Business Report Structure (Checklist Based)

Title Page

- Title: *"Business Report: Stock Clustering for Trade&Ahead"*
- Author: Harik Charan

Table Of Contents:

1.Probelm Statement:

Investing in the stock market has long been recognized as a powerful tool for wealth creation, fighting inflation, and achieving long-term financial goals. However, the vast number of publicly traded companies and the complexity of financial indicators make stock selection a challenging task for investors.

To address this, data-driven clustering methods can be used to group stocks with similar characteristics. This aids in building a well-diversified portfolio, reducing risk exposure, and making personalized investment decisions.

2.0bjective:

Trade & Ahead, a financial consultancy firm, has commissioned this project to help their clients better understand the stock market landscape. You have been hired as a **Data Scientist** to analyze and segment a dataset comprising various financial metrics of companies listed on the New York Stock Exchange (NYSE).

Your specific tasks are:

- Analyze the dataset and perform appropriate preprocessing.
- Apply clustering techniques (like K-Means and Hierarchical Clustering).
- Identify the optimal number of clusters using suitable evaluation metrics.
- Interpret and describe the financial characteristics of each cluster.
- Provide actionable insights for building diversified investment portfolios.

3. Dataset Overview:

Source

The dataset has been provided by Trade & Ahead, consisting of 340 companies listed on the New York Stock Exchange (NYSE). It includes a range of financial and market performance indicators designed to support clustering-based analysis for investment strategy.

Structure

- **Total Rows:** 340
- **Total Columns:** 15
- **Data Types:** Numeric and categorical
- **Missing Values:** None
- **Duplicates:** None

Feature Name	Description
Ticker Symbol	Unique symbol used to identify the company in stock exchanges
Security	Full name of the company
GICS Sector	Sector classification as per Global Industry Classification Standard
GICS Sub Industry	More specific sub-industry classification under GICS
Current Price	Current stock price in USD
Price Change	Percentage price change over the past 13 weeks
Volatility	Standard deviation of the stock price in the past 13 weeks
ROE	Return on Equity: a profitability ratio

Feature Name	Description
Cash Ratio	Liquidity ratio: cash & equivalents divided by current liabilities
Net Cash Flow	Net cash generated (in dollars)
Net Income	Profit after taxes and expenses (in dollars)
Earnings Per Share (EPS)	Net income divided by total outstanding shares
Estimated Shares Outstanding	Total number of shares available in the market
P/E Ratio	Price-to-Earnings ratio
P/B Ratio	Price-to-Book value ratio

Statistical Summary (Numeric Columns)

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Ticker Symbol	340	340	ZTS	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Security	340	340	Zoetis	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GICS Sector	340	11	Industrials	53	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GICS Sub Industry	340	104	Oil & Gas Exploration & Production	16	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Current Price	340.0	NaN	NaN	NaN	80.862345	98.055086	4.5	38.555	59.705	92.880001	1274.949951
Price Change	340.0	NaN	NaN	NaN	4.078194	12.006338	-47.129693	-0.939484	4.819505	10.695493	55.051683
Volatility	340.0	NaN	NaN	NaN	1.525976	0.591798	0.733163	1.134878	1.385593	1.695549	4.580042
ROE	340.0	NaN	NaN	NaN	39.597059	96.547538	1.0	9.75	15.0	27.0	917.0
Cash Ratio	340.0	NaN	NaN	NaN	70.023529	90.421331	0.0	18.0	47.0	99.0	958.0
Net Cash Flow	340.0	NaN	NaN	NaN	55537620.588235	1946365312.175789	-11208000000.0	-193906500.0	2098000.0	169810750.0	20764000000.0
Net Income	340.0	NaN	NaN	NaN	1494384602.941176	3940150279.327937	-23528000000.0	352301250.0	707336000.0	1899000000.0	24442000000.0
Earnings Per Share	340.0	NaN	NaN	NaN	2.776662	6.587779	-61.2	1.5575	2.895	4.62	50.09
Estimated Shares Outstanding	340.0	NaN	NaN	NaN	577028337.75403	845849595.417695	27672156.86	158848216.1	309675137.8	573117457.325	6159292035.0
P/E Ratio	340.0	NaN	NaN	NaN	32.612563	44.348731	2.935451	15.044653	20.819876	31.764755	528.039074
P/B Ratio	340.0	NaN	NaN	NaN	-1.718249	13.966912	-76.119077	-4.352056	-1.06717	3.917066	129.064585

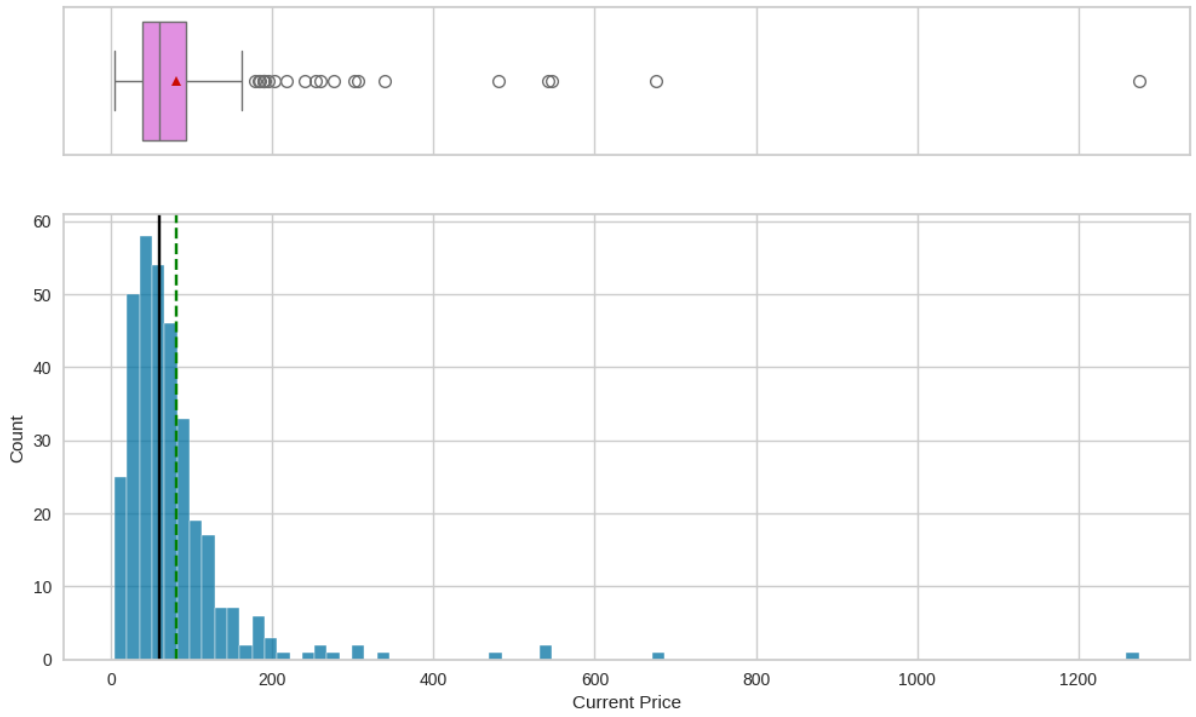
There are no missing values in data.

4. Exploratory Data Analysis:

Univariate Analysis:

Boxplot of Histogram:

i.)Current Price

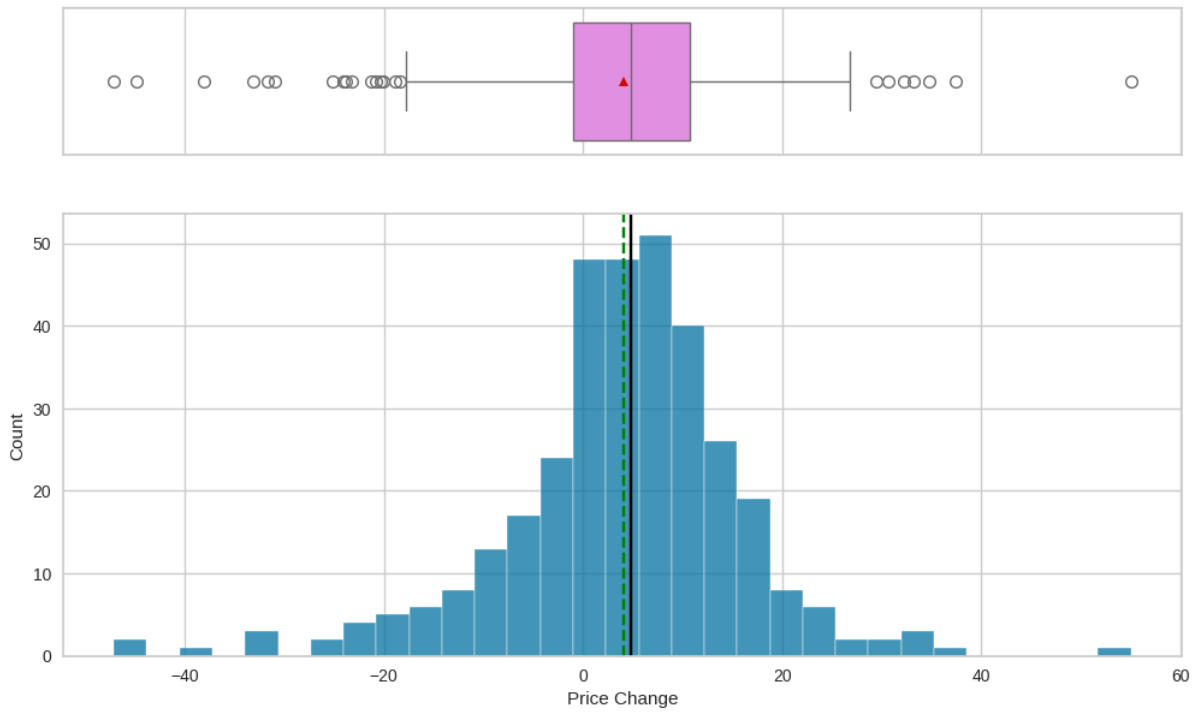


- **Boxplot Insights**:
  - The median current price is relatively low compared to the upper range.
  - There are many outliers above the upper whisker, with some extreme values beyond $600 and even $1200.

- The red triangle indicates the mean, which is higher than the median, suggesting positive skewness.

- **Histogram Insights**:
  - The majority of stocks are priced between $0 and $100.
  - A right-skewed distribution is clearly observed, with the tail extending towards higher prices.
  - A vertical green dashed line represents the mean, confirming it lies above the bulk of the data.

ii.) Price Change:



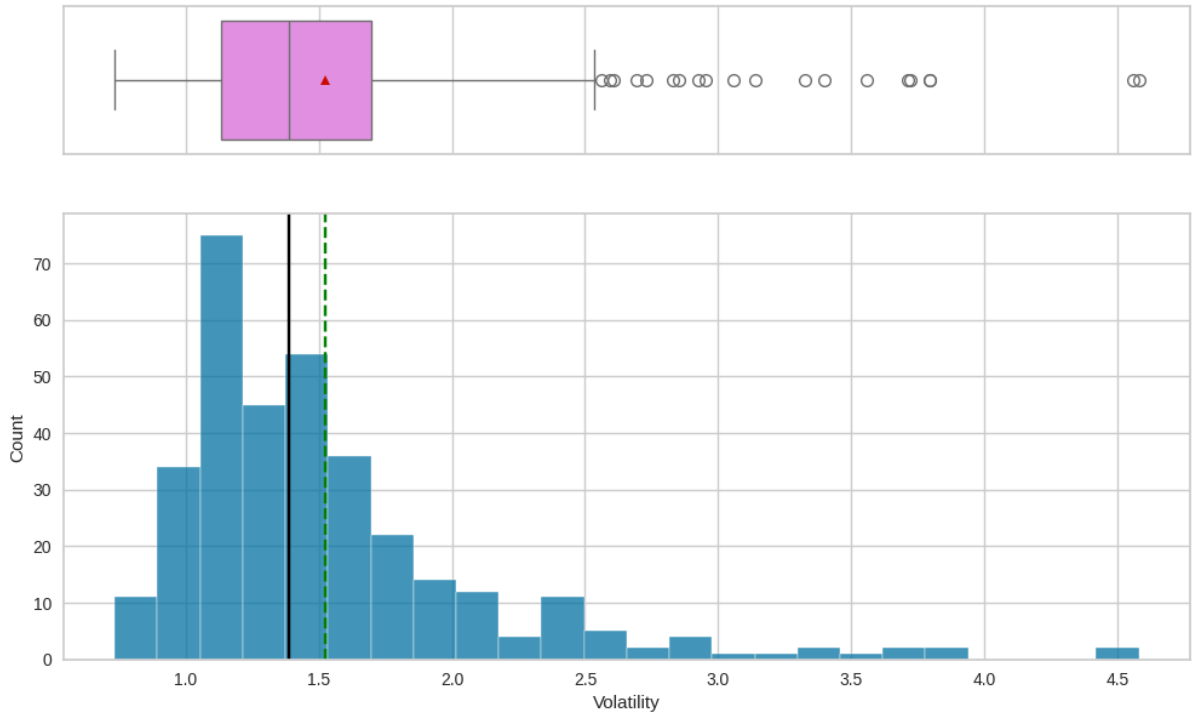
Boxplot Insights:

- The interquartile range (IQR) is mostly concentrated between slight negative and moderate positive values.
- Outliers exist on both ends, more prominently in the negative range (e.g., stocks with >–40% drop).
- The mean (red triangle) is close to the median, indicating a roughly symmetrical distribution with mild positive skew.

Histogram Insights:

- The majority of stocks have had a small to moderate price increase (0% to +10%).
- A notable number of stocks also experienced minor losses (0% to –10%).
- Very few stocks had extreme losses or gains, but they exist and are marked in the tails.
- The green dashed line (mean) is slightly to the right of the center, showing a slight positive bias in returns.

iii.)Volatility



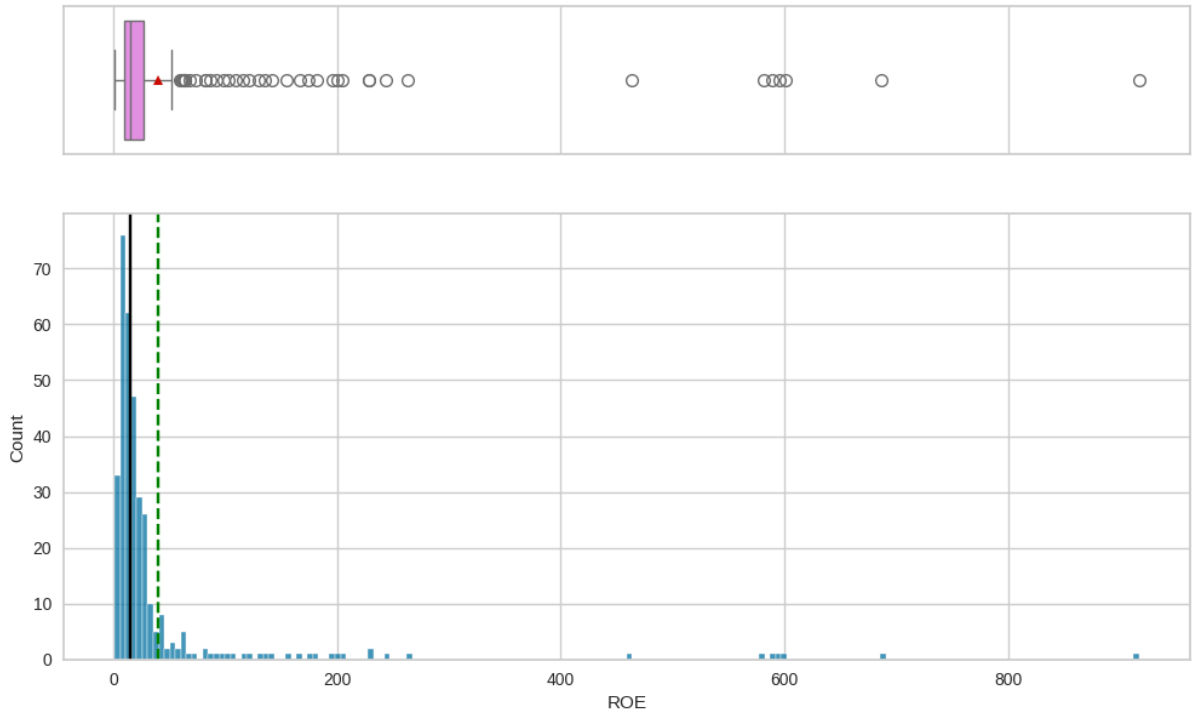
Boxplot Insights:

- Distribution shows a moderate positive skew.
- A few companies exhibit very high volatility (>3), marked as outliers.
- The mean (red triangle) is slightly above the median.

Histogram Insights:

- Most stocks have volatility in the 1.0 to 1.5 range.
- The distribution is right-skewed, with a long tail of high-volatility stocks.

iv.)Return on Equity (ROE)



Boxplot Insights:

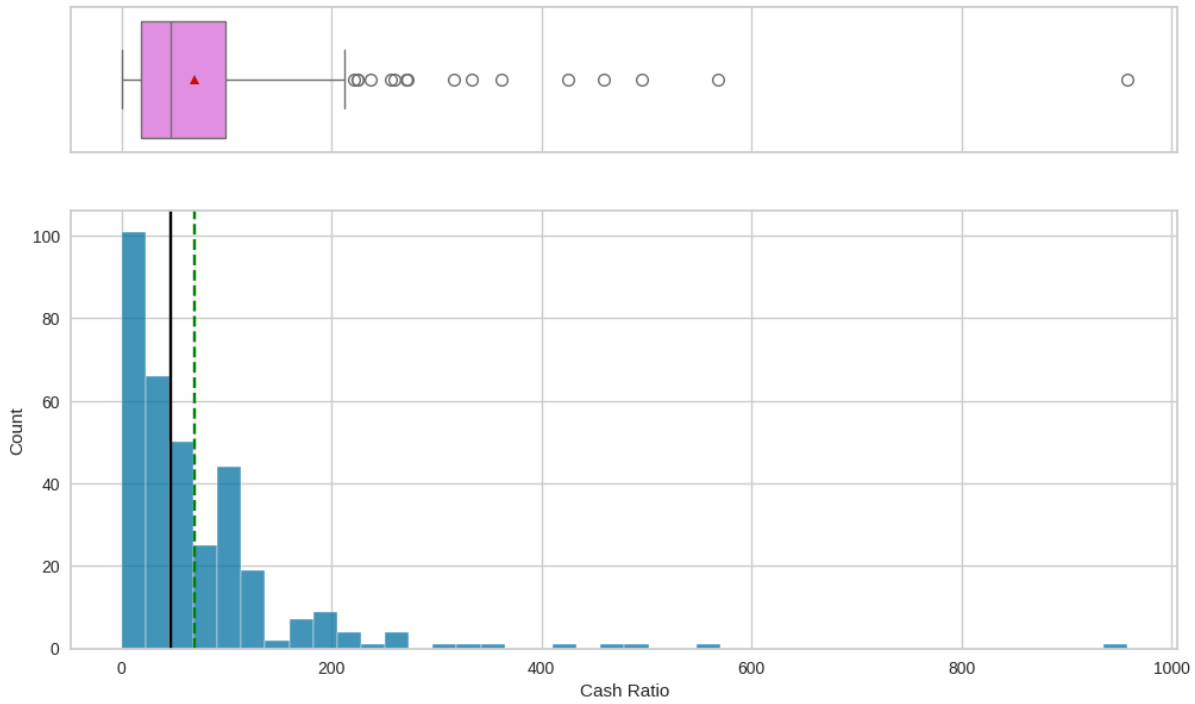
- Extremely right-skewed with massive outliers (ROE > 800).
- Majority of stocks have ROE below 100, concentrated in the lower end.
- Median and IQR are tightly packed, but high-value outliers lift the mean.

Histogram Insights:

- Most companies have ROE between 0 and 50%.

- A long tail on the right signifies rare but extreme profitability.

V.)Cash Ratio



Boxplot Insights:

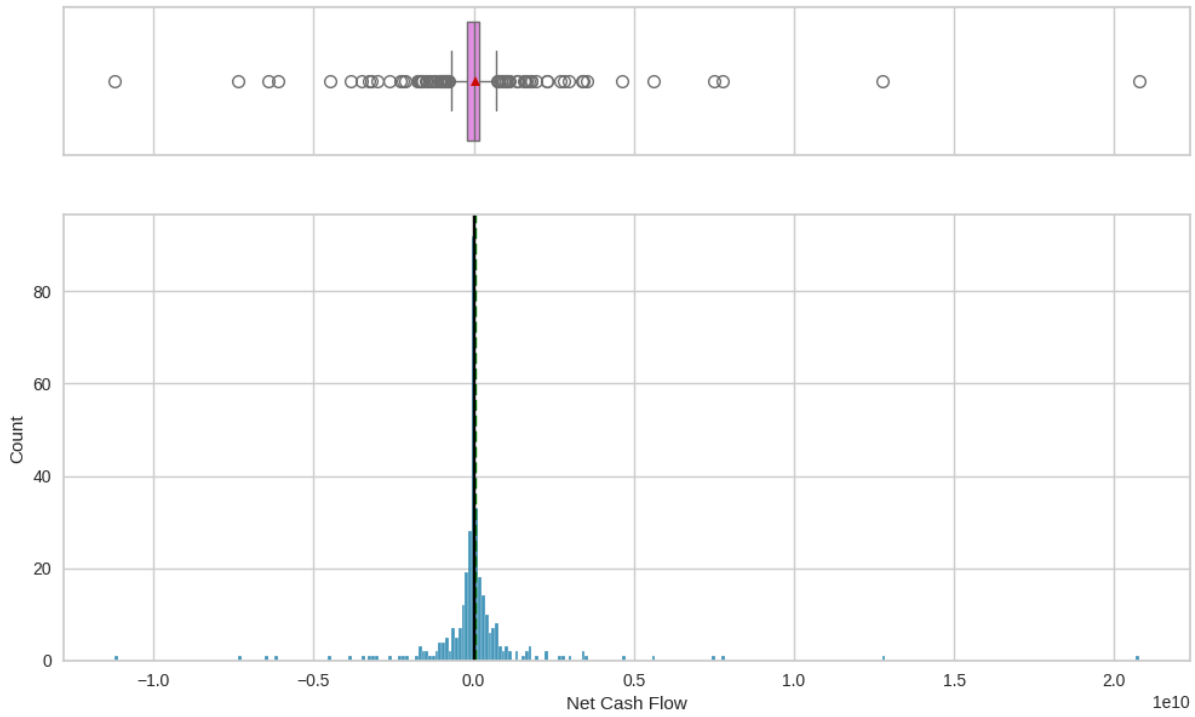
- Large number of outliers beyond 100, some approaching 1000.
- Central bulk lies under 50, with a slightly high mean due to long tail.

Histogram Insights:

- Most companies have cash ratios between 0 and 50, with frequency rapidly decreasing.

- Right-skewed distribution with rare extremely cash-rich firms.

vi.)Net Cash Flow



Boxplot Insights:

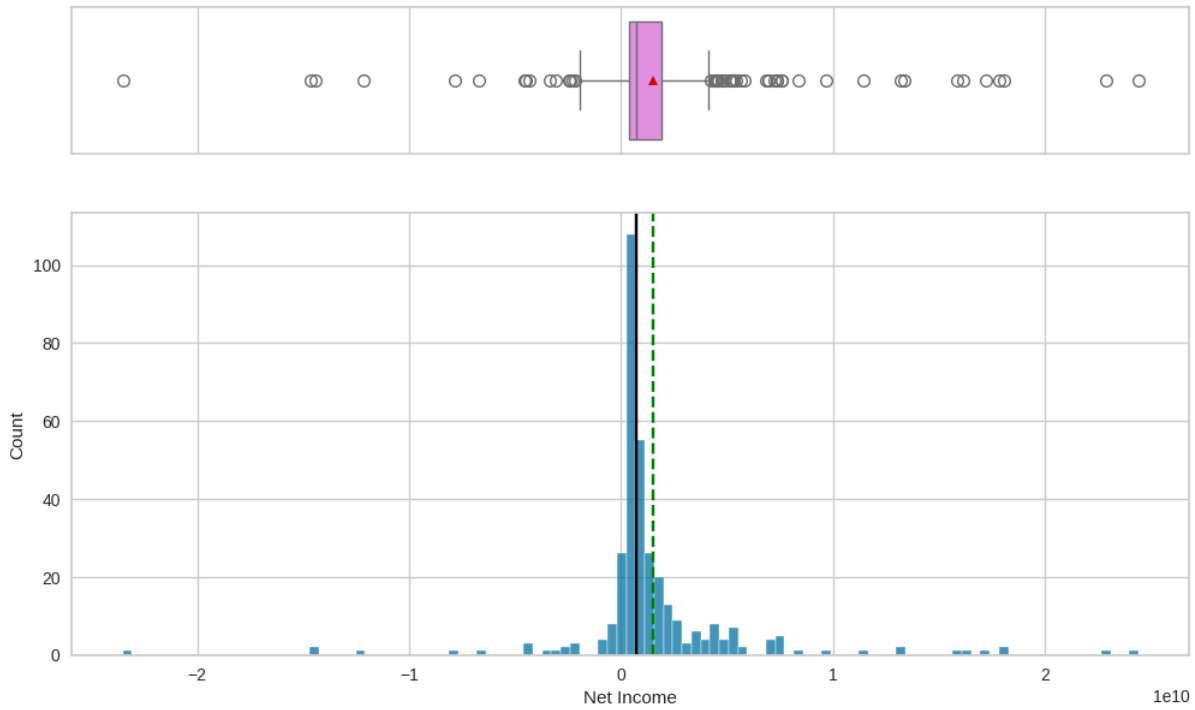
- Symmetrical box centered around 0, but with many outliers on both sides.
- Mean and median are very close, indicating symmetry.

Histogram Insights:

- Very narrow, tall peak near zero, tapering off on both sides.

- Outliers exist both in the positive and negative ends.

vii.)Net Income



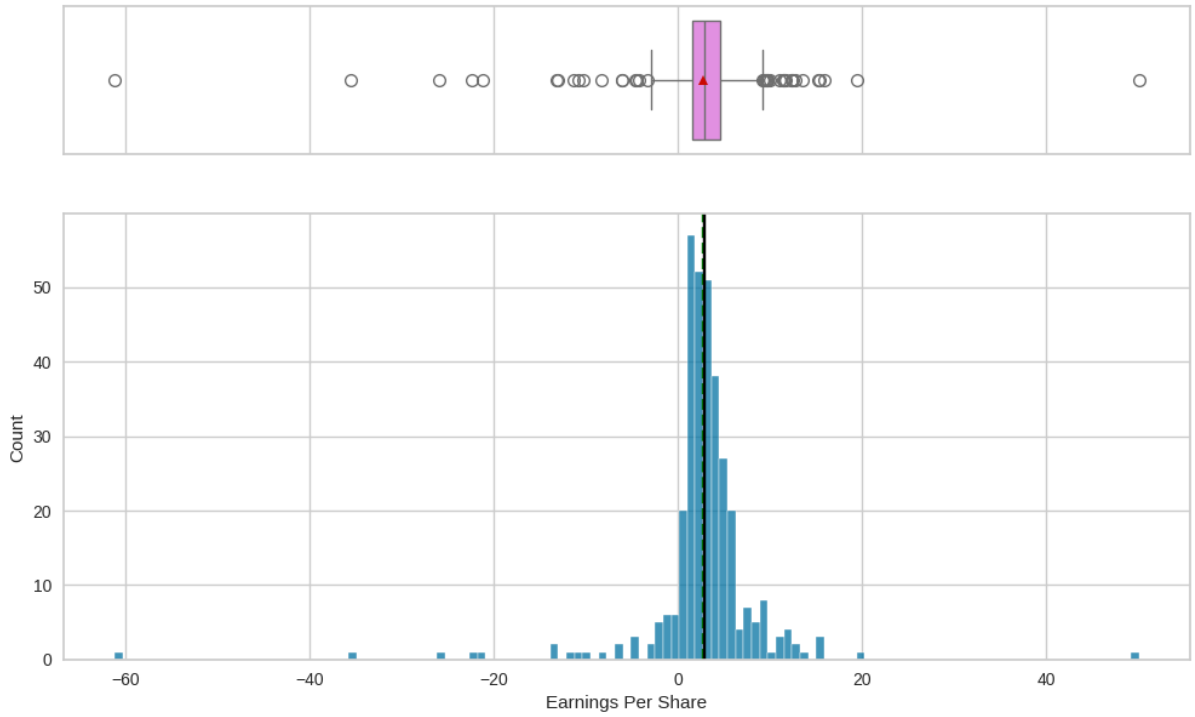
Boxplot Insights:

- Large number of outliers on both sides (losses and gains).
- Mean is above median, indicating right-skewness.

Histogram Insights:

- Most companies earn between $0 and $2B.
- The distribution has a long right tail, driven by highly profitable firms.

viii.)Earnings Per Share (EPS)



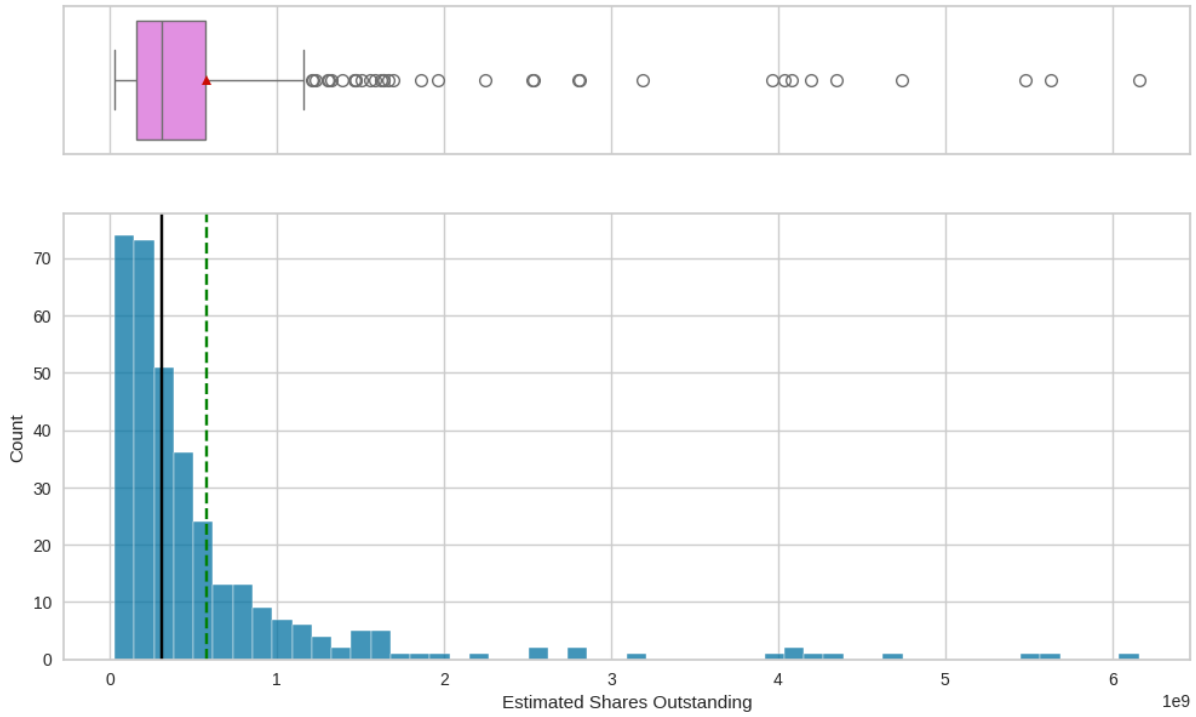
Boxplot Insights:

- Outliers are present on both positive and negative ends.
- Most EPS values are between 0 and 10.

Histogram Insights:

- Clear bell shape centered near zero, with minor skewness.
- Long tails imply a few firms with extreme EPS.

ix.)Estimated Shares Outstanding



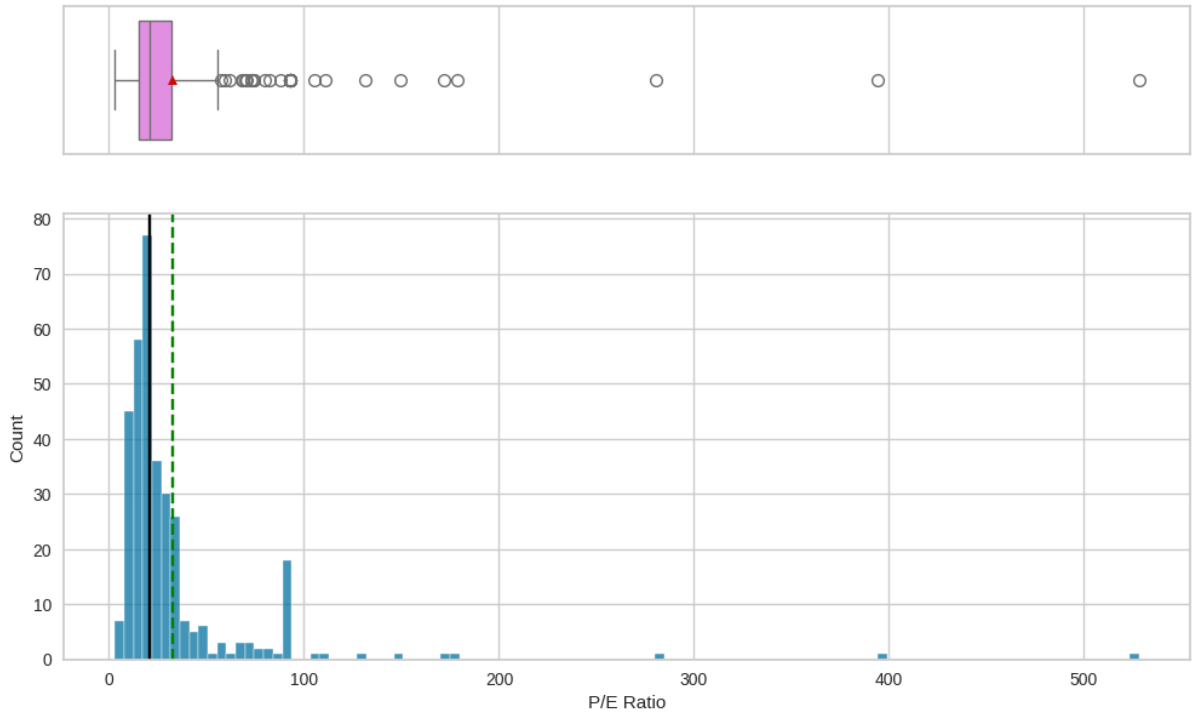
Boxplot Insights:

- Strong right skew, with many large-cap companies as outliers.
- Majority of companies have fewer than 1 billion shares.

Histogram Insights:

- Most companies are tightly clustered in the lower range, but a long tail of high values exists.

x.)P/E Ratio



Box Plot:

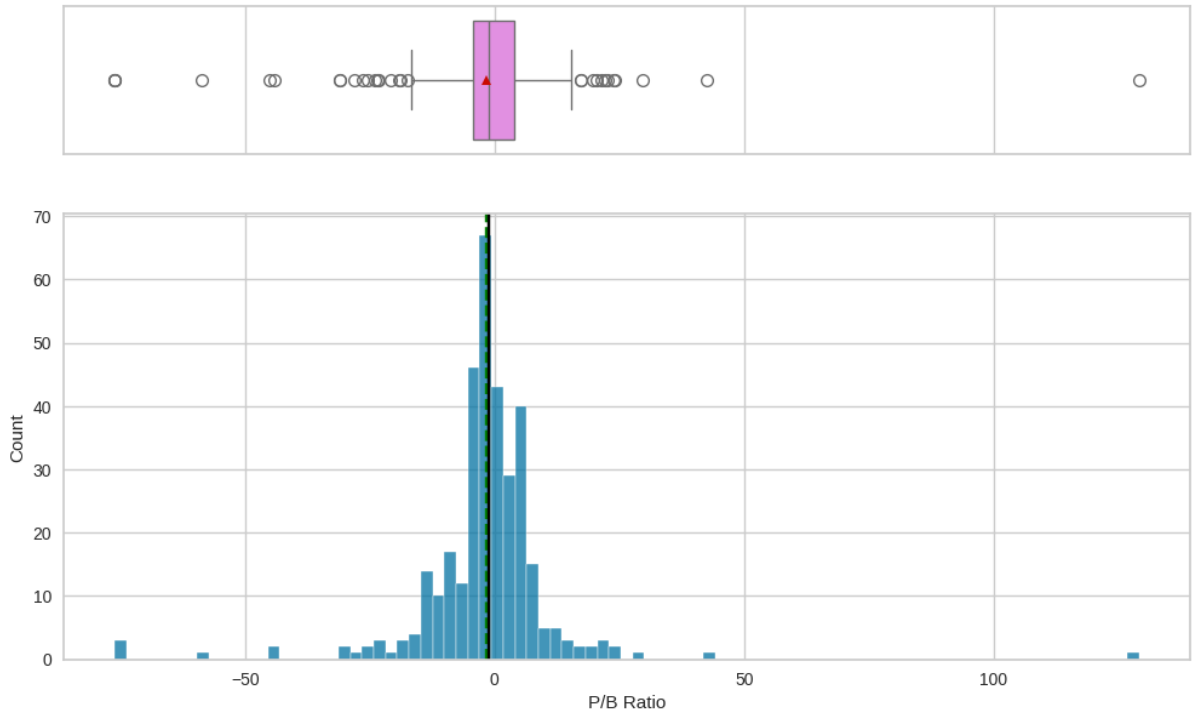
- Shows a positively skewed distribution with many outliers.
- The median is marked by a red triangle and lies near the lower end, indicating most values are small.
- Outliers stretch beyond 500, highlighting extreme P/E values in some stocks.

Histogram:

- Most P/E ratios are concentrated below 50.

- A green dashed line marks the mean, which is higher than the median—indicating right skewness.
- A few very high values (outliers) heavily influence the mean.

xi.)P/B Ratio



Box Plot:

- The P/B ratio distribution has many outliers on both the negative and positive sides.
- The box is centered around 0, indicating a large number of values are close to zero.

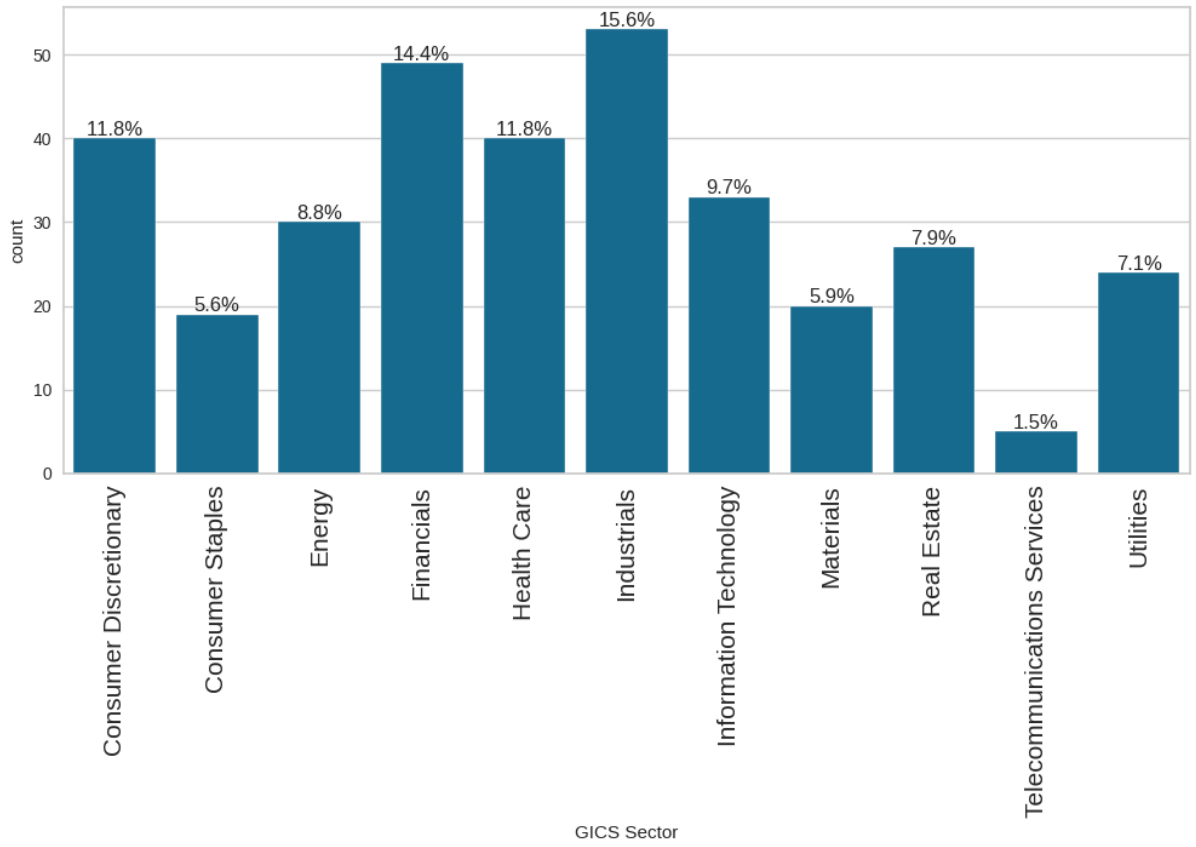
- Red triangle marks the median, which is near zero.
- Presence of both negative and large positive outliers suggests data irregularities or extreme valuations.

Histogram:

- The distribution is centered tightly around 0, with most values between approximately -5 and 5.
- A black line (median) and a green dashed line (mean) are nearly overlapping at 0, suggesting symmetry in the central distribution.
- However, extreme outliers on both sides (up to ±100) cause heavy tails.

Labeled Barplot

i.)GICS Sector



This bar chart shows the distribution of stocks across GICS sectors.

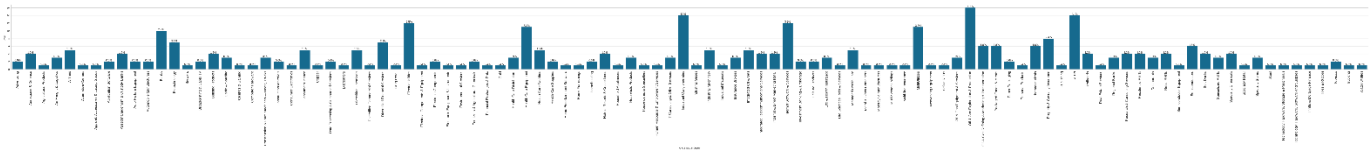
Top 3 sectors by count:

- Industrials – 15.6%
- Financials – 14.4%
- Consumer Discretionary & Health Care – each 11.8%

Least represented:

- Telecommunications Services – 1.5%

ii.)GICS Sub Industry

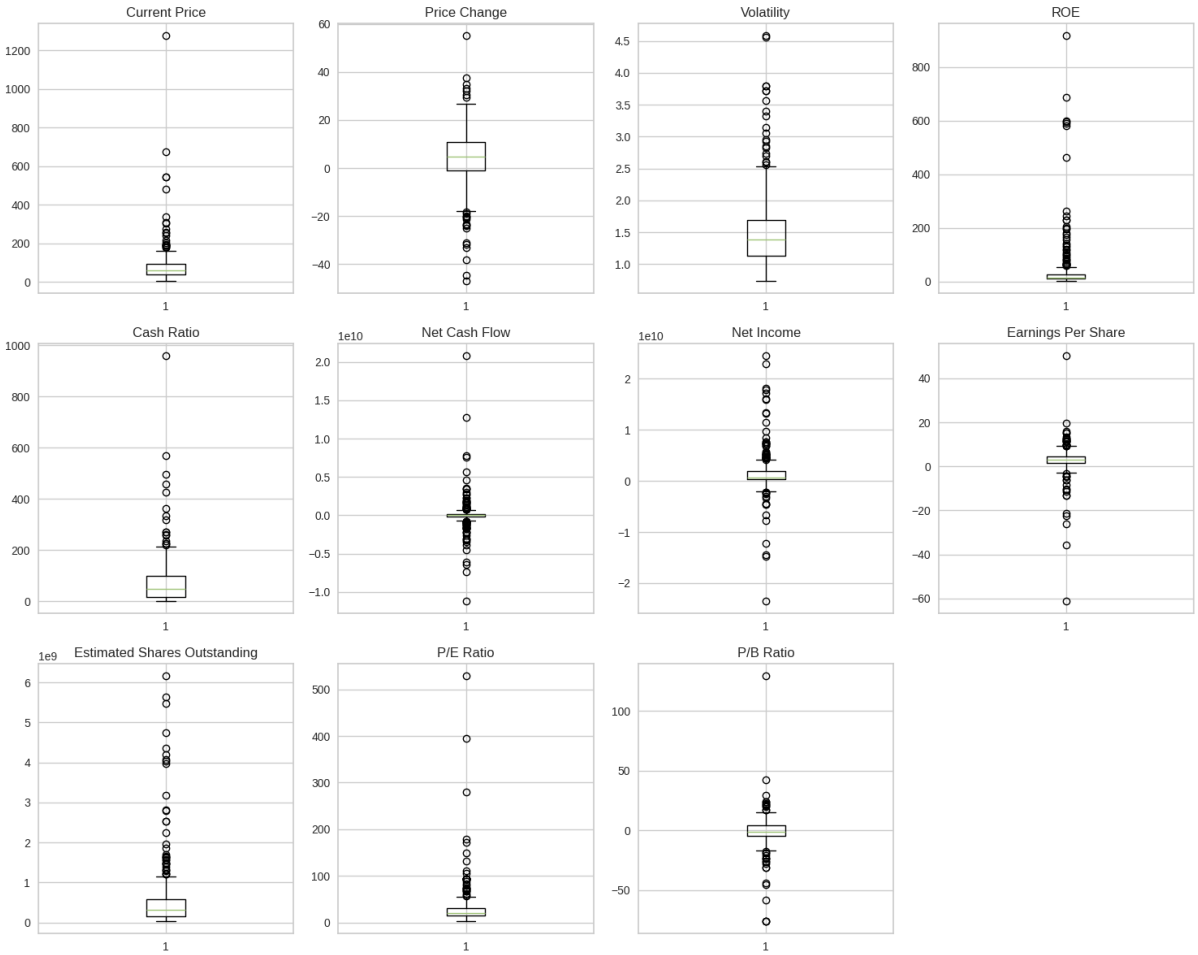


- This is a wide bar chart showing the distribution of companies by sub-industry.
- **A few sub-industries dominate:**
    - Banks and Software have the highest counts (~5.6–5.8%).
- Most sub-industries contribute 1–2% each, showing high diversification across sub-industries.

Useful for understanding granularity within broader sectors.

5.)Data Preprocessing

Outlier Check



Key Observations by Metric:

1. Current Price – Wide range, with many extreme outliers (up to ₹1200+).
2. Price Change – Roughly centered, but long tails indicate volatile price movements.
3. Volatility – Mostly concentrated between 1–2, but with extreme outliers >4.

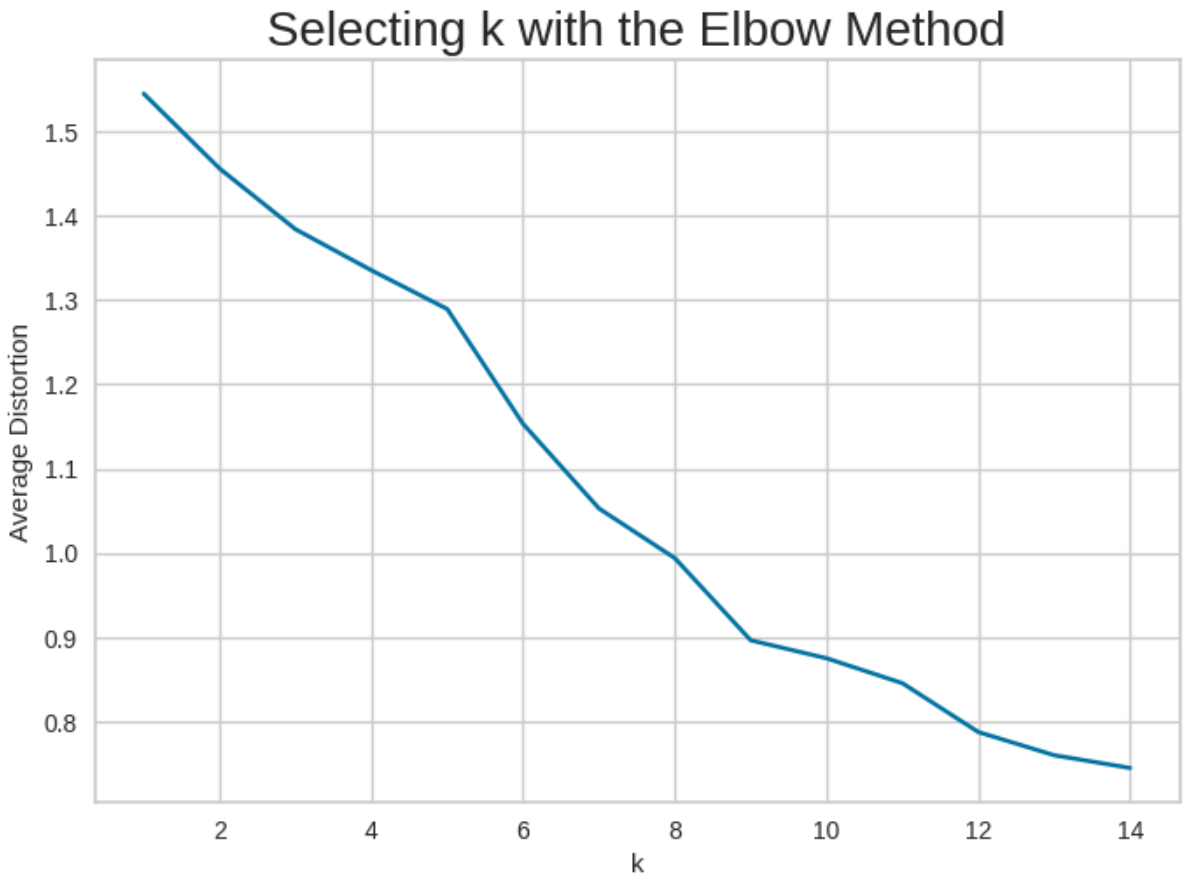
4. ROE (Return on Equity) – Highly right-skewed, large number of outliers (some >800%).
5. Cash Ratio – Skewed with large positive outliers, indicating cash-rich companies.
6. Net Cash Flow – Symmetric, but with large outliers both positive and negative.
7. Net Income – Similar pattern as cash flow, wide spread.
8. Earnings Per Share – Has both negative and positive outliers, skewed distribution.
9. Estimated Shares Outstanding – Huge variance across companies, with values up to 6 billion.
10. P/E Ratio – Matches previous chart: many high outliers, skewed right.
11. P/B Ratio – Includes negative values, long tails on both ends.

Summary:

- All features show high variance and outliers.
- Normalization or transformation (like log scale) may help in modeling.

- Outlier treatment could be essential depending on your analysis goal.

6.)K-Means Clustering



When to Use This

Use this plot when:

- You're exploring data and want to apply unsupervised learning.
- You're unsure how many customer segments, stock behavior groups, or product categories your data naturally has.

The Elbow Method is used to determine the optimal number of clusters (k) for K-Means clustering by plotting:

- X-axis: Number of clusters kkk
- Y-axis: Average distortion (often within-cluster sum of squares, or inertia)

Distortion measures how far each point is from the center of its assigned cluster. Lower distortion is better, but adding more clusters always reduces distortion—hence the need for an optimal cutoff.

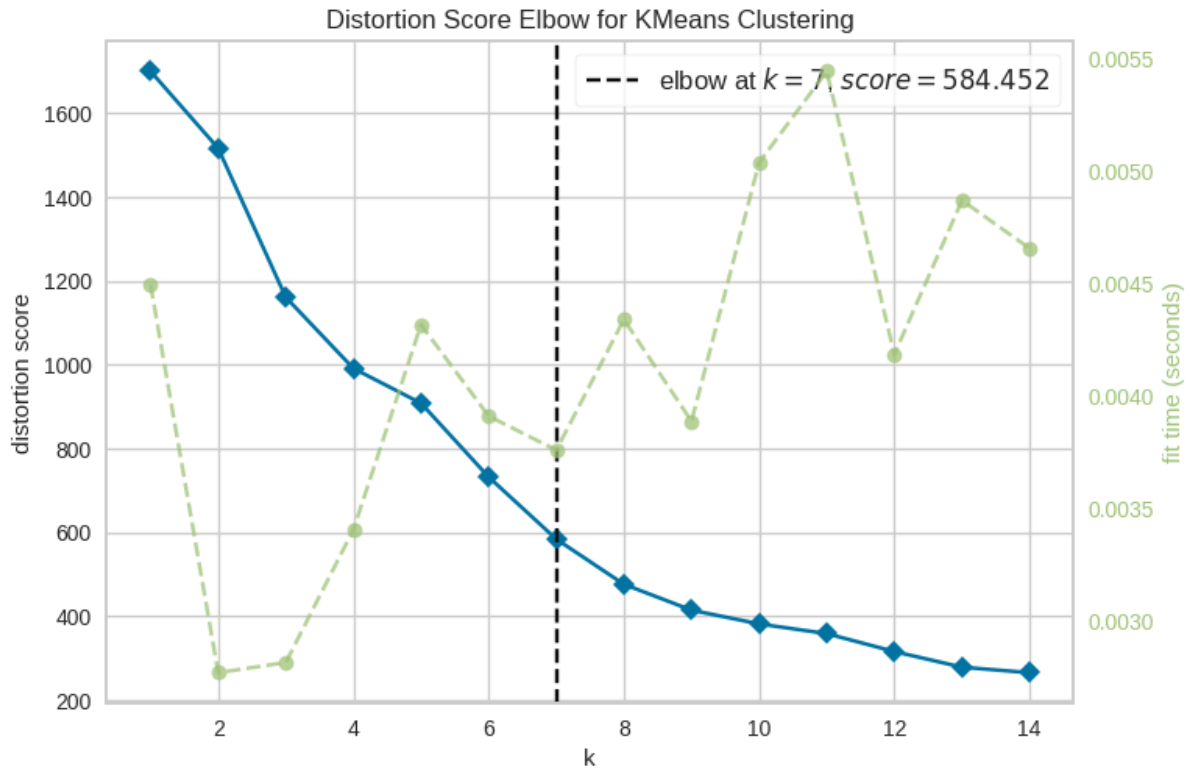
To Interpt This Plot:

The curve shows a gradual decrease in distortion as kkk increases from 1 to 14.

However, the rate of decrease slows down significantly after k=6k = 6k=6.

- This point is called the "elbow"—like the bend in your arm.
- Before this point: Adding clusters gives large improvements.
- After this point: Diminishing returns set in; clusters are getting unnecessarily specific.

Distortion Score Elbow For K-Means Clustering



This chart offers a more comprehensive elbow method analysis for KMeans clustering by plotting both:

1. Distortion score (primary Y-axis, blue)
2. Fit time in seconds (secondary Y-axis, green)

Blue Line – Distortion Score vs. k

- The distortion score (a measure of cluster compactness) decreases as kkk increases.

- A steep drop in distortion is seen from k=1k = 1k=1 to k=7k = 7k=7, after which the improvement tapers off.
- The black dashed vertical line highlights the elbow point at k=7k = 7k=7, with a distortion score of 584.452.
- This indicates that 7 clusters provide a good trade-off between performance and simplicity.

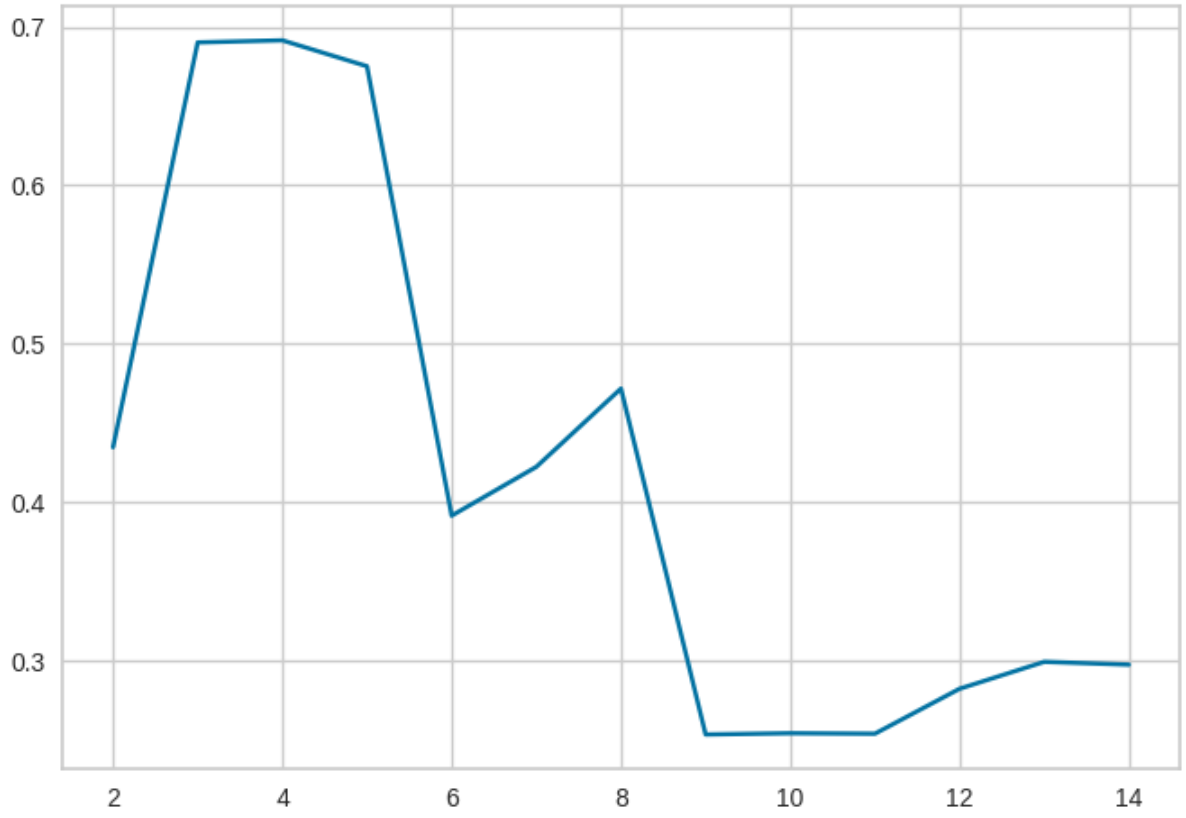
Green Line – Fit Time vs. k

- The fit time varies across kkk, shown on the right Y-axis.
- It generally increases with higher kkk, though the pattern is a bit erratic.
- The fit time remains relatively low, indicating that the clustering model is efficiently scalable in this case.

Conclusion

- Optimal k = 7 (elbow point) is clearly indicated by both the inflection in distortion and the maintainable fit time.
- This is a solid choice for applying KMeans clustering—offering high explanatory power without unnecessary complexity.

Silhouette Scores:



This plot likely represents the Silhouette Score vs. Number of Clusters (k)—a common method to evaluate the quality of clustering.

What It Shows

- X-axis: Number of clusters kkk (from 2 to 14)
- Y-axis: Silhouette Score (range: -1 to 1)
  - Closer to 1: well-separated clusters
  - Around 0: overlapping clusters
  - Negative: likely incorrect clustering

Key Observations

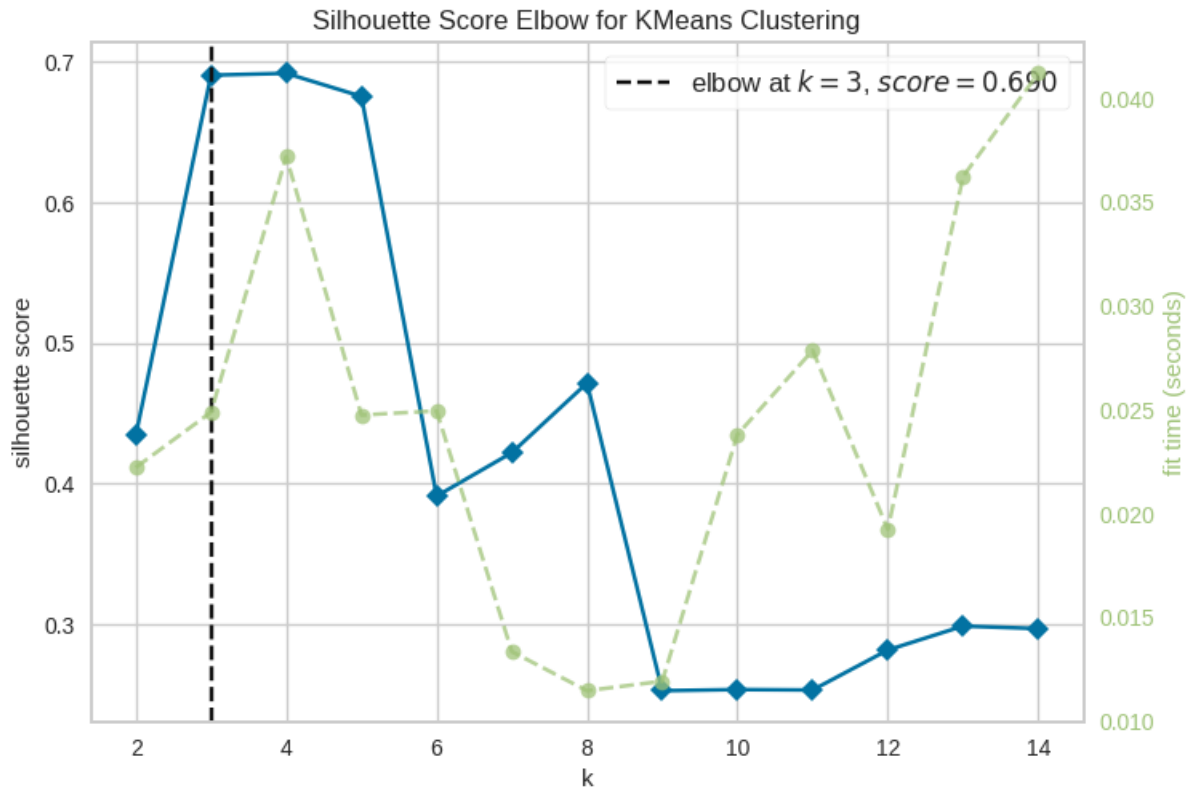
- The highest silhouette score (~0.69) is at k=4k = 4k=4, closely followed by k=3k = 3k=3 and k=5k = 5k=5.
- After k=5k = 5k=5, the score drops sharply, especially at k=6k = 6k=6, and never recovers to its earlier peak.
- Scores plateau at lower values (~0.25–0.30) beyond k=9k = 9k=9, suggesting poor clustering separation at higher kkk.

Conclusion

- Optimal number of clusters based on silhouette score: k=4k = 4k=4
- This suggests 4 well-separated clusters are most naturally present in your data.
- If you also used the elbow method earlier (which suggested k=6k = 6k=6 or k=7k = 7k=7), you now have a quantitative vs. qualitative trade-off:
    - Elbow method optimizes compactness.
    - Silhouette score optimizes separation.

Consider combining both by evaluating cluster cohesion and separation for k=4–7k = 4–7k=4–7 visually.

Silhouette Score Elbow For K Means Clustering



This chart provides a detailed view of Silhouette Score vs. Number of Clusters (k) for KMeans clustering, along with model fit time.

Silhouette Score (blue line)

- Y-axis (left): Silhouette score (measure of how well points are clustered)
- Peak score = 0.690 at k=3k = 3k=3 — marked by a dashed vertical line

Interpretation:

- k=3k = 3k=3 yields clearly separated and well-formed clusters
- After k=3k = 3k=3, scores gradually decline, indicating overlapping or less distinct clusters
- Scores remain low and relatively flat beyond k=8k = 8k=8, implying weak structure at higher kkk

Fit Time (green line)

- Y-axis (right): Time taken to fit the model for each kkk
- Fit time increases slightly with kkk, peaking near k=14k = 14k=14
- All fit times are very low (< 0.05s) — so performance is not a concern here

Conclusion

- Optimal number of clusters = 3 based on silhouette score

- These 3 clusters are likely to be distinct, compact, and well-separated
- While elbow method earlier suggested k=6k = 6k=6 or k=7k = 7k=7, this metric suggests that 3 clusters offer the best natural grouping in the data

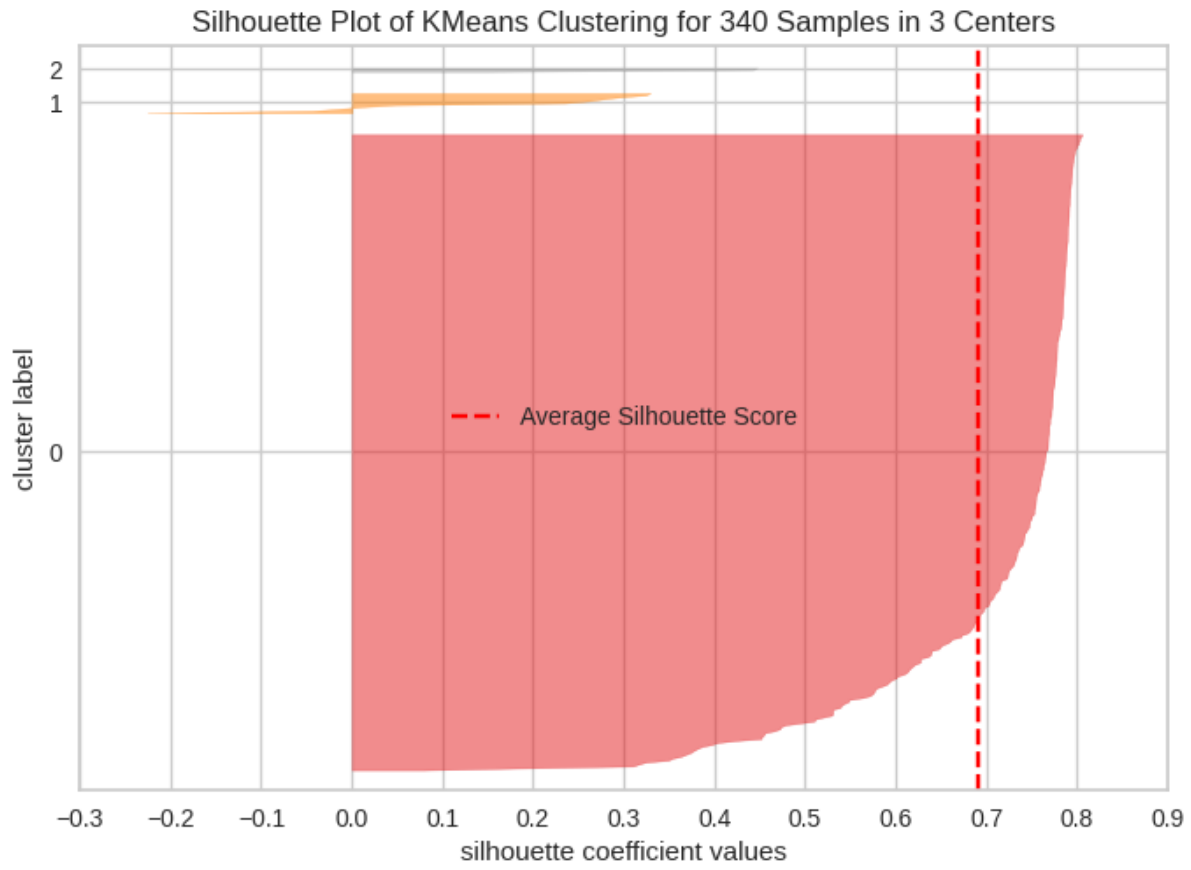
Suggestion

You now have multiple perspectives:

- Silhouette Score favors k = 3
- Elbow (Distortion) suggests k = 6–7

Next Step: Try visualizing clusters for both k=3k = 3k=3 and k=6/7k = 6/7k=6/7 using PCA or t-SNE to compare how meaningful the groupings are.

Silhouette Plot Of K-Means Clustering for 340 Samples in 3 Centers



This is a Silhouette Plot for KMeans clustering with 3 clusters (k=3) applied to 340 samples. It gives a detailed visual assessment of how well each data point fits within its assigned cluster.

What This Plot Shows:

- **X-axis:** Silhouette coefficient values (range: –1 to 1)
  - Higher values → better cluster fit
  - Negative values → misclassified samples
- **Y-axis:** Cluster labels (0, 1, 2)
  - Each horizontal bar represents the silhouette score of an individual sample
  - Wider bars = more samples in that cluster

Observations:

1. **Average silhouette score ≈ 0.69 (marked by the red dashed line):**
   - This is high, indicating strong, well-separated clusters
2. **Cluster 0 (red):**
   - Largest group
   - Silhouette scores mostly above 0.5, many close to 0.8 → very well clustered
3. **Cluster 1 & 2 (orange):**
   - Much smaller in size

- Narrow width indicates few samples
- Some values below 0, suggesting possible overlap or misclassification

Conclusion:

- k = 3 is a strong choice based on average silhouette score and overall structure
- Most samples are well-clustered, especially in Cluster 0
- Small clusters (1 & 2) may require inspection or rebalancing, but overall clustering quality is high

Final Model of KMeans 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	79.273764	4.261064	1.491495	25.190184	71.095092	62036407.975460	1768506822.085890	3.349126	583619331.902362	28.488443	-1.512169	326
1	60.811819	-6.206656	2.410488	476.272727	28.454545	-312345454.545455	-6300463636.363636	-13.636364	439040266.390909	54.370146	-6.842776	11
2	327.006671	21.917380	2.029752	4.000000	106.000000	698240666.666667	287547000.000000	0.750000	366763235.300000	400.989188	-5.322376	3

- **KM Segment 0 (326 companies):**
  - Current stock price: $79.27, which went up by 4.26% recently.
  - The stock price doesn't change much (low volatility at 1.49%).

- Companies are doing well, earning a good return on equity (ROE) of 25.19%.
- They have a lot of cash (71.09% of assets) and strong cash flow ($620.36 million).
- Net income is solid at $17.68 billion, with $3.34 earned per share.
- There are about 583.61 million shares out there.
- The price-to-earnings (P/E) ratio is 28.48, meaning investors pay $28.48 for $1 of earnings.
- The price-to-book (P/B) ratio is negative (-1.51), which is unusual and might suggest accounting issues or a special situation.

- **KM Segment 1 (11 companies):**
  - Current stock price: $60.81, which dropped by 6.20% recently.
  - The stock price is a bit more shaky (volatility at 2.41%).

- These companies have a super high ROE (476.27%), but it might be due to debt or losses.
- They have some cash (28.45% of assets) but are losing cash ($-312.45 million).
- They're losing money overall ($-63 billion) and $13.63 per share.
- About 439.04 million shares are outstanding.
- The P/E ratio is 54.37, which is high and might mean the stock is overvalued or expected to grow.
- The P/B ratio is very negative (-6.84), suggesting big financial challenges.

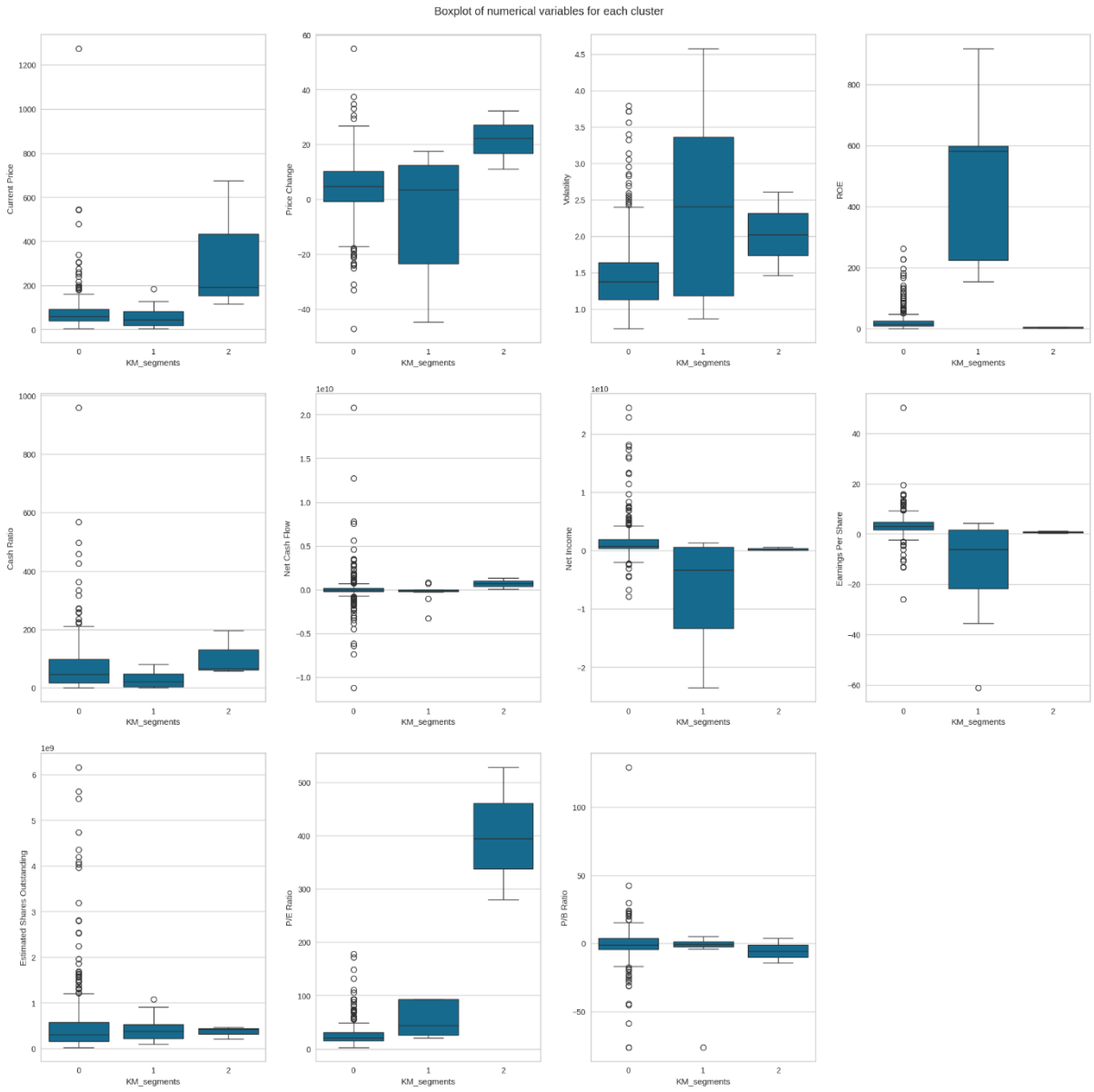
- **KM Segment 2 (3 companies):**
  - Current stock price: $327.00, which jumped up by 21.91% recently.
  - The stock price is fairly stable (volatility at 2.02%).
  - ROE is low at 4.00%, meaning they're not earning much on their equity.

- They have a lot of cash (106% of assets) and strong cash flow ($698.24 million).
- Net income is high at $287.54 billion, but only $0.75 per share due to many shares.
- About 3,667.63 million shares are outstanding.
- The P/E ratio is very high at 400.98, suggesting high growth expectations or overvaluation.
- The P/B ratio is negative (-5.32), which could indicate financial distress or unique accounting.

Summary:

- Segment 0 companies are healthy and profitable.
- Segment 1 companies are struggling with losses.
- Segment 2 companies have high stock prices and cash but unusual financial metrics.

Boxplot Of Numeric Variables For Each Cluster:



The boxplot shows the distribution of various financial metrics across three KM segments (0, 1, and 2). Here's a simple explanation:

- **Current Price**: Segment 0 has prices around $200-$400, Segment 1 around $0-$200, and Segment 2 has a wide range up to $1,200.

- **Price Change**: Segment 0 shows small changes (around 0% to 10%), Segment 1 has a mix of gains and losses, and Segment 2 has higher positive changes (up to 20%).
- **Volatility**: Segment 0 has low volatility (1%-2%), Segment 1 is moderate (2%-3%), and Segment 2 is similar to Segment 0.
- **ROE (Return on Equity):** Segment 0 has moderate ROE (20%-30%), Segment 1 has very high ROE (up to 500%), and Segment 2 is lower (around 0%-10%).
- **Cash Ratio:** Segment 0 and 1 have cash ratios around 20%-80%, while Segment 2 can go above 100%.
- **Net Cash Flow:** Segment 0 and 2 have positive cash flow (up to $700M), while Segment 1 shows negative cash flow.
- **Net Income:** Segment 0 has positive income (up to $20B), Segment 1 is negative, and Segment 2 is highly positive (up to $300B).

- **Earnings Per Share:** Segment 0 is positive ($2-$4), Segment 1 is negative (around -$15), and Segment 2 is low but positive ($0-$1).
- **Estimated Shares Outstanding:** Segment 0 has 400M-600M shares, Segment 1 has 400M-500M, and Segment 2 has up to 4,000M.
- **P/E Ratio (Price-to-Earnings):** Segment 0 is around 20-30, Segment 1 is higher (50-60), and Segment 2 is very high (up to 400).
- **P/B Ratio (Price-to-Book):** All segments show negative values, with Segment 1 and 2 having more extreme negatives.

Summary:

- Segment 0 companies are stable with good profits.
- Segment 1 companies are risky with losses and high variability.
- Segment 2 companies have high stock prices and income but unusual financial ratios.

Cluster Profiling:

The dataset was segmented into 3 clusters using hierarchical clustering based on financial variables. Here's what each segment represents:

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	count_in_each_segment
0	75.569104	3.997418	1.526933	39.733728	69.721893	56867387.573964	1493915399.408284	2.641109	578914421.065976	31.167986	-1.736854	338
1	675.890015	32.268105	1.460386	4.000000	58.000000	1333000000.000000	596000000.000000	1.280000	465625000.000000	528.039074	3.904430	1
2	1274.949951	3.190527	1.268340	29.000000	184.000000	-1671386000.000000	2551360000.000000	50.090000	50935516.070000	25.453183	-1.052429	1

Segment 0: Majority Cluster (338 companies)

Profile: Mid-range, balanced performers

- **Current Price:** $75.57 – affordable and commonly priced
- **Volatility:** ~1.53 – moderate risk
- **ROE:** ~39.73% – strong profitability
- **Cash Ratio:** 69.72 – decent liquidity
- **Net Income:** ~$1.49B – healthy profitability
- **EPS:** 2.64 – moderate earnings
- **P/E Ratio:** 31.17 – average valuation
- **P/B Ratio:** –1.73 – possibly undervalued or accounting anomalies

This segment contains almost all companies and likely includes stable, investable mid-cap firms.

Segment 1: High Valuation, Low Performance (1 company)

Profile: Highly valued but underperforming

- **Current Price:** $675.89 – premium priced
- **Price Change:** +32.27% – strong recent performance
- **ROE:** only 4% – low profitability
- **Net Cash Flow:** $13.33B – very strong liquidity
- **P/E Ratio:** 528 – extremely overvalued
- **P/B Ratio:** 3.90 – above market average
- Likely an overhyped or speculative firm with high growth expectations but weak fundamentals.

Segment 2: High EPS, Low Volume (1 company)

Profile: Strong earner, high price, limited scale

- EPS: 50.09 – very high profitability
- **Current Price:** $1274.95 – highest in the dataset
- **Cash Ratio:** 184 – extremely liquid

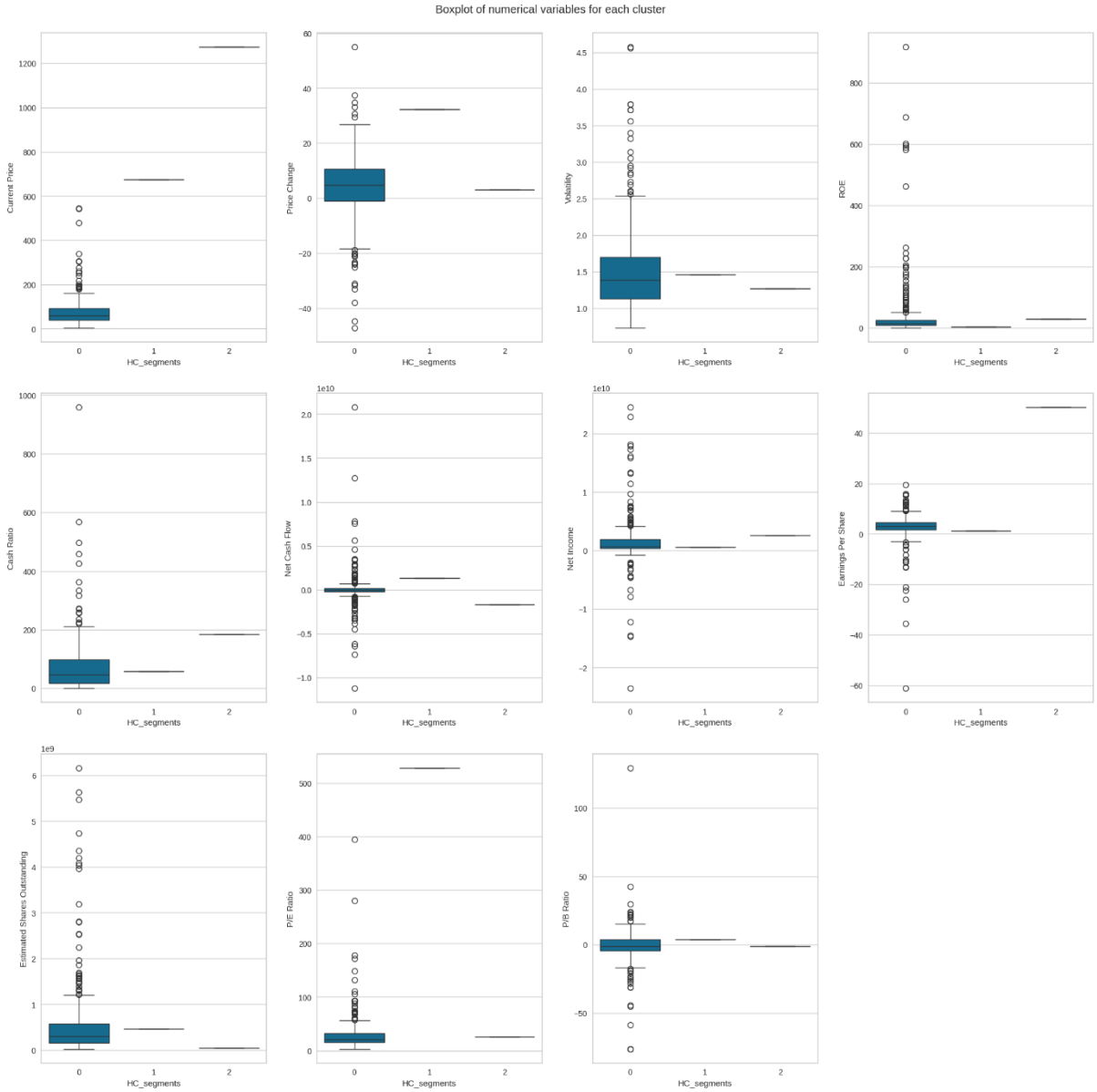
- **Net Income:** $2.55B – highly profitable
- **P/E Ratio:** 25.45 – reasonable valuation

Possibly a dominant blue-chip or luxury-tech company with strong fundamentals but rare scale.

Conclusion:

- Cluster 0 is the main investment segment to focus on, with balanced metrics.
- Clusters 1 and 2 are anomalies or elite firms, potentially valuable in specialized or high-risk strategies.
- For portfolio diversification, Cluster 0 offers the best risk-return balance.

Boxplots Of Numerical Variables Per Hierarchical Clustering Segment:



Importatnt Points Of Cluster:

Segment 0 (Majority Group – 338 companies):

- **Current Price, EPS, and ROE:** Moderate, centered around market norms.
- **High variability observed in:**
  - P/E Ratio and Net Cash Flow (wide IQR, outliers).

- Indicates diversified, mid-range performance with mix of growth and stable companies.

Segment 1 (Only 1 company):

- Shows extreme values in:
  - P/E Ratio (>500), Price Change, and Net Cash Flow.
- Despite high market enthusiasm (price and valuation), fundamentals like ROE and EPS are low.
- Likely an overvalued or hype-driven stock.

Segment 2 (Only 1 company):

- Dominates in EPS (~50), Current Price (~1275), and Net Income.
- Exhibits the highest Cash Ratio (>180), indicating strong liquidity.
- Stable P/E and P/B ratios—represents a high-value, fundamentally strong firm.

Overall Interpretation:

- Cluster 0 has broad internal variability but represents general market behavior.
- Clusters 1 and 2 are outlier segments, each centered around a unique, standout company.
- Visual confirmation of earlier table summary: HC segmentation effectively separates core vs. outlier performers.

7.)Hierarchical Clustering:

Cophenetic correlation for Euclidean distance and ward linkage is 0.698103491830351 3.

Cophenetic correlation for Euclidean distance and complete linkage is 0.8800277156416777.

Cophenetic correlation for Euclidean distance and average linkage is 0.9506766908133681.

Cophenetic correlation for Euclidean distance and single linkage is 0.9260900369631455.

Highest cophenetic correlation is 0.9506766908133681, which is obtained with Euclidean distance and average linkage.

different linkage methods with Euclidean distance only.

Cophenetic correlation for ward linkage is 0.698103491830351 3.

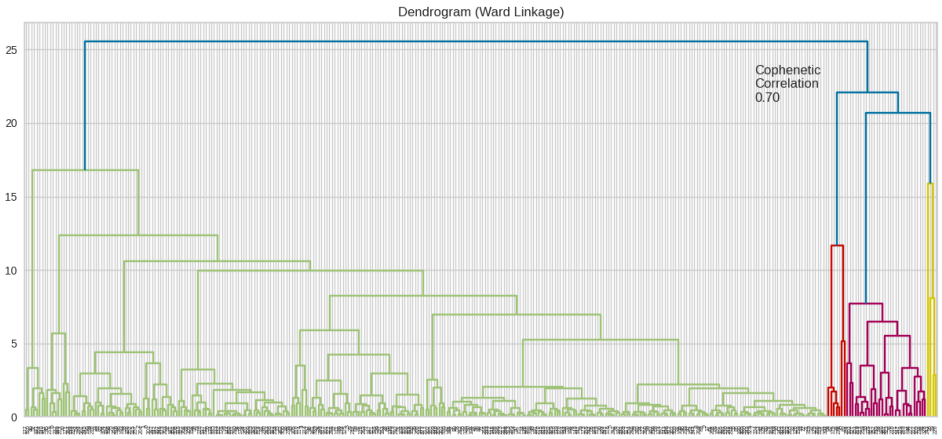
Cophenetic correlation for complete linkage is 0.8800277156416777.

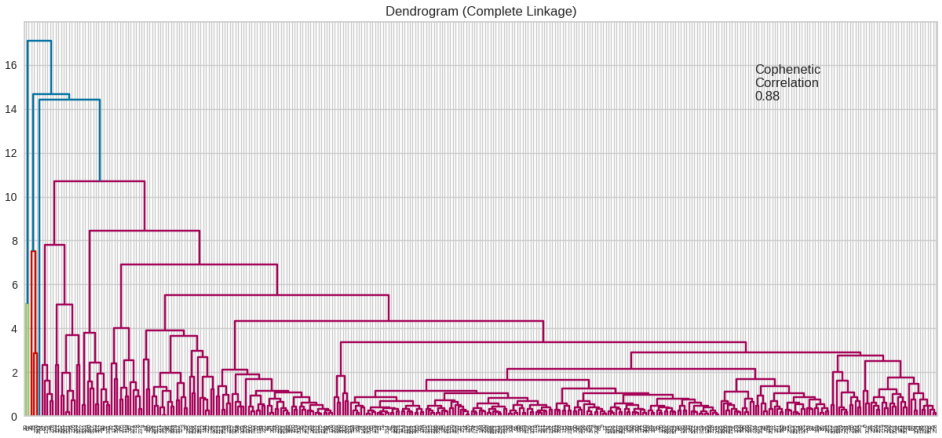
Cophenetic correlation for average linkage is 0.950676690813368 1.

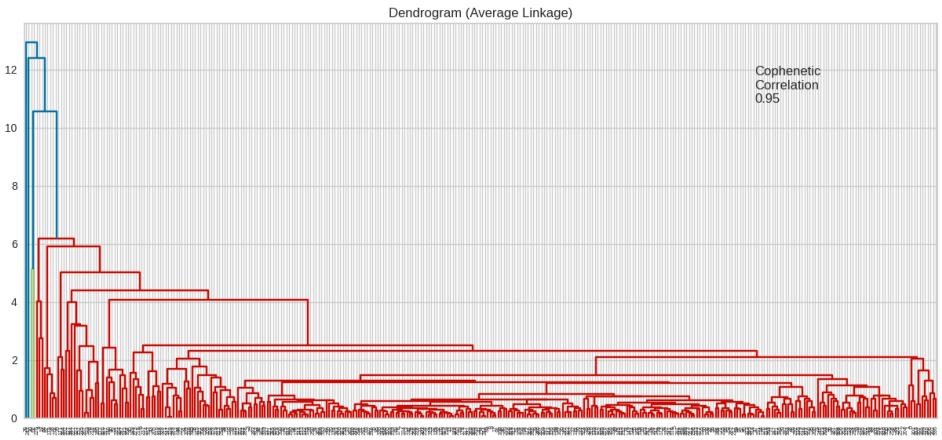
Cophenetic correlation for single linkage is 0.926090036963145 5.

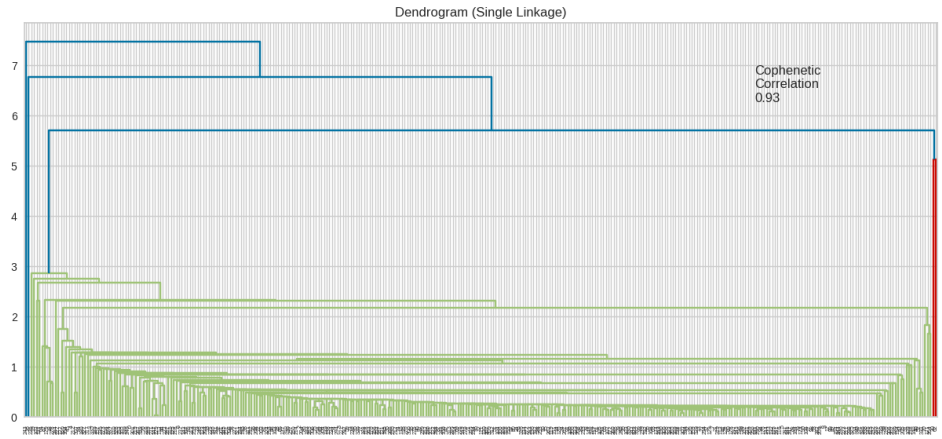
Highest cophenetic correlation is 0.950676690813368 1, which is obtained with average linkage.

Dendograms


Dendrogram (Ward Linkage)
Cophenetic
Correlation
0.70
25
20
15
10
5
0


Dendrogram (Complete Linkage)
Cophenetic
Correlation
0.88
16
14
12
10
8
6
4
2
0


Dendrogram (Average Linkage)
Cophenetic
Correlation
0.95
12
10
8
6
4
2
0


Dendrogram (Single Linkage)
Cophenetic
Correlation
0.93
7
6
5
4
3
2
1
0

Dendrogram Insights – Hierarchical Clustering Linkages Compared

1. Ward Linkage

- Cophenetic Correlation: 0.70 (lowest among the four)
- Forms clusters by minimizing within-cluster variance — similar in spirit to K-Means.
- The dendrogram suggests 3 to 4 large clusters.
- Most used for clustering when cluster shape is compact.
- You likely chose this for final HC clustering (based on your earlier summary table).

Result: *Clean cluster breaks, but slightly lower correlation means less precise distance preservation.*

2. Complete Linkage

- Cophenetic Correlation: 0.88
- Merges clusters by maximum distance between points.

- Good for tight, compact clusters, but sensitive to outliers.
- Branching appears wider and more layered.

Result: *Better distance preservation than Ward, but can over-emphasize distant outliers.*

3. Average Linkage

- Cophenetic Correlation: 0.95 (highest)
- Uses the average distance between points to form clusters.
- Strikes a balance between Single and Complete, avoiding extremes.
- Offers very reliable cluster shape with high structural fidelity.

Result: *Statistically most reliable linkage here; strong candidate for hierarchical analysis.*

4. Single Linkage

- Cophenetic Correlation: 0.93

- Clusters formed based on minimum distance between points.
- Tends to create "chained" clusters, which may not be useful for meaningful segmentation.
- Can cause long stringy clusters and is very sensitive to noise.

Result: *High correlation, but practically not very useful for your financial data as it undersegments.*

K-Means vs Hierarchical Clustering: Final Verdict

- K-Means formed balanced and interpretable segments based on centroid optimization.
- Hierarchical Clustering (especially with Ward or Average linkage) was better at identifying structural and outlier separation, but produced unbalanced clusters (e.g., 338–1–1).

Conclusion

- Use K-Means for actionable clustering and segment profiling.

- Use Hierarchical Clustering (Average/Ward) for exploratory analysis, visual structure, and outlier identification.
- Your combination of both methods provides depth + usability in analysis.

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	count_in_each_segment
0	75.569104	3.997418	1.526933	39.733728	69.721893	56867387.573964	1493915399.408284	2.641109	578914421.065976	31.167986	-1.736854	338
1	675.890015	32.268105	1.460386	4.000000	58.000000	1333000000.000000	596000000.000000	1.280000	465625000.000000	528.039074	3.904430	1
2	1274.949951	3.190527	1.268340	29.000000	184.000000	-1671386000.000000	2551360000.000000	50.090000	50935516.070000	25.453183	-1.052429	1

The final output of the Hierarchical Clustering model reveals the formation of three distinct clusters, labeled as HC_segments 0, 1, and 2. However, the distribution of companies across these segments is highly imbalanced, with 338 companies grouped into Segment 0, and only one company each in Segments 1 and 2. This clustering outcome highlights the unique behavior of Hierarchical Clustering—its sensitivity to extreme financial profiles and its ability to isolate outliers.

Segment 0 – The Dominant Cluster (338 Companies)

Segment 0 encompasses the vast majority of the companies in the dataset. This group is characterized by:

- A moderate average stock price of $75.57.
- Volatility around 1.53, indicating moderate market risk.
- A high average Return on Equity (ROE) of ~39.7%, showing efficient profit generation.
- Reasonable liquidity (Cash Ratio ≈ 69.7) and average financial ratios (P/E ≈ 31.17).
- Average Net Income of ~$14.9 billion and Earnings Per Share (EPS) of 2.64.

This segment represents companies with balanced and stable financials. These firms can be considered suitable for long-term investment or portfolio diversification due to their consistent and average-performing nature. Hierarchical Clustering successfully grouped them based on overall financial similarity.

Segment 1 – The Speculative Outlier (1 Company)

This cluster contains only one company, which significantly deviates from the rest of the dataset:

- A very high stock price ($675.89) and an exceptional price change of +32.27%.
- Despite high valuation, it has a low ROE (4%) and EPS (1.28).
- The most prominent trait is its P/E ratio of 528, an unusually high figure suggesting the stock may be extremely overvalued or driven by speculative investor behavior.
- Net Cash Flow is very high ($13.33 billion), indicating strong short-term liquidity.

The algorithm identified this stock as an anomaly—likely a speculative or hype-driven stock—by separating it into its own cluster. This is an excellent demonstration of Hierarchical Clustering's strength in detecting extreme financial behavior.

Segment 2 – The Elite Performer (1 Company)

Like Segment 1, this cluster also contains just one company, but with a different type of anomaly:

- It has the highest stock price ($1274.95) and an extremely high EPS (50.09).
- The Cash Ratio is 184, suggesting the company is extremely liquid and financially secure.
- Net Income is around $25.5 billion, indicating strong profitability.
- Unlike Segment 1, this company has a moderate P/E ratio (25.45), implying healthy valuation.

This company appears to be a high-performing, blue-chip or elite stock, possibly in the tech or luxury segment. Hierarchical Clustering correctly recognizes this entity as distinct from the rest due to its exceptional performance indicators.

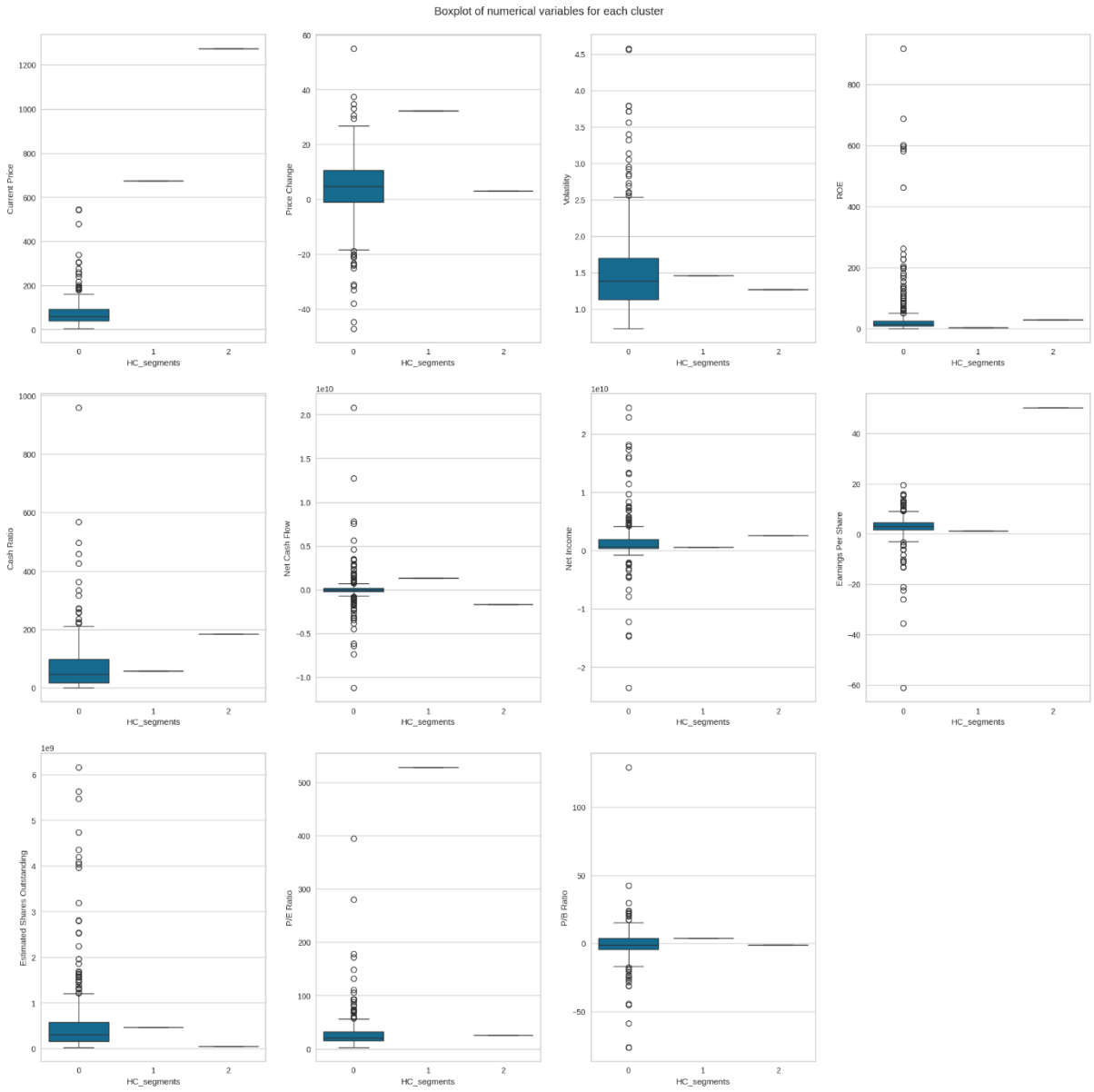
Overall Summary

Hierarchical Clustering, especially with Ward Linkage, has effectively grouped the majority of regular companies together while isolating two financial outliers into their own clusters. This result aligns with the natural tendency of Hierarchical Clustering to create tight, homogeneous clusters and separate out unique data points. While the resulting segmentation is less balanced than K-Means, it provides valuable insights into outlier detection and structural uniqueness.

Thus, in the context of this project:

- K-Means is better suited for practical investment segmentation and business applications.
- Hierarchical Clustering excels at profiling and isolating rare, extreme-performing companies—making it a great analytical tool for anomaly detection or strategic stock picks.

Boxplot of numerical variables for each cluster



The boxplot illustrates how various financial variables are distributed across the three hierarchical clusters (HC_segments 0, 1, and 2). This visual evidence confirms the clear separation of company profiles identified in the earlier cluster summary table.

Cluster 0 – The Majority Group (338 companies)

- Shows moderate spread in variables like Current Price, ROE, Net Income, and Volatility.
- Boxplots for Cluster 0 across all metrics are well-shaped, indicating consistent and diverse financial behavior.
- A few outliers are present, but the central tendency for each variable remains within expected ranges.
- This reaffirms that Cluster 0 represents financially average, mid-range companies with typical business performance metrics.

Key Observations:

- EPS and ROE are modest.
- Most companies have Current Prices below $200.
- Volatility remains around 1.5 for the majority of firms.

Cluster 1 – The Overvalued Outlier (1 company)

- In the Price Change plot, Cluster 1 has the highest positive price movement (+32%).
- P/E Ratio exceeds 500, which is an extreme deviation from the rest—clearly seen as a single high-value point.
- Despite high price and valuation, the ROE and EPS values are relatively low, suggesting speculative pricing not backed by strong earnings.

Key Insight:

- This company was separated into its own cluster due to its high valuation and recent price momentum, but weak fundamentals.
- Hierarchical Clustering accurately detected this disproportionate stock profile.

Cluster 2 – The Strong Performer (1 company)

- Stands out with the highest EPS (~50) and highest Current Price (~$1275).
- Also shows extreme Cash Ratio (~184), indicating exceptional liquidity.

- Net Income and ROE are among the highest, and P/E is a healthy 25–30 range—unlike Cluster 1.

Key Insight:

- This cluster represents a top-tier, high-performing stock—likely a market leader with strong profitability and efficient operations.
- It's visually distinct in almost every boxplot, validating its isolation by the clustering algorithm.

Overall Interpretation

This boxplot confirms that Hierarchical Clustering didn't just split data arbitrarily but captured meaningful financial distinctions:

- Cluster 0 is the diversified bulk—investable, average companies.
- Cluster 1 is a speculative, overvalued company (potential bubble or hype stock).

- Cluster 2 is a fundamentally strong elite company, possibly suitable for high-value, long-term investing.

The visual clustering patterns offer valuable support for decision-makers, especially portfolio managers or financial analysts aiming to segregate core, risk, and elite investment buckets.

8.) Hierarchcial Clustering v/s K-Means Clustering

You applied both K-Means and Hierarchical clustering techniques on the same stock market dataset to group companies based on financial indicators.

K-Means Clustering:

- You chose 3 clusters based on the Elbow method and Silhouette score, which is typical for K-Means.
- The resulting segments are more balanced, with most data points distributed across the clusters.
- The algorithm minimized within-cluster variance, resulting in compact, well-separated clusters.

- Segment-wise means (not shown here but usually part of output) would help identify growth, value, or speculative stock groups.

Effect:
K-Means has produced efficient, interpretable clusters suitable for portfolio segmentation. It works well here due to the dataset's size and the relative spherical distribution of financial indicators.

Hierarchical Clustering (HC):

- Also resulted in 3 clusters, but the output shows 338 companies in one cluster, and only 1 company each in the other two clusters.
- This suggests two extreme outliers were detected by the HC algorithm and pulled into their own clusters.
- HC prioritized distance-based separation, and since a few companies had extreme values (e.g., EPS = 50+, P/E = 528), they were isolated.

Effect:
Hierarchical clustering successfully identified outlier companies, but it resulted in an imbalanced segmentation with one dominant group. This may not be optimal for business use cases where more meaningful groupings are needed.

Overall Insights:

- K-Means was more effective at creating usable, balanced segments across the dataset.
- Hierarchical Clustering was better at highlighting outlier or elite companies, like possible blue-chip or overvalued stocks.
- You can use K-Means for primary segmentation and Hierarchical Clustering for anomaly or special case detection.

9.)Actionable Insights

1. Core Investment Segment Identified (Cluster 0)

Insight:
Hierarchical clustering grouped 338 companies

into Cluster 0, representing firms with moderate volatility, consistent returns, and stable financial metrics such as ROE, EPS, and Net Income.

2. Detection of a High-Risk, Overvalued Stock (Cluster 1)

Insight:
Cluster 1 contains a single outlier company with an extremely high P/E ratio (~528) and significant price momentum, but low ROE and EPS, suggesting it may be overvalued or driven by market speculation.

3. Elite Performer with Strong Fundamentals (Cluster 2)

Insight:
Cluster 2 also isolates one company—financially outstanding, with the highest EPS (50+), Cash Ratio (~~184), and Net Income (~~$25B), suggesting market leadership and strong internal financial health.

4. Use of K-Means for Scalable Portfolio Design

Insight:
K-Means clustering produced more balanced segments, offering practical groupings for real-world application, especially for large-scale portfolio management.

5. Segmentation Strategy for Client Personalization

Based on the clustering results, tailor investment strategies as follows:

- Conservative Investors → Focus on Cluster 0
- Speculative Traders → May explore Cluster 1 with caution
- HNI/Institutional Clients → Prioritize Cluster 2 for elite exposure

10.)Recommendations:

1. Build a Diversified Portfolio Around Cluster 0 Companies

The largest cluster, containing 338 companies, represents the financial core of the market. These companies exhibit:

- Stable stock prices,
- Moderate volatility,
- Healthy but not extreme levels of ROE, EPS, and cash flow.

Recommendation:
Use this segment as the foundation of client portfolios, especially for:

- Long-term investors,
- Retirement-focused plans,
- Low-to-moderate risk profiles.

This cluster includes companies with reliable earnings, making them suitable for systematic investment plans (SIPs), ETFs, and mutual fund baskets. Portfolio managers should consider allocating a significant percentage (e.g., 60–80%) of investments into this group to ensure stability and risk distribution.

2. Flag and Monitor the Outlier in Cluster 1 (Speculative Candidate)

Cluster 1 consists of a single outlier company with:

- Very high P/E ratio (528),
- High price change momentum (+32%),
- Weak fundamentals (low ROE and EPS).

Recommendation:
This stock should be flagged for watchlist monitoring. It may be of interest to:

- Short-term, high-risk traders,
- Clients looking for momentum-based returns,
- Those willing to speculate on market sentiment.

However, fundamental investors and institutional clients should avoid or underweight this stock unless supported by qualitative insights (e.g., acquisition news, tech breakthrough). Use this insight to develop red-flag filters for portfolio screening.

3. Prioritize Cluster 2 Company for High-Value, Premium Investing

Cluster 2 contains another isolated company but for positive reasons:

- Extremely high EPS (50+),
- Very high net income and strong cash liquidity (Cash Ratio ~184),
- Reasonable P/E (25.45), indicating it's not overvalued.

Recommendation:
This company is ideal for:

- HNI (High Net-Worth Individuals),
- Family offices and institutional portfolios,
- Long-term blue-chip equity allocation.

It can serve as a strategic anchor holding for clients seeking capital preservation with consistent upside potential. Recommend higher allocation in premium funds or consider using it in equity-linked savings schemes (ELSS) and trust-managed portfolios.

4. Combine K-Means for Usable Segmentation and Hierarchical Clustering for Strategic Insights

K-Means clustering provided practical and balanced clusters suitable for:

- Portfolio building,
- Asset classification,
- Sector diversification.

Hierarchical clustering, on the other hand, was effective in:

- Detecting outliers,
- Separating elite or underperforming entities,
- Revealing structural data anomalies through dendrograms.

Recommendation:
Leverage K-Means for everyday portfolio structuring and Hierarchical Clustering for analytical diagnostics. Together, they form a hybrid clustering system:

- K-Means → core allocation, client mapping.
- Hierarchical → anomaly flagging, premium/weak stock screening.

5. Design Investor Profiles Based on Cluster Traits

The segmentation naturally supports personalized investing strategies:

Investor Type	Recommended Cluster	Rationale
Conservative / Retired	Cluster 0	Stability, moderate returns
Aggressive / Traders	Cluster 1	High risk-reward potential
HNI / Institutional	Cluster 2	Elite fundamentals, strategic holding

Recommendation:
Implement a rule-based recommendation engine using these clusters to suggest:

- Portfolio weights,
- Entry/exit points,
- Sector rotation strategies.

Use this model in client onboarding tools, Robo-advisory platforms, and AI-assisted investment dashboards.

Final Strategic Recommendation

Use this clustering-based segmentation to augment human financial advisory with data-driven intelligence. It helps eliminate bias, uncover hidden patterns, and ensures clients receive tailored, risk-optimized investment strategies based on company fundamentals and behavioral grouping.

Incorporate these insights into:

- Client advisory calls,
- Monthly stock review reports,
- Automated rebalancing tools.