NYSE Stock Portfolio Optimization

Project 6 – Unsupervised Learning

July 10, 2025

Submitted By:
Alex Kyeremateng Botwe

Contents / Agenda

List of Tables

List of Figures

List of Figures

Executive Summary

- Investing in the stock market offers a powerful path to long term wealth creation and inflation protection, fueled by the benefits of compound interest and potential tax advantages. To maximize returns and mitigate risk, a diversified portfolio is crucial, and cluster analysis can simplify the process of identifying suitable, low correlation stocks across various market segments.
- This project successfully leveraged advanced machine learning clustering techniques, specifically K-Means and Hierarchical Clustering to segment a comprehensive dataset of stock prices based on key financial performance indicators for a few companies listed under the New York Stock Exchange (NYSE). The analysis consistently identified three distinct and actionable stock segments, providing Trade&Ahead with a robust framework to move beyond traditional market classifications. This data-driven approach is poised to significantly enhance the personalization of investment strategies, optimize portfolio construction, and sharpen risk management capabilities, ultimately driving greater client satisfaction and competitive advantage.

- **Business Insights:**

The clustering analysis revealed a clear and consistent market microstructure, categorizing companies into three primary personas that reflect distinct investment characteristics:

a. Cluster 0: Balanced & Diversified Core (Largest Segment):

- This segment represents the vast majority of the market, comprising companies with generally stable financial performance and moderate growth or value metrics. These are the workhorse stocks that form the foundation of diversified portfolios, offering broad market exposure across various sectors.

b. Cluster 1: Growth & High-Performance Leaders (Targeted Segment):

- This group consists of companies exhibiting strong growth, high profitability, and robust financial health. The analysis highlighted both a broader K-Means group and, notably, a singular, extreme outlier (Priceline.com Inc) identified by Hierarchical Clustering, underscoring the method's ability to pinpoint exceptionally high performing entities. These stocks are prime candidates for clients seeking capital appreciation.

c. Cluster 2: Deep Value / Underperforming Assets (Small, High-Risk Segment):

- This critical segment, consistently identified by both algorithms (e.g., Apache Corporation, Chesapeake Energy), comprises companies facing significant financial challenges, characterized by negative profitability, high volatility, and often, distressed balance sheets. These represent high risk opportunities for contrarian investors seeking potential turnaround plays.

- **Business Recommendations:**

To leverage these powerful insights, Trade&Ahead should implement the following recommendations:

a. Recommendation for Clients Seeking Core Portfolio Stability and Growth:

- **Integrate Clusters into Client Advisory:** Embed these three stock segments directly into client portfolio construction and advisory services.

  - **Core Portfolios**: Predominantly allocate to the "Balanced & Diversified Core" for broad market exposure and stability.

  - **Growth-Oriented Clients**: Strategically allocate to the "Growth & High-Performance Leaders" for capital appreciation, potentially isolating extreme outliers for higher conviction bets

  - **High-Risk Clients**: Carefully evaluate opportunities within the "Deep Value / Underperforming Assets" segment, emphasizing due diligence and appropriate risk management.

- **Business Recommendations:**

b. Enhance Targeted Research and Due Diligence:

- Direct research teams to conduct more focused analysis within each cluster, identifying specific sub-trends, sector plays, or individual company catalysts. This optimizes research efficiency.

c. Improve Client Communication and Education:

- Utilize these clear segment personas and the supporting visuals (dendrograms, cluster profiles) to explain portfolio diversification and investment rationale to clients. This transparency builds trust and clarifies the reasons behind recommendations.

d. Implement Continuous Monitoring and Re-Clustering:

- Market dynamics evolve. Establish a schedule either quarterly or semi-annually to re-run the clustering algorithms and update segment profiles to ensure relevance and adapt to changing market conditions.

- **Conclusions:**

  - This clustering analysis demonstrates the power of data-driven segmentation in finance.

  - Both methodologies consistently identified 3 distinct, well-defined stock segments, moving beyond traditional classifications

  - Three critical groups namely "Balanced & Diversified Core“, "Growth & High-Performance Leaders“ and "Deep Value / Underperforming Assets“ were identified.

  - By discerning underlying patterns within stock data, it empowers investors to move beyond superficial observations, enabling more informed, strategic, and risk-aware investment decisions.

  - The findings highlight that even with a simple cluster count, different algorithms can unveil different, yet equally valuable, market insights

  - Trade&Ahead can now lead with intelligence, offering a differentiated and superior investment experience for its clients

Business Problem Overview and Solution Approach

- **Business Problem Overview:**

In today's dynamic and competitive financial landscape, providing generic investment advice is no longer sufficient. Clients, particularly in the high-net-worth segment, increasingly demand highly personalized investment strategies that align precisely with their unique goals, risk tolerance, and financial profiles.

Trade&Ahead faces the challenge of:

Limited Granularity: Relying solely on broad market classifications (e.g., GICS sectors) may obscure deeper, more nuanced relationships and behavioral patterns among stocks

Suboptimal Personalization: Without a granular understanding of how different companies truly group together based on their financial characteristics, delivering truly tailored and optimized portfolios becomes difficult.

Missed Opportunities: Inability to quickly identify and act on distinct pockets of opportunity (e.g., extreme growth, distressed value) that may not fit neatly into traditional categories.

This necessitates a more sophisticated approach to understanding the underlying structure of the stock market to enhance our advisory services and maintain our competitive edge.

- **Solution Approach**

1. **Data Analysis**

   - We utilized comprehensive financial metrics for a broad universe of stocks, including pricing, volatility, profitability (ROE, Net Income), liquidity (Cash Ratio, Net Cash Flow), and valuation ratios (P/E, P/B) for the analysis

2. **Data Cleaning**

   - Handle missing values: Checked and Impute missing data

   - Handle outliers: Identified and addressed extreme values that could skew the model

3. **Exploratory Data Analysis (EDA)**

- Analyze the distribution of individual features.
- Explore relationships between features and the GISC Sector using bar plots
- Identify highly correlated features to ensure that clusters are truly formed on distinct financial characteristics rather than amplified redundant information, ultimately leading to more robust and actionable segments.

4. **Data Preparation & Feature Scaling:**

- Ensured data quality, handled outliers, and performed feature scaling to standardize the influence of each financial metric, preventing disproportionate weighting

5. **Application of Diverse Clustering Techniques**

   i. **K-Means Clustering:** Employed this partitioning method to efficiently identify distinct, non-overlapping groups of stocks based on their centroids

   ii. **Hierarchical Clustering:** Utilized this method to reveal the nested relationships among stocks, providing a visual dendrogram and allowing for exploration of natural groupings at various levels. Different linkage methods (Ward, Complete, Average, Single) were evaluated to find the most representative hierarchy

6. **Robust Cluster Evaluation & Selection**

   i. **Optimal Cluster Determination:** Applied quantitative metrics such as the Elbow Method and Silhouette Scores (for K-Means) and the Cophenetic Correlation Coefficient (for Hierarchical Clustering) to objectively determine the most appropriate number of clusters.

   ii. **Dendrogram Analysis:** Visually inspected hierarchical dendrograms to confirm the interpretability and robustness of the chosen cluster structure.

7. Comprehensive Cluster Profiling

i. **Financial Persona Creation:** Analyzed the average financial characteristics of each identified cluster to develop rich, descriptive "personas" (e.g., "Growth Leaders," "Deep Value Play").

ii. **Company Identification:** Identified specific companies belonging to each segment, providing tangible examples for strategic decision-making.

This systematic approach ensured the development of actionable stock segments that will empower Trade&Ahead to deliver highly personalized and sophisticated investment solutions to its clientele.

EDA – Univariate Analysis

- **Current Price**

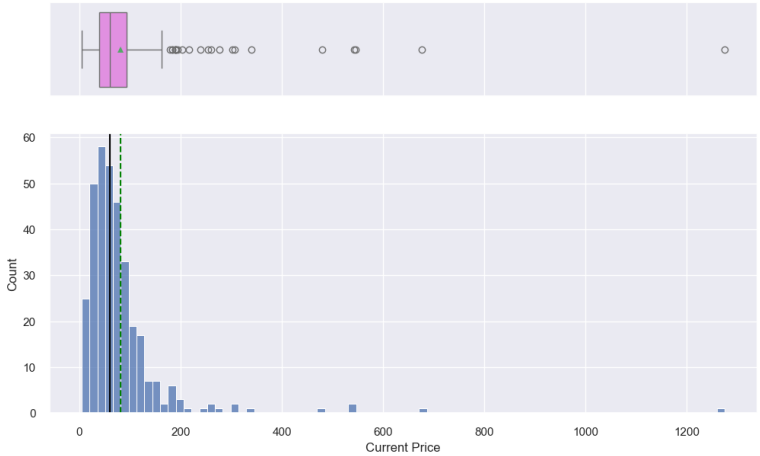


Fig 1: Univariate Analysis: Current Price

Observations:

- Right-skewed distribution with most prices under $100
- Mean is approximately $50–$60 with a lower median, confirming skewness
- The distribution is highly right-skewed, with a long tail extending towards higher prices.
- Most companies have low stock prices with a few having very high prices.

Link to Appendix slide on data background check

- **Price Change**

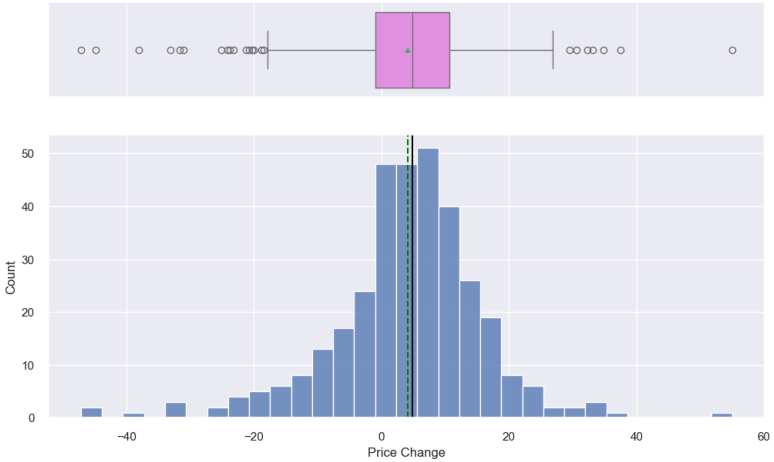


Fig 2: Univariate Analysis: Price Change

Observations:

- Price Change is roughly normally distributed, centered around 0%
- Mean and median are near zero, indicating minimal average change over 13 weeks
- Most price changes fall within -20% to +20%, with a narrow interquartile range
- Outliers on both ends show some stocks had large gains or losses
- The distribution suggests that while many stocks experienced minor price fluctuations around zero, there were also instances of considerable gains and losses

Link to Appendix slide on data background check

- **Volatility**

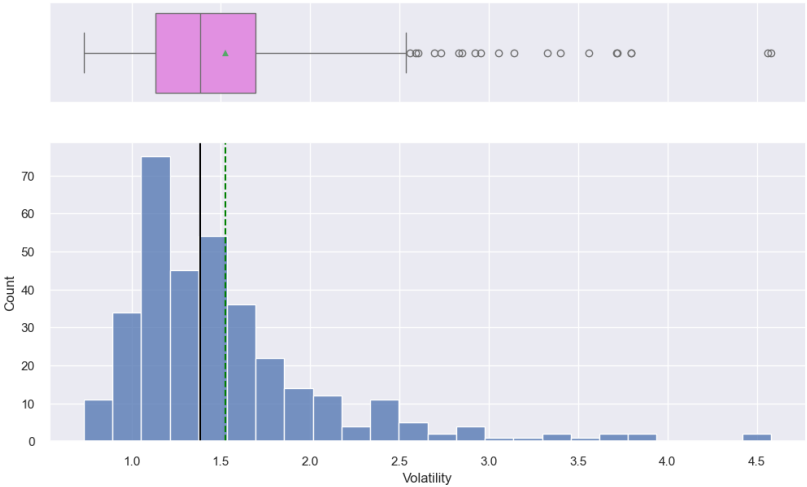


Fig 3: Univariate Analysis: Volatility

Observations:

- Volatility is right-skewed, with most companies having low to moderate volatility indicating that a significant number of stocks are considerably more volatile, suggesting higher risk or more dynamic price movements
- The mean of approximately 1.5 is greater than the median of approximately 1.4, confirming skewness.
- 50% of the companies have volatility within this relatively narrow range
- There are several outliers indicating some stocks have significantly higher volatility
- The distribution therefore suggest that investors should look to diversify their portfolios by managing risk exposure

Link to Appendix slide on data background check

- **ROE**

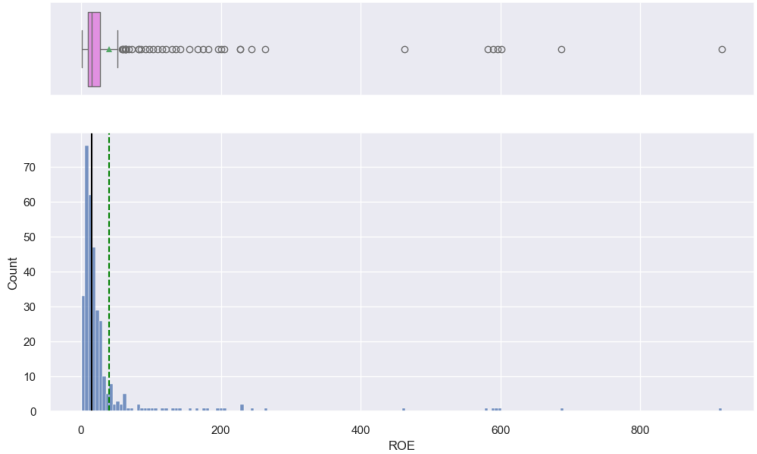


Fig 4: Univariate Analysis: ROE

Observations:

- ROE is extremely right-skewed, with most companies having ROE near zero. This suggests that for a large portion of the companies, the return generated on shareholders' equity is minimal or indicates slight losses

- The ROE values are tightly clustered near zero for most companies, with many extreme outliers on the high end (some over 800) that skew the mean but do not reflect the typical company

- Most companies have low ROE, but a few with extremely high ROE pull up the average, indicating high profitability variability and the importance of identifying top performers.

Link to Appendix slide on data background check

- **Cash Ratio**

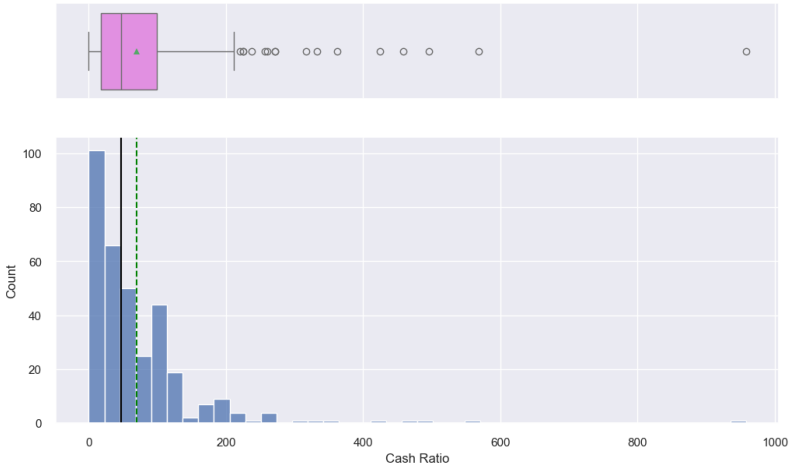


Fig 5: Univariate Analysis: Cash Ratio

Observations:

- Cash ratio distribution is heavily right-skewed, with most companies having low values
- The median cash ratio is low approximately between 50 to 60, while the mean is higher approximately between 100 to 120 due to outliers
- Interquartile range is narrow, showing 50% of companies have cash ratios concentrated in a low range.
- This distribution therefore indicates that most companies have low cash ratios, but a few with extremely high cash reserves skew the average upward, indicating high liquidity variability.

Link to Appendix slide on data background check

- **Net Cash Flow**

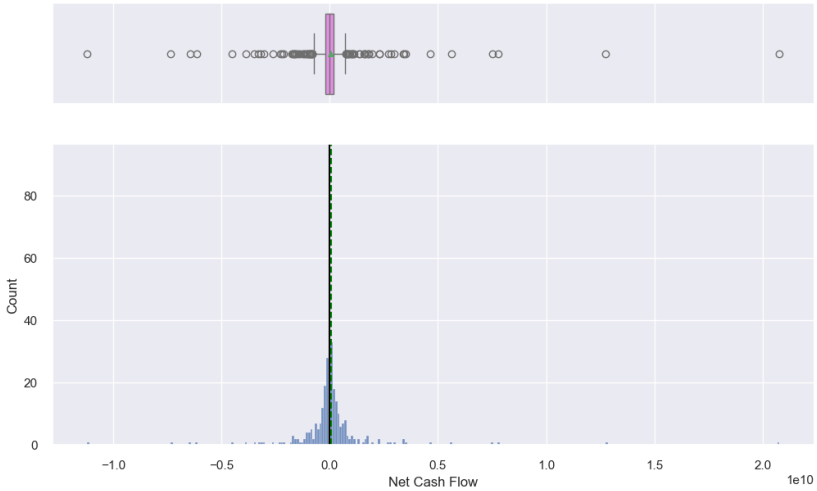


Fig 6: Univariate Analysis: Net Cash Flow

Observations:

- Most companies have net cash flow near zero, with both mean and median almost identical and the middle 50% of values tightly clustered around zero.
- Net Cash Flow values are tightly clustered near zero but have an extremely wide range with many outliers showing massive cash inflows and outflows on both ends
- The distribution thus suggest that most companies have stable or near-zero net cash flow, but extreme positive and negative outliers distort the mean, making it unreliable as a typical measure.

Link to Appendix slide on data background check

- **Net Income**

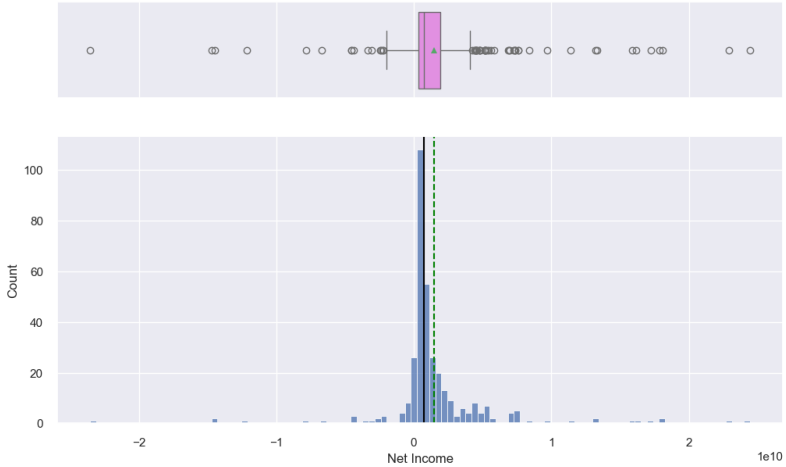


Fig 7: Univariate Analysis: Net Income

Observations:

- Net Income is sharply peaked near zero with long tails on both sides, indicating extreme outliers and a highly kurtotic distribution.
- Most companies have net income near zero or slightly negative, with both median and mean close to zero, indicating typical firms are near breakeven.
- The distribution suggests that most companies have net income near zero, but extreme profits and losses from a few outliers make the average misleading, reflecting high variability in profitability

Link to Appendix slide on data background check

- **Earnings Per Share**

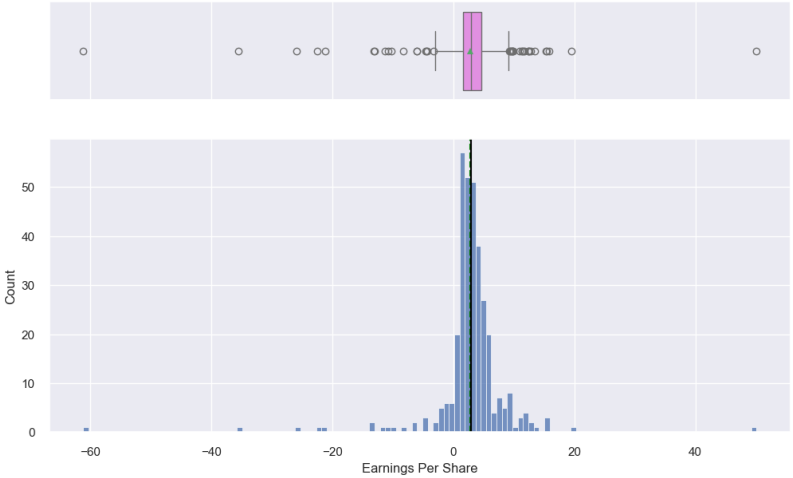


Fig 8: Univariate Analysis Earnings Per Share

Observations:

- The median earnings per share is positioned very close to zero. This indicates that for a significant portion of companies, their earnings per share are minimal or near breakeven.
- The interquartile range, is quite narrow. This suggests that the majority of companies have EPS values clustered very tightly around zero
- Earnings Per Share (EPS), similar to Net Income, exhibits extreme dispersion due to very large positive and negative outliers, highlighting a vast range in per-share profitability and significant financial disparity across the market.
- While many companies show modest or near-zero Earnings Per Share (EPS), the market also includes a few highly profitable firms and some struggling companies with substantial losses per share

Link to Appendix slide on data background check

- **Estimated Shares Outstanding**

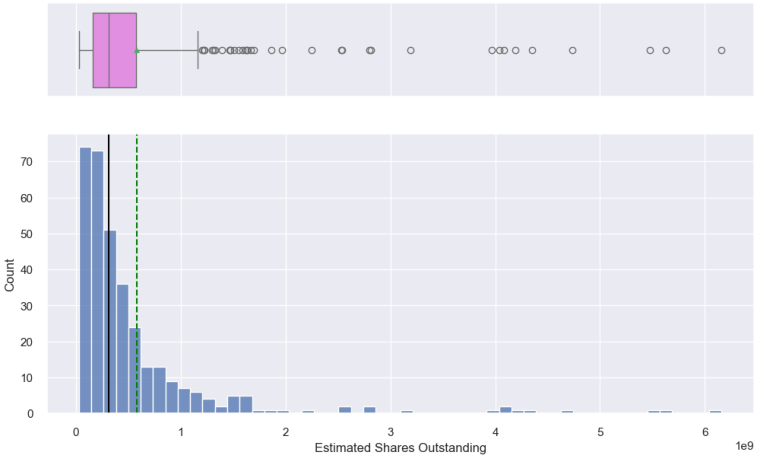


Fig 9: Univariate Analysis Estimated Shares Outstanding

Observations:

- Estimated shares outstanding are heavily right-skewed, with most companies having low values and a few having very high numbers, creating a prominent peak at the lower end

- 50% of companies have a narrow range of low shares outstanding, but numerous outliers exist with extremely high shares, some exceeding 500–600 million.

- The distribution hence indicates that most companies are smaller to mid-sized in shares outstanding, while a few large firms with very high shares skew the mean, making it unrepresentative of the typical company

Link to Appendix slide on data background check

- **P/E Ratio**

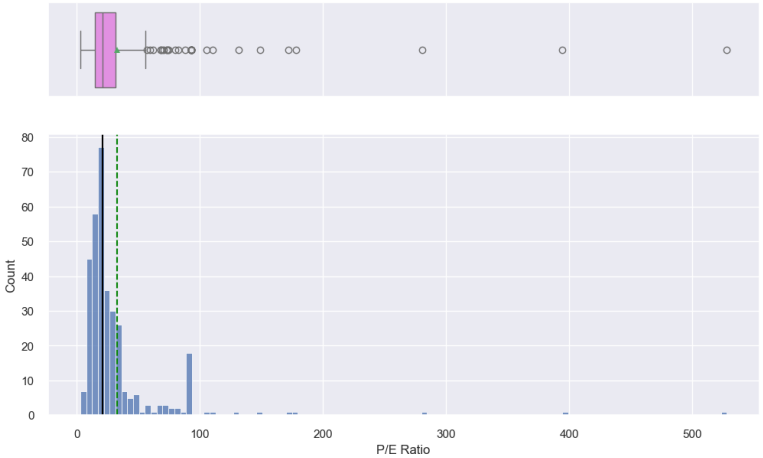


Fig 10: Univariate Analysis P/E Ratio

Observations:

- P/E Ratio distribution is highly right-skewed, with most companies having low P/E ratios and a long tail of much higher values.
- Most companies have a P/E ratio below 50, with the median around 15-20 and the mean higher due to right-skewed outliers, indicating a concentration of lower P/E values
- 50% of companies have low, tightly clustered P/E ratios, while numerous outliers extend far above 500, reflecting highly valued firms or those with very low earnings
- Most companies trade at modest P/E multiples, but a few have extremely high P/E ratios, making the average unrepresentative; the median is a better indicator for typical valuations.

Link to Appendix slide on data background check

- **P/B Ratio**

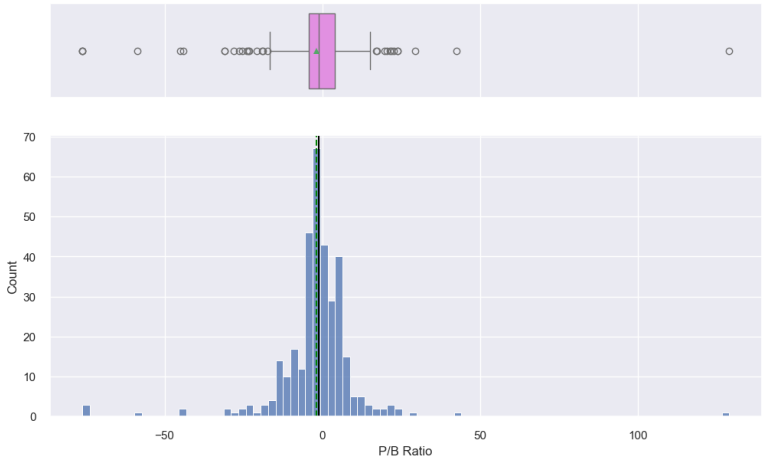


Fig 11: Univariate Analysis P/B Ratio

Observations:

- The distribution in the histogram observes that most P/B ratios are tightly clustered near zero, but numerous outliers on both sides extend from around -60 to 100+, indicating companies trading far below or above book value.

- From the box plot, most companies have P/B ratios near zero or slightly positive, with both median and mean close to zero, indicating typical firms trade near book value

- The P/B Ratio distribution suggests most companies trade near book value, but extreme outliers including negative P/B ratios indicate some are significantly under or overvalued, making the median a better measure than the mean for this dataset

Link to Appendix slide on data background check

- GICS Sector

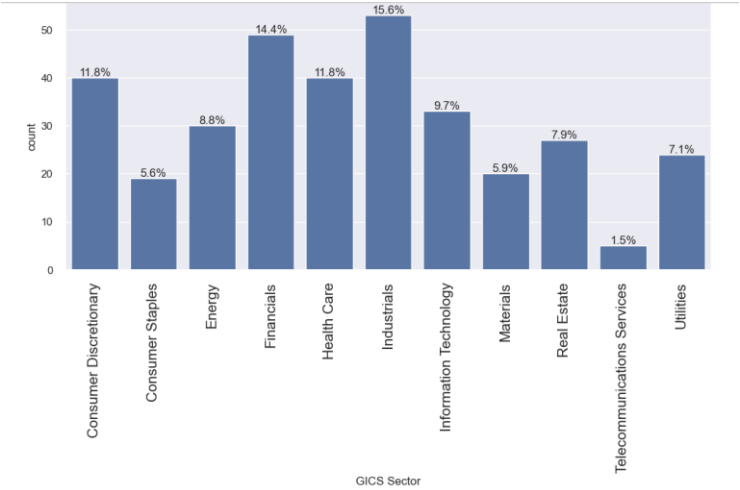


Fig 12: Univariate Analysis GICS Sector

Observations:

- Industrials is the most prevalent GICS sector, accounting for 15.6% of the companies. Financials also have a significant presence at 14.4%.

- Consumer Discretionary, Health Care, and Information Technology each represent a notable portion of the dataset, with 11.8%, 11.8%, and 9.7% respectively. Real Estate and Utilities also show moderate representation at 7.9% and 7.1% respectively

- Energy (8.8%), Materials (5.9%), and Consumer Staples (5.6%) are less represented compared to the top sectors.

- Telecommunications Services is the least represented sector by a significant margin, making up only 1.5% of the companies in the dataset

- The dataset has a varied sector distribution, emphasizing Industrials and Financials while underrepresenting Telecommunications, which should be considered in cluster analysis

Link to Appendix slide on data background check

- **GICS Sub Industry**

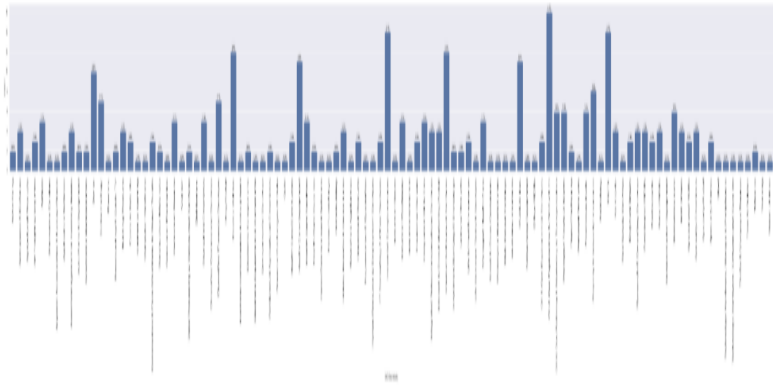


Fig 13: Univariate Analysis GICS Sub Industry

Observations:

- The dataset covers a very large number of distinct GICS Sub Industry categories, highlighting detailed industry classification
- Company counts vary significantly across sub-industries, with some categories (e.g., Oil & Gas Exploration & Production, REITs, Industrial Conglomerates) being highly represented
- Many sub-industries contain only a few companies, indicating the wide breadth and often thin representation of niche segments within the dataset
- The data reflects a diverse and fragmented industry landscape, emphasizing the need for cluster analysis to identify meaningful groupings beyond just broad sector definitions.

Link to Appendix slide on data background check

EDA – Bivariate Analysis

- **Correlation Check**

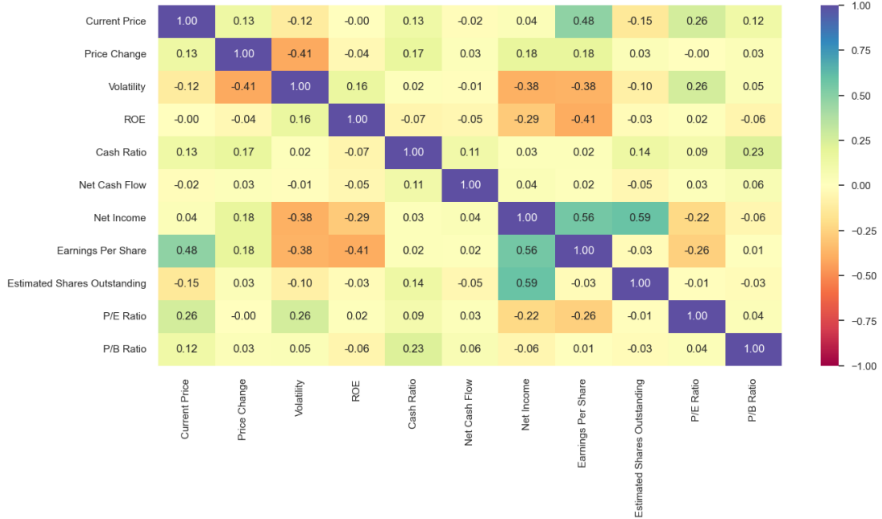


Fig 14: Correlation Plot

Observations:

- Heatmap shows pairwise correlations across numerical features
- Net income strongly correlates with both earnings per share (0.56) and net cash flow (0.59).
- Current price correlates positively with EPS (0.48), confirming earnings influence valuation
- Net income correlates with shares outstanding (0.59), indicating bigger firms generate higher profits
- Volatility negatively correlates with EPS and net income (~ -0.38), reflecting instability
- ROE shows weak or negative correlations with EPS and net income, likely due to outliers or data issues.
- Many variables show weak interdependence: Price change, cash ratio, and P/B ratio have generally low correlations with other metrics

Link to Appendix slide on data background check

- **Price Change vs GICS Sector**

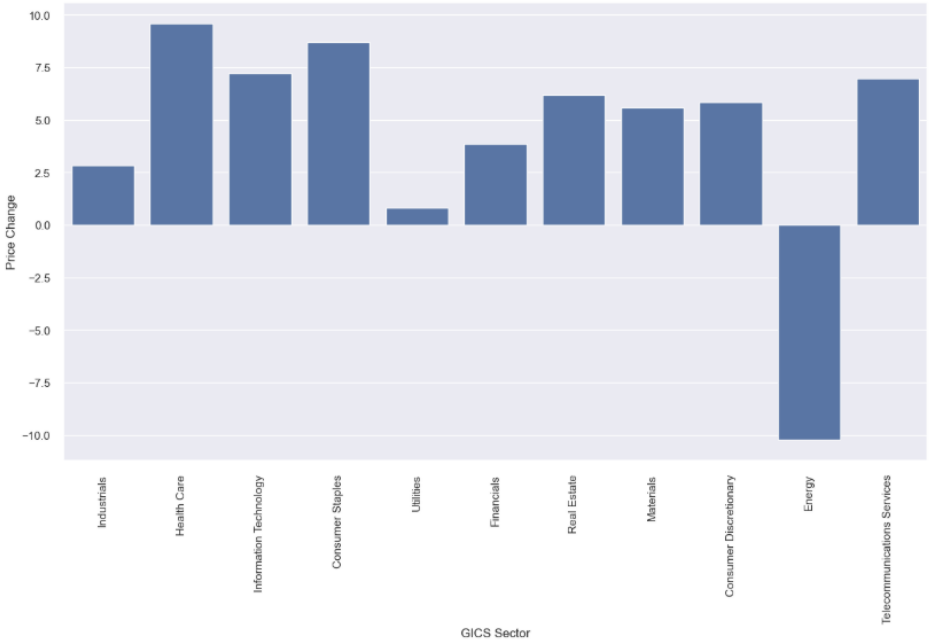


Fig 15: Price Change vs GICS Sector

Observations:

- Health Care shows the highest positive price change, close to 10% followed by Consumer Staples of just under 9%
- Information Technology, Materials and Telecommunication show moderate positive price changes, around 7.5%, 6% and 7% respectively.
- Real Estate and Consumer Discretionary have similar positive price changes, both slightly above 6%.
- The Industrials, Financials, and Utilities sectors all experienced modest positive price changes, ranging from approximately 1% to 3.5%
- Energy stands out with a substantial negative price change, approximately -10%
- In the nutshell most sectors rose, led by Health Care and Consumer Staples, while Energy saw a sharp decline.

Link to Appendix slide on data background check

- **Cash Ratio vs GICS Sector**

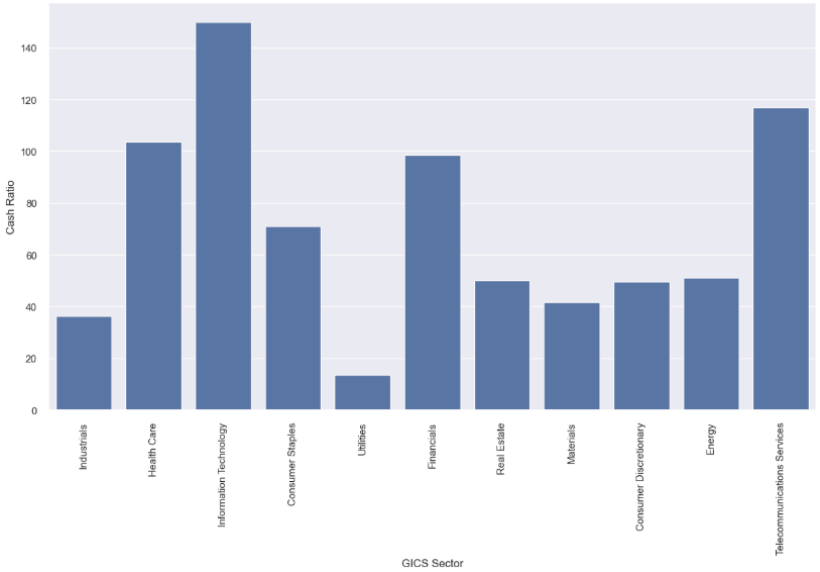


Fig 16: Cash Ratio vs GICS Sector

Observations:

- Information Technology has the highest Cash Ratio, significantly above 140. This suggests that companies in this sector tend to hold a very high proportion of their current liabilities in cash or cash equivalents

- Telecommunications Services, Health Care, and Financials have high Cash Ratios (around 100–120), while other sectors show moderate to low ratios, with Industrials the lowest at around 35.

- Utilities has the lowest Cash Ratio among all sectors, barely above 10. This indicates that utility companies tend to operate with very low cash reserves relative to their current liabilities.

- Information Technology and Telecommunications Services sectors therefore appear to have the strongest liquidity in terms of cash on hand, while the Utilities sector has the weakest.

Link to Appendix slide on data background check

- P/E Ratio vs GICS Sector

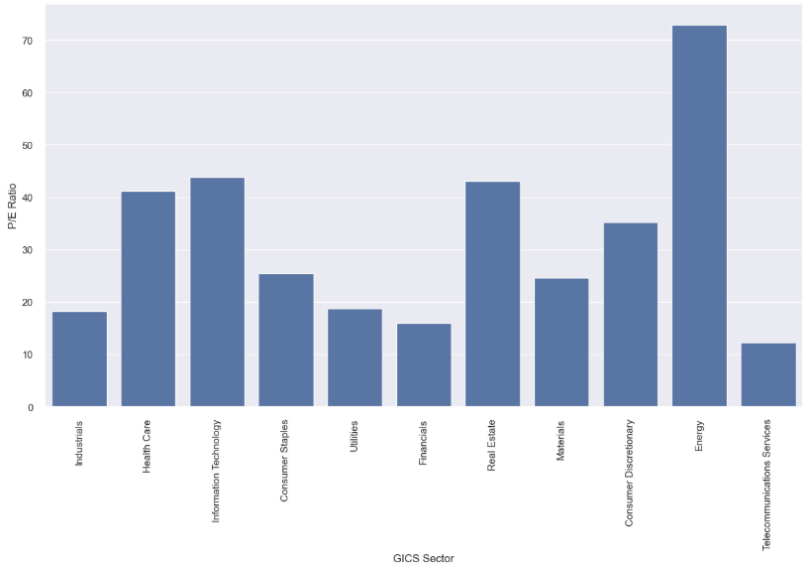


Fig 17: P/E Ratio vs GICS Sector

Observations:

- Energy has by far the highest P/E Ratio, exceeding 70. This could suggest high investor expectations for future earnings growth in this sector, or perhaps a period of depressed earnings relative to stock price

- Real Estate, Information Technology, and Health Care show relatively high P/E Ratios (around or above 40), while Financials show below 20, with other sectors falling in between.

- Telecommunications Services has the lowest P/E Ratio, around 15.

- The Energy sector has the highest P/E Ratio among all sectors, followed by Real Estate, Information Technology, and Health Care, while Telecommunications Services and Financials have the lowest

Link to Appendix slide on data background check

- **Volatility vs GICS Sector**

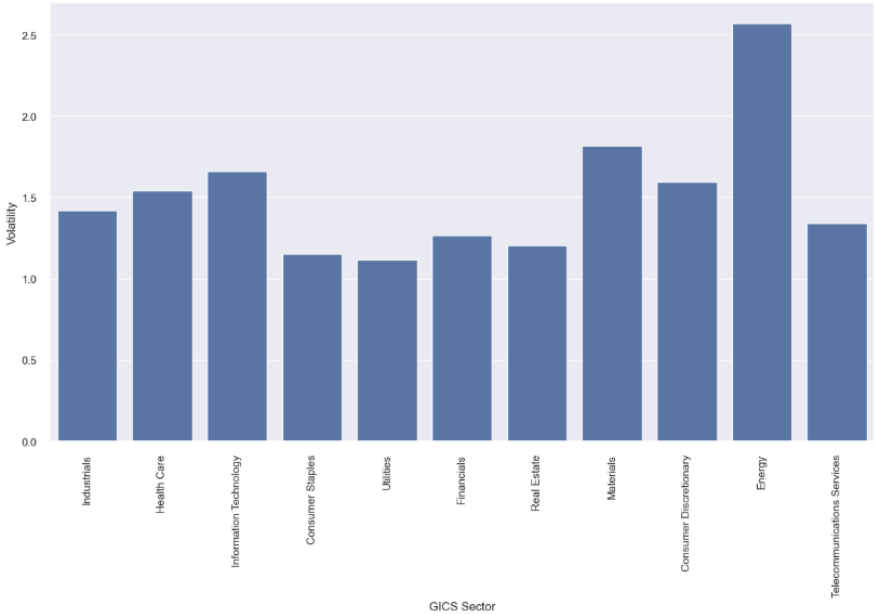


Fig 18: Volatility vs GICS Sector

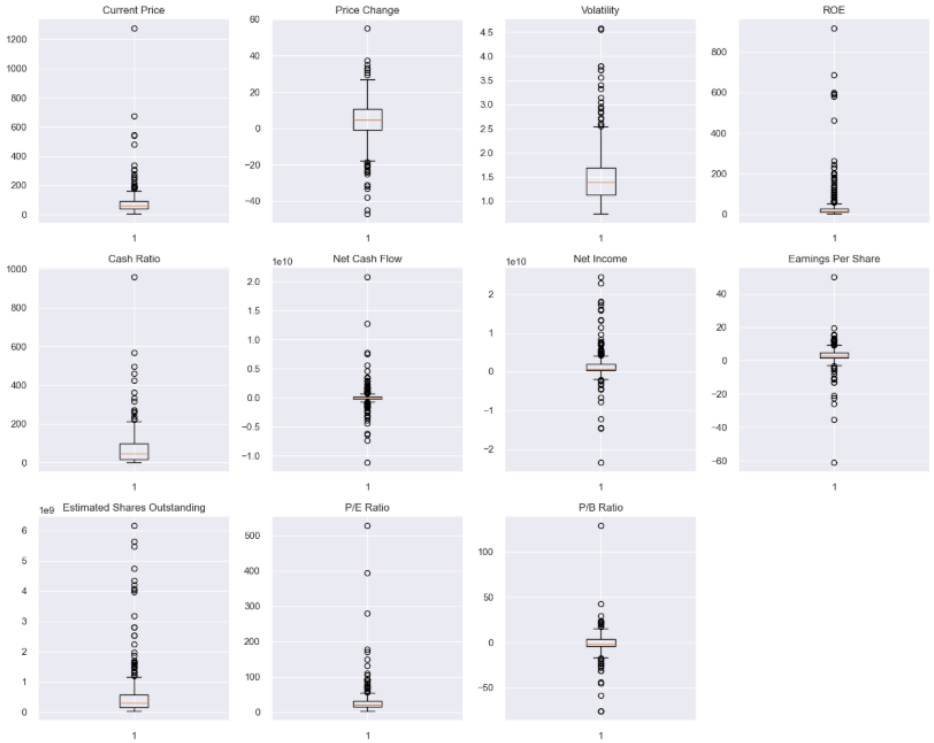
Observations:

- Energy has the highest volatility, significantly above 2.5. This indicates that stock prices in the Energy sector have experienced the largest fluctuations over the past 13 weeks
- Materials exhibits relatively high volatility (above 1.8), followed by Information Technology and Consumer Discretionary (above 1.5), with Health Care, Industrials, and Telecommunications Services showing moderate to slightly lower levels
- Financials and Real Estate show similar and slightly lower volatility, both around 1.2
- Consumer Staples and Utilities have the lowest volatility among all sectors, both just above 1.0

Link to Appendix slide on data background check

Data Preprocessing

- Outlier Check



- **Positive Skewness/Upper Outliers:** Many variables, including Current Price, Volatility, ROE, Cash Ratio, Estimated Shares Outstanding, P/E Ratio, and P/B Ratio, show a significant number of outliers on the higher end, indicating that a small number of companies have exceptionally high values for these metrics.
- **Both Positive and Negative Outliers:** Price Change, Net Cash Flow, Net Income, and Earnings Per Share exhibit outliers on both the positive and negative sides, reflecting companies with extreme positive or negative financial performance

- There are no missing values in the dataset
- There are no duplicate values

- Feature Scaling

	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
0	-0.393341	0.493950	0.272749	0.989601	-0.210698	-0.339355	1.554415	1.309399	0.107863	-0.652487	-0.506653
1	-0.220837	0.355439	1.137045	0.937737	0.077269	-0.002335	0.927628	0.056755	1.250274	-0.311769	-0.504205
2	-0.367195	0.602479	-0.427007	-0.192905	-0.033488	0.454058	0.744371	0.024831	1.098021	-0.391502	0.094941
3	0.133567	0.825696	-0.284802	-0.317379	1.218059	-0.152497	-0.219816	-0.230563	-0.091622	0.947148	0.424333
4	-0.260874	-0.492636	0.296470	-0.265515	2.237018	0.133564	-0.202703	-0.374982	1.978399	3.293307	0.199196

- Materials shows the highest volatility, while Information Technology, Consumer Discretionary, Health Care, Industrials, and Telecommunications Services exhibit moderate levels.
- Standardization removes the original units and direct interpretability of raw values, replacing them with values representing standard deviations from the mean
- Though absolute values change, standardization preserves the relative relationships and general distribution shape within each feature, maintaining the original data's comparative structure.
- While extreme outliers are less visually apparent after scaling, large absolute scaled values still indicate their presence, reflecting original data points far from the mean.
- The feature scaling has successfully normalized the data, making all features contribute equally to distance calculations, which is crucial for the subsequent cluster analysis

K-Means Clustering Summary

- **Objective**: K-Means clustering was used to segment 340 stocks into meaningful investment groups based on standardized financial metrics enabling better client alignment, risk management, and research efficiency
- **Optimal Number of clusters using K-Means:**
  - Three robust clusters (k=3) identified using the Elbow Method and Silhouette Scores.
- **Cluster Profiling:**

i. **Cluster 0 - "Balanced & Diversified Core" Cluster (306 companies)**

- Represents average performers across metrics
- Suitable for most investors seeking diversified, stable exposure
- Ideal for core portfolio construction and low-risk stock picking

Link to Appendix slide on K-Means Clustering

ii. Cluster 1 - "Growth / High Performers" Clusters (32 companies)

- High performance: strong price momentum, profits, and premium valuations.
- Includes tech and leading energy names like Amazon, Netflix
- Best for aggressive, growth-oriented investors

iii. Cluster 2 - "Deep Value / Underperforming Energy" Cluster (2 companies)

- Low volatility, negative profits, undervalued (deep value).
- Includes firms like Apache and Chesapeake Energy
- Appeals to contrarian investors seeking high-risk, high-reward opportunities

Link to Appendix slide on K-Means Clustering

Hierarchical Clustering Summary

- **Objective**: Hierarchical clustering was used to segment 340 publicly traded companies into three distinct groups, based on standardized financial metrics. This method uncovers nested relationships between stocks, offering a layered and interpretable view of market structure beyond sector or size.

- **Optimal Number of clusters using Hierarchical Clustering:**

  - The visual inspection of the Ward's linkage dendrogram, which had the highest Cophenetic Correlation Coefficient also strongly supported cutting the dendrogram to yield 3 distinct clusters.

- **Cluster Profiling:**

  i. **Cluster 0 - "Balanced & Diversified Core" (337 companies)**

  - Largest cluster with moderate price, profitability, and volatility, ideal for balanced, risk-aware investing.
  - Broadly diversified across sectors like Tech, Financials, Health Care, Energy, and Consumer Staples
  - Grouping driven by similar financial behaviors rather than industry classification
  - Key companies featured are Amazon, JPMorgan Chase, Pfizer, Chevron, Procter & Gamble, McDonald's, and others

Link to Appendix slide on Hierarchical Clustering

ii. Cluster 1 - "Growth / High Performers" Clusters (1 company)

- Contains only Priceline.com Inc, isolated due to its exceptional financial performance.
- Characterized by superior profitability, growth, and valuation metrics.
- Too financially distinct to be grouped with other companies in the market.
- Acts as a benchmark for high-growth or standalone investment opportunities.

iii. Cluster 2 - "Deep Value / Underperforming Energy" Cluster (2 companies)

- Comprises Apache Corporation and Chesapeake Energy, marked by financial underperformance.
- Defined by low profitability, low valuation, and signs of financial distress
- Indicates high volatility and poor fundamentals, warranting investor caution
- May attract value or turnaround investors looking for high-risk opportunities.

Link to Appendix slide on K-Means Clustering

APPENDIX

Data Background and Contents

- **Data Overview**

	Ticker Symbol	Security	GICS Sector	GICS Sub Industry	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio
102	DVN	Devon Energy Corp.	Energy	Oil & Gas Exploration & Production	32.000000	-15.478079	2.923698	205	70	830000000	-14454000000	-35.55	4.065823e+08	93.089287
125	FB	Facebook	Information Technology	Internet Software & Services	104.660004	16.224320	1.320606	8	958	592000000	3669000000	1.31	2.800763e+09	79.893133
11	AIV	Apartment Investment & Mgmt	Real Estate	REITs	40.029999	7.578608	1.163334	15	47	21818000	248710000	1.52	1.636250e+08	26.335526
248	PG	Procter & Gamble	Consumer Staples	Personal Products	79.410004	10.660538	0.806056	17	129	160383000	636056000	3.28	4.913916e+08	24.070121
238	OXY	Occidental Petroleum	Energy	Oil & Gas Exploration & Production	67.610001	0.865287	1.589520	32	64	-588000000	-7829000000	-10.23	7.652981e+08	93.089287
336	YUM	Yum! Brands Inc	Consumer Discretionary	Restaurants	52.516175	-8.698917	1.478877	142	27	159000000	1293000000	2.97	4.353535e+08	17.682214
112	EQT	EQT Corporation	Energy	Oil & Gas Exploration & Production	52.130001	-21.253771	2.364883	2	201	523803000	85171000	0.56	1.520911e+08	93.089287
147	HAL	Halliburton Co.	Energy	Oil & Gas Equipment & Services	34.040001	-5.101751	1.966062	4	189	7786000000	-671000000	-0.79	8.493671e+08	93.089287
89	DFS	Discover Financial Services	Financials	Consumer Finance	53.619999	3.653584	1.159897	20	99	2288000000	2297000000	5.14	4.468872e+08	10.431906
173	IVZ	Invesco Ltd.	Financials	Asset Management & Custody	33.480000	7.067477	1.580839	12	67	412000000	968100000	2.26	4.283628e+08	14.814159

Table 1: Top few rows of the Dataset

- The dataset contains 340 rows and 15 columns representing financial data and stock indicators for several companies listed on the New York Stock Exchange.

- It includes information such as Ticker Symbol, Company Name, GICS Sector and Sub-Industry, Current Price, Price Change (13 weeks), Volatility (13 weeks), ROE, Cash Ratio, Net Cash Flow, Net Income, Earnings Per Share, Estimated Shares Outstanding, P/E Ratio, and P/B Ratio.

- **Data Background**

The dataset stock_data is intended for analyzing and grouping stocks to provide personalized investment strategies.

- **Data Contents**

```
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
```

Table 2: Information on the Data Set

There are three datatypes namely: float64(7), int64(4) and object(4) with 11 numerical and 4 categorical .

Data Description

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

Statistical Summary

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Ticker Symbol	340	340	AAL	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Security	340	340	American Airlines Group	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

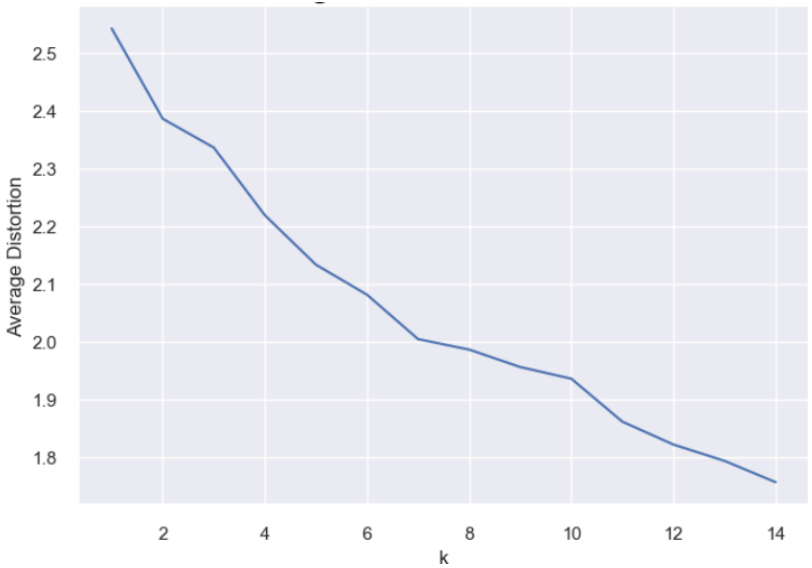
Table 3: Statistical Summary of the Dataset

Observations

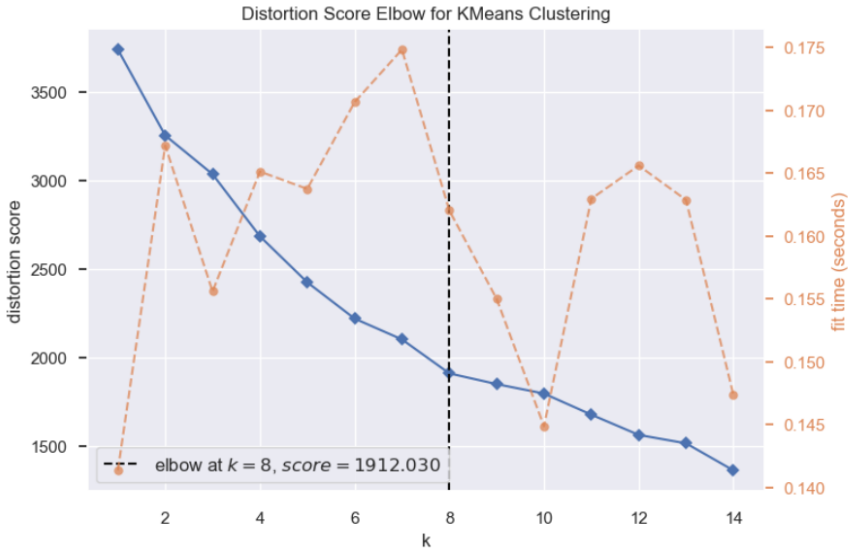
- **Diverse Market Coverage:** Includes 340 unique companies across 11 sectors and 104 sub-industries
- **Wide Price & Performance Range:** Stock prices range from $4.50 to $1,275, with price changes spanning –47% to +55%, reflecting diverse market behavior
- **High Variability in Financial Health:** Large disparities in ROE (1 to 917), net income (–$23.5B to $24.4B), and cash ratios highlight mix of strong performers and distressed firms
- **Valuation Outliers Present**: Extreme P/E (up to 528) and negative P/B ratios indicate presence of growth stocks and deeply undervalued/distressed companies

K-Means Clustering Technique

- **Selecting K with the Elbow Method**

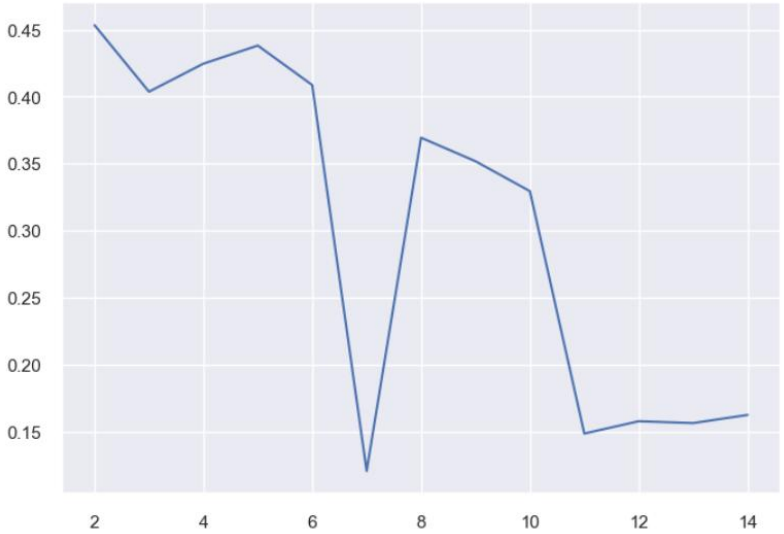


- A sharp WCSS drop from K=1 to K=2 indicates significant compactness gain, with further improvement from K=2 to K=3 suggesting added value in a third cluster
- A clear "elbow" at K=3 indicates diminishing returns in WCSS reduction beyond this point, as further decreases become less significant
- Beyond K=3, WCSS reduction tapers off, indicating diminishing returns and more fragmented clusters with limited added insight.

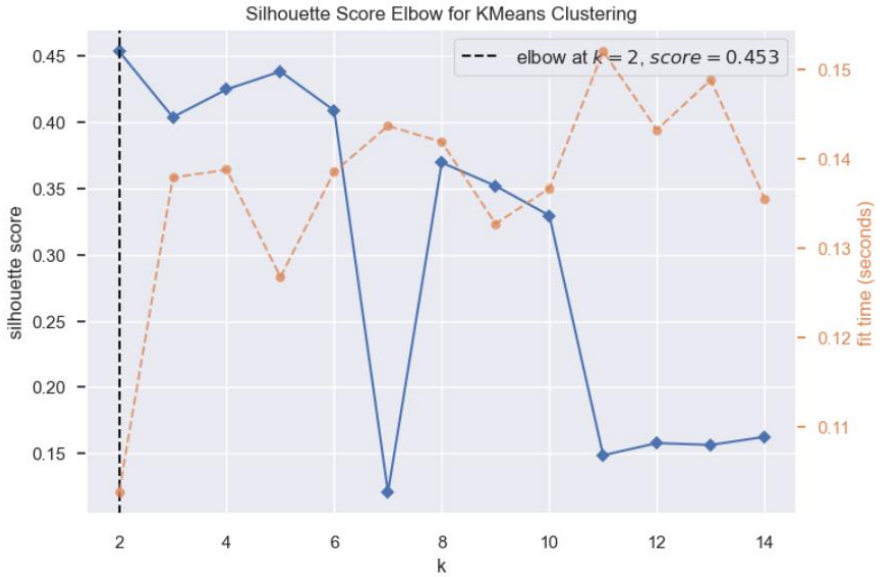


- Optimal cluster count identified at K=8, where the distortion score significantly levels off (elbow point).
- Distortion score drops steeply from K=1 to K=8, indicating improved compactness with more clusters.
- Beyond K=8, improvements in distortion are marginal, suggesting diminishing returns from additional clusters.
- Fit time remains relatively stable, showing clustering efficiency is not significantly affected by increasing K.

Checking the Silhouette scores

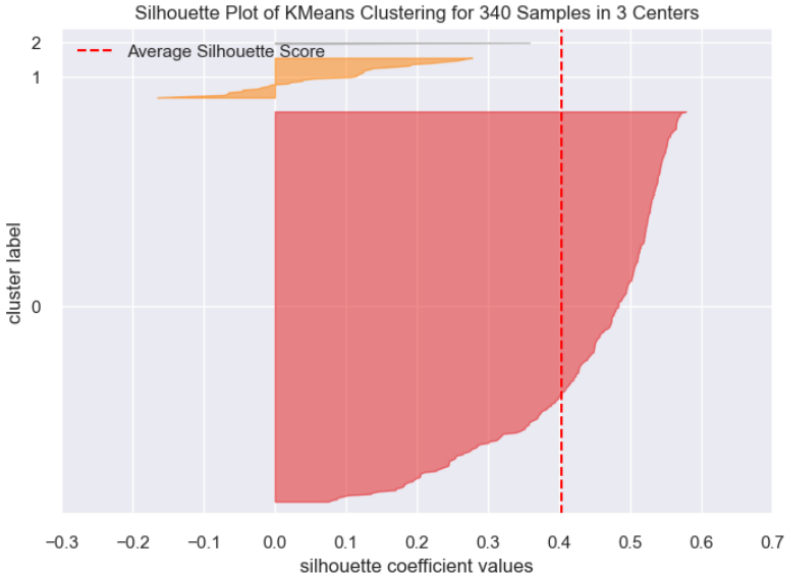


- The plot shows that the highest Silhouette Score is achieved when the number of clusters (K) is 2, indicating the most distinct and well-separated
- Beyond K=2, the Silhouette Score generally decreases as the number of clusters increases, suggesting that the clusters become less compact or less well-separated from each other
- For K=3, the Silhouette Score remains positive (though lower than at K=2), implying that the clusters are still reasonably defined and exhibit some degree of separation and cohesion



- The plot clearly shows that the highest Silhouette Score is achieved with K=2 clusters, suggesting that stocks are most distinctly separated into two main groups.
- As the number of clusters increases beyond 2, the Silhouette Score generally decreases, indicating that the clusters become less well-defined, more overlapping, or less compact
- While not the peak, the Silhouette Score at K=3 (our chosen optimal count from the Elbow Method) remains positive, implying that the three identified stock clusters are still reasonably well-separated and cohesive

Based on the comprehensive analysis using both the Elbow Method and the Silhouette Score plot, K=3 is the recommended optimal cluster count for K-Means clustering since it provides a robust and interpretable segmentation that aligns well with the objective of grouping stocks based on their attributes for personalized investment strategies



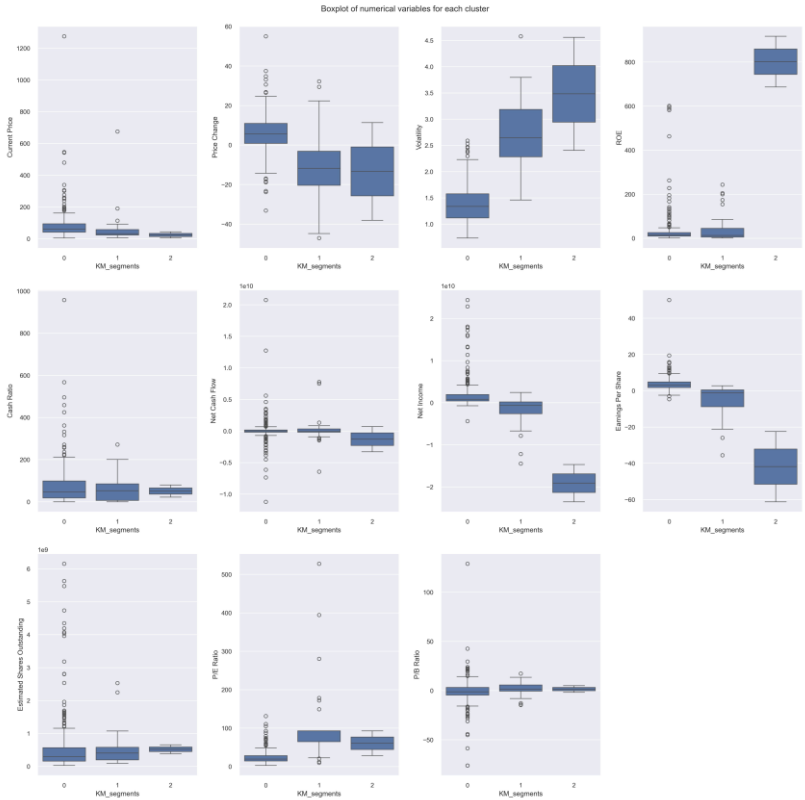
- **Cluster 0:** This cluster contains the vast majority of the samples with high positive silhouette scores, indicating strong internal cohesion and separation.
- **Cluster 1:** This cluster is much smaller than Cluster 0 in terms of the number of samples showing mixed silhouette scores, with negatives (down to -0.15) suggesting some samples fit better in other clusters.
- **Cluster 2:** This cluster is the smallest and shows low clustering quality, with mixed silhouette scores (–0.05 to 0.2) indicating several poorly assigned samples.

Cluster Profiling

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	
KM_segments												
0	82.966657	5.704375	1.387064	33.859477	70.186275	48658872.55	1996387438	3.899886	582707542.9	24.291397	-2.106344	306
1	64.263438	-10.38278	2.732033	46.8125	69.65625	205568000	-2018462219	-5.177187	526311835.9	110.425211	1.787692	32
2	24.485001	-13.351992	3.482611	802	51	-1292500000	-1.9107E+10	-41.815	519573983.3	60.748608	1.565141	2

- **Segment 0 (306 Companies):** Represents the majority, characterized by high current prices, positive price changes, strong profitability (Net Income, EPS), and robust cash flow, suggesting stable and growing companies. (Note: P/E and EPS values may be aggregated).
- **Segment 1 (32 Companies):** Consists of companies experiencing price declines, negative net income/EPS, and higher volatility, indicating a struggling or underperforming group.
- **Segment 2 (2 Companies):** A very small, highly distressed segment with the steepest price declines, massive losses (negative Net Income/EPS), and extremely unusual financial ratios (e.g., very high Cash Ratio, extremely high/negative Net Cash Flow), likely representing outliers or data anomalies.

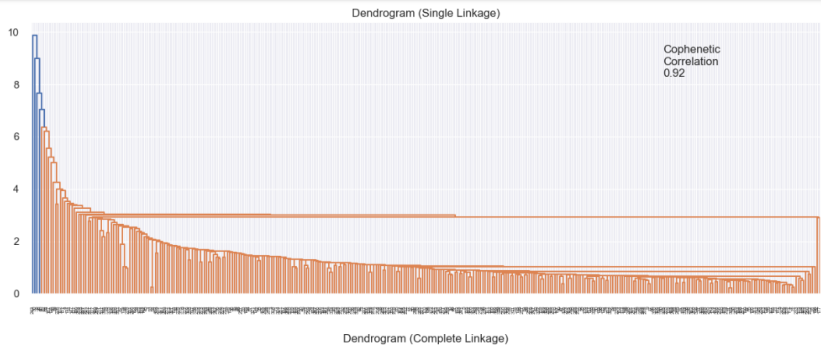
Box Plot for numerical variables for each Cluster

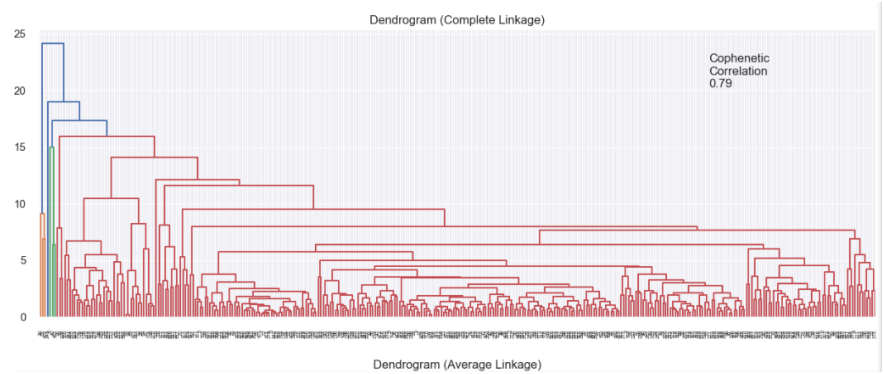


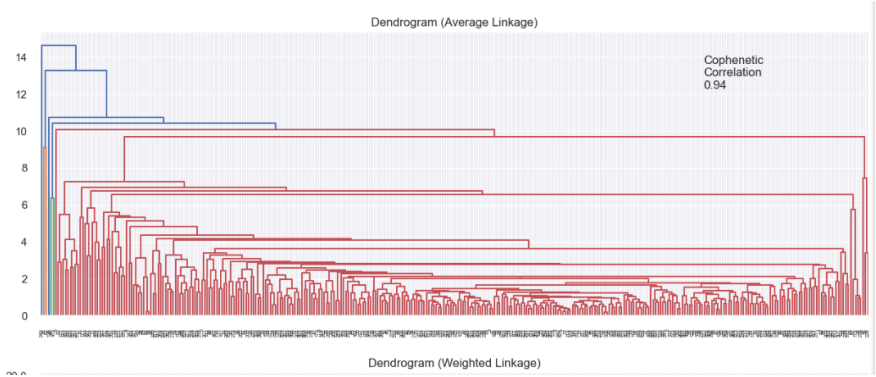
- **KM Segment 0 (Majority):** Characterized by generally stable prices, positive price changes, profitability (Net Income, EPS), and healthy cash flow.
- **KM Segment 1 (Underperforming):** Shows negative price changes, higher volatility, and often negative net income and EPS.
- **KM Segment 2 (Extreme Outliers):** Displays the most severe price declines, highest volatility, and substantial losses (negative Net Income, EPS); however, extremely high ROE and Cash Ratio for this segment are anomalous and strongly suggest data errors or highly unusual financial structures.
- **Distinct Cluster Profiles:** Each segment exhibits unique distributions across financial metrics, clearly separating "healthy," "underperforming," and "extreme outlier" groups identified by KMeans Clustering

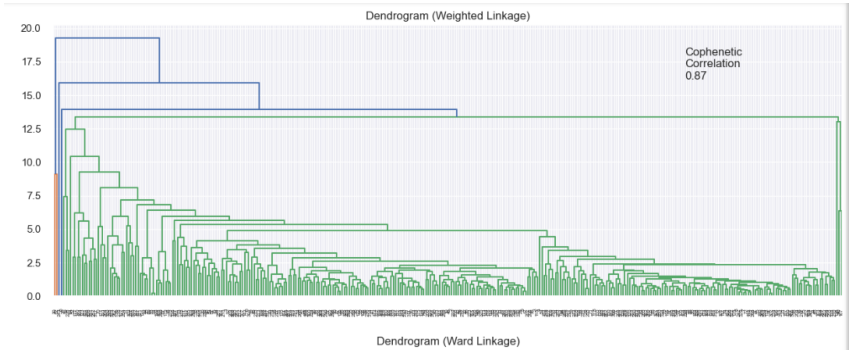
Hierarchical Clustering Technique

- **Checking Dendrograms**



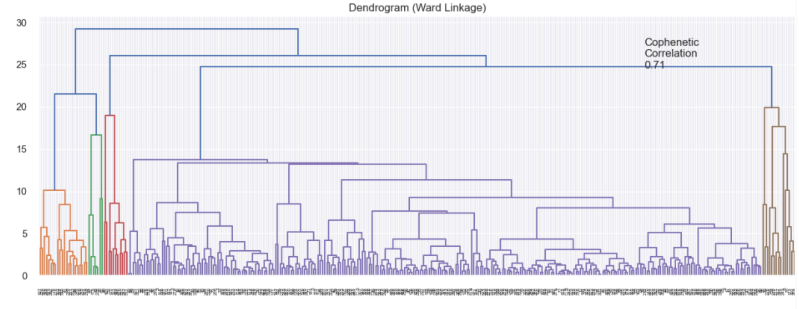






Hierarchical Clustering Technique

- **Checking Dendrograms**



- **correlations for different linkage methods**

	Linkage	Cophenetic Coefficient
4	ward	0.710118
1	complete	0.787328
3	weighted	0.869378
0	single	0.923227
2	average	0.942254

- **Average Linkage is Best:** With a Cophenetic Coefficient of 0.942254, "average" linkage provides the most accurate representation of the original data distances among all methods tested.
- **Single Linkage is Also Strong:** Surprisingly, "single" linkage also shows a very high Cophenetic Coefficient (0.923227), suggesting it's a good fit despite its common "chaining" characteristic.
- **Weighted Linkage is Good:** "Weighted" linkage performs well with a coefficient of 0.869378.
- **Complete and Ward Linkage are Less Optimal:** "Complete" (0.787328) and "ward" (0.710118) linkages have lower coefficients, indicating they preserve the original distances less faithfully compared to average, single, and weighted linkages for the dataset provided.

Hierarchical Clustering Technique

- **Cluster Profiling**

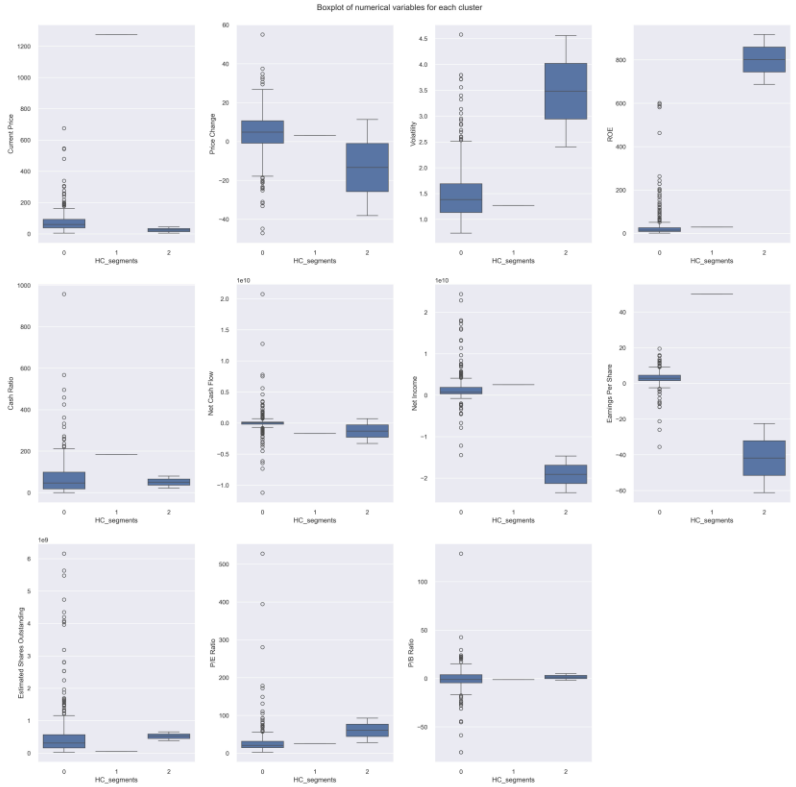
	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P/B Ratio	count_in_each_segment
HC_segments												
0	77.653642	4.184271	1.515129	35.103858	69.79822	68662246.29	1613508620	2.900905	578930419.4	32.466828	-1.739711	337
1	1274.949951	3.190527	1.26834	29	184	-1671386000	2551360000	50.09	50935516.07	25.453183	-1.052429	1
2	24.485001	-13.351992	3.482611	802	51	-1292500000	-1.9107E+10	-41.815	519573983.3	60.748608	1.565141	2

- **Segment 0 (337 Companies):** The largest cluster, characterized by positive price changes, good profitability (Net Income, EPS), strong cash flow, and healthy cash ratios. This segment represents the bulk of the "stable and performing" companies.

- **Segment 1 (1 Company):** An outlier segment representing a single, highly unique company with an extremely high current price, significant net income, but surprisingly negative net cash flow. This indicates a very unusual financial profile warranting individual investigation.

- **Segment 2 (2 Companies):** A very small cluster of highly distressed or extreme outlier companies, evidenced by the most severe price declines, massive net income losses, and highly unusual (e.g., negative, very large) financial metrics.

Box Plot for numerical variables for each Cluster



- **HC Segment 0 (Majority):** Represents the bulk of companies, showing stable prices, positive growth, profitability (Net Income, EPS), and healthy cash flow.
- **HC Segment 1 (Single Outlier):** Isolates one company with an exceptionally high current price and net income, but notably negative net cash flow, making it a unique financial outlier.
- **HC Segment 2 (Extreme Distress):** Comprises a few companies exhibiting the most severe price declines, highest volatility, and substantial losses; extremely high ROE and Cash Ratio in this segment are anomalous and may suggest data issues or highly unusual financial structures.
- **Clear Segment Differentiation:** The plots demonstrate clear distinctions between the healthy core, the unique outlier, and the highly distressed companies identified by hierarchical clustering.

K-Means vs Hierarchical Clustering

- Both K-Means and Hierarchical Clustering are powerful techniques for segmenting stock data. The exploration revealed interesting similarities and subtle differences:

    1. **Time taken for execution:** K-Means clustering generally tends to be faster than Hierarchical Clustering

    2. **Cluster Comparison:**

        - **Distinctness:** Both K-Means and Hierarchical Clustering produced distinct clusters; K-Means provided a broader view of "growth" companies, while Hierarchical Clustering isolated extreme outliers offering a more granular distinction

        - **Similarities in Cluster Identification**: The "Deep Value / Underperforming Energy" cluster was identical in both algorithms, containing two companies, while the "Broad Market / Diversified Core" clusters were highly similar in composition but differed in size, with K-Means having 306 companies and Hierarchical Clustering having 337.

3. **Observations in the similar clusters of both algorithms:**

   i. **"Deep Value / Underperforming Energy" Cluster:**

      - **K-Means Cluster 2:** 2 companies (Apache Corporation, Chesapeake Energy)
      - **Hierarchical Cluster 2:** 2 companies (Apache Corporation, Chesapeake Energy)
      - **Number of Observations in Similar Clusters**: 2 observations (100% overlap)

   ii. **"Balanced & Diversified Core" Cluster:**

      - **K-Means Cluster 0:** 306 companies
      - **Hierarchical Cluster 0:** 337 companies)
      - **Number of Observations in Similar Clusters**: The "Broad Market / Diversified Core" clusters from both algorithms, though not identical, are largely similar, with differences likely due to how each algorithm classified companies with moderate growth.

iii. **"Growth / High Performers" Clusters:**

- **K-Means Cluster 1:** 32 companies (a broader growth segment)
- **Hierarchical Cluster 1:** 1 company (Priceline.com Inc, an extreme outlier)
- **Number of Observations in Similar Clusters**: These clusters represent the "growth" theme, but their sizes differ significantly due to Hierarchical Clustering's ability to isolate the single most extreme outlier

4. **Number of Clusters obtained as the appropriate number of clusters from both algorithms:** Both K-Means and Hierarchical Clustering consistently suggested that 3 clusters are the appropriate number of segments for your stock dataset.

- **K-Means:** The Elbow Method and Silhouette Score analysis both indicated an optimal k = 3
- **Hierarchical Clustering**: The visual inspection of the Ward's linkage dendrogram (which had the highest Cophenetic Correlation Coefficient) also strongly supported cutting the dendrogram to yield 3 distinct clusters

References

Great Learning. (n.d.) *Unsupervised Learning- Clustering*. **Great Learning.** https://olympus.mygreatlearning.com/courses/124969/modules/items/6397206?pb_id=18483

Great Learning

Happy Learning !