Trade&Ahead (Case Study)

Project 7: Unsupervised Learning

Date: February 24th, 2022.

By: Ijeoma Ejem

Contents / Agenda

- Executive Summary pg. 3 - 5.
- Business Problem and Solution Overview pg. 6 - 7.
- Data Overview pg. 8 - 9.
- EDA Results:

  Univariate Analysis pg. 10 - 23.

  Bivariate Analysis pg. 24 - 29.
- Data Preprocessing pg. 30 - 32.
- K-Means Clustering:

  Elbow Method pg. 33 - 36.

  Silhouette Analysis pg. 37 - 51.
- Hierarchical Clustering pg. 52 - 69.
- K-Means Vs. Hierarchical Clustering: Evaluation pg. 70 - 71.

EXECUTIVE SUMMARY

Executive Summary

	Conclusion	Recommendation
1	The Energy sector stands out with the highest P/E ratio across all sectors, approximately 72. The Information Technology sector closely follows with an average P/E ratio of around 44. The Real Estate and Healthcare sectors also show relatively high P/E ratios, with values closely trailing those of the Information Technology sector.	This indicates that investors are willing to pay more for each dollar of earnings generated by companies in the Energy and Information Technology sectors, possibly due to expectations of future growth and profitability. Investment decisions should be made after considering all relevant factors, including the current market conditions, economic trends, and financial performance of individual companies within the sector.
2	While the healthcare sector has demonstrated a positive price change with an average increase of approximately $9 per stock, the energy sector has experienced a significant negative price change with an average decrease of over -$10 per stock.	The Energy sector's high P/E ratio, while attractive to some investors, must be balanced against its high volatility. The fluctuations in this sector's performance signal the need for careful monitoring. Investors who are risk averse may find other sectors such as Healthcare to be less risky options.
3	We utilized both K-Means and Hierarchical Clustering algorithms to develop models for identifying diverse segments of stock options for the firm's client portfolio.	Based on our analysis, I recommend using the K-Means Clustering algorithm (k=6) to build the clustering model, as it generated more diverse segments compared to the Hierarchical Clustering algorithm.

	Conclusion	Recommendation
4	When comparing GICS sectors in terms of volatility, the Energy sector was found to be the most volatile while the Utilities sector was found to have the least volatility.	The suitability of investing in the Energy sector depends on an investor's risk tolerance and investment objectives. While the Utilities sector is the least volatile of all sectors, it may not always be the most suitable option due to the low cash ratios and P/E ratios associated with this sector.
5	Strong financial positions of Information Technology companies are indicated by their high cash ratios, allowing for financial flexibility and potential growth and innovation. Conversely, the low cash ratio of the Utilities sector implies weaker financial positions.	Investors should consider investing in companies within the Information Technology sector due to their strong financial positions. However, investors may want to exercise caution when investing in companies within the Utilities sector, as their weaker financial positions could limit their ability to innovate and grow.
6	Our heatmap shows a strong positive correlation (0.59) between estimated shares outstanding and net income, indicating that as net income increases, the number of outstanding shares is likely to increase as well. Additionally, there is a positive correlation (0.56) between earnings per share and net income that suggests the same.	Investors should carefully consider the potential dilution of their holdings when investing in companies with high estimated shares outstanding. They should also closely monitor changes in net income to assess the impact on earnings per share and overall financial performance.

PROBLEM AND SOLUTION OVERVIEW

Business Problem and Solution Overview

Problem:

- Investing in stocks can provide a good way to fight inflation, create wealth, and grow savings through compound interest, but it's important to maintain a diversified portfolio to maximize earnings and reduce risk.
- Trade&Ahead, financial consultancy firm, needs a clustering model built to help with analyzing stock price and financial indicator data for NYSE-listed companies, grouping stocks based on their attributes, and providing insights on each group's characteristics.

Solution:

- Using exploratory data analysis (EDA) to explore significant influences and patterns in the stock market that will help Trade&Ahead maintain a diversified portfolio to maximize earnings and minimize risk for clients.
- We built a cluster analysis model that will help us identify similar stocks with minimum correlation across different market segments, which can protect against potential losses.
- We will analyze stock price and financial indicators data, group the stocks based on the attributes provided, and share insights about the characteristics of each group to provide personalized investment strategies for Trade&Ahead's clients.

DATA OVERVIEW

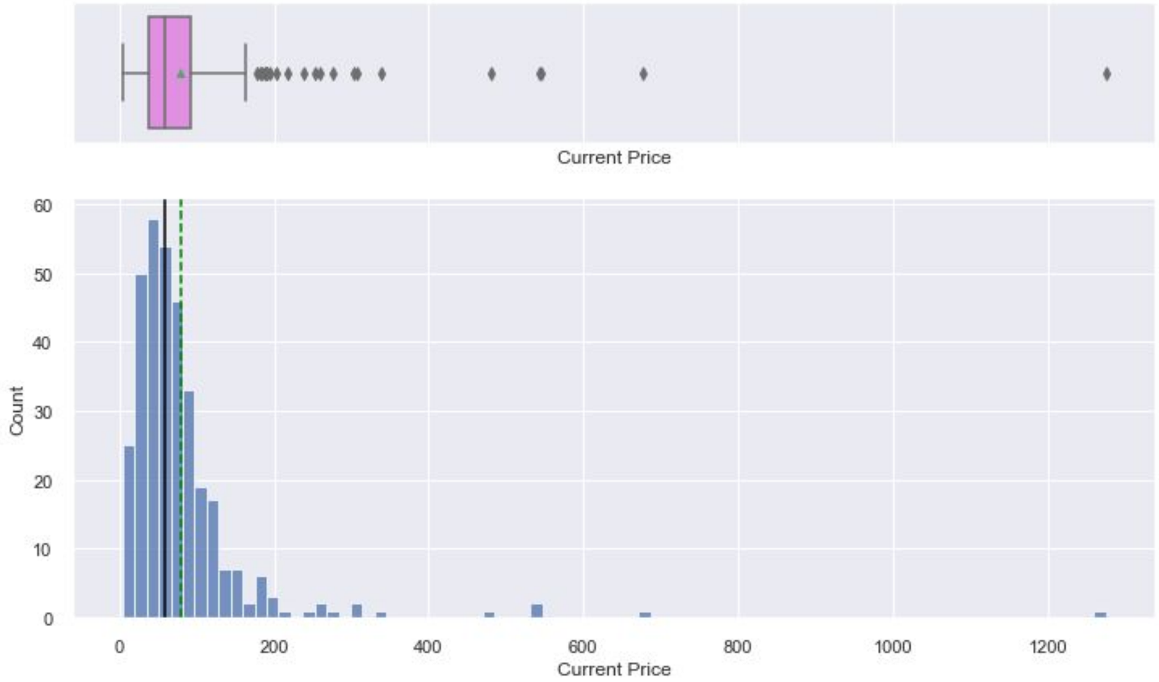
Data Overview

- ➢ There are 340 rows (observations) and 15 columns (attributes) in total.
- ➢ All datatypes are accurately represented; there are 4 objects, 4 integers and 7 float in the dataset. The attributes we are analyzing and building a model around have 4 categorical variables and 11 numerical variables.
- ➢ From the statistical summary of the data, the following is deduced;
  - The GICS sector has 11 unique values, with Industrials being the top sector.
  - The GICS sub-industry has 104 unique values, with Oil & Gas Exploration & Production being the top sub-industry.
  - The average Return on Equity (ROE) is 39.5, the median ROE is 15, and the maximum ROE is 917.
  - The average net cash flow value is $55,537,620 and the average net income is $1,494,384,602.
  - The average earnings per share is 2.77, with a minimum value of -61.2 and a maximum value of 50.09.
- ➢ There are 0 missing values and duplicate values in the data.

EDA: UNIVARIATE ANALYSIS

EDA Results

UNIVARIATE ANALYSIS

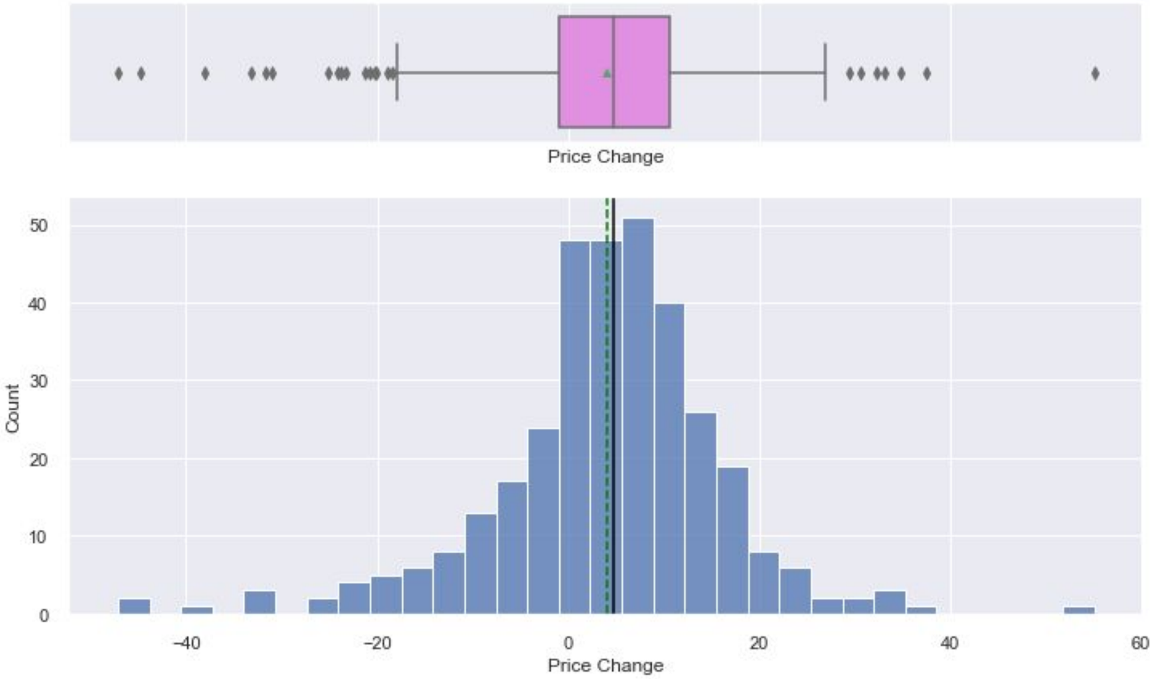


Current Price

Current price of the stock has a right-skewed distribution with a mean value of around $80 and a median value of around $60. This suggests that the majority of the stock prices fall below the mean value, indicating that there may be some expensive stocks that are driving up the average price. Additionally, there are numerous outliers in the data, which may indicate extreme price movements or other factors impacting the stock price. Outliers can have a significant impact on statistical measures such as the mean and standard deviation, so investors should exercise caution when interpreting these values.

EDA Results

UNIVARIATE ANALYSIS

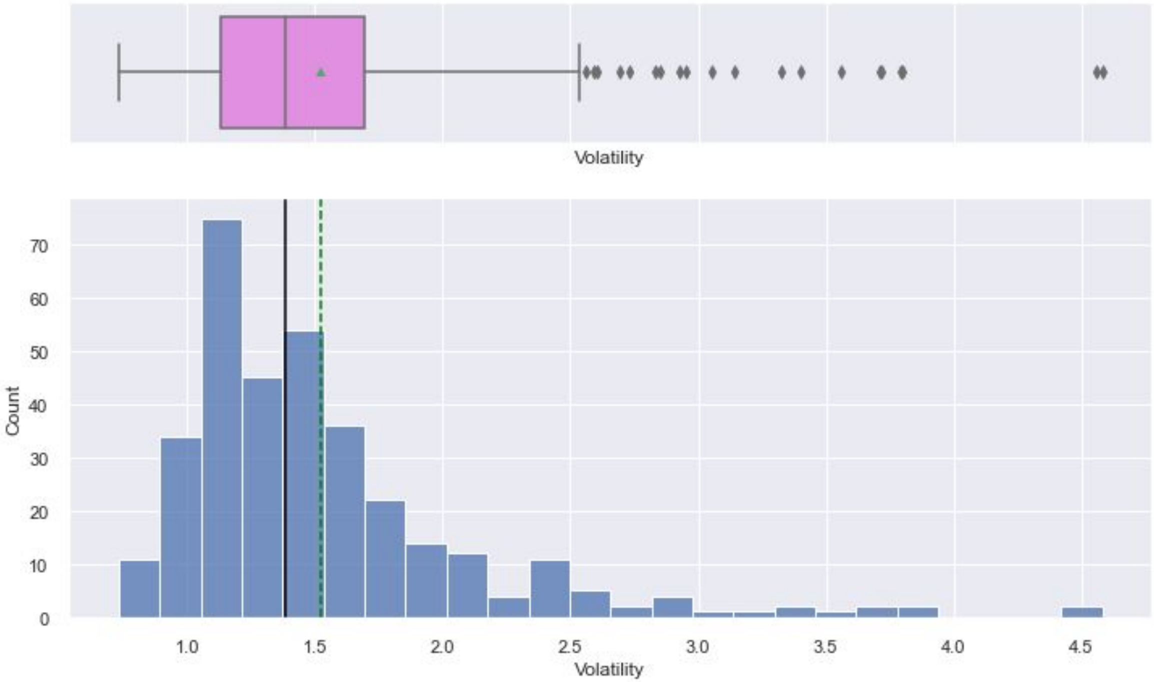


Price Change

The distribution of price change over a 13-week period appears to be normally distributed. The data suggests that the mean and median values of the percentage price change are both around $5, indicating that, on average, the stock prices have increased by a similar percentage over the 13-week period. However, it's worth noting that the normal distribution assumption may not hold true for all stocks, and some may exhibit more extreme price changes than the distribution suggests.

EDA Results

UNIVARIATE ANALYSIS

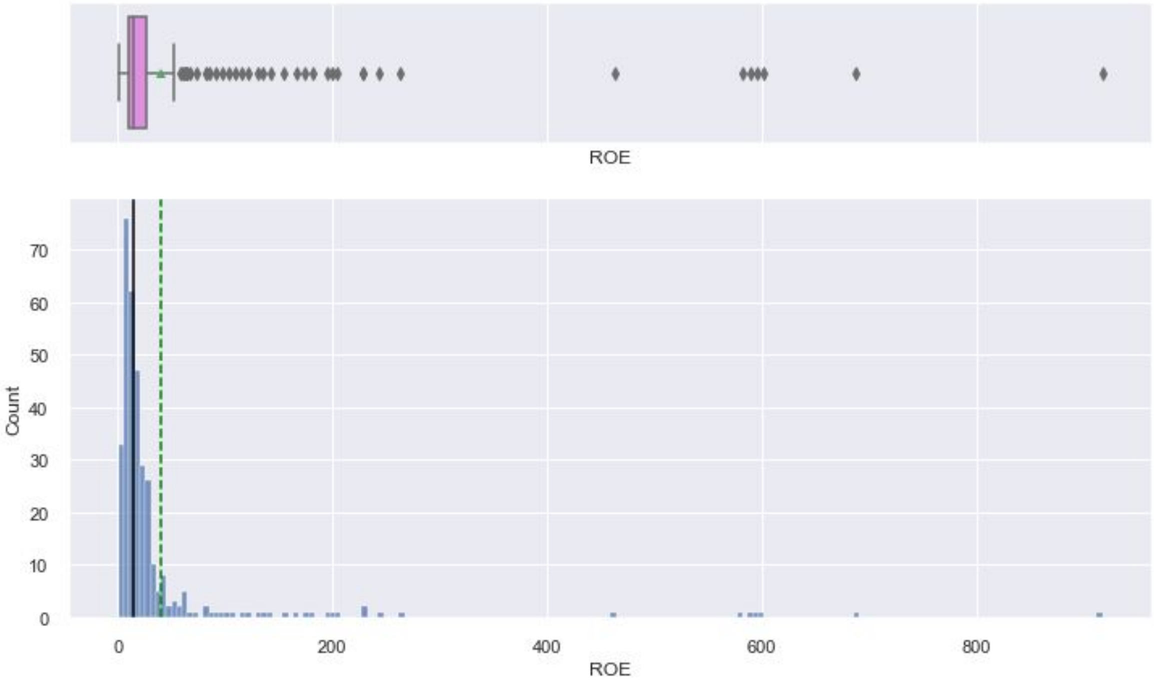


Volatility

Volatility refers to the degree of variation in a stock's price over time (calculated here as the standard deviation of the stock price over the past 13 weeks). The data show an average standard deviation of 1.5, indicating that the stock prices have exhibited moderate volatility over the period. The maximum deviation is 4.58, which suggests that some stocks have experienced significant price swings during the period. Deviations above 1.7 are considered outliers in the data, which may indicate unusual price movements or other factors impacting the stock price.

EDA Results

UNIVARIATE ANALYSIS

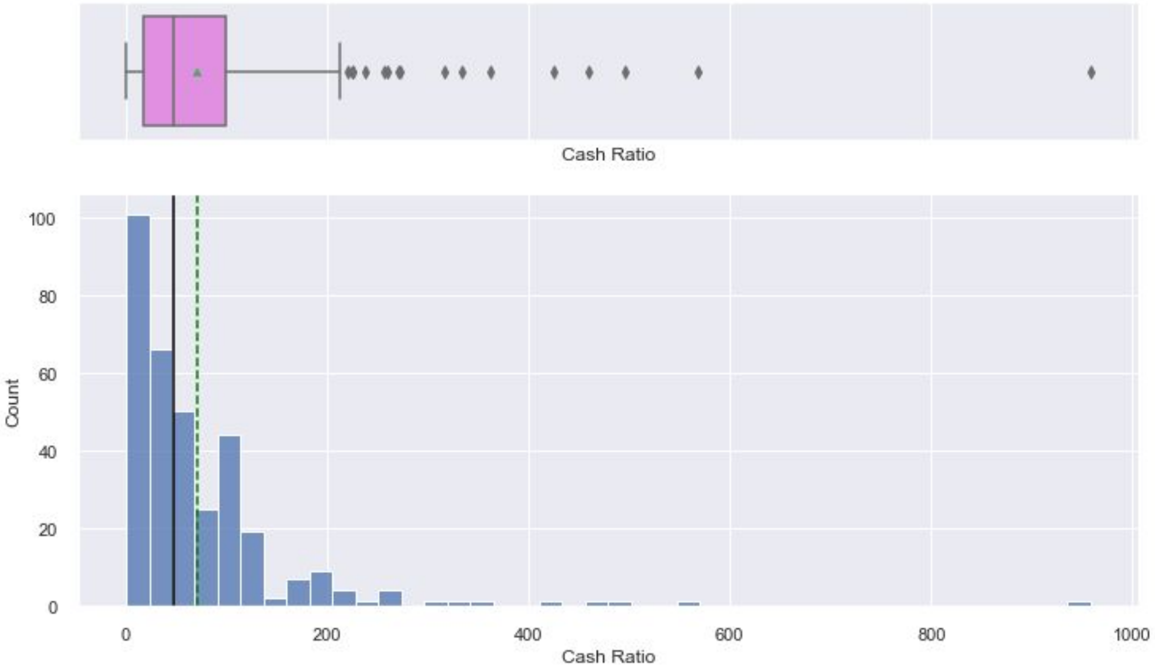


ROE

The distribution of Return On Equity (ROE)(company's profitability relative to its shareholder equity) is positively skewed. The average ROE value is 39.5, indicating that the companies, on average, have a healthy return on equity. However, the median ROE value is 15, suggesting that there are several companies with lower profitability dragging down the median value. The maximum ROE value is 917, indicating that there are a few companies with extremely high profitability relative to their shareholder equity. Overall, the right-skewed distribution suggests that there are several outliers with very high ROE values driving up the average.

EDA Results

UNIVARIATE ANALYSIS

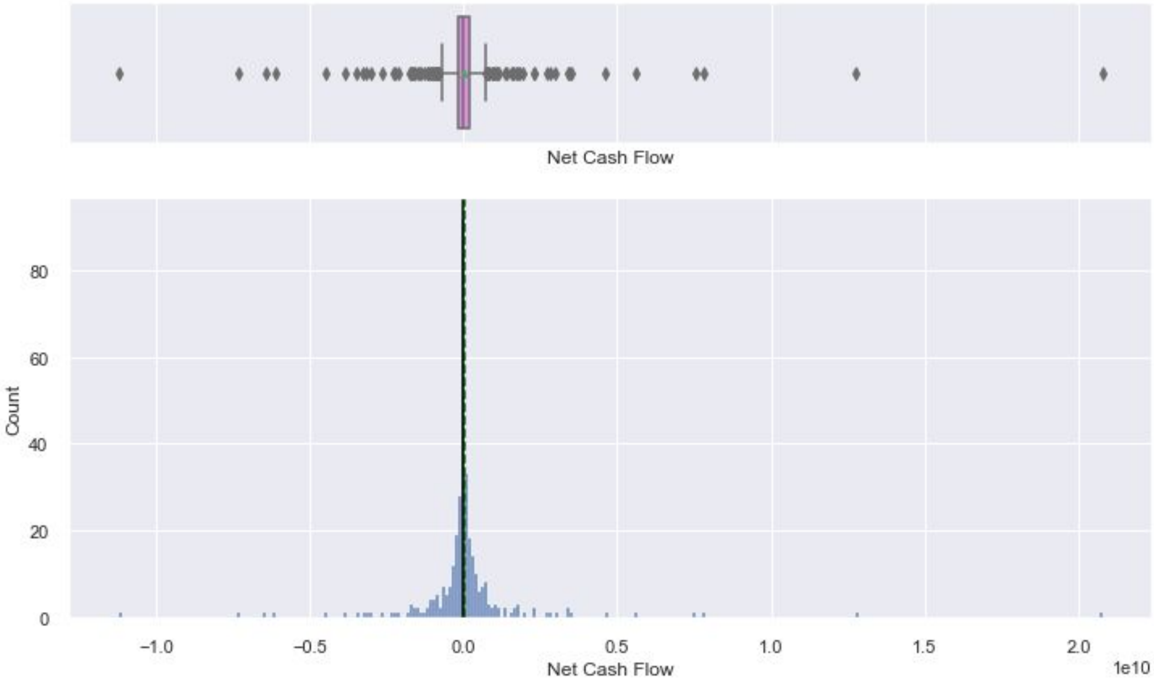


Cash Ratio

Cash ratio measures a company's ability to pay off its short-term liabilities with its available cash reserves (total cash reserves divided by total current liabilities). The average cash ratio across the data set is 70, indicating that the companies, on average, have sufficient cash reserves to cover their current liabilities. The median cash ratio is 47. The standard deviation is 90, implying that there is a significant degree of variability in cash ratios across the companies in the data set.

EDA Results

UNIVARIATE ANALYSIS

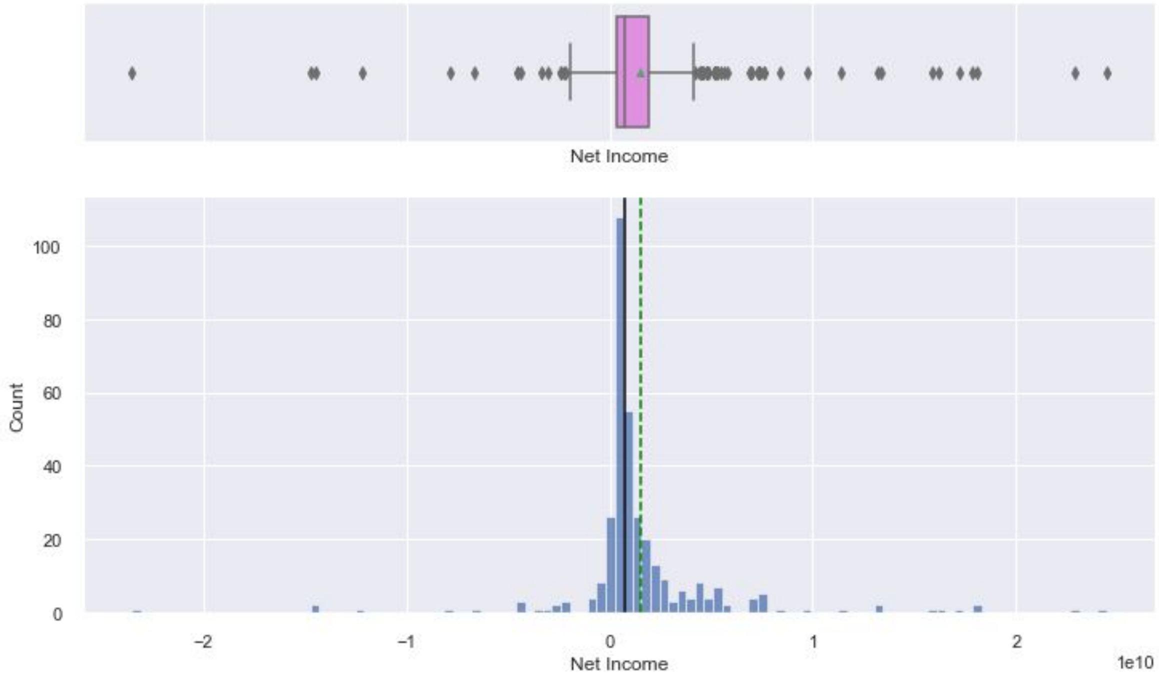


Net Cash Flow

The distribution of net cash flow appears to be normally distributed. The average net cash flow value is $55,537,620. The minimum net cash flow value is -$11,208,000,000, which indicates a significant cash outflow for some companies, potentially due to investments or debt payments. The maximum net cash flow value is $20,764,000,000, indicating that some companies experienced a significant inflow of cash during the period, potentially due to strong operating performance or other factors such as asset sales.

EDA Results

UNIVARIATE ANALYSIS

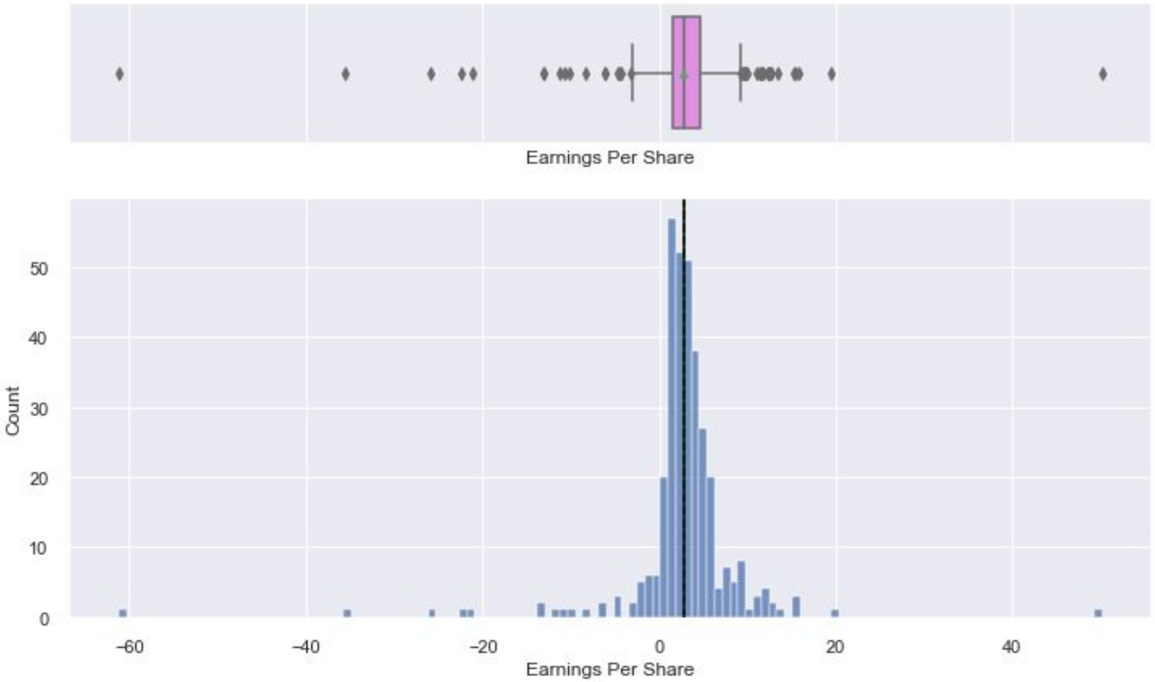


Net Income

The average net income (revenues minus expenses, interest, and taxes) is $1,494,384,602. The minimum net income in the data set is a loss of $23,528,000,000, indicating that some companies experienced significant financial losses during the period. The maximum net income is $24,442,000,000, suggesting that some companies enjoyed high levels of profitability during the period.

EDA Results

UNIVARIATE ANALYSIS

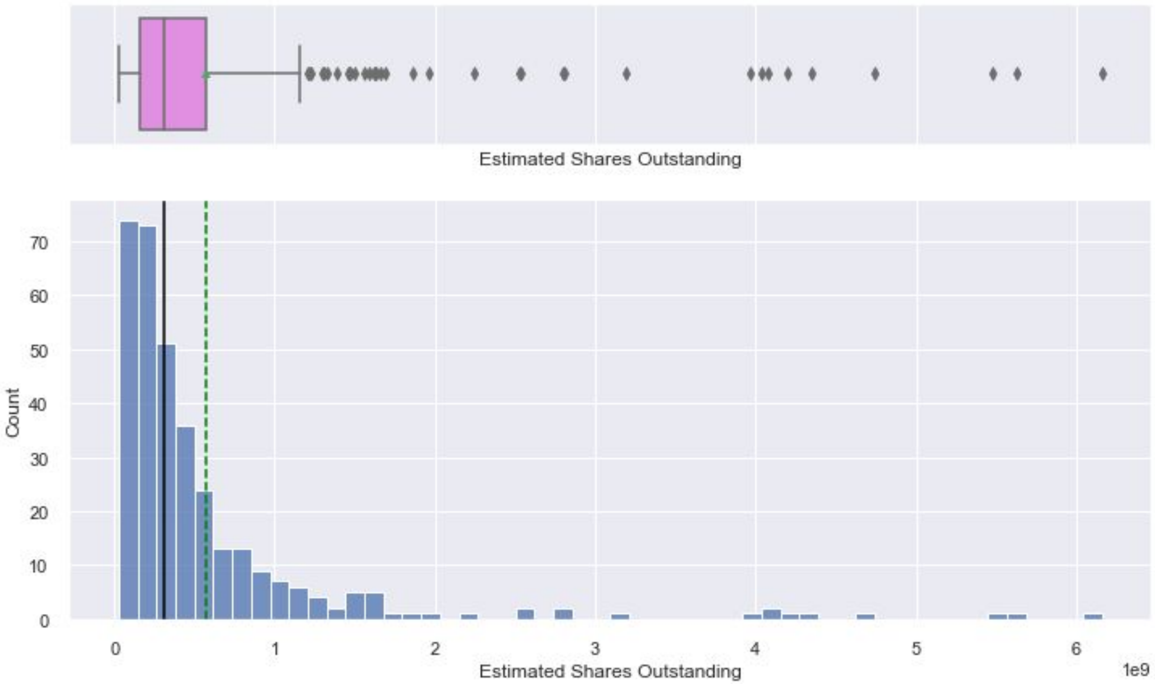


Earnings Per Share

The distribution of earnings per share appears to be normally distributed. The mean and median values are both between $2.7 and $2.9, indicating that the central tendency of the distribution is around this range. The standard deviation is $6.5, suggesting a moderate amount of variability around the mean. The minimum value of earnings per share is $-61.2, indicating a significant loss per share for some companies. The maximum value of earnings per share is $50.1, which may indicate a highly profitable company with exceptional earnings performance.

EDA Results

UNIVARIATE ANALYSIS

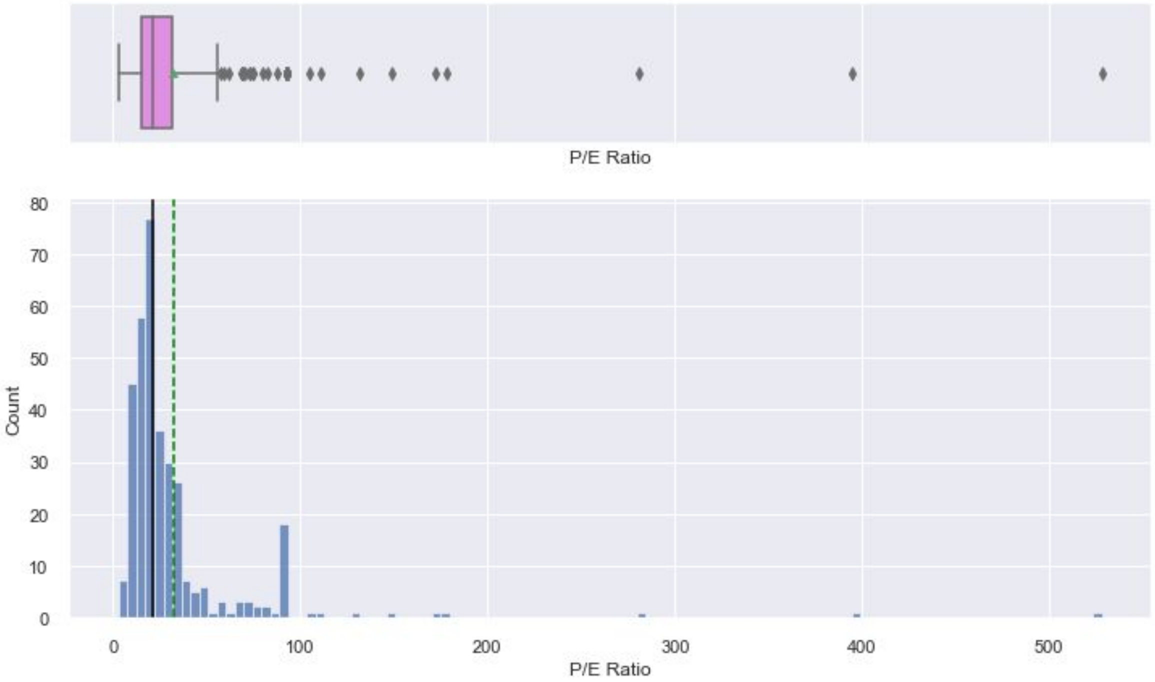


Estimated Shares Outstanding

The distribution of currently held shares by shareholders is positively skewed, with an average of 577,028,337 shares. The minimum number of shares held is 27,672,157. The median value of shares held is 309,675,138. The maximum number of shares held is 6,159,292,035, which may indicate a large shareholder or group of shareholders with significant ownership in the company.

EDA Results

UNIVARIATE ANALYSIS

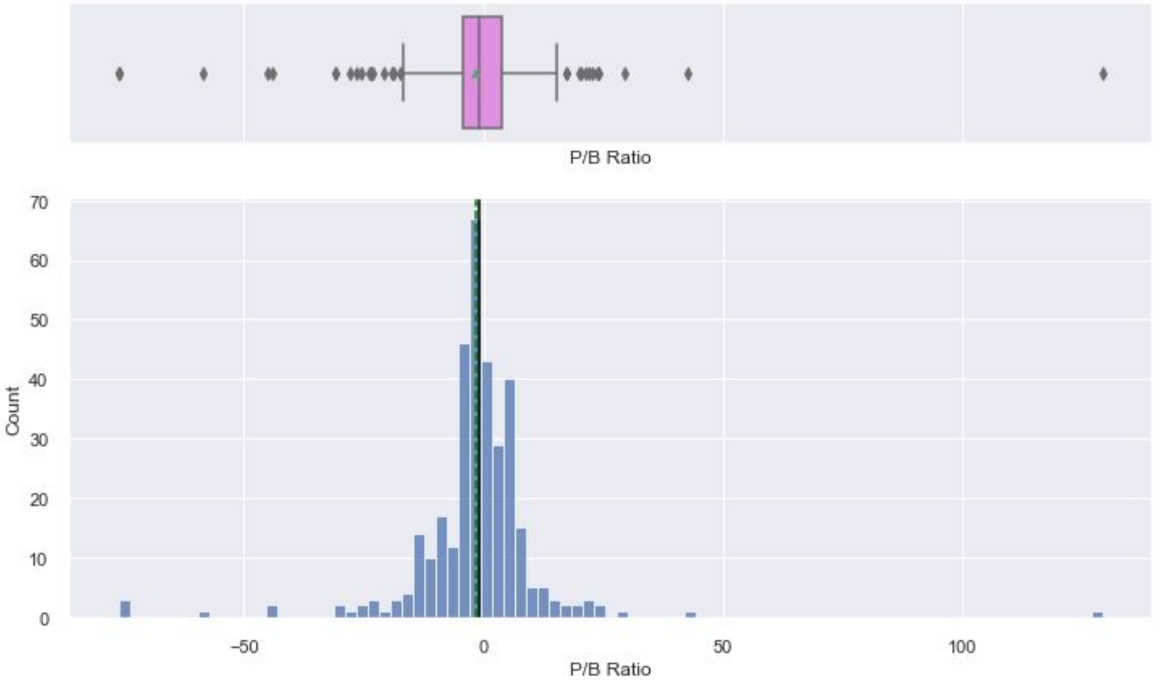


P/E Ratio

The price-to-earnings (P/E) ratio compares a company's current stock price to its earnings per share. The average P/E ratio for the given data set is 33, indicating that investors are willing to pay $33 for every dollar of earnings. The minimum P/E ratio is 2.9, which may suggest an undervalued stock or strong earnings growth potential. The median P/E ratio is 20.8. The maximum P/E ratio is 528, which may indicate an overvalued stock or exceptionally high earnings growth expectations.

EDA Results

UNIVARIATE ANALYSIS

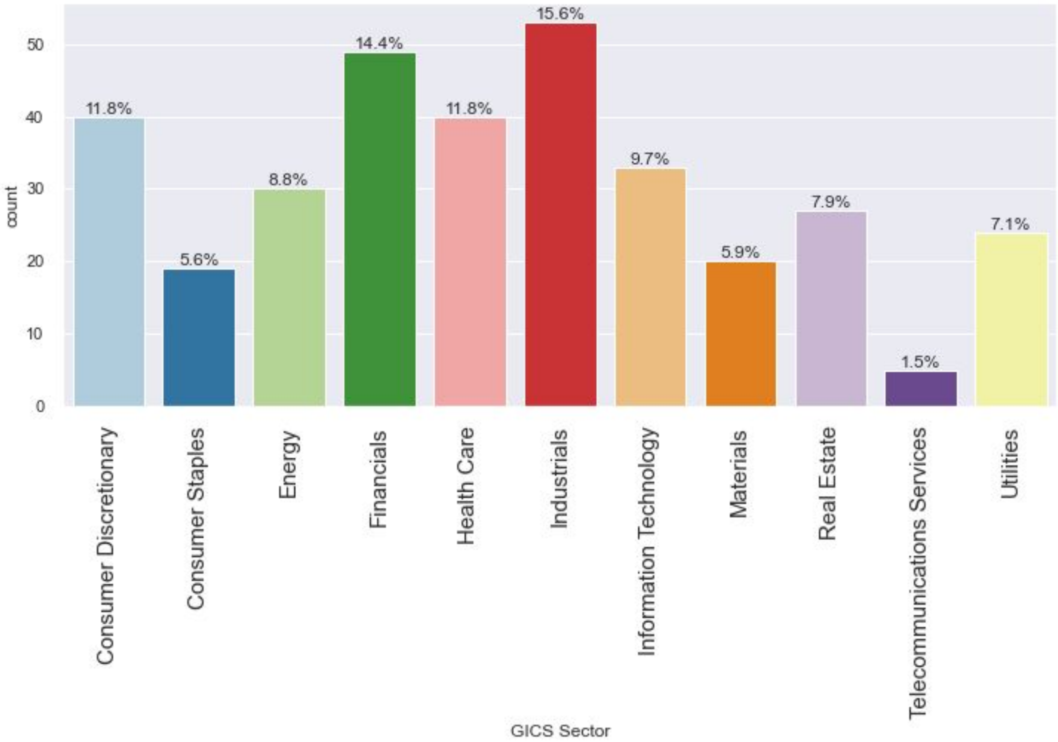


P/B Ratio

The price-to-book (P/B) ratio measures a company's stock price per share relative to its book value per share. The average P/B ratio for the given data set is -1.71, indicating that the average stock price is trading below its book value. The minimum P/B ratio is -76.11, suggesting an undervalued stock or potentially poor financial performance. The median P/B ratio is -1.06, and the other half have ratios below. The maximum P/B ratio is 129.1, which may indicate an overvalued stock or a high level of investor confidence in the company's future prospects.

EDA Results

UNIVARIATE ANALYSIS



GICS Sector

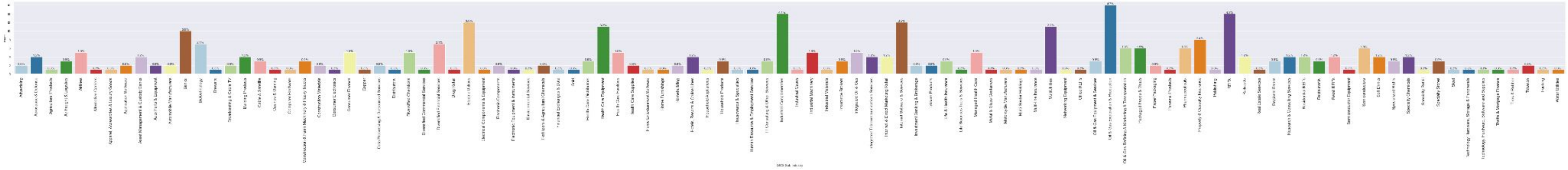
Based on the Global Industry Classification Standard (GICS), the Industrials sector has the largest number of companies, with 53 firms, accounting for 15.6% of all companies in the GICS sectors. The Financials sector comes in second with around 49 companies, comprising 14.4% of the total, followed by Consumer Discretionary and Health Care, both with 40 companies, accounting for 11.8% of all companies in each sector.

EDA Results

UNIVARIATE ANALYSIS

GICS Sub Industry

The plot shows percentage of companies within various GICS Sub Industries, with the largest being Oil & Gas Exploration & Transportation, which has a count of 16 companies, representing 4.7% of the total companies in the dataset. The next largest sub-industries are Industrial Conglomerates and REITs, both with a count of 14 companies each, representing 4.1% of the total companies. Electric Utilities and Internet Software & Services are also prominent sub-industries, with a count of 12 companies each, representing 3.5% of the total companies in the dataset. Together, these five sub-industries make up the largest portion of the dataset, highlighting the importance of these industries in the economy.



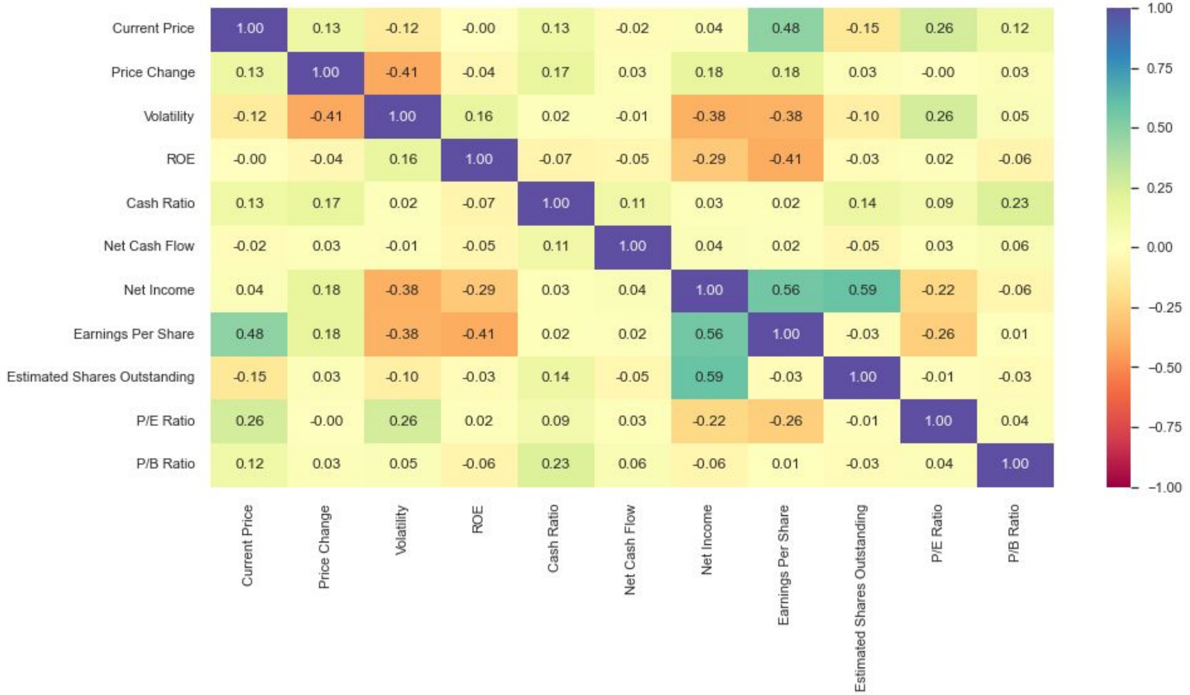
EDA: BIVARIATE ANALYSIS

EDA Results

BIVARIATE ANALYSIS

Heatmap

The heatmap shows that estimated shares outstanding and net income have a strong positive correlation (0.59), followed by a moderately strong positive correlation between earnings per share and net income (0.56). There is also a relatively, moderate positive correlation between earnings per share and current price (0.48), as well as between volatility and price change (0.41) and earnings per share and ROE (0.41).



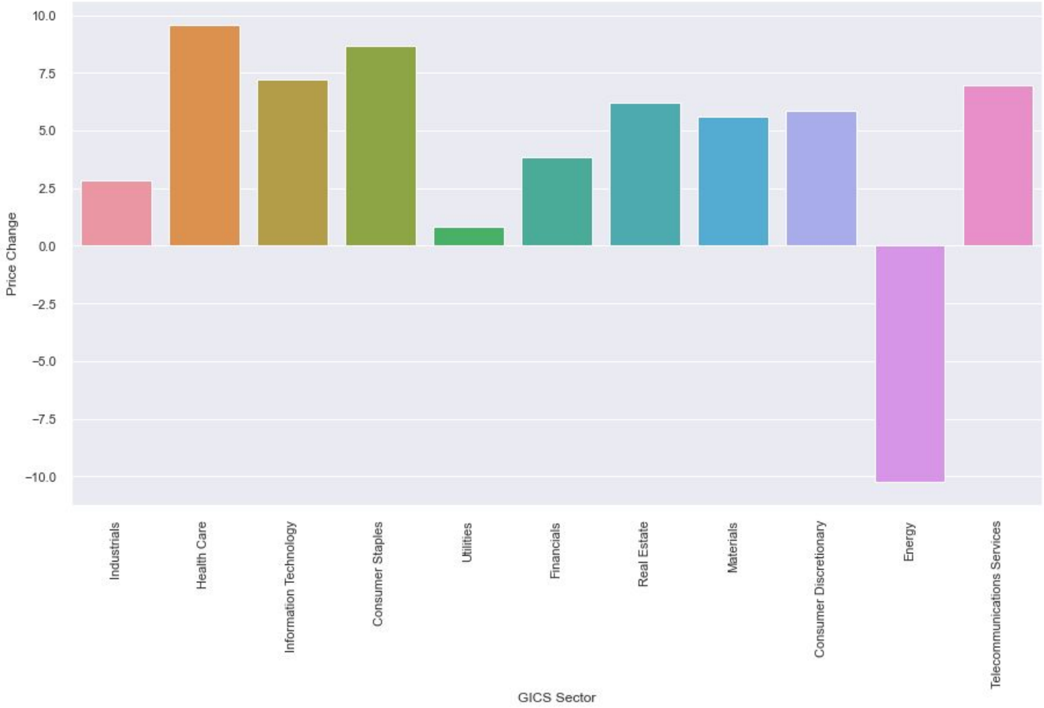
	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio
Current Price	1.00	0.13	-0.12	-0.00	0.13	-0.02	0.04	0.48	-0.15	0.26	0.12
Price Change	0.13	1.00	-0.41	-0.04	0.17	0.03	0.18	0.18	0.03	-0.00	0.03
Volatility	-0.12	-0.41	1.00	0.16	0.02	-0.01	-0.38	-0.38	-0.10	0.26	0.05
ROE	-0.00	-0.04	0.16	1.00	-0.07	-0.05	-0.29	-0.41	-0.03	0.02	-0.06
Cash Ratio	0.13	0.17	0.02	-0.07	1.00	0.11	0.03	0.02	0.14	0.09	0.23
Net Cash Flow	-0.02	0.03	-0.01	-0.05	0.11	1.00	0.04	0.02	-0.05	0.03	0.06
Net Income	0.04	0.18	-0.38	-0.29	0.03	0.04	1.00	0.56	0.59	-0.22	-0.06
Earnings Per Share	0.48	0.18	-0.38	-0.41	0.02	0.02	0.56	1.00	-0.03	-0.26	0.01
Estimated Shares Outstanding	-0.15	0.03	-0.10	-0.03	0.14	-0.05	0.59	-0.03	1.00	-0.01	-0.03
P/E Ratio	0.26	-0.00	0.26	0.02	0.09	0.03	-0.22	-0.26	-0.01	1.00	0.04
P/B Ratio	0.12	0.03	0.05	-0.06	0.23	0.06	-0.06	0.01	-0.03	0.04	1.00

EDA Results

BIVARIATE ANALYSIS

We checked the highest price increase of stocks on average across all economic sectors.

On average, the healthcare sector has shown the highest positive price change, with an increase of around $9 in stock price. The consumer staples sector follows closely behind, with an average stock price change of about $8.5. However, the energy sector has experienced the largest negative price change, with a decrease of over -$10 on average.



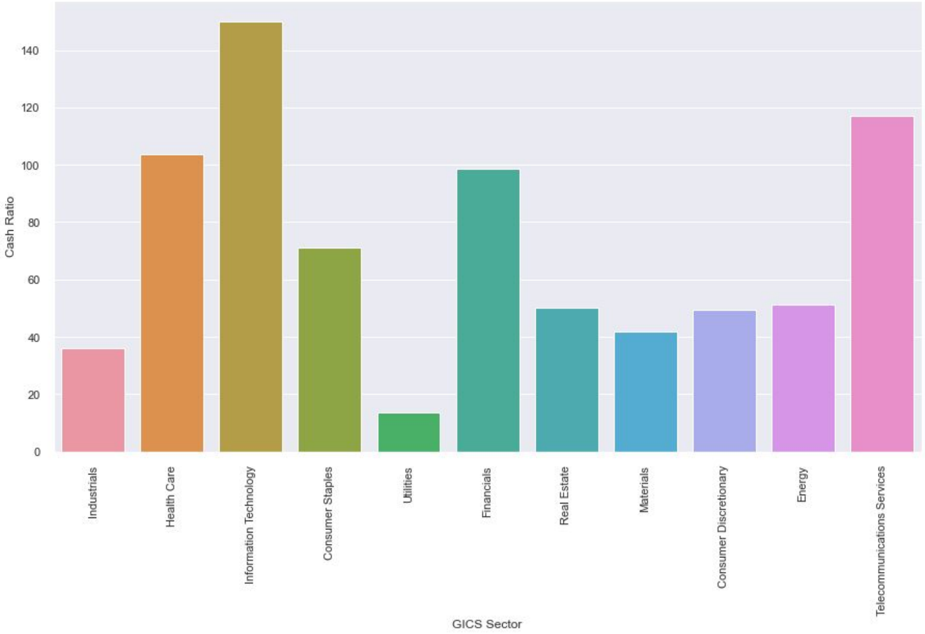
EDA Results

BIVARIATE ANALYSIS

We checked how the average cash ratio varies across economic sectors.

The information technology sector has the highest cash ratio of all GICS sectors at around 150, followed by telecomm services and health care. The sector with the lowest cash ratio is Utilities at just over 10. The high cash ratios of the Information Technology sector suggests that companies within these industries have strong financial positions, with a high level of cash reserves relative to their current liabilities. This provides them with financial flexibility, which can lead to growth and innovation in their respective industries.

On the other hand, the low cash ratio of the Utilities sector indicates that companies in this industry may be more susceptible to financial stress or unexpected events. This could be due to the capital-intensive nature of the Utilities industry, where companies require significant investments in infrastructure and equipment to operate, resulting in less cash available for reserves.



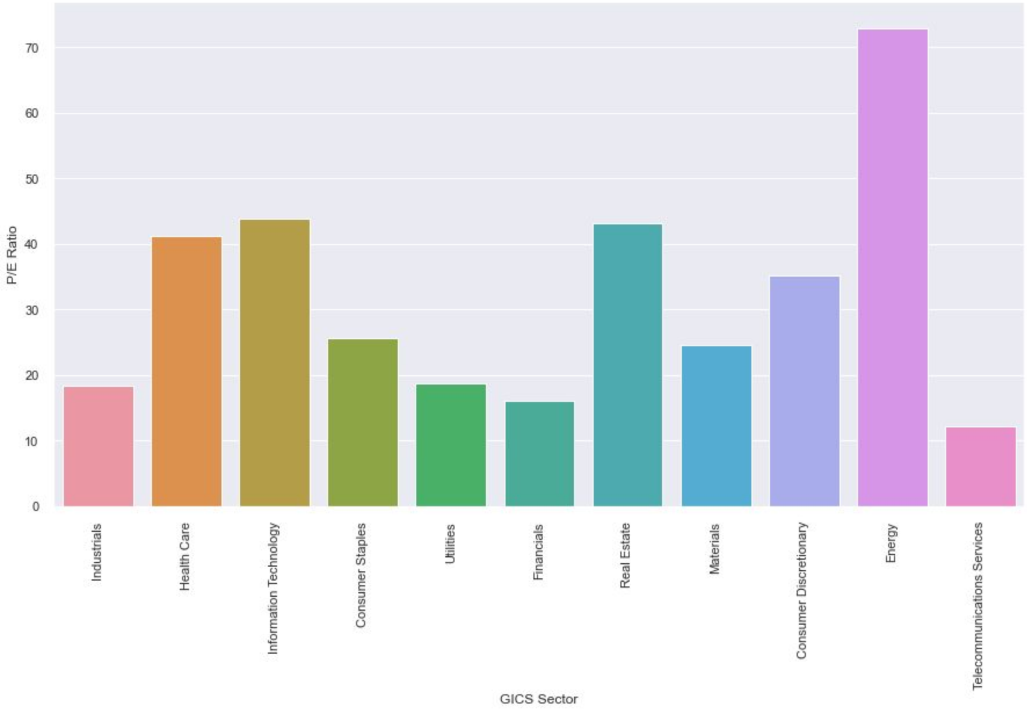
EDA Results

BIVARIATE ANALYSIS

We checked how the P/E ratio varies, on average, across economic sectors.

The Energy sector has the highest P/E ratio across all sectors at around 72, followed by the Information Technology sector at around 44, and Real Estate and Healthcare sectors following closely behind. This indicates that investors are willing to pay more for each dollar of earnings generated by companies in the Energy and Information Technology sectors, possibly due to expectations of future growth and profitability.

On the other hand, the sectors with the lowest P/E ratio across all sectors are Telecommunications at about 12 and Financials at about 16. This suggests that investors are not willing to pay as much for each dollar of earnings generated by companies in these sectors, possibly due to concerns about regulatory challenges or lower expected growth rates.

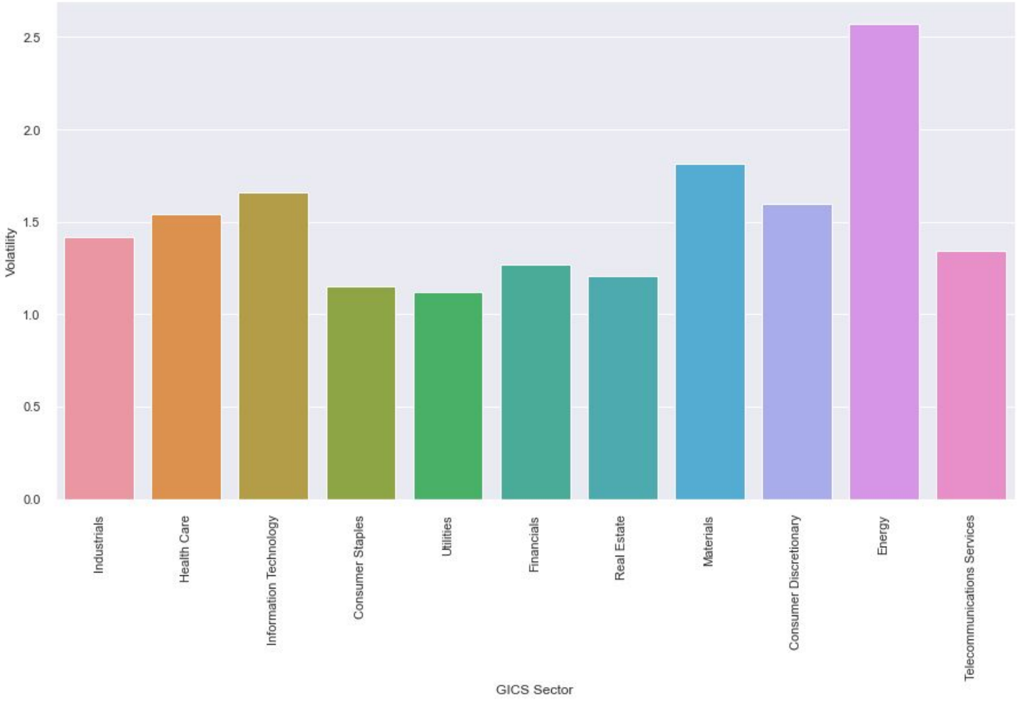


EDA Results

BIVARIATE ANALYSIS

We checked how volatility varies, on average, across economic sectors.

When GICS sectors are compared in terms of volatility, we found that the Energy sector is the most volatile with the highest standard deviation of all sectors at a deviation of about 2.6. Materials is the next highest sector at about 1.8 deviations. The sector with the least volatility is the Utilities sector at about 1.1, followed closely by Consumer Staples. This means that the Energy and Materials sectors are subject to greater fluctuations in stock prices than other sectors, which may make them more attractive to investors seeking higher potential returns but also exposes them to greater risks. On the other hand, the Utilities and Consumer Staples sectors are generally considered less risky investments due to their relatively stable stock prices. However, this may also result in lower potential returns.



DATA PREPROCESSING

Data Processing Overview

Data Preprocessing:

Since there are no missing or duplicate values, we do not have to treat the data for this.

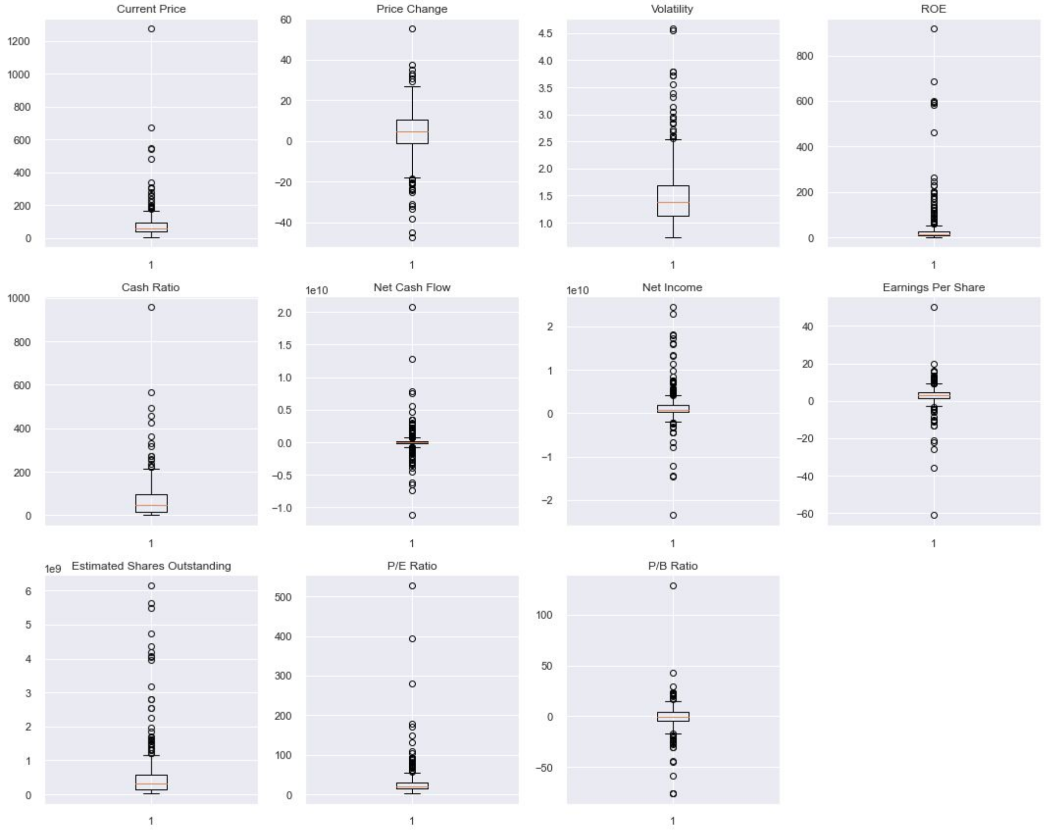
Data Preparation for Clustering:

We scaled the data prior to clustering to ensure that all variables have equal importance in the analysis, and to avoid bias towards variables with larger scales or variances. By scaling the data, we can ensure that each variable contributes equally to the clustering process and avoid any distortion in the results due to differences in the scale of the variables.

Data Processing: Outlier Detection and Treatment

Outlier Check:

We plotted boxplots for each numerical attribute to visualize and detect outliers in the data. We observed numerous outliers in the plots. However, at this time, we have not removed or treated any outliers in the data because we need to determine if they are genuine values or errors. Outliers are a common occurrence in stock data, as fluctuations in the stock market are driven by several factors such as changes in company and industry performance, natural disasters, political events, trends, innovations, and more. Therefore, it is essential to first understand and verify the data before making any decisions about how to treat outliers so for now, we will leave the outliers as is.

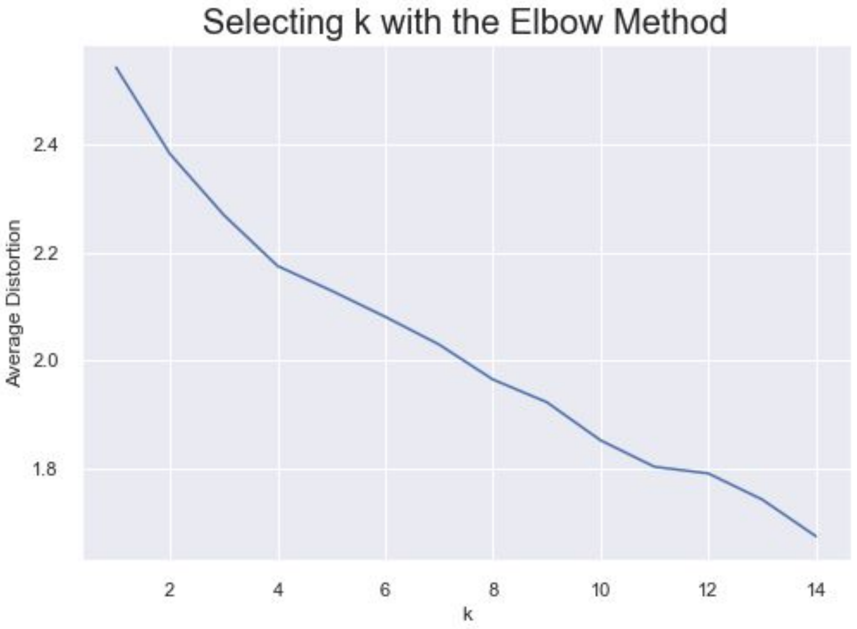


K-MEANS CLUSTERING: ELBOW METHOD

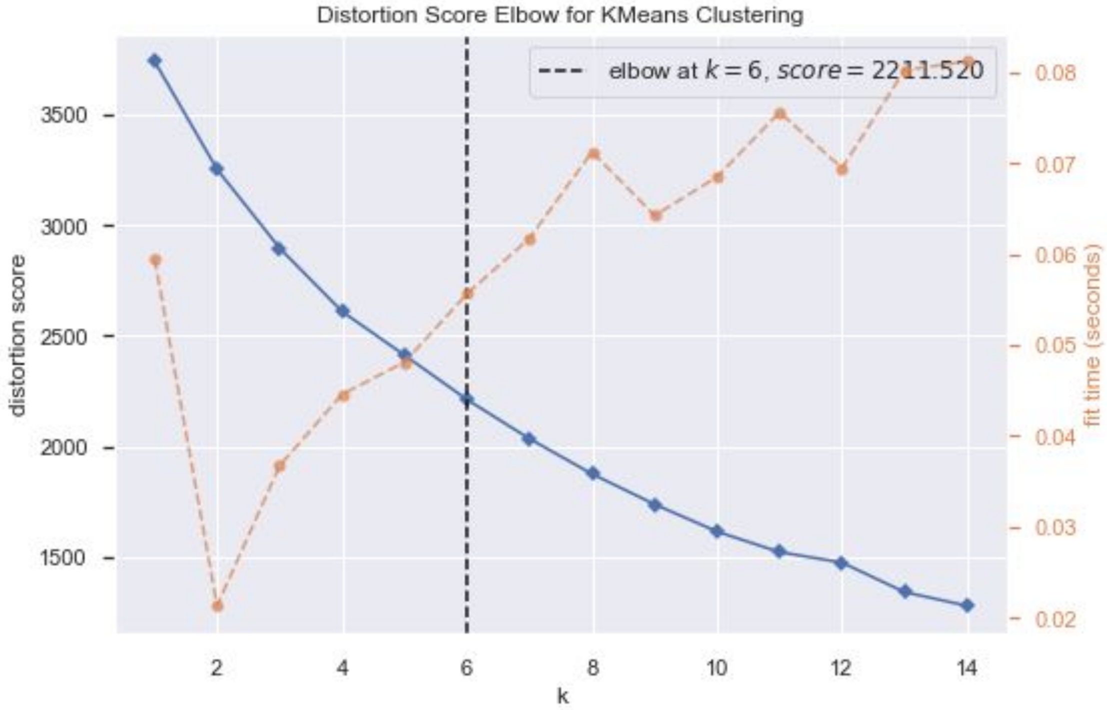
K-Means Clustering: Elbow Plot

Average Distortion Scores:

```
Number of Clusters: 1, Average Distortion: 2.5425
Number of Clusters: 2, Average Distortion: 2.3823
Number of Clusters: 3, Average Distortion: 2.2692
Number of Clusters: 4, Average Distortion: 2.1745
Number of Clusters: 5, Average Distortion: 2.1287
Number of Clusters: 6, Average Distortion: 2.0804
Number of Clusters: 7, Average Distortion: 2.0289
Number of Clusters: 8, Average Distortion: 1.9641
Number of Clusters: 9, Average Distortion: 1.9221
Number of Clusters: 10, Average Distortion: 1.8513
Number of Clusters: 11, Average Distortion: 1.8024
Number of Clusters: 12, Average Distortion: 1.7900
Number of Clusters: 13, Average Distortion: 1.7417
Number of Clusters: 14, Average Distortion: 1.6735
```



Elbow Plot: Distortion Score Elbow for K-Means Clustering



Algorithm Evaluation: K-Means Clustering: Elbow Plot

We employed the K-means clustering algorithm on the scaled dataset with varying numbers of clusters from 1 to 14. To identify the optimal number of clusters for the dataset, we ran code to calculate the average distortion for each cluster size. We then plotted these distortions against the number of clusters and looked for the "elbow" point in the curve, which indicates the optimal number of clusters. We found that the graph had a curved line without any sharp points easily identifiable as our elbow point.

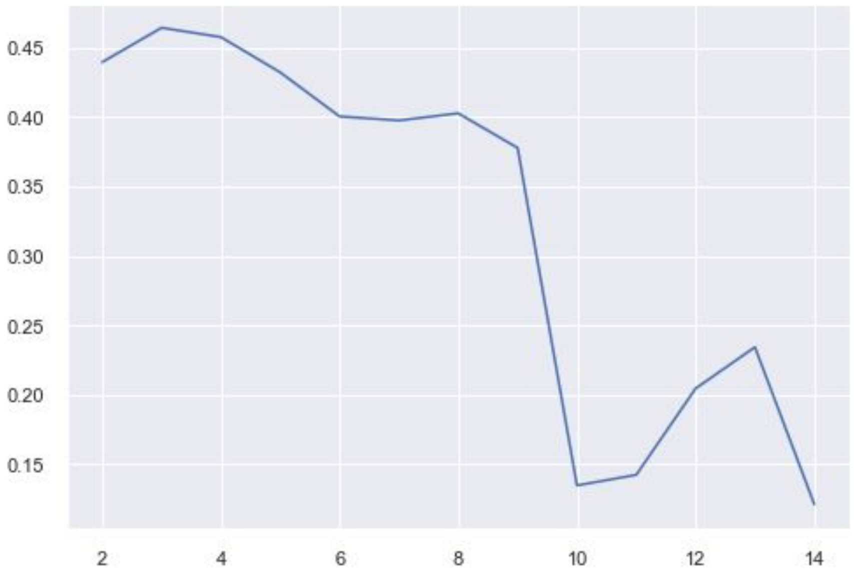
We then utilized the KElbowVisualizer, which fits the K-means model for each k value, computes the distortion score, and plots it on a line graph. We found that the elbow point in the graph occurred at k=6 (with a score of 2211.520), which suggests that six clusters best capture the underlying patterns in the data.

K-MEANS CLUSTERING: SILHOUETTE ANALYSIS

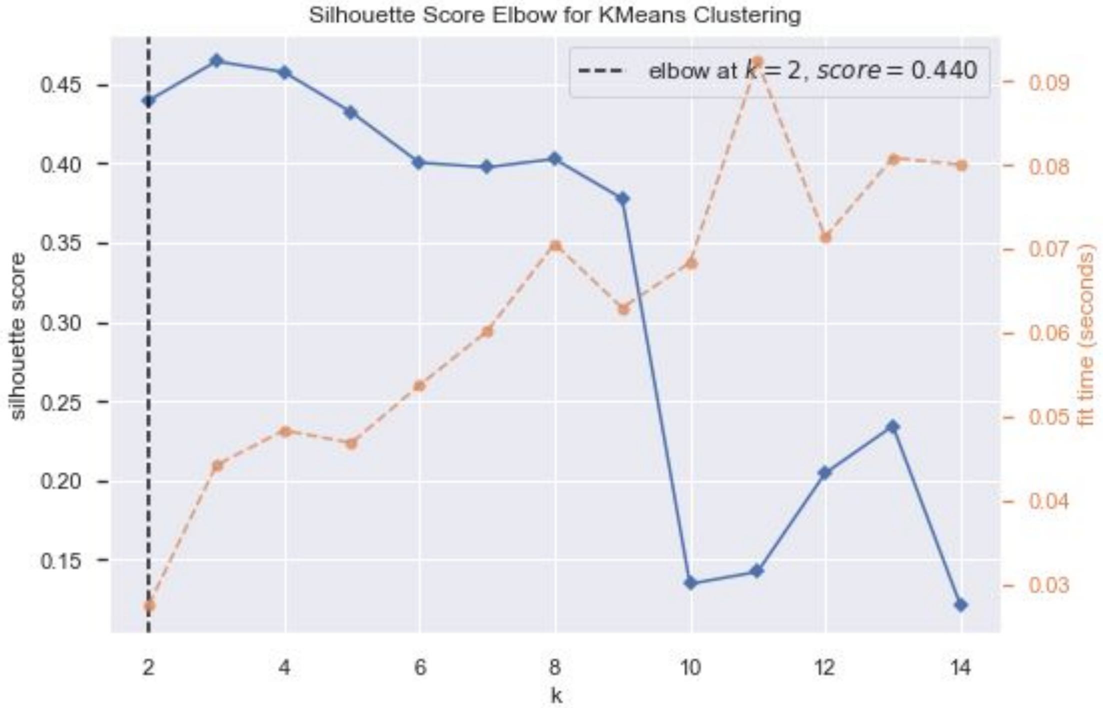
K-Means Clustering: Silhouette Analysis

Silhouette Scores:

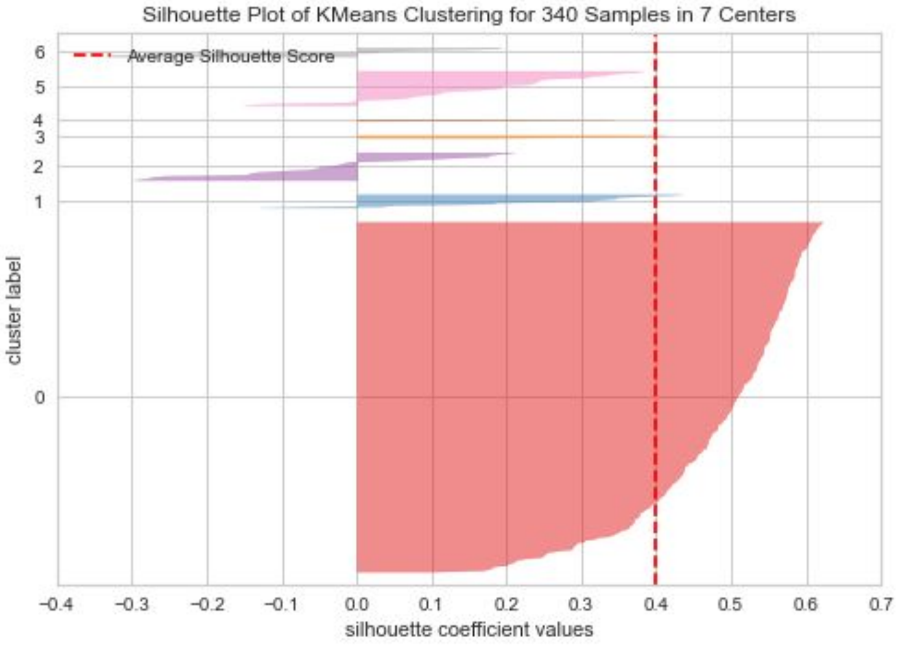
For n_clusters = 2, the silhouette score is 0.4396
For n_clusters = 3, the silhouette score is 0.4644
For n_clusters = 4, the silhouette score is 0.4577
For n_clusters = 5, the silhouette score is 0.4322
For n_clusters = 6, the silhouette score is 0.4005
For n_clusters = 7, the silhouette score is 0.3976
For n_clusters = 8, the silhouette score is 0.4027
For n_clusters = 9, the silhouette score is 0.3778
For n_clusters = 10, the silhouette score is 0.1345
For n_clusters = 11, the silhouette score is 0.1421
For n_clusters = 12, the silhouette score is 0.2044
For n_clusters = 13, the silhouette score is 0.2342
For n_clusters = 14, the silhouette score is 0.1210



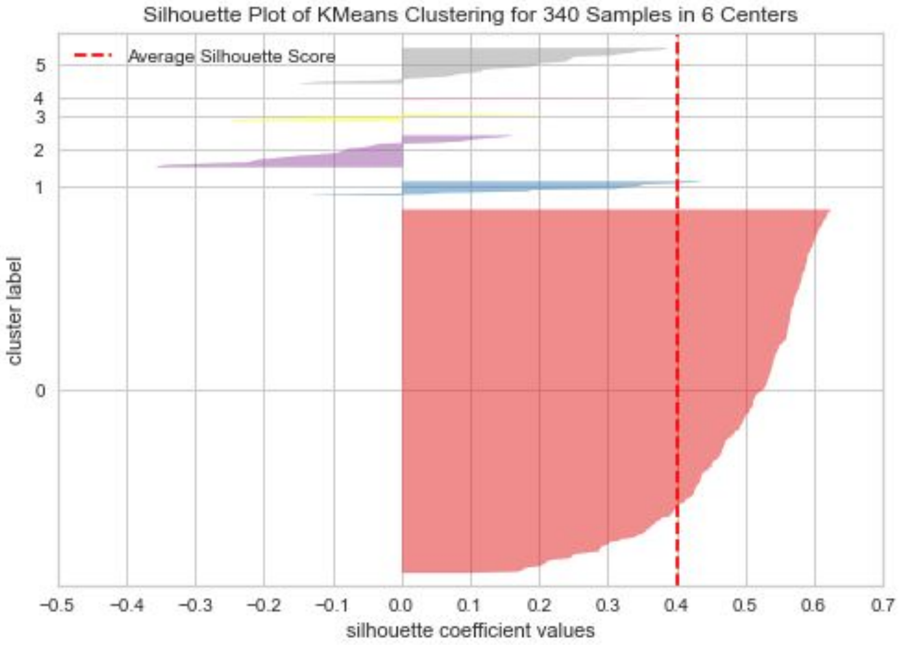
Silhouette Score Elbow for K-Means Clustering



Silhouette Visualizer: Plot of K-Means Clustering

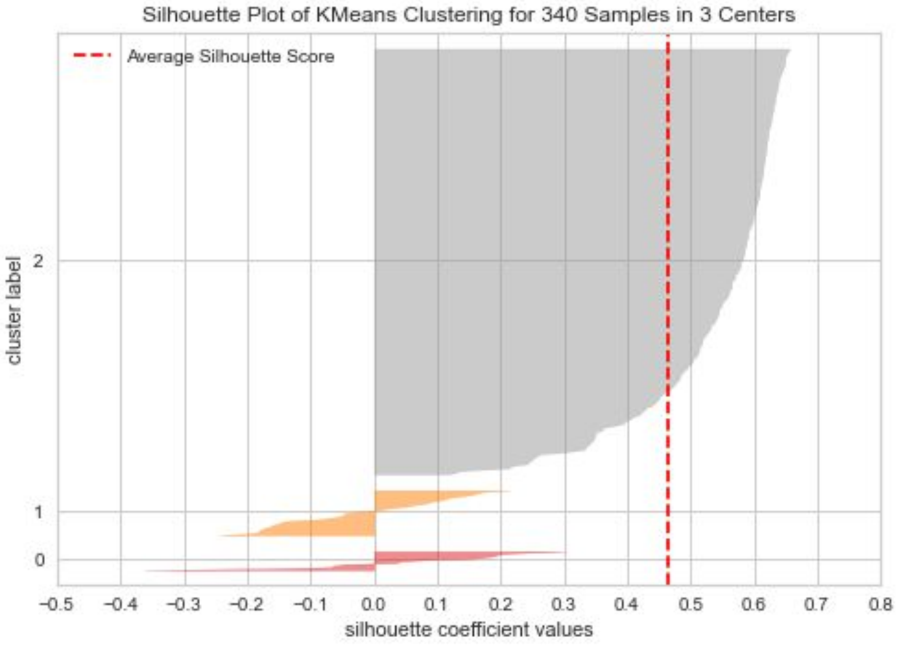


Visualizing 7 Clusters

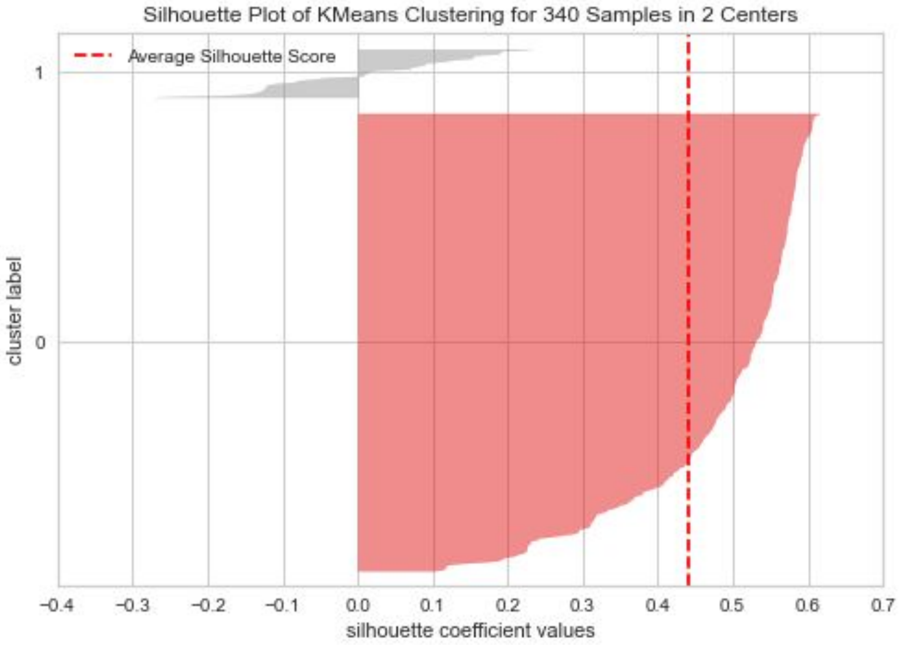


Visualizing 6 Clusters

Silhouette Visualizer: Plot of K-Means Clustering



Visualizing 3 Clusters



Visualizing 2 Clusters

Algorithm Evaluation: K-Means Clustering: Silhouette Score

The elbow method may not always provide a clear optimal number of clusters, especially if the data is complex or there is significant overlap between clusters so we employed the Silhouette score metric to assess the clustering performance for different values of k. Initially, we performed K-means clustering for a range of k values (2 to 14) and then calculated the Silhouette score for each clustering output. We then created a line plot of the Silhouette scores against the number of clusters to identify the optimal number of clusters for the data set. We utilized the KElbowVisualizer algorithm to compute the Silhouette score for each k value and plot the scores on a line graph. The elbow point in the graph, which represents the optimal number of clusters, is determined by the highest Silhouette score.

Our silhouette analysis revealed that the line plot has more than one elbow, with the optimal point at k=2. Notably, the line plot for k values displayed a steep descent after k=2, followed by a slight rise and a steady decline. Although the Silhouette scores for k=3 were the highest among all possible outcomes, we chose k=2 as our optimal number of clusters based on the returned k value from the elbow plot. The results suggest that k=2 can best capture the underlying patterns in the data.

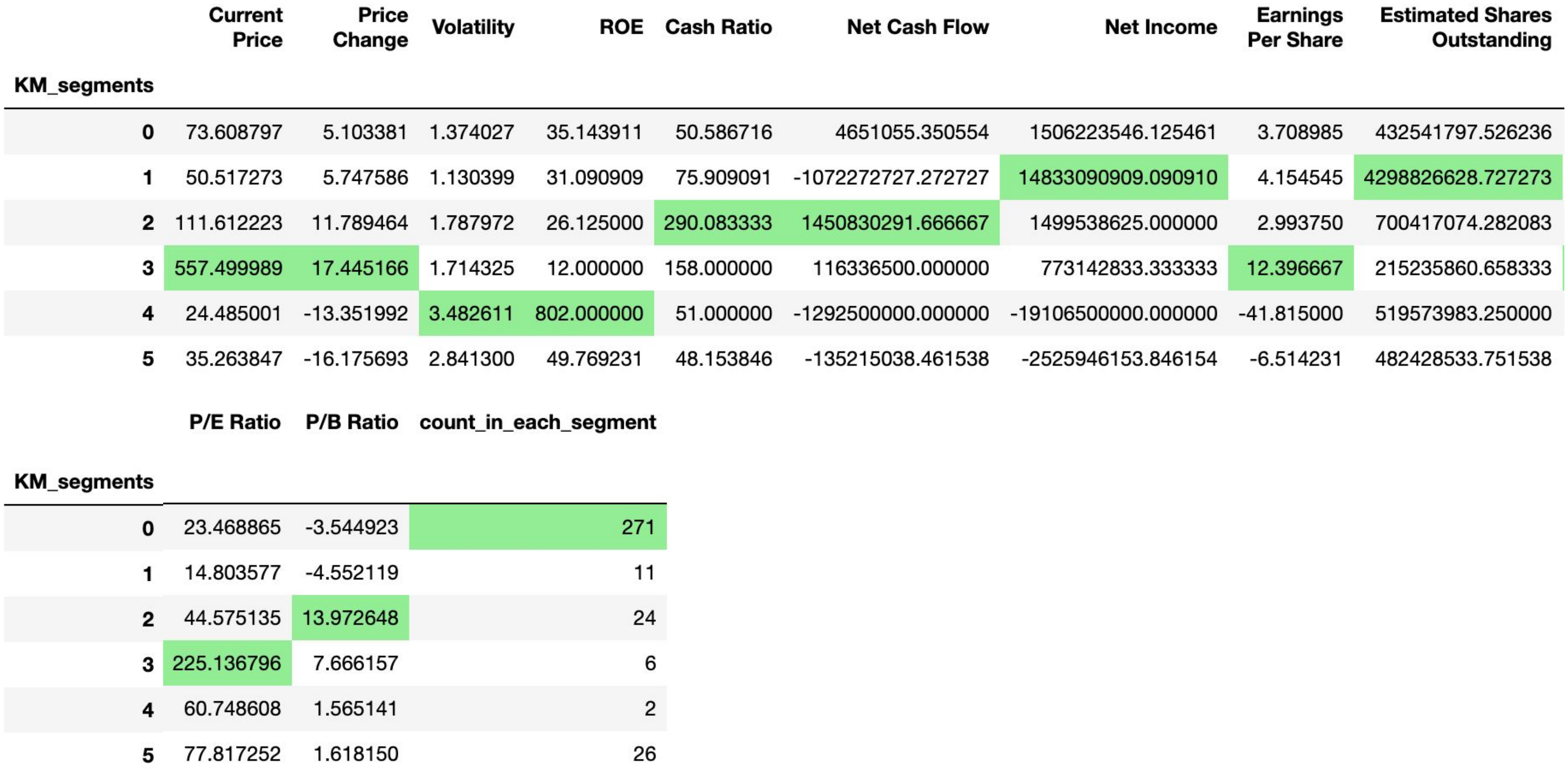
Model Building: K-Means (KM) Clustering

We obtained different optimal values of k using the Elbow Method and Silhouette Score. While the Elbow Method suggests k=6, the Silhouette Score suggests k=2, this means that these methods have different perspectives on what the optimal number of clusters should be.

However, upon further evaluation of the clusters formed by each method and their respective cluster profiles, I identified that reducing the cluster size will significantly affect the computation of averages.

Given the complex and risky nature of stock investment, I determined that using k=2 would oversimplify our data and potentially lead to financial harm for our clients. As such, we will use the Elbow Method's recommendation of k=6. This will allow us to make more informed decisions and minimize risks associated with investing and trading stocks.

Cluster Profile: KM Segments



KM_segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	count_in_each_segment
0	73.608797	5.103381	1.374027	35.143911	50.586716	4651055.350554	1506223546.125461	3.708985	432541797.526236	23.468865	-3.544923	271
1	50.517273	5.747586	1.130399	31.090909	75.909091	-1072272727.272727	14833090909.090910	4.154545	4298826628.727273	14.803577	-4.552119	11
2	111.612223	11.789464	1.787972	26.125000	290.083333	1450830291.666667	1499538625.000000	2.993750	700417074.282083	44.575135	13.972648	24
3	557.499989	17.445166	1.714325	12.000000	158.000000	116336500.000000	773142833.333333	12.396667	215235860.658333	225.136796	7.666157	6
4	24.485001	-13.351992	3.482611	802.000000	51.000000	-1292500000.000000	-19106500000.000000	-41.815000	519573983.250000	60.748608	1.565141	2
5	35.263847	-16.175693	2.841300	49.769231	48.153846	-135215038.461538	-2525946153.846154	-6.514231	482428533.751538	77.817252	1.618150	26

K-Means Cluster Profile Evaluation

Upon analyzing the stock data, we have identified some distinct segments with unique characteristics:

Segment 1, which contains 11 stock options, has the highest net income and estimated shares outstanding, indicating strong financial performance and growth potential.
Segment 2, with 24 stock options, stands out for its high cash ratio, net cash flow, and P/B ratio, suggesting a financially stable and undervalued group of stocks.
Segment 3, consisting of 6 stock options, is characterized by the highest current price, price change, and P/E ratio, indicating strong market performance and growth potential.
Segment 4, which contains only 2 stock options, has the highest volatility and Return on Earnings, making it a high-risk, high-reward investment opportunity.

By identifying pattern within these distinct segments, we can better understand the strengths and weaknesses of each stock option and make more informed investment decisions based on our risk tolerance and investment objectives.

Cluster Segment: Companies & their Clusters

In cluster 5, the following companies are present:
['Anadarko Petroleum Corp' 'Baker Hughes Inc' 'Cabot Oil & Gas' 'Concho Resources' 'Devon Energy Corp.' 'EOG Resources' 'EQT Corporation' 'Freeport-McMoran Cp & Gld' 'Hess Corporation' 'Hewlett Packard Enterprise' 'Kinder Morgan' 'The Mosaic Company' 'Marathon Oil Corp.' 'Murphy Oil' 'Noble Energy Inc' 'Newfield Exploration Co' 'National Oilwell Varco Inc.' 'ONEOK' 'Occidental Petroleum' 'Quanta Services Inc.' 'Range Resources Corp.' 'Spectra Energy Corp.' 'Southwestern Energy' 'Teradata Corp.' 'Williams Cos.' 'Cimarex Energy']

In cluster 4, the following companies are present:
['Apache Corporation' 'Chesapeake Energy']

In cluster 3, the following companies are present:
['Alexion Pharmaceuticals' 'Amazon.com Inc' 'Intuitive Surgical Inc.' 'Netflix Inc.' 'Priceline.com Inc' 'Regeneron']

In cluster 2, the following companies are present:
['Adobe Systems Inc' 'Analog Devices, Inc.' 'Alliance Data Systems' 'Amgen Inc' 'Broadcom' 'Bank of America Corp' 'Celgene Corp.' 'Chipotle Mexican Grill' 'eBay Inc.' 'Equinix' 'Edwards Lifesciences' 'Facebook' 'First Solar Inc' 'Frontier Communications' 'Halliburton Co.' "McDonald's Corp." 'Monster Beverage' 'Newmont Mining Corp. (Hldg. Co.)' 'Skyworks Solutions' 'TripAdvisor' 'Vertex Pharmaceuticals Inc' 'Waters Corporation' 'Wynn Resorts Ltd' 'Yahoo Inc.']

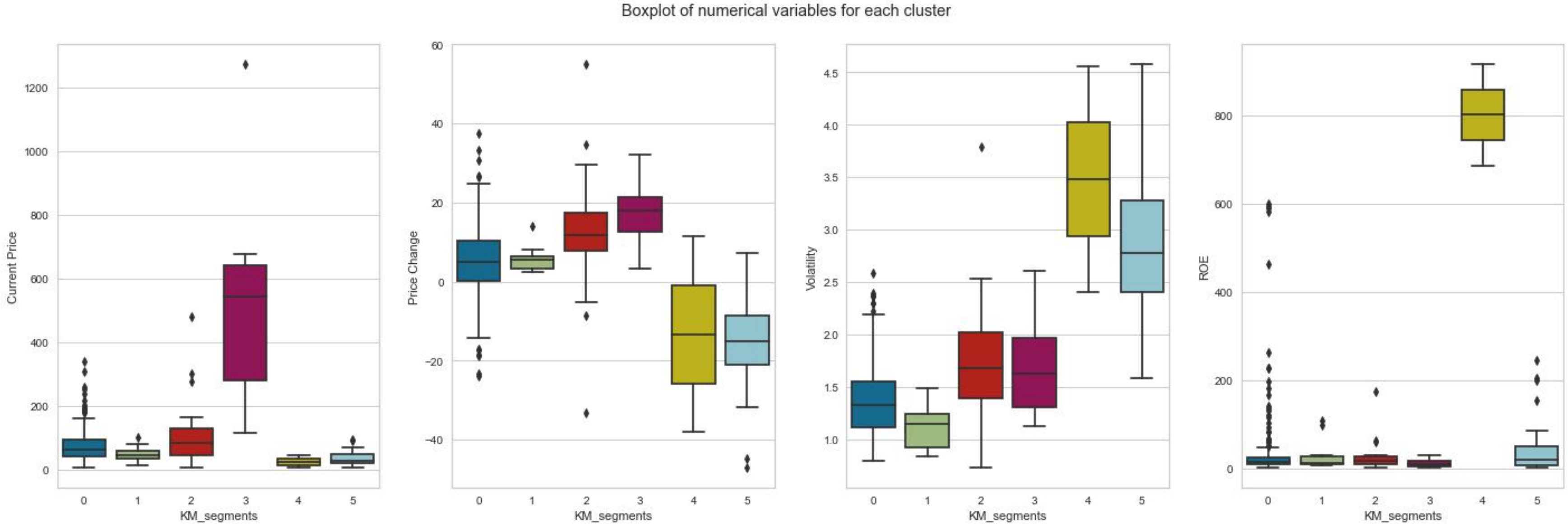
In cluster 1, the following companies are present:
['Citigroup Inc.' 'Ford Motor' 'Gilead Sciences' 'Intel Corp.' 'JPMorgan Chase & Co.' 'Coca Cola Company' 'Pfizer Inc.' 'AT&T Inc' 'Verizon Communications' 'Wells Fargo' 'Exxon Mobil Corp.']

In cluster 0, all other companies not listed in the above clusters are present.

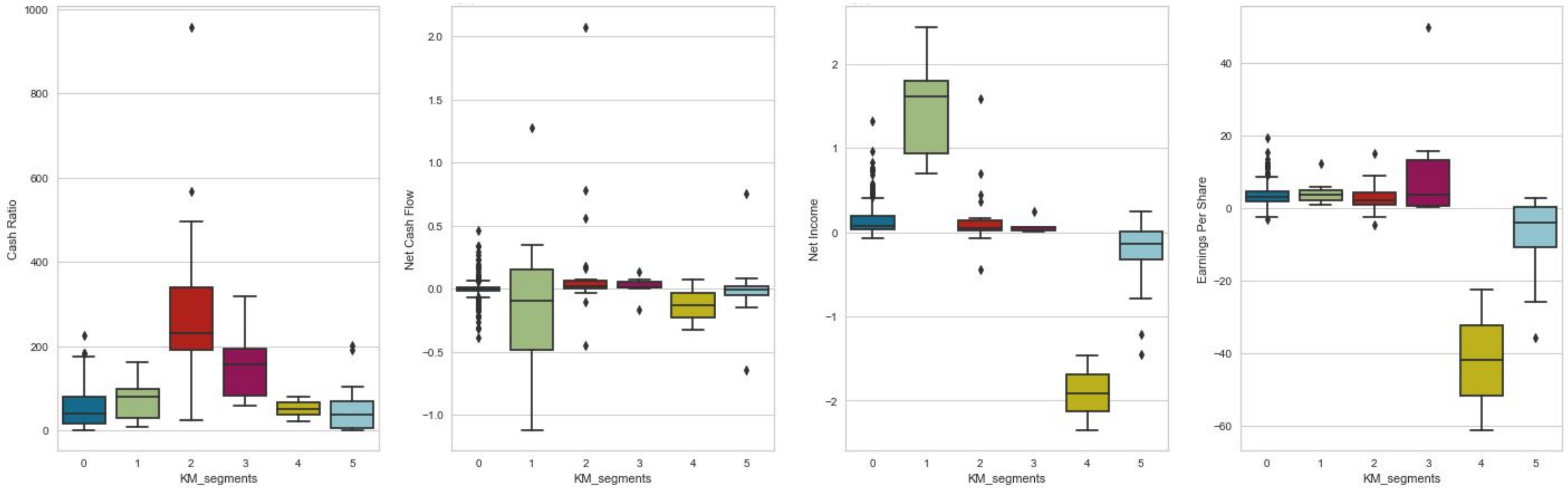
Cluster Segment: KM_Segment by GICS Sector

KM_Segments & GICS Sector	0	1	2	3	4	5
Consumer Discretionary	33	1	4	2	0	0
Consumer Staples	17	1	1	0	0	0
Energy	5	1	1	0	2	21
Financials	45	3	1	0	0	0
Health Care	30	2	5	3	0	0
Industrials	52	0	0	0	0	1
Information Technology	20	1	9	1	0	2
Materials	17	0	1	0	0	2
Real Estate	26	0	1	0	0	0
Telecommunications Services	2	2	1	0	0	0
Utilities	24	0	0	0	0	0

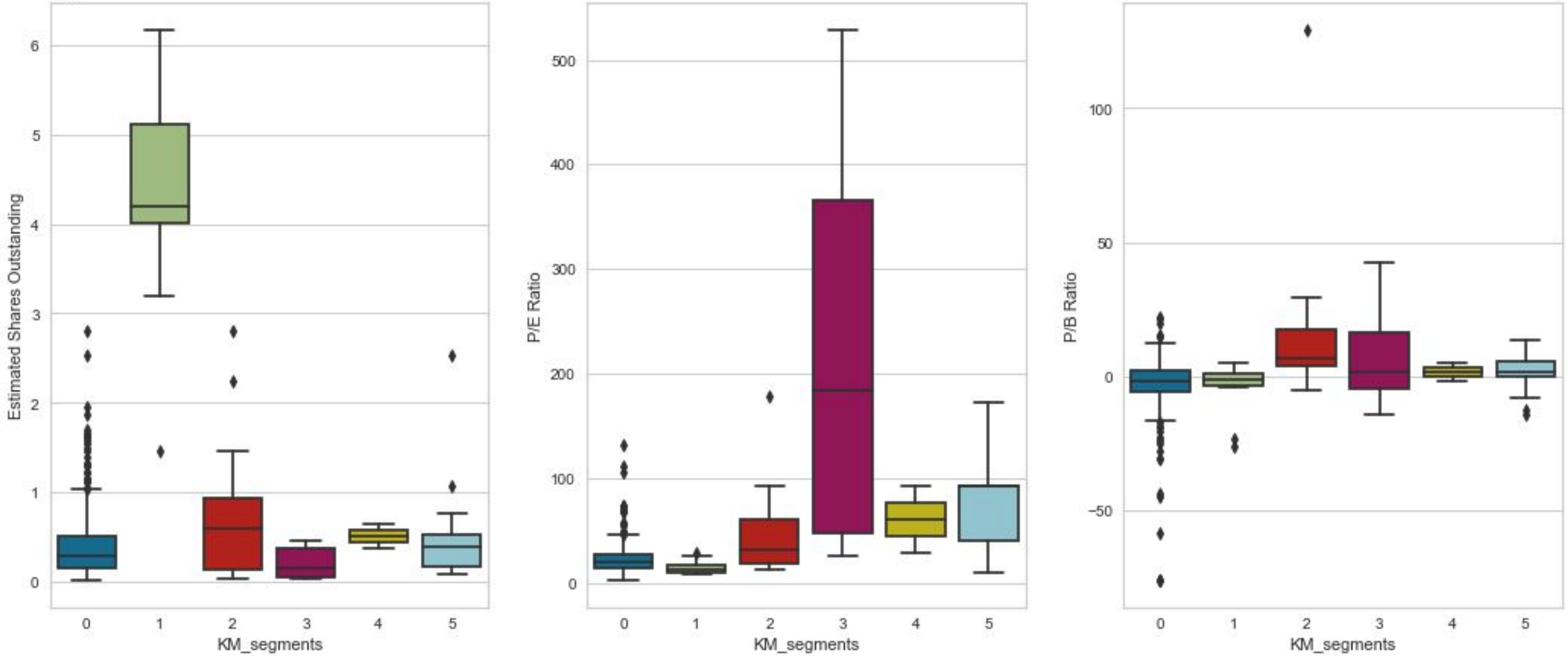
Visualizing Numerical Variables of KM_Segments



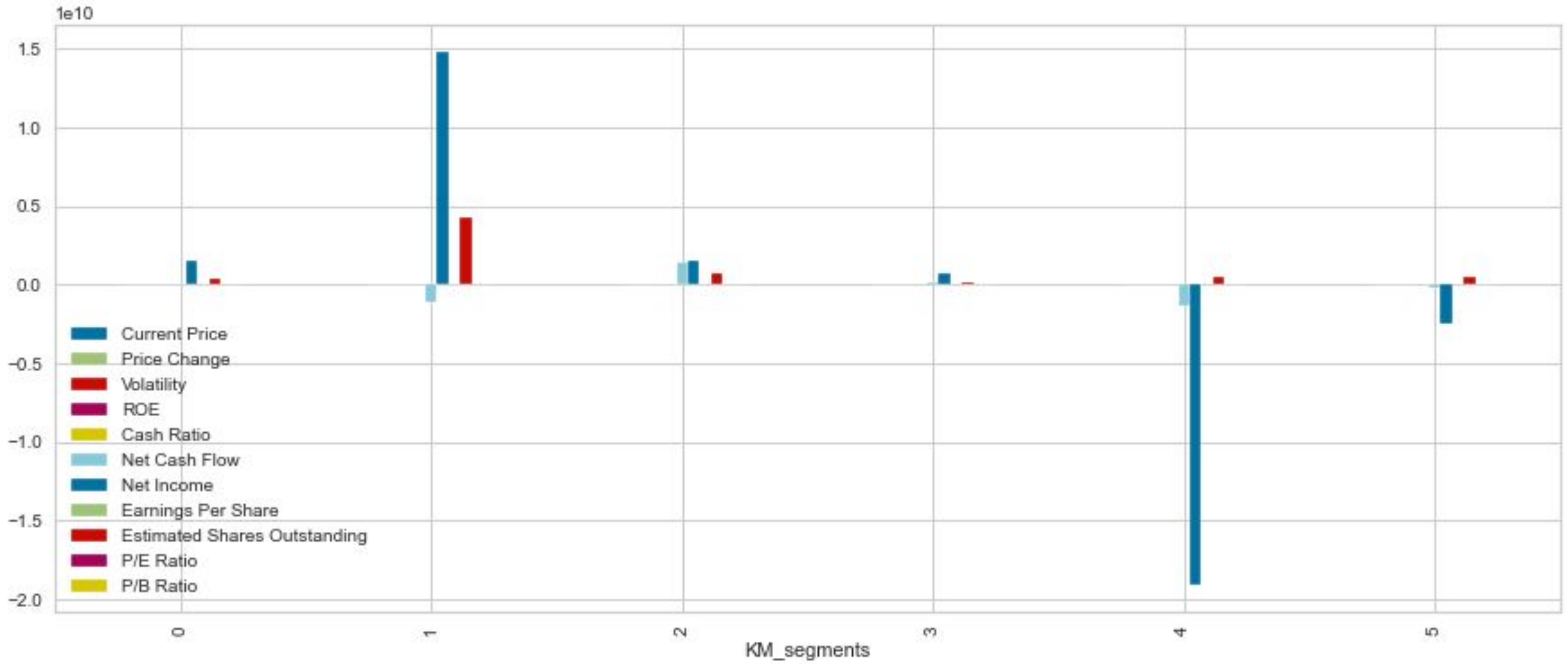
Visualizing Numerical Variables of KM_Segments



Visualizing Numerical Variables of KM_Segments



Visualizing Numerical Variables of KM_Segments



HIERARCHICAL CLUSTERING

Cophenetic Correlation

Combining Distance Metrics and Linkage Methods:

```
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 0.9259195530524591.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.7925307202850002.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.9247324030159736.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.8708317490180427.
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, obtained with Euclidean distance and average linkage.
```

Cophenetic Correlation: Euclidean Distance

We explored different linkage methods with Euclidean distance only:

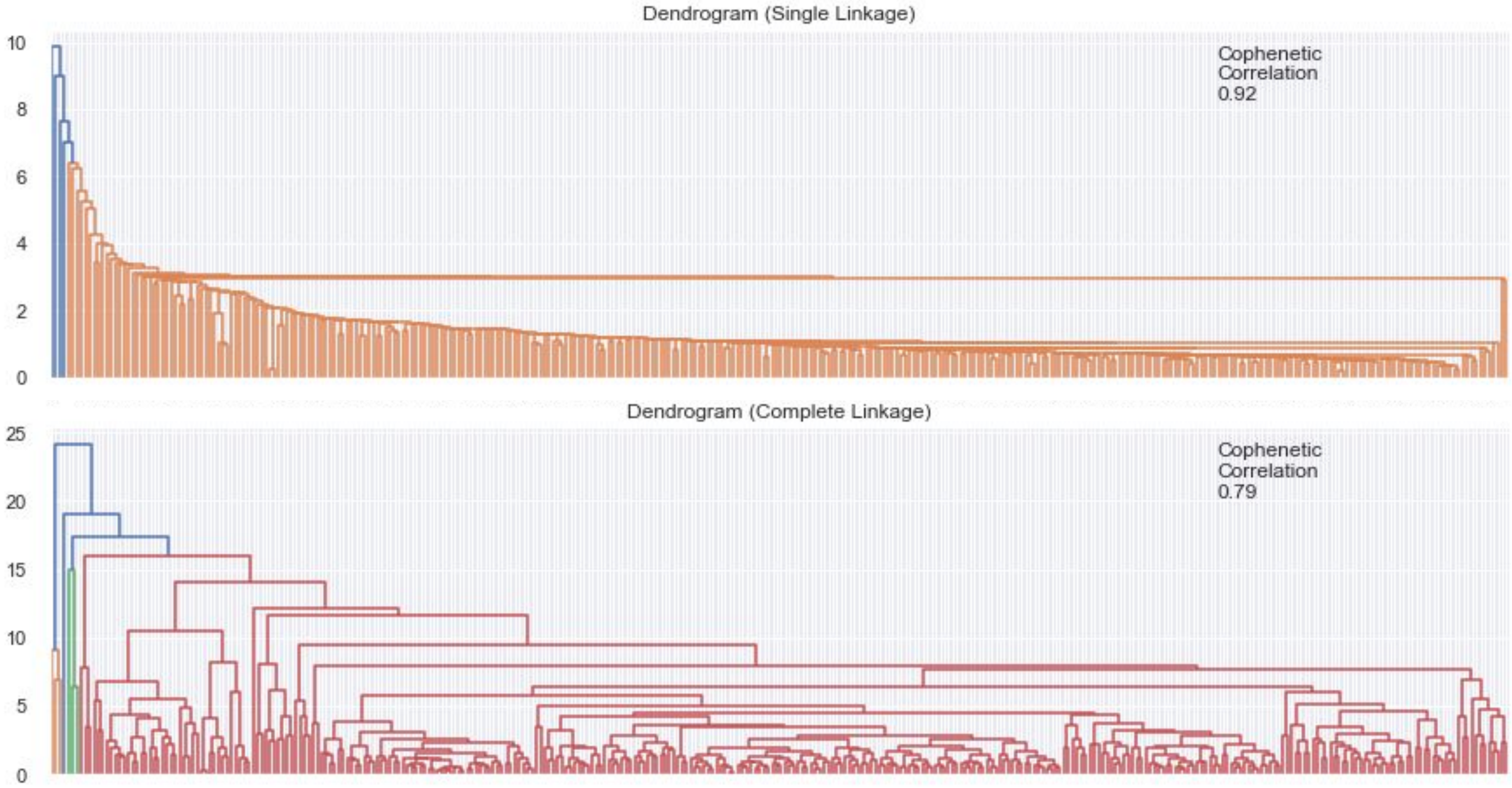
```
Cophenetic correlation for single linkage is 0.9232271494002922.
Cophenetic correlation for complete linkage is 0.7873280186580672.
Cophenetic correlation for average linkage is 0.9422540609560814.
Cophenetic correlation for weighted linkage is 0.8693784298129404.
Cophenetic correlation for centroid linkage is 0.9314012446828154.
Cophenetic correlation for ward linkage is 0.7101180299865353.
Cophenetic correlation for median linkage is 0.9198690668829905.
*****************************************************************************************************************
Highest  cophenetic correlation is still 0.9422540609560814, obtained with average linkage.
```

Algorithm Evaluation: Hierarchical Clustering (HC)

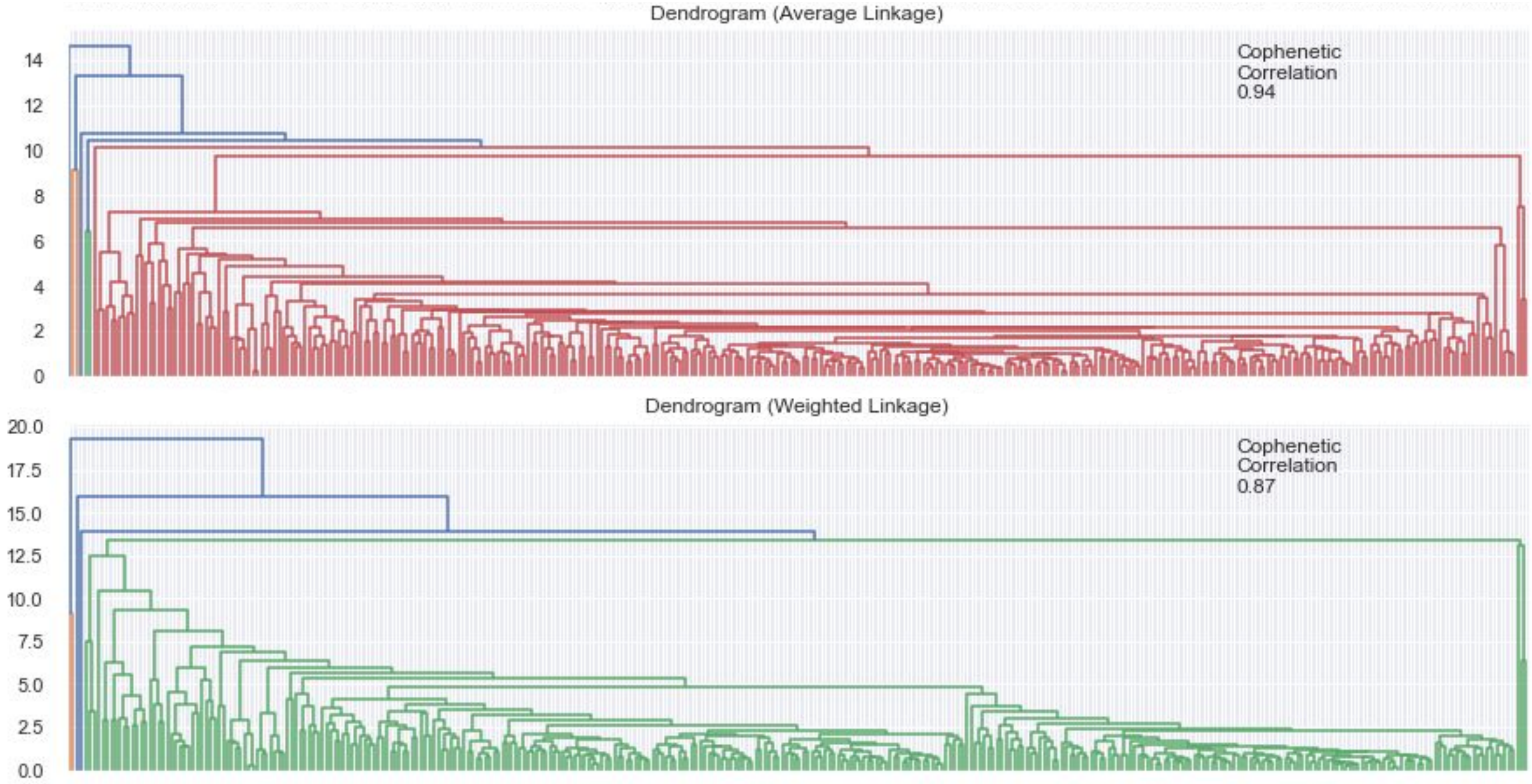
We conducted hierarchical clustering on our scaled dataset, utilizing various distance metrics and linkage methods. We aimed to identify the combination of distance metric and linkage method that produced the highest cophenetic correlation coefficient. After exploring different options, we found that Euclidean distance, in conjunction with average linkage, generated the highest cophenetic correlation coefficient of 0.9422.

This coefficient serves as an indicator of the quality of the clustering results, and suggests that this particular combination of distance metric and linkage method was most effective in preserving the original pairwise distances between data points in our hierarchical clustering.

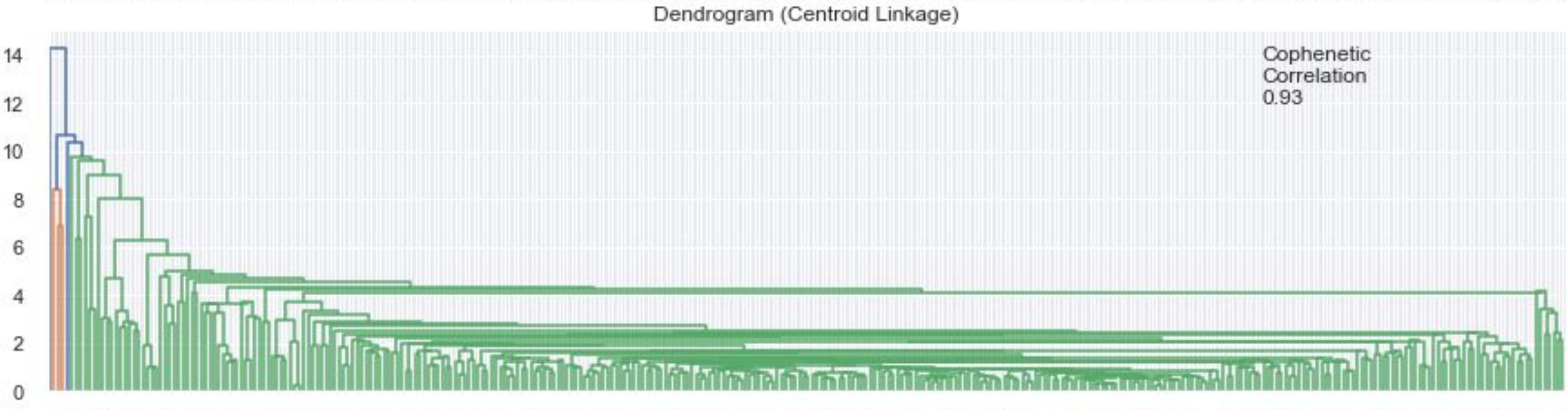
Dendrograms: Different Linkage Methods using Euclidean Distance

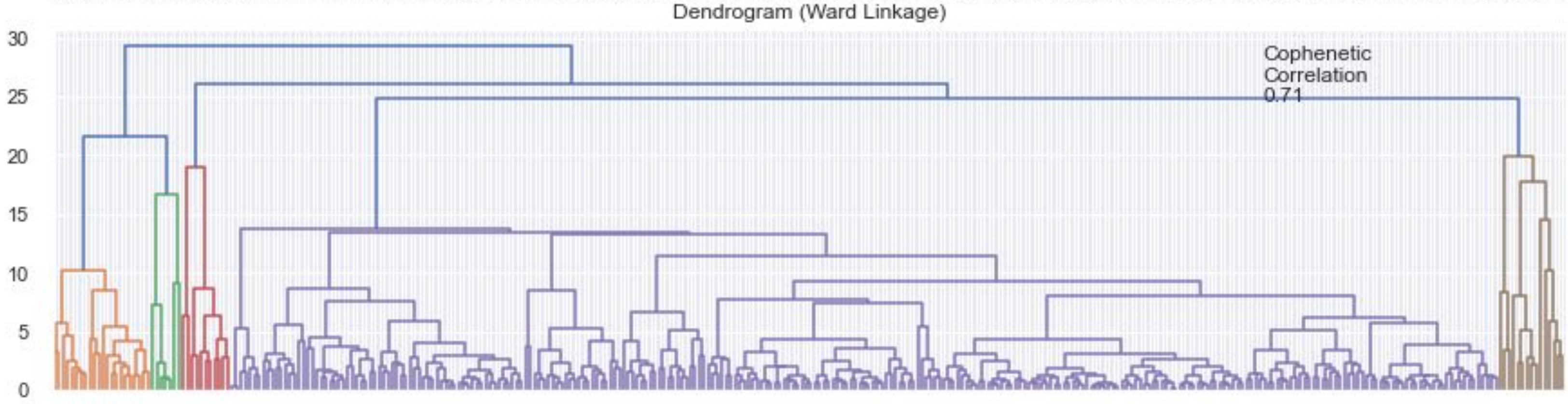


Dendrograms: Different Linkage Methods using Euclidean Distance

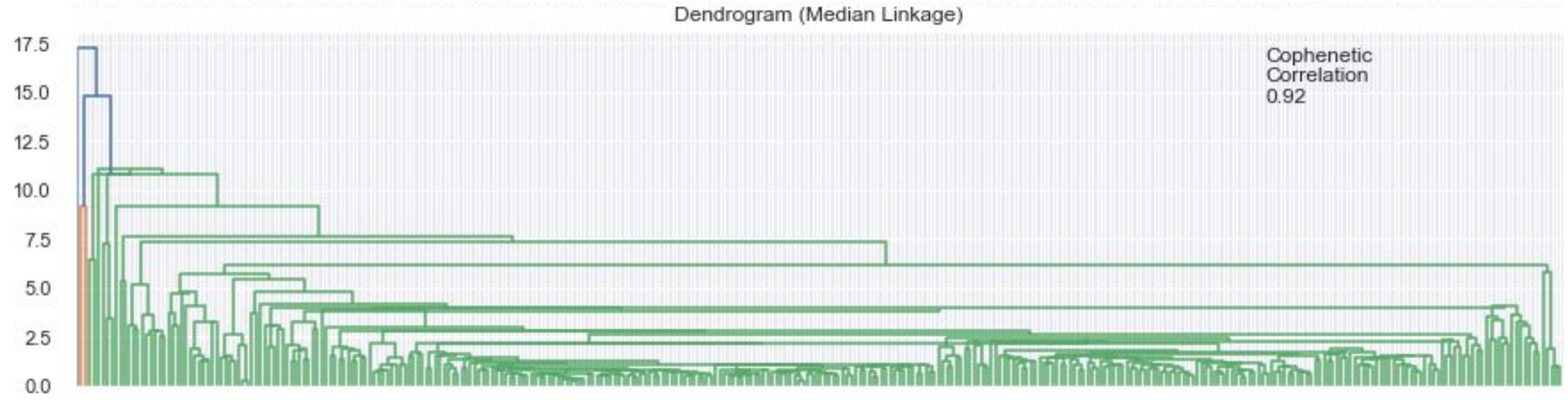


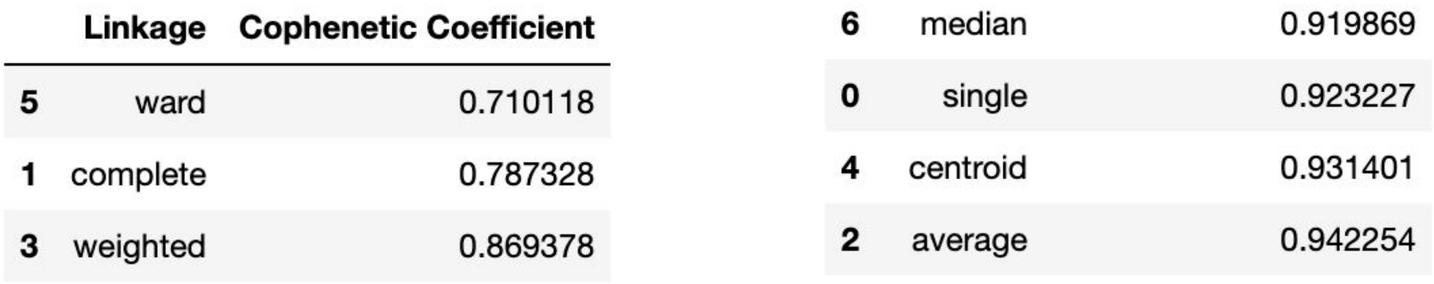
Dendrograms: Different Linkage Methods using Euclidean Distance





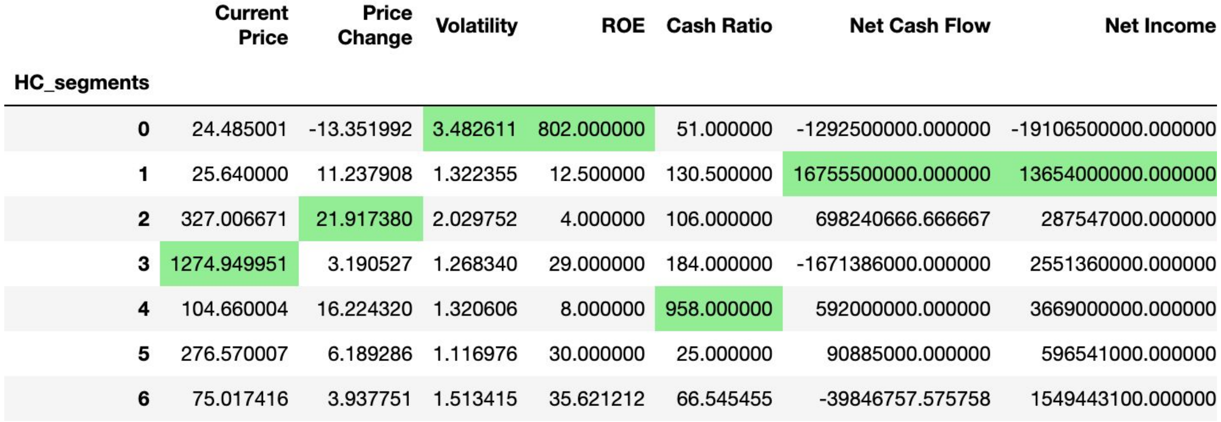
Dendrograms: Different Linkage Methods using Euclidean Distance



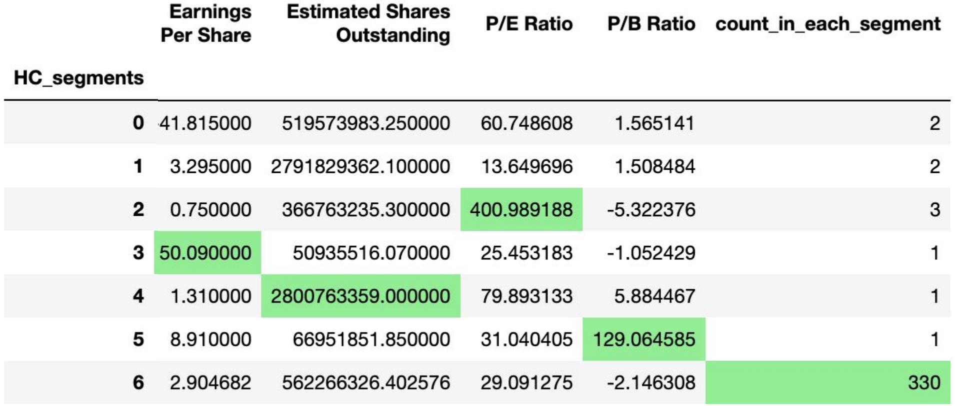


	Linkage	Cophenetic Coefficient
5	ward	0.710118
1	complete	0.787328
3	weighted	0.869378
6	median	0.919869
0	single	0.923227
4	centroid	0.931401
2	average	0.942254

Cluster Profile: HC Segments



HC_segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income
0	24.485001	-13.351992	3.482611	802.000000	51.000000	-1292500000.000000	-19106500000.000000
1	25.640000	11.237908	1.322355	12.500000	130.500000	16755500000.000000	13654000000.000000
2	327.006671	21.917380	2.029752	4.000000	106.000000	698240666.666667	287547000.000000
3	1274.949951	3.190527	1.268340	29.000000	184.000000	-1671386000.000000	2551360000.000000
4	104.660004	16.224320	1.320606	8.000000	958.000000	592000000.000000	3669000000.000000
5	276.570007	6.189286	1.116976	30.000000	25.000000	90885000.000000	596541000.000000
6	75.017416	3.937751	1.513415	35.621212	66.545455	-39846757.575758	1549443100.000000



HC_segments	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	count_in_each_segment
0	41.815000	519573983.250000	60.748608	1.565141	2
1	3.295000	2791829362.100000	13.649696	1.508484	2
2	0.750000	366763235.300000	400.989188	-5.322376	3
3	50.090000	50935516.070000	25.453183	-1.052429	1
4	1.310000	2800763359.000000	79.893133	5.884467	1
5	8.910000	66951851.850000	31.040405	129.064585	1
6	2.904682	562266326.402576	29.091275	-2.146308	330

Hierarchical Cluster Profile Evaluation

After performing clustering analysis on the stock dataset, we identified distinct segments based on the selected attributes. Each segment represents a group of stocks that share similar characteristics. The characteristics of each segment are summarized as follows:

Segment 0: This segment contains 2 stock options with the highest volatility and Return on Equity (ROE) among all the clusters. These stocks may be suitable for investors looking for high-risk, high-reward investments.
Segment 1: This segment contains 2 stock options with the highest net cash flow and net income. These stocks may be suitable for investors seeking stable and profitable investments.
Segment 2: This segment contains 3 stock options with the highest price change and P/E ratio. These stocks may be suitable for investors who prioritize growth potential.
Segment 3: This segment contains 1 stock option with the highest current price and earnings per share. This stock may be suitable for investors who prioritize stable and high-earning investments.
Segment 4: This segment contains 1 stock option with the highest cash ratio and estimated shares outstanding. This stock may be suitable for investors who value financial stability.
Segment 5: This segment contains 1 stock option with the highest P/B Ratio. This stock may be suitable for investors who prioritize value investing strategies.

Overall, these segments provide investors with valuable insights into the characteristics of each stock option, allowing them to make informed investment decisions based on their individual investment goals and risk tolerance levels.

Model Building: Hierarchical Clustering

After obtaining the dendrogram with the highest cophenetic correlation, I decided on the optimal number of clusters for hierarchical clustering. To accomplish this, I examined cluster values ranging from 2 to 9 and assessed them against the cluster profile.

While comparing results, I identified that as the number of clusters gets smaller, we lose accuracy for simplicity, and this could affect the computation of averages significantly. Therefore, I considered how the average scores changed as clusters became smaller, similar to what was done in K-Means clustering. Additionally, I selected six key evaluators from the attributes that are essential in evaluating stock/investment options. I monitored the average scores of these evaluators as we varied the number of clusters and found that n_clusters=7 preserved more accuracy in terms of combined data points, while allowing us to evaluate more segments based on a more accurate understanding of their pros and cons for the investor.

Since investing in stocks can be a risky endeavor, it is best to avoid oversimplifying our data into very small groups in a way that could be financially harmful and irresponsible to our clients. Therefore, I recommend using n_clusters=7 to make more informed decisions and minimize risks associated with investing and trading stocks.

N_Clusters: 6 Key Evaluators

1. Current Price: The current stock price is important because it indicates the cost of purchasing a share of the company's stock.
2. ROE: Return on equity is a key metric that measures a company's profitability and efficiency in using shareholder equity to generate profits.
3. Net Income: A company's net income is an important indicator of its financial health and profitability.
4. Earnings Per Share: Earnings per share (EPS) is a key metric that indicates the amount of profit a company has generated per share of stock outstanding.
5. P/E Ratio: The price-to-earnings ratio (P/E ratio) is a widely used metric that indicates how much investors are willing to pay for each dollar of earnings generated by the company.
6. P/B Ratio: The price-to-book ratio (P/B ratio) compares a company's stock price to its book value per share, which provides an indication of the company's valuation relative to its underlying assets.

Cluster Segment: Companies & their Clusters

In cluster 0, the following companies are present:
['Apache Corporation' 'Chesapeake Energy']

In cluster 1, the following companies are present:
['Bank of America Corp' 'Intel Corp.']

In cluster 2, the following companies are present:
['Alexion Pharmaceuticals' 'Amazon.com Inc' 'Netflix Inc.']

In cluster 3, the following companies are present:
['Priceline.com Inc']

In cluster 4, the following companies are present:
['Facebook']

In cluster 5, the following companies are present:
['Alliance Data Systems']

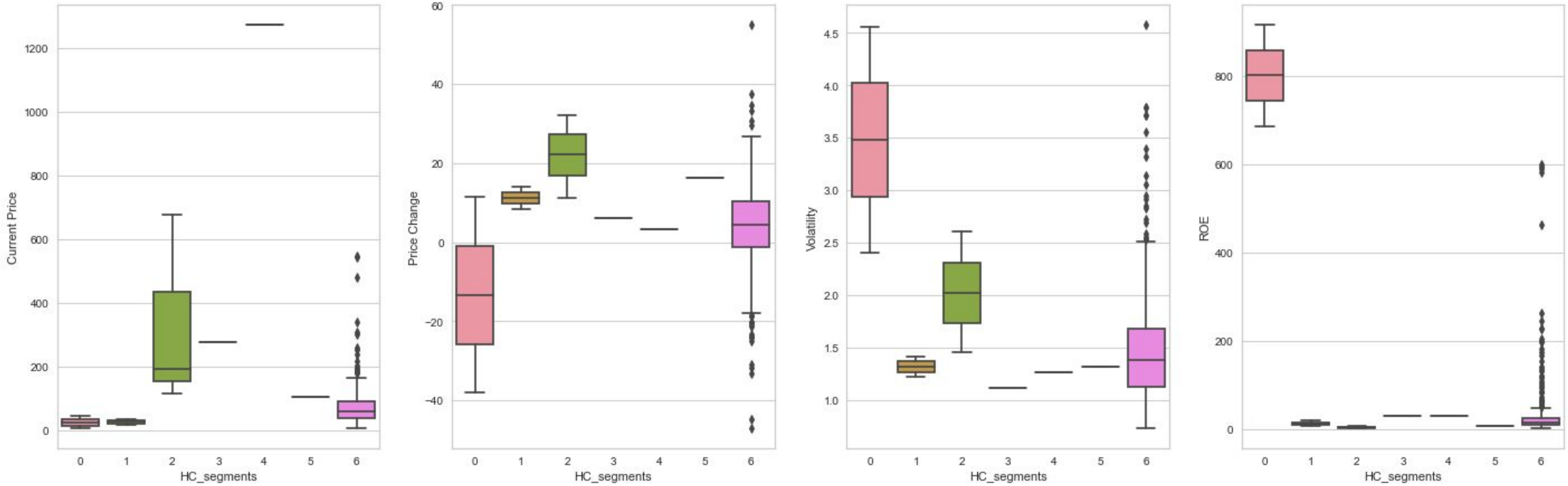
In cluster 6, all other companies not listed in the above clusters are present.

Cluster Segment: HC_Segment by GICS Sector

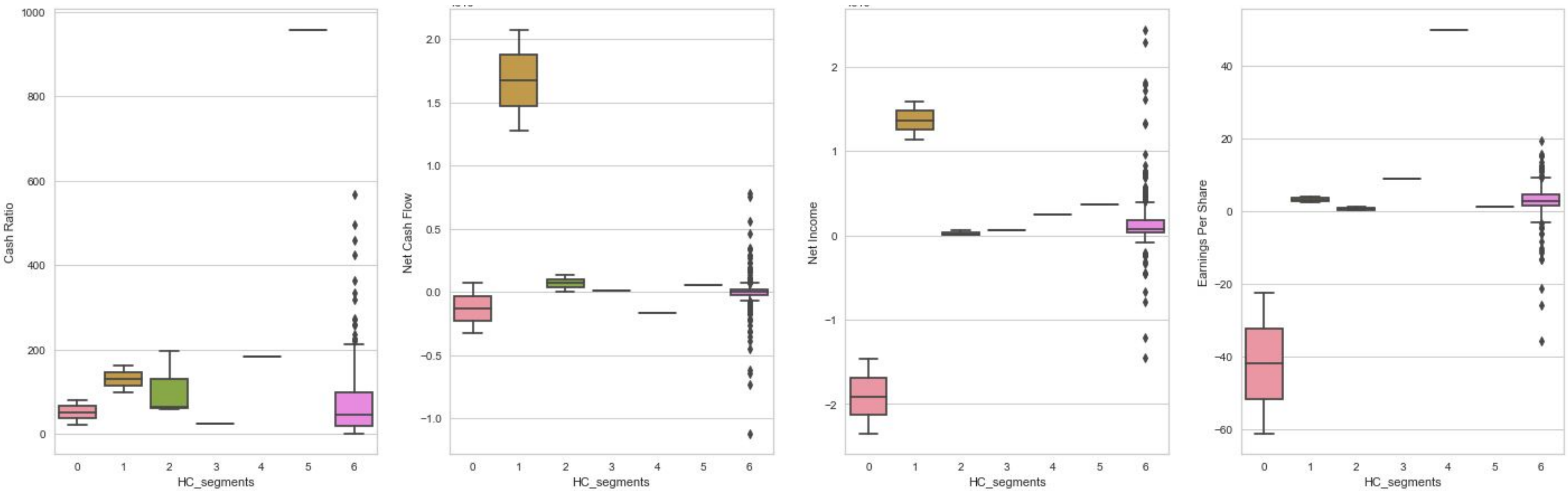
HC_Segments & GICS Sector	0	1	2	3	4	5	6
Consumer Discretionary	0	0	1	0	1	0	38
Consumer Staples	0	0	0	0	0	0	19
Energy	2	0	0	0	0	0	28
Financials	0	1	0	0	0	0	48
Health Care	0	0	1	0	0	0	39
Industrials	0	0	0	0	0	0	53
Information Technology	0	1	1	1	0	1	29
Materials	0	0	0	0	0	0	20
Real Estate	0	0	0	0	0	0	27
Telecommunications Services	0	0	0	0	0	0	5
Utilities	0	0	0	0	0	0	24

Visualizing Numerical Variables of HC_Segments

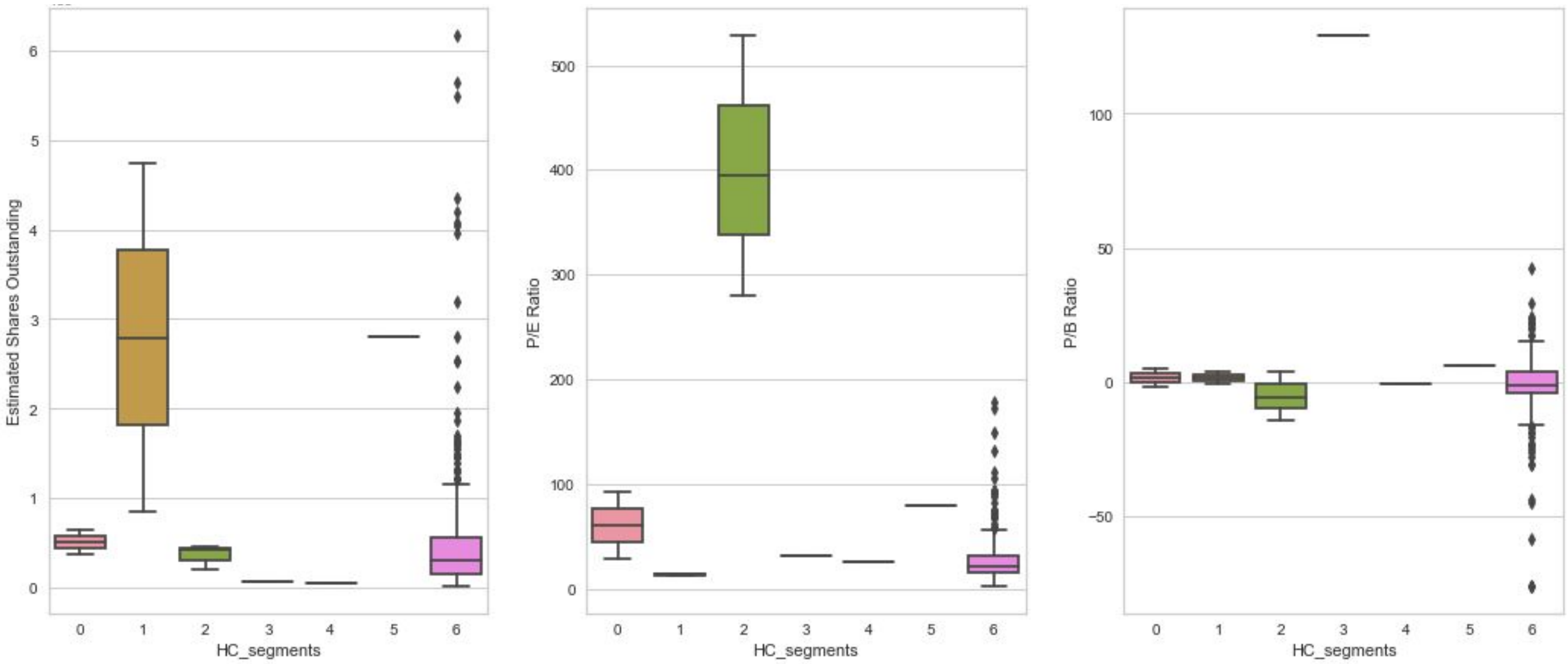
Boxplot of numerical variables for each cluster



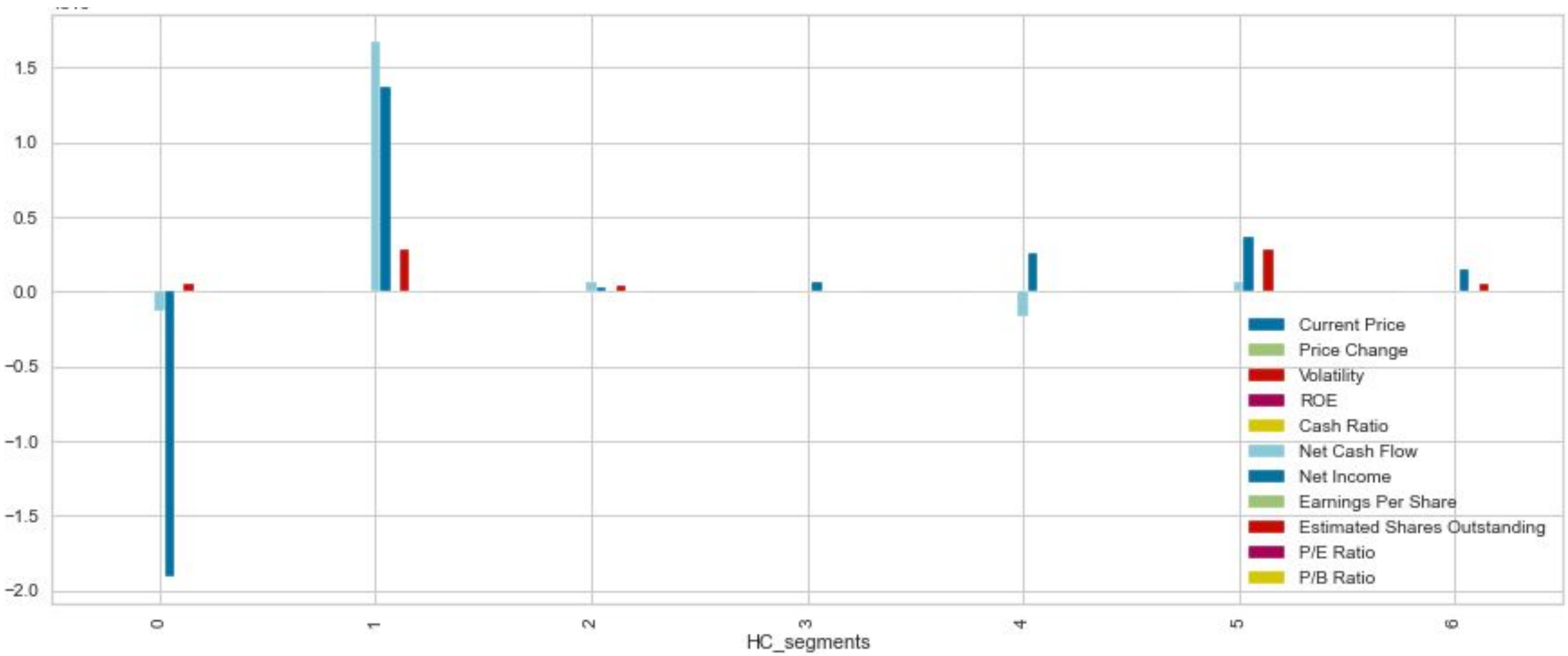
Visualizing Numerical Variables of HC_Segments



Visualizing Numerical Variables of HC_Segments



Visualizing Numerical Variables of HC_Segments



K-MEANS VS. HIERARCHICAL CLUSTERING

A Comparison of K-Means and Hierarchical Clustering

- K-Means took less time to execute, with computation time within a few seconds compared to hierarchical clustering, which took around a minute or two to compute.

- In this particular case, K-means produced more distinct clusters than hierarchical clustering. This is evident when comparing both segments by their GICS sector.

- The largest clusters formed using both algorithms had 271 observations for K-means clustering and 330 observations for hierarchical clustering.

- The appropriate number of clusters chosen for K-means was k=6, while for hierarchical clustering, n_clusters=7 was chosen based on the evaluation of the dendrogram and the cluster profiles.

- The hierarchical clustering model appears to have preserved more accuracy around values within clusters compared to K-means clustering model.

THE END :)