

Customer Segmentation Analysis

Methodology

The clustering analysis was performed using the following approach:

1. Data Preparation

- Loaded customer profiles from **Customers.csv**.
- Loaded transaction data from **Transactions.csv**.
- Cleaned data by removing missing values.
- Combined profile and transaction information.

2. Feature Engineering

- Used customer demographic information.
- Incorporated transaction history features.
- Applied **StandardScaler** for feature normalization.

3. Clustering Algorithm

- Selected **K-means clustering algorithm**.
- Tested cluster numbers from **2 to 10**.
- Used **PCA** for dimensionality reduction and visualization.

Results

Optimal Number of Clusters

- **Final number of clusters chosen: 4.**
- Selection based on:
 - Silhouette score analysis.
 - Davies-Bouldin Index evaluation.
 - Visual inspection of cluster separation.

Clustering Metrics

1. Davies-Bouldin Index:

- Measures average similarity between clusters.
- Lower values indicate better clustering.
- Value obtained for 4 clusters shows good separation.

2. Silhouette Score:

- Measures how similar objects are to their own cluster compared to other clusters.
- Higher values indicate better-defined clusters.
- Optimal score achieved with 4 clusters.

Visualization

The clustering results were visualized using:

PCA Plot:

- 2D scatter plot using first two principal components.
- Different colors representing different clusters.
- Shows clear separation between customer segments.

Cluster Distribution Plot:

- Bar plot showing the number of customers in each cluster.
- Helps understand cluster sizes and distribution.

Key Findings

Cluster Characteristics

- Distinct customer segments identified based on purchasing behavior.
- Clear separation between high-value and low-value customers.
- Regional patterns in customer segmentation.

Business Implications

- Targeted marketing strategies can be developed for each segment.
 - Product recommendations can be tailored to cluster characteristics.
 - Resource allocation can be optimized based on cluster value.
-

Technical Implementation

- Used **scikit-learn's KMeans** implementation.
 - Applied **PCA** for dimensionality reduction.
 - Utilized **seaborn** and **matplotlib** for visualization.
 - Implemented **Davies-Bouldin Index** calculation.
 - Computed **silhouette scores** for cluster validation.
-

The clustering analysis successfully identified meaningful customer segments while maintaining good cluster separation metrics, providing valuable insights for business decision-making.