Trade&Ahead Project

PGP-DSBA

11/01/2023

Contents / Agenda

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

Executive Summary/Actions, Insights and Recommendation

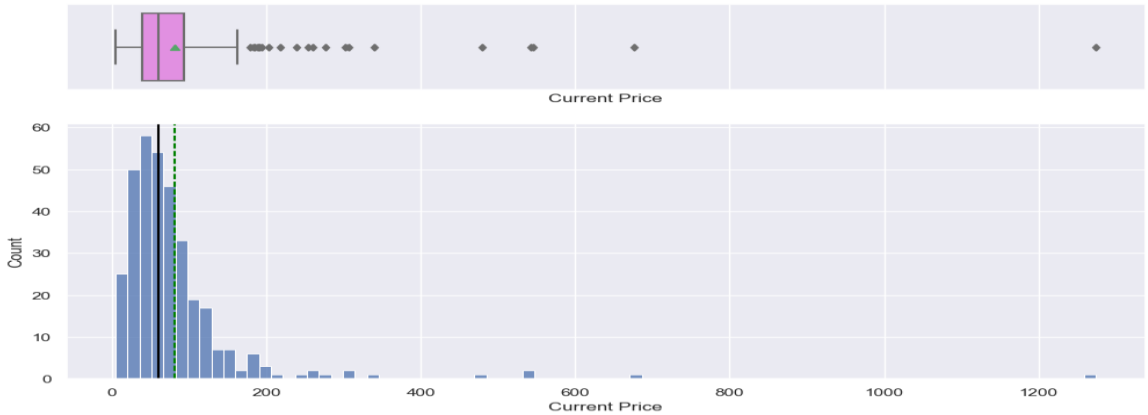
- Trade&Ahead should first identify the financial goals, risk tolerance, and investment behaviors of their clients, then recommend a cluster as a potential portfolio of stocks that will fit these needs
- However, many of these clusters, based on the characteristics of the stocks within them, are essentially substitutes for standard indexes, such as the Dow Jones Industrial Average and the S&P 500, which could more easily achieve these goals
- Alternatively, Trade&Ahead could use these clusters as a starting point for further financial statement analysis, particularly which individual stocks do not fit the "profile" of the cluster
- Assuming selecting individual stocks is a component of a client's investment strategy, Trade&Ahead may then be able to identify stocks that should outperform its peers (i.e., the price will rise = buy recommendation) or likely fall behind its peers (i.e., the price will fall = sell recommendation)

Business Problem Overview and Solution Approach

- Trade&Ahead is a financial consultancy firm that provides their customers with personalized investment strategies.
- It is important to maintain a diversified portfolio when investing in stocks to maximize 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.
- The objective is to analyze the data, grouping the stocks based on the attributes provided, and sharing insights about the characteristics of each group.

EDA Results

- There are different data points on which the exploratory univariate analysis in which we have make different plots on each data point.
- We have used the histogram boxplot of data points such as current price, price change, volatility, ROE, cash ratio, net cash flow, net income, earnings per share, estimated shares outstanding, P/E ratio, and P/B ratio.



- **Current price**
- The distribution is heavily right skewed, with 49 of the 340 stocks having twice the median value of all stocks
- As expected, no stock is listed at less of less than 0 dollars
- **Price change**
- The distribution is biased towards lower volatilities, but long tails do exist both for positive and negative price changes
- The most volatile stocks show as low as a 47% decrease to as high as a 55% increase over 13 weeks
- **Volatility**
- As expected, the distribution of standard deviations is right skewed and not normal

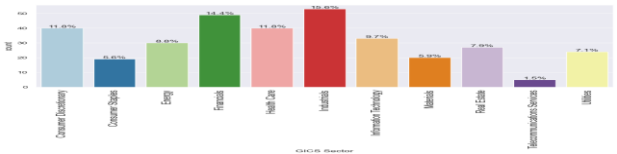
- **Cash Ratio / ROE**
- As expected, both distributions are heavily right skewed, and no stock is listed with either metric with a value of less than 0
- For example, 24 stocks are listed with returns on equity of less than 5, and 25 stocks are listed with returns of over 100 percent
- **Net Income / EPS**
- As expected, net income is shown to be right skewed with both long positive and negative tails I.e., most companies generate meager profits, but some are failing, and some are highly successful
- 32 companies within the dataset are showing a net income of less than 0 dollars
- EPS, as a derivative of Net Income, shows a similar distribution, with most showing low positive values and a few stocks (34) showing negative values
- **Estimated shares outstanding**
- The distribution is highly right-skewed, but no stock has a value of outstanding shares that is unrealistic

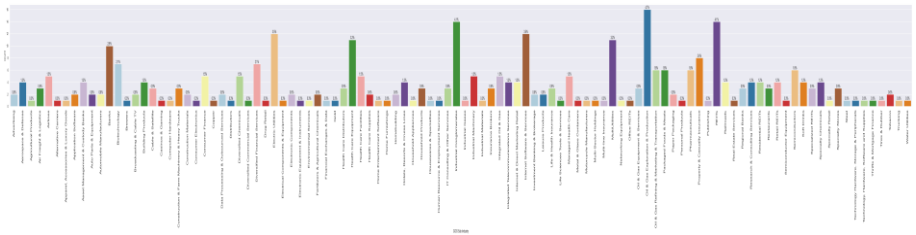
- The distribution of P/E ratios is highly right-skewed. Interestingly, no stock shows a negative ratio, even though several stocks have a negative EPS and no stock has a price list of less than 0
- The distribution for P/B ratios is mostly centered around 0 but with long positive and negative. For example, 175 of the 340 total stocks are shown to be below the 25th percentile and above the 75th percentile and additionally, 31 of the stocks are outliers
- **Conclusions**
- As expected, stocks offer uncertain returns with high upsides, mostly modest returns, and the omnipresent possibility that the value of the stock may become worthless (i.e., the company goes bankrupt)
- All of these variables contain a few or several outliers; however, none of these values appear to be unrealistic given the nature of stock prices and historical expectations

EDA Results

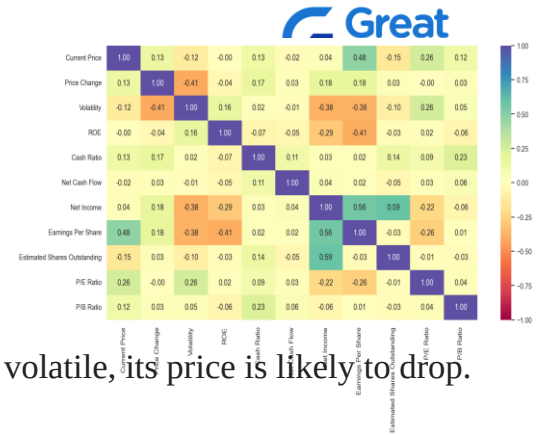
We have created labeled bar plots for the GICS Sector and sub-industry which shows results as:-

- The stocks are drawn from 11 different industrial sectors, with no one sector comprising more than 16% of the dataset
- The top 4 of the 11 sectors (industrials, financials, consumer discretionary, and health care) comprise over half of the total number of stocks
- The dataset is comprised of stocks from 104 different subindustries, with no subindustry having more than 16 stocks in the dataset.
- These observations indicate that the 340 stocks held within the dataset are highly diversified across sectors and subindustries.





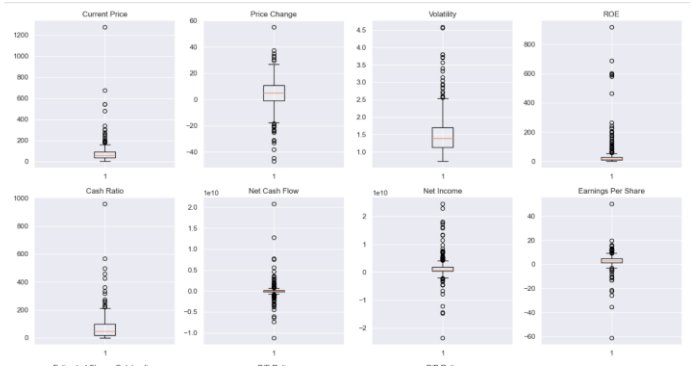
EDA Results



- Several variables are moderately correlated (+/- .40) with one another
  - Volatility is negatively correlated with price change, i.e., as a stock becomes more volatile, its price is likely to drop.
  - Net income is negatively correlated with volatility, i.e. as a company generates higher net income its price is likely less volatile
  - Net income is also positively correlated with earnings per share (EPS) and estimated shares outstanding
  - EPS is positively correlated with current price, i.e. as a company's EPS rises, its prices are also highly likely to increase
  - EPS is also negatively correlated with ROE, i.e. as a company generates more equity for shareholders, an equivalent amount of net income in the following periods will generate a lower return

Data Preprocessing

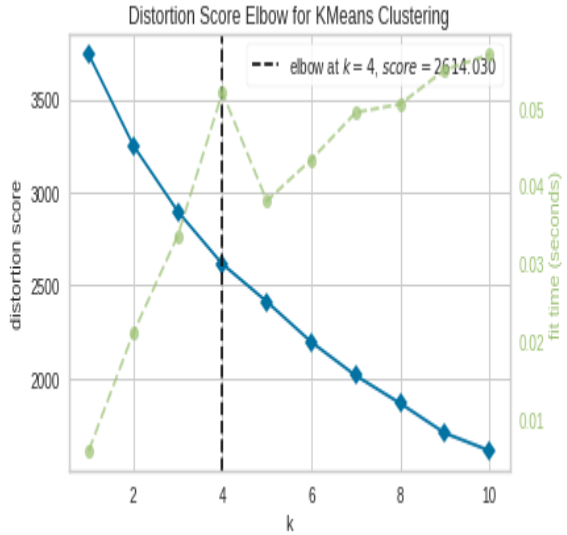
- Dataset has no missing or duplicate values
- All columns with a dtype object should be dtype category to conserve memory.



K-Means Clustering Summary

For checking the elbow plot this shows the number of clusters and and the graph which shows the raw data of average distortion.

```
Number of Clusters: 1   Average Distortion: 2.5425069919221697
Number of Clusters: 2   Average Distortion: 2.382318498894466
Number of Clusters: 3   Average Distortion: 2.2683105560042285
Number of Clusters: 4   Average Distortion: 2.175554082632614
Number of Clusters: 5   Average Distortion: 2.1166317116050672
Number of Clusters: 6   Average Distortion: 2.057226072692816
Number of Clusters: 7   Average Distortion: 2.0349967805679707
Number of Clusters: 8   Average Distortion: 1.9638459131602009
Number of Clusters: 9   Average Distortion: 1.9411480759161681
Number of Clusters: 10  Average Distortion: 1.869218987537272
```

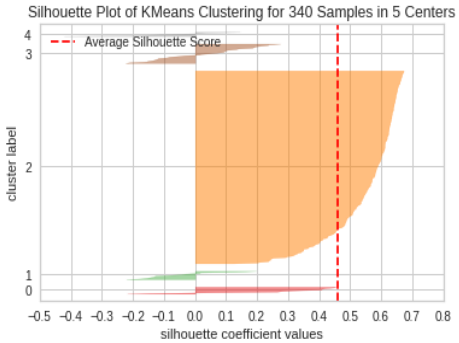


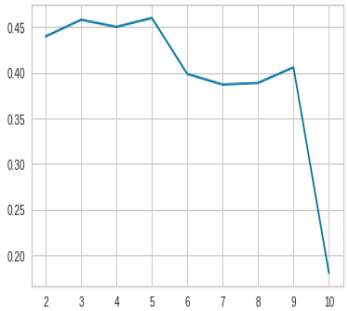
Link to Appendix slide on K-Means Clustering

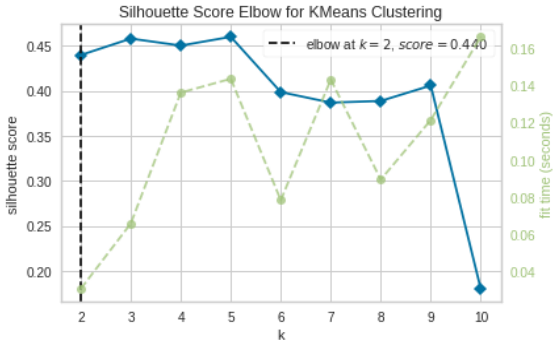
K-Means Clustering Summary

- We have checked here the silhouette scores for the range of k clusters and that is plotted in the graph which makes it clearer.
- Between the Elbow and Silhouette plots, the number of clusters with the best performance appears to be 5

```
For n_clusters = 2, the silhouette score is 0.4396963509980457)
For n_clusters = 3, the silhouette score is 0.45797710447228496)
For n_clusters = 4, the silhouette score is 0.45017906939331087)
For n_clusters = 5, the silhouette score is 0.4599352800740646)
For n_clusters = 6, the silhouette score is 0.3985379248608659)
For n_clusters = 7, the silhouette score is 0.3868475076242907)
For n_clusters = 8, the silhouette score is 0.3886929719130642)
For n_clusters = 9, the silhouette score is 0.40581042332267614)
For n_clusters = 10, the silhouette score is 0.18011528994705786)
```







K-Means Clustering Summary

- Cluster profiling:-
- We have got the cluster profiling of the k_means cluster which shows the different results.

KMeans_clusters	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio
0	50.477272	5.588148	1.141171	31.545455	64.181818	-2581727272.727273	14675545454.545454	4.490909	4012176129.000000	14.010093
1	81.418719	10.536341	1.578634	17.000000	367.538462	3857062692.307693	3129067846.153846	2.197692	1334755181.584615	65.639418
2	73.769121	5.466467	1.392827	34.632143	55.710714	4183132.142857	1443269353.571429	3.629625	430217149.035393	24.132318
3	38.808966	-13.680395	2.938240	106.034483	55.551724	-189825655.172414	-3578126517.241379	-8.657586	463121182.880690	85.946813
4	585.527134	7.752090	1.508020	17.571429	159.142857	210520428.571429	804590428.571429	14.410000	116080574.760000	118.763084

K-Means Clustering Summary

- **Cluster 0 - Large Market Capitalization / Dow Jones Industrial Average**
- 11 stocks, comprised mostly of stocks within the Financials, Health Care, Information Technology (IT), and Consumer Discretionary sectors
- Companies within this cluster have:
- Low volatility
- Most of the companies with the highest outflows of cash
- The highest net incomes
- The highest number of shares outstanding
- **Cluster 1 - "Cash is King"**
- 13 stocks, comprised mostly of stocks within the Healthcare and IT sectors
- Companies within this cluster have:
- Moderate volatility
- Mostly profitable
- Most of the highest cash ratios and cash inflows

K-Means Clustering Summary

- **Cluster 2 - S&P 500 / Diversification**
- 280 stocks (~84% of all stocks in the dataset) drawn from all sectors present in the dataset
- Companies within this cluster have:
- Low P/E ratios
- Most of the outliers on negative P/B ratios
- **Cluster 3 - "Ride the Energy Rollercoaster" portfolio / Growth Mindset**
- 29 stocks, a vast majority of which are from the Energy sector
- Companies within this cluster have:
- Low stock prices, but high ROE
- High beta
- Most of the most volatile stocks, especially those with outliers in price decreases
- Mostly negative net incomes and earnings per share

K-Means Clustering Summary

- **Cluster 4 - High Earnings for a High Price**
- 7 stocks, comprised mostly of stocks from the Health Care and Consumer Discretionary sectors
- Companies within this cluster have:
- Most of stocks with the highest prices
- Favorable cash ratios
- The most favorable P/B ratios
- Most of the highest earnings-per-share

Hierarchical Clustering Summary

We have observed here the values of the cophenetic correlation of different points.

```
Cophenetic correlation for Euclidean distance and single linkage is 0.9232.
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 weighted linkage is 0.8694.
Cophenetic correlation for Chebyshev distance and single linkage is 0.9063.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.5989.
Cophenetic correlation for Chebyshev distance and average linkage is 0.9338.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.9127.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.9259.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.7925.
```

The highest cophenetic correlation is 0.9423 which is obtained with the average linkage.

- The cophenetic correlation is highest for average and centroid linkage methods, but the dendrogram for average appears to provide better clusters
- 5 appears to be the appropriate number of clusters for the average linkage method

Link to Appendix slide on Hierarchical Clustering

Hierarchical Clustering Summary

- After the correlation I have done the cluster profiling which shows results as:-

HC_clusters	Current Price	Price Change	Volatility	ROE	Cash Ratio	Net Cash Flow	Net Income	Earnings Per Share	Estimated Shares Outstanding	P/E Ratio	P
0	77.884243	4.105986	1.516865	35.320359	66.775449	-32825817.365269	1535255703.592814	2.903308	559027333.145509	32.437511	-1
1	25.640000	11.237908	1.322355	12.500000	130.500000	16755500000.000000	13654000000.000000	3.295000	2791829362.100000	13.649696	1
2	24.485001	-13.351992	3.482611	802.000000	51.000000	-1292500000.000000	-19106500000.000000	-41.815000	519573983.250000	60.748608	1
3	104.660004	16.224320	1.320606	8.000000	958.000000	592000000.000000	3669000000.000000	1.310000	2800763359.000000	79.893133	5
4	1274.949951	3.190527	1.268340	29.000000	184.000000	-1671386000.000000	2551360000.000000	50.090000	50935516.070000	25.453183	-1

- There are 2 clusters of one company, 2 clusters of two companies, and a single cluster of the remaining 334 companies
- The clustering of these companies does not solve the business problem at hand, because the clusters do not have enough variability

Hierarchical Clusters

- **Cluster 0 - Growth for a Price**
- 15 stocks, comprised mostly of stocks within the Health Care, Information Technology (IT), and Consumer Discretionary sectors
- Companies within this cluster have:
- Most of stocks with the highest prices
- Significant outliers in price-to-equity ratio
- The most favorable price-to-book (P/B) ratios
- Most of the highest cash ratios
- **Cluster 1 - Short-term Poor, Long-term Rich**
- 7 stocks, comprised mostly of stocks within the Consumer Staples and Energy sectors
- Companies within this cluster have:
- The highest returns-on-equity
- The lowest net incomes
- Mostly negative earnings per share

Hierarchical Clusters

- **Cluster 2- DJIA**
- 11 stocks, comprised mostly of stocks within the Financials and Telecommunications sectors
- Companies within this cluster have:
- Most of the companies with the highest inflows and outflows of cash
- The highest net incomes
- The highest number of shares outstanding
- **Cluster 3 - Diversification**
- 285 stocks (~84% of all stocks in the dataset) drawn from all sectors present in the dataset
- Companies within this cluster have:
- Most of outliers in price increases and some of the outliers in price decreases
- Some of outliers in cash inflows and outflows
- Most of the outliers in P/B ratio

Hierarchical Clusters

- **Cluster 4 - Energy-specific portfolio**
- 22 stocks, a vast majority of which are in the Energy sector
- Companies within this cluster have:
- Most of the most volatile stocks, especially those with outliers in price decreases
- Mostly negative net incomes and earnings per share

K-means vs Hierarchical Clustering

Which clustering technique took less time for execution?

- Both the KMeans model and the Agglomerative Clustering model fit the dataset within ~0.1s

Which clustering technique gave you more distinct clusters, or are they the same?

- Both algorithms give similar clusters, with a single cluster of a 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, 5 clusters provided distinct clusters with sufficient observations in each to reasonably differentiate which "type" of stock is representative of the cluster.

Great Learning

Happy Learning !