Capstone Project

HOTEL BOOKING ANALYSIS (EDA)

INTRODUCTION

We are here to explore a hotel booking dataset to discover important factors that govern the bookings, which contain booking information for a city hotel and a resort hotel.

We will analyse some important aspects of hotel bookings which will helps us identify major loopholes and give us insights which will be helpful to run profitable hotel business.

Data Summary: The dataset contains different columns or variables. Lets have a look at them.

CATEGORICAL COLUMNS

- **Hotel** : There are two types of hotels city hotel and a resort hotel .
- **Arrival_date_month** : month when guests arrived .
- **Reserved_room_type** : room booked .
- **Assigned_room_type** : room allotted.
- **Lead_time** : The time between reservation and actual arrival.
- **Customer_type**: Type of customers(Transient, group, etc.)
- **Meal** : Meal preferences per reservation.(BB,FB,HB,SC,Undefined].
- **Country**: The origin of guests arrival.
- **Distribution_channel**: The medium through booking was made.
- **Market_segment** : This column show how reservation was made and what is the purpose of reservation.

➢ CATEGORICAL COLUMNS

- **Reservation_status** :status of booking.
- **Is_cancelled** :Value indicating if the booking was canceled (1) or not (0).
- **Repeated_guests** : if the booking name was from a repeated guest (1) or not (0).

➢ NUMERIC COLUMNS

- **Arrival_date_year :** year of arrival date.
- **Adults ,Children and babies :** number of children, number of babies and number of adults.
- **Stay_in_week_nights , stays_in_weekend_nights :** No. of week nights(**Monday-Friday**) or no. of weekend nights (**Saturday & Sunday**) the guest stayed or booked to stay at the hotel.
- **Arrival_date_day_of_month, arrival_ date_week_number :** Month of arrival date and Week number of year for arrival date.
- **Booking_changes :** Number of changes/amendments made to the booking.
- **Total_number_of_special_request :** Number of special requests made by the customer**.**
- **Required_car_parking_spaces :** Number of car parking spaces required by the customer.

NUMERIC COLUMNS

- **Adr : Average Daily Rate** as defined by dividing the sum of all lodging transactions by the total number of staying nights.
- **Company :** That made the booking or responsible for paying the booking.
- **Agent :** ID of the travel agency that made the booking.
- **Days_in_waitings_list** : Number of days the booking was in the waiting list before it was confirmed to the customer.

DATA INFORMATION

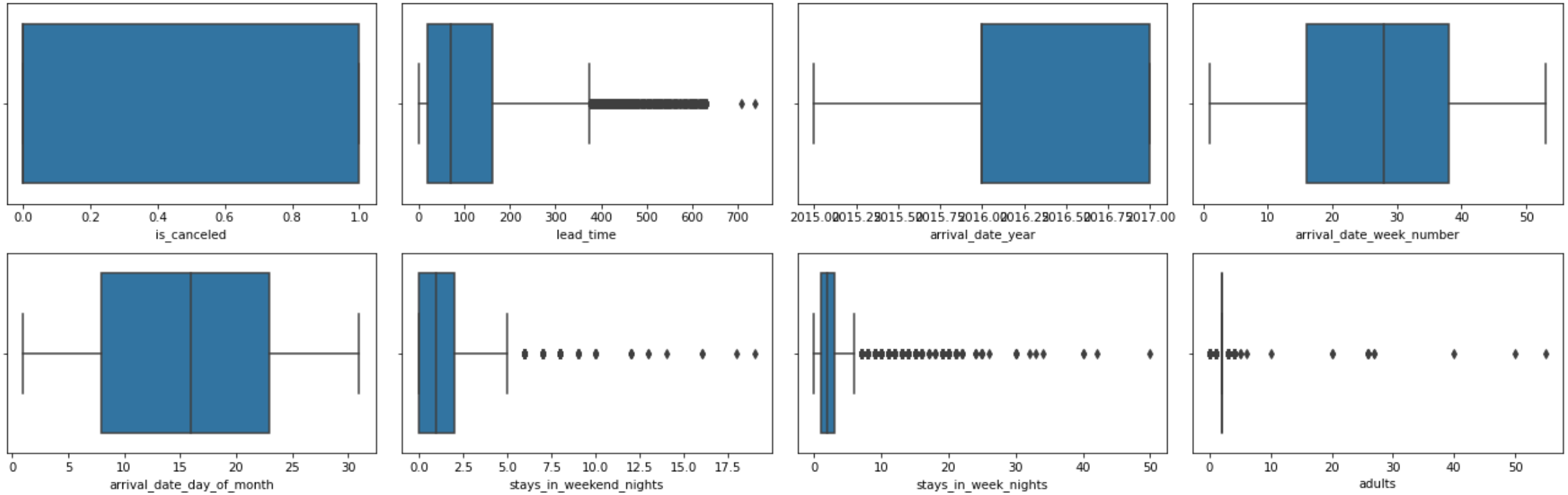
After applying some basic methods , we found out that the data-set is sufficiently large. And here are some of the findings.

- The Data-set contains 119390 rows and 32 features.
- There are 4 columns of float64 data types, 12 columns with object data types, and 16 columns with int64 data types.
- We got many null values in the data-set.

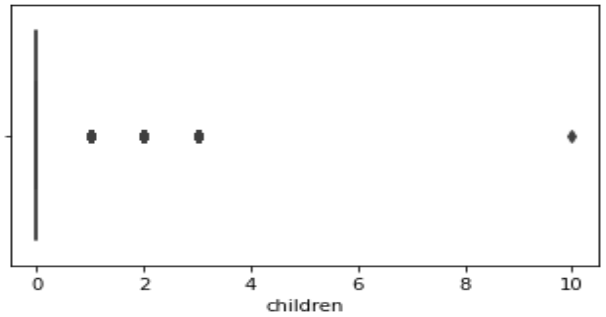
  - ❑ 112593 null values in company column.
  - ❑ 16340 null values in agent column.
  - ❑ 488 null values in country column.
  - ❑ 4 null values in children column.

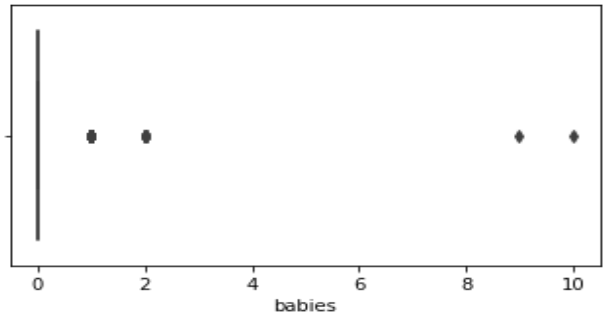
HANDLING OUTLIERS

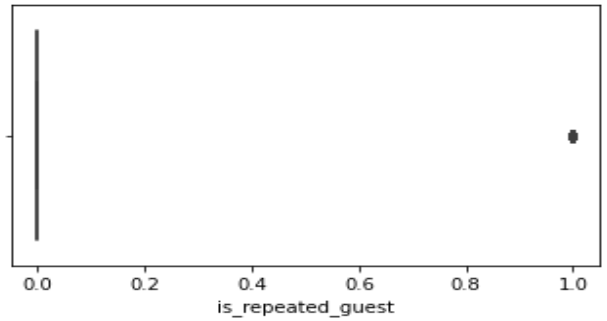
Some columns with outliers

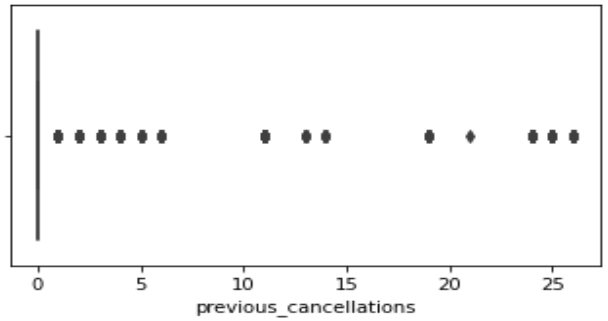


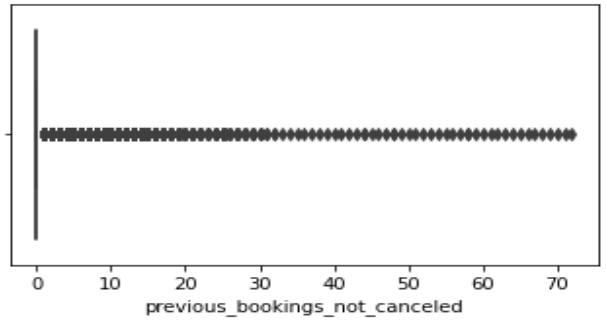
- **Some columns with outliers**

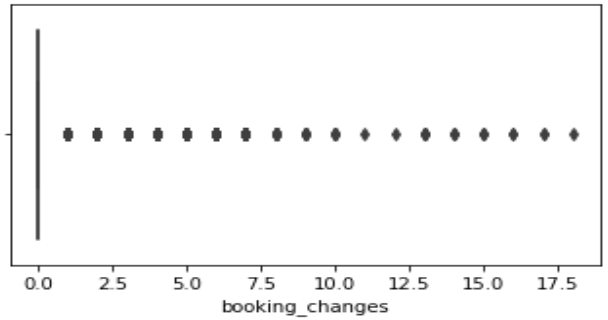


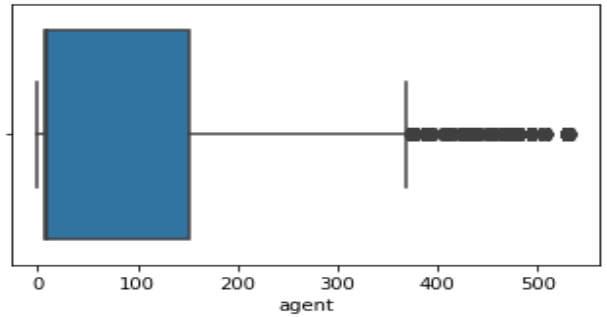


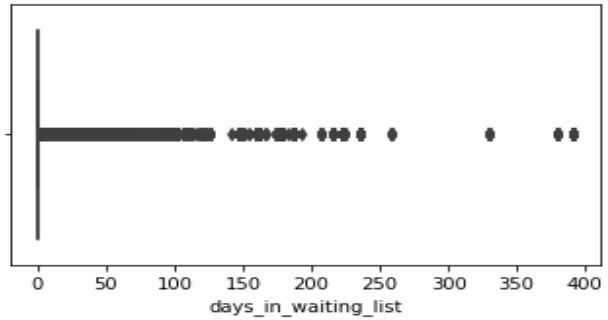


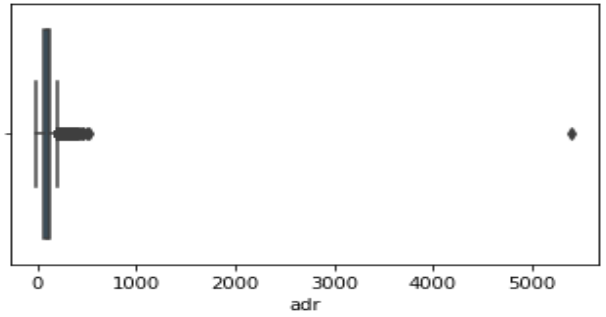


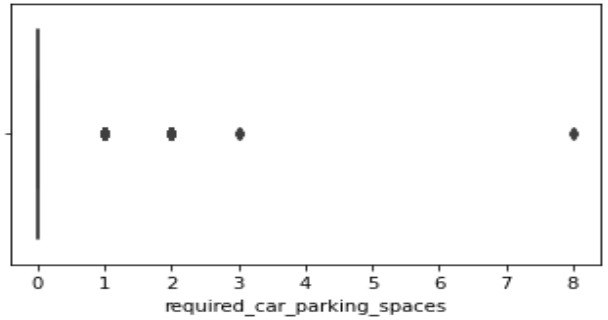


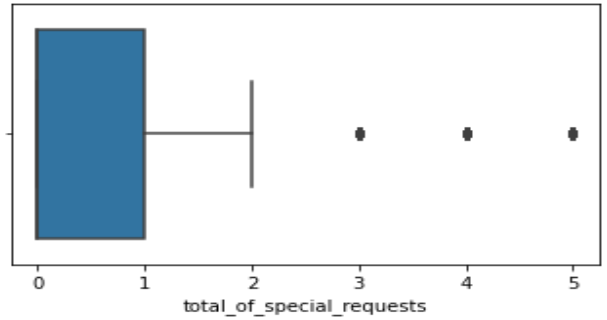






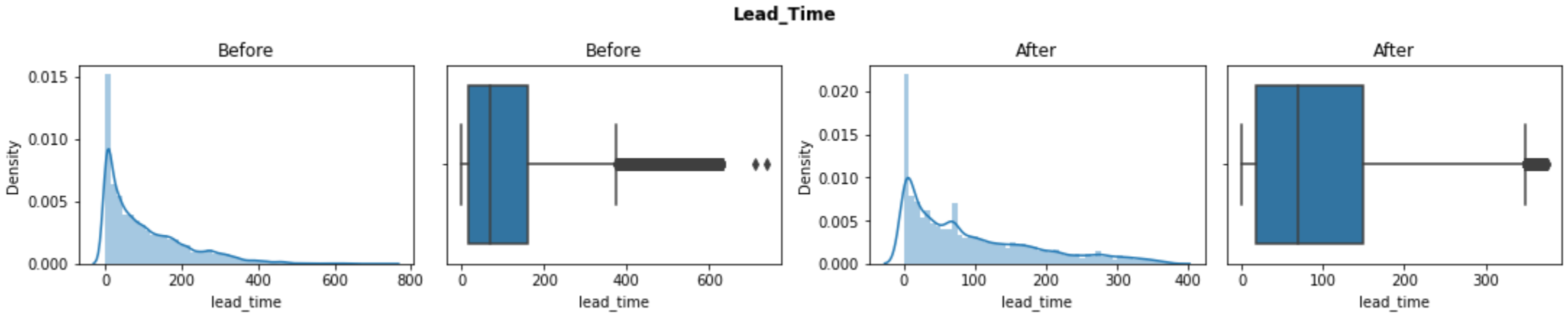




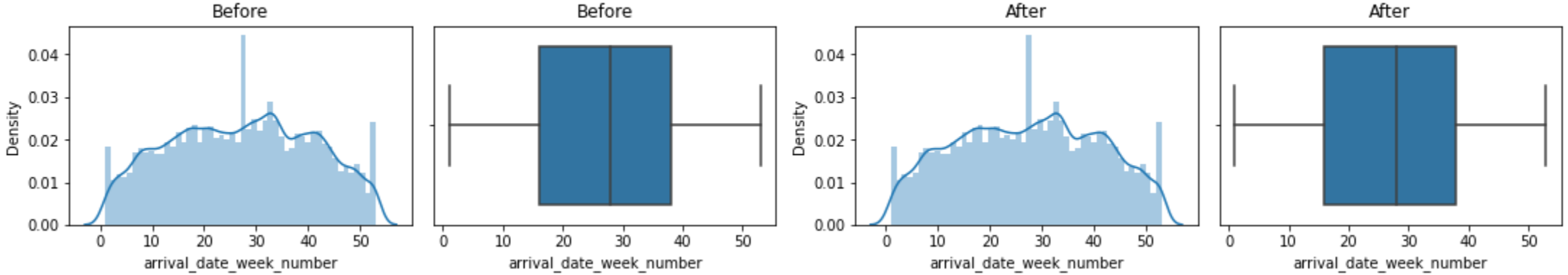


TREATING OUTLIER

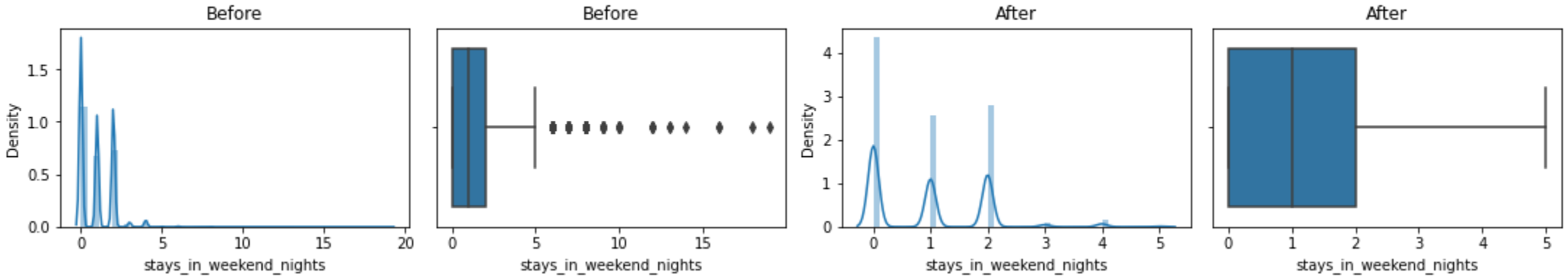
- Firstly we remove outliers by defining threshold based on common understanding.
- Now we removed outliers by standard methods and plotting graph , using IQR method and capping to define the range of inliners.
- Replacing the outliers with median value and capping them.



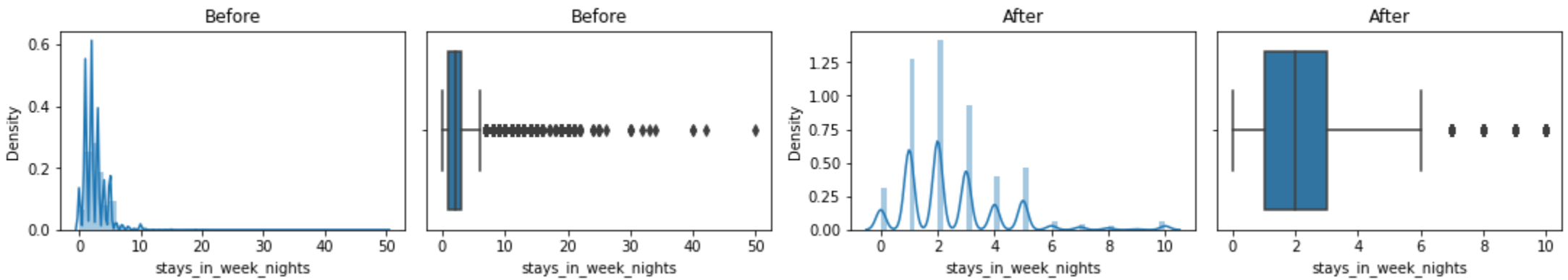
Arrival_Date_Week_Number

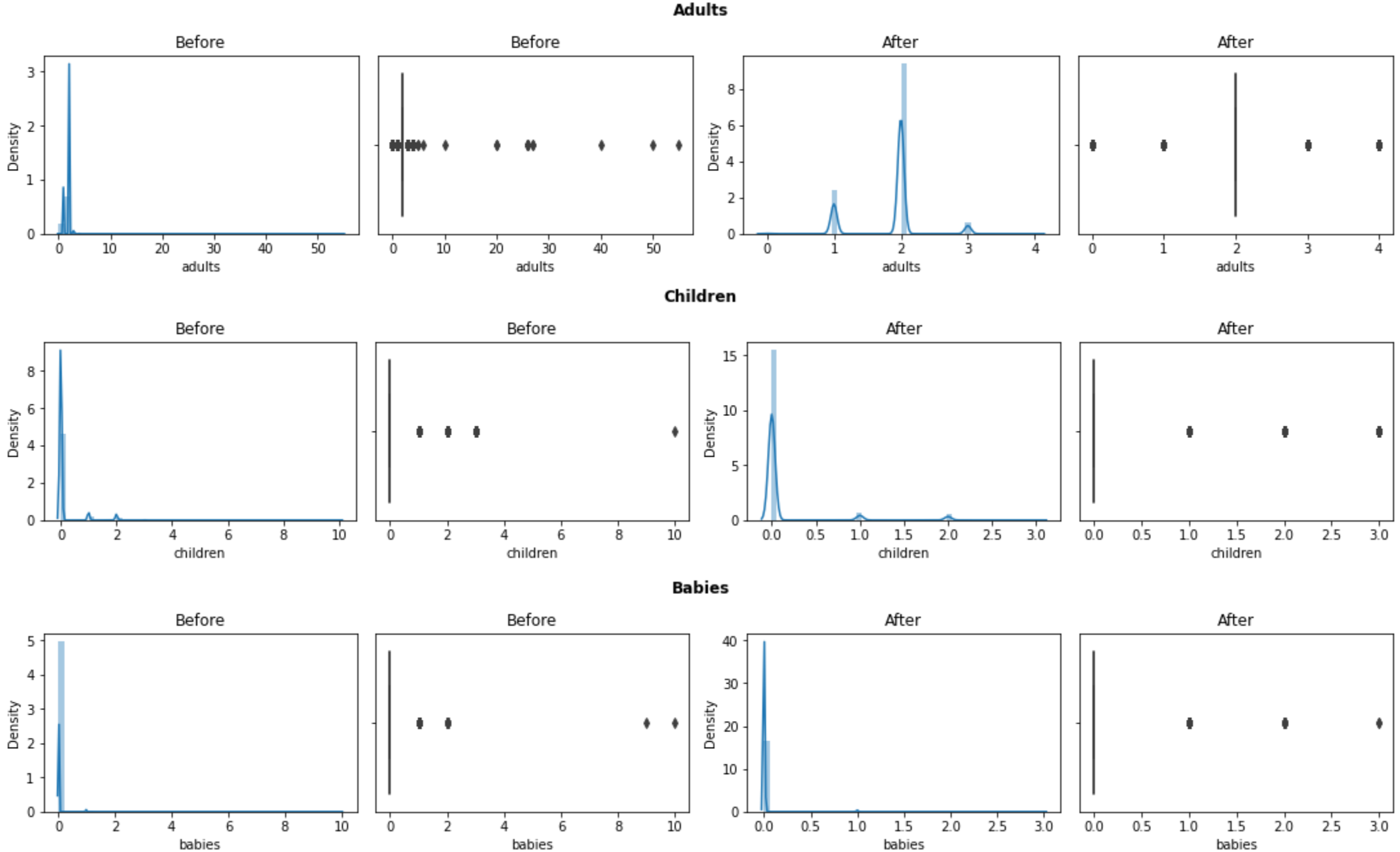


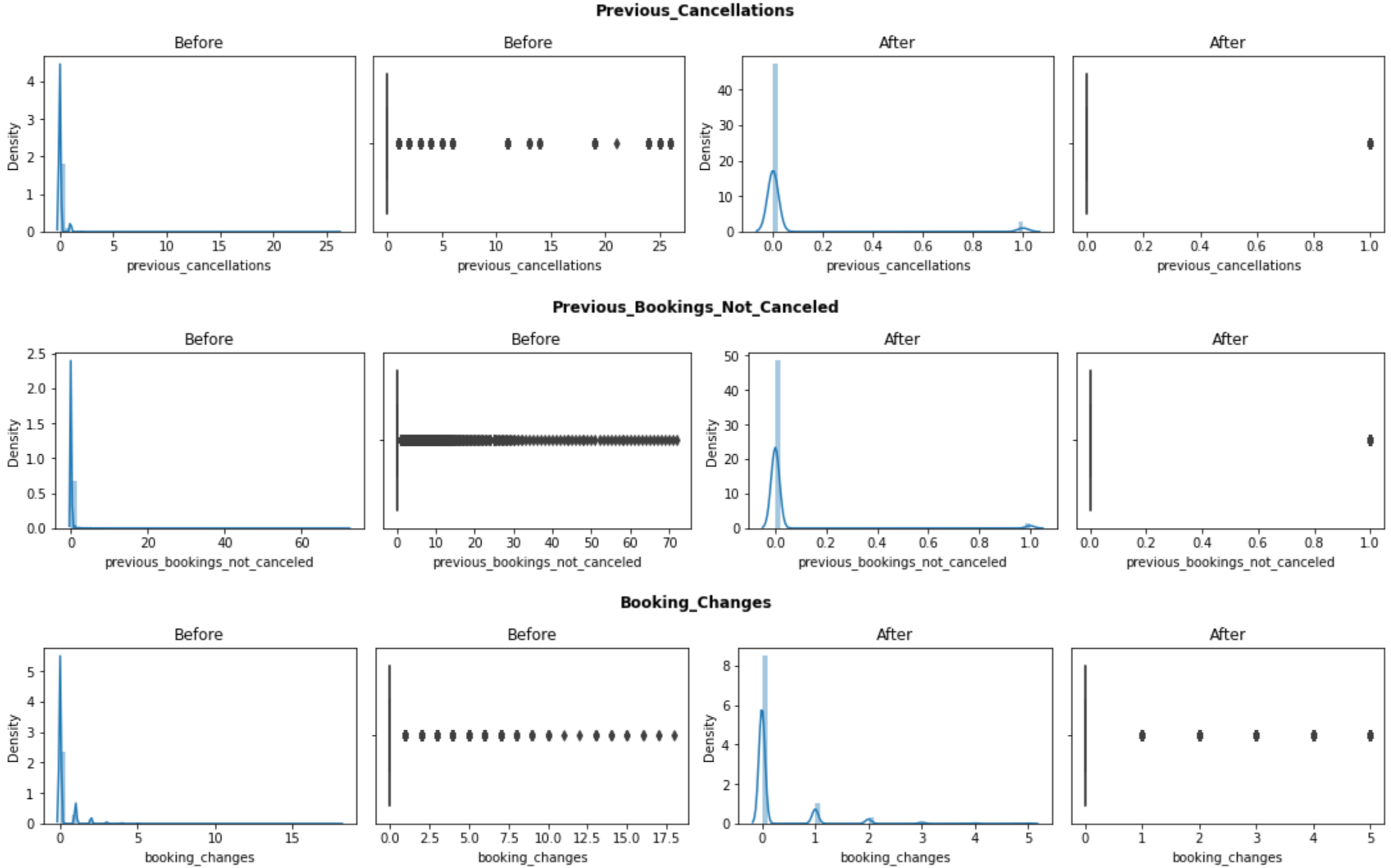
Stays_In_Weekend_Nights

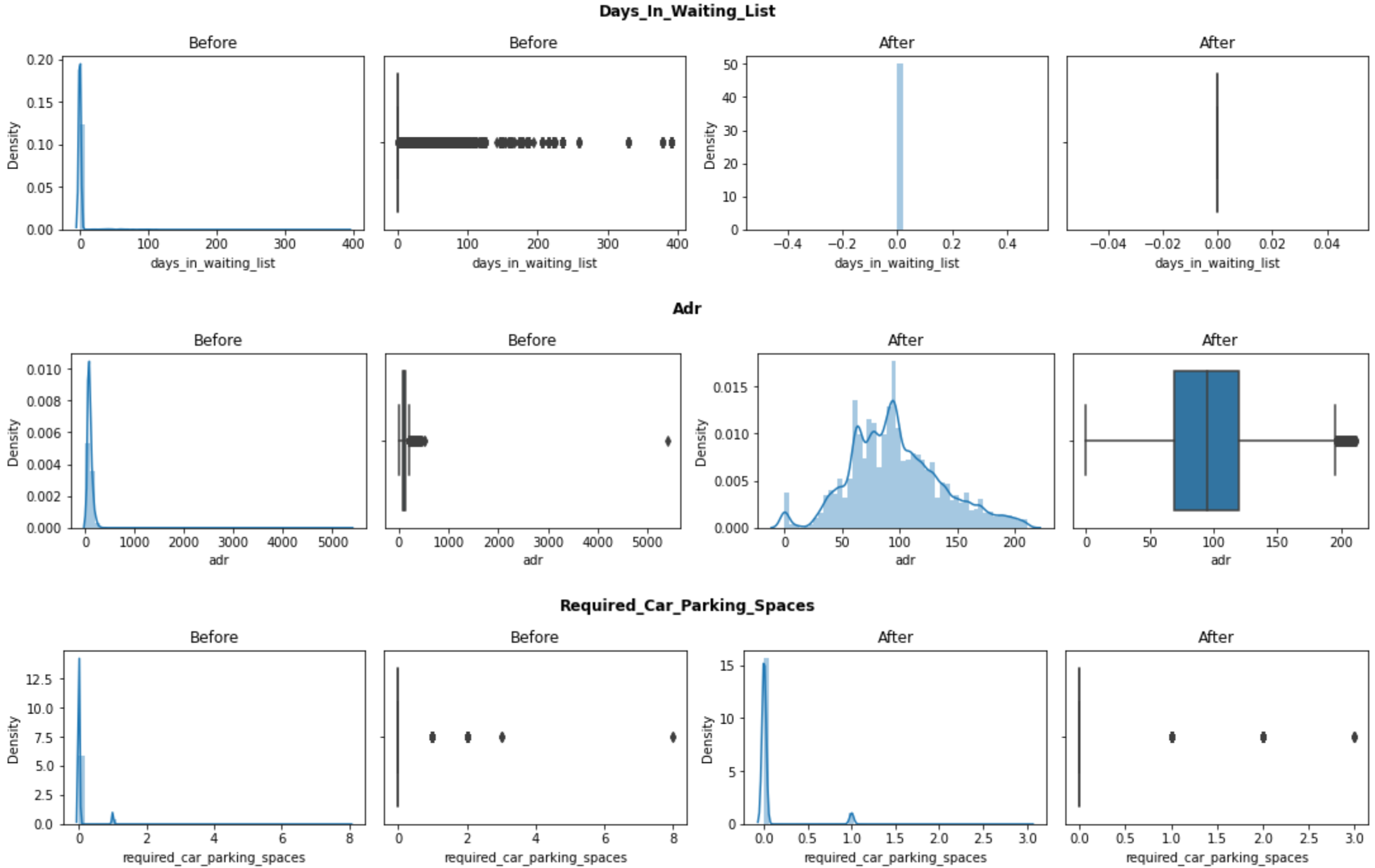


Stays_In_Week_Nights




Adults
Before
Before
After
After
Density
adults
Children
Before
Before
After
After
children
Babies
Before
Before
After
After
babies


Previous_Cancellations
Before
Before
After
After
Density
previous_cancellations
Previous_Bookings_Not_Canceled
previous_bookings_not_canceled
Booking_Changes
booking_changes


Days_In_Waiting_List
Before
Before
After
After
Density
days_in_waiting_list
Adr
Before
Before
After
After
Density
adr
Required_Car_Parking_Spaces
Before
Before
After
After
Density
required_car_parking_spaces

CLEANING AND MANIPULATING DATASETS

- Firstly we checked for duplicates and found 32052 duplicate values , after dropping all of them we are left with 87158 observation and 31 columns.
- Then we converted data type of relevant columns from float to integer, the converted columns were **Children**,**Agent** and **Adr** .
- After that we created some features

1.We combined the features ‘assigned_room_type’ and ‘reserved_room_type’ to make a new feature called “**same_room**”. To verify if the customer has allotted the same room which they reserved.

2.We joined ‘children’ and ‘babies’ column to “**Total_children**”. And also ‘total_children’ and ‘adults’ to “total_members”.

3.We also ‘stays_in_weekend_nights’ and ‘stays_in_week_nights’ to “**Total_nights**”.

4. We made a column called “**money_per_person**” in which we divided ‘total_nights’by ‘total_members’ and multiplied it by ‘adr’.

5. We made a new feature named “**guest_type**”,where if ‘total_members’=1 then it is “single” and if ‘total_members’ =2 then it is “couple”or“family”.

6. Then we added a new column “**lead_time_category**” where we divided it in three parts low, medium and high.

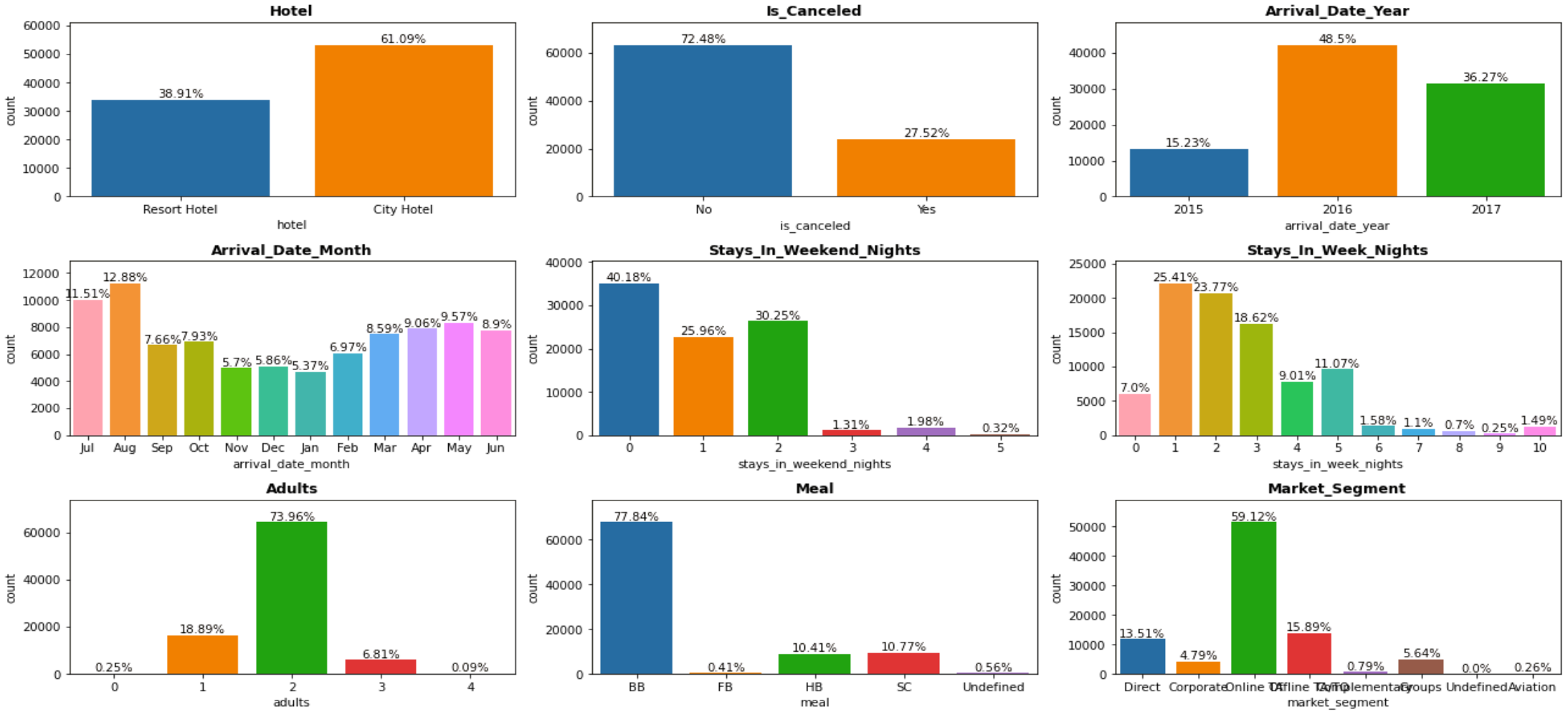
7. Then we dropped the columns which were used to form a feature.

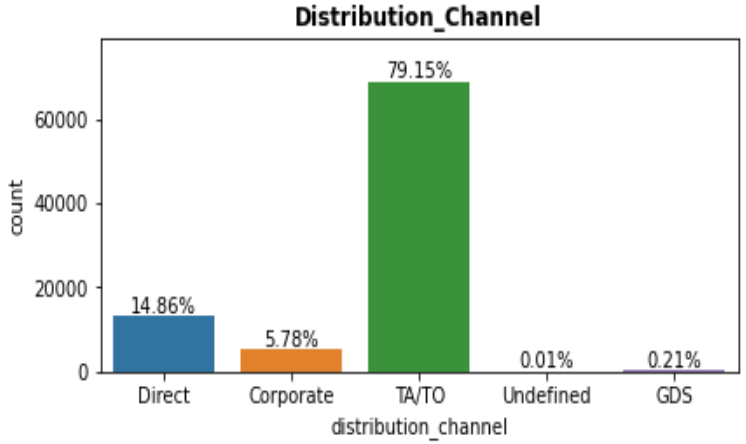
8. After cleaning and manipulating we are left with 87158 observation and 33 columns, on which we can perform our **EDA**.

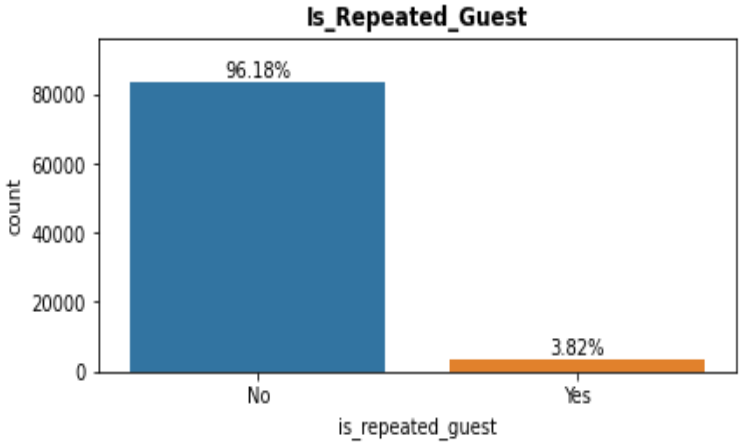
EXPLORATORY DATA ANALYSIS

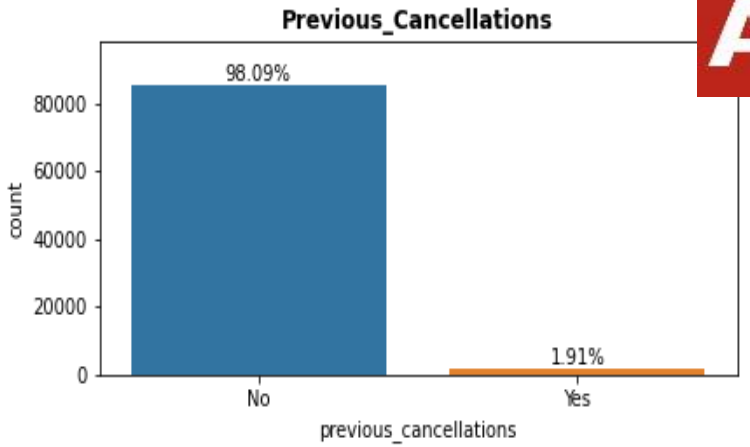
UNIVARIATE ANALYSIS

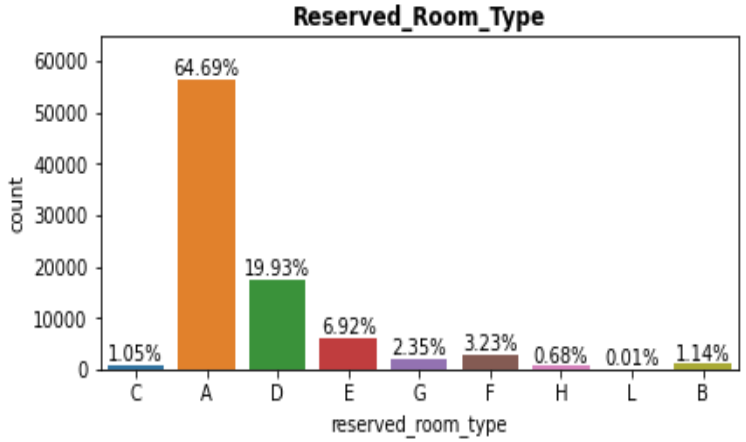
We have plotted the count plots for each target variables

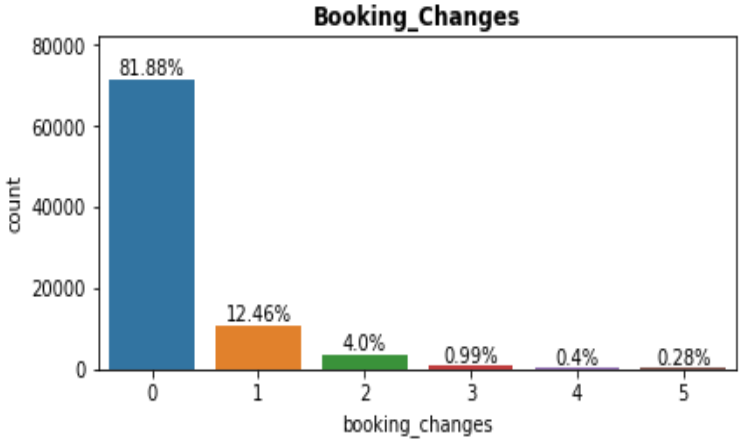


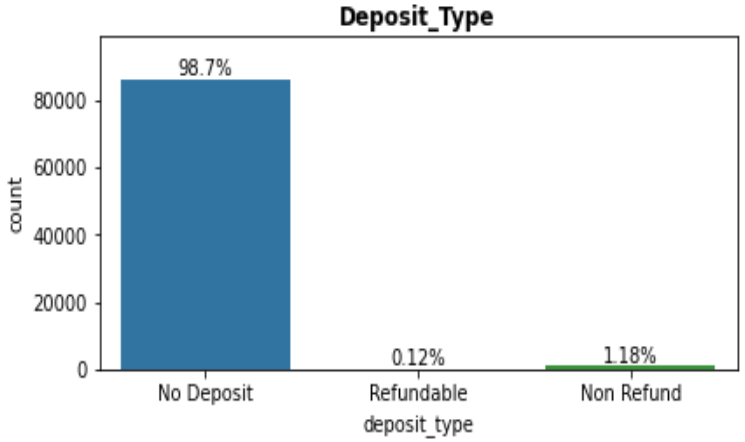

Distribution_Channel
count
distribution_channel
Direct
Corporate
TA/TO
Undefined
GDS
14.86%
5.78%
79.15%
0.01%
0.21%

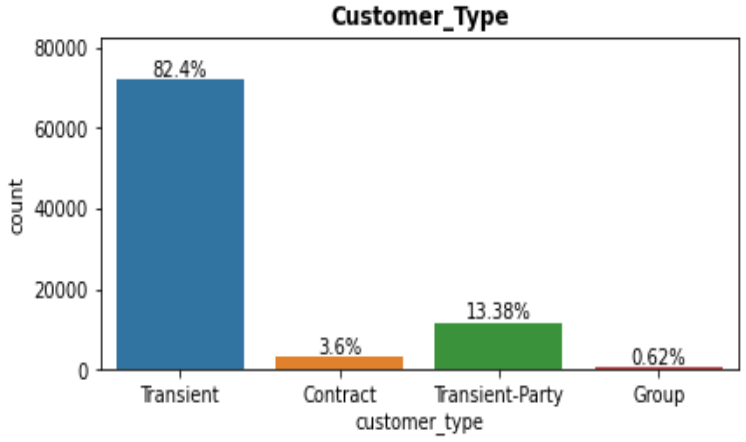

Is_Repeated_Guest
count
is_repeated_guest
No
Yes
96.18%
3.82%

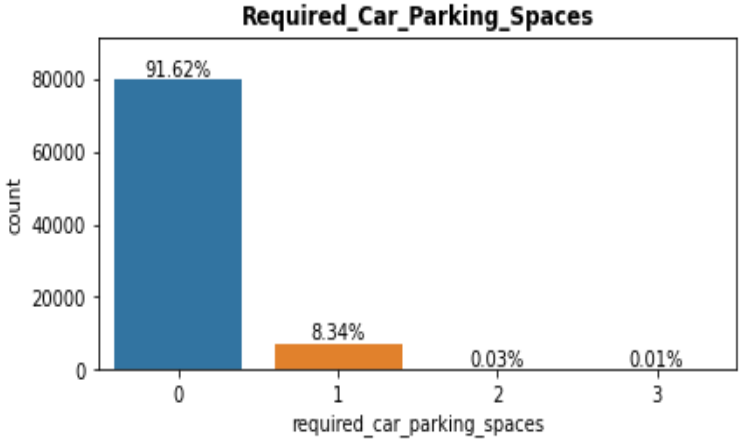

Previous_Cancellations
count
previous_cancellations
No
Yes
98.09%
1.91%

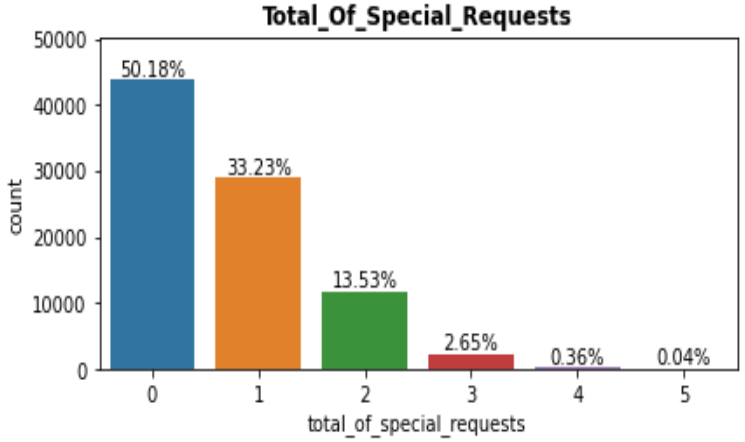

Reserved_Room_Type
count
reserved_room_type
C
A
D
E
G
F
H
L
B
1.05%
64.69%
19.93%
6.92%
2.35%
3.23%
0.68%
0.01%
1.14%

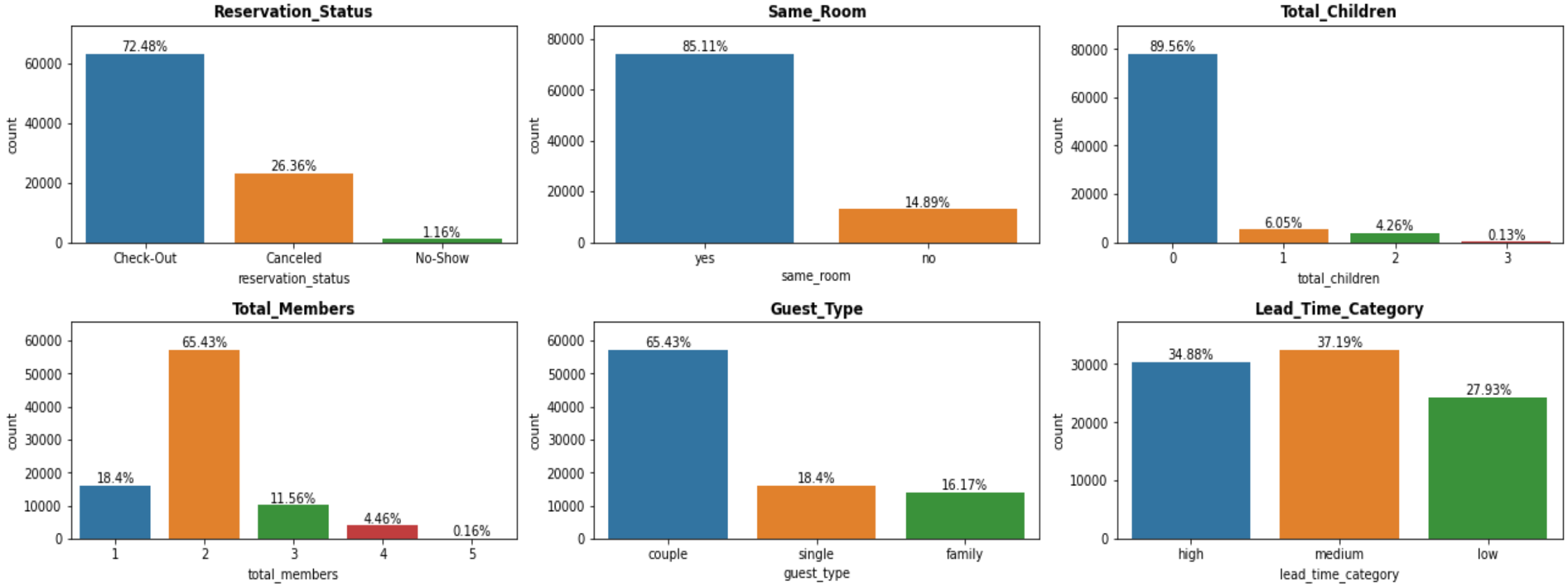

Booking_Changes
count
booking_changes
0
1
2
3
4
5
81.88%
12.46%
4.0%
0.99%
0.4%
0.28%


Deposit_Type
count
deposit_type
No Deposit
Refundable
Non Refund
98.7%
0.12%
1.18%


Customer_Type
count
customer_type
Transient
Contract
Transient-Party
Group
82.4%
3.6%
13.38%
0.62%


Required_Car_Parking_Spaces
count
required_car_parking_spaces
0
1
2
3
91.62%
8.34%
0.03%
0.01%


Total_Of_Special_Requests
count
total_of_special_requests
0
1
2
3
4
5
50.18%
33.23%
13.53%
2.65%
0.36%
0.04%

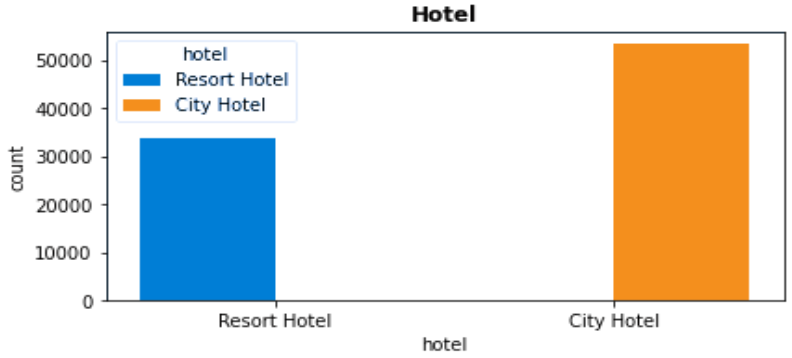

Reservation_Status
72.48%
26.36%
1.16%
Check-Out
Canceled
No-Show
reservation_status
count
Same_Room
85.11%
14.89%
yes
no
same_room
Total_Children
89.56%
6.05%
4.26%
0.13%
total_children
Total_Members
65.43%
18.4%
11.56%
4.46%
0.16%
total_members
Guest_Type
65.43%
18.4%
16.17%
couple
single
family
guest_type
Lead_Time_Category
34.88%
37.19%
27.93%
high
medium
low
lead_time_category

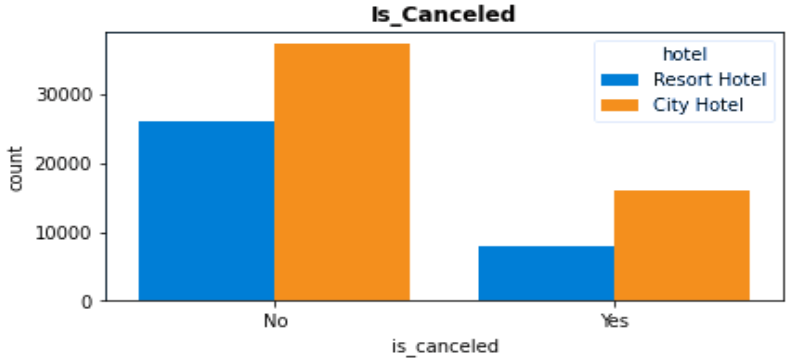
OBSERVATIONS

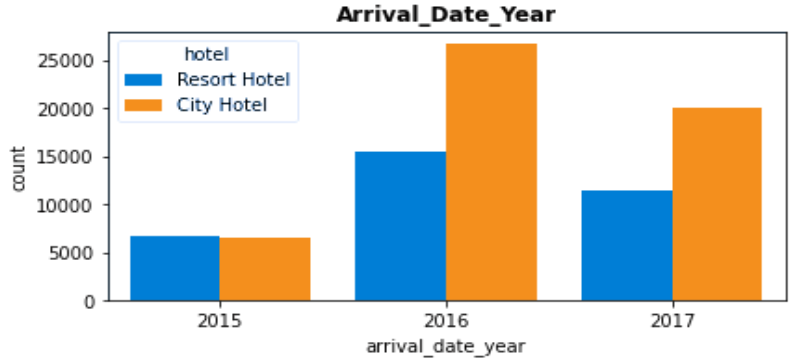
> More than 60% of the bookings are of City Hotel.
> Nearly one third of the bookings are cancelled.
> There was an annual 218.5% rise in hotel bookings in 2016 which dropped down by 25.2% in 2017
> August is the most preferred month by people for bookings.
> Most preferred meal is BB(Bread and Breakfast).
> Most of the customers are coming through Online.
> Top distribution channel is TA/TO
> Only 3.82% of the guests are arriving again.
> People who cancel bookings do not really book again.
> Room type A is the most preferred one.
> People don’t want to pre-deposit the money.
> More than 90% of people don’t require any parking space.
> Around 15% of guests are not assigned with their preferred room.
> Around 10% of the guests arrive with children.
> At least 2 people arrive 80% of the times.
> Around 15% of the people visit with their family.

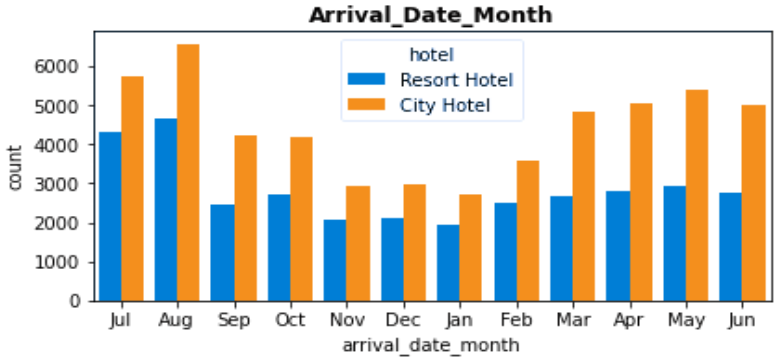
HOTELWISE ANALYSIS

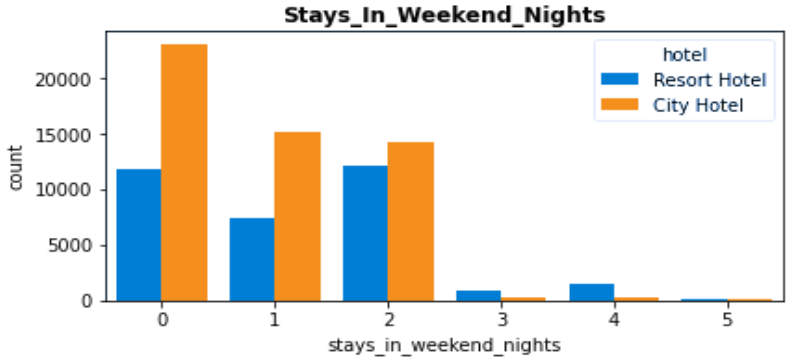
FEATURE VISUALIZATION BASED ON HOTEL

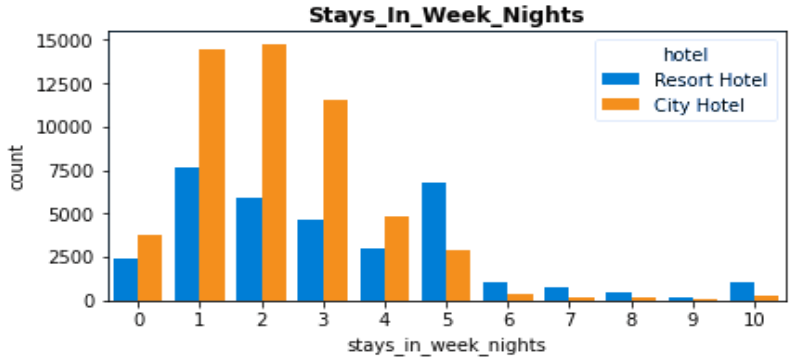


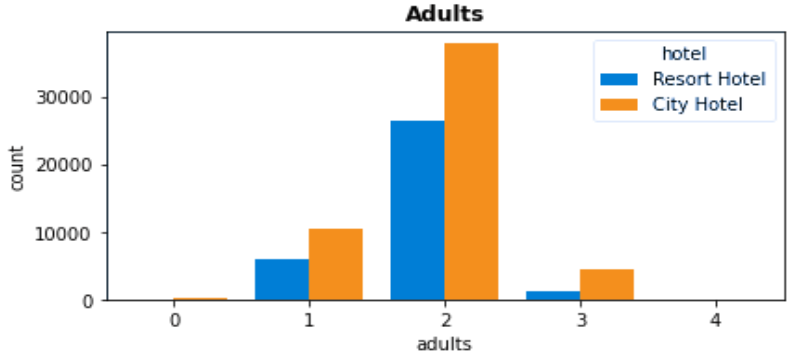


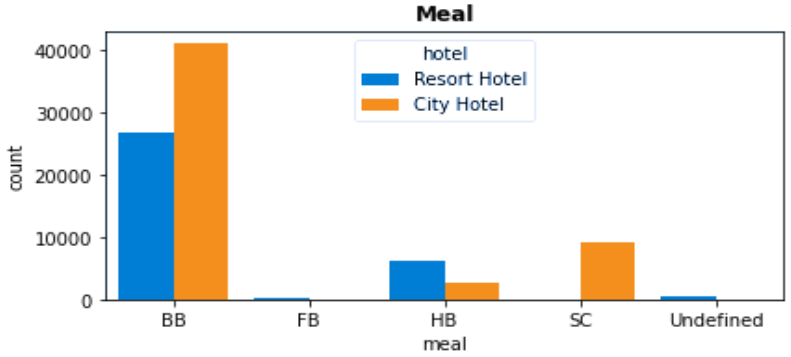


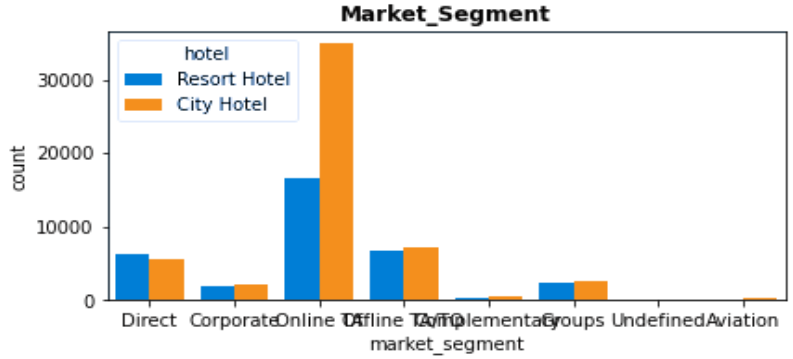




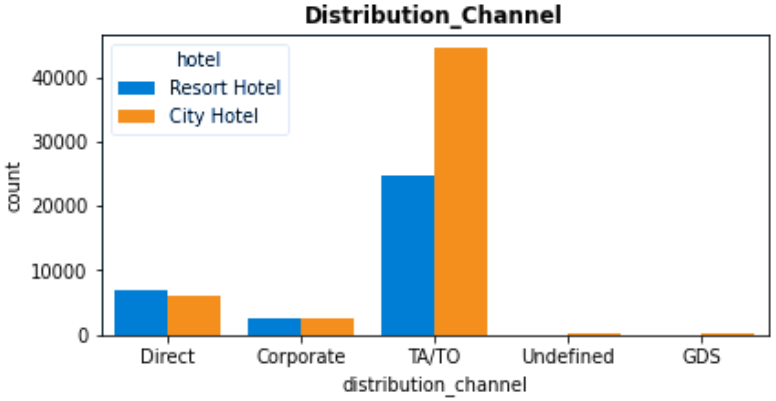


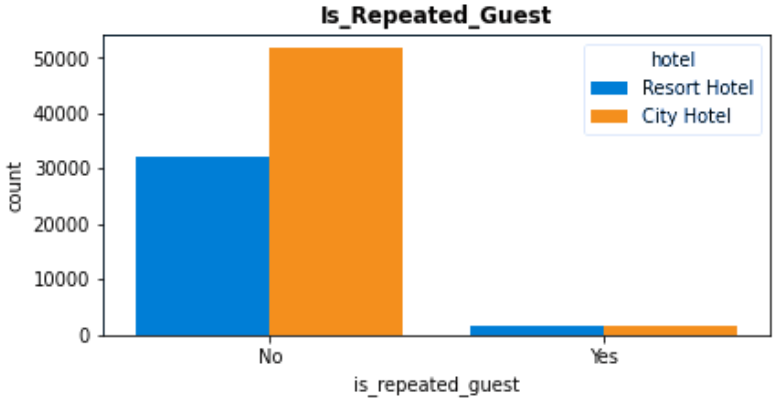


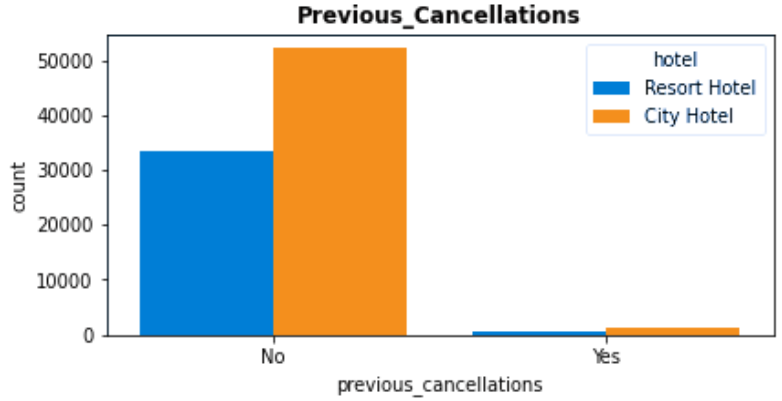


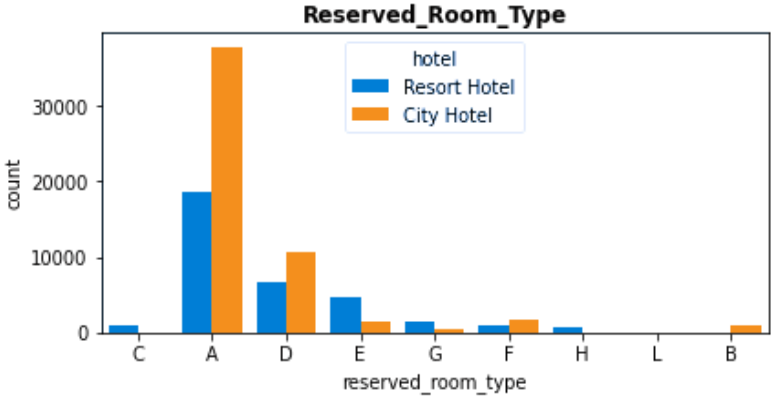


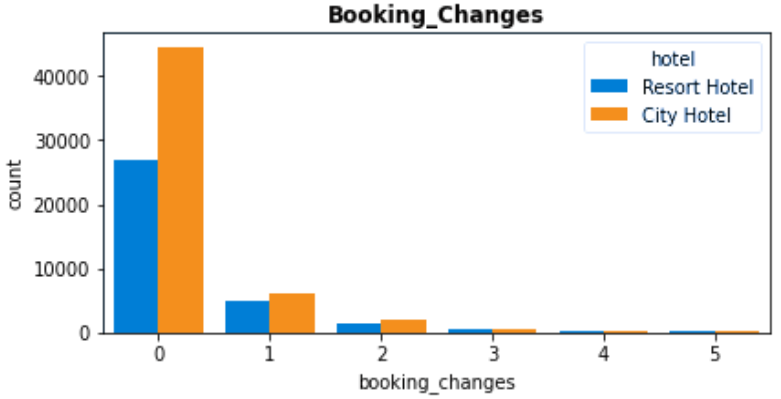
FEATURE VISUALIZATION BASED ON HOTEL

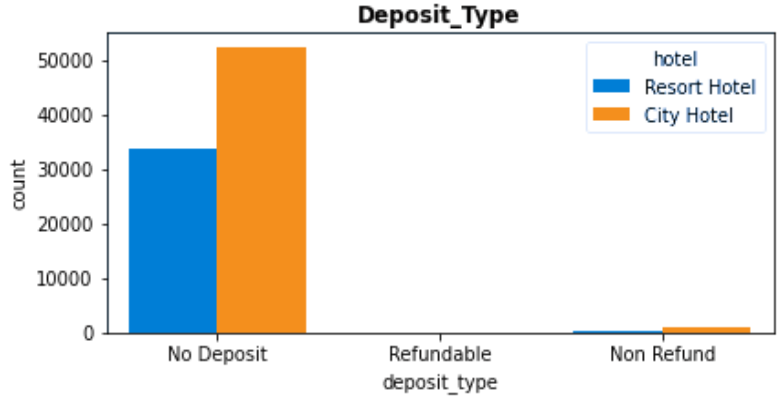


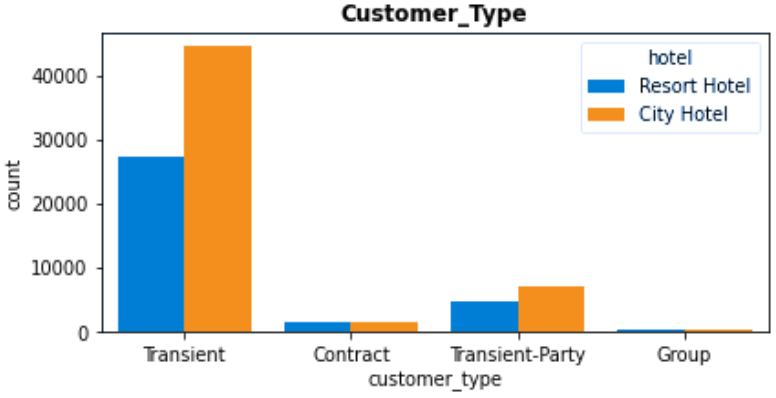


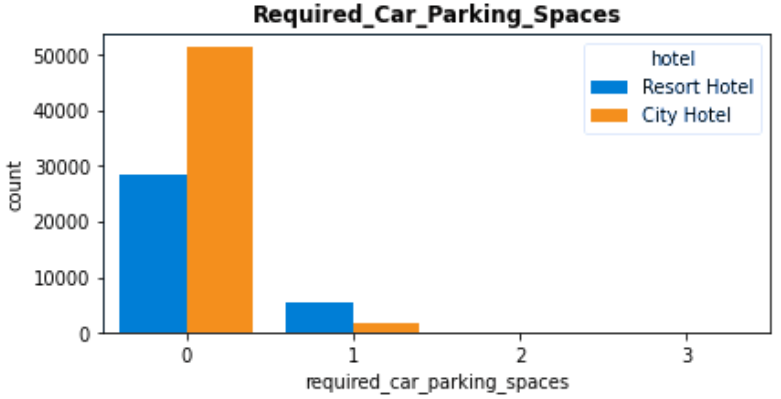


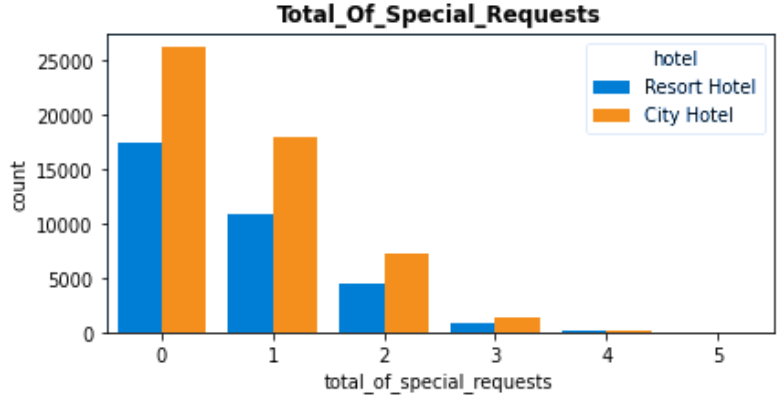




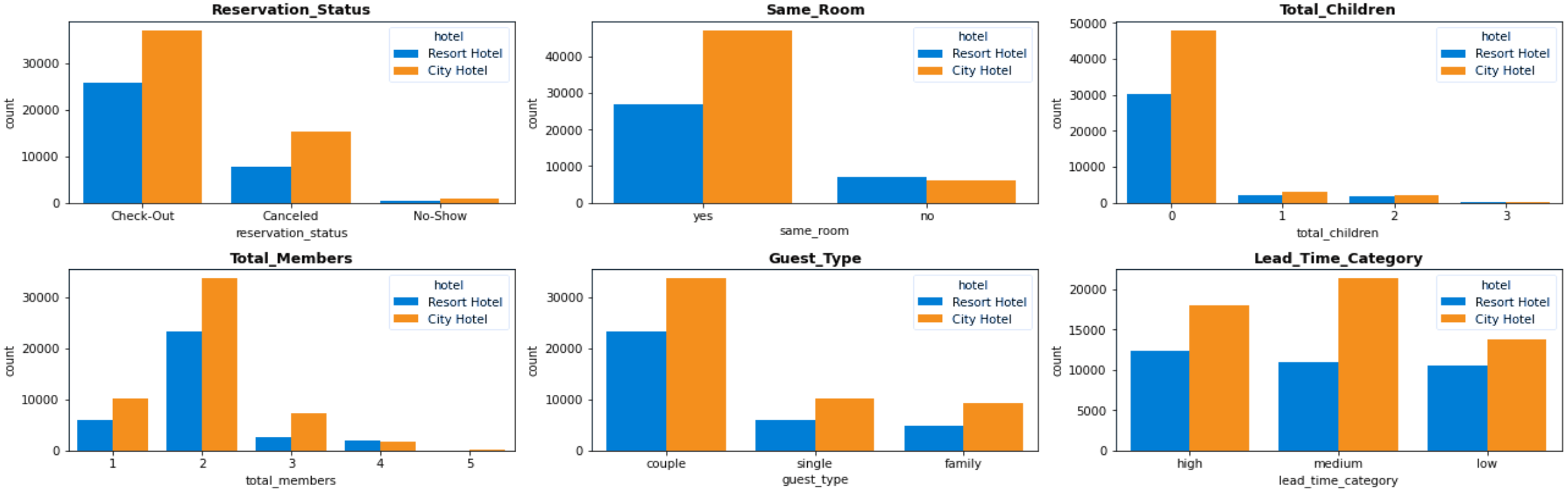








FEATURE VISUALIZATION BASED ON HOTEL

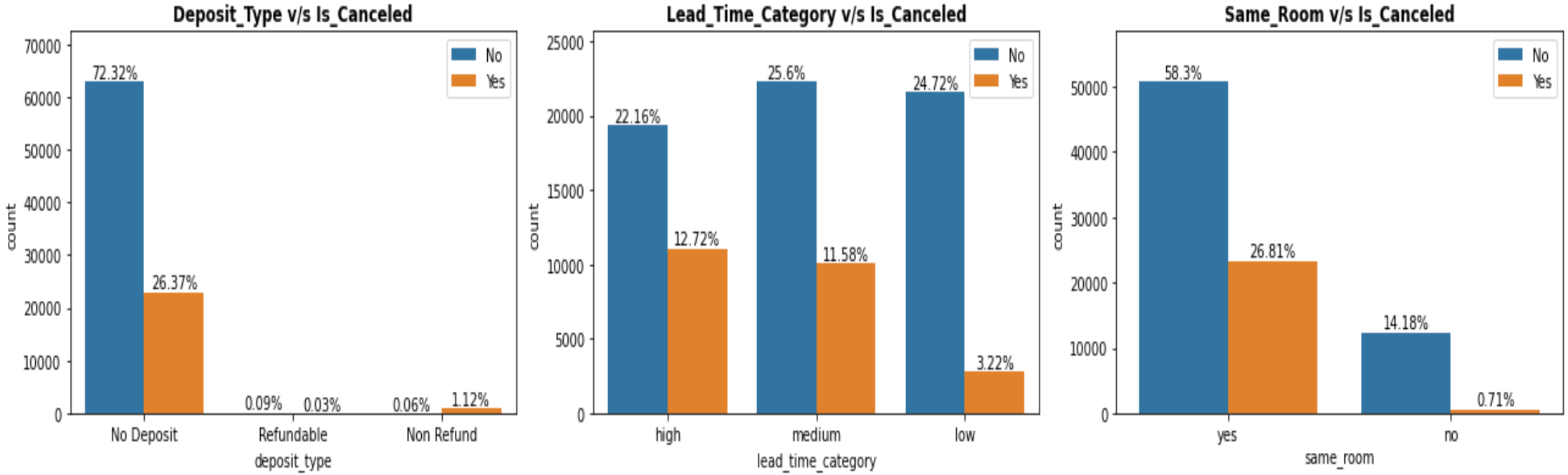


As per the above plots, we can lead to some checks according to their appearances which are as follows :
online booking is most preferred in both the hotel type,
repeated number of guests are coming most from the city hotel,
A room type is mostly popular amongst its category ,
From city prospective people are choosing no deposit as booking their hotel ,
number of special request count is equal or less than three, couples responsible for most of the valid bookings .

HOTELWISE OBSERVATION

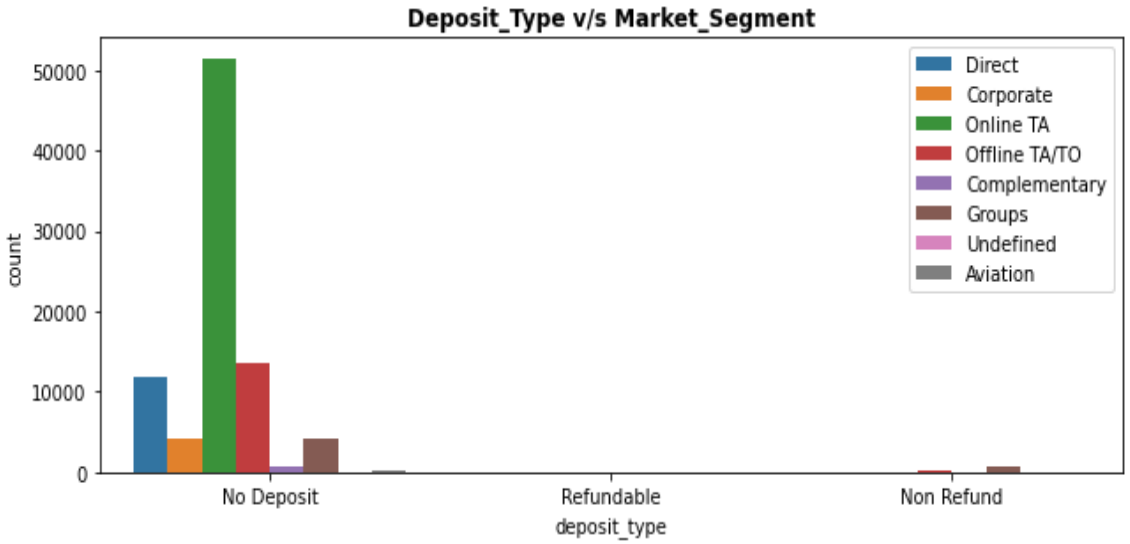
- People prefer City Hotel more as compared to Resort Hotel.
- Cancellation % for city hotel is slightly more than that of city hotel.
- Resort hotel was leading in 2015 in terms of bookings. But City hotel defeated it for next two years.
- People who stay for more than 4 nights prefer Resort hotel the most.
- People who prefer HB(Half board) meal plan (Breakfast and evening meals included) also prefer Resort hotel the most. City hotel is leading in rest all meal plans.
- Resort hotel is leading in Direct marketing but lagging with a huge margin in online segment.
- People who require car parking spaces also prefer Resort hotel. Otherwise City hotel is preferred the most
- City hotel assigns the same room as reserved by the guests most of the time. While Resort hotel fails to do so more for more than 25% of guests

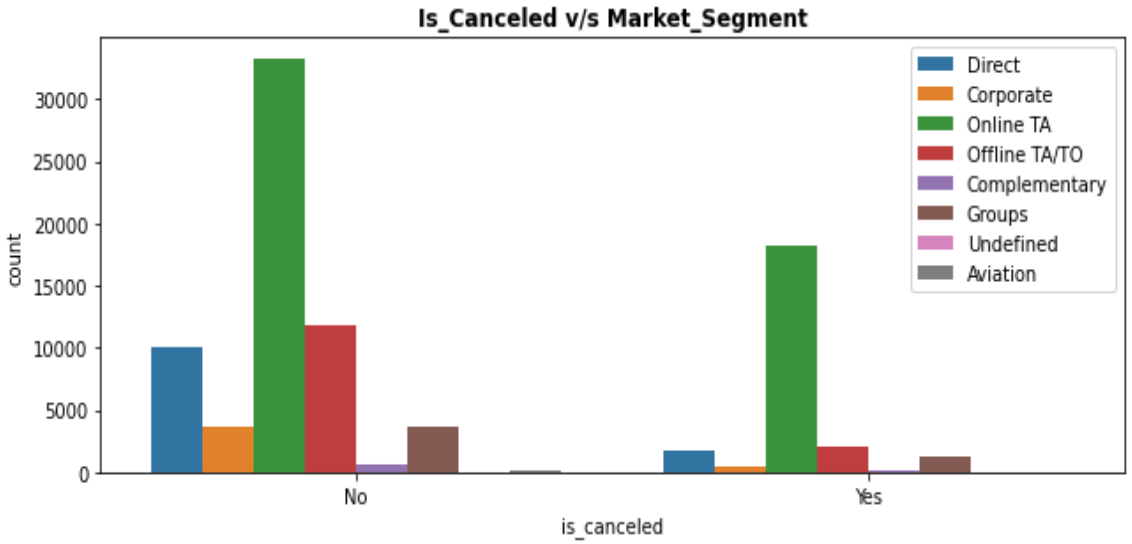
BIVARIATE ANALYSIS



- Most of the bookings are cancelled where there was no deposit made. That's because most of people didn't make deposits. So it's not a reason for cancellation.
- If the lead time is low, less people cancel the bookings. But If the bookings are made more than 15 days in advance, there are comparatively high chances of cancellation. We can conclude this because the data is almost equally distributed among low medium and high lead time.
- Not having assigned the same room is not a reason for cancellation. As only 0.7% of bookings were cancelled when the same room was not assigned.

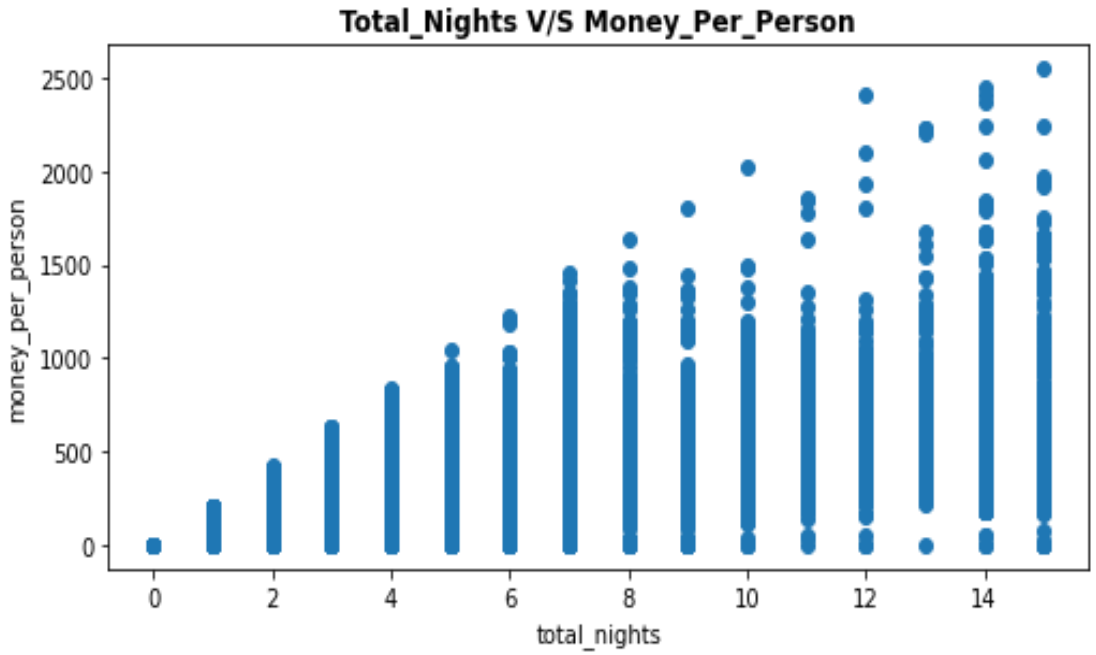
MARKET SEGMENT ANALYSIS

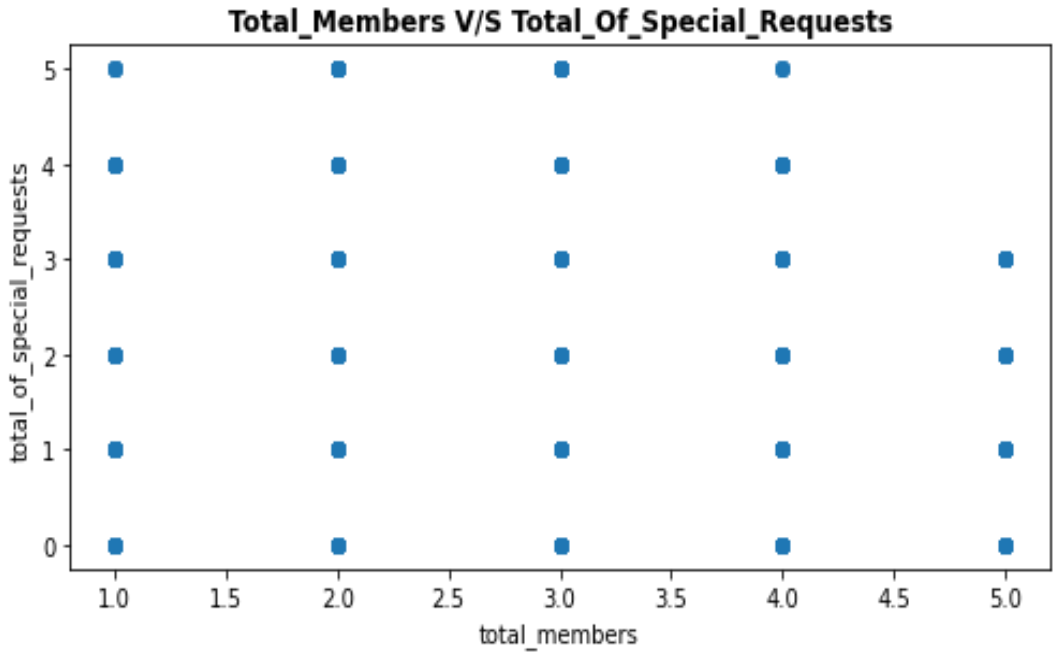




- More than 30% of the online bookings are cancelled. Direct bookings have very less cancellation%.

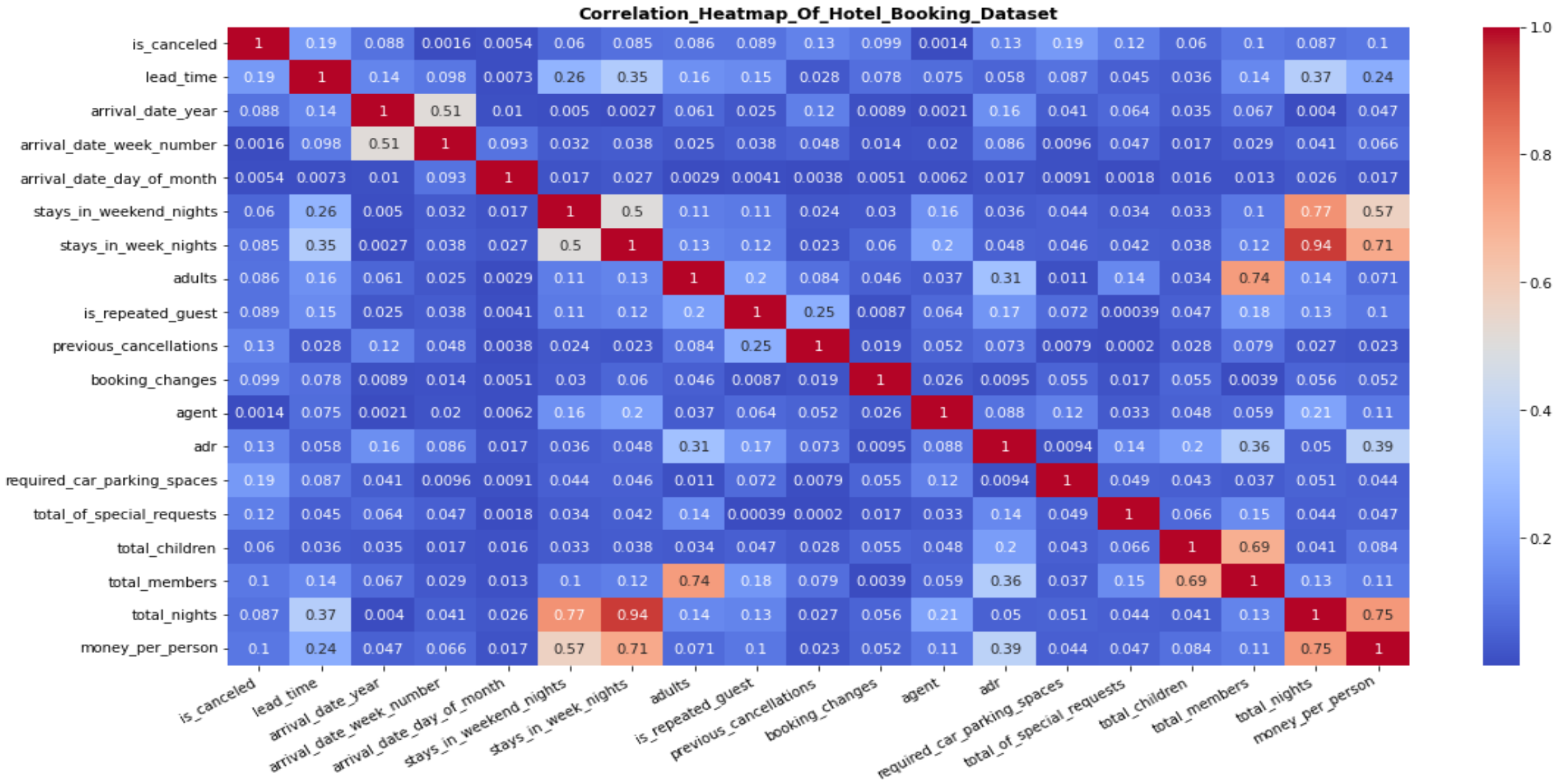
VISUALIZATION OF OPTIMAL LENGTH OF STAY AND RELATION B/W SPECIAL REQUEST AND TOTAL MEMBERS ARRIVED





- 3 Nights seems economical for stay.
- Number of Special requests seems have very less related with total members. So we can simply take the average of it to get the number of special requests

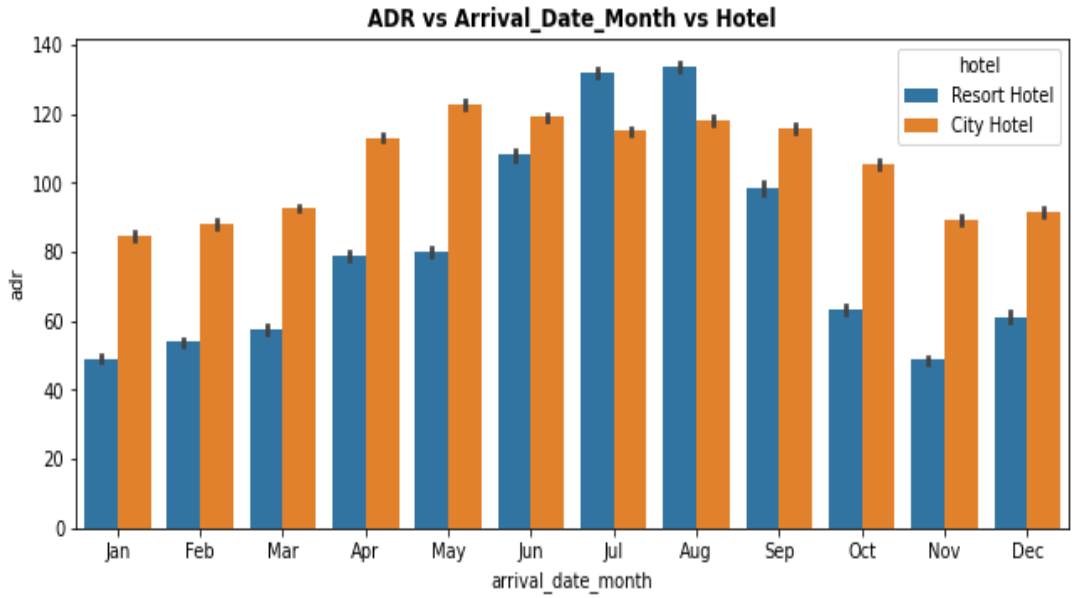
MULTIVARIATE ANALYSIS

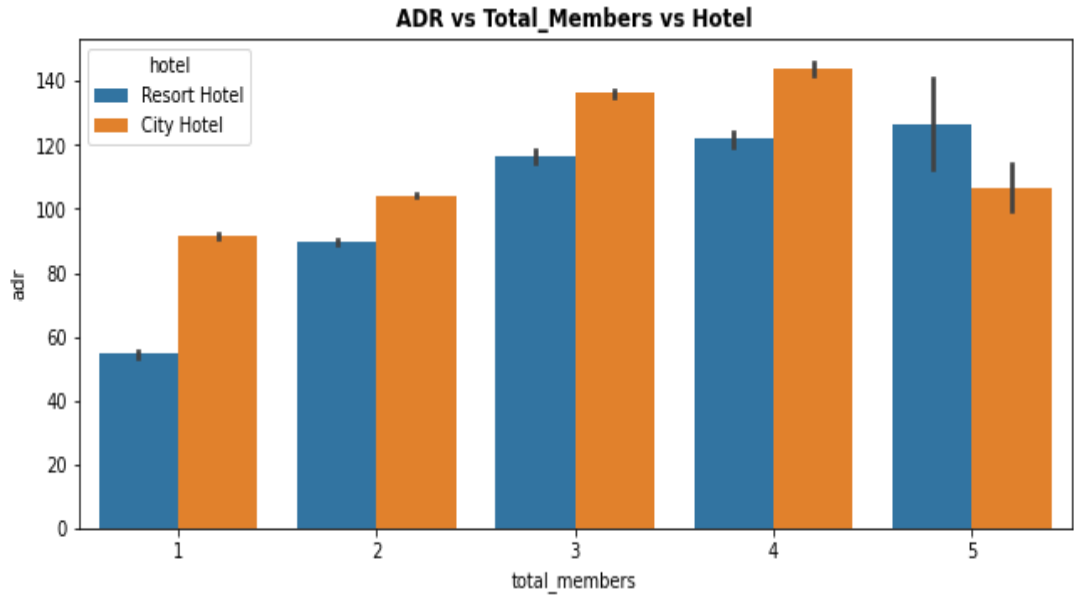


OBSERVATION

- In the heatmap, It shows some high correlations between few variables, that’s because we have created some new columns from existing columns and have not dropped it later.
- Total special requests depends more on total members arrived.
- Average daily revenue depends more on total special requests and total members as compared to other variables.

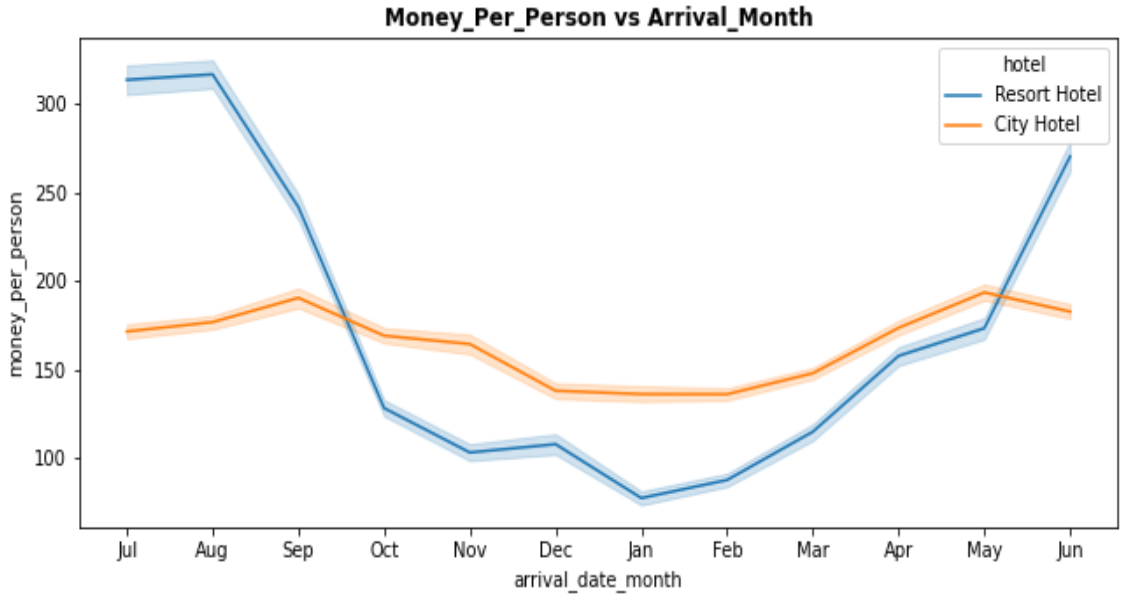
ADR based on arrival month and total_members for both hotel

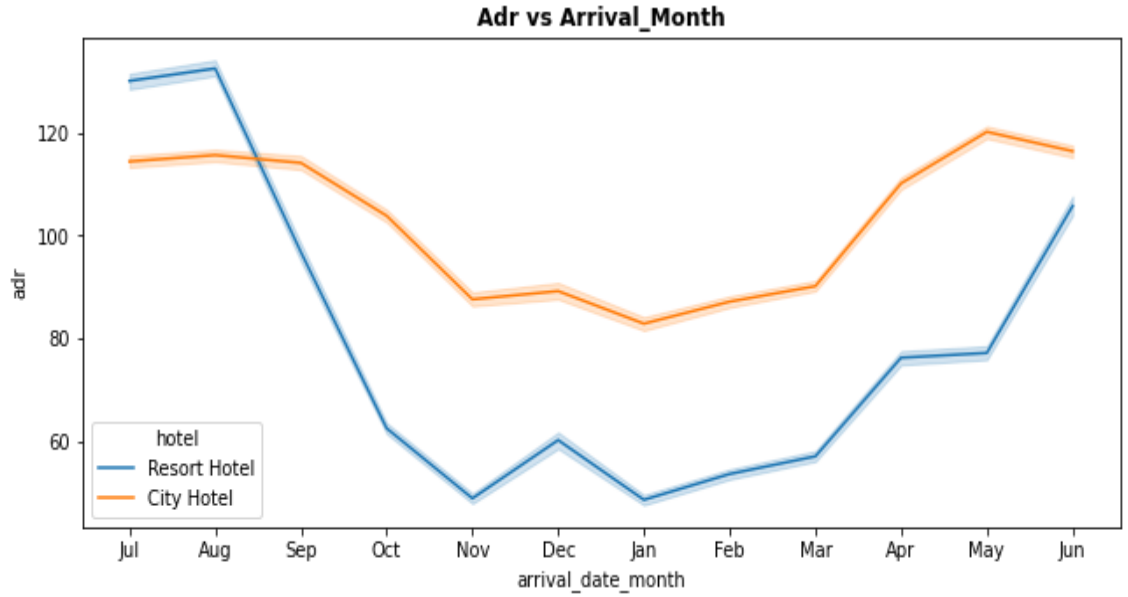




- Average Daily Revenue rises from January to Aug then again falls down from Aug to Dec.
- Resort hotel leads in ADR only during the peak months July and August. Rest all months City hotel has high ADR.
- April to September is the semester with high ADR for both hotels
- ADR for Resort hotel is directly proportional to total members.
- Seems like City hotel is giving some discount offers for members more than 4 because the ADR has drastic decrease for members more than 4.

CHECKING BEST TIME TO BOOK HOTEL





- Best time to book a hotel is in January. As money spent is lesser.
- City hotel seems consistent with the price throughout the year.

ADDITIONAL CHECKS

- Average time taken for customer arrival after making reservation : 77 days
- Average nights spent by visitors: 3
- Average money spent by visitors: 186 Bucks.

AGENTS WITH MOST BOOKINGS

INDEX	AGENT	BOOKINGS
0	9	17193
1	0	10566
2	240	8074
3	7	2858
4	14	2759

COUNTRY'S WITH MOST BOOKINGS

COUNTRY	NO. OF BOOKINGS
PRT	27309
GBR	10421
FRA	8821
ESP	7237
DEU	5385

CONCLUSION

- Top Hotel- City Hotel. Top meal- Bread and Breakfast. Top Agent- Agent No. 9. Top room type- A
- One out of every three bookings are cancelled.
- People prefer to tour more in August.
- Most preferred meal is BB(Bread and Breakfast.
- Online marketing is the best way to attract customers.
- People do not want to pre-deposit the money for booking.
- Only 10% of people require parking space.
- Most of the visitors are couples.
- Resort hotel is preferred mostly for longer stay,day time stays. and when the parking needed.
- More than 15 days advance bookings have high chances of cancellation.
- Assigning different room is not a reason for cancellation.
- Direct bookings have very less cancellation%.
- Best time to book a hotel is in January.
- Average days in advance booking : 77 days

CONCLUSION

- Average nights spent by visitors: 3
- Most visitors are from these countries: Portugal, Britain, France, Spain and Germany.
- Total Special requests and the revenue depends more on total members arrived.