Trade&Ahead

Eugenie Seholm

Contents

Business Problem Overview

Problem:

- Diversified stock portfolio is important
  - Higher returns & lower risk
- Determining the worth of a stock can be overwhelming, and much more for a multitude of stocks

Objective:

- Analyze the data
- Group the stocks based on the attributes provided
- Share insights about the characteristics of each group.

Business Problem Overview

Financial Implications:

- Cluster analysis can identify stocks that exhibit similar characteristics and those that exhibit minimum correlation.
- This will help investors better analyze stocks across different market segments and help protect against risks that could make the portfolio vulnerable to losses.

Solution Approach

- Data preparation
- K-Means Clustering
  - Elbow Method
  - Silhouette Score/Coefficients
- Hierarchical Clustering
  - Explore Distance Metrics, Linkage Method
  - Dendograms
- Compare Cluster Profiles

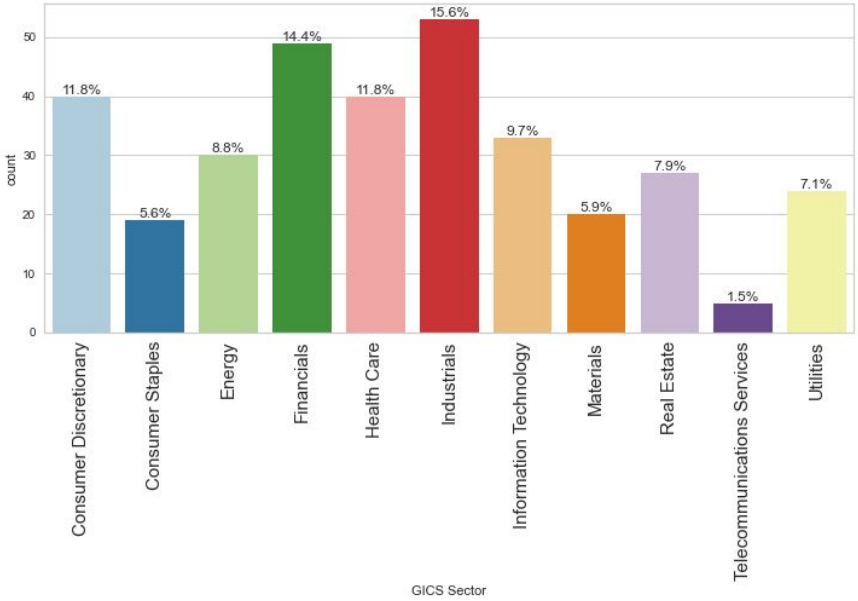
Data Overview

- 340 observations (rows)
- 15 variables
  - **4 categorical**
  - **11 numerical**
- 0 duplicated rows
- 0 missing entries
- Fix column name
  - “Security” → “Company”
- Outlier values assumed to be authentic
- Data scaled for clustering

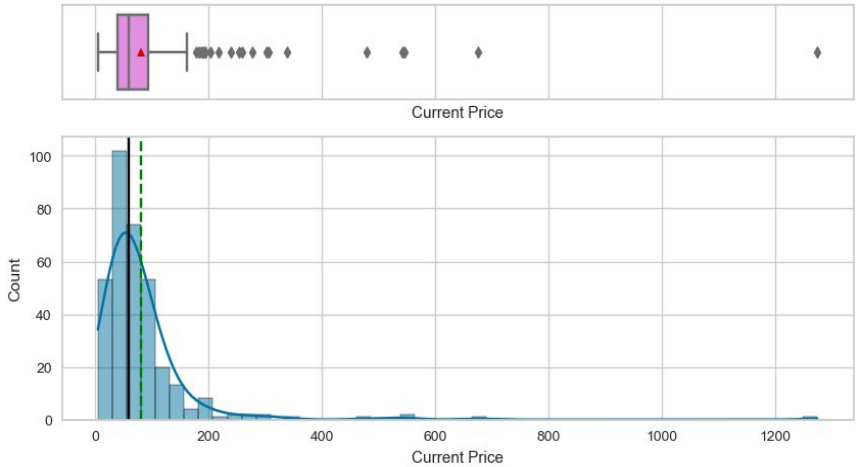
Variable	Description
Ticker Symbol	An abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market
Company	Name of the company
GICS Sector	The specific economic sector assigned to a company by the Global Industry Classification Standard (GICS) that best defines its business operations
GICS Sub Industry	The specific sub-industry group assigned to a company by the Global Industry Classification Standard (GICS) that best defines its business operations
Current Price	Current stock price in dollars
Price Change	Percentage change in the stock price in 13 weeks
Volatility	Standard deviation of the stock price over the past 13 weeks
ROE	A measure of financial performance calculated by dividing net income by shareholders' equity
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 of per share

Exploratory Data Analysis (EDA): Categorical Variables

- Ticker Symbol all unique values
- Company names all unique values
- GICS Sub Industry 104 unique values
  - Oil & Gas Exploration & Production, Industrial Conglomerates, and REITs are the 3 most popular
- GICS Sector
  - Top 3: Industrials, Financials, Consumer Discretionary

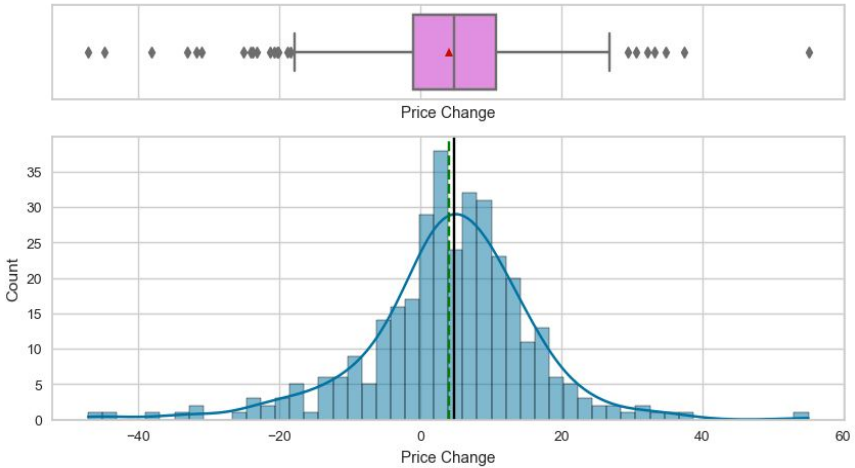


EDA: Current Price & Price Change



Current Price

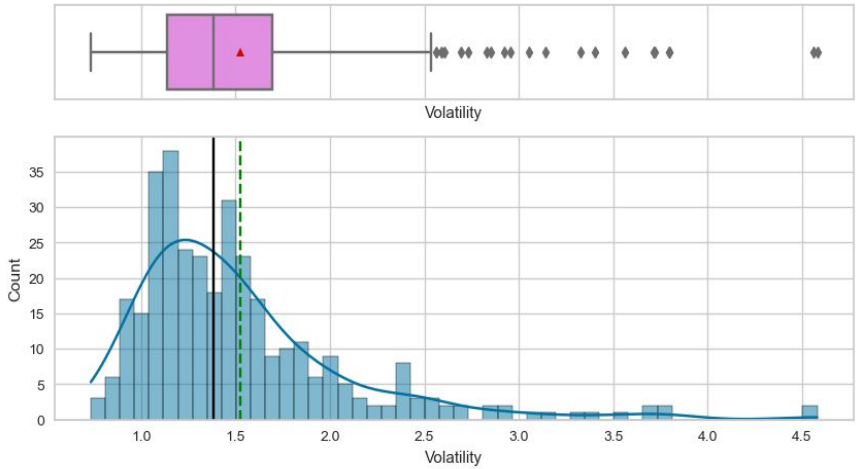
- Extreme values in positive direction
- Positive skew



Price Change

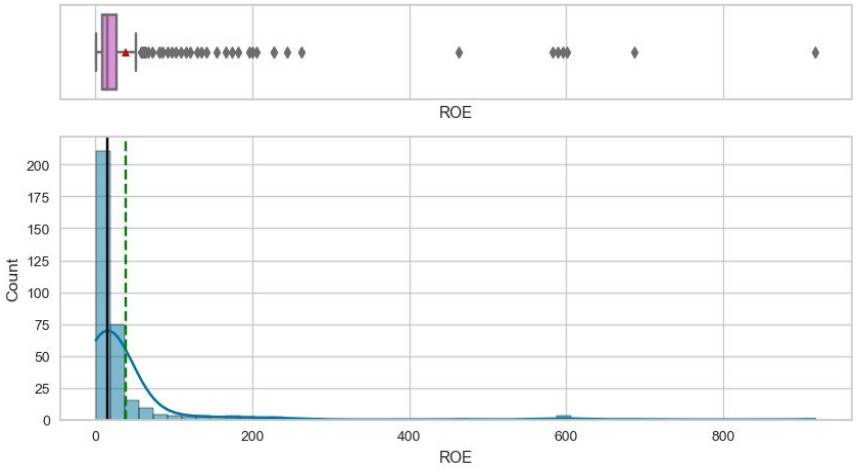
- Extreme values in both directions
- Approximately normal looking distribution

EDA: Volatility & ROE



Volatility

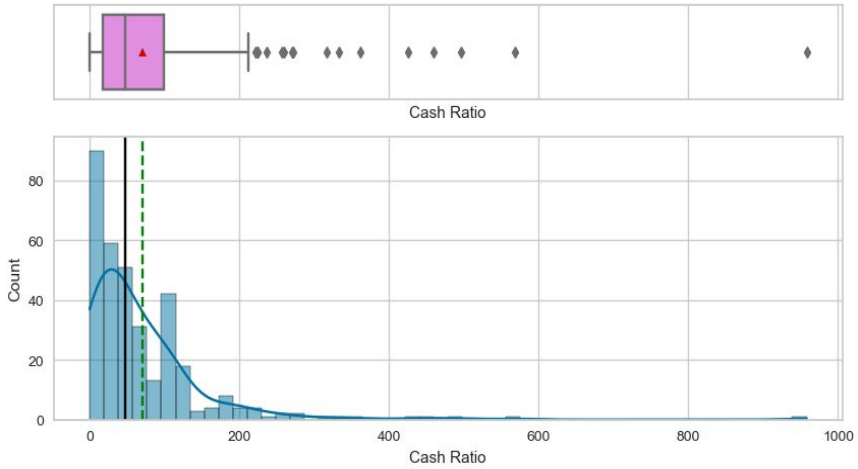
- Extreme values in positive direction
- Positive skew



ROE

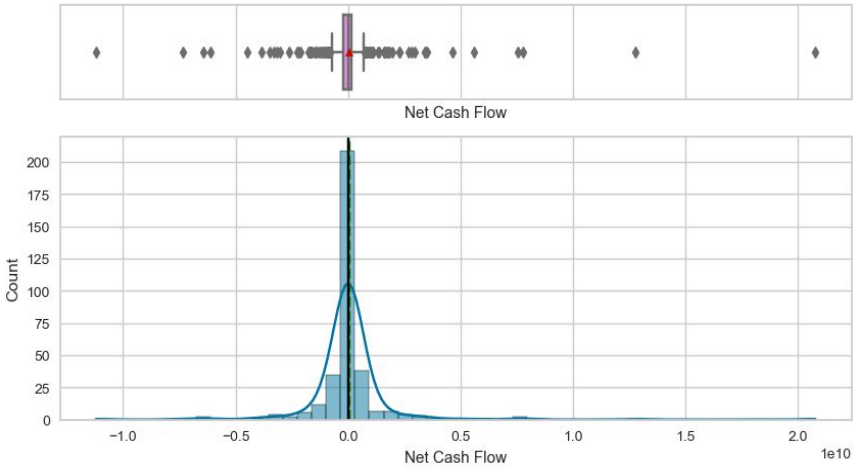
- Extreme values in positive direction
- Positive skew

EDA: Cash Ratio & Net Cash Flow



Cash Ratio

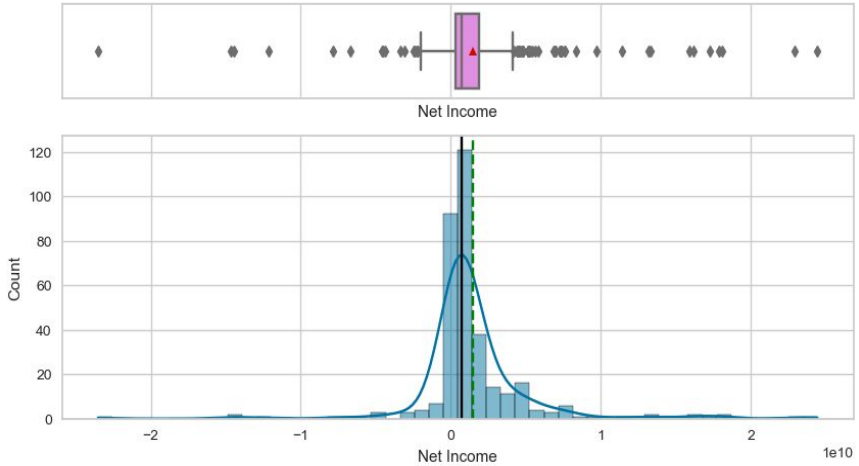
- Extreme values in positive direction
- Positive skew



Net Cash Flow

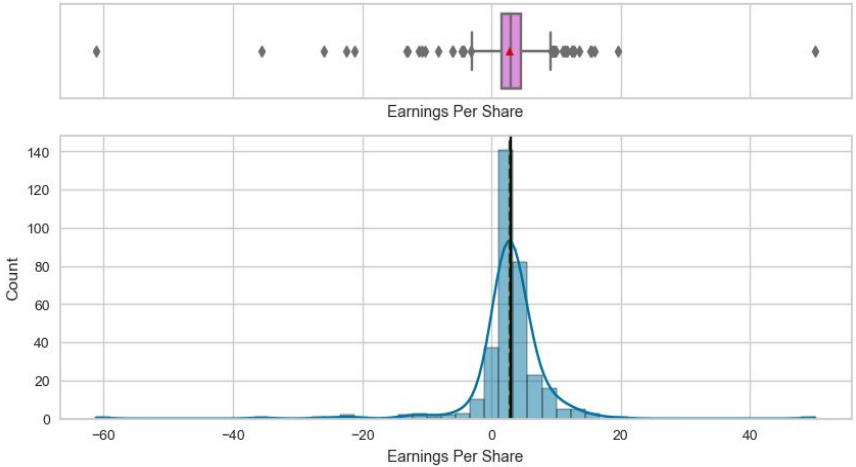
- Extreme values in both directions

EDA: Net Income & Earnings Per Share



Net Income

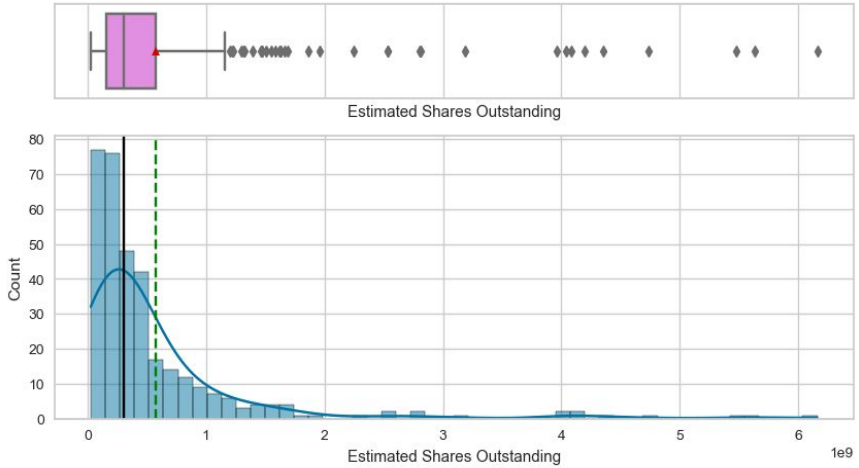
- Extreme values in both directions



Earnings Per Share

- Extreme values in both directions

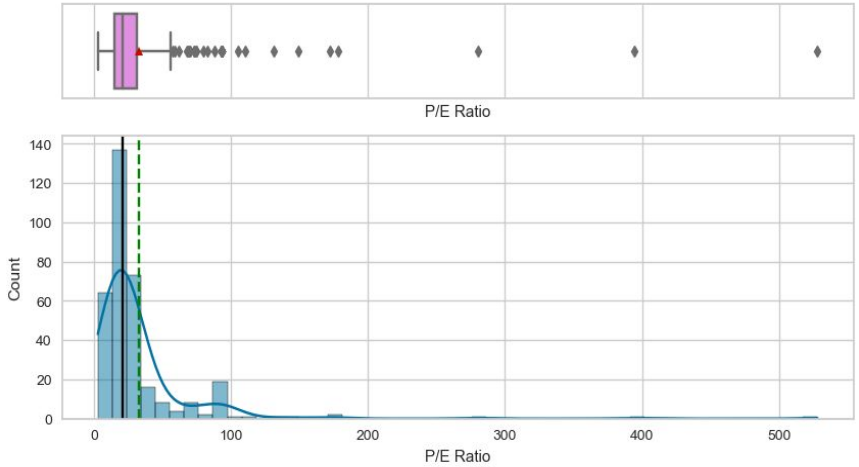
EDA: Estimated Shares Outstanding



Estimated Shares Outstanding

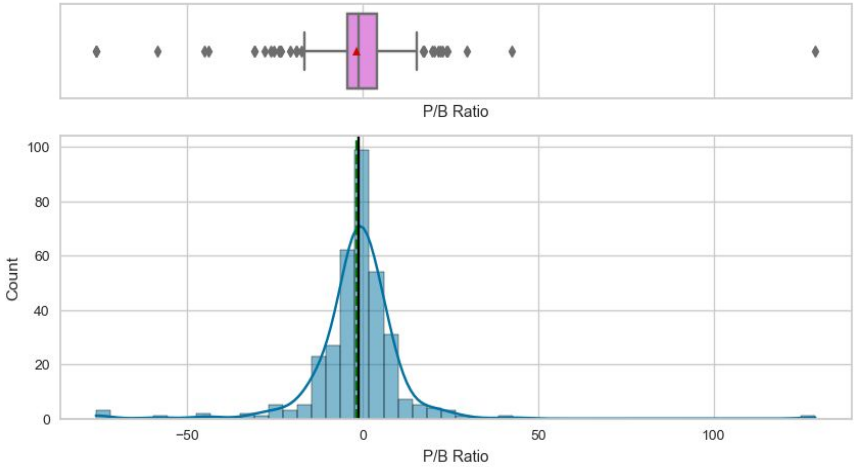
- Extreme values in the positive direction
- Positive skew

EDA: P/E Ratio & P/B Ratio



P/E Ratio

- Extreme values in positive direction
- Positive skew

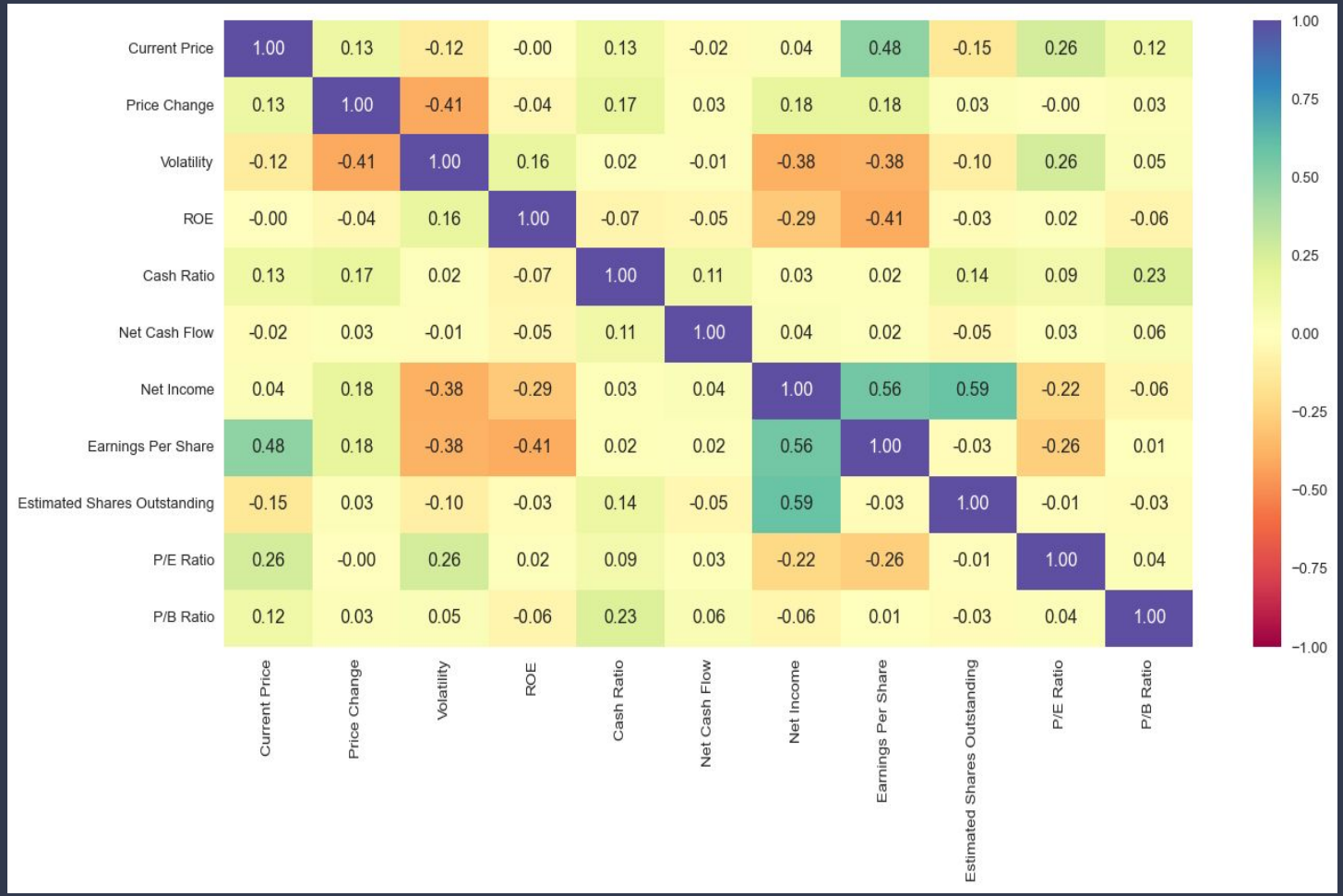


P/B Ratio

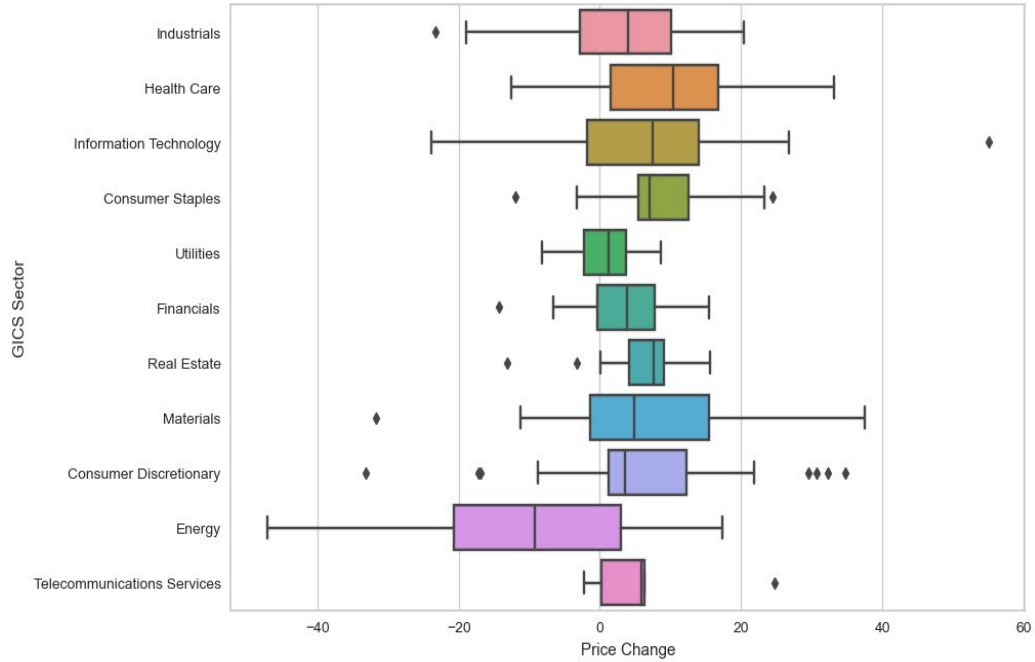
- Extreme values in both directions

EDA: Correlation Heatmap

- No strongly correlated variables
- Net Income has moderate positive correlation with Earnings Per Share ^ Estimated Shares Outstanding
- Earnings Per Share has moderate positive correlation with Current Price
- Volatility and Price Change have weak to moderate negative correlation
- ROE and Earnings Per Share have a weak to moderate negative correlation

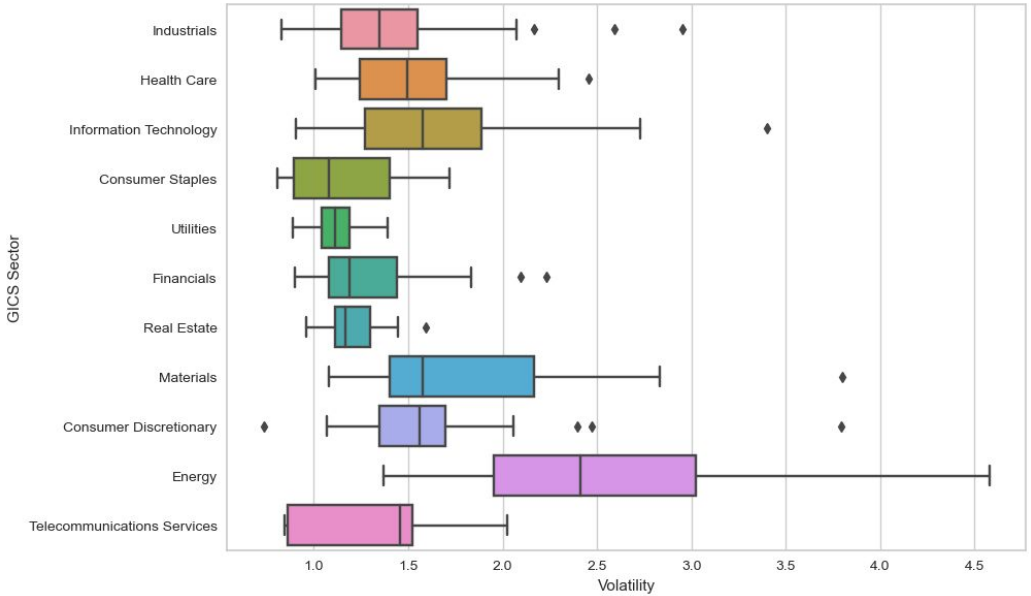


EDA: GICS Sector vs Price Change



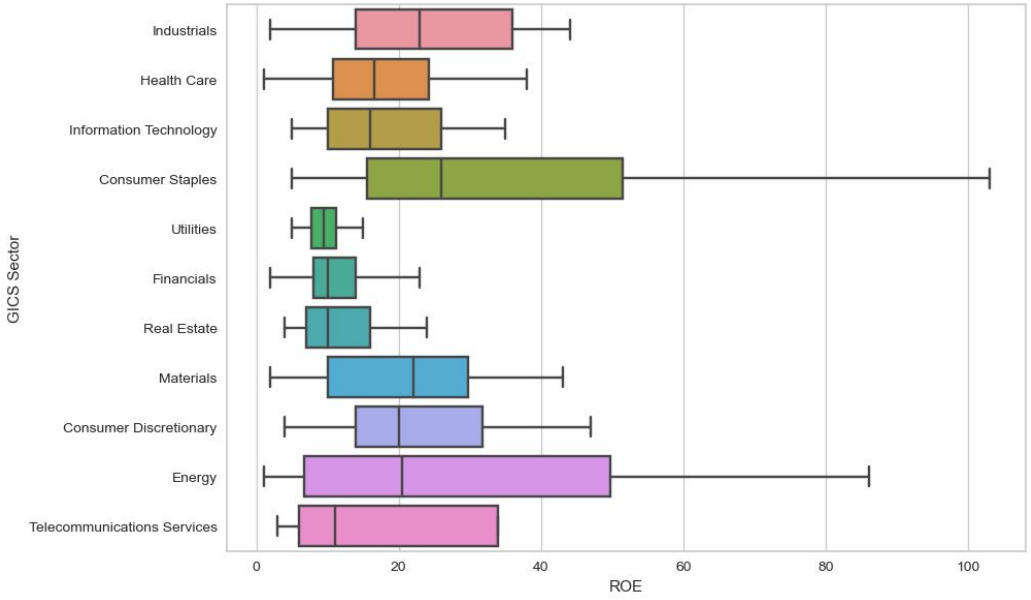
- Energy is the only sector that has mostly negative values and has the minimum value
- Information Technology has maximum value

EDA: GICS Sector vs Volatility



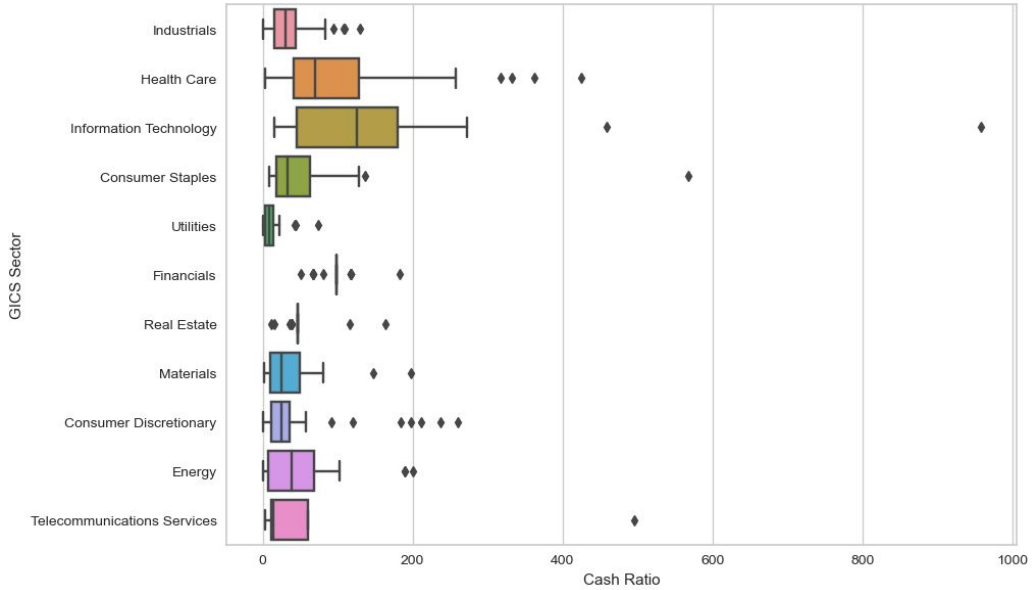
- Energy has the highest values (including maximum)
- Utilities appears to have the lowest overall values

EDA: GICS Sector vs ROE



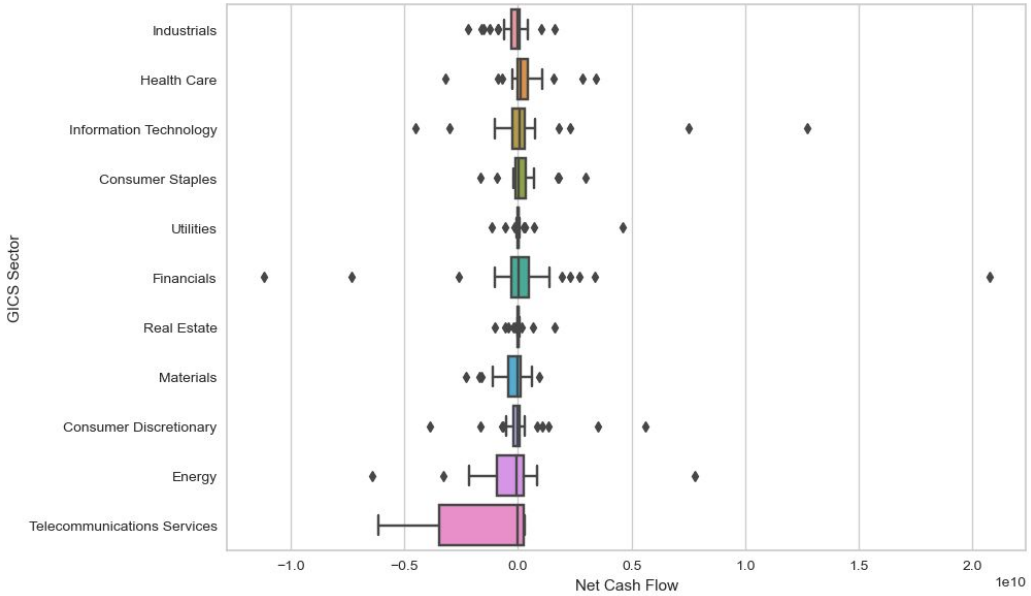
- Consumer Staples appears to have the highest overall values
- Utilities has the lowest overall values
- Telecommunications Services has the lowest median value

EDA: GICS Sector vs Cash Ratio



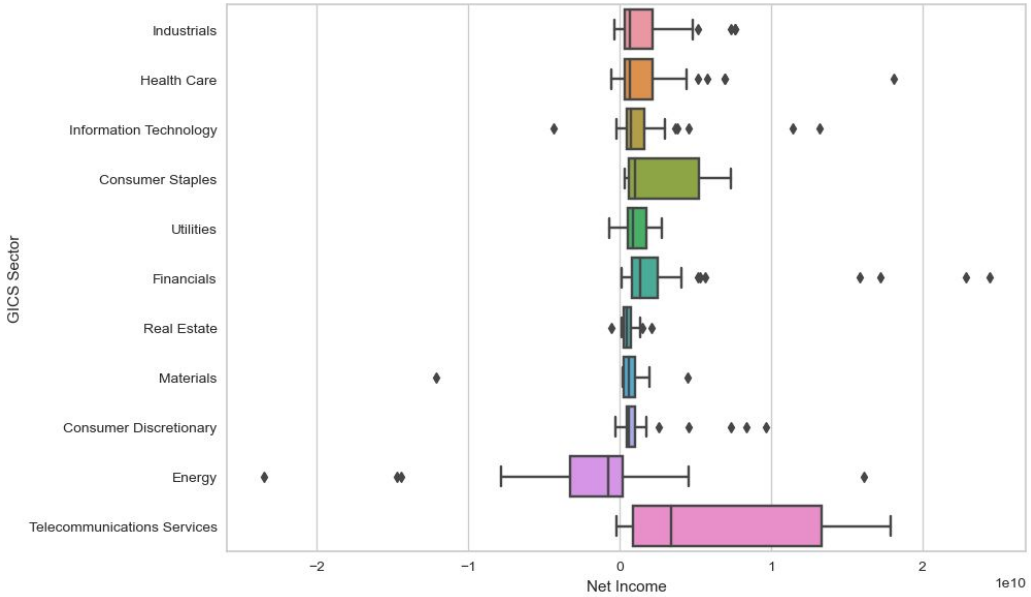
- Information Technology has the maximum and highest median values
- Utilities has overall lowest values

EDA: GICS Sector vs Net Cash Flow



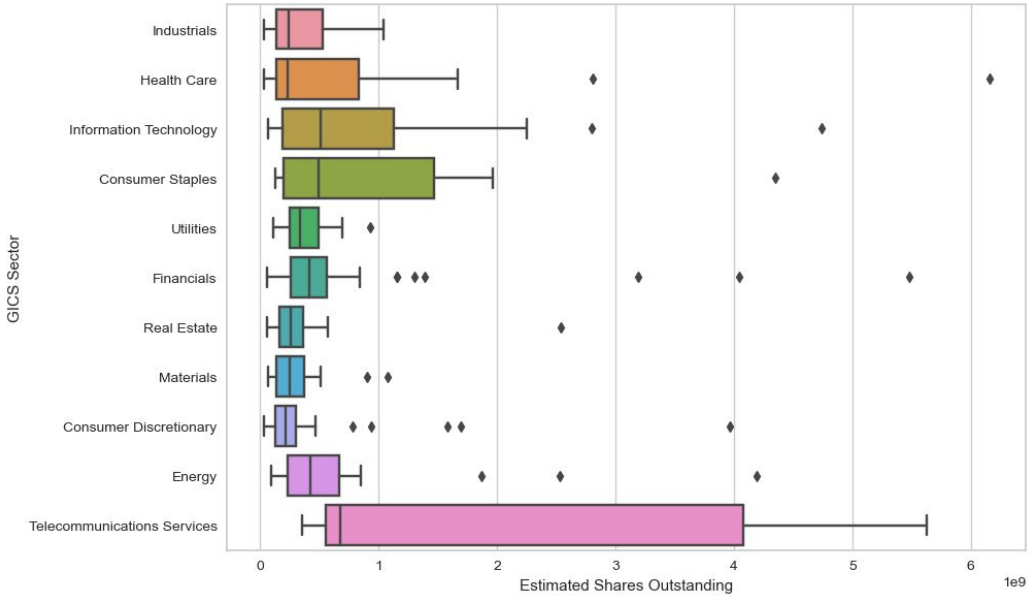
- Financials has the most extreme values in both negative and positive directions
- Telecommunications Services has the biggest spread, aside from extreme values

EDA: GICS Sector vs Net Income



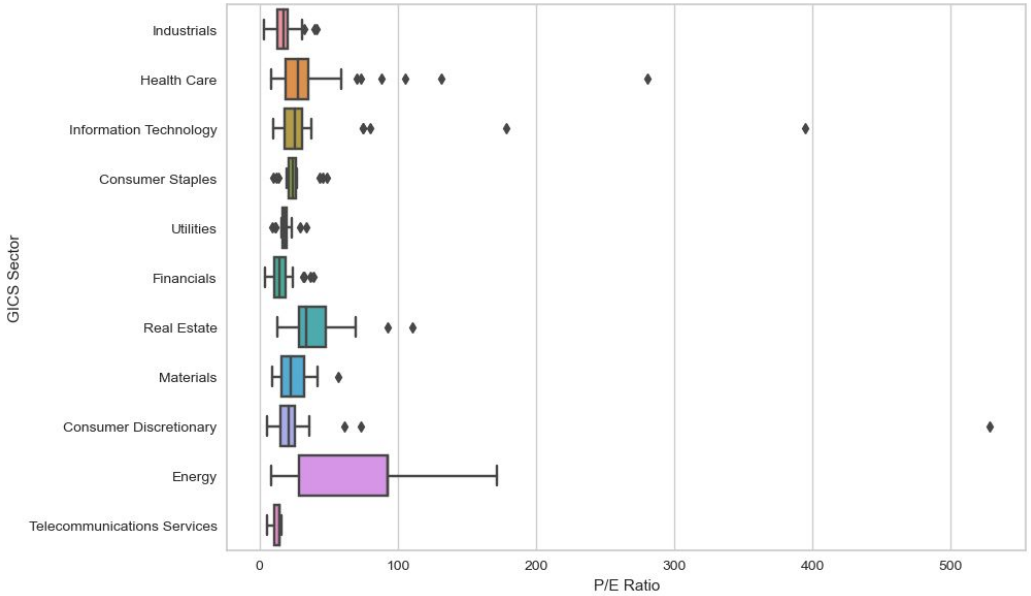
- Energy has the minimum value, lowest overall values, as well as mostly negative values
- Financials has maximum value, but majority of the sector is under 0.5 net income
- Telecommunications Services has largest spread, aside from extreme values, and is mostly positive

EDA: GICS Sector vs Estimated Shares Outstanding



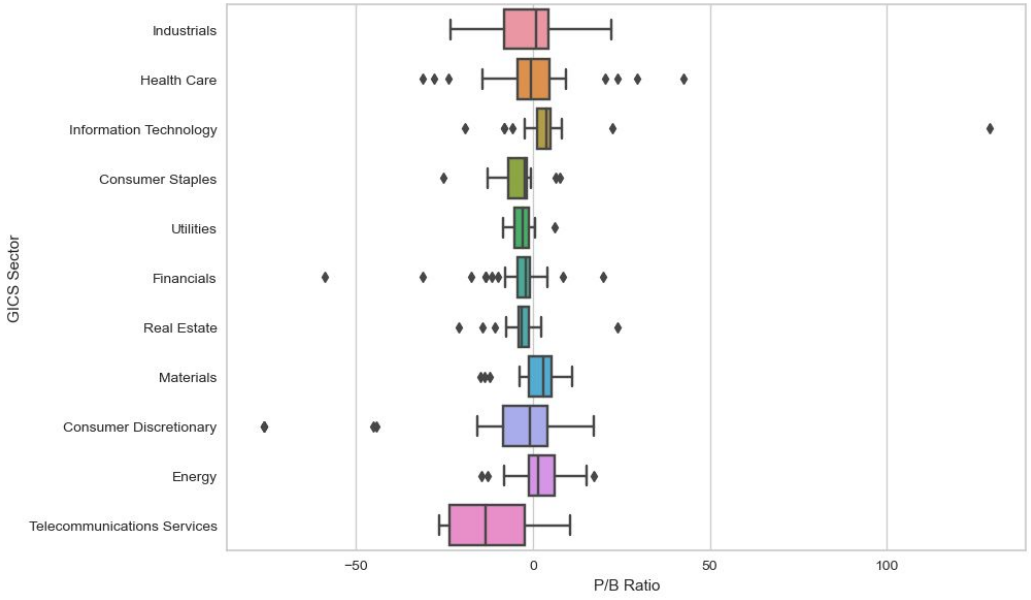
- Telecommunications Services has highest median value and largest spread, aside from extreme values
- Health care has maximum value but most values are much lower

EDA: GICS Sector vs P/E Ratio



- Consumer Discretionary has maximum value but most values are not very high
- Telecommunications Services has very low overall values
- Energy has largest spread, aside from extreme values

EDA: GICS Sector vs P/B Ratio



- Information Technology has the largest median value and max value, although most values are close to the median and positive
- Consumer Staples, Utilities, Financials, Real Estate, and Telecommunications Services have a majority of negative values

EDA: Key Findings and Insights

Many sectors have extreme values for different numerical variables

Numerical variables are not strongly correlated

Energy sector has many negative price changes, high volatility, negative net cash flow, and negative net income

Information Technology has high positive price change, high net cash ratio, and high P/B ratio

Utilities has low volatility, low ROE, low cash ratio, and mostly negative P/B ratio

Consumer Staples has high ROE and mostly negative P/B ratio

Telecommunication Services has low ROE; large spread of net cash flow, net income, estimated shares outstanding; low P/E ratio; and mostly negative P/B ratio

Financials has some of the most extreme values (net cash flow, net income) and mostly negative P/B ratio

Health care has maximum estimated shares outstanding value but most values are much lower

Consumer Discretionary has maximum P/E ratio but most values are farly low

Real Estate has mostly negative P/B ratio

Model Overview and Performance Summary

K-Means Clustering

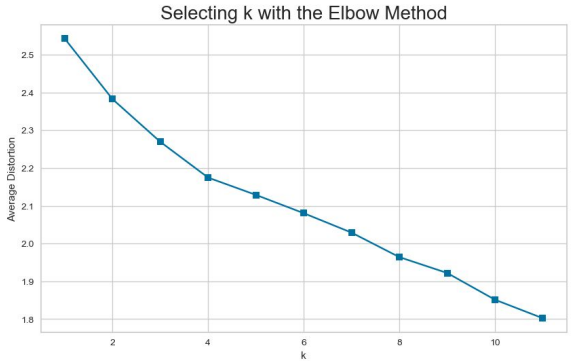
versus

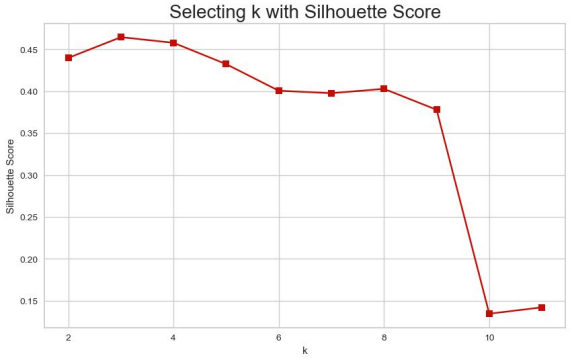
Hierarchical Clustering

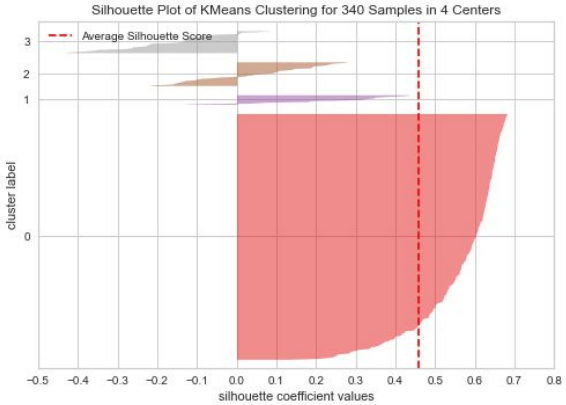
K-Means Clustering

- Fit scaled data to K-Means model
- Use Elbow Method to determine appropriate *k*
- Silhouette Score for choosing k
- Silhouette coefficients for choosing k

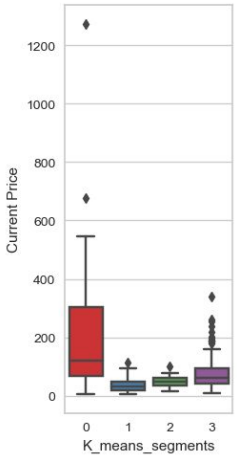
- Choose k=4
- Silhouette score approximately 0.458

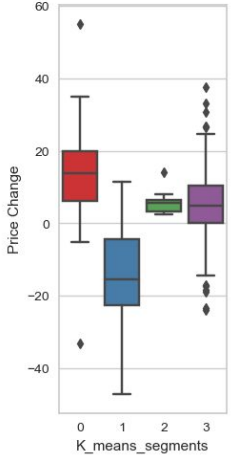


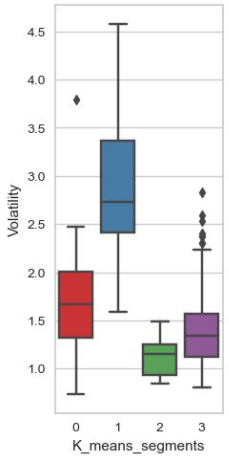


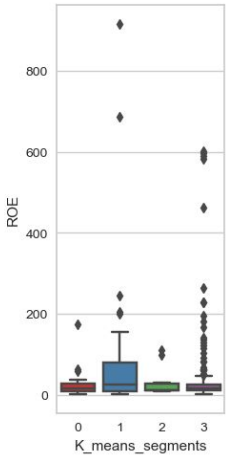


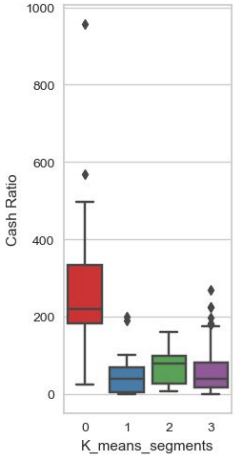
K-Means Clustering

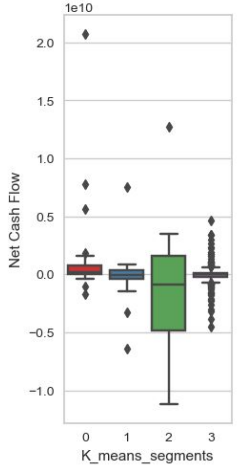


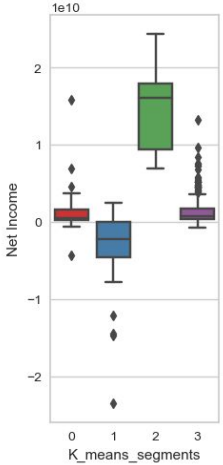


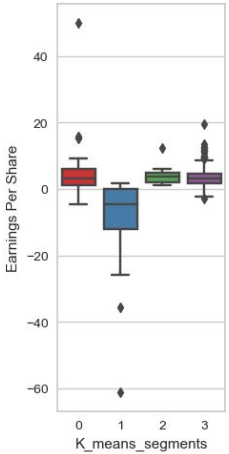


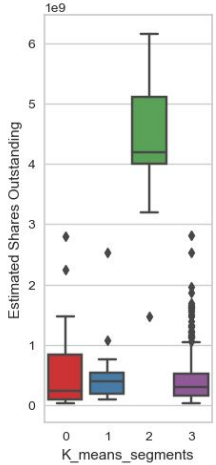


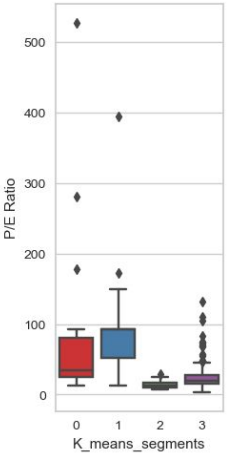


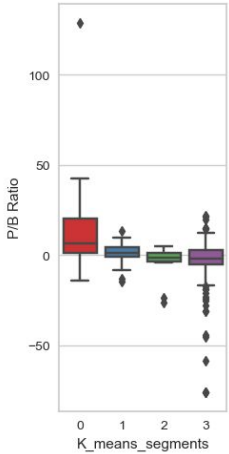






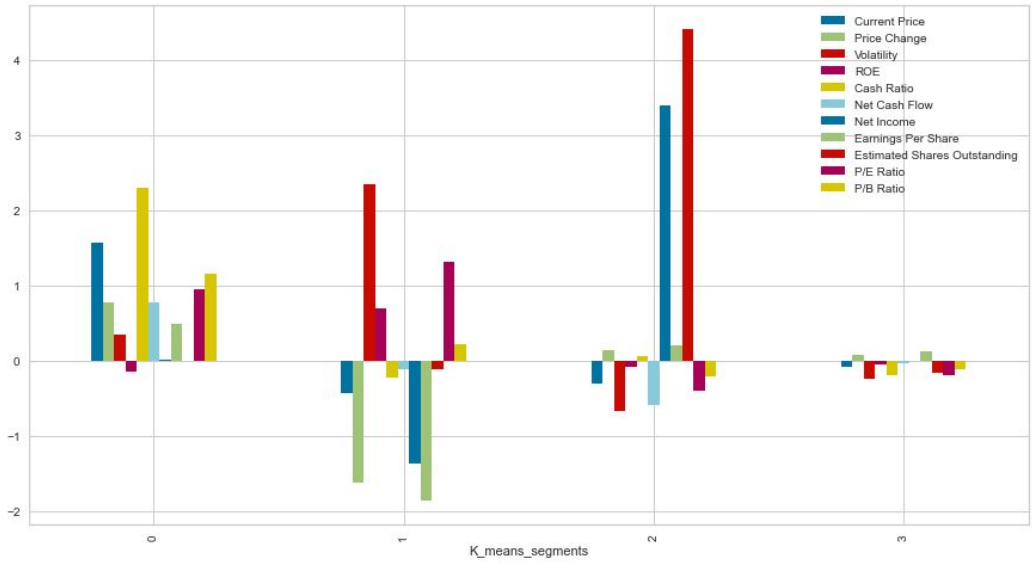






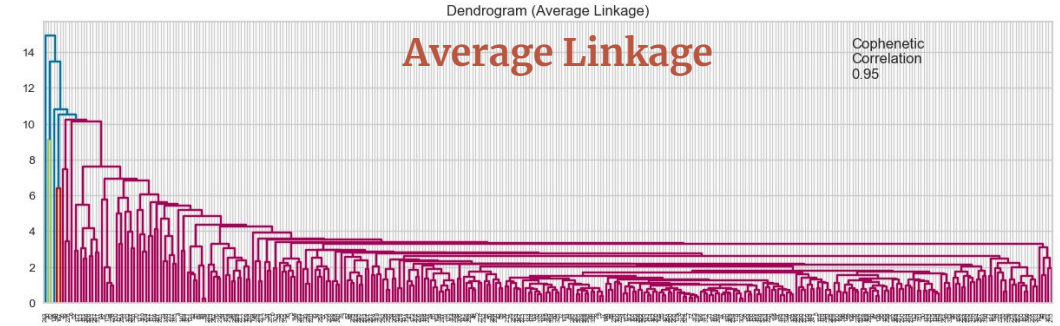
K-Means Clustering

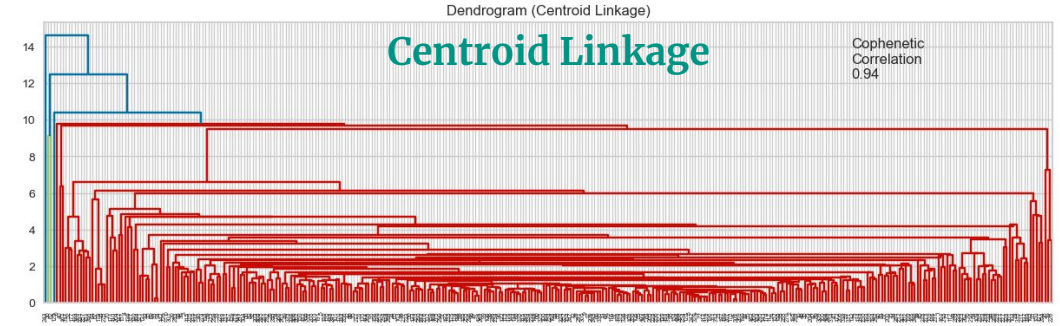
- Cluster 0 (25 companies)
  - High current price, price change, cash ratio, P/B ratio
- Cluster 1 (27 companies)
  - Low current price, price change, net income, earnings per share
  - Moderately high POE, P/E ratio
  - High volatility
- Cluster 2 (11 companies)
  - Low volatility, net cash flow P/E ratio
  - High net income, estimated shares outstanding
- Cluster 3 (277 companies)
  - Most variables close to zero in value

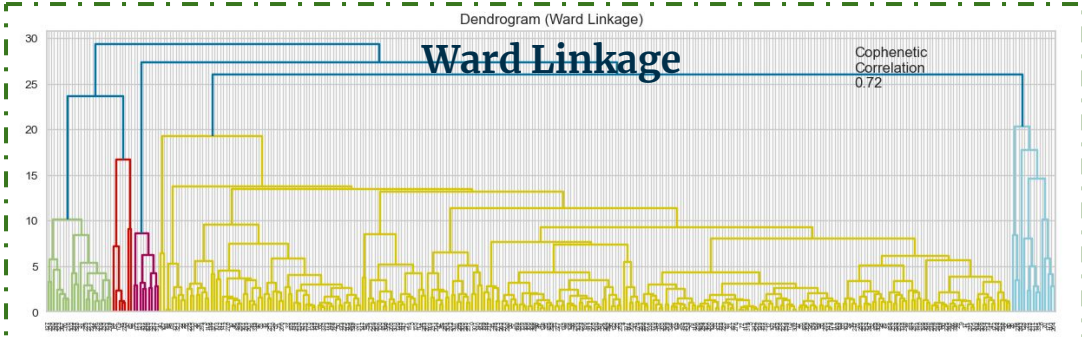


Hierarchical Clustering

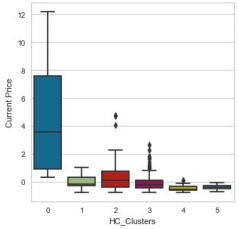
- Explore with different distance metrics
- Explore with different linkage methods
- Use cophenetic correlation to select appropriate distance metric and linkage method
- Examine dendograms of different linkage methods
  - Average has highest cophenetic correlation, followed by centroid
  - **Ward linkage** has more distinct and separated clusters. Cophenetic correlation: 0.72

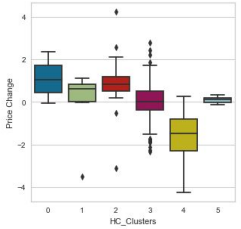


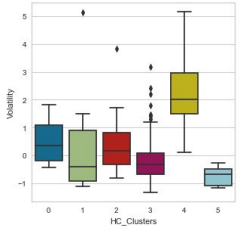


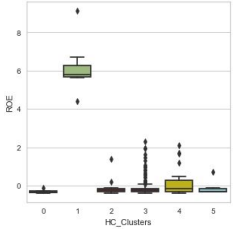


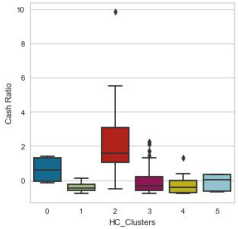
Hierarchical Clustering

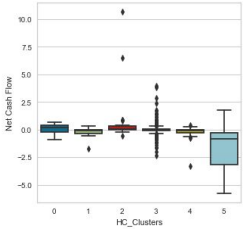


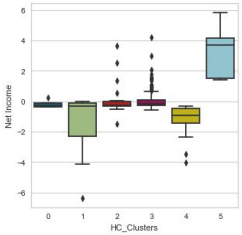


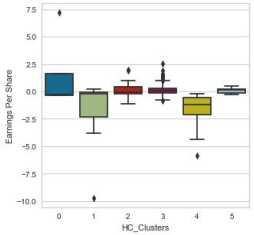


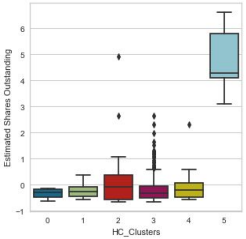


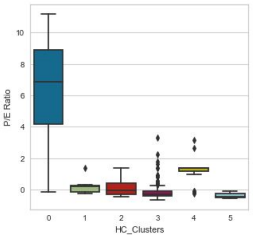


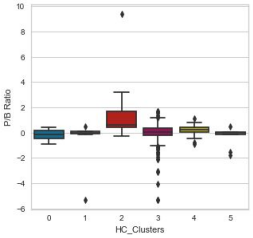






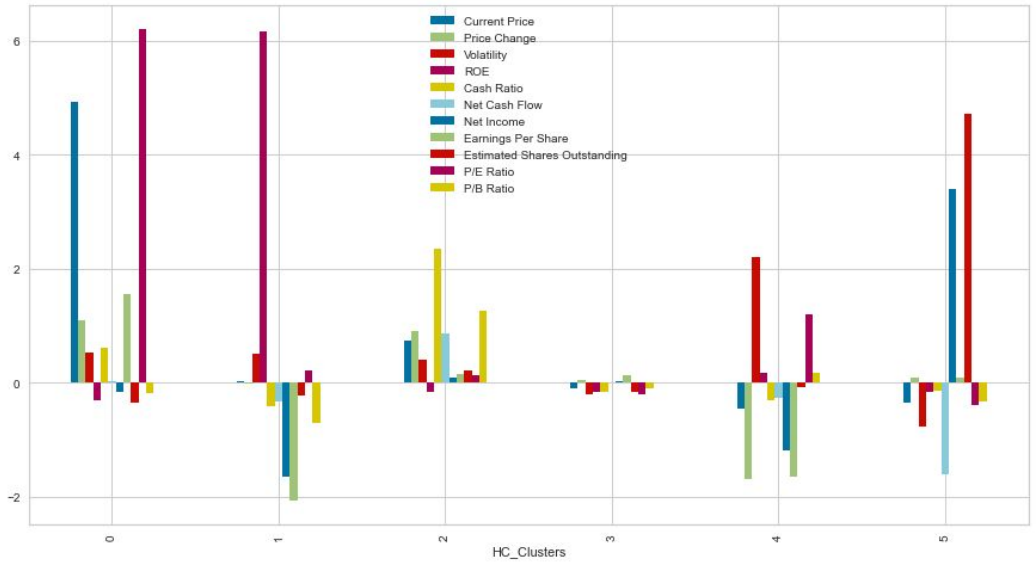






Hierarchical Clustering

- Cluster 0 (4 companies)
  - Moderate to high cash ratio, earnings per share
  - High current price, P/E ratio
- Cluster 1 (7 companies)
  - High ROE
- Cluster 2 (23 companies)
  - High cash ratio, P/B ratio
- Cluster 3 (275 companies)
  - Most variables close to zero in value
- Cluster 4 (22 companies)
  - Low current price, price change (negative), net income (negative), earnings per share (negative)
  - Moderate P/E ratio
  - High volatility
- Cluster 5 (9 companies)
  - Low volatility
  - Somewhat low net cash flow
  - High net income, estimated shares outstanding



K-Means vs Hierarchical Clustering

Largest cluster is approximately same size and has values close to zero across all variables

Hierarchical cluster appears to have slightly more information

K-Means

- Fewer, larger clusters
- Identified 27 companies with high volatility
- Identified 2 more companies than hierarchical clustering with low volatility, low net cash flow, high net income, high estimated shares outstanding
- 277 companies in largest cluster

Hierarchical

- More clusters, but smaller sizes
- Identified 22 companies with high volatility
- Identified somewhat similar companies in a cluster with high cash ratio and high P/B ratio but separated out companies with high current price and high P/E ratio
- 275 companies in largest cluster

Business Recommendations & Conclusions

- Gather information from each individual investor as stock recommendations need to be tailored to each individual
  - Risk tolerance?
  - Specific goals?
  - Short/long term investing?
  - Capital available for investing?
  - Etc.
- Pair clustering model with EDA observations to choose the companies that will work best for the individual investor
  - E.g., if investor is looking for low volatility stock, choose the cluster with low volatility and compare estimated shares outstanding for those companies with their sector averages to give a better indication of long-term patterns
- Recommend a *diversified portfolio* based on their profile, choosing stocks that fit their profile across different sectors

Business Recommendations & Conclusions

- More data needs to be collected for the clustering model
  - E.g., information regarding dividends is valuable since some people *only* invest in dividend stock
  - Debt-to-equity, debt-to-assets, EBITDA, etc. would additionally be valuable information for a more accurate clustering model and to double check for data accuracy
  - We need a dataset that spans a longer period of time for factors such as price changes/volatility since some stocks will be volatile during certain periods only
- Check model against long-term period dataset
  - Use the existing models we have created to determine how well the model captured the clustering