Trade&Ahead

Data based classification : Stay Ahead

Background

1. 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.
2. It is important to maintain a diversified portfolio when investing in stocks in order to maximize earnings under any market condition. It is often easy to get lost in a sea of financial metrics to analyze while determining the worth of a stock.
3. 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.

Business Problem Overview and Solution Approach

Trade&Ahead is a financial consultancy firm who provide their customers with personalized investment strategies.

Objective

The objective is to analyze the data, grouping the stocks based on the attributes provided, and sharing insights about the characteristics of each group.

Data comprising stock price and some financial indicators for a few companies listed under the New York Stock Exchange is available to make the analysis.

Data Overview

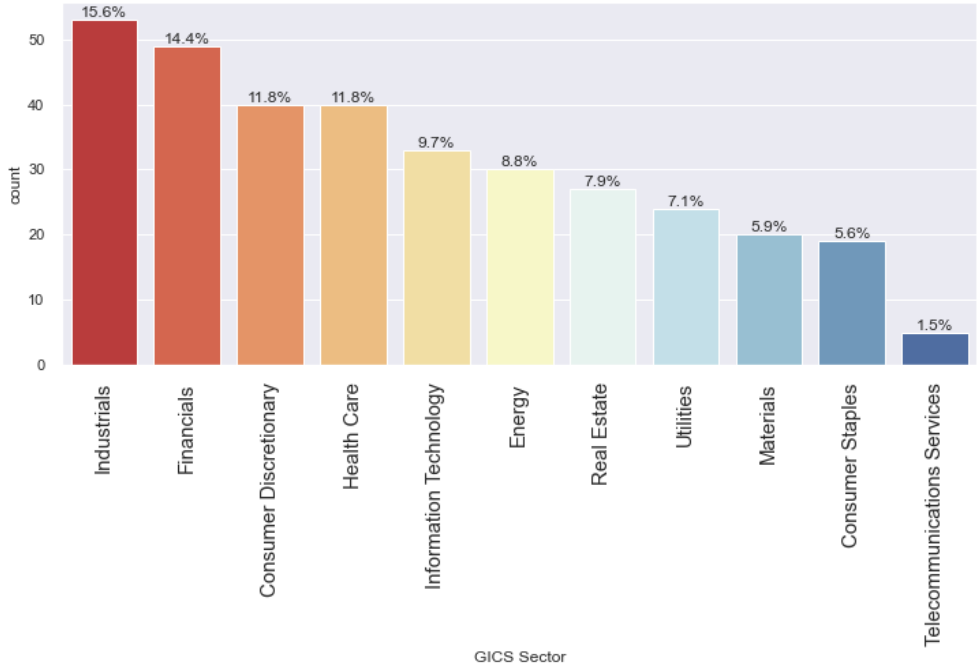
Variable	Description
Ticker Symbol	An abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market
Company	Name of the company
GICS Sector	The specific economic sector assigned to a company by the Global Industry Classification Standard (GICS) that 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 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	Variables
340	15
No Missing Data	
No Duplicates	

Exploratory Data Analysis

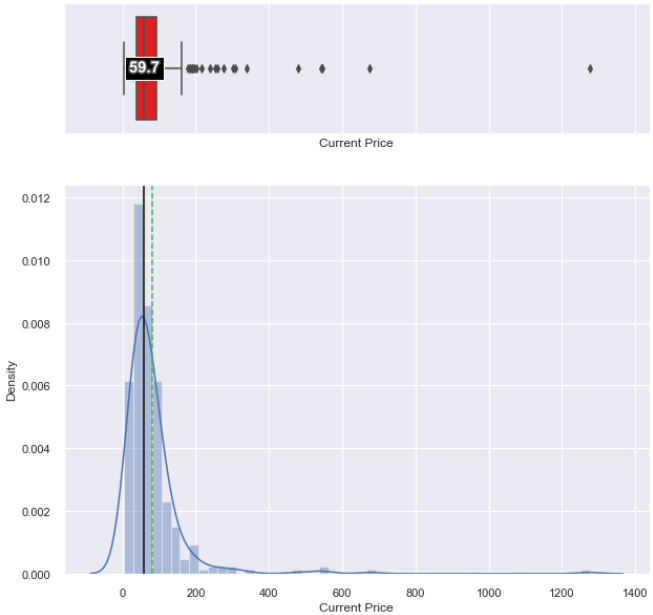
- GICS Sector

- Industrials and Financials are the top GISC Sector, each having 15.6 and 14.4 % respectively

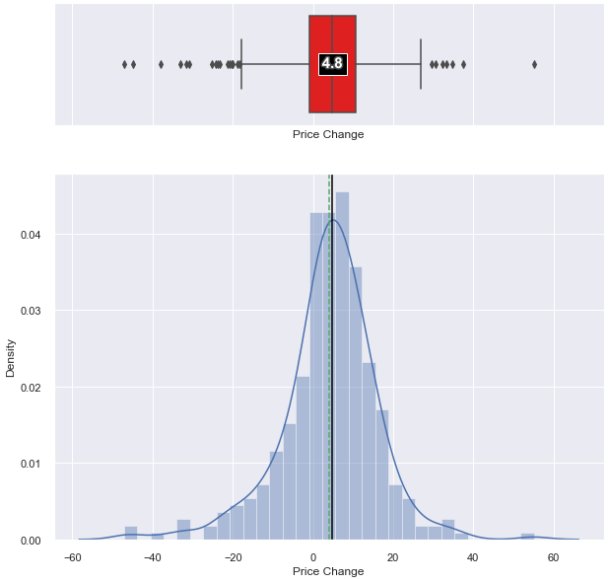


Exploratory Data Analysis

Current Price: Current Price is right skewed as expected, and there are outliers as well as expected

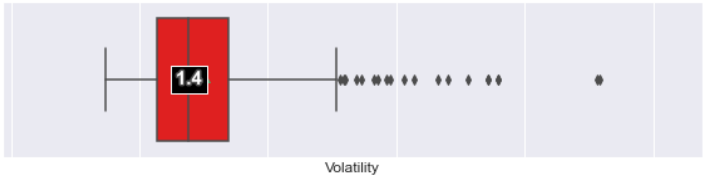


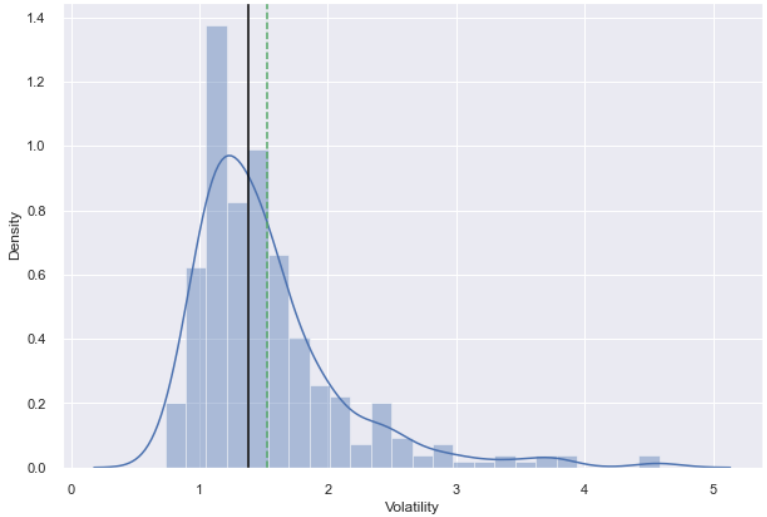
Price Change: Percentage change in the stock price in 13 weeks is almost normally distributed



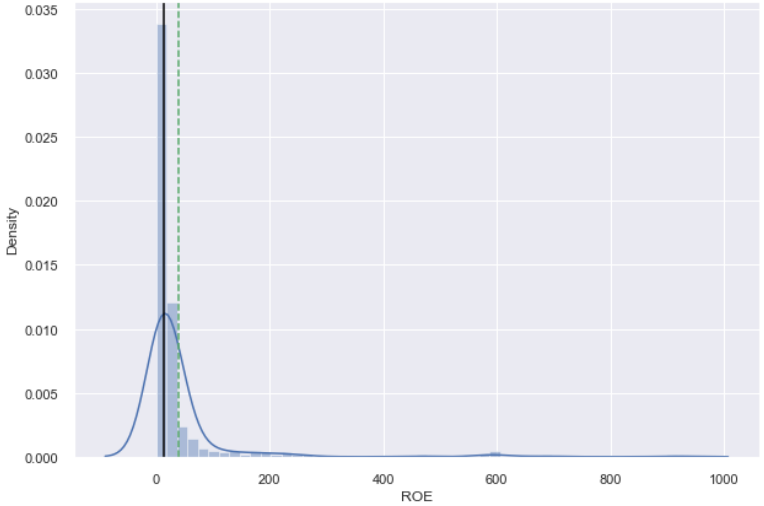
Exploratory Data Analysis

Volatility: Volatility is right tailed; some stock have had large price movements in the last 13 weeks.



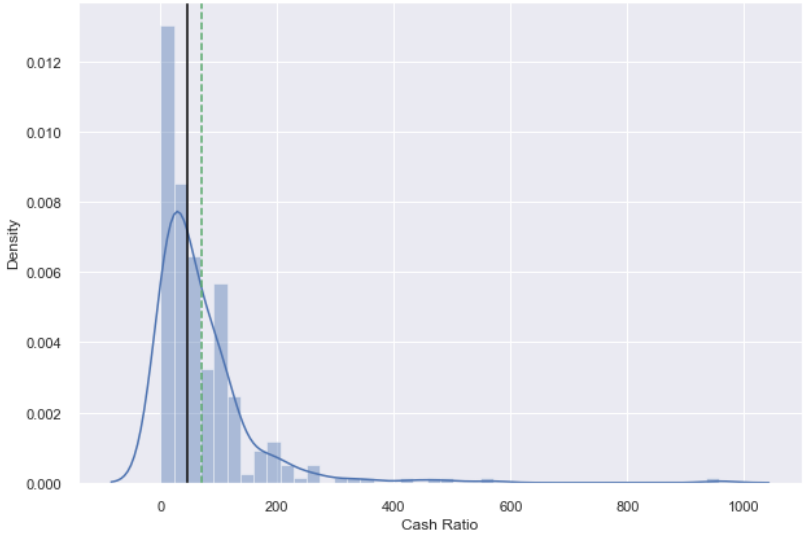


ROE: ROE is right skewed

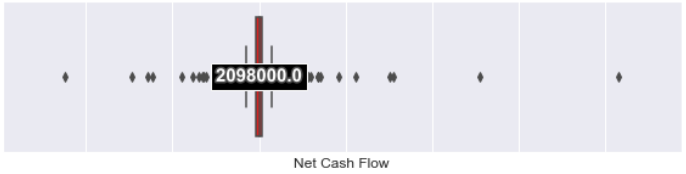


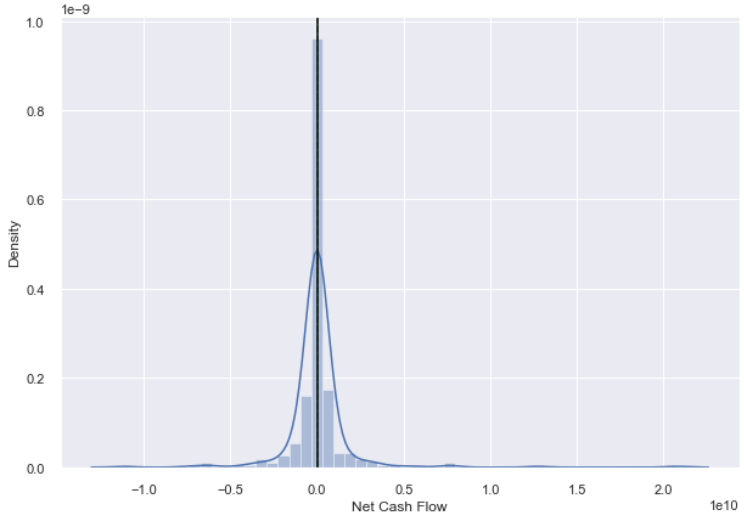
Exploratory Data Analysis

Cash Ratio: Cash Ratio is also right skewed



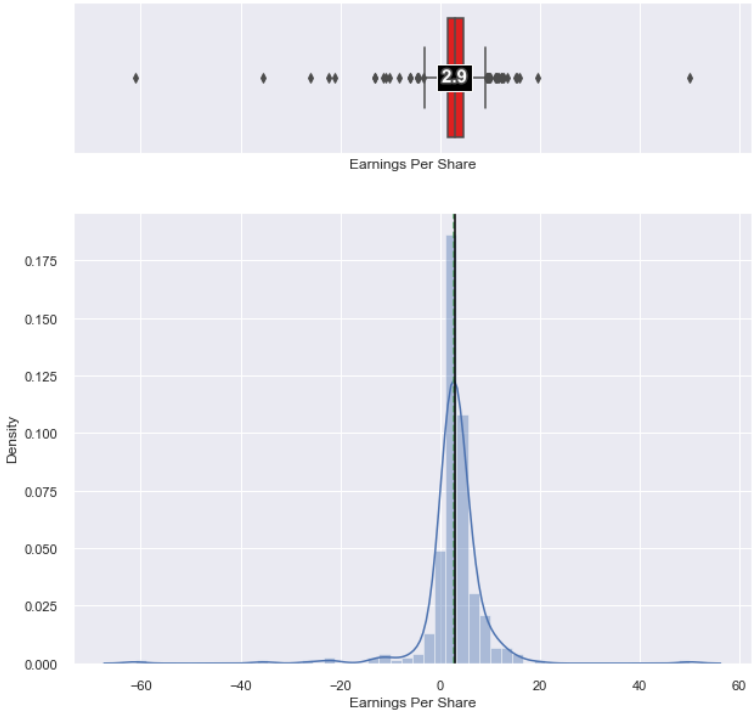
Net Cash Flow: Net Cash Flow is close to Normal distribution with a slight right skewed.



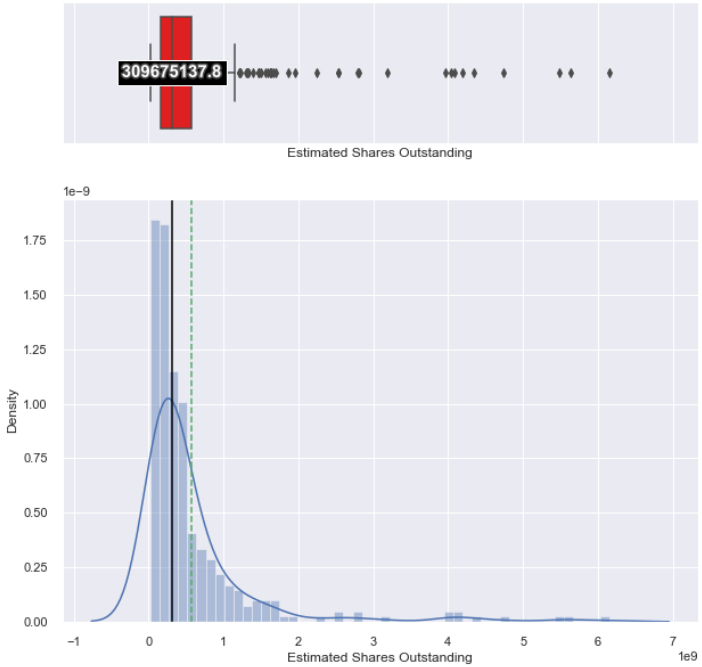


Exploratory Data Analysis

Earnings Per Share: Earnings Per Share

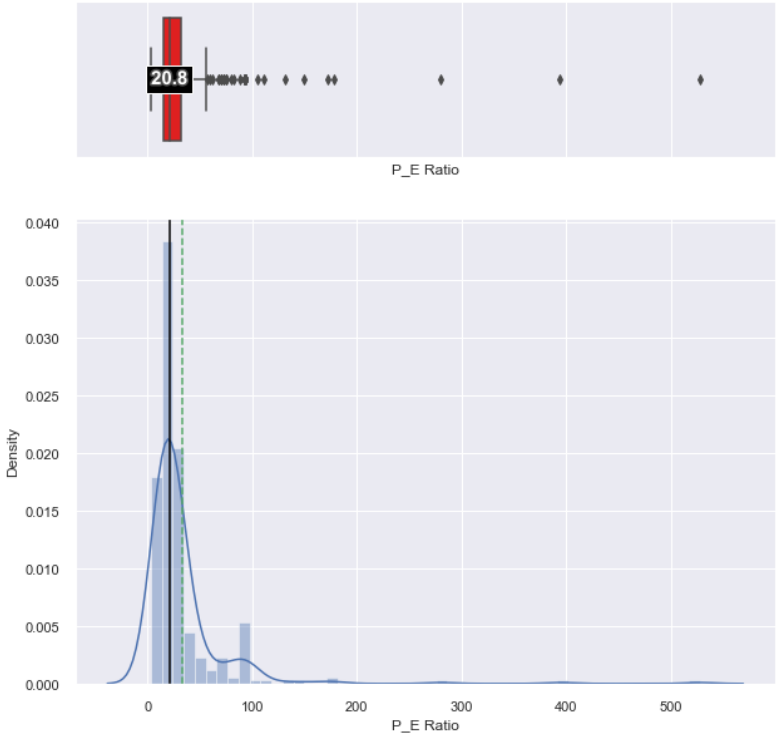


Estimated Shares Outstanding: Estimated Shares Outstanding is right skewed

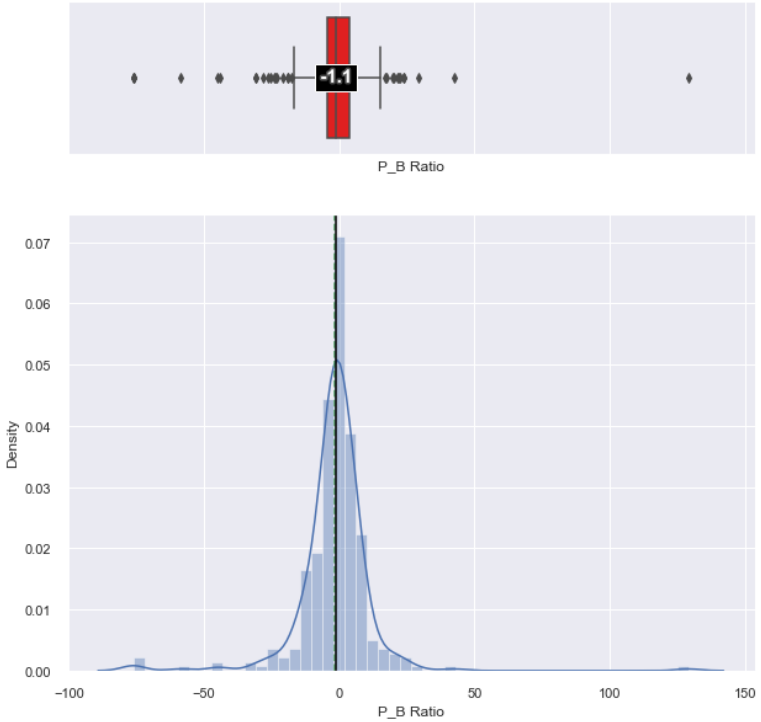


Exploratory Data Analysis

P/E Ratio : P/E ratio is right skewed



P/B Ratio : P/B ratio is close to normal distribution



Exploratory Data Analysis

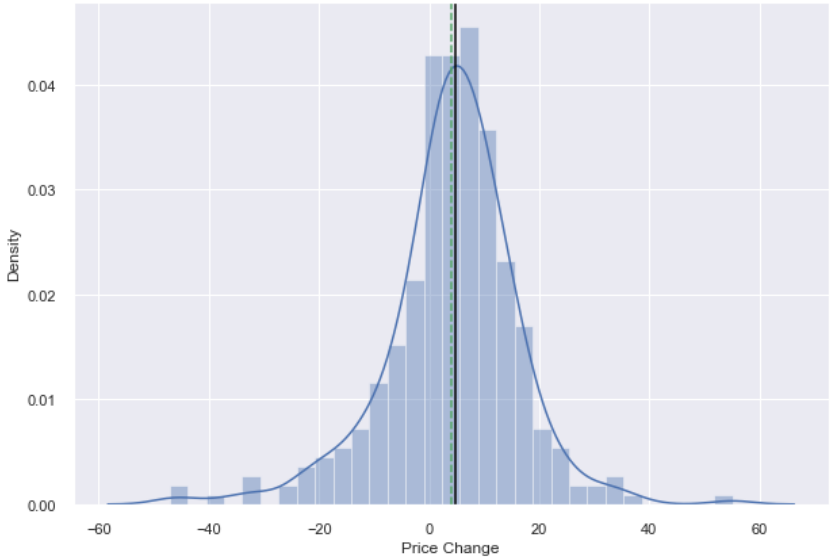
- GICS Sub Industry
- Top 20 GISC Sub Industry by no of stocks

	GICS Sector	GICS Sub Industry	0
0	Energy	Oil & Gas Exploration & Production	16
1	Real Estate	REITs	14
2	Industrials	Industrial Conglomerates	14
3	Information Technology	Internet Software & Services	12
4	Utilities	Electric Utilities	12
5	Health Care	Health Care Equipment	11
6	Utilities	MultiUtilities	11
7	Financials	Banks	10
8	Financials	Property & Casualty Insurance	8
9	Financials	Diversified Financial Services	7
10	Health Care	Biotechnology	7
11	Energy	Oil & Gas Refining & Marketing & Transportation	6
12	Consumer Staples	Packaged Foods & Meats	6
13	Information Technology	Semiconductors	6
14	Health Care	Pharmaceuticals	6
15	Health Care	Managed Health Care	5
16	Industrials	Airlines	5
17	Financials	Consumer Finance	5
18	Materials	Diversified Chemicals	5
19	Energy	Integrated Oil & Gas	5
20	Health Care	Health Care Facilities	5

Insights from EDA..

1. What does the distribution of stock prices look like?

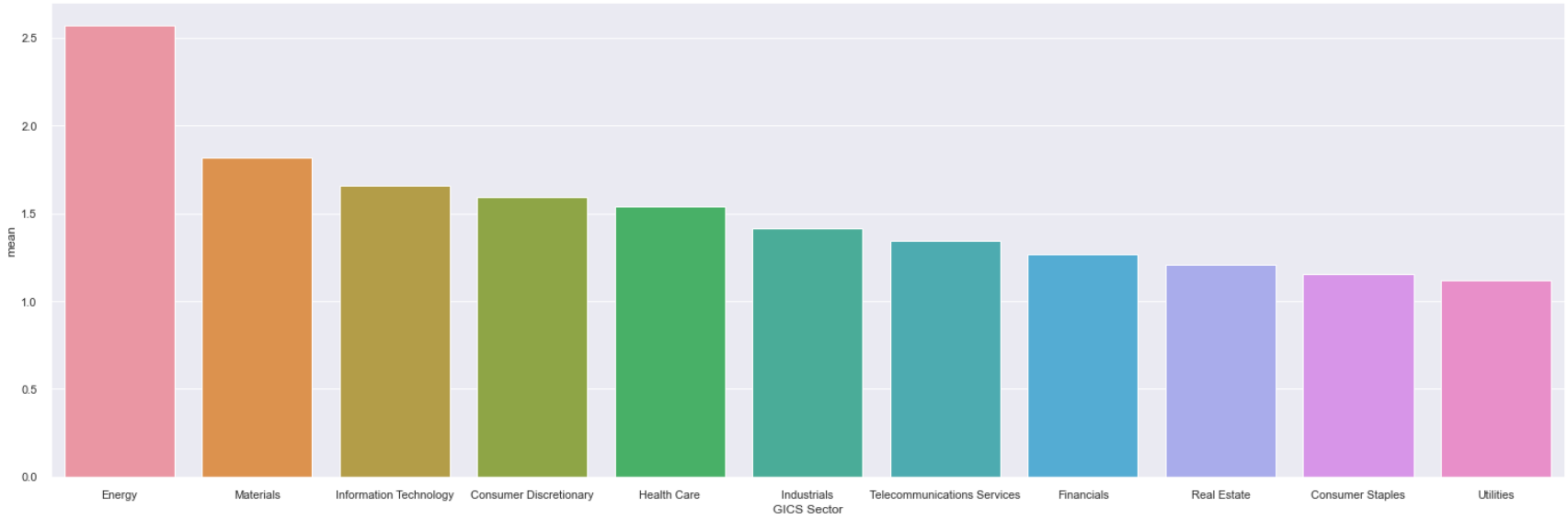
- The distribution of the stock prices almost look like a normal distribution



Insights from EDA..

2. The stocks of which economic sector have seen the maximum price increase on average?

The stocks of "Energy" economic sector have seen the maximum price increase on average.



Insights from EDA..

3. How are the different variables correlated with each other?

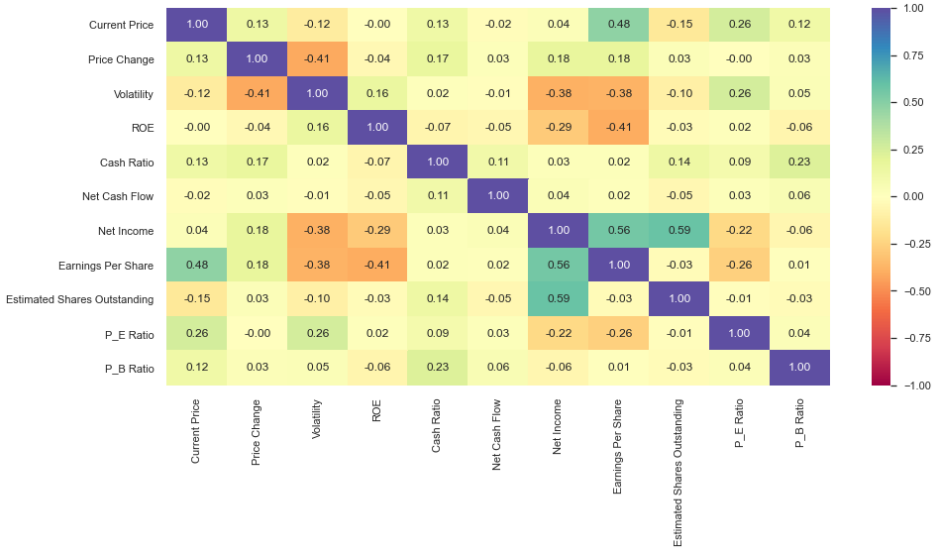
•Net Income and Estimated Shares Outstanding have good positive correlation, similarly Net Income and Earnings Per Share.

•Current Price and Earnings Per Share have strong positive correlation.

•ROE and Earnings Per Share have strong negative correlation.

•Price Change and Volatility have strong negative correlation, this is as expected.

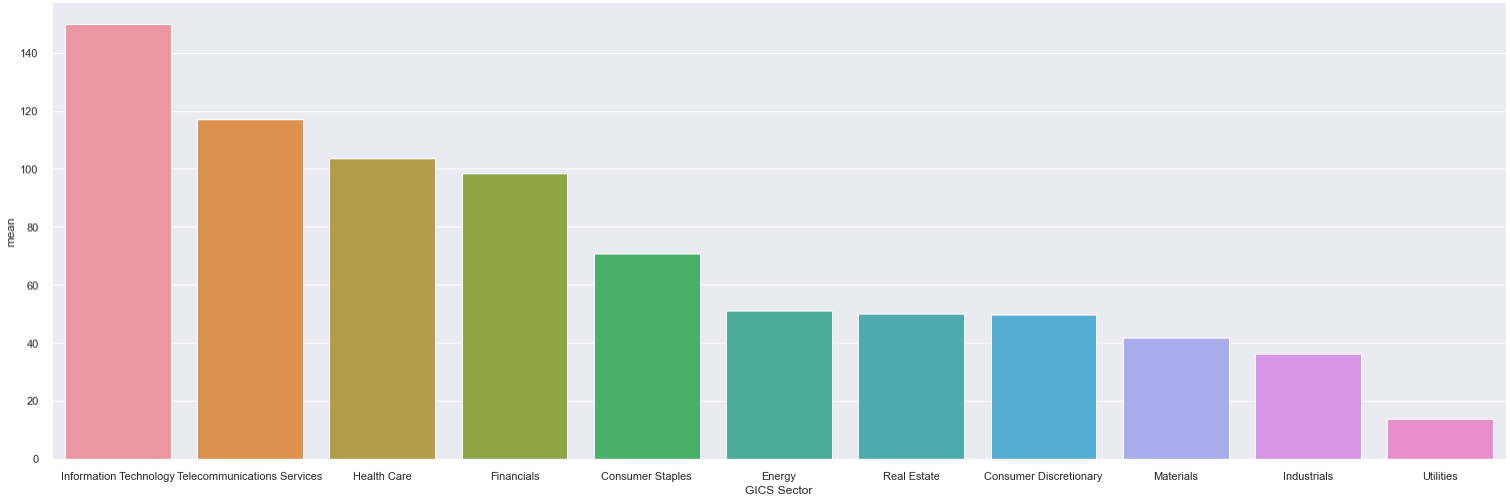
•Earnings Per Share and Volatility, Net Income and Volatility exhibit negative correlation.



Insights from EDA..

4. Cash ratio provides a measure of a company's ability to cover its short-term obligations using only cash and cash equivalents. How does the average cash ratio vary across economic sectors?

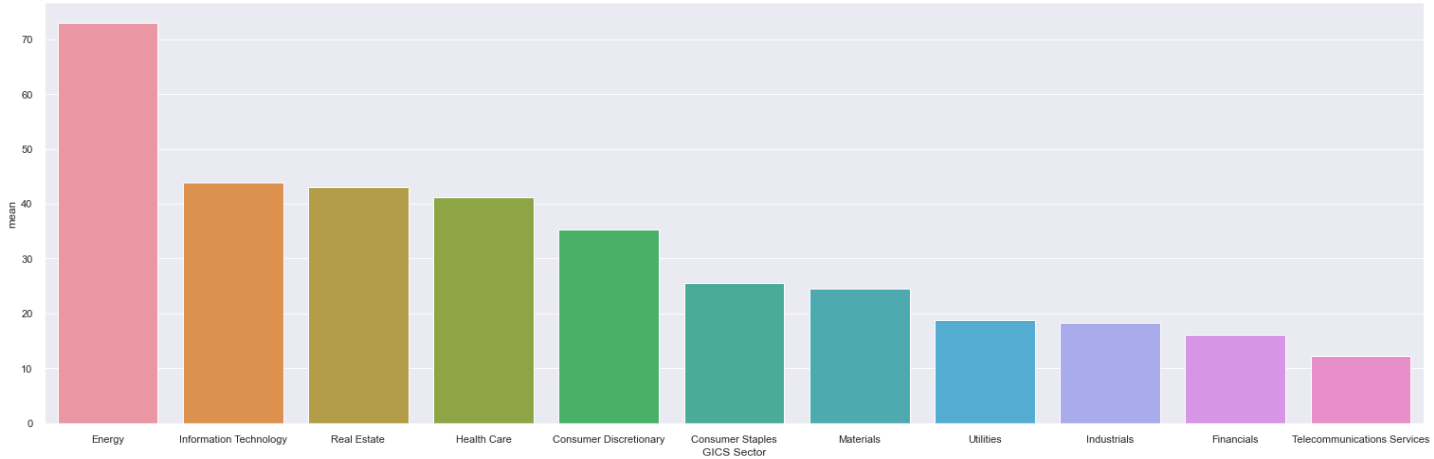
•Information Technology, Telecommunications Services and Health Care have very good cash ratio.
•Materials, Industrials and Utilities have less cash ratio.



Insights from EDA..

5. 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. How does the P/E ratio vary, on average, across economic sectors?

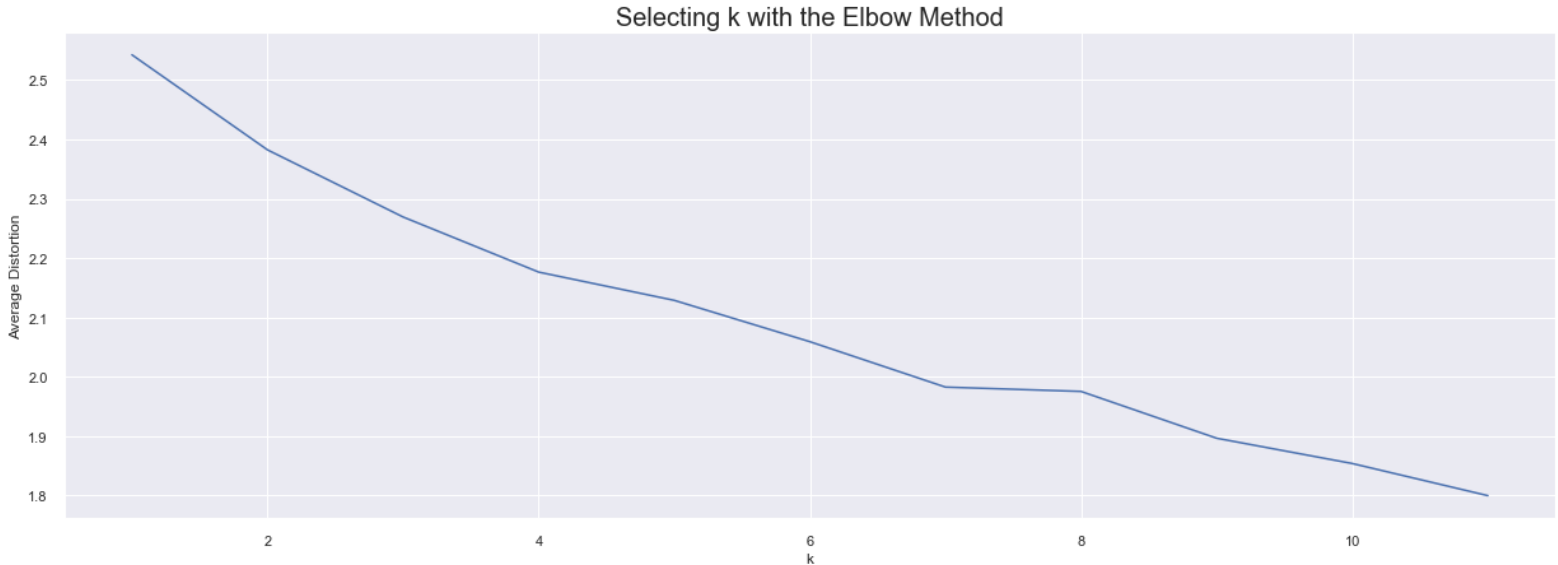
•Energy sector has a high P/E ratio
•Information Technology, Real Estate, Health Care all have High P/E ratio, which may not be good for investors.
•Utilities, Industrials, Financials have good P/E ratio for investors.
•Telecommunications Services sector has a very good P/E ratio.



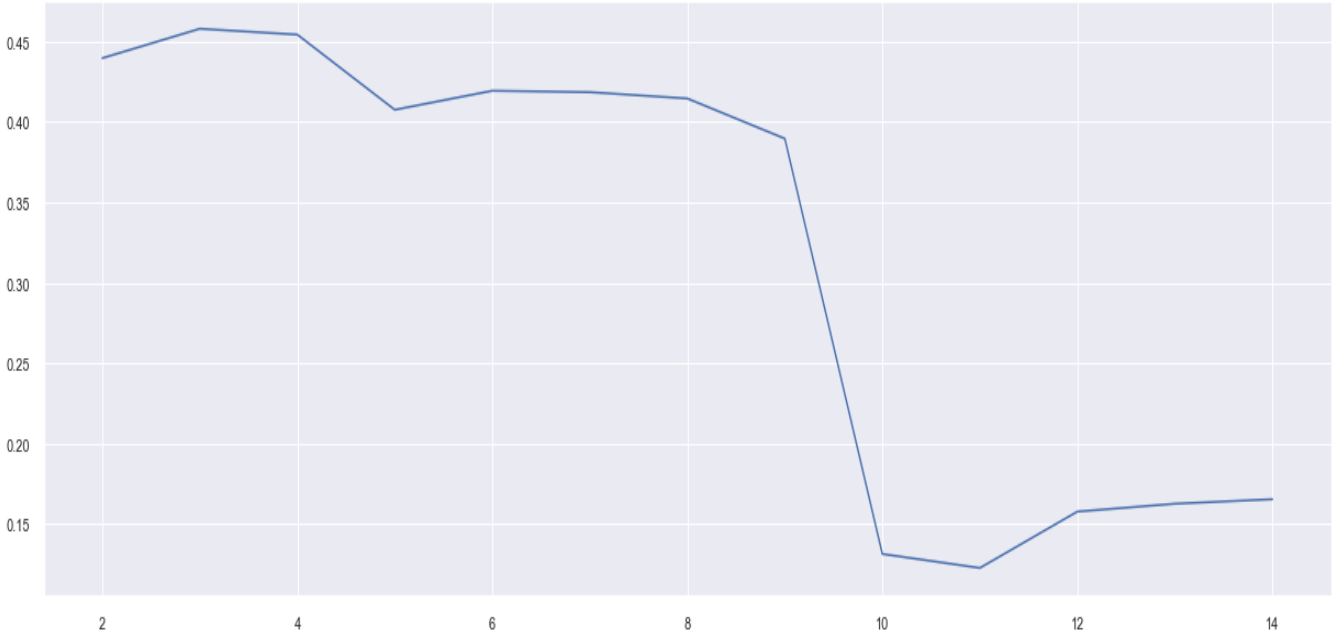
Model Building – K-means Clustering

```
Number of Clusters: 1   Average Distortion: 2.5425069919221697
Number of Clusters: 2   Average Distortion: 2.382318498894466
Number of Clusters: 3   Average Distortion: 2.2692367155390745
Number of Clusters: 4   Average Distortion: 2.176396791566185
Number of Clusters: 5   Average Distortion: 2.128799332840716
Number of Clusters: 6   Average Distortion: 2.0591416288820374
Number of Clusters: 7   Average Distortion: 1.9826333396712665
Number of Clusters: 8   Average Distortion: 1.9753526418461937
Number of Clusters: 9   Average Distortion: 1.8964970616244075
Number of Clusters: 10  Average Distortion: 1.8539123989265462
Number of Clusters: 11  Average Distortion: 1.7997730037404913
```

Appropriate value for k seems to be 8 or 10.



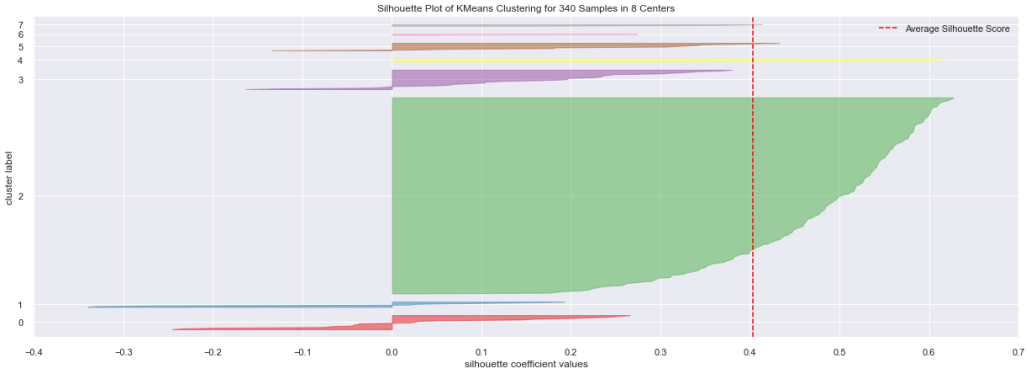
Model Building – K-means Clustering, contd..



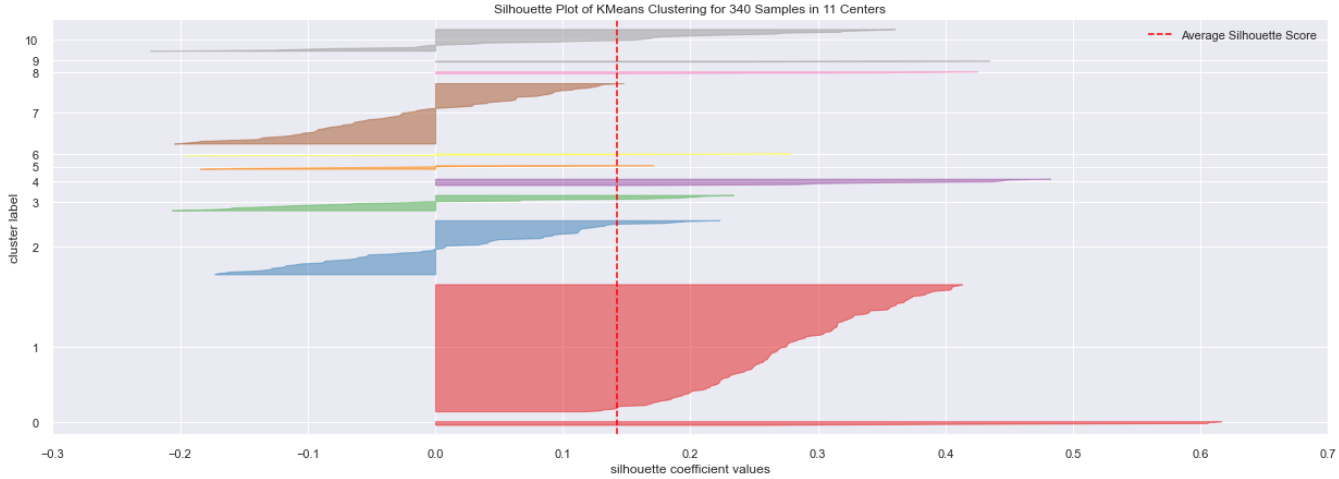
Appropriate value for k seems to be 8 or 10.

Silhouette score for 6, 7 and 8 clusters is higher. So, let us explore different values of k.

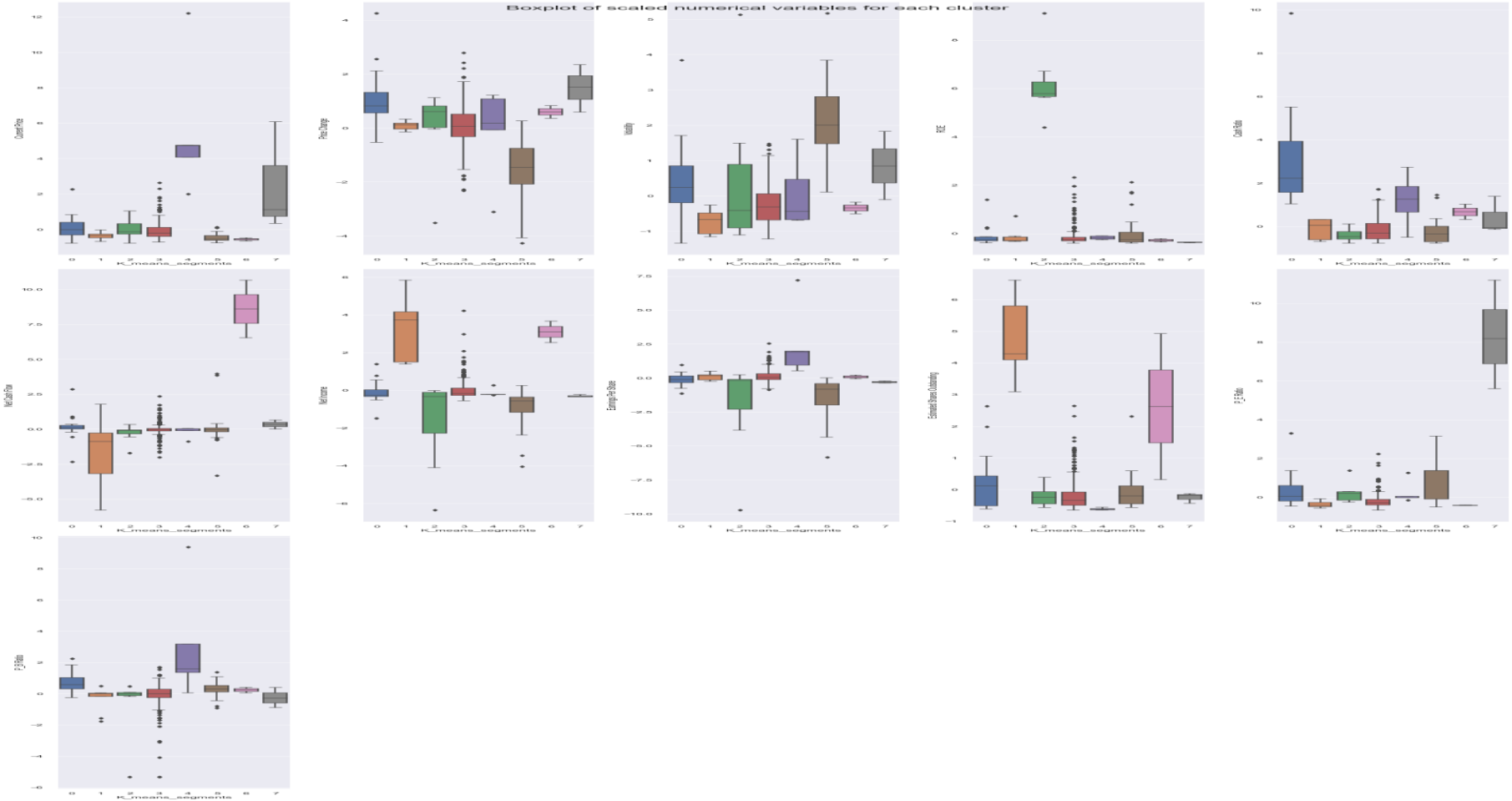
finding optimal no. of clusters with silhouette coefficients



Let us take 8 as the appropriate no. of clusters as the silhouette score is high and visual representation of the clusters at 8 gives an even distribution.



Cluster Profiling:



Insights – Clusters from K-Means

Insights

- **Cluster 0**:
  - This cluster contains stocks with a median price of approximately $91.
  - Price change of about $16 in last 13 weeks and less volatility of about 1.8.
  - The Earning per share is 2.1 and P/E ratio is 46 , which is high and P/B ration is positive 9 indicating the shares are more or less correctly valued.
  - Has the highest Cash Ratio among all clusters.
- **Cluster 1**:
  - This cluster contains stocks with a median price of approximately $46.
  - Price change of about $5 in last 13 weeks and less volatility of about 1.
  - The Earning per share is 3.4 and P/E ratio is 15 , which is high(indication high growth) and P/B ration is -negative 6 indicating the shares are undervalued.
  - Has the highest Net Income and Estimated Shares Outstanding among all clusters.
- **Cluster 2**:
  - This cluster contains stocks with a median price of approximately $84.
  - Price change of about $3.8 in last 13 weeks and less volatility of about 1.8.
  - The Earning per share is -10.8 and P/E ratio is 42.2 , which is high(indication high growth) and P/B ration is -negative 11 indicating the shares are undervalued.
  - Has the ROE among all clusters.
- **Cluster 3**:
  - This cluster contains stocks with a median price of approximately $73.
  - Price change of about $5 in last 13 weeks and less volatility of about 1.3.
  - The Earning per share is 3.7 and P/E ratio is 23.3 , which is high(indication high growth) and P/B ration is -negative 3 indicating the shares are undervalued.
  - Has the highest number of stocks among all clusters.
- **Cluster 4**:
  - This cluster contains stocks with a median price of approximately $624 which is the highest average.
  - Price change of about $2.3 in last 13 weeks and less volatility of about 1.5.
  - The Earning per share is 19.26 and P/E ratio is 42.05 , which is high(indication high growth) and P/B ration is positive 41 indicating the shares are over valued.The hignest P/B ratio among all the clusters.
  - Has the highest Earnings per share.

Insights – Clusters from K-Means, contd..

Insights

- **Cluster 5**:
  - This cluster contains stocks with a median price of approximately $34.
  - Price change of about -$15 in last 13 weeks and volatility of about 2.8, the highest among all the clusters.
  - The Earning per share is -6 and P/E ratio is 76.25 , which is high(indication high growth) and P/B ration is positive 2 indicating the shares are almost correctly valued.
  - Has the second highest number of stocks, highest volatility among all clusters.
- **Cluster 6**:
  - This cluster contains stocks with a median price of approximately $25.
  - Price change of about $11 in last 13 weeks and less volatility of about 1.3, the lowest among all clusters.
  - The Earning per share is 3.29 and P/E ratio is 13.64 , which is good and P/B ration is -negative 1.5 indicating the shares are
    undervalued.
  - Has the highest Net Income among all clusters.
- **Cluster 7**:
  - This cluster contains stocks with a median price of approximately $327.
  - Price change of about $21.9 in last 13 weeks and volatility of about 2.02.
  - The Earning per share is .75 and P/E ratio is 400.89 , which is too high and P/B ration is -negative 5 indicating the shares are undervalued.
  - Has the P/E ratio and price change among all clusters.

Model Building – Hierarchical Clustering

```
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 0.9259195530524589.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.7925307202850003.
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.

Model Building – Hierarchical Clustering, contd..

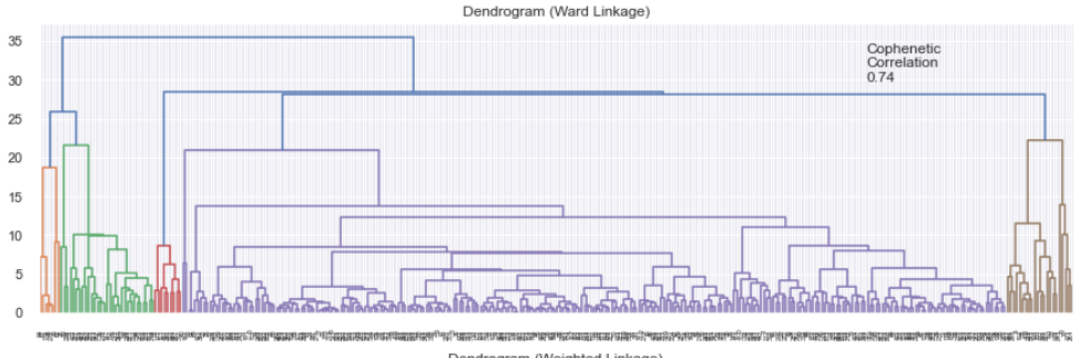
Let's explore different linkage methods with Euclidean distance only.

```
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.

We see that the cophenetic correlation is maximum with Euclidean distance and average linkage.

Hierarchical Clustering - Dendrogram



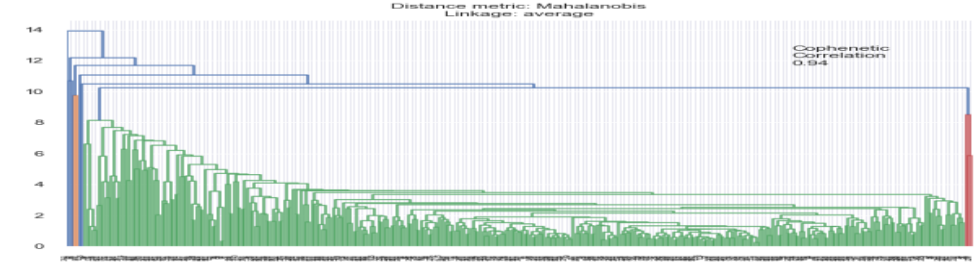
Dendrogram for Ward linkage shows distinct and separate clusters.

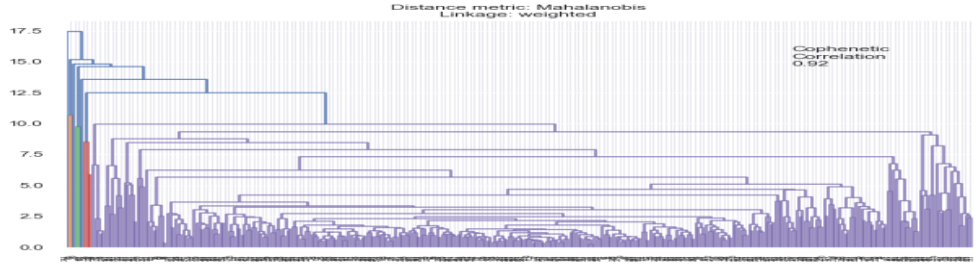
	Linkage	Cophenetic Coefficient
0	single	0.923227
1	complete	0.787328
2	average	0.942254
3	centroid	0.931401
4	ward	0.710118
5	weighted	0.869378

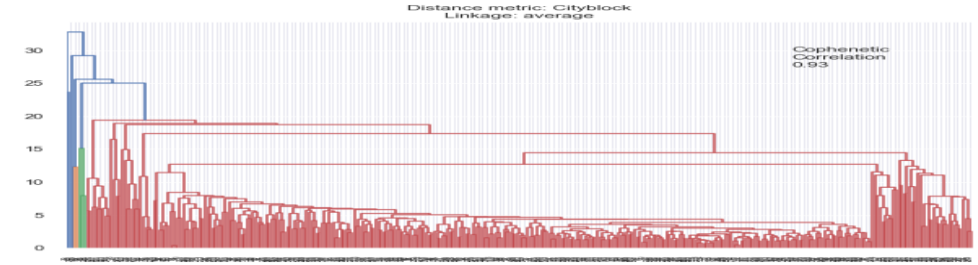
Hierarchical Clustering - Mahalanobis and Manhattan distances with average and weighted linkage methods

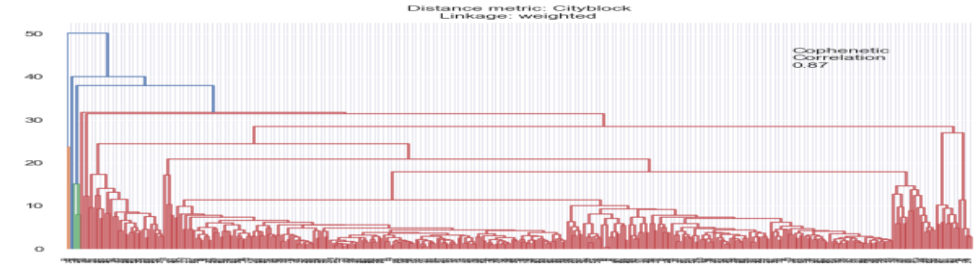
Out of all the dendrograms we saw, the dendrogram with Ward linkage gave us separate and distinct clusters.

Let us create 7 clusters.

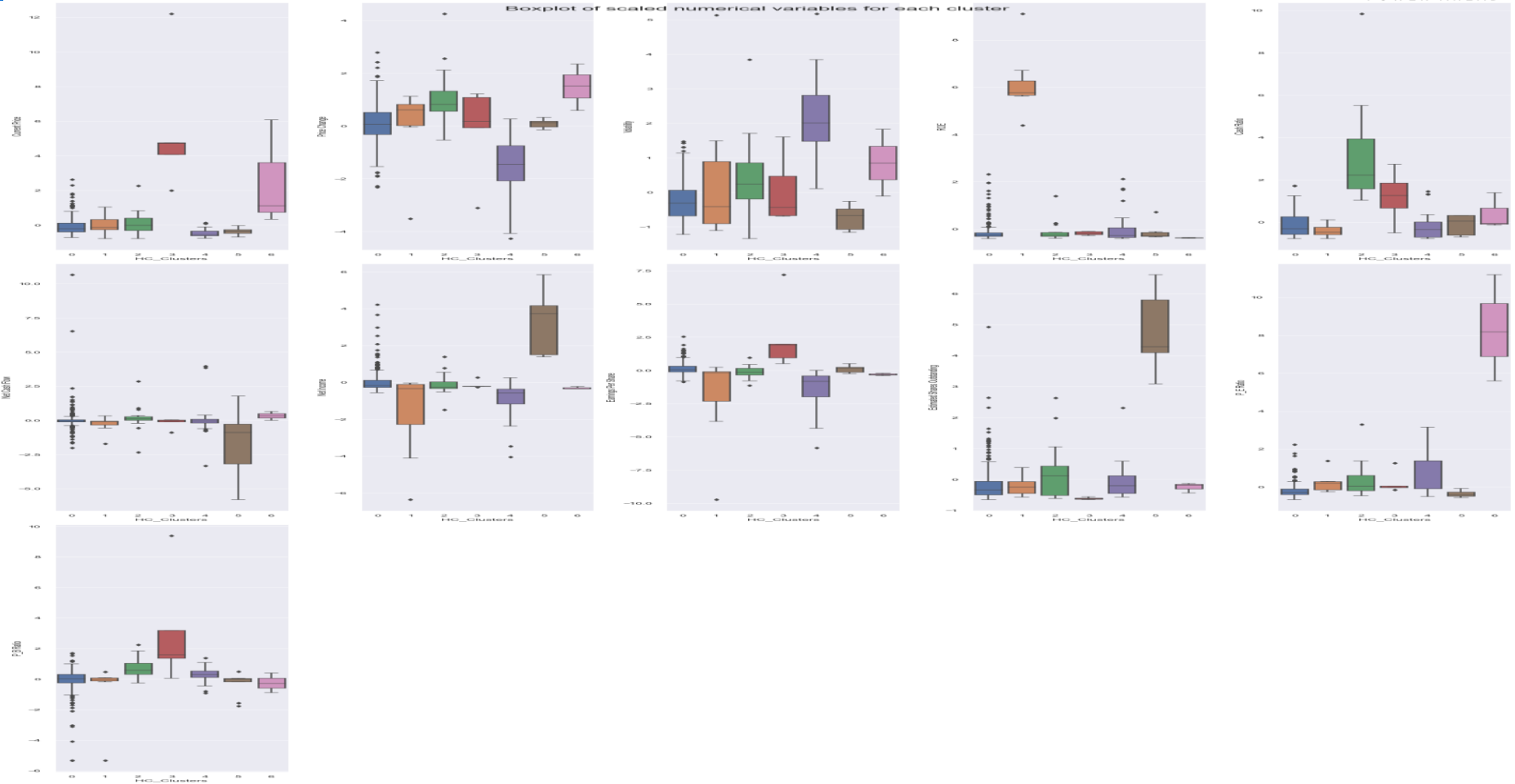








Cluster Profiling:



Insights – Clusters from Hierarchical clustering

Insights

- **Cluster 0**:
  - This cluster contains stocks with a median price of approximately $73.
  - Price change of about $5 in last 13 weeks and less volatility of about 1.3.
  - The Earning per share is 3.78 and P/E ratio is 23 , which is good and P/B ration is -negative indicating the shares are undervalued.
  - Has the highest number of stocks among all clusters.
- **Cluster 1**:
  - This cluster contains stocks with a median price of approximately $84.
  - Price change of about $4 in last 13 weeks and less volatility of about 1.8.
  - The Earning per share is -10.8 and P/E ratio is 42 , which is high(indication high growth) and P/B ration is -negative indicating the shares are undervalued.
  - Has the highest ROE among all clusters.
- **Cluster 2**:
  - This cluster contains stocks with a median price of approximately $91 .
  - Price change of about $16 in last 13 weeks and less volatility of about 1.8.
  - The Earning per share is 2.1 and P/E ratio is 46 , which is high(indication high growth) and P/B ration is positive 9 indicating the shares are almost correctly valued.
  - Has the highest Cash Ratio among all clusters.
- **Cluster 3**:
  - This cluster contains stocks with a median price of approximately $624 which is the highest average.
  - Price change of about $2.3 in last 13 weeks and less volatility of about 1.5.
  - The Earning per share is 19.2 and P/E ratio is 42 , which is high(indication high growth) and P/B ration is positive 41 indicating the shares are over valued.The hignest P/B ratio among all the clusters.

Insights – Clusters from Hierarchical clustering, contd..

- **Cluster 4**:
  - This cluster contains stocks with a median price of approximately $34.
  - Price change of about -$15 in last 13 weeks and volatility of about 2.8, the highest among all the clusters.
  - The Earning per share is -6 and P/E ratio is 76 , which is high(indication high growth) and P/B ration is positive 9 indicating the shares are almost correctly valued.
  - Has the second highest number of stocks among all clusters.
- **Cluster 5**:
  - This cluster contains stocks with a median price of approximately $46.
  - Price change of about $5 in last 13 weeks and less volatility of about 1, the lowest among all clusters.
  - The Earning per share is 3.4 and P/E ratio is 15 , which is good and P/B ration is -negative indicating the shares are undervalued.
  - Has the highest Net Income and Estimated Shares Outstanding among all clusters.
- **Cluster 6**:
  - This cluster contains stocks with a median price of approximately $327.
  - Price change of about $21 in last 13 weeks and volatility of about 2.
  - The Earning per share is .75 and P/E ratio is 400 , which is too high and P/B ration is -negative indicating the shares are undervalued.
  - Has the P/E ratio among all clusters.

K-means vs Hierarchical Clustering

K-means we took 8 clusters and Hierarchical clustering we took 7 clusters.

Looking at the clusters both the models reveal similar cluster information like P/E ratio matches between the K-means and Hierarchical clusters

Business Insights and Recommendations

- Cluster 6 from K-means is promising. Has less volatility and good P/E and P/B ratio.
- Cluster 5 from Hierarchical cluster is also promising. Has less volatility and good P/E and P/B ratio.

- Cluster 3 from K-means is stable cluster. Has less volatility and good P/E and P/B ratio.
- Cluster 0 from Hierarchical cluster is also stable cluster. Has less volatility and good P/E and P/B ratio.

- Cluster 0 from K-means has less volatility and high price change, indication the prices moves often
- Cluster 2 from Hierarchical cluster has less volatility and high price change, indication the prices moves often.

- Cluster 5 from K-means has negative price change.
- Cluster 4 from Hierarchical cluster has negative price change.

- Cluster 6 from Hierarchical clusters has high P/E and high price change may not be a good investment vehicle.
- Likewise, Cluster 7 from K-means has high P/E and high price change may not be a good investment vehicle.

greatlearning
Power Ahead

Happy Learning !