Clustering Algorithms
DS605 - Lab 6 Report

Anusha Singh - 202518041

26 September,2025

Abstract

This report presents an analysis of the Mall Customers dataset using three clustering algorithms: K-Means, Agglomerative Hierarchical Clustering, and DBSCAN. We evaluate the performance of each method, compare the clusters they produce, and justify our choices using silhouette scores, the Elbow Method, and dendrogram visualizations.

1 Introduction

Customer segmentation is a critical task in retail analytics, enabling targeted marketing campaigns and personalized services. We focus on two features: Annual Income (k$) and Spending Score (1–100). Clustering these dimensions helps us discover meaningful customer groups.

2 Methodology

2.1 Clustering Methods

- **K-Means:** To choose an optimal value of k in this lab we use the elbow method and silhouette scores.
- **Hierarchical Clustering:** To choose the choice of cluster we used Dendogram and Ward's linkage.
- **DBSCAN:** Parameter tuning was done for `eps` and `min_samples` with evaluation using silhouette score.

3 Results and Discussion

3.1 1. Optimal Clusters

An optimal cluster n=5 was determined for both K-Means and Hierarchical Clustering. For K-Means, this choice was supported by the Elbow Method and silhouette scores (Figure 1). For Hierarchical Clustering, the dendrogram similarly revealed that a five-cluster split was the most natural data partition (Figure 2).

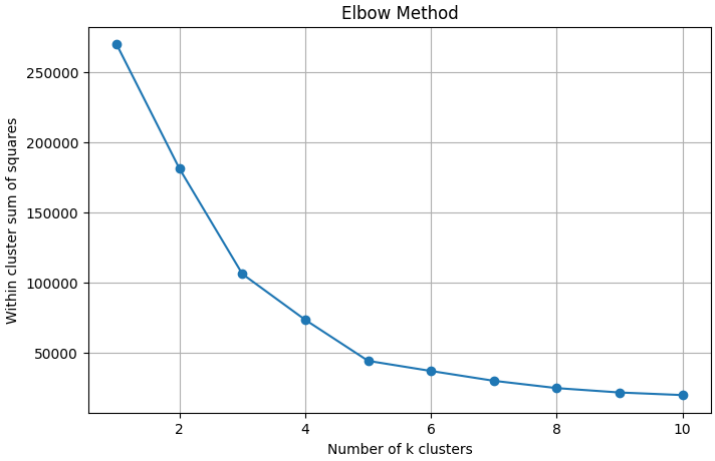

Figure 1: Elbow Method for K-Means showing the optimal $k = 5$.

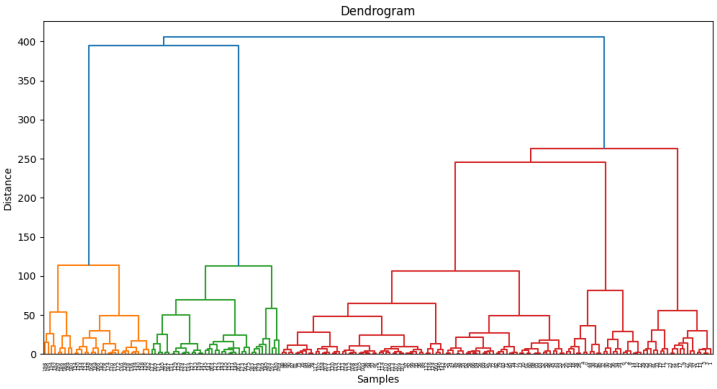

Figure 2: Dendrogram (Ward linkage) suggesting 5 clusters.

3.2 2. Cluster Comparison

As shown in (Figure 3), both K-Means and Hierarchical Clustering successfully identified 5 compact, yielding nearly identical results. In contrast, DBSCAN's density-based approach proved less stable; depending on the parameters, it either merged all points into one cluster or formed 3-4 groups while flagging some data points as outliers (noise), a unique feature of this algorithm.

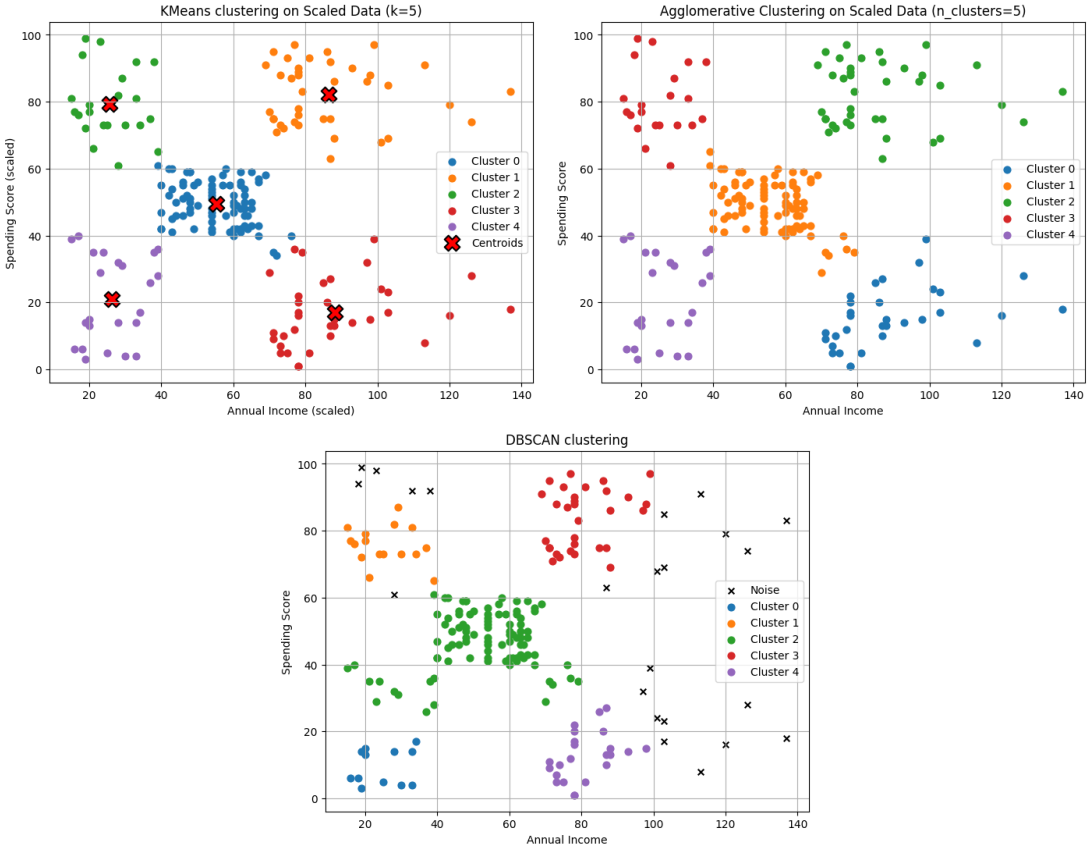

Figure 3: Comparison: K-Means vs. Agglomerative Hierarchical vs DBSCAN Clustering.

3.3 3. DBSCAN Performance

DBSCAN's performance was notably weaker on this dataset (see Figure 4). It failed to discern significant density variations, often collapsing all customers into a single group. While powerful for datasets with arbitrarily shaped clusters or significant noise, DBSCAN was ill-suited for this problem. The natural customer segments here were compact and well-defined, a structure better captured by partitioning methods like K-Means and Hierarchical Clustering, which consistently identified clear boundaries.

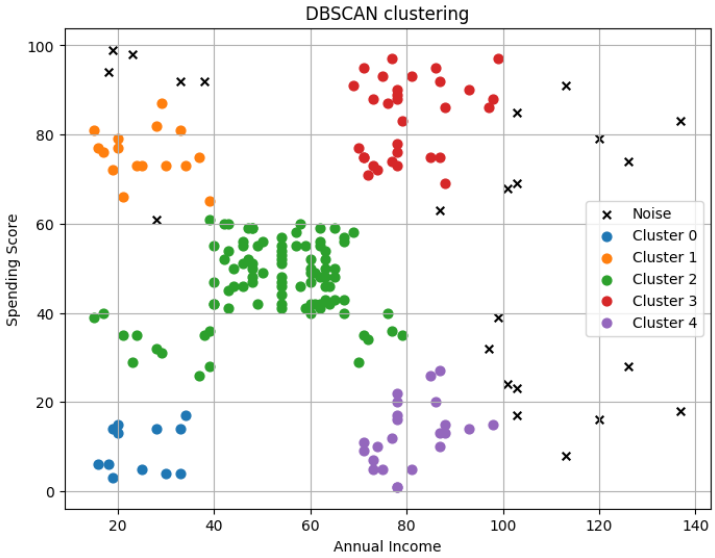

Figure 4: DBSCAN clustering result (example parameters with silhouette > 0.5).

3.4 4. Algorithm Suitability

In summary, K-Means and Hierarchical Clustering were better suited for this dataset due to its compact and clearly separated cluster structure. DBSCAN, which excels at identifying non-spherical clusters and outliers, was less effective because the data lacked the complex density patterns it is designed to detect.

3.5 5. Real-World Application

In finance and security, clustering is used to establish baseline patterns of normal behavior in order to automatically detect suspicious deviations.

- Fraud Detection by flagging transactions that fall outside of a customer's typical spending habits.
- Credit Risk Assessment by grouping customers into different risk profiles to automate loan and credit decisions.
- Network Security by identifying unusual network traffic that could signal a cyber-attack

4 Conclusion

In conclusion, this project successfully demonstrated how clustering algorithms can transform raw data into actionable customer segments. K-Means and Hierarchical Clustering both effectively identified five distinct groups with strong validation scores. In contrast, DBSCAN was less suitable and required extensive tuning. The strategic application of such segments allows for targeted marketing that can simultaneously enhance customer satisfaction and drive revenue.