TRADE & AHEAD Data Analysis

Python Foundations : PGP-DSBA

18th may, 2023

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Overview
- EDA Results
- Data Preprocessing
- K-Means Clustering
- Hierarchical Clustering
- K-Means vs Hierarchical Clustering
- Business Insights and recommendations

Executive Summary

- Trade&Ahead, a financial consultancy firm, has enlisted the services of a Data Scientist to analyze stock data and provide personalized investment strategies to their clients. The dataset consists of information on companies listed on the New York Stock Exchange, including ticker symbols, company names, GICS sectors and sub-industries, current stock prices, price changes, volatilities, ROE (Return on Equity), cash ratios, net cash flows, net incomes, earnings per share, estimated shares outstanding, P/E (Price-to-Earnings) ratios, and P/B (Price-to-Book) ratios.

- The primary objective is to group stocks based on their attributes to gain insights into their characteristics and facilitate more effective investment recommendations. To achieve this, cluster analysis will be employed, enabling the classification of stocks into distinct groups based on similarities in their financial indicators.

- By analyzing attributes such as current price, volatility, ROE, cash ratio, net cash flow, net income, earnings per share, P/E ratio, and P/B ratio, stocks can be organized into clusters with similar characteristics. This approach allows for a better understanding of the diversification potential and risk profiles associated with different stocks across various market segments.

Executive Summary

- The outcomes of the cluster analysis will provide Trade&Ahead with valuable insights. These insights can be used to develop personalized investment strategies for clients, taking into account their risk tolerance, financial goals, and investment preferences. By leveraging the characteristics of each stock group, Trade&Ahead can recommend diversified portfolios that balance risk and potential returns.

- Ultimately, Trade&Ahead aims to assist clients in achieving their financial aspirations by providing well-informed investment strategies. The insights gained from the analysis and grouping of stocks will help clients navigate the complexities of the stock market and make informed investment decisions. By tailoring investment recommendations to each client's individual circumstances, Trade&Ahead can help them optimize their investment returns and protect their portfolios against potential losses.

Business Problem Overview and Solution Approach

- **Problem Definition**: The problem at hand is to analyze stock data and provide personalized investment strategies to clients of Trade&Ahead, a financial consultancy firm. The goal is to group stocks based on their attributes and gain insights into their characteristics, allowing for more effective investment recommendations. The challenge lies in efficiently analyzing the large amount of data and identifying meaningful patterns to guide investment strategies.

- **Solution Approach/Methodology**: The proposed solution approach involves the use of cluster analysis, a statistical technique that classifies data points into groups or clusters based on their similarities. By applying this methodology to the stock data, stocks with similar financial indicators can be grouped together, providing insights into their shared characteristics.

- The methodology starts with data preprocessing, where the raw stock data is cleaned, standardized, and transformed as necessary. This step ensures that the data is in a suitable format for analysis.

- Next, relevant financial indicators such as current price, volatility, ROE, cash ratio, net cash flow, net income, earnings per share, P/E ratio, and P/B ratio are selected as attributes for clustering. These indicators capture key aspects of a company's financial performance and valuation.

Business Problem Overview and Solution Approach

- The cluster analysis is then performed using an appropriate algorithm such as k-means clustering or hierarchical clustering. These algorithms partition the stocks into groups based on their attribute similarities, with the number of clusters determined by evaluating different options and selecting the optimal number based on criteria such as the silhouette coefficient or elbow method.

- Once the stocks are grouped into clusters, further analysis can be conducted to understand the characteristics of each group. This analysis may involve calculating average values or distributions of attributes within each cluster, identifying outliers or exceptional stocks, and examining the correlations between attributes.

- The insights gained from the cluster analysis can be utilized to develop personalized investment strategies for Trade&Ahead's clients. These strategies can consider the risk tolerance, investment goals, and preferences of individual clients, leveraging the characteristics of each stock group to recommend diversified portfolios.

- By employing the solution approach of cluster analysis, Trade&Ahead can effectively analyze the stock data, identify meaningful patterns, and provide personalized investment strategies to clients. This approach helps optimize investment recommendations, enhance portfolio diversification, and support clients in achieving their financial aspirations.

Data Overview

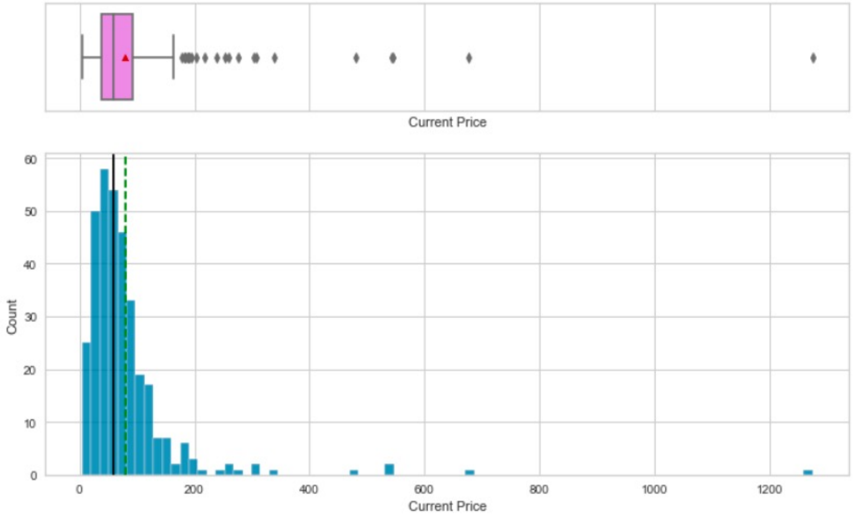
- The data set consist of 340 rows and 15 columns
- We have nine (4) object type, 7 float and 4 integer
- There are no missing or duplicated value in the dataset
- The average current price is 81
- The average price change is 4
- The average volatility rate 1.5
- The average ROE is 39
- The average cash ratio is 70
- The average net cash flow is 55,537,520
- The average net income is 1,494,384,602
- The average earnings per share is 2.7
- The average estimated shares is 577,028,337
- The average P/E Ratio is 32
- The average P/B Ratio is -1.7

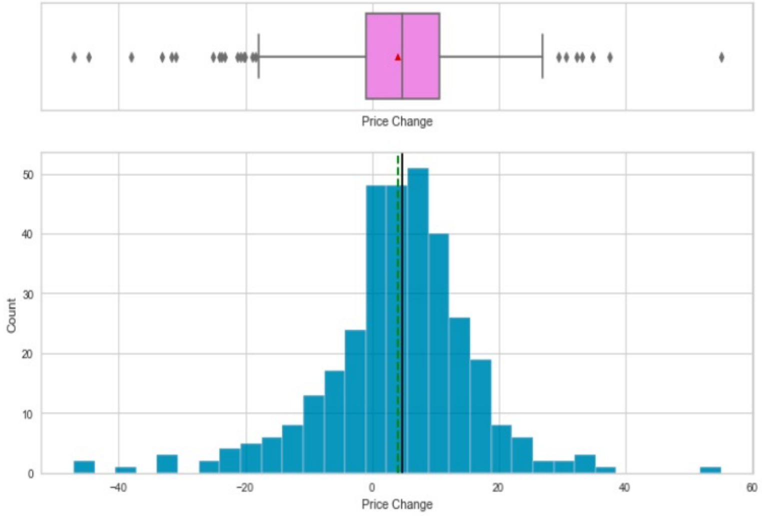
Link to Appendix slide on data background check

EDA Results

- *The distribution for current price is rightly skewed*
- *The boxplot shows that there are a lot of upper outliers to the right for this variable.*
- *The average current price is higher than the median for current price indicating the distribution is skewed to the right*

- *The average price change is almost the same with the median indicating the distribution is nearly symmetrical*
- *There are outliers on both sides for this distribution*
- *The average price change is almost the same as the median price change*

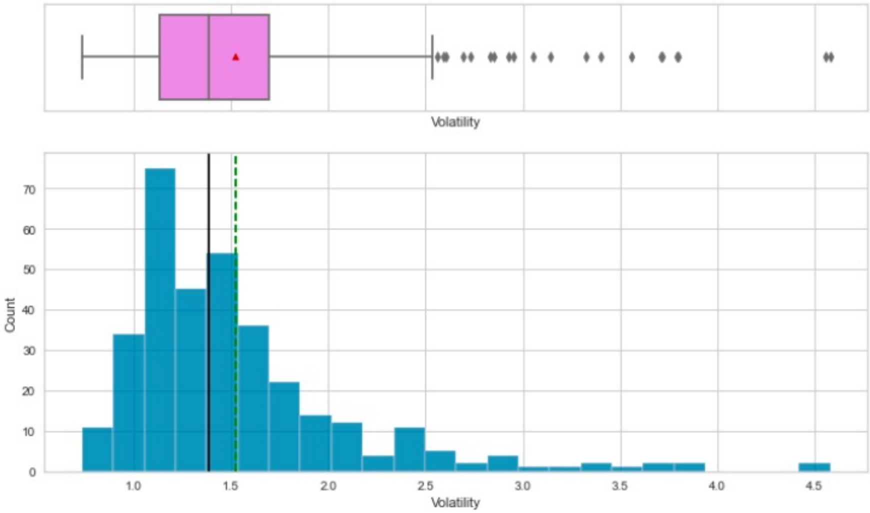


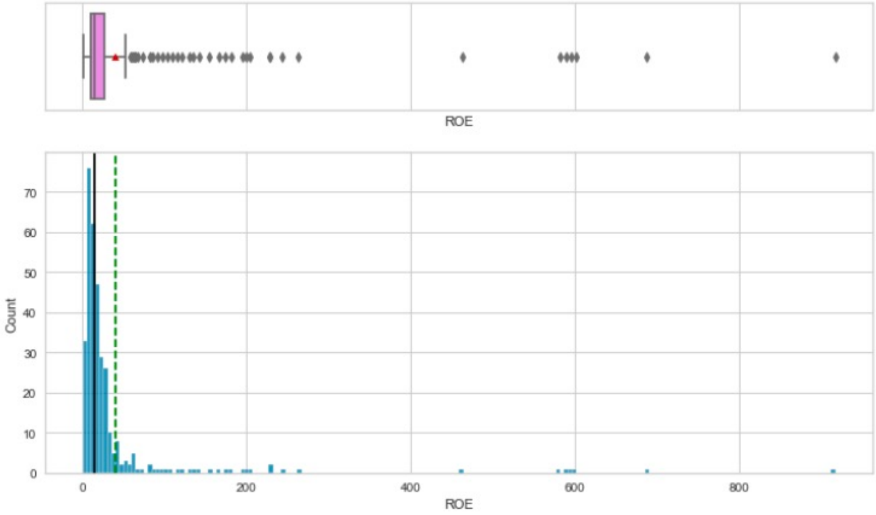


EDA Results

- *The distribution for volatility is rightly skewed*
- *The boxplot shows that there are a lot of upper outliers to the right for this variable.*
- *There is a significant difference in the average and median volatility*

- *The distribution for ROE is highly skewed to the right*
- *The boxplot shows that there are a lot of upper outliers to the right for this variable.*
- *The average ROE is higher than the median ROE*

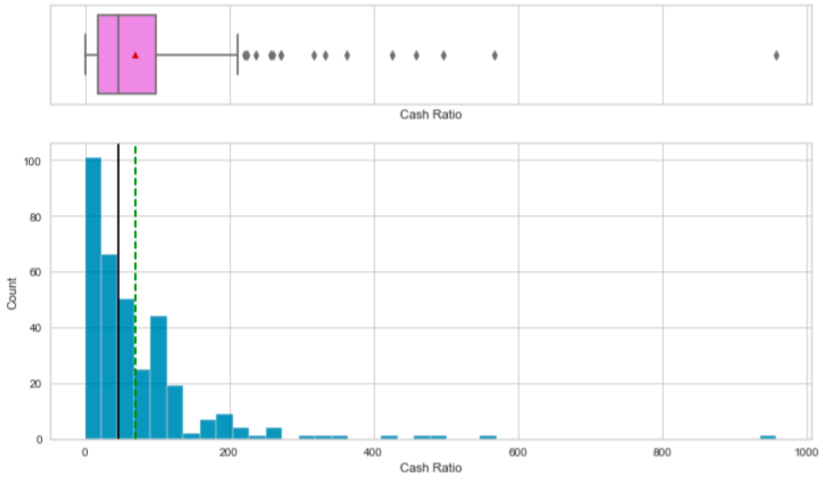


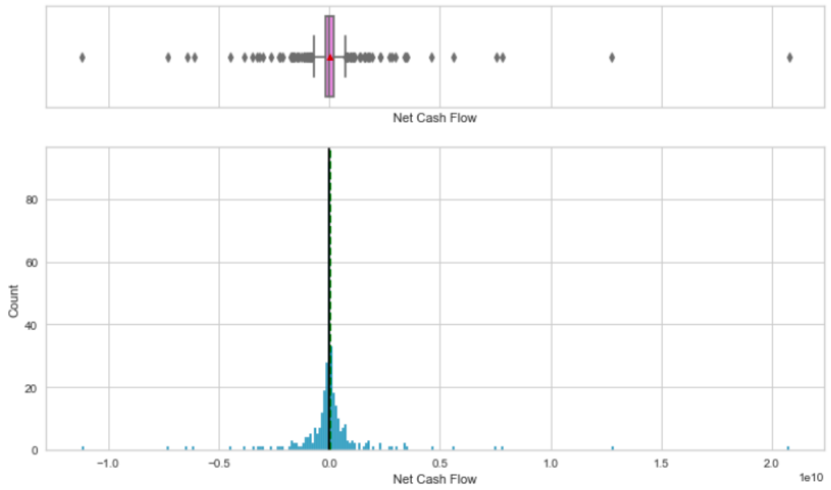


EDA Results

- *The distribution for cash ratio is rightly skewed*
- *The boxplot shows that there are a lot of upper outliers to the right for this variable.*
- *There is a significant difference in the average and median for cash ratio*

- *The average net cash flow is almost the same with the median indicating the distribution is nearly symmetrical*
- *There are outliers on both sides for this distribution*

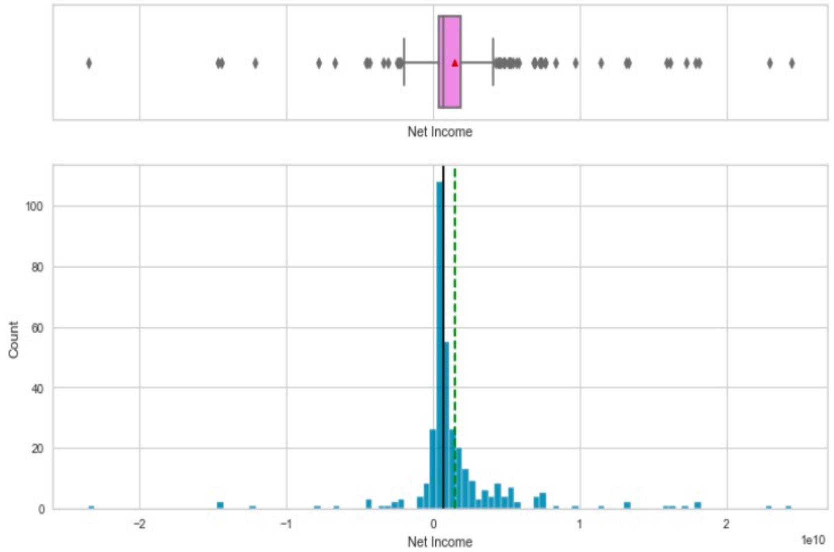


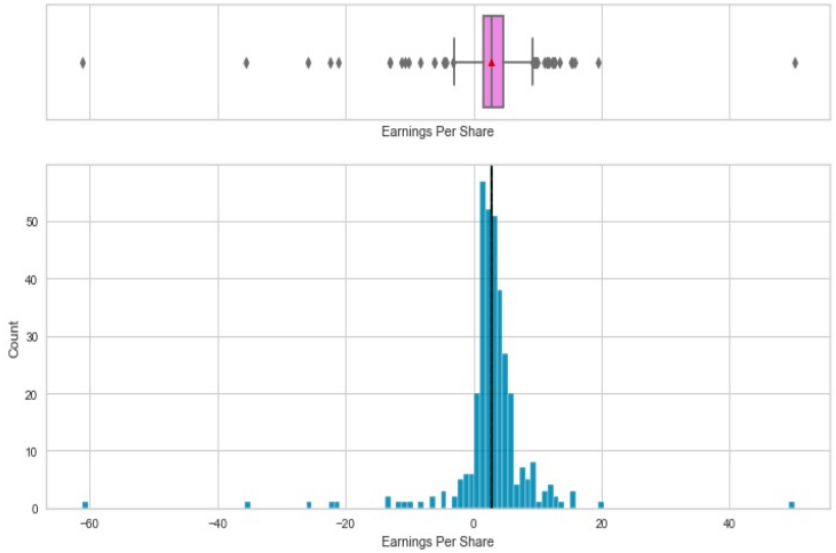


EDA Results

- *The average net income is almost the same with the median indicating the distribution is almost symmetrical*
- *There are outliers on both sides for this distribution*

- *The average earnings per share is the same with the median indicating the distribution is symmetrical*
- *There are outliers on both sides for this distribution*

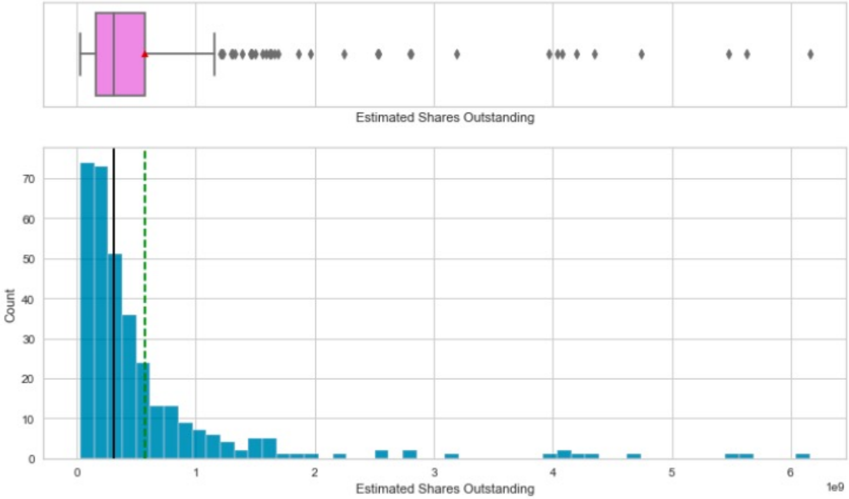


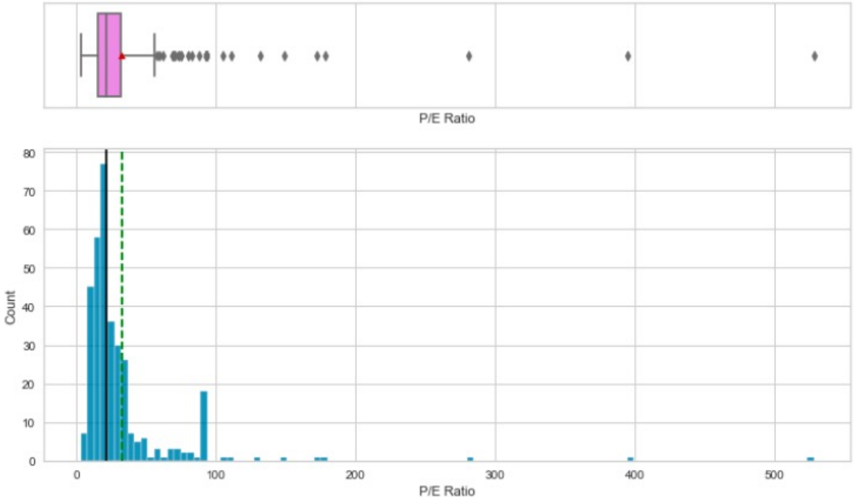


EDA Results

- *The distribution for estimated shares outstanding is rightly skewed*
- *The boxplot shows that there are a lot of upper outliers to the right for this variable.*
- *There is a significant difference in the average and median for estimated shares outstanding*

- *The distribution for P/E Ratio is rightly skewed*
- *The boxplot shows that there are a lot of upper outliers to the right for this variable.*
- *There is a significant difference in the average and median for P/E Ratio*

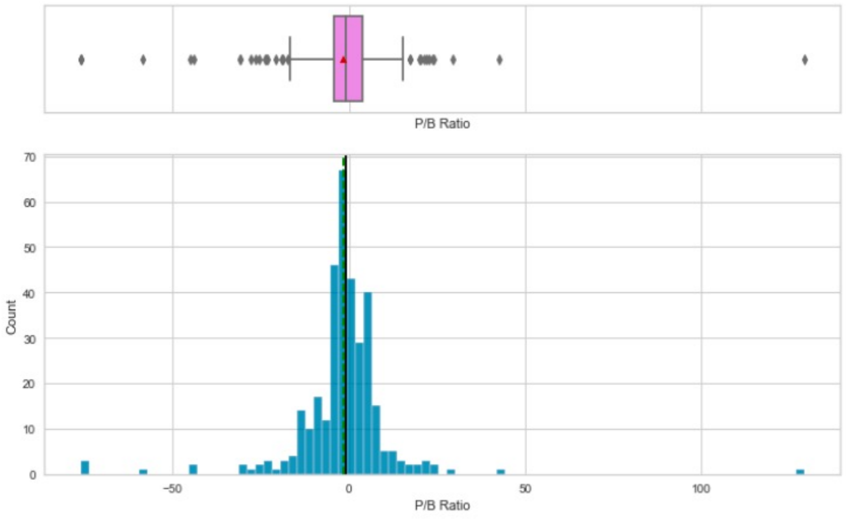


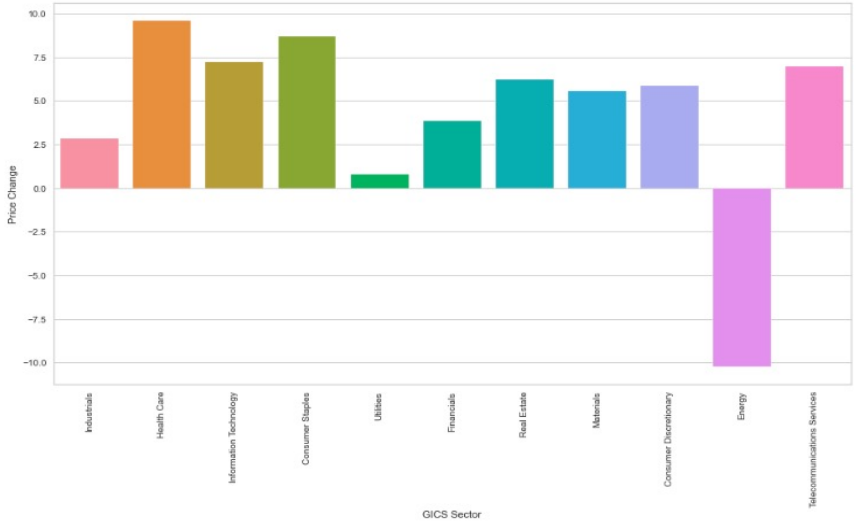


EDA Results

- *The average P/B Ratio is almost the same with the median indicating the distribution is nearly symmetrical*
- *There are outliers on both sides for this distribution*

- *Healthcare, consumer staples followed by information technology has seen the maximum stock price increase*
- *Energy has the least stock price change in 13 weeks*

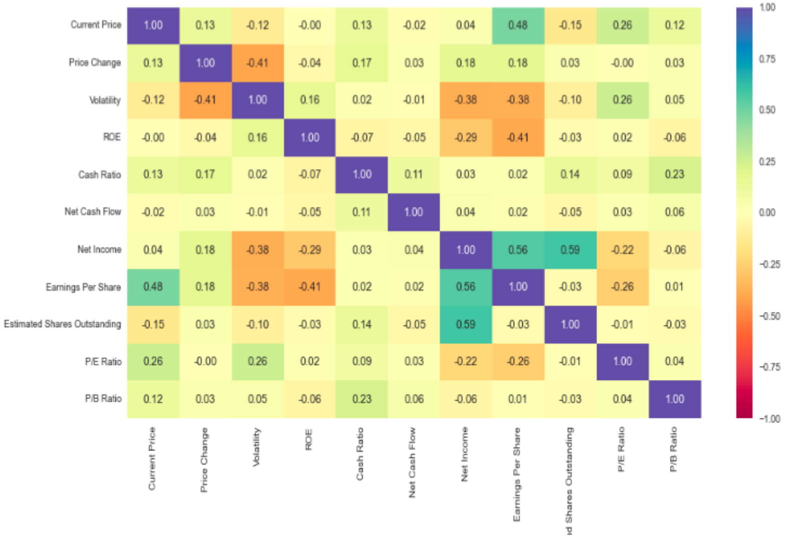


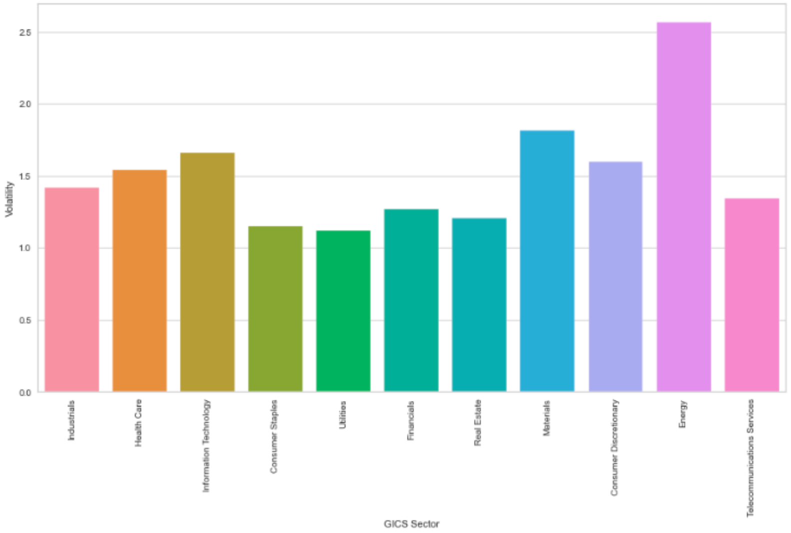


EDA Results

- *The heat map distribution shows there is a high correlation between net income and earnings per share and estimated shares outstanding*
- *The distribution also shows little correlation between price change and current price*
- *There is no correlation between volatility and price change but shows little correlation with P/E Ratio*
- *The current price also indicates a high correlation with earnings per share*

- *The energy industry shows a high volatility on stock prices which becomes risker to invest followed by material industry*
- *The utilities sector presents a low risk invest as a result of low volatility on stock prices*

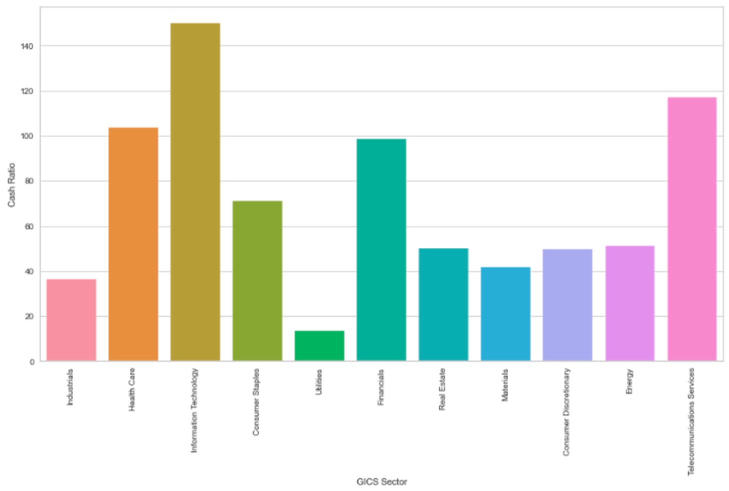


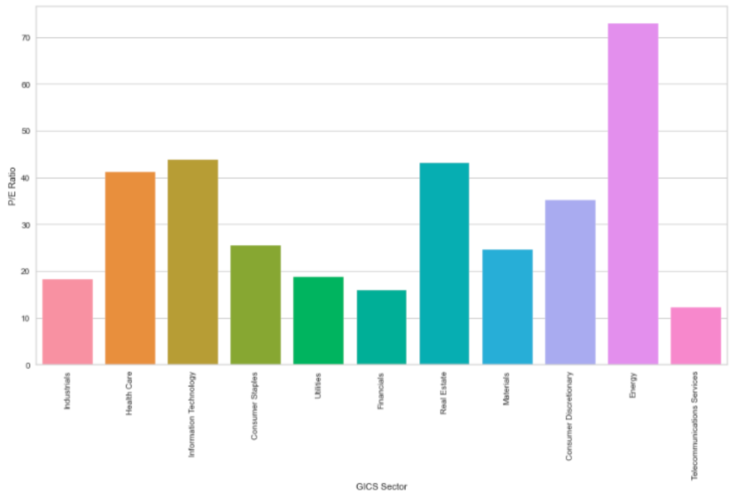


EDA Results

- *Information technology, telecommunication services and health care has shown the highest cash ratio accordingly which gives the ability to cover short term obligations with cash equivalents*

- *The energy industry shows a very high P/E ratio in this distribution compared to the other industry*
- *There is a tie between information technology and real estate while telecommunication services displays the lowest P/E ratio*





Data Preprocessing

- The data set contains no duplicates or missing values.
- The dataset contained outliers.
- The data preparation will be used in data analysis, categorizing stocks based on the qualities provided, and offering insights about each group's features.

K-Means Clustering Summary

- The appropriate k value is 4 or 6.
- Because the silhouette for 4 is larger than that of 6, we choose 4 as the value of k.

Cluster 0

- This cluster suggests that the stock has comparatively and medium prices, indicating whether it is a good or poor investment.

Cluster 1

- This cluster has a low volatility rate, a low P/B ratio, and a low P/E ratio, indicating a very low risk investment.
- A high net cash flow, cash ratio, and net income of 14,833,090,909 indicate a good investment industry.
- The net cash flow of -1,072,272,727 and cash ratio of 50 imply a low probability of the company's survival in the future.

Cluster 2

- This cluster has a high volatility rate, and the P/E Ratio indicates that it is a high risk investment.
- A low net cash flow, cash ratio, and negative net income of -3,887,457,740 indicate a weak investment industry.
- This cluster has the fewest estimated outstanding shares, indicating a low corporate value.
- This cluster has the lowest EPS of -9, suggesting poor profitability.

K-Means Clustering Summary

Cluster 3

- This cluster represents the greatest movement in stock price in 13 weeks, which can be caused by an increase or reduction in earnings.
- There is also a large net cash flow, cash ratio, and net income of 1,572,611,680, indicating that the frequent price changes are the result of increased profits.
- This cluster has a low volatility rate, making it a low-risk investment.
- The estimated number of shares outstanding held by its shareholders is relatively large, which is good for market capitalization and value.
- A high earnings per share of 6 and a ROE of 26 indicate increased profitability.
- A P/B ratio of 14 and a P/E ratio of 74 indicate that future earnings are likely to be high, and the stock may be expensive.

Hierarchical Clustering Summary

- The maximum cophenetic correlation obtained from Euclidean distance and ward linkage is the same as the average linkage.
- The dendrogram using Ward linkage revealed distinct clusters.
- The optimal number of clusters from the dendrogram using the Ward linkage approach would be four.

Cluster 0

- This cluster has a high volatility rate and a relatively high P/E Ratio, indicating that it is a high risk investment.
- By looking at the lowest stock price movement, cash ratio, net cash flow, negative net income, and earnings per share, this cluster shows low profitability.

Cluster 1

- This cluster represents the greatest movement in stock price in 13 weeks, which can be caused by an increase or reduction in earnings.
- There is also a strong net cash flow and cash ratio, indicating that the frequent price changes are due to increased earnings.
- This cluster has a low volatility rate, making it a low-risk investment.
- The estimated number of shares outstanding held by its shareholders is relatively large, which is good for market capitalization and value.
- A P/B ratio of 19 and a P/E ratio of 113 indicate that future earnings are likely to be high, and the stock may be expensive.
- However, this cluster has a low ROE and net income, which does not indicate low profitability and could be due to a big price change or corporate reorganization.

Hierarchical Clustering Summary

Cluster 2

- This cluster has a high volatility rate, and the P/E Ratio indicates that it is a high risk investment.
- A high net cash flow, cash ratio, and net income indicate a great investment industry.
- This cluster has the most estimated outstanding shares, indicating the company's high worth.
- This cluster has a low profits per share and P/B ratio, which may indicate that funds are not being spent efficiently.

Cluster 3

- This clusters consists of low and medium prices/ values which can indicate medium risk invest

K-Means vs Hierarchical Clustering

Which clustering technique took less time for execution?

- Within 0.1s, both the KMeans and the Agglomerative Clustering models matched the dataset.

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

- Both methods produce clusters that are identical, with a single cluster containing the majority of the stocks and the remaining four clusters containing 7-29 stocks.

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

- For both algorithms, four clusters yielded separate clusters with enough observations in each to distinguish which "type" of stock is characteristic of the cluster. Differences and similarities in the cluster patterns produced by both clustering algorithms. Based on the outliers in the 11 variables, both methods produced similar clusters.

Business Insights and Recommendations

The following are the insights and recommendations

- The energy industry has considerable volatility in stock prices, making it riskier to invest, followed by the material industry.
- Because of the low volatility of stock prices, the utilities industry is a low-risk investment.
- Information technology, telecommunications services, and health care have the highest cash ratios, allowing them to satisfy short-term obligations using cash equivalents.
- In this distribution, the energy industry has an extremely high P/E ratio when compared to the other industries.
- Trade&Ahead should first determine their clients' financial goals, risk tolerance, and investment behaviors before recommending a cluster as a prospective portfolio of stocks that will meet these criteria.
- However, given on the characteristics of the stocks within them, many of these clusters are effectively alternatives for standard indexes such as the Dow Jones Industrial Average and the S&amp;P 500, which might more readily fulfill these aims.
- Alternatively, Trade&Ahead could use these clusters as a jumping-off point for additional financial statement analysis, focusing on which individual stocks do not fit the "profile" of the cluster.
    - Assuming that picking individual stocks is part of a client's investing plan, Trade&Ahead may be able to identify equities that are likely to outperform their peers (i.e., price will rise = buy recommendation) or likely to underperform their peers (i.e., price will fall = sell suggestion).

Business Insights and Recommendations

In K-means clustering,

- cluster 3 has the best attributes when It comes to investing,
- cluster 1 and two indices poor investment sector,
- while cluster 0 sights an average for investors

In hierarchical clustering,

- cluster 1 comes in as the best option for investing,
- cluster 2 shows a great investment sector but comes with a high risk,
- cluster 0 shows a bad investor and
- cluster 3 comes in with average

Great Learning

Happy Learning !