Customer Segmentation Using Clustering Techniques

Customer segmentation is a vital process that helps businesses divide their customer base into distinct groups with similar characteristics or behaviors. By grouping customers, businesses can tailor marketing strategies, improve customer experience, and optimize resource allocation. In this project, we'll use clustering techniques to perform customer segmentation by combining profile information from Customers.csv and transaction data from Transactions.csv.

Step-by-Step Approach

---

1. Objective

The goal is to group customers based on their profile and transaction behavior into meaningful clusters. These clusters will represent different customer types, enabling targeted marketing and data-driven business decisions.

Key deliverables:

- Number of clusters formed.
- DB Index value
- Other metrics like silhouette score and Interia Score for additional evaluation.
- Visualizations to illustrate clusters.

---

2. Data Preparation

Datasets Used

1. **Customers.csv**: Contains customer profile data such as:
    - CustomerID: Unique identifier for each customer.
    - CustomerName: Name of the customer.
    - Region: Geographic region of the customer.
    - SignupDate: Date of signup.

2. **Transactions.csv**: Contains transaction data such as:
    - TransactionID: Unique identifier for each transaction.
    - CustomerID: Identifier linking the transaction to a customer.
    - ProductID: Product purchased in the transaction.
    - Quantity: Quantity of the product purchased.
    - TotalValue: Total transaction value.

Feature Engineering

- **Profile Features**:
    - Encode Region into numerical values using one-hot encoding.
    - Extract the signup year or tenure from SignupDate.
- **Transaction Features**:
    - Aggregate transaction data for each customer:
        - **Total Spending**: Sum of TotalValue.
        - **Total Quantity**: Sum of Quantity.
        - **Average Transaction Value**: Mean of TotalValue.
        - **Unique Products Purchased**: Count of distinct ProductID.
        - **Transaction Frequency**: Number of transactions per customer.

Final Dataset

The final dataset will be a customer-level dataset where each row represents a customer, and the columns are engineered features derived from the raw data.

---

3. Clustering Algorithm

Algorithm Choice

For this project, we use **K-Means Clustering** because:

- It is straightforward and widely used for customer segmentation.
- It partitions the data into k clusters based on feature similarity.

- The number of clusters (k) can be selected through evaluation metrics like **Davies-Bouldin Index** or **Silhouette Score**.

---

4. Evaluation Metrics

DB Index (Davies-Bouldin Index)

- The **DB Index** is a clustering evaluation metric where **lower values indicate better clustering**.
- It measures the ratio of within-cluster scatter to between-cluster separation.
- A good clustering minimizes the DB Index by creating compact and well-separated clusters

Silhouette Score

- Evaluates how similar a point is to its own cluster compared to other clusters.
- Range: [-1, 1]. Higher values indicate better-defined clusters.

Inertia

- it is a metric used to evaluate the quality of clusters formed by the K-Means algorithm. It measures the sum of squared distances between each data point and the centroid of the cluster to which it belongs.
- A lower inertia score indicates that the clusters are tighter and closer to their centroids, which typically means better clustering performance.

---

6. Results

Number of Clusters:

- k = 2 clusters were formed, as determined by experimenting with different values (2 to 10) and evaluating metrics.

DB Index:

- A **DB Index** of 0.7316768187412901 was achieved

Silhouette Score:

- The silhouette score was 0.4891735891779627

Interia Score:

- Interia Score was 248.57529667843852

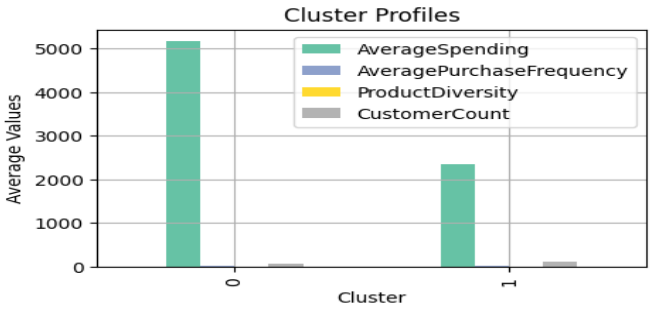
---

7. Deliverables

1. **Clustering Report**:
    - Number of clusters: 2
    - DB Index: 0.73
    - Silhouette Score: 0.48
    - Interia Score: 248.57

2.Visual Representations:

- Scatterplot of clusters on key features (Total Spending vs. Average Transaction Value).



- Visualize Clusters using PCA Visualization