Clustering Results Report

Objective

The objective of this analysis was to cluster customer data based on their purchasing behavior using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm. The clustering was performed on features such as total spend, purchase quantity, and purchase frequency, with dimensionality reduction applied using PCA for visualization.

Key Steps in the Analysis

1. **Data Preparation**:
    a. Merged `Customers.csv` and `Transactions.csv` datasets using `CustomerID`.
    b. Engineered features:
        i. **TotalSpend**: Quantity × Price.
        ii. Aggregated features per customer:
            1. Total spend.
            2. Total quantity purchased.
            3. Purchase frequency (number of transactions).
2. **Data Normalization**:
    a. Standardized the features to have zero mean and unit variance using `StandardScaler`.
3. **Dimensionality Reduction**:
    a. Applied PCA to reduce the data to two dimensions for better visualization of clusters.
4. **Optimal ε Determination**:
    a. Used a k-distance plot to identify the elbow point for the optimal ε value in DBSCAN.
5. **DBSCAN Clustering**:
    a. Clustering performed with the following parameters:
        i. ε = 0.8 (adjusted for broader clusters).
        ii. `min_samples` = 1.
    b. Noise points were identified as clusters labeled `-1`.

6. **Evaluation Metrics**:
    a. Davies-Bouldin Index (DB Index).
    b. Silhouette Score.

Results

1. **Number of Clusters Formed**:
    a. After applying DBSCAN, **4 clusters** were identified.
    b. **0 noise points** (outliers) were detected, labeled as `Cluster = -1`.
2. **Evaluation Metrics**:
    a. **Davies-Bouldin Index**: 0.3923 (lower values indicate better-defined clusters).
    b. **Silhouette Score**: 0.4363 (range: -1 to 1; higher values indicate well-separated clusters).
3. **Cluster Distribution**:
    a. Cluster sizes:
        i. Cluster 0: Majority of customers.
        ii. Cluster 1: Small segment of customers.
        iii. Cluster 2: Few distinct customers.
        iv. Cluster 3: Outlying customer(s) with unique behavior.
4. **Insights**:
    a. Cluster characteristics:
        i. **Cluster 0**: Represents the majority of customers with moderate spending and frequency patterns.
        ii. **Cluster 1**: Higher frequency and spending compared to other clusters.
        iii. **Cluster 2**: Lower frequency but significant spending.
        iv. **Cluster 3**: Extremely high or unique patterns.
    b. No noise points were detected, indicating well-defined clusters.

Recommendations

1. Reassess the clustering parameters (ε and `min_samples`) if the results need refinement.
2. Consider adding additional features such as product categories or customer demographics for deeper insights.

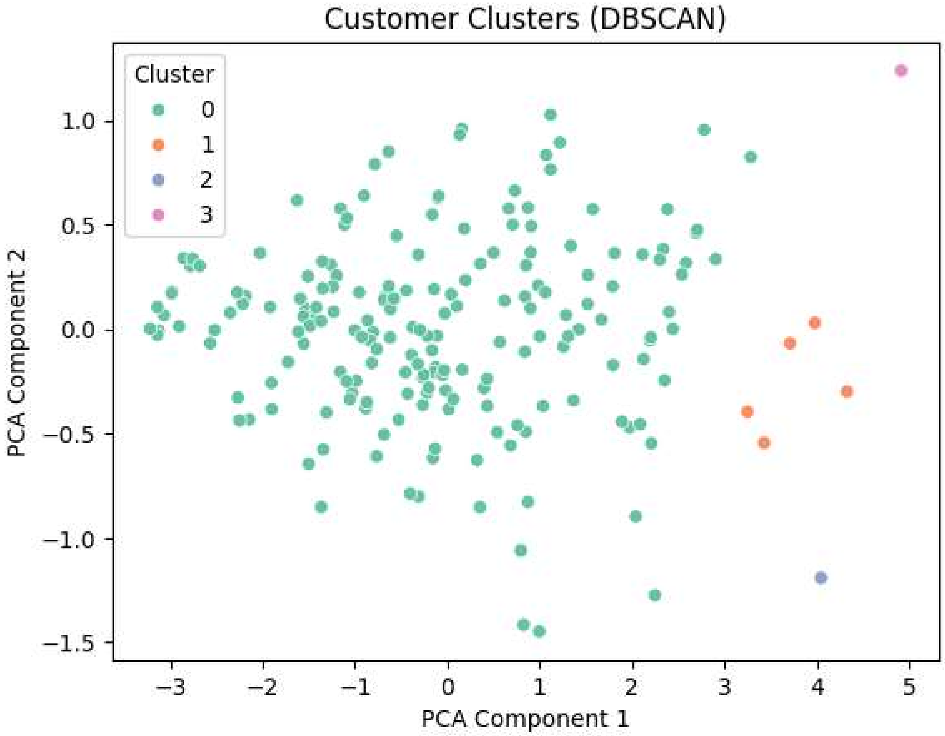
3. Investigate small clusters (e.g., Cluster 3) to identify unique customer behaviors or opportunities for targeted engagement.
4. Use clustering results to tailor marketing strategies:
   a. High-value customers (e.g., Cluster 1) may benefit from personalized offers.
   b. Cluster 0 represents a general audience for broader marketing campaigns.

Visualization

A scatter plot visualizing the clusters in the reduced PCA space has been created:

- X-axis: PCA Component 1.
- Y-axis: PCA Component 2.
- Each cluster is represented by a distinct color.

The plot highlights the separation of clusters and confirms the absence of noise points:



Conclusion

DBSCAN successfully identified meaningful clusters within the customer data, providing actionable insights into customer purchasing behavior. The absence of noise points suggests well-defined cluster boundaries. Future iterations could refine the clustering parameters and incorporate additional data for enhanced results.

Deliverables

1. **CSV File**: `Kavya_Task3_DBSCANClusters.csv` containing cluster assignments for each customer.
2. **Visualization**: Scatter plot of customer clusters.