Clustering Analysis: Mall Customers Dataset

Patel Harsh Satishkumar
202518011

September 26, 2025

Part 5: Analysis and Questions

Plots

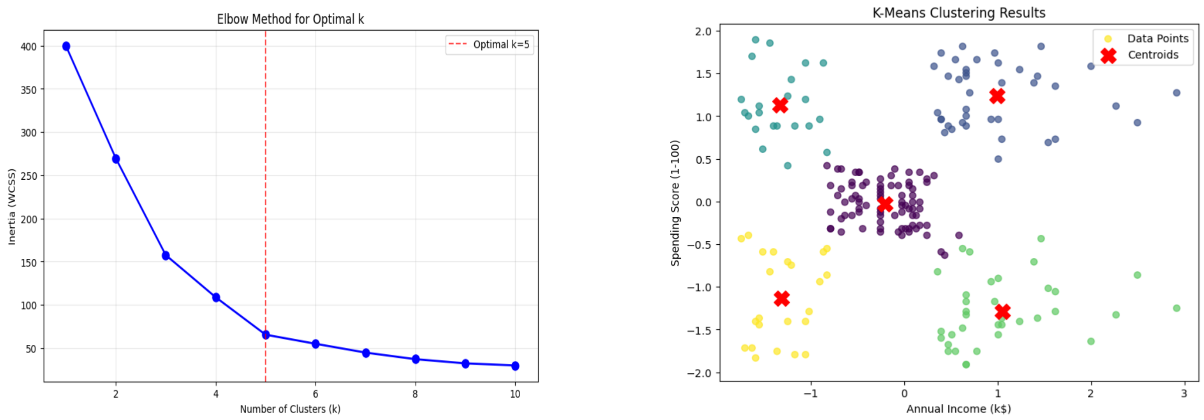

Figure 1: Elbow Method and K-Means Clustering

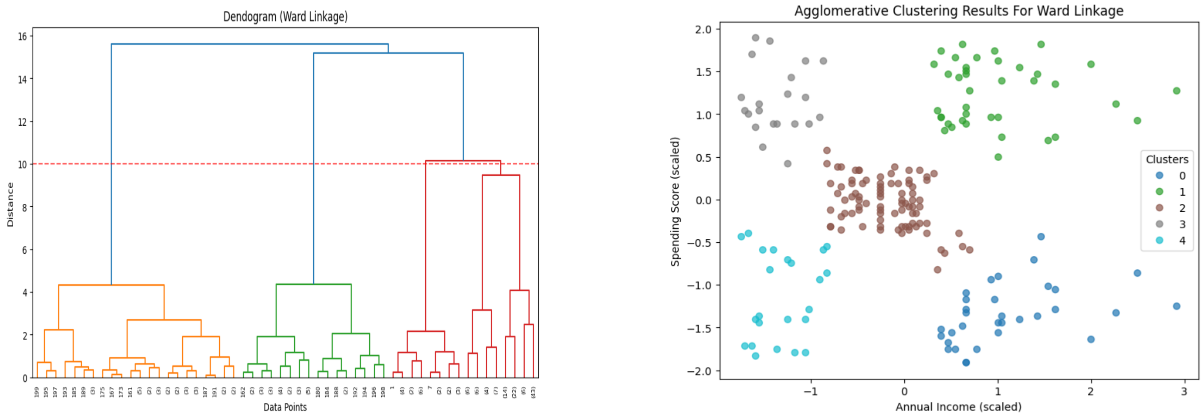

Figure 2: Dendogram and Agglomerative Clustering

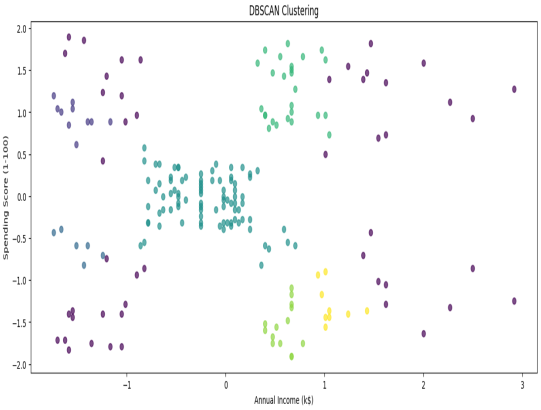

Figure 3: DBSCAN Clustering

1. Optimal Clusters

For the **K-Means algorithm**, the Elbow Method indicated that the optimal number of clusters was **5**. The Within-Cluster Sum of Squares (WCSS) showed a sharp decline up to $k = 5$, after which the reduction rate flattened, forming the "elbow."

For **Hierarchical Clustering**, the dendrogram also supported the presence of **5 clusters**. The longest vertical line without intersections was cut at this level, confirming that five natural groupings exist in the data.

Thus, both methods consistently suggest that **5 clusters** provide the most meaningful segmentation.

2. 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.

3. DBSCAN Performance

DBSCAN performed reasonably well on this dataset. With parameters such as `eps` $= 5$ and `min_samples` $= 5$, it identified the main dense clusters correctly but also labeled a few customers as **noise points** (label -1).

Unlike K-Means and Hierarchical Clustering, which force every customer into a cluster, DBSCAN allows for flexibility by excluding outliers. This makes it more realistic in scenarios where not all customers fit neatly into a segment.

4. Algorithm Suitability

Based on the dataset characteristics:

- **K-Means** was the most suitable algorithm. The clusters were spherical, compact, and balanced, aligning well with the algorithm's assumptions.
- **Hierarchical Clustering** provided similar results but was more computationally expensive and slightly less precise in boundary formation.
- **DBSCAN** was less suitable here because the dataset did not contain complex cluster shapes or high noise. However, it was useful in detecting outliers.

Thus, K-Means is the preferred choice for this dataset.

5. Real-World Application

In a real-world mall marketing scenario, the identified clusters could represent distinct customer segments such as:

- **High Income, Low Spending Score:** These customers are wealthy but cautious. Marketing could target them with premium loyalty programs, exclusive product launches, or personalized offers to encourage spending.
- **Low Income, High Spending Score:** These are enthusiastic spenders. Promotions, discounts, and bundle deals would resonate with them.
- **High Income, High Spending Score:** These are the ideal customers who can be targeted with luxury goods, VIP experiences, and long-term engagement strategies.
- **Low Income, Low Spending Score:** They may be less profitable segments, but targeted campaigns such as seasonal discounts could still drive engagement.

Such segmentation enables the marketing team to design personalized strategies, improving customer satisfaction and increasing mall revenue.