Hotel Bookings & Cancellations

Final Technical Report

Submitted by:
Ameya Mahalaxmikar
Christopher McKinley
Himamshu Chandrashekara
Sazal Sthapit

IST.687.M007.FALL21
INTRODUCTION TO DATA SCIENCE

List of Abbreviations

EDA	Exploratory Data Analysis
ML	Machine Learning
NIR	No Information Rate
SVM	Support Vector Machines
TTC	Time To Complete
URL	Universal Resource Locator
xGB	eXtreme Gradient Boosting

Table of Contents

Abstract

Within the hotel industry, booking cancellations by customers are a great cause of concern. Countless decisions need to be made based on the number of guests staying as well as the duration that they are staying. When bookings are cancelled, they can disrupt well thought out logistical plans. Being able to provide insight into why guests cancel hotel bookings as well as determining what factors may be able to better predict those that may cancel could greatly increase net revenue while also preventing overbooking and poor customer experience. This document demonstrates our analysis into using data science (exploratory data analysis as well as machine learning models) to predict booking cancellations and provide actionable insights to guide hotel management to make more informed decisions. Using the data provided, we were able to determine seven variables that were significant determinants on hotel cancellations, train a machine learning model to more than 75% balanced accuracy, and provided recommendations based on those determinants.

Section I: Introduction

Net predicted hotel bookings (i.e., total bookings minus predicted cancellations) is a key variable in the hospitality industry. However, uncertainties in predicting booking cancellations make total demand forecasts difficult, and the decisions thereof, risky. A direct consequence of this is seen on the revenues and bottom-lines of these hotels, as well as in their operations.

In this project, based on the dataset for a hotel, we have tried to identify a predictive model that helps the hotel make a reasonably accurate booking cancellation prediction through:

a) comparative evaluations of machine learning predictive models, and
b) exploration of data through statistical analyses.

In the real world, such models help hotels improve their demand forecasts, and better understand their actual demands. In this exercise, however, we have not only focused on improving forecast accuracy but also on interpreting the results of these predictive models.

As such, we started by listing down a few business questions regarding booking cancellations that the given data might be able to answer. For example, does longer lead time increase or decrease the chances of cancelling a booking? Are repeat guests more likely to cancel their bookings compared to the first-time guests? Do bookings made through different market segments influence the likelihood of cancelling a booking? Are some types of customers more likely to cancel their bookings than others? (See Appendix-2 for the full set of preliminary questions).

Next, we cleaned and munged the given dataset, and followed up with a correlation analysis of the variables. We then experimented with four types of machine learning algorithms to get an initial idea of which variables might be important in determining the cancellation of bookings. In parallel, we also performed EDA (Exploratory Data Analysis) to discover patterns, spot anomalies, test hypotheses and check assumptions with the help of statistical and graphical methods. Then, we delved deeper into some of the analyses based on the results of preliminary EDA.

For the models, we compared the performances of our four models, and determined that Random Forest (ranger) is the best one for this case (see Section III -> Step 3 for details). We then worked towards fine-tuning that ML model.

The findings and the recommendations based on our EDA and ML model optimization are listed down in Section V of this report. We selected these findings and recommendations based on the overall objective of this exercise, which is to provide the hotel in question with actionable insights so that it can minimize booking cancellations and predict such cancellations more accurately in the coming days.

Section II: Dataset, Variables and Assumptions

What dataset are we using?

```
**
pristine <- data.frame(read_csv("https://intro-datascience.s3.us-east-2.amazonaws.com/Resort01.csv")) #puts CSV into a dataframe called pristine

dim(pristine)
#shows 40,060 rows and 20 columns
**
```

The dataset that we are using consists of booking records for a hotel. It has 40,060 booking records and 20 columns. The full description of these 20 columns can be found in Appendix-1: Metadata. The variable that we are trying to predict is the first column named 'IsCanceled' which denotes whether that booking (the row) was eventually cancelled or not. In other words, we used the rest of the data columns as independent variables to predict this dependent variable.

Which variables did we use?

Initially, we used all the variables to complete our Exploratory Data Analysis (EDA) (see Section III – Step 4 for the full description of our EDA). However, while developing models, we excluded the 'Country' variable for two reasons. First, the svm algorithm flagged it as a near-zero variance variable, which made it go into infinite loops, hence not allowing the model to conclude. Second, we think it is ethically wrong to categorize a behavior such as likelihood to cancel hotel bookings by nationality.

```
**
# Including 'country' variable generate warnings in the svm model:
svm.model1 <- train(IsCanceled ~., data=trainSet,
                    method="svmRadial",
                    trControl=trctrl,
                    preProcess=c("center", "scale"),
                    tuneLength=10)

#Corresponding warning message:
"Warning in preProcess.default(method = c("center", "scale"), x = c(7, 13,   :
  These variables have zero variances: CountryBHS, CountryCOM, CountryEGY, CountryGGY, CountryJOR, CountryMAC, CountryMKD, CountryPLW, CountrySAU, CountrySYR, CountryTUN"
**
```

What assumptions and limitations do we have about the data?

This section describes some of the assumptions and limitations we have about the data that have some bearings on our further EDA and ML works. Those assumptions are as follows:

a) We assume there are no duplicate rows in the dataset. Although some 8,000 rows have exactly same values for all the columns, there is no way for us to determine if those rows are different bookings or just erroneous repetitions.
b) The dataset does not have timestamps on booking records. This is limiting in several ways. For example, we don't know when exactly was a particular booking cancelled (we only know the lead time). We also cannot do an analysis on how booking cancellations pattern may be different in different seasons.

Section III: The Process

This section describes the sequence of steps we took in performing EDA, and in developing the final ML model.

Step 1: Preprocessing

i. **Check for NA and the missing values**

We first downloaded the data from the stipulated URL, and assigned it to the variable 'pristine'. This is the raw dataset (for us). We then created a working copy of this 'pristine' dataset and named it 'hoteldata'. The first thing we did was that we checked if it had an NA anywhere in the database.

```
**
anyNA(hoteldata) #returned FALSE

**
```

Hence, we determined that there are no empty or undefined values anywhere in the dataset. However, we did find 'NULL' as values in 464 rows under the column 'Country' as follows:

```
**
sum(hoteldata$Country==”NULL”) #Result = 464, i.e. 464 rows
do not have any countries assigned to them. This became
another basis to not consider ‘Country’ in our ML
exploration.
**
```

ii. **Change categorical variables into factors**

```
**
hoteldata <- mutate_if(hoteldata, is.character,  factor) #makes
the categorical codes (datatype  characters) as factors
**
```

For example, this code turned the column ‘MarketSegment’ into factor type from character/categorical type.

iii. Change binary variables into factors

```
**
hoteldata$IsCanceled <-  as.factor(hoteldata$IsCanceled) #makes
the binary as factors
hoteldata$IsRepeatedGuest <-  as.factor(hoteldata$IsRepeatedGuest
) #makes the binary as factors
**
```

These were numeric (binary) codes for categorical data. So, we converted them into factors.

iv. Combine values

```
**
levels(hoteldata$Meal) <-
list(UndefinedSC=c("Undefined","SC"),BB=c("BB"),FB=c("FB"),HB=c("
HB")) #combine Undefined and SC into one (since they're the
same)
**
```

Under ‘Meal’ variable, values assigned as ‘Undefined’ and ‘SC’ meant the same thing. So we replaced all ‘Undefined’ and ‘SC’ under column ‘Meal’ as ‘UndefinedSC as follows:

v. Create subsets

We created subsets of ‘hoteldata’ based on the certain criteria. For example, we divided it into repeat customers and first-time customers as thus:

```
**
RepeatGuestData <- hoteldata %>% filter(IsRepeatedGuest == 1)
FirstTimeGuestData <- hoteldata %>% filter(IsRepeatedGuest == 0)
**
```

We also divided the ‘hoteldata’ into subsets of canceled and non-cancelled bookings as thus:

```
**
CanceledBookingData<- hoteldata %>% filter(IsCanceled == 1)
```

```
NonCanceledBookingData <- hoteldata %>% filter(IsCanceled == 0)
**
```

At the end of the Pre-processing, we thus have the following datasets:

S.N.	Dataframe	Description
1	pristine	Raw data (as obtained from the URL provided)
2	hoteldata	Checked for blank and NULL values; binary values converted to factors; categorical values converted to factors
3	CanceledBookingData	Subset of hoteldata that contains only cancelled bookings
4	NonCanceledBookingData	Subset of hoteldata that contains only non-cancelled bookings
5	RepeatGuestData	Subset of hoteldata that contains booking records of only repeat guests
6	FirstTimeGuestData	Subset of hoteldata that contains booking records of only first-time guests

Step 2: Formulating Business Questions

We then started formulating business questions around booking cancellations based on the dataset. These questions are directly related to the ‘IsCanceled’ variable, for example:

i. Do longer lead times increase the probability of cancellation?
ii. Does deposit type have significant influence on determining booking cancellations?
iii. Do bookings made through different means such as Travel Agents and Tour Operators have significant role in determining cancellations?
iv. Are repeat guests more likely to cancel or are new guests more prone to cancel their bookings?
v. Is previous booking cancellation (or no cancellations) a good predictor of future booking cancellations for that customer?
vi. Do many booking changes ultimately lead to increased likelihood of a cancellation?

Note: Please see Appendix 2 for the full list of EDA questions on this report.

Step 3: Experimenting and selecting the Machine Learning (ML) algorithm

Note: This step was done in parallel to *Step 4: Exploratory Data Analysis*. Both of these steps provided inputs to each other. For example, we pursued further EDA on variables determined as important by ML models. On the other hand, we used these models to check whether the variables deemed important through EDA are also ranked as important variables by different ML methods. In other words, we used these ML models to see if they concur with our intuition-based EDA questions.

Phase I: Experimenting with different ML algorithms

We experimented with four ML methods at this stage. This allowed for streamlined testing of multiple models, and simplified the amount of code needed to complete the task.

All models used the same 70%-30% split of data for training and testing, as well as a training control method of repeated k-fold cross validation, with 5 (k) separations and 3 repeats. A set seed was also used in order to create reproducibility.

```
**
trctrl <- trainControl(method="repeatedcv", number=5, repeats=3)
**
```

In order to decrease the amount of time needed to complete the training of each model, the R libraries, “parallel” and “doParallel” were used to enable the effective usage of multicore processors on Windows based operating systems.

```
**
library(parallel)
library(doParallel)
no_cores <- detectCores() - 4 # Calculate the number of cores (subtract
4, so you don't tie it up)

cl <- makePSOCKcluster(no_cores) # create the cluster for caret to use
registerDoParallel(cl)
**
```

The four ML models that we experimented with at this stage are:

1. Support Vector Machines with Radial Basis Function Kernel (svmRadial)
2. Classification and Regression Trees (rpart)
3. Random Forest (ranger)
4. eXtreme Gradient Boosting Trees (xgbTree)

Initially, we used all the variables (except ‘Country’) in the dataset to feed into the four models. In effect, we did not do any feature selection at this point. The codes that we used for the four models along with their parameters are shown below:

```
**
#svmRadial
svm.model <- train(IsCanceled~.,method="svmRadial",
data=trainSet,trControl=trctrl)

#rpart
model_rpart <- train(IsCanceled~.,method="rpart",
data=trainSet,trControl=trctrl)
```

```
#ranger
model_ranger <-
train(IsCanceled~.,method="ranger",data=trainSet,trControl=trctrl,impor
tance = 'permutation') #permutation was added in order to get this
model type to calculate variable importance

#xgbTree
model_xgbtree <- train(IsCanceled~.,method="xgbTree",
data=trainSet,trControl=trctrl)
**
```

A comparison chart that shows the results produced by the four methods is shown below:

Model	Accuracy	Balanced accuracy	95% CI	NIR	P-Value	Kappa	Sensitivity	Specifity	TTC (min)
svmRadial	.8335	0.7520	0.8267, 0.8401	0.7224	< 2.2e-16	0.5484	0.9351	0.5689	10.28
rpart	.8	0.6768	0.7927, 0.8071	0.7224	< 2.2e-16	0.4148	0.9537	0.3999	0.1841
ranger	0.851	0.7962	0.8446, 0.8574	0.7224	< 2.2e-16	0.6147	0.9195	0.6730	17.77
xgbTree	0.842	0.7716	0.8353, 0.8485	0.7224	< 2.2e-16	0.5796	0.9298	0.6133	2.427

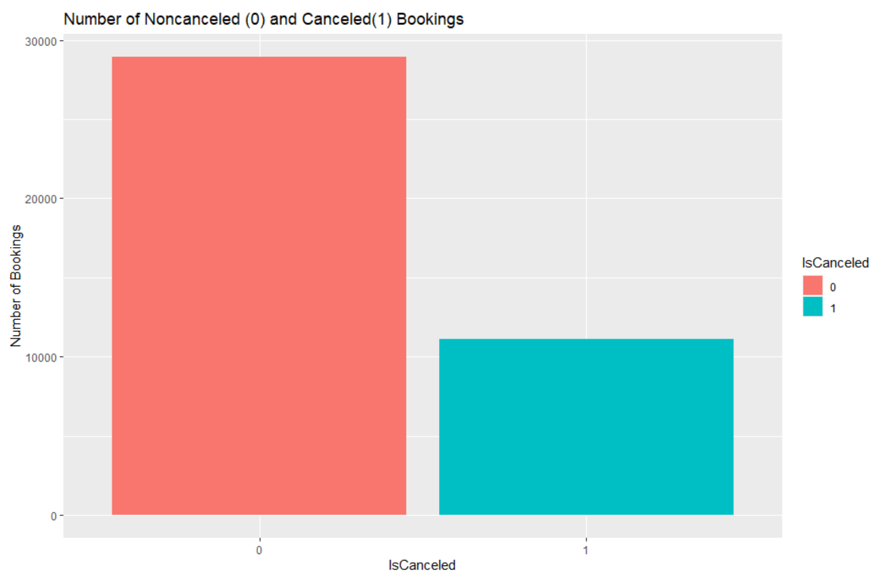
The models were trained again using a 10-fold cross validation which is a more commonly used amount, however there was not a significant difference other than in time to completion (TTC) which differed unexpectedly in some cases. The No Information rate (NIR) and P-Value were the same for all models. Accuracy was slightly skewed due to the imbalanced data which is discussed in detail below. The 95% Confidence Interval, for all models is above the no information rate, which is the same for all models, along with the P-Value. Specificity is how many of the positive cases, (IsCanceled=0 or No) the model could predict correctly. Specificity is how many of the negative cases, (IsCanceled=1 or Yes) the model could predict correctly. Time to completion for the models can be a crucial factor, for instance if we need to perform almost constant model training based on a constant flow of new booking data. The ability to efficiently train a model to perform near real time predictions would be key.

Model	Accuracy	Balanced accuracy	95% CI	NIR	P-Value	Kappa	Sensitivity	Specifity	TTC (min)
svmRadial	0.8336	0.7521	0.8268, 0.8402	0.7224	< 2.2e-16	0.5486	0.9353	0.5689	23.68
rpart	0.8	0.6768	0.7927, 0.8071	0.7224	< 2.2e-16	0.4148	0.9537	0.3999	0.10327456667
ranger	0.8526	0.7991	0.8462, 0.8589	0.7224	< 2.2e-16	0.6195	0.9195	0.6787	39.98483
xgbTree	0.8437	0.7756	0.8371, 0.8502	0.7224	< 2.2e-16	0.5858	0.9288	0.6223	5.356659

Phase II: Selecting the final model and fine-tuning it

We used the following metrics to choose the final model:

Balanced Accuracy – Balanced Accuracy is a preferred metric when dealing with imbalanced data. For our data set 27.8% of the data is cancelled and 72.2% is not canceled indicating an imbalanced dataset also shown in the figure below. Balanced Accuracy includes the Sensitivity and Specificity since imbalanced Accuracy is determined from (Sensitivity+Specificity)/2. For this business case, we are not dealing with trying to determine a life and death situation, so there is not a significantly greater weighting of sensitivity vs specificity. The model with the highest accuracy was the ranger model, followed by xgbTree, svmRadial and rpart.




```
**
canceleddata <- ggplot(hoteldata, aes(x= IsCanceled, binwidth = 15,
fill=IsCanceled)) + geom_histogram(stat = "count")+ labs(y="Number of
Bookings", title = "Number of Noncanceled (0) and Canceled(1)
Bookings")
**
```

Kappa- The Kappa, which compares the observed accuracy with an expected accuracy, also considering random chance is also a preferred metric when utilizing imbalanced datasets. Again, the model with the highest Kappa was the ranger model, followed by xgbTree, svmRadial and rpart. This follows the same trend as the balanced accuracy.

Due to the ranger model being the best model in terms of Balanced Accuracy and Kappa, it was selected as the model to continue tuning.

The ranger model was tuned with the code below

```
**
trctrl <- trainControl(method="repeatedcv", number=10,
repeats=3,classProbs = TRUE, savePredictions = T) #classProbs
and savePredictions were needed to make a model comparison library work
correctly

model_ranger <- train(IsCanceled ~.,
      data=trainSet,method="ranger",trControl=trctrl,
      preProcess=c("center", "scale"),tuneLength=3,
      importance = 'permutation', num.threads = 12)
      #the ranger method is one of the few within caret that work with
      #windows multicore processors by default, num.threads is how it
      #is enabled

predictValues_ranger <- predict(model_ranger, newdata = testSet)
> confusionMatrix(predictValues_ranger, testSet$IsCanceled)
**
```

Notes:

1. The train and testset information were not changed.
2. The model took 34.8817 minutes to run.

****Output****

CONFUSION MATRIX AND STATISTICS

	Reference	
Prediction	**NO**	**YES**
NO	7958	1080
YES	723	2256

Accuracy	: 0.85
95% CI	: (0.8435, 0.8563)
No Information Rate	: 0.7224
P-Value [Acc > NIR]	: < 2.2e-16
Kappa	: 0.6132
Mcnemar's Test P-Value	: < 2.2e-16
Sensitivity	: 0.9167
Specificity	: 0.6763
Pos Pred Value	: 0.8805
Neg Pred Value	: 0.7573
Prevalence	: 0.7224
Detection Rate	: 0.6622
Detection Prevalence	: 0.7521

```
Balanced Accuracy          : 0.7965

'Positive' Class           : NO
 **
```

****ranger variable importance (table)**

```
# only 20 most important variables shown (out of 44)

                             Overall
LeadTime                      100.00
MarketSegmentOnline TA         77.73
DepositTypeNon Refund          66.73
RequiredCarParkingSpaces       46.77
TotalOfSpecialRequests         44.68
StaysInWeekNights              37.47
CustomerTypeTransient          33.20
StaysInWeekendNights           31.42
CustomerTypeTransient-Party    30.73
MarketSegmentOffline TA/TO     29.94
AssignedRoomTypeD              29.64
MarketSegmentGroups            23.77
PreviousCancellations          21.71
ReservedRoomTypeD              20.77
MealBB                         19.30
BookingChanges                 19.13
MealHB                         18.03
MarketSegmentDirect            16.18
Adults                         14.58
AssignedRoomTypeE              10.22
```

The Model was run again on a subset of the data which only included repeat guests (1778 observations). That training only took 46 seconds to complete.

```
Confusion Matrix and Statistics

            Reference
Prediction  NO    YES
       NO   496   10
       YES    4   23

Accuracy                 : 0.9737
95% CI                   : 0.9563, 0.9856)
No Information Rate      : 0.9381
P-Value [Acc > NIR]      : 0.0001129

Kappa                    : 0.7529

Mcnemar's Test P-Value   : 0.1814492
```

```
Sensitivity          : 0.9920
Specificity          : 0.6970
Pos Pred Value       : 0.9802
Neg Pred Value       : 0.8519
Prevalence           : 0.9381
Detection Rate       : 0.9306
Detection Prevalence : 0.9493
Balanced Accuracy    : 0.8445

'Positive' Class     : NO
```

ranger variable importance

```
#only 20 most important variables shown (out of 44)
```

	Overall
PreviousCancellations	100.000
Adults	12.097
LeadTime	11.888
PreviousBookingsNotCanceled	11.114
AssignedRoomTypeD	10.473
StaysInWeekendNights	8.856
MarketSegmentGroups	8.502
MarketSegmentCorporate	8.037
TotalOfSpecialRequests	7.035
MarketSegmentDirect	6.843
StaysInWeekNights	6.724
RequiredCarParkingSpaces	6.650
ReservedRoomTypeD	5.212
CustomerTypeTransient-Party	4.927
ReservedRoomTypeE	4.766
BookingChanges	4.054
ReservedRoomTypeG	3.504
MealHB	3.233
MealBB	3.059
CustomerTypeTransient	2.137

With the Repeat Guest subset, PreviousCancellations and PreviousBookingsNotCancelled were significantly more important than with the full dataset that included first time guests; 100 and 11.11 for the repeat guest, compared to 21.71 and 1.83 for the full dataset. In order to get the best insight from repeat guests, previous cancellations and previous bookings not cancelled they need to be run with a model separate from first time guests.

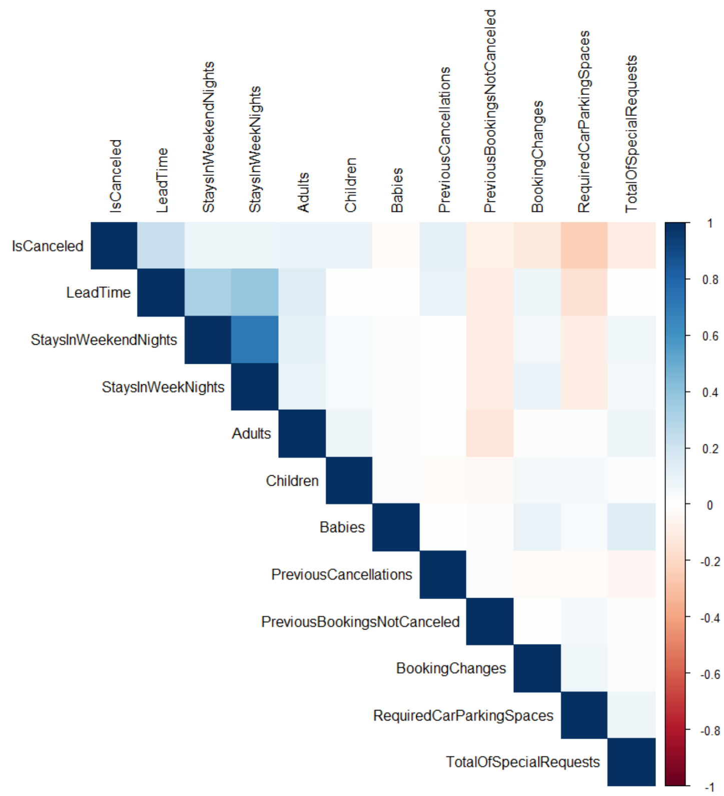
Step 4: Exploratory Data Analysis

Note: This step was done in parallel to *Step 3: Experimenting and selecting the Machine Learning (ML) algorithm*. Both steps provided inputs to each other. For example, we pursued further EDA on variables determined as important by ML models. On the other hand, we used the model results obtained in Step 3 to check whether the variables deemed important through EDA are also ranked as important variables by those ML methods.

Section I: As the first step to give some legitimacy to our intuition-based business questions (see Appendix 2), we performed a dataset-wide correlation test on the numeric variables as follows:

```
**
#Import data from part 1:
dfm=hoteldata

#Create local copy of dfm for manipulation
dfm_temp <- dfm

#Change 'IsCanceled' back to numeric
dfm_temp$IsCanceled <- as.numeric(dfm_temp$IsCanceled)

#Check correlation:
dfnum = dplyr::select_if(dfm_temp, is.numeric) #select only numeric
variables
dfnum = data.frame(lapply(dfnum, function(x)
as.numeric(as.character(x)))) #loop through columns and change factor
variables into numeric
res=cor(dfnum)
dev.new()
corrplot(res, method="color", type="upper", tl.col="black")
**

**Output**
(on the next page)
```

Output



Results

1. IsCanceled and LeadTime have high correlation (positive)
2. IsCanceled and RequiredCarParkingSpaces have high correlation (negative)
3. IsCanceled and PreviousCancellations have high correlation (positive)
4. IsCanceled and Babies have almost no correlation
5. IsCanceled has slightly positive correlation with StaysInWeekendNights, StaysInWeekNights, Adults, and Children
6. IsCanceled has slightly negative correlation with PreviousBookingsNotCancelled, BookingChanges, and TotalSpecialRequests

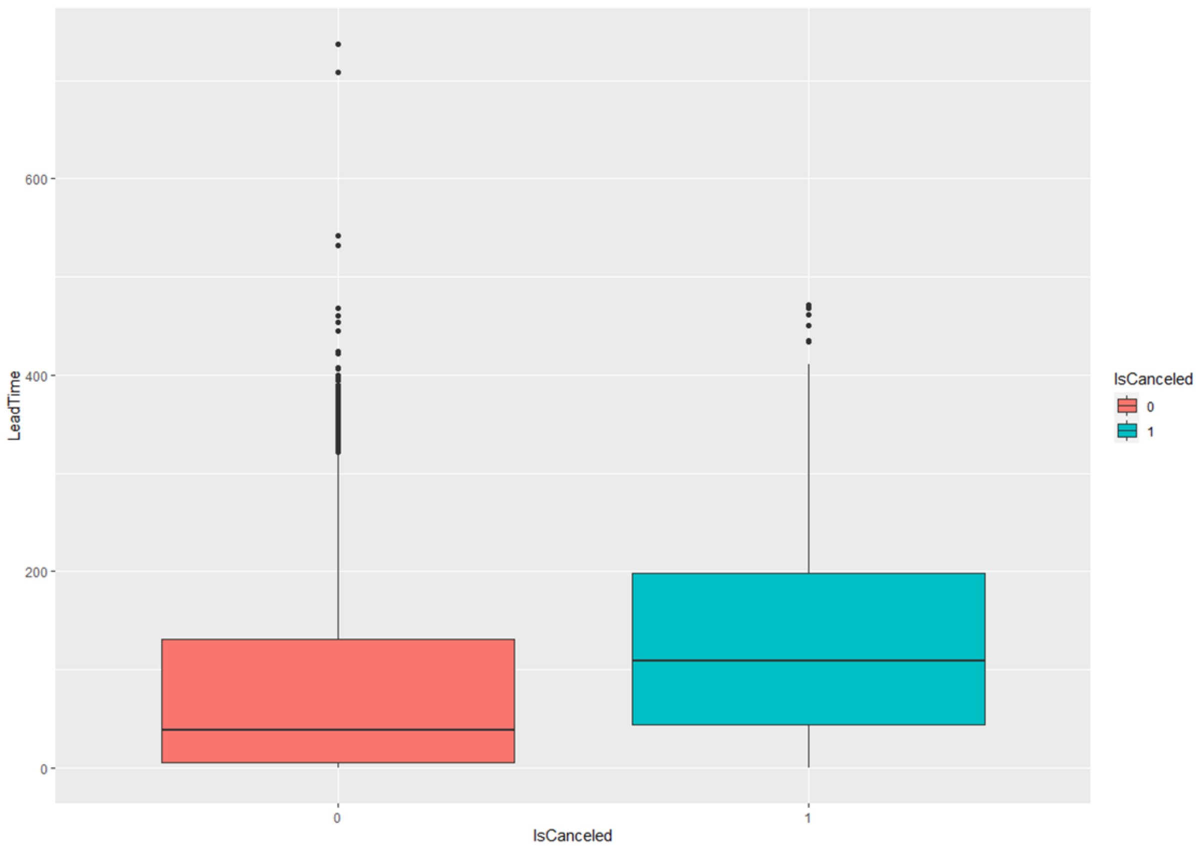
Interpretation

- Through this very preliminary correlation comparison, it seems that LeadTime, RequiredCarParkingSpaces, and PreviousCancellations are important variables in determining IsCanceled variable. So, to start with, we performed further EDA on these variables.

Q. How does lead time vary between cancelled and non-cancelled bookings?

```
**Code
LeadtimeBoxPlot <- ggplot(hoteldata, aes(x = IsCanceled, y = LeadTime,
fill = IsCanceled)) +
  geom_boxplot()
**
```

****Output****



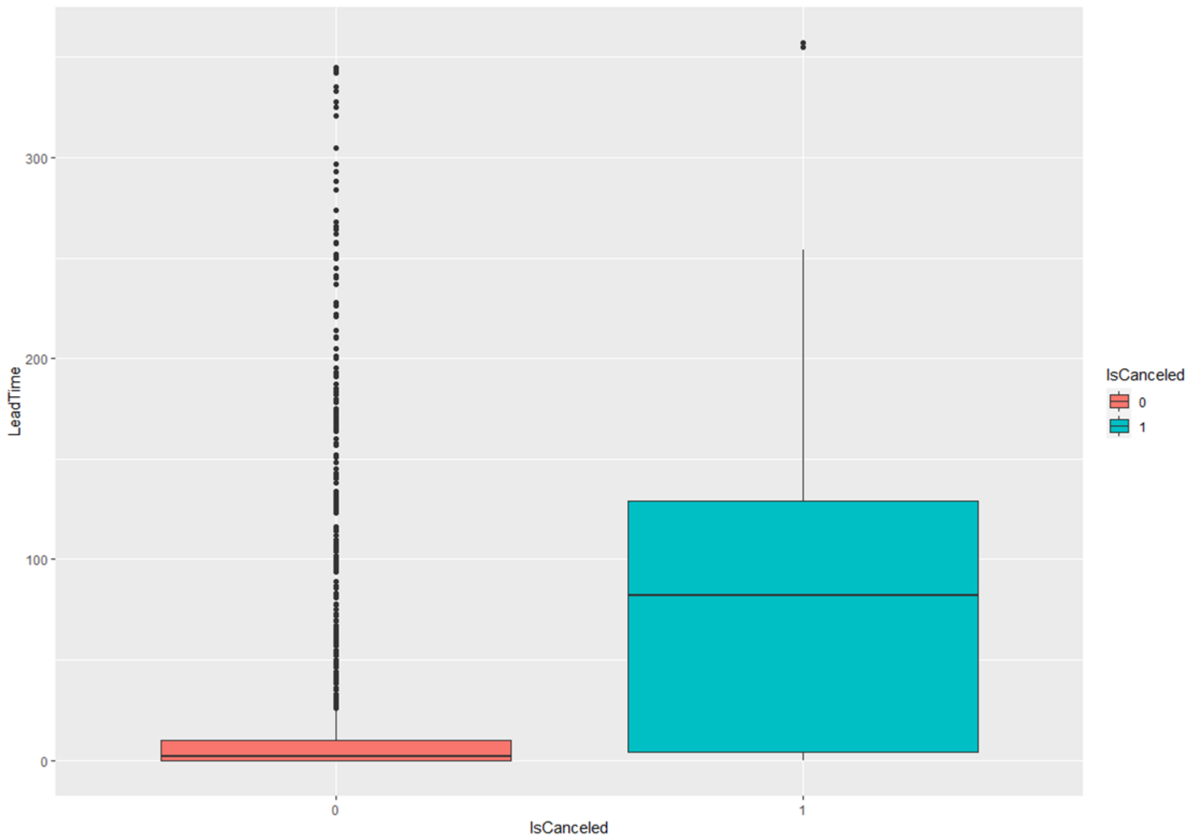
	Non-Canceled Bookings	Cancelled Bookings
Mean	78.84	128.68
Median	38	109
Mode	0 (3079)	0 (157)

Interpretation

- In general, lead time for bookings that do not get cancelled are lower (mean lead time for not-cancelled bookings = 78.84 days) than for the bookings which get cancelled (mean lead time of cancelled bookings = 128.68 days).
- 50% of all non-cancelled bookings have quite a short lead days (just five weeks or less).

Q. How does lead time vary across cancelled and non-cancelled bookings for repeat guests?

```
**
LeadtimeBoxPlot <- ggplot(repeatguestdata, aes(x = IsCanceled, y =
LeadTime, fill = IsCanceled)) +
  geom_boxplot()
**

**Output**
```



	Non-Canceled Bookings	Cancelled Bookings
Mean	21.02	83.77
Median	2	82
Mode	0 (593)	82 (29)

Interpretation

- The same pattern holds for repeat guests as well. Even when repeat guests cancel, the lead time for those bookings is generally higher (mean of 83.77 days) when compared to bookings which are not cancelled (mean of 21.02 days).
- 50% of non-cancelled bookings by repeat guests have a lead time of 2 or less days.
- Mode = 0 days for non-cancelled reservations by repeat guests also suggest that these repeat guests mostly make the reservation right on the day of arriving at the hotel.

Q. How do cancellations compare between first time guests and repeat guests?

We also checked if PreviousBookings and PreviousBookingsNotCancelled have any impact on determining the cancellation of a booking. We did this test by dividing the bookings into those done by repeat guests and those done by first time guests.

```
**Code
hoteldata %>% tabyl(IsCanceled) %>% adorn_totals("row") %>%
adorn_pct_formatting()
RepeatGuestData %>% tabyl(IsCanceled) %>% adorn_totals("row") %>%
adorn_pct_formatting()
FirstTimeGuestData %>% tabyl(IsCanceled) %>% adorn_totals("row") %>%
adorn_pct_formatting()
**

**Output**
```

All Guests	First Time Guests	Repeat Guests
IsCanceled n percent 0 28938 72.2% 1 11122 27.8% Total 40060 100.0%	IsCanceled n percent 0 27271 71.2% 1 11011 28.8% Total 38282 100.0%	IsCanceled n percent 0 1667 93.8% 1 111 6.2% Total 1778 100.0%

Interpretation

- We found that only 6% of repeat guests cancelled their reservations whereas 28% of the first time guests cancelled. It implies that the first time guests are much more prone to cancel their bookings.
- Thus, one recommendation for the hotel to minimize booking cancellation risks is to distribute its bookings among repeat as well as first time guests.

Q. What does BookingChanges tell us about probable cancellations?

```
**Code
hoteldata %>% tabyl(BookingChanges) %>% adorn_totals("row") %>%
adorn_pct_formatting()
NonCanceledBookingData %>% tabyl(BookingChanges) %>%
adorn_totals("row") %>% adorn_pct_formatting()
IsCanceledBookingData %>% tabyl(BookingChanges) %>% adorn_totals("row")
%>% adorn_pct_formatting()
**
```

Output

All Bookings

BookingChanges	n	percent
0	32252	80.5%
1	5469	13.7%
2	1561	3.9%
3	460	1.1%
4	182	0.5%
5	72	0.2%
6	32	0.1%
7	12	0.0%
8	8	0.0%
9	4	0.0%
10	3	0.0%
12	1	0.0%
13	2	0.0%
16	1	0.0%
17	1	0.0%
Total	40060	100.0%

Non-Canceled Bookings

BookingChanges	n	percent
0	22284	77.0%
1	4656	16.1%
2	1326	4.6%
3	403	1.4%
4	155	0.5%
5	64	0.2%
6	24	0.1%
7	10	0.0%
8	6	0.0%
9	4	0.0%
10	2	0.0%
12	1	0.0%
13	2	0.0%
17	1	0.0%
Total	28938	100.0%

Cancelled Bookings

BookingChanges	n	percent
0	9968	89.6%
1	813	7.3%
2	235	2.1%
3	57	0.5%
4	27	0.2%
5	8	0.1%
6	8	0.1%
7	2	0.0%
8	2	0.0%
10	1	0.0%
16	1	0.0%
Total	11122	100.0%

Observations

- Overall, most bookings had no changes (80.5%)
- Non-Cancelled bookings had a higher percentage of changes (23%) than Cancelled bookings (10.4%)

Interpretation

- We could infer that those who are not likely to cancel make more changes to their bookings to make it work (rather than cancelling it altogether).

Q. How does the Amount of Required Car Parking Spaces Affect Cancellations?

```
**Code
hoteldata %>% tabyl(RequiredCarParkingSpaces) %>% adorn_totals("row")
%>% adorn_pct_formatting()

IsCanceledBookingData %>% tabyl(RequiredCarParkingSpaces) %>%
adorn_totals("row") %>% adorn_pct_formatting()

NonCanceledBookingData %>% tabyl(RequiredCarParkingSpaces) %>%
adorn_totals("row") %>% adorn_pct_formatting()

**Output
```

All Bookings

RequiredCarParkingSpaces	n	percent
0	34570	86.3%
1	5462	13.6%
2	25	0.1%
3	1	0.0%
8	2	0.0%
Total	40060	100.0%

Non-Canceled Bookings

RequiredCarParkingSpaces	n	percent
0	23448	81.0%
1	5462	18.9%
2	25	0.1%
3	1	0.0%
8	2	0.0%
Total	28938	100.0%

Cancelled Bookings

RequiredCarParkingSpaces	n	percent
0	11122	100.0%
Total	11122	100.0%

Observation

- Due to the lack of variance in the amount of required car parking spaces (0,1,2,3,8 are the only numbers) it was easier to visualize this data with tables. One thing that immediately stands out is that 100% of bookings that were cancelled did not have 'require a car parking space'. Also of note is that all bookings that had at least one required car parking space were not cancelled. It is also key to note however, that most bookings (86.3%) did not require a car parking space.

Interpretation

- Based on the analysis of the data, it is highly probable that a booking will not be cancelled if there is at least one required parking space associated with that booking. Despite 100% of the cancelled bookings did not require parking spaces, a significant amount of non cancelled bookings did not require parking spaces either, so it is not safe to say that not requiring a parking space is indicative of a potential cancelled booking.

Section II: In this section of EDA, we will delve deeper into the remaining variables deemed as important by our ranger model. Those variables are DepositType, MarketSegment, TotalOfSpecialRequests, and CustomerType.

Q. How does the Deposit Type Affect Cancellations?

To first understand how deposit type may affect cancellations, we first looked at the distribution of the different deposit types.

```
**Code
hoteldata %>% tabyl(DepositType) %>% adorn_totals("row") %>%
adorn_pct_formatting()
NotCancelData %>% tabyl(DepositType) %>%  adorn_pct_formatting(digits =
2, affix_sign = TRUE)
CancelData %>% tabyl(DepositType) %>%  adorn_pct_formatting(digits = 2,
affix_sign = TRUE)
**
```

****Output****

All Bookings

DepositType	n	percent
No Deposit	38199	95.4%
Non Refund	1719	4.3%
Refundable	142	0.4%
Total	40060	100.0%

Non-Canceled Bookings

DepositType	n	percent
No Deposit	28749	99.35%
Non Refund	69	0.24%
Refundable	120	0.41%

Cancelled Bookings

DepositType	n	percent
No Deposit	9450	84.97%
Non Refund	1650	14.84%
Refundable	22	0.20%

Observation

- When looking the full dataset, the subset of non-canceled bookings, and subset of canceled bookings, the "No Deposit" deposit type is the most prevalent. Cancelled Bookings have a higher percentage of non-refundable deposits. When considering the imbalance of the data (72.2% of the data is non canceled bookings) there is still a higher number of canceled bookings with a non-refundable deposit (1650 cancelled, vs 69 non-cancelled) in addition to the non-cancelled bookings having a higher percentage of non-refundable deposits.

Interpretation

- One would assume without fault that if one paid for a non-refundable booking, they would be more likely not to cancel that booking, as they would not be able to get their money back.
- However, based on the analysis done by creating tables of the DepositType categorical variable, as it relates to the IsCanceled categorical variable, it appears that non-refundable deposits did not lead to less canceled bookings. There could be various reasons for this, such as emergency circumstances that caused those making the booking not able to make it to the hotel, or the nonrefundable deposit, despite being the value of the total stay, not being expensive enough to be an adequate deterrent for one to follow through with checking in at the hotel. This is not necessarily a significant problem for the hotel, since the hotel is getting the full amount of money, from that booking, they also potentially incur fewer operating costs for that booking as they wouldn't need to have that room cleaned after that booked customer leaves, and depending on when it was cancelled, they could potentially book another customer in that room. Using the deposit type information, in addition to the lead time information, could potentially yield increased profit. This is due to our interpretation of the data, concluding that the longer the lead time, the more likely a booking would be cancelled. The hotel could set a threshold where if the lead time for a booking is longer than the median lead time for canceled bookings (109 days) then a non-refundable deposit could be required.

Q. Do cancellation patterns vary significantly across market segments?

```
**
hoteldata %>% tabyl(MarketSegment) %>% adorn_totals("row") %>%
adorn_pct_formatting()
NonCanceledBookingData   %>% tabyl(MarketSegment) %>%
adorn_totals("row") %>% adorn_pct_formatting()
CanceledBookingData   %>% tabyl(MarketSegment) %>% adorn_totals("row")
%>% adorn_pct_formatting()
**
```

Output

All Bookings

```
MarketSegment     n percent
Complementary   201    0.5%
    Corporate  2309    5.8%
       Direct  6513   16.3%
       Groups  5836   14.6%
 Offline TA/TO  7472   18.7%
    Online TA 17729   44.3%
        Total 40060  100.0%
```

Non-Canceled Bookings

```
MarketSegment     n percent
Complementary   168    0.6%
    Corporate  1958    6.8%
       Direct  5635   19.5%
       Groups  3362   11.6%
 Offline TA/TO  6334   21.9%
    Online TA 11481   39.7%
        Total 28938  100.0%
```

Cancelled Bookings

```
MarketSegment     n percent
Complementary    33    0.3%
    Corporate   351    3.2%
       Direct   878    7.9%
       Groups  2474   22.2%
 Offline TA/TO  1138   10.2%
    Online TA  6248   56.2%
        Total 11122  100.0%
```

Market Segment	Canceled	Non-canceled	Total bookings	% canceled	% non-canceled
Complementary	33	168	201	16.4179104	83.58208955
Corporate	351	1958	2309	15.2013859	84.79861412
Direct	878	5635	6513	13.4807308	86.51926915
Groups	2474	3362	5836	42.3920493	57.60795065
Offline TA	1138	6334	7472	15.2301927	84.76980728
Online TA	6248	11481	17729	35.2416944	64.7583056

Observation

- Cancellation % for ‘Groups’ is 42% and for ‘Online TA’ is 35%, which are higher than those for the rest of the segments.
- An approximate proportion of cancellations in the remaining 4 out of 6 segments is around 15%. Together, these segments make up 41% of the observations.

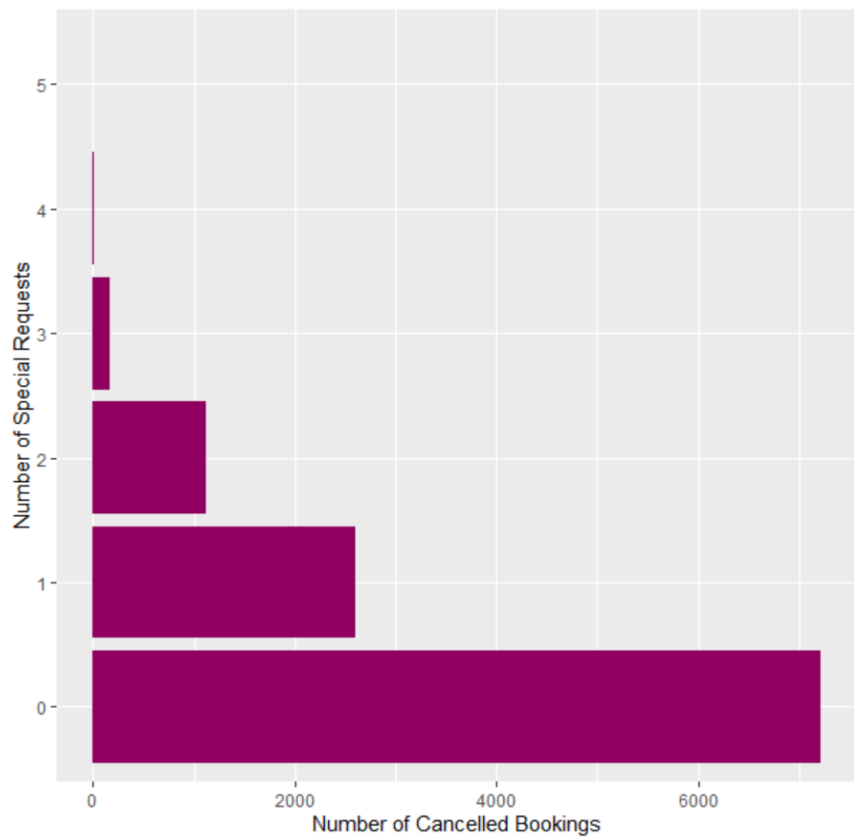
Interpretation

- Across segments, ‘Online TA’ and ‘Groups’ exhibit higher chances of cancellations compared to the other four segments.

Q. Do special requests determine booking cancellations?

```
**Code
cancellationbasedonSpecialRequests<-
data.frame(table(CanceledBookingData$TotalOfSpecialRequests))
dev.new(height= 8, width= 8)
ggplot(cancellationbasedonSpecialRequests, aes(x= Var1, y=Freq)) +
  geom_bar(stat="identity", fill = "#8F0060") + xlab("Number of Special
Requests")+ ylab("Number of Cancelled Bookings")+coord_flip()
dev.off()
```

Output




```
> cancellationbasedonSpecialRequests
  Var1 Freq
1    0 7216
2    1 2597
3    2 1127
4    3  166
5    4   15
6    5    1
>
```

Observation

- Within cancelled bookings, the number of cancellations decreases sharply as the number of special requests increases.

Interpretation

- Like number of booking changes, special requests can be viewed as an attempt to make things work. Hence, it makes sense that bookings with special requests are cancelled less often.

Q. How does the Customer Type Affect Cancellations?

```
**
CanceledBookingData  %>% tabyl(CustomerType) %>% adorn_totals("row")
%>% adorn_pct_formatting()

NonCanceledBookingData  %>% tabyl(CustomerType) %>% adorn_totals("row")
%>% adorn_pct_formatting()
**

**Output**
```

	All Bookings		Non-Canceled Bookings		Cancelled Bookings	
CustomerType	n	percent	n	percent	n	percent
Contract	1776	4.4%	1619	5.6%	157	1.4%
Group	284	0.7%	254	0.9%	30	0.3%
Transient	30209	75.4%	20793	71.9%	9416	84.7%
Transient-Party	7791	19.4%	6272	21.7%	1519	13.7%
Total	40060	100.0%	28938	100.0%	11122	100.0%

Observation

- The highest percentage of Customer Type for all bookings was the Transient booking type (75.4%), meaning the booking was not part of a group. The next highest is Transient-Party, followed by contract and then group. Non Cancelled bookings and cancelled bookings had the same ranking of customer types. Cancelled bookings have a higher percentage of transient bookings (84.7%) compared to non cancelled bookings (71.9%)

Interpretation

- The increase in transient customer type in cancelled bookings (12.8%) compared to non-canceled bookings may be significant enough for ML models to use effectively in determining the likelihood of a booking being cancelled.

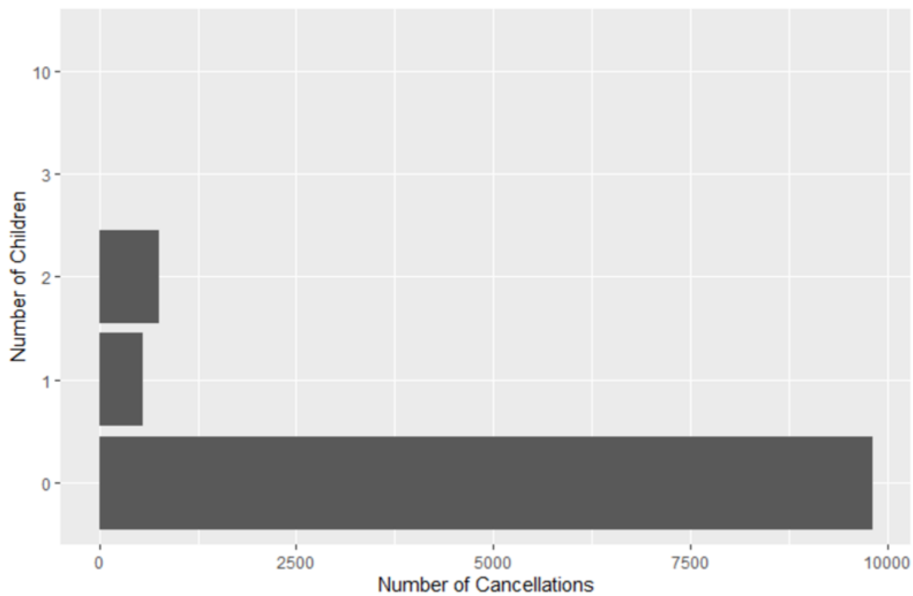
Section III: In this section, we have performed EDAs on some variables deemed unimportant by correlation analysis and ranger model importance table. The reason we did this is to ensure that those correlation analysis and ranger model are not missing out on something important. To that effect, we asked the following questions:

Q. Does number of children affect booking cancellations?

To answer this prompt, we utilized the table function, which allowed us to compartmentalize and compute the number of observations from the cancelled hotel reservations subset of data, in accordance to the number of children in the family that had cancelled their booking. This tabulated data was then converted to a data frame in order to showcase the output graphically, and saved to a variable called, “cancellationbasedonChildren”. In order to obtain a horizontal bar chart for the same, we used the ggplot2 library.

```
**
cancellationbasedonChildren<-
data.frame(table(CanceledBookingData$Children))
ggplot(cancellationbasedonChildren, aes(x = Freq, y = Var1,
main="Number of Cancellations based on Number of Children")) +
geom_bar(stat = "identity") +ylab("Number of Children")+xlab("Number of
Cancellations") +theme_get()
**
```

Output




```
> cancellationbasedonChildren
  Var1 Freq
1    0 9808
2    1  558
3    2  753
4    3    2
5   10    1
> 
```

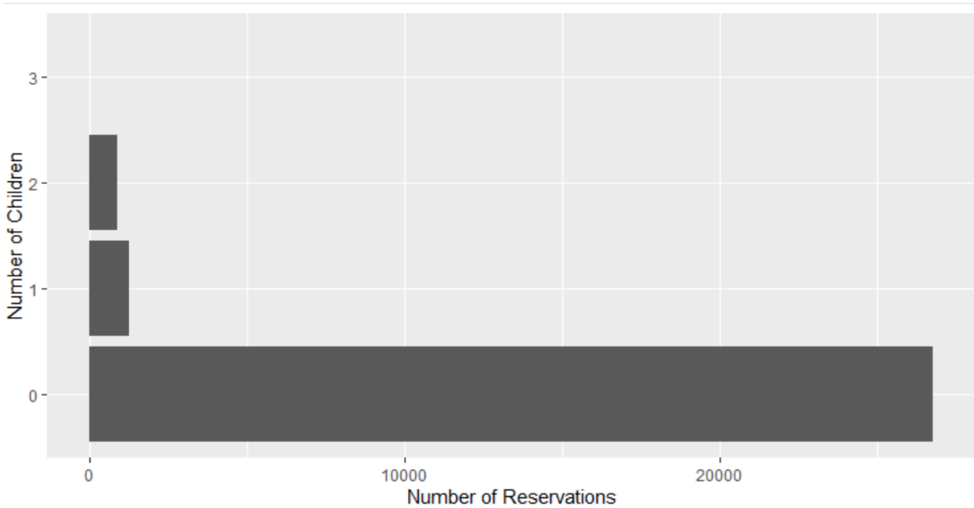
Observation

- We can infer that the maximum number of cancellations are done by families with no children, and the minimum number of cancellations are observed from the family with the greatest number of children i.e., 10. There seems to be an apparent inverse proportionality in the number of children and the number of cancellations, when observing the cancelled hotel reservations subset of data.

But we also checked if the same pattern exists for cancelled bookings as follows:

```
**
noncancellationbasedonChildren<-
data.frame(table(NonCanceledBookingData$Children))
ggplot(noncancellationbasedonChildren, aes(x = Freq, y = Var1,
main="Number of confirmed reservations based on Number of
Children")) + geom_bar(stat = "identity") +ylab("Number of
Children")+xlab("Number of Reservations") +theme_get()
**
```

Output




```
> noncancellationbasedonChildren
  Var1  Freq
1    0 26768
2    1  1280
3    2   875
4    3    15
> 
```

Observation

- From the above plot, we can infer that the maximum number of cancellations are done by families with no children, and the minimum number of cancellations are observed from the family with the greatest number of children i.e., 3. There seems to be an apparent inverse proportionality in the number of children and the number of cancellations, when observing the non-cancelled hotel reservations subset of data.

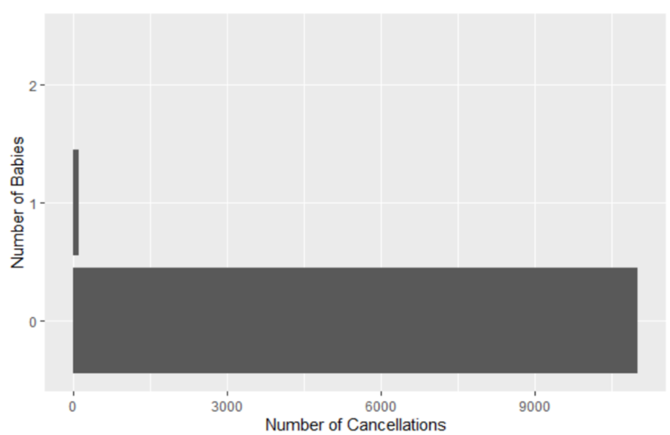
Interpretation

- Since the same patterns exist between number of children and cancellations, and number of children and non-cancellations, we conclude that number of children does not play a significant role in determining booking cancellations.

Q. How does the number of babies affect booking cancellations?

```
**code
cancellationbasedonBabies<-
data.frame(table(CanceledBookingData$Babies))
ggplot(cancellationbasedonBabies, aes(x = Freq, y = Var1,
main="Number of Cancellations based on Number of Babies")) +
geom_bar(stat = "identity") +ylab("Number of
Babies")+xlab("Number of Cancellations") +theme_get()
**
```

****Output****




```
> cancellationbasedonBabies
  Var1  Freq
1    0 11019
2    1   101
3    2     2
> 
```

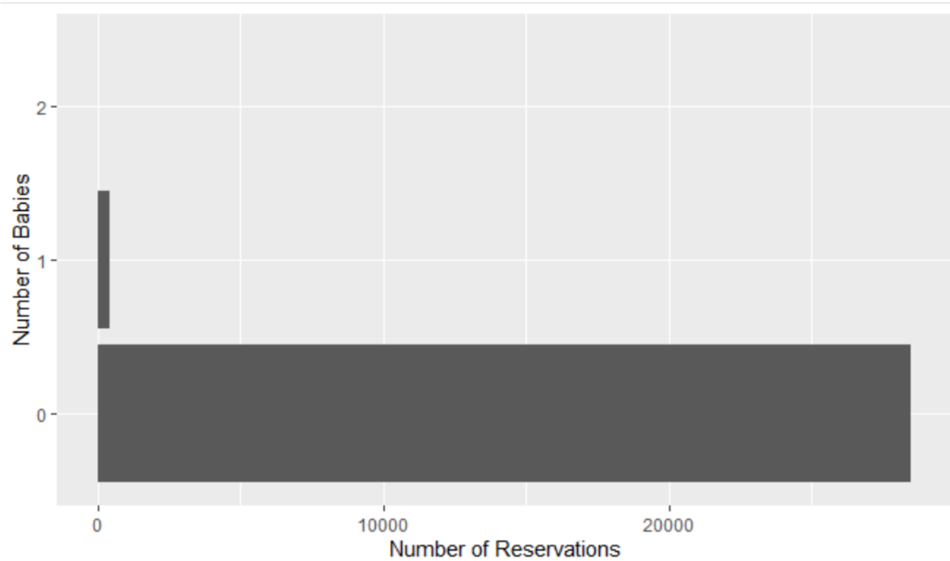
Interpretation

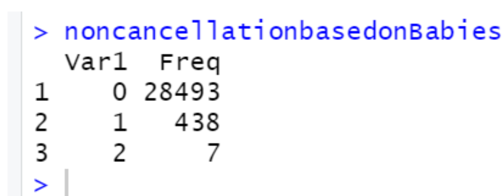
- From the above plot, we can infer that the maximum number of cancellations are done by families with no babies, and the frequency of cancellations decrease with increasing number of babies in the family. Ergo, there seems to be an apparent inverse proportionality in the number of babies and the number of cancellations, when observing the cancelled hotel reservations subset of data.

The same was done for noncancellations

```
**code
noncancellationbasedonBabies<-
data.frame(table(NonCanceledBookingData$Babies))
ggplot(noncancellationbasedonBabies, aes(x = Freq, y = Var1,
main="Number of confirmed reservations based on Number of
Babies")) +
  geom_bar(stat = "identity") +ylab("Number of
Babies")+xlab("Number of Reservations") +theme_get()
**
```

```
**Output**
```





```
> noncancellationbasedonBabies
  Var1  Freq
1    0 28493
2    1   438
3    2     7
>
```

Inference

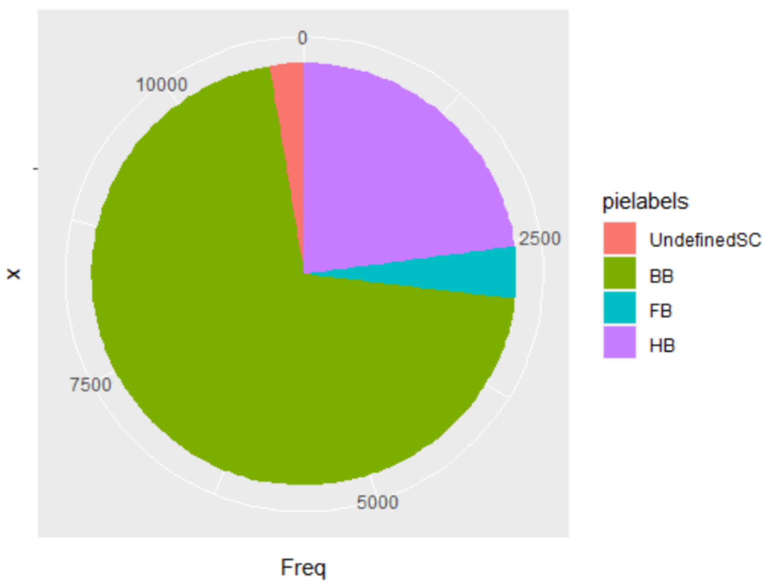
- From the above plot, we can infer that the maximum number of cancellations are done by families with no babies and there seems to be an apparent inverse proportionality in the number of children and the number of cancellations, when observing the non-cancelled hotel reservations subset of data.

Q. How does the FnB facet affect the overall hospitality that is expected?

```
**code
cancellationbasedonMeal<-
data.frame(table(CanceledBookingData$Meal))
```

```
pielabels= cancellationbasedonMeal$Var1
ggplot(cancellationbasedonMeal, aes(x="", y=Freq, fill=
pielabels)) +  geom_bar(stat="identity") + coord_polar("y")

CollateddfFnB<- data.frame(table(hoteldata$Meal))
CollateddfFnB$cancelledpercentage<-
(cancellationbasedonMeal$Freq/CollateddfFnB$Freq)*100
CollateddfFnB$noncancelledpercentage<- 100 -
CollateddfFnB$cancelledpercentage
CollateddfFnB<- rename(CollateddfFnB, Meal_Plans= Var1)
**
```

Output




```
> cancellationbasedonMeal
         Var1 Freq
1 UndefinedSC  289
2          BB 7843
3          FB  443
4          HB 2547
>
```

```
> CollateddfFnB
   Meal_Plans  Freq cancelledpercentage noncancelledpercentage
1 UndefinedSC  1255            23.02789               76.97211
2          BB 30005            26.13898               73.86102
3          FB   754            58.75332               41.24668
4          HB  8046            31.65548               68.34452
>
```

Inference

When we delve into the cancelled hotel reservations subset of data, and group bookings on the basis of their meal preferences, an implication that is observed is that the greatest number of cancellations are observed from guests who had opted for the Bed and Breakfast hospitality meal package.

Q. How does the proportion of bookings vary across countries?

As an aside, although we did not use 'Country' variable for any analysis, we did a quick frequency distribution table by countries.

```
**
arrange(hoteldata %>% tabyl(Country) %>% adorn_totals("row") %>%
adorn_pct_formatting(), desc(n))
**
```

****Output****

Country	n	percent
Total	40060	100.0%
PRT	17630	44.0%
GBR	6814	17.0%
ESP	3957	9.9%
IRL	2166	5.4%
FRA	1611	4.0%
DEU	1203	3.0%
CN	710	1.8%
NLD	514	1.3%
USA	479	1.2%

Interpretation

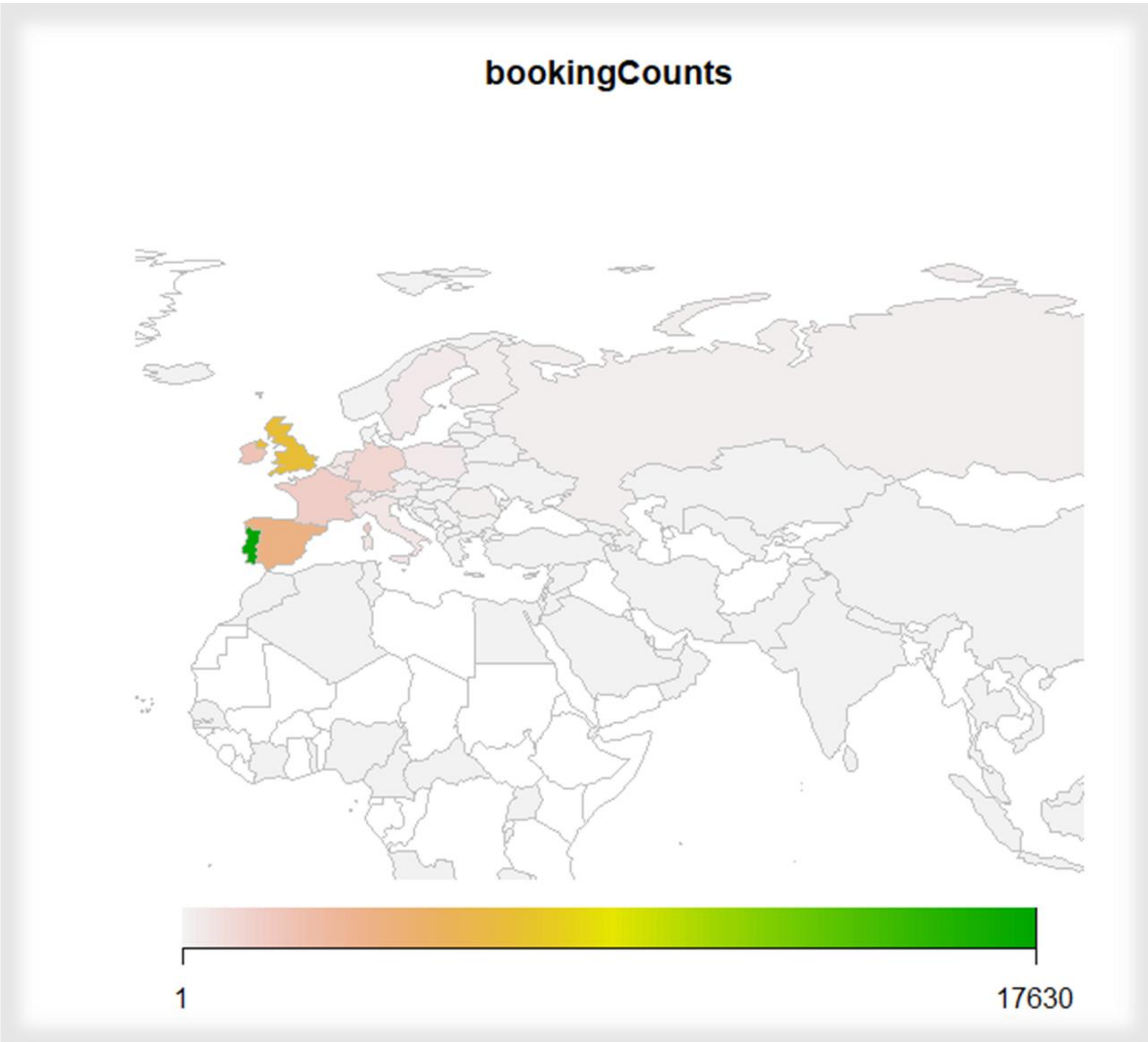
- 44% of bookings were done from Portugal. This implies that the hotel is popular among the Portuguese guests, or it could very well mean that the hotel is in Portugal.
- This hotel gets its most customers from Europe, with CN (China) and USA being the only non-European nations in the top nine countries.

We showed this information on a map too.

```
**
library(rworldmap)
CountryCountData <- hoteldata %>% mutate(count=1) %>% group_by(Country)
%>% summarise(bookingCounts = sum(count))
CountryCountData$group <- ntile(CountryCountData$bookingCounts, 50)
CountryCountData$rank <-rank(CountryCountData$bookingCounts)
mean(CountryCountData$rank) #317.93, the average
mapdata <- joinCountryData2Map(CountryCountData,"ISO3","Country")
```

```
mapCountryData(mapdata,
               nameColumnToPlot = "bookingCounts" ,
               catMethod = "fixedWidth",
               colourPalette = "terrain",
               mapRegion = "eurasia",
               numCats = 126,
               # missingCountryCol = 'Green'
)
**
```

Output



Furthermore, we also created a wordcloud to show this distribution of countries among the bookings:

```
**
countries <- as.character(hoteldata$Country)

#change all CNs to CHNs because they both refer to China:
countries <- replace(countries, which(countries=="CN"), "CHN")

#Change all NULLs to 'Unknown':
countries <- replace(countries, which(countries=="NULL"),"Unknown")

#Replace country codes by country names:
countries <- countrycode(countries, "iso3c", "country.name")

#Remove all spaces (required for the wordcloud to recognize country
names correctly):
countries <- str_replace_all(countries, " ", "")

#Create corpus
docs <- Corpus(VectorSource(countries))

docs <- docs %>% tm_map(stripWhitespace)

dtm <- TermDocumentMatrix(docs) #creates TDM
matrix <- as.matrix(dtm) #changes into a matrix
words<- sort(rowSums(matrix),decreasing=TRUE) #sorts countries
according to count
df<-data.frame(word=names(words), freq=words) #creates a df with
country names and corresponding frequencies

#Draw the wordcloud
dev.new()
wordcloud(words=df$word, freq=df$freq, min.freq=1,
          max.words=50000, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8,"Dark2"))
```

Output

Interpretation

- It confirms that most bookings are done from Portugal, with Great Britain and Spain coming in distant second and third positions respectively.

Section IV: Findings and Recommendations

Based on our EDA and interpretation of ML models, we have found that the variables listed below play significant roles in determining whether a booking is cancelled or not. And this is how they determine it:

1. **Lead Time:** Generally, the longer the lead time, the more likely the booking will get cancelled. 50% of all non-cancelled bookings have quite a short lead time (just five weeks or less). This is also true if we segregate the guests by repeat guests versus first time guests. 50% of non-cancelled bookings by repeat guests have a lead time of 2 or less days.
    a. **Recommendation:** Over any period of time, if you have too many bookings with long lead times, consider overbooking.
2. **Repeat Guests vs First-time Guests:** Repeat guests are far less likely to cancel a booking than the first-time guests. On average, 28.8% of first-time guests cancel whereas only 6.2% of repeat guests cancel their bookings.
    a. **Recommendation**: If you have a very high proportion of first-time guests booking, consider overbooking; or consider prioritizing repeat guests for the remaining bookings.
3. **Deposit Type:** Normally, one would think that bookings with non-refundable deposits will get cancelled less often. However, in this hotel, proportion of cancelled bookings is significantly high (14.84%) compared to that for non-cancelled bookings (0.24%).
    a. **Recommendation**: Revise your deposit policy. Consider raising the amount of your non-refundable deposit because at the moment, the current deposit amount is not a strong deterrent to prevent booking cancellations.
4. **No of booking changes:** 80.5% of bookings have no changes made to it. However, non-cancelled bookings have higher modifications (23%) compared to cancelled-bookings (10.4%). So, if a booking is modified, it has lower chances of getting cancelled in the end.
    a. **Recommendation**: This tells us that booking changes can be seen as an effort to make things work (from hotel's as well as guests' sides). So, incentivize customizations during and after booking (a way to accomplish this could be to allow users to make many choices while booking, and/or allowing them to make no-cost modifications when feasible).
5. **Required car parking space:** All bookings that had at least one required car parking space were not cancelled. All cancelled bookings have exactly zero requests for car parking space.
    a. **Recommendation**: If you could, ask the guests whether they will be needing a parking space. Historically, those who drive themselves to your hotel are less likely to cancel their bookings.

6. **Special requests:** In line with number of booking changes and required car parking space, increase in special requests corresponds to decrease in booking cancellations. Again, this can be seen as an increased commitment to a booking from the guest's side, and hence it leads to less probability of cancellation.
7. **Market segments:** 'Groups' has 42% cancellation rate. 'Online TA' has 35% cancellation rate. Each of the remaining four segments have about 15% cancellation rates only.
    a. **Recommendation**: Bookings in segments 'groups' and 'online TA' are more likely to be cancelled compared to bookings in other segments. So, in the short run, the hotel should minimize its over-dependence on these segments. In the longer run, it needs to unravel and fix the causes that are driving up these segments' cancellation rates.

Appendix-1: Metadata

Variable	Description
IsCanceled	Categorical Value indicating if the booking was canceled (1) or not (0)
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 week nights (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 “TA” means “Travel Agents” and “TO” means “Tour Operators”
IsRepeatedGuest	Categorical, Value indicating if the booking name was from a repeated guest (1) or not (0)
PreviousCancellations	Integer, Number of previous bookings that were cancelled by the customer prior to the current booking
PreviousBookingsNotCancelled	Integer, Number of previous bookings not cancelled by the customer prior to the current booking
ReservedRoomType	Categorical, Code of room type reserved. Code is presented instead of designation for anonymity reasons
AssignedRoomType	Categorical, Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type

	due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
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 Refund – a deposit was made in the value of the total stay cost. Refundable – a deposit was made with a value under the total cost of stay.
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
RequiredCarParkingSpaces	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)

Appendix-2: Full list of business questions

1. How does lead time vary between cancelled and non-cancelled bookings?
2. How does lead time vary across cancelled and non-cancelled bookings for repeat guests?
3. How do cancellations compare between first time guests and repeat guests?
4. What does BookingChanges tell us about probable cancellations?
5. How does the Amount of Required Car Parking Spaces Affect Cancellations?
6. How does the Deposit Type Affect Cancellations?
7. Do cancellation patterns vary significantly across market segments?
8. Do special requests determine booking cancellations?
9. How does the Customer Type Affect Cancellations?
10. Does number of children affect booking cancellations?
11. How does the number of babies affect booking cancellations?
12. How does the FnB facet affect the overall hospitality that is expected?
13. How does the proportion of bookings vary across countries?