Machine Learning (Clustering)

Manya Choradiya

September 26, 2025

1 Analysis and Questions

1.1 Optimal Clusters: What was the optimal number of clusters you identified for K-Means and Hierarchical Clustering? Justify your choices using the Elbow Method plot and the Dendrogram.

K-Means: 5
Using the Elbow Method plot, the WCSS (Within-Cluster Sum of Squares) drops sharply until 5 clusters and then levels off. The "elbow" indicates that adding more clusters beyond 5 does not significantly improve compactness, so 5 clusters are optimal.

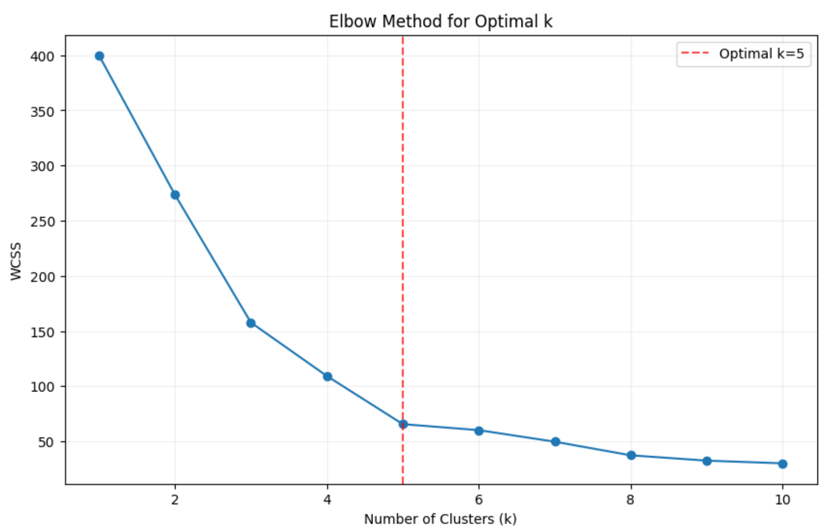

Figure 1: Elbow Method plot for K-Means.

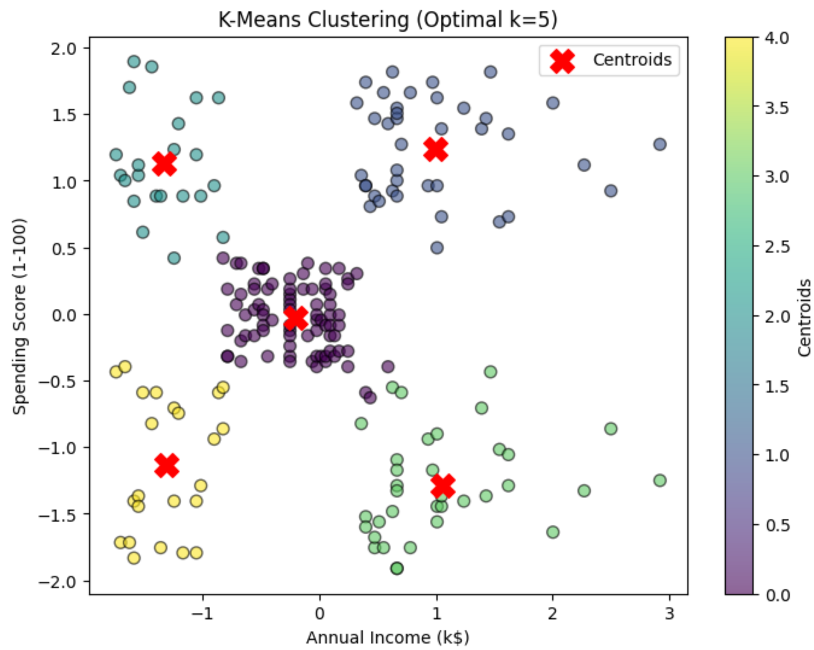

Figure 2: K-Means clustering results.

Hierarchical Clustering: 5

From the Dendrogram, a horizontal cut at a height that separates the data into 5 clusters keeps similar points together without creating too many small clusters. This natural separation makes 5 clusters meaningful.

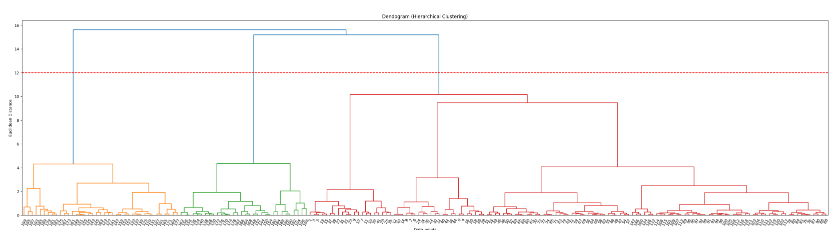

Figure 3: Dendrogram for Hierarchical Clustering.

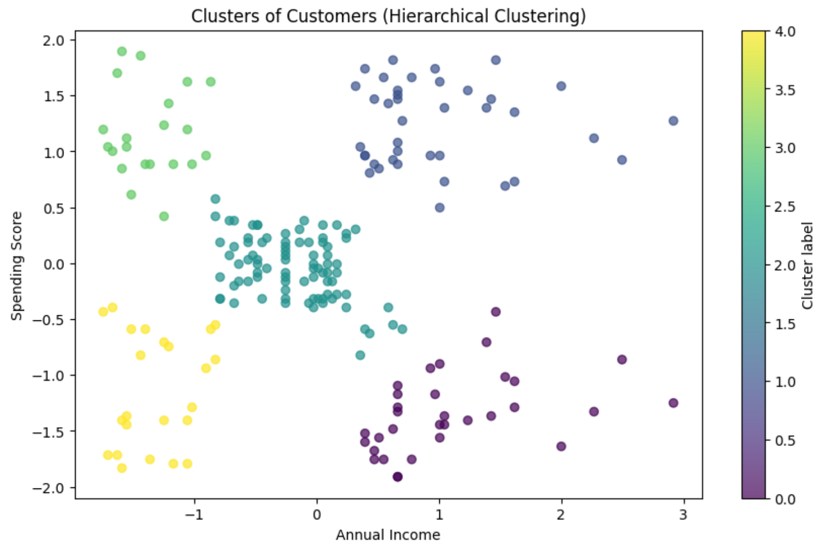

Figure 4: Hierarchical clustering results.

1.2 Cluster Comparison: Visually compare the results of the three algorithms. Did they produce similar clusters? Describe any notable differences.

All three algorithms identified similar customer groups. K-Means and Hierarchical Clustering both produced 5 well-defined clusters that mostly overlap. DBSCAN, however, identified some points as noise and did not force all points into clusters, resulting in slightly irregular-shaped clusters compared to K-Means and Hierarchical Clustering.

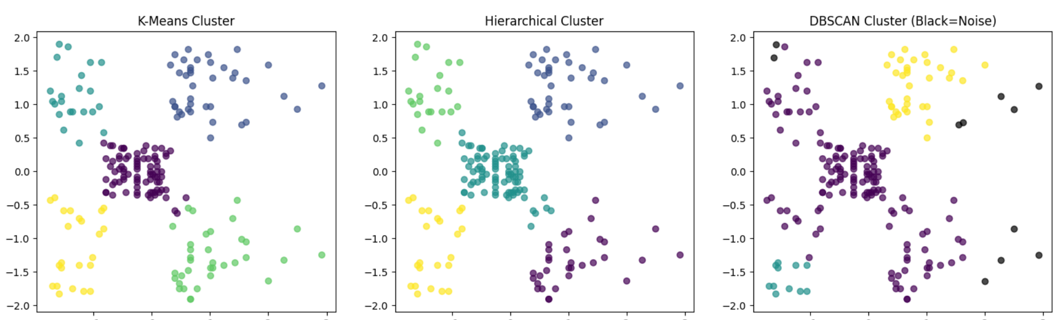

Figure 5: Visual comparison of clusters from K-Means, Hierarchical Clustering, and DBSCAN (black = noise).

1.3 DBSCAN Performance: How did DBSCAN perform on this dataset? Did it identify any noise points? How did its results compare to the other methods which force every point into a cluster?

DBSCAN performed well in identifying dense clusters and separating out noise points. Some points were labeled as noise (cluster -1), which shows that they do not belong to any natural group. Unlike K-Means and Hierarchical Clustering, which assign every point to a cluster, DBSCAN leaves the outliers unassigned, making it more robust to unusual data points.

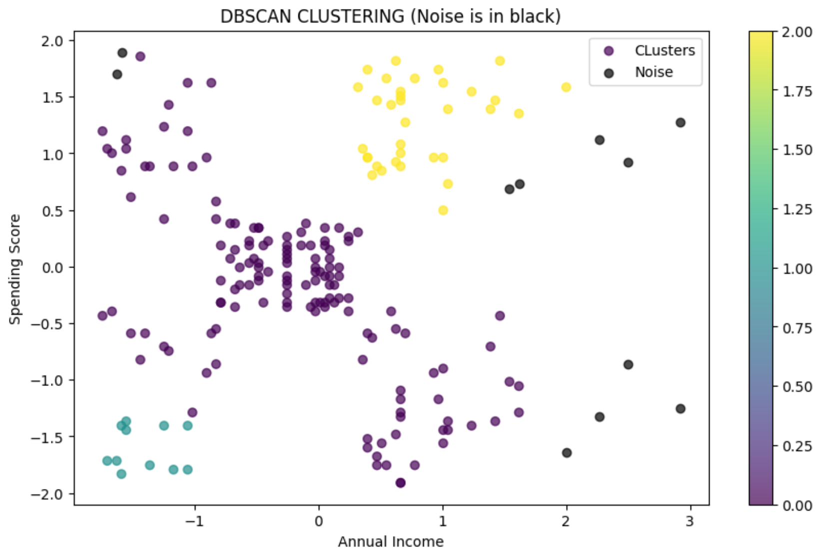

Figure 6: DBSCAN clustering with noise points in black.

1.4 Algorithm Suitability: Based on your results, which algorithm do you think was most suitable for this specific dataset and why? Consider the shape and density of the clusters.

K-Means is the most suitable algorithm for this dataset because the clusters are roughly spherical and well-separated.
Hierarchical Clustering produces similar results but is more computationally costly.
DBSCAN is useful for detecting outliers, but since most groups are of similar density
K-Means provides clear and easy-to-interpret clusters.

1.5 Real-World Application: Describe a hypothetical real-world business scenario where the customer segments you identified could be used by the mall's marketing team. For example, how would you target the group with high income but low spending score?

The mall's marketing team can use these customer segments to tailor promotions. For example, the group with high income but low spending score can be targeted with premium products, exclusive discounts, or personalized offers to encourage more spending. The group with low income but high spending can be offered affordable products or seasonal discounts to maintain high satisfaction. Each cluster helps the team design marketing strategies that match customer behavior and spending patterns.