ML - Lab Assignment 6

Name: Jinal Sasiya
ID: 202518062

September 26, 2025

1 Analysis and Discussion

1.1 Optimal Clusters

For K-Means clustering, the elbow method suggested that 5 clusters is optimal. The WCSS curve displayed a clear bend at this point, indicating a good balance between cluster compactness and model simplicity. In the case of Hierarchical Clustering, the dendrogram suggested that 3 clusters would be the most appropriate.

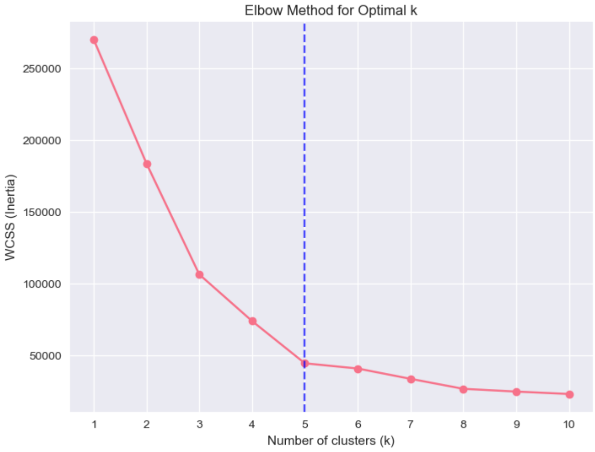

Figure 1: Elbow method for optimal k

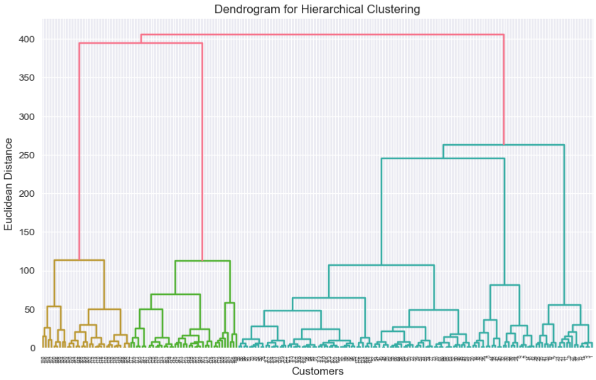

Figure 2: Dendrogram for Hierarchical Clustering

The optimal clusters were identified by cutting the dendrogram at the longest vertical distance without crossing any horizontal lines, which separated the data into three distinct groups. These approaches allowed us to determine the most meaningful number of clusters for each algorithm based on the data's structure.

1.2 Cluster Comparison

When comparing the clustering outcomes, K-Means generated five fairly balanced groups, giving a detailed segmentation of customers. Hierarchical Clustering produced a broader division into three larger categories, providing a high-level summary of the dataset. DBSCAN, on the other hand, detected three compact clusters and labeled around **17 points as noise**. This distinction arises because DBSCAN ignores sparse observations, while K-Means and Hierarchical clustering assign every point to a cluster. Hence, the three methods highlight different perspectives: fine-grained, coarse, and density-based with outlier detection.

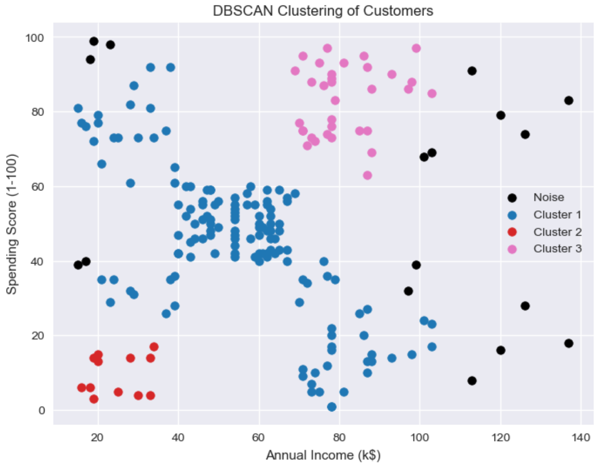

Figure 3: DBSCAN Clustering

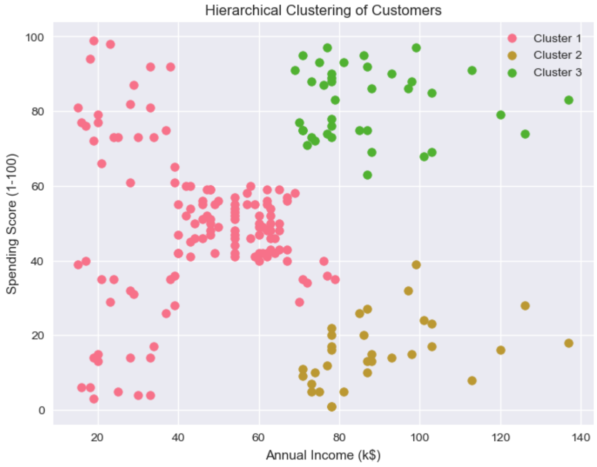

Figure 4: Hierarchical Clustering

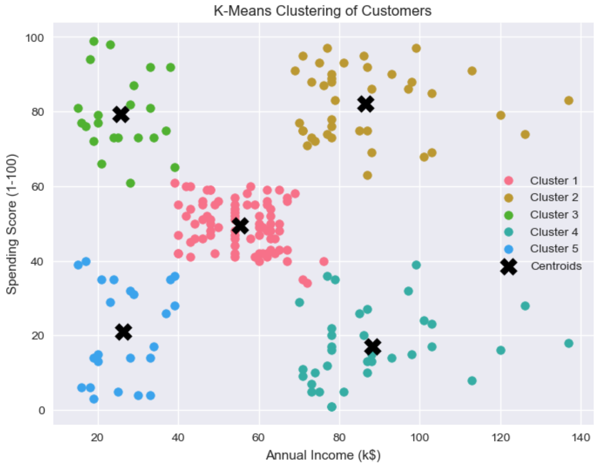

Figure 5: K-Means Clustering

1.3 DBSCAN Performance

DBSCAN was effective in finding dense areas of customers. It produced three main groups of roughly 138, 12, and 33 members, and identified 17 records as noise. Unlike the other two techniques, DBSCAN leaves unusual points unassigned, making it useful for spotting outliers. A limitation, however, is that the results depend heavily on the choice of parameters (e.g., $\varepsilon = 0.5$, $minPts = 14$), and the method can produce small, fragmented clusters if the settings are not chosen carefully.

1.4 Algorithm Suitability

Taking into account the shape of the clusters, their density, and practical application, K-Means appears to be the most reliable option for this dataset. While DBSCAN does highlight anomalies and can form irregular-shaped clusters, its sensitivity to parameters makes it less consistent. Hierarchical Clustering is useful for understanding the overall structure, but with only three groups it is less informative for marketing applications. K-Means, by contrast, gives five interpretable and evenly distributed clusters, which are easier to use for customer segmentation. The silhouette scores were as follows:

- K-Means: 0.554
- Hierarchical: 0.462
- DBSCAN: 0.375

1.5 Real-World Application

The clusters derived here can support targeted marketing strategies for the mall. For instance:

- **Wealthy, high-spending customers**: can be encouraged with premium memberships, loyalty bonuses, or exclusive promotions.
- **Wealthy but low-spending customers**: may respond better to discounts, bundles, or personalized offers designed to boost spending.

- **Lower-income yet high-spending customers**: could be engaged with affordable product ranges, seasonal discounts, or installment schemes.
- **Noise points/outliers**: represent atypical patterns such as bulk buyers or seasonal visitors, which may require individual attention.

Such tailored approaches allow the business to reach different customer segments effectively and maximize overall profitability.