

# CAPSTONE PROJECT - 1

## EDA ON HOTEL BOOKING ANALYSIS

BY -  
RISHANSHU YADAV



## ✓ WORK FLOW :



- EDA will be divided into the following three analyses –
  1. Univariate analysis : Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable.
  2. Bivariate analysis : Bivariate analysis is where you are comparing two variables to study their relationships.
  3. Multivariate analysis : Multivariate analysis is similar to bivariate analysis but you are comparing more than two variables.

# ❖ Problem Statement :

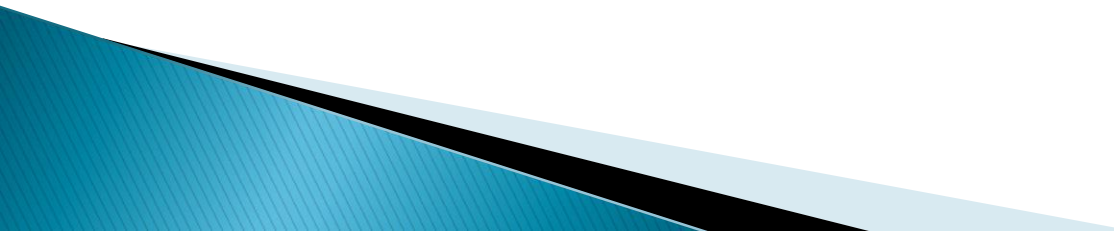
- For this project we will be analyzing Hotel Booking data . This data set contains booking information for a city hotel and a resort hotel , and includes information such as when a booking was made, length of stay , the number of adults , children , and/or babies and the number of available parking spaces .
- Hotel industry is a very volatile industry and the bookings depends on above factors and many more .
- The main objective behind this project is to explore and analyze data to discover important factors that concern the bookings and give insights to hotel management , which can perform various campaigns to boost the business and performance.

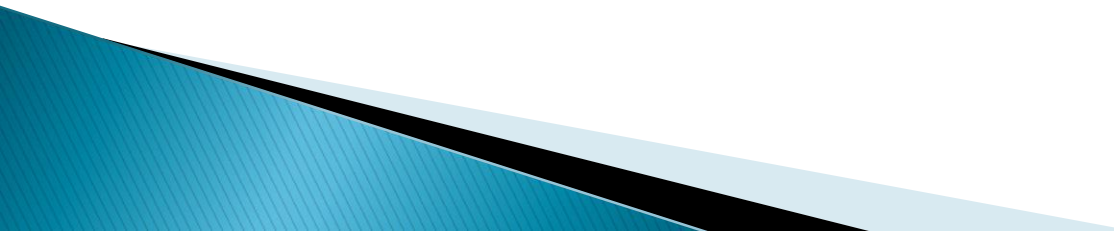
## ❖ DATA COLLECTION AND UNDERSTANDINGS :

–We got here the data. This data contain 119390 rows and 32 columns .  
So its becomes important to understand the columns.

### ❑ Data Description:

- **Hotel** : Type of hotel ( City hotel or Resort hotel )
- **is\_canceled** : Value indicating if the booking was canceled (1) or not (0)
- **lead\_time** : Number of days that elapsed between the entering date of the booking and the arrival date
- **arrival\_date\_year** : Year of arrival date
- **arrival\_date\_month** : Month of arrival date
- **arrival\_date\_week\_number** : Week number of year for arrival date
- **arrival\_date\_day\_of\_month** : Day of arrival date
- **stays\_in\_weekend\_nights** : Number of weekend nights( Saturday or sunday spent at the hotel by guests.
- **stays\_in\_week\_nights** : Number of week nights( Monday to Friday )
- **adults** : Number of adults among guests
- **children** : Number of children among guests

- **babies** : Number of babies
  - **meal** : Type of meal booked.
  - **country** : Country of origin.
  - **market\_segment** : Market segment designation. (TA/TO)
  - **distribution\_channel**: Booking distribution channel.(T/A/TO)
  - **is\_repeated\_guest** : is a repeated guest (1) or not (0)
  - **previous\_cancellations** : Number of previous bookings that were cancelled by the customer prior to the current booking
  - **previous\_bookings\_not\_canceled** : Number of previous bookings not cancelled by the customer prior to the current booking
  - **reserved\_room\_type** : Code of room type reserved.
  - **assigned\_room\_type** : Code for the type of room assigned to the booking.
  - **booking\_changes** : Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
  - **deposit\_type** : No Deposit, Non Refund , Refundable.
  - **agent** : ID of the travel agency that made the booking
- 

- **company** : ID of the company/entity that made the booking .
  - **days\_in\_waiting\_list** : Number of days the booking was in the waiting list before it was confirmed to the customer
  - **customer\_type** : type of customer. Contract,Group,transient,Transient party.
  - **adr** : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
  - **required\_car\_parking\_spaces** : Number of car parking spaces required by the customer
  - **total\_of\_special\_requests** : Number of special requests made by the customer ( e.g. twin bed or high floor)
  - **reservation\_status** : Reservation status (Canceled , check-out or not show )
  - **reservation\_status\_date** : Date at which the last reservation status was uploded.
- 



# ❖ Data Cleaning and Manipulation

➤ After importing required library and modules we move for data cleaning , manipulation and visualization.

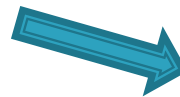
From the given data set first we will see , any missing values (null values) and duplicate value and we need to remove them.

- There were 4 columns company , agent , country , and children with missing values

```
# Checking for null values.
df.isnull().sum().sort_values(ascending=False)[:7]
```

company	112593
agent	16340
country	488
children	4
reserved_room_type	0
assigned_room_type	0
booking_changes	0

dtype: int64



```
#Because of the null value we drop the company and agent columns
#for our simplicity in further operations.
df = df.drop('company' , axis = 1)
```

```
[27] df = df.drop('agent',axis = 1)
```

```
df.isnull().sum().sort_values(ascending=False)[:7]
```

country	488
children	4
hotel	0
is_repeated_guest	0
reservation_status	0
total_of_special_requests	0
required_car_parking_spaces	0

dtype: int64

- You must be wondering why I not remove the country and children column also but that column has less null values comparing to other columns .So , for those remaining null values I choose to remove the null rows not the whole column for the future operations .

```
# to drop null values|
df = df.dropna()

[30] # LETS CHECK THAT NULL VALUES ARE GONE FROM THE THE MAIN DATA SET OR NOT
df.isnull().sum().sort_values(ascending=False)[:7]

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

➤ Handling Duplicates : data had 31984 duplicates values.

```
[ ] # Now in data set there are going to have some duplicate values also lets see that.

Duplicates_in_data = sum(df.duplicated(keep = 'first'))
Duplicates_in_data
```

31984

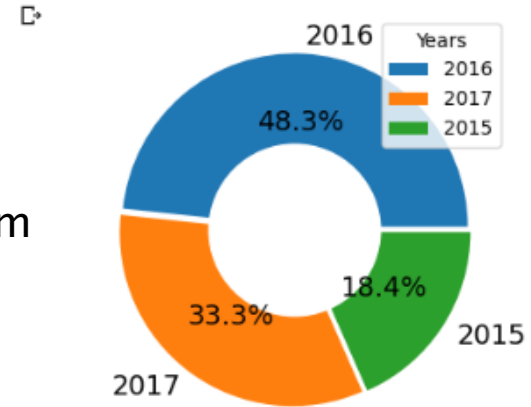
```
[50] #here we drop all duplicates values from our data set.
new_data = df.drop_duplicates()
```



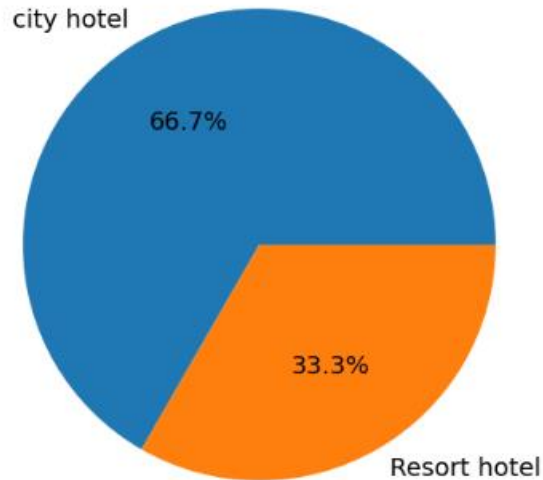
# ❖ EXPLORATORY DATA ANALYSIS ( EDA ) :

## ❑ Exploring the data set

- From the first pie chart we can see we the data set contain three years data. The maximum data is from year 2016.



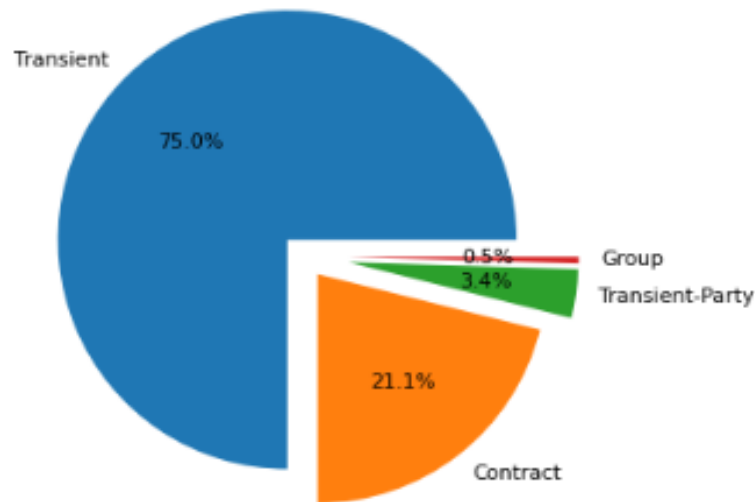
(i)



- From the second pie chart we can see we have main two types of hotel i.e. City hotel and Resort Hotel.
- The maximum data is form City hotel i.e. 66.7% and minimum from Resort hotel i.e. 33.3%.

(ii)

❖ Checking the how many different types of customer are these hotels are having

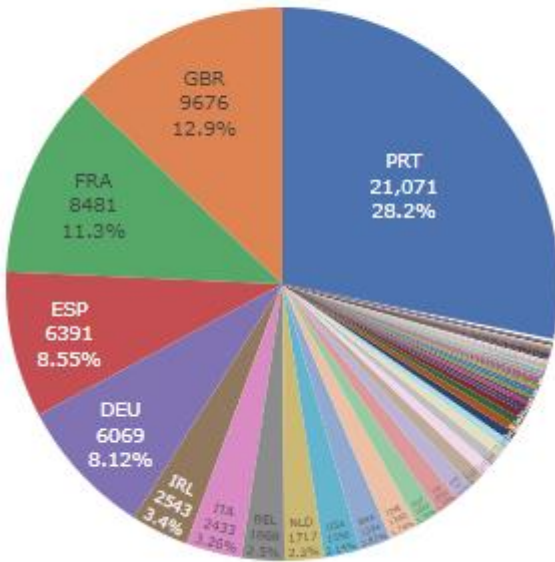


#### ❑ CONCLUSION :

➤ From above chart we can see that 75.0% , 21.1 % , 3.4 % , 0.5 % business are coming from transient, contract, transient-party and group respectively. And here we only considered the conform bookings.

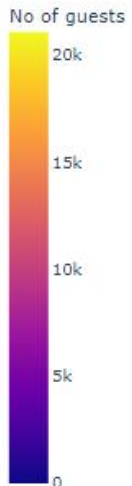
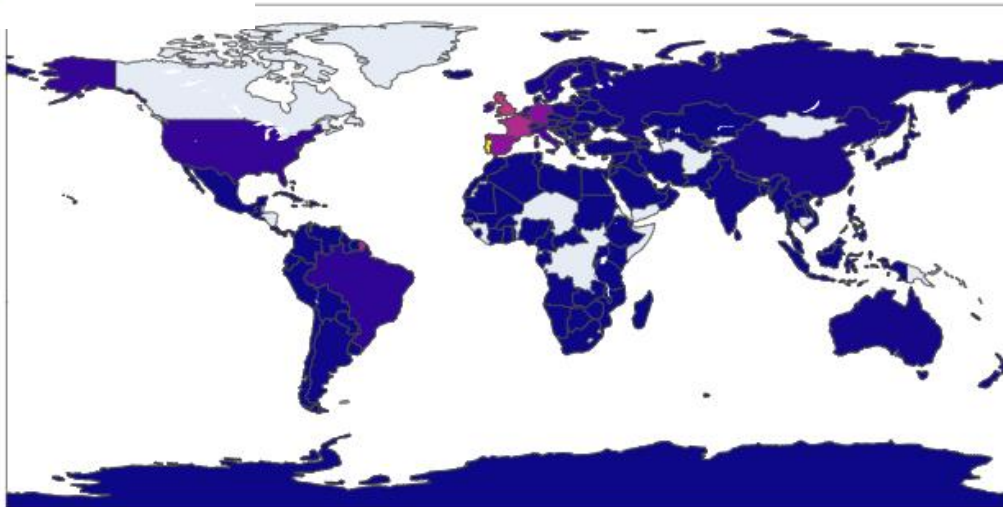
## ❖To analyze home country of guests :

Home country of guests

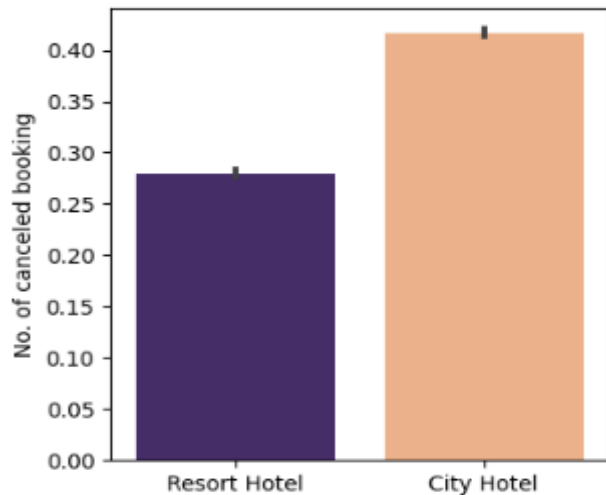


### ❑CONCLUSOION :

- As we see in the side pie chart the maximum number of guests are from Portugal with 21,071 guests, Great Britain with 9,676 guest and France with 8,481 and so on.
- Most tourists come from Europe since the top 5 countries are in Europe.
- And in the next map we expressing the geographical data in much more efficient way.

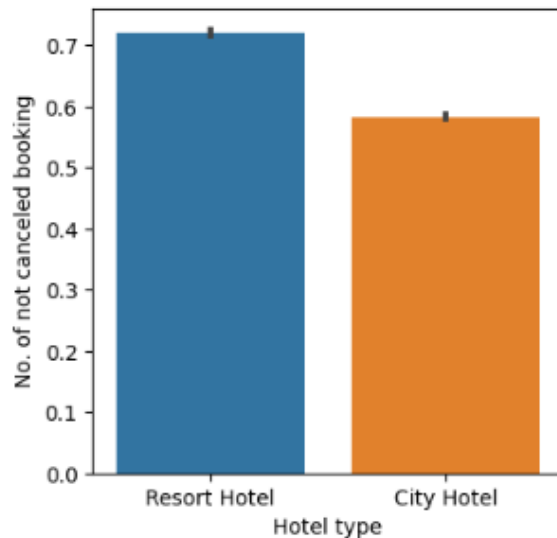


❖ To analyze which hotel gets cancelled the most by the customer :



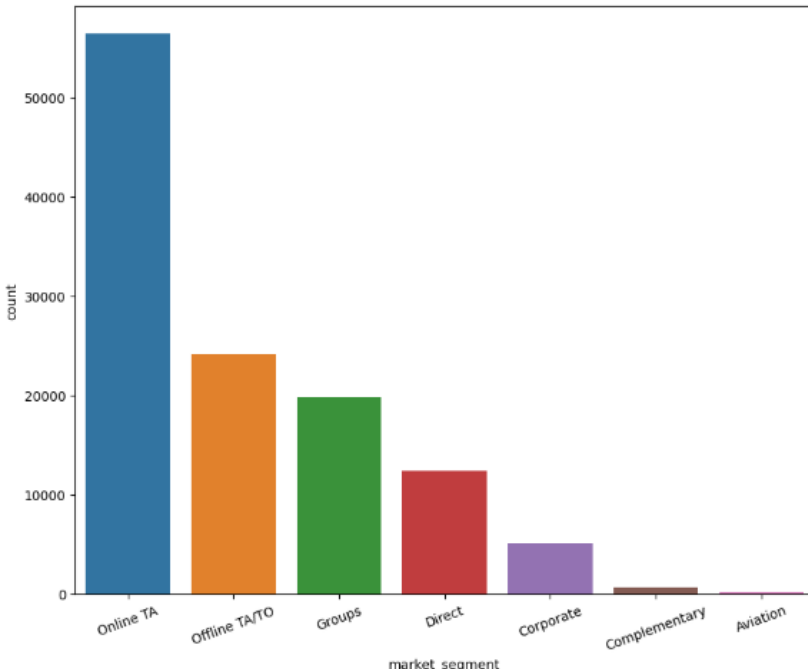
□ OBSERVATIONS :

➤ From the first bar plot we can see the maximum number of bookings are cancelled are from the City hotel that indicates people love to spend their time in peaceful Resort hotel.



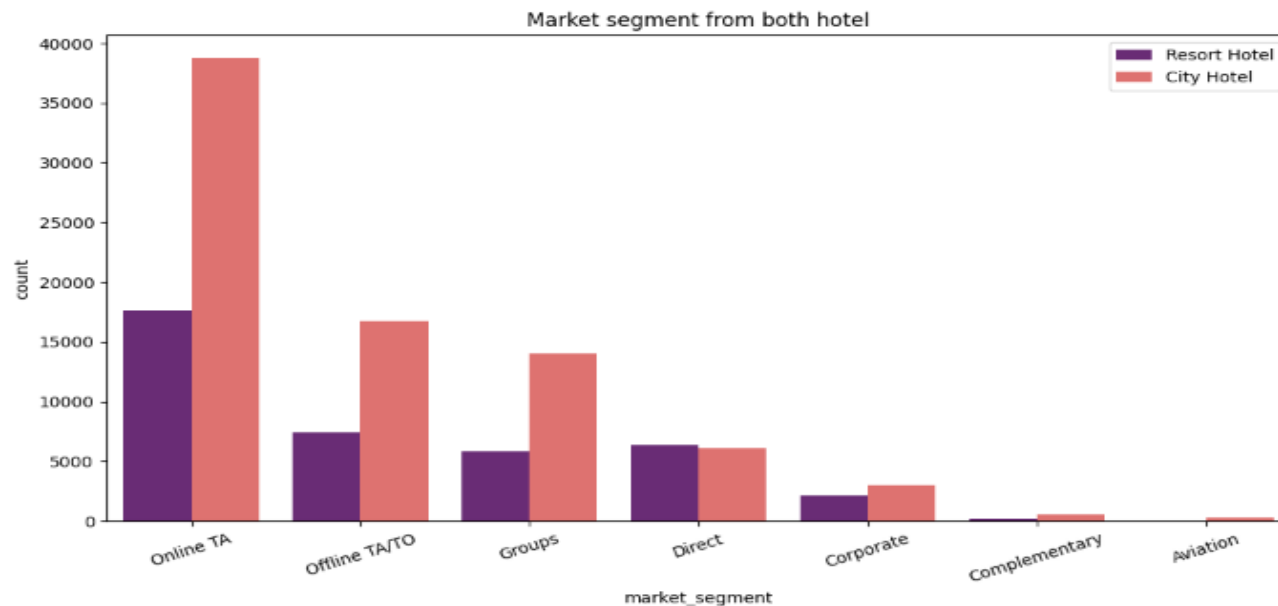
➤ From the Second bar plot we can see the maximum number of booking not cancelled are from Resort hotel which proves our upper statement.

## ❖ Booking by market segment

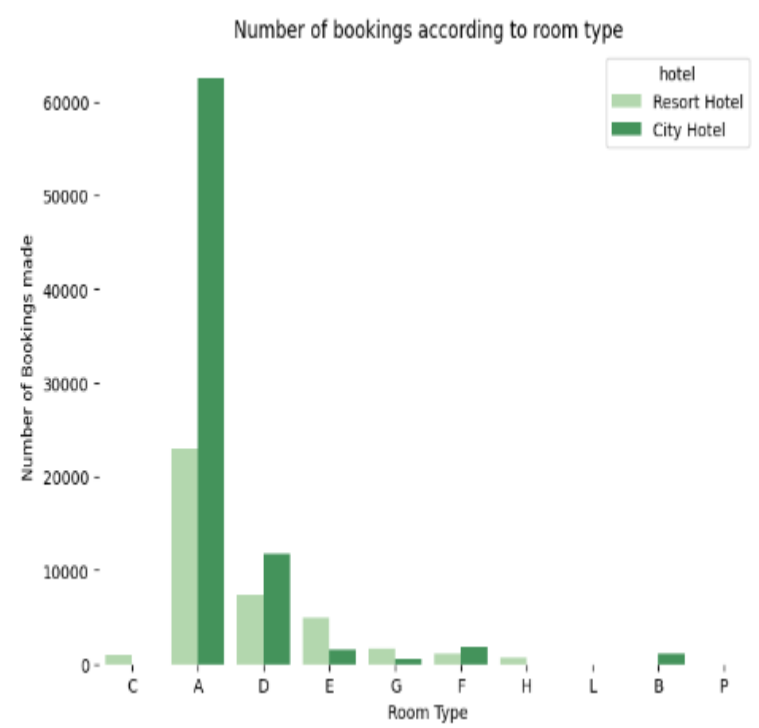
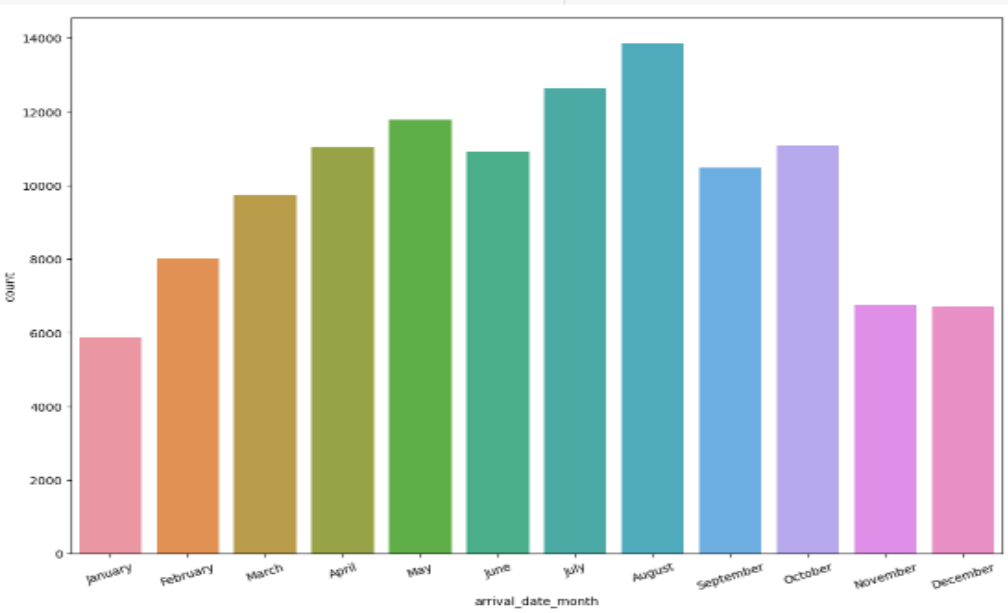


### ❑ OBSERVATIONS :

- From above two chart we can see that the most of the booking are done by online travel agencies and the second most segment is offline travel agencies.
- And the minimum bookings are coming for aviation segment.



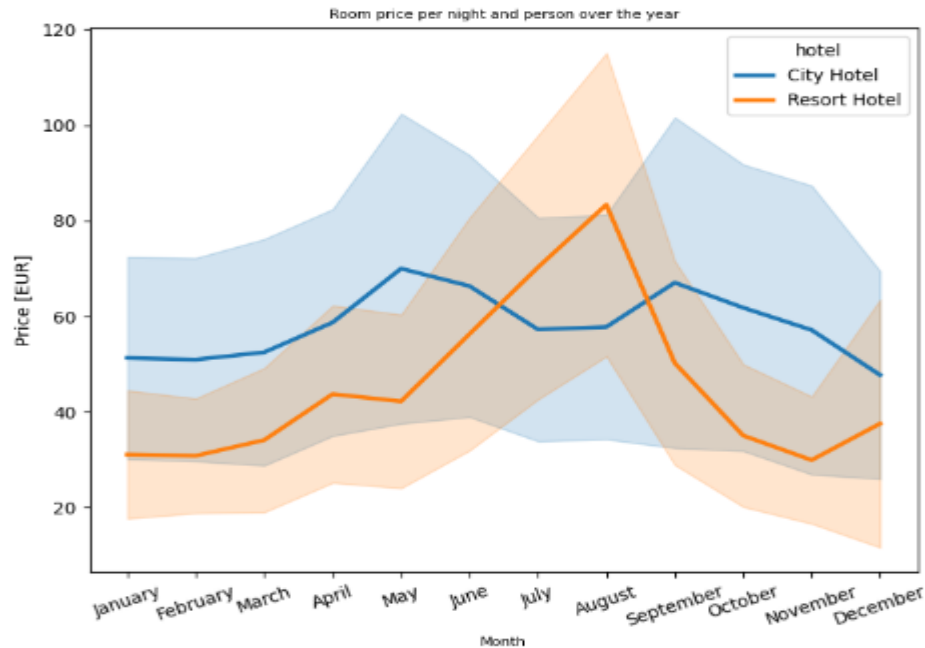
❖ To know the count the visitor in each month and preferred type of hotel rooms



❖ OBSERVATIONS :

- From the above we can see, in the month of August the more number guest are coming and less number guest comes in January over the year.
- From the second graph we see most of the guest preferred type A rooms for staying in both Resort and City hotel. So hotel should more focusing on how to maximize the number of that type.

## ❖ How does the price per night vary over the year ?

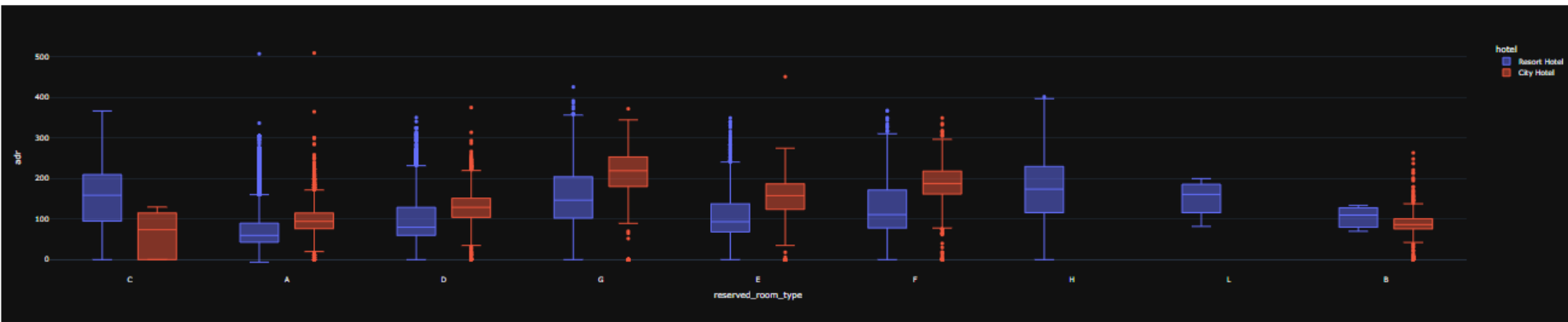


### ❑ OBSERVATION :

- There is a peak on the prices for the Resort hotel is on August.
- There is a peak on the prices for the City hotel is on May and September



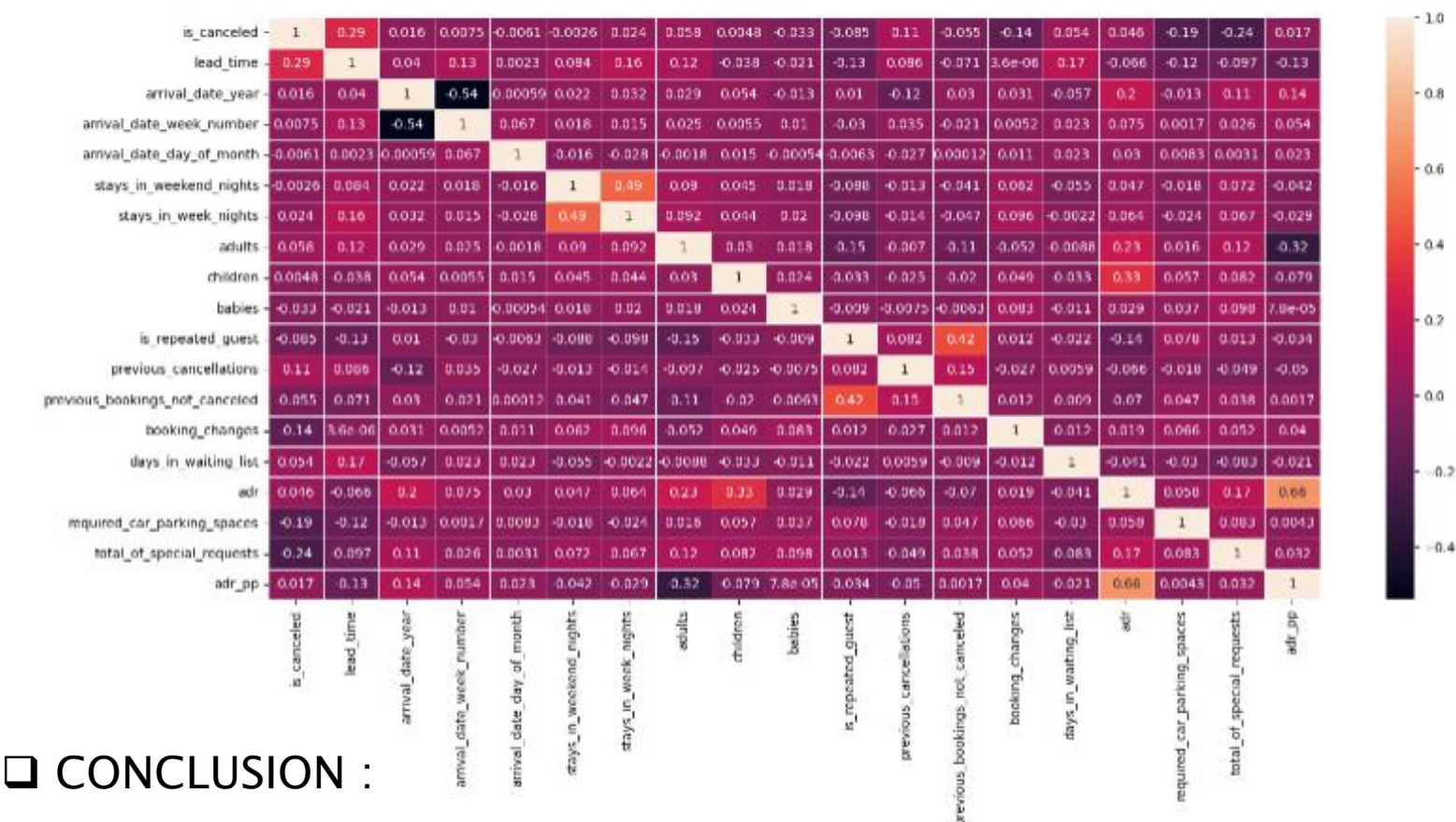
- ❖ Both hotels have different room types and different meal arrangements. Seasonal Factors are also important, So the prices varies a lot. so we are going to create boxplot for better understanding of this.



❑ CONCLUSION :

- The figure shows that the average price per room depends on its type and the standard deviation , And here we can see some outlier of our data.

❖ To know the relation between different variables :



❑ CONCLUSION :

- The canceled and same\_room\_allotted\_or\_not are negatively correlated. Not getting the same room as per reversed room is not the reason for booking cancellations.
- Lead\_time and total stay is positively correlated means more is the stay of customer more will be the lead time.
  - ADR and total people are highly correlated. That means more the people more will be adr. High adr means high revenue.
  - The is\_repeated\_guest and previous\_bookings\_Not\_canceled has strong correlation. May be repeated guest are more likely to cancel their bookings.

## ❖ CONCLUSION :

- City hotels are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
- 27.5 % bookings were got cancelled out of all the bookings.
- 79.1 % bookings were made through TA/TO (travel agents/Tour operators).
- BB( Bread & Breakfast) is the most preferred type of meal by the guests.
- Maximum number of guests were from Portugal, i.e. more than 21000 guests.
- Most of the bookings for City hotels and Resort hotel were happened in 2016.
- Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
- Booking cancellation rate is high for City hotels which almost 30 %.
- Average lead time for resort hotel is high.
- Resort hotels have the most repeated guests.
- Optimal stay in both the type hotel is less than 7 days. Usually people stay for a week.
- Almost 19 % people did not cancel their bookings even after not getting the same room which they reserved while booking hotel. Only 2.5 % people cancelled the booking.



**THANK YOU!**