What is EDA?

EDA stands for "Exploratory Data Analysis," which is a process of analyzing and summarizing data sets to gain insights and make informed decisions. The primary goal of EDA is to discover patterns, relationships, and trends in data, as well as to identify any anomalies or outliers that may affect the accuracy of the analysis.

EDA

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies,and the number of available parking spaces, among other things. Using the booking information about these hotels, the challenge is to build a model to predict booking cancellation based on the following fields:

Data Description:

* **hotel** (str) - Hotel (H1 = Resort Hotel or H2 = City Hotel)
* **lead\_time** (int) - Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
* **arrival\_date\_year** (int) - Year of arrival date
* **arrival\_date\_month** (str) - Month of arrival date
* **arrival\_date\_week\_number** (int) - Week number of year for arrival date
* **arrival\_date\_day\_of\_month** (int) - Day of arrival date
* **stays\_in\_weekend\_nights** (int) - Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
* **stays\_in\_week\_nights** (int) - Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
* **adults** (int) - Number of adults
* **children** (float) - Number of children
* **babies** (int) - Number of babies
* **meal** (str) - 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** (str) - Country of origin. Categories are represented in the ISO 3155–3:2013 format
* **market\_segment** (str) - Market segment designation (TA = Travel Agents, TO = Tour Operators)
* **distribution\_channel** (str) - Booking distribution channel (TA = Travel Agents, TO = Tour Operators)
* **is\_repeated\_guest** (int) - Value indicating if the booking name was from a repeated guest (1) or not (0)
* **previous\_cancellations** (int) - Number of previous bookings that were cancelled by the customer prior to the current booking
* **previous\_bookings\_not\_canceled** (int) - Number of previous bookings not cancelled by the customer prior to the current booking
* **reserved\_room\_type** (str) - Code of room type reserved. Code is presented instead of designation for anonymity reasons
* **assigned\_room\_type** (str) - 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
* **booking\_changes** (int) - 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
* **deposit\_type** (str) - 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)
* **agent** (float) - ID of the travel agency that made the booking
* **company** (float) - ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
* **days\_in\_waiting\_list** (int) - Number of days the booking was in the waiting list before it was confirmed to the customer
* **customer\_type** (str) - 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
* **adr** (float) - Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
* **required\_car\_parking\_spaces** (int) - Number of car parking spaces required by the customer
* **total\_of\_special\_requests** (int) - Number of special requests made by the customer (e.g. twin bed or high floor)
* **reservation\_status** (str) - Reservation last status, assuming one of three categories: (Canceled = booking was canceled by the customer, Check-Out = customer has checked in but already departed, No-Show = customer did not check-in and did inform the hotel of the reason why)
* **reservation\_status\_date** (str) - Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel

**Procedure:**

* Data collection and cleaning the dataset
* Finding variables and understanding
* Finding correlation
* Visualizing and analysing variables
* Model implementation

Step1: Data collection

(R language)

#Preprocessing part

data <- read.csv("C:/Users/rkt\_1/Desktop/EDA final/hotel\_bookings.csv")

#checking first few rows of the data

head(data)

# Check for missing values

colSums(is.na(data))

value=0

# Replace missing values with the value

data$children[is.na(data$children)] <- value

#checking the type of data

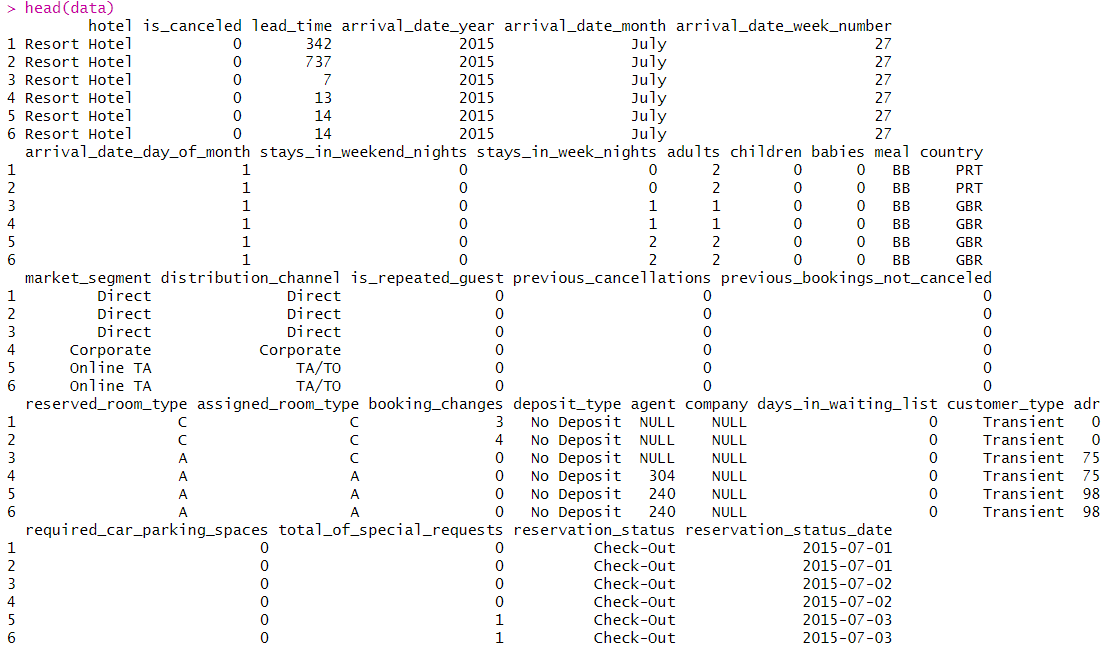
str(data)

#Examining the number of rows and column

num\_rows <- nrow(data)

num\_cols <- ncol(data)

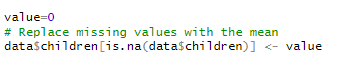
Checking out the first few rows of the dataset to check what value they contain



Checking out if there is any column that has missing values



Only the column children has 4 missing value

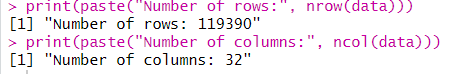


Since when the information about children is not given so we assume they do not have children so replacing the missing value with 0

Checking the structure of the dataset



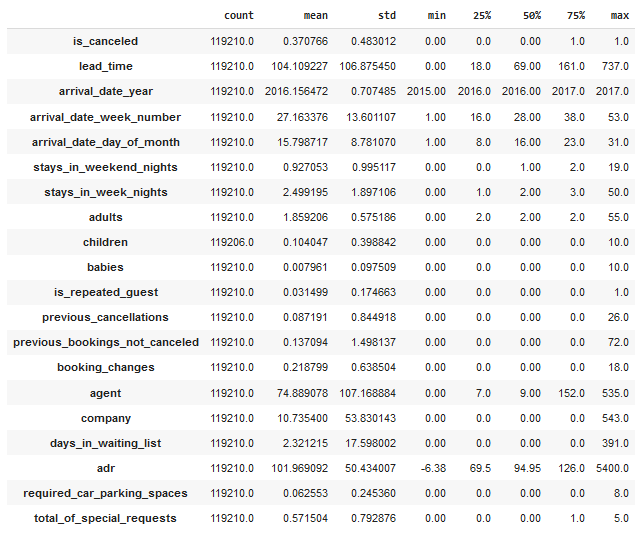
Calculating the total number of rows and columns that we would be operating on



Step2: Finding variables and understanding

Exploring the continuous variables

**Summary of our dataset (continuous features)**



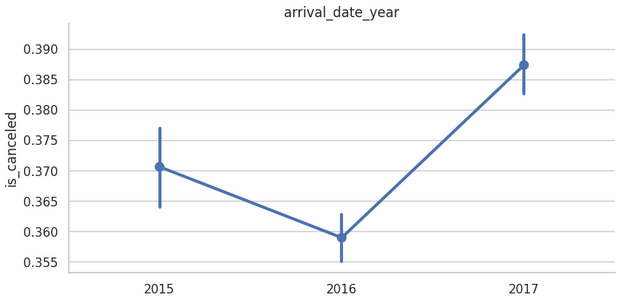
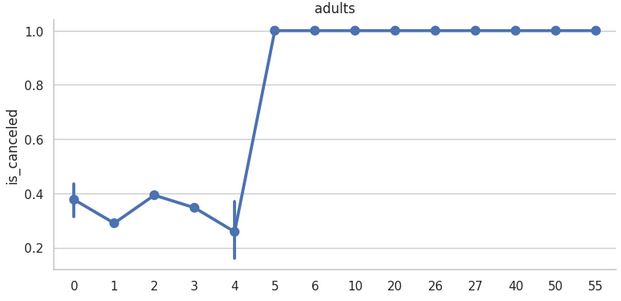
**#Analysis**

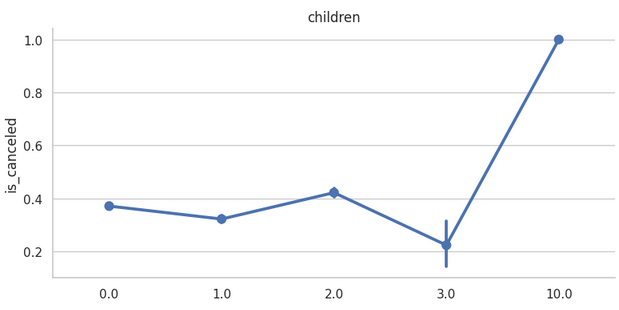
Checking all the continuous feature

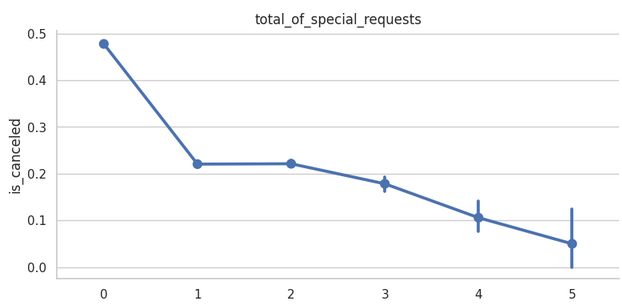
Most of the cancellations happened in 2017

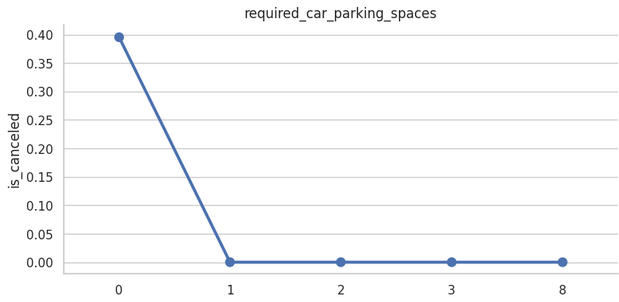
When there are more number of adults then most of the cancellations happens similar is the case with number of children

When more number of special requests are made then less is the chance of cancellation

****

****

****

****

****

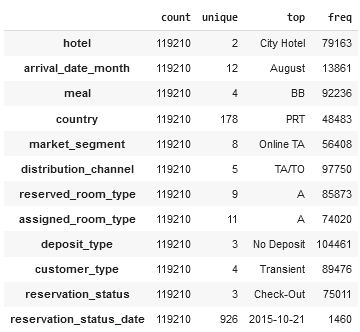
Repeated guest are not cancelling the booking this means the guests were happy with the previous services

New guests are likely to cancel their bookings

Again if the guests made a request for parking space then they are not cancelling the bookings

Exploring the categorical variables

**Summary of our dataset (categorical features)**

****

**#Analysis**

Most of the cancellations are during winters

**Guests who selects** FB (full board i.e, all the meals breakfast, lunch ,dinner included) cancels the most

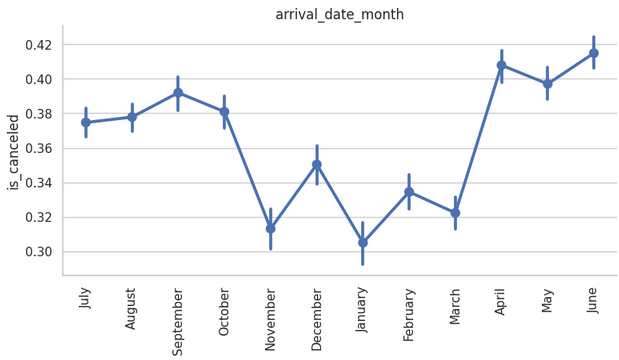
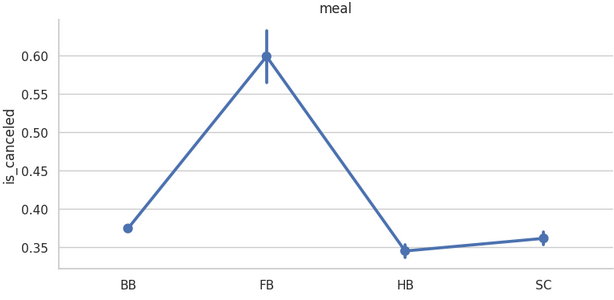
When the market segment is not know they have the maximum number of cancellations

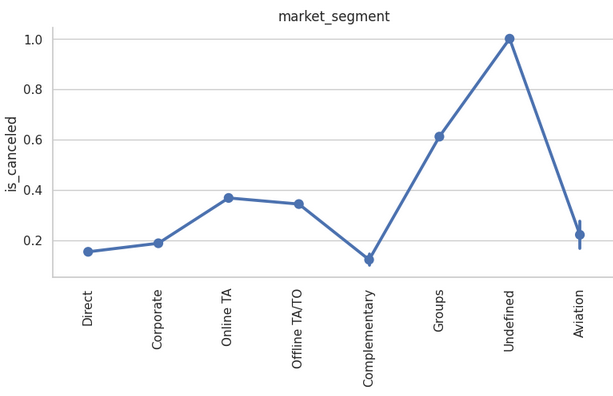
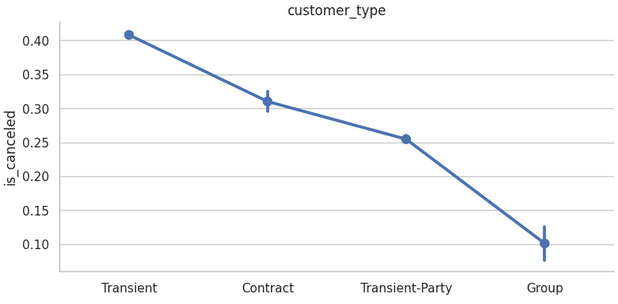
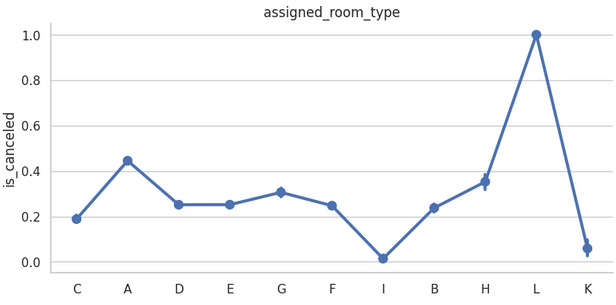
when the booking is not part of a group or contract, and is not associated to other transient booking has most numbers of cancellations

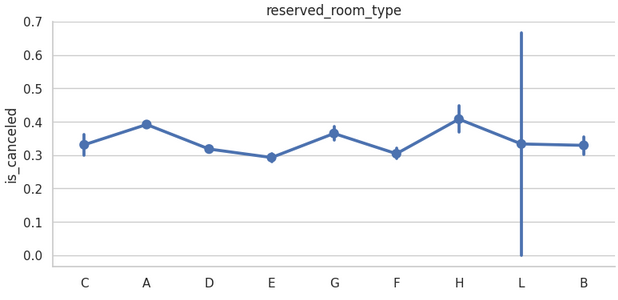
Almost there are cancellations from all the room types

When the assigned room type is “L” then it has the most number of cancellations

We can observe that L room type is not preferred by guests

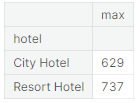
****

****

****

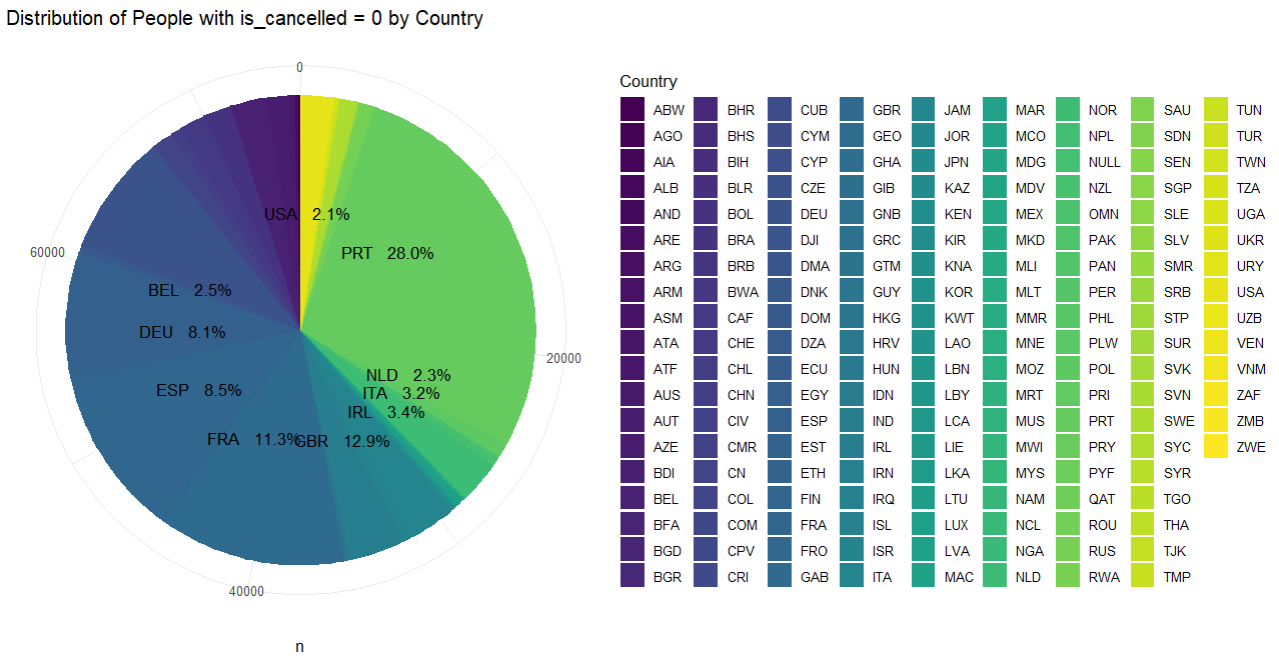
## What was the maximum lead time in each hotels?

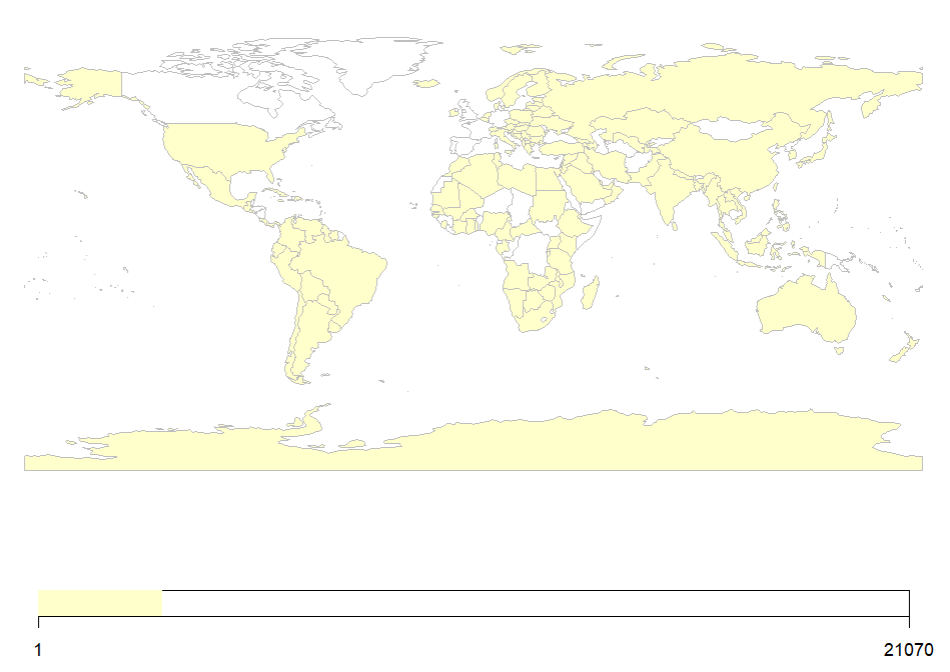
## Resort hotel has maximum lead time

****

**Maximum visitors belong to which country?**

People from all over the world are staying in these two hotels but most of the people who actually visits the two hotel (i.e when the cancellation status is 0) are from Portugal (28.0%) followed up by United Kingdoms (12.9%) France (11.3%) and other European countries.



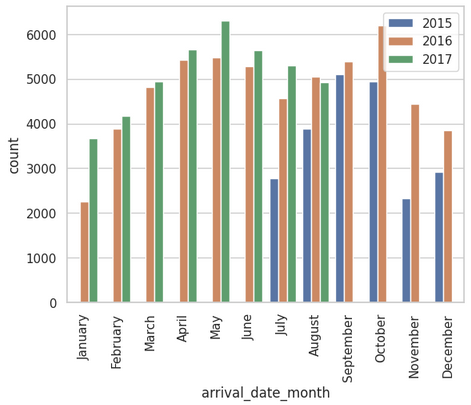


**How many guest arrived each month per year?**

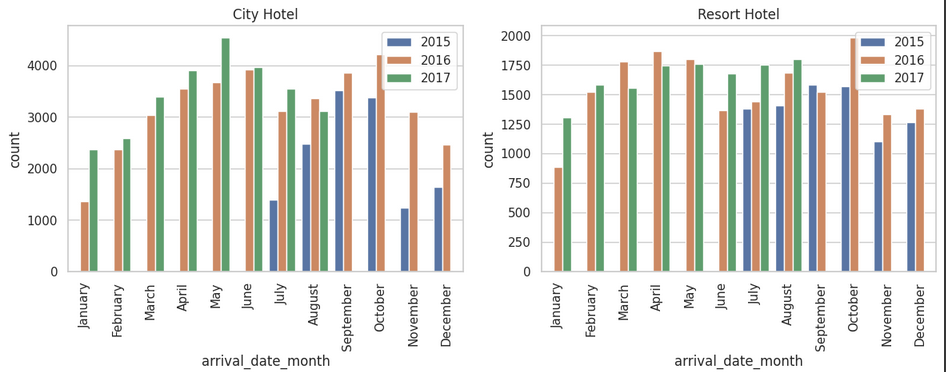
Year: 2015, Month: September

Year: 2016, Month: October

Year: 2017, Month: May



## How many guests arrived each month per year in each hotel?

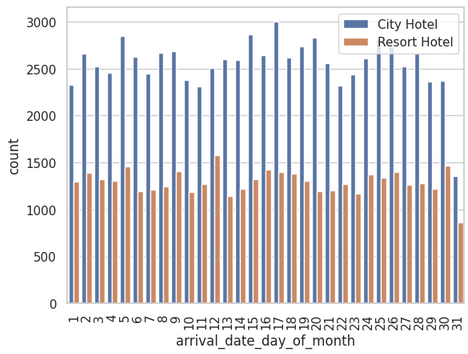


**Peak month in City Hotel  
- 2015: September  
- 2016: October  
- 2017: May**

**Peak month in Resort Hotel  
- 2015: September  
- 2016: October  
- 2017: August**

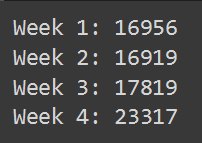
## How many guests arrived by date of month?

## More or less the arrival dates has no effect, as guests are arriving at all the dates in almost same number



## Guest arrival trend

Maximum number of guest visits the hotel in the 4th week of the month (month end)



**How much do the guests pay per night?**

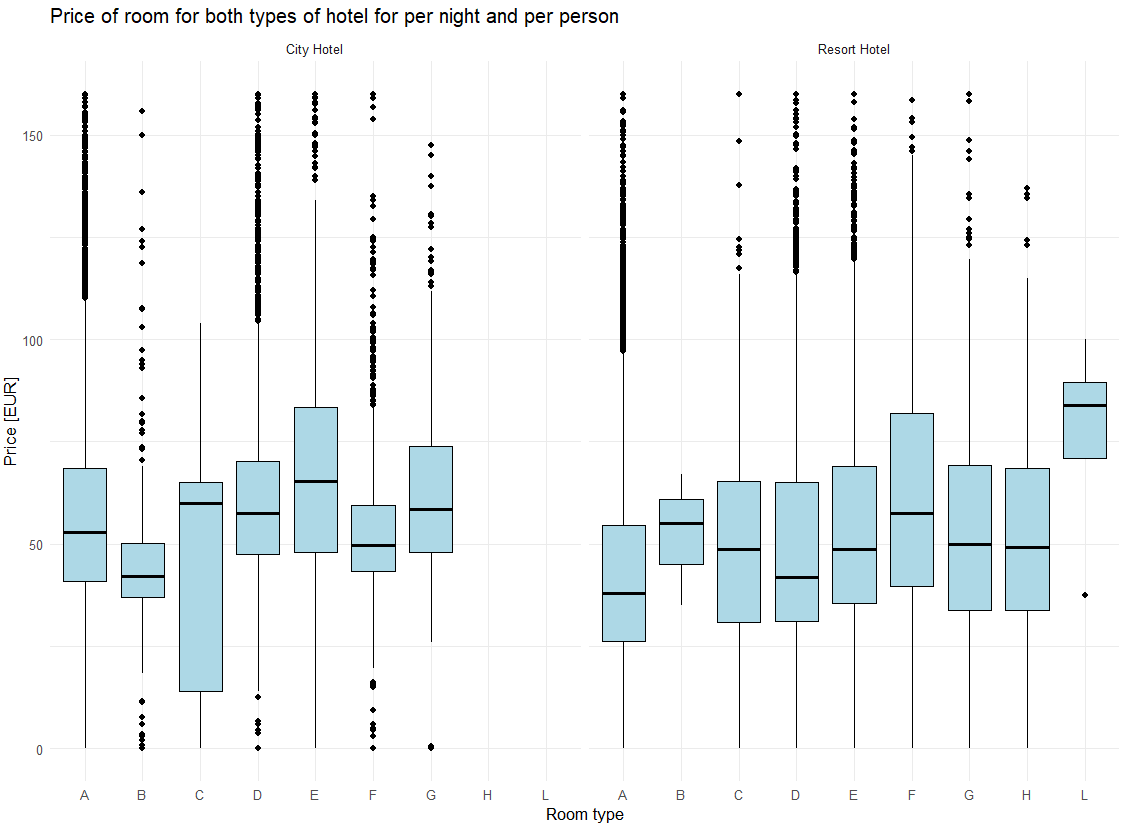
Amount to be paid can depend on various factors such as type of hotel, type of room selected or on the meal plan chosen and the time of visit as well.

(Prices are in EUR)\*

Average Price for Resort Hotel: € 47.57563 per night per person

Average Price for City Hotel: €54.2152 per night per person

The price of city hotel is more than resort hotel on an average



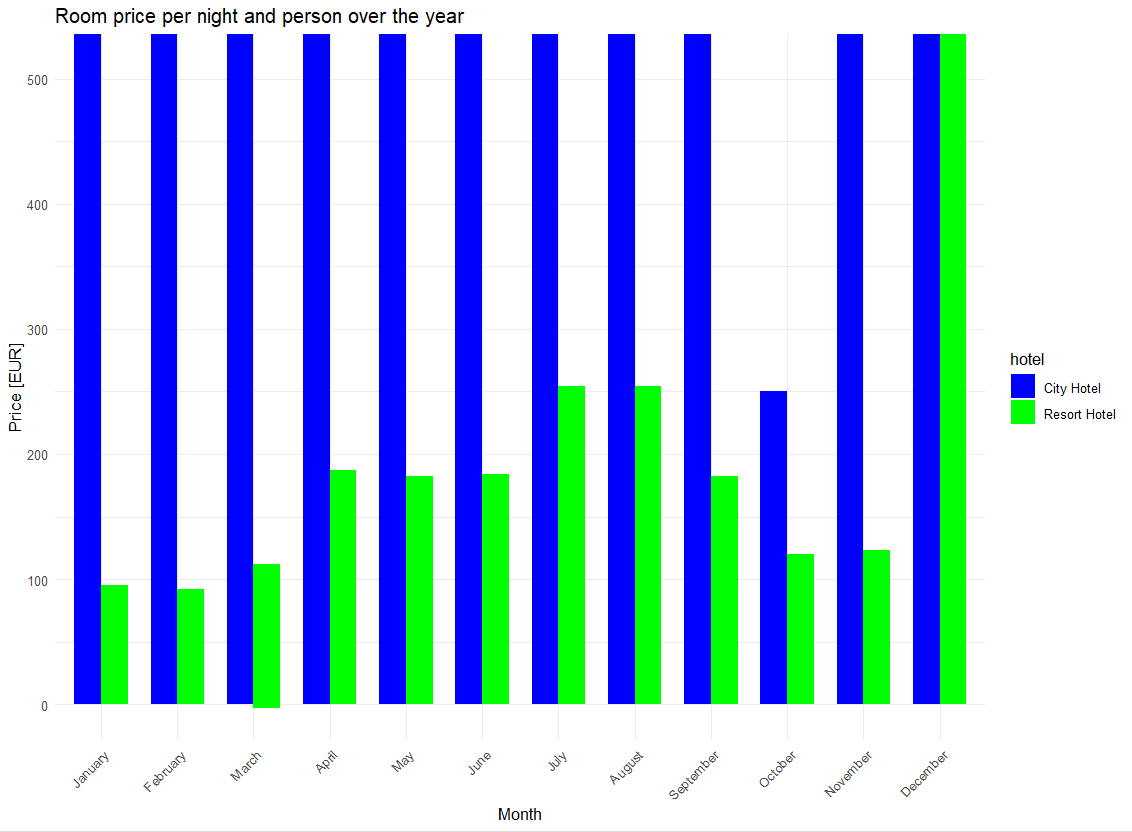
This figure shows the variation of the prices of the same room type for both the hotels. We can observe on an avg the price of rooms for city hotel is more than those of resort hotel

To keep it simple, I'm using the average price per night and person, regardless of the room type and meal.

The price of the city hotel is high all round the year and there is not much variation as well but the price suddenly drops in the month of October

The price of resort hotel rises in the summer season between June-August and is comparatively less expensive as compared to city hotel

At the end of the year during Christmas and New-year the price of both the hotel is high



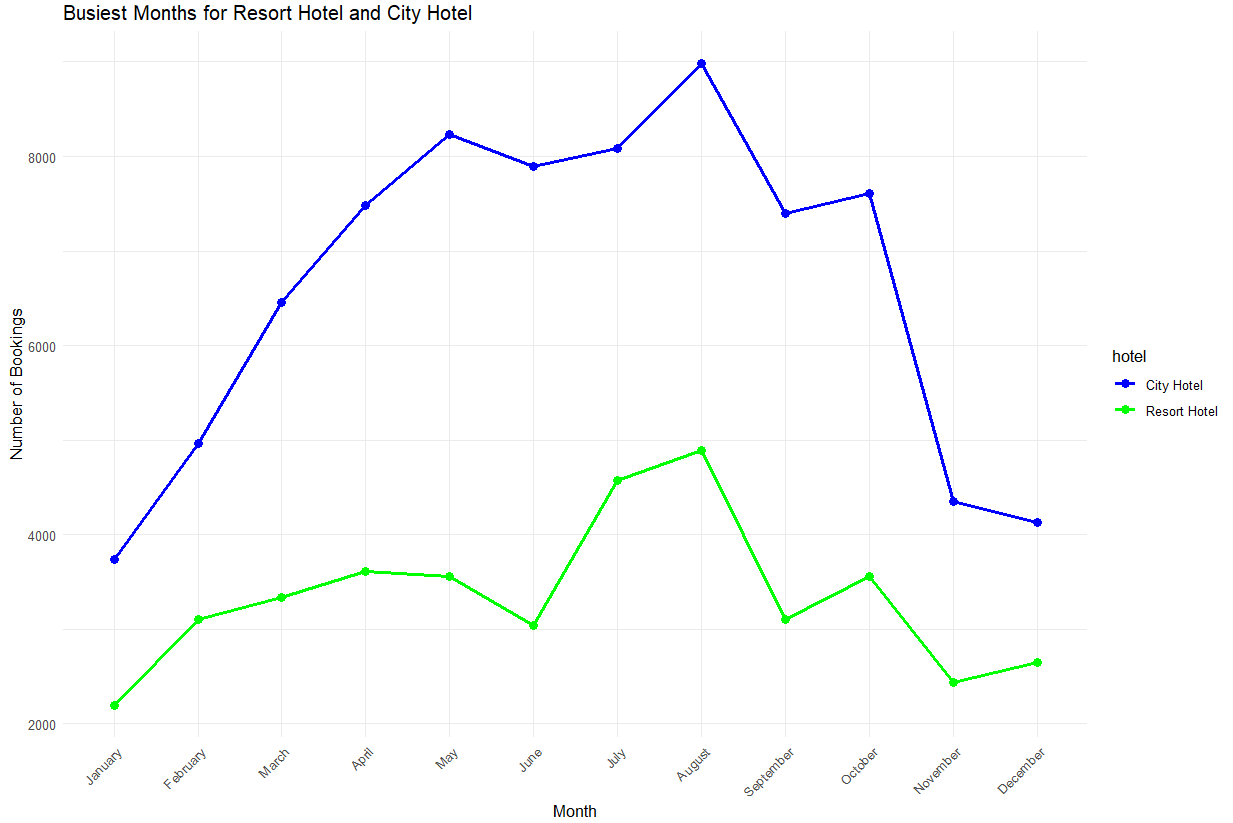
### **Which are the most busy month?**

The city hotel is comparatively busier than resort hotel any time of the month also the prices are high for city hotel

The resort hotel expects more guests during the summer season but overall, the number of booking for resort hotel is low

For both of the hotel August is the busiest month

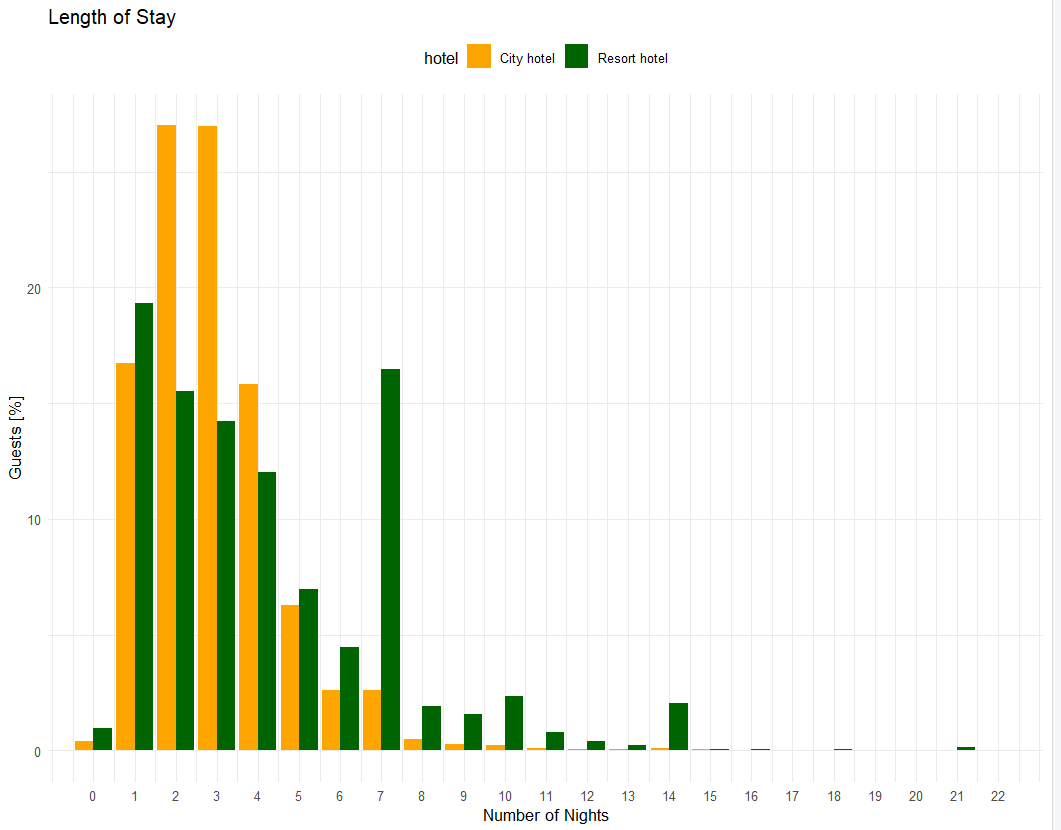
At the time of Christmas and New-year when the price of both the hotels are high, they receive the least number of bookings



### **How long do people stay at the hotels?**

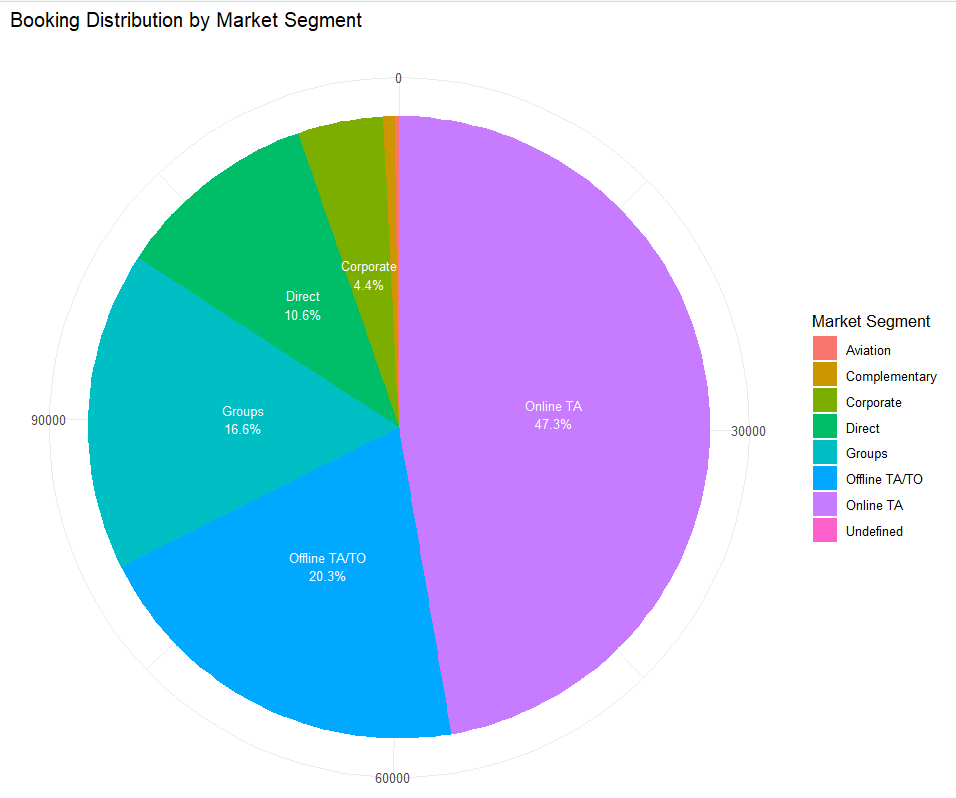
In city hotel most number of guests like to stay for 1-4 nights with a clear preference of 2-3 nights and since the city hotel is more expensive so guests are not staying in for a longer period of time

For resort hotel also 1-4 nights seems an optimal length of stay but there is a sudden preference of 7 nights and since the resort hotel is much less expensive the guests makes it a popular choice for longer stays



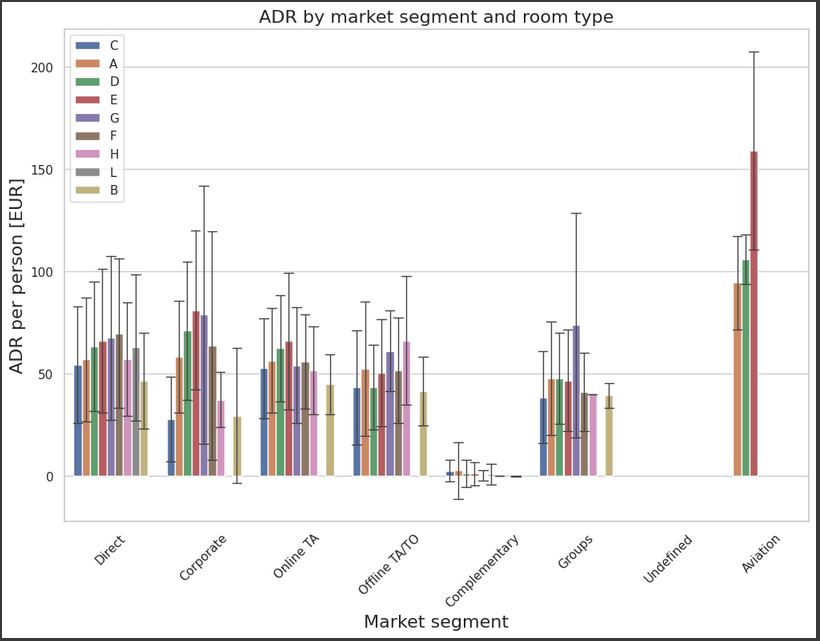
**Booking by market segment?**

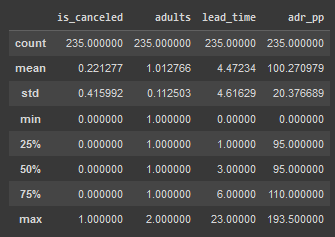
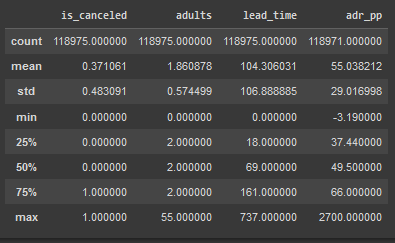
Most of the bookings are by Online TA (travel agent) followed up by offline travel agent or tour operators and the least is from aviation



ADR (Average Daily Rate) is highest for aviation and is least for groups while for rest of them it is more or less same

Groups gets the best price whereas aviation pays twice as much them





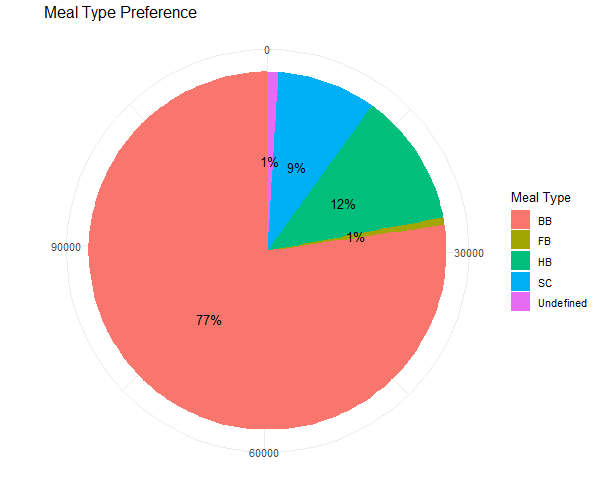
Aviation data Non aviation data

Many reasons could be possible :

1. The lead time for aviation is very low as compared to non aviation one that is aviation people book these hotels in urgency
2. Airplanes need a hotel for their crew to stay
3. Usually one adult in one room so booking more numbers of rooms are required

**Which is the most preferred meal type?**

Bed and Breakfast is most preferred among guests



### **How many bookings were cancelled?**

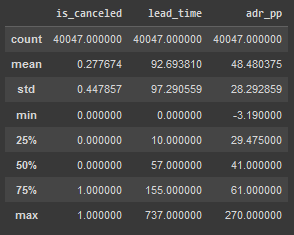
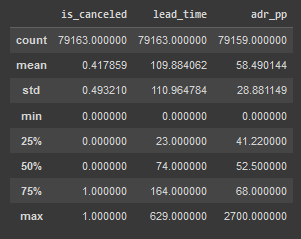
More number of cancellations happens in city hotel

Total bookings cancelled: 44,224 ( 37 %)

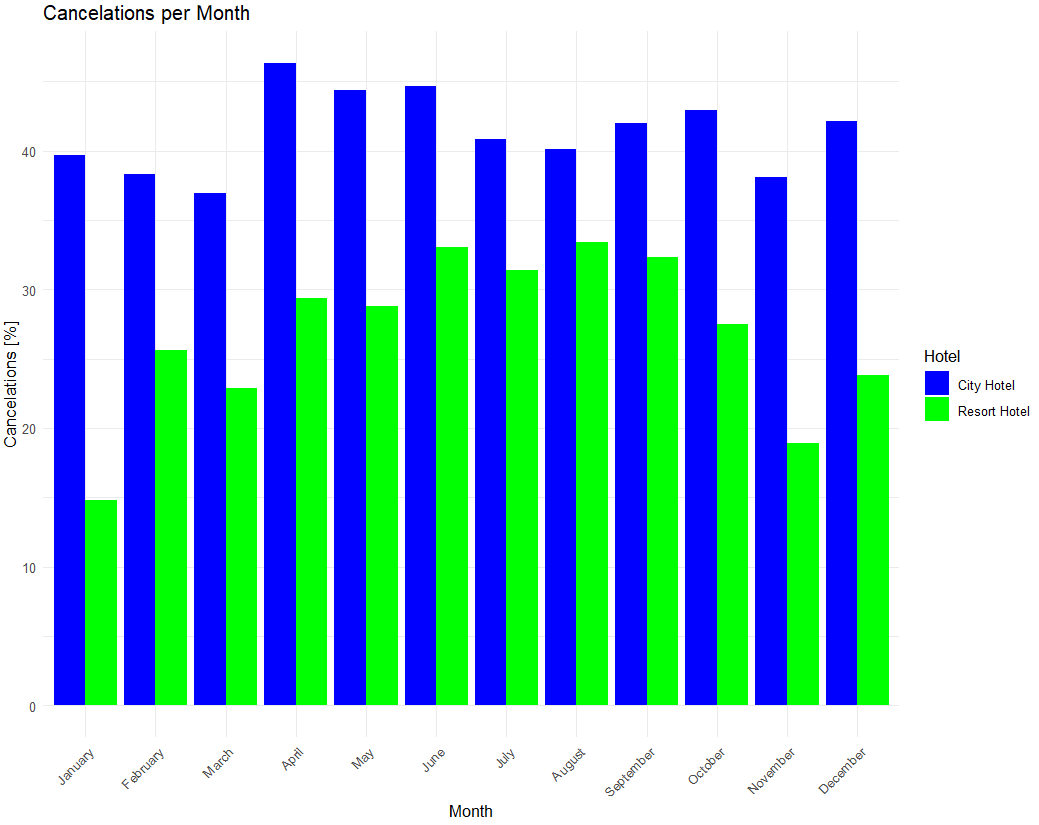
Resort hotel bookings cancelled: 11,122 ( 28 %)

City hotel bookings cancelled: 33,102 ( 42 %)

Since the lead time for city hotel is more and adr per person for city hotel is high also so when people have more time to book the hotel they tend to be more picky in terms of price



Resort Hotel City Hotel



For city hotel the number of cancellations is almost around 40% all-round the year with highest cancellation in the month of April

For resort hotel the maximum number of cancellations is in summer season (when it has the highest price and

high demand) and least number of cancellations is during winter

## How many guests visited hotel more than once?

## 

## In City Hotel 2032 and while in Resort Hotel 1778 guests visited more than once which shows the overall service of the hotels were good

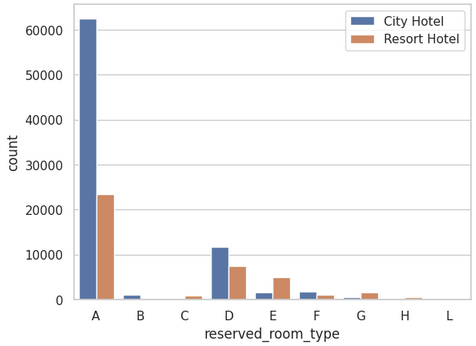
## Which room type was mostly reserved?

## Room type A is most desirable in both the hotels

## Room type B is preferred in city hotel only and not in resort hotel and similar is the case with room type C which is preferred in resort hotel and not in city hotel

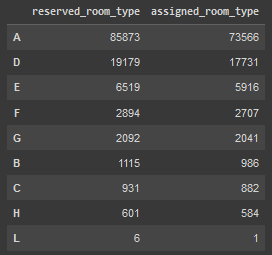
## Room type D is the 2nd most desirable room type in both of the hotels

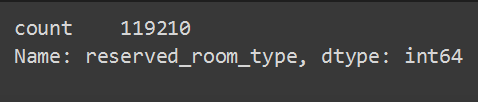
## Room type L is the least desirable one and that is why we have seen above that when the assigned room type is L then maximum number of cancellations occurs

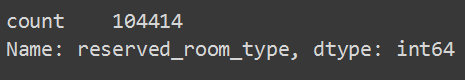


**Was the reserved room type assigned?**

Below is the table indicating not all the guests were assigned their chosen room type

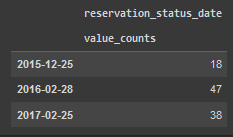






So out of 119210 guests 104414 were allotted with their reserved room type

**Where their any No-Show of guests despite their bookings?**



The year 2016 has maximum number of no-show that is the guests had booked their stay in the hotel but didn’t

show up for the stay

**Effect of lead time?**

Booking that are made just a few days before the trio are rarely cancelled because the guests would be sure to

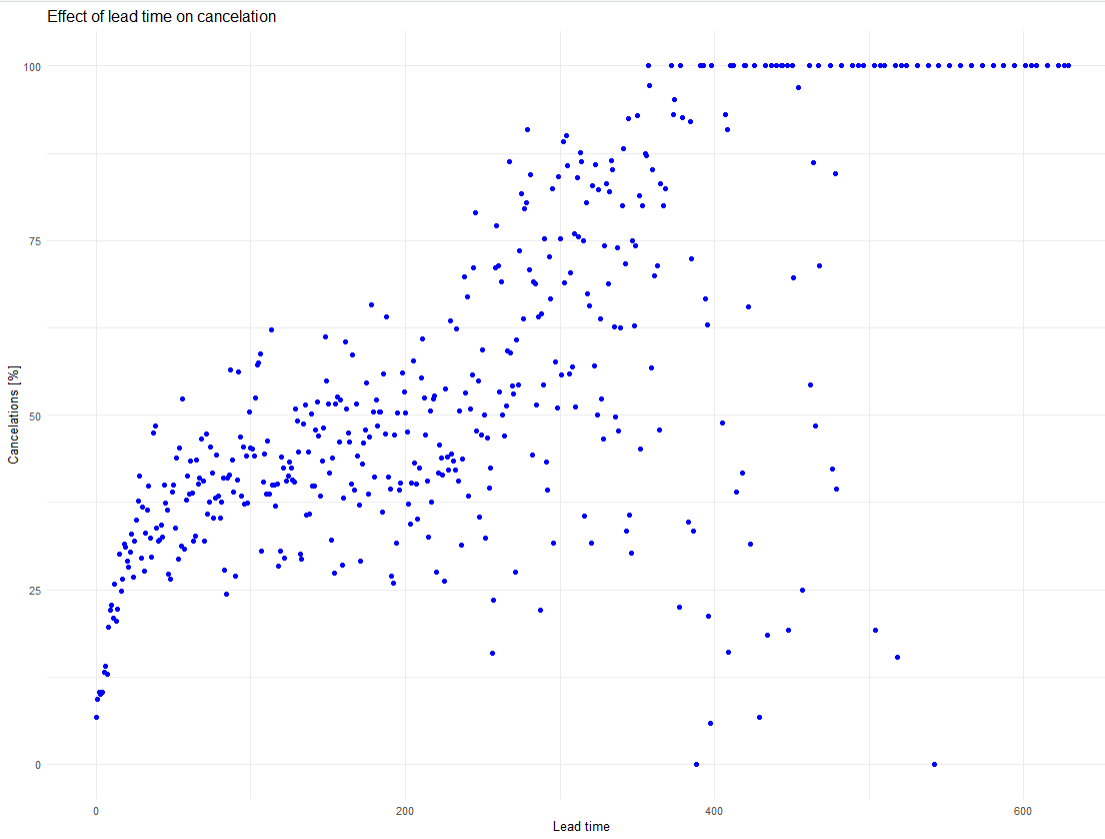
visit the hotel

As the number of days between booking the hotel and actually visiting it increases the number of cancellations

also increases

Whereas the booking that are made a year before the trip are usually cancelled because there is uncertainty

about the trip, of course guests can’t plan their entire year maybe something urgent came up

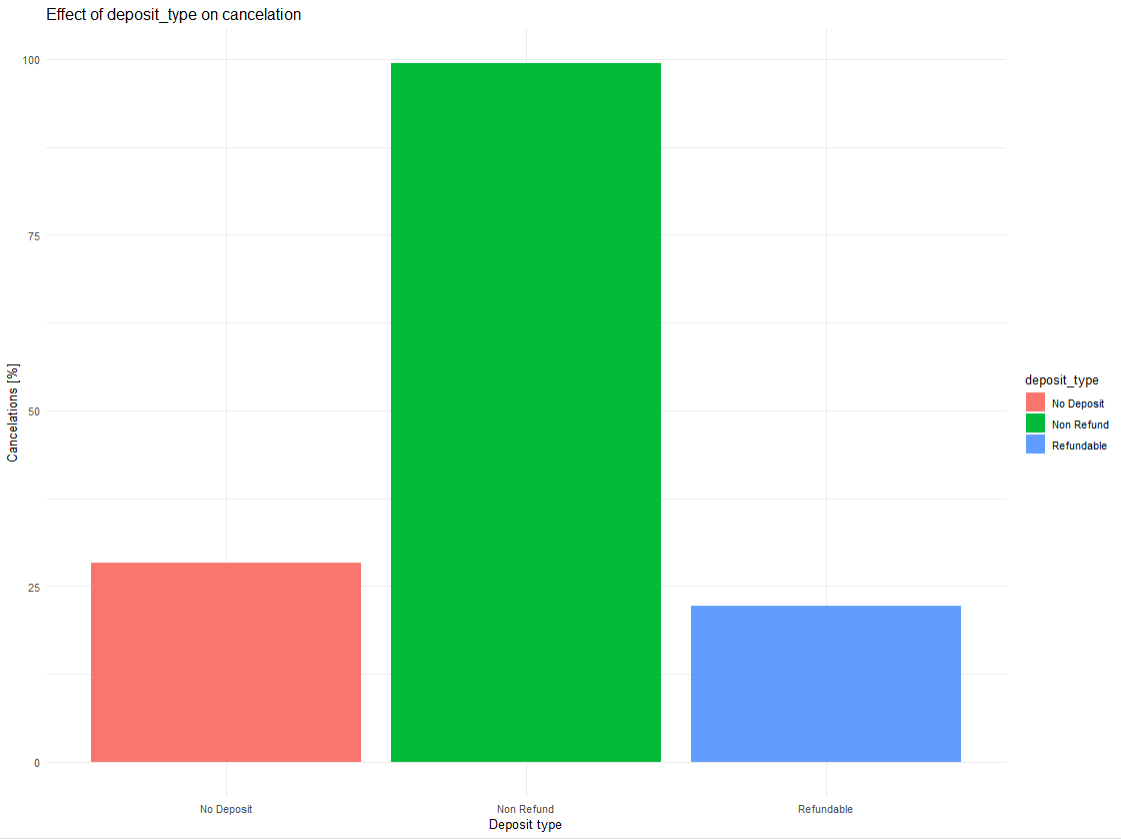


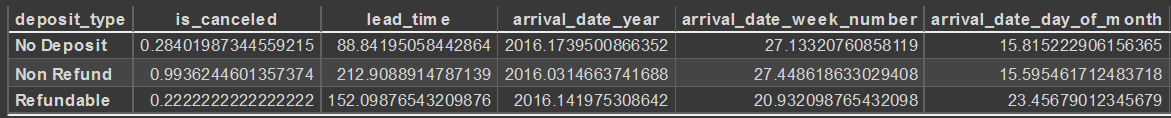
**Effect of deposit type?**

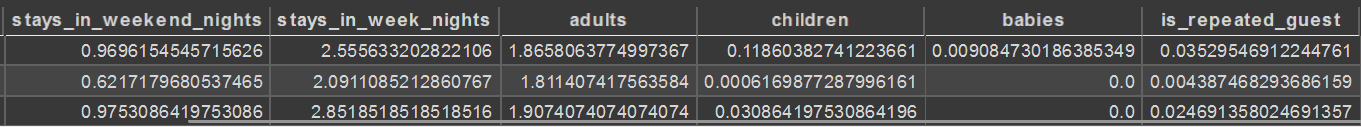
The people who paid the entire amount upfront at the time of booking are the ones who are cancelling the

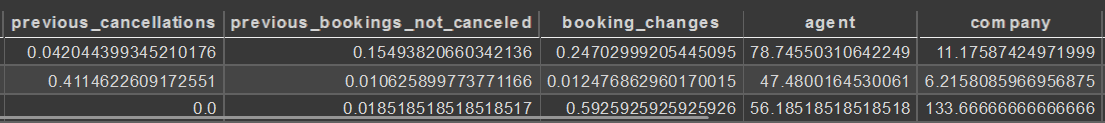
bookings

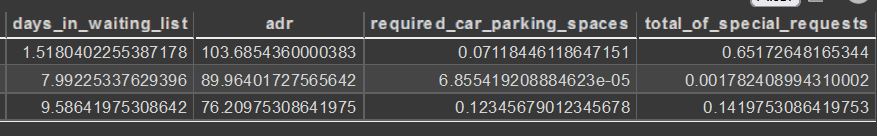
This means that people pay the amount and then cancels the visit to the hotel











Comparing the mean values for Non refund to No Deposit shows the following:

* Non Refund deposits are characterized by > 2x longer lead\_time
* is\_repeated\_guest is ~ 1/10th
* previous\_cancellations is 10x higher
* previous\_bookings\_not\_canceled is 1/15th
* required\_car\_parking\_spaces is almost zero
* special requests are very rare

Based on these findings it seems that especially people who have not previously visited one of the hotels book, pay and cancel repeatedly... this is strange!

**Effect of ADR on cancelation?**



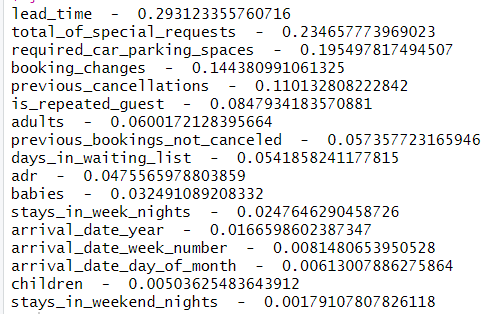
Step3: Finding Correlation

**Which numerical features are most important?**

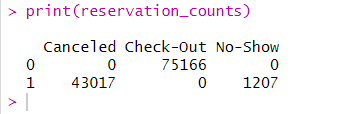
This depicts the correlation of all the column with is\_canceled (our target variable)

Lead time, Number of special requests, total parking space, booking charges, previous cancellations are most important column to determine whether the booking will be cancelled or not

Booking charges is variable it can change over time

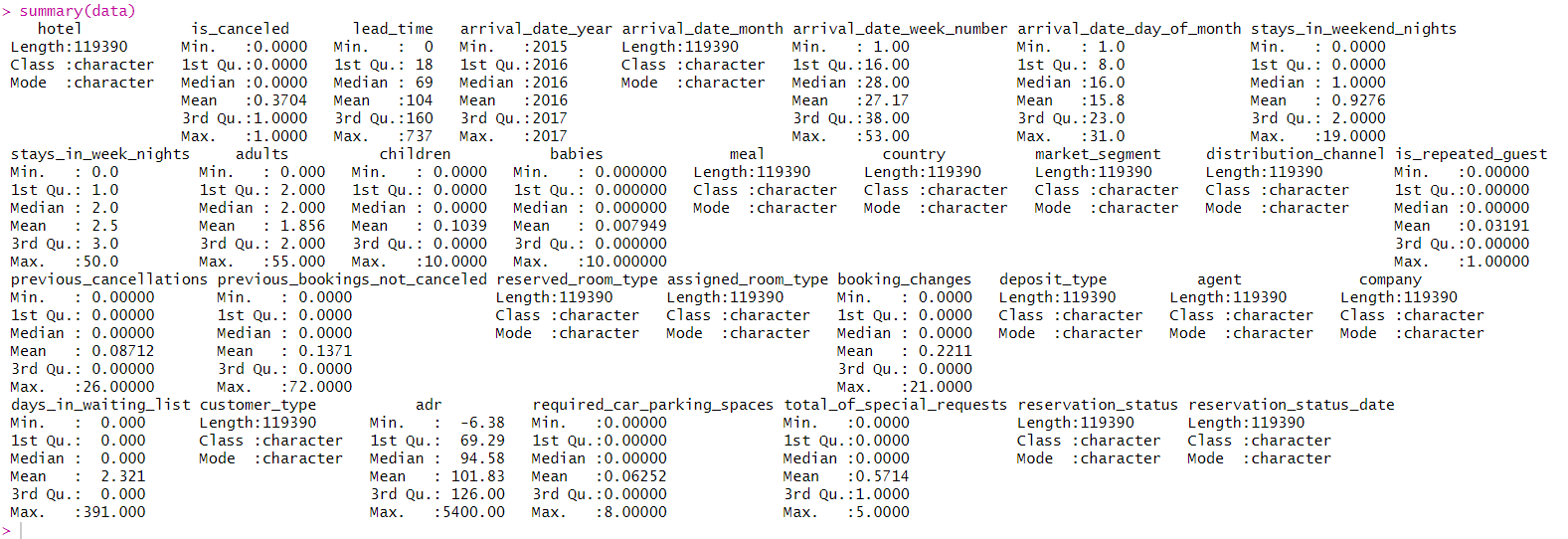


On the basis of is\_canceled column checking the reservation status

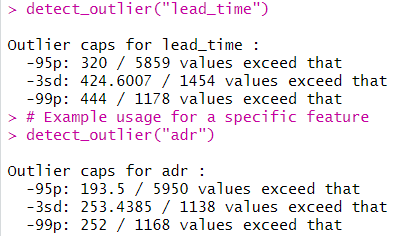


Model implementation

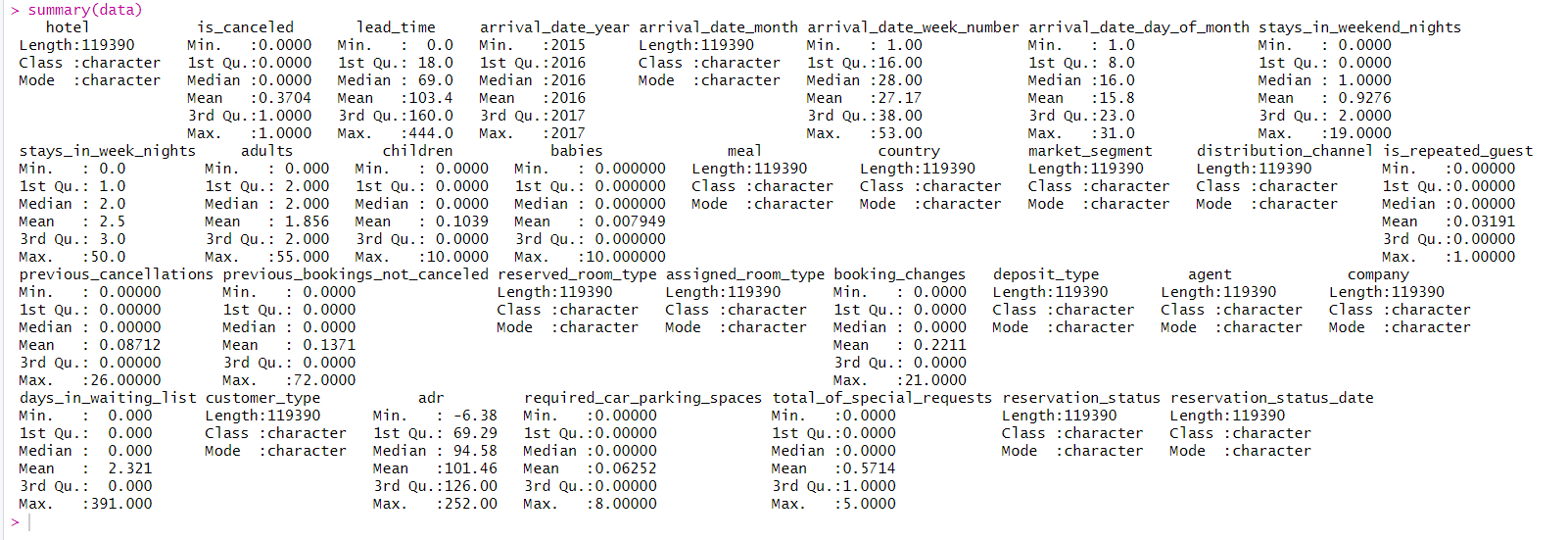
Checking the summary of the entire table



Detecting the outlier in the dataset because they can influence the model



After removing the outlier again checking the summary of the dataset

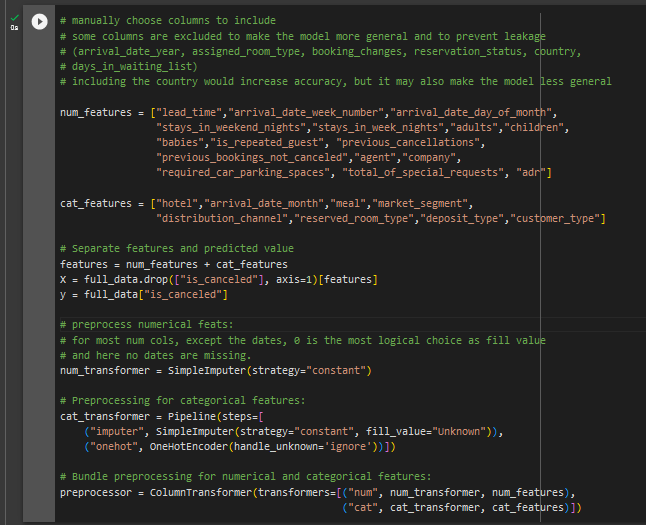


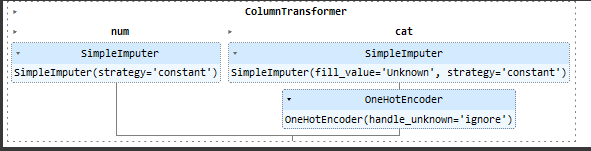
Here we are choosing the numerical and categorical variables that would be help in building the model

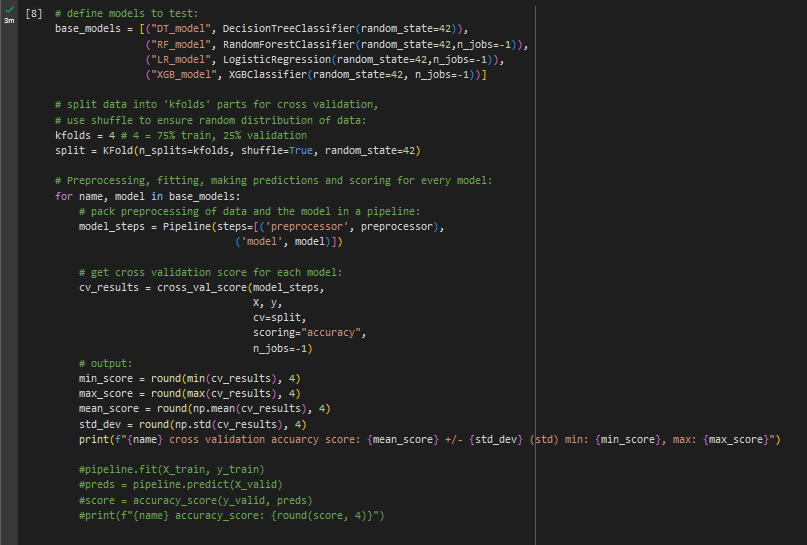
Then we are separating the features column (this will include all the column ) and predict column (which will be is\_canceled)

Then we check if there are any missing values , and if yes then fill it with 0

For the categorical features we are doing one-hot encoding







Results

DT\_model cross validation accuarcy score: 0.8246 +/- 0.0016 (std) min: 0.8221, max: 0.8263

RF\_model cross validation accuarcy score: 0.8664 +/- 0.0012 (std) min: 0.8646, max: 0.8676

LR\_model cross validation accuarcy score: 0.7937 +/- 0.0011 (std) min: 0.7921, max: 0.7951

XGB\_model cross validation accuarcy score: 0.8473 +/- 0.0011 (std) min: 0.8456, max: 0.8487

Random forest is the best model here