Trade & Ahead - Project

Clustering for Stock Recommendation

Contents

Business Problem Overview and Solution Approach

- ## Context of Business Problem

  The stock market has consistently proven to be a good place to invest in and save for the future, 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 and doing the same for a multitude of stocks to identify the right picks for an individual can be a tedious task. 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.

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

- ## Problem to tackle

  "Trade & Ahead" have provided data comprising stock price and some financial indicators for a few companies listed under the New York Stock Exchange. Our objective is analyzing the data, grouping the stocks based on the attributes provided, and sharing insights about the characteristics of each group.

- ## Financial implications

  Not able to cluster and find insights will lead–

  1. Business not able to provide proper suggestion for their customers diversified portfolio.
  2. Wrong suggestions will lead to customer dissatisfaction, and it will lead to business loss.

- ## How to use ML to solve the problem

  With a thorough analysis of the data and use of clustering methods will help group similar traits of data which would help create stock profiles for different segments and types of investors which will suit their needs and help create a diversified portfolio for overall growth of their investment.

Data Overview

- Data Dictionary

Variables	Description
Ticker Symbol	Unique identifier of a stock
Company	Name of the company
GICS Sector	Economic sector assigned to a company
GICS Sub Industry	Business operations of under GICS Sector
Current Price	Current stock price
Price Change	Percentage change in stock price last 13 weeks
Volatility	Standard deviation of the stock over the past 13 weeks
ROE	Financial performance
Cash Ratio	Cash reserve/Current liabilities(equivalent to cash)
Net cash flow	Difference of cash inflows and outflows
Net Income	Revenues minus expenses, interest, and taxes (in dollars)
Earnings Per share	Net profit/Shares outstanding
Estimated shares outstanding	Stocks held by shareholders
P/E ratio	Ratio of stock price to earning per share
P/B ratio	Stock price/Book value per share

Observations	Variables
340	15

Missing value counts	Duplicate counts
0	0

Notes:

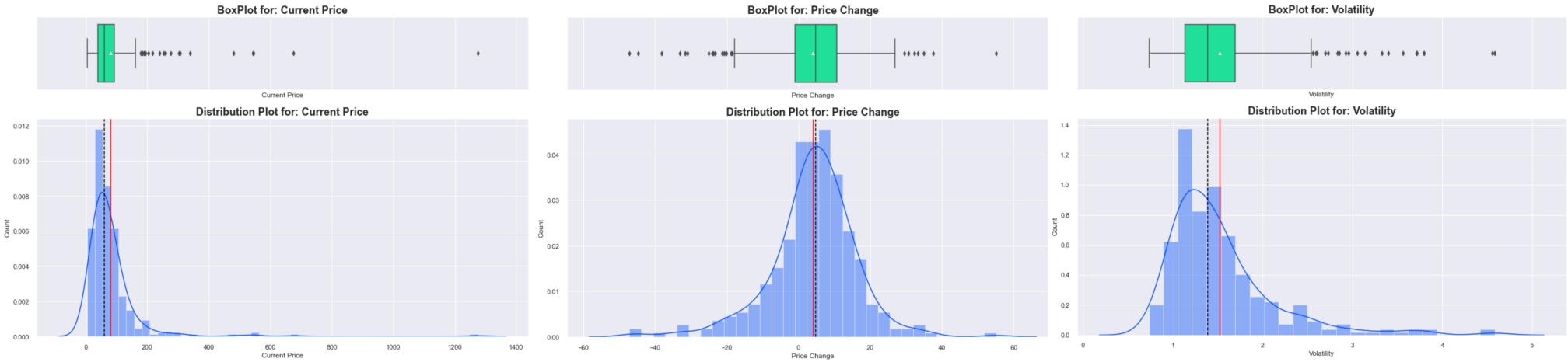
1. Dataset looked consistent with the data definition provided in data dictionary.
2. There are no duplicate values
3. There are no missing values across 15 variables

Brief Description of data manipulation:

1. Data has been scaled before applying clustering
2. Outliers are treated and both outliers treated, and non treated datasets has been explored

Exploratory Data Analysis – Univariate Analysis

- Distribution of numerical variables

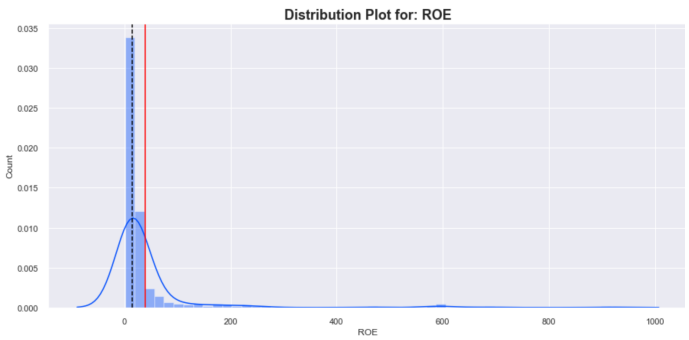


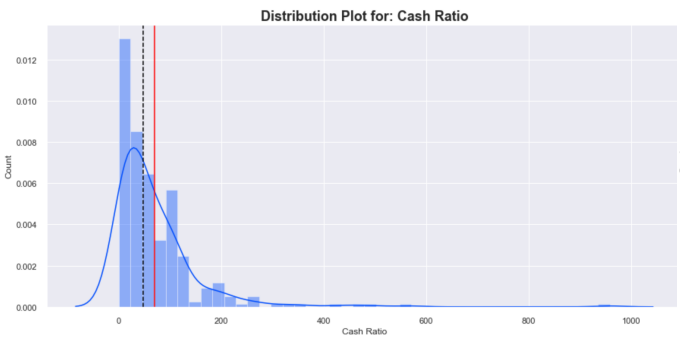
Notes:

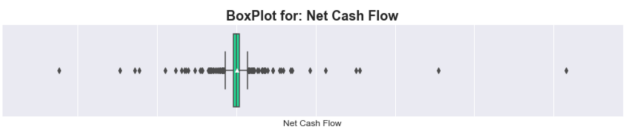
1. We can observe majority of stocks in dataset are below ~100 whereas highest stock price is close to ~1300
2. Distribution of the data in Price Change looks close to normal, with mean and median being close to each other.
3. Price change data is distributed between ~-60 to +65
4. Distribution is skewed to the right for Volatility, with data distributed from ~.7 to ~5.1

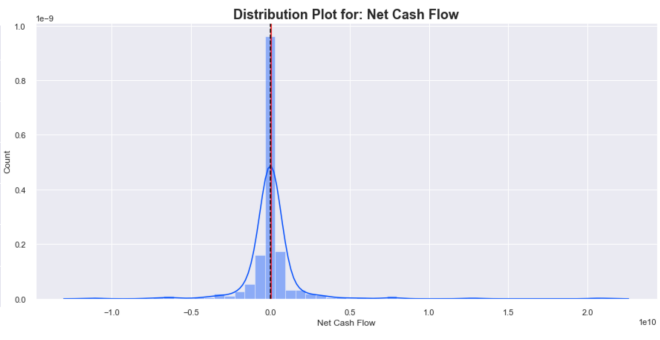
Exploratory Data Analysis – Univariate Analysis

- Distribution of numerical variables







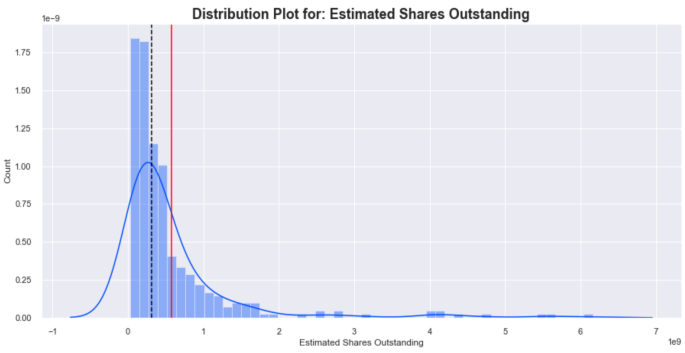


Notes:

1. Data is heavily right skewed in ROE, with data distributed from ~0 to ~1000 but majority is within ~150
2. Cash ratio shows data is skewed to the right and with outliers on the right. Majority of the data is below ~200
3. Distribution of net cash flow shows normally distributed, however has long tails on both sides.

Exploratory Data Analysis – Univariate Analysis

- Distribution of numerical variables

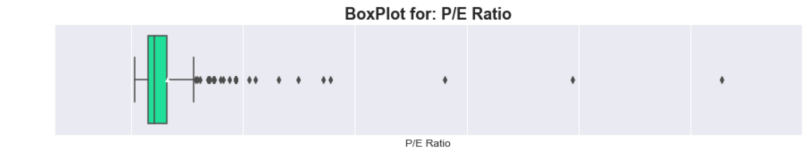


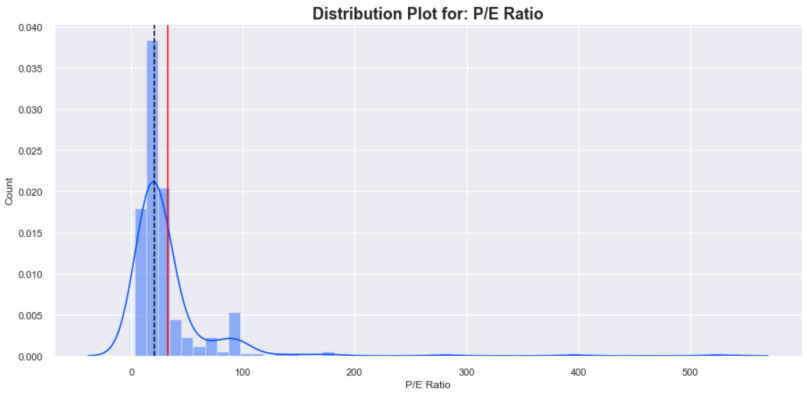
Notes:

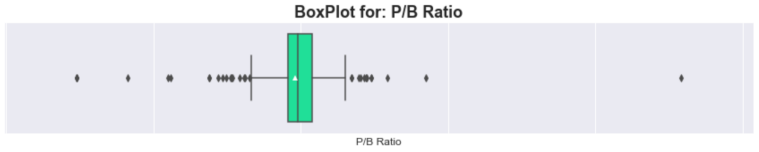
1. Data in Net income looks slightly right skewed and has long tails, which means data is clustered around center
2. Data in earnings per share also looks normally distributed but majority of the data are between ~-20 to ~+20 where is data is spread around ~-60 to ~+60
3. Data in estimated shares outstanding shows heavily right skewed

Exploratory Data Analysis – Univariate Analysis

- Distribution of numerical variables





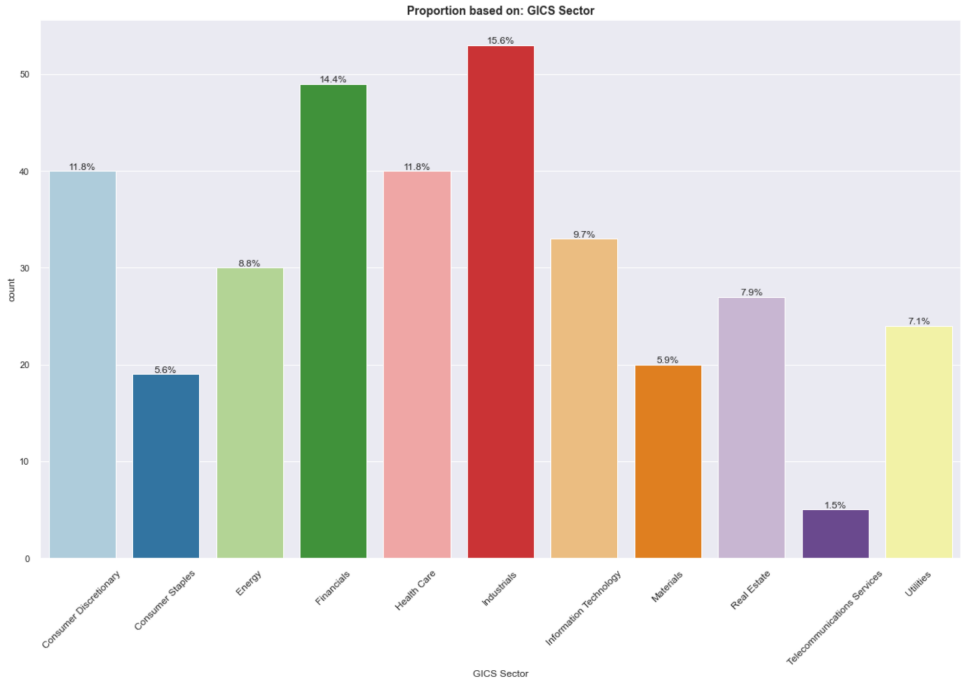


Notes:

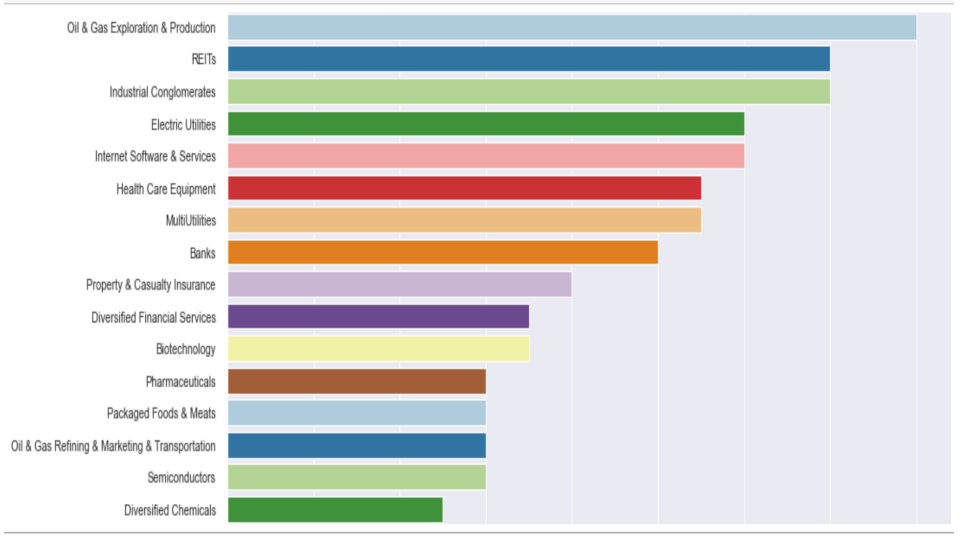
1. Data in P/E ratio is right skewed with majority of the data is below ~100 whereas the data is spread from 0 - ~600
2. P/B ratio shows data are distributed close to normal but has long tails on both sides. Majority of the data are within -50 to +50 whereas the data is spread from -100 to +100

Exploratory Data Analysis – Univariate Analysis

- Categorical variable exploration



Top few GICS sub-Industry

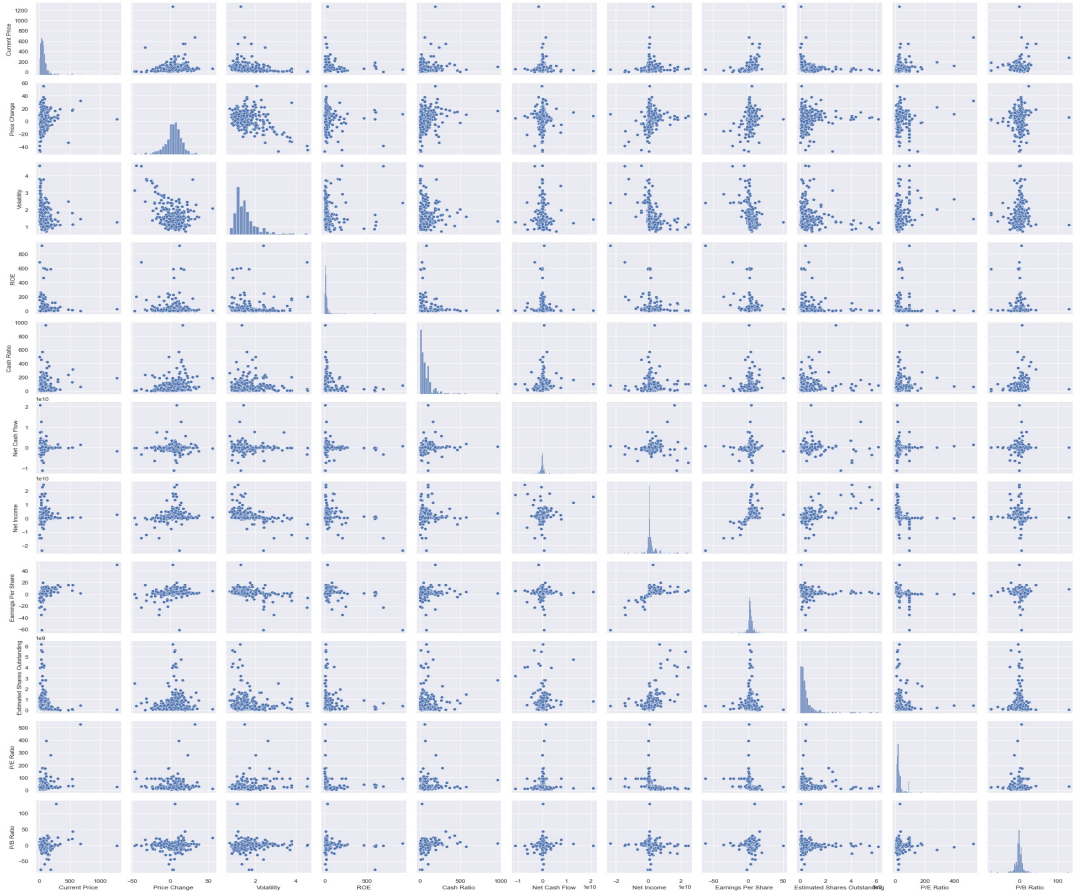


Notes:

1. Industrial sector has the highest observations with ~15.6%
2. Financials and Consumer Discretionary are the next two sectors which has higher observations
3. Telecommunications Services shows the lowest observations in dataset
4. Oil & Gas Exploration and Production shows the highest proportion in dataset with 16 observations.
5. REITs, Industrial Conglomerates are next two highest sub sector with 14 observations.

Exploratory Data Analysis – Bivariate Analysis

Pair plot for all numerical variables

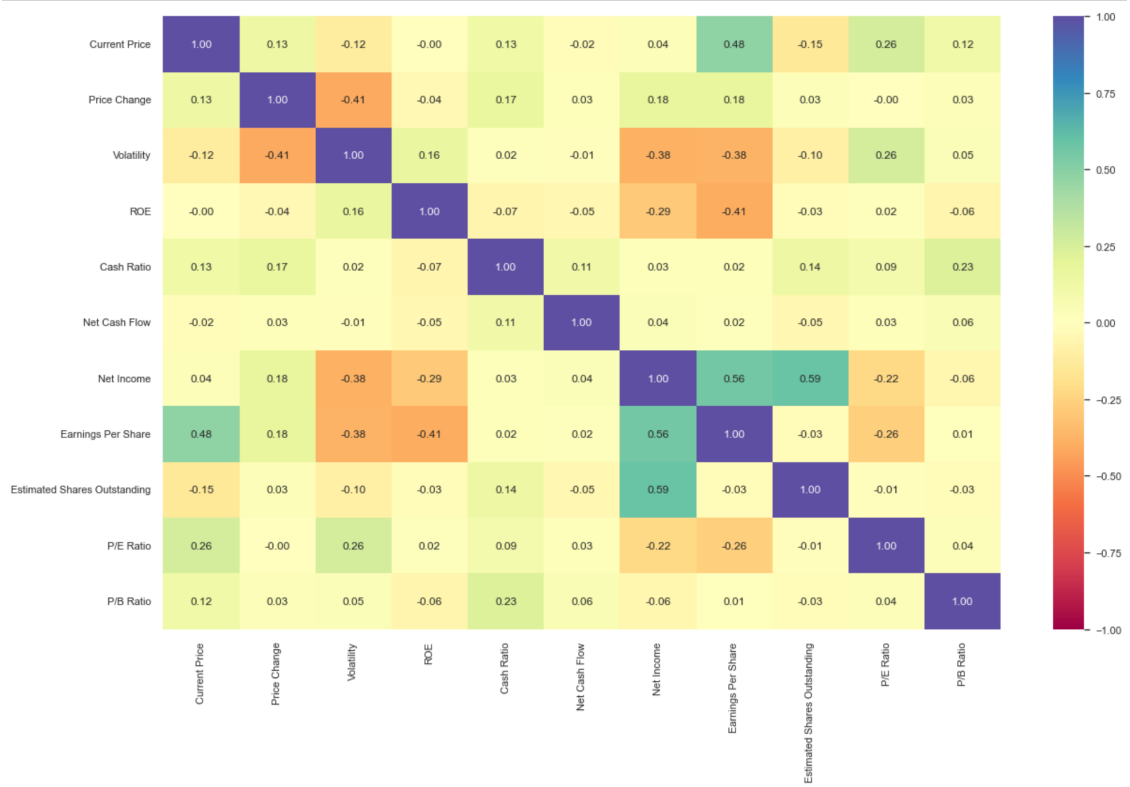


Notes:

1. We can't observe any strong relationship among the variables except the few noted below which visually shows some relationships
2. Current price and Price change doesn't show strong relation ship, however from the plot we can observe for some it has positive linear relationship
3. Price change and volatility shows negative linear relationship, in another words with increasing volatility price changes decreases
4. Estimated shares outstanding and volatility shows negative linear relationship, which mean with increase in outstanding shares volatility decreases
5. Net Income and Earning per share shows positive linear relationship, meaning with increase in Net income the earning per share increases
6. Net Income and Estimated shares outstanding shows positive linear relationship

Exploratory Data Analysis – Bivariate Analysis

Heatmap for all numerical variables



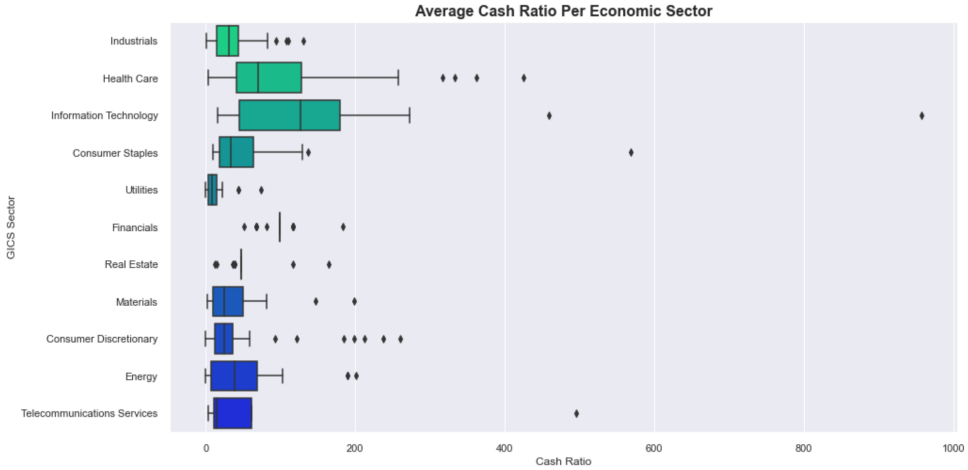
Notes:

1. Volatility and Price change shows negative correlation
2. Net Income shows positive correlation with Earnings per share and Estimated shares outstanding
3. Earnings per share and Current price shows some positive correlation
4. Volatility and Earnings per share shows some negative correlation
5. Earnings per share also shows negative correlation with ROE

Exploratory Data Analysis – Bivariate Analysis

Notes:

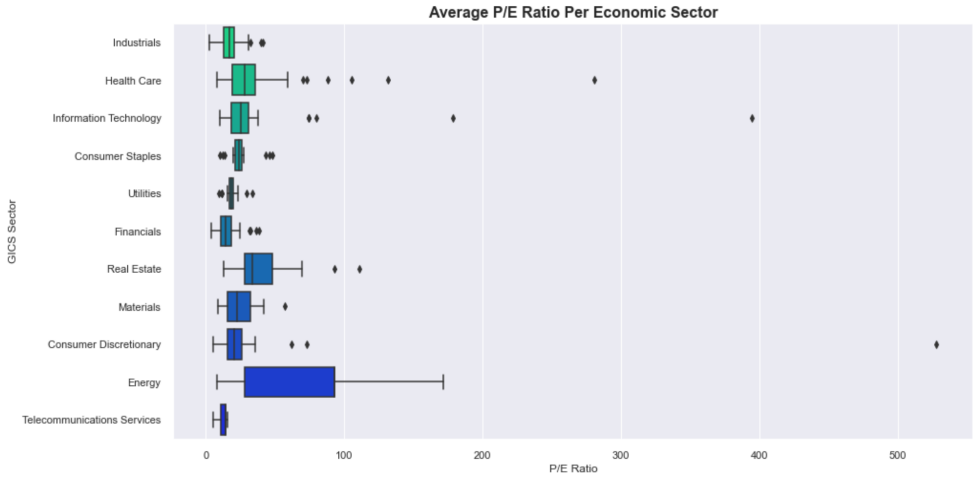
1. We can observe that price changes varies by economic sector, from the above box plot we can see some economic sectors have on average positive/increase in price, some has negative/decrease in price
2. Health care and Materials shows the maximum positive price change.



Notes:

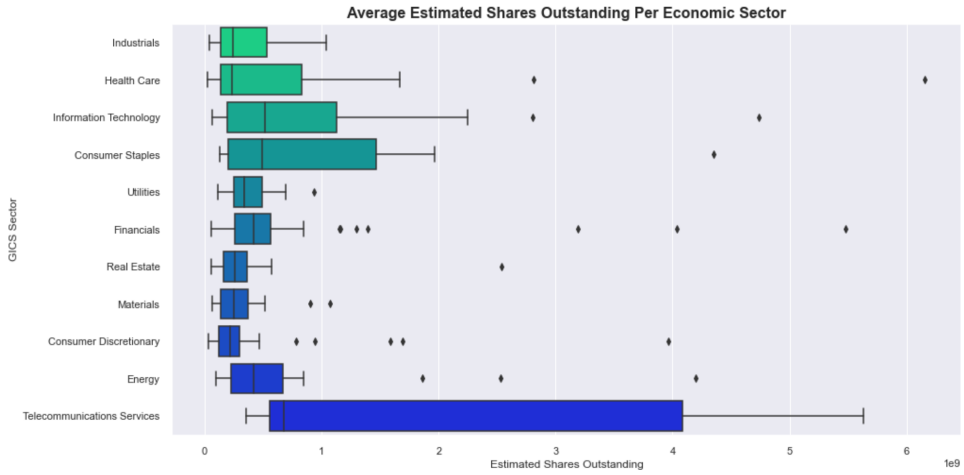
1. We can observe on average Information Technology sector has the highest cash ratio.
2. Utilities sector shows the lowest average cash ratio
3. Health care sector shows the second highest cash ratio on average

Exploratory Data Analysis – Bivariate Analysis



Notes:

1. Energy sector shows the highest P/E ratio on average.
2. Real Estate sector shows the second highest P/E ratio.
3. On average all other sectors have P/E ratio closer to each other, with Telecommunication Services being the lowest

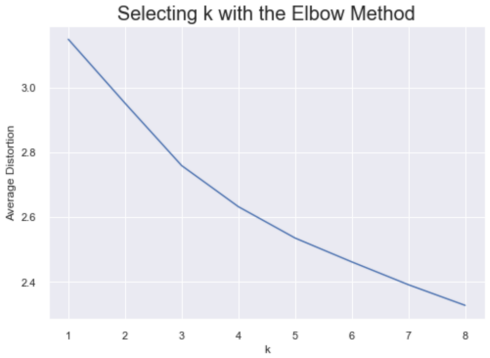


Notes:

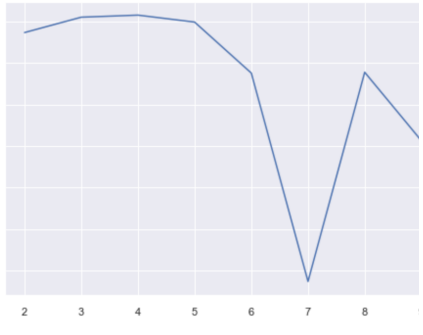
1. Telecommunication services sector shows it has highest estimated shares outstanding on average
2. Consumer Discretionary shows the lowest average estimated shares outstanding

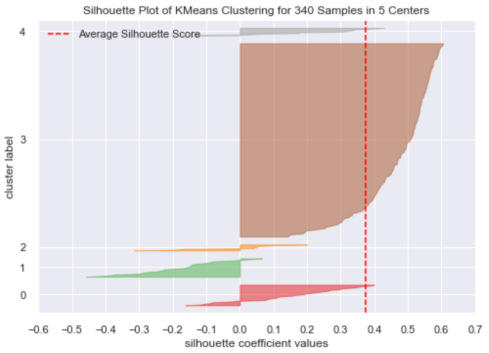
Clustering – K-means

Optimal – K value selection



Silhouette score

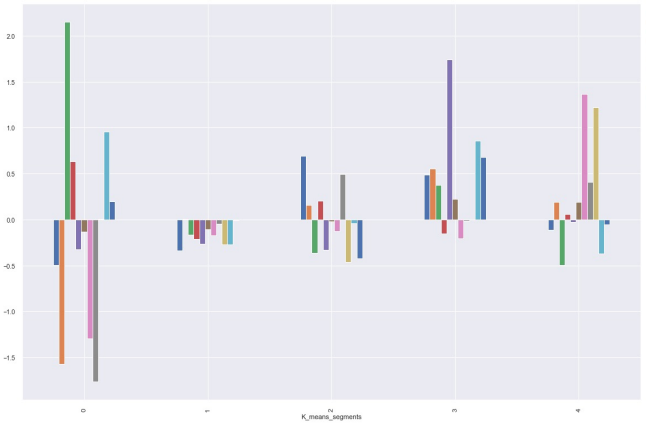


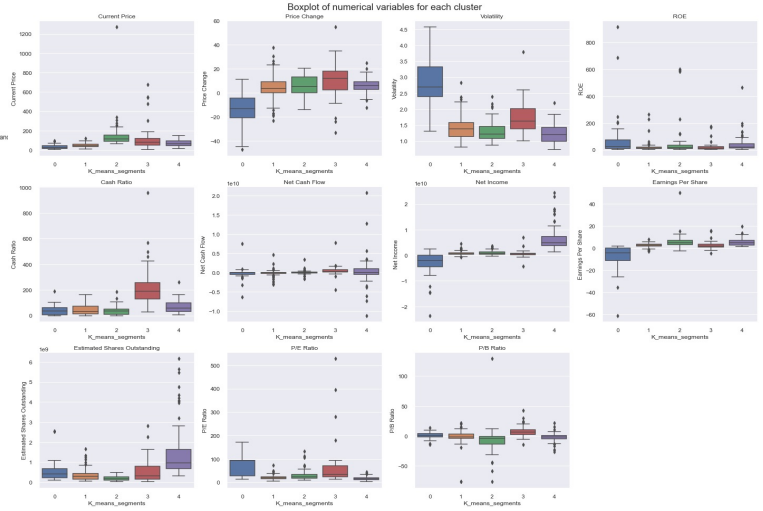


Notes:

1. From the Elbow method we can see optimal value for K is 5
2. Silhouette score also indicates 4/5 as optimal value for K.

Cluster Profiling



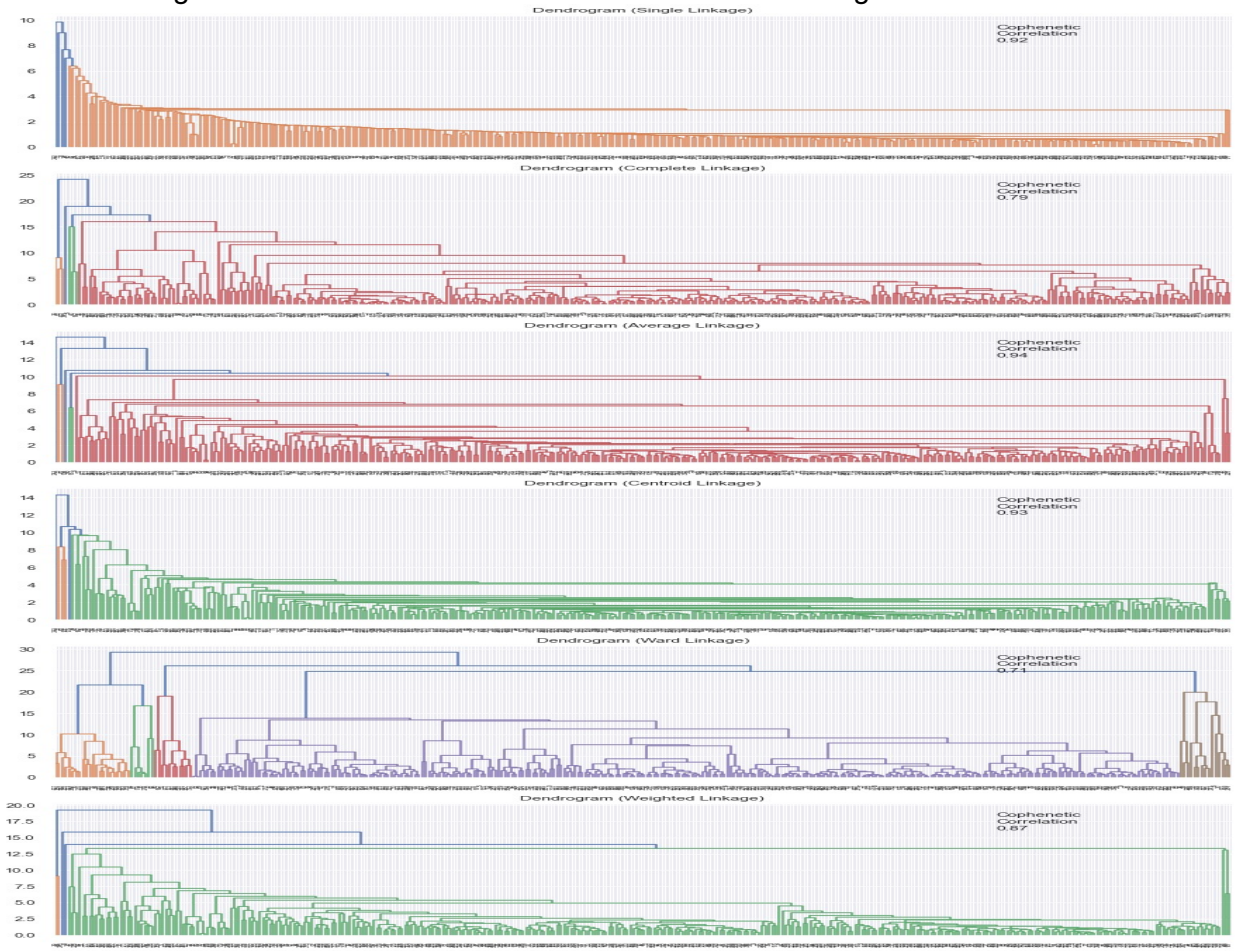


Notes:

1. Detailed Cluster profiling was done based on the obtained clusters in K-Means method.
2. The bar plot and box plot shows how each variable's average value varies across clusters
3. Stocks in each cluster –
   Cluster 0 – 29
   Cluster 1 – 142
   Cluster 2 – 70
   Cluster 3 – 41
   Cluster 4 - 58

Clustering – Hierarchical Clustering

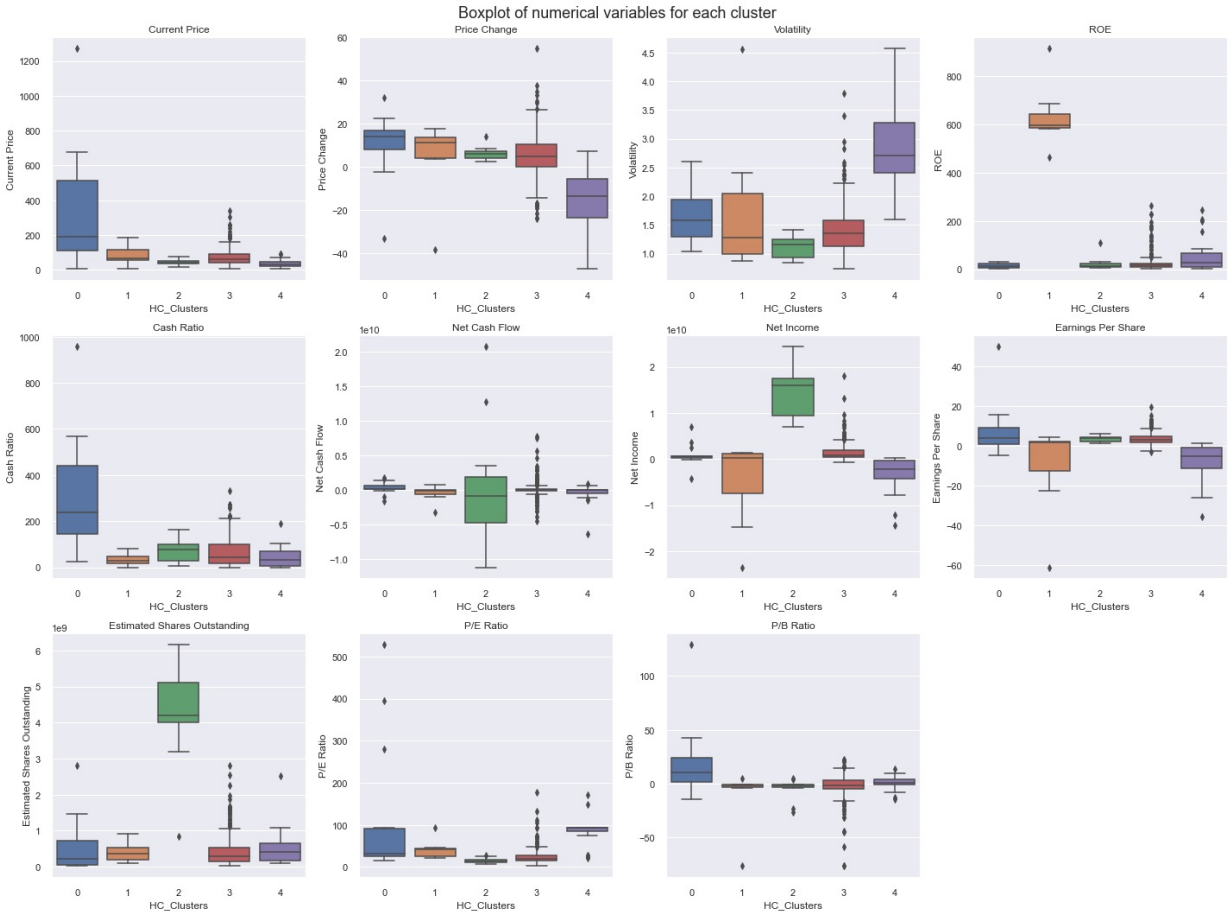
Dendrogram based on Euclidean distance and different Linkage method



Notes:

1. The highest cophenetic correlation was from distance method as Euclidean and Linkage as Average
2. All the dendrograms shows data clustered to left side except Ward linkage method
3. Ward linkage method shows best spread out dendrogram, however has low cophenetic correlation

Clustering – Hierarchical Clustering



Notes:

1. Even though Average linkage method gave the highest cophenetic correlation when we obtained clusters using Agglomerative clustering with the same linkage clusters were not having variability, Cluster 0 got 334 stocks out of 340 so we used Ward method.

2. Ward method provided better clusters where we have better variability than average linkage.

3. Detailed cluster profiling was done using the boxplot, where each variables were explored how they varied across clusters

4. Clusters obtained using ward linkage have stocks count as below
   Cluster 0 – 15
   Cluster 1 – 7
   Cluster 2 – 11
   Cluster 3 – 285
   Cluster 4 – 22

Comparison and Conclusion - K-means vs Hierarchical Clustering

1. Both algorithm K-Means and Agglomerative took almost similar time, since the dataset wasn't big it was very quick
2. From K-means we have seen that optimal number of clusters which provides decent variability was 5, for Hierarchical clustering we have noticed that by ward linkage method with best spread out dendrogram 5 seems to be an optimal number of clusters with decent variability
3. From both clustering methods we can see one of the cluster is coming with majority number of observations/tickers.
4. For K means cluster wise (0-4) number of observations are 29, 142, 70, 41, 58 respectively, whereas for Hierarchical clustering cluster wise (0-4) number of observations are 15, 7, 11, 285, 22, respectively.
5. We can see K-means is providing better variability whereas in hierarchical clustering almost ~83% of data are in one cluster and only rest ~17% is forming other clusters.
6. With K means the current price shows highest for cluster 2 which has around 70 observations, whereas in hierarchical clustering the current price shows highest for cluster 0 which has only 15 observations.
7. For price change K- Means shows cluster 3 has the highest price changes which has 41 observations whereas in hierarchical clustering it shows cluster 0 which has only 15 observations.
8. In K-means cluster 0 shows that it has the highest volatility which has 29 observations and hierarchical clustering shows the highest volatility i cluster 4 which has 22 observations, for this feature both the clustering shows comparable counts of observations
9. For cash ratio K- Means shows cluster 3 has the highest price changes which has 41 observations whereas in hierarchical clustering it shows cluster 0 which has only 15 observations.
10. For Net income K-means shows cluster 4 has the highest net income with 58 observations whereas hierarchical clustering shows cluster 2 which only has 11 observations
11. With K means the earnings per share shows highest for cluster 2 which has around 70 observations, whereas in hierarchical clustering the earnings per share shows highest for cluster 0 which has only 15 observations.
12. For estimated shares outstanding K-means shows cluster 4 has the highest shares outstanding with 58 observations whereas hierarchical clustering shows cluster 2 which only has 11 observations
13. For P/E ratio K-means shows cluster 0 has the highest P/E ratio which is 29 observations, whereas hierarchical clustering shows cluster 4 which has 22 records, for this clusters are more comparable in both methods.

With the above comparison since K-means clustering has provided the best variability among its clusters, we will prepare the recommendation and insights based on that

Business Insights and Recommendations

- **Cluster 0 and Cluster 3:** These are the stocks which has higher volatility, with average price change is on the higher side, P/E ratio is on the higher side, but their net income is lowest or even in red(negative) and their current price is on the low to mid categories. **These stocks are more suited for quick short term volume trading, as their price change is on higher side it will increase the short-term gain. So, these stocks should be recommended for Day traders or short-term investing. Investors and traders should be made aware their nature of volatility which could potentially risk their investment. Low risk traders/Investors should not invest in these stocks**

- **Cluster 4:** These are the stocks which has very less volatility, though average price change is on the lower side, it has higher number of shares outstanding and net income is high and earning per share is slightly higher than others. Current price of these stocks are still low. **These stocks are low risk, low cost with high potential for large growth in long term future, should be recommended to Investors who are looking to invest early in some low-cost stocks which has higher potential of growth and looking for long term investments**

- **Cluster 2:** These are the stocks which are high performing, with current price on higher side, and with very less volatility and earning per share is highest among any other stocks. **These are the stocks which are the best performing large cap stocks which shows consistent growth and returns, should be recommended to Investors who are looking for investments with consistent returns and possibly don't mind average long-term growth**

- **Cluster 1:** These are the stocks which are having current price as low, even though volatility is low, their net income, estimated shares outstanding and earning per share is low. **These stocks are underperforming stocks with no signs of any potential growth, so Investors should be recommended to not invest in these stocks, as they doesn't have any potential of getting any returns in short or long term at this point of time.**

greatlearning
Power Ahead

Happy Learning !