Trade&Ahead

Business Case

Background

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

Objective

To analyze the data, grouping the stocks based on the attributes provided, and sharing insights about the characteristics of each group.

Data Information

The data contains information about the business problem

Variable	Description	Type of Variable
Ticker Symbol	An abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market	Object
Company	Name of the company	Object
GICS Sector	The specific economic sector assigned to a company by the Global Industry Classification Standard (GICS) that best defines its business operations	Object
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	Object
Current Price	Current stock price in dollars	Float64
Price Change	Percentage change in the stock price in 13 weeks	Float64
Volatility	Standard deviation of the stock price over the past 13 weeks	Float64
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)	Int64
Cash Ratio	The ratio of a company's total reserves of cash and cash equivalents to its total current liabilities	Int64
Net Cash Flow	The difference between a company's cash inflows and outflows (in dollars)	Int64
Net Income	Revenues minus expenses, interest, and taxes (in dollars)	Int64
Earnings Per Share	Company's net profit divided by the number of common shares it has outstanding (in dollars)	Float64
Estimated Shares Outstanding	Company's stock currently held by all its shareholders	Float64
P/E Ratio	Ratio of the company's current stock price to the earnings per share	Float64
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)	Float64

Observations	Variables
340	15

Manipulations to Raw Data:

1.Object variables were converted to Category

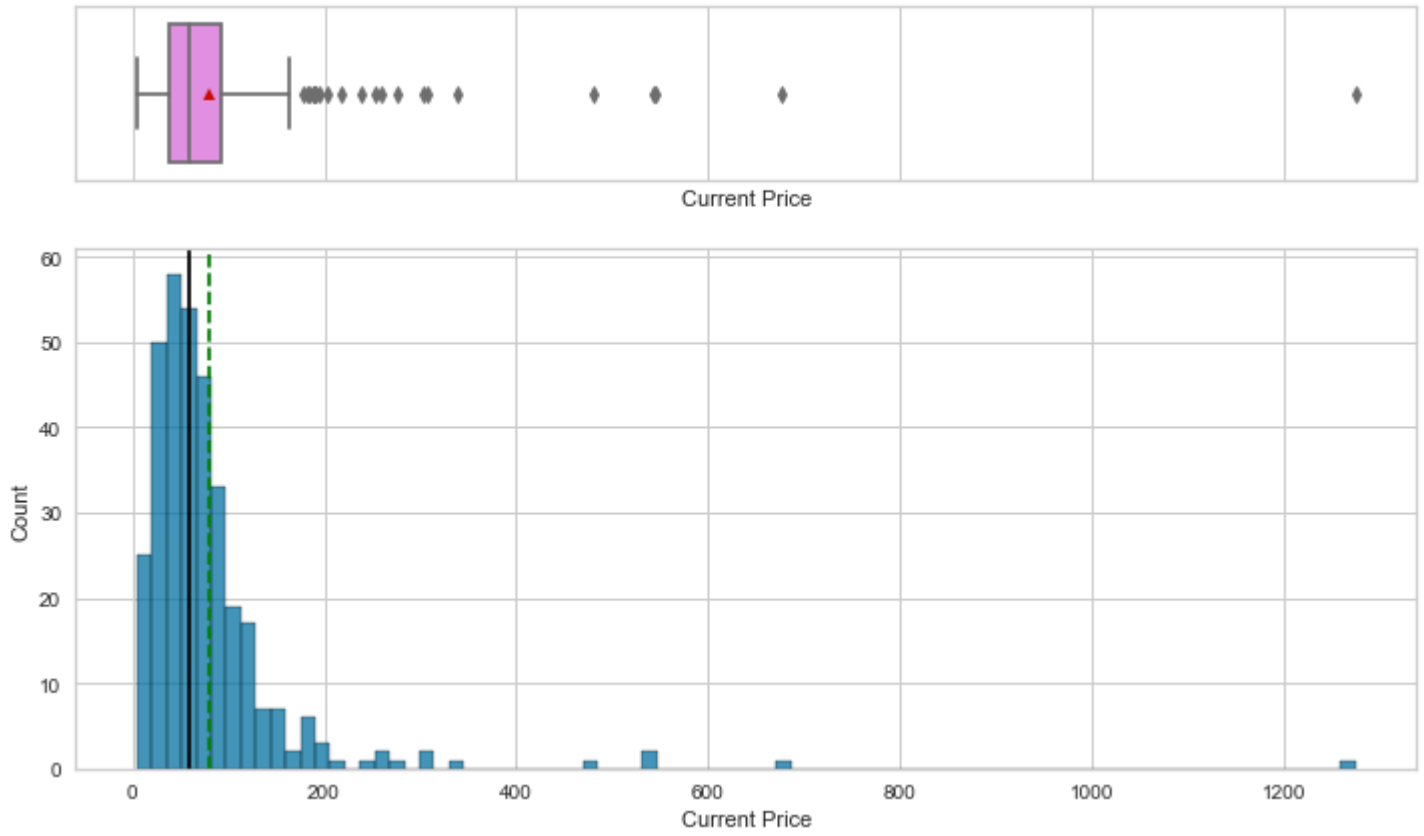
Exploratory Data Analysis – Current Price

This data contains the Current Price

Observations:

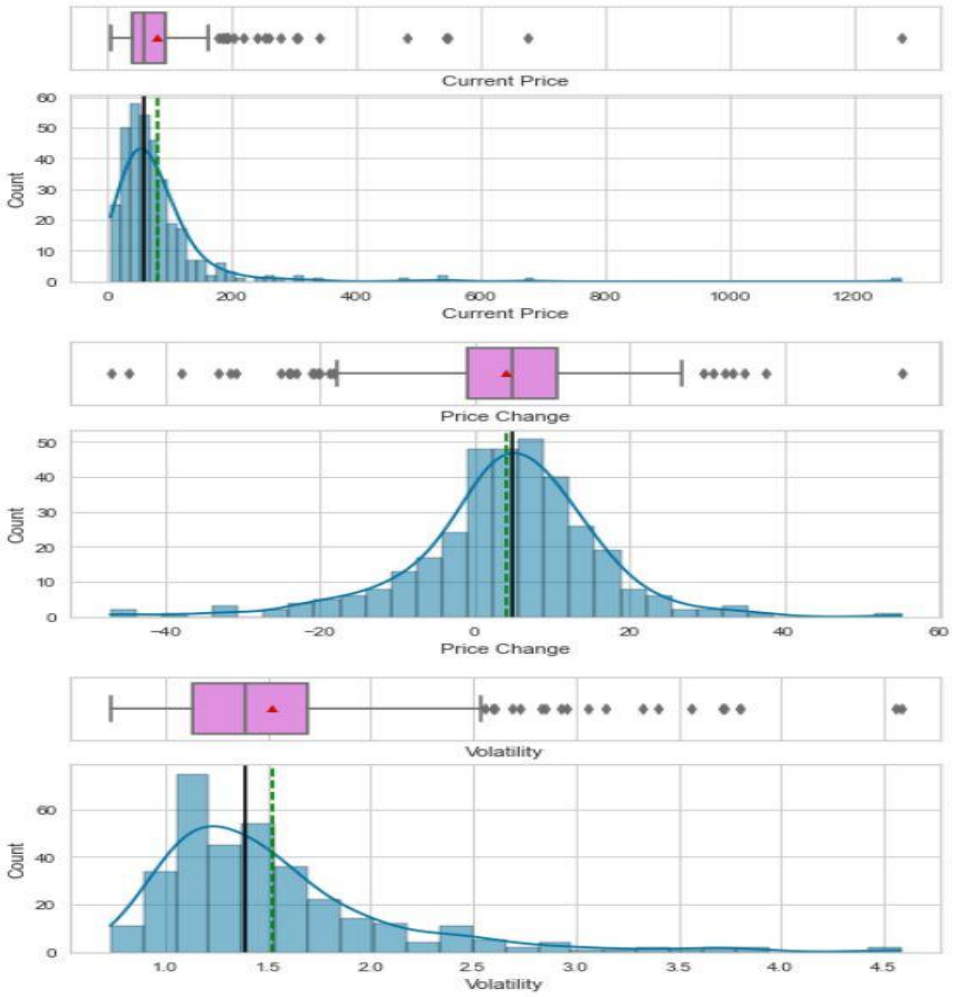
1. The prevailing wage is heavily left skewed
2. The mean and median of the price are very close
3. Most of the prices are clustered around the $100 range
4. The range of the prices is very wide with the highest being around $1250

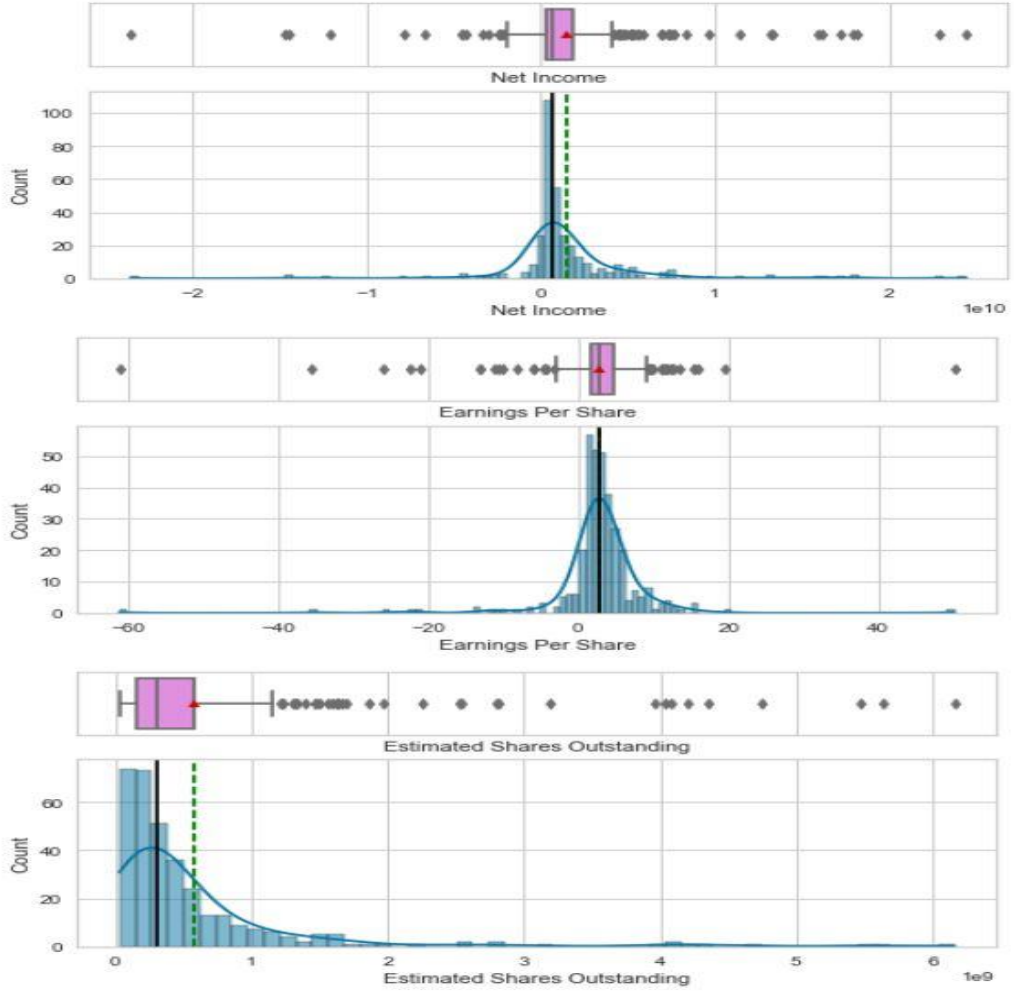
Current Price



Exploratory Data Analysis – Variables

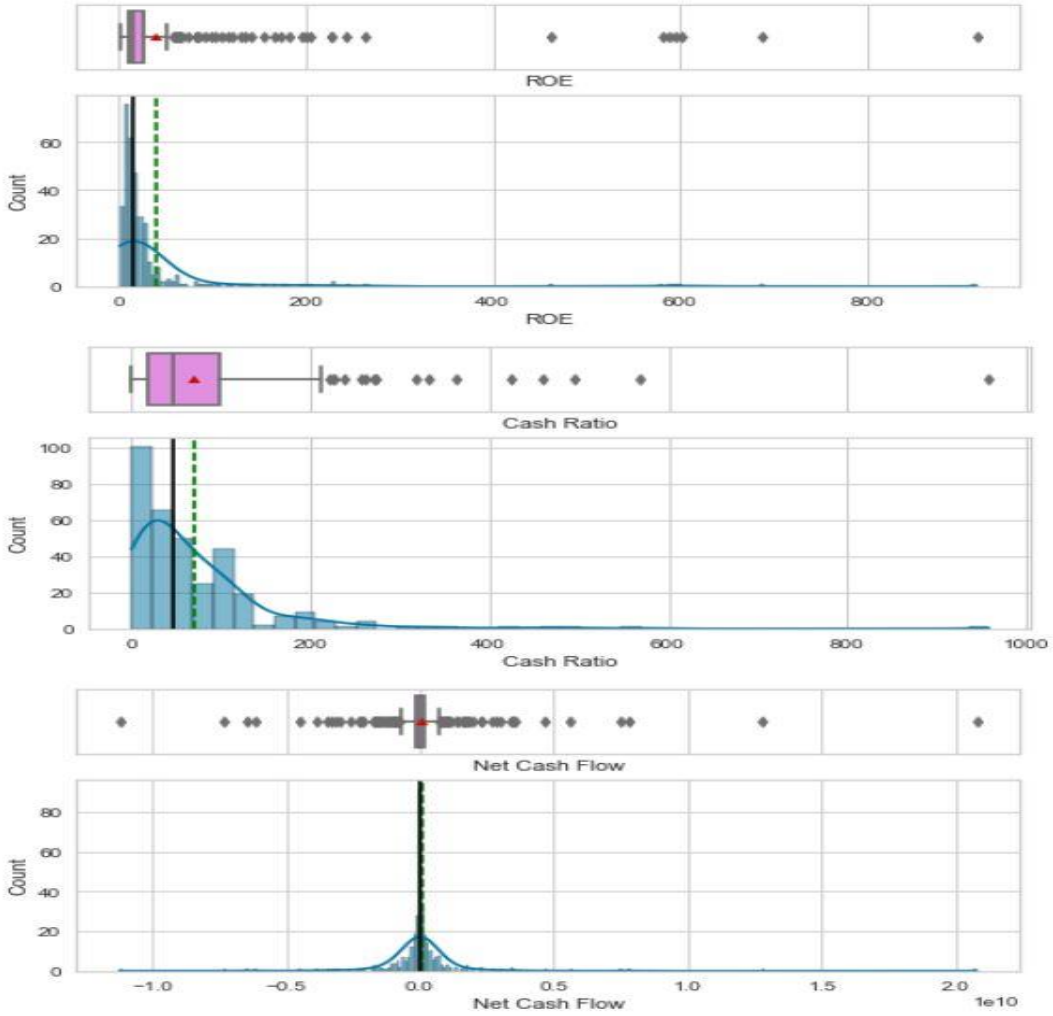
This data contains the different variables for the dataset

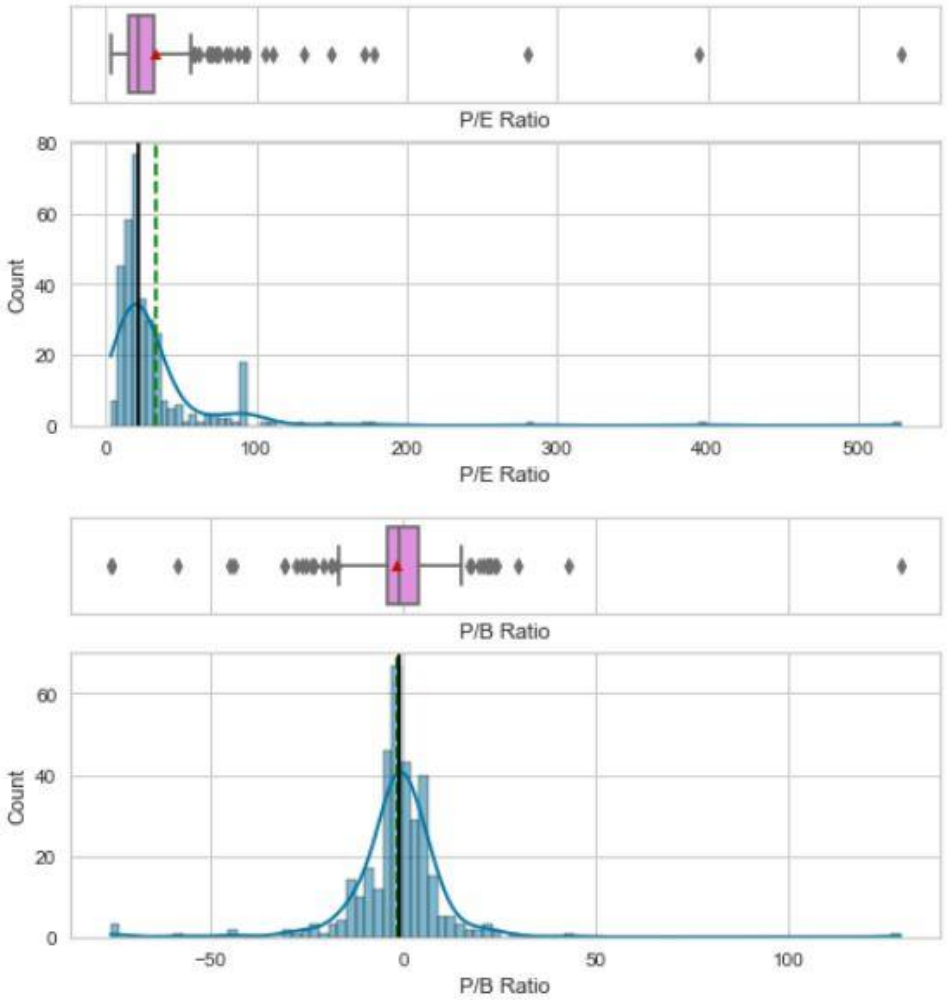




Exploratory Data Analysis – Variables

This data contains the different variables for the dataset

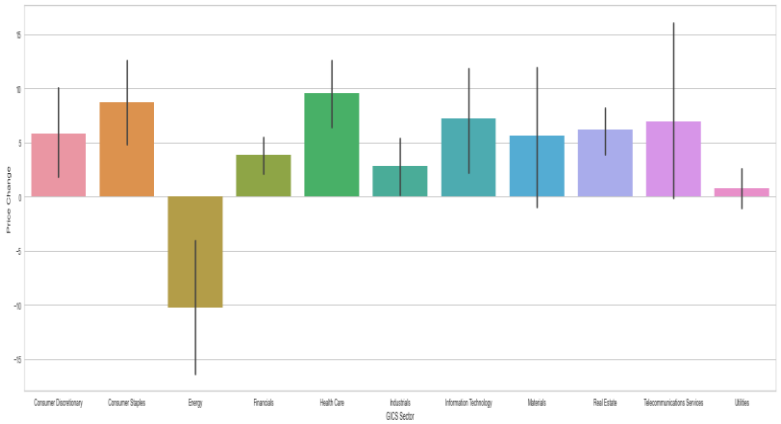




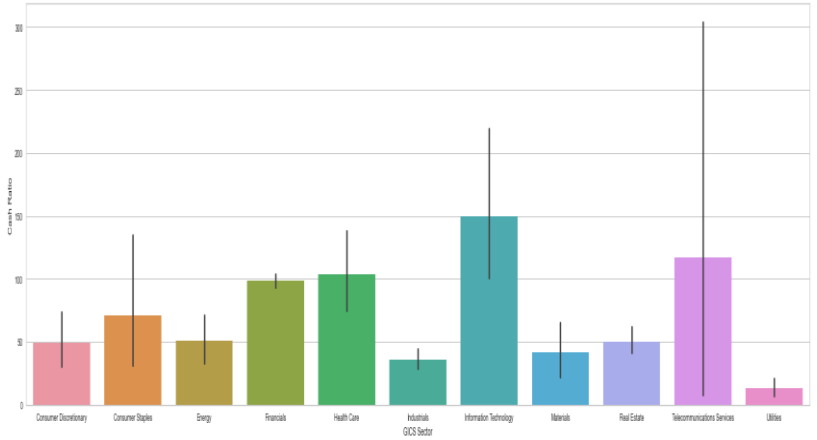
Exploratory Data Analysis – General Information

General Information about the Data

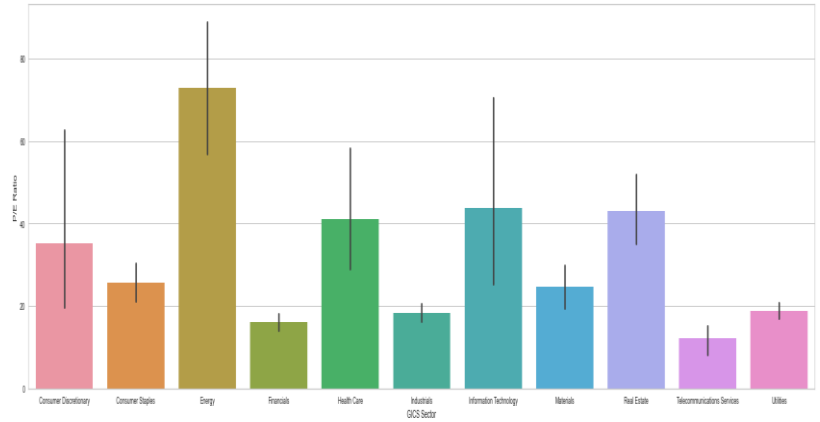
Economic Sector wrt Price Increase



Economic Sector wrt Cash Ration



Percentage split of Employee Region



Observations:

1. The Consumer Staples sector has the highest price increase
2. The Energy sector is the only sector with a negative price increase
3. The Utilities sector has the lowest price increase
4. On average, most of the sectors had an increase in price change

Observations:

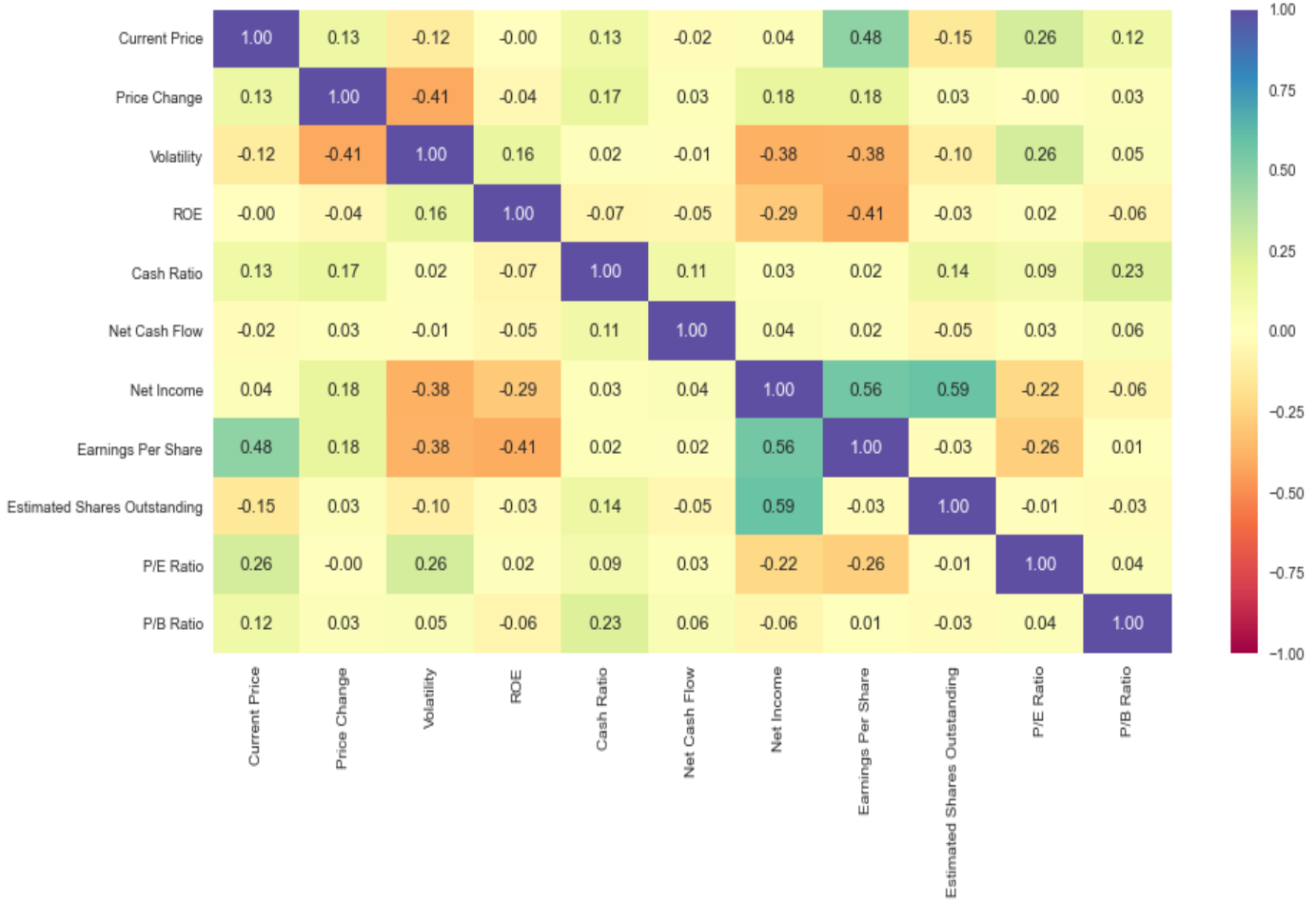
1. Information Technology has the highest cash ratio
2. Utilities has the lowest cash ratio

Observations:

1. Energy has the highest P/E ratio
2. Telecommunications Services has the lowest P/E ratio

Exploratory Data Analysis – General Information

Correlation Matrix

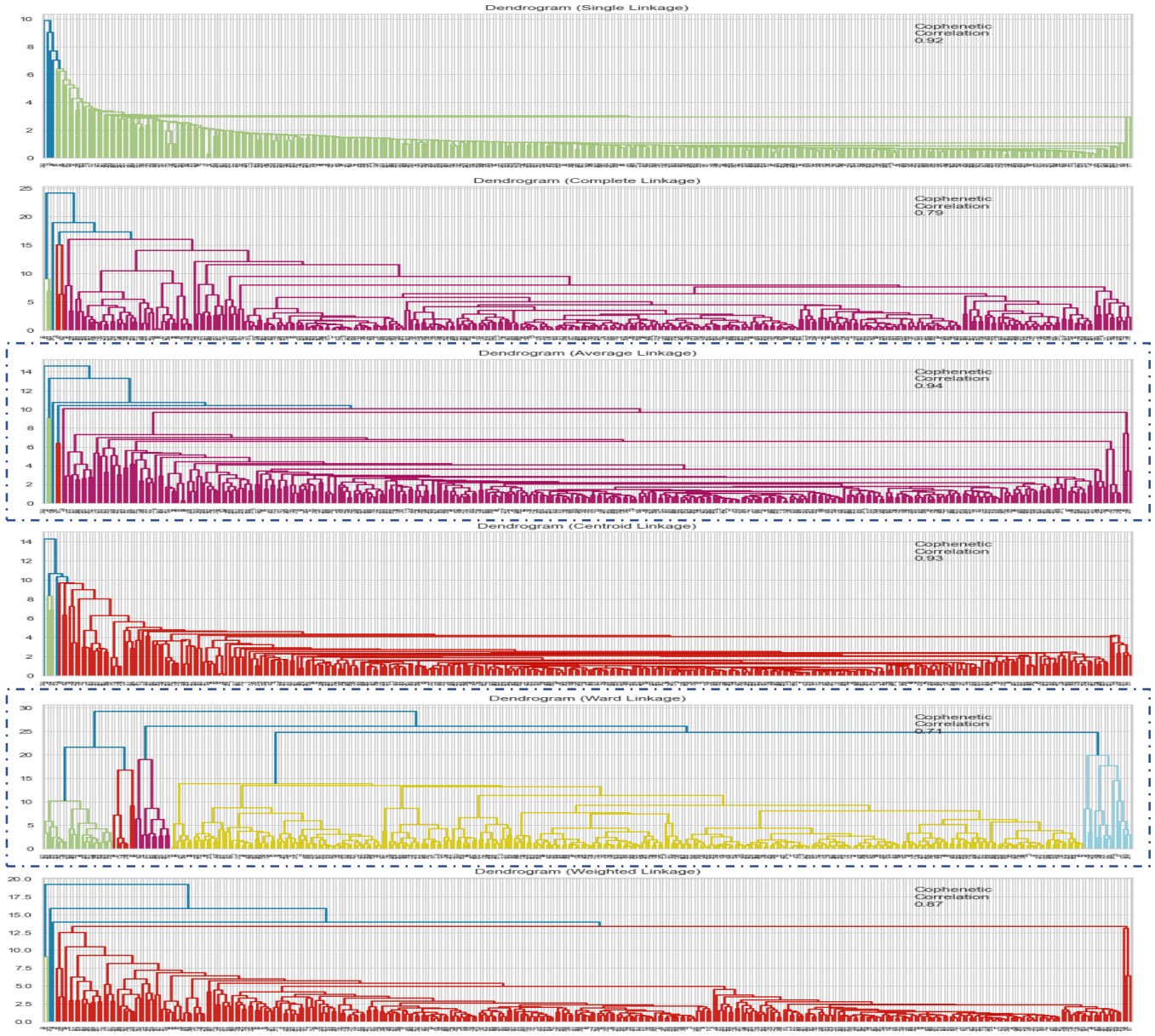


Observations:

1. There are no strong correlations

2. However, the variables with the strongest correlations are:
   a) *Earnings per share* and *Current Price*
   b) *Net Income* and *Estimated Shares Outstanding*
   c) *Net Income* and *Earnings per Share*

Hierarchical Clustering – Summary



Average Linkage

- We see that there are 2 clusters of one company, 2 cluster of two companies, and all the other countries are grouped into another cluster.
- This clustering does not look good as the clusters do not have enough variability.

Ward Linkage

- There is more variation in the clustering.
- We see that there are two cluster 20-plus companies, two clusters of less than 10 companies, and one cluster of 275 companies

Based on these factors, Ward Linkage is what is going to be used for the analysis.

Comparison of Results

The two methods used for the analysis were the K-Means Clustering and the Hierarchical Clustering. Below are the results

K_means_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	72.633251	5.118334	1.383473	34.800000	52.905455	-14092741.818182	1473201570.909091	3.631473	430196050.060400	23.859540	-3.389063	275
1	44.470001	11.397804	2.405408	917.000000	80.000000	698000000.000000	-23528000000.000000	-61.200000	384444444.400000	93.089287	4.970809	1
2	48.103077	6.053507	1.163964	27.538462	77.230769	773230769.230769	14114923076.923077	3.958462	3918734987.169230	16.098039	-4.253404	13
3	34.516154	-17.260440	2.938570	76.076923	48.769231	-261407346.153846	-3098292307.692307	-7.388077	481614743.486154	75.734798	1.646607	26
4	327.006671	21.917380	2.029752	4.000000	106.000000	698240666.666667	287547000.000000	0.750000	366763235.300000	400.989188	-5.322376	3
5	225.945603	12.362269	1.743078	28.545455	299.500000	759552909.090909	1031240545.454545	6.590455	587974943.437273	47.839183	16.875739	22

HC_Clusters	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_segments
0	563.992491	17.235667	1.839399	10.250000	125.500000	105834000.000000	853500250.000000	13.085000	287806305.492500	307.105187	-4.254889	4
1	84.355716	3.854981	1.827670	633.571429	33.571429	-568400000.000000	-4968157142.857142	-10.841429	398169036.442857	42.284541	-11.589502	7
2	152.566666	14.908086	1.769506	24.434783	281.913043	1747221304.347826	1866621956.521739	3.802174	759756952.867391	38.674023	16.027369	23
3	72.421687	4.563230	1.403434	25.218182	55.014545	72801872.727273	1572467469.090909	3.728564	445003946.148763	24.188244	-2.966949	275
4	36.440455	-16.073408	2.832884	57.500000	42.409091	-472834090.909091	-3161045227.272727	-8.005000	514367806.201818	85.555682	0.836839	22
5	46.672222	5.166566	1.079367	25.000000	58.333333	-3040666666.666667	14848444444.444445	3.435556	4564959946.222222	15.596051	-6.354193	9

The Clustering methods will be evaluated on the following criteria:

<u>Time Spent on Technique</u>

The K Means clustering took less time to run than the Hierarchical clustering technique.

<u>Cluster Variation</u>

The variation is very close as both have the same number of clusters

<u>Common Results</u>

As mentioned above, there is a lot of similarity between the two methods

However, considering all these factors, the results from the Hierarchical clustering will be used for the final analysis.

Conclusion

Based on the analysis done on the data, these are the groups with the highest amount of data. Below is a short summary of them:

Cluster 2

There are 23 companies in this cluster.
They have the second highest average share price with the second highest positive price change.
They have an average ROE with the highest net cash flow indicating that the companies are doing very well for their share holders as indicated by the second highest earning per share.

Cluster 3

There are 275 companies in this cluster.
They have the relatively high average share price with a positive price change and a normal volatility.
They have a comparatively high ROE with a positive net cash flow indicating that the companies are well as evidenced by the positive earnings per share.

Cluster 4

There are 22 countries in this cluster.
They have the lowest average share price with a negative price change.
They have the second highest ROE with a negative net cash flow indicating that the companies are not making enough cash and have highest costs than their revenue. They also have a negative earnings per share

Recommendation

Based on the analysis, these are the following recommendations that can help the business determine what cluster should be used for investment purposes:

- Though none of the clusters chosen as representations have the highest *Current Price*, for investment purposes, **Cluster 2** has the best chance of returning the investment as evidenced by the various factors such as it having the second highest *average share price*, the highest *net cash flow* and the second highest *earnings per share*.
- As there are 23 companies in the cluster, it will provide a wide enough range for a diversified enough portfolio if one should so wish.
- The next best cluster to look at is **Cluster 3** based on the above-mentioned factors