CAPTONE PROJECT

EDA ON HOTEL BOOKING ANALYSIS

TEAM MEMBERS

Aditya Tadas

Aishwarya Methe

CONTENT

- ➢Problem Statement
- ➢Data Understanding
- ➢Data Cleaning And Manipulation
- ➢Handling Missing Values
- ➢Feature Engineering
- ➢Uni-variate Analysis
- ➢Handling Outliers
- ➢Bi-variate Analysis
- ➢Multi-Variate Analysis
- ➢Finding Co-Relation.

PROBLEM STATMENT

- For these project we analyzing hotel booking data of three years .
- These dataset consists of booking made by the customer in Resort hotel and City hotel.
- And include information when the hooking was made, number of adults children and babies ,length of stay of customers, average daily rate of hotel, number of special request received by hotel, meal preferred by the customer and number of available parking spaces.
- The main objective of that project is to explore data and identify important factors that governs booking and give insights to hotel management.
- Also other objective is to find out what is the best time to booked hotel to get best average daily rate and high number of special request received by the hotel.

DATA UNDERSTANDING

- hotel : Hotel (Resort Hotel or City 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 into the PMS 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) the guest stayed or booked to stay at the hotel
- Stays in week nights : Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- adults : Number of adults
- children : Number of children
- babies : Number of babies
- meal : Type of meal booked. Categories are presented in standard hospitality meal packages:Undefined/SC – no meal package
- BB – Bed & Breakfast
- HB – Half board (breakfast and one other meal – usually dinner)
- FB – Full board (breakfast, lunch and dinner)
- country : Country of origin.

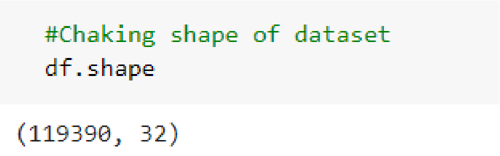
.

DATA UNDERSTANDING

- Market segment : Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- Distribution channel : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- Is repeated guest : Value indicating if the booking name was from 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. Code is presented instead of designation for anonymity reasons.
- Assigned room
  type : Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
- Booking
  changes : Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
- deposit type : Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:
  No Deposit – no deposit was made
- Non Refund * a deposit was made in the value of the total stay cost
- Refundable – a deposit was made with a value under the total cost of stay.
- agent : ID of the travel agency that made the booking

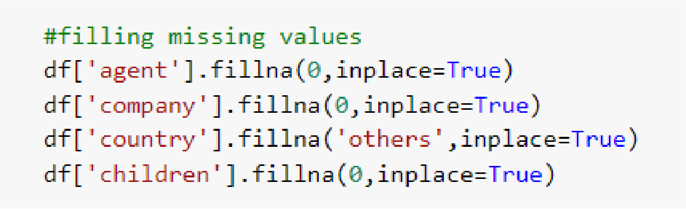
- company : ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
- Days in waiting list : Number of days the booking was in the waiting list before it was confirmed to the customer
- Customer type : Type of booking, assuming one of four categories:
- Contract - when the booking has an allotment or other type of contract associated to it
- Group – when the booking is associated to a group
- Transient – when the booking is not part of a group or contract, and is not associated to other transient booking
- Transient-party – when the booking is transient, but is associated to at least other transient booking
- 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 last status, assuming one of three categories:
- 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
- Reservation status date : Date at which the last status was set. This variable can be used in conjunction with the Reservation
  Status to understand when was the booking canceled or when did the customer checked-out of the hotel

DATA CLEANING AND MANIPULATION



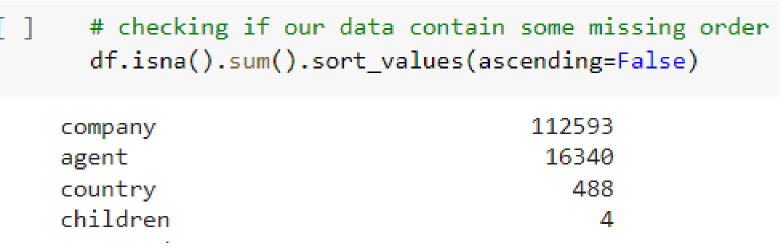
```
#Chaking shape of dataset
df.shape

(119390, 32)
```

- Dataset contains 119390 records and 32 features.



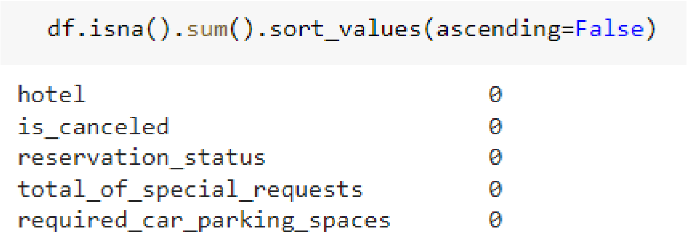
```
#filling missing values
df['agent'].fillna(0,inplace=True)
df['company'].fillna(0,inplace=True)
df['country'].fillna('others',inplace=True)
df['children'].fillna(0,inplace=True)
```

- Filling these missing values with 0.



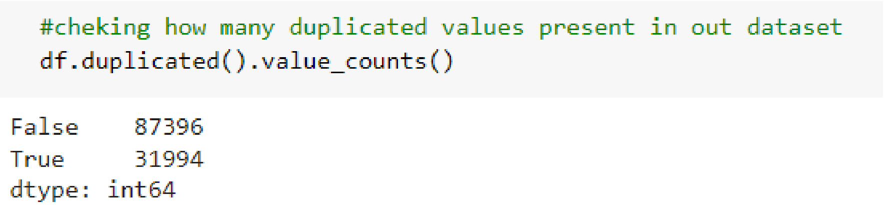
```
[ ]   # checking if our data contain some missing order
      df.isna().sum().sort_values(ascending=False)

 company                         112593
 agent                            16340
 country                            488
 children                             4
```

- There are four columns company, agent, country and children having missing values.



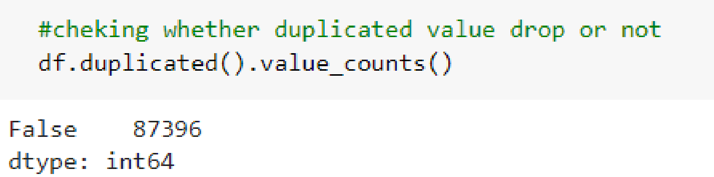
```
df.isna().sum().sort_values(ascending=False)

hotel                           0
is_canceled                     0
reservation_status              0
total_of_special_requests       0
required_car_parking_spaces     0
```

- Now there is no missing values in dataset.

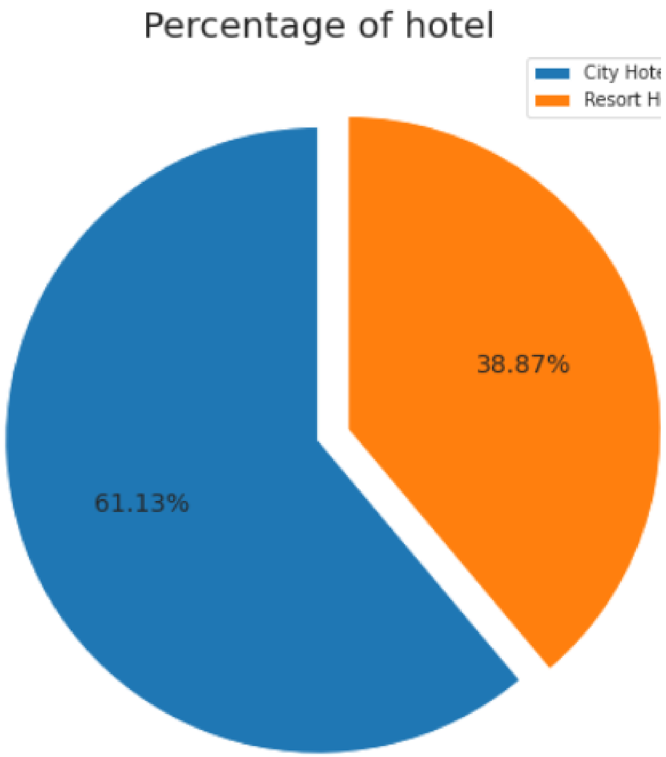


```
#cheking how many duplicated values present in out dataset
df.duplicated().value_counts()

False    87396
True     31994
dtype: int64
```

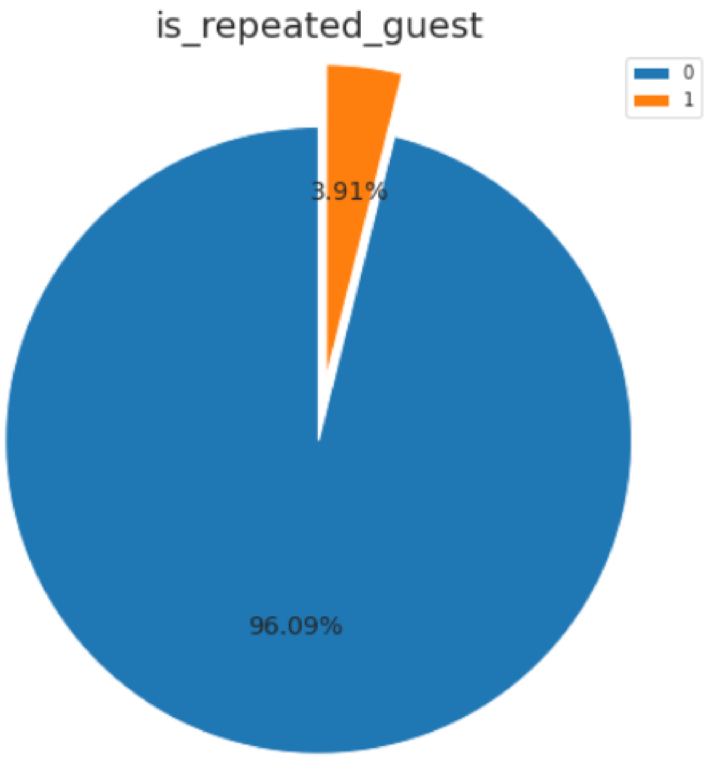
- There are 31994 duplicates value in dataset we need to drop these outliers from the dataset.



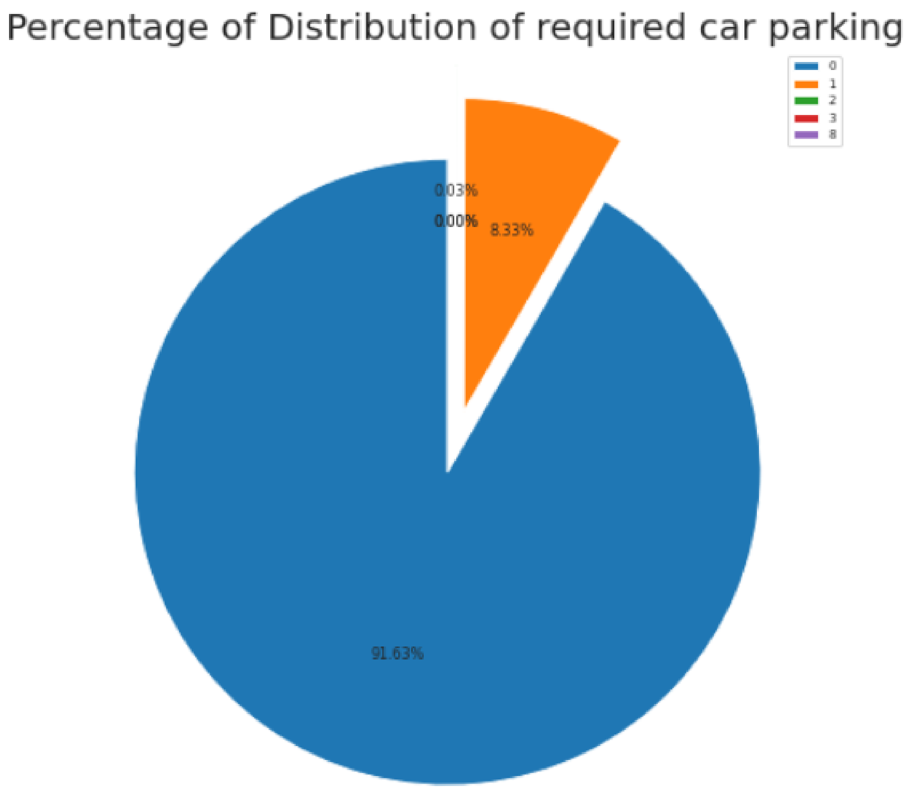
```
#cheking whether duplicated value drop or not
df.duplicated().value_counts()

False    87396
dtype: int64
```



Percentage of hotel

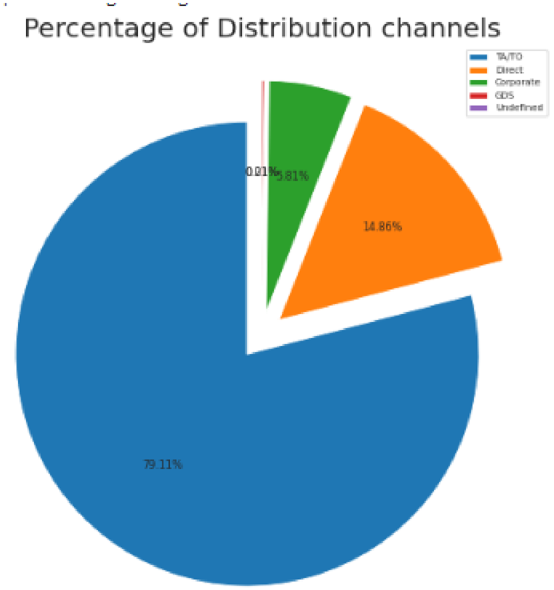


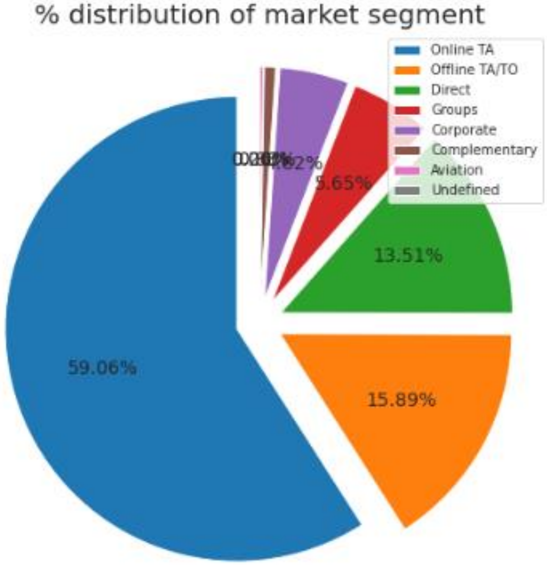
is_repeated_guest

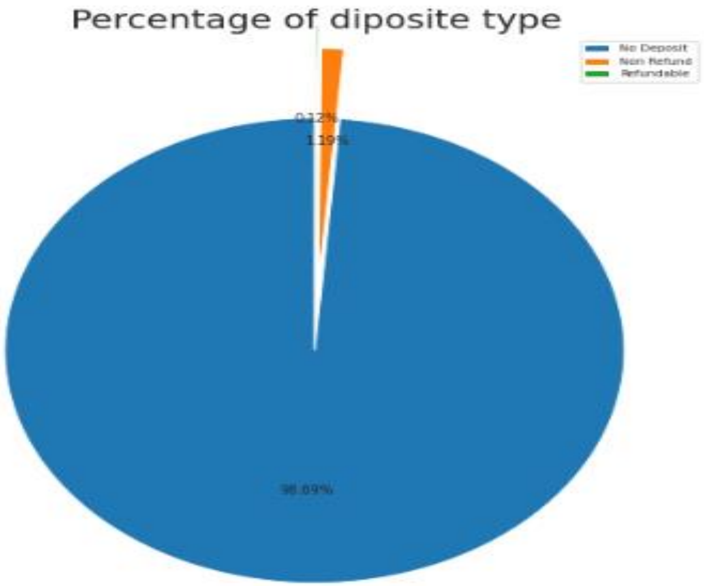


Percentage of Distribution of required car parking

- City hotel is the most preferred hotel type by the customers so we can say that city hotel are in more demand than resort hotel by the customers.
- Only 3.91 % customers are repeated guest in hotel and 96.09 % guest are new so we can say that retenshion rate is very low.
- Over 91.63 %   customers have not required car parking in hotel and 8.33 % customers required only one car parking.


Percentage of Distribution channels


% distribution of market segment

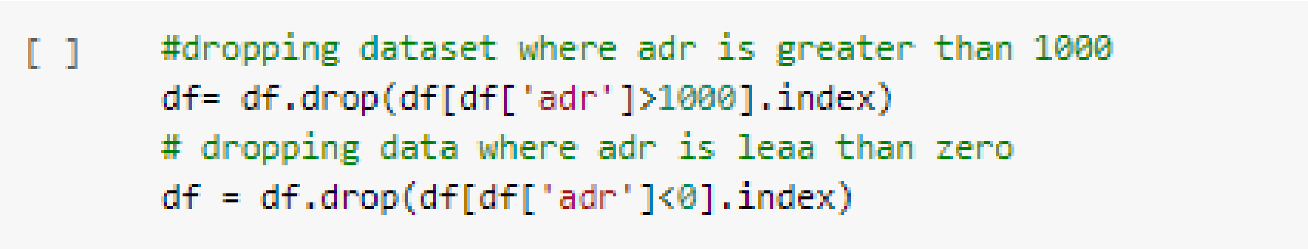

Percentage of diposite type

- Most of the bookings have been done by using online TA market segments and TA/TO distribution channels hence we can conclude that most of the customers prefers to book hotel online through traveling agency after that most of the customer preferred offline travel agency to book hotel or directly contact to hotel for booking.

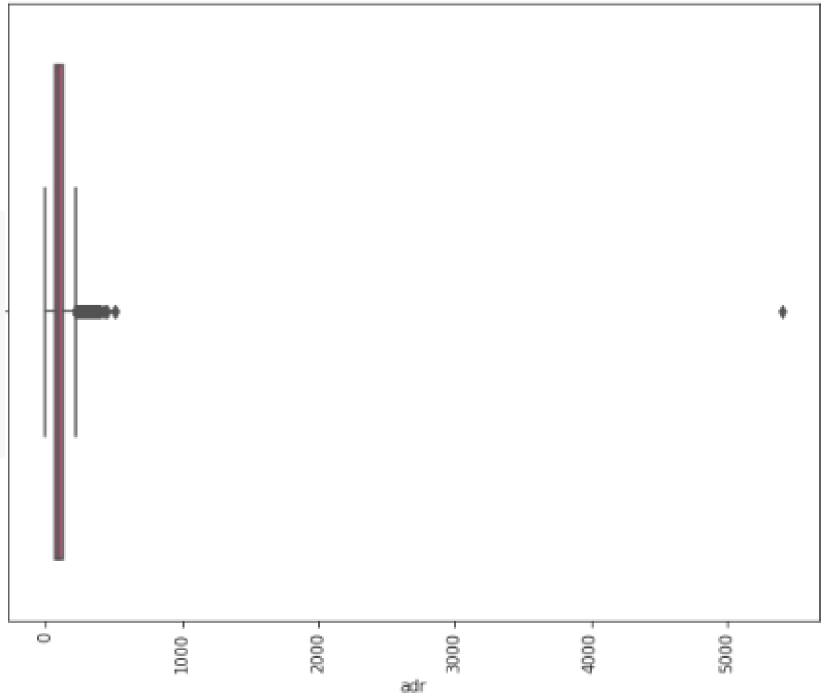
- Also most of the customers did not like to give any deposit to book hotels (98.09%.)

- So we can say that online traveling agency , offline and direct market segment gives facilities to book hotels without any deposite.

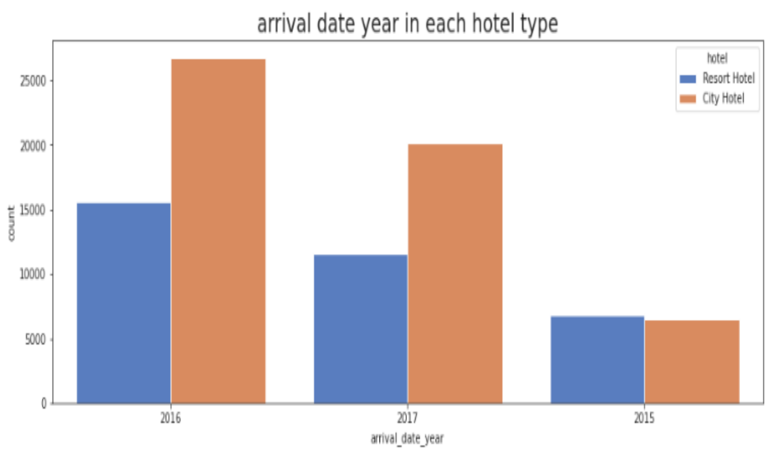
Removing outliers from adr columns

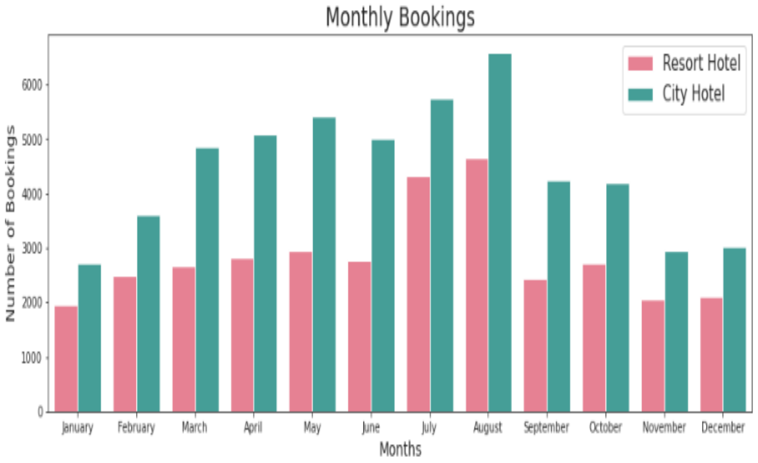


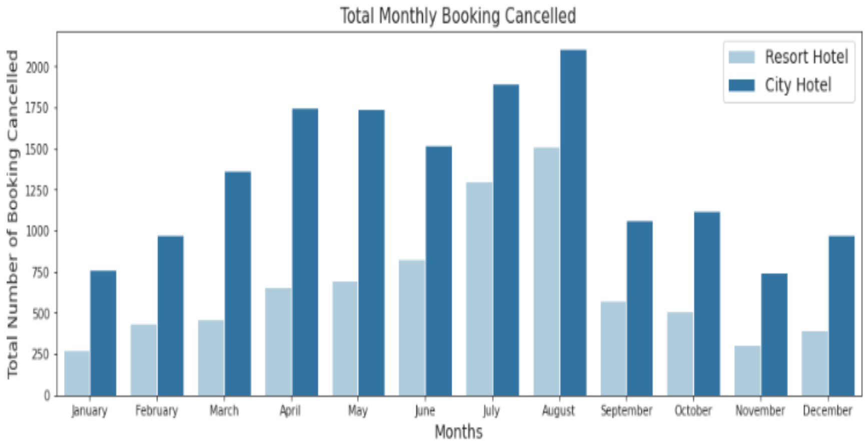
```
[ ]    #dropping dataset where adr is greater than 1000
       df= df.drop(df[df['adr']>1000].index)
       # dropping data where adr is leaa than zero
       df = df.drop(df[df['adr']<0].index)
```



- Visualizing outlier through box plot . There is one big outlier that is shown in the box plot we have removed that outlier from dataset to visualize data in better manner.

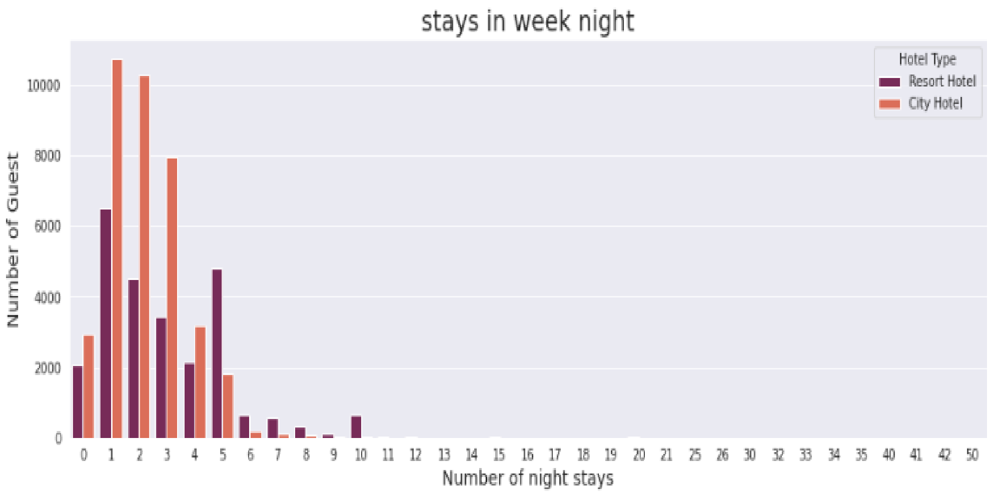


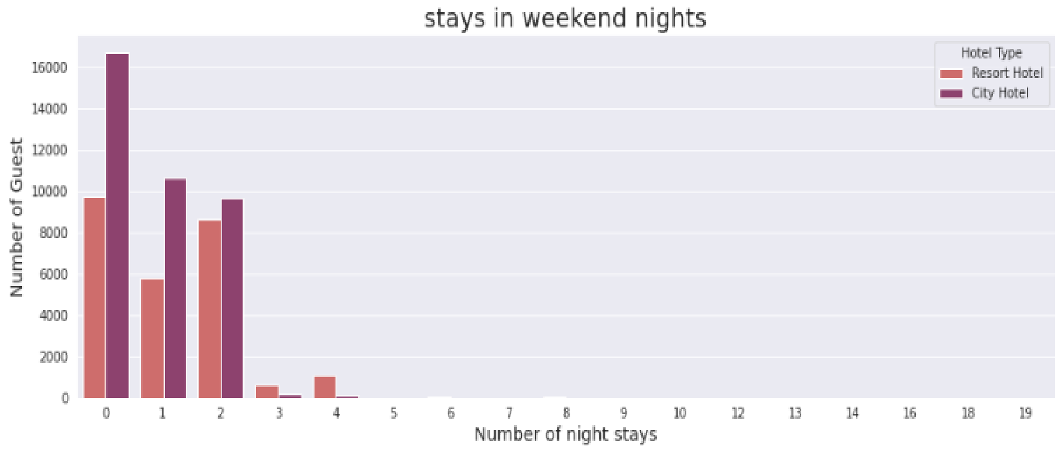




- Most of the bookings have been done in the year 2016 where most number of bookings have been done in city hotel type.
- When we analyze number of booking made in each month we can see that from the month April to August most number of bookings have been made we can also conclude that when the new financial year started customers prefers to book hotel for meeting , traveling or other official purpose also we can say that when there is rainy season most people booked hotel.
- In the month of November to february most of the customers doest like to book hotel.

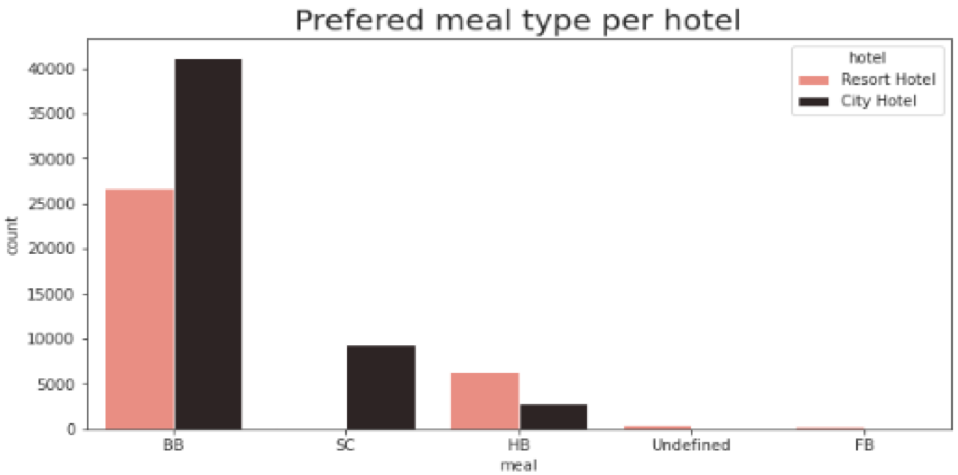
- April To August months have most number of booking also Most number of bookings have been canceled in that months.
- In the month of November to February least number of bookings have been canceled.

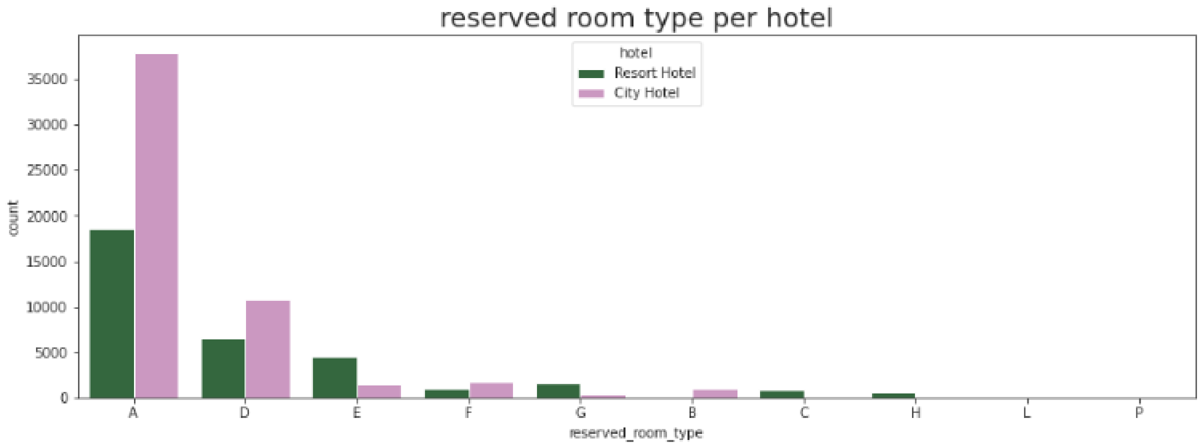


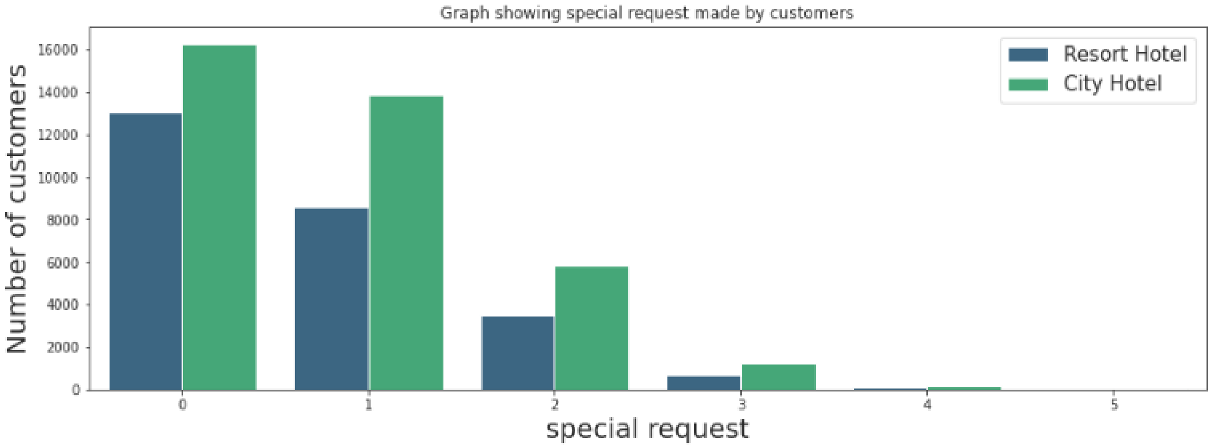


- In a week night most of the customers prefers to stay for 1 and 2 nights also we have some customers prefers to stay in Resort hotel for 10 nights but in city hotel very few customers prefers to stay for more than 5 nights So we can say that customers prefers to stay in week nights in resort hotel more that city hotel.

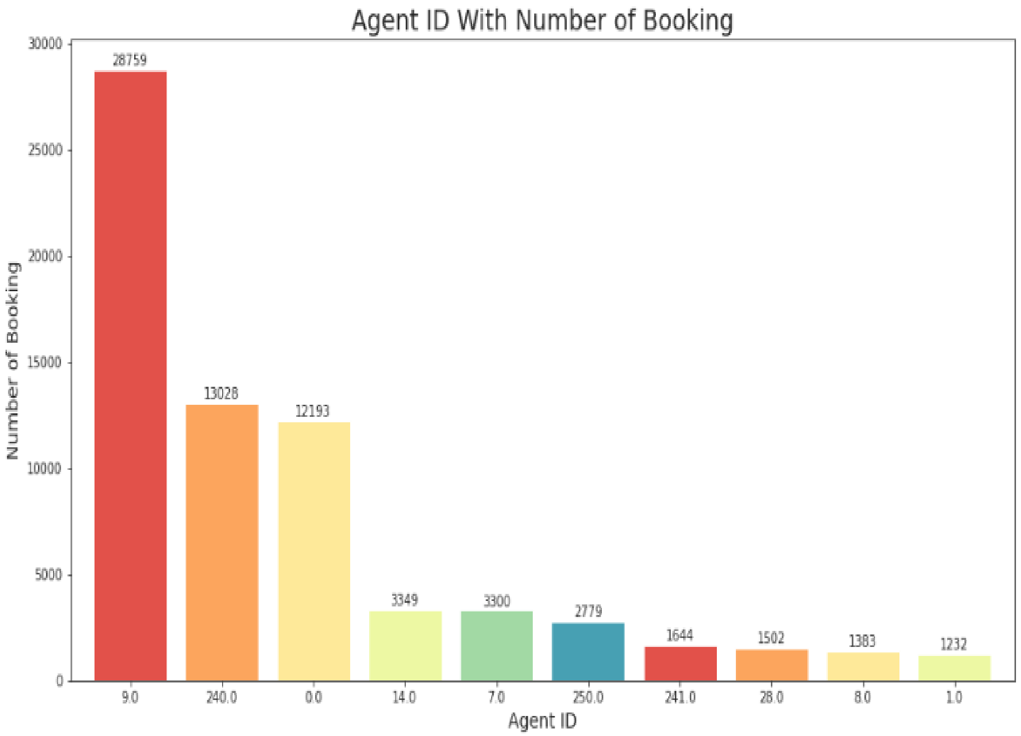
- In weekend night most of the customer does not preferred to stay at night in both hotel type there are some customers who preferred to stay at 1 or 2 week end nights.

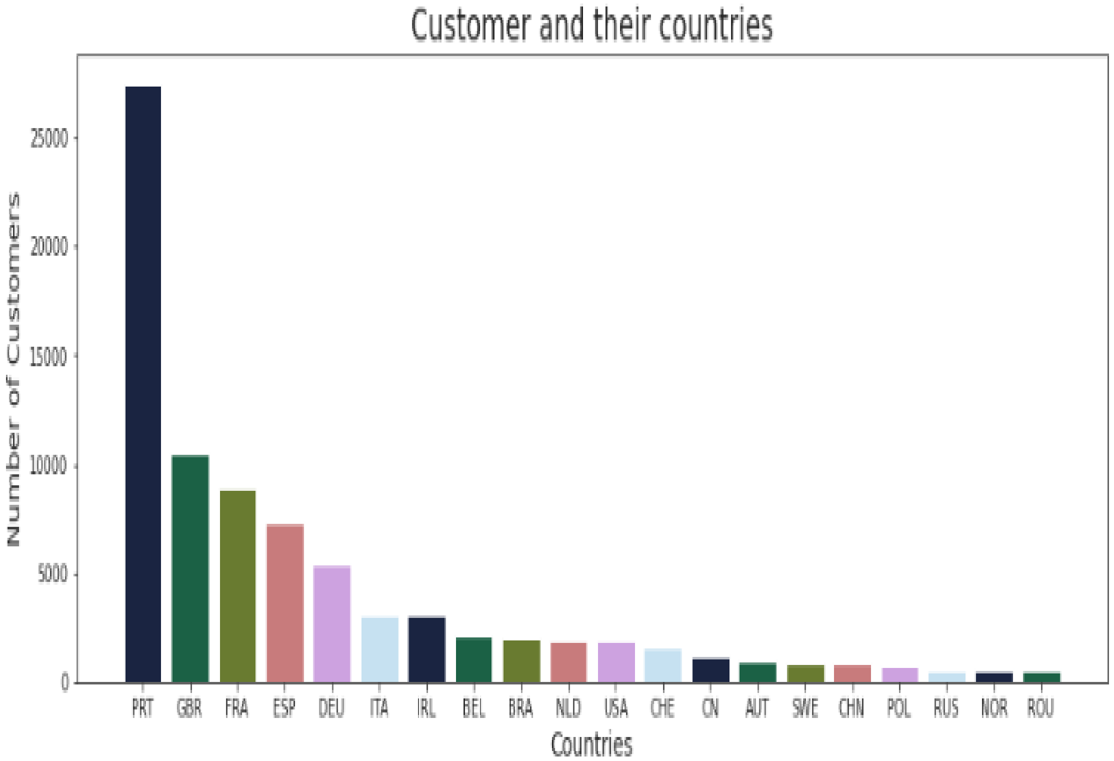






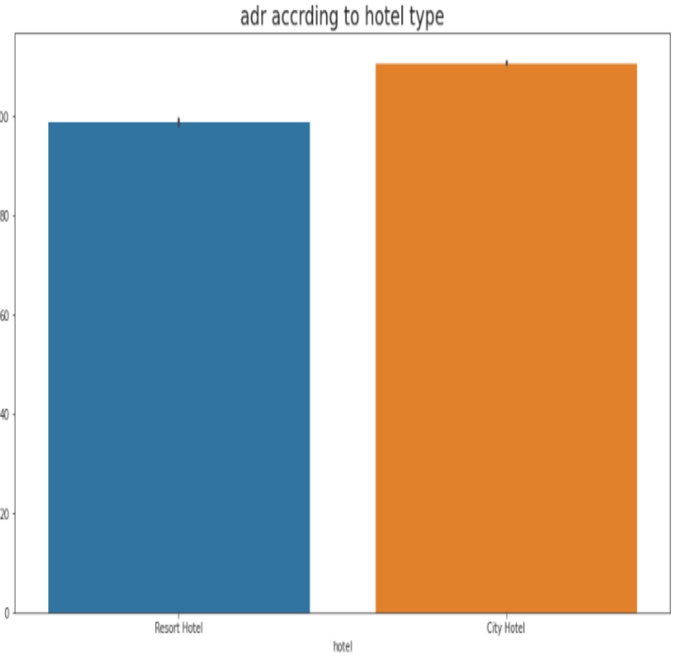
- BB(Bed and Breakfast) meal is mostly preferred by the customers in both type of hotels. Some of the customer in resort hotel preferred SC(self catering) and very few customer preferred HB(half board) and FB(Full board) meal.
- Most number of customer preferred “A” room type in both hotels after that ‘D’ and ‘E’ reserved room type have been preferred.
- When there is no rush in hotel i.e there is very few number of customers booked hotel then hotel will received more number of special request made by the customers.

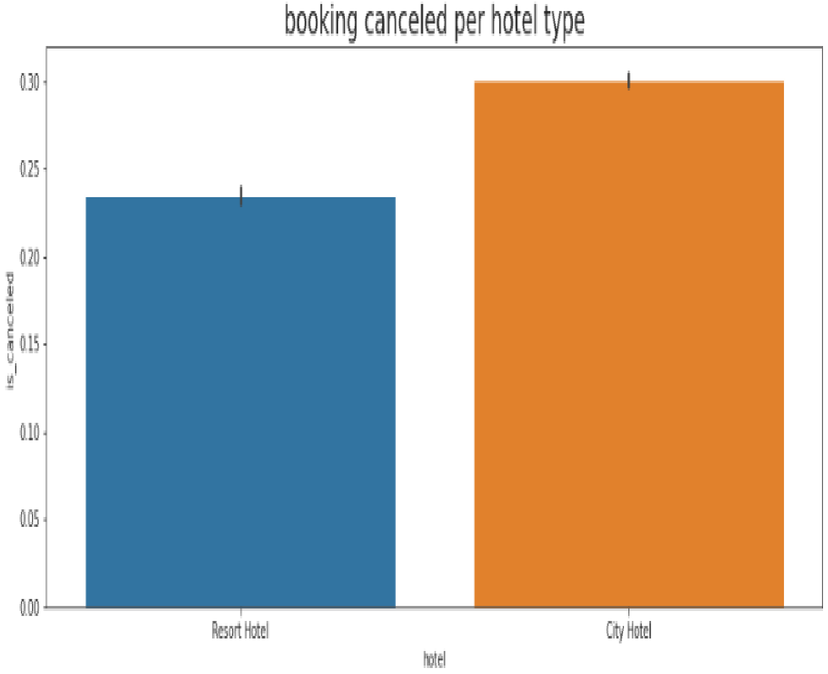


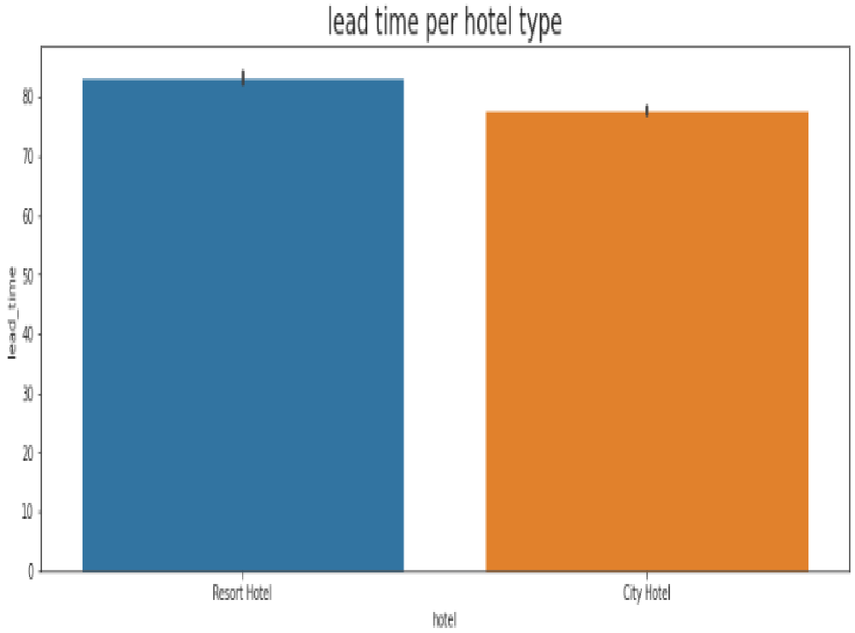


- Most number of booking have been done through the agent number 9 after that agent 240 has done second most number of booking.

- Most of the customers who booked hotel are from the PRT country.


adr accrding to hotel type

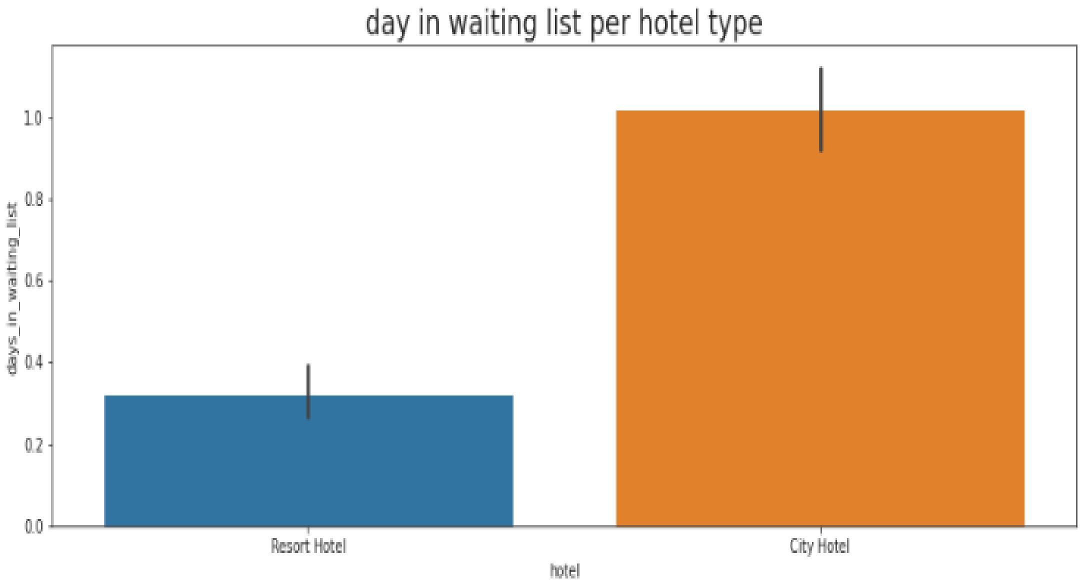

booking canceled per hotel type

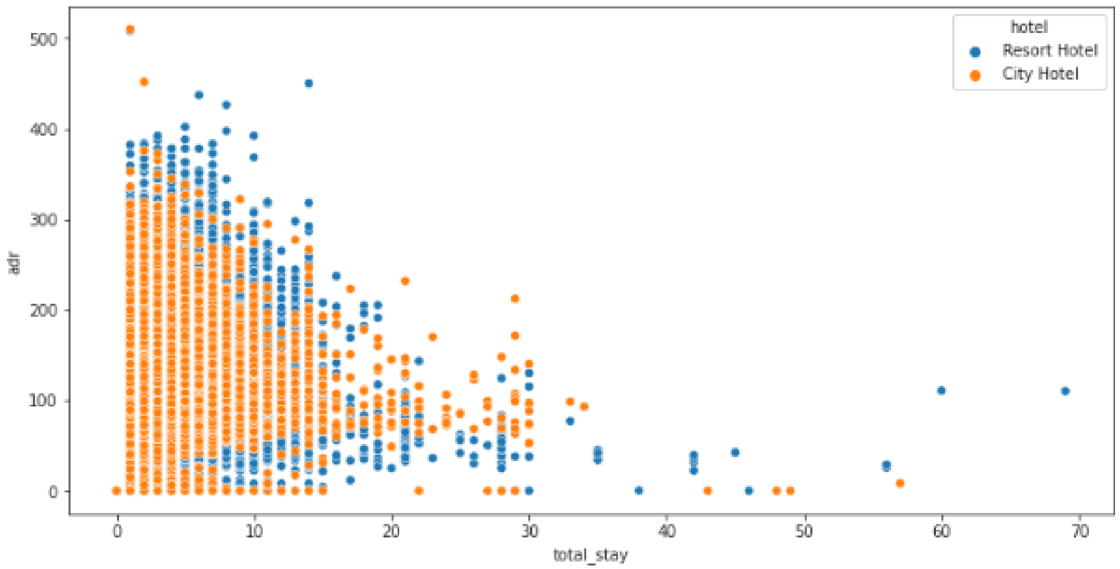

lead time per hotel type

- Most number of customer booked city hotel that’s why demand of city hotel is greater that resort hote so we can see that adr of city hotel is greater than resort hotel.

- City hotel is busy than resort hotel also it has high number of booking soo most number of booking Is also canceled in city hotel than resort hotel.

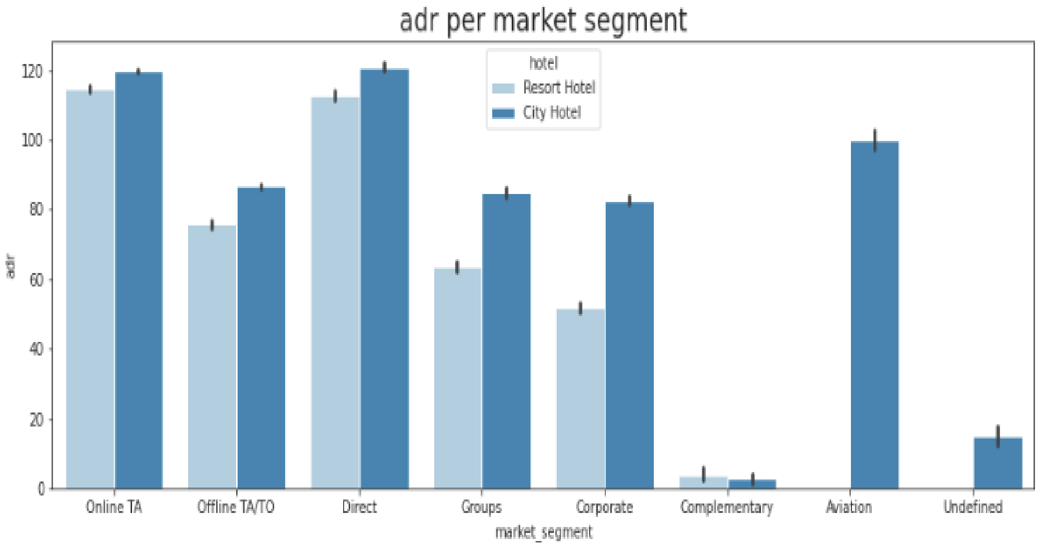
- Most of the customer prefer to stay in resort hotel for longer time than city hotel.

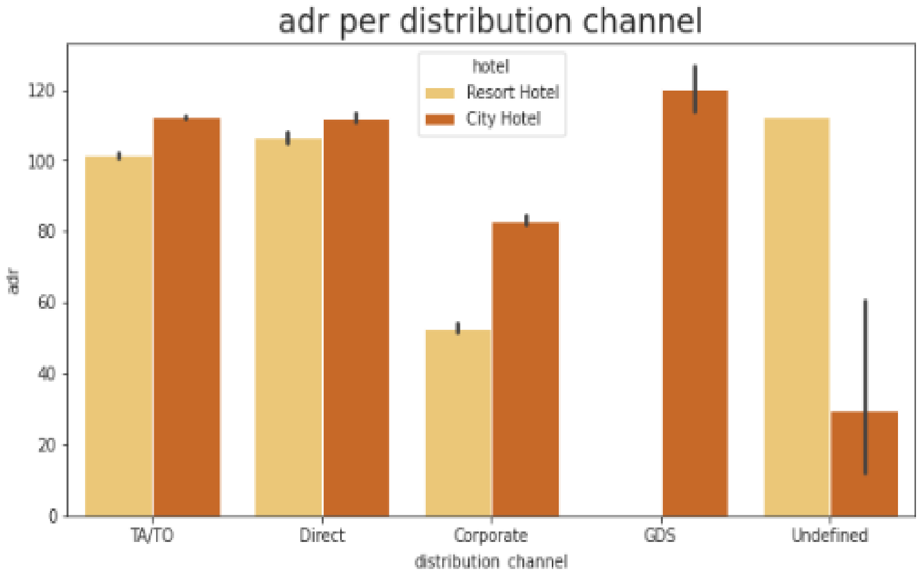




- Most of the customer preferred to book city hotel that’s why demand of city hotel is greater than resort hotel hence we can see that in the above chart waiting day of city hotel is greater that resort hotel.

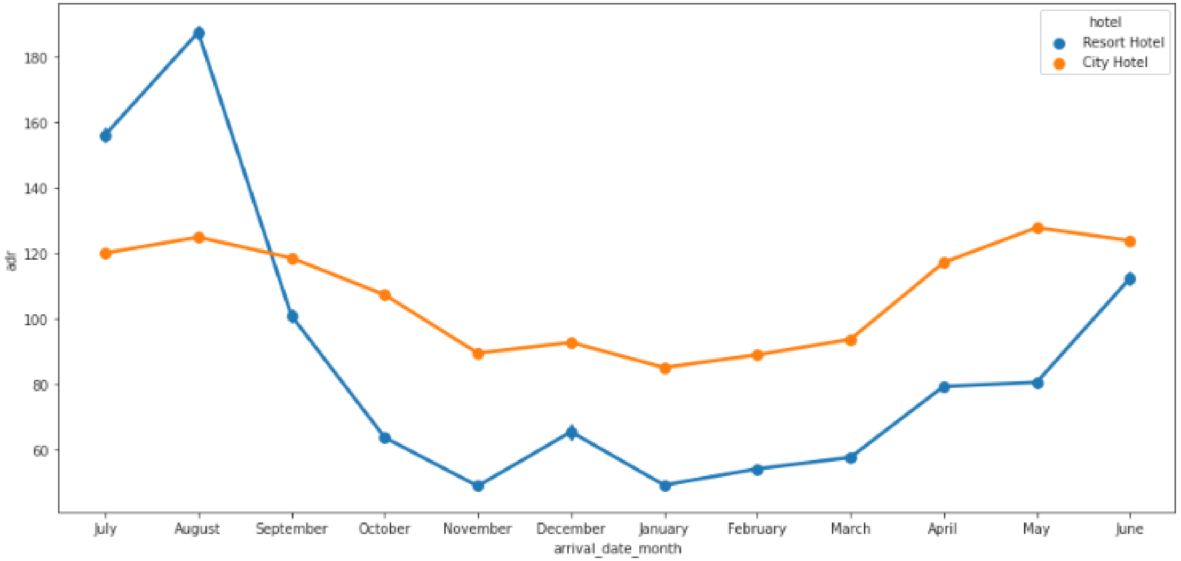
- From the scatter plot when the customer stay for the longer days in hotel then it will get lower adr . Longer the stay of customer best adr customer will get from hotels.

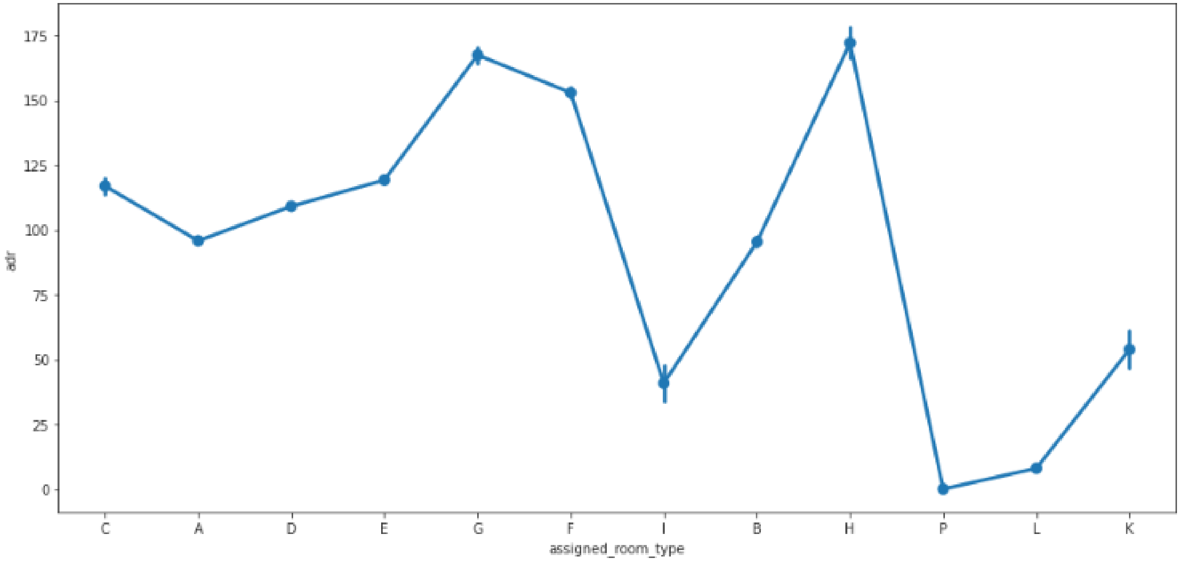




Most number of booking has been made through Online TA and Direct market segment and TA/TO and Direct distribution channel soo these market segment and distribution channel will contribute more to the revenue of hotels that’s why adr of online TA and Direct market segment and TA/TO and Direct distribution channel .

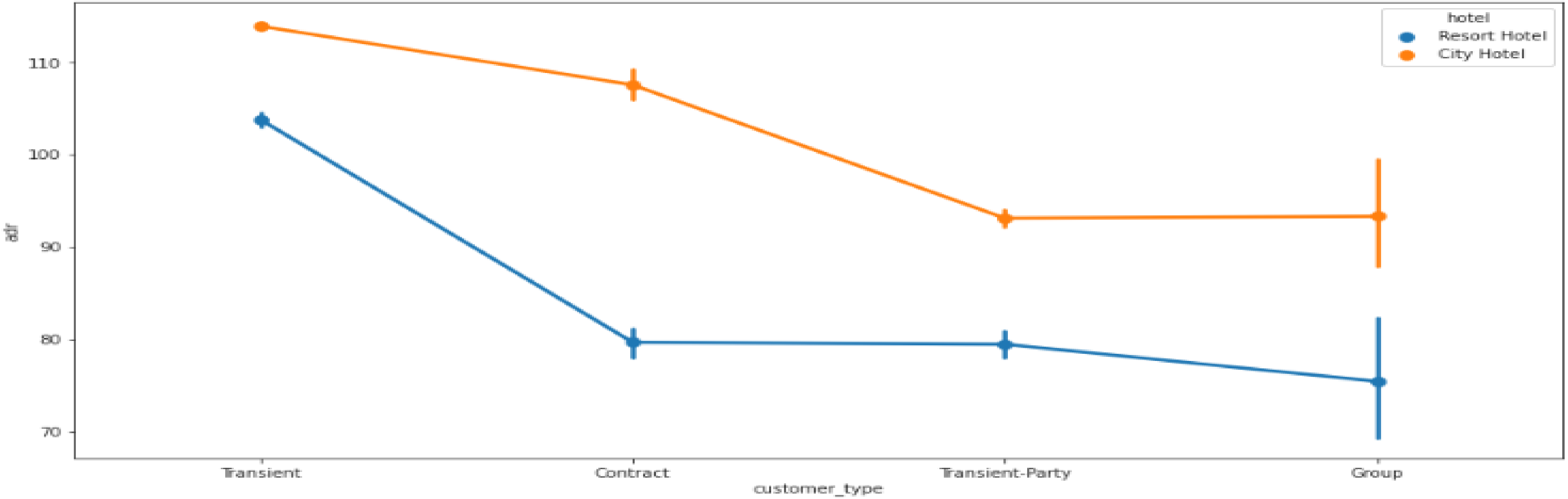
Offline TA/TO and Corporate market segment will contributed mostly to the revenue of hotel that’s why adr of these is high after Online TA and Direct market segment.





- In month July and August resort hotel have high adr than resort hotel soo we can say that there is a huge demand for resort hotel in the month of july and august and resort hotel have generated high revenue in that months.

- The adr of 'G' and "E" and 'H' are greater than all room type that's why we can say that these room types are special room types booked by customers in both hotel type.

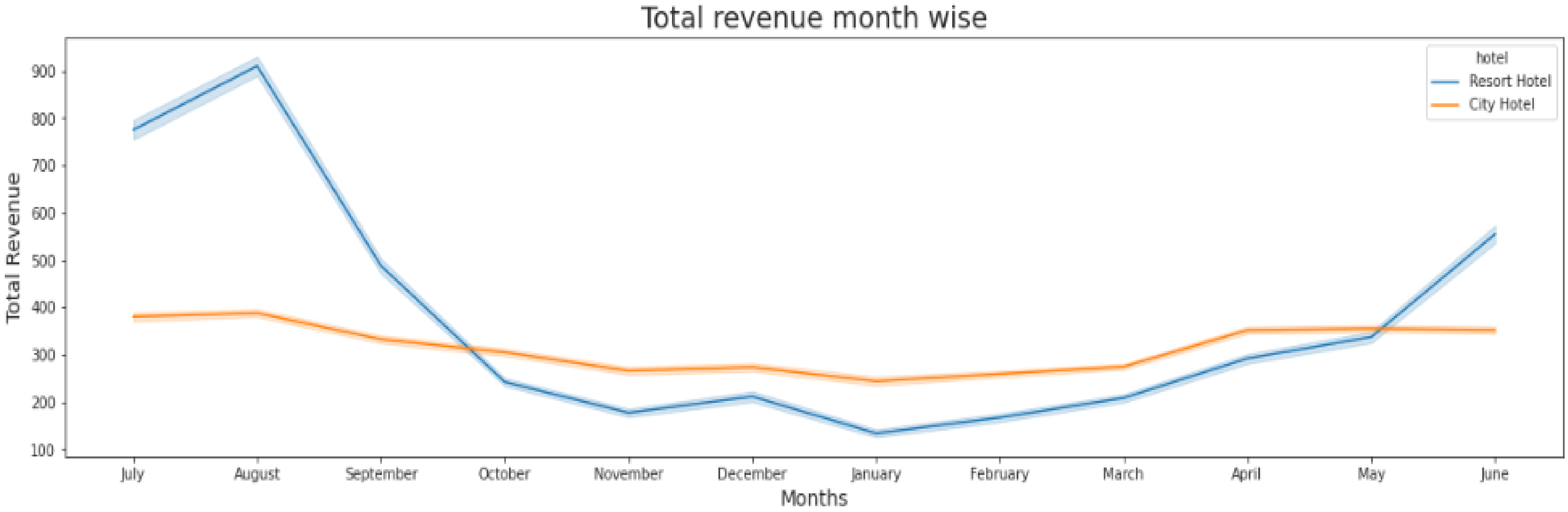


When the customer type is transient then average daily rate (adr) of both hotel is higher than other types of customer.

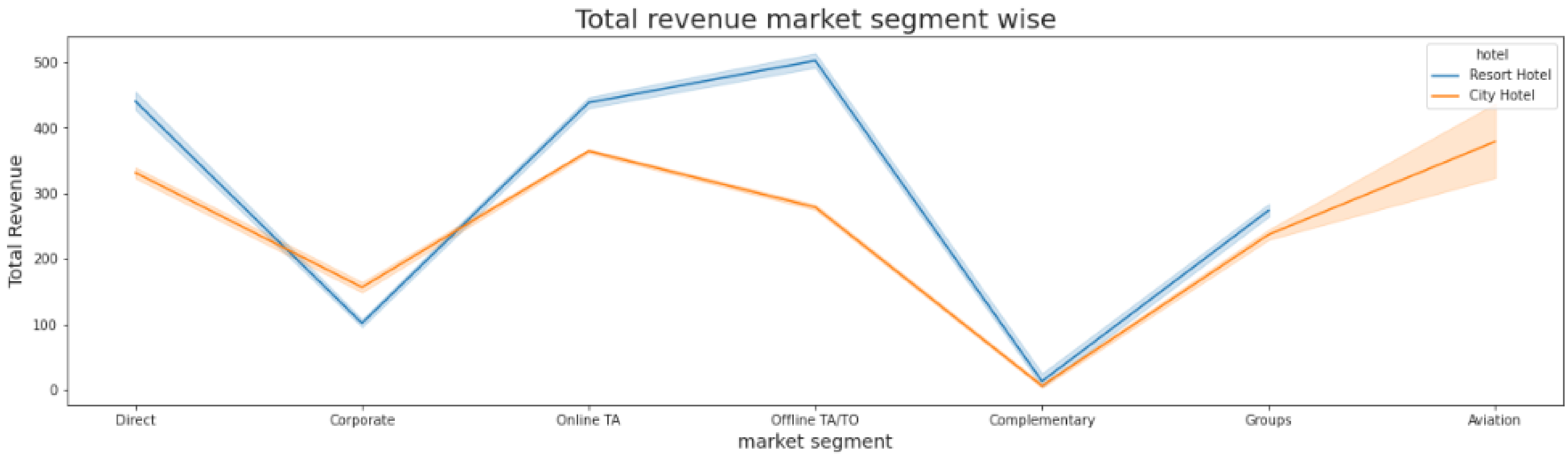
When the customer type is contract then adr of city hotel is higher but in case of resort hotel adr for customer type contract is less as compared to the city hotel type.

In resort hotel when customer types are contract and transient party adr for those customer type is same and in case of group customer type the adr is least in resort hotel.

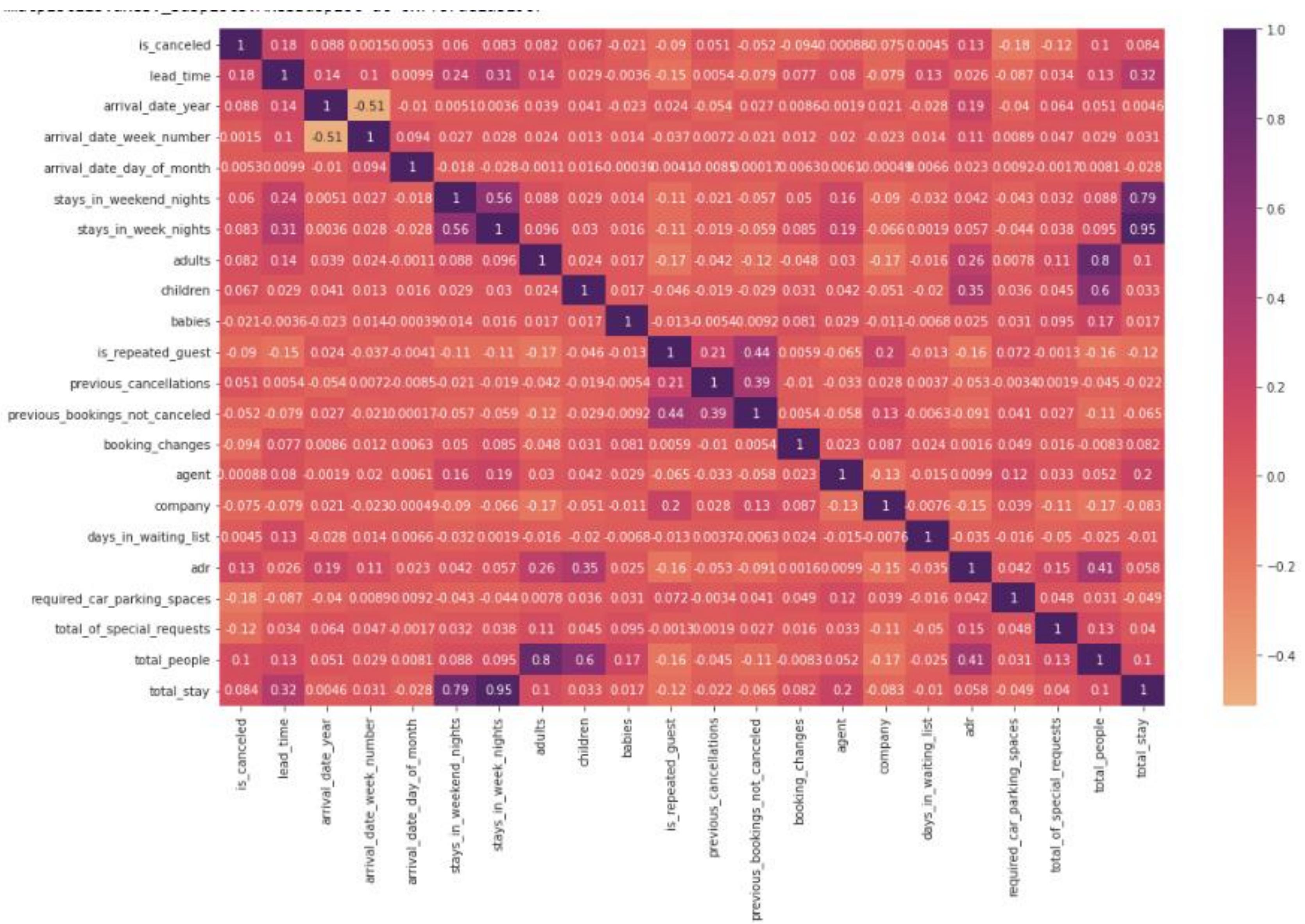
But in case of city hotel adr of transient-party and group are same.



- Graph shows total revenue generated by each hotel in each month .
- In case of city hotel revenue generated in each month is constant there is no fluctuation in that.
- But in case of resort hotel after may resort hotel generated more revenue than city hotel then it is increasing in june and there is huge fluctuation in july and august so we can say that resort hotel generated most revenue in between months of june-August then the revenue was decressing slightly from august and it becomes lowest in January month.



- When bookings have made through online TA and offline TA/TO and Direct resort hotel generate more revenue than city hotel.
- Most of the customer booked resort hotel through this market segment it contributes more to the revenue of resort hotel than city hotel.

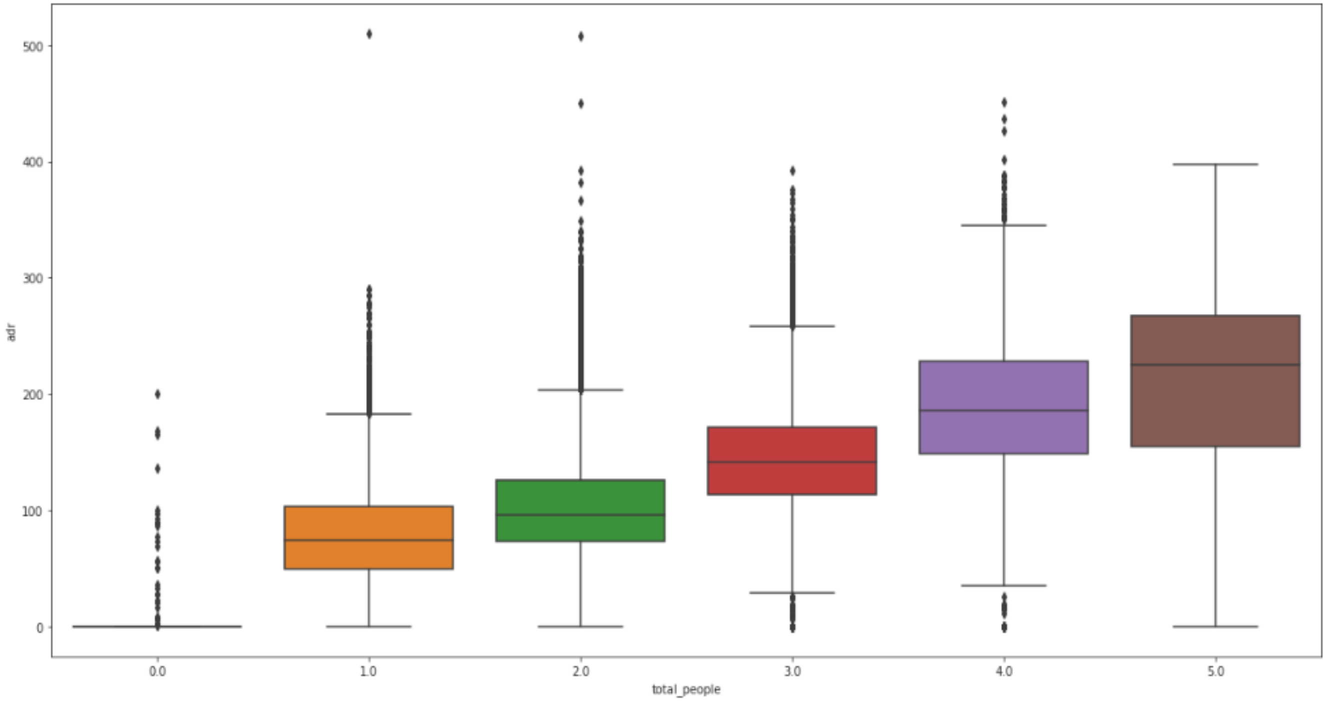


- In the following co-relation plot we can see that when the total people incresses adr also increases that means adr having positive correlation with total people.

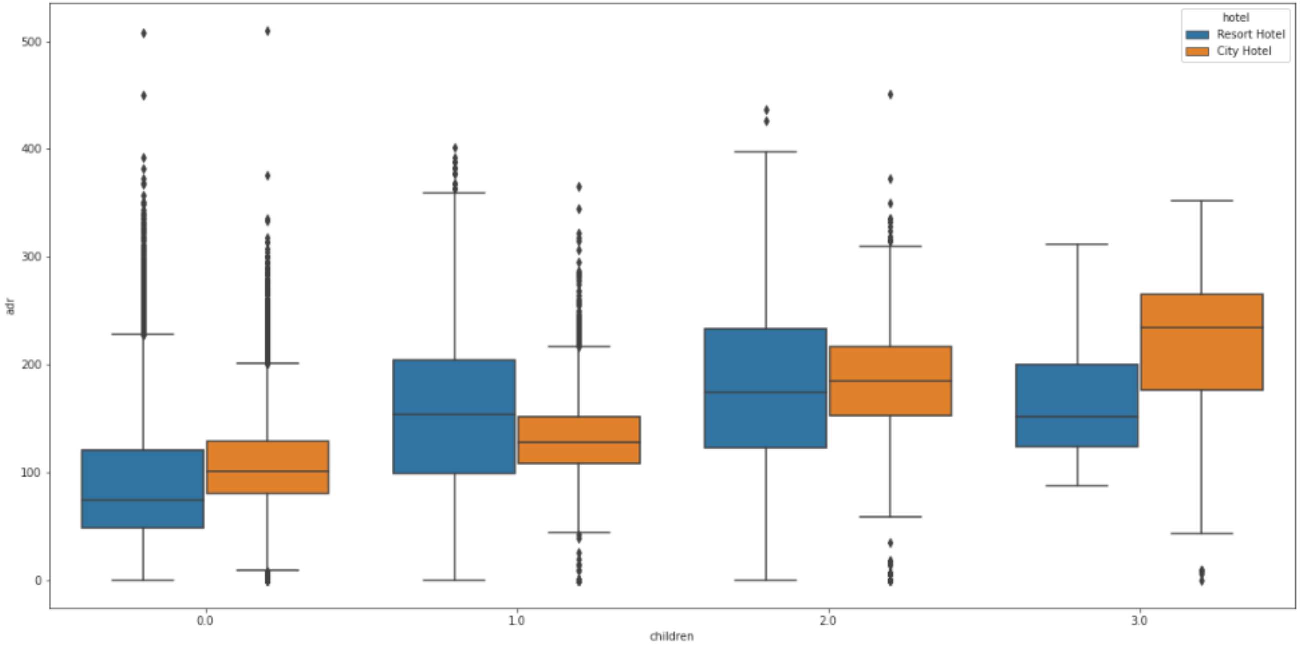
- Also when the total stay increases adr decreases adr is having negative correlation with total stay longer time customer stay lowest adr customer will get.

- Also adr having positive correlation with number of adults and children.

- When the total people increases adr is also increases adr and total people having a positive correlation from the above graph we can see that when there are 2 people adr is 100 but when there are 5 people adr is higher than 200.



- Also as number of children increases adr also increases in city hotel but in resort hotel number of children does not affect on adr.



THANK YOU