

# Customer Segmentation / Clustering Report

---

The data used for clustering includes:

- **Customer Profile** (from the `Customers.csv` file)
- **Transaction Details** (from the `Transactions.csv` file)

We applied the **KMeans** clustering algorithm to segment the customers and used various metrics to evaluate the clustering performance.

---

## 1. Number of Clusters Formed

- **Clustering Algorithm Used:** KMeans.
  - **Optimal Number of Clusters:** Based on the **Elbow Method** (WCSS plot) and the **Silhouette Score**, the optimal number of clusters was chosen to be **2**.
  - **Reason for 2 clusters:** The **Elbow Method** suggested that the optimal number of clusters lies between 2 and 4. The **Silhouette Score** also supported this choice, with a value that indicated moderate clustering quality for 2 clusters.
- 

## 2. Evaluation Metrics

### Silhouette Score

- **Silhouette Score** measures how similar a point is to its own cluster compared to other clusters.
  - A higher score (closer to +1) indicates well-defined clusters, while a score close to 0 means the clusters are overlapping.
  - The silhouette score for this KMeans clustering is **0.288**, which suggests that the clusters are moderately separated, but not perfectly distinct.

**Silhouette Score (KMeans): 0.2878**

- Interpretation: The score indicates moderate clustering quality, suggesting that the two clusters formed are somewhat well-separated but may still have some overlap or fuzzy boundaries.

### Davies-Bouldin Index (DBI)

- The **Davies-Bouldin Index** (DBI) is a metric that evaluates the separation and compactness of clusters.
  - Lower DBI values indicate better clustering (i.e., clusters are more compact and well-separated).

- A higher DBI value suggests that the clusters are overlapping or not clearly defined.

### **Davies-Bouldin Index (KMeans): 3.183**

- Interpretation: This DBI value suggests that the two clusters formed are not very compact and well-separated. The clusters may not be distinct enough, and there could be some overlap. Ideally, a DBI value below 2 would indicate better separation between clusters.
- 

## **3. Other Relevant Clustering Metrics**

### **Cluster Sizes**

- **Cluster Distribution:** The clusters formed by KMeans were relatively balanced, with each cluster containing a reasonable number of customers. There was no extreme imbalance between the sizes of the clusters, which suggests that KMeans was able to effectively partition the data into meaningful groups.

### **PCA Visualization**

- **2D Visualization using PCA:** To visualize the clusters, **Principal Component Analysis (PCA)** was used to reduce the dimensionality of the data to 2 dimensions. A scatter plot was then created, with customers color-coded by their assigned cluster.