Hotel Cancellation prediction model

PROJECT REPORT

Submitted in partial fulfilment of the requirements of the course IST 687 Introduction to Data Science

at

Syracuse University

Information Science School at Syracuse University

Submitted by

Ayyappa Bhuma	258426982
Khushboo Saxena	892552395
Sreelakshmy Cheruvally	387397629

TABLE OF CONTENTS

TITLE **PAGE NO**

1.INTRODUCTION

In this present fast paced world data has become an integral part of everyone's day to day lives. Tons of data is being generated and processed everyday. As we move on the complexity and the volume of data is increasing at an unprecedented scale. To harness these large volumes of data and to draw insight from this ubiquitously available data we use various scientific methods, algorithms and many other processes, which in simple words is Data Science.

Hotels are an industry which has an enormous amount of data being produced everyday with it's bookings, cancellations, reservations, payments etc…Hotels to improve their revenue use Data Science to analyze their past data to understand why customers cancel their reservation and also to predict the chances of cancellation. By analyzing various factors for instance reservation of desired room, applicable cancellation fee, parking facility etc…

1.1 Objective and scope

The main objective of the project is to understand the main drivers for cancellation and reservations of hotel bookings. These factors would eventually help us to come up with a model which will help the hotels to predict the possibility of cancellations. The real life data of a hotel is being analyzed here. The whole process from data cleaning to predicting a model is being used in this project.
Data visualizations is one of the key methods used to analyze the trends. As per the analysis the best possible accurate model will be recommended for the hotels to predict their cancellation chances.

1.2 Problem Statement

The number of cancellations in hotel bookings are increasing, so there is a need to find the trend behind this cancellation. Finding the trend will help to retain the customers from the cancellations and by providing them what they desire out of their stay at the hotel. This will eventually help the hotel industry to improve their revenue by recommending a model by analyzing the available data. A model which is capable enough to predict the customers most likely to cancel their bookings.

2. BUSINESS QUESTIONS

1. How is the cancellation affected when the guests were not allocated the reserved room type.
2. What is the probability of cancellations if a guest has a history of cancellations?
3. Which booking method has the highest likelihood of cancellations?
4. In which part of the world is it most likely that a reservation will be canceled?

3. SUMMARY OF DATA SET

The data set consists of real-life hotel stay data. It has 40,060 rows and 20 columns, with each row representing a hotel booking. The target variable in the data set being IsCanceled which is a categorical variable with 1 for booking canceled and 0 for booking not canceled. It also has other personal information variables such as adults, babies, children and hotel attributes such as meal, previous bookings, reserved and assigned rooms etc…

These variables help us understand the dataset better and also the relevant variables for the model building. It will also give a glimpse of what the customers desire out of their stay at the hotel.

3.1 Data set in detail

This description is in reference to the *Final Project* pdf shared by the instructor

a) **IsCanceled:** Categorical Value indicating if the booking was cancelled or not (1 for cancelled booking and 0 for the bookings not cancelled)

b) **LeadTime:** Integer, Number of days that elapsed between the entering date of the booking into and the arrival date

c) **StaysInWeekendNights:** Integer, Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

d) **StaysInWeekNights:** Integer, Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel

e) **Adults:** Integer, Number of adults

f) **Children:** Integer, Number of children

g) **Babies:** Integer, Number of babies

h) **Meal:** Categorical, Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC - no meal package; BB - Bed and Breakfast; HB - Half board (breakfast and one other meal - usually dinner); FB - -Full board (breakfast, lunch and dinner)

i) **Country:** Categorical, Country of origin. Categories are represented in the ISO 3155 - 3:2013 format

j) **MarketSegment:** Categorical, Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

k) **IsRepeatedGuest:** Categorical, Value indicating if the booking name was from a repeated guest (1) or not (0)

l) **PreviousCancellations:** Integer, Number of previous bookings that were cancelled by the customer prior to the current booking

m) **PreviousBookingsNotCanceled:** Integoer, Number of previous bookings not cancelled by the customer prior to the current booking

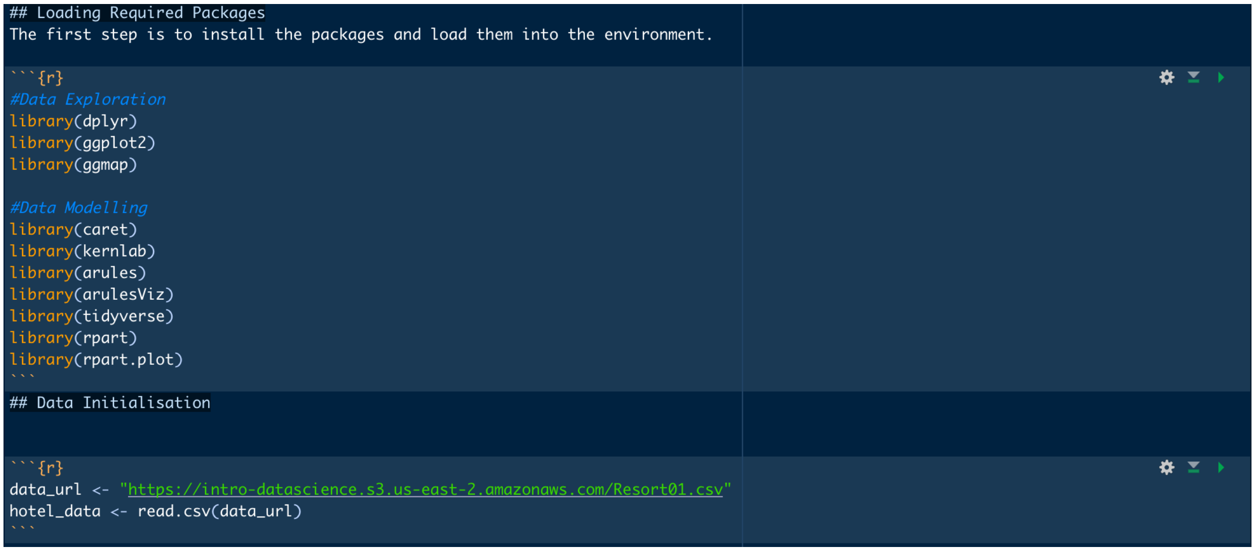
n) **ReservedRoomType:** Categorical, Code of room type preserved. Code is presented instead of designation for anonymity reasons

4. DATA EXPLORATION

Data exploration is the process of uncovering the characteristics and insights from the data using various tools, processes and visualization techniques. Here we follow multiple steps to achieve the end result of building a model.

4.1 Data initialization

Data initialization are the few steps to be followed in order to make the given data set in a readable or understandable form. Here in this project the given dataset is in comma-separated values (csv) format. First this data set requires to be loaded into a dataframe. In order to carry out this conversion of csv files, there are a few necessary libraries which need to be loaded. Figure 4.1 depicts the libraries loaded and the code required to load the csv file into a data frame.



~~~
## Loading Required Packages
The first step is to install the packages and load them into the environment.

```{r}
#Data Exploration
library(dplyr)
library(ggplot2)
library(ggmap)

#Data Modelling
library(caret)
library(kernlab)
library(arules)
library(arulesViz)
library(tidyverse)
library(rpart)
library(rpart.plot)
~~~
Data Initialisation

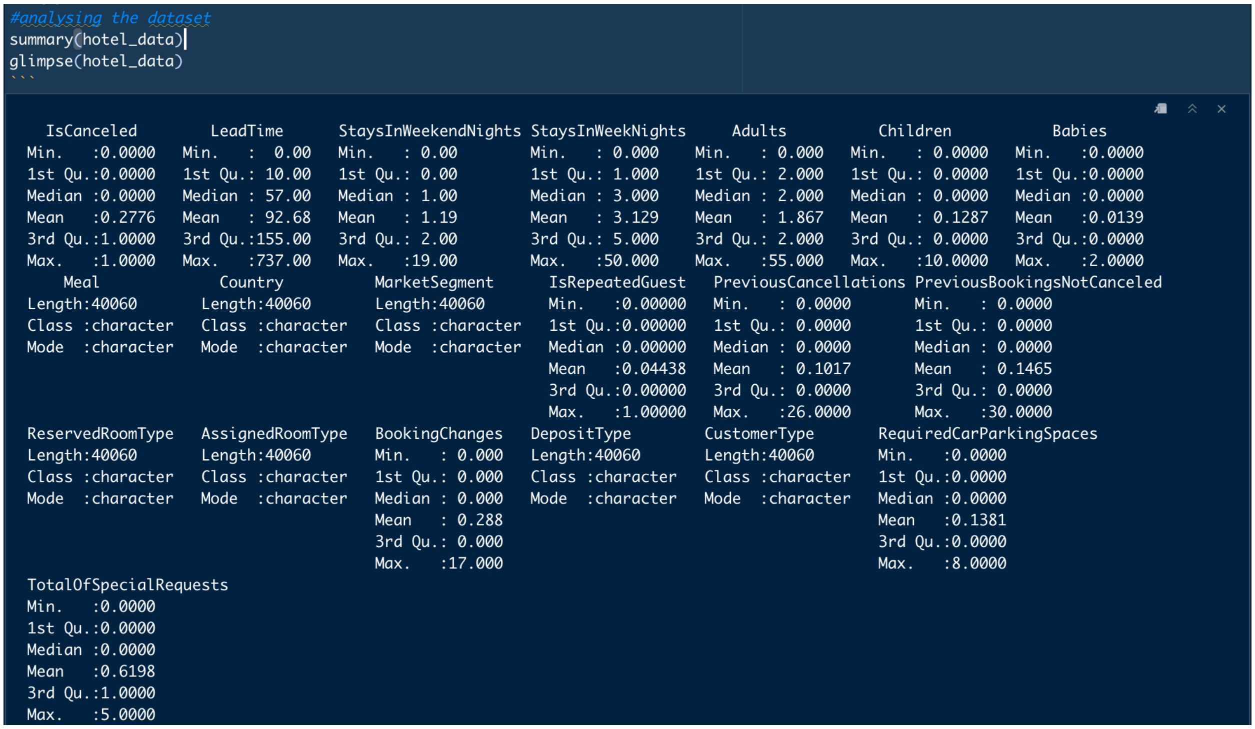
```{r}
data_url <- "https://intro-datascience.s3.us-east-2.amazonaws.com/Resort01.csv"
hotel_data <- read.csv(data_url)
```
```

Figure 4.1
```

## 4.2 Data cleaning

Data cleaning is one of the most important steps before analyzing the data. This is the process where we remove the unwanted data and help in filtering out the relevant data required for analysis. The two cleaning processes used here in this data are to check if the variables  have any NULL values in the data and the other is to check the data types of the variables. After checking the data types of variables, all the variables are labelled as integers though there are some categorical variables. And during the analysis the categorical variables are converted to factor type.

## 4.3 Data Analysis

After completing the cleaning part of the dataset now the data is used for analysing and drawing insights from it. To gather some general insights from the data set we use some functions such as summary() and glimpse() to get some statistical insights from the data set. Figure 4.2 describes the summary(hotel_data).Figure 4.3 depicts the glimpse(hotel_data).After that we also use the method of data visualization to better understand the behaviour of the variables in the data set.

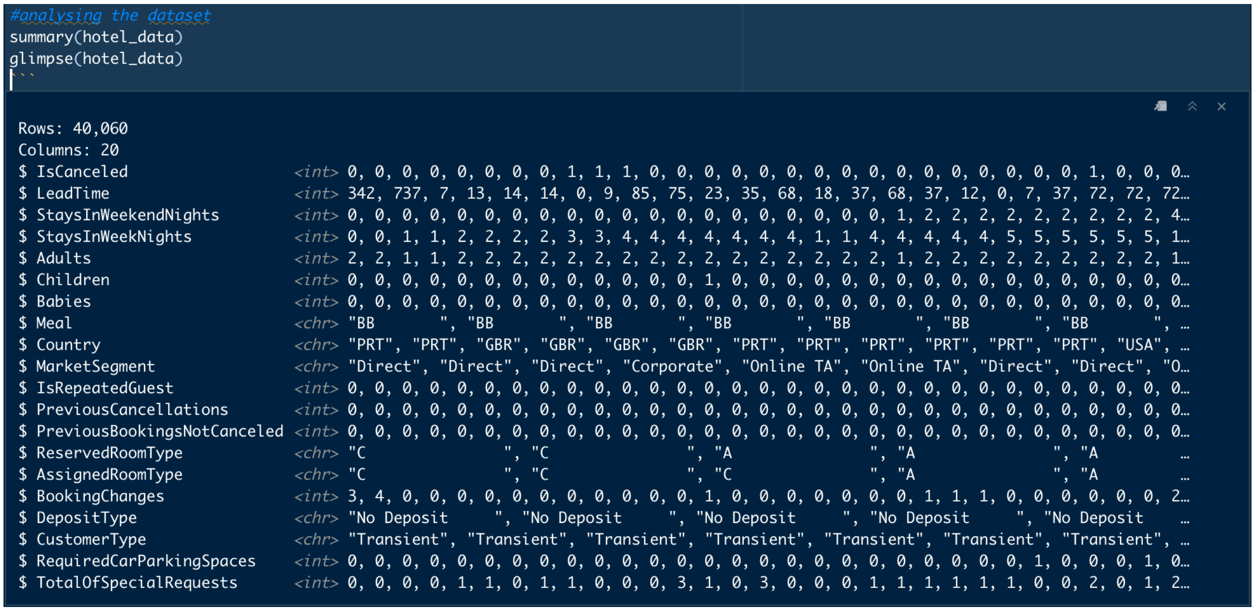


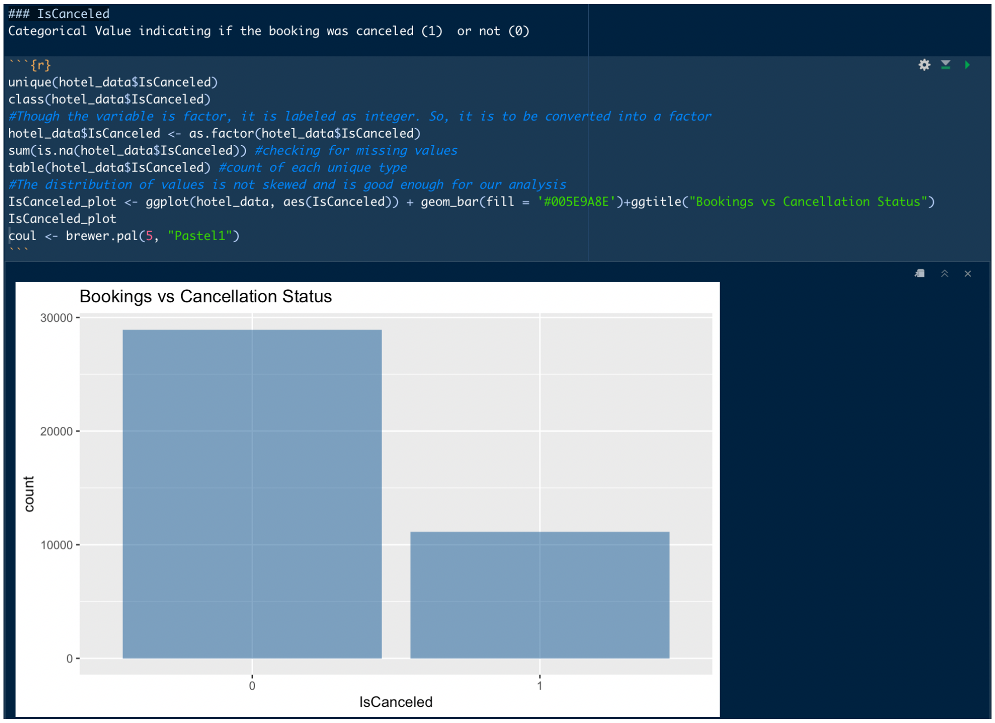
```
#analysing the dataset
summary(hotel_data)
glimpse(hotel_data)
```

```
   IsCanceled       LeadTime      StaysInWeekendNights StaysInWeekNights     Adults          Children          Babies      
 Min.   :0.0000   Min.   :  0.00   Min.   : 0.00        Min.   : 0.000    Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.: 10.00   1st Qu.: 0.00        1st Qu.: 1.000    1st Qu.: 2.000   1st Qu.: 0.0000   1st Qu.:0.0000  
 Median :0.0000   Median : 57.00   Median : 1.00        Median : 3.000    Median : 2.000   Median : 0.0000   Median :0.0000  
 Mean   :0.2776   Mean   : 92.68   Mean   : 1.19        Mean   : 3.129    Mean   : 1.867   Mean   : 0.1287   Mean   :0.0139  
 3rd Qu.:1.0000   3rd Qu.:155.00   3rd Qu.: 2.00        3rd Qu.: 5.000    3rd Qu.: 2.000   3rd Qu.: 0.0000   3rd Qu.:0.0000  
 Max.   :1.0000   Max.   :737.00   Max.   :19.00        Max.   :50.000    Max.   :55.000   Max.   :10.0000   Max.   :2.0000  
     Meal             Country          MarketSegment        IsRepeatedGuest    PreviousCancellations PreviousBookingsNotCanceled
 Length:40060       Length:40060       Length:40060       Min.   :0.00000   Min.   : 0.0000       Min.   : 0.0000            
 Class :character   Class :character   Class :character   1st Qu.:0.00000   1st Qu.: 0.0000       1st Qu.: 0.0000            
 Mode  :character   Mode  :character   Mode  :character   Median :0.00000   Median : 0.0000       Median : 0.0000            
                                                          Mean   :0.04438   Mean   : 0.1017       Mean   : 0.1465            
                                                          3rd Qu.:0.00000   3rd Qu.: 0.0000       3rd Qu.: 0.0000            
                                                          Max.   :1.00000   Max.   :26.0000       Max.   :30.0000            
 ReservedRoomType   AssignedRoomType   BookingChanges   DepositType        CustomerType       RequiredCarParkingSpaces
 Length:40060       Length:40060       Min.   : 0.000   Length:40060       Length:40060       Min.   :0.0000          
 Class :character   Class :character   1st Qu.: 0.000   Class :character   Class :character   1st Qu.:0.0000          
 Mode  :character   Mode  :character   Median : 0.000   Mode  :character   Mode  :character   Median :0.0000          
                                       Mean   : 0.288                                         Mean   :0.1381          
                                       3rd Qu.: 0.000                                         3rd Qu.:0.0000          
                                       Max.   :17.000                                         Max.   :8.0000          
 TotalOfSpecialRequests
 Min.   :0.0000        
 1st Qu.:0.0000        
 Median :0.0000        
 Mean   :0.6198        
 3rd Qu.:1.0000        
 Max.   :5.0000        
```

Figure 4.2

summary(hotel_data) gives the min,max,mean,median and the 3 quartiles of the the dataset



```
#analysing the dataset
summary(hotel_data)
glimpse(hotel_data)
```
```

```
Rows: 40,060
Columns: 20
$ IsCanceled                  <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0…
$ LeadTime                    <int> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, 23, 35, 68, 18, 37, 68, 37, 12, 0, 7, 37, 72, 72, 72…
$ StaysInWeekendNights        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 2, 2, 2, 2, 2, 2, 2, 4…
$ StaysInWeekNights           <int> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 1, 1, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 1…
$ Adults                      <int> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1…
$ Children                    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ Babies                      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ Meal                        <chr> "BB       ", "BB       ", "BB       ", "BB       ", "BB       ", "BB       ", "BB       ", …
$ Country                     <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "GBR", "PRT", "PRT", "PRT", "PRT", "PRT", "PRT", "USA", …
$ MarketSegment               <chr> "Direct", "Direct", "Direct", "Corporate", "Online TA", "Online TA", "Direct", "Direct", "O…
$ IsRepeatedGuest             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ PreviousCancellations       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ PreviousBookingsNotCanceled <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ ReservedRoomType            <chr> "C                  ", "C                  ", "A                  ", "A                  ", "A        …
$ AssignedRoomType            <chr> "C                  ", "C                  ", "C                  ", "A                  ", "A        …
$ BookingChanges              <int> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 2…
$ DepositType                 <chr> "No Deposit     ", "No Deposit     ", "No Deposit     ", "No Deposit     ", "No Deposit     …
$ CustomerType                <chr> "Transient", "Transient", "Transient", "Transient", "Transient", "Transient", "Transient", …
$ RequiredCarParkingSpaces    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0…
$ TotalOfSpecialRequests      <int> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 3, 1, 0, 3, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 2, 0, 1, 2…
```

Figure 4.3

glimpse (hotel_data) gives the number of rows,columns,data type and the first few values



```
### IsCanceled
Categorical Value indicating if the booking was canceled (1)  or not (0)

```{r}
unique(hotel_data$IsCanceled)
class(hotel_data$IsCanceled)
#Though the variable is factor, it is labeled as integer. So, it is to be converted into a factor
hotel_data$IsCanceled <- as.factor(hotel_data$IsCanceled)
sum(is.na(hotel_data$IsCanceled)) #checking for missing values
table(hotel_data$IsCanceled) #count of each unique type
#The distribution of values is not skewed and is good enough for our analysis
IsCanceled_plot <- ggplot(hotel_data, aes(IsCanceled)) + geom_bar(fill = '#005E9A8E')+ggtitle("Bookings vs Cancellation Status")
IsCanceled_plot
coul <- brewer.pal(5, "Pastel1")
```
```

Figure 4.4
```

In the figure 4.4 we can observe that the target variable IsCanceled has been explored using various functions. The unique() is used to know the unique categories of the attribute which in this case is 0 for not canceled and 1 for canceled. The class() helps in understanding the type of the variable. The as.factor is used to convert the categorical variable to a factor variable. is.na() is used to find the NULL values in the variables. table() gives the count of the unique values available in the variable. Data visualization is the method used to gather the general trend of the variable.

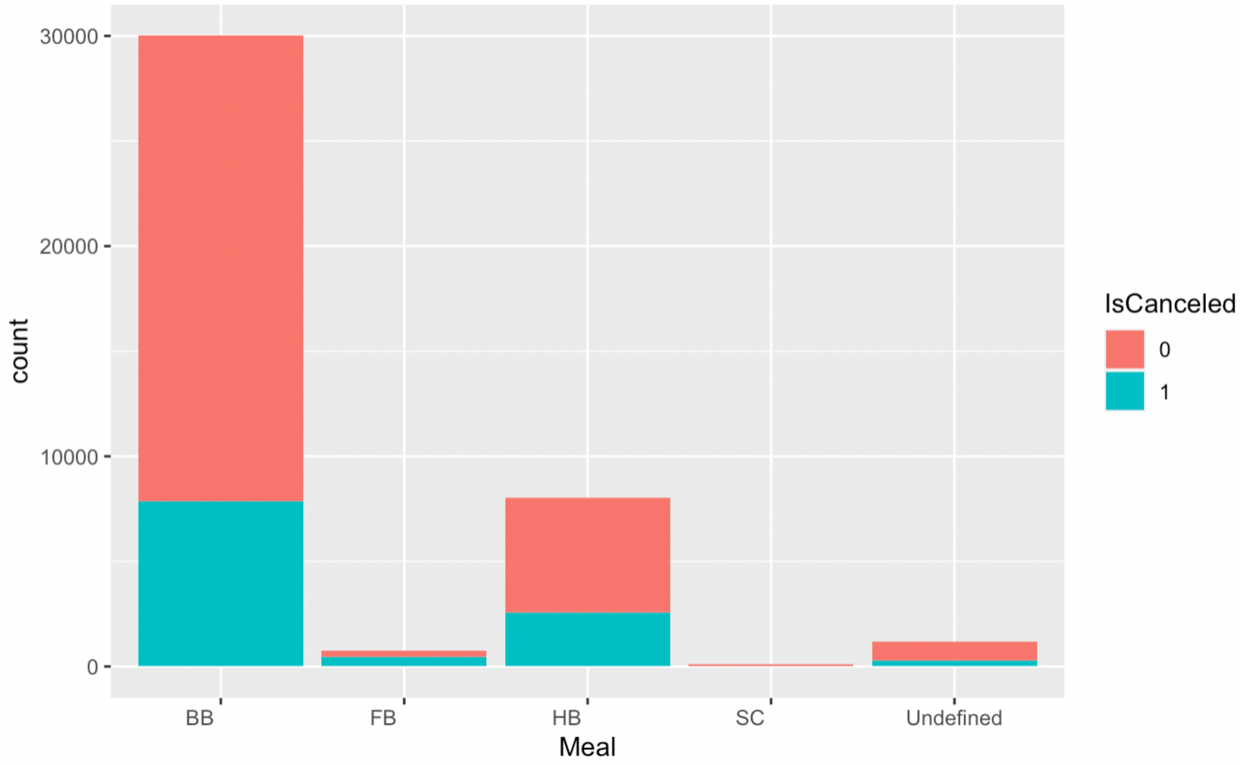
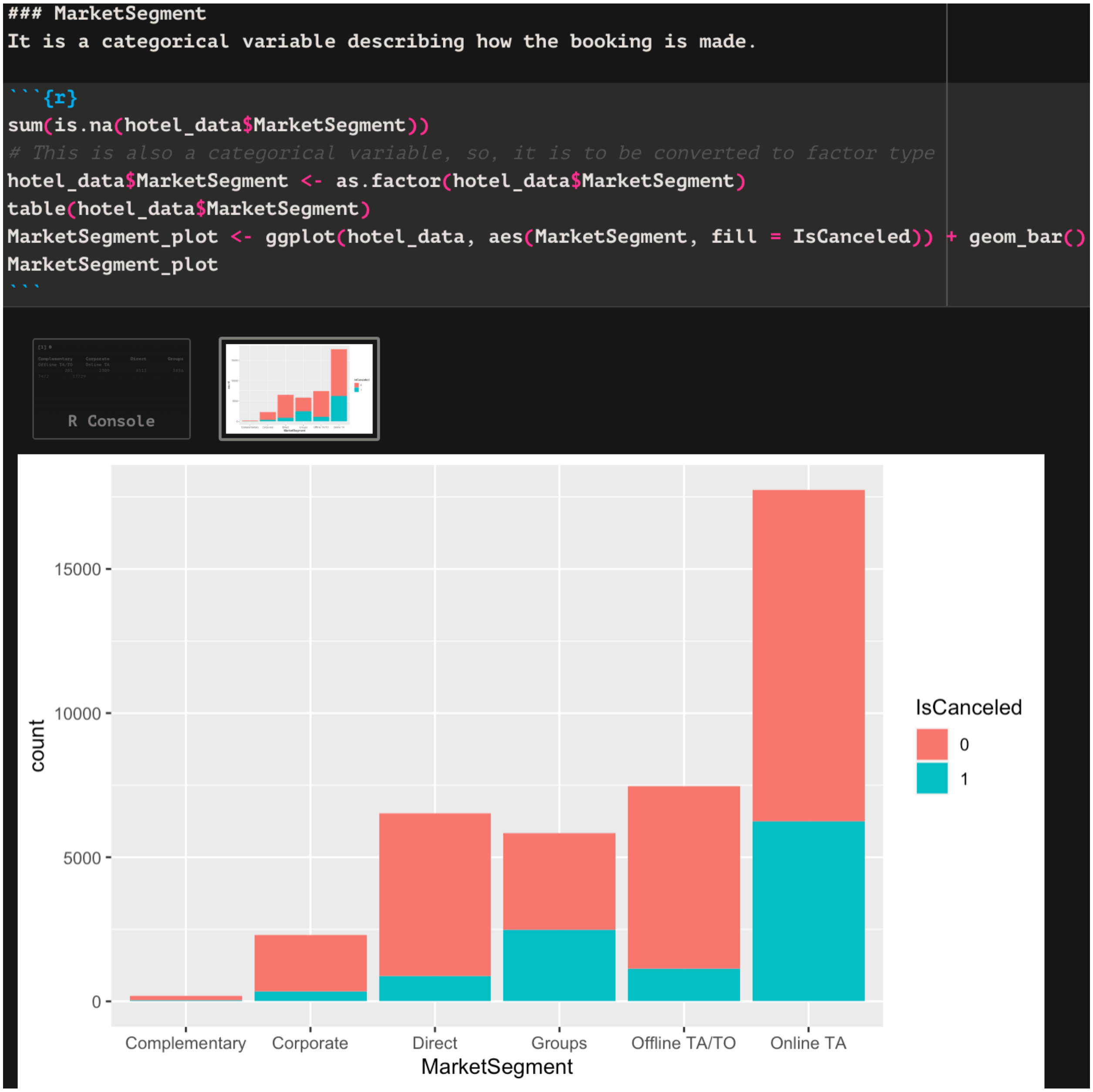


Figure 4.5

From Figure 4.5 we can observe that Bed&Breakfast and Half-breakfast are the most preferred meal plans but also have the maximum number of cancellations.

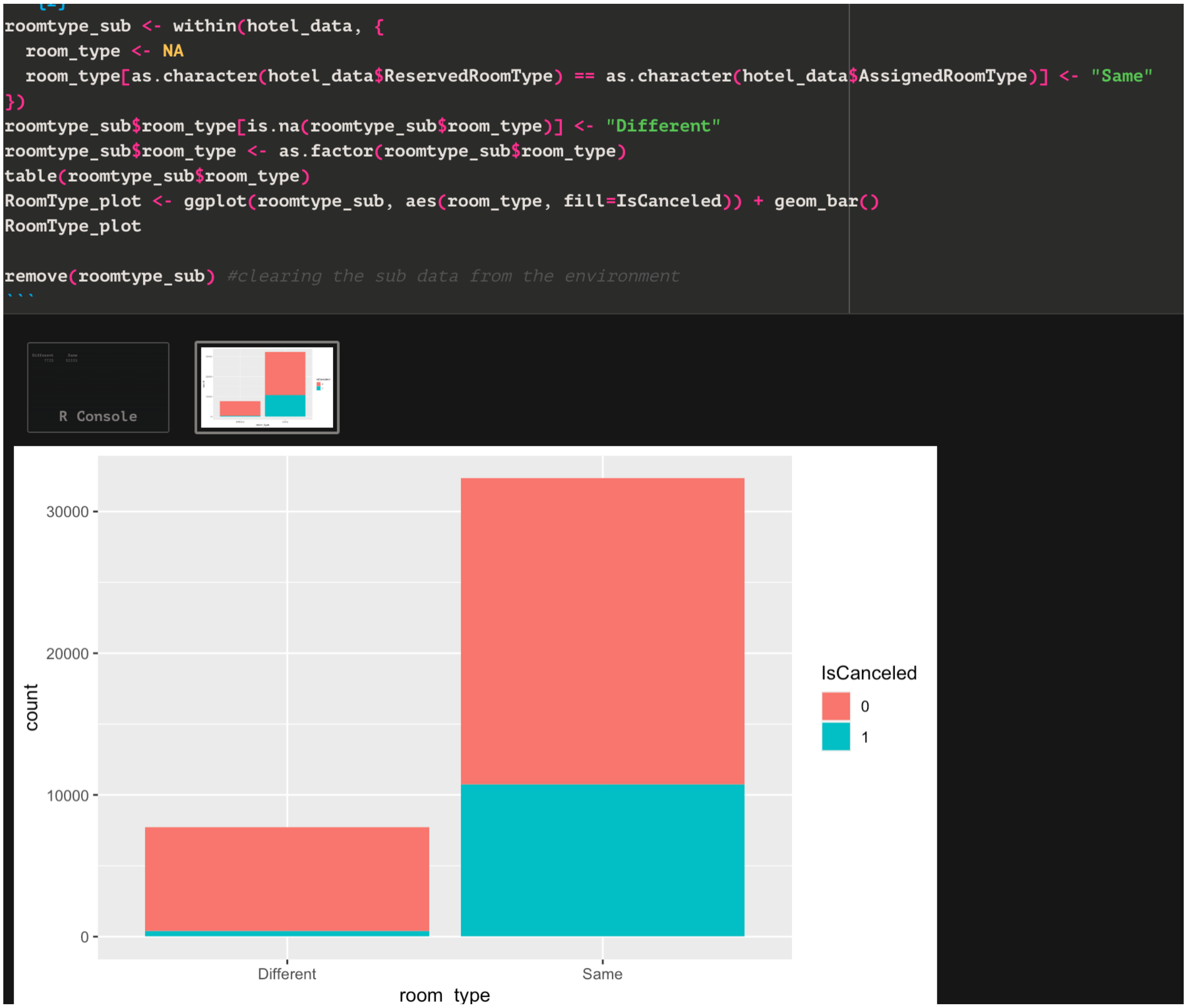


~~~
### MarketSegment
It is a categorical variable describing how the booking is made.

```{r}
sum(is.na(hotel_data$MarketSegment))
# This is also a categorical variable, so, it is to be converted to factor type
hotel_data$MarketSegment <- as.factor(hotel_data$MarketSegment)
table(hotel_data$MarketSegment)
MarketSegment_plot <- ggplot(hotel_data, aes(MarketSegment, fill = IsCanceled)) + geom_bar()
MarketSegment_plot
```
~~~

Figure 4.6

In figure 4.6 we notice that the maximum number of bookings have been made through Online TA and then groups and least bookings are made via complementary and corporate. Most of the cancellations are made when the guests make booking through online TA and groups.

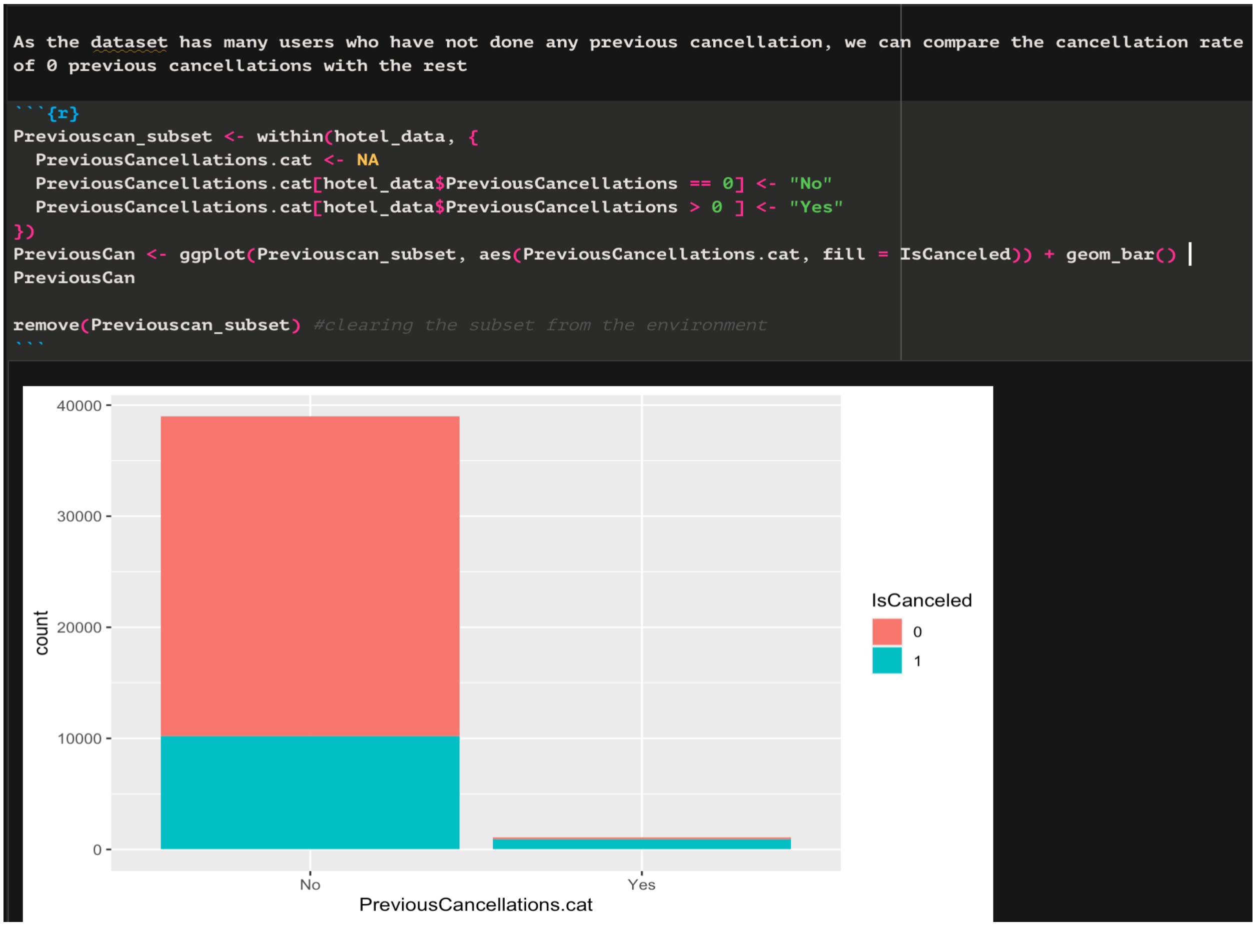


~~~~
roomtype_sub <- within(hotel_data, {
  room_type <- NA
  room_type[as.character(hotel_data$ReservedRoomType) == as.character(hotel_data$AssignedRoomType)] <- "Same"
})
roomtype_sub$room_type[is.na(roomtype_sub$room_type)] <- "Different"
roomtype_sub$room_type <- as.factor(roomtype_sub$room_type)
table(roomtype_sub$room_type)
RoomType_plot <- ggplot(roomtype_sub, aes(room_type, fill=IsCanceled)) + geom_bar()
RoomType_plot

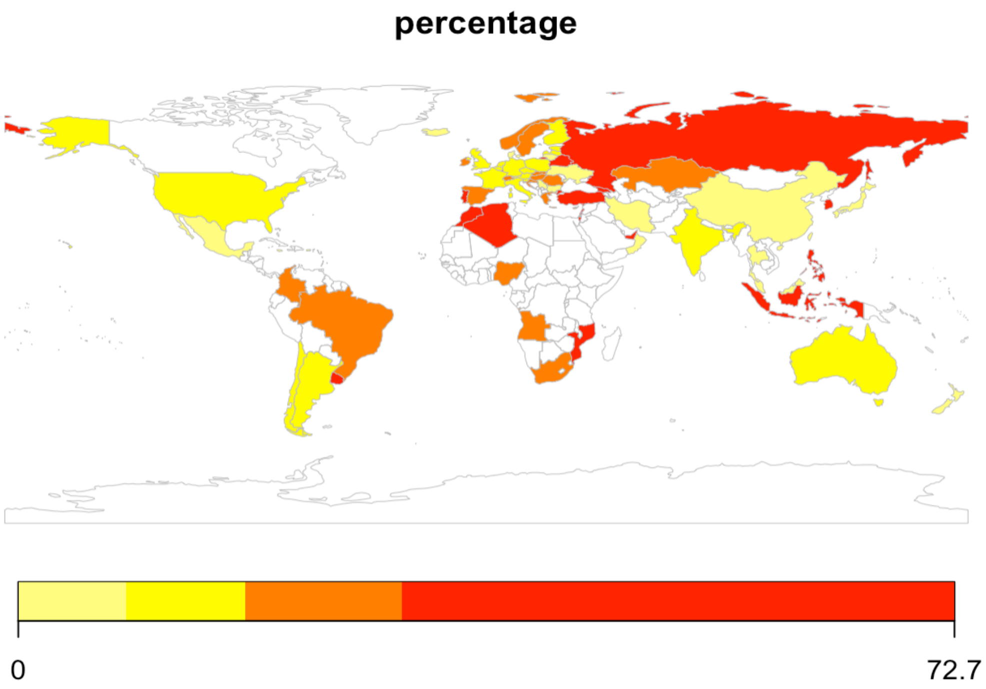
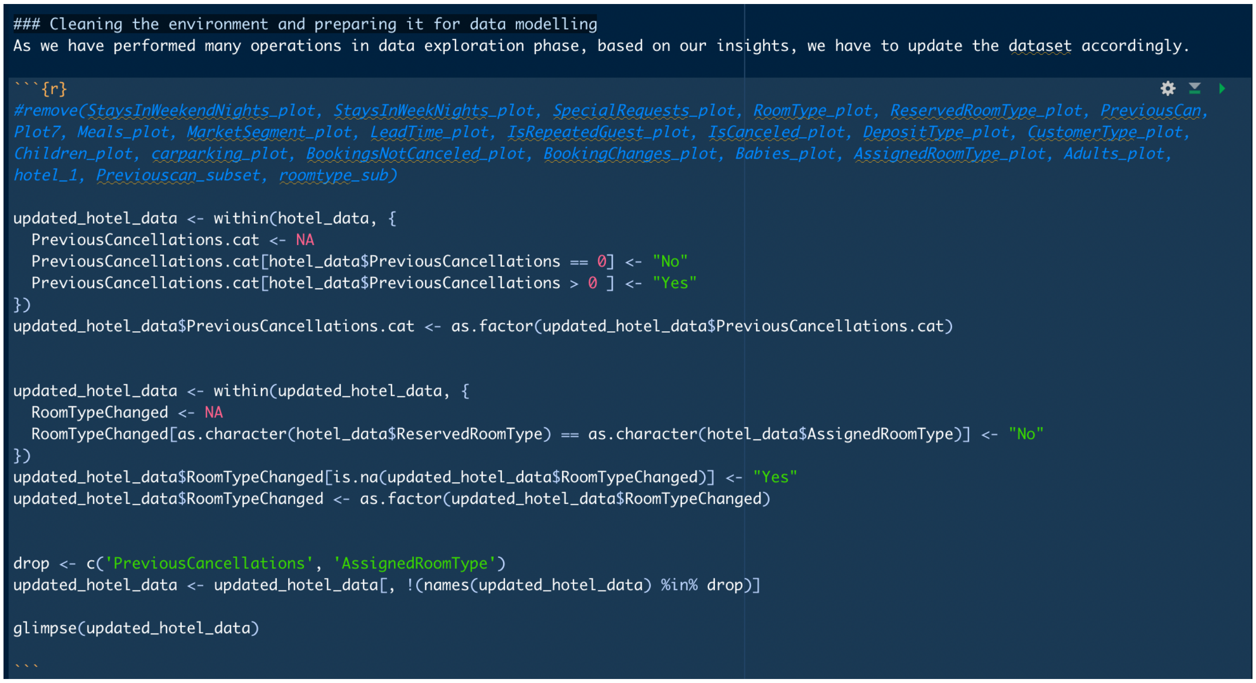
remove(roomtype_sub) #clearing the sub data from the environment
```
~~~~

Figure 4.7

Room type is a variable that we have created to store the assigned and reserved room types, under two categories: Same and Different. First if the room type reserved is the same as the room assigned to the guest then it is assigned to the same otherwise it is assigned to a different room type. As we see in figure 4.7, different room types have fewer cancellations as compared to the same room type. And based on this analysis we have assumed that maybe the hotel offered an upgrade to the guests therefore there are fewer cancellations.



As the dataset has many users who have not done any previous cancellation, we can compare the cancellation rate of 0 previous cancellations with the rest

~~~
```{r}
Previouscan_subset <- within(hotel_data, {
  PreviousCancellations.cat <- NA
  PreviousCancellations.cat[hotel_data$PreviousCancellations == 0] <- "No"
  PreviousCancellations.cat[hotel_data$PreviousCancellations > 0 ] <- "Yes"
})
PreviousCan <- ggplot(Previouscan_subset, aes(PreviousCancellations.cat, fill = IsCanceled)) + geom_bar()
PreviousCan

remove(Previouscan_subset) #clearing the subset from the environment
```
~~~

Figure 4.8

From figure 4.8 we see that there were fewer cancellations made by the guests when there were no previous cancellations.



Figure 4.9

From Figure 4.9 we can deduce that the countries on the east-side have the maximum number of bookings and also see most of the cancellations. From our analysis we have found that Portugal has the most cancellations.

5. ANALYSIS: MODELING TECHNIQUES

After exploring the provided data we have understood which all data are relevant to build the models for prediction and accordingly the data set has been updated by removing the irrelevant variables. Figure 5.1 depicts the preparation for data modeling.



~~~
### Cleaning the environment and preparing it for data modelling
As we have performed many operations in data exploration phase, based on our insights, we have to update the dataset accordingly.

```{r}
#remove(StaysInWeekendNights_plot, StaysInWeekNights_plot, SpecialRequests_plot, RoomType_plot, ReservedRoomType_plot, PreviousCan,
Plot7, Meals_plot, MarketSegment_plot, LeadTime_plot, IsRepeatedGuest_plot, IsCanceled_plot, DepositType_plot, CustomerType_plot,
Children_plot, carparking_plot, BookingsNotCanceled_plot, BookingChanges_plot, Babies_plot, AssignedRoomType_plot, Adults_plot,
hotel_1, Previouscan_subset, roomtype_sub)

updated_hotel_data <- within(hotel_data, {
  PreviousCancellations.cat <- NA
  PreviousCancellations.cat[hotel_data$PreviousCancellations == 0] <- "No"
  PreviousCancellations.cat[hotel_data$PreviousCancellations > 0 ] <- "Yes"
})
updated_hotel_data$PreviousCancellations.cat <- as.factor(updated_hotel_data$PreviousCancellations.cat)


updated_hotel_data <- within(updated_hotel_data, {
  RoomTypeChanged <- NA
  RoomTypeChanged[as.character(hotel_data$ReservedRoomType) == as.character(hotel_data$AssignedRoomType)] <- "No"
})
updated_hotel_data$RoomTypeChanged[is.na(updated_hotel_data$RoomTypeChanged)] <- "Yes"
updated_hotel_data$RoomTypeChanged <- as.factor(updated_hotel_data$RoomTypeChanged)


drop <- c('PreviousCancellations', 'AssignedRoomType')
updated_hotel_data <- updated_hotel_data[, !(names(updated_hotel_data) %in% drop)]

glimpse(updated_hotel_data)

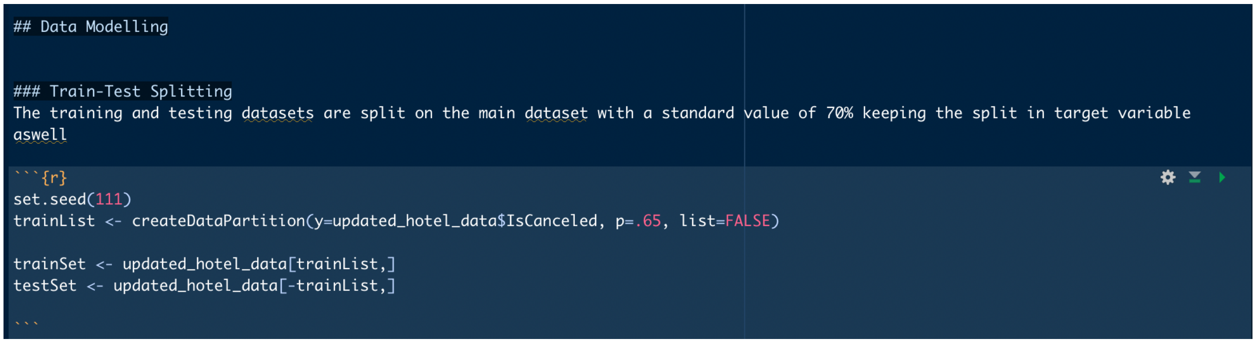
~~~
```

Figure 5.1

To make predictions to determine the future cancellations we have applied the following models: Support Vector Machines, RPart Tree model and Association Rule Mining.

5.1 Support Vector Machines (SVM)

SVM is a supervised learning model that is used to classify the data into different classes. It is used to carry out regression as well as density-estimation. It is also used to find the accuracy of the prediction of a particular model. Here, we split the sampled dataset into 2 parts, training and testing and the train data was used on a model and the test data was used to determine its accuracy of prediction.



```
## Data Modelling

### Train-Test Splitting
The training and testing datasets are split on the main dataset with a standard value of 70% keeping the split in target variable
aswell

```{r}
set.seed(111)
trainList <- createDataPartition(y=updated_hotel_data$IsCanceled, p=.65, list=FALSE)

trainSet <- updated_hotel_data[trainList,]
testSet <- updated_hotel_data[-trainList,]

```
```

Figure 5.2
```

Figure 5.2 depicts the code required to split the data into two parts: train set and test set. After training the set and testing it we got a model with accuracy of 87.36%.

.

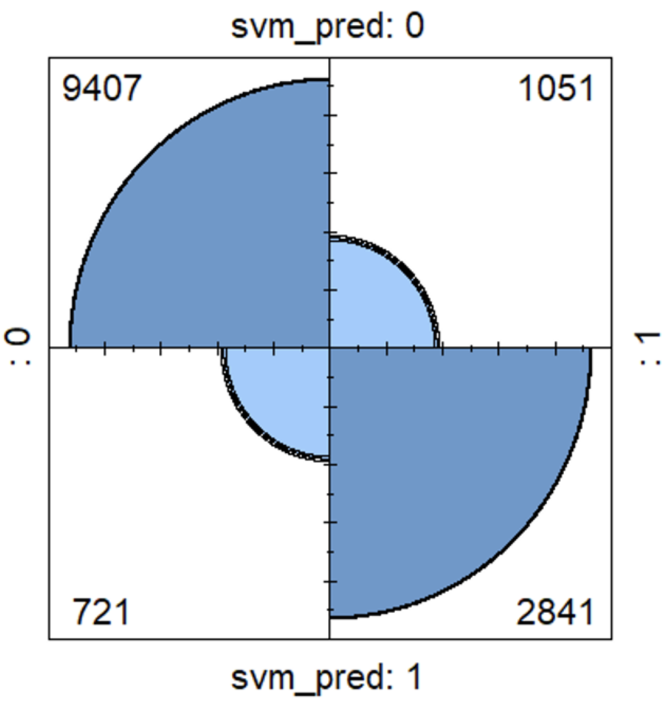
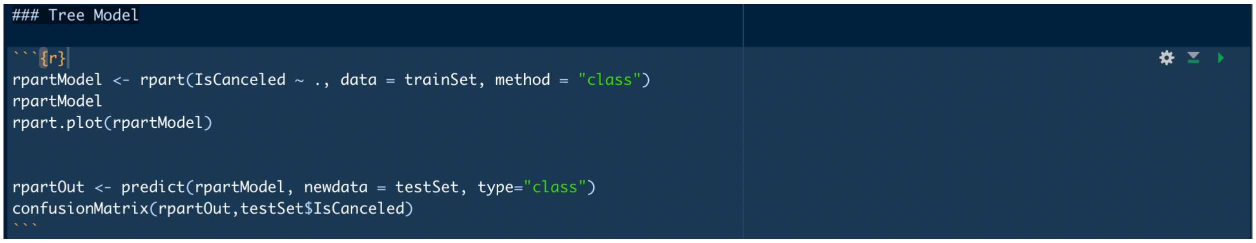


Figure 5.3

## 5.2 Tree (RPart) Model:

Decision trees are a form of supervised machine learning algorithm that can be used to solve a variety of classification problems and also regression problems. This model is a two stage process and the resulting models can be represented as binary trees. The tree is built using the following process: first a single variable is found which can best split the data into 2 groups and once that is done, the process is applied individually to each group and this process is carried out recursively until no improvements can be made. Second stage is using cross-validation to trim the full tree back.  Using this model we have achieved 83.07% accuracy.



~~~
### Tree Model

```{r}
rpartModel <- rpart(IsCanceled ~ ., data = trainSet, method = "class")
rpartModel
rpart.plot(rpartModel)


rpartOut <- predict(rpartModel, newdata = testSet, type="class")
confusionMatrix(rpartOut,testSet$IsCanceled)
```
~~~

Figure 5.4

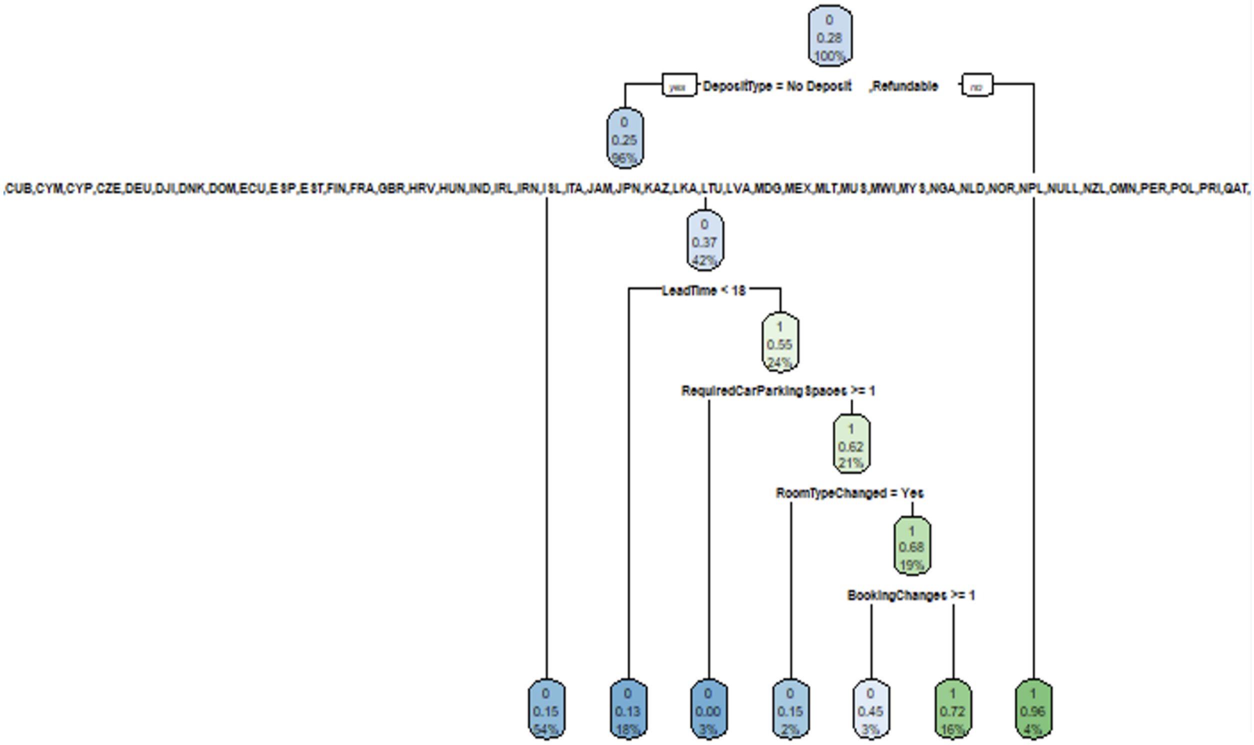

0
0.28
100%
yes
DepositType = No Deposit
,Refundable
no
0
0.25
96%
,CUB,CYM,CYP,CZE,DEU,DJI,DNK,DOM,ECU,ESP,EST,FIN,FRA,GBR,HRV,HUN,IND,IRL,IRN,ISL,ITA,JAM,JPN,KAZ,LKA,LTU,LVA,MDG,MEX,MLT,MUS,MWI,MYS,NGA,NLD,NOR,NPL,NULL,NZL,OMN,PER,POL,PRI,QAT,
0
0.37
42%
LeadTime < 18
1
0.55
24%
RequiredCarParkingSpaces >= 1
1
0.62
21%
RoomTypeChanged = Yes
1
0.68
19%
BookingChanges >= 1
0
0.15
54%
0
0.13
18%
0
0.00
3%
0
0.15
2%
0
0.45
3%
1
0.72
16%
1
0.96
4%

Figure 5.5

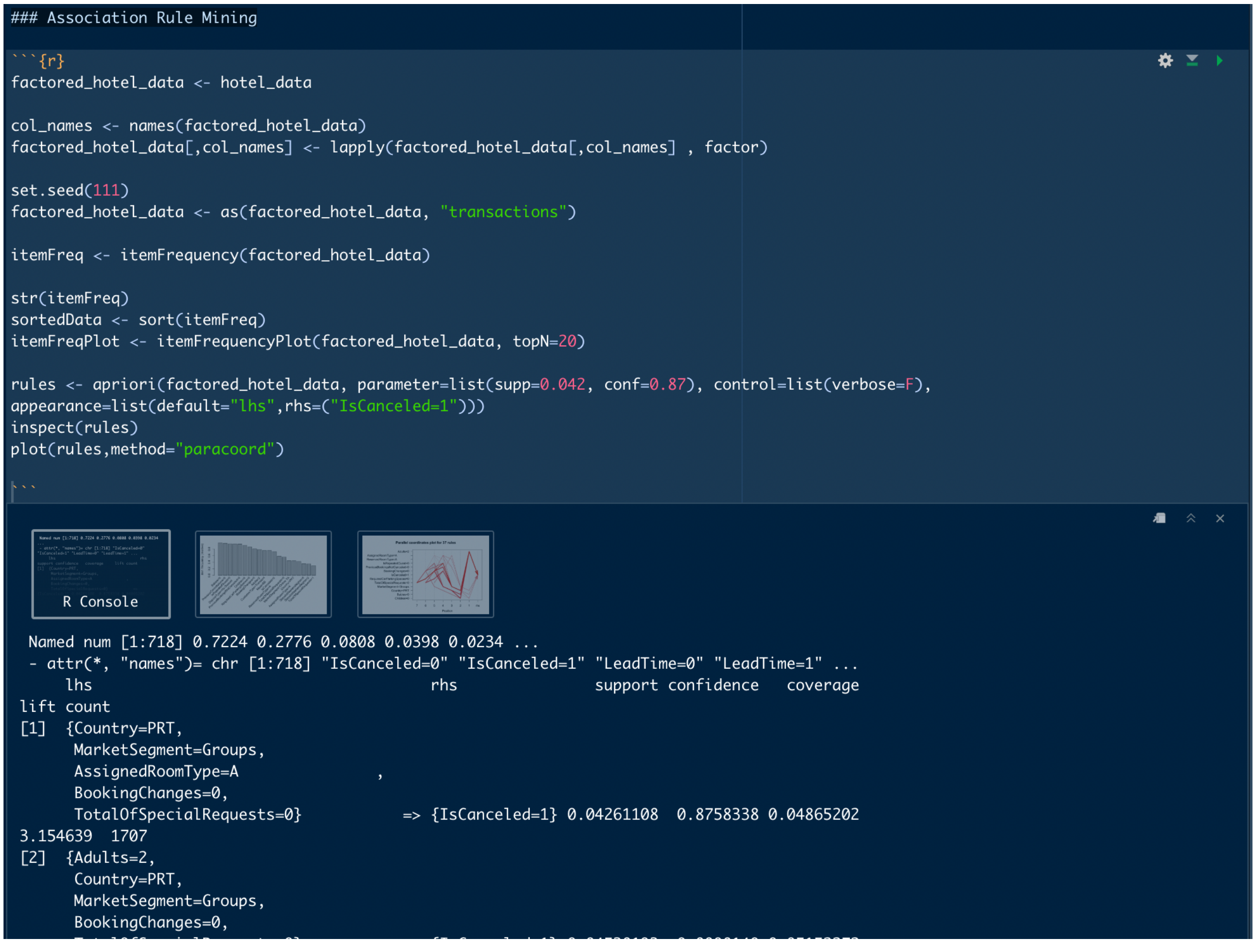
Figure 5.4 and 5.5 depicts the code and the tree model.

5.3 Association Rule Mining

Association rules are created using the apriori function. It is useful in finding variables that often occur together. It is created with a left-hand side that is often together with one variable of the right-hand side.

We converted all integer variables that we plan to use to categorical variables. We then created a data frame of the key variables to be able to perform an apriori analysis on the new data set.

These are some of the rules obtained from the model



~~~
### Association Rule Mining

```{r}
factored_hotel_data <- hotel_data

col_names <- names(factored_hotel_data)
factored_hotel_data[,col_names] <- lapply(factored_hotel_data[,col_names] , factor)

set.seed(111)
factored_hotel_data <- as(factored_hotel_data, "transactions")

itemFreq <- itemFrequency(factored_hotel_data)

str(itemFreq)
sortedData <- sort(itemFreq)
itemFreqPlot <- itemFrequencyPlot(factored_hotel_data, topN=20)

rules <- apriori(factored_hotel_data, parameter=list(supp=0.042, conf=0.87), control=list(verbose=F),
appearance=list(default="lhs",rhs=("IsCanceled=1")))
inspect(rules)
plot(rules,method="paracoord")

```
~~~

```
 Named num [1:718] 0.7224 0.2776 0.0808 0.0398 0.0234 ...
 - attr(*, "names")= chr [1:718] "IsCanceled=0" "IsCanceled=1" "LeadTime=0" "LeadTime=1" ...
     lhs                                           rhs                support confidence   coverage
lift count
[1]  {Country=PRT,
      MarketSegment=Groups,
      AssignedRoomType=A                   ,
      BookingChanges=0,
      TotalOfSpecialRequests=0}            => {IsCanceled=1} 0.04261108  0.8758338 0.04865202
3.154639  1707
[2]  {Adults=2,
      Country=PRT,
      MarketSegment=Groups,
      BookingChanges=0,
```

Figure 5.6

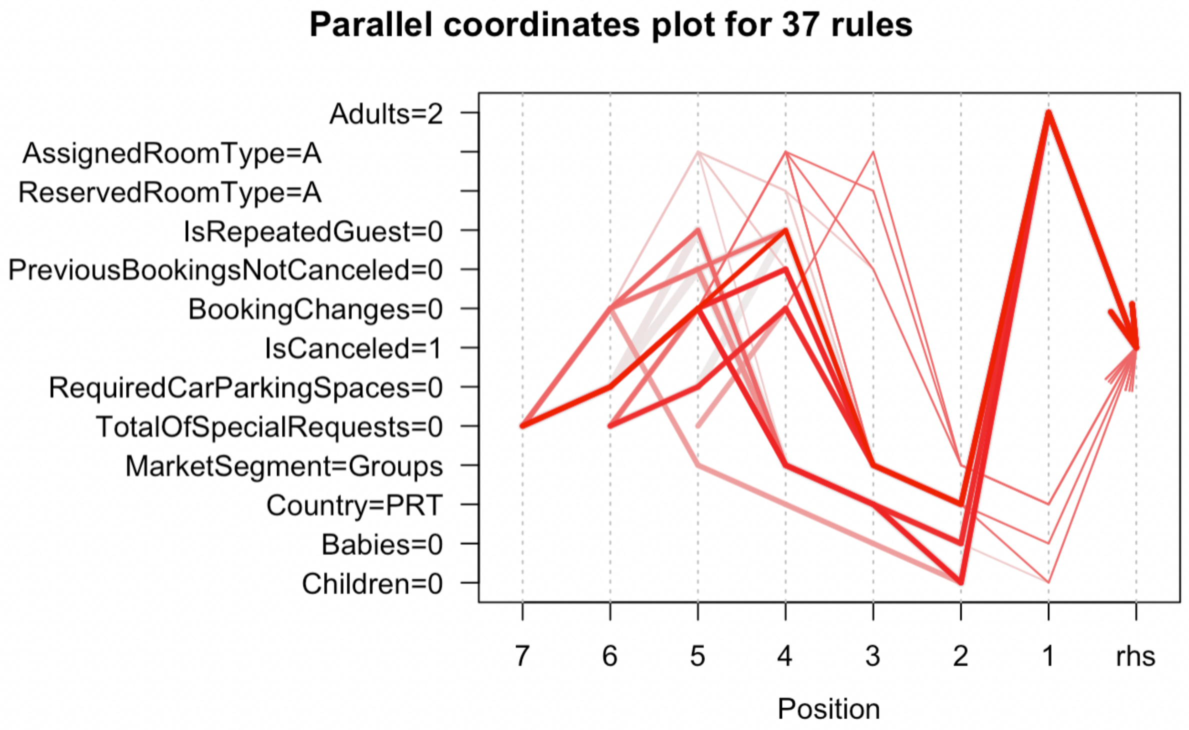


Figure 5.7

6. Insights and Recommendations

6.1 Insights

We have discovered the following trend after examining our dataset:

a) Market Segment

Insight: The number of cancellations is more when the guests book through a third-party agent.

Recommendation: To limit the frequency of cancellations, the hotel can establish a gateway where customers can book their own hotel room.

b) Previous Cancellations

Insight: Guests who canceled their bookings previously are more likely to cancel it again.
Recommendation: Customers who have previously canceled their reservations should be offered bundle packages or discounts.

c) Deposit Type

Insight: The number of cancellation is more incase of no-deposit bookings.
The number of cancellations is high when the booking is non-refundable.
Recommendation: A modest payment for the booking can be paid at the time of booking.

Meal Type

Insight: Bookings with bed & breakfast options have seen more cancellations.
Recommendation: