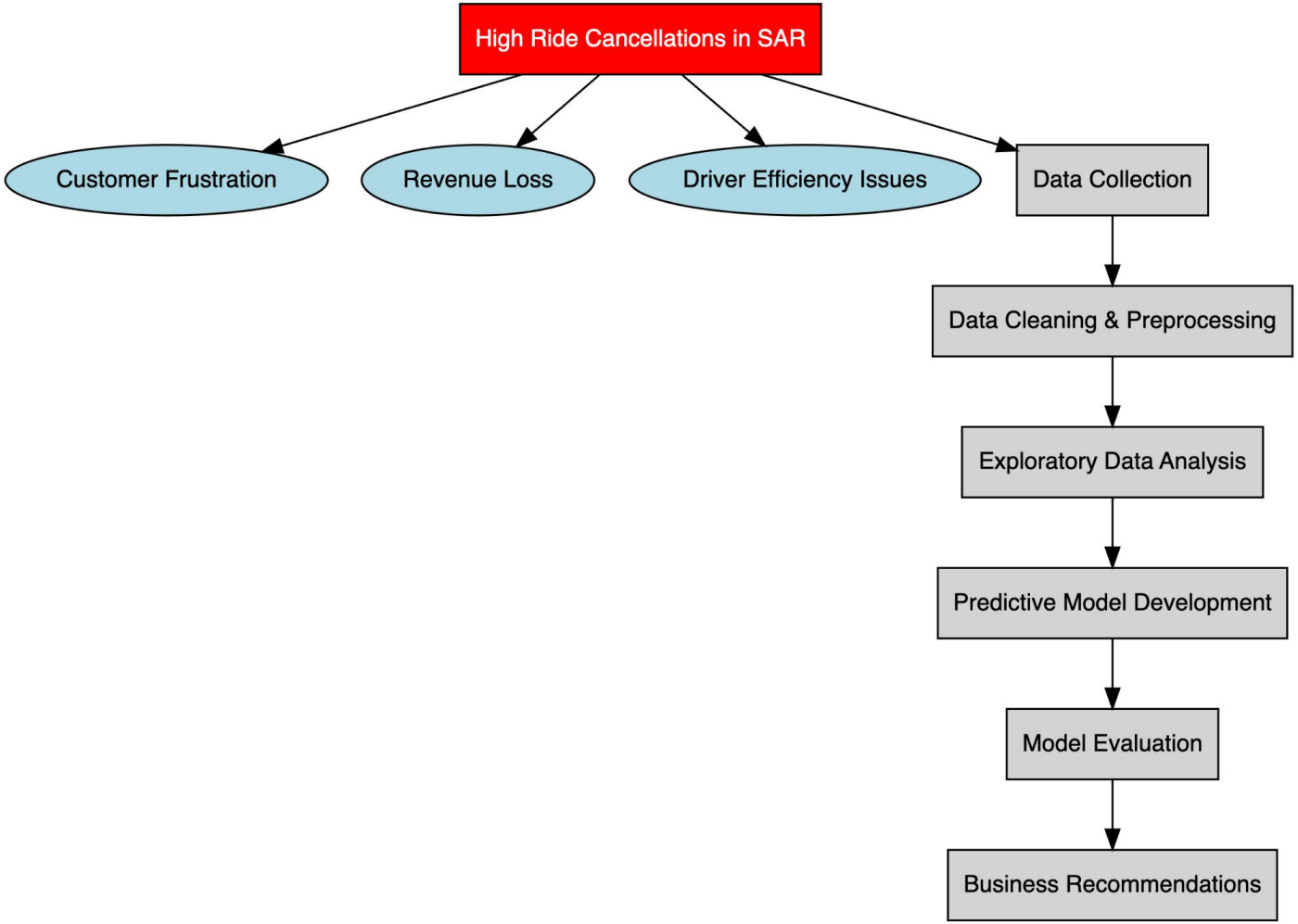
San Francisco Auto Rental Analytics

Name: Sai pranup

Instructor: Dr. Ali Ovlia

Why does ride cancellations matter for SAR?


High Ride Cancellations in SAR
Customer Frustration
Revenue Loss
Driver Efficiency Issues
Data Collection
Data Cleaning & Preprocessing
Exploratory Data Analysis
Predictive Model Development
Model Evaluation
Business Recommendations

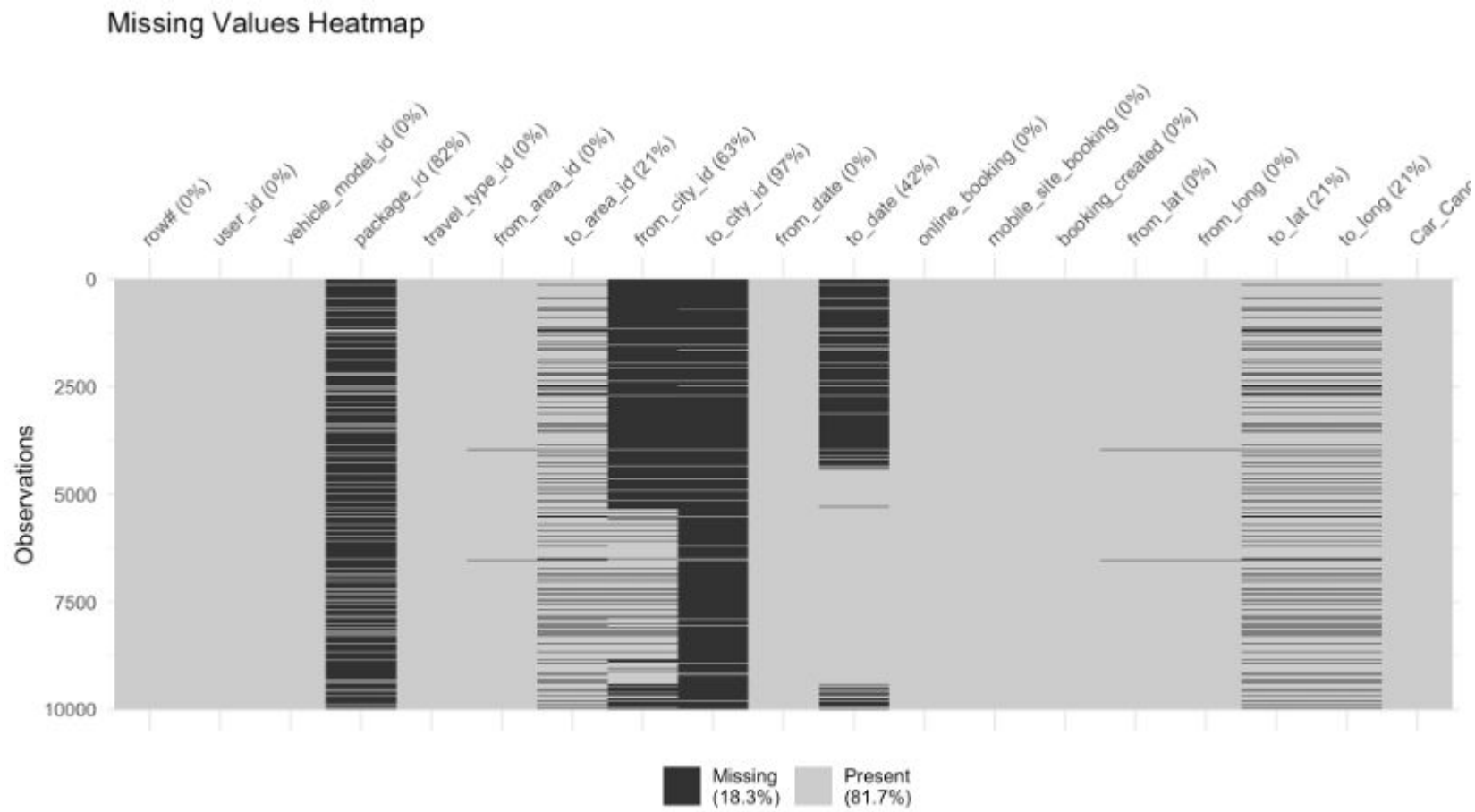
Attribute Name	Description
Row ID	Unique identifier for each record.
User ID	Customer ID, some users booked multiple times.
Vehicle Model ID	Represents the type of vehicle (also identifies driver).
Travel Type ID	Type of travel (long-distance, hourly, etc.).
Package ID	Travel package (3hrs, 4hrs, etc.).
From/To Area & City	Identifies the starting & ending locations.
From/To Date	Timestamp of trip start & end.
Booking Type	Online & Mobile booking (binary).
Car Cancellation	Target variable (0 = not canceled, 1 = canceled).

The dataset consists of **10,000 ride bookings** from SAR in 2013 with **19 key attributes**.

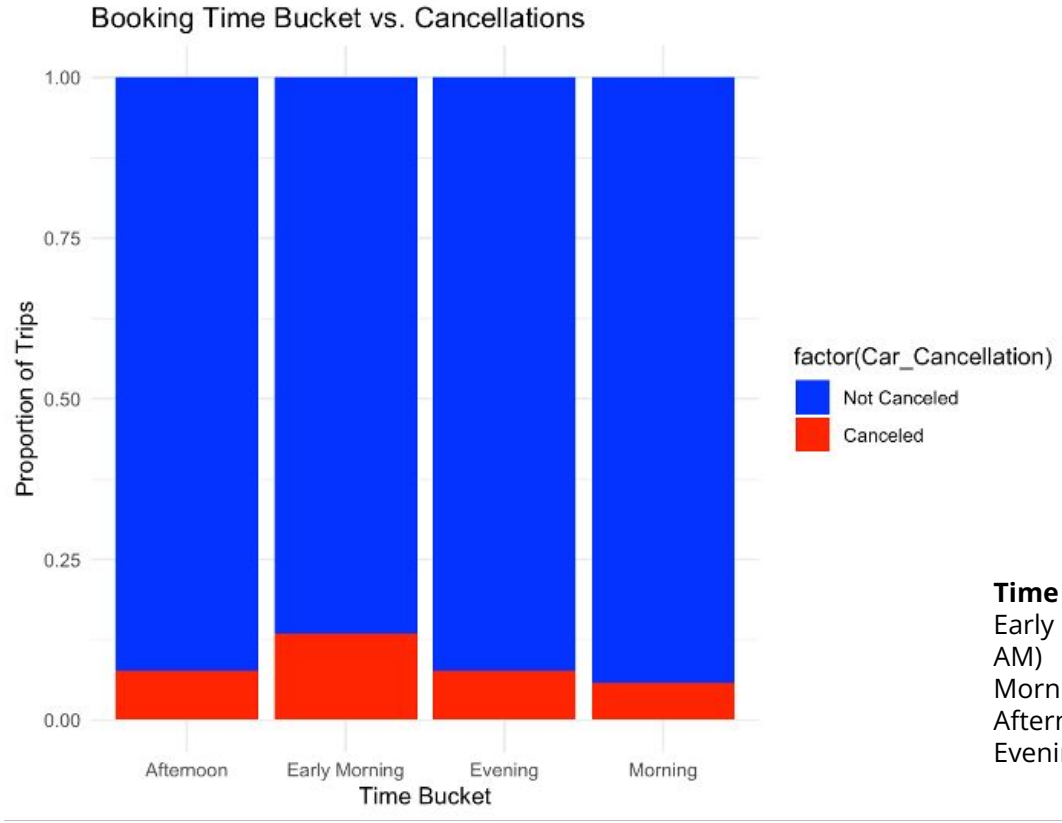
18.26%

Of the total dataset is missing

But how?

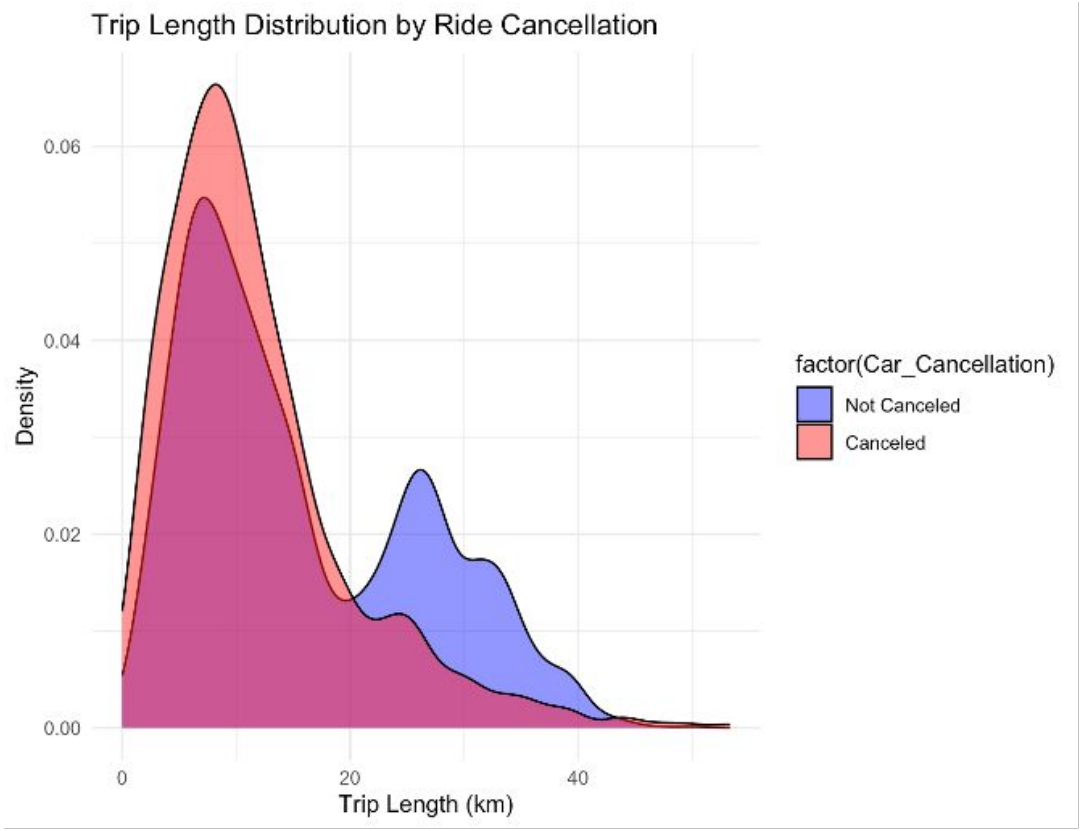

Missing Values Heatmap
row# (0%)
user_id (0%)
vehicle_model_id (0%)
package_id (82%)
travel_type_id (0%)
from_area_id (0%)
to_area_id (21%)
from_city_id (63%)
to_city_id (97%)
from_date (0%)
to_date (42%)
online_booking (0%)
mobile_site_booking (0%)
booking_created (0%)
from_lat (0%)
from_long (0%)
to_lat (21%)
to_long (21%)
Car_Canc
Observations
0
2500
5000
7500
10000
Missing
(18.3%)
Present
(81.7%)

When Do Cancellations Happen?

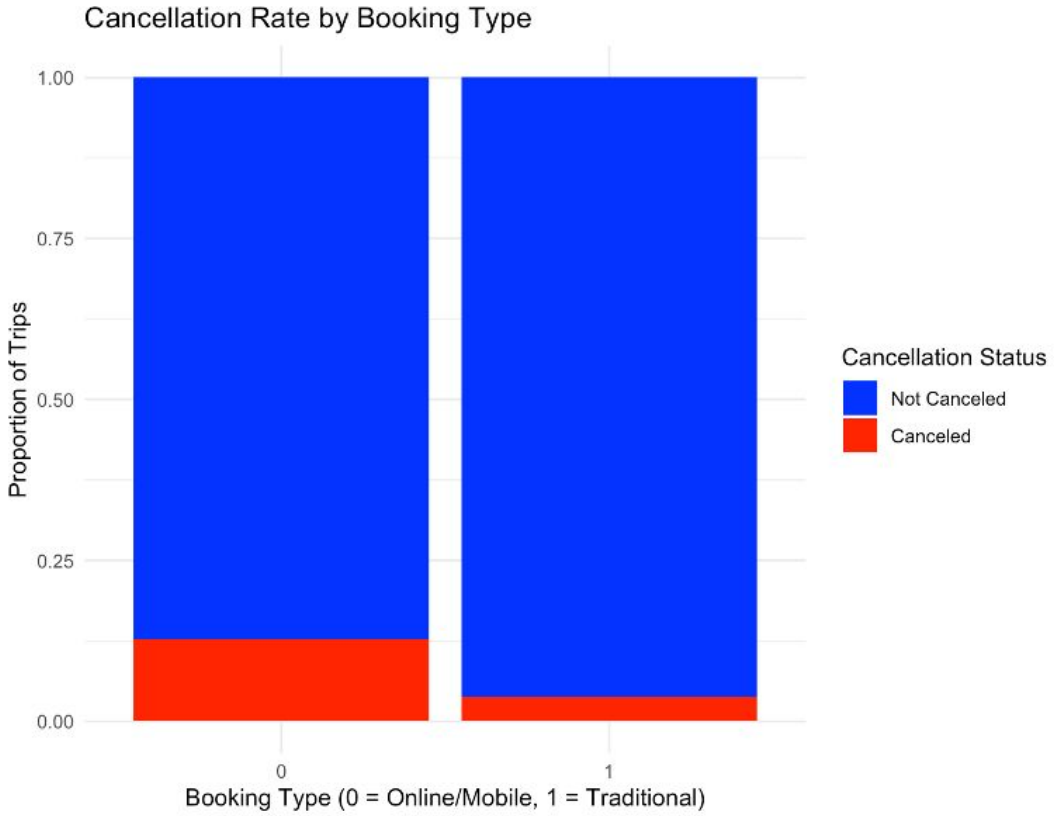


Time slots:
Early Morning (12:00 AM - 5:59 AM)
Morning (6:00 AM - 11:59 AM)
Afternoon (12:00 PM - 5:59 PM)
Evening (6:00 PM - 11:59 PM)

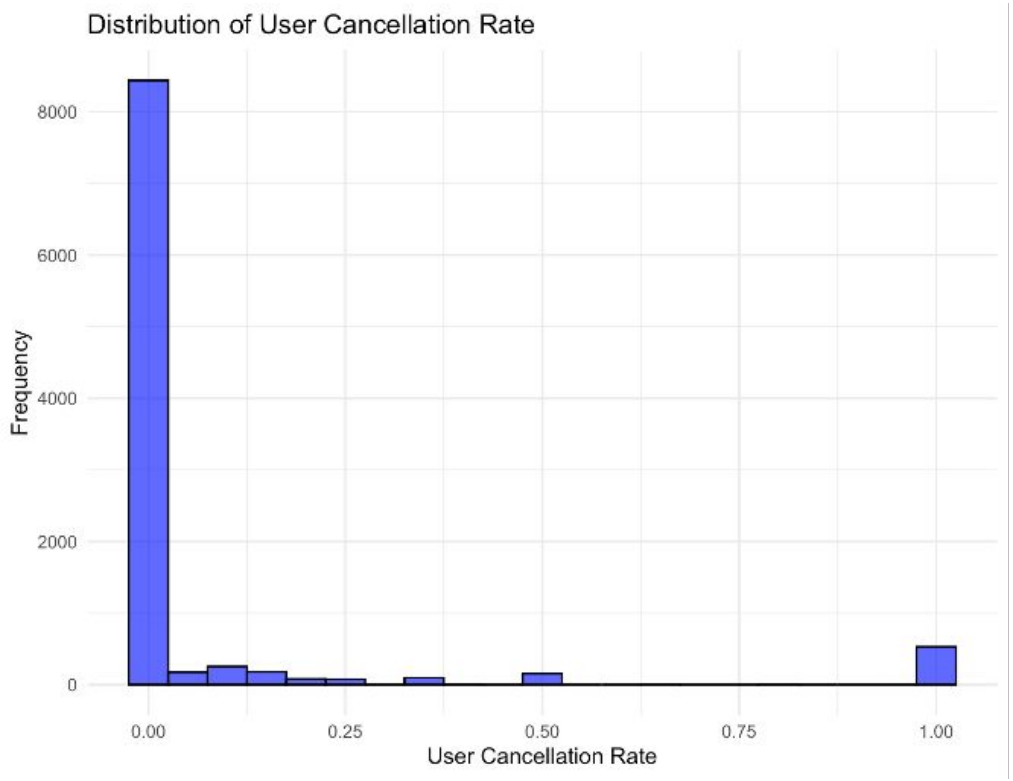
Do Longer Rides Get Canceled More?



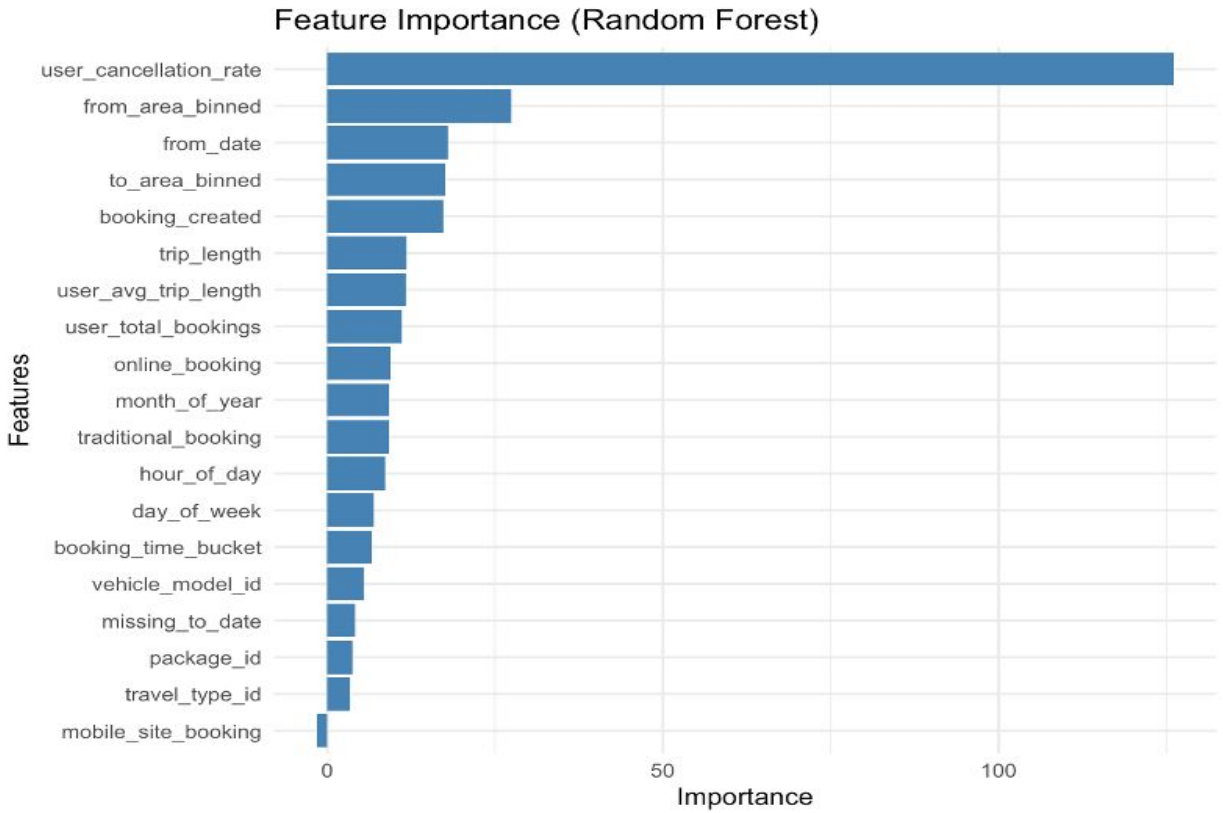
Does Booking Type Affect Cancellations?

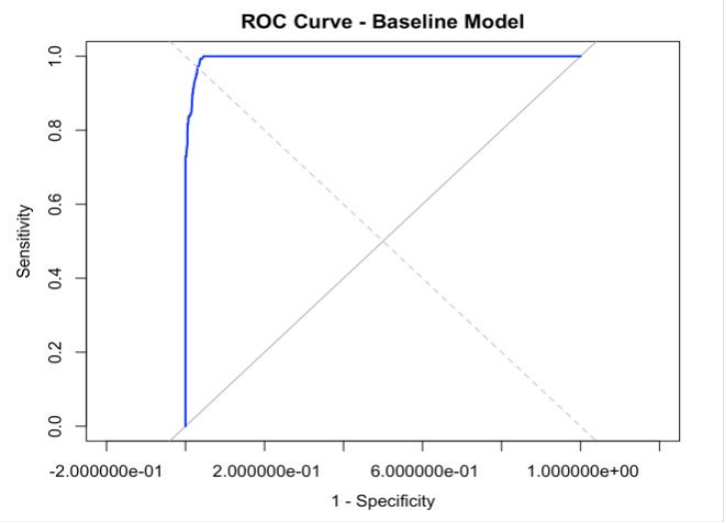


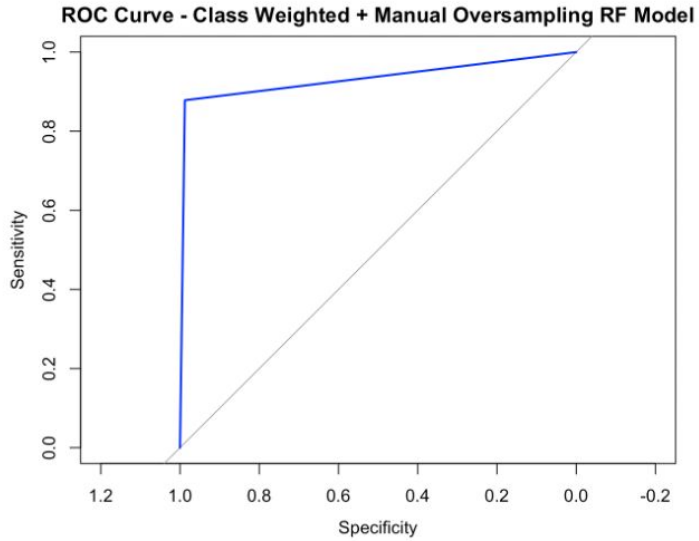
Is user cancelling the rides?

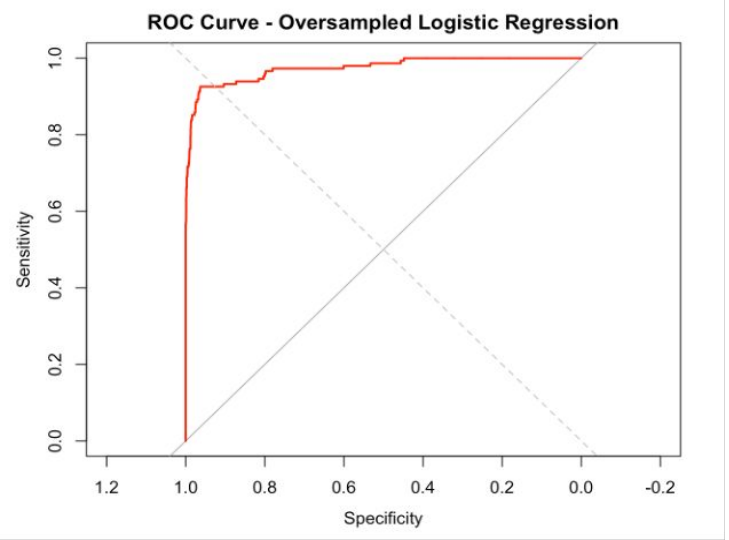


Key Factors Driving Ride Cancellations









Tested three models:
→ Baseline Random Forest – Strong performance but biased towards majority class.
→ Hybrid RF (Class Weighted + Oversampling) – Improved sensitivity for cancellations.
→ Oversampled Logistic Regression – Performed well, but slightly less robust.

- Best Model: Hybrid RF, as it balances false positives and false negatives better.

What SAR can do?

Thank you!