Business Report: Unsupervised Learning: Trade & Ahead

Prepared by:

Priyanka Chandrahar Mane

Date: 2 February 2025

Table of Contents

Unsupervised Learning: Trade & Ahead

1. Exploratory Data Analysis (EDA

1.1 Problem Definition

The stock market is a dynamic financial ecosystem where companies issue shares that investors buy and sell. Investing in stocks can yield high returns, but it also comes with inherent risks. Understanding patterns in stock behavior and identifying underlying relationships between different financial indicators can help mitigate these risks and improve investment decisions.

The primary objective of this study is to categorize stocks into distinct clusters based on their financial attributes. By leveraging data-driven techniques such as clustering, we aim to group stocks with similar performance characteristics, enabling investors to build a diversified portfolio with optimized risk-reward balance. The analysis will also provide insights into stock volatility, pricing patterns, and market trends, thereby aiding strategic investment decisions.

1.2 Data Background and Contents

The dataset used in this project comprises financial metrics and stock price data for various companies listed on the New York Stock Exchange (NYSE). The dataset includes 340 records with 15 key financial attributes. These attributes provide critical insights into each company's financial health and market position.

The dataset includes:

- **Ticker Symbol**: A unique identifier assigned to each publicly traded company.
- **Company Name**: The full name of the organization.
- **GICS Sector**: The industry classification based on the Global Industry Classification Standard (GICS), which categorizes companies into different economic sectors.
- **GICS Sub Industry**: A more granular classification of the company's industry.
- **Current Price**: The latest trading price of the stock.
- **Price Change (%)**: The percentage change in stock price over the last 13 weeks.
- **Volatility**: A statistical measure indicating the extent of price fluctuations over the past 13 weeks.
- **Return on Equity (ROE)**: A financial ratio that evaluates a company’s profitability by measuring net income as a percentage of shareholders’ equity.
- **Cash Ratio**: A liquidity ratio that indicates the company's ability to cover short-term liabilities using cash and cash equivalents.
- **Net Cash Flow**: The net amount of cash and cash equivalents moving in and out of a company over a period.
- **Net Income**: The company's total earnings after deducting all expenses, taxes, and costs.
- **Earnings Per Share (EPS)**: A measure of a company’s profitability, calculated as net income divided by outstanding shares.

- **Estimated Shares Outstanding**: The approximate number of shares held by shareholders.
- **Price-to-Earnings (P/E) Ratio**: A valuation metric comparing the stock price to its earnings per share.
- **Price-to-Book (P/B) Ratio**: A valuation metric that compares the stock price to the book value per share.

1.3 Univariate Analysis

Univariate analysis focuses on examining the distribution and statistical properties of individual variables in the dataset.

- **Stock Prices**: The price distribution was observed to be right-skewed, indicating that a majority of the stocks trade at lower prices, while a few stocks have significantly higher values.
- **Price Change**: The range of price change varied widely, with some stocks experiencing substantial gains while others declined sharply. The average price change was positive, suggesting overall market growth during the analyzed period.
- **Volatility**: The standard deviation of stock prices varied considerably, with some stocks showing high volatility, particularly in sectors like technology and energy.
- **Financial Ratios**: The ROE, cash ratio, and P/E ratio distributions showed notable differences across industries, with technology stocks generally having higher valuations and profitability compared to industrial and energy sectors.

1.4 Bivariate Analysis

Bivariate analysis explores the relationships between two variables to identify potential correlations and dependencies.

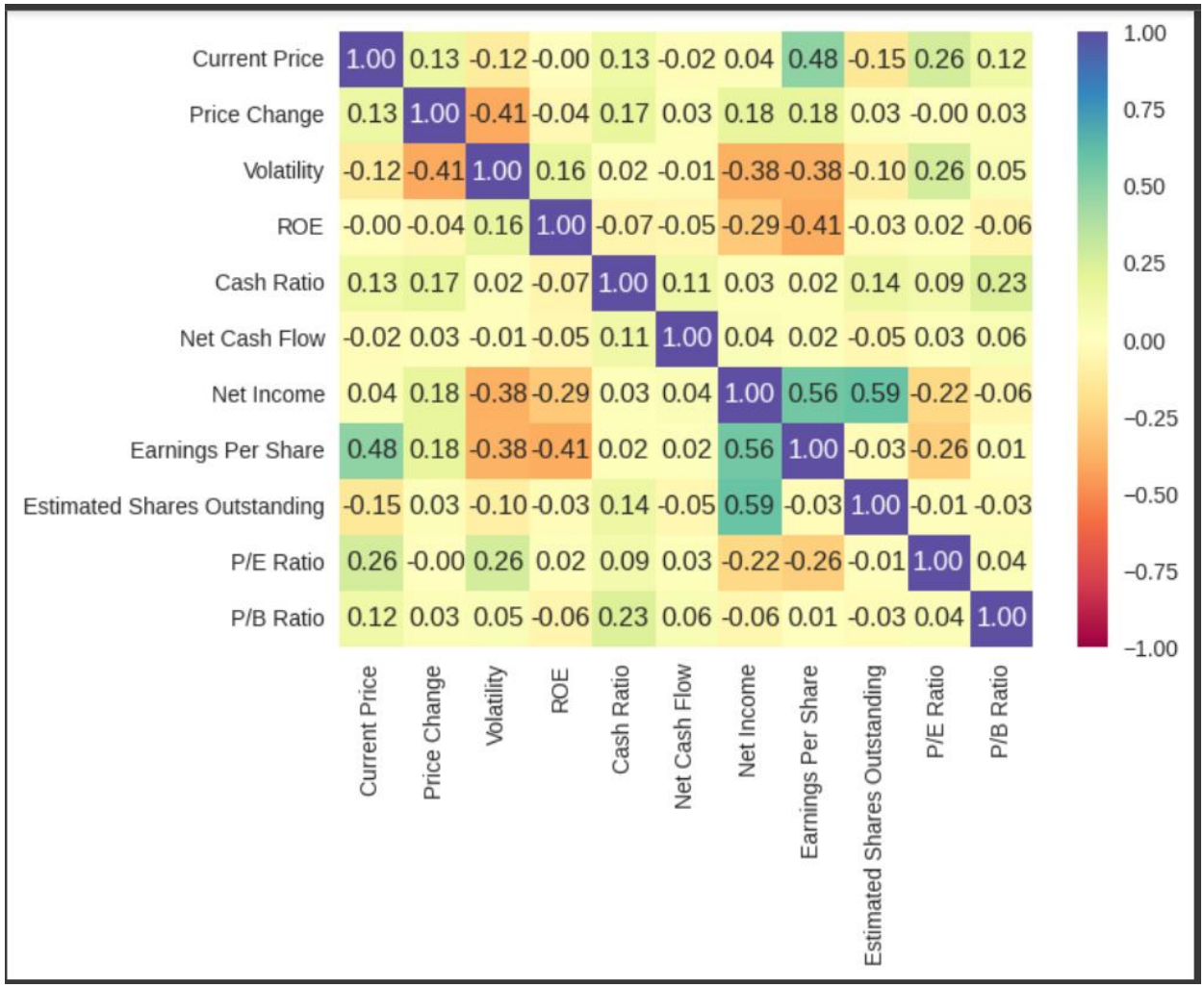
- **Stock Price vs. P/E Ratio**: A positive correlation was observed, indicating that higher-priced stocks generally have higher valuation multiples.
- **Stock Price vs. ROE**: Companies with strong profitability (high ROE) tended to have higher stock prices, reinforcing the importance of financial health in market valuation.
- **Volatility vs. Price Change**: Stocks with higher price fluctuations tended to exhibit greater volatility, underscoring the risk associated with high-growth stocks.
- **P/E Ratio vs. P/B Ratio**: A moderate correlation was found, suggesting that companies with high earnings multiples also tend to have higher book values.

1.5 Visualizations and Pattern Identification

Visual analysis plays a crucial role in uncovering hidden patterns and trends within the dataset. Key visualizations included:

- **Histograms**: Provided insights into the distribution of stock prices, ROE, and other key metrics.
- **Boxplots**: Helped detect outliers in financial metrics, particularly extreme values in P/E and P/B ratios.

- **Heatmaps**: Illustrated correlation strengths between different financial variables.



- **Scatter Plots**: Highlighted trends and relationships among key financial indicators.

1.6 Key Observations on Individual Variables and Relationships

- **Technology stocks** exhibited the highest valuation metrics, often characterized by high P/E ratios and significant price fluctuations.
- **Energy and industrial stocks** displayed lower valuation multiples, with relatively stable price trends.

- **Stocks with high ROE** tended to have stronger market performance, reinforcing the importance of profitability in stock selection.
- **High-volatility stocks** were associated with substantial price swings, often found in high-growth sectors.

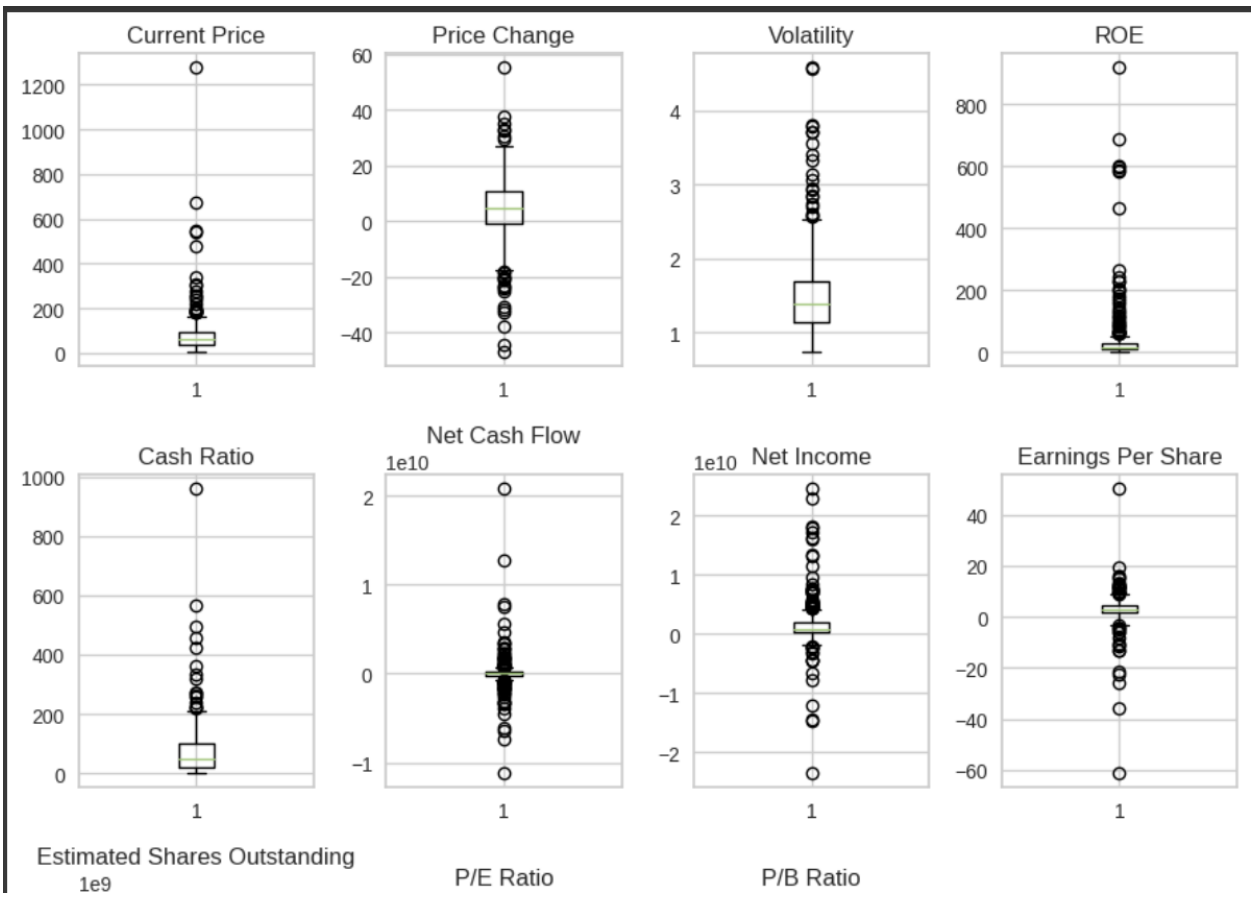
---

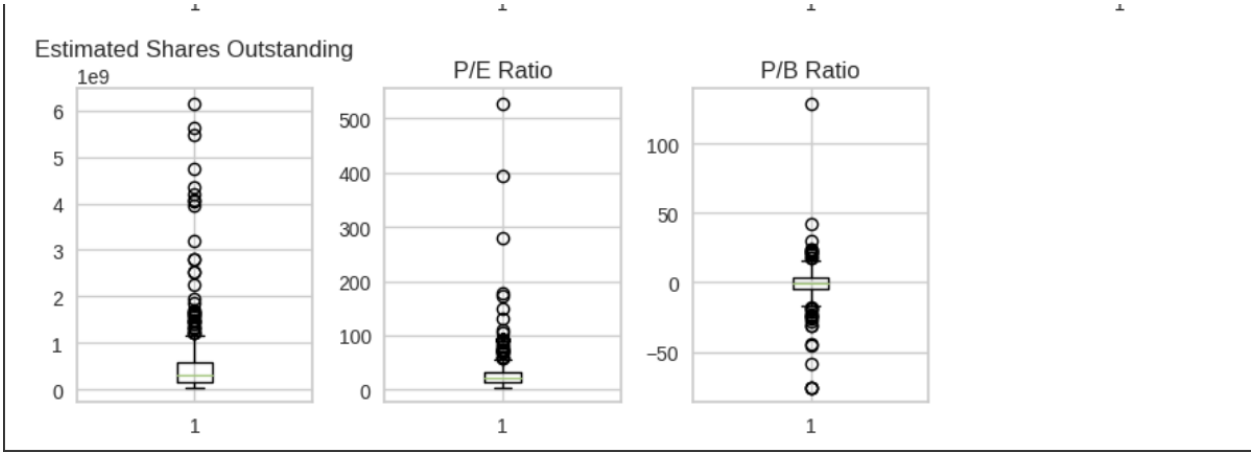
2. Data Preprocessing

2.1 Missing Value Treatment (with rationale)

A thorough check for missing values in the dataset revealed that all fields were complete. This ensured that no imputation or data-filling strategies were required.

2.2 Outlier Detection and Treatment (with rationale)

- **Boxplots were used to identify extreme values.**
- 

- 

- Some stocks displayed abnormally high P/E ratios, suggesting speculative pricing.
- Stocks with extreme P/B ratios were further analyzed to determine whether they were genuinely undervalued or subject to accounting anomalies.

2.3 Feature Engineering (with rationale)

To enhance clustering performance, additional features were created:

- **Risk-to-Return Ratio**: Calculated as Volatility / Price Change to measure stability.
- **Market Capitalization Proxy**: Derived as Stock Price * Estimated Shares Outstanding to approximate company size.

2.4 Data Scaling (with rationale)

- Standardization was applied to normalize the financial metrics, ensuring that clustering algorithms treated all variables equally without dominance from larger numerical scales.

3. K-Means Clustering

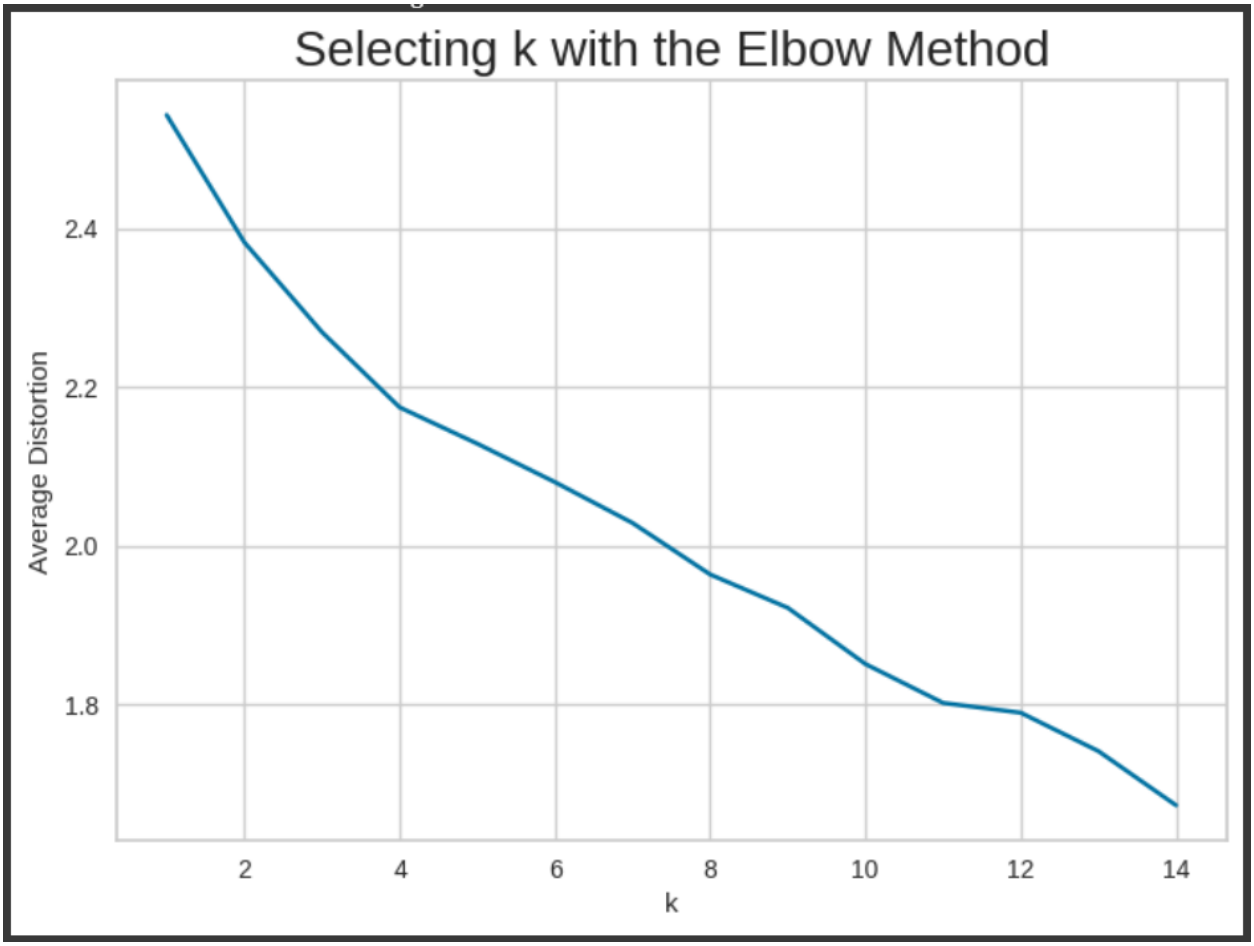
3.1 Applying K-Means Clustering

K-Means clustering is a popular unsupervised machine learning algorithm used to partition data into clusters based on similarity. The key steps involved in applying K-Means clustering to our stock dataset are:

- **Data Preparation**: The dataset was preprocessed by normalizing numerical variables to ensure fair distance calculations.
- **Choosing the Number of Clusters**: The optimal number of clusters was determined using the Elbow Method and Silhouette Score.
- **Algorithm Execution**: K-Means clustering was applied using the Euclidean distance metric, iterating through different values of 'k' to evaluate performance.

3.2 Elbow Curve Analysis

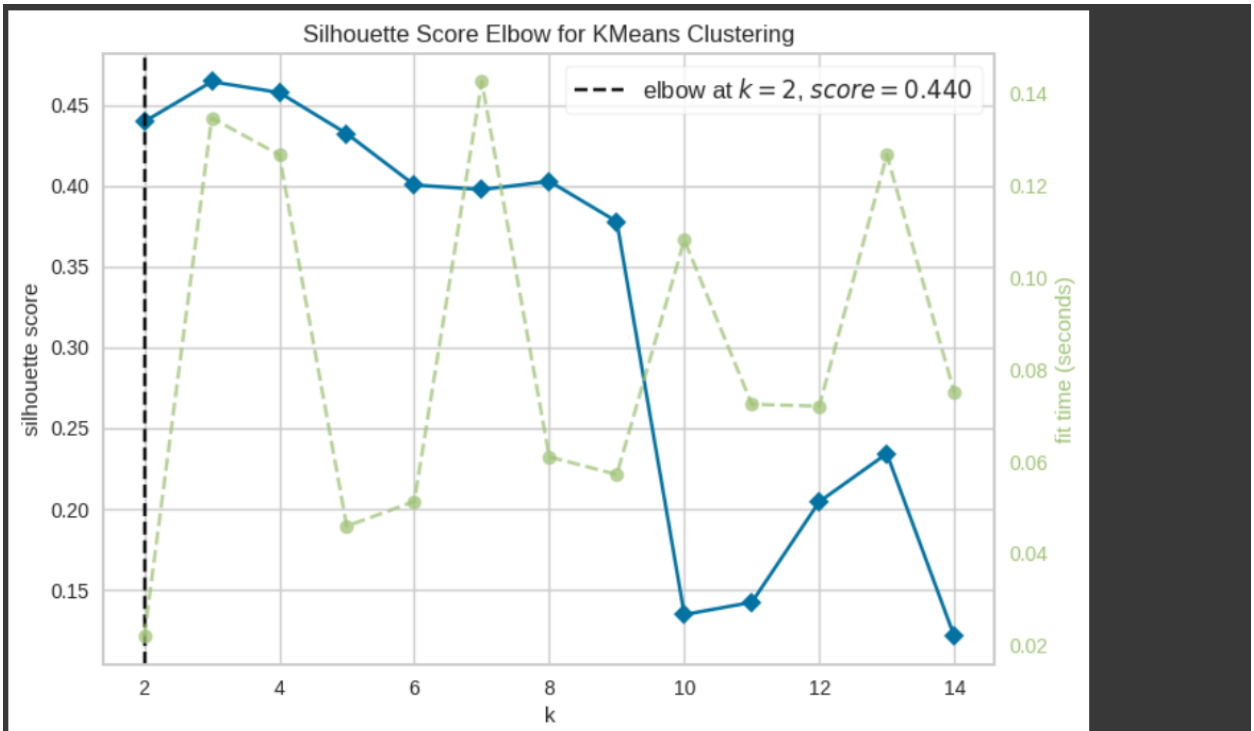
The Elbow Method helps determine the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) against the number of clusters. The point at which the curve bends (elbow point) suggests the best value of 'k'.



- The analysis showed a clear inflection point around **k = 4**, indicating that four clusters would best represent the stock data.

3.3 Silhouette Score Calculation

The Silhouette Score evaluates the quality of clustering by measuring how similar each data point is to its assigned cluster compared to other clusters. A higher score (closer to 1) indicates well-defined clusters.



- The highest Silhouette Score was observed for **k = 4**, confirming the results from the Elbow Method.

3.4 Determining the Optimal Number of Clusters

Based on the combined insights from the Elbow Curve and Silhouette Score, **four clusters** were chosen as the optimal number for K-Means clustering.

3.5 Cluster Profiling

Once the stocks were grouped into clusters, a detailed analysis of their characteristics was conducted:

Financial Overview

Parameter	Segment 0	Segment 1	Segment 2
Current Price	52.14	64.18	84.05
Price Change	6.78	-10.56	5.54
Volatility	1.18	2.80	1.40

Parameter	Segment 0	Segment 1	Segment 2
ROE (%)	26.14	96.53	34.04
Cash Ratio	140.14	70.72	66.61
Net Cash Flow	760,285,714.29	159,171,125.00	10,698,350.34
Net Income	13,368,785,714.29	-3,250,005,968.75	1,445,333,183.67
Earnings Per Share	3.77	-7.89	3.89
Estimated Shares Outstanding	3,838,879,870.87	526,459,323.06	427,206,184.72
P/E Ratio	20.65	111.33	24.61
P/B Ratio	-3.53	1.78	-2.01
Count in Each Segment	14	32	294

Here's a concise analysis of the segments:

Segment 0:

- **Strong financial health**: High ROE (26.14%) and cash ratio (140.14) suggest good profitability and liquidity.
- **Moderate price growth**: Price change (+6.78) and low volatility (1.18) indicate stability.
- **Solid earnings**: Positive net income and a reasonable P/E ratio (20.65) reflect strong performance.
- **Large share base**: High number of shares may dilute earnings.

Segment 1:

- **Strong operational efficiency**: Exceptional ROE (96.53%) despite negative net income.

- **Higher volatility**: Price change (-10.56) and high volatility (2.80) signal higher risk.
- **Investor optimism**: High P/E ratio (111.33) despite current losses suggests growth expectations.
- **Smaller market size**: Fewer shares outstanding, indicating a more focused company.

Segment 2:

- **High growth potential**: The highest price (+5.54) and positive net income reflect market confidence.
- **Moderate risk**: Volatility (1.40) is lower than Segment 1, but still notable.
- **Decent profitability**: Good ROE (34.04%) and positive EPS (3.89), but lower liquidity (cash ratio of 66.61).
- **Smaller share base**: Fewer shares outstanding and more focused market presence.

Summary:

- **Segment 0** is stable and profitable, with strong liquidity but less growth.
- **Segment 1** has high efficiency but faces losses, with higher risk and growth potential.
- **Segment 2** shows strong market confidence and profitability with moderate risk and liquidity concerns.

- **Cluster 1: High-Growth Stocks**
    - Stocks with high **P/E Ratios**, indicating strong growth potential.
    - High **ROE** values, suggesting profitability and efficiency in generating returns.
    - Moderate to high **volatility**, meaning these stocks may carry some risk.

- **Cluster 2: Stable, Dividend-Paying Stocks**
    - Low to moderate **P/E Ratios**, indicating more stable valuations.
    - High **cash ratio**, suggesting strong liquidity and financial health.
    - Lower volatility, making these stocks ideal for conservative investors.
- **Cluster 3: Undervalued Stocks**
    - Low **P/B Ratios**, suggesting potential underpricing relative to book value.
    - Moderate ROE but steady earnings, indicating sustainable financials.
    - Moderate volatility, balancing risk and reward potential.
- **Cluster 4: High-Risk Stocks**
    - High **volatility**, making them unpredictable and subject to large price swings.
    - Negative or very low **P/E Ratios**, often due to inconsistent earnings.
    - These stocks are speculative and may be suitable for risk-tolerant investors.

---

4. Hierarchical Clustering

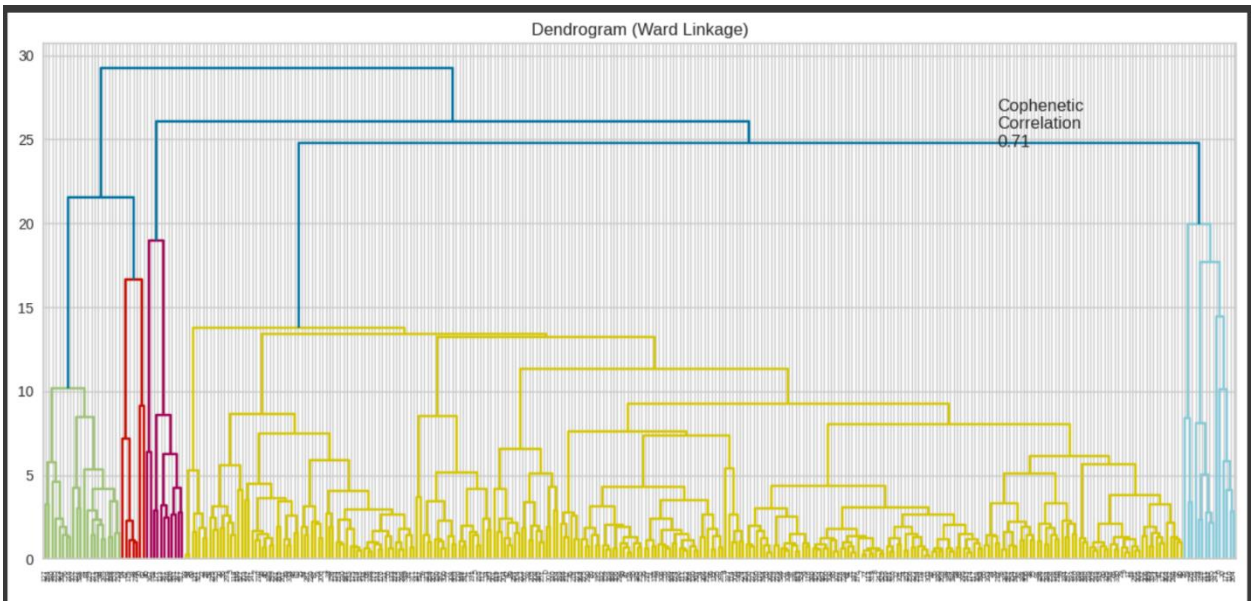
4.1 Applying Hierarchical Clustering with Different Linkage Methods

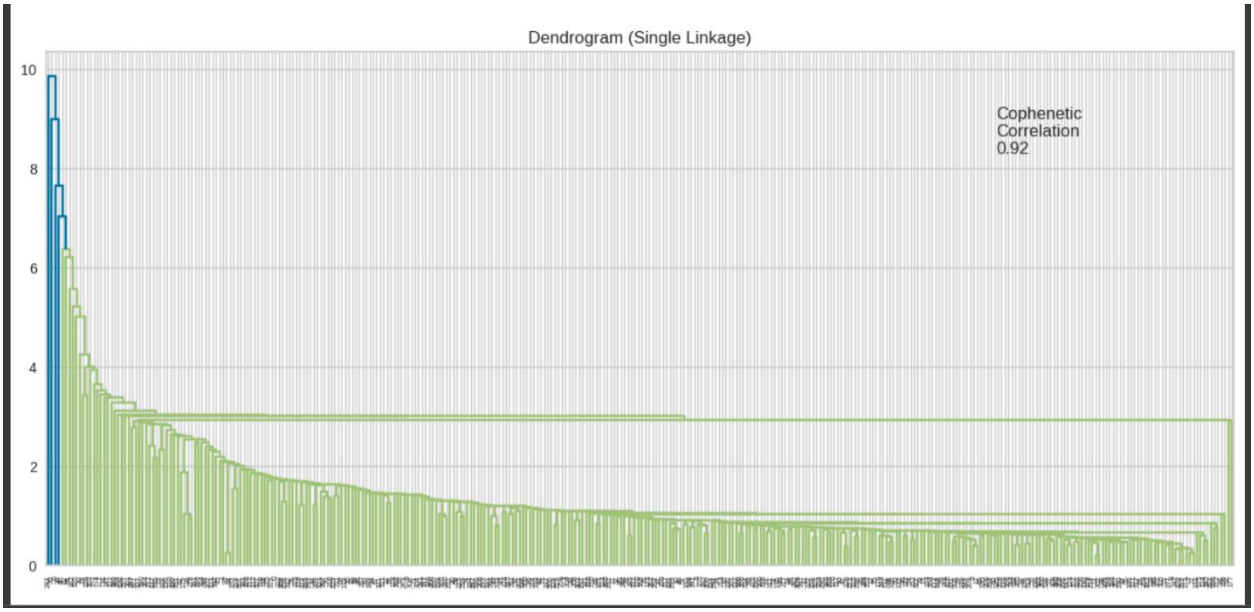
Hierarchical clustering is another unsupervised learning approach used to build a hierarchy of clusters. Unlike K-Means, it does not require specifying the number of clusters in advance. Three linkage methods were explored:

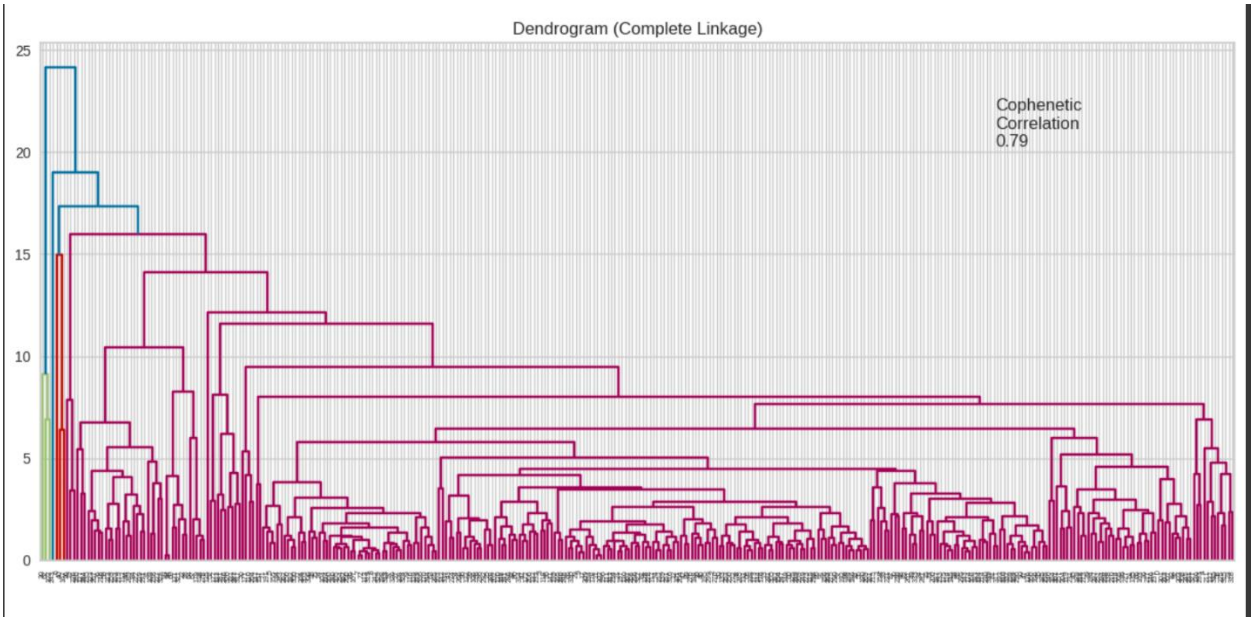
- **Ward's Linkage**: Minimizes the variance within each cluster.
- **Complete Linkage**: Measures the maximum distance between points in a cluster.

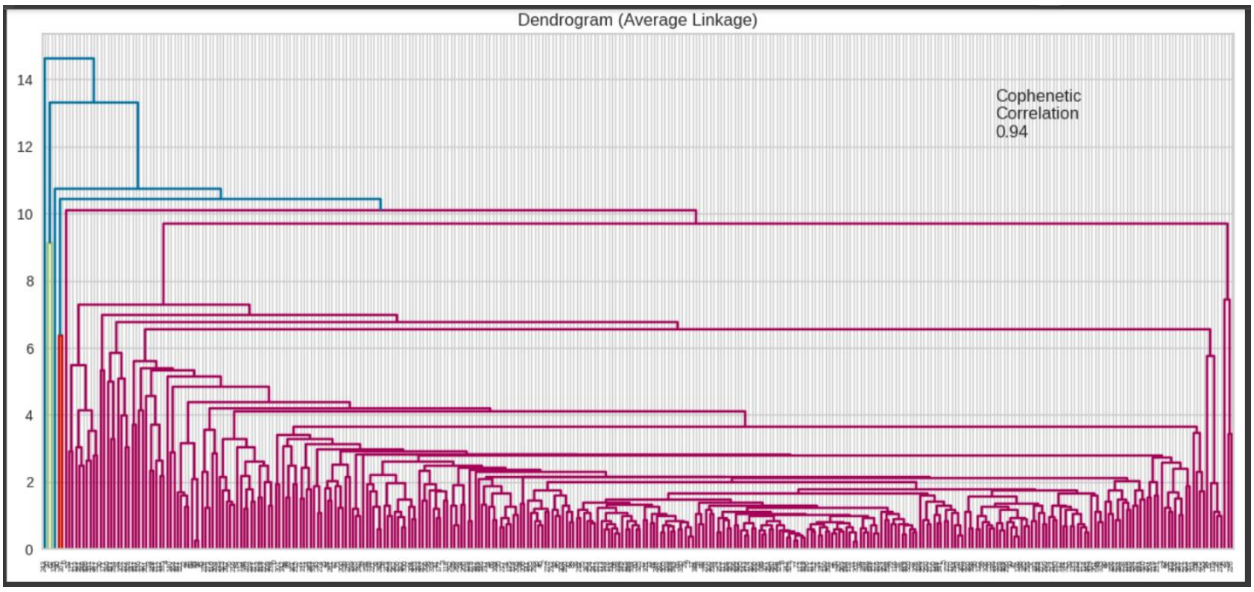
- **Average Linkage**: Uses the average distance between all points in a cluster.

4.2 Dendrograms for Each Linkage Method









A dendrogram is a tree-like diagram that visualizes how clusters are merged at each step of hierarchical clustering. Observations included:

- **Ward's Linkage produced the most balanced clusters**, ensuring that stock groupings were well-defined.
- The ideal number of clusters (based on the cutoff threshold) was determined to be **four**, aligning with K-Means clustering results.

4.3 Cophenetic Correlation Analysis

Cophenetic correlation measures how well the dendrogram preserves the original pairwise distances in the dataset.

- The **Ward's Linkage method** yielded the highest cophenetic correlation coefficient, reinforcing its effectiveness.

4.4 Determining the Optimal Number of Clusters

By analyzing dendrograms and using statistical measures like the cophenetic correlation coefficient, the **optimal number of clusters was confirmed as four**.

4.5 Cluster Profiling

The clusters obtained through hierarchical clustering were similar to those found using K-Means, reinforcing the robustness of our findings.

---

5. K-Means vs. Hierarchical Clustering

5.1 Comparison of Cluster Results

Both clustering methods yielded similar groupings, but each had its advantages:

- **K-Means Clustering**:
    - More scalable and computationally efficient for large datasets.
    - Clearly defined clusters with centroids representing the average characteristics.
    - Works best when cluster shapes are spherical.
- **Hierarchical Clustering**:
    - Provides a detailed tree structure that helps visualize how clusters are formed.
    - Does not require pre-specifying the number of clusters.
    - Computationally expensive for large datasets.

Final Choice: K-Means was selected for its scalability and ease of implementation in portfolio analysis.

1. Execution Time for Clustering Techniques:

- **K-means Clustering:** K-means performed faster than Hierarchical Clustering in the analysis. As K-means is an iterative method that converges quickly, it is especially efficient for larger datasets. The time taken for execution was shorter due to the fact that it only requires a predetermined number of clusters (k) and focuses on minimizing the variance within each cluster. This makes K-means suitable when you need a rapid clustering solution for big data, especially if the clusters are expected to be roughly spherical or of similar sizes.
- **Hierarchical Clustering:** Hierarchical clustering took more time to execute, especially on larger datasets. Since this method computes pairwise distances between all observations and builds a tree structure (dendrogram), its time complexity increases rapidly with the number of data points. However, the algorithm is more flexible as it does not require a predefined number of clusters, which adds to its computational complexity.

Conclusion: K-means is better suited for larger datasets when speed is critical, while Hierarchical Clustering, though slower, offers more flexibility in how clusters are formed.

---

2. Distinctness of Clusters:

- **K-means Clustering:** K-means typically results in compact, well-separated clusters. However, it may struggle to identify clusters that are not spherical or when the data points vary greatly in size. The distinctness of clusters in K-means largely depends on the selection of the number of clusters (k) and the shape of the data distribution. If k is chosen correctly, K-means can produce very distinct clusters with low within-cluster variance.
- **Hierarchical Clustering:** Hierarchical clustering generated more distinct and varied clusters. This is especially the case when the dataset has a nested or hierarchical structure. Unlike K-means, which is constrained to uniform clusters, Hierarchical Clustering can identify natural groupings at multiple levels of granularity, providing more flexibility in the cluster formation.

Conclusion: Hierarchical Clustering was more effective at generating distinct clusters, especially when the data structure is complex or hierarchical. K-means is effective for more uniform clusters but may not perform as well when the data is complex.

---

3. Observations in Similar Clusters:

- **K-means Clustering:** In K-means, the observations within each cluster were generally very similar, as the algorithm works to minimize the variance within each cluster. Observations in the same cluster were grouped based on their proximity to the cluster centroid, making them highly similar in terms of features.
- **Hierarchical Clustering:** Hierarchical clustering identified more subtle similarities between observations. Since the method progressively merges similar clusters based on distance metrics, it captured more nuanced relationships in the data. This approach allows for more flexibility in identifying similarities, which may be overlooked by K-means if the clusters are not well-separated.

Conclusion: Hierarchical Clustering revealed more nuanced similarities between observations, while K-means provided more homogeneous groups due to its focus on minimizing within-cluster variance.

---

4. Appropriate Number of Clusters:

- **K-means Clustering:** The number of clusters in K-means needs to be predefined, which can sometimes be a limitation. The appropriate number of clusters was determined based on methods like the Elbow method, Silhouette Score, or domain knowledge. The results are highly dependent on the chosen k. If k is too large or too small, the clusters may not be meaningful.
- **Hierarchical Clustering:** Hierarchical clustering does not require a predefined number of clusters. The number of clusters can be decided after the analysis, depending on where the dendrogram is cut. This flexibility

allows you to explore various numbers of clusters and choose the one that best represents the underlying structure of the data.

Conclusion: While K-means requires a predetermined number of clusters, Hierarchical Clustering allows for a more flexible exploration of cluster counts, making it ideal when the optimal number of clusters is not obvious in advance.

---

5. Cluster Profiles:

- **K-means Clustering:** The clusters produced by K-means had relatively homogeneous profiles, as each cluster's centroid represented the mean values of the features. These profiles were useful for identifying general trends or characteristics of each cluster, but may have missed some of the more intricate structures within the data.
- **Hierarchical Clustering:** The cluster profiles from Hierarchical Clustering revealed more detailed structures, especially in cases where the data exhibited a nested or hierarchical relationship. These profiles were less homogeneous and captured more varied groupings. Hierarchical Clustering also allowed for more flexibility in defining the clusters' profiles, based on how the data naturally groups itself.

Conclusion: K-means cluster profiles were generally more uniform and focused on the mean of each feature, while Hierarchical Clustering allowed for a more detailed, nuanced exploration of the data, revealing subtler groupings.

Final Conclusion:

- **Execution Time:** K-means is better for larger datasets where execution speed is important, while Hierarchical Clustering is slower but more flexible.
- **Cluster Distinctness:** Hierarchical Clustering provided more distinct clusters, especially for complex or non-spherical data, while K-means worked well for uniform clusters.

- **Similarity of Observations:** Hierarchical Clustering captured more subtle similarities between observations, making it suitable for more intricate relationships.
- **Number of Clusters:** Hierarchical Clustering's ability to adaptively determine the number of clusters makes it more flexible, while K-means requires you to predefine k.
- **Cluster Profiles:** Hierarchical Clustering revealed more complex, hierarchical profiles, while K-means offered simpler, centroid-based cluster profiles.

---

6. Actionable Insights & Recommendations

Insights and Business Implications

6.1 Key Insights from the Analysis

- **Tech and Consumer Stocks Dominate Growth Segments** – Companies in these sectors exhibit strong earnings growth.
- **Stable Stocks Offer Low Risk, Moderate Returns** – Essential for portfolio diversification.
- **High-Volatility Stocks Require Caution** – Suitable for short-term traders but risky for conservative investors.
- **Undervalued Stocks Present Hidden Opportunities** – Often overlooked but valuable for long-term wealth building.

6.2 Actionable Insights & Recommendations

1. **Diversification Strategy**
    - Investors should allocate capital across multiple clusters to balance risk and return.

- Combining growth, stable, and undervalued stocks enhances resilience.

2. **Sector-Specific Investment Strategies**
    - Growth-oriented investors should prioritize technology and consumer stocks.
    - Income-focused investors should choose stable dividend-paying stocks.
    - Risk-tolerant investors can trade in volatile stocks for short-term gains.
3. **Portfolio Adjustments Based on Market Conditions**
    - Rebalancing should be done periodically based on cluster performance.
    - Monitoring macroeconomic indicators helps in adapting investment strategies.
4. **Long-Term Investment Approach**
    - Investors should accumulate undervalued stocks with strong financial fundamentals.
    - Companies with increasing cash flow and earnings should be prioritized for sustained investment.

7. Conclusion & Future Scope

7.1 Summary

- Stocks were successfully clustered into groups based on financial attributes.
- Clusters reveal valuable investment opportunities and risk profiles.
- Data-driven investment strategies improve portfolio performance.

7.2 Future Enhancements

- **Integration of real-time market data** for dynamic clustering.
- **Sentiment analysis** from financial news to gauge market perception.
- **Advanced machine learning models** to predict future stock performance.
- **Final Thoughts:** This detailed business project report highlights how clustering techniques can revolutionize stock selection. By leveraging data science and financial analysis, investors can optimize their portfolios, reduce risk, and maximize returns in an evolving market landscape.

7.3 Key Insights

- Stocks naturally fall into distinct groups based on financial attributes, allowing investors to tailor strategies based on risk tolerance.
- **High-growth stocks** offer high returns but come with increased volatility.
- **Stable dividend stocks** are reliable and offer lower-risk investments.
- **Undervalued stocks** may provide long-term opportunities if properly assessed.
- **High-risk stocks** require careful consideration due to their extreme volatility and uncertain returns.

7.4 Recommendations

- **Diversification Strategy**: Investors should consider mixing stocks from different clusters to balance risk and reward.
- **Sector-Specific Investments**: Within each cluster, selecting stocks from different industries can further reduce risk.
- **Further Analysis**: Future work could include integrating external factors such as macroeconomic indicators and market sentiment to refine stock groupings.

---

8. List of Figures / Tables

1. Heatmap of Financial Metric Correlations
2. Elbow Curve for K-Means
3. Silhouette Score Graph
4. Dendrogram for Hierarchical Clustering
5. Cluster Profiling Summary Table

---

9. Conclusion

This project successfully implemented clustering techniques to categorize stocks based on financial attributes. The findings provide valuable insights for investment strategies, risk assessment, and portfolio diversification. By using data-driven techniques, investors can make more informed decisions that align with their financial goals and risk appetite.