

# Clustering Report

## 1. Clustering Logic

We used K-Means Clustering, a popular unsupervised machine learning algorithm, to group customers into clusters. The steps involved are:

### Feature Engineering:

- Combined customer profile data (Customers.csv) and transaction data (Transactions.csv) to create meaningful features.

### Features included:

- Total spending per customer.
- Average transaction value.
- Number of transactions.
- Favorite product category (most purchased category).
- Customer lifetime (current date - signup date).
- Region (encoded as a categorical feature).

### Normalization:

- Scaled the features using MinMaxScaler to ensure all features contribute equally to the clustering process.

### Optimal Number of Clusters:

- Used the Elbow Method to determine the optimal number of clusters (k).
- Plotted the inertia (within-cluster sum of squares) for k ranging from 2 to 10.
- Choose k=5 as the optimal number of clusters based on the "elbow" point in the plot.

### Clustering:

- Applied K-Means clustering with  $k=5$  to group customers into 4 distinct clusters.

### Evaluation:

- Evaluated the clustering results using the Davies-Bouldin Index (DB Index) and Silhouette Score.

## **2. Clustering Metrics**

The following metrics were used to evaluate the clustering results:

### Davies-Bouldin Index (DB Index):

- Value: 0.81
- Interpretation: A lower DB Index indicates better clustering. A value of 0.81 suggests reasonably well-separated clusters.

### Silhouette Score:

- Value: 0.38
- Interpretation: The Silhouette Score ranges from -1 to 1. A score of 0.38 indicates that the clusters are relatively well-defined and separated.