

# Clustering Report

## Objective

This analysis aims at doing customer segmentation with clustering techniques based on profile information and transaction data. The quality of the segmentation will be evaluated through the calculation of important clustering metrics: the Davies-Bouldin Index (DB Index) and graphically.

## Data Overview

### 1. Customers.csv

The file contains information about the profiles of customers. The columns of this file are CustomerID, CustomerName, Region and SignupDate.

### 2. Transactions.csv

- It contains transactional information such as TransactionID, CustomerID, ProductID, TransactionDate, Quantity, TotalValue, and Price.

### 3. Products.csv:

- Contains product information like ProductID, ProductName, Category, and Price.

## Feature Engineering

Customer-level aggregated features were prepared for this analysis:

- TotalSpending: Overall amount spent by the customer.
- AverageTransactionValue: Average value per transaction.
- PurchaseFrequency: Number of transactions made.
- AverageQuantity: Average quantity bought in a single transaction.
- Region: Categorical variable that indicates which region a customer resides in, encoded.

## Clustering Methodology

- Standardization: Features have been standardized to make them comparable using StandardScaler.
- Algorithm: K-Means is used with varied cluster sizes between 2 and 10.
- Evaluation: The quality of the best clustering is established by the low value of Davies-Bouldin Index (DB Index). Optimal Number of Clusters: The analysis identified 9 clusters with a DB Index of 1.0987, the lowest value among the tested cluster sizes.

## Visualization

### 1. Total Spending vs. Average Transaction Value:

- Clearly differentiates clusters on the basis of spending behavior.
- Shows the big spenders, for example, clusters 4 and 3, against the low spenders, for example, clusters 1 and 6.

### 2. Purchase Frequency vs. Average Quantity:

- Catches the variation in purchase behavior.
- Displays clusters that have a high frequency of purchases but smaller quantities, for example, cluster 8, and clusters that have a low frequency but large quantities, for example, cluster 7.

## Key Metrics

- Number of Clusters: 9
- Davies-Bouldin Index: 1.0987
- Cluster Distribution: Ranges from 11 to 33 customers per cluster.

## Conclusion

The clustering analysis successfully segmented customers based on transaction and profile data. The 9-cluster solution has a good balance of compactness and separation of clusters, as reflected in the low DB Index.

This segmentation can inform targeted marketing strategies, tailored customer engagement, and resource allocation to maximize customer satisfaction and business profitability.