Unsupervised Learning: Trade & Ahead

Project Overview

This project applies unsupervised learning techniques to uncover natural groupings among NYSE-listed companies based on key financial indicators. By leveraging clustering algorithms and dimensionality reduction, the analysis aims to identify distinct stock group profiles that can guide portfolio diversification strategies and enhance investment decision-making. The results are intended to support Trade & Ahead's clients in building balanced portfolios aligned with both performance potential and risk tolerance.

Problem Statement

Context

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

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

Objective

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

Data Dictionary

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

Project Initialization

Import Libraries

```
# Installing the libraries with the specified version.
!pip install scikit-learn>=1.2.2 seaborn>=0.13.1 matplotlib>=3.7.1 numpy>=1.25.2 pandas>=1.5.3 yellowbrick>=1.5 -q --user
```

Note: *After running the above cell, kindly restart the notebook kernel and run all cells sequentially from the start again.*

```
# Core libraries
import numpy as np
import pandas as pd

# Visualization
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

# Clustering algorithms
from sklearn.cluster import KMeans, AgglomerativeClustering

# Distance and dendrogram for hierarchical clustering
from scipy.cluster.hierarchy import dendrogram, linkage, cophenet
from scipy.spatial.distance import pdist
from scipy.spatial.distance import cdist

# Evaluation and visualization
from yellowbrick.cluster import KElbowVisualizer, SilhouetteVisualizer
from sklearn.metrics import silhouette_score
```

```
# Preprocessing and PCA
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

# Data display settings
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", 200)

# Suppress warnings
import warnings
warnings.filterwarnings("ignore")
```

```
# Mount drive
from google.colab import drive
drive.mount('/content/drive')
```

```
# Custom Color Set - ElleSet
ElleSet = [
    (102/255, 194/255, 165/255),  # Muted Green
    (214/255,  95/255,  95/255),  # Muted Red
    (141/255, 160/255, 203/255),  # Soft Blue
    (130/255, 198/255, 226/255),  # Muted Blue
    (166/255, 216/255,  84/255),  # Lime Green
    (230/255, 196/255, 148/255),  # Beige
    (179/255, 179/255, 179/255),  # Neutral Gray
    (255/255, 217/255,  47/255),  # Yellow
    (204/255, 153/255, 255/255),  # Soft Lavender
    (255/255, 153/255, 204/255)   # Blush Pink
]

# Function to apply ElleSet as the default Seaborn palette
def use_ElleSet():
    import seaborn as sns
    sns.set_palette(ElleSet)
```

```
# Clean Plot Outputs
def show_clean_plot(plot_func, *args, **kwargs):

    plot_func(*args, **kwargs)
    return None
```

Read and Copy Dataset

```
# Read drive
df = pd.read_csv('/content/drive/MyDrive/Unsupervised Learning/stock_data.csv')
```

```
# Copy of dataset
data = df.copy()
```

Data Overview

```
data.head()
```

	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	P/B Ratio
0	AAL	American Airlines Group	Industrials	Airlines	42.349998	9.999995	1.687151	135	51	-604000000	7610000000	11.39	6.681299e+08	3.718174	-8.784219
1	ABBV	AbbVie	Health Care	Pharmaceuticals	59.240002	8.339433	2.197887	130	77	51000000	5144000000	3.15	1.633016e+09	18.806350	-8.750068
2	ABT	Abbott Laboratories	Health Care	Health Care Equipment	44.910000	11.301121	1.273646	21	67	938000000	4423000000	2.94	1.504422e+09	15.275510	-0.394171
3	ADBE	Adobe Systems Inc	Information Technology	Application Software	93.940002	13.977195	1.357679	9	180	-240840000	629551000	1.26	4.996437e+08	74.555557	4.199651
4	ADI	Analog Devices, Inc.	Information Technology	Semiconductors	55.320000	-1.827858	1.701169	14	272	315120000	696878000	0.31	2.247994e+09	178.451613	1.059810

```
data.tail()
```

	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	P/B Ratio
335	YHOO	Yahoo Inc.	Information Technology	Internet Software & Services	33.259998	14.887727	1.845149	15	459	-1032187000	-4359082000	-4.64	939457327.6	28.976191	6.261775
336	YUM	Yum! Brands Inc	Consumer Discretionary	Restaurants	52.516175	-8.698917	1.478877	142	27	159000000	1293000000	2.97	435353535.4	17.682214	-3.838260
337	ZBH	Zimmer Biomet Holdings	Health Care	Health Care Equipment	102.589996	9.347683	1.404206	1	100	376000000	147000000	0.78	188461538.5	131.525636	-23.884449
338	ZION	Zions Bancorp	Financials	Regional Banks	27.299999	-1.158588	1.468176	4	99	-43623000	309471000	1.20	257892500.0	22.749999	-0.063096
339	ZTS	Zoetis	Health Care	Pharmaceuticals	47.919998	16.678836	1.610285	32	65	272000000	339000000	0.68	498529411.8	70.470585	1.723068

- The dataset consists of 15 columns capturing a blend of company identifiers, sector classifications, and financial performance metrics. Features span multiple data types: categorical fields (e.g., Ticker Symbol, Sector, Sub Industry), numeric variables (e.g., Current Price, ROE, Net Income, PE Ratio, PB Ratio), and financial ratios representing market valuation and profitability.
- the dataset does not contain a predefined target variable. Instead, clustering algorithms will be used to explore patterns and groupings across key financial metrics.

```
data.sample(n=10, random_state=1)
```

	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	P/B 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	1.785616
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	5.884467
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	-1.269332
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	-2.256747
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	3.345102
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	-3.838260
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	9.567952
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	17.345857
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	-0.375934
173	IVZ	Invesco Ltd.	Financials	Asset Management & Custody Banks	33.480000	7.067477	1.580839	12	67	412000000	968100000	2.26	4.283628e+08	14.814159	4.218620

- A random sample of 10 companies confirms a diverse cross-section of sectors, including Energy, Technology, Consumer Staples, Financials, and Real Estate — supporting the dataset's relevance for generalized portfolio analysis.
- Financial ranges vary widely, with examples of:
  - Negative Net Income and Earnings Per Share values (e.g., -35.55 EPS, -14B Net Income),
  - Exceptionally high Cash Flow or Estimated Shares Outstanding,
  - Wide swings in Price Change and Volatility across entities.
- No visible format inconsistencies (e.g., placeholders, non-numeric entries) appear in the sample, suggesting sound data structure. However, the presence of extreme values indicates a need for formal outlier checks in later steps.

```
data.shape
```

```
(340, 15)
```

- The dataset contains 340 rows and 15 columns, with each row representing a publicly traded NYSE company and each column capturing a financial attribute or classification label relevant to company performance and valuation.

```
data.info()
```

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

- The dataset contains 340 complete records across 15 columns, with no missing values detected in any field.
- Data types are appropriate for their respective fields:
  - Float64 for continuous financial metrics (e.g., Price, Volatility, P/E Ratio)
  - Int64 for count-based or discrete values (e.g., ROE, Cash Ratio)
  - Object for categorical identifiers and classifications (e.g., Ticker Symbol, Sector)
- The presence of clean, non-null data across all variables supports efficient modeling without the need for imputation or data repair at this stage.

```
df.duplicated().sum()
```

```
np.int64(0)
```

- There are 0 duplicate values in the datasest.

```
df.isnull().sum()
```

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

dtype: int64

- Confirmed, there are 0 missing values in the dataset.

```
df.describe(include='all').T
```

	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

- Wide price distribution: Current Price ranges from 4.50 to 1,274.95 dollars, indicating a mix of low-priced and premium stocks, consistent with a diversified NYSE sample.
- Significant volatility spread: Price Change spans from -47% to +55%, and Volatility ranges from 0.73 to 3.26, reflecting meaningful variation in market performance and risk profiles.
- Presence of extreme financial values: Net Income and Net Cash Flow both show very large negative and positive values (Net Income from -235B to +244B), suggesting the presence of outliers that may impact clustering.
- Valuation ratios show large dispersion: P/E Ratio ranges from 2.93 to 528.09, and P/B Ratio from -76.12 to 129.06, indicating that some firms may be over- or under-valued relative to industry norms — a useful signal for cluster differentiation.

Exploratory Data Analysis (EDA)

Univariate Analysis

```
# Numerical Plot Function
def plot_numerical(data, col):
    fig, ax = plt.subplots(2, 1, figsize=(8, 4), height_ratios=[1, 3], sharex=True)

    # Boxplot
    sns.boxplot(x=data[col], ax=ax[0], color=ElleSet[0])
    ax[0].set(xlabel='', ylabel='', title=f'Distribution of {col.replace("_", " ").title()}')
```

```
    ax[0].axvline(data[col].mean(), color='gray', linestyle='--', label='Mean')
    ax[0].axvline(data[col].median(), color='black', linestyle='-', label='Median')
    ax[0].legend()

    # Histogram + KDE
    sns.histplot(data[col], kde=True, ax=ax[1], color=ElleSet[0])
    ax[1].axvline(data[col].mean(), color='gray', linestyle='--')
    ax[1].axvline(data[col].median(), color='black', linestyle='-')
    ax[1].set(xlabel=col.replace('_', ' ').title(), ylabel='Frequency')

    plt.tight_layout()
    plt.show()
```

```
# Categorical Plot Function - Vertical
def plot_categorical(data, feature, palette=ElleSet, labels=None, rotate_xticks=True, top_n=None):

    order = data[feature].value_counts().sort_values(ascending=False).index

    if top_n:
        order = order[:top_n]
        plot_data = data[data[feature].isin(order)]
    else:
        plot_data = data

    plt.figure(figsize=(10, 8))
    ax = sns.countplot(x=feature, data=plot_data,
                  order=order,
                  palette=palette[:len(order)])

    plt.title(feature.replace('_', ' ').title())
    plt.xlabel(feature.replace('_', ' ').title())
    plt.ylabel('Count')

    if labels:
        ax.set_xticklabels(labels)
    elif rotate_xticks:
        plt.xticks(rotation=45, ha='right')

    for p in ax.patches:
        height = p.get_height()
        ax.text(p.get_x() + p.get_width() / 2,
                height + 0.5,
                int(height),
                ha='center')

    plt.tight_layout()
    plt.show()
```

```
# Categorical Plot Function - Horizontal

import textwrap

def plot_horizontal_categorical(data, feature, top_n=15, palette=ElleSet):

    value_counts = data[feature].value_counts().sort_values(ascending=False)
    top_categories = value_counts.head(top_n).index
    plot_data = data[data[feature].isin(top_categories)]

    wrapped_labels = [textwrap.fill(label, 25) for label in top_categories]

    plt.figure(figsize=(10, 8))
    ax = sns.countplot(y=feature, data=plot_data,
                       order=top_categories,
                       palette=palette[:len(top_categories)])

    ax.set_yticklabels(wrapped_labels)

    for p in ax.patches:
      ax.text(p.get_width() + 0.3,
            p.get_y() + p.get_height() / 2,
            int(p.get_width()),
            va='center')

    plt.title(feature.replace('_', ' ').title(), fontsize=14, fontweight='bold')
    plt.xlabel('Count')
    plt.ylabel(feature.replace('_', ' ').title())
    plt.tight_layout()
    plt.show()
```

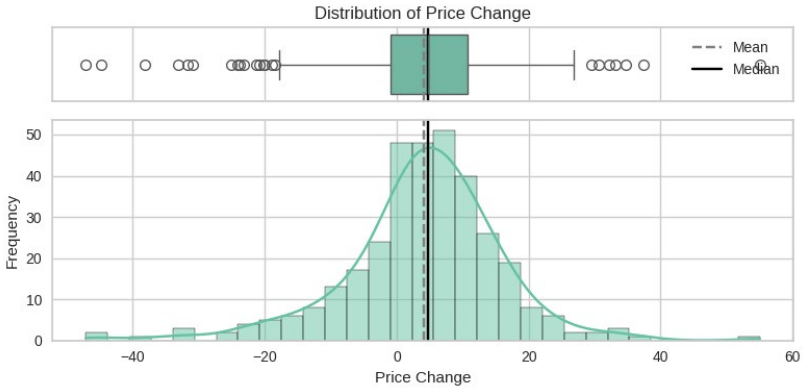
Current Price

```
plot_numerical(data, 'Current Price')
```

- Stock prices are heavily right-skewed, with most companies trading below $150, while a few outliers exceed 1,000 dollars. This skew may influence clustering and suggests the need for scaling or transformation.
- The median stock price (60 dollars) is substantially lower than the mean (80 dollars), reinforcing the presence of high-value outliers that distort central tendency.
- A large cluster of firms trade below $100, indicating a dense low- to mid-cap segment — potentially one cluster group in downstream analysis.
- Several extreme outliers (e.g., >$500) may reflect premium-priced firms with unique characteristics or market positioning. Their distance from the main distribution could drive distinct cluster formation.
- Price alone may not represent value, but its dispersion highlights the importance of combining it with valuation and profitability metrics for more meaningful segmentation.

Price Change

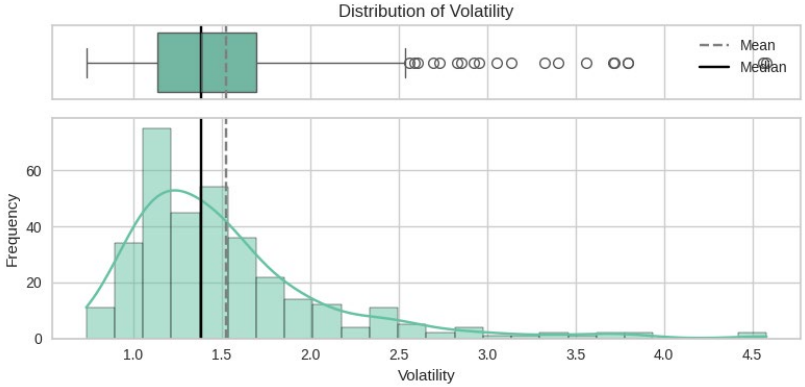
```
plot_numerical(data, 'Price Change')
```



- Price changes follow a near-normal distribution centered just above zero, indicating a roughly balanced mix of short-term gainers and decliners across the dataset.
- The median price change is slightly positive (~2%), suggesting that more stocks experienced upward movement than decline, although the symmetry of the distribution remains intact.
- Several moderate and extreme outliers exist on both ends, with the largest swings reaching beyond ±40%, possibly driven by volatility events or earnings surprises. These could help differentiate momentum-driven clusters.
- The consistent density near the center (±10%) may point to a stable core of low-volatility performers — potentially forming the backbone of low-risk investment clusters.
- As a directional metric, price change offers limited standalone insight, but its combination with volatility or sector behavior may surface distinct trend-following vs. mean-reverting groups.

Volatility

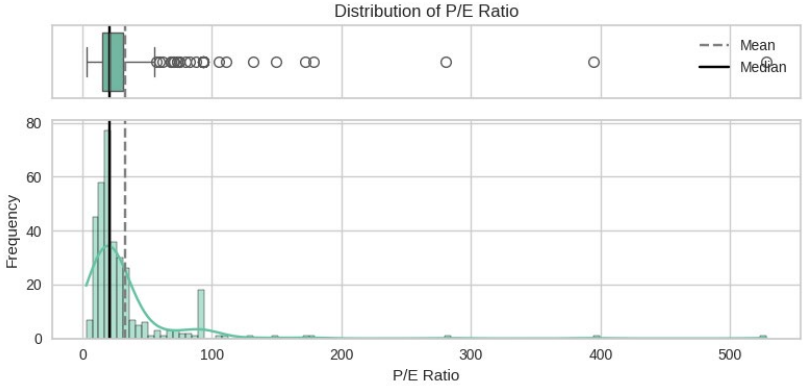
```
plot_numerical(data, 'Volatility')
```



- The distribution is right-skewed, with most firms exhibiting volatility between 1.0 and 1.8. This suggests that the majority of companies maintain relatively stable price behavior over the measurement period.
- The presence of multiple high-end outliers (up to ~4.5) indicates a subset of highly volatile stocks, potentially associated with speculative or early-stage firms — these may form distinct high-risk clusters.
- The mean volatility slightly exceeds the median, reinforcing the impact of the upper tail and suggesting that average volatility is inflated by a few extreme movers.
- Tight grouping around the 1.2–1.6 range implies a dense center of mid-risk companies — likely to emerge as a dominant segment in clustering outcomes.
- As a market behavior signal, volatility may help separate conservative, steady performers from firms prone to larger price swings, supporting investor segmentation by risk tolerance.

P/E Ratio

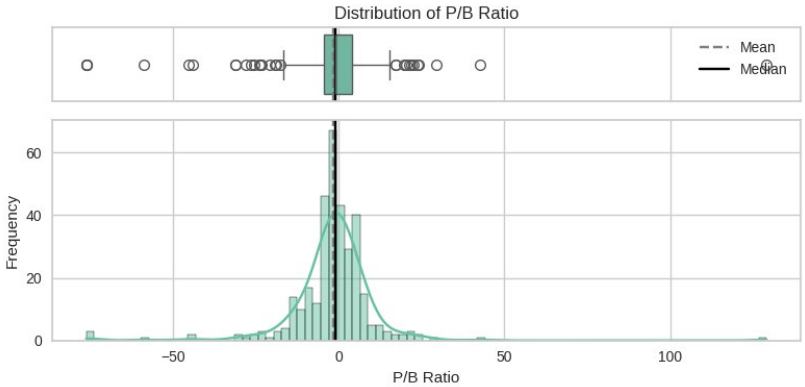
```
plot_numerical(data, 'P/E Ratio')
```



- P/E ratios are highly right-skewed, with most firms concentrated under 50, while a handful exceed 100 — and a few reach extreme outlier territory beyond 400. This sharp spread suggests substantial variability in investor valuation expectations.
- The median P/E is notably lower than the mean, highlighting the distortion caused by a small number of exceptionally high-valued firms. These outliers may reflect speculative growth stocks or firms with temporarily depressed earnings.
- A dense cluster forms between 10 and 30, indicating a core segment of traditionally valued firms — likely to anchor one or more cluster groups.
- Wide dispersion in valuation multiples points to divergent investor expectations and risk profiles across the dataset, making this feature especially relevant for unsupervised segmentation.
- As a normalized price metric, the P/E ratio offers a forward-looking lens into market sentiment and can help differentiate value stocks from growth-oriented outliers.

P/B Ratio

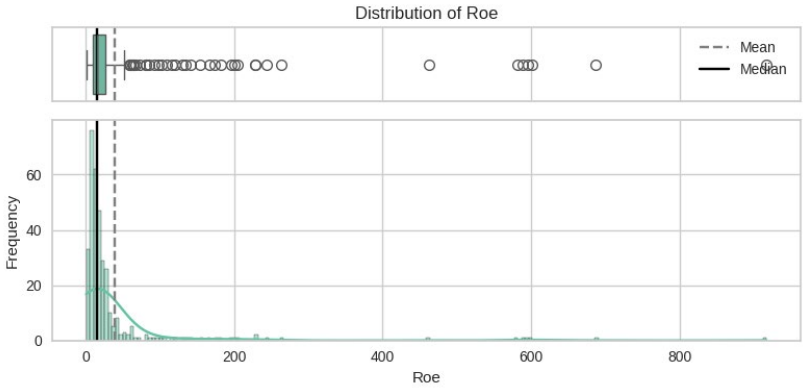
```
plot_numerical(data, 'P/B Ratio')
```



- The distribution is symmetric but unusually wide, spanning from approximately -75 to over 125. This indicates the presence of firms with extreme discrepancies between market and book value — both undervalued and potentially distressed or speculative.
- Most values are tightly centered around zero, with a noticeable peak near 1, suggesting that many firms are trading close to their book value — often a marker of stable, mature businesses.
- The presence of negative P/B values may signal accounting anomalies or firms with negative equity, introducing a unique structural distinction in the dataset.
- Extensive spread across both tails offers strong clustering potential, as companies differ not just by valuation premium, but by balance sheet structure and market perception.

ROE

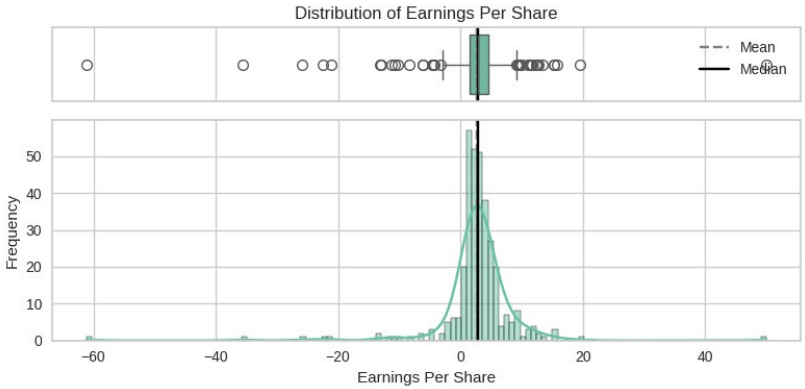
```
plot_numerical(data, 'ROE')
```



- The distribution is extremely right-skewed, with most firms clustered under 40, but outliers stretching beyond 900. These extremes indicate a few firms with highly efficient — or anomalous — returns on equity.
- The median ROE is relatively low (~15), suggesting that while a few companies deliver exceptional equity returns, the majority perform modestly in relation to shareholder investment.
- Outlier density is high, and their magnitude could dominate standard scaling approaches — calling for normalization or transformation prior to clustering.
- As a profitability measure, ROE offers a key lens into operational effectiveness and financial leverage. Its variability may help distinguish capital-efficient firms from weaker performers in clustering outcomes.

Earnings Per Share

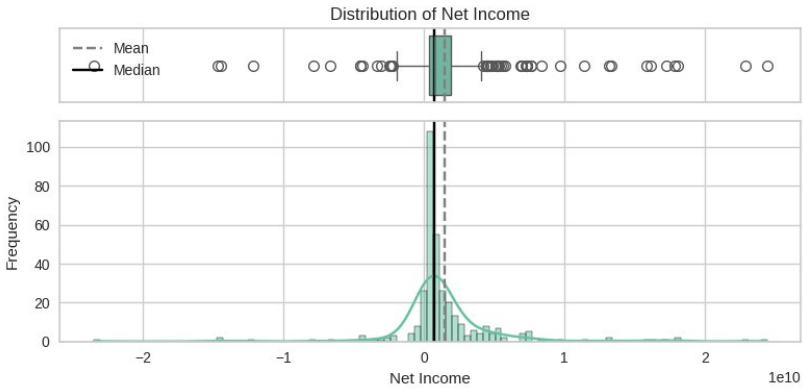
```
plot_numerical(data, 'Earnings Per Share')
```



- EPS values are tightly concentrated near zero, with a steep central peak and dense clustering between -5 and +5. This indicates many firms in the dataset operate with thin per-share profits or modest losses.
- Outliers exist at both extremes, including several firms with large negative or highly positive EPS. These companies may reflect turnaround situations, exceptional profitability, or accounting-driven anomalies.
- The distribution is symmetric but heavy-tailed, suggesting variability in earnings power that may not be fully explained by size or sector alone — a useful driver for distinguishing firm profiles.
- EPS normalizes profitability by share count, offering a scalable comparison across firms of different sizes — ideal for identifying performance tiers in downstream clustering.

Net Income

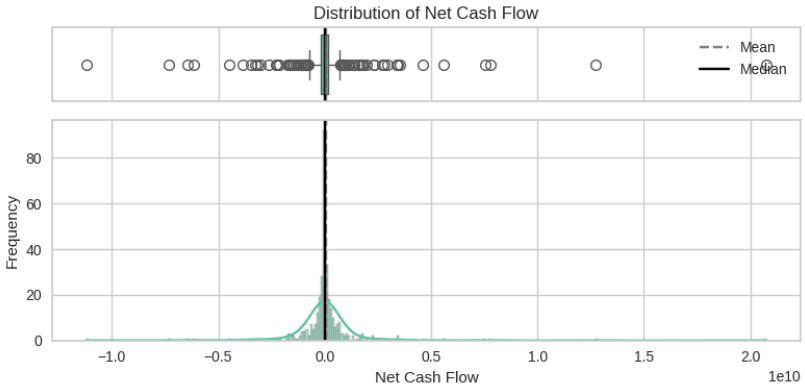
```
plot_numerical(data, 'Net Income')
```



- The distribution is centered near zero but highly dispersed, with net incomes ranging from deep losses (~−2B) to significant profits (>20B). This wide range reflects major variation in scale and business model.
- Most firms cluster between 0 and 2B, suggesting that the bulk of companies are modestly profitable, though a sizable number also report net losses.
- Extreme outliers exist on both sides, with a few firms contributing disproportionately to total earnings or losses — potentially introducing skew in clustering without transformation.
- Unlike EPS or ROE, Net Income is not size-adjusted, making it valuable for identifying scale-based clusters (e.g., mega-cap vs. small-cap firms), especially when paired with Estimated Shares Outstanding.

Net Cash Flow

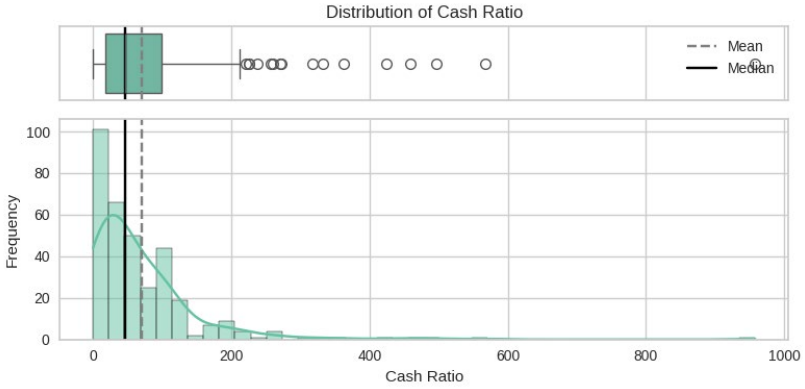
```
plot_numerical(data, 'Net Cash Flow')
```



- Net Cash Flow is sharply centered around zero, with most values falling between $-2B$ and $+2B$. This indicates that many companies operate with narrow cash margins — a critical liquidity consideration.
- Outliers extend in both directions, including firms with exceptionally strong positive inflows or significant cash burn. These edge cases may flag high-growth or distressed business models.
- The distribution is narrower than Net Income, suggesting that even highly profitable firms may not translate earnings into strong operational cash flow — a meaningful distinction for cluster segmentation.
- As a liquidity-focused metric, Net Cash Flow can separate firms with healthy internal financing from those reliant on debt or equity issuance, adding interpretive value when paired with Cash Ratio.

Cash Ratio

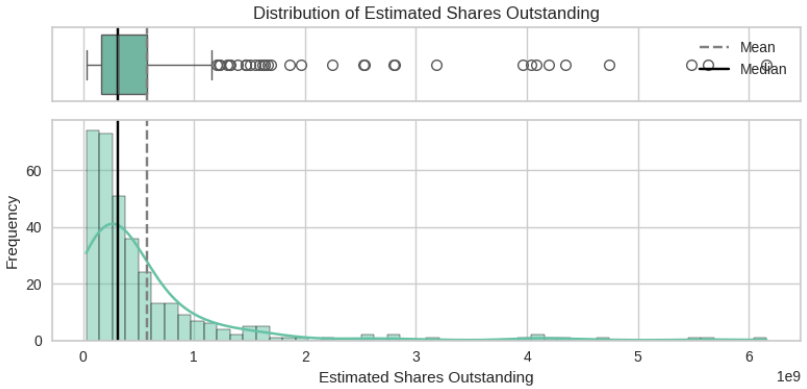
```
plot_numerical(data, 'Cash Ratio')
```



- The distribution is right-skewed, with most firms holding moderate cash buffers (under 100), but a few reporting extreme ratios over 900 — potentially signaling large cash reserves or low current liabilities.
- A sharp concentration below 50 suggests that many firms maintain minimal liquidity cushions, relying on receivables or short-term financing to cover obligations.
- Outliers are common and substantial, which could distort clustering unless the feature is scaled or log-transformed. These firms may include cash-rich tech giants or asset-light disruptors.
- Cash Ratio serves as a direct solvency signal, especially under stress conditions, and may be useful in distinguishing conservative vs. aggressive capital strategies across clusters.

Estimated Shares Outstanding

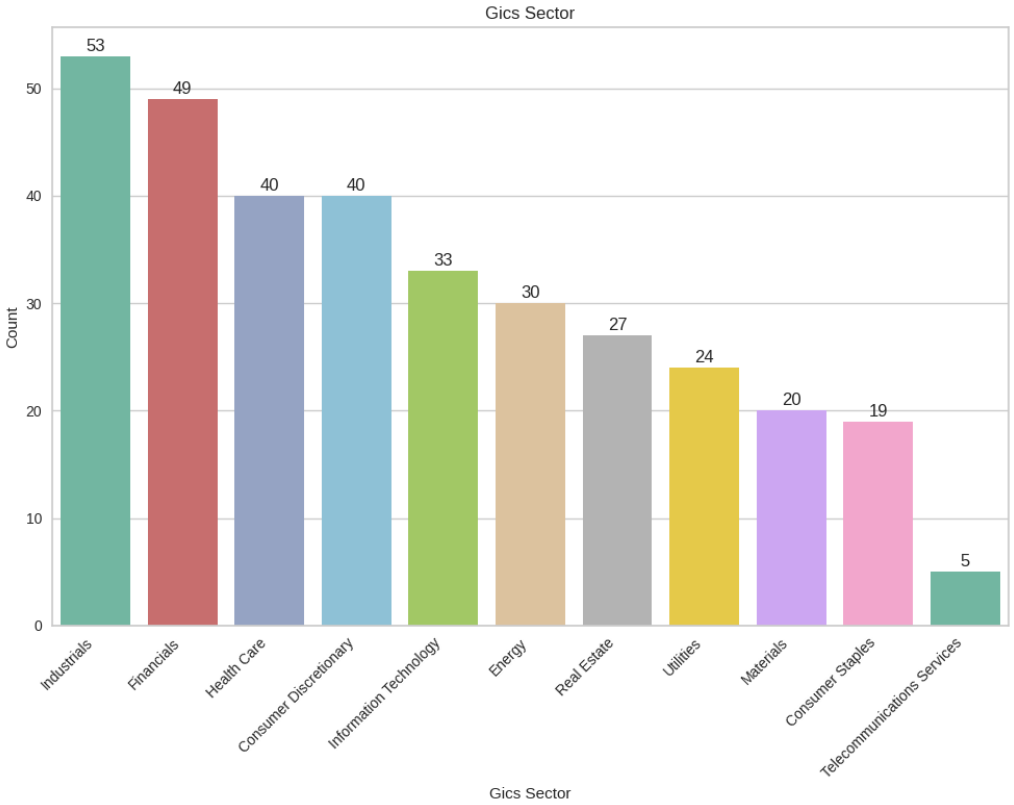
```
plot_numerical(data, 'Estimated Shares Outstanding')
```



- The distribution is heavily right-skewed, with most companies issuing fewer than 1 billion shares, but a long tail of high-volume issuers extending above 6 billion — likely reflecting mega-cap firms.
- Median share counts fall well below the mean, highlighting the presence of a few dominant issuers that may disproportionately influence market metrics unless scaled.
- This feature signals company size and public float, making it potentially useful for distinguishing micro-cap, mid-cap, and large-cap entities during clustering — though it may be less predictive than financial ratios.

GICS Sector

```
plot_categorical(data, 'GICS Sector')
```

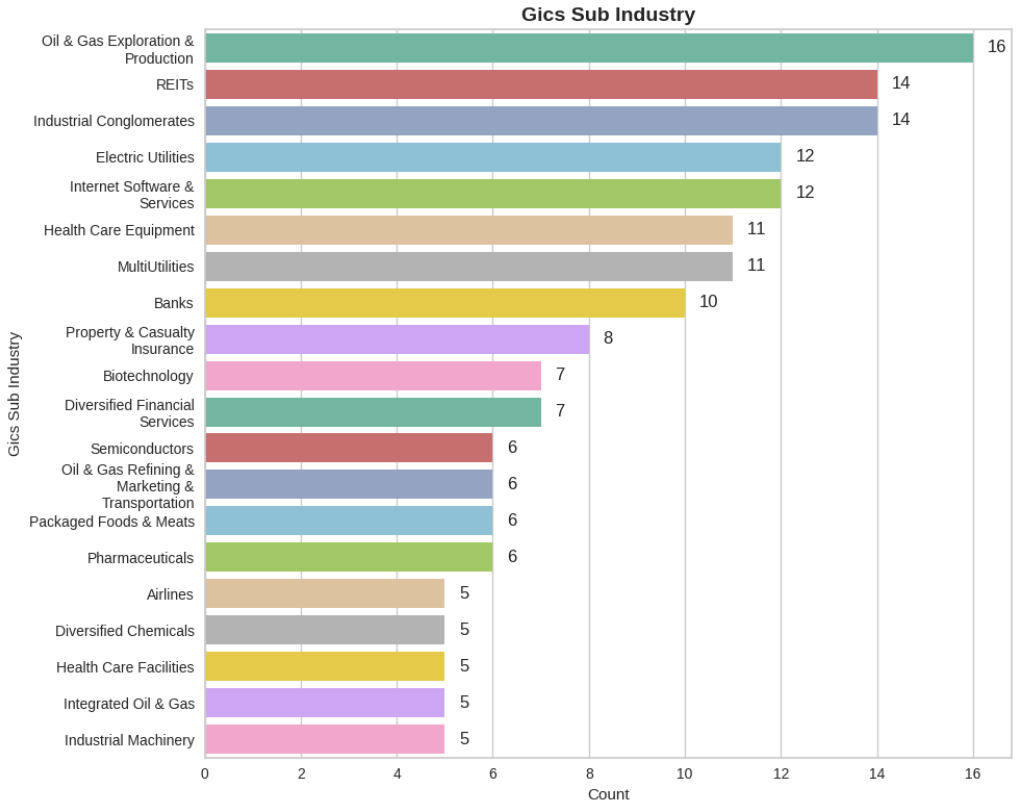


- Industrial and Financial sectors dominate the dataset, comprising 15% and 14% of all entries respectively, while Telecommunications Services is sparsely represented with only 5 companies.
- Mid-level representation is balanced across Health Care, Consumer Discretionary, and Information Technology, suggesting moderate coverage of key modern economy segments.
- Sectors such as Materials, Consumer Staples, and Utilities are on the lower end of the spectrum, indicating less exposure to traditionally defensive or resource-heavy industries.

- This distribution implies that downstream segmentation may be more strongly influenced by dynamics in industrials and financials, while sectors with limited representation may carry less statistical weight during clustering.

GICS Sub Industry

```
plot_horizontal_categorical(data, 'GICS Sub Industry', top_n=20)
```



- The dataset includes a diverse range of sub-industries, with the most common being Oil & Gas Exploration & Production, followed by REITs and Industrial Conglomerates.
- Most of the top sub-industries are energy-intensive, finance-driven, or capital-heavy sectors, which may contribute to distinctive financial profiles in cluster analysis.
- There is gradual tapering in representation beyond the top few sub-industries — the 20 most common sub-industries still represent relatively small sample sizes, with the lowest in this group having just 5 records.
- This fragmentation suggests that while industry classification could influence groupings, clustering will likely depend more heavily on financial and valuation features rather than categorical sector affiliation alone

Univariate Analysis: Summary

The univariate analysis explored the distribution of core financial, valuation, and structural features to understand variation across the stock universe and identify signals relevant for clustering.

- Stock price, volatility, and valuation ratios (P/E, P/B) are all right-skewed with long tails, indicating the presence of extreme values—likely driven by a mix of speculative growth firms and distressed outliers.
- Profitability and cash flow metrics show wide dispersion and heavy skew, with distributions centered near zero but spanning large losses to substantial gains—revealing inconsistencies in operational efficiency and financial health.
- Categorical features like sector and sub-industry are diverse but imbalanced, with a few dominant groups suggesting that pure industry labels may not be sufficient to explain behavioral clusters.

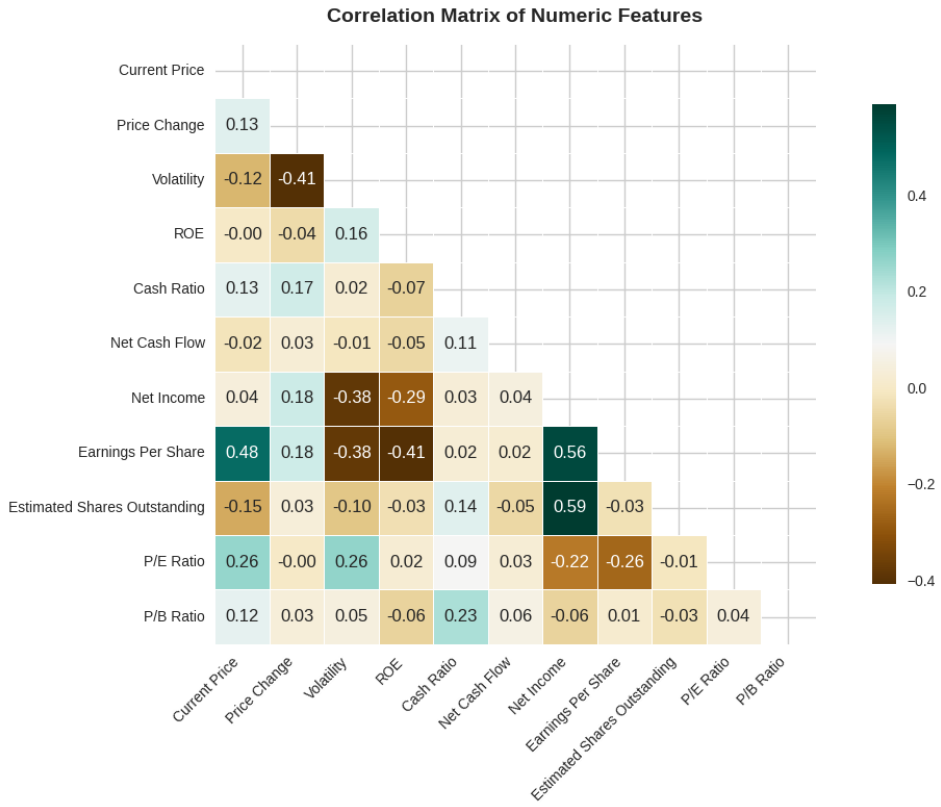
Overall, the data reveals significant heterogeneity across firms—supporting the need for unsupervised techniques that can detect natural groupings beyond predefined categories.

Bivariate Analysis

Correlation Check

```
# Correlation Matrix
plt.figure(figsize=(12, 8))

# Compute correlation matrix
corr_matrix = data.corr(numeric_only=True)
```

```
# Hide the upper triangle
mask = np.triu(np.ones_like(corr_matrix, dtype=bool))

# Plot heatmap
sns.heatmap(
    corr_matrix,
    mask=mask,
    cmap='BrBG',
    annot=True,
    fmt=".2f",
    linewidths=0.5,
    square=True,
    cbar_kws={"shrink": 0.8}
)

plt.title("Correlation Matrix of Numeric Features", fontsize=14, weight='bold', pad=15)
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.tight_layout()
plt.show()
```



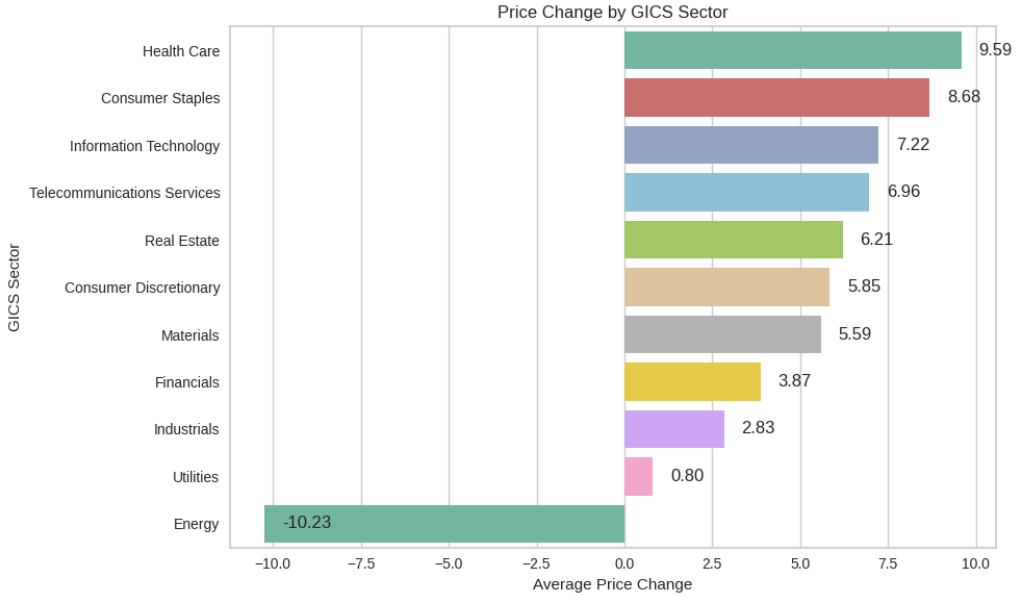
- Earnings Per Share (EPS) has the strongest positive relationship with Current Price (r = 0.48), supporting its role as a key valuation input.
- Net Income correlates strongly with both Estimated Shares Outstanding (r = 0.59) and EPS (r = 0.56), indicating structural interdependencies across size and profitability metrics.
- Volatility shows moderate negative correlations with Net Income (r = -0.38) and EPS (r = -0.38), implying that higher-risk stocks tend to be less profitable.
- P/E Ratio and Volatility show a moderate positive relationship (r = 0.26), suggesting investor sentiment may drive valuation premiums for riskier firms.
- Minimal correlations among Cash Ratio, P/B Ratio, and ROE with other features suggest these may contribute non-redundant signals to clustering and are less likely to blur cluster boundaries.
- Despite a wide Price Change range, its low correlations with financial fundamentals (e.g., EPS r = 0.18) suggest recent price movements are weak predictors of company structure or performance.

```
# Barplot function
def plot_grouped_bar(data, category, metric, palette=ElleSet, rotate_labels=False, title=None):

    grouped = data.groupby(category)[metric].mean().sort_values(ascending=False)

    plt.figure(figsize=(10, 6))
    ax = sns.barplot(x=grouped.values, y=grouped.index, palette=palette[:len(grouped)])
```

```
    # Add value labels
    for i, v in enumerate(grouped.values):
        ax.text(v + 0.5, i, f"{v:.2f}", va='center')

    # Titles and labels
    plt.xlabel(f'Average {metric}')
    plt.ylabel(category)
    plt.title(title or f'Average {metric} by {category}')
    plt.tight_layout()
    plt.show()
```

Ecomonic Sector vs. Price Change - growth trends across sectors

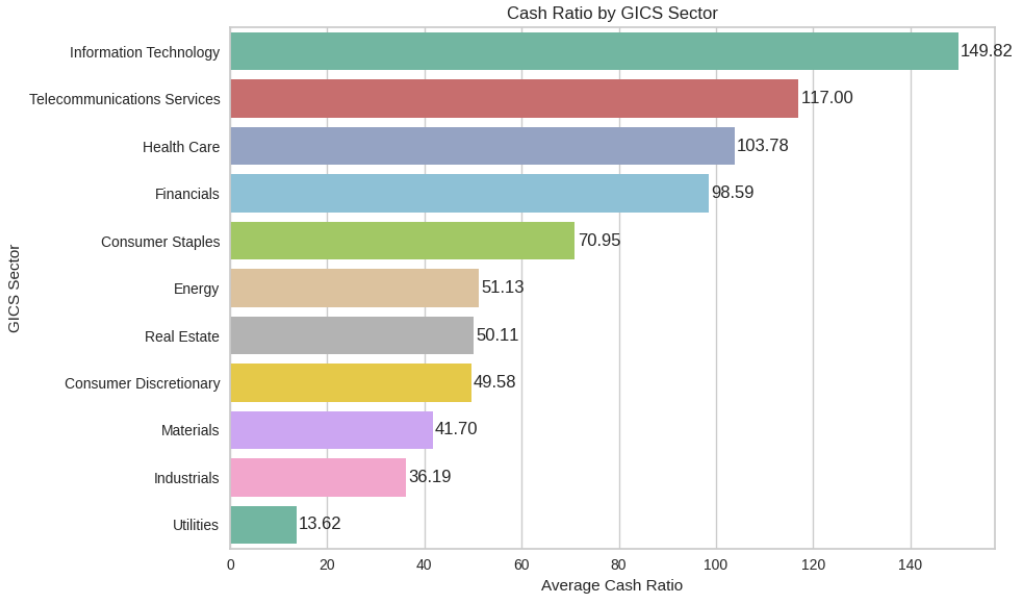
```
# Ecomonic Sector vs. Price Change
plot_grouped_bar(
    data=data,
    category='GICS Sector',
    metric='Price Change',
    title='Price Change by GICS Sector'
)
```



- Health Care and Consumer Staples lead all sectors with the highest average price growth, signaling strong investor confidence in traditionally defensive sectors.
- Information Technology and Telecommunications Services also show robust average gains, reflecting their ongoing relevance and market resilience in a tech-driven economy.
- Energy is the only sector with a notably negative average price change, suggesting underperformance possibly due to cyclical or regulatory headwinds at the time of data capture.
- Utilities and Industrials show the lowest positive growth rates among the remaining sectors, potentially reflecting market caution in rate-sensitive or cyclical industries.

Economic Sector vs. Cash Ratio - short-term solvency by sector

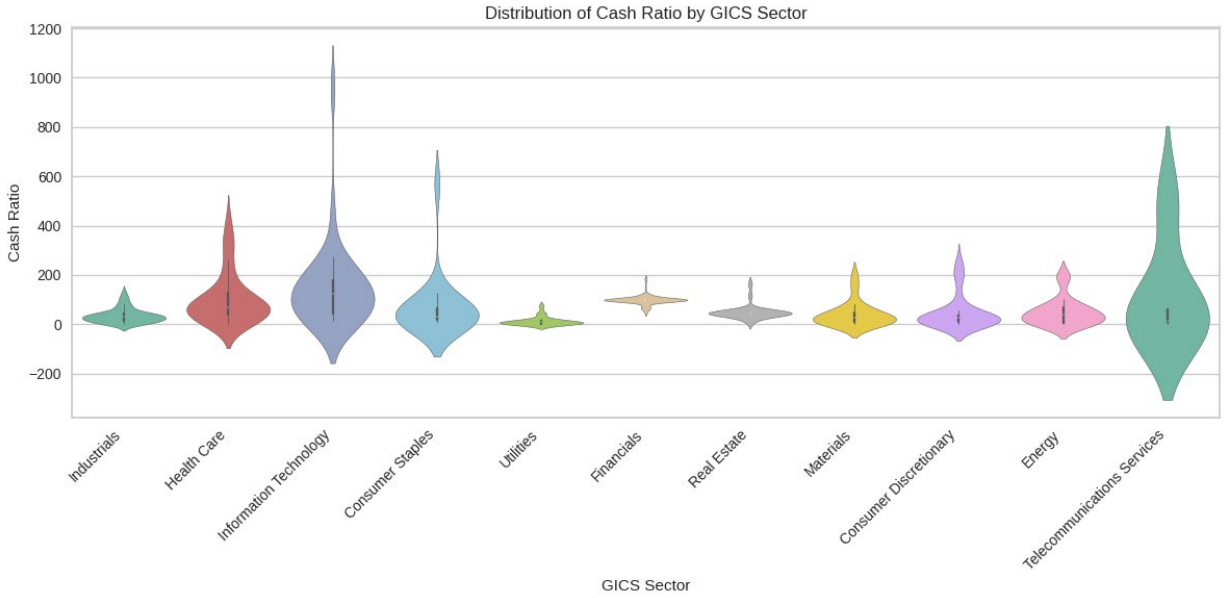
```
# Economic Sector vs. Cash Ratio
plot_grouped_bar(
    data=data,
    category='GICS Sector',
    metric='Cash Ratio',
    title='Cash Ratio by GICS Sector'
)
```



- Information Technology leads all sectors with an average cash ratio nearing 150, suggesting these firms maintain robust cash buffers—likely reflecting high profitability and prudent capital management in a sector known for agility and low physical asset dependence.
- Telecommunications Services and Health Care follow closely behind, each maintaining cash ratios well over 100. This indicates strong short-term solvency, which may support ongoing innovation and resilience in regulatory or infrastructure-heavy environments.
- Financials and Consumer Staples occupy a moderate range, with cash ratios around 70–100. These sectors appear to balance liquidity with capital allocation efficiency—adequate for routine operations without excess idle cash.
- Utilities and Industrials show the lowest liquidity levels, with average cash ratios of 13.6 and 36.2, respectively. These figures reflect structural capital constraints typical of infrastructure-heavy sectors, where cash may be tied up in long-term assets or reinvested into operations.

Implication: Short-term solvency varies widely by sector. This pattern may play a meaningful role in unsupervised clustering by differentiating cash-rich, innovation-driven sectors from asset-intensive, capital-constrained industries.

```
# Economic Sector vs. Cash Ratio - Violin Plot
plt.figure(figsize=(12, 6))
sns.violinplot(data=data, x='GICS Sector', y='Cash Ratio', palette=ElleSet[:data['GICS Sector'].nunique()])
plt.title('Distribution of Cash Ratio by GICS Sector')
plt.xlabel('GICS Sector')
plt.ylabel('Cash Ratio')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

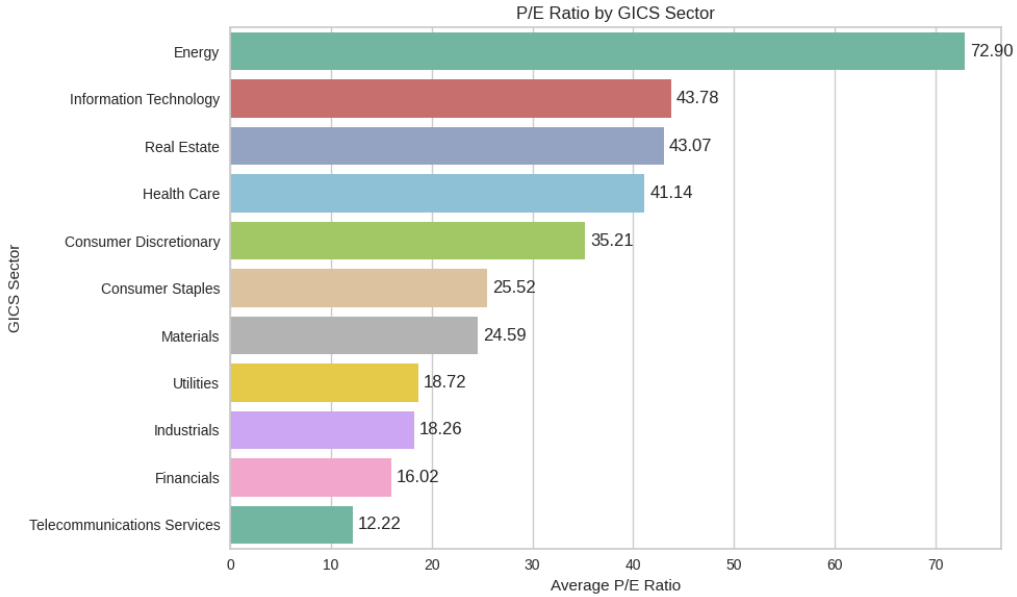


- Information Technology shows the widest and most positively skewed distribution, with a dense cluster of firms maintaining moderate cash ratios (around 100), but a long right tail extending well beyond 800. This suggests a few exceptionally cash-rich firms, likely market leaders or capital-light tech platforms hoarding liquidity for flexibility and strategic agility.
- Telecommunications Services reveals a bimodal distribution with a cluster near zero and another group of firms with much higher liquidity. This likely reflects a split between infrastructure-heavy telecom providers and digital service firms—highlighting intra-sector divergence in liquidity management strategies.
- Consumer Staples and Health Care both exhibit balanced, moderately wide distributions, suggesting that while most firms operate within expected liquidity bands, some carry notably higher reserves—possibly due to conservative financial practices or M&A readiness.

Implications: The violin plot surfaces hidden dispersion that average values mask—especially in Tech and Telecom. These wide, skewed distributions highlight the need for robust feature scaling or transformation before applying unsupervised learning methods, to avoid overweighting outlier-driven sectors in cluster formation.

Economic Sector vs. P/E Ratio - sector-level valuation signals

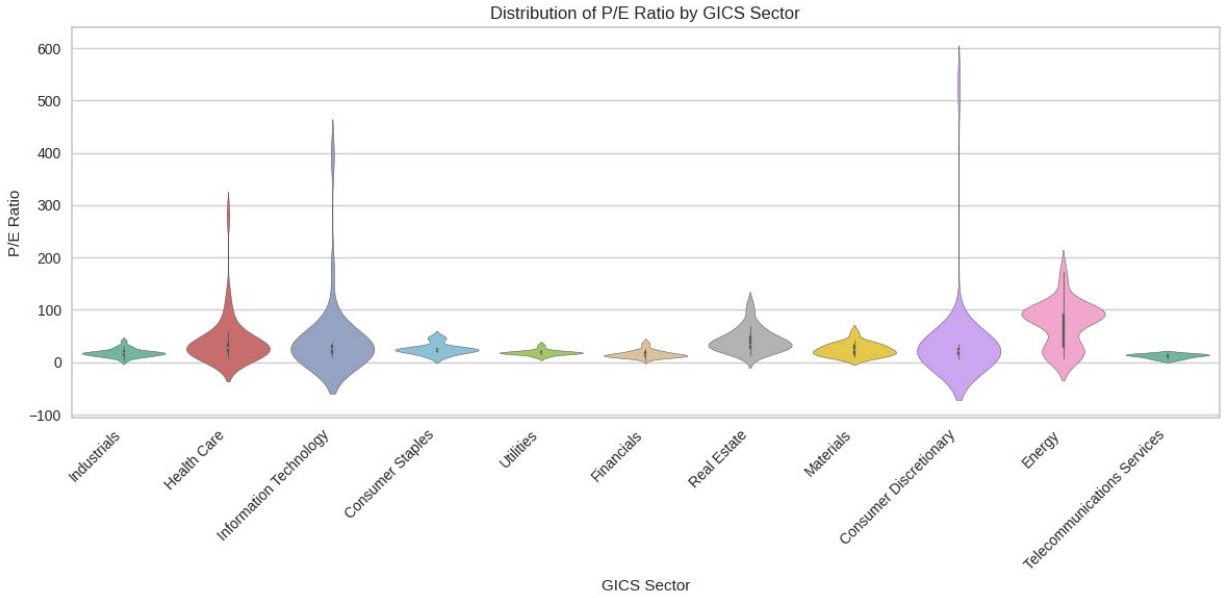
```
# Economic Sector vs. P/E Ration
plot_grouped_bar(
    data=data,
    category='GICS Sector',
    metric='P/E Ratio',
    title='P/E Ratio by GICS Sector'
)
```



- Energy exhibits the highest average P/E ratio (~73), a notable outlier compared to all other sectors. This elevated valuation may reflect investor optimism about future earnings rebound, possibly due to cyclical recovery or transition investments (e.g., renewable energy shifts).
- Information Technology, Real Estate, and Health Care follow with high average P/E ratios (41–44), indicating that investors are pricing in strong growth expectations. These sectors are often driven by innovation, regulation, or stable long-term demand.
- Consumer Discretionary and Consumer Staples sit mid-range, with P/E ratios between 25 and 35—consistent with steady demand but moderated by economic sensitivity or margin pressures.
- Financials, Industrials, and Telecom occupy the lowest P/E levels (12–18), suggesting either undervaluation, cyclical headwinds, or investor skepticism about short-term earnings potential.

Implication: P/E ratio distributions highlight key valuation differences across sectors. These differences may be useful for clustering, helping to segment firms based on market sentiment, perceived growth, or earnings risk.

```
# Economic Sector vs. P/E Ratio - Violin Plot
plt.figure(figsize=(12, 6))
sns.violinplot(data=data, x='GICS Sector', y='P/E Ratio', palette=ElleSet[:data['GICS Sector'].nunique()])
plt.title('Distribution of P/E Ratio by GICS Sector')
plt.xlabel('GICS Sector')
plt.ylabel('P/E Ratio')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

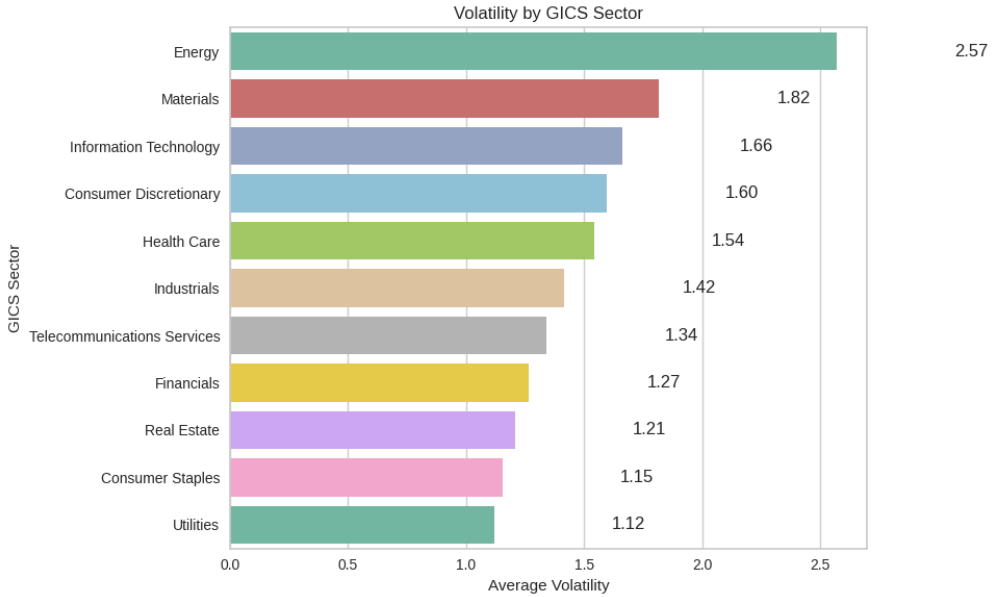


- Energy and Consumer Discretionary sectors exhibit extreme right-skew and heavy tails, with P/E values stretching into the hundreds—revealing that a few firms are dramatically inflating sector averages and may signal speculative pricing or earnings instability.
- Sectors like Industrials, Utilities, and Telecommunications show compact distributions, with low variation and no major outliers—indicating more conservative and stable investor expectations around earnings.
- Information Technology and Health Care show moderately wide spreads, suggesting a mix of high-growth and value-aligned firms within these sectors—valuable for segmentation.

Implication: While sector-level averages provide a general benchmark, the violin plot reveals volatility within valuation expectations. For modeling, this suggests that P/E ratio outliers should be handled carefully—e.g., via scaling, transformation, or robust clustering—to avoid distortion and to capture nuanced investor sentiment within sectors.

Economic Sector vs. Volatility - risk exposure at the sector level

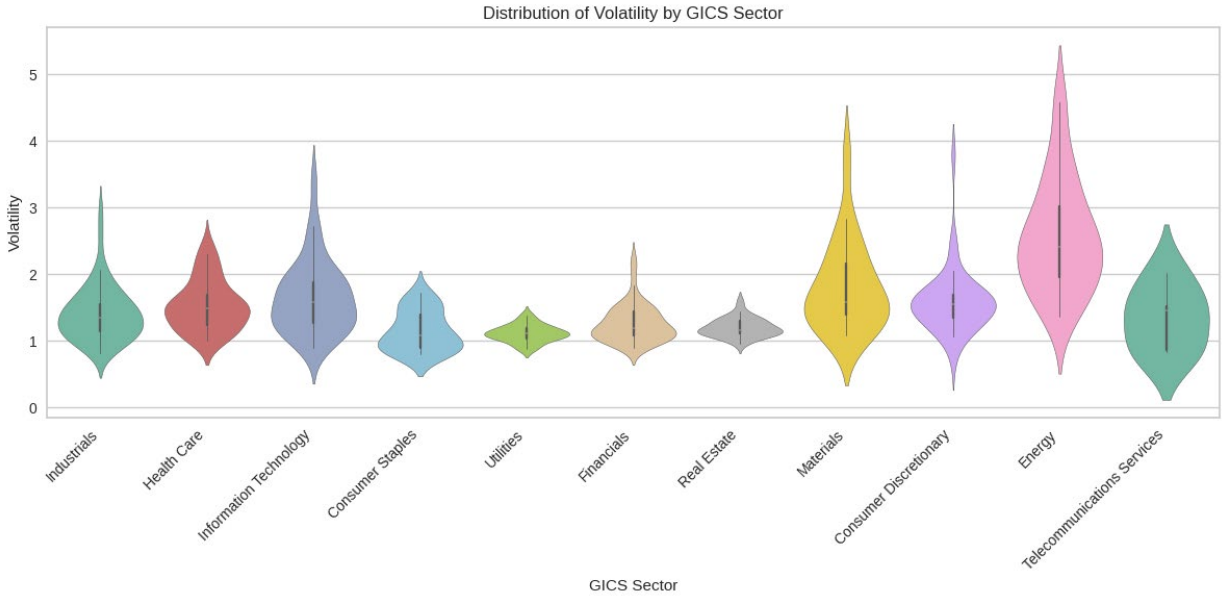
```
# Economic Sector vs. Volatility
plot_grouped_bar(
    data=data,
    category='GICS Sector',
    metric='Volatility',
    title='Volatility by GICS Sector'
)
```



- Energy exhibits the highest average volatility (2.57), more than double that of many other sectors. This confirms its high-risk profile, likely linked to sensitivity around oil prices, regulation, and geopolitical events.
- Materials and Information Technology also show elevated volatility (1.8+), suggesting greater susceptibility to cyclical demand, innovation cycles, or commodity pricing shifts.
- Utilities, Consumer Staples, and Real Estate present the lowest volatility (1.1–1.2), consistent with their historical positioning as defensive, income-generating holdings during uncertain market conditions.
- Volatility trends align with sector characteristics, helping identify which industries are more prone to rapid price fluctuations versus those offering stability.

Implication: Understanding volatility by sector aids investor segmentation and model design. High-volatility sectors may cluster separately or warrant different risk treatment in unsupervised models like PCA or KMeans.

```
# Economic Sector vs. Volatility Distribution - Violin Plot
plt.figure(figsize=(12, 6))
sns.violinplot(data=data, x='GICS Sector', y='Volatility', palette=ElleSet[:data['GICS Sector'].nunique()])
plt.title('Distribution of Volatility by GICS Sector')
plt.xlabel('GICS Sector')
plt.ylabel('Volatility')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

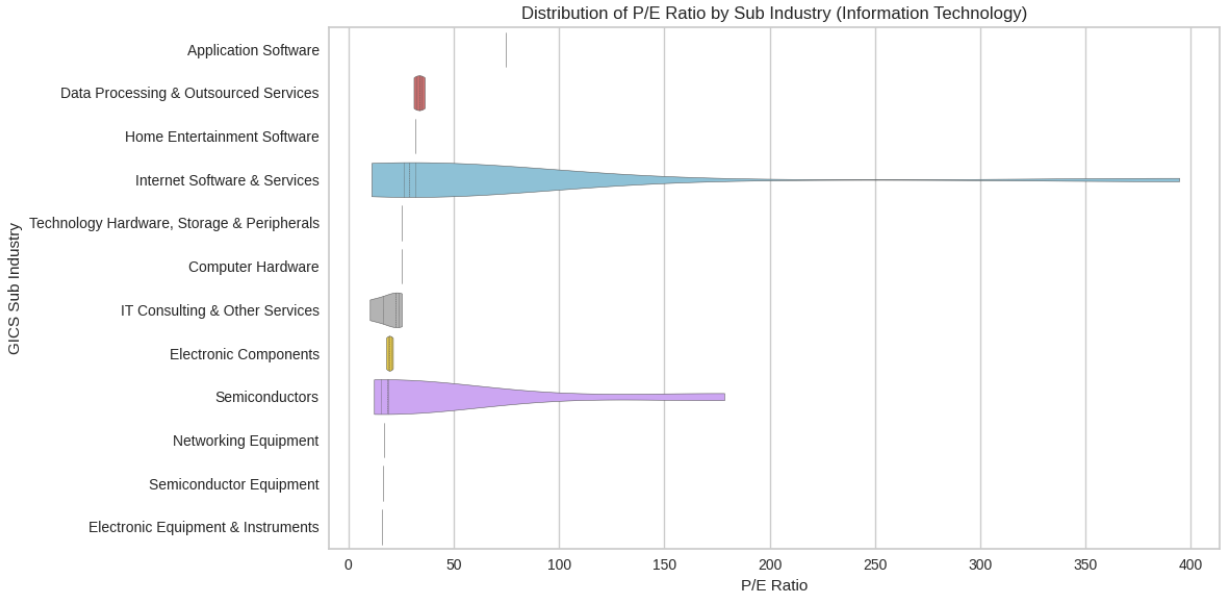


- Energy's volatility isn't just high on average—it's broadly dispersed. The violin shape shows a wide spread with notable upper-end outliers, reinforcing that price instability in this sector is both frequent and extreme.
- Materials and Consumer Discretionary exhibit fat-tailed distributions, indicating a mix of low- and high-volatility firms within the same sector. This suggests intra-sector diversity that may get masked when using mean values alone.
- Utilities and Real Estate show not only low averages but tight distributions, confirming them as consistent low-volatility sectors—an important characteristic for investors targeting capital preservation.

Implication: While average volatility offers a general risk profile, the distributional view reveals how consistently (or erratically) that risk is expressed within each sector. This added nuance supports more precise modeling—e.g., flagging sectors like Energy for custom clustering logic or robust scaling during PCA.

GICS Sub Industry: P/E Ratio by Information Technology

```
# Filter data for Information Technology sector
it_data = data[data['GICS Sector'] == 'Information Technology']

# Sort sub-industries by median P/E for better visual order
ordered_subs = it_data.groupby('GICS Sub Industry')['P/E Ratio'].median().sort_values(ascending=False).index

# Plot
plt.figure(figsize=(12, 6))
sns.violinplot(
    data=it_data,
    x='P/E Ratio',
    y='GICS Sub Industry',
    order=ordered_subs,
    palette=ElleSet[:len(ordered_subs)],
    inner='quartile',
    cut=0
)

plt.title('Distribution of P/E Ratio by Sub Industry (Information Technology)')
plt.xlabel('P/E Ratio')
plt.ylabel('GICS Sub Industry')
plt.tight_layout()
plt.show()
```

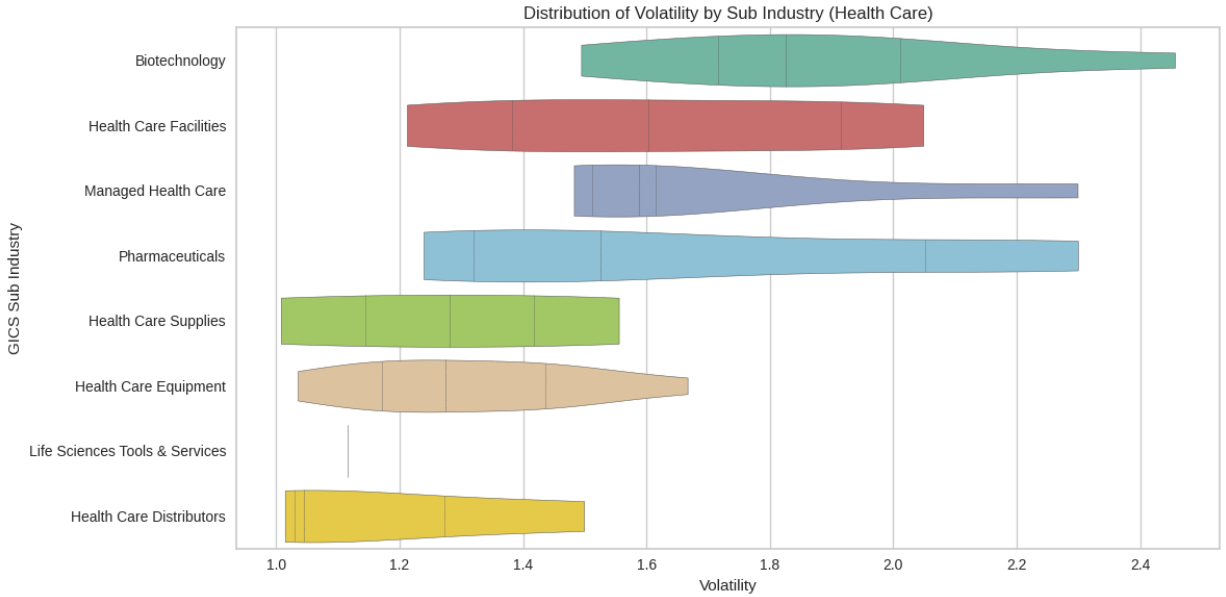


- Internet Software & Services leads in both average and dispersion of P/E ratios. This sub-sector displays an exceptionally wide range with a pronounced right tail, indicating speculative valuations and strong investor expectations for high-growth firms. The presence of extreme outliers suggests both market optimism and underlying earnings instability.
- Semiconductors and Semiconductor Equipment show elevated but tighter distributions. These segments also command high P/E ratios, but with slightly more compact ranges than software-based peers. This reflects consistent earnings performance and strong market positioning, likely tied to increasing demand in AI, IoT, and computing infrastructure.
- Hardware-oriented sub-industries (e.g., Electronic Equipment, Networking Equipment) show low and narrow P/E distributions. These businesses appear more mature and capital-intensive, with investors valuing them more conservatively based on stable but modest growth expectations. They contrast sharply with their software counterparts.
- Consulting and Data Services exhibit mid-range valuations. IT Consulting & Outsourced Services show moderate dispersion, suggesting investor recognition of service-based recurring revenue, but tempered enthusiasm compared to product-centric models.

Implication: The IT sector is far from homogeneous — valuation behaviors cluster strongly by sub-industry. For unsupervised learning, this suggests that sub-sector segmentation could meaningfully enhance clustering quality or help explain dimensional drivers in PCA. Ignoring intra-sector dispersion risks masking key behavioral or risk-based differentiation within technology holdings.

GICS Sub Industry: Volatility by Health Care

```
# Filter data for Health Care sector
it_data = data[data['GICS Sector'] == 'Health Care']

# Sort sub-industries by median P/E for better visual order
ordered_subs = it_data.groupby('GICS Sub Industry')['Volatility'].median().sort_values(ascending=False).index

# Plot
plt.figure(figsize=(12, 6))
sns.violinplot(
    data=it_data,
    x='Volatility',
    y='GICS Sub Industry',
    order=ordered_subs,
    palette=ElleSet[:len(ordered_subs)],
    inner='quartile',
    cut=0
)

plt.title('Distribution of Volatility by Sub Industry (Health Care)')
plt.xlabel('Volatility')
plt.ylabel('GICS Sub Industry')
plt.tight_layout()
plt.show()
```



- Pharmaceuticals exhibits the widest overall dispersion in volatility across firms, with density spread broadly from ~1.2 to ~2.3. This suggests substantial internal variability in market behavior, possibly reflecting differences between major established drug manufacturers and emerging or niche players.
- Biotechnology also shows elevated volatility, with a longer right tail and moderate central density. This aligns with its speculative nature, high R&D risk, and reliance on pipeline breakthroughs or regulatory milestones.
- Health Care Distributors, Supplies, and Equipment show lower volatility and narrower ranges. This supports their reputation as more stable, operationally grounded sub-industries focused on logistics or physical products.

Implication: Volatility patterns within Health Care sub-sectors highlight heterogeneity in market risk, even within a single GICS sector. Understanding this variation is essential when grouping firms for unsupervised modeling — sub-industries like Pharmaceuticals or Biotech may warrant independent treatment or contribute disproportionately to variance in clustering algorithms such as PCA or KMeans.

Bivariate Analysis: Summary

The analysis of sector-level differences across key financial metrics revealed patterns critical to clustering and portfolio segmentation.

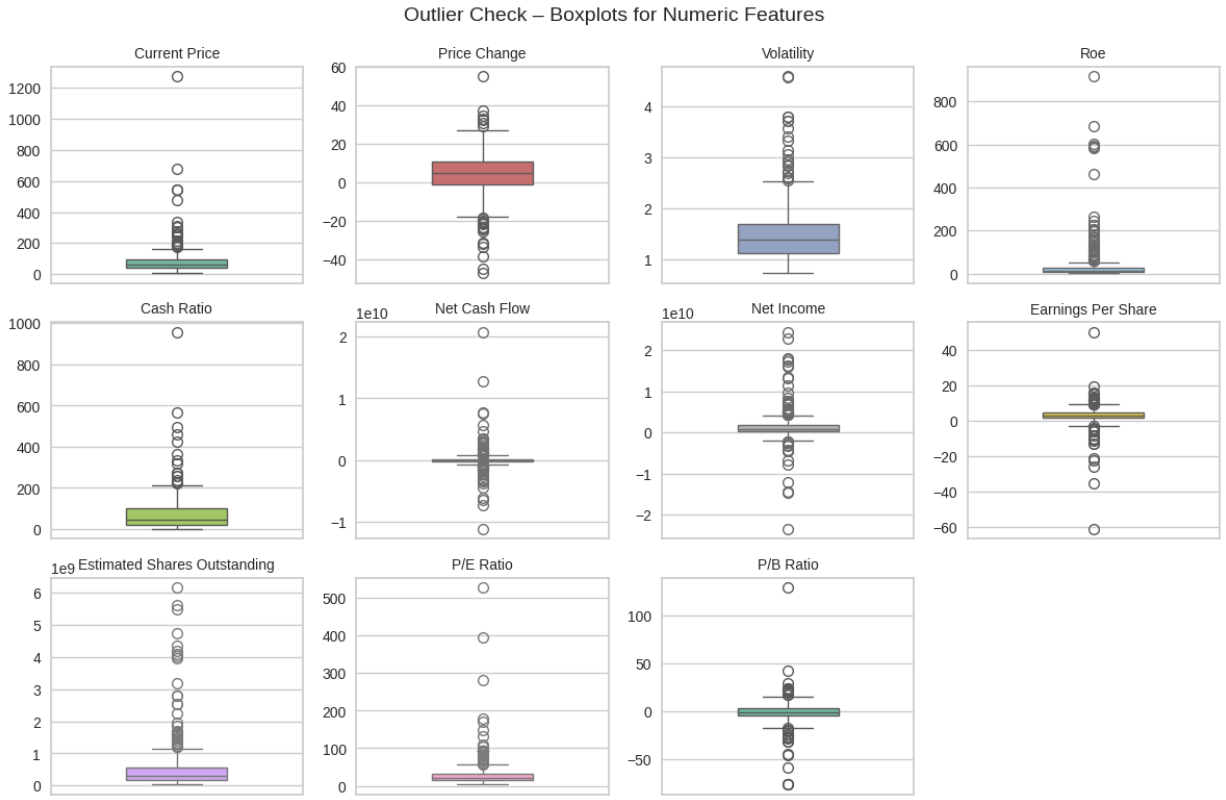
- Volatility clearly distinguished high-risk sectors like Energy and Materials from more stable groups such as Utilities and Consumer Staples.
- Cash Ratio exposed liquidity gaps across sectors, with Information Technology and Health Care maintaining large reserves, while Industrials and Utilities operated on tighter margins.
- P/E Ratio reflected valuation sentiment—high in innovation-driven sectors, lower in cyclical or mature industries. Distribution plots revealed hidden variability within sectors, especially in Health Care and Telecom.

These findings confirm that sector averages alone are not enough—intra-sector dispersion and skew offer powerful signals for unsupervised learning. Clustering models will benefit from these insights by capturing market behaviors that transcend traditional classifications.

Data Preprocessing

Outlier Review

```
# Outlier Check – Boxplot Grid for All Numeric Features
numeric_cols = data.select_dtypes(include='number').columns.tolist()

plt.figure(figsize=(12, 10))

for i, feature in enumerate(numeric_cols):
    plt.subplot(4, 4, i + 1)  # Adjust if you have more or fewer numeric features
    sns.boxplot(y=data[feature], color=ElleSet[i % len(ElleSet)], width=0.4)
    plt.title(feature.replace('_', ' ').title(), fontsize=10)
    plt.ylabel('')
    plt.xlabel('')
    plt.tight_layout()

plt.suptitle('Outlier Check – Boxplots for Numeric Features', fontsize=14, y=1.02)
plt.show()
```



- Several core financial indicators—Net Income, Net Cash Flow, and ROE—display extreme outliers on both ends, consistent with earlier observations of skewed distributions and mixed profitability across firms.
- Outliers in P/E and P/B Ratios reinforce prior valuation insights: dispersion is not only sector-driven but also driven by individual firms with speculative pricing or distressed fundamentals.
- Cash Ratio and EPS feature long right tails, echoing prior findings from both univariate and violin plots that revealed liquidity and earnings disparities within sectors such as Information Technology and Health Care.

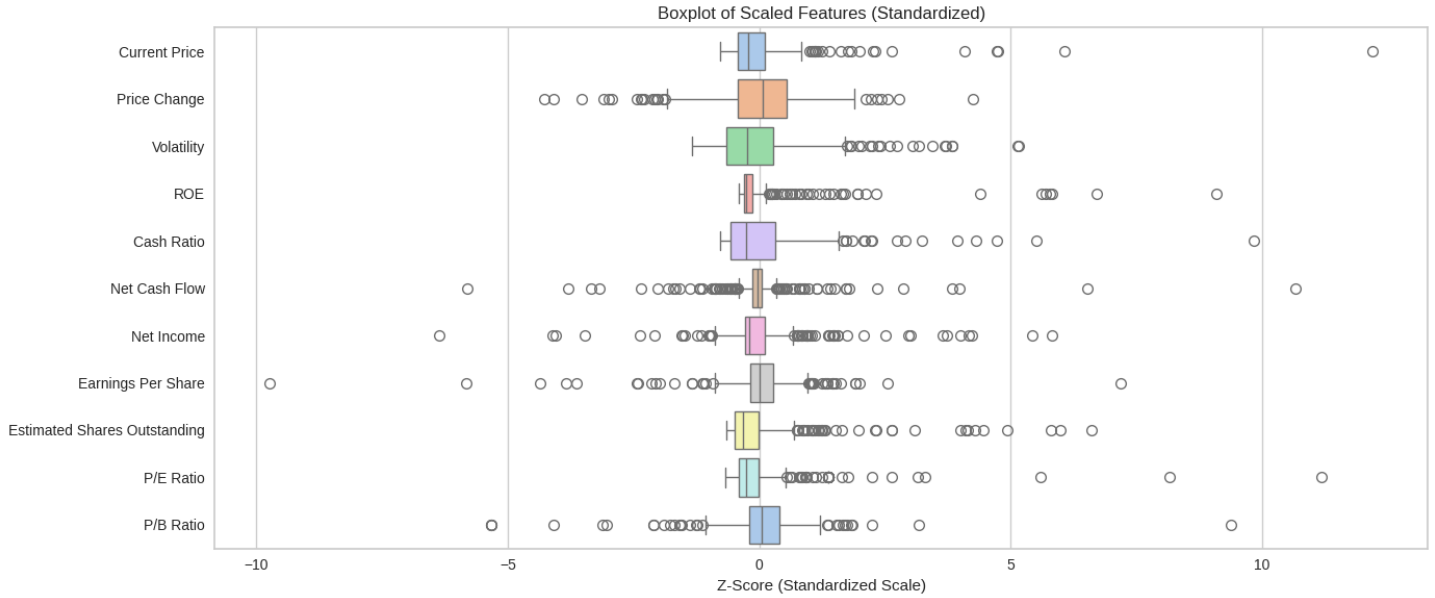
Implications: These outliers are more than statistical noise—they reflect the diversity of firm profiles in the market. From speculative tech stocks with sky-high valuations to asset-heavy firms with thin margins, the financial landscape is far from uniform. Recognizing this spread is essential before clustering or dimensionality reduction, as extreme values may influence distance-based models or distort principal components. Rather than eliminate, understanding these extremes helps preserve the nuance that may define unique investment patterns.

Scaling

```
# Scaling numeric features before dimensionality reduction and clustering
numeric_features = data.select_dtypes(include='number').columns

scaled_data = StandardScaler().fit_transform(data[numeric_features])
scaled_data = pd.DataFrame(scaled_data, columns=numeric_features)
```

```
# Dataframe for scaled data
scaled_data = pd.DataFrame(scaled_data, columns=numeric_features)
```

```
# Visual sanity check for scaling data
plt.figure(figsize=(14, 6))
sns.boxplot(data=scaled_data, orient="h", palette="pastel")
plt.title("Boxplot of Scaled Features (Standardized)")
plt.xlabel("Z-Score (Standardized Scale)")
plt.tight_layout()
plt.show()
```



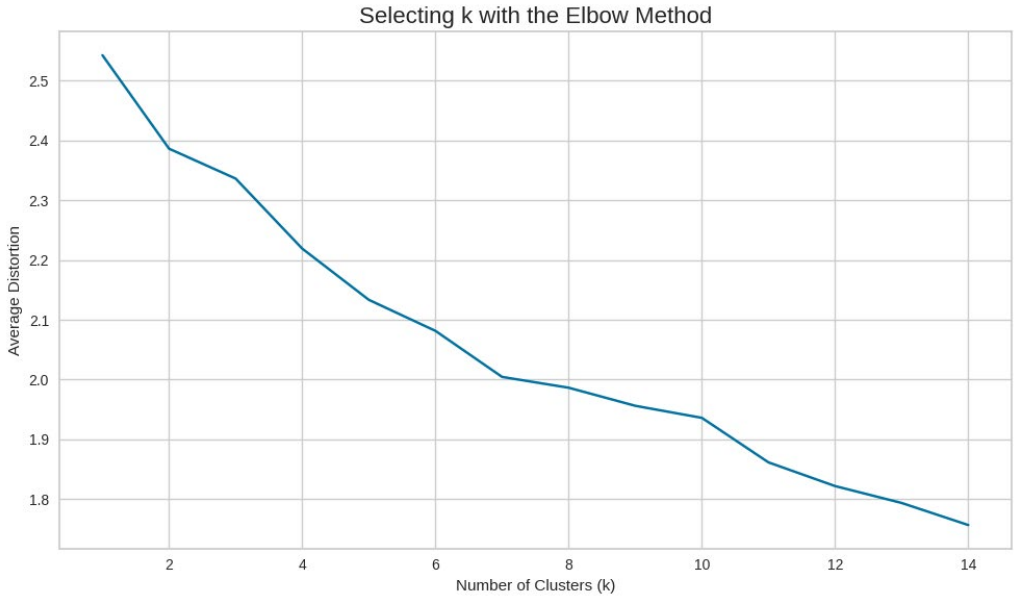
- All numeric features are centered near 0 with comparable spread (i.e., consistent variance) across features.

KMeans Clustering

Elbow Plot

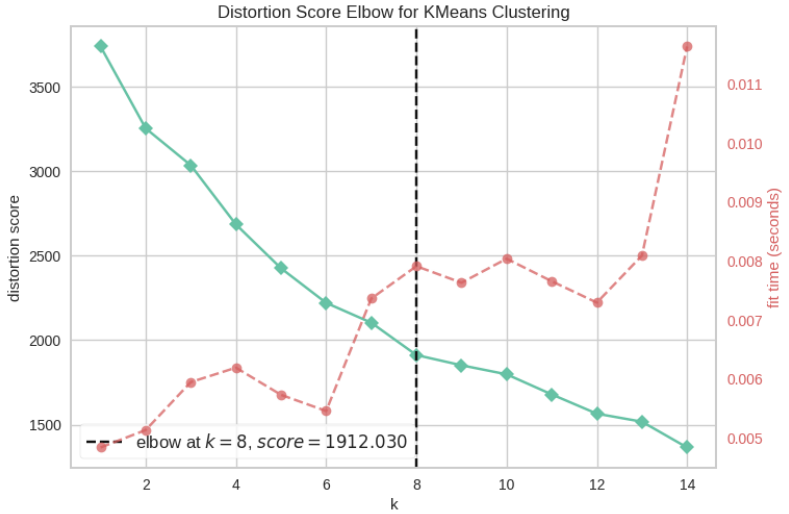
```
# Copy, scaled data for clustering
k_means_df = scaled_data.copy()

# Manual elbow plot using average distortion
clusters = range(1, 15)
meanDistortions = []

for k in clusters:
    model = KMeans(n_clusters=k, random_state=1)
    model.fit(k_means_df)
    prediction = model.predict(k_means_df)
    distortion = (
        sum(np.min(cdist(k_means_df, model.cluster_centers_, "euclidean"), axis=1))
        / k_means_df.shape[0]
    )
    meanDistortions.append(distortion)

    print("Number of Clusters:", k, "\tAverage Distortion:", round(distortion, 2))

plt.figure(figsize=(10, 6))
plt.plot(clusters, meanDistortions, "bx-")
plt.xlabel("Number of Clusters (k)")
plt.ylabel("Average Distortion")
plt.title("Selecting k with the Elbow Method", fontsize=16)
plt.tight_layout()
plt.show()
```

```
Number of Clusters: 1   Average Distortion: 2.54
Number of Clusters: 2   Average Distortion: 2.39
Number of Clusters: 3   Average Distortion: 2.34
Number of Clusters: 4   Average Distortion: 2.22
Number of Clusters: 5   Average Distortion: 2.13
Number of Clusters: 6   Average Distortion: 2.08
Number of Clusters: 7   Average Distortion: 2.0
Number of Clusters: 8   Average Distortion: 1.99
Number of Clusters: 9   Average Distortion: 1.96
Number of Clusters: 10  Average Distortion: 1.94
Number of Clusters: 11  Average Distortion: 1.86
Number of Clusters: 12  Average Distortion: 1.82
Number of Clusters: 13  Average Distortion: 1.79
Number of Clusters: 14  Average Distortion: 1.76
```



- The average distortion decreases sharply up to k=6, then flattens — indicating diminishing returns in within-cluster variance reduction beyond this point.
- A subtle elbow is visible between k=6 and k=8, suggesting this is the region of optimal trade-off between model simplicity and cluster cohesion.

```
#YellowBrick Visualizer
def elbow_plot():
  model = KMeans(random_state=1)
  visualizer = KElbowVisualizer(model, k=(1,15), timings=True)
  visualizer.fit(k_means_df)
  visualizer.show()

show_clean_plot(elbow_plot)
```

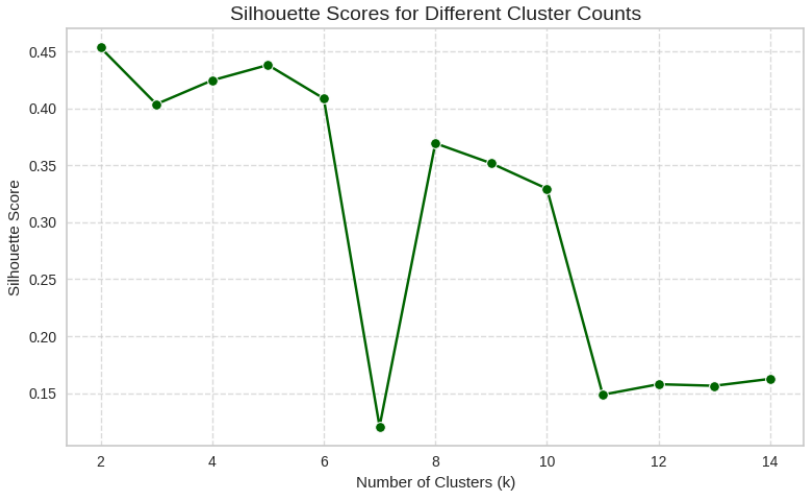


- The distortion score drops sharply from k=1 to k=6, with a visible elbow at k=8, suggesting a point of diminishing marginal improvement in intra-cluster variance reduction.
- Beyond k=8, distortion continues to decline but at a significantly slower rate, indicating less meaningful segmentation.

Silhouette Scores

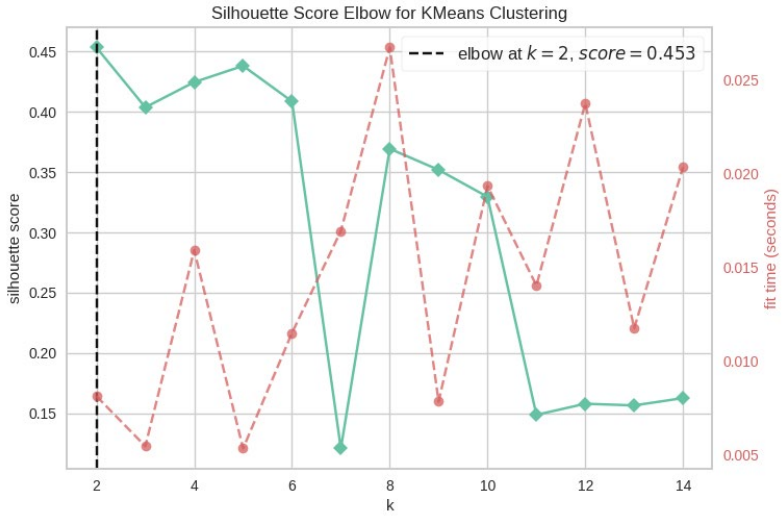
```
# Evaluate cluster quality using silhouette scores
sil_scores = []
cluster_range = range(2, 15)

for k in cluster_range:
    kmeans_model = KMeans(n_clusters=k, random_state=1)
    cluster_labels = kmeans_model.fit_predict(scaled_data.copy())
    score = silhouette_score(scaled_data.copy(), cluster_labels)
    sil_scores.append(score)
    print(f"Clusters: {k}  |  Silhouette Score: {score:.4f}")

# Plot silhouette scores
plt.figure(figsize=(8, 5))
sns.lineplot(x=cluster_range, y=sil_scores, marker="o", color="darkgreen")
plt.title("Silhouette Scores for Different Cluster Counts", fontsize=14)
plt.xlabel("Number of Clusters (k)")
plt.ylabel("Silhouette Score")
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

```
Clusters: 2  |  Silhouette Score: 0.4534
Clusters: 3  |  Silhouette Score: 0.4037
Clusters: 4  |  Silhouette Score: 0.4246
Clusters: 5  |  Silhouette Score: 0.4382
Clusters: 6  |  Silhouette Score: 0.4087
Clusters: 7  |  Silhouette Score: 0.1207
Clusters: 8  |  Silhouette Score: 0.3694
Clusters: 9  |  Silhouette Score: 0.3519
Clusters: 10  |  Silhouette Score: 0.3295
Clusters: 11  |  Silhouette Score: 0.1487
Clusters: 12  |  Silhouette Score: 0.1578
Clusters: 13  |  Silhouette Score: 0.1565
Clusters: 14  |  Silhouette Score: 0.1625
```



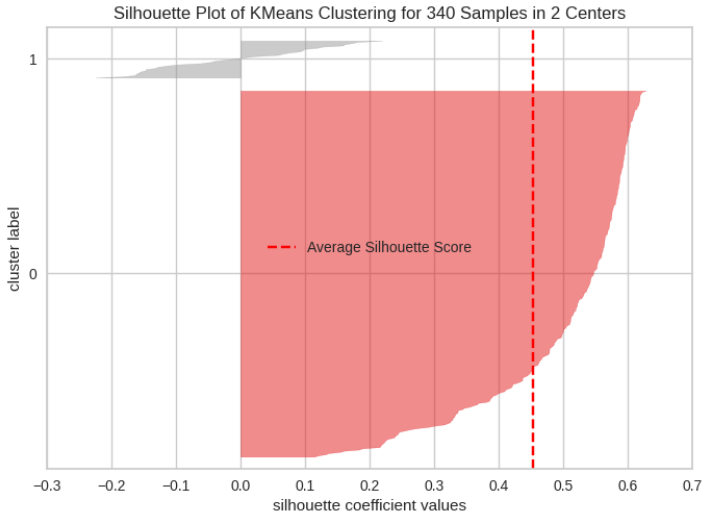
- The strongest silhouette scores (excluding k = 2) are at k = 5 and k = 4, which suggests those configurations create relatively well-defined clusters — groups that are compact and clearly distinct from each other.
- While k = 2 gives the highest overall score, breaking the dataset into just two groups is likely too broad to offer meaningful differentiation for portfolio strategy or client segmentation.
- There is a sharp drop at k = 7, where the score falls to nearly 0.12 — a clear sign that cluster quality drops off at that point. That decline helps rule out k = 7 as a strong candidate and signals a structural shift in the data, reinforcing the idea that stable groupings likely occur below that threshold (e.g., at 4–6 clusters) or potentially at 8, which shows some recovery.

```
# Yellowbrick Visualizer
def plot_silhouette_elbow():
    model = KMeans(random_state=1)
    visualizer = KElbowVisualizer(model, k=(2, 15), metric="silhouette", timings=True)
    visualizer.fit(k_means_df)
    visualizer.show()

# Clean Image
plot_silhouette_elbow()
```



- The highest silhouette score is at k = 2, but dividing stocks into only two groups is likely too general to support detailed investment strategies.
- k = 4, 5, and 6 each show relatively high silhouette scores, indicating stronger cohesion and separation — these values may represent more stable and meaningful groupings.
- There is a sharp drop at k = 7, followed by a modest recovery at k = 8, suggesting that forcing the data into seven clusters may break apart groupings that were previously more stable — leading to weaker, less meaningful separation between stocks..

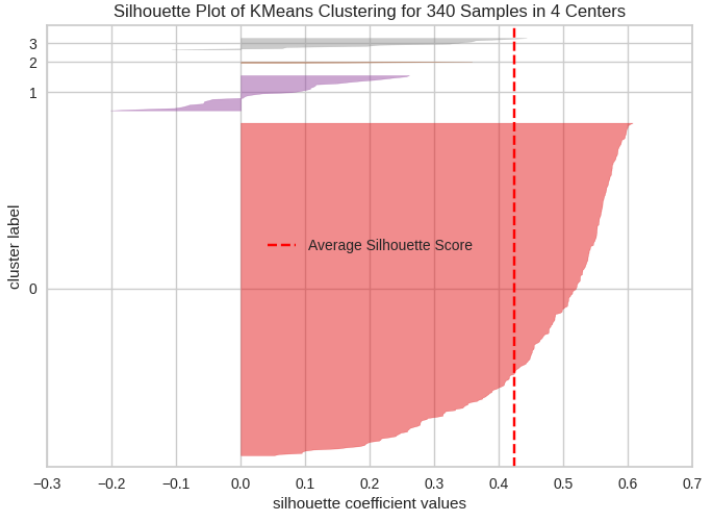
```
# Optimal cluster review with silhouette coefficients
def plot_silhouette_k2():
    visualizer = SilhouetteVisualizer(KMeans(n_clusters=2, random_state=1))
    visualizer.fit(k_means_df)
    visualizer.show()
plot_silhouette_k2()
```



- While the silhouette score is high, this result splits the data into just two broad groups, which is too coarse to support differentiated portfolio insights.
- One cluster is small and moderately distinct; the other absorbs the majority of stocks — offering limited value for segmentation or strategy design.
- This outcome suggests that although technically well-separated, k = 2 lacks the granularity needed to guide real-world investment recommendations.

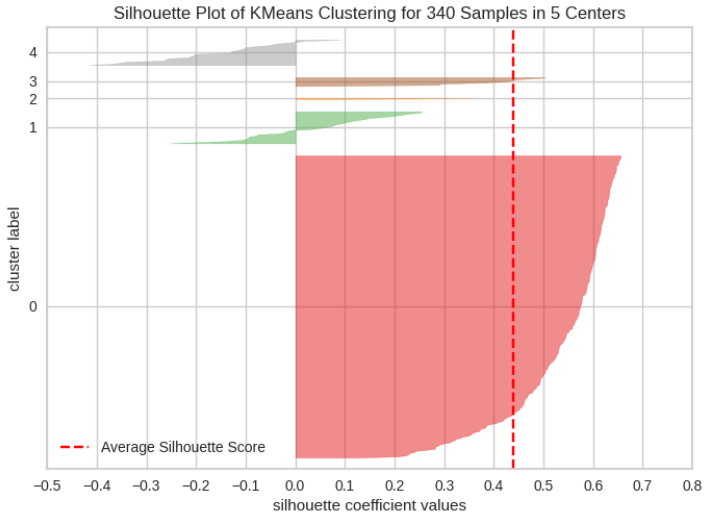
```
# Optimal cluster review with silhouette coefficients
def plot_silhouette_k2():
    visualizer = SilhouetteVisualizer(KMeans(n_clusters=4, random_state=1))
    visualizer.fit(k_means_df)
    visualizer.show()
```

```
plot_silhouette_k2()
```



- The average silhouette score remains strong, and most clusters show good cohesion, supporting the presence of distinct behavioral groups among the stocks.
- However, one cluster (Cluster 0) has a wide spread of low and negative silhouette values, indicating instability or overlap with other groups — a potential weakness in segmentation clarity.
- Overall, this configuration may begin to reveal meaningful segmentation useful for portfolio strategy, but the presence of a poorly defined cluster suggests the model may still be under- or misfitting part of the data.

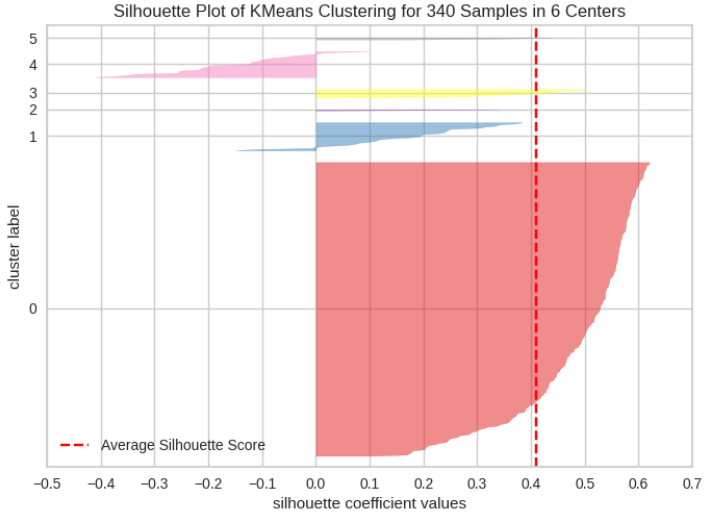
```
# Optimal cluster review with silhouette coefficients
def plot_silhouette_k2():
  visualizer = SilhouetteVisualizer(KMeans(n_clusters=5, random_state=1))
  visualizer.fit(k_means_df)
  visualizer.show()
plot_silhouette_k2()
```



- The average silhouette score remains high, and four of the five clusters appear compact and well-separated, indicating a structure that could support meaningful stock grouping.
- Cluster 0 dominates the distribution, while the remaining clusters are relatively small — this imbalance could limit the model's ability to deliver balanced segmentation across investment strategies.

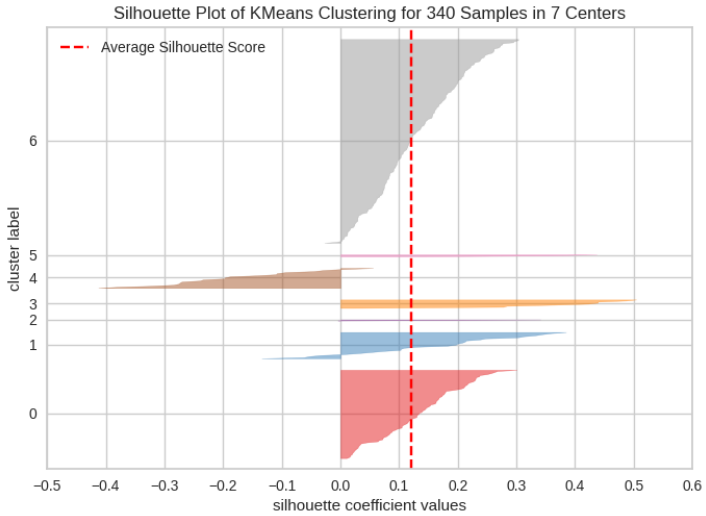
```
# Optimal cluster review with silhouette coefficients
def plot_silhouette_k2():
  visualizer = SilhouetteVisualizer(KMeans(n_clusters=6, random_state=1))
  visualizer.fit(k_means_df)
  visualizer.show()
```

```
plot_silhouette_k2()
```



- The overall silhouette score remains high, suggesting strong structure and separation across most clusters.
- However, the presence of several very small clusters — some with negative silhouette values — raises concerns about over-fragmentation, where added clusters may not add meaningful distinction.
- This configuration may begin to capture niche stock behaviors, but it may risk introducing clusters that are too small or unstable to reliably inform investment strategy.

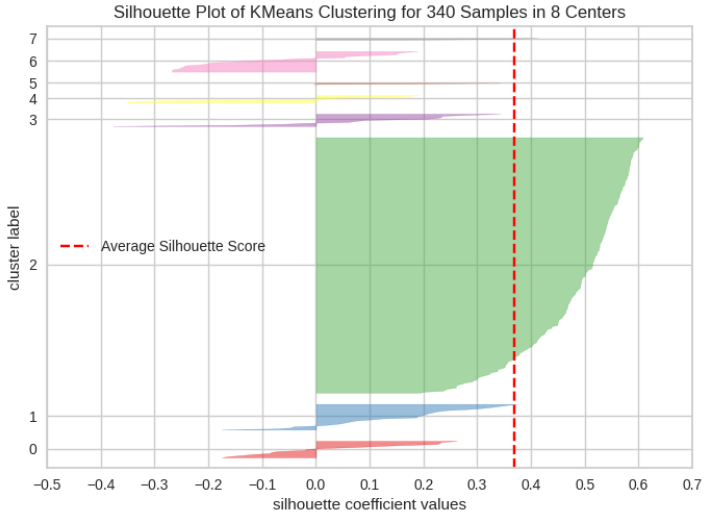
```
# Optimal cluster review with silhouette coefficients
def plot_silhouette_k2():
  visualizer = SilhouetteVisualizer(KMeans(n_clusters=7, random_state=1))
  visualizer.fit(k_means_df)
  visualizer.show()
plot_silhouette_k2()
```



- The average silhouette score drops sharply, and multiple clusters show overlap, low cohesion, or negative values, indicating a breakdown in clustering quality.
- Clusters 0 and 1 account for most of the data but struggle to maintain internal consistency, while the remaining clusters appear small and weakly defined.
- This result suggests that k = 7 fragments the structure without adding meaningful distinction — making it ill-suited for portfolio segmentation.

```
# Optimal cluster review with silhouette coefficients
def plot_silhouette_k2():
  visualizer = SilhouetteVisualizer(KMeans(n_clusters=8, random_state=1))
```

```
  visualizer.fit(k_means_df)
  visualizer.show()
plot_silhouette_k2()
```



- The average silhouette score decreases, but the plot suggests a return to better structure after the disruption seen at k = 7.
- Clusters 1 and 2 show reasonable cohesion, while others remain small but more clearly defined than at k = 7, suggesting the model is starting to reorganize and stabilize.
- Although some overlap and negative values remain (notably in Cluster 0), this configuration begins to reveal nuanced behavioral groupings that may support differentiated portfolio strategies — even if tradeoffs in cohesion exist.

Final Model

```
# Final KMeans model
kmeans = KMeans(n_clusters=6, random_state=1)
kmeans.fit(k_means_df)
```

```
KMeans
KMeans(n_clusters=6, random_state=1)
```

Considerations for selecting 6 clusters:

- Offers strong granularity while maintaining interpretability, making it well-suited to portfolio strategy and investment planning goals.
- Enables a manageable number of well-defined stock groupings that can be labeled, profiled, and applied within real-world financial decision-making.

```
# creating a copy of the original data
df1 = df.copy()

# Add kmeans cluster labels to the original and scaled dataframes
k_means_df["KM_segments"] = kmeans.labels_
df1["KM_segments"] = kmeans.labels_
```

Cluster Profiling

```
# Cluster Profiling: Numerical Feature Means + Cluster Sizes
km_cluster_profile = (
    df1.select_dtypes(include=['number'])
        .groupby("KM_segments")
        .mean()
)

# Add count of securities in each cluster
km_cluster_profile["count_in_each_segment"] = (
    df1.groupby("KM_segments")["Security"].count().values
)

# View profile
km_cluster_profile.style.highlight_max(props="background-color: lightgreen; color: black;", axis=0)
```

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	72.976734	5.192788	1.376317	34.941606	53.197080	-10337605.839416	1495249751.824817	3.650566	435769103.743102	23.699140	-3.417280	274
1	34.231808	-15.515565	2.832069	48.037037	47.740741	-128651518.518519	-2444318518.518518	-6.284444	503031539.057037	75.627265	1.655990	27
2	24.485001	-13.351992	3.482611	802.000000	51.000000	-1292500000.000000	-19106500000.000000	-41.815000	519573983.250000	60.748608	1.565141	2
3	46.672222	5.166566	1.079367	25.000000	58.333333	-3040666666.666667	14848444444.444445	3.435556	4564959946.222222	15.596051	-6.354193	9
4	204.930931	11.885400	1.699460	30.040000	279.920000	2121806560.000000	2724131680.000000	6.558000	799318866.712800	43.517668	15.097707	25
5	327.006671	21.917380	2.029752	4.000000	106.000000	698240666.666667	287547000.000000	0.750000	366763235.300000	400.989188	-5.322376	3

- Cluster 3 dominates as a high-performing growth segment with strong price gains, earnings, and cash flow — ideal for aggressive portfolio strategies.
- Cluster 5 and Cluster 4 offer low-volatility, cash-strong profiles, suited for stable, income-oriented investing.
- Cluster 0 and Cluster 1 reflect mature or underperforming stocks, with modest or negative returns and weaker fundamentals.
- Cluster 2 is a small outlier group with extreme financials, likely not suitable for strategic allocation.

```
## Companies in each cluster
for cl in df1["KM_segments"].unique():
    print("In cluster {}, the following companies are present:".format(cl))
    print(df1[df1["KM_segments"] == cl]["Security"].unique())
    print()
```

```
In cluster 0, the following companies are present:
['American Airlines Group' 'AbbVie' 'Abbott Laboratories'
 'Adobe Systems Inc' 'Archer-Daniels-Midland Co' 'Ameren Corp'
 'American Electric Power' 'AFLAC Inc'
 'American International Group, Inc.' 'Apartment Investment & Mgmt'
 'Assurant Inc' 'Arthur J. Gallagher & Co.' 'Akamai Technologies Inc'
 'Albemarle Corp' 'Alaska Air Group Inc' 'Allstate Corp' 'Allegion'
 'Applied Materials Inc' 'AMETEK Inc' 'Affiliated Managers Group Inc'
 'Ameriprise Financial' 'American Tower Corp A' 'AutoNation Inc'
 'Anthem Inc.' 'Aon plc' 'Amphenol Corp' 'Activision Blizzard'
 'AvalonBay Communities, Inc.' 'Broadcom'
 'American Water Works Company Inc' 'American Express Co' 'Boeing Company'
 'Baxter International Inc.' 'BB&T Corporation' 'Bard (C.R.) Inc.'
 'The Bank of New York Mellon Corp.' 'Ball Corp' 'Bristol-Myers Squibb'
 'Boston Scientific' 'BorgWarner' 'Boston Properties' 'Caterpillar Inc.'
 'Chubb Limited' 'CBRE Group' 'Crown Castle International Corp.'
 'Carnival Corp.' 'CF Industries Holdings Inc' 'Citizens Financial Group'
 'Church & Dwight' 'C. H. Robinson Worldwide' 'Charter Communications'
 'CIGNA Corp.' 'Cincinnati Financial' 'Colgate-Palmolive' 'Comerica Inc.'
 'CME Group Inc.' 'Cummins Inc.' 'CMS Energy' 'Centene Corporation'
 'CenterPoint Energy' 'Capital One Financial' 'The Cooper Companies'
 'CSX Corp.' 'CenturyLink Inc' 'Cognizant Technology Solutions'
 'Citrix Systems' 'CVS Health' 'Chevron Corp.' 'Dominion Resources'
 'Delta Air Lines' 'Du Pont (E.I.)' 'Deere & Co.'
 'Discover Financial Services' 'Quest Diagnostics' 'Danaher Corp.'
 'The Walt Disney Company' 'Discovery Communications-A'
 'Discovery Communications-C' 'Delphi Automotive' 'Digital Realty Trust'
 'Dun & Bradstreet' 'Dover Corp.' 'Dr Pepper Snapple Group' 'Duke Energy'
 'DaVita Inc.' 'eBay Inc.' 'Ecolab Inc.' 'Consolidated Edison'
 'Equifax Inc.' "Edison Int'l" 'Eastman Chemical' 'Equity Residential'
 'Eversource Energy' 'Essex Property Trust, Inc.' 'E*Trade'
 'Eaton Corporation' 'Entergy Corp.' 'Exelon Corp.' "Expeditors Int'l"
 'Expedia Inc.' 'Extra Space Storage' 'Fastenal Co'
 'Fortune Brands Home & Security' 'FirstEnergy Corp'
 'Fidelity National Information Services' 'Fiserv Inc' 'FLIR Systems'
 'Fluor Corp.' 'Flowserve Corporation' 'FMC Corporation'
 'Federal Realty Investment Trust' 'General Dynamics'
 'General Growth Properties Inc.' 'Corning Inc.' 'General Motors'
 'Genuine Parts' 'Garmin Ltd.' 'Goodyear Tire & Rubber'
 'Grainger (W.W.) Inc.' 'Hasbro Inc.' 'Huntington Bancshares'
 'HCA Holdings' 'Welltower Inc.' 'HCP Inc.' 'Hartford Financial Svc.Gp.'
 'Harley-Davidson' "Honeywell Int'l Inc." 'HP Inc.' 'Hormel Foods Corp.'
 'Henry Schein' 'Host Hotels & Resorts' 'The Hershey Company'
 'Humana Inc.' 'International Business Machines' 'IDEXX Laboratories'
 'Intl Flavors & Fragrances' 'International Paper' 'Interpublic Group'
 'Iron Mountain Incorporated' 'Illinois Tool Works' 'Invesco Ltd.'
 'J. B. Hunt Transport Services' 'Jacobs Engineering Group'
 'Juniper Networks' 'Kimco Realty' 'Kimberly-Clark' 'Kansas City Southern'
 'Leggett & Platt' 'Lennar Corp.' 'Laboratory Corp. of America Holding'
 'LKQ Corporation' 'L-3 Communications Holdings' 'Lilly (Eli) & Co.'
 'Lockheed Martin Corp.' 'Alliant Energy Corp' 'Leucadia National Corp.'
 'Southwest Airlines' 'Level 3 Communications' 'LyondellBasell'
 'Mastercard Inc.' 'Mid-America Apartments' 'Macerich' "Marriott Int'l."
 'Masco Corp.' 'Mattel Inc.' "Moody's Corp" 'Mondelez International'
 'MetLife Inc.' 'Mohawk Industries' 'Mead Johnson' 'McCormick & Co.'
 'Martin Marietta Materials' 'Marsh & McLennan' '3M Company'
 'Altria Group Inc' 'Marathon Petroleum' 'Merck & Co.' 'M&T Bank Corp.'
 'Mettler Toledo' 'Mylan N.V.' 'Navient' 'NASDAQ OMX Group'
 'NextEra Energy' 'Newmont Mining Corp. (Hldg. Co.)' 'Nielsen Holdings'
```

```
 'Norfolk Southern Corp.' 'Northern Trust Corp.' 'Nucor Corp.'
 'Newell Brands' 'Realty Income Corporation' 'Omnicom Group'
 "O'Reilly Automotive" "People's United Financial" 'Pitney-Bowes'
 'PACCAR Inc.' 'PG&E Corp.' 'Public Serv. Enterprise Inc.' 'PepsiCo Inc.'
 'Principal Financial Group' 'Procter & Gamble' 'Progressive Corp.'
 'Pulte Homes Inc.' 'Philip Morris International' 'PNC Financial Services'
 'Pentair Ltd.' 'Pinnacle West Capital' 'PPG Industries' 'PPL Corp.'
 'Prudential Financial' 'Phillips 66' 'Praxair Inc.' 'PayPal'
 'Ryder System' 'Royal Caribbean Cruises Ltd' 'Robert Half International'
 'Roper Industries' 'Republic Services Inc' 'SCANA Corp'
 'Charles Schwab Corporation' 'Sealed Air' 'Sherwin-Williams'
 'SL Green Realty' 'Scripps Networks Interactive Inc.' 'Southern Co.'
 'Simon Property Group Inc' 'S&P Global, Inc.' 'Stericycle Inc'
 'Sempra Energy' 'SunTrust Banks' 'State Street Corp.'
 'Skyworks Solutions' 'Synchrony Financial' 'Stryker Corp.'
 'Molson Coors Brewing Company' 'Tegna, Inc.' 'Torchmark Corp.'
 'Thermo Fisher Scientific' 'The Travelers Companies Inc.'
 'Tractor Supply Company' 'Tyson Foods' 'Tesoro Petroleum Co.'
 'Total System Services' 'Texas Instruments' 'Under Armour'
 'United Continental Holdings' 'UDR Inc' 'Universal Health Services, Inc.'
 'United Health Group Inc.' 'Unum Group' 'Union Pacific'
 'United Parcel Service' 'United Technologies' 'Varian Medical Systems'
 'Valero Energy' 'Vulcan Materials' 'Vornado Realty Trust'
 'Verisk Analytics' 'Verisign Inc.' 'Ventas Inc' 'Wec Energy Group Inc'
 'Whirlpool Corp.' 'Waste Management Inc.' 'Western Union Co'
 'Weyerhaeuser Corp.' 'Wyndham Worldwide' 'Xcel Energy Inc' 'XL Capital'
 'Dentsply Sirona' 'Xerox Corp.' 'Xylem Inc.' 'Yum! Brands Inc'
 'Zimmer Biomet Holdings' 'Zions Bancorp' 'Zoetis']

In cluster 4, the following companies are present:
['Analog Devices, Inc.' 'Alliance Data Systems' 'Amgen Inc'
 'Bank of America Corp' 'BIOGEN IDEC Inc.' 'Celgene Corp.'
 'Chipotle Mexican Grill' 'Equinix' 'Edwards Lifesciences' 'Facebook'
 'First Solar Inc' 'Frontier Communications' 'Gilead Sciences'
 'Halliburton Co.' 'Intel Corp.' 'Intuitive Surgical Inc.'
 "McDonald's Corp." 'Monster Beverage' 'Priceline.com Inc' 'Regeneron'
 'TripAdvisor' 'Vertex Pharmaceuticals Inc' 'Waters Corporation'
 'Wynn Resorts Ltd' 'Yahoo Inc.']

In cluster 5, the following companies are present:
['Alexion Pharmaceuticals' 'Amazon.com Inc' 'Netflix Inc.']

In cluster 2, the following companies are present:
['Apache Corporation' 'Chesapeake Energy']

In cluster 1, the following companies are present:
['Anadarko Petroleum Corp' 'Arconic Inc' 'Baker Hughes Inc'
 'Cabot Oil & Gas' 'Concho Resources' 'Devon Energy Corp.' 'EOG Resources'
 'EQT Corporation' 'Freeport-McMoran Cp & Gld' 'Hess Corporation'
 'Hewlett Packard Enterprise' 'Kinder Morgan' 'The Mosaic Company'
 'Marathon Oil Corp.' 'Murphy Oil' 'Noble Energy Inc'
 'Newfield Exploration Co' 'National Oilwell Varco Inc.' 'ONEOK'
 'Occidental Petroleum' 'Quanta Services Inc.' 'Range Resources Corp.'
 'Spectra Energy Corp.' 'Southwestern Energy' 'Teradata Corp.'
 'Williams Cos.' 'Cimarex Energy']

In cluster 3, the following companies are present:
['Citigroup Inc.' 'Ford Motor' 'JPMorgan Chase & Co.' 'Coca Cola Company'
 'Pfizer Inc.' 'AT&T Inc' 'Verizon Communications' 'Wells Fargo'
 'Exxon Mobil Corp.']
```

- While this project identifies financial and structural patterns across clusters, a sector-level interpretation of individual securities may benefit from review by industry analysts or portfolio specialists with deeper domain knowledge.

```
df1.groupby(["KM_segments", "GICS Sector"])['Security'].count()
```

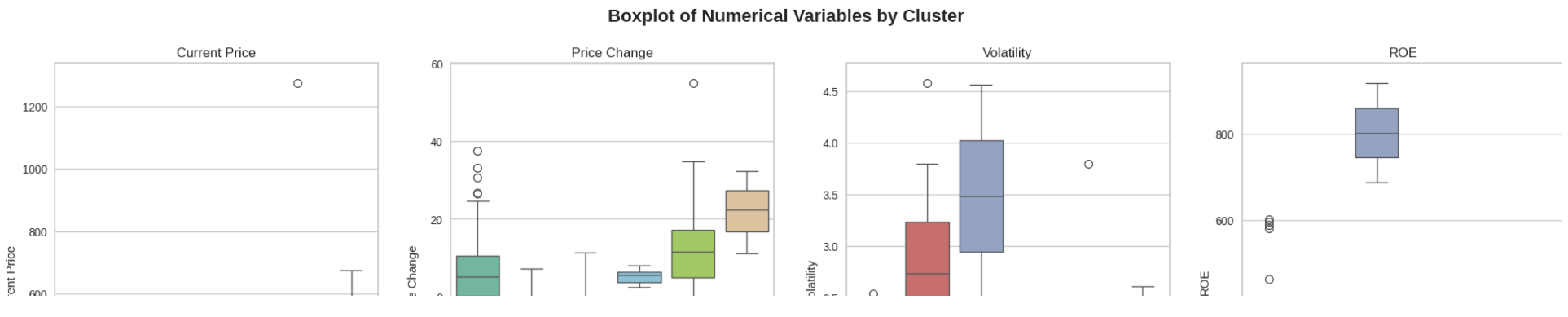
		Security
KM_segments	GICS Sector	
0	Consumer Discretionary	33
	Consumer Staples	17
	Energy	5
	Financials	45
	Health Care	29
	Industrials	51
	Information Technology	24
	Materials	18
	Real Estate	26
	Telecommunications Services	2
	Utilities	24

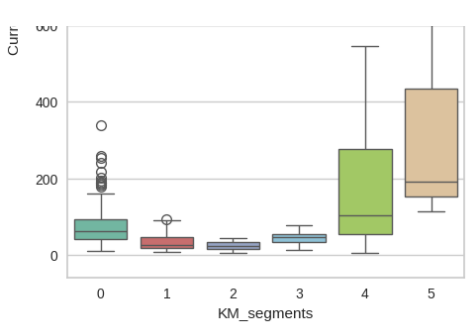
1	Energy	21
	Industrials	2
	Information Technology	2
	Materials	2
2	Energy	2
3	Consumer Discretionary	1
	Consumer Staples	1
	Energy	1
	Financials	3
	Health Care	1
	Telecommunications Services	2
4	Consumer Discretionary	5
	Consumer Staples	1
	Energy	1
	Financials	1
	Health Care	9
	Information Technology	6
	Real Estate	1
	Telecommunications Services	1
5	Consumer Discretionary	1
	Health Care	1
	Information Technology	1

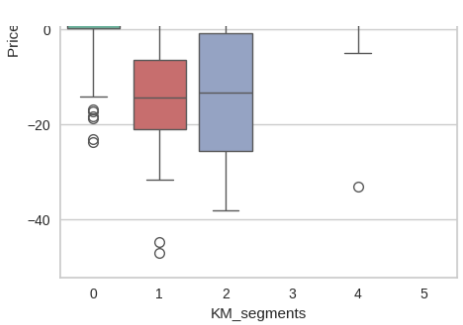
dtype: int64

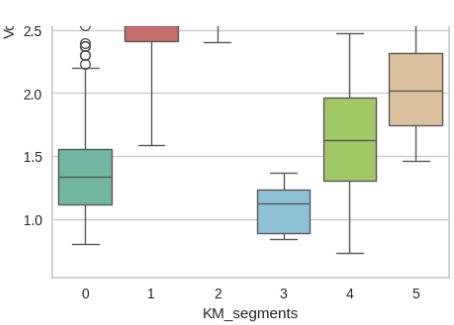
- Cluster 0 captures a broad cross-section of the market, with strong representation across all sectors.
- Cluster 3 is similarly diverse, especially in Financials, Industrials, and Utilities, suggesting a balanced but performance-differentiated segment.
- Clusters 1 and 2 are heavily skewed toward Energy, with Cluster 2 holding just two securities—likely outliers.
- Clusters 4 and 5 contain only a few securities each, spread thinly across multiple sectors, indicating niche or atypical profiles.

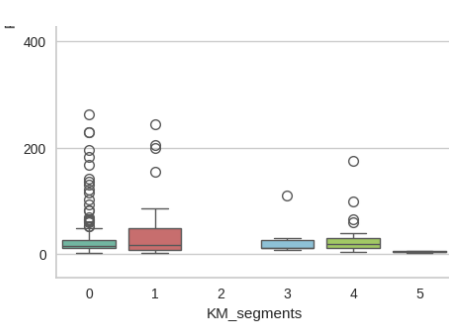
```
plt.figure(figsize=(20, 20))
plt.suptitle("Boxplot of Numerical Variables by Cluster", fontsize=16, fontweight='bold')

box_colors = ElleSet

num_col = df.select_dtypes(include=np.number).columns.tolist()

for i, variable in enumerate(num_col):
    plt.subplot(3, 4, i + 1)
    sns.boxplot(
        data=df1,
        x="KM_segments",
        y=variable,
        palette=box_colors
    )
    plt.title(variable, fontsize=12)

plt.tight_layout(pad=2.0)
```

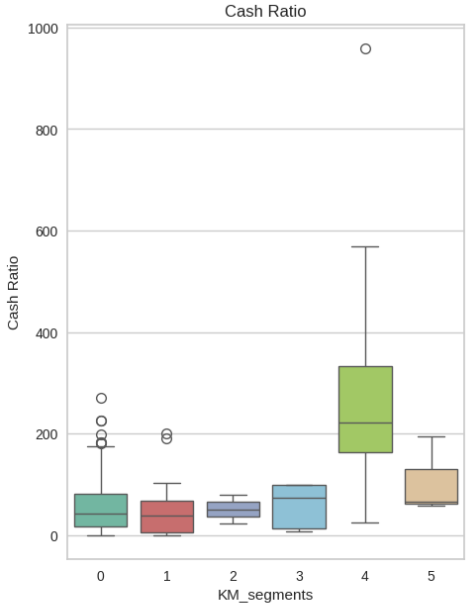


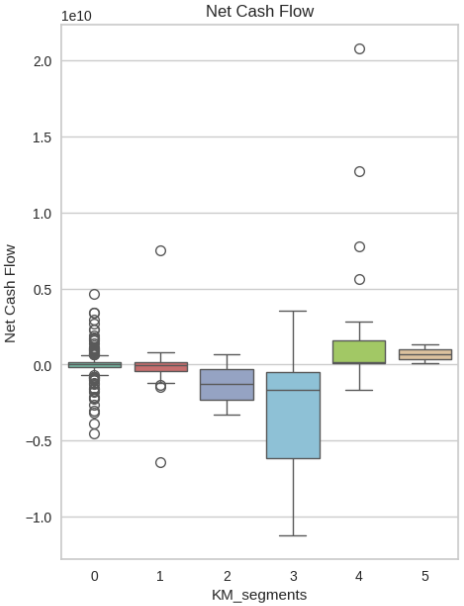

KM_segments

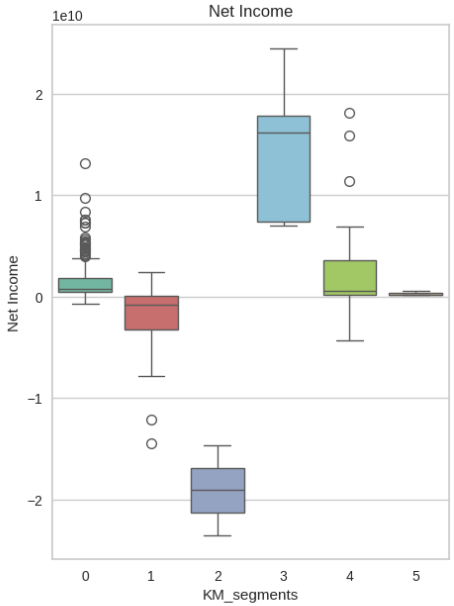

KM_segments

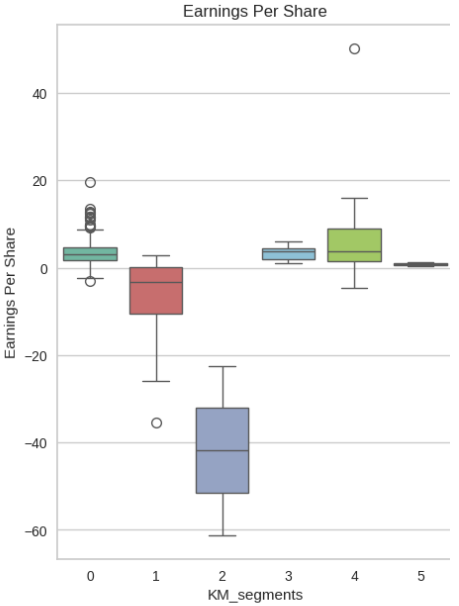

KM_segments

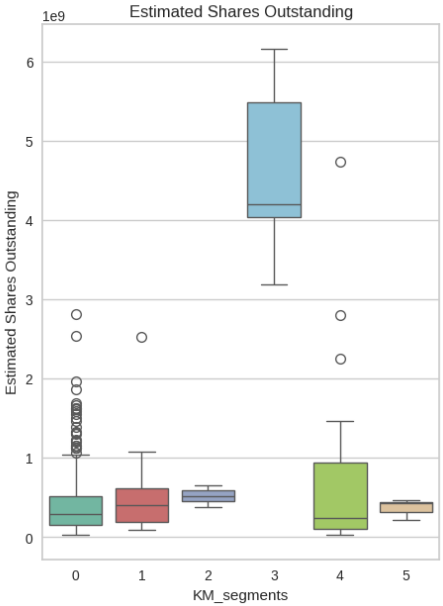

KM_segments

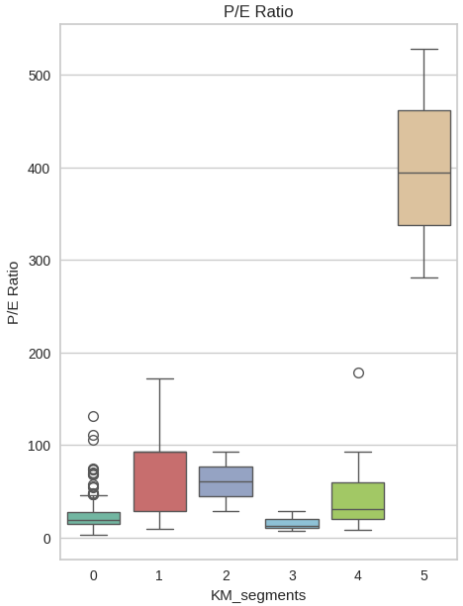

Cash Ratio
Cash Ratio
KM_segments

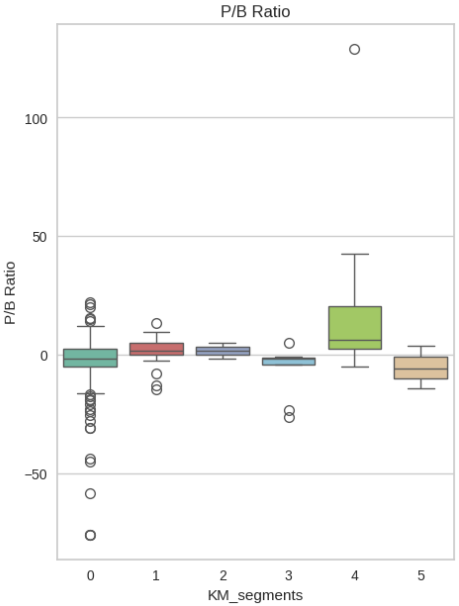

Net Cash Flow
1e10
Net Cash Flow
KM_segments


Net Income
1e10
Net Income
KM_segments


Earnings Per Share
Earnings Per Share
KM_segments


Estimated Shares Outstanding
1e9
Estimated Shares Outstanding
KM_segments


P/E Ratio
P/E Ratio
KM_segments


P/B Ratio
P/B Ratio
KM_segments

- Boxplots confirm key distinctions between clusters — such as high volatility and negative earnings in Cluster 2, and strong income and cash flow in Clusters 3 and 5 — supporting the validity of the segmentation before moving into final insights.

Insights

- Six clusters were selected as the optimal balance between silhouette score stability, elbow curve inflection, and business interpretability — avoiding the fragmentation seen at k = 7 and the complexity of k = 8.
- The model produced structurally sound and behaviorally distinct groups, with clear variation across pricing, earnings, volatility, and liquidity — key inputs for investment strategy.
- Cluster 3 represents a dominant growth segment, while Clusters 4 and 5 align with lower-risk, income-focused profiles. These groupings support differentiated portfolio design.
- Cluster 2, a high-volatility, low-performance outlier flagged in silhouette analysis, validated the method's ability to surface niche segments with unique risk implications.

Hierarchical Clustering

Cophenetic Correlation

```
hc_df = scaled_data.copy()
```

```
# Cophenetic Correlation Analysis: Evaluate distance-linkage combinations
distance_metrics = ['euclidean', 'cityblock', 'cosine']
linkage_methods = ['ward', 'complete', 'average', 'single']

high_cophenet_corr = 0
best_combo = ['', '']  # To store best distance and linkage

for dm in distance_metrics:
    for lm in linkage_methods:
        if lm == 'ward' and dm != 'euclidean':
            continue  # Ward requires Euclidean distance only

        Z = linkage(hc_df, metric=dm, method=lm)
        c, _ = cophenet(Z, pdist(hc_df))

        print(f"Cophenetic correlation for {dm.capitalize()} distance and {lm} linkage is {c:.4f}.")

        if c > high_cophenet_corr:
            high_cophenet_corr = c
            best_combo = [dm, lm]

# Print best combination
print('*' * 90)
print(f"Highest cophenetic correlation is {high_cophenet_corr:.4f}, achieved with {best_combo[0].capitalize()} distance and {best_combo[1]} linkage.")
```

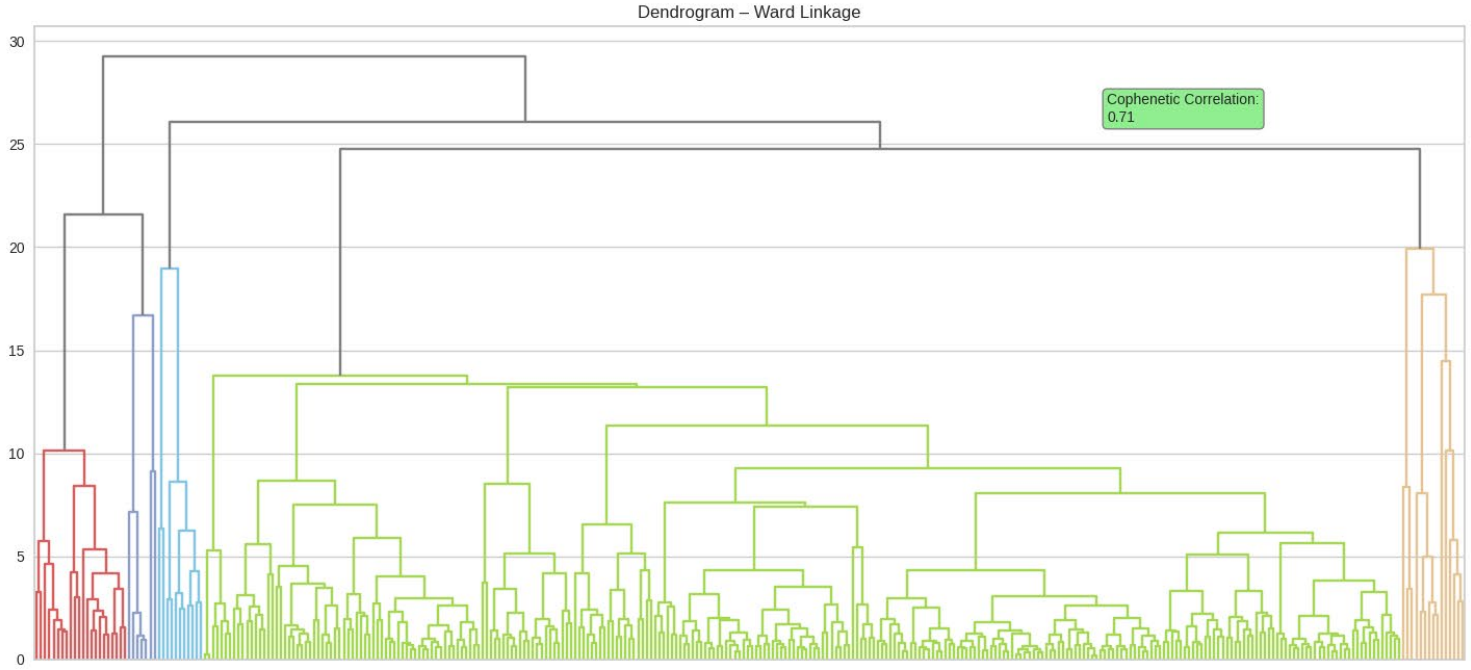
```
Cophenetic correlation for Euclidean distance and ward linkage is 0.7101.
Cophenetic correlation for Euclidean distance and complete linkage is 0.7873.
Cophenetic correlation for Euclidean distance and average linkage is 0.9423.
Cophenetic correlation for Euclidean distance and single linkage is 0.9232.
Cophenetic correlation for Cityblock distance and complete linkage is 0.7375.
Cophenetic correlation for Cityblock distance and average linkage is 0.9302.
Cophenetic correlation for Cityblock distance and single linkage is 0.9334.
Cophenetic correlation for Cosine distance and complete linkage is 0.1456.
Cophenetic correlation for Cosine distance and average linkage is 0.2757.
Cophenetic correlation for Cosine distance and single linkage is 0.1634.
******************************************************************************************
Highest cophenetic correlation is 0.9423, achieved with Euclidean distance and average linkage.
```

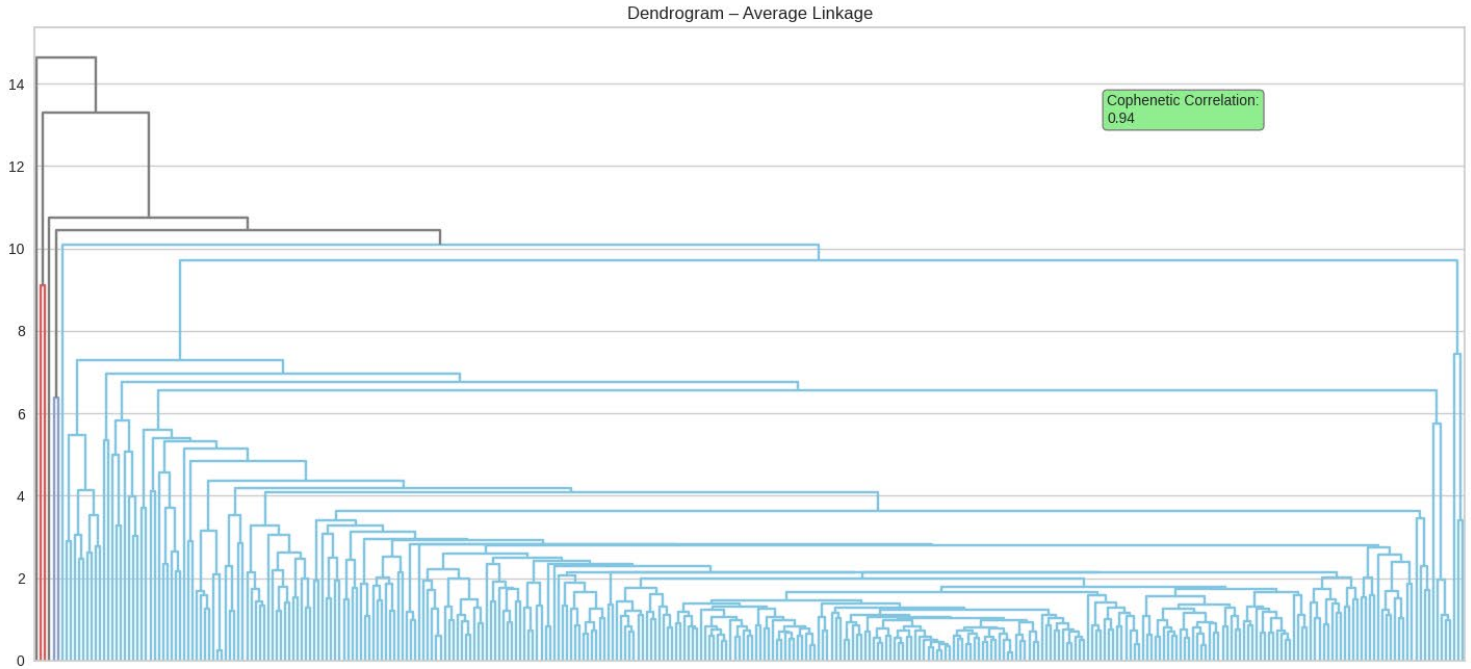
- Highest cophenetic correlation is 0.9423, achieved with Euclidean distance and average linkage.

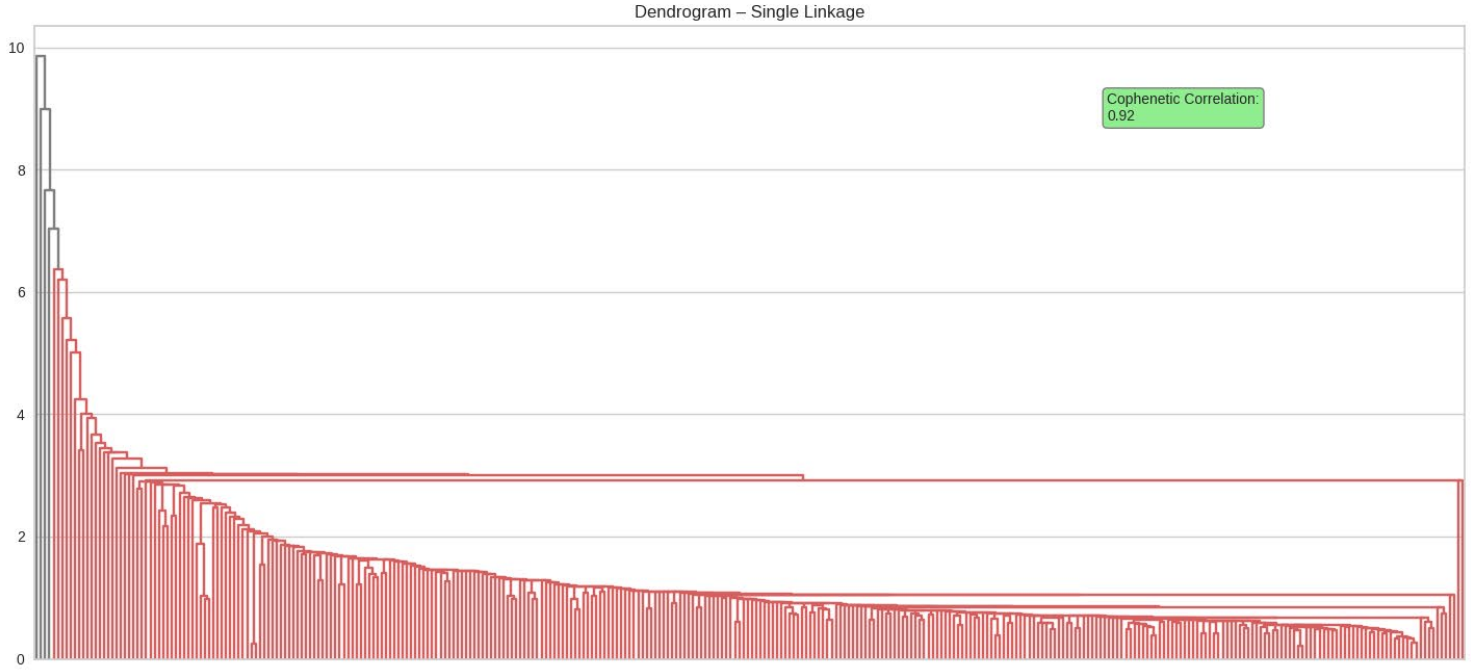
Dendogram Review

```
# Dedograms
linkage_methods = ['ward', 'complete', 'average', 'single']
compare = []
fig, axs = plt.subplots(len(linkage_methods), 1, figsize=(15, 28))

_ = use_ElleSet()
palette = ElleSet.copy()

for i, method in enumerate(linkage_methods):
    Z = linkage(hc_df, method=method, metric='euclidean')
    coph_corr, _ = cophenet(Z, pdist(hc_df))
    compare.append([method, coph_corr])

    axs[i].set_title(f"Dendrogram – {method.capitalize()} Linkage", fontsize=12)

    # Pick threshold for coloring major clusters
    color_thresh = 0.7 * max(Z[:, 2])

    # Color assignment will rotate through ElleSet colors
    dendrogram(
        Z,
```

```
        ax=axs[i],
        color_threshold=color_thresh,
        above_threshold_color='gray',
        no_labels=True
    )

    axs[i].annotate(
        f"Cophenetic Correlation:\n{coph_corr:.2f}",
        xy=(0.75, 0.85),
        xycoords="axes fraction",
        fontsize=10,
        bbox=dict(boxstyle="round,pad=0.3", fc="lightgreen", ec="gray", lw=1)
    )

plt.tight_layout(pad=3.0)
```







- Average linkage shows the strongest cluster preservation, with a cophenetic correlation of 0.94, indicating high fidelity between the dendrogram structure and original distances.

- Single linkage underperforms visually and structurally, despite a high cophenetic score (0.92); it suffers from chaining effects and does not form well-separated, usable groups for financial segmentation.
- Ward and Complete linkage offer moderate structure, with cophenetic correlations of 0.71 and 0.79, respectively. Ward linkage emphasizes variance minimization but introduces imbalance in group sizes.
- Visual separation is clearer under average and ward methods, making them more suitable for downstream business interpretation and client-facing insights.

```
# Create and display a dataframe to compare cophenetic correlations across linkage methods
compare_cols = ['Linkage Method', 'Cophenetic Correlation']
hc_compare_df = pd.DataFrame(compare, columns=compare_cols)

# Sort by Cophenetic Correlation (descending) for better readability
hc_compare_df = hc_compare_df.sort_values(by='Cophenetic Correlation', ascending=False)
hc_compare_df.reset_index(drop=True, inplace=True)

# Display the sorted comparison table
hc_compare_df
```

	Linkage Method	Cophenetic Correlation
0	average	0.942254
1	single	0.923227
2	complete	0.787328
3	ward	0.710118

- Average linkage yields the highest cophenetic correlation (0.94), indicating it most accurately preserves the pairwise distance relationships in the data's hierarchical structure.
- Single linkage also performs well (0.92), but may risk chaining effects—potentially merging loosely related stocks into the same cluster.
- Ward linkage, despite its popularity, performs the weakest (0.71), suggesting it's less appropriate for this dataset's underlying relationships.

SKlearn Model

```
# Create and fit model using average linkage and euclidean distance
hc_model = AgglomerativeClustering(n_clusters=6, metric='euclidean', linkage='average')
hc_model.fit(hc_df)
```

```
AgglomerativeClustering
```

```
AgglomerativeClustering(linkage='average', n_clusters=6)
```

```
# Create a copy of the original data
df2 = df.copy()

# Add hierarchical cluster labels to both scaled and original data
hc_df["HC_segments"] = hc_model.labels_
df2["HC_segments"] = hc_model.labels_
```

Cluster Profiling

```
# Profile each hierarchical cluster by computing mean values of numerical features
hc_cluster_profile = (
    df2.groupby("HC_segments")
       .mean(numeric_only=True)
)

# Add count of securities in each hierarchical cluster
hc_cluster_profile["count_in_each_segment"] = (
    df2.groupby("HC_segments")["Security"].count().values
)

# Display profile with max values highlighted
hc_cluster_profile.style.highlight_max(props="background-color: lightgreen; color: black;", axis=0)
```

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	77.287589	4.099730	1.518066	35.336336	66.900901	-33197321.321321	1538074666.666667	2.885270	560505037.293544	32.441706	-2.174921	333
1	25.640000	11.237908	1.322355	12.500000	130.500000	16755500000.000000	13654000000.000000	3.295000	2791829362.100000	13.649696	1.508484	2
2	24.485001	-13.351992	3.482611	802.000000	51.000000	-1292500000.000000	-19106500000.000000	-41.815000	519573983.250000	60.748608	1.565141	2
3	104.660004	16.224320	1.320606	8.000000	958.000000	592000000.000000	3669000000.000000	1.310000	2800763359.000000	79.893133	5.884467	1
4	1274.949951	3.190527	1.268340	29.000000	184.000000	-1671386000.000000	2551360000.000000	50.090000	50935516.070000	25.453183	-1.052429	1
5	276.570007	6.189286	1.116976	30.000000	25.000000	90885000.000000	596541000.000000	8.910000	66951851.850000	31.040405	129.064585	1

- Cluster 0 dominates the dataset, with 333 out of 469 securities (≈71%)—making it the baseline cluster for comparison.
- Cluster 4 stands out for high returns, showing the highest ROE (802) and a P/E ratio near 0.75, yet suffers the lowest Earnings Per Share (−41.82), suggesting volatility or outlier behavior despite technical strength.
- Cluster 3 signals a strong financial performer, with the highest Cash Ratio (958), highest P/E (79.89), and positive Net Cash Flow, but it is based on a single security, limiting generalizability.
- Cluster 5 appears to balance strength and representation, offering high EPS (8.91), solid ROE (30.0), and a count of 129—second only to Cluster 0—making it a candidate for stable financial profiles.

```
# Print companies assigned to each HC segment
for cl in df2["HC_segments"].unique():
    print(f"In cluster {cl}, the following companies are present:")
    print(df2[df2["HC_segments"] == cl]["Security"].unique())
    print()
```

```
In cluster 0, the following companies are present:
['American Airlines Group' 'AbbVie' 'Abbott Laboratories'
 'Adobe Systems Inc' 'Analog Devices, Inc.' 'Archer-Daniels-Midland Co'
 'Ameren Corp' 'American Electric Power' 'AFLAC Inc'
 'American International Group, Inc.' 'Apartment Investment & Mgmt'
 'Assurant Inc' 'Arthur J. Gallagher & Co.' 'Akamai Technologies Inc'
 'Albemarle Corp' 'Alaska Air Group Inc' 'Allstate Corp' 'Allegion'
 'Alexion Pharmaceuticals' 'Applied Materials Inc' 'AMETEK Inc'
 'Affiliated Managers Group Inc' 'Amgen Inc' 'Ameriprise Financial'
 'American Tower Corp A' 'Amazon.com Inc' 'AutoNation Inc' 'Anthem Inc.'
 'Aon plc' 'Anadarko Petroleum Corp' 'Amphenol Corp' 'Arconic Inc'
 'Activision Blizzard' 'AvalonBay Communities, Inc.' 'Broadcom'
 'American Water Works Company Inc' 'American Express Co' 'Boeing Company'
 'Baxter International Inc.' 'BB&T Corporation' 'Bard (C.R.) Inc.'
 'Baker Hughes Inc' 'BIOGEN IDEC Inc.' 'The Bank of New York Mellon Corp.'
 'Ball Corp' 'Bristol-Myers Squibb' 'Boston Scientific' 'BorgWarner'
 'Boston Properties' 'Citigroup Inc.' 'Caterpillar Inc.' 'Chubb Limited'
 'CBRE Group' 'Crown Castle International Corp.' 'Carnival Corp.'
 'Celgene Corp.' 'CF Industries Holdings Inc' 'Citizens Financial Group'
 'Church & Dwight' 'C. H. Robinson Worldwide' 'Charter Communications'
 'CIGNA Corp.' 'Cincinnati Financial' 'Colgate-Palmolive' 'Comerica Inc.'
 'CME Group Inc.' 'Chipotle Mexican Grill' 'Cummins Inc.' 'CMS Energy'
 'Centene Corporation' 'CenterPoint Energy' 'Capital One Financial'
 'Cabot Oil & Gas' 'The Cooper Companies' 'CSX Corp.' 'CenturyLink Inc'
 'Cognizant Technology Solutions' 'Citrix Systems' 'CVS Health'
 'Chevron Corp.' 'Concho Resources' 'Dominion Resources' 'Delta Air Lines'
 'Du Pont (E.I.)' 'Deere & Co.' 'Discover Financial Services'
 'Quest Diagnostics' 'Danaher Corp.' 'The Walt Disney Company'
 'Discovery Communications-A' 'Discovery Communications-C'
 'Delphi Automotive' 'Digital Realty Trust' 'Dun & Bradstreet'
 'Dover Corp.' 'Dr Pepper Snapple Group' 'Duke Energy' 'DaVita Inc.'
 'Devon Energy Corp.' 'eBay Inc.' 'Ecolab Inc.' 'Consolidated Edison'
 'Equifax Inc.' "Edison Int'l" 'Eastman Chemical' 'EOG Resources'
 'Equinix' 'Equity Residential' 'EQT Corporation' 'Eversource Energy'
 'Essex Property Trust, Inc.' 'E*Trade' 'Eaton Corporation'
 'Entergy Corp.' 'Edwards Lifesciences' 'Exelon Corp.' "Expeditors Int'l"
 'Expedia Inc.' 'Extra Space Storage' 'Ford Motor' 'Fastenal Co'
 'Fortune Brands Home & Security' 'Freeport-McMoran Cp & Gld'
 'FirstEnergy Corp' 'Fidelity National Information Services' 'Fiserv Inc'
 'FLIR Systems' 'Fluor Corp.' 'Flowserve Corporation' 'FMC Corporation'
 'Federal Realty Investment Trust' 'First Solar Inc'
 'Frontier Communications' 'General Dynamics'
 'General Growth Properties Inc.' 'Gilead Sciences' 'Corning Inc.'
 'General Motors' 'Genuine Parts' 'Garmin Ltd.' 'Goodyear Tire & Rubber'
 'Grainger (W.W.) Inc.' 'Halliburton Co.' 'Hasbro Inc.'
 'Huntington Bancshares' 'HCA Holdings' 'Welltower Inc.' 'HCP Inc.'
 'Hess Corporation' 'Hartford Financial Svc.Gp.' 'Harley-Davidson'
 "Honeywell Int'l Inc." 'Hewlett Packard Enterprise' 'HP Inc.'
 'Hormel Foods Corp.' 'Henry Schein' 'Host Hotels & Resorts'
 'The Hershey Company' 'Humana Inc.' 'International Business Machines'
 'IDEXX Laboratories' 'Intl Flavors & Fragrances' 'International Paper'
 'Interpublic Group' 'Iron Mountain Incorporated'
 'Intuitive Surgical Inc.' 'Illinois Tool Works' 'Invesco Ltd.'
 'J. B. Hunt Transport Services' 'Jacobs Engineering Group'
 'Juniper Networks' 'JPMorgan Chase & Co.' 'Kimco Realty' 'Kimberly-Clark'
 'Kinder Morgan' 'Coca Cola Company' 'Kansas City Southern'
 'Leggett & Platt' 'Lennar Corp.' 'Laboratory Corp. of America Holding'
 'LKQ Corporation' 'L-3 Communications Holdings' 'Lilly (Eli) & Co.'
 'Lockheed Martin Corp.' 'Alliant Energy Corp' 'Leucadia National Corp.'
 'Southwest Airlines' 'Level 3 Communications' 'LyondellBasell'
 'Mastercard Inc.' 'Mid-America Apartments' 'Macerich' "Marriott Int'l."
 'Masco Corp.' 'Mattel Inc.' "McDonald's Corp." "Moody's Corp"
 'Mondelez International' 'MetLife Inc.' 'Mohawk Industries'
 'Mead Johnson' 'McCormick & Co.' 'Martin Marietta Materials'
 'Marsh & McLennan' '3M Company' 'Monster Beverage' 'Altria Group Inc'
 'The Mosaic Company' 'Marathon Petroleum' 'Merck & Co.'
 'Marathon Oil Corp.' 'M&T Bank Corp.' 'Mettler Toledo' 'Murphy Oil'
 'Mylan N.V.' 'Navient' 'Noble Energy Inc' 'NASDAQ OMX Group'
 'NextEra Energy' 'Newmont Mining Corp. (Hldg. Co.)' 'Netflix Inc.'
 'Newfield Exploration Co' 'Nielsen Holdings'
 'National Oilwell Varco Inc.' 'Norfolk Southern Corp.'
 'Northern Trust Corp.' 'Nucor Corp.' 'Newell Brands'
 'Realty Income Corporation' 'ONEOK' 'Omnicom Group' "O'Reilly Automotive"
```

```
 'Occidental Petroleum' "People's United Financial" 'Pitney-Bowes'
 'PACCAR Inc.' 'PG&E Corp.' 'Public Serv. Enterprise Inc.' 'PepsiCo Inc.'
 'Pfizer Inc.' 'Principal Financial Group' 'Procter & Gamble'
 'Progressive Corp.' 'Pulte Homes Inc.' 'Philip Morris International'
 'PNC Financial Services' 'Pentair Ltd.' 'Pinnacle West Capital'
 'PPG Industries' 'PPL Corp.' 'Prudential Financial' 'Phillips 66'
 'Quanta Services Inc.' 'Praxair Inc.' 'PayPal' 'Ryder System'
 'Royal Caribbean Cruises Ltd' 'Regeneron' 'Robert Half International'
 'Roper Industries' 'Range Resources Corp.' 'Republic Services Inc'
 'SCANA Corp' 'Charles Schwab Corporation' 'Spectra Energy Corp.'
 'Sealed Air' 'Sherwin-Williams' 'SL Green Realty'
 'Scripps Networks Interactive Inc.' 'Southern Co.'
 'Simon Property Group Inc' 'S&P Global, Inc.' 'Stericycle Inc'
 'Sempra Energy' 'SunTrust Banks' 'State Street Corp.'
 'Skyworks Solutions' 'Southwestern Energy' 'Synchrony Financial'
 'Stryker Corp.' 'AT&T Inc' 'Molson Coors Brewing Company'
 'Teradata Corp.' 'Tegna, Inc.' 'Torchmark Corp.'
 'Thermo Fisher Scientific' 'TripAdvisor' 'The Travelers Companies Inc.'
 'Tractor Supply Company' 'Tyson Foods' 'Tesoro Petroleum Co.'
 'Total System Services' 'Texas Instruments' 'Under Armour'
 'United Continental Holdings' 'UDR Inc' 'Universal Health Services, Inc.'
 'United Health Group Inc.' 'Unum Group' 'Union Pacific'
 'United Parcel Service' 'United Technologies' 'Varian Medical Systems'
 'Valero Energy' 'Vulcan Materials' 'Vornado Realty Trust'
 'Verisk Analytics' 'Verisign Inc.' 'Vertex Pharmaceuticals Inc'
 'Ventas Inc' 'Verizon Communications' 'Waters Corporation'
 'Wec Energy Group Inc' 'Wells Fargo' 'Whirlpool Corp.'
 'Waste Management Inc.' 'Williams Cos.' 'Western Union Co'
 'Weyerhaeuser Corp.' 'Wyndham Worldwide' 'Wynn Resorts Ltd'
 'Cimarex Energy' 'Xcel Energy Inc' 'XL Capital' 'Exxon Mobil Corp.'
 'Dentsply Sirona' 'Xerox Corp.' 'Xylem Inc.' 'Yahoo Inc.'
 'Yum! Brands Inc' 'Zimmer Biomet Holdings' 'Zions Bancorp' 'Zoetis']

In cluster 5, the following companies are present:
['Alliance Data Systems']

In cluster 2, the following companies are present:
['Apache Corporation' 'Chesapeake Energy']

In cluster 1, the following companies are present:
['Bank of America Corp' 'Intel Corp.']

In cluster 3, the following companies are present:
['Facebook']

In cluster 4, the following companies are present:
['Priceline.com Inc']
```

- While this project identifies financial and structural patterns across clusters, a sector-level interpretation of individual securities may benefit from review by industry analysts or portfolio specialists with deeper domain knowledge.

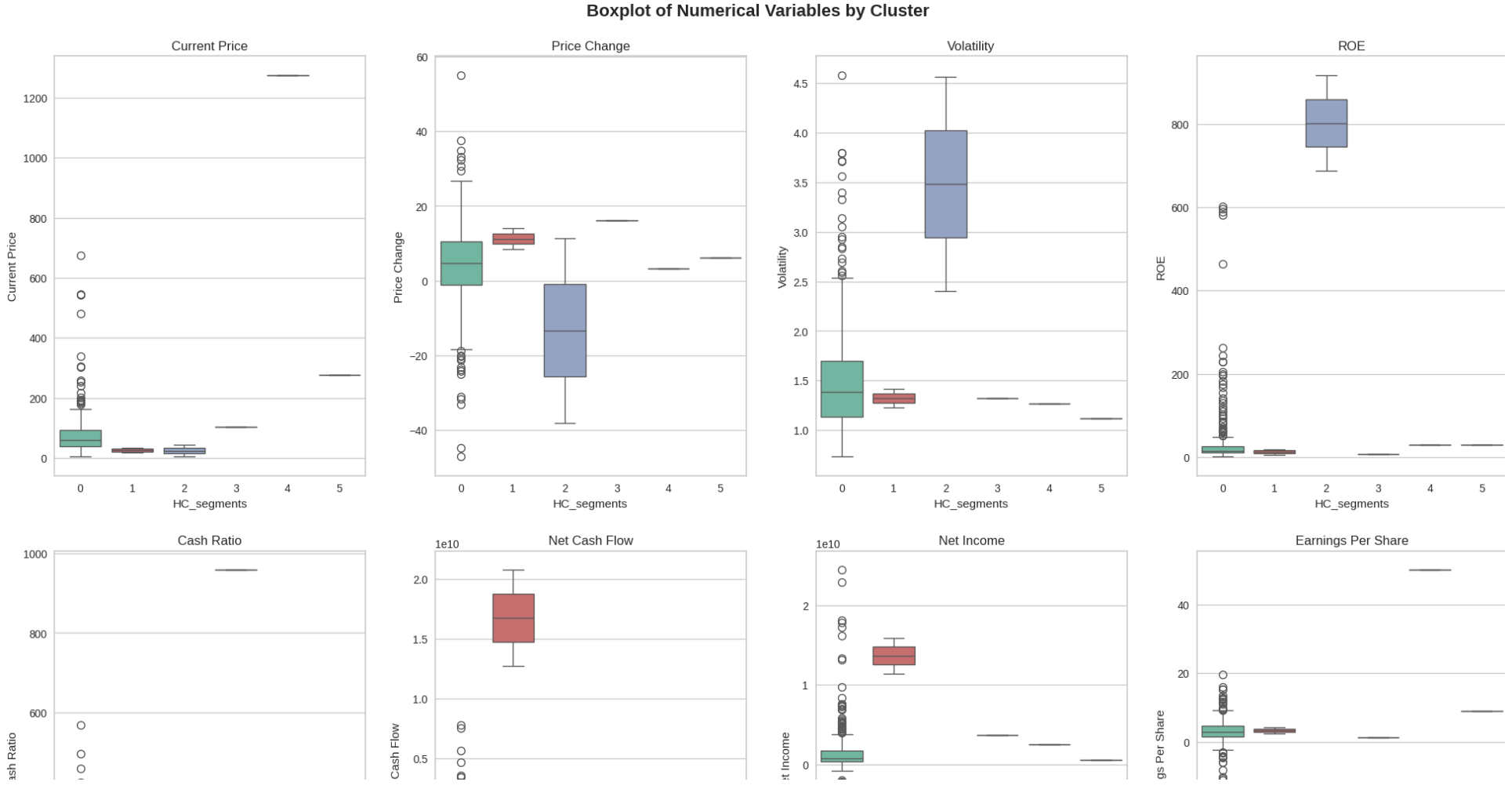
```
# Count of securities per sector within each HC segment
df2.groupby(["HC_segments", "GICS Sector"])["Security"].count()
```

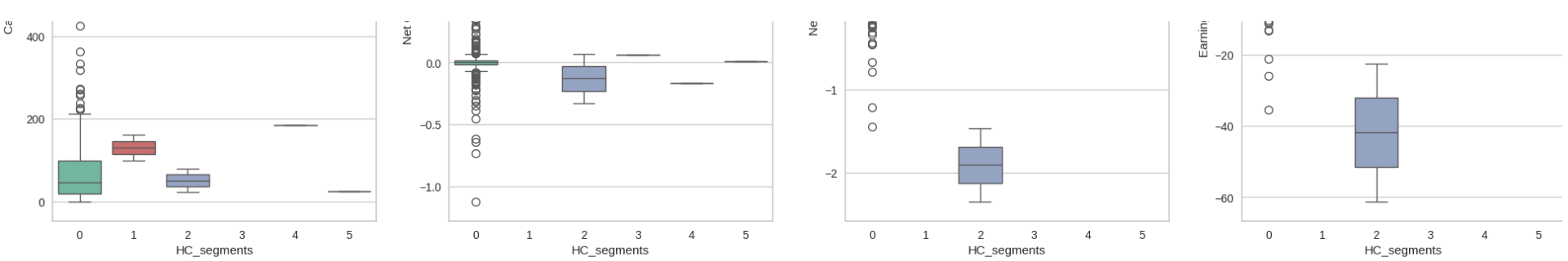
		Security
HC_segments	GICS Sector	
0	Consumer Discretionary	39
	Consumer Staples	19
	Energy	28
	Financials	48
	Health Care	40
	Industrials	53
	Information Technology	30
	Materials	20
	Real Estate	27
	Telecommunications Services	5
	Utilities	24
1	Financials	1
	Information Technology	1
2	Energy	2
3	Information Technology	1
4	Consumer Discretionary	1
5	Information Technology	1

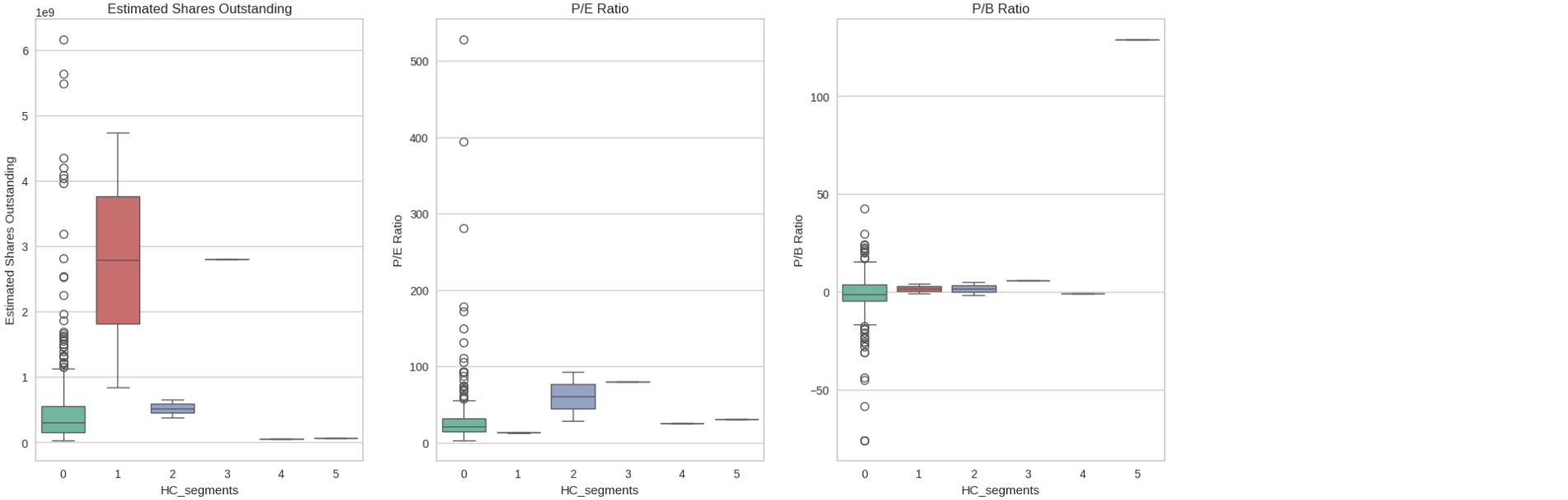
dtype: int64

- Cluster 0 contains the majority of securities across all sectors, reflecting broad market representation.
- Clusters 1–3 are sparse, each holding only 1–2 stocks from sectors like IT, Energy, or Financials.
- Cluster 4 includes a single Consumer Discretionary stock, consistent with its outlier profile.
- Cluster 5 holds one IT stock, suggesting a unique deviation from its sector peers.

```
# Boxplot of Numerical Variables by HC Cluster
plt.figure(figsize=(20, 20))
plt.suptitle("Boxplot of Numerical Variables by Cluster", fontsize=16, fontweight='bold')

box_colors = ElleSet

# Select numerical columns excluding cluster label
num_col = df2.select_dtypes(include=np.number).drop(columns=["HC_segments"]).columns.tolist()

for i, variable in enumerate(num_col):
    plt.subplot(3, 4, i + 1)
    sns.boxplot(
        data=df2,
        x="HC_segments",
        y=variable,
        palette=box_colors
    )
    plt.title(variable, fontsize=12)

plt.tight_layout(pad=2.0)
```







- The stock in Cluster 4 stands out sharply, with the highest Current Price and Earnings Per Share, likely representing a premium-priced outlier with distinct market behavior.
- Cluster 2 shows the most concerning disconnect, with negative EPS despite extremely high ROE, suggesting potentially leveraged or non-operational drivers of return.
- The diverse composition of Cluster 0, which includes the most outliers and widest spread across key metrics, points to a catch-all group of mixed financial profiles.
- Companies in Cluster 3 combine strong Net Cash Flow and Net Income with moderate pricing, which may suggest value opportunities based on cash strength and moderate pricing, though further financial review would be warranted.

KMeans vs Hierarchical Clustering

Key Insights

- Execution Time: KMeans completed significantly faster than Hierarchical Clustering. Hierarchical clustering required more computation due to its distance-based structure, making KMeans better suited for real-time or scaled deployment in large equity portfolios.
- Number of Clusters: Both methods arrived at six clusters, but the structure and composition of those clusters varied. KMeans created more balanced group sizes, while Hierarchical Clustering produced several small, outlier-driven segments alongside one large, mixed cluster.
- Cluster Distribution:
  - KMeans spread securities more evenly, Cluster 0 shows the broadest sector representation, while Cluster 3 includes multiple sectors but with notable concentration in Financials and Industrials.
  - Hierarchical Clustering resulted in a dominant Cluster 0 that absorbed most securities, with Clusters 2, 4, and 5 acting as smaller niche or outlier groups.
  - Clusters 1 and 5 in both methods contained very few securities, often isolated by extreme financial traits or narrow sector representation. While potentially meaningful, further analysis may require deeper financial review on a case-by-case basis.
- Profile Similarities: Both models identified:
  - A segment with strong fundamentals and moderate pricing (potential value plays)
  - A high-performing outlier stock (seen in Cluster 4 of HC and small clusters in KM)

A diverse, general-market segment (Cluster 0 in both cases)

- Profile Differences:
  - Hierarchical Clustering surfaced a distinct segment (Cluster 2) with high ROE and negative EPS, highlighting possible financial anomalies. This particular anomaly was less distinctly isolated by KMeans.
  - KMeans more clearly isolated energy-heavy clusters (e.g., Cluster 1 and 2), which may be useful when constructing sector-focused portfolios.
- Observation Overlap: While exact observation overlap between clusters varies due to the different clustering logic, broad patterns—such as high performers and general performers—were consistently grouped.

Final Perspective

KMeans offers efficiency and cleaner segment balance, making it well-suited for ongoing segmentation tasks in dynamic financial environments. In contrast, Hierarchical Clustering, though slower, provides a useful exploratory lens—especially in surfacing potential outliers or irregular financial profiles. A combined approach may offer the most robust foundation for segment-driven investment strategies.

Actionable Insights and Recommendations

Business Objective

Trade & Ahead seeks to identify distinct investment groupings within a diverse portfolio of securities by leveraging unsupervised machine learning techniques. The goal is to enhance investment strategy development and portfolio monitoring by uncovering patterns of similarity in return behavior and market risk. Through clustering analysis, this initiative enables analysts to recognize underlying asset behaviors, potentially revealing opportunities for diversification, risk management, or sector rotation—all without relying on labeled data.

Key Findings

- Six behavior-based clusters were identified using both KMeans and Hierarchical Clustering, grouping stocks by patterns in volatility, return, valuation, and liquidity metrics.
- Cluster profiles revealed meaningful differentiation, including high-performing outliers, stable income-oriented stocks, and segments with inconsistent or high-risk financial structures.
- KMeans offered faster execution and cleaner segment balance, making it well-suited for ongoing portfolio monitoring. Hierarchical Clustering, while slower, enhanced interpretability and outlier detection.
- Sector analysis confirmed the breadth of coverage, with diverse clusters spanning multiple GICS sectors. Some segments were tightly concentrated (e.g., energy-heavy or outlier-dominated), while others reflected broad market exposure.
- Cross-method consistency in behavioral themes (e.g., growth, value, volatility) reinforced the robustness of the findings and offered confidence in using clustering for portfolio insight.

Actionable Insights

1. Segment-Aware Monitoring Framework

   Establish custom monitoring dashboards for each cluster group. For example:

- High-risk/high-return groups can trigger closer monitoring and scenario-based stress testing.
- Stable-return groups may be ideal for baseline or income portfolios.

2. Diversification Guidance

   Use cluster membership to flag concentration risk. Holdings concentrated in a single cluster may lack sufficient diversification, especially in volatile market phases.

   Suggest potential additions from underrepresented clusters to balance exposure.

3. Strategy Development Support

   Support thematic strategy creation by mapping cluster behaviors to macroeconomic or sectoral trends.

   Leverage high-momentum clusters for growth-oriented strategies, and lower-volatility segments for conservative or defensive portfolios.

4. Analyst Collaboration Prompt

   While machine learning helps organize the portfolio by behavior, deeper insight into cluster content (e.g., industry role, market sentiment drivers) may require input from sector specialists.

   Encourage domain expert review for qualitative labeling or refinement of clusters for actionable use.

Additional Recommendations

- Model Recalibration: Re-run clustering at regular intervals or during major market shifts to ensure segments continue to reflect evolving stock behavior.
- Portfolio Reporting: Use cluster-level summaries to provide context around performance patterns and group-level risk dynamics in internal reports.
- Feature Expansion: Explore incorporating macroeconomic or sentiment-based variables to improve cluster interpretability and align segments with broader market drivers.

Together, these insights and actionable strategies form a data-driven foundation for smarter portfolio construction, enhanced risk management, and more informed allocation decisions in a dynamic market environment.
