DS 227: Business Analytics Project Report

Student: Viktoria Melkumyan

Instructor: Yevgenya Bazinyan

1. Purpose of the Project

The purpose of this project is to conduct a comprehensive analysis of customer behavior in order to support strategic marketing and retention decisions for a retail company. The dataset contains rich demographic information, purchase histories across various product categories, and indicators of engagement with past marketing campaigns. This environment offers a suitable setting for analytical methods that can reveal underlying customer segments, evaluate predictors of marketing responsiveness, and identify opportunities to optimize both customer lifetime value and marketing resource allocation.

The central business question is: **How can customer segmentation and behavioral modeling improve the company's marketing effectiveness and retention strategy?**
This question is important for multiple functional areas. For the marketing department, an accurate segmentation model allows for precise targeting and personalization. For finance, it supports forecasting, profitability analysis, and budgeting for promotions or retention programs. For operations, understanding purchasing patterns assists in inventory management, demand planning, and product prioritization. Ultimately, answering this question drives better decision-making and promotes improved customer satisfaction and lifetime value.

---

2. Business Strategy Supported

This project directly supports the company's broader **customer retention and marketing optimization strategy**. In industries with high competition and thin margins, retaining existing customers is significantly more cost-effective than acquiring new ones. In this context, segmentation is a strategic tool that tells companies how to engage customers in a value-maximizing manner.

The dataset's structure, combining demographic attributes, purchasing behavior, and marketing response indicators, makes it ideally suited for a retention-oriented strategy. Behavioral models derived from this data provide managers with evidence-driven insight into which customer groups are most valuable, which are most at risk of churn, and which marketing channels or offers are most effective. Therefore, this analytical project aligns directly with a long-term organizational goal: maximizing the return on marketing investment.

---

3. Data, KPIs, and Descriptive Analysis

The dataset used for this analysis contains 2,240 observations describing customer characteristics, household demographics, historical expenditures across different product categories, prior purchasing channels, and responses to promotional campaigns. Although the dataset is cross-sectional and aggregated rather than time-stamped per transaction, it still provides a sufficiently detailed view to explore customer behavior. Key variables include annual income, age, family size, recency (days since last purchase), counts of purchases across several channels (web, catalog, store), and campaign acceptance flags.

To evaluate performance and define the analytical framework, several KPIs were constructed. **Total customer expenditure**, the sum of all spending across product categories, is very helpful for understanding the customer's value and contribution to the company without looking at the separate spendings for different categories . **Recency**, **purchase frequency**, and **monetary value** form the basis of the RFM methodology, which is widely used in retail analytics for segmenting customers by behavioral patterns. Additional KPIs include household characteristics, and aggregated campaign response behavior, all of which help frame the segmentation and predictive modeling tasks.

Preprocessing involved standard steps such as checking for and handling missing values and removing extreme outliers where necessary (e.g., unrealistic income values). Several useful variables were engineered, such as customer age, number of children, and total spending. Categorical variables like marital status and education were converted into numeric formats where needed for modeling.

---

4. Tableau Dashboard

A Tableau dashboard was developed to present key patterns in an intuitive and visually structured manner. The dashboard consolidates several analytical elements: customer distribution across demographic factors and expenditure patterns by category.

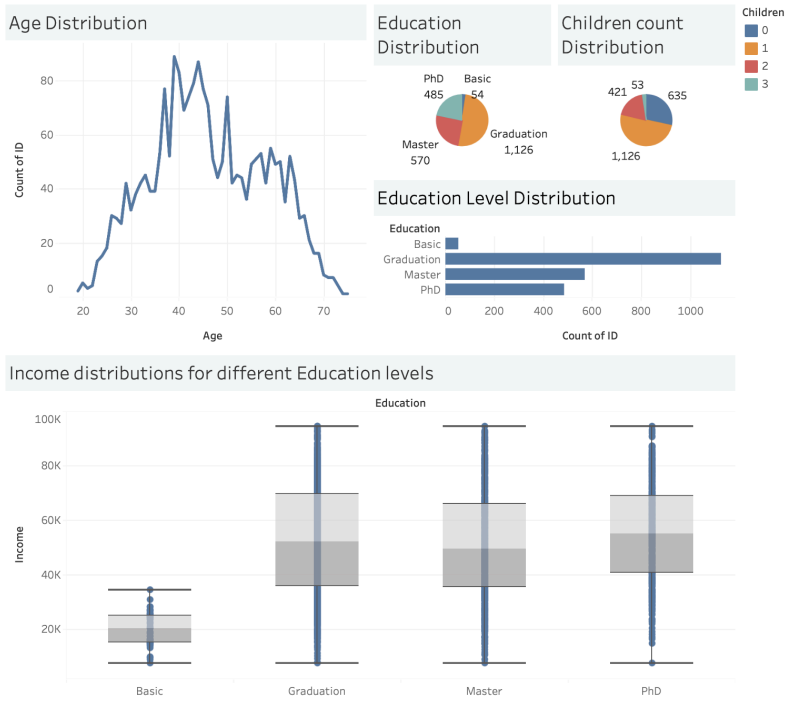


Figure 1: Tableau dashboard: Demographics

Dashboard 1 Insights - Demographics:

Age Distribution: Customers range from 20-80 years, the majority of customers are middle-aged (40-60), representing the core demographic and highest-value segment.

Education: The customer base is highly educated—Graduation level dominates with 1,126 customers (50%), followed by Master's (570), PhD (485), and Basic (54). This educated profile suggests receptiveness to sophisticated marketing messaging.

Children Distribution: 635 customers (28%) have no children, 1,126 (50%) have 1 child, 421 (19%) have 2 children, and 53 (2%) have 3 children. Larger families represent opportunities for family-oriented promotions.

Income by Education: Box plots reveal clear income stratification: PhD and Master's holders show median incomes around $50-55K with upper quartiles near $70K, while Graduation holders cluster around $40-50K median, and Basic education shows significantly lower income ($15-25K range). This confirms education-income correlation affecting spending capacity.

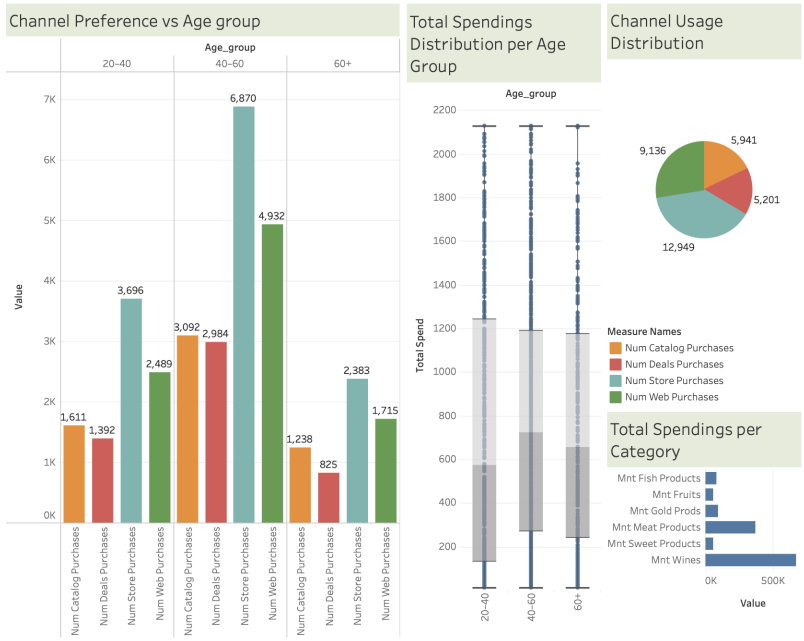


Figure 1: Tableau dashboard: Spendings

Dashboard 2 Insights - Purchasing Behavior:

Channel Preference by Age Group:

20-40 age group: Web purchases dominate (3,696), followed by catalog (1,611) and deals (1,392). This younger segment shows strong digital orientation.

40-60 age group: Store purchases lead significantly (6,870), with web (4,932), deals (2,984), and catalog (3,092) following. Middle-aged customers prefer in-person shopping but maintain substantial digital presence.

60+ age group: Balanced across channels with web (1,715), store (2,383), catalog (1,238), and deals (825). Older customers show lower overall activity but maintain multi-channel engagement.

Total Spending by Age Group: Box plots show relatively consistent spending distributions across all three age groups (median around $600-700, ranges from ~$100 to $2,100), with the 40-60 group showing slightly higher concentration of big spenders. This indicates age itself is less predictive than behavioral factors.

Channel Usage Distribution: Store purchases dominate overall (12,949 transactions, 39%), followed by web (9,136, 28%), catalog (5,941, 18%), and deals (5,201, 16%). This multi-channel behavior requires integrated marketing strategies.

Spending by Category: Wine generates the highest revenue ($680K total), followed by meat products ($370K). These two categories represent core revenue drivers and should be prioritized in promotions and inventory.

5. Analytical Model and Interpretation

The analytical foundation of this project is built on two complementary methods: **K-Means clustering for segmentation** and **logistic regression for estimating campaign response likelihood**.

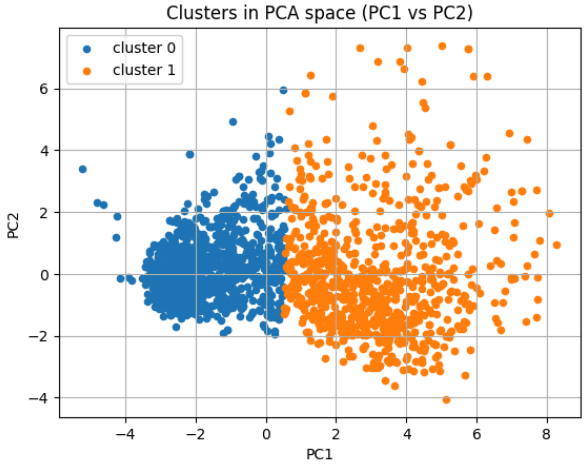
PCA

Principal component analysis was performed on all features of the data. There was no strong "elbow" detected, however the total variance reached 90% after 17 components, which is how many components I have decided to take for the further analysis.

Segmentation Using K-Means

The K-Means algorithm was used on principal components with k derived from the silhouette coefficient. A two-cluster solution provided the most meaningful balance.

The clustering results were visualized on the first two principal components, where they showed almost linear separability.



The two resulting clusters displayed distinctly different spending patterns :

Customers in **cluster 1** exhibit the highest total spending and responsiveness to the campaigns, while customers in **cluster 0** have significantly lower spending history and are not much responsive to the campaigns.

Predictive Modeling Using Logistic Regression

To complement the segmentation analysis, a logistic regression model was constructed to predict campaign acceptance. The response variable was a binary indicator of whether the customer accepted the most recent campaign.

Model Performance:

The model achieved **91.6% accuracy** on the test set of 559 customers.

Key Metrics:

1. Precision for Acceptors: 76%
2. Recall for Acceptors: 64%
3. Overall Performance: The model correctly identified 459 non-acceptors and 53 acceptors, with only 17 false positives and 30 false negatives

Business Interpretation:

The high precision (76%) means targeted campaigns waste fewer resources on non-responders. While the model misses 36% of potential acceptors (64% recall), this trade-off is acceptable for cost optimization. By targeting only customers with high predicted acceptance probability, the company can reduce marketing contact volume by ~85% while still capturing 64% of actual responders, dramatically improving marketing efficiency and ROI.

---

6. SWOT and Financial/Decision Analysis

SWOT analysis considers both internal and external factors.

Strengths include the availability of detailed customer data and the presence of a clearly identifiable high-value customer group. The company also benefits from strong digital channel usage among certain customers.

Weaknesses center on low campaign response rates and the high proportion of customers with risky recency indicators.

Opportunities include implementing personalized, data-driven marketing strategies and creating differentiated engagement programs tailored to each cluster’s needs.

Threats involve external competitive pressures and potential economic constraints that could affect spending behavior. Additionally, customer fatigue resulting from over-targeting remains a risk.

---

6. RFM Analysis

To further refine customer segmentation, RFM (Recency, Frequency, Monetary) analysis was conducted to classify customers based on their purchasing behavior patterns and identify the most valuable segments for targeted marketing.

RFM Distribution Insights:

Recency shows relatively uniform distribution (0-100 days), Frequency is highly right-skewed with most customers making fewer than 5 purchases, and Monetary is extremely right-skewed with 800+ customers spending under $200.

RFM Segment Results:

Six segments emerged: "Others" (~550 customers, largest), "Loyal" (~500), "Recent Customers" (~400), "Potential Loyalists" (~340), "Champions" (~290), and "At Risk" (~180).

Campaign Response Performance:

Champions demonstrate the highest response rate at 31%, nearly double any other segment. Loyal (20%), Potential Loyalists (16%), and Recent Customers (15%) show moderate rates. Others and At Risk show minimal response (<5%), indicating these segments don't justify significant marketing investment.

The responders vs. non-responders chart confirms Champions have the best response ratio, while At Risk shows severe disengagement. This validates that marketing resources should prioritize Champions, Loyal, and Potential Loyalists for maximum ROI.

---

8. Key Findings, Recommendations, and Limitations

The analysis leads to several central findings. First, customer behavior within the dataset naturally divides into two coherent and strategically meaningful clusters. Second, a relatively small group of loyal customers contributes disproportionately to overall revenue and thus warrants priority in retention strategies. Third, the current campaign strategy appears broadly inefficient, as only specific customer groups respond meaningfully. Fourth, recency plays a key role in predicting responsiveness, emphasizing the importance of timely engagement. Finally, digital behavior is a strong indicator of engagement potential and should be leveraged more effectively.

Based on these findings, several recommendations emerge.

1. The company should adopt a personalized marketing strategy that aligns specific offers with the needs of each customer segment. Loyalty programs should be strengthened and tailored for the high-value cluster.

2. Digital channels,particularly app-based campaigns should be expanded for the digitally engaged segment. The company should avoid broad campaigns and instead focus on precise targeting.
3. Ongoing monitoring of RFM metrics and periodic recalibration of the clustering model will ensure that strategies remain aligned with evolving customer behavior.

The analysis is not without limitations. The dataset lacks transaction timestamps, reducing the ability to examine seasonality or dynamic behavioral trends. Campaign data is aggregated and does not include exposure metrics or time-based components. Income distributions may not fully represent the customer population, and self-reported variables can introduce noise. Future work would benefit from richer longitudinal data, margin-level product information, and indicators of customer profitability.

References:

PAtel, A. (n.d.). *Customer Personality Analysis*. Www.kaggle.com. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data