Syracuse University
School of Information Studies

IST687 Introduction to Data Science

Hotel Booking Data Analysis

Asim Kazi, Kevin Harmer, Megha Banerjee, Pranav Sharma, Shruti Varma

Final Project Group 4

I. Introduction

Hotels are the biggest part of the modern travel industry. Any reason to travel is enough justification to make a hotel reservation. With the growth of hotel industries all over the world, it is necessary to start using data to maximize profits. One major way to go about maximizing hotel profits with data is analyzing trends or patterns in reservation cancellations.

While there are infinite possibilities of why someone would want to cancel their hotel reservation, there may be some data characteristics that can foreshadow cancellations. So, we will take the data from https://intro-datascience.s3.us-east-2.amazonaws.com/Resort01.csv and try to find any patterns to better understand and later predict cancellations.

II. Team Members Involved

This project report is created by the individuals listed below for the IST 687- Introduction to Data Science Class.

- Asim Kazi
- Kevin Harmer
- Megha Banerjee
- Pranav Sharma
- Shruti Varma

III. Data Set Description

The data set provided for this project contains real-life hotel stay data, with each row representing a hotel booking. It has a total of 40060 observations of 20 variables. The further description of every data is listed below.

Column Name	Type	Description
IsCanceled	Categorical	Indicating if the booking was ● 1- Canceled ● 0- Not canceled
LeadTime	Integer	Number of days that elapsed between the entering date of the booking into and the arrival date

StaysInWeekendNights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
StaysInWeekNights	Integer	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
Adults	Integer	Number of adults
Children	Integer	Number of children
Babies	Integer	Number of babies
Meal	Categorical	Type of meal booked. Categories are presented in standard hospitality meal packages: ● Undefined/SC – no meal package ● BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner) ● FB – Full board (breakfast, lunch and dinner)
Country	Categorical	Country of origin. Categories are represented in the ISO 3155– 3:2013 format
MarketSegment	Categorical	Market segment designation. In categories, the term ● Complimentary- Complimentary Bookings (usually for promotions) ● Corporate- Bookings through companies ● Direct- Direct bookings from customers ● Groups- Bookings for a large group ● Online TA/TO- Online Travel Agents or Travel Operators ● Online TA- Online Travel Agents
IsRepeatedGuest	Categorical	Value indicating if the booking name was from a ● 1- Repeated guest ● 0- Not repeated guest

PreviousCancellations	Integer	Number of previous bookings that were cancelled by the customer prior to the current booking
PreviousBookingsNotCanceled	Integer	Number of previous bookings not cancelled by the customer prior to the current booking
ReservedRoomType	Categorical	Code of room type reserved.
AssignedRoomType	Categorical	Code for the type of room assigned to the booking.
BookingChanges	Integer	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
DepositType	Categorical	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-Refundable – 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.
CustomerType	Categorical	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
RequiredCardParkingSpaces	Integer	Number of car parking spaces required by the customer

TotalOfSpecialRequests	Integer	Number of special requests made by the customer (e.g. twin bed or high floor)

IV. Preliminary Analysis

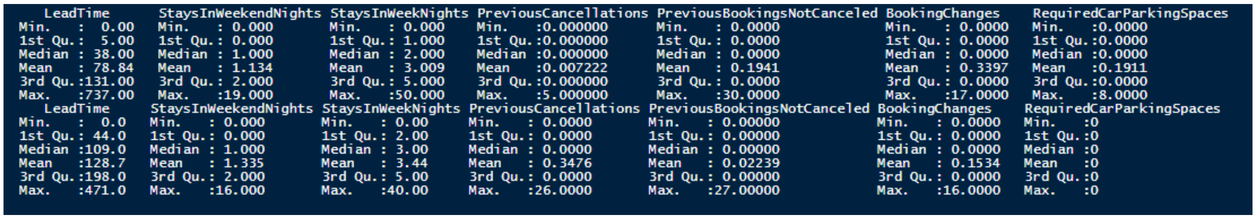
Upon examining the structure of the data set, we see that there are several different avenues of analysis that can be taken. Based on the amount of data and 20 variables, we can conduct several different types of analysis on the data set, including linear and nonlinear trends on the integer data, interactions between integer variables, support vector machines, association rule mining in categorical data and more.

But before we dive into our potential insights that can be gathered from our data set, it is important to clean our data set. Upon thorough examination of the data set, each column has correct data types and reasonable factors that can be used for analysis, with the exception of the Country variable. The Country variable had 464 "NULL" values, we excluded from our analysis.

V. Problem Formulation

A. Numerical Analysis

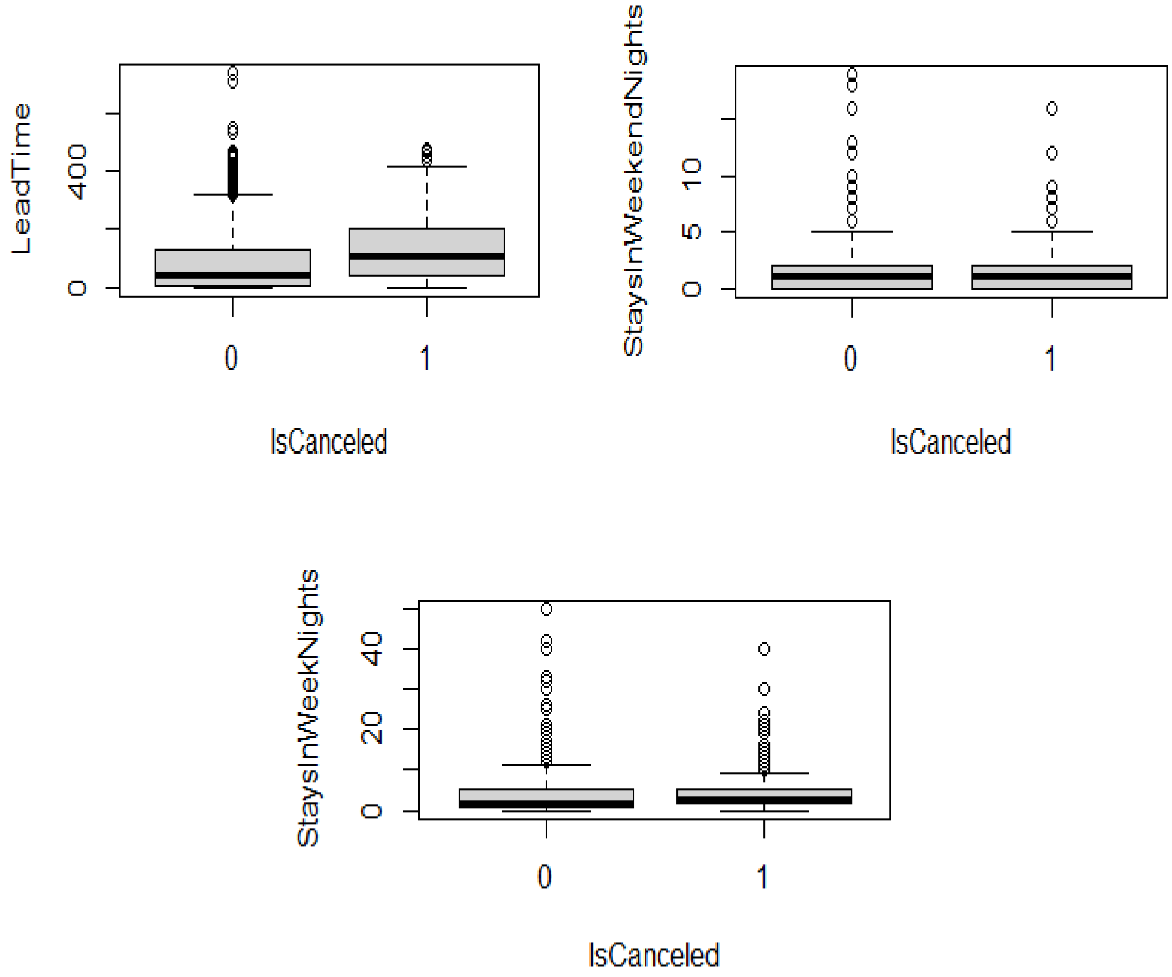
Although there are 20 different variables in the data set, only 7 of those variables are numerics. To analyze their differences, we can split the data set into cancellations (which we will call cancel) and non-cancellations (which we will call stay). With two distinct data sets analyzing the same numeric variables, we can begin to search for differences in the data set. The top row includes non-cancellation observations and the second row includes cancellation observations.



```
    LeadTime      StaysInWeekendNights StaysInWeekNights PreviousCancellations PreviousBookingsNotCanceled BookingChanges    RequiredCarParkingSpaces
 Min.   :  0.00   Min.   : 0.000       Min.   : 0.000    Min.   :0.000000      Min.   : 0.0000             Min.   : 0.0000   Min.   :0.0000
 1st Qu.:  5.00   1st Qu.: 0.000       1st Qu.: 1.000    1st Qu.:0.000000      1st Qu.: 0.0000             1st Qu.: 0.0000   1st Qu.:0.0000
 Median : 38.00   Median : 1.000       Median : 2.000    Median :0.000000      Median : 0.0000             Median : 0.0000   Median :0.0000
 Mean   : 78.84   Mean   : 1.134       Mean   : 3.009    Mean   :0.007222      Mean   : 0.1941             Mean   : 0.3397   Mean   :0.1911
 3rd Qu.:131.00   3rd Qu.: 2.000       3rd Qu.: 5.000    3rd Qu.:0.000000      3rd Qu.: 0.0000             3rd Qu.: 0.0000   3rd Qu.:0.0000
 Max.   :737.00   Max.   :19.000       Max.   :50.000    Max.   :5.000000      Max.   :30.0000             Max.   :17.0000   Max.   :8.0000
    LeadTime     StaysInWeekendNights StaysInWeekNights PreviousCancellations PreviousBookingsNotCanceled BookingChanges    RequiredCarParkingSpaces
 Min.   :  0.0   Min.   : 0.000       Min.   : 0.00     Min.   : 0.0000       Min.   : 0.00000            Min.   : 0.0000   Min.   :0
 1st Qu.: 44.0   1st Qu.: 0.000       1st Qu.: 2.00     1st Qu.: 0.0000       1st Qu.: 0.00000            1st Qu.: 0.0000   1st Qu.:0
 Median :109.0   Median : 1.000       Median : 3.00     Median : 0.0000       Median : 0.00000            Median : 0.0000   Median :0
 Mean   :128.7   Mean   : 1.335       Mean   : 3.44     Mean   : 0.3476       Mean   : 0.02239            Mean   : 0.1534   Mean   :0
 3rd Qu.:198.0   3rd Qu.: 2.000       3rd Qu.: 5.00     3rd Qu.: 0.0000       3rd Qu.: 0.00000            3rd Qu.: 0.0000   3rd Qu.:0
 Max.   :471.0   Max.   :16.000       Max.   :40.00     Max.   :26.0000       Max.   :27.00000            Max.   :16.0000   Max.   :0
```

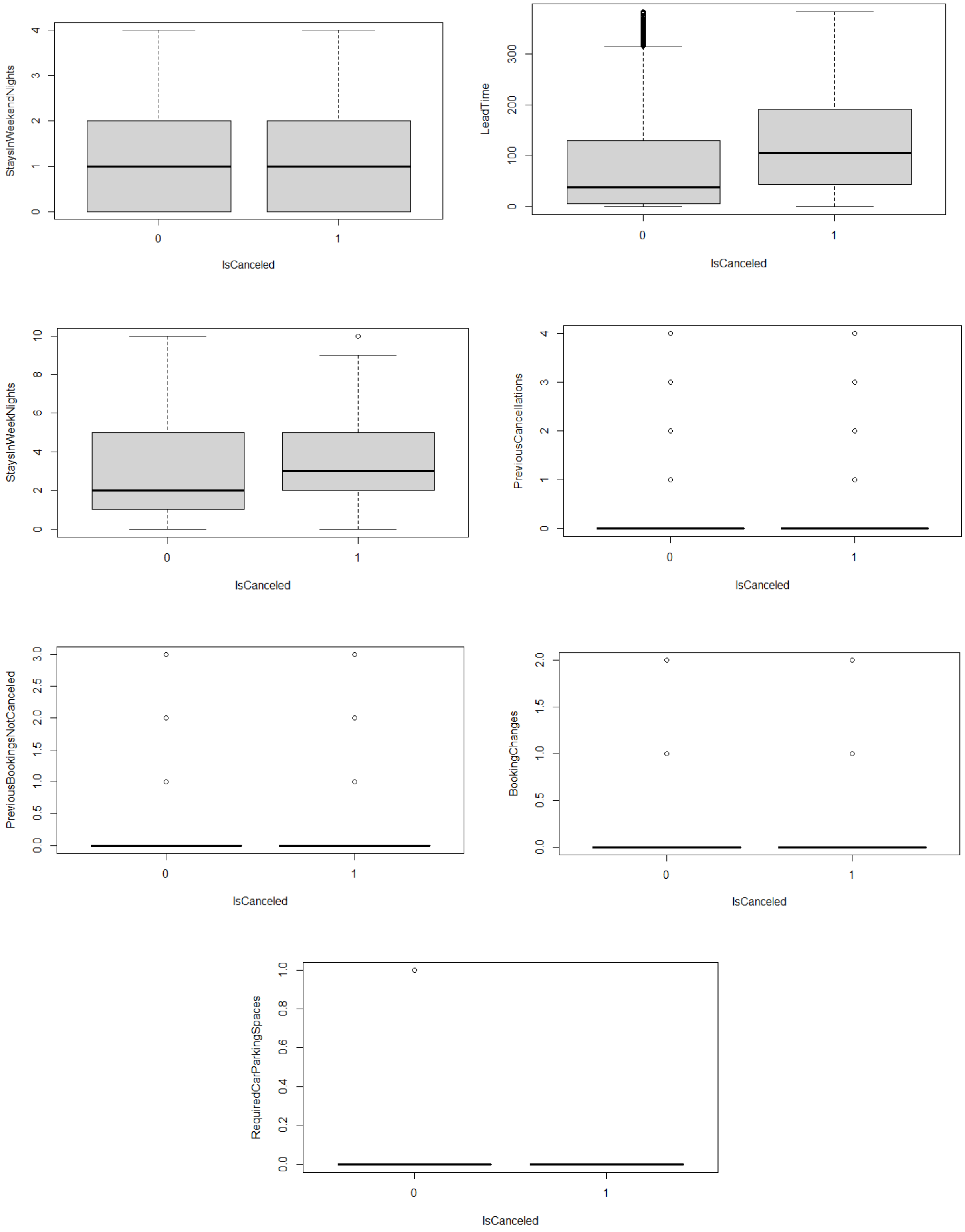
Descriptive Statistics of Numerical Variables

The descriptive statistics show that there may be some differences in the two data sets. Lead time was the strongest difference, showing much larger observations in the cancellation data set. Otherwise, there were some differences in means, specifically in PreviousCancellations (favoring cancellations), PreviousBookingsNotCancelled (favoring non-cancellations), and BookingChanges (favoring non-cancellations). We can analyze these distributions further by examining the boxplots (0 meaning not canceled and 1 meaning canceled).



Raw Boxplots of Numerical Data

Four of the boxplots are not shown because they do not show anything substantial. Overall, the boxplots are hard to analyze due to the large number of outliers. Consequently, we will subset our data sets further and remove any extreme values just for analyzing the distributions. We decided to remove any value above the mean plus three standard deviations to get a better idea of the data distribution. We will incorporate the points back into the data set in the future when analyzing our models. The results are shown below.



Adjusted Boxplots of Numerical Data

The results of these boxplots are much clearer than the previous values. As expected there is a large difference in lead time, showing much higher lead times in the cancellation data set. Due to that difference, there is promising evidence that there is a relationship between LeadTime and whether the hotel reservation is canceled, but we will explore that more later. There also appears to be a minor trend in the StaysInWeekendNights, but no clear difference in StaysInWeekNights.

Otherwise, the data appears non-conclusive or centered around 0. This suggests that there may be some other factor relationship between the last four variables and cancellation result. As a result, we may be able to analyze this behavior later by subsetting those variables into 2-factor variables (zero and non-zero) and check to see if there is any underlying relationship.

Furthermore, we may be able to view some of these variables in correspondence with the interactions between variables. For example, if someone were to cancel 1 booking, but attend 20 other hotel reservations, they may be more likely to not cancel despite the possible correlation of their cancellation to their next reservation being canceled.

B. Analyzing the Relationship of Numerics as Factors

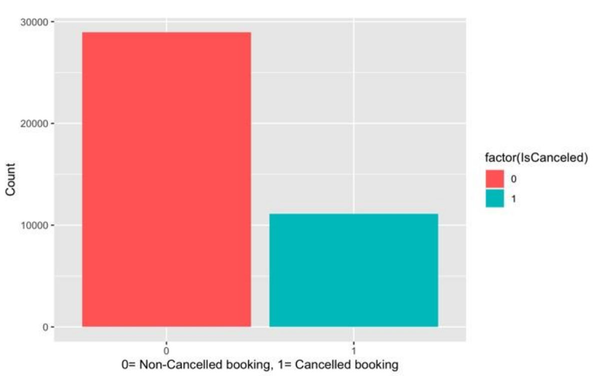
As discussed in the previous section, there were some differences in the numeric descriptive statistics, but were not really observed in the boxplots. The abnormal variables include PreviousCancellations, PreviousBookingsNotCanceled, BookingChanges, and RequiredCarParkingSpaces. Consequently, we will take the variables with abnormal boxplots and split the variables into factors, with 0 representing 0 and 1 representing any non-zero value. By treating our numerical variables as binary inputs, we can include them in the factor analysis with the rest of the categorical variables. The subset coding is shown in Section VI.

C. Variable Interactions

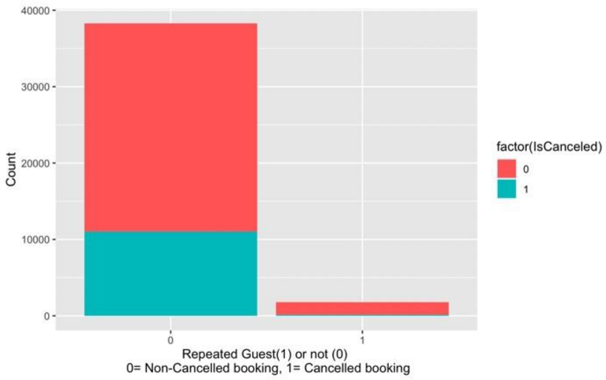
To expand on the point made in Section V-A., there may be underlying relationships in the data outside direct relationships between independent variables and the dependent variable. Instead, there are some observations that have more information that is given. Specifically, we will look at the interaction between StaysInWeekNights/StaysInWeekendNights and PreviousCancellations/PreviousBookingsNotCanceled, as well as their combined impact on the cancellation of hotel reservations.

D. Barplots and Significant Percentages

After the initial analysis, we further tried to focus on some variables deeply to find out the contributing factors. This time, we used bar plots as a visualization tool.The bar plot presented below is the distribution graph for the IsCanceled variable (0 for Non-Cancelled bookings, 1 for the Cancelled ones). Among 40060 instances of hotel bookings, we have 11122 cancellations, which counts for around 28% cancellations. As the data set is not balanced, we have decided to analyze on the basis of percentages instead of counts. Next, we tried to look at what proportion of the guests in the hotel are repeated and whether this variable contributes to the cancellations or not. Around 71% guests are first timers (Not Repeated) guests. Repeated guests tend to cancel a booking a lot less than the Non-repeated guests.

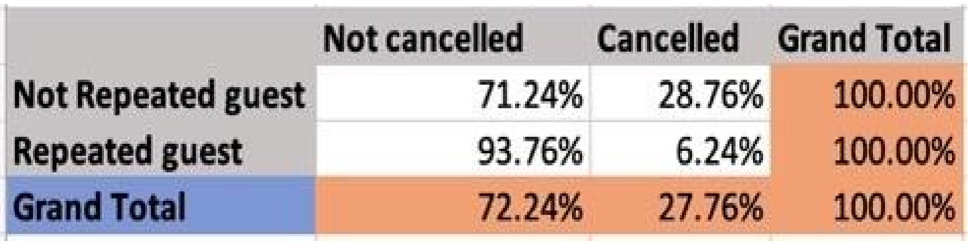


Frequency distribution of Non-cancelled and cancelled instances



Frequency distribution of Repeated guest variable in Non-cancelled and cancelled instances

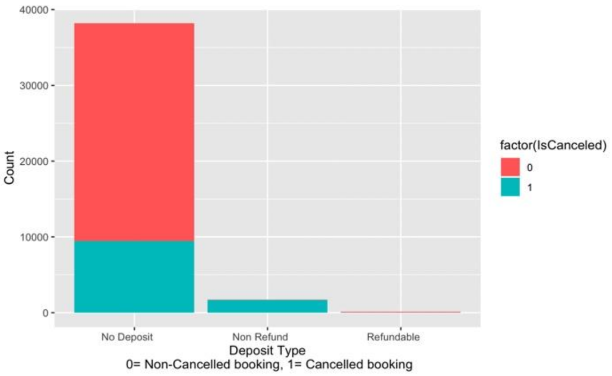
The percentages of repeated guests are significantly lower than the percentages of first timers. This can be because of the imbalanced data set. But if it is not, there must be some significant factors causing guests not to stay in this same hotel.



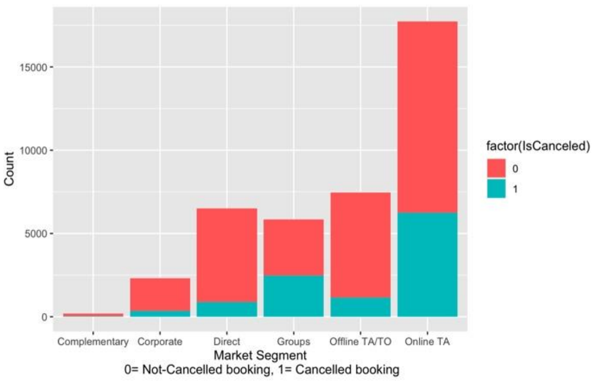
	Not cancelled	Cancelled	Grand Total
Not Repeated guest	71.24%	28.76%	100.00%
Repeated guest	93.76%	6.24%	100.00%
Grand Total	72.24%	27.76%	100.00%

Percentage distribution of repeated and non repeated guest with respect to cancels and non cancellations

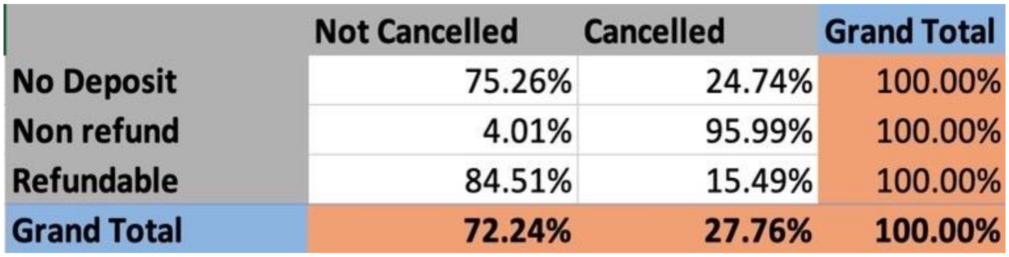
We further focused on another significant variable, DepositType. Out of all the 'Non refund' bookings, around 96% got cancelled. Also out of total 'No deposit' booking, around 25% got cancelled, which is significant as well. Another significant finding is that, out of all the group bookings, approximately 43% got cancelled. For online Tour Agents bookings, 35% got cancelled. According to the data, the total booking of the hotel comes via Online travel agents. Also, the highest percentages of cancellations also comes from the online TAs. The direct bookings account for highest percentages of non-cancellation which is almost 87%.



Frequency distribution of Deposit Type for Non-cancelled and cancelled instances

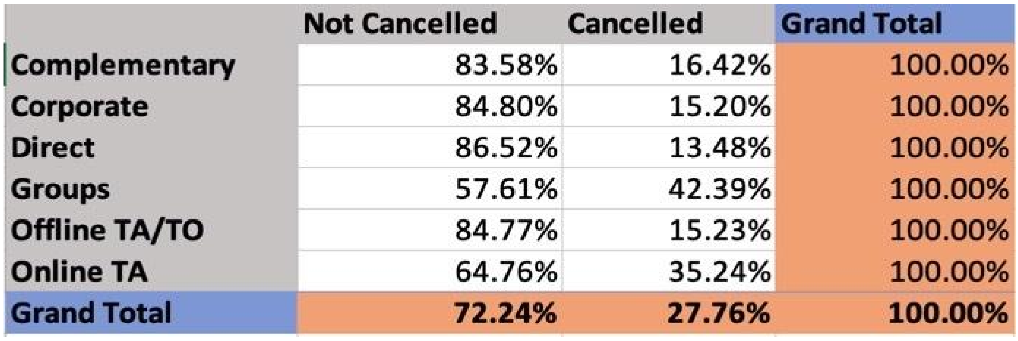


Frequency distribution of Market Segment for Non-cancelled and cancelled instances



	Not Cancelled	Cancelled	Grand Total
No Deposit	75.26%	24.74%	100.00%
Non refund	4.01%	95.99%	100.00%
Refundable	84.51%	15.49%	100.00%
Grand Total	**72.24%**	**27.76%**	**100.00%**

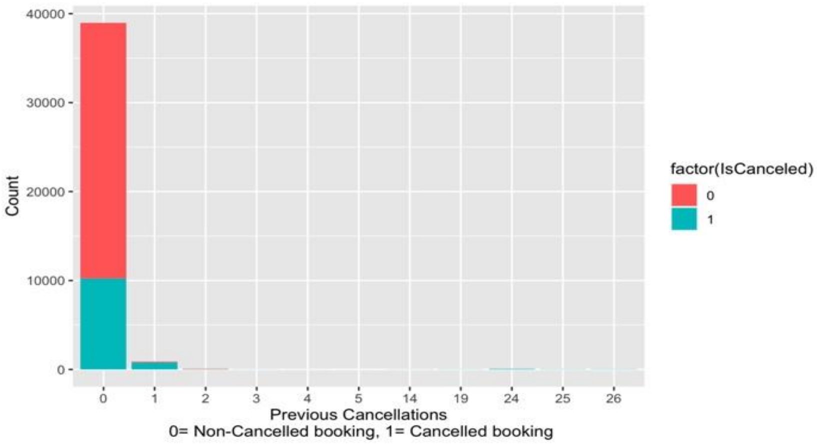
Percentage distribution of Deposit Type with respect to cancels and non cancellations



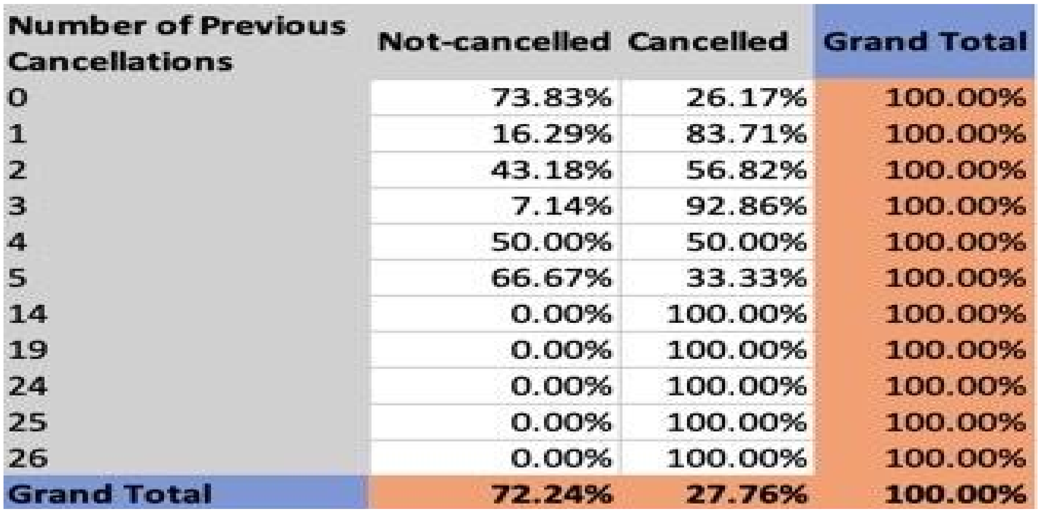
	Not Cancelled	Cancelled	Grand Total
Complementary	83.58%	16.42%	100.00%
Corporate	84.80%	15.20%	100.00%
Direct	86.52%	13.48%	100.00%
Groups	57.61%	42.39%	100.00%
Offline TA/TO	84.77%	15.23%	100.00%
Online TA	64.76%	35.24%	100.00%
Grand Total	**72.24%**	**27.76%**	**100.00%**

Percentage distribution of Market Segment with respect to cancels and non cancellations

The previous cancellation is another variable that we focused on. The percentages suggest that a customer with a history of one previous cancellation is more likely to cancel the next booking. According to the data, it happened 84% of the time. Eventually, customers with a higher number of previous cancellations are more likely to cancel.



Frequency distribution of Previous Cancellations for Non-cancelled and cancelled instances



Number of Previous Cancellations	Not-cancelled	Cancelled	Grand Total
0	73.83%	26.17%	100.00%
1	16.29%	83.71%	100.00%
2	43.18%	56.82%	100.00%
3	7.14%	92.86%	100.00%
4	50.00%	50.00%	100.00%
5	66.67%	33.33%	100.00%
14	0.00%	100.00%	100.00%
19	0.00%	100.00%	100.00%
24	0.00%	100.00%	100.00%
25	0.00%	100.00%	100.00%
26	0.00%	100.00%	100.00%
Grand Total	72.24%	27.76%	100.00%

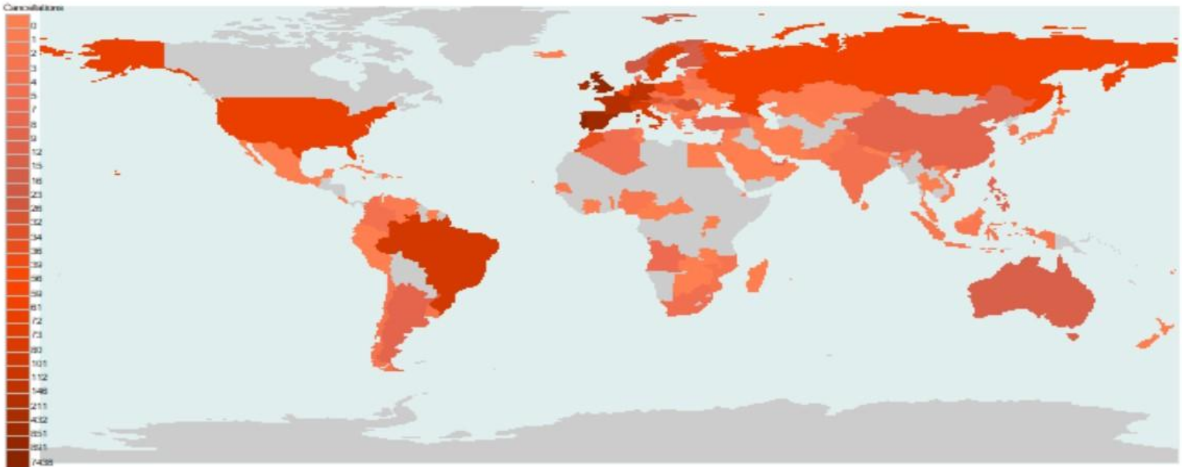
Percentage distribution of Previous Cancellations history with respect to cancels and non cancellations

These percentages provide interesting results. In Section VI., we will dive into the strongest contributing factors which lead to a cancellation.

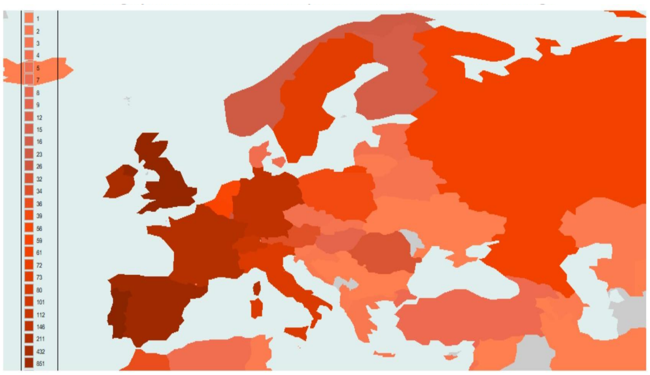
E. Geographic Information

In this section we used geographical information to help understand and analyze any spatial patterns and relationships. We grouped the countries together to find the number of bookings cancelled by each country and used that information to construct a world map and further looked into every region.

The map constructed shows Portugal to be the country with the highest number of cancellations followed by countries in the United Kingdom and Latin America. We also focused on every region to get a better comprehension of which region was subjected to maximum cancellation. The darkest color in the legend shows the highest number of cancellations.

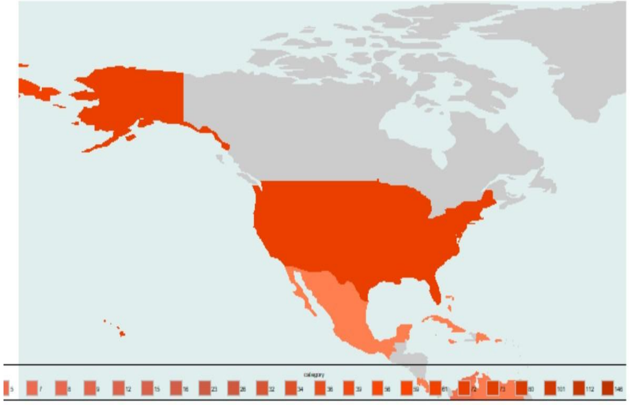


Geographical Information of Countries based on cancellations

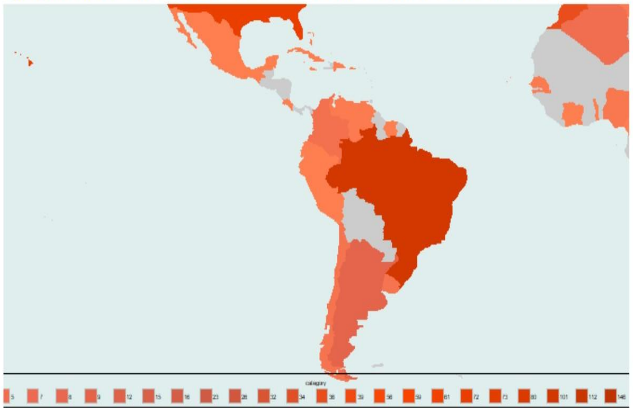


Geographical Information of the European Countries that cancelled bookings

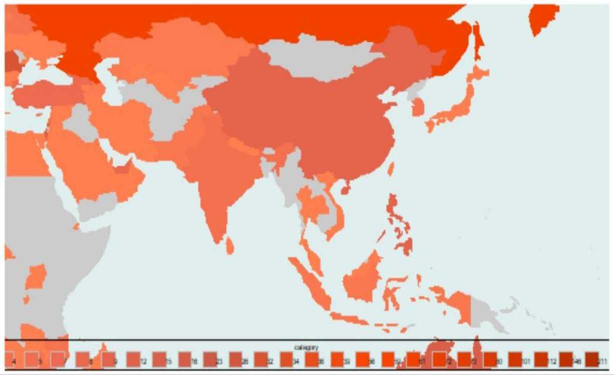
Geographical Information of the UK Countries that cancelled bookings



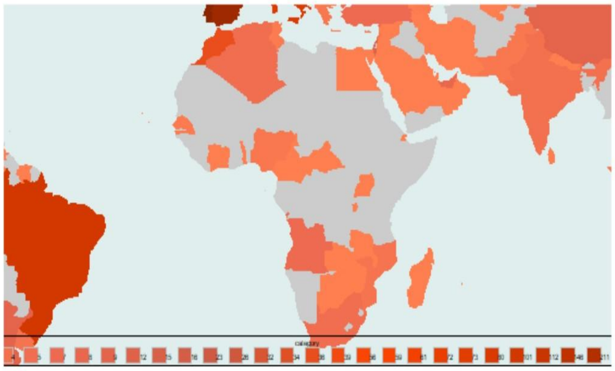
Geographical Information of the North American Countries that cancelled bookings



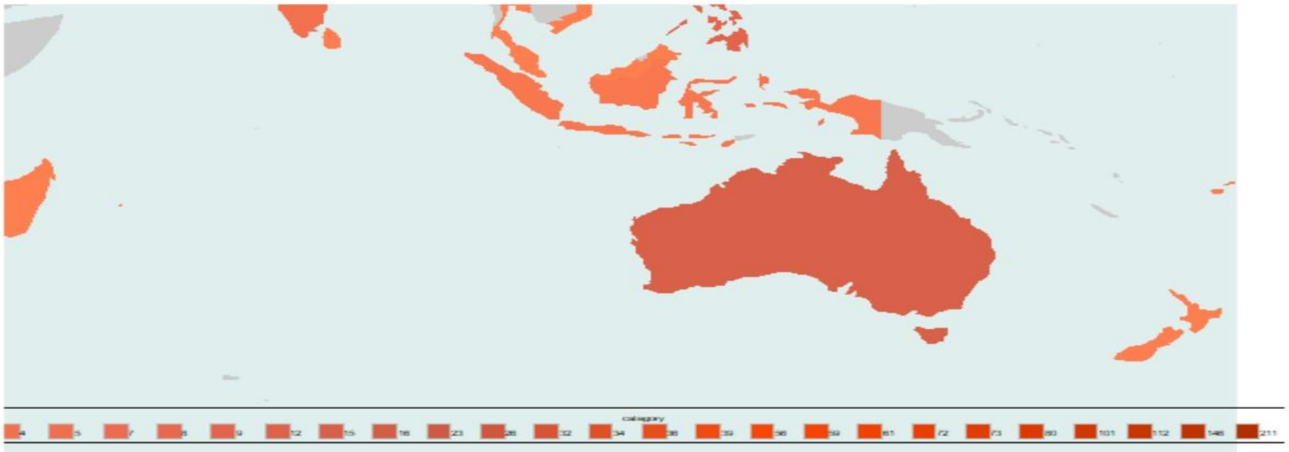
Geographical Information of the Latin American Countries that cancelled bookings



Geographical Information of the Asian Countries that cancelled bookings



Geographical Information of the African Countries that cancelled bookings

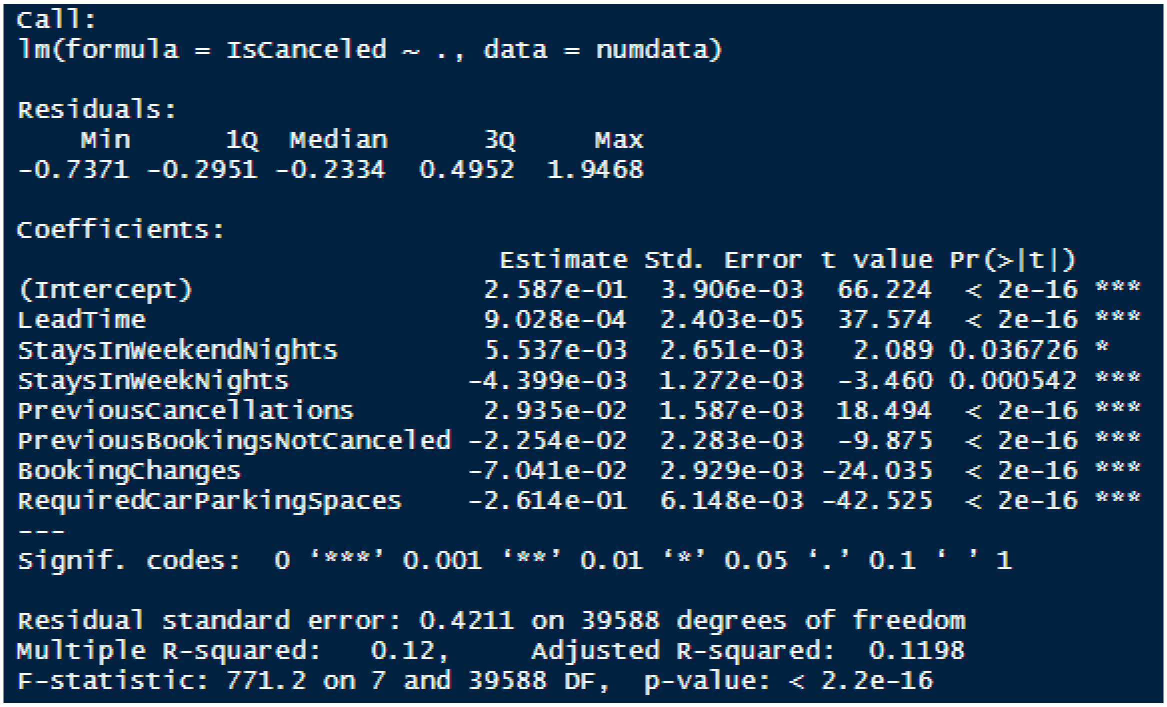


Geographical Information of the Oceania Countries that cancelled bookings

VI. Analysis & Results

A. Numerical Trends

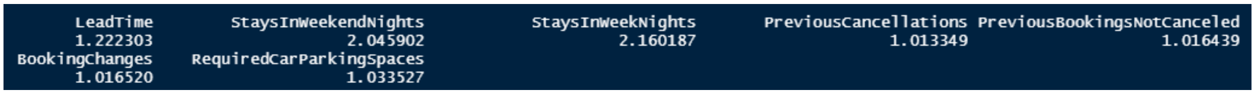
In this section, we will take a look at the raw numerical data. We will begin by conducting a linear regression on the IsCanceled variable and numerical data. Specifically, we will use R to run a lm() function (from the stats package) with IsCanceled as the independent variable and LeadTime, StaysInWeekendNights, StaysInWeekNights, PreviousCancelations, PreviousBookingsNotCanceled, BookingChanges, and RequiredParkingSpaces as the dependent variables. The summary of the results is shown below:



```
Call:
lm(formula = IsCanceled ~ ., data = numdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7371 -0.2951 -0.2334  0.4952  1.9468 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  2.587e-01  3.906e-03  66.224  < 2e-16 ***
LeadTime                     9.028e-04  2.403e-05  37.574  < 2e-16 ***
StaysInWeekendNights         5.537e-03  2.651e-03   2.089 0.036726 *  
StaysInWeekNights           -4.399e-03  1.272e-03  -3.460 0.000542 ***
PreviousCancellations        2.935e-02  1.587e-03  18.494  < 2e-16 ***
PreviousBookingsNotCanceled -2.254e-02  2.283e-03  -9.875  < 2e-16 ***
BookingChanges              -7.041e-02  2.929e-03 -24.035  < 2e-16 ***
RequiredCarParkingSpaces    -2.614e-01  6.148e-03 -42.525  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4211 on 39588 degrees of freedom
Multiple R-squared:   0.12,	Adjusted R-squared:  0.1198 
F-statistic: 771.2 on 7 and 39588 DF,  p-value: < 2.2e-16
```

Linear Model Summarization

As seen by the statistics, there does appear to be a series of relationships between our IsCanceled variable and our dependent variables. With an alpha level of 0.05, we see each of our dependent variables have significant results. Furthermore, the F-Statistic is extremely high and corresponds to an extremely low p-value. We will also check for collinearity between our variables, using the vif() command in R's car package.

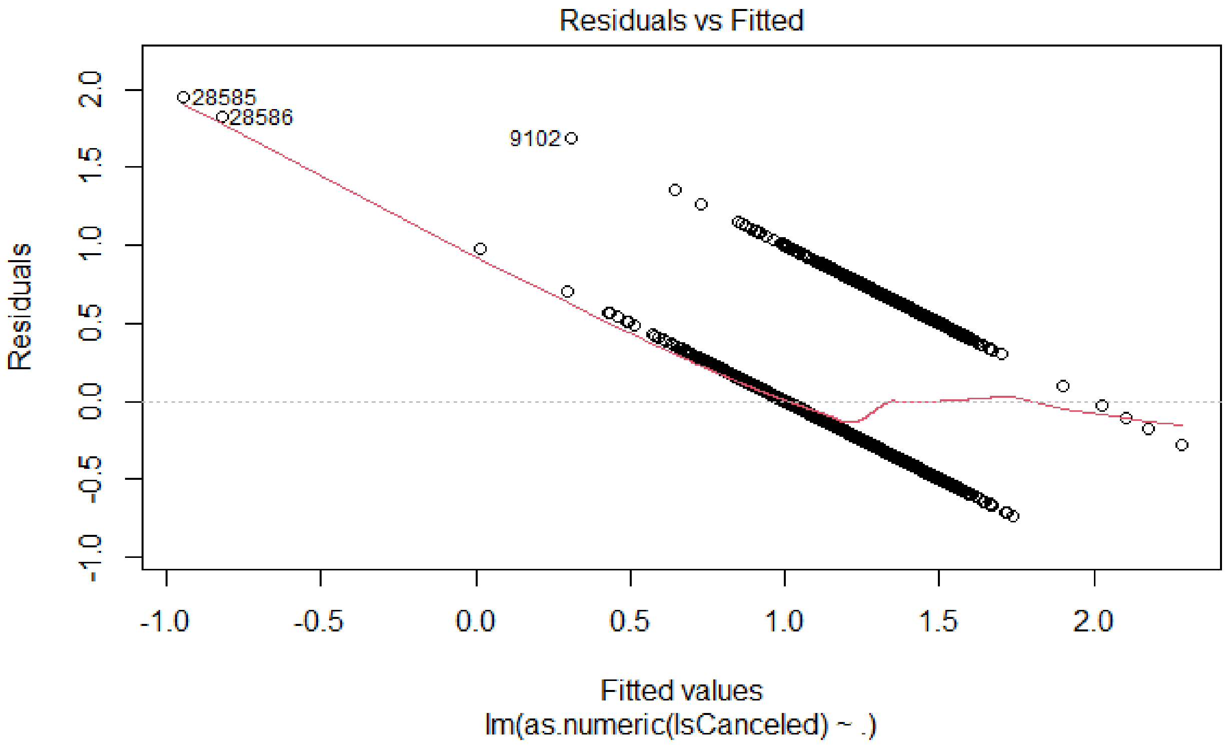


```
       LeadTime         StaysInWeekendNights              StaysInWeekNights        PreviousCancellations PreviousBookingsNotCanceled 
       1.222303                     2.045902                       2.160187                     1.013349                     1.016439 
 BookingChanges     RequiredCarParkingSpaces 
       1.016520                     1.033527 
```

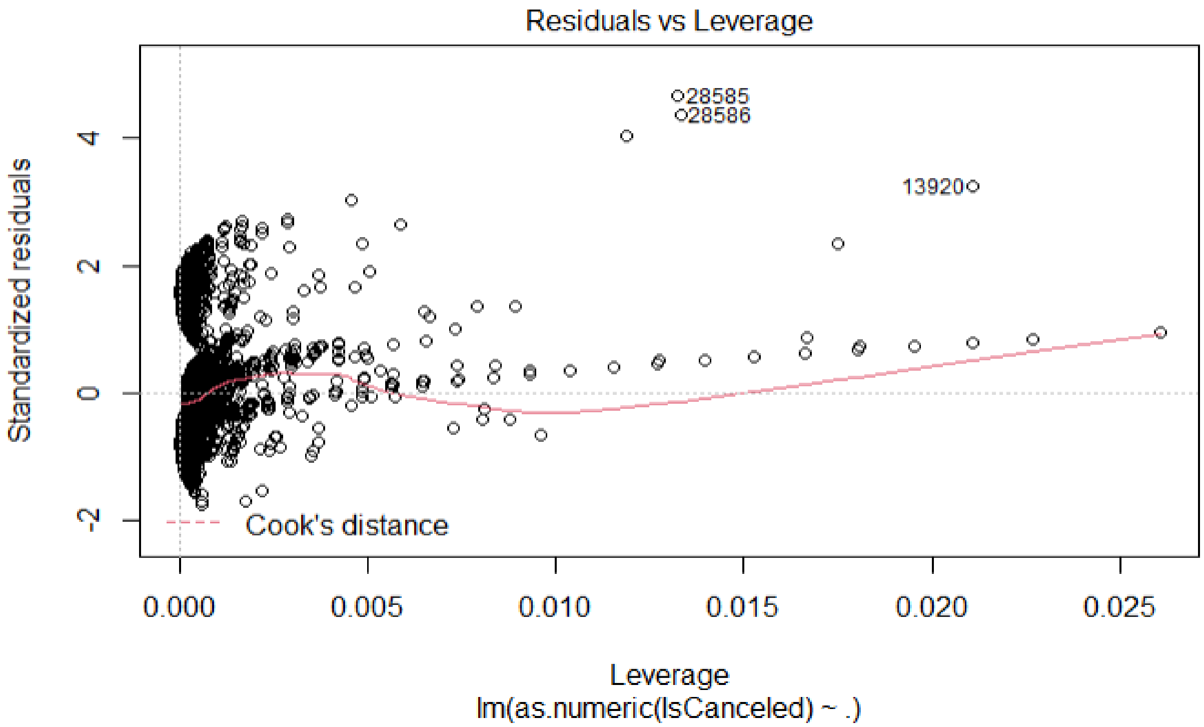
Linear Model Collinearity

The highest results for vif were in StaysInWeekendNights (2.046) and StaysInWeekNights (2.160), which were not high enough to suggest any collinearity. However, it is interesting to see that those two variables were the highest, meaning they may have an interaction component that we have not yet uncovered. But we will look at that more in the next section.

Despite the model yielding significant results, the actual results are not as promising. Specifically, the adjusted r-squared value came to 0.1198, which is relatively small for our expected data trend. Furthermore, our beta-weights do not appear to be overly influential on the data set. LeadTime, the only variable with input above 40, contributes to a 0.33 increase in IsCanceled with every 365 days. Requiring a parking space came in with the highest beta weight only produced a decrease of 0.261 in IsCanceled with every additional parking space. Next we will use the plot() function to analyze the residual behaviors of our model.



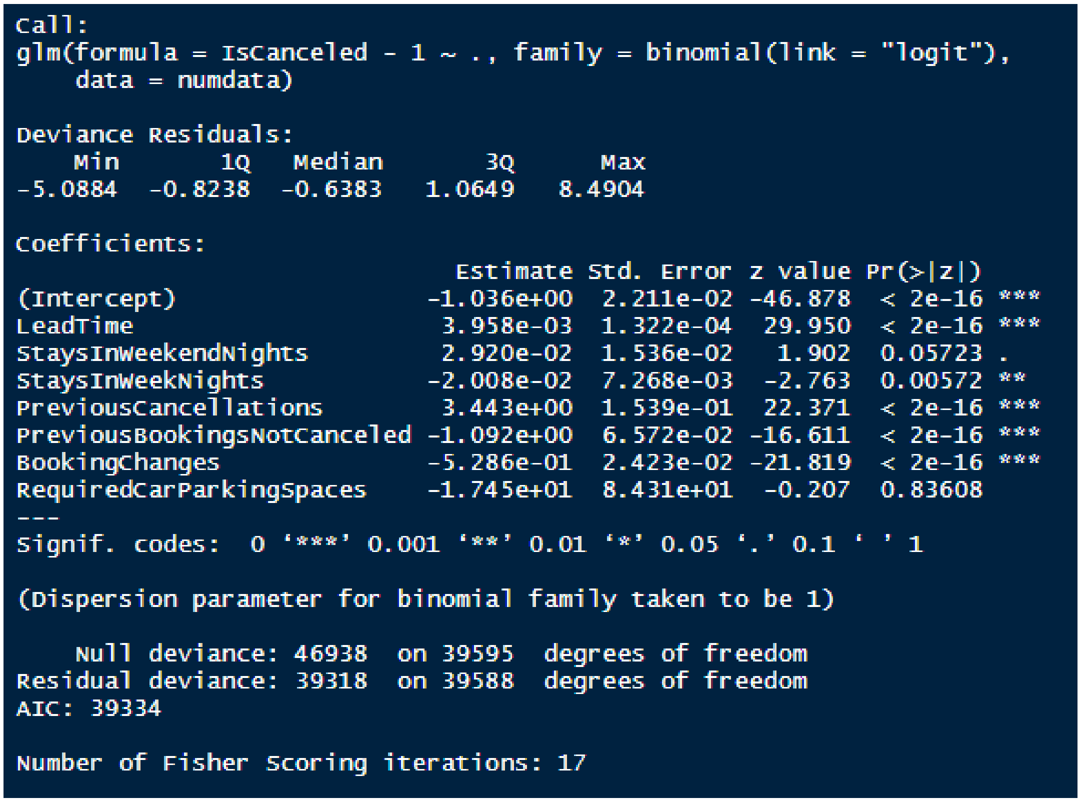
Residuals of Fitted Values



Standardized Residuals

The first plot shows the residuals in correspondence with fitted values. In the plot, there are visibly two lines for the data points and our linear model appears to be covering both of these subsets. Understandably, these two sets are the cancellation and non-cancellation data points. The second plot shows standardized residuals. The goal of this plot should have to have random points distributed around the 0 line, but the red line does show that there is some underlying trend in the data. Both of these plots are evidence that the relationship between our variables are not actually linear.

Therefore, our linear model only accounts for 12% of the variation in IsCanceled and the beta-weights are pretty weak. The plots suggest that the relationship is not actually linear. When considering the fact that our IsCanceled variable is a binary factor and not actually numeric, these are still decent results. These linear tests show that there is some relationship between IsCanceled and the other variables, but cannot necessarily describe it. So, we will move into a model that can. Binary functions can normally be explained by a logarithmic function, which predicts 1 or 0 based on the influence of other factors. Jumping right into it, we use the generalized linear model in the binomial family to analyze our relationship. The results of the model are shown below.

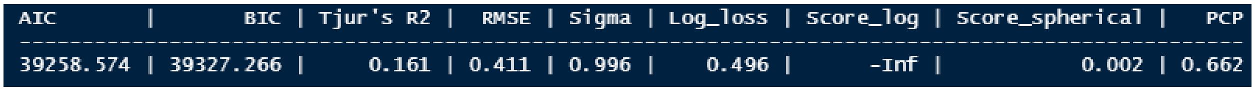


```
Call:
glm(formula = IsCanceled - 1 ~ ., family = binomial(link = "logit"),
    data = numdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0884  -0.8238  -0.6383   1.0649   8.4904

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.036e+00  2.211e-02 -46.878  < 2e-16 ***
LeadTime                     3.958e-03  1.322e-04  29.950  < 2e-16 ***
StaysInWeekendNights         2.920e-02  1.536e-02   1.902  0.05723 .
StaysInWeekNights           -2.008e-02  7.268e-03  -2.763  0.00572 **
PreviousCancellations        3.443e+00  1.539e-01  22.371  < 2e-16 ***
PreviousBookingsNotCanceled -1.092e+00  6.572e-02 -16.611  < 2e-16 ***
BookingChanges              -5.286e-01  2.423e-02 -21.819  < 2e-16 ***
RequiredCarParkingSpaces    -1.745e+01  8.431e+01  -0.207  0.83608
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46938  on 39595  degrees of freedom
Residual deviance: 39318  on 39588  degrees of freedom
AIC: 39334

Number of Fisher Scoring iterations: 17
```

Logistic Regression Summarization

The variable results of the logistic regression proved to be slightly different from the linear model. If we maintain our alpha level of 0.05, we now find that StaysInWeekendNights and RequiredCarParkingSpaces are no longer significant. It is also important to note that we can calculate a chi-squared from the resulting data. The Null deviance (46938) subtracted by the Residual deviance (39318) yields a total chi-squared value of 7620. We will break this value down further in the section.

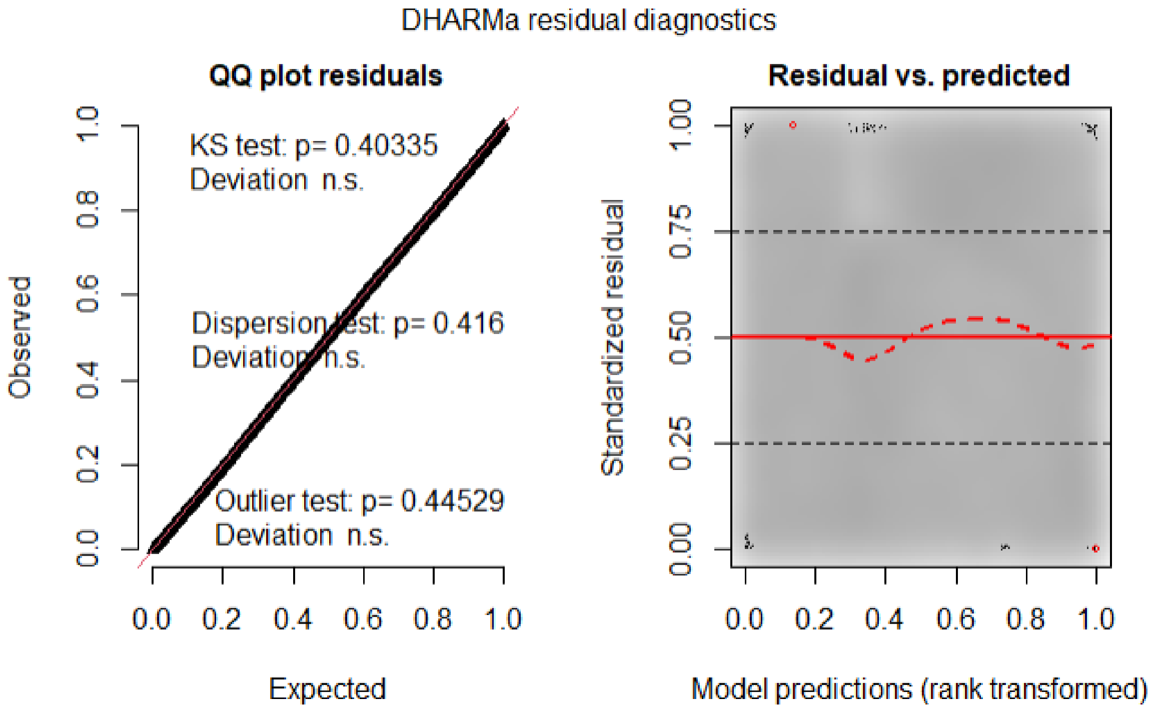
First, we will formulate an r-squared value using our logistic model. The performance library has a function, model_performance(), which produces a few statistics on models. This includes a pseudo r-squared value.The results of the function are shown below.



```
AIC       |       BIC | Tjur's R2 |  RMSE | Sigma | Log_loss | Score_log | Score_spherical |   PCP
-----------------------------------------------------------------------------------------------------
39258.574 | 39327.266 |     0.161 | 0.411 | 0.996 |    0.496 |      -Inf |           0.002 | 0.662
```

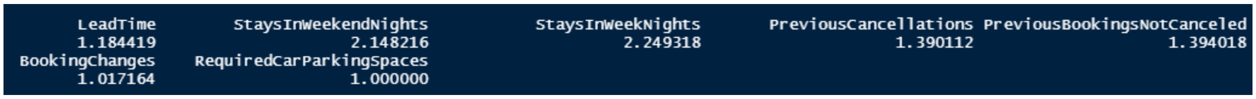
R-Squared Analysis

Tjur's r-squared provides a value of 0.161, which is stronger than our linear model's r-squared value. Although it is not necessarily high, it does give us a better indication of the data. Next, we will take a look at the regression fit using the DHARMa package. We can use the simulateResiduals() function to produce a data set of a simulated set of possible residuals then plot their behavior using the plot() function. The results are shown below.



Simulated Residual Analysis

The simulated residual values were consistent with the residual model. There was a small unexpected pattern in the plot on the right. Still, relatively random and normally behaving residuals support the idea that the underlying relationship between IsCanceled and the other variables have a logistic behavior. Before we move on to chi-squared analysis, we will take a look at the covariance between the dependent variables. Similar to the linear model, we will use the vif() function from the car package, with the corresponding results:

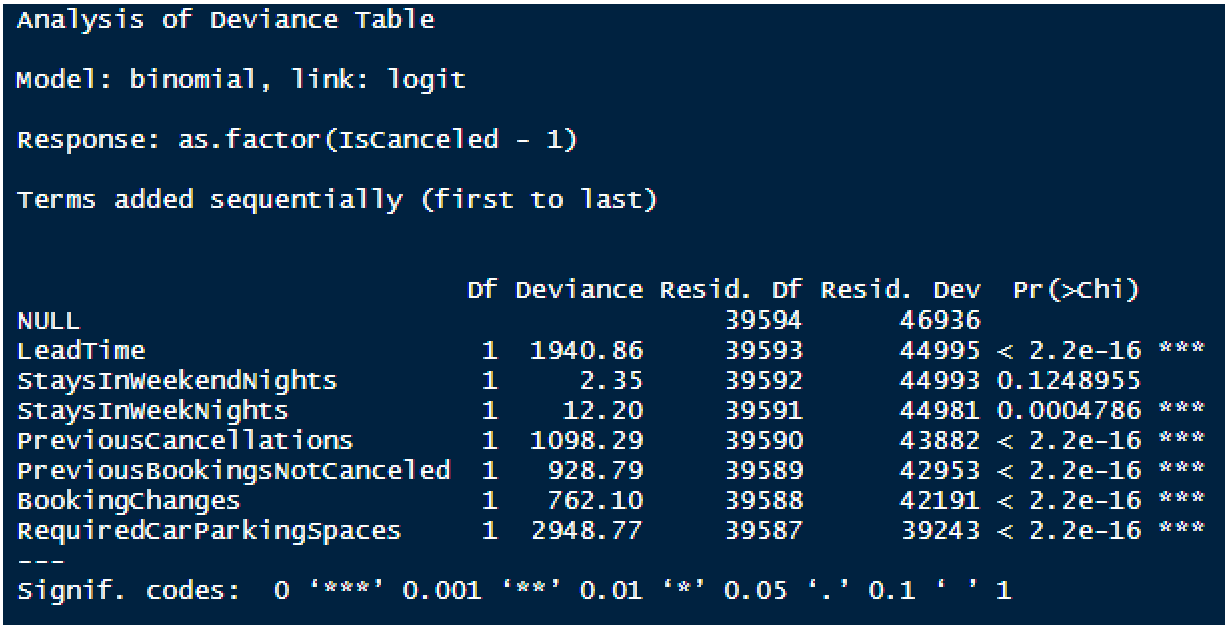


```
      LeadTime      StaysInWeekendNights         StaysInWeekNights      PreviousCancellations PreviousBookingsNotCanceled
      1.184419                  2.148216                  2.249318                   1.390112                    1.394018
BookingChanges  RequiredCarParkingSpaces
      1.017164                  1.000000
```

Logistic Model Collinearity

The results were very similar to the linear model. With StaysInWeekNights and StaysInWeekendNights being the only variables above 2.

Lastly, we will take the chi-squared statistics from the linear regression and analyze. Using the anova() function from the stats package, we can test the significance of our chi-squared value, which is shown below.

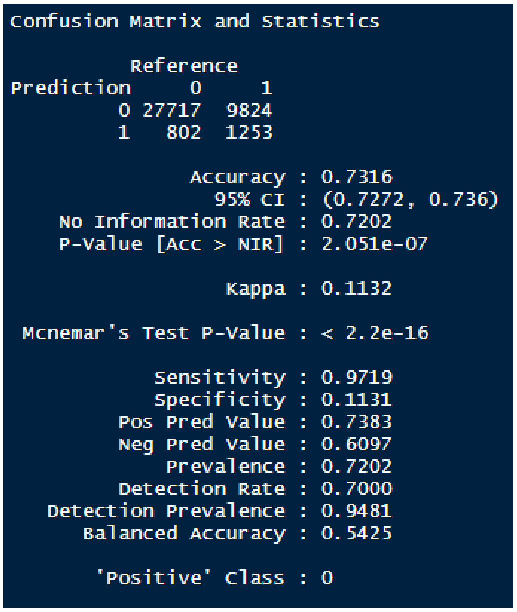


```
Analysis of Deviance Table

Model: binomial, link: logit

Response: as.factor(IsCanceled - 1)

Terms added sequentially (first to last)


                            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                        39594      46936              
LeadTime                     1  1940.86     39593      44995 < 2.2e-16 ***
StaysInWeekendNights         1     2.35     39592      44993 0.1248955    
StaysInWeekNights            1    12.20     39591      44981 0.0004786 ***
PreviousCancellations        1  1098.29     39590      43882 < 2.2e-16 ***
PreviousBookingsNotCanceled  1   928.79     39589      42953 < 2.2e-16 ***
BookingChanges               1   762.10     39588      42191 < 2.2e-16 ***
RequiredCarParkingSpaces     1  2948.77     39587      39243 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Chi-Squared Analysis of Logistic Regression

The chi-squared analysis breaks up the total chi-squared deviance and breaks it up into the variable that is responsible for the relationship. The p-values still have the same meaning, concluding that each variable except for StaysInWeekendNights is statistically significant.

Now that we have analyzed our linear and logistic regression models and concluded that there are statistical relationships embedded in the dataset, we can move onto predictive analytics. Starting with our linear model, we can create a prediction data set that uses the predict() function to get artificial numerical values for IsCanceled. Once predictions are made, we then create a confusion matrix which shows the predicted model values and the actual referential values in the data set. The results are shown below.

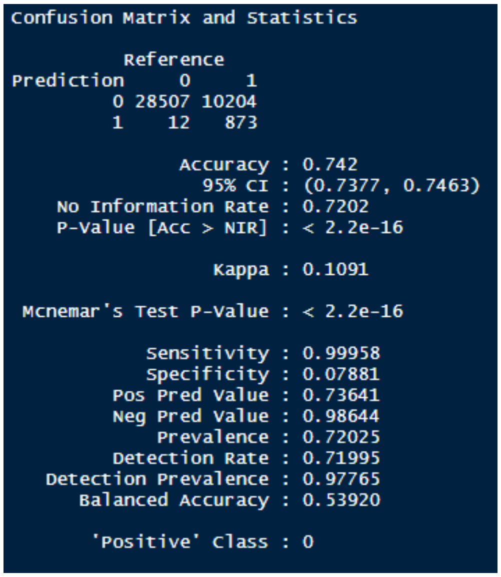


```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 27717  9824
         1   802  1253

               Accuracy : 0.7316
                 95% CI : (0.7272, 0.736)
    No Information Rate : 0.7202
    P-Value [Acc > NIR] : 2.051e-07

                  Kappa : 0.1132

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9719
            Specificity : 0.1131
         Pos Pred Value : 0.7383
         Neg Pred Value : 0.6097
             Prevalence : 0.7202
         Detection Rate : 0.7000
   Detection Prevalence : 0.9481
      Balanced Accuracy : 0.5425

       'Positive' Class : 0
```

Linear Model Confusion Matrix

Our linear model produced a confusion matrix which yielded 73.16% accuracy. Although this is good, it is not much higher than our No Information Rate of 72.02%. So, despite our model producing significant values, it was only 1.14% better than predicting values randomly.

Next, we will run the same analysis on our logistic regression. Its corresponding confusion matrix is shown below.



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 28507 10204
         1    12   873

               Accuracy : 0.742
                 95% CI : (0.7377, 0.7463)
    No Information Rate : 0.7202
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1091

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.99958
            Specificity : 0.07881
         Pos Pred Value : 0.73641
         Neg Pred Value : 0.98644
             Prevalence : 0.72025
         Detection Rate : 0.71995
   Detection Prevalence : 0.97765
      Balanced Accuracy : 0.53920

       'Positive' Class : 0
```

Logistical Model Confusion Matrix

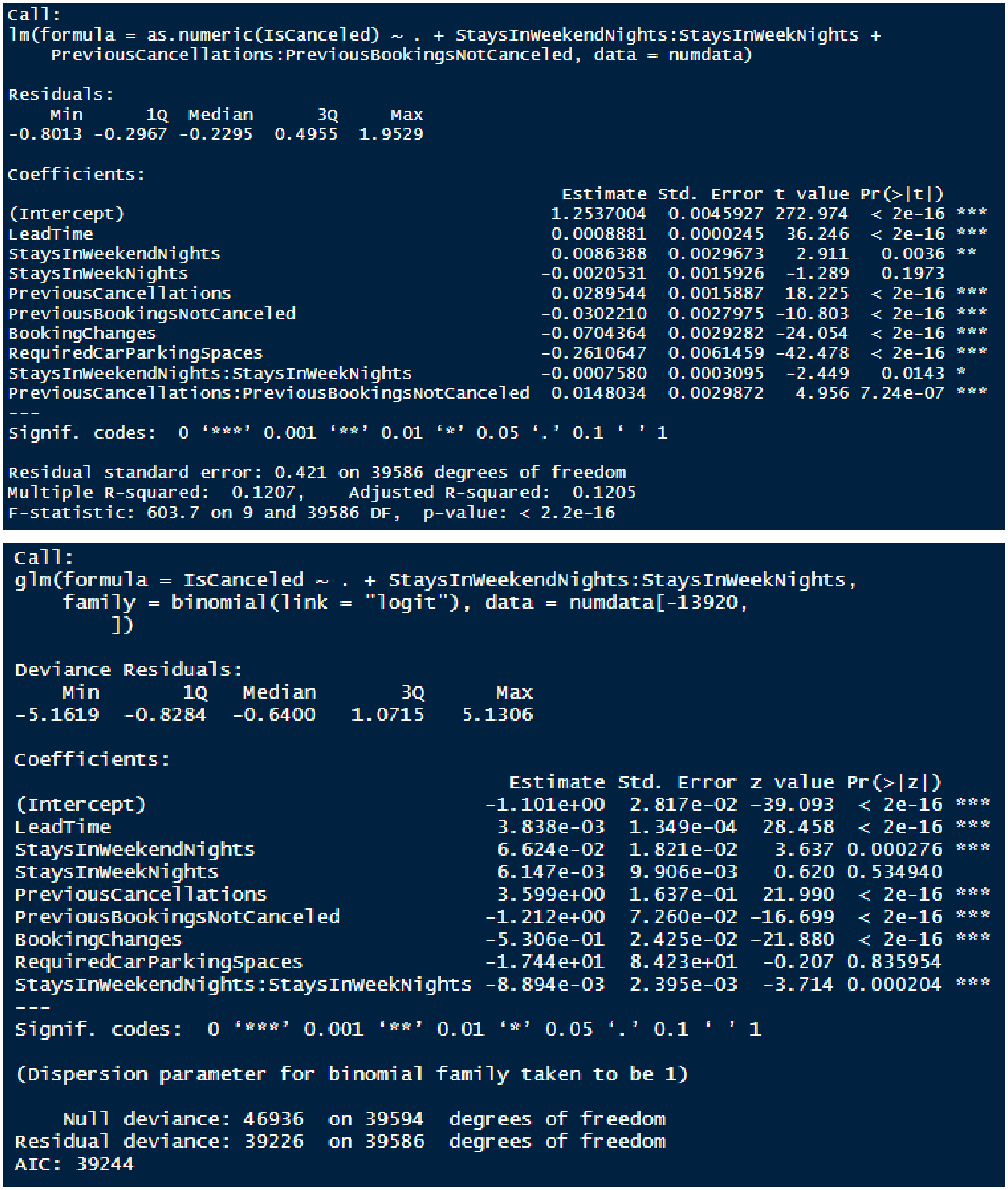
The confusion matrix showed that our model predicted IsCanceled with 74.2% accuracy. Although the model is significant, the difference between the accuracy and the No Information Rate (72.02%) is not very large, meaning that our model is only 2.18% better than predicting values randomly.

Shown by the confusion matrices, the logistic regression proved to be a stronger predictor of the IsCanceled variable. Not only did it account for twice the increase in accuracy, but it produced only 12 instances where the model predicted a reservation was going to be canceled when it was not. It is very important that this error be avoided; hotels would likely try to avoid this type I error over missing a cancellation prediction. It is much better to predict a non-cancellation that ends up being canceled (type II error) as opposed to predicting a cancellation when a guest does not cancel (type I error). For implementation scenarios, a guest who does not show up is significantly better than a guest arriving with no available rooms.

But overall, the dataset's numerical data was not the best predictor of IsCanceled. Although there were evident relationships in the data, they were not strong enough to significantly change the potential prediction of a hotel reservation being canceled by more than a few percentage points. Still, incorporating the logistic regression model may be a good idea to help avoid Type I error based on numerical input.

B. Integer Interactions

In this section, we will take a closer look at the interactions between the StaysInWeekNights/StaysInWeekendNights and the PreviousCancellations/PreviousBookingsNotCanceled variables. Because we found significant results from the previous section, we will add interaction terms to the existing linear and logistic models. The modified models are shown below (top: linear regression; bottom: logistic regression).

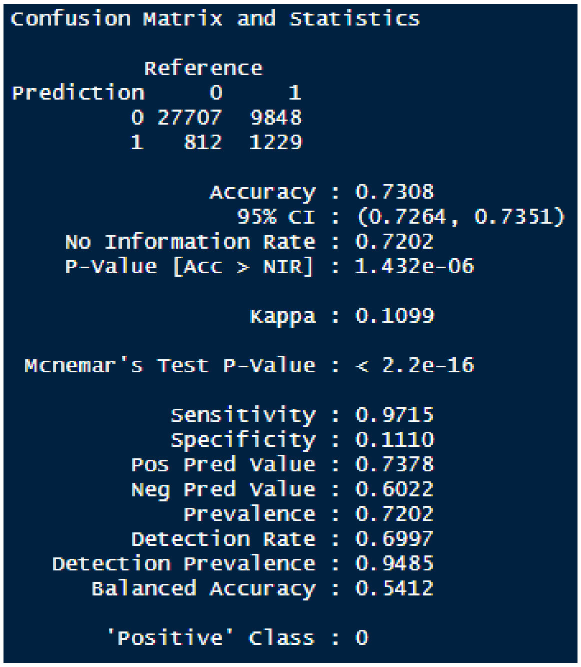


```
Call:
lm(formula = as.numeric(IsCanceled) ~ . + StaysInWeekendNights:StaysInWeekNights +
    PreviousCancellations:PreviousBookingsNotCanceled, data = numdata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8013 -0.2967 -0.2295  0.4955  1.9529

Coefficients:
                                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       1.2537004  0.0045927 272.974  < 2e-16 ***
LeadTime                                          0.0008881  0.0000245  36.246  < 2e-16 ***
StaysInWeekendNights                              0.0086388  0.0029673   2.911   0.0036 **
StaysInWeekNights                                -0.0020531  0.0015926  -1.289   0.1973
PreviousCancellations                             0.0289544  0.0015887  18.225  < 2e-16 ***
PreviousBookingsNotCanceled                      -0.0302210  0.0027975 -10.803  < 2e-16 ***
BookingChanges                                   -0.0704364  0.0029282 -24.054  < 2e-16 ***
RequiredCarParkingSpaces                         -0.2610647  0.0061459 -42.478  < 2e-16 ***
StaysInWeekendNights:StaysInWeekNights           -0.0007580  0.0003095  -2.449   0.0143 *
PreviousCancellations:PreviousBookingsNotCanceled  0.0148034  0.0029872   4.956 7.24e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.421 on 39586 degrees of freedom
Multiple R-squared:  0.1207,	Adjusted R-squared:  0.1205
F-statistic: 603.7 on 9 and 39586 DF,  p-value: < 2.2e-16
```

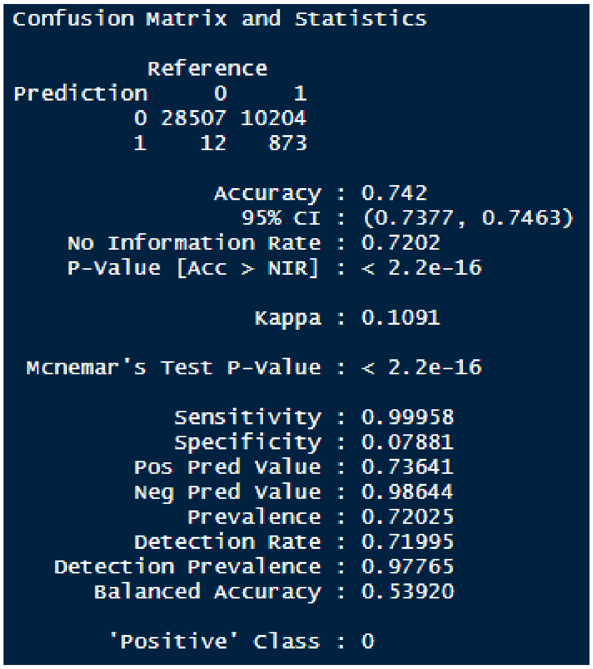
```
Call:
glm(formula = IsCanceled ~ . + StaysInWeekendNights:StaysInWeekNights,
    family = binomial(link = "logit"), data = numdata[-13920,
        ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.1619  -0.8284  -0.6400   1.0715   5.1306

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -1.101e+00  2.817e-02 -39.093  < 2e-16 ***
LeadTime                               3.838e-03  1.349e-04  28.458  < 2e-16 ***
StaysInWeekendNights                   6.624e-02  1.821e-02   3.637 0.000276 ***
StaysInWeekNights                      6.147e-03  9.906e-03   0.620 0.534940
PreviousCancellations                  3.599e+00  1.637e-01  21.990  < 2e-16 ***
PreviousBookingsNotCanceled           -1.212e+00  7.260e-02 -16.699  < 2e-16 ***
BookingChanges                        -5.306e-01  2.425e-02 -21.880  < 2e-16 ***
RequiredCarParkingSpaces              -1.744e+01  8.423e+01  -0.207 0.835954
StaysInWeekendNights:StaysInWeekNights -8.894e-03  2.395e-03  -3.714 0.000204 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46936  on 39594  degrees of freedom
Residual deviance: 39226  on 39586  degrees of freedom
AIC: 39244
```

Adjusted Model Summaries with Interactions (Top: Linear; Bottom: Logistic)

Notice that the logistic regression does not have the PreviousCancellations:PreviousBookingsNotCancelled term. The model was initially made with the interaction, but did not produce significant results. The implementation of the new interaction variables did not significantly change the numerical analysis depicted in the previous section, so we will exclude that from this analysis. Instead, we will jump right into predictions. Confusion matrices were created using the same methodology, with results shown below (left: linear; right: logistic).



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 27707  9848
         1   812  1229
                                          
               Accuracy : 0.7308          
                 95% CI : (0.7264, 0.7351)
    No Information Rate : 0.7202          
    P-Value [Acc > NIR] : 1.432e-06       
                                          
                  Kappa : 0.1099          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.9715          
            Specificity : 0.1110          
         Pos Pred Value : 0.7378          
         Neg Pred Value : 0.6022          
             Prevalence : 0.7202          
         Detection Rate : 0.6997          
   Detection Prevalence : 0.9485          
      Balanced Accuracy : 0.5412          
                                          
       'Positive' Class : 0               
```



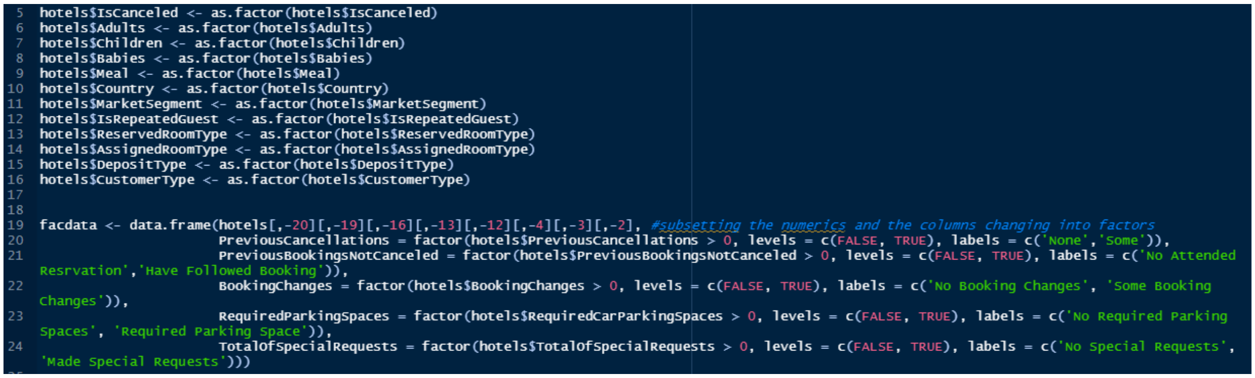
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 28507 10204
         1    12   873
                                          
               Accuracy : 0.742           
                 95% CI : (0.7377, 0.7463)
    No Information Rate : 0.7202          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.1091          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.99958         
            Specificity : 0.07881         
         Pos Pred Value : 0.73641         
         Neg Pred Value : 0.98644         
             Prevalence : 0.72025         
         Detection Rate : 0.71995         
   Detection Prevalence : 0.97765         
      Balanced Accuracy : 0.53920         
                                          
       'Positive' Class : 0               
```

Corresponding Confusion Matrices with Interactions (Left: Linear; Right: Logistic)

Although the interaction terms were statistically significant in the analysis of cancellations, they had no effect on the accuracy of the logistic confusion matrix and actually decreased the calculated accuracy of the linear confusion matrix. Consequently, we will not incorporate these results into our action insights.

C. Analyzing Factors Using Support Vector Machine

Moving away from numerics, we will now analyze the parts of our data that can be characterized as factors. The following code can be used to create a new subset of the hotels data with all of the natural factors and numerical factor transformations alike.

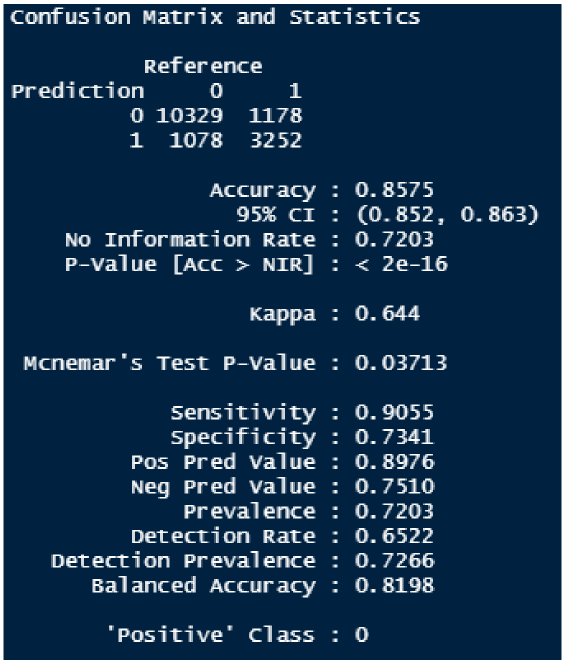


```r
hotels$IsCanceled <- as.factor(hotels$IsCanceled)
hotels$Adults <- as.factor(hotels$Adults)
hotels$Children <- as.factor(hotels$Children)
hotels$Babies <- as.factor(hotels$Babies)
hotels$Meal <- as.factor(hotels$Meal)
hotels$Country <- as.factor(hotels$Country)
hotels$MarketSegment <- as.factor(hotels$MarketSegment)
hotels$IsRepeatedGuest <- as.factor(hotels$IsRepeatedGuest)
hotels$ReservedRoomType <- as.factor(hotels$ReservedRoomType)
hotels$AssignedRoomType <- as.factor(hotels$AssignedRoomType)
hotels$DepositType <- as.factor(hotels$DepositType)
hotels$CustomerType <- as.factor(hotels$CustomerType)


facdata <- data.frame(hotels[,-20][,-19][,-16][,-13][,-12][,-4][,-3][,-2], #subsetting the numerics and the columns changing into factors
                      PreviousCancellations = factor(hotels$PreviousCancellations > 0, levels = c(FALSE, TRUE), labels = c('None','Some')),
                      PreviousBookingsNotCanceled = factor(hotels$PreviousBookingsNotCanceled > 0, levels = c(FALSE, TRUE), labels = c('No Attended Resrvation','Have Followed Booking')),
                      BookingChanges = factor(hotels$BookingChanges > 0, levels = c(FALSE, TRUE), labels = c('No Booking Changes', 'Some Booking Changes')),
                      RequiredParkingSpaces = factor(hotels$RequiredCarParkingSpaces > 0, levels = c(FALSE, TRUE), labels = c('No Required Parking Spaces', 'Required Parking Space')),
                      TotalOfSpecialRequests = factor(hotels$TotalOfSpecialRequests > 0, levels = c(FALSE, TRUE), labels = c('No Special Requests', 'Made Special Requests')))
```

Factor Subsetting/Cleaning

With a data frame that encompasses all of the possible factors, we can see if there is any predictive way of determining cancellations based on the rest of the variables. We will begin by partitioning the data set into two subsets, one for model training and the other for model testing. To do this, we will use the createDataPartition() function from the caret package. Specifically, the function randomly generates indices for a data set which correspond to the same percentage of outputs. It also can break the data sets into proportions. Consequently, we will create a training data set with 60% of the cancellations and non-cancellations and a testing data set with the remaining 40% of data observations.

Using the training set, we can generate a support vector machine model. Now that we have a model making predictions based on the given categorical data, we can use it to analyze the testing data set. Using the predict function, we first obtain the predicted values of the testing data set based on the other parameters. Then, we compare the predictions to the actual IsCanceled variable in the testing data set. The resulting confusion matrix from their comparison is shown below.



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 10329  1178
         1  1078  3252

               Accuracy : 0.8575
                 95% CI : (0.852, 0.863)
    No Information Rate : 0.7203
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.644

 Mcnemar's Test P-Value : 0.03713

            Sensitivity : 0.9055
            Specificity : 0.7341
         Pos Pred Value : 0.8976
         Neg Pred Value : 0.7510
             Prevalence : 0.7203
         Detection Rate : 0.6522
   Detection Prevalence : 0.7266
      Balanced Accuracy : 0.8198

       'Positive' Class : 0
```

SVM Confusion Matrix

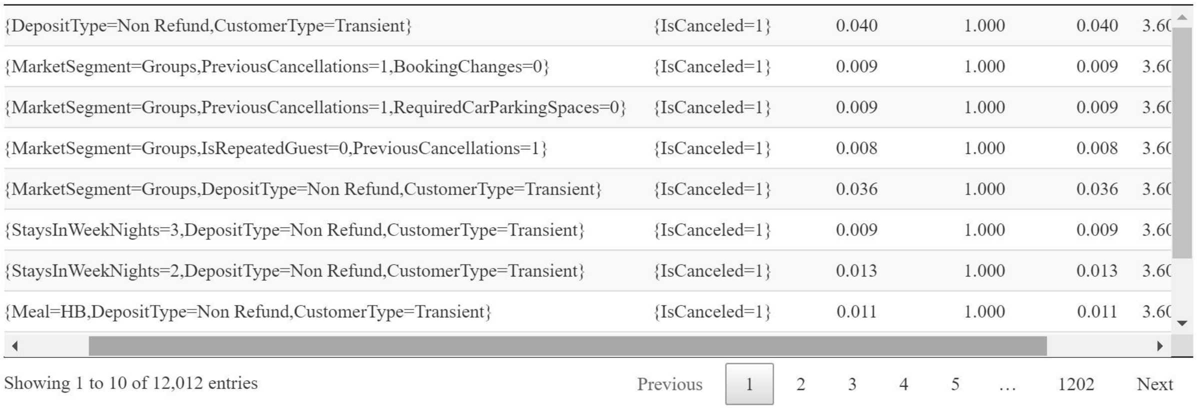
The resulting confusion matrix is promising, with an overall accuracy of 85.75% and significant p-values supporting the underlying relationship. This is 13.72% better than the no information rate (ability to predict values without model), giving us a promising model to analyze hotel cancellations.

Despite the strong accuracy, the type I error is relatively high for this model. The model predicted 9.5% of non-cancellations incorrectly, leading to a problem when using this model. Although incorporating it may accurately predict cancellations and non-cancellations, there is still a strong chance that a predicted cancellation will not actually be a cancellation.

D. Association Rule Mining

We deployed association rule mining techniques to find out those instances that are more likely to get cancelled. We set the confidence to various values to see which factors were common in the instances that lead to the cancellation of the booking.

The screenshot below shows the result set when the confidence was set to 0.6 and the support was set to 0.008. We see instances when “Deposit type = Non Refund” and “Customer Type = Transient” that were common in the result set. This helps us just get an idea on which factors could be important decision makers. The factors found in association mining were also backed by other models.

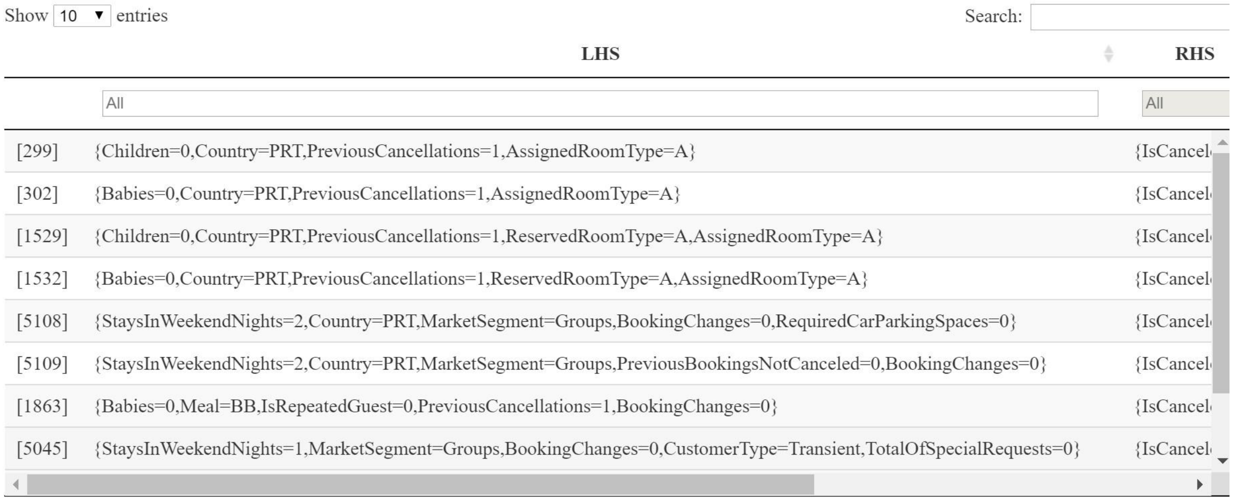


{DepositType=Non Refund,CustomerType=Transient}	{IsCanceled=1}	0.040	1.000	0.040	3.60
{MarketSegment=Groups,PreviousCancellations=1,BookingChanges=0}	{IsCanceled=1}	0.009	1.000	0.009	3.60
{MarketSegment=Groups,PreviousCancellations=1,RequiredCarParkingSpaces=0}	{IsCanceled=1}	0.009	1.000	0.009	3.60
{MarketSegment=Groups,IsRepeatedGuest=0,PreviousCancellations=1}	{IsCanceled=1}	0.008	1.000	0.008	3.60
{MarketSegment=Groups,DepositType=Non Refund,CustomerType=Transient}	{IsCanceled=1}	0.036	1.000	0.036	3.60
{StaysInWeekNights=3,DepositType=Non Refund,CustomerType=Transient}	{IsCanceled=1}	0.009	1.000	0.009	3.60
{StaysInWeekNights=2,DepositType=Non Refund,CustomerType=Transient}	{IsCanceled=1}	0.013	1.000	0.013	3.60
{Meal=HB,DepositType=Non Refund,CustomerType=Transient}	{IsCanceled=1}	0.011	1.000	0.011	3.60

Showing 1 to 10 of 12,012 entries

Association Rule Mining with confidence= 0.9 & support= 0.008

For the next screenshot the confidence was set to 0.9 and the support was set to 0.008. We now found instances where the Country was Portugal, the Assigned Room type : 'A' and the customer had a history of cancellation, would be more likely to cancel their booking.



	LHS	RHS
	All	All
[299]	{Children=0,Country=PRT,PreviousCancellations=1,AssignedRoomType=A}	{IsCancel
[302]	{Babies=0,Country=PRT,PreviousCancellations=1,AssignedRoomType=A}	{IsCancel
[1529]	{Children=0,Country=PRT,PreviousCancellations=1,ReservedRoomType=A,AssignedRoomType=A}	{IsCancel
[1532]	{Babies=0,Country=PRT,PreviousCancellations=1,ReservedRoomType=A,AssignedRoomType=A}	{IsCancel
[5108]	{StaysInWeekendNights=2,Country=PRT,MarketSegment=Groups,BookingChanges=0,RequiredCarParkingSpaces=0}	{IsCancel
[5109]	{StaysInWeekendNights=2,Country=PRT,MarketSegment=Groups,PreviousBookingsNotCanceled=0,BookingChanges=0}	{IsCancel
[1863]	{Babies=0,Meal=BB,IsRepeatedGuest=0,PreviousCancellations=1,BookingChanges=0}	{IsCancel
[5045]	{StaysInWeekendNights=1,MarketSegment=Groups,BookingChanges=0,CustomerType=Transient,TotalOfSpecialRequests=0}	{IsCancel

Association Rule Mining with confidence= 0.9 & support= 0.008

VII. Actionable Insights

On the basis of above data analysis, our team has come up with the following actionable insights as a proposal to reduce the number of cancellations.

- Use numeric data and logistic regression model to predict a good amount of guaranteed cancellations
- Take in the other given data and use a support vector machine model to analyze the factors and predict the rest of the cancellations/non-cancellations. However, approximately 25% of customers that are predicted to cancel do not end up canceling, so leave about one fourth of the rooms reserved. Furthermore, reach out to canceled predictions in advance and see if they have decided to cancel, then re-analyze the svm model to potentially increase accuracy.
- The world map diagram shows that most of the cancellations are from European countries, especially from Portugal. Whenever the hotel receives a reservation from this region, they should reach out to those customers to notify them regarding their upcoming bookings ahead of time. If they intend to cancel, they can do it way ahead of the arrival date.
- Make an effort to get special requests from the customers, to be very safe , at least 4 or more. Customers making special requests hardly cancel their bookings
- When people make an effort to change their bookings greet them with more importance because it shows inclination towards coming to stay in the hotel. Which means there is less chance of cancellation

VIII. Future Exploration

- The data set is highly imbalanced. So, in future we would focus on collecting more data (specifically Instances for cancellations) to balance both parts of the variables (Cancelled or not) to understand the factors affecting hotel business better. We would also like to look at the reviews, testimony or customer experience column and run sentiment analysis to understand if there are any other issues apart from these variables that we have worked on.
- We can see a huge number of cancellations from Portugal. This is an unlikely situation because there are guests coming from all around the world. We would like to focus on the data in particular from that country and try to reduce the cancellations.
- In this data exploration and analysis, we have discovered the most unlikely situation- 96% cancellations for non-refundable bookings, which forces us to think that there are some other major issues behind these cancellations.

IX. Acknowledgement

We are sincerely grateful to our professors Jeffery S. Saltz, Erik Anderson and Teaching Assistant Abhijit Shamkant Gokhale for guiding us in the best possible way for completion of this project. As a coursework, we performed all the analysis by ourselves. We believe everything presented in the report is true to the best of our knowledge.

X. References

- Jeffrey S. Saltz and Jeffrey Morgan Stanton, “Data Science for Business With R”; ISBN 978-1544370453