Customer Analytics for Grocery Store

ISOM 690M - Marketing Analytics for R

Haley Hoang, Keerthana Katragadda, Elsie Uwera

I. Introduction

In an increasingly competitive retail grocery landscape, data-driven decision-making is becoming an essential part of daily operations. The grocery retail industry is undergoing a fundamental transformation driven by a convergence of changing consumer behaviors, technological innovations, and macroeconomic pressures. As food prices have risen sharply over recent years, many consumers are becoming more price-conscious, while still demanding convenient, quality, and personalized shopping experiences.

At the same time, the way consumers shop is evolving: many expect a seamless blend of online and in-store experiences. Grocery retailers have responded by investing heavily in omnichannel capabilities, integrating e-commerce, loyalty programs, delivery services, and digital marketing. In parallel, the rise of advanced data analytics, machine learning, and real-time data processing is enabling grocers to turn massive volumes of transactional, and supply-chain data into actionable insights. Retailers no longer rely solely on intuition or broad segmentation; instead, they can analyze customer-level purchase histories, channel usage and more.

According to industry trends, the most successful grocery retailers in coming years will be those that shift from mass-market, one-size-fits-all strategies to micro-targeted, personalized marketing and operations. Organizations must leverage their rich customer dataset to build targeted strategies that maximize marketing return on investment, and strengthen customer loyalty.

This report applies analytical techniques including exploratory data analysis (EDA), customer segmentation, and predictive modeling; to provide targeted strategies and recommendations.

II. Project Objectives

The core objectives of this project are to:

- Understand customer behavior through exploratory data analysis
- Segment customers using clustering methods to identify high-potential groups
- Analyze marketing campaign effectiveness, including predictors of campaign response
- Develop data-driven strategies to increase customer engagement, campaign success, and overall profitability

III. Exploratory Data Analysis

Exploratory Data Analysis was conducted to understand the structure of the dataset, identify key customer characteristics, detect patterns in purchasing behavior, and uncover factors associated with campaign response. The dataset consists of 2,240 customers and includes demographic information, spending amounts across six product categories, purchasing channels, website activity, and marketing campaign history.

1) Customer Demographics

Customers span a wide range of age groups, with most falling between their mid-20s and early 70s. Education levels are generally high, with the majority holding a graduate degree, followed by PhD and Master's degrees. Married and partnered individuals represent the largest share of the customer base, and these groups exhibit slightly higher campaign response rates than single or divorced customers. Household income shows significant variability with a right-skewed distribution due to high-income outliers. The median income, approximately $51,000, provides a more representative measure of the typical customer.

Household composition was analyzed using both Kidhome and Teenhome variables, which were combined into a total children count. Most customers have zero to two children, and households with fewer children show higher responsiveness to campaigns.

2) Spending Behavior Across Product Categories

Customer spending is concentrated primarily in Wines and Meat Products, followed by Gold (premium) products. Fruits, Fish, and Sweets represent smaller portions of total spending. A constructed Total Spending variable (sum of all category spending) revealed an average spend of approximately $562 per customer over the two-year period.

A significant distinction emerged between responders and non-responders. Responders consistently spent more on premium categories, especially Wines, Meat Products, and Gold Products. This indicates that high-value customers are substantially more likely to accept marketing offers.

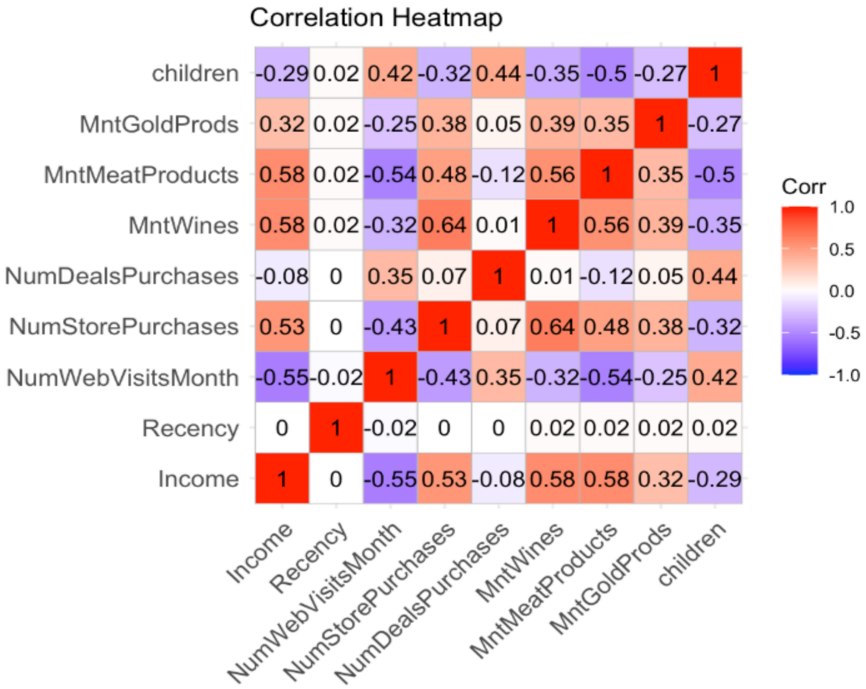
3) Channel Usage Patterns

Store purchases dominate the dataset, with nearly 80% of customers preferring the physical store. Web purchases and catalog purchases account for smaller but strategically important segments. Although catalog shoppers represent only a small fraction of the customer base, they exhibit the highest average total spending and the highest campaign response rate (approximately 34%). Web-preferred customers show the highest website visit frequency but comparatively lower spending, suggesting a tendency toward browsing rather than purchasing. Store-preferred customers represent the largest group but display the lowest average spending and the weakest response rates.

4) Behavioral Predictors of Response

Several behavioral patterns distinguish responders from non-responders. Responders tend to have:

- Lower recency values (they purchased more recently),
- Fewer web visits (indicating they are buyers rather than browsers),
- Higher spending in premium product categories,
- Smaller household sizes.



	Income	Recency	NumWebVisitsMonth	NumStorePurchases	NumDealsPurchases	MntWines	MntMeatProducts	MntGoldProds	children
children	-0.29	0.02	0.42	-0.32	0.44	-0.35	-0.5	-0.27	1
MntGoldProds	0.32	0.02	-0.25	0.38	0.05	0.39	0.35	1	-0.27
MntMeatProducts	0.58	0.02	-0.54	0.48	-0.12	0.56	1	0.35	-0.5
MntWines	0.58	0.02	-0.32	0.64	0.01	1	0.56	0.39	-0.35
NumDealsPurchases	-0.08	0	0.35	0.07	1	0.01	-0.12	0.05	0.44
NumStorePurchases	0.53	0	-0.43	1	0.07	0.64	0.48	0.38	-0.32
NumWebVisitsMonth	-0.55	-0.02	1	-0.43	0.35	-0.32	-0.54	-0.25	0.42
Recency	0	1	-0.02	0	0	0.02	0.02	0.02	0.02
Income	1	0	-0.55	0.53	-0.08	0.58	0.58	0.32	-0.29

Our correlation analysis indicates strong positive relationships between in-store purchases and spending on Wines and Meat Products. Interestingly, income shows only weak direct correlation with spending, suggesting that engagement and product preferences are stronger drivers of purchasing behavior than income alone.

Overall, the EDA reveals that premium consumption behavior, recency of purchase, channel preference, and household size play a far more important role in predicting campaign response than basic demographic variables.

IV. Variable Selection and Data Preparation

To construct a parsimonious and behaviorally meaningful model, we selected five predictors: total_spend, Recency, children, channel_pref, and NumWebVisitsMonth. These variables were chosen based on evidence uncovered in Phase I through exploratory analysis and statistical testing.

Rationale for Variable Selection

- total_spend — Customer Value & Purchase Intensity

Total spending is one of the strongest indicators of customer value and engagement. Phase I analyses showed that responders consistently display higher historical spending. This variable captures cumulative buying behavior in a single, interpretable metric.

- Recency — Recency of Purchase Activity

  Recency is a foundational component of RFM modeling and a powerful predictor of marketing responsiveness. Customers who purchased recently are more engaged with the brand and exhibit a significantly higher response rate.

- children — Household Structure

  Household size influences consumption volume, shopping patterns, and sensitivity to promotions. Phase I findings demonstrated that customers with more children respond differently compared to smaller households, making this a valuable demographic predictor.

- channel_pref — Preferred Shopping Channel

  Channel preference (Store, Web, Catalog) reflects distinct behavioral differences. Catalog customers were found to have the highest response rates, whereas Store-centered customers were less responsive. Including this variable captures meaningful channel-based heterogeneity in responsiveness.

- NumWebVisitsMonth — Digital Engagement & Intent

  Monthly web visits capture browsing behavior, product interest, and early-stage consideration. The strong positive relationship between website visitation and response suggests that digital engagement is a robust behavioral predictor.

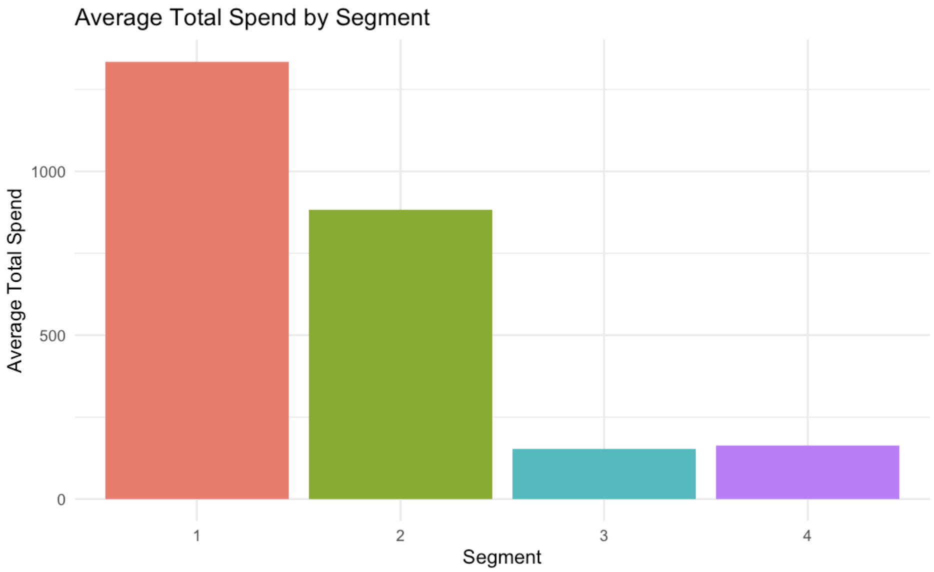
Collectively, these five variables capture customer value, purchasing recency, household composition, shopping preferences, and digital engagement—five dimensions strongly tied to marketing response behavior.

V. Market Segmentation: K-Means clustering

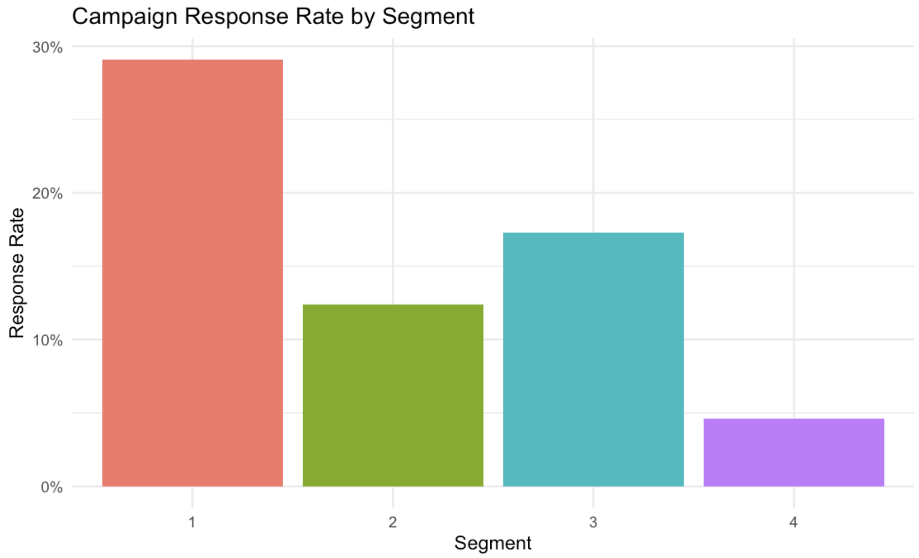
To create actionable customer groups, K-Means clustering was applied to standardized numerical variables including income, recency, total spending, number of children, channel purchase counts, and web visit frequency. Collectively, they represent customer value, recent purchasing activity, household structure, preferred shopping channels, and online engagement—all of which are highly relevant to predicting likelihood of response in grocery marketing campaigns.

Standardization ensured that no single variable dominated the clustering process due to scale differences. The optimal number of clusters was selected using the Elbow Method, which identified a clear inflection point at four clusters.

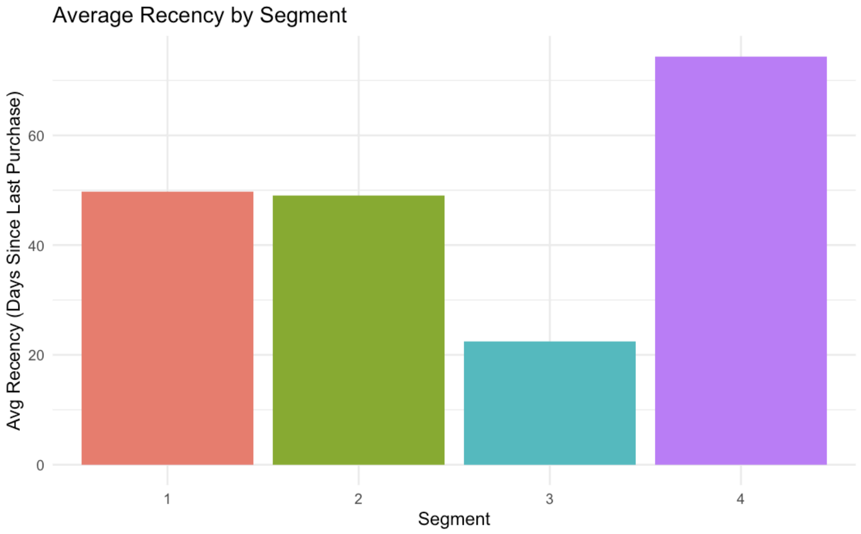
Results:



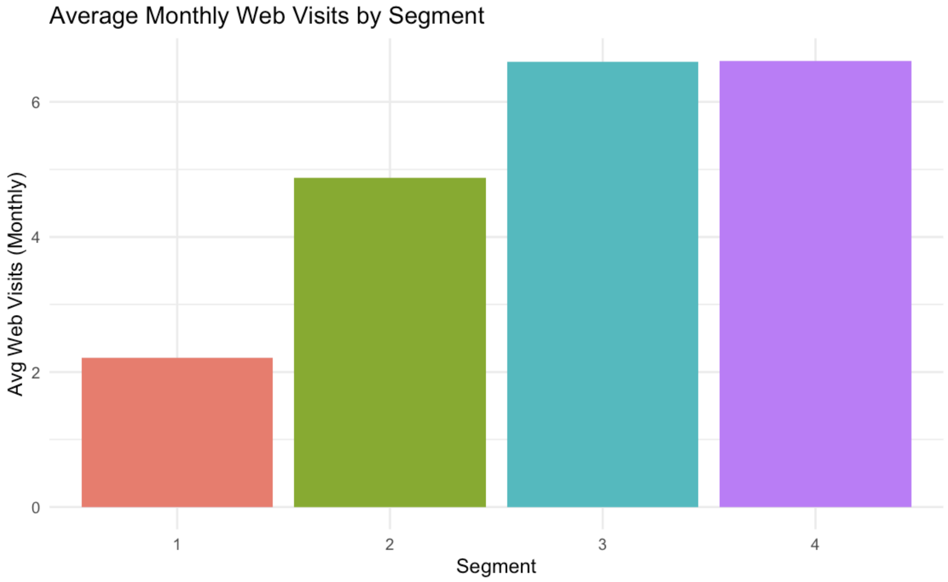
The segmentation results reveal a strong distinction in spending behavior across the four customer groups. Segment 1 stands out as the highest-value segment, with average spending well above the other groups, indicating a cluster of customers who consistently make large purchases and likely contribute disproportionately to overall revenue. Segment 2 represents a mid-value segment, with average spending below Segment 1 but still significantly higher than the remaining segments, suggesting a solid group of shoppers with moderate purchasing power. In contrast, Segments 3 and 4 show substantially lower average spend, each falling well below the mid-tier level. These customers likely shop less frequently or purchase lower-priced items, making them lower-priority targets for high-cost or premium campaigns. Overall, the pattern highlights clear opportunities for differentiated marketing within each segment.



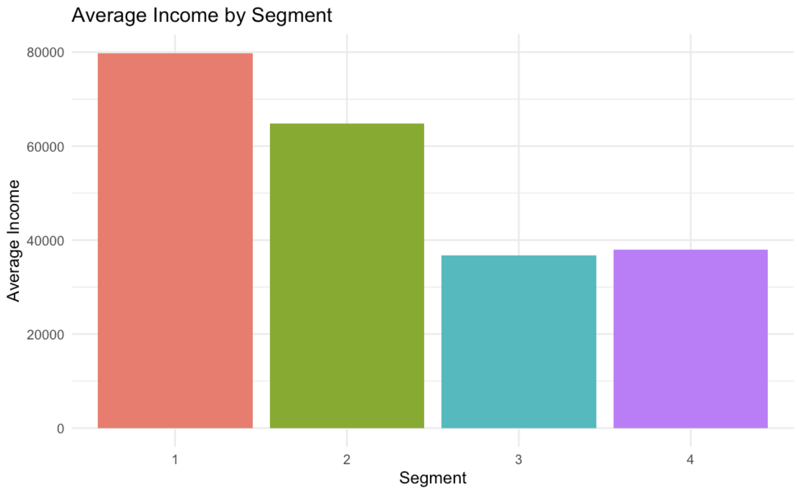
Segment 1 shows the highest response rate—nearly 30%—indicating that these customers are highly receptive to marketing outreach and represent the most promising group for future campaigns. Segment 3 also responds relatively well, with a moderate response rate around 17%, suggesting that these customers are worth targeting, especially with well-timed or personalized offers. In contrast, Segment 2 produces only a modest response rate of roughly 13%, indicating limited but still measurable engagement. Segment 4 has the lowest response rate, falling below 5%, signaling that this segment is the least responsive and may not justify heavy investment in promotional spend.



Segment 3 shows the lowest recency—around 22 days—which means customers in this group shop most frequently and are the most actively engaged with the brand. Segments 1 and 2 follow closely behind with recency values just below 50 days, indicating moderately regular purchasing behavior. In contrast, Segment 4 has the highest recency at nearly 70 days, suggesting that this segment is the least active and may be drifting toward churn. These patterns indicate that Segments 1 and 3 contain more recently engaged customers who are likely to be receptive to ongoing campaigns, while Segment 4 may require reactivation strategies.



The number of monthly web visits varies noticeably across customer segments, highlighting differences in digital engagement. Segments 3 and 4 are the most digitally active, each averaging around 6–7 web visits per month, suggesting that these customers frequently browse or interact with the company's online platform. Segment 2 shows moderate online activity, averaging about 5 monthly visits, indicating a group that is comfortable online but not as highly engaged as Segments 3 and 4. In contrast, Segment 1 has the lowest online engagement, with only about 2 web visits per month, implying a more traditional shopping pattern and a lower likelihood of responding to purely digital marketing efforts..



Segment 1 shows the highest average income indicating an affluent group with strong purchasing power and greater capacity for premium spending. Segment 2 follows with a solid mid-level income, suggesting a financially stable segment that may be responsive to value-driven yet quality-focused promotions. In contrast, Segments 3 and 4 have significantly lower average incomes, positioning them as more budget-conscious groups. These income differences help explain variations in spending and responsiveness observed in other analyses and underscore the importance of tailoring product offerings and promotional strategies to the financial profiles of each segment.

Table 1. Segments Summary and Assigned Labels

Segment	Income	Spend	Recency	Web Visits	Response Rate	Label

1	High	Very High	Moderate	Very Low	30%	Affluent traditional shoppers
2	Mid	Mid	Moderate	Mid	13%	Mixed channel spenders
3	Low	Very Low	High	High	17%	Digital but low spend
4	Low	Low	Very Low	High	5%	Browsers but disengaged

VI. Predictive Modelling: Logistic Regression

1) Coefficient & Odds Ratios Interpretation

The logistic regression model reveals several statistically significant predictors that align with expected marketing behaviors. Table 1 summarizes the key coefficients, odds ratios, and managerial interpretations.

Table 1. Predictor Coefficients, Odds Ratios, and Business Interpretation

Predictor	Coefficient	Odds Ratio	Interpretation
total_spend	+0.001456	1.0015	A $1,000 increase in historical spending raises the odds of response by ~15%. High spenders are especially responsive.

NumWebVisitsMonth	+0.2056	1.228	Each additional web visit increases the odds of responding by 22.8%. Digital browsers are strong prospects.
Recency	–0.0262	0.974	Each additional inactive day reduces response odds by ~2.6%. Recent shoppers are significantly more responsive.
children	–0.480	0.619	Each additional child decreases response odds by ~38%. Households with children are less engaged with direct campaigns.
channel_pref: Store	–0.874	0.418	Store-only shoppers are 58% less likely to respond than Catalog/Web customers.
channel_pref: Web	–0.091	0.913	Slightly less responsive than Catalog customers, but more responsive than Store customers.

These results confirm that **customer value, digital engagement, and recent purchasing behavior** are the strongest drivers of responsiveness, while store-only shoppers and larger households are least likely to respond.

Key Target Segment Identified by the Model

The combined model outputs highlight a clear ideal target - **“High-Value Digital Engagers”:**

- High historical spend

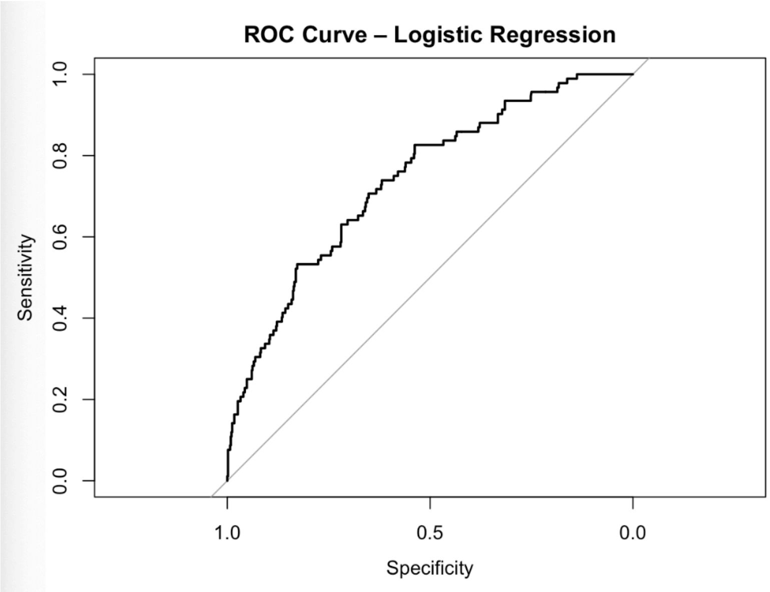
- Recent purchases (last 3–6 months)
- Strong website engagement (6+ visits/month)
- Few or no children at home
- Prefer Web or Catalog channels

This group exhibits the greatest predicted lift in response and provides the highest potential marketing ROI.

2) Model Performance Evaluation

2.1) ROC Curve and AUC

Figure 1. ROC Curve for Logistic Regression Model

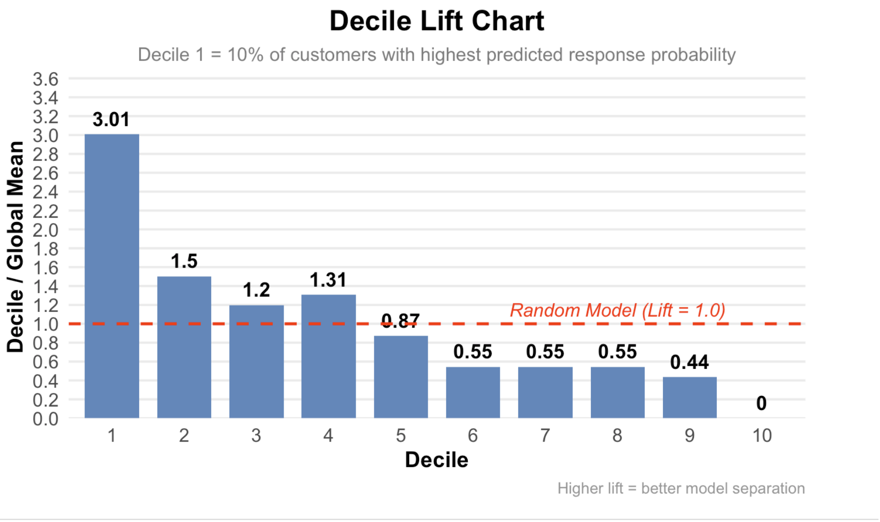


This plot shows that the model achieves an AUC of 0.73, indicating good predictive separation between responders and non-responders. Importantly, AUC is threshold-independent, meaning it reflects performance across all possible probability cutoffs.

2.2) Decile Lift Analysis & Optimal Cutoff Selection

Instead of selecting a cutoff arbitrarily (e.g., 0.20), we determine the optimal threshold through decile (lift) analysis, a method widely used in marketing analytics to maximize campaign profitability.

Figure 2. Decile Lift Chart for Optimal Cutoff Selection

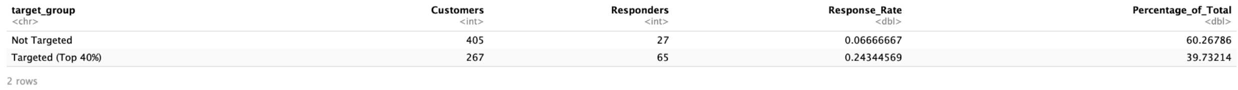


This chart ranks customers into ten deciles based on predicted response probability. The top decile achieves a lift of 3.01× over baseline, confirming strong model discrimination. Deciles 1–4 yield lifts above 1.0, supporting the recommended cutoff around the top 40% of customers.

- Customers were ranked by predicted probability and divided into 10 equal-sized deciles.
- The top 40% of customers (Deciles 1–4) showed the greatest lift in response rate.
- The optimal cutoff corresponding to Deciles 1–4 is 0.1340.

2.3) Campaign Impact of the Optimized Cutoff

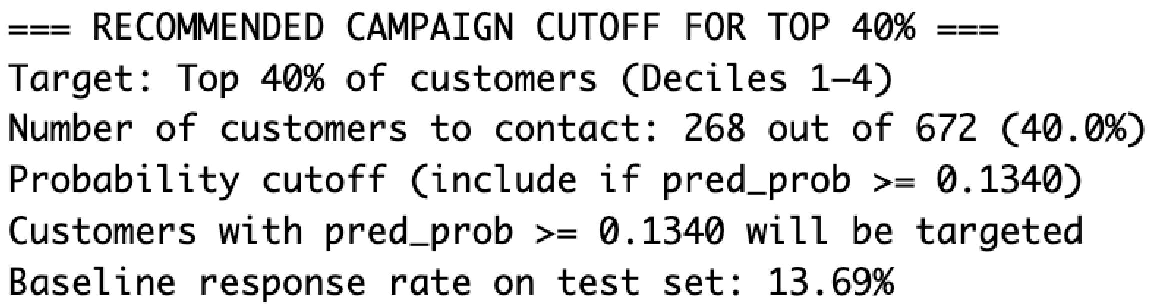
Figure 4. Campaign Targeting Summary Table (Top 40% Cutoff ≥ 0.1340)



target_group <chr>	Customers <int>	Responders <int>	Response_Rate <dbl>	Percentage_of_Total <dbl>
Not Targeted	405	27	0.06666667	60.26786
Targeted (Top 40%)	267	65	0.24344569	39.73214

2 rows

Figure 5. Recommended Campaign Cutoff for Top 40% of Customers



```
=== RECOMMENDED CAMPAIGN CUTOFF FOR TOP 40% ===
Target: Top 40% of customers (Deciles 1-4)
Number of customers to contact: 268 out of 672 (40.0%)
Probability cutoff (include if pred_prob >= 0.1340)
Customers with pred_prob >= 0.1340 will be targeted
Baseline response rate on test set: 13.69%
```

Based on these 2 figures, targeting only the top 40% of customers:

- Achieves a 24.3% response rate (vs. 13.69% baseline → 1.78× lift)
- Captures 70.7% of all responders
- Reduces outreach volume by 60%
- Significantly improves ROI and marketing efficiency

This demonstrates the powerful operational value of using predictive scoring to prioritize high-potential customers.

3) Confusion Matrix and Classification Metrics

Using the optimized cutoff (0.1340), Figure 6 shows the confusion matrix and Table 2 summarizes the model's performance on the test set.

Figure 6. Confusion Matrix for Top 40% Targeting Strategy

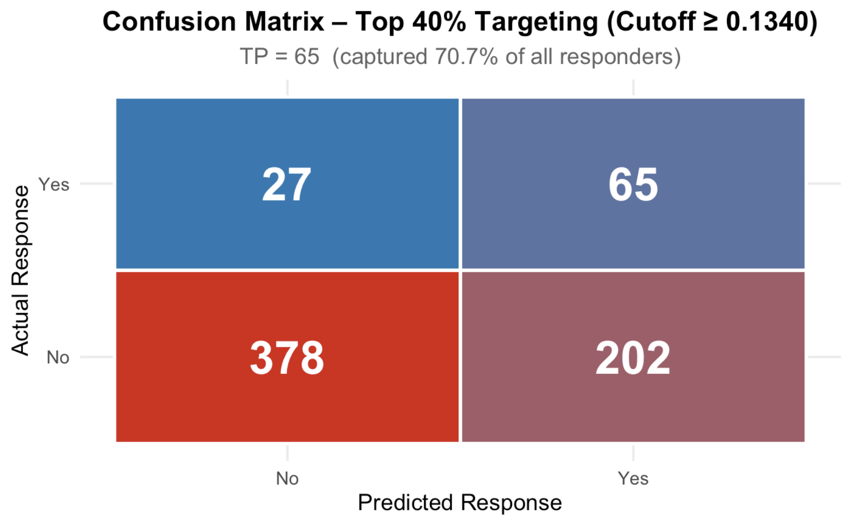


Table 2. Classification Metrics at Optimal Cutoff

Metric	Value	Interpretation
Accuracy	65.9%	Correctly predicts ~2/3 of outcomes; less important in imbalanced data.
Recall (Sensitivity)	70.7%	Captures most true responders — critical for maximizing sales lift.
Specificity	65.2%	Excludes ~65% of non-responders to reduce campaign waste.
Precision	24.3%	Of targeted customers, 24.3% respond; 1.78× better than baseline (13.7%).
F1-Score	36.2%	Balanced measure; acceptable in a setting where recall is prioritized.

Overall, the model balances recall and cost efficiency effectively, ensuring the campaign reaches most potential responders without excessive waste.

VII. Recommendations:

1. Prioritize High-Value, High-Response Customer Segments

Segmentation results identify Segment 1 as the most valuable customer group, with the highest average spending and a campaign response rate near 30%. Segment 3 also demonstrates meaningful engagement, with a response rate of approximately 17% and very recent purchasing behavior. These two segments represent the strongest revenue and return-on-investment potential.

The company should make Segments 1 and 3 the primary focus of future marketing campaigns. Greater investment should be allocated toward loyalty rewards, premium product bundles, and personalized promotions for these customers. Retention and upselling strategies targeted at these high-performing segments will maximize revenue while improving long-term customer lifetime value.

2. Match Marketing Channels to Customer Behavior

Significant differences in digital and in-store behavior were observed across customer segments. Segments 3 and 4 exhibit the highest levels of digital engagement, while Segments 1 and 2 are predominantly store-focused. Because channel preference is strongly associated with response behavior, marketing strategies must be aligned with how customers naturally interact with the retailer.

For digitally active customers, marketing efforts should emphasize personalized emails, mobile notifications, website-exclusive offers, and retargeting advertisements. For store-dominant customers, in-store coupons, printed mailers, and register-triggered promotions should be prioritized. Aligning promotional delivery with customer behavior will significantly improve engagement and response efficiency.

3. Use Predictive Modeling to Prioritize Customers

The logistic regression model developed in this study demonstrates strong predictive performance with an AUC of 0.73. The strongest positive predictors of campaign response include high total spending, recent purchasing activity, and frequent web engagement. Negative predictors include having children in the household, a preference for store-only shopping, and long periods of inactivity.

The retailer should incorporate predictive response scores directly into campaign decision-making. Customers should be ranked by predicted probability of response, with the highest-scoring customers receiving the most aggressive and personalized marketing outreach. Lower-probability customers should be targeted using low-cost or automated communication methods to conserve marketing budget and improve overall campaign efficiency.

4. Reactivate At-Risk Customers

Segment 4 represents the highest churn-risk group. These customers display frequent website visits but very poor recency and extremely low campaign response rates, indicating engagement without conversion. This pattern reflects customers who are browsing but not purchasing and may be close to permanently disengaging from the brand.

The company should deploy structured reactivation campaigns for this segment using time-limited discounts, free delivery incentives, and personalized product reminders. These win-back offers should be automatically triggered when customer inactivity exceeds 45 to 60 days. Early intervention will increase the likelihood of recovery before full churn occurs.

VIII. **Conclusion**:

This project demonstrates how data-driven analytics can significantly improve marketing effectiveness in the grocery retail environment. Through exploratory analysis, customer segmentation, and predictive modeling, clear patterns of customer behavior and campaign response were identified.

The EDA revealed that premium spending, recent purchasing behavior, and channel preference are the strongest drivers of campaign response. Market segmentation uncovered three distinct customer segments with fundamentally different purchasing behaviors and profitability profiles. Predictive modeling further validated these patterns and provided quantifiable evidence of the variables most strongly associated with marketing success. We also believe that the improvements and business recommendations will be advantageous in increasing revenue and customer retention.