Report_group14

Executive Summary

This project focuses on customer segmentation for a large retail business, aiming to identify distinct customer profiles based on demographic characteristics, behavioral patterns, and purchasing preferences. The core business problem addressed is the need to better understand customer heterogeneity in order to support more effective marketing strategies, personalization efforts, and resource allocation.

To achieve this, we employed an unsupervised learning approach using KMeans clustering. The dataset included over 34,000 customers and featured variables such as age, gender, household composition, loyalty card usage, store visit behavior, and lifetime spending across multiple product categories. Extensive data preprocessing was performed, including feature engineering (e.g., age and customer tenure), missing value imputation, and standardization. Dimensionality reduction via Principal Component Analysis (PCA) was used both for visualization and to support interpretation of clustering results.

As a result, we identified four distinct customer segments. These clusters differed in terms of age distribution, total spending, product category preferences (e.g., alcohol, hygiene, or groceries), and engagement metrics such as store visits and loyalty card usage. Some segments were characterized by younger, promotion-sensitive customers with low spending, while others included older, high-value customers with broader purchasing behavior. These insights form the foundation for developing targeted promotion strategies tailored to each group.

Exploratory Data Analysis

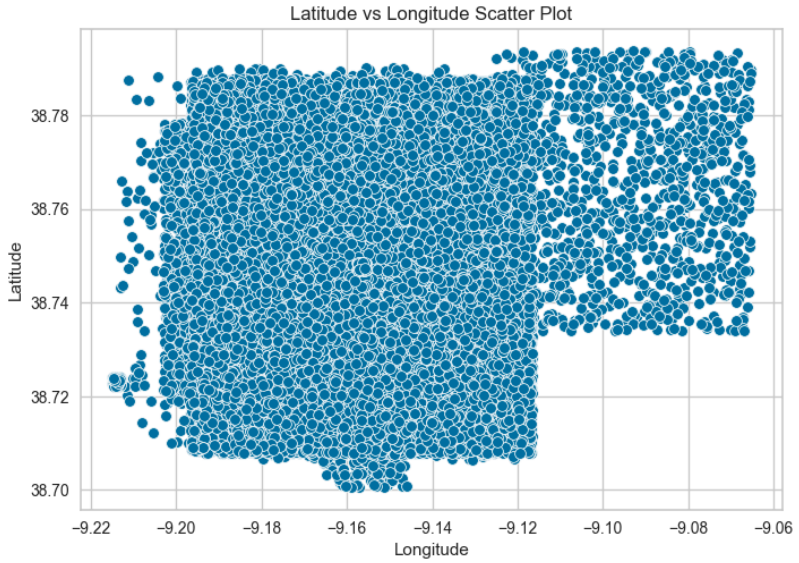
For this project, we used a dataset comprising demographic, behavioral, and geographic data on retail customers. The original dataset included variables such as gender, birthdate, year of first transaction, loyalty card number, geographic coordinates, in-store behavior, and transaction history.

Data Cleaning and Feature Engineering

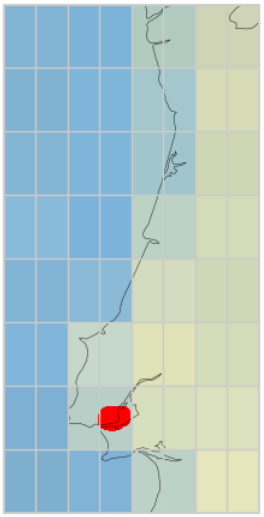
Irrelevant columns, such as customer names and export-specific indices, were removed at the initial stage. The gender variable was transformed into a binary indicator, and a new age variable was constructed based on the customer's year of birth. Similarly, a tenure variable reflecting the number of years since the first recorded transaction was created. Both birthdate and transaction year were discarded after feature construction.

Loyalty card status was derived from the presence of a card number and converted into a binary variable indicating whether the customer possessed a loyalty card.

To capture geographic aspects, latitude and longitude were converted into spatial points. The customer distribution across coordinates was visualized using the following scatter plot:



A second visualization used cartographic projection to display customer positions on a regional map:



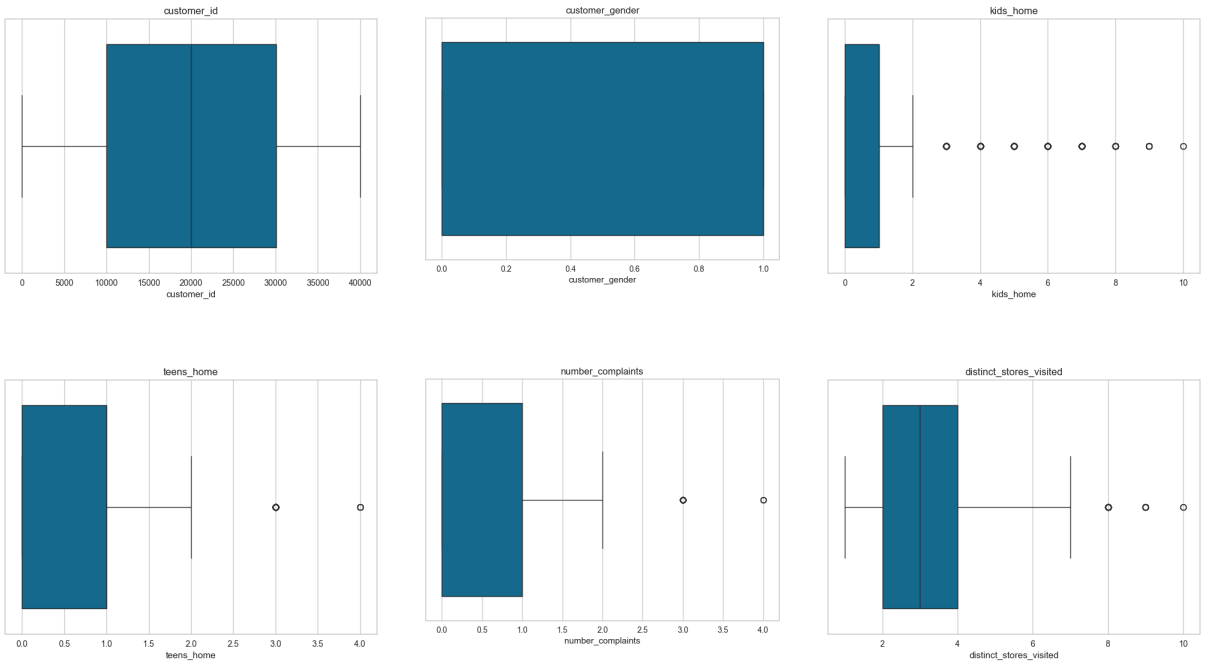
A new feature, distance from a central reference point (Lisbon), was computed to measure customer proximity to the business's central region. This feature serves as a proxy for geographic accessibility and regional segmentation. Afterward, the original latitude, longitude, and geometry columns were removed.

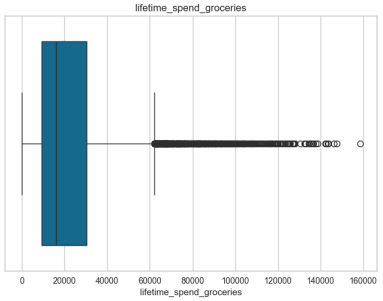
Handling Missing Values

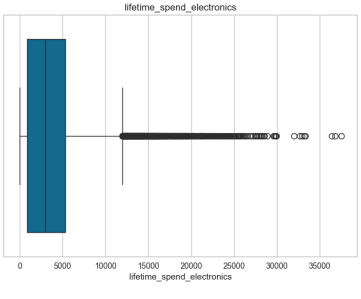
Several columns contained missing values. Binary features such as presence of children, teenagers, or complaints were imputed with zeros, assuming absence of corresponding events. Store visit count was imputed with one as the minimal plausible activity level. Lifetime spending variables were filled with zeros, under the assumption that missing entries corresponded to no purchase activity. Continuous variables like age and typical shopping hour were imputed using the median value across the dataset. After these steps, no missing values remained.

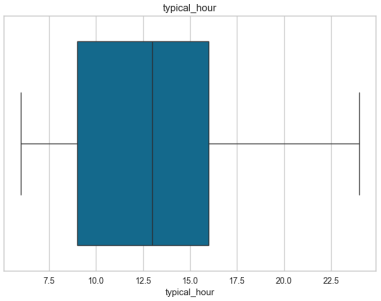
Outlier Treatment

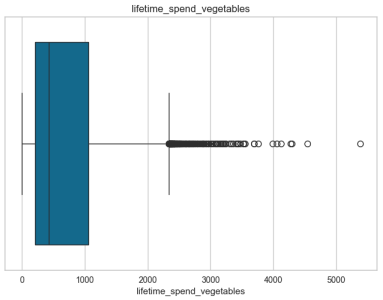
Outliers were identified through visual inspection of boxplots:

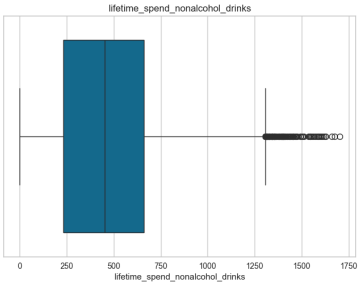


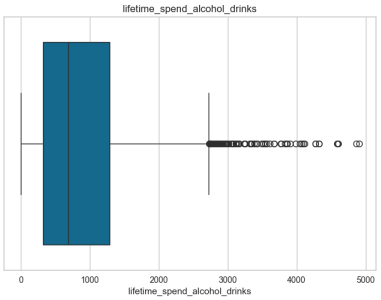

lifetime_spend_groceries
0
20000
40000
60000
80000
100000
120000
140000
160000
lifetime_spend_groceries

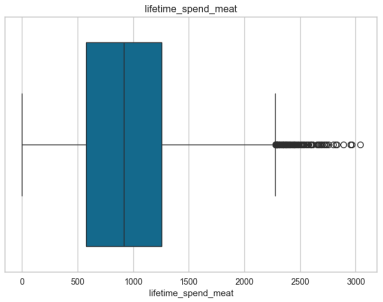

lifetime_spend_electronics
0
5000
10000
15000
20000
25000
30000
35000
lifetime_spend_electronics

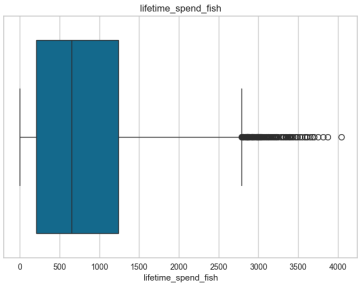

typical_hour
7.5
10.0
12.5
15.0
17.5
20.0
22.5
typical_hour

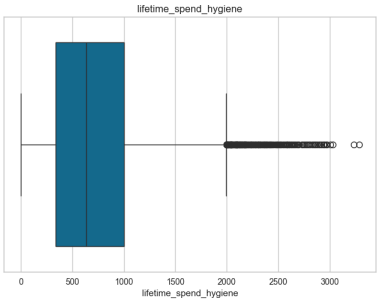

lifetime_spend_vegetables
0
1000
2000
3000
4000
5000
lifetime_spend_vegetables

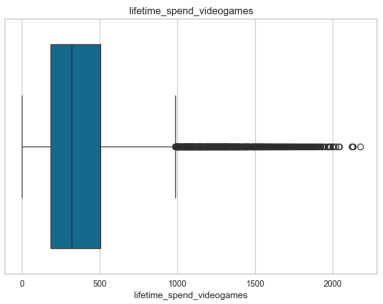

lifetime_spend_nonalcohol_drinks
0
250
500
750
1000
1250
1500
1750
lifetime_spend_nonalcohol_drinks

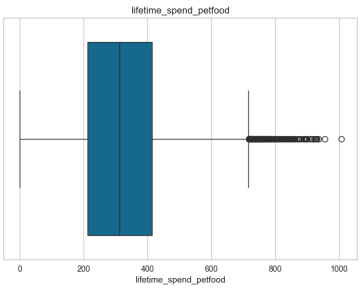

lifetime_spend_alcohol_drinks
0
1000
2000
3000
4000
5000
lifetime_spend_alcohol_drinks

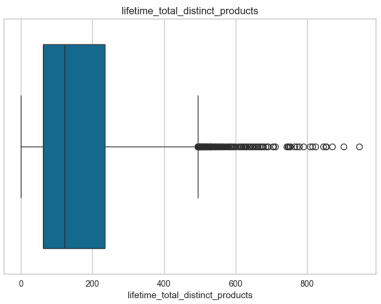

lifetime_spend_meat
0
500
1000
1500
2000
2500
3000
lifetime_spend_meat

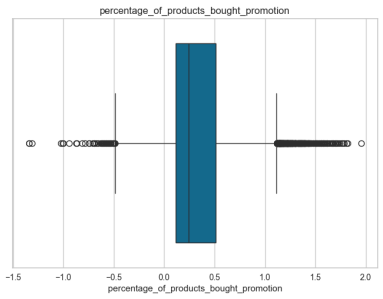

lifetime_spend_fish
0
500
1000
1500
2000
2500
3000
3500
4000
lifetime_spend_fish

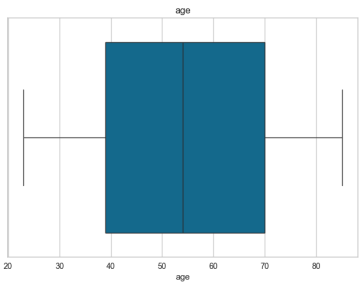

lifetime_spend_hygiene
0
500
1000
1500
2000
2500
3000
lifetime_spend_hygiene

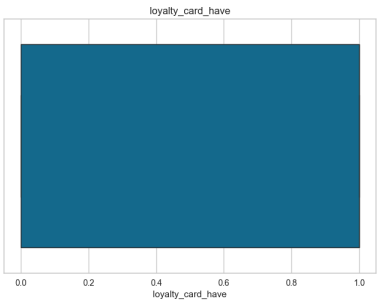

lifetime_spend_videogames
0
500
1000
1500
2000
lifetime_spend_videogames

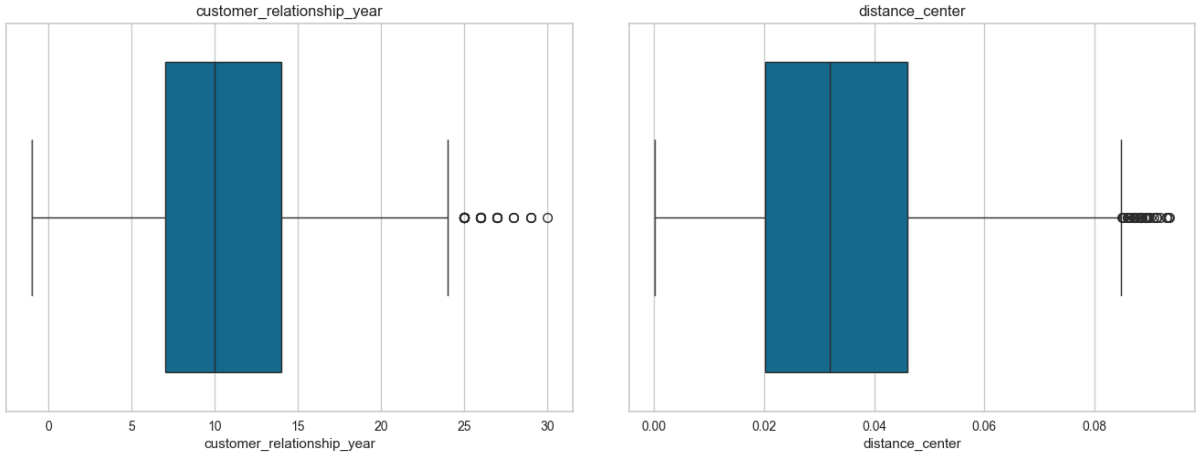

lifetime_spend_petfood
0
200
400
600
800
1000
lifetime_spend_petfood


lifetime_total_distinct_products
0
200
400
600
800
lifetime_total_distinct_products


percentage_of_products_bought_promotion
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
percentage_of_products_bought_promotion


age
20
30
40
50
60
70
80
age


loyalty_card_have
0.0
0.2
0.4
0.6
0.8
1.0
loyalty_card_have

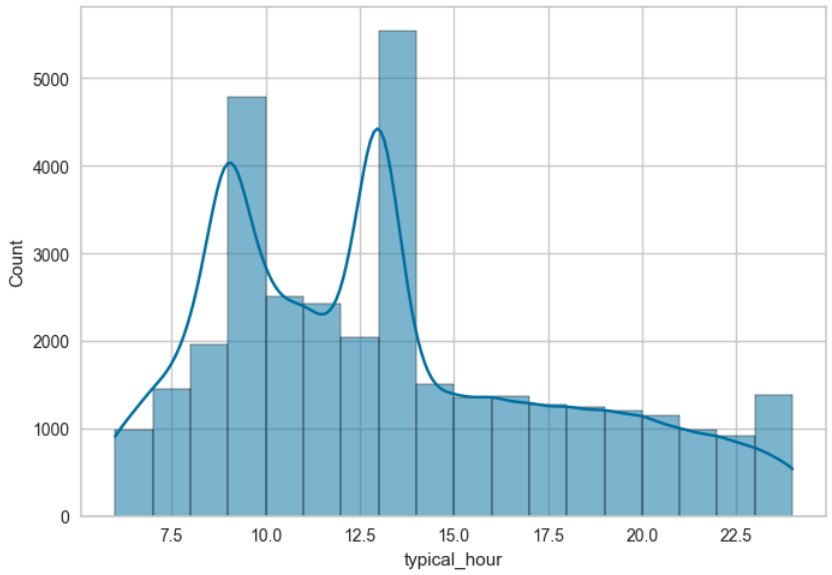


Outlier proportions were estimated, and features exhibiting high proportions—such as number of complaints or stores visited—were winsorized to cap extreme values and reduce skewness. This allowed for improved robustness in downstream modeling.

Additional Features

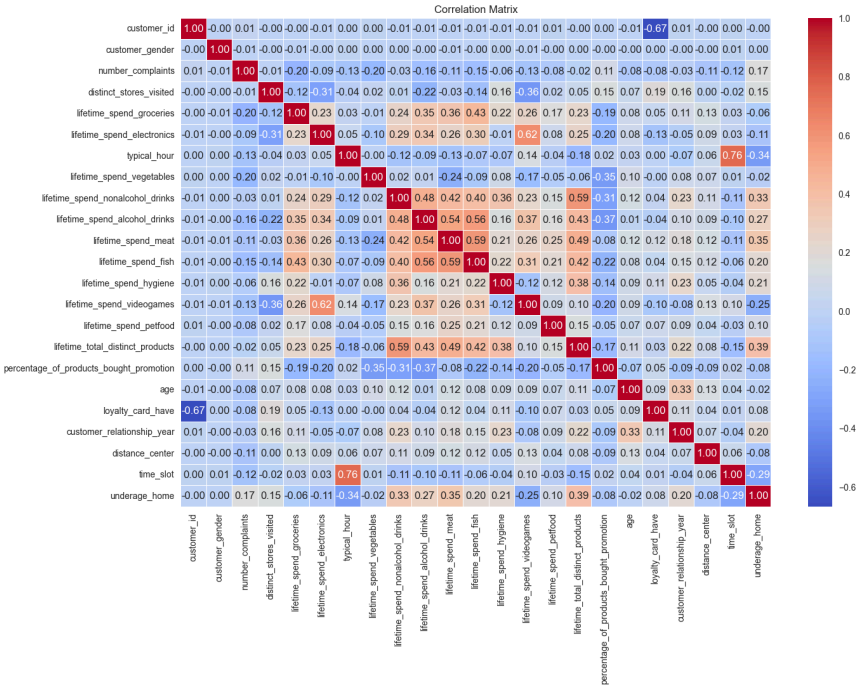
The variable `typical_hour` was converted into a categorical feature representing daily time slots (e.g., morning, afternoon, evening). Another feature was engineered by summing the number of children and teenagers in the household into a single variable indicating total underage household members.

A histogram was created to inspect the distribution of shopping hours:



Correlation Analysis

A correlation matrix was computed and visualized to examine interdependencies among numerical variables:

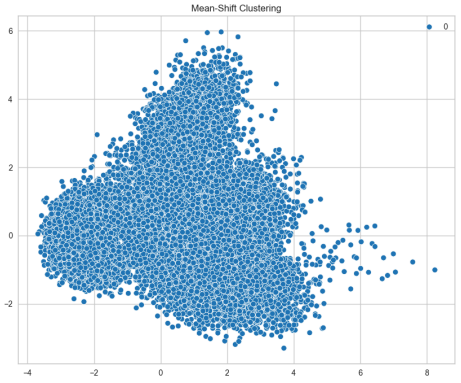


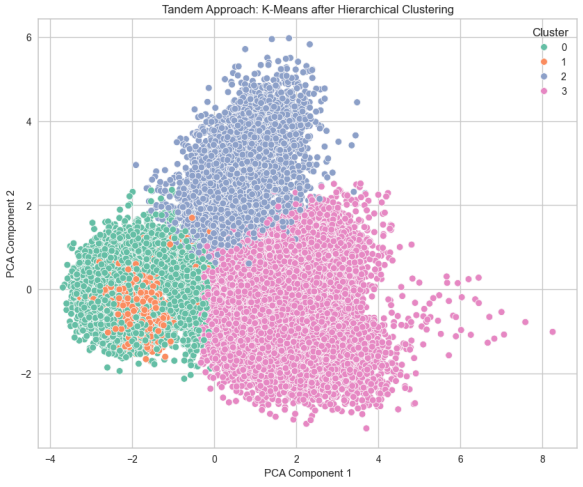
Strong positive correlations were observed among the different lifetime spending categories, suggesting consistency in customer value across product groups. Weaker correlations were noted between demographic and behavioral variables.

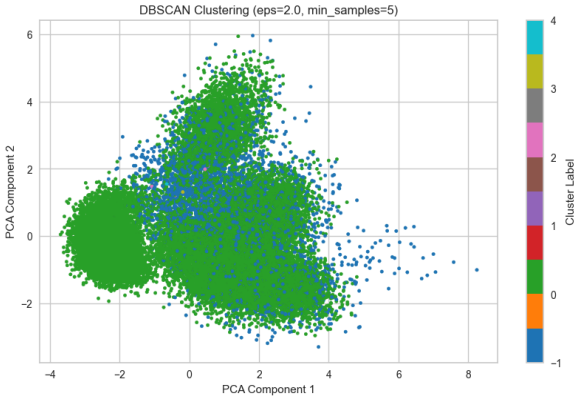
Customer Segmentation

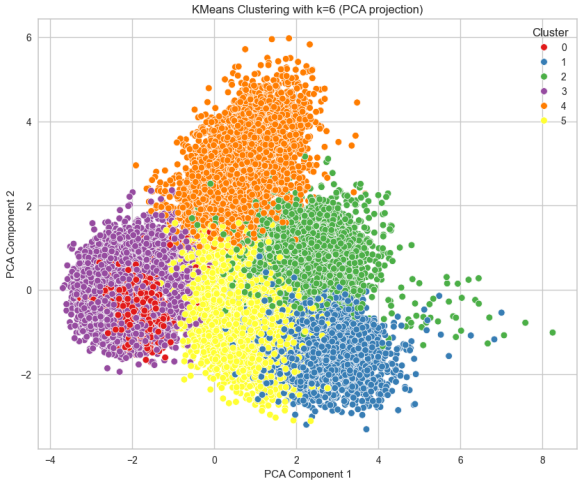
Segmentation Approach

To uncover distinct groups of customers with shared characteristics, we employed a multi-method segmentation strategy combining KMeans clustering, DBSCAN, Mean Shift, and a tandem clustering approach using PCA followed by KMeans. This hybrid methodology enabled both quantitative robustness and visual interpretability. KMeans served as the primary segmentation tool due to its efficiency and scalability with large customer datasets. To ensure the validity of the KMeans results, we utilized the Elbow Method and Silhouette Score to determine the optimal number of clusters. In parallel, DBSCAN was introduced to identify non-spherical, density-based clusters and detect outliers that might skew centroid-based algorithms. Additionally, we explored Mean Shift clustering—a non-parametric technique that automatically determines the number of clusters based on data density. By applying Mean Shift to a PCA-reduced feature space, we were able to uncover dense customer groupings without relying on pre-specified cluster counts, capturing organic behavioral patterns and validating the presence of cohesive segments discovered through KMeans. Finally, we incorporated a tandem clustering technique by reducing the dimensionality of the standardized features via PCA before applying KMeans again. This enhanced the separability of clusters in reduced feature space and improved cluster visualization.









Pre-Segmentation Heuristics

The data was thoroughly preprocessed before segmentation. We engineered the `age` feature from birthdates and computed customer tenure from the first transaction year. Demographic variables such as gender, household composition, and purchasing hours were included. Spending variables covered ten distinct product categories, such as groceries, hygiene, and electronics. All numerical variables were standardized using `StandardScaler`, and missing values were imputed with domain-relevant defaults (e.g., median for age, zeros for complaints or spending). Outliers were not removed due to the robustness of KMeans, though scaling helped mitigate their impact.

Cluster Summary and Characteristics

We choose KMeans clustering with 6 clusters using a standardized feature set including age, household characteristics, shopping behavior, and category-specific spending. Prior to clustering, we imputed missing values using mean imputation and standardized the data. PCA was used for visualization but not for clustering itself.

Summary of customer segments:

- **Cluster 0:** Older adults (~56) with small-to-medium baskets and low promotion sensitivity (9%). Spend primarily on groceries, hygiene products, and pet food. Represent stable and loyal customers with consistent shopping habits.
- **Cluster 1:** Older adults (~56) with very high lifetime spend across all categories, including electronics, groceries, hygiene, alcohol, and pet food. Moderate promotion sensitivity (~20%). Represent high-income, family-oriented power shoppers.
- **Cluster 2:** Older VIP customers (~56), with the highest grocery spend and overall large baskets. Low promotion sensitivity (10%). Represent the most loyal and valuable segment in terms of lifetime value.
- **Cluster 3:** Younger adults (~29) with the highest promotion sensitivity (62%) and lower overall spend. Their baskets are smaller, with frequent use of discounts and promotions. Represent price-conscious, promotion-driven consumers.
- **Cluster 4:** Older customers (~56), electronics-heavy shoppers with significant spend on alcohol and hygiene. Promotion sensitivity ~31%. Represent affluent customers with diverse category interests.
- **Cluster 5:** Older adults (~56), large balanced baskets with strong spend in groceries, hygiene, electronics, and pet food. High promotion sensitivity (~40%). Represent affluent, high-frequency shoppers with cross-category buying behavior.

Each segment varies in terms of size, average spend, promotion sensitivity, and basket composition, offering valuable insights for personalized marketing strategies.

Segment Size Comparison

The clusters produced by KMeans were well balanced in terms of customer volume, with each cluster comprising approximately 20–30% of the overall population. Based on its stability, scalability, and interpretability, **KMeans was selected as the primary segmentation method** for downstream analysis and campaign design. DBSCAN identified a limited number of core dense regions but flagged over 4,000 customers as noise, revealing its limited suitability for this high-dimensional retail dataset and highlighting a subset of potential behavioral outliers. Mean Shift, in contrast, generated fewer but denser clusters by merging similar behavioral profiles, especially across medium- and high-spending groups. While more computationally intensive, its results largely reinforced the presence of core customer types uncovered via KMeans. The tandem PCA + KMeans approach yielded visually well-separated clusters with consistent segment definitions, further validating the reliability of the chosen segmentation structure. Though exploratory, these alternative methods confirmed the internal consistency of the final KMeans-based segmentation.

Targeted Promotion

After performing customer segmentation with KMeans (6 clusters), we identified distinct customer segments with unique profiles based on age, basket composition, promotion sensitivity, and category preferences.

We propose the following **targeted promotion plan**:

Segment 1 (KMeansCluster = 0)

Profile:

- Older customers (~54 y.o.)
- Balanced spend on **groceries**, **hygiene**, **pet food**
- Low promotion sensitivity (9%)
- Small-to-medium basket

Promotions Approach: Stimulate cross-sell in non-promotional categories.

Mechanics:

- Buy 1 hygiene product → get 50% off pet food
- Bundle: groceries + non-alcoholic drinks → 10% discount

Segment 2 (KMeansCluster = 1)

Profile:

- Older customers (~54 y.o.), many with children (underage_home ≈ 4)
- High total spend, very large basket
- High interest in **groceries**, **hygiene**, **alcohol**, **pet food**, **electronics**
- Moderate promotion sensitivity (20%)

Promotions Approach: Encourage premium product upsell.

Mechanics:

- Buy premium groceries → get discount on electronics
- Combo: hygiene + alcohol → 15% discount

Segment 3 (KMeansCluster = 2)

Profile:

- Older "VIP" customers (~54 y.o.)
- Very high total spend → top segment
- Very loyal → low promo sensitivity (10%)
- Massive spend on **groceries**, **fish**, **meat**

Promotions Approach: Loyalty rewards to retain VIPs.

Mechanics:

- Buy meat → get 25% off fish
- Exclusive offers: 15% discount on full basket over X€

Segment 4 (KMeansCluster = 3)

Profile:

- Young adults (~29 y.o.)
- High promotion sensitivity (62%)
- Small basket, occasional buyers
- Focused on **groceries**, **hygiene**, **non-alcoholic drinks**

Promotions Approach: Drive frequency and habit-building.

Mechanics:

- 50% off entire basket for next visit
- Buy 2 drinks → get 1 free snack

Segment 5 (KMeansCluster = 4)

Profile:

- Older customers (~54 y.o.)
- Low underage_home → possibly single or couple households
- High spend on **electronics**, **alcohol**, **fish**, **hygiene**
- High promotion sensitivity (31%)

Promotions Approach: Stimulate complementary category purchases.

Mechanics:

- Buy electronics → get 30% off alcohol
- Bundle: fish + hygiene → 20% discount

Segment 6 (KMeansCluster = 5)

Profile:

- Older customers (~54 y.o.)
- Large basket, high spend on **groceries**, **hygiene**, **video games**, **electronics**
- High promotion sensitivity (40%)

Promotions Approach: Cross-category bundle promotions.

Mechanics:

- Buy 1 video game → get 50% off hygiene product
- Buy groceries + electronics → 20% off total basket

Summary of the Promotion Approach

- **Personalization:** Promotions aligned with age, family structure, basket composition.
- **Mechanics:** Combination of *bundle offers*, *slash prices*, *loyalty rewards*, and *habit-building promotions*.
- **Goals:** Increase loyalty in VIP segments, stimulate basket growth, re-engage promotion-driven segments.

Conclusions and Recommendations

In this project, we applied several unsupervised machine learning techniques to perform customer segmentation based on customer demographics and purchasing behavior. We explored KMeans, Tandem KMeans, DBSCAN, and MeanShift clustering methods. After detailed evaluation of the segment profiles and their business interpretability, we selected **KMeans with 6 clusters** as the final segmentation model. KMeans provided well-balanced and stable clusters with clear behavioral differences, suitable for actionable marketing strategies. While Tandem KMeans and DBSCAN offered interesting alternative views of the customer base, and MeanShift produced poor results with highly unbalanced clusters, KMeans was ultimately chosen for its simplicity, stability, and ease of implementation in business processes.

Based on the identified KMeans segments, we developed a targeted promotion strategy aligned with the specific characteristics of each customer group. These strategies aim to foster customer loyalty, drive cross-category purchases, and re-engage promotion-sensitive segments. The integration of Association Rules within each segment provided further actionable insights to support personalized campaign design. We recommend adopting the KMeans-based segmentation for operational use and suggest exploring enhancements through temporal analysis and additional data enrichment in future iterations.