Capstone Project-1

Hotel Booking Analysis

By

Akshay Nikam

Points for Discussion

Problem Statement

- This project contains the real world data record of hotel bookings of a city and the resort hotel containing booking details made by customers from 2015 to 2017.
- Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? This hotel booking dataset can help you explore those questions!
- We are going to explore and analyse the data to discover important factors that govern the bookings.

Work Flow

We Will divide our work flow into three following steps-

Data Collection and understanding

After collecting data it's very important to understand your data. So we had hotel Booking analysis data. Which had 119390 rows and 32 columns. So let's understand this 32 columns.

Data Description:

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', 'agent',
       'company', 'days_in_waiting_list', 'customer_type', 'adr',
       'required_car_parking_spaces', 'total_of_special_requests',
       'reservation_status', 'reservation_status_date'],
      dtype='object')
```

Data Collection and Understanding

Data Description:

Basic Summary

```
<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
```

Data Cleaning and Manipulation

Missing value handling : There were 4 columns company, agent, country and children with missing values. I have fill these missing values with appropriate values.

```
# Let see which of the columns ar having missing values in it.
hotel_booking_data.isnull().sum().sort_values(ascending=False)[:5]
```

```
company                 112593
agent                    16340
country                    488
children                     4
reserved_room_type           0
dtype: int64
```

```
# And column 'Country' null values with 'others'
hotel_booking_data['country']=hotel_booking_data['country'].fillna('other')

# Lets fill columns 'Agent', 'Company', 'Children' with '0'
hotel_booking_data[['agent','company','children']]=hotel_booking_data[['agent','company','children']].fillna(0)
```

```
# As we can see we have succsessfully fill the null values with required values.
hotel_booking_data.isnull().sum().sort_values(ascending=False)[:5]
```

```
hotel                           0
is_canceled                     0
reservation_status              0
total_of_special_requests       0
required_car_parking_spaces     0
dtype: int64
```

Data Cleaning and Manipulation

Handling Duplicates: Data had 31994 duplicates values. So we dropped it from

```
# Now lets see if any duplicate values are present in our DataFrame.
hotel_booking_data.duplicated().value_counts()
```

```
False    87396
True     31994
dtype: int64
```

```
# As we can see that there are 31994 duplicate row values, so lets drop these duplicate values.
hotel_booking_data.drop_duplicates(inplace=True)
```

```
# Now check again if the duplicate values droped or not
hotel_booking_data.duplicated().value_counts()
```

```
False    87396
dtype: int64
```

Data Cleaning and Manipulation

Manipulation: 1) Two new columns created :
'Total_People' = from the Children, adults, babies, 'Total_stay' = From weekend nights and weekdays night.

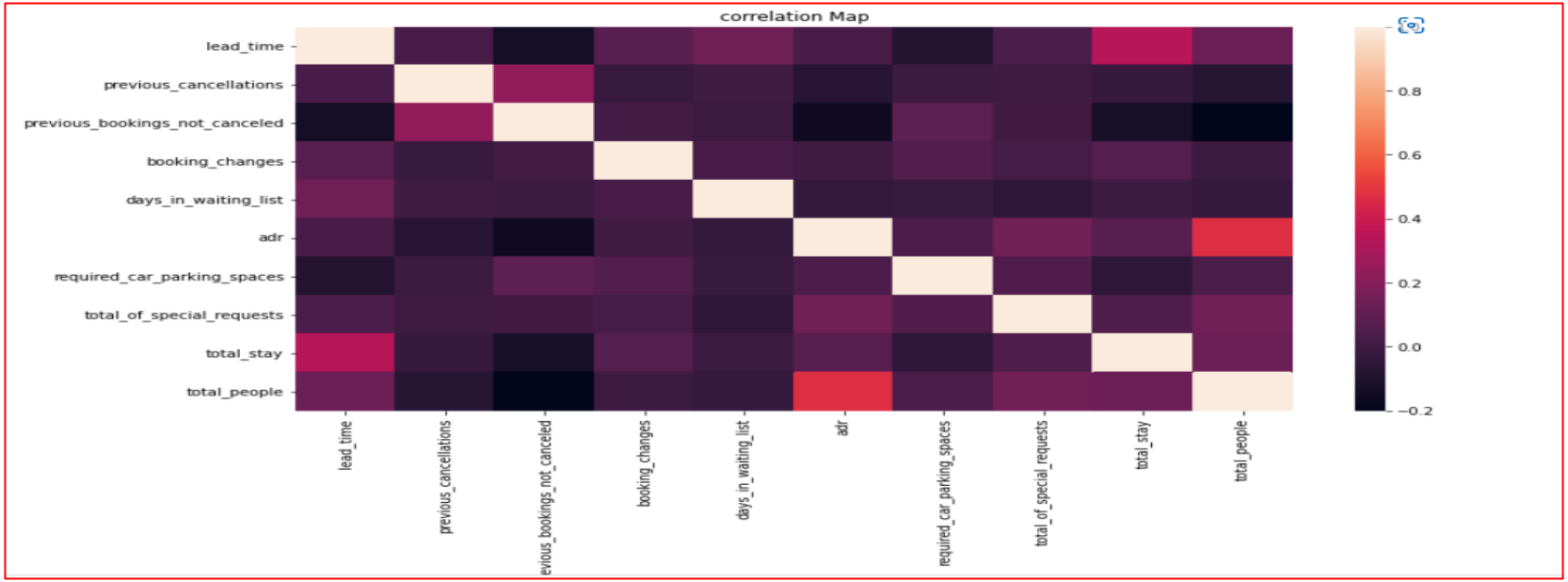
```
# Lets make new column by adding 'children' and 'babies for our need.
hotel_booking_data['kids'] = hotel_booking_data.children + hotel_booking_data.babies
# Combine total members by adding kids and adults this will give us total member per booking.
hotel_booking_data['total_members'] = hotel_booking_data.kids + hotel_booking_data.adults
```

2) Datatype conversion : Here also I have converted data type of following columns into string
And 'Arrival_date' column to datetime format

```
#convert the datatypes of required columns to string.
hotel_booking_data['arrival_date_year'] = hotel_booking_data['arrival_date_year'].astype('str')
hotel_booking_data['arrival_date_month'] = hotel_booking_data['arrival_date_month'].astype('str')
hotel_booking_data['arrival_date_day_of_month'] = hotel_booking_data['arrival_date_day_of_month'].astype('str')
hotel_booking_data['is_canceled'] = hotel_booking_data['is_canceled'].astype('int')
hotel_booking_data['is_repeated_guest'] = hotel_booking_data['is_repeated_guest'].astype('str')
```

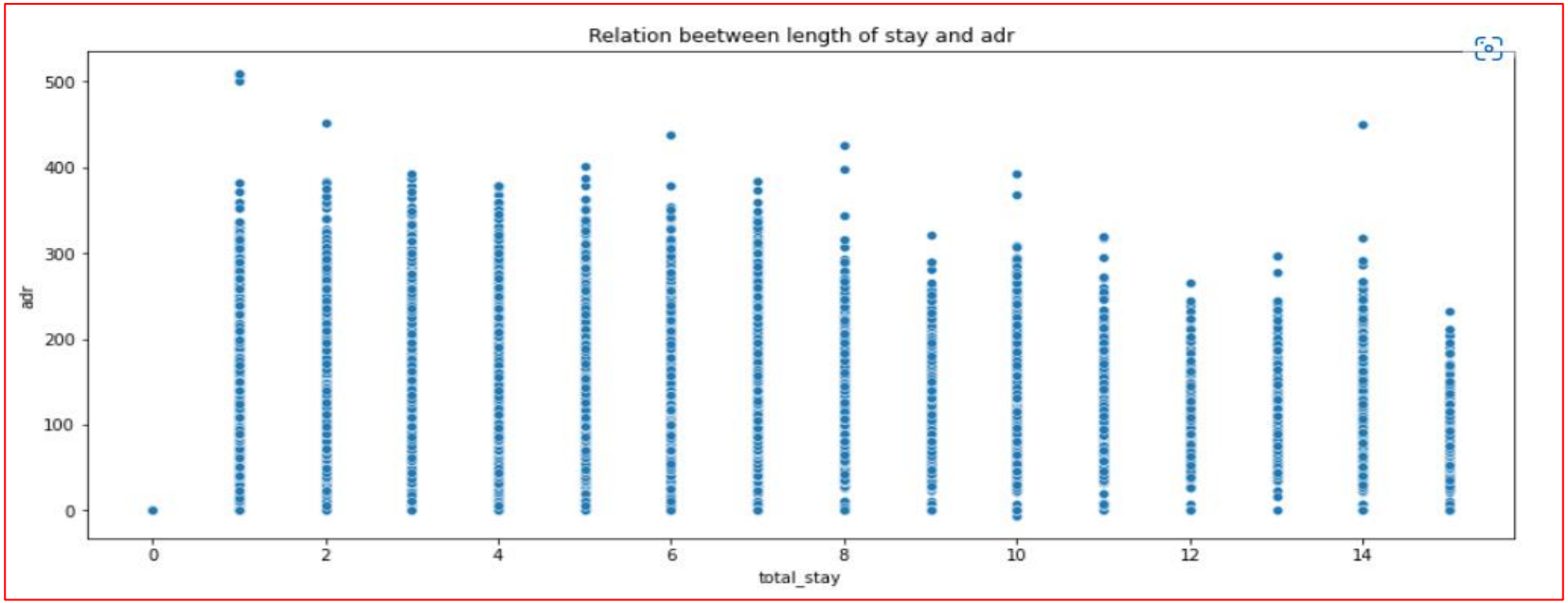
```
# Lets import datetime to convert arrival time column to datetime datatype.
from datetime import datetime
from datetime import date
# Lets add new column contains date of arrival of customers and convert it to datetime datatype.
hotel_booking_data['arrival_date'] = hotel_booking_data['arrival_date_day_of_month'] + '-' + hotel_booking_data['arrival_date_month'] + '-' + hotel_booking_data['arrival_date_year']
hotel_booking_data['arrival_date'] = hotel_booking_data['arrival_date'].apply(lambda x: datetime.strptime(x, '%d-%B-%Y'))
```

EDA



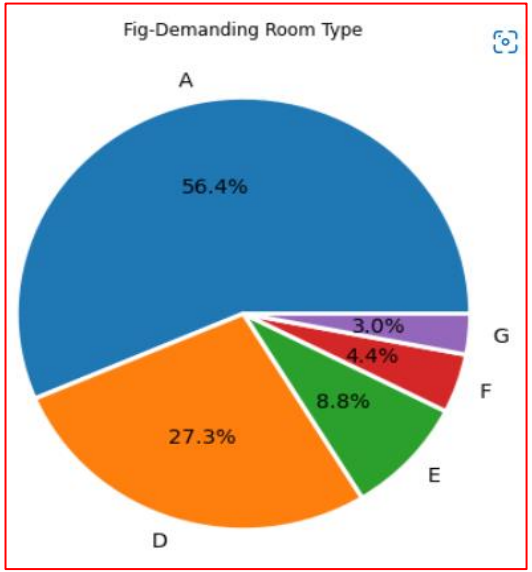
- Total stay length and lead time have slight correlation. This may means that for longer hotel stays people generally plan little before the actual arrival.
- ADR is slightly correlated with total people, which makes sense as more no. of people means more revenue, therefore more ADR.

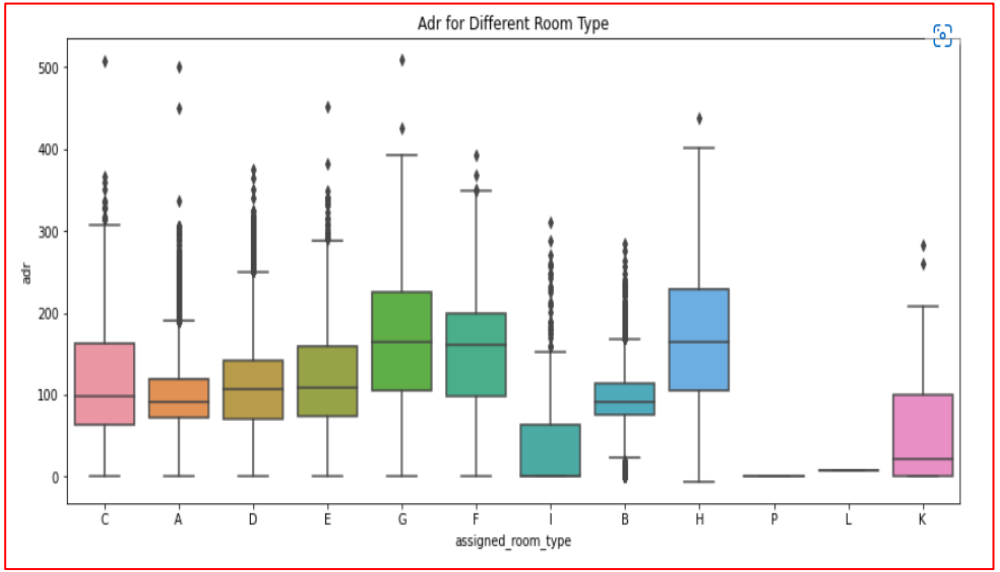
EDA



- From the scatter plot we can see that as length of total stay increases the ADR decreases. This means for longer stay, the better deal for customer can be finalized.

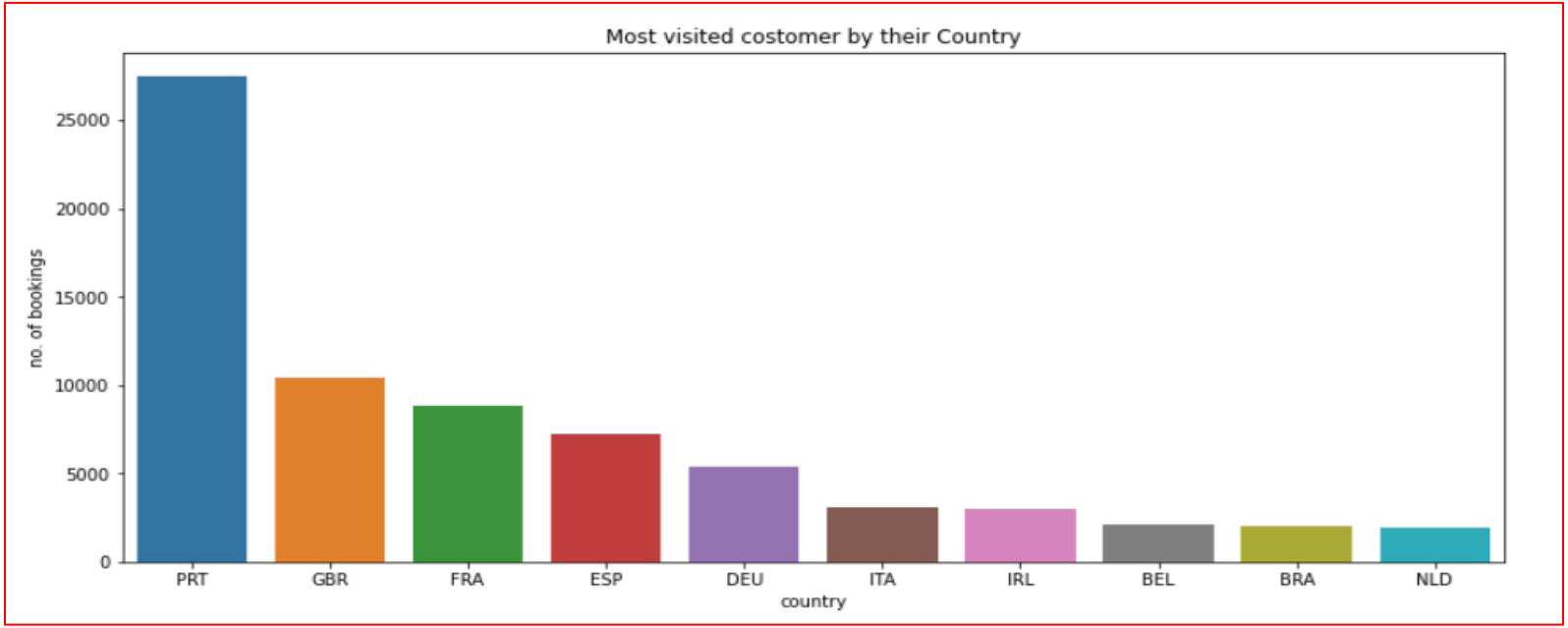
EDA





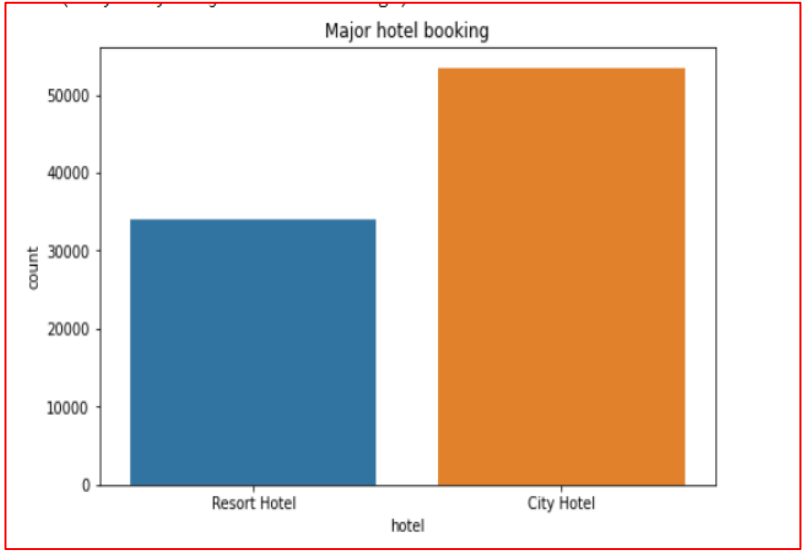
- From above plot we can see that the Most demanded room type is A as most people prefer it, but from box plot we can see that the better ADR rooms for hotels are of type H, G and F also. Hotels should increase the no. of room types A and H to maximize revenue.

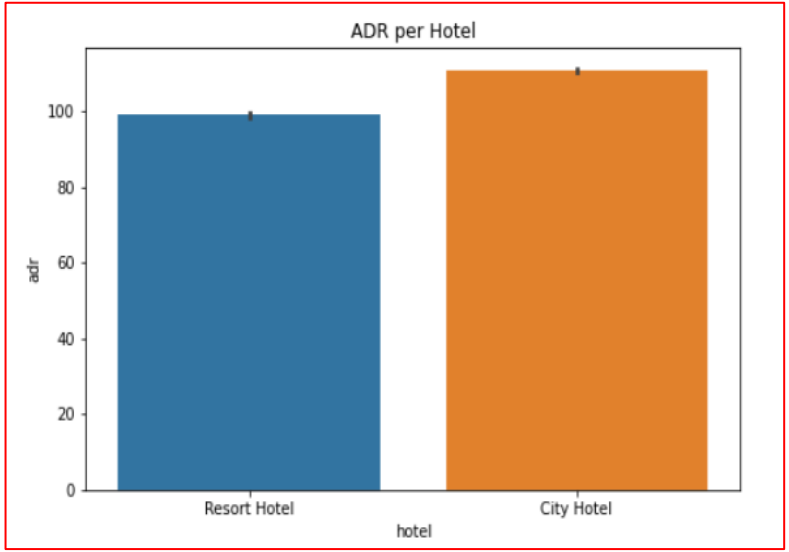
EDA



- From above plot we can see that. Most of the customers come from European countries such as Portugal, Great Britain, France and Spain

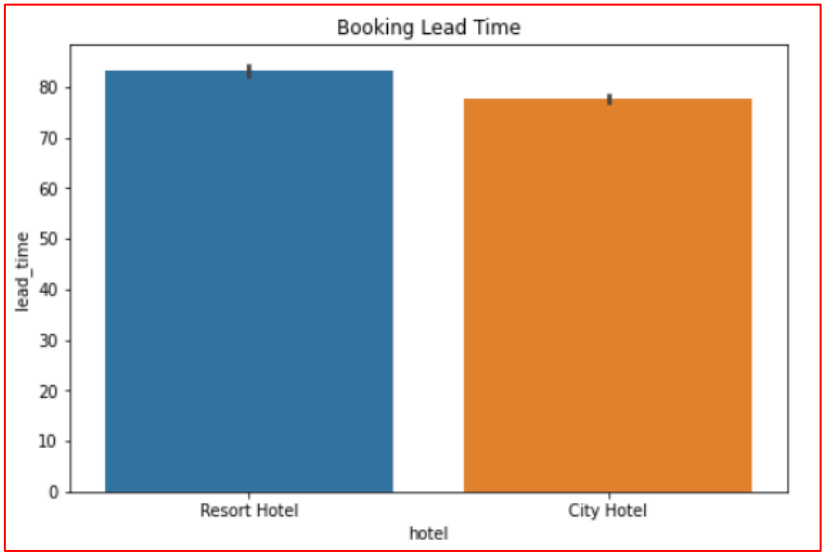
EDA

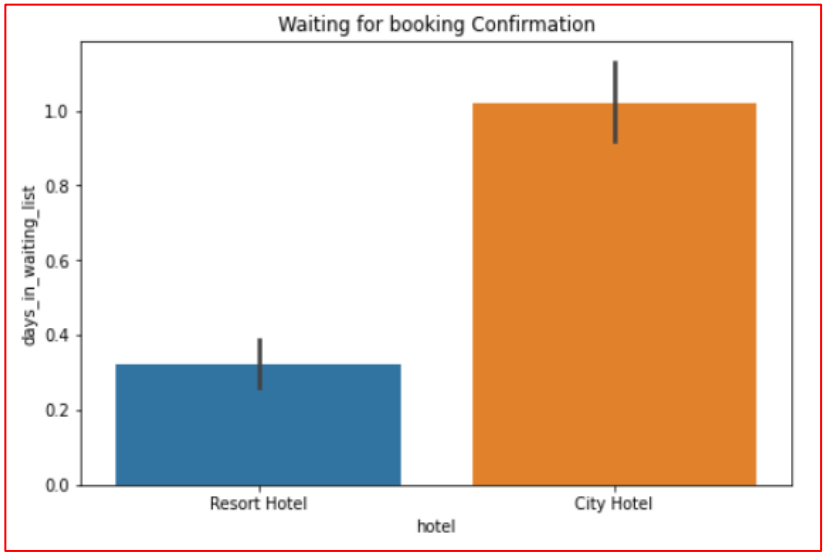




- From above plot we can say that city hotel booking is more than resort hotel bookings.
- ADR(average daily rate) of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue than resort hotels.

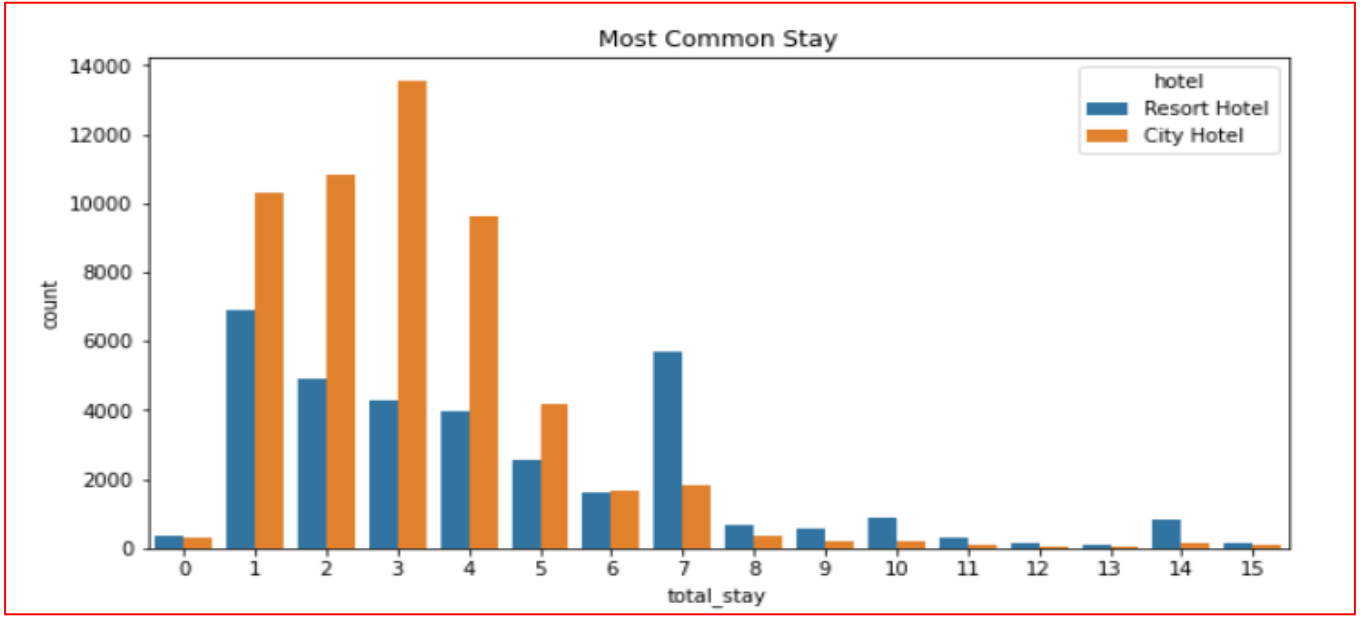
EDA





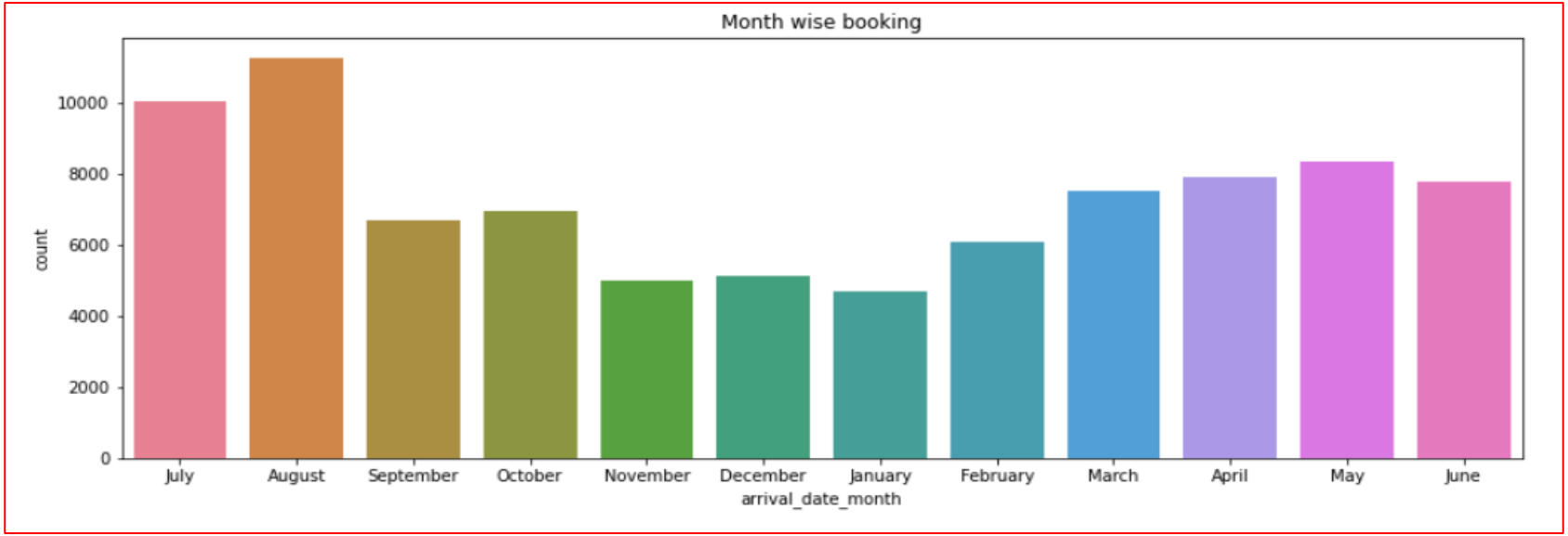
- Resort hotel has slightly higher lead time that means people used to book resort hotel in advance. Also one thing to notice here is lead time is significantly higher in each case, this means customers generally plan their hotel visits way to early.
- City hotel has significantly longer waiting time due to because it is having the high rush of customers, hence City Hotel is much busier than Resort Hotel.

EDA



- Most common stay length is less than 4 days and generally people prefer City hotel for short stay, but for long stays, Resort Hotel is preferred.

EDA

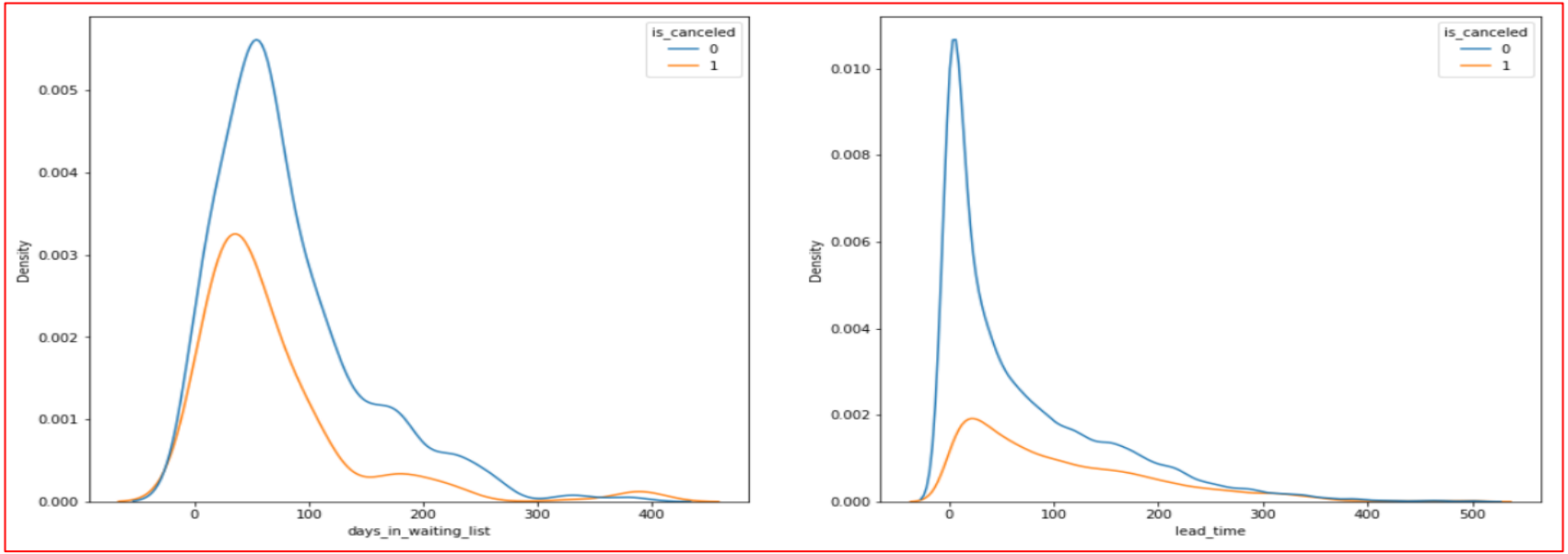


- From the graph, we can see July and August is the most occupied (busiest) month in the year and November, December and January is the most unoccupied month in the year. Also we can say that during these months (November, December and January) customer could get great offers and discounts on bookings

EDA

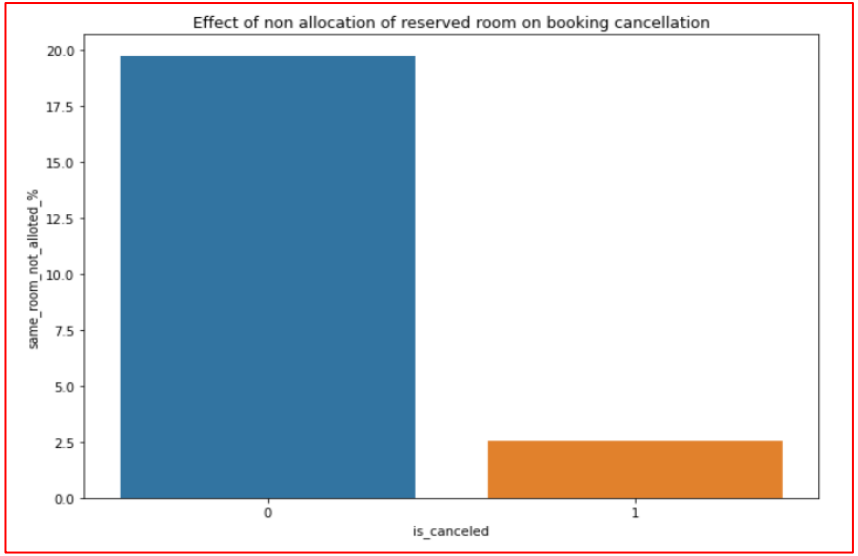
- In comparison city hotel bookings got cancelled more than resort hotel.

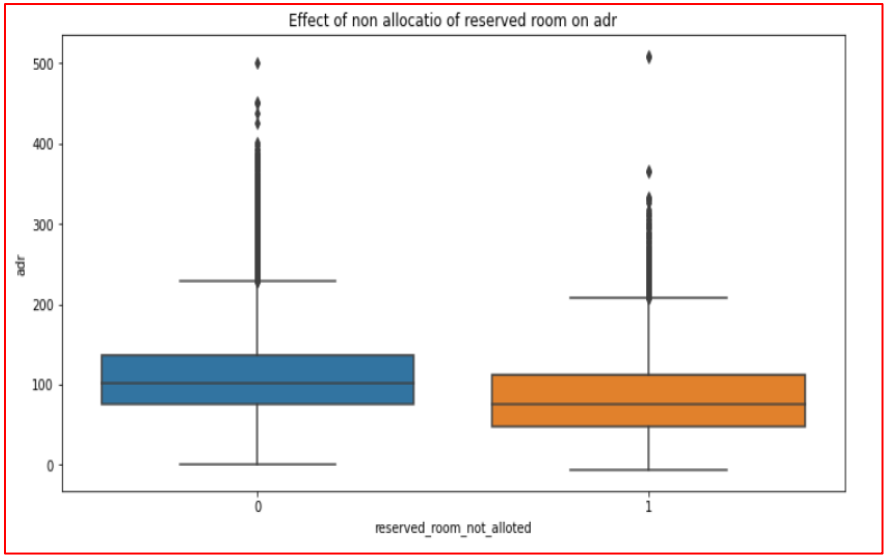
EDA



- We see that most of the bookings that are cancelled have waiting period of less 150 days but also most of bookings that are not cancelled also have waiting period less than 150 days. Hence this shows that waiting period has no effect on cancellation of bookings. Also, lead time has no affect on cancellation of bookings, as both curves of cancelation and not cancelation are similar for lead time too.

EDA





- We see that not getting same room as demanded is not the case of cancellation of rooms. A significant percentage of bookings are not cancelled even after getting different room as demanded.
- So not getting same room do affects the adr, people who didn't got same room have paid a little lower adr, except for few exceptions.

Conclusion

Conclusion :

- Most demanded room type is A as most people prefer it, but the better ADR rooms for hotels are of type H, G and F. Hotels should increase the no. of room types A and H to maximize revenue.
- Most of the customers come from European countries such as Portugal, Great Britain, France and Spain Hotel businesses can attract more European travelers.
- City hotels are in high demand as the majority of reservations are for city hotels this high demand makes them costlier and busier than the resort hotels.
- People who planned long stay or vacation (more than 5 days) prefers resort hotels and people prefer city hotel majorly for short stays (less than 4 days).
- Peoples more often go out and book hotels in Summer and Rainy seasons these lead the maximum ADR in these seasons. If you want better offers and discount try to book hotel in winter season. Hopefully it will work!.