Customer Segmentation Using KMeans Clustering

1. Introduction:

This project involves the application of clustering techniques on customers' profiles as well as their transactional data. It is aimed at discovering different groups of customers who could be approached to offer customized marketing strategies or tailored services.

2. Dataset:

Customers.csv: It is a customer profile data containing customer ID, region, and signup date.

Transactions.csv: Transactions data like total spend, transactions count, etc.

These two datasets were joined on CustomerID, and finally, a complete dataset was formed including customer demographic as well as the transaction details.

3. Preprocessing Tasks:

The preprocessing tasks involved in this task:

Data merging: The data was merged with the CustomerID.

Feature Engineering: The features calculated are listed below:

TotalValue: Sum of the spending by the customer

TransactionID: Number of times the customer has made a transaction (frequency)

Quantity: Sum of the total quantity bought

Scaling: Features were standardized using StandardScaler such that the contribution of all the features to the clustering was uniform.

4. Clustering Algorithm:

Algorithm Used: KMeans Clustering

Optimal Number of Clusters: After evaluating the clustering metrics, the optimal number of clusters was chosen to be 2.

Cluster Evaluation Metrics: The performance of the clustering was evaluated using the Silhouette Score and the Davies-Bouldin Index (DB Index).

5. Clustering Results:

Optimal Number of Clusters: The optimal number of clusters formed was 2, based on the lowest DB Index and highest Silhouette Score.

Silhouette Score: The best Silhouette Score was 0.4949 for the optimal clusters. It reflects that there is a reasonable degree of cohesiveness as well as separation among the clusters.

DB Index: The value of Davies-Bouldin Index of the obtained clustering is 1.016. It indicates the measure of separation among the clusters; smaller the values are better the separation.

6. Cluster Interpretation:

Clusters that are created can be depicted as below.

Cluster 1 (Low to Moderate Spend, Low Frequency):

This category includes customers with low total spendings and less number of transactions.

Probably an infrequent shopper or someone who is at the nascent stage of engagement with the firm.

Cluster 2: High Spend, High Frequency:

Such a cluster consists of customers who spend a lot and shop more often.

They are probably loyal or premium customers who frequently interact with the brand and have a large amount of spend.

7. Metric Value

Number of Clusters 2

DB Index 1.016

Silhouette Score 0.4949

8. Conclusion:

Customer segmentation led to 2 clusters that are associated with two types of customers: low engagement and low spend, and high engagement with frequent purchases. The DB Index and silhouette score indicate a moderate separation between these clusters.

Though the clusters look pretty well defined, there is always room for further refinement in the feature space or through an alternative clustering method such as DBSCAN or hierarchical clustering for more refined segmentation.