Recommendations for Hotels based on the cancellation data

IST 687: Introduction to Data Science

By

Aditya Shah - 747686910
Manan Vora - 349846253
Yash Shimpi - 313092853

Table of Contents

Introduction *1*

Data Acquisition *1*

Primary Questions/ Considerations *3*

Descriptive Statistics and Data Pre-Processing *4*

Exploratory Analysis and Data Munging *6*

Predictive Analysis *18*

Recommendation *25*

Introduction

We are given a dataset that contains hotel stay data, with each row representing a hotel booking.

Throughout this project we achieve two main objectives:

1. Trying to understand key metrics for people canceling hotel reservations.
2. Predict conditions that make customers cancel their reservations.

We thoroughly analyze the dataset which contains a total of 20 columns/variables and 40060 rows/entries. The variable that we will be predicting is **IsCanceled** which gives us information on whether the booking is canceled by the user. We will implement the techniques taught during lectures and lab to come out with the best possible solution.

Data Acquisition

The data is made available to us by the instructor. The data is available in the form of CSV file which we read with the help of tidyverse function as shown below

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

The variables mentioned in the dataset are as follows: -

Variable Name	**Meaning**	**Data Type**
IsCanceled	Indicates if the booking was canceled or not	Categorical
LeadTime	Number of days between booking and the arrival date	Integer
StaysInWeekendNights	Number of weekend nights that the guests are staying	Integer
StaysInWeekNights	Number of weeknights that the guests are staying	Integer
Adults	Number of adults in the booking	Integer
Children	Number of children in the booking	Integer
Babies	Number of babies in the booking	Integer

Meal	Type of meal booked like bed and breakfast, Full Board, Half board and No meals	Categorical
Country	The country of origin	Categorical
MarketSegment	How the rooms were booked, channels specifically like Offline Travel Agents, Online Tour Operators and so on.	Categorical
IsRepeatedGuest	Indicates if the guest is a returning one or not.	Categorical
PreviousCancellations	Number of previous cancellations prior to the current booking	Integer
PreviousBookingsNotCanceled	Number of previous bookings not canceled prior to this booking.	Integer
ReservedRoomType	Code of room type.	Categorical
AssignedRoomType	Code of room type assigned to the customer.	Categorical
BookingChanges	Number of changes/amendments from the time when booking was entered.	Integer
DepositType	Indicating if the customer made a depositor not. If yes, is it refundable or not.	Categorical
CustomerType	Type of booking: allotment/contract associated to the booking, booking associated to a group, transient booking, transient party booking - transient booking associated with another transient booking.	Categorical
RequiredCardParkingSpaces	Number of car parking spaces required by the customer.	Integer
TotalOfSpecialRequests	Number of special requests made by the customer.	Integer

Primary Questions/ Considerations

There were some questions that were taken into consideration so as to lay a pathway into more effective exploration and an exhaustive analysis of data. Some of these questions were thought to be rational explanations of cancellations based on the different variables available to us. These are listed below.

1. Does the **lead time** affect cancellation? Does a higher lead time mean more possibility of cancellation?
2. Does **staying in a weeknight** contribute to the cancellations owing to the shift and work schedule?
3. Does the number of **adults** translate to **Required car parking spaces**?
4. Do **meals** have a relationship with cancellation?
5. **Online tour agents** provide security and flexibility to customers. Does this contribute to cancellation?
6. **Transient customers** are more flexible with their schedule and tend to change their plans quicker than a group. Does this trend is reflected in the data as well?
7. The higher the number **of special requests** made by the customers, the more inclined the customer is to come to a hotel. So, if the person does not make any request or only a few, is he more likely to cancel?
8. In the case of **no deposit**, the person is not liable in any terms for the bookings. Does this encourage cancellation?
9. Is a **repeated customer** less inclined to cancel than someone who is new?
10. If the **assigned room** is different from the **reserved room**, does that make the customers cancel their reservation? Hgc have top cancellations among all - maybe fix that

Descriptive Statistics and Data Pre-Processing

The variables present in the dataset is understood with the help of **str** and **summary** function as shown below:

1. **str(hotel_data)**

```
spec_tbl_df [40,060 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ IsCanceled                : num [1:40060] 0 0 0 0 0 0 0 0 1 1 ...
 $ LeadTime                  : num [1:40060] 342 737 7 13 14 14 0 9 85 75 ...
 $ StaysInWeekendNights      : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ StaysInWeekNights         : num [1:40060] 0 0 1 1 2 2 2 2 3 3 ...
 $ Adults                    : num [1:40060] 2 2 1 1 2 2 2 2 2 2 ...
 $ Children                  : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ Babies                    : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ Meal                      : chr [1:40060] "BB" "BB" "BB" "BB" ...
 $ Country                   : chr [1:40060] "PRT" "PRT" "GBR" "GBR" ...
 $ MarketSegment             : chr [1:40060] "Direct" "Direct" "Direct" "Corporate" ...
 $ IsRepeatedGuest           : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ PreviousCancellations     : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ PreviousBookingsNotCanceled: num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ ReservedRoomType          : chr [1:40060] "C" "C" "A" "A" ...
 $ AssignedRoomType          : chr [1:40060] "C" "C" "C" "A" ...
 $ BookingChanges            : num [1:40060] 3 4 0 0 0 0 0 0 0 0 ...
 $ DepositType               : chr [1:40060] "No Deposit" "No Deposit" "No Deposit" "No Deposit" ...
 $ CustomerType              : chr [1:40060] "Transient" "Transient" "Transient" "Transient" ...
 $ RequiredCarParkingSpaces  : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
 $ TotalOfSpecialRequests    : num [1:40060] 0 0 0 0 1 1 0 1 1 0 ...
```

The str functions gives the information about different variable types present in the dataset

2. **summary(hotel_data)**

```
summary(hotel_data)
```

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

Summary function helps us to know that there are no null values present.

Now, we convert the variables with type <chr> to factors as shown below

```r
hotel_data$CustomerType <- as.factor(hotel_data$CustomerType)
hotel_data$DepositType <- as.factor(hotel_data$DepositType)
hotel_data$MarketSegment <- as.factor(hotel_data$MarketSegment)
hotel_data$Meal <- as.factor(hotel_data$Meal)
hotel_data$IsRepeatedGuest <- as.factor(hotel_data$IsRepeatedGuest)
hotel_data$ReservedRoomType <- as.factor(hotel_data$ReservedRoomType)
hotel_data$AssignedRoomType <- as.factor(hotel_data$AssignedRoomType)
```

Since our goal is to find whether the booking is canceled or not, we will check the total number of entries for canceled and not canceled.

```r
table(hotel_data$IsCanceled)
```

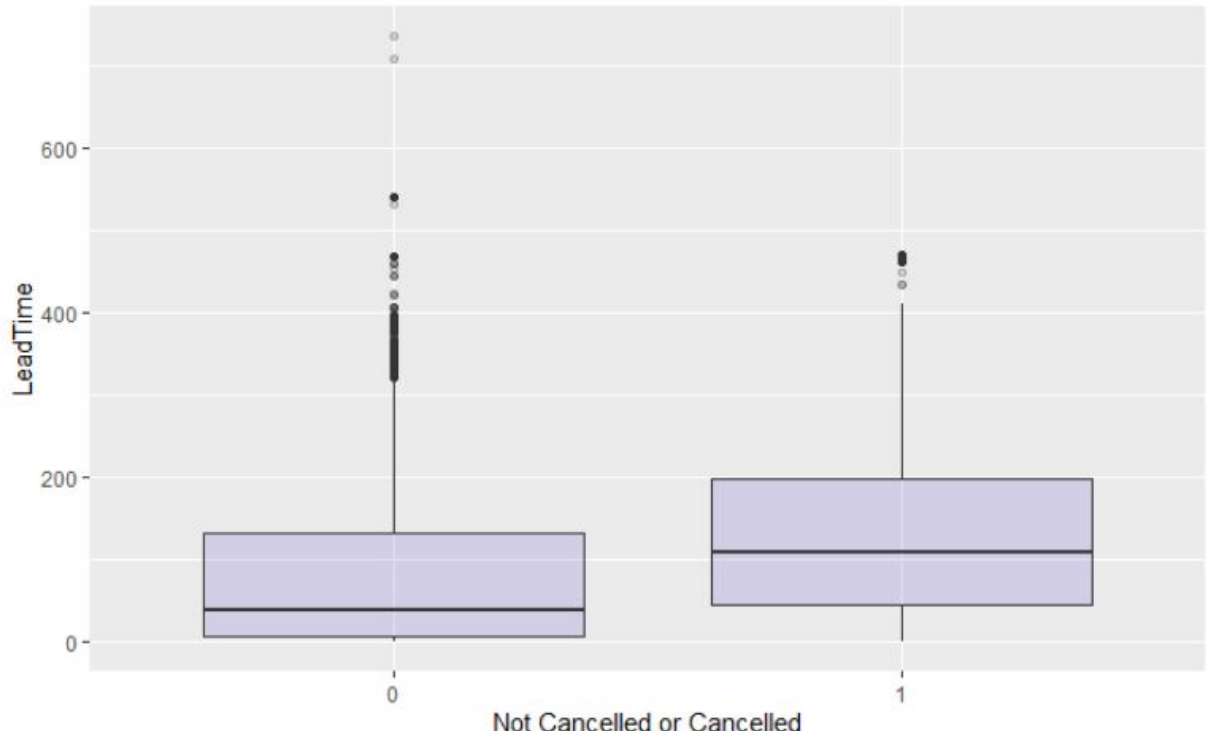
```
    0     1
28938 11122
```

We have 11122 entries with canceled bookings and 28938 of those which were not canceled.

Exploratory Analysis and Data Munging

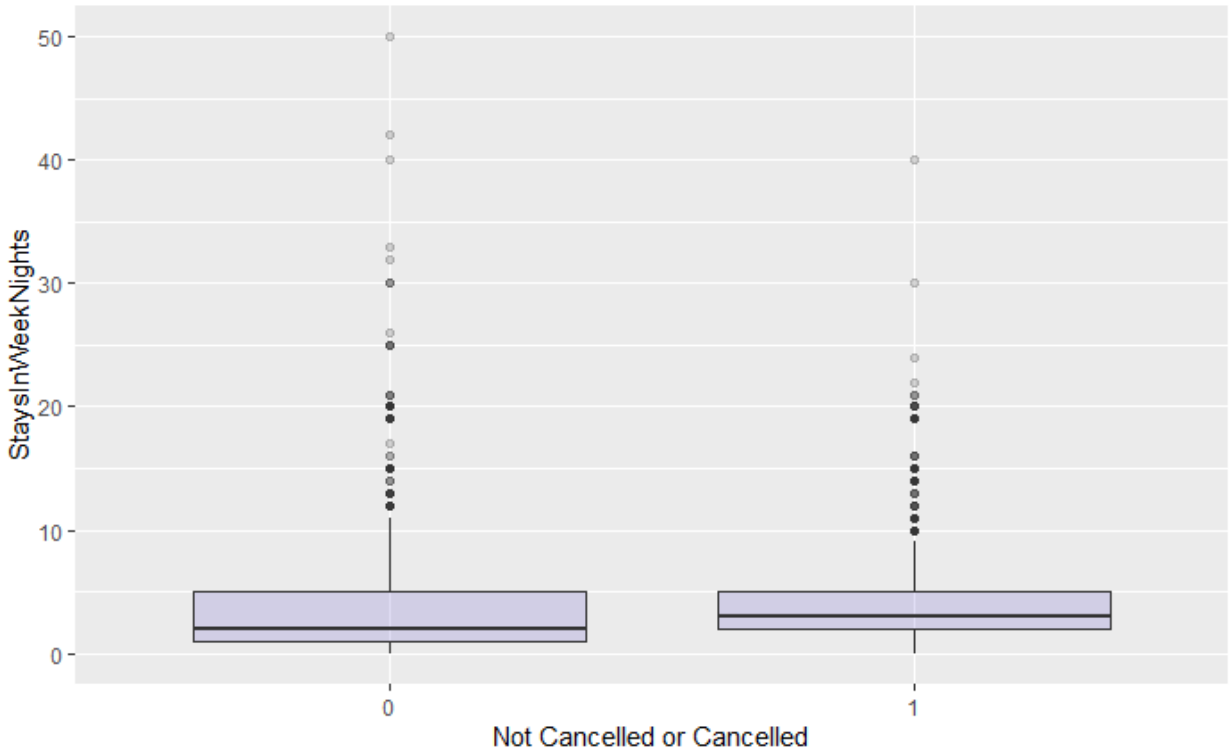
To understand the trends in the dataset, we go forward by finding Cross Tables and Visualizations like Histograms, Barplot, Box Plots, and World Map of the numeric and/or categorical variables. Through our visualization we try to answer the questions raised earlier. We plan to find the key metrics by inferring the results to try find some insightful patterns. We analyze the dataset in its entirety and by dividing the dataset into two - canceled and not canceled sets.

1. **Lead Time w.r.t. "IsCanceled"**

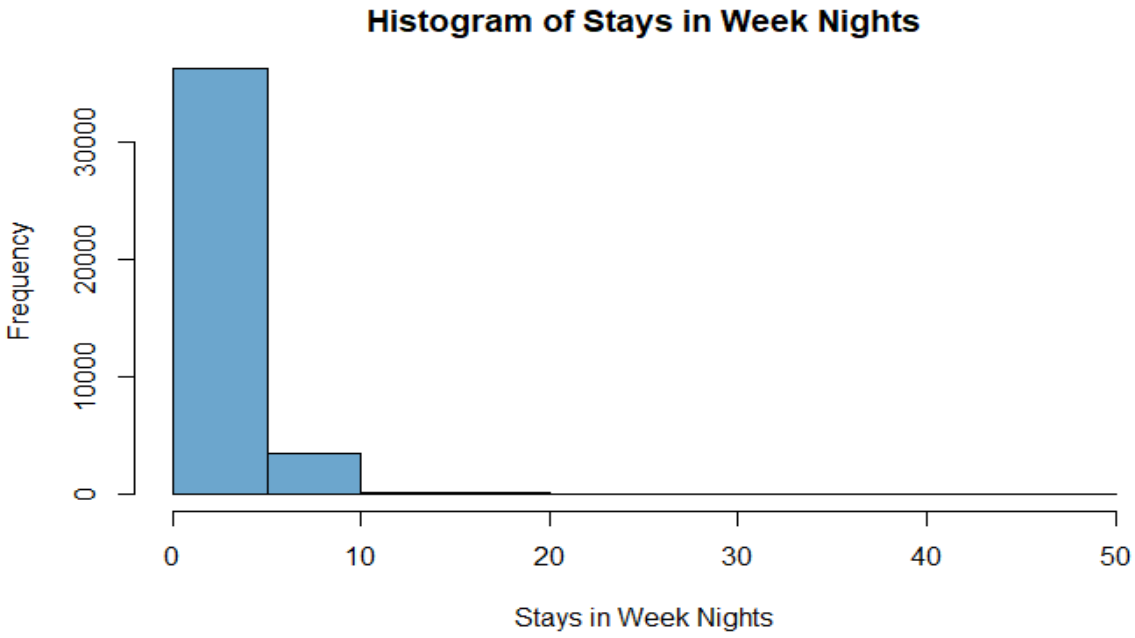


While there are a ton of outliers in the dataset, we can observe that there is a difference between the mean lead times of cancelled and not cancelled bookings. This can be interpreted as higher lead times leading to more cancellations from customers.

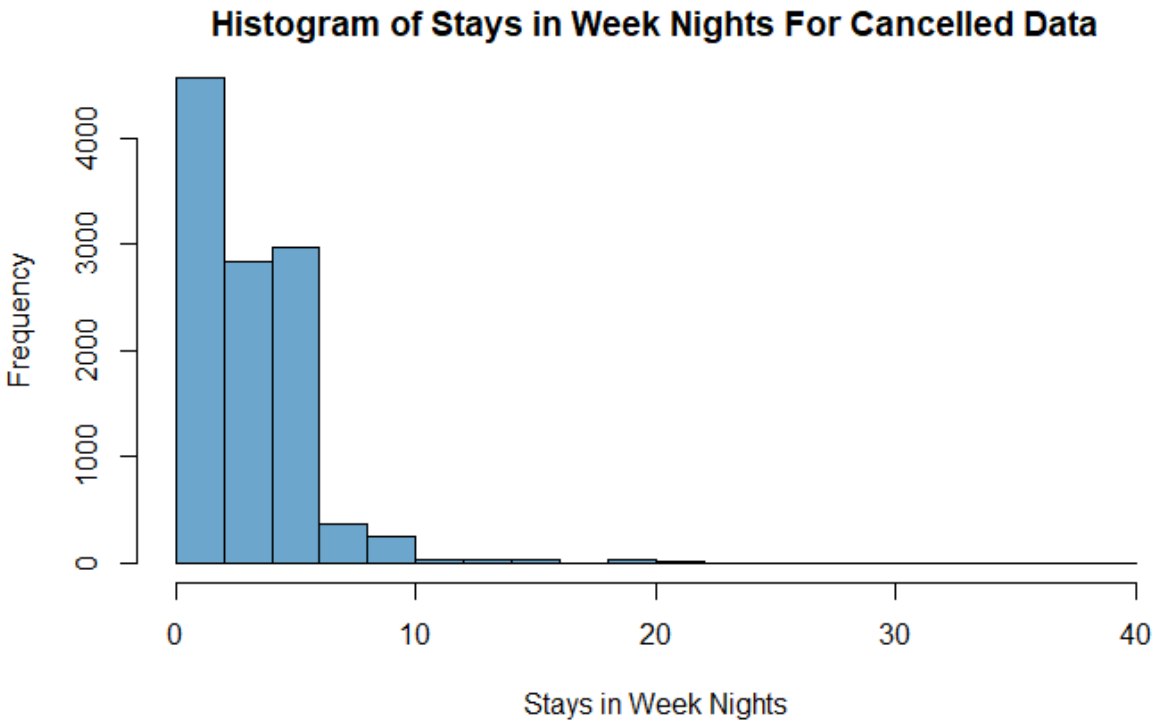
2. **Staying in Weeknights vs Canceled**



The mean of cancellations and those who don't, when customers stay in weeknights is very slightly different. However, even in this minuscule difference, the mean of cancellations is greater. This might be explained by the fact that plans change according to work schedules unexpectedly.

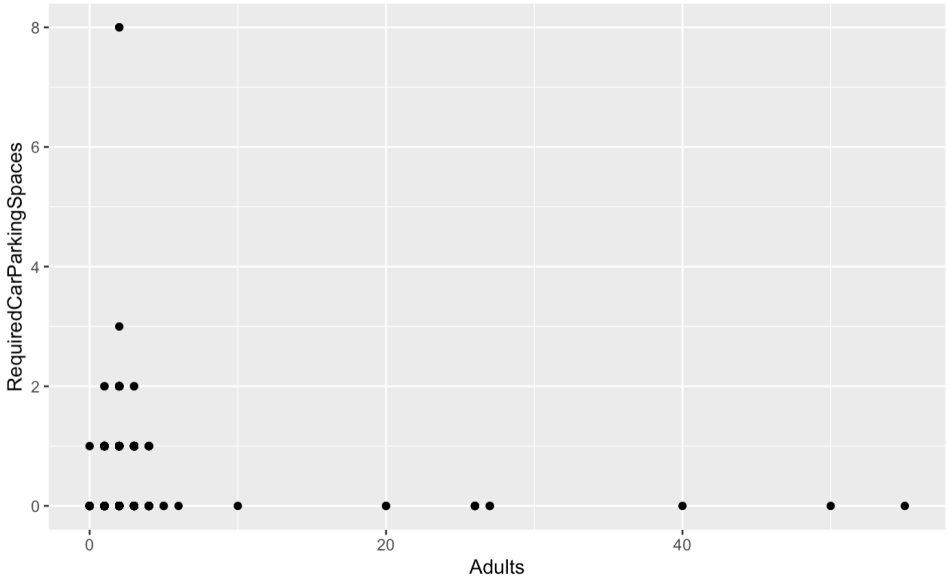


There is little to be known from the histogram of data where people stay in weeknights. This shows that a major proportion of bookings were made for 0-5 weeknights.



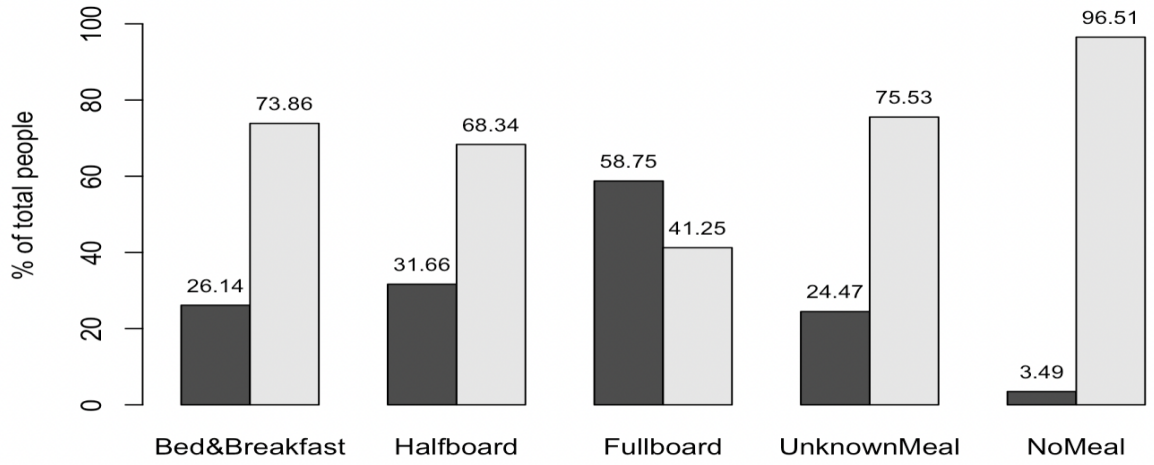
Just like the previous data showed us, as majority of bookings were made for 0-5 weeknights, there is similar trend seen in the cancellations made: the cancellations are maximum for bookings made for 0-5 weeknights

3. Adults vs Car Parking Space



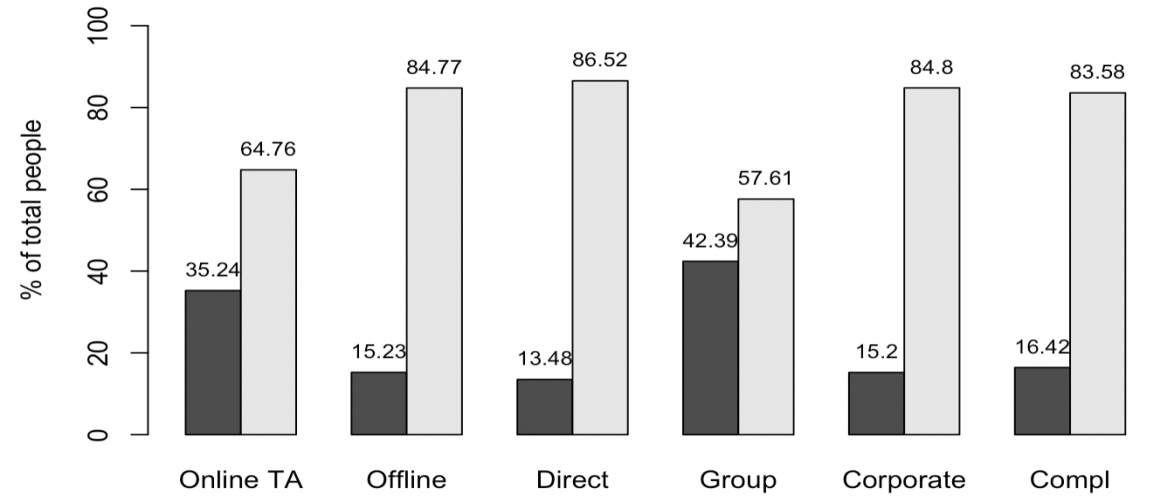
This plot shows that the number of adults is only decisive factor for number of car parking spaces required but only up to a certain number.

4. Meals vs Cancellation



The representation shows that the more the number of meals, the greater is the chance of cancellations. This is a trend that is seen in the data which is a little difficult to explain logically. It could be a consequence of having to commit greater amount of time to staying in hotel which might make customers not be able to abide to a booking in case of change in plans.

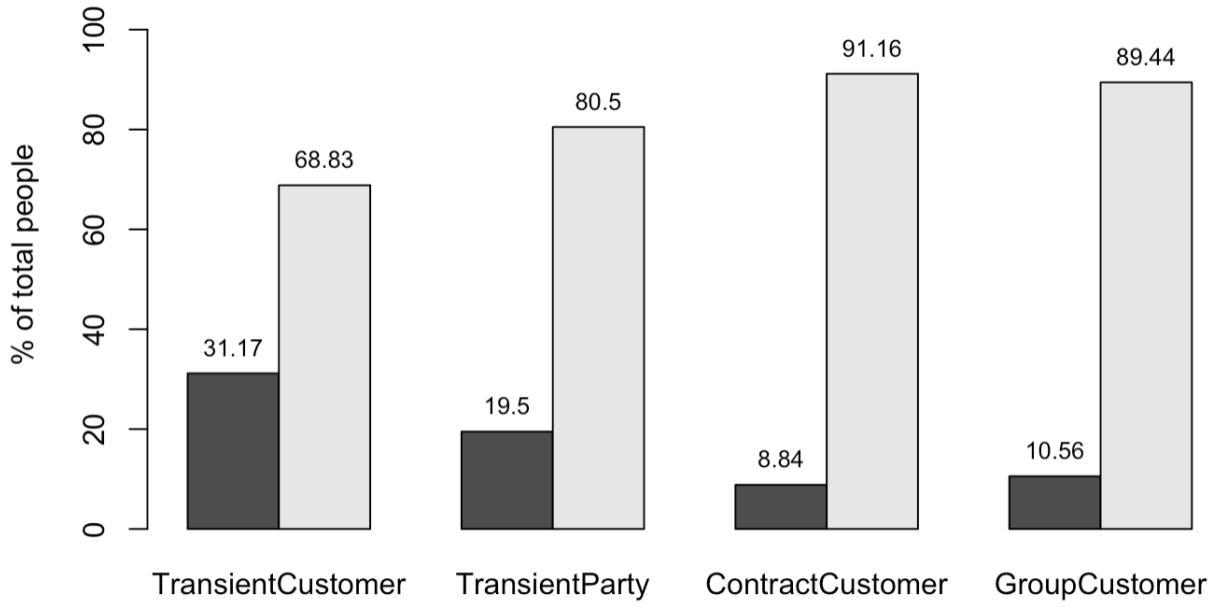
5. Market Segment vs Cancellation



Online travel agents such as Agoda, Bookings.com offer a lot of customer flexibility offers to attract customers. This might end up leading customers using those flexibilities, like cancellation at no cost, and make it easier for customers from the online segment to be more likely to cancel a booking. Other trend seen is that a group segment booking is more

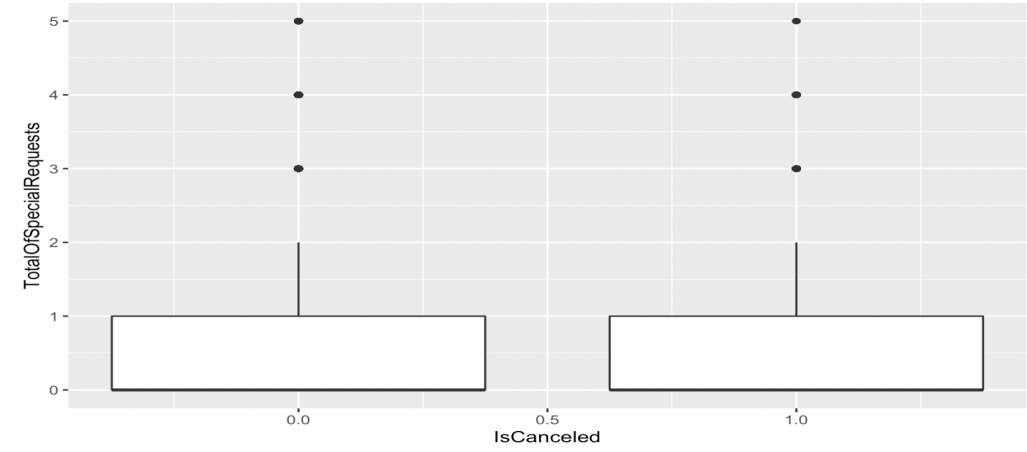
pronounced in the data as the cancellation of a group results in the cancellations of a large number of individual bookings that constituted the said group.

6. Customer Segment vs Cancellation



A transient customer is one that is an individual booking that is not associated with another transient booking. These bookings have shown the maximum number of cancellations out of all the customer segments.

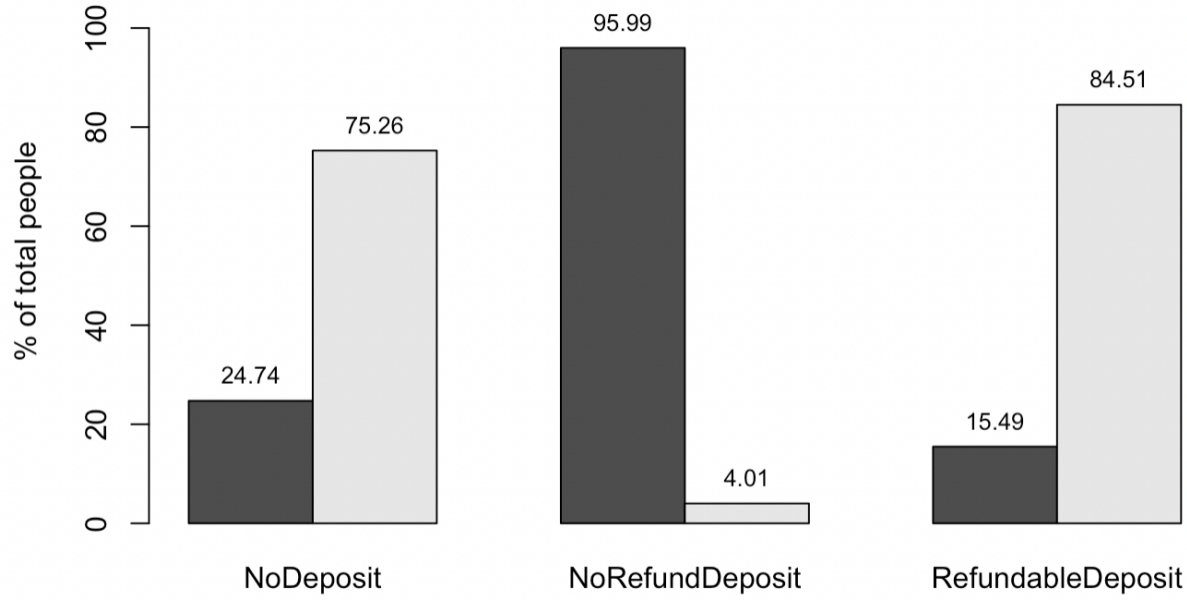
7. Requested Number of Special Requests vs Cancellation



It can be thought that fulfilling as many special requests as possible of the customers could convince them to not cancel their bookings. However, the data shows that there is no difference between the cancellation trends between customers with various

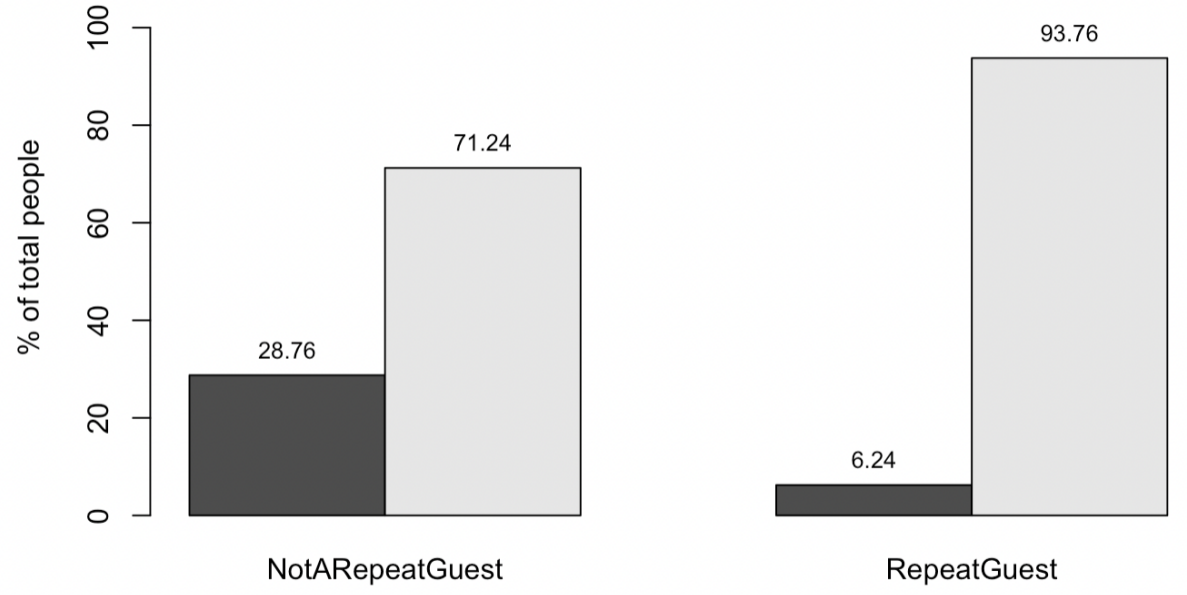
number of special requests. The means of special requests of those customers who have cancelled and those who haven't is the same.

8. Deposit Type vs Cancellation



No Refund deposit bookings are generally cheaper than those which have refundable deposits. In an event where a customer finds a better deal or better services, we can assume that they can choose to trade that deposit loss in favor of a better booking.

9. Repeated customer vs IsCanceled



Is a repeated customer less inclined to cancel than someone who is new? The representation says yes. There is a massive difference in the percentages of cancellations of first-time customers and that of a returning customer.

10. Assigned+Reserved vs Cancellation

	IsCanceled	
CompareReserveAssign	0	1
FALSE	7341	384
TRUE	21597	10738

CompareReserveAssign is the attribute that says if the reserved room type and assigned room type were the same in a given booking. TRUE means they were and FALSE means otherwise. Logic says that the customers will be satisfied if their reserved and assigned room types are the same, and eventually would be less inclined to cancel. The data agrees to this logic and seems like a large proportion of bookings: about 50% were given the same type of room that they reserved. However, the cancellation in the same case is also very high.

Analysis of Complete Dataset

A. CrossTable

	IsCanceled=0	IsCanceled=1
IsRepeatedGuest=0	27271	11011
DepositType=No Deposit	28749	9450
CustomerType=Transient	20793	9416
Meal=BB	22162	7843
IsCanceled=0	28938	0
MarketSegment=Online TA	11481	6248
IsCanceled=1	0	11122
Meal=HB	5499	2547
CustomerType=Transient-Party	6272	1519
MarketSegment=Offline TA/TO	6334	1138
MarketSegment=Direct	5635	878
MarketSegment=Groups	3362	2474
MarketSegment=Corporate	1958	351
IsRepeatedGuest=1	1667	111
CustomerType=Contract	1619	157
DepositType=Non Refund	69	1650
Meal=Undefined	883	286
Meal=FB	311	443
CustomerType=Group	254	30
MarketSegment=Complementary	168	33
DepositType=Refundable	120	22
Meal=SC	83	3

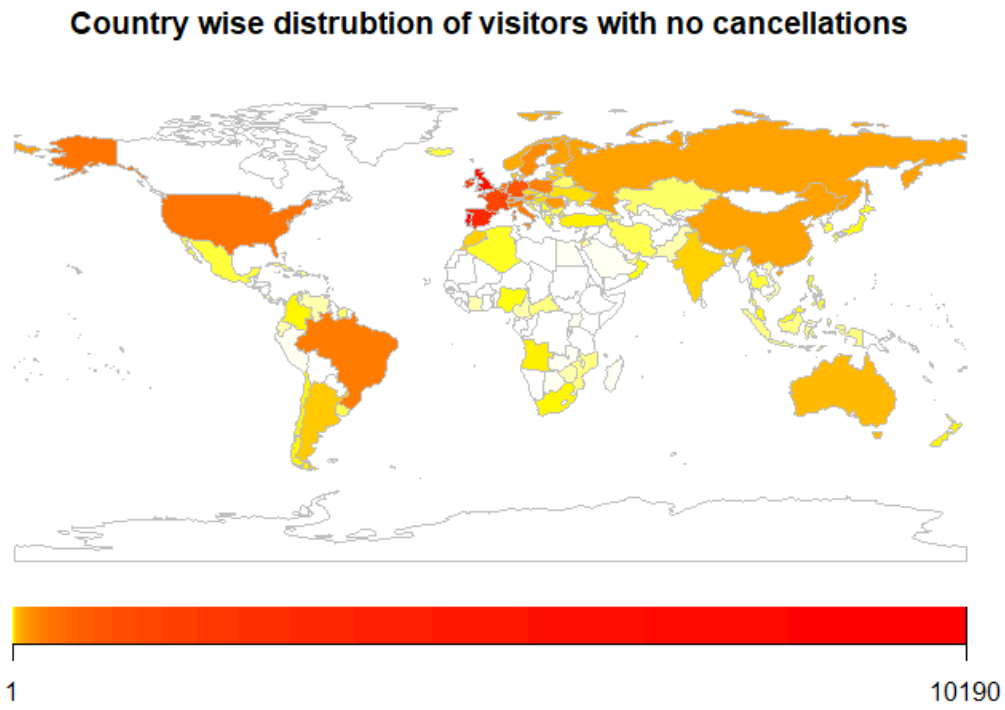
Since the data is vastly different in many different categories of the cross table, a conversion to percentage can essentially regularize the data.

Hence, a crosstable with data converted to percent w.r.t. the total number of transactions was made as shown below:

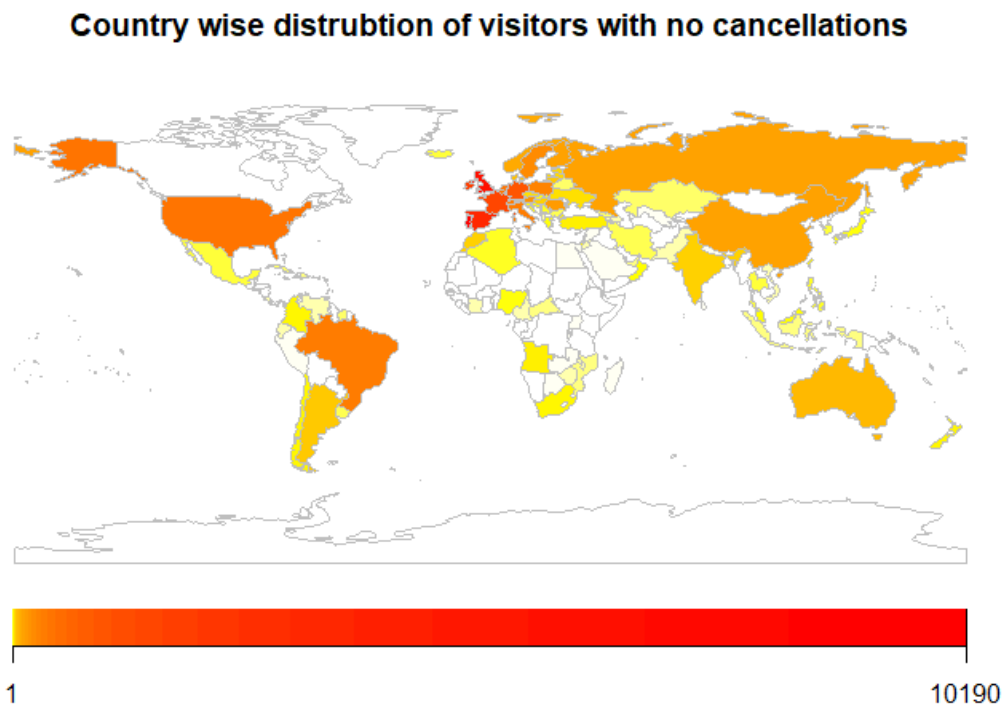
	IsCanceled.0	IsCanceled.1
IsRepeatedGuest=0	71.237135	28.762
DepositType=No Deposit	75.261132	24.738
CustomerType=Transient	68.830481	31.169
Meal=BB	73.861023	26.138
IsCanceled=0	100.000000	0.000
MarketSegment=Online TA	64.758306	35.241
IsCanceled=1	0.000000	100.000
Meal=HB	68.344519	31.655
CustomerType=Transient-Party	80.503145	19.496
MarketSegment=Offline TA/TO	84.769807	15.230
MarketSegment=Direct	86.519269	13.480
MarketSegment=Groups	57.607951	42.392
MarketSegment=Corporate	84.798614	15.201
IsRepeatedGuest=1	93.757030	6.242
CustomerType=Contract	91.159910	8.840
DepositType=Non Refund	4.013962	95.986
Meal=Undefined	75.534645	24.465
Meal=FB	41.246684	58.753
CustomerType=Group	89.436620	10.563
MarketSegment=Complementary	83.582090	16.417
DepositType=Refundable	84.507042	15.492
Meal=SC	96.511628	3.488

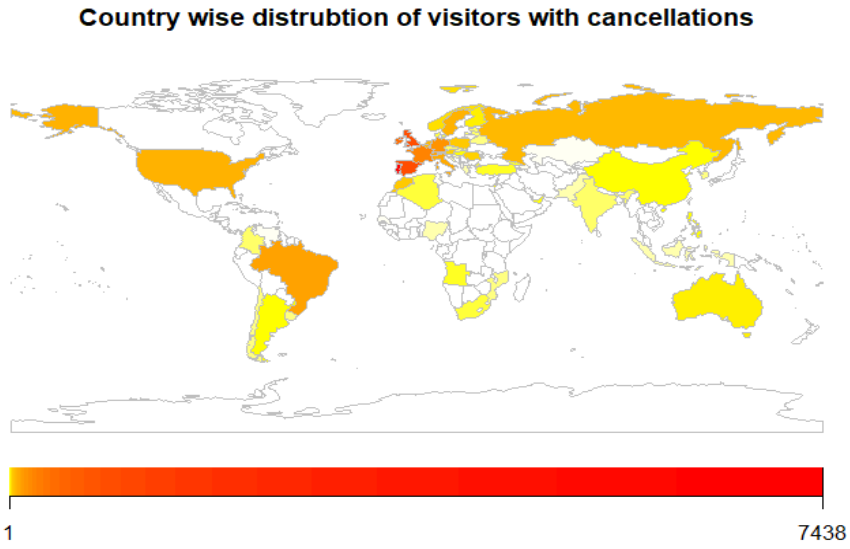
B. Map Visualizations

1. World Map



The above map shows that customers from Western Europe had the least number of cancellations

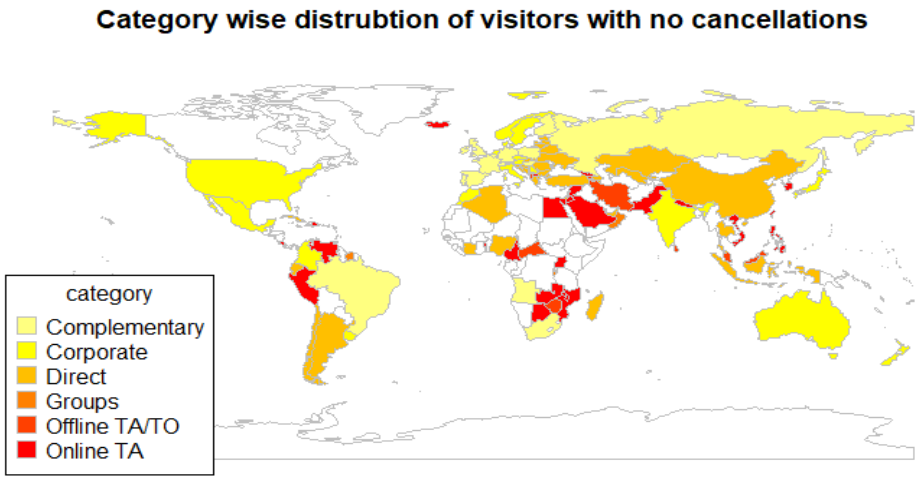


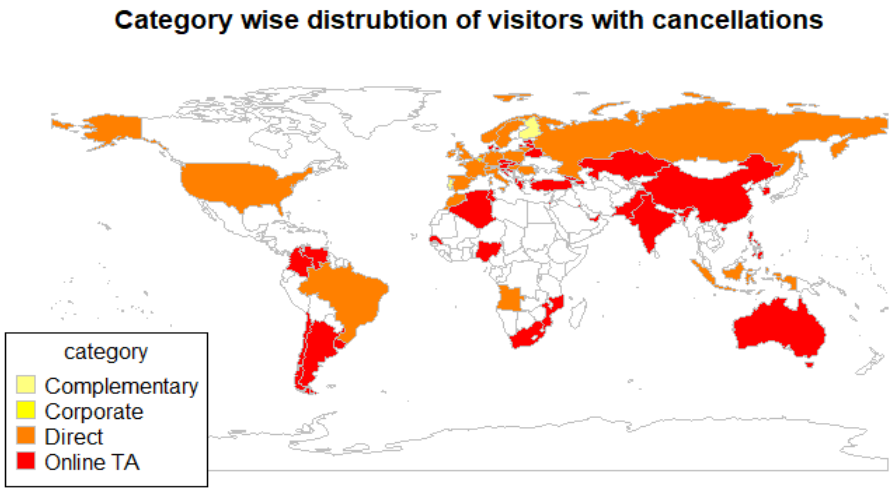


The above map shows that customers from Spain and Portugal had most number of cancellations.

The following maps will depict category wise distribution of visitors with

a. Cancellations
b. No Cancellations





After looking at the visualization we can conclude that the key metrics are:

- RepeatedGuest
- CustomerType
- MarketSegment
- LeadTime

Predictive Analysis

The 2nd stage of the project involves predicting whether the customer will cancel the booking or not based on the key metrics. We will showcase 3 machine learning techniques and based on the results that we obtain; we can decide as to which model should be preferred.

Before using various machine learning techniques, we split the data into training and testing sets. We will train our model with trainset and predict it on our test set.

Training and Testing Sets

```{r}
library(kernlab);library(caret)

trainList <- createDataPartition(y=hotel_data$IsCanceled,p=2/3,list=FALSE)

trainSet <- hotel_data[trainList,]
testSet <- hotel_data[-trainList,]
```

The typical strategy is to divide data into a training set and a test set (for example, two-thirds training and one-third test), train the model on the training set, and then see how well the model does on the test set.

1. **Linear Regression**

   Linear regression is a linear model, e.g., a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

   Firstly, we create a model with dependent variables as IsCanceled while the independent variables are all numeric.

```
hotel_data_n = hotel_data[,c('IsCanceled','LeadTime','StaysInWeekendNights','StaysInWeekNights','Adults',
                             'Children','Babies','IsRepeatedGuest','PreviousCancellations',
                             'PreviousBookingsNotCanceled','BookingChanges','RequiredCarParkingSpaces',
                             'TotalOfSpecialRequests')]
lmOut1 <- lm(IsCanceled ~ ., data = hotel_data_n)
summary(lmOut1)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-1.2223 -0.2962 -0.1997  0.4712  2.0230

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.2231052  0.0066855  33.371  < 2e-16 ***
LeadTime                      0.0008669  0.0000238  36.432  < 2e-16 ***
StaysInWeekendNights          0.0051141  0.0026053   1.963 0.049655 *
StaysInWeekNights            -0.0046266  0.0012472  -3.710 0.000208 ***
Adults                        0.0308355  0.0030538  10.097  < 2e-16 ***
Children                      0.0943377  0.0046891  20.118  < 2e-16 ***
Babies                        0.0085863  0.0176726   0.486 0.627074
IsRepeatedGuest1             -0.1041790  0.0112708  -9.243  < 2e-16 ***
PreviousCancellations         0.0289930  0.0015652  18.523  < 2e-16 ***
PreviousBookingsNotCanceled -0.0084511  0.0023049  -3.667 0.000246 ***
BookingChanges               -0.0703261  0.0028956 -24.288  < 2e-16 ***
RequiredCarParkingSpaces     -0.2587448  0.0060364 -42.864  < 2e-16 ***
TotalOfSpecialRequests       -0.0457740  0.0025946 -17.642  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4151 on 40047 degrees of freedom
Multiple R-squared:  0.1409,    Adjusted R-squared:  0.1407
F-statistic: 547.5 on 12 and 40047 DF,  p-value: < 2.2e-16
```

From the results of the Linear Model, we can infer a few points:

a. P-value is <2.2e-16 which is a good value

b. The adjusted R-squared value is 0.1407 which signifies low accuracy.

c. Numerical Variables although are seen to have high significance does not contribute much towards the model’s accuracy.

Firstly, we create a model with dependent variables as IsCanceled while the independent variables are all categorical.

```
lmOut2 <- lm(IsCanceled ~ ., data = hotel_data_c)
summary(lmOut2)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-0.9762 -0.2403 -0.1177  0.1283  1.1805

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.122809   0.084833   1.448 0.147720
MealFB                    0.047132   0.015134   3.114 0.001845 **
MealHB                    0.043696   0.005183   8.430  < 2e-16 ***
MealSC                   -0.142384   0.042375  -3.360 0.000780 ***
MealUndefined            -0.100161   0.012573  -7.966 1.68e-15 ***
CountryALB                0.037104   0.239453   0.155 0.876861
CountryAND                0.226148   0.192206   1.177 0.239365
```

```
MarketSegmentOffline TA/TO    0.094849   0.028170   3.367 0.000761 ***
MarketSegmentOnline TA        0.279199   0.027831  10.032  < 2e-16 ***
DepositTypeNon Refund         0.534867   0.012216  43.784  < 2e-16 ***
DepositTypeRefundable        -0.043064   0.033426  -1.288 0.197626
CustomerTypeGroup            -0.123818   0.025462  -4.863 1.16e-06 ***
CustomerTypeTransient         0.017121   0.011005   1.556 0.119784
CustomerTypeTransient-Party -0.061949   0.012131  -5.107 3.29e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.391 on 39920 degrees of freedom
Multiple R-squared:  0.2405,    Adjusted R-squared:  0.2379
F-statistic: 90.95 on 139 and 39920 DF,  p-value: < 2.2e-16
```

From the model we can infer that:

a. P-value is <2.2e-16 which is a good value

b. The adjusted R-squared value is 0.2379 which signifies low accuracy but greater than just the numeric type.

c. Some Categorical Variables although are seen to have high significance.

Lastly, we have the dependent variable that we will predict which is IsCanceled while we are using all remaining variables as independent variables.

Linear Model

```
```{r}
lmOut <- lm(IsCanceled ~ ., data = hotel_data)
summary(lmOut)
```
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.92744 -0.25199 -0.07552  0.22255  2.27017

Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.604e-01  7.661e-02   3.399 0.000677 ***
LeadTime                    8.883e-04  2.266e-05  39.206  < 2e-16 ***
StaysInWeekendNights        1.808e-02  2.252e-03   8.028 1.02e-15 ***
StaysInWeekNights           6.126e-03  1.085e-03   5.648 1.63e-08 ***
Adults                      2.300e-02  2.716e-03   8.468  < 2e-16 ***
Children                    3.987e-02  6.002e-03   6.643 3.11e-11 ***

BookingChanges             -3.638e-02  2.554e-03 -14.246  < 2e-16 ***
DepositTypeNon Refund       3.915e-01  1.134e-02  34.539  < 2e-16 ***
DepositTypeRefundable      -3.534e-02  3.009e-02  -1.175 0.240185
CustomerTypeGroup           1.363e-03  2.347e-02   0.058 0.953699
CustomerTypeTransient       9.094e-02  1.004e-02   9.061  < 2e-16 ***
CustomerTypeTransient-Party 5.220e-03  1.111e-02   0.470 0.638305
RequiredCarParkingSpaces   -2.518e-01  5.231e-03 -48.130  < 2e-16 ***
TotalOfSpecialRequests     -7.217e-02  2.456e-03 -29.387  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3518 on 39890 degrees of freedom
Multiple R-squared:  0.3857,    Adjusted R-squared:  0.3831
F-statistic: 148.2 on 169 and 39890 DF,  p-value: < 2.2e-16
```

From the results of the Linear Model, we can infer a few points:

a. P-value is <2.2e-16 which is a good value

b. The adjusted R-squared value is 0.3831 which signifies low accuracy.

c. The result has a lot of variables out of which there are some variables like a country which signifies the low significance in the model.

d. Categorical variables have high significance on the overall impact.

2. Support Vector Machine

Support-vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.

```
library(kernlab);library(caret)

hotel_data$IsCanceled <- as.factor(hotel_data$IsCanceled)
trainList <- createDataPartition(y=hotel_data$IsCanceled,p=0.7,list=FALSE)
trainSet <- hotel_data[trainList,]
testSet <- hotel_data[-trainList,]

svmModel <- train(IsCanceled ~ ., data=trainSet, method='svmRadial')
svmPred <- predict(svmModel, newdata=testSet)

confusion <- confusionMatrix(svmPred, testSet$IsCanceled)
confusion
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 8271 1569
         1  410 1767

               Accuracy : 0.8353
                 95% CI : (0.829, 0.8427)
    No Information Rate : 0.7204
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5405

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

            Sensitivity : 0.9398
            Specificity : 0.5328
         Pos Pred Value : 0.8294
         Neg Pred Value : 0.7977
             Prevalence : 0.7205
         Detection Rate : 0.6927
   Detection Prevalence : 0.8035
      Balanced Accuracy : 0.7301

       'Positive' Class : 0
```

The accuracy of the SVM is greater than that of Linear Model by a huge margin.

3. Association Mining - Apriori

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in a dataset. It is intended to identify strong rules discovered in the dataset using some measures of interestingness.

	LHS	RHS	support	confidence	coverage	lift	count
	All	All	All	All	All	All	All
[18]	{AssignedRoomType=H,ReservedRoomType=H}	{IsCanceled=1}	0.006	0.420	0.015	1.511	245.000
[3]	{ReservedRoomType=H}	{IsCanceled=1}	0.006	0.408	0.015	1.468	245.000
[22]	{AssignedRoomType=G,ReservedRoomType=G}	{IsCanceled=1}	0.016	0.407	0.039	1.467	643.000
[7]	{ReservedRoomType=G}	{IsCanceled=1}	0.016	0.400	0.040	1.441	644.000
[4]	{AssignedRoomType=H}	{IsCanceled=1}	0.006	0.353	0.018	1.270	251.000
[9]	{AssignedRoomType=G}	{IsCanceled=1}	0.016	0.351	0.046	1.265	651.000
[20]	{AssignedRoomType=C,ReservedRoomType=C}	{IsCanceled=1}	0.007	0.344	0.022	1.239	300.000
[5]	{ReservedRoomType=C}	{IsCanceled=1}	0.008	0.330	0.023	1.189	303.000

The following screenshots show that the strongest variables that are contributing towards cancellations across the dataset are the AssignedRoomType and ReservedRoomType.

4. Decision Tree

A decision Tree is a supervised learning algorithm that works both categorical and continuous input and output variables. Below we can see that variables like AssignedRoomType, BookingChanges, CustomerType, DepositType, LeadTime, MarketSegment, PreviousCancellations, RequiredCarParkingSpaces and TotalofSpecialRequests are important as compared to other variables for predicting the cancelations.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 9370 2159
         1  468 1622

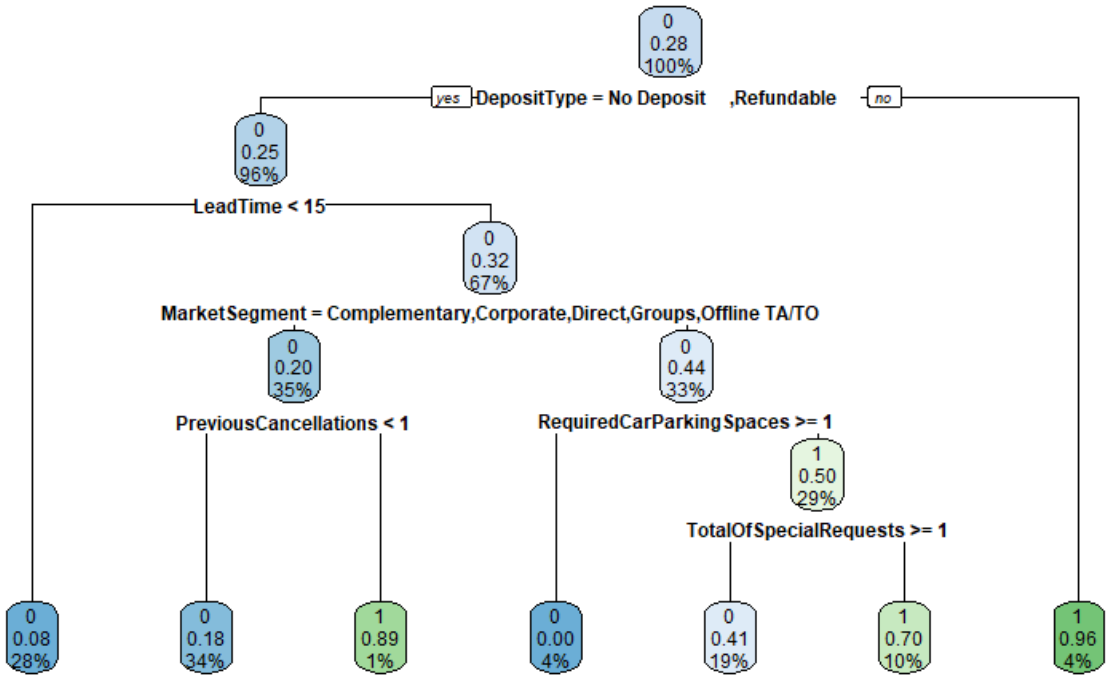
               Accuracy : 0.8071
                 95% CI : (0.8004, 0.8137)
    No Information Rate : 0.7224
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4423

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

            Sensitivity : 0.9524
            Specificity : 0.4290
         Pos Pred Value : 0.8127
         Neg Pred Value : 0.7761
             Prevalence : 0.7224
         Detection Rate : 0.6880
   Detection Prevalence : 0.8465
      Balanced Accuracy : 0.6907

       'Positive' Class : 0
```



```
> varImp(cartTree)
                                Overall
AssignedRoomType              281.0172
BookingChanges                443.0938
CustomerType                  514.3913
DepositType                  1108.2640
LeadTime                     1507.7840
MarketSegment                1663.5734
PreviousCancellations        1202.8846
RequiredCarParkingSpaces     2197.1721
TotalOfSpecialRequests        624.6983
StaysInWeekendNights            0.0000
StaysInWeekNights               0.0000
Adults                          0.0000
Children                        0.0000
Babies                          0.0000
Meal                            0.0000
IsRepeatedGuest                 0.0000
PreviousBookingsNotCanceled     0.0000
ReservedRoomType                0.0000
```

Recommendation

After looking at the exploratory analysis and predictive learning model we are able to identify some key recommendations that can help the hotel in identifying why people cancel and how they can try to avoid that. Given below are our recommendations:

1. Setting up a loyalty program for new customers so that they continue coming and do not cancel.
2. For transient customers we can run a promotional deal or offer them stay at a cheaper price so that they choose our hotel.
3. Allowing bookings to be made only 3 months prior to check in date as majority of our cancellation are due to time frame being large.
4. For minimizing the cancellation of groups that are booked via Travel Agents, we can set up free amenities like a gym, pool area, recreational area, free high-tea, food or beverages to a certain limit so that they can enjoy their stay in the hotel.
5. Although assigned rooms and reserved rooms do not constitute a huge part in cancellation, we found that many cancellations happen with people who have rooms H, G, and C assigned to them compared to other rooms. Maybe changing the layout of room can be useful here.
6. Since our hotel has guest from all around the world, we can have themes or food that is consistent with that country so that they feel more convinced to come at our hotel and they enjoy their stay.