ML-Lab Assignment 6

Name: Sanjana Nathani
ID: 202518002

September 26, 2025

1 Analysis and Discussion

1.1 Optimal Clusters

For K-Means clustering, as the elbow method showed the optimal clusters as 5. As the WCSS curve showed a clear bend at that point, indicating a good balance between compactness and simplicity. For Hierarchical Clustering, the dendrogram indicated 3

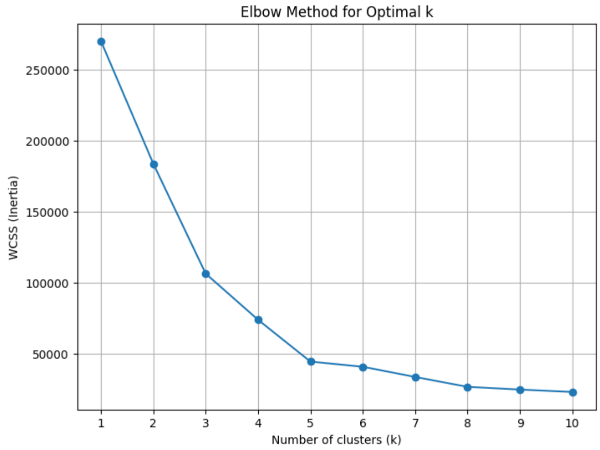

Figure 1: Elbow method for Optimal k

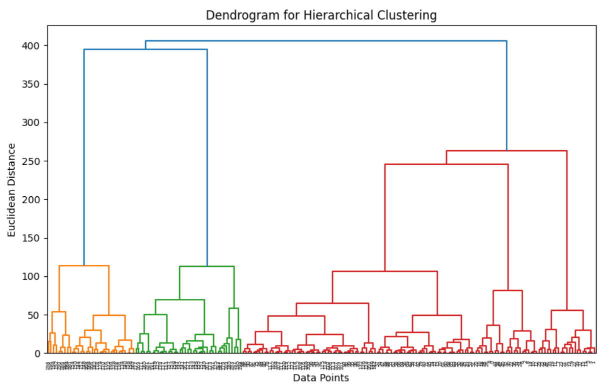

Figure 2: Dendrogram for Hierarchical Clustering

clusters as the best choice, identified by cutting the tree at the longest vertical distance without crossing any horizontal lines, which separated the data into three distinct groups. These methods helped us determine the most meaningful number of clusters for each algorithm based on the structure of the data.

1.2 Cluster Comparison

Visually comparing the results of the three algorithms, we observe that K-Means produced five distinct clusters with relatively balanced sizes, capturing finer granularity in customer segmentation. Hierarchical Clustering resulted in three larger clusters, giving a broader overview of the data. DBSCAN identified three clusters along with 17 noise points, focusing on dense regions and labeling sparse points as noise. Overall, the shape and density of clusters differ between algorithms, with DBSCAN highlighting anomalies that the other methods assign to clusters.

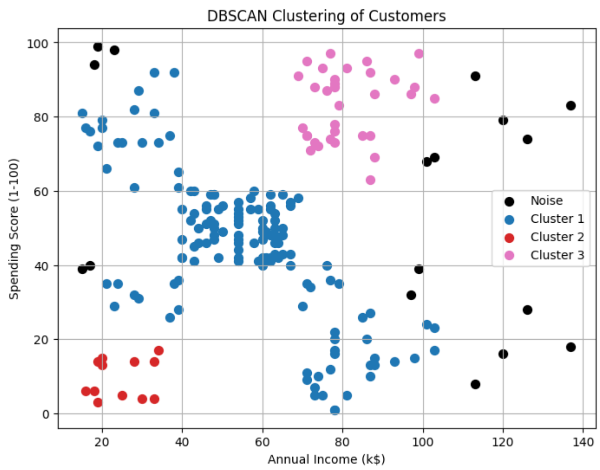

Figure 3: DBSCAN Clustering

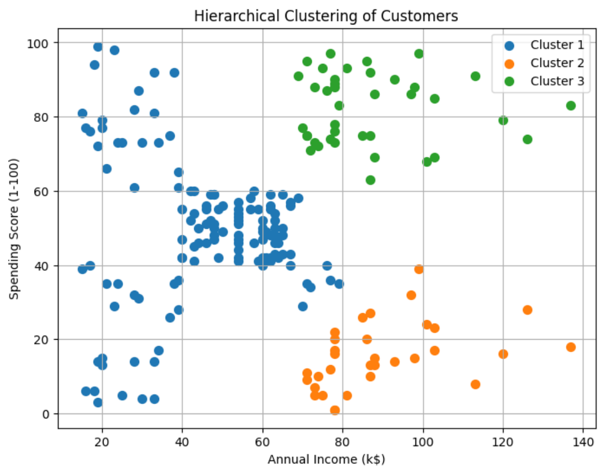

Figure 4: Hierarchical Clustering of Customers

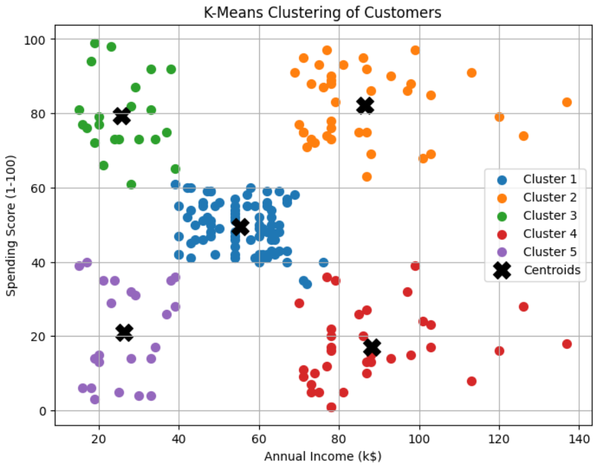

Figure 5: KMeans Clustering

1.3 DBSCAN Performance

DBSCAN performed well in identifying dense clusters, producing three main clusters containing 138, 12, and 33 points respectively, while marking 17 points as noise. Unlike K-Means and Hierarchical Clustering, which force all points into clusters, DBSCAN effectively isolates outliers, providing insights into unusual customer behavior. However, it is sensitive to parameter selection ($\epsilon = 0.5$, min_samples = 14), and some clusters may be smaller than expected.

1.4 Algorithm Suitability

Considering cluster shapes, density, and practical business application, K-Means is the most suitable algorithm for this dataset. Although DBSCAN achieved a slightly higher silhouette score for dense clusters, K-Means provides a predictable number of clusters (5), assigns all customers to a cluster, and produces balanced, interpretable clusters. Hierarchical Clustering, while useful for visualization and understanding overall data structure, produces fewer clusters (3) and lacks the granularity needed for actionable segmentation. The silhouette scores for each algorithm are as follows:

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

1.5 Real-World Application

The identified clusters can be leveraged by the mall's marketing team for targeted campaigns. For example:

- **High-income, high-spending customers:** Offer premium products, loyalty programs, and exclusive promotions.
- **High-income, low-spending customers:** Send personalized discounts or bundled offers to encourage higher spending.

- **Low-income, high-spending customers:** Focus on budget-friendly deals or installment options to maintain loyalty.
- **Outliers (DBSCAN noise points):** Investigate individually to understand unusual purchasing behaviors, which may represent bulk buyers or seasonal visitors.

This segmentation allows the mall to design marketing strategies tailored to each group, optimizing customer engagement and revenue.