Analysis and Questions

Harsh Jethwani: 202518055

September 26, 2025

1) Optimal Clusters

The optimal number of clusters identified is 5 for both K-Means and Hierarchical Clustering.

Justification:

- K-Means (Elbow Method): The "Elbow Method for Optimal k" plot shows the Within-Cluster Sum of Squares (WCSS) versus the number of clusters (k). The "elbow" point, where the rate of decrease in WCSS significantly slows down, occurs at k=5. This suggests that adding more than five clusters doesn't provide a substantial reduction in the within-cluster variance.
- Hierarchical Clustering (Dendrogram): The Dendrogram shows the Euclidean distances at which clusters are merged. To determine the optimal number of clusters, one typically draws a horizontal line across the largest vertical drop that doesn't intersect a pre-existing cluster. In the "Dendrogram," if a horizontal line is drawn around a distance of 200 to 250, it would intersect the vertical lines at 5 points, thus separating the data into 5 main clusters (the two large initial clusters on the left and the three large initial clusters on the right).

2) Cluster Comparison

Visual Comparison and Differences:

- K-Means: Clusters are roughly spherical and centered around the red Centroids. The Notable difference here is that the clusters are forced to be similar in size and density due to the distance-to-centroid measure.
- Hierarchical: Five distinct, well-separated, and visually interpretable clusters (Red, Blue, Green, Cyan, Magenta). This is the most visually coherent and well-defined clusters that match the natural groupings in the scatter plot. It clearly separates the data into the five corners/regions and the center.

- DBSCAN: A large number of points are classified as Noise (black dots). It fails to cleanly define the five distinct clusters seen in the other methods, instead focusing heavily on the central density.
- The Similarity is that both K-Means and Hierarchical produced 5 clusters that strongly align with the five characteristic regions on the scatter plot.

3) DBSCAN Performance

DBSCAN performed poorly at identifying the five expected, relatively distinct groupings in this dataset.

- Noise Points: Yes, DBSCAN identified a large number of points as Noise (black dots), which are essentially the points that are not part of any sufficiently dense region. This includes many points that K-Means and Hierarchical clustering assigned to the corners of the plot.
- Comparison to Other Methods:
  - K-Means and Hierarchical Clustering force every point into one of the k=5 clusters, leading to complete segmentation of the customer base.
  - DBSCAN's density-based approach only clustered the most dense part of the data (the center) effectively. The sparser, but still distinctly grouped, customer segments in the corners (like the "Careful Spenders" and "Target Group") were largely labeled as noise, making the segmentation less useful for a complete customer profile.

4) Algorithm Suitability

Hierarchical Clustering was the most suitable algorithm for this specific dataset.

- Reasoning: Hierarchical Clustering (as seen in the "Clusters of Customers" plot) successfully identified the five natural, well-separated, and non-spherical groupings of the data.
- While K-Means also produced 5 clusters, its tendency to form spherical clusters is less ideal for the slightly elongated or box-like shape of the central cluster.
- DBSCAN failed to cluster the sparser, yet important, corner groups.

5) Real-World Application

The five segments identified (best seen in the Hierarchical/K-Means results) represent distinct customer types for the mall:

- The Segments are:

  1. Low Income, Low Spending (Cluster: Magenta/Yellow): Cautious Customers
  2. Low Income, High Spending (Cluster: Cyan/Green-left): Impulsive Customers
  3. Average Income, Average Spending (Cluster: Blue/Purple): Moderate Customers
  4. High Income, Low Spending (Cluster: Red/Green-bottom): Rich but Careful Spenders
  5. High Income, High Spending (Cluster: Green/Blue-top): Target Group/Big Spenders

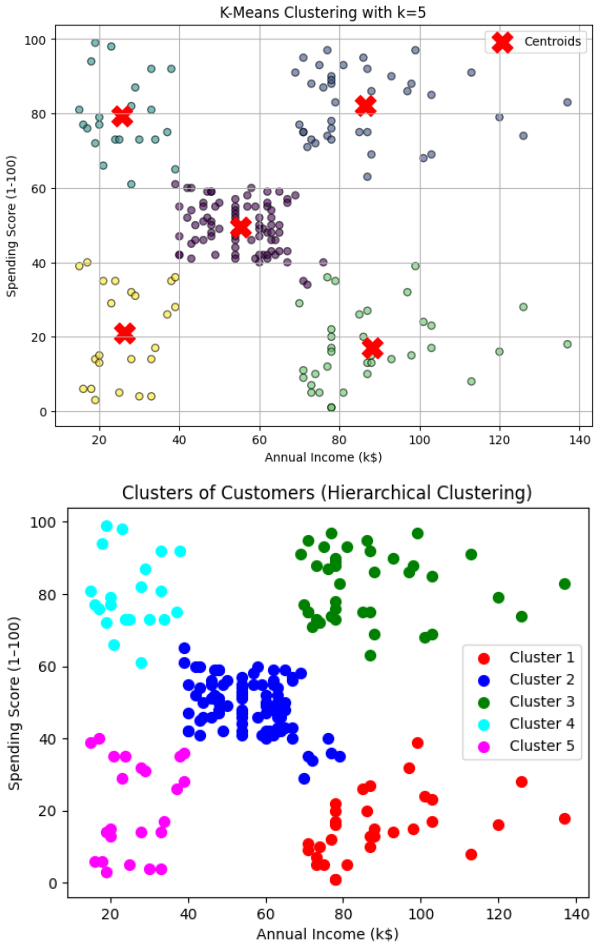

Figure 1: Cluster visualization results

- Targeting the High Income, Low Spending Group:

– Given this group's high income, we need to incentivize a purchase, so we will be employing the Fear of Missing Out (FOMO) strategy. Time-limited services or products that appeal to their inherent desire for quality and value (not cheapness) can be a good tactic to convert their income into spending by emphasizing scarcity.

– Offering loyalty programs that provide high-value perks that focuses on the value of a product rather than mere discounts is another interesting strategy worth exploring.