A REPORT ON CLUSTERING MODEL

Introduction

This report documents the findings of the customer segmentation analysis using KMeans clustering on customer and transaction data. The goal was to identify distinct customer segments among customers based on their purchasing behavior along with demographic information.

Datasets Used

- **Customers.csv**: Contains customer profile information, including CustomerID, CustomerName, Region, and SignupDate.
- **Transactions.csv**: Contains transaction details including TransactionID, CustomerID, ProductID, TransactionDate, Quantity, TotalValue, and Price.

Method of Work

1. **Data Merging**: The two datasets were merged so that one could have all the details on the customers and their transaction history.
2. **Feature Engineering**: Key features were derived while aggregating transaction data:
    - **Total Quantity Purchased**: sum of all quantities each have been purchased by.
    - **Total Spending**: the sum of transaction value for each customer.
3. **Data Preprocessing**: The feature data was normalized with StandardScaler in order to make every feature contribute equally to the distance calculations in clustering.
4. **Clustering Algorithm**: Customer segmentation was done using the KMeans clustering algorithm. The best number of clusters is given by the elbow method.

Findings

- **Number of clusters**:
    - Number of Clusters: 2, DB Index: 0.63, Silhouette Score: 0.55
    - Number of Clusters: 3, DB Index: 0.70, Silhouette Score: 0.45
    - Number of Clusters: 4, DB Index: 0.72, Silhouette Score: 0.45
    - Number of Clusters: 5, DB Index: 0.75, Silhouette Score: 0.43

    - Number of Clusters: 6, DB Index: 0.82, Silhouette Score: 0.39
    - Number of Clusters: 7, DB Index: 0.88, Silhouette Score: 0.37
    - Number of Clusters: 8, DB Index: 0.83, Silhouette Score: 0.38
    - Number of Clusters: 9, DB Index: 0.84, Silhouette Score: 0.39
    - Number of Clusters: 10, DB Index: 0.80, Silhouette Score: 0.37
- **Optimal number of clusters(k)**: Based on the metrics above, the optimal number of clusters is determined to be 2, as it has the lowest DB Index (0.63) and a relatively high silhouette score (0.55).

Metrics Interpretation

- **Davies-Bouldin Index**: A lower value of the DB index indicates good separation between clusters. With a value of 0.63, the two clusters are reasonably distinct from each other.
- **Silhouette Score**: A silhouette score close to +1 indicates points being well clustered. The score of 0.55 shows a good level of cohesion in relation to the clusters.

Visualizations

- **Elbow Method Plots**: These plots would show the DB Index and silhouette scores for different numbers of clusters, from 2 to 10, providing a visual mechanism to estimate the optimal number of clusters.
- **Customer Segmentation Cluster Plot**: This scatter plot is capable of depicting the clusters that would be formed from the quantity purchased and the total amount spent**.**