# Team #28 Members

1- Yasmine Mahmoud Ahmed.

2- Manar Alaa Mahmoud.

3- Menna Ayman ElSayed.

4- Nada Roshdy Abd El-Mohsen.

5- Ahmed Hazem Raafat.

## A)The project description:

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.

## B) The dataset and variables description:

- □ Dataset : Hotel Booking Demand
  - □ Variables :
  - □ id: unique identifier of each booking
- □ no_of_adults: Number of adults
- □ no_of_children: Number of Children
- □ no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- □ no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- □ type_of_meal_plan: Type of meal plan booked by the customer:
- □ required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
- □ room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.
- □ lead_time: Number of days between the date of booking and the arrival date

- □ arrival_year: Year of arrival date
- □ arrival_month: Month of arrival date
- □ arrival_date: Date of the month
- □ market_segment_type: Market segment designation.
- □ repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- □ no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- □ no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- □ avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- □ no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)

## C) The problem definition and project objectives:

- **Problem Definition:**

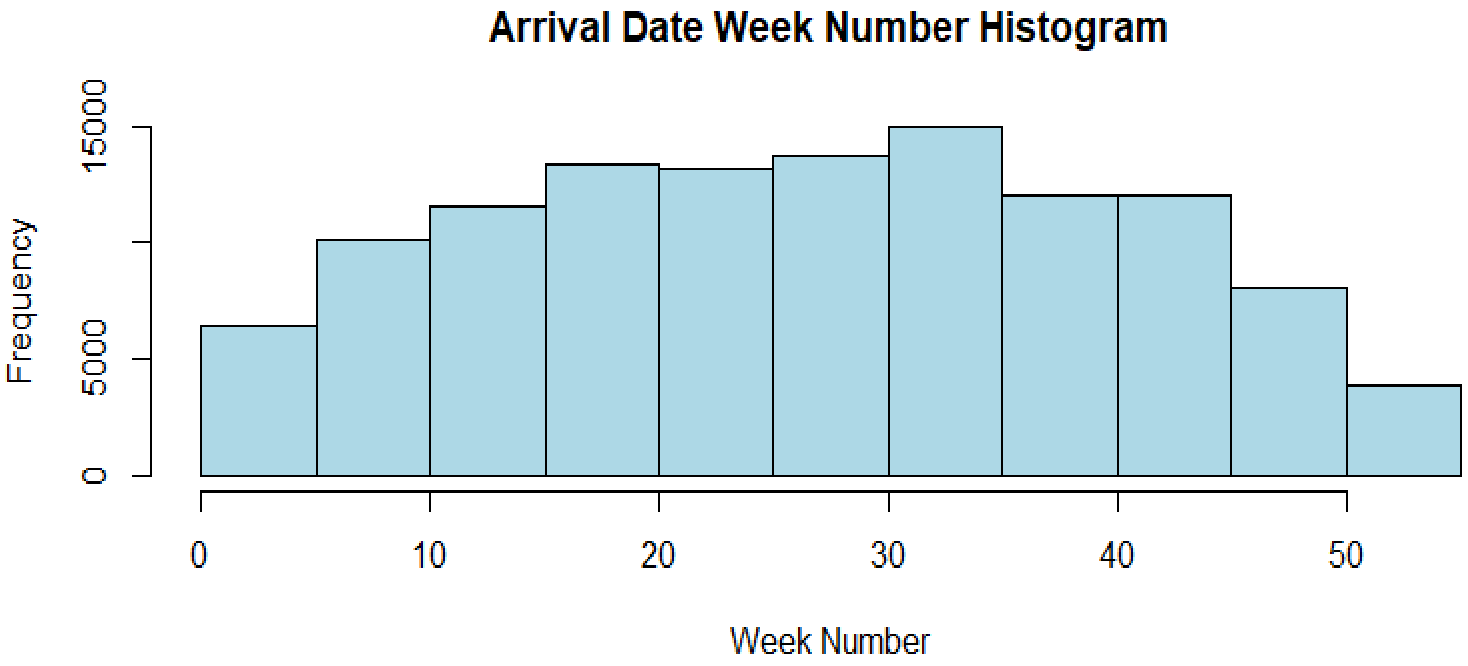
Sometimes one wonder when’s the best time to take a vacataion with minimal waiting days or which month in the season is better for a holiday wether it’s a city hotel or a resort, what is the best time to make a reservation,also how to avoid the hotel making overbookings and thus getting other rooms than the one you choose.

- **Project Objectives:**

The goal of this data anlaysis is to provide insight into the workings of the hotel and to give the guest an idea on the best time to make their reservation, also for the hotel to have a prediction if the guest is most likely to cancel their reservation.

**D) The data visualisation graphics with observations and interpretations of each chart:**

## 1st Graph



## Observation:

We can observe that for the 53 weeks in the year, the frequencies vary slightly.

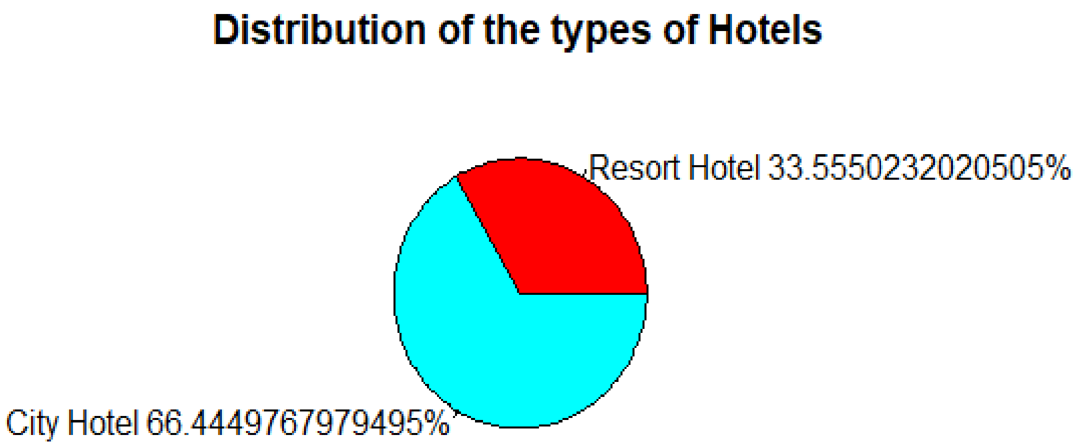
## Interpretation:

There is a major arrival throughout the weeks that decreases at the beginning and the ending of the year.

With people arriving the most in weeks #30 : #35

And arrive the least in the last weeks #50 : #53

# 2nd Graph


Distribution of the types of Hotels
Resort Hotel 33.5550232020505%
City Hotel 66.4449767979495%

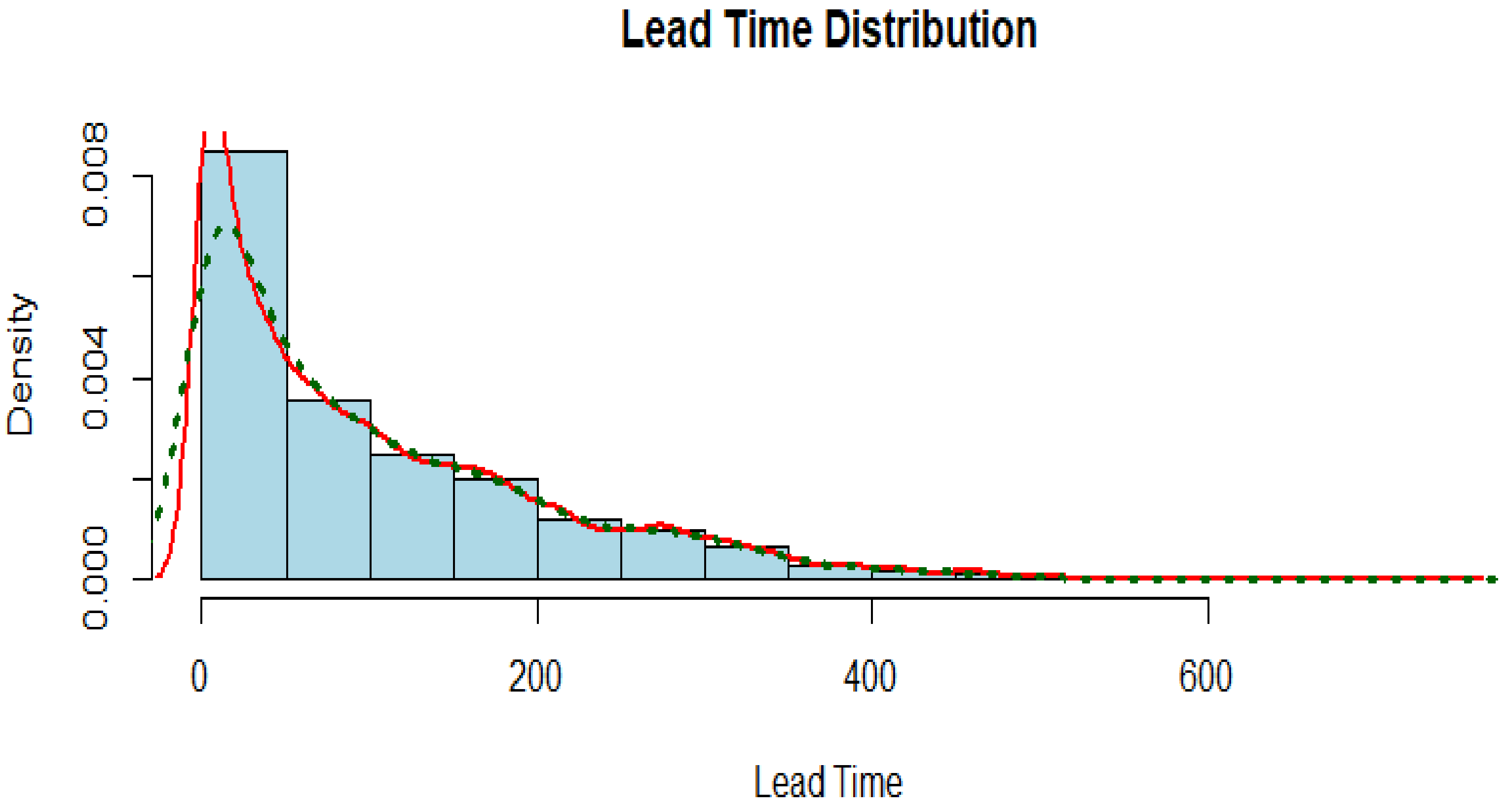
**Observation:**

Observe that the City Hotel possesses the greater portion than the Resort Hotel in this pie chart.

**Interpretation:**

This means that people prefer City Hotel and tend to go and book for the City Hotel more than the Resort Hotel.

## 3rd Graph



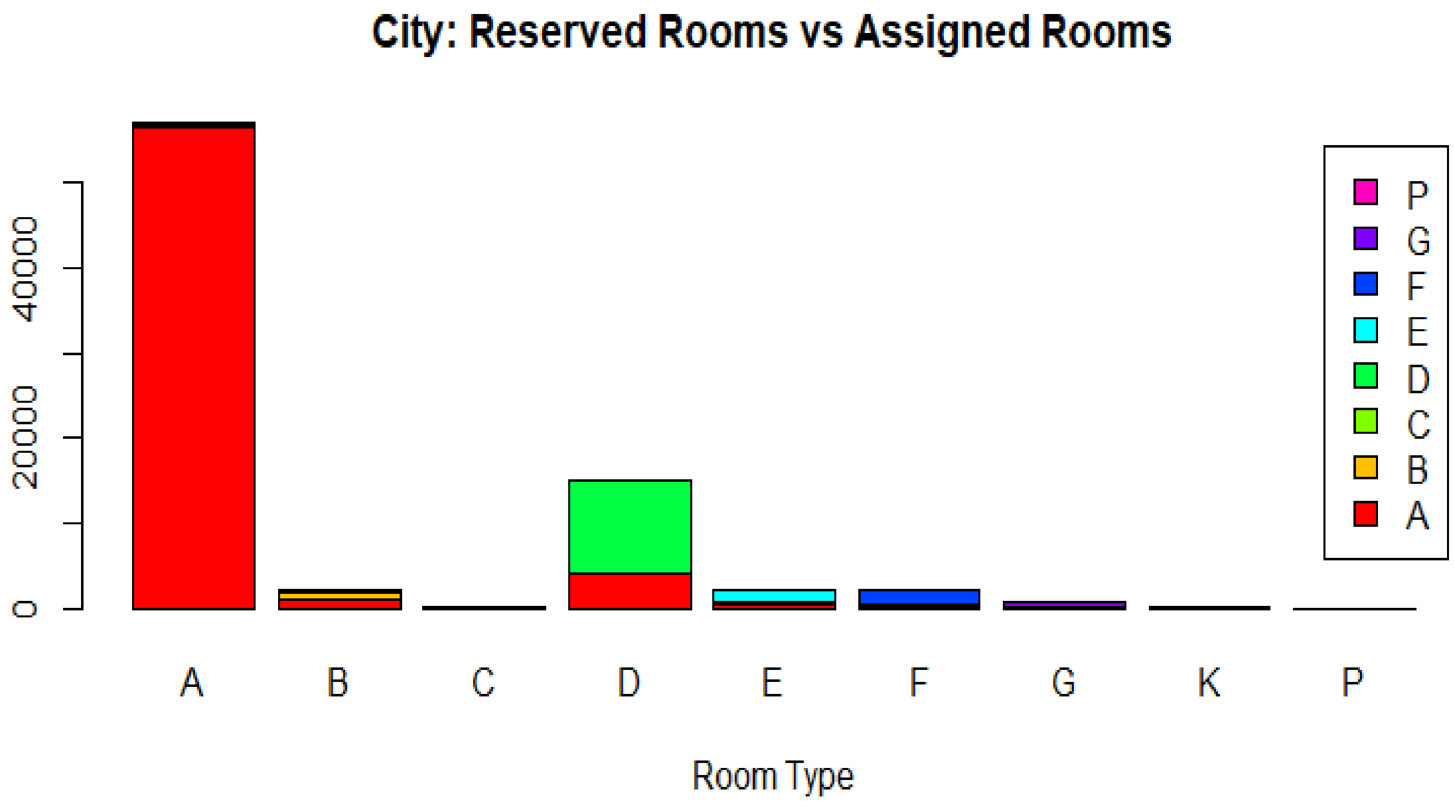
### Observation:

The density of the lead time (Number of days that elapsed between the entering date of the booking into the PMS and the arrival date) when it is equal to 0 is very high and decreases gradually as lead time increases in value.

### Interpretation:

Most bookings are done near or on the date of arrival.

## 4th Graph



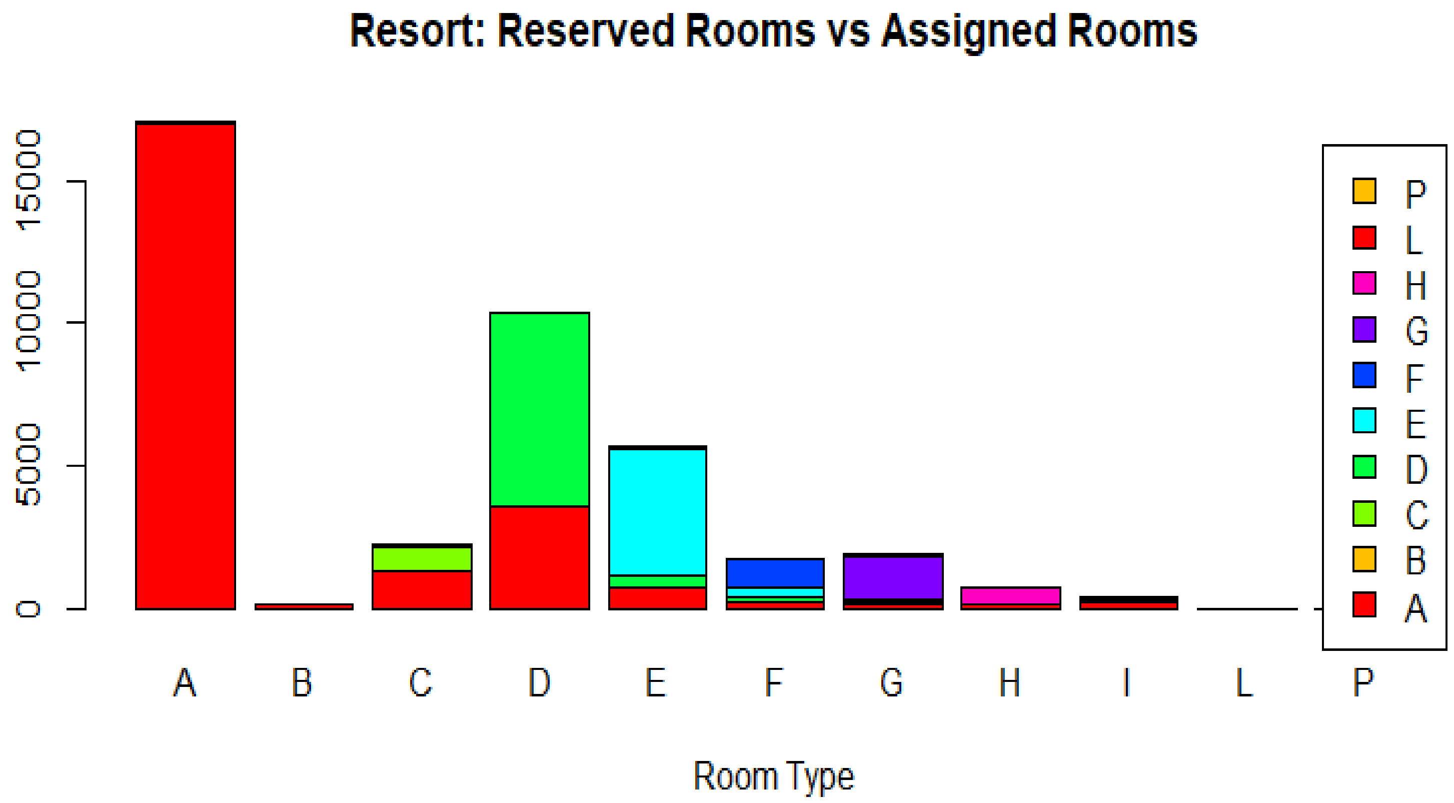
## Observation:

Customers mostly favour room type A, followed by D.
In the booking,most customers are assigned for room type A, followed by D.

## Interpretation:

There is an overbooking in the City Hotel especially with people reserving A type room the most, it can be difficult to assign it as it was reserved.

# 5th Graph


Resort: Reserved Rooms vs Assigned Rooms
0
5000
10000
15000
A
B
C
D
E
F
G
H
I
L
P
Room Type
P
L
H
G
F
E
D
C
B
A

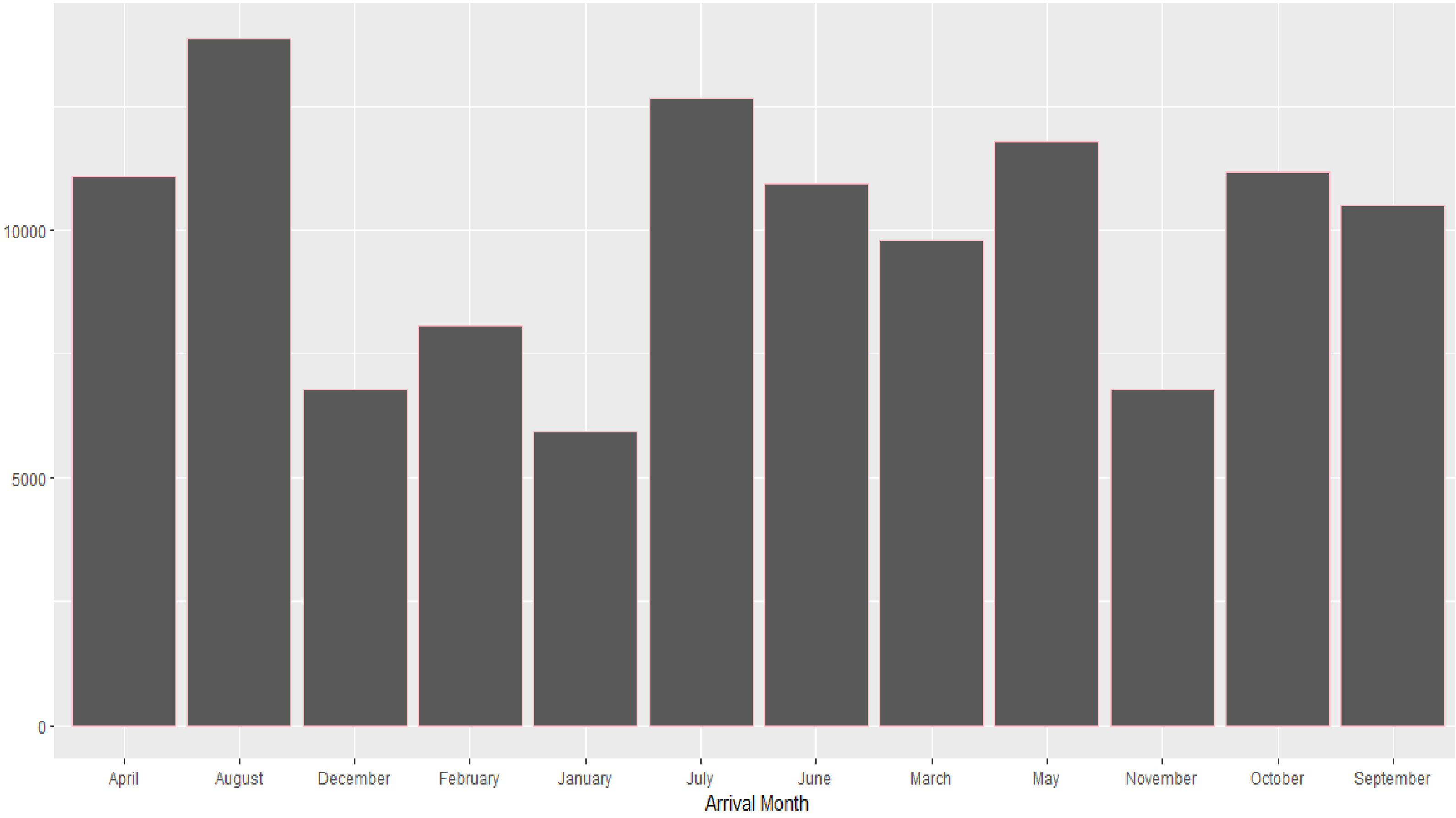
## Observation:

Customers mostly favour room type A, followed by D .
In the booking,most customers are assigned for room type A, followed by D.

## Interpretation:

There is an overbooking in the Resort Hotel especially with people reserving A, D and E type rooms the most, it can be difficult to assign them as they were reserved.

# 6th Graph
# (Bar Plot)



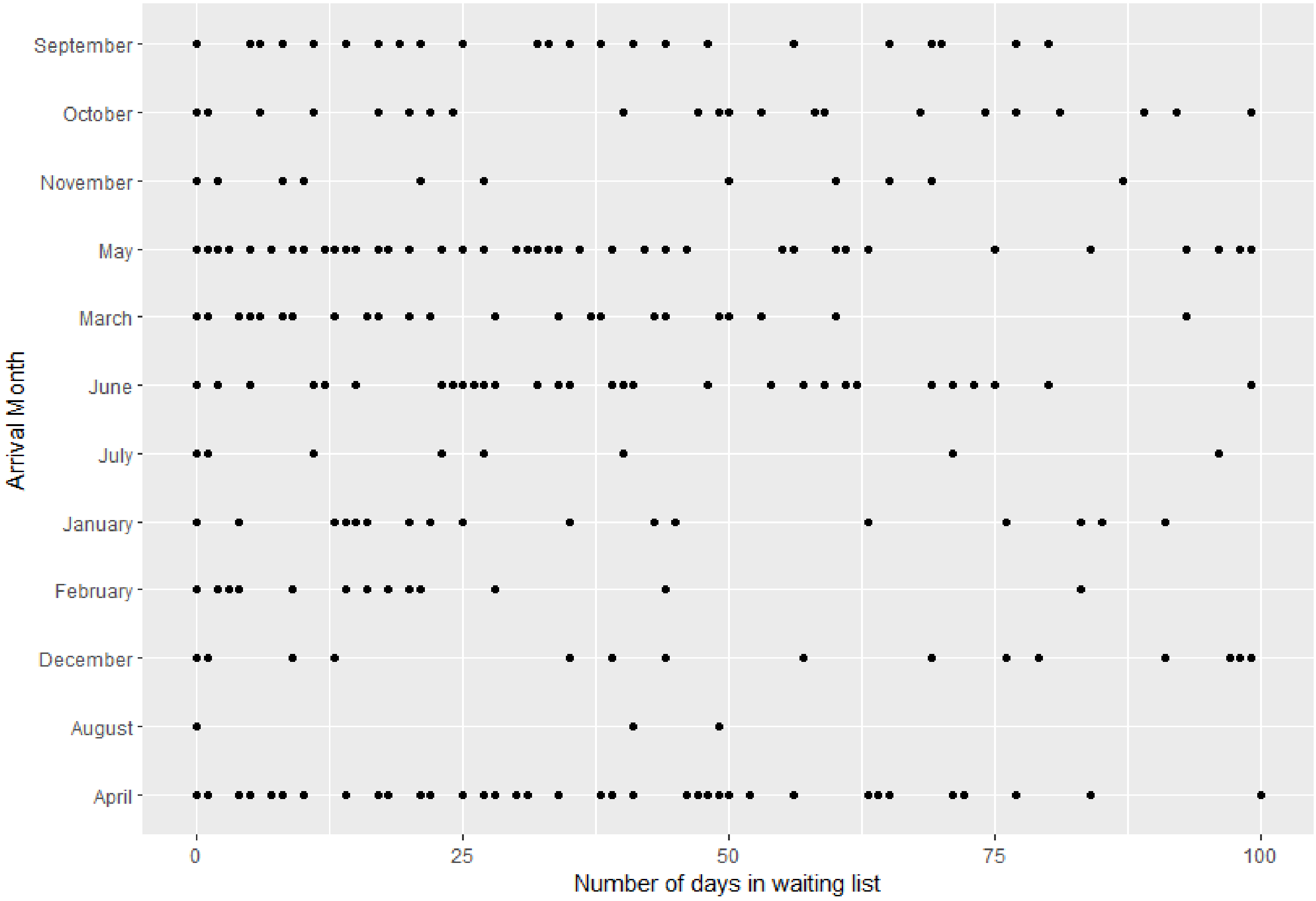
**Observation:**

The plot shows the month against the number of guests, with August having the most guests visiting the city and resort hotel, it’s also obvious how July seems to come second.

**Interpertation:**

This means that people are eager to spend their days at the hotel in these months.

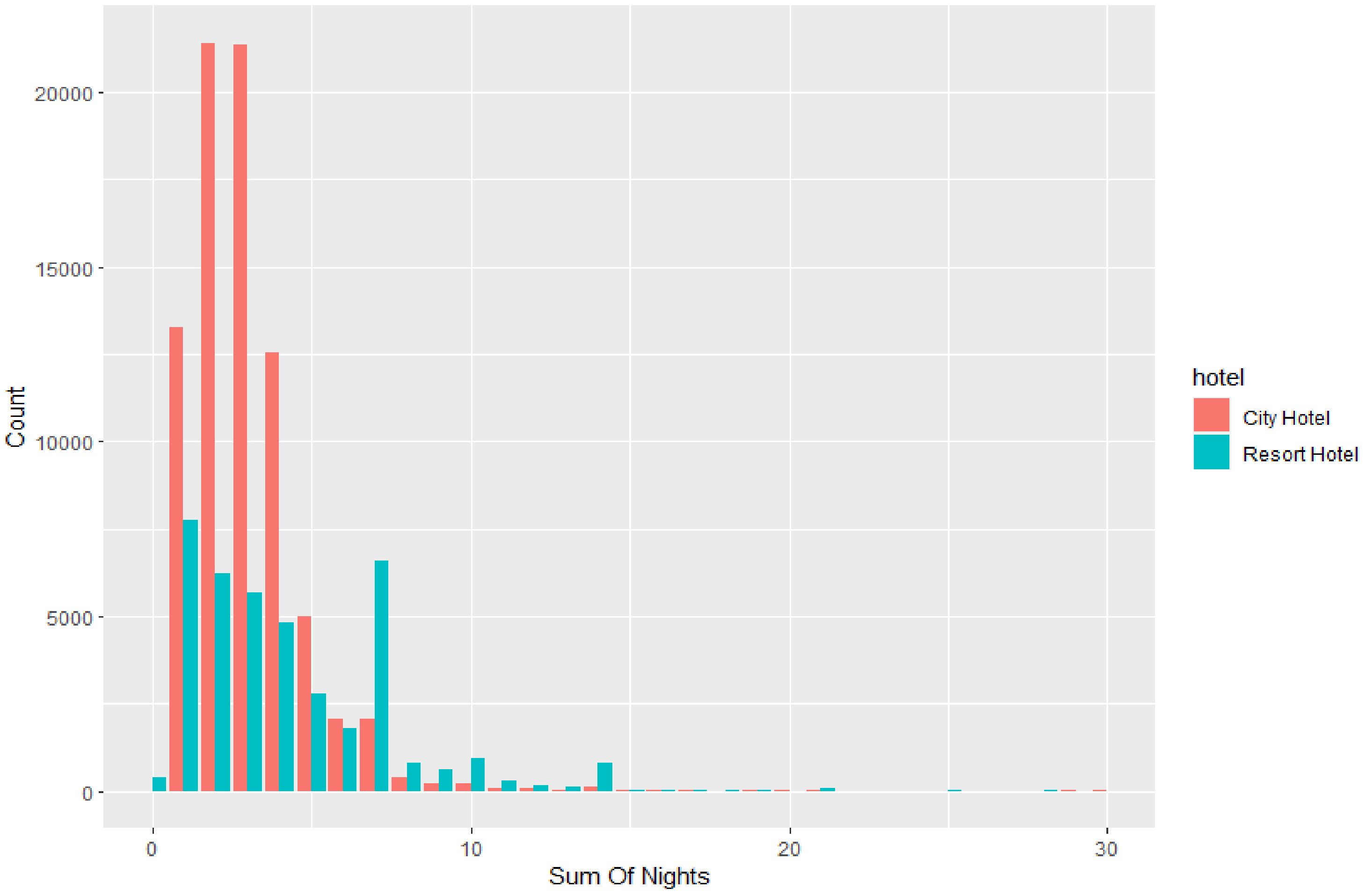
## 7th Graph (Dot Plot)



**Observation:**
The plot shows the number of days in the waiting list that the customer had to wait before their booking was confirmed against the month.

**Interpertation:**
This confirms the 6th plot about August and July being the most busy months in the hotels, as the waiting days are actually longer in these specific months.

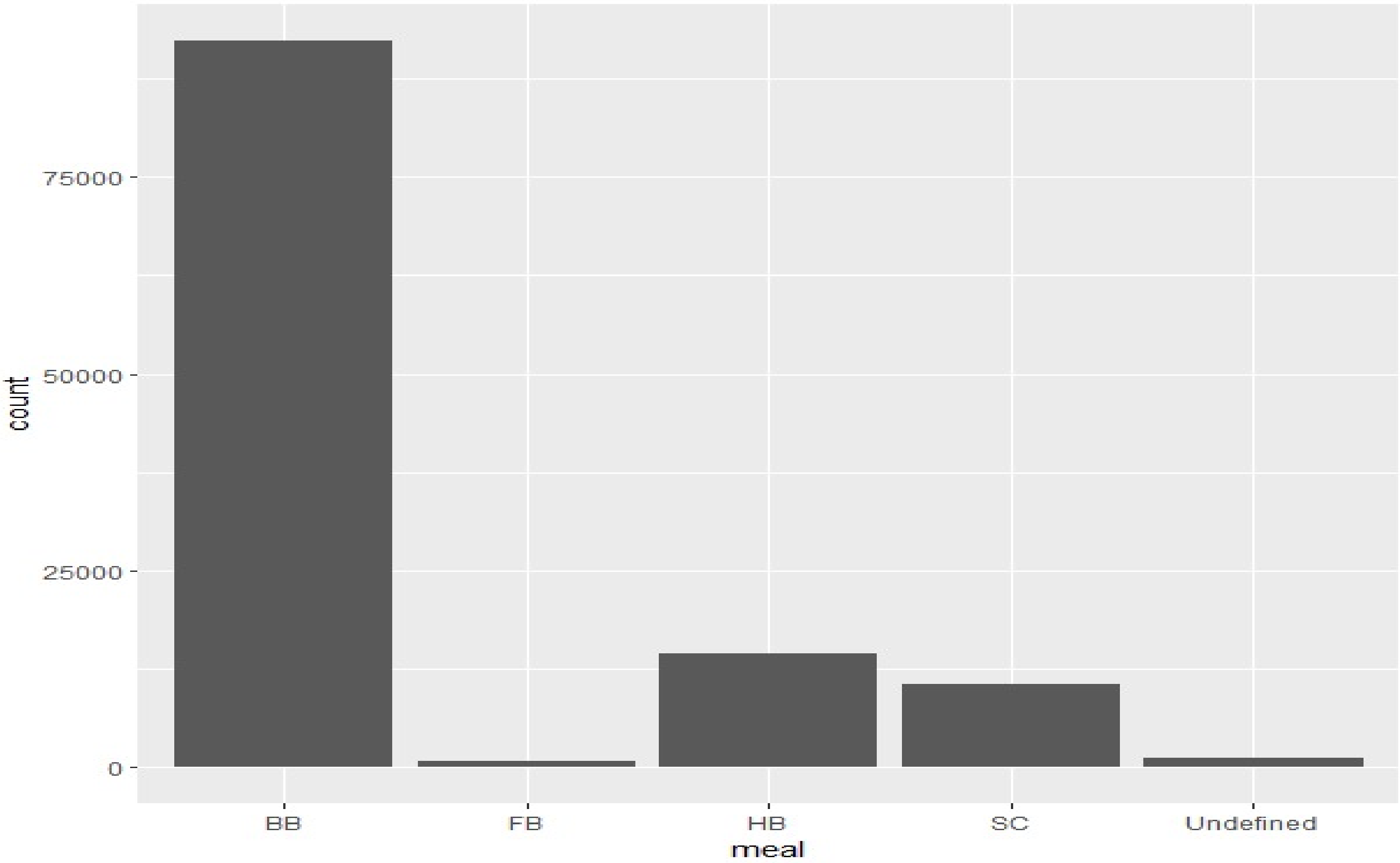
## 8th Graph (Bar Plot)



**Observation:**
The plot shows the number of days the guest spent at the specific hotel, categorised against the actual count of the guests.

**Interpertation:**This means that when guests want to spend longer days at a hotel (more than 7 days) They're most likely reserving at a resort hotel.

## 9th Graph



### Observation :

The plot shows the different meals types against the number of guests choosing that type of meal (BB,FB,HB,SC,Undefined).

### Interpretation:

Most customers prefer to book in BB (bed and breakfast) followed by HB (half board) and SC (without meal package).

## 10th Graph

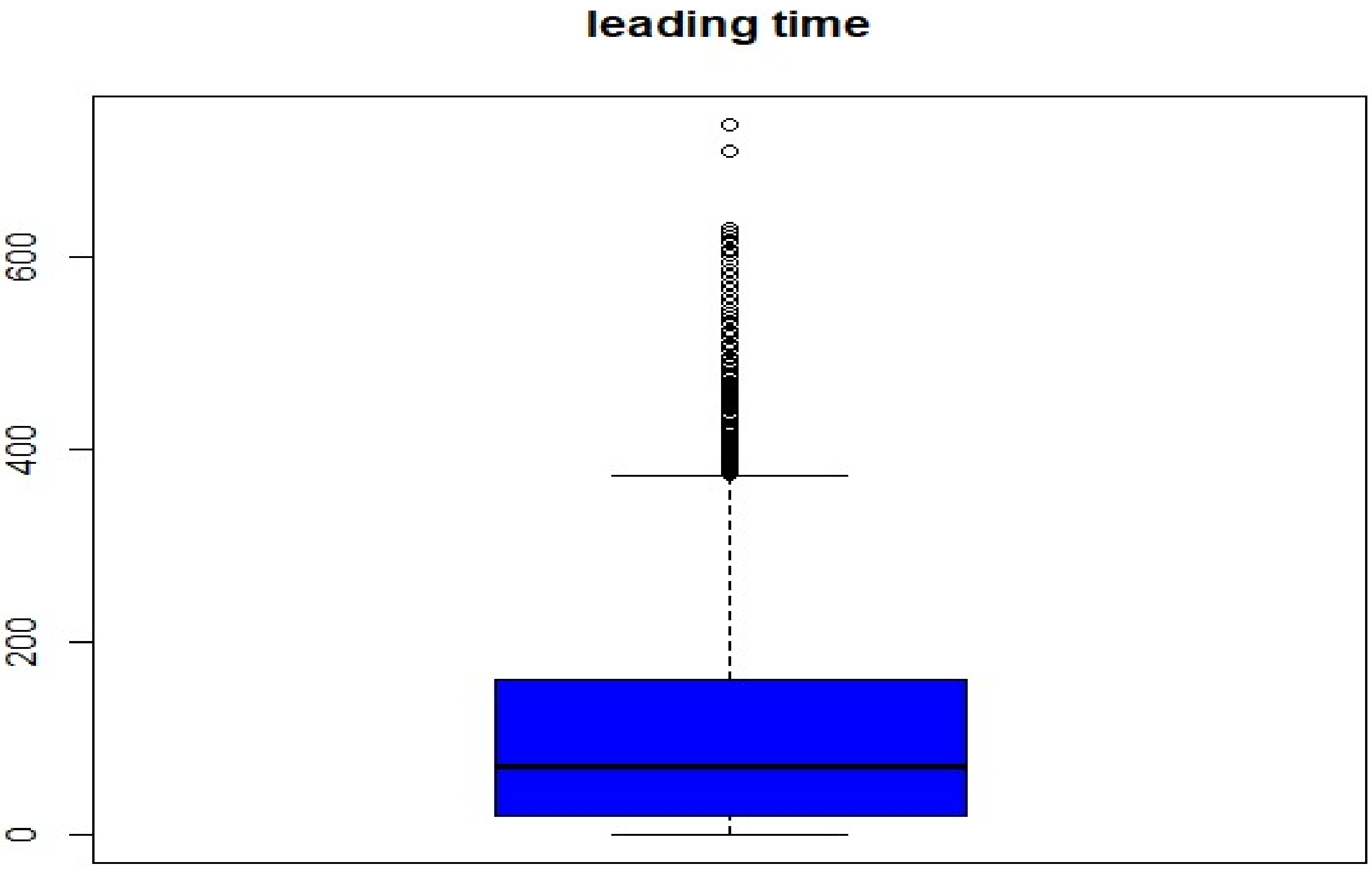
### Observation :

The plot shows the type of reservation the guests had made against the number, types are( No deposit, Non Refund, Refundable).

### Interpretation:

This means that most guests had preferred to book in hotels where it is not necessary to pay a deposit before the reservation.

# 11th Graph



## Observation:

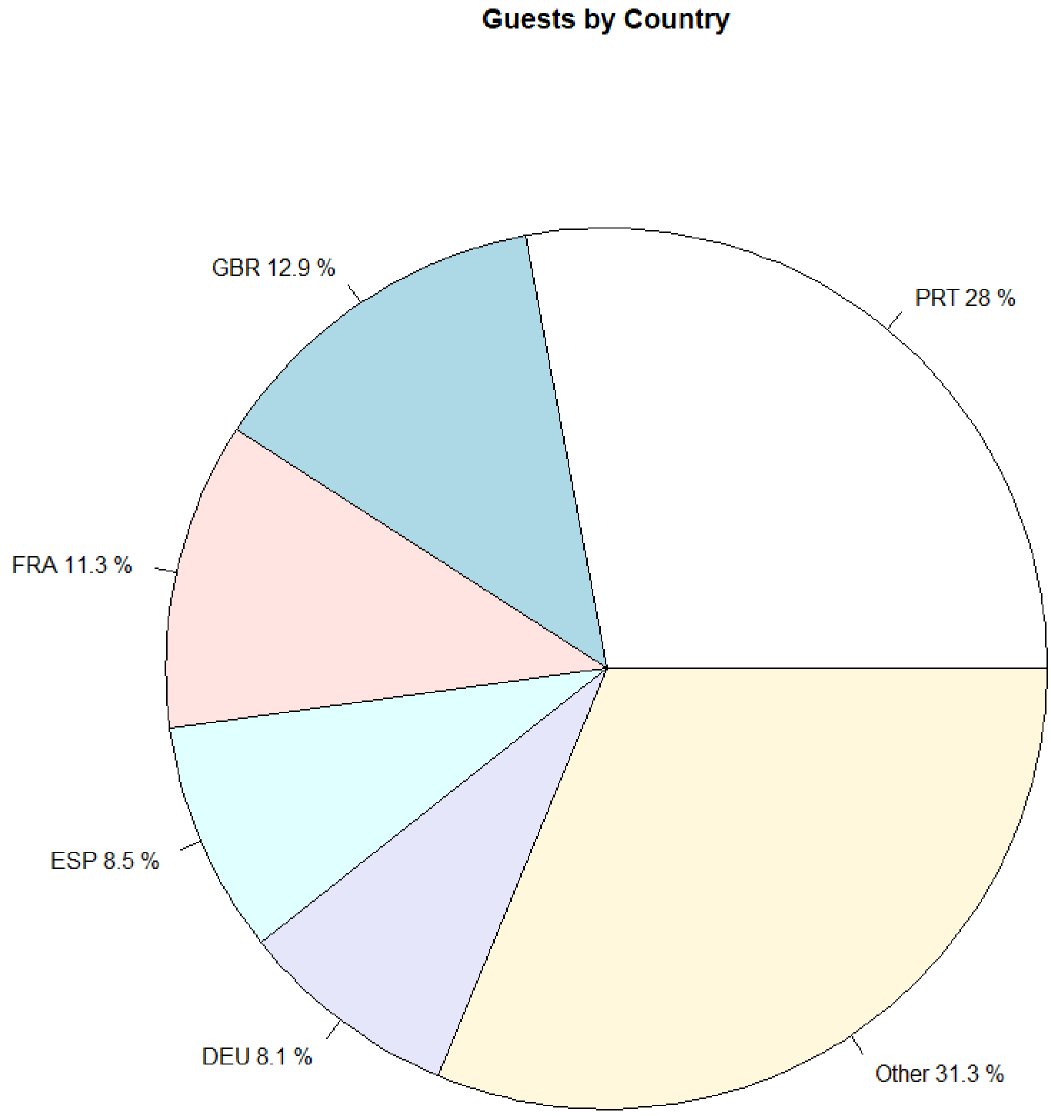
Right skewness (+ve), Find outliers from 400 to up ,The mean Almost at 60,Q1 Almost at 10 and Q3 Almost at 180.

- Maximum=400.
- Minimum=0.

## Interpretation:

The leading time ranges from 0 to 400, but most of them range from 0 to 90.

# 12th Graph


Guests by Country
GBR 12.9 %
PRT 28 %
FRA 11.3 %
ESP 8.5 %
DEU 8.1 %
Other 31.3 %

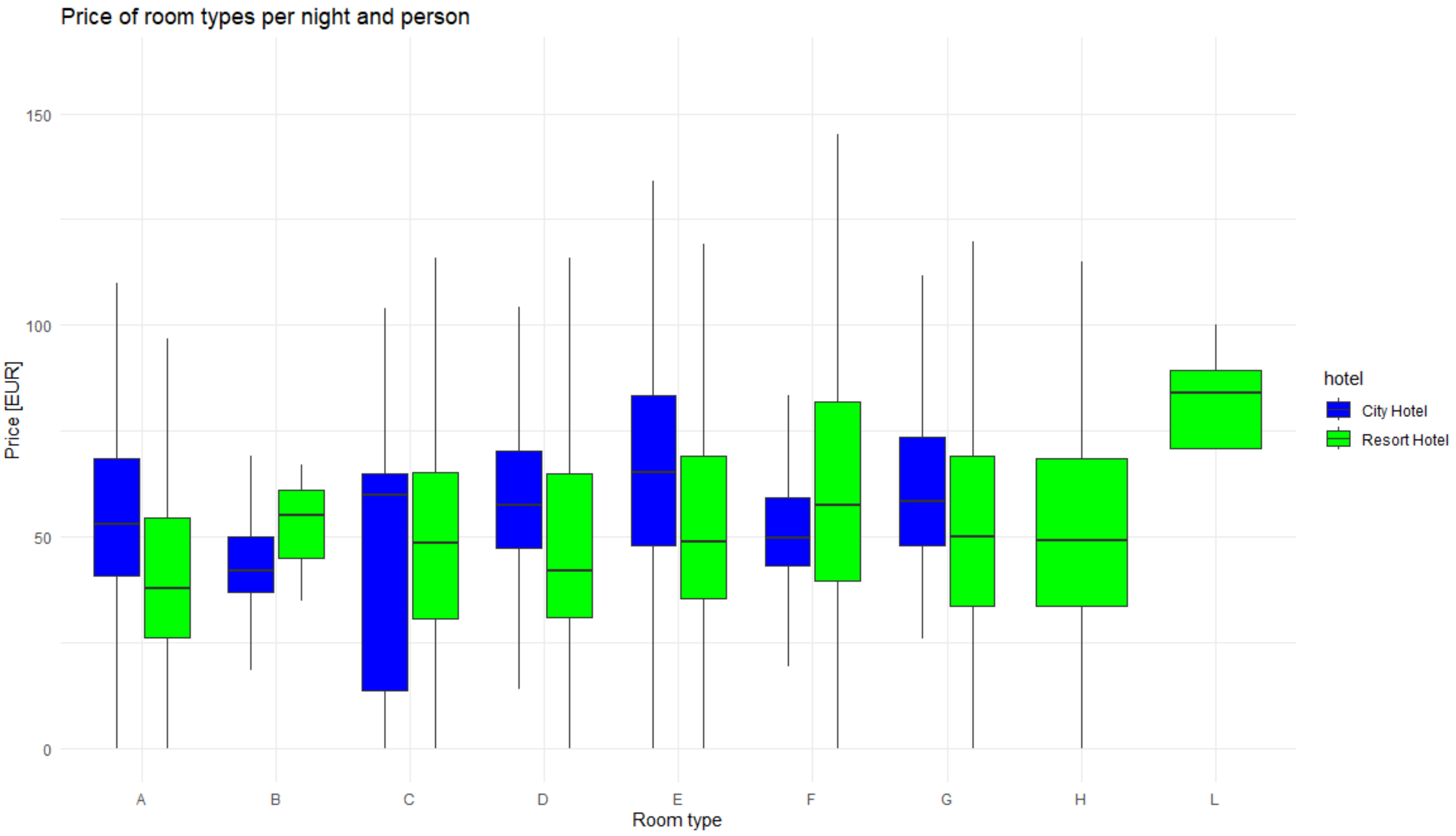
## Observation:

- It appears that PRT “aka Portugal” has the largest portion in the pie chart.

## Interpretation:

- That means that people from all over the world are staying in these two hotels. Most guests are from Portugal and other countries in Europe.

**13th Graph**



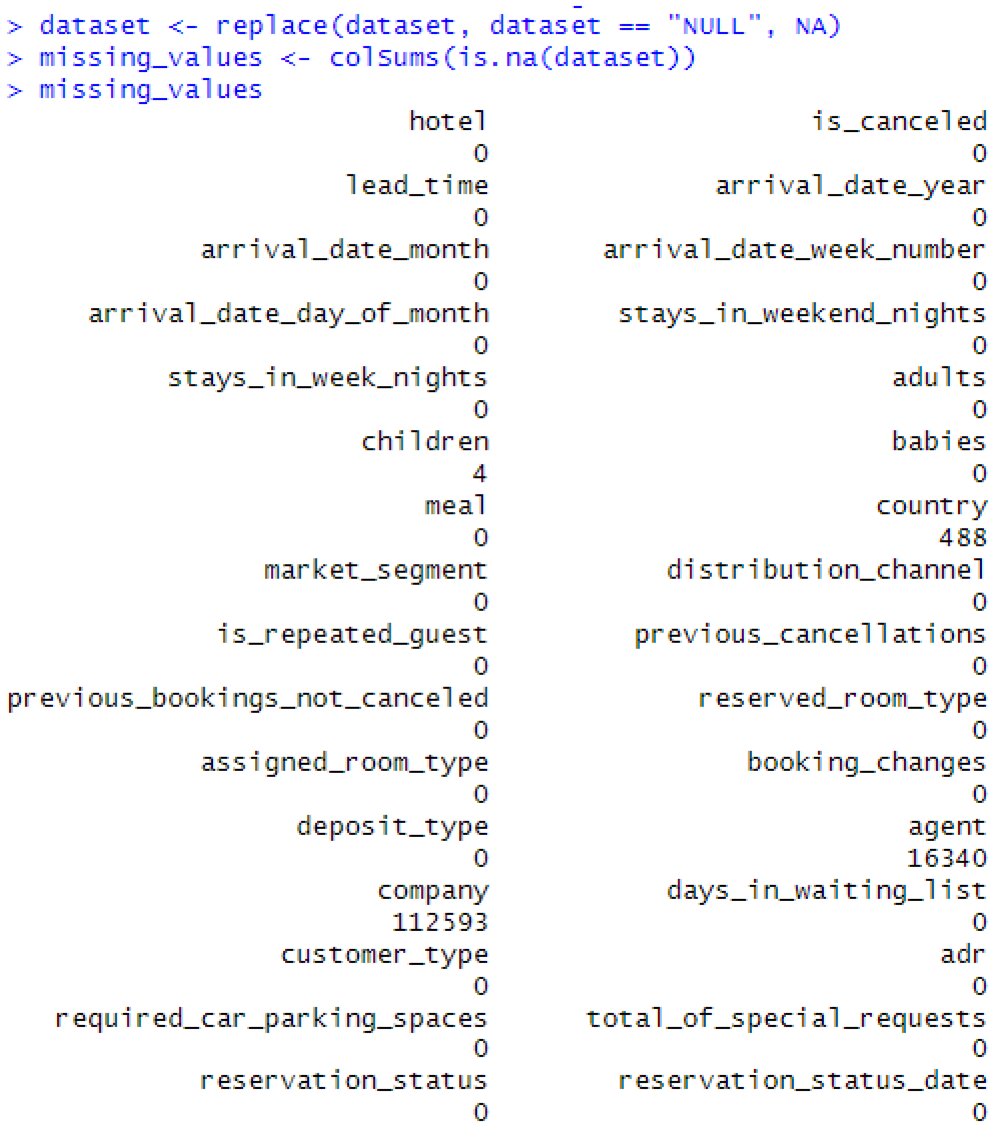
## Observations:

- There are no rooms booked from types of type ‘H’ and ‘L’ in the City Hotel
- It appears from the graph that the average price for a room for the room types is equal to 50

## Interpretations:

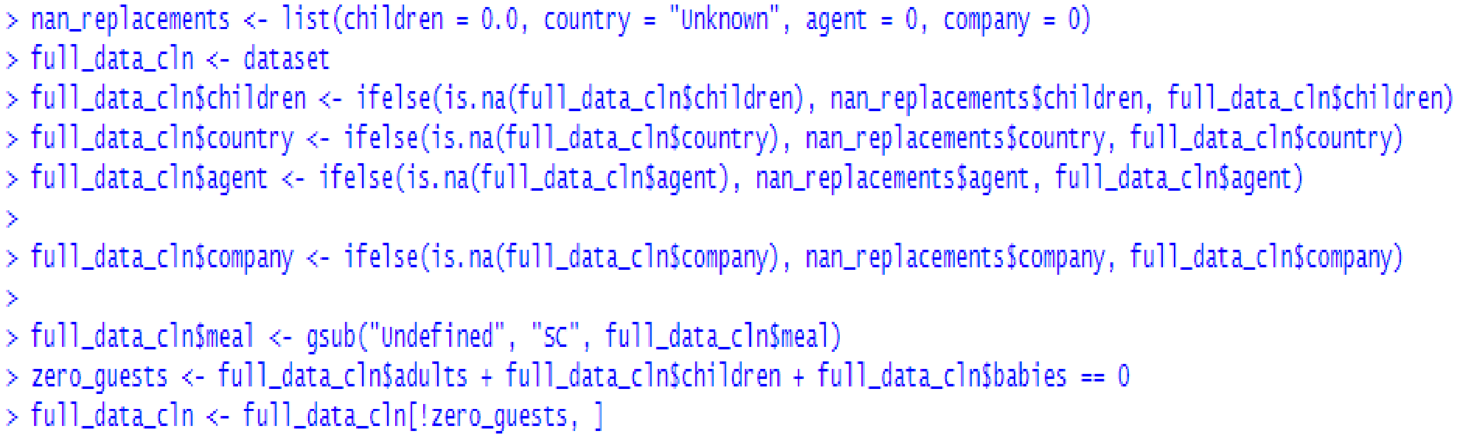
- The Price of room per night and person is always more in the City Hotel except in room type "B" and "F"

## E)Any applied data cleaning



```
> dataset <- replace(dataset, dataset == "NULL", NA)
> missing_values <- colSums(is.na(dataset))
> missing_values
                         hotel                    is_canceled 
                             0                              0 
                     lead_time              arrival_date_year 
                             0                              0 
            arrival_date_month       arrival_date_week_number 
                             0                              0 
     arrival_date_day_of_month        stays_in_weekend_nights 
                             0                              0 
          stays_in_week_nights                         adults 
                             0                              0 
                      children                         babies 
                             4                              0 
                          meal                        country 
                             0                            488 
                market_segment           distribution_channel 
                             0                              0 
             is_repeated_guest         previous_cancellations 
                             0                              0 
previous_bookings_not_canceled             reserved_room_type 
                             0                              0 
            assigned_room_type                booking_changes 
                             0                              0 
                  deposit_type                          agent 
                             0                          16340 
                       company           days_in_waiting_list 
                        112593                              0 
                 customer_type                            adr 
                             0                              0 
   required_car_parking_spaces      total_of_special_requests 
                             0                              0 
            reservation_status        reservation_status_date 
                             0                              0 
```

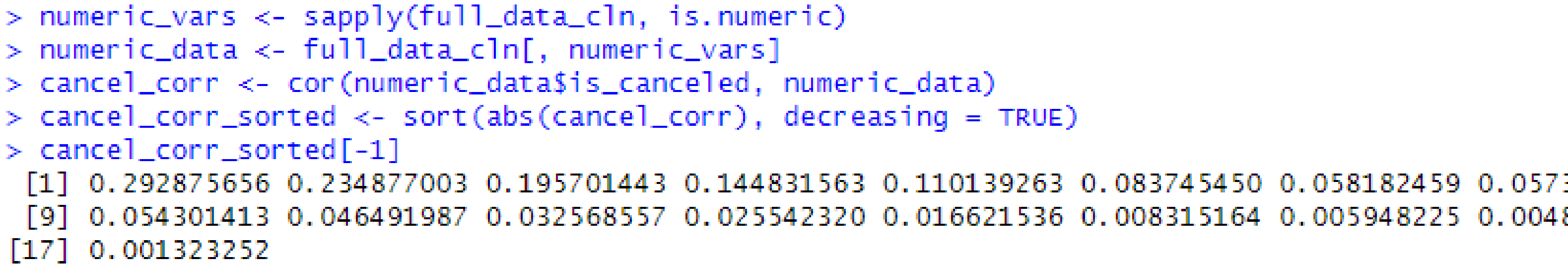
- **Here we found that not all the null values are recorded by “NA” in the dataset, So we replaced the “NULL” with “NA” to be able to omit them.**



```
> nan_replacements <- list(children = 0.0, country = "Unknown", agent = 0, company = 0)
> full_data_cln <- dataset
> full_data_cln$children <- ifelse(is.na(full_data_cln$children), nan_replacements$children, full_data_cln$children)
> full_data_cln$country <- ifelse(is.na(full_data_cln$country), nan_replacements$country, full_data_cln$country)
> full_data_cln$agent <- ifelse(is.na(full_data_cln$agent), nan_replacements$agent, full_data_cln$agent)
> 
> full_data_cln$company <- ifelse(is.na(full_data_cln$company), nan_replacements$company, full_data_cln$company)
> 
> full_data_cln$meal <- gsub("Undefined", "SC", full_data_cln$meal)
> zero_guests <- full_data_cln$adults + full_data_cln$children + full_data_cln$babies == 0
> full_data_cln <- full_data_cln[!zero_guests, ]
```

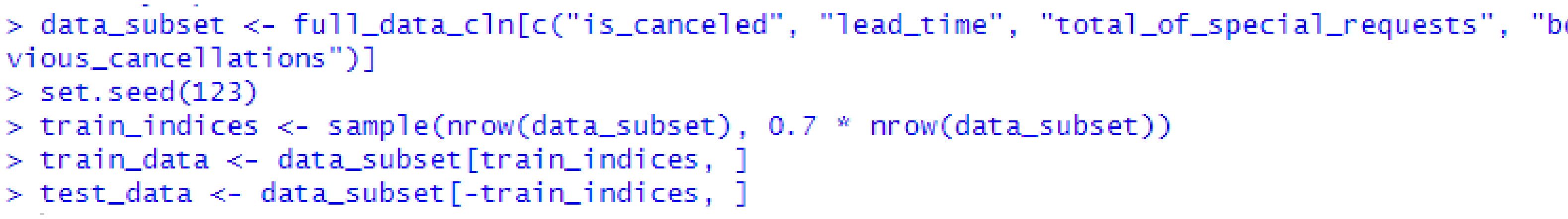
- **And we put default values instead of the Null values if found in the column**
- **Also handling the “undefined” term in the meal with “SC” which is the universal form for undefined meal**
- **We checked for anonymous reservations with (zero adults, children and babies) and removed them**

## F,G)The used data analytic technique “Decision tree”, Dataset preparation



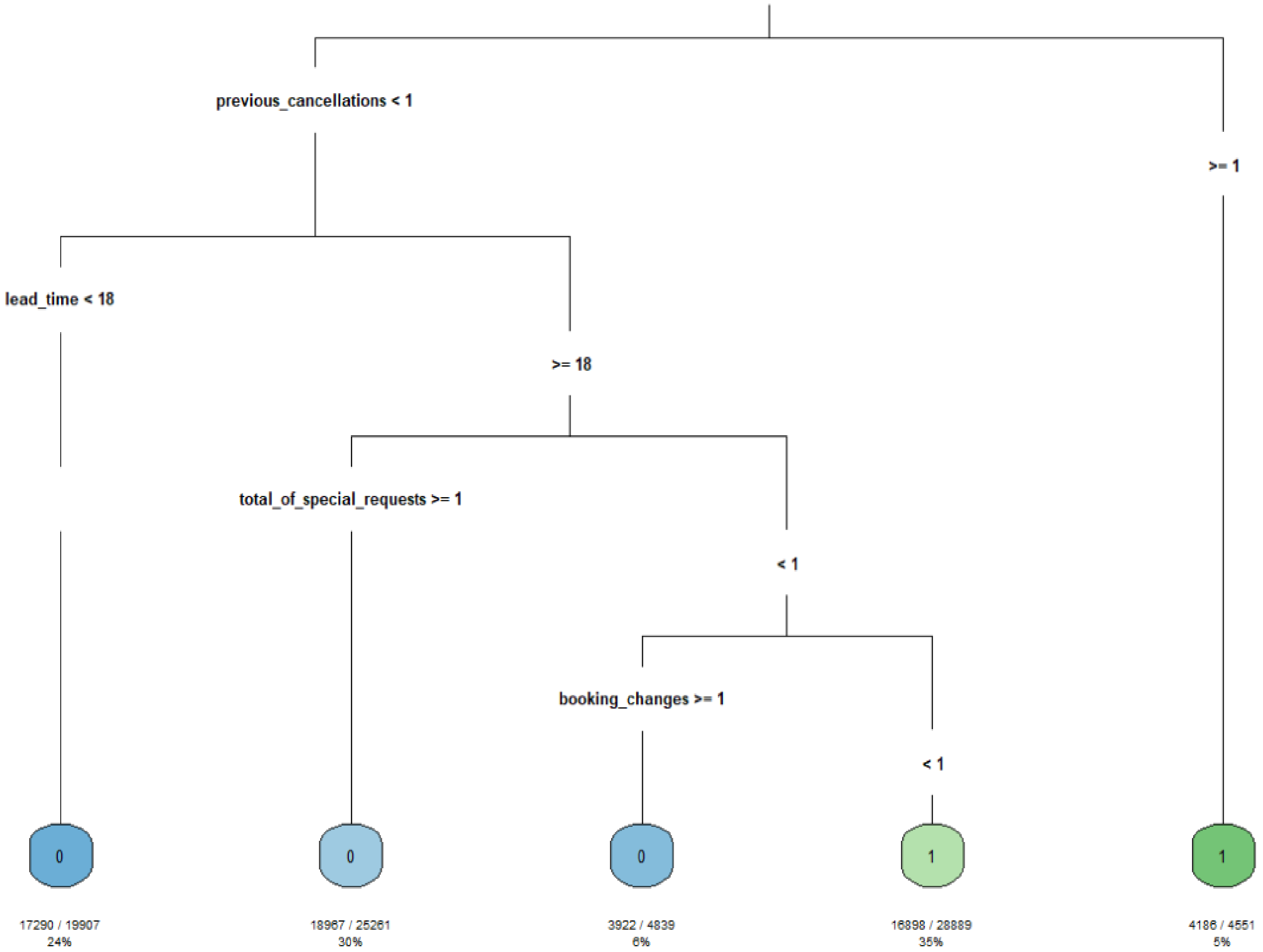
```
> numeric_vars <- sapply(full_data_cln, is.numeric)
> numeric_data <- full_data_cln[, numeric_vars]
> cancel_corr <- cor(numeric_data$is_canceled, numeric_data)
> cancel_corr_sorted <- sort(abs(cancel_corr), decreasing = TRUE)
> cancel_corr_sorted[-1]
 [1] 0.292875656 0.234877003 0.195701443 0.144831563 0.110139263 0.083745450 0.058182459 0.0573
 [9] 0.054301413 0.046491987 0.032568557 0.025542320 0.016621536 0.008315164 0.005948225 0.0048
[17] 0.001323252
```

- We found that the strongest correlation between columns and the is_cancelled column are (lead_time,total_of_special_requests,booking_changes,previous_cancellations).
- So we used them as predictors attributes in the decision tree to predict if the guest will cancel the reservation.



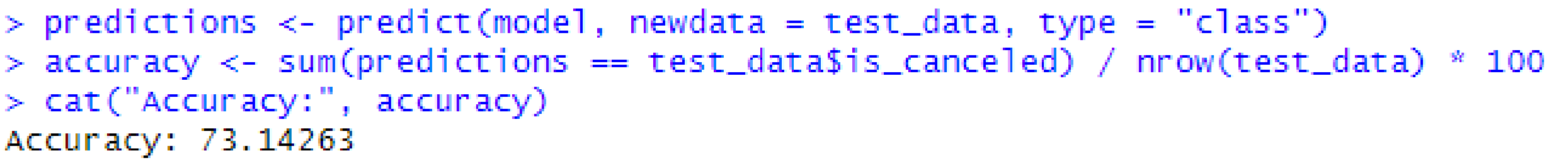
```
> data_subset <- full_data_cln[c("is_canceled", "lead_time", "total_of_special_requests", "bo
vious_cancellations")]
> set.seed(123)
> train_indices <- sample(nrow(data_subset), 0.7 * nrow(data_subset))
> train_data <- data_subset[train_indices, ]
> test_data <- data_subset[-train_indices, ]
```

- Here, we prepared the needed part from the data in the “data_subset”, and initialised 0.7 of the data for the training set, and the rest “0.3” for the testing set.



- **Here is the decision tree with the percentages of each scenario from the whole records in the training set, which predicts if the guest will cancel the reservation or not.**

## H)Performance measure of the analytic technique



```
> predictions <- predict(model, newdata = test_data, type = "class")
> accuracy <- sum(predictions == test_data$is_canceled) / nrow(test_data) * 100
> cat("Accuracy:", accuracy)
Accuracy: 73.14263
```

- **The accuracy of the decision tree is 73.14%, which is the highest accuracy that could be achieved from the other analytic techniques we've tried**

## I)Discussion for relevant project findings

- Seasonal Patterns and Hotel Preference: The observation of seasonal patterns in hotel bookings, with peak demand in certain months like August and July, can be connected to the finding that the City Hotel is more popular among guests compared to the Resort Hotel. Understanding seasonal demand can help hotels allocate resources and optimise pricing strategies accordingly.

- Lead Time and Overbooking Challenges: The finding that most bookings are made close to the arrival date (short lead time) is connected to the challenge of overbooking, especially for popular room types. Short lead times can make it difficult for hotels to manage inventory and assign rooms, leading to potential overbooking issues.
- Longer Stays and Resort Hotel Preference: The finding that guests staying longer than 7 days tend to book more at the Resort Hotel indicates a connection between the preference for longer stays and the choice of the Resort Hotel. Guests looking for extended vacations or leisure trips may prefer the amenities and offerings of a resort setting.
- Seasonal Demand and Waiting List: The observation that waiting list days are longer in months with higher demand, such as August and July, suggests a connection between seasonal demand and the need for guests to wait for confirmation. Higher demand can lead to a higher number of reservations and potentially longer waiting times for guests.

- Meal Packages and Price Variation: The preference for bed and breakfast (BB) meal packages and the higher room prices in the City Hotel are interconnected findings. Guests who opt for the bed and breakfast package may be willing to pay higher prices for a more inclusive experience, which could explain the price variation between hotels.

- Room Type Preference and Overbooking Challenges: The finding that certain room types, such as A and D, are most preferred by guests in both hotels, connects to the challenge of overbooking. Overbooking issues may arise specifically for these popular room types, as there might be more requests than available rooms, leading to potential difficulties in assigning rooms as reserved.
- Lead Time and Price Variation: The distribution of lead time, with most bookings made within 90 days and lead times ranging from 0 to 400, connects to the variation in room prices. The timing of bookings and lead time may impact pricing strategies, with last-minute bookings potentially resulting in different pricing structures compared to bookings made well in advance.

- Prediction Model Accuracy and Decision-Making: The accuracy of the decision tree model in predicting reservation cancellations connects to the project objectives of enhancing decision-making. A reliable predictive model can assist hotel management in identifying potential cancellations and taking appropriate actions, such as adjusting inventory, optimising staffing, or implementing targeted marketing strategies.