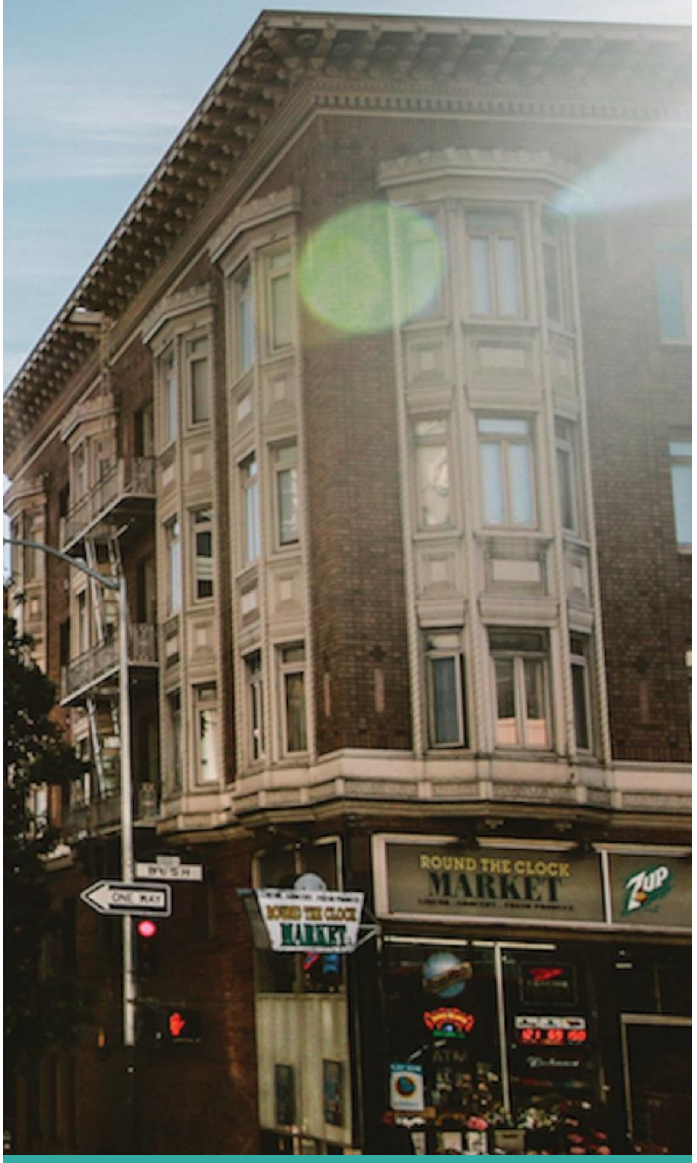
LAB 5: HIERARCHICAL CLUSTERING



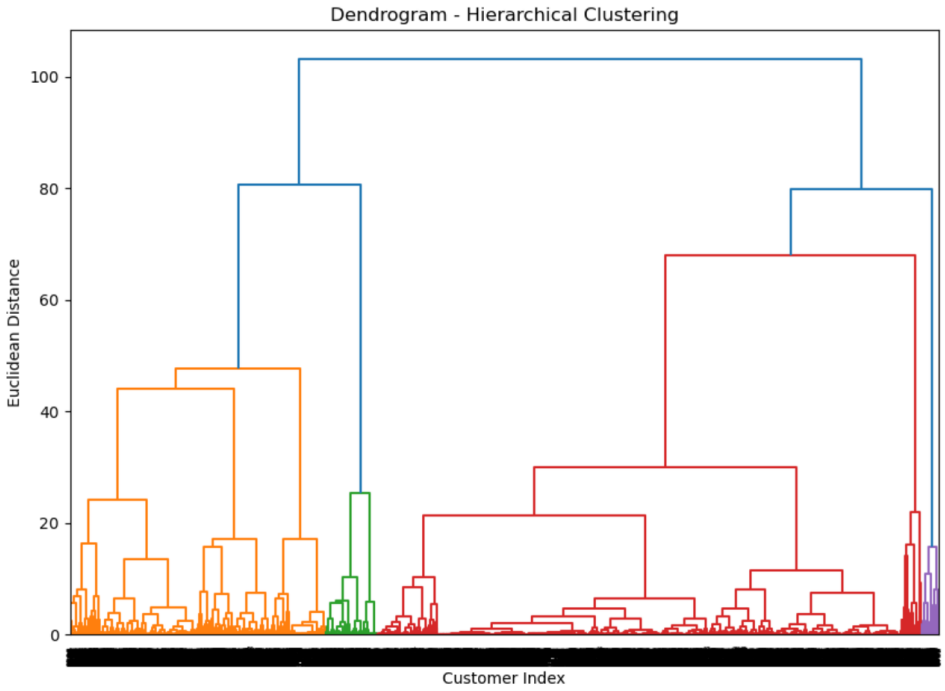
OCTOBER 22

BSAN6070 INTRODUCTION TO MACHINE LEARNING

Authored by: Andres Sebastian Gaibor Heredia

1. Hierarchical Clustering

We will continue to use our financial dataset which comprises 26,154 companies from fiscal year 2000-2024. We are dropping the categorical variables that have missing values as we reviewed, they are not crucial for the analysis. The resulting dataframe contains 235,996 records and 301 fields. From there we select 10,000 records to be able to generate the linkage matrix for the dendrogram. We then Winsorized the two selected features for the hierarchical cluster analysis, earnings per share and working capital. We use the 1st and 95th percentile to trim down outliers in the data. Then we standardized the features using StandardScaler() for better clustering performance. We use the Ward method to generate the linkage matrix for the dendrogram. Then we plot the resulting dendrogram to determine the optimal number of clusters.



From the plot we can see that the optimal number of clusters is 4, so we employ it to generate the Hierarchical Agglomerative Clustering. We compute the silhouette score for the clustering and find that it's 0.46. The Silhouette Score is a metric that evaluates how well clusters are formed in a clustering algorithm, such as K-Means or Hierarchical Clustering. It measures how similar a data point is to its own cluster (cohesion) compared to other clusters (separation).
The Silhouette Score ranges from -1 to 1, where:

1.0 → Perfect clustering (each point is well within its cluster and far from others).

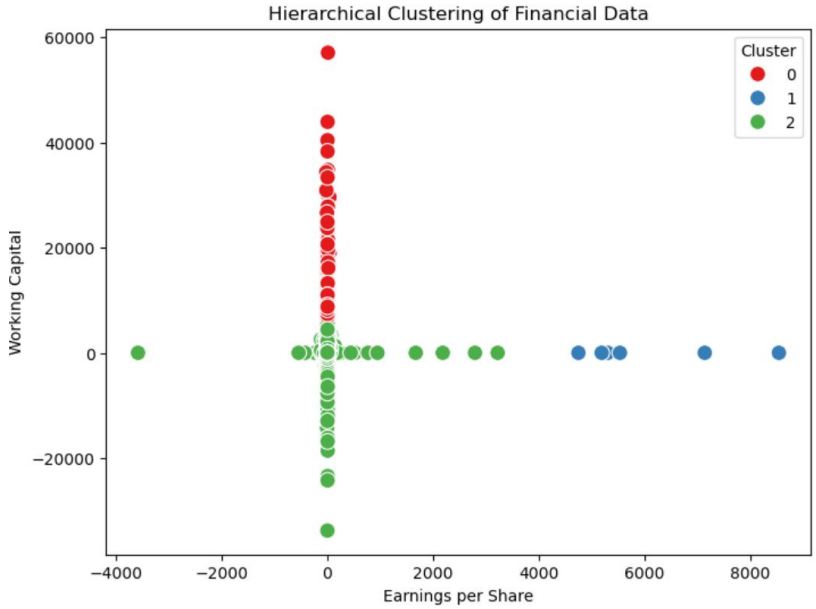
0.5 - 1.0 → Well-defined clusters.

0.2 - 0.5 → Somewhat meaningful clustering, but there may be some overlap.

0.0 - 0.2 → Weak clustering, with many points close to other clusters.

Negative values → Poor clustering (many points are misclassified).

If we use a scatterplot to visualize the clusters before applying winsorization we obtain the following plot for the hierarchical clustering.



To be able to discard the outliers from the data, the Winsorized features generate a different plot, where x is earnings per share Winsorized, and y is working capital Winsorized.

Observations from the Clustering Results

Cluster 0 (Red): Contains companies with negative Earnings per Share (EPS) and moderate to high Working Capital. Likely represents financially struggling companies or those with negative profitability but still managing working capital.

Cluster 1 (Blue): Covers a wide range of companies with moderate to positive EPS and working capital. The largest and most densely populated cluster, indicating financially stable or moderately profitable firms.

Cluster 2 (Green): Consists of companies with high EPS and high working capital. Represents financially strong and highly profitable companies.

Cluster 3 (Purple): Includes companies with low to moderate EPS but highly negative Working Capital. Represents financially risky firms, possibly experiencing liquidity issues.

Key Takeaways & Business Insights

Financially Strong vs. Weak Firms: Cluster 2 (Green) represents the strongest firms in terms of profitability and liquidity. Cluster 3 (Purple) represents the weakest firms, potentially facing financial distress.

***High-Risk vs. Low-Risk Companies*:** Companies in Cluster 0 (Red) have negative earnings but varying working capital, requiring further analysis to determine whether they are recovering or in financial trouble. Companies in Cluster 3 (Purple) have both negative earnings and poor liquidity, making them the highest-risk group.

Market Trends & Investment Strategies: Investors might prefer companies in Cluster 2 (Green) due to strong profitability and liquidity. Cluster 1 (Blue) represents a diverse set of companies, potentially including growth stocks with moderate financial stability. Cluster 0 (Red) and Cluster 3 (Purple) may contain turnaround stocks or high-risk investments.

Recommendations

For Investors: Target Cluster 2 (Green) for stable investments.
Exercise caution with Cluster 3 (Purple) due to high financial risk.
Monitor Cluster 0 (Red) firms for potential recovery or further decline.