Hyperiondev

TASK

Exploratory Data Analysis on Hotel Booking Analysis Dataset

Visit our website


HOTEL BOOKING

Introduction

The hospitality and tourism industry is a vast sector that includes all the economic activities that directly or indirectly contribute to, or depend upon, travel and tourism. This industry sector includes Hotels & Resorts. There are many decisions that a hotel operator needs to make. They need to choose where to promote their hotel and how resources must be allocated to engaging with various distribution channels. Prices need to be set for the guest accommodations and knowing what languages are needed at a hotel is helpful during the hiring process. Finally, revenue from the bookings needs to be determined.

The operator's questions can be summarized into many questions: Where are the hotel bookings coming from? What is the frequency of cancellations for my guests, and how likely are they to be no-shows? What is the trend in the rates that I can charge my guests? How much revenue is my hotel earning per booking, and how much is coming from each country?

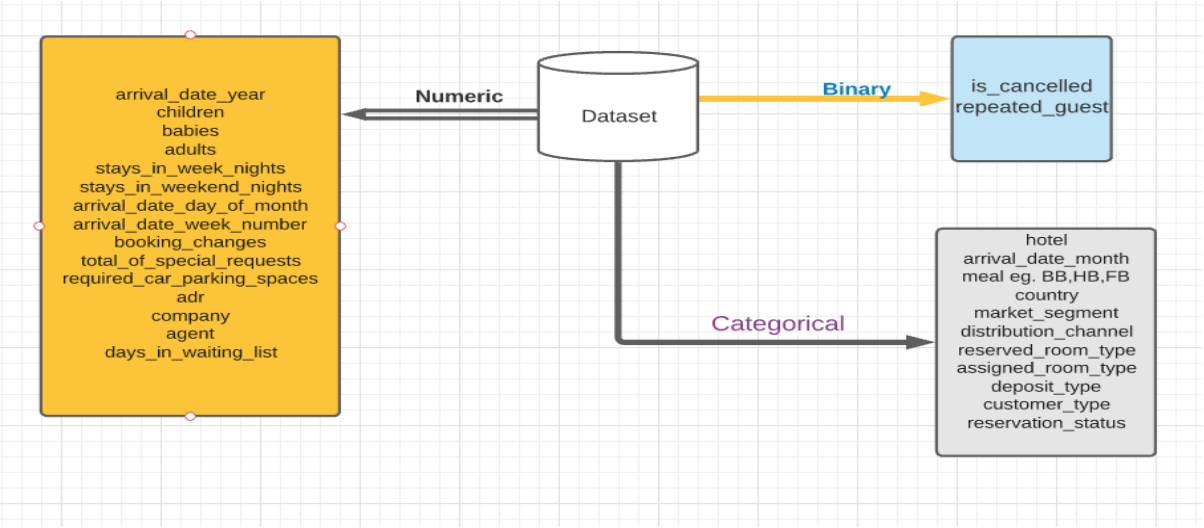
Agenda

To discuss the analysis of given hotel bookings data set from 2015 to 2017, we will be doing analysis of given data set in following ways:

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

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

Data Summary



market_segment: This column show how reservation was made and what is the purpose of reservation. Eg , corporate means corporate trip , TA/TO for travel agency or tour operators.

distribution_channel: The medium through booking was made. [Direct,Corporate,TA TO,undefined,GDS for global distribution systems].

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.)

DATA CLEANING

Cleaning data is crucial step before Exploratory Data Analysis (EDA) as it will remove the ambigous data that can affect the outcome of EDA. First and foremost, the data needs to be read and understood before data cleaning and analysis can occur.

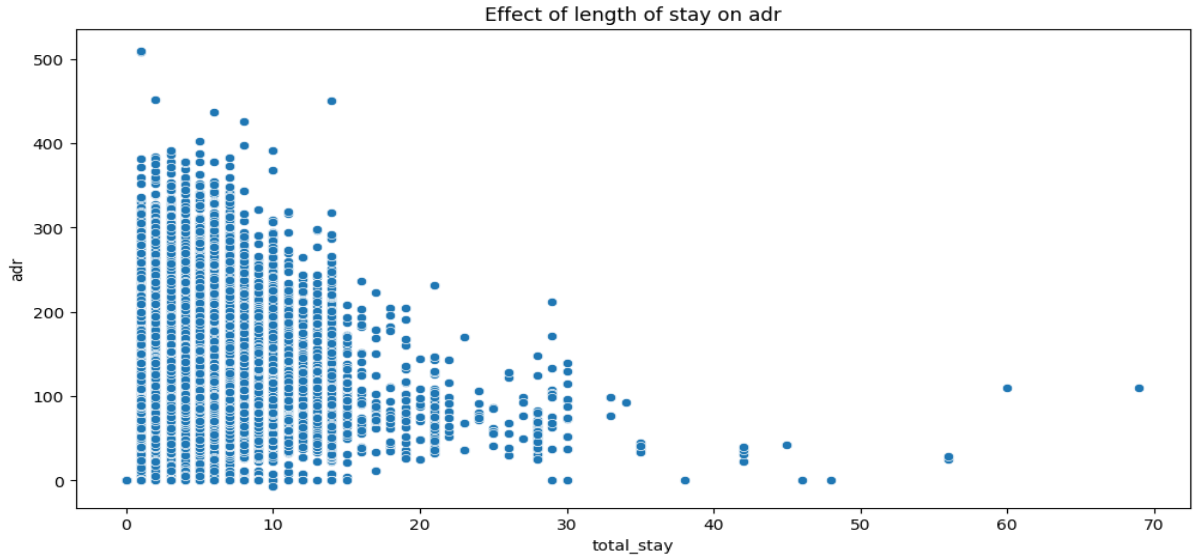
While cleaning data we will perform following steps:

1. Remove duplicate rows.
2. Handling-missing values.
3. Convert columns to appropriate datatypes.
4. Adding important columns.
5. Check if there are any outliers and see how we treat the outliers.

Firstly, we look at our dataset and display the first and last 5 records in the dataset to see the dataset fields and data populate. Secondly, we get information on the dimension, structure and description of the dataset. While cleaning data, I firstly had to check if there are any duplicates and we notice that there were 31994 duplicates and we simply dropped them so solve the problem.

I created new column total_stay by adding stays_in_weekend_nights + stays_in_week_nights so that we can analyse the stay length at hotels. Moreover, I also created new column total_people by adding adults + children + babies.

In addition, I had to identify which features are classified as continuous and categorical variables in the dataset and the notebook analysis has all of them. Lastly, I had to check if there are any outliers in the dataset and I notice that there were outliers in the dataset and this pose a problem. Therefore, I had to impute all the outliers found in the categorical and continuous variables. However, when I did an analysis to see if the length of stay affects adr, we notice that there is an outlier in adr, so we will remove that for better scatter plot.



MISSING DATA

Number of missing values found in each column, and we notice that the company column has 82 137 missing values, agent has 12 193 missing values, country has 452 missing values and children has 4 missing values. This poses a problem to our dataset and the missing values must be populated. I had to fill all these columns to solve the problem.

DATA STORIES AND VISUALISATIONS

Visualization is an important part of data analysis as it can be used to tell a story about your data and visualize data findings. It is important to perform an EDA in the running of business so that you will have a clear idea of how the business is running. It is important for hotel owners to do an EDA since its main purpose is to help look at data before making any assumptions. In performing an EDA on the two hotels, I will try to answer many questions.

Univariate Analysis

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

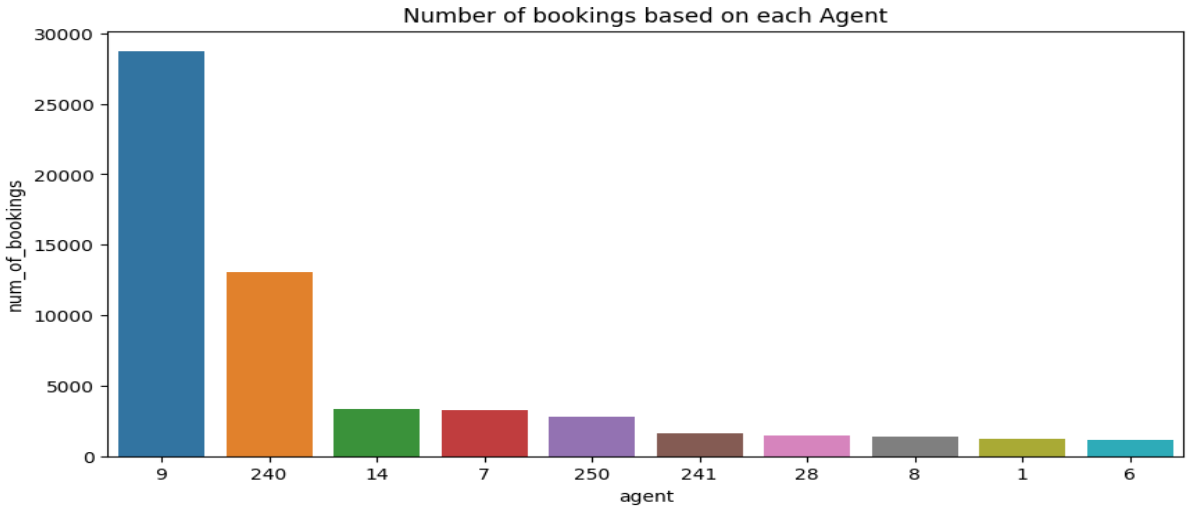
(1) Which agent made most of bookings?

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

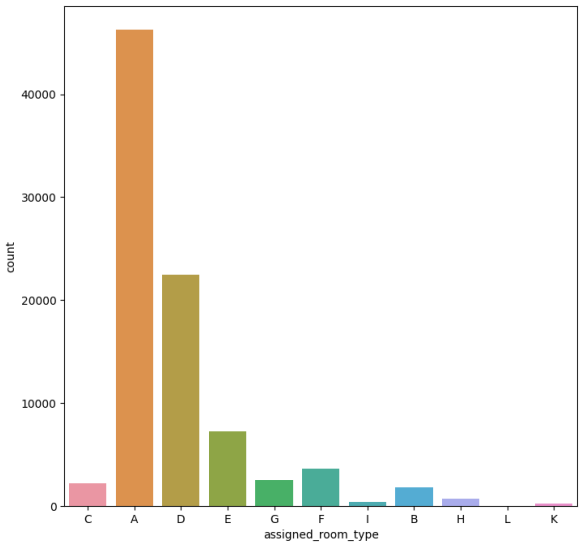
(3) From which country most of the customers are coming?

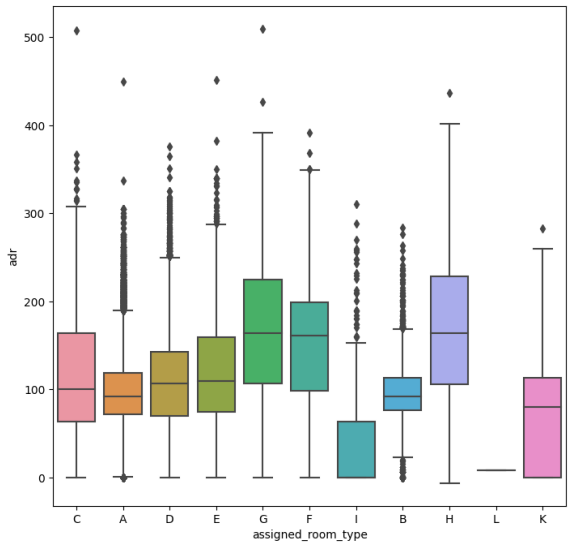
(4) What is the most preferred meal by customers?

(5) How long do people stay at the hotels?

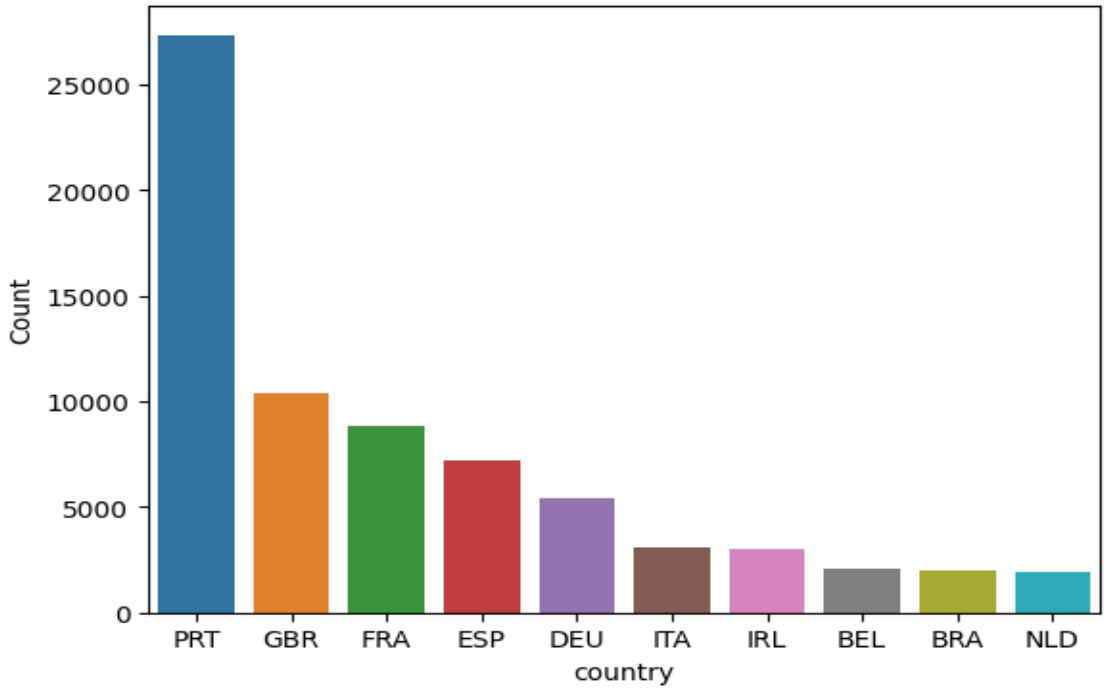


Agent no. 9 has made most number of bookings.

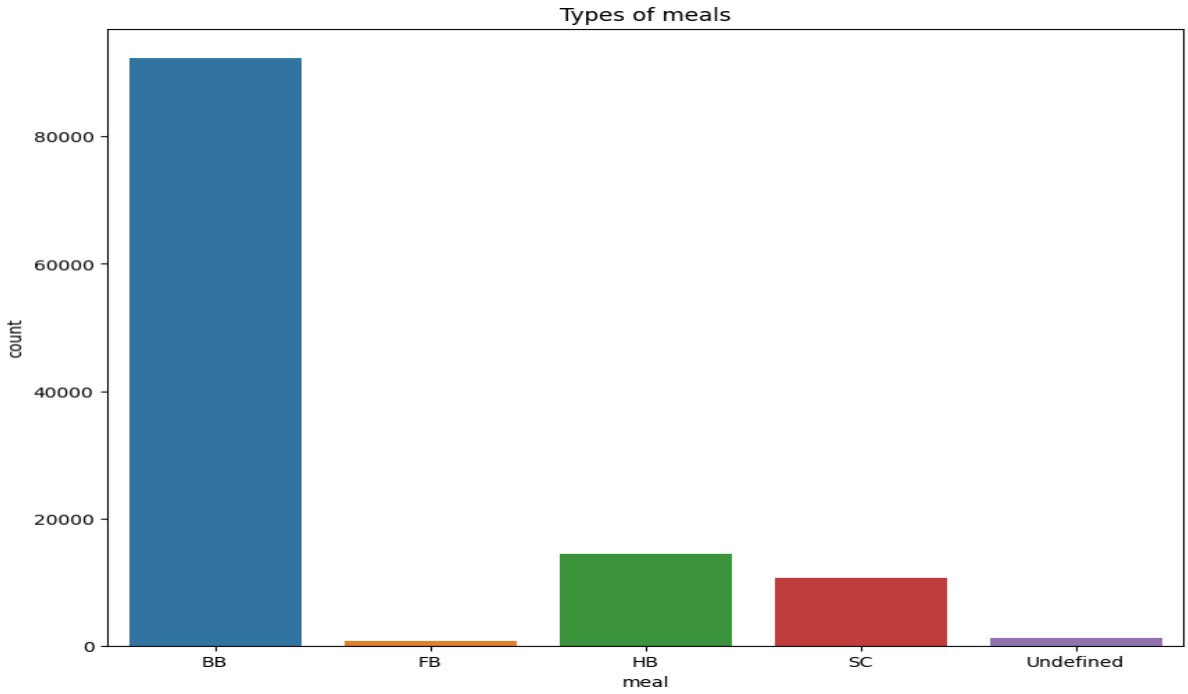




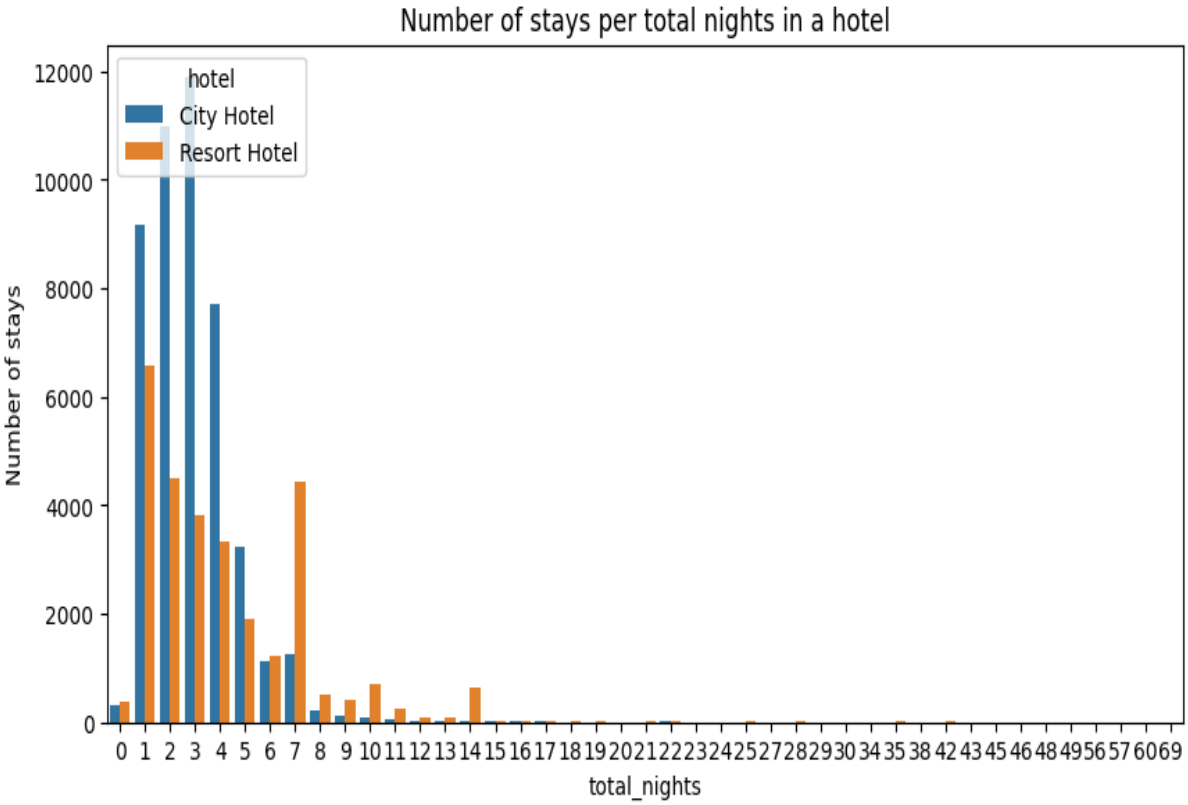
Most demanded room type is A, but better adr rooms are of type H, G and F also. Hotels should increase the no. of room types A and H to maximise revenue.



Most guests/visitors are from Western Europe, namely Portugal, Great Britain, France and Spain being the highest.



Most preferred meal type is BB(Bed and breakfast).



Most people prefer to stay at the hotels for <=5 days.

Hotel wise Analysis

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

(1) What is the percentage of bookings in each hotels?

(2) Which hotel makes more revenue?

(3) Which hotel has higher lead-time?

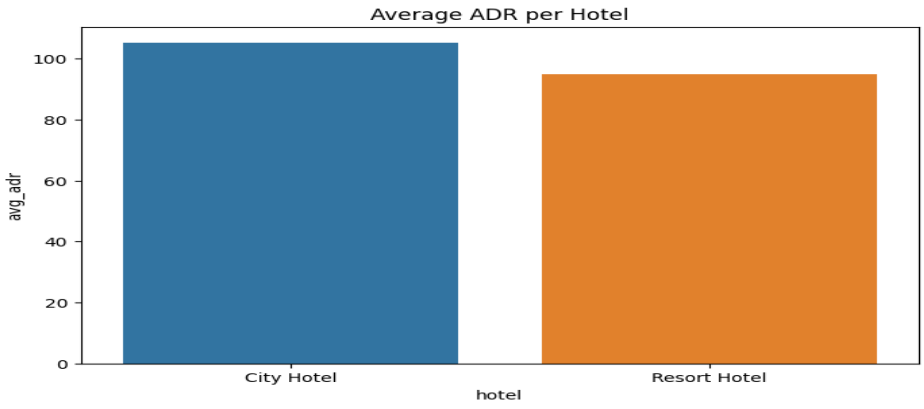
(4) What is most preferred stay length in each hotel?

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

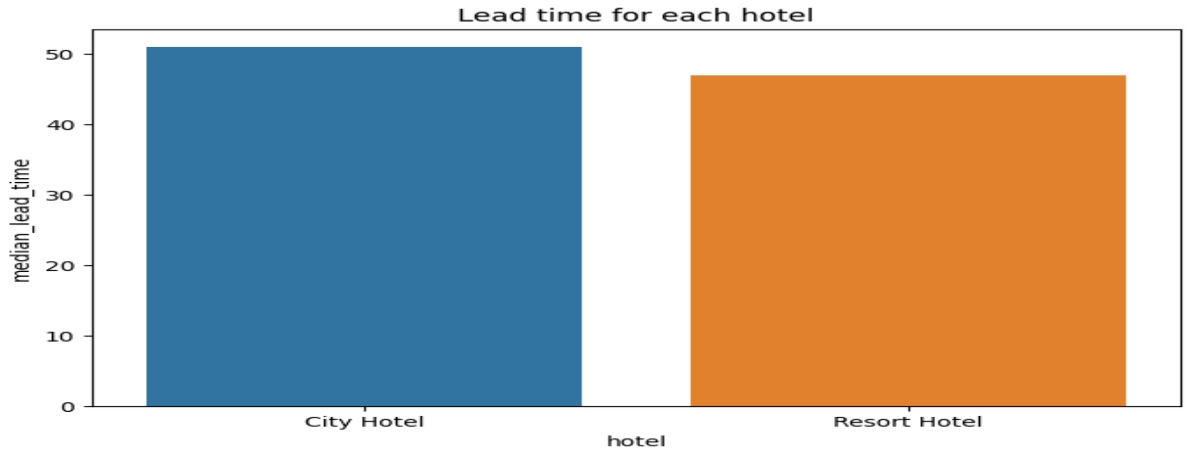
(6) Which hotel has higher booking cancellations rate?

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

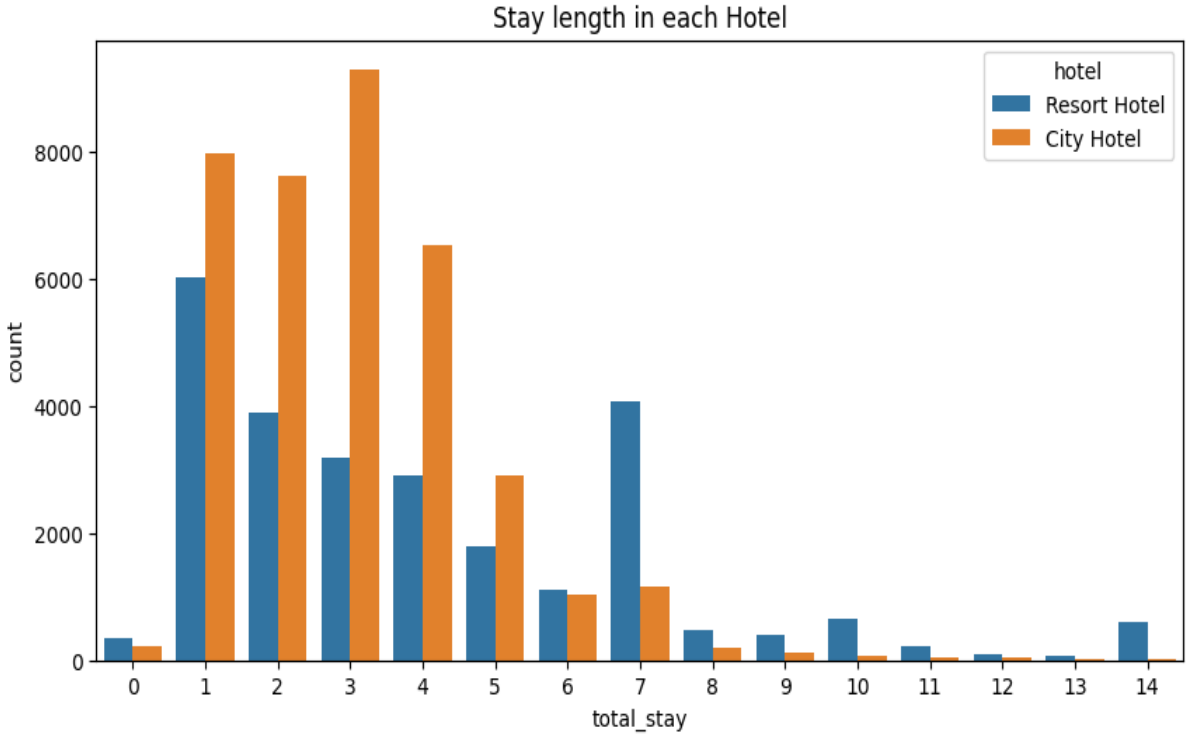
Around 62% bookings are for City hotel and 38% bookings are for Resort hotel.



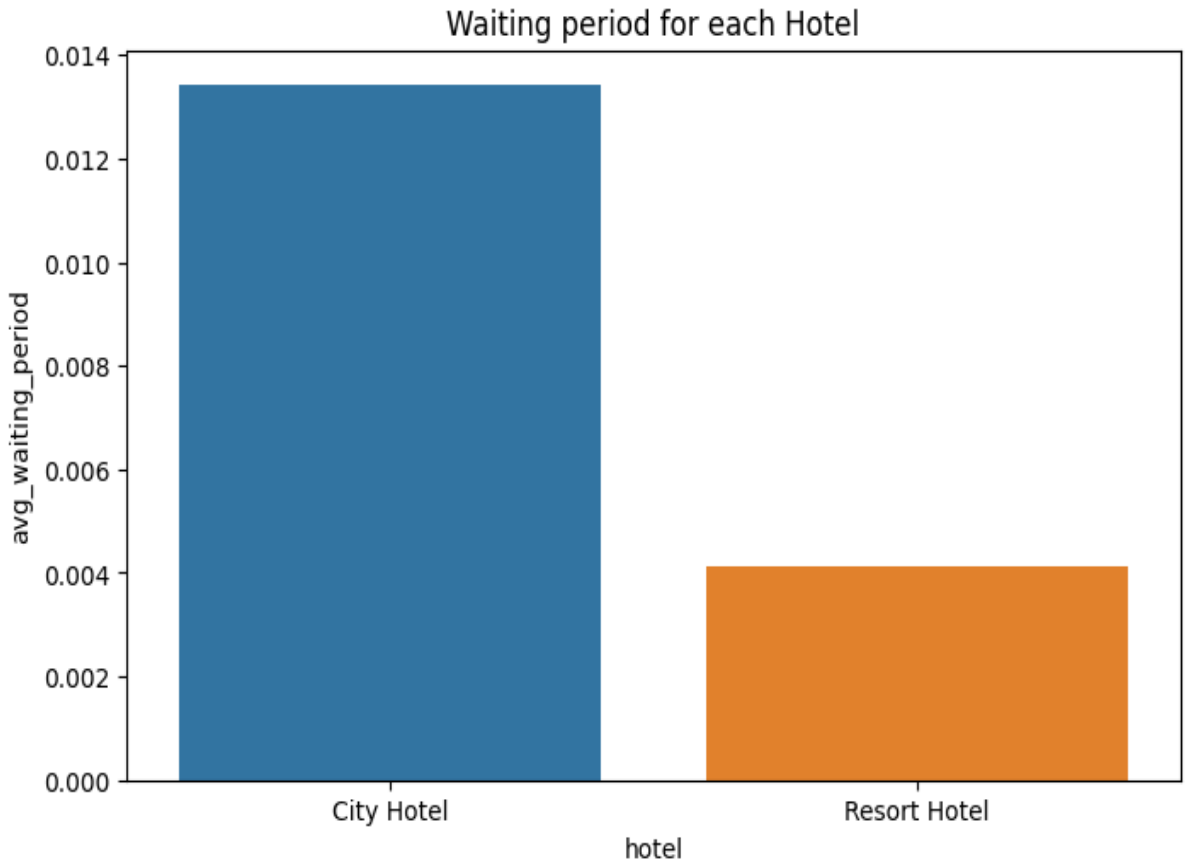
Avg adr of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue.



City hotel has slightly higher median lead_time. In addition, median lead_time is significantly higher in each case, this means customers generally plan their hotel visits excessively early.



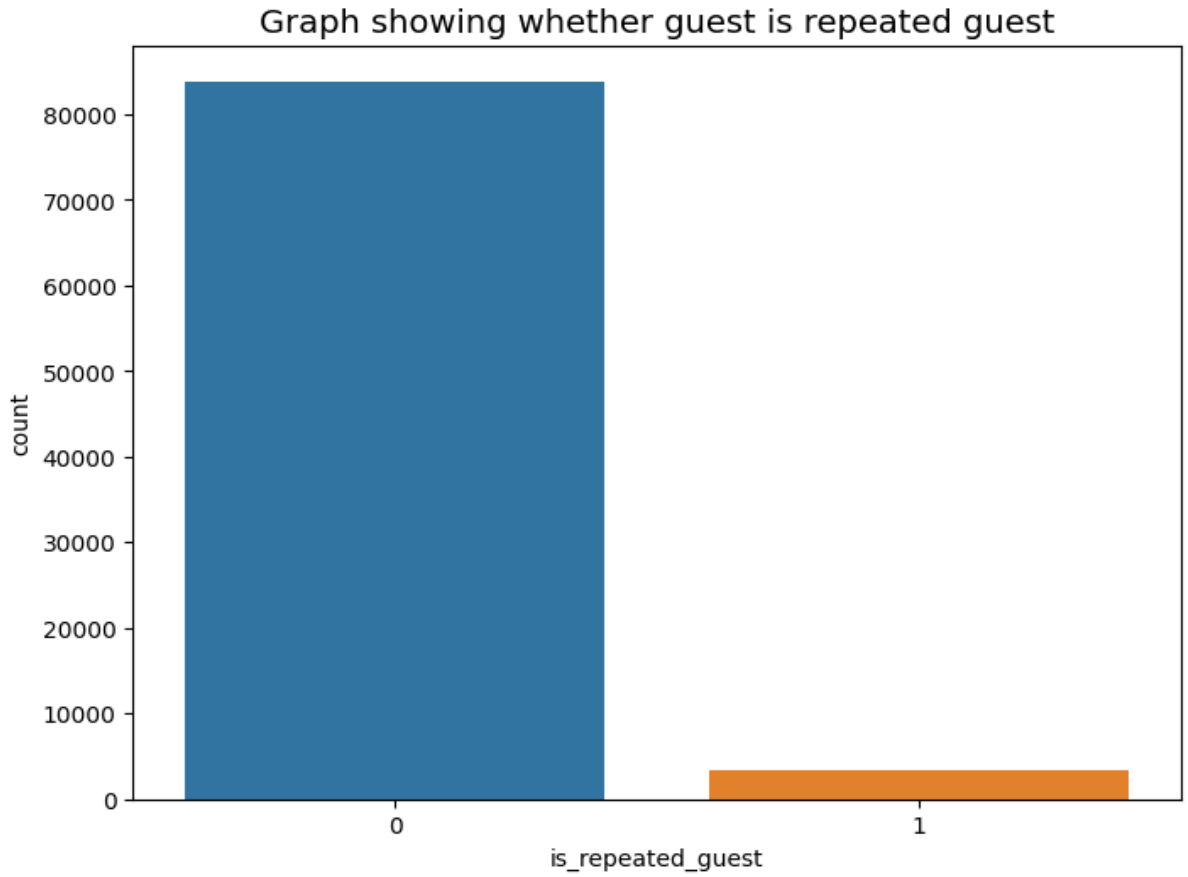
Most common stay length is less than 4 days and generally, people prefer City hotel for short stay, but for long stays, Resort Hotel is preferred.



City hotel has significantly longer waiting time; hence, City Hotel is much busier than Resort Hotel.

Almost 30% of City Hotel bookings and 24% of Resort hotel bookings got cancelled.

Both hotels have very small percentage that customer will repeat, but Resort hotel has slightly higher repeat percentage than City Hotel.



Most customers are not repeating their booking.

Distribution channel wise Analysis

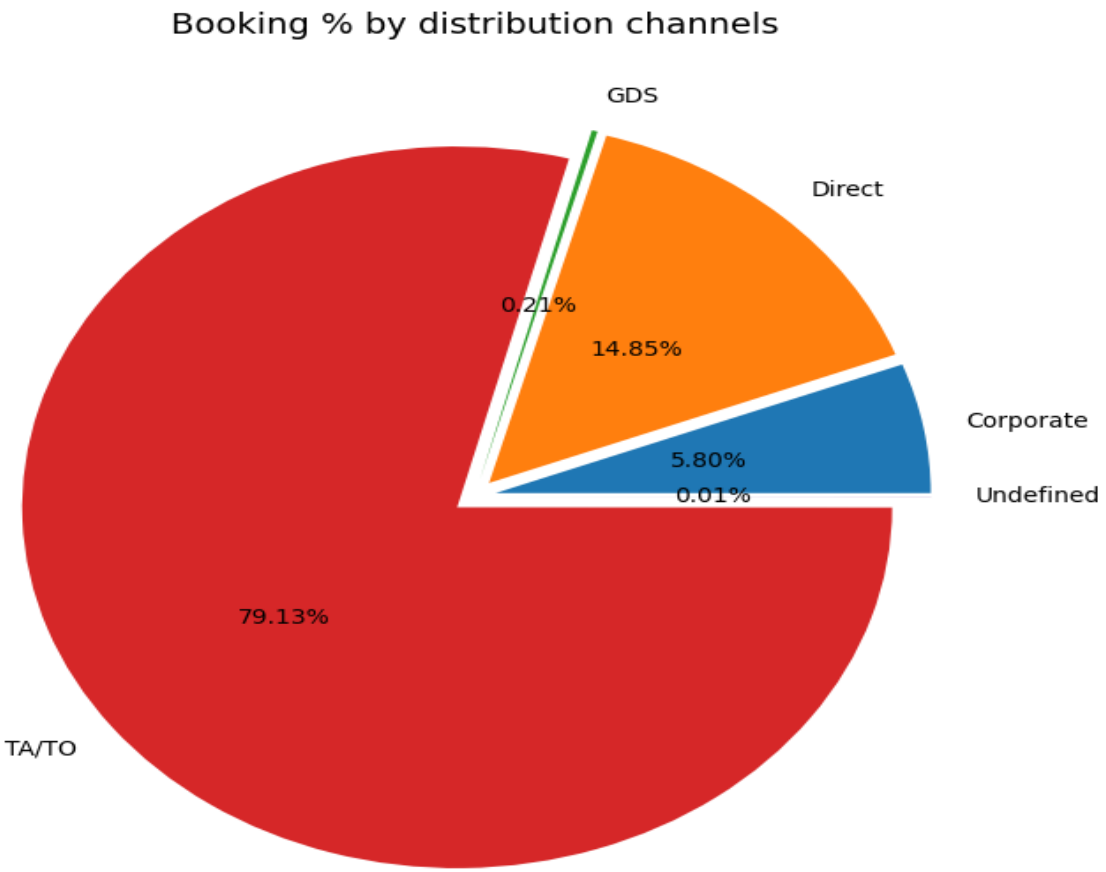
While doing Distribution channel wise analysis of given hotel booking dataset, we answered following questions:

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

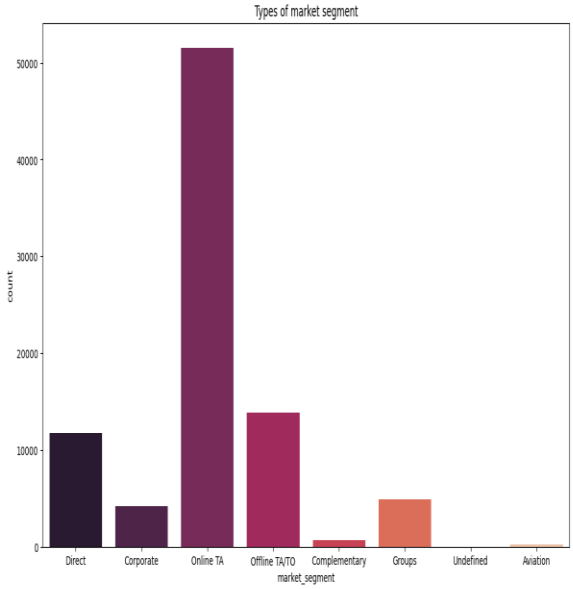
(2) Which channel is mostly used for early booking of hotels?

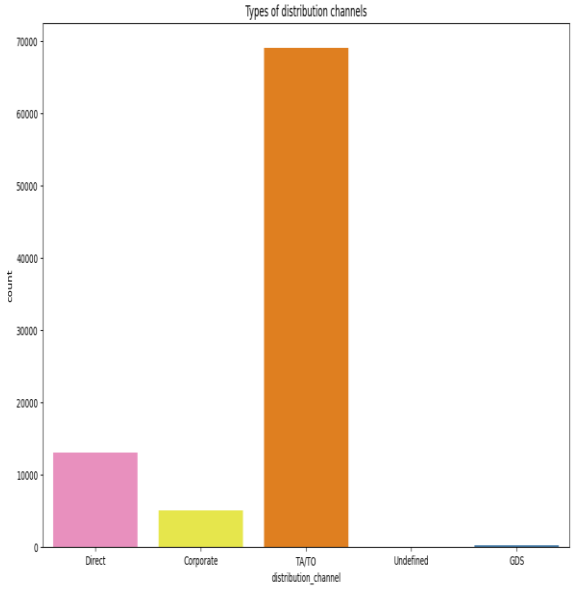
(3) Which channel has longer average waiting time?

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

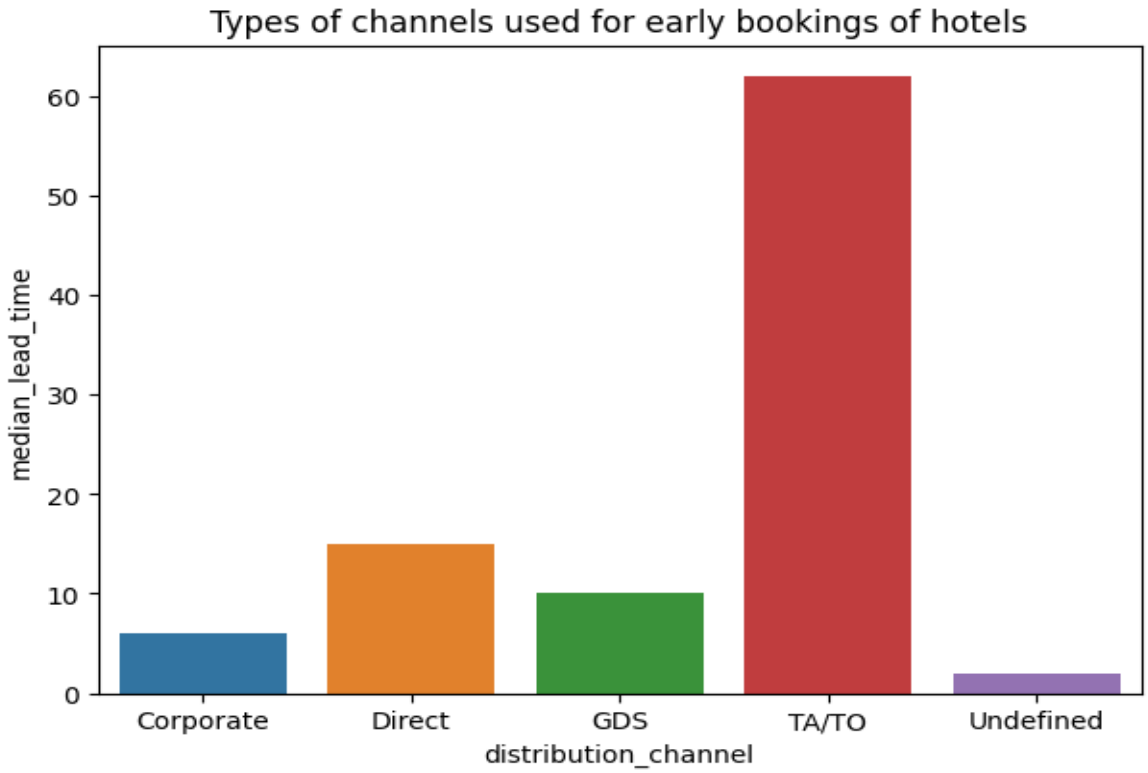


Here we can see that the most of guest are making reservation through TA/TO channels, which is travel agency and tour operator. Then the second most used channel is direct.

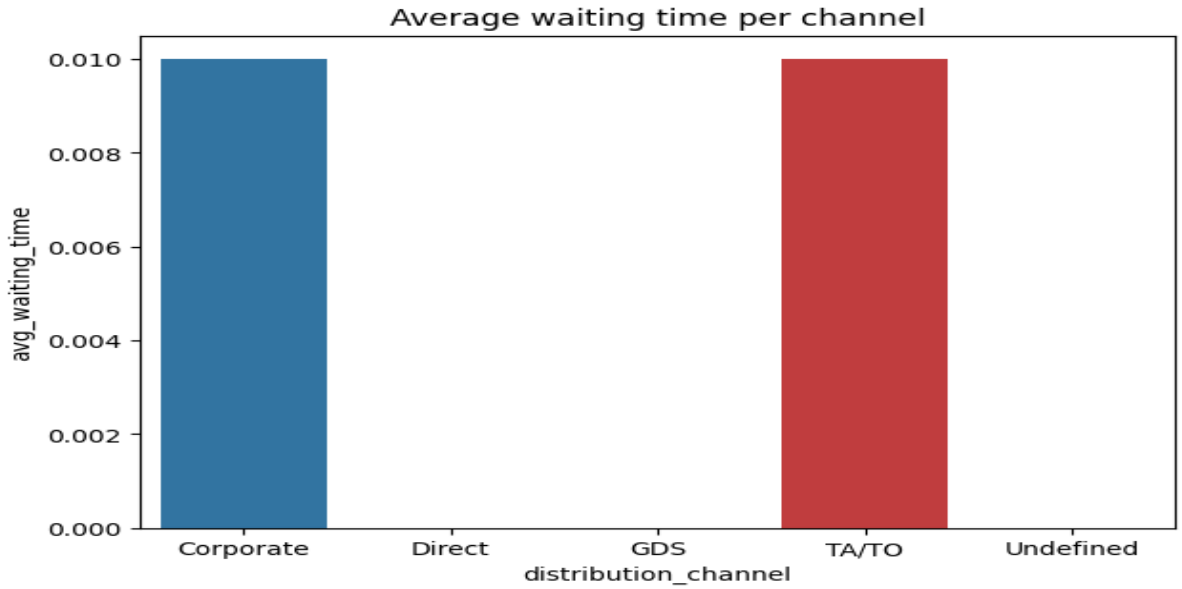




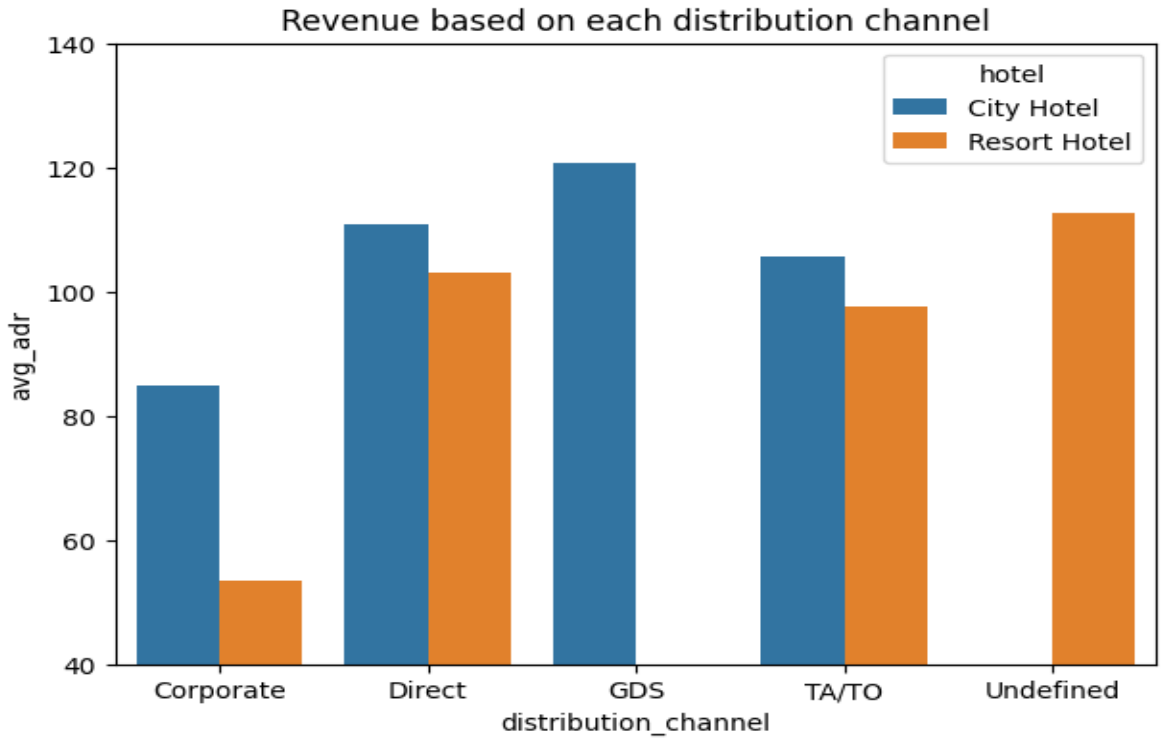
Majority of the bookings were made through online travel agent and the travel agents whether online/offline have the highest distribution rates.



TA/TO is mostly used for planning Hotel visits ahead of time. But for sudden visits other mediums are most preferred. Then the second most used channel is direct.



While booking via TA/TO and Corporate ones may have to wait a little longer to confirm booking of rooms.



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 Undefined channels to get 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:

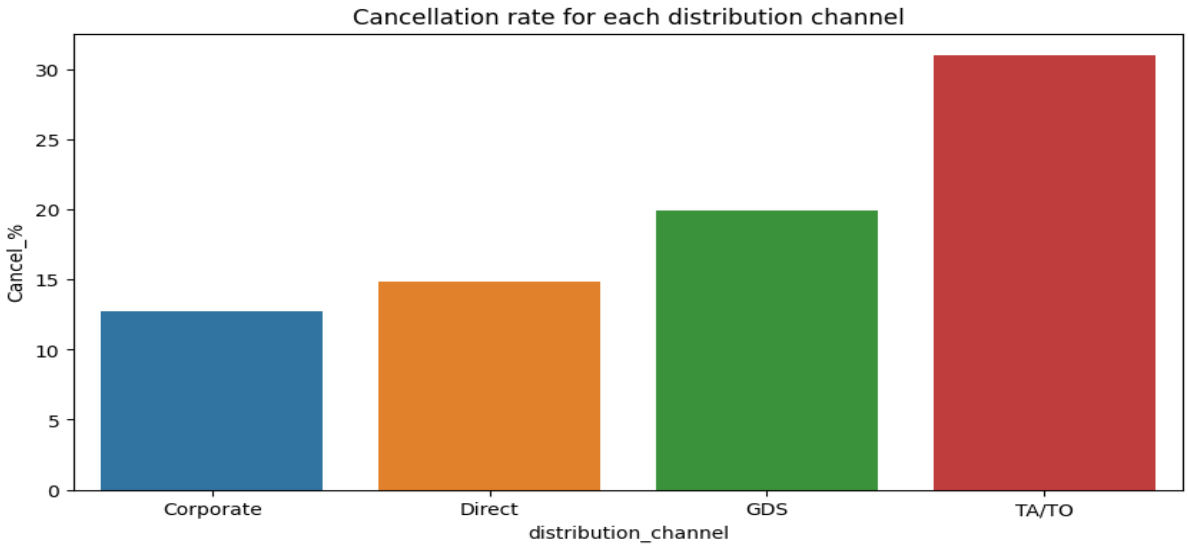
(1) Which significant distribution channel has highest cancellation percentage?

(2) Does a longer waiting period or longer lead-time causes the cancellation of bookings?

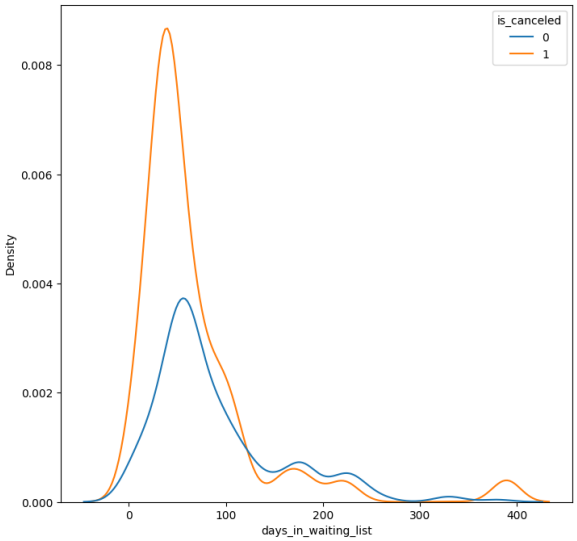
(3) Whether not getting allotted the same room type as demand is the main cause of cancellation for bookings?

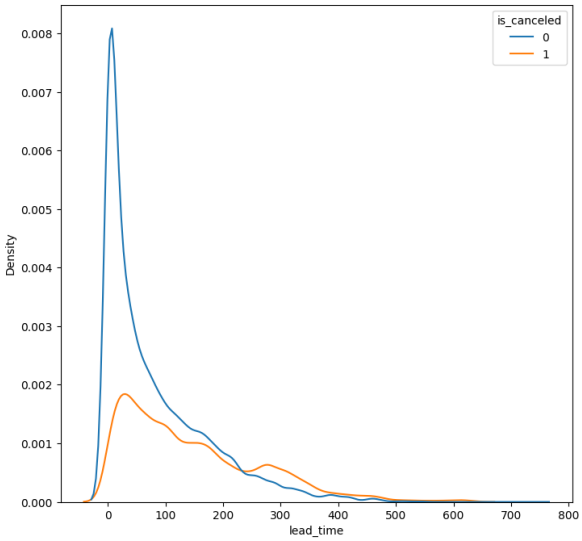
(4) Does not alloting the same room as demanded affect adr?

(5) Which deposit_type have a majotity of bookings?



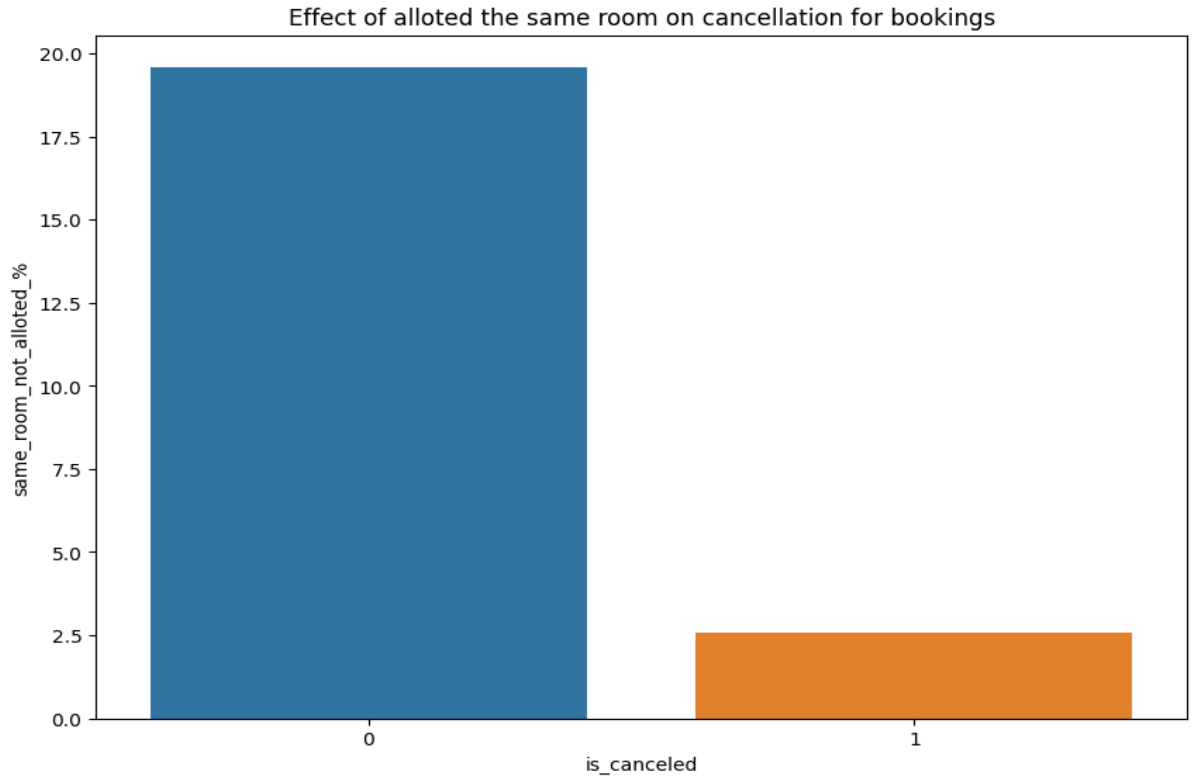
TA/TO has highest booking cancellation percentage. Therefore, a booking via TA/TO is 30% likely to get cancelled.



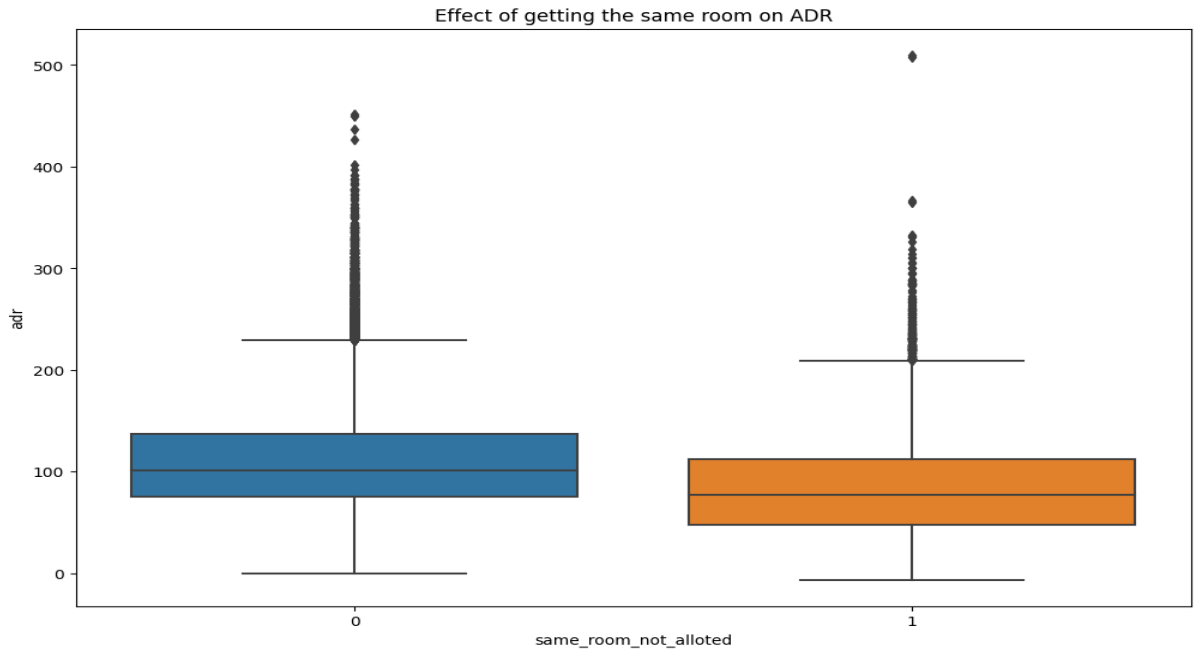


We see that most of the bookings that are cancelled have waiting period of less 50 days but also most of bookings that are not cancelled also have waiting period less than 50 days. Hence this shows that waiting period has no effect on cancellation of bookings.

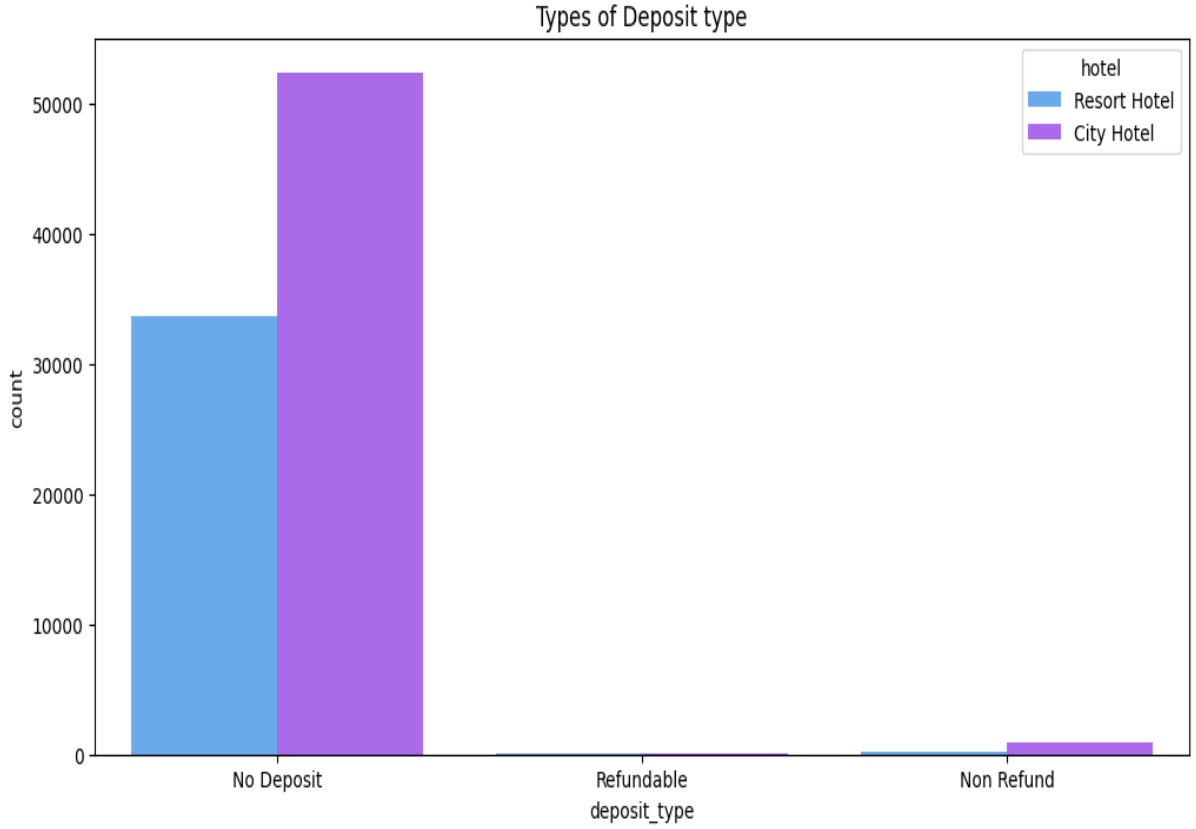
Also, lead time has a smaller affect on cancellation of bookings when lead_time is greater than 50. At 250 the cancellation is higher. On average,both curves of cancelation and not cancelation are similar for lead time too.



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.



Majority of the bookings do not have a deposit towards it.

Time wise Analysis

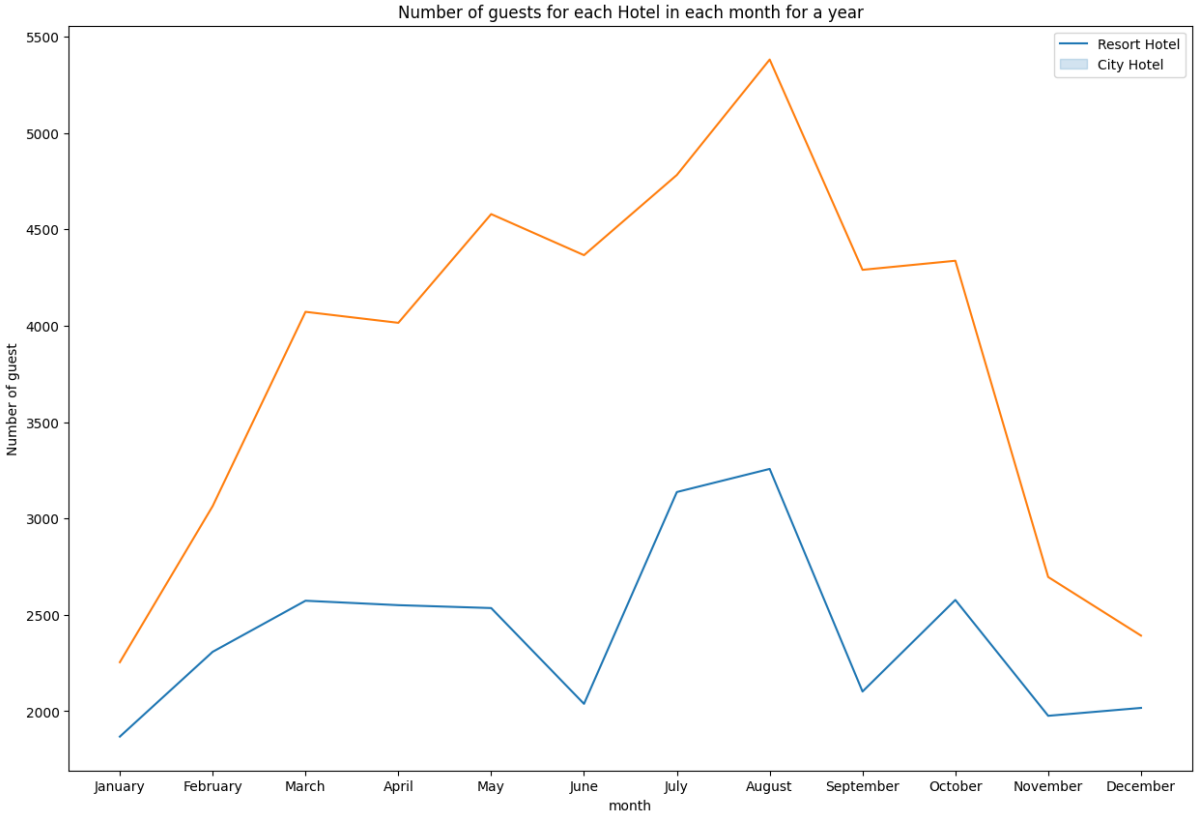
While doing time wise analysis of given hotel booking dataset, we answered following questions:

(1) What are the busiest months for hotels?

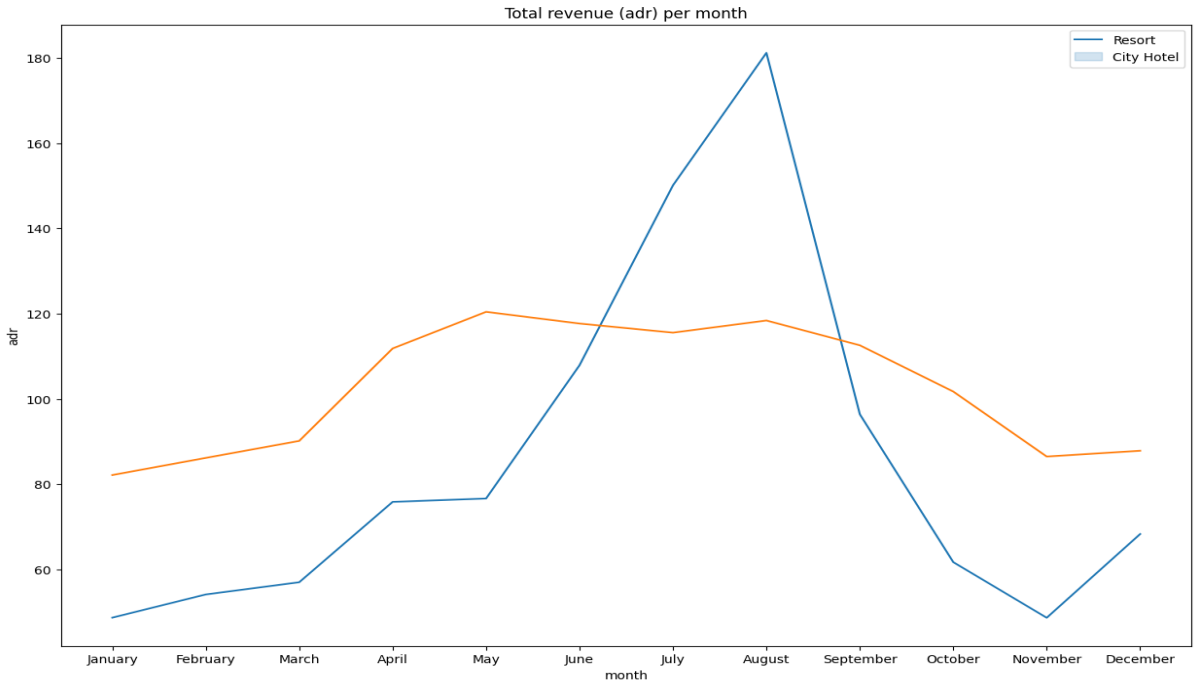
(2) In which months hotels charges higher adr? / Which month results in high revenue?

(3) What is the trend of bookings within a month? / How does booking numbers and adr changes within a month?

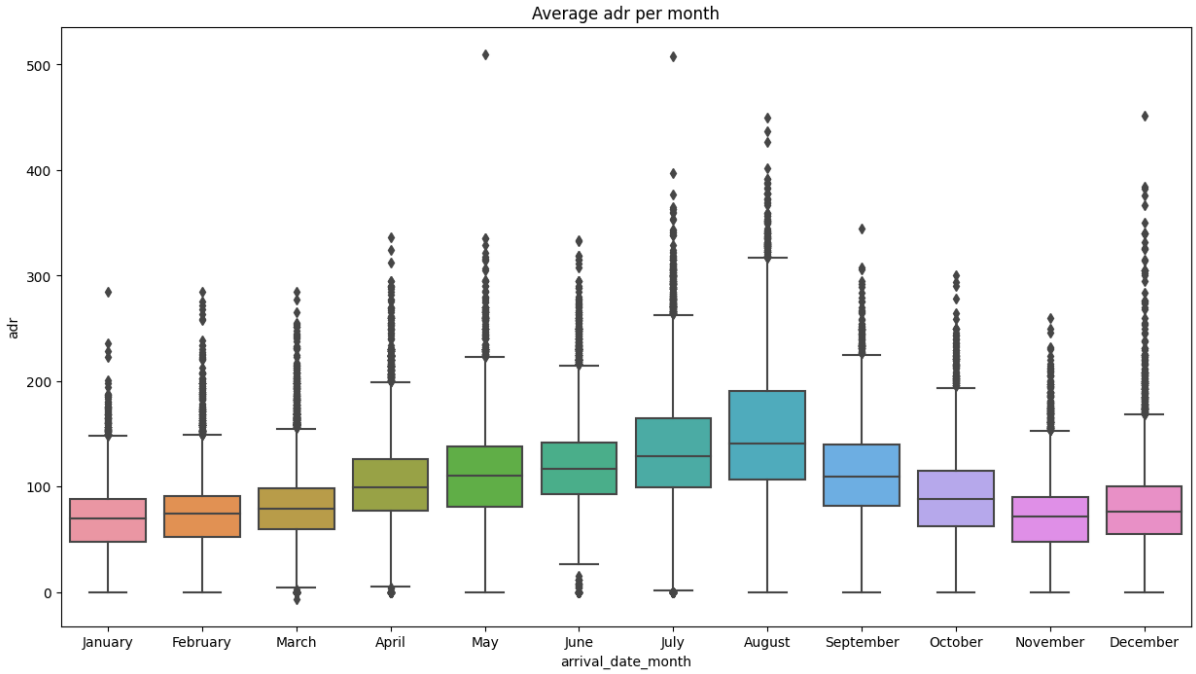
(4) Which types of customers mostly make bookings?



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

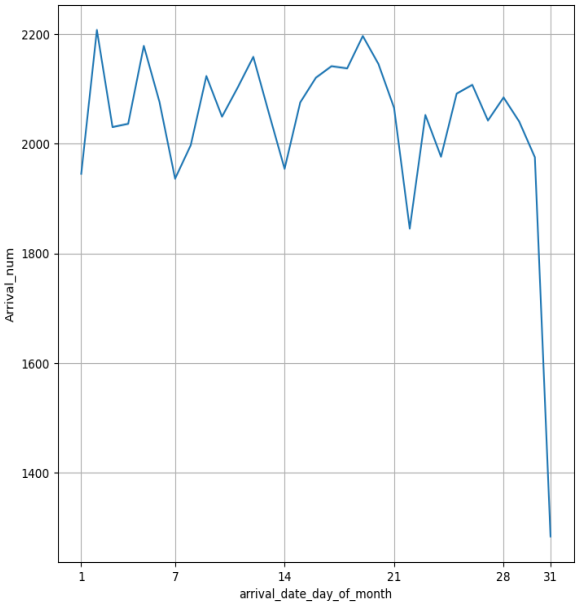


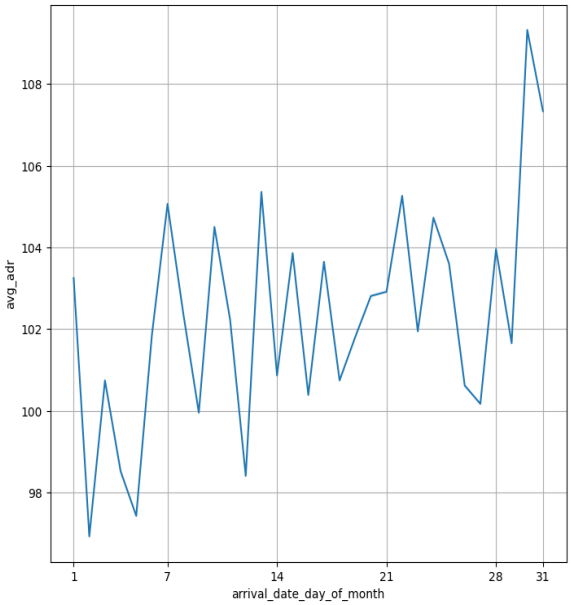
The revenue aspect looks different; the Resort Hotels receives more revenue with respect to City Hotel. From May to August, there was rapid increase in adr. August recorded the highest.



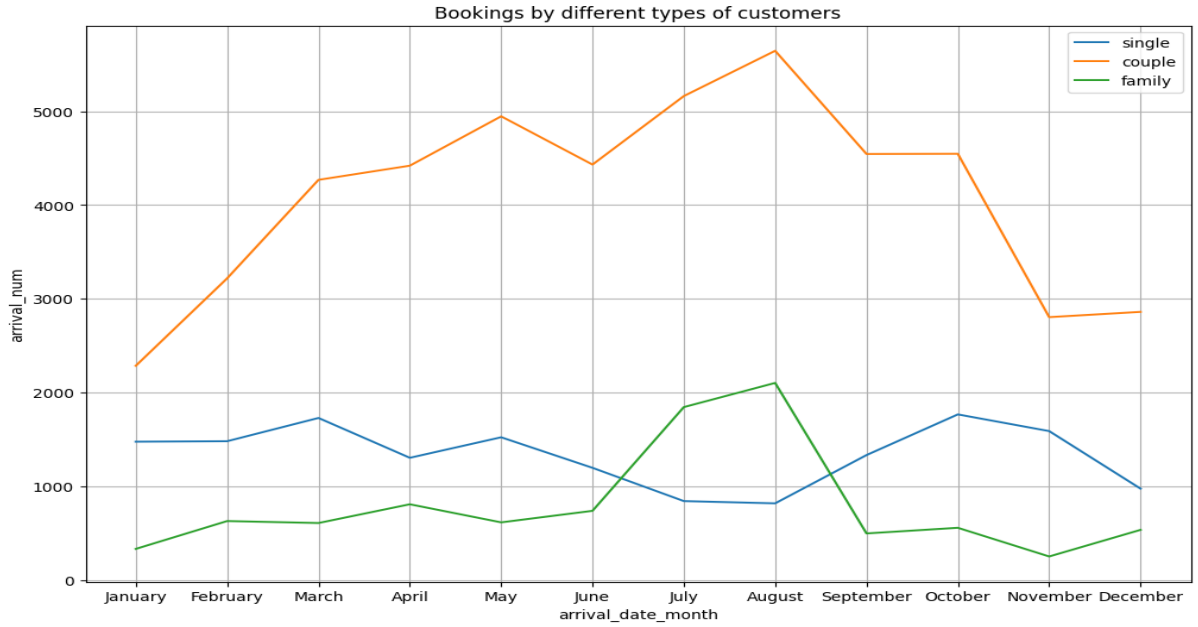
Avg adr rises from beginning of year upto middle of year, reaches peak at August, and then lowers to the end of year. However, hotels do make some good deals with high adr at end of year also.

For both City and Resort hotels, Nov to Jan have cheaper average daily rates.





We can see that graph Arrival_num has small peaks at regular interval of days. This can be due to increase in arrival weekend. In addition, the avg adr tends to go up as month ends. Therefore, charges are more at the end of month.



Couples do most of the bookings (although we are not sure that they are couple, as data does not tell about that).

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.

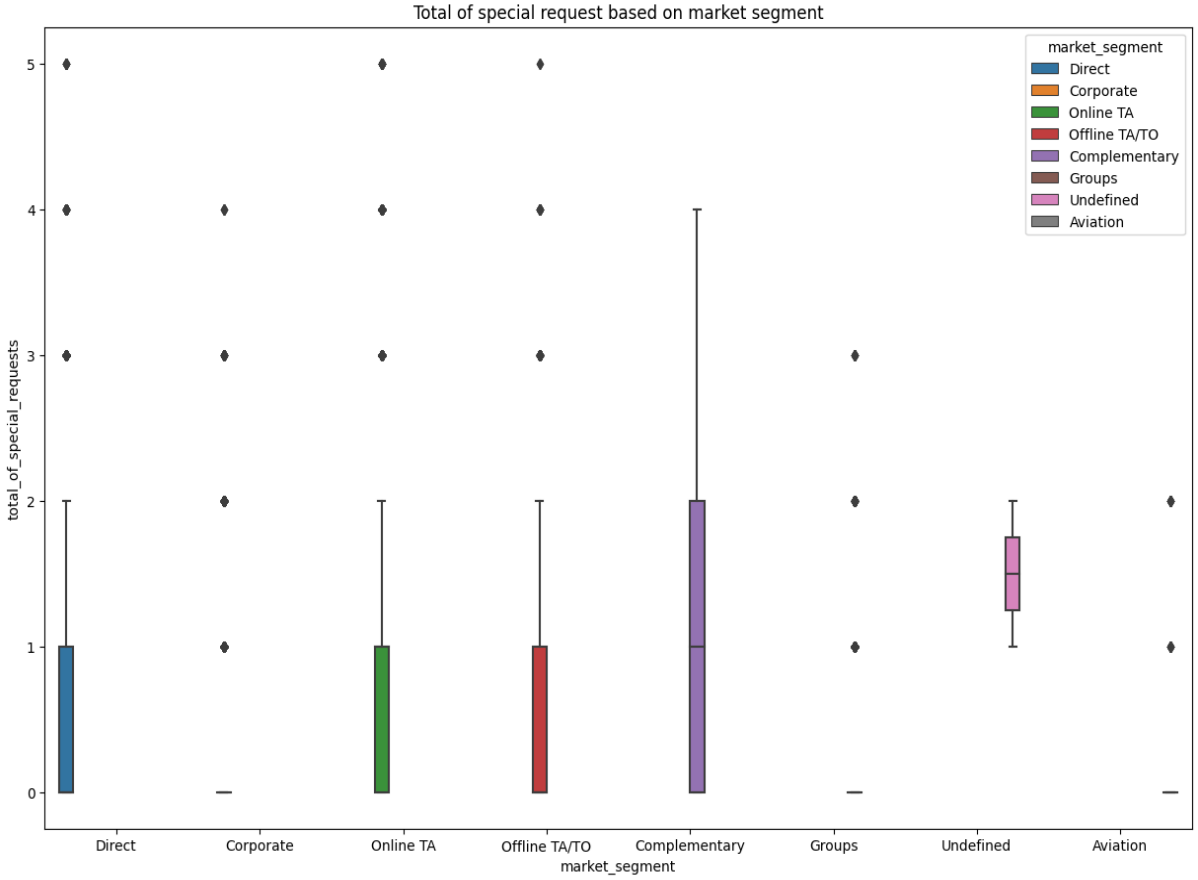
Some important questions

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

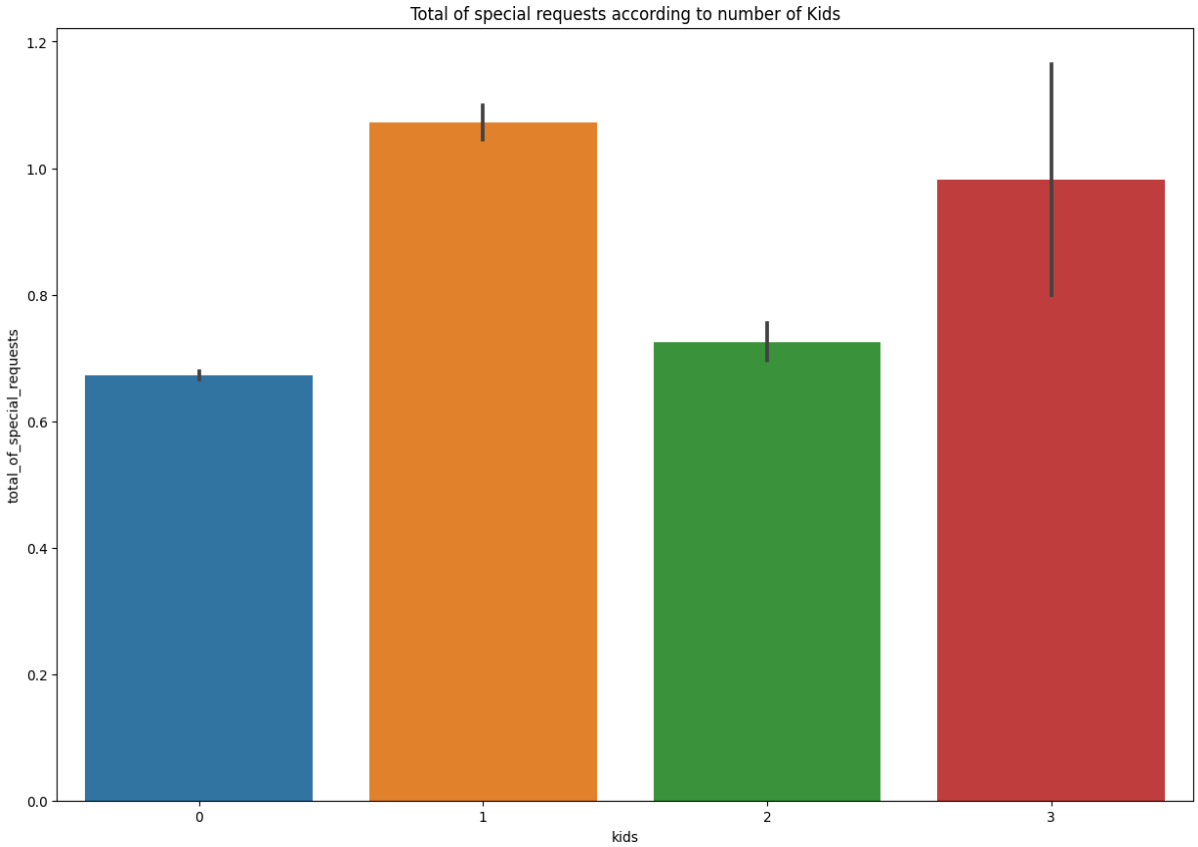
(1) What are the different reason for special requests?

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

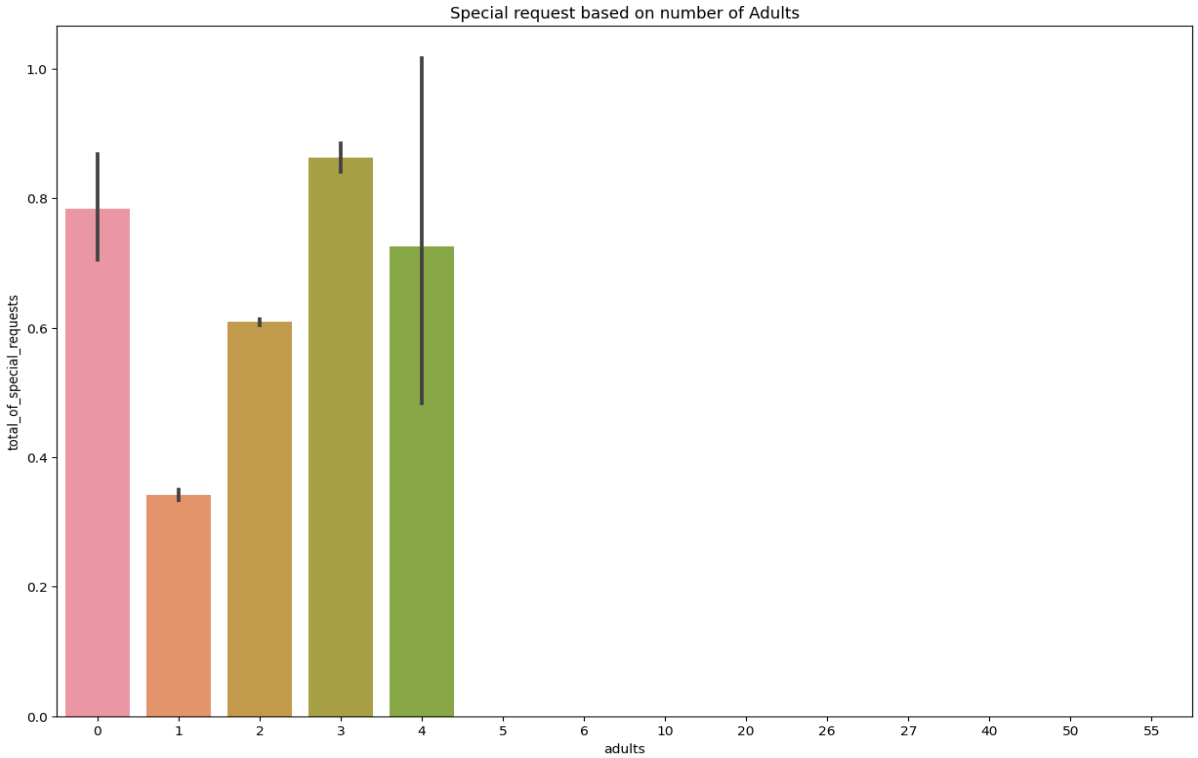
(3) How adr is affected by total staying period in hotels?



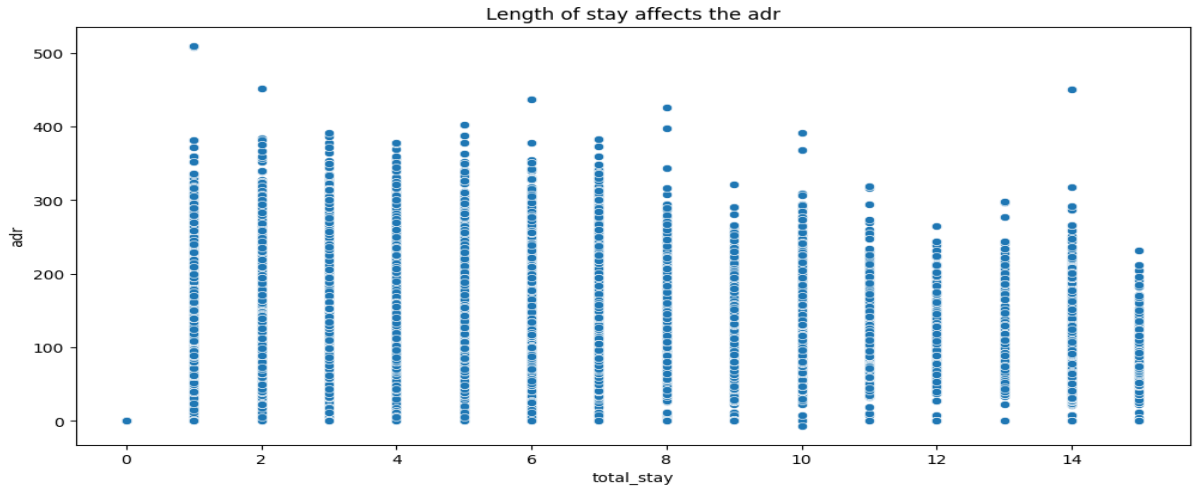
Here we can see that all market segment mostly have special request. There is one segment, which is complementary, having more than average number of special request.



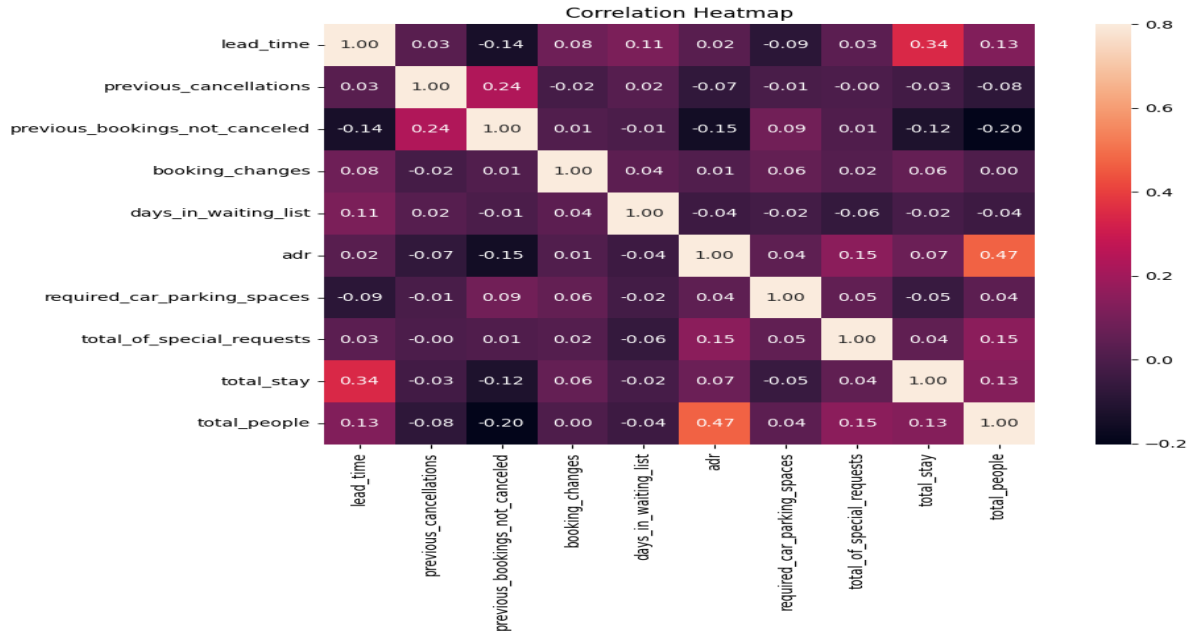
The number of special request are almost the same in the kids section, but more for the first and third kid.



We can see that if the adults are more than two there are more chances that hotels will receive more special requests.



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.



Total stay length and lead time have slight correlation. This may means that for longer hotel stays people generally plan little before the the actual arrival.

adr is slightly correlated with total_people, which makes sense as more no. of people means more revenue, therefore more adr.

CONCLUSION

● Around 62% bookings are for City hotel and 38% bookings are for Resort hotel, therefore City Hotel is busier than Resort hotel. In addition, the overall adr of City hotel is slightly higher than Resort hotel.

● Mostly guests stay for less than 5 days in hotel and for longer stays, Resort hotel is preferred.

● Both hotels have significantly higher booking cancellation rates and very few guests less than 3 % return for another booking in City hotel. 5% guests return for stay in Resort hotel.

● Most of the guests came from European countries, with most number of guests coming from Portugal.

● Guests use different channels for making bookings out of which most preferred way is TA/TO.

● For hotels higher adr deals come via GDS channel, so hotels should increase their popularity on this channel.

● Almost 30% of bookings via TA/TO are cancelled.

● Almost 30% of City Hotel bookings and 24% of Resort hotel bookings got cancelled.

● Not getting same room as reserved, longer lead time and waiting time do not affect cancellation of bookings. Although different room allotment do lowers the adr.

● July-August are the most busiest and profitable months for both of hotels.

● Within a month, adr gradually increases as month ends, with small sudden rise on weekends.

● Couples are the most common guests for hotels; hence, hotels can plan services according to couples needs to increase revenue.

● More number of people in guests results in more number of special requests.

● Bookings made via complementary market segment and adults have on average high no. of special request.

● For customers, generally the longer stays (more than 15 days) can result in better deals in terms of low adr.

Challenges

(1) There was a lot of duplicate data.

(2) Data was present in wrong datatype format.

(3) Choosing appropriate visualization techniques to use was difficult.

(4) A lot of null values were there in the dataset.

THIS REPORT WAS WRITTEN BY: VINCENT MUKOMBA