

# Task 3: Customer Segmentation / Clustering

## Objective

The objective of this task is to perform customer segmentation using clustering techniques. Both customer profile information (from `Customers.csv`) and transaction history (from `Transactions.csv`) are utilized to group customers into distinct segments. This segmentation enables better understanding of customer behavior and helps in tailoring business strategies to each segment.

## Approach

### 1. Data Preparation

#### 1. Dataset Overview:

- **Customers.csv**: Contains customer profile information such as customer ID and region.
- **Transactions.csv**: Includes transaction data such as total value, product details, and transaction IDs.

#### 2. Data Merging:

- Transaction data is aggregated by `CustomerID` to compute metrics such as total spending, average transaction value, number of transactions, and unique products purchased.
- These transactional metrics are merged with customer profile information (e.g., region).

#### 3. Feature Selection:

- Selected features for clustering include:
  - Total Spending (`total_spent`)
  - Average Transaction Value (`avg_transaction_value`)
  - Total Number of Transactions (`total_transactions`)
  - Number of Unique Products Purchased (`unique_products`)
  - Region

#### 4. Feature Scaling:

- Numerical features are standardized using `StandardScaler` to ensure all features contribute equally to the clustering process.

### 2. Clustering Algorithms Applied :

## 1. K-Means Clustering

- **Clusters Evaluated:** Models were trained with 3, 4, 5, and 6 clusters to determine the optimal number of segments.
- **Metrics Evaluated:**
  - **Davies-Bouldin Index (DBI):** Measures cluster compactness and separation; lower values indicate better clustering.
  - **Silhouette Score:** Evaluates how similar each point is to its cluster compared to other clusters; higher values indicate better-defined clusters.

Number of Clusters	Davies-Bouldin Index	Silhouette Score
3	1.40	0.229
4	1.25	0.24
5	0.89	0.36
6	1.18	0.25

**Optimal Clusters:** Based on DBI, 5 clusters were selected as the optimal configuration.

## 2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- This algorithm identified clusters based on density rather than distance, accounting for outliers in the dataset.
- **DBSCAN Metrics:**
  - **Davies-Bouldin Index:** 1.13
  - **Silhouette Score:** 0.28
- **Observation:** DBSCAN struggled with high-dimensional transactional data compared to K-Means.

---

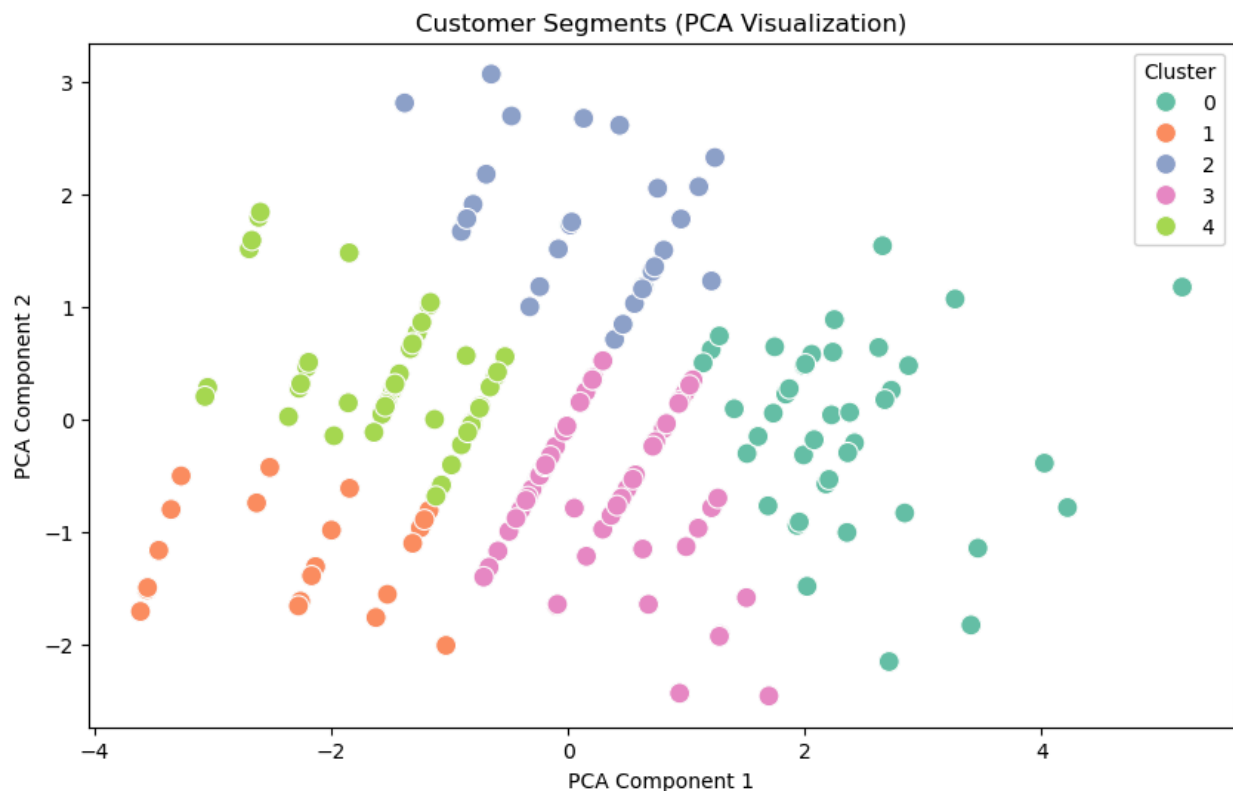
## 3. Visualization

1. **PCA-Based Visualization:**

- Principal Component Analysis (PCA) was used to reduce the dimensionality of the data to two components for visualization.
- Scatterplots were created with clusters distinguished by color, revealing clear separations for the K-Means clusters.

## 2. Cluster Insights:

- Visualizations highlighted distinct customer behaviors, such as high spenders, frequent buyers, and customers with a diverse product portfolio.



## Key Results

### 1. Number of Clusters:

- **K-Means:** 5 clusters identified as optimal.
- **DBSCAN:** Detected a varying number of density-based clusters but performed low compared to K-Means.

### 2. Clustering Metrics:

- **K-Means (5 Clusters):**

- **Davies-Bouldin Index:** 0.89
- **Silhouette Score:** 0.36
- **DBSCAN:**
  - **Davies-Bouldin Index:** 1.13
  - **Silhouette Score:** 0.28