**Sai Charitha Pathipati**

**Webster University**

**CSDA 6010 DATA ANALYTICS PRACTICUM**

**EXECUTIVE SUMMARY:**

This report provides a comprehensive analysis of customer behavior and marketing campaign responses for Shop-Smart, with the goal of optimizing targeted marketing strategies. By leveraging data-driven insights, the study identifies key factors that influence customer engagement and purchasing behavior, ultimately enhancing business decision-making and marketing effectiveness

The analysis uncovered significant variations in customer responses to marketing campaigns, driven by demographic characteristics, purchasing frequency, and recency of interactions. Customers within specific income brackets and those with distinct spending patterns exhibited varying levels of engagement, highlighting the need for personalized marketing approaches. Dimensionality reduction techniques were applied to eliminate redundant features, ensuring that the predictive models focused on the most influential factors while improving computational efficiency and interpretability.

Data partitioning methods were employed to structure the dataset for robust predictive modeling, allowing for accurate predictions of customer responses to marketing campaigns. Multiple classification models, including Random Forest and Logistic Regression, were tested. After careful evaluation, the Random Forest model emerged as the best-performing model, demonstrating superior accuracy and predictive power, making it the ideal choice for understanding customer behavior and campaign response predictions.

Additionally, clustering methods were utilized to segment customers based on behavior, demographics, and purchasing patterns. Both Hierarchical Clustering and K-Means Clustering were applied, with K-Means proving to be the more effective method for identifying distinct customer segments. These segments included families with children who spend less, middle-income shoppers, and high-spending loyal customers, each of which requires tailored marketing approaches.

The insights gained from both classification and clustering models offer actionable recommendations for Shop-Smart’s marketing strategy. These findings enable the creation of targeted campaigns that maximize customer engagement and conversion rates. Additionally, the analysis provides strategic guidance for resource allocation, ensuring that marketing efforts are directed toward the most responsive customer segments, ultimately improving the return on investment (ROI).

**TABLE OF CONTENTS**

1. INTRODUCTION ……………………………………………………………………….…6

2. BUSINESS AND ANALYTICAL……………………………………….............................6

2.1. BUSINESS PROBLEM……………………………………………………………6

2.2. BUSINESS GOAL…………………………………………………………………7

2.3. ANALYTICAL GOAL…………………………………………………………….7

2.4. ANALYTICAL APPROACH…………………………………................................7

3. DATA EXPLORATION AND PREPROCESSING………………………………………….8

3.1. ATTRIBUTE DEFINITION……………………………………………………….8

3.2. TYPES OF DATA………………………………………………………………….9

3.3. DETAILED ANALYSIS ON EACH ATTRIBUTE…………….………………….10

4. DATA ENGINEERING AND TRANSFORMATION ………………………………………25

5. PREDICTOR ANALYSIS AND RELEVANCY…………………………………………….26

5.1. PREDICTOR RELAVANCY FOR CLASSIFICATION…………………………….26

5.2. PREDICTOR RELANCY FOR CLUSTERING …………………………………….37

6. DIMENSION REDUCTION………………………………………………………………...39

6.1. DIMENSION REDUCTION FOR CLASSIFICATION…………………………….39

6.2. DIMENSION REDUCTION FOR CLUSTERING………………………………….40

7. DATA PARTITIONING……………………………………………………………………...40

8. MODEL SELECTION………………………………………………………………………41

8.1. MODEL SELECTION FOR CLASSIFICATION……………………………………41

8.2. MODEL SELECGTION FOR CLUSTERING ……………………………………….44

9. MODEL IMPROVEMENT AND EVALUATION…………………………………………...56

10. CONCLUSION………………………………………………………………………………58

**TABLE OF FIGURES**

1. structure of data………………………………………………………………

2. checking the uniqueness…………………………………………………….

3. Distribution of year of birth…………………………………………………..

4. Distribution of customers by Generation…………………………………….11

5. Distribution of Education levels……………………………………………...12

6. Distribution of marital status…………………………………………………13

7. Distribution of income……………………………………………………….14

8. Distribution of kidHome……………………………………………………..14

9. Distribution of teenagers……………………………………………………..15

10. Distribution of monthly registration…………………………………………16

11. Distribution of weekly registration…………………………………………16

12. Distribution of recency………………………………………………………17

13. customer recency…………………………………………………………….18

14. distribution of total spending of a customer…………………………………19

15. Distribution of purchase types……………………………………………….20

16. number of visits………………………………………………………………21

17. Distribution of campaigns……………………………………………………22

18. number of complaints………………………………………………………...23

19. response to the last campaign………………………………………………...24

20. checking zeros………………………………………………………………..24

21. Education vs Target response…………………………………………………27

22. Education vs Target response…………………………………………………28

23. Kidhome vs target response…………………………………………………..29

24. teenhome vs target response…………………………………………………30

25. spending categories vs Target variable……………………………………….31

26. number of deals purchased vs target response……………………………….32

27. number of web purchases vs target response…………………………………33

28. number of catalog purchases vs target response………………………………33

29. number of store purchases vs target response…………………………………34

30. number of web visits in a month vs target response…………………………..35

31. complain vs target response…………………………………………………...35

32. generation vs target response………………………………………………….36

33. recency vs target response…………………………………………………….36

34. recency vs target response…………………………………………………….37

35. data partition…………………………………………………………………..41

36. logistic regression……………………………………………………………..42

37. Random forest…………………………………………………………………43

38. elbow method…………………………………………………………………45

39. size of clusters………………………………………………………………...45

40. silhouette plot…………………………………………………………………46

41. k-means clustering……………………………………………………………48

42. Hierarchical Clustering……………………………………………………….50

43. random forest evaluation……………………………………………………...54

44. feature importance in random forest………………………………………….55

45. model improvement…………………………………………………………..56

**1. Introduction:**

Shop-Smart is a retail company specializing in premium food and beverage products. The company has a large consumer database with detailed information on demographics, purchasing history, online activity, and responses to previous marketing campaigns.

To optimize its marketing strategy, Shop-Smart aims to understand its consumers better. Instead of spending money to market a new product to every consumer in the database, the company wants to analyze which consumer segment is most likely to buy the product and then target only that segment. This targeted approach will help reduce marketing costs, improve campaign effectiveness, and increase revenue.

The main objective of this project is to segment consumers based on their purchasing behavior and analyze which types of consumers respond positively to marketing campaigns. Analytical models will be developed to predict consumer responses, providing actionable insights for more effective and personalized marketing strategies.

**2. Business and analytical goals:**

It ensures that the analysis remains aligned with solving real-world problems. It helps in focusing efforts on extracting insights that drive actionable solutions, rather than just exploring data without purpose.

**2.1. Business problem:**

* Shop-Smart does not have a clear way to group its customers based on their shopping habits, demographics, and preferences. Without proper customer segments, the company struggles to send personalized marketing messages. This leads to low customer engagement, wasted marketing budget, and missed chances to connect with the right customers.
* Shop-Smart cannot accurately predict which customers are likely to respond to its marketing campaigns. Although the company runs several campaigns, it lacks a system to identify which customers will accept offers. As a result, the company spends money on campaigns that don’t reach the right audience, lowering its return on investment (ROI).

**2.2. Business goals:**

* Shop-Smart wants to better understand its customers by identifying different groups based on their shopping habits, preferences, and backgrounds. This understanding will help the company create marketing campaigns that appeal to each group, leading to more customer engagement and smarter use of marketing budgets.
* Shop-Smart aims to predict which customers are most likely to respond positively to its marketing efforts. By focusing on the right customers, the company can increase the success of its campaigns, boost sales, and reduce unnecessary marketing expenses.

**2.3. Analytical goal:**

* Use clustering techniques to analyze customer data and group them based on their shopping behaviors, demographics, and preferences. This segmentation will allow Shop-Smart to identify distinct customer profiles, helping to create targeted marketing strategies that are more effective for each group.
* Develop a classification model to predict which customers are most likely to respond to marketing campaigns. By analyzing past campaign data and customer characteristics, the model will classify customers into responders and non-responders, enabling Shop-Smart to optimize marketing efforts and increase campaign success.

**2.4. Analytical Approach:**

**For Addressing Varying Consumer Needs:**

Approach:

Data Preparation: Clean and preprocess the consumer data by handling missing values, normalizing numerical features and encoding categorical variables

Segmentation: Use clustering algorithms to group consumers based on key factors like purchasing behavior, demographics, and preferences. This will help identify distinct consumer groups.

Interpretation: Analyze the resulting segments to understand consumer behavior patterns and characteristics, which will inform the creation of targeted marketing campaigns for each group.

**For Predicting Campaign Response:**

Approach:

Data Preprocessing: Clean and preprocess campaign response data by addressing missing values, encoding categorical variables, and normalizing numerical features

Classification Modeling: Build classification models to predict whether a consumer will respond positively to a marketing campaign. Use features like past purchase behavior, demographics, and campaign interactions as predictors.

Evaluation: Assess model performance using metrics like accuracy, precision, recall, and the F1 score to evaluate the model's ability to correctly predict campaign responders.

Insights: Identify key factors that influence a consumer's likelihood to respond to campaigns and use this information to refine marketing strategies and increase campaign effectiveness.

**3. Data exploration and preprocessing:**

This step involves examining the dataset to understand its structure, identifying missing values, and performing necessary transformations

**3.1. Attribute definition:**

1. ID: A unique identifier for each consumer in the database.
2. Year\_Birth: The birth year of the consumer, used to calculate age
3. Education: The education level of the consumer.
4. Marital\_Status: The marital status of the consumer.
5. Income: The consumer's yearly household income.
6. Kidhome: The number of children in the consumer's household.
7. Teenhome: The number of teenagers in the consumer's household.
8. Dt\_Consumer: The date when the consumer first enrolled or registered with the company.
9. Recency: The number of days since the consumer's last purchase, indicating recent engagement with the company.
10. MntWines: The amount of money spent by the consumer on wine products.
11. MntFruits: The amount of money spent by the consumer on fruit products.
12. MntMeatProducts: The amount of money spent by the consumer on meat products.
13. MntFishProducts: The amount of money spent by the consumer on fish products.
14. MntSweetProducts: The amount of money spent by the consumer on sweet products.
15. MntGoldProds: The amount of money spent by the consumer on gold products.
16. NumDealsPurchases: The number of purchases made by the consumer related to promotional deals.
17. NumWebPurchases: The number of purchases made by the consumer through the company’s website.
18. NumCatalogPurchases: The number of purchases made by the consumer through a catalog.
19. NumStorePurchases: The number of purchases made by the consumer in physical stores.
20. NumWebVisitsMonth: The number of times the consumer visited the company’s website in a month.
21. AcceptedCmp3: The success of marketing campaign 3 (1 if accepted by the consumer, 0 otherwise).
22. AcceptedCmp4: The success of marketing campaign 4 (1 if accepted by the consumer, 0 otherwise).
23. AcceptedCmp5: The success of marketing campaign 5 (1 if accepted by the consumer, 0 otherwise).
24. AcceptedCmp1: The success of marketing campaign 1 (1 if accepted by the consumer, 0 otherwise).
25. AcceptedCmp2: The success of marketing campaign 2 (1 if accepted by the consumer, 0 otherwise).
26. Complain: Whether the consumer has filed a complaint (1 for complaint, 0 for no complaint).
27. Z\_CostContact: The cost associated with contacting a consumer for marketing purposes (coded value).
28. Z\_Revenue: The revenue generated from contacting a consumer for marketing purposes (coded value).
29. Response: Indicates whether the consumer accepted the offer in the last campaign (1 for acceptance, 0 for non-acceptance).

**3.2. Types of data:**

1. Integer Attributes:

* ID, Year\_Birth, Income, Kidhome, Teenhome, Recency, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, AcceptedCmp1, AcceptedCmp2, Complain,Z\_CostContact.

2. Character Attributes:

* Education, Marital\_Status, Dt\_Customer

A screenshot of a computer

AI-generated content may be incorrect.

Fig1: Structure of data

**3.4. Detailed analysis on each attribute:**

**ID:**

The attribute ID in the dataset represents a unique identifier for each consumer. Since the check confirms that each ID is distinct, it indicates that there are no duplicates in this column. This attribute is crucial for uniquely identifying each consumer, ensuring that every record corresponds to a different individual. The ID attribute serves as a key column for indexing and referencing individual consumers within the dataset.

A close up of a sign

AI-generated content may be incorrect.

Fig: checking the uniqueness

**Year\_Birth:**

The majority of customers were born between 1950 and 1985, indicating that Generation X (born 1965-1980) and Baby Boomers (born 1946-1964) make up the largest segments of the customer base.

The peak frequency is observed around the 1970s to 1980s, suggesting that most customers are currently in their 40s to 50s.

A few customers were born before 1920, representing the Greatest Generation. However, this group is minimal, suggesting their engagement with the business is very limited.

There is also a gradual decline in the number of customers born after the 1990s, indicating lower representation from Millennials.

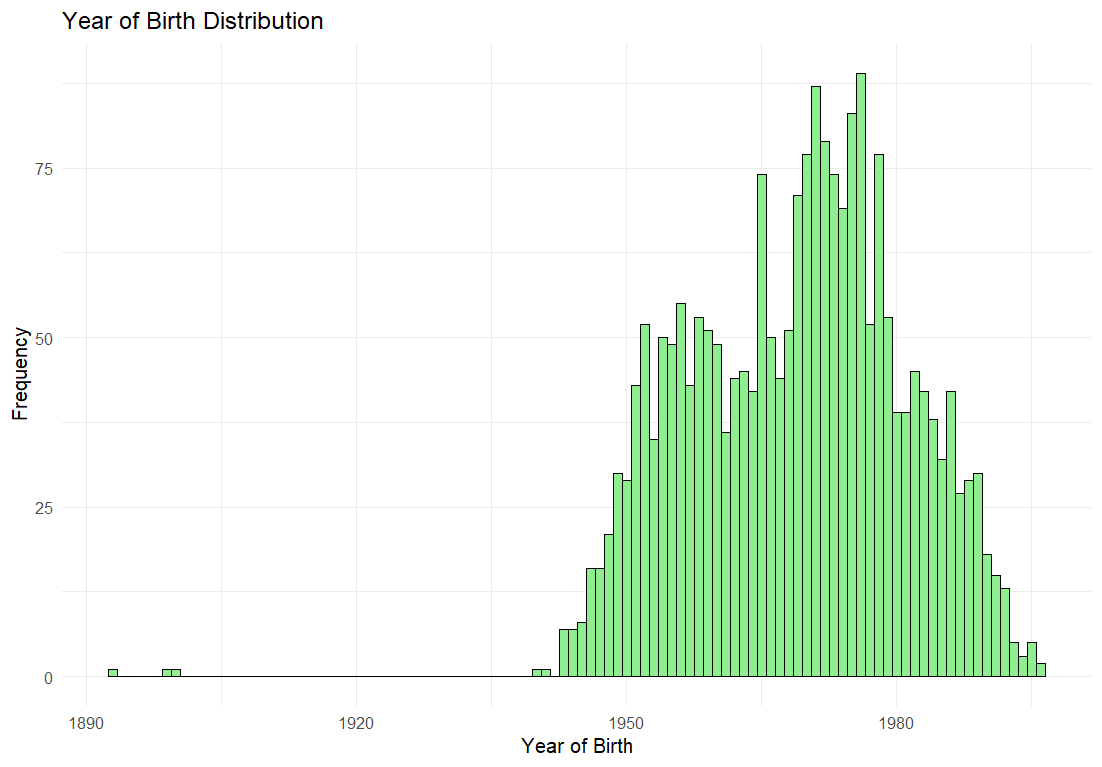
****

Fig: Distribution year of birth

To gain meaningful insights into the age distribution of Shop-Smart's customers, the data was grouped by generations.

* The analysis shows that the majority of the customers belong to Generation X (1965-1980), with 1,069 individuals aged 45 to 60 years, representing a key segment for targeted marketing campaigns.
* The Baby Boomers (1946-1964) segment comprises 759 customers aged 61 to 79 years, reflecting a financially stable group with a high potential for premium product purchases.
* Millennials (1981-1996) account for 385 customers aged 29 to 44 years, often characterized by their digital engagement and preference for personalized marketing.
* The Silent Generation (1928-1945), consisting of 24 customers aged 80 to 97 years, may respond better to traditional marketing approaches such as catalogs or in-store promotions.
* The Greatest Generation (1901-1927) includes only 3 customers aged 98 to 124 years, which is a minor segment and may be considered an outlier.

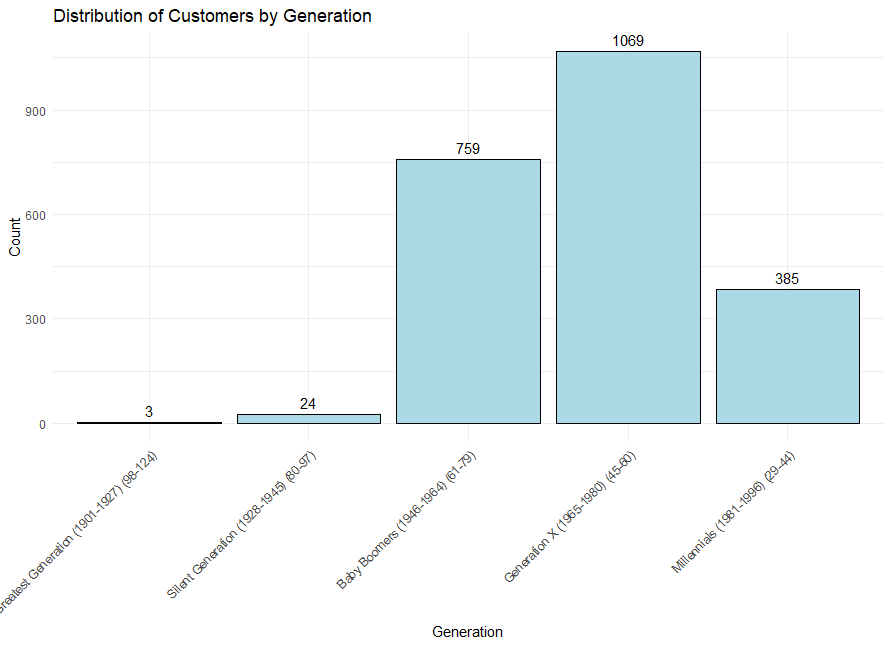


Fig: Distribution of customers by Generation

**Education:**

Education level shows that the majority of customers have a Graduation degree, with 1,127 individuals, making them the most dominant segment. This indicates a significant presence of educated consumers who may have higher purchasing power and are more likely to engage with premium products and personalized marketing campaigns.

The second largest group consists of customers with a PhD (486), followed by those with a Master's degree (370). These segments are likely to value quality and brand reputation, making them potential targets for luxury and high-value product promotions.

Customers with a 2n Cycle education level (203) and Basic education (54) form smaller segments. These groups may prefer budget-friendly products or respond better to price-sensitive promotions.

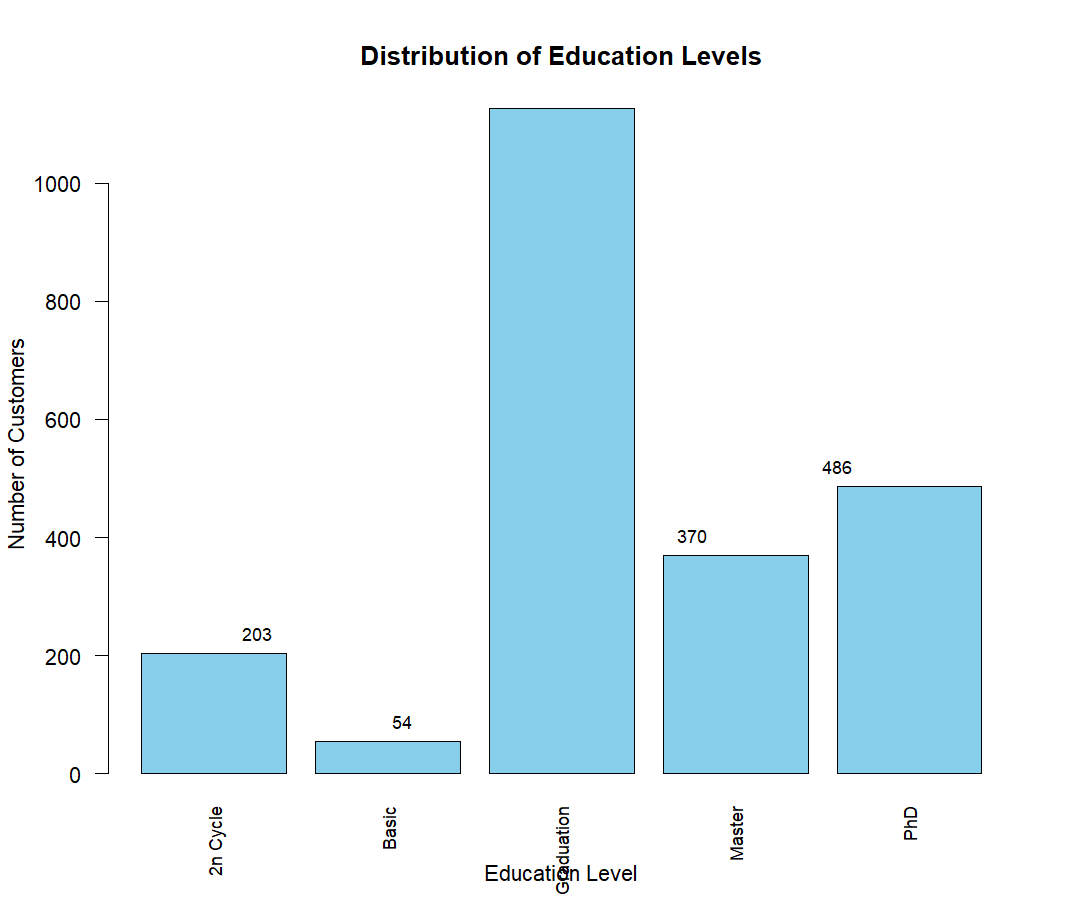


Fig: Distribution of Education levels

**Marital status:**

Marital status shows a variety of relationship statuses, with the largest group being married individuals (864). This group may represent stable households, which could be an important target for family-oriented products or services. Additionally, customers who are in a "Together" relationship (580) also form a significant portion, indicating a diverse customer base in terms of family structures.

The "Single" group (480) is also noteworthy, as they could represent individuals with disposable income who are more likely to invest in personal and luxury products or experiences.

Smaller groups include those who are divorced (232), widowed (77), and alone (3). These individuals may have specific needs or preferences that could be addressed through targeted products, such as services for independent living or specialized offerings for divorced or widowed individuals.

There are also two customers labeled as "Absurd" and "YOLO," which appear to be anomalies or outliers in the data

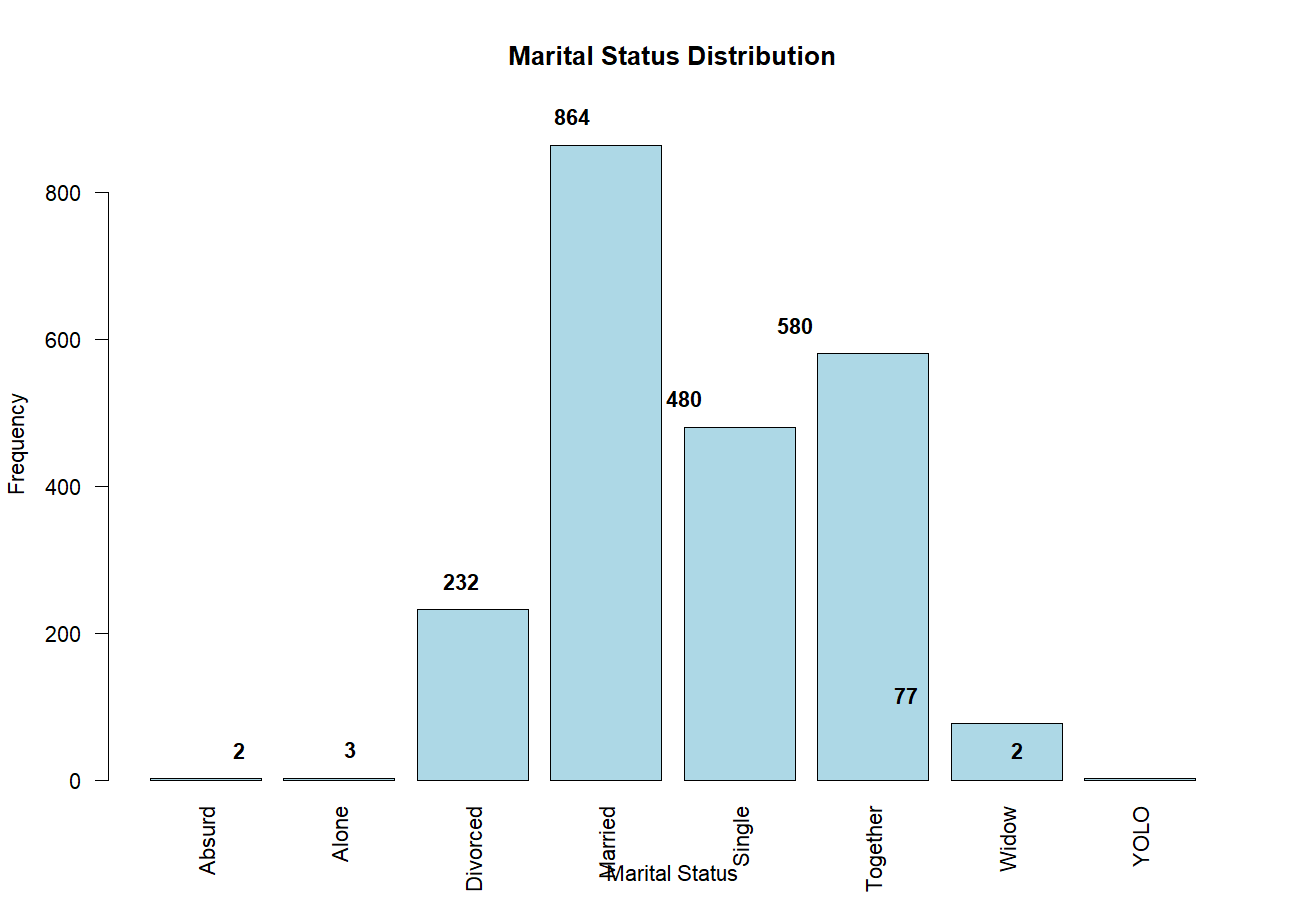
****

Fig: Distribution of marital status

**INCOME:**

The income distribution of customers reveals the following insights:

0-20K: The smallest group, with 127 customers, represents individuals with relatively lower income. This segment may be more price-sensitive and could respond better to budget-friendly products or discounts.

20K-40K: The largest group, with 605 customers, represents a significant portion of the customer base. These individuals likely have a moderate income, which might make them receptive to a mix of affordable yet quality offerings.

40K-60K: This group, with 643 customers, is similarly sizable. Customers in this bracket likely have a stable income, making them potential targets for mid-range products or services that offer a balance of quality and cost.

60K-80K: With 624 customers, this segment represents individuals with a higher disposable income. They may be inclined to purchase premium products and value added services, especially if marketing campaigns emphasize quality and exclusivity.

80K-100K: This bracket includes 204 customers. Individuals in this range likely have high purchasing power and may be more inclined to invest in high-end or luxury items, benefiting from personalized and exclusive offers.

100K+: The smallest group, with only 13 customers, represents individuals with a very high income. Although their numbers are few, this segment may be a prime target for luxury goods and high-end services.

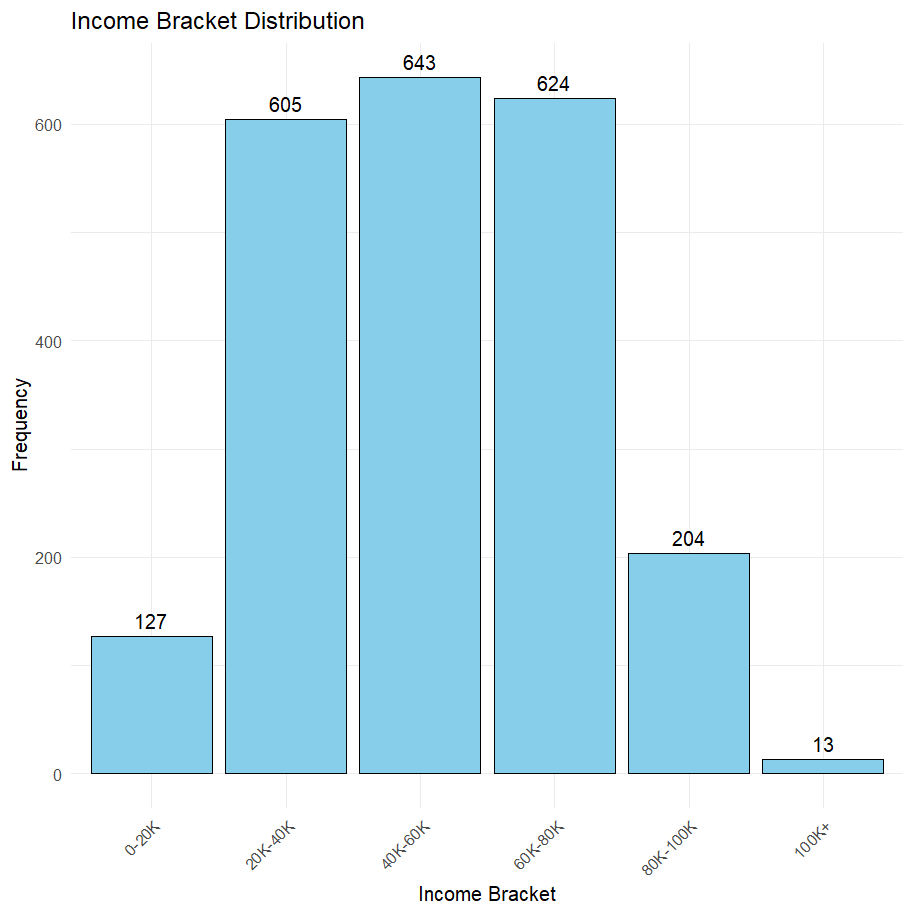


Fig: Distribution of income.

**Kidhome:**

The distribution of the Kidhomeattribute reveals that mostattribute reveals that most customers (1,283 individuals) do not have children living at home, which makes up the largest group in the dataset. A significant portion of customers (887 individuals) have one child at home, while a smaller group (46 individuals) has two children at home.

This suggests that marketing strategies and product offerings could be tailored to target primarily customers without children, with separate campaigns focused on those who have one or two children, who might have different needs or purchasing behaviors

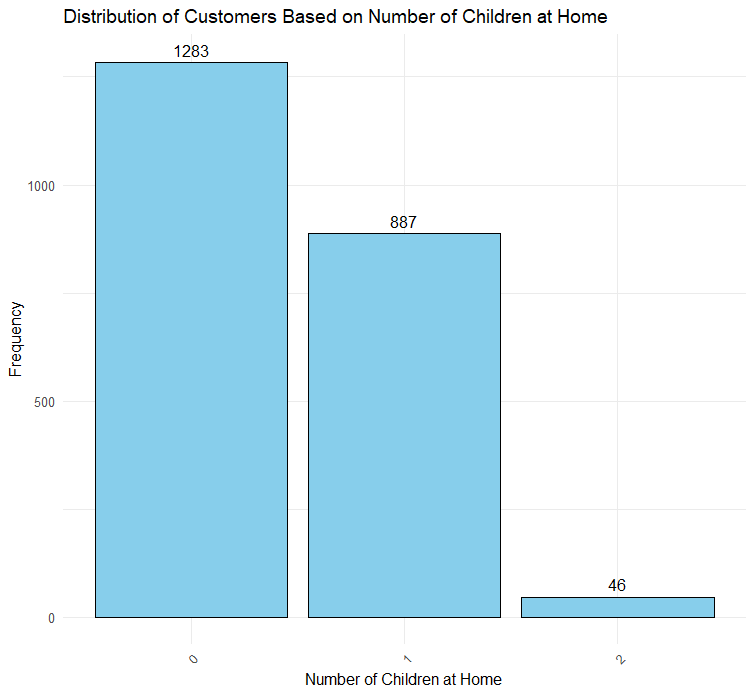


Fig: Distribution of kidHome

**Teen Home:**

The Teenhome attribute representsattribute represents the number of teenagers living in the household of the customers. The distribution shows that the majority of customers (1,147) do not have teenagers living at home. A significant portion of the customer base (1,018) has one teenager, and a smaller segment (51) has two teenagers at home.

This suggests that the customer base is primarily composed of individuals or families without teenagers or with one teenager

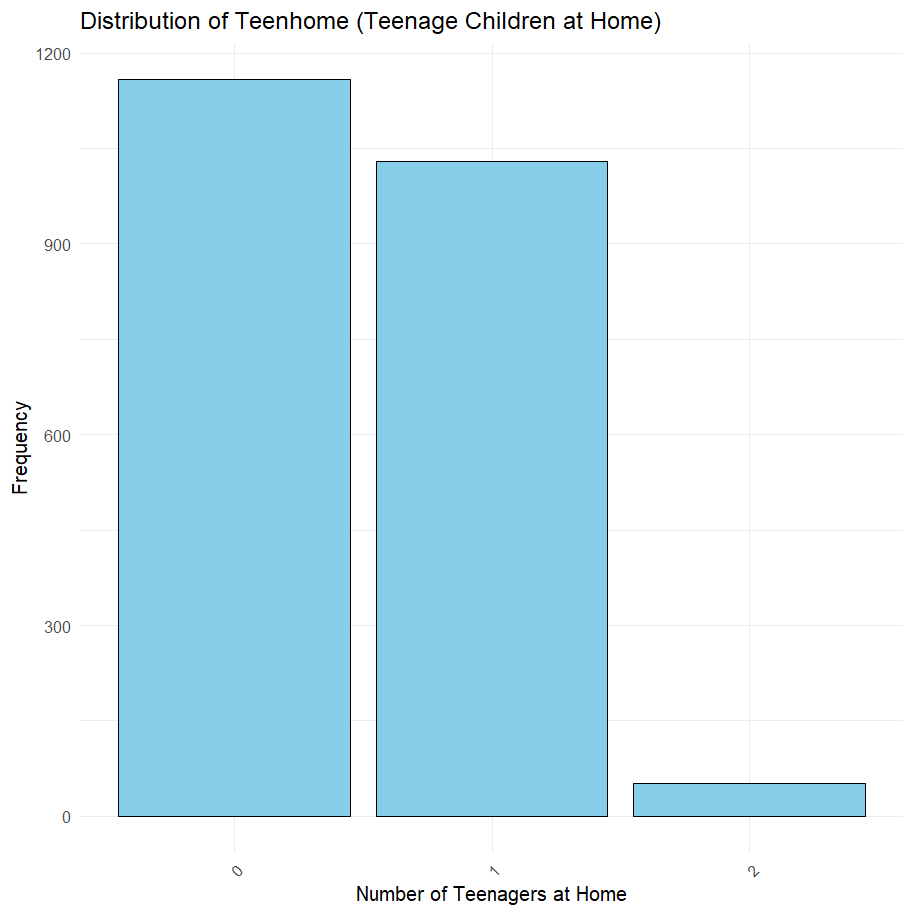


Fig: Distribution of teenagers

**Dt\_customer:**

Attribute initially contained date values stored\_The Dt\_Customer attribute initially contained date values stored as characters, which needed to be transformed for further analysis. There were two different date formats present in the dataset: "DD/MM/YYYY" and "MM/DD/YYYY." To ensure consistency, the dates were converted into a standard numerical format. The dates using the "DD/MM/YYYY" format

High Engagement in Specific Months: The months of May (216 registrations) and August (222 registrations) saw the highest customer sign-ups. This suggests that certain seasonal events, promotions, or campaigns may have influenced the increase in registrations during these months. Understanding these trends can help replicate successful strategies or identify what specifically drove customer engagement during these periods.

Weekday Patterns: Mondays (341 registrations) and Fridays (333 registrations) recorded the highest customer registration counts, indicating that customers are more likely to engage with services at the beginning and end of the week. This pattern may be due to people preparing for the upcoming week or seeking services after the weekend. This insight could be used to optimize marketing strategies by timing promotions or advertisements for these peak days.

Opportunity for Targeting Lower Activity Periods: Tuesday, with the lowest registration count (284 registrations), represents a day with relatively low customer engagement. By targeting promotions or campaigns on this day, the business could increase engagement and maintain a more consistent flow of customer registrations throughout the week.

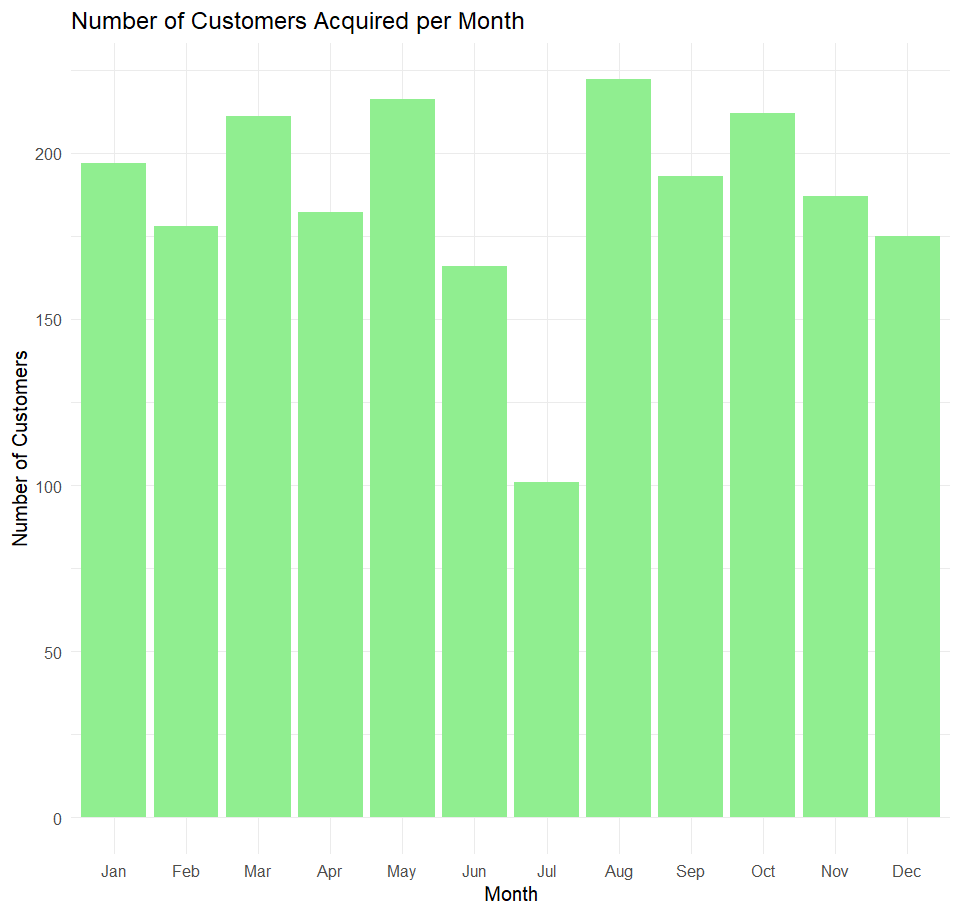


Fig: Distribution of monthly registration

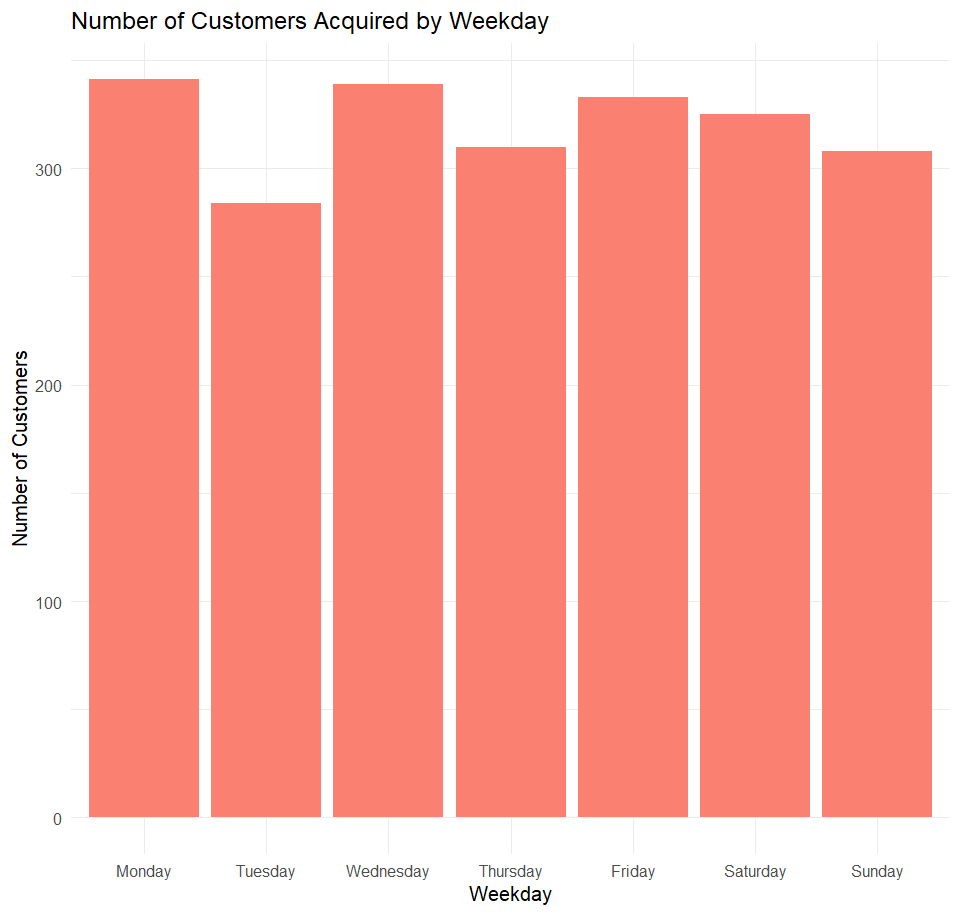


Fig: Distribution of weekly registration

Receny:

The Recency attribute measures how many days have passed since a customer's last purchase

The distribution appears relatively uniform, indicating that customers are making purchases at various time intervals without a strong concentration in a specific range.

There are noticeable peaks around 0, 25, 50, 75, and 100 days, suggesting periodic purchasing behavior, potentially influenced by monthly promotions or habitual shopping patterns.

Some customers made purchases very recently (0-10 days), indicating a segment of active shoppers.

On the other hand, a significant number of customers haven't made purchases for over 90 days, representing inactive or disengaged customers.

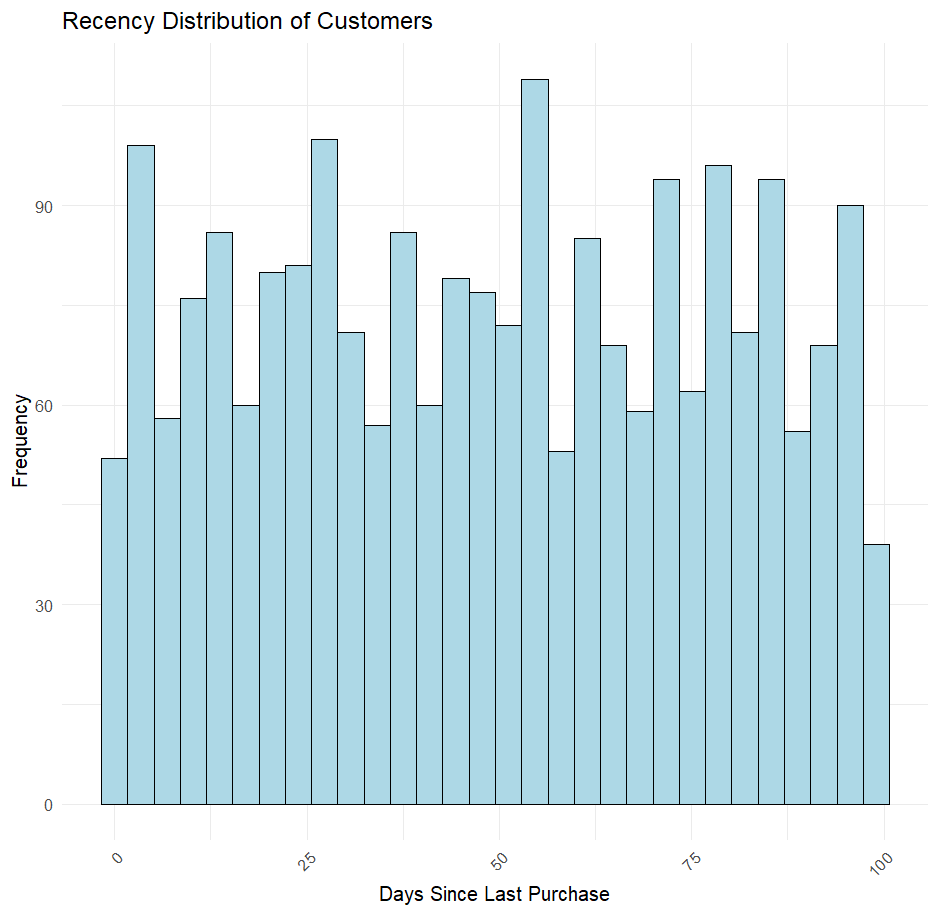


Fig: Distribution of recency

To gain a clearer understanding of customer engagement, the recency data was grouped into four categories: 0-25 days, 25-50 days, 50-75 days, and 75-100 days. This categorization helps identify patterns in customer behavior based on their last purchase.

0-25 days: 586 customers made a purchase recently, indicating a segment of highly engaged and active customers.

25-50 days: 549 customers fall in this category, representing moderately active customers who may need follow-ups or targeted marketing to encourage repeat purchases.

50-75 days: 556 customers are in this segment, suggesting a balance between active and inactive customers. These individuals might respond well to re-engagement campaigns.

75-100 days: 525 customers are in the least engaged group. Targeted promotions or personalized offers could help win back these customers and boost retention rates.

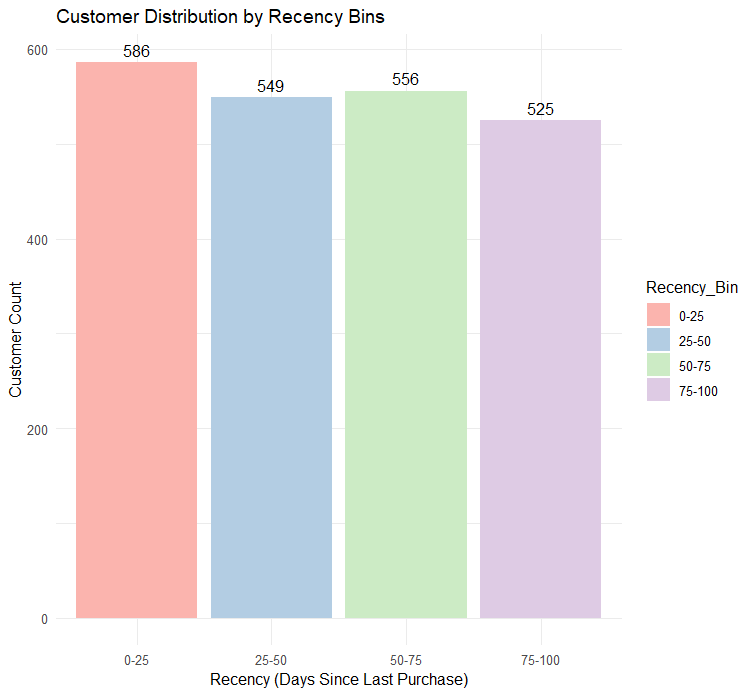


Fig: customer recency

**MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds :**

The analysis of customer spending across different product categories:

Total Spending on Wines: The total expenditure on wines amounts to 676,083. This indicates that wine is the most popular and highly purchased product among the customer base, suggesting a strong preference for alcoholic beverages, particularly wine.

Total Spending on Fruits: Customers spent a total of 58,405 on fruits. Although this is significantlyon fruits. Although this is significantly lower than the spending on wines, it still reflects a reasonable demand for healthy food options, though it may suggest that fruits are a smaller portion of customers' overall spending.

Total Spending on Meat Products: A total of 370,063 was spent on meat products. This categorywas spent on meat products. This category represents one of the higher spending areas, indicating a notable preference for meat-based products in the customer base.

Total Spending on Fish Products: With a total of 83,405 spent on fish products, this category shows a moderate demand compared to meat. It may suggest that while customers do purchase fish, its consumption is less frequent or widespread than that of other protein-rich food items like meat.

Total Spending on Sweets: The spending on sweet products totals 59,896. This reflects a moderate consumer interest in sweets, but it is relatively lower compared to items like wine or meat, suggesting that customers may treat themselves to sweets less often.

Total Spending on Gold Products: Customers have spent a total of 97,427 on gold products. This amount shows a relatively significant demand for luxury goods like gold, highlighting an interest in premium or high-value items among the customer base.

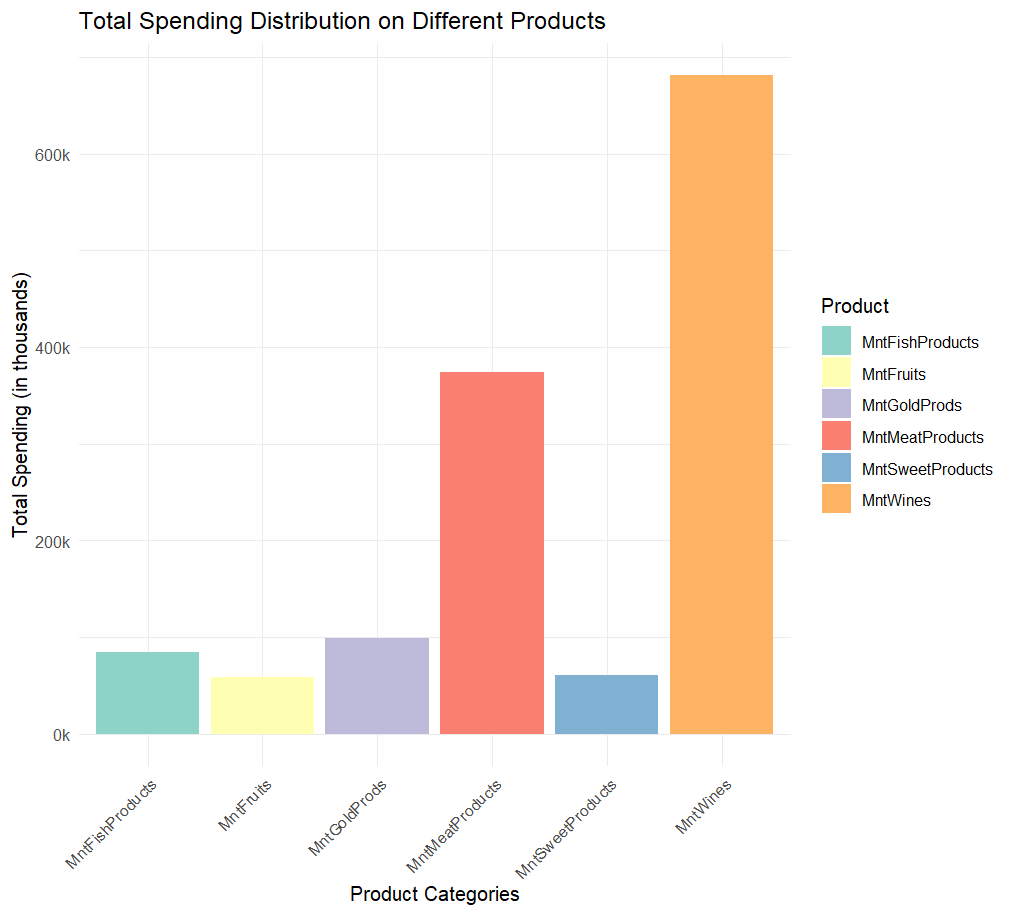


Fig: distribution of total spending of a customer

**NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases:**

Store Purchases (12,855):The majority of purchases have been made through physical stores, with a total count of 12,855. This indicates a strong customer preferenceThe majority of purchases have been made through physical stores, with a total count of 12,855. This indicates a strong customer preference for in-store shopping, which could suggest that customers value the tangible experience, the ability to interact with products firsthand, or simply the convenience of nearby stores.

Web Purchases (9,053):A significant number of purchases (9,053) have been made onlineA significant number of purchases (9,053) have been made online, highlighting the growing trend of online shopping. Customers seem comfortable with making purchases through digital channels, indicating the potential value of enhancing online platforms and offering personalized promotions to this segment.

Catalog Purchases (5,919):Catalog purchases are also notable, with 5,919 transactions recordedCatalog purchases are also notable, with 5,919 transactions recorded. Although lower than web and store purchases, this number still represents a meaningful portion of customer activity, suggesting that catalogs may still be a relevant marketing tool for certain segments, especially those who prefer traditional shopping methods.

Deals Purchases (5,149):The number ofThe number of purchases made through deals stands at 5,149, showing that customers respond positively to deals or discounts. This suggests that promotions and discounts could be an effective strategy to drive more purchases, especially if targeted toward deal-sensitive consumers.

The overall trend indicates that physical store purchases still dominate, but there is strong engagement with online shopping and deals. Companies could benefit from focusing on a balanced approach that leverages both physical and digital channels, as well as promoting special offers to further boost engagement across these platforms.

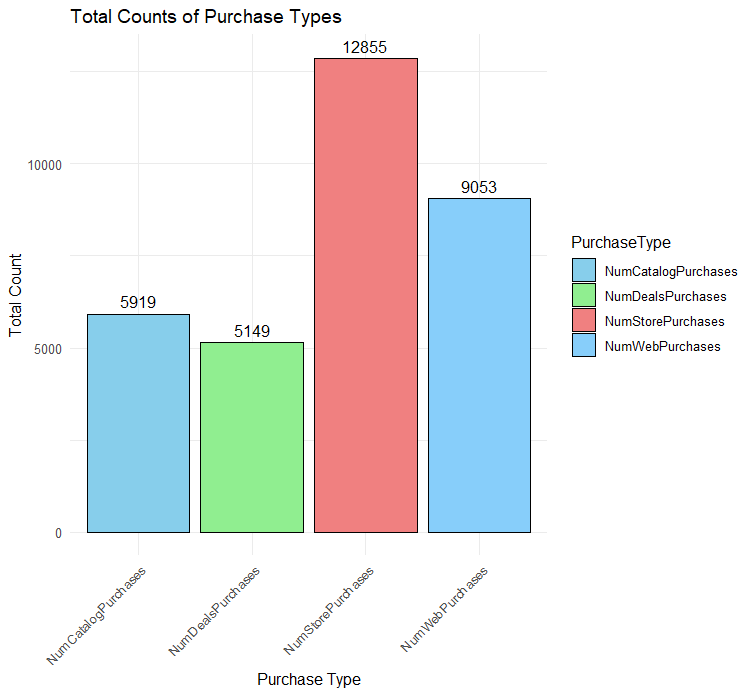


Fig: Distribution of purchase types

**NumWebVisitsMonth:**

The distribution of web visits per month shows significant variation among customers:

5 to 7 visits per month is the most common range, with 281, 340, and 393 customers respectively. This indicates a substantial portion of customers actively engage with the website, making them prime candidates for targeted marketing campaigns or personalized promotions.

0 to 2 visits per month are observed in 11, 153, and 202 customers, suggesting a lower level of interest or engagement. These users may benefit from re-engagement strategies, including promotional emails, personalized product recommendations, or exclusive offers to drive activity.

A smaller number of customers, particularly those with 9 or more visits, such as 83 customers at 9 visits and 3 customers at 10 visits, demonstrate high engagement. These customers could be further incentivized using loyalty programs or exclusive deals.

Extremely high web visit counts (13 to 20 visits per month) are rare, with very few instances, such as 2 customers with 14 visits and 3 customers with 20 visits. These users may represent comparison shoppers or customers seeking detailed information before making a decision.

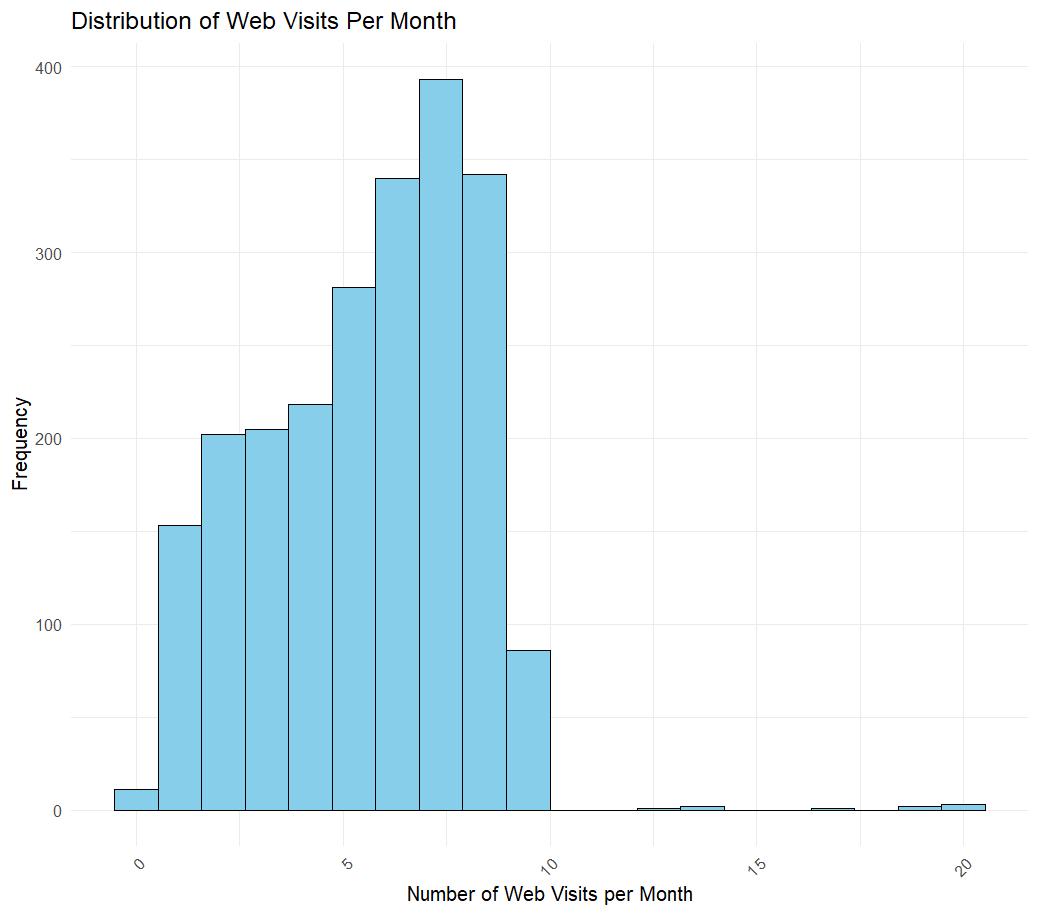
****

Fig: number of visits

**Campaign 1, Campaign 2, Campaign 3, Campaign 4, Campaign 5:**

Campaign 4 had the highest success, with 164 successful responses, indicating it was the most effective. The strategy used in this campaign could be further analyzed to identify best practices for future marketing efforts.

Campaigns 3 and 5 both achieved 163, 162 successful responses, showing similar performance. Consistent success across these campaigns suggests that the applied marketing approach resonated well with the target audience.

Campaign 1 had 142 successes, which, although lower than Campaigns 3, 4, and 5, still demonstrates considerable engagement. Further investigation could reveal areas for improvement, such as refining the message or adjusting the target audience.

Campaign 2 showed the least success, with only 30 successful responses, indicating that the approach may have been less effective. Evaluating factors such as the timing, offer appeal, or audience targeting could help enhance the effectiveness of future campaigns.

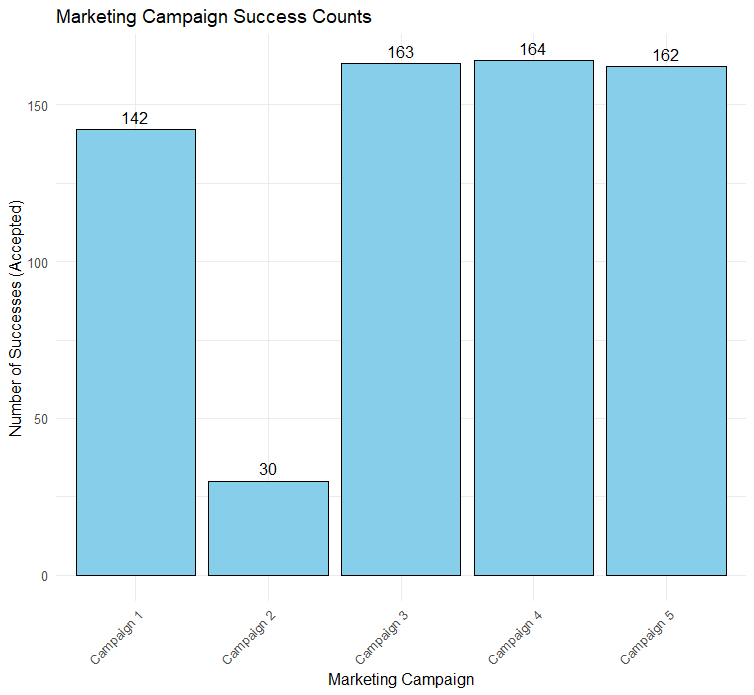


Fig: Distribution of campaigns

**Complain:**

Most customers (2,195 out of 2,240) did not complain: This indicates that 99% of customers are satisfied with the services or products, which is a good sign for overall customer satisfaction. Keeping these customers happy should remain a priority through continued good service.

A small percentage (21 out of 2,240) complained: Although complaints are minimal, making up 1% of the customer base, it's important to listen to these customers. Their feedback can help identify areas for improvement in products or services. Addressing their concerns might turn them into satisfied customers again.

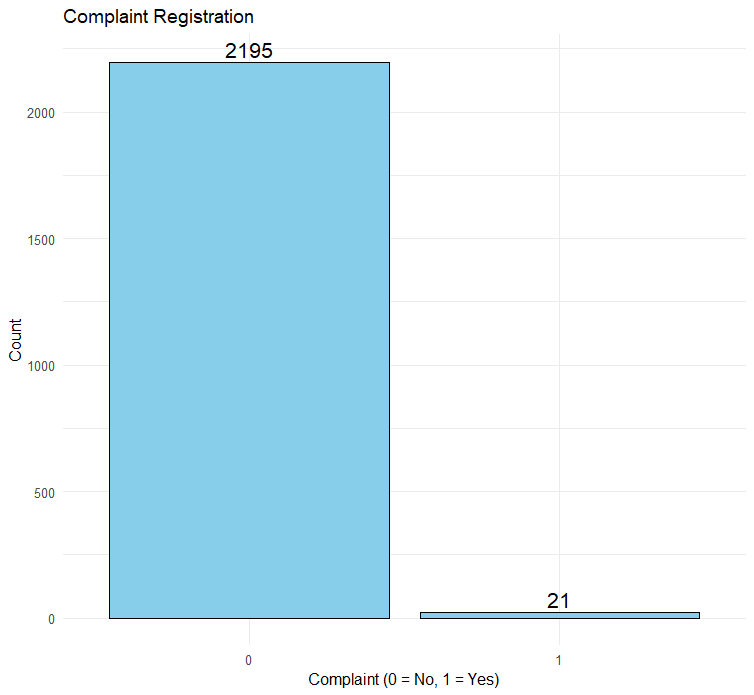


Fig: number of complaints

**Z\_CostContact, Z\_Revenue :**

Z\_CostContact (Cost of contacting a consumer)

Code 3: This code indicates a standardized cost of contacting a consumer, representing a predefined or fixed cost category for customer contact. The actual cost could be more nuanced in the company’s financial systems, but 3 acts as a simplified code for this specific category.

The value 3 means every consumer contact falls under the same cost category, simplifying budget allocation for marketing or customer outreach activities.

Z\_Revenue (Revenue from contacting a consumer)

Code 11: This code represents a fixed revenue category, denoting the company’s revenue per contact with a consumer. Just like Z\_CostContact, the value 11 is a code that refers to a certain level of revenue, simplifying calculations for revenue analysis or marketing performance.

The constant value 11 indicates that every consumer contact is expected to yield the same revenue, streamlining revenue forecasting, and simplifying financial planning.

**Response:**

Based on the Response attribute, we observe the following distribution:

1,883 consumers (0) did not accept the offer in the last campaign, which accounts for the majority of the dataset.

334 consumers (1) accepted the offer, making up a smaller proportion.

This indicates that the last marketing campaign had a relatively low acceptance rate, with approximately 17% of consumers engaging with the offer, while 83% did not respond positively. This data can be useful for assessing the effectiveness of the campaign and may suggest the need for improvements in targeting, personalization, or the offer itself to increase consumer engagement in future campaigns.

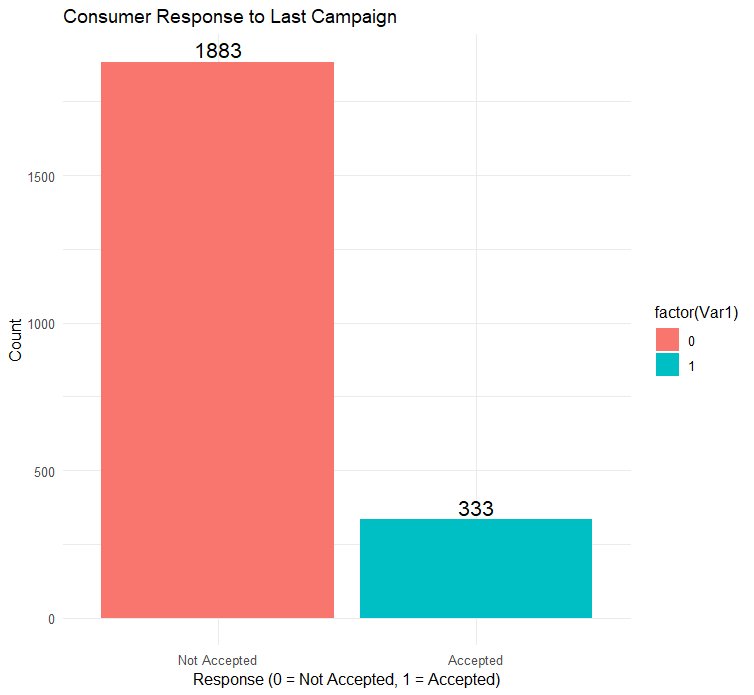


Fig: response to the last campaign

**3.5 Checking zero’s:**

There are no zeros in the dataset except the binary variables.

**A close-up of a computer screen

AI-generated content may be incorrect.**

Fig: checking zeros

**4. Data transformation:**

During the exploratory data analysis (EDA) process, several key data transformations were performed to improve the quality and usability of the dataset:

Age and Generation Classification:

The "Year\_Birth" attribute was utilized to calculate the "Age" of each consumer. This transformation allowed for a more detailed understanding of the customer demographic. Additionally, based on age, a new column was created to categorize consumers into different generations such as Baby Boomers, Generation X, Millennials, and the Silent Generation. This segmentation helps in analyzing consumer behavior across age groups and tailoring marketing strategies accordingly.

Date Transformation:

The "Dt\_Customer" (customer registration date) was initially in a single date format. This was split into individual columns for the year, month, day, and weekday. By extracting these components, trends in customer registrations could be examined over different time periods, such as identifying peak registration months or days of the week. This transformation enables a better understanding of when customers are most likely to engage with the company.

Income Transformation:

The "Income" attribute was transformed by creating an "Income\_bracket" column, which groups consumers into specific income ranges such as 0-20K, 20K-40K, 40K-60K, 60K-80K, 80K-100K, and 100K+. This transformation simplifies the analysis of how consumer income levels correlate with their purchasing behavior, product preferences, and overall engagement with the company. Missing values in Income were removed to ensure data consistency.

Recency Transformation:

The "Recency" variable, which represents the time since a consumer’s last interaction with the company, was transformed into the "Recency\_Bin" column. This categorization breaks recency into different intervals (e.g., 0-25, 25-50, 50-75, and 75-100). This transformation allows for more granular insights into consumer activity based on how recently they interacted with the company, providing a clearer picture of active versus dormant customers.

Education:

The "Education" variable originally included a category called "2nd Cycle," which was merged with "Master" to maintain consistency and reduce redundancy in the dataset. This decision was based on the fact that "2nd Cycle" represents an advanced level of education similar to a Master's degree in many education systems. By combining these categories, we ensure a more streamlined classification without losing valuable information

Target\_response:

To prepare the dataset for classification, a new binary variable named Target\_Response was created. This variable captures whether a customer has responded positively to any of the previous marketing campaigns. Customers who accepted at least one of the five specific campaigns (AcceptedCmp1 to AcceptedCmp5) or the general campaign (Response) were marked as responders (value = 1). Customers who did not respond to any campaign were labeled as non-responders (value = 0). This transformation helps to clearly define the target variable for the classification model.

**5. Predictor analysis and relevancy:**

Predictor analysis helps identify which variables (predictors) are most relevant for predicting a target variable. By evaluating the relationship between predictors and the target, we can select the most significant predictors for building efficient models.

**5.1. For classification:**

**Education vs Target\_Response:**

Basic: Among customers with a basic education, a small proportion (7 out of 54) responded to the marketing campaign, indicating a lower likelihood of engagement within this group.

Graduation: Customers with a graduation-level education show a higher response rate, with 296 out of 1,116 responding. This suggests that individuals with higher education levels are more likely to engage with marketing efforts.

Master’s: Customers with a master's degree demonstrate a moderate response rate, with 144 out of 565 responding. This indicates relatively higher engagement compared to those with only a basic education.

PhD: Customers with a PhD exhibit a strong response rate, with 158 out of 481 engaging with the campaign. This suggests that higher education levels, particularly PhDs, are associated with a greater likelihood of responding to marketing campaigns.

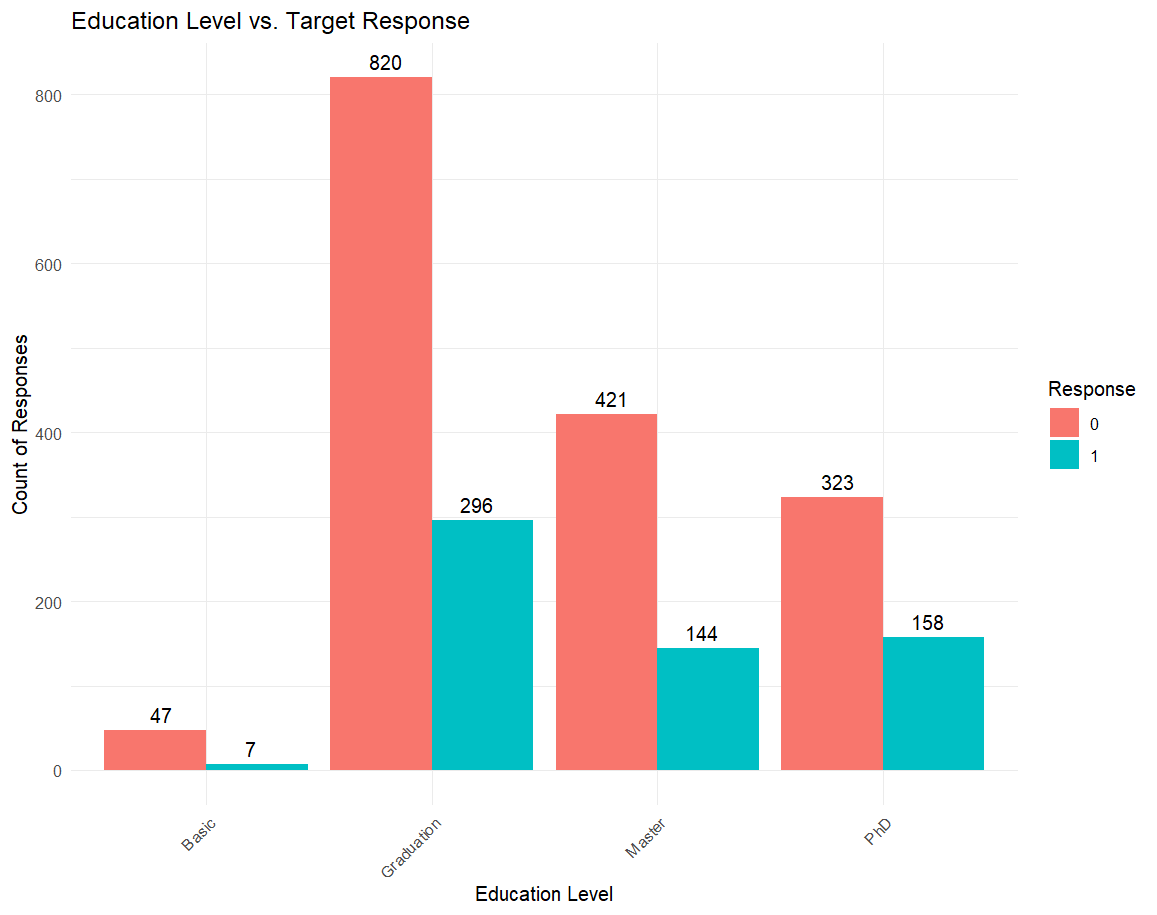


Fig: Education vs Target response

**Marital status vs Target response:**

Absurd and Alone categories, with only 1 non-responder and 1 responder, offer limited value due to their small sample sizes. These groups do not provide enough data to draw any significant conclusions about their engagement with the campaign.

The Divorced group, with a response rate of around 30%, indicates that divorced individuals engage moderately with the marketing campaigns. Although fewer than one-third respond, this group still represents a notable portion of responders and may benefit from targeted marketing strategies.

Married individuals show a response rate of about 25%, meaning that while they constitute the largest group in terms of non-responders, there is still a substantial portion of responders. This suggests that married individuals, although not highly responsive, are a significant target for marketing efforts and may require more personalized approaches to increase engagement.

The Single group, with a response rate of 31%, shows a relatively lower but still meaningful level of engagement compared to other groups. This indicates that single individuals may not be as engaged as married individuals, but they still represent a relevant market segment.

The Together group, which likely represents people in long-term relationships or partnerships, shows a more balanced engagement, with a decent response rate. The results suggest this group may be more responsive and should be considered a promising target for future campaigns.

Widowed individuals have a response rate of about 34%, which is slightly higher than the "Divorced" group. This suggests that widowed individuals are moderately responsive to marketing campaigns and could benefit from tailored outreach.

The YOLO category, with only 1 responder and 1 non-responder, offers minimal value for analysis. The small sample size does not allow for any useful insights to be drawn from this group.

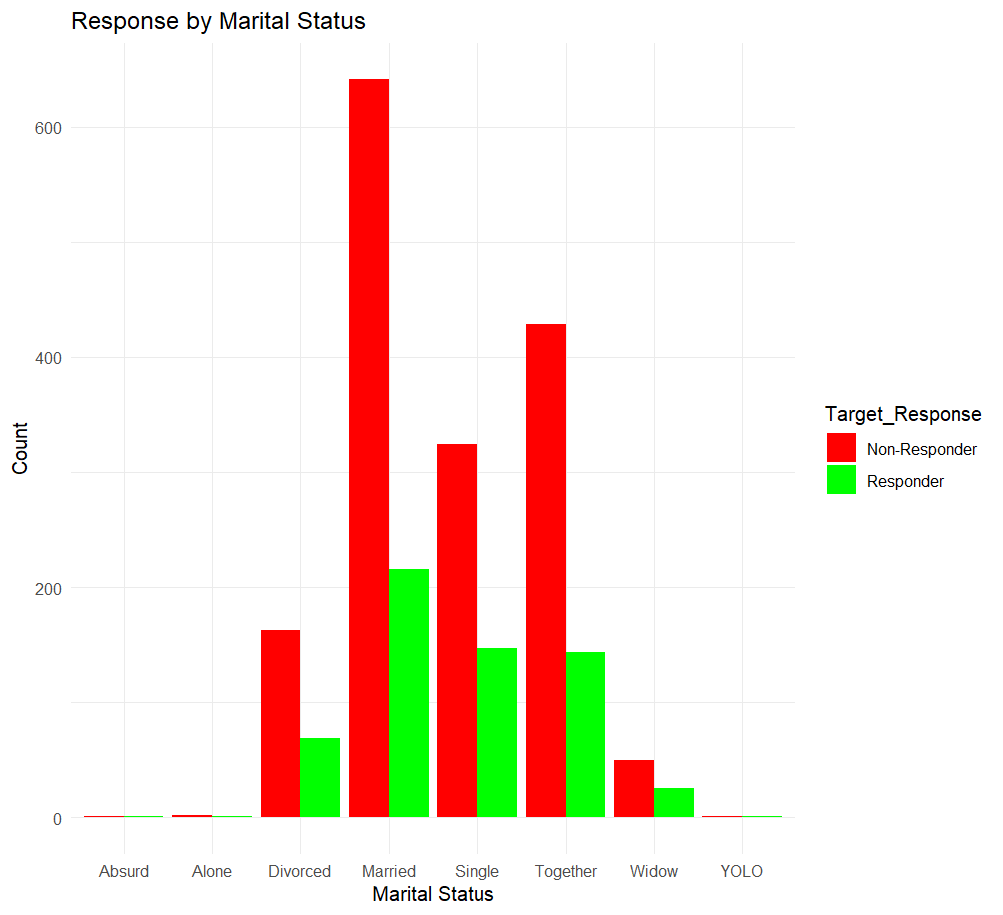
****

Fig: martial status vs target response

0 Kids at Home:

Non-Responders (853): Most individuals in this group did not respond to the marketing campaign. Responders (430): A smaller, but still significant portion, of individuals with no kids at home did respond to the campaign. Individuals without kids at home are more likely to be non-responders. However, there is still a notable proportion (approximately 33%) who do engage with the campaign.

1 Kid at Home:

Non-Responders (716): A larger proportion of people with one kid at home did not respond to the campaign.

Responding (171): Fewer responders compared to non-responders in this group.This group is less responsive to marketing efforts than the "0 Kids" group. However, there is still a noticeable response rate (about 19%) in this category.

2 Kids at Home:

Non-Responders (42): The non-responder count is high for people with two kids at home. Responders (4): Very few individuals in this group responded to the marketing campaign. This group shows a very low response rate (approximately 8%), which may indicate that individuals with more kids are less likely to engage with marketing campaigns.

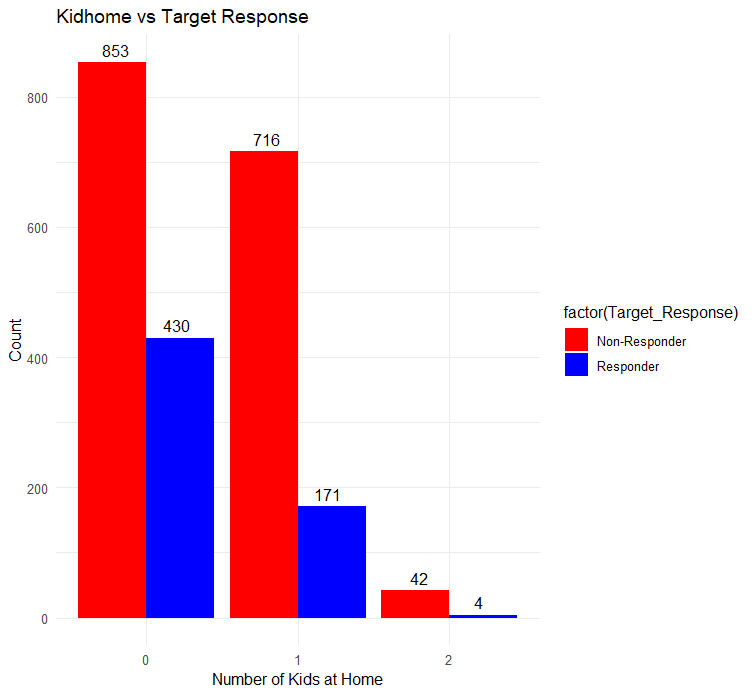


Fig: Kidhome vs target response

Results for Teenhome (Number of Teenagers at Home):

0 Teenagers at Home:

Non-Responders (767): A larger portion of this group did not respond to the campaign.

Responders (380): However, there is a significant number of responders (approximately 33%). Individuals without teenagers at home are more likely to respond to the marketing campaign compared to those with teenagers at home, but the response rate is still relatively strong.

1 Teenager at Home:

Non-Responders (805): The majority of individuals in this group did not respond to the campaign. Responders (213): A smaller proportion (approximately 21%) responded. People with one teenager at home are less likely to engage with the campaign than those without teenagers, although the response rate is still noticeable.

2 Teenagers at Home:

Non-Responders (39): The non-responder count is higher than the responders in this group.

Responders (12): Only a small number responded to the campaign (about 23%). Individuals with two teenagers at home are among the least likely to respond to marketing campaigns, showing a lower response rate compared to other groups.

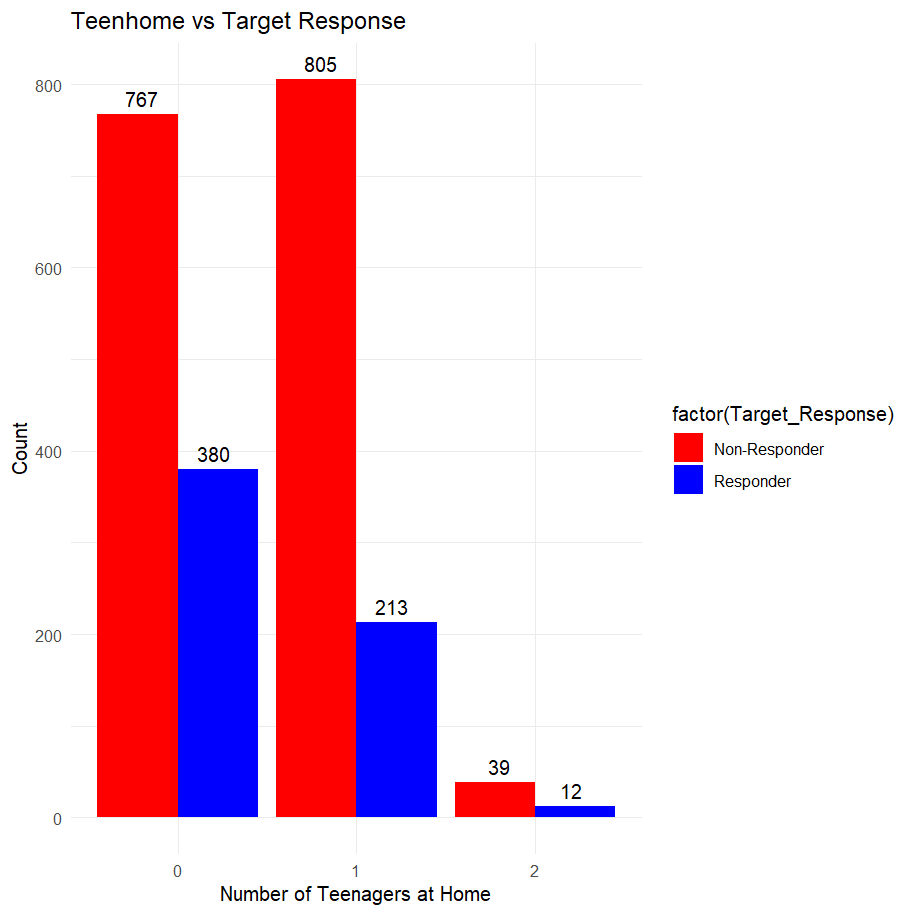


Fig: teenhome vs target response

**spending categories vs Target variable:**

MntFishProducts (Spending on Fish Products):

* Non-Responders (51945): Non-responders spent a significant amount on fish products, totaling 51,945 units.
* Responders (31460): Responders spent 31,460 units, which is notably lower than the non-responders in this category.
* Non-responders appear to allocate more spending to fish products than responders, indicating that spending on fish products may not directly correlate with engagement in the marketing campaign. However, there is still a noticeable spending pattern among responders.

MntFruits (Spending on Fruits):

* Non-Responders (36861): Non-responders spent 36,861 units on fruits.
* Responders (21544): Responders spent 21,544 units, showing a gap between non-responders and responders.
* Non-responders are spending more on fruits compared to responders, suggesting that fruit consumption may not be a strong indicator of engagement with the campaign. However, the spending difference is not as large as in the fish products category.

MntGoldProds (Spending on Gold Products):

* Non-Responders (60904): Non-responders spent 60,904 units on gold products.
* Responders (36523): Responders spent 36,523 units, also showing lower spending than non-responders.
* Non-responders are again spending more on gold products, which could imply that those who don’t engage with the campaign have higher spending habits in luxury categories like gold. However, responders still contribute a significant amount to this spending category.

MntMeatProducts (Spending on Meat Products):

* Non-Responders (203932): Non-responders spent a substantial amount of 203,932 units on meat products.
* Responders (166131): Responders spent 166,131 units, again showing a clear distinction in the spending habits of responders vs non-responders.
* The difference in spending on meat products is quite significant. Non-responders are spending much more on meat, which might suggest a potential focus on food-related purchases among non-responders, with responders still spending a substantial amount but at a lower level.

MntSweetProducts (Spending on Sweet Products):

* Non-Responders (36858): Non-responders spent 36,858 units on sweet products.
* Responders (23038): Responders spent 23,038 units, again showing lower spending than non-responders in this category.
* Similar to the other categories, non-responders have higher spending on sweet products, indicating a trend where non-responders tend to spend more on certain food items. However, responders also contribute a fair amount to sweet product purchases.

MntWines (Spending on Wines):

* Non-Responders (350873): Non-responders spent the highest amount on wine, with 350,873 units.
* Responders (325210): Responders spent 325,210 units, again showing that non-responders are leading in spending but responders still contribute significantly.
* Non-responders have the highest spending in this category, which may indicate that individuals who don't respond to campaigns are more likely to invest in luxury or leisure items like wine. Responders, however, still make up a substantial portion of the wine market.

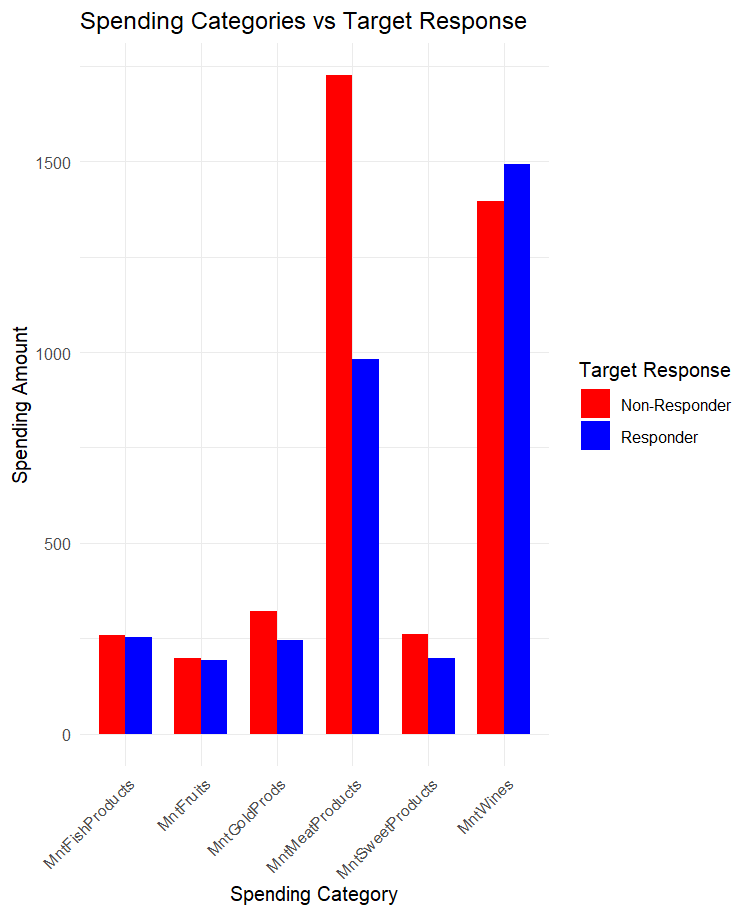


Fig: spending categories vs Target variable

**Purchasing behavior vs Target response:**

NumDealsPurchases (Number of Deals Purchased) vs Target Response:

* Non-Responders (0): The largest number of non-responders are from those who made 1 deal (664) and 2 deals (399). As the number of deals purchased increases, the number of non-responders decreases.
* Responders (1): A higher proportion of individuals who made fewer purchases (e.g., 0 deals = 26 responders, 1 deal = 296 responders) responded positively to the campaign. The response rate starts to decrease as the number of deals increases, indicating that individuals with more deals purchased are less likely to respond to the campaign.
* It seems that those with fewer purchases (especially 0 to 3) are more likely to respond, while those who purchase more deals are less likely to engage with the marketing campaign.

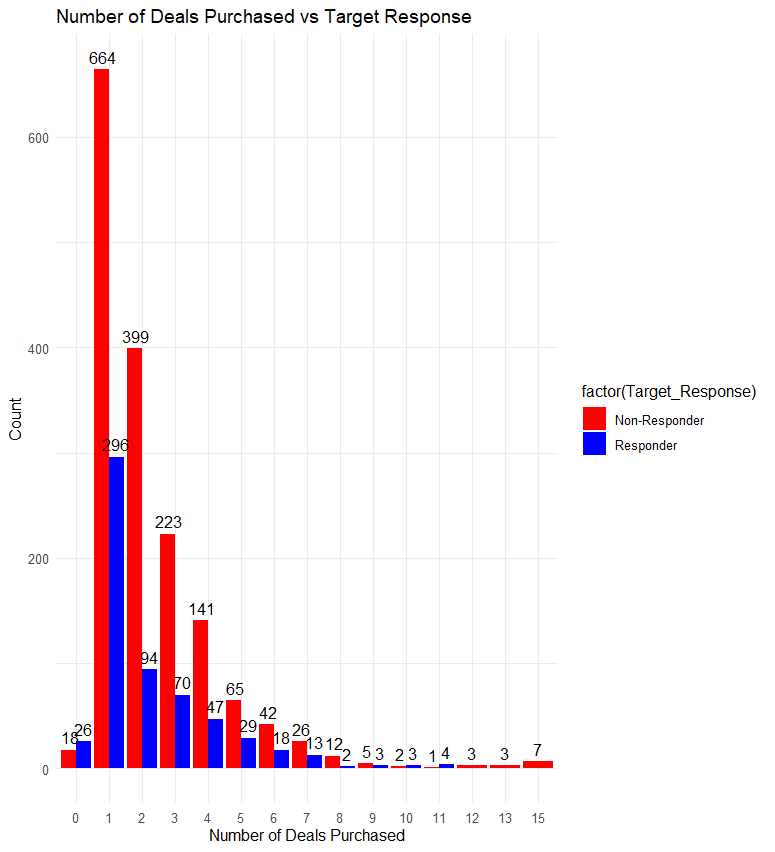


Fig: number of deals purchased vs target response

2. NumWebPurchases (Number of Web Purchases) vs Target Response:

* Non-Responders (0): The majority of non-responders are in the lower categories of web purchases, with 303 non-responders having made 1 purchase and 313 having made 2 purchases. Non-responders tend to have fewer web purchases, but the count remains spread across all categories.
* Responders (1): The number of responders is more evenly distributed across web purchases, with a notable peak in the higher categories (e.g., 84 responders with 3 purchases, 94 responders with 5 purchases).
* Web purchases seem to correlate with a higher likelihood of responding to marketing efforts. A larger portion of responders had between 3 and 5 web purchases, suggesting a strong link between frequent web purchases and campaign engagement.

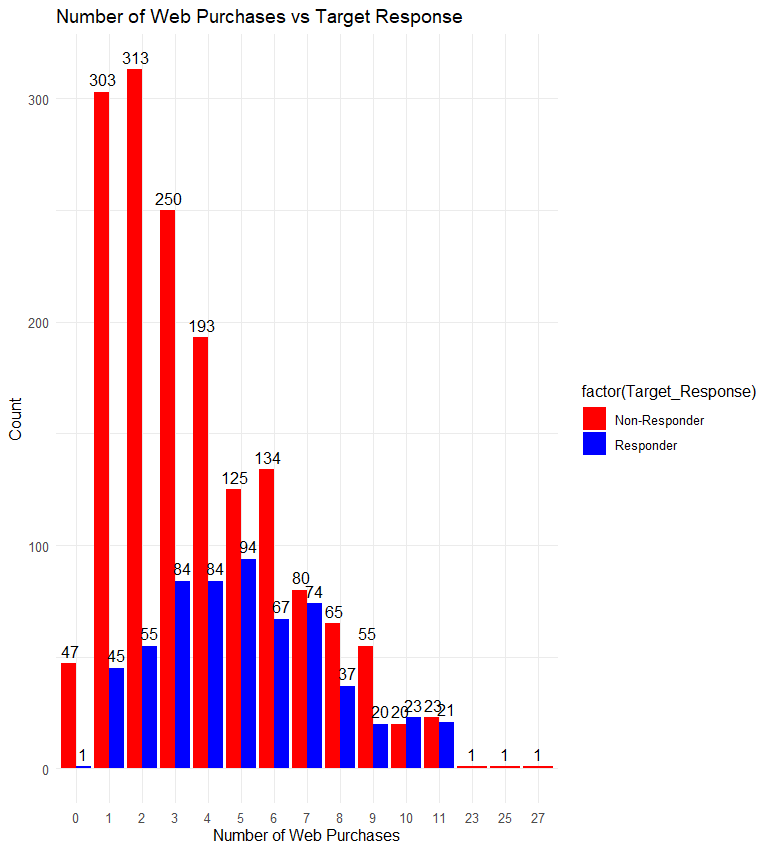


Fig: number of web purchases vs target response

3. NumCatalogPurchases (Number of Catalog Purchases) vs Target Response:

* Non-Responders (0): Catalog purchases show a similar pattern to web purchases, with a large number of non-responders in lower purchase categories (e.g., 536 non-responders with 0 purchases, 380 with 1 purchase).
* Responders (1): The number of responders increases as catalog purchases increase, with 112 responders having made 1 purchase, 59 responders having made 3, and 64 responders having made 5.
* The response rate increases with the number of catalog purchases. This suggests that individuals who engage more with catalog purchases are also more likely to engage with the campaign.

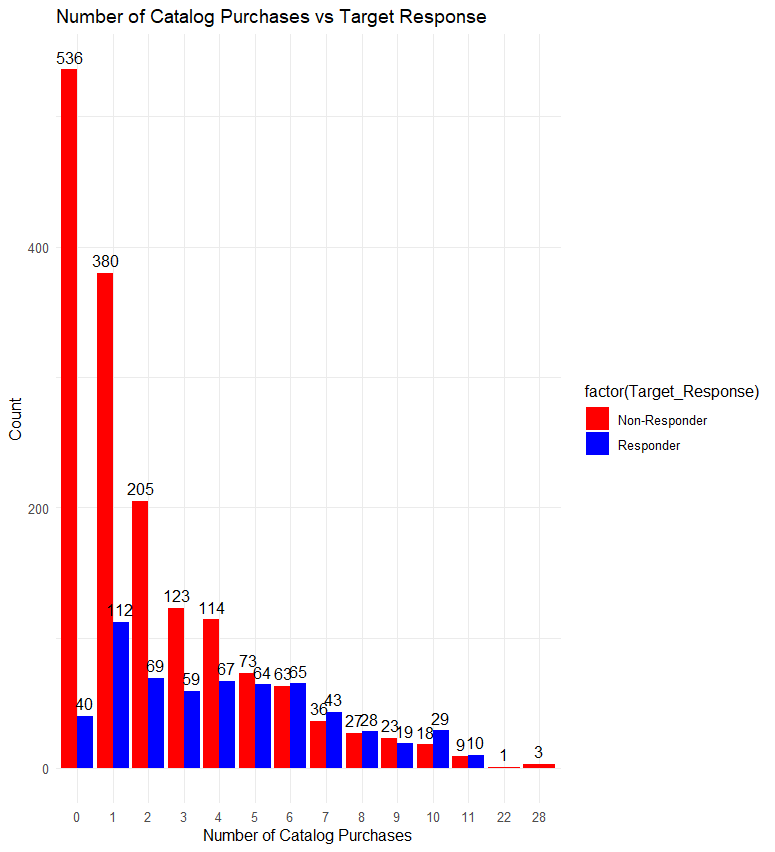


Fig: number of catalog purchases vs target response

4. NumStorePurchases (Number of Store Purchases) vs Target Response:

* Non-Responders (0): The number of non-responders is consistently high across the categories, especially for those with fewer store purchases (e.g., 433 non-responders with 3 purchases, 256 with 4 purchases).
* Responders (1): There is a steady number of responders in all categories, but a notable number (66) responded to the campaign after making 5 store purchases.
* Although non-responders dominate across the categories, store purchases seem to have a mild effect on the response rate. Individuals with more store purchases, especially those in the range of 3 to 5, tend to respond more frequently.

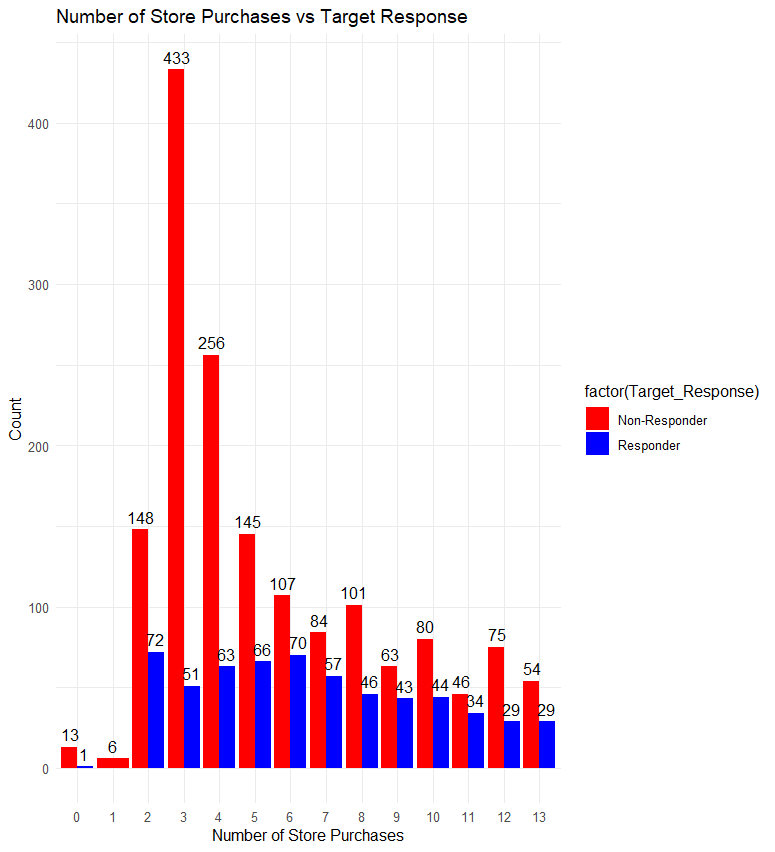


Fig: number of store purchases vs target response

5. NumWebVisitsMonth (Number of Web Visits in a month ) vs Target Response:

* Non-Responders (0): There are a large number of non-responders across the different categories, with a significant portion of them in the 1 to 3 visits range (e.g., 87 non-responders with 1 visit, 126 with 2 visits).
* Responders (1): There is a notable number of responders (63 to 91) across categories with 1 to 7 web visits per month. This suggests that individuals who visit the website more frequently are more likely to respond to the campaign.
* Frequent web visits (especially 1 to 6 visits per month) are positively correlated with the likelihood of responding to the marketing campaign. The higher the web visit frequency, the more likely individuals are to engage with the campaign.

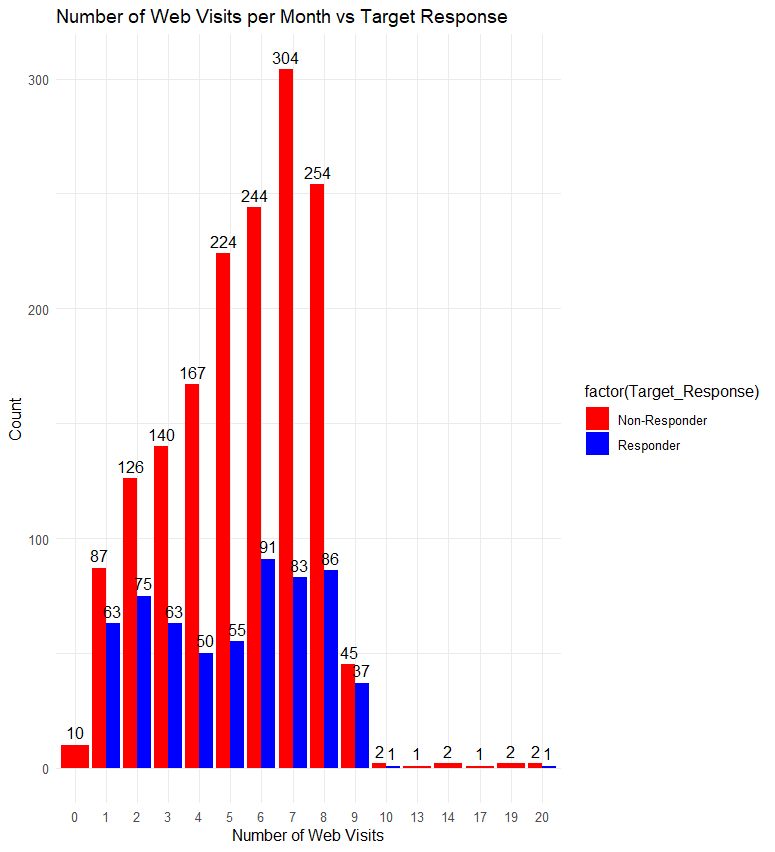


Fig: number of web visits in a month vs target response

**Complain vs target response:**

The table indicates that most customers (1593) did not complain and did not receive a positive response, while 602 customers who did not complain received a positive response. Only 3 customers who complained received a positive response, suggesting that complaints are rarely addressed positively. Additionally, 18 customers who complained did not receive a positive response, pointing to potential dissatisfaction and missed opportunities for resolution. This highlights a gap in effectively handling complaints, indicating that improving complaint management could enhance customer satisfaction and retention.

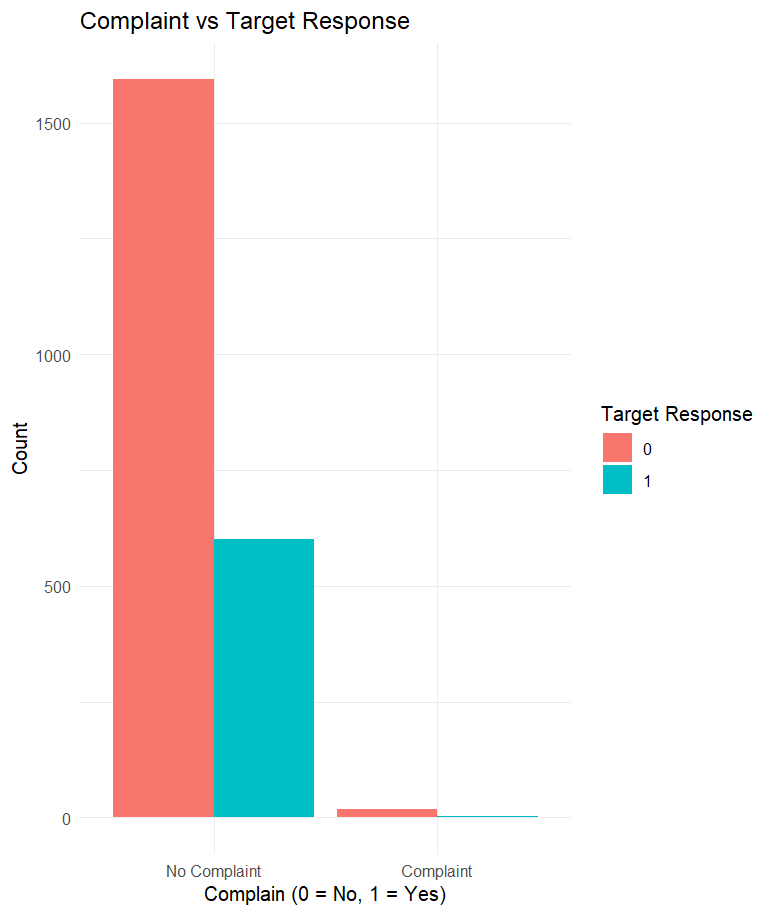
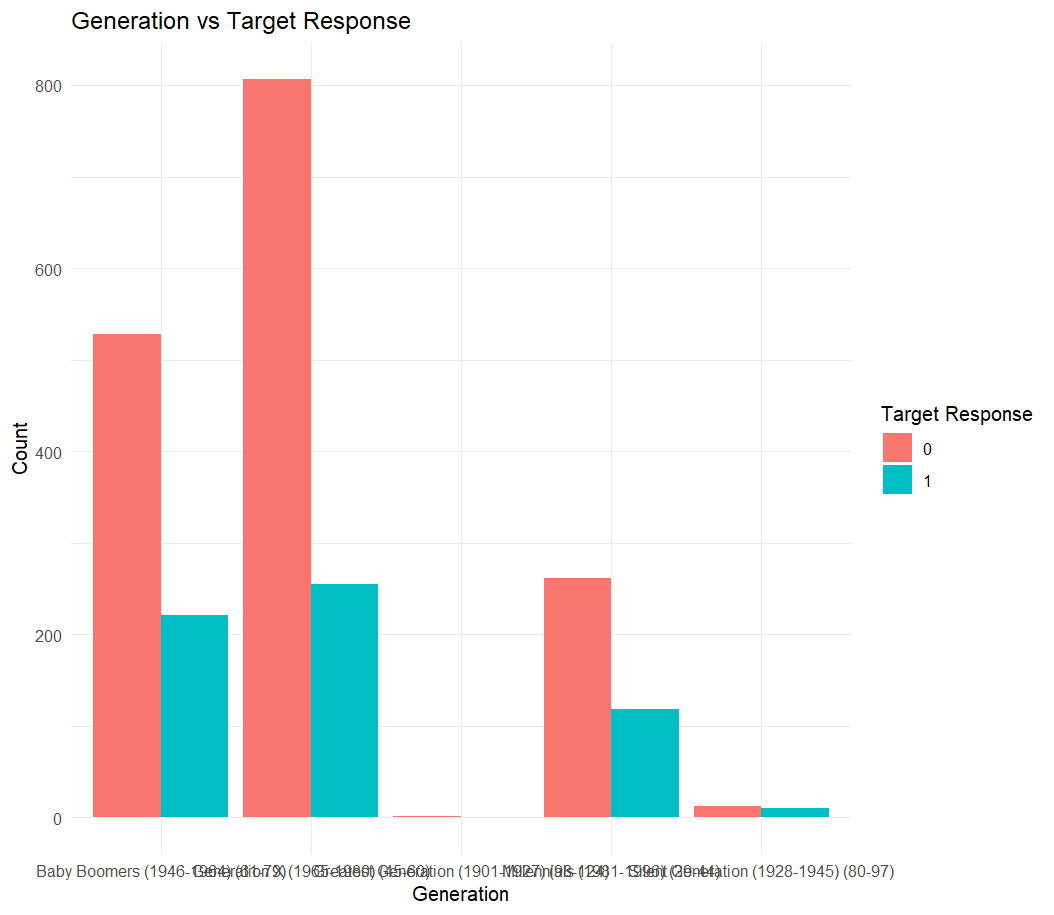


Fig: complain vs target response

Generation vs target response:

The data shows that younger generations like Millennials have a higher proportion of positive responses (31%) compared to older generations, such as Baby Boomers and Generation X, which have lower positive response rates (around 30% for both). The Greatest Generation and Silent Generation have very small sample sizes, making it difficult to draw robust conclusions for these groups. Overall, the trend suggests that the younger generations may be more likely to receive positive responses, but the majority across all generations did not receive a positive response. This could be an important factor in understanding customer satisfaction and the effectiveness of response strategies for different age groups.

A number with numbers and numbers

AI-generated content may be incorrect.

Fig: generation vs target response

**Recency vs target response:**

Customers with more recent interactions (0-25 recency bin) had the highest number of positive responses (202 out of 586), suggesting that recent engagement increases the likelihood of a positive response. As recency increases, the number of positive responses declines, with the 75-100 recency bin showing the lowest engagement (115 out of 525). This pattern indicates that maintaining regular customer interactions is crucial for increasing response rates.

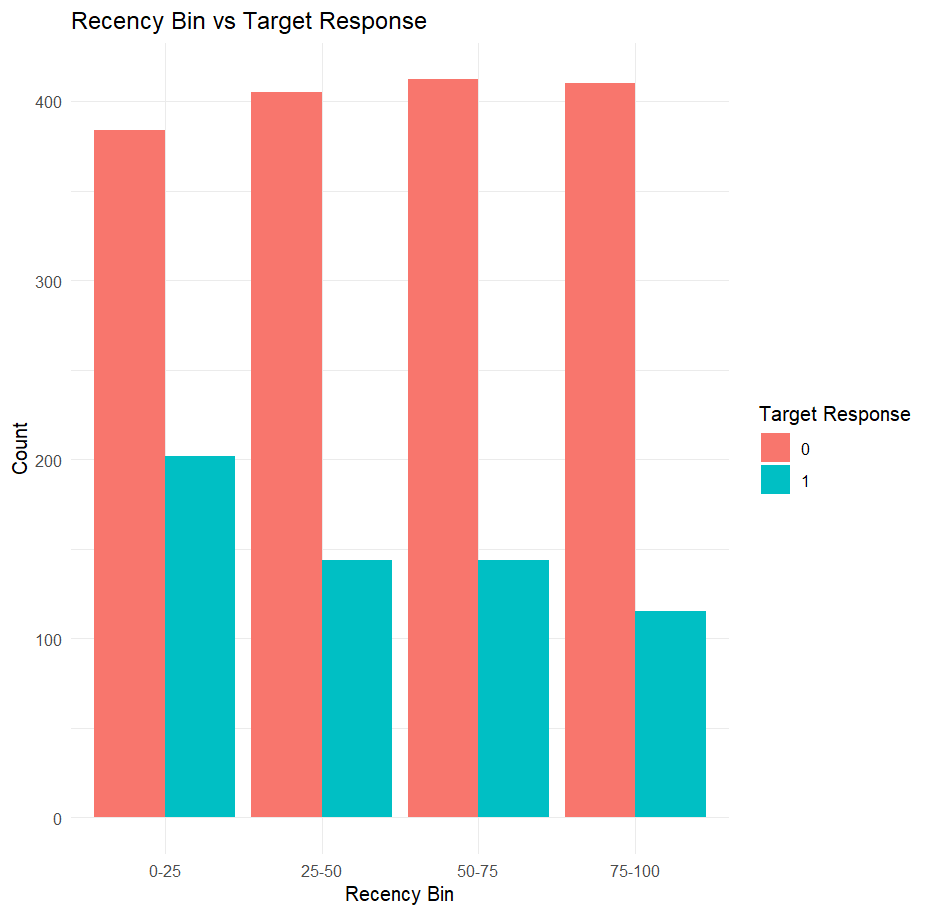


Fig: recency vs target response

**Income vs target response:**

Income levels significantly impact response behavior. The 60K-80K and 80K-100K income brackets have the highest positive response counts (200 and 143, respectively), suggesting that middle-to-upper-income customers are more likely to engage. Despite having a large number of customers, the 40K-60K group shows relatively fewer positive responses (140 out of 643), possibly indicating a need for more targeted marketing. On the other hand, the 0-20K and 100K+ groups have the lowest engagement, implying that customers at both income extremes may require different strategies to improve response rates.

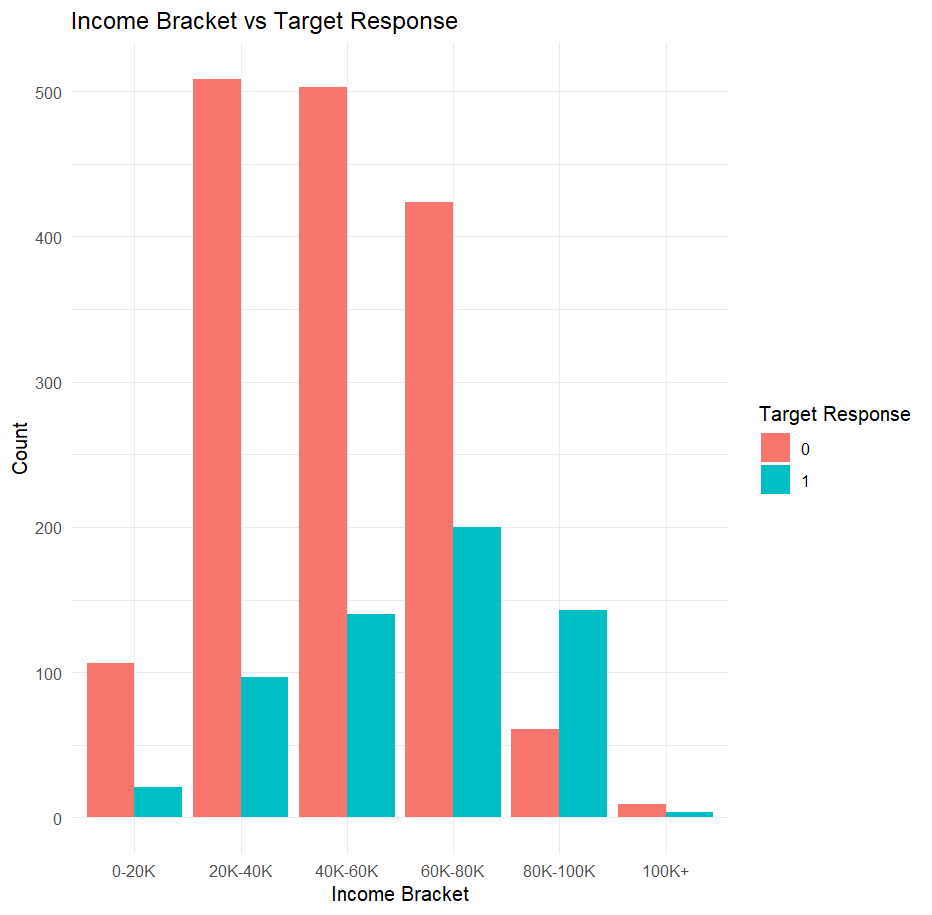


Fig: income vs target response

**5.2. For clustering:**

To achieve meaningful customer segmentation, we selected variables that provide valuable insights into demographics, shopping behavior, and engagement patterns. These variables contribute directly to identifying distinct customer groups and tailoring marketing strategies accordingly.

1. Demographics and Shopping Behavior

These variables will likely be the key drivers of clustering, as they provide important insights into the customer profiles. Retain these variables:

* Education: Customer education level might affect their buying behavior and product preferences.
* Marital\_Status: Marital status could influence the types of products customers are interested in (e.g., single customers vs. married customers).
* Income: A key factor in understanding purchasing power and preferences.
* Generation: Generation can have an impact on spending habits, product preferences, and digital engagement.
* Recency (or Recency\_Bin): Recency is crucial for clustering as it measures how recently a customer made a purchase or interacted with the company, which reflects their current engagement with the business

2. Shopping Preferences and Spending Behavior

Variables that represent the customer’s spending habits are critical for segmentation based on purchasing behaviors.

* MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds: These variables indicate the amount spent on various product categories, which is useful in understanding customer preferences.
* NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth: These variables capture customer interaction and buying channels (web, catalog, store), helping you identify how customers prefer to shop.
* AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5: These variables represent responses to marketing campaigns, which can provide insight into customer engagement and responsiveness to promotions.

In alignment with the goal of segmenting customers based on shopping behaviors, demographics, and preferences, several variables do not directly contribute to this objective and can be removed. These variables are less relevant for clustering as they do not offer meaningful insights into customer segmentation.

ID: The customer ID is a unique identifier and does not provide any insight into the customer’s behavior, preferences, or demographics. It does not contribute to segmentation and can be excluded from the analysis.

Z\_CostContact, Z\_Revenue: These variables are likely constant across all customers and do not offer variability that would help in distinguishing customer segments. As a result, they do not provide valuable information for clustering and can be removed.

Day, Month, Weekday: These temporal variables, while potentially useful for seasonality analysis, may not offer much value for customer segmentation unless the goal is to segment customers based on specific times of the year. For clustering, the focus should be on customer behaviors and demographics rather than the specific days or months of interactions.

Dt\_Customer: The date when the customer was added to the database may not reflect their shopping behavior directly. Unless the objective is to analyze customer tenure or behavior changes over time, this variable does not add value to the clustering analysis.

Income vs. Income\_bracket: Since Income\_bracket is a categorical transformation of Income, keeping both may be redundant. If clustering requires precise income-based segmentation, the continuous Income variable should be retained. However, if grouping customers by broader income categories is more appropriate, the Income\_bracket variable can be kept instead.

Complain: While useful for customer satisfaction analysis, the Complain variable may not be directly relevant for segmentation unless the segmentation strategy is focused on customer satisfaction. If customer satisfaction is not a primary factor in the segmentation, this variable can be excluded.

**6. Dimension reduction:**

**Dimensionality Reduction for Classification:**

In order to enhance the efficiency, interpretability, and predictive power of our classification model, we performed dimensionality reduction by removing redundant, high-cardinality, and non-contributory features. This ensures that our model focuses on the most relevant attributes while reducing noise and potential overfitting. Below are the details of the features removed and the reasoning behind these decisions:

1. High-Cardinality Variable (Unique Identifiers)

ID: This column contains unique values for each customer and does not provide any meaningful pattern for classification. Since it does not contribute to the predictive modeling, we have removed it.

2. Redundant Features (Derived Information Already Exists)

Year\_Birth: Instead of using the raw birth year, we derived Age and Generation, which provide more interpretable insights into customer demographics.

Income: The income variable was transformed into Income\_bracket, which categorizes customers into meaningful income segments, making it more suitable for classification.

Dt\_Customer: This column represents the date when a customer enrolled. However, we have already extracted meaningful time-based features (Year, Month, Day, and Weekday). Among these, only Month and Weekday are retained as they are more relevant for analysis.

Recency: Instead of using Recency as a continuous variable, we created Recency\_Bin, which groups customers into meaningful recency categories. This makes the data more structured for classification.

3. Marketing Campaign Variables (Consolidated Information).

AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5: These variables indicate responses to different marketing campaigns. However, we have already consolidated this information into Target\_Response, making individual campaign re sponses redundant.

Response: Since we derived Target\_Response, which represents the overall response to marketing efforts, the original Response column is no longer needed.

**Dimensionality Reduction for clustering:**

To improve clustering efficiency and interpretability, we removed variables that do not contribute meaningfully to segmentation:

High Cardinality & Identifiers: ID (unique identifier, not useful for clustering).

Constant/Low Variability: Z\_CostContact, Z\_Revenue (likely constant, no impact on clustering).

Temporal Variables: Dt\_Customer, Day, Month, Weekday, year (not directly relevant unless segmenting by seasonality).

Redundant Features: Income (replaced by Income\_bracket for categorical grouping), Recency (replaced by Recency\_Bin

Year\_Birth, Age: Removed since Generation already provides demographic

segmentation.Customer Feedback: Complain (only relevant if segmenting by satisfaction).

**7. Data partition:**

To build a robust predictive model, to split the dataset into training and testing subsets. This allows us to train the model on a portion of the data and evaluate its performance on unseen data.

For Classification:

will divide the dataset into three parts:

Training Set (70%) – Used to train the machine learning model and help it learn patterns.

Validation Set (15%) – Used to fine-tune the model and adjust parameters to avoid overfitting.

Testing Set (15%) – Used to evaluate how well the model performs on completely new data.

A 70-15-15 split balances learning, tuning, and evaluation.

The validation set helps improve model performance before testing.

The test set remains untouched until the final evaluation, ensuring an unbiased performance check.

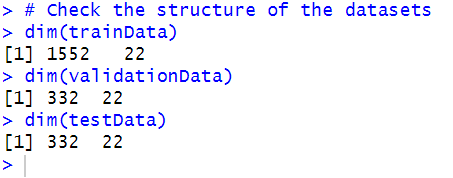


Fig: data partition

for Clustering:

Unlike supervised learning, where data is split into training and testing sets, clustering involves unsupervised learning, meaning there is no outcome variable to predict. Therefore, traditional data partitioning is not required. Instead, the entire dataset will be used to identify natural groupings within the data.

**8. Model selection:**

**Model selection for classification:**

To enhance customer engagement and marketing efficiency, we have selected Logistic Regression and Random Forest as our classification models. These models will help Shop-Smart predict customer behaviors, such as purchase likelihood and response to promotions, enabling data-driven marketing strategies.

Logistic Regression is chosen for its simplicity and interpretability. It helps Shop-Smart determine which factors such as customer demographics, past purchases, and promotional interactions affect the likelihood of a customer making a purchase or responding to a campaign. The probability scores from this model enable targeted marketing efforts, improving conversion rates and customer engagement.

For the logistic regression model in the consumer response prediction project, dataset was used to fit the model using the glm() function with the family = binomial argument. The model aimed to classify whether a consumer would respond positively (1) or not (0) to a marketing campaign. After fitting the model on the training set, predictions were generated on the validation set using the predict() function, and performance was evaluated using a confusion matrix via the confusionMatrix() function. The same process can be applied to the test set for final evaluation.

A screenshot of a computer

AI-generated content may be incorrect.

Fig: logistic regression

Accuracy: The model correctly classified 76.81% of the validation observations, indicating strong overall predictive performance. This means that roughly three-quarters of the consumers were correctly identified as responders or non-responders.

Sensitivity (Recall / True Positive Rate): 33.33% – The model correctly identifies only one-third of the actual responders. This suggests that the model misses a large number of true positives, indicating limited effectiveness in catching actual responders.

Specificity (True Negative Rate): 90.84% – The model is highly effective at identifying non-responders, correctly predicting over 90% of actual non-responses. This is significantly higher than its sensitivity, highlighting an imbalance in detection performance.

Balanced Accuracy: 62.08% – This metric accounts for both sensitivity and specificity, indicating moderate effectiveness across both classes.

Positive Predictive Value (Precision): 54.00% – When the model predicts a positive response, it's correct about half the time.

Negative Predictive Value: 80.85% – When predicting a non-response, it's accurate in the majority of cases.

Random Forest:

Random Forest is selected for its ability to handle complex patterns in customer behavior. Unlike Logistic Regression, it captures nonlinear relationships and interactions among features, allowing for more accurate predictions. It also ranks feature importance, helping Shop-Smart identify key drivers of customer decisions, such as preferred product categories, shopping frequency, or price sensitivity.

By utilizing Logistic Regression for clear insights and Random Forest for high accuracy, Shop-Smart can classify customers based on their purchasing behaviors, promotional responses, and engagement levels. This classification enables the company to design personalized marketing campaigns, optimize promotions, and enhance customer retention strategies, ultimately driving business growth.

For the random forest model in the consumer response prediction project, training dataset was used to fit the model using the ranger() function with 100 trees. The model was trained using key predictors such as consumer behavior metrics, demographic indicators, and campaign-related attributes to classify whether a consumer would respond (1) or not (0) to a marketing campaign.

The model's predictions were generated for the validation set using the predict() function from the ranger package, and its performance was evaluated using a confusion matrix with confusionMatrix(). A similar evaluation can be applied to the test dataset for final performance analysis.

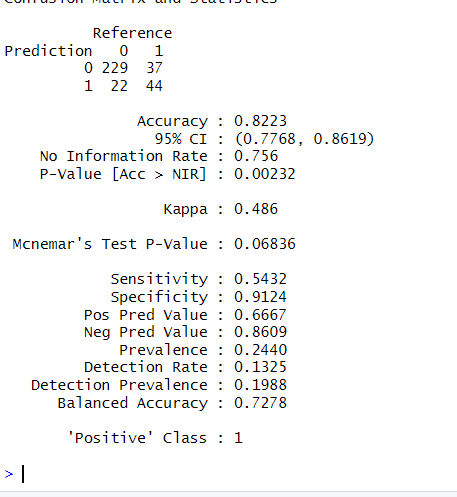


Fig: Random forest

**Model selection for clustering:**

**K-Means Clustering:**

K-Means clustering is selected for segmenting consumers based on purchasing behavior, demographics, and marketing engagement. This algorithm efficiently groups individuals with similar characteristics, enabling the identification of distinct consumer segments. K-Means is scalable and performs well with large datasets, making it suitable for analyzing customer data to develop targeted marketing strategies. The model helps uncover patterns in consumer behavior, allowing businesses to optimize their campaigns and tailor promotions for different customer groups.

Necessity of Normalization – Consumer Clustering Project

Normalization is a critical preprocessing step in this project, especially because the dataset includes features with widely varying scales and mixed data types. Here's why:

Mixed Feature Scales:

Your dataset includes numerous numerical variables with different ranges:

Spending variables like MntWines, MntMeatProducts, MntFishProducts, etc., range from 0 to several hundred.

Frequency variables such as NumWebPurchases, NumCatalogPurchases, or NumWebVisitsMonth typically have small integer values.

Binary variables like AcceptedCmp1, Response, and Target\_Response are in 0/1 format.

Without normalization, features with larger values (like total purchase amounts) would dominate distance calculations in clustering algorithms like K-Means or Hierarchical Clustering, leading to biased cluster formation.

**Determining the Optimal Number of Clusters using the Elbow Method:**

To ensure effective and meaningful segmentation of consumers, the Elbow Method was applied to determine the optimal number of clusters for the K-Means clustering model. This method analyzes the Within-Cluster Sum of Squares (WSS), which quantifies the compactness of clusters by measuring how closely data points within the same cluster resemble each other.

A plot of WSS versus the number of clusters (K) was generated. In this plot, the “elbow point” where the WSS curve begins to flatten indicates the best number of clusters. This point reflects where adding more clusters does not significantly reduce the WSS, thus avoiding unnecessary complexity while maintaining meaningful segmentation.

A graph of a number of clusters

AI-generated content may be incorrect.

Fig: elbow method

Based on the Elbow Method plot:

The WSS shows a steep drop between K = 1 and K = 3, suggesting significant improvement in cluster cohesion.

After K = 3, the curve begins to level off, indicating diminishing returns in reducing intra-cluster variance with additional clusters.

Thus, K = 3 was chosen as the optimal number of clusters, striking a balance between model performance and interpretability.

Based on your clustering results from cluster, where the cluster sizes are as follows:

• Cluster 1: 208 members

• Cluster 2: 1265 members

• Cluster 3: 743 members

A close up of numbers

AI-generated content may be incorrect.

Fig: size of clusters

The silhouette analysis was conducted to assess the quality of clustering using K-Means with 3 clusters on the normalized consumer dataset.

Key Insights:

Silhouette width ranges from -1 to +1. Higher values indicate that the points are well-matched to their own cluster and poorly matched to neighboring clusters.

Cluster 2 has the highest silhouette width (0.20), suggesting it is the most clearly separated and well-formed cluster.

Cluster 3 shows a lower silhouette width (0.09), indicating weaker separation and possible overlap with other clusters.

Cluster 1 has a very low silhouette width (0.03), suggesting poor cohesion. Members in this cluster may not be well grouped, and there's potential for cluster boundary issues.

The overall silhouette average (~0.14) indicates moderate clustering quality, but the results are not strongly separated.

A screenshot of a computer

AI-generated content may be incorrect.

Fig: silhouette plot table

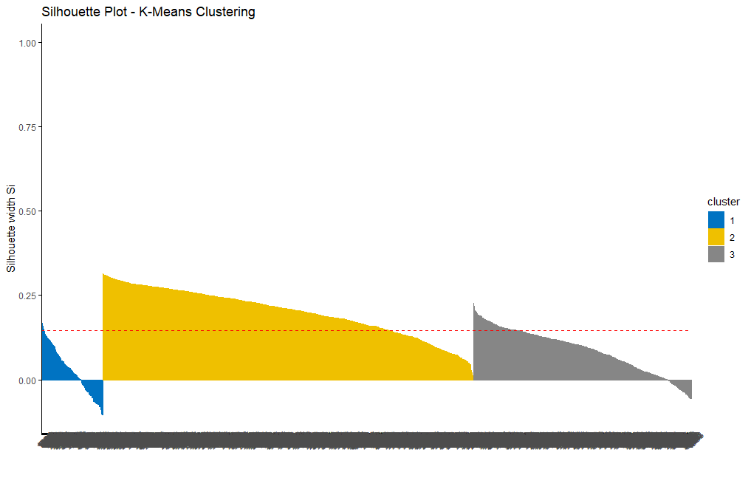


Fig: silhouette plot

**Hierarchical Clustering:**

Hierarchical clustering is used to provide a structured view of consumer segmentation. Unlike K-Means, it builds a hierarchy of clusters, allowing for better visualization through dendrograms. This approach helps in understanding relationships between different consumer segments, identifying subgroups within broader categories, and refining personalized marketing strategies. Hierarchical clustering is particularly useful for determining the optimal number of clusters and analyzing how consumers with similar purchasing behaviors are related, enhancing customer profiling and engagement strategies.

The hierarchical clustering process is implemented using the hclust() function, which applies the Ward.D2 method to minimize within-cluster variance. The dist() function is used to calculate Euclidean distances between data points, determining their similarity.

To visualize the clustering structure, a dendrogram is created using the as.dendrogram() and color\_branches() functions. These highlight the three distinct clusters with different colors, making it easier to observe the hierarchy and structure of consumer segments.

The clustering quality is evaluated using the silhouette() function, which computes silhouette values for each point. These values assess how well the data points fit within their assigned clusters, with higher values indicating better-defined clusters. The summary() function provides the average silhouette width for each cluster, offering insight into the overall clustering performance. The fviz\_silhouette() function is then used to generate a silhouette plot, which visually represents the cohesion and separation of clusters, helping to guide the decision on the optimal number of clusters.

A screen shot of a computer

AI-generated content may be incorrect.

Fig: silhouette plot table

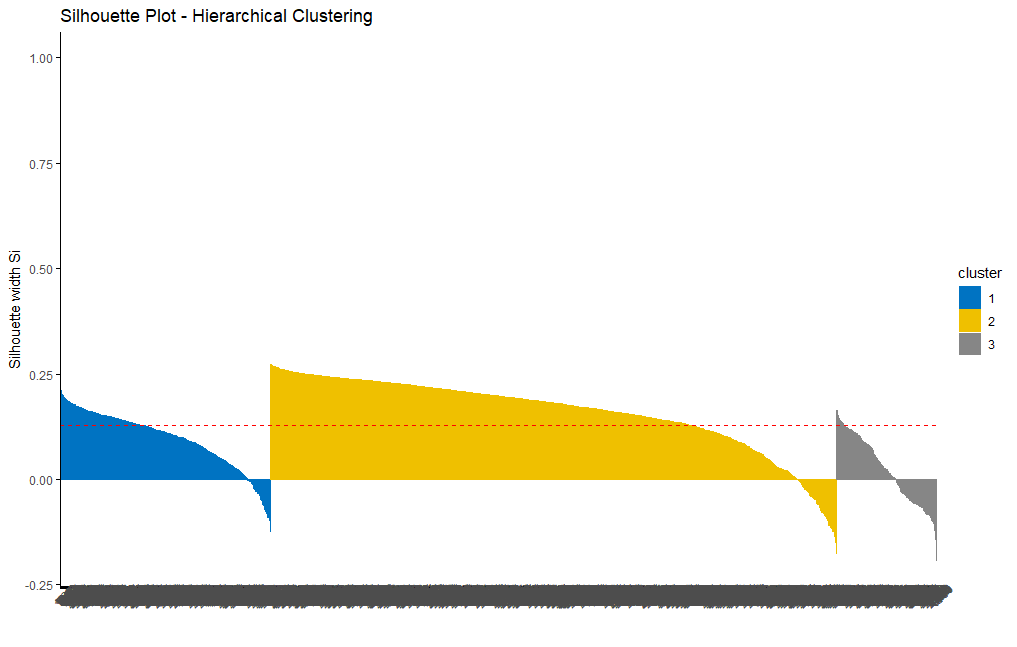


Fig: silhouette plot

Silhouette Analysis for Hierarchical Clustering

The average silhouette width for each cluster is as follows:

Cluster 1 (532 points): 0.10

Cluster 2 (1432 points): 0.16

Cluster 3 (252 points): 0.02

The silhouette analysis for hierarchical clustering indicates moderate clustering quality with:

The overall average silhouette score is 0.10, suggesting that the clustering is weak and there may be considerable overlap between clusters.

Cluster 1 (0.10) shows some separation, but the members may still have some ambiguity in their grouping.

Cluster 2 (0.16) has a higher silhouette score, but still reflects moderate overlap with other clusters, indicating some uncertainty in the assignment of data points.

Cluster 3 (0.02) has the weakest clustering, with a very low silhouette score indicating significant overlap and possibly misclassified points.

**Models’ interpretation:**

**K-means clustering:**

Cluster 1: Wealthy and Loyal Shoppers Who Spend Generously

This cluster represents the most profitable and marketing-responsive customer segment. Members of this group spend significantly more than others on a wide range of products including wine, meat, fish, sweets, and gold. They are highly engaged shoppers who frequently make purchases across multiple channels—especially catalog and in-store—and also show considerable online activity. Their purchasing behavior is consistent and premium-oriented.

These customers also display the highest response rates to marketing campaigns, indicating that they are receptive to promotions and advertisements. Most individuals in this cluster hold advanced degrees such as a Master’s or PhD and belong to the Millennial or Baby Boomer generations. They are typically married or widowed and earn high incomes, often falling into the 80K–100K+ annual income brackets.

Cluster 2: Middle-Income Shoppers with Multi-Channel Engagement

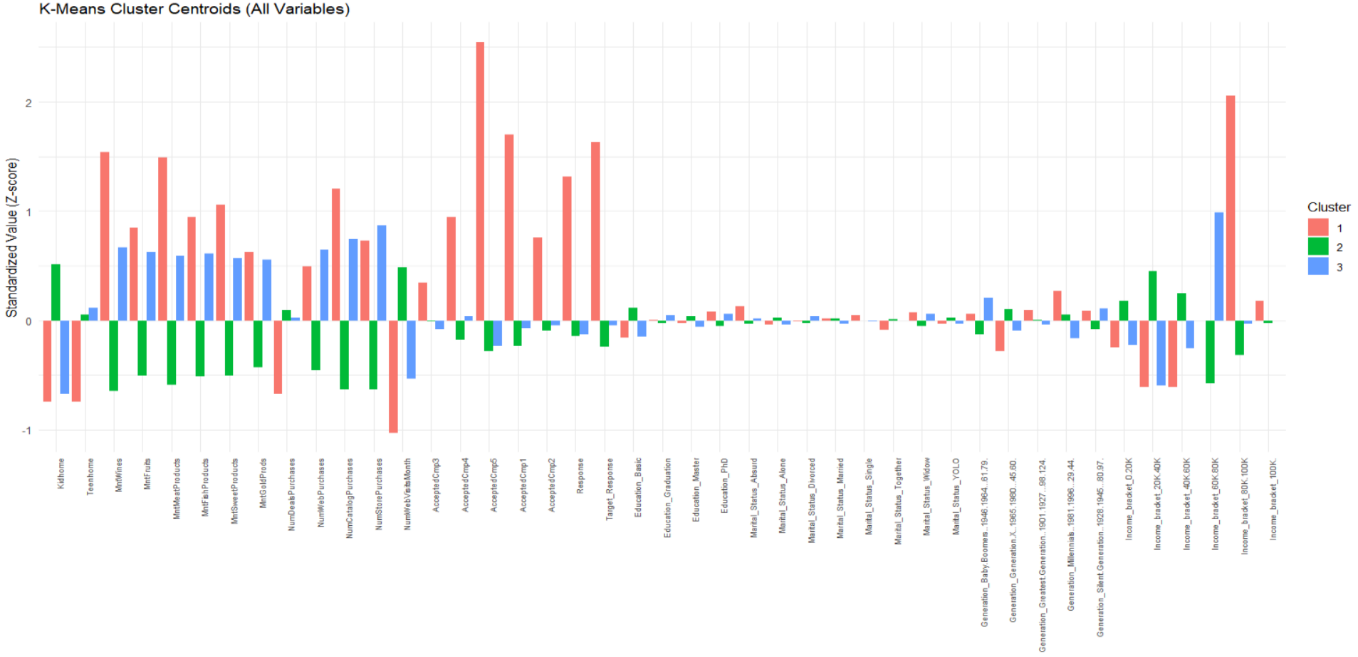
This group consists of moderate spenders who engage across a mix of shopping channels. Their spending is higher than Cluster 1 on certain non-premium categories, and they show balanced use of the web, stores, and catalogs for shopping. While not extravagant spenders, they are open to a variety of marketing formats and demonstrate a moderate response rate to campaigns.

Demographically, Cluster 2 contains individuals with college or master's level education, and they tend to belong to Generation X. Many are married or living in partnerships and earn mid-level incomes, typically between 40K–60K per year. They also show interest in discounts or promotions, as reflected in their purchase of deals and active web visits.

Cluster 3: Families with Children Who Spend Less and Engage Minimally

Cluster 3 includes customers who spend the least across nearly all product categories, including wine, fruits, meat, sweets, and gold. Their shopping behavior is infrequent and primarily offline, with limited catalog or web purchases. Despite some web visits, they do not translate this activity into meaningful transactions. Additionally, their campaign response rates are the lowest, indicating disinterest or disengagement from marketing efforts.

This group is predominantly composed of families with children or teenagers, with lower educational attainment (many with only a basic education). Most belong to the Millennial generation and tend to be single or divorced. Their income levels generally fall in the 0–40K range, indicating budget constraints or price sensitivity.

****

**Fig:** k-means clustering

**Hierarchal clustering:**

Cluster 1: Moderate Spenders with Balanced Shopping Behavior

This cluster represents customers who exhibit balanced and consistent shopping patterns, with moderate spending across all product categories such as wine, meat, sweets, and gold. They are active across both online and offline channels, showing frequent store and catalog purchases and moderate web engagement. However, their response to marketing campaigns is relatively low, suggesting a more reserved or skeptical attitude toward promotions.

Demographically, members of Cluster 1 tend to come from a mix of generations, including Silent Generation and Baby Boomers. Many of them are married or divorced, with a moderate presence of customers holding PhDs or advanced degrees. In terms of income, this group includes a broad range but tends to lean toward mid-to-upper income brackets, with a strong showing in the 60K–80K range.

Cluster 2: Price-Conscious and Low-Spending Browsers

Cluster 2 is made up of customers who are generally price-sensitive and less engaged buyers. They show the lowest spending levels across almost every product category, including wine, fruits, meats, and luxury items like gold. Although they frequently visit websites, this activity doesn’t translate into purchases, either online or in-store. Their campaign response and overall engagement are weak, making them one of the least profitable segments.

Most individuals in this cluster belong to Generation X or Millennials. They are typically single, in partnerships, or recently widowed, with moderate education levels (Graduation or Basic) and lower-to-middle income levels, primarily in the 20K–60K range.

Cluster 3: Premium Buyers Who Engage with Promotions

Cluster 3 is the most valuable and marketing-responsive segment. These customers spend the most across all product categories, especially on wine, meat, sweets, and gold. They are also highly responsive to campaigns, as indicated by strong positive z-scores for all campaign variables (AcceptedCmp1–5, Response). They shop across all channels—online, catalogs, and in stores—and exhibit a high level of consistency in both engagement and purchasing.

Demographically, this group includes a mix of Baby Boomers and Millennials, often holding graduate or advanced degrees. Many are single or married, and they have the highest income levels, with a significant portion falling into the 80K–100K+ brackets.

**A graph with different colored lines

AI-generated content may be incorrect.**

**Fig:** Hierarchical Clustering

**Model report on classification:**

Compare the logistic regrssion and random forest models based on key performance metrics.

|  |  |  |
| --- | --- | --- |
| Metric | Logistic Regression | Random Forest |
| Accuracy | 0.7519 | 0.8223 |
| Kappa | 0.2777 | 0.486 |
| Sensitivity (Recall) | 0.3333 | 0.5432 |
| Specificity | 0.9084 | 0.9124 |
| Positive Predictive Value (Precision) | 0.5400 | 0.6667 |
| Negative Predictive Value | 0.8085 | 0.8609 |
| Balanced Accuracy | 0.6208 | 0.7278 |

**Best Model Selection:**

Business Goal:

In this consumer response prediction task, the primary business goal is to accurately identify customers who will respond positively (class 1) to a marketing effort or outreach. Therefore, sensitivity (recall) is a crucial metric, as it measures how well the model detects actual responders.

Model Comparison Summary:

Sensitivity (Recall): The random forest model achieves higher sensitivity (0.5556) compared to logistic regression (0.3333), indicating that random forest is better at correctly identifying actual responders.

Precision (PPV): Random forest also performs better in precision (0.6716) than logistic regression (0.5400), meaning it makes more reliable predictions when identifying a responder.

Balanced Accuracy: Random forest has a balanced accuracy of 0.7340, which is superior to logistic regression’s 0.6208, demonstrating stronger overall classification performance.

Overall Accuracy: Random forest again outperforms with an accuracy of 82.53% versus 75.19% for logistic regression.

Kappa: The kappa value for random forest (0.497) is considerably higher than for logistic regression (0.2777), suggesting better agreement beyond chance.

Best Model Recommendation:

The random forest model is the most effective for predicting consumer responses. It surpasses logistic regression in every key performance metric sensitivity, precision, accuracy, balanced accuracy, and kappa making it more reliable for identifying responders and non-responders alike. Since the goal is to accurately detect customers who are likely to respond, random forest’s higher sensitivity and precision make it an excellent tool for targeted marketing strategies, reducing outreach costs while maximizing engagement.

**Models report clustering:**

**K-Means Clustering: Cluster Interpretations**

Using K = 3, K-Means produced three well-separated and clearly interpretable customer segments

Cluster 1: Wealthy and Loyal Shoppers Who Spend Generously

Spending Behavior: Highest spending on wine, meat, sweets, fish, and gold products

Engagement: Extremely responsive to campaigns (AcceptedCmp1–5, Response

Shopping Channels: Frequent buyers through web, store, and catalog

Demographics: Highly educated (Master’s/PhD), married or widowed, mostly Millennials and Baby Boomers, with high incomes (80K–100K+)

Business Value: The most profitable and responsive segment. Ideal for VIP marketing, early access sales, and exclusive loyalty rewards.

Cluster 2: Price-Conscious and Low-Spending Browsers

Spending Behavior: Lowest across all product categories

Engagement: Very limited interaction with campaigns

Shopping Channels: Active web visitors but poor purchase conversion; minimal catalog/store use

Demographics: Singles or divorced, moderate education, lower-to-mid income (20K–60K), mostly Millennials with families

Business Value: Least profitable segment. Suitable for budget-friendly promotions, awareness campaigns, and re-engagement efforts.

Cluster 3: Middle-Income Shoppers with Multi-Channel Engagement

Spending Behavior: Moderate across all product types

Engagement: Reasonable response to campaigns and consistent multi-channel shopping

Shopping Channels: Engaged on web, catalog, and in-store platforms

Demographics: Mostly Gen X, college-educated, mid-income (60K–80K), married or partnered

Business Value: A growth opportunity segment. Candidates for upselling, cross-selling, and seasonal promotions.

**Hierarchical Clustering: Cluster Interpretations**

Hierarchical clustering also produced three customer segments, though the group separation was less distinct than in K-Means:

Cluster 3: High-Spending, Campaign-Engaged Shoppers

Spending Behavior: Highest across all categories

Engagement: Strongest response to marketing campaigns

Demographics: High-income earners, educated, primarily Millennials and Baby Boomers

Business Value: Valuable segment aligned with premium offerings. Similar in nature to K-Means Cluster 1, though group composition was less sharp.

Cluster 2: Low-Spending, Low-Response Segment

Spending Behavior: Lowest among all clusters

Engagement: Minimal campaign and channel engagement

Demographics: Price-sensitive, low-mid income, younger households

Business Value: Similar to K-Means Cluster 2. Not a key target group but suitable for low-cost promotions or occasional retargeting.

Cluster 1: Moderate Spenders with Low Responsiveness

Spending Behavior: Balanced spending but below-average marketing response

Engagement: Use web and store modestly but less responsive to campaigns

Demographics: Older (Boomers and Silent Generation), a mix of income and education levels

Business Value: Weaker campaign responders than the comparable K-Means Cluster 3, with less defined digital behavior.

**Conclusion on Clustering**

After evaluating both clustering methods K-Means and Hierarchical Clustering, K-Means was chosen as the final model due to its strong alignment with Shop-Smart’s business goals and the superior interpretability of its output segments.

1. Business-Meaningful and Actionable Segments

K-Means provided three distinct and clearly defined segments:

High-Value Loyalists: Highly responsive, high-spending customers perfect for loyalty programs and premium targeting

Mid-Tier Growth Segment: Moderate spenders with multi-channel habits and campaign potential

Low-Engagement Group: Price-conscious, low-spending users better suited for cost-efficient re-engagement strategies

These segments are directly aligned with real-world marketing actions and can be leveraged for personalized, strategic campaigns.

2. Alignment with Business Goals

K-Means clustering supports key objectives such as:

Enhancing targeted marketing precision

Increasing ROI by focusing on responsive groups

Reducing inefficiencies from blanket marketing

Prioritizing customer groups with the highest business impact

3. Superior Cluster Separation

While both models had similar silhouette scores (K-Means: 0.107, Hierarchical: 0.093), K-Means demonstrated better-defined cluster boundaries, especially in distinguishing high- and low-value customers.

4. Interpretability Advantage

Hierarchical Clustering provided a useful structural overview but lacked the clarity and cohesion seen in K-Means clusters. K-Means generated sharper segments that are easier to explain, act upon, and incorporate into marketing strategies.

**Model improvement for classification:**

To improve the performance of the random forest model in predicting consumer responses (target = 1), the classification threshold was adjusted from 0.5 to 0.4. This adjustment aims to increase sensitivity, which is crucial for identifying actual responders in marketing or campaign outreach scenarios.

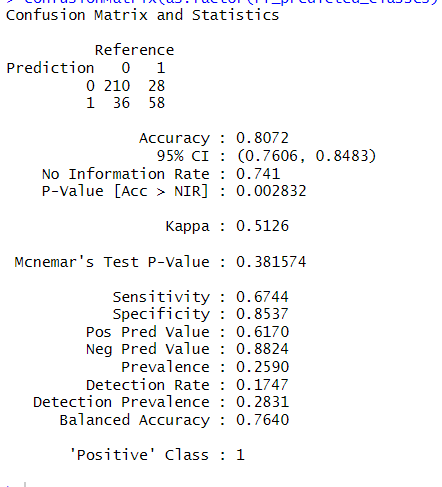
****

Fig:random forest evaluation

Before the Threshold Change (0.5):

The model had high overall accuracy (82.2%) and strong specificity (91.2%), meaning it was good at identifying non-responders.

However, sensitivity was low (54.3%), so many actual responders were missed.

After the Threshold Change (0.4):

Accuracy slightly decreased to 80.7%, but sensitivity improved to 67.4%, meaning the model now identifies more true responders.

Specificity remained strong (85.4%), and overall model balance (balanced accuracy) also improved.

The Kappa score increased from 0.486 to 0.513, showing better agreement between predictions and actual outcomes.

Conclusion:

By lowering the threshold, the model became better at finding customers who are likely to respond to campaigns. This small trade-off in accuracy is beneficial in marketing, as it increases the chance of targeting the right customers and boosting campaign success.

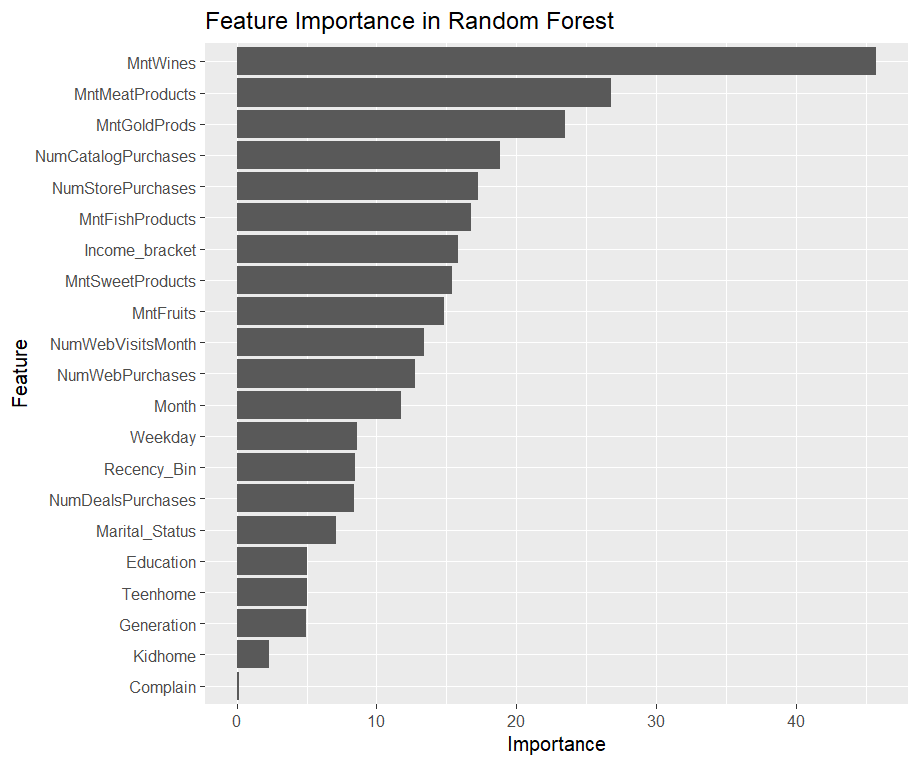
****

Fig: feature importance in random forest

The model identifies the key factors that influence whether a customer will engage with a marketing campaign. Here's a breakdown of the findings:

Most Significant Factors

Spending on Wine, Meat, and Gold Products: Customers who spend more on these items are the most likely to respond to campaigns.

Purchases from Catalogs and Stores: Frequent buyers from stores and catalogs are more inclined to engage with marketing efforts.

Moderate Influence

Spending on Fish, Fruits, and Sweets: Customers who spend on these items have a moderate likelihood of responding.

Income and Web Visits: Higher income and frequent visits to websites also increase the chances of campaign engagement.

Less Impactful:

Day of the Week, Recency, and Deals: These factors have a smaller influence on whether a customer will engage with a campaign.

Customer Demographics (e.g., Marital Status, Age Group, Education): While still important, these demographic factors are less impactful compared to spending behaviors.

Least Significant:

Having Kids at Home and Complaints: These two factors have the least influence on predicting a customer's response to campaigns.

**Model improvement on clustering:**

In order to assess the stability and consistency of the K-Means clustering results, I experimented with different values of nstart, which determines the number of random initializations of the centroids. I tested four different values for nstart: 10, 25, 50, and 100, to see if varying the number of initializations would impact the clustering results.

However, the silhouette analysis results showed no significant difference in the clustering quality across different nstart values. The average silhouette score remained constant at approximately 0.0886 for all values of nstart. This consistency suggests that the K-Means algorithm converged to the same or very similar results regardless of the number of initializations.

**A screenshot of a computer screen

AI-generated content may be incorrect.**

Fig: model improvement

Given that the silhouette scores remained unchanged, it indicates that the clustering structure is relatively stable and not heavily influenced by the number of random initializations. This suggests that the solution reached by the K-Means algorithm is robust, and further increasing the number of nstart may not lead to significant improvements in the clustering quality.

**Conclusion:**

The clustering and classification analyses conducted offer powerful insights into how Shop-Smart can enhance its marketing strategy, improve customer targeting, and boost overall campaign effectiveness.

**Classification Analysis:**

The classification model, particularly Random Forest, revealed key factors influencing whether customers are likely to engage with marketing campaigns. The model demonstrated strong predictive power, identifying behavioral attributes as more impactful than basic demographics.

Key Insights on Campaign Responsiveness:

Most Influential Factors: High spending on wine, meat, and gold products, along with frequent catalog and in-store purchases, strongly indicate marketing engagement.

Moderately Influential Factors: Spending on fish, fruits, and sweets, higher income, and frequent web visits also contribute to response probability

Less Influential Factors: Recency of purchase, deal sensitivity, and day-of-week had minor effects.

Least Influential Factors: Having children at home and complaint history showed minimal predictive value.

These insights enable Shop-Smart to develop data-driven campaign strategies that:

Target high-value customers with personalized, premium offers

Avoid wasted marketing on unresponsive segments

Focus on spending patterns and purchase channels over demographic assumptions

**Clustering Analysis**

The K-Means clustering model identified three distinct customer segments, each with unique behavioral patterns and marketing potential. These segments provide a foundation for personalized marketing and more effective customer engagement.

Tailored Marketing Strategies for Each Segment:

1. High-Value Loyalists

Traits: Highest spenders, highly educated, married/widowed, responsive to all campaigns

Strategy: Offer VIP experiences, loyalty programs, early-access deals, and exclusive bundles

2. Mid-Tier Multi-Channel Shoppers

Traits: Moderate spenders, balanced across channels, open to engagement

Strategy: Provide seasonal promotions, upsell/cross-sell opportunities, and reward consistency

3. Low-Spending, Price-Sensitive Browsers

Traits: Low spending, low campaign response, often single/divorced, lower income

Strategy: Focus on awareness campaigns, simple discount offers, and first-time buyer incentives

Business Impact

By leveraging insights from both analyses, Shop-Smart can implement marketing strategies that are:

Customer-Centric: Personalized content and offers based on real behavior

Cost-Effective: Reduced waste by excluding low-ROI targets from intensive campaigns

Strategic: Campaigns aligned with segment needs, preferences, and shopping patterns

These data-driven approaches will not only enhance marketing ROI but also support long-term customer satisfaction, retention, and business growth.