Hotel Booking Analysis

EDA Capstone Project
AlmaBetter

KAMATAM HARSHITH
LAKSHMI KEERTHANA
TITO VARGHESE
ANMOL RAJ

Points to Discuss

- Agenda
- Data summary
- Data cleaning
- Hotel wise analysis
- Distribution Channel wise analysis
- Cancellation related analysis
- Time and Stay related analysis
- Heat Correlation
- Challenges

Agenda

To extract,observe and analyse the given hotel bookings data set from 2015-2017.

The analysis of given data set in following ways :

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

Data Summary

Given data set has different columns of variables crucial for hotel bookings:

hotel: The category of hotels, which has two values 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]
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.)

Data information

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	hotel	119390 non-null	object
1	is_canceled	119390 non-null	int64
2	lead_time	119390 non-null	int64
3	arrival_date_year	119390 non-null	int64
4	arrival_date_month	119390 non-null	object
5	arrival_date_week_number	119390 non-null	int64
6	arrival_date_day_of_month	119390 non-null	int64
7	stays_in_weekend_nights	119390 non-null	int64
8	stays_in_week_nights	119390 non-null	int64
9	adults	119390 non-null	int64
10	children	119386 non-null	float64
11	babies	119390 non-null	int64
12	meal	119390 non-null	object
13	country	118902 non-null	object
14	market_segment	119390 non-null	object
15	distribution_channel	119390 non-null	object
16	is_repeated_guest	119390 non-null	int64
17	previous_cancellations	119390 non-null	int64
18	previous_bookings_not_canceled	119390 non-null	int64
19	reserved_room_type	119390 non-null	object
20	assigned_room_type	119390 non-null	object
21	booking_changes	119390 non-null	int64
22	deposit_type	119390 non-null	object
23	agent	103050 non-null	float64
24	company	6797 non-null	float64
25	days_in_waiting_list	119390 non-null	int64
26	customer_type	119390 non-null	object
27	adr	119390 non-null	float64
28	required_car_parking_spaces	119390 non-null	int64
29	total_of_special_requests	119390 non-null	int64
30	reservation_status	119390 non-null	object
31	reservation_status_date	119390 non-null	object

dtypes: float64(4), int64(16), object(12)
No. of Rows : 119390 entries, 0 to 119389
No. of Data columns : 32 columns

Data Cleaning

Data Cleaning is a crucial step before EDA as it will remove the ambiguous data that can affect the outcome of EDA.

While cleaning data we will perform the following steps:

1) Remove duplicate rows *(df1[df1.duplicated()].shape)+df1.drop_duplicates(inplace = True)*

No. of duplicate rows : 31980

2) Handling missing values. (`hotelbookings.isnull().sum().sort_values(ascending=False)`

```
hotelbookings[['company','agent']] = hotelbookings[['company','agent']].fillna(0)
hotelbookings['children'].fillna(hotelbookings['children'].mean(), inplace = True)
hotelbookings['country'].fillna('others', inplace = True)
```

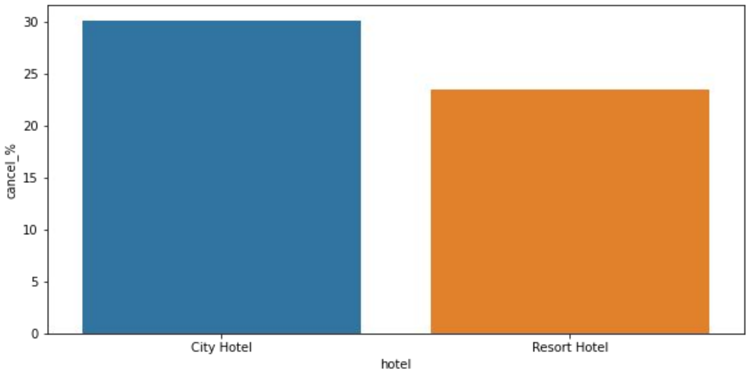
3) Convert columns to appropriate data types. (df1[['children', 'company', 'agent']] = df1[['children', 'company', 'agent']].astype('int64'))

4) Removing the Outliers (adr,lead_time,days_in_waiting_list,required_car_parking_space)

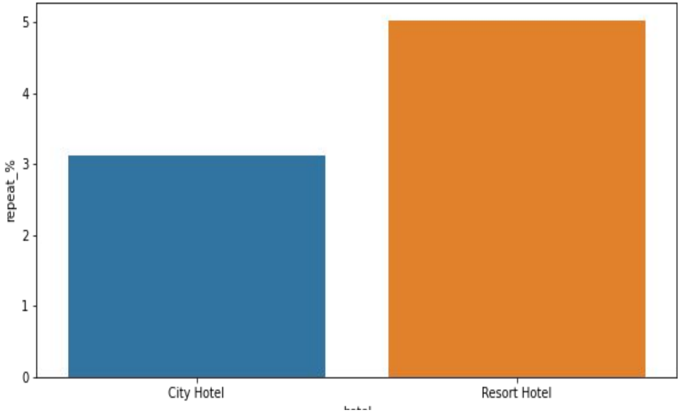
Hotel wise analysis

- Hotel with higher bookings cancellation rate.
- Hotel with longest waiting time
- Hotel with most revenue.
- Chances of customer returning to hotel for another stay
- Factors Governing Booking
- Special requests by the guests

Hotel with higher bookings cancellation rate



Hotel with most revenue

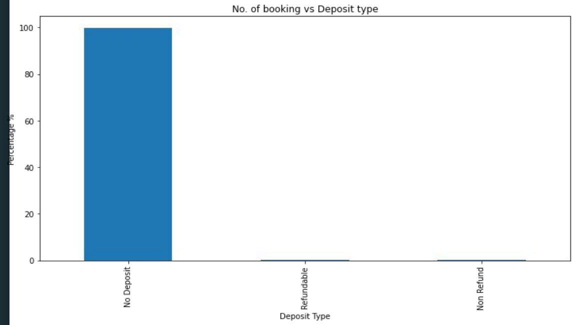


Chances of customer returning to hotel for another stay

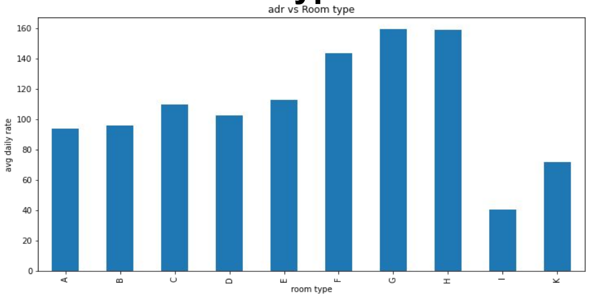
- 30% of customers of City Hotel have cancelled their booking.Whereas 20-25% of customers have cancelled their booking in Resort Hotel.
- City hotel generates most revenue.
- Both the customers have less chances of its customer returning for the stay.

Factors governing booking

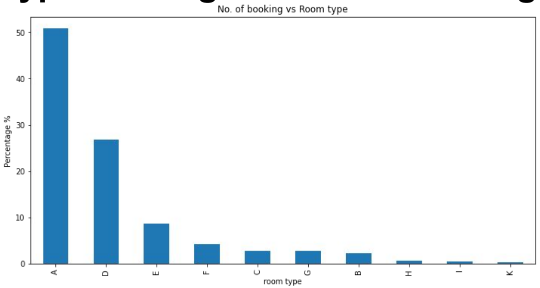
Deposit type

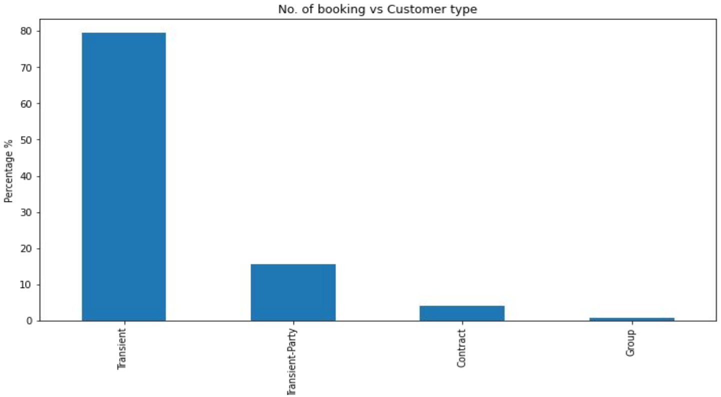


Room type vs adr



Room type with highest no. of bookings

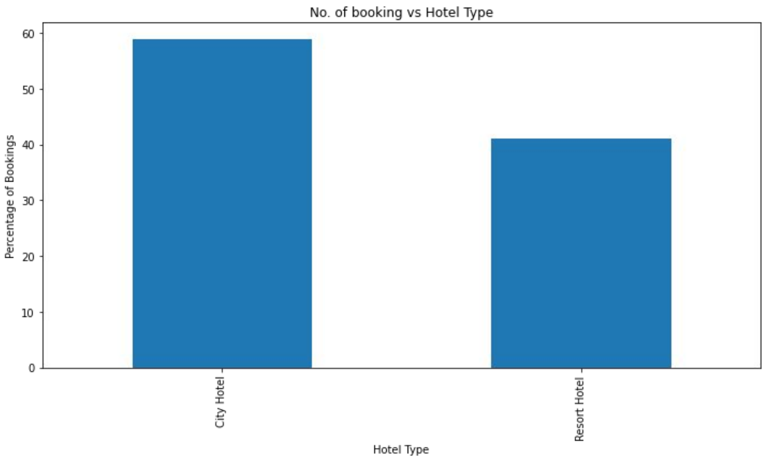




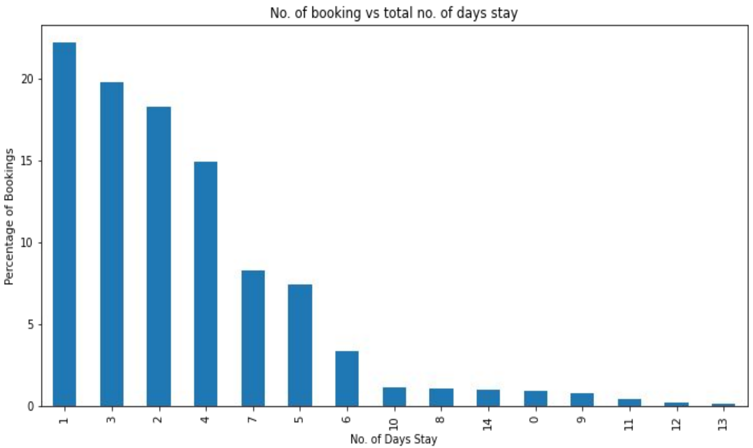
Customer Type

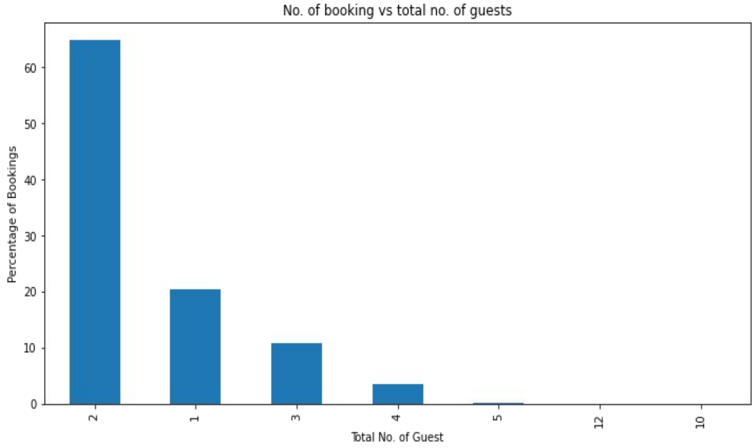
- Most number of customers used No Deposit option
- Room type A has the highest number of bookings compared to the other room types.
- The most number of bookings was made by Transient Customer Type and the least was by Group customer type..

Hotel Type



Total number of days stays



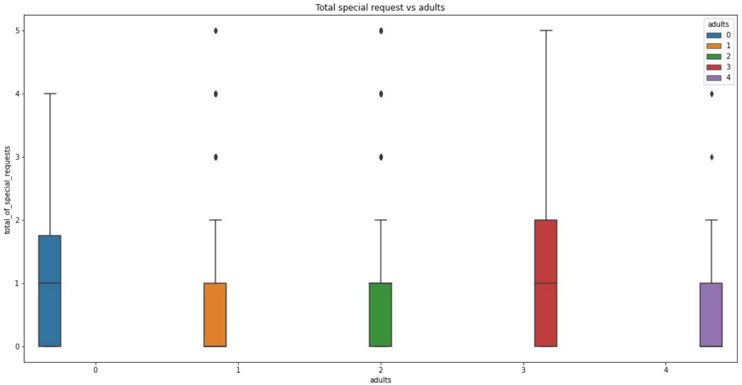


Total Number of Guest

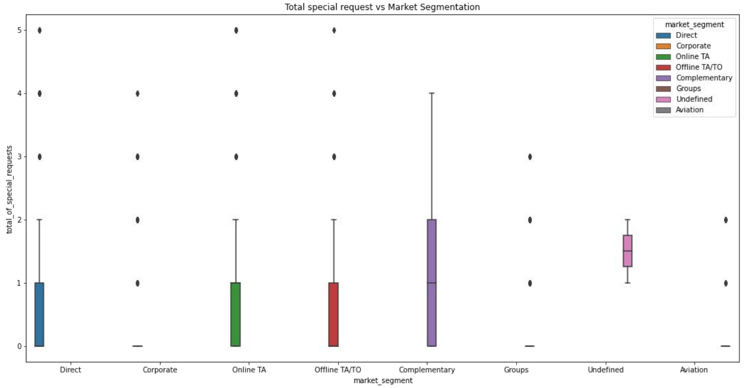
- Most preferred hotel was City Hotel.
- The number of days stay was mostly 1
- Most number of bookings was done by couples.

Special requests by the guests

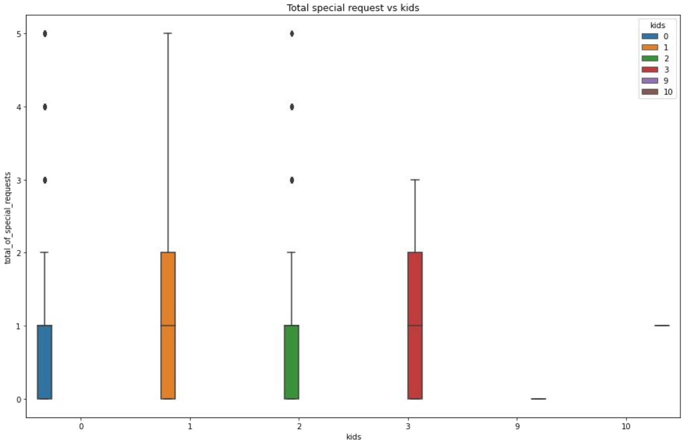
Special Requests According to Adults



Special Requests According to Market Segmentation



Special Requests According to Kids

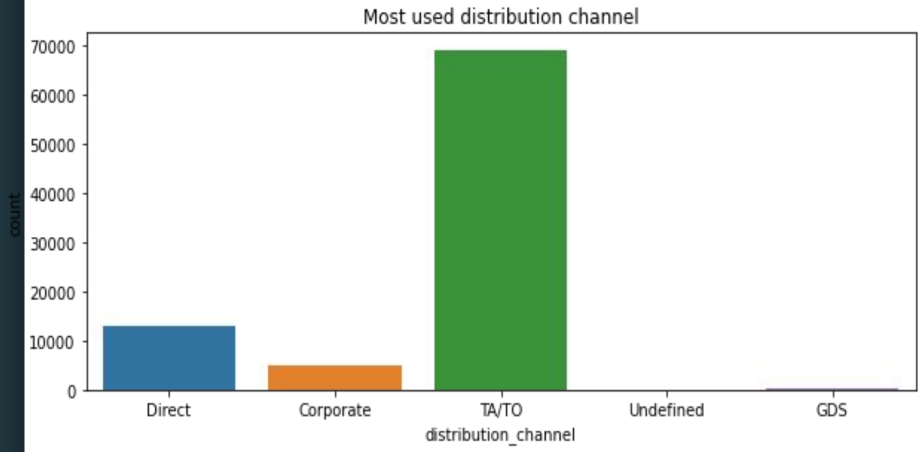


- The most number of special request demand was from Complementary market segment.
- The cases where the number of adults is more than 3 ,there was a high demand of special requests.
- When the no. of kids were 1 and 3, we can expect more special requests.

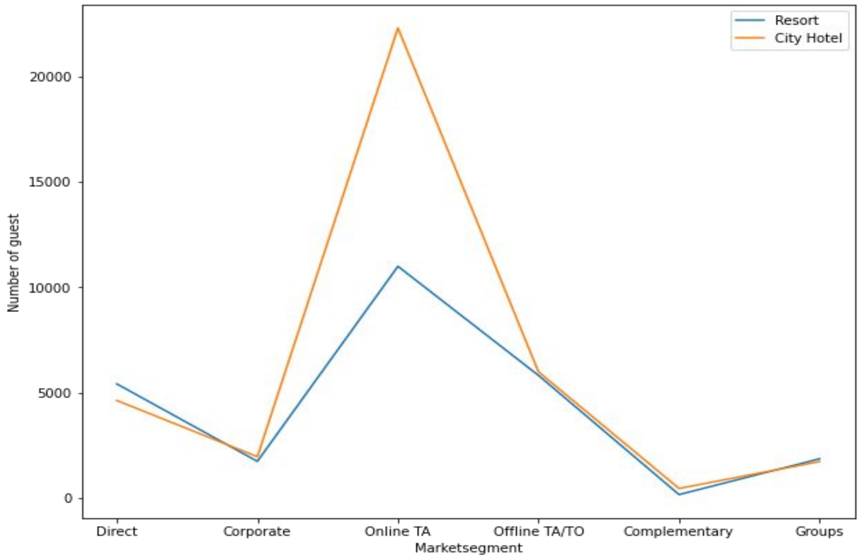
Distribution Channel wise analysis

- Most used Distribution Channel
- Which distribution channel brings better revenue generating deals for hotels?
- Market segments used by the guests
- Distribution Channel with highest cancellation

Most used distribution channel



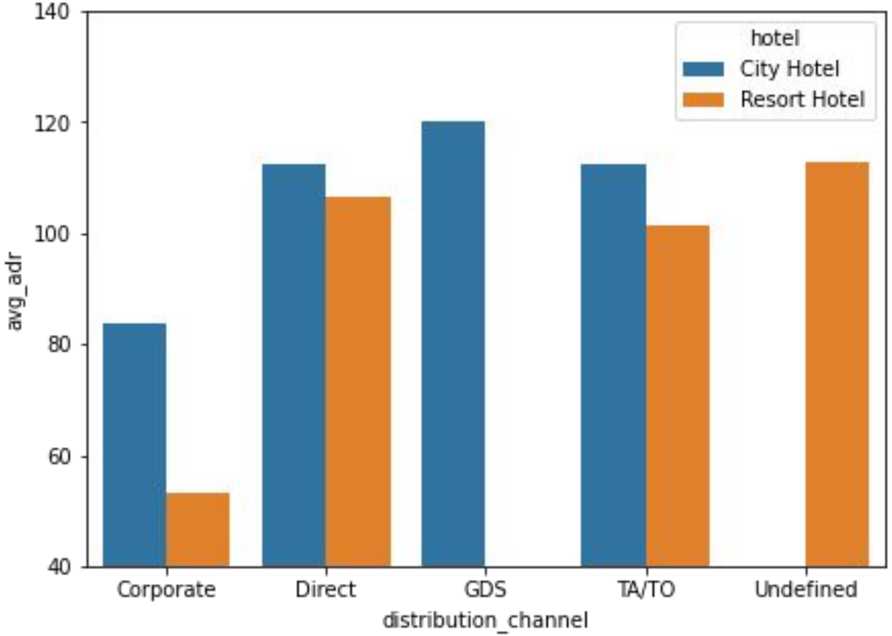
Most used market segment



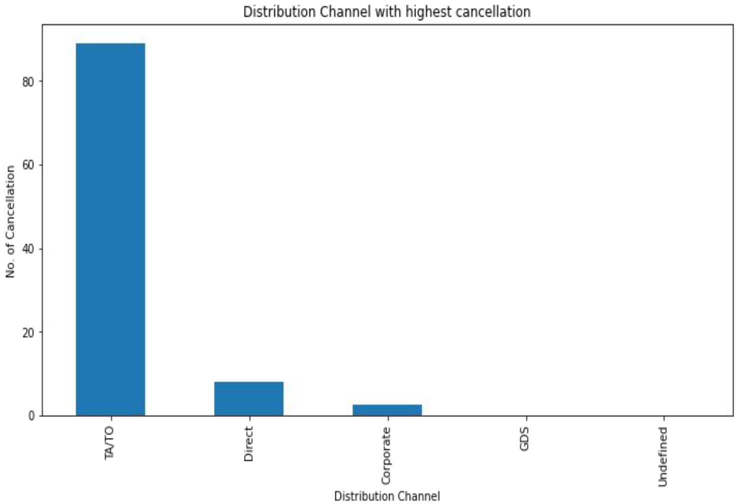
Most number of customers have used TA/TO(Travel Agency/Travel Operator) distribution channel for hotel bookings.

Mostly used market segment by the guests was Online TA to book City hotel and Resort hotel.

Distribution channel bringing highest revenue generating deals



Distribution Channel with highest cancellation



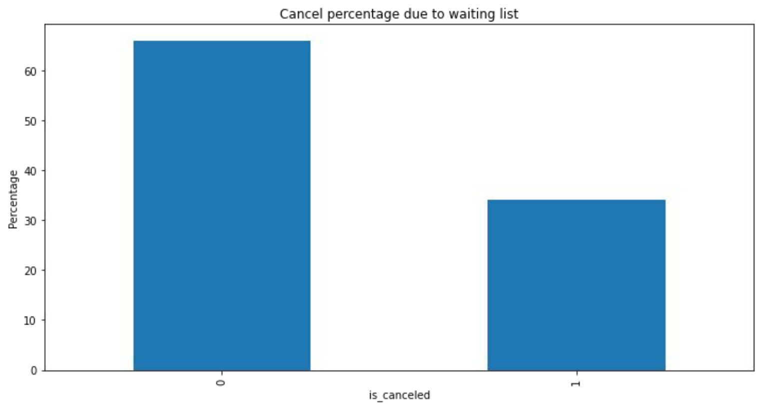
- GDS channel brings higher revenue for City hotel. Whereas for Resort hotel gets more revenue by direct and TA/TO channel.

Cancellation related Analysis

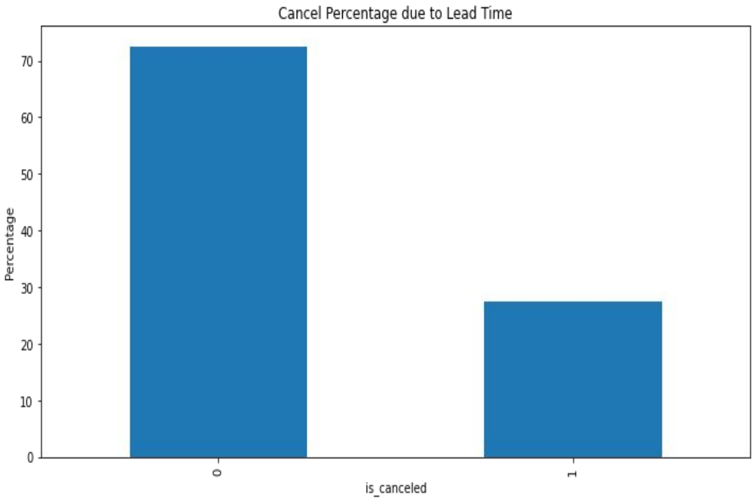
- Waiting time(days)
- Lead Time
- Cancellation for not assigning same room
- Car parking space

Cancellation related analysis

Waiting time(days)

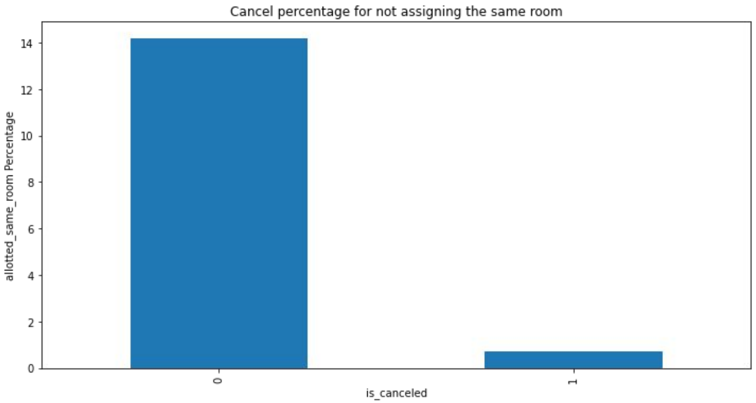


Lead Time



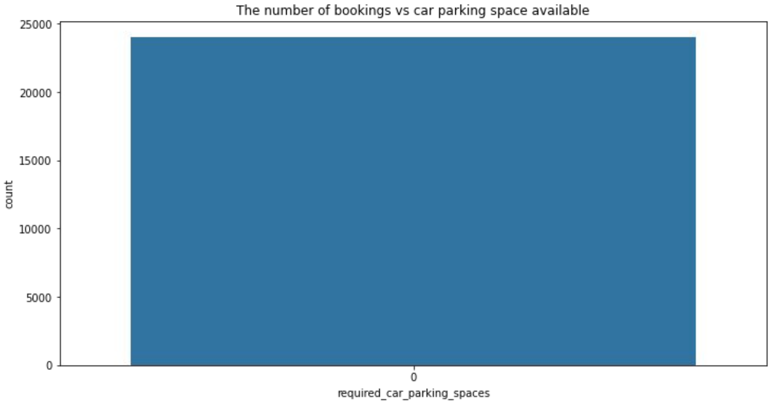
- The parameters like lead time and days in waiting list have no significant impact on the cancellation rate.

Cancellation for not assigning same room



- The booking change from assigned room to reserved room parameter not had any influence on cancellation of bookings.

Car parking space

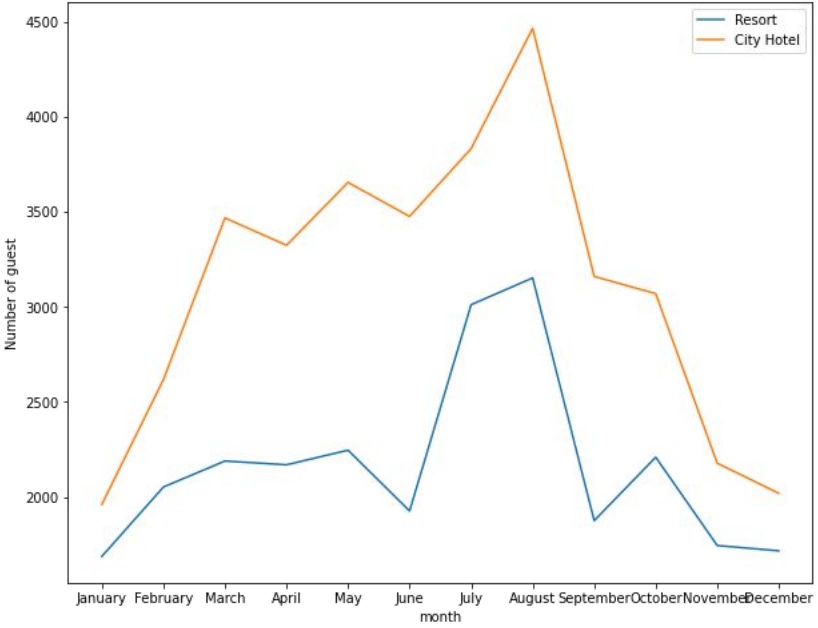


- Main reason for cancellation has been because of no car parking space.

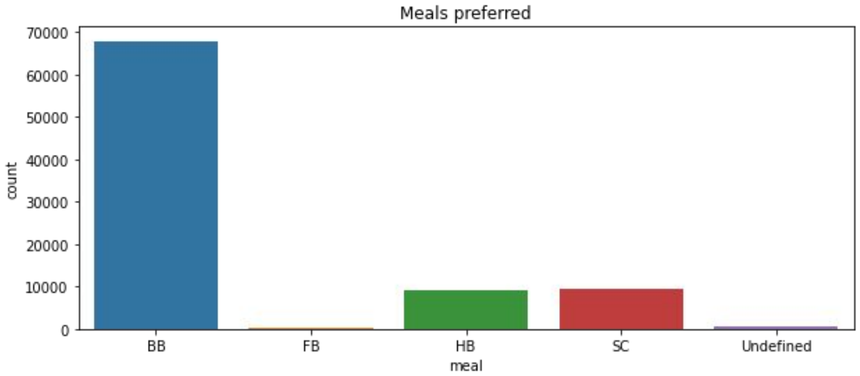
Time and Stay related Analysis

- Customer type with maximum Average Daily Rate
- Type of customers booking the most
- Best time to book a hotel room
- Countries from which most customers are coming

Best time to book a hotel room

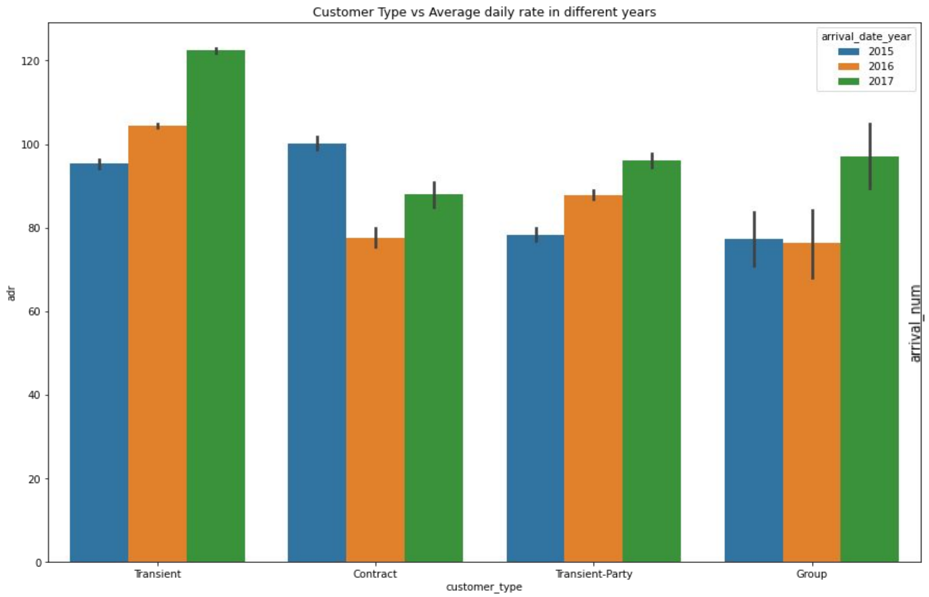


Meal preferred



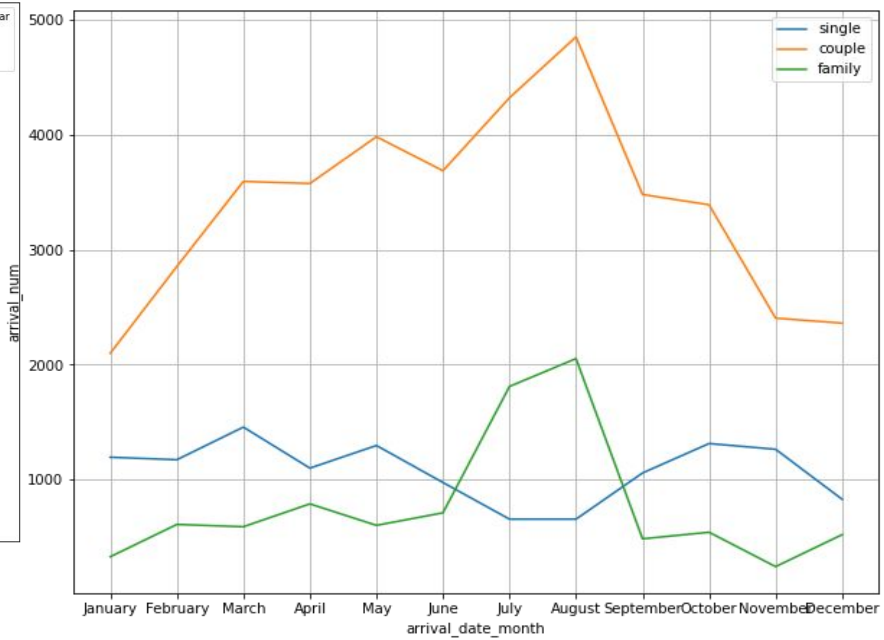
- As per the plot graph, we can conclude that most of the customers visit in the month of August.
- In this analysis, we have concluded that the most preferred meal type by the customer is Bed and Breakfast(BB).

Customer type with maximum Average Daily Rate



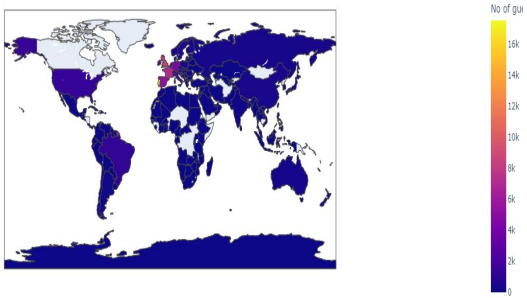
We can conclude that, transient customer type generates maximum adr. The adr of transient-party has been increasing with the year. The group customer type does not show much of a progression.

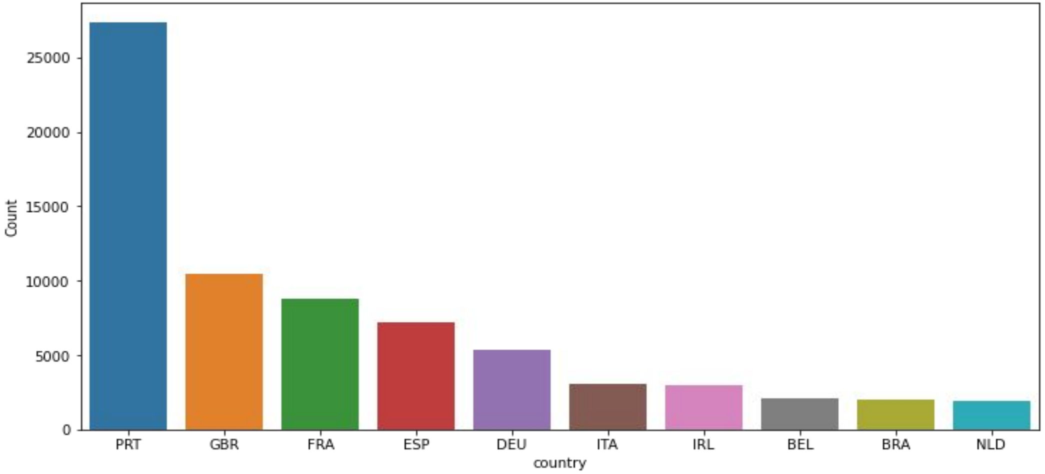
Type of customers booking the most



According to the above graph, it shows that mostly couples or families that has been visiting during the month of July and August.

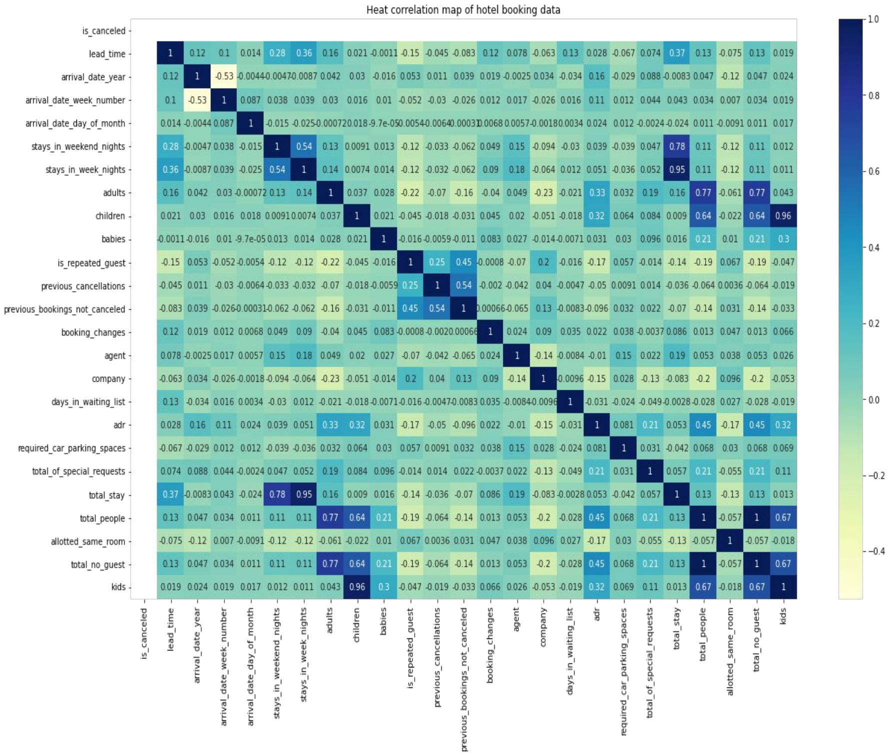
Countries from which most customers are coming





- Most of the guests were from Portugal with 25k customers. Second by Great Britain with 10k customers. Followed by France, Spain and Germany respectively.

Heat Correlation Map



The average daily rate is positively correlated with total number of guest.Hence we can understand that when number of guest increases ,the average daily rate of hotel also increases with it.

The average daily rate is positively correlated with the number of special request.Hence we can understand that when the number of special request increases the revenue of the hotel increases.

The total number of days stay and lead time have slight positive correlation to each other. Thus we can say that higher number of days stay result in higher lead time.

Challenges

A lot of null values were present in the dataset.

There were a lot of duplicate data.

Removing the outliers from the given dataset.

Selecting appropriate visualization techniques was a tedious job.

Thank You !!!