CAPSTONE PROJECT

HOTEL BOOKING ANALYSIS (EDA)

INTRODUCTION

We are here to explore a hotel booking dataset to discover important factors that govern the bookings. This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

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

ATTRIBUTE INFORMATION

- **hotel:** name of hotel whether City hotel or Resort Hotel
- **is_canceled:** (0 or 1) Indicates whether booking was cancelled or not.
- **lead_time:** The time between reservation and actual arrival.
- **arrival_date_year:** Year of arrival date.
- **arrival_date_month:** Month name of arrival date.
- **arrival_date_week_number:** Week number on arrival date.
- **arrival_date_day_of_month:** Day of the month of arrival date.
- **stays_in_weekend_nights:** Number of weekend nights the guest stayed or booked to stay at the hotel.
- **stays_in_week_nights:** Number of week nights the guest stayed or booked to stay at the hotel.
- **Adults, Children, Babies :** Number of adults, children and babies arriving.
- **Meal:** Type of meal booked.

ATTRIBUTE INFORMATION

- **Country:** The origin country of the guests.
- **market_segment:** Shows how the reservation was made and what is the purpose of reservation.
- **distribution_channel:** The medium through which the booking was made.
- **is_repeated_guest:** (0 or 1) Indicates whether or not the booking is of a repeated guest.
- **previous_cancellations:** (0 or 1) Indicates whether or not the guest has previous cancellations.
- **reserved_room_type:** Type of room booked.
- **assigned_room_type:** Type of room allotted /assigned.
- **booking_changes:** Number of changes/amendments made to the booking.
- **deposit_type:** Whether refundable/non-refundable/No-deposit made.
- **Agent:** ID of the travel agency that made the booking.

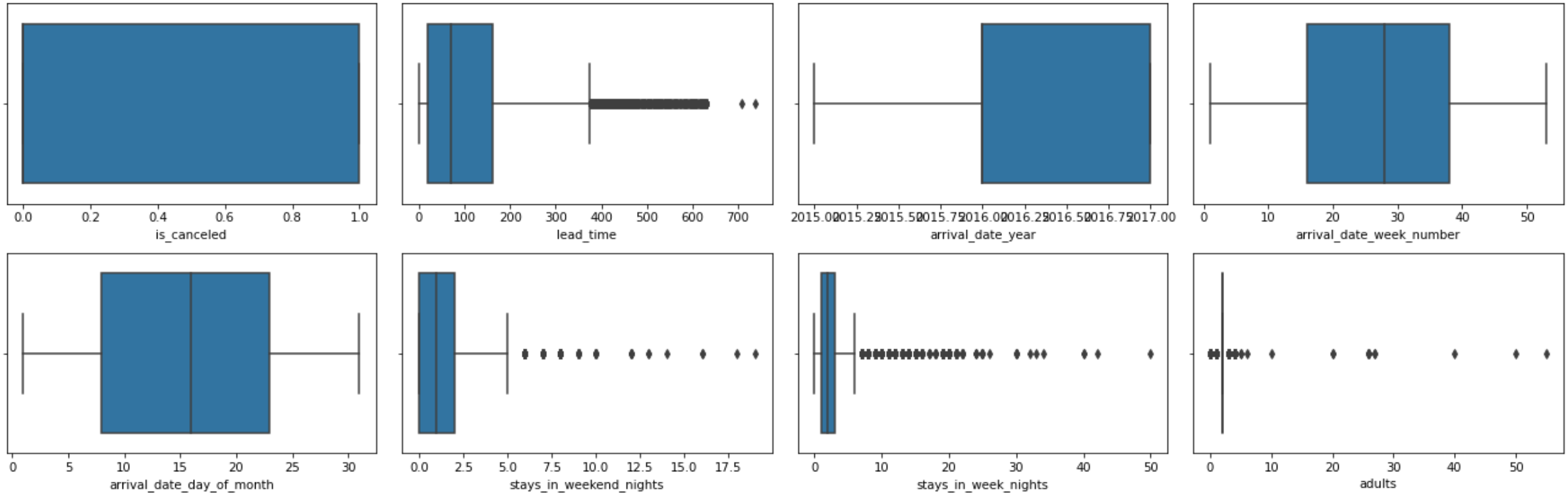
ATTRIBUTE INFORMATION

- **Company:** The name of company that made the booking or is responsible for paying for the booking.
- **days_in_waiting_list:** Number of days the booking was in the waiting list before it was confirmed to the customer.
- **customer_type:** Type of customers(Transient, group, etc.)
- **Adr:** Average daily rate is the average revenue that a hotel receives for each occupied guest room per day.
- **required_car_parking_spaces:** Number of car parking spaces required.
- **total_of_special_requests:** Number of special requests made.
- **reservation_status:** Self explanatory.
- **reservation_status_date:** Self explanatory.

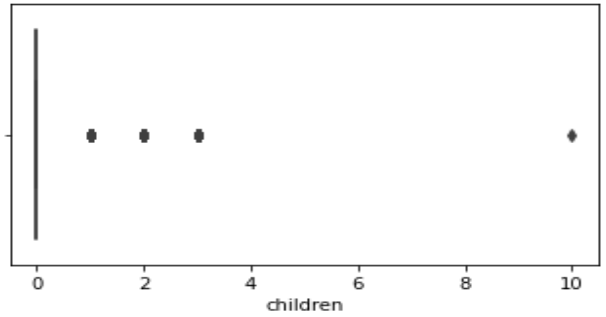
DATA INSPECTION

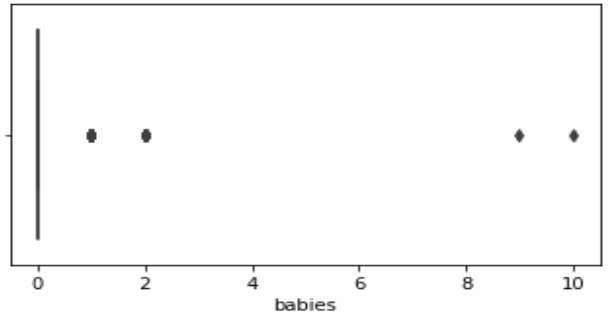
- The Dataset contains 119390 rows and 32 features.
- There are missing values in some columns. Their count and percent in each are as follows:
  - **company – 112593 (94.31%)**
  - **agent – 16340 (13.69%)**
  - **country – 488 (0.41%)**
  - **children – 4 (Negligible)**
- We remove the company column because 94.3% of data was missing.
- If no of children and agent is null we replaced it with 0.
- For the missing values in the country column, we replaced it with mode (value that appears most often).

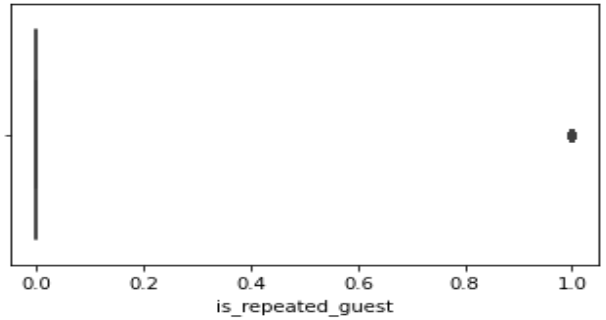
CHECKING OUTLIERS

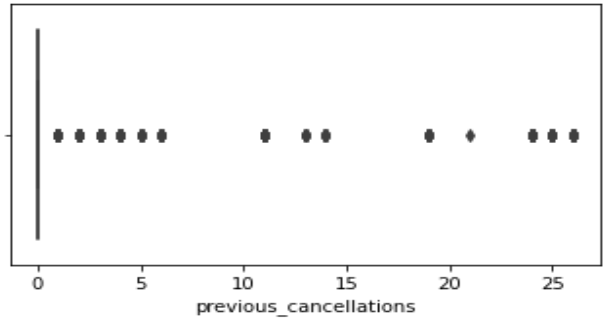


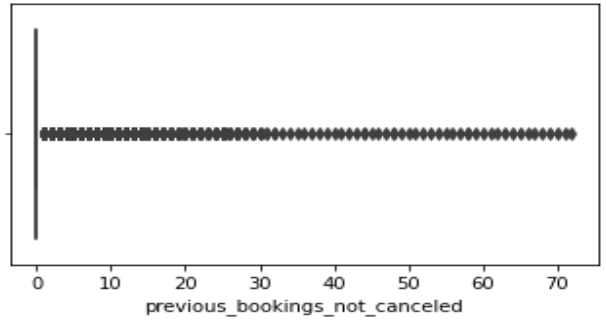
CHECKING OUTLIERS

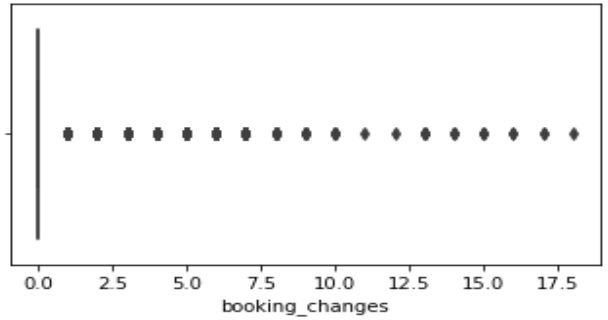


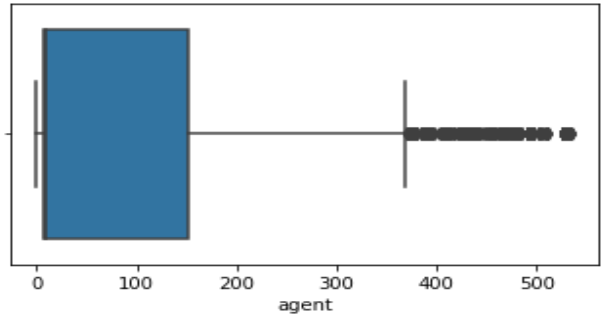


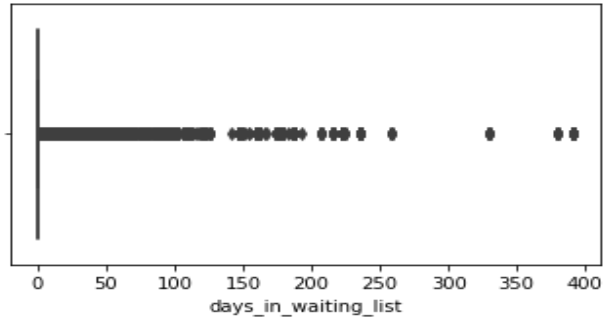


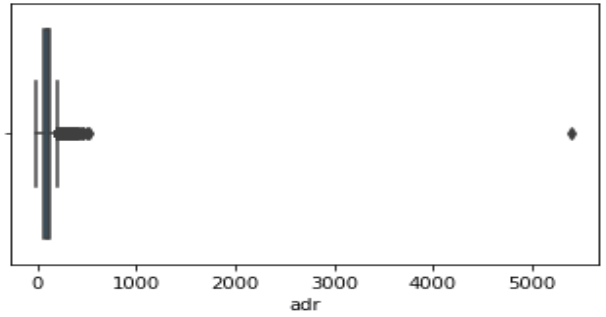


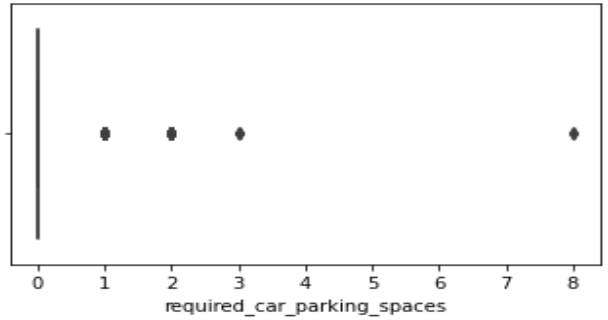


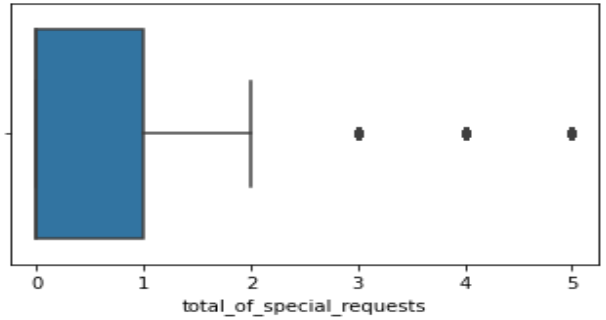






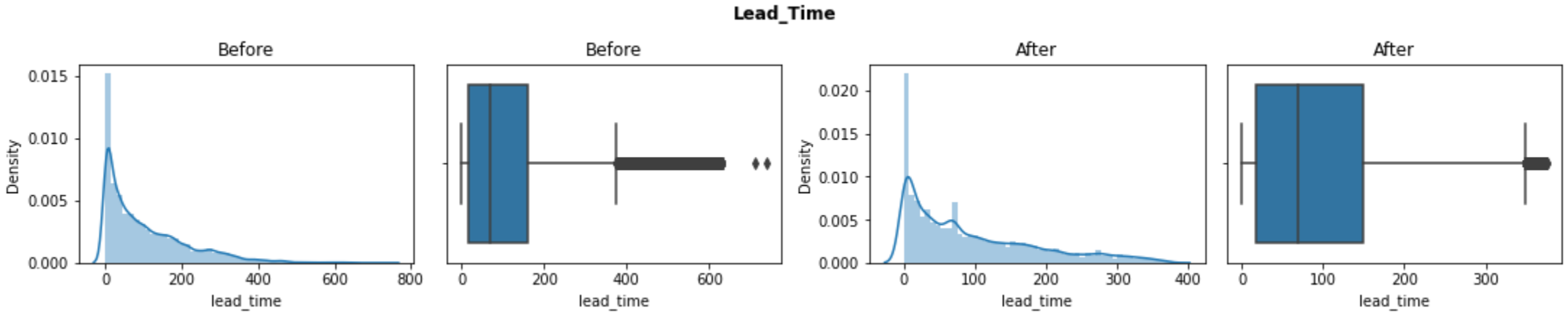






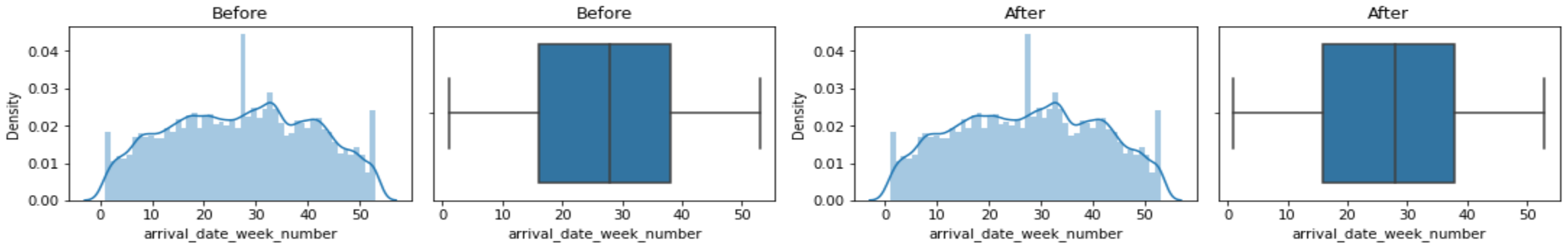
HANDLING OUTLIERS

- Firstly we removed outliers by defining thresholds based on common understanding.
- Then we removed outliers in remaining columns by IQR method and replacing them with the median values.
- We also did percentile capping to remove the outliers. Lets compare the plots before and after the outlier treatment.

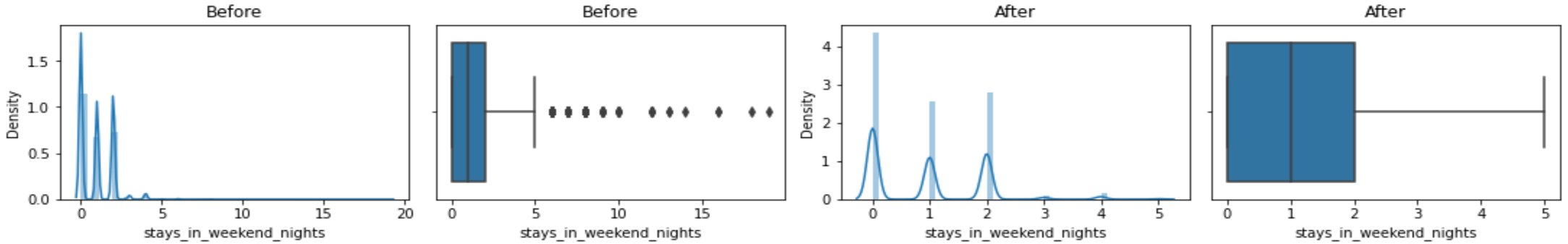


HANDLING OUTLIERS

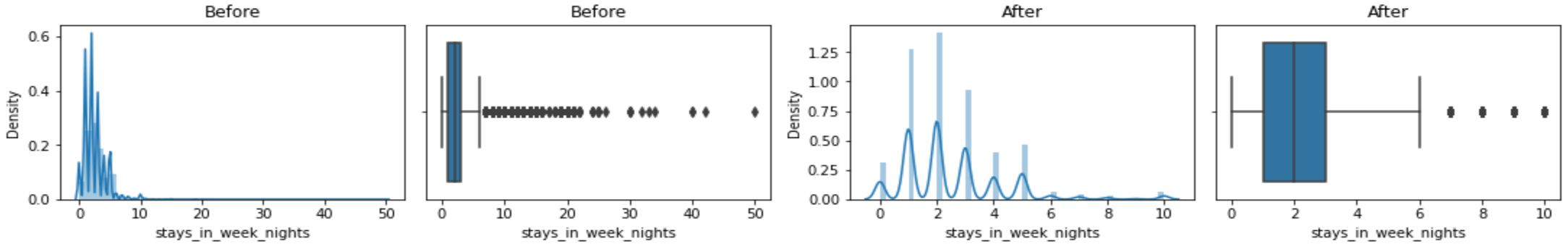
Arrival_Date_Week_Number



Stays_In_Weekend_Nights

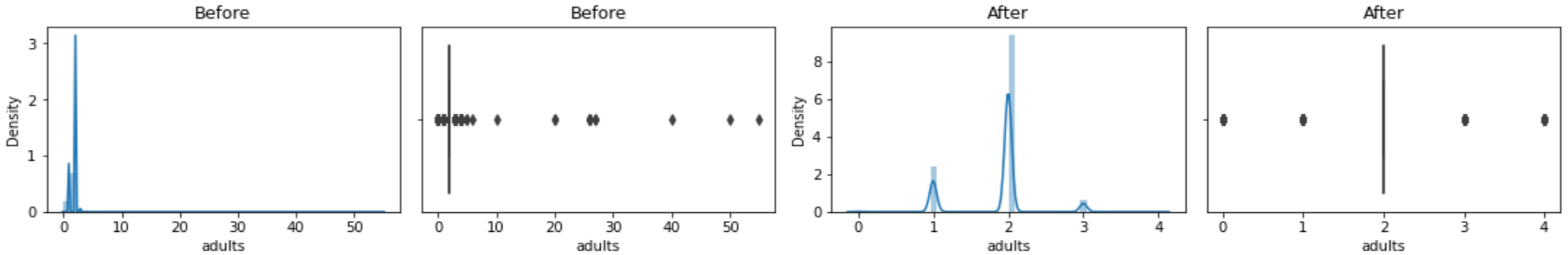


Stays_In_Week_Nights

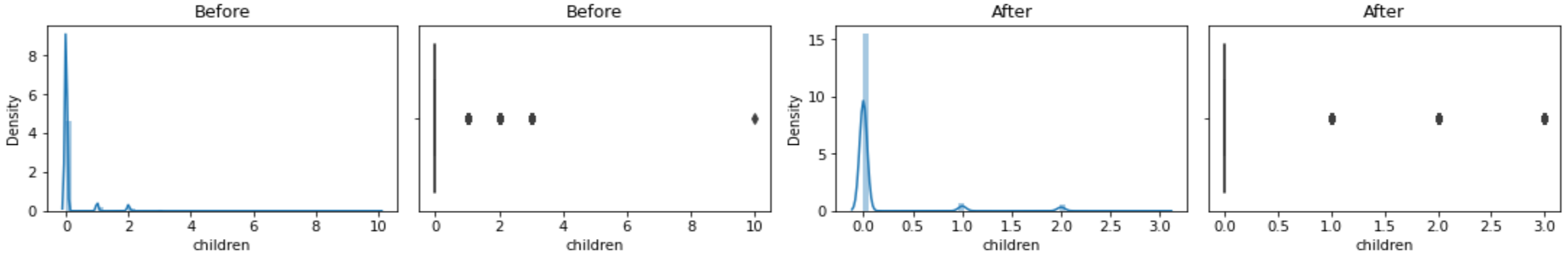


HANDLING OUTLIERS

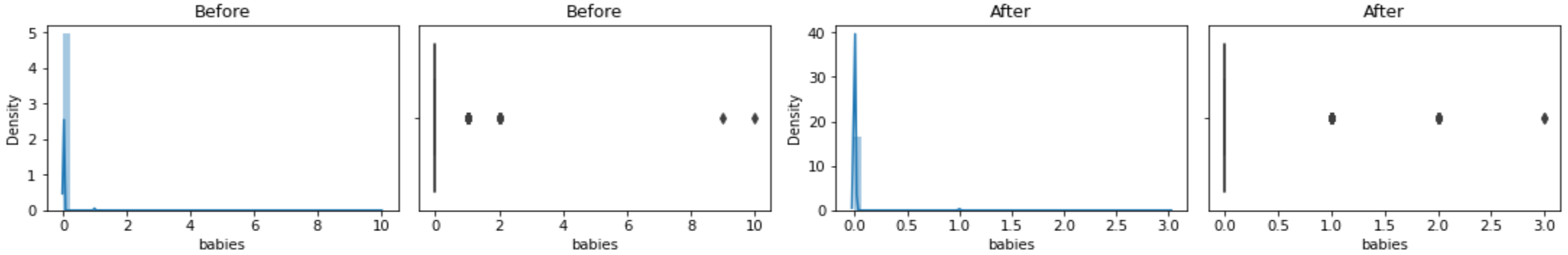
Adults



Children

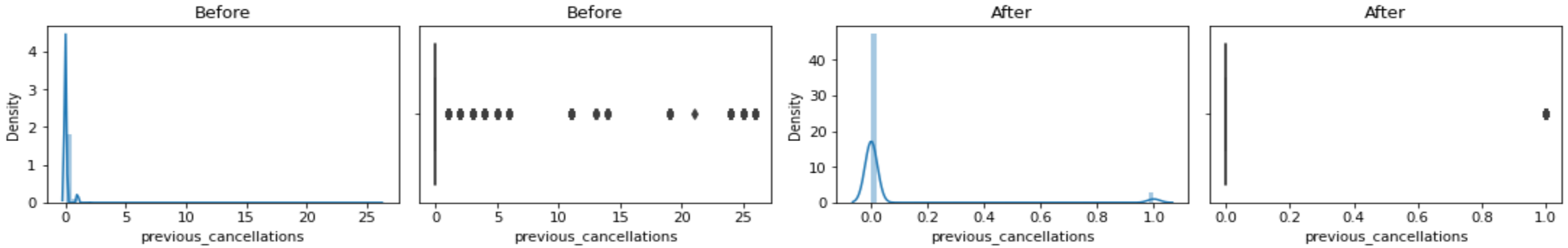


Babies

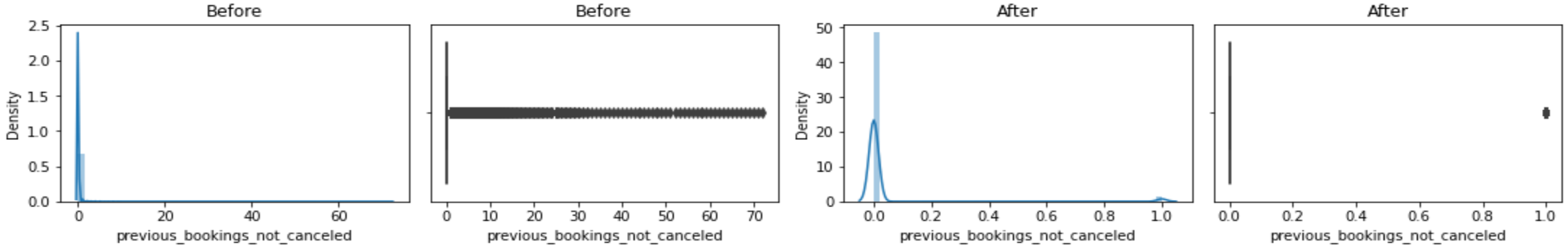


HANDLING OUTLIERS

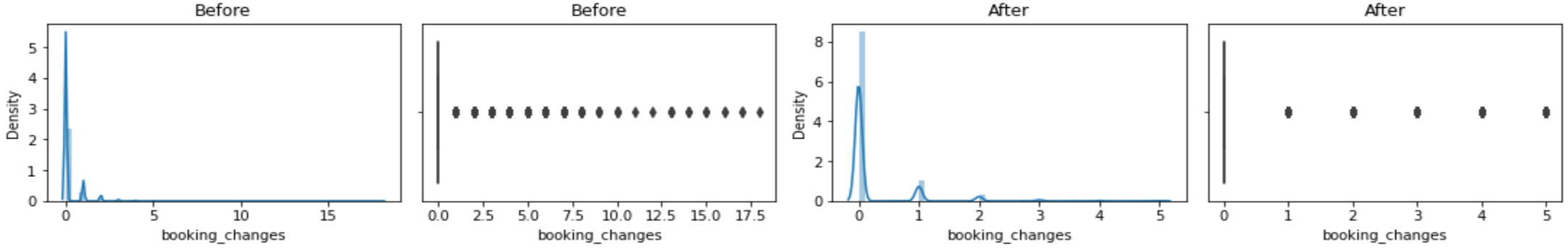
Previous_Cancellations



Previous_Bookings_Not_Canceled

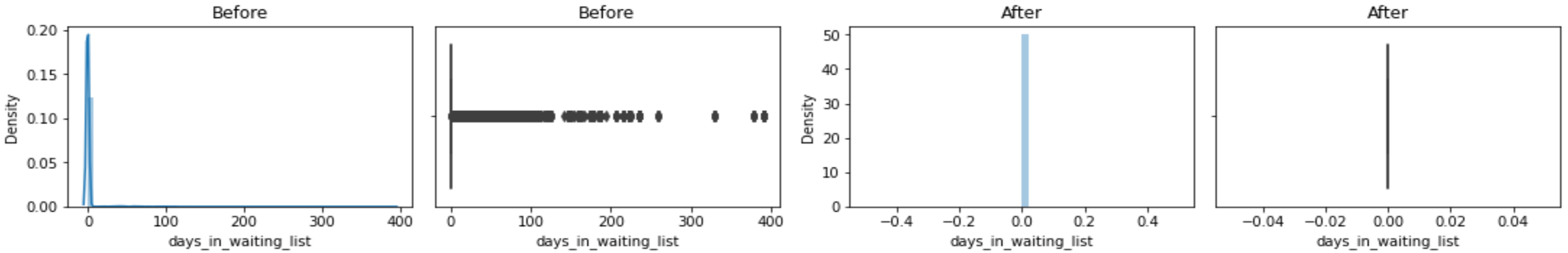


Booking_Changes

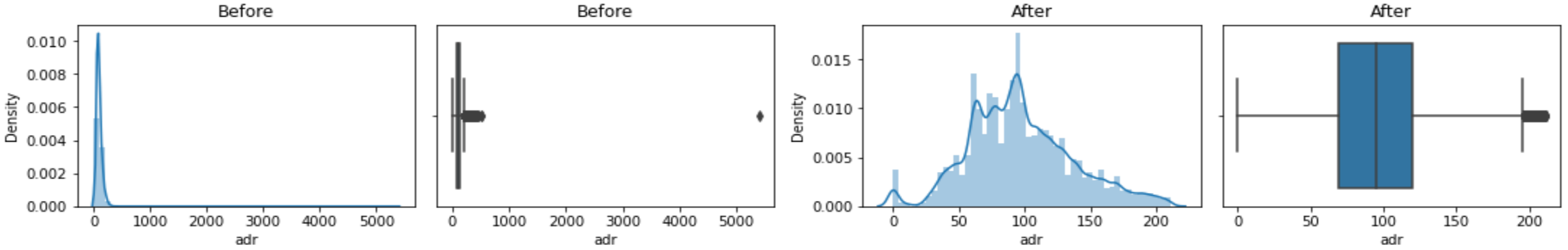


HANDLING OUTLIERS

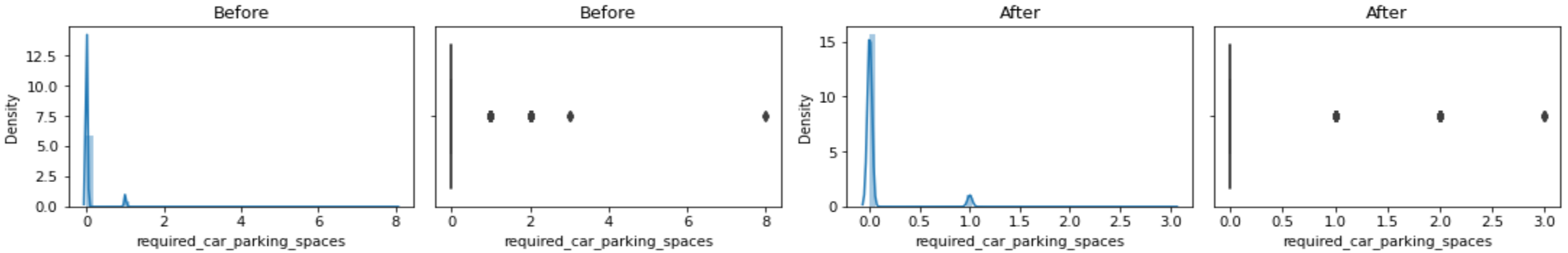
Days_In_Waiting_List



Adr



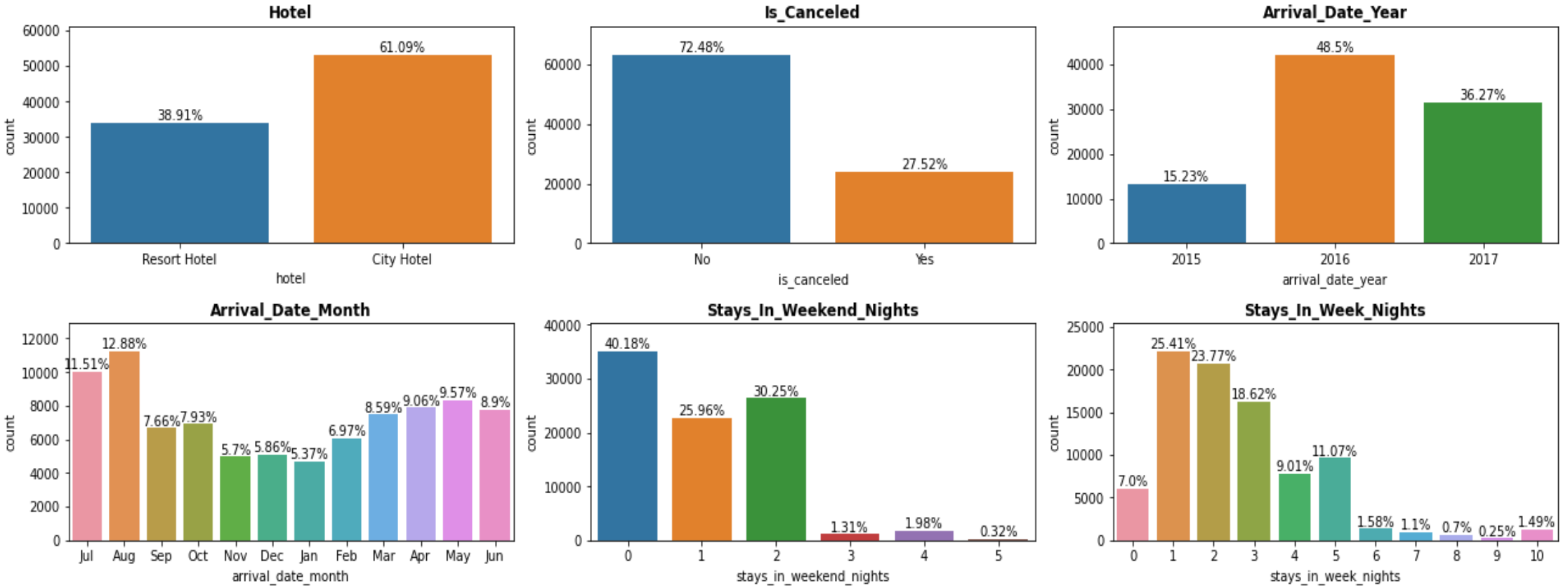
Required_Car_Parking_Spaces



MANIPULATING DATASET

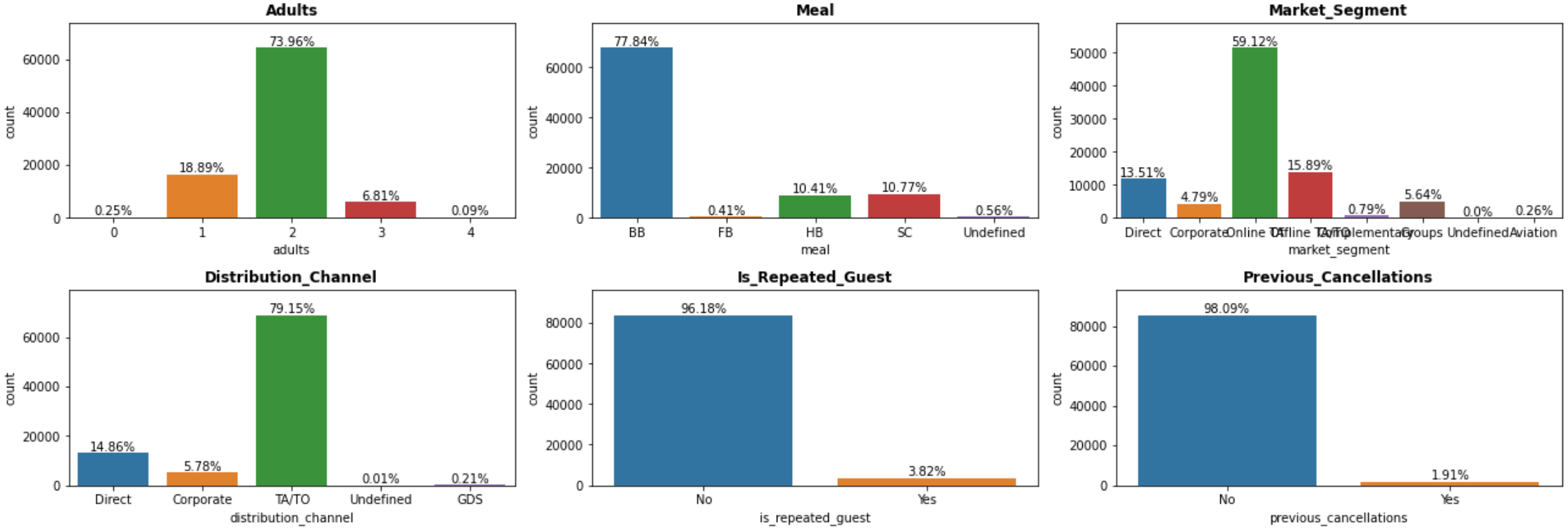
- We found 32052 duplicate rows in the dataset we dropped all such rows and we started doing some feature engineering.
- Created a new feature named **same_room** to indicate whether the customer's preferred room type is allotted or not.
- Created a new feature named **total_children** by adding 'children' and 'babies'.
- Created a new feature named **total_members** by adding 'total_children' and 'adults'.
- Created a new feature named **total_nights** by adding 'stays_in_weekend_nights' and 'stays_in_week_nights' column values.
- Created a new feature named **money_per_person** by multiplying adr with total_nights and dividing the result by total_members.
- Created a new feature named **guest_type** which categorize guests in single, couple or family based on total members.
- Also created a feature named **lead_time_category** which categorize **lead_time** in low, medium or high based on number of days.

UNIVARIATE ANALYSIS



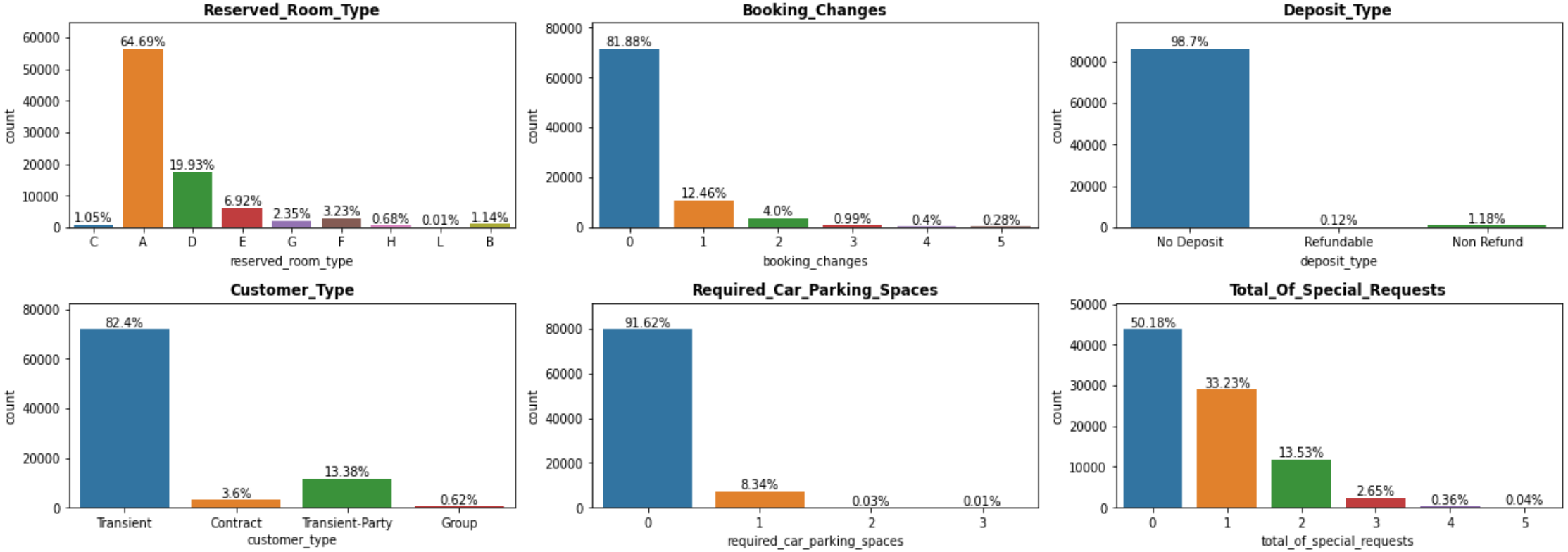
- More than 60% of the bookings are of City Hotel.
- Nearly one third of the bookings are cancelled.
- There was an annual 218.5% rise in hotel bookings in 2016 which dropped down by 25.2% in 2017.
- August is the most preferred month by people for bookings.

UNIVARIATE ANALYSIS



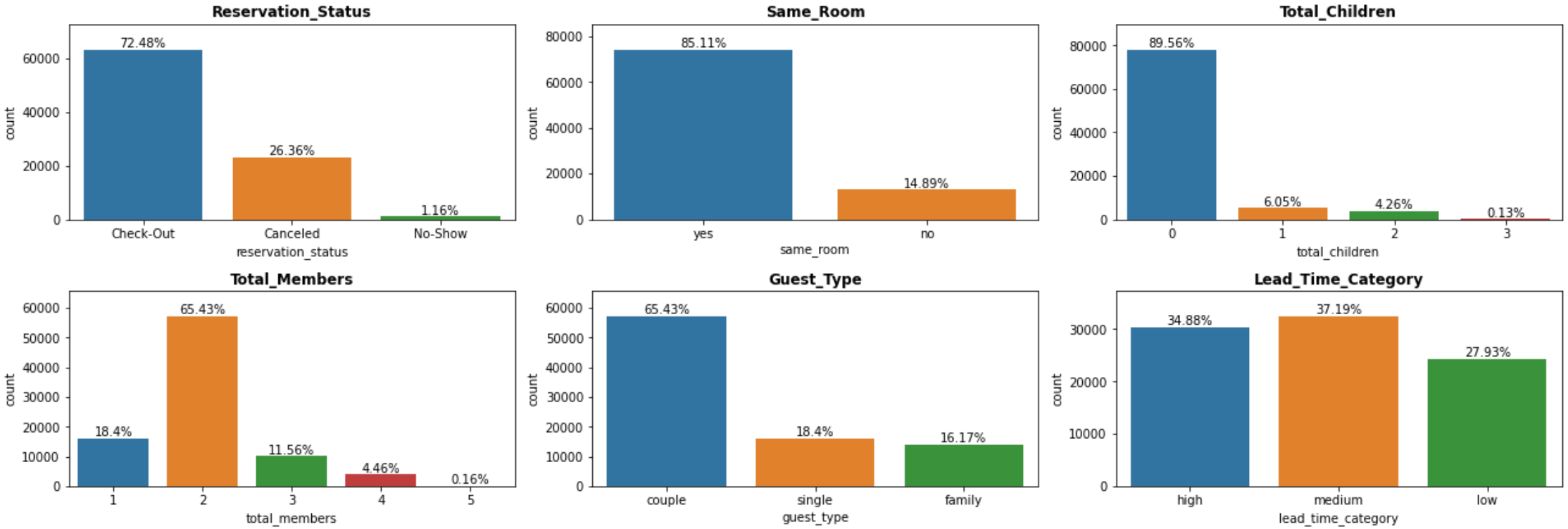
- Most of the bookings are of two adults.
- Most preferred meal is BB(Bread and Breakfast).
- Most of the customers are coming through Online.
- Top distribution channel is TA/TO
- Only 3.82% of the guests are arriving again.
- People who cancel bookings do not really book again.

UNIVARIATE ANALYSIS



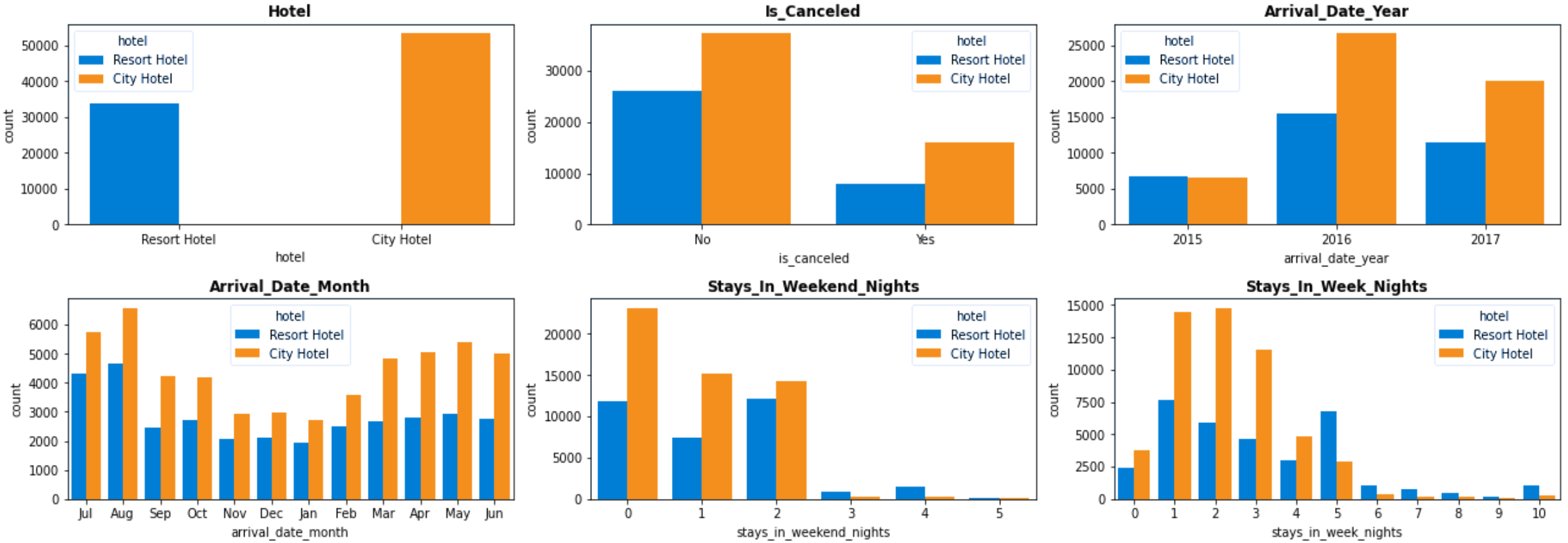
- Room type A is the most preferred one.
- People don’t want to pre-deposit the money.
- Most customers are of type Transient.
- More than 90% of people don’t require any parking space.

UNIVARIATE ANALYSIS



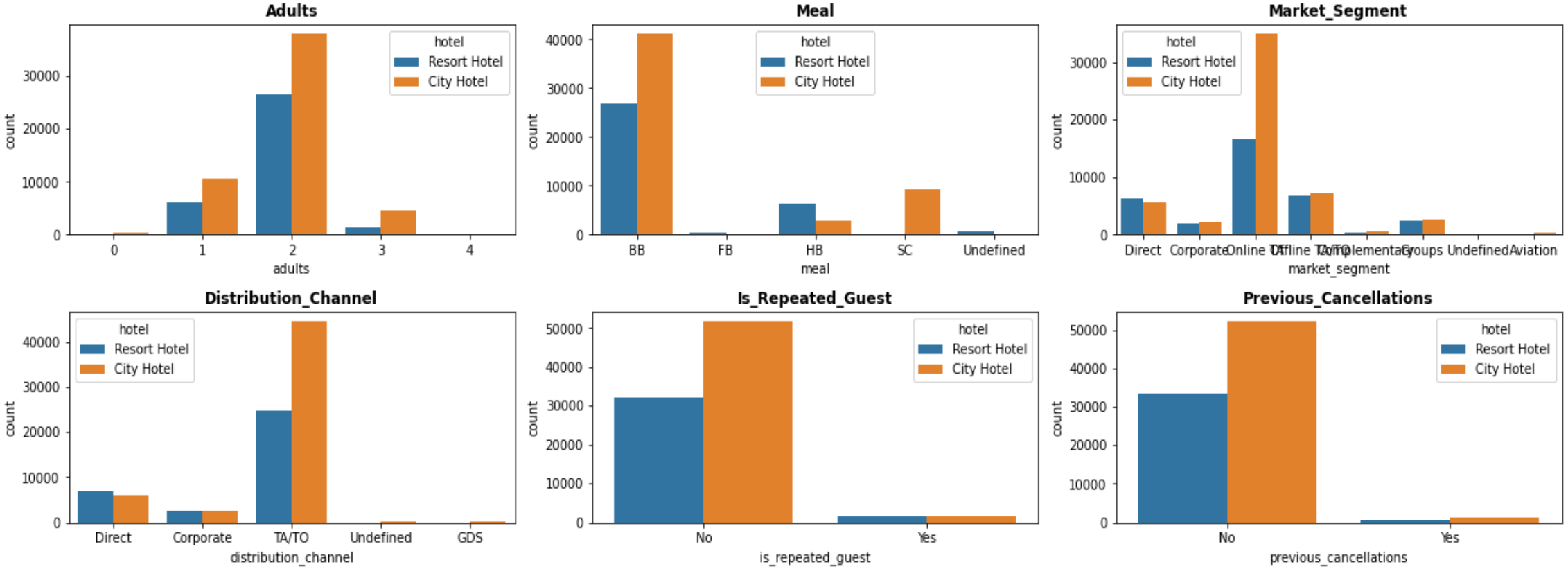
- Around 15% of guests are not assigned with their preferred room.
- Around 10% of the guests arrive with children.
- At least 2 people arrive 80% of the times.
- Around 15% of the people visit with their family. While most of the visitors are couples.

HOTELWISE COMPARISON



- 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.

HOTELWISE COMPARISON

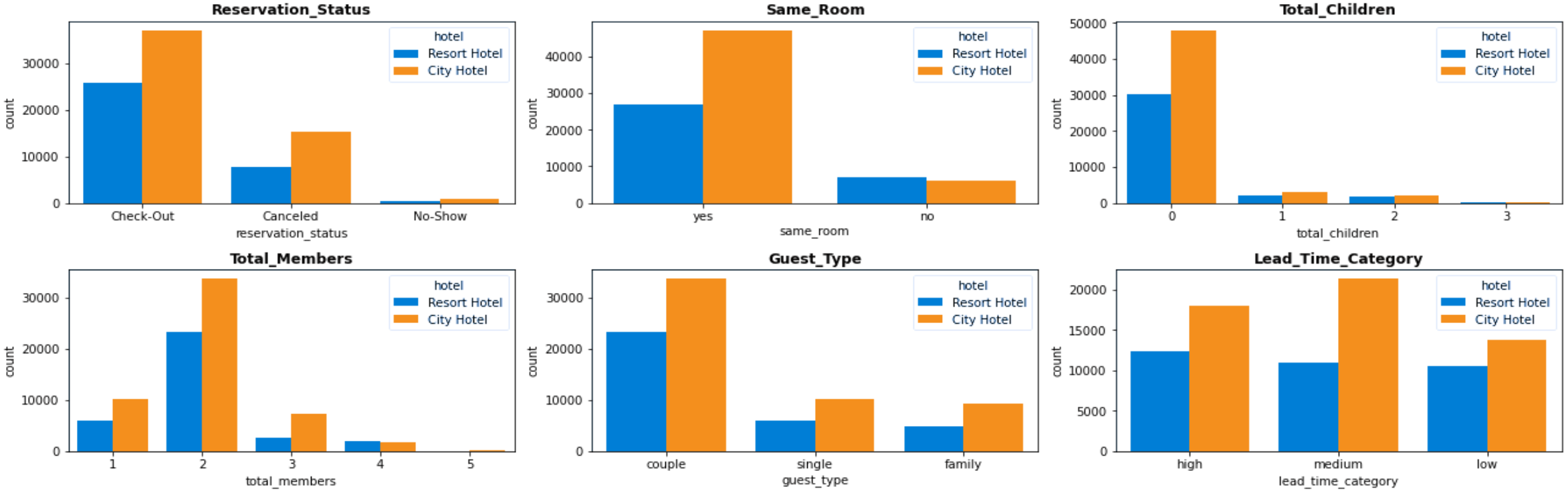


- People who prefer HB(Half board) meal plan (Breakfast and evening meals included) also prefer Resort hotel the most. City hotel is leading in rest all meal plans.
- Resort hotel is leading in Direct marketing but lagging with a huge margin in online segment.

HOTELWISE COMPARISON

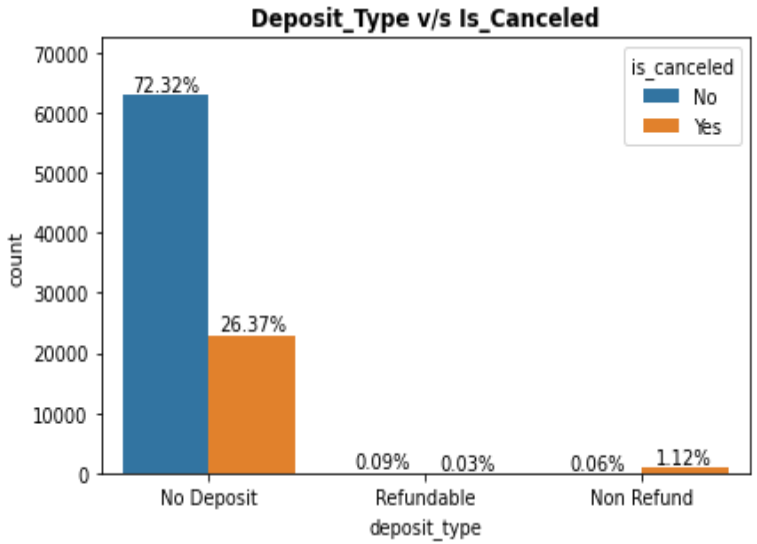
- People who require car parking spaces also prefer Resort hotel. Otherwise City hotel is preferred the most

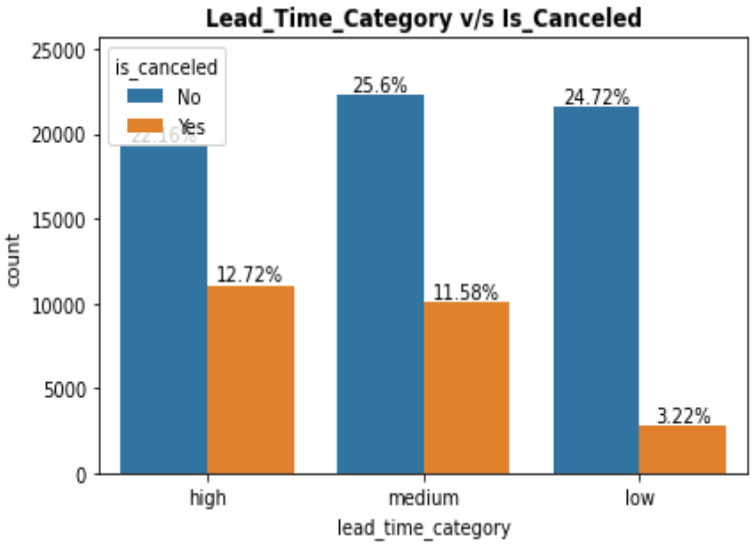
HOTELWISE COMPARISON

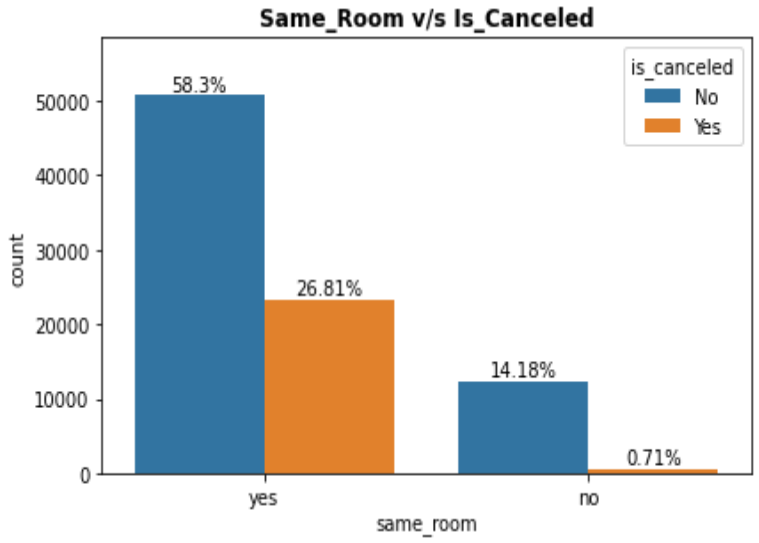


- City hotel assigns the same room as reserved by the guests most of the time. While Resort hotel fails to do so more for more than 25% of guests.

BIVARIATE ANALYSIS



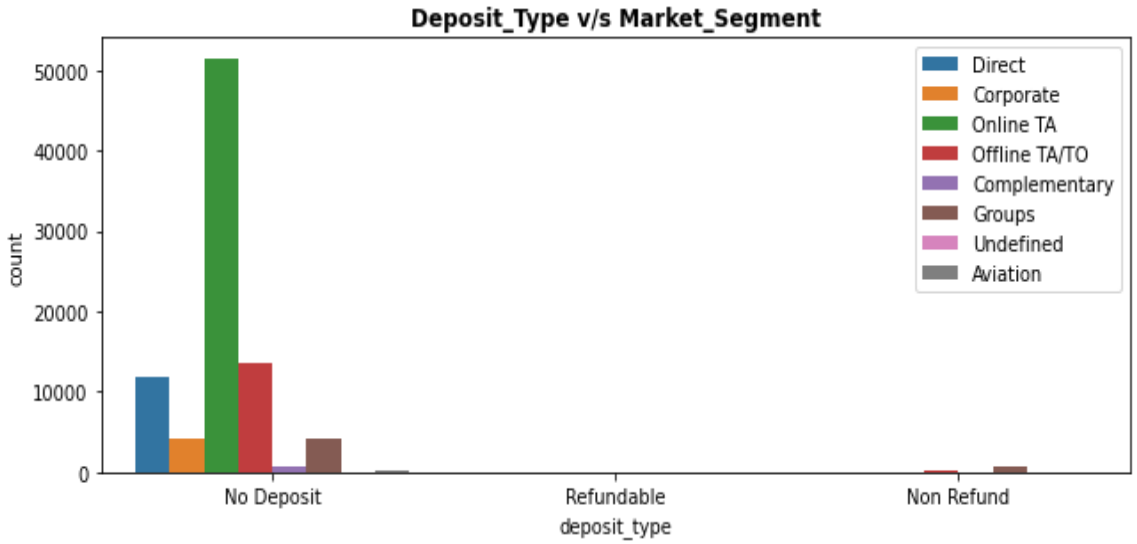


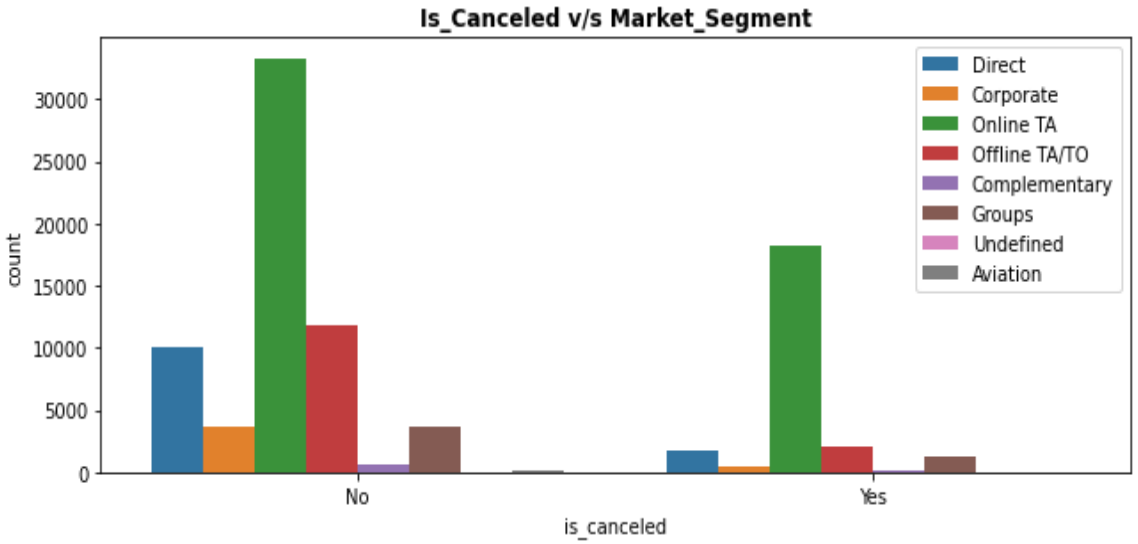


- Most of the bookings are cancelled where there was no deposit made. That's because most of people didn't make deposits. So it's not a reason for cancellation.
- If the lead time is low, less people cancel the bookings. But If the bookings are made more than 15 days in advance, there are comparatively 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.

BIVARIATE ANALYSIS

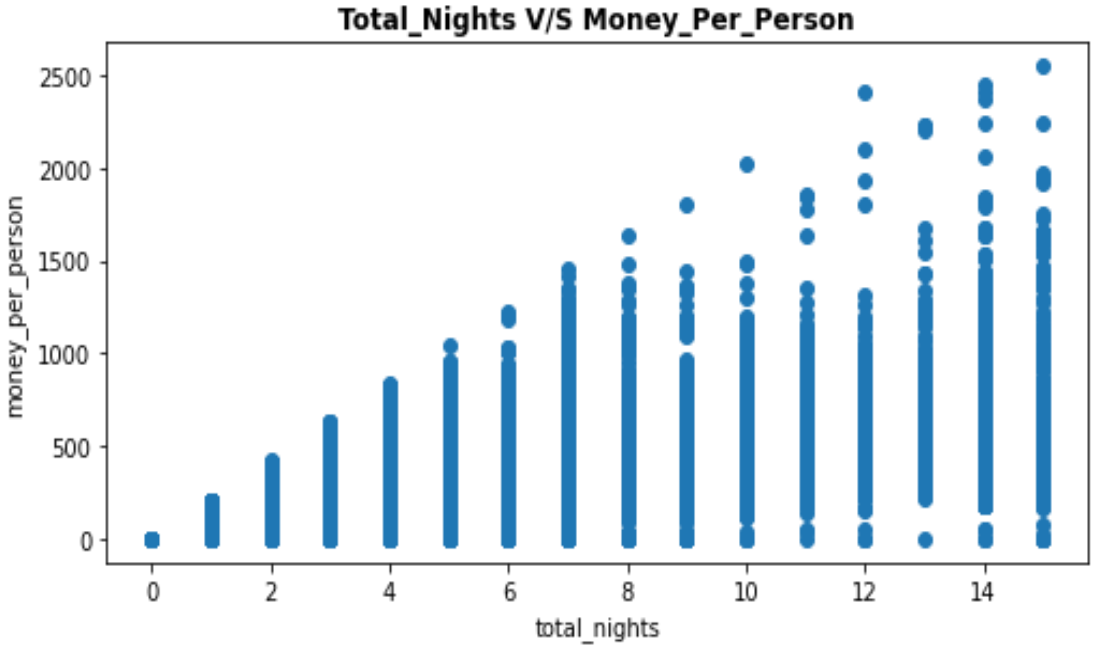
MARKET SEGMENT ANALYSIS

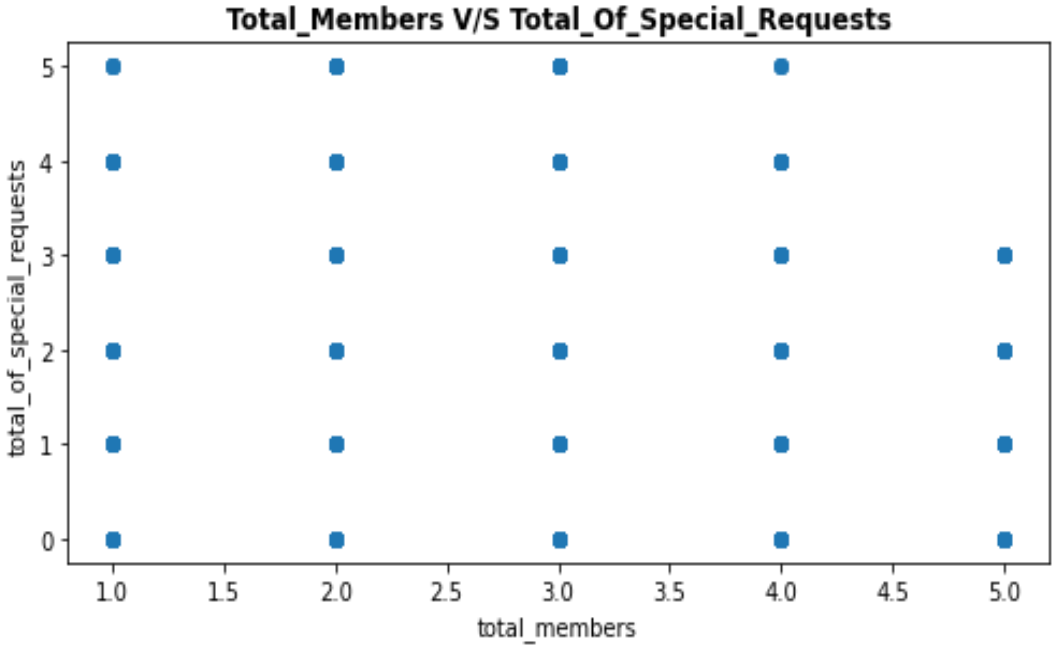




- More than 30% of the online bookings are cancelled. Direct bookings have very less cancellation%.

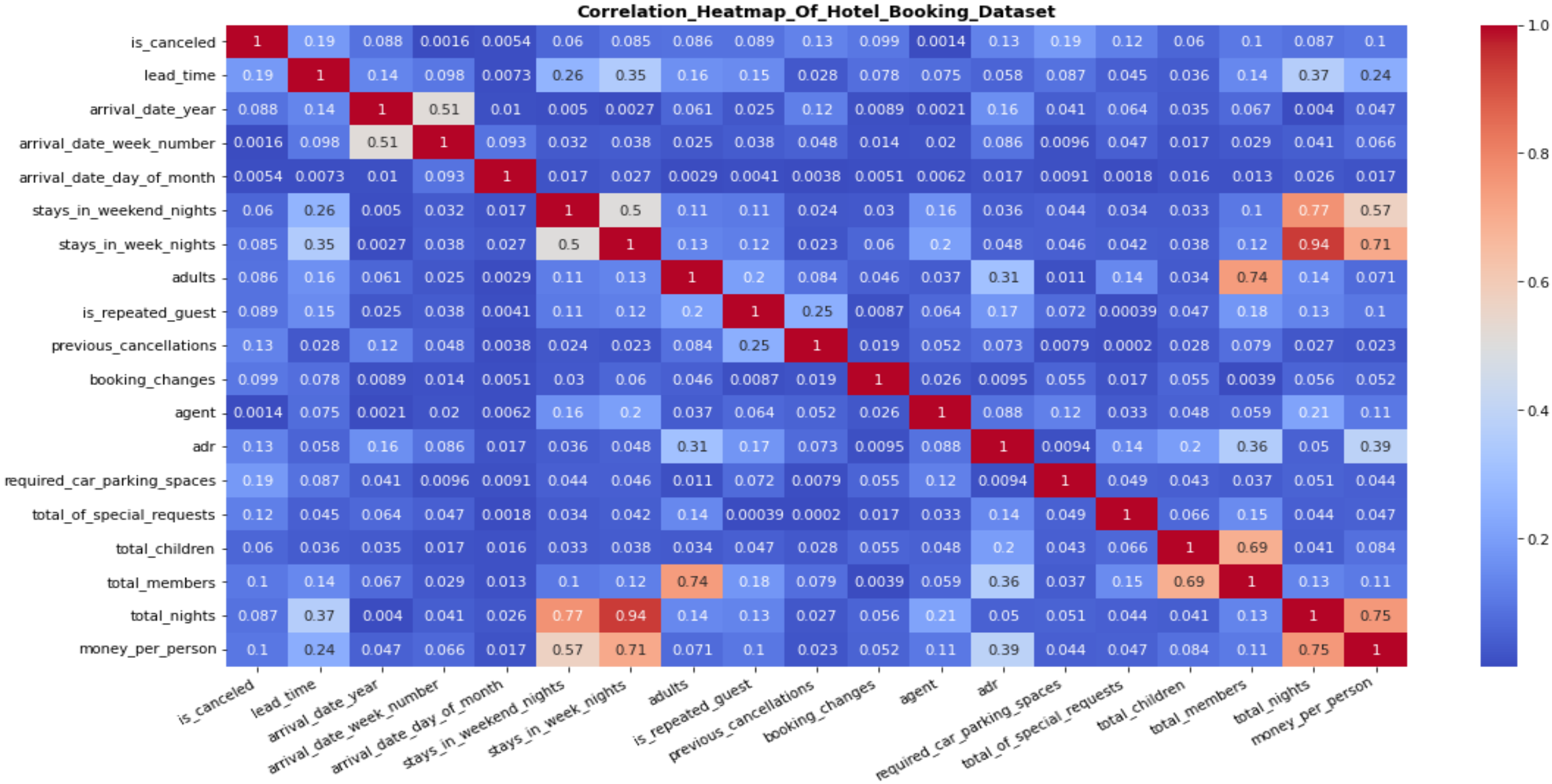
BIVARIATE ANALYSIS





- 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.

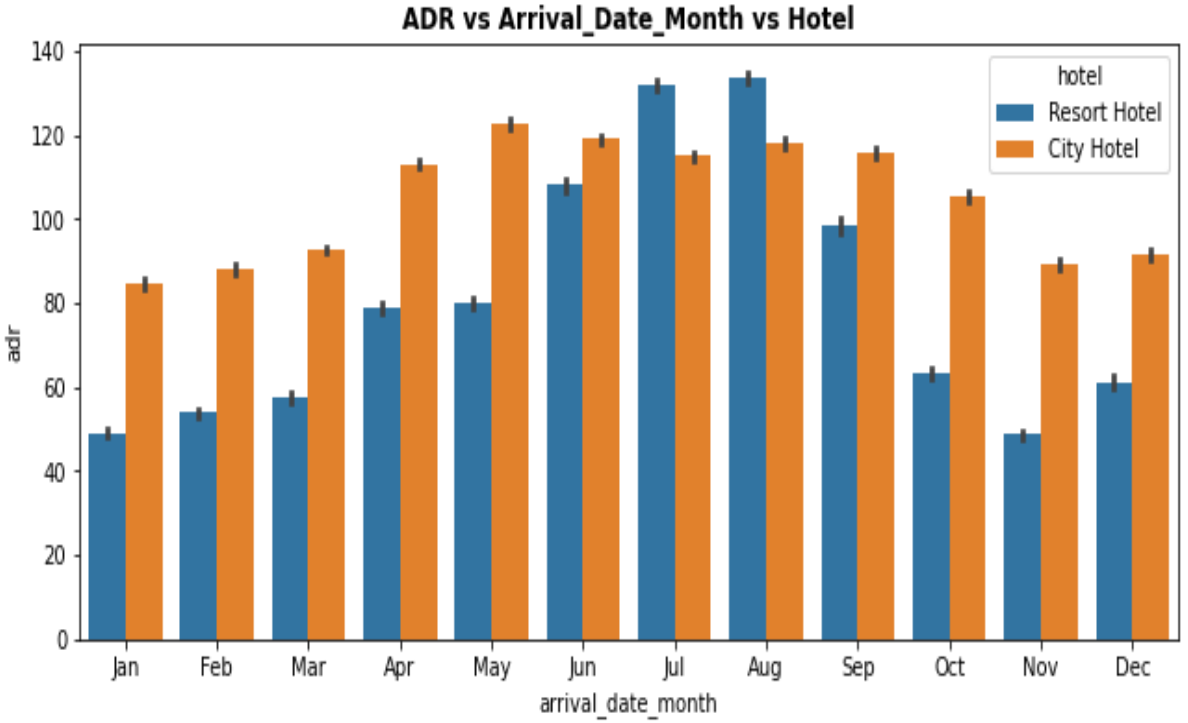
MULTIVARIATE ANALYSIS

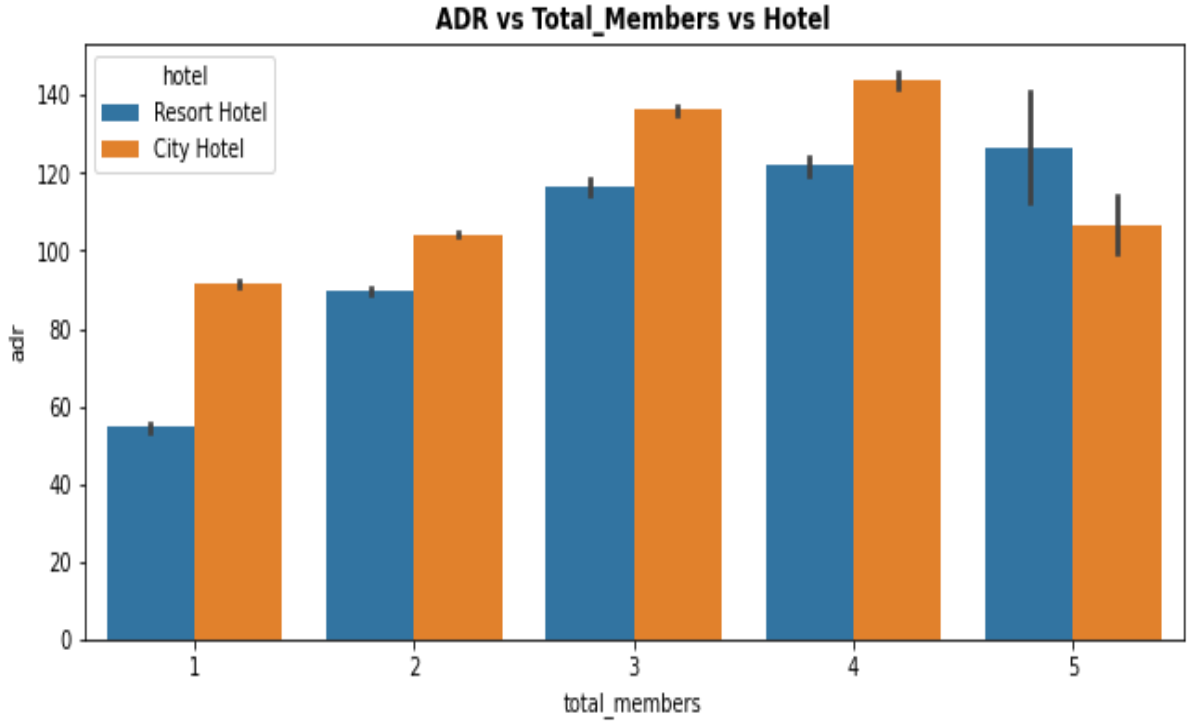


MULTIVARIATE ANALYSIS

- 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.

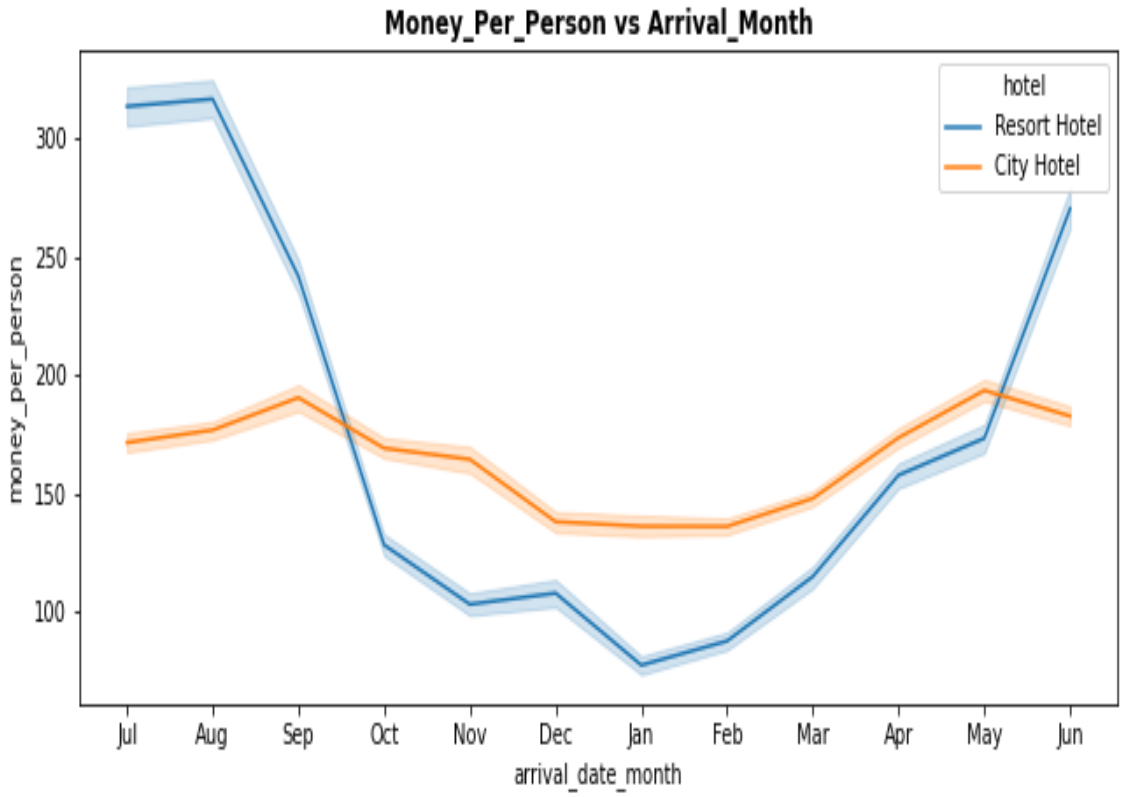
MULTIVARIATE ANALYSIS

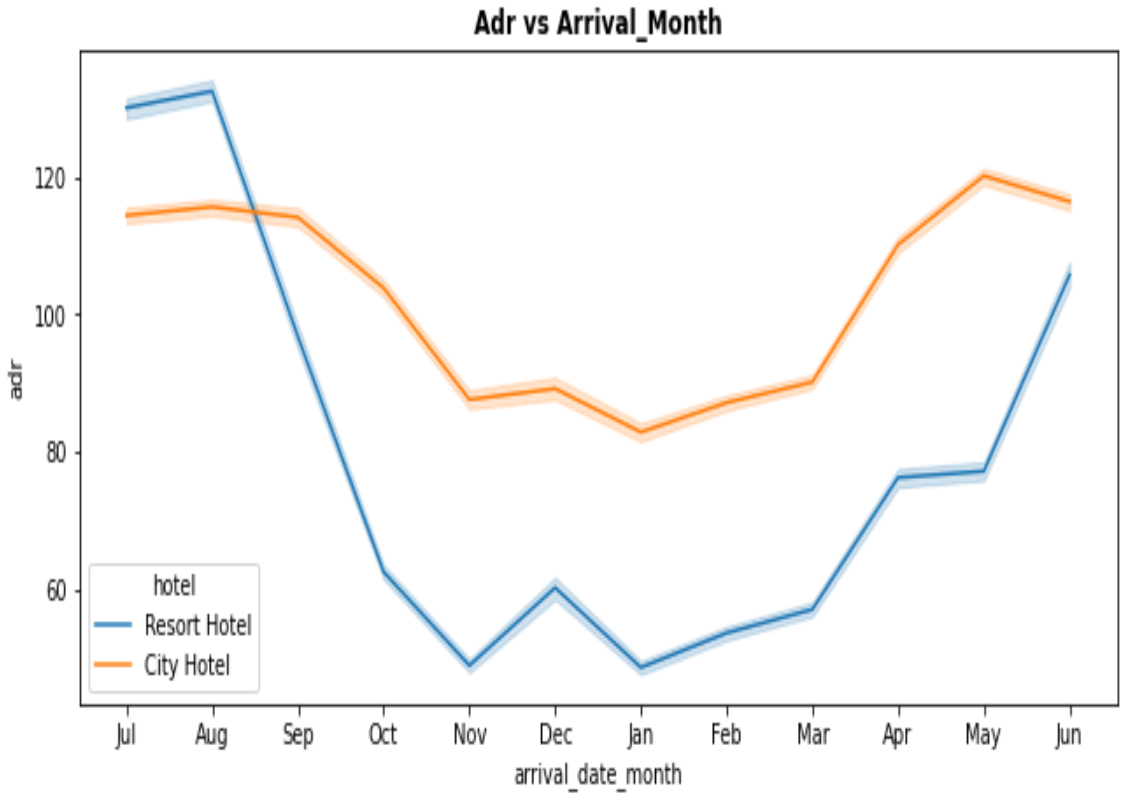




- Average Daily Revenue rises from Jan to Aug then again falls down from Aug to Dec.
- Resort hotel leads in ADR only during the peak months July and August. Rest all months City hotel has high ADR.
- April to September is the semester with high ADR for both hotels
- ADR for Resort hotel is directly 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 number of bookings and country where the highest number of guests arrive from are as follows:

AGENT	NO. OF BOOKINGS
9	17193
0	10566
240	8074
7	2858
14	2759

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 space is 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%.

CONCLUSION

- Best time to book a hotel is in January.
- Average days in advance booking : 77 days
- 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.