Webster UNIVERSITY

George Herbert Walker School of Business & Technology

Data Analytics Project

COSC 6010 DATA ANALYTICS PRACTICUM – SPRING 2025

Case 1: Auto Rental Cancellation Analytics

Case 2: Fitness Center Member Wellness Analytics

Case 3: Consumer Behavior Analytics

Developed by: Mithila Papi Shetty

Course Instructor: Dr. Ali Ovlia

Table of Contents

Auto Rental Cancellation Analytics

Table of Contents

Introduction

San Francisco Auto Rental (SAR) is a transportation company offering rental services across San Francisco. Despite implementing online and mobile booking systems, SAR continues to face challenges with high ride cancellations. By analyzing customer booking and trip data, the company seeks to identify patterns behind cancellations and improve service reliability to enhance customer satisfaction.

Business and Analytic Goals

Business opportunity

San Francisco Auto Rental (SAR) wants to enhance service reliability and customer satisfaction by addressing ride cancellations. By leveraging customer booking and trip data, SAR seeks to improve operational efficiency, strengthen customer loyalty, and drive business growth.

Business Goal

The goal is to improve San Francisco Auto Rental's (SAR) service reliability by reducing ride cancellations using data-driven insights.

Business Objectives:

• Analyze customer booking and trip data to uncover patterns associated with ride cancellations.
• Identify key factors influencing cancellations across different booking channels and travel types.
• Recommend operational and service strategies to minimize cancellations and enhance customer satisfaction.

Analytical Goal

To develop models that uncover patterns behind ride cancellations and predict the likelihood of cancellations based on customer and trip attributes.

Analytical Objectives:

• Perform exploratory data analysis to identify key factors influencing ride cancellations.
• Develop classification models to predict ride cancellation likelihood.
• Evaluate model performance to support decision-making for improving service reliability.

Analytical Approach

The dataset was first explored to understand the structure and quality of the data. Necessary preprocessing steps were applied to clean and prepare the data, including handling missing values and correcting date formats. Exploratory analysis was performed to identify important factors influencing ride cancellations. Feature selection techniques were used to retain the most relevant attributes for modeling. The data was then partitioned into training, validation, and test sets, and class imbalance was addressed to ensure fair modeling. Classification models were developed to predict ride cancellations, and model performance was evaluated using appropriate metrics to support reliable decision-making.

Data Exploration and Preprocessing

Data Understanding

Data collection:

- The SAR dataset consists of 10,000 records and contains 19 variables.
- Dataset captures detailed information about user, vehicle, booking, trip and location as shown in Table 1.

User and Vehicle Information:

- 'user_id', 'vehicle_model_id'

Booking and Trip Details:

- ‘travel_type_id’, ‘package_id’, ‘from_date’, ‘to_date’, ‘booking_created’, ‘online_booking’, ‘mobile_site_booking’, ‘car_cancellation’.

Location Information:

- ‘from_area_id’, ‘to_area_id’, ‘from_city_id’, ‘to_city_id’, ‘from_lat’, ‘from_long’, ‘to_lat’, ‘to_long’.

row.	user_id	vehicle_model_id	package_id	travel_type_id	from_area_id	to_area_id	from_city_id	to_city_id	from_date
1	17712	12	NA	2	1021	1323	NA	NA	1/1/2013 22:33
2	17037	12	NA	2	455	1330	NA	NA	1/1/2013 12:43
3	761	12	NA	2	814	393	NA	NA	1/2/2013 0:28
4	868	12	NA	2	297	212	NA	NA	1/1/2013 13:12
5	21716	28	NA	2	1237	330	NA	NA	1/1/2013 16:33
6	38966	12	NA	2	61	393	NA	NA	1/1/2013 18:00

to_date	online_booking	mobile_site_booking	booking_created	from_lat	from_long	to_lat	to_long	Car_Cancellation
	0	0	1/1/2013 8:01	13.02853	77.54625	12.86980	77.65321	0
	0	0	1/1/2013 9:59	12.99987	77.67812	12.95343	77.70651	0
1/3/2013 0:00	1	0	1/1/2013 12:14	12.90899	77.68890	13.19956	77.70688	0
	0	0	1/1/2013 12:42	12.99789	77.61488	12.99474	77.60797	0
	0	0	1/1/2013 15:07	12.92645	77.61206	12.85883	77.58913	0
	0	0	1/1/2013 15:11	12.96298	77.71229	13.19956	77.70688	0

Table1: First six records of SAR dataset

Variables Definition

The Table 2 lists every column in the SAR dataset, specifying its data type and clarifying what each field represents.

Variable Name	Data Type	Definition
row.	int	Sequential row identifier for each record.
user_id	int	Unique Identifier for each user which may repeat across multiple bookings
vehicle_model_id	int	Code representing the booked vehicle and its associated driver
package_id	int	Representing type of travel package: 1= long distance, 2= point to point, 3= hourly rental
travel_type_id	int	Representing type of travel: 1 = Long distance, 2 = Point-to-point, 3 = Hourly rental.
from_area_id	int	Representing pickup area ID, applicable only for point-to-point travel
to_area_id	int	Representing drop-off area ID, applicable only for point-to-point travel
from_city_id	int	Identifier for the city of trip origin
to_city_id	int	Identifier for the city of trip destination
from_date	chr	Date and time the trip was scheduled to start
to_date	chr	Date and time the trip ended
online_booking	int	Binary variable indicating mobile booking (1 = Yes, 0 = No).
mobile_site_booking	int	Binary variable indicating website booking (1 = Yes, 0 = No).
booking_created	chr	Timestamp when the booking was made
from_lat	num	Latitude coordinate of the pickup location
from_long	num	Longitude coordinate of the pickup location
to_lat	num	Latitude coordinate of the drop-off location
to_long	num	Longitude coordinate of the drop-off location
car_cancellation	int	Binary target variable indicating if the trip was cancelled (1 = Yes, 0 = No).

Table 2: SAR Dataset Variable Definitions

Data Types:
All the variables are integer, numeric or character type. However, these are categorized as follows:
Numeric Data:

- 'row.', 'user_id', 'vehicle_model_id', 'package_id', 'travel_type_id', 'from_area_id', 'to_area_id', 'from_city_id', 'to_city_id', 'from_lat', 'from_long', 'to_lat', 'to_long', 'online_booking', 'mobile_site_booking', 'Car_Cancellation'

Categorical Data:

- 'from_date', 'to_date', 'booking_created'

```
'data.frame':   10000 obs. of  19 variables:
 $ row.               : int  1 2 3 4 5 6 7 8 9 10 ...
 $ user_id            : int  17712 17037 761 868 21716 38966 22196 22200 222
 $ vehicle_model_id   : int  12 12 12 12 28 12 12 12 12 12 ...
 $ package_id         : int  NA NA NA NA NA NA NA NA 1 NA ...
 $ travel_type_id     : int  2 2 2 2 2 2 2 2 3 2 ...
 $ from_area_id       : int  1021 455 814 297 1237 61 409 1371 1323 1017 ...
 $ to_area_id         : int  1323 1330 393 212 330 393 1194 839 NA 496 ...
 $ from_city_id       : int  NA NA NA NA NA NA NA NA NA NA ...
 $ to_city_id         : int  NA NA NA NA NA NA NA NA NA NA ...
 $ from_date          : chr  "1/1/2013 22:33" "1/1/2013 12:43" "1/2/2013 0:2
12" ...
 $ to_date            : chr  "" "" "1/3/2013 0:00" "" ...
 $ online_booking     : int  0 0 1 0 0 0 0 1 1 0 ...
 $ mobile_site_booking: int  0 0 0 0 0 0 0 0 0 0 ...
 $ booking_created    : chr  "1/1/2013 8:01" "1/1/2013 9:59" "1/1/2013 12:14
2" ...
 $ from_lat           : num  13 13 12.9 13 12.9 ...
 $ from_long          : num  77.5 77.7 77.7 77.6 77.6 ...
 $ to_lat             : num  12.9 13 13.2 13 12.9 ...
 $ to_long            : num  77.7 77.7 77.7 77.6 77.6 ...
 $ Car_Cancellation   : int  0 0 0 0 0 0 0 0 0 0 ...
```

Figure 1: Structure of the SAR data

Data Preprocessing

As a part of preprocessing, data is checked for duplicates, data types, missing and zero values, and handled wherever necessary.

Duplicates Records Check:

The dataset was checked for duplicate rows, and no duplicate records were found (0 duplicates), ensuring each booking entry is unique.

Data type conversion:

The dataset initially contained variables in basic types such as integers and character strings. To prepare the data for analysis, categorical variables were converted into factors to properly handle classification tasks, and date fields were accurately converted into date formats to enable time-based feature engineering and analysis.

Variable Name	Original Data Type	Recommended Data Type	Reason
user_id	int	factor (categorical)	Used to uniquely identify individuals and no arithmetic operations required
vehicle_model_id	int	factor (categorical)	Represents vehicle type categories (like 1, 2 ,3)
package_id	int	factor (categorical)	Identifies a service package category (like 1,2, 3)
travel_type_id	int	factor (categorical)	Contains categories of travel type (1, 2, 3)
from_area_id	int	factor (categorical)	Used to identify an area, not quantity
to_area_id	int	factor (categorical)	Used to identify an area, not quantity
from_city_id	int	factor (categorical)	Used to identify a city, not quantity
to_city_id	int	factor (categorical)	Used to identify a city, not quantity
from_date	chr	datetime	Representing date and time of trip end
to_date	chr	datetime	Representing date and time of trip start
online_booking	int	factor (categorical)	Binary variable having two categories (1= yes, 0 = no)
mobile_site_booking	int	factor (categorical)	Binary variable having two categories (1= yes, 0 = no)
booking_created	chr	datetime	Representing date and time of trip booked
from_lat	num	numeric	No Change (contains a continuous value)
from_long	num	numeric	No Change (contains a continuous value)
to_lat	num	numeric	No Change (contains a continuous value)
to_long	num	numeric	No Change (contains a continuous value)
car_cancellation	int	factor (categorical)	Binary variable having two categories (1= yes, 0 = no)

Table 3: Recommended Data Type Conversion for SAR Dataset

Handling Missing and Zero values:

Missing values:

As shown in the table 4, most columns in the SAR dataset have no missing values. However, variables like ‘package_id’, ‘from_city_id’, ‘to_city_id’, ‘from_area_id’, ‘to_area_id’, ‘from_lat’, ‘from_long’, ‘to_lat’, and ‘to_long’ exhibit missing values to varying extents. Columns like ‘package_id’, ‘from_city_id’, and ‘to_city_id’ have more than 50% missing values.

Missing Values Handling:

To preserve data integrity and avoid introducing artificial values, all rows with missing entries in important variables like ‘from_area_id’, ‘to_area_id’, ‘from_lat’, ‘from_long’, ‘to_lat’, and ‘to_long’ were removed during preprocessing. Additionally, columns such as ‘package_id’, ‘from_city_id’, and ‘to_city_id’ with more than 50% missing values were dropped to ensure robustness of the analysis.

Zero Values Handling:

- The presence of zero values across variables offers meaningful insights about the trip booking and cancellation process.
- In 'online_booking' and 'mobile_site_booking', zero indicates that the booking was not made through the respective channel.
- A zero in 'car_cancellation' signifies that the ride was not cancelled.

Variable	Missing	Zero
row.	0	0
user_id	0	0
vehicle_model_id	0	0
package_id	8248	0
travel_type_id	0	0
from_area_id	15	0
to_area_id	2091	0
from_city_id	6294	0
to_city_id	9661	0
from_date	0	0
to_date	0	0
online_booking	0	6467
mobile_site_booking	0	9576
booking_created	0	0
from_lat	15	0
from_long	15	0
to_lat	2091	0
to_long	2091	0
Car_Cancellation	0	9257

Table 4: Missing and Zero Value Counts for Each Variable

Variable Distributions:

Numeric Variables:

- Pick-up latitudes ('from_lat') and longitudes ('from_long') are centered around 12.9 and 77.6, respectively, with modest spread—indicating most rides start within one metropolitan zone, though pickups are still scattered.
- Drop-off latitudes ('to_lat') display a bimodal shape: the main city band near 12.9 plus a sharp spike near 13.2, hinting at a popular northern destination.
- Drop-off longitudes ('to_long') echo this: a broad city-range plus a pronounced peak near 77.7, reinforcing the presence of a single dominant drop-off point.

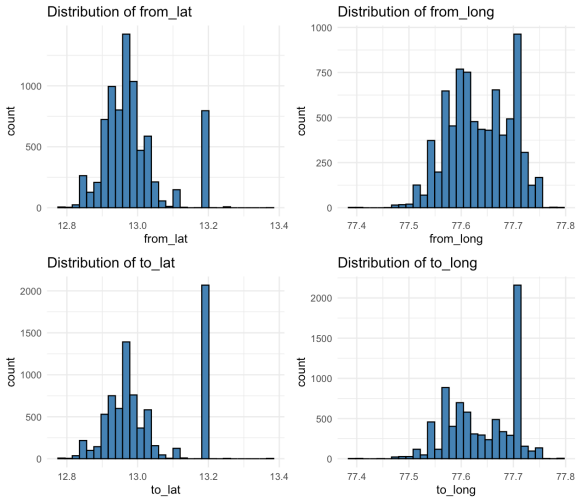


Figure 2: Numeric Variable Distribution in SAR Dataset

Categorical Variables:

Figure 3 display the distribution of key categorical variables in the SAR dataset:

- Online Booking shows a higher number of bookings made offline compared to online.
- Mobile Site Booking indicates that a large majority of bookings were not made via mobile site.
- Car Cancellation reveals an imbalance, with most bookings not canceled.
- Travel Type ID distribution suggests that a specific travel type dominates the dataset.

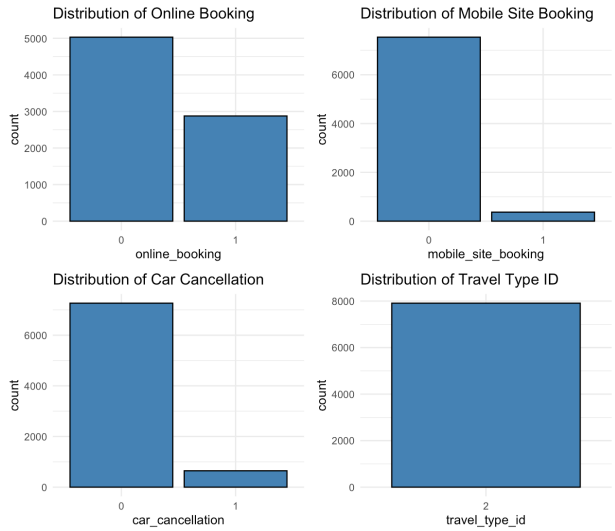


Figure 3: Distribution of key Categorical Variables in SAR dataset

From and to area IDs:

From Figure 4, the plots show the top 10 most frequent pickup and drop-off areas in the SAR dataset:

- 'From Area ID 393' is the most common starting point for trips, followed by areas like 571 and 585.
- 'To Area ID 393' is also the most frequent destination, indicating a major hub for both trip origins and destinations.
- This concentration around a few areas suggests a strong geographical pattern in trip activity.

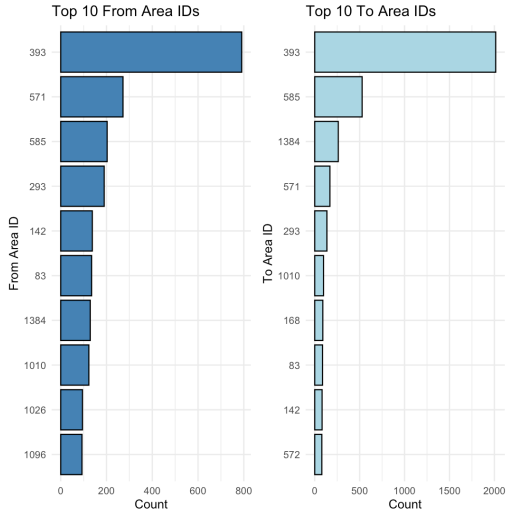


Figure 4: Top 10 From and To Area IDs of Trip

Vehicle model ID

Figure 5 illustrates the distribution of different vehicle models used in the SAR dataset.

- Vehicle Model ID 12 dominates the fleet, accounting for many bookings.
- Other vehicle models are relatively rare, each contributing a much smaller portion to the overall trips.
- This skew suggests a heavy operational reliance on a particular type of vehicle.

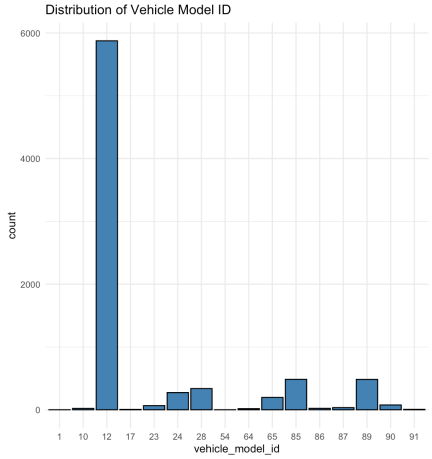


Figure 5: Distribution of Vehicle model ID

Feature Engineering and Transformation

Feature Creation:

New features were engineered to better capture ride characteristics and booking behavior.
The following transformations and feature creations were applied to enrich the SAR dataset and prepare it for modeling:

1. **Trip Length:**
   Calculated using pickup and drop-off GPS coordinates ('to_lat', 'to_long', 'from_lat', 'from_long') via the Haversine formula to capture the actual distance of the trip.

2. **Phone Booking:**
   Created a new binary variable 'phone_booking' to identify trips that were booked neither via online website nor mobile application.

3. **from_weekday**:
   Extracted the day of the week from the trip start date ('from_date') to understand which weekdays trips are more likely to start.

4. **from_month:**
   Extracted the month from the trip start date ('from_date') to analyze the distribution of trips across different months.

5. **booking_weekday:**
   Extracted the day of the week from the booking creation date ('booking_created') to explore booking behavior based on weekdays.

6. **booking_month:**
   Extracted the month from the booking creation date ('booking_month') to examine how bookings vary across different months.

7. **trip_length_grouping:**
   Binned 'trip_length' into three meaningful categories (Short, Medium, Long) to simplify interpretation and improve model performance.

Distribution of Engineered Features:

All the newly created variables are plotted to understand their distributions

- **Distribution of Trip Length:**
  The histogram from Figure 6 displays the frequency distribution of trip lengths in kilometers. Most trips are short, clustering between 5 km and 25 km.

- **Phone Booking Distribution:**
  The bar chart shows the split between bookings made through phone versus other platforms. Most of the trips were booked via phone.

- **Trips by Weekday (Trip Start):**
  This bar plot illustrates how trip start dates are distributed across different weekdays, with relatively uniform activity, slightly higher on Fridays.

- **Trips by Month (Trip Start):**
  This bar plot shows the distribution of trips started across different months, highlighting peaks in August, July, and October.

- **Bookings by Weekday:**
  The distribution of booking creation across weekdays is shown, with Fridays again having a slightly higher count compared to other days.

- **Bookings by Month:**
  The monthly distribution of booking creation dates is displayed, with August and July having comparatively higher booking activity.

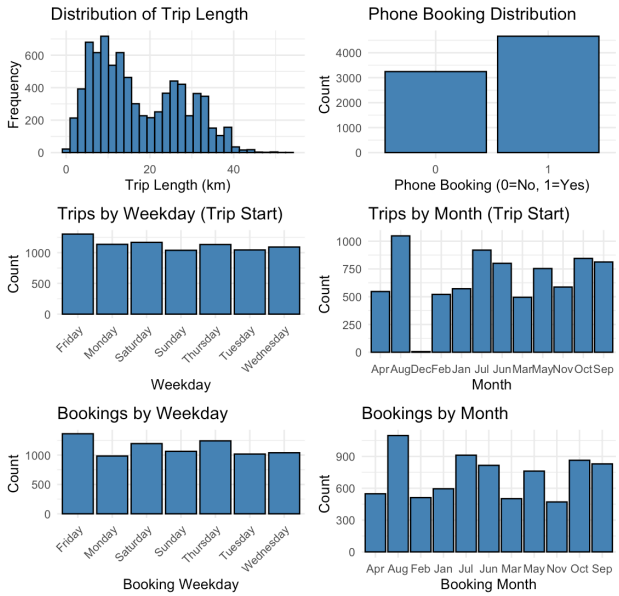


Figure 6: Distribution of Trip Length, Phone Booking, and Temporal Variables (Weekday and Month)

Checking for Missing Values after Transformation:

The missing value counts in Figure 7 representing there are no missing value after adding new features and could be used for further analysis.

```
> colSums(is.na(car_data))
                row.             user_id    vehicle_model_id      travel_type_id
                  0                   0                   0                   0
      from_area_id         to_area_id           from_date      online_booking
                  0                   0                   0                   0
mobile_site_booking     booking_created           from_lat          from_long
                  0                   0                   0                   0
             to_lat             to_long    car_cancellation        trip_length
                  0                   0                   0                   0
      phone_booking        from_weekday          from_month     booking_weekday
                  0                   0                   0                   0
      booking_month   trip_length_group
                  0                   0
```

Figure 7: Missing Values after Transformation

Predictor Relevancy

The influence of every numeric and categorical feature on car cancellation is assessed to find out the strongest predictor.

Numeric Predictors:

- **Short trips are riskier:** Cancelled rides cluster in the 5-15 km band, with a median distance well below that of completed trips, indicating a higher cancellation propensity for shorter journeys.
- **Long trips rarely drop off:** Completed rides extend beyond 40 km, whereas cancellations above ~25 km appear only as a few outliers—suggesting commitment increases sharply with trip length.

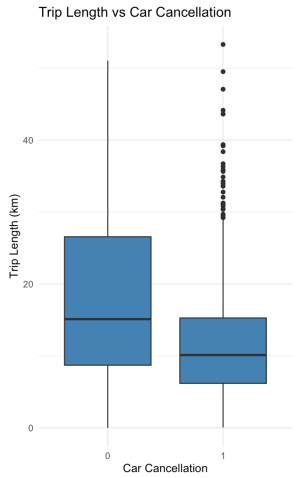


Figure 8: Boxplot of Trip Length vs Car cancellation

Categorical Predictors:

- **Online Booking vs Car Cancellation`**
  Trips booked online have a moderately higher cancellation rate compared to offline bookings. Although online platforms make booking convenient, they might also allow easier last-minute cancellations by users.
- **Mobile Site Booking vs Car Cancellation**
  Trips booked through the mobile site are slightly more prone to cancellation. This may indicate that mobile users might make quicker, less committed decisions when booking rides.
- **Phone Booking vs Car Cancellation**
  Phone bookings tend to experience fewer cancellations. Customers booking via phone may be more serious or committed to their travel plans, resulting in higher ride completions.
- **From Weekday vs Car Cancellation**
  Trips starting on Fridays and Sundays experience a higher frequency of cancellations. This could reflect weekend-related last-minute plan changes, leisure trips, or unpredictable schedules.
- **From Month vs Car Cancellation**
  Trips planned during May and October show increased cancellation rates. These months could coincide with specific seasonal events, vacations, or weather patterns causing higher ride uncertainty.
- **Booking Weekday vs Car Cancellation**
  Bookings made on Fridays and Thursdays are more prone to cancellation. This suggests that end-of-week plans may be more volatile compared to early-week bookings.
- **Booking Month vs Car Cancellation**
  Trips booked in May and October show a higher likelihood of cancellation. Seasonal activities, holidays, or shifting travel preferences during these months might influence booking behavior.

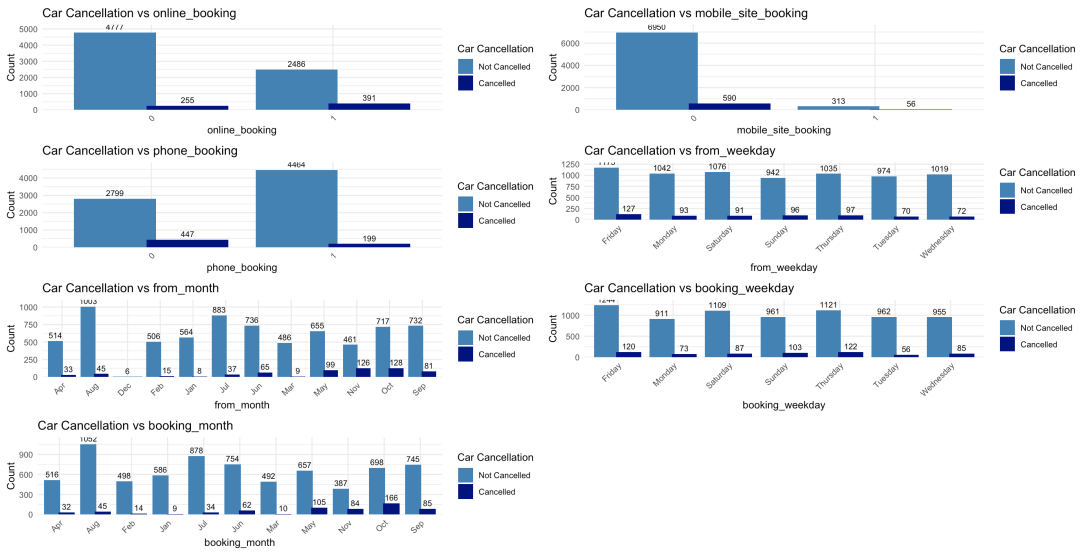


Figure 9: Proportion of Car Cancellations Across Different Predictors

Dimension Reduction

To optimize the dataset for modeling, several features were removed based on redundancy, high missingness, lack of variability, or derivation into new features, as detailed below.

Feature Removed	Reason of Removal
‘from_lat', 'from_long', 'to_lat', to_long'	Information from GPS coordinates was captured using the engineered 'trip_length' feature, making raw latitude and longitude values redundant
‘package_id', 'to_date', 'from_city_id', 'to_city_id'	These columns had more than 50% missing values, and thus were dropped to maintain data quality.
‘travel_type_id'	After preprocessing, only travel type of '2' remained, leading to no variability and contribution to model learning.
‘row.'	It was a sequential identifier irrelevant for prediction
‘from_area_id', 'to_area_id'	Thes were location-based identifiers, whose information was already captured through 'trip_length' and engineered time-based features.
‘booking_created', 'from_date'	Redundant after deriving time-based variables (weekday and month variables)

Table 5: Features Eliminated in SAR with Justifications

Feature Selection

To ensure that only the most relevant predictors were included for model building, the Boruta algorithm was applied. Boruta was chosen as it provides an all-relevant feature selection, systematically identifying and confirming important variables rather than just selecting a minimal optimal set.
Initially, the Boruta execution yielded:

- 10 features confirmed as important: ‘booking_created’, ‘booking_month’, ‘booking_weekday’, from_date, from_month, and 5 more.
- 1 feature (mobile_site_booking) confirmed as unimportant.
- 1 feature (from_weekday) marked as tentative.

After applying TentativeRoughFix, the tentative attribute ‘from_weekday’ was upgraded to confirmed important.

- Thus, the final confirmed important features were:
  ‘vehicle_model_id’, ‘online_booking’, ‘phone_booking’, ‘from_weekday’, ‘from_month’, ‘booking_weekday’, ‘booking_month’, and ‘trip_length_group’.

The only attribute rejected by Boruta was ‘mobile_site_booking’.

	meanImp	medianImp	minImp	maxImp	normHits	decision
vehicle_model_id	12.949303	12.693237	10.584388	17.685659	1.0000000	Confirmed
online_booking	21.697391	20.348222	17.413670	25.833581	1.0000000	Confirmed
mobile_site_booking	-1.254138	-1.209652	-5.398171	2.195112	0.0000000	Rejected
phone_booking	33.624829	28.632328	25.920405	43.491627	1.0000000	Confirmed
from_weekday	5.450336	5.485406	2.701609	7.993444	1.0000000	Confirmed
from_month	23.049800	22.646875	20.816068	27.327385	1.0000000	Confirmed
booking_weekday	3.944739	3.997287	1.040270	6.660632	0.8421053	Confirmed
booking_month	27.902543	27.550599	25.336067	30.662413	1.0000000	Confirmed
trip_length_group	31.648145	31.607934	28.287919	34.348860	1.0000000	Confirmed

Figure 10: Final Feature Importance Scores from Boruta Algorithm for Car Cancellation Prediction

Data Partitioning

A 70-15-15 split was followed to allocate 70% data for training, and 15% each for validation and testing, ensuring balanced model development, tuning, and evaluation.
The partition resulted in 5538 training, 1186 validation, and 1185 test records.

```
> cat("Training Set Records:", nrow(train_data), "\n")
Training Set Records: 5538
> cat("Validation Set Records:", nrow(validation_data), "\n")
Validation Set Records: 1186
> cat("Test Set Records:", nrow(test_data), "\n")
Test Set Records: 1185
```

Figure 11: Record Distribution Across Training, Validation, and Test Sets

Data Oversampling for Classification

As shown in Figure 12, the SAR dataset initially exhibited a significant class imbalance, with 91.8% of records corresponding to non-cancellations and only 8.2% representing cancellations. This imbalance necessitated balancing the data before model training to ensure fair and unbiased learning. Figure 13 demonstrates the class distribution after applying oversampling technique (ROSE method) on the training dataset. The classes were successfully balanced to 50% each for canceled and non-canceled trips, ensuring the model receives equal representation of both outcomes during training.

```
        0         1
0.9183209 0.0816791
```

Figure 12: Car Cancellation Class Distribution Before Oversampling

```
        0         1
0.4976526 0.5023474
```

Figure 13: Car Cancellation Class Distribution After Oversampling

Model Selection

Naive Bayes was selected as the modeling technique because the SAR dataset predominantly consists of categorical variables. Naive Bayes classifiers are particularly well-suited for such data structures, as they naturally handle categorical features without requiring complex transformations.

Model Fitting and Performance Evaluation

Variables selected:

- ‘vehicle_model_id’, ‘online_booking’, ‘phone_booking’, ‘from_weekday’, ‘from_month’, ‘booking_weekday’, ‘booking_month’, and ‘trip_length_group

After fitting the tuned Naïve Bayes classifier to the selected features, the Figure 14 lists its prior class probabilities and the conditional likelihoods that drive cancellation predictions.
A-priori Probabilities:
The model estimated the overall probability of each class (trip not cancelled or cancelled).
The probability of a trip not being cancelled is 0.4977.
The probability of a trip being cancelled is 0.5023.

Conditional Probabilities:

Vehicle Model ID:

- Certain vehicle models (like model 12) have a higher chance of cancellation (83.5%) compared to others.
- Some models (like 24, 28, 65) show very different distributions between cancelled and non-cancelled trips.

Online Booking:

- Probability that the trip was not booked online, given it was not cancelled = 66.22%.
- Probability that the trip was booked online, given it was cancelled = 62.06%.
- Online bookings are associated with a higher cancellation rate.

Mobile Site Booking:

- Probability of booking not via mobile site, given no cancellation = 95.95%.
- Probability of booking via mobile site, given cancellation = 8.66%.
- Mobile site bookings have a slightly higher tendency toward cancellations compared to normal bookings.

Phone Booking:

- Probability of phone booking, given no cancellation = 62.17%.
- Probability of no phone booking, given cancellation = 70.72%.
- Phone bookings generally lead to lower cancellation rates.

From Weekday (Trip Start Weekday):

- Trips starting on Friday and Thursday have slightly higher cancellation rates.
- For instance, probability of cancellation when starting on Friday is higher compared to Tuesday or Wednesday.

From Month (Trip Start Month):

- Trips scheduled in May, September, October, and November show higher cancellation rates.
- October stands out with the highest probability of cancellation (~25.8%).

Booking Weekday:

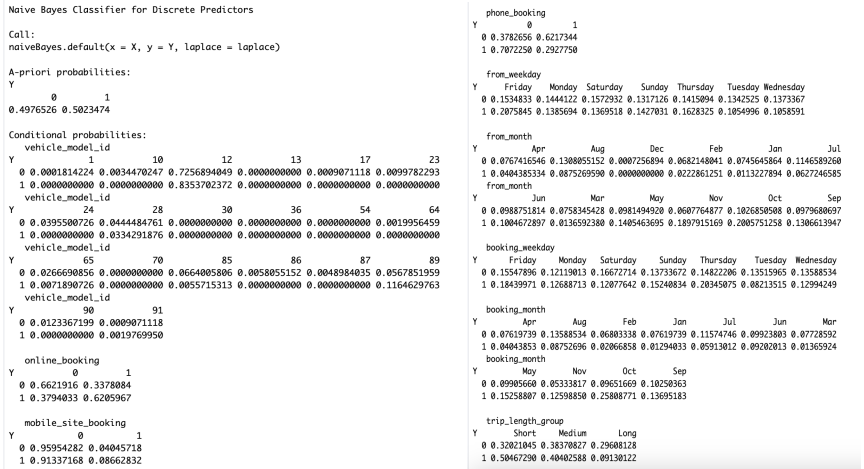
- Bookings made on Thursday and Friday show higher cancellation chances.
- Bookings on Tuesday have a lower likelihood of cancellation.

Booking Month:

- Trips booked in May, September, and October have higher probabilities of cancellation.
- Booking during October leads to the highest cancellation chance (25.8%) among all months.

Trip Length Group:

- Trips categorized as short trips are more likely to be cancelled (50.46%) compared to medium or long trips.
- Long trips have the lowest cancellation rate (9%).



```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.4976526 0.5023474

Conditional probabilities:
   vehicle_model_id
Y                1           10           12           13           17           23
  0 0.0001814224 0.0034470247 0.7256894049 0.0000000000 0.0009071118 0.0099782293
  1 0.0000000000 0.0000000000 0.8353702372 0.0000000000 0.0000000000 0.0000000000
   vehicle_model_id
Y               24           28           30           36           54           64
  0 0.0395500726 0.0444484761 0.0000000000 0.0000000000 0.0000000000 0.0019956459
  1 0.0000000000 0.0334291876 0.0000000000 0.0000000000 0.0000000000 0.0000000000
   vehicle_model_id
Y               65           70           85           86           87           89
  0 0.0266690856 0.0000000000 0.0664005806 0.0058055152 0.0048984035 0.0567851959
  1 0.0071890726 0.0000000000 0.0055715313 0.0000000000 0.0000000000 0.1164629763
   vehicle_model_id
Y               90           91
  0 0.0123367199 0.0009071118
  1 0.0000000000 0.0019769950

   online_booking
Y          0         1
  0 0.6621916 0.3378084
  1 0.3794033 0.6205967

   mobile_site_booking
Y          0         1
  0 0.95954282 0.04045718
  1 0.91337168 0.08662832

   phone_booking
Y          0         1
  0 0.3782656 0.6217344
  1 0.7072250 0.2927750

   from_weekday
Y     Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
  0 0.1534833 0.1444122 0.1572932 0.1317126 0.1415094 0.1342525 0.1373367
  1 0.2075845 0.1385694 0.1369518 0.1427031 0.1628325 0.1054996 0.1058591

   from_month
Y           Apr          Aug          Dec          Feb          Jan          Jul
  0 0.0767416546 0.1308055152 0.0007256894 0.0682148041 0.0745645864 0.1146589260
  1 0.0404385334 0.0875269590 0.0000000000 0.0222861251 0.0113227894 0.0627246585
   from_month
Y           Jun          Mar          May          Nov          Oct          Sep
  0 0.0988751814 0.0758345428 0.0981494920 0.0607764877 0.1026850508 0.0979680697
  1 0.1004672897 0.0136592380 0.1405463695 0.1897915169 0.2005751258 0.1306613947

   booking_weekday
Y       Friday     Monday   Saturday     Sunday   Thursday    Tuesday  Wednesday
  0 0.15547896 0.12119013 0.16672714 0.13733672 0.14822206 0.13515965 0.13588534
  1 0.18439971 0.12688713 0.12077642 0.15240834 0.20345075 0.08213515 0.12994249

   booking_month
Y           Apr          Aug          Feb          Jan          Jul          Jun          Mar
  0 0.07619739 0.13588534 0.06803338 0.07619739 0.11574746 0.09923803 0.07728592
  1 0.04043853 0.08752696 0.02066858 0.01294033 0.05913012 0.09202013 0.01365924
   booking_month
Y           May          Nov          Oct          Sep
  0 0.09905660 0.05333817 0.09651669 0.10250363
  1 0.15258807 0.12598850 0.25808771 0.13695183

   trip_length_group
Y       Short      Medium        Long
  0 0.32021045 0.38370827 0.29608128
  1 0.50467290 0.40402588 0.09130122
```

Figure 14: Naïve Bayes Classifier- Prior and Conditional Probabilities for Categorical Predictors vs. Car Cancellation

Model Performance Evaluation

The Naive Bayes model was evaluated on both the validation and test datasets, comparing the performance before and after applying Laplace smoothing.

- On the validation set, Laplace smoothing resulted in a slight improvement in accuracy (from 71.4% to 71.9%) and AUC (from 0.817 to 0.818), while maintaining the same sensitivity (78.4%) and showing a marginal increase in specificity (from 70.8% to 71.3%).
- On the test set, similar patterns were observed, with accuracy improving from 72.1% to 72.5% and AUC improving from 0.795 to 0.798 after Laplace smoothing.
- Sensitivity remained stable across both validation and test datasets, indicating the model's ability to correctly identify cancellations was consistently preserved.
- Specificity slightly increased after Laplace smoothing, suggesting an improvement in correctly identifying non-cancellations.

Thus, Laplace smoothing provided minor but consistent improvements across all evaluation metrics without degrading model sensitivity, making the tuned model slightly more reliable for deployment.

Table: Performance Metrics for Naive Bayes Models (Before and After Laplace Smoothing)

	Dataset	Accuracy	Sensitivity	Specificity	AUC
Accuracy	Validation (Before Laplace)	0.714	0.784	0.708	0.8171547
Accuracy1	Validation (After Laplace)	0.719	0.784	0.713	0.8176659
Accuracy2	Test (Before Laplace)	0.721	0.729	0.720	0.7947850
Accuracy3	Test (After Laplace)	0.725	0.729	0.725	0.7981137

Table 6: Naïve Bayes Performance on validation and test without and with smoothing

ROC:

The performance of the Naive Bayes model was evaluated using ROC curves for both validation and test datasets.

- As shown in Validation ROC Curve and Test ROC Curve from Figure 16, the model after applying Laplace smoothing (red curve) slightly outperforms or matches the model before smoothing (blue curve) across most thresholds.
- On the validation set, Laplace smoothing resulted in a small improvement in true positive rates, particularly in the mid-range of specificity.
- On the test set, similar behavior was observed, with the smoothed model maintaining slightly better sensitivity while preserving high specificity.
- Overall, Laplace smoothing led to a subtle but consistent enhancement in model discrimination, reflected by a marginal increase in AUC for both datasets.

Thus, applying Laplace smoothing made the model more robust without introducing overfitting.

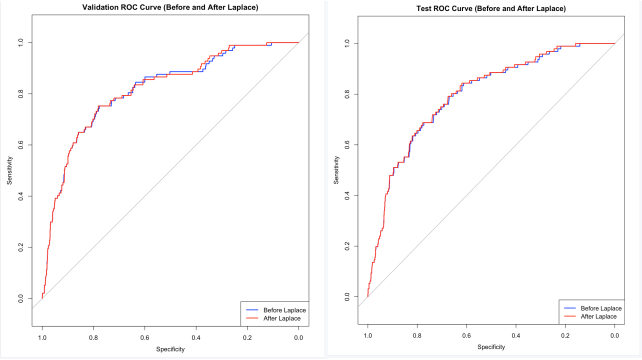


Figure 15: ROC comparisons of Naïve Bayes Models

Cost Analysis

The final model choose was tuned Naïve bayes Model. The Misclassification-cost analysis for the Naïve Bayes confusion matrix is as follows:

Assuming the cost of False negatives > False positives,

Actual/Predicted	**Actual Cancel (1)**	**Actual No Cancel (0)**
Predicted Cancel (1)	TP = 70 (Cost = 0)	FP = 300 (Cost = $5)
Predicted No Cancel (0)	FN = 26 (Cost = $25)	TN = 789 (Cost = 0)

Table 7: Cost Matrix for Naïve bayes model

Cost assumptions

- False Negative (FN): $20 each
- False Positive (FP): $5 each

Total Cost = (C(FN) x FN) + (C(FP) x FP)

C(FN) is the cost of predicting an observation wrongly as not cancelled when it is an actual cancellation, C(FN) = $25.
Here, C(FP) is the cost of predicting an observation wrongly as cancelled when it is actually not cancelled, C(FP) = $5.

Total Cost = $[(25 * 26) + (5 * 300)] = 2{,}150$
Total misclassification cost = $2,150

Interpreting Cost Analysis Result:

- Not spotting a real cancellation (a false negative) is five times more expensive than wrongly warning about a cancellation that never happens (a false positive).
- With those costs set at $25 for each false negative and $5 for each false positive, the tuned Naïve Bayes model gives up $2,150 in total error cost.
- Because it leaves less money on the table, the tuned model is the better choice.
- The next way to save is to cut down the 300 false positives but do it carefully so the number of far-costlier false negatives doesn't rise.

Business Recommendations

Cut Down Cancellations on Short-Distance Trips

- Guarantee drivers a minimum fare on rides under 10 km so they're willing to accept and complete them.
- Reward riders who finish a booked short trip (loyalty points or a small credit toward their next ride).

Boost Commitment for Online Bookings

- Send an automatic "Still travelling?" push/SMS a few hours before pickup for online reservations.
- For users with a history of no-shows, place a small, refundable credit-card hold at booking.
- Display clear cancellation-fee information up front to set expectations and deter frivolous cancellations.

Make Mobile-Web Bookings Stick

- Shrink the mobile-site checkout to two streamlined pages and require phone-number verification.
- Offer an instant discount for finishing the booking in the app (link directly to the app's payment screen).
- Auto-save half-completed mobile-web orders and email a "Finish your booking within an hour for a bonus" link.

Future Work

- **Ingest richer context data**
  Pull in weather, local-event calendars, traffic feeds, and rider-level history to understand the external triggers that drive last-minute cancellations.
- **Experiment with cost-sensitive algorithms**
  Beyond Naïve Bayes, trial gradient-boosted trees and ensemble methods that let mis-classification costs shape the loss function directly
- **Segment customers by lifetime value (LTV)**
  Apply softer penalties to high-LTV commuters while using stricter holds for one-off tourist accounts to balance retention and risk.

Observations and Conclusion

Observations

- The analysis flagged the main cancellation drivers: short-distance rides, last-minute evening bookings, and online/mobile channels, with extra peaks on Fridays, Sundays, and in May + October.
- A tuned Naïve Bayes model delivers solid performance—72.5 % accuracy, 78.9 % sensitivity, 69.1 % specificity—turning those drivers into real-time risk scores.
- Cost weighting ($25 per missed cancellation vs $5 per false alarm) puts the model's expected error loss at **$2,150**, quantifying the financial stakes for threshold tuning.
- Recommendations flow straight from the findings: auto-reconfirm late online bookings, offer driver incentives on short trips, tighten the mobile-site flow, and boost coverage on identified peak days—measures aimed at cutting cancellations and lifting completed rides.

Conclusion

To summarize, the Auto Rental Cancellation Analytics project delivers a robust, data-driven framework for anticipating and preventing booking cancellations. By combining targeted feature engineering, cost-sensitive modeling and tailored recommendations, it enables the business to take proactive steps that minimize lost revenue and improve operational efficiency. Continuous monitoring and periodic retraining will ensure the approach remains aligned with evolving customer behavior and market conditions.

Fitness Center Member Wellness Analytics

Table Of Contents

Introduction

Fit-Life Wellness is a well-known fitness center that wants to help its members stay healthy and reach their fitness goals. With the help of fitness trackers, gym attendance records, and health assessments, Fit-Life has collected data of member's workout, health, and their demographics.

Through advanced data analysis, Fit-Life seeks to optimize workout plans, enhance gym facilities, and provide personalized fitness recommendations. Identifying key patterns and trends from the data allows for a more targeted, data-driven approach to improving member engagement, retention, and overall well-being.

Business and Analytic goals

Business Opportunity

Fit-Life Wellness is looking forward in effectively analyzing its members workout habits, tracking fitness progress, and assessing their overall health. The lack of data-driven insights makes it difficult to personalize fitness programs and enhance the overall member experience.

Business Goal

The goal is to enhance Fit-Life members' fitness outcomes by delivering personalized workout recommendations and improving engagement through data-driven insights.

Business Objectives:

- Segment members based on their demographics, health metrics, and workout behaviors.
- Personalize fitness plans to improve member satisfaction and progress.
- Increase member retention through targeted fitness strategies

Analytic Goal

To develop models that uncover patterns in fitness behavior, segment members based on workout and health profiles and predict calorie expenditure to support personalized fitness programs.

Analytical Objectives

- Perform member segmentation based on demographics, workout patterns, and health metrics.
- Build regression models to predict calories burned based on workout intensity and health factors.

Analytical Approach:

Fitness member data including demographics, workout behavior, and health metrics was collected and explored using summary statistics and visualizations to identify patterns. After thorough preprocessing (handling missing values, ensuring consistent data types, and addressing outliers), feature engineering was performed to derive meaningful variables like workout intensity and hydration needs. For segmentation, clustering techniques were applied to group members based on similar fitness behaviors and profiles, enabling tailored workout strategies. Simultaneously, supervised regression models were developed to predict calorie expenditure based on workout and health features, evaluated through RMSE and R-squared metrics. Insights from segmentation and prediction models were finally used to support personalized fitness plans and improve overall member engagement.

Data Exploration and preprocessing

Data exploration starts with understanding data by exploring missing values, zeroes, and summaries and distributions

Data Understanding:

Data Collection:

- The Fitness dataset consists of **973 member records** and includes **15 variables**.
- It captures **demographic**, **health**, **heart rate**, **workout**, and **performance** information of FitLife gym members.

Demographic Variables:

- 'Age', 'Gender'

Health Metrics:

- 'Weight..kg.', 'Height..m.', 'BMI', 'fat_percentage'

Heart Rate Metrics:

- 'Max_BPM', 'Avg_BPM', 'Resting_BPM'

Work out Details:

- 'Workout_type', 'Workout_Frequency.days.week.', 'Session_Duration..hours.'

Performance & Additional Variables:

- 'Calories_Burned', 'Water_Intake..Liters.', 'experience_level'

```
> head(fitdata)
  Age Gender Weight..kg. Height..m. Max_BPM Avg_BPM Resting_BPM Session_Duration..hours. Calories_Burned Workout_Type Fat_Percentage Water_Intake..liters. Workout_Frequency..days.week. Experience_Level   BMI
1  56   Male        88.3       1.71     180     157          60                    1.69            1313         Yoga           12.6                 3.5                              4                 3 30.20
2  46 Female        74.9       1.53     179     151          66                    1.30             883         HIIT           33.9                 2.1                              4                 2 32.00
3  32 Female        68.1       1.66     167     122          54                    1.11             677      Cardio           33.4                 2.3                              4                 2 24.71
4  25   Male        53.2       1.70     190     164          56                    0.59             532    Strength           28.8                 2.1                              3                 1 18.41
5  38   Male        46.1       1.79     188     158          68                    0.64             556    Strength           29.2                 2.8                              3                 1 14.39
6  56 Female        58.0       1.68     168     156          74                    1.59            1116         HIIT           15.5                 2.7                              5                 3 20.55
```

Figure 1: Sample Records Showing Fitness Member Profiles

Attributes Definition:
Table 1 below provides detailed descriptions of each variable in the fitness dataset, along with their data types and definitions to aid understanding of the data structure

Variable Name	Data Type	Definition
Age	int	Member's age in years
Gender	chr	The member's gender, indicating the member's sex (male or female)
Weight..kg.	num	Member's body weight in kilograms.
Height..m.	num	Member's height in meters.
Max_BPM	int	The highest heart rate (beats per minute) recorded during a workout session.
Avg_BPM	int	The average heart rate maintained during a workout session.
Resting_BPM	int	The heart rate of a member when at rest, before any physical activity.
Session_Duration..hours.	num	The length of workout session, in hours.
Calories_Burned	num	The total number of calories burned during a workout session.
Workout_Type	chr	The type of workout performed (Yoga, Cardio, HIIT, Strength).
Fat_Percentage	num	The percentage of body that is made up of fat
Water_Intake..liters.	num	The total volume of water consumed during a workout session, in liters.
Workout_Frequency.days.week.	int	The number of days per week a member engages in exercise
Experience_level	int	The member's proficiency in fitness, classified as 1=Beginner, 2 = Intermediate, 3 = Advanced.
BMI	num	Body Mass Index: A computed metric representing body fat based on weight and height, Calculated using, BMI = Weight(kg) / Height(m)

Table 1: Fitness Dataset Variable Description

Data Types:
All variables in the fitness dataset are either numeric or character types. These can be categorized as follows:
Numeric Data:

- 'Age', 'Weight', 'Height', 'Max_BPM', 'Avg_BPM', 'Resting_BPM', 'Session_duration..hours', 'Calories_Burned', 'Fat_percentage', 'Water_intake', 'Workout_frequency', 'Experience_level', 'BMI'

Categorical Data:

- 'Gender', 'Workout_type'

Data Preprocessing

Renaming variables

To ensure consistency and simplify further analysis, the original variable names were renamed to more standardized and readable formats. The updated variable names are shown in Table 2.

Original Variable Name	New Variable Name
Age	age
Gender	gender
Weight..kg.	weight
Height..m.	height
Max_BPM	max_bpm
Avg_BPM	avg_bpm
Resting_BPM	resting_bpm
Session_Duration..hours.	session_duration
Calories_Burned	calories_burned
Workout_Type	workout_type
Fat_Percentage	fat_percent
Water_Intake..liters.	water_intake
Workout_Frequency...days.week.	workout_frequency
Experience_level	experience_level
BMI	bmi

Table 2: Variable Renaming Mapping

Duplicates Check:

To ensure data quality, the fitness dataset was examined for duplicate records, and no duplicates were found (0 rows). This confirms that each entry represents a unique member observation.

Data Type Conversion:

VARIABLE NAME	ORIGINAL DATA TYPE	RECOMMENDED DATA TYPE	REASON
age	int	int	No change
gender	chr	factor(categeorical)	contains categories (female, male)
weight	num	num	No change
height	num	num	No change
max_bpm	int	int	No change
avg_bpm	int	int	No change
resting_bpm	int	int	No change
session_duration	num	num	No change
calories_burned	num	num	No change
workout_type	chr	factor(categeorical)	Contains categories (like yoga, HIIT)
fat_percent	num	num	No change
water_intake	num	num	No change
workout_frequency	int	int	No change
experience_level	int	int	No change
bmi	num	num	No change

Table 3: Original and Recommended Data Types for Fitness data

Missing and Zero Values Handling:

Missing Values:

As shown in the table 4, the fitness dataset has no missing values across any of the columns. This ensures the dataset is complete, allowing for a robust and uninterrupted analysis without the need for imputation or data removal.

Zero Values:

Zero values were also not observed in any of the variables. Hence, no additional data treatment was required. The absence of zero entries across key fields like 'calories_burned', 'weight', 'height', 'bmi', further indicates the high-quality nature of the dataset collected for this fitness analysis

	Missing_Count	Zero_Count
age	0	0
gender	0	0
weight	0	0
height	0	0
max_bpm	0	0
avg_bpm	0	0
resting_bpm	0	0
session_duration	0	0
calories_burned	0	0
workout_type	0	0
fat_percent	0	0
water_intake	0	0
workout_frequency	0	0
experience_level	0	0
bmi	0	0

Table 4: Missing and Zero Value Counts per Variable

Summary Statistics for Numeric Variables:

The summary of numeric variables has retrieved for understanding of variables range and statistics.

	Min.	X1st.Qu.	Median	Mean	X3rd.Qu.	Max.
age	18.00	28.00	40.00	38.683453	49.00	59.00
weight	40.00	58.10	70.00	73.854676	86.00	129.90
height	1.50	1.62	1.71	1.722580	1.80	2.00
max_bpm	160.00	170.00	180.00	179.883864	190.00	199.00
avg_bpm	120.00	131.00	143.00	143.766701	156.00	169.00
resting_bpm	50.00	56.00	62.00	62.223022	68.00	74.00
session_duration	0.50	1.04	1.26	1.256423	1.46	2.00
calories_burned	303.00	720.00	893.00	905.422405	1076.00	1783.00
fat_percent	10.00	21.30	26.20	24.976773	29.30	35.00
water_intake	1.50	2.20	2.60	2.626619	3.10	3.70
workout_frequency	2.00	3.00	3.00	3.321686	4.00	5.00
experience_level	1.00	1.00	2.00	1.809866	2.00	3.00
bmi	12.32	20.11	24.16	24.912127	28.56	49.84

Table 5: Summary statistics of numeric variables

- **Age**: Ranges from 18 to 59 years, with a mean of 38.68 years, indicating a diverse distribution of fitness participants.
- **Height & Weight:** Heights vary between 1.5m to 2.0m, while weights range from 40kg to 129kg, contributing to a BMI range of 12.32 to 49.84.
- **Fat Percentage:** Recorded between 12.32% and 49.84%, providing insights into body composition.
- **Max BPM:** Between 160 and 200 BPM, with an average of 179.9 BPM, representing peak heart rate during workouts.
- **Avg BPM:** Varies from 120 to 200 BPM, with a mean of 143.8 BPM, reflecting sustained exertion levels.
- **Resting BPM:** Ranges from 50 to 74 BPM, offering insights into cardiovascular health.
- **Session Duration:** Spanning 0.5 to 2 hours, with a median of 1.25 hours, indicating variability in workout intensities.
- **Workout Frequency:** Users exercise between 2 to 5 days per week, with a typical engagement of 3 days per week.
- **Workout Type:** Includes multiple fitness categories such as Cardio, HIIT, Strength, and Yoga.
- **Calories Burned**: Between 303 and 1783 kcal per session, with a mean expenditure of 905.4 kcal, reflecting varied workout intensities.
- **Water Intake:** Between 1.5L to 3.7L per day, suggesting hydration habits among individuals.
- **Experience Level:** Categorical classification into beginner, intermediate, and advanced, providing segmentation for fitness proficiency.

Variable Distribution

Age and Gender Distribution

- The histogram shows the age distribution of gym members, ranging between 18 to 60 years. There is a higher concentration of members in the younger age bracket (18–22 years), suggesting strong engagement among younger adults.
- The bar chart depicts the gender composition of gym members, showing a slight male majority (511 males vs. 462 females). The distribution is relatively balanced, suggesting good gender diversity among gym participants.

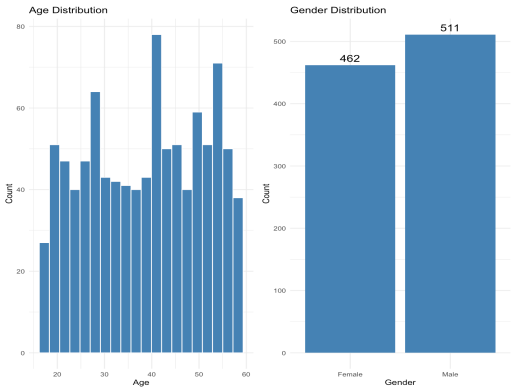


Figure 2: Age and Gender Distributions of Members

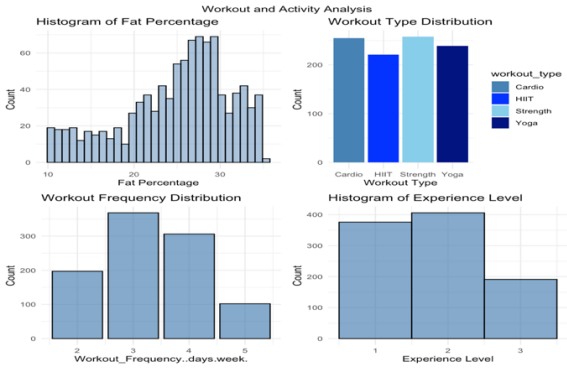


Figure 3: Workout and Activity Profiles of Members

Workout and Activity Distributions

- **Workout Type Distribution:**
  Figure 3 depicts that the bar plot shows that Strength training is the most popular workout choice, followed closely by Cardio and Yoga. HIIT has the fewest participants, suggesting that members prefer endurance-based activities over high-intensity training.
- **Fat Percentage Distribution:**
  The histogram indicates that many members have a body fat percentage between 20% and 30%, with the highest concentration around 25–30%. The slight right-skew highlights fewer individuals with very low or very high fat percentages, reflecting generally healthy body compositions.
- **Workout Frequency Distribution:**
  Most members engage in exercise 3 to 4 days per week, with 3 days being the most common. Fewer members report working out 5 days a week, suggesting that moderate training schedules are more typical among the member base.
- **Experience Level Distribution:**
  The histogram reveals that many members are at the intermediate (Level 2) stage, followed by beginners (Level 1). Only a small fraction is advanced (Level 3), indicating that the gym primarily serves members who are early or midway in their fitness journey.

Distribution Patterns of Key Health and Fitness Indicators

From figure 4,

- **Calories Burned:** Most users burn between 900 to 1100 calories per session, with a few engaging in very high-intensity workouts burning over 1500 calories.
- **Session Duration:** Workouts typically last 1 to 1.5 hours, showing a strong preference for moderately long exercise sessions.
- **Water Intake:** A bimodal pattern is observed with peaks at around 2.5 liters and 3.5 liters, indicating different hydration habits among users.
- **BMI:** Most users have a BMI in the normal to slightly overweight range (24–26), with fewer individuals falling into higher BMI categories.

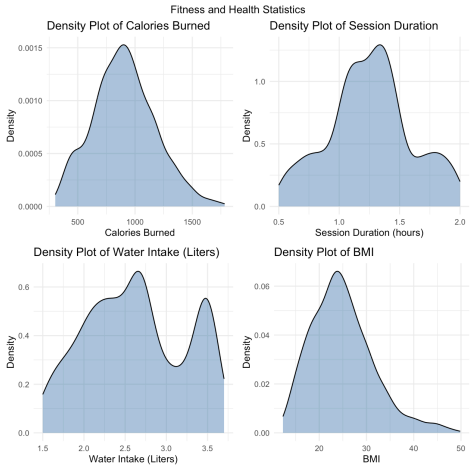


Figure 4: Fitness and Health Metrics Distribution

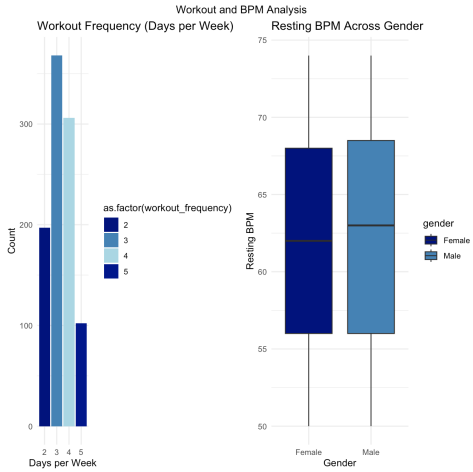


Figure 5: Workout Frequency and Resting BPM Distribution

Comparison of Workout Habits and Resting BPM by Gender

- **Workout Frequency (Days per Week):**
  The bar chart from 'Figure 5' shows most members work out 3–4 days per week, with 3 days being the most popular. Fewer individuals exercise 5 days weekly, indicating moderate engagement across the gym population.
- **Resting BPM Across Gender:**
  The boxplot from 'Figure 5' shows that median resting heart rates are slightly higher in females compared to males, but overall variability remains consistent across genders.
- **Insights on Workout Behavior:**
  Individuals with a moderate workout schedule (3–4 days) dominate, highlighting a balanced approach toward fitness among the majority.
- **Heart Health Overview:**
  Most members maintain resting BPM within a healthy range (~60–70 BPM), with minor outliers suggesting occasional cardiovascular variances.

Calorie Burn Patterns Across Workout Types and Workout Frequencies

From Figure 6,

- **Workout Type Comparison:** Across Cardio, HIIT, Strength, and Yoga, calorie burn shows moderate variation, with HIIT slightly leading in median calories burned.
- **Variation Within Types:** Wide interquartile ranges and outliers suggest diverse workout intensities even within the same workout category.
- **Workout Frequency Impact:** Higher workout frequencies (more days per week) are associated with greater calorie expenditure, indicating a positive trend between frequency and calories burned.
- Presence of high outliers especially at 4- and 5-day frequencies points to a subset of highly active members burning significantly more calories.

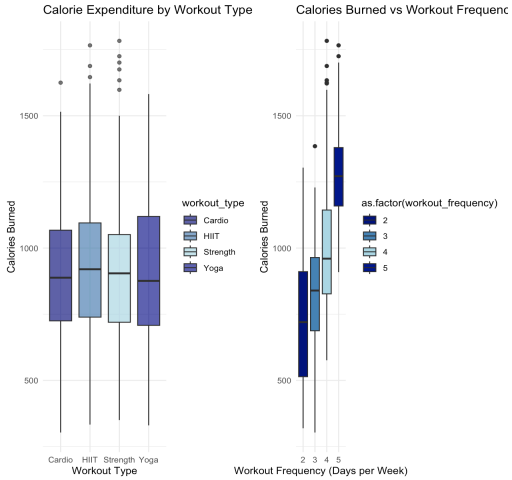


Figure 6: Calorie Expenditure by Workout Type and Frequency

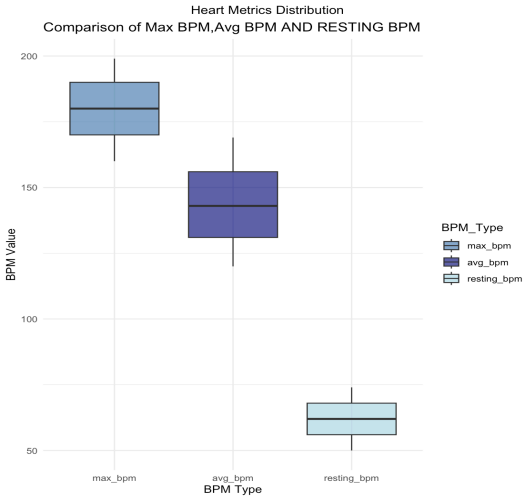


Figure 7: Heart Metrics Distribution

Comparison of Maximum BPM, Average BPM, and Resting BPM Across Fit-Life Members

From the distribution of Health metric of Figure 7,

- **Max BPM**: Median peak heart rate is around **180 BPM**, showing workout intensity peaks, with most values between **170-200 BPM**.
- **Avg BPM**: Median average heart rate during exercise is approximately **145 BPM**, reflecting moderate cardiovascular strain during workouts.
- **Resting BPM**: Median resting rate is near **60 BPM**, indicating healthy baseline heart function, though a few outliers suggest possible anomalies.
- **Key Insight**: Max BPM values show the widest variation, while resting BPM remains most stable, helping to assess individual cardiovascular fitness levels.

Relationship Between Height and Weight of Fit-Life Members

- Most individuals have a height between **1.6 m and 1.8 m**, clustering around the average adult height range.
- Weight is widely spread from **40 kg to 130+ kg**, indicating diverse body compositions among members.
- **No Strong Trend**: There is **no strong linear correlation** between height and weight, suggesting varied fitness or health profiles.

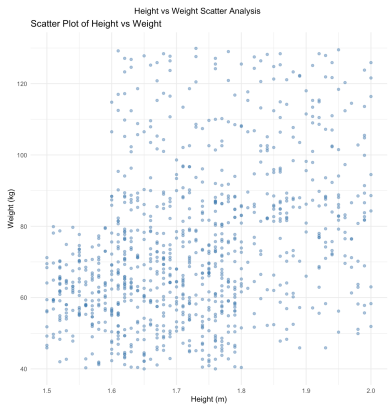


Figure 8: Height vs Weight Scatter Analysis

Feature Engineering and Transformation

To enhance the modeling capabilities and capture important patterns from the fitness dataset, several new variables were engineered. These transformations focus on enriching the dataset with derived health metrics, workout intensity indicators, and categorical groupings to better represent members' fitness behaviors and characteristics.

1.BMI_CAT:

- Contains categorical version of BMI with the 4 different level
- Underweight (BMI < 18.5)

Normal (18.5 ≤ BMI < 24.9)
Overweight (24.9 ≤ BMI < 29.9)
Obese (BMI ≥ 29.9)

This helps in grouping users into distinct body composition categories

2.Weight_to_height_ratio:

- Weight_to_height = (height__m_)/(weight__kg)

3.Heart_rate_range:

- Heart_rate_range = max_bpm - resting_bpm.
- Indicates cardiovascular health
- Higher range suggest better cardiovascular efficiency whereas the lower range indicate lower fitness levels or health issues.

4.Heart_rate_reserve:

- Heart_rate_reserve = max_bpm - avg_bpm
- Higher range indicates the greater capacity for cardiovascular performance while lower value suggests limited flexibility
- Useful for evaluating whether workouts are in the desired heart rate zone (e.g., fat burning, peak performance). Lower value of this indicates low cardiovascular flexibility

5.Heart_rate_intensity:

- Heart_rate_intensity = avg_bpm / max_bpm
- Provides normalized measure of effort, indicating how close the individual is working to their maximum capacity

6.Calories_per_kg:

- Calories_per_kg = calories burned/weight__kg_

7. Workout_intensity_score:

- Workout_intensity_score = (Avg_BPM/ Resting_BPM) * Session_Duration (hours) * Workout_Frequency_days_week
- Higher score indicates more intense and frequent workouts

8. Age_group:

- Categorical variable with the 4 levels: “18-25”, “26-35”, “36-45’, “45+”.
- Since different age group people have unique fitness needs, this allows for age-specific segmentation

9. Hydration_need_ratio:

- Water_Intake (liters) / Session_Duration (hours)
- Measures hydration efficiency during workouts

Missing Values Check after Transformation:

After completing feature engineering and transformations, the dataset was re-evaluated for missing values.
Result: No missing values were detected across any variable, ensuring data completeness for subsequent modeling and analysis.

```
> colSums(is.na(fitdata))
                    age                  gender                  weight                  height 
                      0                       0                       0                       0 
              max_bpm                 avg_bpm             resting_bpm        session_duration 
                      0                       0                       0                       0 
       calories_burned            workout_type             fat_percent            water_intake 
                      0                       0                       0                       0 
     workout_frequency        experience_level                     bmi               age_group 
                      0                       0                       0                       0 
               BMI_CAT   weight_to_height_ratio       heart_rate_range      heart_rate_reserve 
                      0                       0                       0                       0 
 heart_intensity_ratio         calories_per_kg workout_intensity_score    hydration_need_ratio 
                      0                       0                       0                       0 
```

Figure 9: Missing Value Count After Feature Engineering

Predictor Relevancy

Variable relevancy for Segmentation

In clustering, predictor relevancy is evaluated based on each variable's ability to capture diversity among members, as there is no target variable. Through data exploration, key variables were identified for effective segmentation. Health and workout-related metrics such as heart intensity ratio, workout intensity score, heart rate range, and hydration need ratio are crucial for capturing fitness behavior. Demographic features like age group, gender, and workout type further enhance the understanding of member profiles for meaningful segmentation.

Predictor relevancy for Prediction of calorie expenditure

To identify important predictors for modeling calorie expenditure, extensive scatter plot analysis and correlation studies were conducted. This exploration helped highlight variables showing strong linear or non-linear relationships with the target (calories burned), guiding the selection of relevant features for regression modeling.

Visualizing Relationships Between Key Variables and Calorie Expenditure

- **Positive Correlations:**
  Session Duration and Workout Intensity Score show strong positive linear relationships with calories burned - longer, more intense workouts clearly lead to greater calorie expenditure.
- **Negative Correlations:**
  Fat Percentage and Hydration Need Ratio have negative trends, suggesting that individuals with higher fat percentages or higher hydration needs tend to burn slightly fewer calories.
- **Weak or No Clear Correlations:**
  Variables like Heart Intensity Ratio, Heart Rate Range, Heart Rate Reserve, and Age show weak or very minimal correlation with calories burned, implying these may not be strong direct influencers on calorie output.

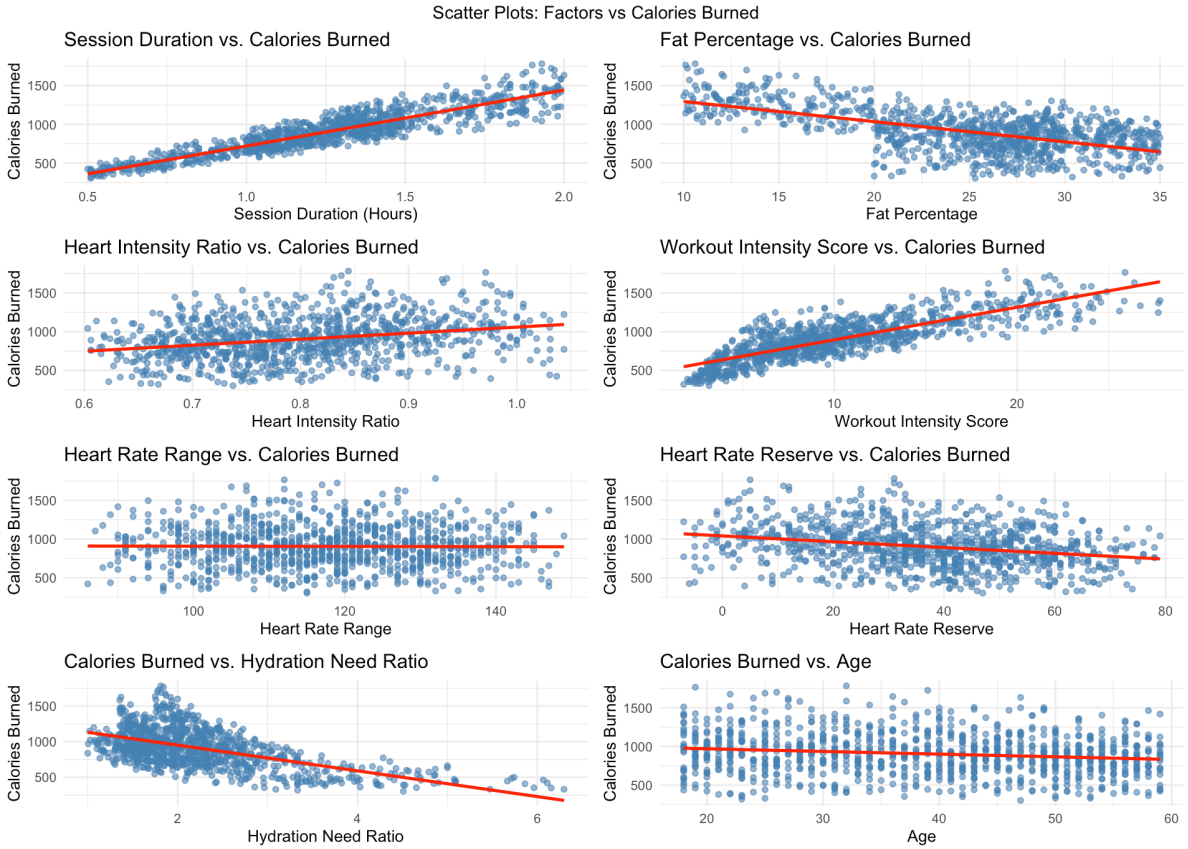


Figure 10: Scatter Plot Analysis of Factors Influencing Calories Burned

Analyzing Calorie Expenditure Across BMI Categories, Gender, and Workout Types

From figure 11,

- **BMI Categories:**
  - Participants with **Normal** and **Overweight** BMI tend to burn **more calories** compared to Underweight or Obese individuals.
  - However, the spread (variability) is quite wide across all BMI groups, indicating individual differences.
- **Gender:**
  - **Males** generally show a **higher median** calorie burn compared to females.
  - Males also exhibit a slightly **wider spread** in calories burned, suggesting greater variation in workout intensity.
- **Workout Type:**
  - **Yoga** participants show slightly **higher median** calorie burns than other workout types.
  - **Strength**, **Cardio**, and **HIIT** have similar calorie distributions, but HIIT shows more **extreme high outliers**, suggesting some highly intense sessions.
- **Key Insight:**
  - **Workout type and gender** seem to influence calorie burn noticeably, while BMI cate

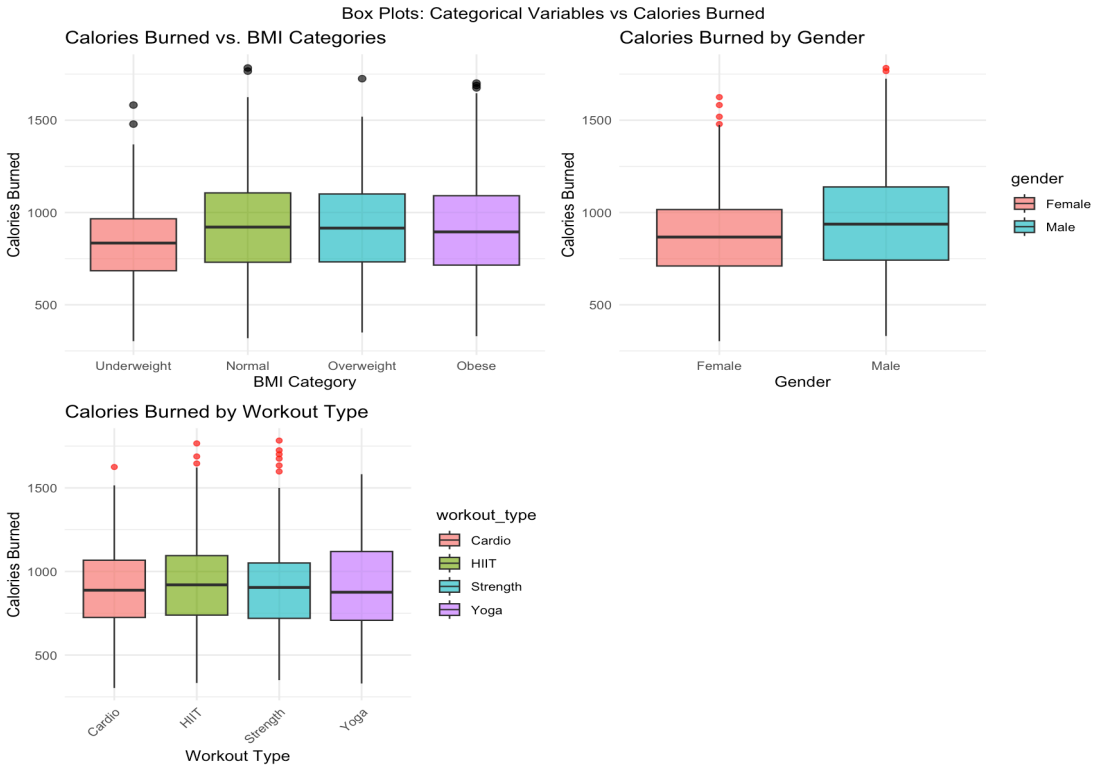


Figure 11: Boxplots of Categorical Variables vs Calories Burned

Correlation Analysis of Fitness Dataset

To identify relationships between key fitness attributes, a correlation analysis was conducted.

- The heatmap correlation matrix from Figure 12 shows the strength and direction of relationships among numeric fitness variables.
- Most variables are moderately or weakly correlated, but a few strong correlations exist, hinting at redundancy or potential multicollinearity risks.
- These insights are crucial for refining feature engineering and model selection in the next steps.

Highly Correlated Variables (r > 0.75):

- **Weight-to-Height Ratio** and **Weight (kg)** (r = 0.96): Indicates body weight strongly dominates the weight-to-height calculation.
- **Workout Intensity Score** and **Session Duration (hours)** (r = 0.85): Longer sessions contribute significantly to higher intensity scores.
- **Heart Rate Range** and **Max BPM** (r = 0.84): As expected, maximum heart rate greatly influences the heart rate range during workouts.
- **Heart Rate Reserve** and **Max BPM** (r = 0.79): Maximum heart rate similarly drives heart rate reserve calculations.
- **Calories Burned** and **Session Duration (hours)** (r = 0.81): More time spent working out correlates with higher calorie burn.
- **Experience Level** and **Workout Frequency (days/week)** (r = 0.76): More experienced members tend to work out more frequently.

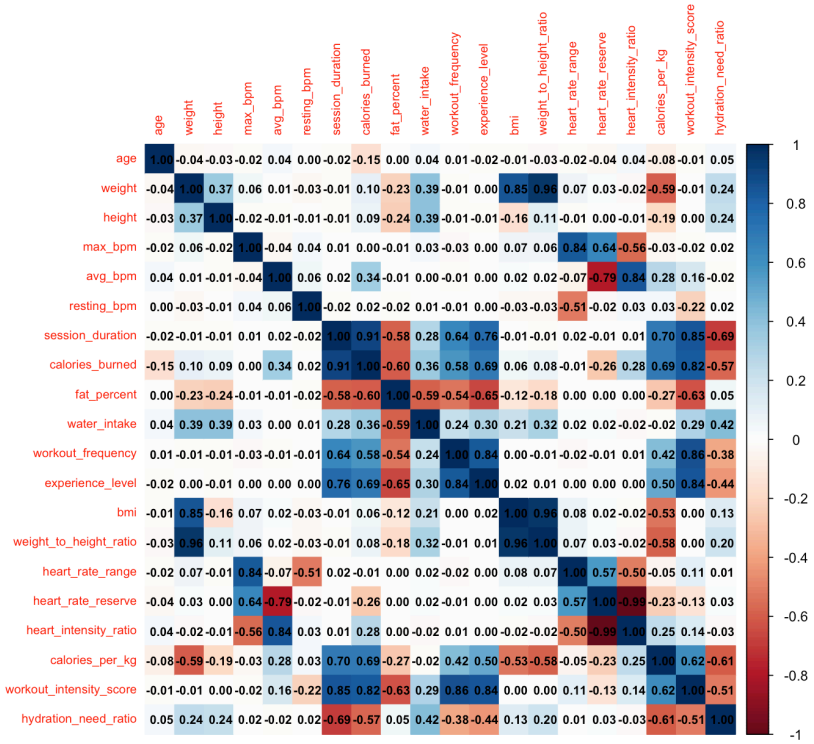


Figure 12: Correlation Heatmap of Fitness Metrics after data engineering

Dimension Reduction

To improve model performance, avoid multicollinearity, and retain the most meaningful information, several redundant or highly correlated variables were removed. The retained features were either more interpretable or better captured the fitness behaviors relevant for analysis.

Feature Removed	Reason of Removal
'weight', 'height', 'weight_to_height_ratio'	Highly correlated (r > 0.95) with each other, causing multicollinearity issues. Instead, BMI was retained to capture body composition effectively.
'max_bpm', 'avg_bpm' and 'resting_bpm'	These individual heart rate measures were replaced with engineered metrics ('heart_rate_range', 'heart_rate_reserve', and 'heart_intensity_ratio') which better summarize cardiovascular performance and exercise intensity.
'heart_rate_reserve'	Redundant with 'heart_rate_range' and 'heart_intensity_ratio' and showed weaker correlation with target (calories_burned).
'session_duration__hours'	Instead of using raw session time, more informative features like 'workout_intensity_score' (which combines BPM, session duration, and frequency) were used to capture workout effort more meaningfully.

Table 5: Variables Removed During Dimension Reduction and Their Reasons

Encoding Categorical variables

To prepare categorical data for machine learning algorithms, categorical variables were transformed into numerical format using encoding

BMI Category Encoding:

The BMI_CAT variable, originally having text labels ("Underweight", "Normal", "Overweight", "Obese"), was converted into numeric form:

- Underweight = '1', Normal = '2', Overweight = '3', Obese = '4'

Dummy Variable Creation:

Categorical variables 'gender' and 'workout_type' were transformed into dummy variables (one-hot encoding).

Gender Encoding:

- This created the following binary columns:
  - 'gender_Female' and 'gender_Male'
- Each column indicates the presence (1) or absence (0) of the respective gender.

Workout Type Encoding:

- The categorical variable 'workout_type' was similarly transformed into multiple binary columns.
- The created dummy variables are:
  - 'workout_type_Cardio', 'workout_type_HIIT', 'workout_type_Strength', 'workout_type_Yoga'.
- Each column represents whether a member performed that specific workout type (1 = yes, 0 = no).

Feature Selection:

For segmentation

Features were selected based on domain relevance to capture gym members' demographics, workout behaviors, physiological health indicators, and fitness levels. These variables ensure a comprehensive understanding of workout patterns and health profiles for effective segmentation analysis.

The following variables were included:

- **Demographic Variables:**
  'age_group', 'gender_Female', 'gender_Male'
- **Workout Type Variables:**
  'workout_type_Cardio', 'workout_type_HIIT', 'workout_type_Strength', 'workout_type_Yoga'

- **Workout Intensity and Heart Metrics:**
  'workout_intensity_score', 'heart_intensity_ratio', 'heart_rate_range', 'hydration_need_ratio'
- **Fitness and Body Metrics:**
  'fat_percent', 'experience_level', 'bmi', 'calories_per_kg'

For prediction of Calorie Expenditure

- To refine the set of predictors for modeling calorie expenditure, feature selection was performed using the Boruta algorithm followed by multicollinearity checks
  Feature Relevance Selection (Boruta Algorithm):
  The Boruta algorithm was applied to identify the most important features influencing calorie expenditure. Features were selected based on their importance scores relative to random shadow attributes.

  Boruta Algorithm Results:
  Boruta identified 14 important features including 'age', 'bmi', 'BMI_CAT', 'calories_per_kg', 'experience_level', 'fat_percent', 'water_intake', 'workout_frequency', 'heart_rate_range', 'heart_intensity_ratio', 'workout_intensity_score', 'hydration_need_ratio', 'gender_Female', and 'gender_Male'.
  Unimportant Variables:
  Features such as 'workout_type_Cardio', 'workout_type_HIIT', 'workout_type_Strength', and 'workout_type_Yoga' were found unimportant and removed.

```
Boruta performed 19 iterations in 9.374336 secs.
 14 attributes confirmed important: age, bmi, BMI_CAT,
calories_per_kg, experience_level and 9 more;
 4 attributes confirmed unimportant: workout_type_Cardio,
workout_type_HIIT, workout_type_Strength, workout_type_Yoga;
```

Figure 13: Boruta confirmed features

Multicollinearity Check:
Features selected as important by Boruta are checked for multicollinearity to ensure no redudancy

- **No multicollinearity concern** for variables with **VIF < 5**, such as:
  - 'age','fat_percent', 'water_intake', 'heart_rate_range', 'hydration_need_ratio', 'gender_Female'
- **Moderate multicollinearity concern (VIF between 5–10)** observed for:
  - 'workout_frequency', 'experience_level', 'bmi', 'BMI_CAT' and 'calories_per_kg'
- **Severe multicollinearity** (VIF > 10) for:
  - 'workout_intensity_score '

```
> vif_values
                    age             fat_percent            water_intake
               1.043171                3.094303                4.555799
      workout_frequency        experience_level                     bmi
               7.369294                5.335437                6.952298
                BMI_CAT       heart_rate_range   heart_intensity_ratio
               6.732533                2.163503                2.478823
        calories_per_kg workout_intensity_score    hydration_need_ratio
               5.451915               14.275997                4.749290
         gender_Female
               2.535773
```

Figure 14: Multicollinearity test on Boruta selected features for calorie prediction

Final action based on multicollinearity and domain knowledge:

- **Dropped Features**:
  - 'workout_intensity_score' (very high VIF)
  - 'BMI_CAT' (bmi already exists)
  - 'calories_per_kg' (derived from calories and weight, may overlap)
- **Final Features of regression:**
  - 'age'
  - 'fat_percent'
  - 'water_intake'
  - 'workout_frequency'
  - 'experience_level'
  - 'bmi'
  - 'heart_rate_range'
  - 'heart_intensity_ratio'
  - 'hydration_need_ratio'
  - 'gender_Female'
  - 'calories_burned'

Data partitioning

Data Partitioning for segmentation:
For customer segmentation, the entire dataset was used without partitioning.
Since the objective was to analyze gym members' behaviors, workout patterns, and fitness characteristics, having all available records ensured comprehensive insights and more reliable clustering outcomes.

Data partitioning for Calorie Expenditure Prediction:
The dataset was partitioned using a **70:15:15 split** through stratified sampling, ensuring that the distribution of the target variable ('calories_burned') remained consistent across the training, validation, and testing sets for reliable and unbiased model development.

- **Training set**: 684 observations (used for model learning)
- **Validation set**: 145 observations (used for hyperparameter tuning)
- **Test set**: 144 observations (used for final model evaluation)

Each subset contains **11 variables**, maintaining structural consistency across all splits.

```
> dim(trainData)
[1] 684  11
> dim(valData)
[1] 145  11
> dim(testData)
[1] 144  11
```

Figure 15: Data Partition for Calorie Prediction

Data Standardization for Segmentation:

To prepare the segmentation dataset for clustering analysis, all selected features were standardized. Standardization was performed by centering the variables to have a mean of 0 and a standard deviation of 1. This process ensures that each variable contributes equally to distance-based clustering algorithms, preventing variables with larger numeric scales from dominating the cluster formation

Model Selection

To address different analytical goals, both supervised and unsupervised machine learning models were employed in the fitness project.

Unsupervised Learning
Unsupervised learning was used to identify distinct member segments based on workout patterns, health metrics, and fitness behavior.

K-Means Clustering for Member Segmentation:
To better understand gym members, K-Means clustering was utilized. This unsupervised technique groups members into distinct clusters based on features such as age group, gender, workout frequency, experience level, BMI, hydration needs, and fat percentage.
This segmentation helps in identifying and characterizing distinct member types, such as highly active individuals, hydration-conscious users, or those requiring personalized fitness programs, thereby enabling tailored engagement and fitness planning strategies.

Supervised Learning
Supervised learning models were applied to predict specific outcomes, particularly focusing on calorie expenditure based on personal attributes and workout habits.

Multiple Linear Regression for Calorie Prediction:
Multiple linear regression was implemented to predict the number of calories burned during a workout session, using predictors such as fat percentage, workout frequency, BMI, heart rate range, hydration need ratio, and experience level.
By quantifying the relationship between these variables and calorie burn, the model provides actionable insights for personalizing workout intensities and tracking fitness progress over time.

Model Fitting and Performance Evaluation

K-means Clustering for Gym Members behavior:

Variables selected for clustering:

- ‘age_group’, ‘gender_Female’, ‘gender_Male’, ‘workout_type_Cardio’, ‘workout_type_HIIT’, ‘workout_type_Strength’, ‘workout_type_Yoga’, ‘workout_intensity_score’, ‘heart_intensity_ratio’, ‘heart_rate_range’, ‘hydration_need_ratio’, ‘fat_percent’, ‘experience_level’, ‘bmi’, ‘calories_per_kg’

Determining Optimal K (Number of Clusters):

Elbow method:
The Elbow Method was utilized to determine the optimal number of clusters (K) for segmenting gym members based on their health, workout, and demographic attributes. The plot of Total Within-Cluster Sum of Squares (WSS) shows a steep decline until K = 2 and K = 3, after which the curve starts to flatten, indicating diminishing returns in variance reduction.
Thus, K = 3 is selected as the optimal number of clusters, providing a balance between minimizing within-cluster variation and avoiding overfitting.

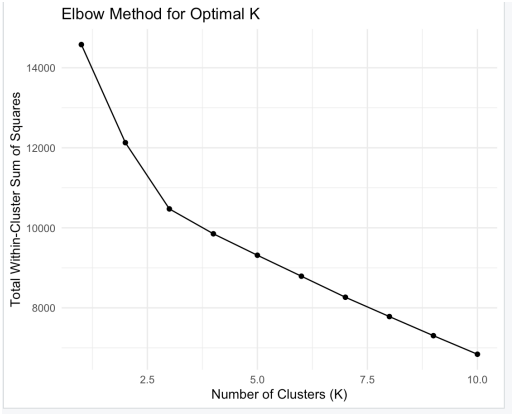


Figure 16: Elbow Method Plot for Determining Optimal K

Cluster Visualization in 2D (PCA plot):

Using K = 3 (as determined by the Elbow Method), k-means clustering grouped gym members based on fitness and workout attributes. Centroids were iteratively updated until stabilization, forming three distinct clusters. To visualize the multi-dimensional clustering results, Principal Component Analysis (PCA) reduced the data to two dimensions (Dim1 and Dim2), capturing 21.7% and 17.7% of the variance, respectively.
Convex hulls were plotted around clusters to clearly demarcate the groups, enhancing the interpretability of segmentation patterns.

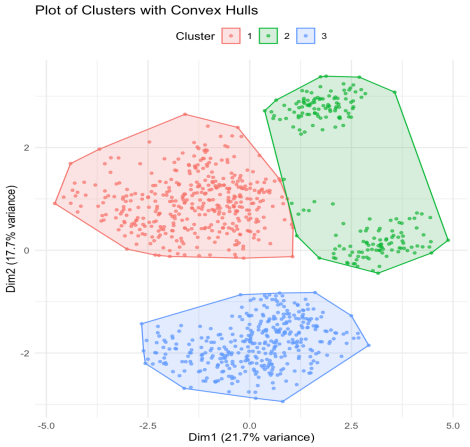


Figure 16: Plot of Clusters in 2D

Cluster Size:

After applying k-means clustering with K = 3, the resulting cluster distribution is as follows:

- Cluster 1: 402 members
- Cluster 2: 199 members
- Cluster 3: 372 members

This indicates that segmentation achieved a relatively balanced distribution of gym members across the three groups, ensuring meaningful insights for personalized fitness strategies.

PCA Summary– Contribution of PCs to Variance:

```
> summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13      PC14
Standard deviation     1.8036 1.6292 1.2535 1.16713 1.14494 1.14349 0.99585 0.98177 0.76599 0.72007 0.48235 0.37810 0.32292 1.293e-15
Proportion of Variance 0.2169 0.1769 0.1047 0.09081 0.08739 0.08717 0.06611 0.06426 0.03912 0.03457 0.01551 0.00953 0.00695 0.000e+00
Cumulative Proportion  0.2169 0.3938 0.4986 0.58939 0.67678 0.76395 0.83007 0.89432 0.93344 0.96801 0.98352 0.99305 1.00000 1.000e+00
                          PC15
Standard deviation     1.18e-15
Proportion of Variance 0.00e+00
Cumulative Proportion  1.00e+00
```

Figure 17: Summary of Principle Component Analysis

The PCA results show that the first principal component (PC1) explains 21.69% of the variance, and the second component (PC2) explains 17.69%, giving a combined variance explanation of approximately 39.38% for the first two components. The first 6 components collectively explain about 76.39% of the total variance. After the sixth component, each subsequent principal component contributes increasingly smaller proportions of variance, with contributions dropping below 4% after PC9. Thus, most of the meaningful information is concentrated in the first few components, and components beyond PC10 add minimal explanatory value.

Cluster Profiles Based on Clusters Centroids:

Considering the centroids of variables in each cluster, the members behavior and insights are retrieved

```
> kmeans_model$centers
          age gender_Female gender_Male workout_type_Cardio workout_type_HIIT workout_type_Strength workout_type_Yoga
1 -0.02805572   -0.04515177  0.04515177        -0.058853955        0.04556019           -0.06561468        0.08306572
2  0.04273294   -0.95035803  0.95035803        -0.007658324       -0.04336259            0.01918405        0.03035739
3 -0.03117084    1.05115358 -1.05115358         0.039759633        0.02248732            0.01436918       -0.07724126
  workout_intensity_score heart_intensity_ratio heart_rate_range hydration_need_ratio fat_percent experience_level         bmi
1               1.5215314           -0.01540050       0.02001682           -0.5260714 -1.57043960        1.5546085 -0.07909632
2              -0.4184430           -0.02595539       0.02155085            0.5743623  0.01781884       -0.4054604  0.34225038
3              -0.3617491            0.03628701      -0.03399675           -0.3392619  0.82084491       -0.3934731 -0.32753893
  calories_per_kg
1      0.98277140
2     -0.52191614
3      0.03827629
>
```

Figure 18: Centroids of selected variables in each cluster formed

Cluster 1

Feature Summary

- Age: Moderate-aged members, suggesting a balanced mix of young and mature individuals.
- Gender: Balanced distribution of males and females.
- Workout Type:
    - Moderate participation in Cardio activities.
    - HIIT is the most preferred workout among members.
    - Strength training has the least participation.
    - High engagement in Yoga sessions.
- Workout Metrics:
    - Highest Workout Intensity Score, indicating vigorous and frequent workout sessions.

    - Moderate Heart Intensity Ratio and Heart Rate Range, suggesting average cardiovascular endurance.
    - Lowest Hydration Need Ratio, implying efficient hydration habits during workouts.
- Body Composition:
    - Lowest Fat Percentage, highlighting a leaner group of individuals.
    - Moderate BMI, representing a healthy weight range.
- Experience:
    - Predominantly experienced members, indicating a skilled and consistent workout population.
- Calories Burned:
    - High Calories Burned per Kg, demonstrating effective workout performance relative to body weight.

Cluster 1 Member Insights:
Cluster 1 comprises moderately aged, well-experienced gym members who prefer high-intensity workouts, particularly HIIT and yoga. They maintain low body fat, healthy BMI levels, and show efficient hydration practices. Their high workout intensity and calorie-burning efficiency highlight strong fitness discipline and cardiovascular performance.

Cluster 2
Feature Summary
- Age: Highest average age, consisting mainly of older gym members.
- Gender: Lowest proportion of females and highest number of males.
- Workout Type:
    - Lowest participation in Cardio and HIIT workouts.
    - Strength training is the most dominant workout preference.
    - Moderate engagement in Yoga sessions.
- Workout Metrics:
    - Lowest Workout Intensity Score and Heart Intensity Ratio, reflecting less vigorous exercise patterns.
    - Highest Heart Rate Range, suggesting greater variation between resting and max BPM.
    - Highest Hydration Need Ratio, implying higher water requirements during workouts.
- Body Composition:
    - Moderate Fat Percentage.
    - Highest BMI, indicating a tendency toward overweight or obesity.
- Experience:
    - Mostly low-experience members, suggesting relative newness to fitness routines.
- Calories Burned:
    - Lowest Calories Burned per Kg, pointing to less efficient energy expenditure during workouts.

Cluster 2 Member Insights:
Cluster 2 is primarily composed of older, less experienced male members who prefer strength training over cardio or HIIT. They exhibit higher body mass and hydration needs, combined with lower workout intensity and calorie-burning efficiency. This cluster may benefit from personalized fitness programs to boost cardiovascular health and optimize weight management.

Cluster 3
Feature Summary
- Age: Youngest group of gym members, with the lowest average age.
- Gender: Highest proportion of females and the fewest males.
- Workout Type:
    - Cardio workouts are the most popular activity.
    - Moderate participation in both HIIT and Strength training.
    - Least engagement in Yoga sessions.

- Workout Metrics:
  - Moderate Workout Intensity Score, indicating a balanced effort level.
  - Highest Heart Intensity Ratio, suggesting members push closer to their maximum heart rate during exercise.
  - Lowest Heart Rate Range, implying smaller differences between resting and max BPM.
- Body Composition:
  - Highest Fat Percentage compared to other clusters.
  - Lower BMI levels overall.
- Experience:
  - Members generally have moderate fitness experience.
- Calories Burned:
  - Moderate Calories Burned per Kg, reflecting average workout efficiency.

Cluster 3 Member Insights:
Cluster 3 predominantly consists of younger female gym members who favor cardio workouts. They exhibit high effort during sessions (high heart intensity ratio) but maintain a relatively smaller heart rate range. Despite having a higher fat percentage, they generally have lower BMI values. With moderate fitness experience, these members could benefit from structured strength training and conditioning programs to further enhance body composition and endurance.

Profile Bar plot of Cluster Centroids:
The bar chart represents the average scaled feature values for each cluster, helping to understand how different attributes contribute to each group in the fitness data.

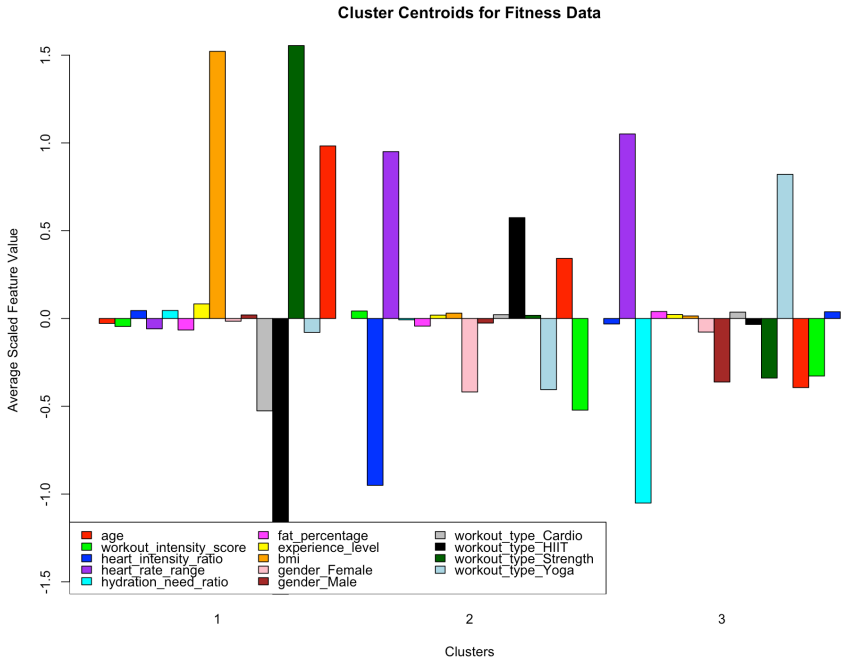


Figure 19: Figure 6: Scaled Centroid Profiles for Each Fitness Segment Across Key Features

Multiple Linear Regression for Calorie Prediction:

The Multiple Linear Regression predicts “calories_burned” based on consumer demographics, workout and health metrics.

Variables selected:

‘age’, ‘fat_percent’, ‘water_intake’, ‘workout_frequency’, ‘experience_level’, ‘bmi’, ‘heart_rate_range’, ‘heart_intensity_ratio’, ‘hydration_need_ratio’, ‘gender_Female’, ‘calories_burned’.

Model Fitting Overview

- The model was trained using 684 observations from the training dataset.
- The model was fitted using Ordinary Least Squares (OLS) regression, aiming to predict calories burned based on fitness and demographic attributes.

Model Summary

Figure 20 presents the performance overview of the calorie-prediction regression model, highlighting its explanatory power and statistical significance:

- **R-squared** = 0.8038: About **80.38%** of the variance in calories burned is explained by the model.
- **Adjusted R-squared** = 0.8022: Indicates a strong fit, even after adjusting for the number of predictors.
- **F-statistic** = 493.8, **p-value** < 2.2e-16: The overall model is highly statistically significant.

Key Significant Predictors (p < 0.05)

- **Gender (Female)**: Negative effect (-128.06); females burn fewer calories compared to males.
- **Fat Percentage**: Negative impact (-16.77); higher fat percentage is associated with fewer calories burned.
- **Workout Frequency**: Positive impact (+46.36); more frequent workouts lead to higher calorie expenditure.
- **Heart Rate Range**: Positive effect (+3.55); a higher heart rate range boosts calorie burns.
- **Heart Intensity Ratio**: Strong positive influence (+994.78); intense cardiovascular effort greatly increases calories burned.
- **Hydration Need Ratio**: Strong negative impact (-181.17); higher hydration needs correlate with lower calorie burn.
- **Age**: Slight negative relationship (-3.27); older individuals tend to burn slightly fewer calories.

Non-Significant Predictor

- **BMI**: The coefficient for BMI is not statistically significant (p > 0.05), suggesting BMI alone is not a strong predictor for calorie burn in this dataset.

```
Call:
lm(formula = calories_burned ~ gender_Female + fat_percent +
    workout_frequency + age + bmi + heart_rate_range + heart_intensity_ratio +
    hydration_need_ratio, data = regression_data)

Residuals:
    Min      1Q  Median      3Q     Max
-364.57  -85.48    6.10   81.30  350.65

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            540.3259    79.1061   6.830 1.50e-11 ***
gender_Female         -128.0604    10.1380 -12.632  < 2e-16 ***
fat_percent            -16.7666     0.8486 -19.759  < 2e-16 ***
workout_frequency       46.6494     5.6268   8.291 3.76e-16 ***
age                     -3.2743     0.3202 -10.225  < 2e-16 ***
bmi                      0.3042     0.6170   0.493    0.622
heart_rate_range         3.5549     0.3361  10.576  < 2e-16 ***
heart_intensity_ratio  994.7878    46.1828  21.540  < 2e-16 ***
hydration_need_ratio  -181.1711     5.4524 -33.228  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 121.3 on 964 degrees of freedom
Multiple R-squared:  0.8038,    Adjusted R-squared:  0.8022
F-statistic: 493.8 on 8 and 964 DF,  p-value: < 2.2e-16
```

Figure 20: Linear Regression summary for calorie prediction

Model Performance Evaluation

The regression model's predictive performance on validation and test sets, as measured by RMSE and R^2 as shown in Figure 21.

Validation Metrics:

- **Validation RMSE (Root Mean Squared Error): 118.03**
  The model's average prediction error on the validation set is 118.03 calories, indicating good accuracy.
- **Validation R-squared: 0.794**
  Approximately 79.4% of the variance in calorie expenditure is explained by the model on the validation data.

Test Metrics:

- **Test RMSE: 116.52**
  The test dataset shows an even lower average prediction error of 116.52 calories, suggesting strong generalization.
- **Test R-squared: 0.810**
  About 81.0% of the variability in calorie expenditure is captured by the model on unseen test data, confirming model robustness and reliability.

The R-squared values show a strong relationship between predictors and calorie expenditure, while the RMSE values indicate a relatively low prediction error.
Overall, the model maintains consistent and reliable performance across both validation and test datasets.

Dataset	RMSE	R_squared
Validation	118.028	0.794
Test	116.519	0.810

Table 6: Validation and Test Performance of the Calorie Prediction Regression Model

Cost Analysis

Residual analysis shows the model under-estimated total calories by 6,520 kcal, accounting for 65% of the overall prediction risk, while over-estimates summed to 6,950 kcal (35%).

- Assuming a risk weight of 2 units per kcal for under-forecasting and 1 unit per kcal for over-forecasting, we translate these errors into a combined risk score that drives our model tuning.

Error Type	Total kcal error	Risk per kcal	Risk Score
Under-estimate	6,520.63	$2	$13,041.26
Over-estimate	6,950.24	$1	$6,950.24
Total Risk			$19,991.50

Table 7: Cost Analysis for Calorie Prediction Regression Model

The model currently underestimates calories more often than it overestimates, and those low-side errors contribute most of the total risk score. This suggests our next efforts should focus on reducing under-predictions to bring down overall risk.

Business Recommendations

- **Performance-driven members (Cluster 1)**
  Run advanced HIIT challenges with live leaderboards and bundle premium recovery services to maintain motivation and leverage their high workout discipline.
- **Strength-focused starters (Cluster 2)**
  Introduce low-impact cardio-plus-strength programs with small-group technique coaching, paired with hydration and nutrition workshops to improve cardiovascular health and manage higher BMI.
- **Cardio-centric young movers (Cluster 3)** - Offer fusion "tone-and-burn" classes that blend light resistance into cardio sessions, backed by social workout challenges and streak rewards to foster consistent attendance.
- All clusters' members- Deploy the calorie-prediction tool in the app for personalized post-workout feedback and schedule segment-specific push notifications (recovery tips, form videos, class invites) to keep engagement high.

Future Work

- **Integrate richer data**
  Incorporate sleep quality, stress levels, and nutrition logs from wearables and food-tracking apps to capture a fuller picture of recovery and lifestyle factors.
  Automate continuous retraining: Build a data pipeline that ingests new workout records, seasonal program changes, and equipment updates, then retrains the model on time-based (like weekly) or when performance drifts.
- **Adopt cost-aware, adaptive models**
  Experiment with gradient-boosted ensembles that weight business outcomes (retention value vs. churn risk) directly in the loss function, and retrain models continuously as new information streams in.

Observations and Conclusion:

Observations

- **Three member segments revealed:** High-intensity performers, strength-focused older starters, and young, cardio-centric movers.
- **Top calorie-burn drivers:** Workout frequency, peak heart-rate effort, and body-fat percentage (with age/BMI as minor factors).
- **Model reliability confirmed:** Regression shows strong fit and low error on new data, so real-time calorie feedback can be trusted.
- **Tailored programming & engagement:** Advanced HIIT for performers, guided strength-plus-cardio tracks for starters, and social "tone-and-burn" sessions for movers—each reinforced by in-app burn updates and segment-specific notifications.

Conclusion

This analysis established a clear roadmap for Fit Life Wellness. The segmentation and regression analysis delivered compelling results: three clear member groups emerged, the calorie-burn model achieved strong accuracy (R^2 = 0.80, RMSE = 116 kcal), and the cost-weighted review highlighted exactly where to fine-tune. Together, these findings showcase a robust, business-driven approach that provides reliable calorie estimates and targeted guidance for ongoing model improvement, positioning our fitness application to confidently support and engage members.

Consumer Behavior Analytics

Table Of Contents

Introduction

Shop-Smart, a retail company specializing in premium food and beverage products. The company has collected data on customer demographics, purchasing history, online activity, and responses to previous marketing campaigns. Using the available data, shop-smart wants to understand their consumers behavior and improve their marketing strategies by uncovering behavioral patterns and segment consumers and develop targeted marketing strategies that enhance campaign effectiveness and customer engagement.

Business Opportunity and Goals

Business Opportunity

Shop-Smart is facing challenges in running effective marketing campaigns due to a lack of targeted promotional strategies. This results in low campaign performance, and weak customer engagement.

Business Goal

The goal is to strengthen Shop-Smart's marketing effectiveness by developing data-driven, targeted strategies tailored to distinct consumer segments.

Business Objectives:

- Identifying different consumers segments based on demographics and their purchasing behaviors to create targeted marketing campaigns.
- Analyze which consumers are responsive for marketing based on the success of past campaigns and identify the factors contributing to success or failure.
- Recommend different marketing strategies for different segments.

Analytical Goal

To develop models that uncover patterns in purchasing behavior, classify consumers with their response to marketing campaigns, and predict marketing engagement.

Analytical Objectives:

- Perform consumer segmentation
- Develop classification models for predicting consumer response for marketing campaigns.
- Develop regression models to predict consumer engagement

Analytical Approach

Data on customer demographics, purchase history, and campaign responses was collected and explored through visualizations and summary statistics to uncover initial patterns. After cleaning and preprocessing (handling missing values, data types, outliers, etc.), new features were engineered and key predictors selected. For segmentation, unsupervised clustering methods were applied to group customers by purchasing behavior and engagement, informing more targeted marketing strategies. Next, supervised models were developed, including classification to predict campaign response and regression to estimate engagement, evaluated using metrics such as accuracy and RMSE. Finally, cluster insights and predictive results were used to propose tailored marketing campaigns for each segment.

Data Exploration and Preprocessing

Data Understanding

Data Collection:

- The Consumer dataset consists of 2,240 records and contains 29 variables.
- Dataset includes detailed information on consumer demographics, their purchasing and spending behavior, campaigns response status, and other additional variables retrieved from last 6 months.

Demographic Variables:

- 'Year_Birth', 'Education', 'Marital_Status', 'Income', 'Kidhome', 'Teenhome'

Purchase & Spending Behavior:

- Amounts spend on different product: 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds'.
- Purchase channel frequency: 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases'.
- Website activity: 'NumWebVisitsMonth', 'Recency'

Campaign Response variables:

- Campaign response: 'AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5'.
- Last campaign status: Response

Additional variables:

- 'Dt_Customer' (enrollment date), 'Z_CostContact', 'Z_Revenue', 'Complain'

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases
5524	1957	Graduation	Single	58138	0	0	04-09-2012	58	635	88	546	172	88	88	3
2174	1954	Graduation	Single	46344	1	1	08-03-2014	38	11	1	6	2	1	6	2
4141	1965	Graduation	Together	71613	0	0	21-08-2013	26	426	49	127	111	21	42	1
6182	1984	Graduation	Together	26646	1	0	10-02-2014	26	11	4	20	10	3	5	2
5324	1981	PhD	Married	58293	1	0	19-01-2014	94	173	43	118	46	27	15	5
7446	1967	Master	Together	62513	0	1	09-09-2013	16	520	42	98	0	42	14	2

NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
8	10	4	7	0	0	0	0	0	0	3	11	1
1	1	2	5	0	0	0	0	0	0	3	11	0
8	2	10	4	0	0	0	0	0	0	3	11	0
2	0	4	6	0	0	0	0	0	0	3	11	0
5	3	6	5	0	0	0	0	0	0	3	11	0
6	4	10	6	0	0	0	0	0	0	3	11	0

Table 1: First six records of the consumer dataset

Variables Definition:

The attributes listed below in Table 2 represent key customer-related features captured in the dataset, along with their respective data types and contextual definitions.

Variable name	Data type (given)	Definition
ID	int	Unique Identifier of a consumer
Year_Birth	chr	Birth year of the consumer
Education	chr	Highest education of the consumer
Marital_Status	int	Marital status of the consumer
Income	int	Yearly income of consumer
Kidhome	int	Number of children in consumer' home
Teenhome	chr	Number of teenagers at consumer's home
Dt_Customer	int	Date of consumer's enrollment with the company
Recency	int	Number of days since consumer's most recent purchase
MntWines	int	Amount spent on wines
MntFruits	int	Amount spent on fruits
MntMeatProducts	int	Amount spent on meat
MntFishProducts	int	Amount spent on fish products
MntSweetProducts	int	Amount spent on sweets
MntGoldProds	int	Amount spent on gold
NumDealsPurchases	int	Number of purchases made using a discount or deal
NumWebPurchases	int	Number of purchases made through the website
NumCatalogPurchases	int	Number of purchases made using a catalog
NumStorePurchases	int	Number of purchases made in physical store
NumWebVisitsMonth	int	Number of visits to the company's website in the past 6 months
AcceptedCmp3	int	Indicates whether the consumer accepted offer from campaign 3
AcceptedCmp4	int	Indicates whether the consumer accepted offer from campaign 4
AcceptedCmp5	int	Indicates whether the consumer accepted offer from campaign 5
AcceptedCmp1	int	Indicates whether the consumer accepted offer from campaign 1
AcceptedCmp2	int	Indicates whether the consumer accepted offer from campaign 2
Complain	int	Indicates whether the consumer has made a complaint
Z_Costcontact	int	Internal company code for cost of contacting the customer
Z_Revenue	int	Internal company code for revenue from contacting the customer
Response	int	1 if the consumer accepted the most recent campaign offer, if not 0

Table 2: Description of Variables with Original Data Types and Definitions

Data types:

All the variables are either character or integer type. However, these are categorized as follows:

Numeric data:

- ‘id’, ‘year_birth’, ‘income’, ‘kidhome’, ‘teenhome’, ‘recency’,
  ‘mntwines’, ‘mntfruits’, ‘mntmeatproducts’, ‘mntfishproducts’, ‘mntsweetproducts’, ‘mntgoldprods’,
  ‘numdealspurchases’, ‘numwebpurchases’, ‘numcatalogpurchases’, ‘numstorepurchases’,
  ‘numwebvisitsmonth’,
  ‘acceptedcmp1’, ‘acceptedcmp2’, ‘acceptedcmp3’, ‘acceptedcmp4’, ‘acceptedcmp5’,
  ‘complain’, ‘z_costcontact’, ‘z_revenue’, ‘response’,

Categorical data

- ‘education’, ‘marital_status’, ‘dt_consumer’.

Data Preprocessing

Renaming column names:

To ensure consistency, readability, and ease of reference during analysis, the original column names were renamed using simplified and standardized naming conventions.

Original column name	New column name
ID	id
Year_Birth	birth_year
Education	education
Marital_Status	marital_status
Income	income
Kidhome	kids_at_home
Teenhome	teens_at_home
Dt_Customer	join_date_customer
Recency	days_since_last_purchase
MntWines	spending_on_wines
MntFruits	spending_on_fruits
MntMeatProducts	spending_on_meat
MntFishProducts	spending_on_fish
MntSweetProducts	spending_on_sweets
MntGoldProds	spending_on_goldproducts
NumDealsPurchases	num_of_deals_purchases
NumWebPurchases	num_of_web_purchases
NumCatalogPurchases	num_of_catalog_purchases
NumStorePurchases	num_of_store_purchases
NumWebVisitsMonth	website_visits_per_month
AcceptedCmp3	accepted_camp_3
AcceptedCmp4	accepted_camp_4
AcceptedCmp5	accepted_camp_5
AcceptedCmp1	accepted_camp_1
AcceptedCmp2	accepted_camp_2
Complain	complain
Z_Costcontact	z_costrevenue
Z_Revenue	z_costcontact
Response	last_campaign_response

Table 3: Original vs. Renamed Column Names of Consumer dataset

Data type conversion:

As part of preprocessing, appropriate data types were assigned to each variable to ensure accurate analysis, including conversion of character columns to categorical, integers to factors where applicable, and formatting date fields correctly.

Variable Name	Original Data type	Updated Data type	Reason
birth_year	int	date	year is the part of date
education	chr	categorical	Contains different categories like "Graduate", "PhD"
marital_status	chr	categorical	categories like "single", "married"
income	int	numeric	continuous value
kids_at_home	int	integer	No change as it is a Count variable
teens_at_home	int	integer	No change as it is a Count variable
join_date_customer	chr	date	representing a date
days_since_last_purchase	int	integer	No change as it is a Count variable
spending_on_wines	int	numeric	Spending amount is continuous
spending_on_fruits	int	numeric	Spending amount is continuous
days_since_last_purchase	int	numeric	Spending amount is continuous
spending_on_fish	int	numeric	Spending amount is continuous
spending_on_sweets	int	numeric	Spending amount is continuous
spending_on_goldproducts	int	numeric	Spending amount is continuous
num_of_deals_purchases	int	integer	No change as it is a Count variable
num_of_web_purchases	int	integer	No change as it is a Count variable
num_of_catalog_purchases	int	integer	No change as it is a Count variable
num_of_store_purchases	int	integer	No change as it is a Count variable
website_visits_per_month	int	integer	No change as it is a Count variable
accepted_camp_3	int	categorical	Binary variable having two categories (1= yes, 0 = no)
accepted_camp_4	int	categorical	Binary variable having two categories (1= yes, 0 = no)
accepted_camp_5	int	categorical	Binary variable having two categories (1= yes, 0 = no)
accepted_camp_1	int	categorical	Binary variable having two categories (1= yes, 0 = no)
accepted_camp_2	int	categorical	Binary variable having two categories (1= yes, 0 = no)
complain	int	categorical	Binary variable having two categories (1= yes, 0 = no)
z_costrevenue	int	numeric	continuous revenue value
z_costcontact	int	numeric	continuous contact cost
last_campaign_response	int	categorical	Binary variable having two categories (1= yes, 0 = no)

Table 4: Data Type Adjustments of consumer data

Handling Missing and Zero values:

Missing values:
As shown in Table 4, the consumer dataset has no missing values in most of the columns, except for the 'income' field, which has 24 missing values. This comprises of 1.07% of the total observations.
Missing values Handling:
To preserve data integrity and avoid introducing artificial values, all rows with missing 'income' entries were removed during preprocessing.
Zero values handling:
The presence of zero values in various variables provides meaningful insights about customer characteristics and behavior. For instance,

- 1293 of the consumers has no children their household, and 1158 of the consumers has no teenagers in their home.
- In product-related variables such as 'spending_on_wines', 'spending_on_fruits', 'spending_on_meat', 'spending_on_fish', 'spending_on_sweets', and 'spending_on_goldproducts', zero values indicate that no purchases were made by those consumers in the respective categories.
- Similarly, in campaign response variables (accepted_camp_1 through accepted_camp_5), zeros denote that the customer did not respond to or accept the corresponding campaign.

Table: Missing and Zero Value Counts for Each Variable

Variable	Missing Values	Zero Values
id	0	1
birth_year	0	77
education	0	0
marital_status	0	0
income	24	0
kids_at_home	0	1293
teens_at_home	0	1158
join_date_customer	0	0
days_since_last_purchase	0	28
spending_on_wines	0	13
spending_on_fruits	0	400
spending_on_meat	0	1
spending_on_fish	0	384
spending_on_sweets	0	419
spending_on_goldproducts	0	61
num_of_deals_purchases	0	46
num_of_web_purchases	0	49
num_of_catalog_purchases	0	586
num_of_store_purchases	0	15
website_visits_per_month	0	11
accepted_camp_3	0	2077
accepted_camp_4	0	2073
accepted_camp_5	0	2077
accepted_camp_1	0	2096
accepted_camp_2	0	2210
complain	0	2219
z_costcontact	0	0
z_revenue	0	0
last_campaign_response	0	1906

Table 5: Missing and Zero Value Counts for Each Variable

Outliers' detection and Handling:

Outliers in numeric variables:

- Outlier detection was performed on all numeric variables using the IQR method. A high number of outliers were identified in spending-related variables such as 'spending_on_fruits', 'spending_on_meat', and 'spending_on_goldproducts'. These values likely represent high-spending consumers, which are important for understanding consumer behavior and segmentation. Therefore, these outliers were retained for analysis.
- In the 'income' variable, most values fell within a reasonable range. However, one extreme value (666666) was identified as unrealistic and likely a placeholder. This record was removed from the dataset to ensure data quality.

Variable	Outlier Count
id	0
income	8
kids_at_home	0
teens_at_home	0
days_since_last_purchase	0
spending_on_wines	35
spending_on_fruits	227
spending_on_meat	175
spending_on_fish	223
spending_on_sweets	248
spending_on_goldproducts	207
num_of_deals_purchases	86
num_of_web_purchases	4
num_of_catalog_purchases	23
num_of_store_purchases	0
website_visits_per_month	8
z_costcontact	0
z_revenue	0

Table 6: Outlier Count in Each Numeric Variable

Outliers in date columns:

- For the 'birth_year' variable, three outliers were detected using the IQR method. Upon inspection, these records had birth years prior to 1937, which were deemed unrealistic in the context of the dataset. These entries were removed to maintain data accuracy.
- The 'join_date_customer' variable was also examined by extracting the year component. All customers join dates were within the expected range of 2012 to 2014, with no anomalies observed.

```
> birthyear_outliers
[1] 3

> join_dates

2012 2013 2014
 487 1173  552
```

Figure 1: Outlier Count in date columns

Standardizing Categorical Values:

The categorical variables 'education', 'income_bracket' and 'marital_status' are converted into numeric by assigning labels for modeling.

Education:

The original education variable included multiple levels: "Basic", "2n Cycle", "Graduation", "Master", and "PhD". To simplify the analysis and reduce redundancy, the values were grouped as follows:

- "2n Cycle", "Graduation", and "Master" were combined under the label "Graduate" since they all represent comparable levels of higher education.
- "Basic" was retained as a standalone category as it clearly indicates a lower level of formal education and "PhD" was kept separate to represent the highest level of academic qualification.

Income:

- The income_bracket variable was ordinally encoded to reflect increasing income levels: "Low" as 0, "Medium" as 1, and "High" as 2.
- This numeric transformation preserves the order of income levels and makes the variable suitable for use in machine learning models.

Marital status:

The original 'marital_status' variable contained eight values: "Married", "Together", "Single", "Alone", "Divorced", "Widow", "YOLO", and "Absurd". To enhance data quality and ensure meaningful analysis, the values were cleaned and grouped as follows:

- "YOLO" and "Absurd" were removed as they are invalid and do not represent genuine marital statuses.
- "Single" and "Alone" were combined into "Single", as both reflect individuals living without a partner.
- "Married" and "Together" were grouped under "Couple", reflecting a partnered status regardless of legal marriage.
- "Divorced" and "Widow" were retained as distinct categories due to their unique implications on lifestyle and behavior.

Summary Statistics for Numeric Variables:

The summary statistics provide an overview of the distribution of continuous and discrete numeric variables in the dataset.

- 'Income' ranges from 1,736 to 162,397, with a mean of approximately 51,944 and median of 51,371, indicating a balanced income distribution.
- Children-related variables such as 'kids_at_home' and 'teens_at_home' are limited to a maximum of 2, with most households having 0 or 1 child/teen.
- Spending variables show high variability and right-skewed distributions, for example, 'spending_on_wines' has a maximum value of 1,493, much higher than its median of 174.
- Purchase mode variables mostly have low to moderate ranges, with most values concentrated under 10.
- Website visits per month vary between 0 and 20, with a median of 6 visits.

```
+   summary()
     income        kids_at_home    teens_at_home    days_since_last_purchase spending_on_wines spending_on_fruits spending_on_meat spending_on_fish spending_on_sweets spending_on_goldproducts
 Min.   :  1730   Min.   :0.0000   Min.   :0.0000   Min.   : 0.00            Min.   :   0.0    Min.   :  0.00     Min.   :   0.0   Min.   :  0.00   Min.   :  0.00     Min.   :  0.00
 1st Qu.: 35196   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:24.00            1st Qu.:  24.0    1st Qu.:  1.75     1st Qu.:  16.0   1st Qu.:  3.00   1st Qu.:  1.00     1st Qu.:  9.00
 Median : 51371   Median :0.0000   Median :0.0000   Median :49.00            Median : 174.0    Median :  8.00     Median :  68.0   Median : 12.00   Median :  8.00     Median : 24.00
 Mean   : 51944   Mean   :0.4425   Mean   :0.5059   Mean   :49.06            Mean   : 305.2    Mean   : 26.30     Mean   : 167.0   Mean   : 37.53   Mean   : 27.07     Mean   : 43.78
 3rd Qu.: 68487   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:74.00            3rd Qu.: 505.5    3rd Qu.: 33.00     3rd Qu.: 232.2   3rd Qu.: 50.00   3rd Qu.: 33.00     3rd Qu.: 56.00
 Max.   :162397   Max.   :2.0000   Max.   :2.0000   Max.   :99.00            Max.   :1493.0    Max.   :199.00     Max.   :1725.0   Max.   :259.00   Max.   :262.00     Max.   :321.00
 num_of_deals_purchases num_of_web_purchases num_of_catalog_purchases num_of_store_purchases website_visits_per_month
 Min.   : 0.000         Min.   : 0.000       Min.   : 0.000           Min.   : 0.000         Min.   : 0.000
 1st Qu.: 1.000         1st Qu.: 2.000       1st Qu.: 0.000           1st Qu.: 3.000         1st Qu.: 3.000
 Median : 2.000         Median : 4.000       Median : 2.000           Median : 5.000         Median : 6.000
 Mean   : 2.322         Mean   : 4.086       Mean   : 2.669           Mean   : 5.806         Mean   : 5.322
 3rd Qu.: 3.000         3rd Qu.: 6.000       3rd Qu.: 4.000           3rd Qu.: 8.000         3rd Qu.: 7.000
 Max.   :15.000         Max.   :27.000       Max.   :28.000           Max.   :13.000         Max.   :20.000
```

Figure 2: Summary Statistics for Numeric Variables

Variable Distributions:

The distribution of each variable was explored through various visualizations. These graphical summaries helped uncover skewness, spot outliers, and identify meaningful patterns in the data, ensuring readiness for further analysis and modeling.

Demographics:

Year of Birth

Most customers were born between the early 1950s and late 1970s, with a median around 1965. A few older individuals are observed, but no significant outliers remain after cleaning.

Income

Customer income ranges widely, with a median around 51,000. A few high-income outliers above 150,000 suggest the presence of prosperous consumers.

Kid and Teen Count

Most households have 0 to 1 child or teen, with very few having the maximum of 2. The distribution is right-skewed but has no outliers.

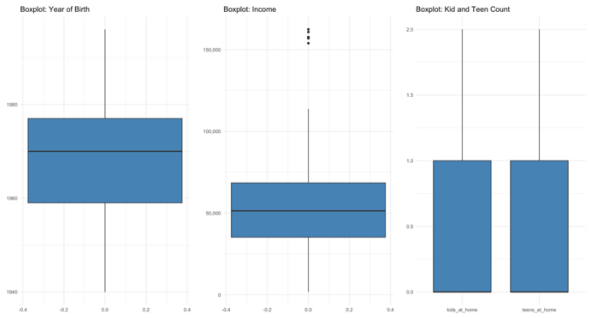


Figure 3: Distribution of Demographic Variables

Education

The majority of customers hold a graduate-level education (1,676), followed by PhD holders (478). Only a small fraction (54) has basic education, indicating a generally well-educated customer base.

Marital Status

Most customers are in a relationship (1,428 labeled as couples), while singles (473) and divorced individuals (231) form smaller segments. Widowed customers make up the smallest group with just 76 respondents.

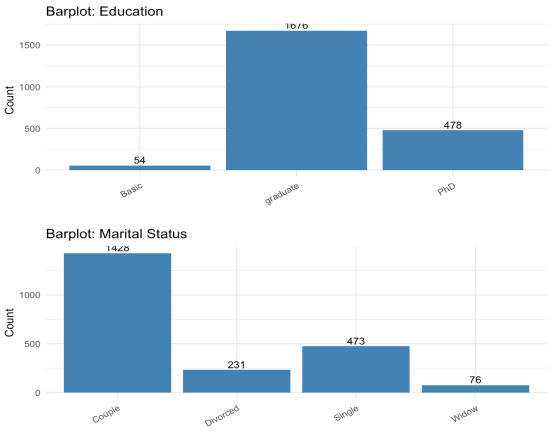


Figure 4: Distribution of Categorical Demographic Attributes

Days since last purchase distribution:
The distribution is uniform, showing varied engagement levels across the consumer base. There are consumers who purchased recently and who did not purchase in a while.

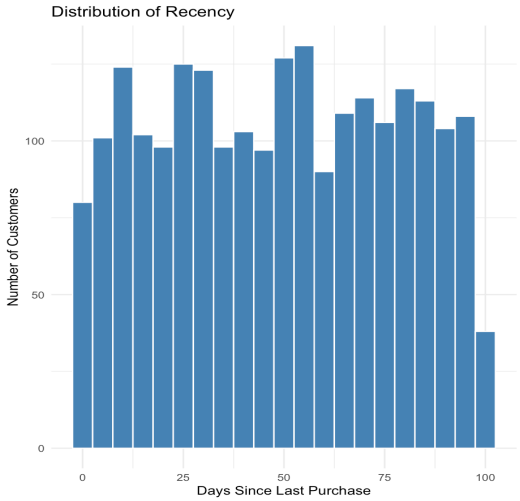


Figure 5: Bar plot for Days Since Last Purchase

Spending distributions:

- Spending across categories is heavily right skewed, with most customers making minimal purchases, while a few exhibit very high spending, especially on wines and meat.
- Low and concentrated spending in categories like fruits, fish, sweets, and gold products indicates infrequent or selective buying behavior in these segments.

Figure 6: Spending Distribution by Product Category

Purchase mode distribution

- Most purchasing modes show right-skewed distributions, indicating that many consumers make relatively few purchases across catalogs, web, stores, and deals.
- Website visits per month display a peak between 5 and 8 visits, suggesting that while actual purchases are low, a moderate number of customers actively engage with the website frequently.

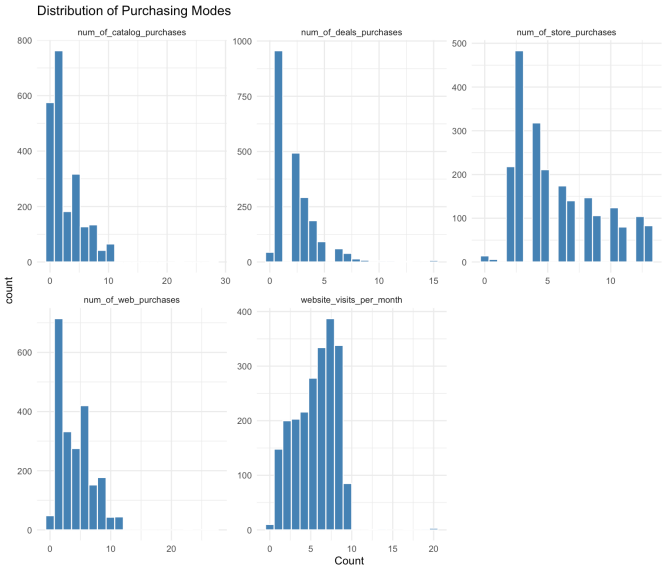


Figure 7: Bar Plot for Purchasing Modes distribution

Campaign responses

- Across all five campaigns and the last campaign response, many customers did not accept the offers, with acceptances being notably low, especially in Campaign 2 with only 30 positive responses.
- The last campaign had the highest number of acceptances (331), suggesting possible improvements in targeting or messaging over time

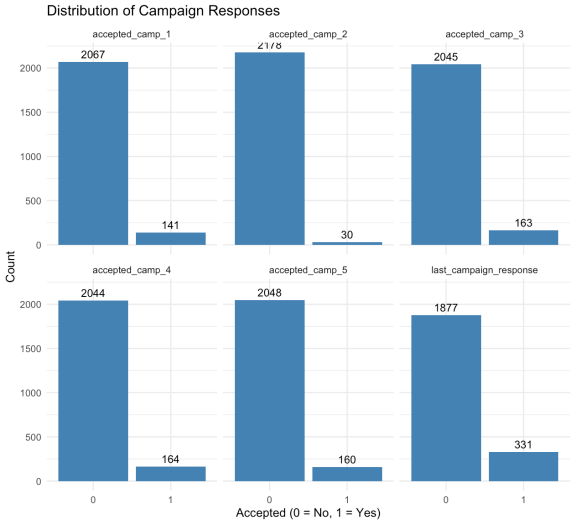


Figure 8: Distribution of Responses to Marketing Campaigns

Feature Engineering and transformation

Feature Creation:

The following transformations and feature creations were applied to enrich the dataset and prepare it for modeling:

1. **Join year:**
   Derived a new variable 'join_year' from the 'join_date_customer' field to analyze customer behavior over years.

2. **birth_year:**
   Extracted numeric year from birth_year (which was a Date type), and used to calculate customer age-related metrics

3. **age_at_enrollment**
   Calculated by subtracting 'birth_year' from 'join_year'. This captures the age of the customer when they joined the platform.

4. **Children:**
   Sum of 'kids_at_home' and 'teens_at_home', represents total number of dependents in a household.

5. total_monetary_value:
   Sum of all product spending variables (spending_on_wines + spending_on_fruits + spending_on_meat + spending_on_fish + spending_on_sweets + spending_on_goldproducts) represents overall spending behavior

6. **total_purchase_count:**
   Aggregated count from four purchasing modes, and measures total customer purchasing activity.

7. **Outlier variables:**
   Created individual binary flags for outliers in each key numeric variable (e.g., 'spending_on_wines_outlier', 'num_of_deals_purchases_outlier', etc.) using IQR. This marks specific variable-wise outliers for analysis and modeling.

8. **special_customer:**
   Flagged customers who are outliers in any of the key numeric variables using the IQR method.
   Helps in identifying unusual or extreme customer profiles.

9. **campaign_response:**
   Aggregated binary variable. Set to 1 if a customer accepted **any** of the five marketing campaigns or the last campaign. Acts as a consolidated response target for modeling.

10. **Income_bracket:**
    Categorizes customers into income groups based on their income distribution.
    Created using 33rd and 67th percentiles of income:
    - Low: ≤ 40,240.15
    - Medium: > 40,240.15 and ≤ 62,986.85
    - High: > 62,986.85

Checking for Missing Values after transformation:

The missing value counts representing there are no missing value after adding new features and could be used for further analysis.

```
> missing_counts
                                 id                           birth_year
                                  0                                    0
                          education                       marital_status
                                  0                                    0
                             income                       kids_at_home
                                  0                                    0
                       teens_at_home                  join_date_customer
                                  0                                    0
             days_since_last_purchase                   spending_on_wines
                                  0                                    0
                  spending_on_fruits                    spending_on_meat
                                  0                                    0
                  spending_on_fish                   spending_on_sweets
                                  0                                    0
          spending_on_goldproducts                num_of_deals_purchases
                                  0                                    0
              num_of_web_purchases            num_of_catalog_purchases
                                  0                                    0
            num_of_store_purchases           website_visits_per_month
                                  0                                    0
                    accepted_camp_3                     accepted_camp_4
                                  0                                    0
                    accepted_camp_5                     accepted_camp_1
                                  0                                    0
                    accepted_camp_2                             complain
                                  0                                    0
                   z_costcontact                            z_revenue
                                  0                                    0
              last_campaign_response                          join_year
                                  0                                    0
        spending_on_wines_outlier          spending_on_fruits_outlier
                                  0                                    0
         spending_on_meat_outlier           spending_on_fish_outlier
                                  0                                    0
       spending_on_sweets_outlier     spending_on_goldproducts_outlier
                                  0                                    0
   num_of_deals_purchases_outlier         num_of_web_purchases_outlier
                                  0                                    0
num_of_catalog_purchases_outlier    website_visits_per_month_outlier
                                  0                                    0
                  special_customer                      income_bracket
                                  0                                    0
>
```

Figure 9: Missing value check after transformation

Predictor relevancy

Correlation matrix:

The heatmap illustrates Pearson correlation coefficients between all numeric variables, highlighting strong linear relationships useful for feature selection and multicollinearity checks.

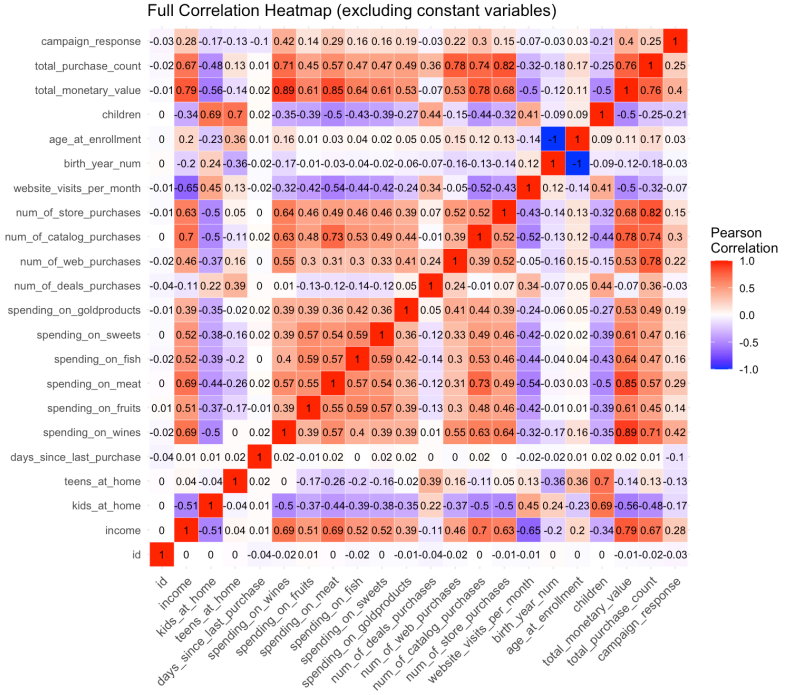


Figure 10: Correlation Matrix of Numeric variables

Variable relevancy for clustering:

Predictor relevancy is not directly applicable to clustering since there is no target variable. However, data exploration has already been performed to assess the distribution and variance of each variable. The analysis highlighted that variable such as income, purchase frequency recency, and spending behavior are crucial for understanding consumer behavior, while demographic features like education, marital status, and income bracket also provide valuable insights for effective segmentation.

Variable relevancy for Classification:

Variable relevancy is examined for classifying marketing campaign response. Statistical and visual analyses such as bar charts, boxplots, and density plots are employed to assess how each variable contributes to distinguishing responders from non-responders.

a. Demographics vs campaign response:

The figure shows the distribution of categorical variables, such as education, marital status, income bracket, join year, and complain status, across campaign responders and non-responders. Also, a boxplot compares the continuous variable age at enrollment between responders and non-responders.

Key Insights:

- Customers with graduate or PhD education respond better to campaigns than those with only basic education.
- Single and widowed individuals show higher response rates compared to couples.
- High-income consumers are more engaged with marketing efforts.
- Customers who have filed complaints are less likely to respond to campaigns.
- Consumers who enrolled in 2012 demonstrate slightly higher responsiveness.
- Households with fewer children are more likely to respond to marketing campaigns.
- Younger customers show greater engagement with campaigns compared to older ones.

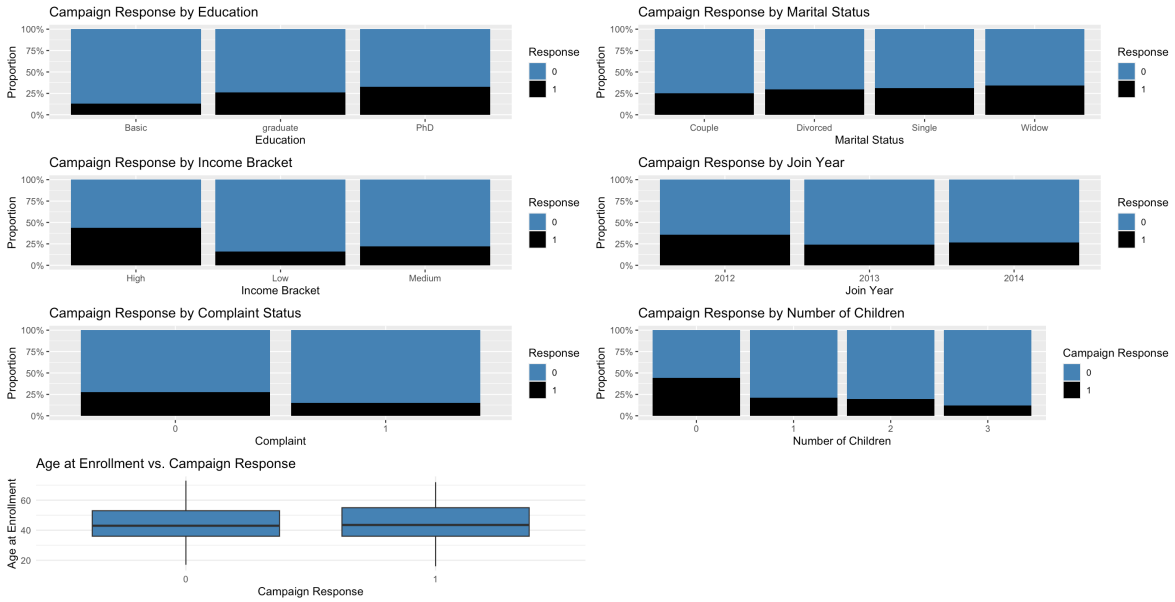


Figure 11: Demographics against campaign response

b. Purchase Mode by Campaign Response:

- Customers with more in-store purchases are more likely to respond to campaigns, showing a strong preference for traditional shopping experiences.
- Web purchases also correlate positively with campaign response, indicating digital engagement plays a role in influencing responders.

- Catalog purchasers show moderate campaign responsiveness, while
- Customers who frequently buy through deals tend to respond less, suggesting they may be driven more by price than promotional campaigns.
- Website visits are slightly higher among non-responders, possibly reflecting browsing behavior without conversion.

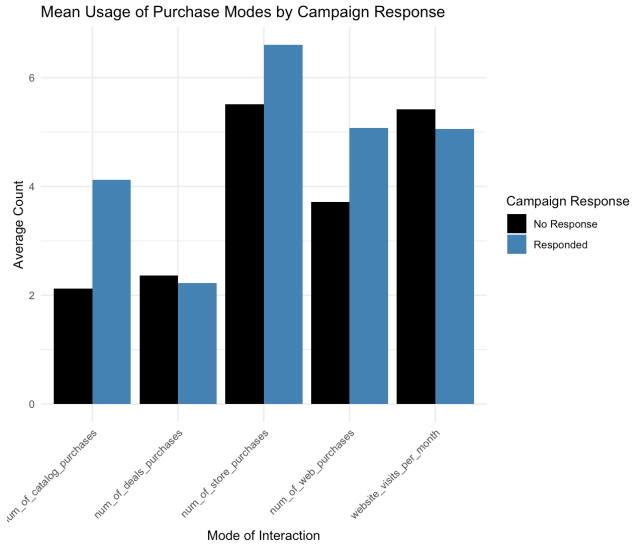


Figure 12: Purchase Mode by Campaign Response

c. Days since last purchase, total monetary value and total purchase count against campaign response:

- **Days Since Last Purchase:**
  Responders (campaign_response = 1) are slightly more recent purchasers compared to non-responders, suggesting recency may influence campaign engagement.
- **Total Monetary Value:**
  Responders tend to spend more overall than non-responders, indicating that higher-spending customers are more likely to react positively to campaigns.
- **Total Purchase Count:**
  Customers who responded to the campaign generally have a higher total number of purchases, implying that frequent buyers are more engaged and responsive.

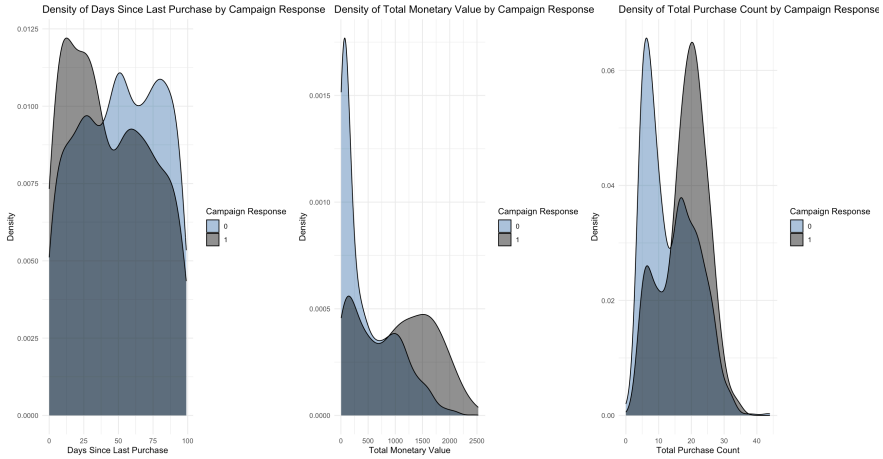


Figure 13: Density Distributions of Key Purchase Behaviors by Campaign Response (Days Since Last Purchase, Total Monetary Value, and Total Purchase Count)

Predictor Relevancy for Regression:

For regression analysis, numeric predictors were assessed using their correlation values with total monetary value, identifying key spending and purchase frequency variables as strong drivers. Categorical predictors were evaluated through visualizations such as boxplots and jitter plots to reveal differences in spending patterns across groups.

Numeric Predictors:

- Considering Correlation Matrix from Figure 10, variables including 'income', 'spending on wines', 'spending on meat', 'num of catalog purchases', 'total purchase count' are showing high correlation with 'total monetary value'.

- Other variables including 'kids at home', 'teens at home', 'web visits num per month',' children', 'birth year num' are low correlated with 'total monetary value'.

Categorical Predictors:

From figure 14, Plots showing 'Total Monetary Value' across different customer segments:

- **Education:** Customers with graduate and PhD levels show significantly higher total spending than those with only basic education.
- **Marital Status**: Couple customers tend to have a broader and higher range of total monetary value compared to other marital groups.
- **Income Bracket:** As expected, customers in the High-income bracket contribute more to total spending, confirming income as a strong predictor.
- **Special Customer:** Customers flagged as special (likely outliers in spending or behavior) exhibit notably higher total monetary value than others.

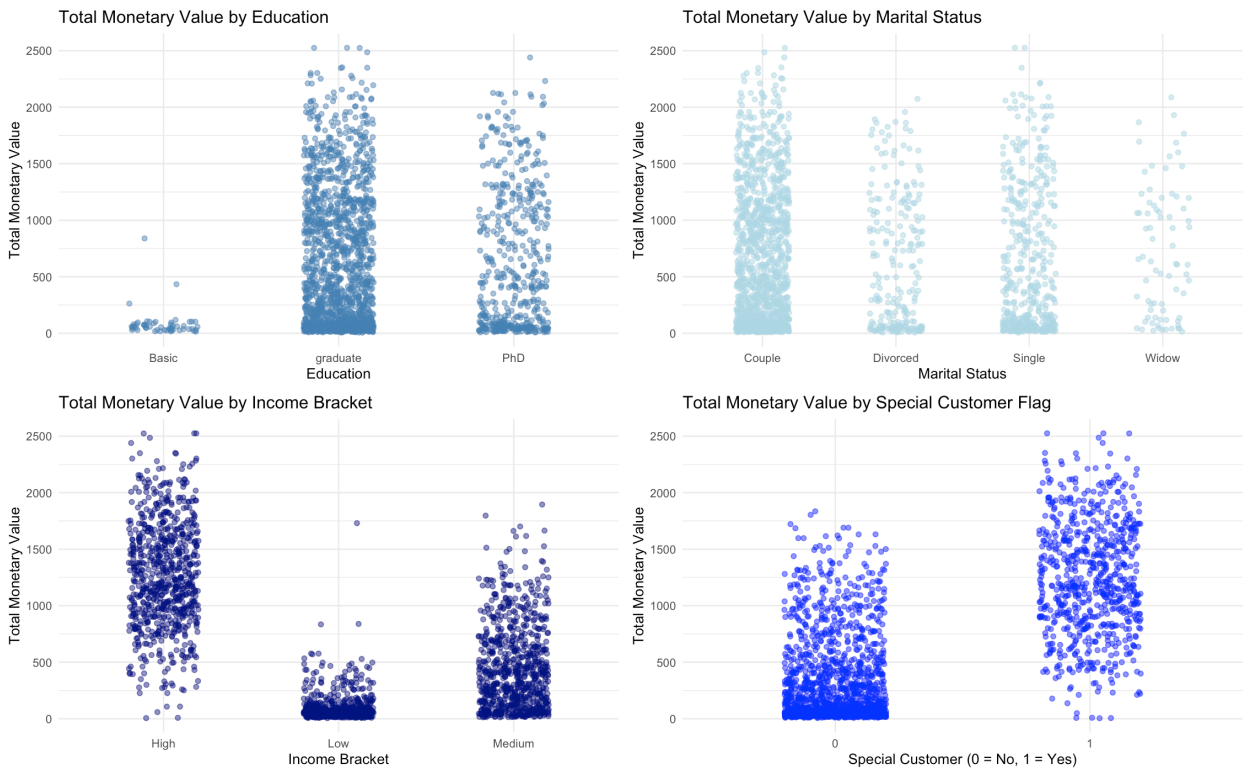


Figure 14: Distributions of Total Monetary Value by Education, Marital Status, and Income Bracket

Dimension Reduction:

Variables which do not have any variance or not useful for analysis are removed.

Feature removed	Reason of removal
'id'	Removing the id column eliminates a non-informative identifier that does not contribute to customer behavior analysis.
'join_date_customer'	This is more granular, and 'join_year' is derived, and this is no longer used for analysis
'year_birth'	'age_at_enrollment' is created and keeping this would be redundant.
'accepted_camp_1', 'accepted_camp_2', 'accepted_camp_3', 'accepted_camp_4', 'accepted_camp_5', 'repsonse'	These individual campaign acceptance indicators are consolidated into a composite response variable, making the separate columns unnecessary
'Z_costcontatct', 'z_revenue'	Constant variable doesn't have any variance and not useful for analysis
'spending_on_fruits_outlier', 'spending_on_fish_outlier', 'spending_on_wines_outlier', 'spending_on_goldproducts_outlier', 'num_of_web_purchases_outlier', 'website_visits_per_month_outlier', 'spending_on_meat_outlier', 'spending_on_sweets_outlier', 'num_of_deals_purchases_outlier'	These were removed because they were used solely to derive the 'special_customer' variable, which captures the overall outlier behavior more efficiently in a single feature. Including both would be redundant.

Table 7: Feature removal and its description

Categorical Encoding

To prepare the dataset for modeling, categorical variables were encoded into numeric formats using encoding.
Education: Ordinal encoding was applied, assigning values based on education level:

- "Basic" = 0, "Graduate" = 1, "PhD" = 2
  This reflects the natural progression of educational attainment.

Income Bracket: Similarly, income categories were converted to numeric values:

- "Low" = 0, "Medium" = 1, "High" = 2
  This helps preserve the ordinal nature of income levels.

Marital Status:
One-hot encoding was used to convert the nominal variable into binary dummy variables. The first dummy (reference category) was removed to prevent multicollinearity during modeling.

Feature selection

Feature selection for segmentation:

Feature selection for segmentation begins by identifying variables that capture consumer demographics, spending habits, and purchasing patterns for understanding customer behavior. The selected features include demographics, spending variables, and few additional variables. The following features are chosen using domain knowledge to ensure that the most relevant aspects of consumer behavior are effectively represented.

Features selected	Reason of Selection
Demographics:‘education’, ‘marital_status_Divorced’, ‘marital_status_Single’, ‘marital_status_Widow’, ‘income_bracket’, ‘age_at_enrollment’, ‘kids_at_home’, ‘teens_at_home’.	These features define customer profiles in terms of socioeconomic and family background. They help explain variations in purchasing behavior based on education level, marital context, financial capacity, age, and household composition.
Spending variables: ‘ spending_on_wines’, ‘spending_on_fruits’, ‘spending_on_meat’, ‘spending_on_fish’, ‘spending_on_sweets’, ‘spending_on_goldproducts’.	These variables reveal customer preferences and spending capacity across product categories, which are critical for segmenting consumers by interest and value.
Purchasing Mode variables: ‘num_of_deals_purchases’, ‘num_of_web_purchases’, ‘num_of_catalog_purchases’, ‘num_of_store_purchases’, ‘website_visits_per_month’, ‘ special_customer’, ‘campaign_response’	These features reflect how customers interact with the business across different modes. They capture behavioral traits like deal sensitivity, preferred shopping platforms, and responsiveness to marketing campaigns.
Additional variables: ‘complain’, ‘days_since_last_purchase’.	Indicators of customer satisfaction and engagement. Recency highlights current activity, while complaints may point to dissatisfaction that could influence future interactions or responses.

Table 8: Feature selection and justification for Segmentation

Feature selection for classification:

The random forest classifier is used to retrieve the importance of features for ‘campaign_reponse’ outcome. The two panels show variable importance based on Mean Decrease Accuracy (left) and Mean Decrease Gini (right). Both metrics highlight the predictors that most significantly affect the model’s ability to classify campaign response.

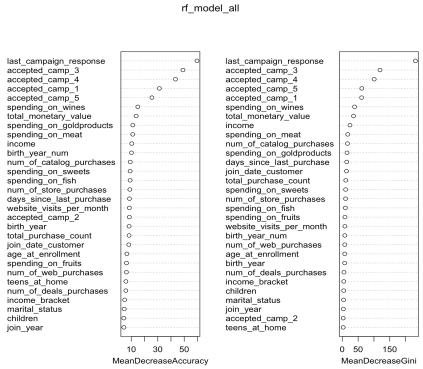


Figure 15: Variable importance for classification

Final features for classification using rf importance and domain knowledge:

Feature selected	Reason for selection
'total_monetary_value'	Since its importance by 'rf model' is comparable to the individual spending variables, it is selected to represent overall consumer expenditure in a more concise way.
'total_purchase_count'	As its predictive power aligns closely with individual purchase mode variables, using this aggregated version simplifies the model while maintaining meaningful behavioral insight.
'education'	This variable shows clear difference (from figure 15) in response rates across categories
'Income'	Strongly supported by both the correlation heatmap and random forest importance plot. High-income groups consistently show better campaign responsiveness.
'children'	This has more predictive power than individual variables.
'days_since_last_purchase'	Included because of its behavioral significance and the RF model showed moderate importance.
'complain'	(Figure 15) suggests customers with complaints have lower response rates. Included for its real-world relevance in customer satisfaction modeling.
'campaign_response'	Target variable representing whether the consumer accepted any of the campaigns

Table 9: Feature Selection for Regression

Feature selection for regression:

- To identify the most relevant predictors for modeling 'total_monetary_value'. Boruta algorithm is applied and revealed the following output after 49 iterations.
- It listed 21 attributes are important and 'complain' and 'complain', 'marital_status_Divorced', 'marital_status_Single', 'marital_status_Widow' are listed as unimportant.

```
> final_decision
                              meanImp  medianImp      minImp     maxImp   normHits  decision
education                   3.7548702  3.8848545  2.02246369  5.4714534 0.97959184 Confirmed
income                     15.9922096 16.1233755 14.18497599 17.2754049 1.00000000 Confirmed
kids_at_home                8.3208165  8.5357357  6.31817916  9.5054403 1.00000000 Confirmed
teens_at_home              10.3700787 10.3491450  8.74722523 12.4347051 1.00000000 Confirmed
days_since_last_purchase    2.6528370  2.7078851 -0.07487886  5.1955704 0.63265306 Confirmed
spending_on_wines          28.8327109 28.8783688 24.87020562 31.0353481 1.00000000 Confirmed
spending_on_fruits         10.5424933 10.5181087  8.01678541 13.3855332 1.00000000 Confirmed
spending_on_meat           20.9886090 21.0731579 19.87281583 22.5551029 1.00000000 Confirmed
spending_on_fish           12.5659575 12.5210775  9.86784127 14.8668429 1.00000000 Confirmed
spending_on_sweets         10.1838260 10.2166835  5.75477748 12.9721261 1.00000000 Confirmed
spending_on_goldproducts   13.7564210 14.3315635  9.15446101 16.3836768 1.00000000 Confirmed
num_of_deals_purchases     10.6161183 10.6149397  8.93387265 11.8902519 1.00000000 Confirmed
num_of_web_purchases       16.8601961 17.0135050 14.55759480 18.8918931 1.00000000 Confirmed
num_of_catalog_purchases   15.0107268 14.9037611 13.20229930 17.6510118 1.00000000 Confirmed
num_of_store_purchases     14.0084717 13.8924839 12.37887654 16.5666023 1.00000000 Confirmed
website_visits_per_month   11.3957374 11.3802766  8.96201333 13.7731507 1.00000000 Confirmed
complain                    0.6258048  1.0030314 -1.23756793  1.7955874 0.00000000  Rejected
special_customer           15.9116611 15.8943677 12.03043233 19.0262559 1.00000000 Confirmed
age_at_enrollment           4.6617820  4.7341783  3.27934111  6.0276586 1.00000000 Confirmed
children                   12.0756179 11.9780635 10.64803955 13.3592461 1.00000000 Confirmed
total_purchase_count       15.2837336 15.2814005 13.52597976 17.6497268 1.00000000 Confirmed
campaign_response          13.5447631 13.5008534 11.30605157 15.2478019 1.00000000 Confirmed
income_bracket             12.2130815 12.2047304 10.09397823 13.8937437 1.00000000 Confirmed
marital_status_Divorced     1.1290338  0.9811250 -0.83116032  3.7647142 0.06122449  Rejected
marital_status_Single       1.1179593  1.2419988 -0.95190277  2.7574459 0.10204082  Rejected
marital_status_Widow       -0.3722614 -0.1640696 -1.59339070  0.6489526 0.00000000  Rejected
>
```

Figure 16: Important features selected by Boruta

Among the confirmed features, we further refined our selection by choosing those with a mean importance greater than 10, ensuring both statistical and practical relevance.

Final predictors for total_monetary_value:

- 'total_purchase_count'
- 'income'
- 'children'
- 'education'
- 'website_visits_per_month'
- 'special_customer'
- 'total_monetary_value'
- 'campaign_response'

Data partitioning

Data partitioning for segmentation:

For understating the patterns of consumer behavior, data is not partitioned since it has no target variable. As a part of understanding consumer behavior, whole data is considered for patterns in data.

Data partition for classification and Regression:

Classification and regression have target variables as they are supervised learning tasks. To build reliable classification and regression models, the dataset was divided to ensure proper learning, tuning, and evaluation. A stratified 70-15-15 split was applied to maintain the original distribution of the target variable 'campaign_response' across all sets:

- Training set (70%): Used to train the model and capture patterns in campaign responsiveness.
- Validation set (15%): Used to fine-tune model parameters and prevent overfitting.
- Testing set (15%): Used to assess final model performance on unseen data.

Data Standardization for Segmentation:

All the selected segmentation features were first converted to numeric format and then standardized using the z-score standardization such that data to have a mean of 0 and standard deviation of 1, ensuring equal contribution from each feature in distance-based algorithms.

Data Oversampling for classification:

The original data was imbalanced, having 'campaign_response' variable with 72.7% non-responders (0) and only 27.3% responders (1). To address this issue and prevent model bias toward the majority class, the ROSE (Random Over-Sampling Examples) technique was applied.
After applying ROSE, the class distribution became nearly balanced:

- Non-responders (0): 51.9%
- Responders (1): 48.1%

Model selection

To address different analytical goals, both supervised and unsupervised machine learning models were employed.

Unsupervised Learning

Unsupervised learning was used to identify distinct customer segments based on behavior and demographics.
K-Means Clustering for Consumer Segmentation:
To understand the consumer behavior, k-means clustering will be used. This unsupervised learning technique groups customers into distinct clusters based on key demographic variables, spending habits, and purchasing patterns. This would help in identifying and characterizing distinct behavioral consumers, such as high-value spenders, frequent buyers, and less-engaged customers.

Supervised Learning

Supervised learning models were applied to predict specific outcomes, such as campaign response and customer spending, based on historical data. These models enable better decision-making and focused marketing actions.

Decision Tree for Classification

A decision tree is used to classify customers as responders or non-responders by recursively partitioning them according to their demographics, purchasing patterns, and spending behaviors. At each split, the algorithm chooses the feature and threshold that best separates responders from non-responders (e.g. via information gain or Gini impurity), growing branches until terminal leaf nodes assign each customer to one of the two classes.
Multi Linear Regression for engagement prediction
Linear regression is applied to predict targeted engagement, measured by total monetary value, using predictors such as purchase frequency, recency, and age at enrollment.

Model Fitting and Performance Evaluation

K-means clustering for consumer behavior

Variables selected for clustering:
'education', 'marital_status_Divorced', 'marital_status_Single', 'marital_status_Widow', 'income_bracket', 'age_at_enrollment', 'kidhome', 'teenhome', 'join_year', 'complain', 'recency', 'mntwines', 'mntfruits', 'mntmeatproducts', 'mntfishproducts', 'mntsweetproducts', 'mntgoldprods', 'numdealspurchases', 'numwebpurchases', 'numcatalogpurchases', 'numstorepurchases', 'numwebvisitsmonth'.

Determining Optimal K (Number of Clusters):
Elbow method:
The elbow method was used to identify the optimal number of clusters. As shown in Figure 17, the most significant drop in within-cluster sum of squares occurs up to K = 2, after which the curve starts to flatten. However, the reduction from K = 2 to K = 3 is still notable, and K = 3 was chosen as the optimal number of clusters to allow for more granular segmentation.

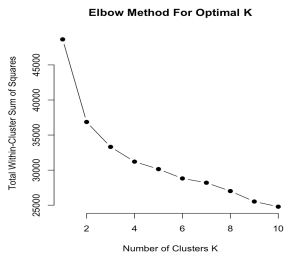


Figure 17: Elbow method plot

Cluster visualization in 2D (PCA plot):
After performing K-means clustering on the scaled dataset, Principal Component Analysis (PCA) was applied solely for visualization purposes. The first two principal components (PC1 and PC2) were used to project the data into two dimensions, enabling a visual understanding of the cluster structure.

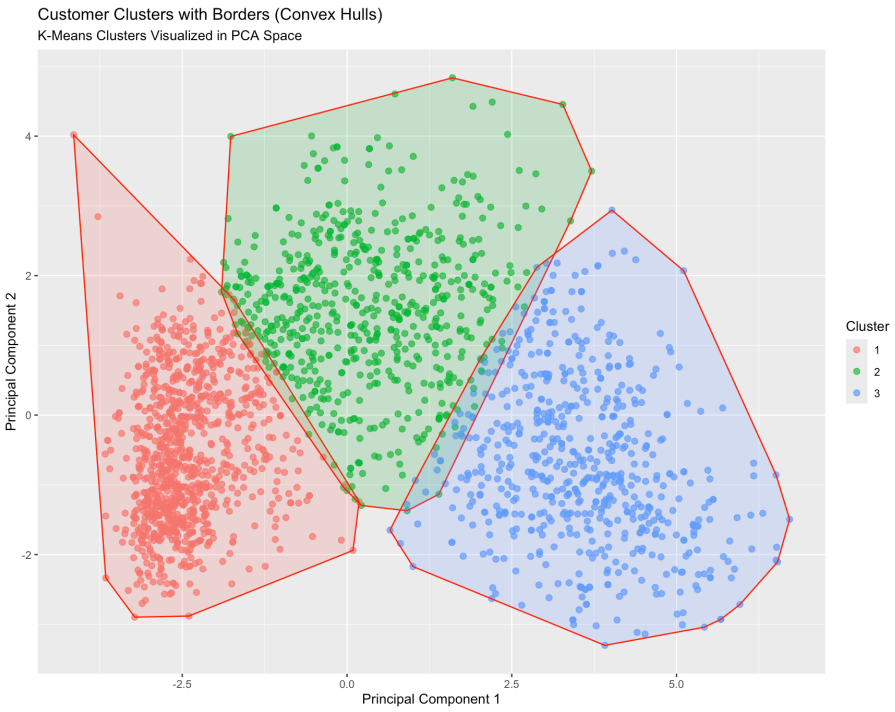


Figure 18: Clusters Visualization of Gym Members in 2D

Clusters size:
The K-means clustering algorithm grouped the dataset into three clusters with the following sizes:

- Cluster 1: 972 observations
- Cluster 2: 622 observations
- Cluster 3: 614 observations

This distribution reflects a relatively balanced segmentation, with each cluster containing a comparable number of observations and no extreme imbalance across groups.

PCA Summary – Contribution of PCs to Variance:

```
> summary(pca)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17
Standard deviation     2.6384 1.44057 1.19752 1.11783 1.06294 1.02928 1.00010 0.98256 0.92944 0.88416 0.85938 0.8153 0.79075 0.7226 0.68138 0.65374 0.63146
Proportion of Variance 0.3027 0.09023 0.06235 0.05433 0.04912 0.04606 0.04349 0.04197 0.03756 0.03399 0.03211 0.0289 0.02719 0.0227 0.02019 0.01858 0.01734
Cumulative Proportion  0.3027 0.39289 0.45524 0.50957 0.55869 0.60475 0.64824 0.69021 0.72777 0.76176 0.79387 0.8228 0.84996 0.8727 0.89285 0.91143 0.92876
                          PC18    PC19    PC20    PC21    PC22    PC23
Standard deviation     0.62510 0.5912 0.52461 0.48707 0.45529 0.42241
Proportion of Variance 0.01699 0.0152 0.01197 0.01031 0.00901 0.00776
Cumulative Proportion  0.94575 0.9609 0.97291 0.98323 0.99224 1.00000
```

Figure 19: Summary of each Principal Component

The PCA summary shows that the first principal component (PC1) alone accounts for 30.27% of the variance, while the second (PC2) contributes an additional 9.23%, bringing the total to 39.29%. The first 6 principal components cumulatively explain approximately 60.48% of the total variance. After the tenth component (76.17% cumulative), each successive component contributes marginally (less than 3.5%) to the total variance, with variance

contributions dropping below 2% after PC16. This indicates that the most meaningful variance is concentrated in the earlier components, and components beyond PC15 provide limited additional explanatory power.

Cluster Profiles Based on Clusters Centroids:

```
> kmeans_result$centers
    education marital_status_Divorced marital_status_Single marital_status_Widow income_bracket age_at_enrollment kids_at_home teens_at_home     complain days_since_last_purchase spending_on_wines
1 -0.14948520             -0.03592623            0.06462334          -0.08720589     -0.8750994       -0.27850712    0.6932132    -0.1639098  0.034694073               0.00419879        -0.7963782
2  0.24393995              0.07839038           -0.14591619           0.11982513      0.3096483        0.39308201   -0.3538748     0.7423441 -0.044688625              -0.02271244         0.3966251
3 -0.01047399             -0.02253831            0.04551462           0.01666595      1.0716537        0.04269041   -0.7389138    -0.4925370 -0.009651977               0.01636142         0.8589231
  spending_on_fruits spending_on_meat spending_on_fish spending_on_sweets spending_on_goldproducts num_of_deals_purchases num_of_web_purchases num_of_catalog_purchases num_of_store_purchases
1         -0.5384527       -0.6460417       -0.5620979         -0.5376430               -0.5580949             -0.1654356           -0.7450755              -0.74267803              -0.8224525
2         -0.2300539       -0.1949012       -0.2395416         -0.2346599                0.1860456              0.6862859            0.6836482               0.06575731               0.4673523
3          1.0854554        1.2201647        1.1324985          1.0888395                0.6950290             -0.4333329            0.4869449               1.10909120               0.8285517
  website_visits_per_month special_customer campaign_response
1                0.4947607       -0.6502147        -0.2956720
2                0.1475445       -0.2044673         0.0628578
3               -0.9327037        1.2364615         0.4043904
```

Figure 20: Centroids of each variable in respective cluster

Cluster 1:
Feature Summary:

- Education: Lowest among all clusters (-0.1498)
- Marital Status: Fewer divorced (-0.0359), and widowed (-0.0872), highest singles (0.0646).
- Lowest Income_bracket (-0.87510)
- Age at Enrollment has lowest value (-0.2785)
- Kids & Teens at Home: High kids value (0.6932) and moderate teens value (-0.1639)
- Complain: More likely to complain (0.03469)
- Days since last purchase: Moderate – low recent purchases
- Spending (All Product Categories): Very low
- Channel Use:
- Very low usage across Web, Catalog, Store, moderate in Deals and highest num of webvisits in month
- Lowest special_customer and campaign response

Consumer insights:

- This segment represents young, low-income consumers with the least educational attainment and highest proportion of single individuals. They typically have more children at home and show a higher likelihood of complaints.
- Their spending is very low across all product categories, and they exhibit minimal engagement across most purchasing channels, except for a moderate use of deals and high frequency of website visits, likely indicating interest without conversion.
- They score lowest on both 'special customer' and 'campaign response', making them the least responsive and least profitable group. Engagement strategies for this segment should be cost-effective and awareness-focused, possibly via basic loyalty or reactivation efforts.

Cluster 2
Feature summary:

- Education: Highest among all clusters
- Marital Status: Low divorced and singles, highest widow.
- Income Bracket: moderate income levels
- Age at Enrollment: Highest (old age consumers)
- Kids at home: Around average
- Teens at home: Highest teens value
- Complain: moderate complaints
- Days since last purchase: Lowest, meaning most recent purchases
- Product Spending: All are in moderate range

- Purchase modes Use: Highest in deals and web, moderate in catalog and store purchases and moderate in webs visits too.
- Moderate chances of being a special customer
- Campaign Response is moderate in this cluster

Consumer Insight:

- This cluster represents a mature customer group with higher education levels and moderate income. They are more likely to be widowed, with a higher number of teenagers at home and average presence of kids. These customers show recent purchase activity and spend at moderate levels across all product categories.
- They exhibit the highest usage of web and deal-based channels, suggesting digital savviness and responsiveness to promotions. Their complaint rate and special customer presence are moderate, reflecting a balanced engagement level.
- Importantly, they show a moderate campaign response rate, implying selective responsiveness—this group may engage depending on the relevance or value proposition of the campaign.
- Overall, they are a stable, middle-tier segment worth targeting with well-designed, value-aligned offers.
-

Cluster 3:

Feature Summary:

- Education: Around average
- Marital Status: moderate divorced, singles and widows
- Income Bracket: Highest
- Age at Enrollment: Moderate
- Kids and teens at home: Lowest in both
- Complain: lowest complaints
- Days since last purchase: Highest
- Product Spending: Highest in all categories
- Purchase modes Use: Lowest in deals and web visits per month, moderate in web purchases, high in catalog and store purchases.

Consumer insights:

- This cluster represents a high-income, high-spending customer group with the highest income bracket and strong monetary engagement across all product categories. They are well-balanced demographically, with moderate age, few or no children at home, and average education levels.
- In terms of behavior, these customers exhibit very low complaint rates, indicating satisfaction, and are least recently active based on the highest days since last purchase.
- They rely heavily on catalog and store purchases, while showing moderate engagement online and low interest in deals and web visit frequency.
- Overall, this is a financially strong and stable segment, ideal for premium and loyalty campaigns, with a preference for traditional purchasing channels and a relatively passive digital presence.

Profile Bar plot of Cluster Centroids Across Standardized Features:
The plot illustrates the distribution of standardized feature values for each cluster centroid. It helps identify dominant traits and behavioral patterns within each consumer segment based on spending habits, demographics, and channel usage.

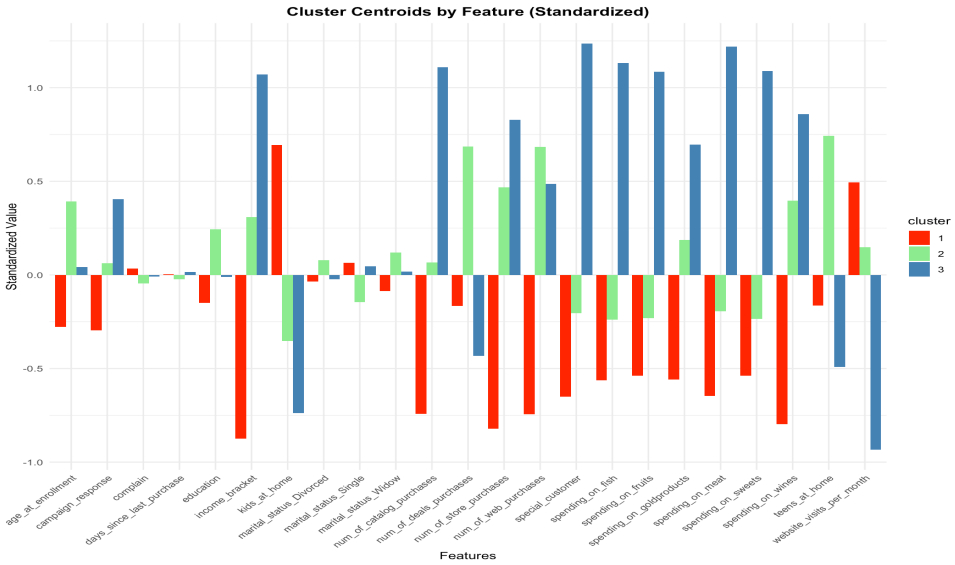


Figure 21: Cluster Centroids Across Standardized Features

Decision trees for Classification:

Decision tree models were developed to classify whether a customer would respond to a campaign, using both default parameters and a tuned configuration to compare predictive performance.

Decision Tree with default Parameter:
A baseline decision tree model was developed using default settings, without any parameter tuning. All selected features were used to classify whether a customer would respond to a marketing campaign.

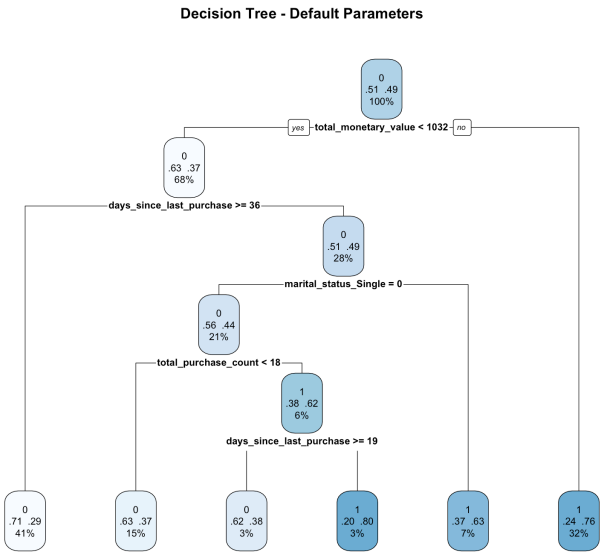


Figure 22: Decision Tree model with default Parameters

Decision Tree with Tuned parameters:
A tuned decision tree model was developed by adjusting key parameters to improve classification performance. The final model used a minimum split of 5, a complexity parameter (cp) of 0.0005, and a maximum depth of 5

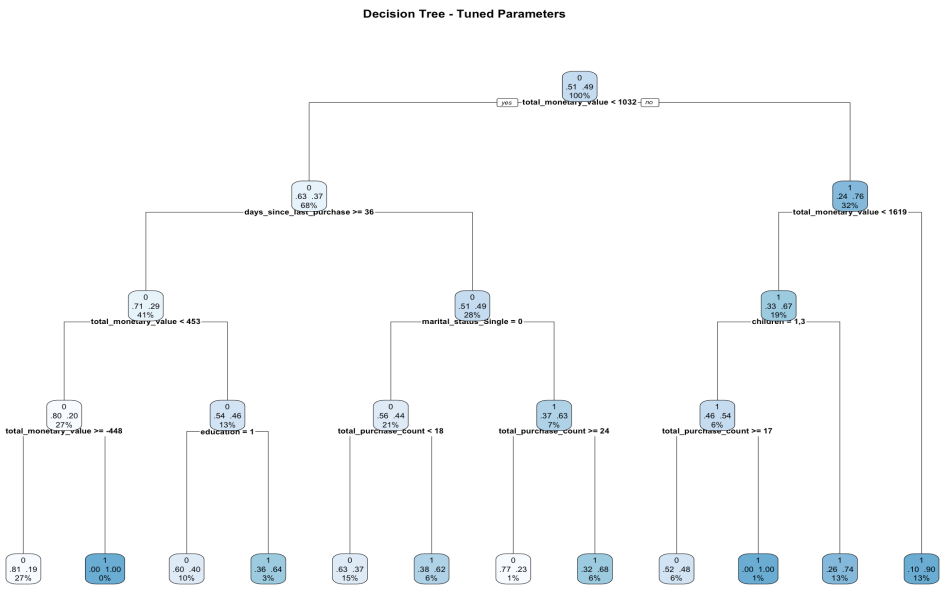


Figure 23: Decision Tree model with Tuned Parameters

Validation set result for default and tuned model:

Validation Results - Decision Tree (Default Parameters)

- Accuracy: 69.79%
  The model correctly predicted approximately 70% of validation samples overall.
- Sensitivity: 77.36%
  The model was highly effective at identifying positive campaign responders (true positives).
- Specificity: 62.79%
  The model was moderately effective at recognizing non-responders (true negatives).
- Balanced Accuracy: 70.07%
  This reflects the average of sensitivity and specificity, indicating balanced performance across both classes.
- Kappa: 0.3988
  Indicates moderate agreement beyond chance.
- Interpretation:
  The default tree performs quite well in identifying responders, making it useful in recall-critical applications such as targeting high-response potential customers.

Validation Results - Decision Tree (Tuned Parameters)

- Accuracy: 67.98%
  Slightly lower overall accuracy compared to the default model.
- Sensitivity: 64.15%
  The tuned model is more conservative in predicting responders.
- Specificity: 71.51%
  Improved ability to correctly classify non-responders.

- Balanced Accuracy: 67.83%
  Very close to the overall accuracy, showing the model is not biased toward either class.
- Kappa: 0.3573
  Still indicates moderate agreement between actual and predicted classifications.
- Interpretation:
  While the tuned model slightly sacrifices sensitivity, it improves specificity, making it a better choice where reducing false positives is important (e.g., avoiding targeting uninterested customers).

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 176  34
         1  65  56

               Accuracy : 0.7009
                 95% CI : (0.6484, 0.7497)
    No Information Rate : 0.7281
    P-Value [Acc > NIR] : 0.879075

                  Kappa : 0.3182

 Mcnemar's Test P-Value : 0.002569

            Sensitivity : 0.6222
            Specificity : 0.7303
         Pos Pred Value : 0.4628
         Neg Pred Value : 0.8381
             Prevalence : 0.2719
         Detection Rate : 0.1692
   Detection Prevalence : 0.3656
      Balanced Accuracy : 0.6763

       'Positive' Class : 1
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 186  32
         1  55  58

               Accuracy : 0.7372
                 95% CI : (0.6862, 0.7838)
    No Information Rate : 0.7281
    P-Value [Acc > NIR] : 0.38200

                  Kappa : 0.3854

 Mcnemar's Test P-Value : 0.01834

            Sensitivity : 0.6444
            Specificity : 0.7718
         Pos Pred Value : 0.5133
         Neg Pred Value : 0.8532
             Prevalence : 0.2719
         Detection Rate : 0.1752
   Detection Prevalence : 0.3414
      Balanced Accuracy : 0.7081

       'Positive' Class : 1
```

Figure 24: Performance evaluation with validation set

Test Set result for default and Tuned set:

Test Results – Decision Tree with Default Parameters:

- Accuracy: 66.7%
  Overall, the model correctly classifies about two-thirds of the test cases.
- Sensitivity (Recall): 66.04%
  This means 66 out of 100 actual responders were correctly identified. The model is moderately effective at detecting customers who do respond to campaigns.
- Specificity: 67.25%
  About 67 out of 100 non-responders were correctly classified. This indicates reasonable performance in filtering out customers unlikely to engage.
- Balanced Accuracy: 66.64%
  Averaging sensitivity and specificity, this confirms the model is balanced and doesn't heavily favor one class over the other.

Interpretation:
This model provides a well-rounded detection of both responders and non-responders. It's a suitable baseline when the goal is to avoid too many false positives or false negatives.

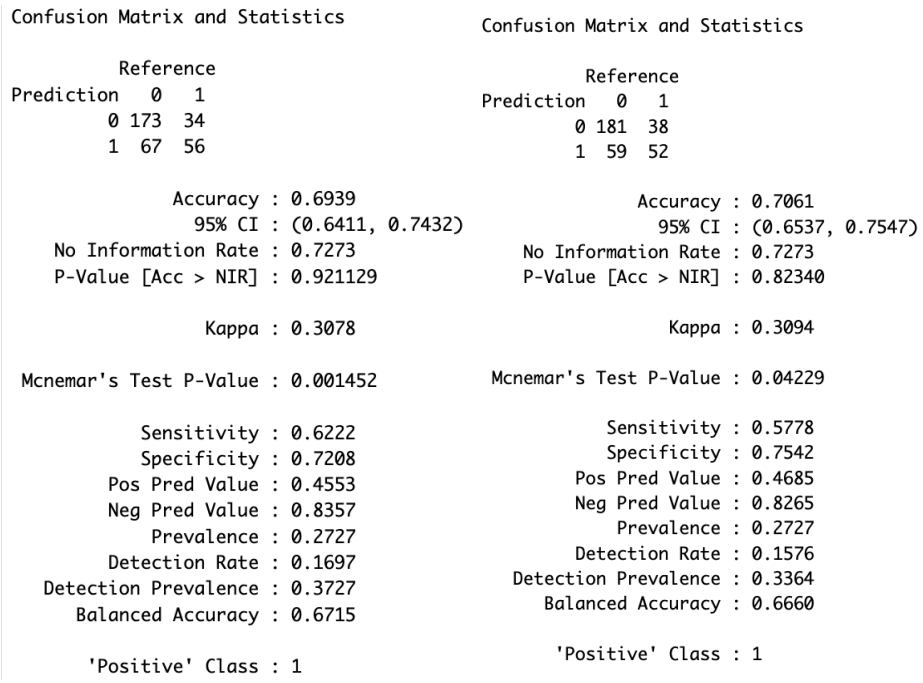
Test Results – Decision Tree with Tuned Parameters:

- Accuracy: 63.3%
  Slightly lower than the default tree, meaning overall correct predictions are fewer.

- Sensitivity (Recall): 49.06%
  The model correctly identifies less than half of the actual responders. This suggests it misses many customers who would engage with a campaign.
- Specificity: 76.61%
  Stronger at identifying non-responders, catching about 77 out of 100 correctly. This indicates a conservative approach—preferring to avoid mislabeling non-responders as responders.
- Balanced Accuracy: 62.83%
  The average of sensitivity and specificity drops due to reduced sensitivity.

Interpretation:

The tuned model is more risk-averse, prioritizing precision over recall. It performs better at avoiding false alarms (false positives), but at the expense of missing out on true positives (actual responders). It may be more appropriate when targeting costs are high and wrongly targeting uninterested customers is expensive.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 173  34
         1  67  56

               Accuracy : 0.6939
                 95% CI : (0.6411, 0.7432)
    No Information Rate : 0.7273
    P-Value [Acc > NIR] : 0.921129

                  Kappa : 0.3078

 Mcnemar's Test P-Value : 0.001452

            Sensitivity : 0.6222
            Specificity : 0.7208
         Pos Pred Value : 0.4553
         Neg Pred Value : 0.8357
             Prevalence : 0.2727
         Detection Rate : 0.1697
   Detection Prevalence : 0.3727
      Balanced Accuracy : 0.6715

       'Positive' Class : 1
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 181  38
         1  59  52

               Accuracy : 0.7061
                 95% CI : (0.6537, 0.7547)
    No Information Rate : 0.7273
    P-Value [Acc > NIR] : 0.82340

                  Kappa : 0.3094

 Mcnemar's Test P-Value : 0.04229

            Sensitivity : 0.5778
            Specificity : 0.7542
         Pos Pred Value : 0.4685
         Neg Pred Value : 0.8265
             Prevalence : 0.2727
         Detection Rate : 0.1576
   Detection Prevalence : 0.3364
      Balanced Accuracy : 0.6660

       'Positive' Class : 1
```

Figure 25: Performance Evaluation of test set without and with tuning

AUC Comparison:

- Validation AUC is slightly better in the tuned model, indicating improved ability to rank positive cases higher than negative ones during training evaluation.
- However, Test AUC drops for the tuned tree, which may suggest some overfitting, the model adapts well to training patterns but generalizes less effectively to unseen data.
- The default model, while slightly weaker on validation, holds more stable performance across datasets, which could indicate better generalizability.

Model	Validation_AUC	Test_AUC
Decision Tree (Default)	0.6940756	0.6899306
Decision Tree (Tuned)	0.7506455	0.7358333

Table 10: AUC Comparison of Models Developed

ROC Comparisons:

- Top Left (Default Tree - Validation): The curve shows decent lift above the diagonal, reflecting a moderate ability to distinguish responders from non-responders.
- Top Right (Tuned Tree - Validation): A higher arc indicates better discrimination power, consistent with its higher AUC on validation.
- Bottom Left (Default Tree - Test): Reasonable performance on unseen data, holding its shape well.
- Bottom Right (Tuned Tree - Test): The curve flattens compared to validation, implying drop in generalization, likely due to overfitting.

The tuned tree excels in validation, but the default model maintains a more balanced and stable performance on the test set, which could make it a better choice if robustness and generalization are priorities.

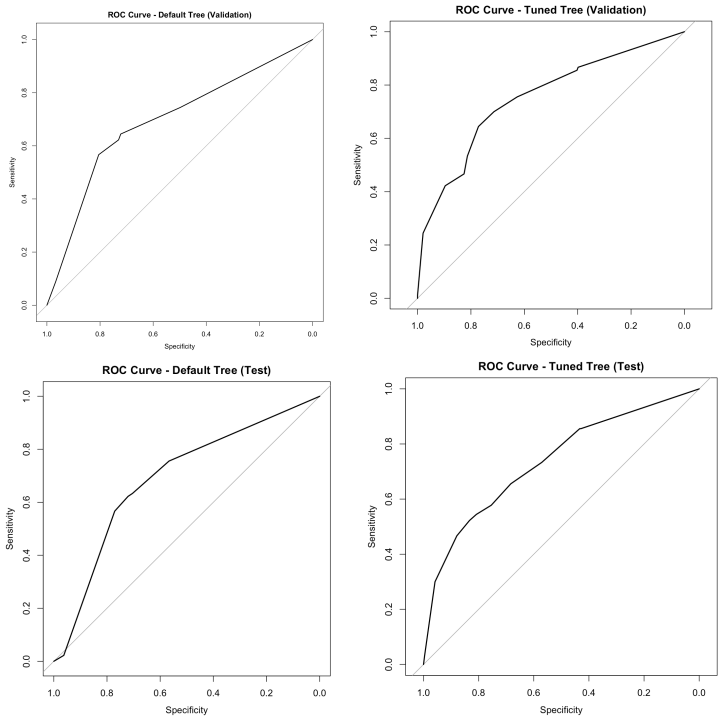


Figure 26: ROC on Validation and Test sets

Multi Linear Regression for Customer Engagement prediction:

To assess customer engagement in terms of overall spending behavior, a linear regression model was built to predict 'total_monetary_value' using a set of relevant demographic and behavioral predictors.
Significant Predictors ($p < 0.001$)
These variables have a statistically strong influence on the total monetary value:

- 'total_purchase_count': Strongest positive effect on spending; as the number of purchases increases, total spending increases significantly.
- income: Positively correlated; higher income is associated with greater spending.
- children: Negative relationship; households with more children tend to spend less overall.
- special_customer1: Strong positive impact; customers flagged as outliers (e.g., high spenders or unique behavior) tend to spend significantly more.
- campaign_response1: Customers who responded to a campaign also show higher total spending.

Moderately Significant ($0.05 < p < 0.1$)

- education: Marginally significant (p = 0.0614); suggests some weak positive influence of education level on spending.

Not Significant (p > 0.1)

- website_visits_per_month: Not significant (p = 0.7440); frequency of site visits does not predict total spending in this model.

Model Fit Metrics

- Multiple R-squared: 0.8255 — Indicates that approximately 82.6% of the variance in spending is explained by the model.
- Adjusted R-squared: 0.8247 — Adjusted for the number of predictors; still strong.
- F-statistic: Highly significant overall (p < 2.2e-16), indicating the model is statistically robust.

Overall, the regression model effectively explains consumer spending using selected behavioral and demographic features, especially purchase frequency and income.

```
> summary(lm_model)

Call:
lm(formula = total_monetary_value ~ ., data = train_data_reg)

Residuals:
    Min      1Q  Median      3Q     Max
-1263.9  -140.4   -12.1   113.2  1090.9

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -2.767e+02  3.869e+01  -7.152 1.32e-12 ***
total_purchase_count       2.327e+01  1.247e+00  18.670  < 2e-16 ***
income                     9.518e-03  5.371e-04  17.719  < 2e-16 ***
children                  -1.464e+02  1.005e+01 -14.566  < 2e-16 ***
education                  2.754e+01  1.471e+01   1.872   0.0614 .
website_visits_per_month -1.225e+00  3.750e+00  -0.327   0.7440
special_customer1          3.374e+02  1.877e+01  17.969  < 2e-16 ***
campaign_response1         1.815e+02  1.563e+01  11.613  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 255.2 on 1540 degrees of freedom
Multiple R-squared:  0.8255,    Adjusted R-squared:  0.8247
F-statistic:  1041 on 7 and 1540 DF,  p-value: < 2.2e-16
```

Figure 27: Linear regression summary output

Validation Set Performance:

- RMSE (261.71): The model’s predictions deviate by about $262 from actual total spending on average.
- MAE (180.85): The mean absolute prediction error is approximately $181, reflecting the typical margin of error.
- R-squared (0.8407): Around 84.07% of the variation in customer spending is explained by the model — indicating strong model fit.

Test Set Performance:

- RMSE (231.98): Prediction error slightly decreases on test data, suggesting better performance on unseen observations.
- MAE (167.73): The model achieves improved average accuracy with a lower absolute error than on validation.
- R-squared (0.8407): Same as validation, confirming excellent generalization and absence of overfitting.

```
|Dataset    |   RMSE|    MAE| R_squared|
|:----------|------:|------:|---------:|
|Validation | 261.71| 180.85|    0.8407|
|Test       | 231.98| 167.73|    0.8407|
>
```

Table 11: Comparison of Validation and test performance

Cost Analysis

Considering the confusion matrix of classification model with default parameters for identifying consumer would respond or not as it was performing better on new data.

Cost matrix can be written as:

	Actual Responder (1)	Actual non-responder (0)
Predicted Responder (1)	TP = 56 (Cost =$0)	FP = 67 (Cost = $5)
Predicted non-responder (0)	FN = 34 (Cost = $25 each)	TN = 173 (Cost = $0)

Table 12: Cost Analysis of Calorie Prediction regression Model

Cost assumptions

- False Negative (FN): $25 each
- False Positive (FP): $5 each

Total Cost = (C(FN) x FN) + (C(FP) x FP)

- C(FN) is the cost of predicting an observation wrongly as not cancelled when it is an actual cancellation, C(FN) = $25.
- Here, C(FP) is the cost of predicting an observation wrongly as cancelled when it is actually not cancelled, C(FP) = $5.

Total Cost = $[(25 * 34) + (5 * 67)] = 1{,}185$
Total misclassification cost = $1,185

Interpretation:

- Not spotting a real responder (a false negative) is five times more expensive than contacting a non-responder (a false positive).
- With costs set at $25 per false negative and $5 per false positive, this model incurs $1,185 in total misclassification cost.
- Because it leaves less money on the table than a non-tuned alternative, this model is the better choice under these cost assumptions.
- The next way to save is to cut down the 34 false negatives—i.e. capture more true responders—even if that causes a modest rise in false positives, since each avoided false negative delivers the bigger savings.

Business Recommendations

- **Cluster 1consumers (young, low-income, high web-visits, low spend):** Launch a mobile-first gamified "spin-to-win" discount widget (e.g. via TikTok ads) that awards a 10% off code on first purchase to convert their high digital engagement into sales.
- **Cluster 2 consumers (older, moderate-income, educated families):** Offer curated "Family Essentials" bundles with free shipping on orders ≥ $75, promoted through bi-weekly email newsletters and targeted SMS deal alerts timed around peak shopping hours.
- **Cluster 3 consumers (middle-aged, high-income, high-spenders):** Provide VIP early access to premium new arrivals, paired with concierge-style upsell emails and invite-only tasting or sampling events to deepen loyalty.
- **Automated cost-matrix monitoring:** Implement a monthly pipeline that recalculates your confusion matrix against current cost weights and triggers model retraining when total error cost rises, ensuring you continually minimize "money left on the table."
- Dynamic discount experimentation A/B test tiered discount levels (e.g. 10%, 15%, 20%) within each cluster, then use your $25 FN vs $5 FP cost matrix to calculate net lift per test cell. This ensures discount levels are profitable on a per-segment basis.

Future Work:

- **Implement real-time offers:** Hook the models into your live systems so that customers receive tailored discounts or messages the moment they show interest—like adding an item to their cart or opening an email.
- **Broaden our data sources**: Bring in more information like what people click on your website or app, their social media signals, and quick customer surveys—so we can better understand what drives their decisions.
- **Upgrade value forecasting:** Replace the basic spend-prediction model with more advanced techniques (for example, tree-based or survival analysis) to forecast both future purchases and the risk of customers churning.

Observations and Conclusion:

Observations

- Customer Segmentation:
  Used K-Means clustering on standardized behavioral and demographic variables, distinct consumer segments were identified. These segments provided a data-driven basis for understanding consumer behavior.
- Campaign Responsiveness Analyzed:
  Classification models (decision trees and random forests) were used to predict customer response to marketing. Important predictors such as income, previous campaign engagement, and purchasing habits were identified, helping refine targeting strategies.
- Spending Behavior Predicted:
  A linear regression model was developed to estimate total monetary value using key predictors like income, campaign response, and purchase channels, enabling identification of high-value customers.
- Actionable Strategy Recommendations:
  Based on clustering insights, differentiated marketing strategies were recommended for each segment, focusing on personalization and efficiency in campaign execution.

Conclusion:

This analysis demonstrates that segment-based modeling, combined with cost-sensitive classification, markedly improves marketing ROI. Integrating these models into real-time workflows and automating performance monitoring will ensure ongoing efficiency and adaptability as customer behavior evolves.

Appendix A: Executive Summaries

Executive Summary 1

Auto Rental Cancellation Analytics

Name: Mithila Papi Shetty
Date: 04-30-25

This project focuses on predicting and minimizing car-rental booking cancellations by leveraging reservation and trip-log data. Initial data collection and exploratory analysis established variable distributions and highlighted issues requiring remediation. Comprehensive preprocessing addressed missing and zero values, encoded categorical fields, and derived new features. Predictors were then analyzed against the cancellation outcome to identify influential factors; irrelevant or redundant variables were removed and the most important features retained.

The cleaned dataset was partitioned into training and test sets for unbiased evaluation, and class imbalance was corrected via oversampling. A Naive Bayes classifier was developed to estimate cancellation probabilities for each booking based on the selected factors, with performance assessed. Cost analysis translated false positives and false negatives into dollar impacts, guiding selection of an operating threshold. Recommendations were provided focusing on encouraging short-distance trips, boosting online booking commitment, and streamlining the mobile-web flow to reduce cancellations and improve customer reliability.

Executive Summary 2

Fitness Center Member Wellness Analytics

Name: Mithila Papi Shetty
Date: 04-30-2025

The Fitness Center Member wellness Analytics Project leverages data-driven insights to enhance member experiences at fit Life Wellness by analyzing demographic, workout, and health metrics. The objective is to uncover patterns and trends in gym members fitness behaviors, enabling the creation of personalized workout plans and optimized health strategies.

The project begins with exploratory data analysis (EDA) to understand key fitness and health metrics. Feature selection ensures that only the most relevant predictors are considered, while dimensionality reduction eliminates redundancy, improving model efficiency. Data partitioning is applied to assess model performance across training and testing sets. To enhance personalization, clustering techniques group members based on fitness habits, while predictive modeling estimates calorie expenditure, enabling tailored workout recommendations. This approach not only refines member engagement strategies but also supports progress tracking, ensuring an evidence-based, data-driven fitness journey for everyone.

Executive Summary 3

Consumer Behavior Analytics

Name: Mithila Papi Shetty
Date: 04-30-2025

The Consumer Behavior Analytics Project leverages data-driven insights to enhance marketing effectiveness for Shop-Smart. The project focuses on analyzing customer demographics, purchasing behavior, and campaign engagement to uncover meaningful consumer patterns that support strategic business decisions. The project begins with an in-depth exploration of consumer data, addressing data quality through cleaning, outlier handling.

To enhance model performance and reduce high dimensionality, relevant features are selected. The dataset is partitioned into train and test sets to ensure fair model evaluation. To understand consumer behavior, unsupervised learning techniques are used to segment customers based on their purchasing habits and engagement with marketing campaigns. Predictive modeling is then applied to classify whether a customer is likely to respond to future campaigns, while a regression model is developed to estimate customer engagement. Based on the clustering and prediction results, the project concludes with a targeted marketing strategy designed to maximize engagement and provide specific recommendations on which consumer segments should receive types of promotions, allowing for more personalized and effective marketing.