# Customer Marketing Data Analysis for FMCG Retail Company

Data-Driven Marketing Strategy Development Using Python and Machine Learning

## 1. Introduction

Customer data analytics refers to the process of collecting, analyzing, and interpreting data related to customer behavior and interactions with a business. In the fast-moving consumer goods (FMCG) sector, understanding customer preferences and purchasing patterns is crucial for developing effective marketing strategies. According to Kotler and Keller (2016), customer-centric approaches help businesses modify products according to specific needs, behaviors, and concerns of different customer types.

This report analyzes customer data from an FMCG retail company using Python programming. The dataset contains 2,240 customer records with information about demographics, purchasing behavior across product categories, channel preferences, and responses to marketing campaigns. The objective is to leverage this data to identify key insights and trends that enable development of targeted marketing strategies for optimizing customer engagement, retention, and conversion rates.

The analysis employs descriptive statistics using pandas library, data visualization through matplotlib and seaborn, and machine learning algorithms from scikit-learn. The marketing framework follows the Segmentation-Targeting-Positioning (STP) model, which Wedel and Kamakura (2000) describe as fundamental to modern marketing practice.

## 2. Data Preprocessing and Feature Engineering

### 2.1 Data Cleaning

The raw dataset required several preprocessing steps before analysis. First, I examined missing values and found 24 records with missing income data. Following the recommendation of Hair et al. (2019) for handling missing data, I used median imputation since income distributions are typically skewed. One extreme outlier with income value of $666,666 was identified and removed as it clearly represented a data entry error.

Age was calculated from the Year_Birth column using 2024 as reference year. Records showing unrealistic ages (below 18 or above 100 years) were filtered out. After cleaning, the final dataset contained 2,236 valid customer records.

### 2.2 Feature Engineering

To better capture customer behavior patterns, I created several derived features. Total_Spending was calculated as the sum of spending across all five product categories (wines, fruits, meat, fish, and sweets). Total_Purchases combined purchases from all channels (web, catalog, and store). Deal_Sensitivity measured the proportion of purchases made using discounts, indicating price consciousness. Web_Engagement_Ratio captured the proportion of web purchases to total purchases. These engineered features proved valuable for both segmentation and prediction tasks.

## 3. Exploratory Data Analysis

### 3.1 Customer Demographics

The analysis revealed a mature customer base with average age of 55 years and median household income of $51,382. Approximately 71.5% of customers have children (either young kids or teenagers) living at home, suggesting family-oriented purchasing decisions. Education levels are relatively high, with 50% holding graduation degrees and 22% possessing advanced degrees (Master's or PhD). This demographic profile suggests that marketing messages should be sophisticated and value-focused rather than simplistic.

### 3.2 Product Category Analysis

Spending analysis across product categories revealed a clear hierarchy. Wines dominate at 54.1% of total revenue ($680,029), followed by meat products at 29.7% ($373,375). Fresh categories including fruits, fish, and sweets collectively account for only 16.2% of total spending. This concentration in premium categories (wines and meat representing 84% of revenue) indicates the business essentially operates as a specialty retailer rather than a general FMCG store.

| Product Category | Total Revenue | Percentage | Avg per Customer |
|---|---|---|---|
| Wines | $680,029 | 54.1% | $304 |
| Meat Products | $373,375 | 29.7% | $167 |
| Fish Products | $83,931 | 6.7% | $38 |
| Sweet Products | $60,552 | 4.8% | $27 |
| Fruits | $58,753 | 4.7% | $26 |

### 3.3 Purchase Channel Distribution

Store purchases remain the dominant channel with 46% of total transactions (12,959 purchases), demonstrating the continued importance of physical retail despite digital transformation trends. Web purchases account for 33% (9,140 purchases), while catalog orders represent 21% (5,955 purchases). Interestingly, customers visit the company website an average of 5.3 times per month but complete only about 4 web purchases, suggesting potential friction in the online conversion process or comparison shopping behavior.

### 3.4 Campaign Response Analysis

Marketing campaign effectiveness showed progressive improvement over time. Campaign 1 achieved 6.4% response rate, Campaign 2 dropped to only 1.3%, Campaign 3 recovered to 7.3%, and the most recent campaign reached 14.9%. This pattern suggests the marketing team learned from early failures and refined targeting strategies. The poor Campaign 2 results likely triggered a strategy review that benefited subsequent efforts.

# 4. Customer Segmentation Using K-Means Clustering

## 4.1 Algorithm Selection and Methodology

For customer segmentation, I selected K-Means clustering algorithm. As described by Jain (2010), K-Means is an iterative algorithm that partitions data into K distinct clusters by minimizing within-cluster variance. The algorithm assigns each customer to the nearest cluster center (centroid), then updates centroids as the mean of assigned points, repeating until convergence. This method is computationally efficient and works well for customer segmentation tasks.

Before clustering, all numerical features were standardized using StandardScaler to ensure equal contribution from each variable. Without standardization, high-magnitude variables like income would dominate the distance calculations. The clustering was performed on seven variables: Recency (days since last purchase), Total_Purchases, Total_Spending, Income, Age, Deal_Sensitivity, and Web_Engagement_Ratio.

## 4.2 Determining Optimal Number of Clusters

Two methods were used to determine the optimal cluster count. The elbow method plots within-cluster sum of squares (WCSS) against number of clusters, looking for the point where improvement diminishes. The silhouette score, developed by Rousseeuw (1987), measures how similar customers are to their own cluster compared to other clusters, with values closer to 1 indicating better-defined clusters. Both methods indicated K=4 as optimal, yielding a silhouette score of 0.185.

## 4.3 Segment Profiles

| Segment Name | Size | Avg Income | Avg Spending | Response Rate | Key Characteristics |
|---|---|---|---|---|---|
| High-Value | 20.8% | $55,809 | $579 | 15.3% | Older loyal customers |
| Premium Shoppers | 31.8% | $74,732 | $1,242 | 22.9% | Affluent, highly engaged |
| Budget Conscious | 23.7% | $33,486 | $85 | 14.9% | Price-sensitive families |
| At-Risk | 23.7% | $36,476 | $112 | 4.0% | Disengaged customers |

The Premium Shoppers segment, comprising about one-third of customers, generates the highest revenue with average spending of $1,242 and shows highest campaign responsiveness at 22.9%. High-Value customers are older loyal customers who spend moderately but consistently. Budget Conscious customers have lower incomes and show high deal sensitivity. The At-Risk segment represents nearly a quarter of the customer base but shows minimal engagement with only 4% campaign response rate.

# 5. Campaign Response Prediction Using Random Forest

## 5.1 Model Selection

To predict which customers are likely to respond to marketing campaigns, I implemented a Random Forest classifier. According to Breiman (2001), Random Forest is an ensemble method that builds multiple decision trees using bootstrap samples and combines their predictions through majority voting. Each tree considers a random subset of features at each split, which reduces overfitting and improves generalization. This algorithm handles mixed data types well and provides feature importance rankings.

### 5.2 Model Implementation and Results

The model was trained to predict the 'Response' variable (whether customer accepted the last campaign). Features included Income, Age, Total_Spending, Total_Purchases, Recency, Deal_Sensitivity, Total_Children, and spending on wines and meat products. The data was split 75-25 for training and testing, with stratified sampling to maintain class balance. The Random Forest was configured with 100 trees and maximum depth of 8 to prevent overfitting.

Model performance was evaluated using 5-fold cross-validation, achieving an AUC (Area Under ROC Curve) score of 0.777, indicating reasonable predictive ability. The model achieved 87% overall accuracy, correctly identifying customers likely to respond to campaigns and enabling more efficient marketing resource allocation.

### 5.3 Feature Importance Analysis

| Rank | Feature | Importance | Marketing Implication |
|---|---|---|---|
| 1 | Recency | 20.2% | Recent purchasers respond better |
| 2 | Total Spending | 16.8% | High spenders are more engaged |
| 3 | Income | 15.7% | Affluent customers more responsive |
| 4 | Wine Purchases | 12.8% | Wine buyers show high engagement |
| 5 | Meat Purchases | 11.7% | Meat buyers respond well |

Recency emerged as the strongest predictor at 20.2%, aligning with the RFM (Recency, Frequency, Monetary) framework commonly used in direct marketing (Hughes, 1994). Customers who purchased recently are significantly more likely to respond to new campaigns. Total spending and income follow as important predictors, suggesting that engaged, affluent customers form the most responsive target audience.

## 6. Marketing Strategy Recommendations

### 6.1 Premium Shoppers Strategy (31.8% of customers)

For the Premium Shoppers segment with highest spending ($1,242) and response rate (22.9%), I recommend implementing an exclusive loyalty program with premium benefits. Given their strong wine and meat purchasing patterns, offer curated pairing bundles and early access to new products. Private tasting events and sommelier consultations would appeal to this affluent segment. Their high campaign response rate justifies increased marketing investment. However, messaging should emphasize exclusivity and quality rather than discounts to avoid brand devaluation.

### 6.2 High-Value Customers Strategy (20.8% of customers)

High-Value customers are older loyal customers (average age 61) who value quality and personal service. Catalog campaigns showcasing product provenance and craftsmanship will resonate with this segment. Consider subscription boxes featuring carefully selected premium products. In-store experiences like cooking demonstrations and wine education sessions can deepen engagement. Avoid aggressive discounting which may cheapen brand perception for these quality-conscious customers.

### 6.3 Budget Conscious Customers Strategy (23.7% of customers)

The Budget Conscious segment shows high deal sensitivity, indicating price-focused purchasing behavior. However, deep discounts erode margins and train customers to wait for sales. Instead, emphasize value through family-sized packages, bulk purchase savings, and loyalty points accumulating toward meaningful rewards. Flash sales and mobile deal alerts can drive urgency without permanent price reductions. Frame offers around 'smart shopping' rather than cheap buying to maintain customer dignity.

### 6.4 At-Risk Customers Strategy (23.7% of customers)

The At-Risk segment presents the greatest challenge with only 4% campaign response rate. Before abandoning these customers, attempt reactivation through personalized 'we miss you' campaigns offering one-time comeback incentives based on their past purchase history. Survey a sample to understand disengagement reasons—the insights could reveal fixable service issues. However, as Reinartz and Kumar (2002) demonstrated, not all customer relationships are worth saving. If customers don't respond after three win-back attempts, reduce marketing investment and reallocate budget to more responsive segments.

## 7. Channel Optimization Strategy

The analysis reveals distinct channel preferences across customer segments. Store purchases remain dominant at 46%, particularly among older and premium segments. Maintain strong in-store presence with trained staff capable of making personalized recommendations. Consider hosting regular tasting events and product demonstrations.

For the web channel (33%), the gap between monthly visits (5.3) and actual purchases suggests conversion optimization opportunities. Streamline the checkout process, implement abandoned cart recovery emails, and add customer reviews to build confidence. Personalized product recommendations based on browsing history could improve conversion rates.

Catalog marketing (21%) should target the 45+ demographic who appreciate physical browsing. Include QR codes linking to online ordering for customers who prefer browsing catalogs but ordering digitally. Track

catalog-driven web visits to measure cross-channel effectiveness.

## 8. Conclusion

This analysis demonstrates how Python-based data analytics can transform customer data into actionable marketing strategies. Using K-Means clustering, four distinct customer segments were identified, each requiring different marketing approaches. The Random Forest model achieved 77.7% AUC accuracy in predicting campaign response, enabling more efficient targeting and resource allocation.

Key recommendations include: developing premium experiences for high-value customers who drive most revenue, respecting traditional preferences of older premium shoppers, offering value propositions (not just discounts) to budget-conscious families, and taking decisive action on at-risk customers—either winning them back or reallocating resources to more responsive segments.

For implementation, I suggest starting with the premium shoppers segment given their immediate revenue potential and campaign responsiveness. Monitor key metrics including response rates, customer lifetime value, and segment migration. Retrain the predictive model quarterly as new campaign data becomes available to maintain accuracy. With disciplined execution, these data-driven strategies should meaningfully improve customer engagement and overall marketing return on investment.

## References


Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Hughes, A. M. (1994). *Strategic database marketing*. Probus Publishing.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

Kotler, P., & Keller, K. L. (2016). *Marketing management* (15th ed.). Pearson.

Reinartz, W. J., & Kumar, V. (2002). The mismanagement of customer loyalty. *Harvard Business Review, 80*(7), 86-94.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Kluwer Academic Publishers.