Clustering Companies by Financial Health Using JPMorgan Chase QuantChallenge 2023 Dataset

Methods and Approach

The dataset provided by JPMorgan Chase during the 2023 Quant Challenge contains financial data for multiple stocks over time, including metrics like liabilities, equity, revenue, net income, and more. While the dataset is organized around individual stocks, the features reflect the financial condition of underlying companies. The main goal of this project is to cluster companies based on their financial health and performance using stock-level data. Since there were no predefined labels, clustering was the best choice to group companies in a meaningful way.

Data preprocessing included:

- Dropping irrelevant columns such as Date and Stock because clustering is based on financial health, not time or stock symbol.
- Normalizing the continuous features using StandardScaler to ensure all financial variables contributed equally.

For modeling, I applied two unsupervised learning algorithms:

- K-Means Clustering: I used the elbow method to find the optimal number of clusters and fit the model accordingly.
- Hierarchical Clustering: I used dendrograms to visually assess cluster distances and determine the right number of groups.

Results and Analysis

K-Means Clustering:

- Using the elbow method, the optimal number of clusters was found.
- After fitting the K-Means model, companies were assigned to different clusters based on the similarities in their financial metrics.
- A visualization of the clusters showed clear groupings based on financial characteristics like liabilities, equity, and revenue.

Hierarchical Clustering:

- A dendrogram was plotted to observe the linkage distances.
- This method provided another visualization of how companies relate to each other financially and suggested a natural separation into clusters, validating the findings from K-Means.

Feature Importance Observations:

- No feature importance ranking (since clustering is unsupervised), but through visualization and exploration, liabilities, net income, and current assets appeared to influence clustering separation the most.

K-Means Silhouette Score: 0.1681088062988105

Hierarchical Silhouette Score: 0.12725926833984824

What this means:

- Both scores are below 0.25, which indicates the clusters are not very well-separated.
- But K-Means performed slightly better than Hierarchical Clustering (0.168 vs 0.127). This also visually matches the PCA plot: K-Means clusters looked more compact and distinct.

Comprehensive Conclusion

In this project, I applied unsupervised learning techniques (KMeans and Hierarchical Clustering Algorithms) to cluster companies based on their financial data. Although the dataset was organized around stocks, these features reflect the financial behavior of the companies themselves.

Using K-Means and Hierarchical Clustering, I identified four distinct groups of companies, Each cluster represents different financial profiles — from high-performing firms with strong revenue and dividend payouts to lower-performing companies with negative net income and minimal returns. This grouping was validated using PCA visualizations, Cluster Summaries, and Silhouette Scores which revealed meaningful insights into company performance. They showed clear separation among clusters in K-Means and slightly more overlap in Hierarchical Clustering.

The main findings were:

- Companies naturally form groups based on financial indicators like liabilities, equity, and revenue.
- The elbow method and dendrograms both supported consistent cluster separations.
- K-Means and Hierarchical Clustering produced complementary insights, giving confidence in the groupings.

These clusters can serve as a foundation for further investment analysis, helping investors target companies that match specific financial strategies, risk levels, or growth potential as resulted by the three validation methods.