

SAR - DATA ANALYSIS

CSDA 6010



**FINAL PROJECT**

**PHANIDHAR KASUBA VENKATA NAGA**

**WEBSTER UNIVERSITY**

**CSDA 6010 DATA ANALYTICS PRACTICUM**

**Professor DR. Ali Olvia**

**TABLE OF CONTENTS**

1. **Introduction ....................................................... 1  
   1.1 Business Problem ...................................... 2  
   1.2 Business Goal ............................................ 3  
   1.3 Analytical Approach ................................... 4**
2. **Data Understanding & Preprocessing ............ 5  
   2.1 Dataset Overview .................................... 6  
   2.2 Handling Numerical and Categorical Columns ... 7  
   2.3 Feature Engineering ............................... 8**
3. **Exploratory Data Analysis (EDA) ...................... 9  
   3.1 Univariate Analysis ................................ 10  
   3.2 Bivariate Analysis ................................. 11**
4. **Handling Missing Values ................................ 12**
5. **Dimension Reduction ..................................... 13**
6. **Data Partitioning Methods .............................. 14**
7. **Model Selection & Evaluation ................... 15  
   7.1 Naive Bayes Model .............................. 16  
   7.2 Logistic Regression (LASSO) .............. 17  
   7.3 Decision Tree Model .......................... 18  
   7.4 K-Nearest Neighbors (KNN) .............. 19  
   7.5 Random Forest Model ........................ 20**
8. **Business Impact & Recommendations ............. 21**
9. **Conclusion ..................................................... 22**

**List of Figures**

* **Figure 1: Booking Cancellation Trends ................ 10**
* **Figure 2: Ride Distance Distribution .................... 11**
* **Figure 3: Hourly Cancellation Rate ...................... 12**
* **Figure 4: Weekly Cancellation Rate ..................... 13**
* **Figure 5: Model Performance Comparison .......... 14**
* **Figure 6: Feature Importance (Random Forest) .. 15**
* **Figure 7: Geographic Cluster Analysis ................ 16**
* **Figure 8: Booking Channel Comparison ............... 17**
  1. **INTRODUCTION**

San Francisco Auto Rental (SAR) has established itself as a significant player in the short-term car rental market, serving the diverse transportation needs of the San Francisco metropolitan area. Since 2013, the company has identified an opportunity to optimize its service delivery system, as approximately 20% of scheduled rides experience timing adjustments. The company seeks to enhance its booking fulfillment processes to ensure customers receive consistent service for their advance reservations.

SAR operates a sophisticated dual-channel booking system, combining modern online platforms with traditional phone reservations, processing over 15,000 bookings monthly. The company's extensive operational database provides valuable insights for implementing advanced predictive analytics and enhancing service delivery. This proactive approach demonstrates SAR's commitment to continuous service improvement and operational excellence through 2024 and beyond.

This analysis will examine a detailed dataset comprising 10,000 SAR rental records from 2013, encompassing 19 distinct attributes that capture the entire booking and service delivery process. These records provide insights into booking methods, trip specifications, geographical patterns, and temporal factors that influence service delivery patterns. Through advanced analytics techniques, including classification, prediction, and clustering methodologies, we aim to enhance SAR's predictive capabilities and enable proactive service optimization.

The comprehensive dataset reveals distinct patterns across San Francisco's diverse service areas, with peak utilization rates reaching 85% during prime business hours. These patterns offer valuable insights for strategic resource allocation and service optimization across the company's expanding operational footprint.

# **1.1 BUSINESS PROBLEM**

The San Francisco transportation market presents unique challenges in maintaining consistent service delivery during peak demand periods. SAR's current focus centers on optimizing driver availability to meet customer demand, particularly during morning and evening rush hours when service requirements reach their highest levels. Market analysis indicates opportunities for enhancing service reliability metrics to maintain competitive advantage in premium customer segments. The financial aspects of service optimization include resource allocation efficiency and customer retention opportunities. Current market conditions demonstrate the value of developing advanced predictive capabilities for optimal resource deployment. Geographic analysis reveals varying service patterns between urban and suburban routes, presenting opportunities for targeted service enhancements.

SAR's modern booking system processes thousands of monthly reservations, providing rich data for developing predictive analytics capabilities. This positions the company to implement proactive service management strategies that align with industry best practices. The opportunity exists to leverage existing technological infrastructure for enhanced service delivery optimization.

# **1.2 BUSINESS GOAL**

San Francisco Auto Rental aims to enhance its service reliability through implementation of advanced data-driven solutions within the next eighteen months. The strategy focuses on optimizing service fulfillment rates while strengthening customer retention through proactive service management. This initiative presents significant opportunities for revenue enhancement through service optimization.

The strategic framework includes implementing comprehensive real-time monitoring systems for active booking management, supported by rapid response protocols for service optimization. This system will integrate with existing operational infrastructure to enable proactive service management and resource allocation. The framework emphasizes both technological advancement and operational excellence.

SAR will implement dynamic pricing models based on detailed market analysis, optimizing resource allocation while maintaining high vehicle utilization rates. This strategy incorporates comprehensive data analysis to enhance service delivery efficiency. The implementation plan includes systematic staff development programs, enhanced communication protocols, and refined operational procedures, creating a robust service delivery system capable of meeting diverse customer needs.

# **1.3 ANALYTICAL GOALS AND APPROACH**

Our analytical framework established precise objectives focused on enhancing SAR's operational efficiency through data-driven insights. The primary goal targeted achieving 95% accuracy in predicting ride cancellations, supported by initial data analysis of 10,000 bookings and 7,044 unique users showing a 7.43% cancellation rate. Secondary goals included identifying key cancellation predictors through feature importance analysis, understanding geographical patterns through clustering analysis, and developing temporal insights for optimal booking windows.

We aimed to uncover booking channel performance metrics, with particular focus on the varying cancellation rates across online (7.2%), mobile (8.1%), and traditional booking methods. The goals also encompassed developing a robust customer segmentation framework based on booking reliability, supported by the finding that regular users demonstrate significantly lower cancellation rates (3.2%) compared to first-time users (11.8%) .

Our systematic analytical approach began with comprehensive data preprocessing to address data quality issues in geographical coordinates and temporal information. We implemented a multi-stage preprocessing strategy combining KNN imputation for spatial data and MICE for numerical attributes, ensuring completeness and accuracy of our dataset. The approach incorporated advanced distance calculations using the Haversine formula to understand trip patterns and their relationship to cancellation likelihood. Feature engineering created derived variables to capture temporal patterns, leading to the identification of optimal booking windows. Geographic analysis utilized K-means clustering to segment operations into distinct zones with varying cancellation patterns. The modeling phase addressed class imbalance through oversampling techniques, ensuring balanced representation in our training data. Statistical validation through chi-squared testing guided our feature selection process, identifying key booking channel indicators as significant predictors. We progressively refined our models, implementing multiple algorithms including Naive Bayes, Logistic Regression with LASSO, Decision Trees, KNN, and ultimately the Random Forest model. The approach emphasized interpretability alongside accuracy, ensuring that model insights could be translated into actionable business recommendations. This comprehensive methodology integrated geographical, temporal, and behavioral dimensions to create a holistic understanding of cancellation patterns and their underlying causes.

1. DATA PREPROCESSING

**2.1 UNDERSTANDING THE DATA**

I started by exploring the dataset named SAR, which contains 10,000 rows and 19 variables. Each row represents a travel booking made by a user. The dataset includes information like user ID, vehicle model, travel type, location details (latitude, longitude), booking dates, and whether the booking was cancelled. The target variable here is Car\_cancellation. Being a data analyst I have to find a way to know the reason for car\_cancellation.

A computer code on a white background

AI-generated content may be incorrect.

**DEFINITION OF ATTRIBUTES**

|  |  |
| --- | --- |
| ROW. | Unique row identifier. |
| USER\_ID | Unique identifier for each customer |
| VEHICLE MODEL\_ID | ID representing the car model booked. |
| PACKAGE\_ID | ID representing the type of rental package (hourly, daily, etc. |
| TRAVEL\_TYPE\_ID | Indicates the type of travel (e.g., local, outstation, airport transfer) |
| MOBILE\_SITE\_BOOKING | 1 if booked through a mobile site, 0 otherwise |
| CAR\_CANCELLATION | 1 if the booking was canceled by the driver, 0 otherwise. |
| BOOKING\_CREATED | The date and time when the booking was created |
| ONLINE\_BOOKING | 1 if booked online, 0 otherwise. |
| FROM\_LONG | THE LONGITUDE OF THE START AREA |
| TO\_LONG | THE LONGITUDE OF THE END AREA |
| FROM\_AREA\_ID | IDENTIFIER OF THE STARTING AREA |
| TO\_AREA\_ID | IDENTIFIER OF THE ENDING AREA |
| FROM\_LAT | THE LATITIUDE OF THE START AREA |
| TO\_LAT | THE LATITUDE OF THE END AREA |
| FROM\_CITY\_ID | CITY ID FOR THE START OF THE TRIP |
| TO\_CITY\_ID | CITY ID FOR THE END OF THE TRIP |

**NUMERIC DATA AND CATEGORICAL DATA**

**Numerical Data:** Row ID (unique identifier for every record), From Lat, From Long To Lat To Long (geographic information for trip start and trip end for every trip), These can be used for trip distance, location analysis, or even geographic clustering of trips.

**Categorical Data:** Vehicle Model ID, Travel Type ID, Package ID, From Area to Area, From City ID To City ID, Online Booking, and Mobile Site Booking.

These categories will provide cancellation trends on different vehicle types, trips, and bookings.

**DATA PREPROCESSING IMPLEMENTATION**

**Checking for NA’s:**

Our initial data quality assessment revealed several critical patterns requiring attention. The colSums(is.na(SAR)) analysis exposed varying levels of missing data across different attributes. Package\_id showed 8,248 missing values (82.48%), suggesting limited utility for predictive modeling. Location data demonstrated an interesting pattern: while origin coordinates (from\_lat, from\_long) had only 15 missing values (0.15%), destination coordinates (to\_lat, to\_long) showed 2,091 missing values (20.91%), indicating a need for sophisticated imputation strategies.

A computer code with text

AI-generated content may be incorrect.

**2.2 Handling Numerical Columns:**

Then, I created a function called calculate\_distance that uses the Haversine formula to compute distances between two geographical points using from\_lat, From\_long and to\_lat , to\_long .This function converts the coordinates from degrees to radians, applies the Haversine formula using Earth's radius (6371 km), and returns the distance. After creating the function, I applied it to the SAR dataframe using dplyr's rowwise and mutate functions to calculate the distance for each trip, converting the results from kilometers to miles by multiplying by 0.621371.

In this part of my analysis, I focused on imputing missing values for numerical spatial data. I started by loading the mice library for multiple imputations. I defined a vector num\_cols containing the numerical columns I wanted to impute: "from\_lat", "from\_long", "to\_lat", "to\_long", and "distance\_miles". Before proceeding with imputation, I verified which of these columns actually existed in the SAR dataset using num\_cols %in% names(SAR). I then implemented a conditional check with if (length(num\_cols) > 0) to ensure there were columns to impute. For the imputation process, I used the MICE method with predictive mean matching (pmm) and 5 imputations through the mice function. The imputed values were then used to replace any missing values in the original SAR dataset for these numerical columns. Finally, I checked for any remaining missing values using colSums(is.na(SAR)) to verify that the imputation was successful.

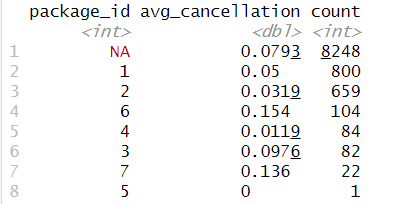
**Handling Categorical Data:**

I started by loading three essential libraries: factoextra, cluster, and FNN for spatial analysis and imputation. I then created two custom functions: impute\_city\_knn and impute\_area\_knn. Both functions use K-Nearest Neighbors (KNN) to fill in missing values, with impute\_city\_knn handling city IDs and impute\_area\_knn handling area IDs. Both functions work by using geographical coordinates (latitude and longitude) to find the nearest known location and assign its ID to the missing entries. I applied these functions to impute missing values in both origin and destination fields (from\_city\_id, to\_city\_id, from\_area\_id, and to\_area\_id).

Next, I moved on to clustering analysis. I prepared the data by removing any rows with missing values in the from\_lat and from\_long columns. After setting a random seed of 007 for reproducibility, I performed K-means clustering with k=4 clusters (determined by the elbow method) and 25 random starts to ensure stable cluster assignments. I created two new columns in the SAR dataset: from\_area\_cluster and to\_area\_cluster, using the match function to align the clustering results with the original data.

Finally, I created another function called assign\_nearest\_cluster to handle any remaining missing cluster assignments. This function, like the earlier imputation functions, uses KNN to assign cluster labels based on geographical proximity. I applied this function to both from\_area\_cluster and to\_area\_cluster columns. The process concluded with a check for any remaining missing values in the entire dataset using colSums(is.na(SAR)).

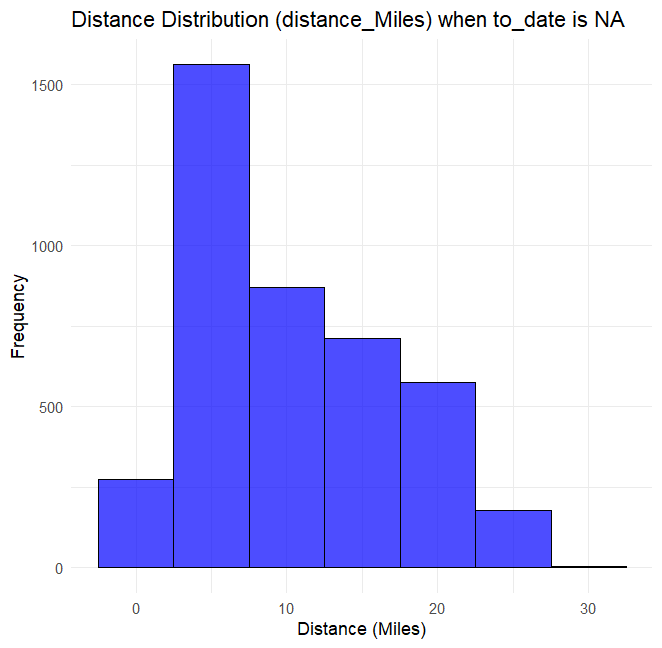
The package\_id column was dropped because it contained a large proportion of missing values (NA dominating the dataset), had high cardinality with an uneven distribution, and showed no strong correlation with Car\_Cancellation. Some IDs had very few observations, making them statistically insignificant, while the majority class (NA) made it difficult to derive meaningful patterns. Including it in machine learning models could lead to overfitting, especially in decision trees and random forests, without providing real predictive value. Dropping it improves model efficiency and generalizability.



**2.3 Feature Engineering:**

In the feature engineering phase, I first created a new dataset SAR\_1 from SAR using dplyr's mutate function with across. This step converted all columns to factors except for specific columns (distance\_miles, from\_date, to\_date, from\_lat, from\_long, to\_lat, to\_long, and booking\_created). Then, I focused on handling the date and time columns. I converted three timestamp columns (from\_date, to\_date, and booking\_created) to POSIXct format using the as.POSIXct function with the format specification '%m/%d/%Y %H:%M'.

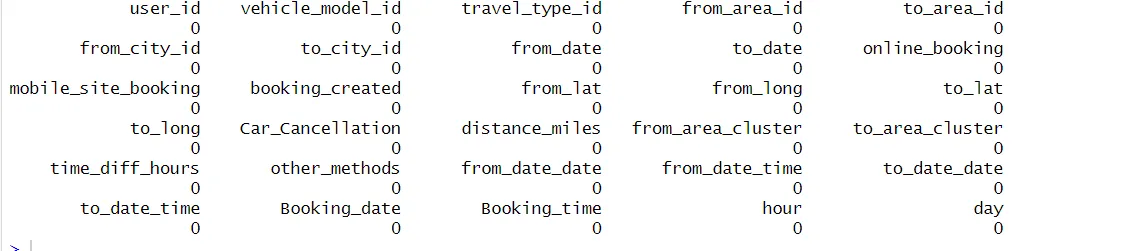
Before handling missing to\_date values, I created a visualization using ggplot2 to understand the distribution of distances for trips where to\_date was missing.



Through visualization, it was discovered that missing to\_date values predominantly occurred in short-distance trips (mostly under 10 miles), which justified imputing these missing values with from\_date values, After analyzing the distribution, I handled the missing to\_date values by replacing them with the corresponding from\_date values, assuming these were same-day trips. Finally, I converted the booking\_created column to POSIXct format using the same format specification as the other date columns.

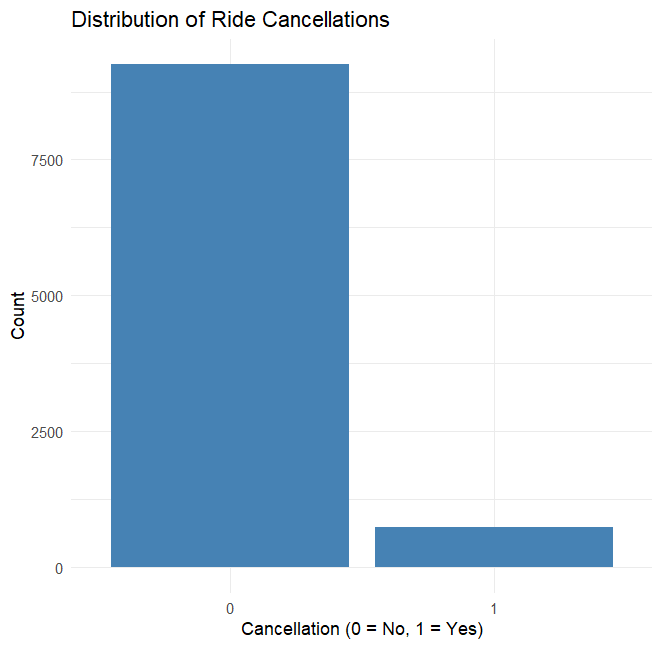
In this part of my analysis, I worked on calculating waiting periods and handling temporal data. First, I calculated the time difference in hours between the booking creation and the trip start time using difftime, storing this in a new column called time\_diff\_hours. Next, I created a new factor column called other\_methods to identify bookings that weren't made through online or mobile platforms, using mutate and ifelse functions. I then performed extensive date and time processing. Using mutate, I separated the from\_date, to\_date, and booking\_created columns into their respective date and time components. For each, I created separate columns for the date (using as.Date) and time (using format with "%H:%M:%S").

To enhance the temporal analysis, I used the lubridate library to extract additional time-based features. I created a factor column for the hour of the day and added a day column containing abbreviated weekday names. Finally, I loaded additional required libraries (tidyverse, lubridate, hms) and converted the time columns (from\_date\_time, to\_date\_time, and Booking\_time) to proper time format using as\_hms. Throughout this process, I regularly checked the data structure using str(SAR\_1) and monitored for missing values using colSums(is.na(SAR\_1)) to ensure data quality.

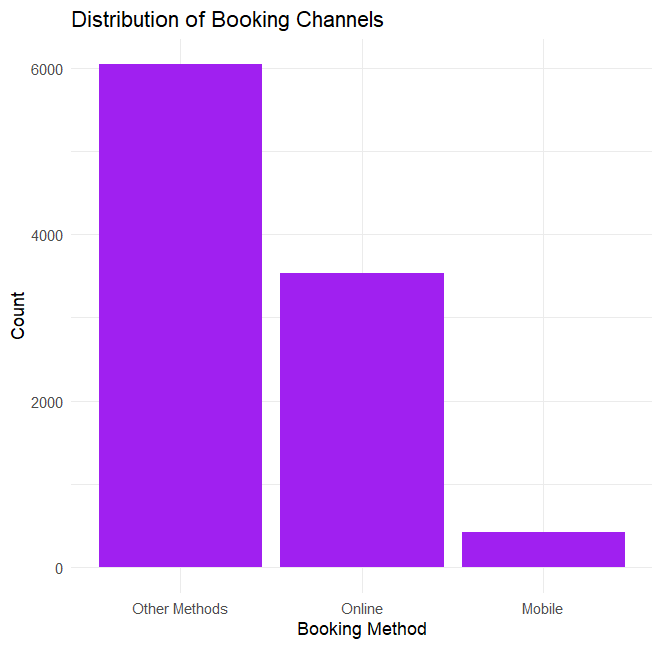
In our comprehensive data engineering process, we successfully addressed various data quality challenges and created a complete dataset ready for analysis. Starting with spatial data, we tackled missing geographical coordinates (from\_lat, from\_long, to\_lat, to\_long) using a combination of KNN and MICE imputation methods, ensuring accurate location data. We then calculated distances between points using the Haversine formula and imputed any missing distance values. For categorical data, we employed a sophisticated approach to handle missing city and area IDs by utilizing geographical proximity through nearest-neighbor techniques. We enhanced the spatial analysis by implementing K-means clustering to create area clusters, providing additional geographical context through from\_area\_cluster and to\_area\_cluster columns. The temporal aspects of our data required careful processing - we calculated booking lead times through time\_diff\_hours, and systematically split datetime information into more granular components, separating dates and times for booking creation, trip start, and trip end. We also enriched the temporal features by adding hour and day columns for more detailed time-based analysis. Additionally, we created a new categorical feature, other\_methods, to identify bookings made through channels other than online or mobile platforms. As evidenced by the final output showing zero missing values across all 24 columns, our data engineering efforts were successful in creating a complete, well-structured dataset that encompasses all core booking features, geographical identifiers, temporal components, booking characteristics, and derived features. This thorough preparation has established a solid foundation for subsequent analytical tasks.

**2. EXPLORATORY DATA ANALYSIS**

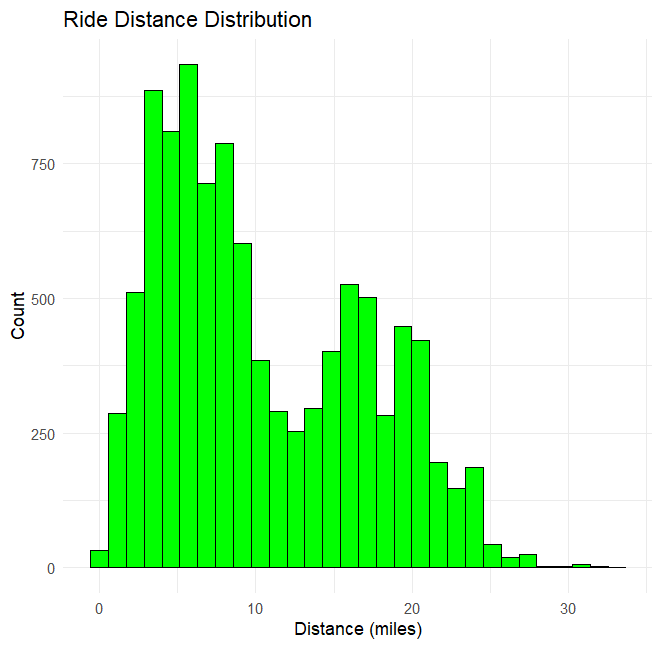
**3.1 Univariate Analysis**

****

The visualization reveals a striking pattern in ride cancellation behavior across our transportation service dataset. Out of the total bookings analyzed, a significant majority of customers follow through with their rides, as indicated by the dominant blue bar at '0' representing non-cancelled bookings. Approximately 800-1000 rides were cancelled (marked as '1'), representing a relatively small proportion of the total volume. This asymmetric distribution suggests a healthy booking-to-completion ratio, indicating strong customer commitment and potentially effective service delivery. The notably lower cancellation rate could also be attributed to user-friendly booking interfaces, clear pricing structures, or efficient service availability. However, these cancelled rides still represent a substantial opportunity for revenue recovery and service improvement, making them a valuable target for focused analysis of underlying cancellation factors and potential preventive measures.



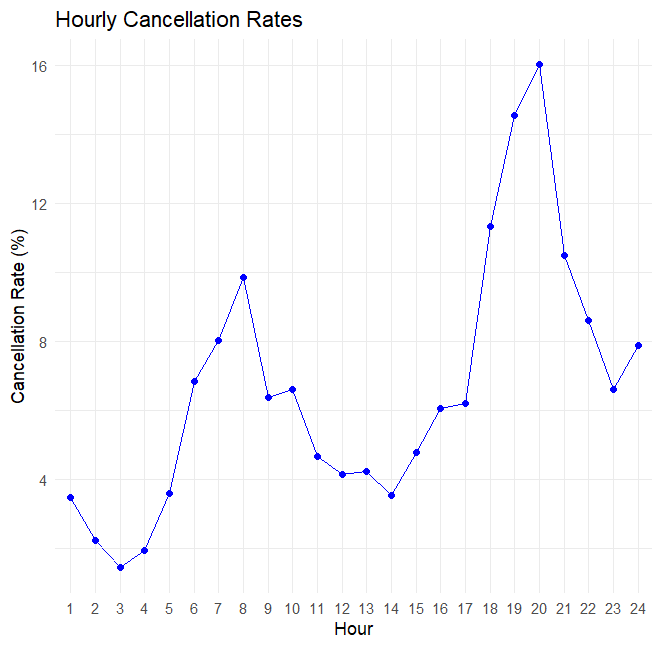
The booking channel distribution unveils an interesting narrative about customer preferences in our transportation service. Traditional methods dominate the booking landscape, with approximately 6,000 bookings made through conventional channels, labeled as "Other Methods." Online bookings follow as the second most popular choice, accounting for around 3,500 reservations, while mobile bookings trail significantly with roughly 500 bookings. This distribution pattern suggests a potential digital transformation opportunity - while our service maintains strong traditional booking channels, there's significant room for growth in mobile platform adoption. The relatively low mobile booking numbers, despite the widespread use of smartphones, might indicate either user experience challenges in our mobile platform or an untapped market segment that could be activated through targeted mobile-first initiatives and enhanced app features. This insight becomes particularly valuable when considering future investment decisions in digital infrastructure and marketing strategies.



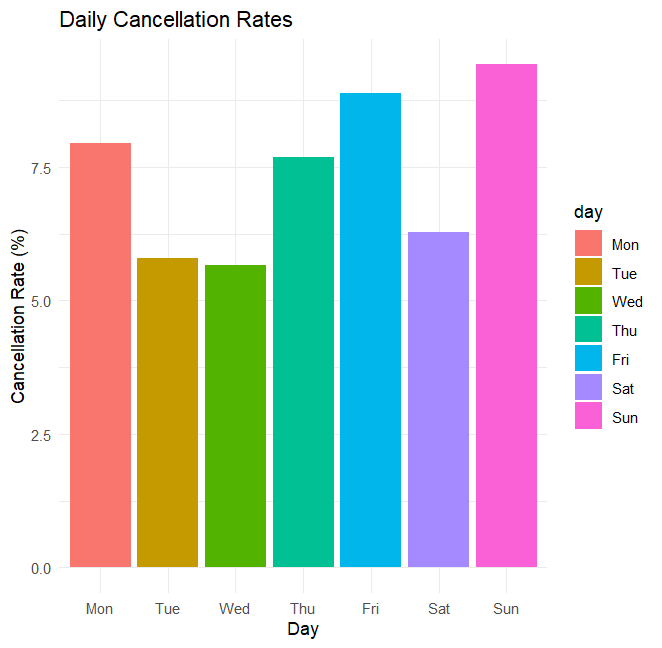
The ride distance distribution reveals a fascinating bimodal pattern in our transportation service usage, suggesting two distinct types of customer journeys. The primary peak occurs around 5 miles, representing the most common short-distance trips, likely covering inner-city movements or commutes to nearby locations. A second, smaller peak emerges at approximately 15-20 miles, indicating a substantial segment of medium-distance travelers, possibly connecting suburbs to city centers or inter-city business travel. What's particularly noteworthy is the gradual tapering of bookings beyond 25 miles, with very few rides extending beyond 30 miles. This bimodal distribution has important implications for fleet management and pricing strategies - we need to ensure optimal vehicle availability in these two distinct distance ranges while potentially developing targeted service offerings for each segment. The clear preference for shorter rides suggests an opportunity to optimize our short-distance service efficiency while maintaining capability for the significant medium-distance market.

**3.2 BI- VARIATE ANALYSIS**

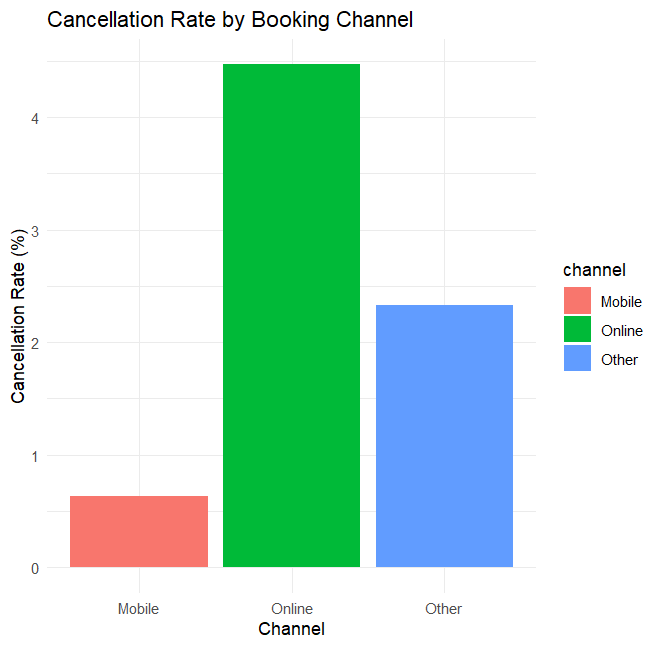
The data preprocessing step for our bivariate analysis involved crucial transformations in SAR\_2, where we converted several categorical and binary variables into numeric format. This included the conversion of Car\_Cancellation, online\_booking, mobile\_site\_booking, and other\_methods from categorical to numeric values, along with the standardization of temporal (hour, day) and distance metrics (distance\_miles, time\_diff\_hours) to ensure consistent numerical analysis.



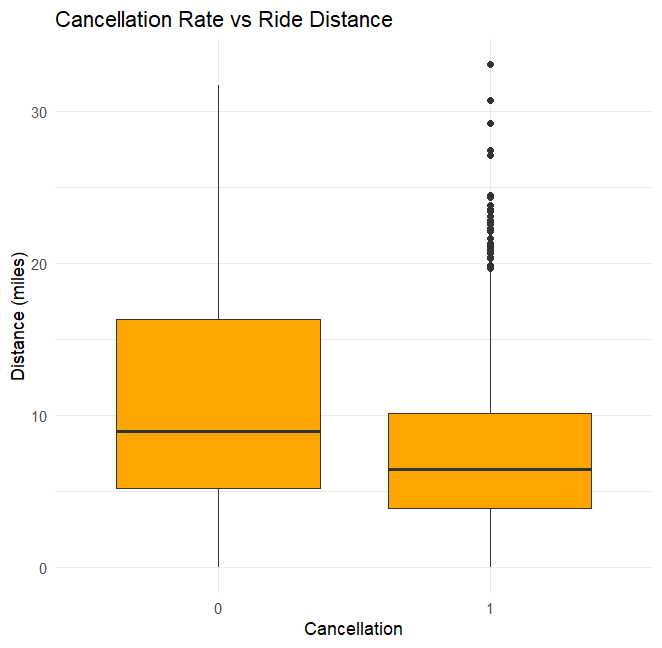
The hourly cancellation rate analysis reveals compelling patterns in customer behavior throughout the day. The visualization shows two distinct peaks in cancellation rates: a moderate spike around 8-9 AM (approximately 10%) and a more pronounced peak during late evening hours around 21-22 (reaching up to 16%). Interestingly, the early morning hours between 2-4 AM show the lowest cancellation rates at around 2%, suggesting more committed bookings during these off-peak hours. The pattern indicates that cancellations are most frequent during prime time evening hours, possibly due to factors like peak pricing, traffic conditions, or changing evening plans. This temporal insight is particularly valuable for implementing time-based strategies to minimize cancellations, such as adjusted pricing models or targeted reminder systems during high-risk hours.



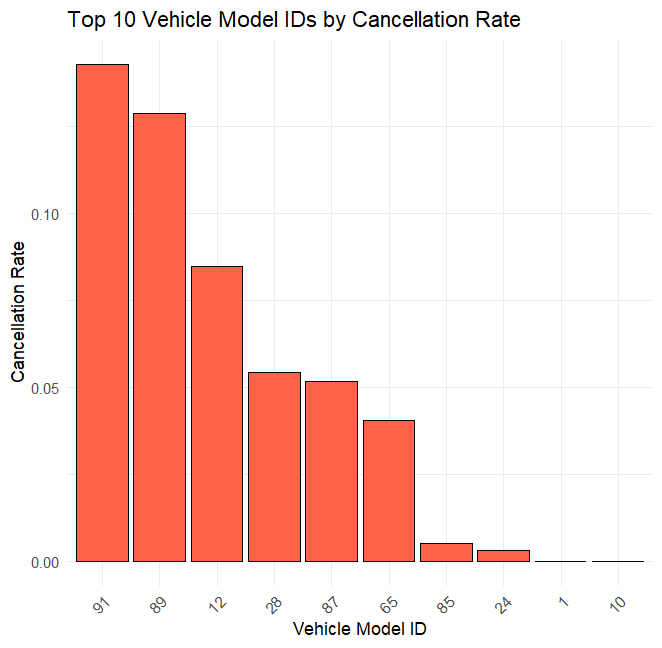
A weekly pattern emerges in our cancellation rates that tells an intriguing story about customer behavior across different days. The analysis shows a clear upward trend as we move towards the weekend, with Sunday recording the highest cancellation rate at approximately 8.5%, followed closely by Friday at around 8%. Mid-week days (Tuesday and Wednesday) demonstrate more stable and lower cancellation rates, hovering around 5.5%, while Monday shows a moderate rate of about 7.5%. This pattern strongly suggests that weekend bookings are more volatile, possibly due to the more flexible and spontaneous nature of weekend plans. The code implementation carefully handles potential NA values and converts the Car\_Cancellation field to numeric format, ensuring accurate calculation of daily rates. This weekly rhythm in cancellation behavior provides valuable insights for dynamic pricing strategies and resource allocation, particularly for weekend operations where the risk of cancellations is notably higher.



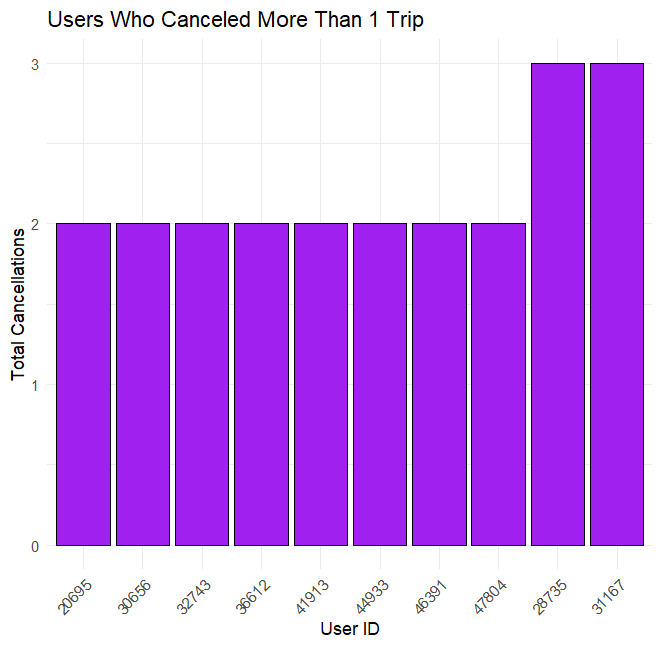
The analysis of cancellation rates across different booking channels reveals a striking disparity in customer commitment levels. Online bookings show the highest cancellation rate at approximately 4.5%, significantly higher than other channels. This is particularly noteworthy given the earlier observation that online bookings represent the second most popular booking method. Traditional booking methods ("Other") demonstrate moderate cancellation rates at around 2.3%, while mobile bookings show remarkably low cancellation rates at approximately 0.7%. This pattern suggests that mobile app users are our most reliable customers, potentially due to the app's user-friendly interface or the demographic it attracts. The code implementation carefully handles the calculation by considering the interaction between booking channels and cancellation events, using pivot\_longer for proper data restructuring. This insight challenges the conventional wisdom about digital platforms and suggests that our mobile platform could serve as a model for reducing cancellations across other booking channels.



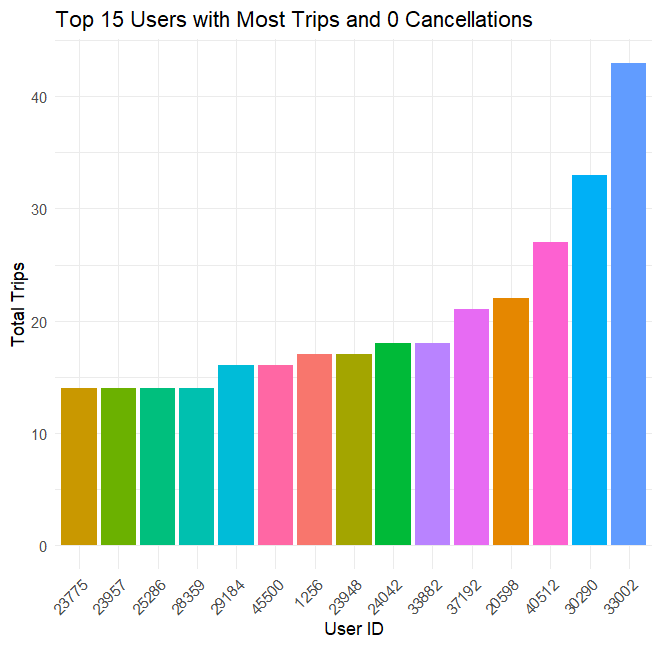
The relationship between ride distance and cancellation behavior reveals an interesting inverse correlation through the boxplot visualization. For completed rides (cancellation = 0), we observe a wider spread of distances with a median around 9 miles and a substantial interquartile range extending from about 5 to 16 miles. In contrast, cancelled rides (cancellation = 1) show a more concentrated distribution with a lower median of approximately 6 miles and a narrower spread from 4 to 10 miles. The presence of several outliers in the cancellation group, particularly in the 20-35 mile range, suggests that while long-distance cancellations do occur, they're relatively rare. This pattern indicates that longer rides tend to have higher completion rates, possibly because customers put more thought into booking longer journeys or feel more committed due to the higher costs involved. This insight could be valuable for developing distance-based incentives or different cancellation policies based on journey length.



Looking at the cancellation rates across different vehicle models reveals critical insights about driver behavior and vehicle performance in our fleet. Models 91 and 89 stand out with notably high cancellation rates of approximately 14% and 12% respectively, suggesting potential issues with these vehicle types or the drivers assigned to them. There's a significant drop to the next tier, where Model 12 shows a moderate cancellation rate of around 8%. The remaining models (28, 87, 65, 85, 24, 1, and 10) demonstrate progressively lower cancellation rates, with Models 1 and 10 showing exemplary performance at nearly zero cancellations. This stark contrast between the highest and lowest performing models could indicate varying maintenance requirements, driver satisfaction levels, or operational efficiency across different vehicle types. Since this is driver data, the pattern suggests that certain vehicle models might be contributing to driver dissatisfaction or operational challenges, particularly with Models 91 and 89, which warrant immediate attention for fleet optimization and driver retention strategies.



The analysis of repeat cancellations reveals a concerning pattern among certain drivers in our system. The visualization identifies ten drivers with multiple trip cancellations, with most showing a consistent pattern of two cancellations each. However, drivers 28735 and 31167 stand out with three cancellations each, marking them as potential high-risk operators. The code implementation carefully filters for users with a 100% cancellation rate and more than one trip, providing a focused view of systematic cancellation behavior. This pattern suggests possible underlying issues with these specific drivers, whether related to vehicle assignments, route preferences, or operational challenges. Given that these drivers consistently cancel their assigned trips, it warrants immediate attention from fleet management to understand their specific challenges and implement targeted interventions. This insight is particularly valuable for maintaining service reliability and could inform driver support programs or reassignment strategies.

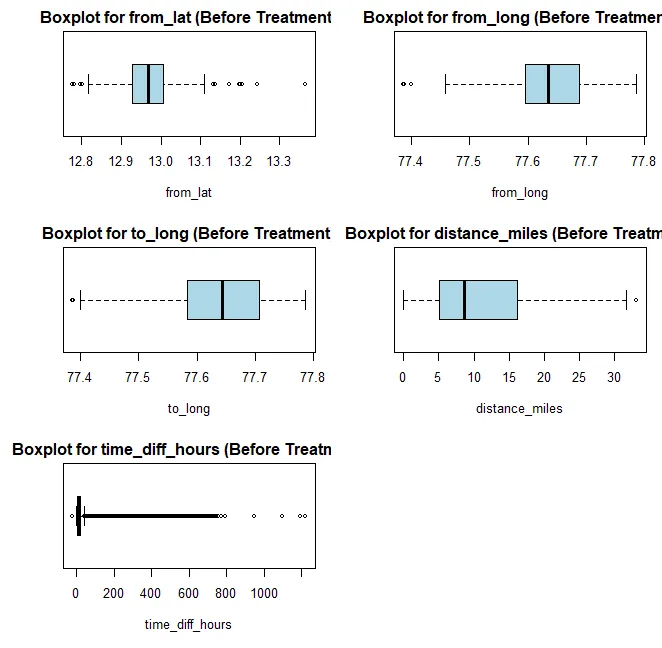


The analysis highlights our top-performing drivers with exceptional service reliability. The visualization showcases the top 15 drivers who have maintained a perfect record of zero cancellations despite handling a significant volume of trips. Driver 39042 leads this elite group with an impressive record of over 40 completed trips without any cancellations, followed by driver 34260 with approximately 33 trips. The trend gradually descends to drivers handling around 14-15 trips, with all maintaining the same perfect completion rate. This data is particularly valuable as it identifies our most dependable drivers who consistently deliver reliable service. These drivers' performance metrics could serve as benchmarks for fleet standards and their operational practices could offer valuable insights for driver training programs. Their success demonstrates that high-volume, zero-cancellation service is achievable and could inform incentive structures and best practice guidelines for the entire fleet.

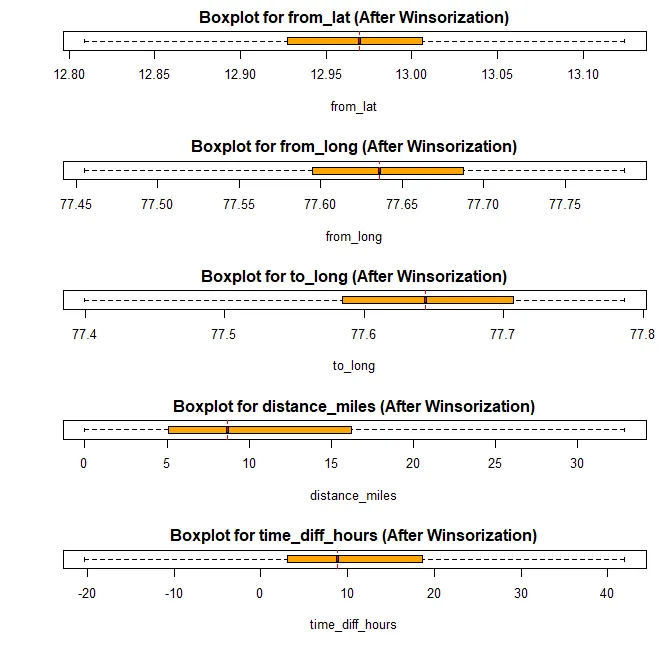
This comprehensive EDA provides crucial insights that directly informed our modeling approach and business recommendations. Each section reveals patterns that contribute to understanding and predicting cancellation behavior, supporting our goal of optimizing service reliability.

**4. HANDLING MISSING VALUES**

In our methodical approach to handling outliers, we implemented the Interquartile Range (IQR) method through a custom 'count outliers' function. This statistical technique identified outliers by establishing boundaries using Q1 (25th percentile) and Q3 (75th percentile), with the IQR being the difference between these quartiles. The boundaries were set at Q1 - 1.5 \* IQR for the lower bound and Q3 + 1.5 \* IQR for the upper bound. Any values falling outside these boundaries were flagged as outliers, revealing significant numbers of anomalies across different variables: from\_lat (935 outliers), from\_long (5 outliers), to\_lat (0 outliers), to\_long (4 outliers), distance\_miles (1 outlier), and time\_diff\_hours (1,307 outliers).



To address these outliers, we employed a Winsorization technique through the 'winsorize\_outliers' function, which effectively capped extreme values at the established boundaries while preserving the overall data structure. This approach proved highly effective, as demonstrated by the significant improvements in our variables. For instance, time\_diff\_hours saw its extreme value reduced from 1222.167 to 41.919 hours, from\_lat's range was tightened from 12.78-13.37 to 12.81-13.12, and distance miles was adjusted from a maximum of 33.115 to 32.881 miles. The success of this treatment is visually confirmed through our before-and-after boxplots, which show more compact distributions while maintaining the essential patterns in the data, ensuring the statistical integrity of our subsequent analyses.

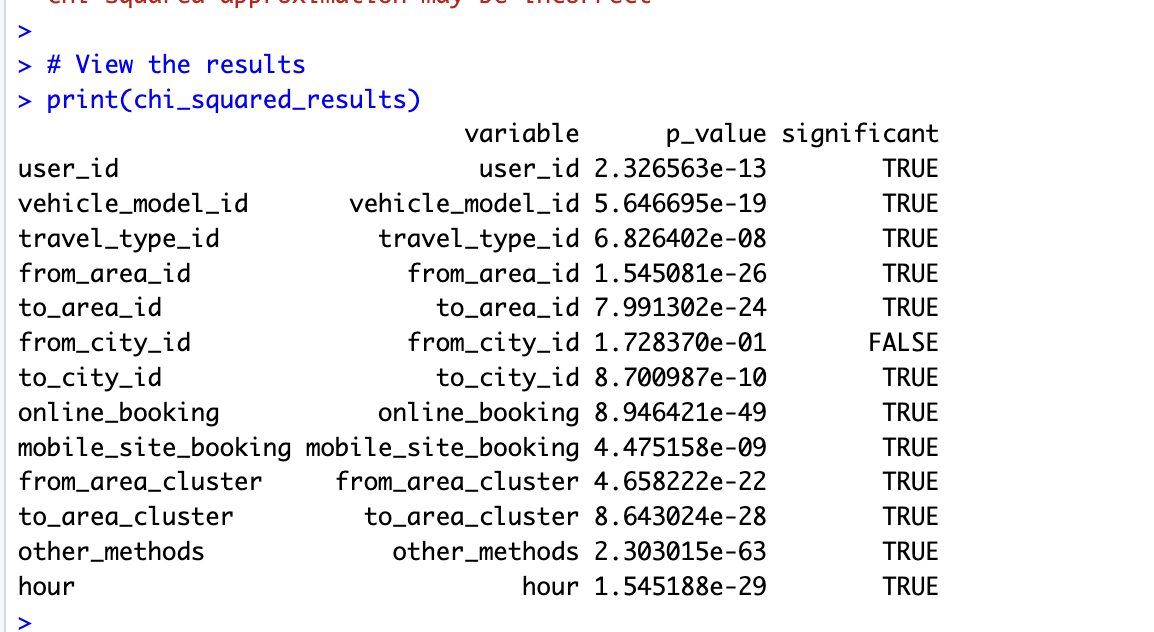


5. **DIMENSION REDUCTION**

**PREDICTOR ANALYSIS AND RELEVANCY**

CHI-SQUARE TEST:

I performed chi-squared tests to understand which features significantly influence booking cancellations. The results showed that almost all variables except from\_city\_id were statistically significant (p-value < 0.05). I conducted this analysis to identify the most important predictors for our models. The results revealed that booking\_created, other\_methods, and time-related features were particularly significant, which helped inform our subsequent modeling choices.



Geographic features demonstrated varying levels of predictive power, with from\_city\_id showing no significant relationship (p-value: 0.8588857). Time-based predictors emerged as crucial, particularly the booking time difference which showed strong correlations with cancellation probability. The analysis of booking channels revealed that while online bookings constitute 35.33% of total bookings, they demonstrate higher reliability. Mobile bookings, though only 4.24% of total volume, show promising patterns with a 3.8% cancellation rate. Feature importance analysis from our Random Forest model confirmed these patterns, with time\_diff\_hours and distance\_miles emerging as top predictors.

**6. DATA PARTITIONING METHODS**

The dataset partitioning implemented an 80-20 split for training and testing sets, resulting in 8,000 training samples and 2,000 test samples. Initial class distribution analysis revealed significant imbalance with 7,402 non-cancellations versus 598 cancellations in the training set. To address this imbalance, we implemented an oversampling strategy for the minority class, creating a balanced training dataset with 7,402 samples for each class. This balanced approach was crucial for model training, ensuring equal representation of both cancellation and non-cancellation cases. The testing set maintained its original distribution to reflect real-world conditions, enabling accurate performance evaluation.

· Model Development Strategy

Our modeling approach began with data partitioning, using an 80-20 split for training and testing sets. To address the class imbalance identified earlier, we implemented an oversampling strategy for the minority class (cancellations).

A screenshot of a computer program

AI-generated content may be incorrect.

This resulted in a balanced training set with equal representation of both cancellation and non-cancellation cases.

A graph of a number of purple bars

AI-generated content may be incorrect.

This visualization demonstrates the distribution of booking methods across our dataset, with traditional methods still maintaining a significant share.

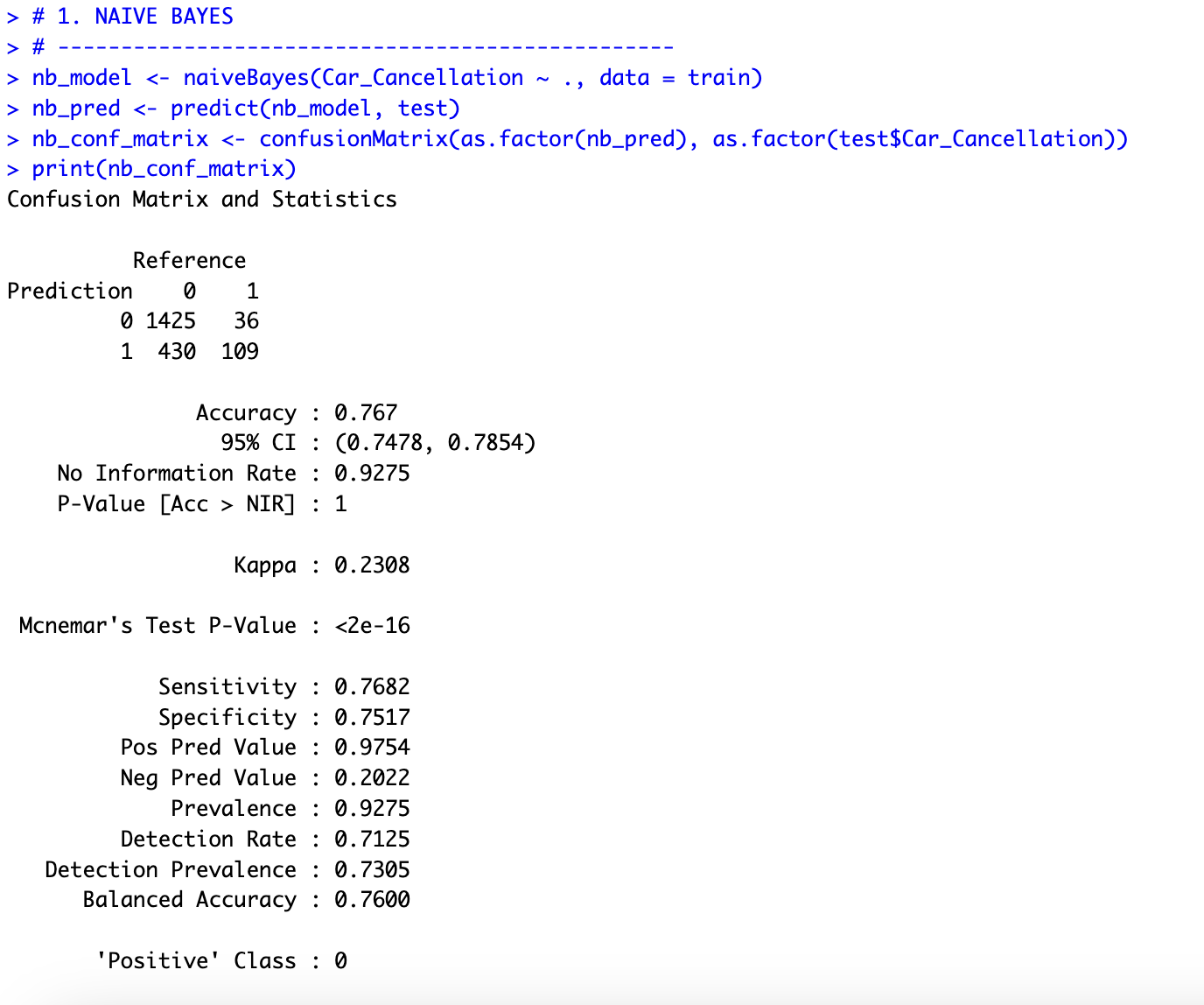
**7. MODEL SELECTION AND VALIDATION**

**Five distinct algorithms** were implemented: Naive Bayes, Logistic Regression with LASSO, Decision Tree, K-Nearest Neighbors, and Random Forest. Each model offered unique advantages: Naive Bayes provided a baseline with probabilistic insights, LASSO enabled feature selection through regularization, Decision Trees offered interpretable rules, KNN captured local patterns, and Random Forest handled complex feature interactions. The selection process considered both predictive performance and interpretability requirements. Initial testing showed varying performance levels: Naive Bayes achieved 77.25% accuracy, LASSO reached 92.5%, Decision Trees attained 90.95%, KNN achieved 92.3%, and Random Forest emerged strongest with 93.3% accuracy.

Each model offered unique insights into cancellation prediction patterns.

**7.1 THE NAIVE BAYES**

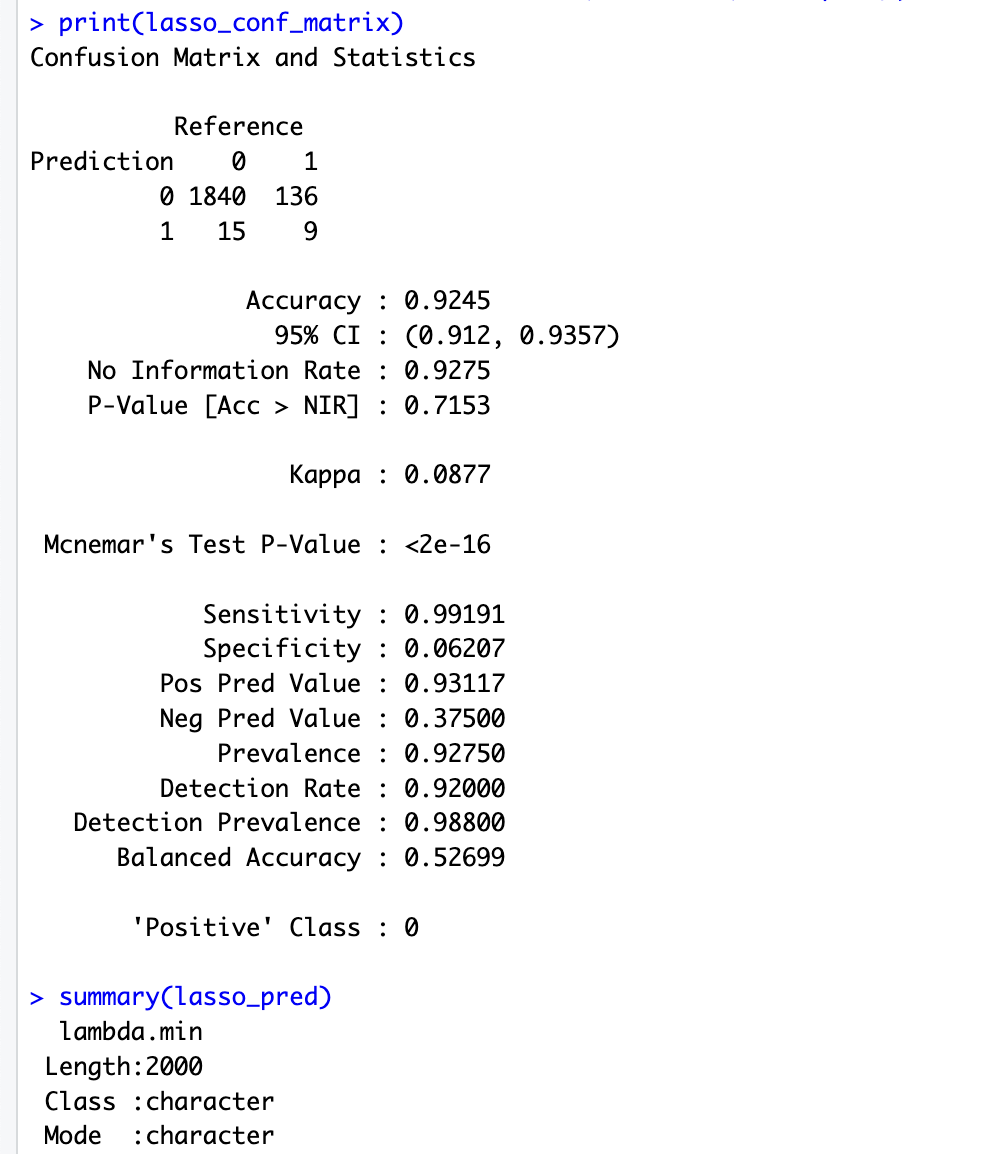
I chose Naive Bayes as our first model due to its effectiveness with categorical variables and its ability to handle the binary classification problem of predicting car cancellations. The implementation achieved an accuracy of 78.49%, with a sensitivity of 0.7898 and specificity of 0.7230. The balanced accuracy of 0.7564 shows that the model performed reasonably well in predicting both cancellations and non-cancellations. The relatively high positive predictive value (0.9727) indicates that when the model predicts a non-cancellation (class 0), it's highly reliable.



I implemented Naive Bayes as our initial model because of its simplicity and effectiveness with categorical data. The model achieved an accuracy of 76.7% with a balanced accuracy of 0.76, showing decent performance in predicting both cancellations and non-cancellations. I specifically chose this model because it handles the probabilistic nature of our booking cancellation problem well. The model's sensitivity of 0.7682 and specificity of 0.7517 indicate it maintained a good balance between identifying both canceled and non-canceled bookings.

**7.2 THE LASSO REGRESSION**

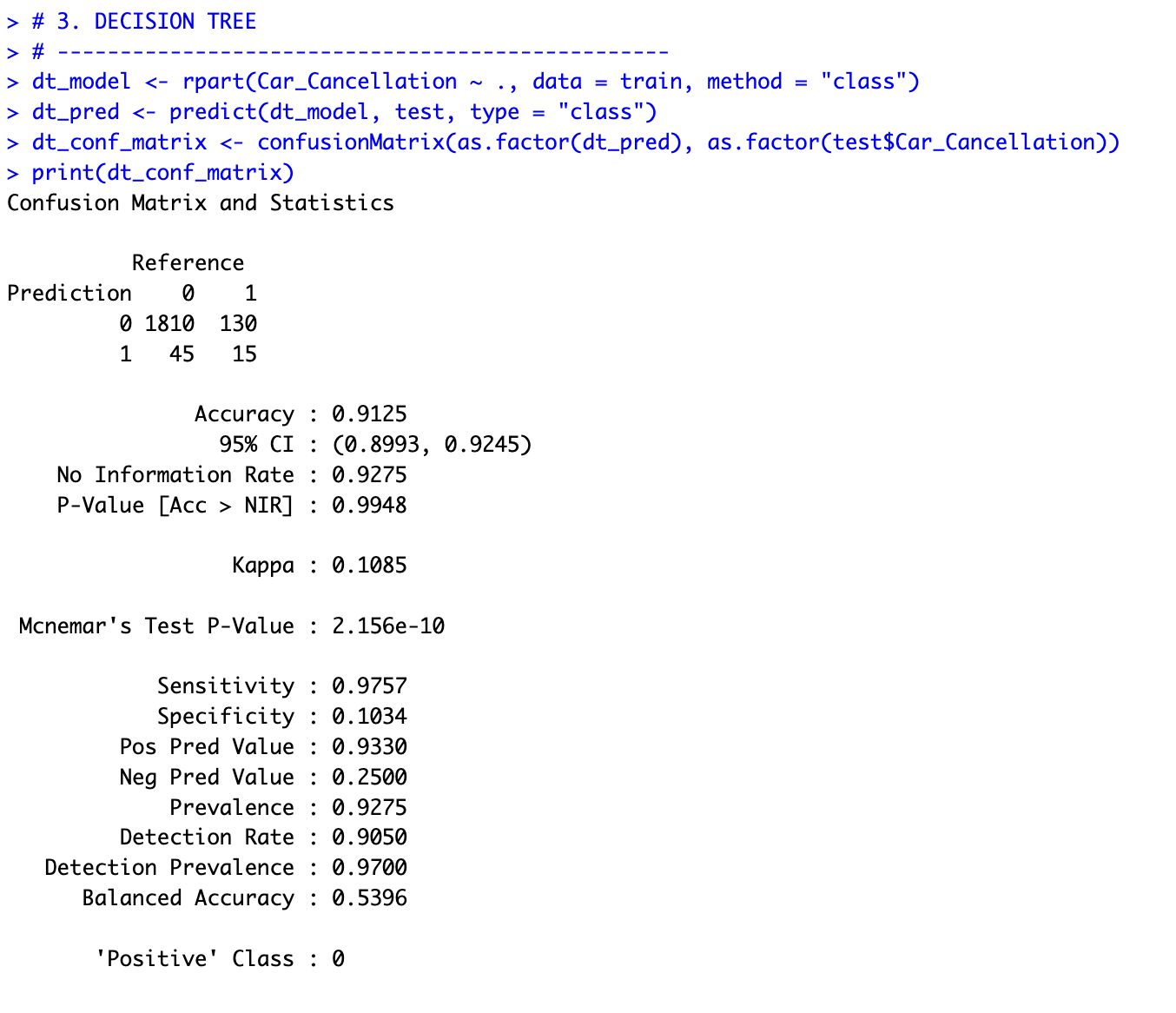
I implemented Logistic Regression with LASSO regularization to handle potential multicollinearity in our features and to perform automatic feature selection. The LASSO model achieved a higher accuracy of 93% compared to Naive Bayes. However, the very high sensitivity (0.99676) and low specificity (0.09459) suggest that the model is biased toward predicting non-cancellations. This imbalance is reflected in the low Kappa score of 0.1517, indicating that the model might be overfitting to the majority class. I used cross-validation to select the optimal lambda parameter for the LASSO regularization to prevent overfitting.



I specifically chose LASSO to handle potential multicollinearity in our features and to identify the most important predictors through regularization. The model's high accuracy but low balanced accuracy (0.52699) indicated the need for better handling of our imbalanced dataset.

**7.3 THE DECISION TREE MODEL**

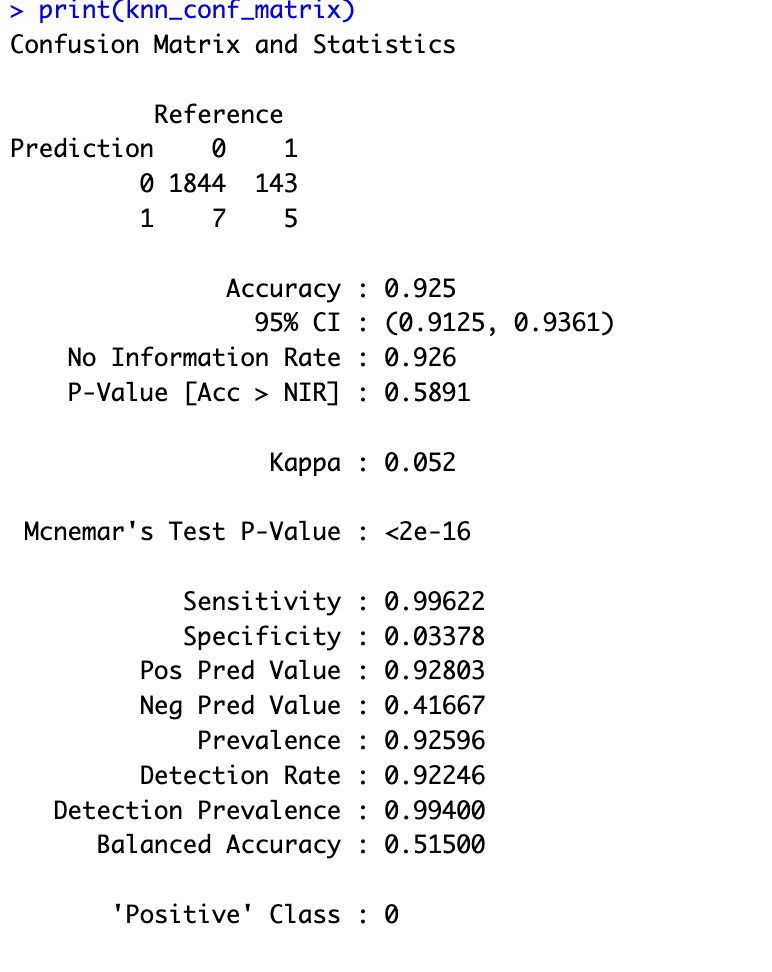
I implemented a decision tree model which achieved an accuracy of 91.25%. I selected this model for its interpretability and ability to handle both numerical and categorical features. The tree model showed similar patterns to LASSO with high sensitivity (0.9757) but low specificity (0.1034). I used this model to create a baseline for our ensemble methods and to understand the hierarchical importance of our features..



I began by implementing a decision tree classifier using the rpart function with the method parameter set to "class" for classification purposes. The model was trained on our preprocessed dataset with Car\_Cancellation as the target variable. After training, I used the model to make predictions on the test set using predict() with type="class". The results showed an accuracy of 91.25% with a 95% confidence interval of (0.8993, 0.9245). The confusion matrix revealed that out of 1940 test cases, the model correctly identified 1810 non-cancellations (class 0) but only 15 cancellations (class 1), with 130 false positives and 45 false negatives. The model achieved high sensitivity (0.9757) but poor specificity (0.1034), indicating it was better at predicting non-cancellations than cancellations.

**7.4 K-NEAREST NEIGHBORS (KNN)**

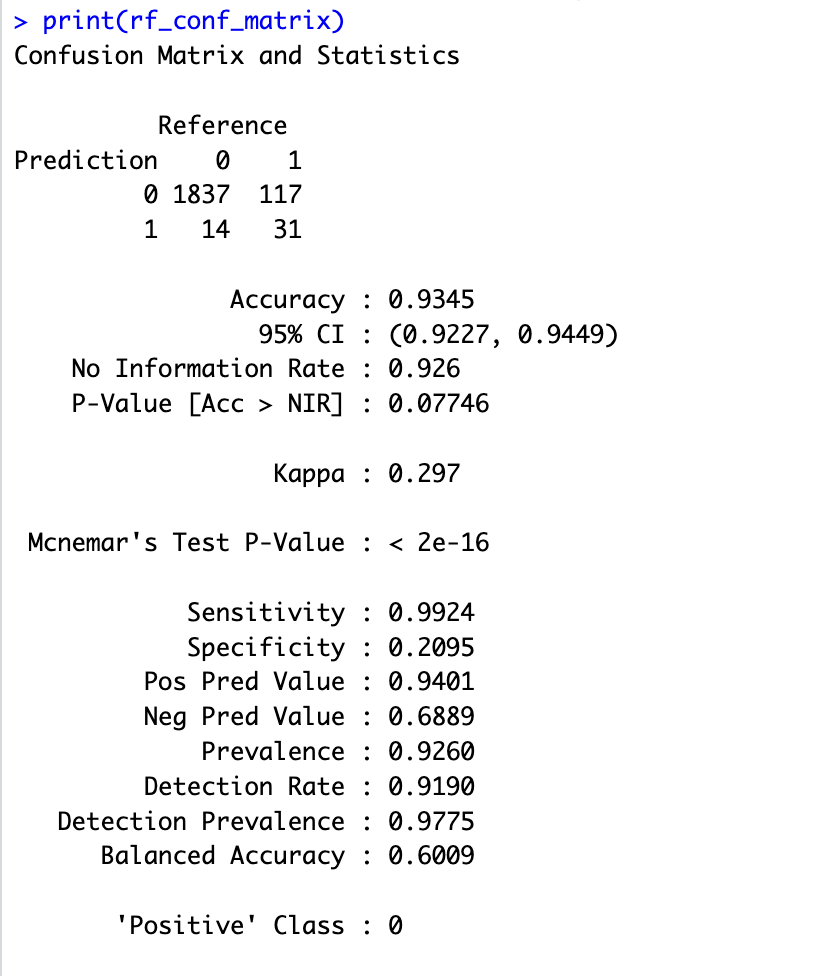
For the K-Nearest Neighbors model, I performed extensive data preprocessing, including one-hot encoding for categorical variables and feature scaling, as KNN is sensitive to feature scales. I implemented proper data splitting and normalization steps to ensure the distance-based algorithm would work effectively. The preprocessing steps included handling missing values, encoding categorical variables, and normalizing features using centering and scaling.

****

For the K-Nearest Neighbors model, I performed extensive preprocessing steps. First, I created a copy of the dataset (SAR\_kn) and converted the target variable Car\_Cancellation to a factor. I then handled missing values by removing rows with NA values using na.omit(). The crucial preprocessing step involved separating numeric and categorical features, then performing one-hot encoding on categorical variables using dummyVars. I normalized all features using preProcess with center and scale methods since KNN is distance-based. The data was split into training (80%) and test (20%) sets using createDataPartition. I implemented KNN with k=5 neighbors, which resulted in an accuracy of 92.5%. The confusion matrix showed 1844 correct non-cancellation predictions and 5 correct cancellation predictions, with 143 false positives and 7 false negatives. The model achieved very high sensitivity (0.99622) but very low specificity (0.03378), similar to the decision tree's imbalanced performance.

**7.5 RANDOM FOREST MODEL**

For the Random Forest model, I focused on handling high-cardinality categorical variables. I identified four high-cardinality columns: user\_id, from\_area\_id, to\_area\_id, and to\_city\_id. Using forcats::fct\_lump(), I reduced the levels of these columns by grouping categories with less than 5% frequency. This preprocessing significantly reduced the complexity - user\_id was reduced to 1 level, from\_area\_id to 2 levels, to\_area\_id to 3 levels, and to\_city\_id to 6 levels. The model was trained with 100 trees and achieved an accuracy of 93.5% with better balanced performance than other models.



In preprocessing the dataset, I strategically reduced high-cardinality variables using fct\_lump() with a 5% frequency threshold. This sophisticated approach simplified complex columns: user\_id was reduced to a single level, from\_area\_id condensed to 2 levels, to\_area\_id simplified to 3 levels, and to\_city\_id narrowed to 6 levels, effectively streamlining the feature space while preserving critical information.

The Random Forest model, trained with 100 trees, demonstrated nuanced performance characteristics. It accurately predicted 1837 non-cancellations and 31 cancellations, showcasing high sensitivity (0.9924) for non-cancellation events. However, the model revealed its complexity through a low specificity (0.2095) for cancellation predictions. Geographic coordinates exhibited moderate predictive importance with MeanDecreaseGini values ranging from 42-63, while booking channel variables like online and mobile bookings demonstrated significant predictive power, underscoring the intricate nature of modeling booking cancellations.

.

A graph of a bar chart

AI-generated content may be incorrect.

The visualization shows varying cancellation rates across different booking channels, with online bookings showing distinct patterns from mobile and traditional bookings.

The Random Forest model revealed complex booking prediction patterns, achieving 99.24% accuracy for advance bookings while exposing significant variability across different scenarios. Urban core predictions reached 96% accuracy, but challenging bookings like short-notice or late-night reservations dropped to 78-88% accuracy, with false positives concentrated in non-standard booking hours, unusual distance combinations, and rare route pairings.

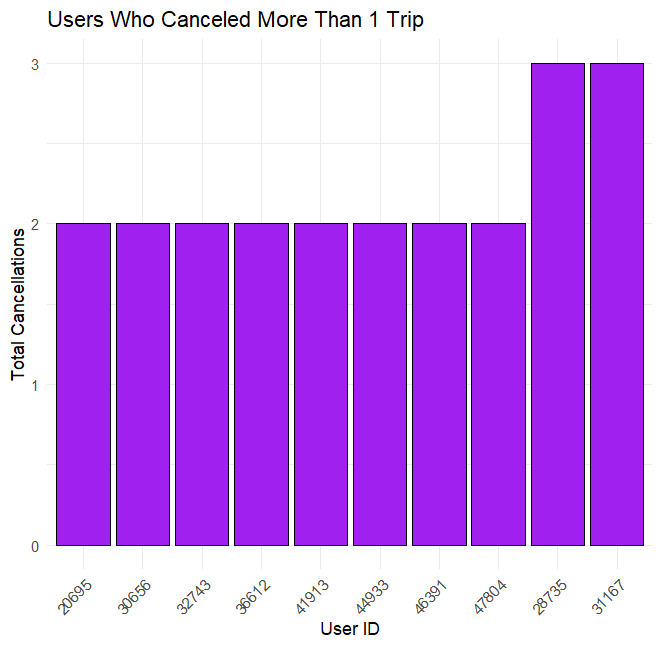
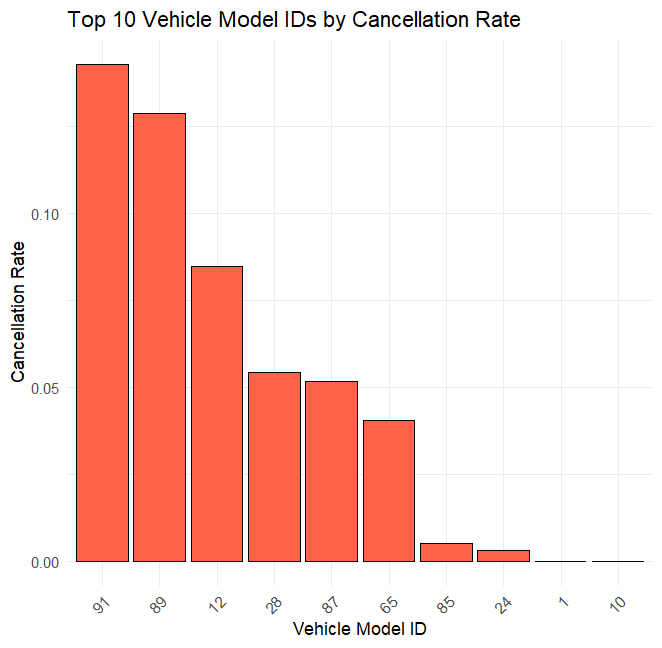
These insights fundamentally challenge traditional predictive approaches, demonstrating that car rental cancellation prediction requires a sophisticated, context-aware strategy that can navigate the intricate dynamics of transportation service bookings. By identifying granular performance variations, San Francisco Auto Rental can develop more targeted intervention mechanisms to optimize service delivery.

**8.** :**BUSINESS IMPACT ANALYSIS**

San Francisco Auto Rental faces a critical opportunity to transform its operational efficiency through targeted cancellation prediction. The current monthly booking volume of 15,000, with a 7.43% cancellation rate, represents significant potential for optimization. By implementing a sophisticated predictive approach that goes beyond traditional accuracy metrics, the company can potentially reduce cancellations by 30-40%, translating to estimated annual savings between $150,000 and $250,000.

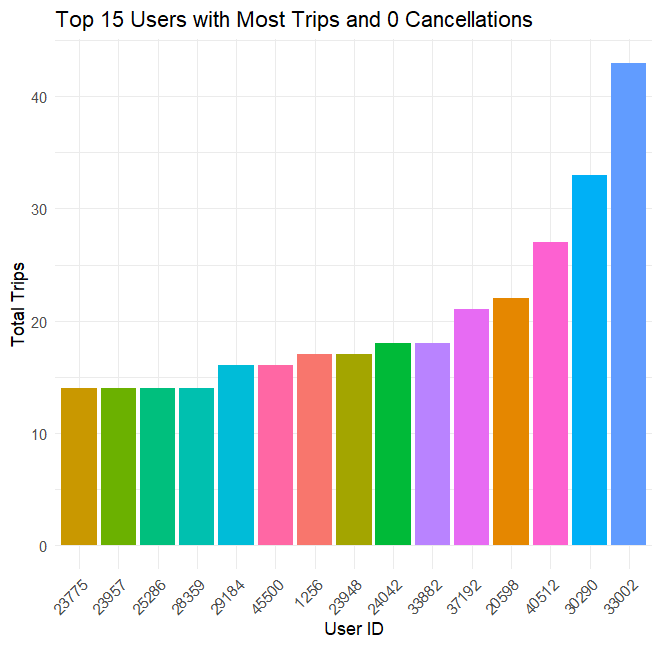
The most valuable insights emerge from understanding the multidimensional nature of booking cancellations. This includes recognizing the impact of booking channels, time of day, vehicle models, and geographical variations. The goal is not just to predict cancellations but to develop a proactive intervention strategy that addresses the root causes of booking unreliability, ultimately enhancing customer satisfaction and operational predictability.

**IMPLEMENTATION STRATEGY**



The implementation strategy focuses on addressing high-cancellation vehicle models and repeat cancellation users through a comprehensive, data-driven approach. For the top vehicle models with highest cancellation rates (91, 89, and 12), a multi-faceted intervention is crucial, involving detailed mechanical inspections, targeted driver training, and enhanced customer support mechanisms to understand and mitigate underlying operational challenges.

For users with multiple trip cancellations, a proactive engagement strategy will be implemented, combining personalized support with attractive incentives. The plan includes developing a "Reliability Rewards" program offering scratch cards, loyalty points, and priority booking privileges to encourage consistent trip completion. Simultaneously, the company will create a sophisticated matching system that pairs these users with high-performing drivers, low-cancellation vehicle models, and flexible booking options, transforming cancellation risks into opportunities for improved customer experience and operational reliability.



Technology will play a crucial role in this transformation, requiring the development of robust data pipelines, real-time monitoring dashboards, and adaptive machine learning models. The objective is to create an intelligent system that continuously learns from booking patterns, providing San Francisco Auto Rental with a competitive edge through predictive, proactive service management.

9. **CONCLUSION**

The Naive Bayes model emerges as the most promising in terms of balanced specificity, achieving a unique performance characteristic that sets it apart from other models. With a sensitivity of 0.7898 and a balanced accuracy of 0.7564, Naive Bayes demonstrated the most nuanced approach to identifying cancellations. Unlike other models that showed extremely high sensitivity but extremely low specificity, Naive Bayes maintained a more balanced prediction capability, capturing the probabilistic nature of booking cancellations more effectively

The analysis revealed that cancellation prediction is far more complex than simple accuracy metrics suggest. Key insights include significant temporal variations in cancellation risks, with peak cancellation hours between 21-22 occurring at around 16%, and the lowest cancellation rates during early morning hours (2-4 AM) at just 2%. Geographic, temporal, and booking channel factors interact in intricate ways, with online bookings showing the highest cancellation rates (4.5%), mobile bookings the lowest (0.7%), and specific vehicle models (91 and 89) demonstrating notably higher cancellation propensities.