Stock Market Clustering

Trade & Ahead – Unsupervised Learning

4/17/2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- K-Means Clustering
- Hierarchical Clustering
- Appendix

Executive Summary

Actionable Insights

- Use clusters to diversify, spreading investments across varied stock characteristics.
- Identify sector-focused clusters for targeted strategies.
- Opt for KMeans for balanced sector representation in portfolios.
- Apply diverse clusters to reduce sector-specific risks.

Recommendations:

- Analyze clusters for deeper sector insights.
- Personalize strategies based on client risk profiles and cluster data.
- Monitor and adjust portfolios with market changes.
- Use sector rotation based on cluster trends.
- Diversify within preferred sectors using cluster analysis.
- Educate clients on diversification benefits and economic impacts on sectors.

Business Problem Overview and Solution Approach

Objective

- Analyze and group NYSE-listed stocks using financial indicators to enhance investment strategies.

Challenges

- Complex data processing and feature selection.
- Need for meaningful and actionable stock clusters.
- Clear communication of insights to stakeholders.

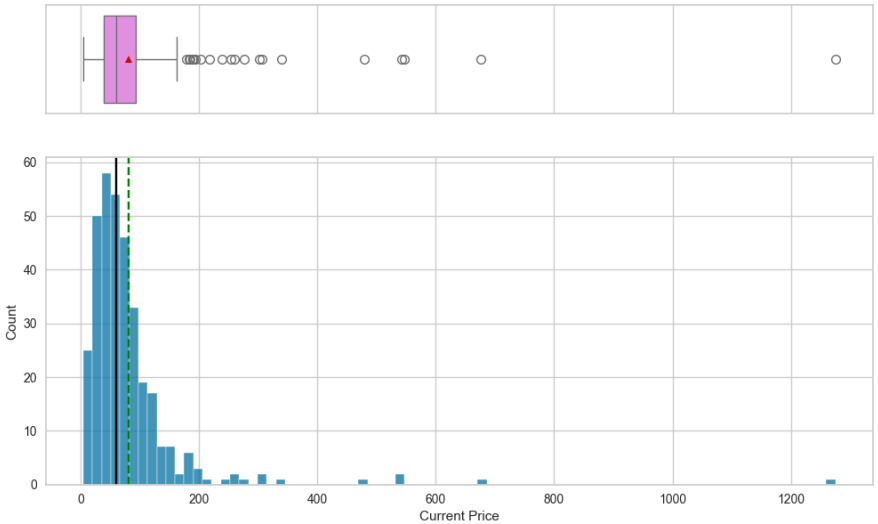
Solution Approach

1. Data Preprocessing: Clean and normalize data for comparability.
2. Feature Selection: Use PCA and correlation analysis to identify key indicators.
3. Cluster Analysis: Apply K-means or Hierarchical clustering, determine optimal clusters.
4. Evaluation: Assess cluster homogeneity and separation.
5. Reporting: Describe cluster characteristics, provide strategic recommendations, and use visual aids for clarity.

EDA Results

Univariate Analysis: Current Price

- Most stock prices are under $200, peaking just below $100.
- Prices show a right-skewed distribution with fewer high-priced stocks.
- The median price is lower than the mean, due to the skewness.
- Several outliers indicate stocks with much higher prices.
- A narrow IQR shows that the middle 50% of prices are tightly grouped.

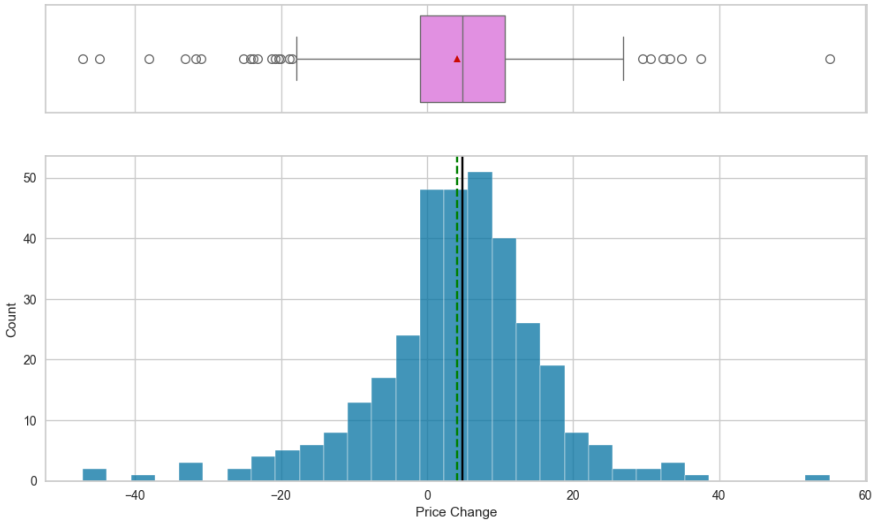


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: Price Change

- Price change is symmetrically distributed around a central value close to 0.
- Median change is around 0, suggesting no strong upward or downward trend.
- Outliers on both ends indicate some significant price drops and gains.
- Distribution is bell-shaped, resembling a normal distribution.
- Majority of changes are within a narrow range, showing modest fluctuations.

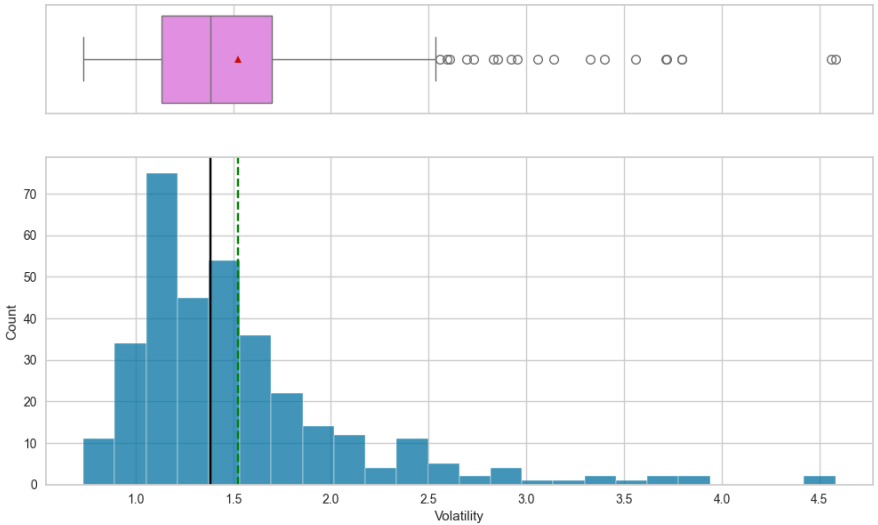


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: Volatility

- Volatility is concentrated around 1.5, with most data falling below 2.0.
- Distribution is right-skewed, fewer stocks with high volatility.
- Median is less than the mean, skewed by high-volatility outliers.
- Numerous outliers suggest some stocks are much more volatile.
- Most stocks have moderate volatility, implying steadier price movements.

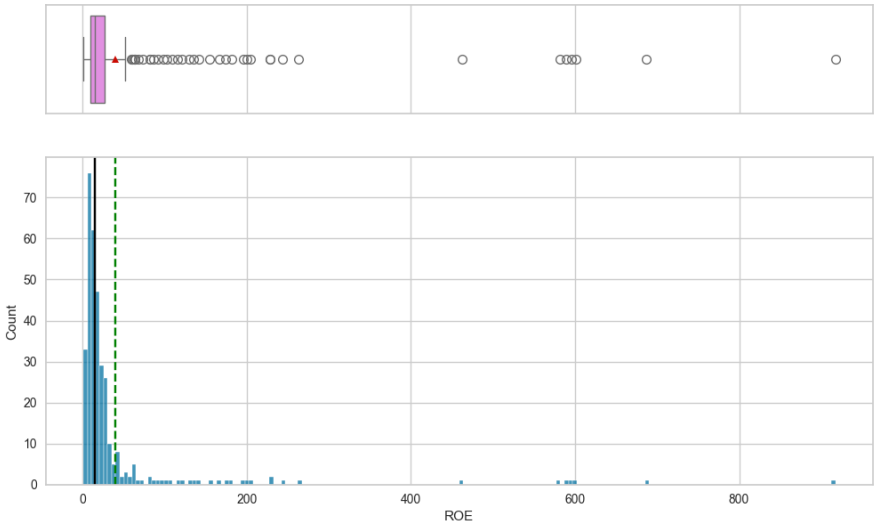


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: ROE

- ROE (Return on Equity) values are mostly clustered close to 0.
- The distribution is highly right-skewed, indicating few firms with very high ROE.
- The median ROE is near 0, suggesting that more than half of the firms have low or no earnings.
- There are numerous outliers, showing some firms with extremely high ROE.
- The data suggests that high ROE is rare, with most firms achieving much lower returns.

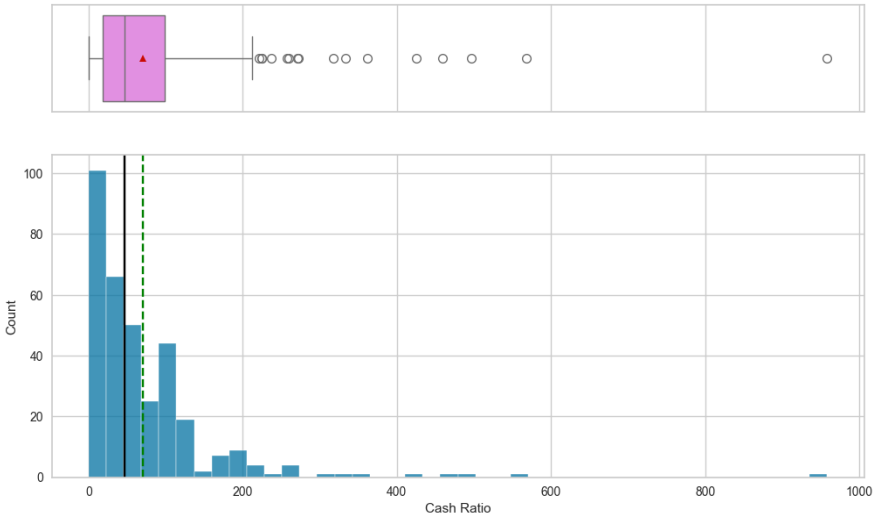


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: Cash Ratio

- Cash Ratio is predominantly low; most firms have a ratio under 50.
- Distribution is right-skewed with a few firms having very high cash ratios.
- Median is close to zero, hinting at generally low liquidity across firms.
- Outliers indicate some firms have unusually high liquidity.
- Data points to a typical low cash-on-hand situation in most firms.

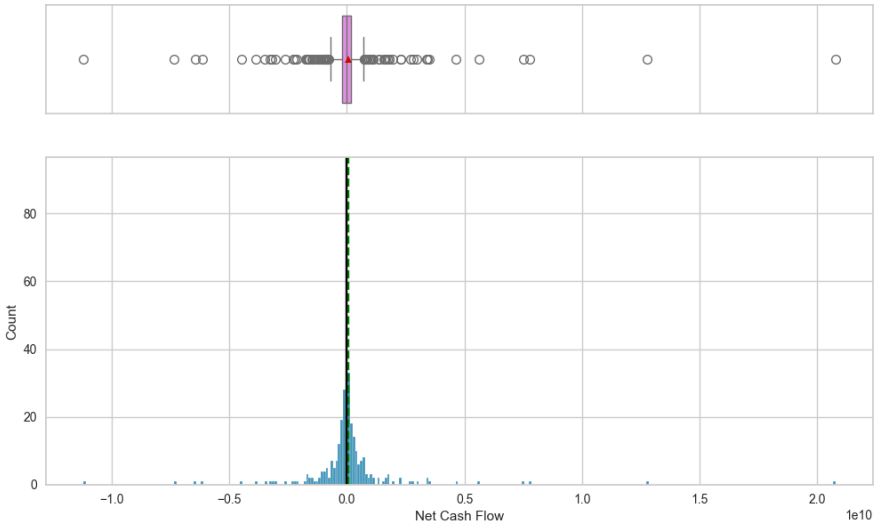


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: Net Cash Flow

- Net Cash Flow is concentrated around 0, indicating balanced cash flow for most firms.
- The distribution is slightly right-skewed, suggesting some firms have higher positive cash flow.
- Median is at 0, consistent with many firms having neutral cash flow.
- A few outliers show extreme positive net cash flow.
- Majority of firms have minimal cash flow variation, indicating stability.

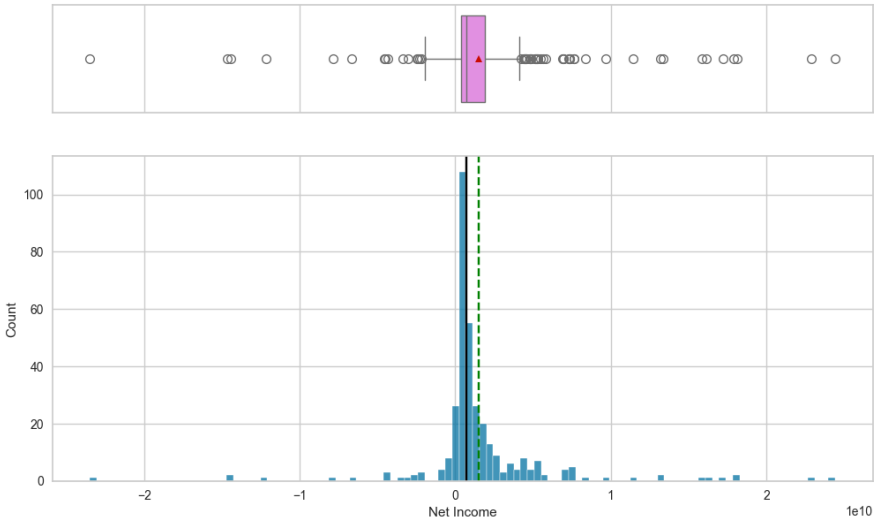


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: Net Income

- Net Income is mostly around 0, with a balance between losses and gains.
- Slight right-skew suggests a few firms with higher profits.
- Median near 0, indicating that half the firms are not significantly profitable.
- Outliers present on both sides, showing some firms with high losses or profits.
- Majority of firms have net incomes close to breakeven.

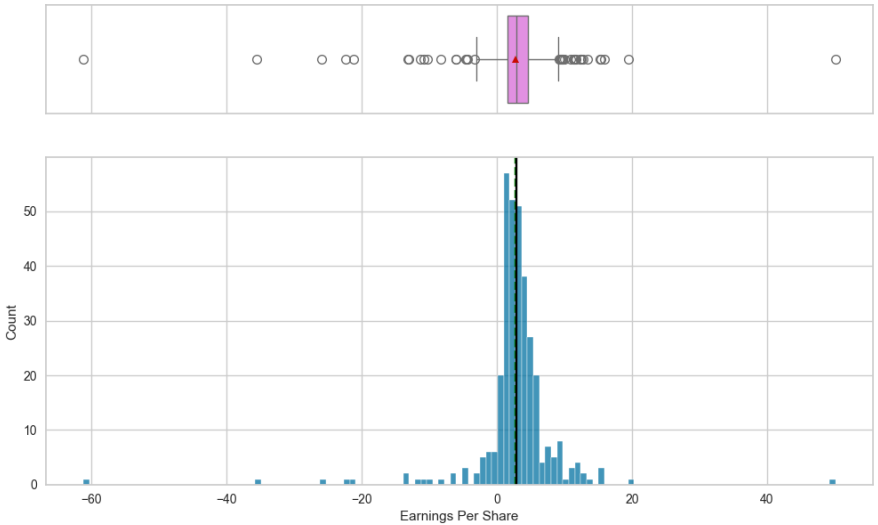


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: Earnings Per Share

- EPS mostly clusters near 0, with many firms showing little to no earnings.
- Distribution is slightly right-skewed with some firms having higher EPS.
- Median EPS is close to 0, indicating typical firm earnings are low.
- Notable outliers on both sides indicate some firms with significantly high or low EPS.
- Majority of firms have EPS within a narrow range, suggesting uniform earnings performance.

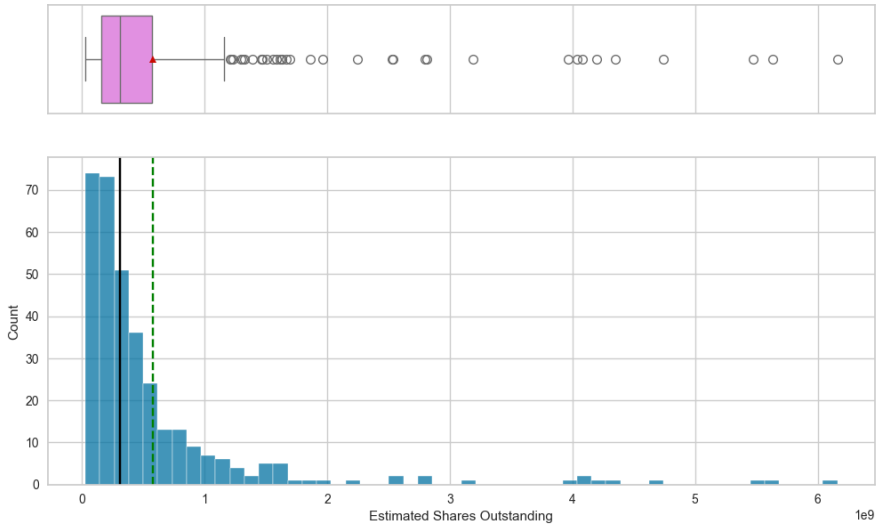


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: Estimated Shares Outstanding

- Shares outstanding are largely under 1 billion, with a concentration near the lower end.
- The distribution is right-skewed, indicating a few companies with a very large number of shares.
- Median is less than 1 billion, suggesting that over half of the companies have fewer shares.
- Several outliers are present, indicating companies with extremely high shares outstanding.
- The bulk of companies have a relatively small number of shares outstanding, implying smaller equity bases.

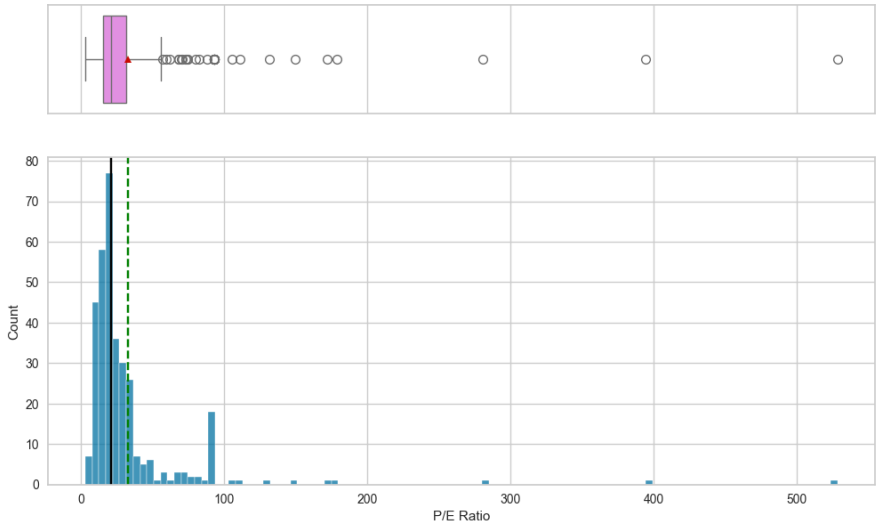


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: P/E Ratio

- P/E Ratios mostly fall below 50, indicating a concentration of lower-valued stocks.
- The distribution shows a long right tail, suggesting some stocks with very high P/E ratios.
- Median P/E Ratio is low, which may suggest stocks are priced lower relative to earnings.
- There are outliers indicating some stocks with extremely high P/E Ratios.
- The spread of P/E Ratios suggests varied investor expectations and valuations.

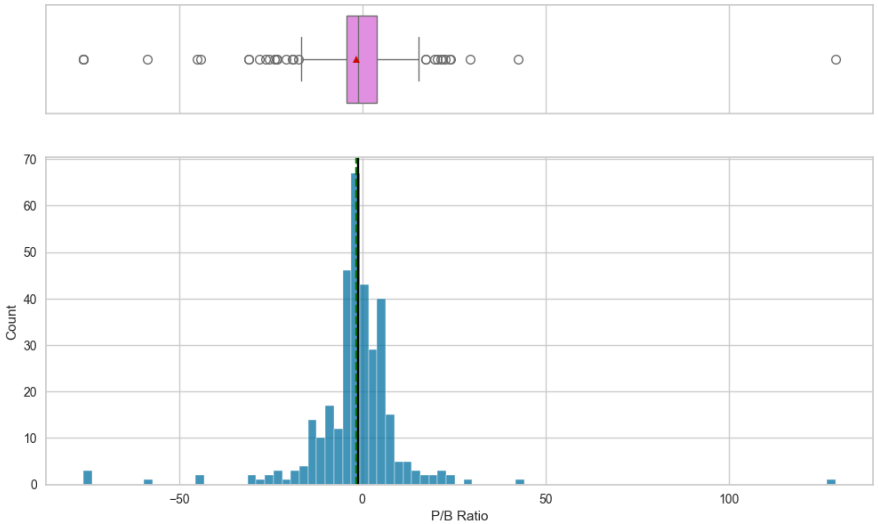


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: P/B Ratio

- P/B (Price to Book) Ratios are centered near 1, indicating fair valuation of most stocks.
- The distribution is right-skewed, with few stocks having very high P/B ratios.
- Median is close to 1, suggesting typical stock is priced around its book value.
- Outliers show some stocks with very high P/B ratios, far from the average.
- Most stocks have P/B ratios suggesting they are not overvalued.

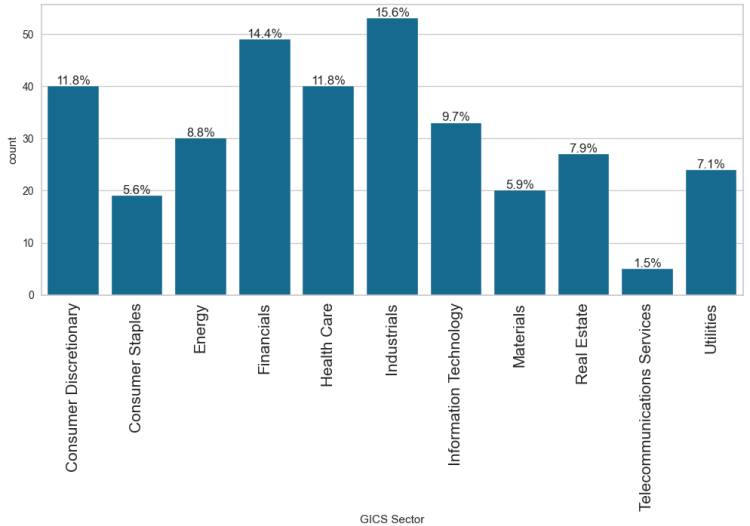


Link to Appendix slide on data background check

EDA Results

Univariate Analysis: GICS Sector

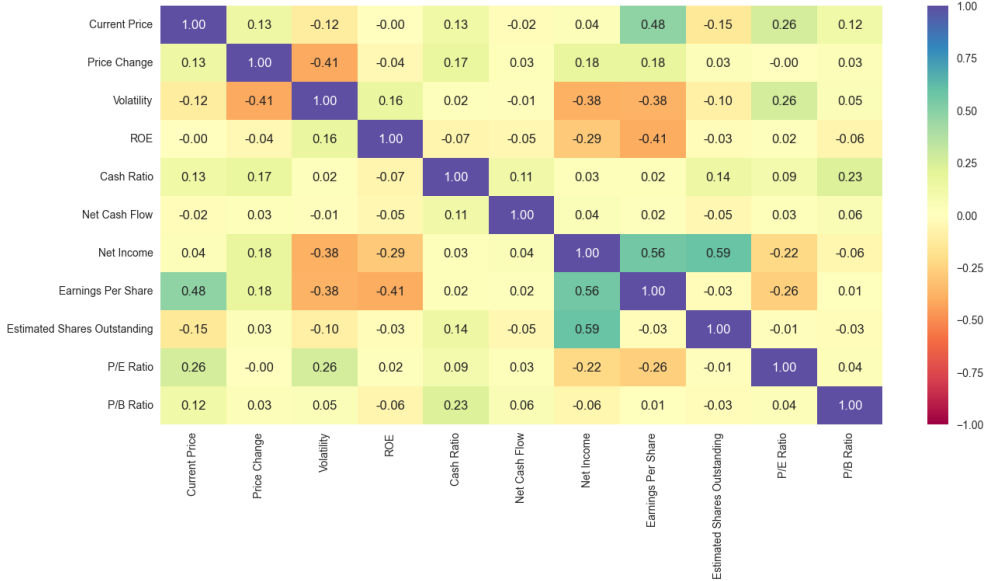
- Industrials and Financials are the most represented sectors, at 15.6% and 14.4% respectively.
- Consumer Staples (5.6%) and Telecommunications (1.5%) have the smallest representations.
- The distribution across sectors is uneven, suggesting sector-specific investment opportunities.



Link to Appendix slide on data background check

EDA Results

Bivariate Analysis: Correlation

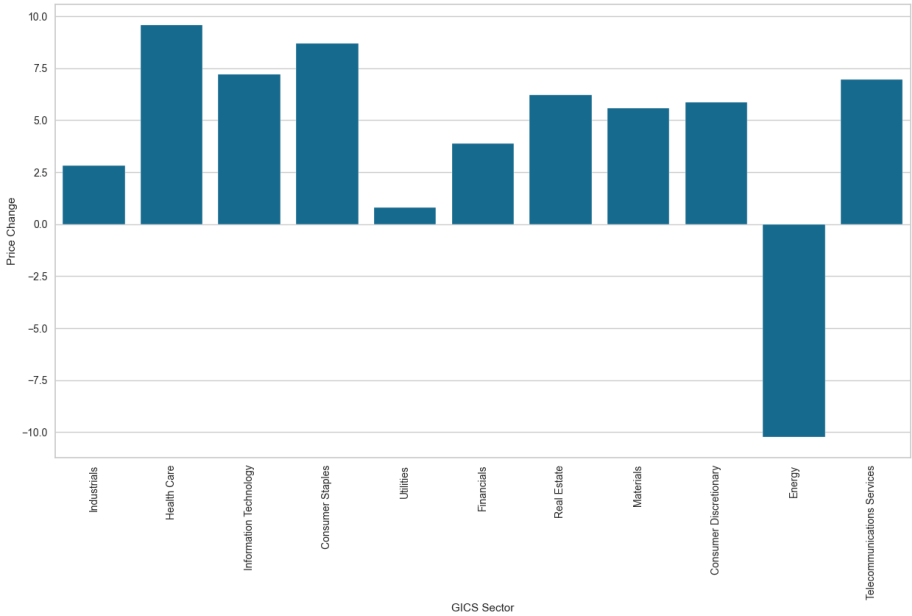


Link to Appendix slide on data background check

EDA Results

Bivariate Analysis: Economic Sector vs. Price Change

- Energy sector displays a significant negative price change, indicating a decrease in stock prices.
- Health care and Consumer Staples has the highest positive price change
- The variation across sectors may indicate different market conditions affecting each sector.

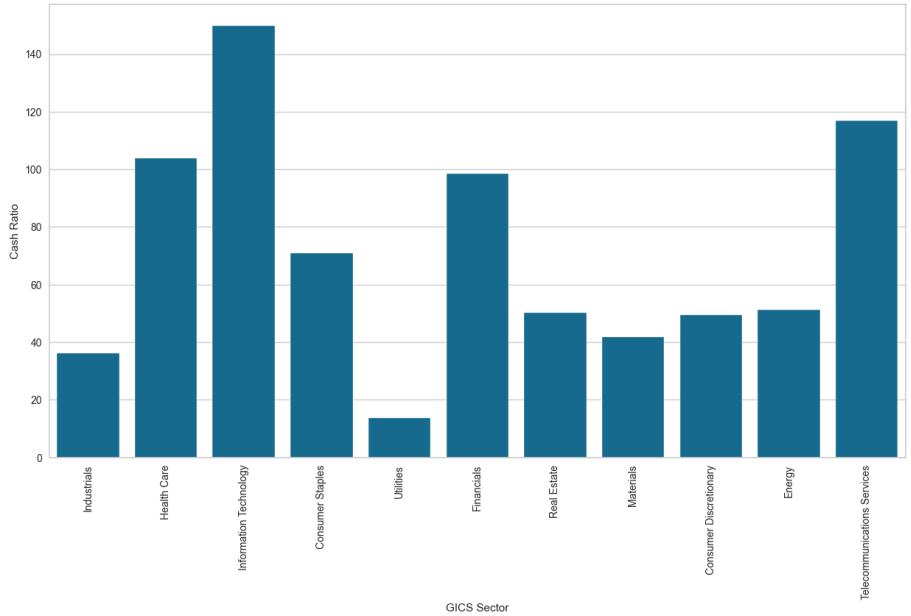


Link to Appendix slide on data background check

EDA Results

Bivariate Analysis: Economic Sector vs. Cash Ratio

- Information Technology has the highest cash ratio indicating high liquidity
- Utilities has the lowest cash ratio
- The chart shows significant variation in liquidity across different sectors.

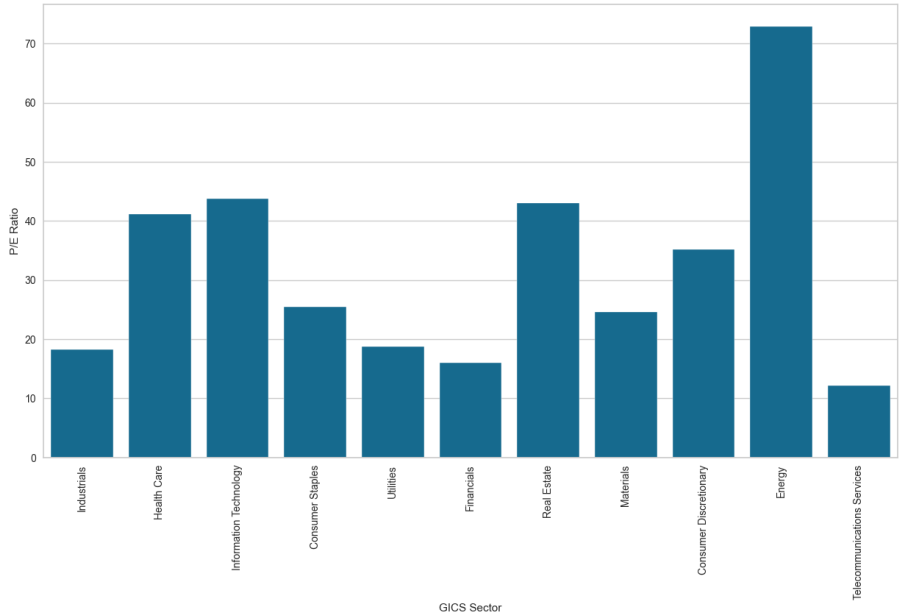


Link to Appendix slide on data background check

EDA Results

Bivariate Analysis: Economic Sector vs. P/E Ratio

- Energy has the highest P/E Ratio, which may suggest higher growth expectations or overvaluation.
- Industrials have the lowest P/E Ratio, which could imply undervaluation or lower growth expectations.
- The variation in P/E Ratios can inform investors about market sentiment and valuation levels in each sector.

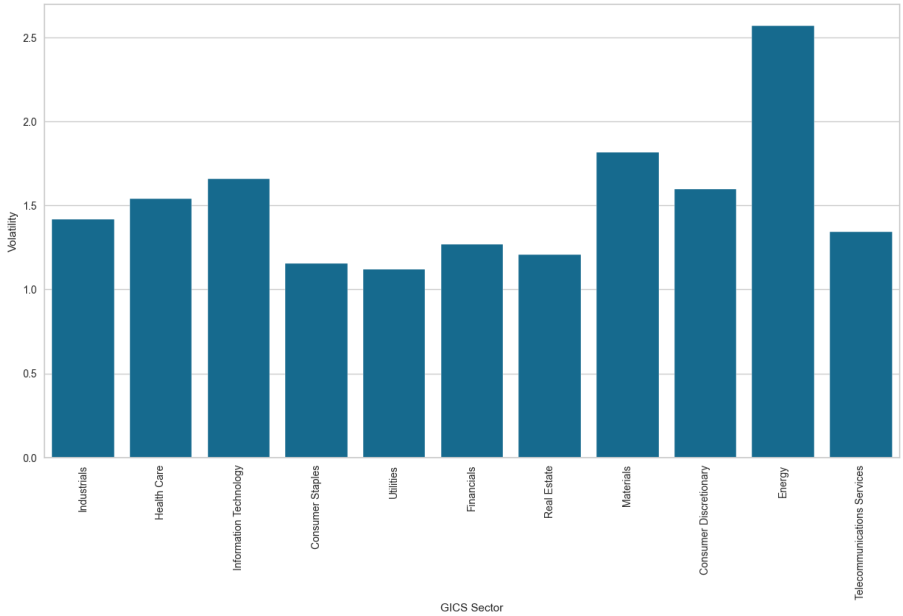


Link to Appendix slide on data background check

EDA Results

Bivariate Analysis: Economic Sector vs. Price Change

- Energy shows the highest volatility, indicating greater degree of price fluctuation
- The rest of the sectors are relatively close in volatility
- The chart illustrates varying risk profiles across sectors based on volatility, which can guide investment decisions according to risk tolerance.



Link to Appendix slide on data background check

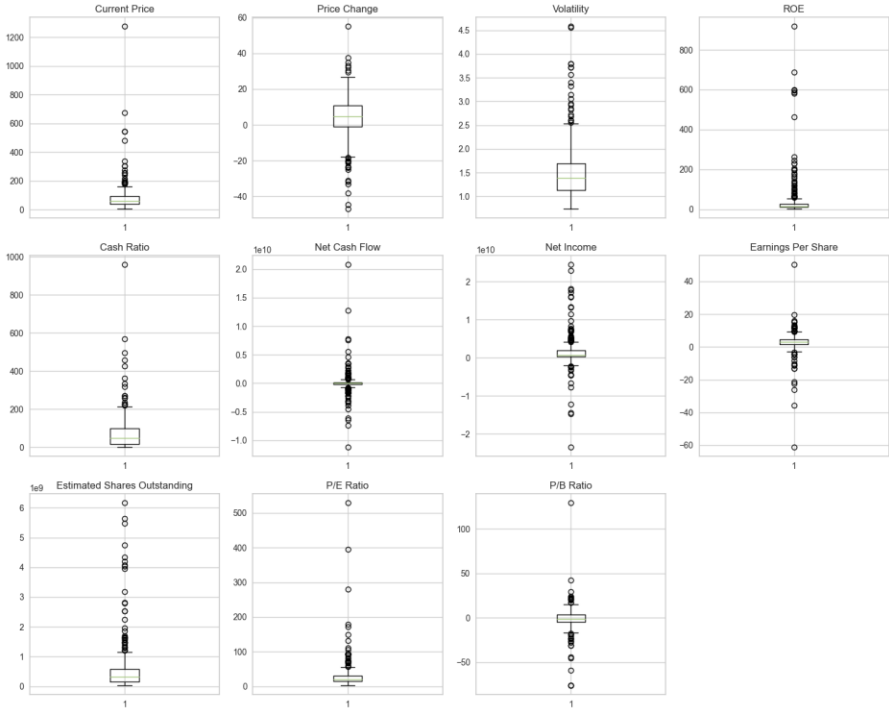
Data Preprocessing

- Duplicate value check
  - There are no duplicate values
- Missing value treatment
  - There are no missing values

Data Preprocessing

Outlier Check

- There are outliers present for every column therefore, scaling will be necessary for each one.
- StandardScaler will be used to scale the data to prevent undue influence on the model



K-Means Clustering Summary

- Optimal Number of clusters using K-Means
  - The optimal number of clusters was found using the silhouette and elbow method
  - After using a silhouette plot, the optimal number of clusters to be found was 4
- Cluster Profiling

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	72.399112	5.066225	1.388319	34.620939	53.000000	-14046223.826715	1482212389.891697	3.621029	438533835.667184	23.843656	-3.358948	277
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	38.099260	-15.370329	2.910500	107.074074	50.037037	-159428481.481481	-3887457740.740741	-9.473704	480398572.845926	90.619220	1.342067	27
3	234.170932	13.400685	1.729989	25.600000	277.640000	1554926560.000000	1572611680.000000	6.045200	578316318.948800	74.960824	14.402452	25

Link to Appendix slide on K-Means Clustering

Hierarchical Clustering Summary

- Optimal Number of clusters using Hierarchical Clustering
  - Highest cophenetic correlation was achieved with Euclidean distance and average linkage
  - 4 clusters were used
- Cluster Profiling

HC_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	77.573266	4.148438	1.515708	35.184524	67.154762	67104693.452381	1607391086.309524	2.905640	572317821.413095	32.325679	-1.762402	336
1	1274.949951	3.190527	1.268340	29.000000	184.000000	-1671386000.000000	2551360000.000000	50.090000	50935516.070000	25.453183	-1.052429	1
2	24.485001	-13.351992	3.482611	802.000000	51.000000	-1292500000.000000	-19106500000.000000	-41.815000	519573983.250000	60.748608	1.565141	2
3	104.660004	16.224320	1.320606	8.000000	958.000000	592000000.000000	3669000000.000000	1.310000	2800763359.000000	79.893133	5.884467	1

Link to Appendix slide on Hierarchical Clustering

APPENDIX

Data Background and Contents

Market Investment Benefits:

- Fights inflation
- Wealth creation
- Tax benefits
- Compounded long-term returns
- Early investments bolster retirement funds

Investment Strategy:

- Diversified portfolio for maximized earnings
- Balance between high returns and risk mitigation
- Clustering analysis to identify similar stocks and minimize correlation

Trade & Ahead Project:

- Cluster analysis of NYSE-listed company stocks
- Group stocks based on financial indicators
- Provide investment insights for personalized strategies

Data Overview:

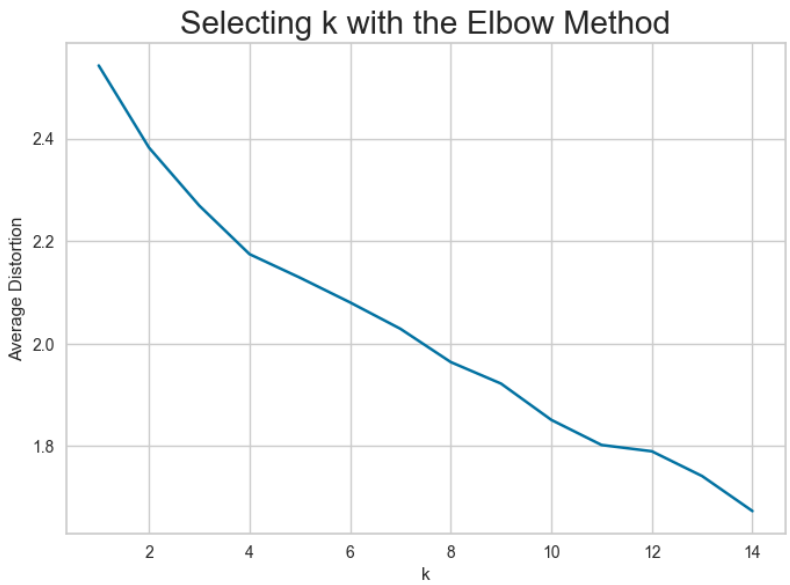
- Ticker Symbol: Unique stock identifier
- Company: Name of the listed entity
- GICS Sector/Sub Industry: Economic category of business
- Current Price: Stock value in USD
- Price Change: 13-week percentage variation
- Volatility: 13-week price stability measure
- ROE: Return on Equity ratio
- Cash Ratio: Liquidity measurement
- Net Cash Flow: Financial health indicator
- Net Income: Profitability after expenses
- Earnings Per Share: Profit allocation per stock unit
- Estimated Shares Outstanding: Shares available to public
- P/E Ratio: Price to Earnings comparison
- P/B Ratio: Price to Book value analysis

K – Means Clustering Technique

- Application of K – Means
  - Efficiency for Large Datasets: KMeans is computationally faster and more efficient for large datasets, making it suitable for stock market data with many companies.
  - Ease of Interpretation: The clusters produced by KMeans are easy to interpret, aiding in clear categorization of stocks based on their attributes.
  - Centroid-Based Clustering: KMeans provides clear centroids for clusters that can be used as the profile of the typical stock in each cluster, useful for quick comparisons.
  - Adaptability to Feature Scaling: KMeans responds well to feature scaling, ensuring that financial indicators with varying scales contribute equally to the analysis.

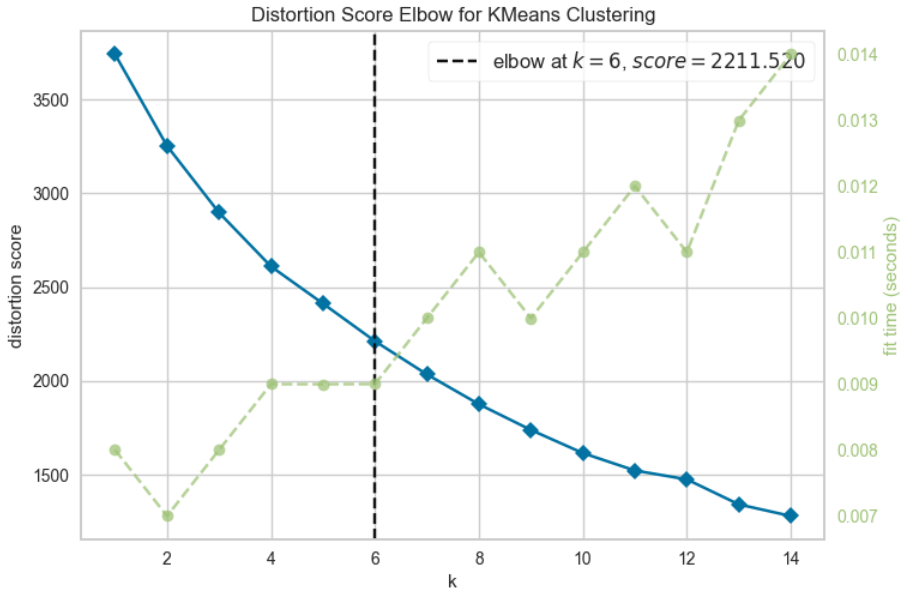
K-Means Clustering Technique

- Elbow Curve
  - Slight bend at K = 4
  - There is no obvious elbow, more information needed to determine optimal number of clusters



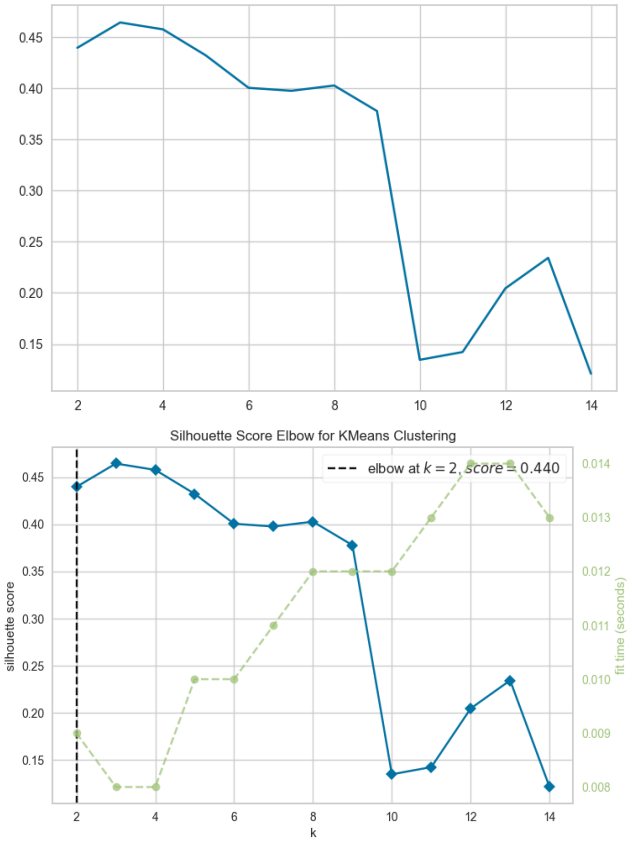
K-Means Clustering Technique

- Distortion Score
  - A clearer distinction can be made that the optimal number of clusters is likely at k=6.



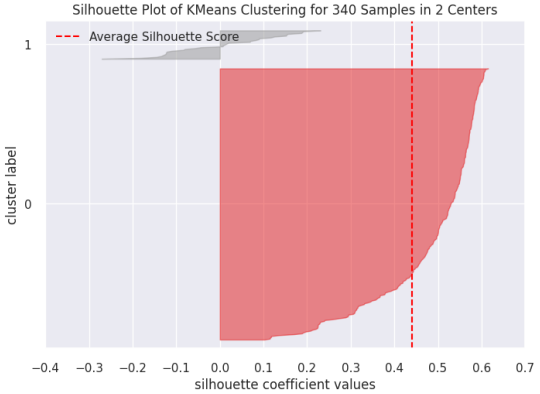
K-Means Clustering Technique

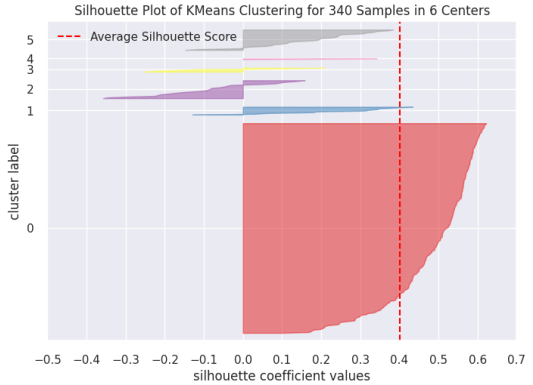
- Silhouette Scores
  - High scores for 2 clusters suggest good cluster definition.
  - Scores decline from 2 to 6 clusters, indicating decreasing cluster quality.
  - A significant dip at 6 clusters may suggest poor clustering.
  - Slight improvement after 6 clusters, but not as high as initial scores.
  - Fluctuating scores with more clusters, but no return to the high initial values.



K-Means Clustering Technique

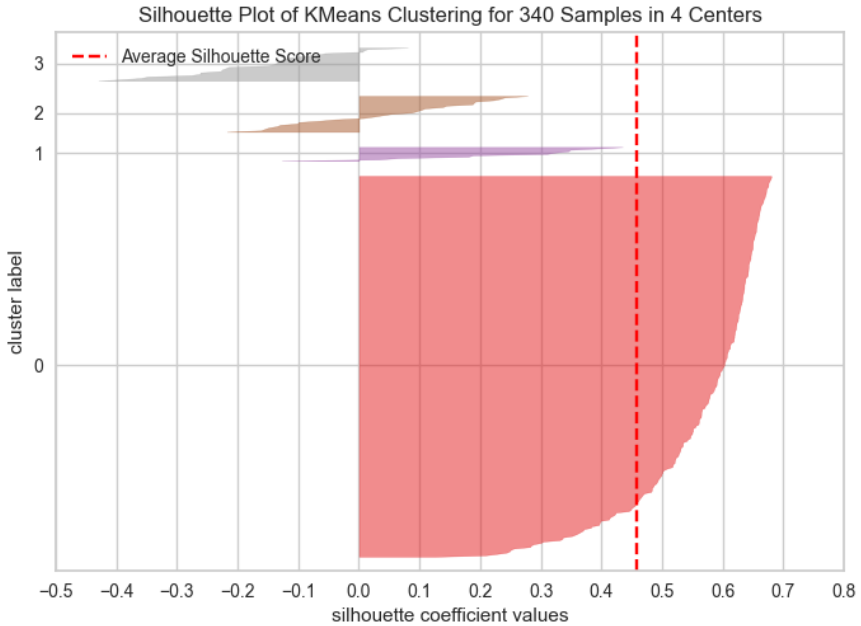
- Silhouette Plots
  - 2 centers show good separation and cohesion but indicates room for improvement. Clusters may not be highly distinct which can cause overlap
  - 6 centers shows less-than-ideal clustering which can be seen by a lower average score. Centers with negative scores indicate low distinction or maybe even incorrect clustering.





K-Means Clustering Technique

- Silhouette Plot
  - Plot with 4 centers shows more balanced cluster sizes
  - Each cluster mainly has positive values with wide plots, indicating good definition
  - Moderate average score
  - 4 centers appears to be the optimal number



Hierarchical Clustering Technique

- Application of Hierarchical Clustering
  - Intuitive Dendrogram Representation: Hierarchical clustering produces a dendrogram that allows for a visual representation of the clusters, offering an intuitive understanding of the data structure.
  - No Need to Specify Cluster Number: It does not require pre-specification of the number of clusters, which can be advantageous when the ideal number of clusters is unknown.
  - Flexibility in Cluster Formation: Hierarchical clustering can produce a diverse set of cluster shapes and sizes, which can be more representative of the underlying market structures.
  - Detailed Cluster Analysis: It allows for a detailed cluster analysis at different levels of data granularity, which can be helpful for investors who require different levels of investment detail.

Hierarchical Clustering Technique

Cophenetic Correlation for different distance metrics and linkages

```
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 single linkage is 0.9232271494002922.
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 single linkage is 0.9062538164750717.
Cophenetic correlation for Cosine distance and complete linkage is 0.14558585821013348.
Cophenetic correlation for Cosine distance and average linkage is 0.27565476401031014.
Cophenetic correlation for Cosine distance and single linkage is 0.16336378418310402.
********************************************************************************************
Highest cophenetic correlation is 0.9422540609560814, which is obtained with Euclidean distance and average linkage.
```

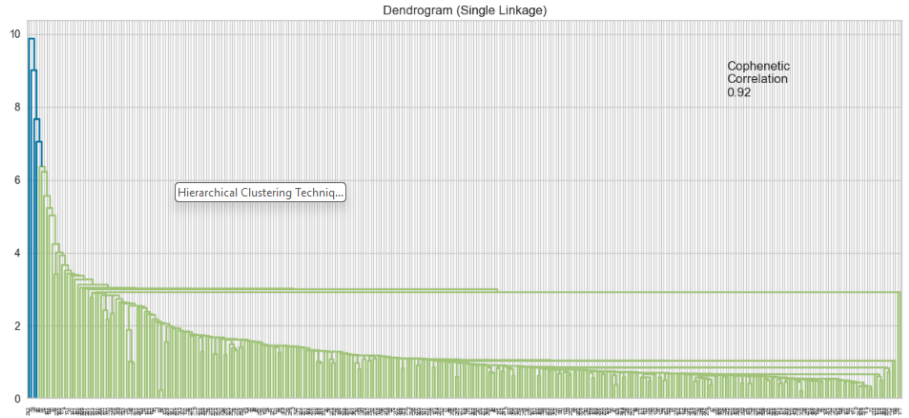
Hierarchical Clustering Technique

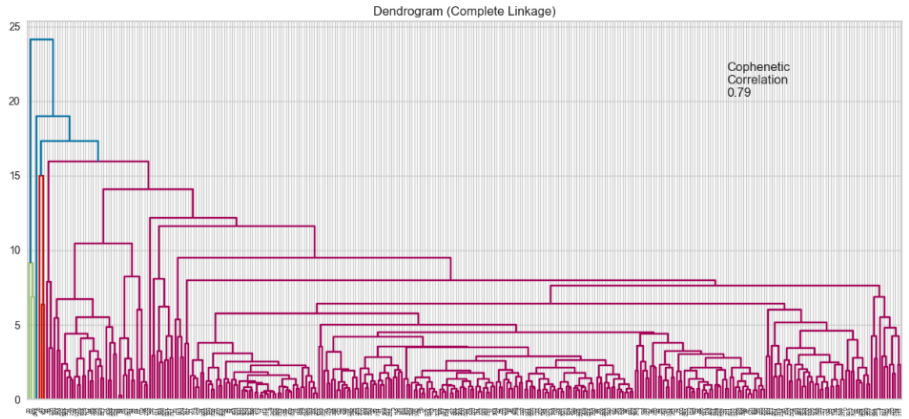
- Cophenetic correlation for different linkages using only Euclidean distance

```
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 ward linkage is 0.7101180299865353.
**********************************************************************************************************
Highest cophenetic correlation is 0.9422540609560814, which is obtained with average linkage.
```

Hierarchical Clustering Technique

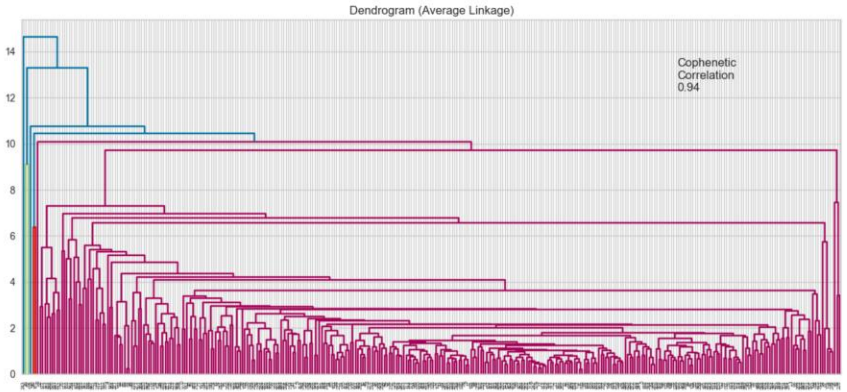
- Dendrogram (Single Linkage)
  - Shows a chaining effect with a gradual merging of points into clusters.
  - High cophenetic correlation (0.92) suggests accurate distance preservation.
  - Many small clusters merging at low distances, indicating high point similarity.
- Dendrogram (Complete Linkage)
  - Shows a chaining effect with a gradual merging of points into clusters.
  - High cophenetic correlation (0.92) suggests accurate distance preservation.
  - Many small clusters merging at low distances, indicating high point similarity.

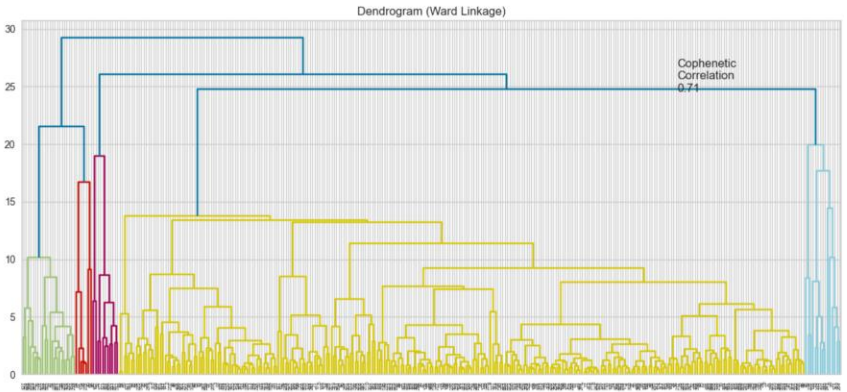




Hierarchical Clustering Technique

- Dendrogram (Average Linkage)
  - Balanced cluster formation with intermediate distances between merges.
  - High cophenetic correlation (0.94) indicating reliable representation of distances.
  - Lack of a chaining effect suggests moderate similarity and diversity within clusters.
- Dendrogram (Ward Linkage)
  - Large-scale structure with clear, distinct clusters suggested by long vertical lines.
  - Cophenetic correlation of 0.74 indicates moderate preservation of data point distances.
  - Substantial differences in merging levels suggest clusters with significant variance.





K-Means vs Hierarchical Clustering

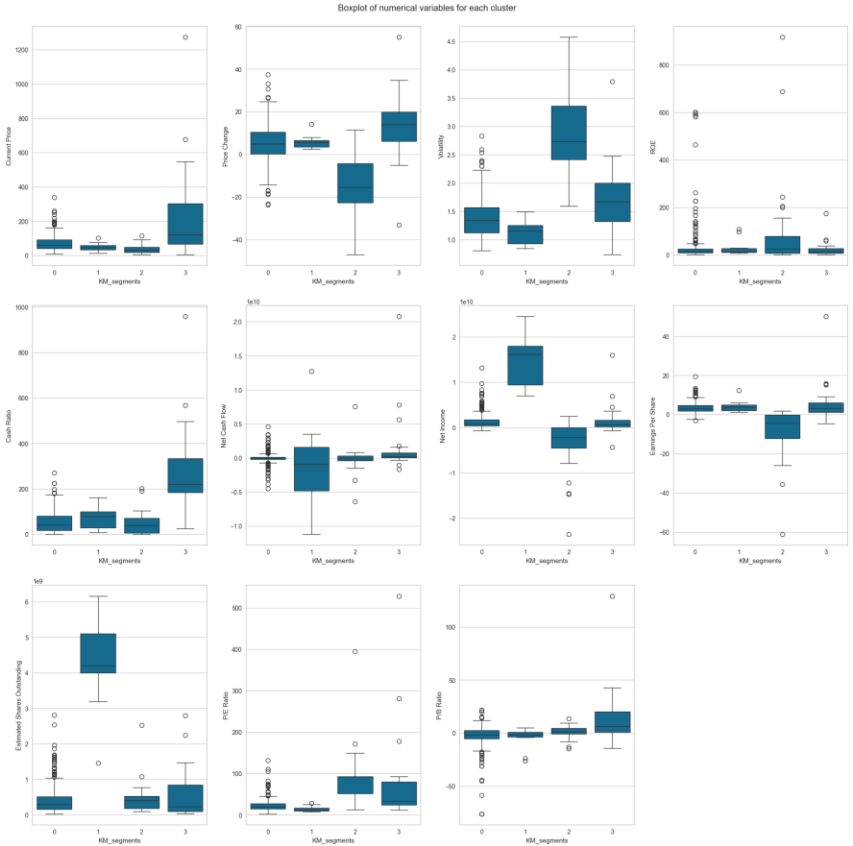
- Both methods did not take long to execute, however Hierarchical seemed easier to execute
- According to count, K-Means offers a balanced distribution across the different sectors while Hierarchical can be considered more distinct by creating highly specialized clusters
- 4 Clusters were used for both clustering algorithms

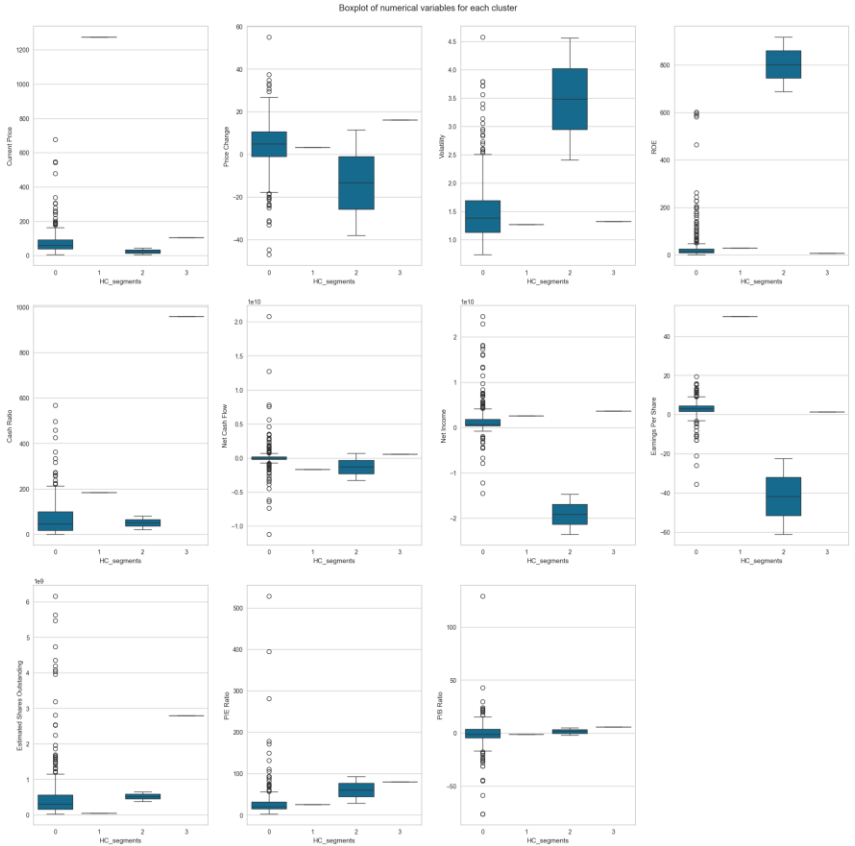
KM_segments	GICS Sector	
0	Consumer Discretionary	33
	Consumer Staples	17
	Energy	6
	Financials	45
	Health Care	29
	Industrials	52
	Information Technology	24
	Materials	19
	Real Estate	26
	Telecommunications Services	2
	Utilities	24
1	Consumer Discretionary	1
	Consumer Staples	1
	Energy	1
	Financials	3
	Health Care	2
	Information Technology	1
	Telecommunications Services	2
2	Energy	22
	Industrials	1
	Information Technology	3
	Materials	1
3	Consumer Discretionary	6
	Consumer Staples	1
	Energy	1
	Financials	1
	Health Care	9
	Information Technology	5
	Real Estate	1
	Telecommunications Services	1

HC_segments	GICS Sector	
0	Consumer Discretionary	39
	Consumer Staples	19
	Energy	28
	Financials	49
	Health Care	40
	Industrials	53
	Information Technology	32
	Materials	20
	Real Estate	27
	Telecommunications Services	5
	Utilities	24
1	Consumer Discretionary	1
2	Energy	2
3	Information Technology	1

Name: Security, dtype: int64

K-Means vs Hierarchical Clustering





Great Learning

Happy Learning !