EDA
Capstone Project-1
Hotel Booking Analysis

Made by:
SURAJ KUMAR

Content:

- Problem Statement
- Data Set
- Data Summary
- Data Wrangling
- Exploratory data analysis(EDA)
- Conclusion

Problem Statement:

1. What is the month with the most guest arrivals?
2. What is the year with the most guest arrivals?
3. How does the price vary per night over the year?
4. Which countries do customers come from?
5. What is the strongest market segment and distribution channel?

Data Set :

The dataset includes data on reservations made for two hotels, a resort and a city hotel, both of which are expected to open between July 1, 2015, and August 31, 2017.

The identical data was gathered for both hotels: 32 variables were used to describe 40,060 observations for the resort and 79,330 observations for the city hotel. 119,390 hotel reservations, including those that were cancelled, are included in the dataset. All components that could identify hotels or clients were removed because this is genuine data.

Cleaning data:

The removal of oversized data that could affect the outcome of EDA has made data cleaning a critical aspect in current EDA.

We will take the following actions while cleaning the data:

1. Get rid of redundant rows.
2. Taking care of missing values.
3. Change the column data types to the proper ones.
4. Including significant columns.

Data Summary:

Hotel- The category of resort ,which is Resort and City hotel.

Is_cancelled- The value of column show the cancellation type. If the booking was cancelled or not. Values[0,1],where 0 indicates not cancelled.

Stayed_in_weekend_nights- The number of weekend nights stay per reservation.

Stayed_in_weekday_nights- The number of weekday nights stay per reservation.

Meal- Meal preferences per reservation [BB, FB, HB, SC, Undefined].

Country- The origin country of guest.

Data Summary:

Market_segment- This column show how reservation was made and what is the purpose of reservation. Eg. corporate means corporate trip, TA for travel agent, TO for Tour Operator.

Distribution_channel- The median through which booking is made [Direct, TA/TO, corporate, undefined, GDS].

Is_repeated_channel- Shows if the guest is who has arrived earlier or not. Values[0,1]→0 indicates no and 1 indicated yes person is repeated guest.

Days_in_waiting_list- Number of days between actual booking and transact.

Customer_type- Type of customers(Transient, group, etc)

Libraries use:

- import numpy as np
- import pandas as pd
- import matplotlib.pyplot as plt
- import seaborn as sns
- import missingno as msno
- import folium
- from folium.plugins import HeatMap
- import plotly.express as pltx

Data Wrangling:

Shape of dataset-

```
Df.shape
```

```
(119390, 32)
```

Data Set Information-

```
 #   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)
```

Data Wrangling(contd.):

Finding Null Values

```
hotel                               0
is_canceled                         0
lead_time                           0
arrival_date_year                   0
arrival_date_month                  0
arrival_date_week_number            0
arrival_date_day_of_month           0
stays_in_weekend_nights             0
stays_in_week_nights                0
adults                              0
children                            4
babies                              0
meal                                0
country                           488
market_segment                      0
distribution_channel                0
is_repeated_guest                   0
previous_cancellations              0
previous_bookings_not_canceled      0
reserved_room_type                  0
assigned_room_type                  0
booking_changes                     0
deposit_type                        0
agent                           16340
company                        112593
days_in_waiting_list                0
customer_type                       0
adr                                 0
required_car_parking_spaces         0
total_of_special_requests           0
reservation_status                  0
reservation_status_date             0
dtype: int64
```

Data Wrangling(contd.):

```
Df1=Df.fillna(value=0)
```

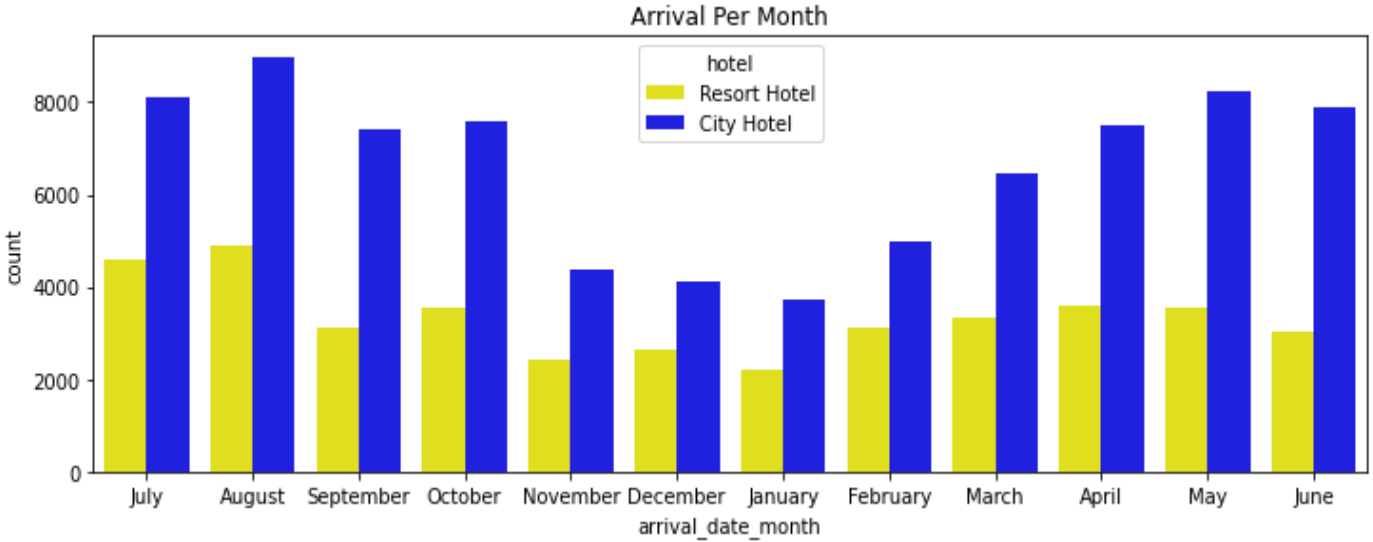
Removing Null Values →

```
hotel                           0
is_canceled                     0
lead_time                       0
arrival_date_year               0
arrival_date_month              0
arrival_date_week_number        0
arrival_date_day_of_month       0
stays_in_weekend_nights         0
stays_in_week_nights            0
adults                          0
children                        0
babies                          0
meal                            0
country                         0
market_segment                  0
distribution_channel            0
is_repeated_guest               0
previous_cancellations          0
previous_bookings_not_canceled  0
reserved_room_type              0
assigned_room_type              0
booking_changes                 0
deposit_type                    0
agent                           0
company                         0
days_in_waiting_list            0
customer_type                   0
adr                             0
required_car_parking_spaces     0
total_of_special_requests       0
reservation_status              0
reservation_status_date         0
dtype: int64
```

Exploratory data analysis:

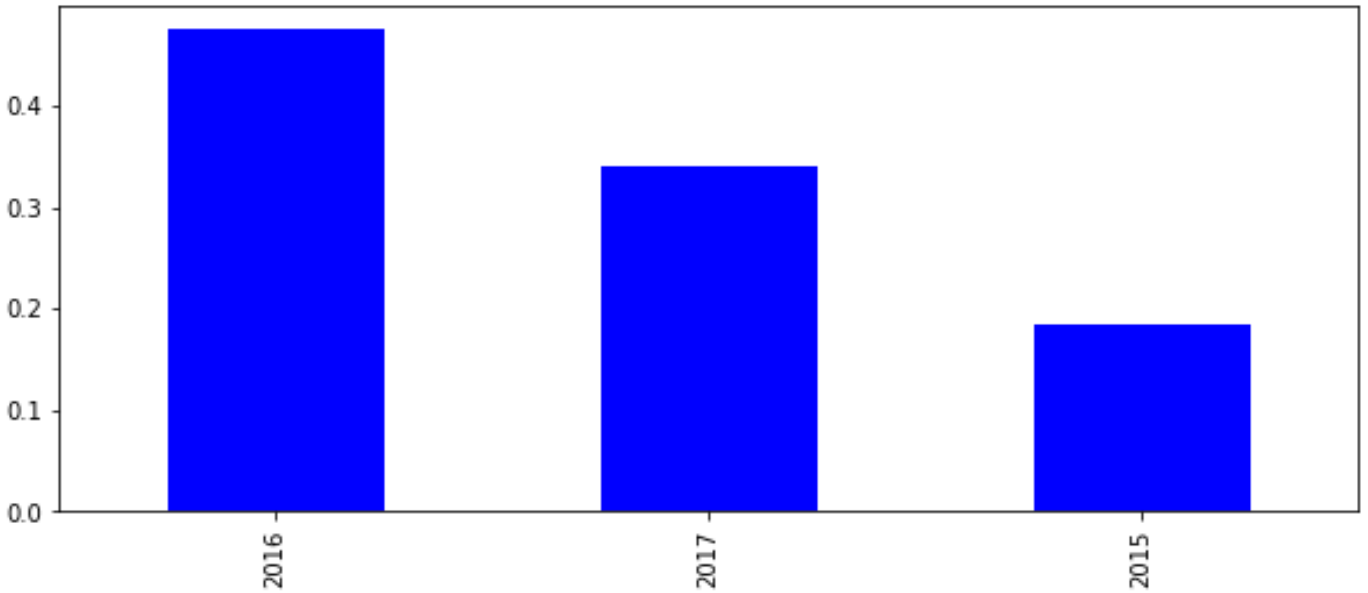
- What is the month with the most guest arrivals?

August	13877
July	12661
May	11791
October	11160
April	11089
June	10939
September	10508
March	9794
February	8068
November	6794
December	6780
January	5929



August and July are the months with the highest number of arrivals for both hotels.

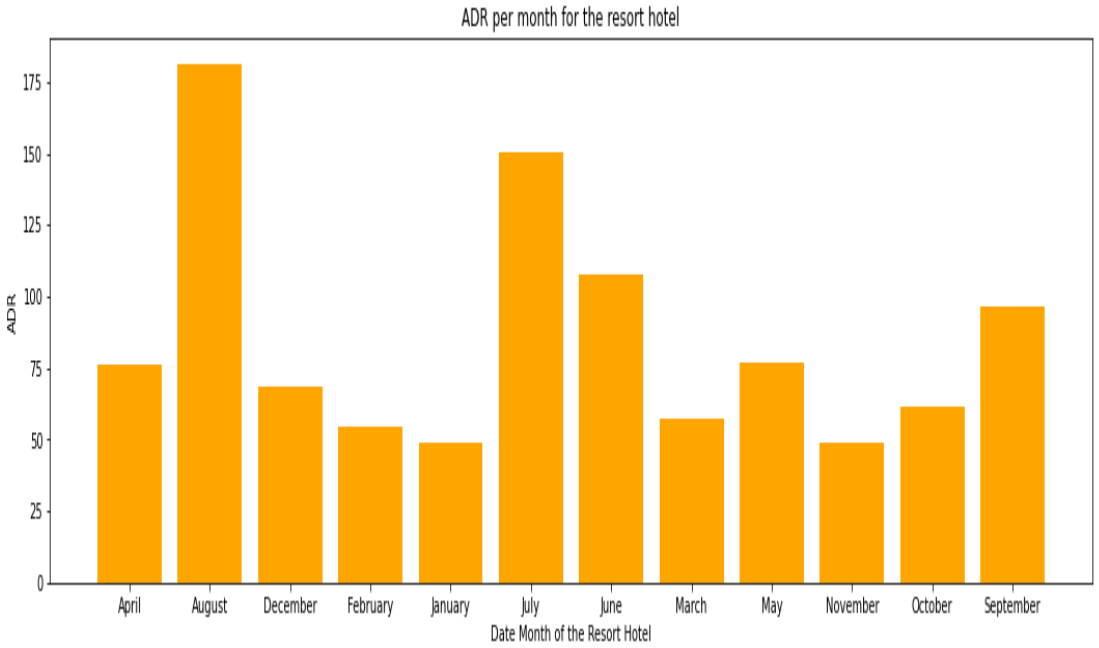
➢ What is the year with the most guest arrivals?



2016	56707
2017	40687
2015	21996

2016 is the year in which most of the guest have visited

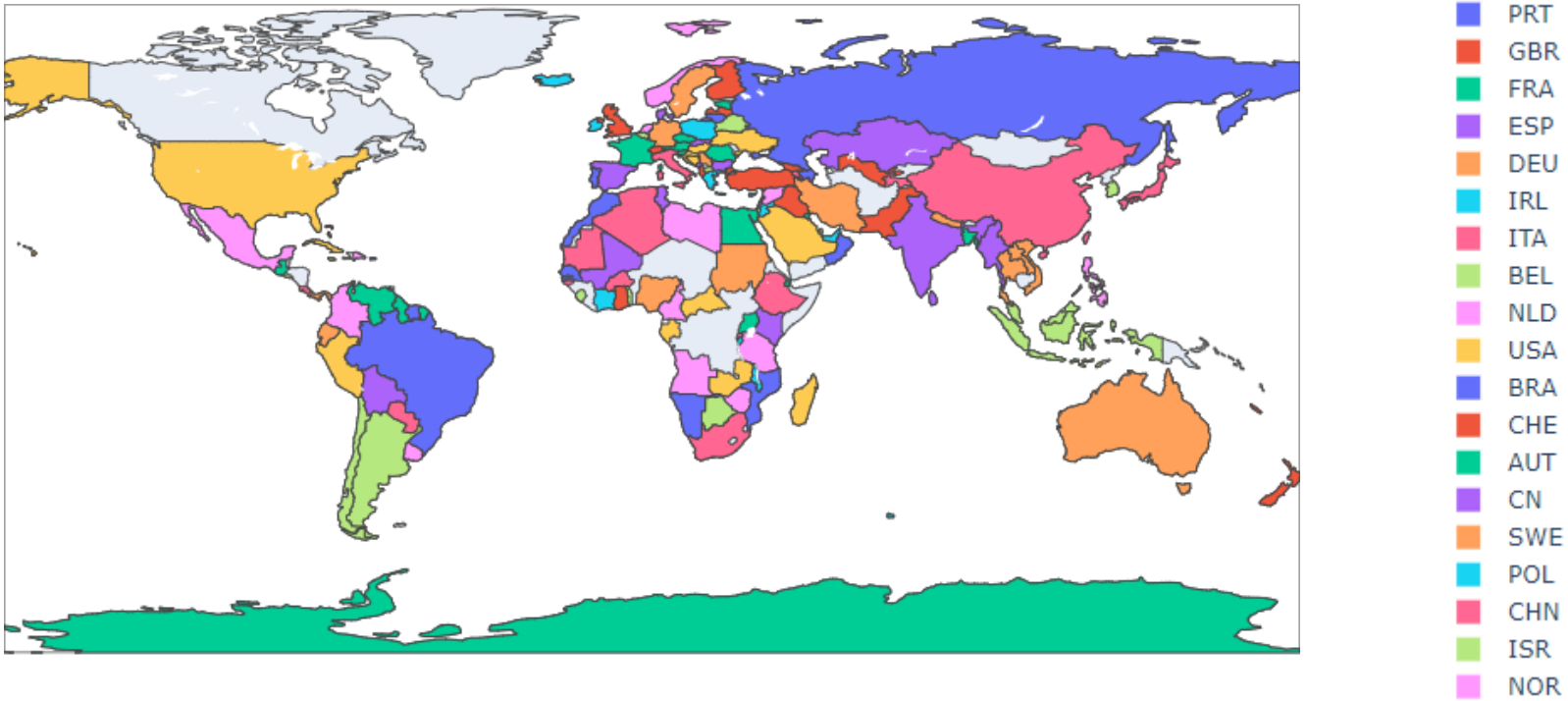
- How does the price vary per night over the year?



	arrival_date_month	adr
0	April	75.867816
1	August	181.205892
2	December	68.322236
3	February	54.147478
4	January	48.708919
5	July	150.122528
6	June	107.921869
7	March	57.012487
8	May	76.657558
9	November	48.681640
10	October	61.727505
11	September	96.416860

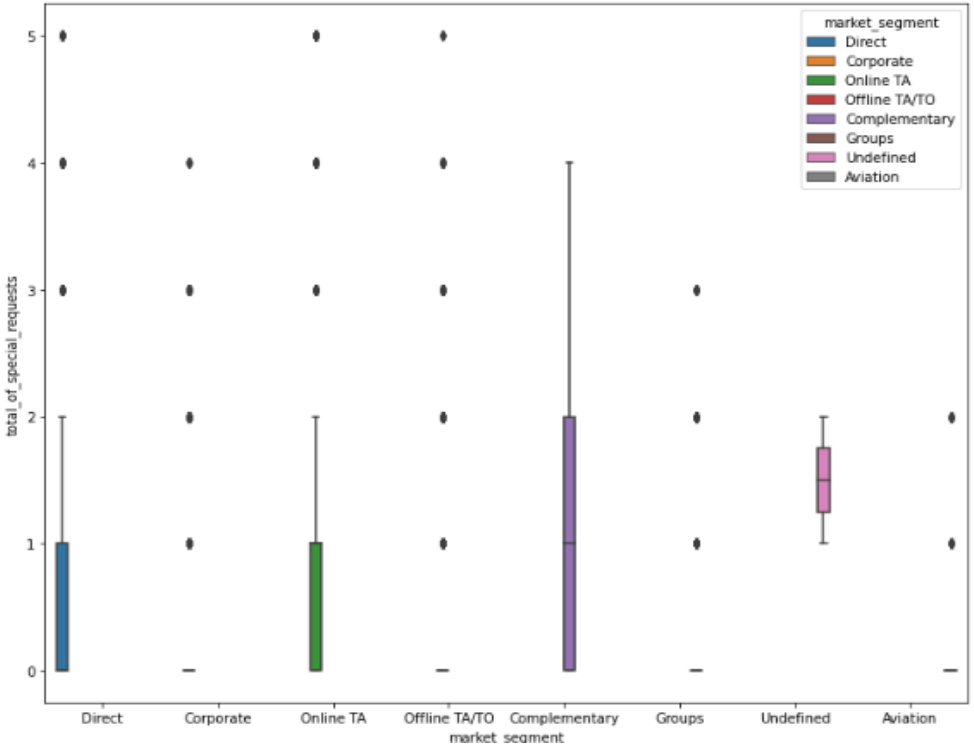
August is having a highest average distributed rate followed by July and June.

➢ Which countries do customers come from?



Most customers come from Europe, mainly from Portugal and neighboring

What is the strongest market segment and distribution channel?



	index	market_segment
0	Online TA	56477
1	Offline TA/TO	24219
2	Groups	19811
3	Direct	12606
4	Corporate	5295
5	Complementary	743
6	Aviation	237
7	Undefined	2

Conclusion:

- The majority of the reservations are for city hotels. There is no doubt that those hotels require the most targeted spending.
- We also understand that there may be high no deposit policies to blame for the high cancellation rate. We should also aim for the months of May through August. Due to the summer season, those are the busiest months.
- Western Europe is where the majority of the visitors are from. We ought to allocate a large portion of our money to those areas.
- Given that we don't have many repeat customers, we should focus our advertising on them to obtain more of them.
- TA and TO have emerged as the most powerful market segments and distribution channels, followed by the direct channel with the hotel. In the final scenario, a special offer could be used to encourage the use of this channel.

THANK YOU