Trade&Ahead Presentation

Business Problem Overview and Solution Approach

Context

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.

Business Objective

Objective

Trade&Ahead is a financial consultancy firm who provide their customers with personalized investment strategies. They have hired you as a Data Scientist and provided you with data comprising stock price and some financial indicators for a few companies listed under the New York Stock Exchange. They have assigned you the tasks of analyzing the data, grouping the stocks based on the attributes provided, and sharing insights about the characteristics of each group.

Data 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)

Data Overview:

#	Column	Non-Null Count	Dtype
0	Ticker Symbol	340 non-null	object
1	Security	340 non-null	object
2	GICS Sector	340 non-null	object
3	GICS Sub Industry	340 non-null	object
4	Current Price	340 non-null	float64
5	Price Change	340 non-null	float64
6	Volatility	340 non-null	float64
7	ROE	340 non-null	int64
8	Cash Ratio	340 non-null	int64
9	Net Cash Flow	340 non-null	int64
10	Net Income	340 non-null	int64
11	Earnings Per Share	340 non-null	float64
12	Estimated Shares Outstanding	340 non-null	float64
13	P/E Ratio	340 non-null	float64
14	P/B Ratio	340 non-null	float64

Data is comprised of 15 columns and 340 rows with no missing values. The numerical columns are integer and float types with 4 object columns.

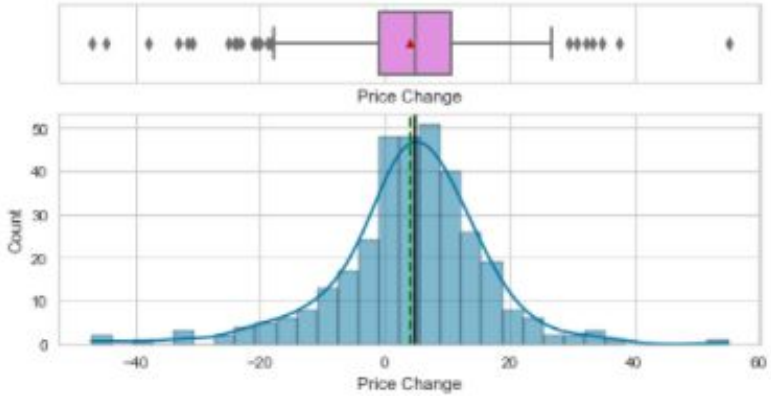
Data Overview Continued:

	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	340.000000	340.000000	340.000000	340.000000	340.000000	3.400000e+02	3.400000e+02	340.000000	3.400000e+02	340.000000	340.000000
mean	80.862345	4.078194	1.525976	39.597059	70.023529	5.553762e+07	1.494385e+09	2.776662	5.770283e+08	32.612563	-1.718249
std	98.055086	12.006338	0.591798	96.547538	90.421331	1.946365e+09	3.940150e+09	6.587779	8.458496e+08	44.348731	13.966912
min	4.500000	-47.129693	0.733163	1.000000	0.000000	-1.120800e+10	-2.352800e+10	-61.200000	2.767216e+07	2.935451	-76.119077
25%	38.555000	-0.939484	1.134878	9.750000	18.000000	-1.939065e+08	3.523012e+08	1.557500	1.588482e+08	15.044653	-4.352056
50%	59.705000	4.819505	1.385593	15.000000	47.000000	2.098000e+06	7.073360e+08	2.895000	3.096751e+08	20.819876	-1.067170
75%	92.880001	10.695493	1.695549	27.000000	99.000000	1.698108e+08	1.899000e+09	4.620000	5.731175e+08	31.764755	3.917066
max	1274.949951	55.051683	4.580042	917.000000	958.000000	2.076400e+10	2.444200e+10	50.090000	6.159292e+09	528.039074	129.064585

- Data contains wide range of numbers from negative values to numerical to the power of 10.
- Data will be scaled to perform clustering as the range of values would create difficulties when grouping.

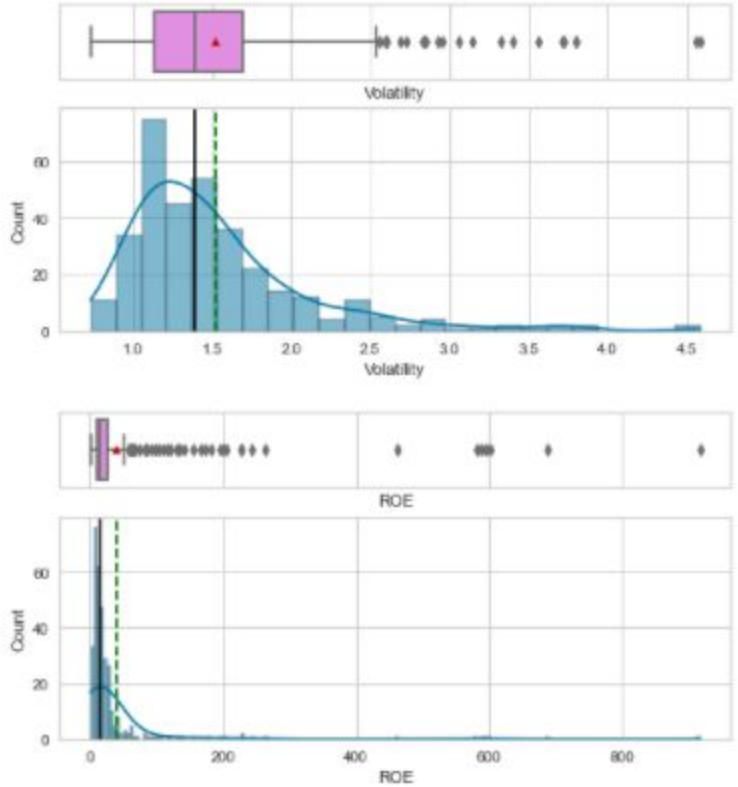
Univariate Analysis: Distribution of Numeric Columns

Current price is very heavily skewed due to very high outliers of over $1200 per share with most of the shares between $38 and $92 per share.



Price change has a very normal distribution curve with 75% of stocks having a price change of -$0.9 to $11 showing some stable market performance during this time.

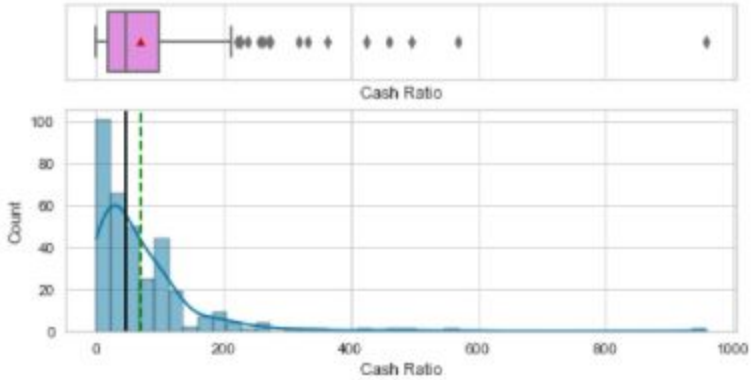
Univariate Analysis: Distribution of Numeric Columns



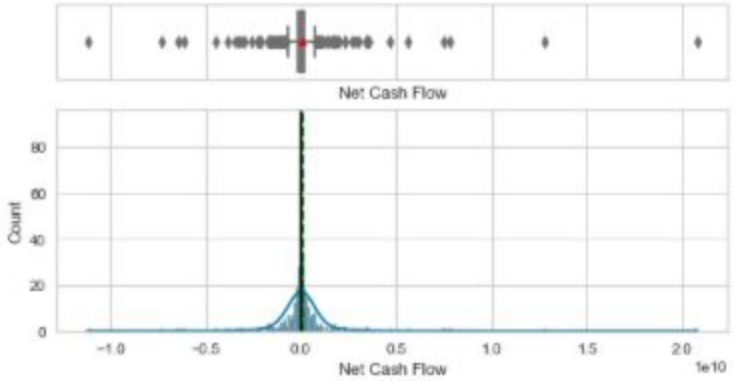
Volatility shows a heavily skewed curve with 75% of stocks having a change of 1.13 to 1.69 standard deviations over the previous 13 weeks.

ROE has ¾ of the stocks between 9.75 and 27 and again shows extreme skewness to the right indicating many companies having increased profit generation without relying on capital as much.

Univariate Analysis: Distribution of Numeric Columns

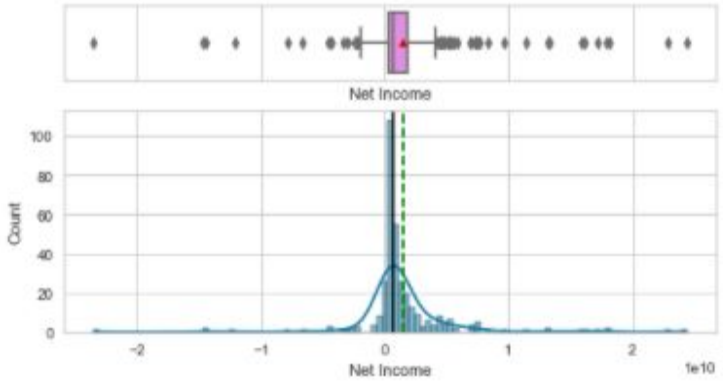


The overall cash ratio exhibits a right skewness to the normal distribution curve with 75% of the stocks between 18 and 99.

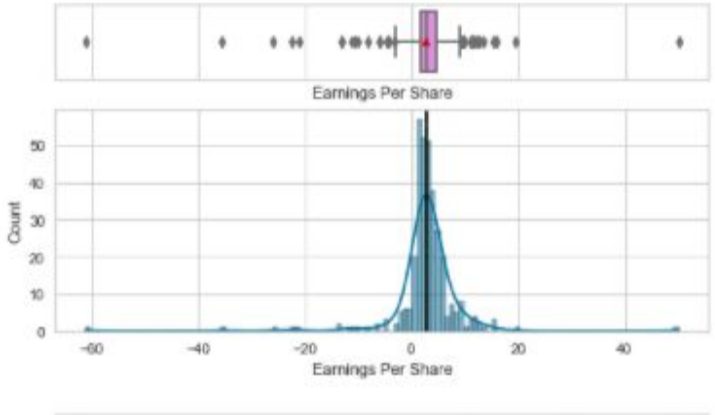


Net Cash Flow shows a normal distribution with almost a 50/50 split of stocks in the positive and those in the negative. Companies in the positive show a higher extreme outlier. The money going into and out of the companies shows just how much movement the stock market has on a day-to-day basis.

Univariate Analysis: Distribution of Numeric Columns

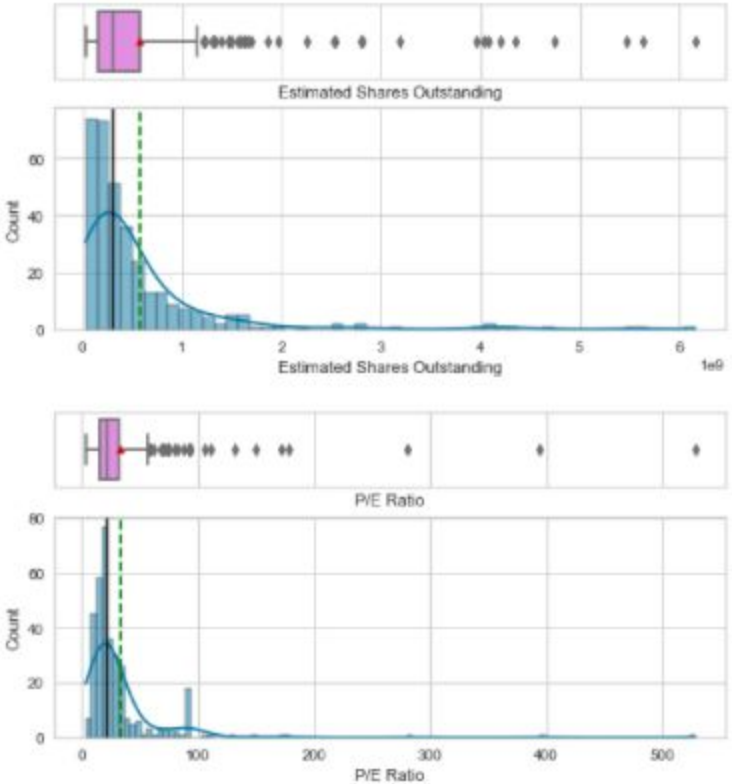


Net Income follows a normal distribution with 75% of the stocks in the positive, but a fair number of negative outliers telling of stocks with revenue not able to keep up with expenses.



Earnings per Share has a normal distribution curve with 75% of the stocks above 0. Range is from 1.5 to 4.6 with the extreme being in the negative range.

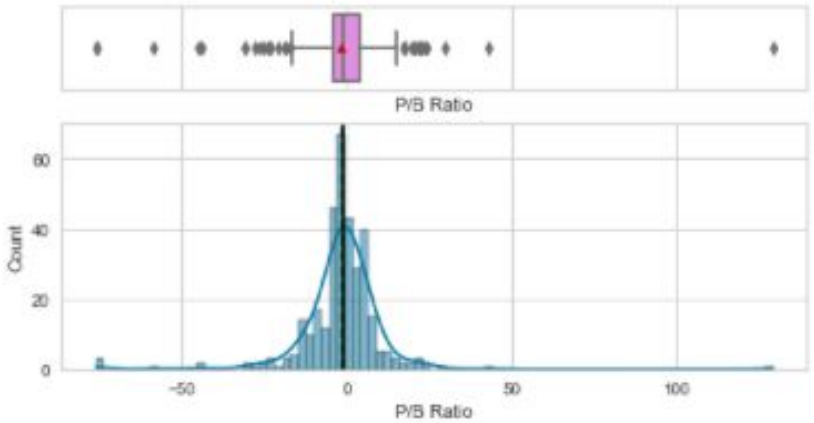
Univariate Analysis: Distribution of Numeric Columns



Estimated Shares Outstanding is positive across all stocks with extreme right skewness, over 75% of the stocks are within 1.5 - 5x10^8 shares.

P/E Ratio is heavily skewed to the right with ¾ of the stocks between a ratio of 15-31, but with very extreme outliers. The P/E ratio can vary across industries as to what a “good” or “bad” ratio is.

Univariate Analysis: Distribution of Numeric Columns



P/B ratio shows very normal distribution curve with extreme positive and negative tailing. 75% of the stocks are in the ratio range of -4 to 3.9 showing a wide range which is to be expected given the various industries present in the data set. Again, this distribution has extreme outliers on either end

Data Pre-Processing:

- Missing values (those with N/A) were treated by applying the median column value to the missing values.

- The set was divided into training and validation sets in order to maintain data integrity and prevent data leakage.

```
Training
0    0.945321
1    0.054679
Name: Target, dtype: float64
**********************************
Validation
0    0.945333
1    0.054667
Name: Target, dtype: float64
```

Training data set has 28,000 data points while validation has 12,000 data points.

Rate of failure (1) and rates of not failing (0) were consistent between both groups.

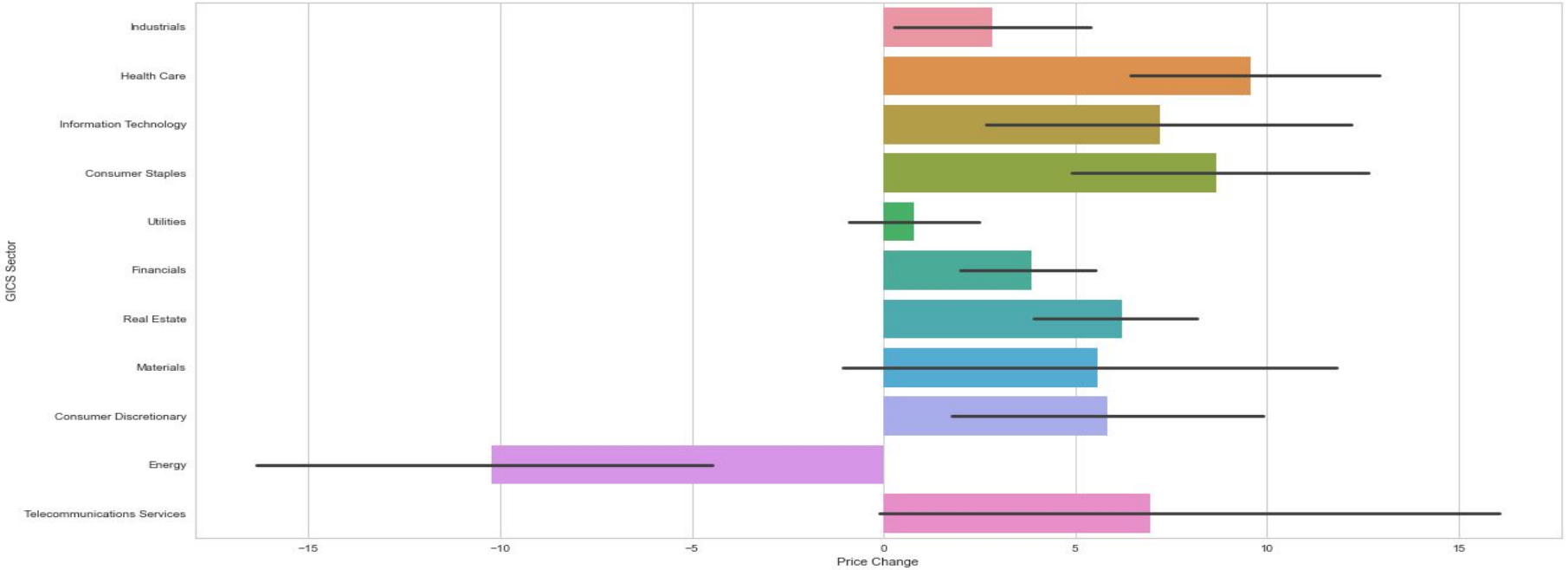
Bivariate Analysis: Correlation Matrix

Bivariate Analysis: Correlation Matrix Observations

How are the different variable correlated to each other?

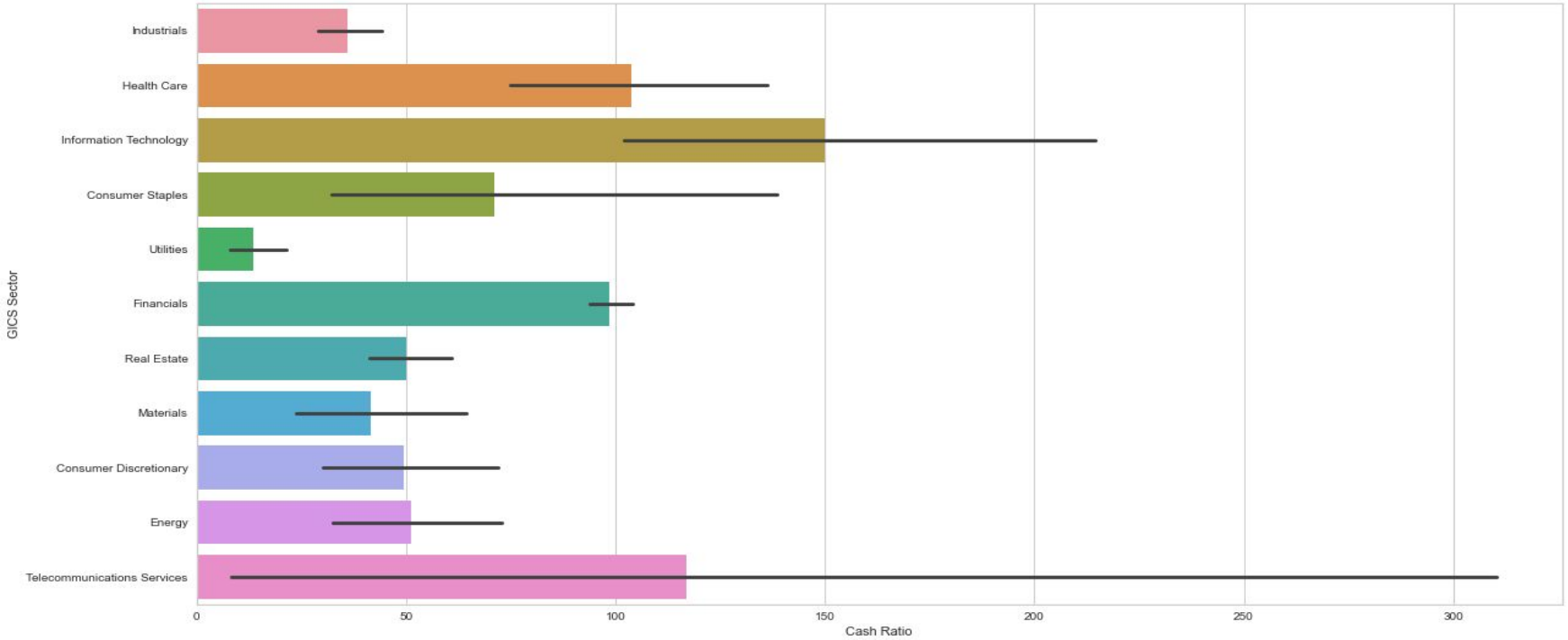
- Majority of the correlations are very low, or even negative showing little to no correlation
- Highest positive correlations are:
  - Earnings per Share vs Current Price
  - Earnings per Share vs Net income
  - Estimated Shared Outstanding vs Net income
- Highest negative correlations are:
  - Net Income and Volatility
  - Earnings per share and ROE
  - Volatility and Price Change

Bivariate Analysis: GICS Sector vs Price Change



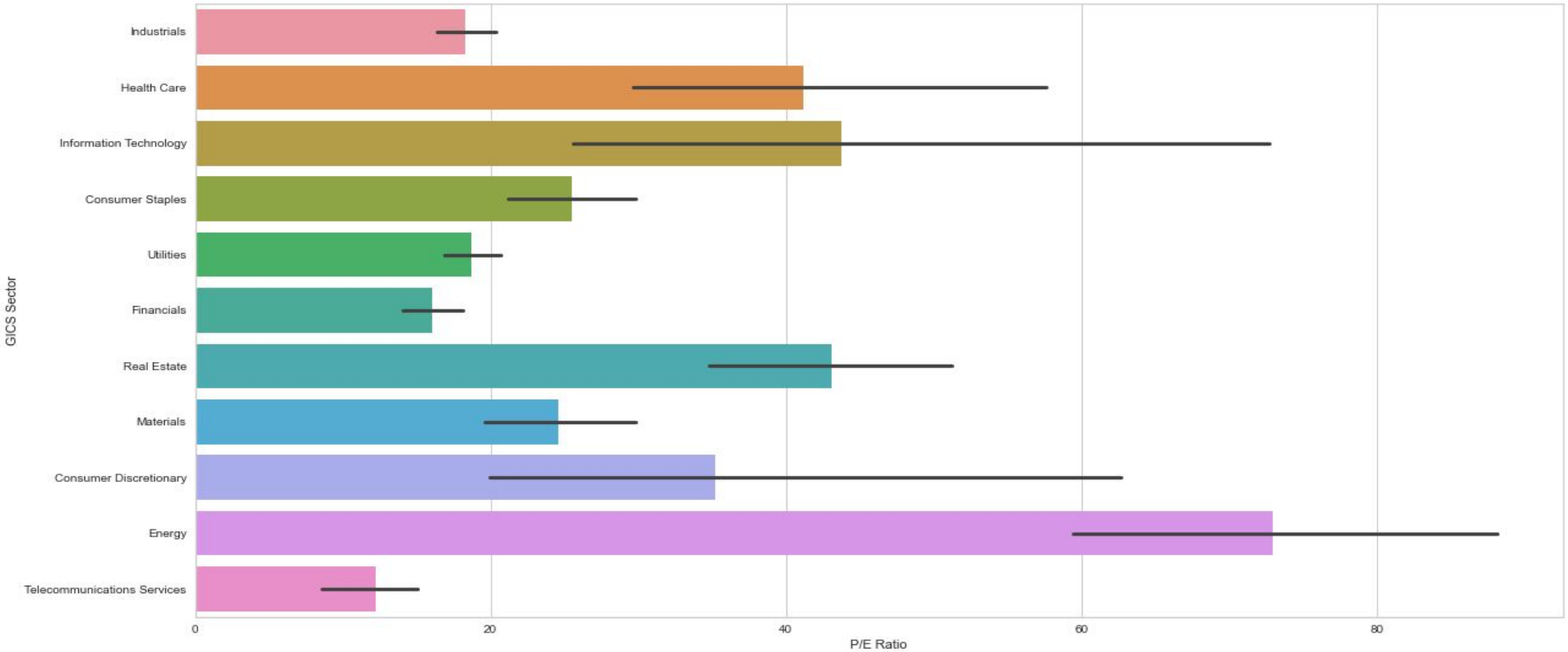
-Health Care stocks on average have seen the largest stock price increase over the last 13 weeks with an average $9.6

Bivariate Analysis: GICS Sector vs Cash Ratio



The information technology stocks have the highest average cash ratio across the economic sectors, followed by Telecommunications Services and Health Care

Bivariate Analysis: GICS Sector vs P/E Ratio



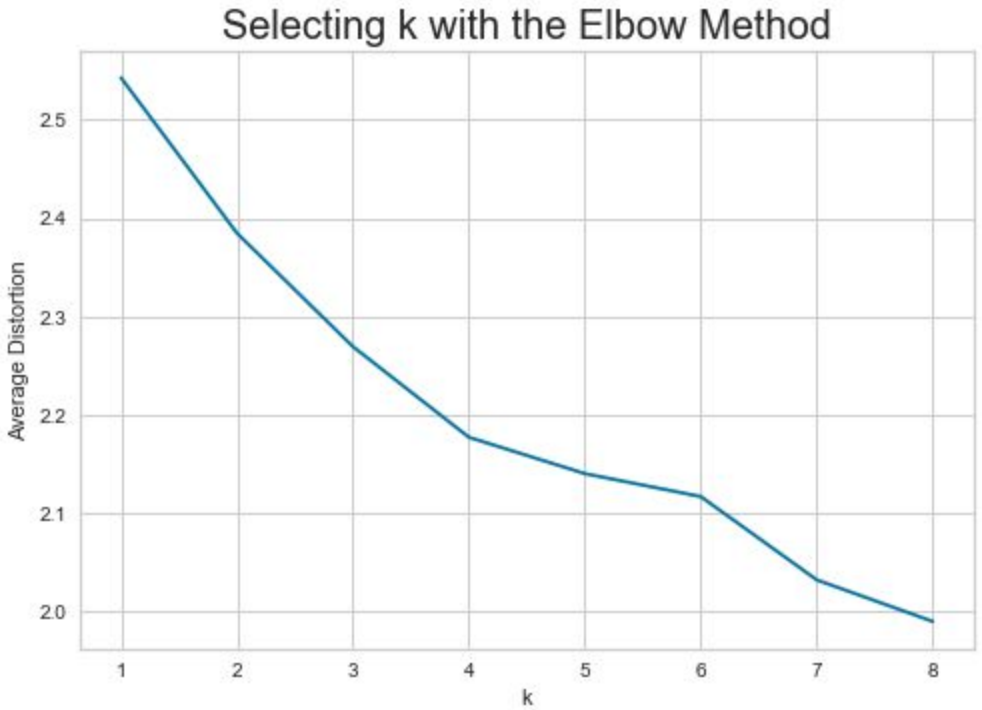
- The highest P/E ratio average belongs to the Energy Sector at 72.89
- The next three highest P/E ratios belong to the Information Technology, Real Estate and Health Care Sectors.

Modeling Plan:

The data will be scaled as the numeric columns range from negative to positive ^10. After scaling the data will be clustered using K-means clustering and Hierarchical clustering. Once both have been completed the clusters will be analyzed and compared to one another to determine which grouping method offers the most insight into the stocks and provides the most business acumen for modeling.

Modeling Plan: K- Means Clustering

Clusters will be determined using the elbow technique as well as the Silhouette scoring technique.

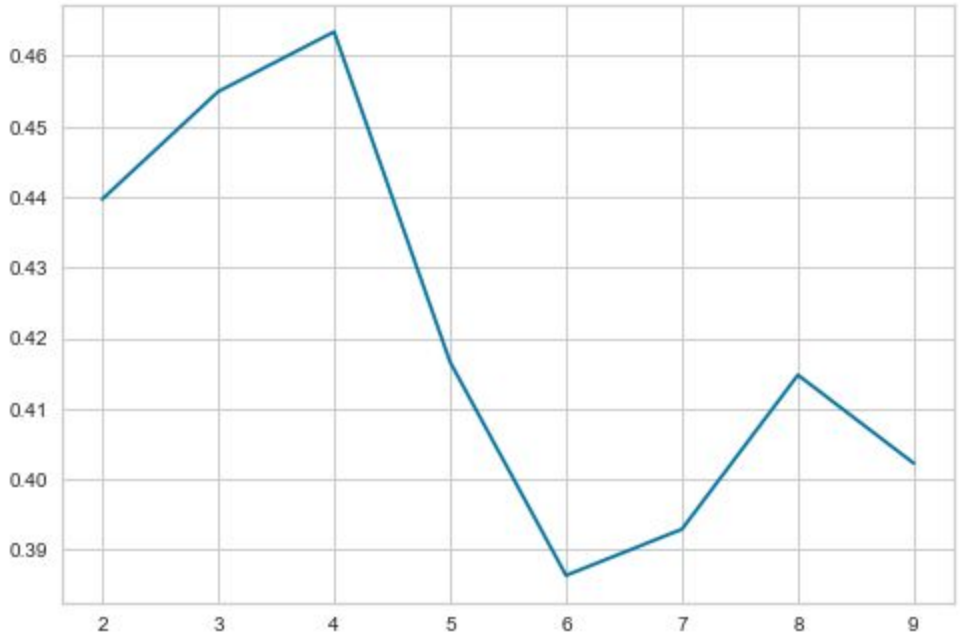


- Elbows at k=4, k=6, and k=7

```
Number of Clusters: 1   Average Distortion: 2.5425069919221697
Number of Clusters: 2   Average Distortion: 2.384499097487295
Number of Clusters: 3   Average Distortion: 2.2692367155390745
Number of Clusters: 4   Average Distortion: 2.1770961257432995
Number of Clusters: 5   Average Distortion: 2.140142891213095
Number of Clusters: 6   Average Distortion: 2.116827259798435
Number of Clusters: 7   Average Distortion: 2.0321372574873173
Number of Clusters: 8   Average Distortion: 1.9899047212918002
```

Modeling Plan: K- Means Clustering

Silhouette Scoring

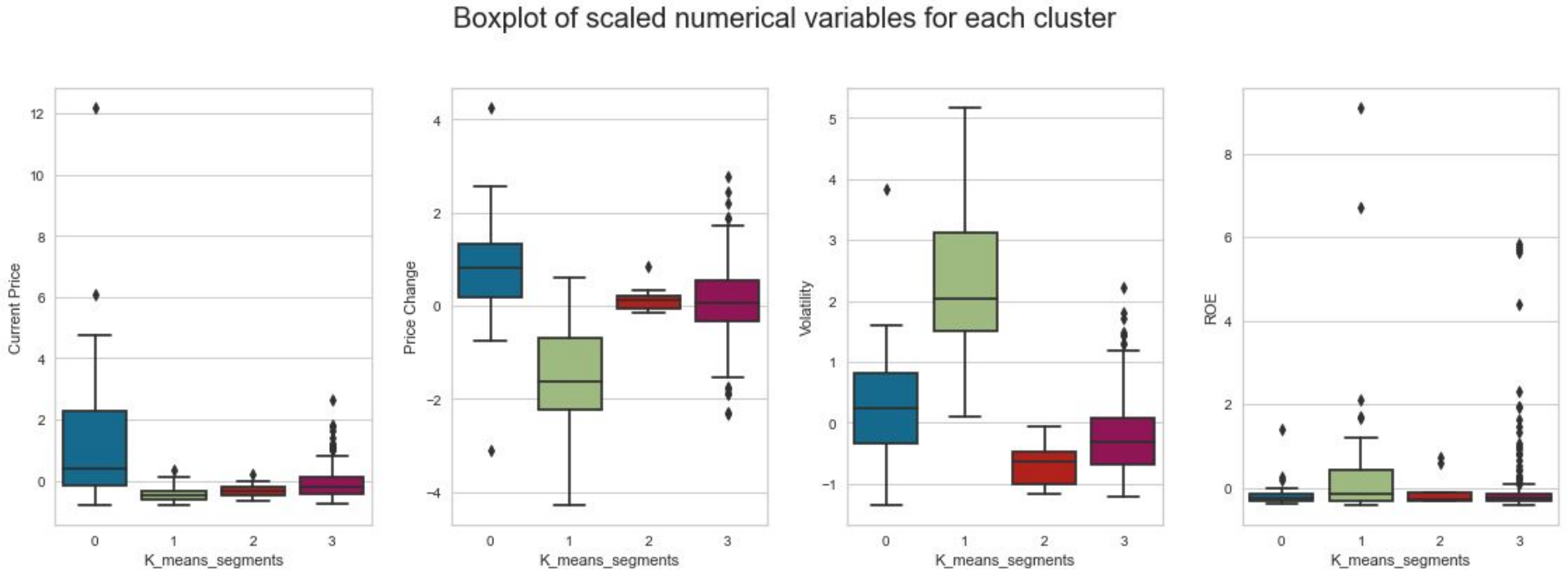


- Clear peak at k=4 and the silhouette score is the highest which indicates optimal K value for clustering

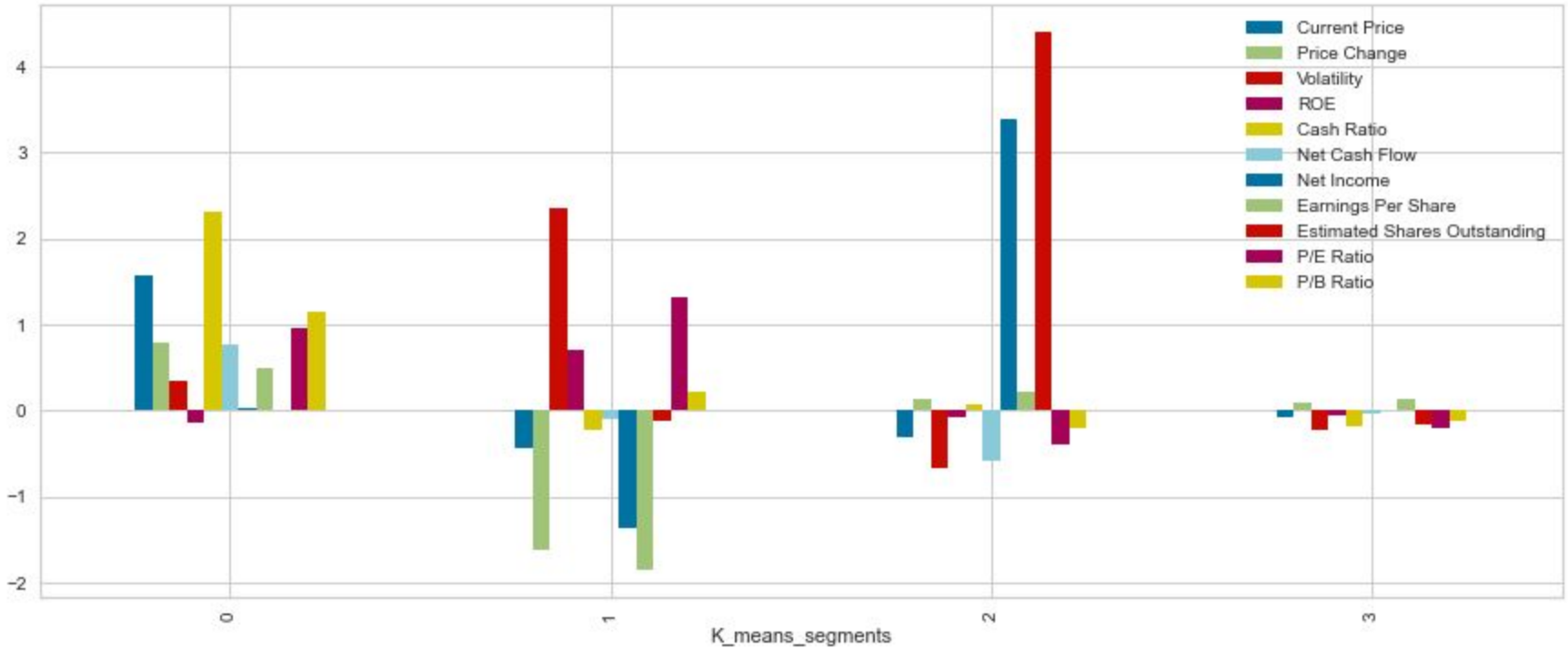
```
For n_clusters = 2, silhouette score is 0.43969639509980457
For n_clusters = 3, silhouette score is 0.45494915445064915
For n_clusters = 4, silhouette score is 0.4634280755871751
For n_clusters = 5, silhouette score is 0.4166170771600757
For n_clusters = 6, silhouette score is 0.3863465606304045
For n_clusters = 7, silhouette score is 0.39290126833986105
For n_clusters = 8, silhouette score is 0.41476612263987034
For n_clusters = 9, silhouette score is 0.40224227420635367
```

Modeling Plan: K- Means Clustering

- K-means analysis will proceed using a cluster of 4, with a cluster of 8 as a potential replacement if 4 does not offer adequate insight.



Modeling Plan: K- Means Clustering



Modeling Plan: K- Means Clustering Observations

K-Mean Clustering with k = 4 Observations

- Cluster 0:
  - Contains stocks with average price
  - Stocks have highest price change
  - Stocks have the highest volatility
  - ROE is negative
  - Stocks have the highest cash ratio
  - Highest net cash flow
  - net income is positive
  - Highest earnings per share
  - Estimated shares outstanding is zero
  - Second highest P/E ratio
  - Highest P/B ratio
- Cluster 1:
  - Most negative current price
  - Highest price change (negative)
  - most volatile stocks
  - Highest ROE
  - Negative Cash ratio
  - Negative net clash flow
  - Most negative Net Income
  - Most negative earnings per share
  - Negative estimated shares outstanding
  - Highest P/E ratio
  - 2nd highest P/B ratio
- Cluster 2:
  - Negative current price
  - Small positive price change
  - Highest negative volatility
  - Slightly negative ROE
  - 2nd highest cash ratio
  - Highest negative cash flow
  - Companies have the highest net income
  - 2nd highest earning per share
  - Highest number of outstanding shares
  - Negative P/E and P/B ratio
- Cluster 3:
  - Contains 277 companies
  - 6 factors are negative:
    - Current Price
    - Volatility
    - Cash Ratio
    - Estimated Shares Outstanding
    - P/E and P/B ratios

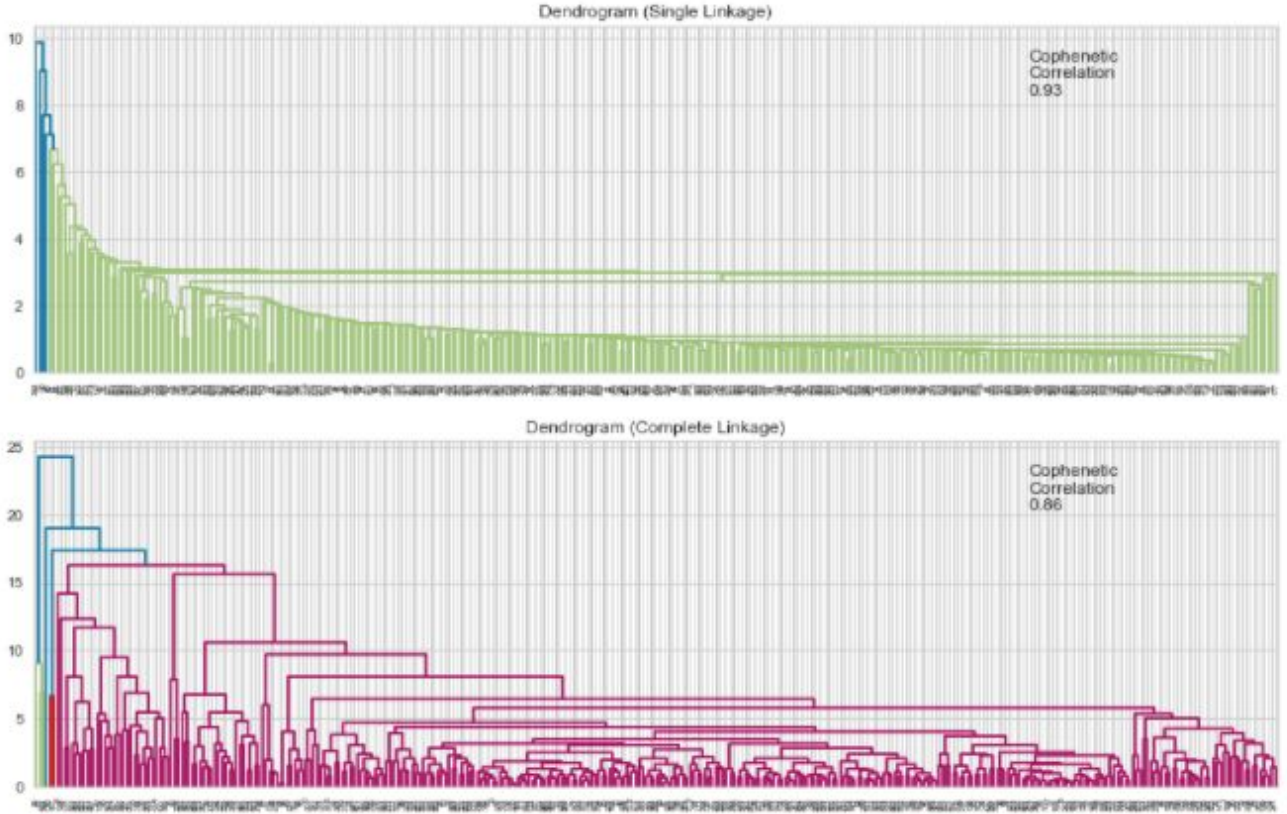
Modeling Plan: Hierarchical Clustering

Hierarchical clustering will be performed by taking multiple distance metrics and linkage methods to determining which one offers the highest Cophenetic correlation, indicating a potential optimal clustering formula.

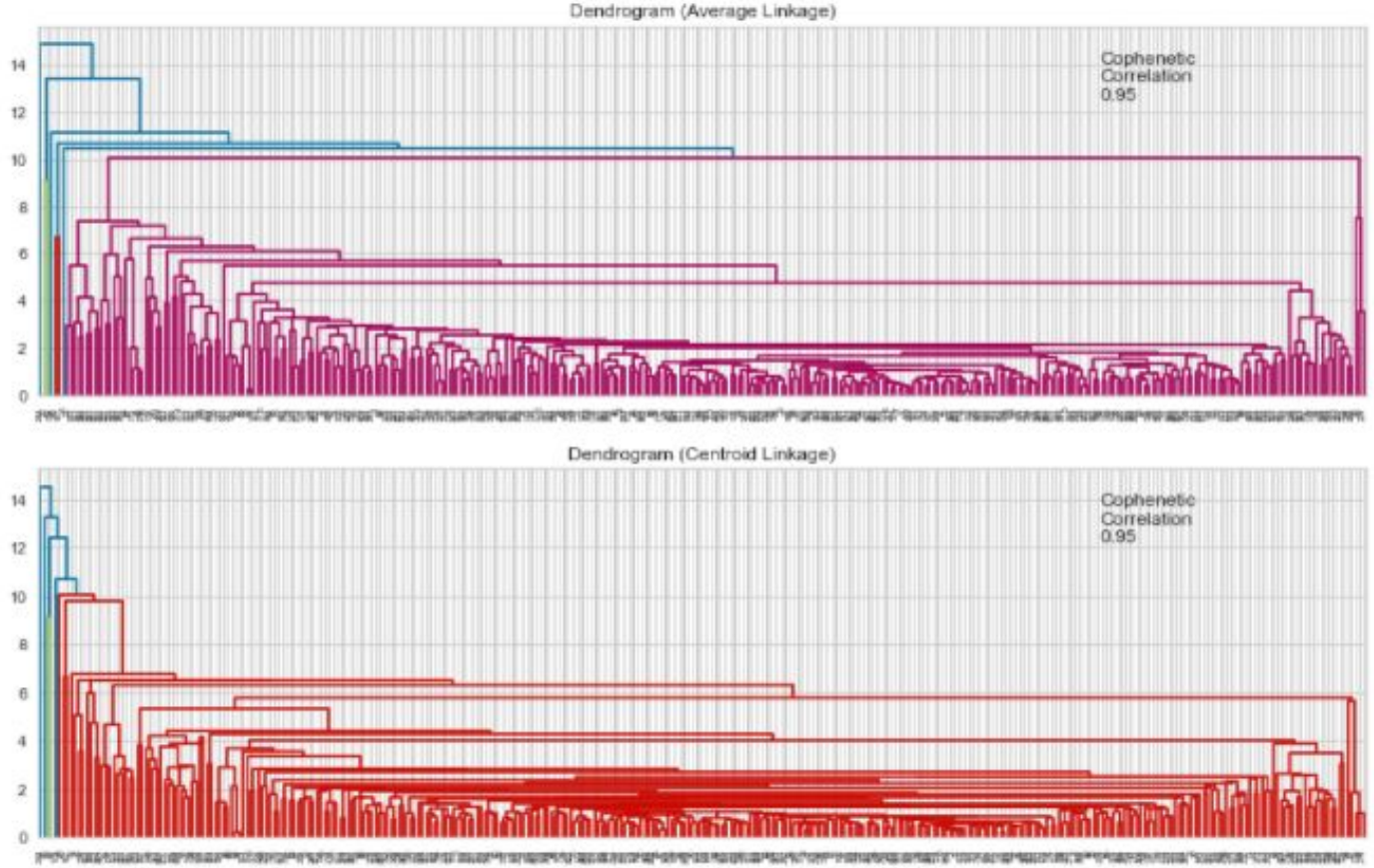
```
Cophenetic correlation for Euclidean distance and single linkage is 0.9304469769832866.
Cophenetic correlation for Euclidean distance and complete linkage is 0.8559480642212798.
Cophenetic correlation for Euclidean distance and average linkage is 0.946403836884538.
Cophenetic correlation for Euclidean distance and weighted linkage is 0.7508819056084053.
Cophenetic correlation for Chebyshev distance and single linkage is 0.9161627445317929.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.8225020941532581.
Cophenetic correlation for Chebyshev distance and average linkage is 0.9379218754329659.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.9153206618543515.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.9348505176633235.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.6881861661402053.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.9360657692078036.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.8810701545336995.
Cophenetic correlation for Cityblock distance and single linkage is 0.938373245895409.
Cophenetic correlation for Cityblock distance and complete linkage is 0.8124007660644492.
Cophenetic correlation for Cityblock distance and average linkage is 0.9168123859372298.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.866729262879581.
```

Highest cophenetic correlation is 0.946403836884538, which is obtained with Euclidean distance and average linkage.

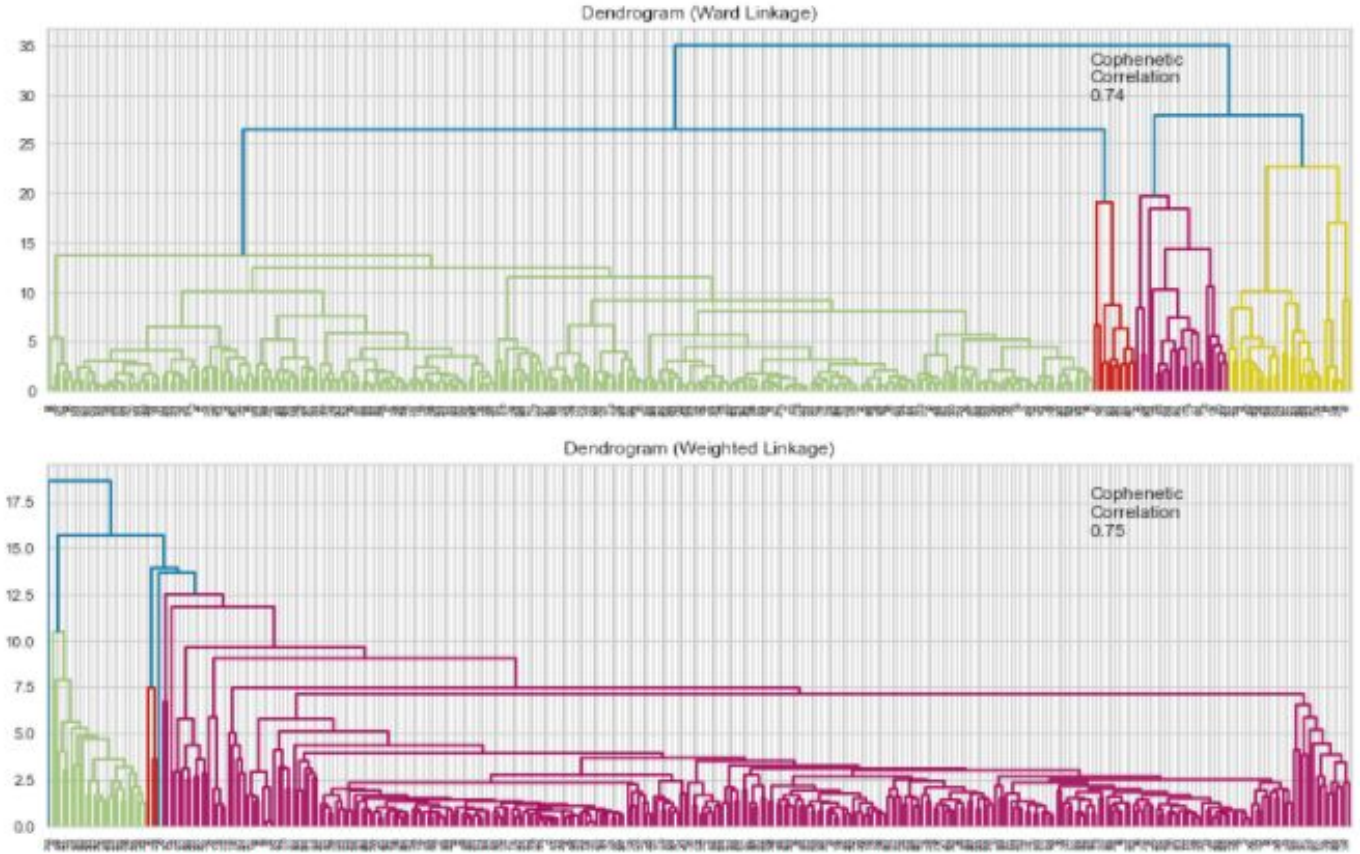
Modeling Plan: Hierarchical Clustering Dendrograms



Modeling Plan: Hierarchical Clustering Dendrograms



Modeling Plan: Hierarchical Clustering Dendrograms



Modeling Plan: Hierarchical Clustering Model

- Average and Centroid have the highest cophenetic correlation of 0.95
- Moving forward we will use Average linkage with 6 clusters as the Centroid method gets rather messy

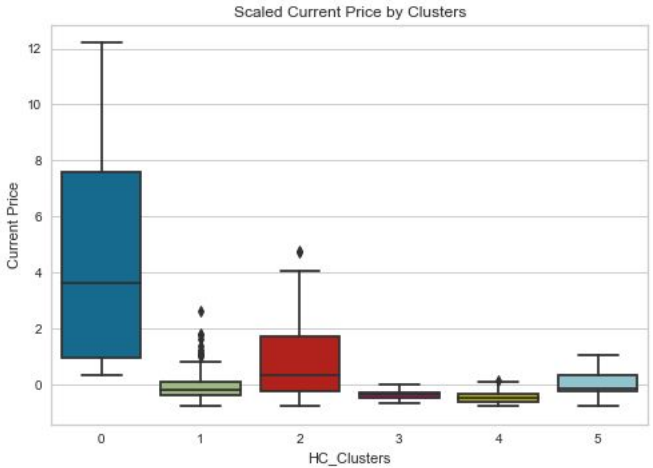
Average Linkage and 6 clusters

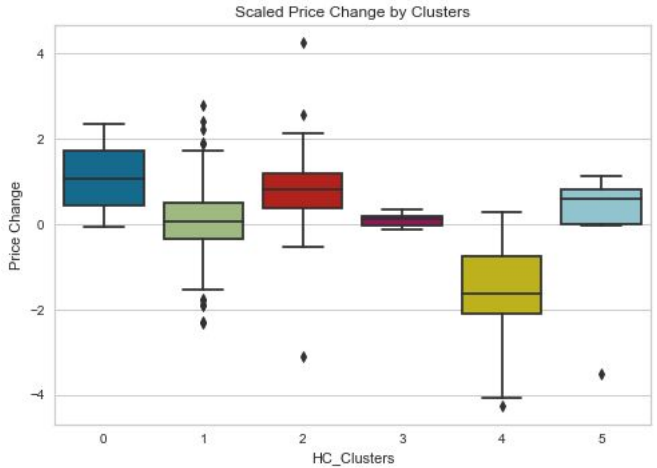
- The clustering using Average linkage produces clusters that do not have much variability
- Cluster 0 dominates with the remaining 5 clusters only totaling 7 companies out of over 300
- Looking at the dendrograms for a method offering better variability and grouping the Ward Linkage method has good grouping albeit a lower cophenetic correlation.

Ward Linkage and 6 clusters

- Ward Linkage with 6 clusters has made a little more variability in the clusters although cluster 0 still dominates in total companies.

Modeling Plan: Hierarchical Clustering Observations

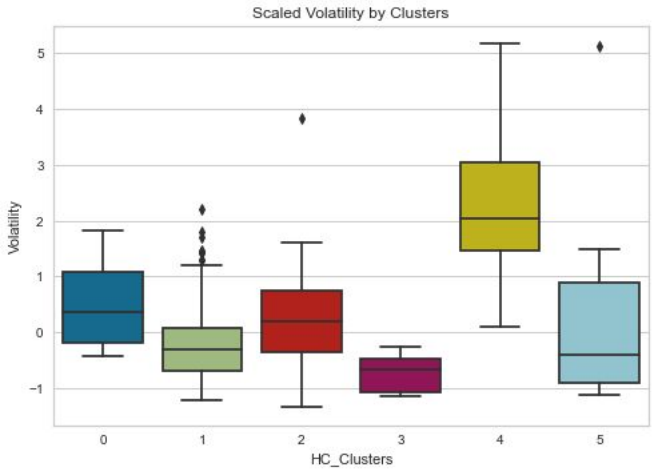


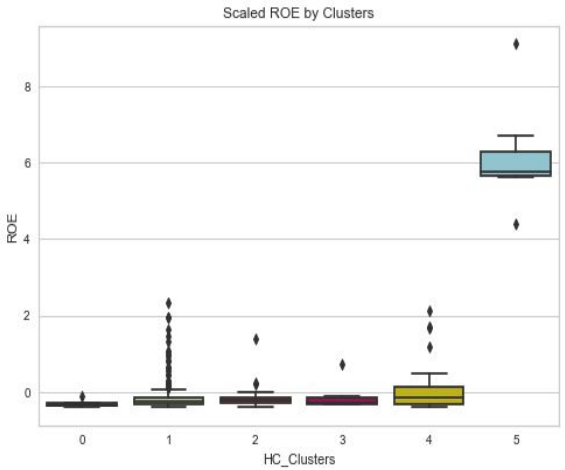


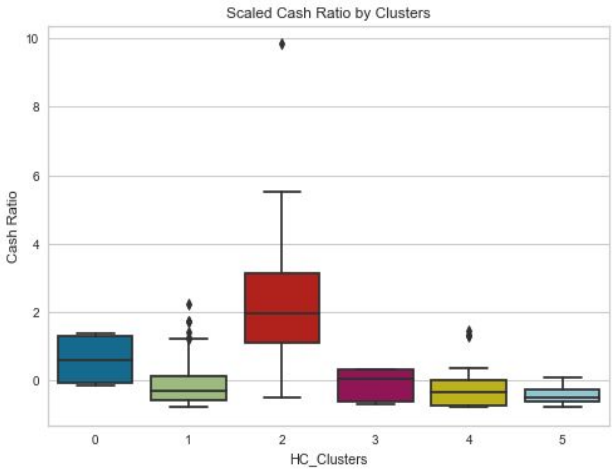
- Cluster 0 contains the highest priced stocks, followed by cluster 2
- Clusters 3 and 4 contain stocks with the lowest price

- Cluster 0 contains the stocks with the highest average price change
- Cluster 4 contains the most stocks with a negative price change

Modeling Plan: Hierarchical Clustering Observations





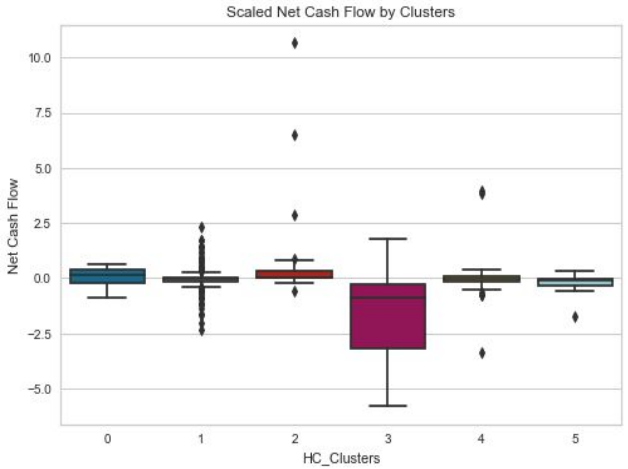


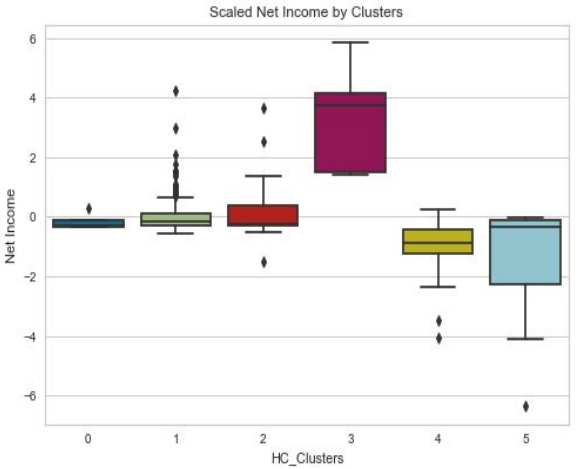
- Cluster 4 contains the stocks with the most volatility, which aligns with that cluster having the most negative price change
- Cluster 3 contains stocks with the lowest volatility

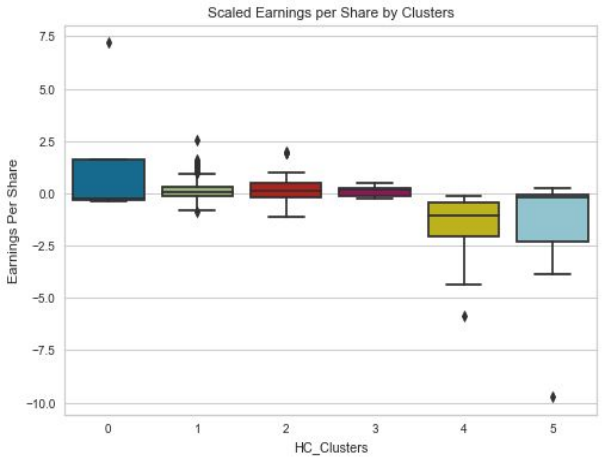
- Cluster 5 contains the highest ROE, which can be both positive and negative..
- Cluster 0 has the lowest ROE which could indicate companies having inefficient business plans for generating profit

- Cluster 2 contains the highest cash ratios, followed by cluster 0
- Clusters 4 and 5 have negative cash ratios
- A higher cash ratio could mean that the companies can more readily pay off debts if necessary

Modeling Plan: Hierarchical Clustering Observations





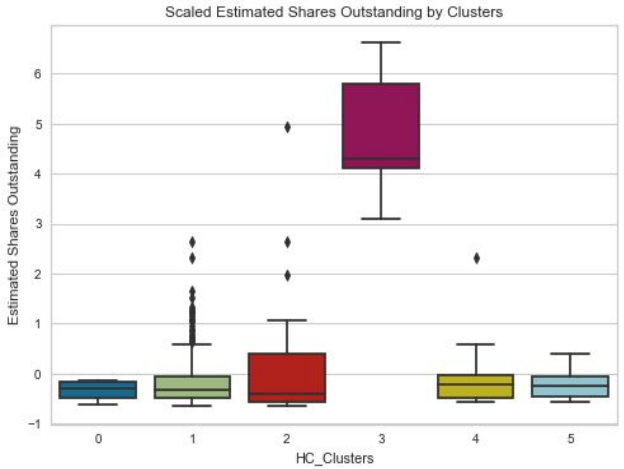


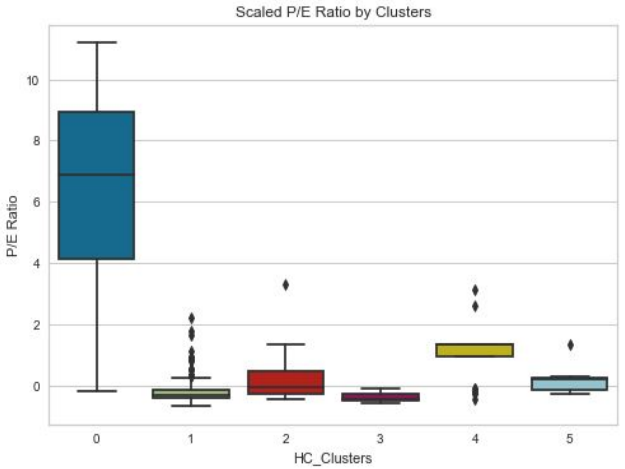
- The highest net cash flow belongs to cluster 0, which could indicate the company is bleeding money and has more bills than revenue
- cluster 3 has the most negative cash flow
- Positive cash flow indicates the company has more money coming in then out

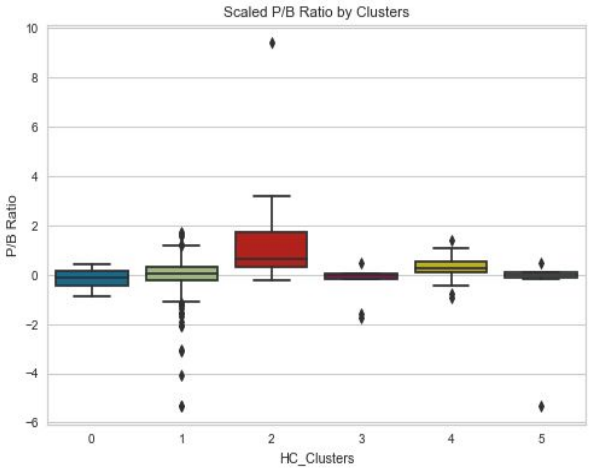
- Cluster 3 has the highest average net income across all clusters
- Clusters 0, 4 and 5 have negative average net income indicating their revenue may not be enough to support their expenses, interest and taxes

- Cluster 2 has the highest earnings per share
- Clusters 0,1,2,3 have positive earnings per share, while clusters 4 and 5 have negative earnings per share

Modeling Plan: Hierarchical Clustering Observations





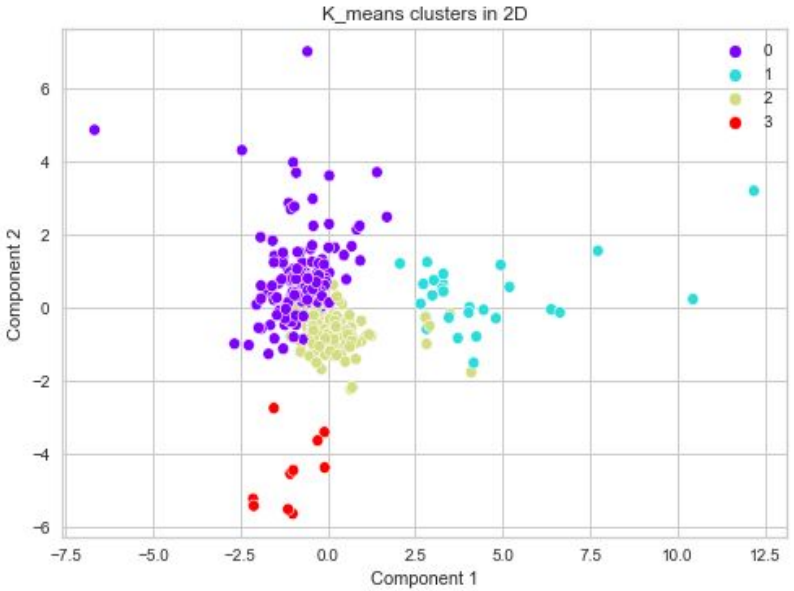


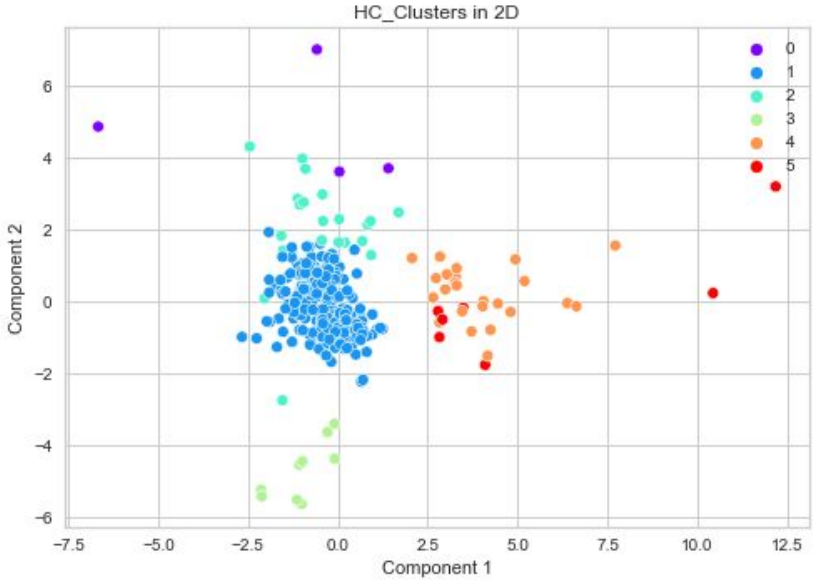
- Cluster 3 has the highest average number of outstanding shares
- Clusters 0,1,2,4 and 5 have negative average shares outstanding

- Cluster 0 has the highest P/E ratio by a wide margin
- A higher P/E ratio can be an indication of investors taking a chance on a stock due to potential future increase in stock price which can change across industries

- Cluster 2 has the highest average P/B ratio
- Cluster 0 has the lowest average P/B ratio
- P/B ratio can indicate a stock being overvalued (high P/B) or undervalued (low P/B)

K-means vs Hierarchical Clustering





Cluster Decision

- Comparing the 2D models of both k-means and Hierarchical clustering, the Hierarchical clustering using 6 cluster better represents the data as the outliers are grouped together better than with the 4 cluster from k-means. The hierarchical clustering has a little bit of overlap, between clusters 4 and 5, but the separation between clusters 0,1,2 and 3 could offer better insight for business decision making.

Business Insights for Trade&Ahead

Actionable Insights and Recommendations

- Using the clusters from Hierarchical clustering we have a generalized idea of shared characteristics across the groups
  - Cluster 0:
    - Represents the "hot" stocks
    - Smallest cluster of companies represented
    - Has the highest average stock price as well as the highest stock change of any group
    - Has low P/B ratio indicating current market values may be under expectations
    - Also has the highest P/E ratio , which may indicate that investors expect growth in the future and are willing to buy shares of the companies
    - This cluster also has the highest earnings per share of any cluster , indicating that the companies are profitable and has more profits to give our to shareholders
  - Cluster 1:
    - This cluster represents the largest group of companies and is an example of "average" market movement
    - Contain average priced stocks, low volatility and low price change
    - Companies contain the 3rd highest net income
    - Has a very high ROE indicating strong stock performance
    - Cluster has a low P/B ratio indicating the market may deem them "under valued", offering good "bang-for-your-buck"
  - Cluster 2:
    - Contains the group of companies with the 2nd highest average stock prices and price change of stock
    - High ROE indicates good stock performance
    - Has the highest net cash flow of any cluster (and it is positive), indicating money is being invested into the company
    - Has the highest cash ratio (and it is positive) indicating that the company revenue exceeds expenditures
    - Does have the highest average P/B ratio which could indicate that the stock is overvalued by investors and could mean in the future the price could change for the worse.
    - These are the "high-risk-high-reward" stocks
  - Cluster 3:
    - Cluster contains companies with below average stock price, stock price change, and volatility
    - Clust has the highest number of outstanding shares which could be from companies trying to raise capital
    - These companies have the highest net income, but also the most negative cash flow. This could be from company management investing company revenue into long-term plans that have not produced expected profit as of yet

Business Insights for Trade&Ahead

- Cluster 4:
  - Smaller cluster with the highest volatility of any grouping and also the most negative stock price change
  - Has very negative net income and also negative earnings per share indicating large loses
  - Cluster has a positive P/B ratio indicating the markey may deem them "overvalued" at the time and coupled with the very negative volatility and price change inherent with this cluster is a good combination to stay away from
  - This cluster should be marked as "Do not touch"
- Cluster 5:
  - Has average stock price, stock price change, and volatility
  - These companies have negative net income, and net cash flow
  - With such a low P/B ratio it could indicate that the stock price is trading at a lower price relative to the companies assets
  - This is cluster shares similarities according to Hierarchical clustering with cluster 4 and as such should be on the "Do not buy" list

Recommendations:

Summary of Recommendations

- Cluster 0 represents the "hot" stocks
- Cluster 1 includes the "safe" investing stocks
- Cluster 2 is for the investors seeking excellent return, but must be wary of market change and assessment of industry values
- Cluster 3 includes companies that may be good long term investments as their characteristics indicate the companies are not flashy, but maybe have long term plans that could move the cluster into the "hot" cluster at some point.
- Clusters 4 and 5 should not be considered for investing as the poor company metrics and outlier nature of the clusters would not bode well for investors.