Great Learning
POWER AHEAD

PROJECT | UNSUPERVISED LEARNING

DATA-DRIVEN STRATEGIES FOR Trade&Ahead

SHAISHAV MERCHANT

AGENDA

Executive Summary	03
Business Problem & Solution Approach	05
Data Overview	08
Exploratory Data Analysis (Summary)	10
Data Preprocessing	11
K-Means Cluster Summary	13
Hierarchical Cluster Summary	15
Insights & Recommendations	17
Appendix	20

Trade&Ahead

EXECUTIVE SUMMARY

This executive summary encapsulates our analysis and strategic insights from clustering NYSE-listed companies using unsupervised learning. Our objective is to refine investment strategies, enhance portfolio diversification, and optimize performance through detailed financial analysis and data-driven segmentation.

EXECUTIVE SUMMARY

- **Strategic Objective:** Engaged by Trade&Ahead to leverage unsupervised learning for clustering NYSE stocks, enhancing portfolio diversification and optimizing performance.
- **Data Robustness:** Examined a comprehensive dataset of 340 NYSE companies, no missing values, ensuring a thorough financial analysis.
- **EDA Insights:** Univariate and Bivariate analysis revealed significant financial variability and correlations, crucial for investment strategies.
- **Clustering Methodology:** Determined optimal clusters (K=6) using K-Means based on Elbow Method and Silhouette Scores, with Hierarchical Clustering offering depth in structural data analysis.
- **K-Means Cluster Dynamics:** K-Means clustering distinguished 6 distinct segments, from stable Industrials and Financials to volatile Energy stocks, suggesting varied investment profiles.
- **Hierarchical Clustering Detail:** Provided granular cluster formations with 6 segments, highlighting financial traits from high liquidity to high risk.
- **Consistent Findings:** Both K-Means and Hierarchical Clustering aligned on a 6-cluster solution, indicating robust market segmentation.
- **Divergent Insights:** While K-Means offered efficiency and uniform clusters, Hierarchical Clustering excelled in revealing complex market relationships.
- **Actionable Insights:** Recommended cluster-specific investment approaches, with a focus on strategic outliers and risk characterization.
- **Forward-Looking Recommendations:** Advocate for continuous model updates, stakeholder education, and investment in technology for predictive market analytics, enhancing future decision-making processes.

BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH

Trade&Ahead has engaged us to refine investment strategies by clustering NYSE listed companies based on key financial indicators. Our approach uses unsupervised learning to identify similar yet minimally correlated stocks, aiding in diversified portfolio construction. This method enhances risk management and optimizes portfolio performance.

PROBLEM OVERVIEW

Context

The stock market offers a lucrative avenue for investment, promising substantial returns over time. However, constructing a well-diversified portfolio is complex due to the myriad financial metrics available. Effective diversification not only maximizes returns but also mitigates risks during market downturns, making the process of selecting the right stocks based on their financial performance and correlation a critical challenge for investors.

Objective

1. Review financial indicators of companies listed on the New York Stock Exchange (NYSE) to gain insights into their economic behaviours.
2. Group stocks based on similarities in their financial attributes to identify patterns and relationships.
3. Provide investors with data-driven clusters to aid in constructing diversified investment portfolios.
4. Utilize the clustering of stocks to minimize risks associated with volatility by spreading investments across various sectors and financial profiles.
5. Leverage insights from clustered data to make informed decisions that aim to enhance the performance of investment portfolios.

SOLUTION APPROACH

Our approach methodically extracts actionable insights from financial data to refine investment strategies by segmenting stocks based on their characteristics using various clustering methods:

EDA (Exploratory Data Analysis):

- We start by examining patterns, relationships, and distributions in the dataset, identifying key variables and outliers.

Data Preprocessing:

- We prepare the dataset for clustering by normalizing data, handling missing values, and inspect outliers to ensure effective analysis.

K-Means Clustering:

- Using the Elbow Method and Silhouette Scores, we determine the optimal number of clusters, evaluating within-cluster consistency and separation.

Hierarchical Clustering:

- We build dendrograms to observe the formation of clusters at various similarity levels, utilizing different distance metrics and linkage methods. We also validate cluster quality using the cophenetic correlation coefficient.

Identifying Number of Clusters and Cluster Profiling:

- This final evaluation phase confirms the optimal number of clusters to capture natural groupings within the data and proceeds with detailed profiling of each cluster, crucial for developing tailored investment strategies that are both robust and actionable.

DATA OVERVIEW

The dataset contains financial metrics for NYSE-listed companies, including stock price, price changes, volatility, ROE, cash ratio, net cash flow, net income, earnings per share, and valuation ratios like P/E and P/B. It spans various sectors, offering a broad basis for investment analysis and strategy development.

DATA OVERVIEW

COLUMN	TYPE	REMARKS
Ticker Symbol	object	Stock Identifier
Security	object	Name of the company
GICS Sector	object	Economic sector assigned to company
GICS Sub Industry	object	Sub industry group assigned to company
Current Price	float64	Current stock price in dollars
Price Change	float64	Percentage change in stock price – 13 weeks
Volatility	float64	Standard deviation of stock price – 13 weeks
ROE	int64	Measure of stock performance
Cash Ratio	int64	Ratio of company’s cash to current liabilities
Net Cash Flow	int64	Difference between cash inflows and outflows
Net Income	int64	Revenue *less* Expenses
Earnings Per Share	float64	Net profit *divided by* Outstanding shares
Estimated Shares Outstanding	float64	Stock currently held by shareholders
P/E Ratio	float64	Ratio of stock price to earnings per share
P/B Ratio	float64	Ratio of stock price to its book value

Dataset Information:

Key Insights:

- **Dataset Scope:** Includes 340 entries and 15 attributes, with no missing values, ensuring comprehensive coverage.
- **Uniqueness:** Each entry is uniquely identified by 'Ticker Symbol' and 'Security', covering a wide range of companies.
- **Sector Variety:** Features 11 unique 'GICS Sectors', dominated by 'Industrials'.
- **Statistical Diversity:** Shows significant variability in financial metrics such as 'Current Price' and 'Net Income'.
- **Financial Stability Metrics:** Attributes like 'Volatility', 'ROE', and 'P/E Ratio' highlight financial stability and performance.

Conclusion:

This robust dataset is well-suited for advanced financial analysis and clustering, ideal for developing nuanced investment strategies based on detailed company metrics.

RECORDS	COLUMNS	DATA TYPE
340	15	float64(7), int64(4) and object(4)

EXPLORATORY DATA ANALYSIS – SUMMARY

Univariate Analysis Summary:

- **Price Variability:** Current Price ranges from $4.50 to $1274.95, showing a diverse market. Price Change highlights potential for high-risk and high-return investments.
- **Risk and Liquidity:** Volatility and Cash Ratio exhibit significant variability, essential for assessing risk and liquidity which are critical in downturns.
- **Financial Health:** Net Cash Flow and Net Income show extensive ranges, key for evaluating company operations and profitability.
- **Profitability Metrics:** Earnings Per Share and ROE indicate varied profitability, affecting valuation and investment attractiveness.
- **Market Valuation Tools:** P/E and P/B Ratios vary widely, providing insights into stock overvaluation or undervaluation, crucial for investment decisions.

Bivariate Analysis Summary:

- **Profitability and Market Value:** There's a consistent positive correlation between profitability metrics (Net Income and Earnings Per Share) and stock performance, suggesting that more profitable companies typically enjoy higher stock prices and greater market capitalization.
- **Financial Stability and Volatility:** Higher Net Income and Earnings Per Share correlate with reduced market volatility, indicating that financially stable companies generally experience fewer price fluctuations, offering more predictable investment outcomes.

Feature Relation with GICS (Economic) Sector:

- **Sector Influence:** Industrials and Financials dominate; Energy struggles with declines, while Healthcare and Consumer Staples show strong growth.
- **Financial Health:** Information Technology boasts high liquidity; Utilities display lower cash ratios, suggesting financial challenges.
- **Valuation and Volatility:** High P/E in Energy implies overvaluation or growth expectations; Telecommunications shows low P/E; Energy most volatile, Utilities least.

Refer to Appendix section for detailed EDA

DATA PREPROCESSING

Data preprocessing for Trade&Ahead's dataset ensures accurate, uniform inputs, crucial for unsupervised learning to effectively detect patterns and group similar stocks, thereby aiding in the development of informed, data-driven investment strategies.

DATA PREPROCESSING

1. Duplicate Record Verification:

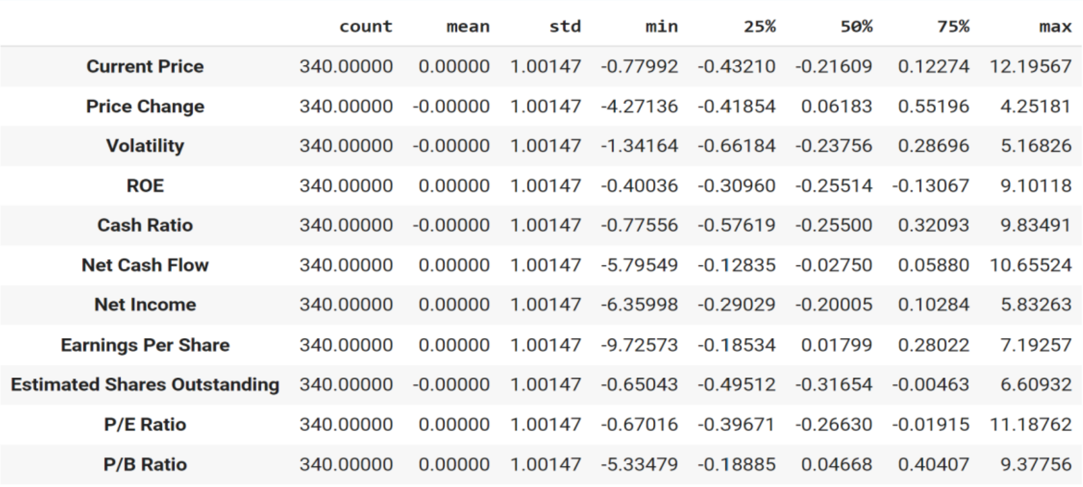
- Confirmed that no duplicate records exist, ensuring data integrity.

2. Missing Value Analysis:

- Verified absence of missing and exceptional values, confirming dataset completeness.
- Our thorough checks didn't identify any misrepresented data or typo error.

4. Feature Scaling with StandardScaler:

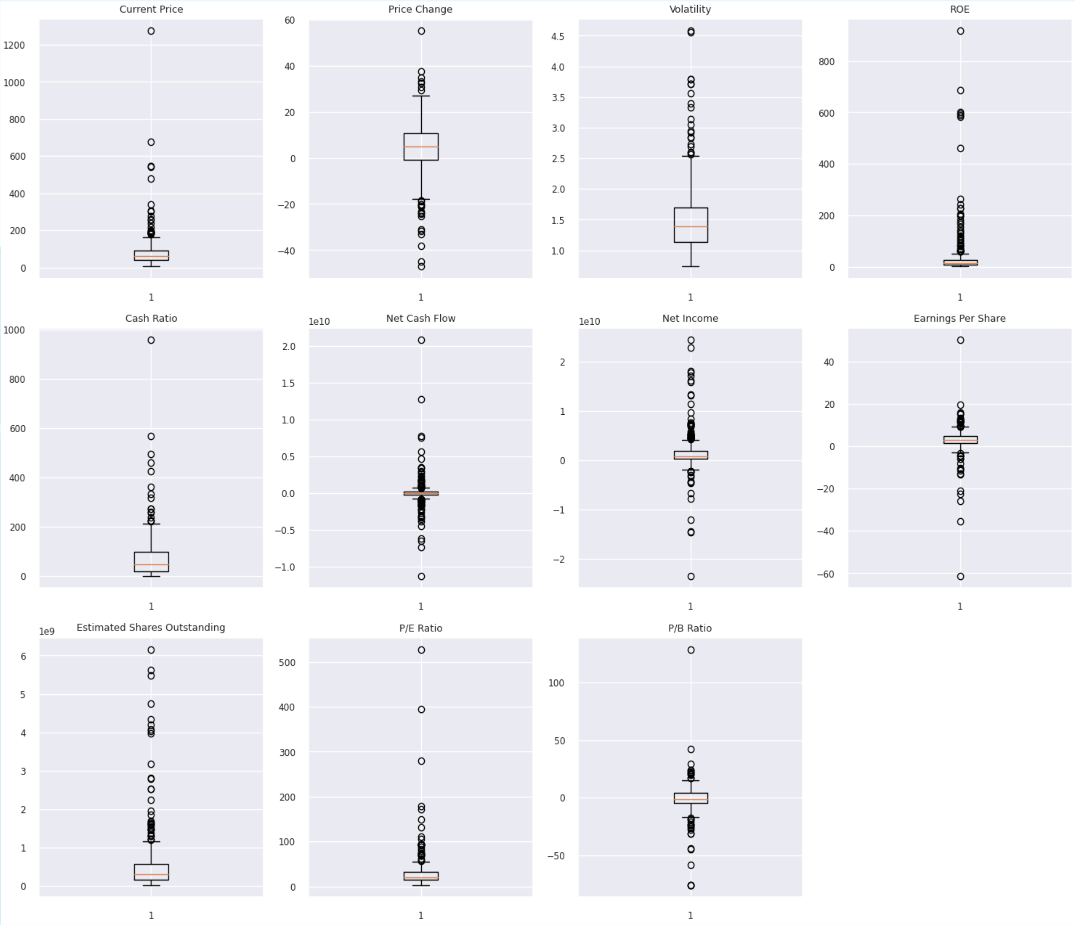
- Applied SKLearn's StandardScaler to normalize feature scales, essential for effective distance computation in clustering algorithms.



	count	mean	std	min	25%	50%	75%	max
Current Price	340.00000	0.00000	1.00147	-0.77992	-0.43210	-0.21609	0.12274	12.19567
Price Change	340.00000	-0.00000	1.00147	-4.27136	-0.41854	0.06183	0.55196	4.25181
Volatility	340.00000	-0.00000	1.00147	-1.34164	-0.66184	-0.23756	0.28696	5.16826
ROE	340.00000	0.00000	1.00147	-0.40036	-0.30960	-0.25514	-0.13067	9.10118
Cash Ratio	340.00000	-0.00000	1.00147	-0.77556	-0.57619	-0.25500	0.32093	9.83491
Net Cash Flow	340.00000	0.00000	1.00147	-5.79549	-0.12835	-0.02750	0.05880	10.65524
Net Income	340.00000	0.00000	1.00147	-6.35998	-0.29029	-0.20005	0.10284	5.83263
Earnings Per Share	340.00000	0.00000	1.00147	-9.72573	-0.18534	0.01799	0.28022	7.19257
Estimated Shares Outstanding	340.00000	-0.00000	1.00147	-0.65043	-0.49512	-0.31654	-0.00463	6.60932
P/E Ratio	340.00000	0.00000	1.00147	-0.67016	-0.39671	-0.26630	-0.01915	11.18762
P/B Ratio	340.00000	0.00000	1.00147	-5.33479	-0.18885	0.04668	0.40407	9.37756

3. Outlier Identification:

- Identified significant to moderate outliers; these will be retained as they represent valid variations.



K-MEANS CLUSTERING

Decoding Optimal Clustering: Why K=6 Stands Out for K-Means?

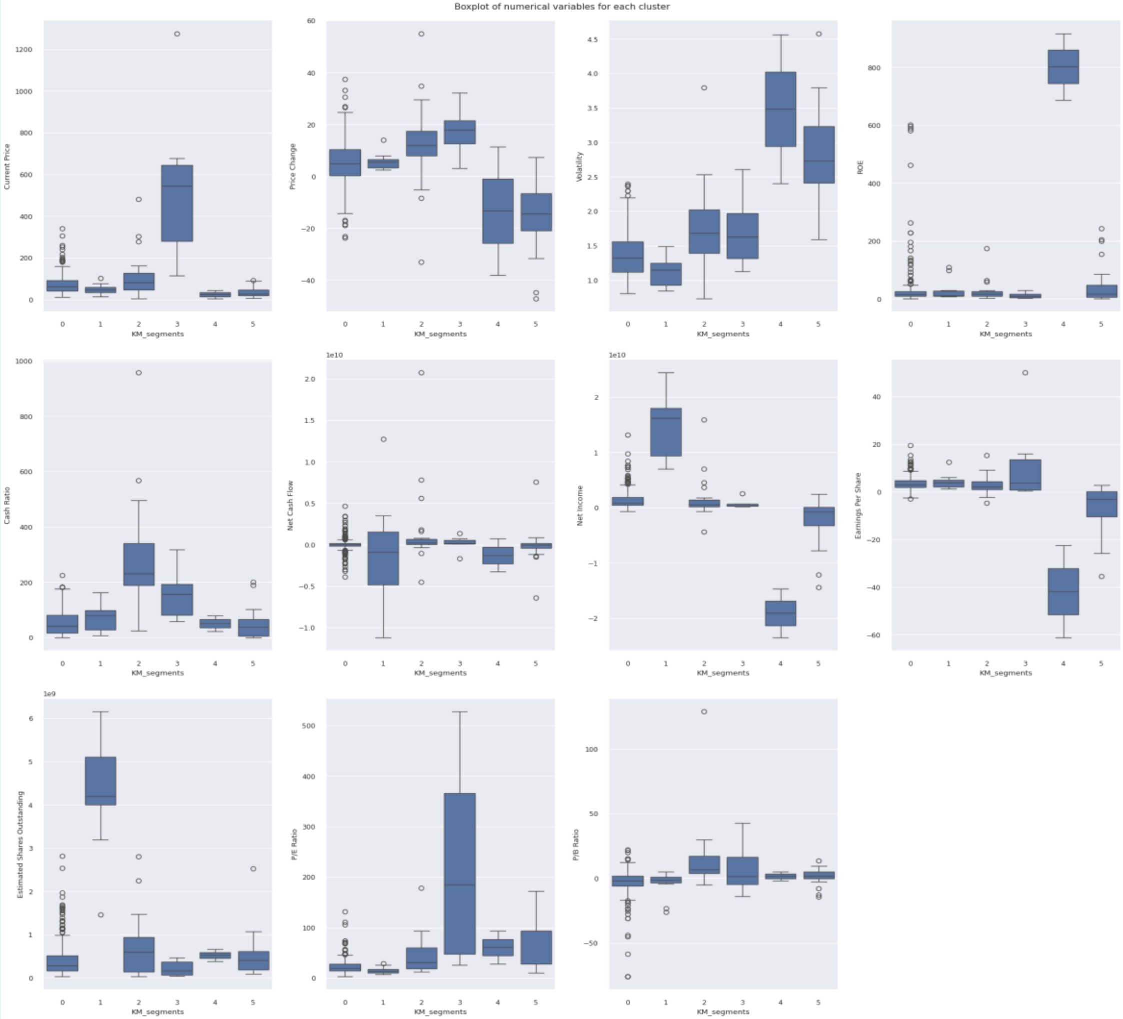
1. **Elbow Method:** Although the Elbow chart doesn’t show a clear bend at K=6, data reveals that additional clusters beyond this point yield minimal improvement in reducing distortion, suggesting K=6 as the optimal cluster number.

2. **Silhouette Scores:** While the silhouette score peaks earlier, it remains adequate at K=6 (0.400), supporting effective balance between cluster cohesion and separation.

3. **Analysis Consistency:** Both Elbow Method and Silhouette Scores corroborate that K=6 provides a practical balance, ensuring meaningful data segmentation without unnecessary complexity.

Cluster Profiling – Next Page
Refer to Appendix section for detailed K-Means Cluster working

K- MEANS CLUSTERING – CLUSTER SUMMARY



Cluster 0 (Largest, Balanced Performance)

- Most common with 270 members.
- Moderate values in financial metrics, no extremes.
- Dominated by Industrials and Financials.
- Good cluster for investors seeking diversified, stable stocks.

Cluster 1 (High Net Worth, Negative Cash Flow)

- Smaller cluster with only 11 stocks.
- Highest Net Income, negative Net Cash Flow.
- Financials and Health Care lead.
- High risk due to negative cash flow despite high net worth.

Cluster 2 (High Valuation, Tech-Lead)

- 24 stocks with high P/E and P/B ratios.
- Strong presence in Information Technology.
- Suitable for growth-focused investors due to high valuation metrics.

Cluster 3 (Premium Pricing, Health Care Inclined)

- Only 6 stocks, very high Current Price and P/E Ratio.
- Health Care centric.
- Likely overvalued stocks or with high investor expectations.

Cluster 4 (Extreme Negative Features, Energy Sector)

- 2 stocks with extreme negative values across several features.
- Both stocks in Energy sector.
- Bad cluster indicating financially troubled companies.

Cluster 5 (Negative Performers, Energy Dominated)

- 27 stocks, most with negative Price Change and ROE.
- Predominantly Energy sector stocks.
- Another bad cluster indicating underperforming companies.

Concluding Remarks:

K-Means clustering highlights diverse profiles, from the stable, large cluster (Cluster 0) in Industrials and Financials, to the financially troubled smaller clusters (Cluster 4 and 5) in Energy. This distinction aids in directing investments towards reliable sectors and avoiding high-risk areas.

HIERARCHICAL CLUSTERING

Euclidean Distance and Ward Linkage to Define 6 Hierarchical Clusters

1. **Euclidean Distance Metric:** Euclidean distance was favored for its relatively high Cophenetic Correlation Coefficients across linkages, particularly with 'Average', ensuring distinct and effective cluster formation.

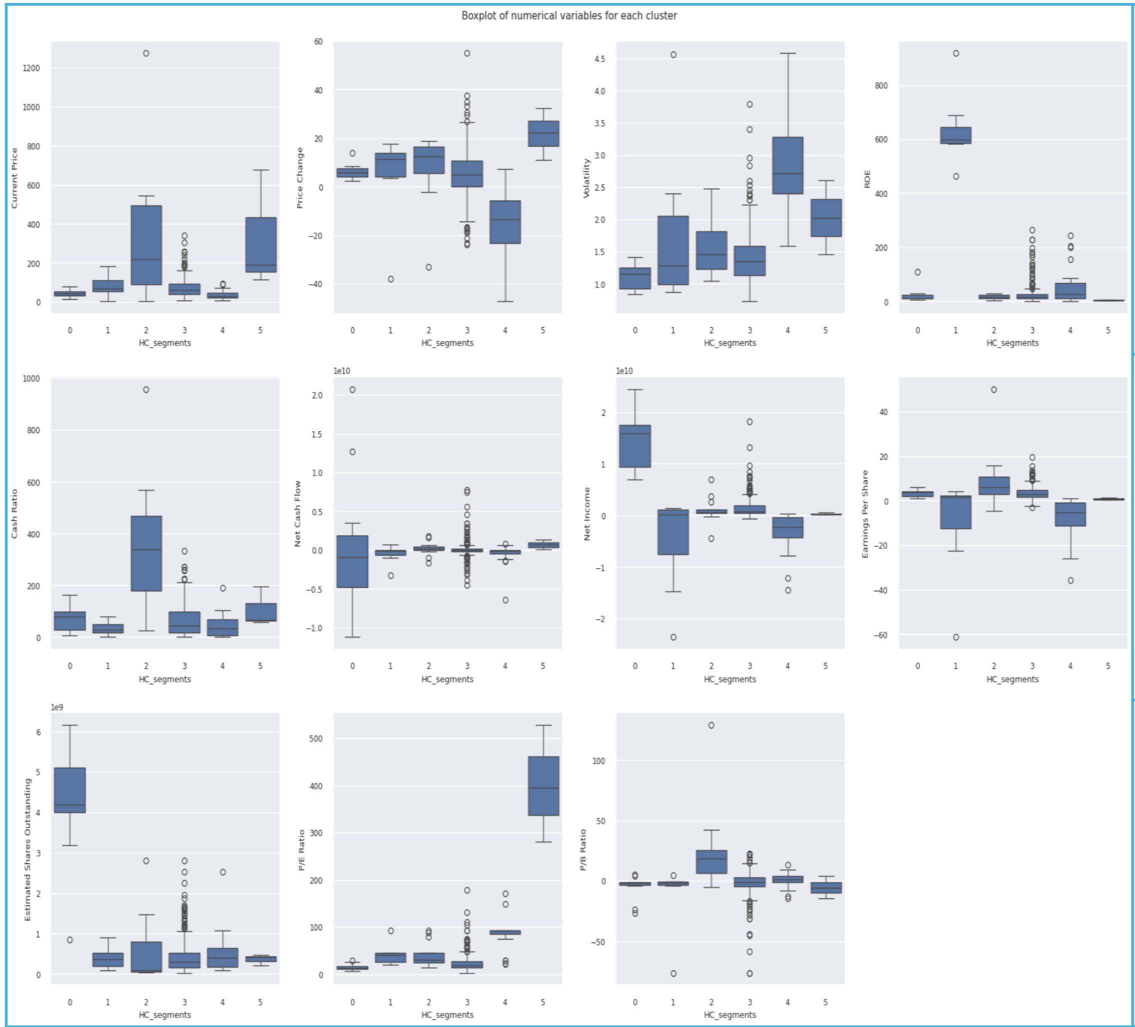
2. **Ward Linkage over Average:** Despite Average linkage's high CCC, Ward linkage was chosen for its clear, distinct cluster separations in dendrograms and superior ability to minimize intra-cluster variance.

3. **Justification for 6 Clusters:** The choice of 6 clusters, guided by natural groupings in the Ward linkage dendrogram, ensures precise differentiation and maintains a balance, avoiding over-simplification or excessive complexity.

Cluster Profiling – Next Page

Refer to Appendix section for detailed Hierarchical Cluster working

HIERARCHICAL CLUSTERING – CLUSTER SUMMARY



Cluster 0 (Diverse Financial Focus)

- 11 securities, relatively small cluster.
- Moderate prices; high Price Change, Volatility.
- Dominated by Financials and Telecom.
- Indicates a variety of companies with significant recent growth or recovery.

Cluster 1 (Negative Anomalies)

- Very small cluster with 7 stocks.
- Moderate price with negative Price Change.
- Negative Net Income; lowest in Cash Ratio.
- Includes Consumer Staples and Energy; may represent undervalued or distressed stocks.

Cluster 2 (Stable, High-Liquidity Sector)

- A small cluster with 12 stocks.
- High Current Price, substantial Price Change.
- Very high Cash Ratio; solid Net Income, ROE.
- Led by Health Care and Information Technology; suggests well-established, financially robust companies.

Cluster 3 (Largest, Industrially Diverse)

- Largest cluster with 285 securities.
- Varied sectors; lower Price.
- ROE and Cash Ratio are moderate, indicating consistent performance.
- Primarily Industrials and Financials, reflecting traditional market stability.

Cluster 4 (Energy Dominated, High Risk)

- 22 industry specific stocks.
- Lowest Current Price and substantial negative Price Change.
- Very negative Net Income; highest P/E Ratio.
- Energy sector concentration, potentially high-risk investment opportunities.

Cluster 5 (High Value, High Growth Expectations)

- Smallest cluster with 3 stocks.
- Highest Current Price and Price Change.
- Extremely high P/E Ratio and lowest ROE.
- Suggests exclusive, potentially overvalued stocks with high growth expectations.

Concluding Remarks:

These clusters exhibit a broad spectrum from highly liquid, stable stocks (Cluster 2) to high-risk, volatile sectors (Cluster 4). The largest cluster (Cluster 3) reflects a cross-section of the market's backbone industries, while the smallest (Cluster 5) represents outlier companies with exceptional stock metrics.

INSIGHTS & RECOMMENDATIONS

As Trade&Ahead's appointed data scientists, we've derived key insights and tailored recommendations from our clustering analysis to enhance the company's investment strategies, ensuring precise market positioning and optimized financial outcomes.

ACTIONABLE INSIGHTS

Segment Identification:

- Focus investment strategies on clusters dominated by impactful sectors such as Industrials, Financials, Health Care, and Energy to capitalize on their market influence.

Outlier Utilization:

- Analyze clusters containing extreme outliers or companies with unique financial characteristics to uncover high-reward investment opportunities.

Risk Management:

- Utilize the uniformity in financial behaviors observed in K-Means clusters to predict and manage investment risks more effectively.

Market Dynamics:

- Leverage the detailed data structuring from Hierarchical clustering to gain insights into complex market dynamics and inter-sector relationships, crucial for strategic decision-making.

Risk Characterization and Management:

- Identify similar risk profiles among stocks through clustering, aiding in the creation of diversified portfolios that balance risk across different sectors.

Dynamic Portfolio Diversification:

- Employ low correlation clusters for stable investments and high correlation clusters for potentially higher returns, enhancing portfolio diversification dynamically.

Trend Identification and Adaptation:

- Detect and adapt to trends within sectors using cluster analysis to strategically modify client portfolios in response to natural market fluctuations and predictions.

RECOMMENDATIONS

Cluster-Specific Strategies:

- Tailor investment strategies to the distinct characteristics of each cluster, focusing on sectors or companies predicted to outperform based on the clustering analysis.

Further Analysis on Outliers:

- Conduct in-depth financial analysis on outlier companies to assess whether their market positions are due to disruptions, innovations, or anomalies that might introduce risks or opportunities.

Dynamic Cluster Review:

- Continuously update and review the clustering models to reflect the latest market data and trends, ensuring the investment strategies remain aligned with current market conditions.

Leverage Hierarchical Insights:

- Use insights from hierarchical clustering to formulate a multi-layered investment approach that adjusts dynamically based on evolving market conditions and cluster data.

Educate Stakeholders:

- Enhance stakeholder engagement through visualizations and comprehensive reports that outline the strategic value of clustering-based investment decisions.

Technological Investment:

- Invest in advanced analytics to refine clustering techniques continually, incorporating real-time data and predictive analytics to anticipate future market movements effectively.

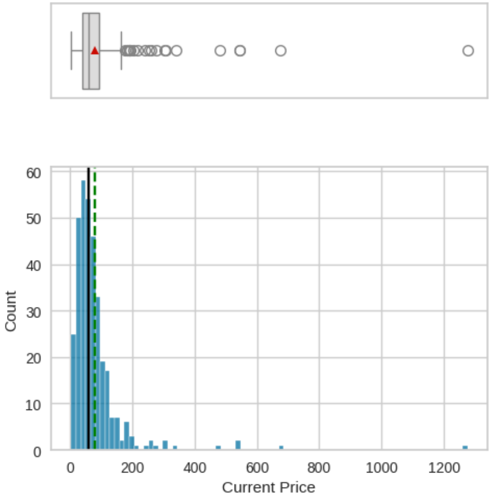
APPENDIX

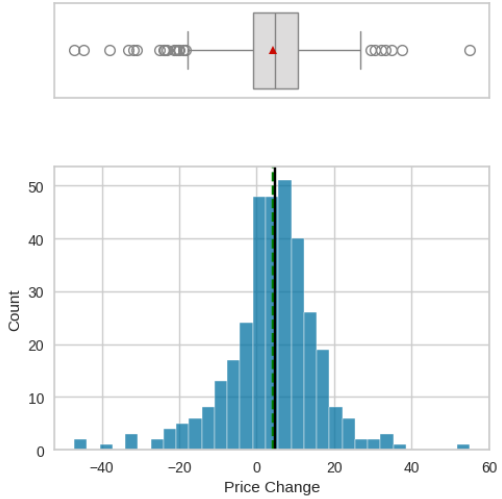
Trade&Ahead

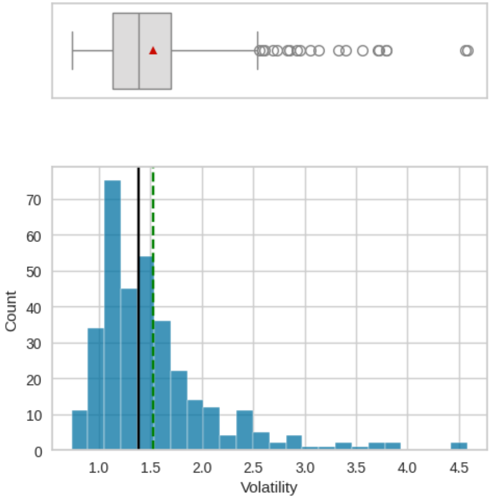
EXPLORATORY DATA ANALYSIS (EDA)

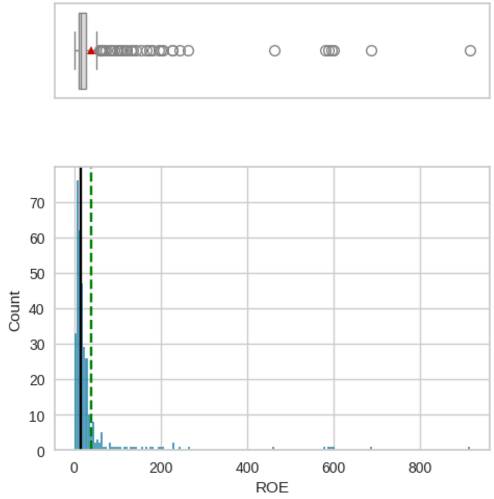
EDA (Exploratory Data Analysis) will reveal trends and key features of the dataset through univariate and bivariate analysis, providing a deeper understanding of the underlying patterns and relationships within the data.

EDA - UNIVARIATE ANALYSIS (1 OF 3)



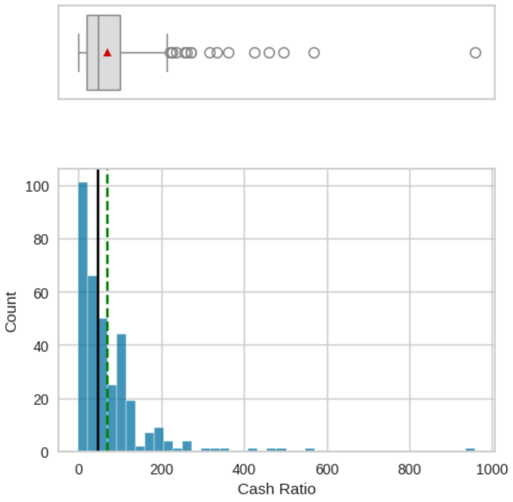


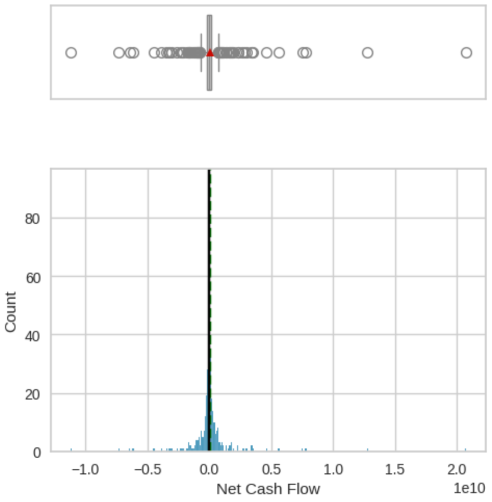


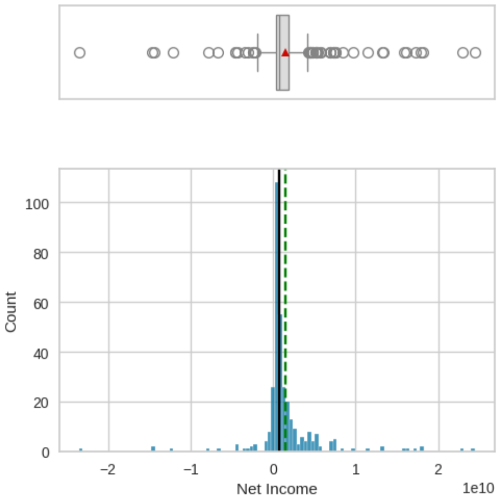


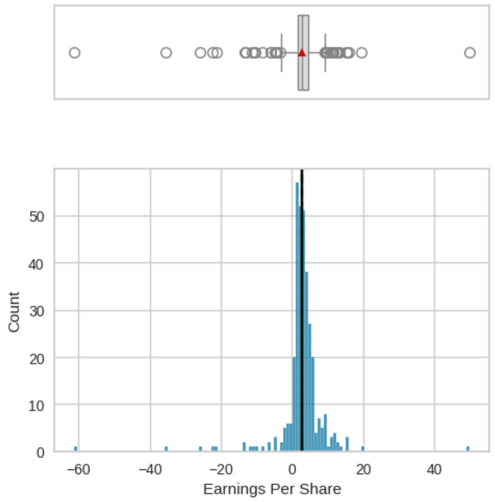
Current Price	Price Change	Volatility	ROE
Range and Variability: • Prices range from $4.50 to $1274.95 with a mean of $80.86, indicating a diverse set of companies from various sectors and sizes. • Exhibits a right-skewed distribution with outliers indicating notably high-priced stocks. **Investment Insight:** • Categorises stocks in various price bands, essential for portfolio diversification.	**Identify Volatility:** • Exhibits a range from -47.13% to 55.05%, with a mean of about 4.08%, highlighting potential high-risk and high-return stocks. • Normally distributed with outliers, particularly showing extreme decreases in some stock prices. **Strategy Impact:** • Crucial for momentum strategies that capitalize on trends in stock price movements.	**Measure of Risk:** • Standard deviation ranges from 0.73% to 4.58%, with stocks showing higher volatility offering higher returns at increased risk. • Right-skewed with a few high outliers, suggesting occasional abnormal price fluctuations. **Portfolio Planning:** • Essential for risk assessment, aiding in the selection of stocks that align with the investor's risk tolerance.	**Performance Indicator:** • Ranges dramatically from 1% to 917%, with a mean of 39.60%, indicating varying levels of company efficiency in generating profits. • Highly skewed with substantial outliers, indicating some exceptionally profitable companies. **Strategic Allocation:** • Higher ROE values indicates well-managed companies, potentially more attractive for long-term investment.

EDA - UNIVARIATE ANALYSIS (2 OF 3)



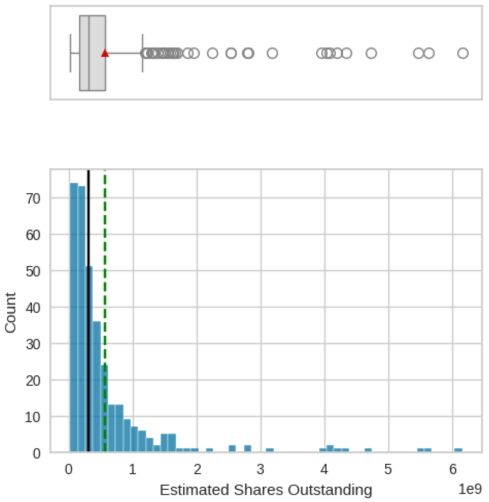


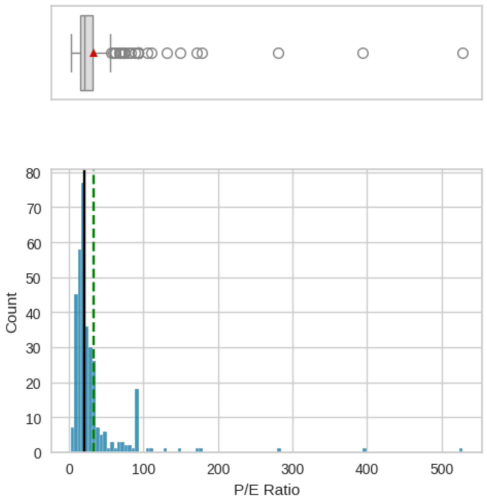


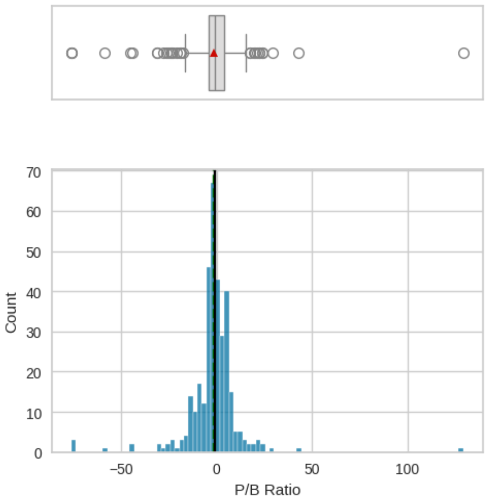


Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share
Liquidity Measurement: • Ranges from 0 to 958 with a median of 47, showing varying levels of immediate financial stability among companies. • Skewed towards lower values with extreme outliers reflecting very high liquidity. **Investment Decision Making:** • Crucial for risk-averse investors helps evaluating companies with high liquidity, essential during economic downturns.	**Financial Health:** • Varies widely from -11.21 billion to 20.76 billion, reflecting different operational and financial activities. • Approx normal distribution with significant outliers, reflecting major operational or financial activities. **Strategy Impact:** • Indicates companies' abilities to generate cash, which is vital for assessing investment sustainability and growth potential.	**Profitability Scale:** • Ranges from -23.53 billion to 24.44 billion, with significant deviations indicating profitability or financial struggles. • Slightly Right-skewed with both positive and negative outliers, indicating varied profitability and losses. **Core Analysis:** • Directly impacts dividend payments and reinvestment potential, critical for income-focused investment strategies.	**Profitability per Share:** • Ranges from -61.20 to 50.09, with a mean of 2.78, providing a basis for comparing profitability on a per-share basis. • Skewed with notable negative outliers, indicating companies with significant losses per share. **Valuation Metrics:** • Often used in conjunction with P/E ratio to evaluate whether a stock is over or under-valued relative to its earnings.

EDA - UNIVARIATE ANALYSIS (3 OF 3)







Estimated Shares Outstanding

Company Size Indicator:

- Ranges significantly from about 27.67 million to over 6.16 billion shares, illustrating the scale of companies.
- Right-skewed with outliers at higher values, pointing to companies with a large number of shares.

Market Impact:

- Higher shares outstanding can affect stock liquidity and volatility, important for assessing market impact risk.

P/E Ratio

Valuation Metric:

- Ranges from 2.94 to 528.04, with a mean of 32.61, indicating how much investors are willing to pay per dollar of earnings.
- Right-skewed with extreme high outliers, suggesting overvalued stocks.

Investment Relevance:

- Critical for determining stock valuation, helping investors decide if a stock is overpriced or underpriced in the market.

P/B Ratio

Asset Valuation:

- Fluctuates from -76.12 to 129.06, indicating diverse valuations of company assets relative to market prices.
- Normal distribution, with extreme values in both directions, indicating significant variances in valuation relative to book value.

Strategic Importance:

- Useful for value investing strategies, focusing on companies potentially undervalued relative to their actual asset base.

Key Insights:

High Variability:
Several metrics like Net Cash Flow and Net Income show extreme ranges, indicative of diverse financial strategies and outcomes.

Skewness and Outliers:
Most features are right-skewed with notable outliers, especially in metrics such as P/E Ratio and Cash Ratio, which can impact investment decisions.

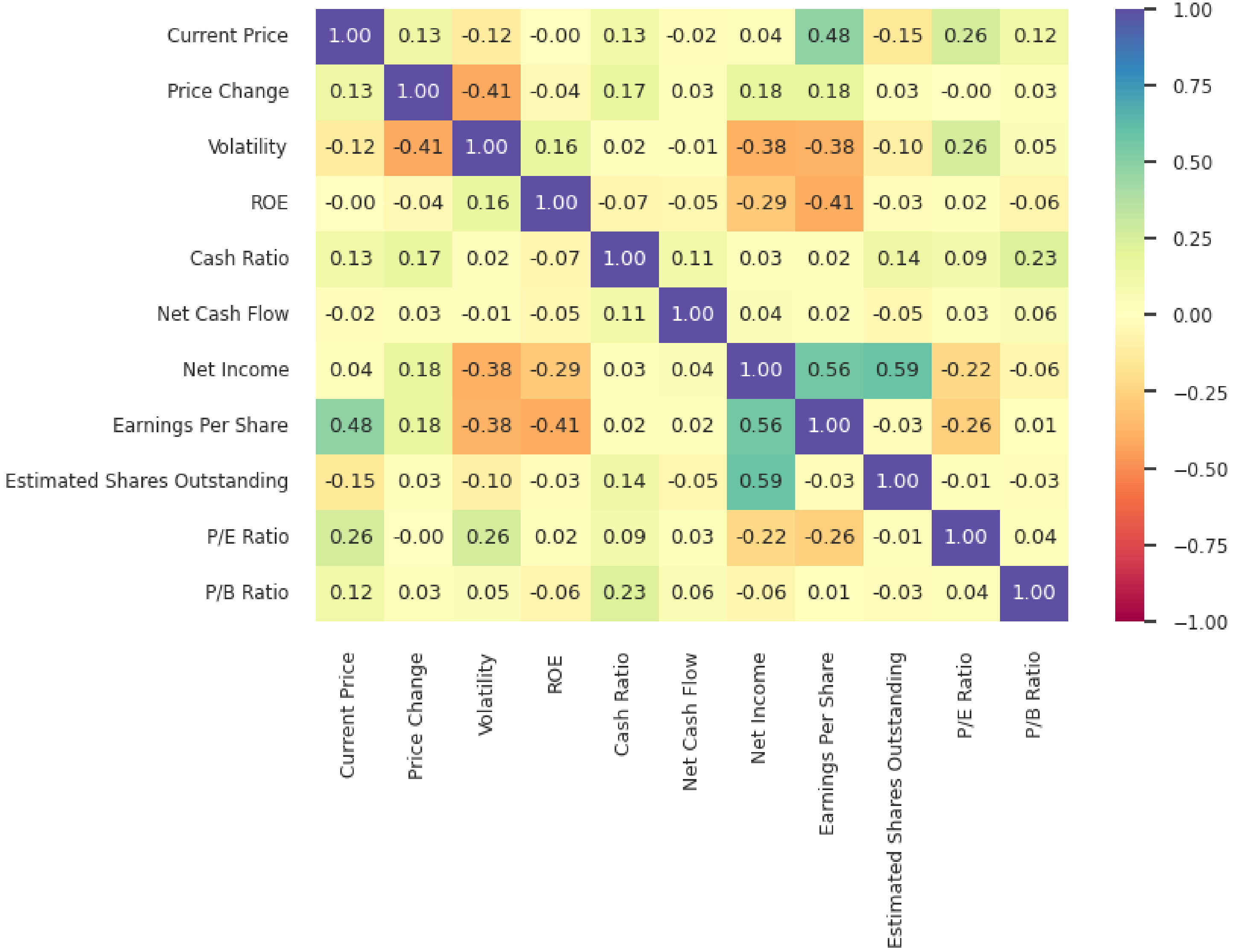
Investment Implications:
Features with wide distributions and outliers provide opportunities for targeted investment strategies, focusing on either growth-oriented or stability-focused stocks.

Further Actions:

Outlier Analysis:
Investigate extreme outliers to understand their causes and implications for risk.

Segmentation:
Use clustering to segment companies based on financial characteristics for more nuanced investment approaches.

EDA – BIVARIATE (HEATMAP)



	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

Key Insights:

Current Price and Earnings Per Share:

Moderate positive correlation indicates stock price tends to rise with higher earnings.

Net Income and Estimated Shares Outstanding:

Moderate positive relationship, suggesting that companies with higher net income tend to have more shares outstanding..

Net Income and Earnings Per Share:

Strong positive correlation highlights that higher net income results in higher earnings per share.

Net Income and Volatility:

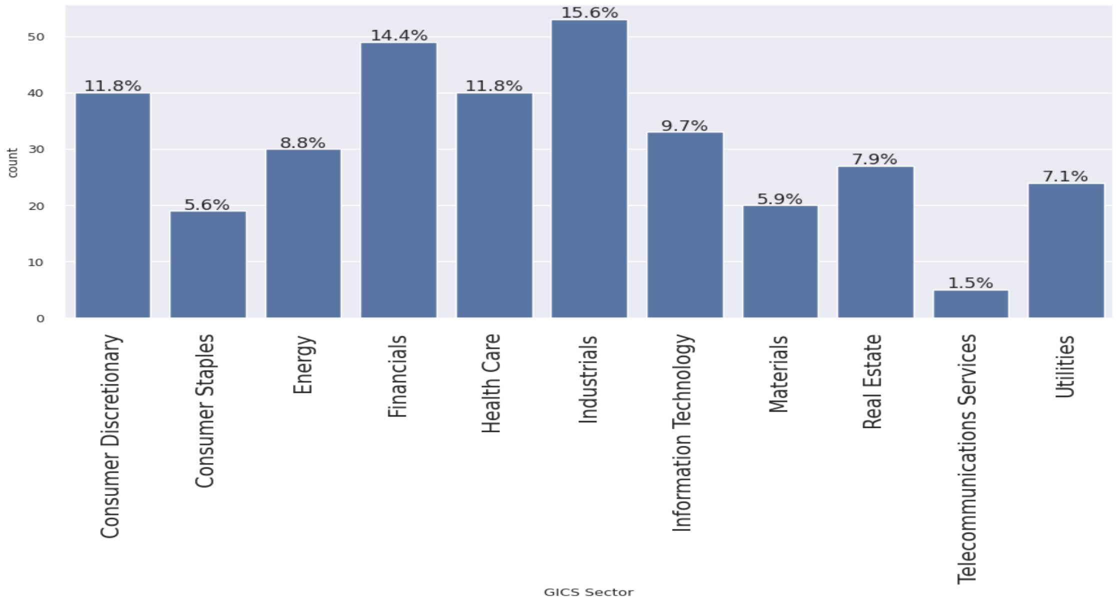
Higher Net Income or Earnings Per Share typically results in reduced volatility, stabilizing price changes.

Concluding Remarks:

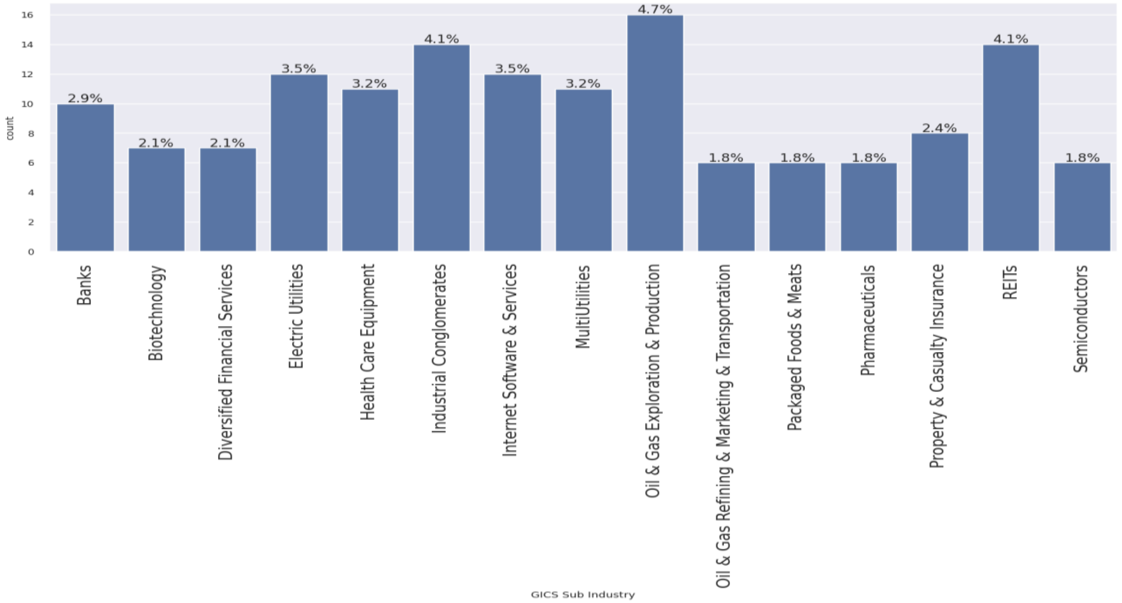
Notable correlations between key financial metrics provide valuable insights for investment strategies. These bivariate relationships indicate that stock value, company profitability, and financial health are intertwined.

EDA – ECONOMIC SECTORS (1 OF 3)

Economic Sectors Summary



Top 15 Sub-Industries

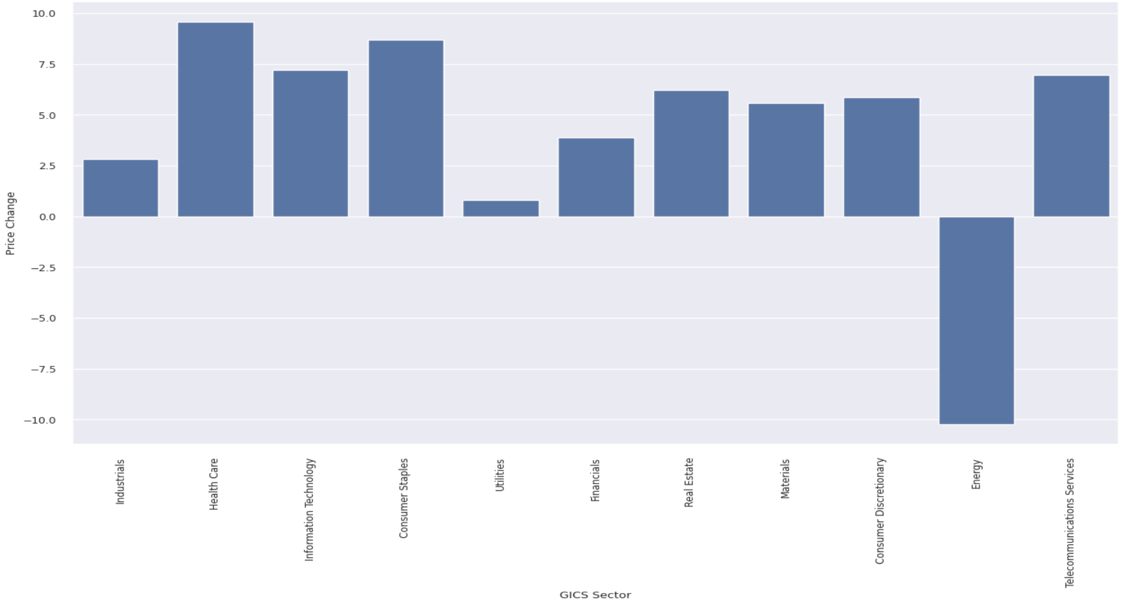


- Industrials and Financials sectors represent the largest portions of the dataset, with Industrials having the highest representation at 15.59% and Financials at 14.41%.
- Telecomm sector is the least represented sector at just 1.47%, indicating it might be a more specialized or smaller sector compared to others.
- Consumer Discretionary and Healthcare, both are well-represented, each comprising a noticeable share of the dataset at around 11.8%.
- The dataset reveals a market led by Industrials and Financials, with varied engagement across sectors from Tech to Telecommunications, shaping a multifaceted economic landscape.

- Oil & Gas Exploration & Production is the top sub-industry, indicating a strong energy sector focus.
- REITs and Industrial Conglomerates are significantly represented, showing the importance of real estate and industrial diversity.
- The data highlights a strong presence of energy, real estate, and industrial sectors, alongside a diverse mix from utilities, tech, and healthcare, crucial for informing market strategies.

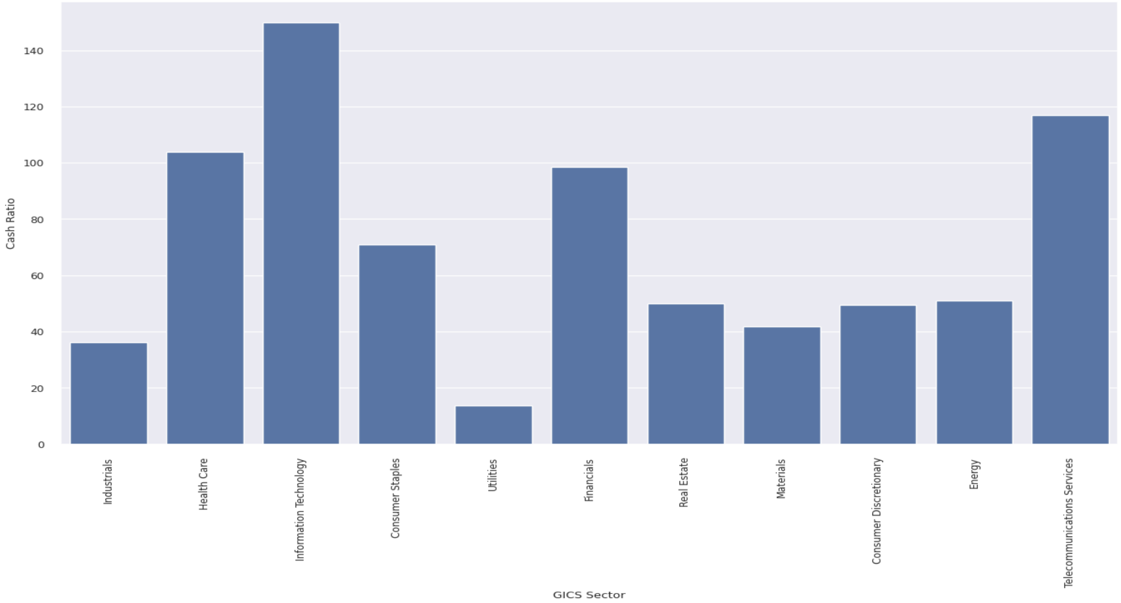
EDA – ECONOMIC SECTORS (2 OF 3)

Economic Sector vs Price Change



- There's a notable variation in price change among different GICS Sectors.
- Most sectors seem to show a positive price change, suggesting a period of general growth or recovery in those sectors.
- The Energy sector particularly stands out with a substantial negative price change of more than 10%, indicating a decrease in stock prices which could reflect sector-specific challenges or a market downturn.
- Healthcare and Consumer Staples sectors are leading with higher price changes, over 8% increase, hinting at strong sector performance or favourable market conditions during the measured timeframe.

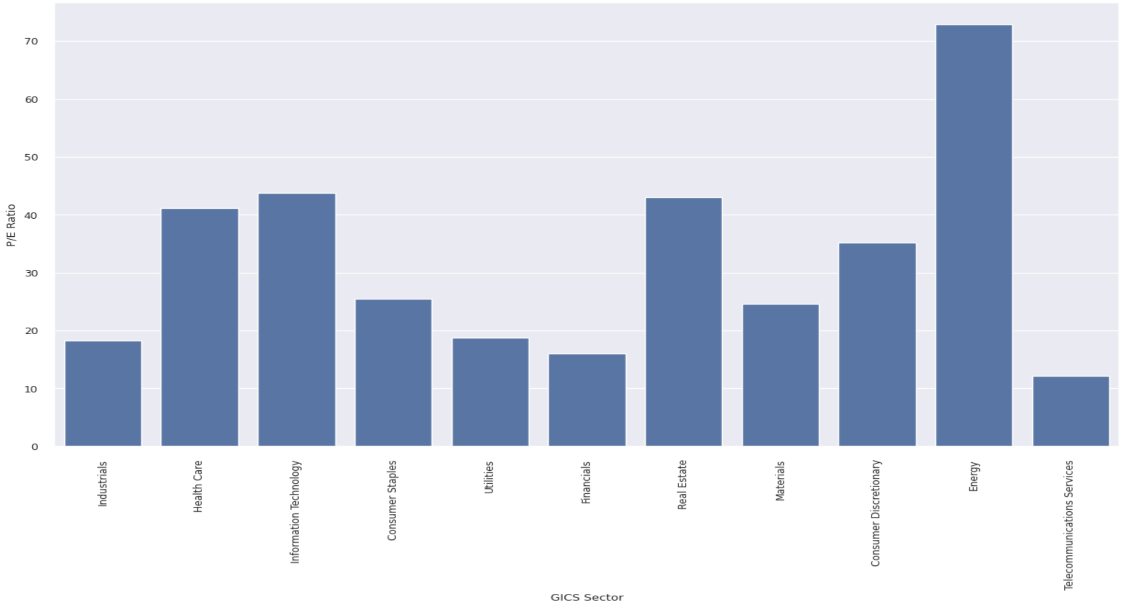
Average Cash Ratio across Economic Sectors



- Information Technology sector exhibits a notably high average cash ratio of over 140%, suggesting strong liquidity and the ability to cover short-term liabilities.
- Telecomm, Financials and Healthcare sectors also show higher-than-average cash ratios of around 100%, indicating healthy financial buffers.
- Utilities sector stands out with a significantly lower cash ratio, which could signal potential liquidity concerns or a different capital structure that relies less on cash reserves.
- Other sectors show cash ratio on lower side.

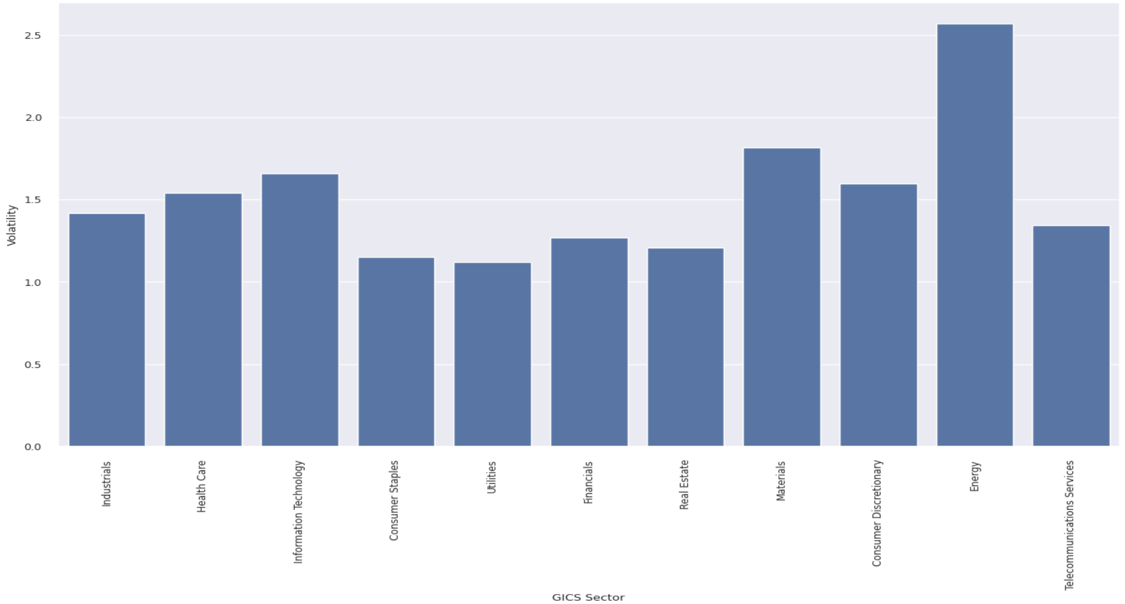
EDA – ECONOMIC SECTORS (3 OF 3)

Economic Sector vs P/E Ratio



- Energy sector exhibits the highest P/E Ratio, which may suggest expectations of significant earnings growth or that stocks are currently overvalued relative to earnings.
- Telecomm sector has the lowest P/E Ratio, potentially indicating undervaluation relative to earnings or conservative growth expectations by the market.
- Real Estate, Information Technology, Health Care and Consumer Discretionary sectors show moderately high P/E Ratio.

Volatility in Stock Prices by Economic Sectors

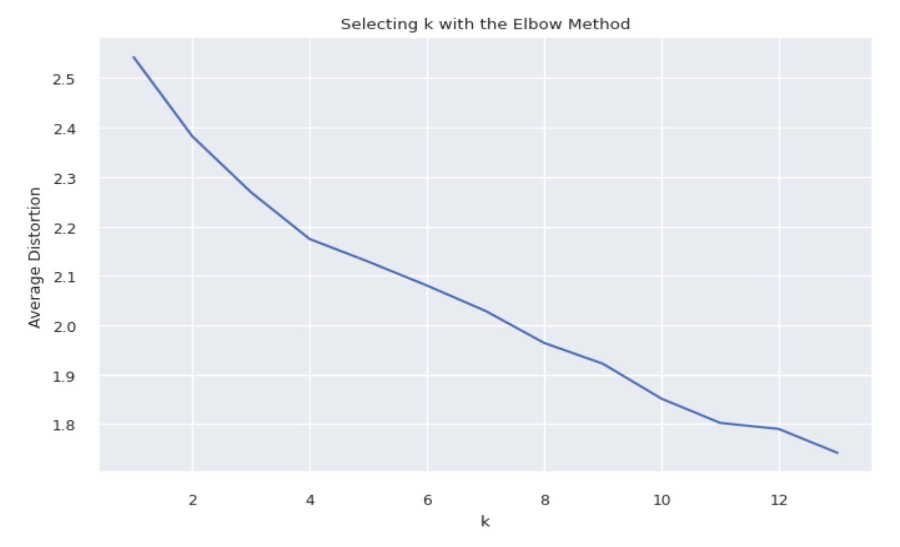


- Energy sector exhibits the highest average volatility, suggesting that stocks in this sector experience more significant price fluctuations.
- Consumer Staples and Utilities stock show the lowest volatility, a feature of defensive sectors, more stable during various market conditions.
- Materials, IT, Consumer Discretionary and Healthcare show moderately high volatility indicating greater degree of price variations.
- Other sectors show low or moderate volatility in stock prices, which could point to a balanced mix of stable companies and those sensitive to economic cycles within the sector.

K-MEANS CLUSTERING

K-Means clustering is employed to stratify NYSE-listed companies into distinct groups based on financial characteristics, aiding Trade&Ahead to tailor investment strategies and enhance portfolio diversification effectively.

K-MEANS - ELBOW METHOD



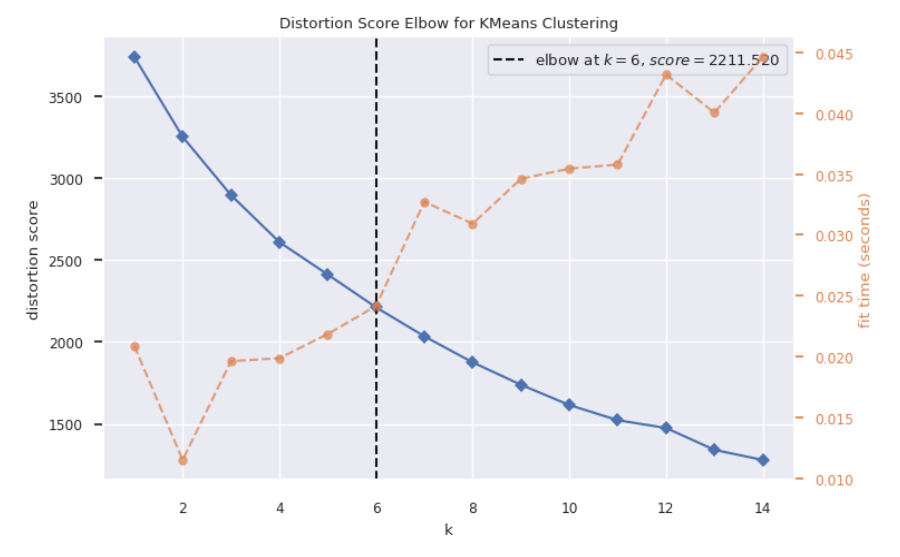
Finding Best Cluster Value using Elbow Method

Determining Optimal Clusters using Elbow Method:

- Applied K-Means for K values from 1 to 14 on scaled data.
- Captured Average Distortion for each Cluster size.
- Plotted Cluster size and Average Distortion to identify the best cluster size using Elbow Method.

Observations:

- From the Elbow chart, the appropriate value for K seems to be 6.



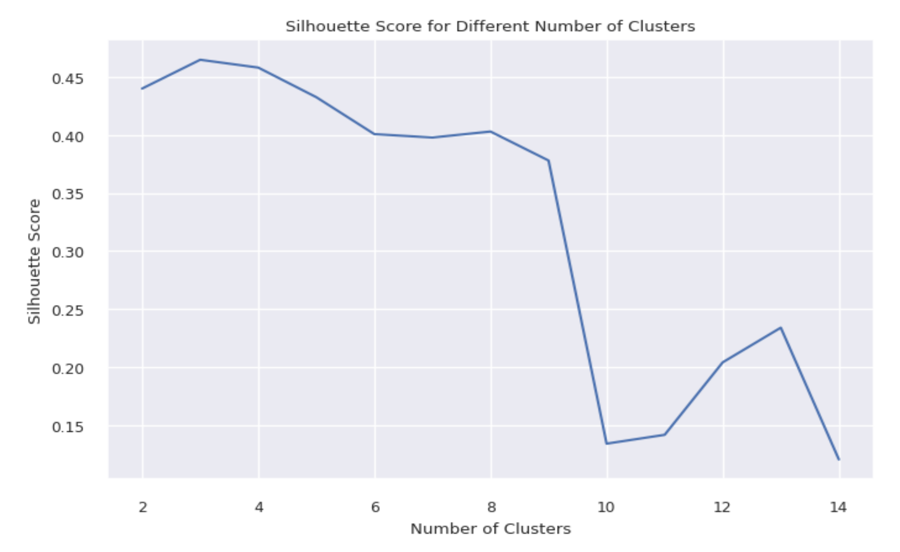
Validate Cluster (K) selection using Distortion Score and Fit-time:

- Used KElbowVisualiser() function to identify distortion score and fit time for each cluster size.

Observations:

- Confirmed elbow point at K=6 for K-means clustering.
- Beyond K=6, the distortion score, which stands at 2211.52, sees minimal reduction.
- This suggests that six clusters are sufficient for meaningful data segmentation.

K-MEANS - SILHOUETTE METHOD

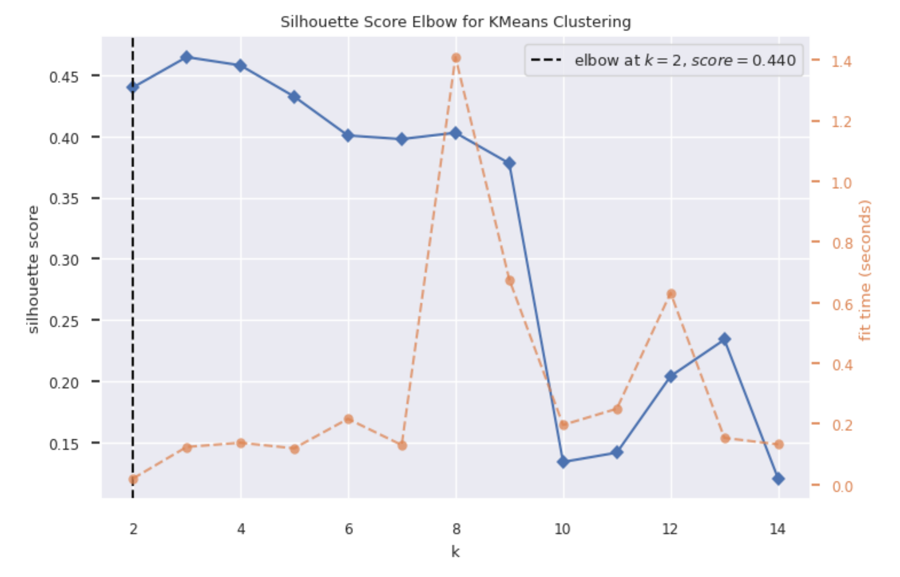


Determining Optimal Clusters using Silhouette Method:

- Applied K-Means for K values from 1 to 14 on scaled data.
- Captured Silhouette Score for each Cluster size.
- Plotted Cluster size and Silhouette Score for easy interpretation.

Observations:

- Highest silhouette score at 3 clusters, indicating strong intra-cluster cohesion.
- Scores diminish post-6 clusters, reinforcing the choice of K=6.
- Additional clusters beyond 6 do not enhance cluster definition significantly.



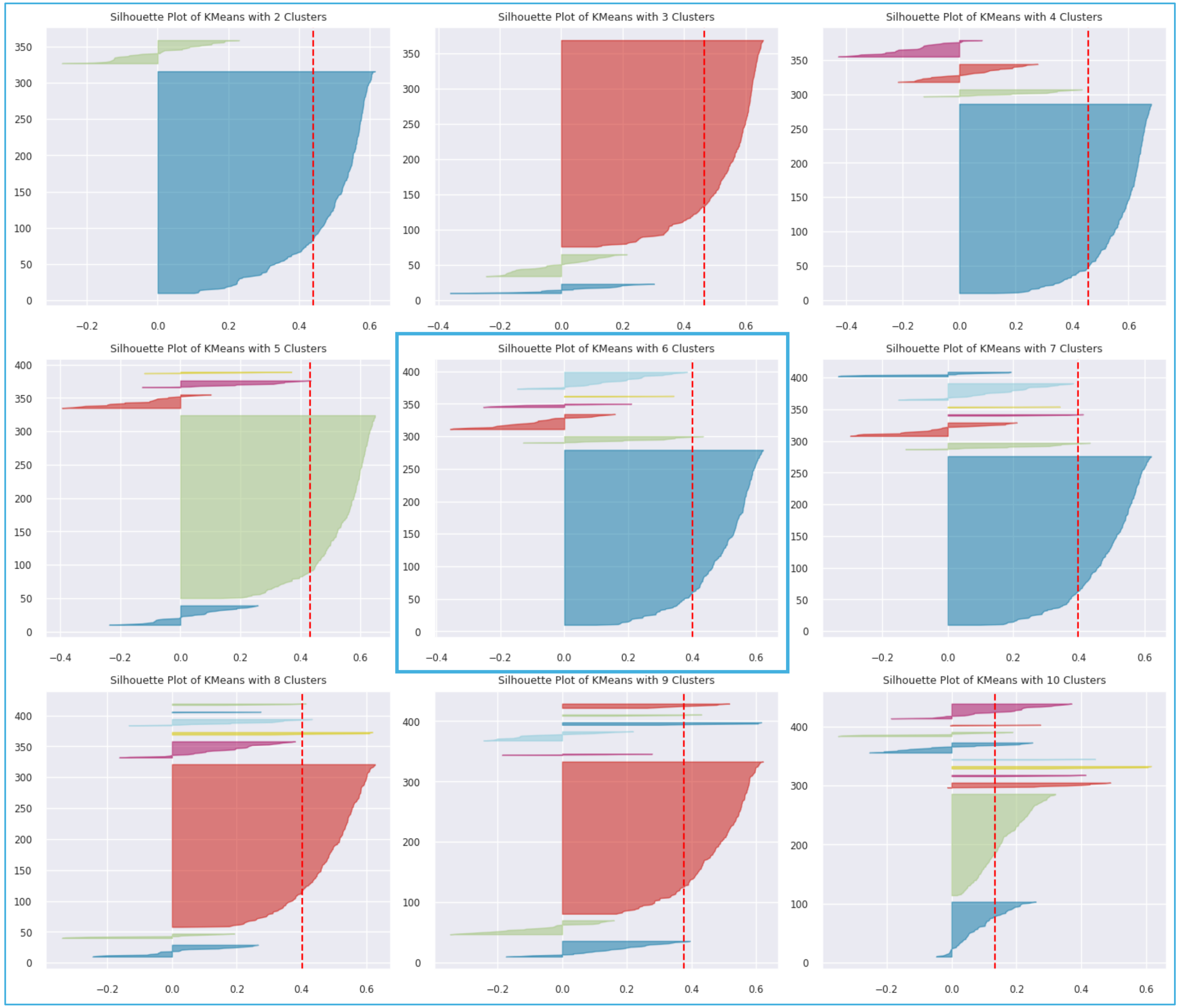
Validate Cluster (K) selection using Silhouette Score and Fit-time:

- Used KElbowVisualiser() function to identify silhouette score and fit time for each cluster size.

Observations:

- Peak silhouette score at K=2, with a value of 0.440, suggests optimal distinctiveness between two clusters.
- Contrary to previous K=6 indication, K=2 provides clearer cluster separation.
- The data is best divided into two well-defined clusters for maximum distinction.
- Next, we will compare SilhouetteVisualizer to compare Silhouette Coefficients.

K- MEANS COMPARING SILHOUETTE COEFFICIENTS


Silhouette Plot of KMeans with 2 Clusters

Silhouette Plot of KMeans with 3 Clusters

Silhouette Plot of KMeans with 4 Clusters

Silhouette Plot of KMeans with 5 Clusters

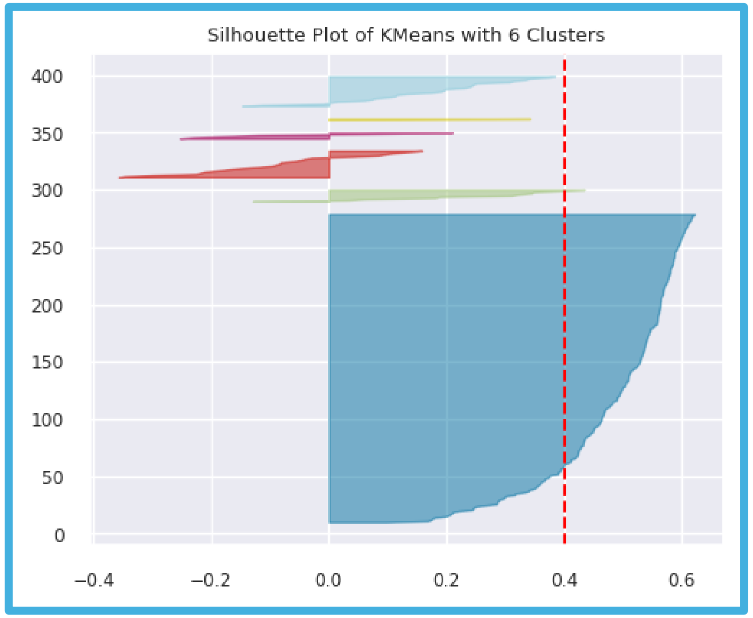
Silhouette Plot of KMeans with 6 Clusters

Silhouette Plot of KMeans with 7 Clusters

Silhouette Plot of KMeans with 8 Clusters

Silhouette Plot of KMeans with 9 Clusters

Silhouette Plot of KMeans with 10 Clusters


Silhouette Plot of KMeans with 6 Clusters

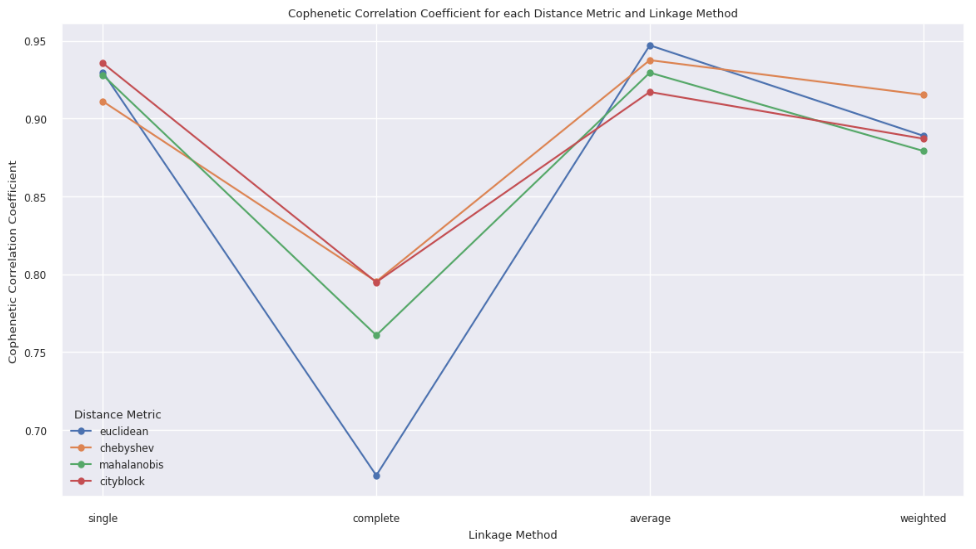
Observations:

- We will visualize Silhouette score for 2 to 10 clusters.
- Clusters K=2, 3, 4, and 5 displayed suboptimal silhouette scores with considerable score fluctuations.
- At K=6, clusters exhibited closer to average silhouette scores, suggesting more cohesive and separated groups.
- This justifies selecting **K=6** for a more stable clustering solution.

HIERARCHICAL CLUSTERING (HC)

Hierarchical clustering will unveil intrinsic groupings within the Trade&Ahead dataset based on financial similarities. Utilizing the cophenetic correlation coefficient and dendrogram, we'll assess cluster validity across different distance and linkage metrics, refining our clustering strategy for more precise investment insights.

HC - COPHENETIC CORRELATION

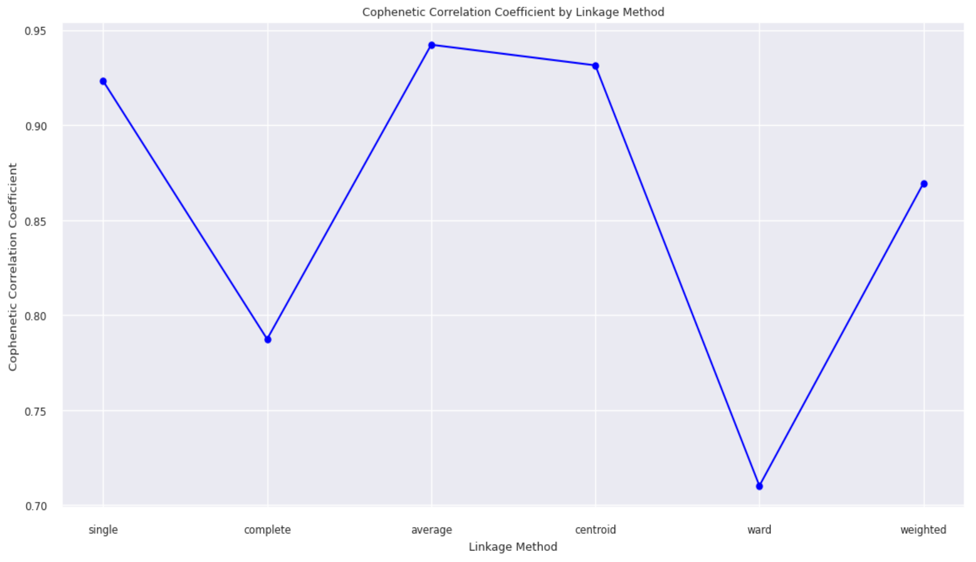


Determining Optimal Cophenetic Correlation Coefficient (CCC):

- We calculated and analyzed the CCC across Euclidean, Chebyshev, Mahalanobis, and Cityblock distance metrics, applying Single, Complete, Average, and Weighted linkage methods.

Observations:

- Average linkage with Euclidean distance yields the highest CCC at 0.94697, indicating excellent cluster representation.
- High CCCs for Single linkage suggest sensitivity to outliers.
- Additional linkage methods will be explored with Euclidean distance.



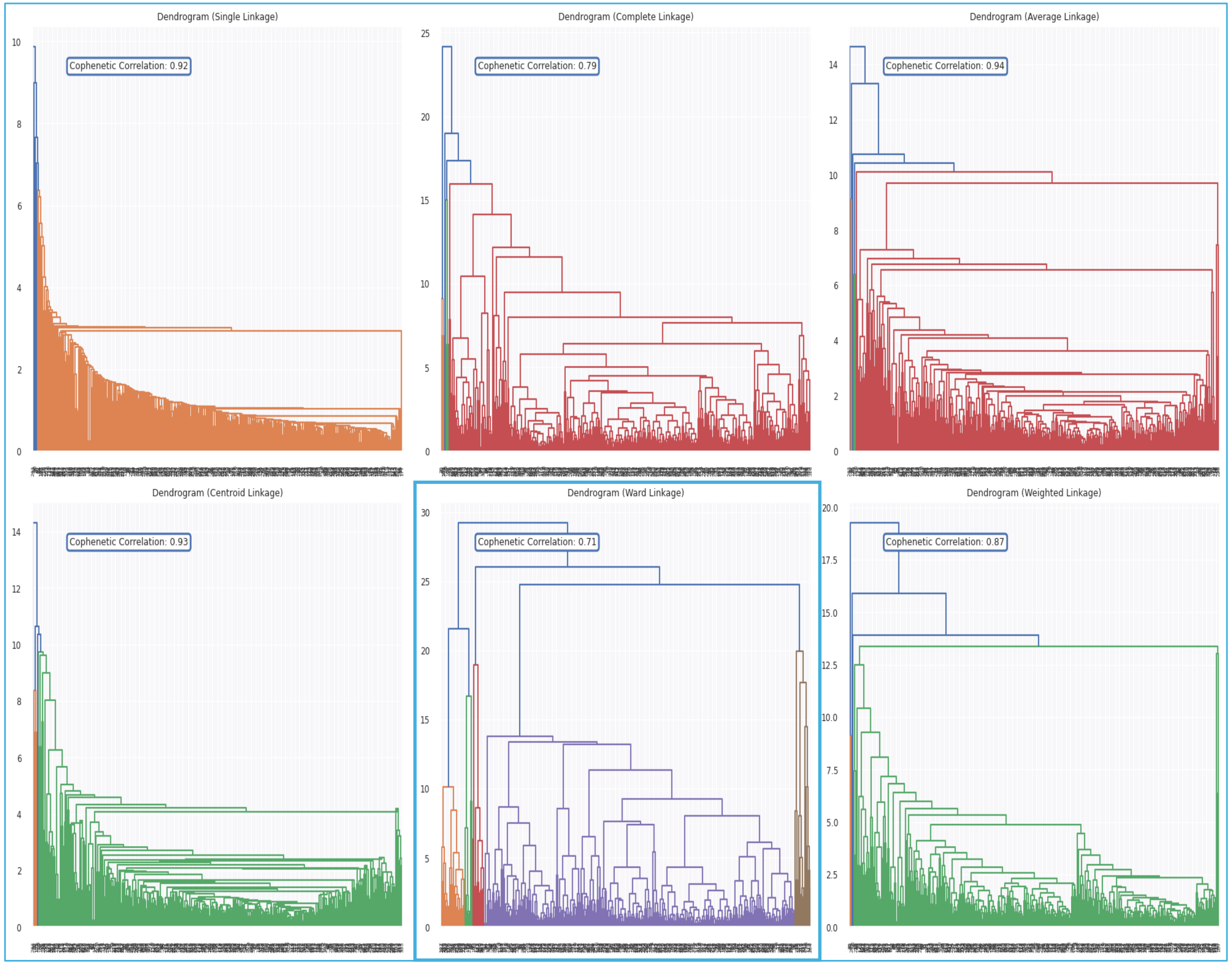
CCC using Euclidean distance metric and various Linkage methods

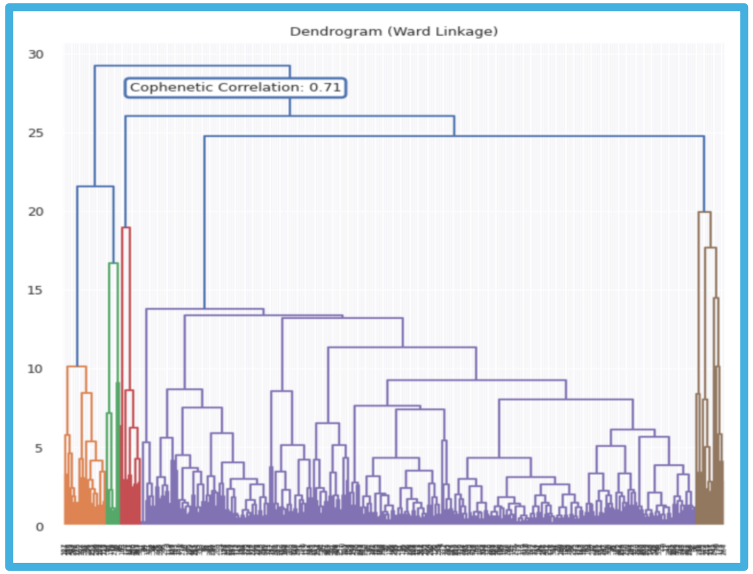
- Assuming Euclidean distance metric gave better results, we have computed CCC along with additional Linkage methods.

Observations:

- The 'average' and 'centroid' linkage methods show the highest CCC, suggesting optimal preservation of original distances.
- In contrast, ‘ward’ and 'complete’ linkage shows the low efficiency, indicating potential distortions.
- We will plot dendrograms to visually determine the most effective cluster.

HC - DENDROGRAM





Observations:

- While **Average linkage** with **Euclidean distance** exhibits the highest Cophenetic Coefficient, Ward's method with a lower CCC may indicate less compact but more distinct and meaningful clusters.
- Dendrograms from **Ward linkage** method visually demonstrate clearer separation among clusters, essential for discernible classification and actionable insights.
- The choice of **6 clusters** aligns with the natural grouping observed in the Ward method dendrogram, balancing granularity with interpretability.

K-MEANS VS HIERARCHICAL CLUSTERING

Which clustering technique took less time for execution?

- K-Means took around 16 seconds to execute as compared to Hierarchical Clustering's 33 seconds.
- Hierarchical clustering was slower primarily due to its complex dendrogram creation.

Which clustering technique gave you more distinct clusters, or are they the same?

- K-Means and Hierarchical Clustering both suggested 6 segments.
- K-Means, guided by silhouette and elbow methods, provided more distinct clusters.
- Hierarchical clustering's distinctness depends on the linkage method and may vary.

How many observations are there in the similar clusters of both algorithms?

- Both gave similar clustering split, with one very large cluster.
- The largest cluster also shows uncanny similarities for Current Price, Price Change as well economic sector distribution.
- This consistency also supports the reliability of clustering outcome.

How many clusters are obtained as the appropriate number of clusters from both algorithms?

- Both algorithms determined **6** as the appropriate number of clusters, as indicated by various metrics including the silhouette score and dendrogram analysis.

Differences and Similarities in Cluster Profiles:

Similarities:

- Both K-Means and HC predominantly identify clusters in sectors like Industrials, Financials, Health Care, and Energy.
- Each method highlights outliers with unique financial traits, crucial for strategic investment decisions.

Differences:

- K-Means tends to produce clusters of uniform sizes due to its centroid-based approach, unlike HC.
- HC offers a more detailed view of data structuring, useful for deciphering complex market dynamics.

Conclusion:

Both clustering methods bring valuable insights to the dataset, with K-Means excelling in efficiency and distinct cluster formation, and HC providing depth in data structure understanding.

Trade&Ahead

THANK YOU

SHAISHAV MERCHANT
PROJECT | UNSUPERVISED LEARNING