Trade & Ahead - Problem Statement

PGP - Data Science & Business Analytics
February 1, 2025

Leslieane Beltran

Contents/Agenda

- Executive Summary
- Business Overview/Solution Approach
- EDA Results
- Data Preprocessing
- K Means Clustering
- Hierarchical Clustering
- Insights & Recommendations
- Appendix

Executive Summary

- The stock market is a great way to grow wealth, fight inflation, and save for the future, especially when starting early.
- A well-diversified portfolio reduces risk and helps maximize returns across different market conditions.
- Trade&Ahead has hired a Data Scientist to analyze stock data, group stocks based on key attributes, and provide investment insights.
- Using financial metrics like stock price, volatility, ROE, and valuation ratios, stocks will be clustered to identify trends and strong performers.
- This analysis will help investors make smarter decisions by balancing risk and reward for better financial growth.

Business Problem Overview

- Investing in stocks can be complex, with countless financial metrics to analyze, making it difficult to identify the right opportunities.
- Without proper diversification, investors risk higher losses when the market fluctuates.
- Trade&Ahead needs a data-driven approach to help clients make smarter investment decisions.
- Grouping stocks based on shared characteristics can simplify analysis and improve portfolio strategy.
- Identifying low-correlation stocks helps reduce risk and maximize returns over time.
- A structured stock classification system can provide clearer insights and better investment recommendations.

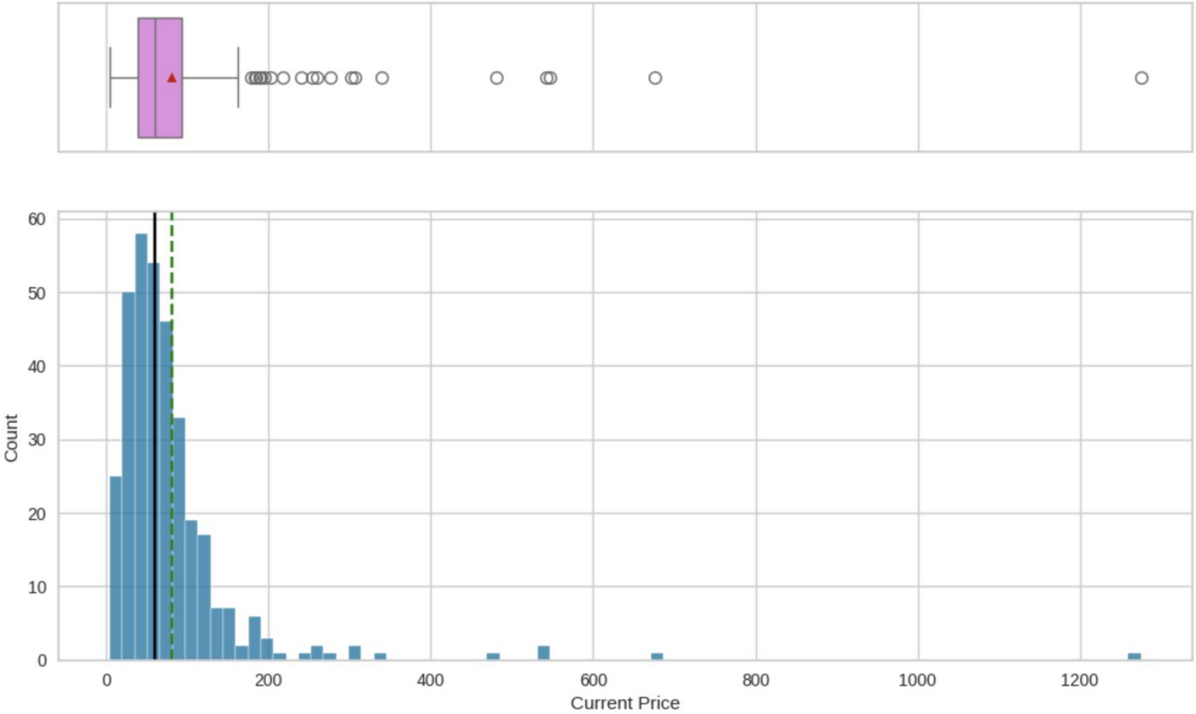
Solution Approach

The following describe the solution approach:

- Analyze stock data using key financial indicators like price, volatility, ROE, and valuation ratios.
- Apply cluster analysis to group stocks with similar characteristics and low correlation.
- Identify patterns and trends within each cluster to simplify decision-making for investors.
- Provide insights on stock groupings to help clients build diversified, risk-balanced portfolios.
- Use data visualization and statistical modeling to enhance understanding and strategy development.
- Deliver actionable recommendations to Trade&Ahead for smarter, data-driven investment strategies.

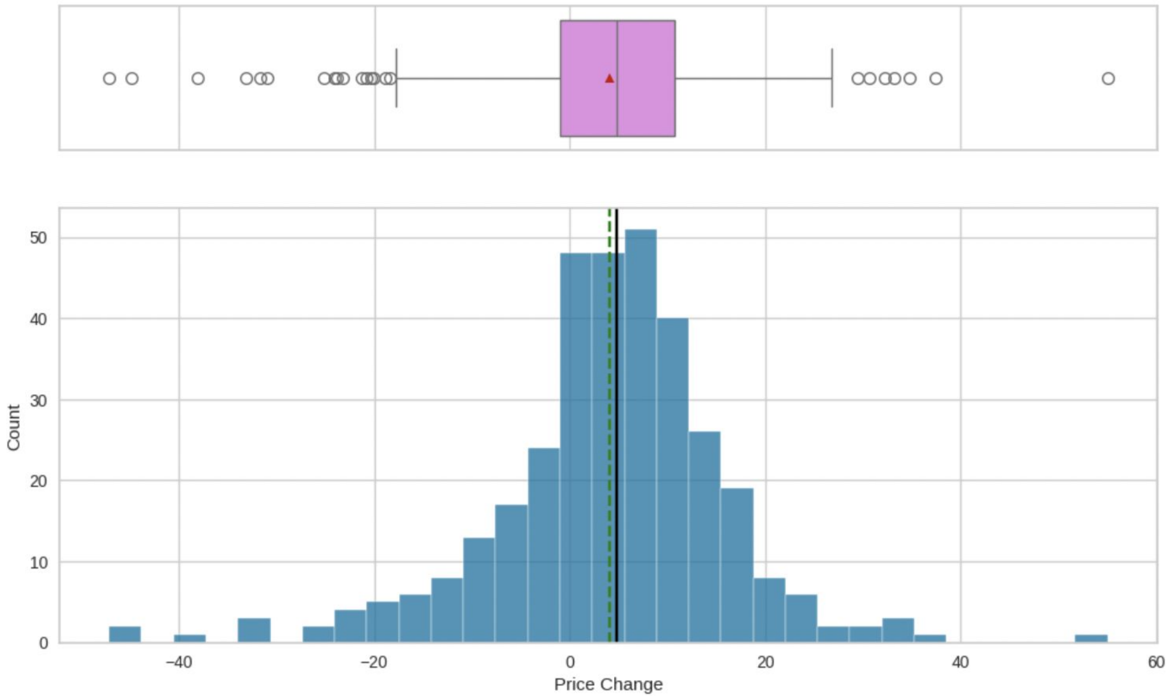
EDA Results - Univariate Analysis

Current Price



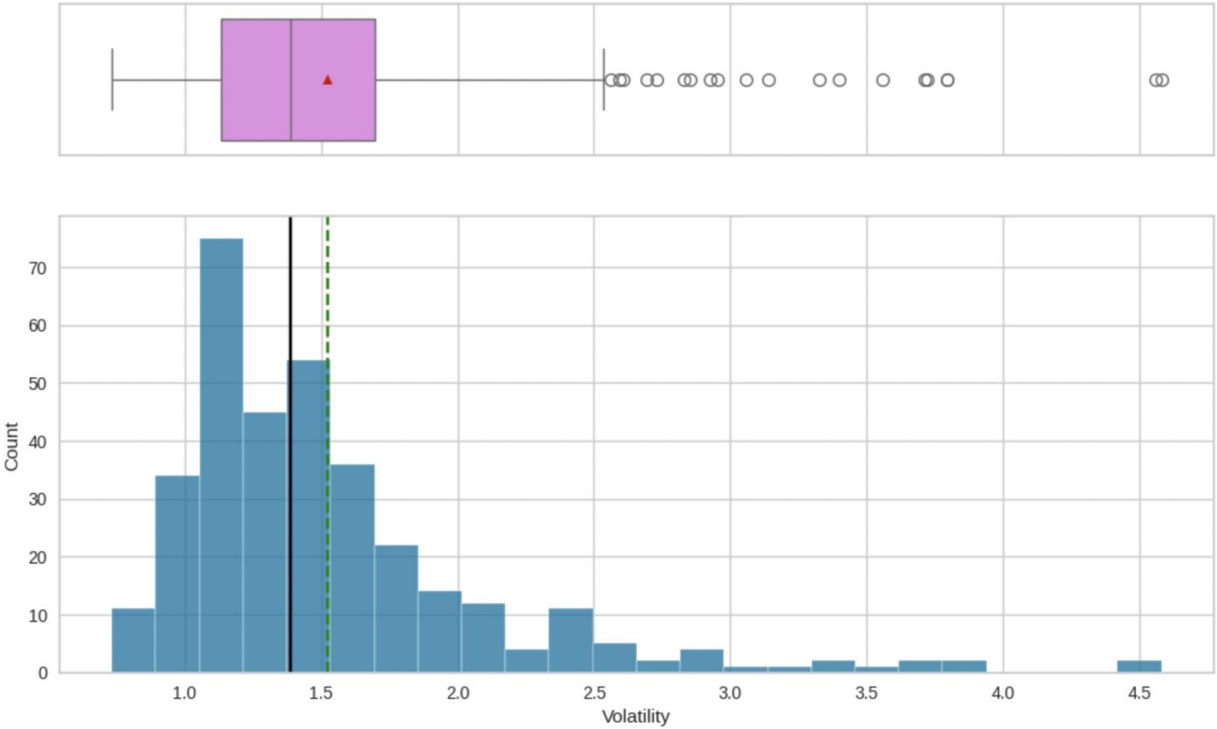
Most stocks are priced below $200, with a few outliers at much higher prices. The data is right-skewed, highlighting a wide range of valuations.

Price Change



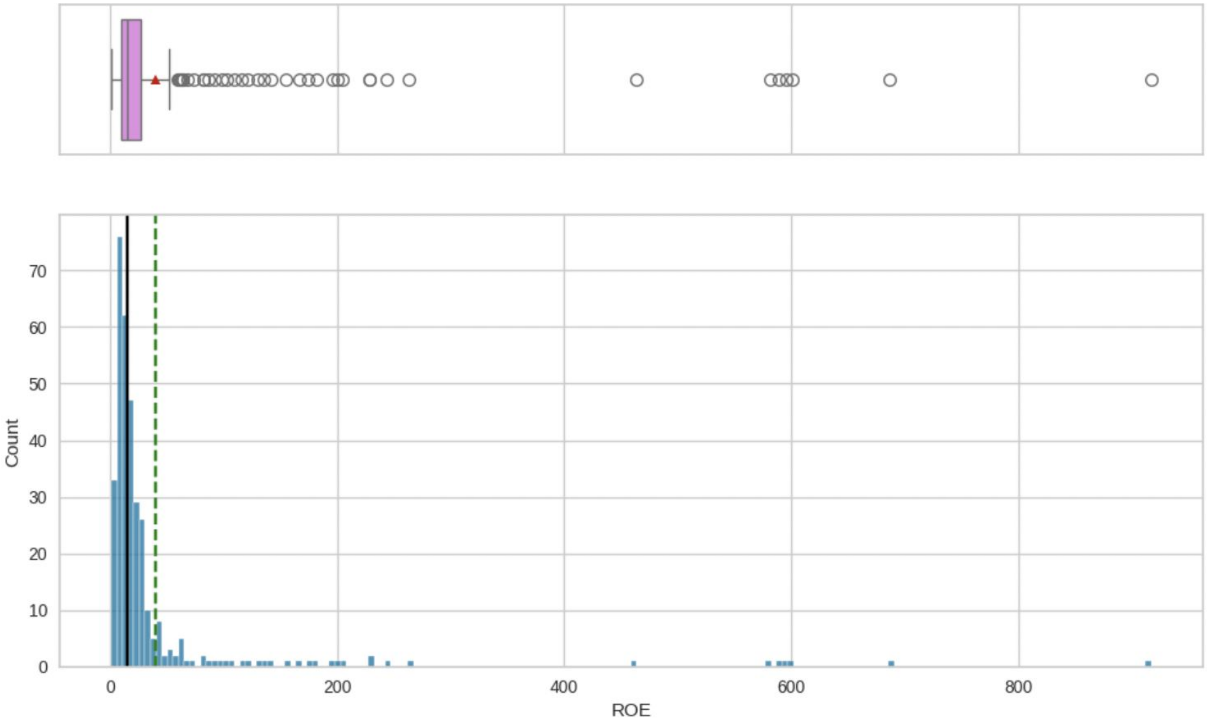
Most stocks experienced a small price change, centered near 0%, indicating minimal overall movement. The distribution is roughly symmetrical, with a few extreme outliers on both the positive and negative ends.

Volatility



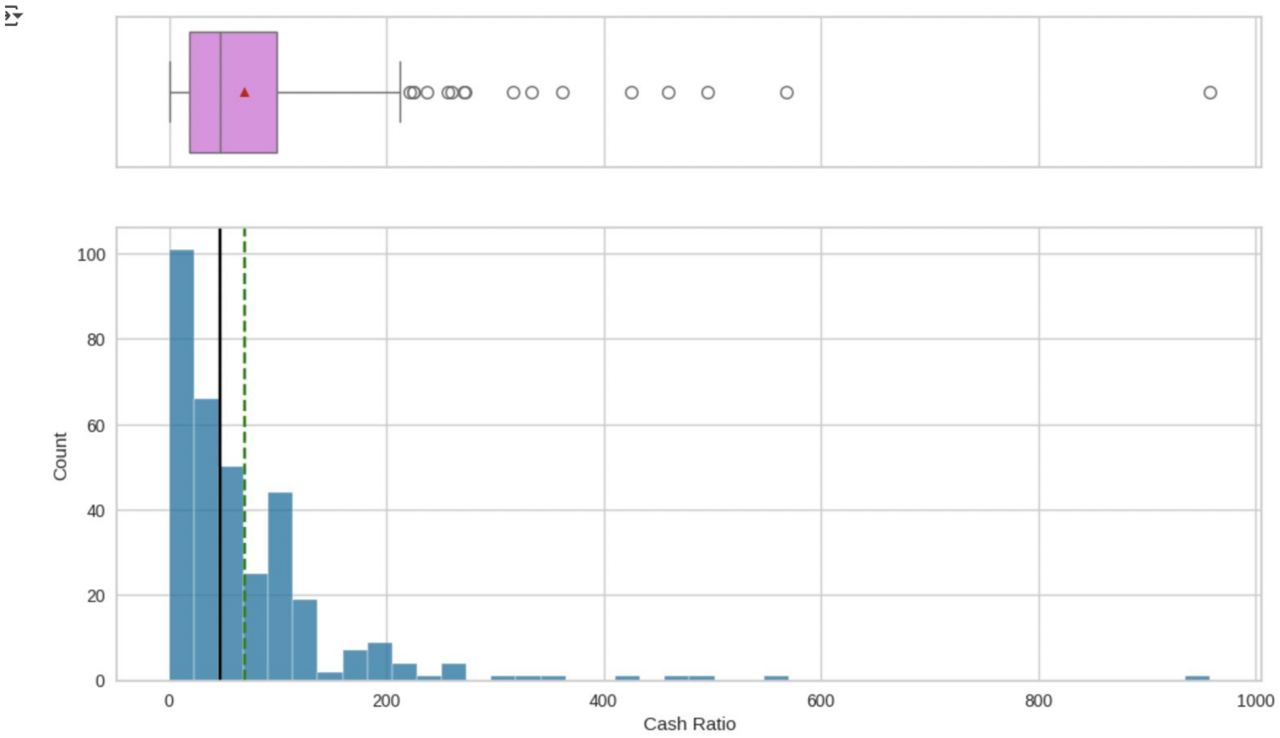
Most stocks have low volatility, with few outliers. Dataset are relatively stable, but some show higher price swings.

ROE



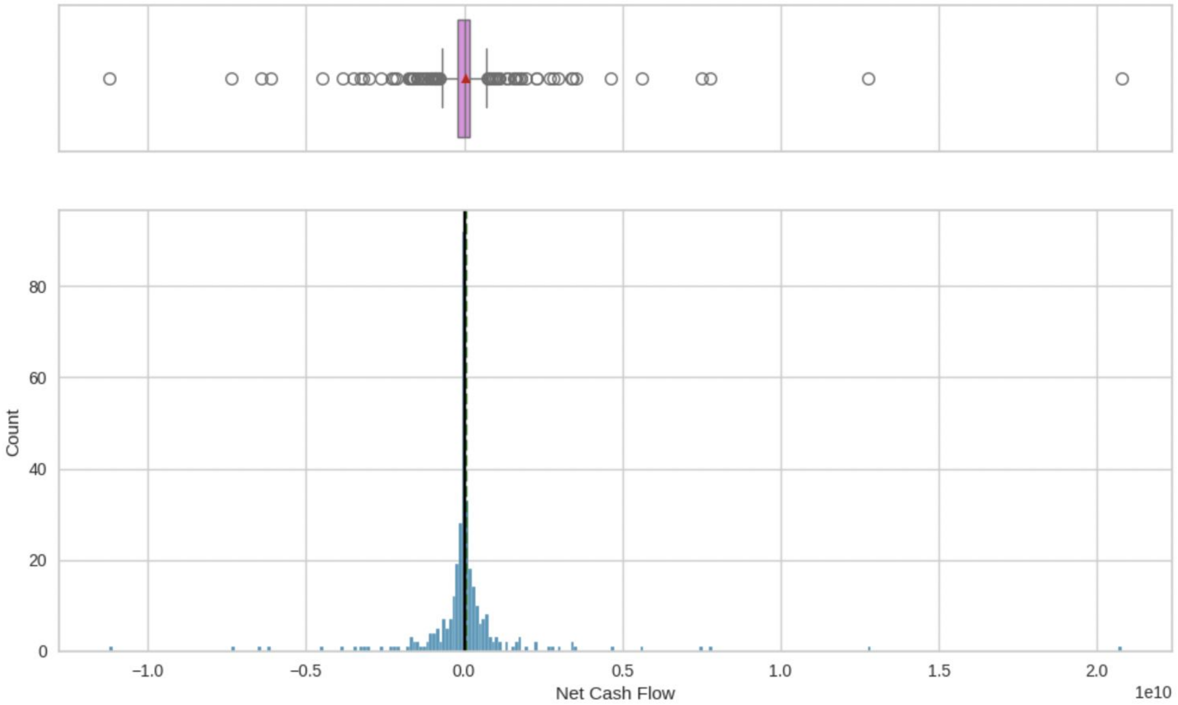
Most companies have a low ROE, clustered near zero, few extreme outliers with very high ROE values. This suggests that while some companies are highly profitable.

Cash Ratio



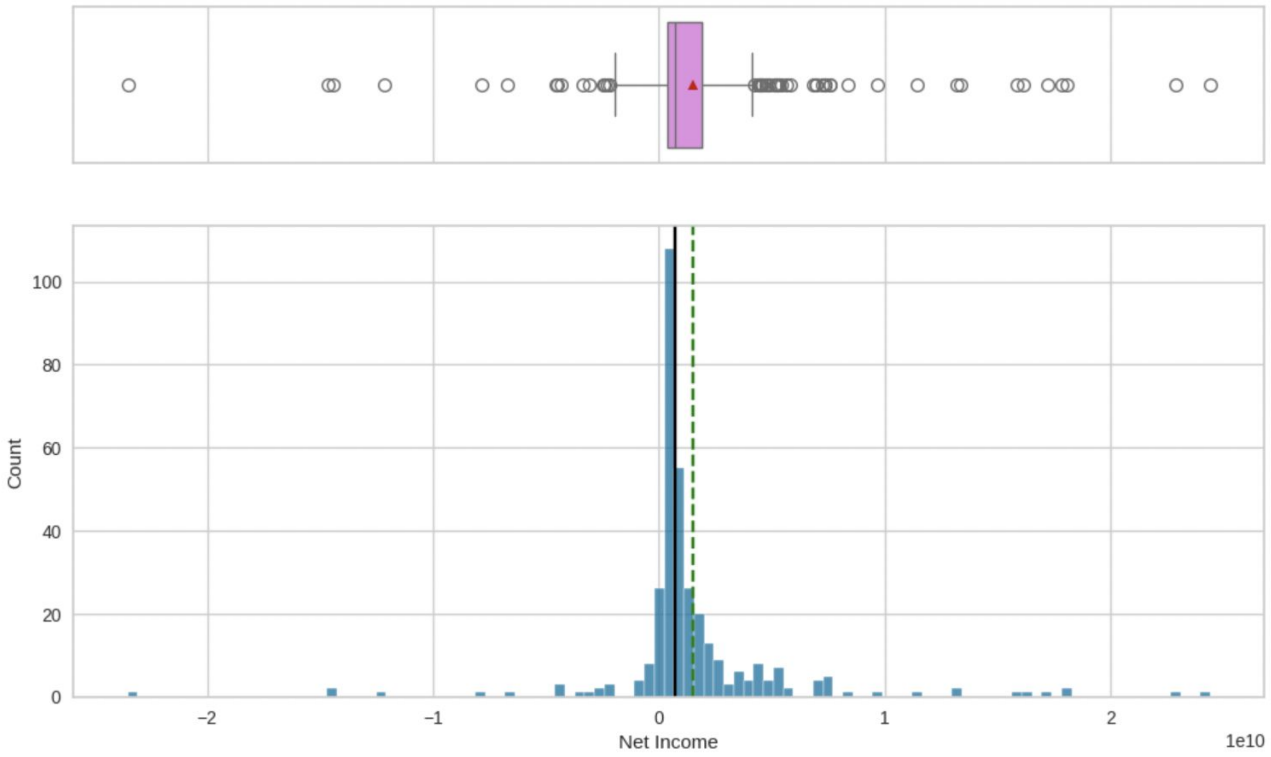
Most companies have a low cash ratio, with just a few having really high values. This means most aren’t holding a lot of cash compared to their liabilities.

Net Cash Flow

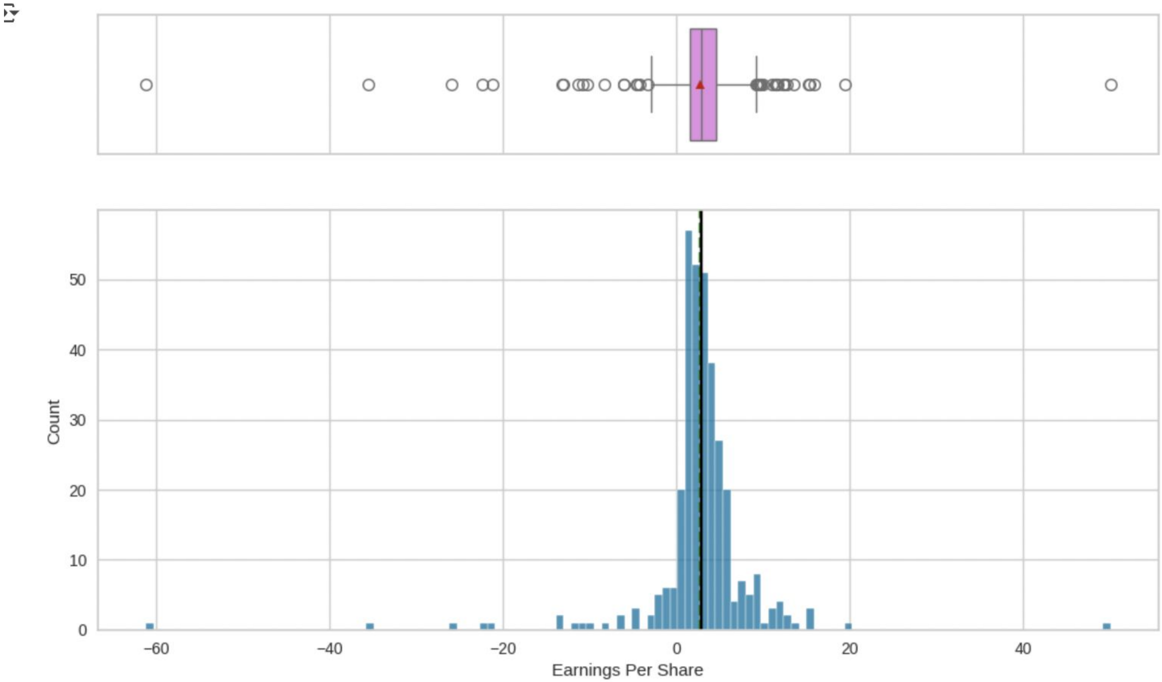


Most companies have a net cash flow close to zero, with a few showing big gains or losses. It looks like cash flow is steady for most, but some outliers stand out.

Net Income

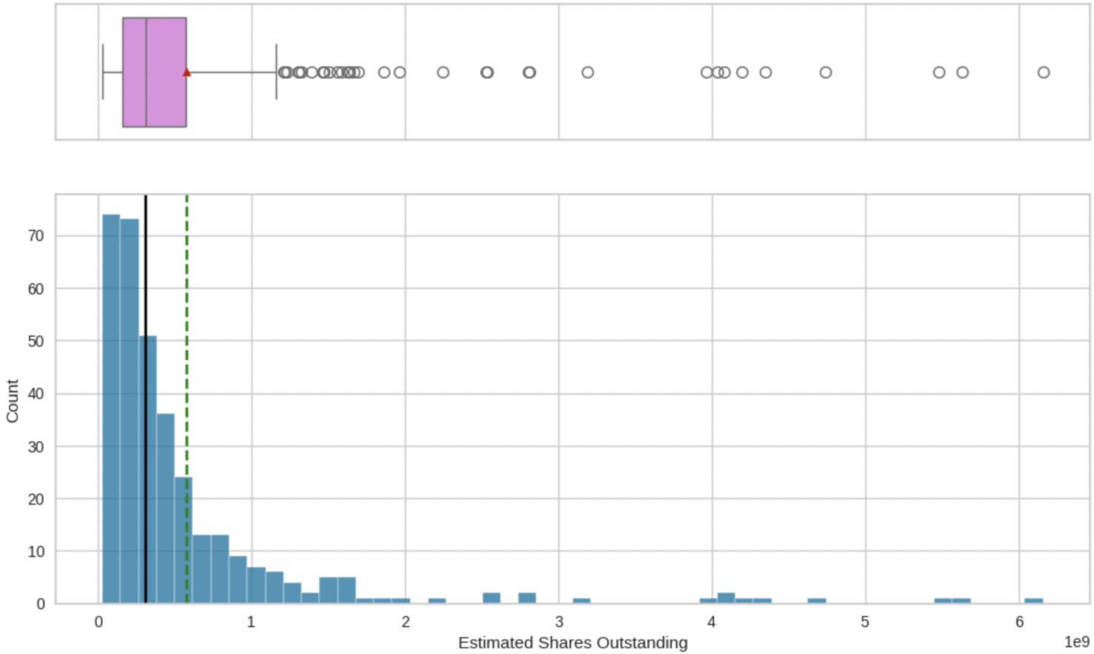


Earnings Per Share



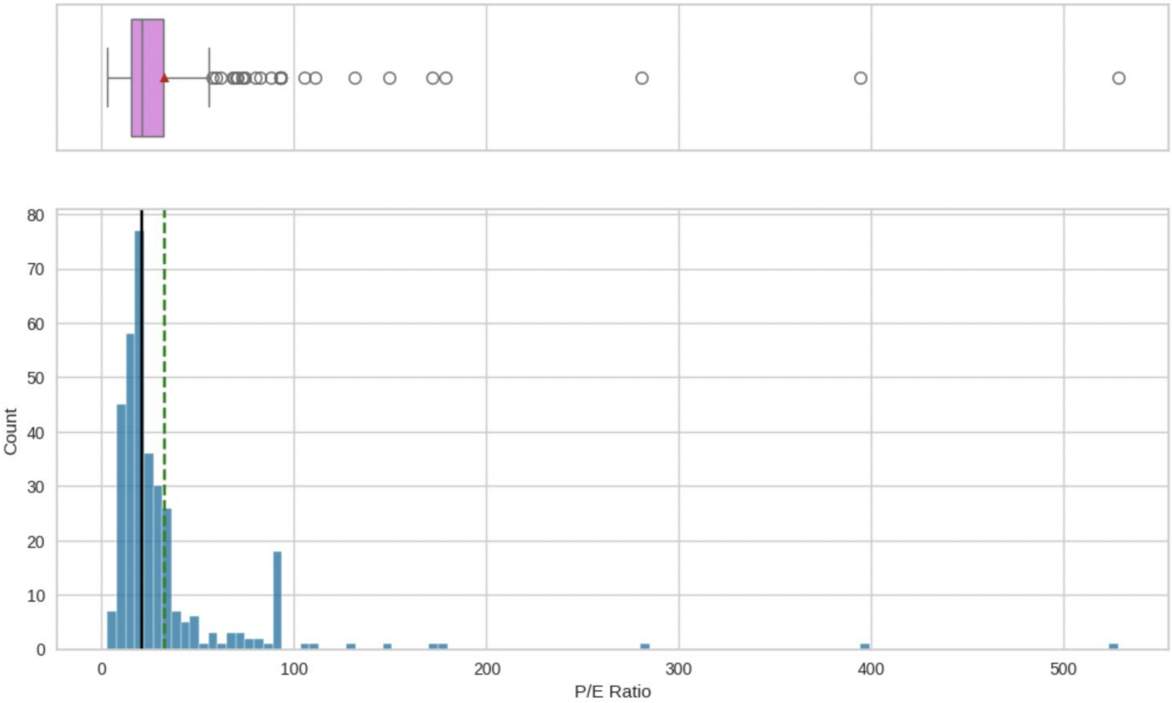
Most companies have earnings per share (EPS) close to zero, with a few outliers on both ends.

Estimated Shares Outstanding



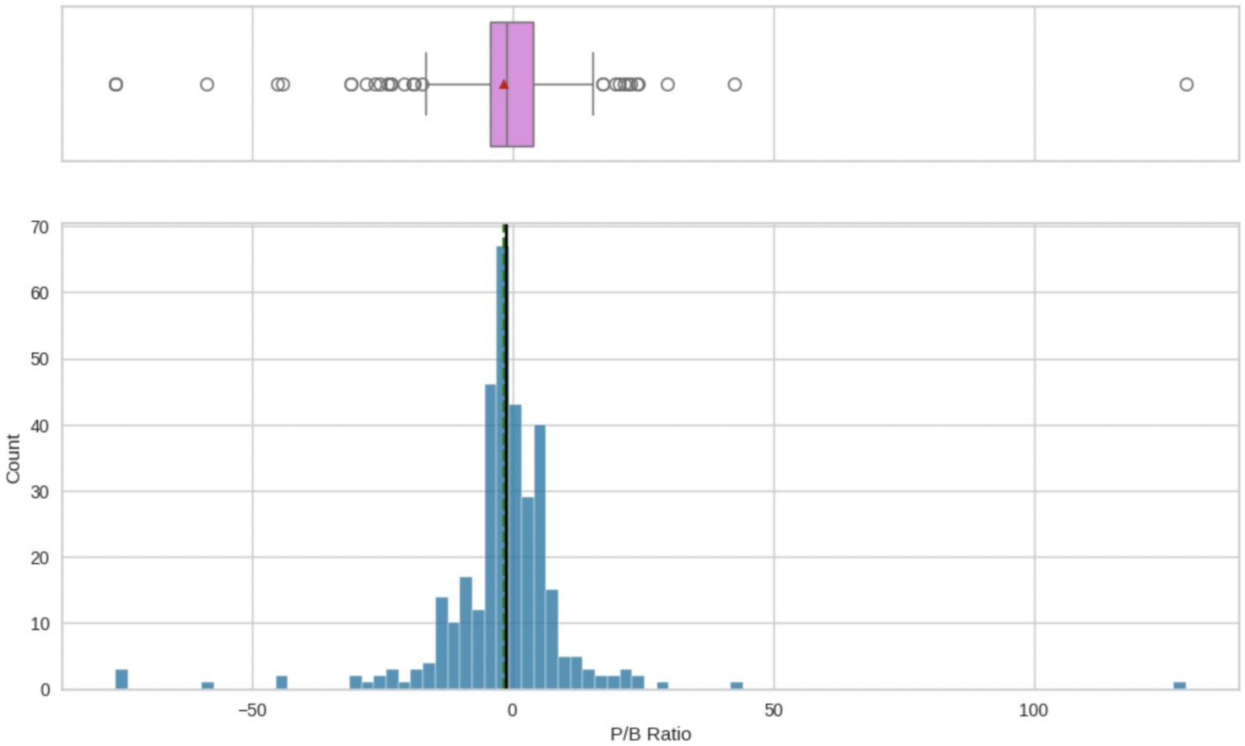
Most companies have a low number of estimated shares outstanding, but there are a few with much higher counts. This shows a big range in company sizes.

P/E Ratio



Most companies have a low P/E ratio, but a few outliers go way higher. This means most stocks are fairly valued, while some are priced much higher compared to earnings.

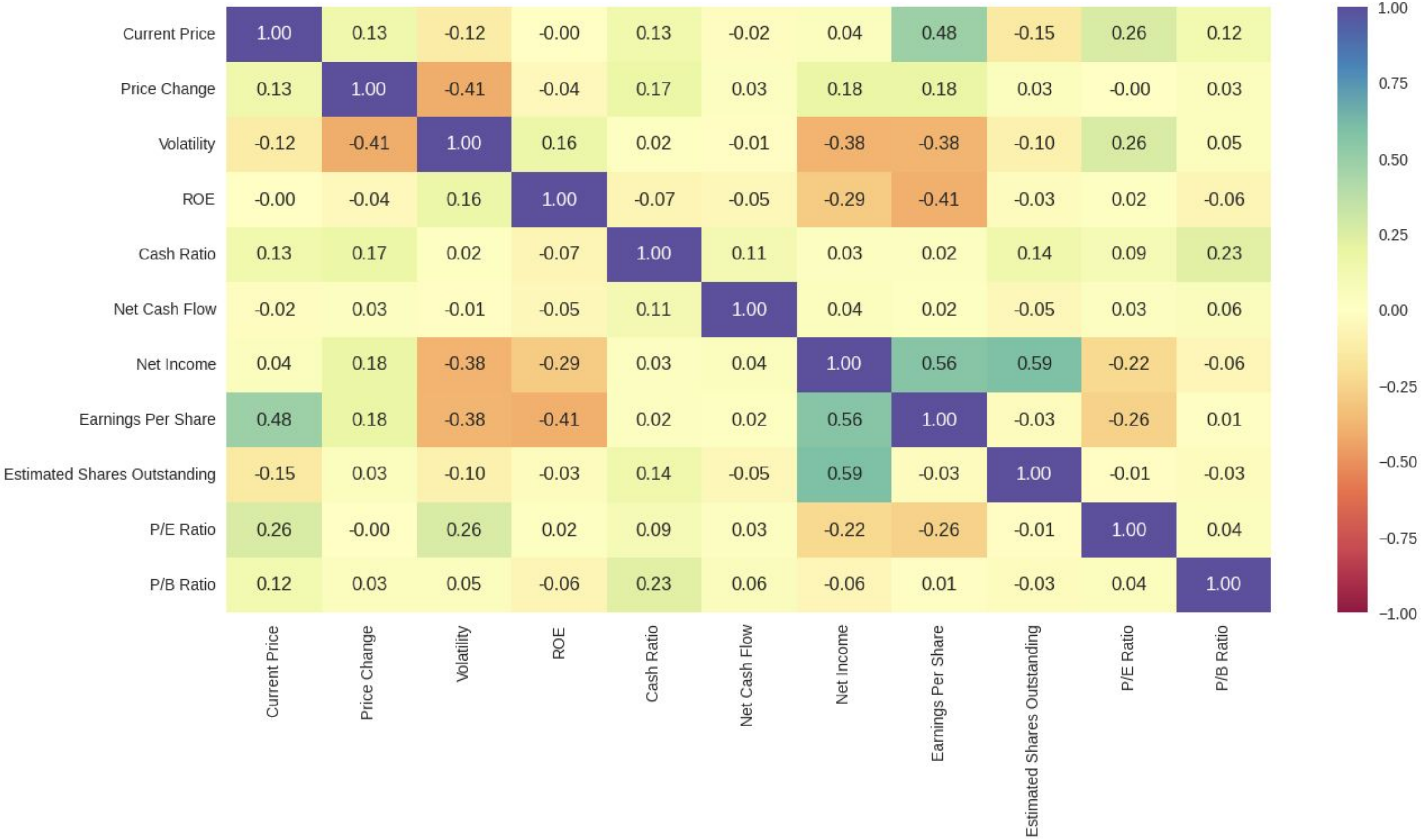
P/B Ratio



Most companies have a P/B ratio close to zero, with a few outliers far from the norm.

Bivariate Analysis

Correlation Check

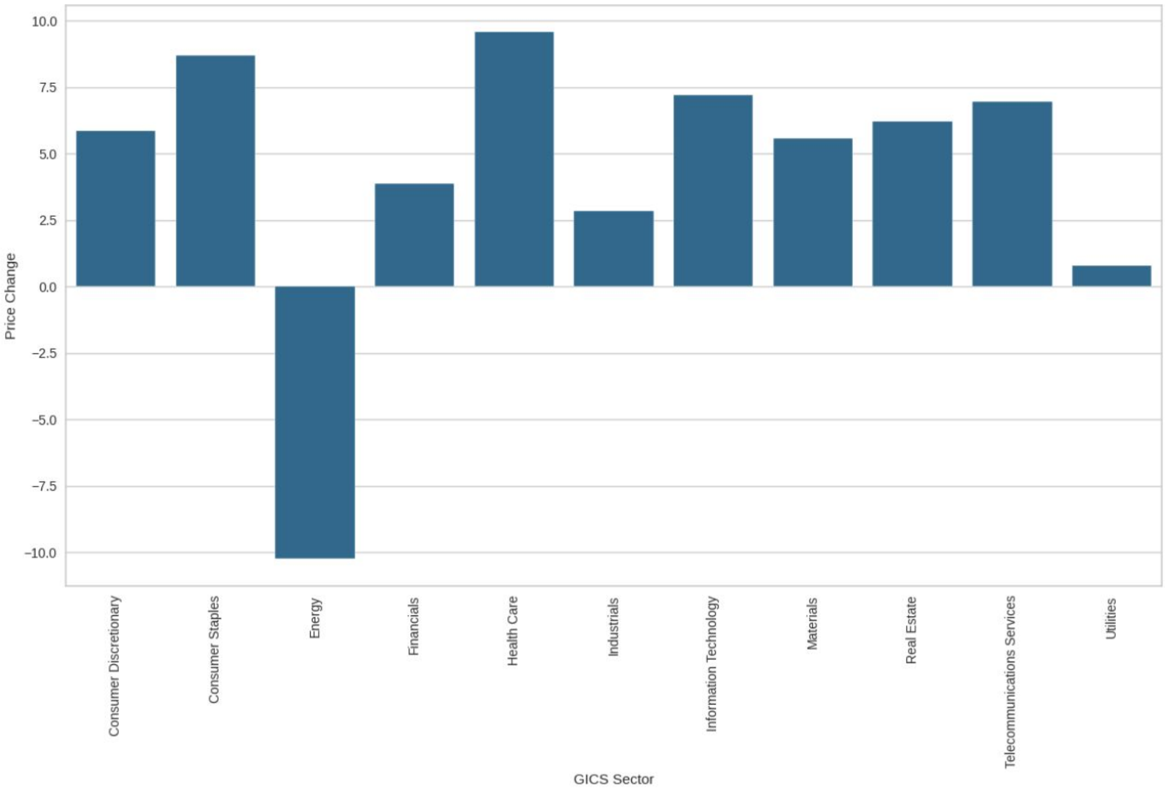


	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio
Current Price	1.00	0.13	-0.12	-0.00	0.13	-0.02	0.04	0.48	-0.15	0.26	0.12
Price Change	0.13	1.00	-0.41	-0.04	0.17	0.03	0.18	0.18	0.03	-0.00	0.03
Volatility	-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

Volatility is negatively correlated with both price change and net income

Net income is positively correlated with EPS and estimated shares outstanding, EPS is positively correlated with current price but negatively correlated with ROE

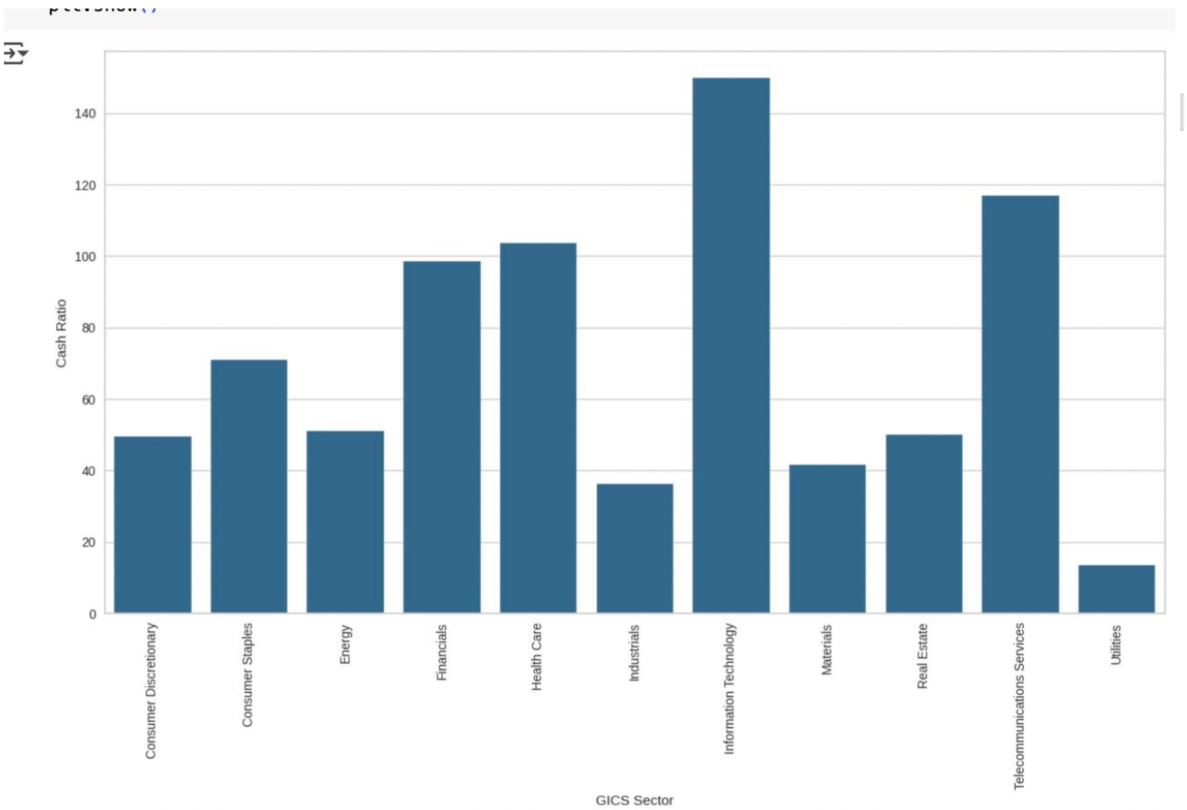
Max Price Increase on Average



Most sectors have seen positive price changes, with Information Technology and

Energy shows significant price drop, indicating a rough patch for that industry.

Avg Cash Ratio Across Economic Sectors

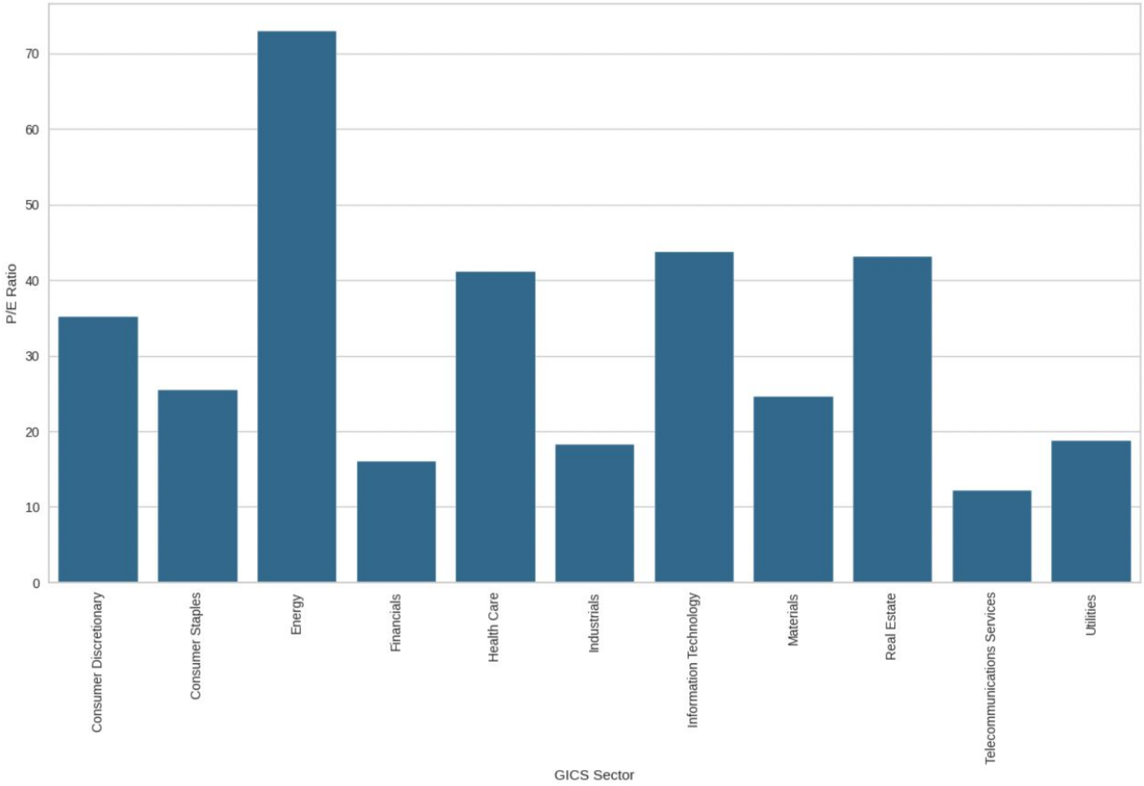


```
plt.show()
```

Information Technology and Telecommunications Services sectors have the highest cash ratios.

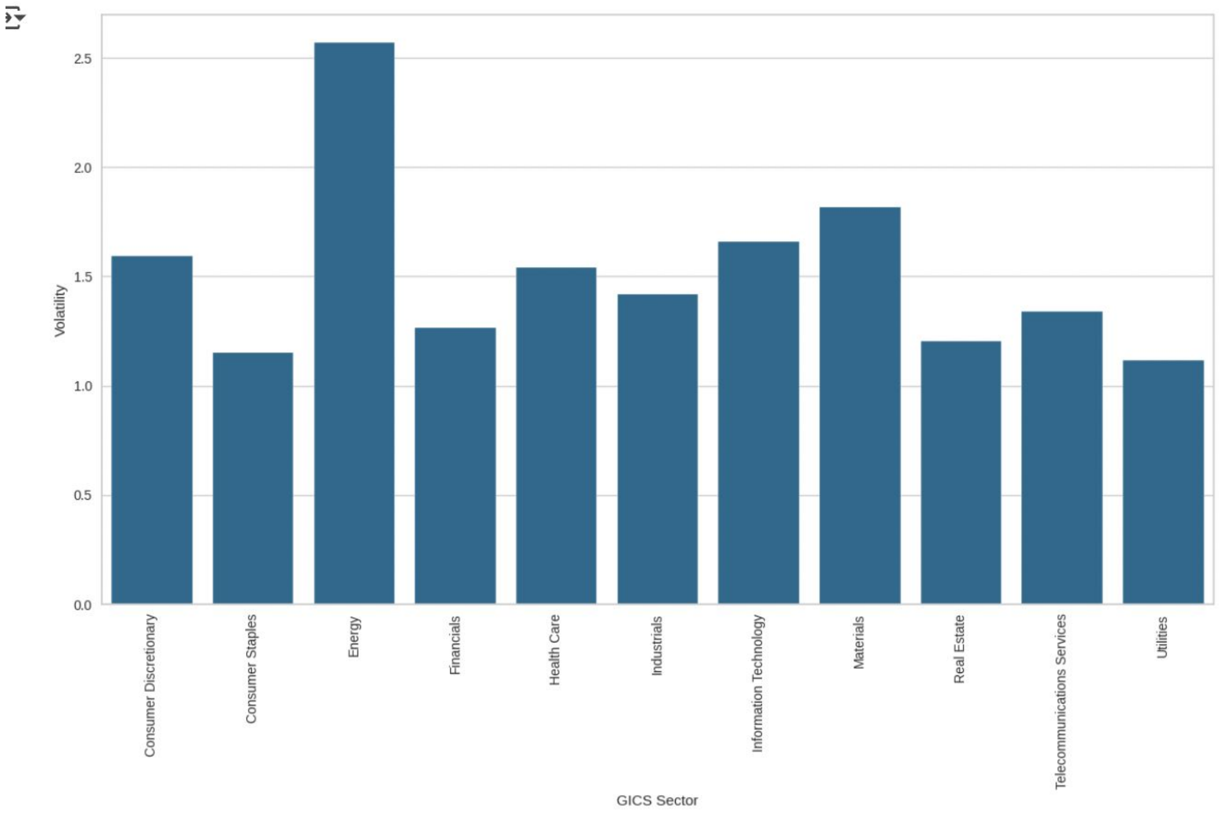
Utilities sector has the lowest cash ratio, indicating tighter cash reserves.

P/E Ratio Across Sectors



Energy sector has the highest P/E ratio, indicating potentially overvalued stocks, while Industrials and Utilities have relatively low P/E ratios, suggesting more conservative valuations.

Volatility Across Sectors



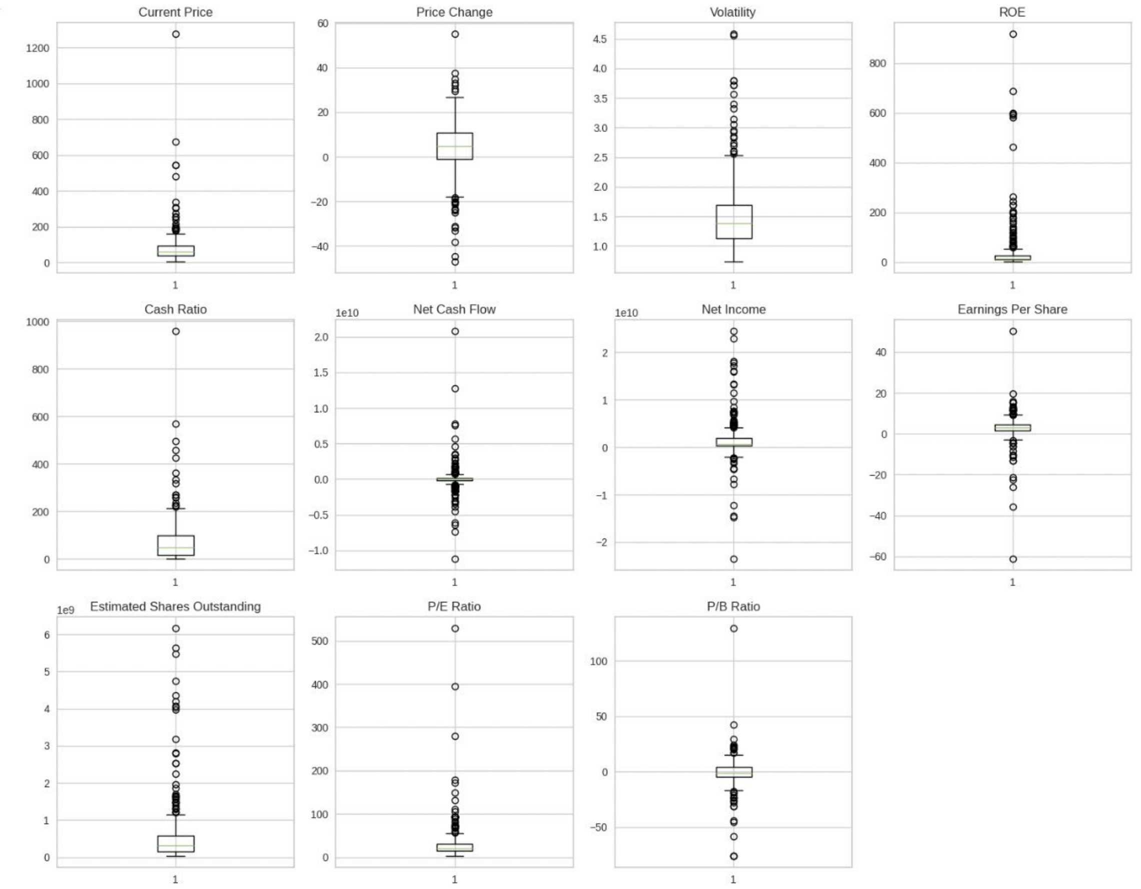
Energy sector has the highest volatility, indicating greater price fluctuations, while Consumer Staples and Utilities have lower volatility, reflecting more stability in their stock prices.

Data Preprocessing

Data Preprocessing

- Zero duplicate values
- Zero missing values

Outlier Check



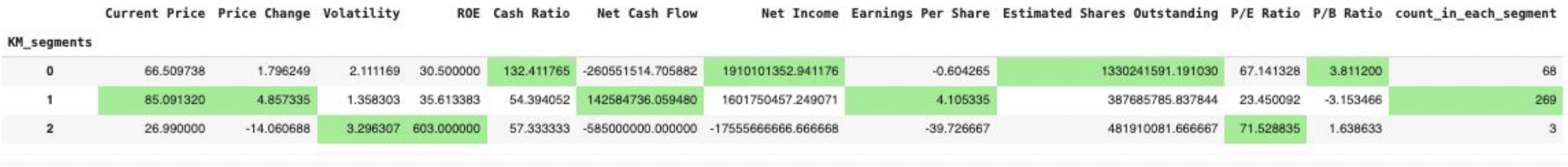
Outliers were present but they were treated.

K-Means Clustering

Summary

Optimal Number of clustering using K-Means: Based on Elbow and Silhouette plots, the number of clusters with the best performance appears to be 3

Cluster Profiling:



KM_segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	count_in_each_segment
0	66.509738	1.796249	2.111169	30.500000	132.411765	-260551514.705882	1910101352.941176	-0.604265	1330241591.191030	67.141328	3.811200	68
1	85.091320	4.857335	1.358303	35.613383	54.394052	142584736.059480	1601750457.249071	4.105335	387685785.837844	23.450092	-3.153466	269
2	26.990000	-14.060688	3.296307	603.000000	57.333333	-585000000.000000	-17555666666.666668	-39.726667	481910081.666667	71.528835	1.638633	3

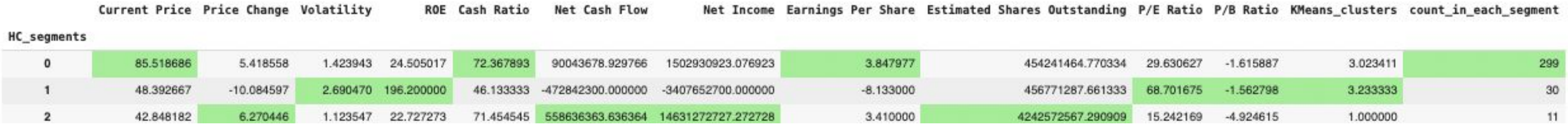
Cluster 0: stable, profitable companies with high net income and strong cash reserves.
Cluster 1: high-performing, stable stocks with significant growth potential and the largest representation of stocks.
Cluster 2: risky stocks with declining prices, low earnings, and a small representation.

Hierarchical Clustering Summary

Summary

Optimal Number of clusters using Hierarchical Clustering: The optimal number of clusters appears to be around 3–4 clusters.

Cluster Profiling



HC_segments	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	KMeans_clusters	count_in_each_segment
0	85.518686	5.418558	1.423943	24.505017	72.367893	90043678.929766	1502930923.076923	3.847977	454241464.770334	29.630627	-1.615887	3.023411	299
1	48.392667	-10.084597	2.690470	196.200000	46.133333	-472842300.000000	-3407652700.000000	-8.133000	456771287.661333	68.701675	-1.562798	3.233333	30
2	42.848182	6.270446	1.123547	22.727273	71.454545	558636363.636364	14631272727.272728	3.410000	4242572567.290909	15.242169	-4.924615	1.000000	11

Cluster 0 has the most stable and high-value stocks with the highest Current Price, Earnings Per Share, and the largest size (299 stocks).
Cluster 1 includes underperforming stocks with high volatility, negative Price Change and Net Income, while Cluster 2 has smaller stocks with notable Net Cash Flow and Estimated Shares Outstanding.

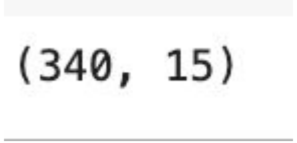
Insights & Conclusions

Insights & Conclusions

- Trade&Ahead should figure out clients' goals, risk tolerance, and how they like to invest to recommend clusters that match their needs.
- Some clusters are basically substitutes for big indexes like the Dow Jones or S&P 500, so they could help clients hit their goals more easily.
- Clusters can be a starting point for digging deeper into financials, especially to spot stocks that don’t quite fit the cluster.
- If clients want to pick individual stocks, Trade&Ahead can look for ones that might beat others in the cluster (buy) or fall behind (sell).
- Over time, these insights could help Trade&Ahead fine-tune their strategies and offer more personalized investment advice.

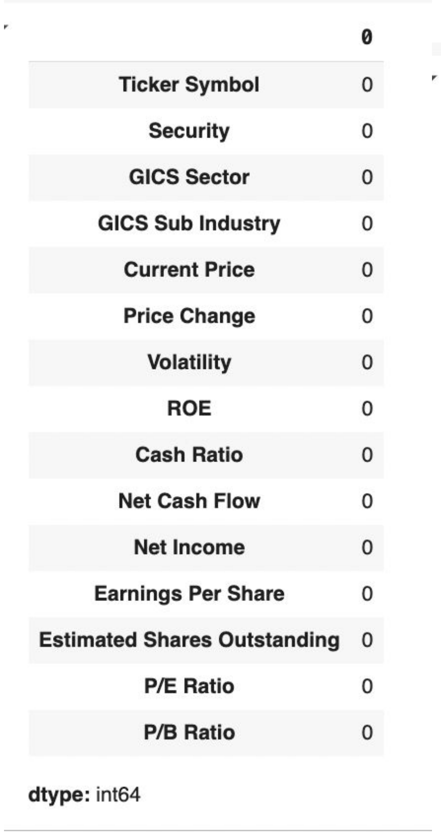
Appendix

Data Background and Contents



```
(340, 15)
```

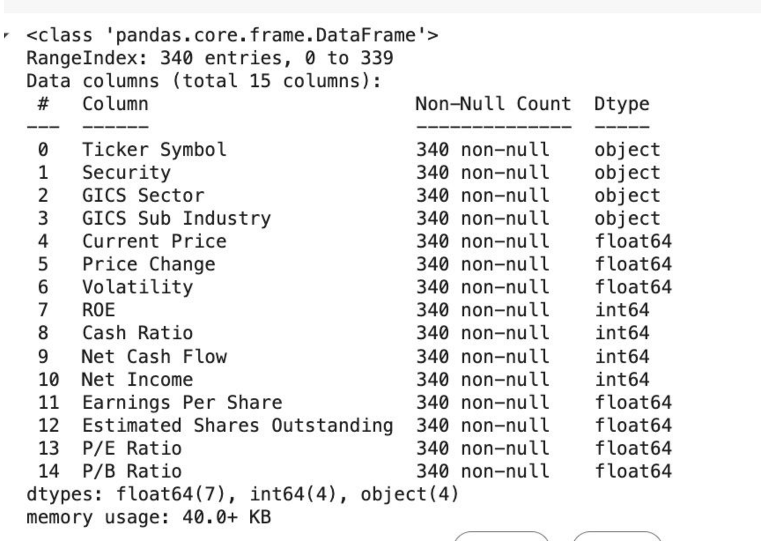
Shape of dataset



	0
Ticker Symbol	0
Security	0
GICS Sector	0
GICS Sub Industry	0
Current Price	0
Price Change	0
Volatility	0
ROE	0
Cash Ratio	0
Net Cash Flow	0
Net Income	0
Earnings Per Share	0
Estimated Shares Outstanding	0
P/E Ratio	0
P/B Ratio	0

dtype: int64

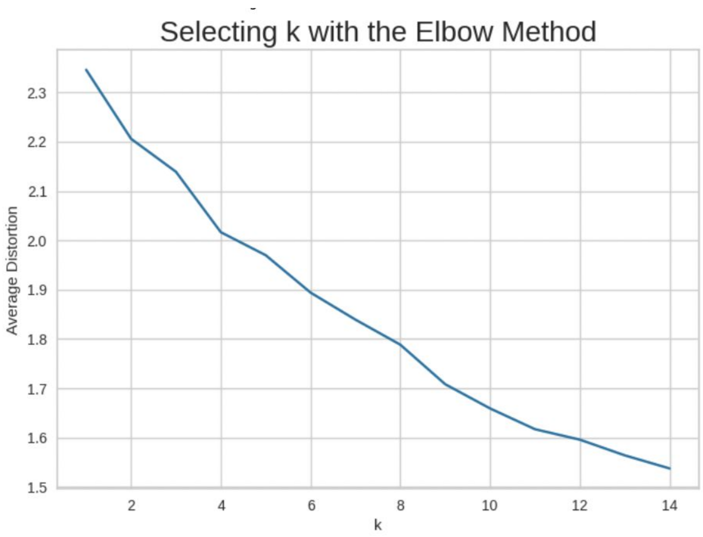
No missing values

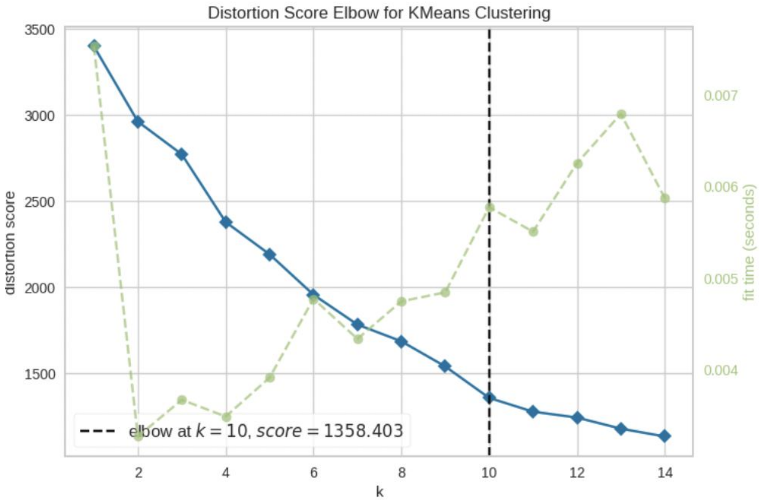


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 340 entries, 0 to 339
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Ticker Symbol                 340 non-null    object 
 1   Security                      340 non-null    object 
 2   GICS Sector                   340 non-null    object 
 3   GICS Sub Industry             340 non-null    object 
 4   Current Price                 340 non-null    float64
 5   Price Change                  340 non-null    float64
 6   Volatility                    340 non-null    float64
 7   ROE                           340 non-null    int64  
 8   Cash Ratio                    340 non-null    int64  
 9   Net Cash Flow                 340 non-null    int64  
 10  Net Income                    340 non-null    int64  
 11  Earnings Per Share            340 non-null    float64
 12  Estimated Shares Outstanding  340 non-null    float64
 13  P/E Ratio                     340 non-null    float64
 14  P/B Ratio                     340 non-null    float64
dtypes: float64(7), int64(4), object(4)
memory usage: 40.0+ KB
```

Data types & columns

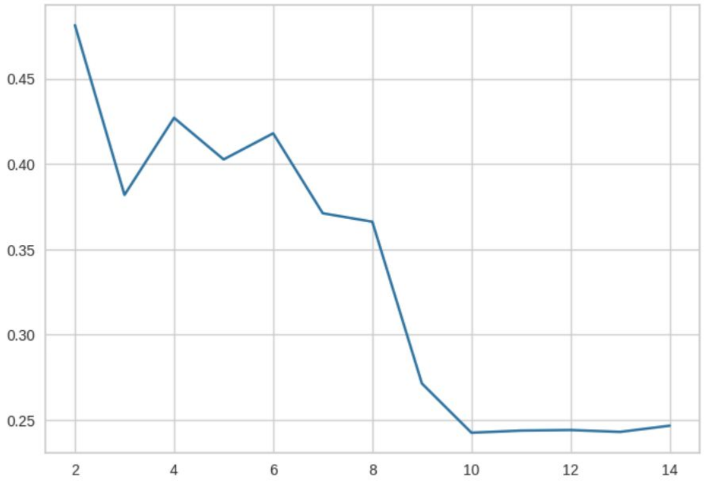
K-Means Clustering Technique

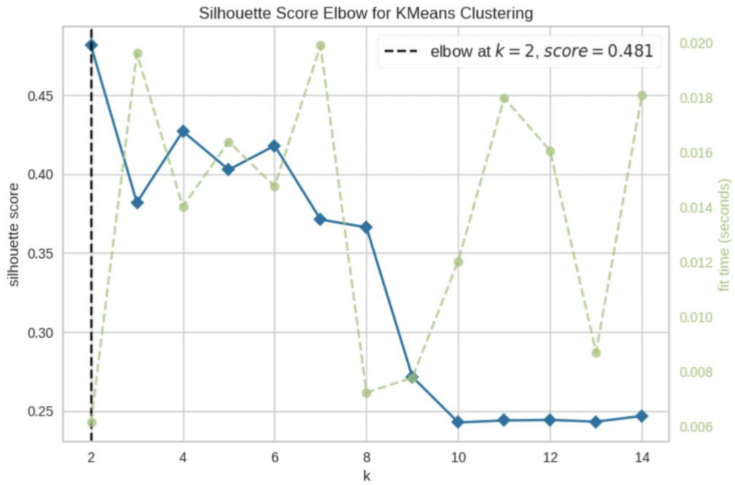


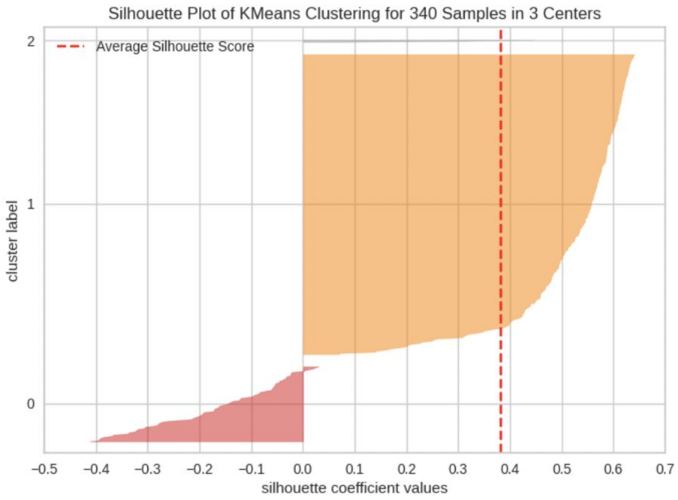


Looks like k=10 is a good choice for the number of clusters since the distortion score levels off there.

The second chart backs this up, showing a clear "elbow" at k=10, so it feels like the sweet spot.





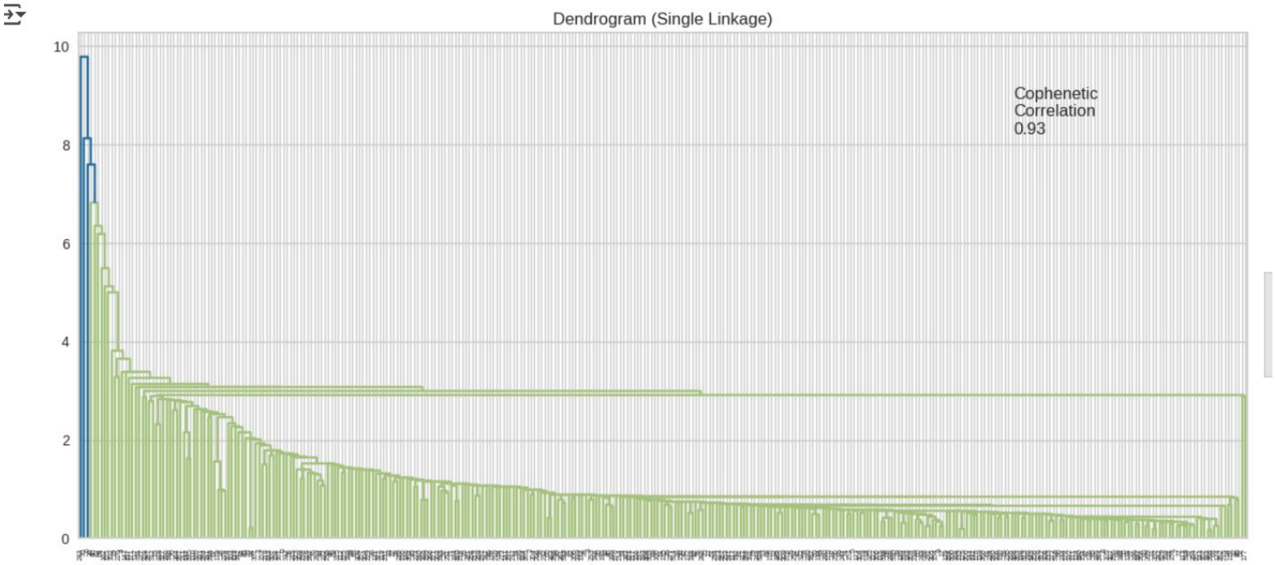


The silhouette score shows that 2 clusters work best, with the highest score and clear separation between groups.

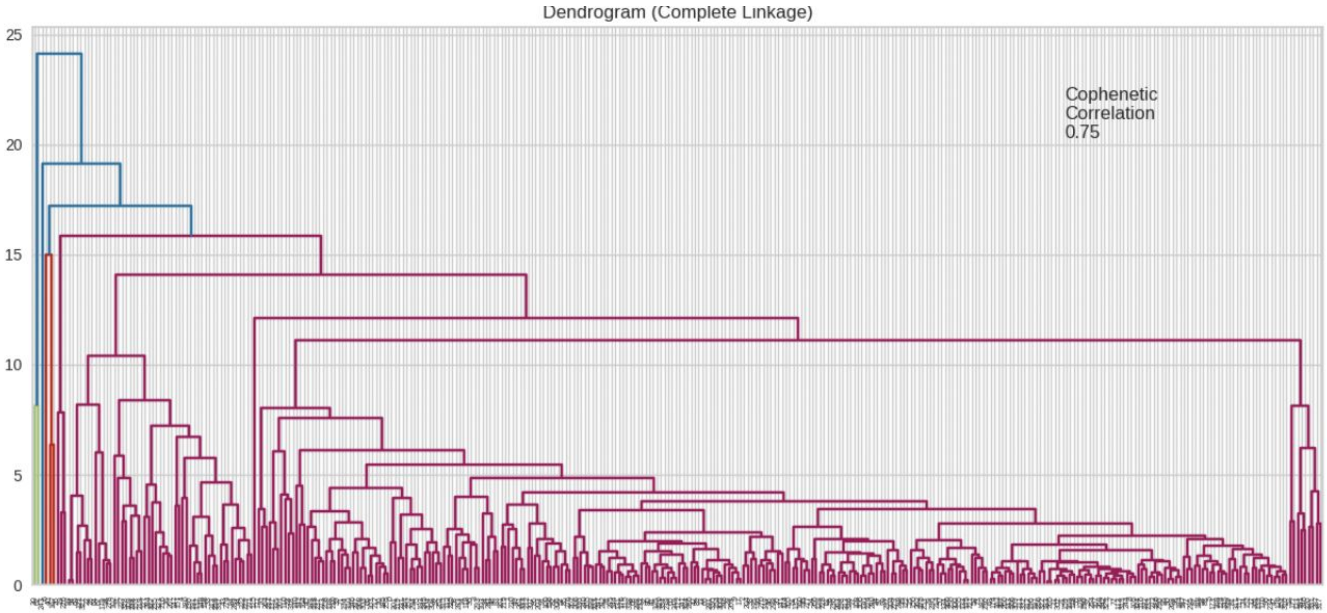
After 2 clusters, the quality drops, so adding more clusters doesn’t improve much.

Most data points fit nicely into their clusters with 2 groups, making it the best choice.

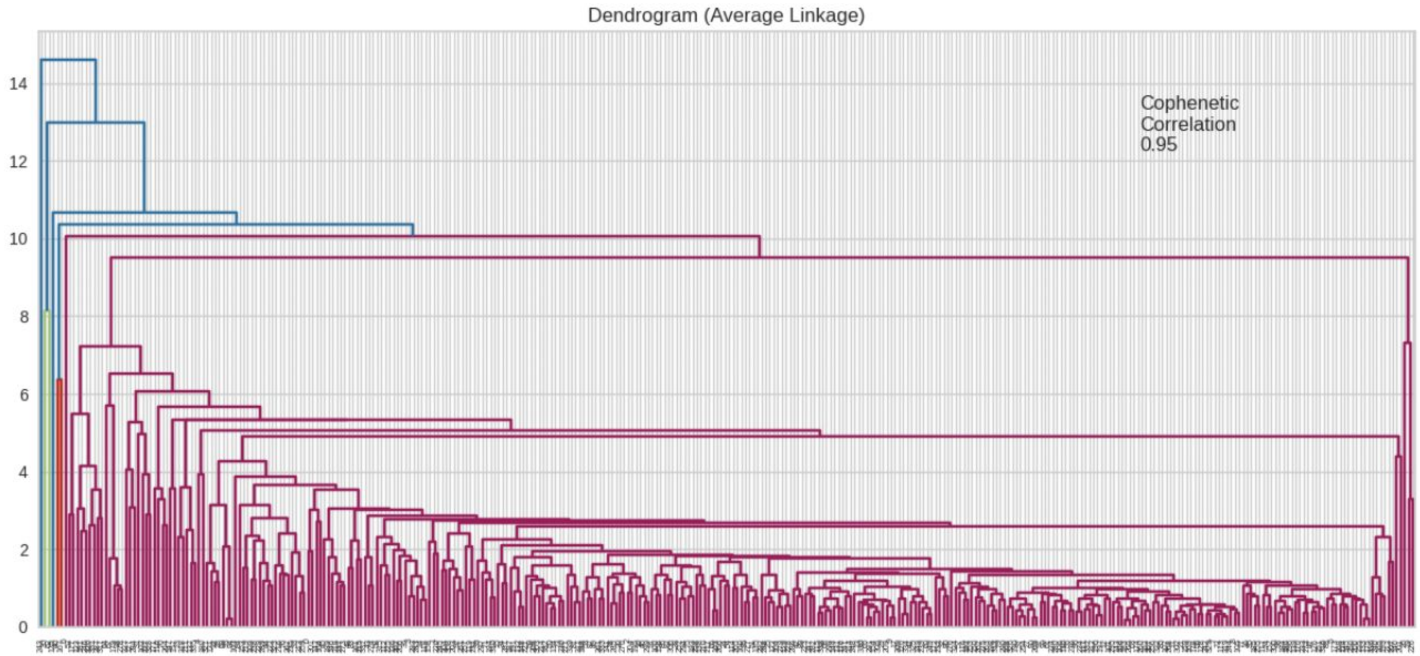
Hierarchical Clustering Technique



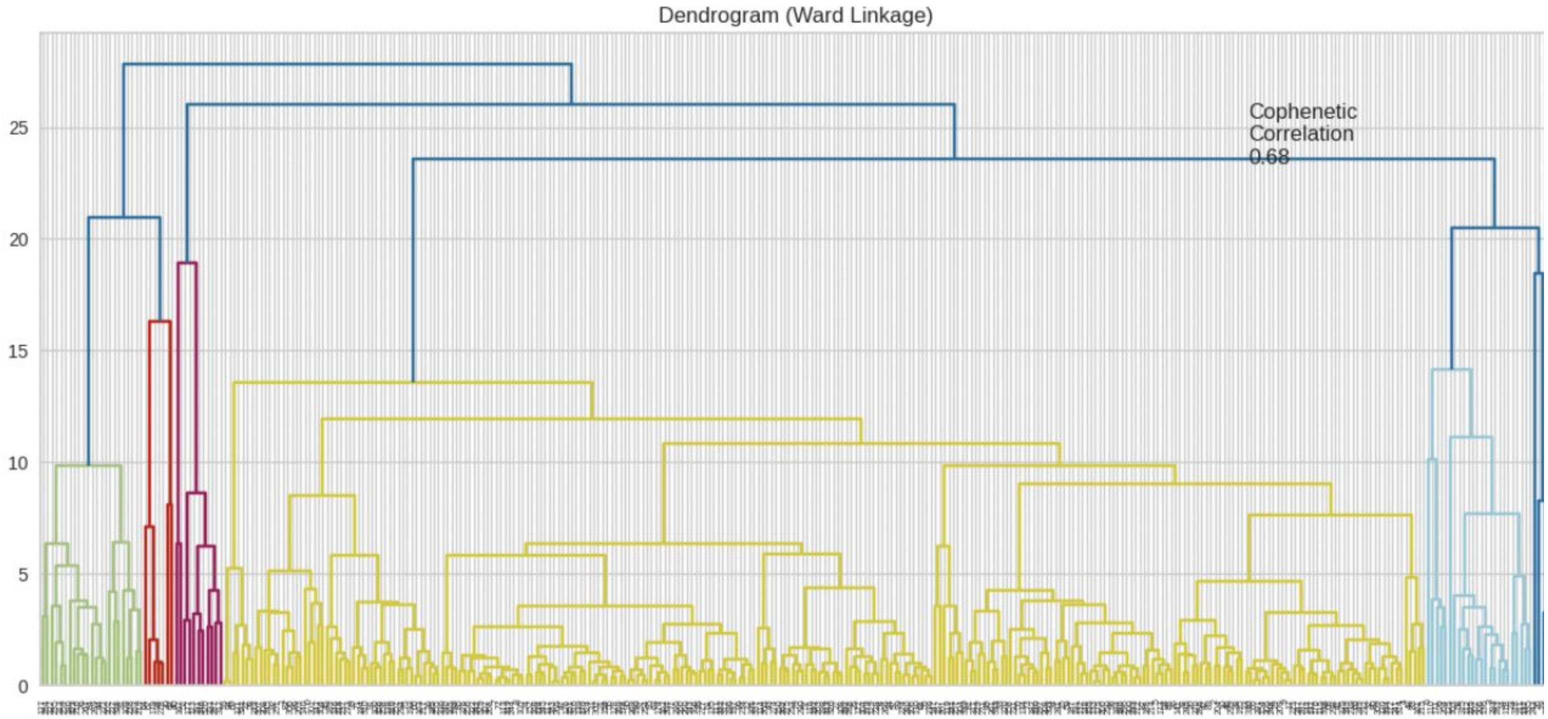
This dendrogram shows single linkage clustering with a solid fit (cophenetic correlation of 0.93). To find meaningful clusters, you can cut the tree at a point where clear groupings stand out.



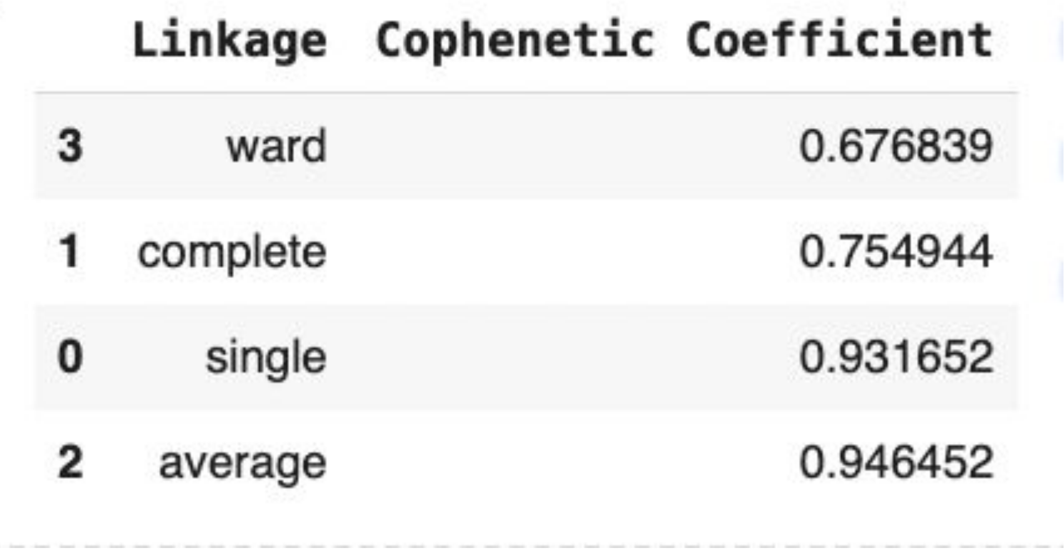
This dendrogram uses complete linkage clustering with a lower cophenetic correlation of 0.75. It’s not as strong of a fit, but still useful



This dendrogram uses average linkage clustering with a strong cophenetic correlation of 0.95, suggesting a very good fit.



Cophenetic correlation of 0.68,
which is moderate.



	Linkage	Cophenetic Coefficient
3	ward	0.676839
1	complete	0.754944
0	single	0.931652
2	average	0.946452

Among the linkages, "average" (0.946) and "single" (0.931) perform best, meaning they are more reliable for hierarchical clustering in this dataset. "Ward" linkage has the lowest coefficient (0.677), indicating it may not be as effective

K-Means vs Hierarchical Clustering

Which clustering technique took less time for execution?

- Both models fit the dataset within less than 0.3s

Which clustering technique gave you more distinct clusters, or are they the same? How many observations are there in the similar clusters of both algorithms?

- K-Means seems to give clearer and more distinct clusters compared to Hierarchical Clustering, making it a bit better for this dataset. Both methods found similar patterns, like identifying high and low-performing stocks. The number of observations in the biggest clusters was the same (e.g., 299).

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

- Both algorithms agree on **3 clusters** being the best fit for this dataset.

Differences or similarities in the cluster profiles from both the clustering techniques

- Both algorithms identify 3 clusters as the optimal number.
- The overall cluster characteristics are consistent between the two methods
- No differences present