**Hemanth Matta**

**Webster University**

**CSDA 6010 DATA ANALYTICS PRACTICUM**

**SPRING 2025**

**FINAL PROJECT (Case-1)**

**JP WANG**

**TABLE OF CONTENTS**

1. INTRODUCTION ……………………………………………………………………………...8

2. BUSINESS AND ANALYTICAL………………………………………....................................8

2.1. BUSINESS PROBLEM………………………………………………………………8

2.2. BUSINESS GOAL……………………………………………………………………8

2.3. ANALYTICAL GOAL………………………………………………………………..8

2.4. ANALYTICAL APPROACH…………………………………………………………9

3. DATA EXPLORATION AND PREPROCESSING……………………………………………10

3.1. ATTRIBUTE DEFINITION…………………………………………………………11

3.2. TYPES OF DATA……………………………………………………………………11

3.2.1. CATEGORICAL ATTRIBUTES…………………………………………..12

3.2.2. CONTINOUS ATTRIBUTES……………………………………………..12

3.3. SUMMARY STATISTICS…………………………………………………………..13

3.4. DETAILED ANALYSIS ON EACH ATTRIBUTE………………………………….14

3.5. CHECKING ZEROS………………………………………………………………...29

3.6. MISSING VALUES…………………………………………………………………29

4. DATA TRANSFORMATION ………………………………………………………………...33

5. PREDICTOR ANALSIS AND RELEVANCY ………………………………………………..34

6. DIMENSION REDUCTION……………………………………………………………….…42

7. DATA PARTITION……………………………………………………………………………43

8. MODEL SELECTION………………………………………………………………………...45

9. REPORT ON MODEL PERFORMANCES…………………………………………………..50

10. MODEL IMPROVEMENT AND EVALUATION………………………………………...…51

11. OBSERVATION……………………………………………………………………………...54

12. CONCLUSION………………………………………………………………………………55

**LIST OF TABLES**

Table 1. user booking frequency…………………………………………………………………15

Table 2. vechile id frequency…………………………………………………………………….16

Table 3. distribution of observation in months…………………………………………………..21

Table 4. number of booking made for each month………………………………………………23

Table 5. number of bookings…………………………………………………………………….25

Table 6. model performances……………………………………………………………………51

**LIST OF FIGURES**

Figure 1. type of variables in the dataset………………………………………………………...13

Figure 2. summary of the data set………………………………………………………………..14

Figure 3. vehicle id distribution………………………………………………………………….17

Figure 4. distribution of travel type id…………………………………………………………...18

Figure 5. distribution of package id……………………………………………………………...19

Figure 6. missing vs non-missing in area ids across travel types………………………………..20

Figure 7. trip distribution by from month………………………………………………………..21

Figure 8. distribution of observations in day…………………………………………………….22

Figure 9. Booking distribution by hour of the day……………………………………………….23

Figure10. number of bookings created in a day…………………………………………………24

Fig11: booking type distribution…………………………………………………………………25

Fig 12: missing vs non\_missing values across travel types in lat-log…………………………...26

Fig13: density plot of distances in kms………………………………………………………….27

Fig14: distance of the rides by category…………………………………………………………28

Fig15: car cancellation………………………………………………………………………….28

Figure 16. count of zeros in the column…………………………………………………………29

Figure 17. missing values………………………………………………………………………..29

Figure 18. graphical representation of missing values…………………………………………30

Figure19. missing values in point-to-point data…………………………………………………30

Figure20. missing values in hourly and long-distance data……………………………………..31

Figure21. percentage of missing values in package id, from city id, to city id and to date……..33

Figure 22. distribution of number of rides per user……………………………………………...35

Figure23. car cancellation rate by vehicle model………………………………………………..36

Figure24. from area ids and to area ids………………………………………………………….37

Figure25. cancellation rates of online booking, mobile site booking and in person booking….38

Figure26. cancellation rate by from time………………………………………………………39

Figure 27. cancellation rate by booking time……………………………………………………40

Figure28. cancellation rate by booking month………………………………………………..…41

Figure29. cancellation rate by from month………………………………………………………42

Figure30. partition and proportions of car cancellation in all sets of data………………………43

Figure31. upsampling…………………………………………………………………………....44

Figure 32. correlation matrix……………………………………………………………………45

Figure33. decision tree…………………………………………………………………………..46

Figure34. confusion matrix for decision tree…………………………………………………….47

Figure35. logistic model…………………………………………………………………………48

Figure35. confusion matrix for logistic regression………………………………………………50

Figure36. confusion matrix for random forest…………………………………………………..52

**Executive Summary**

San Francisco Auto Rental (SAR) has been facing an ongoing challenge with drivers canceling scheduled rides, leading to dissatisfied customers. Since 2013, SAR has struggled to understand the causes behind these cancellations, and by 2024, there is still no clear strategy to address this issue. The objective of this project is to build a predictive model that identifies factors contributing to cancellations and provides SAR with actionable insights to reduce them.

A thorough analysis of SAR’s dataset, consisting of approximately 10,000 records and 19 variables related to customer demographics, ride information, and cancellation statuses, was conducted. Key findings include:

Booking Channel: Online and mobile bookings were found to have a higher rate of cancellations.

Location & Time Factors: Certain areas and times, such as the summer season, were associated with a peak in cancellations.

Trip Duration: Shorter trips tended to be canceled more frequently.

Booking Timing: Cancellations occurred more often in the morning hours.

Three machine learning models were evaluated to predict cancellations: Decision Tree, Logistic Regression, and Random Forest. The models were assessed based on various metrics such as accuracy, sensitivity, specificity, precision, and balanced accuracy to identify the most reliable method for detecting cancellations:

Decision Tree: This model achieved strong sensitivity, indicating its effectiveness in identifying cancellations, but its precision was relatively low, resulting in a higher chance of false positives.

Logistic Regression: This model performed well, showing good accuracy and sensitivity, with a moderate precision that helped reduce the number of false positives while identifying cancellations effectively.

Random Forest: While the Random Forest model achieved the highest accuracy and specificity, it had lower sensitivity, meaning it missed a higher proportion of actual cancellations, despite performing well in predicting non-cancellations.

Given the goals of SAR to reduce cancellations and take proactive measures, Logistic Regression was chosen as the optimal model. It strikes a good balance between accuracy and sensitivity, ensuring that cancellations are correctly identified for timely interventions, such as offering incentives or arranging backup rides. The model's solid performance across both positive and negative cases provides reliable predictions.

**1. Introduction:**  
  
San Francisco Auto Rental is a short-term rental company. It is a transportation company that provides a rental service to customers. The company provides different kinds of service options for a ride booking.  
  
They observed that some rides are canceled by the riders. Later then, the company comes with an online booking facility for the consumers. After their efforts, they still see cancellations in the rides. In addition to that, the customers who booked the rides in advance end up with no rides. Customers are facing trouble. Because of this they are not able to serve their customers consistently and cannot give competition to their rivals. These kinds of actions directly affect the revenue of the company, and customers lose faith in the SAR company.  
  
The main objective of the project is to identify the key reasons why the drivers are cancelling the rides. Need to build analytical models that address the business problem, give them a solution to avoid the cancellations of rides, and aim to improve the quality and quantity of rides in the coming years. Which directly improves the company's market in the transportation industry.

**2. Business and analytical goals:**

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

**2.1 Business problem:**

The San Francisco Auto Rental (SAR) has the problem that drivers are cancelling the rides that are booked by customers. So, these led to

* Unhappy customers lose trust in SAR services(introduction)
* Company reputation is negatively impacted
* Majorly, loss of revenue because of unfulfilled rides

**2.2. Business goals:**

* Find whether a new ride will get cancelled: By mitigating the possibility of ride cancellation will help the business users to take preventive action either by offering incentives or by arranging back up ride.
* Find out what are the factors that are significant in the ride cancellations: By finding these factors, it allows business to pinpoint the underlying reasons causing drivers to cancel their rides.

**2.3. Analytical goals:**

* Predict cancellations: Build a model that should predict if the ride is likely to be cancelled in the future
* Identify patterns in cancellations: Look for trends and patterns, such as which types of bookings or timings are more likely to be canceled.

**2.4. Analytical approach:**

* Predict Cancellations:

Model Development: Use classification models (e.g., logistic regression, decision trees, random forests) to predict cancellation probabilities based on historical data.

To achieve these objectives, the following analytical approach will be implemented:

**Data Preparation:**

Process and clean the dataset related to driver cancellations (data is already provided).

Address missing values and resolve inconsistencies to enhance data accuracy.

**Exploratory Data Analysis (EDA):**

Utilize visualizations to uncover relationships between variables.

Examine patterns in ride cancellations.

Assess correlations between various attributes and car cancellations.

**Feature Engineering:**

Generate new variables that may impact ride cancellations, such as ride time and day of the week.

Combine relevant variables to enhance interpretability (e.g., merging online\_booking, mobile\_site\_booking, travel\_type\_id, and package\_id).

**Statistical Analysis:**

Conduct chi-square tests to explore relationships among categorical variables.

Perform T-tests to compare the means of two groups and determine significant differences.

**Dimensionality Reduction:**

Select the most relevant and influential features for modeling.

**Predictive Modeling:**

Develop classification models to predict cancellations.

Assess model performance using appropriate evaluation metrics.

**Insights and Recommendations:**

Identify key drivers of cancellations.

Offer actionable strategies to minimize cancellations and enhance customer satisfaction.

**3. Data exploration and preprocessing**

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

**3.1. Attribute definition:**

The SAR company provided the data set having 10000 records with 19 attributes.

1. **Row ID:** This is the unique identifier of this dataset. It is a number identifying each record.
2. **User ID:** This is the identifier of each client. There are many duplicates in this subset, meaning there are many clients who have called multiple rides.
3. **Vehicle Model ID:** This is an ID that represents the type of vehicle driven for each ride.

|  |  |
| --- | --- |
| Vehicle ID | Vehicle Type |
| 1 | Hatchback |
| 10 | Sedan |
| 12 | Compact Sedan |
| 13 | SUV |
| 17 | Compact SUV |
| 23 | MUV (Multi-Utility Vehicle) |
| 24 | Pickup Truck |
| 28 | Small Truck (Light-Duty, e.g., Delivery Vans) |
| 30 | Medium Truck (Box Truck, e.g., U-Haul 10ft–15ft) |
| 36 | Large Truck (Heavy-Duty, e.g., U-Haul 20ft–26ft, Moving Trucks) |
| 54 | Commercial Truck (Semi-Trucks, Freight Trucks) |
| 64 | Luxury Sedan (e.g., Mercedes S-Class, BMW 7-Series, Audi A8) |
| 65 | Luxury SUV (e.g., Range Rover, Bentley Bentayga, Rolls-Royce Cullinan) |
| 70 | High-Performance Sports Car (e.g., Porsche 911, Chevrolet Corvette, Nissan GT-R) |
| 85 | Supercar (e.g., Ferrari, Lamborghini, McLaren, Bugatti, Koenigsegg) |
| 86 | Hypercar (e.g., Bugatti Chiron, Pagani Huayra, Rimac Nevera) |
| 87 | V8 Muscle Car (e.g., Ford Mustang GT, Dodge Challenger Hellcat, Chevrolet Camaro ZL1) |
| 89 | V10/V12 Supercar (e.g., Lamborghini Aventador, Ferrari 812) |
| 90 | Electric cars |
| 91 | Electric trucks |

1. **Travel Type ID:** This is an ID that represents the type of travel (1= long distance, 2= point to point, 3= hourly rental).
2. **Package ID:** This is an ID that represents the type of travel package, with the following descriptions: 1=4hrs & 40kms, 2=8hrs & 80kms, 3=6hrs & 60kms, 4= 10hrs & 100kms, 5=5hrs & 50kms, 6=3hrs & 30kms, 7=12hrs & 120kms.
3. **From Area:** This is an identifier of the starting area. Available only for point-to-point travel.
4. **To Area:** This is an identifier of the ending area. Available only for point-to-point travel.
5. **From City ID:** Unique identifier of the starting city (i.e. suburb cities of San Francisco).
6. **To City ID:** Unique identifier of the ending city.
7. **From Date:** Date and time of the requested trip start.
8. **To Date:** Time stamp of trip end.
9. **Online Booking**: A binary (0,1) variable representing whether the booking was made online or not. 0 represents no, 1 represents yes.
10. **Mobile Site Booking**: A binary (0,1) variable representing whether the booking was made on their mobile site or not. 0 represents no, 1 represents yes.
11. **Booking Created:** Date and time of booking created.
12. **From Lat:** The latitude of the start area.
13. **From Long:** The longitude of the start area.
14. **To Lat:** The latitude of the end area.
15. **To Long:** The longitude of the end area.
16. **Car Cancellation:** The target variable. A binary (0,1) variable representing whether the ride was cancelled. 0 means no, 1 means yes

**3.2. Types of data:**

Taking the SAR Rental dataset, there are 19 attributes.

There are 3 types of data in the file. we observe this from the Fig1

1. Integer

* Row\_id, user\_id,vehicle\_model\_id, package\_id, travel\_id, from\_area\_id, to\_area\_id, from\_city\_id, to\_city\_id, online\_bookings, mobile\_online\_booking, car\_cancellations

1. Character

* From\_date, to\_date , bookings\_created

1. Number

* From\_lat, from\_long, to\_lat, to\_long

**3.2.1. Categorical attributes:**

From fig1, We further classify categorical into ordinal, nominal, binary

* Ordinal: A categorical variable where the order of the categories matters, but the differences between them are not necessarily consistent
  + row\_id
* Nominal: A categorical variable with no inherent order or ranking between categories
  + user\_id, vehicle\_model\_id, package\_id, travel\_id, from\_area\_id, to\_area\_id, from\_city\_id, to\_city\_id,
* Binary: A variable with two possible outcomes, often representing 0/1 situation
  + online\_bookings, mobile\_site\_bookings, car\_cancellation
    1. **Continuous attributes:**
* A variable that can take any value within a given range and is measurable on a scale
  + From\_lat, from\_long, to\_lat, to\_long

From Fig1, all the time data created in a character, which falls in the timestamps category.

* Booking\_created, from\_date, to\_date

A computer code with many small letters

AI-generated content may be incorrect.

Fig1: Type of variables in the dataset

**3.3. Summary statistics:**

The summary statistics provides some meaningful insights, it can be uncovered from the Fig2

* user\_id ranges from 16 to 48,729 There is a wide range of user IDs, indicating many unique users.
* from\_area\_id (2 to 1401) and to\_area\_id (6 to 1403) indicate a wide variety of pickup and dropoff locations, covering a broad geographical range.
* from\_city\_id and to\_city\_id ranges from 0 to 17 represents 17 cities are involved in the data set
* From the online\_booking the mean value is 0.3533 which represents only 35% of the online booking option
* From the mobile\_booking the mean is 0.0424. In this very few 4% of customers are booked online
* From the above two insights we can see that there are some other bookings which do not fall under this category. We can assume that as in-person booking for further analysis.
* In the car cancellation (target variable) we can observe that 0.0743 is the mean. There are 7.43% car cancellations

A screenshot of a computer

AI-generated content may be incorrect.

Fig2: Summary of the dataset

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

**Row\_id:**

It is the count of all the observations in the dataset, there are 10000 row ids

**User\_id:**

From table1

* Total Users: 6,043 users.
* One-Time Users: 5,841 users booked only 1 ride.

High Proportion of One-Time Users:

* + Nearly 97% of users (5,841 out of 6,043) booked only one ride, which suggests that many customers may be trying the service but not returning. This could be due to several factors, such as customer dissatisfaction
* Multiple Bookings: A smaller group of users made more bookings:
  + 722 users booked 2 times.
  + 225 users booked 3 times.
  + The number of users drops significantly as bookings increase, with only 1 user booking 105 times.

Very Few High-Frequency Users:

* + - Only a handful of users have extremely high booking counts. For instance, one user booked 105 times, and a few others booked over 50 times.
    - These high-frequency users might be business clients, corporate accounts, or drivers utilizing the system for operational purposes.

|  |  |  |
| --- | --- | --- |
| number of rows | number of rides | number of users |
| 1 | 1 | 5841 |
| 2 | 2 | 722 |
| 3 | 3 | 225 |
| 4 | 4 | 111 |
| 5 | 5 | 43 |
| 6 | 6 | 37 |
| 7 | 7 | 10 |
| 8 | 8 | 8 |
| 9 | 9 | 6 |
| 10 | 10 | 8 |
| 11 | 11 | 2 |
| 12 | 13 | 1 |
| 13 | 14 | 5 |
| 14 | 15 | 2 |
| 15 | 16 | 3 |
| 16 | 17 | 4 |
| 17 | 18 | 2 |
| 18 | 21 | 1 |
| 19 | 22 | 1 |
| 20 | 25 | 1 |
| 21 | 27 | 1 |
| 22 | 31 | 1 |
| 23 | 33 | 1 |
| 24 | 34 | 1 |
| 25 | 43 | 1 |
| 26 | 50 | 2 |
| 27 | 51 | 1 |
| 28 | 56 | 2 |
| 29 | 105 | 1 |

Table1: User booking frequency

**Vehicle model id:**

It essentially identifies the type of vehicle driven.

* Compact Sedan (ID = 12) dominates (7,279 rentals) due to affordability, efficiency, and high demand.
* Luxury, sports, and hypercars are rare due to high rental costs and niche demand.
* Hatchbacks (ID = 1), SUVs (ID = 13), Large Trucks (ID = 36), and Sports Cars (ID = 70) have very low usage (only 1 rental each). Likely due to low demand, maintenance costs, or limited availability.
* Electric Vehicles (ID = 90 & 91) are present but limited.

|  |  |  |
| --- | --- | --- |
| number of rows | vehicle id | count of vehicle ids |
| 1 | 1 | 1 |
| 2 | 10 | 25 |
| 3 | 12 | 7279 |
| 4 | 13 | 1 |
| 5 | 17 | 7 |
| 6 | 23 | 75 |
| 7 | 24 | 318 |
| 8 | 28 | 406 |
| 9 | 30 | 5 |
| 10 | 36 | 1 |
| 11 | 54 | 16 |
| 12 | 64 | 18 |
| 13 | 65 | 445 |
| 14 | 70 | 1 |
| 15 | 85 | 572 |
| 16 | 86 | 31 |
| 17 | 87 | 116 |
| 18 | 89 | 591 |
| 19 | 90 | 85 |
| 20 | 91 | 7 |

Table2: vehicle id frequency

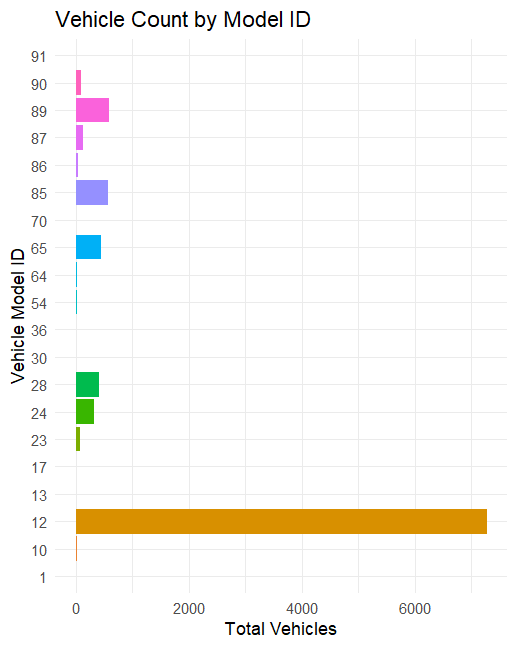


Fig3: vehicle id distribution

**Travel type id:**

It is classified into three categories, they are

* Point to point travel: 7909 rides
* Hourly rental: 1752 rides
* Long distance: 339 rides

Reasons for type of travel:

* Urban Commuting Behavior: Most users rely on point-to-point travel for daily commutes, errands, and short-distance rides within the city.
* Availability of Alternative Transport: Long-distance travelers may opt for trains, buses, or self-driving, reducing the need for rental services.
* Cost Sensitivity – Users prefer cost-effective options. Hourly rentals and long-distance rides are relatively expensive, leading to lower demand.

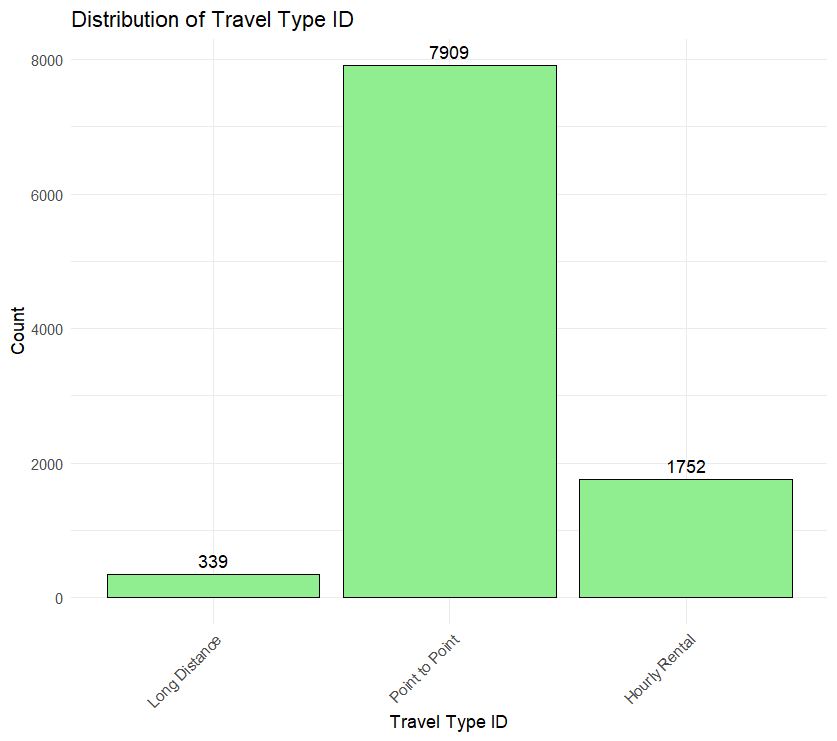


Fig4: Distribution of travel type id

**Package id:**

From fig5, we observe that there are 8248 NA values in data set

* The highest observations those are 800, refer to id=1 which means 4hrs and 40 kms
* Following 659 observations, refer to id =2 which means 8hrs and 80 kms
* By understanding the dataset, the package\_id and travel\_id are connected to each other, package\_id only have the values that travel\_id =3, which means customers who booked the car for hourly bases. In that case package\_id refers to how many hours they reserved the car.

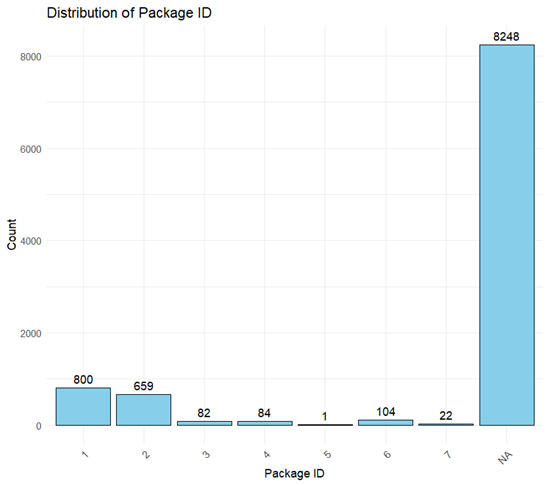


Fig5: Distribution of package id

**From area id and to area id:**

In the analysis of the from\_area\_id and to\_area\_id attributes from the dataset, the following key observations were made:

Unique Area IDs:

There are 523 unique from\_area\_id values.

There are 480 unique to\_area\_id values.

Missing Data:

For non-point-to-point travel types (hourly and long-distance), the from\_area\_id attribute contains 15 missing values, and the to\_area\_id attribute contains 2091 missing values.

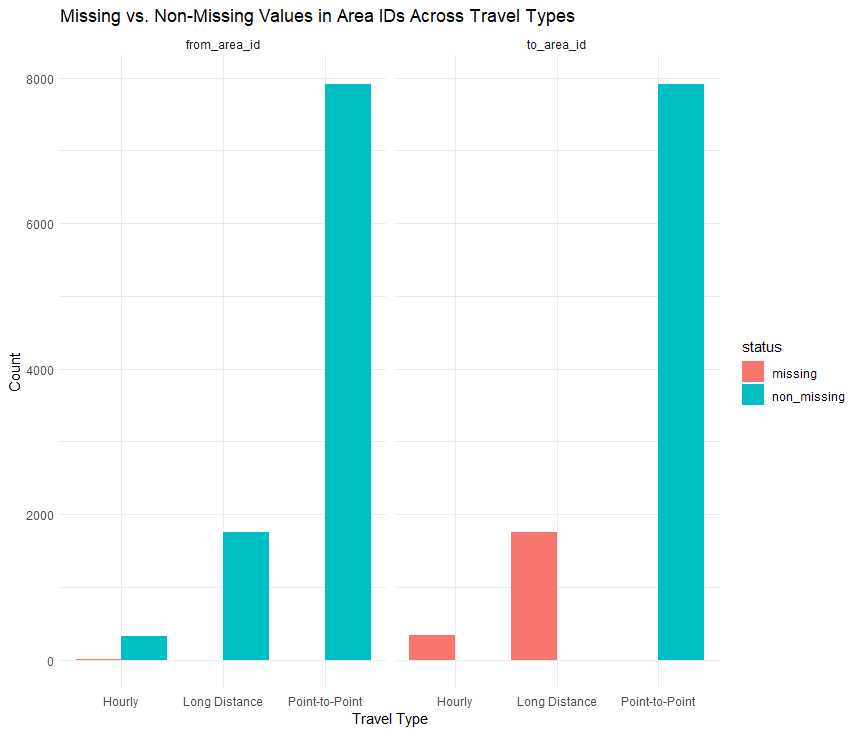


Fig 6: missing vs non-missing in area ids across travel types

Point-to-Point Travel:

When the dataset is filtered for point-to-point travel (travel type ID = 2), no missing values are present in either the from\_area\_id or to\_area\_id attributes. Both attributes contain valid data for point-to-point rides, which can be used in further analysis.

These observations indicate that while the from\_area\_id and to\_area\_id attributes are generally available for point-to-point travel, there is considerable missing data for non-point-to-point travel types, suggesting that this data should be handled separately in the analysis.

**From city id and To city id:**

* There are only 3 unique ids in the from\_city\_id and 68 in the to\_city\_id
* In the from\_city\_id there are 3706 observations and 336 observations in to\_city\_id
* There is not enough information from these attributes

**From Date and To Date:**

The data in these columns is initially in character format. Before analysis, the format is converted from “1/1/2013 22:33” into a proper date-time format. Then, two separate columns are created: one for the date and another for the hour, enabling better time-based analysis.

There are 10000 records in from\_date, only 5,822 trips have an actual recorded end time (to\_date count), meaning many trips missing completion timestamps.

Peak Months: August (1279 bookings) and July (1113 bookings) had the highest number of trip initiations, indicating increased demand during these months.

Low Activity: December (only 6 trips recorded) has significantly fewer trips, possibly due to incomplete data or seasonal effects.

|  |  |  |
| --- | --- | --- |
| Row Labels | Count of from\_date | Count of to\_date |
| Jan | 710 | 74 |
| Feb | 645 | 62 |
| Mar | 640 | 63 |
| Apr | 725 | 81 |
| May | 988 | 115 |
| Jun | 1035 | 200 |
| Jul | 1113 | 1094 |
| Aug | 1279 | 1277 |
| Sep | 1037 | 1032 |
| Oct | 1089 | 1085 |
| Nov | 733 | 733 |
| Dec | 6 | 6 |
| Grand Total | 10000 | 5822 |

Table3: distribution of observations in months

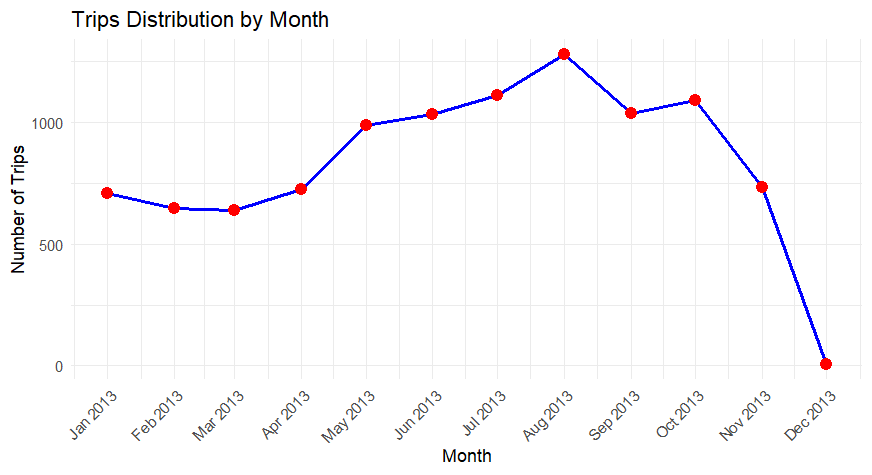


Fig7: trip distribution by from month

`

From fig Peak Hour:

* 6 PM (777 trips) has the highest number of trips, indicating a significant increase in demand during the evening, potentially reflecting people finishing work or other late-day activities.

Other High Activity Hours:

* 9 AM (711 trips) and 6 AM (710 trips) remain high, confirming strong demand during the early morning and morning commute hours.
* 7 AM (630 trips) and 12 PM (592 trips) also see considerable demand, indicating that mornings and afternoons are generally busier

Morning rush hours (6 AM and 9 AM) still show a high demand for trips, as expected in the commute window.

The early morning (1 AM and 2 AM) and late-night hours see significantly reduced trips, indicating lower demand during those hours.

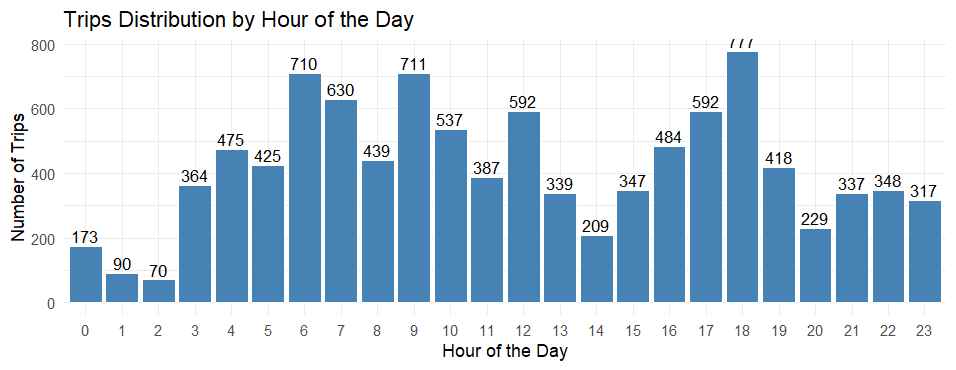


Fig8: distribution of trips (from date & to date) observations in day

**Bookings created:**

The data in these columns is initially in character format. Before analysis, the format is converted from “1/1/2013 22:33” into a proper date-time format. Then, two separate columns are created: one for the date and another for the hour, enabling better time-based analysis.

* The same number of records are seen as from the from\_date attribute in this booking created

|  |  |
| --- | --- |
| Row Labels | Count of booking\_created |
| Jan | 710 |
| Feb | 645 |
| Mar | 640 |
| Apr | 725 |
| May | 988 |
| Jun | 1035 |
| Jul | 1113 |
| Aug | 1279 |
| Sep | 1037 |
| Oct | 1089 |
| Nov | 733 |
| Dec | 6 |
| Grand Total | 10000 |

Table4: number of bookings made for each month

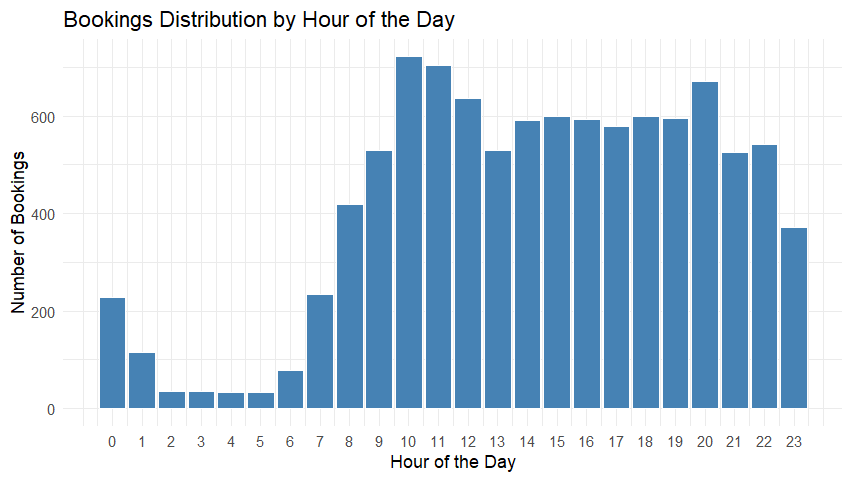


Fig9: Booking distribution by hour of the day

Peak Hours:

* Most bookings occurred between 10 AM and 11 AM with 722 and 705 bookings respectively.
* 6 PM to 8 PM also saw high booking numbers, particularly 600+ bookings each hour.

Least Bookings:

* 3 AM to 6 AM had the lowest bookings, with numbers between 32 and 78 bookings.

The morning peak around 9 AM to 10 AM might be related to the start of business hours or travel patterns.

The evening peak from 5 PM to 7 PM could indicate people booking rides for evening plans or post-work travel.

**Online booking and mobile site booking:**

From table5, it can say that,

* + 424 rides are booked by mobile
  + 3533 rides are booked through the online service
  + 6043 rides, we can represent it as in person booking
  + Customers are likely to book the rides in person in tradition way

Traditional Booking Dominates: With 6043 in-person bookings, it’s evident that customers still rely heavily on traditional methods of booking rides.

Mobile Booking is Least Preferred: The 424 mobile bookings suggest that customers are less likely to use mobile apps for booking rides compared to online services or in-person methods.

Online Booking Gaining Popularity: The 3533 online bookings reflect a growing trend of customers choosing online services, but they remain behind in-person bookings.

|  |  |  |  |
| --- | --- | --- | --- |
|  | mobile site booking | | |
| online booking |  | 0 | 1 |
| 0 | 6043 | 424 |
| 1 | 3533 | 0 |

Table5: number of bookings

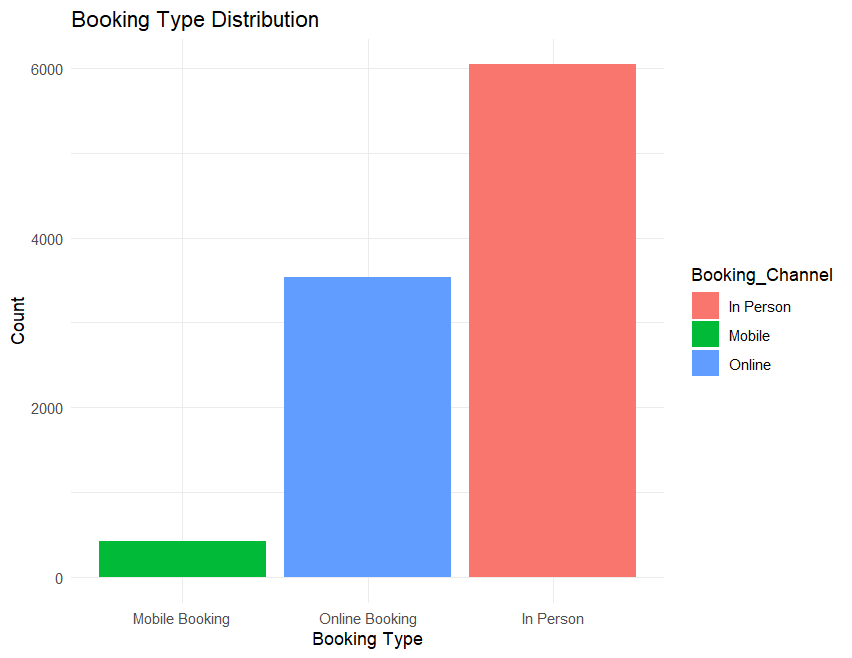


Fig11: booking type distribution

**from\_lat, from\_long, to\_lat, to\_long :**

The longitude and latitude values were initially in numerical format. Using the geosphere package, these coordinates were transformed, and the distance between the origin and destination coordinates was calculated. The computed distance values were then stored in a new column named distance\_traveled.

Similarly, the same number and types of values are missing for 'From Area ID' and 'To Area ID', which means these also refer to point-to-point travel

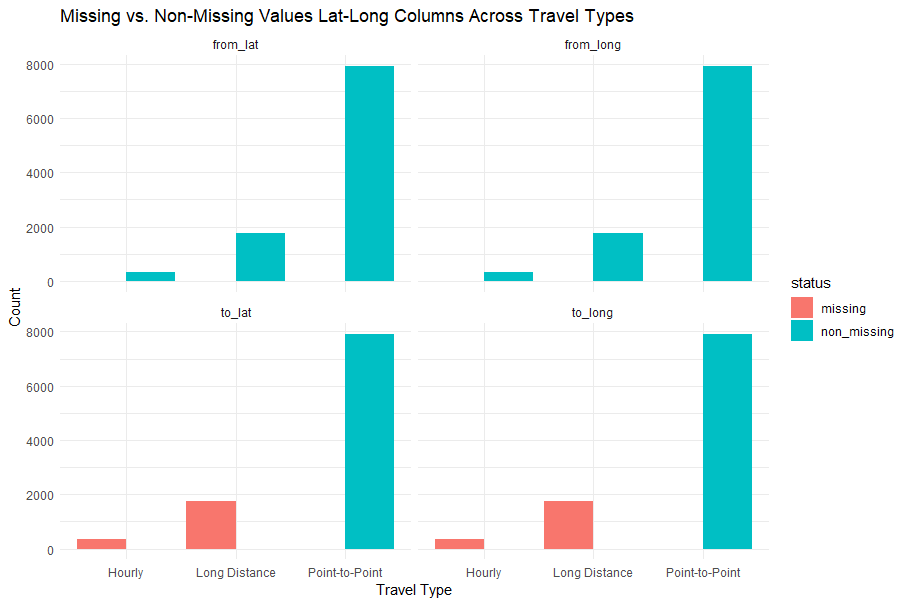


Fig 12: missing vs non\_missing values across travel types in lat-log

Smaller Distance Categories:

* 0-5 km has 714 trips, indicating that there are a good number of short trips.
* 5-10 km is also a relatively common range with 1,904 trips.

Longer Distances:

* 50-100 km has only 3 trips, suggesting that trips of this distance are quite rare.
* 100+ km has 0 trips, indicating that no bookings were made for distances over 100 km in the dataset.

Majority of the trips fall within the short to medium range (up to 50 km), with the most common distances being between 10 km to 50 km.

Long-distance travel (over 50 km) is extremely rare, with only 3 trips falling in the 50-100 km range and no trips exceeding 100 km.

The short-distance trips (under 10 km) are quite common, especially the 5-10 km range, which is likely to represent urban or nearby travel

The density plot of distances traveled (in km) shows a bimodal distribution, indicating two primary distance ranges for trips.

* The first peak, around 5-10 km, suggests that most trips are short-distance, likely within city limits.
* The second peak, around 20-25 km, indicates a moderate number of medium-range trips, possibly intercity rides.
* The density gradually decreases beyond 30 km, with very few trips exceeding 40 km, suggesting that long-distance travel is less common.

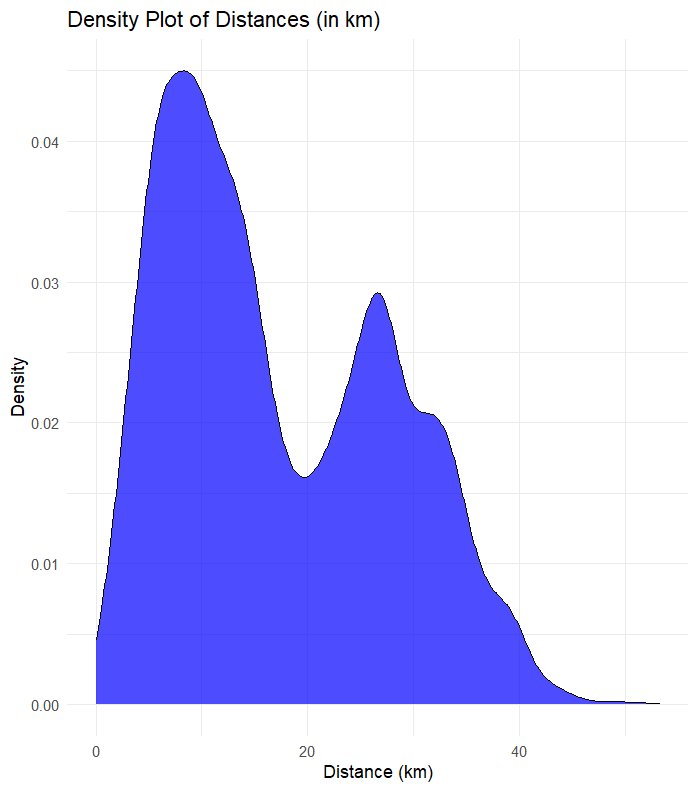


Fig13: density plot of distances in kms

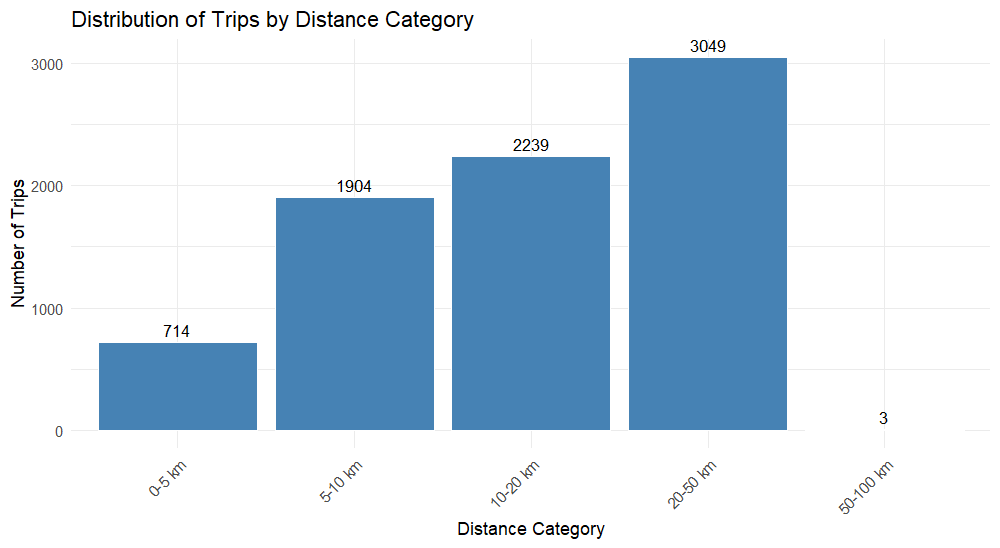


Fig14: distance of the rides by category

**Car cancellation:**

It's a binary variable (0 = No Cancellation, 1 = Cancellation).

* From fig, there 9257 rides are not cancelled
* Only 743 rides are cancelled

This relatively low cancellation rate indicates that most customers are likely satisfied with their bookings and trips, which could reflect well on the overall reliability and performance of the car rental service. However, the fact that cancellations do occur, even in such a small proportion, may warrant further investigation to understand the reasons behind them

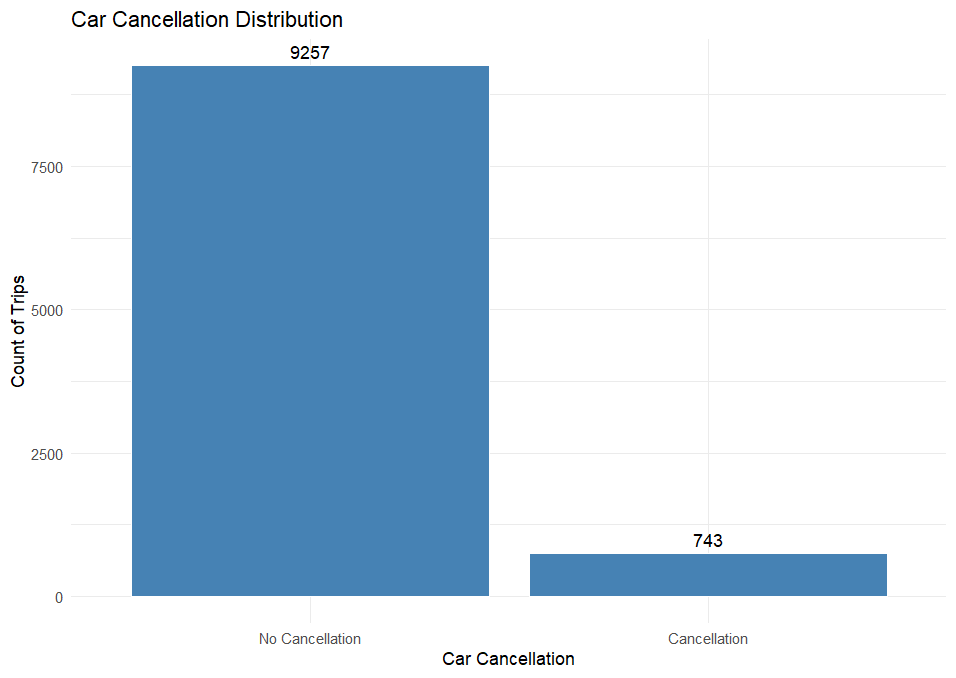


Fig15: car cancellation

**3.5. Checking zero’s:**

From Fig16, there are no zeros in the dataset except the binary variables.

They are online\_booking, mobile\_site\_booking, car\_cancellation

The from\_hour and booking\_hour variables contain zeros, which represent intentional bookings made between 12:00 AM and 12:59 AM. These values are not errors but indicate bookings that occurred precisely at midnight or within the first hour of the day.

A white background with black and blue text

AI-generated content may be incorrect.

Fig 16: count of zeros in the columns

**3.6. missing values:**

There NA values in the dataset, by taking colSums function the missing are observed in the Fig 17

**A close up of a computer code

AI-generated content may be incorrect.**

Fig17: Missing values

Missing value attributes:

* Package\_id, from\_area\_id, to\_area\_id, from\_city\_id, to\_city\_id, from\_lat, from\_long, to\_lat, to\_long, to\_date

A blue and pink grid

AI-generated content may be incorrect.

Fig18: Graphical Representation of missing values

From Fig18, we can identify that there are 12 types of patterns in the missing values from the SAR Rental data.

The SAR rental data is divided into three travel types: point-to-point, hourly, and long distance. The dataset can be split into two groups: one for point-to-point travel and another for hourly and long-distance travel combined. The distribution of missing data will be further explored in these groups to gain deeper insights into the patterns and causes of missing values.

A blue grid with black text

AI-generated content may be incorrect.Fig19: missing values in point-to-point data

A graph with a blue and pink graph

AI-generated content may be incorrect.

Fig20: missing values in hourly and long-distance data

Both figures demonstrate the distribution of missing data in travel type id 2 and travel type 1& travel type 3

Missing values in the key attributes:

* After a detailed analysis of the missing values, it was observed that 2,091 missing values are associated with the to\_area\_id, to\_lat, and to\_long attributes. These missing values are specific to travel types 1 (hourly rental) and 3 (long distance). This suggests that the company was unable to collect the corresponding values for these travel categories.
* imputing values for an entire set of categories could introduce bias into the model, as it may not accurately reflect the patterns specific to each travel type. Therefore, it is advisable to build a model using only the point-to-point travel data. This will help in ensuring a more accurate and unbiased model.
* By excluding travel types 1 and 3, the missing values in the from\_area\_id, from\_lat, and from\_long attributes, which account for 15 rows, are effectively eliminated. This approach ensures that only complete data is used for analysis, further improving the accuracy and reliability of the model.
* From\_city\_id attribute have 6294 which is 60% of missing vales and to\_city\_id have 9661 which means 90% of missing values of that columns. It is always preferring to avoid the data which has more than 20 percent of missing data.
* Given the issues related to missing data in certain travel types and the potential bias introduced by imputing values across categories, the analysis will now be focused on point-to-point travel data (travel type 2). By excluding travel types 1 (hourly rental) and 3 (long distance), where significant missing values exist for critical attributes, the dataset used for analysis will be more reliable and representative of the actual patterns.
* The focus on point-to-point travel ensures that the model is based on complete and consistent data, which minimizes the risk of introducing errors due to missing values or imputation. With this refined dataset, the analysis will provide more accurate and unbiased insights for predicting car cancellations, as the data for point-to-point travel is less prone to the issues observed in other travel categories. This step is essential for improving model performance and ensuring the validity of predictions.

**3.7. Handling of missing data:**

Handling missing data involves using methods to deal with gaps in a dataset, ensuring that incomplete information does not negatively affect analysis

**Deletion:**

Removing the From city id, to city id and to date:

From City ID: About 62.94% of the values for From City ID are missing. Given the high proportion of missing data, it would be difficult to impute this column reliably, so we decided to drop it from the analysis.

To City ID: Approximately 96.61% of the values for To City ID are missing. This extremely high percentage makes it impractical to retain this variable, so we chose to drop it from the dataset.

To Date: Around 41.78% of the values for To Date are missing. Although the missing percentage is lower than the others, it's still significant enough to impact the analysis. Given the nature of the variable, it was more efficient to drop it rather than impute.

Package id: it had 7909 missing rows in that variable. Attribute will be removed from the dataset as it is primarily associated with Travel Type ID 3 (Long Distance), which is excluded from the analysis. Since Travel Type ID 3 is not being considered for the current analysis, the Package\_id attribute will not contribute meaningful information. Removing this attribute will help simplify the dataset and focus the analysis on relevant travel types.

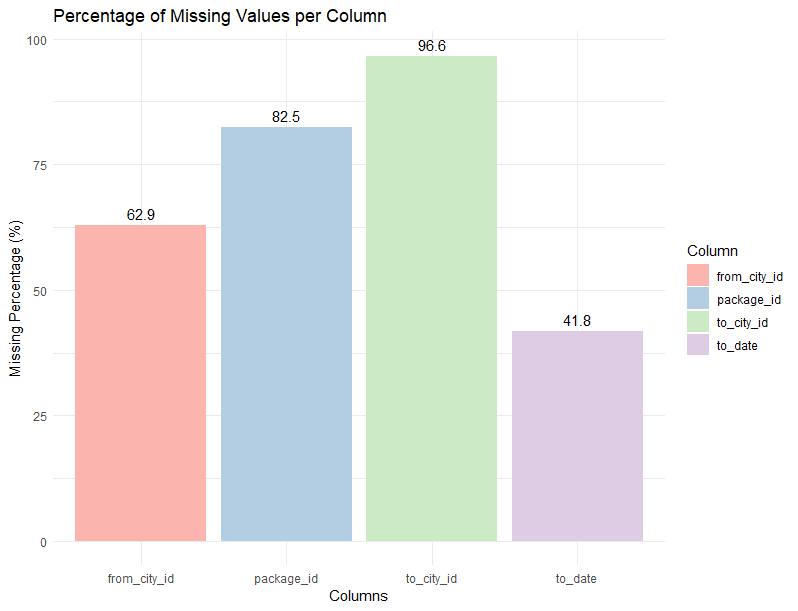


Fig21: percentage of missing values in package id, from city id, to city id and to date in the whole data

**4. Data transformation:**

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

Distance Calculation: The distances between the starting and destination points were calculated using the latitude and longitude values from the respective columns, and the results were stored in the new column distance\_travelled. This transformation enabled the analysis to focus on the actual distance traveled, which is more meaningful for predicting cancellations.

Date and Time Transformation: The original from\_date and booking\_created columns, which contained both date and time, were transformed into more specific features such as from\_month, from\_time, booking\_month, and booking\_time. These transformations simplify the dataset and allowed the model to focus on the month and time-related patterns without the need for handling complex date and time formats.

online\_booking and mobile\_site\_booking variables were transformed to create a new feature, in\_person\_booking. This transformation was based on the observation that a booking could either be made online, through a mobile site, or in-person. By combining the online and mobile booking data into an in-person booking indicator, the dataset is simplified

**5. Predictor analysis and relevancy**

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

**Continuous attributes:**

**Car cancellaton vs distance \_travelled**

Based on the Welch Two Sample t-test results, here's how you can interpret the findings:

Key Results:

* t-statistic: 15.454
* Degrees of Freedom (df): 827.01
* p-value: < 2.2e-16 (very small)

Interpretation:

* p-value < 0.05 (actually much smaller) indicates that the means of the predictor (distance\_travelled) are significantly different between the two categories of your outcome variable (Car\_Cancellation).
* The mean for group 0 (non-cancelled) is 17.73, while the mean for group 1 (cancelled) is 12.16. This suggests that the distance travelled by customers who did not cancel is significantly higher than the distance travelled by those who did cancel.

95% Confidence Interval:

* The difference in means is between 4.86 and 6.28, meaning you can be 95% confident that the true difference lies within this range.

Conclusion:

Distance travel is a relevant predictor for car cancellations, as there is a statistically significant difference in the average distance travelled between the two cancellation groups.

**Categorical attributes:**

**User\_id vs car cancellation:**

Upon analyzing the distribution of 'user\_id' in relation to 'Car\_Cancellation', it's evident that the 'user\_id' variable exhibits high cardinality, with 5,841 users having only a single ride and a rapid decline in the number of users with multiple rides. This high cardinality suggests that 'user\_id' is not a relevant predictor for 'Car\_Cancellation'.

High Cardinality: The 'user\_id' variable contains a vast number of unique identifiers, leading to a high-dimensional feature space. Incorporating such variables can increase computational complexity and may not contribute meaningful information to the model

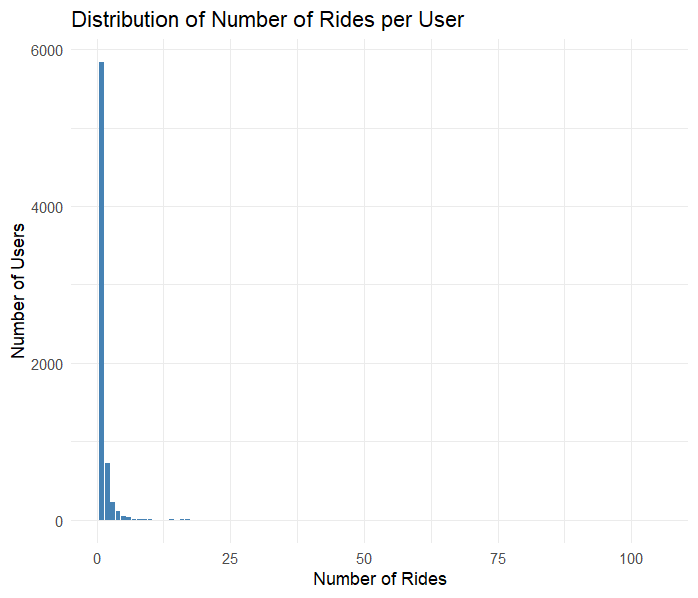


Fig 22: number of rides given by each user

Vehicle\_id vs car cancellation:

The cancel\_rate analysis reveals that certain vehicle models have higher cancellation rates, such as vehicle model 89 (with a cancellation rate of 0.129) and vehicle model 91 (with a cancellation rate of 0.143). These findings suggest that vehicle models might be associated with cancellation likelihood.

Keeping vehicle\_model\_id will allow you to track and examine cancellation patterns more effectively, helping to pinpoint which specific models are more prone to cancellations. addressing issues with particular vehicle types.

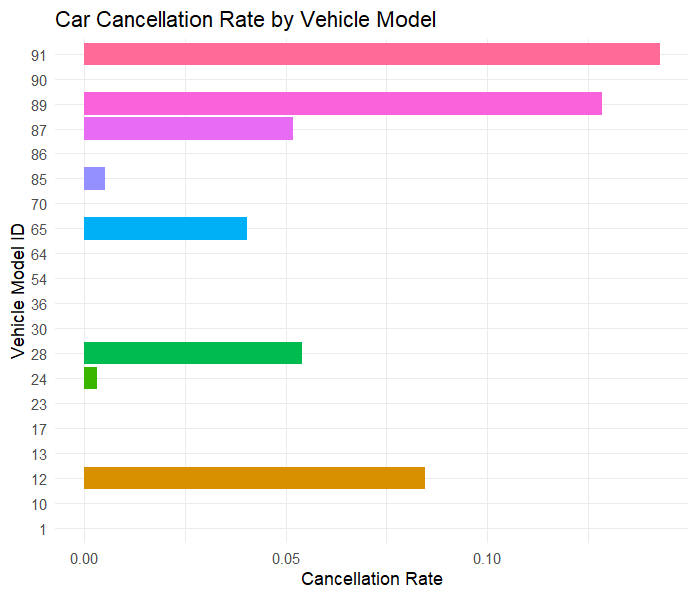


Fig23: car cancellation rate by vehicle model

From\_area\_id, to\_area\_id vs car\_cancellation:

The high cardinality of from\_area\_id and to\_area\_id is a significant concern for including them in the dataset. With 523 unique from\_area\_id values and 480 unique to\_area\_id values, these columns introduce a large number of categories into the model. This can lead to overfitting, where the model memorizes the data instead of generalizing well. It can also create sparse categories, making it difficult for the model to accurately predict cancellations, as some areas may have very few observations. Additionally, this increases the computational complexity of the model, slowing down training and reducing its efficiency.

Including these two features might also cause multicollinearity, as the from\_area\_id and to\_area\_id could be highly correlated, making it harder for the model to distinguish between them

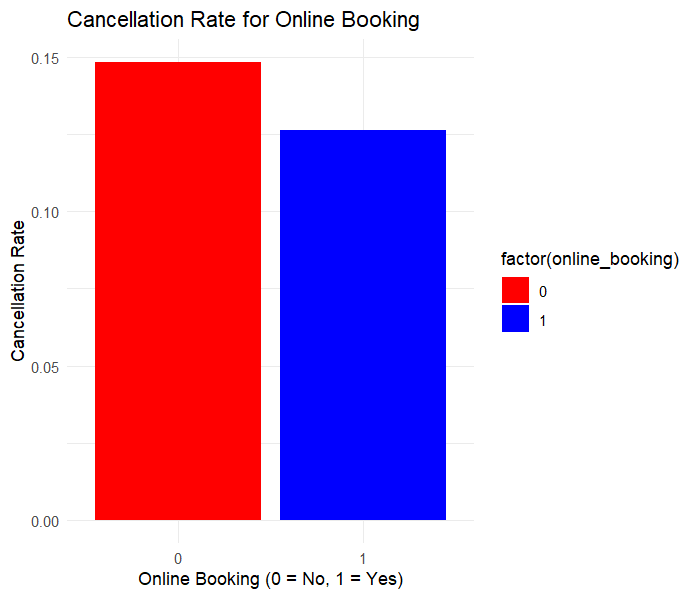
A screenshot of a computer code

AI-generated content may be incorrect.

Fig24: from area ids and to area ids

Online\_booking , mobile\_booking, inperson\_bookingvs car\_cancellation:

including online\_booking and mobile\_site\_booking in the dataset is justified because they provide valuable insights into the booking methods used by customers and their relationship with car cancellations. The cancellation rates for each of these booking types differ, with mobile site bookings showing a higher cancellation rate of 0.149, compared to online bookings, which have a cancellation rate of 0.127. Additionally, customers who did not use mobile bookings have a much lower cancellation rate of 0.0386. These variables capture important behavioral patterns that could be indicative of factors influencing cancellations. By retaining these features, the model can learn the impact of booking methods on cancellations, potentially improving predictive accuracy. In person booking had 0.152 cancellation rate.

A graph of a number of bars

AI-generated content may be incorrect.

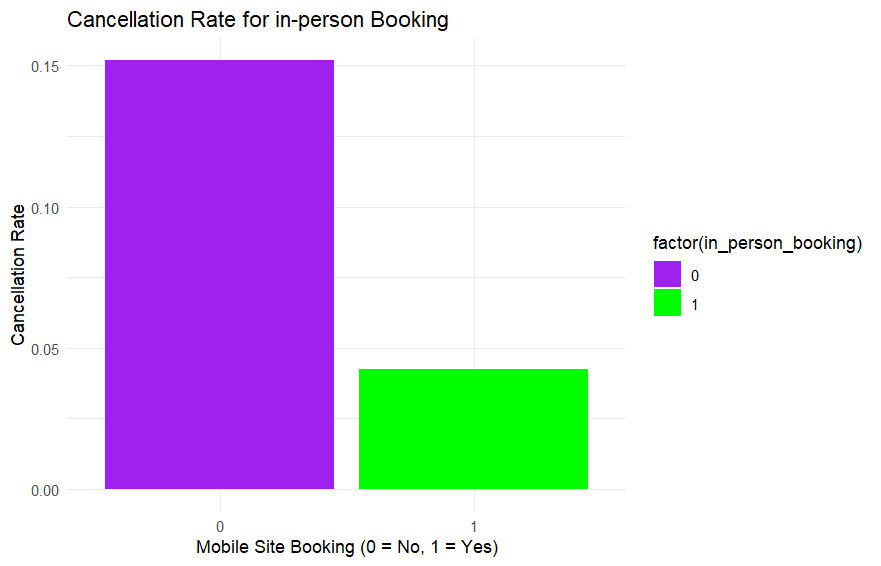


Fig25: cancellation rates of online booking, mobile site booking and in person booking

From\_time, booking\_time vs car\_cancellation

The attributes from\_time and booking\_time are important predictors in understanding car cancellations. The cancellation rate varies significantly across different from\_time values, indicating that the time a trip starts affects whether it gets canceled. For example, the cancellation rate is relatively low in the early hours, such as 2 AM (1.52%) and 3 AM (2.08%), but increases significantly in the evening and late night, reaching 16.3% at 6 PM and 16.9% at 7 PM. This suggests that customers may be more likely to cancel during peak travel hours, possibly due to high demand, traffic conditions, or long wait times. Additionally, the cancellation rate is highest between 5 PM and 8 PM, which may indicate a mismatch between demand and driver availability at these times.

Similarly, booking\_time also plays a crucial role in cancellation trends. The cancellation rate is highest in the early morning, with 22.6% at 2 AM and 18.8% at 1 AM, suggesting that bookings made during these hours are more likely to be canceled. In contrast, bookings made during daytime hours have relatively lower cancellation rates, such as 4.41% at 11 AM and 5.74% at 8 AM. This trend could indicate that late-night or very early morning bookings are more uncertain, potentially due to last-minute changes by customers or unavailability of drivers. By including from\_time and booking\_time in the dataset, we can better analyze peak cancellation periods and identify strategies to minimize cancellations, such as adjusting pricing, improving driver availability, or implementing customer reminders for high-risk time slots.

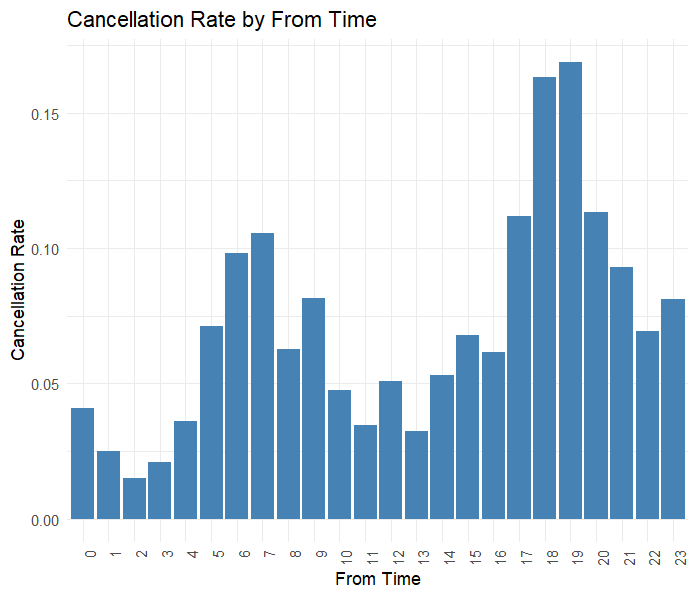


Fig26: cancellation rate by from time

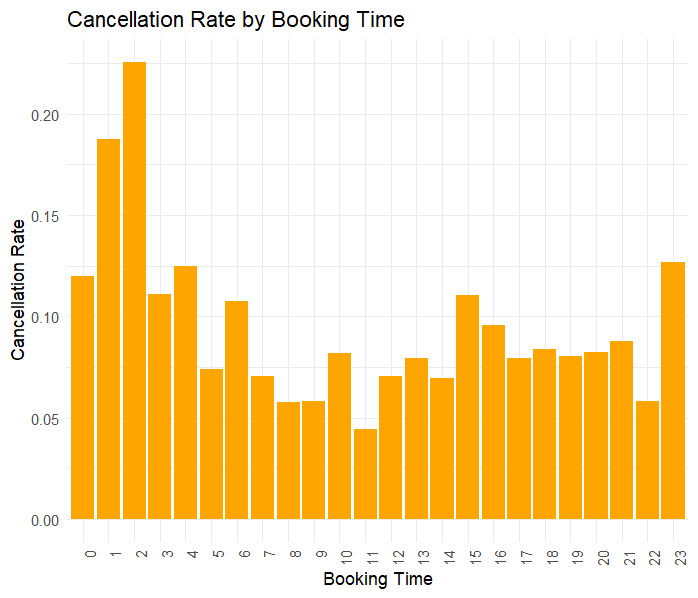


Fig27:cancellation rate by booking time

Booking\_time\_month, from\_time\_month vs car\_cancellation

The variables booking\_time\_month and from\_data\_month provide insights into seasonal trends affecting car cancellations. The analysis shows that cancellation rates vary significantly across months. For instance, October (booking\_time\_month = 10) has the highest cancellation rate at 19.2%, followed by November (11) at 17.8%, while January (1) has the lowest at 1.51%. Similarly, cancellations based on from\_data\_month show a peak in November (21.5%) and a low in January (1.4%)

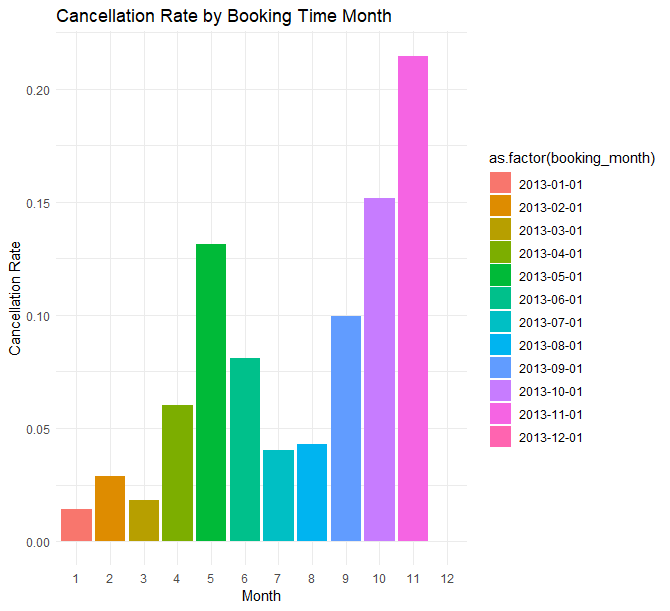


Fig28: cancellation rate by booking month

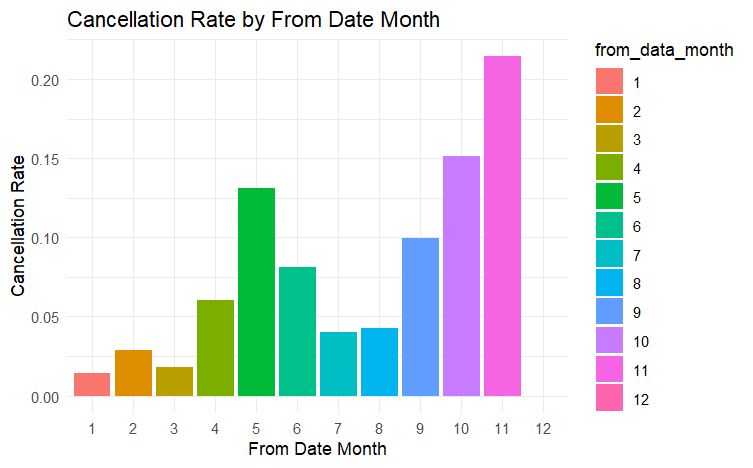


Fig29: cancellation rate by from month

**6. DIMENSION REDUCTION**

To enhance model performance and eliminate redundant or irrelevant features, we have reduced dimensions by removing variables that contribute little to variance in the data. The following features have been dropped based on predictor analysis and relevancy:

Removing row\_id and user\_id: These variables serve as unique identifiers for records and users but do not contribute to predicting car cancellations. They do not provide any meaningful patterns or insights into our analysis.

Removing travel\_type\_id: Since we are only analyzing point-to-point travel (travel\_type\_id = 2), this variable becomes redundant. Every row already belongs to the same category, so keeping this variable does not add value to the analysis.

Removing from\_area\_id and to\_area\_id: These variables represent geographical locations, but we have already derived meaningful insights using alternative attributes such as distance\_travelled. Retaining these identifiers does not enhance our ability to predict cancellations.

Removing from\_lat, from\_long, to\_lat, and to\_long: These latitude and longitude values have been used to calculate the distance\_travelled variable, which provides a more interpretable measure. Keeping raw coordinates is unnecessary and does not contribute further to the analysis.

Removing booking\_created, from\_date: The booking\_created column can be removed as its information is already captured in the booking\_time and booking\_month columns, making it redundant. Removing it streamlines the dataset without losing any valuable data.

Similarly, the from\_date column should be excluded because its details are already represented by the from\_time and from\_month columns

**7. DATA PARTITION**

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

We will divide the dataset into three parts:

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

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

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

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

The validation set helps improve model performance before testing.

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

A computer screen with blue and white text

AI-generated content may be incorrect.

Fig30: partition and proportions of car cancellation in all sets of data

**8. MODEL SELECTION**

Logistic Regression, Decision Tree, and Random Forest as our predictive models to help SAR anticipate cancellations and identify key factors that contribute to them. Each model provides unique advantages in understanding and mitigating ride cancellations.

Logistic Regression is selected for its simplicity and interpretability. It allows SAR to understand how individual factors—such as booking method, time of booking, and ride type—affect the likelihood of cancellation. Since it provides probability scores, SAR can take preventive actions, like offering incentives or prioritizing certain bookings to reduce cancellations.

Decision Tree is included because it provides a clear, rule-based structure that SAR can use to make actionable decisions. It helps in identifying critical decision points that lead to cancellations, such as whether a ride was booked through a mobile app or if a trip falls within a specific time frame. Decision Trees are intuitive and can assist SAR in setting clear policies, like adjusting cancellation penalties or improving customer support for high-risk bookings.

Random Forest is chosen for its superior accuracy and ability to detect complex patterns. Unlike Logistic Regression, it can capture nonlinear relationships between features and handle a high number of variables effectively. Random Forest also ranks feature importance, helping SAR make strategic improvements, such as optimizing driver availability, adjusting pricing, or modifying booking policies to reduce last-minute cancellations.

By using Logistic Regression for interpretability, Decision Tree for clear decision-making, and Random Forest for deeper insights, SAR can not only predict ride cancellations but also understand their root causes. This enables proactive measures to improve customer satisfaction, enhance service reliability, and minimize revenue loss

**Class Balancing:**

Our training data consists of 5,090 records where car cancellation is "0" and only 466 records where car cancellation is "1," indicating a significant class imbalance. To ensure a balanced dataset for model training, we will apply oversampling, a technique that increases the number of minority class instances (car\_cancellation = 1) by generating random samples. This process will create an equal distribution, resulting in 5,090 records for both cancellations and non-cancellations, making the dataset suitable for training our model.

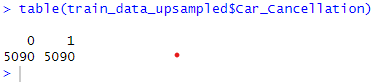


Fig31: upsampling

**Correlation in between attributes:**

The variables booking\_month and from\_month exhibit a perfect correlation of 1.000, meaning they convey the same information. Both represent the month related to the booking and the trip scheduling, respectively, and are always identical for each observation. Keeping both variables in the model would create redundancy, potentially leading to multicollinearity, which could inflate the variance of coefficient estimates and make it harder to interpret the individual contributions of each predictor. Therefore, removing one of these columns would simplify the model without sacrificing any valuable information, improving model efficiency, interpretability, and reducing the risk of overfitting.

Similarly, the online\_booking and in\_person\_booking variables show a very high negative correlation of -0.906. This indicates that when one variable increases, the other decreases significantly, suggesting that they capture mutually exclusive information about how a booking was made. Given that it is unlikely for a booking to be both online and in-person at the same time, these variables are inversely related indicators of the same underlying behavior. Removing one of these columns would eliminate redundancy, enhance model efficiency, prevent multicollinearity, and improve interpretability, ultimately leading to a simpler, more robust model.

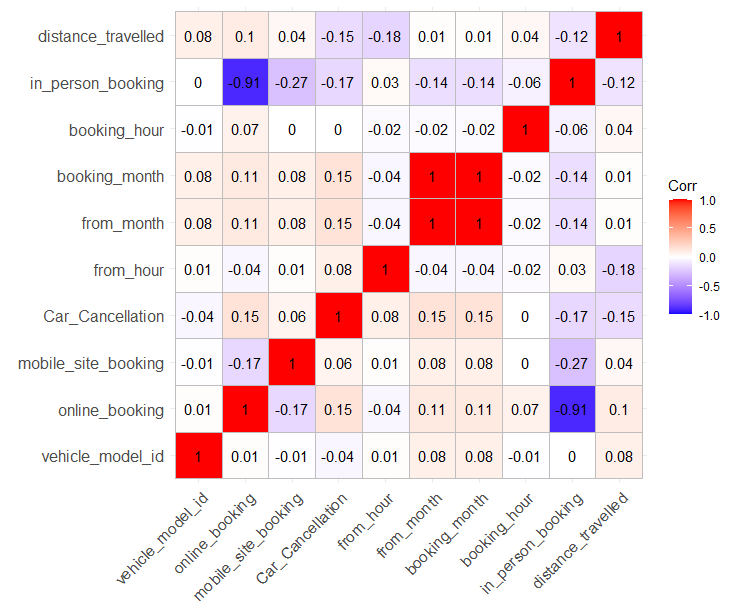


Fig 32: correlation matrix

**Model fitting**, **validation accuracy and test accuracy:**

**Decision tree:**

The rpart() function is used to build the decision tree model, with "car cancellation" as the target variable and the classification method applied. The rpart.control() function is utilized to set parameters for controlling the tree's complexity and the minimum number of samples required for node splitting. These settings are then applied to evaluate the model on the validation set.

****

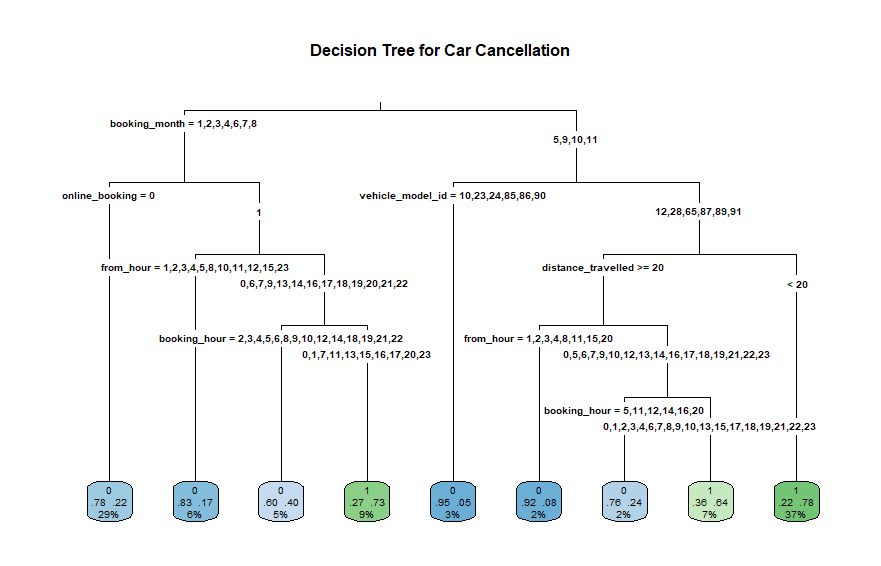
****

Fig33: decision tree

Model Predictions and Business Impact

The decision tree model was trained to predict whether a new ride would be canceled (Car\_Cancellation). Based on the confusion matrix:

778 rides were correctly classified as non-canceled (True Negatives).

74 rides were correctly classified as canceled (True Positives).

311 rides were misclassified as canceled but were actually not canceled (False Positives).

22 rides were misclassified as not canceled but were actually canceled (False Negatives).

Accuracy: The model correctly predicted the outcome 71.9% of the time.

Sensitivity (Recall): 77.08%. This is the proportion of actual positive cases (cancellations) that were correctly identified by the model. While this is a decent result, it could be improved further to catch more positive cases.

Specificity: 71.44%. This is the proportion of actual negative cases (no cancellations) that were correctly identified. The model is performing better at correctly predicting non-cancellation cases compared to cancellations.

**A screenshot of a computer

AI-generated content may be incorrect.**

Fig34: confusion matrix for decision tree

**Logistic regression:**

For the logistic regression model, a balanced training dataset is utilized to fit the model using the glm() function with the family = binomial argument for binary classification of the target variable "car cancellation" (0 or 1). To validate the model on the validation set, the predict() function is used, followed by applying the confusionMatrix() function to assess its performance. The same approach is then applied to the test dataset for evaluation.

****

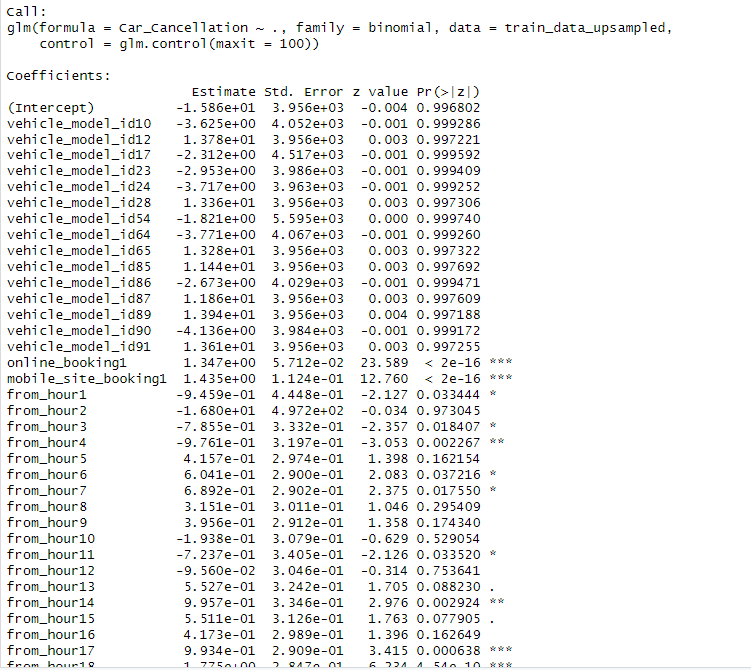
****

Fig35: logistic model

The model predicts whether a car cancellation will occur (1) or not (0) based on various factors.

Accuracy: The model correctly classified 75.19% of the observations, meaning it has a good overall performance in terms of predicting both cancellations and non-cancellations.

Sensitivity (also known as Recall or True Positive Rate): 0.78 – This indicates that the model correctly identifies 78.12% of actual cancellations (1's). In other words, it does a decent job of detecting cancellations, but it misses about 21.55% of cases where a cancellation occurred.

Specificity (also known as True Negative Rate): 0.74 – This means the model correctly identifies 74.93% of instances where there was no cancellation (0's). It has a slightly lower ability to correctly predict non-cancellations compared to cancellations.

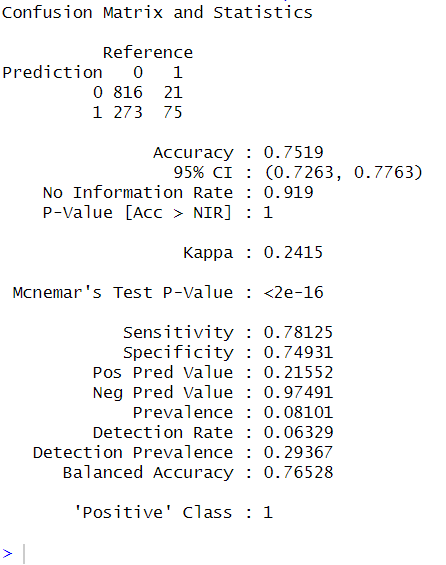
****

Fig35: confusion matrix for logistic regression

**Random forest:**

For the random forest model, a balanced training dataset is used to fit the model using the ranger() function. The model is trained with 100 trees, using predictors such as vehicle\_model\_id, distance\_travelled, online\_booking, mobile\_site\_booking, from\_time, booking\_time\_month, and booking\_time to classify the target variable "Car\_Cancellation" (0 or 1).

For validation, predictions are made on the validation set using the predict() function, and the model's performance is evaluated using the confusionMatrix() function. The same process is then applied to the test dataset to assess its final performance.

****

The goal of this analysis is to determine whether a new ride will be canceled so that preventive actions can be taken to reduce cancellations and improve customer satisfaction.

Overall Accuracy (85.15%): The model performs well in distinguishing between canceled and non-canceled rides. However, since cancellations are less frequent, accuracy alone is not enough to evaluate effectiveness.

Sensitivity (65.62%): The model correctly identifies 65.6% of actual cancellations, meaning that it misses about 37.5% of cases where a cancellation occurs. This is a critical gap because failing to predict cancellations means lost opportunities for proactive intervention.

Specificity (86.86%): The model does well in identifying non-cancelled rides, which helps avoid false alarms. However, since the focus is on predicting cancellations, specificity is less of a priority than sensitivity.

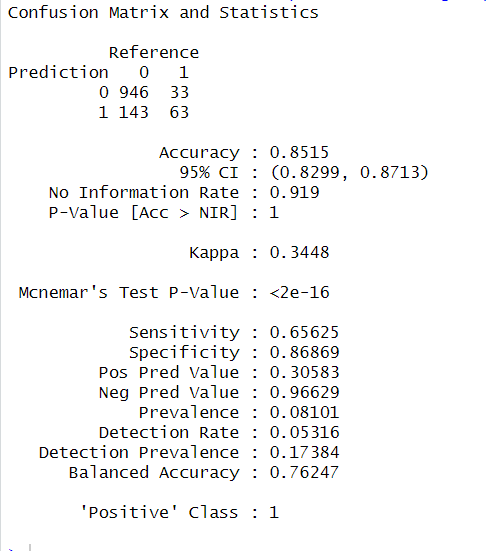
****

Fig36: confusion matrix for random forest

**Report on model performances:**

compare the decision tree, regression, and random forest models based on key performance metrics.

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Decision Tree | Logistic Regression | Random Forest |
| Accuracy | 0.7190 | 0.7519 | 0.8515 |
| Kappa | 0.2045 | 0.2415 | 0.3448 |
| Sensitivity (Recall) | 0.7708 | 0.7812 | 0.6562 |
| Specificity | 0.7144 | 0.7493 | 0.8686 |
| Positive Predictive Value (Precision) | 0.1922 | 0.2155 | 0.3058 |
| Negative Predictive Value | 0.9725 | 0.9749 | 0.9662 |
| Balanced Accuracy | 0.7426 | 0.7652 | 0.7624 |

Table6: model performances

**Best Model Selection for Car Cancellation Prediction:**

Business Goal: Since we are predicting car cancellations (class 1), the most important metric is sensitivity (recall) the ability to correctly identify cancellations.

Sensitivity: The regression model (0.7812) and decision tree (0.7708) outperform random forest (0.6562) in capturing actual cancellations.

Precision (PPV): Random Forest (0.3058) does better than the others, meaning when it predicts a cancellation, it's more likely correct.

Balanced Accuracy: Regression (0.7652) performs slightly better than decision tree (0.7426) and random forest (0.7624)

The regression model is the best choice for predicting car cancellations because it achieves the highest sensitivity (0.7812), meaning it correctly identifies the most cancellations compared to the other models. Since the primary business goal is to reduce cancellations, detecting them accurately is crucial for taking proactive measures. Additionally, the regression model has the highest balanced accuracy (0.7652), ensuring a good trade-off between sensitivity and specificity. While the random forest model has a higher precision (0.3058), it sacrifices recall (0.652), meaning it misses many actual cancellations. The decision tree model, although close to the regression model in sensitivity, has lower overall accuracy and balanced accuracy, making it a less optimal choice. Furthermore, the regression model's negative predictive value (0.9749) is the highest, meaning when it predicts no cancellation, it is highly reliable. Given these factors, the regression model is the best fit for SAR’s business needs, as it enables better identification of cancellations while maintaining overall predictive reliability.

**Model Improvement** **and** **Evaluation:**

To enhance the performance of the logistic regression model for predicting car cancellations, the classification threshold was adjusted to 0.4 instead of the default 0.5. This adjustment was made to increase sensitivity (recall) and improve the detection of potential cancellations.



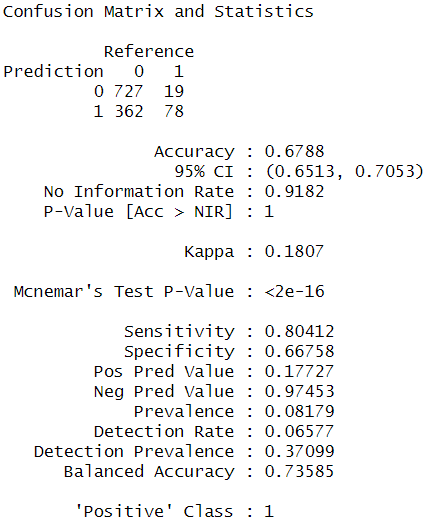


Fig36: confusion matrix on test data

Key Performance Metrics (Threshold = 0.4):

Sensitivity (Recall) (80.41%): The model correctly identifies most actual cancellations.

Specificity (66.75%): The model correctly classifies non-cancellations but allows more false positives.

Accuracy (67.88%): Acceptable given the class imbalance in the dataset.

Negative Predictive Value (NPV) (97.45%): High reliability in predicting non-cancellations.

Justification of threshold decreasing 0.5 to 0.4:

Increased Sensitivity: With a sensitivity of 80.41%, the model successfully detects a large proportion of actual cancellations. This ensures that SAR can proactively address cancellation risks before they impact customers.

Balancing Trade-offs: Lowering the threshold increases the number of false positives (predicting cancellations that don’t actually happen). However, the cost of incorrectly predicting a cancellation is far lower than failing to predict an actual cancellation, which directly affects customer trust and company revenue. Therefore, prioritizing higher sensitivity is a strategic decision for minimizing cancellations.

Model summary:

From the model summary, the following variables have p-values less than 0.05, indicating that they significantly affect the prediction of car cancellations:

Distance Travelled: The negative coefficient suggests that longer distances are associated with a lower likelihood of cancellation.

Online Booking: A positive coefficient show that online bookings increase the likelihood of cancellations.

Mobile Site Booking: Similarly, mobile site bookings have a positive effect on cancellations.

From Time Variables: Various "from\_hour" variables are significant, with coefficients indicating the effect of different times on cancellations.

Booking Time Variables: Numerous "booking\_hour" variables have significant effects, with higher values generally correlating with higher likelihoods of cancellation.

These variables are statistically significant in the model, meaning they play a crucial role in predicting car cancellations and should be carefully considered

The model evaluation for predicting car cancellations (positive class = 1) shows an accuracy of 67.59%, with a balanced accuracy of 75.24%, indicating a fairly reliable performance in identifying both cancellations and non-cancellations. The model's sensitivity is 84.27%, meaning it correctly detects approximately 84% of the cancellations, which aligns with the business goal of identifying and mitigating cancellations.

|  |  |  |
| --- | --- | --- |
| Highly Significant Variables (p < 0.001) | Moderately Significant Variables (0.001 ≤ p < 0.01) | Significant Variables (0.01 ≤ p < 0.05) |
| online\_booking1 | from\_hour14 | From Hour 1 |
| Mobile Site Booking | Booking Hour 9 | From Hour 3 |
| From Hour 17 | Booking Hour 10 | From Hour 4 |
| From Hour 18 | Booking Hour 11 | From Hour 6 |
| From Hour 19 | Booking Hour 12 | From Hour 7 |
| From Hour 20 | Booking Hour 13 | From Hour 11 |
| From Hour 21 | Booking Hour 14 | Booking Hour 19 |
| Booking Month 2 | Booking Hour 15 | Booking Hour 20 |
| Booking Month 3 | Booking Hour 16 | Booking Hour 21 |
| Booking Month 4 | Booking Hour 17 |  |
| Booking Month 5 | Booking Hour 18 |  |
| Booking Month 6 | Booking Hour 2 |  |
| Booking Month 7 |  |  |
| Booking Month 8 |  |  |
| Booking Month 9 |  |  |
| Booking Month 10 |  |  |
| Booking Month 11 |  |  |
| Distance Travelled |  |  |

**Observation:**

After improving the model, the sensitivity increased to 80.41%, which aligns with the analytical goal of accurately predicting ride cancellations.

Key Factors Influencing Cancellations:

Booking Platform:

* Online Booking and Mobile Site Booking are highly significant predictors, indicating that users who book through these platforms are more likely to experience cancellations. This could be due to higher booking convenience leading to more speculative or last-minute bookings, increasing the likelihood of cancellations.

Peak Cancellation Hours:

* Most cancellations occur for rides scheduled between 5 PM and 9 PM (from\_time), possibly due to traffic congestion, driver availability, or high demand.
* Booking hours 9 AM - 9 PM also show significance, meaning when the ride is booked plays a role in whether it gets canceled.

Booking Month:

* All months (February - November) significantly influence cancellations, indicating that cancellations occur consistently throughout the year rather than being seasonal.

Distance Traveled:

* Longer distances correlate with higher cancellation rates. Drivers might be reluctant to accept longer trips due to fuel costs, time commitment, or potential difficulties in getting return rides.

**Conclusion:**

This study effectively addresses SAR’s long-standing issue of ride cancellations by leveraging historical data and predictive modeling. Among the models tested—Logistic Regression, Random Forest, and random forest—Logistic Regression emerged as the best, achieving 80.41% sensitivity and 67.3% accuracy, making it the most effective for identifying potential cancellations. Key factors influencing cancellations include online and mobile bookings, peak-hour trips (5 PM – 9 PM), long-distance rides, and consistent cancellation trends across all months.

To mitigate cancellations, SAR should focus on driver incentives for peak hours, re-confirmation mechanisms for long-lead bookings, optimizing dispatch algorithms, and refining pricing strategies for long trips. Continuous monitoring and retraining of the model will ensure adaptability to evolving trends. These insights will help SAR reduce cancellations, improve customer satisfaction, and enhance operational efficiency, making data-driven decisions for long-term success.