Clustering Analysis of Mall Customers Dataset

Urvi Kava

202518006

September 26, 2025

1 Plots

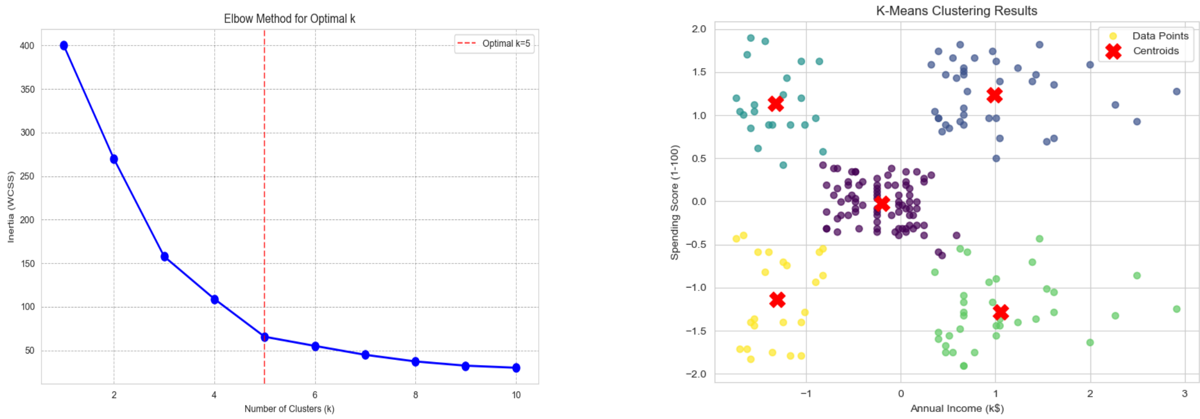

Figure 1: Elbow Method and K-Means Clustering

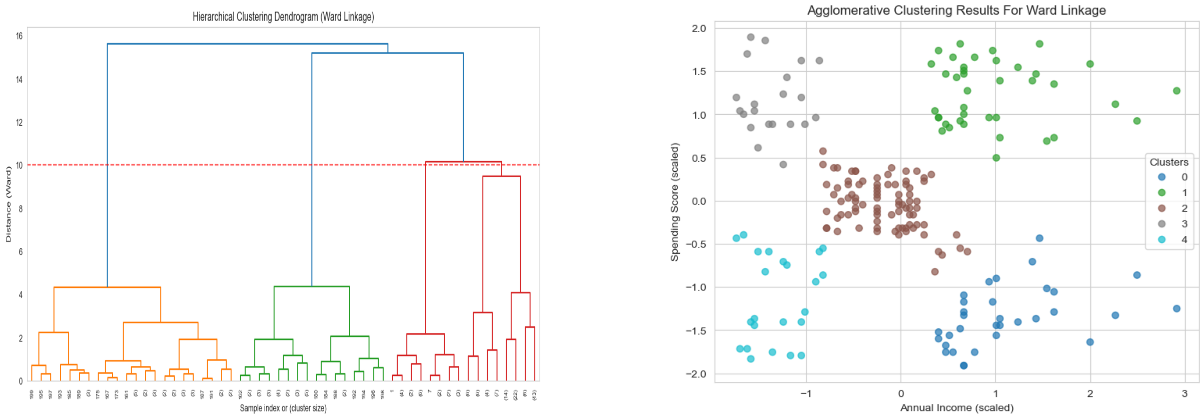

Figure 2: Dendrogram and Agglomerative Clustering

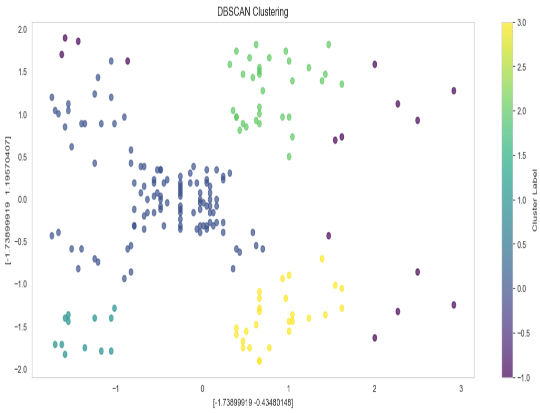

Figure 3: DBSCAN Clustering

2 Optimal Clusters

To determine the optimal number of clusters for **K-Means**, the *Elbow Method* was applied. The Within-Cluster Sum of Squares (WCSS) was plotted for values of k between 1 and 10. The plot exhibited a clear elbow at **5 clusters**, which was selected as the optimal value.

For **Agglomerative Hierarchical Clustering**, the dendrogram was generated using Ward linkage. By examining the longest vertical line that does not intersect any horizontal linkages, the optimal number of clusters was also found to be **5**. This consistency with the K-Means result reinforces the suitability of choosing five distinct customer segments.

3 Cluster Comparison

When visually comparing the clustering results:

- **K-Means** produced well-separated, compact clusters with clear boundaries.
- **Hierarchical Clustering** resulted in clusters similar to K-Means but with slight overlaps on the edges. This is due to its bottom-up merging strategy.
- **DBSCAN** identified clusters of varying shapes and densities. Unlike K-Means and Hierarchical, DBSCAN detected some points as noise (unclustered), which is useful when data contains outliers.

Overall, K-Means and Hierarchical produced broadly similar results, while DBSCAN highlighted the density variations.

4 DBSCAN Performance

Unlike K-Means and Hierarchical Clustering, **DBSCAN** does not require the number of clusters to be specified. Instead, it identifies dense regions based on two parameters: $eps = 0.367$ and $min\_samples = 4$. For this dataset, DBSCAN detected fewer clusters (around 3–4) and classified several customers as **noise points** (labelled as -1).

This differs from K-Means and Hierarchical Clustering, which force every data point into a cluster. DBSCAN was effective in identifying customers with atypical spending and income patterns, but it was less effective at producing evenly distributed customer segments suitable for marketing applications.

5 Algorithm Suitability

Based on the analysis, **K-Means** proved to be the most suitable algorithm for this dataset. The reasons are:

- The dataset exhibits relatively spherical and well-separated clusters, which align with K-Means' assumptions.
- The elbow method provided a clear and interpretable choice of cluster number.
- The resulting clusters were balanced and practical for customer segmentation, unlike DBSCAN, which produced uneven clusters and noise.

Hierarchical clustering also performed well, but K-Means offers greater scalability and interpretability for larger datasets.

6 Real-World Application

The customer segments identified can directly support a mall's marketing strategy. For example:

- **High Income, Low Spending Score:** These customers have financial capacity but spend cautiously. Targeted promotions, loyalty rewards, or exclusive events could encourage higher engagement.
- **High Income, High Spending Score:** Premium customers who are likely to respond to luxury product launches, VIP experiences, and personalized offers.
- **Low Income, High Spending Score:** Budget-conscious but active spenders who can be targeted with discounts, seasonal sales, and affordable product lines.
- **Moderate Segments:** Middle-income groups with varying spending patterns can be approached with general mall-wide offers and family-oriented promotions.

Such segmentation allows the marketing team to allocate resources efficiently and tailor campaigns for maximum impact.

7 Conclusion

The clustering analysis of the Mall Customers dataset revealed five optimal clusters using K-Means and Hierarchical Clustering, while DBSCAN identified fewer clusters with some noise. K-Means was determined to be the most suitable algorithm for this dataset, providing clear and actionable customer segments that can enhance targeted marketing strategies.