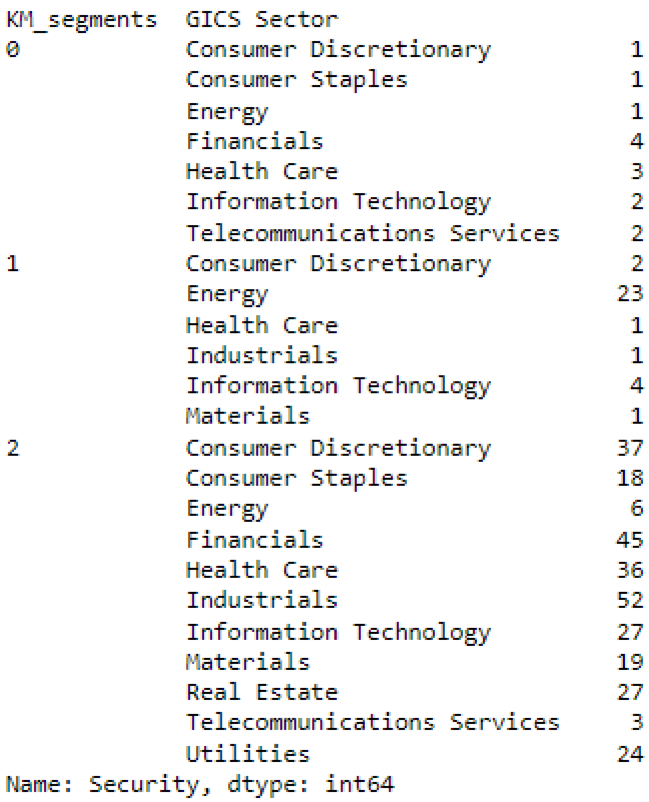


- The plot shows us the average silhouette score is about 0.48.
- Therefore, we will continue to try to perform cluster profiling.

K-Means Clustering Summary – Cluster Profiling

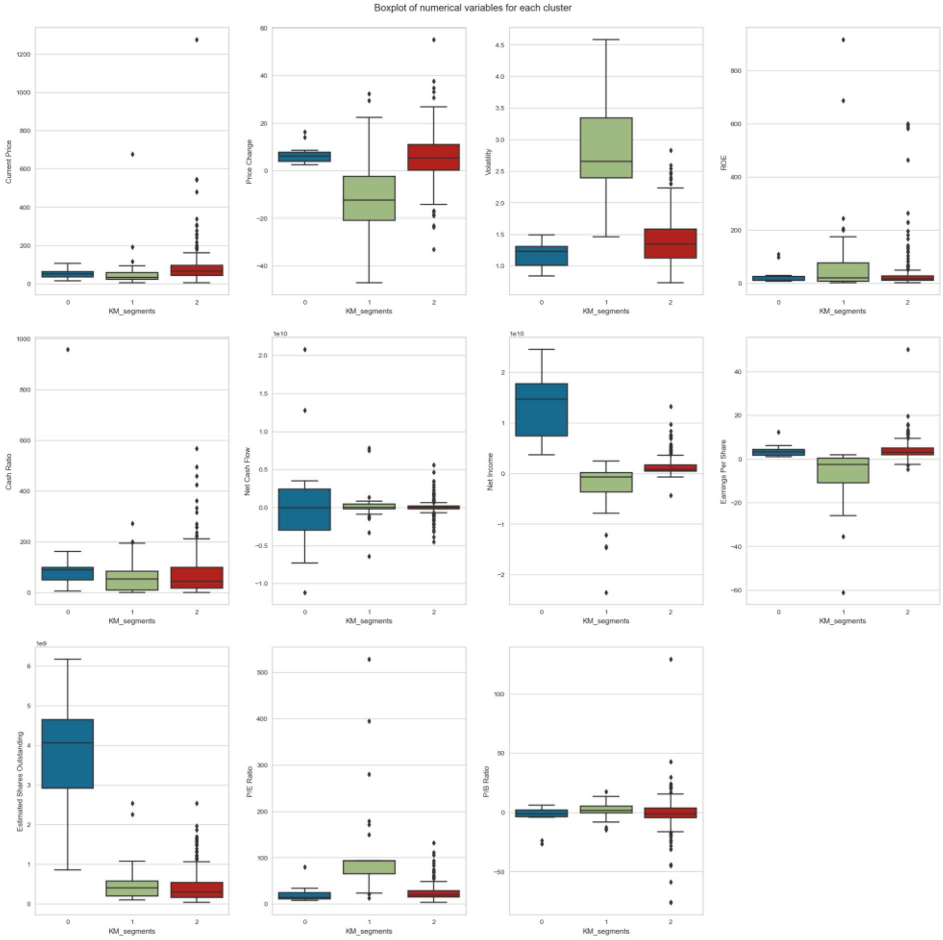
- Cluster profiling as below:

KM_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_each_segment
0	52.142857	6.779993	1.175153	26.142857	140.142857	760285714.285714	13368785714.285715	3.769286	3838879870.871428	20.654832	-3.529270	14
1	64.183438	-10.557046	2.797776	96.531250	70.718750	159171125.000000	-3250005968.750000	-7.886875	526459323.057500	111.333230	1.783445	32
2	84.045331	5.542488	1.404255	34.040816	66.608844	10698350.340136	1445333183.673469	3.890051	427206184.715408	24.613743	-2.013147	294



```
KM_segments  GICS Sector
0            Consumer Discretionary             1
             Consumer Staples                   1
             Energy                             1
             Financials                         4
             Health Care                        3
             Information Technology             2
             Telecommunications Services        2
1            Consumer Discretionary             2
             Energy                            23
             Health Care                        1
             Industrials                        1
             Information Technology             4
             Materials                          1
2            Consumer Discretionary            37
             Consumer Staples                  18
             Energy                             6
             Financials                        45
             Health Care                       36
             Industrials                       52
             Information Technology            27
             Materials                         19
             Real Estate                       27
             Telecommunications Services        3
             Utilities                         24
Name: Security, dtype: int64
```

K-Means Clustering Summary

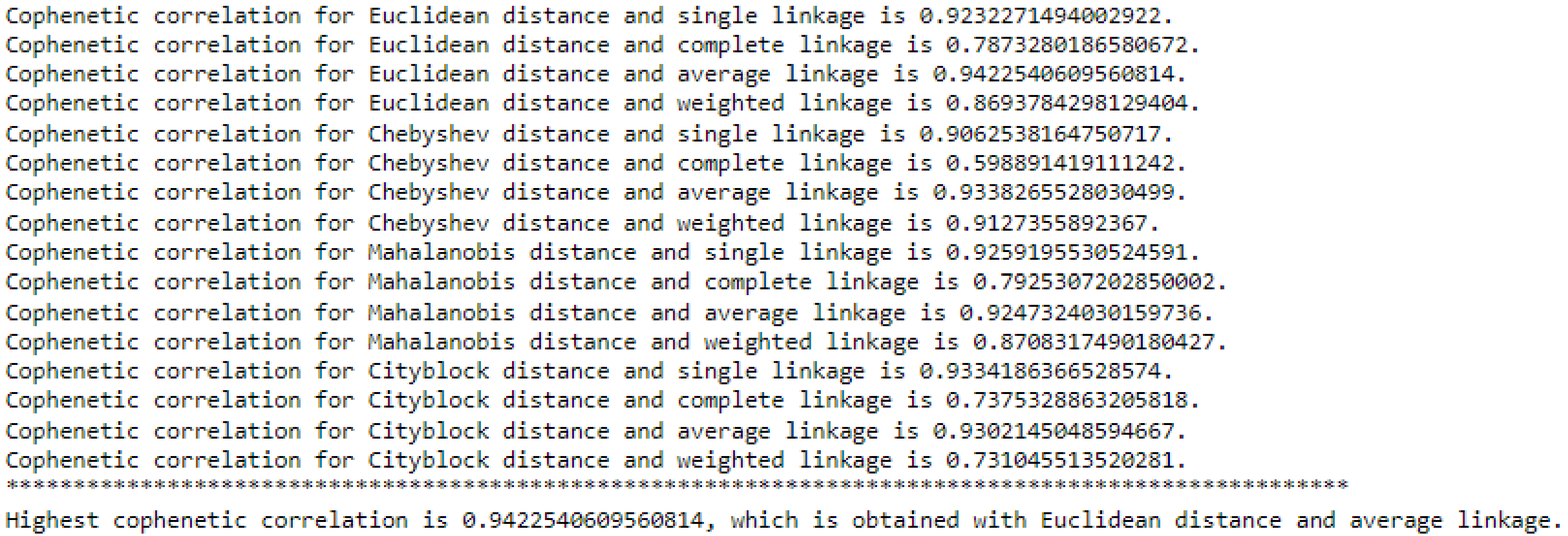


- Most KM segments are comparable even though we observe some outliers.
- However, the most distinct results on KM segments vs parameter involved are price change, volatility, net income and estimated shares outstanding.

K-Means Clustering Summary

- Optimal Number of clusters using K-Means is found to be k=3.
- Sector who is in cluster 0 has a large market capitalization, and most of them has a low volatility, highest net income, highest estimated shares outstanding and having wider range of net cash flow.
- Sector who is in cluster 1 has a wider range of price change and highest volatility. Their volatility is also wide. However, they are having consistent net cash flow and P/B ratio. They are also having highest P/E ratio.
- Sector who is in cluster 2 has a modest result on all parameters except price change. They are having wider range on price change and cluster 2 having the most outlier of P/E ratio.
- We can proceed for another method on hierarchical clustering .

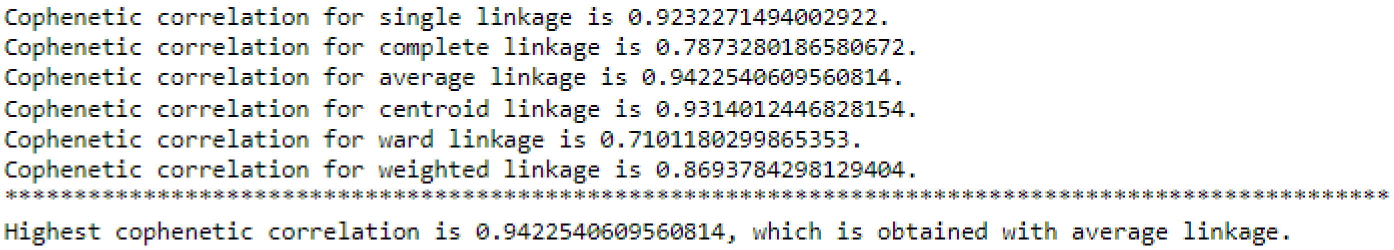
Hierarchical Clustering – Computing Cophenetic Correlation



```
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 0.9259195530524591.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.7925307202850002.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.9247324030159736.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.8708317490180427.
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.
```

- Highest cophenetic correlation using Hierarchical Clustering was observed from Euclidean distance and average linkage.
- Next step we will try to run on Euclidean distance using all linkage available and compare it.

Hierarchical Clustering – Euclidean Distance



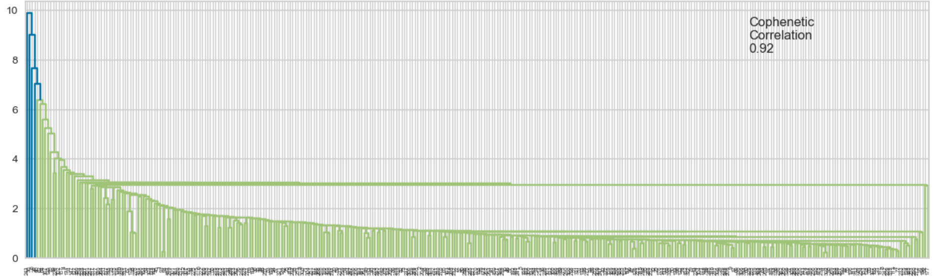
```
Cophenetic correlation for single linkage is 0.9232271494002922.
Cophenetic correlation for complete linkage is 0.7873280186580672.
Cophenetic correlation for average linkage is 0.9422540609560814.
Cophenetic correlation for centroid linkage is 0.9314012446828154.
Cophenetic correlation for ward linkage is 0.7101180299865353.
Cophenetic correlation for weighted linkage is 0.8693784298129404.
****************************************************************************************************
Highest cophenetic correlation is 0.9422540609560814, which is obtained with average linkage.
```

- Highest cophenetic correlation using Euclidean distance is on average linkage.

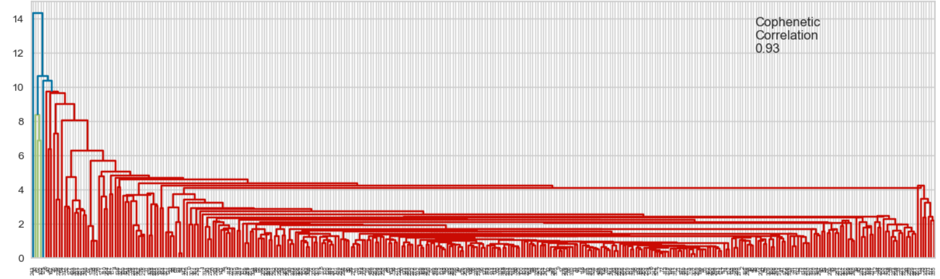
- Next step we will change a dendogram for Euclidean distance and compare it all available linkage.

Hierarchical Clustering – Checking Dendogram

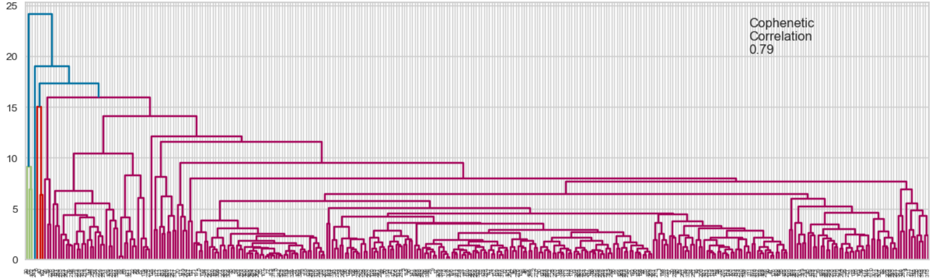
Dendrogram (Single Linkage)



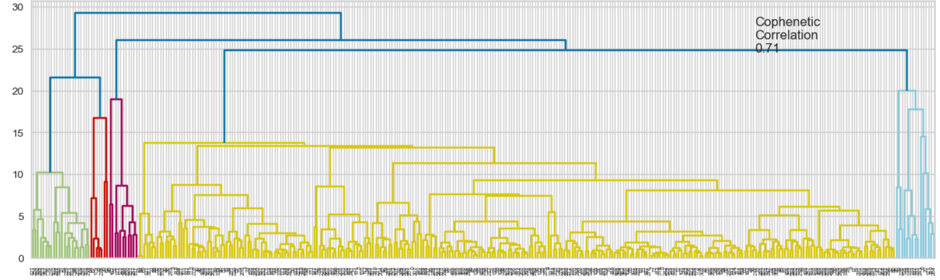
Dendrogram (Centroid Linkage)



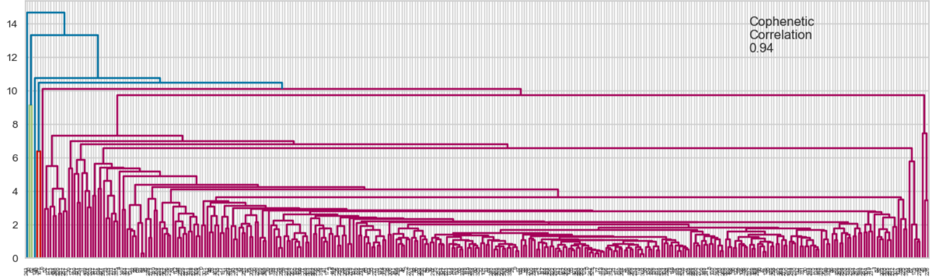
Dendrogram (Complete Linkage)



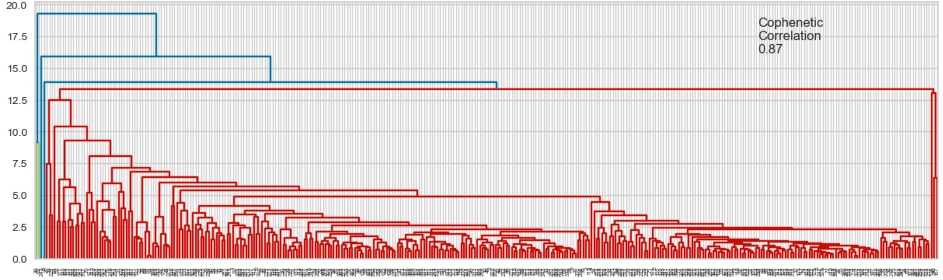
Dendrogram (Ward Linkage)



Dendrogram (Average Linkage)



Dendrogram (Weighted Linkage)



Hierarchical Clustering – Checking Dendogram

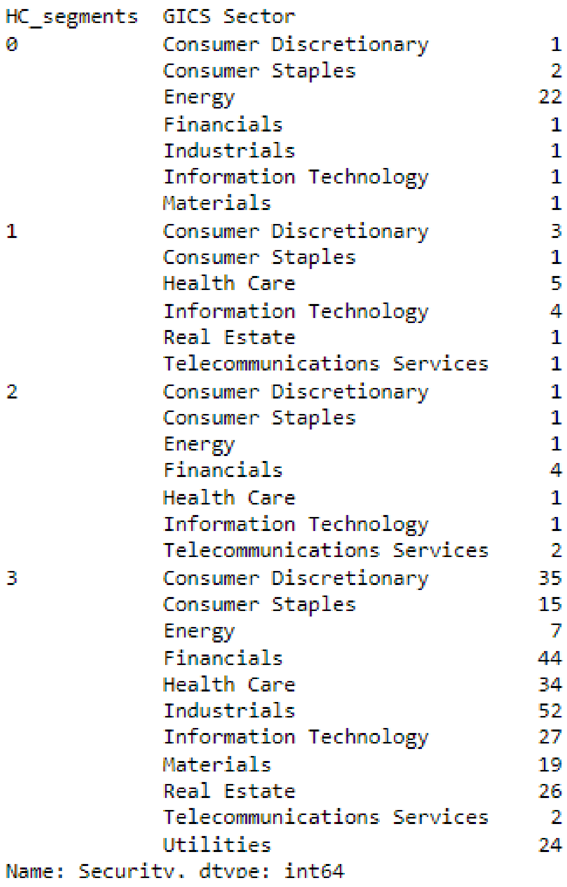
- From observation (previous slide), it is found that Ward linkage gives us the clearest linkage possible.
- K=4 looks possible on the Ward linkage.
- Even though its cophenetic coefficient is lowest among other tested linkage, it provides us the most clear dendogram.

	Linkage	Cophenetic Coefficient
4	ward	0.710118
1	complete	0.787328
5	weighted	0.869378
0	single	0.923227
3	centroid	0.931401
2	average	0.942254

Hierarchical Clustering – Cluster Profiling

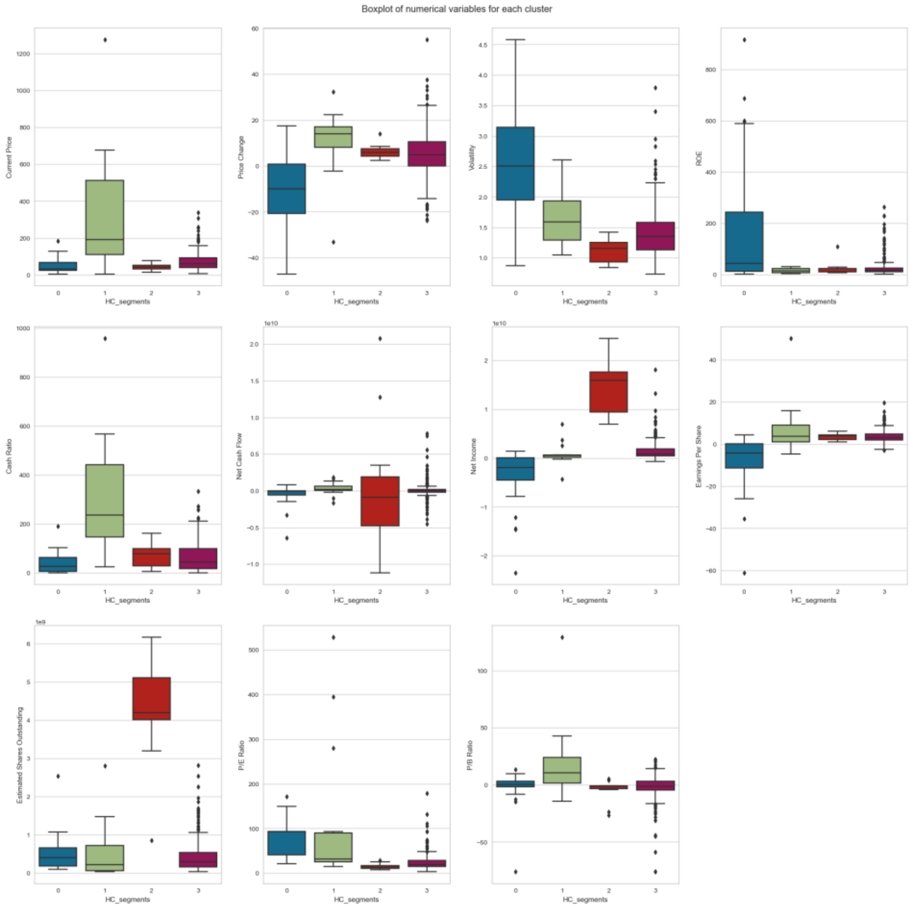
- Cluster Profiling

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_each_segment
0	48.006208	-11.263107	2.590247	196.551724	40.275862	-495901724.137931	-3597244655.172414	-8.689655	486319827.294483	75.110924	-2.162622	29
1	326.198218	10.563242	1.642560	14.400000	309.466667	288850666.666667	864498533.333333	7.785333	544900261.301333	113.095334	19.142151	15
2	42.848182	6.270446	1.123547	22.727273	71.454545	558636363.636364	14631272727.272728	3.410000	4242572567.290909	15.242169	-4.924615	11
3	72.760400	5.213307	1.427078	25.603509	60.392982	79951512.280702	1538594322.807018	3.655351	446472132.228456	24.722670	-2.647194	285



```
HC_segments  GICS Sector
0            Consumer Discretionary          1
             Consumer Staples                2
             Energy                         22
             Financials                      1
             Industrials                     1
             Information Technology          1
             Materials                       1
1            Consumer Discretionary          3
             Consumer Staples                1
             Health Care                     5
             Information Technology          4
             Real Estate                     1
             Telecommunications Services     1
2            Consumer Discretionary          1
             Consumer Staples                1
             Energy                          1
             Financials                      4
             Health Care                     1
             Information Technology          1
             Telecommunications Services     2
3            Consumer Discretionary         35
             Consumer Staples               15
             Energy                          7
             Financials                     44
             Health Care                    34
             Industrials                    52
             Information Technology         27
             Materials                      19
             Real Estate                    26
             Telecommunications Services     2
             Utilities                      24
Name: Security, dtype: int64
```

Hierarchical Clustering Summary



- Most KM segments having wider range and observed having some outliers.
- However, the most distinct results on KM segments vs parameter involved are price change, volatility, net income and estimated shares outstanding.

Summary

- Optimal Number of clusters using K-Means is found to be k=4.
- Sector who is in cluster 0 has a wider range of price change, highest volatility and wider ROE. They are also having the lowest earnings per share and net income, and also observed having an outlier towards negative value.
- Sector who is in cluster 1 has a highest and wider range of current price and cash ratio. However, their ROE, net cash flow and earnings per share is consistent and stable. They are also having highest P/B ratio
- Sector who is in cluster 2 has the most highest net income and estimated shares outstanding compared to other clusters. They are having consistent price change, ROE, earnings per share and P/E ratio. Moreover, their volatility and cash ratio is quite stable.
- Sector who is in cluster 3 is having the most outliers compared to other clusters. Most possible the sector in cluster 3 is not stable and having a business up-down more compared with other clusters.

Summary

- K means vs Hierarchical Clustering:
  - Both technique is able to fit on the dataset very well.
  - Both technique able to give similar cluster group after profiling.
  - Hierarchical clustering able to give more distinct value with good observations in each parameter.

Great Learning

Happy Learning !