BA 820: Unsupervised Machine Learning
Spring 2021

Assignment #1 – Executive Summary

Overview

Upon receiving the dataset, we first conducted data exploration. During this time, we discovered that the dataset contains numerous missing values for features that could be important to our analysis. Therefore, we took several approaches to clean the dataset before proceeding to the next step. Then, we performed clustering analysis on the clean dataset and discovered that the data could potentially be segmented into around 6 like-groups. Lastly, we provided the client with some insights on the clusters and provided recommendations based on our findings.

Exploratory Data Analysis, Data Cleaning, and Preprocessing

Our dataset consists of 755 rows and 43 columns. Each row represents an individual stock and the columns are mostly financial indicators. 41 of the columns are numeric while 2 of them consist of textual data – the ticker of a stock and the end date of a quarter. Since there is no duplicate stock, we set the unique tickers as the indices. After taking a look at all the columns, we decided to remove 3 of them:

- *"Shares split adjusted"* – The values in this column are the same as the values in *"Shares"*. Therefore, we chose to remove this duplicate column.
- *"Split factor"* – Every stock has the same value 1 in this column, which makes it less informative when it comes to distinguishing different stocks.
- *"Quarter end"* – This is an empty column.

That leaves us with 38 numeric columns. Next, we moved on to inspecting the missing values in the dataset. We found that 462 stocks have at least one missing value. Given that is almost 40% of the data, we could not simply drop all stocks with missing values. Therefore, we needed to fix the missing values column by column. There are a few approaches we took:

1. The number of missing values for a stock range from 1 to 12. While most (277) stocks have less than 8 missing values, there are 16 stocks with 8 or more missing values. We chose to discard these stocks because they are missing information on over 20% of the features.
2. After we dropped 16 stocks, there remain 18 columns that contain missing values. Using the rest of the available data, we could approximate the values for some of them. For example, we estimated *"P/B ratio"* by dividing *"Price"* by *"Book value of equity per share"*.[1] Eventually, we were able to approximate the values for 10 columns. That includes *"ROE"*[2], *"Net margin"*, *"Equity to assets ratio"*, *"Dividend payout ratio"*[3], *"P/E ratio"*[4], and *"Long-term debt to equity ratio"*.
3. For columns such as *"Current Assets"*, we used a similar column *"Assets"* as an estimate for it. This rule applies to *"Current Liabilities"* and *"Current ratio"* as well.
4. After we imputed missing values for most stocks, there are only 2 stocks left – "R" and "GD" – that still have missing values. The information missing is on the companies' cash flow, which we could not calculate using the data at hand. Therefore, we decided to discard these 2 stocks.

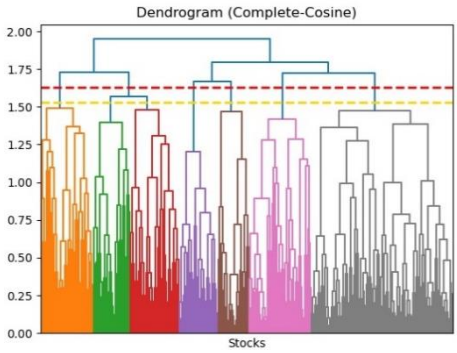
[1] Jason Fernando, "Price-To-Book (P/B Ratio)," Investopedia, last updated January 24, 2021, https://www.investopedia.com/terms/p/price-to-bookratio.asp
[2] Ryan Fuhrmann, "Return on Equity (ROE) vs. Return on Assets (ROA)," Investopedia, last updated November 25, 2020, https://www.investopedia.com/ask/answers/070914/what-are-main-differences-between-return-equity-roe-and-return-assets-roa.asp
[3] Adam Hayes, "Dividend Payout Ratio," Investopedia, last updated August 23, 2020, https://www.investopedia.com/terms/d/dividendpayoutratio.asp
[4] Jason Fernando, "Price-to-Earnings Ratio – P/E Ratio," Investopedia, last updated February 8, 2021, https://www.investopedia.com/terms/p/price-earningsratio.asp

When we completed the data cleaning process, we managed to keep over 93% of the data. There are now 705 stocks, 38 columns, and no missing values. The last step to take before we conduct cluster analysis is to standardize the data. We rescaled features so that they all have a mean of 0 and a standard deviation of 1. This way, no feature has more influence on the clustering algorithm than the others.

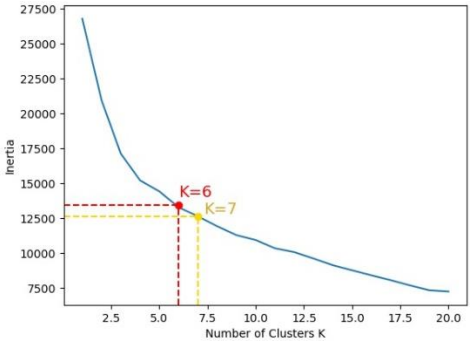
Hierarchical Clustering and K-Means Clustering

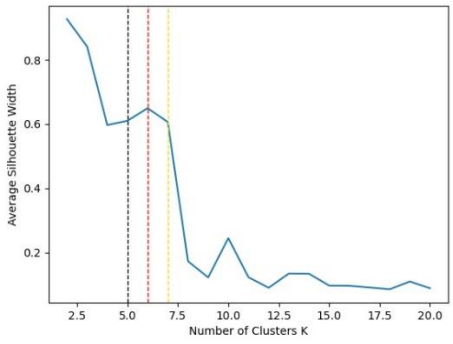
Since the goal is to segment stocks into like-groups, we decided to try out two clustering techniques – hierarchical clustering and K-Means clustering – and examine which one yields better results.

1. **We used a dendrogram to visualize the hierarchical relationship.**
   After trying out different combinations of linkage methods and distance metrics, we found that the dataset could potentially be segmented into 6 or 7 clusters (as illustrated in the dendrogram on the right). Looking at the length of the vertical lines, we could say that creating 6 or 7 clusters generally ensures that the clusters are at a sufficient distance apart from each other. The red dashed line represents the scenario where we put stocks into 6 clusters, and the yellow line represents 7 clusters.

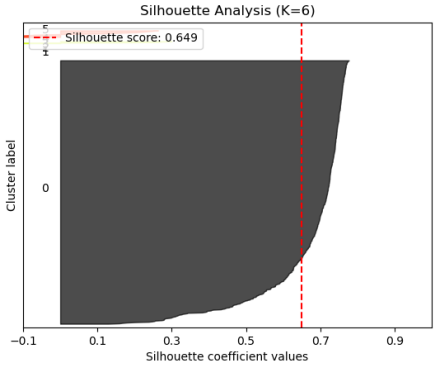


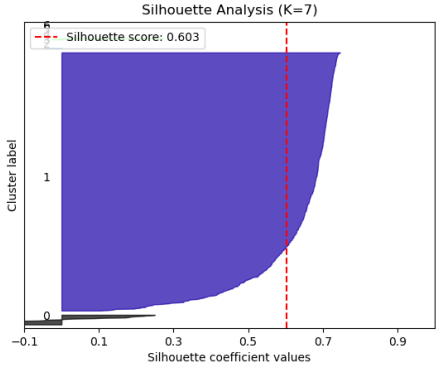
2. **We recorded the inertia and the average silhouette score for different numbers of clusters before conducting K-Means clustering.**
   The graph on the left demonstrates how the inertia decreases as the number of clusters *K* increases, while the graph on the right shows how the average silhouette score decreases as *K* increases. From these two graphs, we could see that the optimal number of clusters is also around 6 or 7, where the inertia is low enough (but does not converge yet), and the average silhouette score is still high.





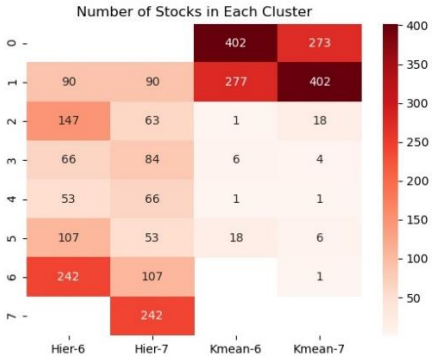
3. **We further inspected the silhouette score for each stock.**
   We could see from the graphs below that the majority of stocks have a positive silhouette score, indicating that they are similar to the other stocks in their assigned clusters. However, it's worth noting that most stocks seem to be in one cluster, even though there are 6 or 7 clusters in total. This could be a sign that K-Means is not the most suitable clustering method for this dataset.





4. **We examined the distribution of stocks in clusters based on different clustering methods.**

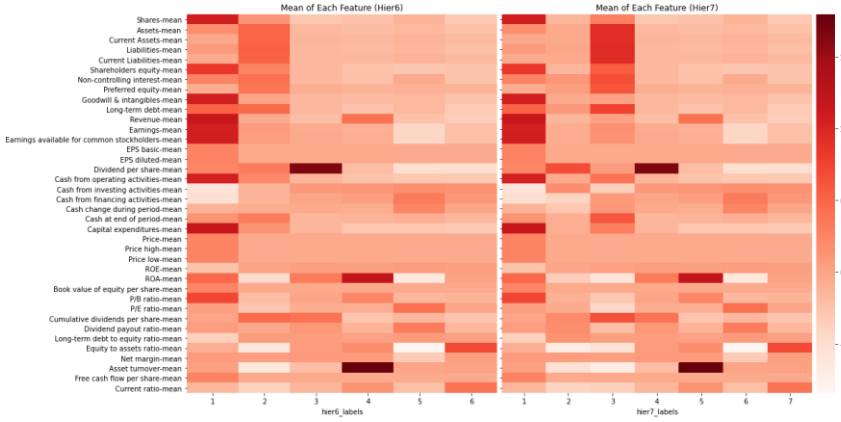
   From the heatmap below, we could see that the distribution of stocks is highly uneven among K-Means clusters – around 96% of the stocks are assigned to 2 of the 6 (or 7) clusters. On the other hand, the stocks appear to be more evenly-distributed among clusters, when we used the hierarchical clustering method on the dataset. Therefore, for this dataset, we would choose hierarchical clustering as the method to segment observations.

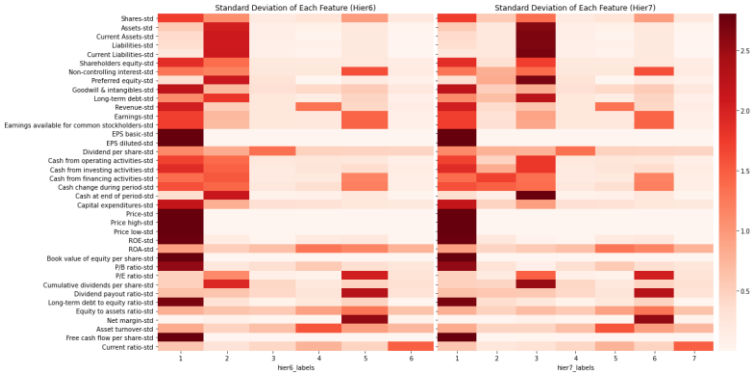


Conclusion and Recommendations

After selecting a clustering method, we assigned each stock to a cluster, took a look at the summary statistics, and tried to identify the characteristics of each cluster. The four graphs here illustrate the mean and standard deviation of each feature, when we created 6 or 7 clusters.

Looking at the mean of each feature, we could tell that stocks in Cluster3 have a particularly high average *"Dividend per share"*, while stocks in Cluster4 have a particularly high *"Asset turnover"* on average. On the other hand, stocks in Cluster1 generally have higher *"Revenue"* and *"Earnings"*. Last but not least, stocks in Cluster5 have a particularly low *"Equity to assets ratio"* on average.





While stocks in Clusters1 generate higher revenue, we noticed that the standard deviation of these stocks' price is also pretty high. This could mean that Cluster1 consists of stocks that are volatile. In addition, the stocks in Cluster2 (or Cluster3, if there are 7 clusters) appear to have more *"Assets"* and *"Liabilities"*, even though the standard deviation for these features are high as well.

In general, whether we segment the stocks into 6 or 7 clusters, these observed patterns seem to apply to both cases. For example, one might want to avoid stocks in Cluster5 (or Cluster6, if there are 7 clusters) as low equity-to-assets ratio often indicates a company is at greater financial risk. Earlier in the dendrogram, we saw that the stocks can be segmented into 2 to 7 clusters. However, considering the nature of investment, we would recommend that the client segment the stocks into at least 5 or 6 clusters. This way, the client can build a more diverse portfolio and reduce risk.