Zeotap Data Science Assignment | Clustering

Objective

The goal of the analysis was to cluster customers based on their transactional behaviors and regional data to uncover meaningful customer segments. The clustering results were evaluated using **Davies-Bouldin Index (DB)** and **Silhouette Score** to determine the quality of clustering.

Dataset Description

The dataset includes customer transactions and demographic information. After preprocessing, the following features were used for clustering:

- **Numerical Features**:
  - `TotalValue`: Total transaction value per customer.
  - `Quantity`: Total number of items purchased by each customer.
  - `TransactionID`: Number of transactions per customer.
- **Categorical Feature**:
  - `Region`: One-hot encoded as `Region_` features.

Preprocessing Steps

1. Merged `Customers.csv` and `Transactions.csv` on `CustomerID`.
2. Aggregated transaction data per customer.
3. One-hot encoded the `Region` column.
4. Scaled all numerical features using `StandardScaler`.

Clustering Algorithms Evaluated

Five clustering algorithms were compared:

1. **KMeans**: Partitional clustering that minimizes intra-cluster variance.
2. **Agglomerative Clustering**: Hierarchical clustering approach.
3. **DBSCAN**: Density-based clustering that identifies dense regions and outliers.
4. **Spectral Clustering**: Graph-based clustering that captures non-linear relationships.

5. **Gaussian Mixture Model (GMM)**: Probabilistic clustering based on Gaussian distributions.

Evaluation Metrics

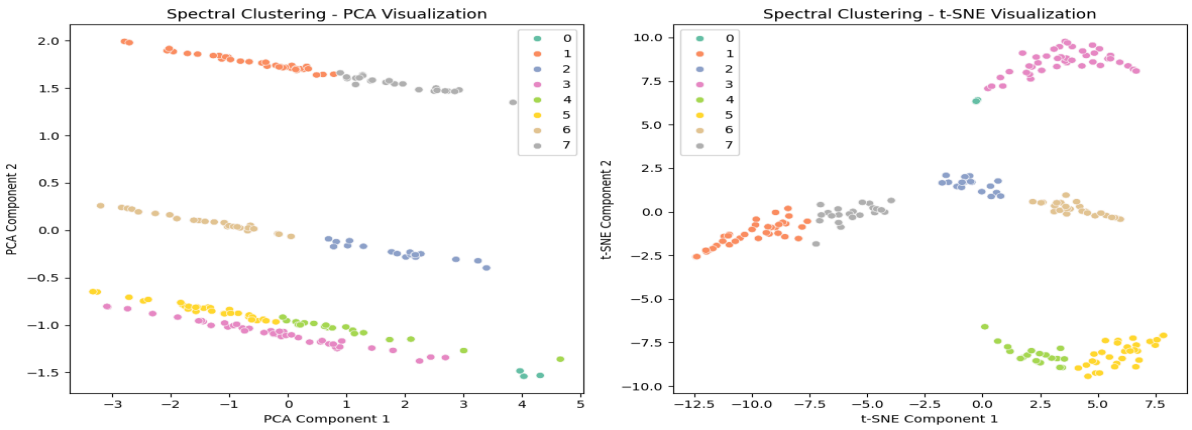
1. **Davies-Bouldin Index (DB)**:
    - Measures the average similarity ratio between each cluster and its most similar cluster. Lower values indicate better-defined clusters.
2. **Silhouette Score**:
    - Measures how similar a point is to its cluster compared to other clusters. Scores range from -1 (poor clustering) to 1 (excellent clustering).

Analysis of Different Algorithms :

Results

Algorithm	Number of Clusters	Davies-Bouldin Index (DB)	Silhouette Score
KMeans	10	0.785	0.442
Agglomerative	2	1.393	0.250
DBSCAN	8	1.395	0.016
Spectral	8	0.743	0.432
GMM	8	1.248	0.307

Visualization of Spectral Clustering Results :



Analysis and Justification for Spectral Clustering

1. **Performance**:
   - Spectral Clustering achieved the lowest **Davies-Bouldin Index (0.743)**, indicating well-separated and compact clusters.
   - The **Silhouette Score (0.432)** is comparable to KMeans and better than other methods, suggesting reasonable cohesion and separation of clusters.
2. **Handling Overlaps**:
   - Unlike KMeans or Agglomerative Clustering, Spectral Clustering can capture non-linear relationships in the data, making it more robust to overlapping clusters.
3. **Number of Clusters**:
   - Spectral identified 8 clusters, which aligns well with the GMM and DBSCAN results, suggesting that the data naturally supports this segmentation.
4. **Challenges with Other Algorithms**:
   - **KMeans**: While it achieved a similar Silhouette Score, its higher DB Index (0.785) indicates poorer cluster separation.
   - **Agglomerative Clustering**: Poor performance with only 2 clusters, showing oversimplified segmentation.
   - **DBSCAN**: Very low Silhouette Score (0.016) due to excessive noise or unassigned points.
   - **GMM**: Higher DB Index (1.248) indicates overlapping clusters not well-separated.