Customer Segmentation Analysis Using K-Means Clustering

Data Science Team

January 27, 2025

1 Executive Summary

This report presents the results of a customer segmentation analysis using K-means clustering. The analysis identified 9 distinct customer segments based on purchasing behavior and regional characteristics. The clustering solution demonstrates good stability and separation between clusters, as evidenced by multiple validation metrics.

2 Methodology

2.1 Data Preprocessing

The analysis integrated three primary datasets:

- Customer demographic data
- Transaction records
- Product information

Key features engineered for the analysis include:

- Total spending per customer
- Number of transactions
- Regional indicators (one-hot encoded)

All features were standardized using z-score normalization to ensure equal weight in the clustering process.

2.2 Clustering Approach

The K-means algorithm was employed with the following specifications:

- Algorithm: K-means clustering
- Distance metric: Euclidean distance
- Number of clusters: 9 (determined through silhouette analysis)
- Random state: 42 (for reproducibility)

3 Results

3.1 Cluster Validation Metrics

The clustering solution achieved the following validation scores:

Metric	**Value**
Davies-Bouldin Index	0.727
Silhouette Score	0.459
Calinski-Harabasz Index	145.547

Table 1: Clustering Validation Metrics

3.2 Silhouette Analysis

Silhouette score analysis across different cluster numbers:

Number of Clusters	**Silhouette Score**
2	0.257
3	0.414
4	0.415
5	0.449
6	0.455
7	0.460
8	0.472
9	0.477
10	0.473

Table 2: Silhouette Scores for Different Cluster Numbers

4 Interpretation

4.1 Clustering Quality

The clustering solution demonstrates good quality based on multiple metrics:

- **Davies-Bouldin Index** (0.727): A relatively low value indicates good cluster separation.
- **Silhouette Score** (0.459): A moderate positive value suggests reasonable cluster cohesion and separation.
- **Calinski-Harabasz Index** (145.547): A high value indicates well-defined clusters.

4.2 Optimal Cluster Selection

The selection of 9 clusters is supported by:

- Peak silhouette score at n=9 (0.477)
- Diminishing returns in silhouette score beyond 9 clusters
- Balance between granularity and interpretability

5 Conclusions

The 9-cluster solution provides a robust segmentation of the customer base. The validation metrics suggest well-separated clusters with moderate cohesion. The silhouette analysis shows that 9 clusters represent an optimal balance between cluster definition and solution complexity.