Technical Report

MegaMart

Sofia Knutas, A01831481
Kristina Kristiansen Stapnes, A01830792

Aplicación de métodos multivariados en ciencia de datos (Gpo 601)
Group 6

1. Objective

The objective of this analysis is to segment customers into distinct groups using multivariate clustering methods, based on behavioral variables, and determine the optimal number of clusters through statistical validation techniques.

2. Data Overview

The dataset contains 3000 observations and 10 variables, where one is an identifier and nine are numerical behavioral features. No missing values were detected. Descriptive statistics confirmed valid ranges and variability across features.

3. Exploratory Data Analysis

Histograms were used to examine variable distributions, showing varying spread and skewness across spending and activity metrics.

Correlation analysis identified strong dependencies, particularly:

- total_spend and avg_basket_size (r = 0.94)
- total_spend and monthly_transactions (r = 0.76)
- avg_basket_size and monthly_transactions (r = 0.69)

Boxplots were used to identify outliers in key variables, and scatterplots illustrated feature relationships.

4. Data Preprocessing

Since clustering is distance-based, all numerical variables were standardized using StandardScaler, resulting in transformed variables with mean = 0 and standard deviation = 1. The scaled dataset was used in all subsequent clustering procedures.

5. Hierarchical Clustering

Hierarchical clustering was executed using four linkage criteria:

- single
- complete

- average
- ward

Ward linkage produced the most distinct separation and minimized chaining effects. Dendrogram inspection suggested cluster separations at higher aggregation heights.

6. Determining Optimal Cluster Count (Hierarchical)

Silhouette scores were computed for k = 3, 4, 5, and 6 clusters. The highest score occurred at k = 4, indicating the strongest cohesion and separation structure.

7. K-Means Cluster Evaluation

K-Means clustering was applied for k = 2 to 10.

Two criteria were analyzed:

- Inertia (Elbow method)
- Silhouette Score

Both measures indicated that k = 4 was optimal, with diminishing improvement beyond this point. This confirmed the hierarchical results.

8. Final Clustering Model

A K-Means model with k = 4 was fitted on the standardized data. Cluster assignments were added to the dataset, and cluster sizes were calculated.

Mean variable profiles were computed for each cluster and visualized using a heatmap to compare group characteristics.

9. Cluster Validation

Silhouette coefficients were computed for all observations.

Findings:

- Overall silhouette score was clearly positive

- Clusters 0 and 3 showed the strongest cohesion
- Cluster 1 showed moderate cohesion
- Cluster 2 showed the weakest separation but remained viable

Only a small number of points exhibited negative silhouette values, indicating few misclassified observations.

10. PCA-Based Visualization

Principal Component Analysis (PCA) was performed to reduce dimensionality for visualization. The first two components explained a meaningful proportion of total variance. Cluster assignments were plotted in the 2D PCA space along with centroid positions. The projection provided an interpretable visualization, acknowledging that clustering itself occurred in 9-dimensional space.

11. Technical Conclusion

The analysis demonstrates that:

- Standardization was required due to scale differences
- Ward linkage produced the clearest hierarchical structure
- Both hierarchical and K-Means validation methods converged on k = 4 as the optimal cluster solution
- Silhouette statistics confirmed acceptable separation and internal cohesion
- PCA projection provided visual confirmation of distinguishable cluster groupings