Customer Segmentation Using K-Means Clustering

Objective:

The goal of this task was to segment customers based on their purchasing behavior using K-Means clustering. The segmentation helps identify groups of customers who exhibit similar spending patterns, enabling targeted marketing strategies for each group.

Data Overview:

The dataset used for segmentation consists of customer and transaction data. The transaction data provides information about customer purchases, including transaction amounts and dates, while the customer data contains demographic information such as the region of each customer. After merging the two datasets, we created aggregated features representing the transaction behaviors of each customer.

Feature Engineering:

To characterize each customer's purchasing behavior, we computed the following features:

- Total Transactions: The total number of transactions made by each customer.
- Total Spent: The total amount spent by each customer across all transactions.
- Average Purchase Value: The average transaction value for each customer.

Data Scaling:

Since the features represent different scales (e.g., total amount spent versus average purchase value), we applied Min-Max scaling to normalize the data. This ensures that all features contribute equally to the clustering process.

Optimal Number of Clusters:

To determine the ideal number of clusters, the Elbow Method was used. We ran the K-Means algorithm for cluster values between 2 and 10 and calculated the Within-Cluster Sum of Squares (WCSS) for each. The optimal number of clusters is typically chosen where the reduction in WCSS slows down, creating an "elbow" in the plot.

Based on the elbow curve, we identified 3 clusters as the optimal choice.

Clustering:

With 4 clusters selected, we applied the K-Means algorithm to the data. Each customer was assigned to one of the 4 clusters based on their purchasing behavior.

Cluster Evaluation:

To evaluate the quality of the clusters, we used the **Davies-Bouldin Index (DB-Index)**. The DB-Index measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower DB-Index indicates better-separated clusters. The calculated DB-Index for this segmentation is:

DB-Index: 0.9008

A DB-Index of **0.9008** indicates that the clusters are well-separated with relatively distinct boundaries between them.

Data Visualization: Customer Segmentation:

One of the key visualizations used is a scatter plot that compares the **Total Amount Spent** by customers against their **Average Purchase Value**. This helps visualize the relationship between these two key features and how customers are grouped based on these characteristics.

Conclusion:

The customer segmentation process using K-Means clustering was successful, and the data was effectively partitioned into 3 clusters based on purchasing behavior. The DB-Index of **0.9008** indicates a reasonable separation between the clusters, suggesting that the algorithm has correctly identified distinct groups within the customer base.

The cluster visualization confirms that the customers within each segment exhibit similar purchasing patterns, which can be useful for personalized marketing strategies.