# Customer Segmentation Clustering Report

## 1. Introduction

Customer segmentation is a crucial step in understanding customer behavior and optimizing marketing strategies. This report presents the results of clustering customers based on both profile and transaction data using K-Means clustering. The primary goal was to identify distinct customer groups and evaluate clustering quality using relevant metrics.

## 2. Data Overview

The clustering analysis was performed using two datasets:

- **Customers.csv** (CustomerID, CustomerName, Region, SignupDate)
- **Transactions.csv** (TransactionID, CustomerID, ProductID, Quantity, TotalValue, Price)

### Preprocessing Steps:

1. Merged customer and transaction data.
2. Extracted aggregated features such as total spend, average spend, number of transactions, and total quantity purchased.
3. Encoded categorical variables (Region) using One-Hot Encoding.
4. Standardized the features using **RobustScaler** to reduce the influence of outliers.

## 3. Clustering Algorithm & Optimal K Selection

To determine the optimal number of clusters (**K**), we tested different values (2 to 10) and evaluated clustering performance using the **Davies-Bouldin Index (DBI)** and **Silhouette Score**.

- **Optimal number of clusters: 4** (Chosen based on the lowest DB Index)

### Final Clustering Results:

- **Davies-Bouldin Index (DBI):** 0.6965 (Lower is better)
- **Silhouette Score:** 0.5527 (Higher is better)
- **Calinski-Harabasz Score:** 237.2063 (Higher is better)

## 4. Clustering Evaluation Metrics

### 4.1 Davies-Bouldin Index (DBI)

- Measures cluster compactness and separation.
- A lower DBI value (closer to 0) indicates well-separated clusters.
- Our result: **0.6965**, indicating good cluster separation.

### 4.2 Silhouette Score

- Measures how similar each point is to its own cluster vs. other clusters.
- Ranges from **-1** (poor) to **1** (good).
- Our result: **0.5527**, indicating moderately well-defined clusters.

### 4.3 Calinski-Harabasz Index

- Measures the variance ratio between clusters and within clusters.
- A higher value indicates better-defined clusters.
- Our result: **237.2063**, confirming strong clustering quality.

## 5. Visualization of Clusters

To interpret clustering results, we applied **Principal Component Analysis (PCA)** to reduce dimensionality to 2D and plotted the clusters:

- Clusters show **clear separation**, reinforcing the effectiveness of K-Means.
- Customers with similar transaction patterns were grouped together.

## 6. Recommendations & Next Steps

### 6.1 Further Improvements

1. **Feature Engineering**
   - Incorporate **Recency, Frequency, and Monetary (RFM)** analysis.
   - Extract customer lifecycle metrics such as **loyalty duration**.
   - Consider session-based behaviors for more dynamic segmentation.
2. **Try Different Clustering Approaches**
   - **Gaussian Mixture Model (GMM)** for probabilistic clustering.
   - **DBSCAN** for density-based clustering.
   - **Agglomerative Hierarchical Clustering** for hierarchical grouping.
3. **Refine K-Means Parameters**
   - Experiment with different distance metrics (e.g., **Manhattan, Cosine**).
   - Use **K-Means++ initialization** with more iterations for stability.

## 7. Conclusion

The clustering model successfully grouped customers into **4 distinct segments**, achieving a **low Davies-Bouldin Index (0.6965)** and a **good Silhouette Score (0.5527)**. These results indicate well-separated and meaningful customer clusters. Future enhancements could involve **advanced feature engineering** and **alternative clustering techniques** to further refine segmentation.