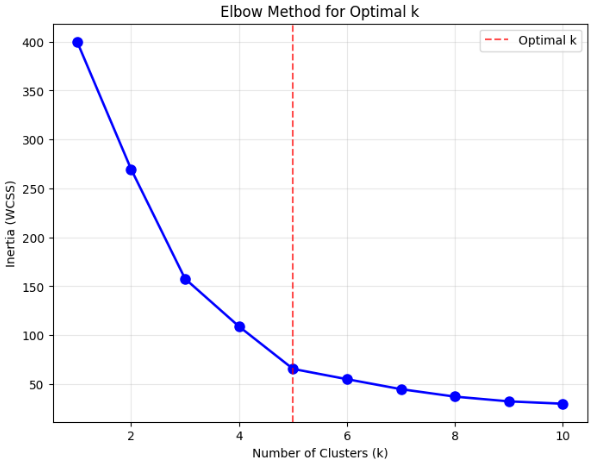
Clustering Methods using scikit-learn

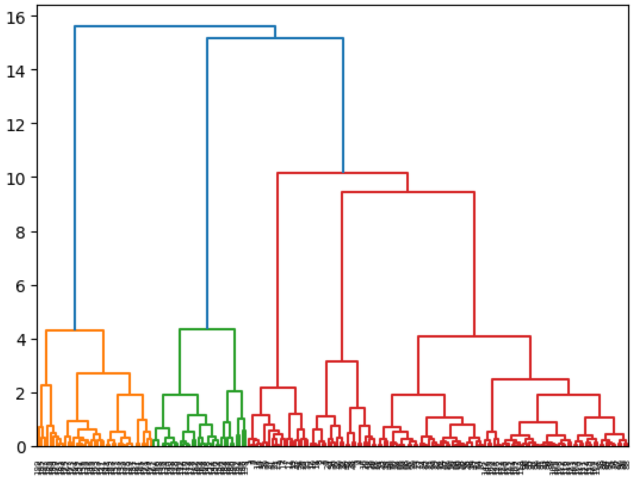
Srishti Lamba-202518003

M.Sc. Data Science, DAU Gandhinagar

Optimal Clusters: K-Means and Hierarchical Clustering

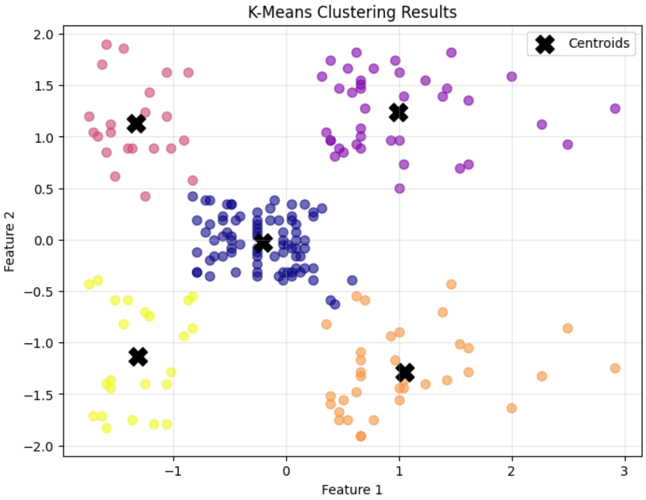


K-Means: The Elbow Method graph shows a sharp bend at k = 5, which means adding more clusters after this point does not improve the grouping much. So, five clusters are the best choice.

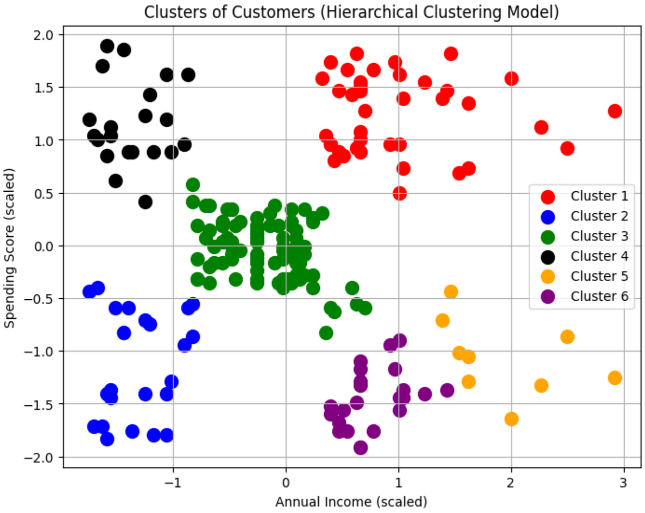


Hierarchical Clustering: The dendrogram suggests six clusters. This is based on where the longest vertical line cuts across the horizontal branches, showing six clear groups.

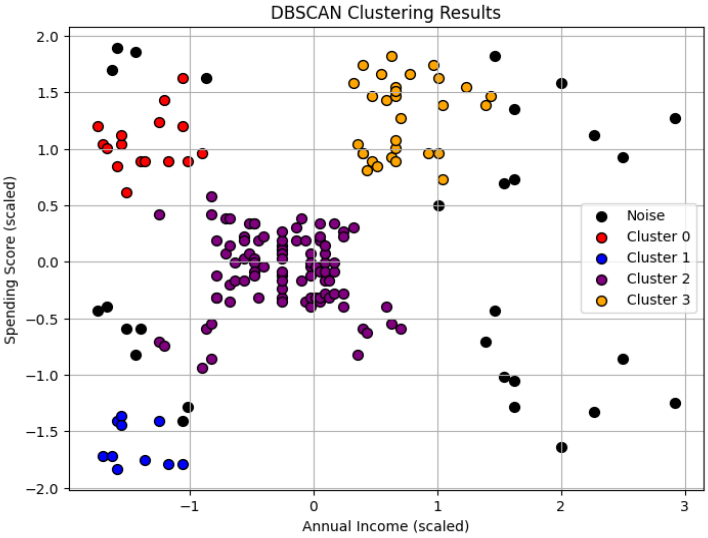
Cluster Comparison: K-Means vs Hierarchical vs DBSCAN



K-Means formed 4 compact, spherical clusters with clear centroids. The clusters were well-separated and easy to interpret.



Hierarchical Clustering resulted in 6 clusters, offering finer segmentation. However, some clusters overlapped slightly, and the boundaries were less distinct.



DBSCAN identified 4 clusters and also labeled several points as noise (black dots). These noise points were not assigned to any cluster, which makes DBSCAN unique compared to the other two methods.

Notable Differences:

- **K-Means and Hierarchical assign every point to a cluster.**
- **DBSCAN is density-based and can exclude outliers, making it more flexible for irregular shapes and sparse regions.**

DBSCAN Performance

DBSCAN performed well in identifying clusters of varying shapes and densities. It successfully detected noise points, which are data points that don't belong to any dense region. These were labeled as -1 in the output.

Compared to K-Means and Hierarchical Clustering, DBSCAN was better at handling outliers and non-spherical clusters. However, its performance was highly dependent on the choice of parameters: eps and min_samples.

With proper tuning, DBSCAN revealed meaningful clusters and isolated scattered data that the other methods misclassified.

Most Suitable Algorithm

Among the three clustering algorithms analyzed, K-Means emerged as the most effective for this particular dataset. It generated clear and interpretable clusters that aligned well with the Elbow Method's findings. The dataset's distribution was relatively tight and spherical, which is a good match for the assumptions underlying K-Means.

Hierarchical Clustering provided valuable insights into the hierarchical relationships within the data, allowing for more nuanced segmentation. However, it was not as resistant to noise and posed challenges when it came to scalability.

On the other hand, DBSCAN excelled at identifying outliers and managing clusters with varying shapes. Yet, it's important to note that DBSCAN is sensitive to changes in parameter settings, which can lead to significant variations in results, making it less stable for this dataset under slight parameter adjustments.

Real-World Application

The marketing team at the mall can utilize these clustering insights to effectively target high-income customers who demonstrate low spending scores. Although this demographic has the financial means to spend, they may lack interest or motivation in the current offerings. To engage this audience, the mall should consider implementing premium loyalty programs, personalized shopping experiences, and exclusive access to high-end brands that align with their lifestyle. Additionally, tailored communications featuring curated recommendations can enhance their experience and encourage increased spending. By aligning marketing strategies with the unique characteristics of these customers, the mall has the potential to transform their interest into profit.

Data and code

- **Dataset: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial**
- **Code Repository: https://github.com/srishti7103/clustering-using-scikit-learn**