# Clustering Results Report

---

## 1. Overview

Dataset Description: The dataset includes customer transaction data, with information on customer purchases, product details, and customer demographics (e.g., region, signup date). The features considered for clustering include:

- Total transactions
- Total quantity purchased
- Total revenue
- Average transaction value
- Region-based one-hot encoding for geographical segmentation.

Clustering Algorithm Used: K-Means Clustering with K=2 clusters.

Number of Clusters Formed: 2 clusters were formed after applying KMeans clustering on the normalized feature data.

---

## 2. Number of Clusters Formed

Clusters Created: 2 clusters.

Cluster Size Distribution:

- Cluster 0: X0 customers
- Cluster 1: X1 customers

The clusters represent customer segments based on their transaction patterns, revenue, and geographical information.

---

## 3. Clustering Performance Metrics

Davies-Bouldin (DB) Index:

- Value: Y (DB index score)

- Interpretation: The Davies-Bouldin Index quantifies the separation and compactness of clusters. A lower value indicates better clustering results with well-separated and tight clusters.

Silhouette Score (optional, not calculated but can be added):

- If calculated, this score would show how well-separated the clusters are. A higher value close to +1 indicates that the clusters are well-separated and meaningful.

Inertia (within-cluster sum of squares):

- Value: W (inertia score)
- Interpretation: Inertia measures how well the clusters fit the data. Lower inertia values indicate that the clusters are compact and well-defined.

---

## 4. Cluster Centroids

Centroid Locations for K-Means Clustering (if available from kmeans.cluster_centers_):

- Cluster 0 Centroid: (x0, y0)
- Cluster 1 Centroid: (x1, y1)

---

## 5. Cluster Visualization

The customer clusters are visualized using a scatter plot of the first two principal components of the normalized feature data, with each point colored according to its cluster assignment.

Visualization Notes:

- Different colors and markers are used to represent the two clusters, with the axes corresponding to the first two principal components.

*(Plot not generated here—this should appear in the analysis)*

---

## 6. Cluster Separation

Euclidean Distance Between Centroids:

- Distance between Cluster 0 and Cluster 1: D (calculated from the centroids)

This metric helps evaluate how distinct the clusters are from each other.

## 7. Summary of Clustering Results

The clustering analysis grouped customers into 2 distinct clusters based on their transactional behavior, revenue, and geographical features. Below is a summary of each cluster:

- Cluster 0:
  - Number of Customers: X0
  - Average Revenue: Y0
  - Average Quantity: Z0
- Cluster 1:
  - Number of Customers: X1
  - Average Revenue: Y1
  - Average Quantity: Z1

---

## 8. Conclusion and Recommendations

The clustering process has successfully grouped customers into two segments that appear to differ in transaction behavior, quantity purchased, and total revenue.

Possible Applications:

- The clusters can be used for targeted marketing strategies, where Cluster 0 and Cluster 1 represent two distinct customer profiles (e.g., high-revenue vs. low-revenue customers, or regional differences).

Potential Improvements:

- Adjusting the number of clusters or trying a different clustering algorithm (e.g., DBSCAN or hierarchical clustering) might yield more refined results, particularly in the case of overlapping clusters.

---