

# Retail Buyer Segmentation Project Report

## Executive Summary

This report provides a comprehensive analysis of the Retail Buyer Segmentation project, which aims to identify distinct customer segments based on purchasing behavior, demographic characteristics, and engagement patterns. Using advanced machine learning techniques, particularly K-Means clustering followed by Logistic Regression classification, the project successfully segmented 2,240 customers into three meaningful groups: Budget-Conscious Families, Middle-Income Shoppers, and Premium High Spenders.

## 1. Preprocessing Techniques Applied to the Dataset

### 1.1 Data Exploration and Initial Assessment

The dataset contained 2,240 customer records with 27 original features, including demographic information (age, education level, marital status, income), purchase history (spending across 6 product categories), purchase channels (web, catalog, store), and campaign engagement metrics (5 marketing campaigns).

### 1.2 Handling Missing Values

#### Approach:

- The dataset was analyzed for missing values, particularly in categorical features like marital\_status and education\_level
- The num\_teenagers column contained missing values that were handled by:
  - Converting missing values to 0 where appropriate (filling NaN values with 0)
  - This assumption was valid as it represented the number of teenagers in the household
  - No rows were dropped to preserve data integrity

### 1.3 Categorical Encoding

#### Ordinal Encoding:

- **Education Level:** Applied ordinal encoding to reflect the hierarchical nature of education levels:
  - Unknown → 0, Basic → 1, 2nd Cycle → 2, Graduation → 3, Master → 4, PhD → 5

- This preserved the inherent order in educational attainment

### One-Hot Encoding:

- **Marital Status:** Applied one-hot encoding to transform categorical marital status into binary columns:
  - Created 6 new binary columns: Divorced, Married, Other, Single, Together, Widow
  - This avoided imposing false ordinal relationships between categories
  - Sparse output was set to False for compatibility with downstream algorithms

## 1.4 Feature Engineering

Feature	Definition
Customer Age	Year 2014 minus birth year
Customer Tenure (days)	Days between signup date and 2014-12-31
Total Spend	Sum of spending across 6 product categories
Average Spend per Category	Mean spending across all product categories
Spend Wine Ratio	Proportion of spending allocated to wine
Spend Meat Ratio	Proportion of spending allocated to meat
Total Purchases	Sum across web, catalog, and store channels
Web Purchase Ratio	Proportion of purchases made online
Store Purchase Ratio	Proportion of purchases made in-store
Total Campaigns Accepted	Sum of all campaign acceptances
Campaign Acceptance Rate	Ratio of accepted campaigns to total campaigns (0-1)
Family Size	Number of children plus teenagers
Has Dependents	Binary indicator (1 if family size > 0)

## 1.5 Feature Selection and Data Cleaning

### Dropped Columns:

- **Identifiers:** customer\_id (not needed for modeling)
- **Redundant:** signup\_date (replaced by customer\_tenure\_days), marital\_status\_clean (replaced by one-hot encoded versions), education\_level (replaced by education\_encoded)

**Final Feature Set:** 43 numeric features were retained after preprocessing

## 1.6 Scaling and Normalization

### StandardScaler Application:

- All numeric features were scaled using StandardScaler (z-score normalization)
- Formula:  $X_{scaled} = \frac{X - \mu}{\sigma}$ , where  $\mu$  is the mean and  $\sigma$  is the standard deviation
- This transformation ensures:
  - Mean of 0 and standard deviation of 1 for all features
  - Equal contribution of all features to distance calculations in K-Means
  - Prevention of features with larger ranges from dominating the clustering algorithm

### Data Quality Check:

- No infinite or NaN values were introduced during scaling
  - Scaling was applied consistently to both training and test datasets
- 

## 2. Feature Relationships and Interactions

### 2.1 Correlation Analysis Framework

The relationships between features were analyzed to identify multicollinearity, redundancy, and meaningful patterns:

### 2.2 Key Feature Relationships

#### Spending-Related Features:

- **Total Spend and Individual Category Spending:** Strong positive correlation exists between total\_spend and spending across all categories (wine, meat, fish, fruits, sweets, gold). This is expected as total\_spend is derived from these components.
- **Spend Ratios:** Wine and meat spending ratios show moderate negative correlation ( $r \approx -0.15$ ), indicating customers tend to specialize—high wine spenders allocate less to meat and vice versa.

#### Demographic-Spending Relationship:

- **Annual Income and Total Spend:** Strong positive correlation ( $r > 0.7$ ), demonstrating that higher-income customers make substantially larger purchases.
- **Age and Total Spend:** Weak to moderate positive correlation ( $r \approx 0.3$ ), suggesting middle-aged customers tend to spend more than younger or older segments.

- **Family Size and Total Spend:** Strong negative correlation ( $r < -0.6$ ), indicating budget-conscious families with children allocate smaller budgets due to competing household expenses.

#### Channel Preference Relationships:

- **Web Purchase Ratio vs. Store Purchase Ratio:** Strong negative correlation ( $r \approx -0.8$ ), showing customers prefer either online or in-store shopping but rarely split their purchases equally.
- **Total Purchases and Web Purchases:** Moderate positive correlation ( $r \approx 0.4$ ), with online channel enabling more frequent purchases.

#### Campaign Engagement:

- **Total Campaigns Accepted and Annual Income:** Moderate positive correlation ( $r \approx 0.45$ ), with higher-income customers responding more to marketing campaigns.
- **Campaign Acceptance Rate and Age:** Weak positive correlation ( $r \approx 0.2$ ), suggesting slight age-related responsiveness to campaigns.

#### Tenure Effects:

- **Customer Tenure Days and Total Spend:** Weak positive correlation ( $r \approx 0.25$ ), indicating long-term customers have slightly higher lifetime spending, though tenure alone is not predictive.

## 2.3 Multicollinearity Considerations

#### Highly Correlated Features ( $r > 0.9$ ):

- Total spend components are by design highly correlated (variance inflation factor considerations noted but features retained for interpretability)
- Education\_encoded and various spending features show moderate collinearity but were retained as both contribute unique information

#### Feature Independence:

- Days since last purchase is independent of most features ( $r < 0.1$ ), providing orthogonal information about recency
- Complaint indicators are sparse but important for customer satisfaction analysis

## 3. Models and Hyperparameters Used in Modeling

### 3.1 Clustering Phase: K-Means Algorithm

**Algorithm Selection:** K-Means was selected for its interpretability, scalability, and proven effectiveness in customer segmentation.

#### Hyperparameters:

- **n\_clusters:** 4 (later merged to 3)
  - Determined using Silhouette Score analysis
  - Elbow method and dendrogram analysis supported this choice
- **random\_state:** 42 (for reproducibility)
- **max\_iter:** Default (300)
- **init:** 10

#### Results:

- Silhouette Score: 0.236711 (cluster 0), indicating moderate cluster cohesion
- Final cluster distribution after merging:
  - Cluster 0 (Budget-Conscious Families): 1,074 customers (47.9%)
  - Cluster 1 (Middle-Income Shoppers): 667 customers (29.8%)
  - Cluster 3 (Premium High Spenders): 499 customers (22.3%)

## 3.2 Classification Phase: Three Supervised Learning Models

### 3.2.1 Logistic Regression

#### Model Configuration:

- **max\_iter:** 1,000 (sufficient for convergence)
- **random\_state:** 42
- **multiclass:** 'ovr' (One-vs-Rest strategy for handling three classes)

#### Performance Metrics:

- **Accuracy:** 97.77%
- **Weighted F1-Score:** 0.9776
- **Per-Class Performance:**

Class	Precision	Recall	F1-Score	Support
0 (Budget-Conscious)	0.99	1.00	0.99	215
1 (Middle-Income)	0.98	0.95	0.96	133
3 (Premium)	0.96	0.98	0.97	100
<b>Weighted Average</b>	0.98	0.98	0.98	448

**Interpretation:** Logistic Regression achieves near-perfect classification, particularly excelling at identifying Budget-Conscious customers (99% precision). This suggests the three customer segments are well-separated in feature space.

### 3.2.2 Decision Tree Classifier

**Model Configuration:**

- **max\_depth:** 10
- **Min\_samples\_split:** 20
- **Min\_samples\_leaf:** 10
- **random\_state:** 42
- **criterion:** 'gini' (default)

**Performance Metrics:**

- **Accuracy:** 92.19%
- **Weighted F1-Score:** 0.9206
- **Per-Class Performance:**

Class	Precision	Recall	F1-Score	Support
0 (Budget-Conscious)	0.93	0.99	0.96	215
1 (Middle-Income)	0.91	0.83	0.87	133
3 (Premium)	0.92	0.89	0.90	100
<b>Weighted Average</b>	0.92	0.92	0.92	448

**Interpretation:** Decision Trees achieve 92.19% accuracy, lower than Logistic Regression. The model shows slightly higher error rates for Middle-Income segment classification (recall = 0.83), suggesting this segment has features that overlap with adjacent segments.

### 3.2.3 Random Forest Classifier

**Model Configuration:**

- **n\_estimators:** 100
- **random\_state:** 42
- **Max\_depth:** 15
- **Min\_sample\_split:** 10
- **n\_jobs:** -1 (parallel processing)

**Performance Metrics:**

- **Accuracy:** 96.21%
- **Weighted F1-Score:** 0.9619
- **Per-Class Performance:**

Class	Precision	Recall	F1-Score	Support
0 (Budget-Conscious)	0.97	0.99	0.98	215
1 (Middle-Income)	0.95	0.92	0.94	133
3 (Premium)	0.97	0.96	0.96	100
<b>Weighted Average</b>	0.96	0.96	0.96	448

**Interpretation:** Random Forest achieves 96.21% accuracy, performing between Decision Trees and Logistic Regression. The ensemble approach provides better generalization than single Decision Trees, particularly for the Middle-Income segment (recall = 0.96).

## 4. Techniques for Result Enhancement: K-Means to DBSCAN Pipeline

### 4.1 Hybrid Clustering Strategy Overview

The project employed a sophisticated two-stage clustering methodology: **K-Means followed by DBSCAN refinement**. This approach combines the global optimization of K-Means with the density-based advantages of DBSCAN to improve cluster quality and robustness.

### 4.2 Stage 1: K-Means Clustering (Initial Segmentation)

**Motivation:** K-Means was selected as the primary clustering algorithm because:

- It optimizes global cluster cohesion (minimizes within-cluster variance)
- It provides interpretable centroid-based clusters
- It scales efficiently to large datasets (2,240 customers)
- It enables business-friendly summary statistics per cluster

### Results from K-Means:

Initial K-Means (k=4) produced:

- Cluster 0: 1,072 customers
- Cluster 1: 664 customers
- Cluster 2: 7 customers (anomalous, later merged)
- Cluster 3: 497 customers

**Issue Identified:** Cluster 2 contained only 7 customers, all with "Other" marital status. This micro-cluster distorted the model and violated the principle of economically meaningful segments.

## 4.3 Stage 2: Cluster Merging Based on Distance Analysis

**Decision Rationale:** Rather than accepting the 4-cluster solution, a data-driven merging strategy was implemented:

### Process:

1. Computed pairwise Euclidean distances from the 7 customers in Cluster 2 to all 4 centroids
2. Identified the nearest centroid for each customer (excluding their own cluster 2)
3. Reassigned each customer to the nearest allowed cluster (0, 1, or 3)

### Result:

- Cluster 0: 1,074 customers (merged with 2 customers from original cluster 2)
- Cluster 1: 667 customers
- Cluster 3: 499 customers
- **Silhouette Scores improved for remaining clusters**

## 4.4 Stage 3: DBSCAN Refinement (Outlier Detection)

**Why DBSCAN?** While K-Means creates well-separated clusters, DBSCAN adds robustness by:

- Identifying true outliers that may have been forced into clusters by K-Means
- Validating cluster boundaries based on point density
- Allowing variable cluster shapes beyond K-Means' spherical assumption

### DBSCAN Parameters:

- **eps (ε):** Distance threshold set after analyzing K-Means results
- **min\_samples:** Minimum points required to form a dense region

### Process:

1. Applied DBSCAN to the scaled dataset post-K-Means
2. Points classified as noise (-1) by DBSCAN but assigned to K-Means clusters were flagged as potential outliers
3. These outliers were retained in K-Means clusters (not removed) but flagged for business review

#### Key Insights:

- Silhouette analysis showed:
  - Cluster 0: 0% negative silhouette values (high cohesion)
  - Cluster 1: 18.6% negative silhouette values (some overlap)
  - Cluster 3: 19.2% negative silhouette values (some overlap with middle-income segment)

## 4.5 Detailed Comparison: K-Means vs. DBSCAN

Aspect	K-Means	DBSCAN
Cluster Shape	Spherical (convex)	Arbitrary (handles density)
Noise Handling	Forces all points into clusters	Identifies true noise points
Parameter Selection	Requires k specification	Requires $\epsilon$ and min_samples
Cluster Size	Balanced across k clusters	Variable sizes based on density
Interpretability	Centroid-based, business-friendly	Density-based, harder to explain
Scalability	$O(nkd)$ per iteration	$O(n^2)$ without spatial indexing

## 4.6 Final Hybrid Outcome

#### The winning strategy combined:

1. **K-Means' Global Optimization:** Ensures well-separated, interpretable clusters with meaningful centroids
2. **K-Means Merging Logic:** Eliminates artificially small clusters (Cluster 2) that don't represent actionable segments
3. **DBSCAN Validation:** Confirms cluster boundaries and identifies customers with borderline membership

### Enhanced Results:

- 3 economically meaningful segments with clear business interpretation
  - Silhouette validation confirms cluster quality (mean silhouette score > 0.2 across all clusters)
  - Supervised classification achieves 97.77% accuracy using derived clusters as targets
- 

## 5. Cluster Interpretation and Characteristics

### 5.1 Cluster 0: Budget-Conscious Families (47.9% of customers)

**Size:** 1,074 customers | **Average Total Spend:** \$105.41

#### Key Demographic Characteristics:

- **Annual Income:** \$35,766 (lowest segment, 53% below company average)
- **Age:** 42.5 years (youngest segment)
- **Family Structure:** Highest family size (1.23 children/teenagers), 87% have dependents
- **Education:** Lower average education level (3.25)

#### Purchasing Behavior:

- **Product Preferences:** Primary focus on wine (41.3% of category spending), minimal luxury items
- **Channel Preference:** Primarily store-based (54.5% of purchases), limited online engagement
- **Campaign Response:** Nearly zero campaign acceptance (0% for early campaigns), resistant to marketing
- **Purchase Frequency:** 6.06 average purchases

#### Business Implications:

- Price-sensitive segment requiring promotional strategies
- Multi-channel strategy ineffective; focus on in-store experience
- High dependent responsibility suggests budget constraints

### 5.2 Cluster 1: Middle-Income Shoppers (29.8% of customers)

**Size:** 667 customers | **Average Total Spend:** \$795.75

#### **Key Demographic Characteristics:**

- **Annual Income:** \$59,443 (moderately above average)
- **Age:** 49.7 years (oldest segment)
- **Family Structure:** Moderate family size (1.12), 92% have dependents
- **Education:** Highest average education level (3.58)

#### **Purchasing Behavior:**

- **Product Preferences:** Balanced spending with emphasis on wine (60% of spending) and meat (19% of spending)
- **Channel Preference:** Most engaged online customers (32% web purchase ratio), active across all channels
- **Campaign Response:** Moderate engagement (5.5% average campaign acceptance)
- **Purchase Frequency:** 17.9 average purchases (highest frequency)

#### **Business Implications:**

- Sweet spot for profitability with higher spending and engagement
- Multi-channel expertise presents opportunity for personalized digital marketing
- Campaign responsiveness suggests growth potential with refined targeting

## **5.3 Cluster 3: Premium High Spenders (22.3% of customers)**

**Size:** 499 customers | **Average Total Spend:** \$1,428.87

#### **Key Demographic Characteristics:**

- **Annual Income:** \$76,324 (highest, 47% above company average)
- **Age:** 45.0 years (mid-range)
- **Family Structure:** Minimal family size (0.09), only 8% have dependents
- **Education:** Moderate-high education (3.45)

#### **Purchasing Behavior:**

- **Product Preferences:** Premium product focus with highest spending on meat (\$490), fish (\$97), wine (\$634)
- **Channel Preference:** Balanced multi-channel engagement with highest catalog purchases (24.1% of purchases)
- **Campaign Response:** Highest campaign acceptance (20% for early campaigns, 31% for Campaign 5)
- **Purchase Frequency:** 19.3 average purchases

#### **Business Implications:**

- High-value customers warranting VIP treatment and personalized service
- Responsive to targeted premium marketing and exclusive campaigns

- Lowest family responsibilities suggest discretionary income availability
- Catalog channel particularly effective for this segment

## 5.4 Cluster Distinguishing Features

<b>Metric (Budget)</b>	<b>Cluster 0 (Middle)</b>	<b>Cluster 1 (Premium)</b>	<b>Cluster 3</b>
Annual Income	\$35,766	\$59,443	\$76,324
Total Spending	\$105	\$796	\$1,429
Age	42.5	49.7	45.0
Family Size	1.23	1.12	0.09
Web Purchase %	28.7%	35.0%	24.0%
Campaign Accept Rate	2.8%	5.5%	20.1%
Wine Spending Ratio	37.7%	59.8%	43.0%