Customer Segmentation Using Clustering

Introduction:

One of the most important methods for comprehending consumer behavior and preferences is customer segmentation. It facilitates the categorization of clients according to shared traits or transactional patterns. In this exercise, we divide up our customer base into two categories using clustering techniques:

Details about the profile: taken from the Customers.csv file.
The Transactions.csv file contains the transaction information.

We will create clusters using the proper clustering technique and assess the quality of the clustering using a number of metrics, such as the Davies-Bouldin (DB) Index.

Data Preprocessing:

1.1 Data Loading

First, we load the data from Transactions and Customers.csv.csv files into the proper data structures (such as Python's pandas DataFrames). This makes it possible to process and analyze the datasets effectively.

1.2 Feature engineering and data merging

Based on a distinct customer identity, the transaction data and customer profile information will be combined to produce an extensive dataset for clustering. Additionally, we might design new features like:

- Average Spend: Based on transaction data, the average amount spent by each consumer.
- Total Transactions: The sum of all the transactions that each client has completed.
- Recentness: How many days have passed since the last transaction?
- Demographics: Data from the consumer profile, such as location, income, and age.

1.2 Standardization

We will use methods like Min-Max Scaling or Standardization to standardize the data so that every feature is on the same scale.

Clustering Method:

1.1 Selection of Algorithms

K-Means Clustering, a popular unsupervised machine learning method, will be employed for this task's client segmentation. Within the range of two to ten clusters, the number will be determined by evaluation measures like the Elbow Method and Silhouette Score.

1.2 The quantity of clusters

The DB Index, which gauges cluster separation and compactness, will be assessed in order to establish the ultimate number of clusters. Visualizing clusters to identify the number of clusters that best describe the data using methods such as t-SNE or PCA (Principal Component Analysis).

Clustering Evaluation:

1.1 The Davies-Bouldin Index (DB Index) for Clustering Evaluation:
The quality of clustering will be assessed using the DB Index. Better-defined clusters are indicated by a lower DB Index. This index will be computed upon clustering.

1.2 Additional Evaluation Criteria:
The Silhouette Score calculates an object's similarity to its own cluster in relation to other clusters.
The total squared distances between each location and the designated cluster centroid is known as inertia.

Findings from Clustering:

1.1 Clusters Created
The number of clusters that develop will be ascertained following the application of K-Means. For example, we will have four different client segments if K=4.

1.2 Centroids of Clusters
The average feature values of the customers in each cluster will be represented by the centroid. These centroids can shed light on the distinctive traits of every client group.

1.3 Value of the DB Index
To evaluate the quality of the clusters created, the clustering's DB Index will be computed. As the number of clusters rises, we anticipate that the DB Index value will fall; however, there will come a time when more clusters won't result in better segmentation quality.

Visualization of Clusters:

1.1 PCA Visualization
We will reduce the dimensionality of the data to two or three dimensions using Principal Component Analysis (PCA) in order to visualize the clusters. We may then plot the clusters on a 2D or 3D plot thanks to this.

1.2 Visualization of t-SNE
To clearly see the distinction between the clusters, the high-dimensional data will be projected into two dimensions using t-SNE, another visualization technique.
In conclusion, it will be discussed how many clusters were produced and what traits each cluster possessed. For instance, a group of high-income clients who buy from you often and a group of low-income, infrequent buyers.
• To evaluate the clustering quality, the DB Index value will be displayed together with additional evaluation metrics like the Silhouette Score.
• Visualizations will give an intuitive sense of how nicely the clientele is divided up.

In conclusion, it will be discussed how many clusters were produced and what traits each cluster possessed. For instance, a group of high-income clients who buy from you often and a group of low-income, infrequent buyers.
• To evaluate the clustering quality, the DB Index value will be displayed together with additional evaluation metrics like the Silhouette Score.
• Visualizations will give an intuitive sense of how nicely the clientele is divided up.