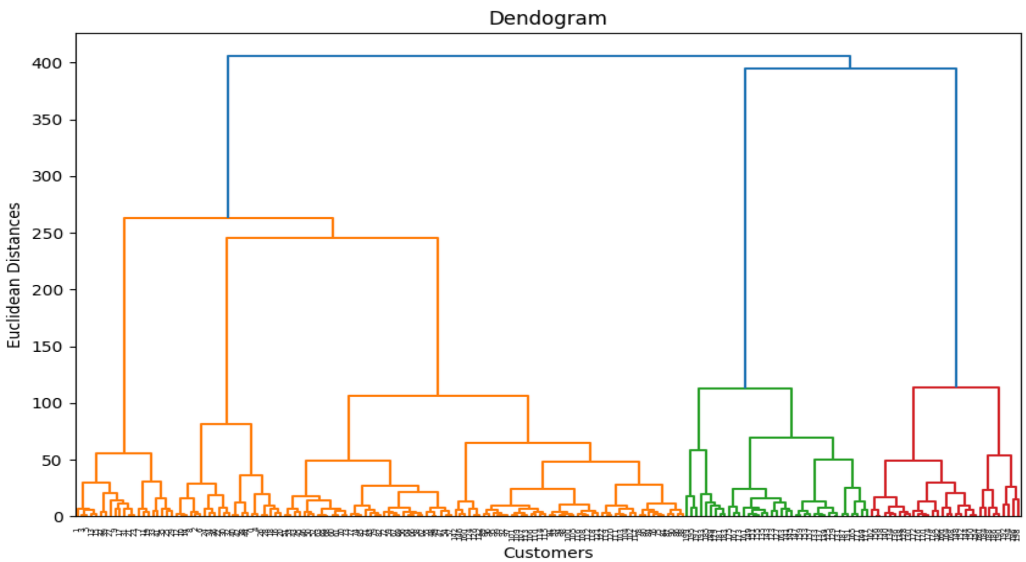
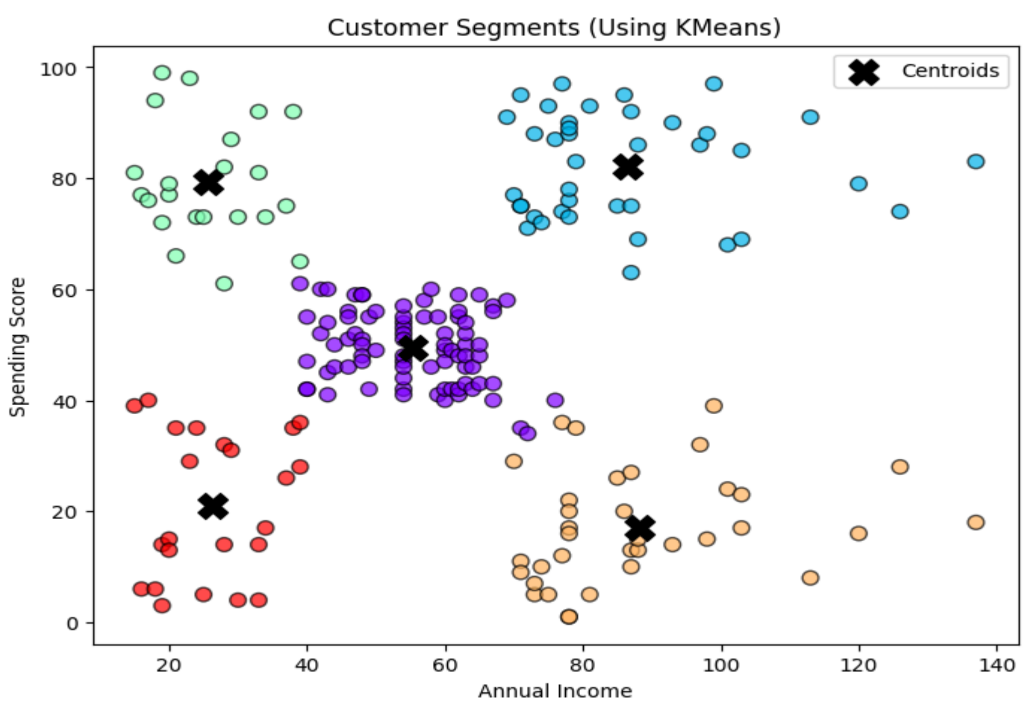
DS605: Lab 06 - Clustering using Scikit-learn

STUDENT ID: 202518052

PART 5

1. **Optimal Clusters**

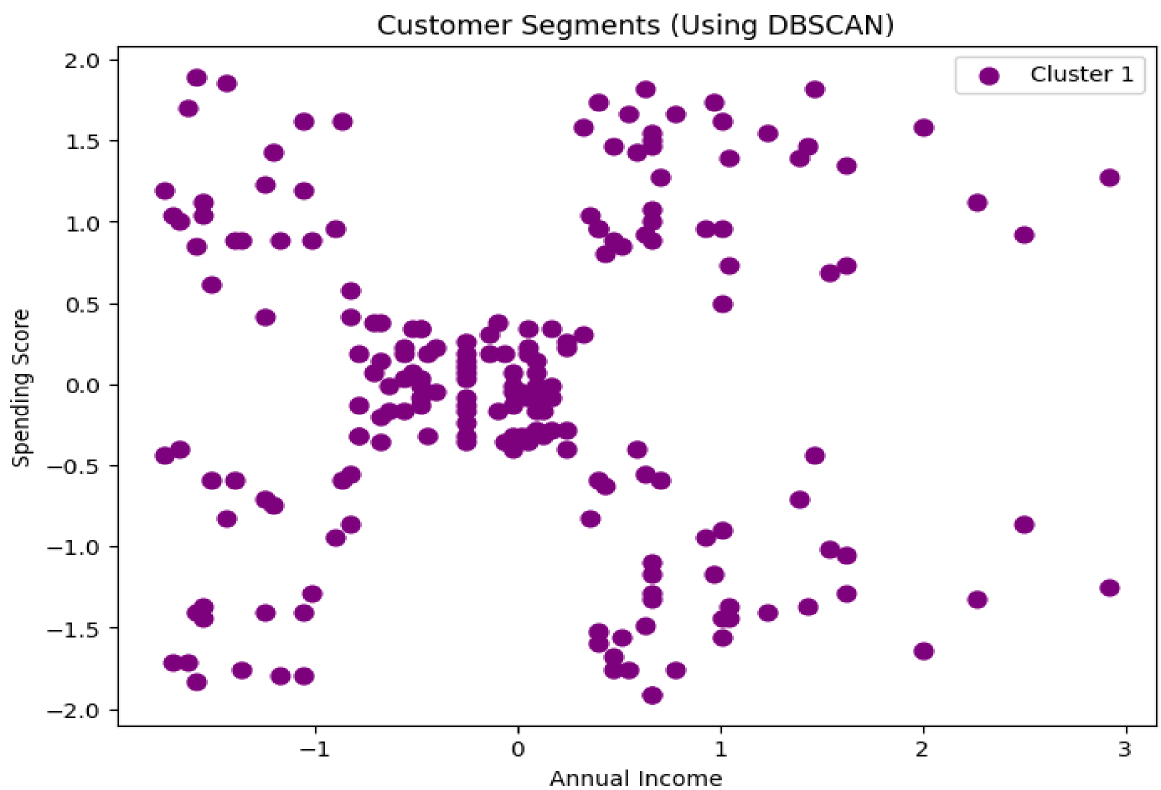


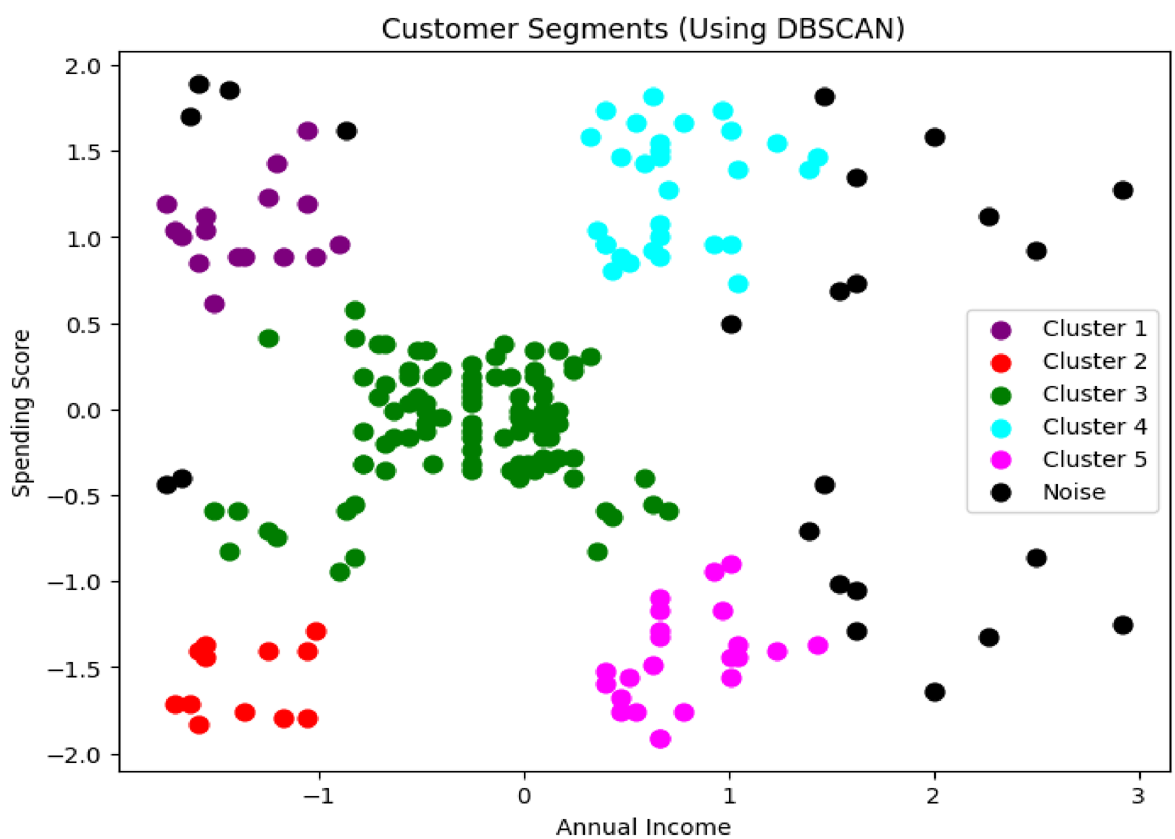


After applying the Elbow Method for K-Means, the Within-Cluster Sum of Squares (WCSS) curve showed a clear bend (elbow) at k = 5. This indicates that 5 clusters is the optimal balance between minimizing intra-cluster distance and avoiding overfitting.
For Hierarchical Clustering (Ward linkage), the dendrogram displayed a significant vertical distance (large merge step) when reducing from more than 5 clusters. Cutting the dendrogram at this height naturally results in 5 well-separated clusters.

2. Cluster Comparison

K-Means and Hierarchical (Ward) clustering produced very similar results, both identifying five compact, spherical clusters in the dataset and assigning every data point to a group. While K-Means forms clusters around centroids and Ward relies on iterative merging, their outcomes were visually almost identical. DBSCAN, however, behaved differently as it grouped points only in high-density regions, left some points as noise or unclustered outliers, and produced clusters that were not constrained to spherical shapes, making its results distinct from those of K-Means and Hierarchical.


Customer Segments (Using DBSCAN)
Cluster 1
Spending Score
Annual Income
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1
0
1
2
3


Customer Segments (Using DBSCAN)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Noise
Spending Score
Annual Income
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1
0
1
2
3

3. **DBSCAN Performance**

DBSCAN successfully detected dense customer segments and separated a few noise points (customers whose behavior didn't align with any dense group). Unlike K-Means and Hierarchical (which force all data points into clusters), DBSCAN left some points unassigned (`label=-1`). This makes DBSCAN powerful for outlier detection, though less useful if every customer needs to belong to a segment.
On this dataset, DBSCAN performed reasonably but produced fewer, denser clusters compared to the 5 clear segments found by the other two methods.

4. **Algorithm Suitability**

For this dataset, the most suitable algorithm was K-Means (or Ward Hierarchical) since the data naturally forms compact, spherical clusters when plotted using Annual Income and Spending Score. K-Means is both efficient and easy to interpret through its centroids, while also ensuring that every customer is assigned to a group. Ward Hierarchical produces a similar cluster structure and adds the advantage of a dendrogram for exploratory analysis. Although DBSCAN is powerful for detecting outliers and handling irregularly shaped clusters, the clear and compact structure of this dataset makes K-Means with five clusters the most appropriate choice.

5. **Real-World Applications**

The mall can leverage these customer segments to design tailored marketing strategies that maximize engagement and sales. For instance, high-income but low-spending customers can be targeted with exclusive VIP promotions, premium experiences, and personalized invitations to luxury events to encourage them to spend more. Customers with both high income and high spending scores could be nurtured through premium memberships, loyalty perks, and early access to new collections. Those with lower income but high spending scores may respond better to discounts, coupons, and referral programs that help maintain their frequent visits. Medium segments can be encouraged to spend more through seasonal offers, bundled deals, and targeted upselling strategies. Finally, outliers identified by DBSCAN, often one-time visitors or unusual shoppers, can be reached with short-term campaigns designed to convert them into repeat buyers.