Customer Segmentation Using DBSCAN

Objective:
The goal of this task was to perform customer segmentation using DBSCAN clustering based on customer profile information (Total Spending, Average Spending, and Transaction Count). Additionally, clustering evaluation metrics were computed, and clusters were visualized using dimensionality reduction (PCA).

Clustering Algorithm: DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was chosen as the clustering algorithm. This algorithm works by grouping points that are closely packed together while marking points in low-density regions as noise. DBSCAN has two main parameters:

1. **eps(epsilon)**: The maximum distance between two points to be considered as neighbors.
2. **min_samples**: The minimum number of points required to form a dense region (cluster).

For this analysis:

- **eps=0.5**
- **min_samples= 5**

These values were selected based on trial and error, aiming for meaningful segmentation.

Data Preprocessing

1. **Features Extracted for Clustering:**

   - **Total Spending:** The sum of all transactions made by a customer.
   - **Average Spending:** The average transaction value for a customer.
   - **Transaction Count:** The total number of transactions made by a customer.
2. **Normalization:**
   The features were normalized using **StandardScaler** to ensure each feature has a mean of 0 and a standard deviation of 1, making them comparable during clustering.

Clustering Evaluation Metrics

The clustering was evaluated using the following metrics:

1. **Davies-Bouldin Index:**

    - **Value:** 0.8381
    - The Davies-Bouldin Index measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower value indicates better clustering. In this case, a value of 0.8381 suggests the clusters are reasonably well-separated.
2. **Silhouette Score:**

    - **Value:** 0.3678
    - The silhouette score measures how similar a point is to its own cluster compared to other clusters. A score close to 1 indicates well-defined clusters. The score of 0.3678 suggests moderate clustering quality, implying that some clusters might overlap or have outliers.

Clusters and Noise Points

1. **Number of Clusters Formed:**
   Total clusters: 5
   The DBSCAN algorithm found 5 distinct clusters of customers, based on the features provided.
2. **Number of Noise Points:**
    - **Total noise points:** 37
    - The noise points are those that DBSCAN could not fit into any of the identified clusters, which are labeled as -1. These are customers whose transaction patterns do not align well with the clusters.

Cluster Visualization

A 2D visualization of the clusters was created using **Principal Component Analysis (PCA)**, which reduced the features into two components for ease of plotting.

- **X-axis:** Principal Component 1
- **Y-axis:** Principal Component 2

Clusters are color-coded, and noise points are marked separately. The visualization helped in identifying the spread of customer groups and any possible outliers.

Conclusion

- **Clusters:** DBSCAN successfully identified 5 customer segments based on their spending and transaction behavior. These clusters represent groups with distinct transaction patterns, which can be useful for targeted marketing, personalized offers, or product recommendations.

- **Noise Points:** A total of 37 customers were identified as noise, suggesting that they do not fit neatly into any of the clusters. These customers might require further analysis to understand their unique behavior.

- **Evaluation Metrics:** The relatively low Davies-Bouldin index and moderate silhouette score indicate that the clustering quality is fair, with room for improvement. Further parameter tuning and feature selection could potentially improve the results.