Customer Segmentation/Clustering

About:

This is a report on the customer segmentation/clustering results performed using both customer profile information (from Customers.csv) and transaction information (from Transactions.csv). The clustering algorithm used is KMeans and the optimal number of clusters is identified between 2 and 10. The clustering metrics used are Davies Bouldin (DB) Index and Silhouette score.

Methodology:

The customer profile information (from Customers.csv) and transaction information (from Transactions.csv) are merged from which 3 features are created for each customer:

- TotalTransactions – Number of transactions made by the customer,
- TotalQuantity – Number of products bought by the customer,
- TotalAmount – Total amount spent by the customer.

This customer purchase information is merged with customer profile information. One-hot encoding is performed on Region and features such as Customer ID, Customer Name and Signup date are dropped. Min-max scaling is performed on the numerical features to ensure data is ready for clustering.

KMeans clustering is performed with the number of clusters ranging from 2 to 10. DB Index and Silhouette score are calculated for each value of number of clusters. The one which gives the lowest DB Index is chosen as the optimal value for the number of clusters. The cluster labels assigned for the customers are stored as results.

Results:

Optimal number of clusters = 4

Davies-Bouldin Index = 0.4136

Silhouette score = 0.7368