STAT 4630 Final Report

Group 34

**Justin Chuderewicz** - jlc2bgd

**Mitchell Flattum** - mef8ueq

**Andrew Hunter** - abh8sxf

**William Peritz** - wgp3aq

**Section 1: Executive Summary for the Regression Question**

Our analysis was focused on answering the question: which factors influence the number of nights that a person will stay in hotels? This question is worth exploring because the hotel industry, and hospitality industry in general, is one of many that could benefit from using data to better understand consumer habits, which they can use to increase their efficiency and profit. Understanding the factors that influence how many nights a consumer may stay in a hotel can allow the business to adapt their prices, advertising strategy, or implement special targeted promotions to different groups of consumers depending on the consumer’s characteristics, as well as current events on a national economic scale. Relevant stakeholders include:

1. Hotel management and owners, who are directly affected by booking habits and economic performance,
2. Hotel marketing teams, who require insights for effective campaign targeting.
3. Hotel financial analysts, who must understand customer behavior for the purposes of financial planning and forecasting
4. Tourism Boards, who are interested in tourism trends and hotel occupancy for regional planning and development.

Over the course of previous milestones, our group has been successful in determining which factors influence how many nights a customer will stay in a hotel. We were able to discover a couple of key factors that were important. There are several factors that tend to reflect and increase in the number of nights stayed as they increase, including: the average price that the consumer pays per night to stay in the hotel, the number of adults in the booking party, the number of children in the booking party, and the time in between when the booking was made and when the reservation date was set for.

We also found a number of factors that tend to reflect a decrease in nights stayed as they increase. The Consumer Price Index (CPI) for the hotel industry is one such factor. CPI is an economic indicator that essentially measures the amount of money that consumers have to spend in order to buy goods or services from a particular industry. We found that as the CPI for the hotel industry increased and consumers were forced to spend more money, they tended to book shorter stays in hotels, which makes sense intuitively. We also found that an increase in the United States’ average fuel prices led to a decrease in nights stayed at hotels, which also makes sense because consumers are forced to spend more money to travel in these conditions.

Based on our findings, we have a number of recommendations for relevant stakeholders:

1. Implement pricing strategies that are responsive to specific conditions under which customers will stay for shorter or longer periods of time. For example, perhaps implement lower prices during times when customers tend to book shorter stays to persuade them to purchase one or more extra nights.
2. Adjust marketing strategies and appeal to guest demographics who are less likely to book long stays. It could also be beneficial to distribute marketing budgets throughout the year in such a way where there is heavier marketing during times when customers are less likely to book longer stays.
3. Tailor packages to attract and retain specific customer segments. For example, offering a free night with the purchase of several nights could incentivize customers to book longer stays
4. Develop plans to mitigate the impact of economic downturns that decrease the average number of nights stayed

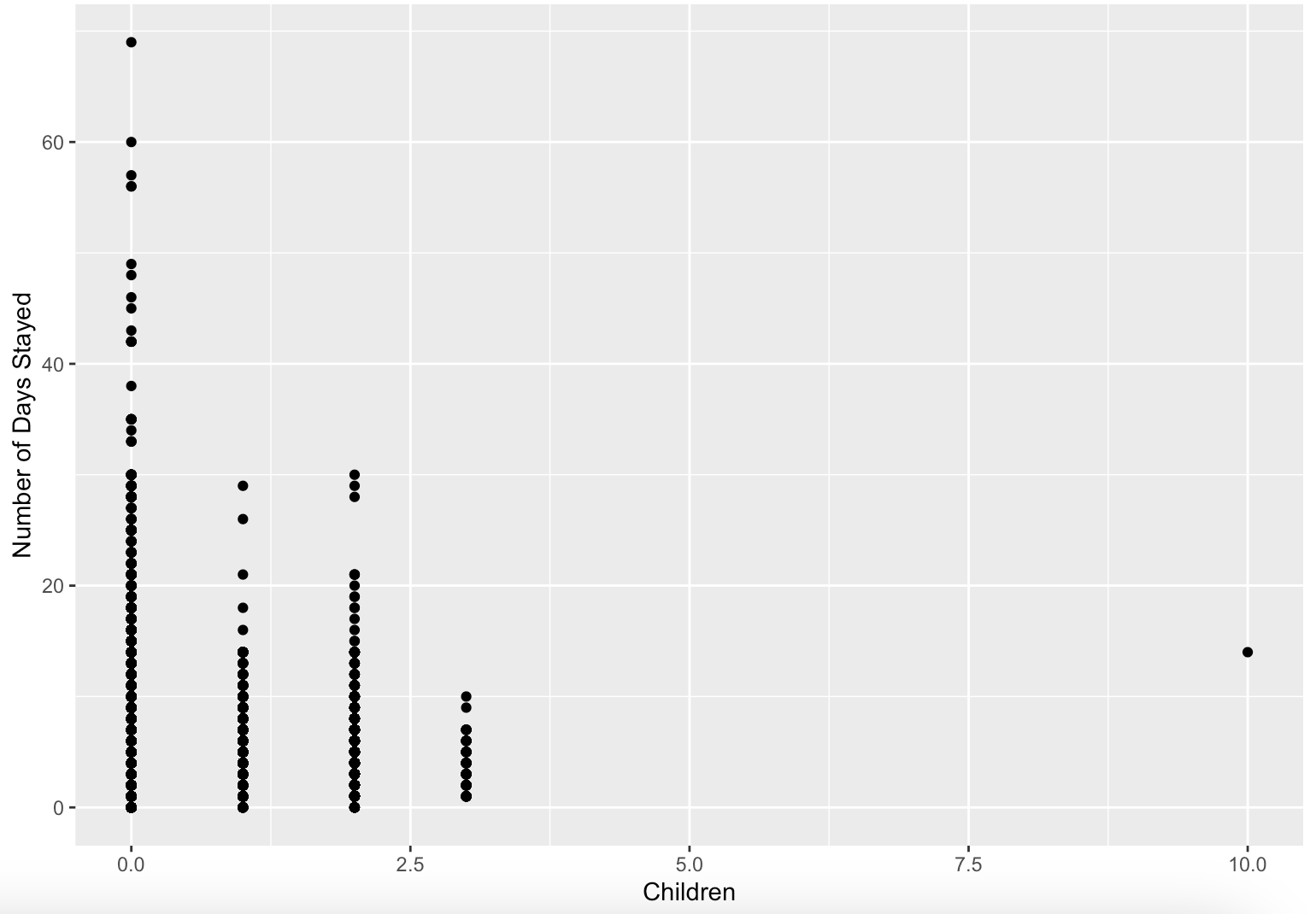
**Section 2: Data and Variable Description for Regression**

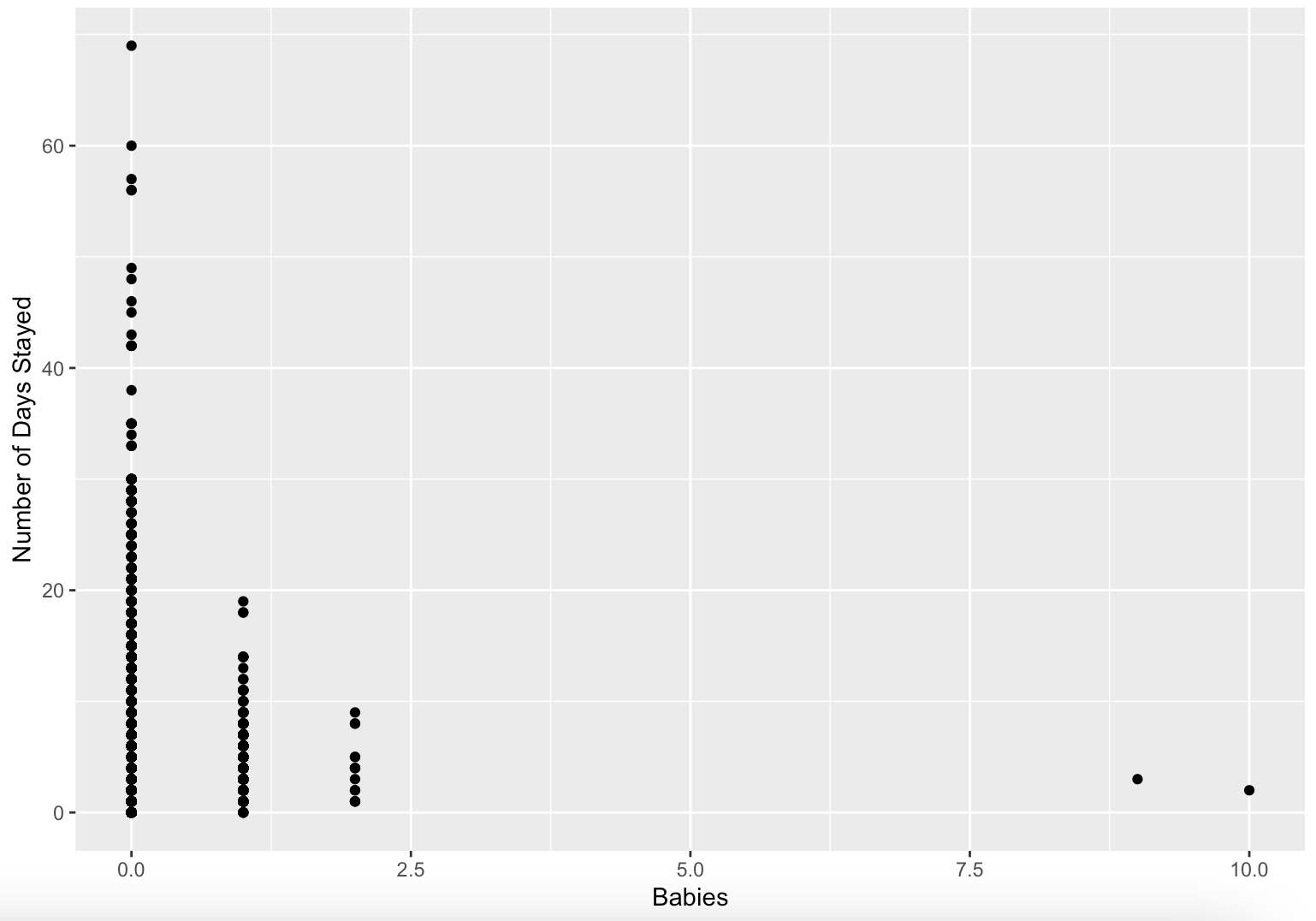
1. Our data is about hotel reservations. Each observation represents one distinct hotel reservation, and contains information about the person booking the hotel, the hotel being booked, and certain economic factors that were occurring at the time of the booking
2. Our dataset was found on Kaggle, and is called “Hotel Booking Demand With Economic Indicators.” It is a modification of the Hotel Booking Demand dataset from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019, and published on the ScienceDirect website. The original dataset contained only hotel information, until Marquis Lardinois modified it by adding columns pertaining to economic data that he obtained from the US Federal Reserve Economic Data site, and uploading it to Kaggle.
3. These are our variables of interest for our regression question:

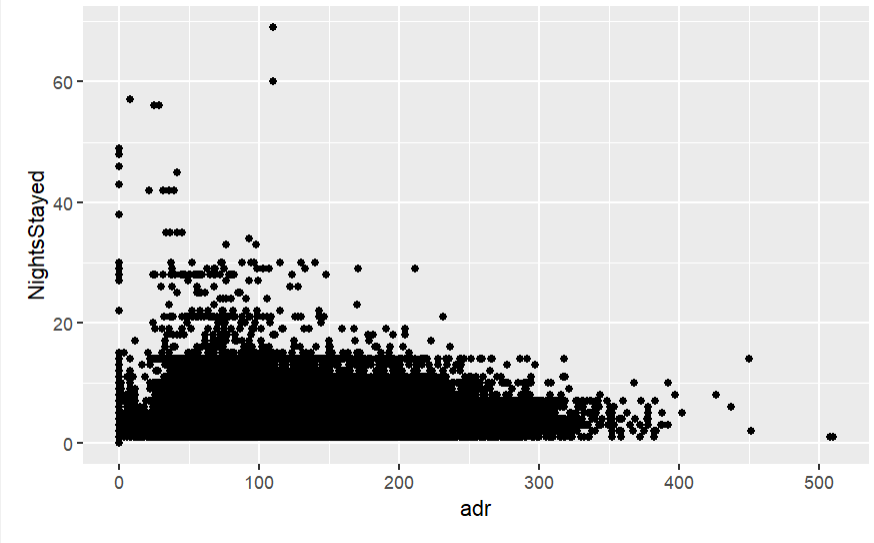
| **Variable Name in Dataset** | **Description** | **Type** |
| --- | --- | --- |
| NightsStayed | The number of nights that the customer stayed in the hotel. We created this variable, it was not in the original dataset. We created it by adding together the values of the variables *StaysInWeekendNights* and *StaysInWeekNights* for each row. | Numeric |
| LogNightsStayed  **(Response Variable)** | A log transformation of NightsStayed that was necessary to create in order to use nights stayed as the response variable | Numeric |
| StaysInWeekendNights | The number of weekend nights that the customer stayed in the hotel during the course of the reservation. Note that we do not include this variable in any analysis, but we are defining it to better contextualize the variable NightsStayed | Numeric |
| StaysInWeekNights | The number of week nights that the customer stayed in the hotel during the course of the reservation. Note that we do not include this variable in any analysis, but we are defining it to better contextualize the variable NightsStayed | Numeric |
| adr | The Hotel’s Average Daily Rate | Numeric |
| adults | The number of adults on the reservation | Numeric |
| children | The number of children on the reservation | Numeric |
| babies | The number of babies on the reservation | Numeric |
| lead\_time | Number of days that elapsed between when the booking was made and when the reservation was set for | Numeric |
| CPI\_HOTELS | The Consumer Price Index (a family of indexes that measure the change in price experienced by urban consumers) for the hotel industry at the time of booking. | Numeric |
| FUEL\_PRICES | US fuel prices at the time of booking | Numeric |
| DIS\_INC | The disposable income per capita in the US at the time of booking | Numeric |

**Section 3: Regression Question**

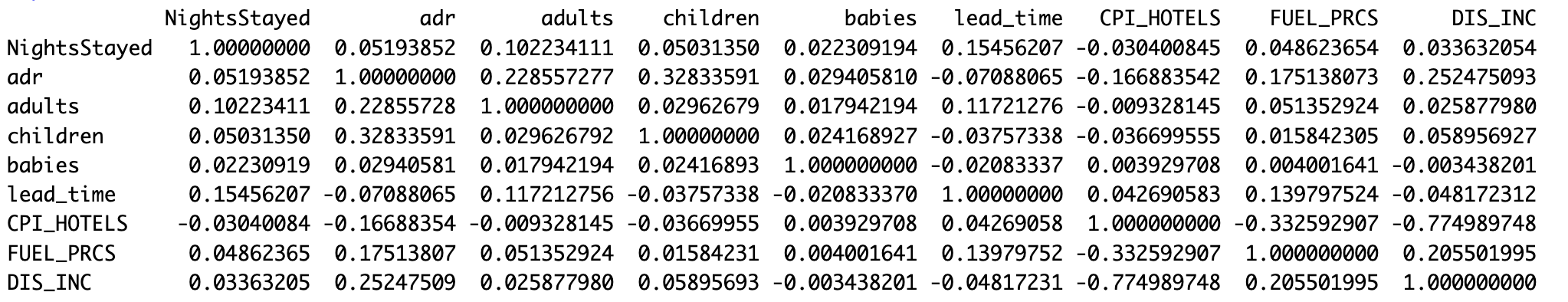
3.1 - Exploratory Data Analysis

Figure 1

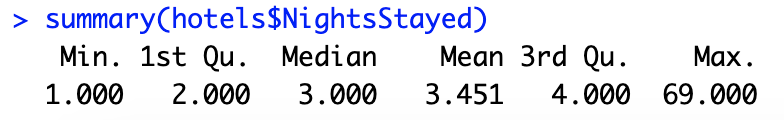
Figure 2

Figure 3

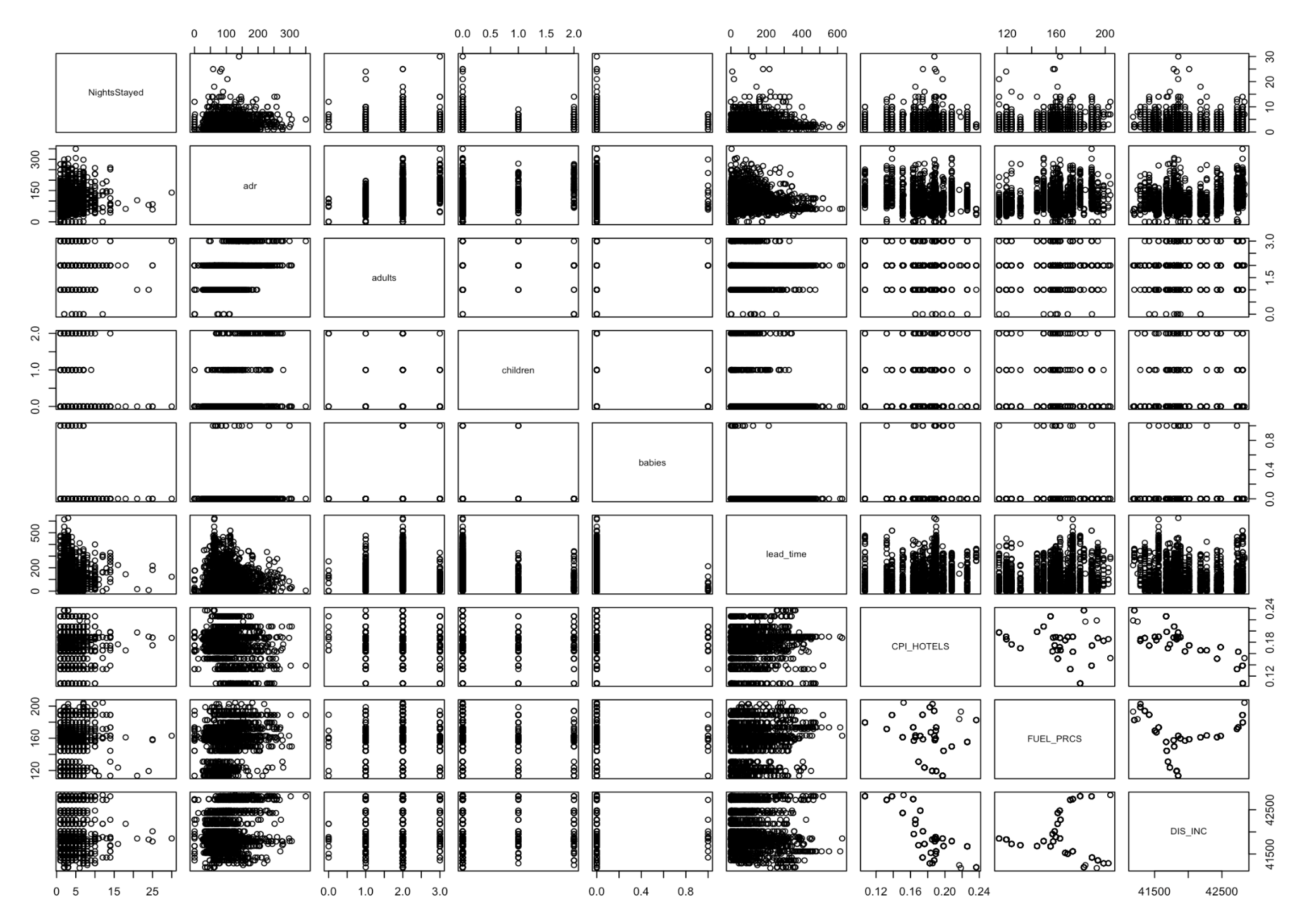
Correlation Table of all relevant variables (Figure 4):



Summary of response variable NightsStayed (Figure 5):



Scatterplot Matrix (Figure 6):



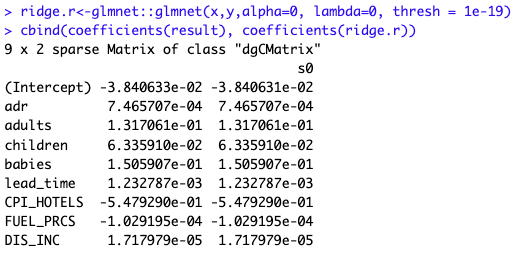
1. Our EDA for our regression questions revealed that a lot of the variables of interest did not appear to have a strong relationship with our response variable. The relationships of note that were found via scatterplot were the relationships of the children, babies, Lead time, and ADR.
2. For both Children and Babies, we can see that customers with fewer children/babies tend to book hotels for more nights per reservation than customers with more children/babies. We also see something interesting- there is a slight downward trend between Lead Time and Nights Stayed. In the context of our data, this means that customers who book reservations further out are more likely to book shorter trips than those who book closer to their reservation date. This is surprising, because we would have expected people who are taking longer trips to plan further ahead and book their reservations earlier.
3. Although these relationships were of note, they aren't the most clear or strong predictors and will definitely require further investigation in future sections.
4. We can also see the summary of our response variable. The median is 3 nights stayed and the mean is 3.451, which is not surprising. The maximum value definitely appears to be an outlier, and we will watch out for that in our future analysis and make sure it does not cause any issues.

3.2 - Shrinkage Methods

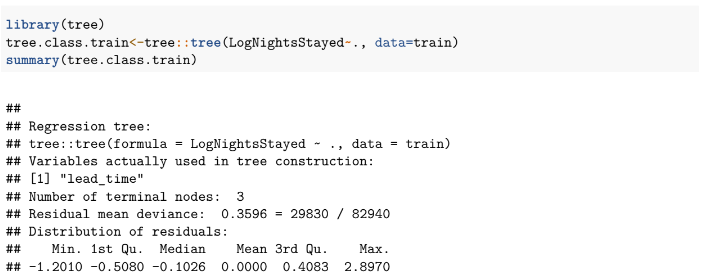
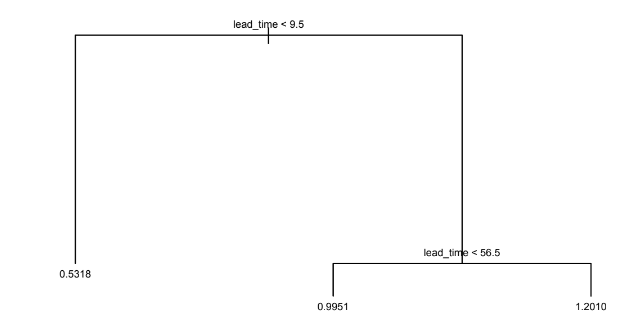
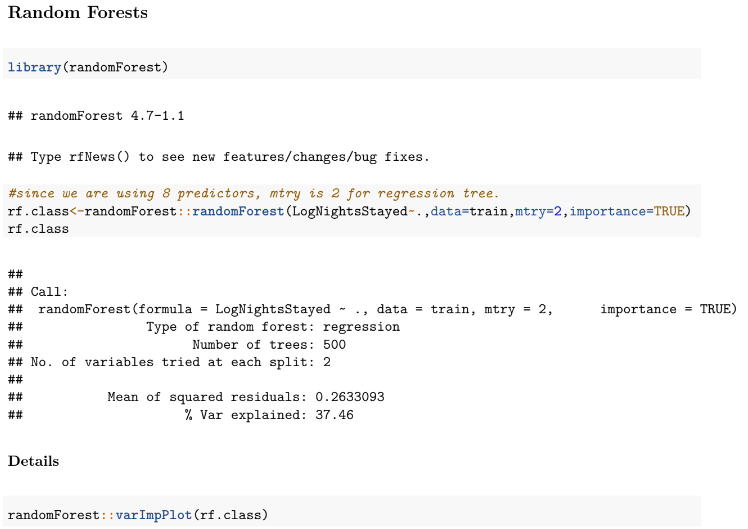
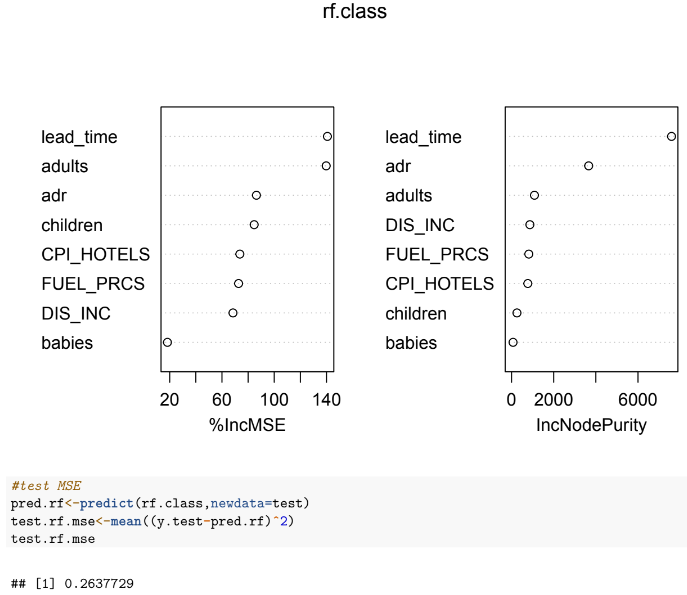
1. We included the following predictors:
   1. ADR
   2. Adults
   3. Children
   4. Babies
   5. Lead\_time
   6. CPI\_HOTELS
   7. FUEL\_PRICES
   8. DIS\_INC

For all of these predictors, we believed that they logically could impact the length of a stay at a hotel. ADR, CPI\_HOTELS, FUEL\_PRICES, and DIS\_INC all measure certain aspects of the economy / hotel industry, so they may impact vacation or travel plans for the average consumer. Especially ADR and CPI\_HOTELS, which report economic data about hotels specifically. Adults, Children, and Babies could be good predictors as well, because different types of families/individuals could tend to take different types of vacations/trips, which would lead to a different number of nights stayed in a hotel. Lead\_time could indicate how much planning went into the trip, so perhaps a shorter lead time results in a less planned-out trip, and could be shorter on average.

1. We choose a value of 1e-19 for our threshold. This was the largest threshold where the estimated coefficients had all the same values, so we believed it to be a good choice:



3.3 - Regression Trees

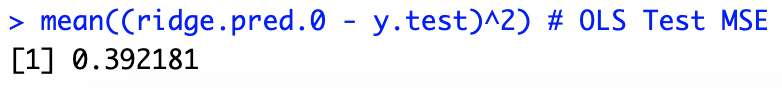
1. In our analysis, we found that the tree built by recursive binary splitting and the tree built by pruning turned out to be the exact same. Since our pruned regression tree shows the same output, plot, and test MSE as our recursive binary splitting tree, it does not matter which we present, so we will present relevant output from our tree built with recursive binary splitting.
2. Output from the summary() function:****
3. We can see here that our tree has 3 terminal nodes
4. Graphical representation of the tree:****
5. Output from the varImpPlot() functions with random forests:****

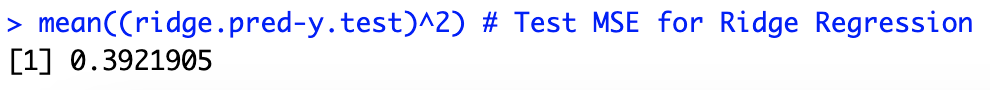
3.4 - Summary of Findings

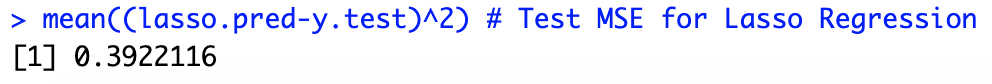
1. A table comparing Test MSE of various methods

| **Method Used** | **Test MSE** |
| --- | --- |
| Linear Regression | 0.392181 |
| Ridge Regression | 0.3921905 |
| Lasso Regression | 0.3922116 |
| Regression Tree (Recursive Binary Splitting) | 0.3586182 |
| Random Forests | 0.2637729 |

Relevant R output:

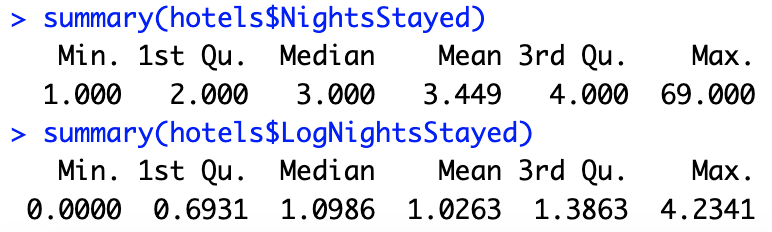






****

1. To understand if our test MSE indicates that our models work well or not, we need to remember that our response variable has been log-transformed. This results in the values for our response variable being lower than they were before. The following piece of R output illustrates this idea:



Our test MSE values range from 0.2637729 to 0.3922116. In the context of our transformed response variable, these test MSE values are rather high, but not terrible. The test MSEs are not small in relation to the data, but not too large. However, based on the scale of the response variable, we can see that Random Forests perform much better than all of the other methods.

1. Our models have helped us answer our question of interest by showing us which variables are most important in determining how many nights a customer will stay in a hotel.
   1. Shrinkage Methods
      1. Our ridge and lasso regression models showed us that the most important factors in predicting nights stayed are ADR, Adults, Children, and lead\_time
      2. We expect a 0.00075 increase in LogNightsStayed for a one unit increase in ADR, holding all other variables constant
      3. We expect a 0.1317 increase in LogNightsStayed for a one unit increase in Adults, holding all other variables constant
      4. We expect a 0.0634 increase in LogNightsStayed for a one unit increase in Children, holding all other variables constant
      5. We expect a 0.00123 increase in LogNightsStayed for a one unit increase in lead\_time, holding all other variables constant
   2. Regression tree
      1. Based on our regression tree, it seems like lead time is the most important predictor in determining the length of days stayed at a hotel
      2. If the lead time is less than 9.5 days, then the customer will spend an average of exp(0.5318) = 1.7 nights at the hotel
      3. If the lead time is greater than 9.5 days but less than 56.5 days, then the customer will spend an average of exp(.9951) = 2.7 nights at the hotel
      4. If the lead time is greater than than 56.5, the customer will spend an average of exp(1.201) = 3.3 nights at the hotel
      5. This decision tree helps answer our research question as it reveals that lead time is likely the most important variable in determining the number of nights stayed. This idea is supported by the results of our shrinkage methods as well, which contained lead time as an important predictor. In the shrinkage methods it was not the variable that had the biggest effect on the response with a single unit change holding all other variables constant, but the range of lead\_time is much higher than adults or children for example. This means that there are more opportunities for a one unit increase in lead\_time compared to adults or children in the booking party.
2. For our regression question, Random Forests performed the best out of all of the models. One thing that we found interesting is that OLS, Ridge, Lasso, and Recursive Binary Splitting were all very close to each other in terms of test MSE, with Recursive Binary Splitting being a bit lower than the rest. However, refining the regression tree with Random Forests yielded a significantly lower test MSE. Random Forests was able to identify several variables that played an important role in predicting the response. The variables that it identified were generally consistent with the other methods we used, but as previously mentioned it was able to achieve a much lower test MSE in the process.

3.5 Address Previous Comments

These comments come from our feedback for Milestone 4. We received no comments about our regression question in Milestone 2, and Milestone 3 was exclusively focused on our classification question.

| **Comment** | **How it was addressed** |
| --- | --- |
| In some of the descriptions, there is an acronym called "PMS". I don't think this acronym was ever defined. Please do so in future milestones | This was an acronym used by the original source of our dataset, however it is not important to understanding the data. We have removed the acronym from our report while preserving its meaning in the data description section |
| Section 2 looks to be a copy and paste of your EDA from the previous milestone, which is EDA for the classification. For this milestone, you need the EDA for the regression question. There is no EDA for the regression question as such. | We corrected this oversight and added EDA specifically for our regression question |
| Remember that your response variable is log transformed, so when interpreting the predicted response you need to convert back to the original units by exponentiating. Eg: when lead time is less than 9.5 days, the predicted stay is exp(0.5318) = 1.7 days. Please edit for future milestones. | We addressed this comment by revising our language to reflect the fact that our response variable was log transformed. Throughout section 3 of this report, whenever we refer to predictors relating to the response variable, we make sure to account for the fact that the response is log transformed. |

**Section 4: Executive Summary for Classification Question**

The question of interest that we will be answering using classification methods is: Which factors influence hotel reservation cancellations?

Motivation:

The focus of this analysis is to understand the factors influencing hotel reservation cancellations. Various aspects, such as demographic data, booking channels, time-related factors (ex. season and booking lead time), and broad economic indicators, can be crucial in understanding what leads to hotel cancellations. For example, certain demographics might have a higher/lower tendency to cancel, or bookings made through specific channels might be more subject to cancellation. Moreover, economic conditions and seasonal factors could also significantly impact cancellation rates. Our exploration is significant for its potential to enhance economic and marketing strategies, customer segmentation, operational efficiency, and adaptability to macroeconomic conditions. Answering this research question could allow hotels to adjust their prices for certain groups of customers if the hotel thinks that they would be more likely to cancel their reservation, or focus their advertising more on the types of customers that they are confident will not cancel. Hotels could also make adjustments to pricing and advertising based on economic factors, knowing that more customers are more likely to cancel during certain economic times in the United States.

Insights from Previous Analysis:

Our previous analyses have involved both regression and classification models (Ridge and Lasso, Regression Trees, Random Forests, etc.) to examine and categorize guest behavior. Our models included factors like guest demographics, booking methods, hotel’s average daily rate, and other economic indicators. The goal we have is to identify significant predictors for hotel cancellations, offering a comprehensive view of guest booking patterns. In terms of this classification question, we have found that an increase in the time between when the booking was made and when the reservation was set for leads to an increase in likelihood of cancellation. In other words, the further in advance a person makes their reservation, the more likely they are to cancel. We also found that deposit type was important in predicting cancellations, with nonrefundable deposits being more likely to lead to cancellations, perhaps counterintuitively.

Stakeholders that are affected:

1. Hotel Management and Owners: Directly affected by booking habits and economic performance.
2. Hotel Marketing Teams: Require insights for effective campaign targeting.
3. Hotel Financial Analysts: In the hospitality industry, understanding customer behavior is crucial for financial planning and forecasting.
4. Tourism Boards: Interested in tourism trends and hotel occupancy for regional planning and development.

Recommendations for Relevant Stakeholders:

Relevant Stakeholders can use the information compiled from our models to:

* Implement pricing strategies that are responsive to specific, relevant predictors of cancellation behaviors.
* Develop marketing strategies focused on guest demographics with higher / lower cancellation rates
* Adjust marketing strategies and appeal to guest demographics who are less likely to cancel reservations
* Tailor packages to attract and retain specific customer segments
* Modify cancellation policies during high-risk periods to minimize revenue loss
* Develop plans to mitigate the impact of economic downturns

Conclusion:

This analysis holds significant value for stakeholders in the hospitality industry, offering insights that can enhance economic performance, marketing strategies, and operational efficiency. By understanding the factors that influence reservation cancellations, hotels can tailor their strategies to better meet market/individual demands and adapt to changing economic conditions. This could result in more effective pricing models, targeted marketing campaigns, and improved overall financial health of the hospitality entities involved.

**Section 5: Variable Description for Classification Question**

The data used in our classification question is about hotel reservations. Each observation represents one distinct hotel reservation, and contains information about the person booking the hotel, the hotel being booked, and certain economic factors that were occurring at the time of the booking. A table consisting of all the variables included in our classification question is listed below:

The response variable is ***IsCanceled***, which is a binary variable indicating whether or not the customer canceled their hotel booking after they placed the reservation. The variables that were included in Logistic Regression and the Classification trees were:

| **Variable Name in Dataset** | **Description** | **Type** |
| --- | --- | --- |
| Is\_canceled  **(Response variable)** | Value indicating if the booking was canceled (1) or not (0) | Categorical, binary |
| adr | Hotel’s Average Daily Rate | Numeric |
| arrival\_date\_week\_number | Week number of the arrival date | Numeric |
| booking\_changes | Number of changes/amendments made to the booking from the moment the booking was made until the moment of check-in or cancellation | Numeric |
| days\_in\_waiting\_list | Number of days the booking was in the waiting list before it was confirmed to the customer | Numeric |
| is\_repeated\_guest | Value indicating if the booking name was from a repeated guest (1) or not (0) | Categorical, Binary |
| lead\_time | Number of days that elapsed between when the booking was made and the arrival date | Numeric |
| previous\_bookings\_not\_canceled | Number of previous bookings not canceled by the customer prior to the current booking | Numeric |
| previous\_cancellations | Number of previous bookings that were canceled by the customer prior to the current booking | Numeric |
| CPI\_HOTELS | The Consumer Price Index (a family of indexes that measure the change in price experienced by urban consumers) for hotels at the time of booking. | Numeric |
| FUEL\_PRICES | US fuel prices at the time of booking | Numeric |
| DIS\_INC | disposable income per capita in the US at the time of booking | Numeric |
| Deposit\_type | Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay. | Categorical |
| customer\_type | Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking | Categorical |

**Section 6: Classification Question**

6.1 - Exploratory Data Analysis

Figure 1: Scatterplot Matrix

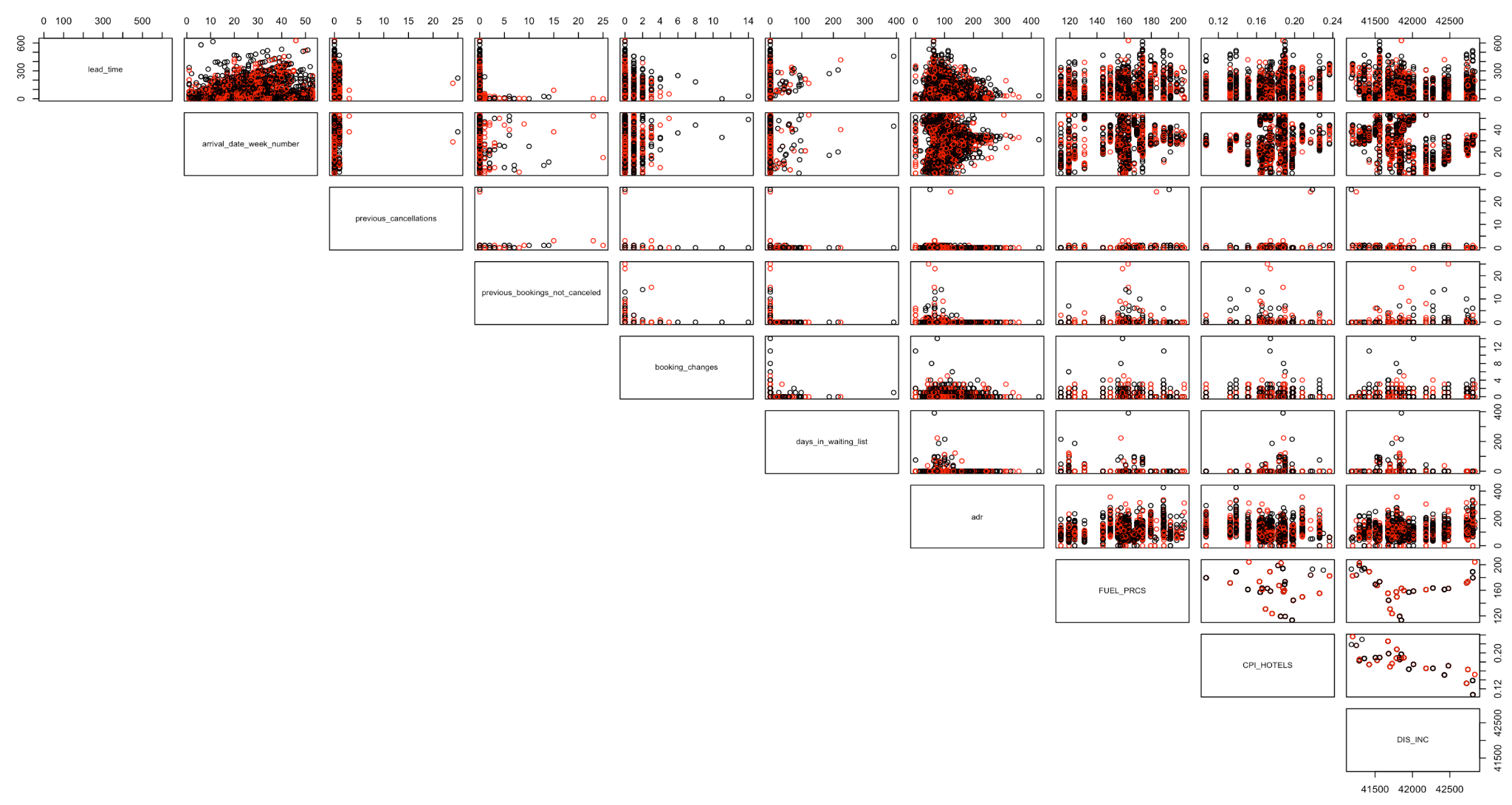
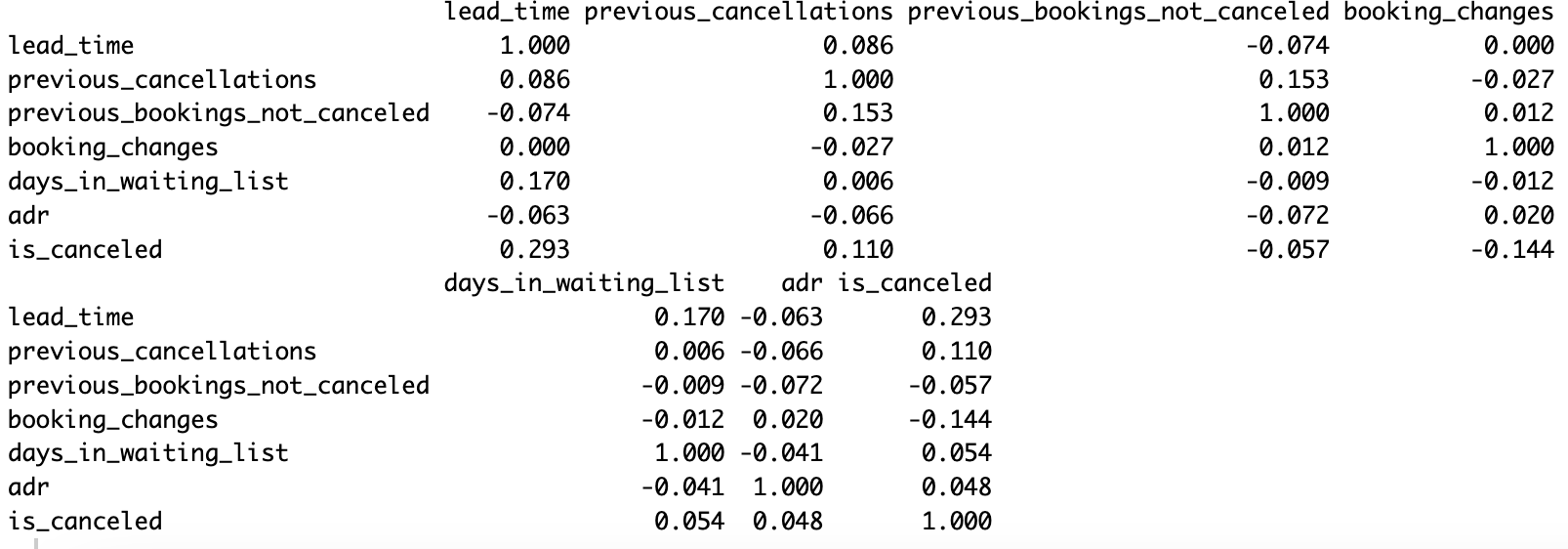


Figure 2: Correlation Chart



Figures 3 and 4: Boxplots of Lead time and Hotel CPI:

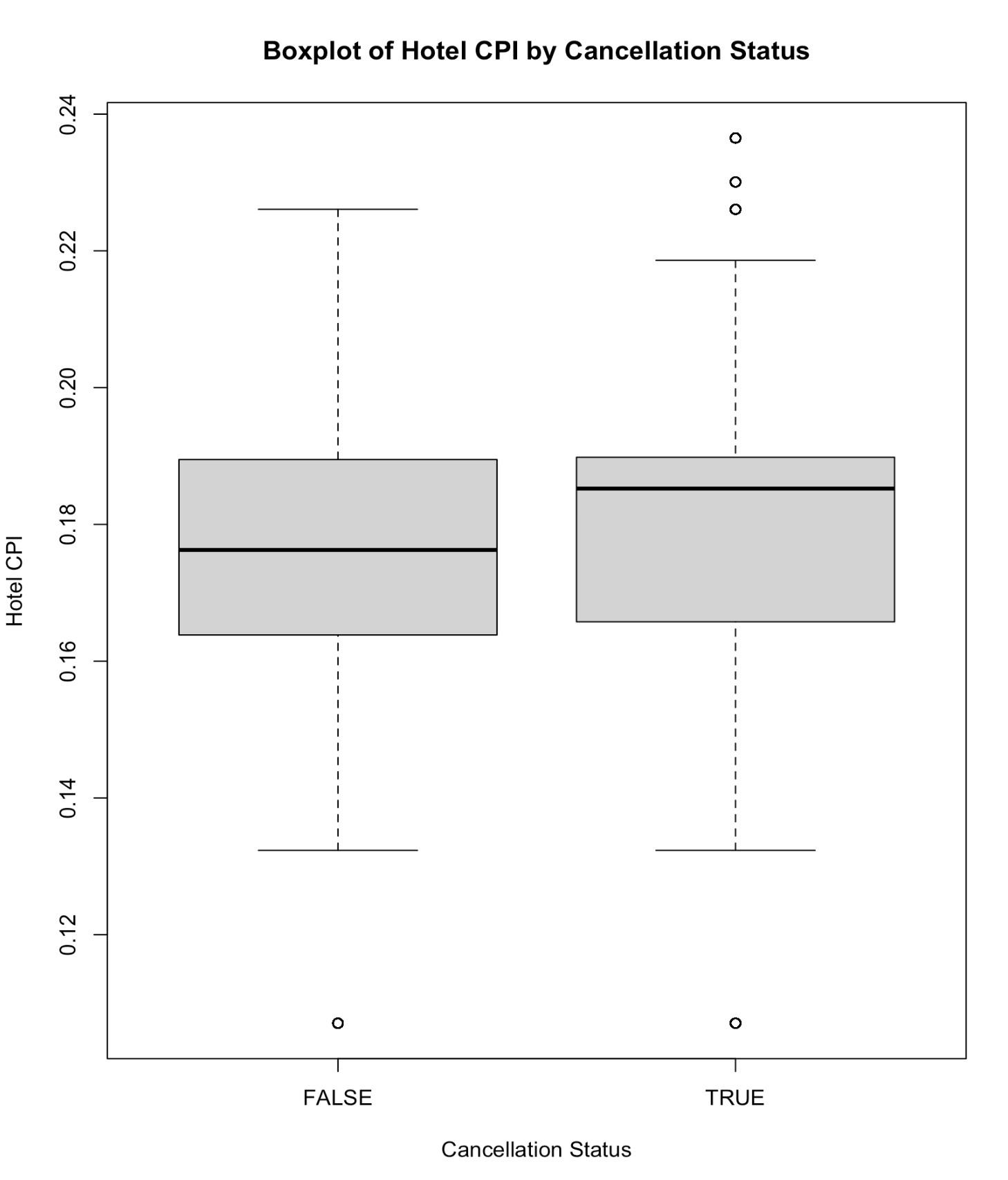
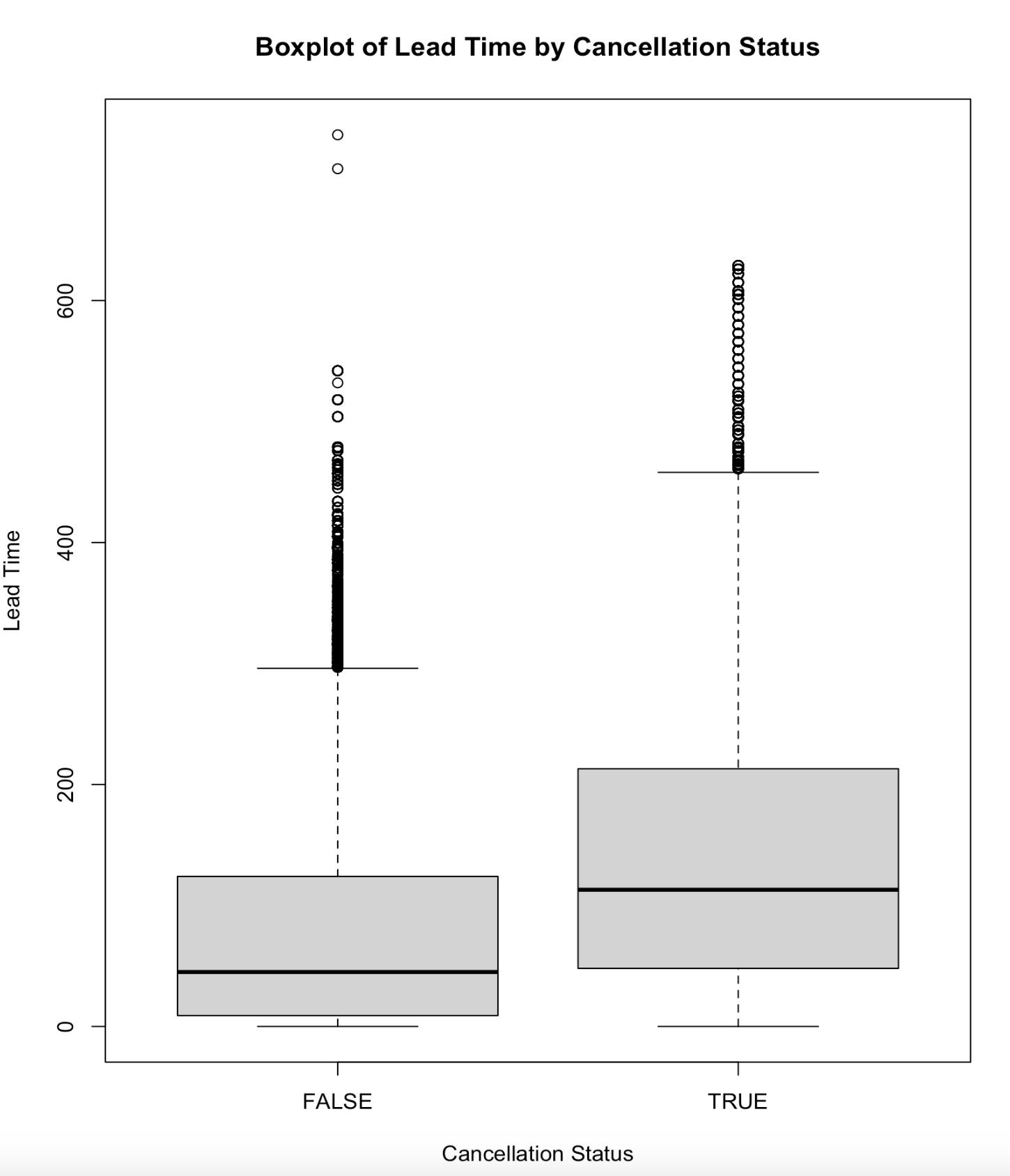


Figure 1 consists of a scatterplot matrix of our preliminary predictors. Our data set contains over 119,000 rows, so this scatterplot matrix was made from a random sample of 2,000 rows to make it easier to interpret. A red dot indicates that the booking was canceled, while a black dot indicates that it was not canceled.

Figure 2 consists of a correlation chart between several of our predictors of interest and “is\_canceled”. This correlation chart shows that Lead Time and Previous Cancellations have the strongest correlation with is\_canceled, however these predictors are still not very strong. Our scatterplot matrix did not reveal very much to us, as there doesn’t seem to be any overtly strong separation between the black and red dots in the scatterplot matrix, although there are some subtle differences.

Figures 3 and 4 highlight the differences between a canceled reservation and one that was not canceled in terms of Lead Time and CPI\_Hotels. From the first one, we can see that the median value for Lead Time among canceled reservations is higher than the median value for Lead Time among non-canceled reservations. Generally, we see that canceled reservations tend to have higher Lead Times than non-canceled reservations. In the context of our data, this means that customers who book hotel reservations further out are more likely to cancel them. This makes sense, because it is natural for peoples’ far-future plans to fall through more often than their near-future plans.

Figure 4 shows us a slightly more subtle, but still present, difference. We can see that the median value for Hotel CPI is higher for canceled reservations than non-canceled ones. In the context of our data, that means that a higher Consumer Price Index for the hotel industry can be associated with an increased number of cancellations. This makes sense, because when the Consumer Price Index rises, it means that consumer prices are rising and the hotels would likely be more expensive.

6.2 - Logistic Regression Model

1. Included predictors:

In our initial model for Logistic Regression, we choose to include the following 10 variables:

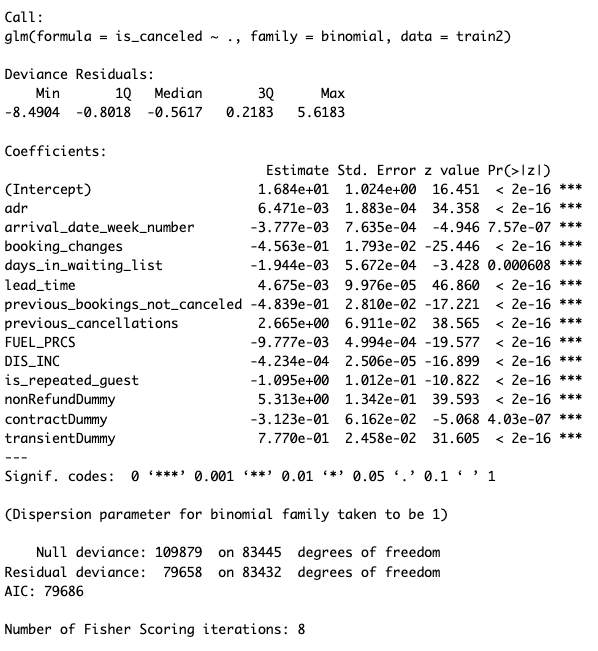
* “adr”,
* "arrival\_date\_week\_number",
* "booking\_changes",
* "days\_in\_waiting\_list",
* "lead\_time",
* "Previous\_bookings\_not\_canceled",
* "previous\_cancellations",
* "CPI\_HOTELS",
* "FUEL\_PRCS",
* "DIS\_INC"

We chose these variables because, based on our intuition and the results of our EDA, we decided that they may have a significant relationship with our response variable. We wanted to select a large amount of predictor variables, because we knew that we would be improving and reducing the model later, and we could then remove any insignificant variables. After testing out our initial Logistic Regression model, we decided to enhance the model by conducting wald tests on all of the variables above. In addition, we also added three more variables into the Logistic model, and confirmed their significance via Likelihood Ratio Tests. These 3 variables are:

* “Is\_repeated\_guest”
* “Deposit\_type”
* “Customer\_type”

These 13 variables were additionally used in building our classification trees.

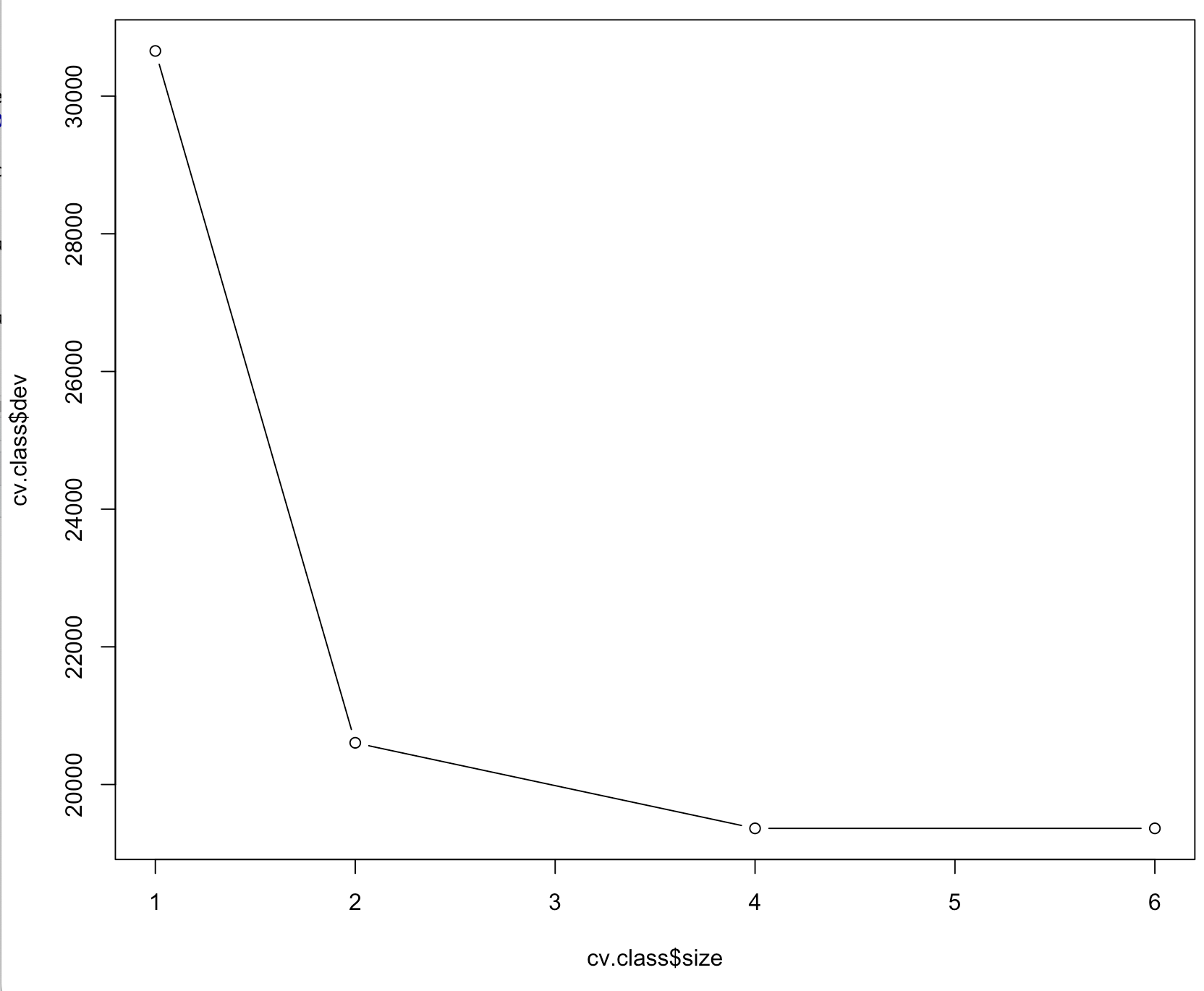
1. Output from the summary() function for logistic regression model:

****

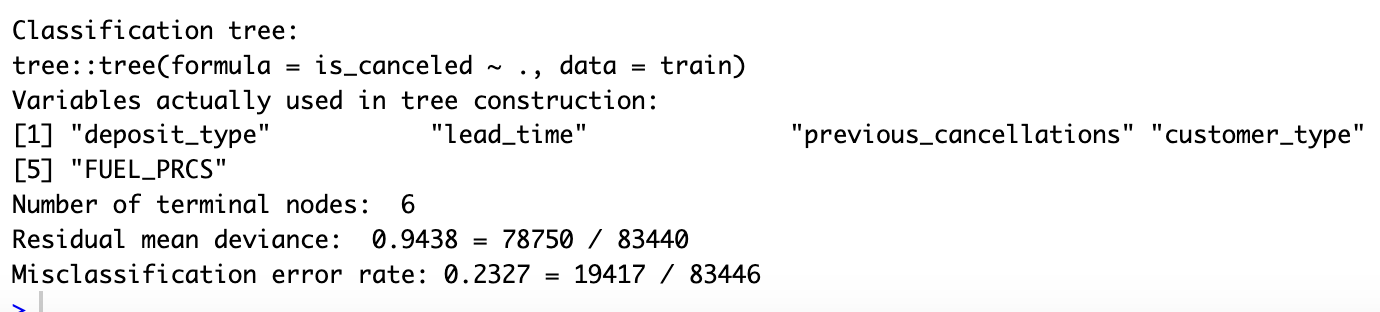
Overall, all of the predictors on our original model had low p-values, indicating that they were significant in predicting hotel cancellations. However, running the Wald test on each predictor in the original logistic regression model informed us that we should remove the CPI\_HOTELS variable. To verify that we had made the right choice, we conducted a likelihood ratio test to compare the model with all possible predictors and the reduced model. The p value of the Likelihood Ratio Test did not provide enough evidence to reject the null hypothesis, meaning the reduced model was preferred. The CPI\_HOTELS variable did not improve the model, therefore this predictor was dropped.

6.3 - Classification Trees:

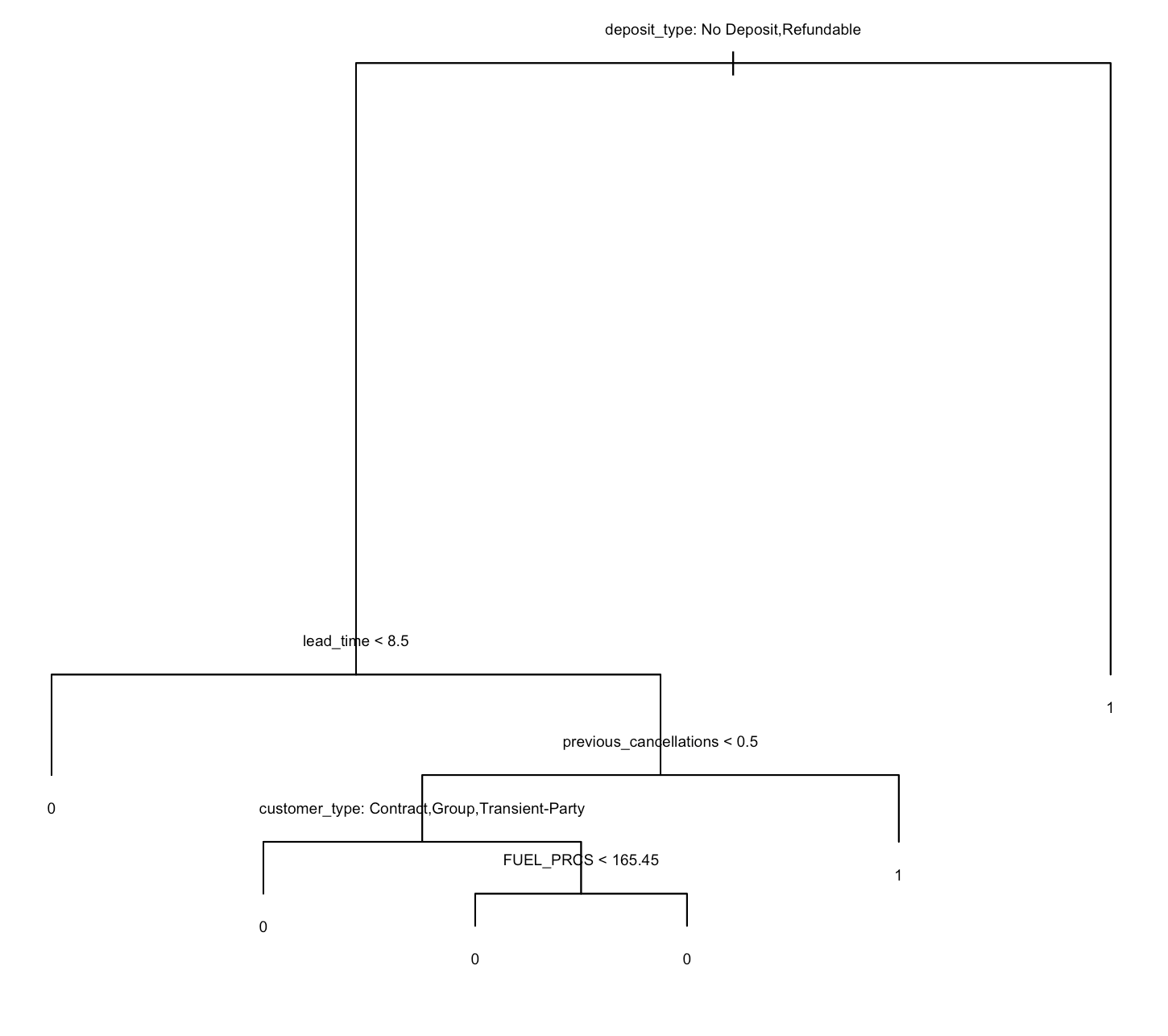
1. We chose to go ahead with the recursive binary splitting tree instead of the pruned tree because when we conducted 10 fold cross validation on our pruned tree output, the ideal number of terminal nodes was found to be 6 on the pruned tree. This can be seen from the plot below that shows 6 terminal nodes having the smallest deviance. Since the recursive binary splitting tree has 6 terminal nodes already, fitting the pruned tree results in the exact same tree as from recursive binary splitting. In addition to this, the residual mean variance and misclassification error rates are the exact same for both trees. Therefore there is no benefit to using the pruned tree instead.



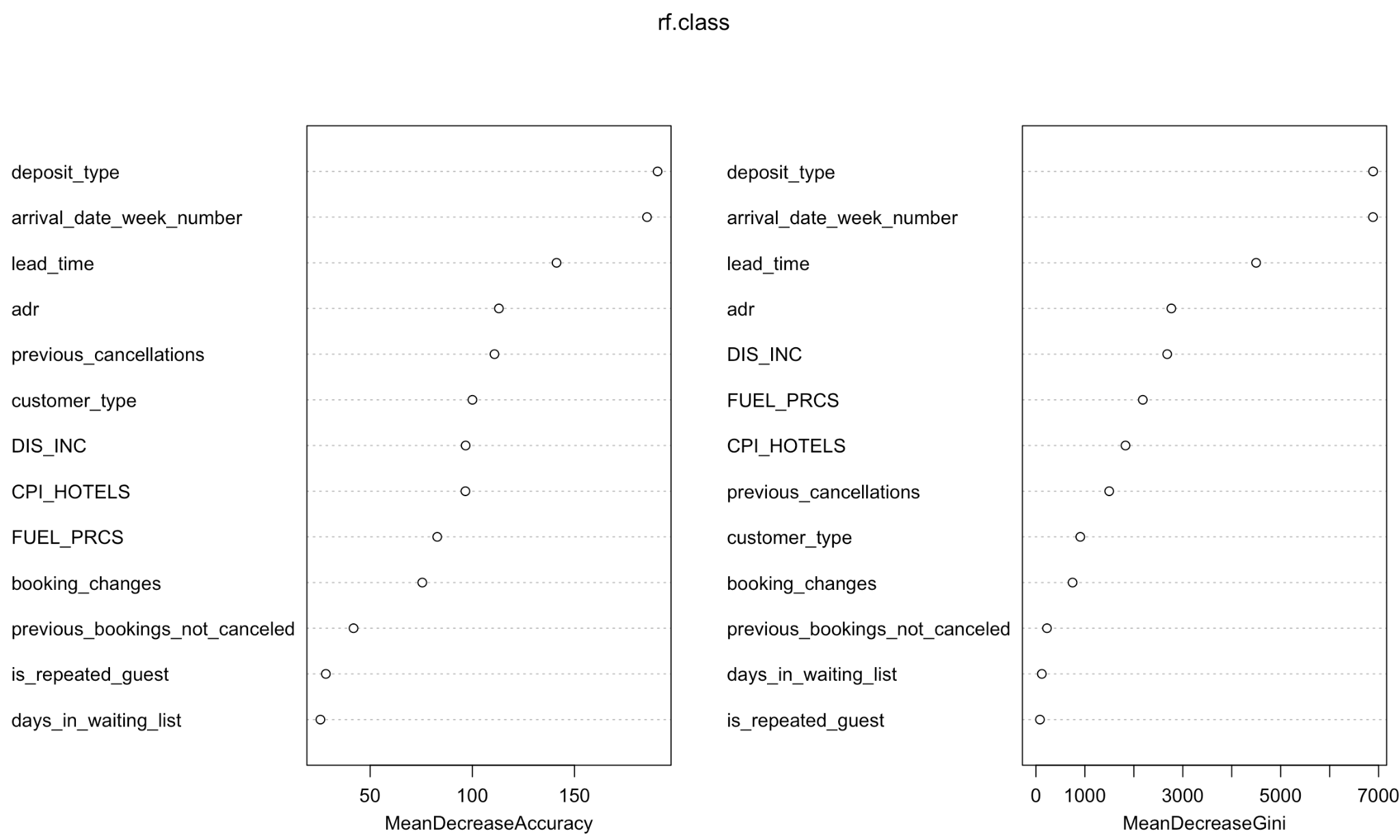
1. Output summary:



1. Our tree has 6 terminal nodes.
2. Graphical Output:



1. Variable importance from random forest:

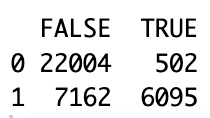


Deposit type and the week number of the arrival date were found to be the most important predictors. Following these two, lead time appears to also be an important predictor.

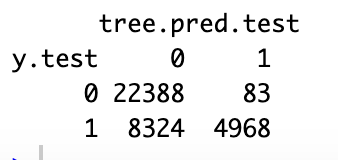
6.4 - Summary of Findings

1. Confusion matrices:

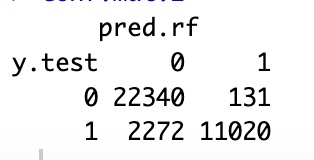
Confusion Matrix for Logistic Regression (0.5 threshold):



Confusion Matrix for Recursive Binary Splitting (0.5 threshold):



Confusion Matrix for Random Forests (0.5 threshold):

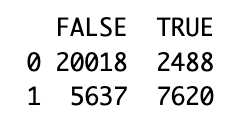


1. Test Error, FNR, FPR (at 0.5 threshold):

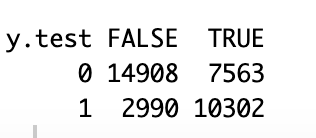
| **Model** | **False Negative Rate** | **False Positive Rate** | **Test Error** |
| --- | --- | --- | --- |
| Logistic Regression | 0.540 | 0.0223 | 0.2782 |
| Recursive Binary Splitting | 0.6262 | 0.0037 | 0.2351 |
| Random Forests | 0.1709 | 0.0058 | 0.0672 |

1. The confusion matrices and rates of the logistic regression and recursive binary splitting models show a trend where the False Negative Rate is significantly higher than the False Positive Rate. In order to have a more accurate model, these rates should be more balanced, therefore in order to solve this problem the threshold should be decreased from 0.5 to 0.4. Lowering the threshold will increase the false positive rate, however it should reduce the False Negative Rate at a much higher magnitude. Although there is a slight imbalance between the FPR and FNR for random forests, both error rates are already really low, so the threshold of 0.5 is good to use for random forests.
2. Confusion matrices with adjusted thresholds

Confusion Matrix for Logistic Regression (0.4 threshold):



Confusion Matrix for Recursive Binary Splitting (0.4 threshold):



1. Table with adjusted thresholds

| **Model** | **Threshold** | **False Negative Rate** | **False Positive Rate** | **Test Error** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.4 | 0.4252 | 0.1105 | 0.3707 |
| Recursive Binary Splitting | 0.4 | 0.2249 | 0.3366 | 0.2951 |
| Random Forest | 0.5 | 0.1709 | 0.0058 | 0.0672 |

1. Discussion:

In our analysis of Hotel Reservation Data, the models we created are aimed to provide insights into the factors that influence hotel reservation cancellations and how many nights people stay in a given hotel. We used Logistic Regression, classification trees, and random forests to examine the significance of these predictors.The results from the Logistic Regression models demonstrated that all predictors, with the exception of the CPI\_HOTELS variable, were significant in predicting hotel cancellation odds. This implied that each factor had a role in influencing whether a hotel reservation would be canceled or not. Our updated logistic regression model tells us the change in the log odds for a hotel cancellation with a one-unit increase in the predictor while holding all other predictors constant, which helps to answer our question of interest. The recursive binary splitting and random forests models helped answer our question of interest by allowing us to see which variables are important in predicting if a hotel reservation will be canceled. Not only this, but both models enabled us to find out the magnitude of importance that each predictor had, and were able to separate certain predictors that had a very large significance from other predictors who had a lesser significance in determining a cancellation. Across all models, deposit type and lead time were shown to be two of the most important predictors.

1. Out of all three models, the random forest model best addressed our question of interest because it had by far the lowest overall error rate out of the three models, however each model had its strengths with determining the significance of predictors. Although the logistic regression model was not able to provide the lowest error rate, it did give us an excellent step to start off with our predictors. The logistic regression model was able to handle a large number of predictors, and combined with the Wald test and LRT tests was able to show us a list of significant predictors. When looking at the output of the classification trees and random forests it is easy to assume that predictors who are at the bottom of the tree, or lowest on the forest output are completely insignificant. The logistic regression model was able to show us that although certain predictors have more significance than others, they are all still significant in some sort of way.

6.5 - Address Previous Comments

| You are writing about the Wald tests in the original logistic regression correct? So all the predictors are quantitative. You write that you should remove CPI Hotels (which is significant), no deposit dummy (which is categorical and not part of the original logistic regression), and group dummy (which is categorical and not part of the original logistic regression). Please clear this up or edit as needed (2 demerits) - | The wald tests were conducted in our original model with 10 variables only, not in our extended model with 13 variables. Since the two dummy variables are categorical instead of quantitative, only CPI HOTELS was removed from the logistic regression model. |
| --- | --- |
| I have no idea if your data is unbalanced. If it is, you need to check the confusion matrix, FPR and FNR as well in this instance. Test error, ROC and AUC may look fine but these other other measures may not. (3 demerit) | The imbalance of FPR and FNR in our logistic regression confusion matrix was fixed by introducing a lower threshold of 0.4 |
| Section 5d: this output doesn't prove to me the pruned tree is the same as recursive binary splitting. You need to show me the deviance vs the size of the tree. (2 demerits) | The deviance and size of the recursive binary tree compared to the pruned tree was included in part 6.3 |

**Section 7- Further Work**

* If our group had more time to work on this project, we would consider several strategies to enhance our findings and data. Such strategies include:
  + Deeper Analysis of Demographic Data: Expanding the analysis to include a more detailed examination of demographic data such as income level and geographical/regional location of the guests. Adding additional variables to our guest demographics could potentially provide a more detailed, accurate understanding of customer behaviors and preferences. The deeper analysis into the demographic information of hotel bookings and cancellations could improve upon our logistic regression by testing more predictors than in our initial analysis and other categorical models.
  + Analysis of Customer Reviews: Incorporating and analyzing text data into our exploration from customer reviews to gauge customer satisfaction could reveal a correlation with booking habits. This idea essentially, too, a deeper analysis of guest data; however, customer reviews, specifically, are likely able to determine whether or not a guest would cancel a hotel reservation in the future. Sentiment analysis can provide qualitative insights that numerical data might miss.
  + Time Series Analysis: Incorporating a time series analysis to understand seasonal patterns and trends over a longer period of time. This could help in predicting future booking behaviors and cancellations based on historical patterns over a greater period of time.
  + Interaction Development: Creating a dynamic or interactive visualization for viewers, specifically hotel stakeholders, using R tools such as ‘plotly’, to dynamically explore data and predictions in a single plot, rather than 2 static visualizations. Dynamic / interactive plots enable the viewer to have more manipulation and control over the data. A dynamic / interactive visualization could better enable viewers to make informed decisions based on real-time insights, while also revealing insights that may have otherwise been unknown.
  + Impact of External Factors: Further exploration into the impact of external factors such as the knowledge of global events, policy changes, or major economic shifts could better predict hotel cancellations. Analyzing the impacts of a health crisis, like the COVID-19 pandemic, on hotel bookings and cancellations could better accurately predict changes in a given hotel’s total booking volume and cancellation rates.

**Section 8- Reflection on Learning**

The STAT 4630 final project has allowed us to apply statistical concepts and machine learning theories learned in class to a real-world scenario. Applying the knowledge learned from lecture to a real-world scenario bridges the gap between theoretical knowledge and practical application. The project helped reinforce our learning in this class by giving us a sandbox for practical application of our accumulated knowledge to a real-world situation. STAT 4630 first explained the lecture material, demonstrated how to carry out technical procedures, guided our knowledge through homeworks, quizzes, & midterms, and ultimately enabled us to research and explore our own data. Building and validating multiple machine-learning and statistical models provided us with hands-on experience in model development, selection, and performance evaluation.

Working in a group taught us the importance of teamwork, communication, collaboration, and problem-solving. We learned to leverage each other's strengths / weaknesses to work efficiently towards a common goal. The project required us to manage a large dataset with multiple variables. This helped us improve our data handling and manipulating skills, which are crucial in any data-driven field. Lastly, addressing feedback and revising initial project submissions taught us the values that come from constructive criticism and continuous improvement in our research and analysis.

Overall, this project was an invaluable step in our learning experience. The project ultimately served as an example for a comprehensive and practical understanding of data analysis using machine learning and its applications in the real world.