Capstone Project-1

Topic: Hotel Booking Analysis

Team : overview archives

Pradeep Kumar Yadav

Y Ishwar Rao

Ganesh Patil

Shashank Mishra

➢ Introduction :

Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week and many more.

This makes analyzing the patterns available in the past data more important to help the hotels plan better. Using the historical data, hotels can perform various campaigns to boost the business.

We can do EDA to predict the future bookings, most engaged months of coming year, additional facilities which can attract more customers and based upon the data, we can raise the revenue.

➢ Points to Discuss :

- ❑ Understanding the problem
- ❑ Agenda
- ❑ ETL pipeline
- ❑ Variable/Column(s)
- ❑ Data summary
- ❑ Univariate analysis
- ❑ Hotel wise analysis
- ❑ Distribution Channel wise analysis
- ❑ Booking cancellation analysis
- ❑ Timewise analysis
- ❑ Some important questions
- ❑ Correlation heat map
- ❑ Conclusion

➢ Understanding the problem :

While doing hotel-wise analysis of given hotel booking dataset, we answered following questions:

1. Percentage of bookings in each hotels?
2. Which hotel makes more revenue?
3. Which hotel has higher lead time?
4. What is most preferred stay length in each hotel?
5. For which hotel, does people have to wait longer to get a booking confirmed?
6. Which hotel has higher booking cancellations rate?
7. Which hotel have higher and how much customer returning rate?

Agenda :

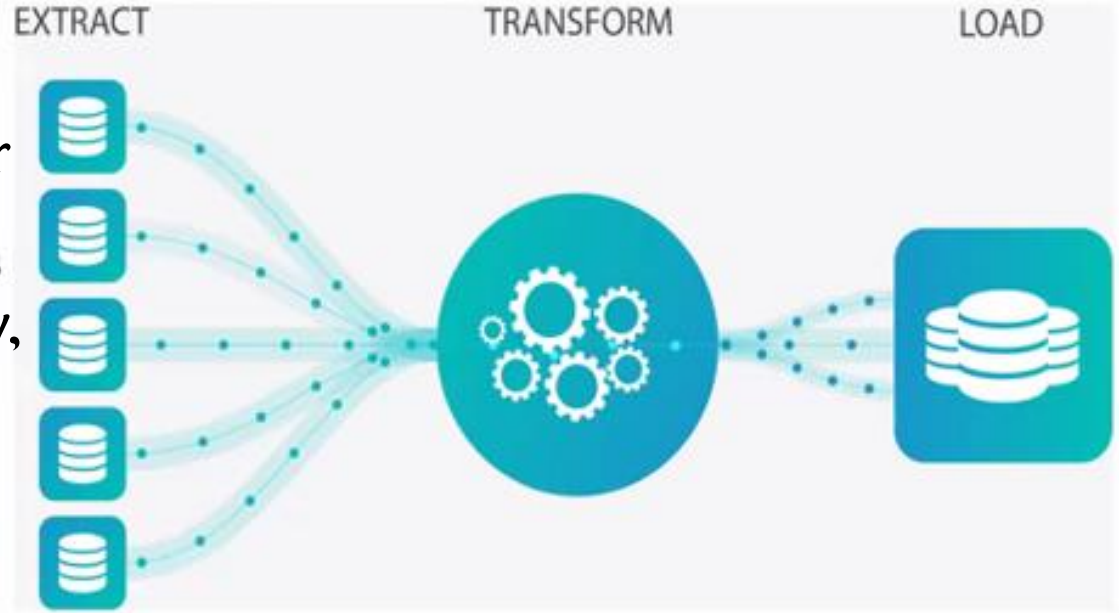
To discuss the analysis of given hotel bookings data set from 2015-2017.

We'll be doing analysis of given data set in following ways :

- Univariate analysis
- Hotel wise analysis
- Distribution Channel wise analysis
- Booking cancellation analysis
- Timewise analysis

By doing this we'll try to find out key factors driving the hotel bookings trends.

➢ ETL pipepline :



Extracting data from the source databases.
Transforming data to match a unified format for specific business purposes.
Loading reformatted data to the storage (mainly, data warehouses).

❑ A **data pipeline** is basically a set of tools and processes for moving data from one system to another for storage and further handling.

❑ Captures datasets from multiple sources and inserts them into some form of database.

❑ Constructing data pipelines in the core responsibility of data engineering.

We also created our project based upon this process as it is being part of Big data, very useful these days.

➢ Variable/Column(s) :

Let us try to analyse the type of data stored in various columns:

1. Columns/Variables having textual values(Categorical data):

i. hotel: type of hotel ('Resort Hotel', 'City Hotel')

ii. arrival_date_month : arrival calendar month.

iii. meal: The values include
- RO: Room only
- BB: Bed & Breakfast
- HB: Half Board (Breakfast and Dinner normally)
- FB: Full Board (Beakfast, Lunch and Dinner)
- SC: Self-catering

iv. Country: Names of countries in short form.

v. market_segment: 'Direct', 'Corporate', 'Online TA'(Traveling Agency), 'Offline TA/TO', 'Complementary', 'Groups', 'Undefined', 'Aviation'

vi. distribution_channel: It shows us the sector through which we got the bookin g like, 'Direct', 'Corporate', 'TA/TO', 'Undefined', 'GDS'.

vii. is_repeated_guest: guest repaeted or not. 1 for yes, 0 for No.

viii. reserved_room_type: Wing of the room like 'A', 'B", etc. for reserved rooms.

ix. assigned_room_type: Wing of the room like 'A', 'B", etc. for assigned rooms.

x. deposit_type: type of deposit like 'No Deposit', 'Refundable', 'Non Refund'.

xi. customer_type : type of the customer like transient, etc.

xii. reservation_status: values are 'Check-Out', 'Canceled', 'No-Show'.

xiii. reservation_status_date: date of reservation in textual format.

All other columns have data is in numerical form i.e. Numerical data.

*'adr' stands for average daily rate, which measures the average rental revenue earned for an occupied room per day.

Data Summary :

Given data set has different columns of variables crucial for hotel bookings. Some of them are:

- Hotel : The category of hotels, which are two resort hotel 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.
- lead_time : The time between reservation and actual arrival.
- 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]

Data Summary(contd..) :

- Country : The origin country of guest.
- market_segment : This column show how reservation was made and what is the purpose of reservation. Eg, corporate means corporate trip, TA for travel agency.
- distribution_channel : The medium through booking was made.[Direct,Corporate,TA/TO,undefined,GDS]
- Is_repeated_guest : 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.)

➢ Univariate Analysis :

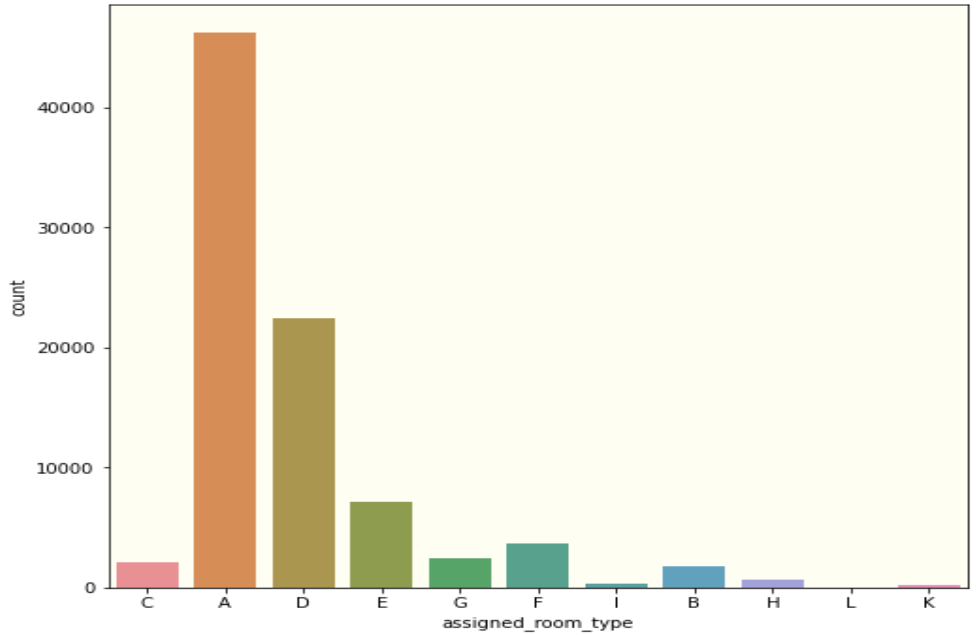
While doing univariate analysis of given hotel booking dataset, we answered following questions:

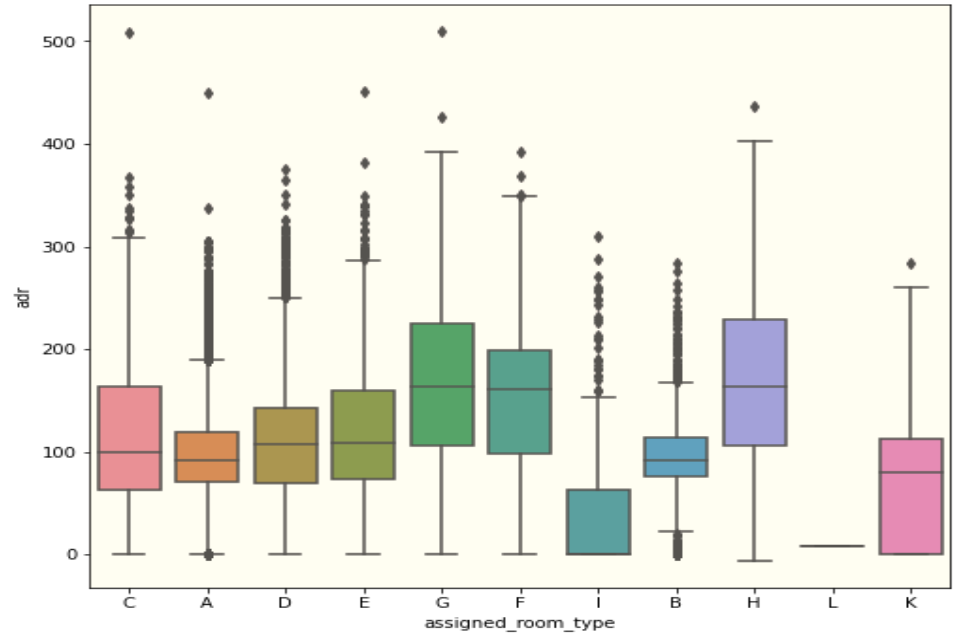
(1) Which agent made most of bookings?

(2) Which room type is in most demand and which room type generates highest adr?

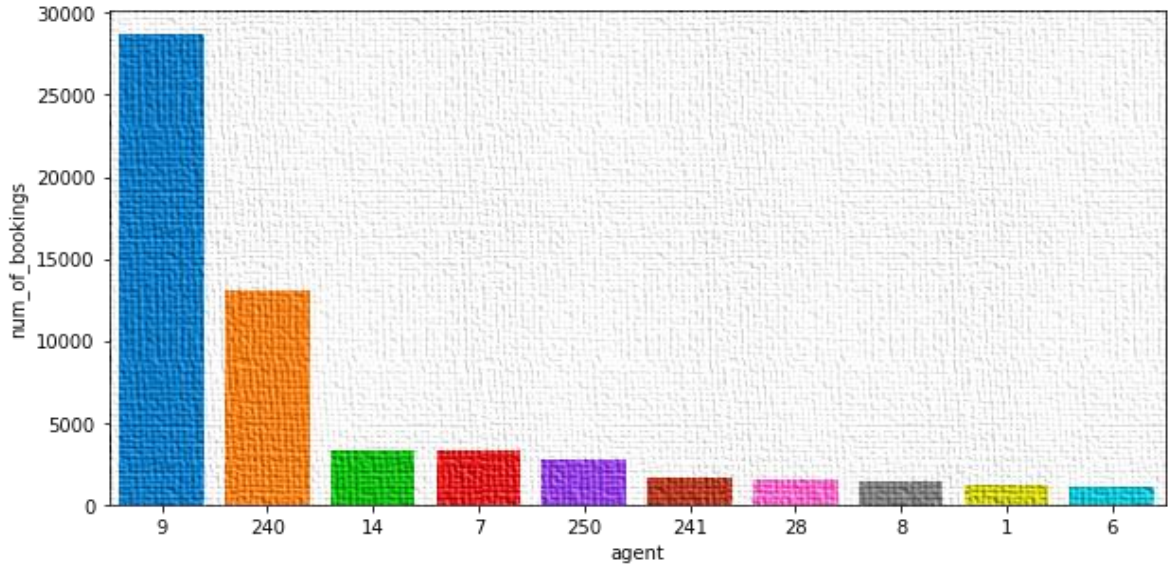
(3) From which country most of the customers are coming?

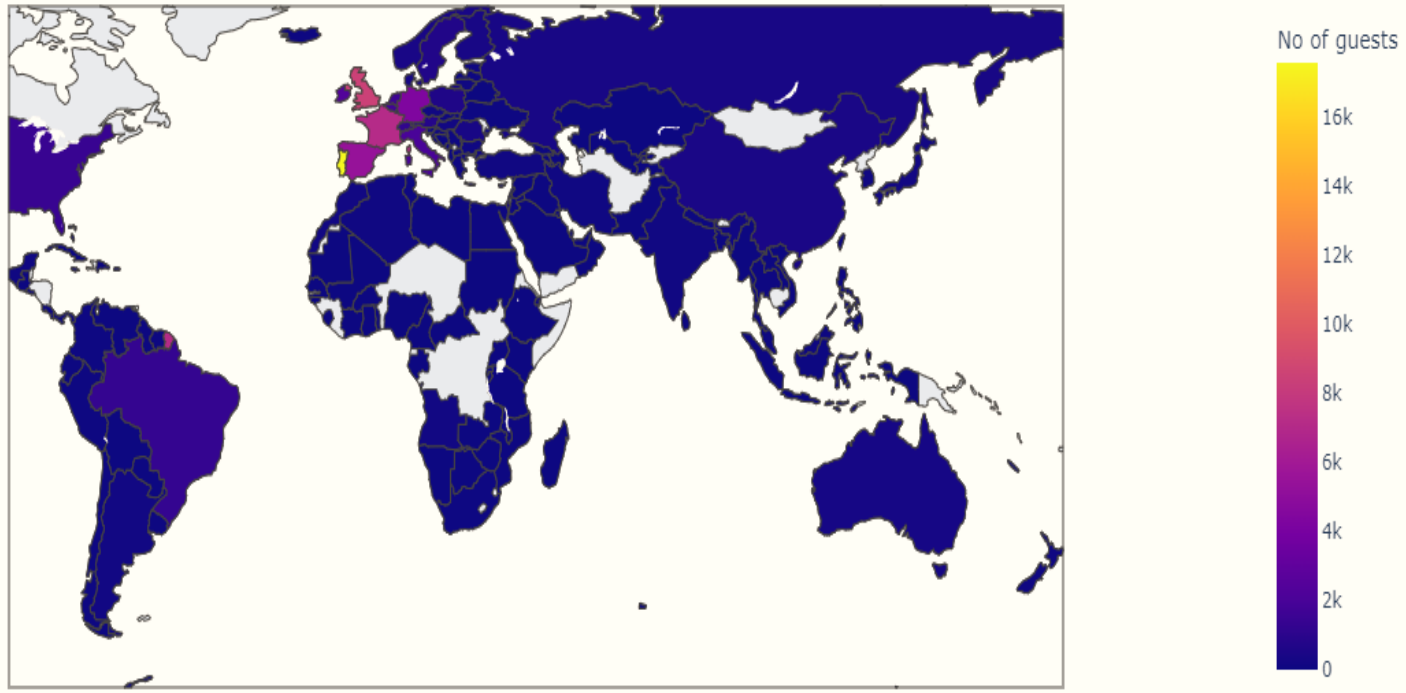
(4) What is the most preferred meal by customers?



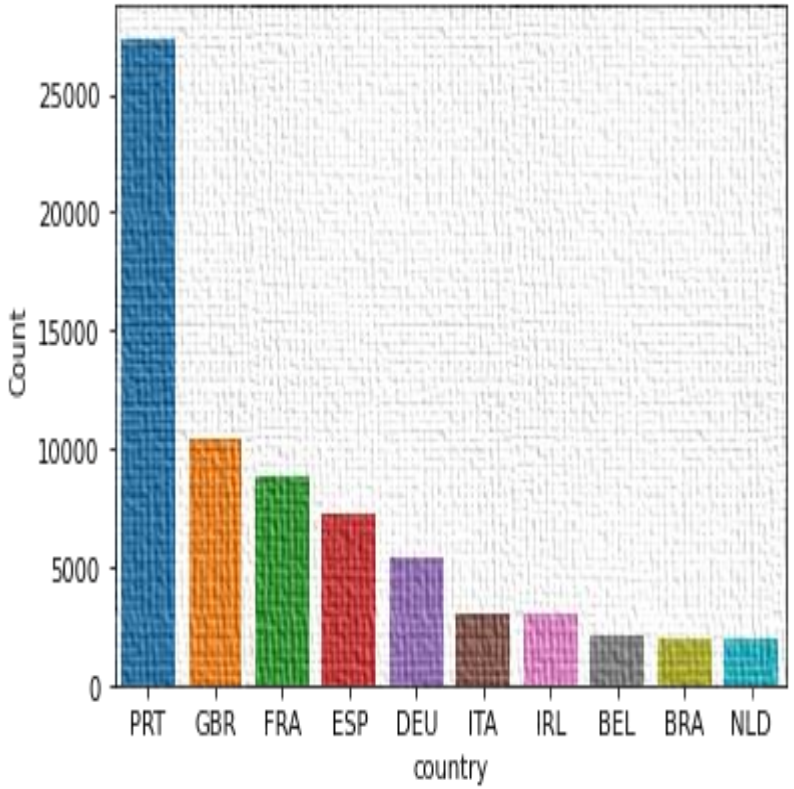


- Type A room is most demanded by customers.
- Room types C, G and H are some of the highest adr(average daily rate) generating rooms.
- Agent with id no. 9 made most of bookings.





- Most of the customers from European countries like Portugal, Great Britain, France and Spain.
- Most preferred meal type is BB(Bed and breakfast).



➢ Hotel wise Analysis :

While doing hotel-wise analysis of given hotel booking dataset, we answered following questions:

(1) Percentage of bookings in each hotels?

(2) Which hotel makes more revenue?

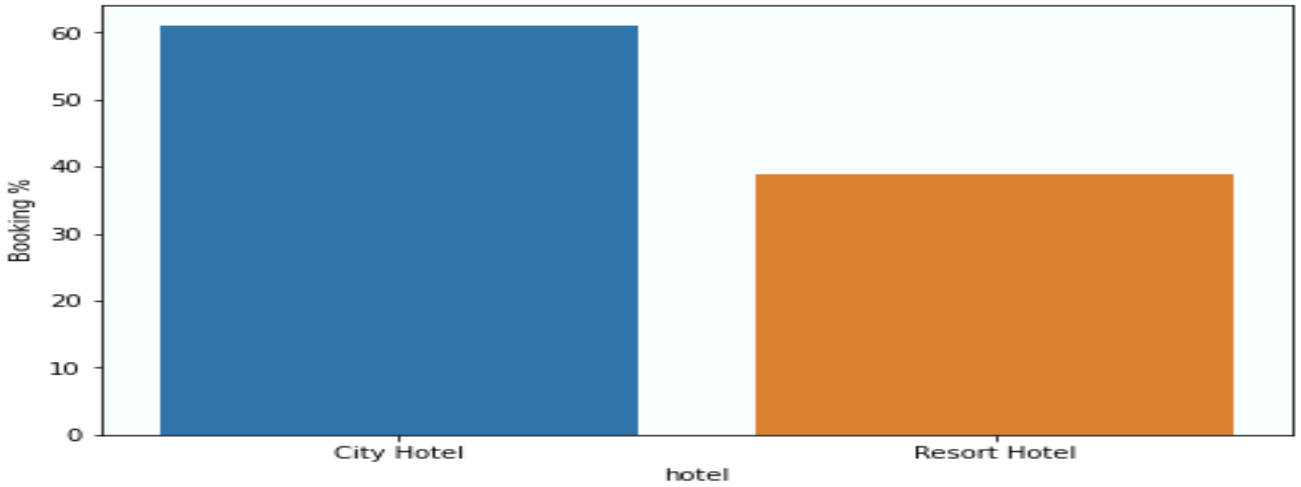
(3) Which hotel has higher lead time?

(4) What is most preferred stay length in each hotel?

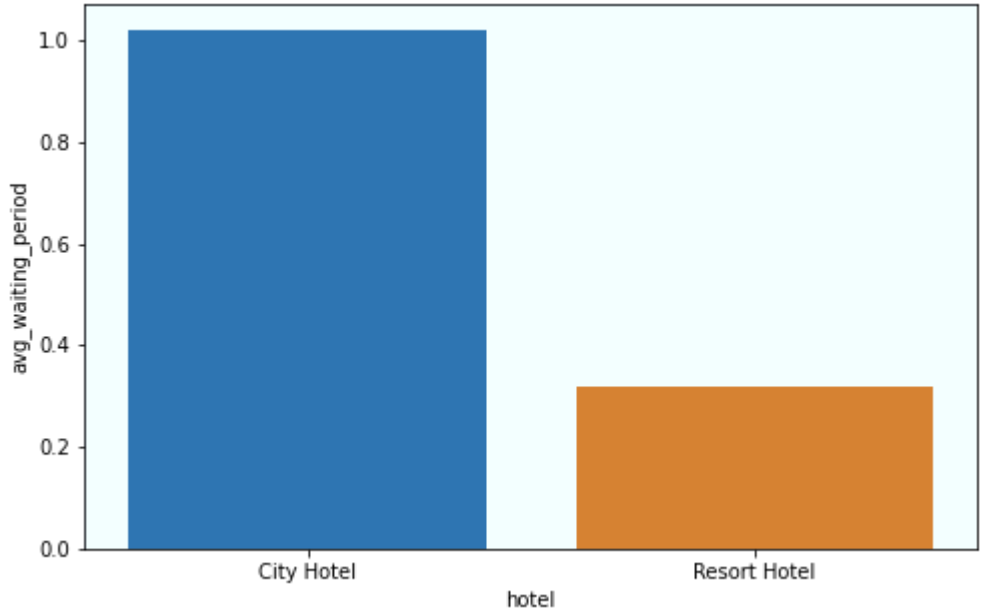
(5) For which hotel, does people have to wait longer to get a booking confirmed?

(6) Which hotel has higher booking cancellations rate?

(7) Which hotel have higher and how much customer returning rate?



- Around 60% bookings are for City hotel and 40% bookings are for Resort hotel.
- Avg adr of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue.
- City hotel has significantly longer waiting time, hence City Hotel is much busier than Resort Hotel.



- City hotel has slightly higher median lead time. Also median lead time is significantly higher in each case, this means customers generally plan their hotel visits way to early.

➢ Distribution channel wise Analysis :

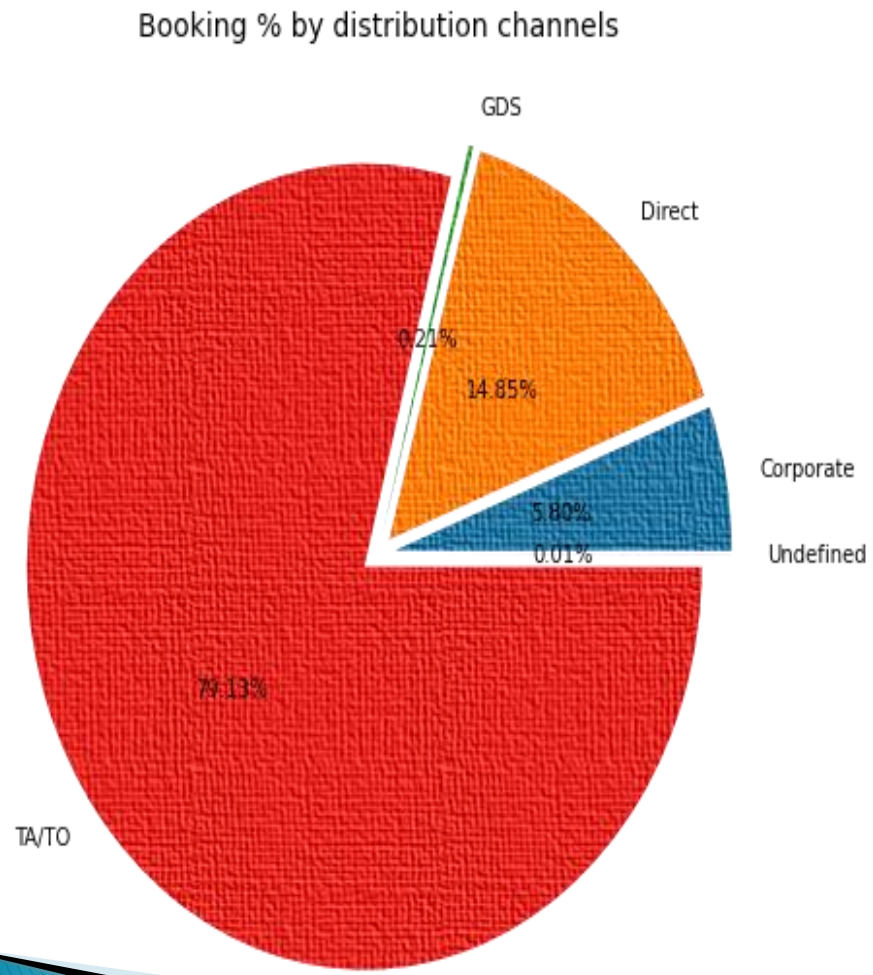
While doing Distribution channel wise analysis of given hotel booking dataset, we answered following questions:

(1)Which is the most common channel for booking hotels?

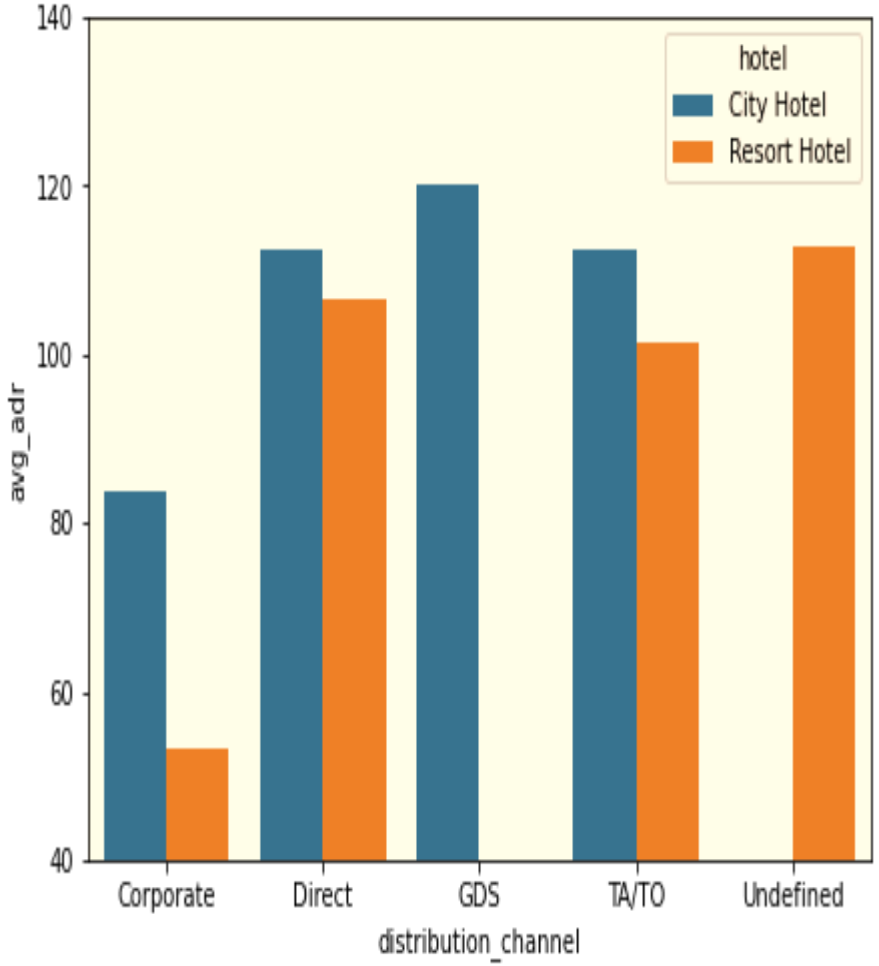
(2)Which channel is mostly used for early booking of hotels?

(3)Which distribution channel brings better revenue generating deals for hotels?

➢ Distribution channel wise Analysis :



- Here we can see that the most of guest are making reservation through TA/TO channels which is travel agency and tour operator.
- Than the second most used channel is direct.
- Channel which is mostly used for early booking of hotels is also TA/TO.



- GDS channel brings higher revenue generating deals for City hotel, in contrast to that most bookings come via TA/TO. City Hotel can work to increase outreach on GDS channels to get more higher revenue generating deals

- Resort hotel has more revenue generating deals by direct and TA/TO channel. Resort Hotel need to increase outreach on GDS channel to increase revenue.

➢ Booking cancellation Analysis :

We analyze the following possible reasons for booking cancellations:

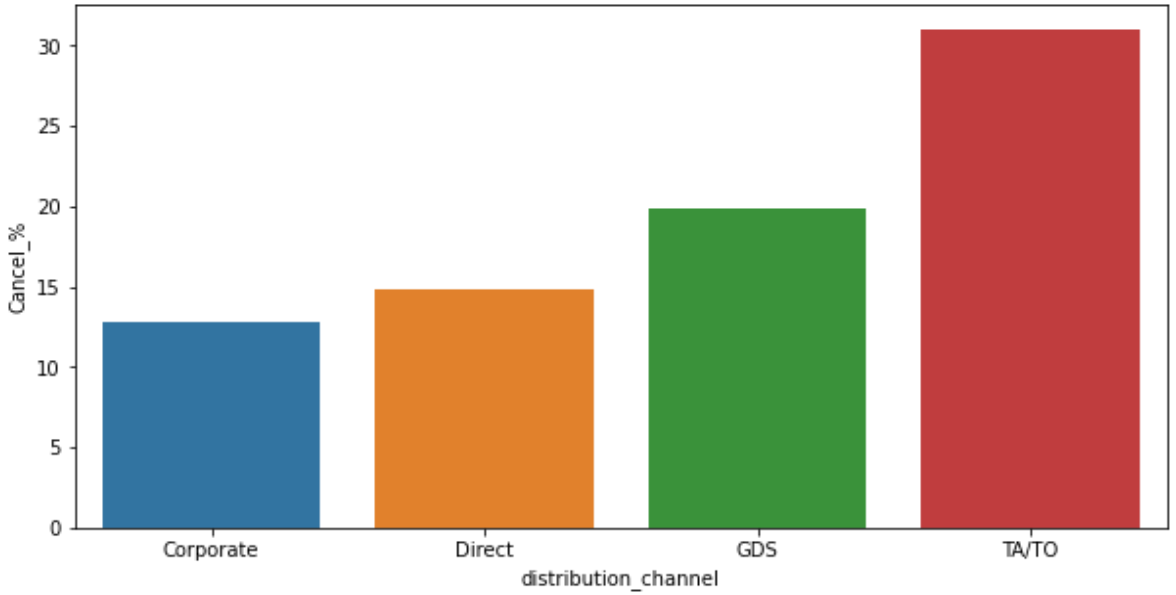
(1) Which significant distribution channel has highest cancellation percentage?

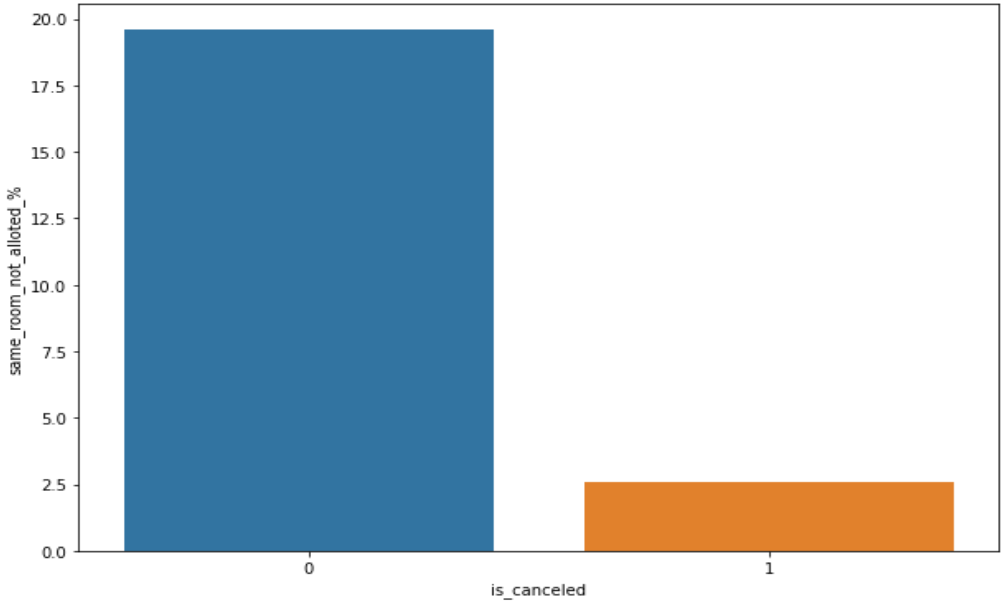
(2) Longer lead time.

(3) Longer time (in days) in waiting list.

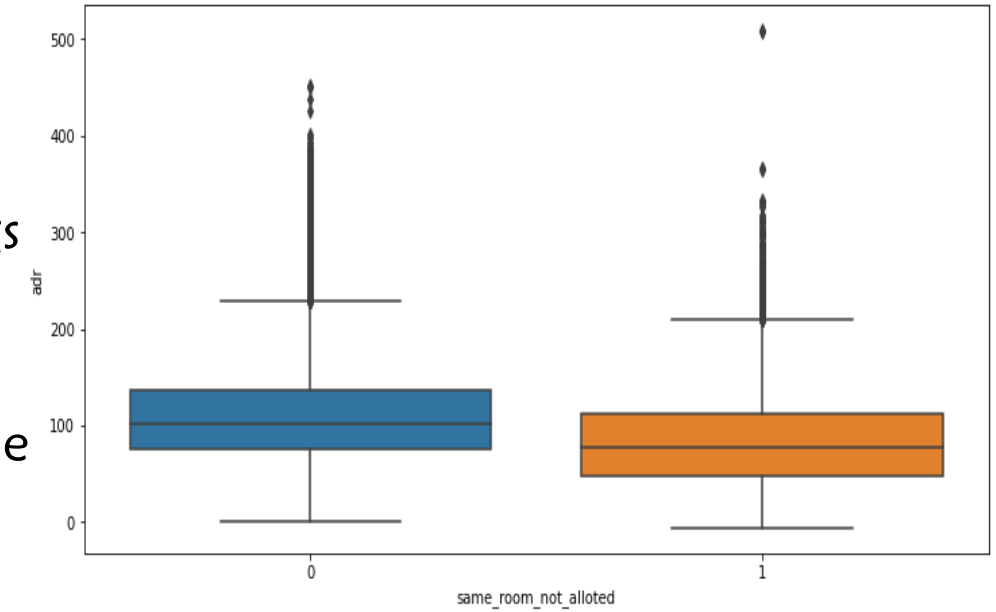
(4) Not getting same room as reserved.

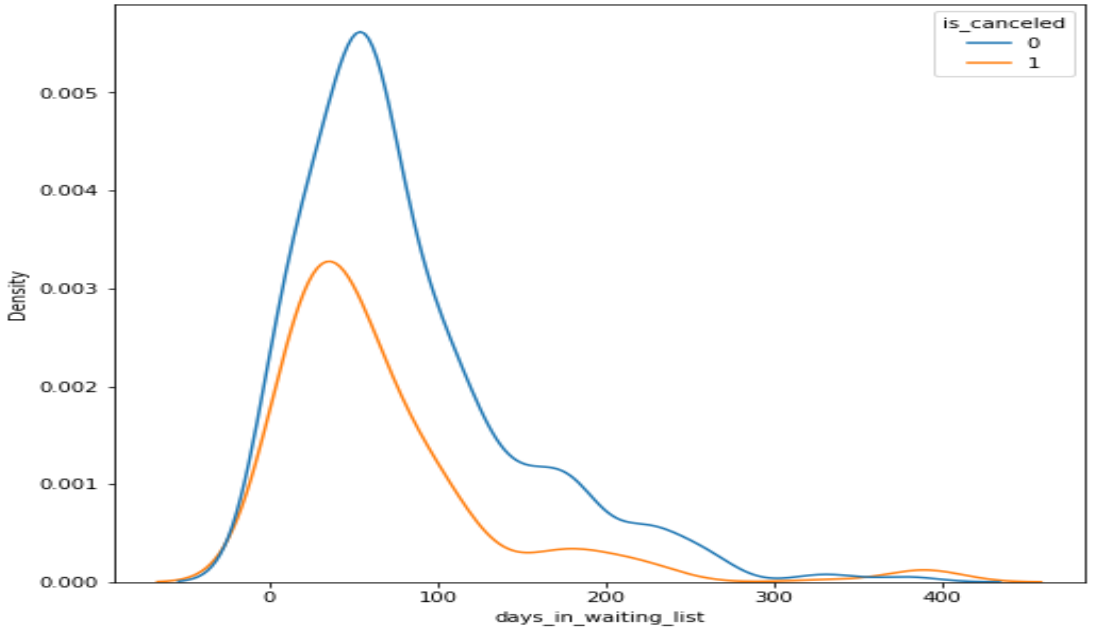
(5) Does not getting same room as reserved effects adr?

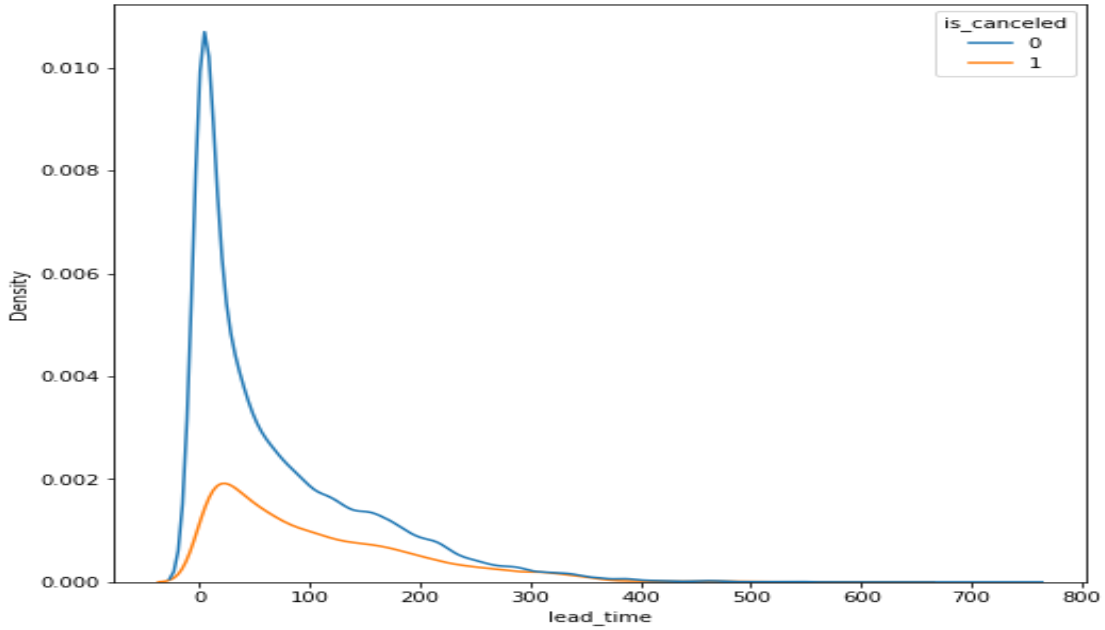




- TA/TO has highest booking cancellation %. Therefore, a booking via TA/TO is 30% likely to get cancelled.
- Not getting same room as demanded is not the case of cancellation of rooms. A significant percentage of bookings are not cancelled even after getting different room as demanded.
- But, customers who didn't got same room have paid a little lower adr, except for few exceptions.







- 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 of less than 150 days. Hence this shows that waiting period has no effect on cancellation of bookings.
- Also, lead time has no effect on cancellation of bookings, as both curves of cancellation and not cancellation are similar for lead time too.

➢ Time-wise Analysis :

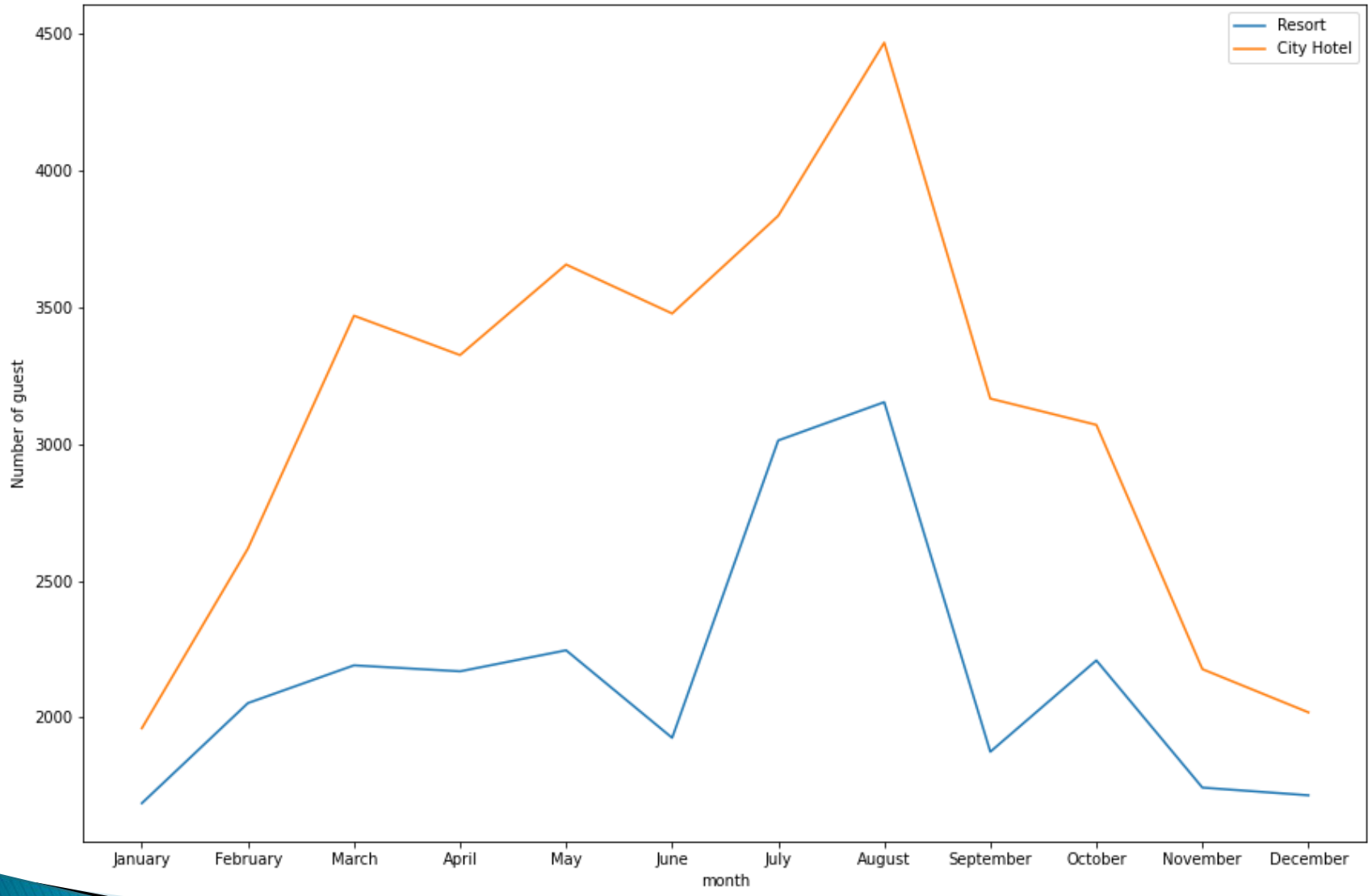
While doing time-wise analysis of given hotel booking dataset, we answered following questions:

(1)What are the most busy months for hotels?

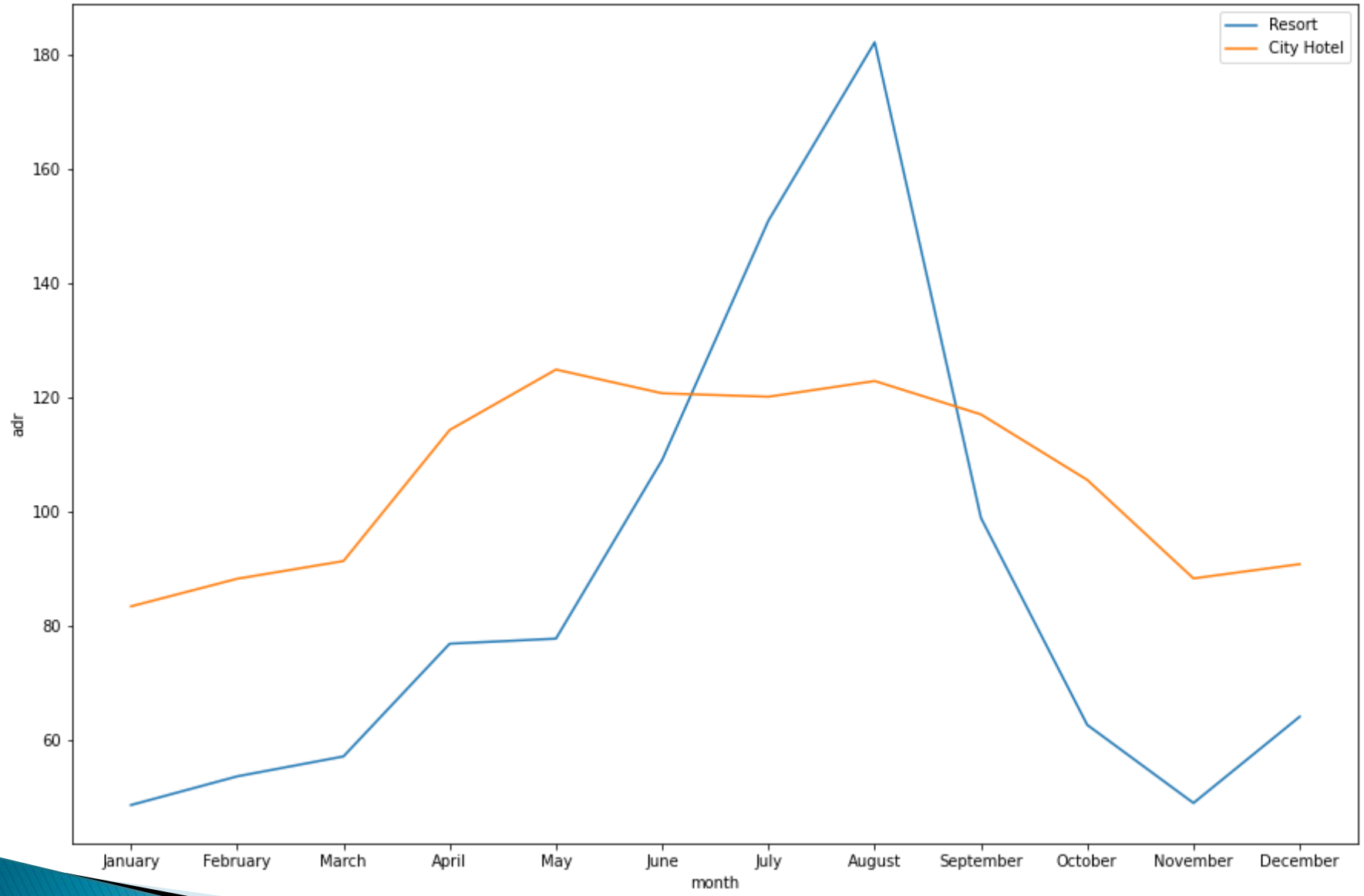
(2) In which months hotels charges higher adr?

(3)How does booking numbers and adr changes within a month?

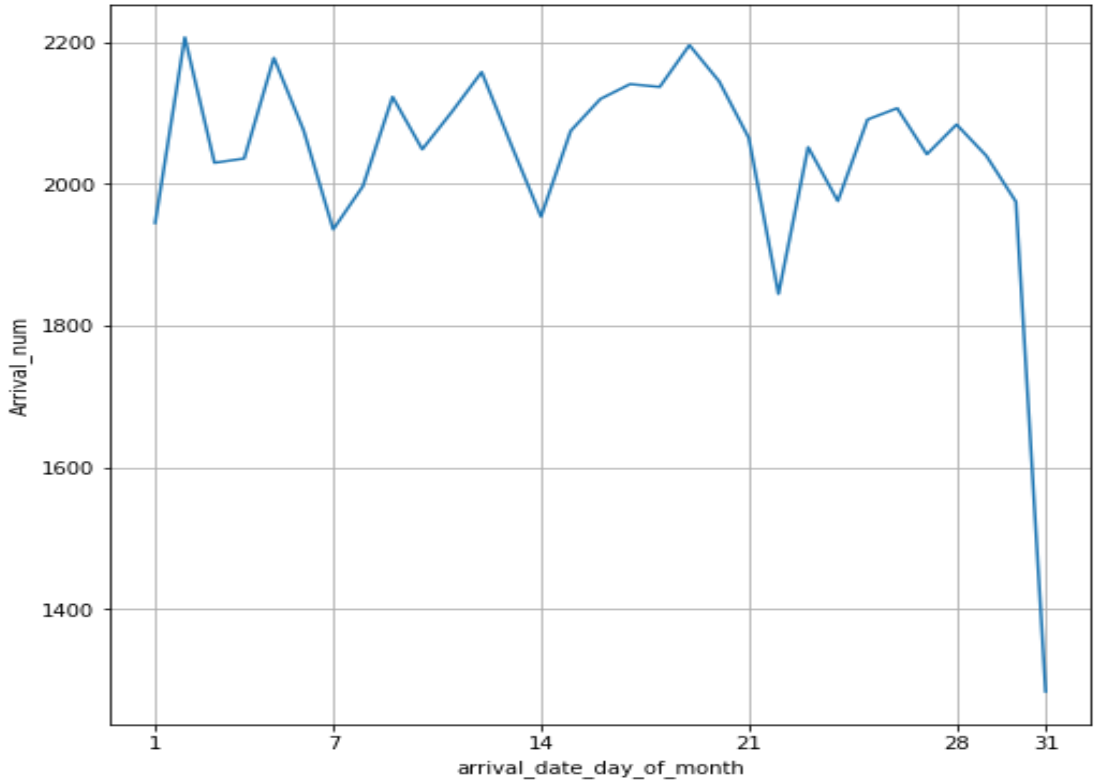
(4)How does bookings varies along year for different types of customers.

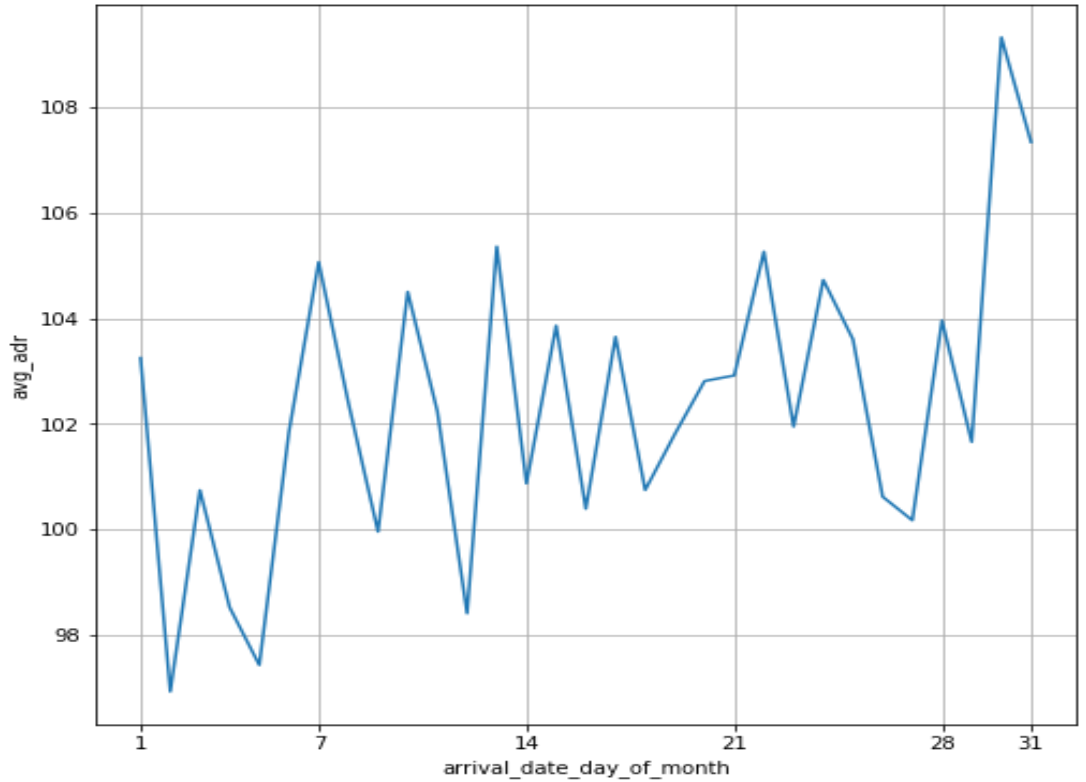


- From the month of July to August the number of bookings increased and in August, City Hotel got most number of guests.



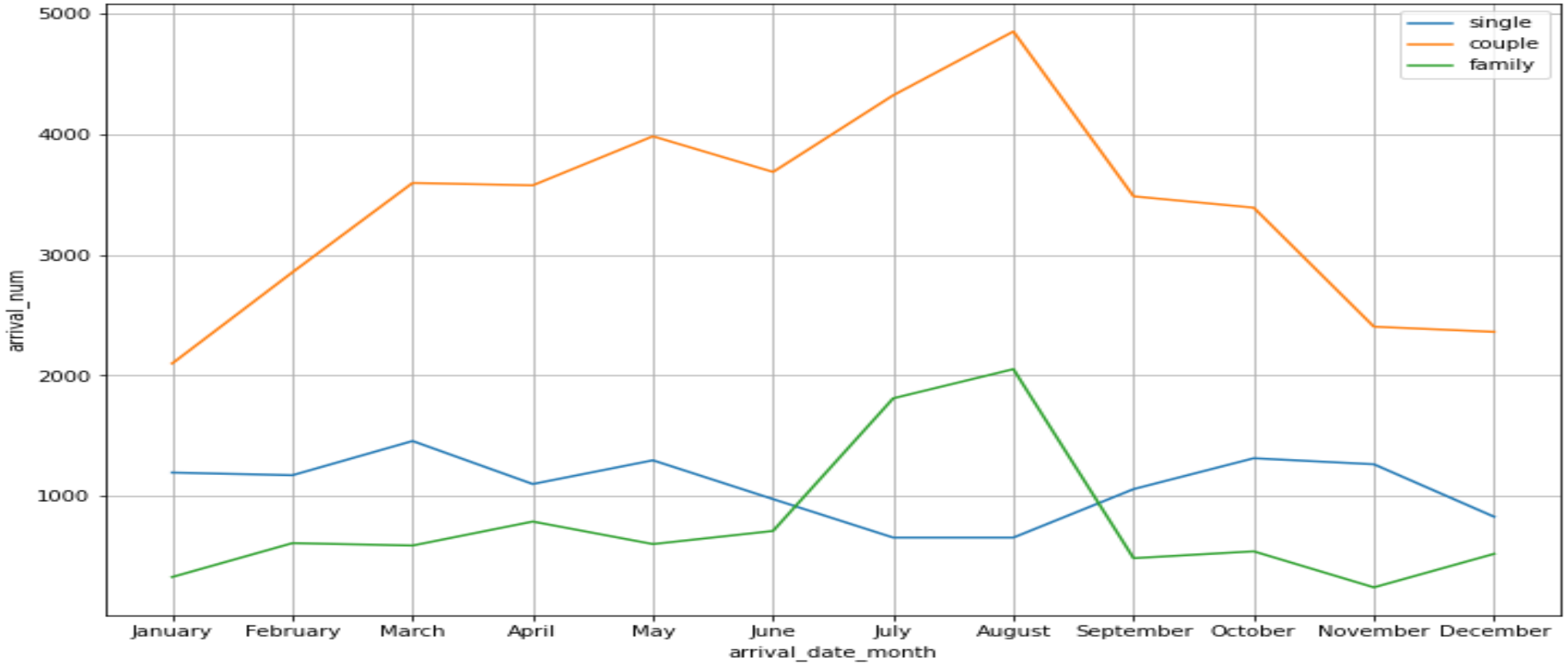
- The revenue aspect looks different, the Resort Hotels receives more revenue with respect to City Hotel. From May to August there was rapid increase in adr. August recorded the highest.





We can see that graph Arrival_num has small peaks at regular interval of days. This can be due to increase in arrival weekend.

Also, the avg adr tends to go up as month ends. Therefore charges are more at the end of month.



Mostly bookings are done by couples.

It is clear from graph that there is a sudden surge in arrival num of couples and family in months of July and August. So better plans can be planned accordingly at that time for these type of customers.

Some important questions

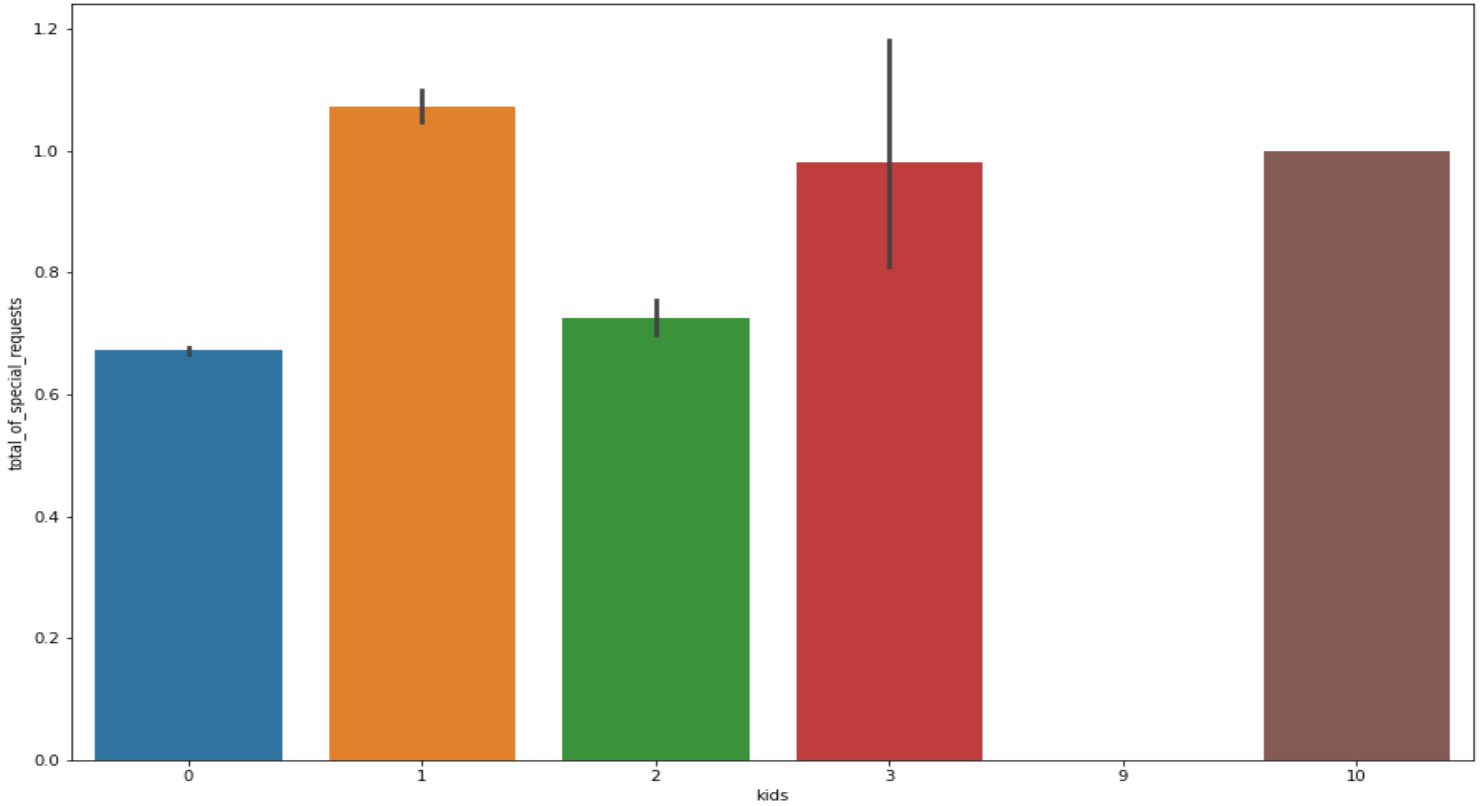
Some other analysis are also done, which are as follows:

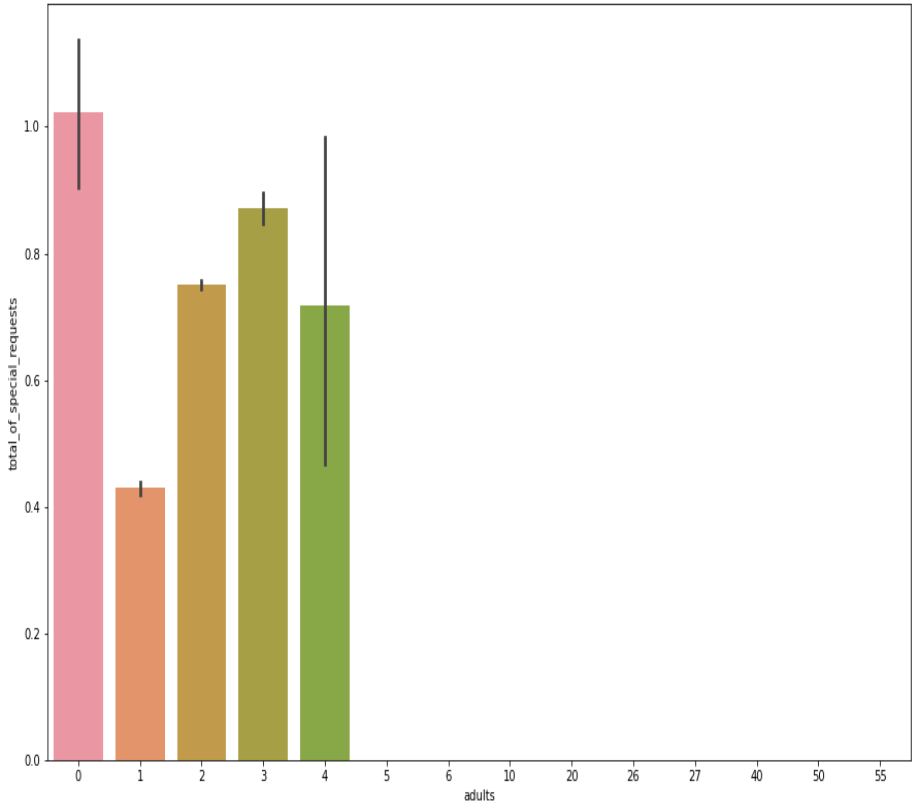
(1)What are the different reason for special requests?

(2)What is the optimal stay length for better deal for customers?

(3)How adr is affected by total staying period in hotels?

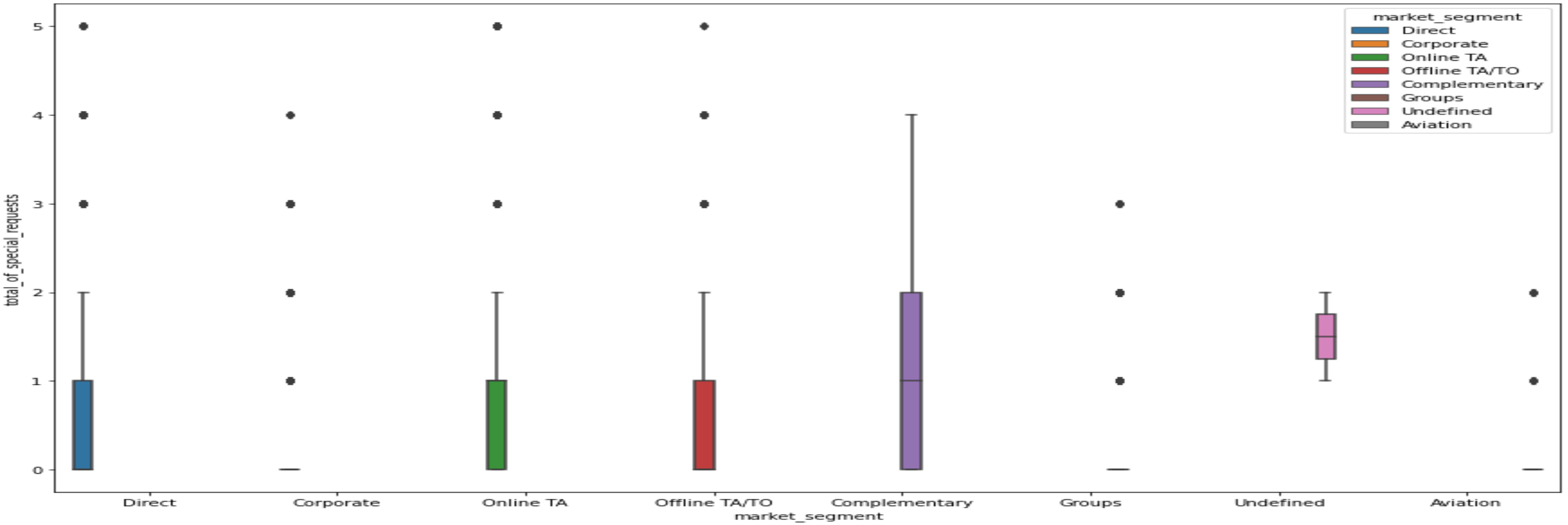
➢ Reasons for special requests :





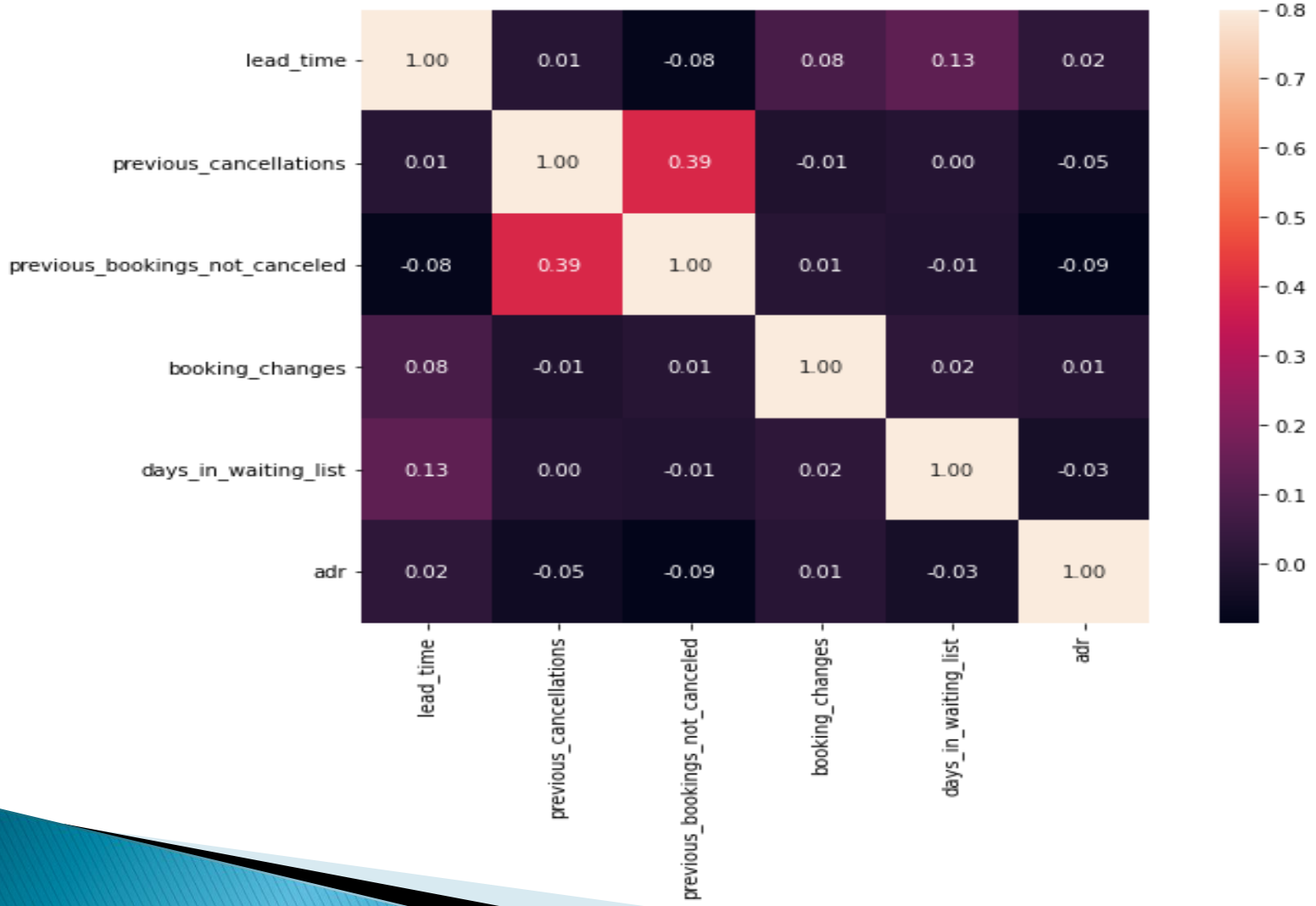
The number of special request are almost the same in the kids section. But, we can see that if the adults are more than 2 there are more chances that hotels will receive more special requests.

➢ Reasons for special requests(cont.) :



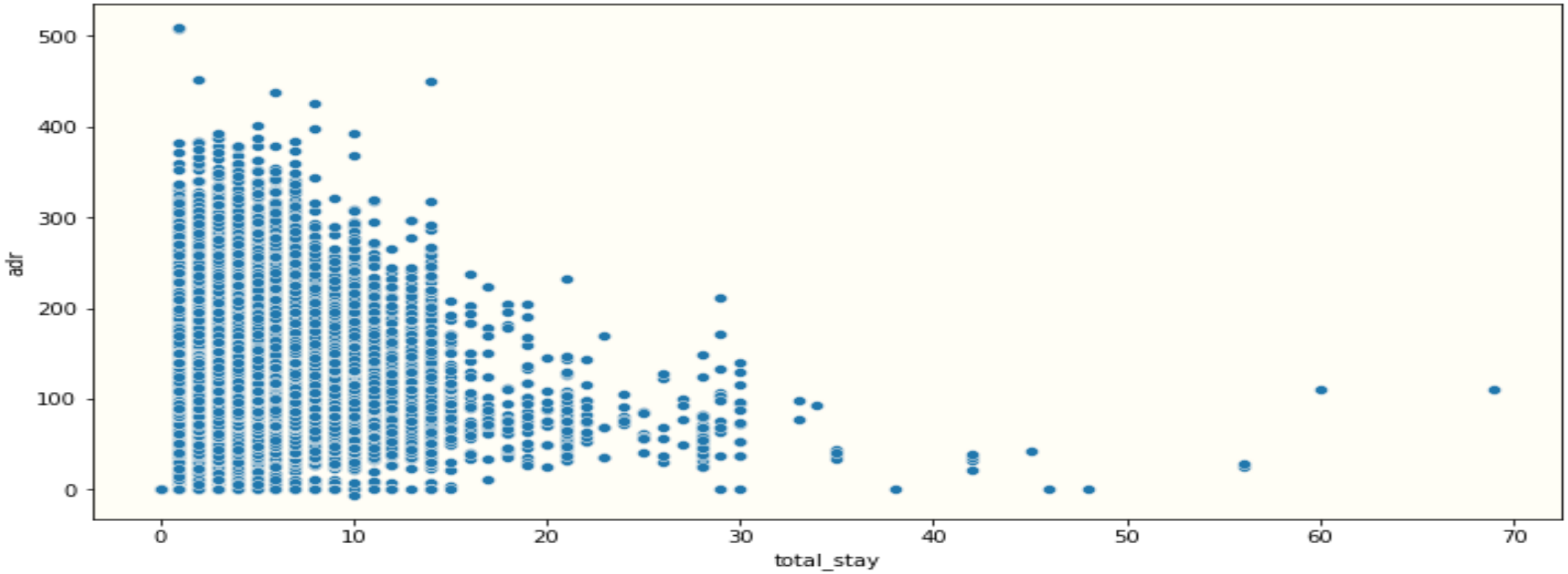
Here we can see that all market segment mostly have special request.
There is one segment which is complementary, having more than average number of special request.

Correlation Heatmap :



- Total stay length and lead time are slightly correlated. 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 service to deliver, therefore more adr.

➢ Optimal stay length for better deals in adr :



For shorter stays the adr(average daily rate varies greatly) but for longer stays (> 15 days) adr is comparatively very less. Therefore, customers can get better deal for longer stays more than 15 days

➢ Conclusion :

1. Around 60% bookings are for city hotels and around 40% bookings are for resort hotels therefore city hotel is busier than resort hotel , also overall adr of city hotel is slightly higher than resort hotel.
2. The majority of reservations are for city hotels.
3. The number of repeated guests is too low.
4. Most of the bookings either in the canceled or checkout done by online TA.
5. City hotels and resort hotels maximum number of bookings by online TA.
6. That aviation industry has the minimum number of days on the waiting list.
7. More visitors are from western europe, namely Portugal, France, Great Britain, and Spain being the highest.
8. Families with children have no particular preference for the hotel type.

➢ Few more Conclusions… :

9. August and July are the most profitable and busiest months for both the hotels.
10. The confirmed bookings goes from their lower value in january to their highest value in august.
11. Transients are the most common customer type, they represent 75% of the total customers.
12. There is a disproportionate amount of cancellations on hotel bookings . Bookers are not required to send in a deposit in most bookings which could explain the high rate of cancellations.
13. Data suggests that hotel business could be improved by targeting working travelers or improving daily rates for weekdays.

Thank You

For Your Attention!

Any Questions