Customer Segmentation using Clustering Algorithms

Katrodiya Het Rupeshbhai
202518005

September 26, 2025

Abstract

This report presents an analysis of customer segmentation using three clustering techniques: K-Means, Hierarchical Clustering, and DBSCAN. The study explores the determination of optimal clusters, compares algorithm performance, and evaluates the suitability of each method for mall customer data. Practical insights are also discussed by linking cluster outcomes with real-world marketing applications.

1 Introduction

Clustering is a widely used unsupervised learning technique to group customers with similar purchasing behavior. In this assignment, I applied K-Means, Hierarchical Clustering, and DBSCAN to mall customer data. The aim was to identify meaningful clusters, analyze their performance, and consider real-world applications of these insights.

2 Results and Discussion

1. Optimal Clusters

For K-Means, the Elbow Method clearly indicated an optimal cluster count of five, as the within-cluster sum of squares showed a noticeable bend at $k = 5$. Similarly, the dendrogram from Hierarchical Clustering also supported five clusters, based on the longest vertical distance before merges. Hence, both methods consistently suggested five groups.

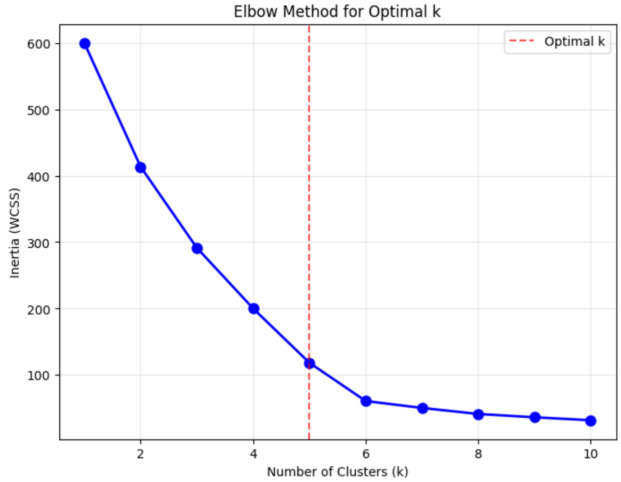

Figure 1: Elbow Method showing optimal cluster number

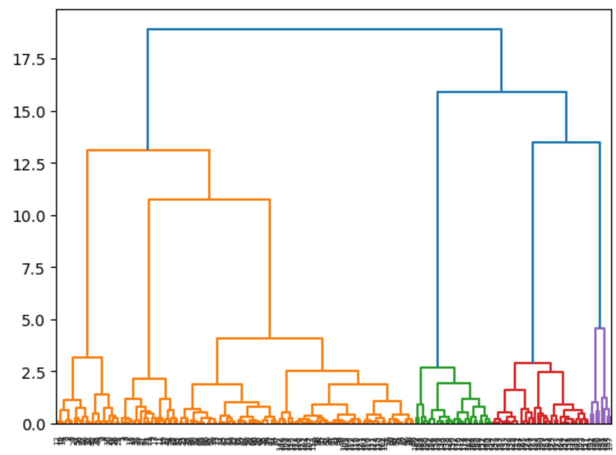

Figure 2: Hierarchical Clustering dendrogram indicating 5 clusters

2. Cluster Comparison

When comparing the visual results, K-Means and Hierarchical Clustering produced fairly similar and compact clusters. However, minor differences in boundary allocation were observed. DBSCAN, on the other hand, formed clusters of varying shapes and marked a few data points as noise, which neither of the other two algorithms did. This indicates DBSCAN's flexibility but also its sensitivity to parameter choices.

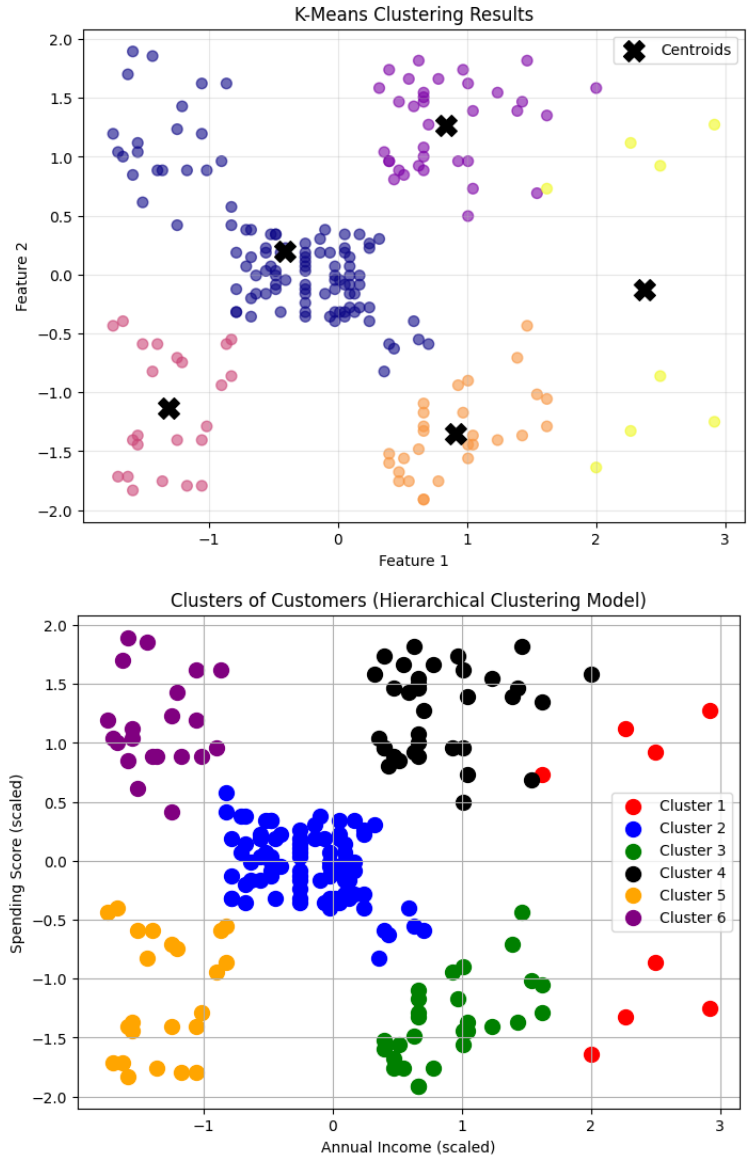

Figure 3: Comparison of K-Means and Hierarchical Clustering results

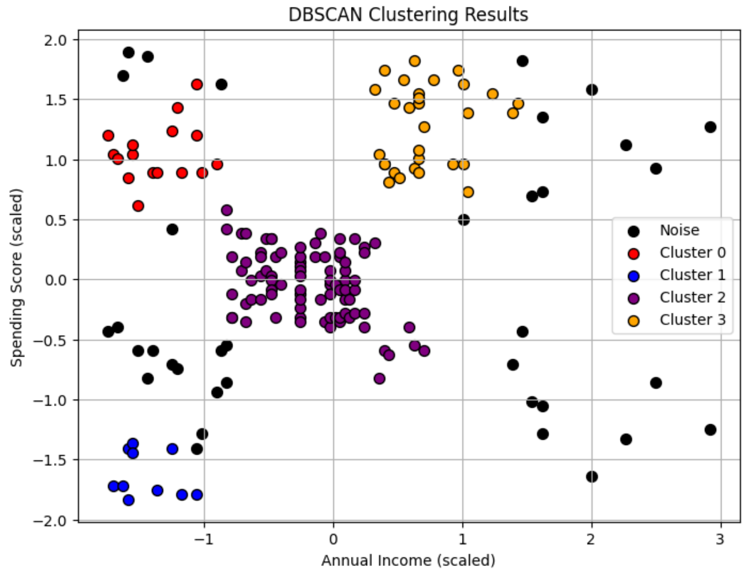

Figure 4: DBSCAN clustering with noise points identified

3. DBSCAN Performance

DBSCAN managed to detect clusters with irregular structures and successfully labeled some customers as outliers. Unlike K-Means or Hierarchical Clustering, it does not force every observation into a cluster, making it useful for identifying unusual customer behavior. However, it required careful tuning of `eps` and `min_samples` to produce meaningful clusters.

4. Algorithm Suitability

Considering the results, K-Means turned out to be the most appropriate for this dataset because the clusters were well-separated, compact, and easy to interpret. Hierarchical Clustering performed almost equally well but is computationally more intensive. DBSCAN was valuable for outlier detection but less consistent due to sensitivity in parameter settings.

5. Real-World Application

The segmentation results have practical marketing implications. For instance:

- **High Income, Low Spending Score:** Target with loyalty programs, premium offers, or personalized promotions to increase engagement.
- **Low Income, High Spending Score:** Attract with seasonal discounts and bundle offers to maximize spending potential.
- **High Income, High Spending Score:** Position luxury products, VIP memberships, and early-access sales.
- **Low Income, Low Spending Score:** Use basic campaigns such as product sampling or festival discounts to encourage purchases.

3 Conclusion

The analysis demonstrated that five clusters best represented the customer groups in the dataset. K-Means provided the most balanced and interpretable results, while DBSCAN added the advantage of detecting noise and irregularly shaped clusters. These insights can help mall management and marketing teams design targeted strategies for different customer groups.