# Exploratory Data Analysis On Hotel Booking

# Sudhanshu Chouhan,Kapil Narayan Singh, Nimisha Nooti

# Data scientist Trainee,Alma better

Abstract:

The hotel industry is considered to be the backbone of the tourism sector. We were provided with the dataset of the trend of hotel bookings.

Exploratory Data Analysis helps us understand the data

In this analysis, we have used Numpy, Pandas, EDA, Dataframes, and Visualizations.

1. Problem Statement

From the data, we have to know the trend in bookings. Bookings depend upon some factors known as variables. We have to check the dependency of factors affecting the Hotel bookings and draw conclusions from our analysis.

These variables are

1. Hotel: Two categories of hotels are Resort hotels or City hotels.

2. is\_canceled: Value indicating if the booking was canceled ()1 or not(0)

3.  lead\_time : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.

4. arrival\_date\_year: Year of booking arrival date .

5. arrival\_date\_month : Month of booking arrival date.

6. arrival\_date\_week\_number: Week number of the booking arrival date.

7. arrival\_date\_day\_of\_month : Day of booking arrival date.

8. stays\_in\_weekend\_nights: Number of weekend nights(Saturday or

Sunday) the guest stayed or booked to stay at the hotel.

9. stays\_in\_week\_nights : Number of week nights(Monday or Friday) the guest stayed or booked to stay at the hotel.

10. adults: Number of adults reserved the hotel stay.

11. children: Number of children.

12. babies : Number Of babies.

13. meal : kind of meal opted for.

14. country : Country code.

15.  market\_segment : Market segment designation .

16. distribution\_channel : Booking distribution channel.

17.  is\_repeated\_guest : Is a repeated guest is binary info (1) yes or (0) no.

18.  previous\_cancellations : Number of previous cancellation not cancelled by current booking.

19.  previous\_bookings\_not\_canceled : Number of previous bookings not canceled by current booking.

20.  reserved\_room\_type : Code of room type reserved.

21.  assigned\_room\_type : Code for the type of room assigned.

22.  deposit\_type  : No deposit, Non Refund, Refundable

24. Agent : ID of the travel agency that made the booking.

25. Company : ID of the company/entity that made the booking.

26. days\_in\_waiting\_list : Number of days the booking was in the waiting list before it was confirmed to the customer.

27.  customer\_type : Type of customer. contact, group, transient, transient party.

28.  adr : Average daily rate as defined by dividing the sum of all lodging transaction by the total number of staying nights.

29.  required\_car\_parking\_spaces : Is parking required.

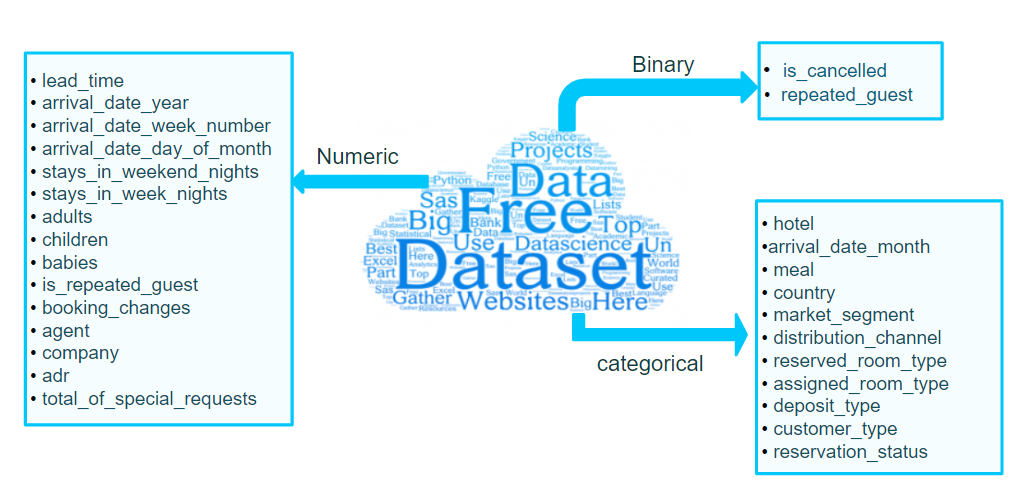
30. total\_of\_special\_requests : Number of additional special requirement.

31.  reservation status : status of reservation

32.  reservation\_status\_date : Date of the specific status.

Introduction:

From the given Data provided we will get the information about that data ,with the help of python we get to know that there are 119930 rows and 32 columns. These columns are known as independent variables which affect the bookings.After that we will try to find out the redundant variables having large number of null values .From the analysis we found out that there are 4 variables which shows null values that are company, agent,country and children.We have clean the data for our analysis .

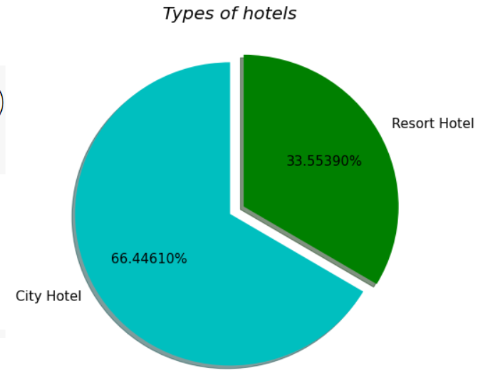


Data has been divided into three categories, namely Numeric, Binary, and categorical.

Exploratory Data Analysis:

1.**TYPES OF BOOKING**

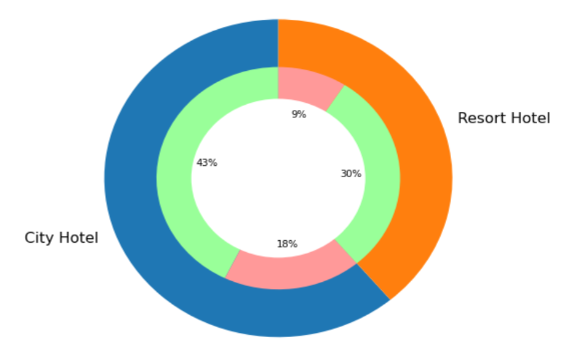
In the first analysis we get to know about booking types of hotels, there are two types of hotels ,City hotels and Resort hotels.



66.4% of City Hotels and 33.5% of Resort Hotels were booked. Therefore City Hotels are more preferred by guests compared to Resort hotels.

**2.PERCENTAGE OF BOOKING CANCELED**

From this analysis we get to know about the cancellation pattern of various types of hotels.

****We can observe from the above pie visualization that the max bookings and cancellations are happening in city hotels.

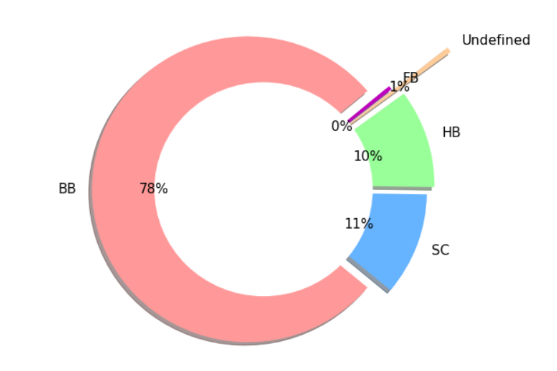
total bookings that are canceled = (18%+9%) = 27% ( 66.6% of cancellation happening in city hotel)

total bookings that are not canceled = (43%+30%) = 73%

* green = bookings not canceled
* pink = bookings canceled

**3.HOTEL BOOKING BASED ON MEAL**

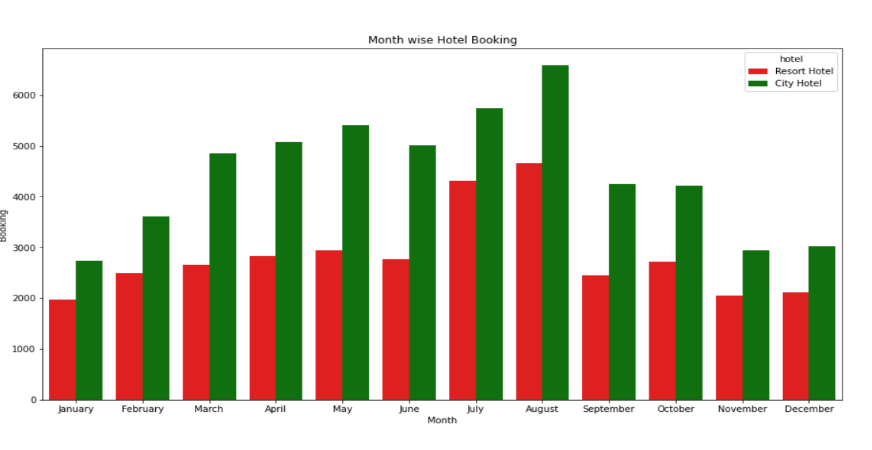
In this analysis we have divided our analysis on the basis of types of meal type generally customer take and did bookings.

****

From the above pie visualization we can conclude that 78% of Hotel Bookings are happening on 'BB' meal type i.e., 'BB: Bed & Breakfast‘.

**4.MONTH WISE HOTEL BOOKINGS**

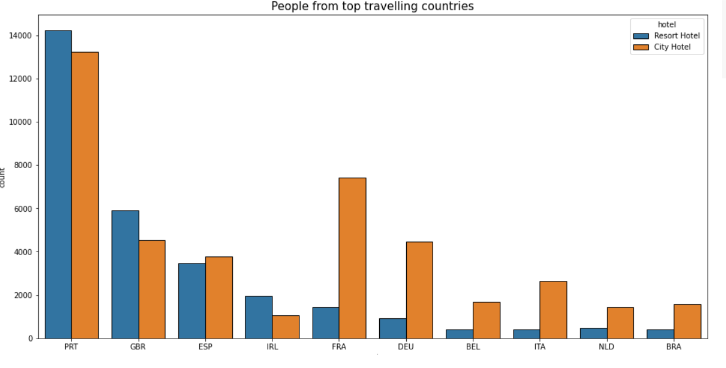
In this analysis we will now focus on the month that has maximum bookings and minimum bookings.



Most of the city and resort bookings are happening in the month of **August**  followed by July. Least bookings are happening in the month of January, November and December.

**5.BOOKING ANALYSIS BASED ON COUNTRY**

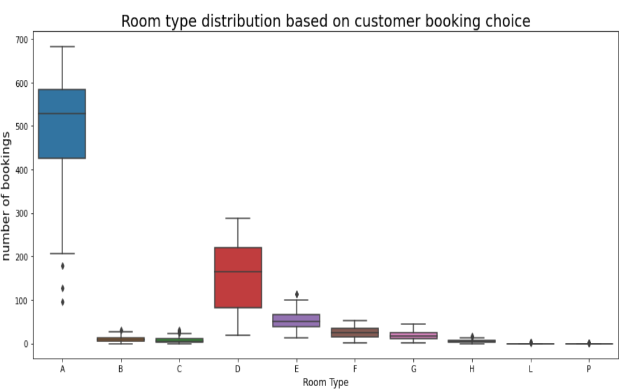
In this analysis we focuses on the country from where maximum booking has been done.



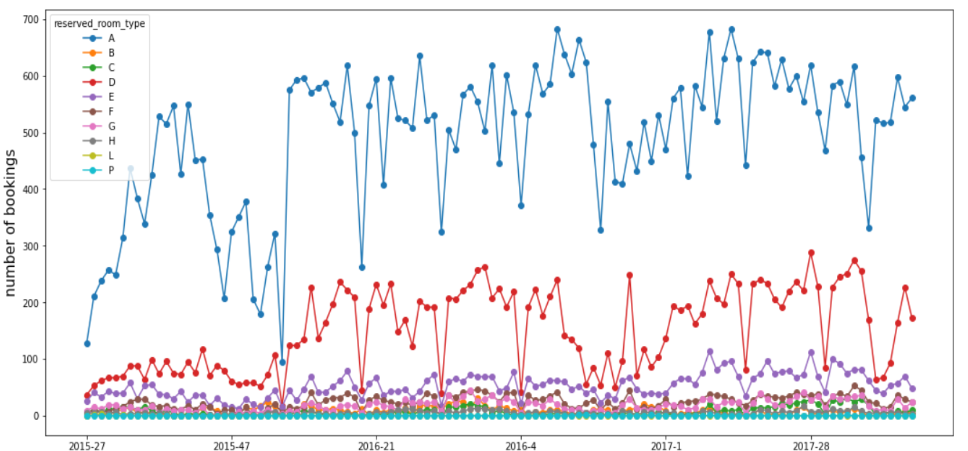
From the above bar chart visualization we can notice that most of the hotel bookings are happening in **"PTR(Portugal)"** country. We can also observe that the maximum people are preferring city hotels compared to Resort Hotels.

**6.DEMAND OF ROOM TYPE WITH RESPECT TO WEEK OF YEARS**

In this analysis we have gone through bookings that have been done by customers.



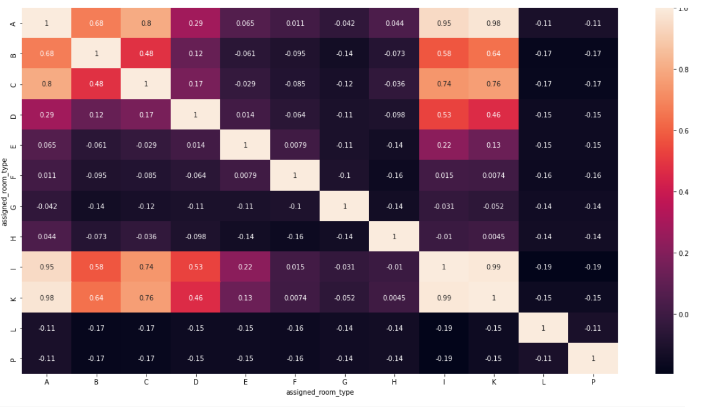
Most demanded room types are A next comes D and least bookings are done for room type P and L.



Maximum bookings happened in the 9th and 12th week of every year.

**7.POSSIBILITY OF GETTING RESERVE ROOM TYPES**

In this analysis we will analyze the possibility of getting reserve room types and assigned room types.



The lighter color indicates the more probability of getting the reserved type of room and the darker color indicates the less/no probability of getting the room of customer choice.

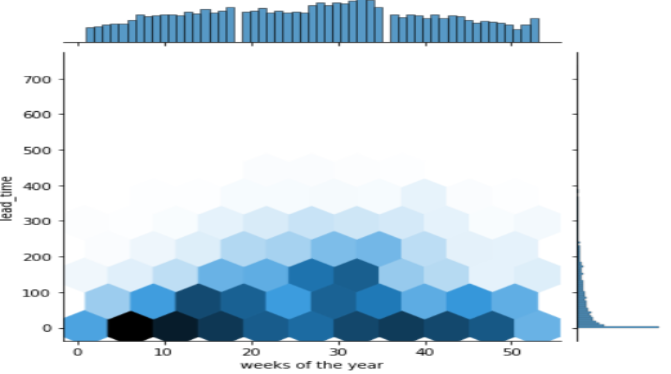
Probability of room allocation for the customer choice

reserved type A is in the

order - A, K, I, C, B.

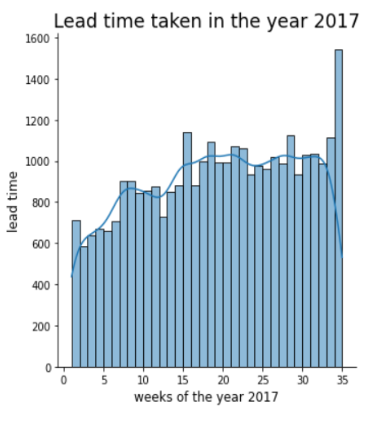
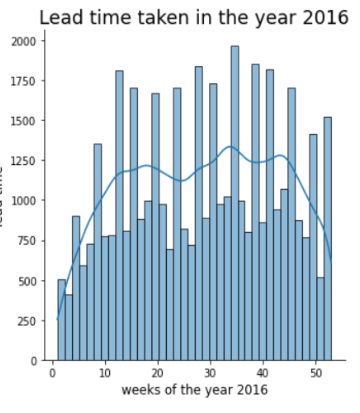
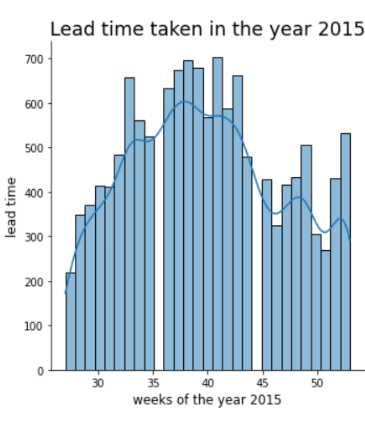
**8.ANALYZING LEAD TIME OF BOOKING**

The bar plots represents max bookings done in the weeks in X-axis i.e., (max bookings done b/w week 30 to 35) and max lead\_time taken for booking in Y-axis i.e., (max lead time taken is 0-immediate booking).

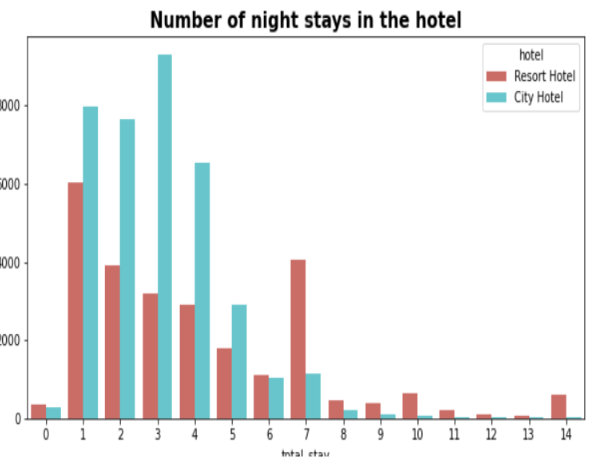
****Hex plot represents max lead time taken in the b/w 11th - 18th weeks of the year.

**9.ANALYZING LEAD TIME YEAR BY YEAR.**

In this analysis we will found out the lead time with respect to years and evaluate our conclusions on that.

****From the above displot analysis we can conclude that maximum lead time taken in bookings are in the year 2016.

**10. NUMBER OF DAYS PEOPLE STAYS IN HOTEL.**

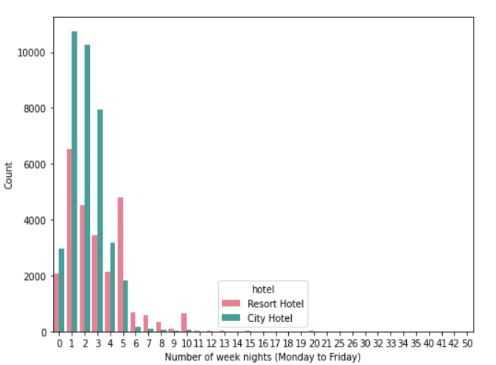
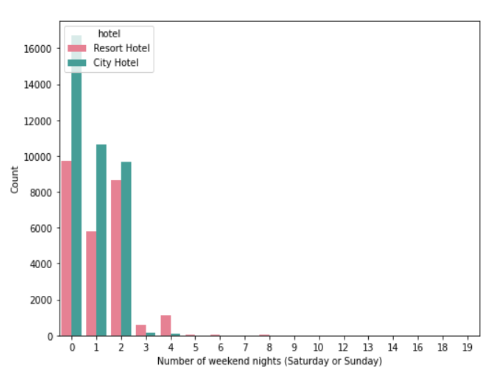
**** We can notice that the majority of people stay or do a booking of ‘7’ or less than ‘7’ days.

* Maximum night bookings are happening in the city hotels and the max length of stay is

‘3’ days.

* Maximum night bookings length of stay happening in the Resort Hotel are for one night stay.
* We can also observe that if the stay is longer than 7 days then guests are preferring to book Resort Hotels only.

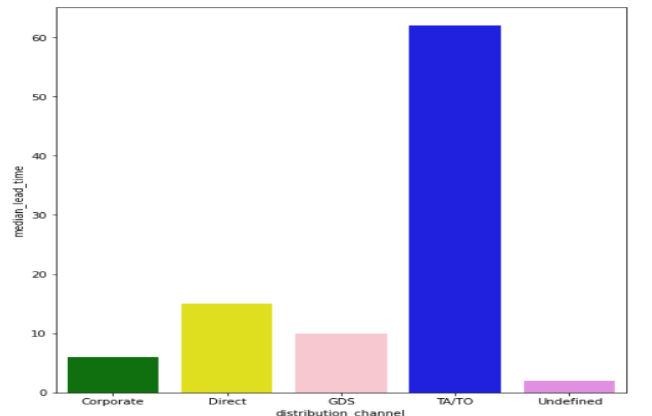
**11.NUMBER OF WEEKEND NIGHT/WEEKDAY NIGHT BOOKED.**

****We can see that the majority of people stay or do a booking of 5 or less than 5 days. Now, we can say the optimal length of stay to get the best daily rate is '5' for weeknights and '2' for weekend nights.

Max night bookings are happening in the city hotels in weekdays and the max length of stay is 1 to 2 days.

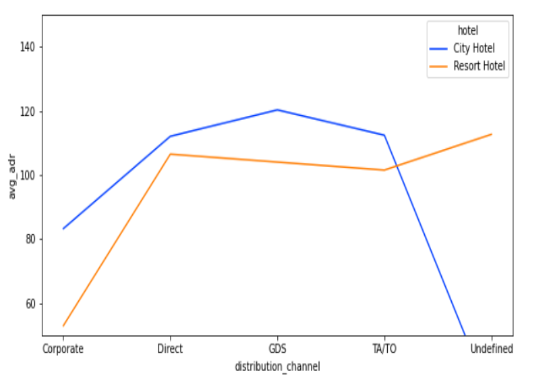
**12.ANALYZING ON THE BASIS OF DISTRIBUTION CHANNEL**

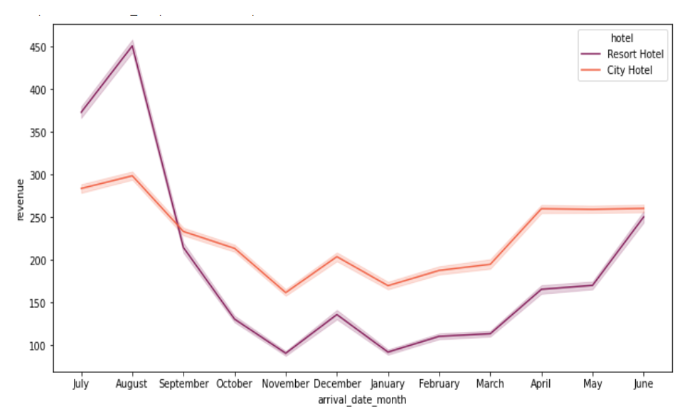
In this analysis we will analyze the distribution channel used for bookings.



Through TA/TO distribution channels, bookings were with high lead time i.e., they are booking early compared to other distribution channels.

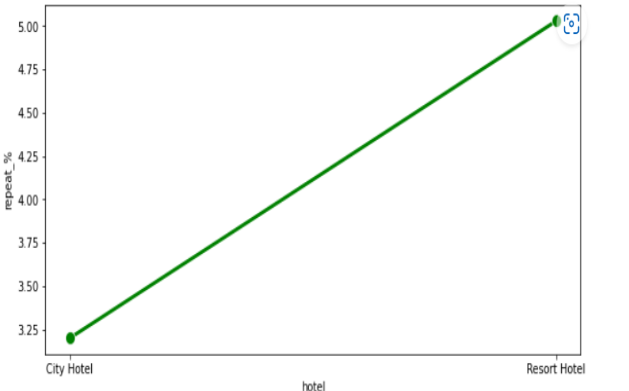
**13.ADR GENERATED THROUGH VARIOUS DISTRIBUTION CHANNEL**

**Average daily rate (ADR)**, one of the three key hotel performance indicators (along with occupancy and Revenue per available room(RevPAR)), is the measure of the average paid for rooms sold in a given time period. The metric covers only revenue-generating guestrooms.

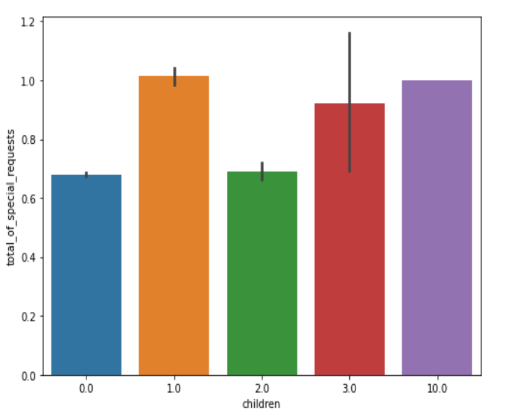
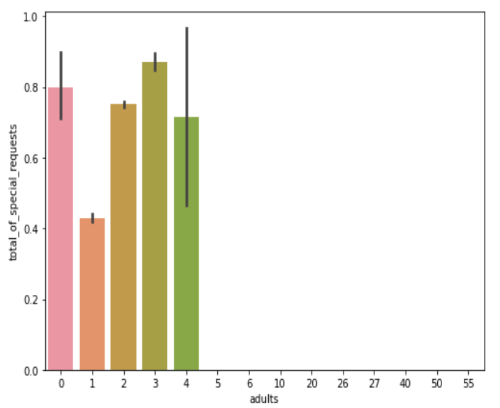
From the above analysis we can notice that the Resort hotels are getting the highest revenue in the month of 'august', 'july' and then decrease drastically. City hotel's revenue is almost constant all year.

**14.CHANCES OF CUSTOMER WILL RETURN**

In this analysis we will analyze chances of customers returning after occupying the same room.

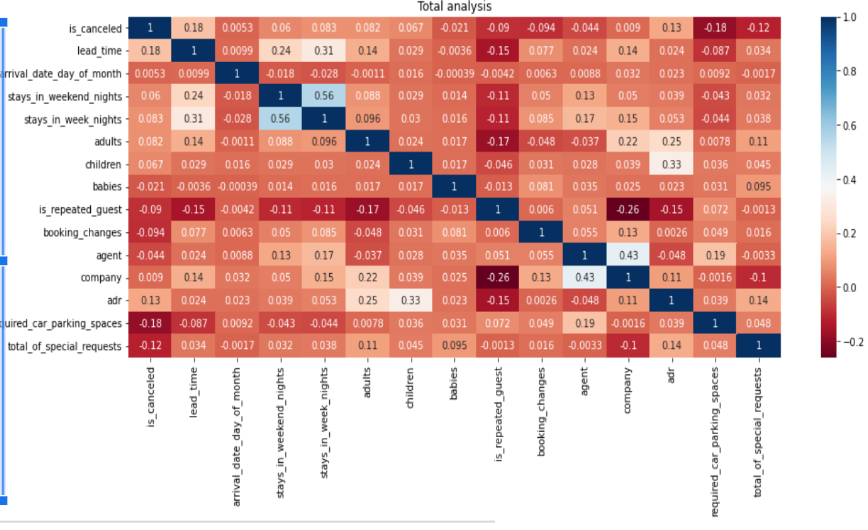
There is a very low probability that the customer will repeat. But the return percentage of a resort is slightly greater than that of a city hotel.

**15.SPECIAL REQUEST MADE BY ADULT AND CHILDREN.**

****We can see that if the adults are more than 2, there are high chances that the hotel receives more special requests and the no. of special requests for children has no much variation.

**16.EDA OF ALL TOTAL NUMERICAL DATA OF GIVEN DATA SET.**

From this analysis we will find out the dependency of variables on the hotel bookings and focus on our target area.

****

The blue color shows the highest probability of the column feature’s influence with respect to that of the row feature in the correlation matrix.

From analysis we will find out that lead time and adr has maximum impact on bookings.

**CONCLUSION**

At last we will sum up our conclusion from our analysis.

* Around 61% bookings are of City hotel and 39% bookings are of Resort hotel, therefore City hotels are busier than the Resort Hotels.
* Around 27% of total bookings are canceled, in that 66.6% cancellations are happening in City hotels.

In both resort and city hotels most of the bookings are happening in "PTR(Portugal)" country. We can also observe that the maximum people are preferring city hotels compared to Resort Hotels.

* Most of the city and resort bookings are happening in the month of August. Followed by July. Least bookings are happening in the month of January, November and December.
* The Resort hotels are getting highest revenue in the month of 'august', 'july' and then decreasing drastically. City hotel's revenue is almost constant all year.
* Maximum bookings are happening in the 9th and 12th week of every year.
* 78% of Hotel Bookings are happening on 'BB' meal type i.e., 'BB: Bed & Breakfast'.
* Most demanded room types are A next comes D and least demanded are of room type P and L.
* Probability of room allocation for the customer choice reserved type A is in the order - A, K, I, C, B.

High probability of lead time taken is '0' i.e., immediate booking is happening but high lead time taken is in the b/w 11th - 18th weeks of the year. Hence, this time is the busiest time of the year.

* The maximum lead time taken in bookings is in the year 2016.
* Majority of people stay or do a booking of 5 or less than 5 days. Now, we can say the optimal length of stay to get the best daily rate is '5' for weeknights and '2' for weekend nights.
* Max night bookings are happening in the city hotels on weekdays and the max length of stay is 1 to 2 days.
* Through TA/TO distribution channels, bookings happened with high lead time i.e, they are booking early compared to other distribution channels.
* As (Average daily rate) ADR is the revenue determining factor 'GDS distribution channel' of city hotel bookings achieving high adr (revenue).

**REFERENCES**

* NUMPY, PANDAS, MATPLOTLIB,SEABORN DOCUMENTATION.
* ALMA BETTER CLASS.
* ALMABETTER STUDENT COMMUNITY.