Executive Summary
Prepared By: Ruchika Venkateswaran

Segmentation of Publicly Traded Stocks as of September 2018

Introduction

Cluster analysis enables us to group a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). One can preset the number of clusters before grouping variables into their respective clusters. An advantage of using cluster analysis over other regression techniques is its ability to classify sets of like-minded groups. This report explores the possibility of clustering publicly traded stocks into a set of like groups.

Data Cleaning

About the Data

This dataset consists of 755 observations and 43 features, with some of the most important features spanning across current assets and liabilities, P/E ratio, earnings per share (EPS), share price, stock price at high and low, ROA and dividend payout ratio. Most of the features are floats/integers except for 'ticker' and 'quarter_end', which are objects.

Filtering by Date

Each ticker is unique (i.e., each ticker only has one observation and falls under only one date). After converting the 'quarter_end' variable's datatype to 'datetime', we can observe that approximately 81% of the stock price data has been taken from 2018, while the remaining 19% of the data falls within the 2006-2017 timeframe. Since the stock market is extremely volatile, including 12 years of data will not yield accurate segmentations of the data. For this reason, the data frame has been filtered to exclude observations from 2006-2017. 87% of the data in 2018 belongs to Q3, and 97% of the 2018-Q3 data includes share prices for the month of September. The final resulting data frame includes data as of September 2018.

Handling Missing Values

Observations with missing values for current assets also have null values for current liabilities and current ratio (current assets/ current liabilities). The assumption made in this scenario is that null values of current ratio are 0. Since current ratio is the ratio of current assets to current liabilities, we have substituted null values for current assets and current liabilities with 0 as well. Observations with null values for net margin, cash from operating, investing and financing activities have been dropped. The missing value treatment for the remaining features are described below using accounting ratios to impute values:

- Equity to Assets Ratio: Shareholders Equity/Total Assets
- Book Value Per Share = (Shareholders Equity – Preferred Equity)/Shares Outstanding
- Stock Price = P/B ratio *Book value of equity per share
- Net Income = Net Margin*Revenue
- ROE = Net Income/Shareholders Equity
- Long Term Debt to Equity Ratio = Long-term debt/Shareholders equity
- P/E Ratio = Share Price/EPS

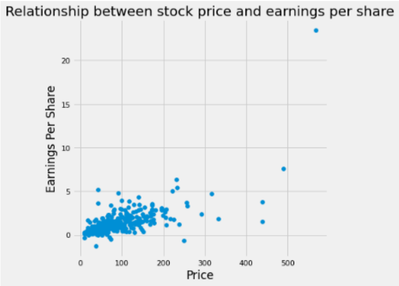
Outlier Treatment

Features such as share price have its maximum value at $ 308080, while EPS has a minimum value at ($-15.52) and the P/E ratio maximum value is 1076. Top and bottom 1 percentile outliers have been removed. The final cleaned dataset consists of 454 observations.

Exploratory Data Analysis

Stocks with current prices close to their 'price at high' generally have higher EPS as well, and the ratio of the current stock price and its price at high helps in determining if the stock price would increase in future. Therefore, an additional column on 'current/high price ratio' has been created. As displayed in the figure below, generally, there is a linear relationship between the share price and the EPS, with the EPS increasing as the stock price increases. The relationship between share price and EPS plays a crucial role in segmenting stocks which will be explained in the next section. The data has been scaled in order to standardize all values before clustering.

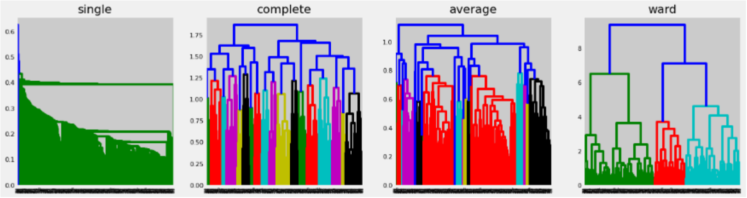
Executive Summary
Prepared By: Ruchika Venkateswaran



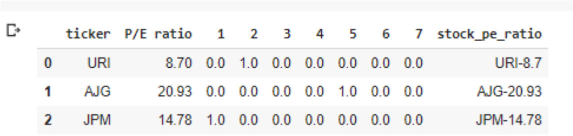
Unsupervised Machine Learning Using Hierarchical Clustering K-Means Clustering

Hierarchical Clustering

This method is based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. In the image below, each linkage method starts with classifying all data points into separate clusters and then aggregating them as the distance decreases. The linkage methods work by calculating the distances or similarities between all objects. Then the closest pair of clusters are combined into a single cluster, reducing the number of clusters remaining. The process is then repeated until there is only a single cluster left.



Clusters have finally been profiled below. However, a more intuitive manner of clustering similar groups of stocks is using K-Means clustering, as described in the following section.



	ticker	P/E ratio	1	2	3	4	5	6	7	stock_pe_ratio
0	URI	8.70	0.0	1.0	0.0	0.0	0.0	0.0	0.0	URI-8.7
1	AJG	20.93	0.0	0.0	0.0	0.0	1.0	0.0	0.0	AJG-20.93
2	JPM	14.78	1.0	0.0	0.0	0.0	0.0	0.0	0.0	JPM-14.78

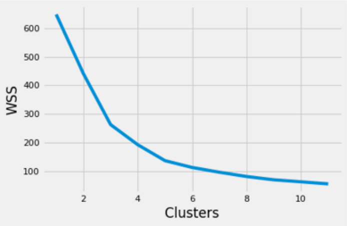
K-Means Clustering

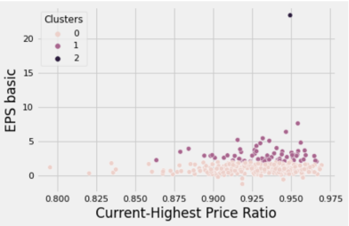
Finally, K-Means Clustering has been implemented to analyze similarities between different clusters of stocks. Segments have been clustered based on i) EPS and Share Price ii) EPS and Current/High Share Price Ratio iii) P/E Ratio and Share Price. The objective of K-means is to group similar data points together and discover underlying patterns. To achieve this, K-means looks for a fixed number (k) of clusters in a dataset, which we generally specify. To identify k, there are several methods that can be used. For this dataset, the Elbow Plot and Silhouette methods have been used to identify the ideal value of k.

A function has been created which loops through values of k starting from 1 to 12. The function also calls the WSS (total within-cluster sum of square) score for each value of k. The total WSS measures the compactness of the clustering and we want it as minimal as possible. The Silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990).

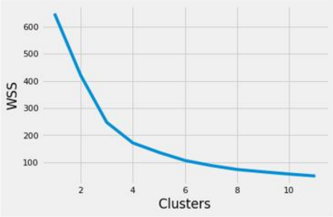
- EPS and Current-Highest Price Ratio: K as specified in both methods is 3. The company that belongs to cluster 2(dark purple) with higher current-high price ratioand a higher EPS value is a well established firm. When the stock price is high and EPS is also higher, then the probability of stock appreciation is low. Investors can consider investing in those stocks that have more potential to increase in price in the future.

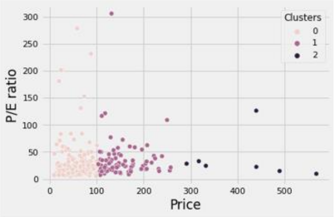
Executive Summary
Prepared By: Ruchika Venkateswaran



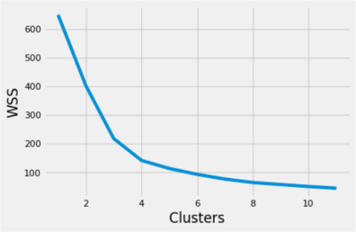


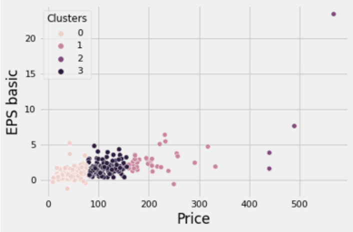
- Share price and P/E ratio: k is 3 in this scenario. There is evidence of misclassification of cluster values. This is because extremely high P/E ratios have been classified along with lower values of price and P/E ratios. Generally, when the P/E ratio and share price is high, the stock is already valued a high price and has low chances of appreciation.





- EPS and Share Price: Both the Elbow (displayed below) and Silhouette methods identify k as 4. When plotting the share price by earnings per share, we note that the light colored clustered (Cluster 0) has stocks with low price and EPS. This cluster segments those companies that are not profitable, as reflected in the stock price value. Cluster 2 (purple) displays values of high stock price and higher EPS, which groups companies that are high performing and well established. Clusters 1 and 2 in the middle have moderately priced share prices with slightly lower EPS values, which are well performing companies that have potential to grow.





Conclusion

To conclude, the third chart (displayed above) with share price and earnings per share is a great visual representation of segmentation of similar groups of stocks. The final recommendation is to invest in companies with more potential of stock appreciation. Stocks in the above chart which are reasonably priced (between $100-$200) and with lower EPS have greater potential to increase in the longer run. Investors should consider investing in stocks which will yield greater returns in future. However, additional qualitative factors such as industry type, employee turnover, etc should also be taken into consideration before making investment decisions.