Capstone Project - 1

EDA on Hotel Booking Analysis

BY

Akshay Dhakate

(Cohort – Azaadi)

Problem Statement

- For this project we will be analyzing Hotel Booking data. 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.

- Hotel industry is a very volatile industry and the bookings depends on above factors and many more.

- The main objective behind this project is to explore and analyze data to discover important factors that govern the bookings and give insights to hotel management ,which can perform various campaigns to boost the business and performance.

Work Flow

I am dividing this work flow into following 3 steps.

1. Data Collection and Understanding

2. Data Cleaning and Manipulation

3. Exploratory Data Analysis(EDA)...

EDA will be divided into following 3 analysis.

1)Univariate analysis: Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable.

2)Bivariate analysis: Bivariate analysis is where you are comparing two variables to study their relationships.

3)Multivariate anlysis: Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables.

Data Collection and Understanding:

- After collecting data it’s very important to understand your data. So we had hotel Booking analysis data. Which had 119390 rows and 32 columns. So let’s understand this 32 columns.

Data Description:

hotel :Resort Hotel or City Hotel
is_canceled : Value indicating if the booking was canceled (1) or not (0)
lead_time : Number of days that elapsed between the entering date of the booking and the arrival date
arrival_date_year : Year of arrival date
arrival_date_month : Month of arrival date
arrival_date_week_number : Week number of year for arrival date
arrival_date_day_of_month : Day of arrival date
stays_in_weekend_nights : Number of weekend nights
stays_in_week_nights : Number of week nights.
adults : Number of adults
children : Number of children
babies : Number of babies
meal : Type of meal booked.
country : Country of origin.

market_segment : Market segment designation. (TA/TO)
distribution_channel : Booking distribution channel.(T/A/TO)
is_repeated_guest : is a repeated guest (1) or not (0)
previous_cancellations : Number of previous bookings that were cancelled by the customer prior to the current booking
previous_bookings_not_canceled : Number of previous bookings not cancelled by the customer prior to the current booking
reserved_room_type : Code of room type reserved.
assigned_room_type : Code for the type of room assigned to the booking.
booking_changes : Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
deposit_type : No Deposit, Non Refund , Refundable.
agent : ID of the travel agency that made the booking
company : ID of the company/entity that made 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 customer. Contract,Group,transient,Transient party.
adr : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
required_car_parking_spaces : Number of car parking spaces required by the customer
total_of_special_requests : Number of special requests made by the customer (e.g. twin bed or high floor)
reservation_status : Reservation last status.

Data Cleaning and Manipulation:

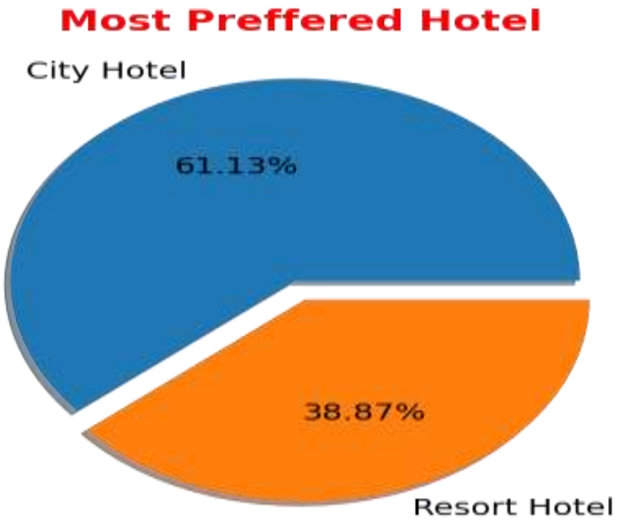
1. company, agent, country and children columns with missing values. I replaced missing values as per requirement.
2. Dropping company column because more then 90% data is missing
3. Data had 31994 duplicates values. So I dropped it from the data.

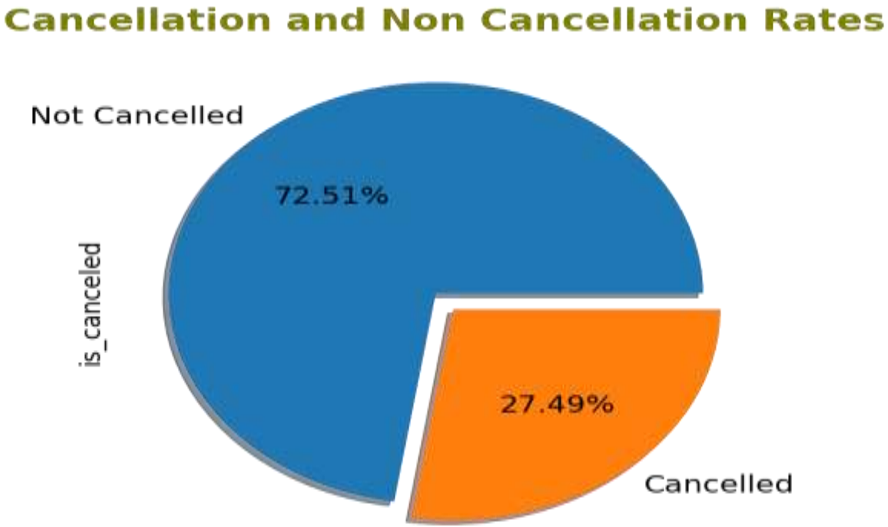
4. I created 2 new columns

- A)‘total_People’ = from the Children, adults, babies.
- B) ‘total_stay ’ = From weekend nights and weekdays night.

Exploratory Data Analysis (EDA) :

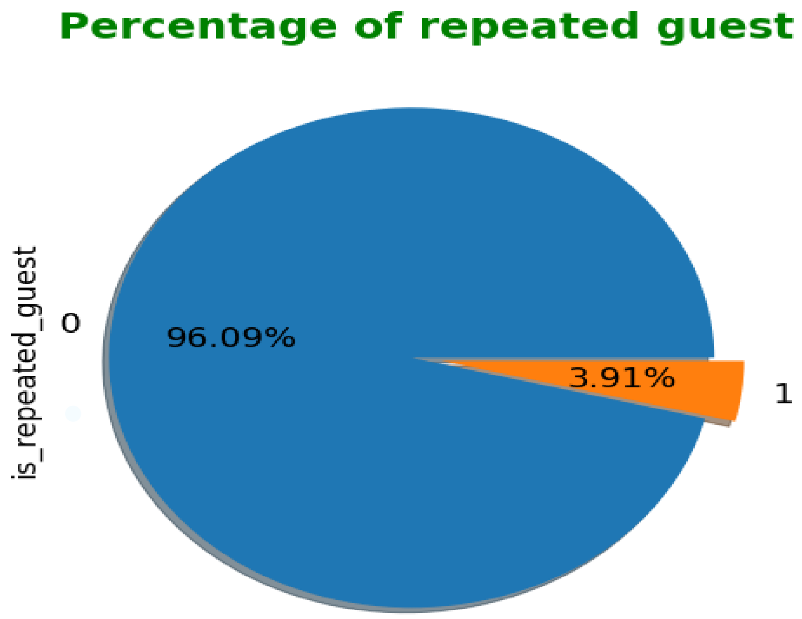
Univariate Analysis



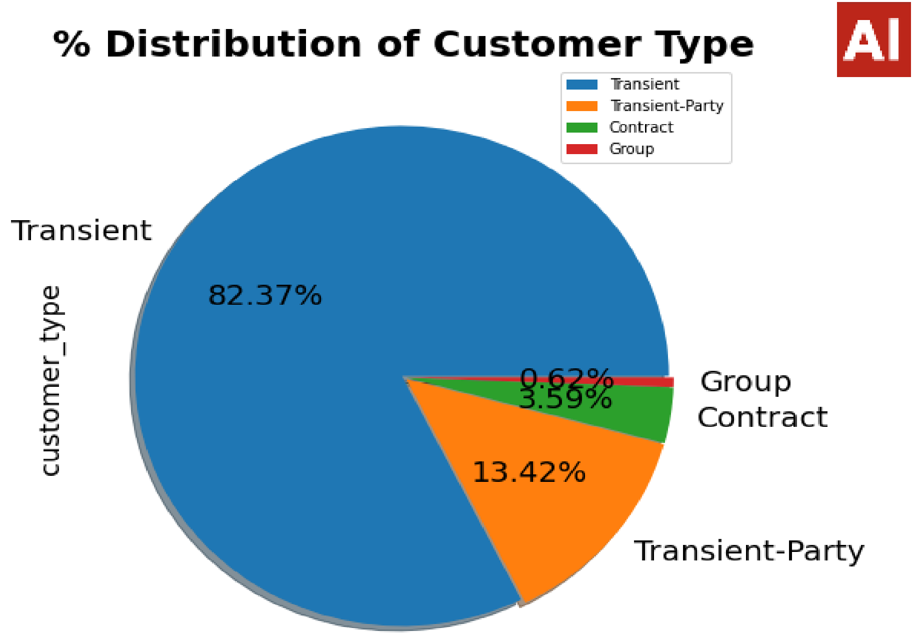


Conclusions:

- City hotels is the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
- 27.49 % bookings were got cancelled out of all the bookings



Percentage of repeated guest

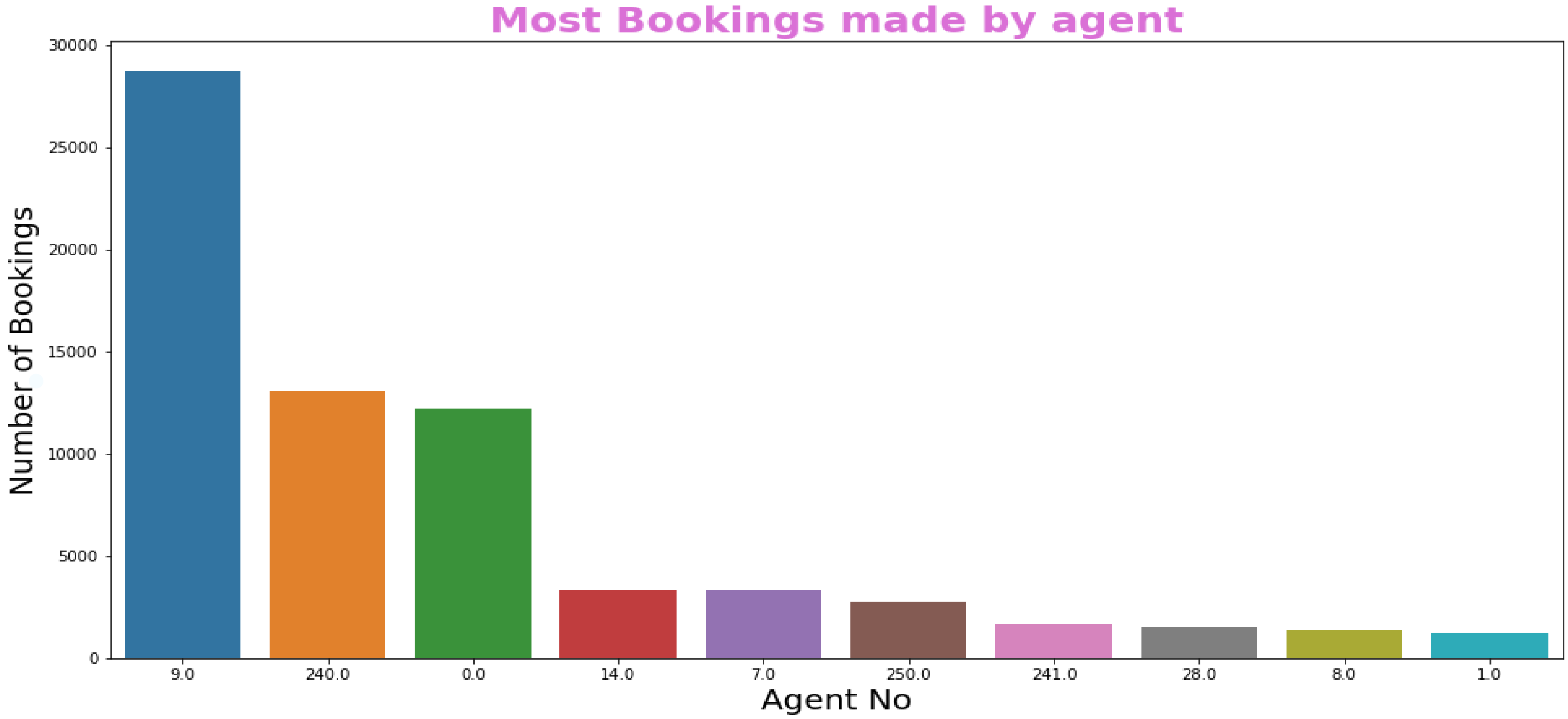


% Distribution of Customer Type

Conclusions:

➢.Only 3.9 % people were revisited the hotels. Rest 96.09 % were new guests. Thus retention rate is low.

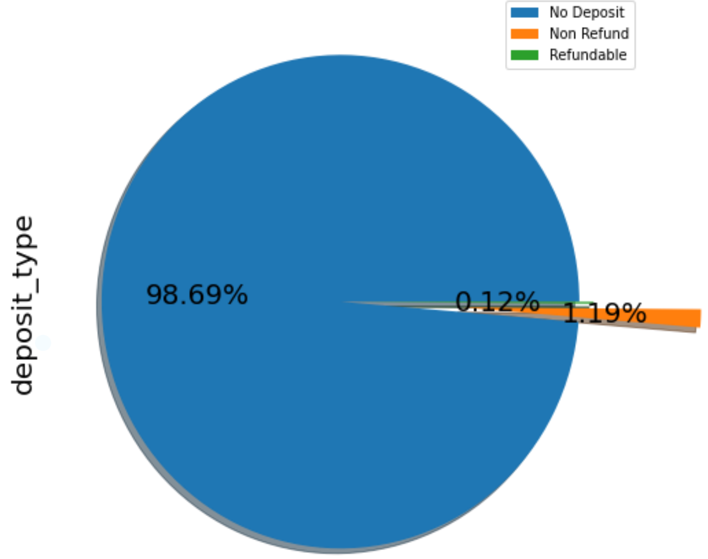
➢Most of the customers/guests were Transient type(82.37%). And transient party were 13.42% and 0.62% belongs to group. Remaining guests belongs to Contract type.



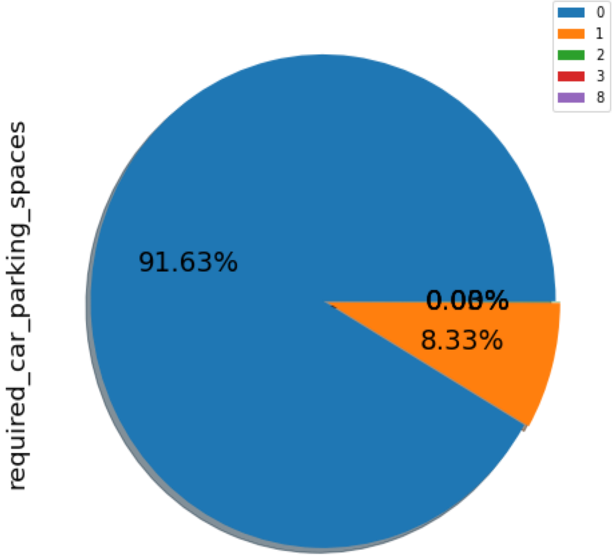
Conclusions:

➢. Agent ID no: 9.0 made most of the bookings

% Distribution of deposit type



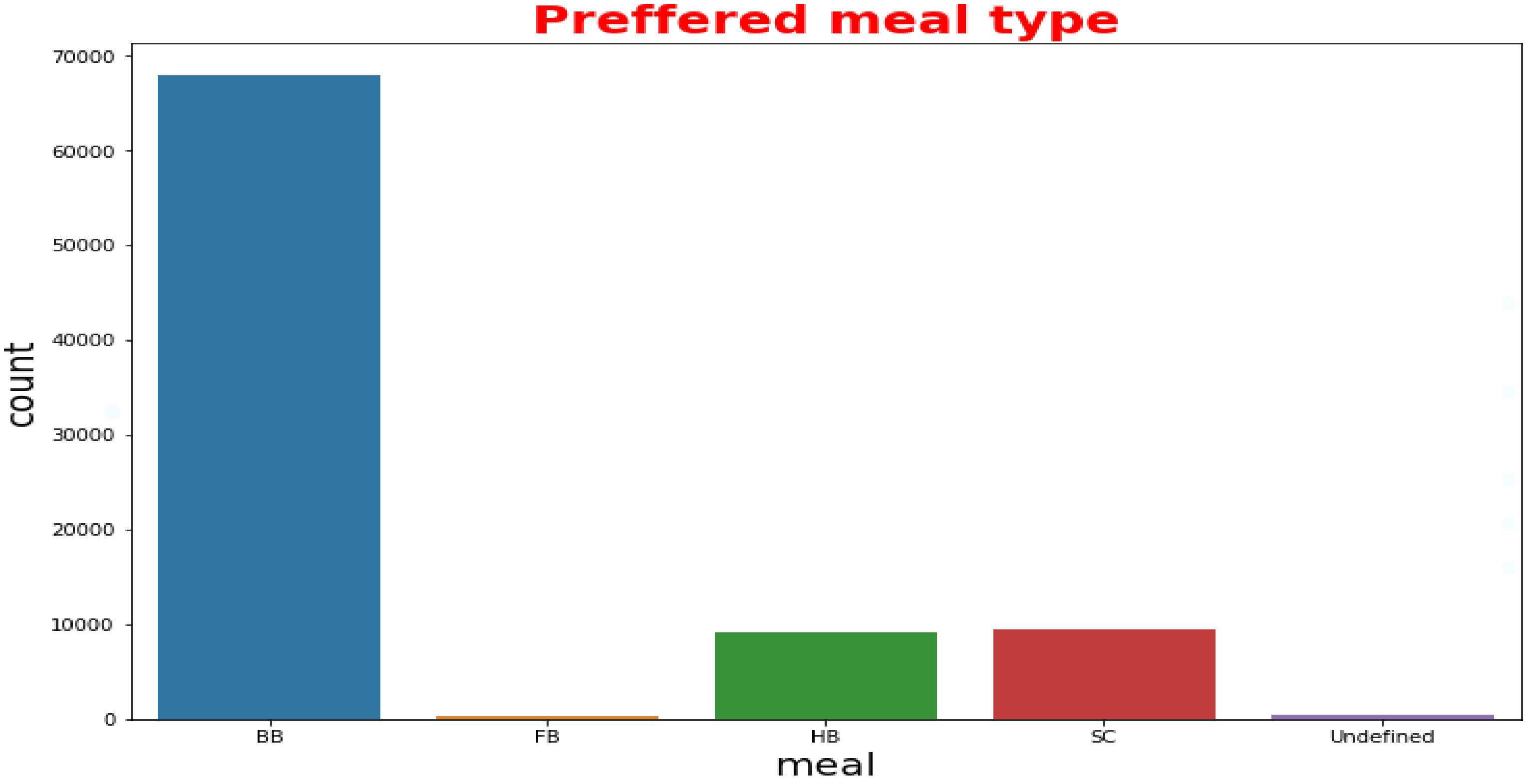
% Distribution of required car parking spaces



Conclusions:

➢. 98.69 % of the guests prefer "No deposit" type of deposit.

➢91.63 % guests did not required the parking space. only 8.33 % guests required only 1 parking space.



Types of meal in hotels:

BB - (Bed and Breakfast)

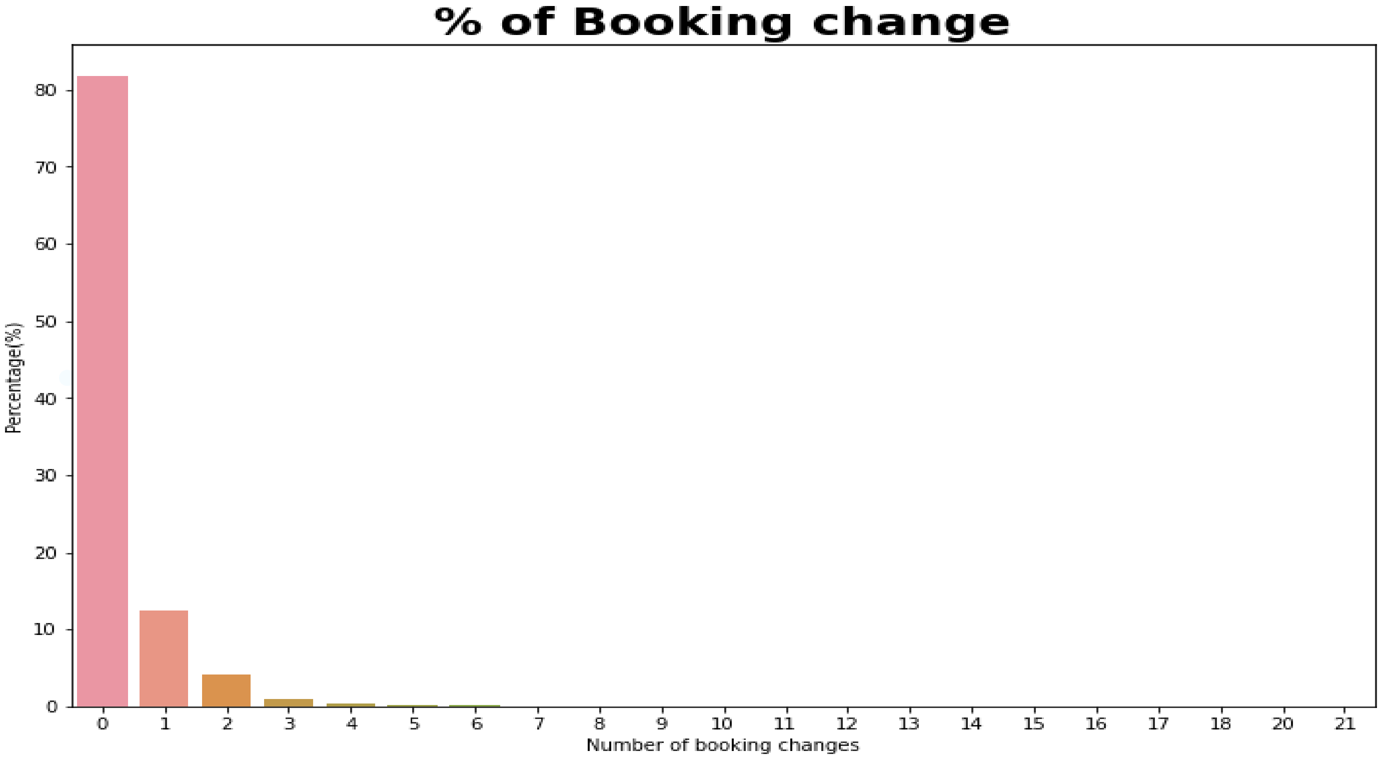
HB- (Half Board)

FB- (Full Board)

SC- (Self Catering)

Conclusions:

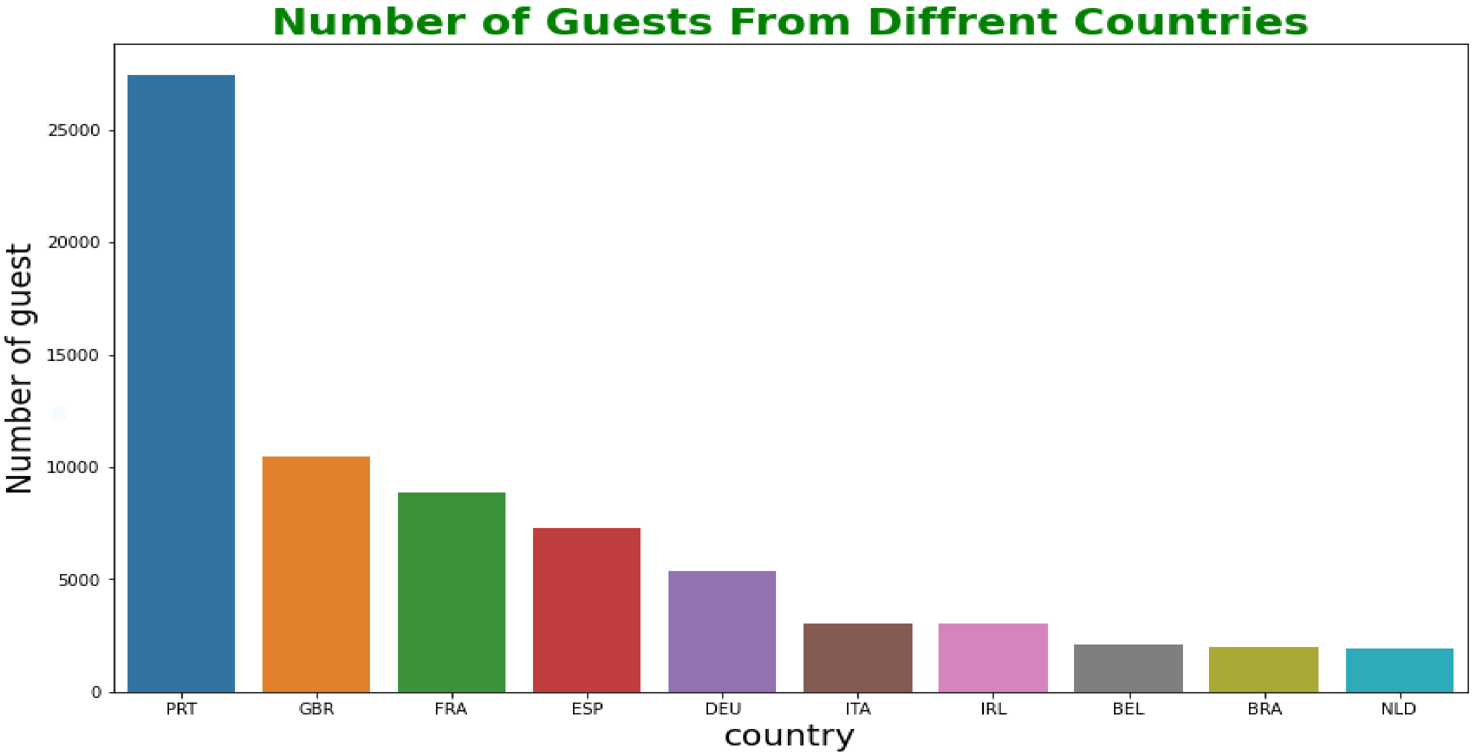
- most preferred meal type by the guests is BB(Bed and Breakfast) HB- (Half Board) and SC- (Self Catering) are equally preferred.



0 = 0 changes made in the bookings
1 = 1 changes made in the bookings
2 = 2 changes made in the bookings
3 = 3 changes made in the bookings

Conclusions:

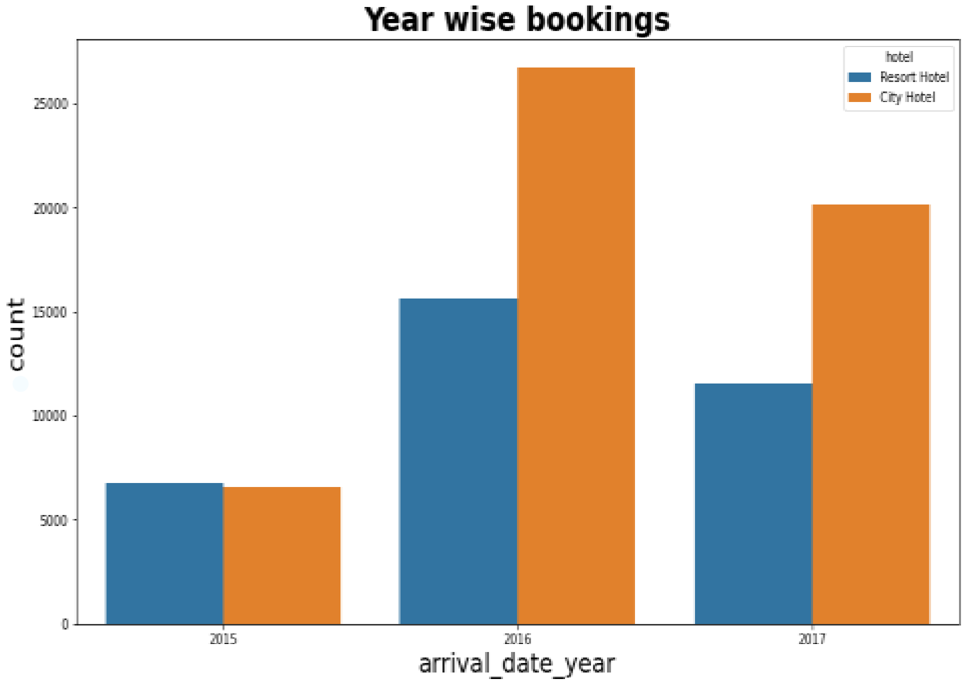
- . 80% -83% of the bookings were not changed by the customer.



PRT- Portugal
GBR- United Kingdom
FRA- France
ESP- Spain
DEU - Germany
ITA -Italy
IRL - Ireland
BEL -Belgium
BRA -Brazil
NLD-Netherlands

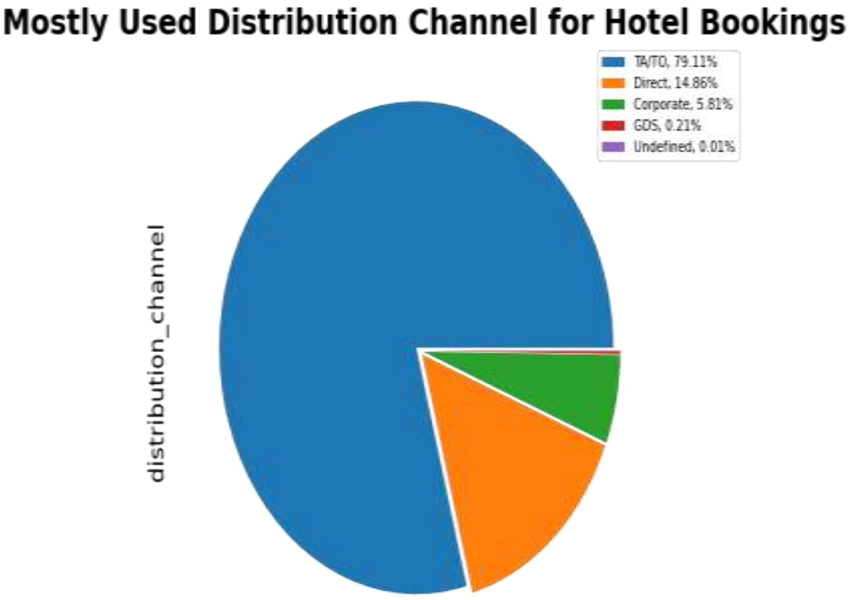
Conclusions:

➢ Most of the guests are coming from Portugal i.e. more 25000 guests are from Portugal



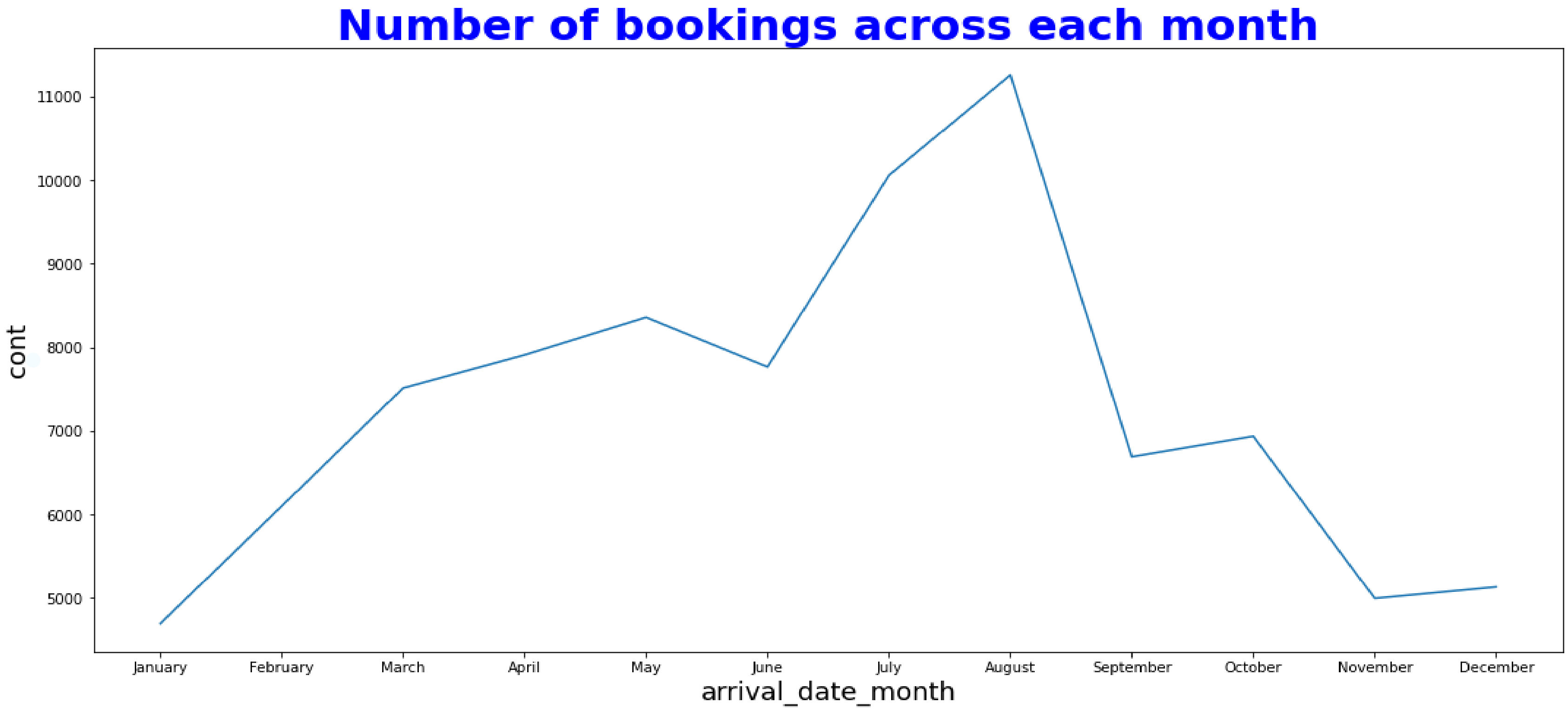
Conclusions:

- ➢ 2016 had the highest bookings.
- ➢ 2015 had less bookings.
- ➢ City hotels had the most of the bookings.



Conclusions:

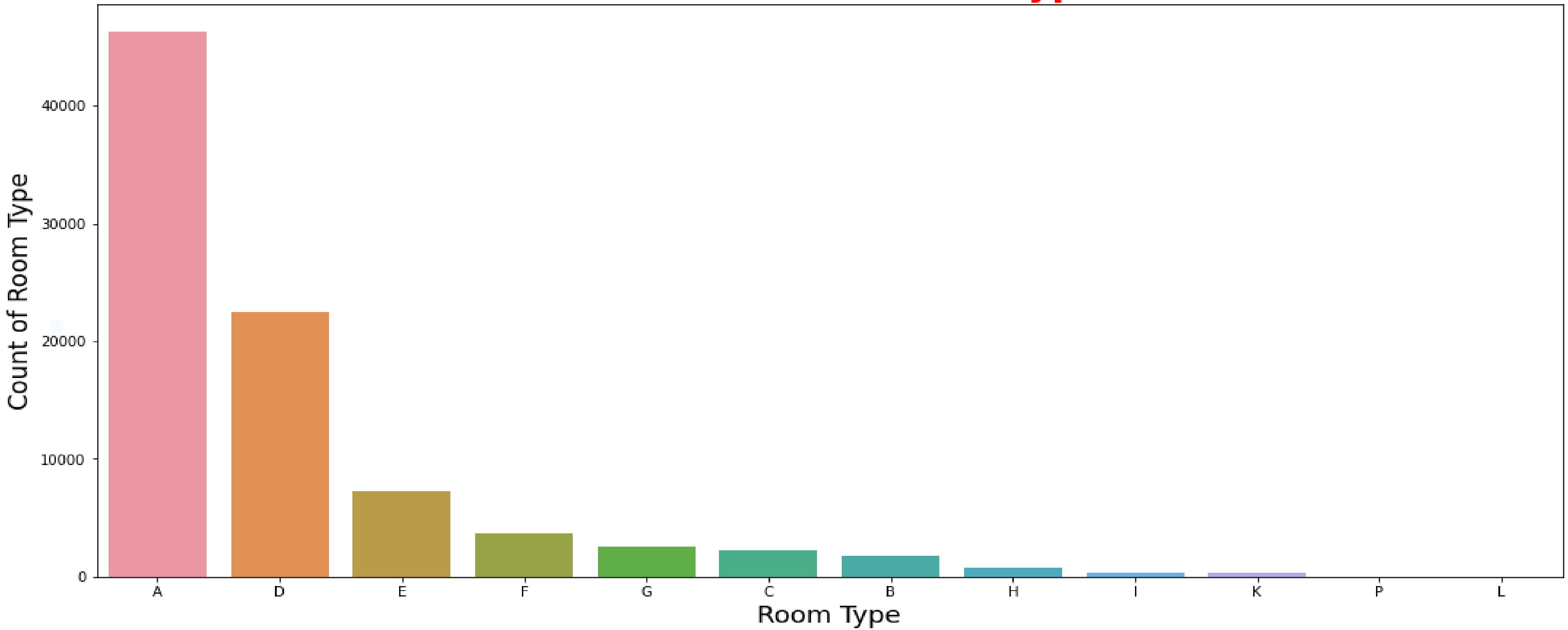
- ➢ 'TA/TO' is mostly(79.11%) used for booking hotels



Conclusions:

- July and August months had the most Bookings. Summer vacation can be the reason for the bookings.

Most Preffered Room Type

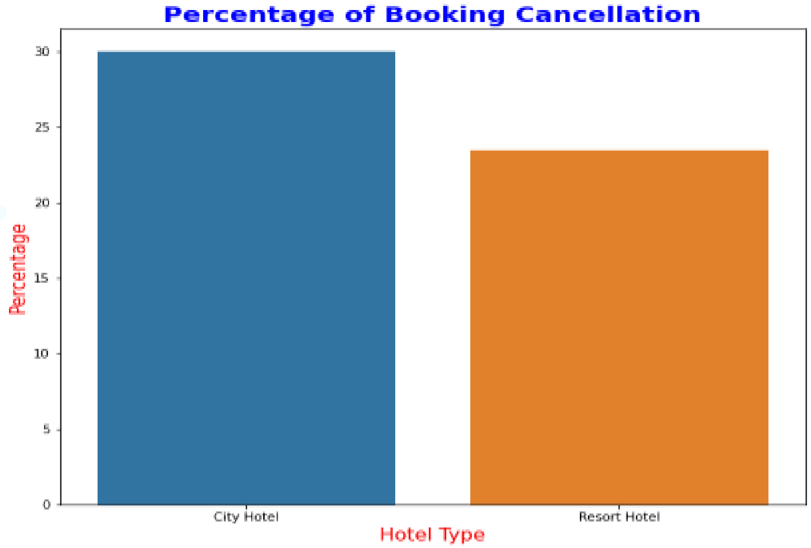


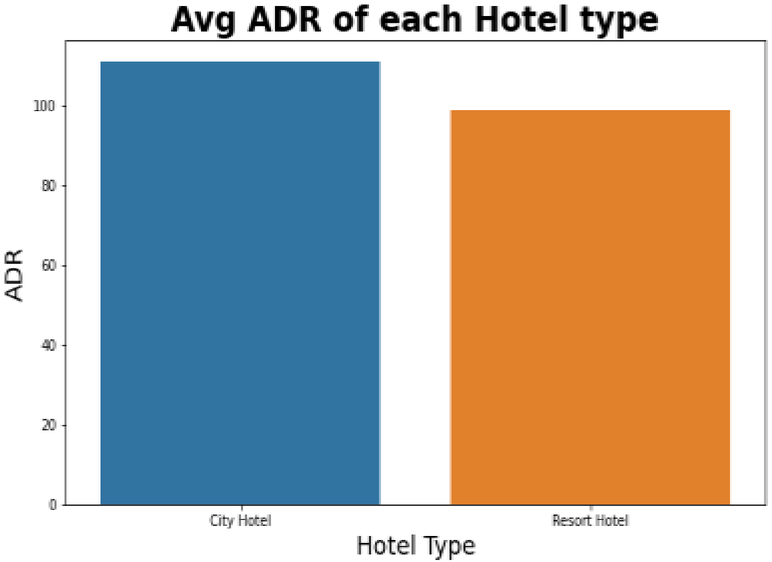
Conclusions:

➢ The most preferred Room type is "A".

Exploratory Data Analysis (EDA) :

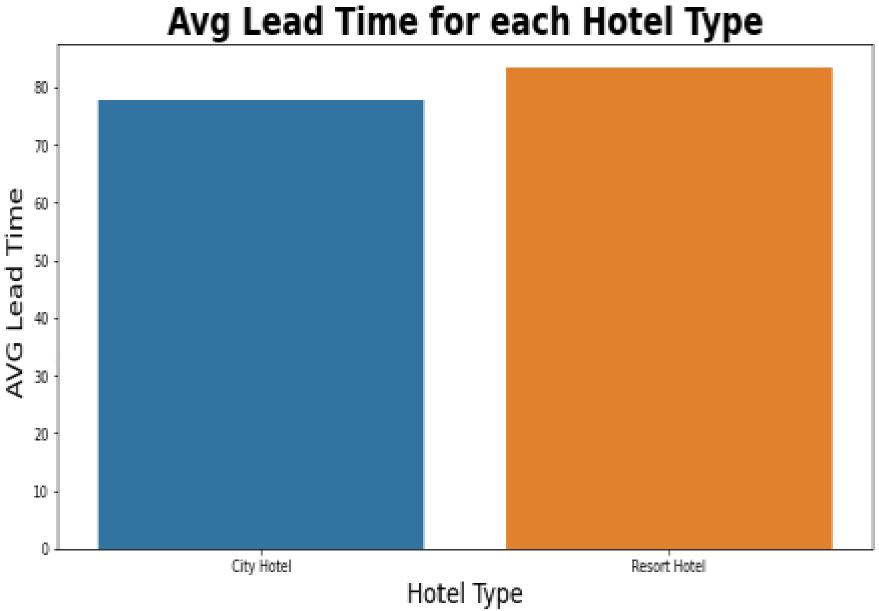
Bivariate and Multivariate Analysis

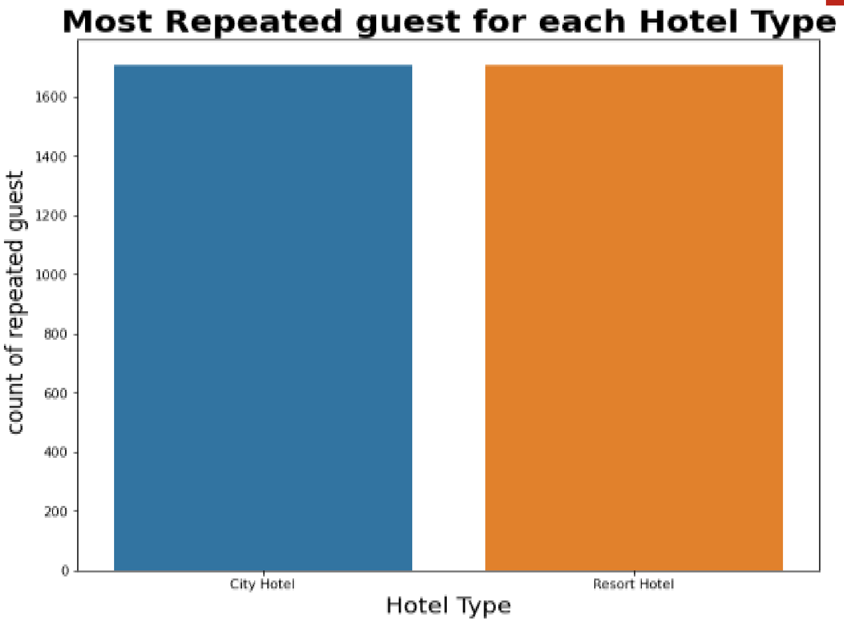




Conclusions:

- City Hotel has highest percentage of cancellation which is almost 30%
- City hotel has the highest ADR. That means city hotels are generating more revenues than the resort hotels. More the ADR more is the revenue.

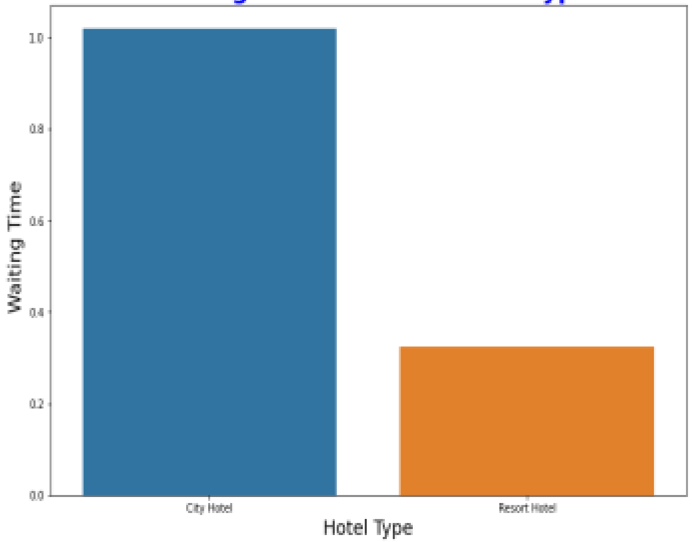




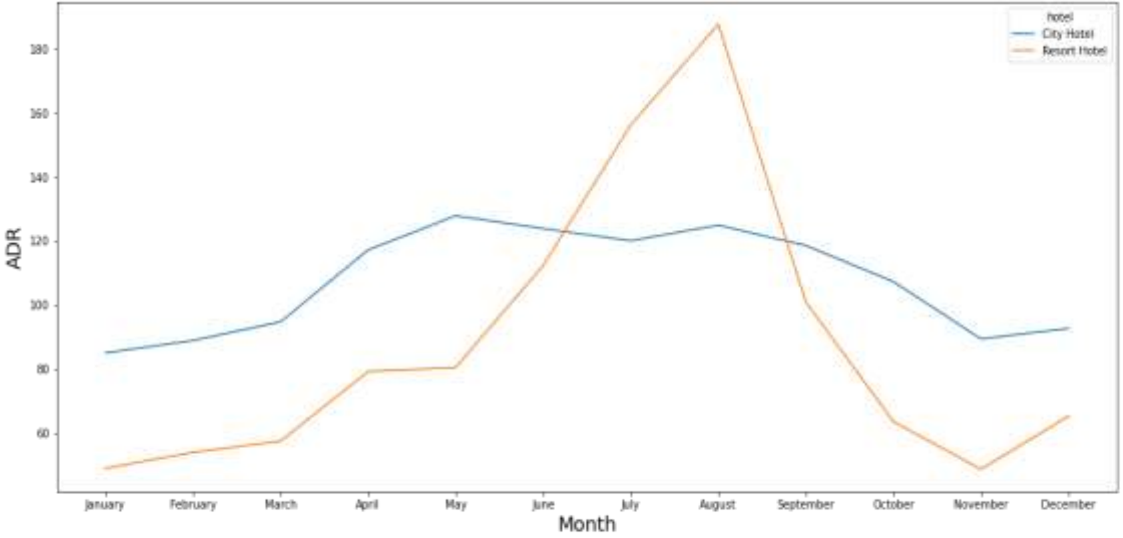
Conclusions:

- Resort hotels has slightly high avg lead time. That means customers plan their trips very early.
- The repeated guest is almost similar for both hotels.

Waiting Time for each Hotel Type



ADR across each month

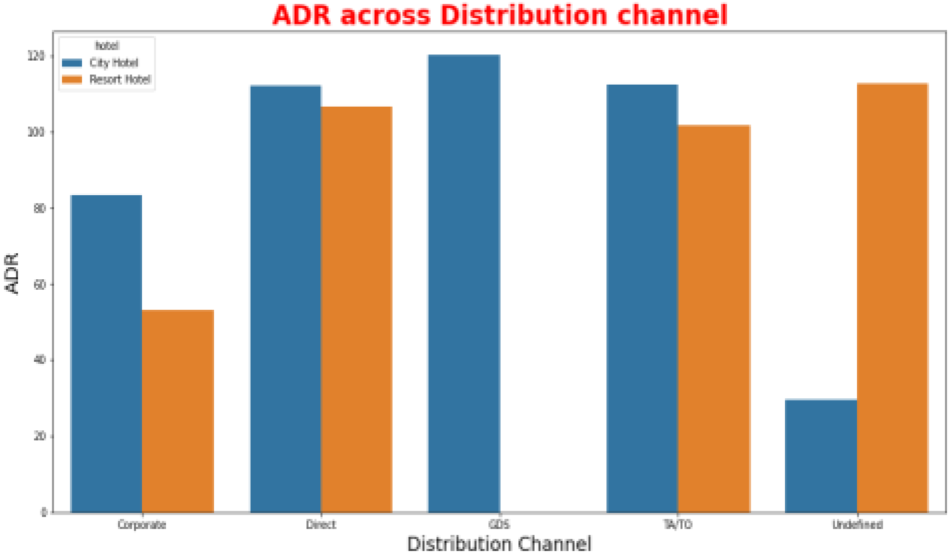


Conclusions:

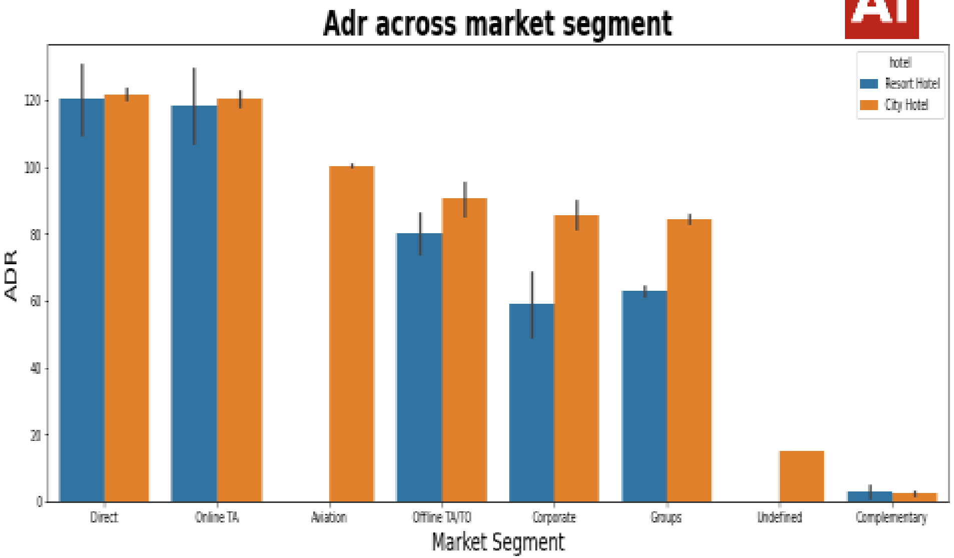
- So the City Hotels has longer waiting period than the Resort Hotels. Thus we can say that City Hotels are much busier than the Resort Hotels.

- For Resort hotel ADR is high in the months June, July, August as compared to City Hotels. May be Customers/People wants to spend their Summer vacation in Resorts Hotels.
- The best time for guests to visit Resort or City hotels is January, February, March, April, October, November and December as the average daily rate in this month is very low.

ADR across Distribution channel



Adr across market segment



Conclusions:

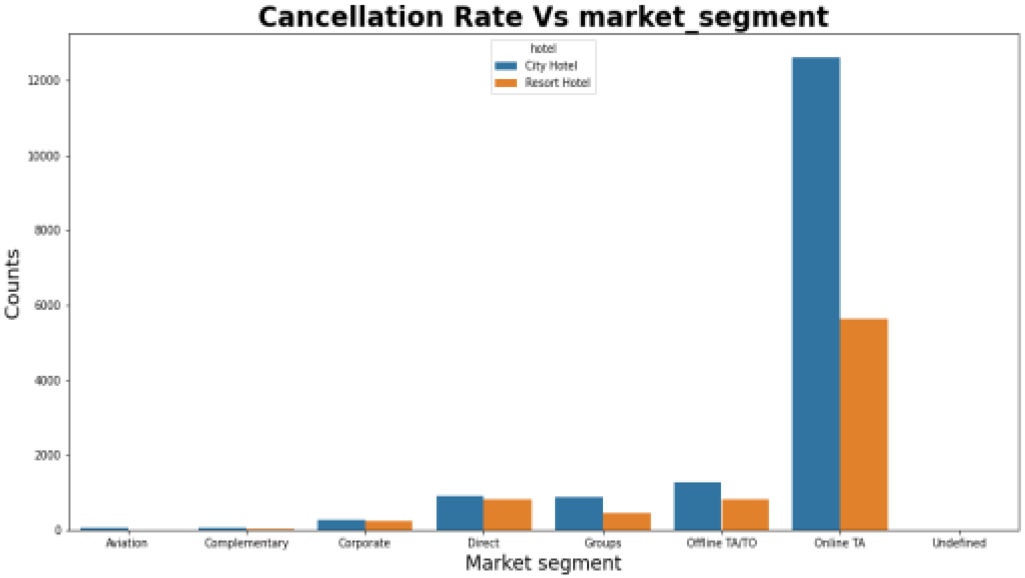
Distribution channel:

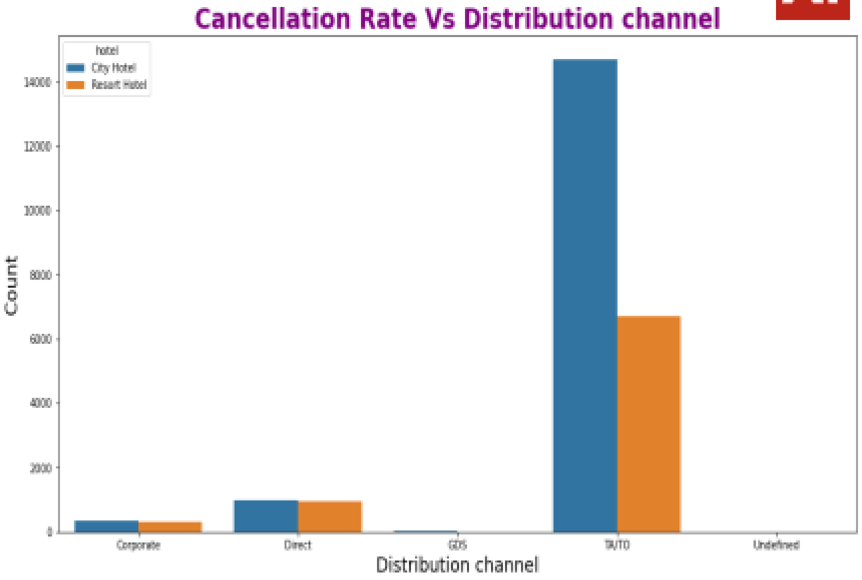
➢'Direct' and 'TA/TO' has almost equal adr in both type of hotels which is high among other channels.

➢GDS has high adr in 'City Hotel' type. GDS needs to increase Resort Hotel bookings. From this we can say that “Direct” and ‘TA/TO’ are generating more revenue than the other channels.

Market Segment:

➢Here “Direct” and ‘Online Travel Agency’ has high adr for both hotel types. Aviation segment needs to increase Resort hotel bookings.

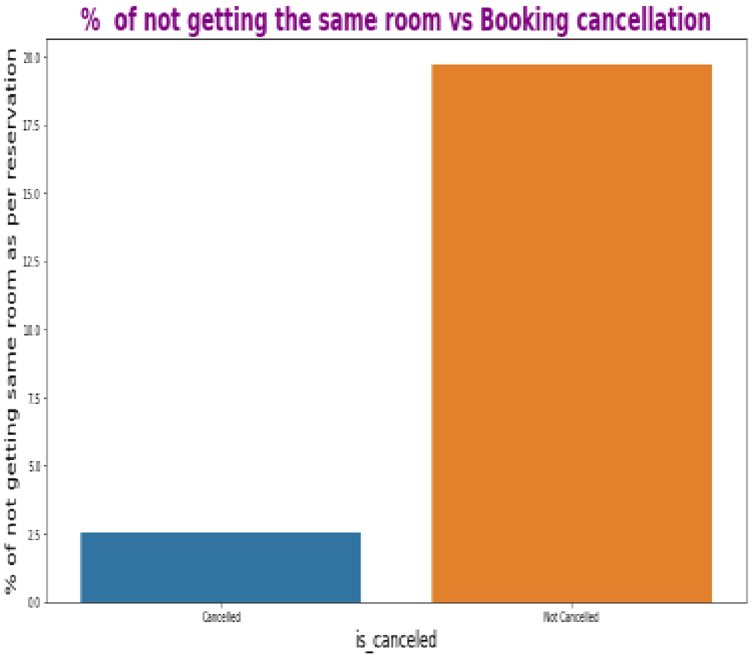


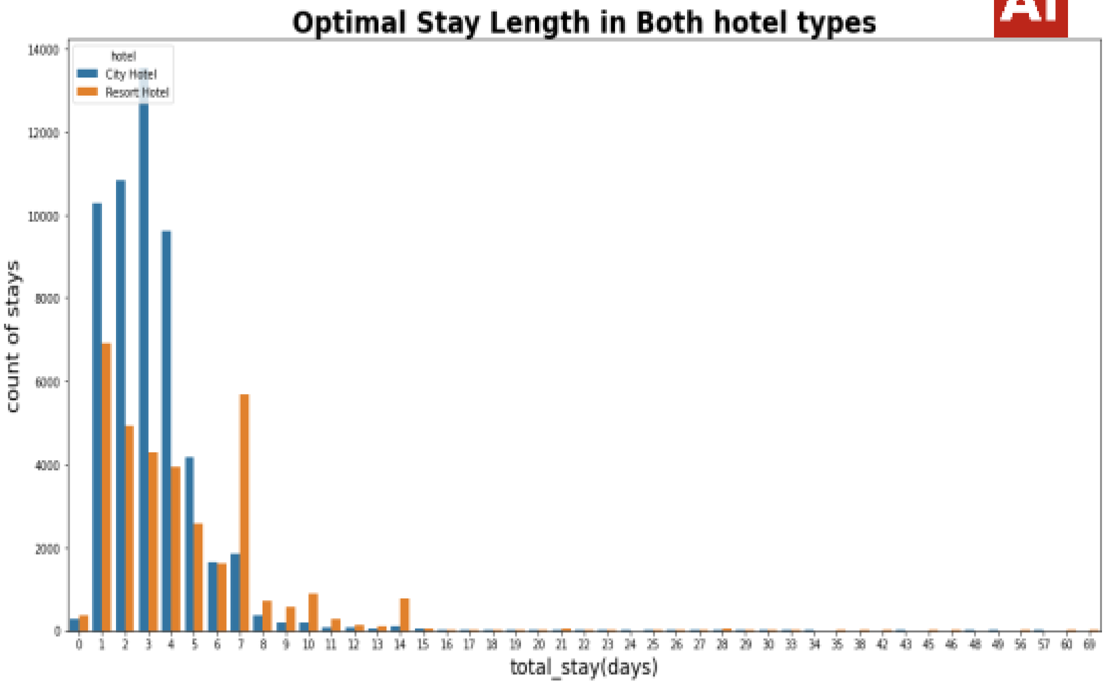


Conclusions:

- 'Online T/A' has the highest cancellation in both type of Hotel
- In order to reduce the booking cancellations hotels need to set the refundable/ no refundable and deposit policies

- In "TA/TO", City hotels has the high cancellation rate compared to resort hotels.
- In "direct" both the hotels has almost same cancellation rate..

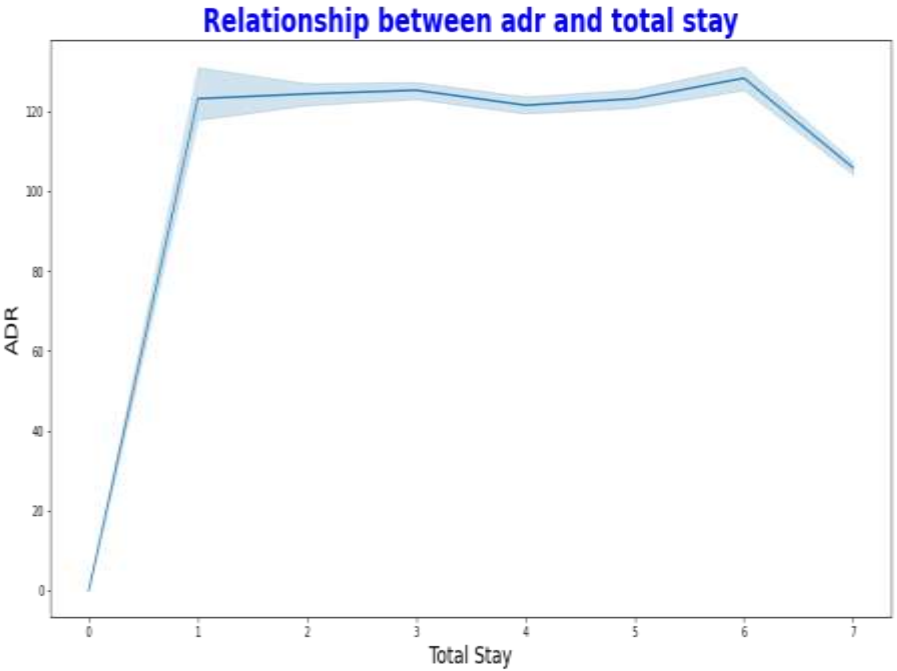




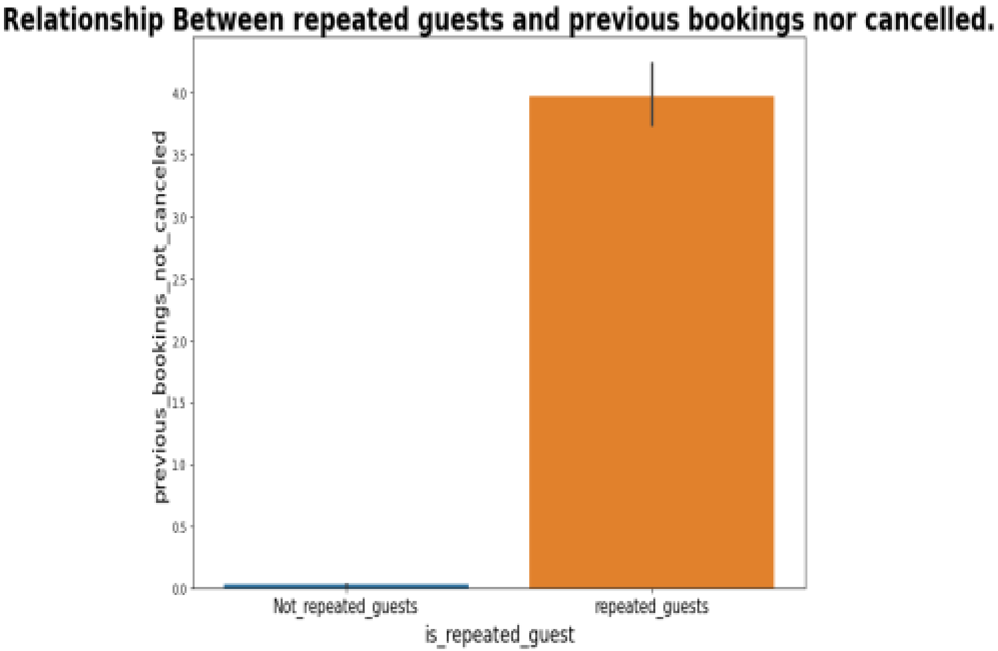
Conclusions:

- Its is clear that there is no much(2.5%) effect on cancellation of the bookings even if the guests are not assigned with rooms which they reserved during booking.

- Optimal stay in both the type hotel is less than 7 days.



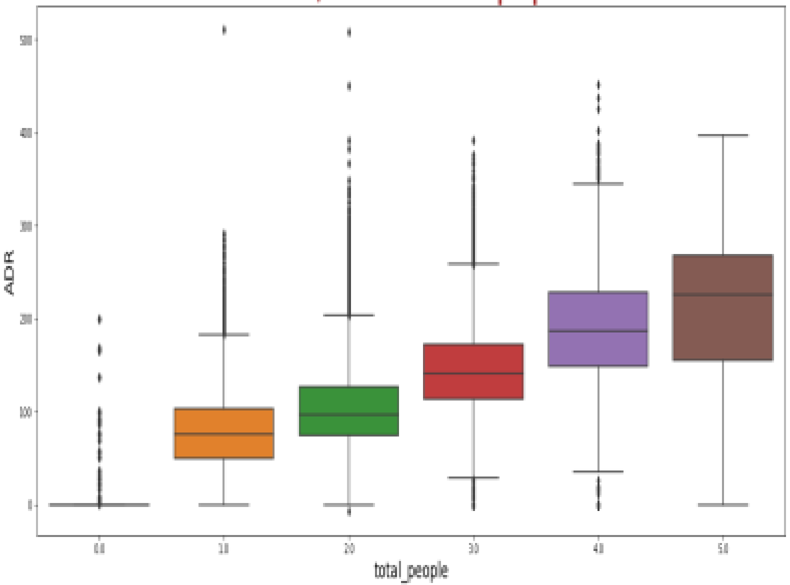
Relationship between adr and total stay

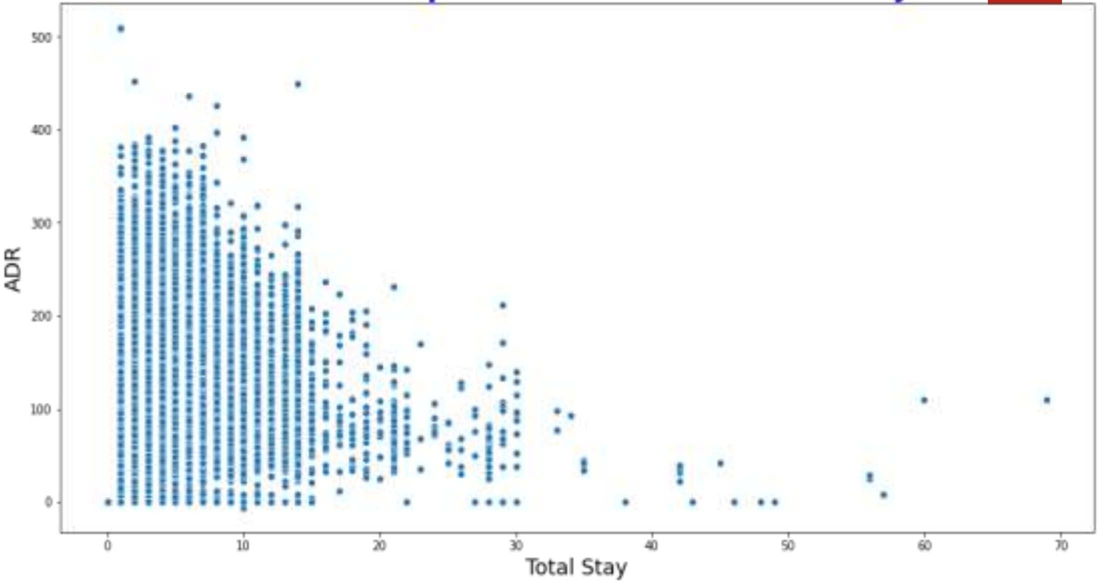


Relationship Between repeated guests and previous bookings nor cancelled.

Conclusions:

- As the total stay increases the adr also increases.
- Not Repeated guests are more likely to cancel their bookings.


ADR v/s Total Number of people

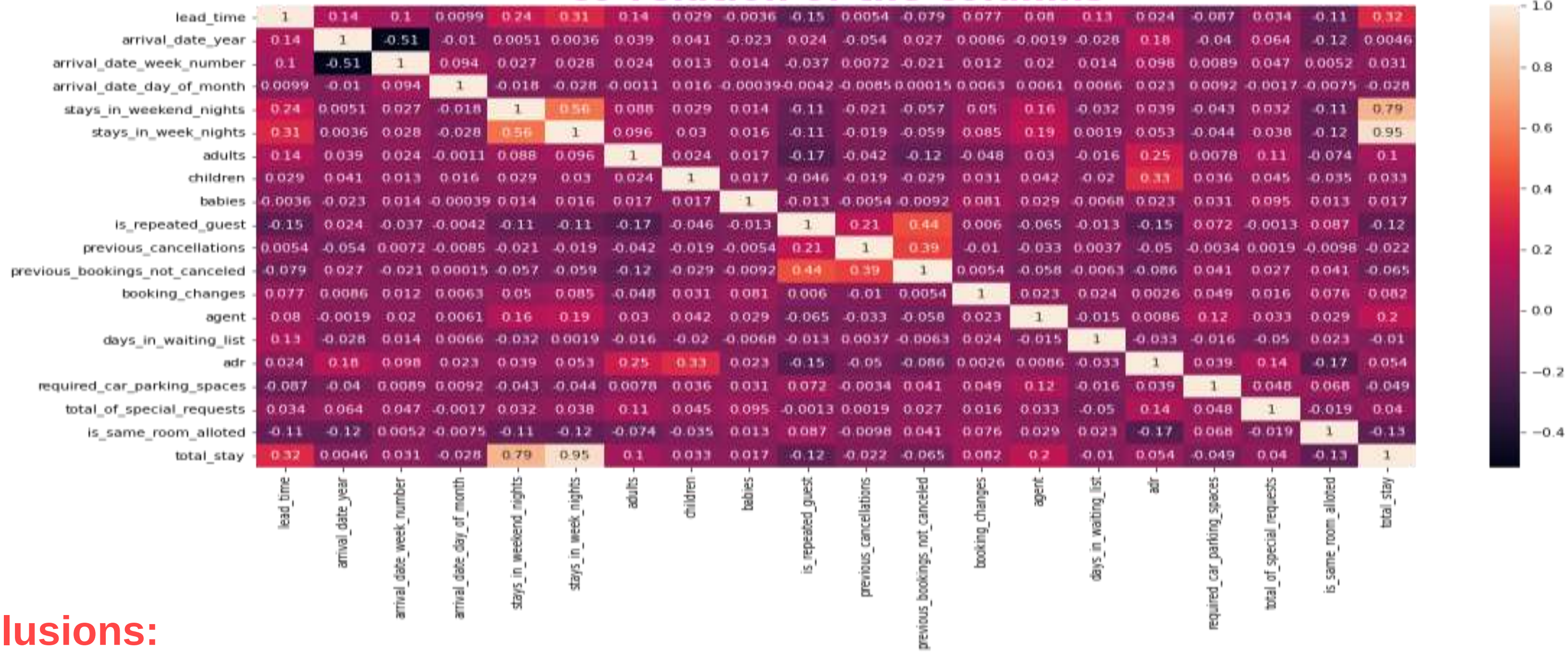

Relationship between adr and total stay

Conclusions:

- As the total number of people increases adr also increases.
- Thus adr and total people are directly proportional to each other.

- From above scatter we can say that as the stay increases adr is decreasing. Thus for longer stays customer can get good adr.

co-relation of the columns



Conclusions:

- is_canceled and same_room_alloted_or_not are negatively corelated. That means customer is unlikely to cancel his bookings if he don't get the same room as per reserved room. We have visualized it above.
- lead_time and total_stay is positively correlated. That means more is the stay of customer more will be the lead time.
- adults, childrens and babies are corelated to each other. That means more the people more will be adr.
- is_repeated guest and previous bookings not canceled has strong correlation. may be repeated guests are not more likely to cancel their bookings.

THANK YOU