Technical Report — Customer Segmentation Using Clustering Analysis

Dataset: `retail_customer_data.csv`
Notebook Source: `retail.ipynb`
Authors: Sophia Gabriela Martínez Albarrán, Sibyla Vera Ávila, Regina Pérez Vázquez
Date: 25/11/2025

Executive Summary

Objective: Identify natural customer segments within MegaMart's behavioral dataset to support personalized marketing, improve retention, and guide resource allocation.

Key Finding: A four-cluster solution provides the best balance of cohesion and separation (Silhouette ≈ 0.317). The segments differ clearly in spending, engagement, browsing activity, tenure, and recency.

The segmentation enables meaningful financial opportunities: reactivating 5% of inactive customers, increasing conversions among browsers, and retaining a small percentage of high-value shoppers results in an estimated **$285K+ USD** annual impact.

1. Data & Preprocessing

Dataset Overview:

- 3,000 customer observations
- 9 behavioral variables: monthly transactions, basket size, total spend, session duration, email open rate, product views per visit, return rate, tenure, recency
- No missing values and no formatting issues

Standardization: All numeric variables were standardized using `StandardScaler` to ensure equal weight in clustering algorithms.

2. Exploratory Data Analysis (EDA)

Distribution Patterns:

- Spending variables show right-skewed behavior.
- Engagement varies widely among customers.
- Email open rate shows many extreme values (0 or 0.95).

Correlations:

- Basket size and total spend: $r = 0.94$
- Monthly transactions and total spend: $r = 0.76$
- Monthly transactions and recency: negative correlation

Outliers: Outliers in spending, session duration, and browsing activity were kept because they represent valid high-value customers.

3. Hierarchical Clustering

Methods compared: Single, Complete, Average, and Ward linkage. Ward linkage showed the cleanest separation and compact clusters.

Cluster Counts Tested: $k = 3, 4, 5, 6$

Silhouette Scores:

- $k = 3$: 0.295
- $k = 4$: **0.316 (best)**
- $k = 5$: 0.300
- $k = 6$: 0.247

Conclusion: Four clusters offer the best balance of interpretability and separation.

4. K-Means Clustering & Validation

Using $k = 4$:

Cluster Sizes:

- Cluster 0: 525 customers (17.5%)
- Cluster 1: 929 customers (31.0%)
- Cluster 2: 433 customers (14.4%)
- Cluster 3: 1113 customers (37.1%)

Model Validation:

- Average Silhouette Score: 0.317
- Clusters 0 and 3 show strongest cohesion.
- Few borderline or poorly assigned customers.

5. Cluster Profiles & Interpretation

Cluster 0 — High-Value Loyalists (17.5%) Very high spend, large baskets, high browsing, low return rate. Highly engaged and consistent.

Cluster 1 — At-Risk Minimal Shoppers (31%) Lowest spend, long periods of inactivity, minimal engagement. High churn risk.

Cluster 2 — Established Steady Buyers (14.4%) High basket size, high spend, long tenure. Predictable steady customers.

Cluster 3 — Browsers with Moderate Spend (37.1%) High product views, moderate spending, medium engagement. Strong candidates for conversion campaigns.

6. PCA Visualization

Principal Component Analysis (PCA) reduced nine dimensions into two components:

- PC1: 41% of variance
- PC2: 21% of variance

Total variance captured: **61.98%**

The PCA scatterplot shows clear separations, though it is only a projection of the full 9D structure.

7. Business Insights

- Customer behavior naturally divides into four meaningful segments.
- Spending and engagement are primary drivers of customer value.
- Recency strongly predicts churn risk.
- Browsing activity signals potential growth through targeted conversion.

8. Strategic Recommendations

1. Retain High-Value Loyalists VIP programs, exclusive access, recognition messages. *Financial impact: protects $104K/year.*

2. Reactivate At-Risk Shoppers Seasonal offers, reminders, personalized discounts. *Impact: $19K/year recovered.*

3. Convert Browsers with Moderate Spend Browse-triggered promotions, bundles, tailored suggestions. *Impact: $161K/year potential.*

4. Strengthen Steady Buyers Curated recommendations, upselling aligned to behavior patterns.

9. Limitations & Next Steps

Limitations:

- Clustering sensitive to scaling and variable choices.
- PCA is only an approximation of multidimensional space.
- No categorical data included.

Next Steps:

- Add product-category features for deeper segmentation.
- Build a predictive churn model incorporating segment labels.
- Track customer transitions between segments.
- Conduct A/B testing for targeted marketing actions.

Appendix — Code Reference

Full reproducible analysis, including dendrograms, heatmaps, Silhouette plots, PCA figures, and profiling tables, is available in the notebook `retail.ipynb`.