Unlocking Investment Insights: Clustering Analysis for Portfolio Optimization

Unsupervised Learning

Pam Lozano

March 2, 2024

Contents / Agenda

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

Objective

Develop and optimize classification models using machine learning techniques on the provided sensor data to accurately predict wind turbine generator failures. The goal is to minimize maintenance costs by identifying failures early, differentiating between true positives (repair costs), false negatives (replacement costs), and false positives (inspection costs).

Key Focus Areas:

1. Classification Model Development: Implement and fine-tune various classification algorithms to effectively predict wind turbine generator failures based on the 40 predictor variables.
2. Cost Optimization: Prioritize model performance metrics that minimize overall maintenance costs, considering the significant difference in costs between repairing, replacing, and inspecting generators.
3. Interpretation of Model Outputs: Analyze and interpret model predictions to understand the trade-offs between true positives, false negatives, and false positives in the context of minimizing maintenance expenses for wind turbine generators.

Data Information

The data provided is of stock prices and some financial indicators like ROE, earnings per share, P/E ratio, etc.

.

Ticker Symbol	An abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market
Company	Name of the company
GICS Sector	The specific economic sector assigned to a company by the Global Industry Classification Standard (GICS) that best defines its business operations
GICS Sub Industry	The specific sub-industry group assigned to a company by the Global Industry Classification Standard (GICS) that best defines its business operations
Current Price	Current stock price in dollars
Price Change	Percentage change in the stock price in 13 weeks
Volatility	Standard deviation of the stock price over the past 13 weeks
ROE	A measure of financial performance calculated by dividing net income by shareholders' equity (shareholders' equity is equal to a company's assets minus its debt)
Cash Ratio	The ratio of a company's total reserves of cash and cash equivalents to its total current liabilities
Net Cash Flow	The difference between a company's cash inflows and outflows (in dollars)
Net Income	Revenues minus expenses, interest, and taxes (in dollars)
Earnings Per Share	Company's net profit divided by the number of common shares it has outstanding (in dollars)
Estimated Shares Outstanding	Company's stock is currently held by all its shareholders
P/E Ratio	Ratio of the company's current stock price to the earnings per share
P/B Ratio	Ratio of the company's stock price per share by its book value per share (book value of a company is the net difference between that company's total assets and total liabilities)

Business Problem Overview

- **Investment Portfolio Diversification:** Creating a diversified investment portfolio is crucial for maximizing returns and mitigating risks in the stock market.
- **Cluster Analysis for Stock Selection:** Utilizing cluster analysis helps identify stocks with similar characteristics and low correlation, aiding in the selection of diverse stocks for investment portfolios.
- **Data-Driven Investment Strategies:** Leveraging data science techniques enables the identification of stocks exhibiting desirable financial metrics and performance indicators.
- **Global Industry Classification Standard (GICS):** Classifying companies based on GICS sectors and sub-industries provides insights into their business operations and sector-wise performance.
- **Financial Metrics Evaluation:** Analyzing financial metrics such as price change, volatility, ROE, P/E ratio, and cash flow aids in evaluating the performance and potential of individual stocks.
- **Optimizing Investment Decisions:** By grouping stocks into clusters based on their attributes, investors can optimize investment decisions, enhance portfolio diversification, and achieve financial goals effectively.

Utilizing cluster analysis and financial metrics evaluation, Trade&Ahead aims to construct diversified investment portfolios tailored to individual risk appetites and financial objectives, optimizing investment decisions and maximizing returns in the stock market.

Solution Approach

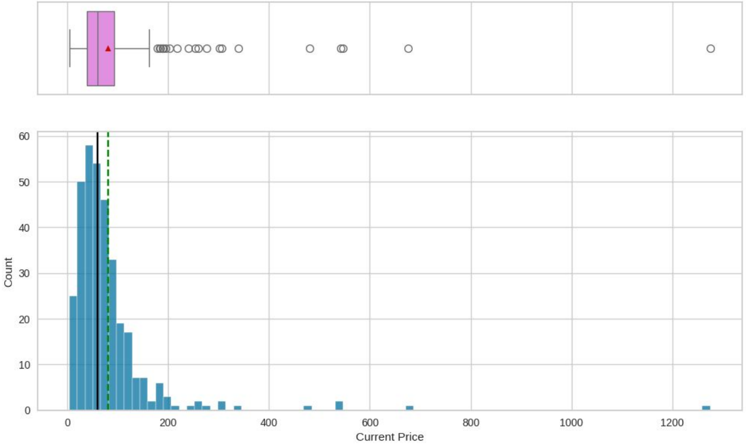
- **Data Acquisition:** Gather stock market data including ticker symbols, financial indicators, and price-related metrics from reliable sources such as financial databases or APIs.
- **Data Preprocessing:** Cleanse the data by handling missing values, removing outliers, and standardizing numerical features to ensure consistency and accuracy in analysis.
- **Feature Engineering:** Select relevant financial indicators and metrics conducive to clustering analysis, such as price change, volatility, ROE, and P/E ratio, to capture key aspects of stock performance.
- **Clustering Analysis:** Apply clustering algorithms like K-means or hierarchical clustering to group stocks based on their financial attributes, aiming to identify clusters of stocks exhibiting similar characteristics.
- **Evaluation and Interpretation:** Assess the quality of clusters using metrics like silhouette score or cluster cohesion, and interpret the clusters to gain insights into the underlying patterns and characteristics of each group of stocks.
- **Portfolio Construction:** Utilize the clustering results to construct diversified investment portfolios by selecting stocks from different clusters, aiming to optimize portfolio performance while managing risk effectively.

Data Background and Contents

- Our dataset comprises of 340 rows and 41 columns providing a comprehensive foundation for our analysis.
- The data types in our dataset are objects, integers and floats, giving the information available for our analysis.
- There are no missing or duplicate values in our dataset.
- The average price for the current price is 80.8623.
- The median price is 59.71, and the maximum number of employees is 1274.95.

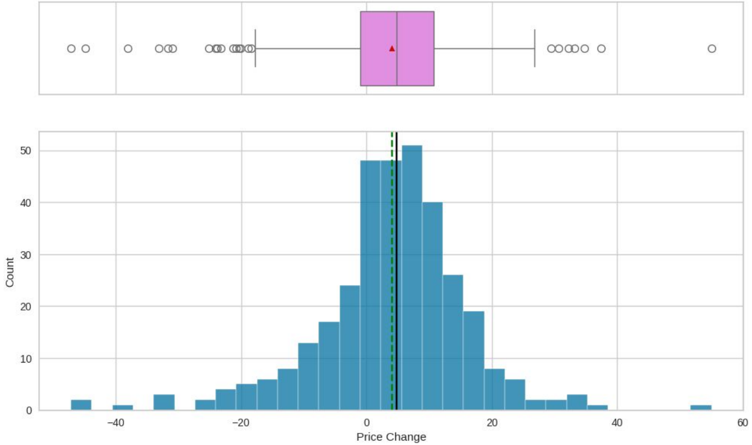
Exploratory Data Analysis

Univariate Analysis - Current Price



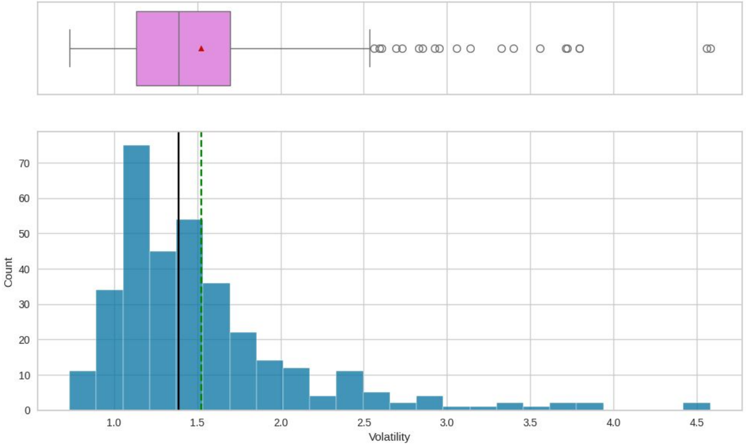
- Outliers are evident on the right of the distribution, suggesting potential extreme values.
- The distribution is right-skewed, with a median of approximately $50 dollars in current price.

Univariate Analysis - Price Change



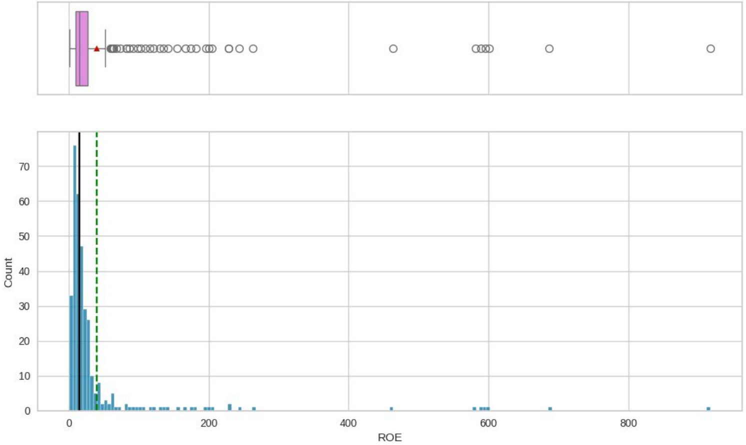
- The distribution is fairly normal, and outliers are present on both tails in price change.

Univariate Analysis - Volatility



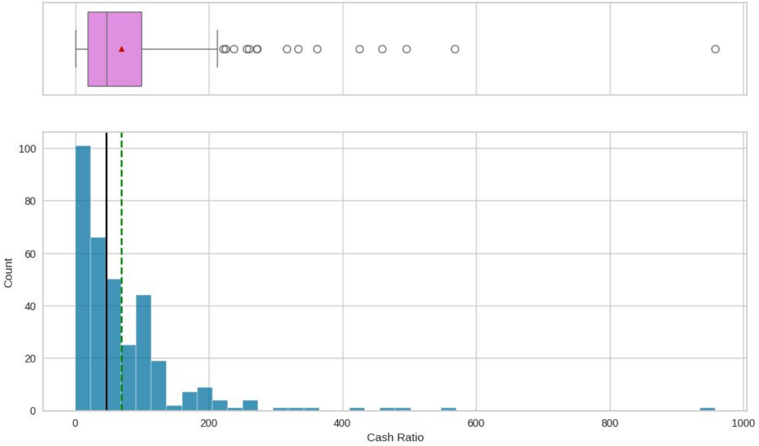
- Outliers are evident on the right of the distribution, suggesting potential extreme values.
- The distribution is right-skewed, with a median of approximately 1.4 in volatility.

Univariate Analysis - ROE



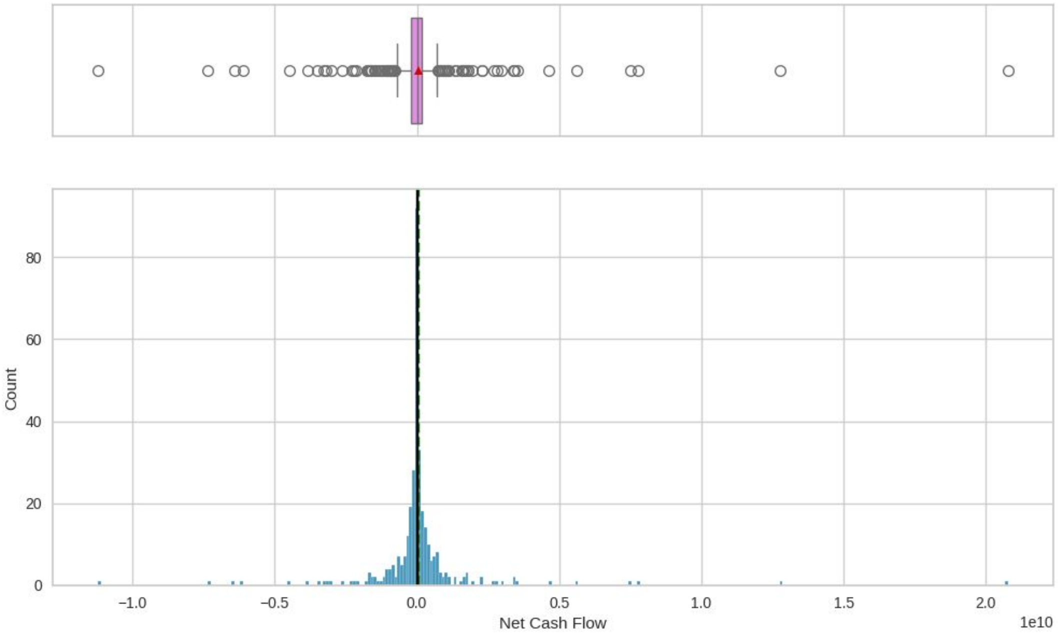
- Outliers are evident on the right tail, suggesting potential extreme values.
- This distribution is right skewed for ROE.

Univariate Analysis - Cash Ratio



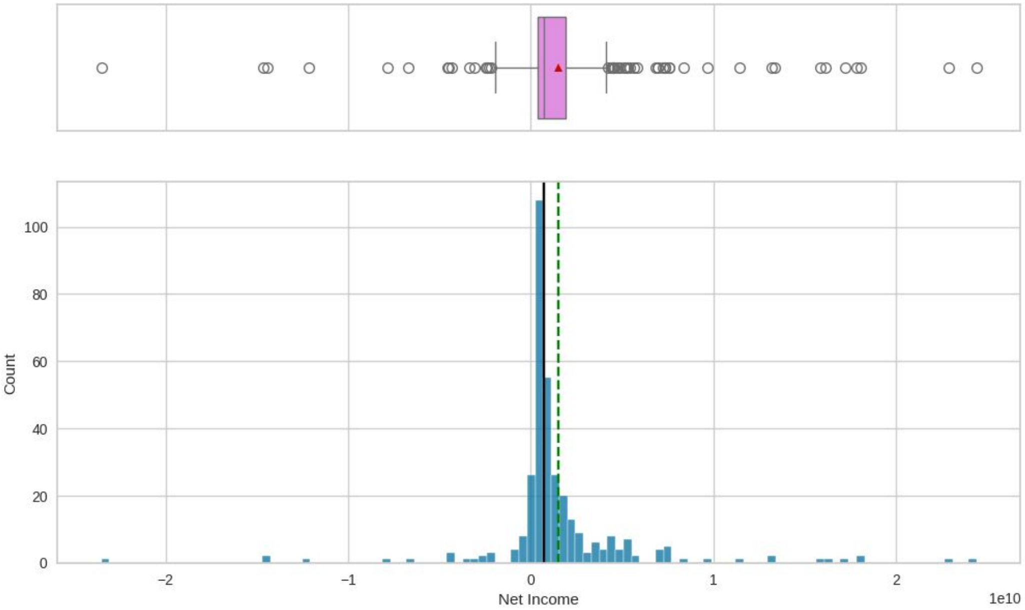
- Outliers are evident on the right tail, suggesting potential extreme values.
- This distribution is right skewed for Cash Ratio.

Univariate Analysis - Net Cash Flow



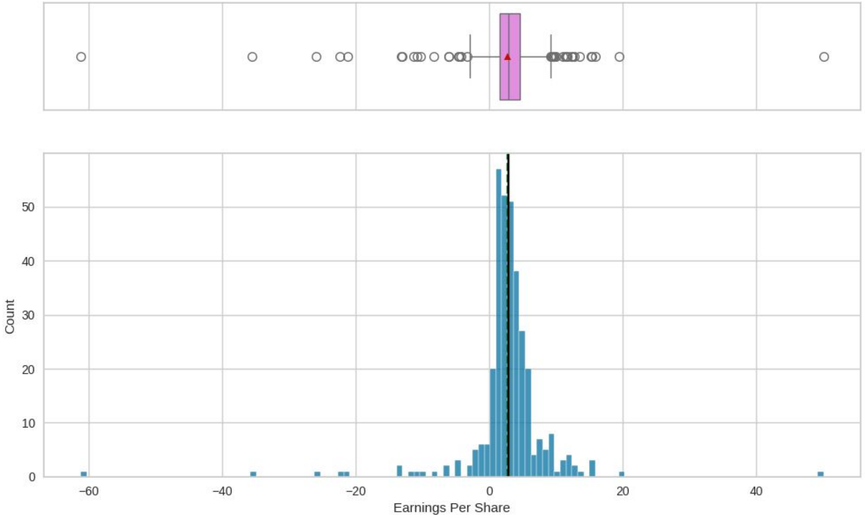
- The distribution is fairly normal, and outliers are present on both tails for net cash flow.

Univariate Analysis - Net Income



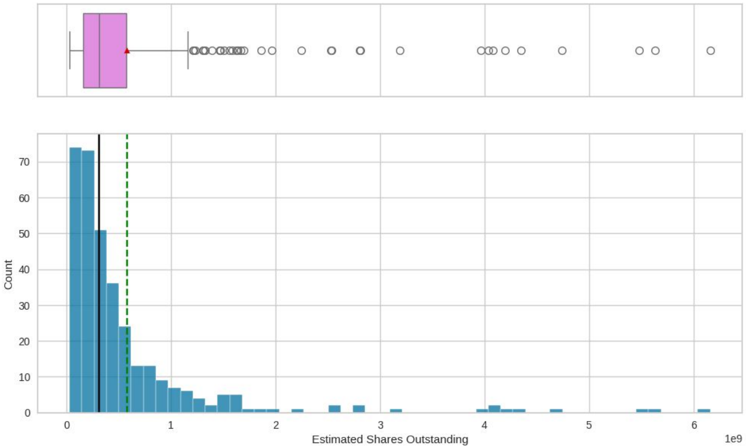
- The graph for Net Income reflects a fairly normal distribution with outliers on both tails.

Univariate Analysis - Earnings Per Share



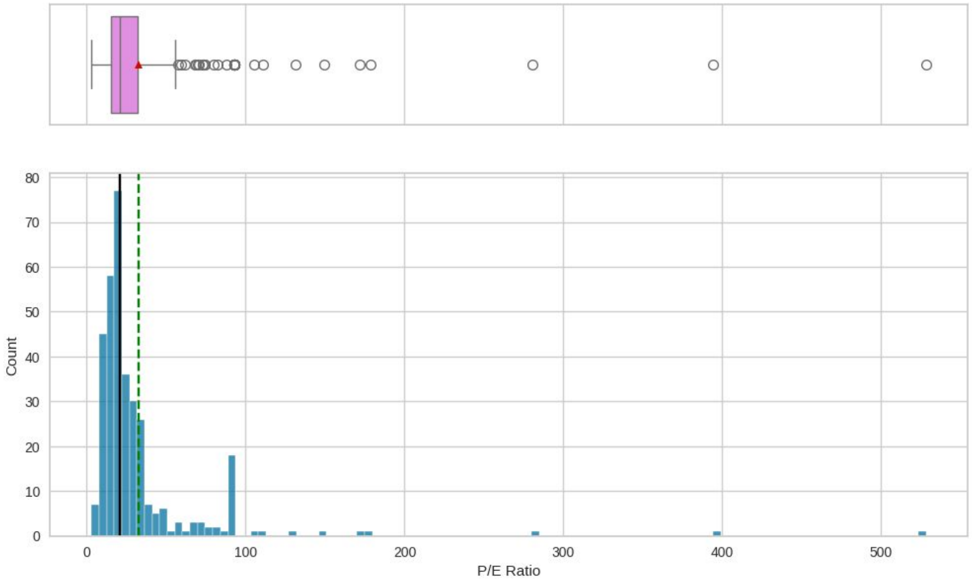
- The distribution is fairly normal, and outliers are present on both tails for Earnings per share.

Univariate Analysis - Estimated Shares Outstanding



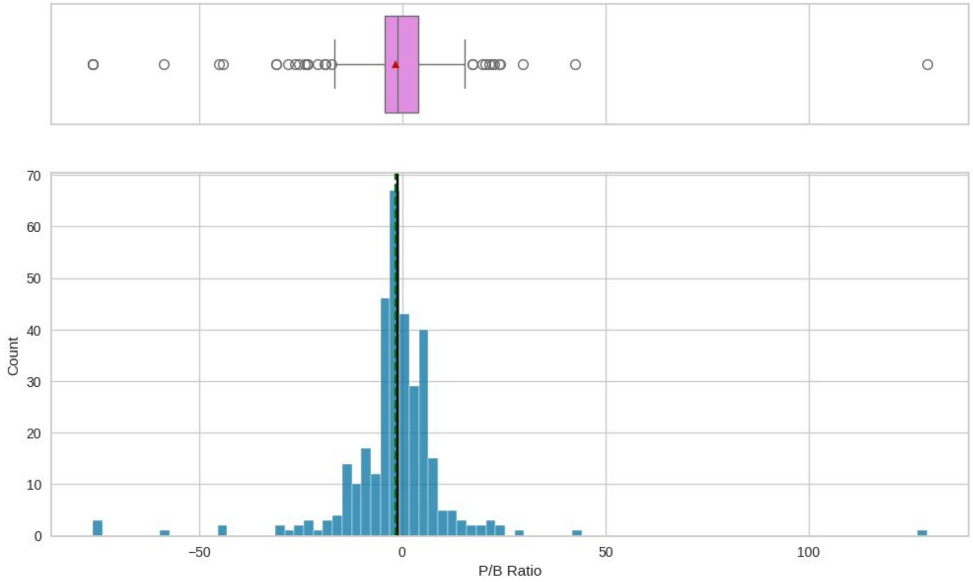
- Estimated Shares Outstanding reflects a right skewed distribution.
- Outliers are evidently present on the right tail.

Univariate Analysis - P/E Ratio



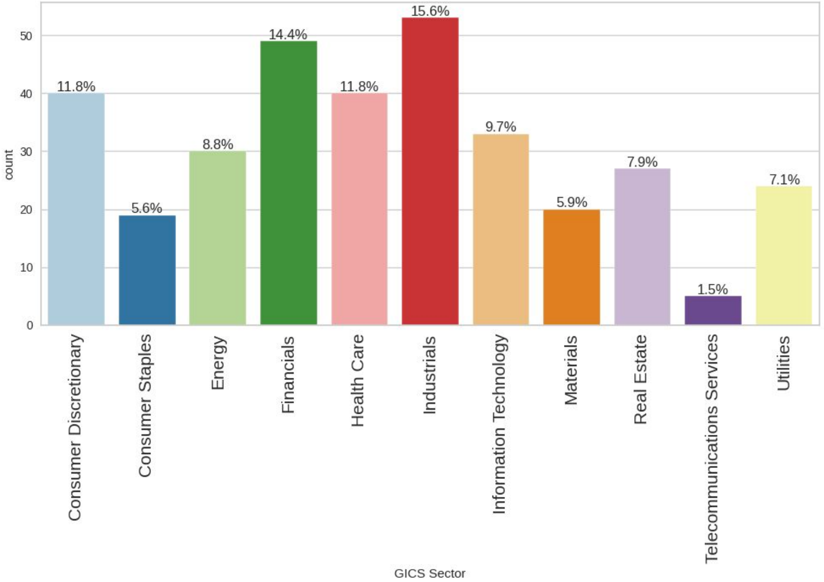
- The distribution for P/E Ratio is right skewed.
- Outliers are clearly present on the right tail.

Univariate Analysis - P/B Ratio



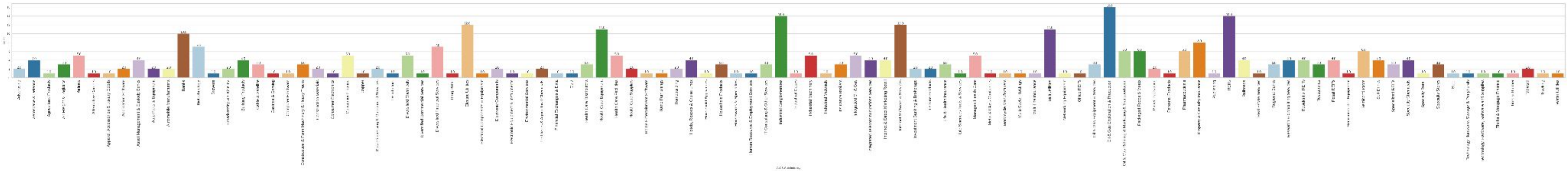
- The distribution for P/B Ratio is fairly normal.
- Outliers are evident on both tails.

Univariate Analysis - GICS Sector



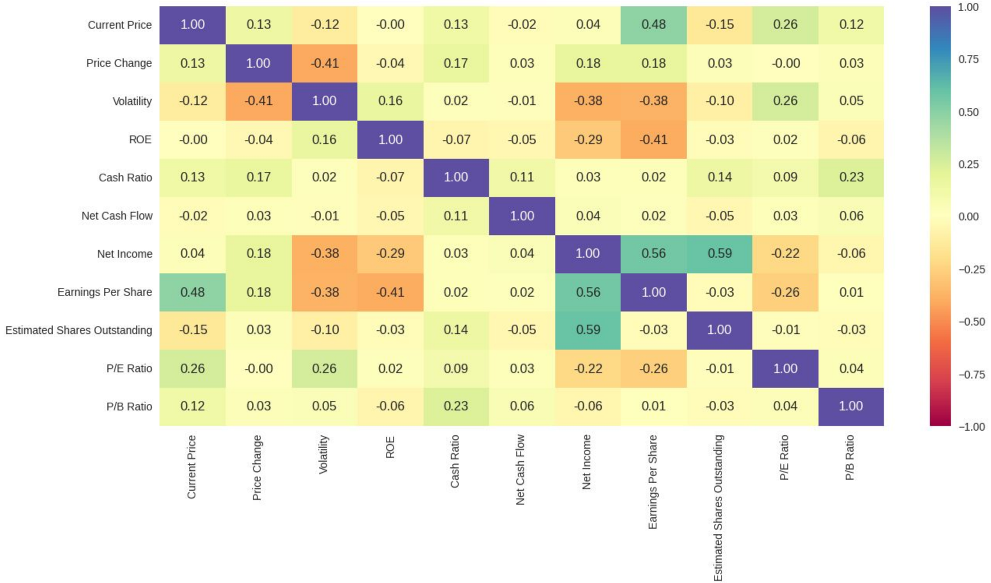
- The majority of the economic sector is in Industrials at 15. 6%.
- The least economic sector is in the Telecommunications sector at 1.5%

Univariate Analysis - GICS Sub Industry



- The top Sub Industries are in Oil & Gas Exploration & Production at 16.0.
- Reits and Industrial Conglomerates are the second sub industries at 14.0.
- Internet software and services and electric utilities come in third at 12.0.

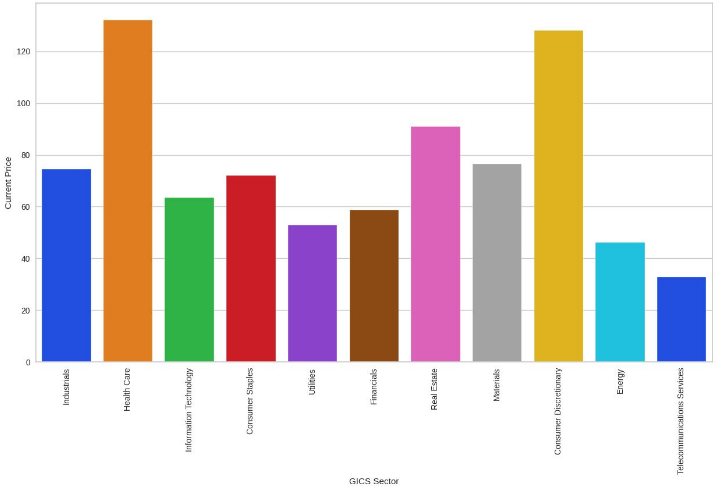
Bivariate Analysis



	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

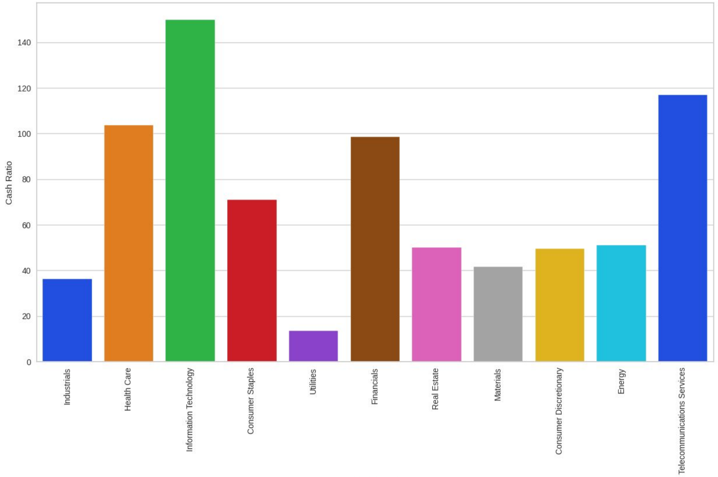
- Net Income is positively correlated with estimated shares, and earnings per share.
- ROE and Earnings per share are negatively correlated.
- Volatility is negatively correlated with Earnings per share and net income.

Bivariate Analysis - GICS Sector vs Current Price



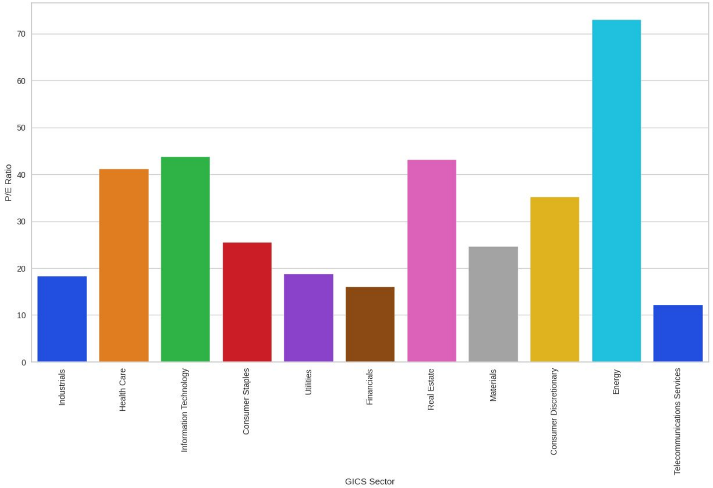
- The economic sectors that has seen the maximum increase on average are healthcare and consumer discretionary.

Bivariate Analysis - GICS Sector vs Cash Ratio



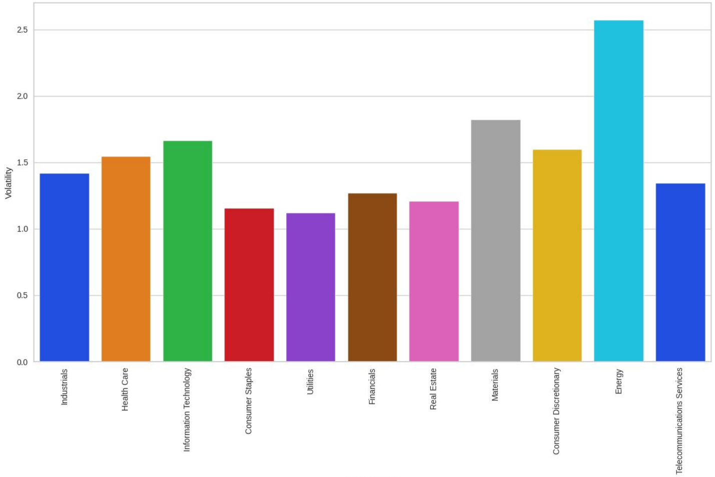
- The economic sectors that have the highest average cash ratio are Information Technology and Telecommunications Services.

Bivariate Analysis - GICS Sector vs P/E Ratio



- The economic sector with the highest P/E ratio is Energy.

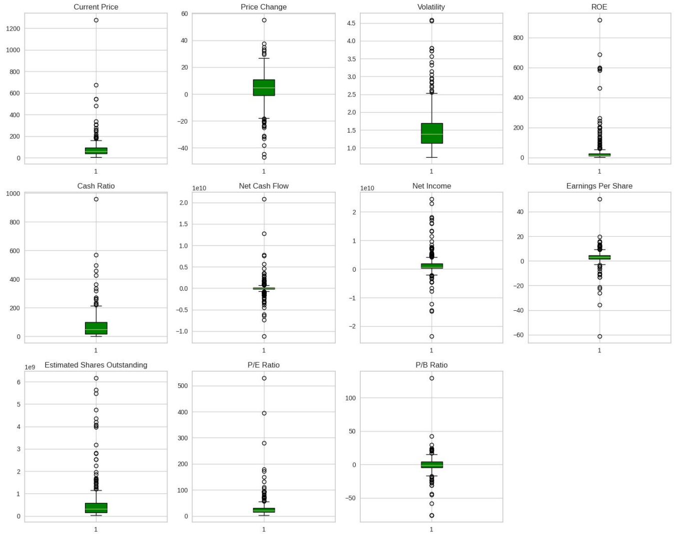
Bivariate Analysis - GICS Sector vs Volatility



- The economic sector with the highest volatility is Energy thus making it a riskier investment.

Data Preprocessing

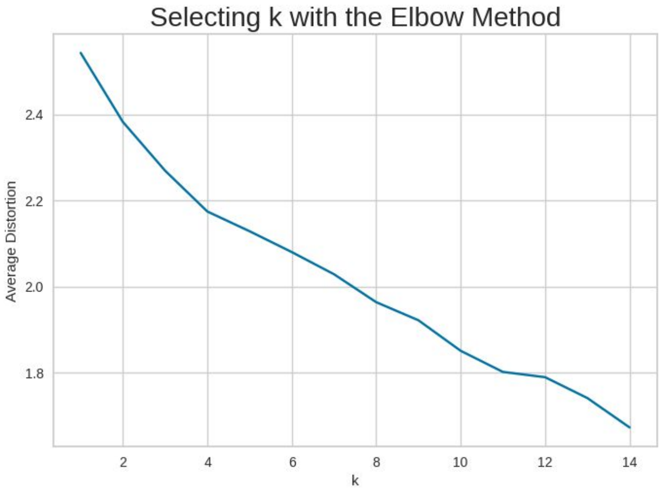
Data Preprocessing



- Several outliers have been identified in the data, but no treatment will be applied.

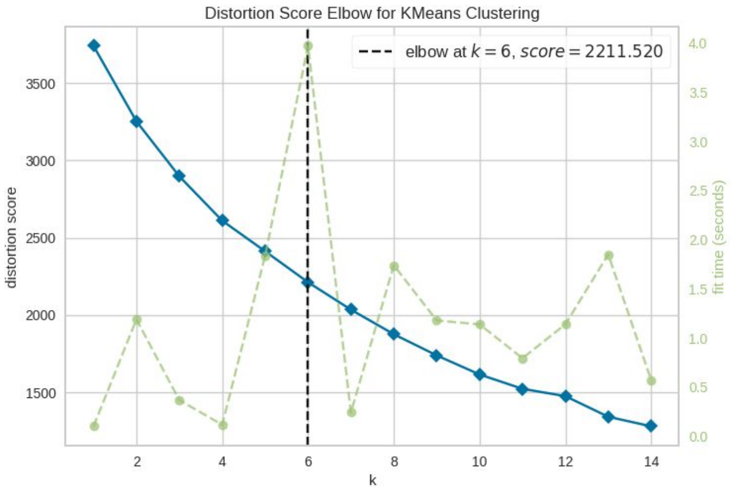
K-Means Clustering

Checking Elbow Plot



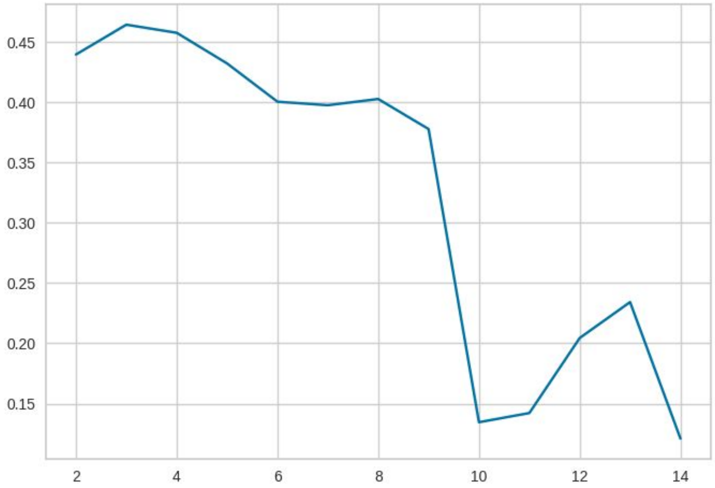
- The appropriate value of k from the elbow curve seems to be 4.

Distortion Score Elbow for K-Means Clustering



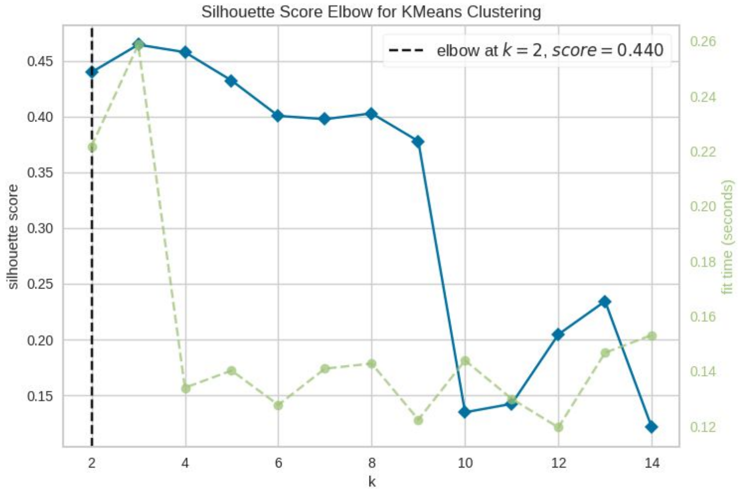
- The distortion score elbow shows 6 for the value of k, and the score is 2211.520.

Checking Silhouette Scores



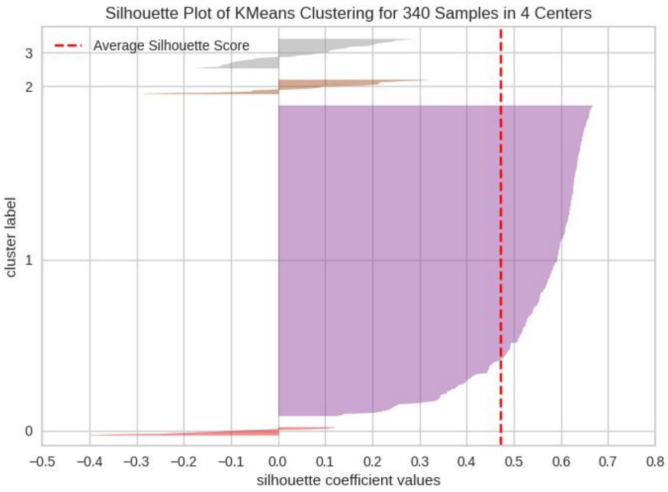
The silhouette scores show that a good value for k is 6.

Checking Silhouette Scores



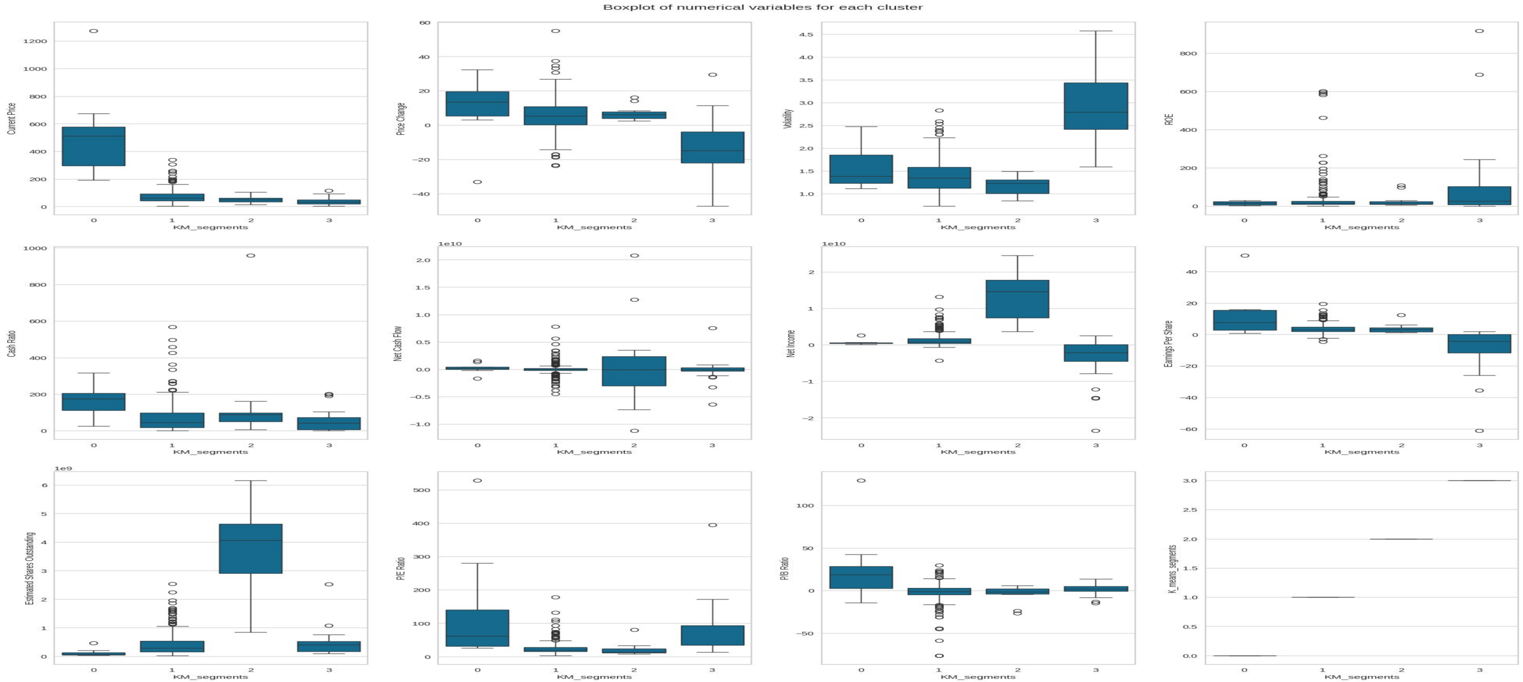
The silhouette score elbow show that a value for k at 2 has a score of .440.

Creating Final Model



- Based off the silhouette scores, the final K-means model has 4 clusters.

Cluster Profiling



- Net Cash Flow shows the clusters average about the same.

Insights

- Cluster 0:
  - Current price is high
  - Price change is high
  - Volatility is low to moderate
  - ROE is low
  - Cash Ratio is high
  - Net Cash Flow is moderate
  - Net Income is moderate
  - Earnings per share is high
  - Estimated shares outstanding is low
  - P/E Ratio is high
  - P/B Ratio is high

- Cluster 1:
  - Current price is moderate
  - Price change is low to moderate
  - Volatility is moderate
  - ROE is moderate
  - Cash Ratio is low
  - Net Cash Flow is low to moderate
  - Net Income is moderate
  - Earnings per share is moderate
  - Estimated shares outstanding is moderate
  - P/E Ratio is low
  - P/B Ratio is low

Insights

- Cluster 2:
  - Current price is low
  - Price change is moderate
  - Volatility is low
  - ROE is low
  - Cash Ratio is low
  - Net Cash Flow is high
  - Net Income is high
  - Earnings per share is moderate
  - Estimated shares outstanding is high
  - P/E Ratio is low
  - P/B Ratio is low

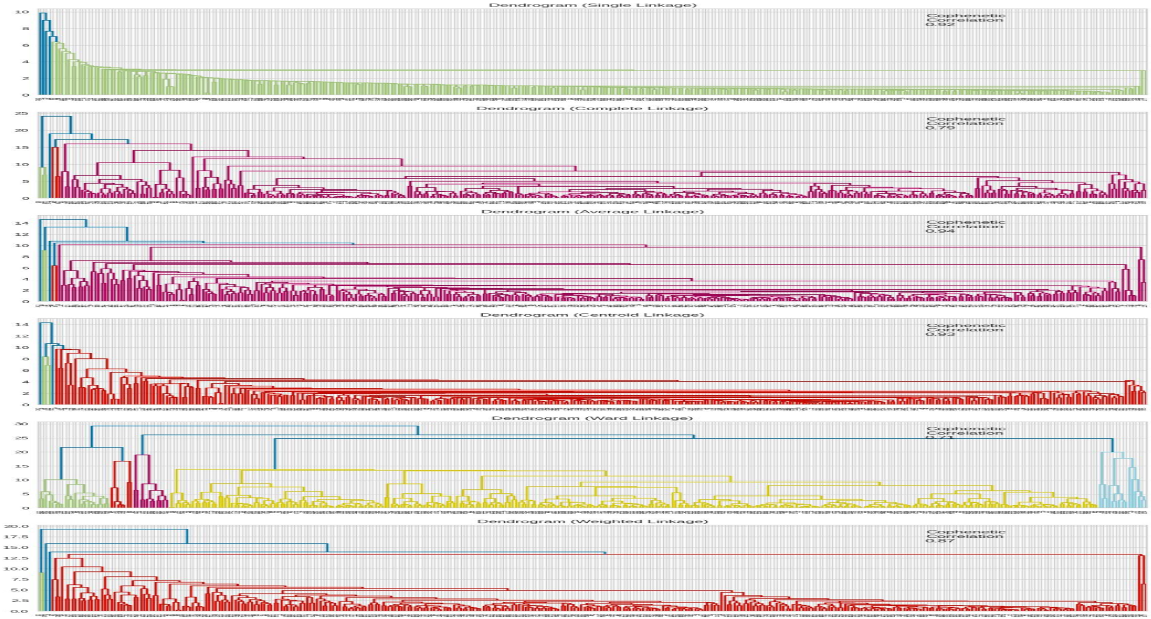
- Cluster 3:
  - Current price is low
  - Price change is low
  - Volatility is high
  - ROE is high
  - Cash Ratio is low
  - Net Cash Flow is low
  - Net Income is low
  - Earnings per share is low
  - Estimated shares outstanding is low
  - P/E Ratio is moderate
  - P/B Ratio is moderate

Hierarchical Clustering

Computing Cophenetic Correlation

- The highest cophenetic correlation is 0.9422540609560814, which is obtained with Euclidean distance and average linkage.

Checking Dendrograms

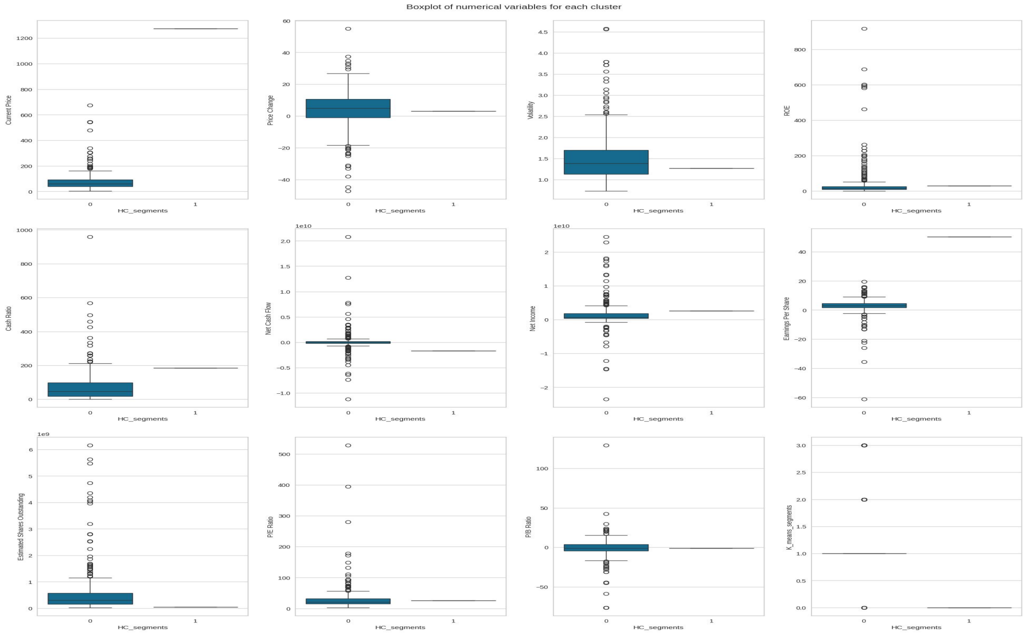


- Dendrogram for Ward Linkage shows clusters more clearly.
- The highest cophenetic correlation is the average linkage method.

Creating Final Model

- Out of all the dendrograms, it is clear that the Ward linkage dendrogram have clearer clusters.

Cluster Profiling



- Net Cash Flow, Net Income, and P/B Ratio show clusters averaging close to 0.

Insights

- Cluster 0:
    - Current price is low
    - Price change is high
    - Volatility is high
    - ROE is high
    - Cash Ratio is low
    - Net Cash Flow is high
    - Net Income is low
    - Earnings per share is low
    - Estimated shares outstanding is low
    - P/E Ratio is high
    - P/B Ratio is low

- Cluster 1:
    - Current price is high
    - Price change is low
    - Volatility is low
    - ROE is low
    - Cash Ratio is high
    - Net Cash Flow is low
    - Net Income is high
    - Earnings per share is high
    - Estimated shares outstanding is low
    - P/E Ratio is low
    - P/B Ratio is high

K-means vs Hierarchical Clustering

Creating Final Model

- From the dendrograms, Ward linkage was able to give clearer and distinct clusters.
- 4 would be the appropriate number of clusters from the dendrogram using the Ward linkage method.
- K-means Clustering was able to be a quicker technique providing more clusters.

Executive Summary

Executive Summary - Conclusions

- Price and Volatility: K-means clustering reveals distinct clusters characterized by varying levels of current price and volatility, with Cluster 0 exhibiting high prices and low to moderate volatility, while Cluster 3 shows low prices and high volatility.
- Financial Performance Metrics: Hierarchical clustering highlights clusters with contrasting financial performance metrics, such as Cluster 0's low net income and high P/E ratio compared to Cluster 1's high net income and low P/E ratio.
- Cash Management: K-means clustering identifies clusters with differing cash management strategies, with Cluster 0 exhibiting high cash ratios, indicating strong liquidity, while Cluster 3 shows low cash ratios, suggesting potential cash flow challenges.
- Investment Potential: Hierarchical clustering showcases clusters with varying investment potential, such as Cluster 0's high price change and high P/E ratio, indicating growth potential, compared to Cluster 1's low price change and low P/E ratio, suggesting stable investment opportunities.
- Earnings and Shares: Both clustering methods reveal clusters with contrasting earnings and shares characteristics, with Cluster 2 exhibiting low current prices and high net cash flow, indicating potential undervaluation and growth opportunities.
- Risk Assessment: The clustering results provide insights into risk assessment, with Cluster 3 in K-means clustering showing high volatility and moderate P/E ratio, indicating higher risk compared to other clusters identified in the analysis.

These conclusions provide valuable insights into the financial characteristics and investment potential of the clustered stocks, aiding stakeholders in making informed decisions regarding portfolio construction and investment strategies.

Executive Summary - Recommendations

- **Diversify Portfolio:** Incorporate stocks from multiple clusters to mitigate risk and optimize returns, leveraging the insights gained from cluster analysis.
- **Sector Allocation:** Allocate investments strategically across economic sectors, considering their performance and volatility, with emphasis on sectors exhibiting favorable trends.
- **Monitor Outliers:** Continuously monitor outliers in financial metrics to identify potential investment opportunities or risks and adjust portfolio allocations accordingly.
- **Explore Sub-Industries:** Explore investment opportunities in top-performing sub-industries such as Oil & Gas Exploration & Production, REITs, and Industrial Conglomerates.
- **Capitalize on Correlations:** Capitalize on positive correlations between key financial metrics while mitigating negative correlations, ensuring a balanced portfolio composition.
- **Stay Informed:** Stay informed about market trends, sector developments, and correlation dynamics to adapt investment strategies in response to changing market conditions.

The recommendations derived from cluster profiling and exploratory data analysis offer actionable strategies for constructing resilient investment portfolios, optimizing sector allocation, and capitalizing on correlation patterns to achieve long-term investment objectives in the dynamic stock market landscape.