UL Project - Guided

Thursday, 24th October 2024

Shamshia Taj

Table of Contents

List of Figures

Context

The stock market has consistently proven to be a good place to invest in and save for the future. There are a lot of compelling reasons to invest in stocks. It can help in fighting inflation, create wealth, and also provide some tax benefits. Good steady returns on investments over a long time can also grow a lot more than seems possible. Also, thanks to the power of compound interest, the earlier one starts investing, the larger the corpus one can have for retirement. Overall, investing in stocks can help meet life's financial aspirations.

It is important to maintain a diversified portfolio when investing in stocks to maximize earnings under any market condition. Having a diversified portfolio tends to yield higher returns and face lower risk by tempering potential losses when the market is down. It is often easy to get lost in a sea of financial metrics to analyze while determining the worth of a stock and doing the same for a multitude of stocks to identify the right picks for an individual can be a tedious task. By doing a cluster analysis, one can identify stocks that exhibit similar characteristics and ones that exhibit minimum correlation. This will help investors better analyze stocks across different market segments and help protect against risks that could make the portfolio vulnerable to losses.

Objective

Trade&Ahead is a financial consultancy firm that provides its customers with personalized investment strategies. They have hired you as a Data Scientist and provided you with data comprising stock prices and some financial indicators for a few companies listed under the New York Stock Exchange. They have assigned you the tasks of analyzing the data, grouping the stocks based on the attributes provided, and sharing insights about the characteristics of each group.

Data Description

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

Data Dictionary

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: The 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)

Exploratory Data Analysis

Univariate analysis

1. Current Price

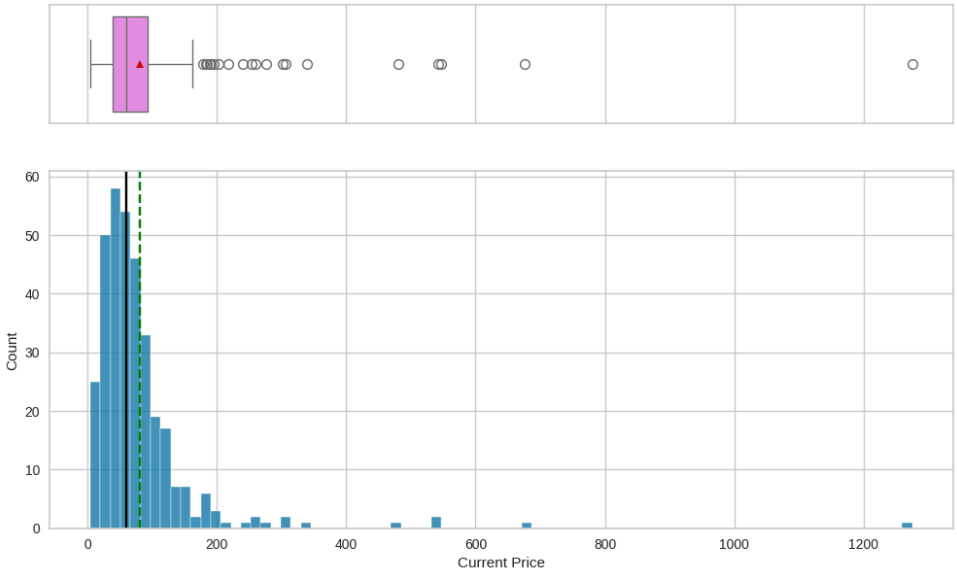

Fig 1

2. Price Change

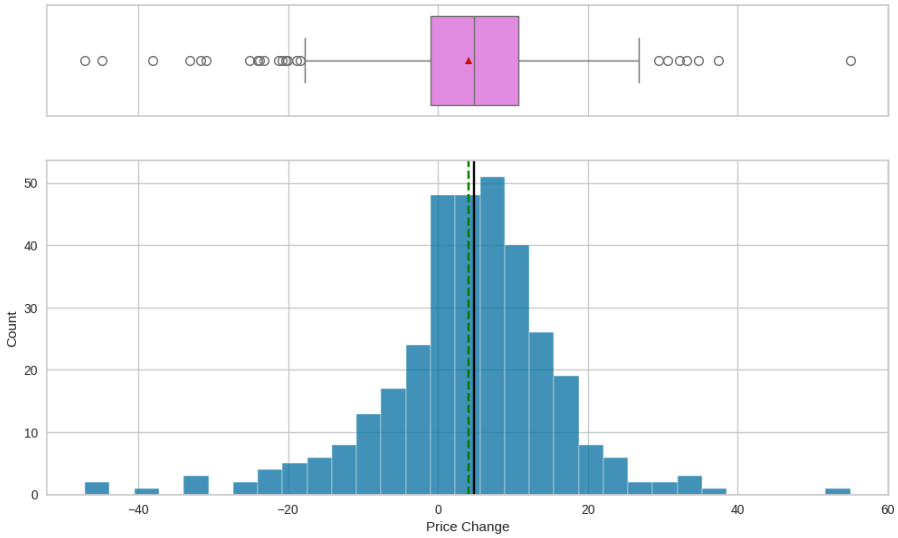

Fig 2

3. Volatility

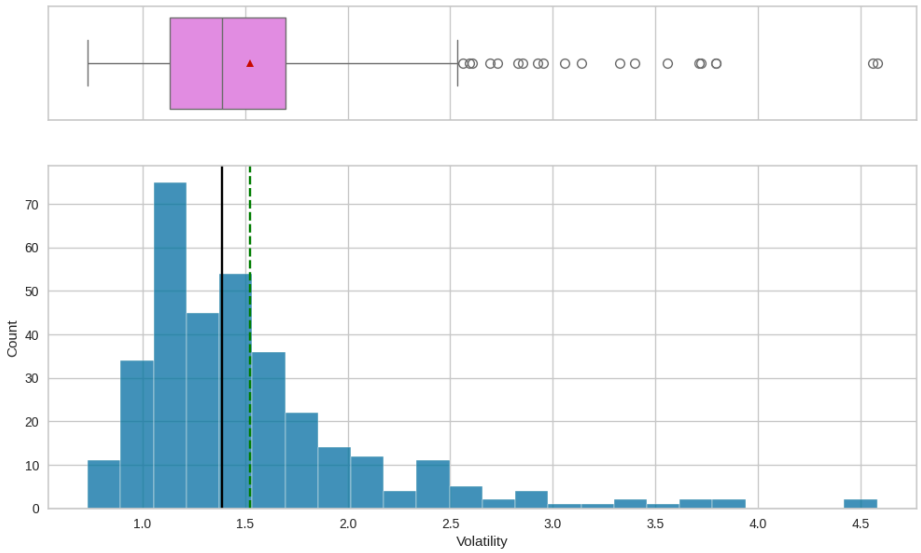

Fig 3

4. ROE

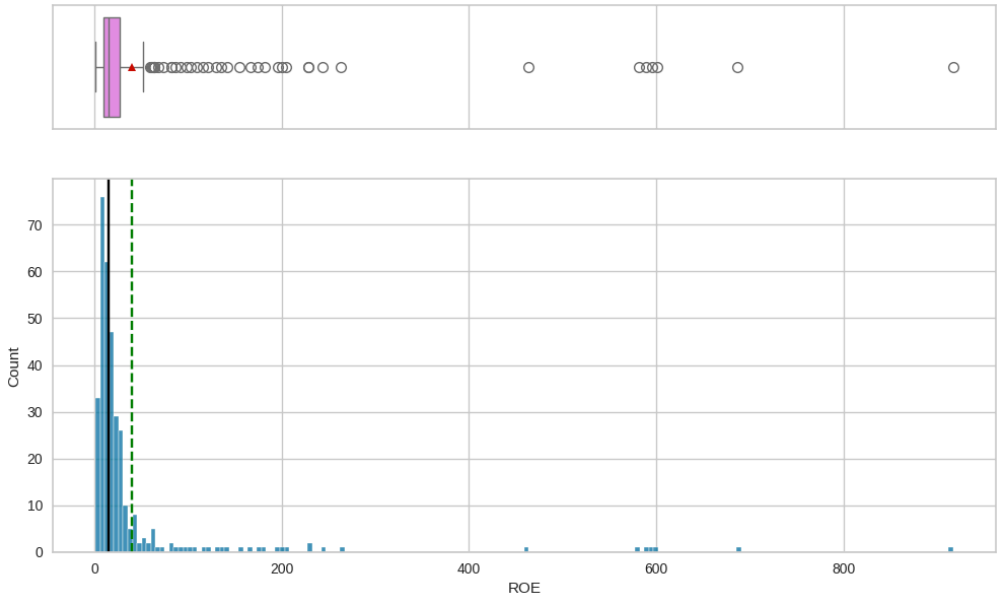

Fig 4

5. CASH RATIO

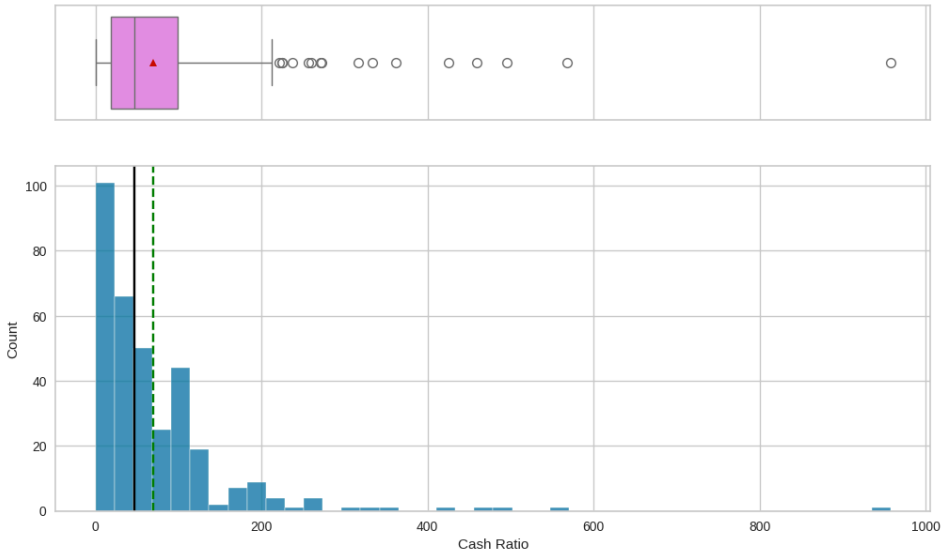

FIG 5

6. NET CASH FLOW

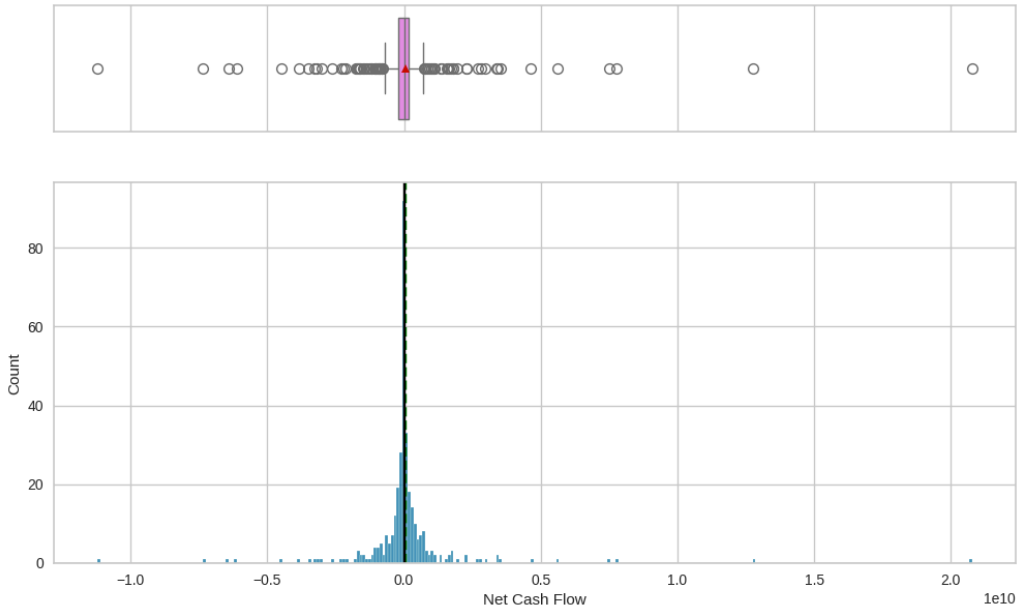

FIG 6

7. NET INCOME

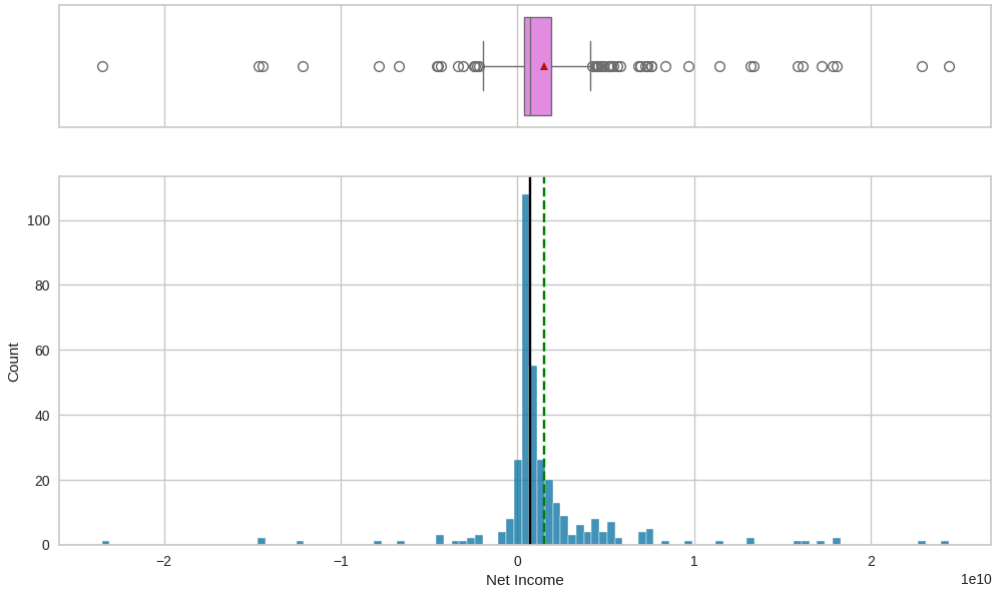

FIG 7

8. EARNING PER SHARE

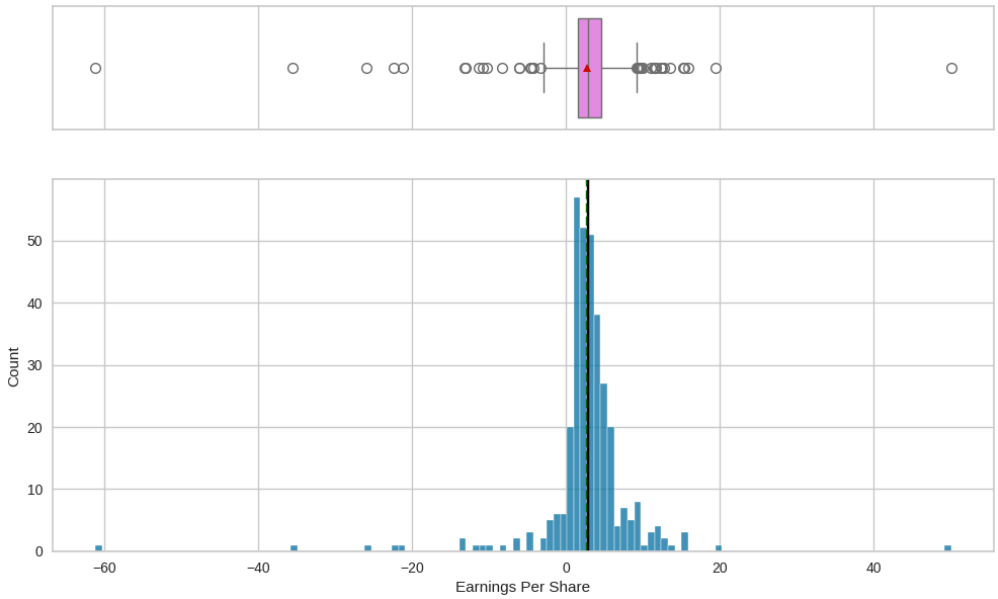

FIG 8

9. Estimated Shares Outstanding

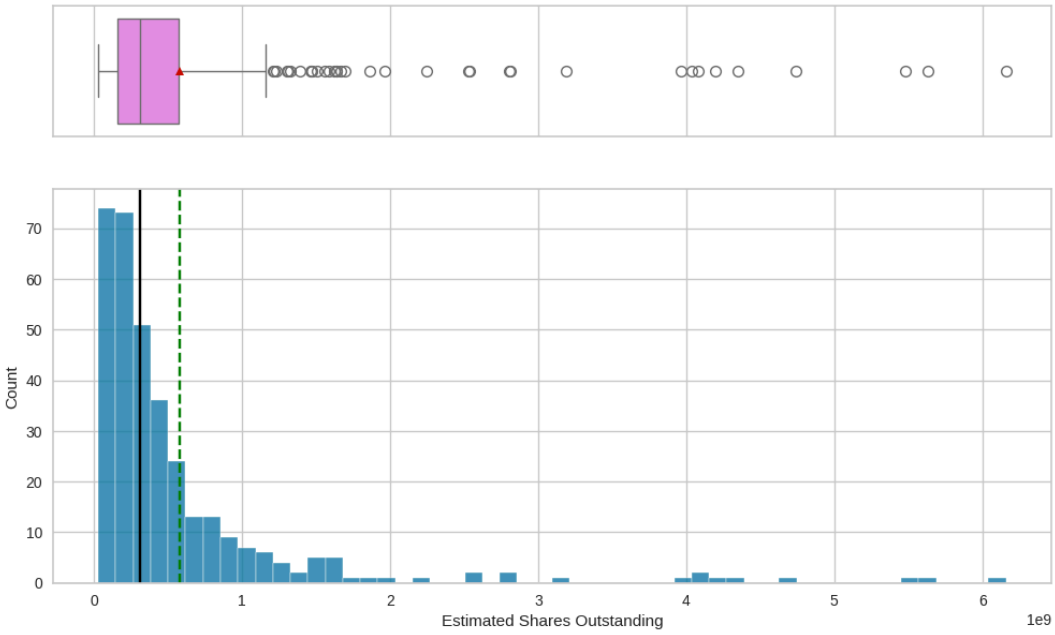

FIG 9

10. P/E Ratio

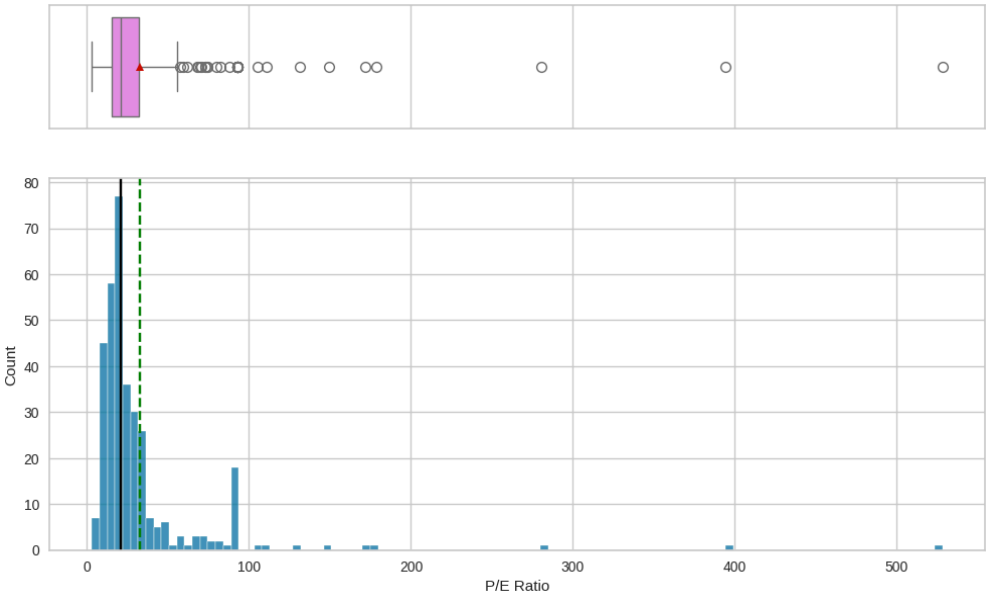

FIG 10

11. P/B Ratio

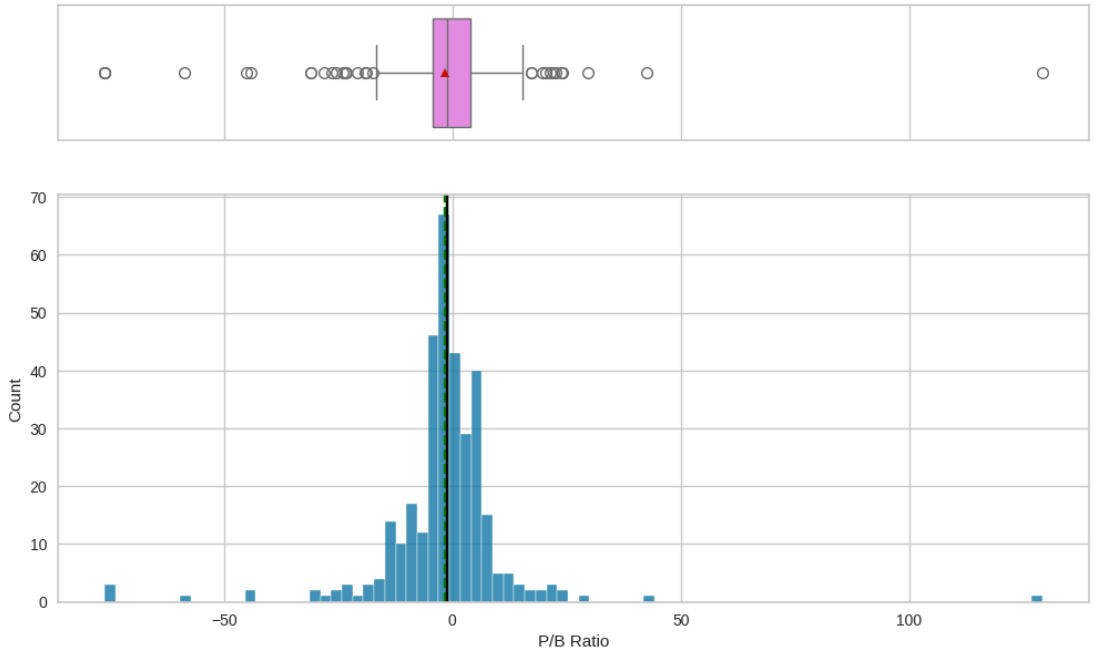

FIg 11

12. GICS Sector

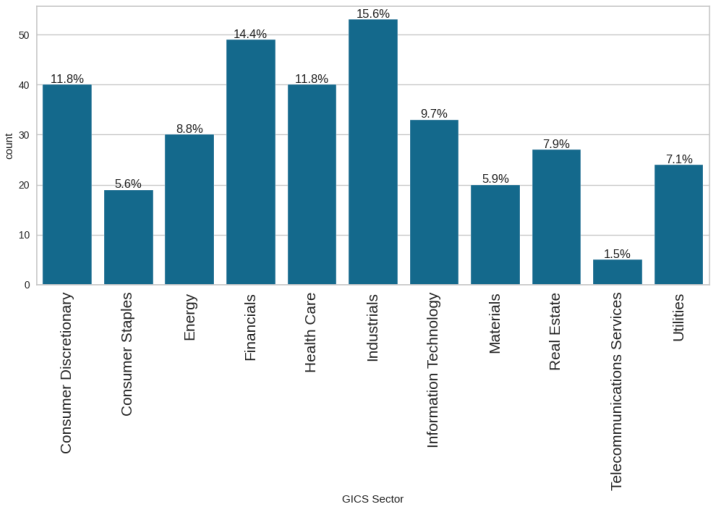

FIg 12

13. GICS Sub Industry

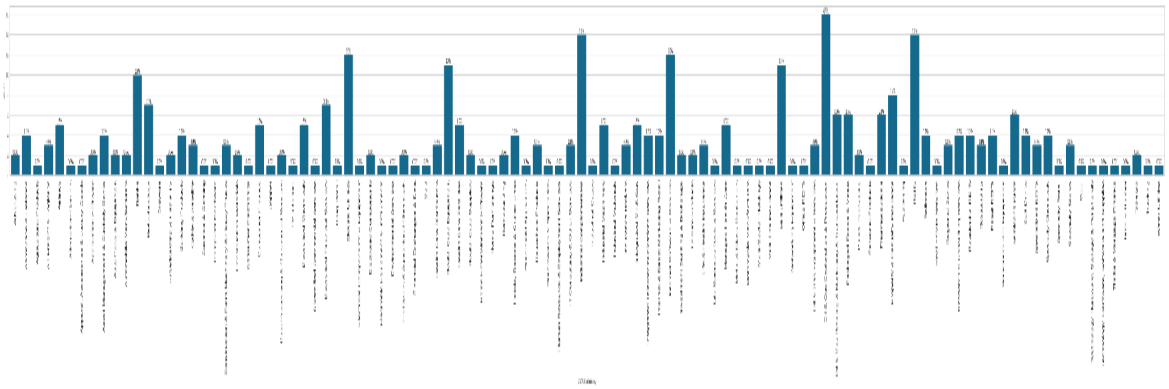

Fig 13

Bivariant Analysis

14. correlation check

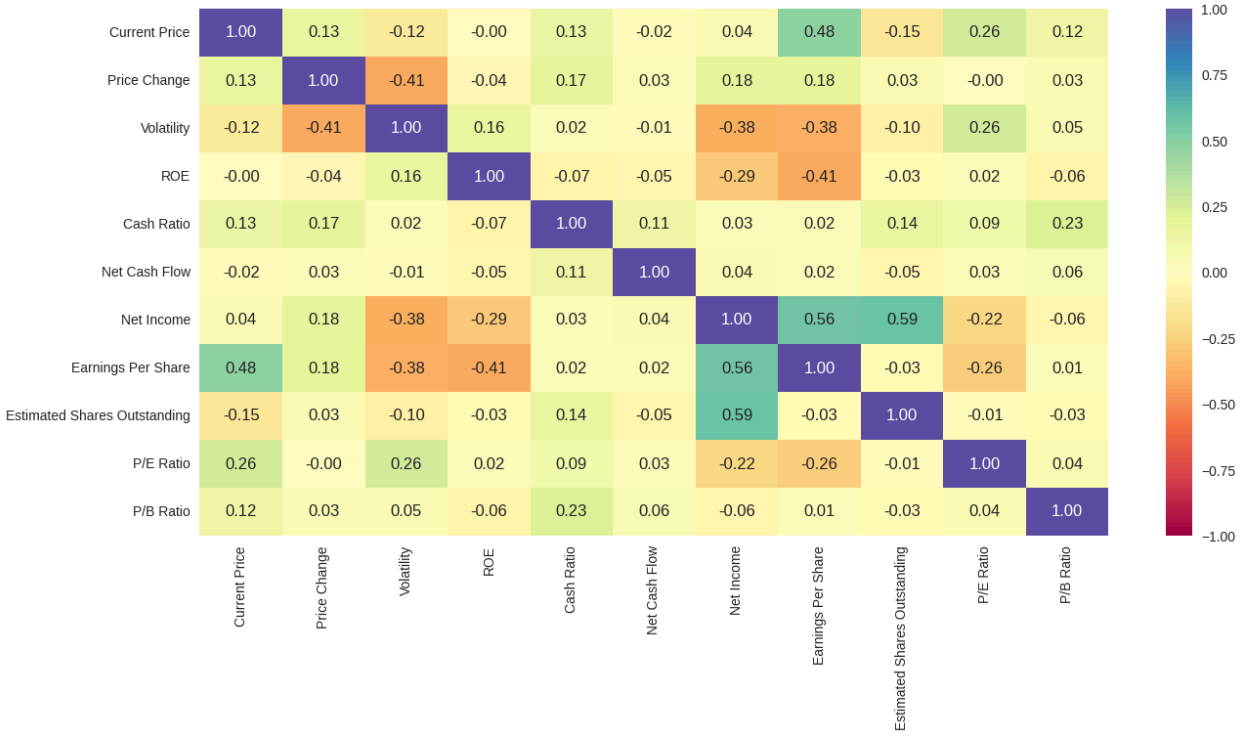

Fig 14

The stocks of which economic sector have seen the maximum price increase on average

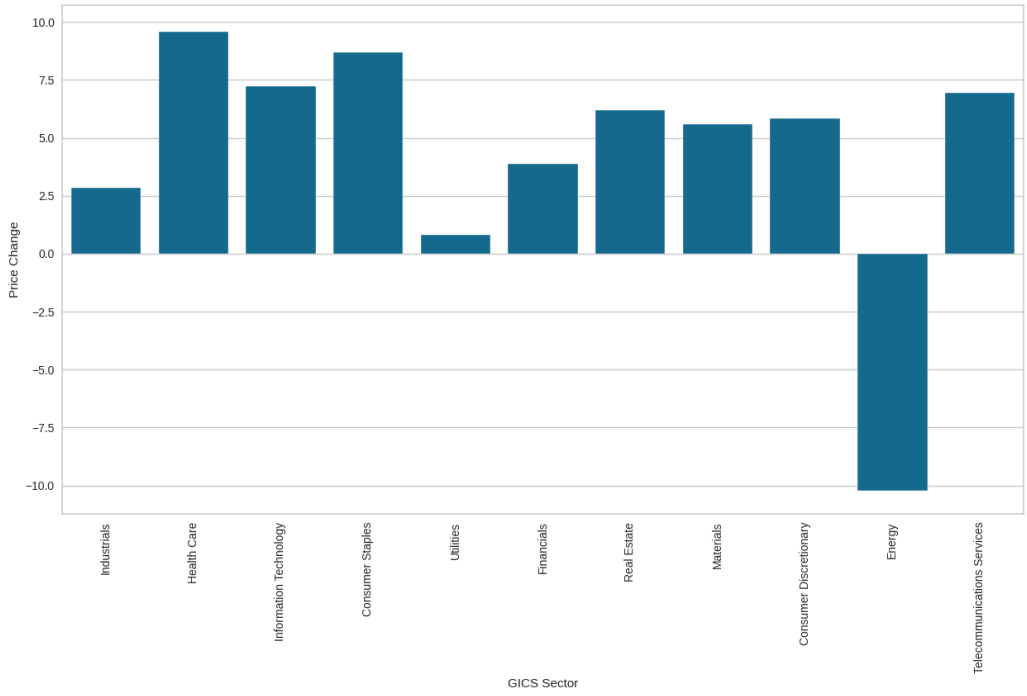

Fig 15

The bar plot shows the average percentage price change for different GICS sectors over a period of 13 weeks. Here's a breakdown:

Health Care, Consumer Staples, and Information Technology sectors have seen the highest average price increases, with Health Care leading at around 10%.

Industrials, Financials, Materials, Consumer Discretionary, and Telecommunications Services have also experienced positive average price changes, though to a lesser degree.

The energy sector is the only one with a significant negative price change, indicating a decrease in stock prices on average for companies in that sector.

Utilities show minimal change, hovering around zero, indicating stability in stock prices for this sector.

This analysis helps in understanding which sectors are performing well and which are underperforming based on recent stock price trends.

Cash ratio provides a measure of a company's ability to cover its short-term obligations using only cash and cash equivalents

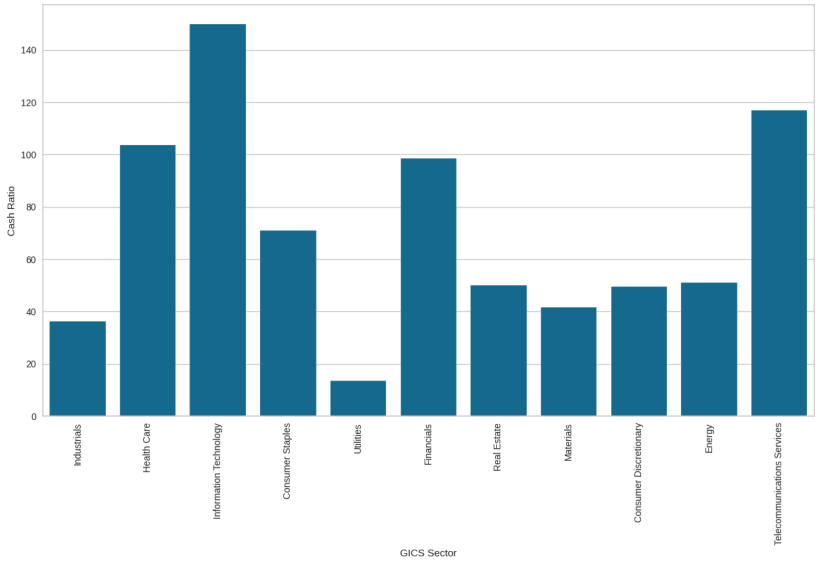

Fig 16

P/E ratios can help determine the relative value of a company's shares as they signify the amount of money an investor is willing to invest in a single share of a company per dollar of its earnings

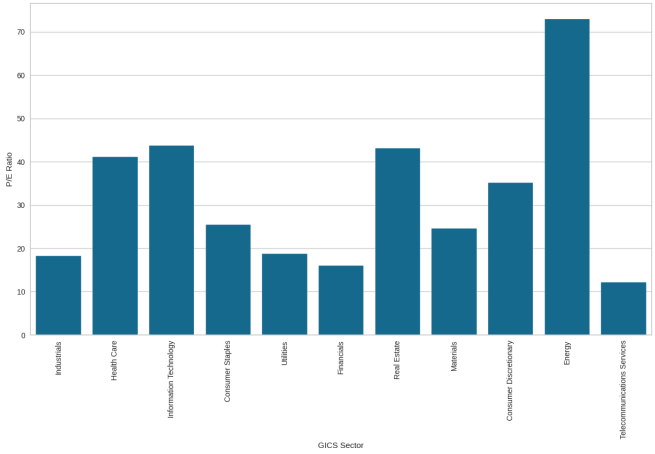

Fig 17

Volatility accounts for the fluctuation in the stock price. A stock with high volatility will witness sharper price changes, making it a riskier investment

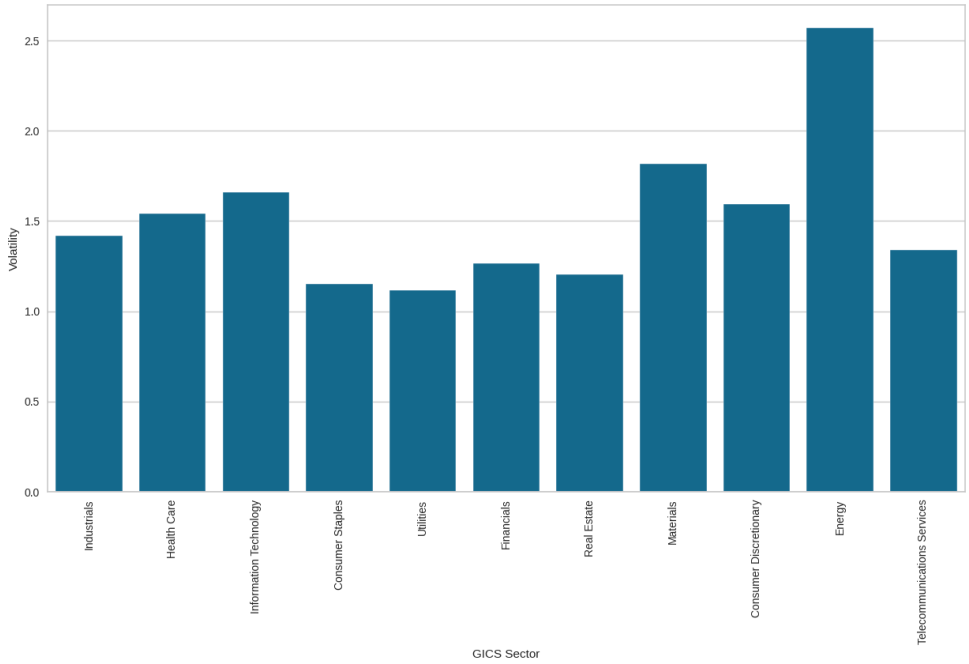

Fig 18

Data Processing

Outlier Check :

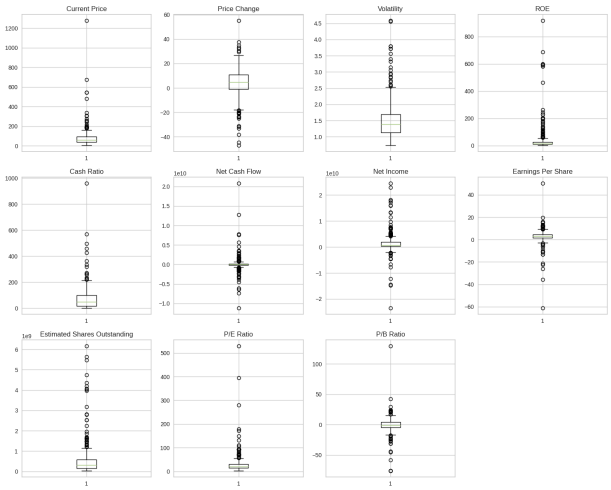

Fig 19

K- Means

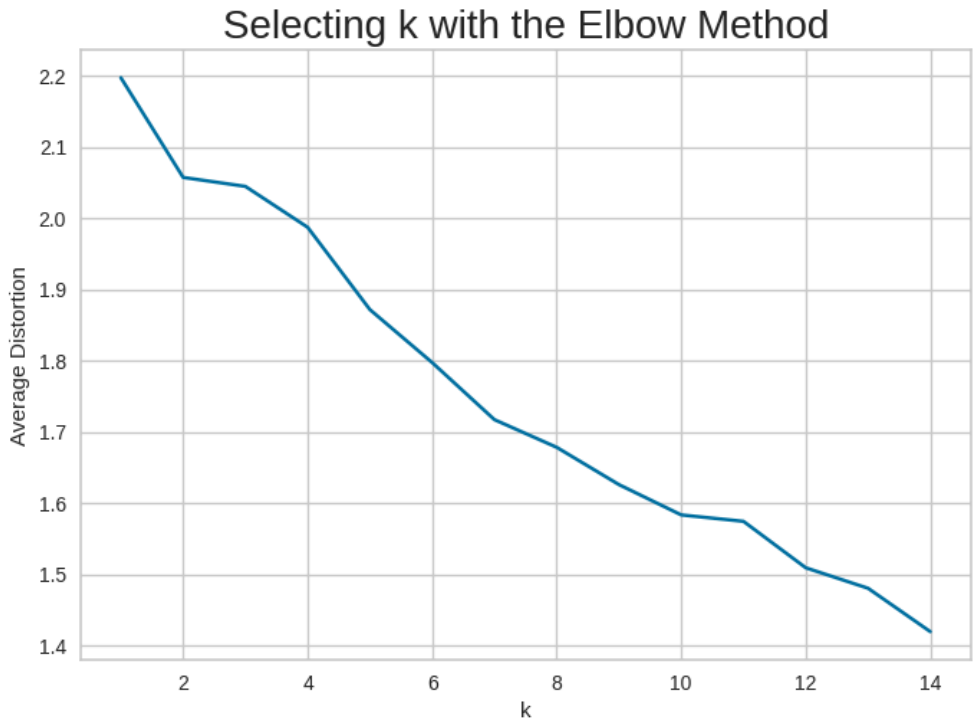

Fig 20

This plot shows the Elbow Method for selecting the optimal number of clusters (k) in a clustering algorithm, typically K-means.

The x-axis represents the number of clusters (k).
The y-axis shows the average distortion (also known as inertia or sum of squared distances), which measures how well the clusters fit the data.
As k increases, the distortion decreases because adding more clusters improves how well the model fits the data. However, after a certain point, the reduction in distortion slows down, creating an "elbow" shape in the curve.

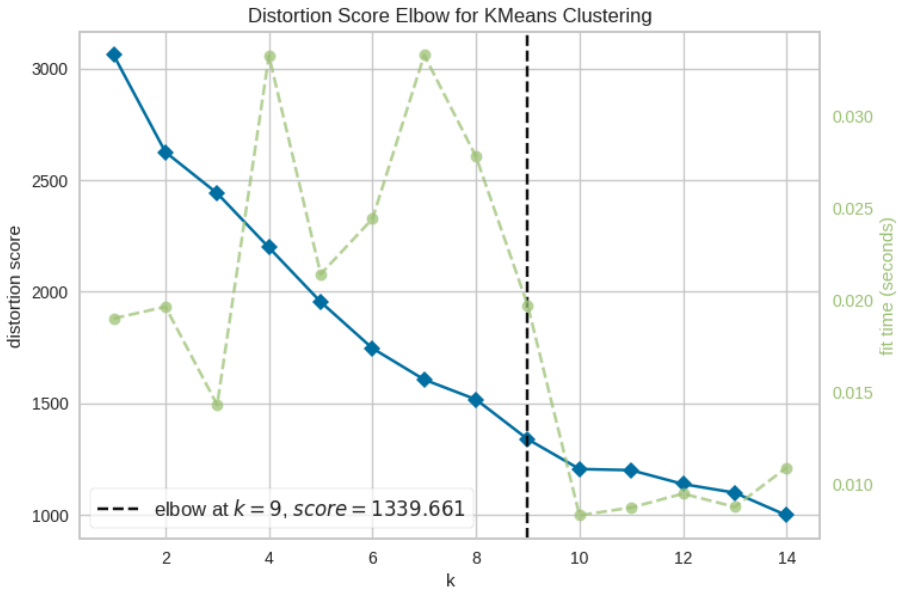

Fig 21

<Axes: title={'center': 'Distortion Score Elbow for KMeans Clustering'}, xlabel='k', ylabel='distortion score'>

The optimal number of clusters is 9, as suggested by the elbow point at k = 9, where the distortion score is 1339.661, balancing between a lower distortion score and a reasonable fit time.

silhouette scores

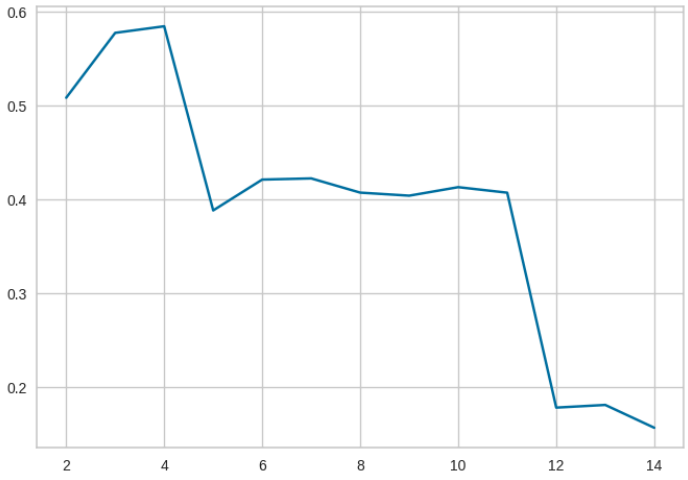

Fig 22

The score peaks around k = 3 or 4, which indicates that the clustering model performs best when using 3 or 4 clusters.

After k = 4, the score gradually declines, with a sharp drop around k = 12, indicating that higher numbers of clusters may lead to overfitting or poorly defined clusters.
This suggests that the optimal number of clusters based on the Silhouette Score might be 3 or 4.

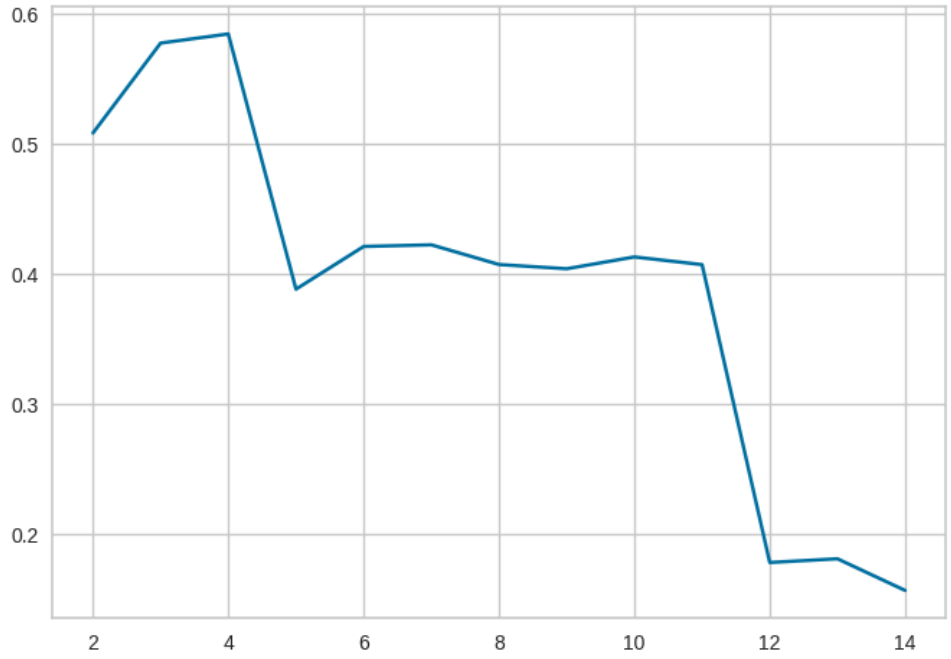

Fig 23

The score peaks around k = 3 or 4, which indicates that the clustering model performs best when using 3 or 4 clusters.
After k = 4, the score gradually declines, with a sharp drop around k = 12, indicating that higher numbers of clusters may lead to overfitting or poorly defined clusters.
This suggests that the optimal number of clusters based on the Silhouette Score might be 3 or 4.

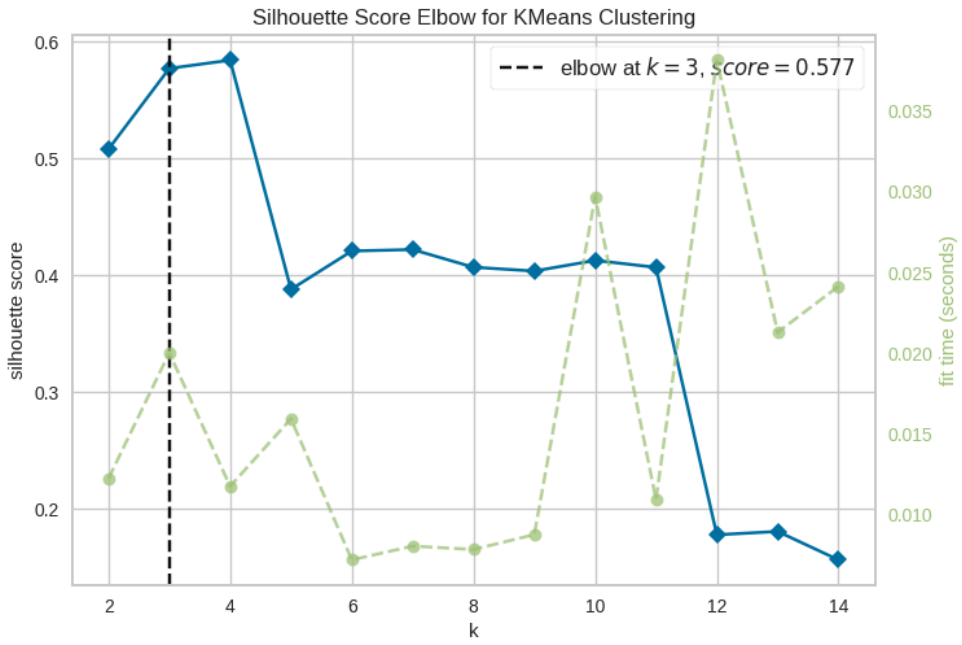

Fig 24

This graph shows the Silhouette Score Elbow Method for determining the optimal number of clusters (k) in KMeans clustering.

Silhouette score (y-axis, left): A metric that measures how well the points are clustered. Higher scores indicate better-defined clusters.
k (x-axis): The number of clusters tested.
The blue solid line represents the silhouette score for different values of k.
The dashed black line indicates the "elbow point" at k = 3, where the silhouette score is 0.577, suggesting that 3 clusters is the optimal choice.

The green dashed line with circular markers shows the fit time (y-axis, right), representing the time it took to compute the KMeans clustering for each k value.

In this case, the "elbow" at k = 3 indicates the best trade-off between compact clusters and computational efficiency.

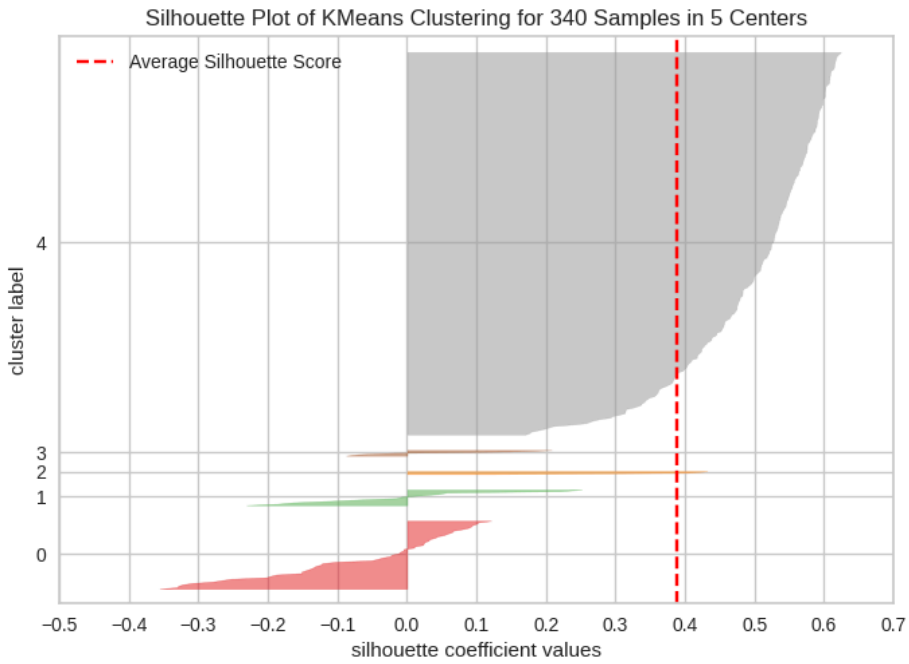

Fig 25

Silhouette Coefficient: Measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). The values range from -1 to 1.

Positive values (close to 1): Indicate that the samples are well clustered.
Negative values (close to -1): Suggest that the samples might be assigned to the wrong clusters.
Clusters: Each color represents a different cluster (there are 5 in total, as indicated by the y-axis labels 0-4).

Red vertical dashed line: This marks the average silhouette score, which summarizes the clustering performance across all samples.

Wide cluster shapes: Cluster 4 (grey) seems to have a high silhouette score, indicating good clustering. Cluster 0 (red) has many points with negative values, suggesting it may not be well-separated.

Creating Final Model

KMeans(n_clusters=4, random_state=1)

Cluster Profiling

Cluster 0: Has 317 entities, moderate price and volatility, strong ROE, positive net income and EPS, but a negative P/B ratio.
Cluster 1: Contains 15 entities, lower price and ROE, high volatility, negative net income, and cash flow.
Cluster 2: Includes 3 entities with high stock price and volatility, positive ROE and cash ratio, and an extremely high P/E ratio.
Cluster 3: Comprises 5 entities, very high price and P/E ratio, strong net income and EPS.

In cluster 0, the following companies are present:
Column 'Company' not found in DataFrame.

In cluster 3, the following companies are present:
Column 'Company' not found in DataFrame.

In cluster 1, the following companies are present:
Column 'Company' not found in DataFrame.

In cluster 2, the following companies are present:
Column 'Company' not found in DataFrame.

Security

KM_segments GICS Sector

0 Consumer Discretionary 35
Consumer Staples 17
Energy 22
Financials 48
Health Care 37
Industrials 52
Information Technology 31
Materials 19
Real Estate 27
Telecommunications Services 5
Utilities 24
1 Consumer Discretionary 2
Consumer Staples 2
Energy 8
Financials 1
Industrials 1
Materials 1
2 Consumer Discretionary 1

Health Care 1
Information Technology 1
3 Consumer Discretionary 2
Health Care 2
Information Technology 1

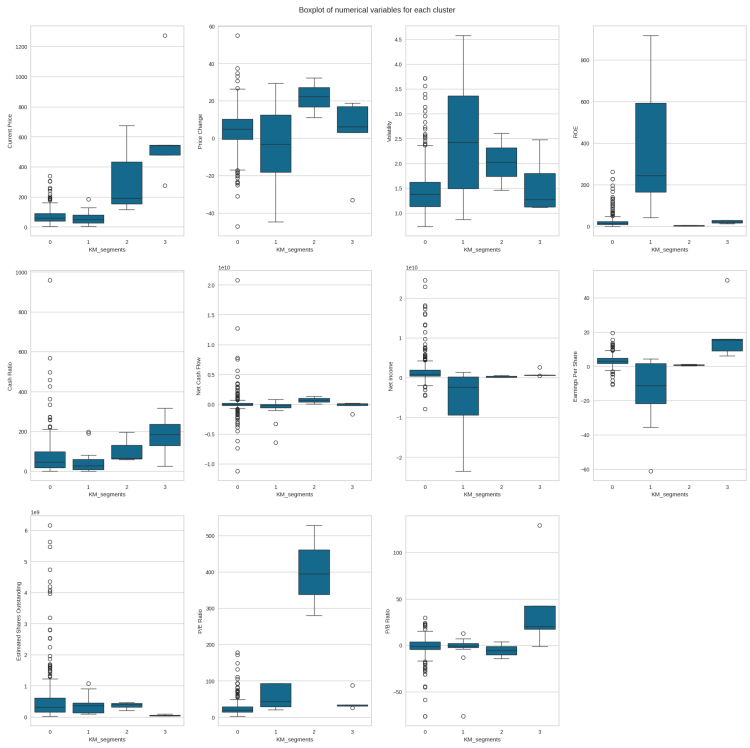

Fig 26

Cophenetic correlation for single linkage is 0.9327510147514677.
Cophenetic correlation for complete linkage is 0.8290907943563686.
Cophenetic correlation for average linkage is 0.9514610013676482.
Cophenetic correlation for ward linkage is 0.6654907779941053.
****

Hierarchical Clustering

Checking Dendrograms

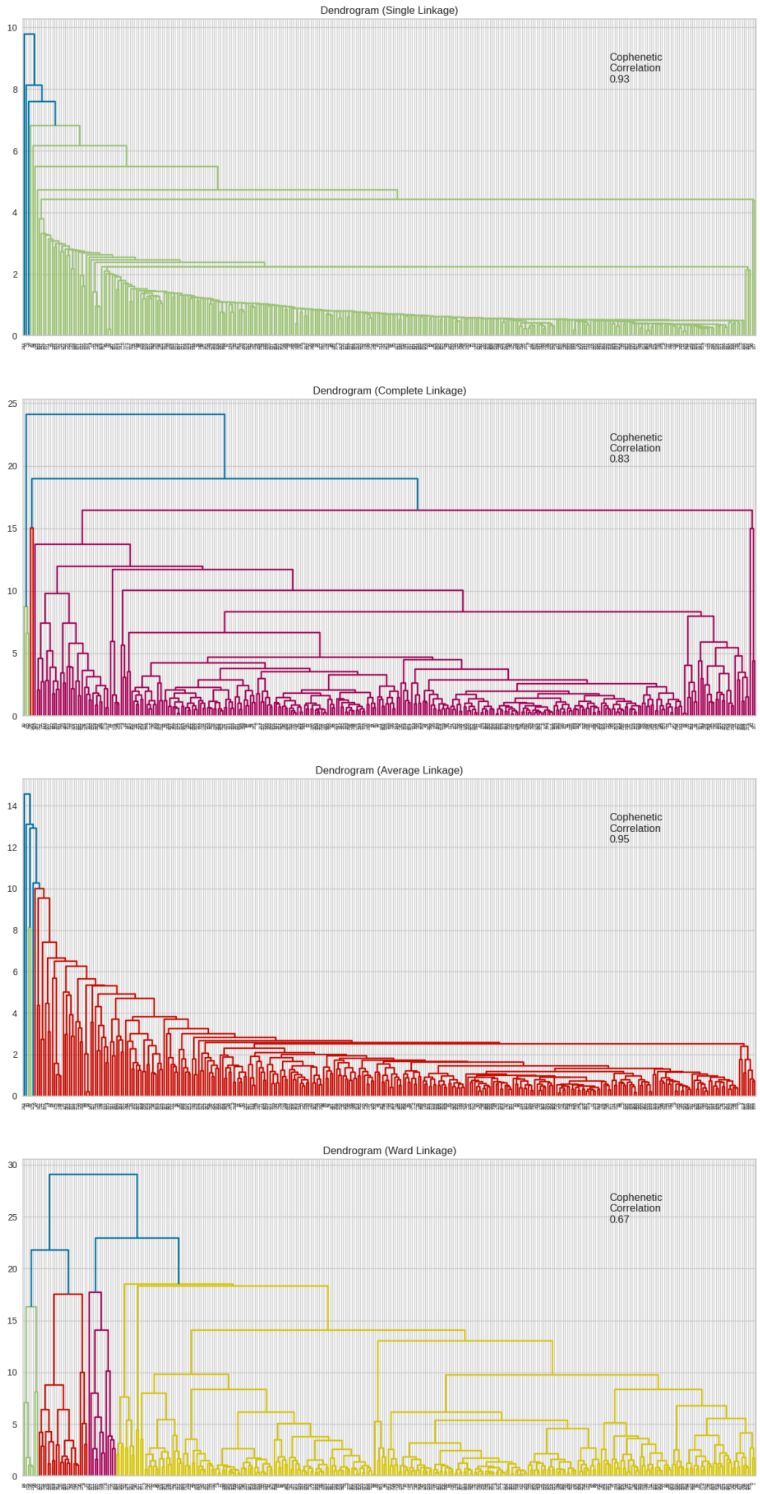
Dendrogram (Single Linkage)
Cophenetic Correlation 0.93
Dendrogram (Complete Linkage)
Cophenetic Correlation 0.83
Dendrogram (Average Linkage)
Cophenetic Correlation 0.95
Dendrogram (Ward Linkage)
Cophenetic Correlation 0.67

Fig 27

	Linkage	Cophenetic Coefficient
3	ward	0.665491
1	complete	0.829091
0	single	0.932751
2	average	0.951461

Creating Model Using Skeleton

In cluster 0, the following companies are present:
Column 'Company' not found in DataFrame. Check your data.

In cluster 1, the following companies are present:
Column 'Company' not found in DataFrame. Check your data.

In cluster 2, the following companies are present:
Column 'Company' not found in DataFrame. Check your data.

In cluster 3, the following companies are present:
Column 'Company' not found in DataFrame. Check your data.

In cluster 4, the following companies are present:
Column 'Company' not found in DataFrame. Check your data.

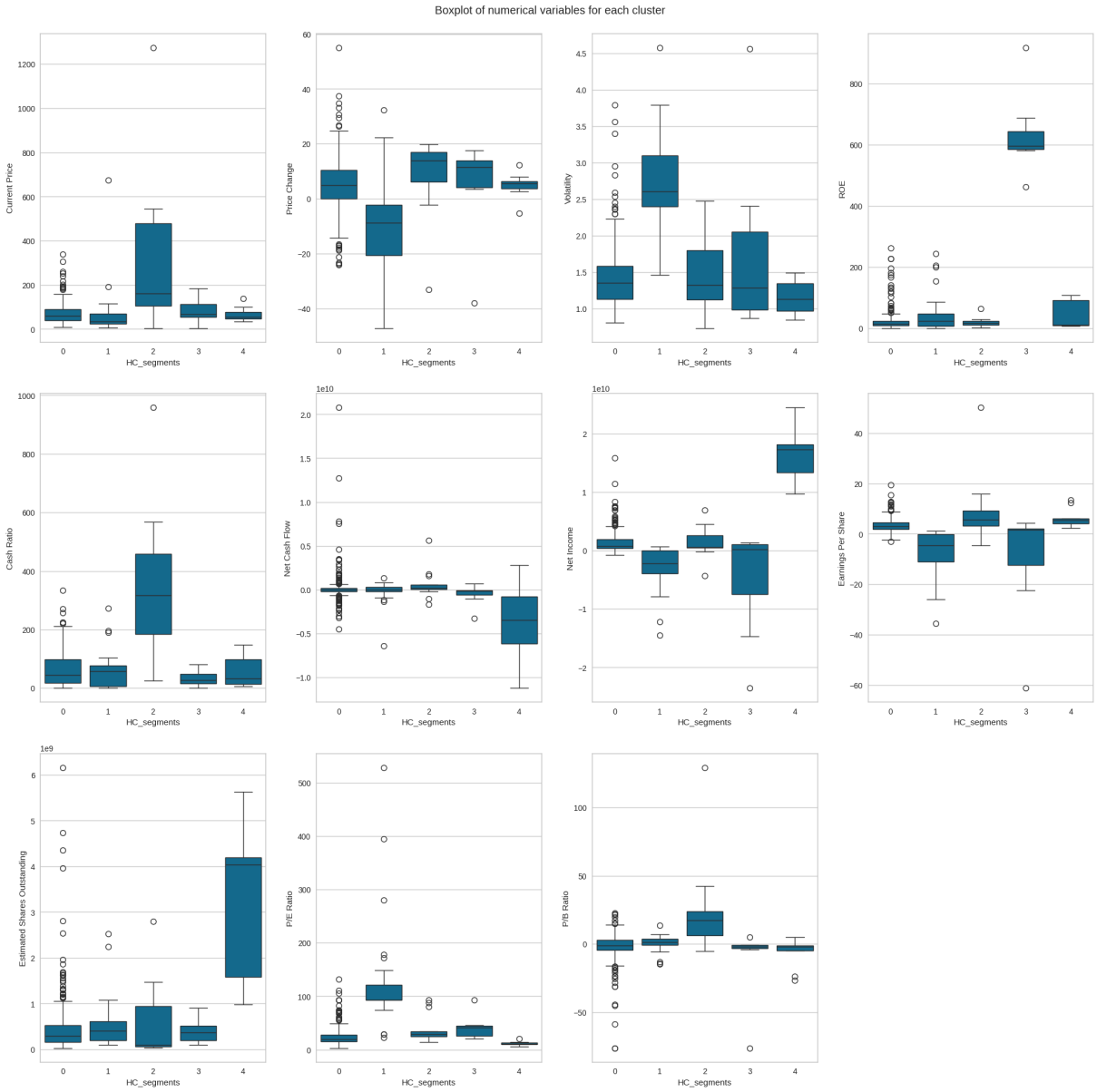
Boxplot of numerical variables for each cluster
Current Price
Price Change
Volatility
ROE
Cash Ratio
Net Cash Flow
Net Income
Earnings Per Share
Estimated Shares Outstanding
P/E Ratio
P/B Ratio
HC_segments

Fig 28

K-means vs Hierarchical Clustering

1. Which clustering technique took less time for execution?
The time efficiency of clustering techniques can vary based on the algorithm's complexity:

K-Means Clustering: This is typically faster because it relies on simple distance calculations (like Euclidean) and iteratively refines the cluster centroids.
Hierarchical Clustering: Usually slower than K-Means, especially for large datasets, because it involves computing pairwise distances between data points at every level of the hierarchy.
In your case, if both methods were applied, K-Means Clustering likely took less time to execute compared to Hierarchical Clustering, which tends to be more computationally intensive.

2. Which clustering technique gave you more distinct clusters, or are they the same?
The number of clusters and the distinction between them can depend on several factors:

K-Means Clustering: Produces a pre-defined number of clusters (based on the value of k), and the distinctiveness of clusters depends on the data's spread and the choice of k. K-Means often creates clusters that are more distinct in terms of minimizing within-cluster variance.

Hierarchical Clustering: Can create more natural clusters and doesn’t require pre-specifying the number of clusters. However, the clusters may be less distinct compared to K-Means because it merges clusters in a stepwise fashion, even when clusters may not be entirely distinct.

Depending on the dataset’s nature, K-Means likely gave you more distinct clusters, whereas Hierarchical might have produced more gradual or overlapping cluster distinctions.

3. How many observations are there in the similar clusters of both algorithms?
To answer this, you would need to compare the cluster labels produced by both algorithms. Here’s a general approach:

K-Means Clustering: Assigns each observation to a single cluster, based on centroids.
Hierarchical Clustering: You can decide the number of clusters by cutting the dendrogram at a certain level.
You can cross-check which observations fall into the same clusters in both algorithms. If the clusters are similar in both techniques, a large proportion of observations should belong to similar clusters. In some cases, these might differ slightly, particularly in boundary cases where Hierarchical might group differently.

4. How many clusters are obtained as the appropriate number of clusters from both algorithms?
The number of clusters is typically decided using techniques like the Elbow Method or Silhouette Score for K-Means, or by analyzing the dendrogram in Hierarchical clustering.

K-Means: The optimal number of clusters is generally determined by observing where the inertia (within-cluster variance) begins to level off (Elbow point). Based on the context provided (for example, in the graph you shared), it looks like the optimal number of clusters might be 3.

Hierarchical Clustering: The number of clusters is decided by selecting a height at which to "cut" the dendrogram, which represents a clustering solution. This can lead to a similar or slightly different number of clusters compared to K-Means.

In most cases, both methods should suggest a similar number of clusters, but there could be slight differences due to the way each method handles distance and merging.

5. Differences or similarities in cluster profiles from both techniques
K-Means: Typically creates more compact, evenly sized clusters. This method works well when clusters are spherical and equally sized.

Hierarchical: Can result in more irregularly shaped clusters and may preserve more structure from the data (e.g., outliers may be placed into separate clusters). It’s better when you don’t have a clear idea of how many clusters to expect.

The profiles of clusters from both methods could differ slightly in terms of cluster shapes and member inclusion, especially if the data contains noise or outliers. Hierarchical clustering might capture more nuanced relationships between stocks, while K-Means would create more clearly separated clusters based on the primary attributes.

Summary:
Execution Time: K-Means is faster.
Distinct Clusters: K-Means often yields more distinct clusters.
Similar Observations: There could be some overlap, but boundary cases might differ.
Number of Clusters: Both methods may yield a similar number of clusters, but Hierarchical might have some flexibility depending on where the dendrogram is cut.
Cluster Profiles: K-Means focuses on minimizing variance within clusters, while Hierarchical might preserve more of the dataset’s natural hierarchy, potentially leading to more nuanced but less distinct clusters.

Actionable Insights & Recommendations

Strong Profitability:

Companies with high ROE like American Airlines (135%) and AbbVie (130%) are effectively generating profits. These stocks could be good candidates for long-term investment.
Overvalued Stocks:

Adobe (P/E: 74.56) and Analog Devices (P/E: 178.45) appear to be overvalued based on P/E ratio. Investors should analyze future growth prospects or wait for a price correction.
Liquidity Strength:

Companies with high cash ratios like Adobe (180) and Analog Devices (272) are in a strong liquidity position. They can cover short-term liabilities, making them more stable investments during uncertain periods.
Price Momentum:

Adobe (Price Change: 13.98) and Abbott (Price Change: 11.30) show recent strong positive momentum, indicating market optimism. These may be suitable for momentum-based investment strategies.