Customer Segmentation Analysis Report

Executive Summary

This report presents the results of a customer segmentation analysis using multiple clustering algorithms. The analysis was performed on customer profile and transaction data to identify distinct customer groups. Among the three clustering methods tested—K-Means, DBSCAN, and Gaussian Mixture Model (GMM)—K-Means emerged as the most effective approach, producing five distinct clusters.

Methodology

Data Preparation

1. Merged customer profile data with transaction information.
2. Engineered key features, including:
    - Total spent
    - Total quantity purchased
    - Number of transactions
    - Average transaction value
    - Recency (days since last purchase)
    - Frequency
    - Monetary value

Feature Selection and Preprocessing

Selected features were standardized using StandardScaler to ensure equal contribution to the clustering process. This step was crucial to handle features with varying scales and units.

Clustering Algorithms Evaluated

1. **K-Means Clustering**
2. **DBSCAN (Density-Based Spatial Clustering)**
3. **Gaussian Mixture Model (GMM)**

Results

Clustering Metrics

Davies-Bouldin Index (DB Index)

- **K-Means:** 1.029
- **DBSCAN:** 3.667
- **GMM:** 1.171

Note: Lower DB Index values indicate better cluster separation.

Silhouette Scores

- **K-Means:** 0.300
- **DBSCAN:** -0.280
- **GMM:** 0.209

Note: Higher Silhouette scores indicate better-defined clusters.

Optimal Clustering Solution

K-Means with 5 clusters was selected as the optimal solution based on the following:

1. Lowest Davies-Bouldin Index (1.029).
2. Highest Silhouette Score (0.300).
3. Clear visual separation of clusters in both PCA and t-SNE visualizations.

Visualization Analysis

The clustering results were visualized using two dimensionality reduction techniques:

1. **Principal Component Analysis (PCA)**
2. **t-Distributed Stochastic Neighbor Embedding (t-SNE)**

Both methods revealed distinct cluster formations, supporting the validity of the selected clustering solution.

Conclusions

1. K-Means clustering with 5 clusters provided the most robust customer segmentation.
2. The clustering solution achieved a DB Index of 1.029, indicating reasonable cluster separation.
3. The positive Silhouette Score of 0.300 suggests moderately well-defined clusters.

4. Both PCA and t-SNE visualizations confirmed the presence of distinct customer segments.

Technical Implementation Details

- **Implementation Language:** Python
- **Key Libraries:** scikit-learn, pandas, numpy, matplotlib
- **Data Processing:** StandardScaler for feature normalization
- **Dimensionality Reduction:** PCA and t-SNE for visualization

Recommendations

1. Use the K-Means clustering solution with 5 clusters for customer segmentation.
2. Consider the DB Index value of 1.029 as a baseline for future segmentation efforts.
3. Utilize the generated segments for targeted marketing and customer relationship management strategies.