Clustering Algorithm: DBSCAN

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was selected for this task. DBSCAN clusters points that are closely packed together and labels those in low-density areas as noise.

Key Parameters for DBSCAN:

- **eps (epsilon)**: Defines the maximum distance between two points for them to be considered
  Value: 0.5
- **min_samples**: Specifies the minimum number of points required to form a dense region (cluster).
  Value: 5

These parameter values were chosen through experimentation to produce meaningful customer segments.

Data Preprocessing

1. **Extracted Features for Clustering**:
   - **Total Spending**: The overall value of all transactions made by a customer.
   - **Average Spending**: The average amount spent per transaction by the customer.
   - **Transaction Count**: The number of transactions a customer has made.
2. **Normalization**: The features were standardized using the **StandardScaler** to ensure all features have a mean of 0 and a standard deviation of 1, which helps avoid bias in the clustering process due to differing scales between features.

Clustering Evaluation Metrics

The clustering performance was measured using the following two metrics:

1. **Davies-Bouldin Index**:
    - **Score**: 0.8381
    - This metric calculates the average similarity between each cluster and its most similar cluster. A lower score indicates well-separated clusters. A value of **0.8381** suggests that the clusters are adequately distinct from one another.
2. **Silhouette Score**:
    - **Score**: 0.3678
    - The silhouette score quantifies how similar a point is to its own cluster compared to other clusters. Scores close to **1** suggest well-separated clusters. The score of **0.3678** suggests moderate clustering quality, indicating some overlap or presence of outliers.

Clusters and Noise Points

1. **Total Number of Clusters**:
    - **Clusters identified**: 5
    - DBSCAN identified **5 distinct customer clusters** based on spending patterns and transaction behaviors, revealing groups with varying levels of activity.
2. **Total Number of Noise Points**:
    - **Noise points**: 37
    - The algorithm classified **37 customers as noise**, meaning these customers did not exhibit transaction patterns that were consistent with any of the identified clusters.

Cluster Visualization

A 2D visualization was produced by applying **Principal Component Analysis (PCA)** to reduce the feature space into two components, making it easier to plot.

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

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

The clusters were displayed in different colors, and noise points were marked separately. This visualization helped in understanding how the customer segments are distributed and highlighted any potential outliers.

Conclusion

- **Cluster Identification**: DBSCAN successfully identified **5 customer segments** based on their spending and transaction patterns. These segments are valuable for targeted marketing strategies, personalized offers, and product recommendations.
- **Noise Points**: **37 customers** were categorized as noise, suggesting that their transaction behaviors do not align with the broader customer segments. These individuals may require additional analysis to identify the unique aspects of their purchasing behavior.
- **Evaluation Metrics**:
    - The **Davies-Bouldin index** being relatively low indicates that the clusters are reasonably distinct.
    - The **moderate silhouette score** indicates some overlap between clusters, suggesting there is room to fine-tune the clustering process for better results.
    - **Next Steps**: Improving the clustering results could involve tuning DBSCAN parameters and exploring additional features for segmentation.