Capstone Project

EDA Of Hotel Booking Analysis

Ravali Muta

Content:

- Introduction
- Loading Libraries (Numpy,Pandas,Matplotlib,Seaborn)
- Reading input dataset by df.info(), df.describe(), df.columns()
- Cleaning the Data and dropping unnecessary columns
- Analyzing the data and plotting the charts from the given Dataset
- Conclusion

Introduction :

- Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week and many more.

- This makes analyzing the patterns available in the past data more important to help the hotels plan better.

- Using the hotels data . We can use the patterns to predict the bookings

Loading Libraries

```python
# Importing the libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

import datetime as dt
```

- Numpy is python library used for working with arrays
- Pandas allows importing data from various file format such as CSV, Excel, Json, SQL
- Matplotlib is a plotting library for the Python Programming Language and its numerical Mathematics extension library
- Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structure in python
- DateTime is used to import to work with the date as well as time

Reading the information of Dataset by .info and .columns

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

```python
df.columns
```

```
Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
       'arrival_date_month', 'arrival_date_week_number',
       'arrival_date_day_of_month', 'stays_in_weekend_nights',
       'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
       'country', 'market_segment', 'distribution_channel',
       'is_repeated_guest', 'previous_cancellations',
       'previous_bookings_not_canceled', 'reserved_room_type',
       'assigned_room_type', 'booking_changes', 'deposit_type',
       'customer_type', 'adr', 'required_car_parking_spaces',
       'total_of_special_requests', 'reservation_status',
       'reservation_status_date'],
      dtype='object')
```

Most important step is to clean the data.

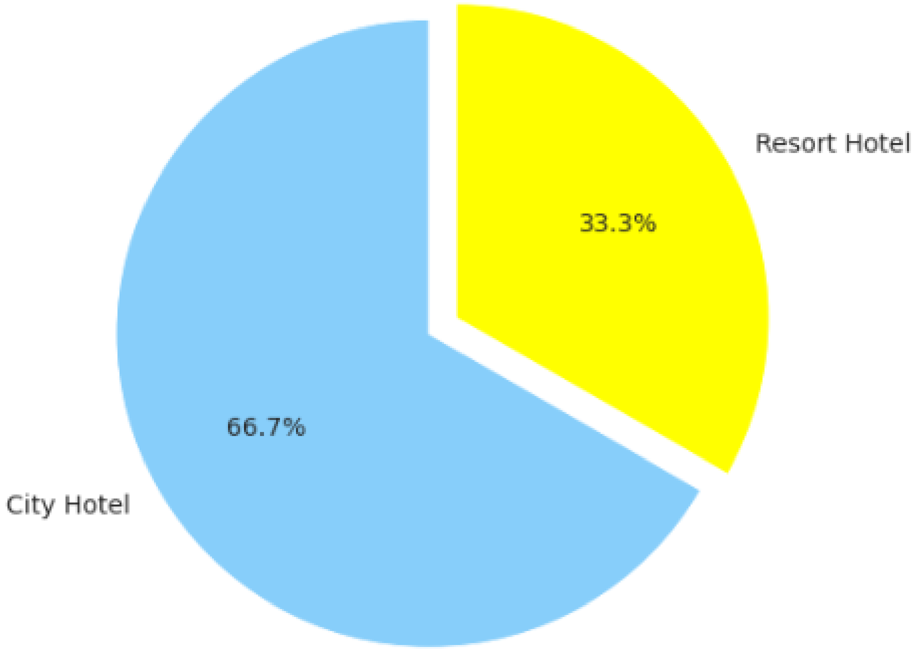
deposit_type	agent	company
No Deposit	NaN	NaN
No Deposit	NaN	NaN
No Deposit	NaN	NaN
No Deposit	304.0	NaN
No Deposit	240.0	NaN

- We have NaN values in Agent and Company but that columns are negligible so we drop that columns.

```
[11] #Lets drop columns with high missing values "agent" and "company".
     df.drop(['agent','company'], axis = 1 , inplace = True)
```

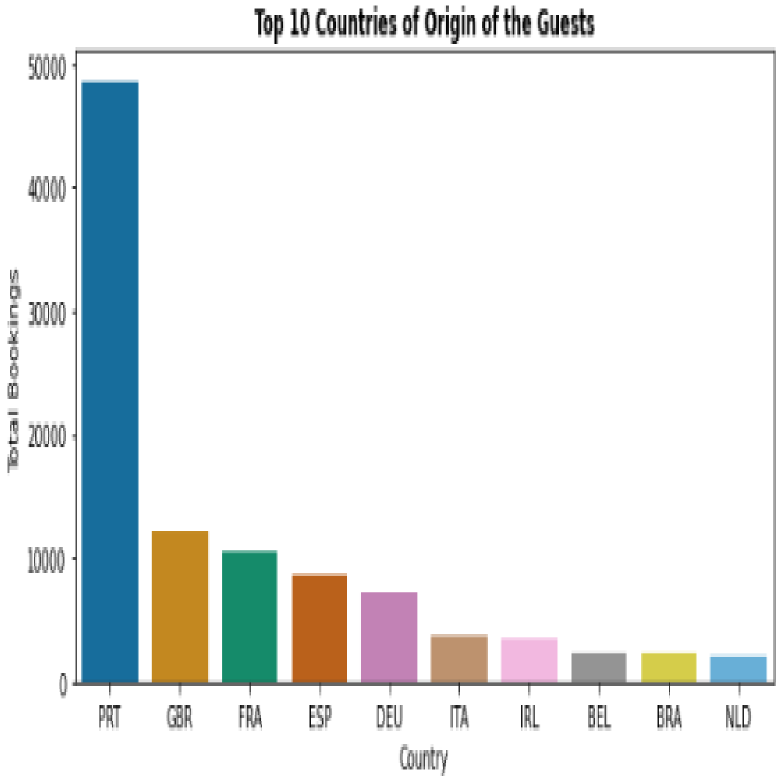
EDA
1. Hotel Comparison :

Here we just compared the Resort hotel and City Hotel In terms of bookings.



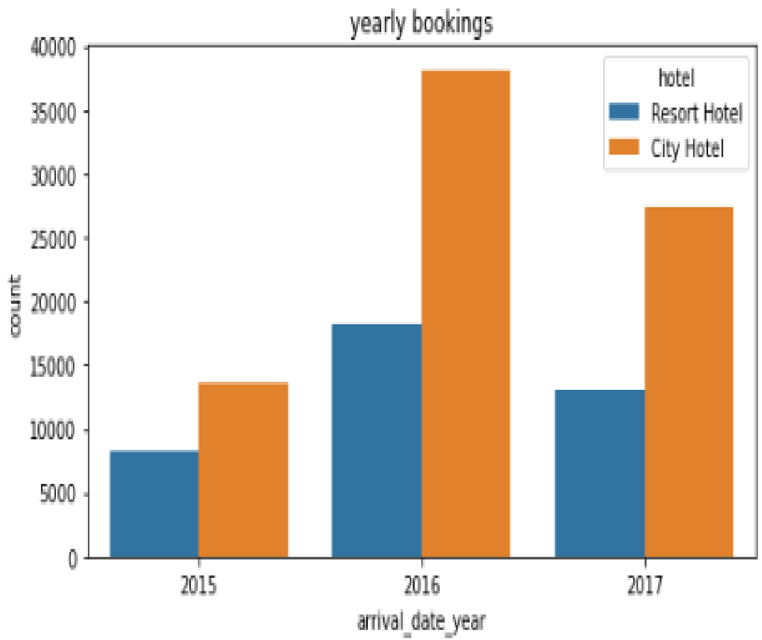
2. Country wise Guests :

➢ **Portugal, UK and France, Spain and Germany are the top countries from most guests come, more than 80% come from these 5 countries.**



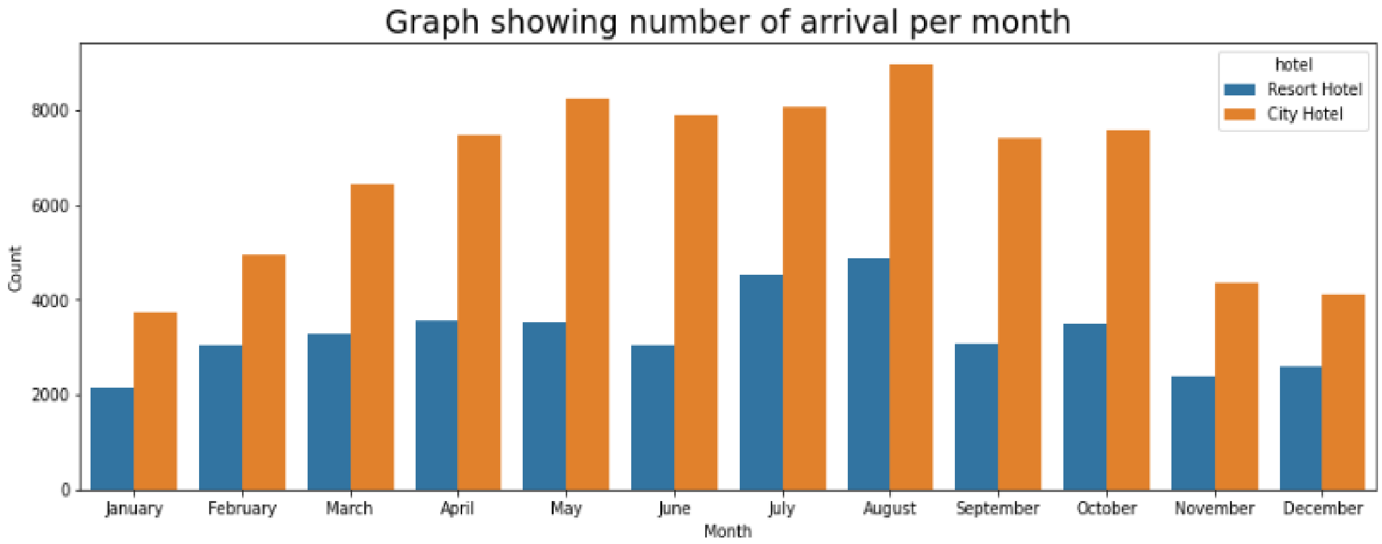
3. Year-wise and hotel-wise bookings :

➢ **Here we can see that 2016 seems to be the year where hotel booking is at its highest.**



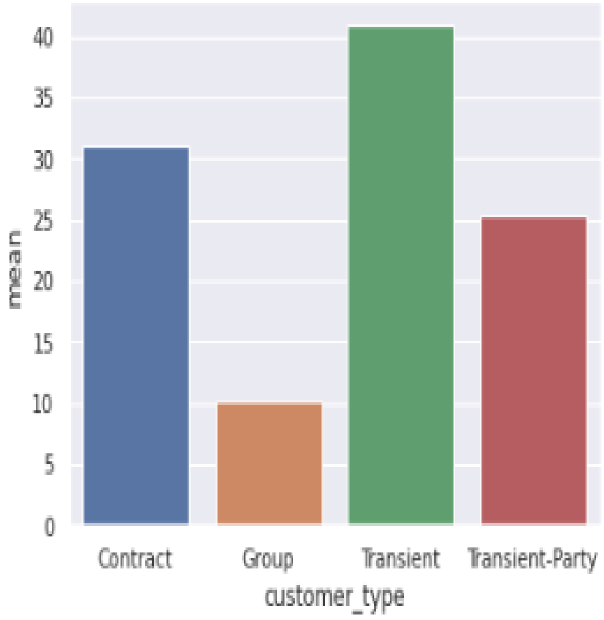
4. Month-wise and hotel-wise arrivals :

We see an increasing trend in booking around the middle of the year, with August being the highest



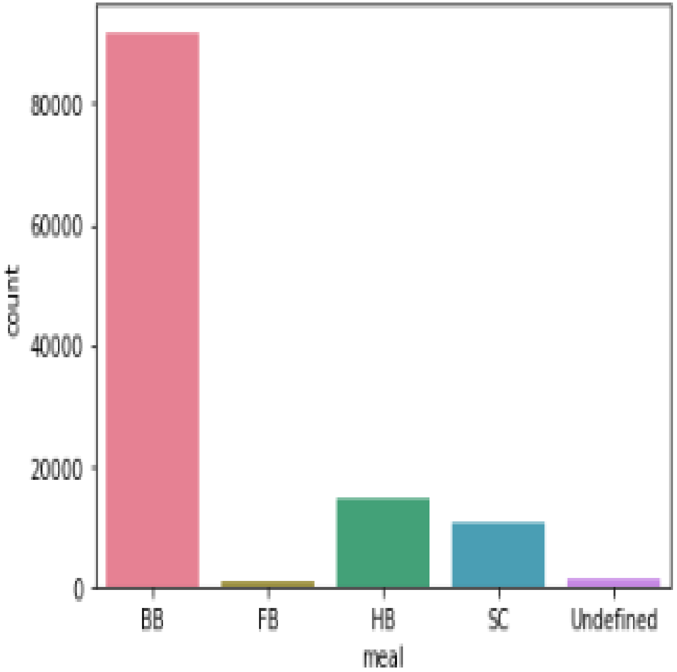
5. Customer Type Bookings:

- **Majority of the bookings are transient. with the ease of booking directly from the website, most people tend to skip the middleman to ensure quick response from their booking.**



6. Meal Type :

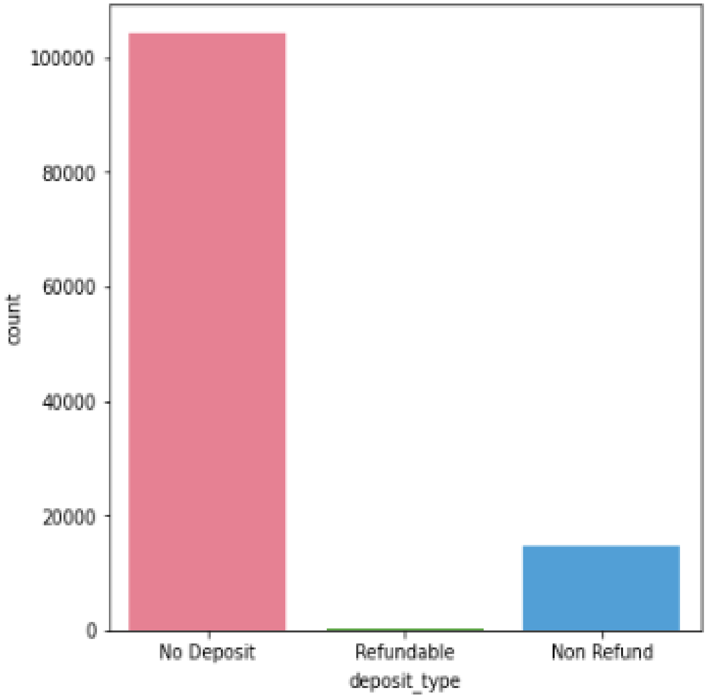
```
BB           0.772620
HB           0.121398
SC           0.089472
Undefined    0.009798
FB           0.006712
Name: meal, dtype: float64
```



As we can see that bed and breakfast as the high count with respect to other meal type

7. Deposit Type :

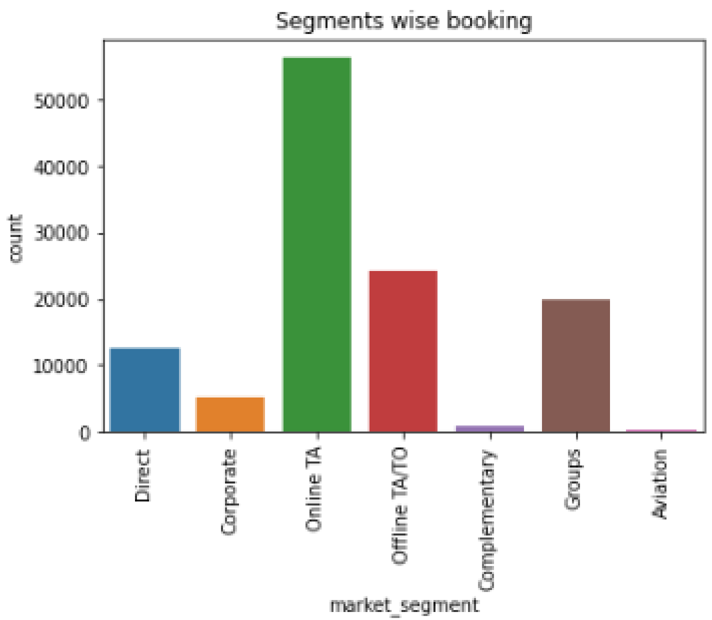
➢ Majority of the booking does not require deposit. That could explain why cancellation rate was actually 50% of non-cancellation rate.



8. Market Type Segment :

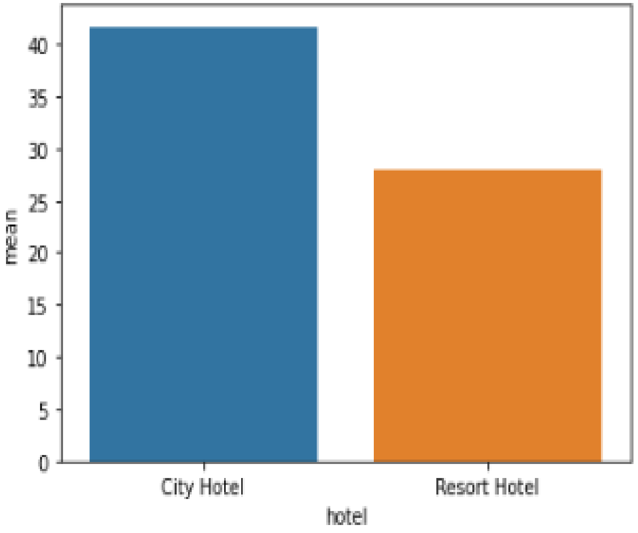
```
Online TA          0.474373
Offline TA/TO      0.203199
Groups             0.166580
Direct             0.104695
Corporate          0.042986
Complementary      0.006173
Aviation           0.001993
Name: market_segment, dtype: float64
```

➢ **Indirect bookings through online and offline travel agents are higher compared to direct bookings and same is the case with group bookings which are also high.**



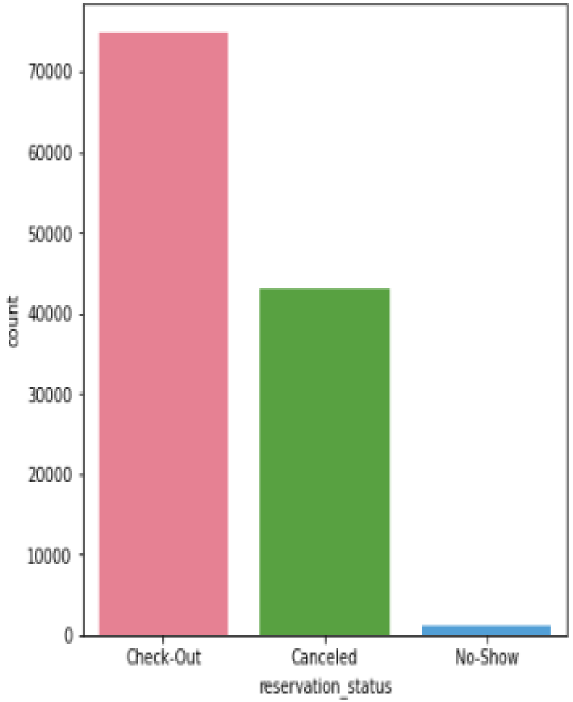
9. Bookings Cancelled :

➢ Here we have seen a huge proportion of cancellation from city hotel.
This was expected since 3/4 of the hotel . bookings belong to city hotels.



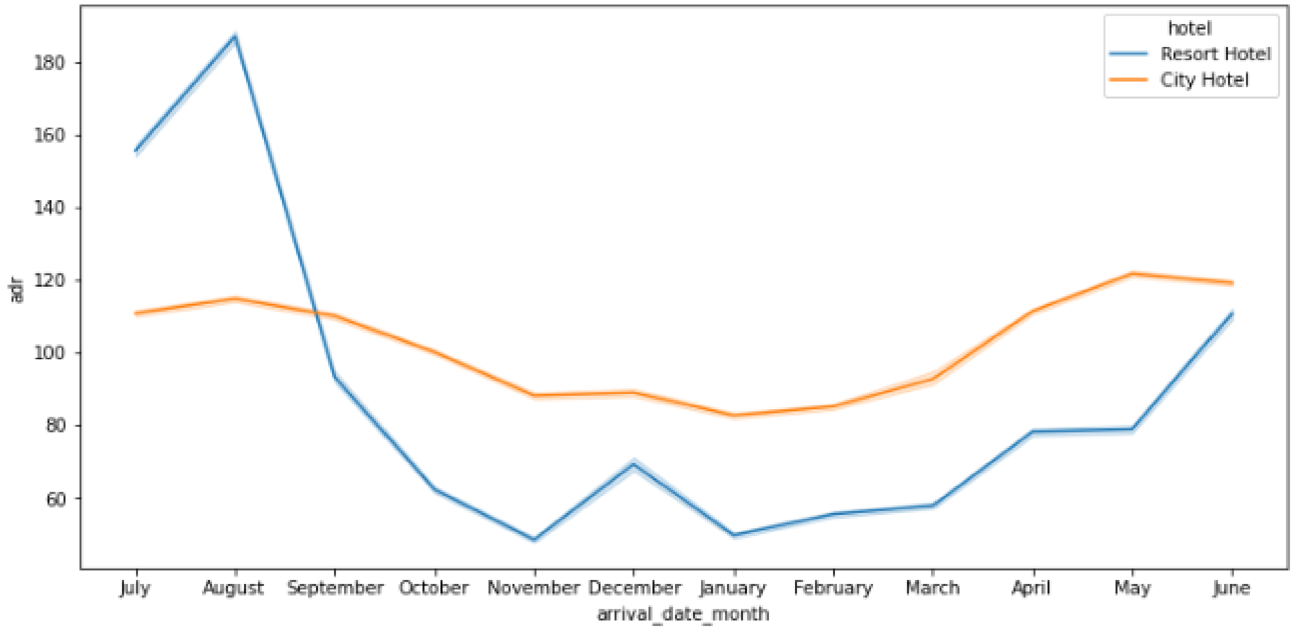
10. Reservation status :

- 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.



11. Average Daily Rate (ADR) :

- Prices of resort hotel are much higher. It seems that that is definitely the case since resort hotels specialize in that.
- Prices of city hotel do not fluctuate that much.



Conclusion :

- Majority of the hotels booked are city hotel.
- We also realize that the high rate of cancellations can be due high no deposit policies.
- The target months between May to Aug. Those are peak months due to the summer period.
- Majority of the guests are from Western Europe.
- Given that we do not have repeated guests.

Q & A