

Athens University of Economics and Business ---------------------------------------------------

Master of Science (MSc) in Business Analytics

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Statistics for Business Analytics IΙ - FT

-Project I -

Instructors: Dimitris Karlis

Student Name: Eleni Ralli

Student ID: f2822312

15 February, 2024

Contents

[**1. Introduction** 2](#_Toc159092494)

[**2. Descriptive analysis and exploratory data analysis** 3](#_Toc159092495)

[**3. Pairwise comparisons** 6](#_Toc159092496)

[**4. Descriptive model** 8](#_Toc159092497)

[**5. Conclusions** 14](#_Toc159092498)

[**6. List of Tables and Figures.** 15](#_Toc159092499)

[**7. Appendix** 16](#_Toc159092500)

# **1. Introduction**

This study focuses on analyzing cancellation behaviors for room bookings in a hotel, aiming to explain the factors influencing the decision to cancel rather than predicting it. We utilize a random sample of bookings that includes a variety of variables such as the number of adults and children, room type, price, special requests, and others. This approach offers a deeper understanding of the dynamics shaping cancellation decisions.

For this study, we utilize a dataset comprising 2000 observations of individual room bookings in some hotel, encompassing 17 diverse variables (including nominal, ordinal, continuous, and discrete types).

Also, we create 3 more variables that will be useful for our study the number of total quests (the sum of total number of adults and total number of children), the number of total nights (the sum of total number of weekend nights and total number of week nights) and the reservation month.

Table 1: Data Table

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Number | Name | Type of Variable | Description |
|  | Booking\_ID | Nominal | Unique identifier for each booking |
|  | Number of Adults | Discrete | Number of adults included in the booking |
|  | Number of children | Discrete | Number of children included in the booking |
|  | number of weekend nights | Discrete | Number of weekend nights included in the booking |
|  | number of week nights | Discrete | Number of week nights included in the booking |
|  | type of meal | Nominal | Τype of meal included in the booking |
|  | car parking space | Nominal | Indicates whether a car parking space was requested or included in the booking |
|  | room type | Nominal | Type of room booked |
|  | lead time | Discrete | Number of days between the booking date and the arrival date |
|  | market segment type | Nominal | Type of market segment associated with the booking |
|  | repeated | Nominal | Indicates whether the booking is a repeat booking |
|  | P-C | Discrete | Number of previous bookings that were canceled by the customer prior to the current booking |
|  | P-not-C | Discrete | Number of previous bookings not canceled by the customer prior to the current booking |
|  | average price | Continuous | Average price associated with the booking |
|  | special requests | Discrete | Number of special requests made by the guest |
|  | date of reservation | Nominal | Date of the reservation |
|  | booking status | Nominal | Status of the booking (canceled or not canceled) |
|  | number of total quests | Discrete | The sum of total number of adults and total number of children |
|  | number of total nights | Discrete | The sum of total number of weekend nights and total number of week nights |
|  | reservation month | Nominal | The month of the reservation |

# **2. Descriptive analysis and exploratory data analysis**

The analysis will be performed using the R statistical package. The dataset initially contained ‘Not Available' (NA) values in two observations. Due to this, these observations would provide incomplete information for our analysis and could cause issues in conducting the analysis using R package. So, they were removed.

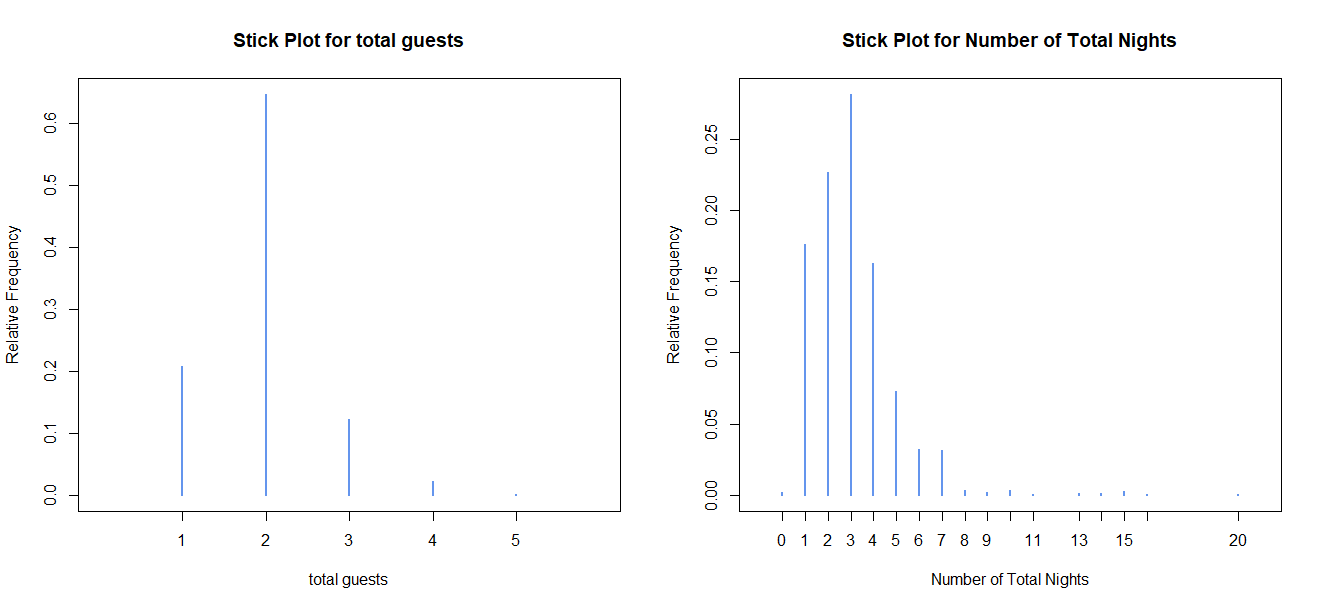
We are examining certain variables that we find interesting on an individual basis to understand the values they hold and to perform some descriptive measures. These measures help us in better understanding each variable. For the quantitative variables like number of total quests, number of total nights, lead time, number of week nights, average price, and special requests, we are looking at their mean, standard deviation, median, minimum and maximum value, skewness, and kurtosis (see Table 2).

From this numeric variable that we examine no one is close to a normal distribution. Number of total quests, number of total nights, lead time, number of week nights, and special requests are discrete variables so this is one more reason to doesn’t be close to a normal distribution. All these numeric variables that we choose to examine show a tendency to have more values on the left side of the distribution (right skew- in a perfect normal distribution, skewness is 0). Also, number of total nights and number of week nights have higher kurtosis values in contrast to a normal distribution (in a perfect normal distribution kurtosis is 3) and variables number of total quests, lead time average price, and special requests have lower kurtosis in contrast to a normal distribution (fewer extreme outliers than a normal distribution) (see Figure 1).

We will examine these hypotheses by conducting normality tests and creating the corresponding QQplots (see Appendix - Figure 2). The tests reveal that none of the numeric variables follows a normal distribution (Shapiro-Wilk and Kolmogorov-Smirnov p-value < 2.2e-16), a finding that is also graphically confirmed by observing both the probability density diagrams (see Figure 1) and the QQplots (see Appendix - Figure 2).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Numeric Variables | Mean | Median | Sd | Minimum | Maximum | Skewness | Kurtosis |
| Number of total quests | 2 | 2 | 1 | 1 | 5 | 0.6 | 4.2 |
| Number of total nights | 3 | 3 | 2 | 0 | 20 | 2.2 | 14.4 |
| lead time | 83 | 56 | 83 | 0 | 418 | 1.3 | 4.2 |
| number of week nights | 2 | 2 | 1 | 0 | 14 | 1.5 | 9.7 |
| average price | 103 | 98 | 35 | 0 | 350 | 0.6 | 6.0 |
| special requests | 1 | 0 | 1 | 0 | 4 | 1.1 | 3.5 |

Table 2: Descriptive Statistics (rounded) Table for some Quantitative Variables



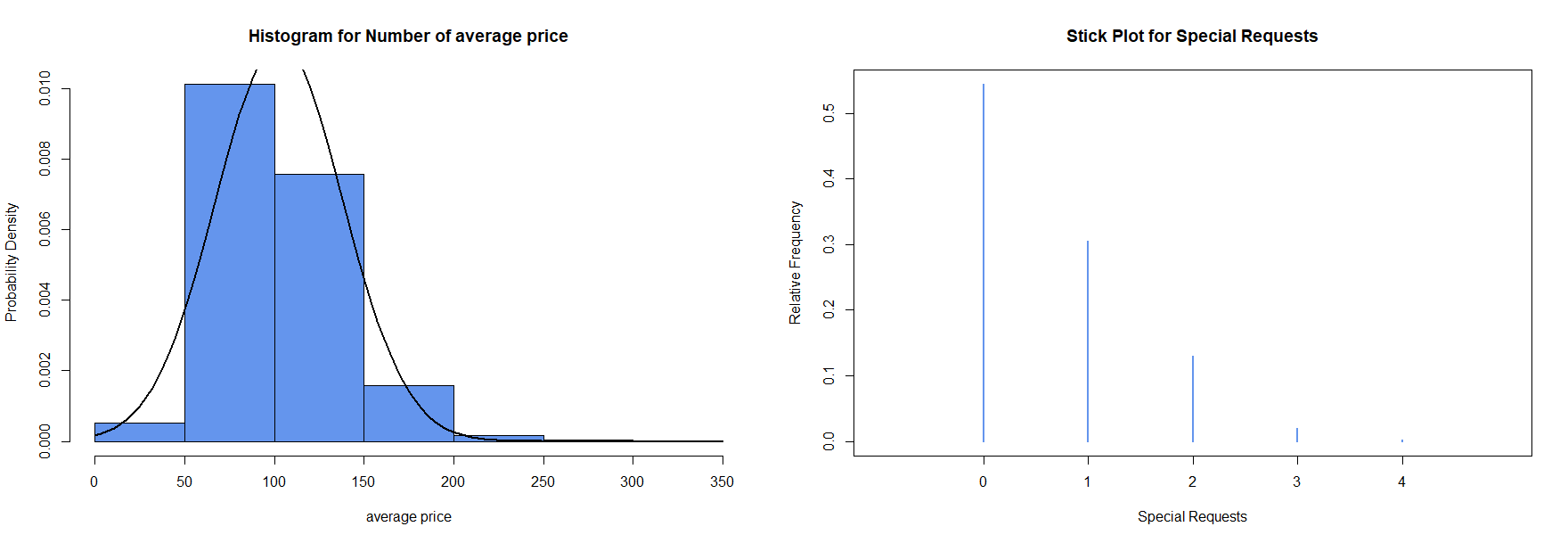
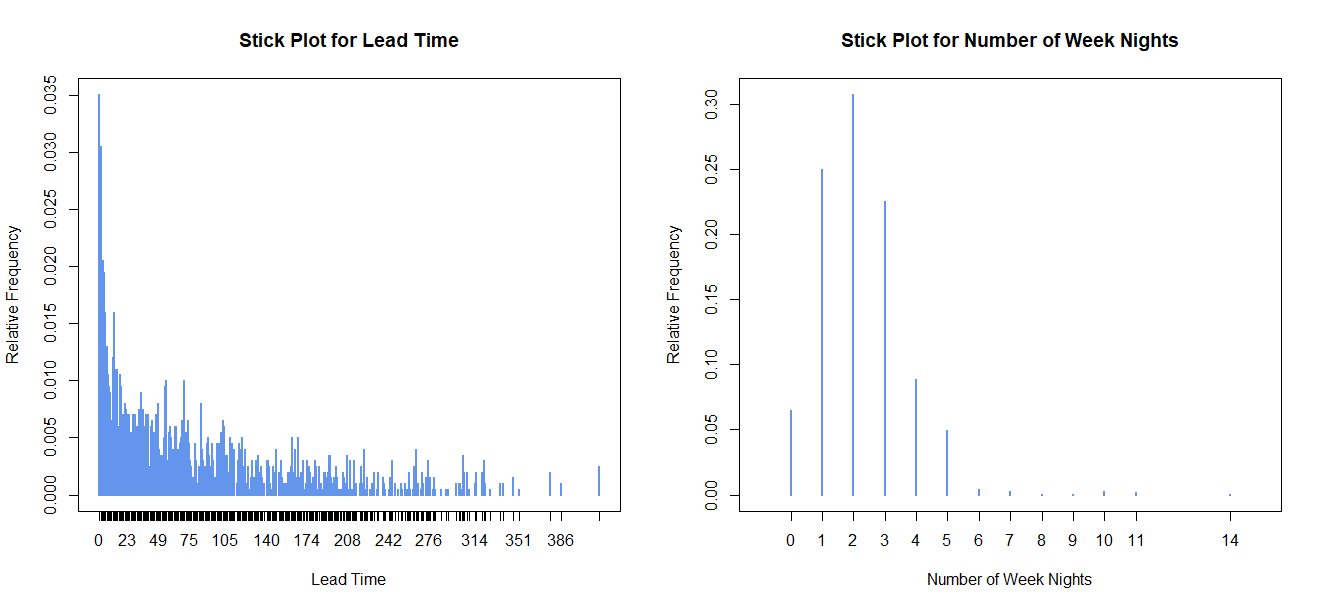


Figure 1: Probability Density Diagrams/ Relative Frequency Diagrams of some Quantitative Variables

We will examine 3 categorical variables from the dataset. The variables market segment type, reservation month and booking status and we see the frequency and percentage distribution of them. These variables show a distribution skewed towards higher levels (see Figure 3 and Table 3).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Categorical Variable | Level:  Aviation | Level: Complementary | Level:  Corporate | Level:  Offline | Level:  Online |
| market segment type | 7  0.4% | 23  1.2% | 110  5.5% | 590  29.5% | 1270  63.4% |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Categorical Variable | Level  01 | Level  02 | Level  03 | Level  04 | Level  05 | Level  06 | Level  07 | Level  08 | Level  09 | Level  10 | Level  11 | Level  12 |
| reservation month | 101  5.1% | 137  6.8% | 139  7.0% | 167  8.4% | 161  8.0% | 187  9.4% | 146  7.3% | 183  9.2% | 245  12.2% | 230  11.5% | 133  6.6% | 171  8.5% |

|  |  |  |
| --- | --- | --- |
| Categorical Variable | Level: Canceled | Level: Not\_Canceled |
| Booking status | 678  33.9% | 1322  66.1% |

Table 3: Frequency table and percentages(rounded) of the levels of the categorical variable market segment type and reservation month

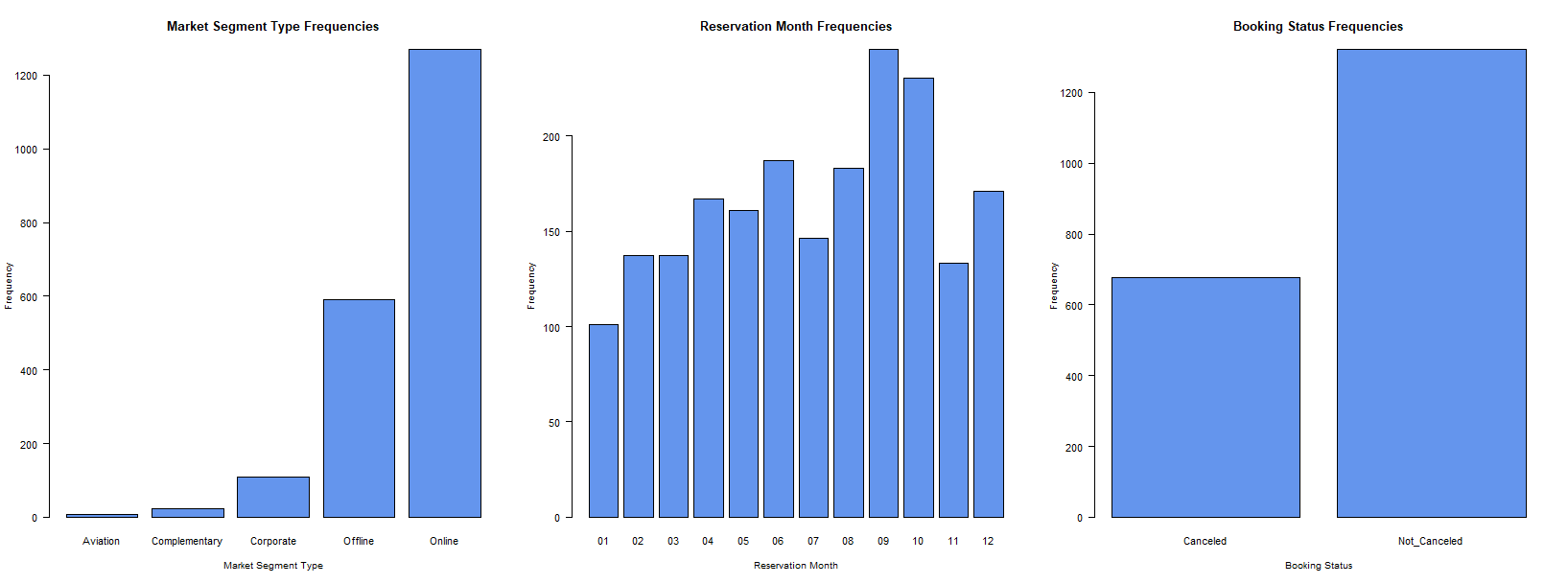


Figure 3: Bar charts of the categorical variable market segment type and reservation month

# **3. Pairwise comparisons**

An important thing is to examine the relationship of the variable (booking status) for which we want to draw conclusions about how it is affected by the other variables in the dataset. Initially, we will examine the relationship of some numeric variables in pairs to get a first picture of how they relate to each other and how each relates to the booking status.

Pairwise comparison between numeric variables

To explore the relationship between certain numerical variables of interest, we apply the Spearman correlation matrix. This reveals that some of the numerical variables we are examining do not have a monotonic relationship with each other (Pearson linear correlation coefficient lower than 0.2 in each pair - see Figure 4). Also, as we have seen earlier, the distributions of these variables do not follow a normal distribution, so we check whether they have a monotonic relationship using a non-parametric test. For the variables average price and special requests (Spearman correlation test p-value < 2.483e-15 < 0.05), the number of weeknights with lead time (Spearman correlation test p-value < 2.2e-16 < 0.05), and the number of weeknights with special requests (Spearman correlation test p-value < 0.04597 < 0.05), there is an indication of a monotonic relationship between them. For average price with number of weeknights (Spearman correlation test p-value = 0.1175 > 0.05) or lead time (Spearman correlation test p-value = 0.3095 > 0.05), as well as lead time with special requests (Spearman correlation test p-value = 0.07097 > 0.05), there is no indication of a monotonic relationship between them.

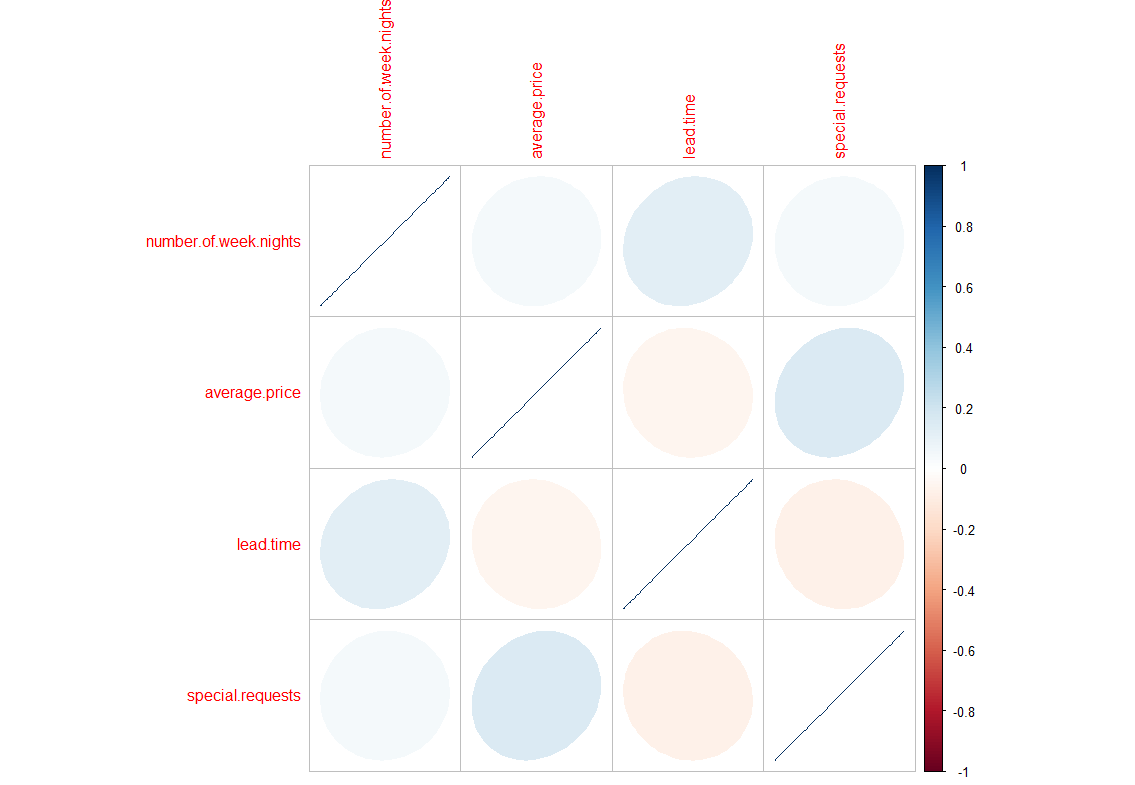


Figure 4: Spearman Correlation Matrix of some Hotel Booking Variables: Number of Week Nights, Average Price, Lead Time, and Special Requests

Pairwise comparison between categorical and numeric variables

We will examine the relationship between the categorical variable booking status and some numerical variables that appear to be of interest. For the relationship of booking status with average price, the hypothesis of normality is rejected (S-W p-value and K-S p-value < 2.2e-16 < 0.05), and the hypothesis of equality of means among groups is also rejected (Welch's t-test p-value < 2.671e-14< 0.05) These differences indicate that the average price has a significant impact on the booking status. (see Appendix -Figure 5). For the relationship of booking status with lead time, the hypothesis of normality is rejected (S-W p-value and K-S p-value < 2.2e-16 < 0.05), and the hypothesis of equality of medians among groups is also rejected (Wilcoxon test p-value < 2.2e-16 < 0.05) These differences indicate that the lead time has a significant impact on the booking status. (see Appendix -Figure 6). For the relationship of booking status with special requests, the hypothesis of normality is rejected (S-W p-value and K-S p-value < 2.2e-16 < 0.05), and the hypothesis of equality of medians among groups is also rejected (Wilcoxon test p-value < 2.2e-16 < 0.05) These differences indicate that the special requests have a significant impact on the booking status.

Pairwise comparison between categorical variables

We will examine the relationship between the categorical variable booking status and some categorical variables that seem interesting to us. For the relationship of booking status with market segment type I have indications that there is a dependency between the market segment type and booking status (Pearson's Chi-squared simulate p-value = 0.0005< 0.05) (see Appendix -Figure 7). For the relationship of booking status with car parking space I have indications that there is dependency between booking status and car parking space (Pearson's Chi-squared p-value = 1.039e-05<0.05) (see Appendix -Figure 8).

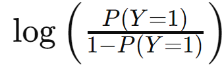
# **4. Descriptive model**

Our goal in this chapter is to estimate how and which of our variables affect the sale booking status. We want to construct reasonable model selecting covariates and being able to use the model to explain the behavior of cancelations. We 're focused on explaining the behavior rather than predicting it, the model coefficients provide insights into how each factor might influence the probability of cancellation.

Initially, for modeling a binary outcome such as booking cancellation (where the response variable is binary either "canceled" or "not canceled") we will use Generalized Linear Model. The GLM is a flexible generalization of ordinary linear regression that allows for the response variable that have a distribution other than a normal distribution. The logit link function is specifically used for binary logistic regression, where the outcome is binary- coded as 0 (“not canceled”) or 1 (“canceled”).

First of all, due to the large number of variables, we will conduct a method used for the selection of variables in regression models. This method is called LASSO (Least Absolute Shrinkage and Selection Operator). The model proposed by the LASSO method suggests including the variables number of weekend nights, car parking space, lead time, market segment type, average price, special requests, reservation month, and total nights in the model. Nonetheless, the coefficients derived from this method are biased, so the model is not reliable for interpreting the behavior of cancellations. Moreover, the null deviance (which is a measure of how well the response variable is predicted by a model with no predictors) in our model is quite high (2561.5 on 1999 degrees of freedom), which suggests that the outcome is not well predicted by the intercept alone. The residual deviance (which is a measure of how well the response variable is predicted by the model including the predictors) (1720.7 on 1977 degrees of freedom) is much lower than the null deviance, indicating that our model provides a better fit to the data than the null model. Our model has an AIC (Akaike Information Criterion - it balances model fit (likelihood) with model complexity (number of parameters). A lower AIC suggests a better model) of 1766.7, which doesn't mean much on its own but can be used to compare with other models built on the same dataset (see Appendix -Figure 9).

Next, we will retain the variables suggested by the LASSO method and perform another method called Stepwise Regression, in which we will use the critical value of the Bayesian Information Criterion (BIC) to determine which variables will be added to or removed from the model. This process selects the model that provides the best balance between complexity and explanation of the variability of the data. The model proposed by Stepwise Regression with BIC criterion suggests including the variables car parking space, lead time, market segment type, average price, special requests in the model.



= *- 2.77 - 2.34* × car.parking.space1 + *0.02*​ × lead.time *- 12.82* ​× market.segment.typeComplementary *- 0.8* ​× market.segment.typeCorporate *-1.65* ​× market.segment.typeOffline + *0.34* ​× market.segment.typeOnline +  *0.02* × average.price  *-1.52* × special.requests .

The residual deviance here (is a measure of how well the response variable is predicted by the model including the predictors) (1755.4 on 1977 degrees of freedom) is much lower than the null deviance, indicating that our model provides a better fit to the data than the null model. Our model has a AIC (it balances model fit (likelihood) with model complexity (number of parameters)- a lower AIC suggests a better model) of 1773.4 and BIC of 1823.79.

The market segment type does not show a statistically significant effect, except possibly for the 'Offline' category (which has a Wald test p-value= 0.053 -close to 0.05). So, we will make a comparison between two Nested Generalized Linear Models (applying a Chi-squared test). Essentially, we will compare the difference in deviance between the two models to determine if the more complex version of the model (i.e., the one that includes the market.segment.type variable) significantly reduces the residual deviance compared to the simpler model that does not include it. It turns out (Chi-squared p-value < 2.2e-16 < 0.05) that the more complex model with the market segment type provides a significantly better fit than the simpler model without it, and we should keep the more complex model.

Additionally, we conducted a Goodness Of Fit test (Chi-squared p-value = 0.99 > 0.05), which is quite high, thereby suggesting that we do not have statistically significant evidence to question the quality of the model fit. In other words, there is no reason to believe that the model does not fit the data well based on the observed deviance. This was logical and expected since rarely in logistic regression do we have a problem with goodness of fit as the dependent variable can only take two values, 0 (“not canceled”) or 1 (“canceled”). Also, our model offers a significant improvement in predicting the dependent variable compared to the null model (McFadden's Pseudo R² = 0.315). Finally, the log likelihood measures the likelihood of the observed data under the given model, with higher values indicating a better fit of the model to the data. For the model we selected (log Lik. -877.6896), it results in a significantly better fit to the data compared to the model that contains only the intercept (log Lik. -1280.74). The ratio of the two log likelihoods from two models is often used to create a measure called "Likelihood Ratio Test" (LRT). This test compares the likelihood of two models to determine if the more complex model provides a statistically significant better fit to the data compared to the simpler model (Deviance Difference = 806.1 and Chi-squared test p-value < 2.2e-16 < 0.05).

We conducted a test for multicollinearity and found that there is no problem with variables being strongly correlated with each other to such an extent that it's difficult to discern the independent effect of each variable on the dependent variable (GVIF<3.16). We also checked if our observations are independent in the sample provided and found no such issue (scatter plot of the standardized residuals against the number of observations – we do not see any patterns or systematic trends) (see Appendix -Figure 10). Finally, we examined the residuals - the differences between the actual observations and the predicted values from the model (person residuals) through diagrams and found that there are some issues with heteroscedasticity and several influential points (data that may have a disproportionate impact on the model's estimates), which, however, we will not remove from our data as after observing them, we do not consider them to be erroneous entries and they provide useful information (see Appendix -Figure 11, 12, 13).

This model shows that certain predictors are statistically significant and have a clear association with the likelihood of a booking being canceled.

According to the standardized coefficients of the model, the 'lead time' variable appears to have the strongest effect, followed by 'special requests' and 'average price'. Conversely, the categories of the 'market.segment.type' variable do not seem to have a significant effect on the probability of cancellation, as their values are closer to zero (see Figure 14).

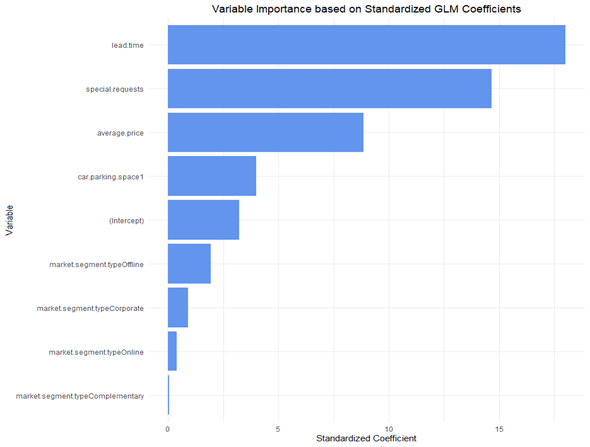


Figure 14: Variable Importance based on Standardized GLM Coefficients

It's also important to consider the context and practical significance, not just statistical significance, when interpreting these results. So, instead of trying to predict the booking status is canceled (the probability that booking 𝑥 has canceled 𝑝(𝑥)), we’ll focus on the probability that a booking has canceled vs the probability that has not canceled (see Table 4).

The ratio is called the odds ratio.

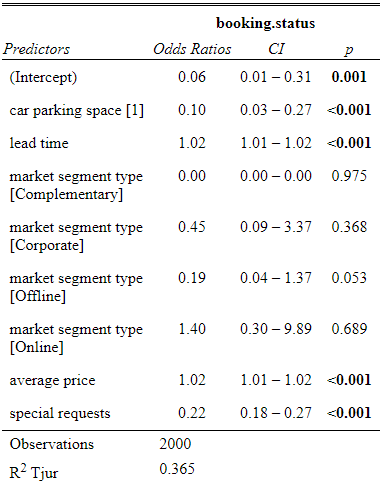


Table 4: Logistic Regression Analysis, Odds Ratios and Confidence Intervals for Booking Status

intercept: The probability of booking cancelation when all numeric variables are equal to 0 and there is no car parking space (car parking space=0) and the market segment type equal to aviation is equal to 0.059 (1/ (1 + exp (2.77))). This explanation is no meaningful because the average price for a room booking can’t be equal to 0.

**Car parking space:** The negative coefficient value (-2.34) for the variable car.parking.space1 in the model indicates that the probability of a booking cancellation decreases when the booking includes a car parking space. In other words, bookings that do not include a parking space are more likely to be cancelled compared to those that do include one. Additionally, the odds ratio is close to 0.10 (exp(-2.34)), which suggests that bookings with a parking space have about 1/10 the probability of being cancelled compared to bookings without a parking space. This indicates a reduction of 90% in the odds when bookings have a parking space (car parking space=1) compared to when they do not have a parking space (car parking space=0). For both conclusions, we assume that the numeric independent variables remain constant and other categorical variables are at their reference level (market segment type = Aviation) (see Figure 15).

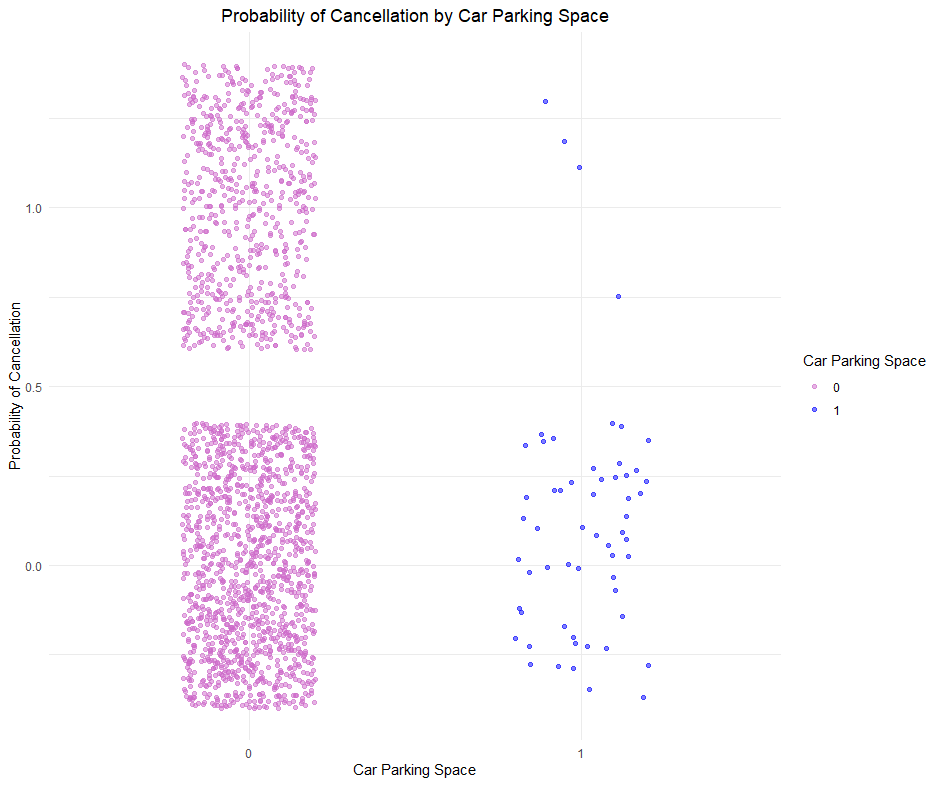


Figure 15: Probability of Cancellation by Car Parking Space

**Lead time**: The positive coefficient value (0.02) indicates that the probability of cancellation increases as the time from booking to the arrival date increases. Additionally, the odds ratio is close to 1.02 (exp(0.02)), which suggests that for each additional day between booking and arrival, the probability of cancellation slightly increases by 2% or we can say that one unit increase in lead time will increase the odds of booking cancellation by 1.02 times. For both conclusions, we assume that other numeric independent variables remain constant and other categorical variables are at their reference level (car parking space =0, market segment type = Aviation) (see Figure 16).

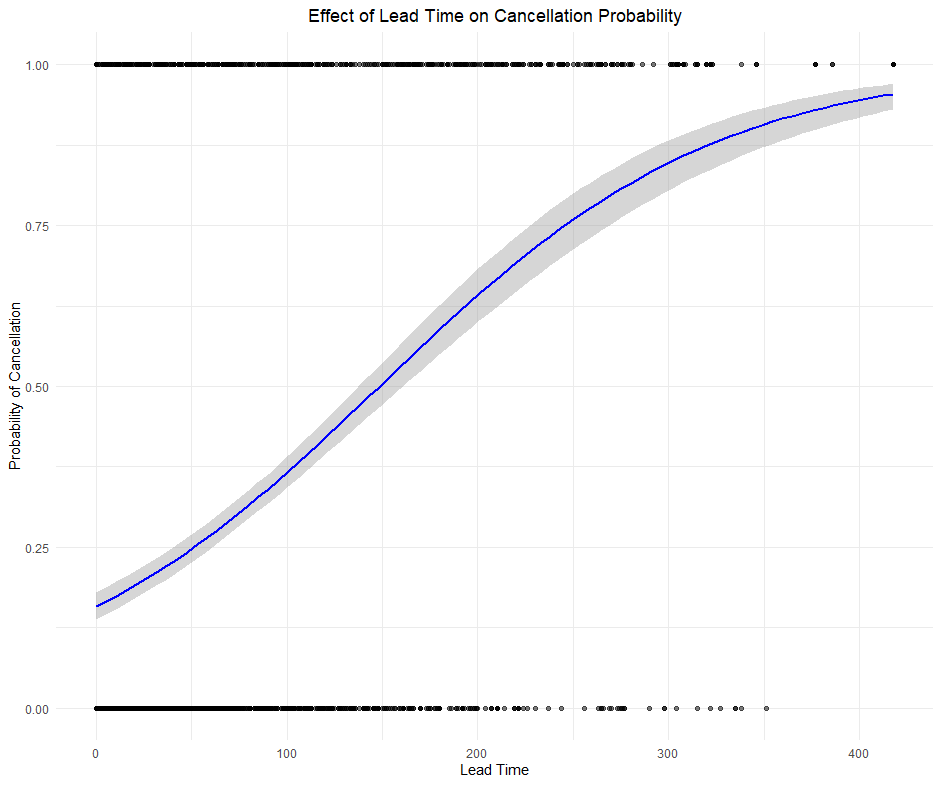


Figure 16: Effect of Lead Time on Cancellation Probability

**Market segment type**: Different categories of the market segment have different coefficients, but none are statistically significant except for the 'Offline' category. Therefore, there is not enough evidence to assume that there is a specific relationship or difference in observations corresponding to these categories.

**Market segment type [Offline]**: which has a negative coefficient (-1.65) indicating a potential reduction in the probability of cancellation when the booking is made offline compared to having market segment type = Aviation. Additionally, the odds ratio is close to 0.19 (exp(-1.65)), suggesting that bookings made offline have about 1/5 the probability of being canceled compared to those with market segment type = Aviation. In other words, the probability of cancellation decreases by about 81% when market segment type = offline compared to when market segment type = Aviation. For the above conclusions, assuming that the numeric independent variables remain constant and other categorical variables are at their reference level (car parking space =0)

**Average price**: The positive coefficient value (0.02) for the variable average.price in the model indicates that higher prices are associated with an increased probability of cancellation. Additionally, the odds ratio is close to 1.02 (exp(0.02)), which suggests that for each additional unit of price, the probability of cancellation increases by approximately 2% or one unit increase in the average price will increase the odds of booking cancellation by 1.02 times. Assuming that other numeric independent variables remain constant and other categorical variables are at their reference level (car parking space =0, market segment type = Aviation) (see Figure 17).

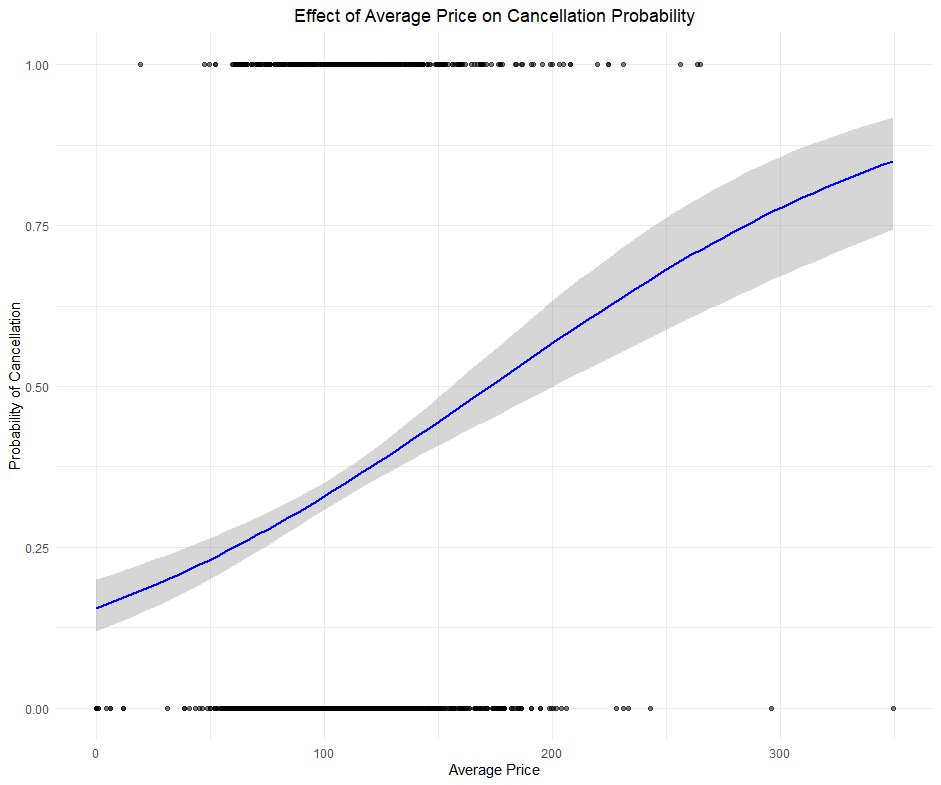


Figure 17: Effect of Average Price on Cancellation Probability

**Special requests**: The negative coefficient value (-1.52) for the variable special.requests in the model indicates that more special requests are associated with a decreased probability of cancellation. Additionally, the odds ratio is close to 0.22 (exp(-1.52)), which suggests that with each additional special request, the probability of cancellation decreases by approximately 78%. Assuming that other numeric independent variables remain constant and other categorical variables are at their reference level (car parking space =0, market segment type = Aviation) (see Figure 18).

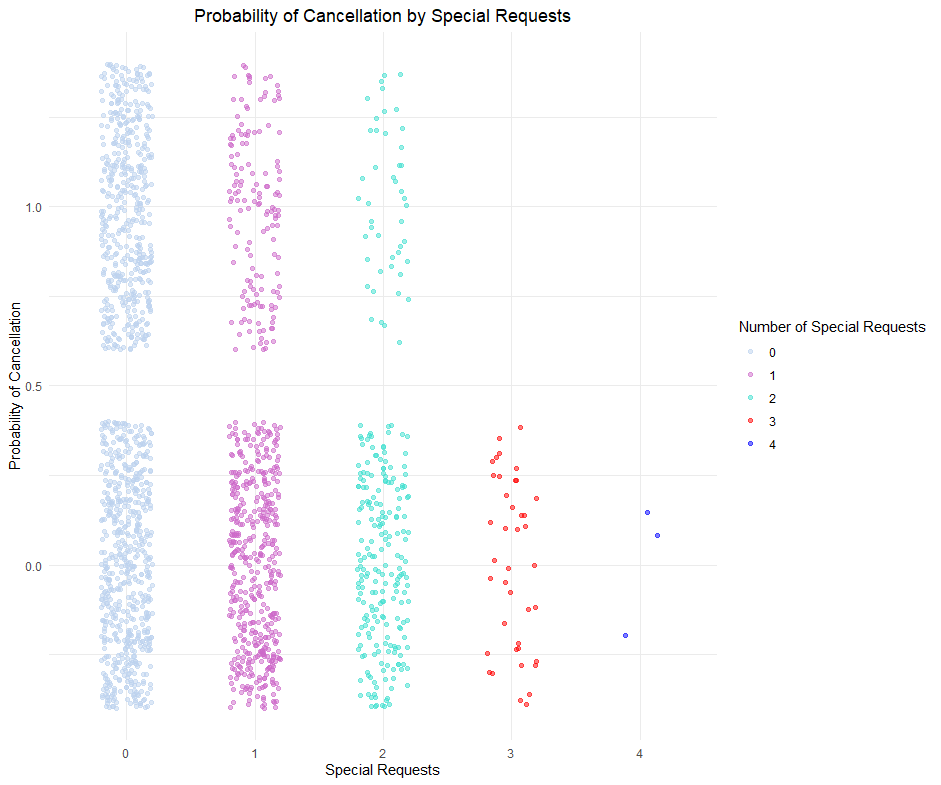


Figure 18: Probability of Cancellation by Special Requests

# **5. Conclusions**

The study provides a comprehensive analysis of the factors influencing hotel booking cancellations. Using logistic regression and other statistical methods, it is found that variables such as lead time, availability of parking space, special requests, and average price have a statistically significant correlation with the probability of cancellation. A model is proposed that offers a better fit to the data compared to a model without predictors, with significant indications for the quality of the fit. The paper concludes on the importance of understanding these factors for improving reservation management and pricing policies.

# **6. List of Tables and Figures.**

Table 1: Data Table

Table 2: Descriptive Statistics Table for some Quantitative Variables

Table 3: Frequency table and percentages(rounded) of the levels of the categorical variable market segment type and reservation month

Table 4: Logistic Regression Analysis, Odds Ratios and Confidence Intervals for Booking Status

Figure 1: Probability Density Diagrams/ Relative Frequency Diagrams of some Quantitative Variables

Figure 2: QQplots for number of total quests, number of total nights, lead time, number of week nights, average price, and special requests variables

Figure 3: Bar charts of the categorical variable market segment type and reservation month

Figure 4: Spearman Correlation Matrix of some Hotel Booking Variables: Number of Week Nights, Average Price, Lead Time, and Special Requests

Figure 5: Error bar of Average Price Grouped by booking status Levels

Figure 6: Boxplot of lead time Grouped by booking status Levels

Figure 7: Barplot for the Proportion of booking status by market segment type

Figure 8: Barplot for the Proportion of booking status by car parking space

Figure 9: LASSO Regression

Figure 10: Standardized Residuals for Logistic Regression Model

Figure 11: Plot of Pearson residuals vs. Fitted Values and binned residual plot

Figure 12: Pearson Standardized Residuals for each Predictor Variable

Figure 13: Analysis of Leverage Points and Outliers in Logistic Regression Model

Figure 14: Variable Importance based on Standardized GLM Coefficients

Figure 15: Probability of Cancellation by Car Parking Space

Figure 16: Effect of Lead Time on Cancellation Probability

Figure 17: Effect of Average Price on Cancellation Probability

Figure 18: Probability of Cancellation by Special Requests

Figure 19: Effect of Average Price and Parking Space Demand on Booking Cancellation Probability

Figure 20: Effect of Lead Time and Parking Space on Cancellation Probability

Figure 21: Effect of Market Segment Type and Average Price on Cancellation Probability

Figure 22: Effect of Special Requests and Car Parking Space on Cancellation Probability

# **7. Appendix**

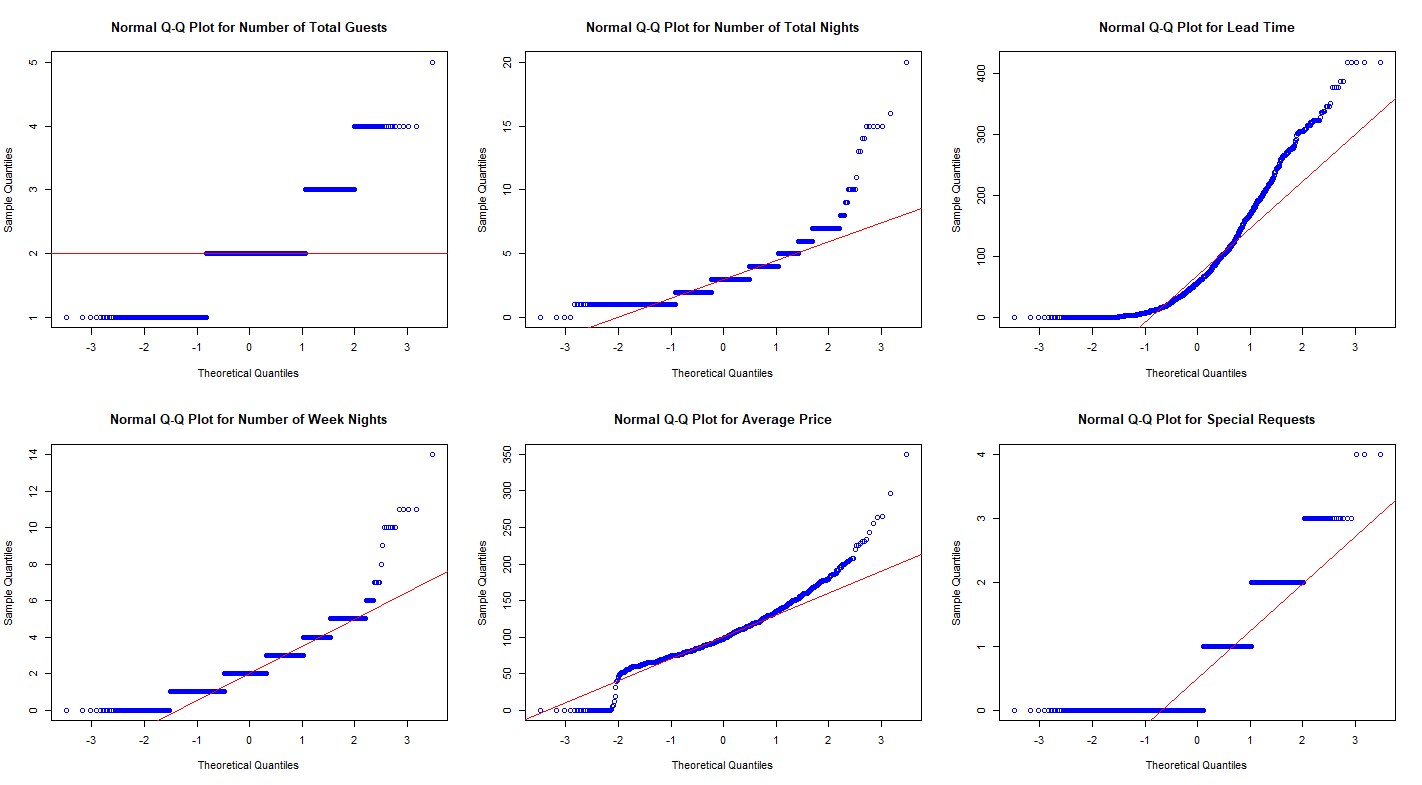
****

Figure 2: QQplots for number of total quests, number of total nights, lead time, number of week nights, average price, and special requests variables

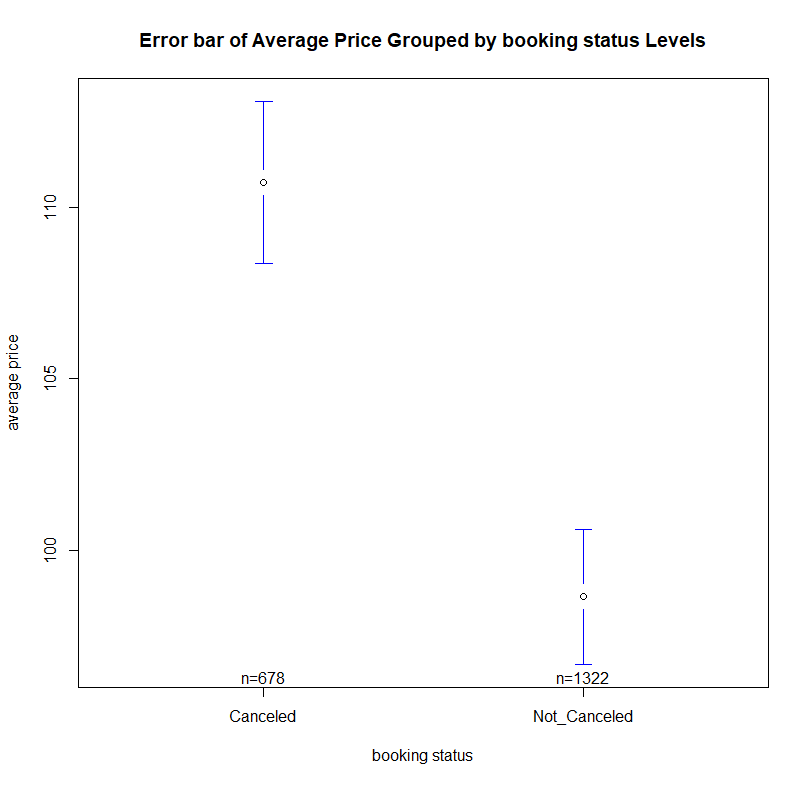
****

Figure 5: Error bar of Average Price Grouped by booking status Levels

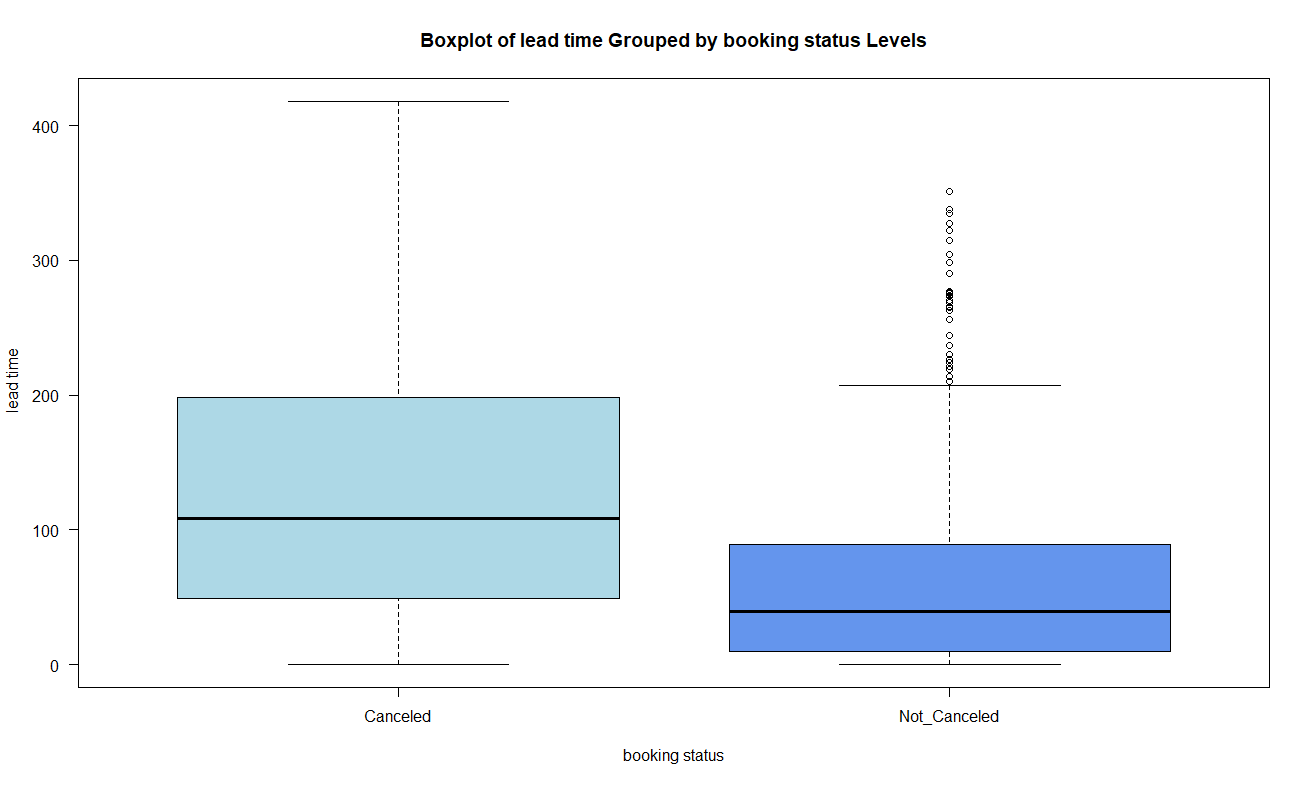


Figure 6: Boxplot of lead time Grouped by booking status Levels

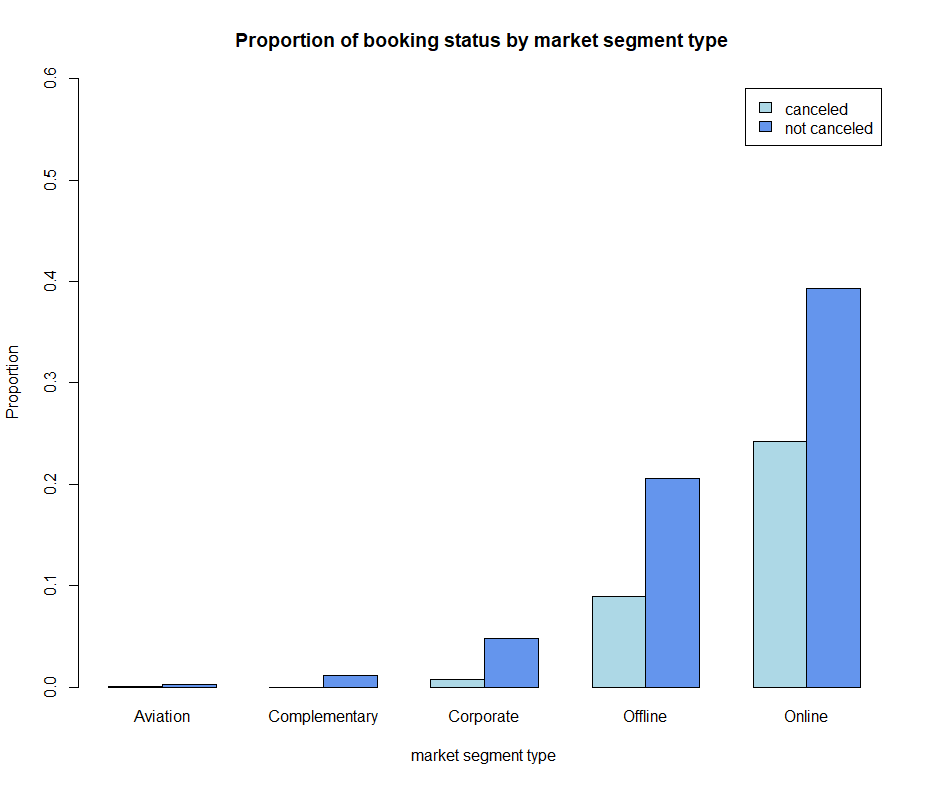


Figure 7: Barplot for the Proportion of booking status by market segment type

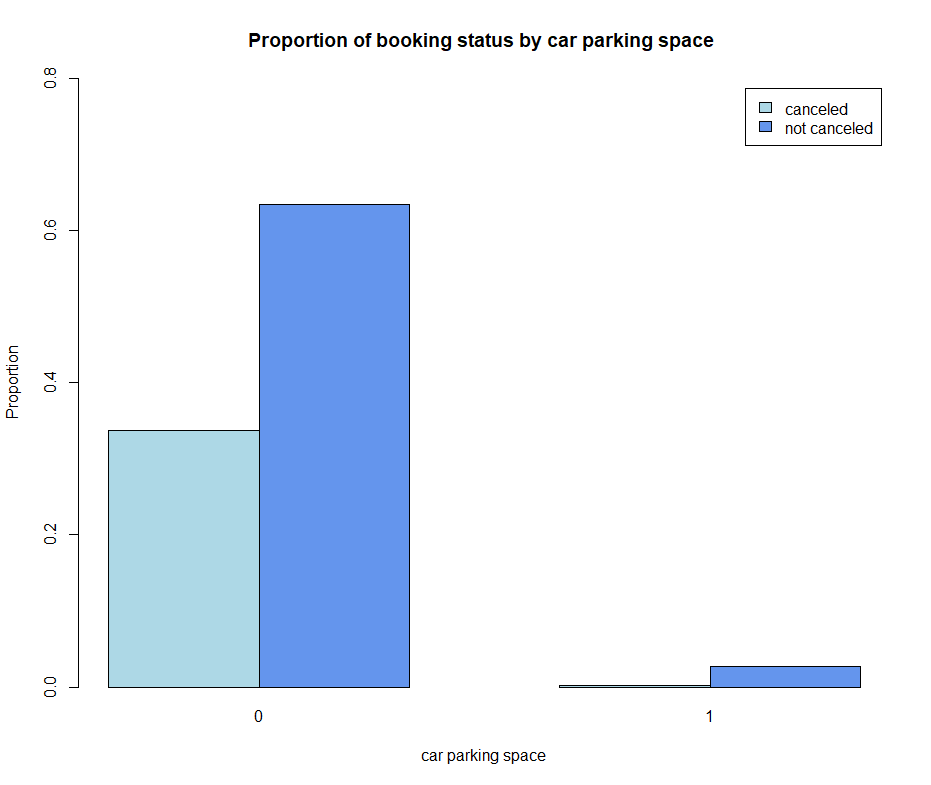


Figure 8: Barplot for the Proportion of booking status by car parking space

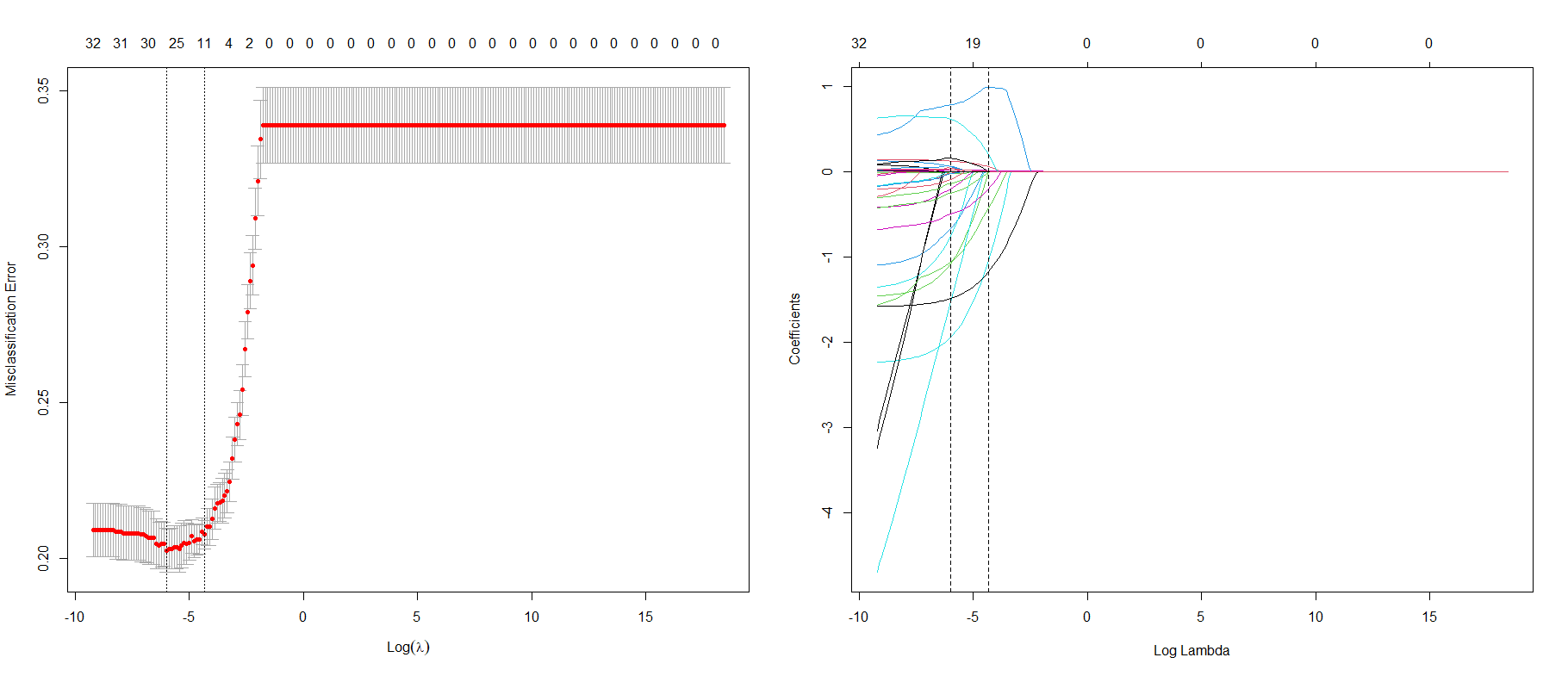
****

Figure 9: LASSO Regression

Left Graph: Cross-Validation for Optimal Lambda in LASSO Regression. This graph shows the cross-validation curve, plotting misclassification error against the logarithm of lambda values used in the LASSO. The vertical dotted lines represent the lambda values that result in the minimum error and the most regularized model within one standard error of the minimum.

Right Graph: Coefficient Path of LASSO Regression. It illustrates how the coefficients of the predictors shrink towards zero as the regularization penalty (lambda) increases. Each line represents a coefficient from the regression model, with the log lambda on the x-axis and the coefficient value on the y-axis. The vertical line indicate the selected value of lambda based on cross-validation.

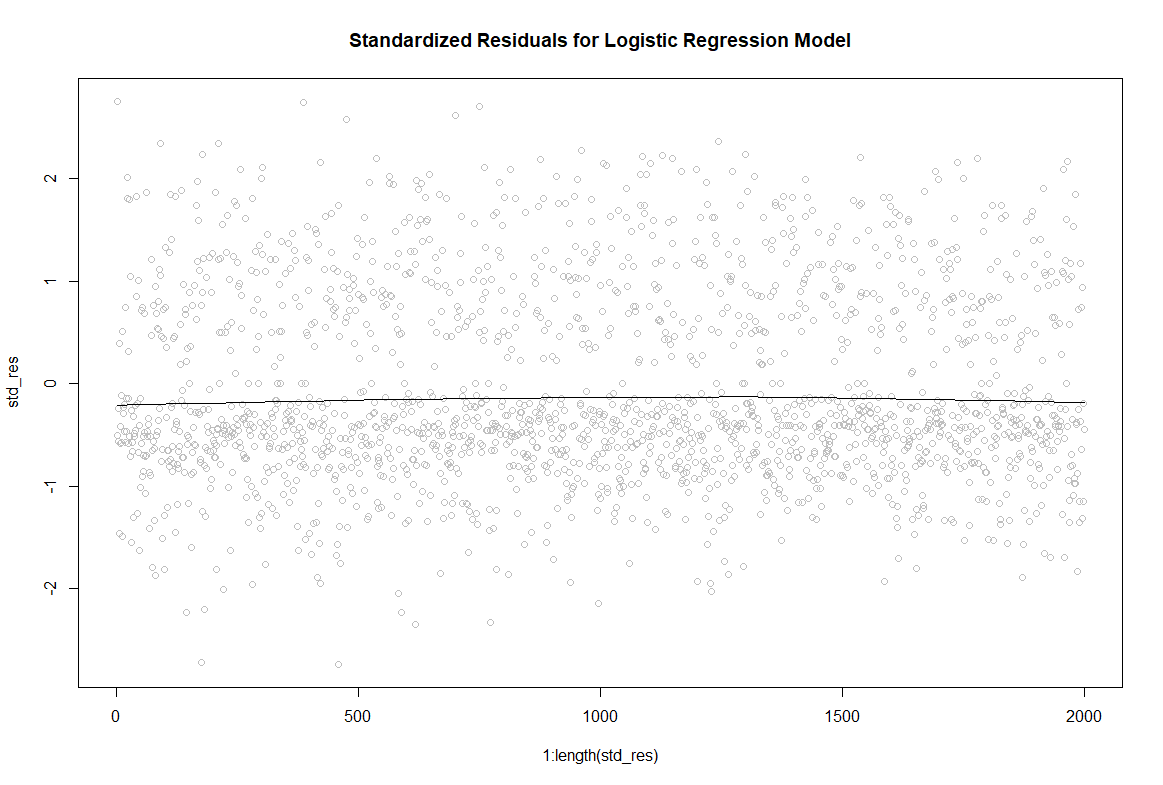
****

Figure 10: Standardized Residuals for Logistic Regression Model

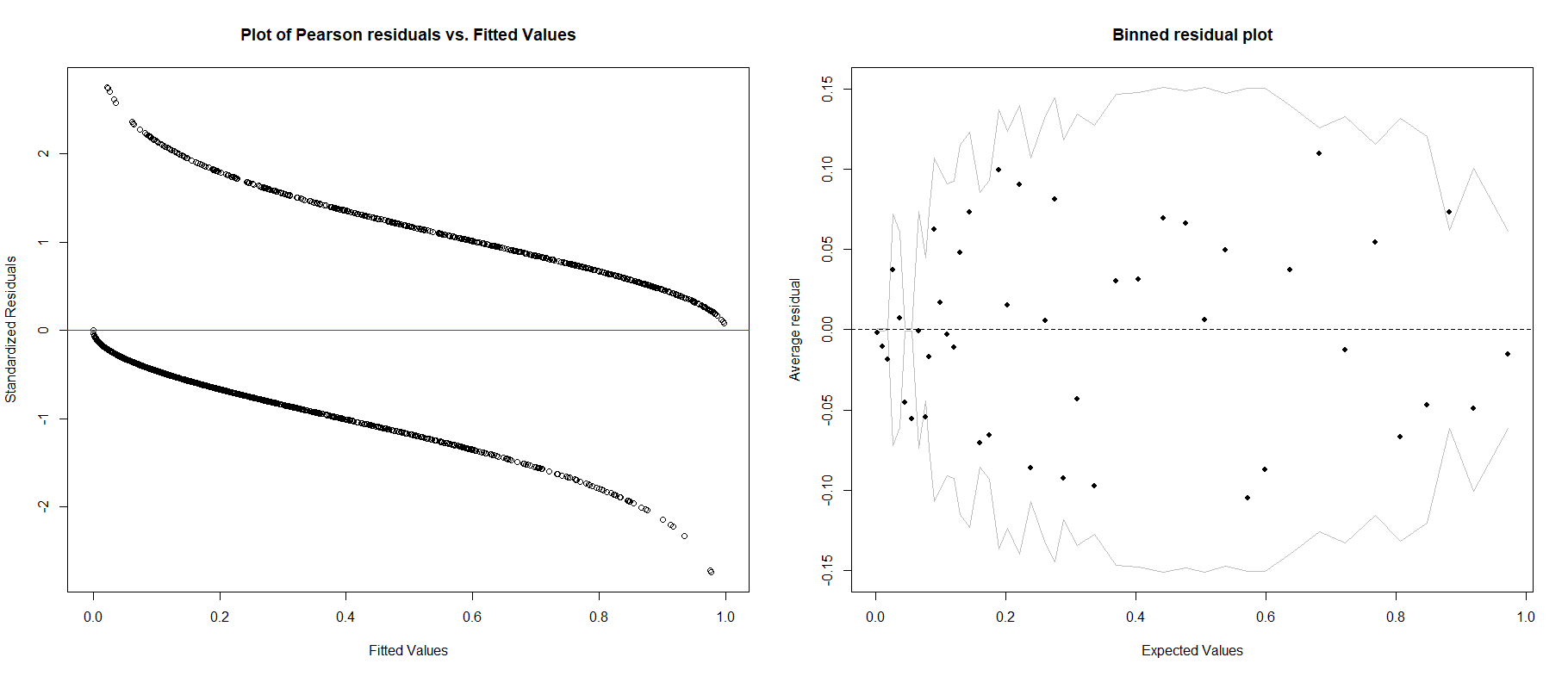
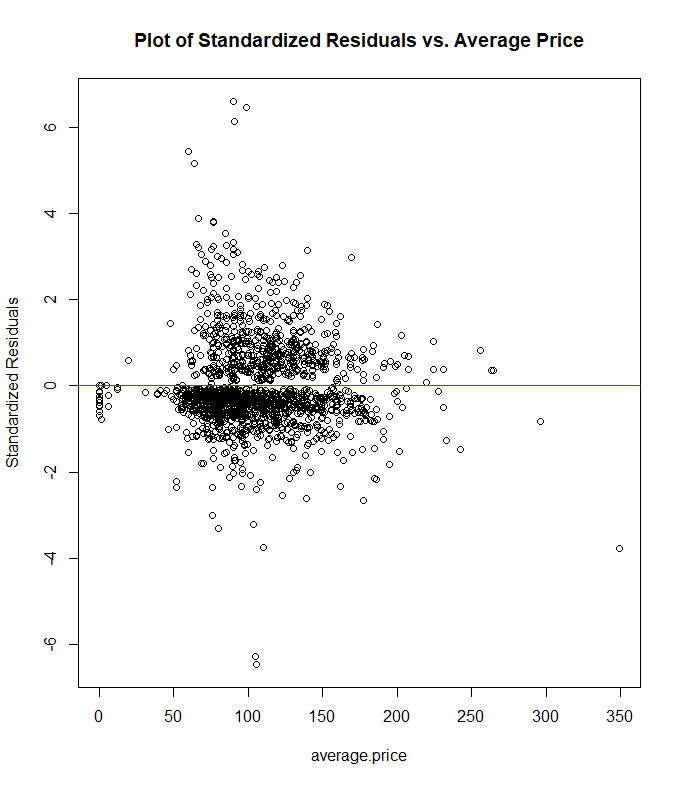
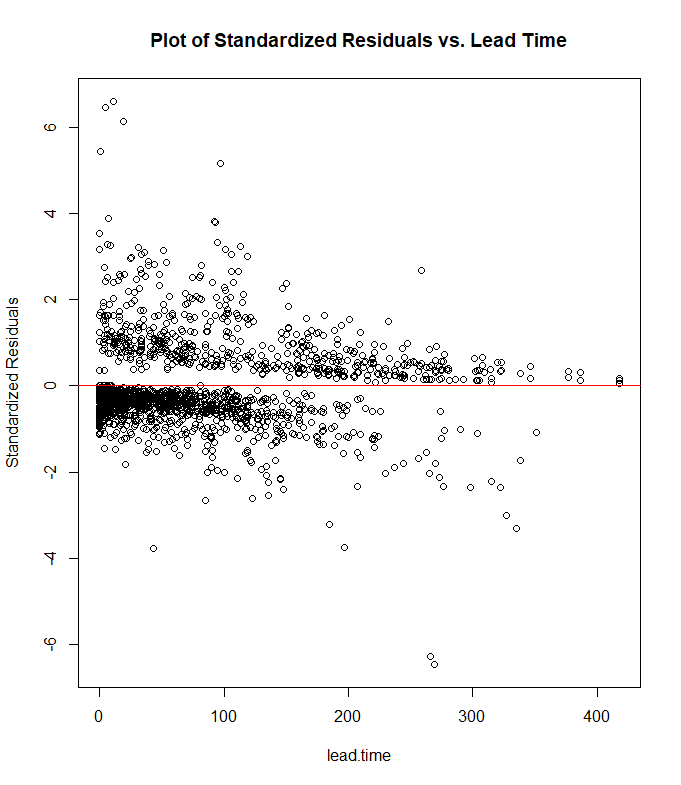
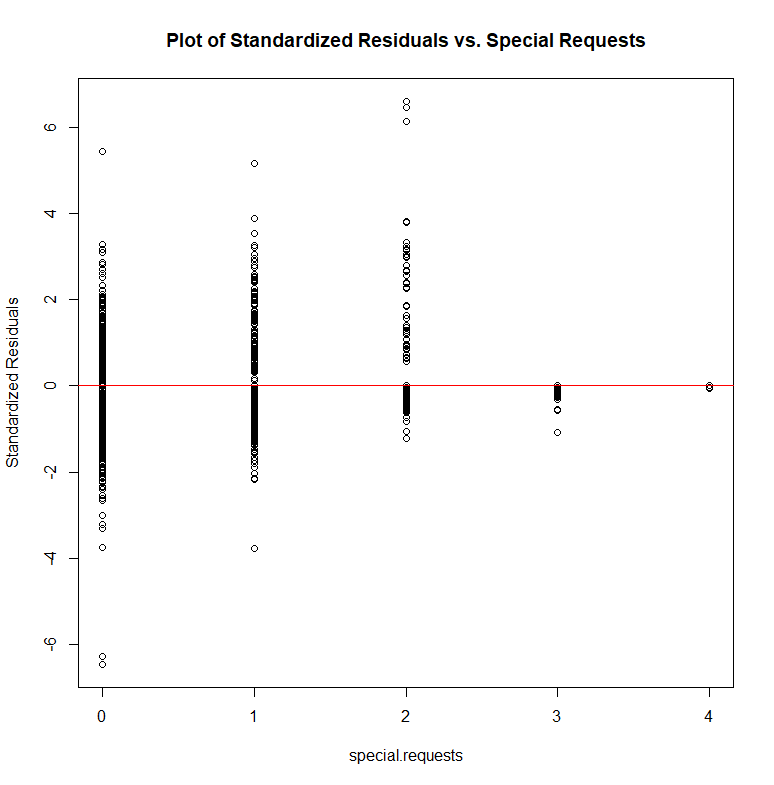
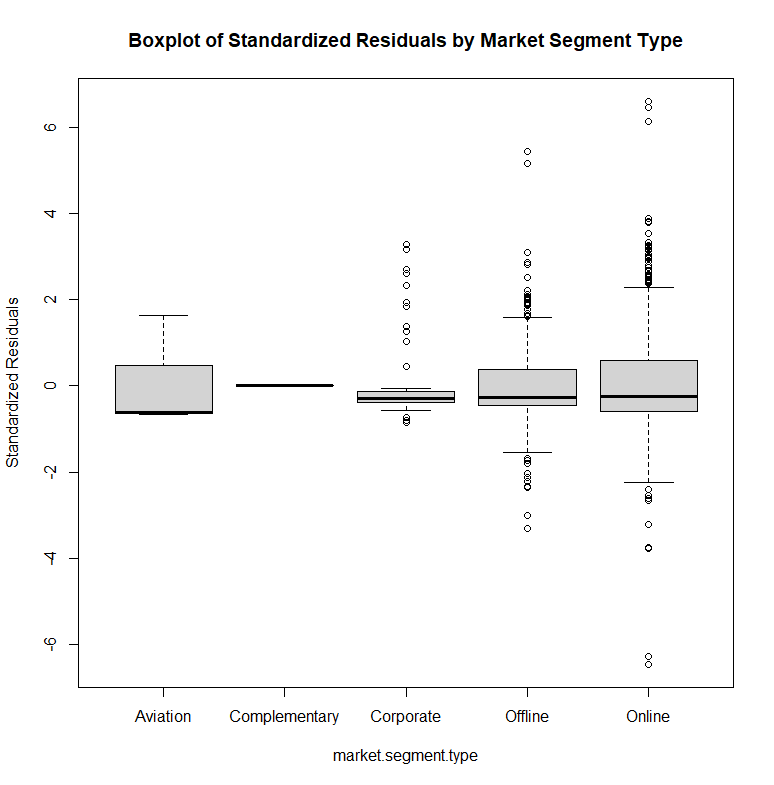
****

Figure 11: Plot of Pearson residuals vs. Fitted Values and binned residual plot

****



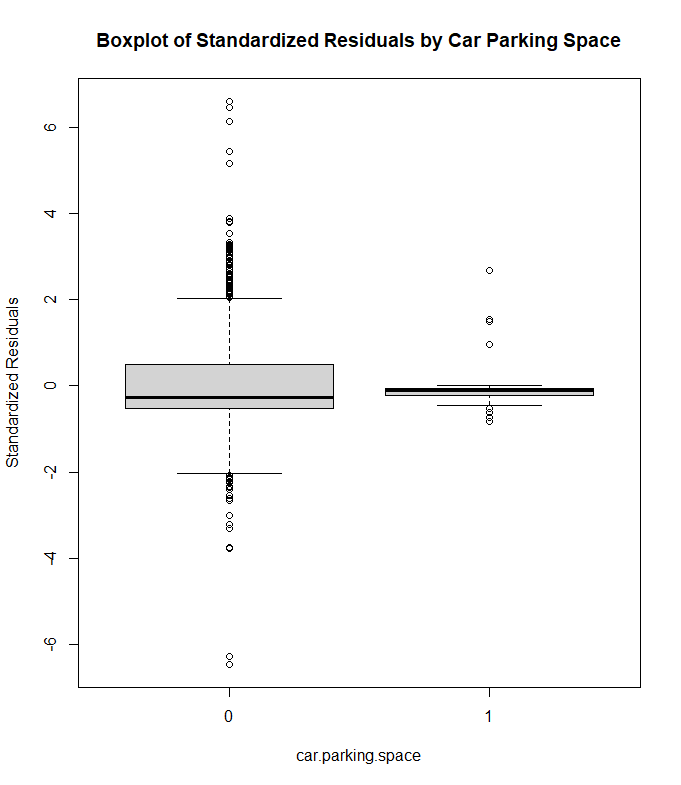


Figure 12: Pearson Standardized Residuals for each Predictor Variable

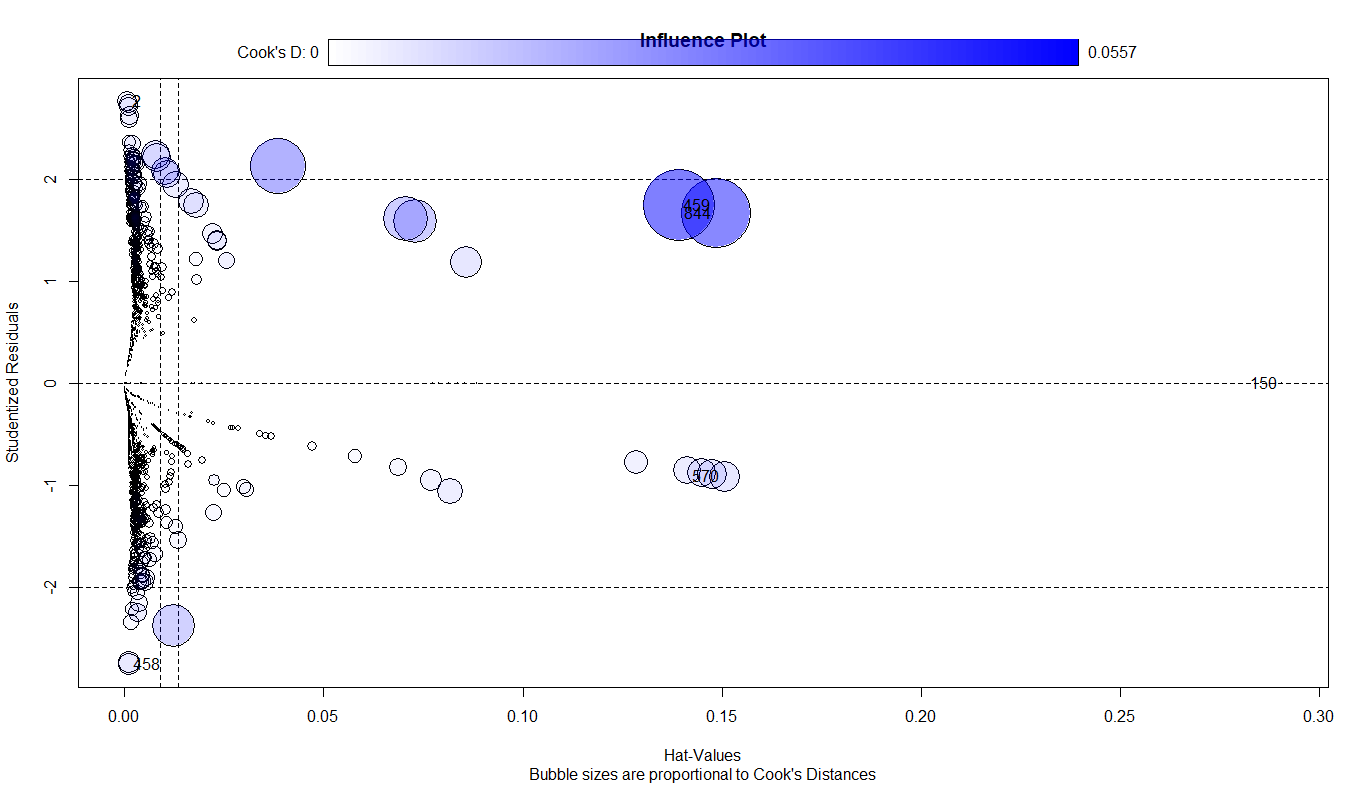
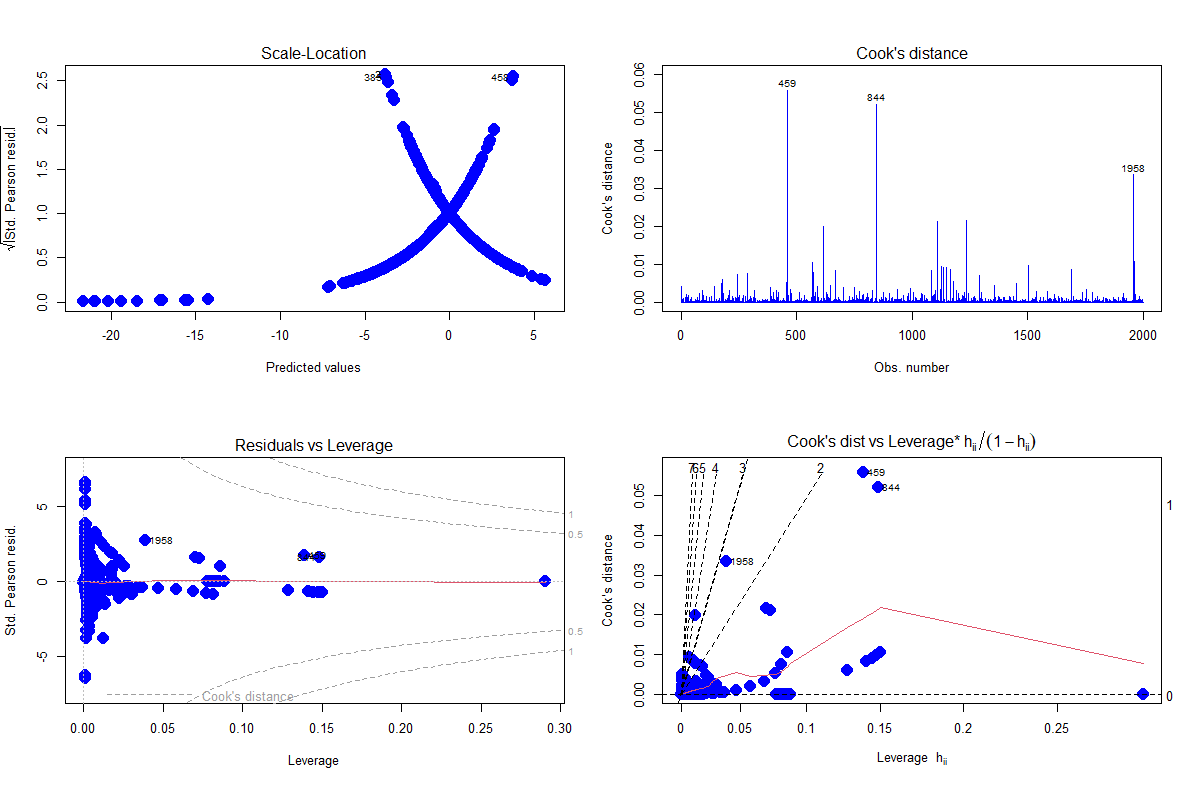


Figure 13: Analysis of Leverage Points and Outliers in Logistic Regression Model

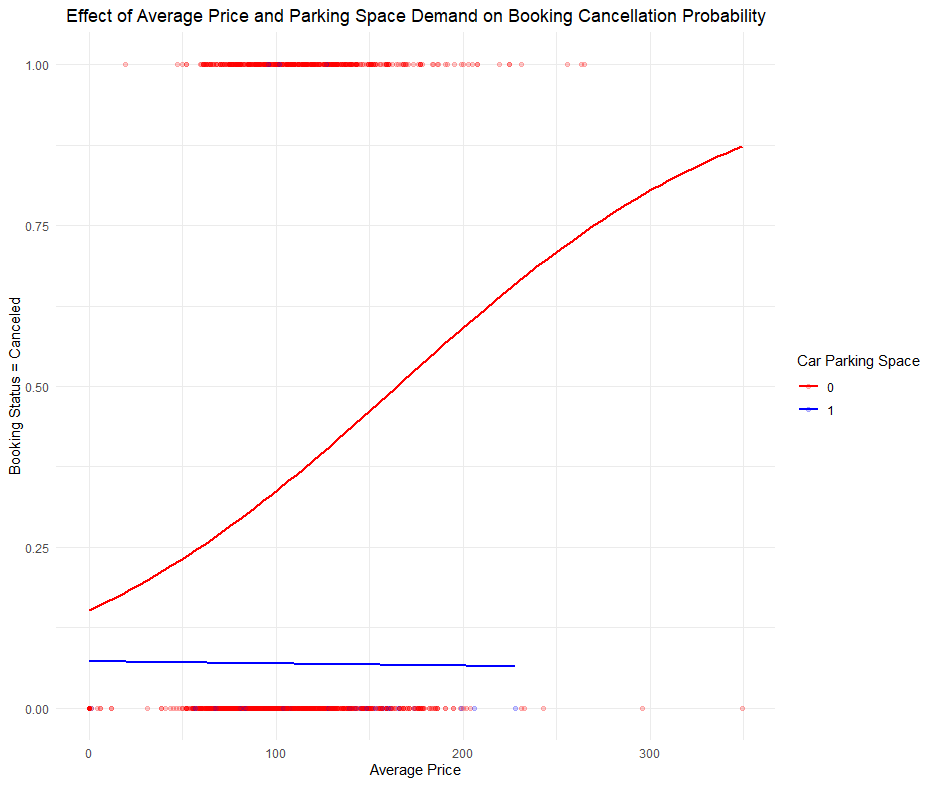


Figure 19: Effect of Average Price and Parking Space Demand on Booking Cancellation Probability

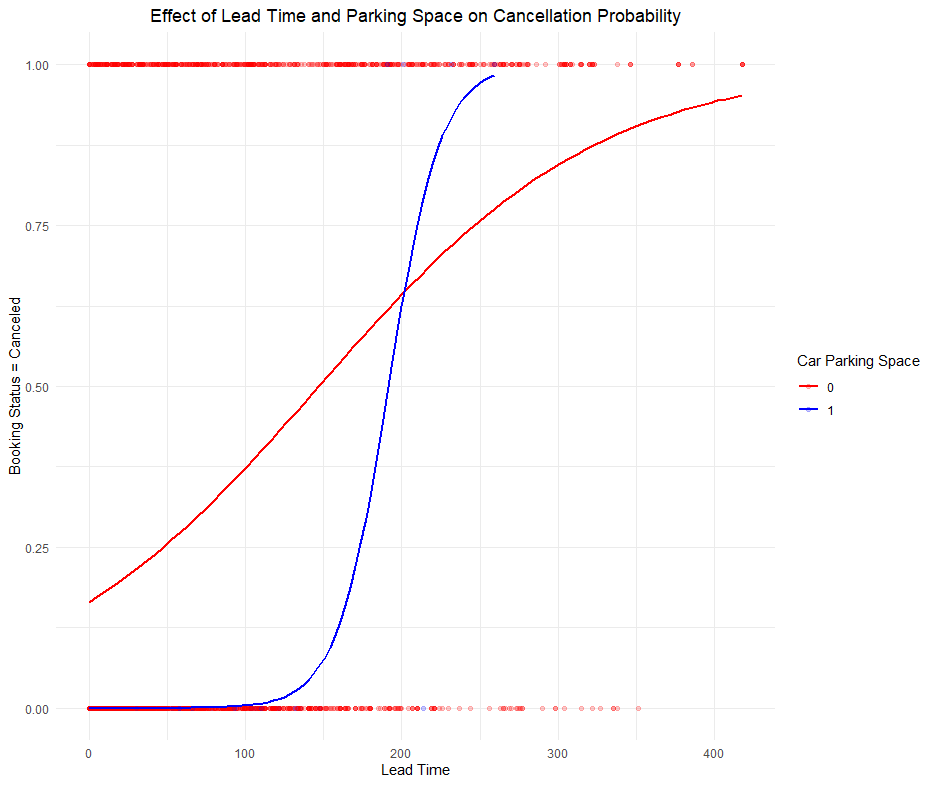


Figure 20: Effect of Lead Time and Parking Space on Cancellation Probability

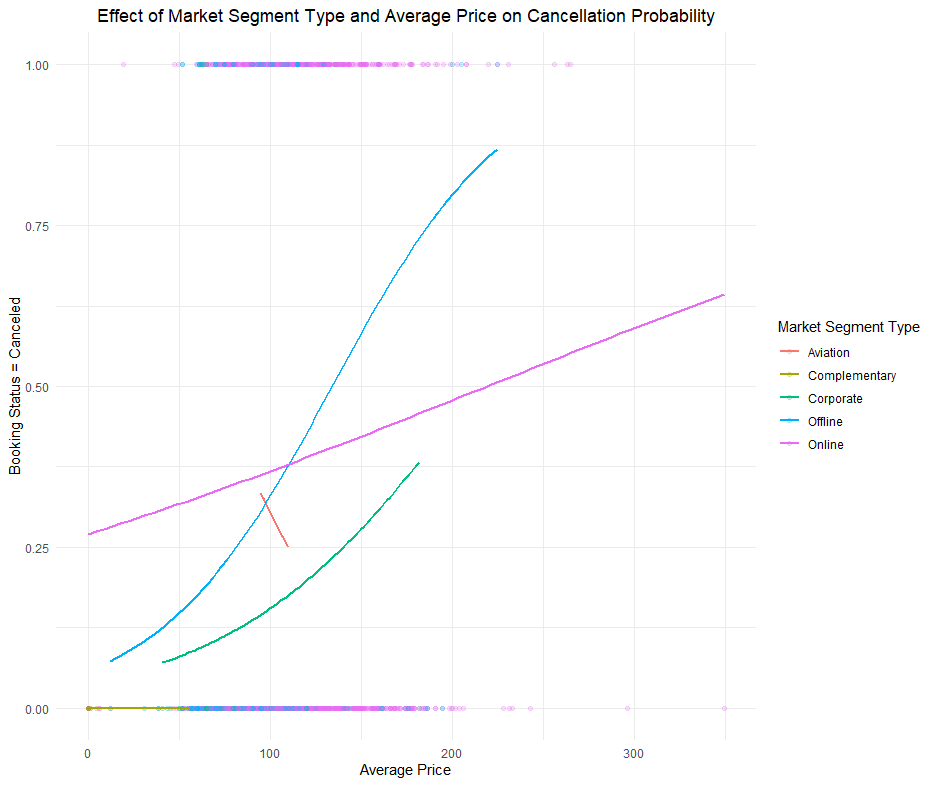


Figure 21: Effect of Market Segment Type and Average Price on Cancellation Probability

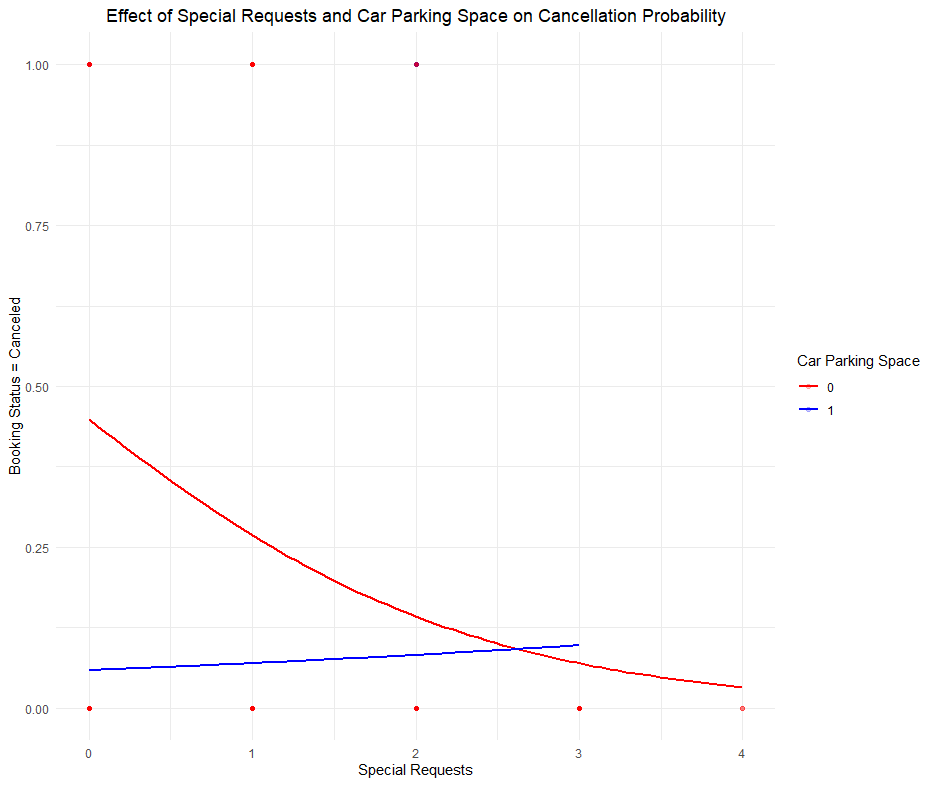


Figure 22: Effect of Special Requests and Car Parking Space on Cancellation Probability