Zeotap Assignment

Name: Ayush Bhatt

University: VIT Bhopal University

Task 3: Customer Segmentation / Clustering

Objective:

This assignment required analyzing customer data containing transactional and profile information through multiple clustering strategies to achieve segmentation. Our examination of clustering models focused on their performance metrics through the Davies-Bouldin Index (DBI). This measure represents clustering quality factors. Clustering quality improves with decreasing numbers on the Davies-Bouldin Index scale.

Methodology:

To perform the segmentation, we used the following steps:

1. **Data Preparation:**
    - Customer Demographics: We used the Customers.csv file, which contains customer profiles, including CustomerID, CustomerName, Region, and SignupDate.
    - Transaction Data: We used the Transactions.csv file to calculate the total spend per customer and the number of unique products purchased by each customer. This data was aggregated and merged with customer profiles.
2. **Feature Engineering:**
    - We calculated the Total Spend per customer (sum of the total transaction values).
    - We calculated the Product Count (number of unique products purchased by each customer).
    - The features used for clustering included customer demographics (Region, SignupDate), Total Spend, and Product Count.
3. **Data Preprocessing:**
    - We standardized the data using StandardScaler to scale all features to a similar range.
    - Missing values were filled with 0.
4. **Clustering Models Evaluated:** We evaluated the following clustering techniques:

- **KMeans Clustering:** A popular algorithm that partitions the data into a predefined number of clusters.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** A density-based clustering algorithm that can identify outliers as noise.
- **Hierarchical Clustering:** This method builds a tree of clusters and is evaluated using different linkage methods, including:
  - **Single Linkage**
  - **Complete Linkage**
  - **Average Linkage**
  - **Ward Linkage**

5. **Model Evaluation:** To evaluate the clustering models, we calculated the Davies-Bouldin Index (DBI) for each model. A **lower DBI indicates better clustering results**. The models were evaluated with different numbers of clusters (for KMeans and hierarchical clustering) and different settings (for DBSCAN).

Results:

1. Davies-Bouldin Index for Different Clustering Models:

Clustering Model	Davies-Bouldin Index
KMeans (k=4)	0.90
DBSCAN	0.98
Hierarchical (Single Linkage)	0.38
Hierarchical (Complete Linkage)	0.72
Hierarchical (Average Linkage)	0.66
Hierarchical (Ward Linkage)	0.81

2. **Visualizing the Clusters (KMeans, DBSCAN, and Hierarchical):** The clustering results were visualized using PCA (Principal Component Analysis), which reduced the data to two dimensions for easy visualization.
   - **KMeans Clustering (k=4):** This method formed distinct clusters but had a relatively higher DBI, indicating that the separation between clusters could be improved.
   - **DBSCAN:** This method identified noise points, but it produced fewer clusters (a limitation when handling sparse data).

- **Hierarchical Clustering (Single Linkage):** This method had the best DBI score (0.38), suggesting that it had the most effective clustering with the fewest overlapping clusters.

3. **Cluster Visualization:** We used PCA to reduce the data to two principal components and visualized the clusters. The scatter plot below shows the clusters formed by KMeans and Hierarchical (Single Linkage) clustering.

Conclusion:

- **Best Model:** Based on the **Davies-Bouldin Index**, the **Hierarchical Clustering (Single Linkage) method** performed the best, with the **lowest DBI score (0.38)**.
- **Future Work:** Further refinement of DBSCAN parameters and exploring other clustering algorithms such as Gaussian Mixture Models could be beneficial for improving the segmentation quality.

Key Insights:

- Customers were successfully **segmented based on both their transaction behavior (total spend, product count) and their demographic information (region)**.
- Hierarchical Clustering with **Single Linkage** is the most suitable for this dataset.
- Further **tuning and testing** with different linkage methods or increasing the number of clusters could potentially improve the segmentation.