TRADE&AHEAD PROJECT

NEHA BHARTIYA

MAY 2022

AGENDA

Background & Objective

Data Insights

Exploratory Data Analysis

Data Preprocessing

K-Means Clustering

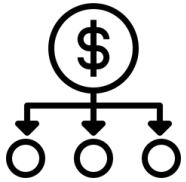
Hierarchical Clustering

K-Means vs Hierarchical Clustering

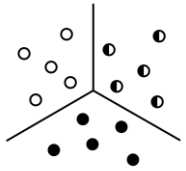
Actionable Insights & Recommendations

BACKGROUND & OBJECTIVE

The stock market has consistently proven to be a good place to invest in and save for the future. There are a lot of compelling reasons to invest in stocks. It can help in fighting inflation, create wealth, and also provides some tax benefits. Good steady returns on investments over a long period of time can also grow a lot more than seems possible. Also, thanks to the power of compound interest, the earlier one starts investing, the larger the corpus one can have for retirement. Overall, investing in stocks can help meet life's financial aspirations.



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.



By doing a cluster analysis, one can identify stocks that exhibit similar characteristics and ones that 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.

The objective is to analyze the given data containing financial indicators for a few companies listed under New York Stock Exchange, group the stocks based on the attributes provided, and share insights about the characteristics of each group.

DATA INSIGHTS

The data contains details of 340 companies having 15 attributes/features.

Variable	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
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)

Observations(rows)	Variables(columns)
340	15

Other notable points about the data:

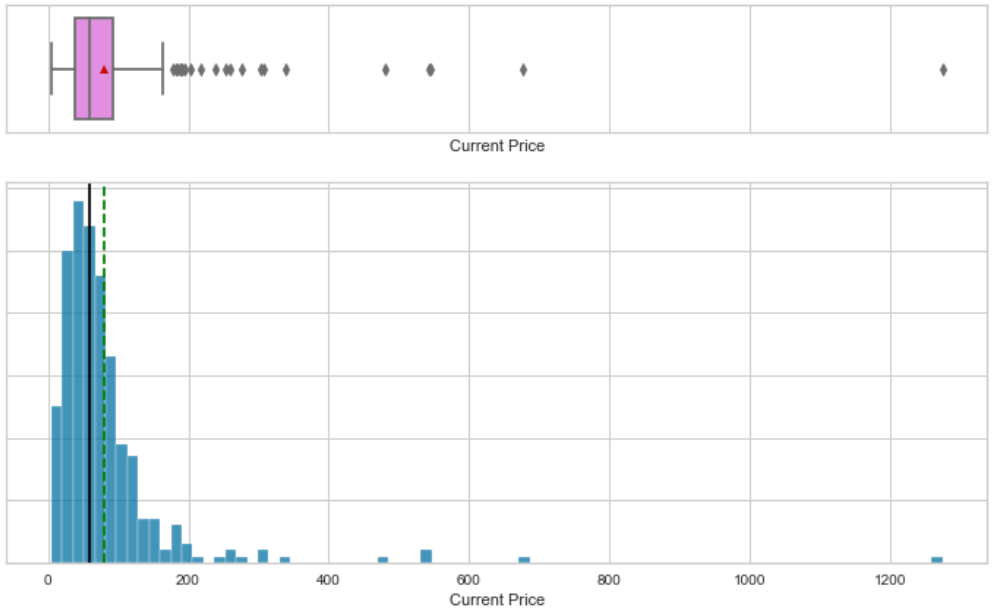
- There are no missing values in any of the columns.
- Ticker Symbol, Security, GICS Sector, GICS Sub Industry are object type, Current Price, Price Change, Volatility, Earnings per share, Estimated share outstanding, P/E ratio, P/B ratio are float type, ROE, Cash Ratio, Net Cash Flow, Net Income are integer type variables.
- Ticker Symbol, Security are unique.
- GICS Sector, GICS Sub Industry are categorical variables and can take fixed number of possible values.

EXPLORATORY DATA ANALYSIS (EDA) – 1/10
CURRENT PRICE, PRICE CHANGE

Current Price

Observations:

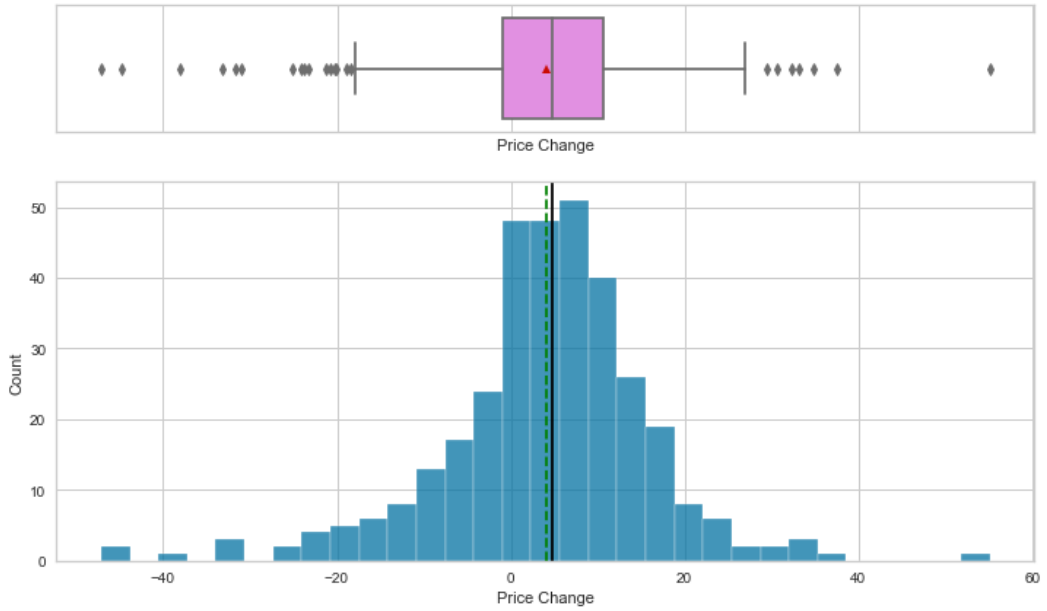
- The distribution is right skewed as mean > median.
- The minimum value is 4.5 and the maximum value is 1274.94.
- There are outliers in the variable.



Price Change

Observations:

- The distribution is near symmetrical.
- There are outliers in the variable.
- The minimum is -47.12 and the maximum is 55.05.

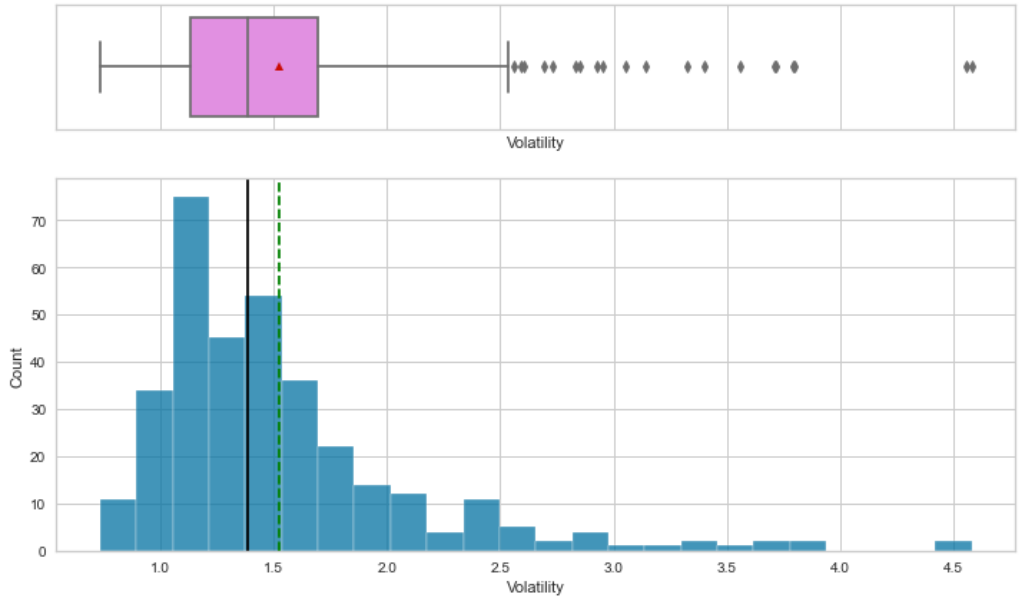


EXPLORATORY DATA ANALYSIS (EDA) – 2/10
VOLATILITY, ROE

Volatility

Observations:

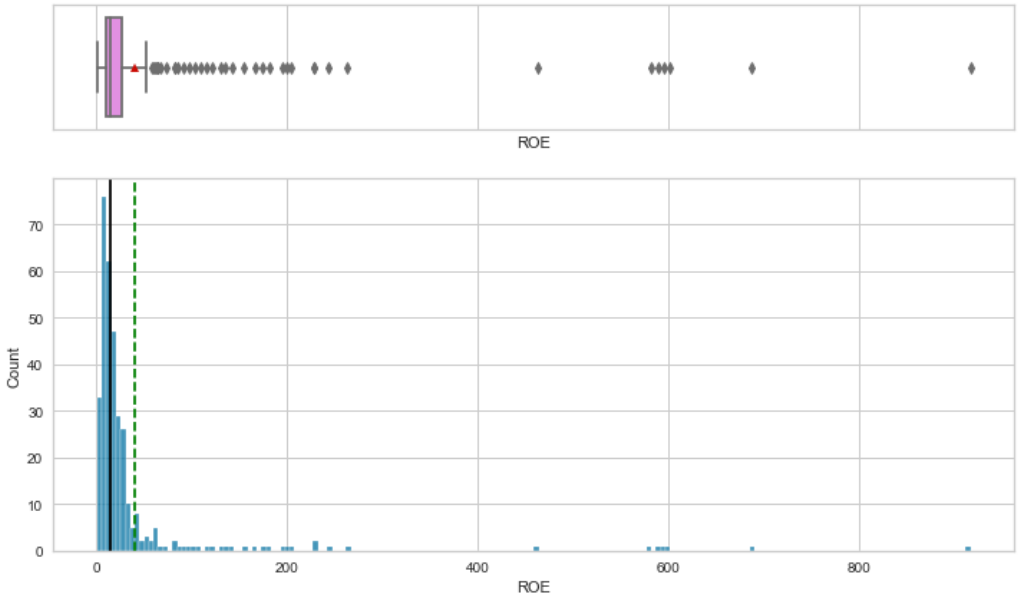
- The distribution is right skewed as mean > median.
- Upper 75% of the values vary from 1.69 to 2.6.
- The maximum value is 4.58.
- There are outliers in the variable.



ROE

Observations:

- The distribution is right skewed as mean > median.
- The minimum value is 1 and the maximum value is 917.
- There are outliers in the variable.

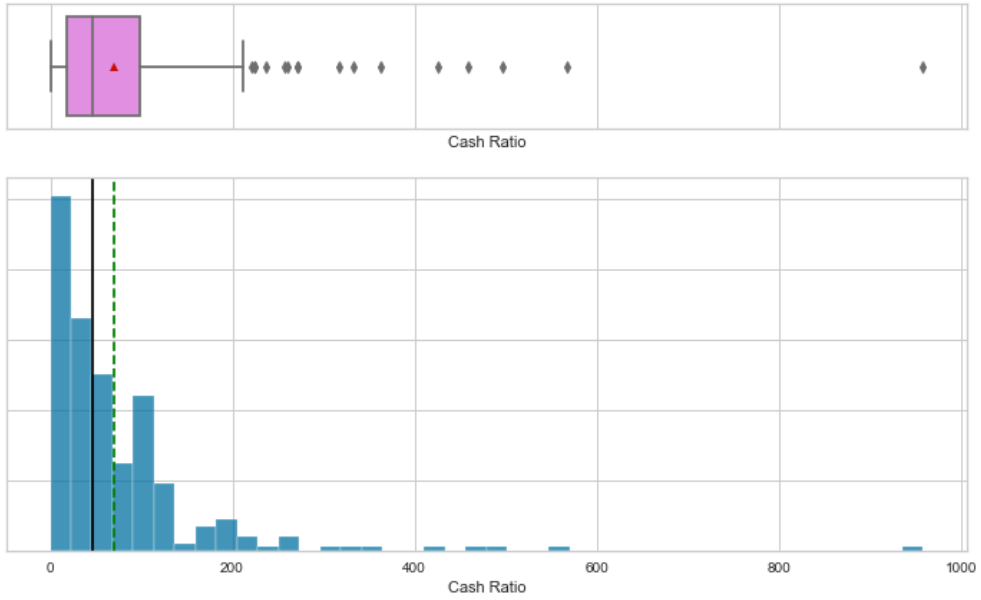


EXPLORATORY DATA ANALYSIS (EDA) – 3/10
CASH RATIO, NET CASH FLOW

Cash Ratio

Observations:

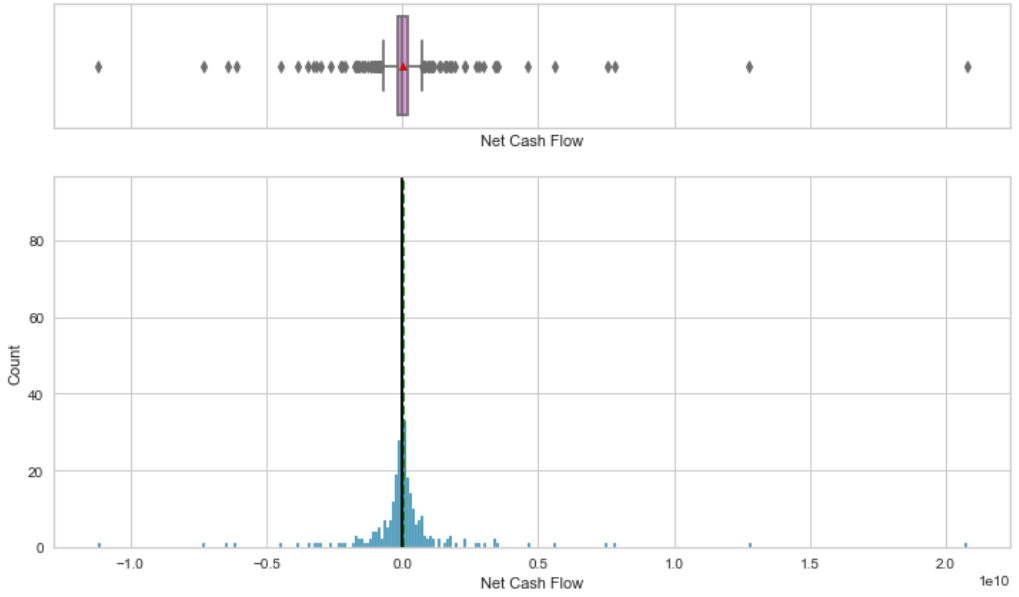
- The distribution is right skewed as mean > median.
- Upper 75% of the values vary from 99 to 220.
- The maximum value is 958.
- There are outliers in the variable.



Net Cash Flow

Observations:

- The distribution is symmetrical.
- There are outliers in the variable.
- The minimum is -11208000000.0 and the maximum is 20764000000.0.

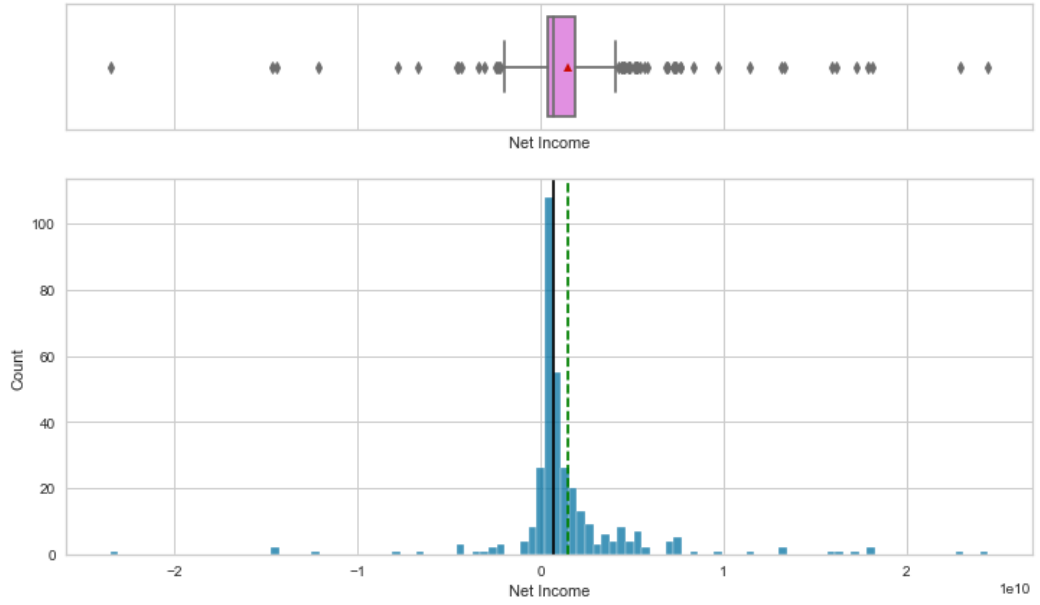


EXPLORATORY DATA ANALYSIS (EDA) – 4/10
NET INCOME, EARNINGS PER SHARE

Net Income

Observations:

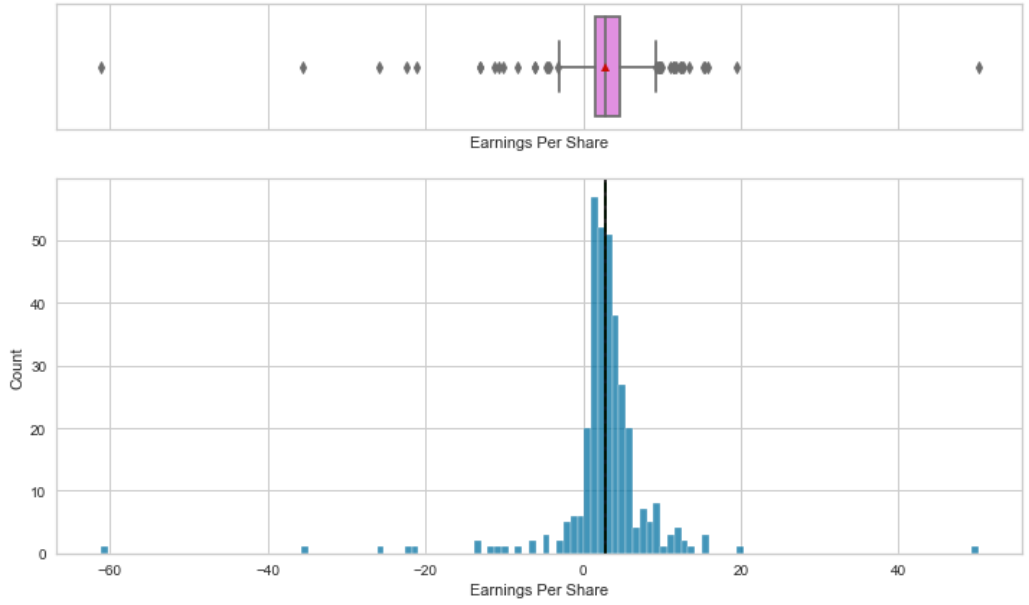
- The distribution is right skewed as mean > median.
- The maximum value is 24442000000.0.
- There are outliers in the variable.



Earnings Per Share

Observations:

- The distribution is symmetrical.
- There are outliers in the variable.
- The minimum is -61.2 and the maximum is 50.90.

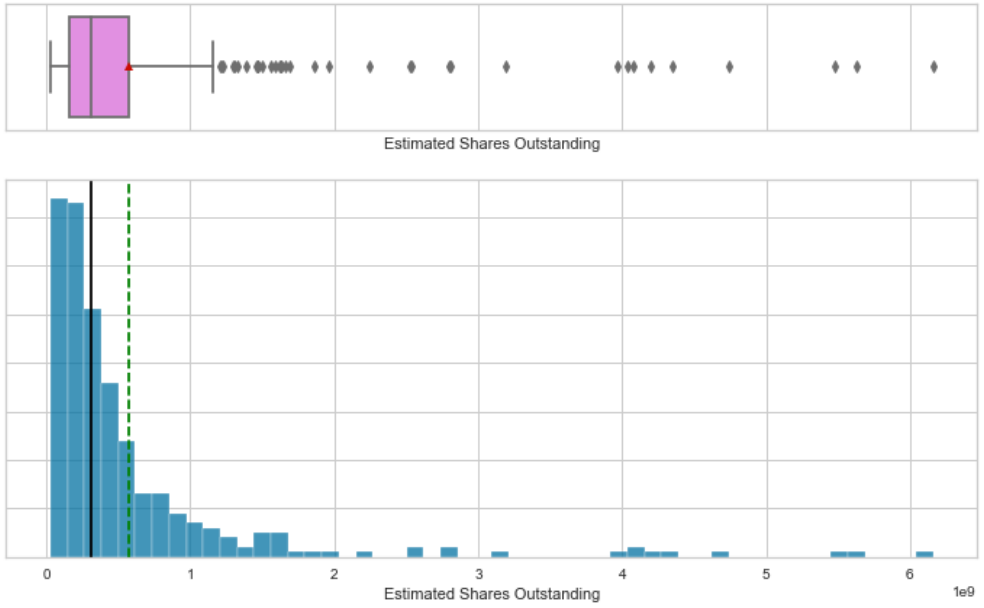


EXPLORATORY DATA ANALYSIS (EDA) – 5/10
ESTIMATED SHARE OUTSTANDING, P/E RATIO

Estimated Share Outstanding

Observations:

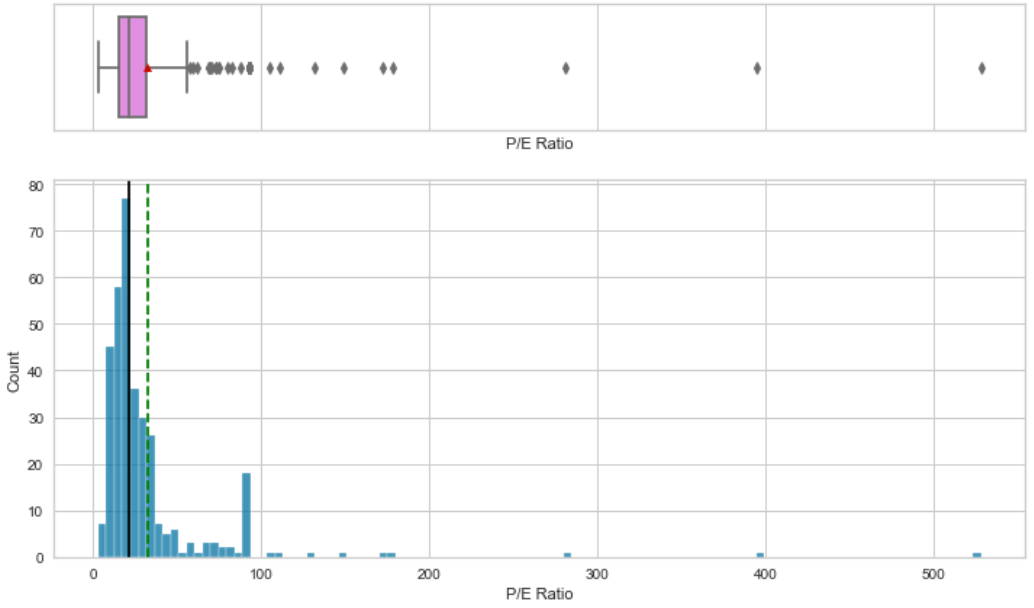
- The distribution is right skewed as mean > median.
- The minimum value is 27672156.86 and the maximum value is 6159292035.0.
- There are outliers in the variable.



P/E Ratio

Observations:

- The distribution is right skewed as mean > median.
- The minimum value is 2.93 and the maximum value is 528.03.
- There are outliers in the variable.

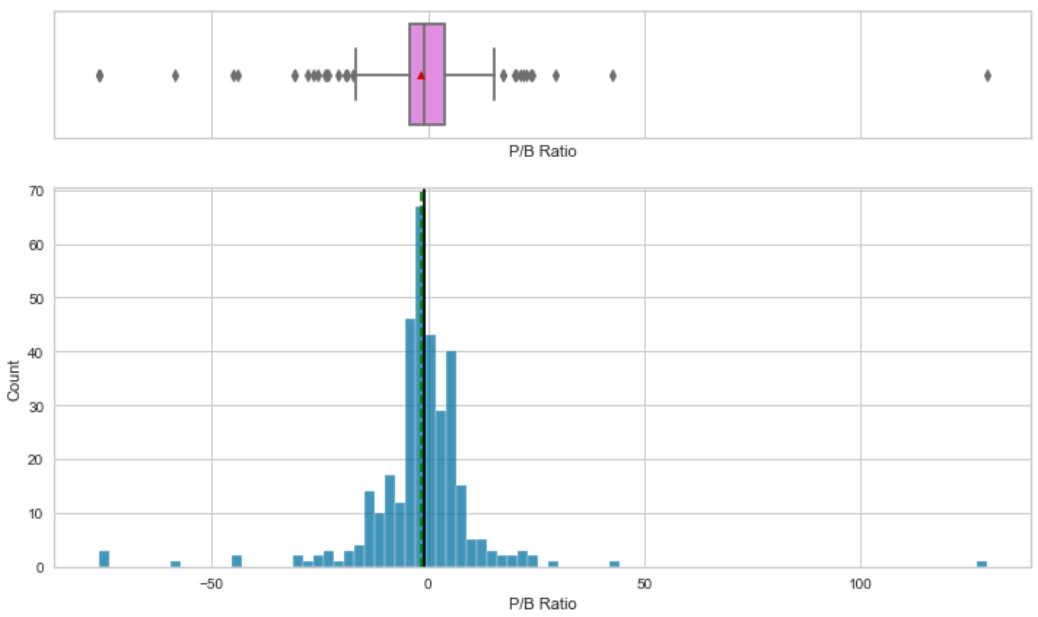


EXPLORATORY DATA ANALYSIS (EDA) – 6/10
P/B RATIO, GICS SECTOR

P/B Ratio

Observations:

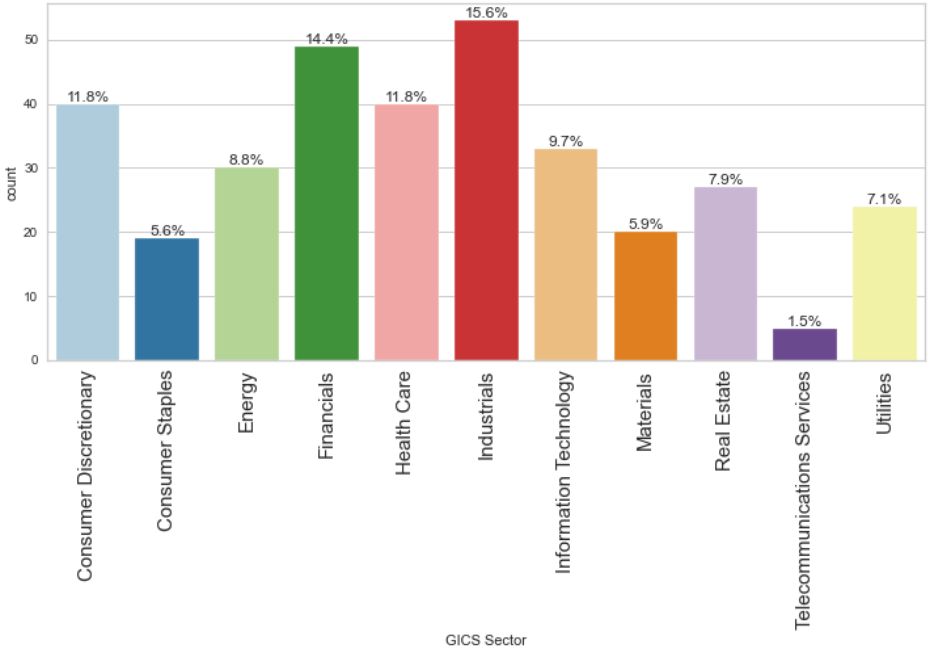
- The distribution is symmetrical.
- The minimum is 76.11 and the maximum value is 129.06.
- There are outliers in the variable.



GICS Sector

Observations:

- 15.6% of the companies are from the industrial sector followed by Financial with 14.4% and consumer Discretionary, health care with 11.8%.
- The lowest is from telecommunications sector with 1.5%.

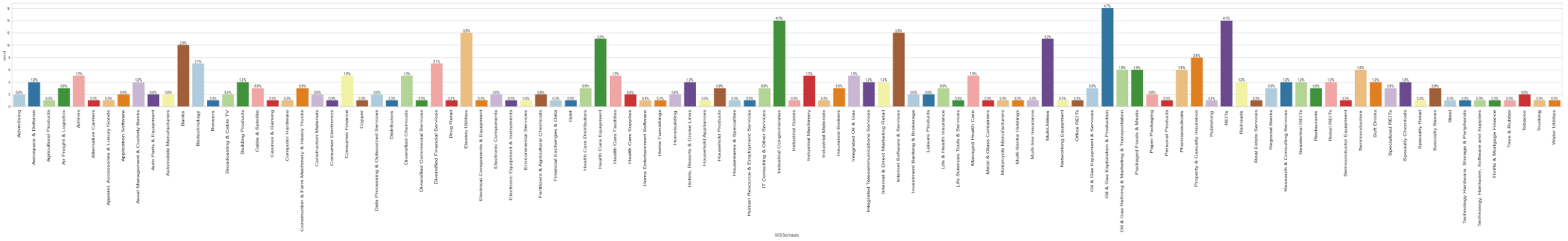


EXPLORATORY DATA ANALYSIS (EDA) – 7/10
GICS SUB INDUSTRY

GICS Sub Industry

Observations:

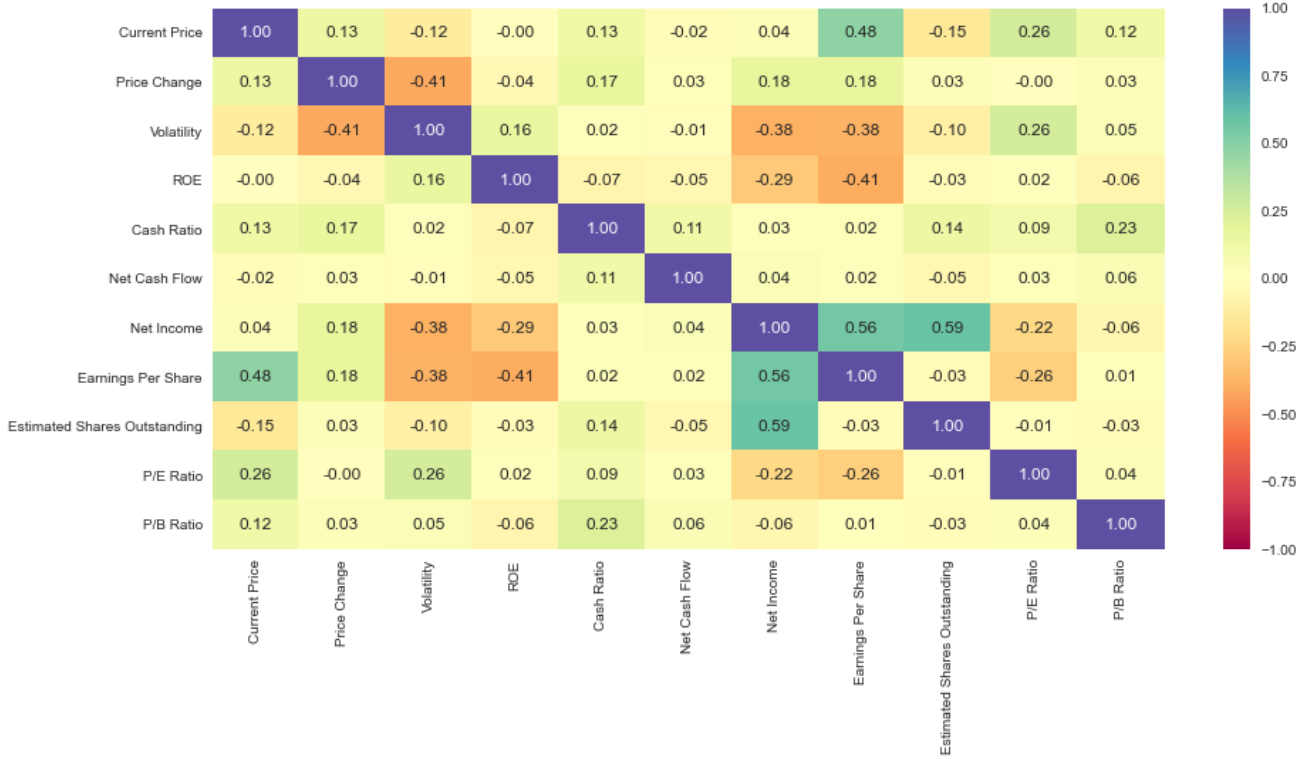
- 4.7% of the companies are from the Oil & gas Exploration & Production sector followed by Industrial Conglomerates with 4.1% and REITs with 4.1%.
- The lowest number of companies are from various segments with 0.3% each.



EXPLORATORY DATA ANALYSIS (EDA) – 8/10
BIVARIATE ANALYSIS OF NUMERICAL VARIABLES

Observations:

- The variables with positive correlation tend to move in similar direction and the variables with negative correlation tend to move in opposite directions.
- There is a positive correlation between Earnings per share and the current price.
- There is good positive correlation between earnings per share and Net Income and Estimated Shares outstanding and Net Income.
- There is negative correlation between Volatility and Price Change.
- There is negative correlation between Earnings per share and ROE.
- There is negative correlation between Net income and volatility also between earnings per share and volatility.

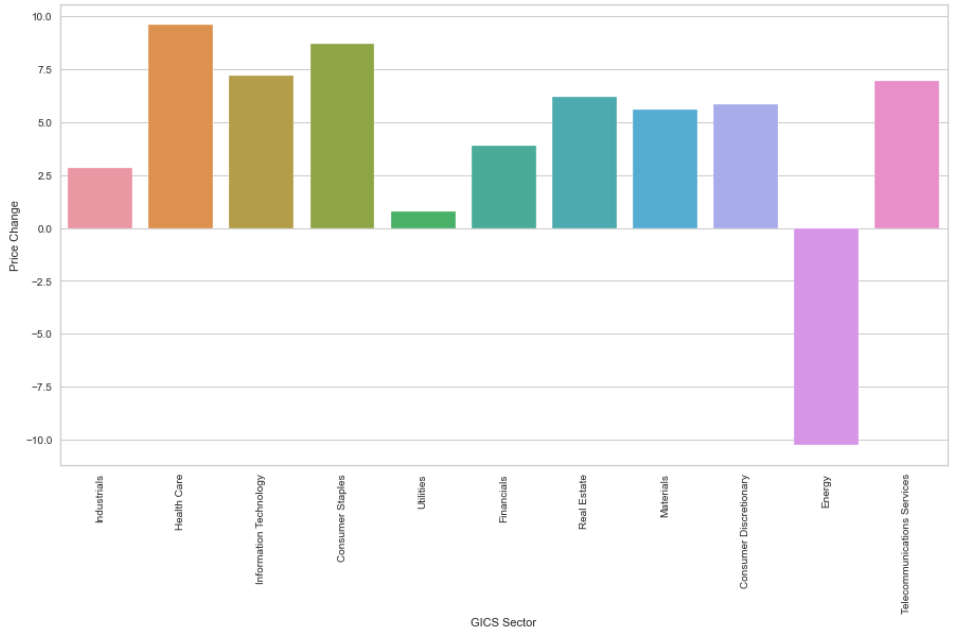


EXPLORATORY DATA ANALYSIS (EDA) – 9/10
PRICE CHANGE, CASH RATIO ACROSS GICS SECTOR

Price Change across GICS Sector

Observations:

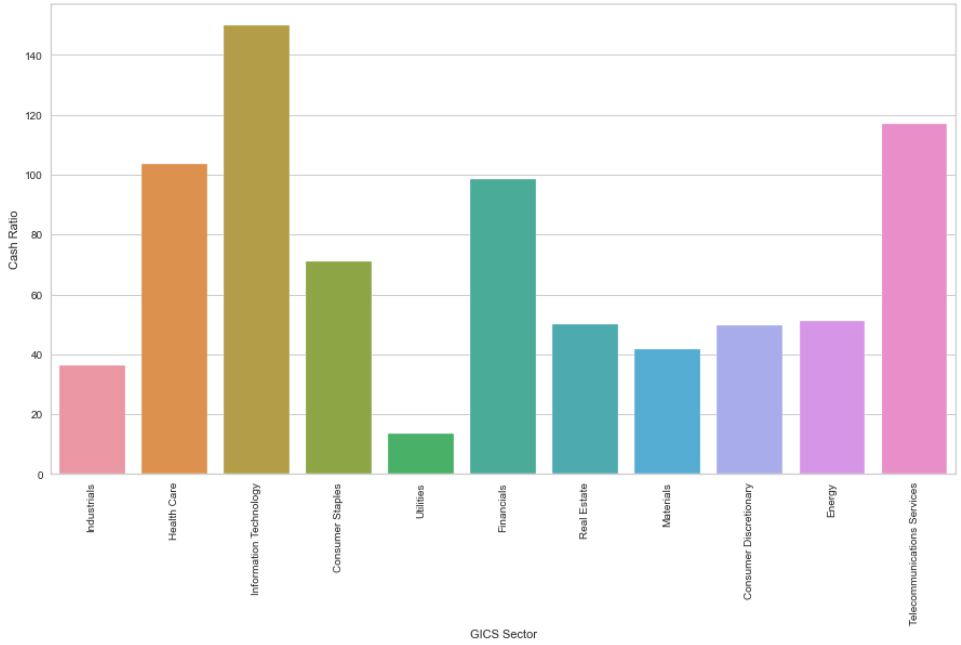
- Average maximum price increase is in the stocks of health care sector followed by consumer staples and then Information technology.
- There is a decline in the stock prices of Energy sector.



Cash Ratio across GICS Sector

Observations:

- Cash ratio provides a measure of a company's ability to cover its short-term obligations using only cash and cash equivalents.
- Maximum average cash ratio is in Information technology sector followed by Telecommunications services then by financials.

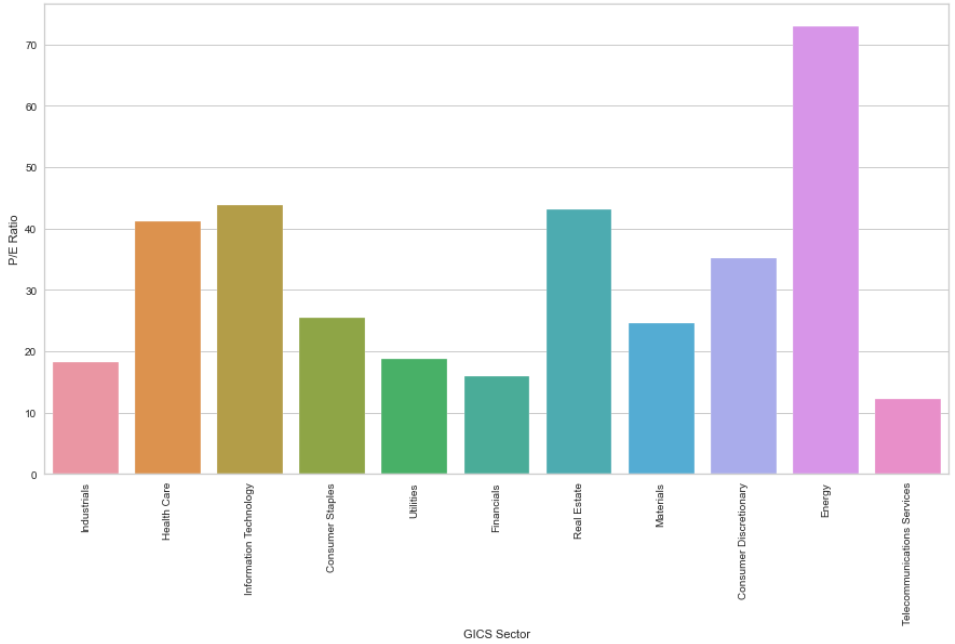


EXPLORATORY DATA ANALYSIS (EDA) – 10/10
P/E RATIO, VOLATILITY ACROSS GICS SECTOR

P/E Ratio across GICS Sector

Observations:

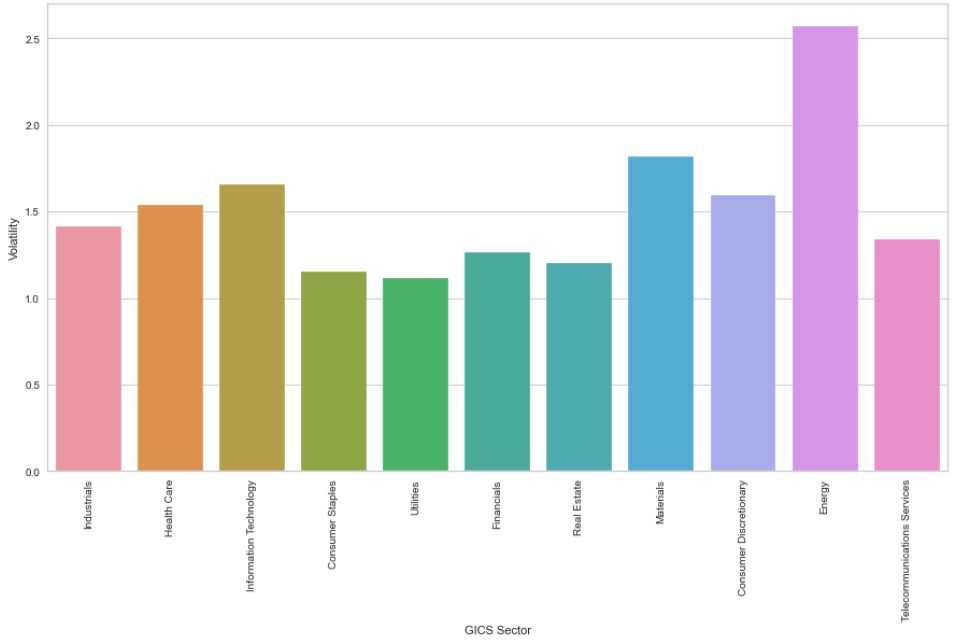
- P/E ratios can help determine the relative value of a company's shares as they signify the amount of money an investor is willing to invest in a single share of a company per dollar of its earnings.
- P/E ration is maximum for Energy sector followed by Information Technology sector.



Volatility across GICS Sector

Observations:

- Volatility accounts for the fluctuation in the stock price. A stock with high volatility will witness sharper price changes, making it a riskier investment.
- Volatility is maximum in Energy sector followed by Materials then Information Technology sector.



DATA PREPROCESSING

DUPLICATE VALUE CHECK, MISSING VALUE TREATMENT, OUTLIER DETECTION, FEATURE ENGINEERING

Observations:

- There are no duplicate values in the data set.
- There are no missing values.
- The columns do have outliers in them.
- Despite this, it is not acceptable to drop the observations as they are legitimate observations. This has been verified by the google search on the outlier companies in each category.

  eg: The further most outlier in “Current Price” was above 1200 which was found to be Priceline.com Inc. Upon google search it has been verified that the value is legit and can not be dropped.

  The further most outlier in “ROE” was above 800 which was found to be Apache Corporation. Upon google search it has been verified that the value is legit and can not be dropped.

  Similarly, the other outliers have been checked and verified.

Data preparation:

- It is observed that the attributes Net Cash Flow, Net Income, Estimated Shares Outstanding have very large scales as compared to the other attributes. In order to avoid the range of values in each feature from acting as a weight when determining how clusters are formed, we will standardize the features.
- Standardization prevents variables with larger scales from dominating how clusters are defined.
- The standardization has been done by using the Standardscaler function and is ready to use.

K-MEANS CLUSTERING (1/5)

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid. Steps for K-means Clustering are as below:

1. Choose the number of clusters, k
2. Select k random points from the data as centroids
3. Assign all the points to the closest cluster centroid
4. Recompute the centroids of newly formed clusters
5. Repeat steps 3 and 4

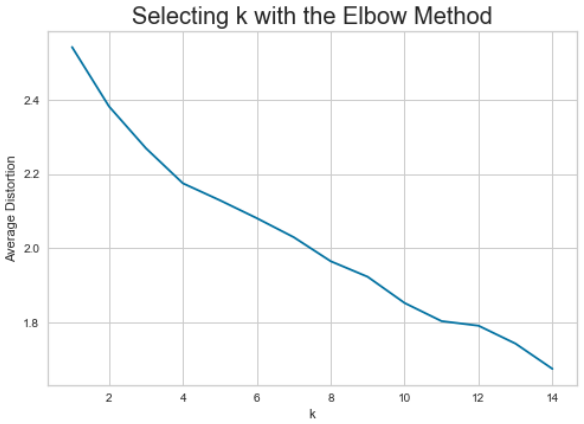
Stopping Criteria for K-Means Clustering

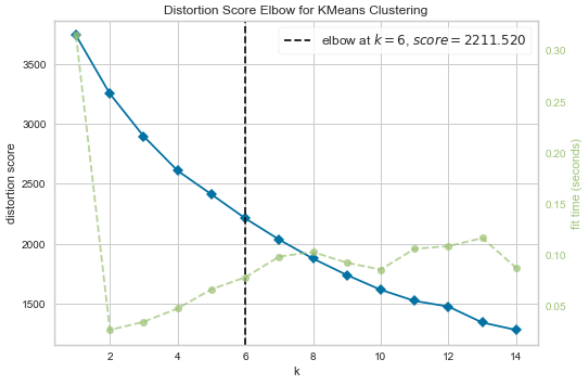
There are essentially three stopping criteria of K-means algorithm:

1.Centroids of newly formed clusters do not change

2.Points remain in the same cluster

3.Maximum number of iterations are reached





Selecting k with Elbow Method

We would choose a value of k where the distortion (sum of the squared distances between the observations and the corresponding centroid) begins to flatten out and we see an inflection point.

There is an elbow forming at 2, 4 and 8.

The distortion score elbow for K-means clustering shows the optimum number as 6.

The appropriate value seems to be 4 as the slope of decline reduces after 4.

K-MEANS CLUSTERING (2/5)

Checking silhouette score

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

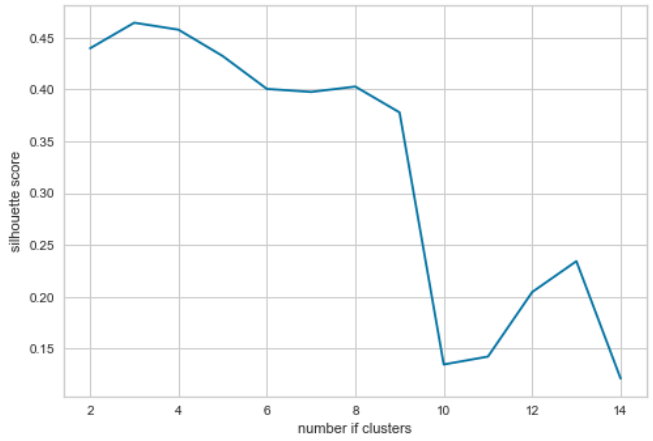
-1: Means clusters are assigned in the wrong way.

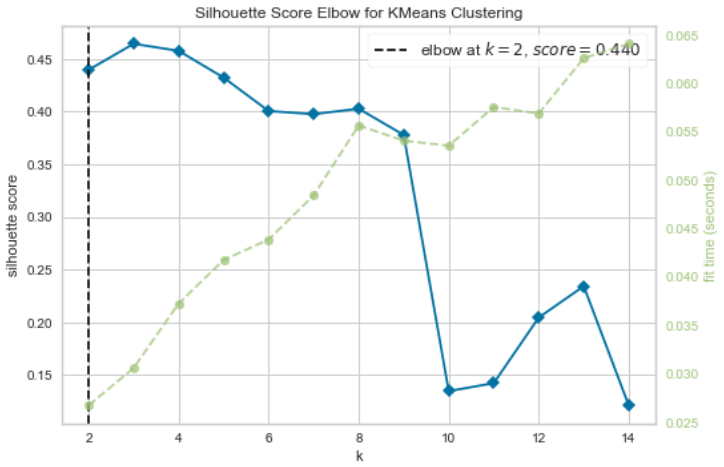
The Silhoutte score starts declining at 3 and gets flat at 6 so we can choose the cluster numbers between 3 and 6.

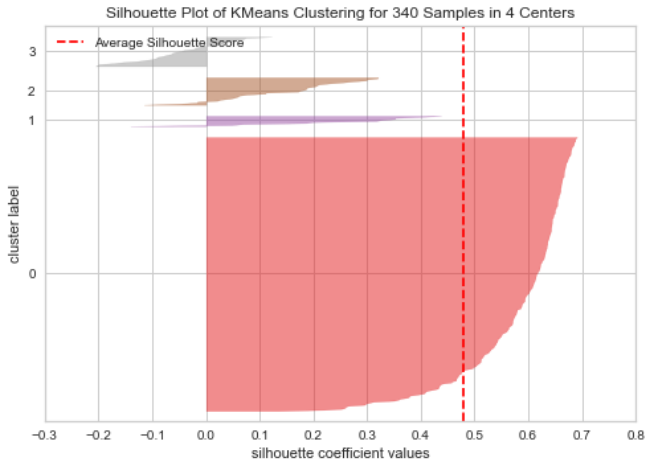
The Silhoutte score graph through KElbowVisualizer shows the elbow at 2.

The Silhoutte plot of K-Means clustering for 340 samples in 4 centers show few samples as being wrongly classified however majority of the samples will be correctly classified through it.

We would choose the number of clusters as 4.







K-MEANS CLUSTERING (3/5)

Cluster Profiles

- The data has been grouped into 4 clusters and the group with the maximum feature value has been highlighted.

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	73.710198	5.094506	1.374468	35.375	50.944853	15016308.8	1567259489	3.740827	436333424	23.412657	-3.520275	272
1	45.45	6.053419	1.094033	24.4	68.7	-1461900000	1.4506E+10	3.333	4582322873	15.465907	-5.323276	10
2	37.248621	-15.03884	2.877413	100.137931	48.931034	-186374793	-3578102034	-8.714138	482937581	87.557226	1.362139	29
3	203.769424	12.981809	1.744529	23.896552	270.517241	1200766345	1396729069	5.032414	609604309	69.869318	13.346229	29

- The cluster 0 contains 272 companies, cluster 1 contains 10, cluster 2 contains 29 and cluster 3 contains 29 companies.
- The high current price cluster is the Custer 3 in K-Means which contain 29 companies.
- The highest price change, volatile, ROE, lowest earnings per share cluster is 2 which contain 29 companies.
- The highest price change, volatile, ROE cluster is 2 which contain 29 companies.
- The lowest net income and heist income with lowest shares outstanding is the cluster 1 with 10.
- Cluster 0 is dominated by Industrials followed by Financials.
- Cluster 2 is dominated by energy sector.

KM_segments	GICS Sector	Count
0	Consumer Discretionary	33
	Consumer Staples	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
	Telecommunications Services	2
2	Energy	22
	Industrials	1
	Information Technology	3
	Materials	1
3	Consumer Discretionary	6
	Consumer Staples	1
	Energy	1
	Financials	1
	Health Care	9
	Information Technology	5
	Real Estate	1
	Telecommunications Services	1

K-MEANS CLUSTERING (4/5)

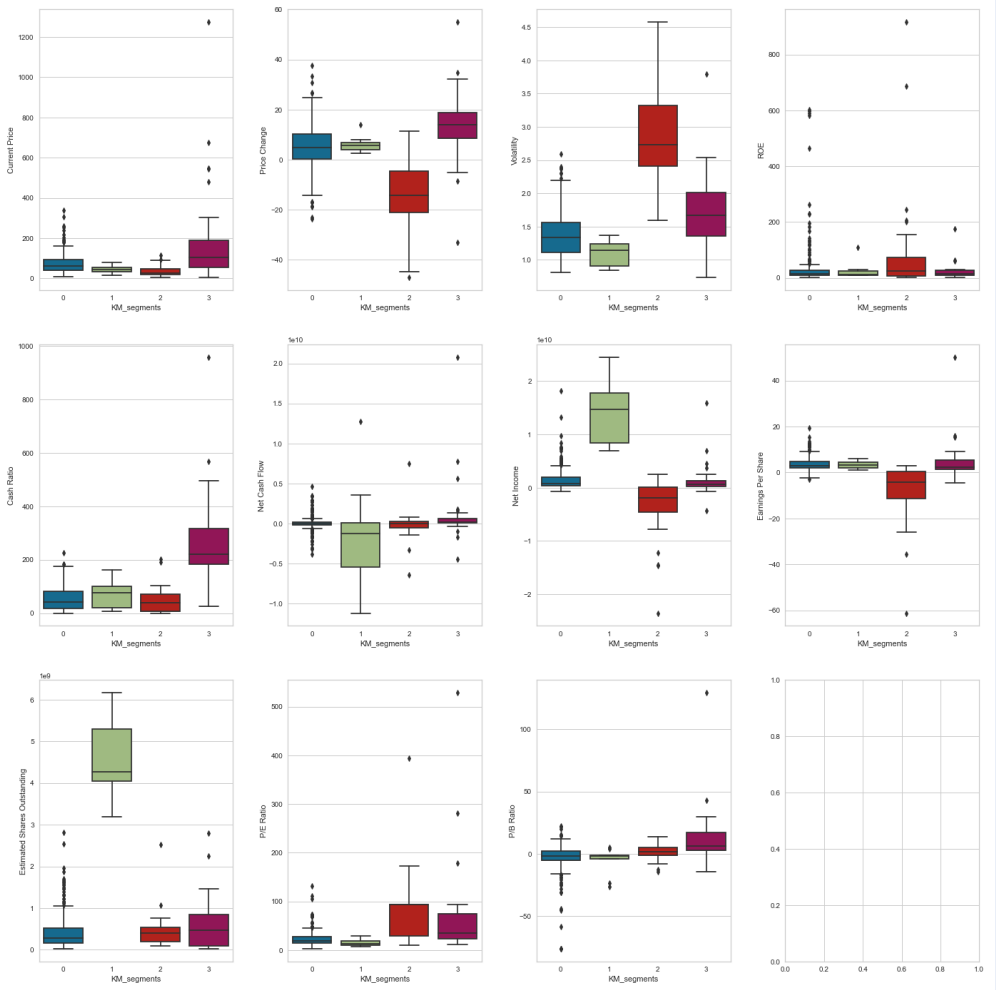
Plots of numerical features across the clusters

Cluster 0

- The current price of these cluster companies fall in the medium range.
- There is less price change thereby being less volatile hence less risky with few exception.
- The cash ratio is less which means the company is at risk of having financial difficulty. However, a low cash ratio may also be an indicator of a company's specific strategy that calls for maintaining low cash reserves, because funds are being used for expansion.
- The Net cash flow, net income, earnings per share is less.
- The estimated shares outstanding is less which indicate that the company has lot of closely held shares.
- The P/E and P/B ratios are also low.

Cluster 1

- The current price of these cluster companies fall in the low range.
- There is less price change thereby being less volatile hence less risky with few exception.
- The cash ratio is less which means the company is at risk of having financial difficulty. However, a low cash ratio may also be an indicator of a company's specific strategy that calls for maintaining low cash reserves, because funds are being used for expansion.
- The Net cash flow is negative, net income, earnings per share is less since a negative net income means a company has a loss, and not a profit, over a given accounting period. While a company may have positive sales, its expenses and other costs will have exceeded the amount of money taken in as revenue.
- The estimated shares outstanding is large which indicate that the company has lot of shares in the market and liquidity of shares is more.
- The P/E and P/B ratios are also low.



K-MEANS CLUSTERING (5/5)

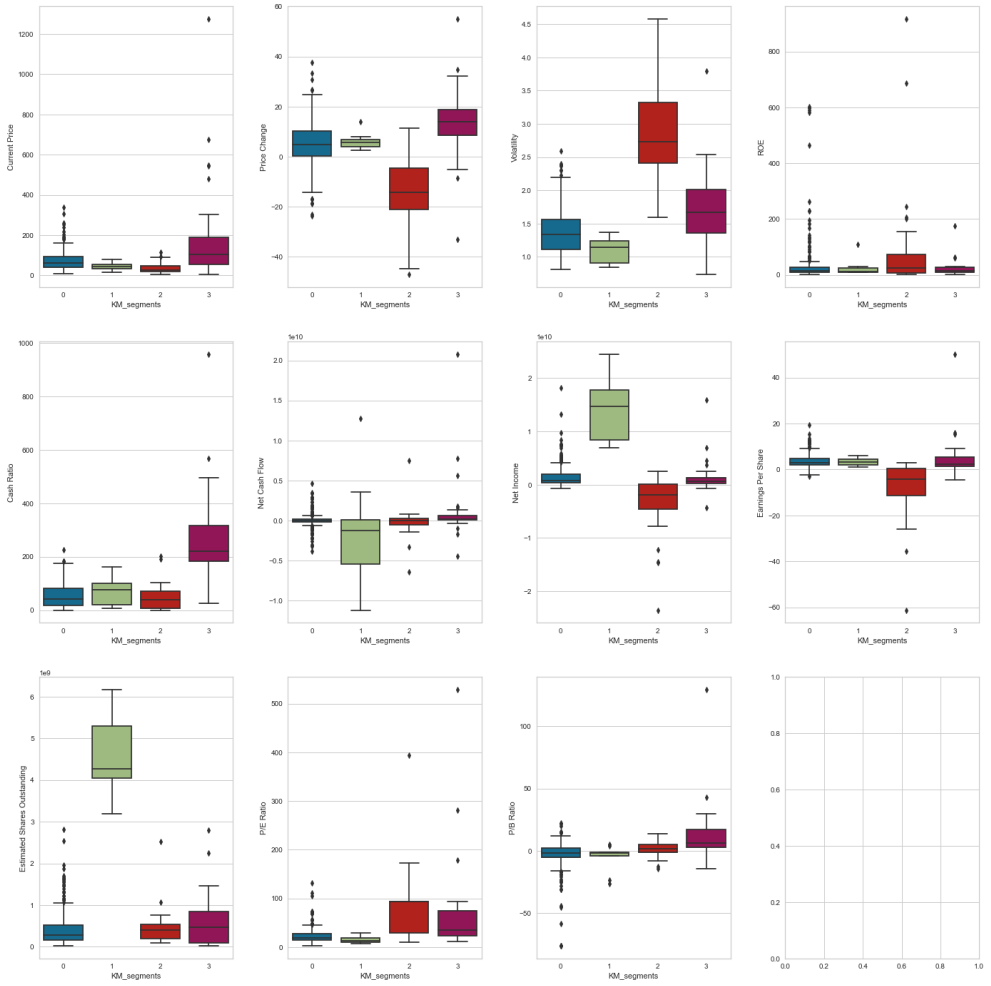
Plots of numerical features across the clusters

Cluster 2

- The current price of these cluster companies fall in the low range.
- There is high price change but to the negative side ie the price has reduced thereby being volatile hence riskier.
- The cash ratio is less which means the company is at risk of having financial difficulty. However, a low cash ratio may also be an indicator of a company's specific strategy that calls for maintaining low cash reserves, because funds are being used for expansion.
- The Net cash flow has a mean of 0.
- The net income, earnings per share is negative.
- The estimated shares outstanding is less which indicate that the company has lot of closely held shares.
- The P/E ratio is more, and P/B ratios is low.

Cluster 3

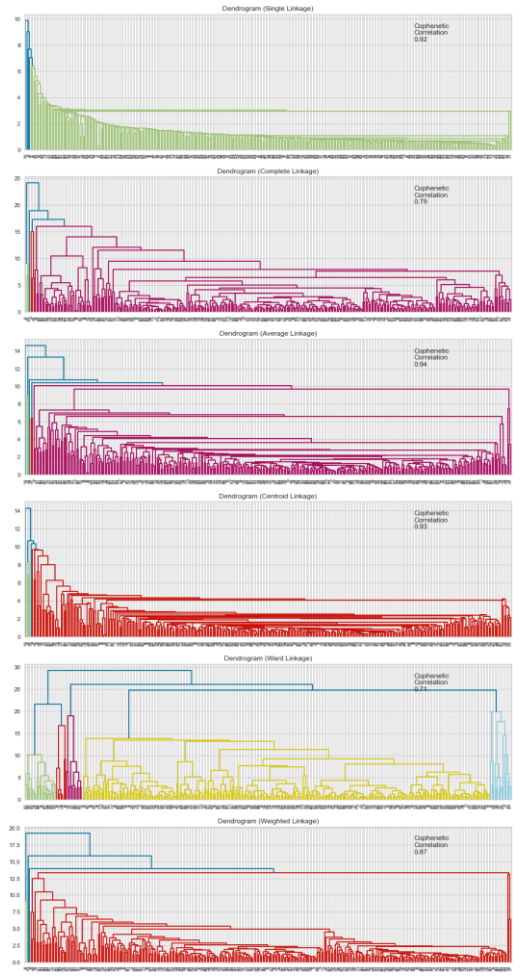
- The current price of these cluster companies fall in the high range.
- There is price change to the positive side thereby being volatile hence riskier.
- The cash ratio is less which means the company is at risk of having financial difficulty. However, a low cash ratio may also be an indicator of a company's specific strategy that calls for maintaining low cash reserves, because funds are being used for expansion.
- The Net cash flow, net income, earnings per share is less.
- The estimated shares outstanding is larger than cluster 0 and 2 which indicate that the company has lot shares in the market and liquidity of shares is more.
- The P/E and P/B ratios are also good.



HIERARCHICAL CLUSTERING (1/4)

- Cophenetic correlation (more precisely, the cophenetic correlation coefficient) is a measure of how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data points. The magnitude of this value should be very close to 1 for a high-quality solution.
  Distance matrics considered: euclidean, chebyshev, mahalanobis, cityblock
  Linkage methods considered: single, complete, average, weighted
- Cophenetic correlation of above distance metrics have been found with the combinations of linkage methods.
- Highest cophenetic correlation is 0.9422540609560814, which is obtained with Euclidean distance and average linkage.
- By observing the dendrograms for the different linkage methods with Euclidean distance it can be seen that the dendrogram with ward linkage shows distinct and separate clusters and are more interpretable.
- 4 would be the appropriate number of the clusters from the dendrogram with Ward linkage method.

	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 (2/4)

Cluster Profiles

- The data has been grouped into 4 clusters and the group with the maximum feature value has been highlighted.

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

- The cluster 0 contains 29 companies, cluster 1 contains 15, cluster 2 contains 11 and cluster 3 contains 285 companies.
- The high current price cluster is the Custer 1 in K-Means which contain 15 companies.
- The highest price change, volatile, ROE, lowest earnings per share cluster is 0 which contain 29 companies.
- The highest price change, volatile, ROE cluster is 0 which contain 29 companies.
- The lowest net income and highest income with lowest shares outstanding is the cluster 2 with 11 companies.

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

HIERARCHICAL CLUSTERING (3/4)

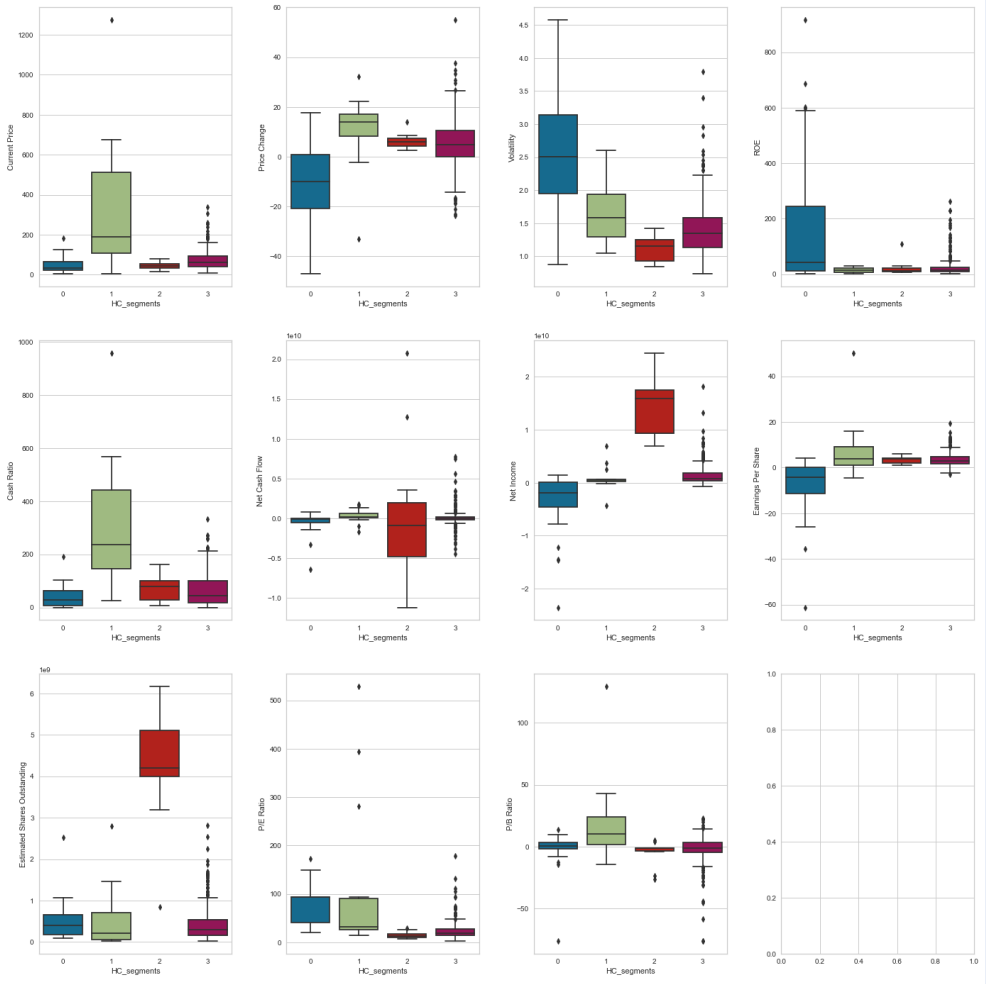
Plots of numerical features across the clusters

Cluster 0

- The current price of these cluster companies fall in the low range.
- There is a lot of price change thereby being highly volatile hence risky.
- The ROE is very high as compared to all the other clusters.
- The cash ratio is less which means the company is at risk of having financial difficulty. However, a low cash ratio may also be an indicator of a company's specific strategy that calls for maintaining low cash reserves, because funds are being used for expansion.
- The Net cash flow is negative, net income, earnings per share is less since a negative net income means a company has a loss, and not a profit, over a given accounting period. While a company may have positive sales, its expenses and other costs will have exceeded the amount of money taken in as revenue.
- The estimated shares outstanding is less which indicate that the company has lot of closely held shares.
- The P/E and P/B ratios is high however the P/B ratio has mean of zero.

Cluster 1

- The current price of these cluster companies fall in the high range.
- There is price change to the positive side thereby being medium volatile hence riskier.
- The ROE is very less.
- The cash ratio is very high which means the company has the means to pay off its current debts with funds leftover. The higher the ratio, the better it is, as it indicates that the company is generating more revenue per rupee spent on the asset.
- The Net cash flow, net income is less however the earnings per share is more as compared to the other clusters.
- The estimated shares outstanding is less which indicate that the company has lot of closely held shares.
- The P/E and P/B ratios are good.



HIERARCHICAL CLUSTERING (4/4)

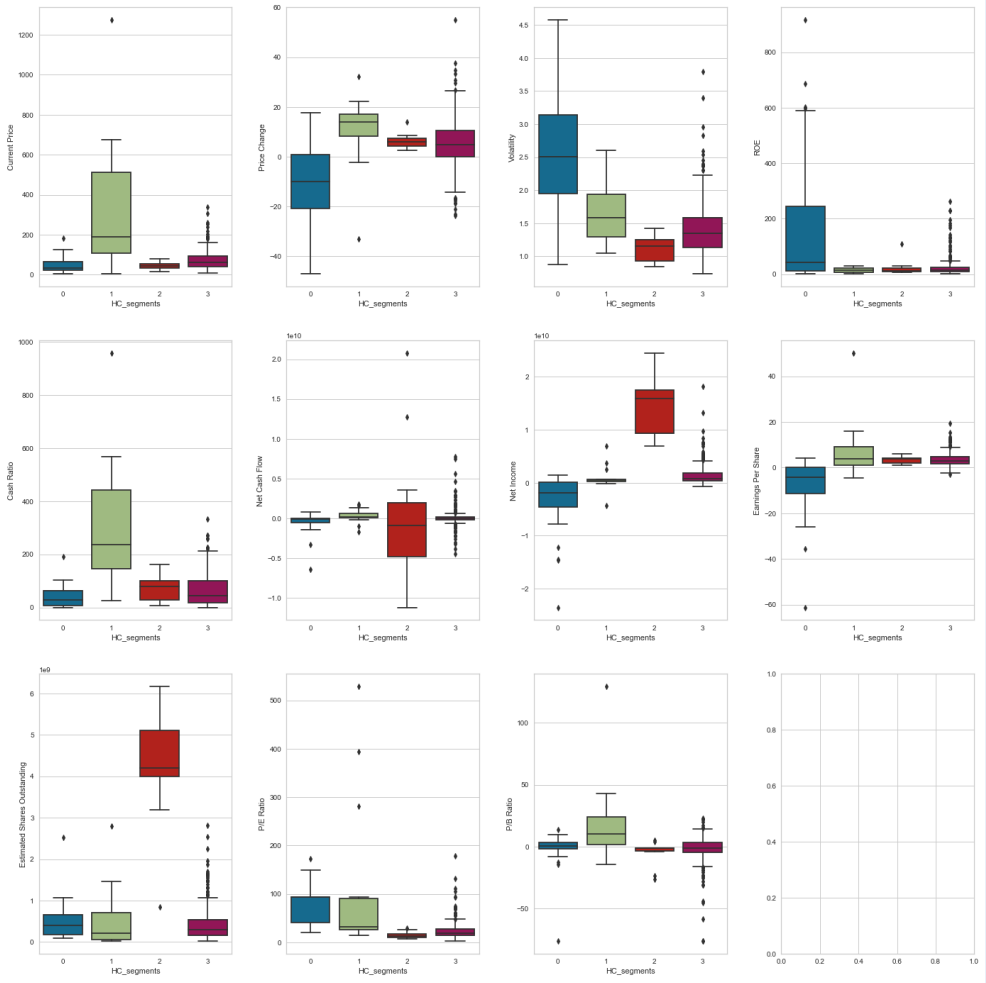
Plots of numerical features across the clusters

Cluster 2

- The current price of these cluster companies fall in the low range.
- There is very less price change thereby being very less volatile hence not risky.
- The ROE is very low.
- The cash ratio is less which means the company is at risk of having financial difficulty. However, a low cash ratio may also be an indicator of a company's specific strategy that calls for maintaining low cash reserves, because funds are being used for expansion.
- The Net cash flow is negative, net income is high, earnings per share is less since a negative net income means a company has a loss, and not a profit, over a given accounting period. While a company may have positive sales, its expenses and other costs will have exceeded the amount of money taken in as revenue.
- The estimated shares outstanding is large which indicate that the company has lot shares in the market and liquidity of shares is more.
- The P/E and P/B , P/B ratio are low.

Cluster 3

- The current price of these cluster companies fall in the low range.
- There is price change to the positive side thereby being medium volatile hence riskier.
- The ROE is very less.
- The cash ratio is less which means the company is at risk of having financial difficulty. However, a low cash ratio may also be an indicator of a company's specific strategy that calls for maintaining low cash reserves, because funds are being used for expansion.
- The Net cash flow, net income , earnings per share is less.
- The estimated shares outstanding is less which indicate that the company has lot of closely held shares.
- The P/E and P/B ratios are low.



K-MEANS VS HIERARCHICAL CLUSTERING

- When comparing the execution time for both the clustering methods, K-Means clustering took less time for execution.
- Hierarchical clustering gave more distinct clusters as compared to the K-Means clustering.
- The high current price cluster is the Custer 3 in K-Means which contain 29 companies as compared to the 15 in cluster 1 which is a high price cluster in Hierarchical Clustering.
- The highest price change, volatile, ROE, lowest earnings per share cluster is 2 in K-means which contain 29 companies and is same as the cluster 0 with highest price change, volatile, ROE, lowest earnings per share in Hierarchical clustering with 29 companies.
- The highest price change, volatile, ROE cluster is 2 in K-means which contain 29 companies and is same as the cluster 0 with highest price change, volatile, ROE in Hierarchical clustering.
- The lowest net income and higest income with lowest shares outstanding is the cluster 1 in K-means clustering with 10 companies and the lowest net income and heist income with lowest shares outstanding is the cluster 2 with 11 companies in Hierarchical clustering.
- The clustering results from both the methods are more or less similar.
- By using K-Means clustering we concluded to have 4 clusters and by using Hierarchical clustering it was concluded to have 4 clusters.

INSIGHTS AND RECOMMENDATIONS

- Cluster 0 has lower current price with higher price change and higher volatility. High volatility will witness sharper price changes, making it a riskier investment. The investors having risk appetite and looking for higher gains can be suggested investment from this group.
- The Cluster 0 shares are appropriate for short term investments to gain quick returns by utilizing its high volatility.
- Cluster 0 contains majority companies from energy sector.
- Cluster 1 has higher current price with moderate price change and moderate volatility making it a medium risk cluster.
- The investors seeking a reliable long term investment can be suggested this cluster as it has moderate risk and returns.
- Cluster 2 companies have lower current price, low price change, low volatility, low ROE, low Cash Ratio, Negetive Net Cash Flow, high Net Income, Low earnings per Share, High estimated shares outstanding, low P/E and P/B ratio.
- The cluster 2 companies are those which have highly traded/floating shares in the market and are big conglomerates with expansions in process.
- The cluster 3 companies have lower current price, moderate price change, moderate volatility, low ROE, low Cash Ratio, zero mean Net Cash Flow, low Net Income, Low earnings per Share, low estimated shares outstanding, low P/E and P/B ratio.
- The cluster 3 has maximum number of companies from Industrials followed by Financials and then Consumer Discretionary.
- These stocks can be referred to the investors who are not willing to take large risk and want to invest for long terms.