Capstone Project-1

EDA On Hotel Booking

BY

Saidul Mondal

(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 I 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. Data had 31994 duplicates values. So I dropped it from the data.
3. I created 2 new columns

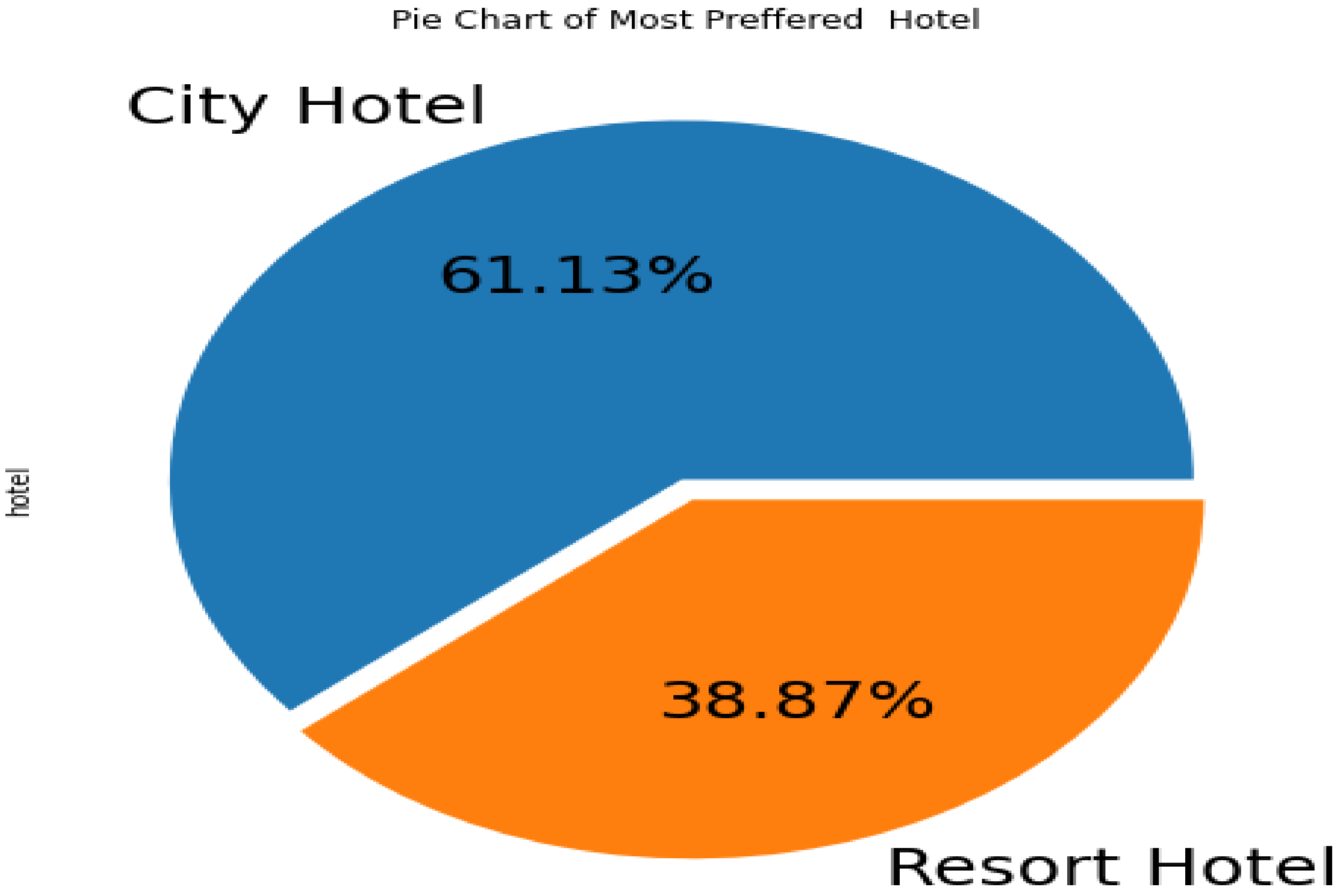
A)‘total_People’ = from the Children, adults, babies.

B) ‘total_stay’ = From weekend nights and weekdays night.

Exploratory Data Analysis (EDA) :

Univariate Analysis

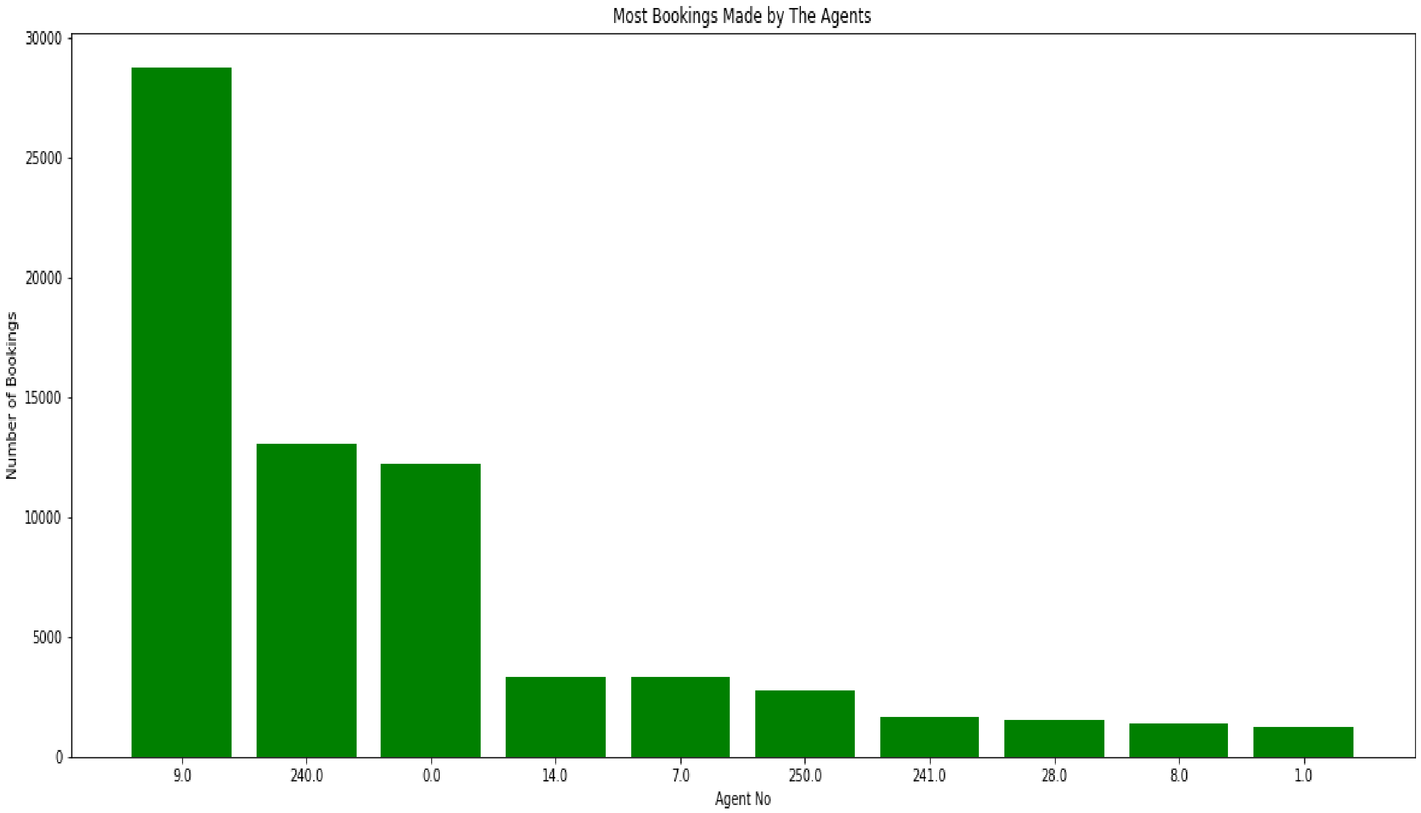
1.Which type of hotel is most prefered by people?



Conclusions:

City Hotel is most preffered hotel by people beacuse it has 61.13% bookings.

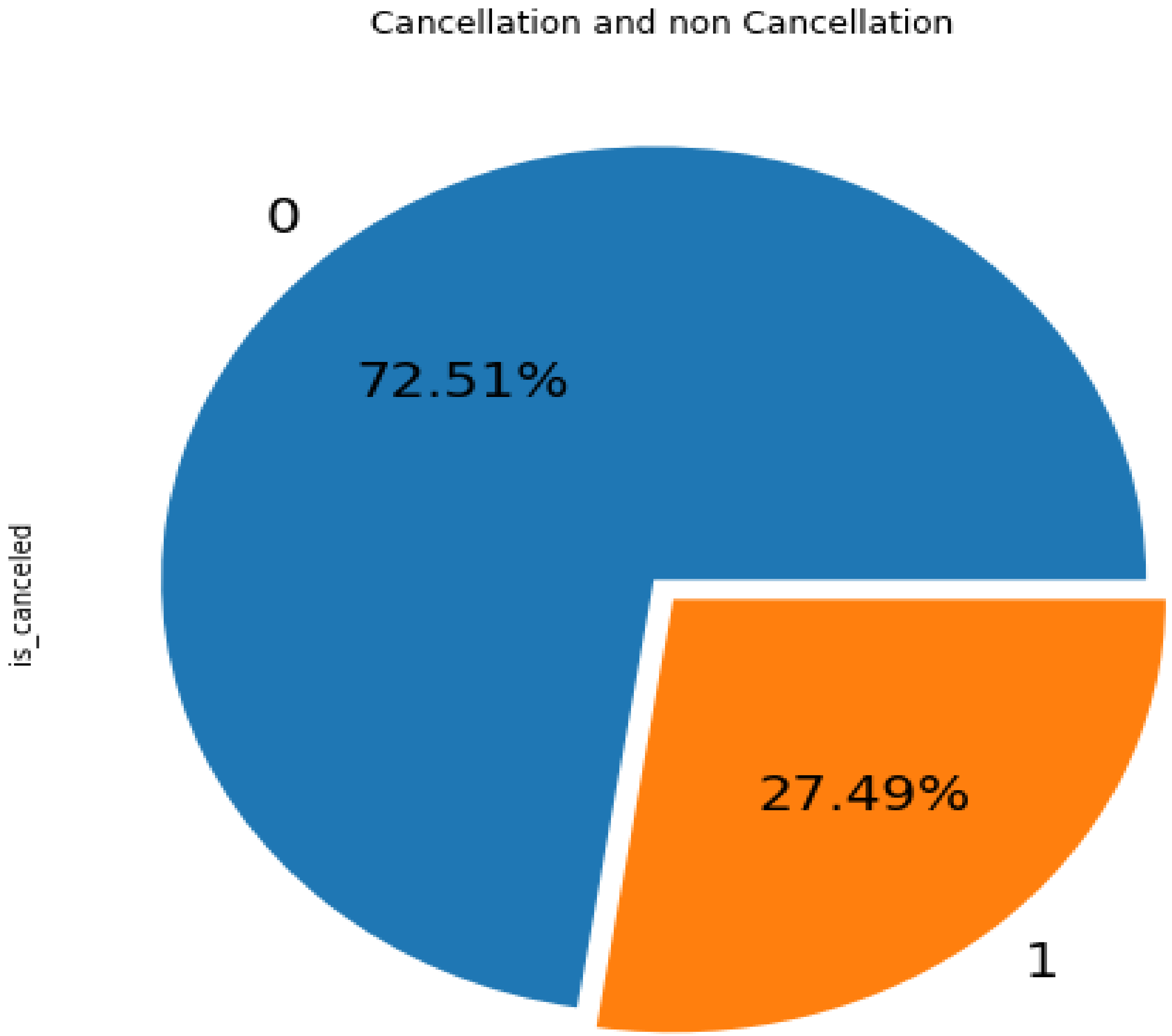
2.Which Agent made the most bookings?



Conclusions:

Agent ID no: 9.0 made most of the bookings

3.What is the pecentage of cancellation?

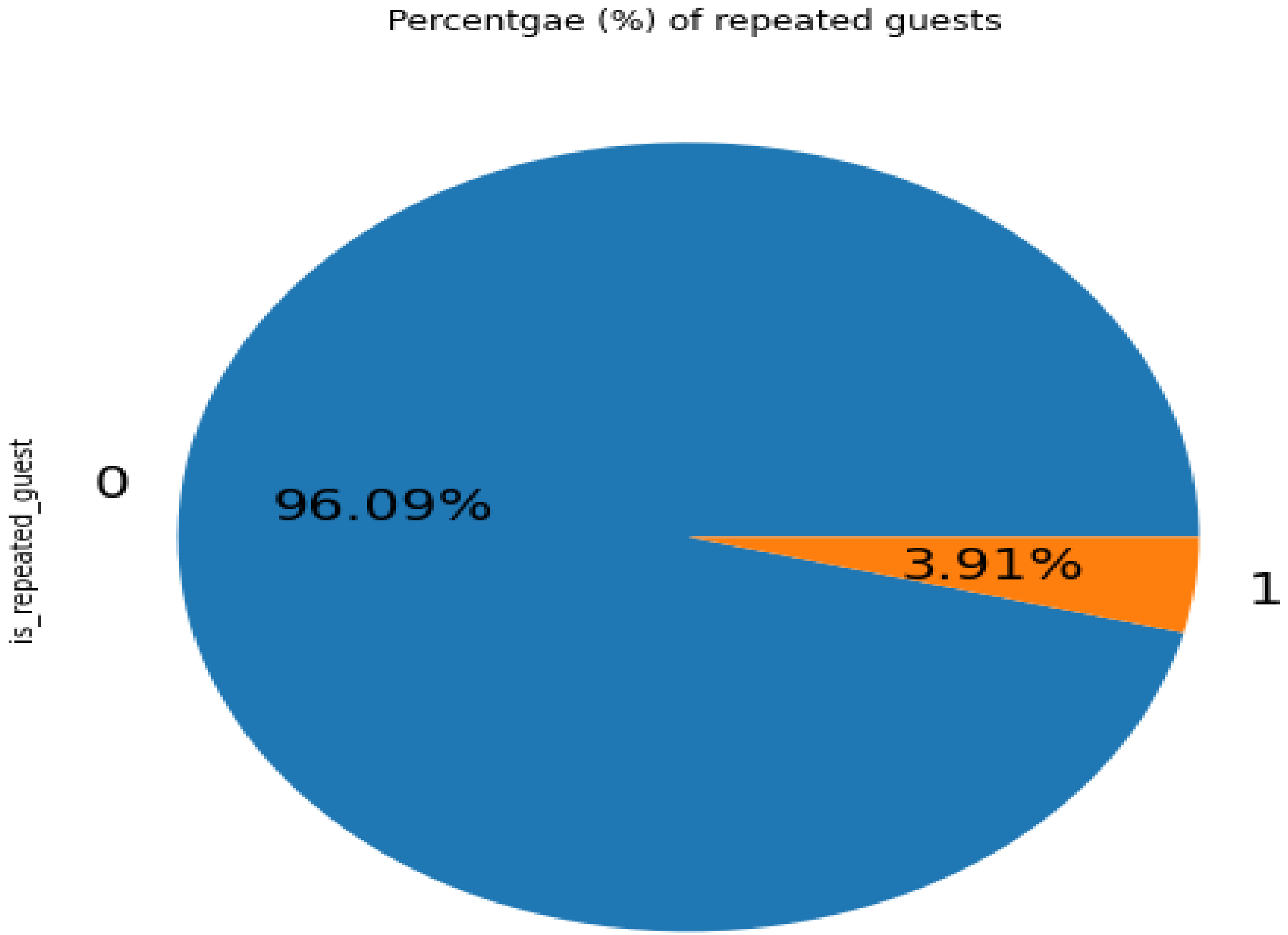


Conclusions:

0= not cancled
1= canceled

27.49 % of the bookings were cancelled.

4.What is the Percentage of repeated guests?

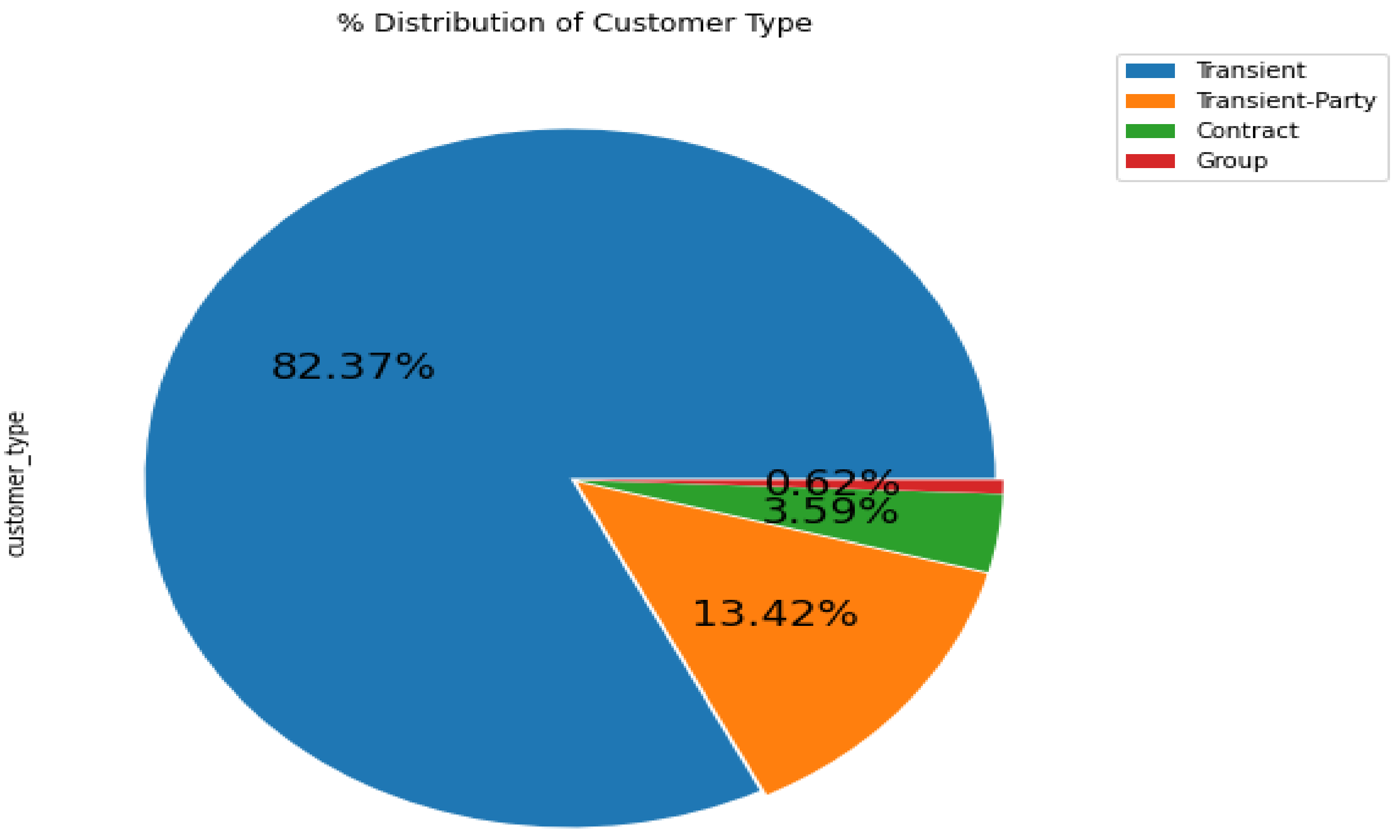


Conclusions:

Repeated guests are very few which only 3.91 %.

Suggestion: Guests management should take feedbacks from guests and try to imporve the services.

5.What is the percentage distribution of "Customer Type"?



Observation:

1. Contract

When the booking has an allotment or other type of contract associated to it

2. Group

When the booking is associated to a group

3. Transient

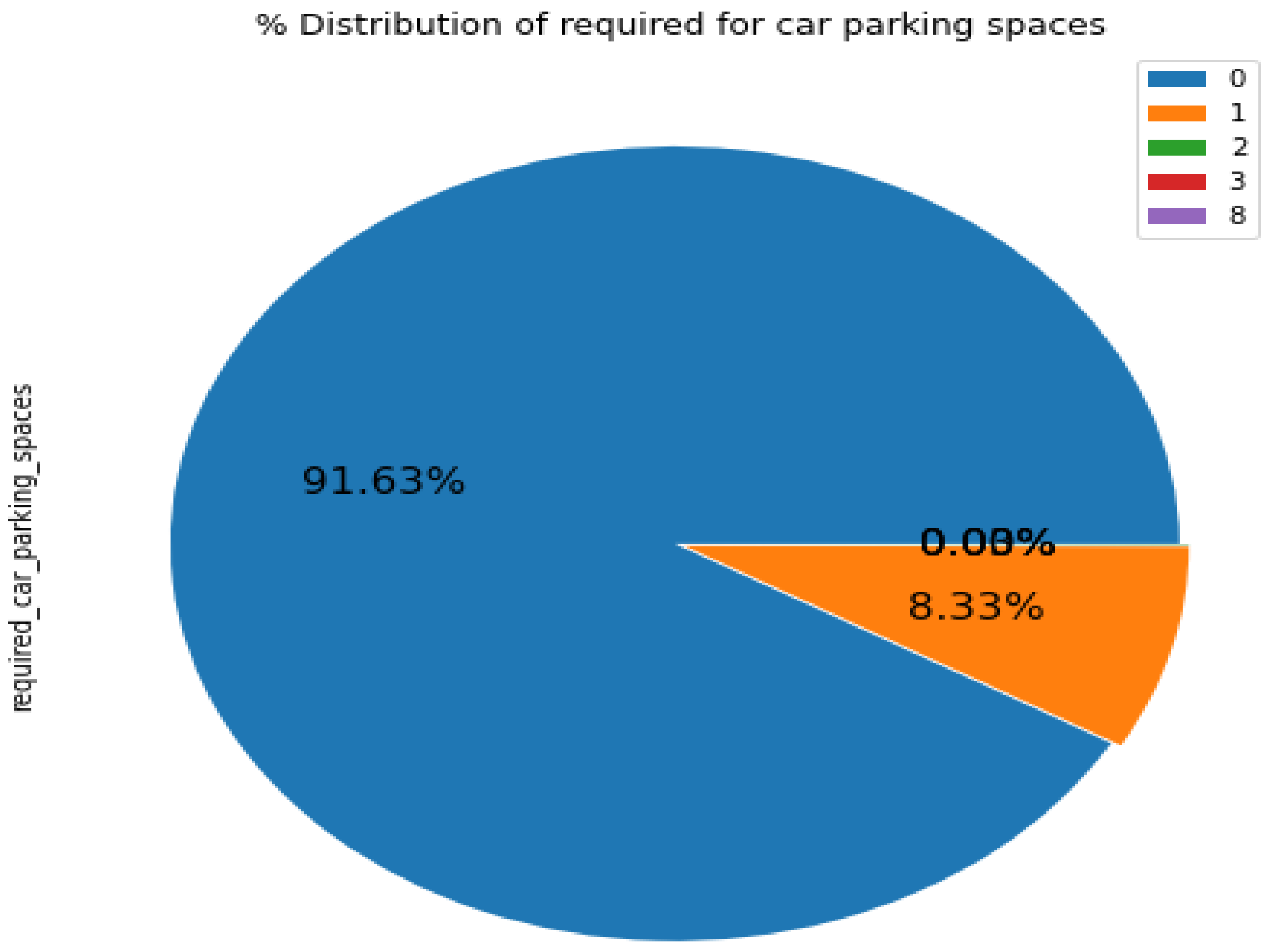
When the booking is not part of a group or contract, and is not associated to other transient booking

4. Transient-party

When the booking is transient, but is associated to at least other transient booking

Conclusions: Transient customer type is more whcih is 82.37 %. percentage of Booking associated by the Group is vey low.

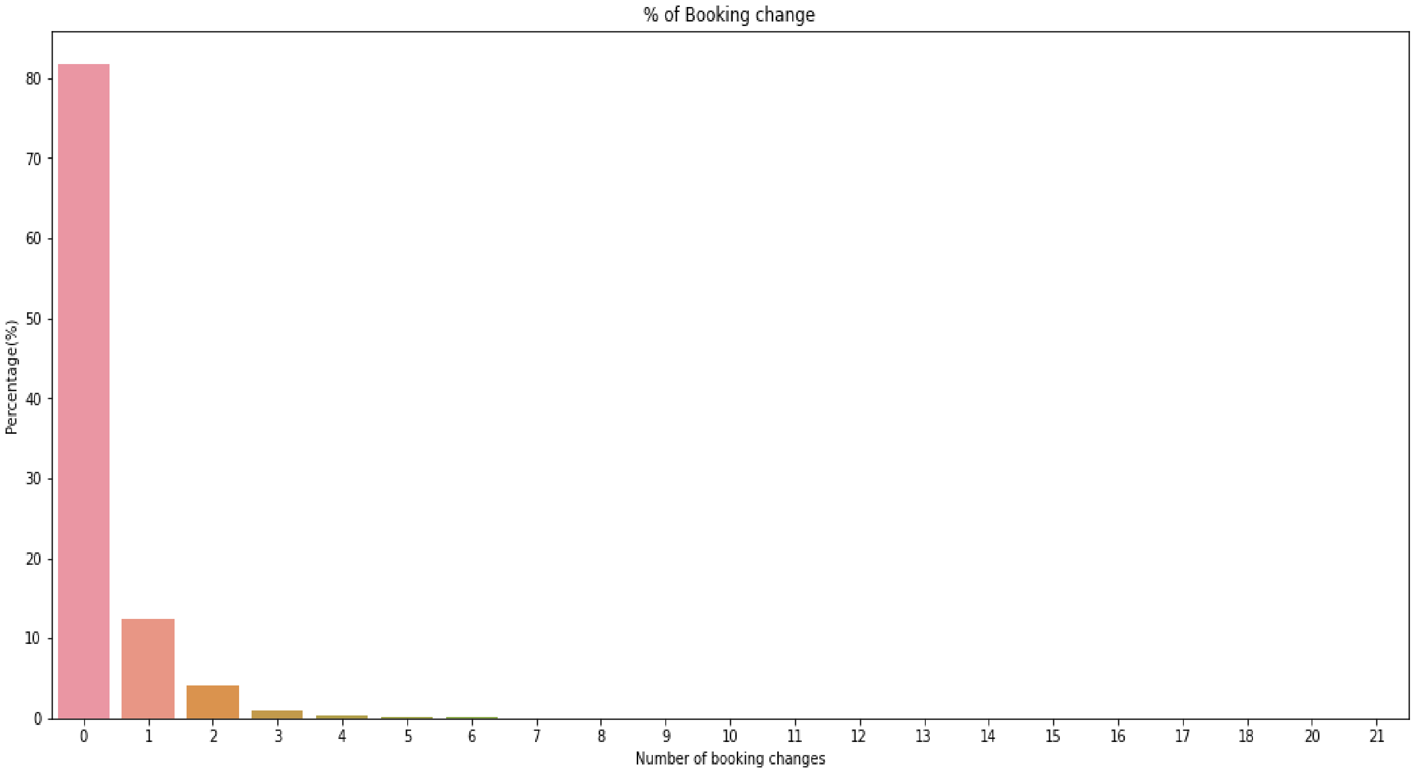
6.What is the percentage distribution of required car parking spaces?



Conclusions:

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

7.What is the percentage of booking changes made by the customer?



Observation:

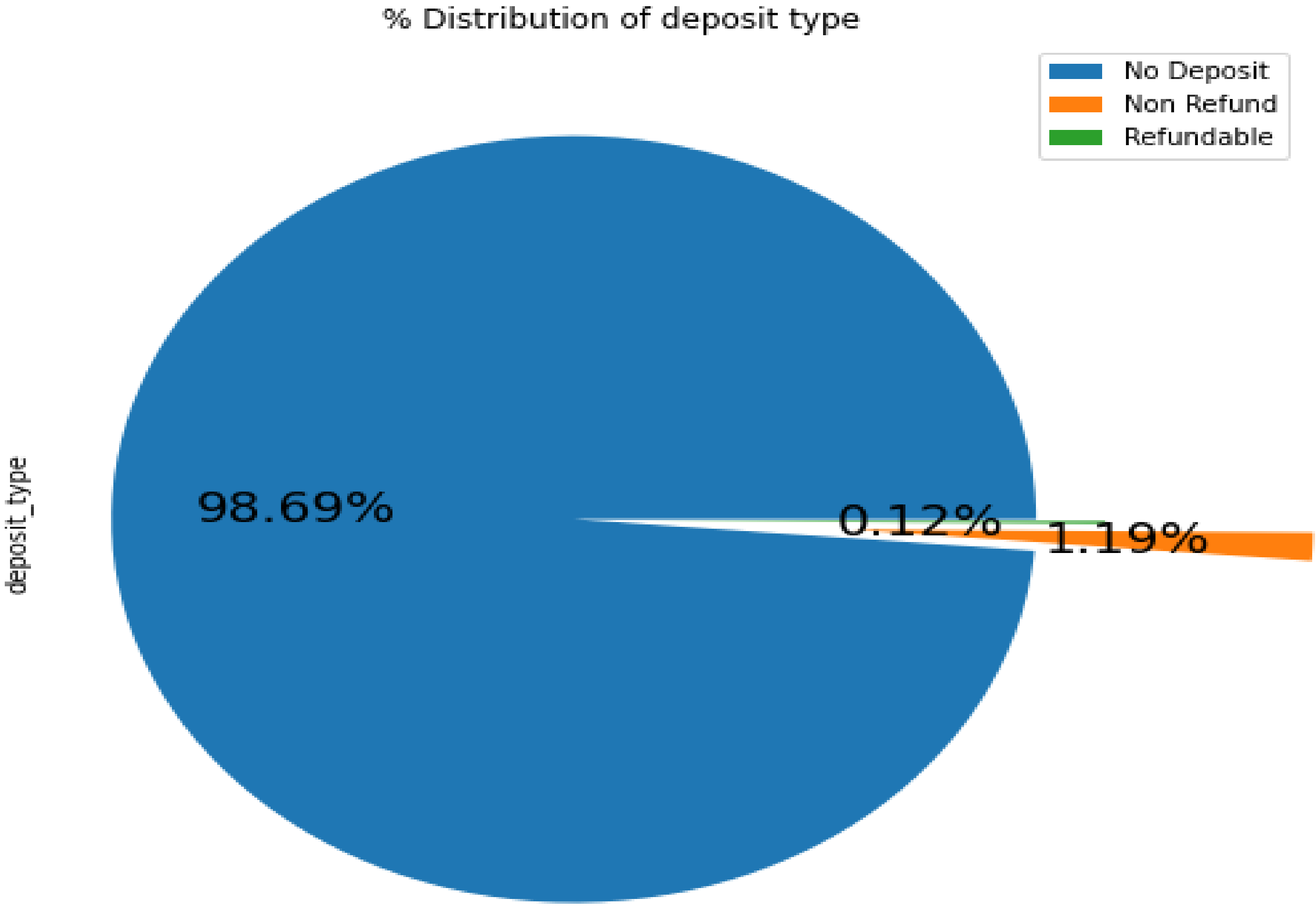
0 = 0 changes made in the booking

1 = 1 changes made in the booking

*2 = 2 changes made in the booking *

Conclusions: 80% -83% of the bookings were not changed by the people.

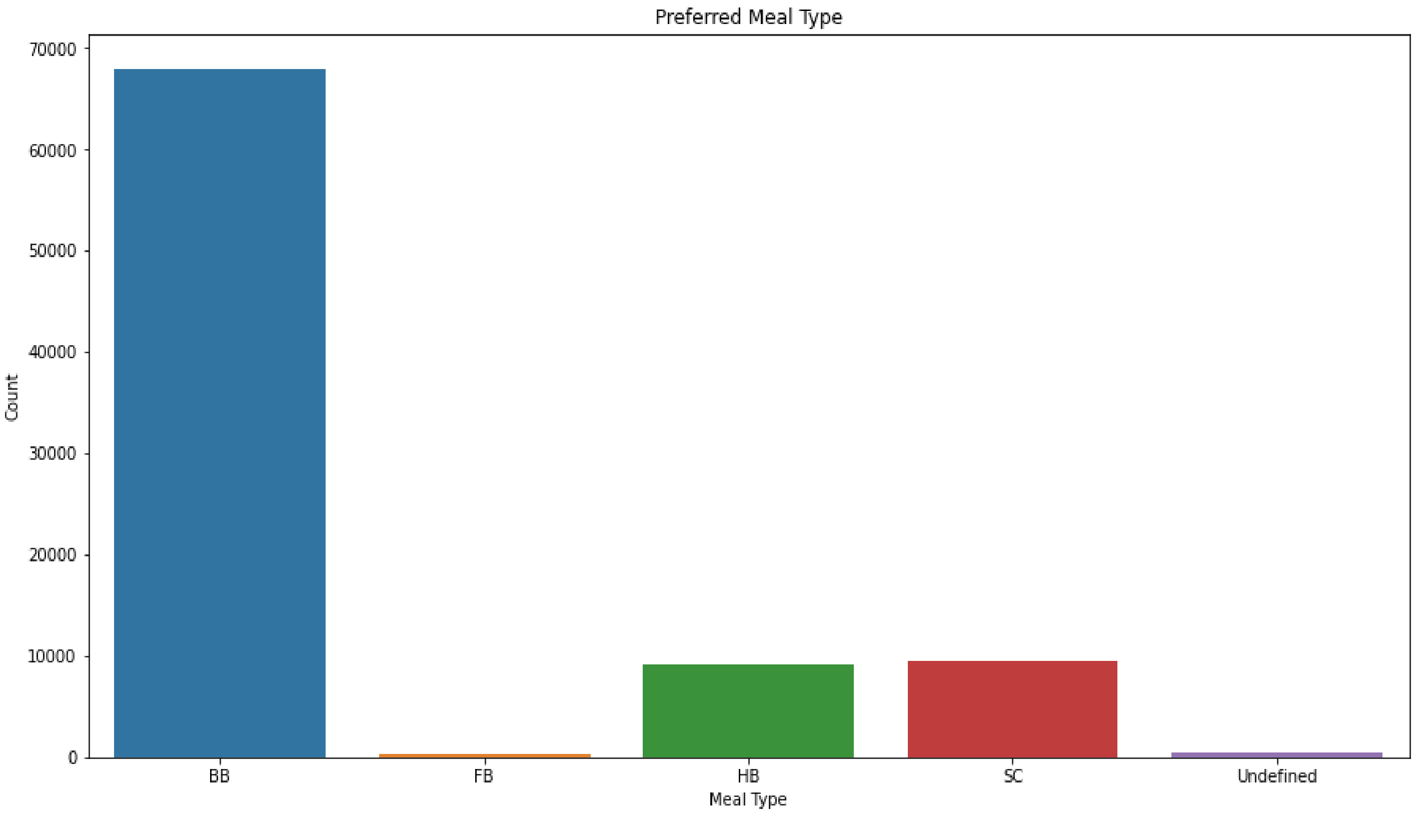
8.What is Percentage distribution of Deposite type ?



Conclusions:

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

9.Which type of food is mostly preferred by the guests?



Observation:

Types of meal in hotels:

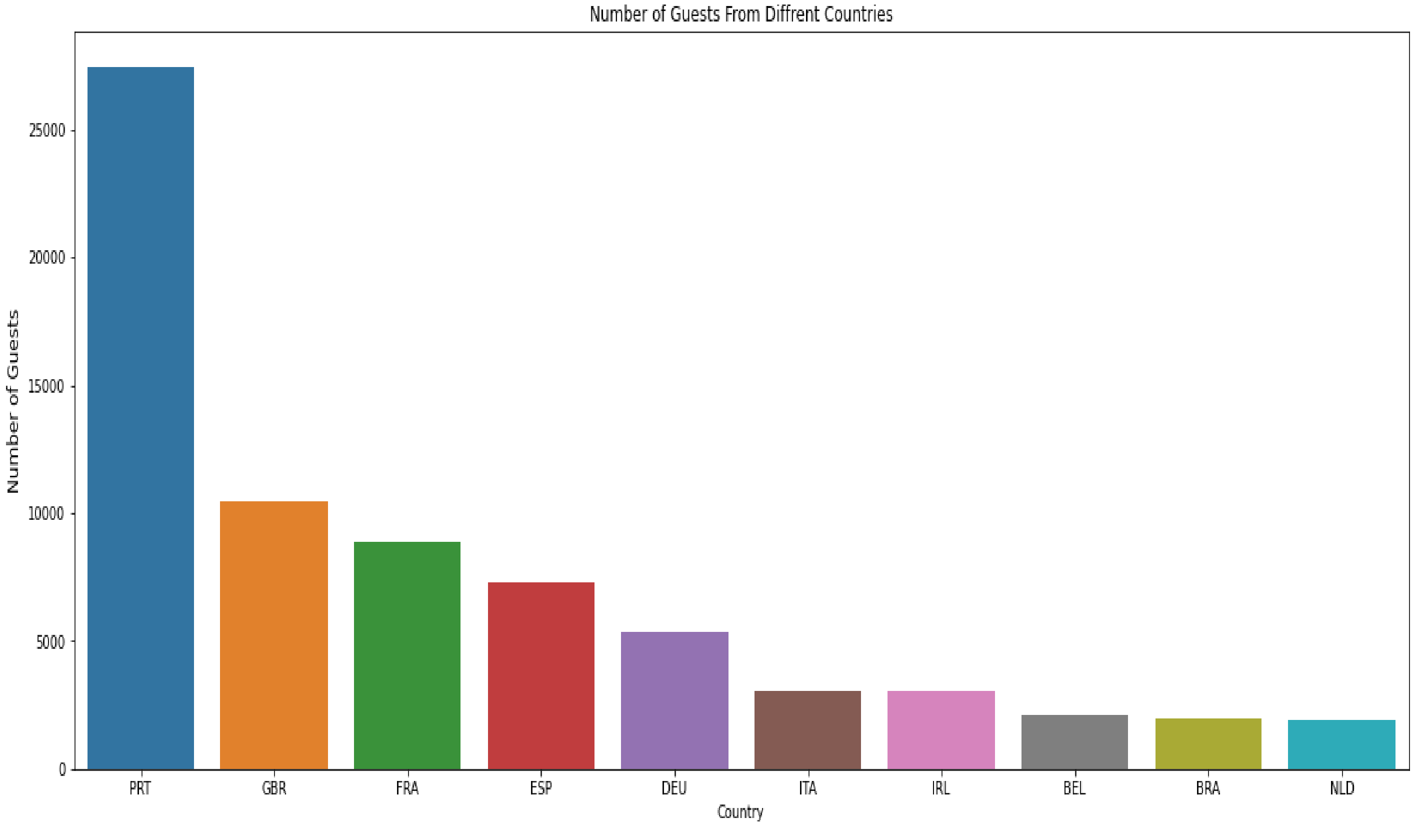
1. BB - (Bed and Breakfast)
2. HB- (Half Board)
3. FB- (Full Board)
4. SC- (Self Catering)

Conclusions:

So the most preferred meal type by the guests is BB(Bed and Breakfast)

HB- (Half Board) and SC- (Self Catering) are equally preferred.

10.From which country the most guests are coming?



Observation

Abbreevations for countries-

PRT- Portugal

GBR- United Kingdom

FRA- France

ESP- Spain

DEU - Germany

ITA -Itlay

IRL - Ireland

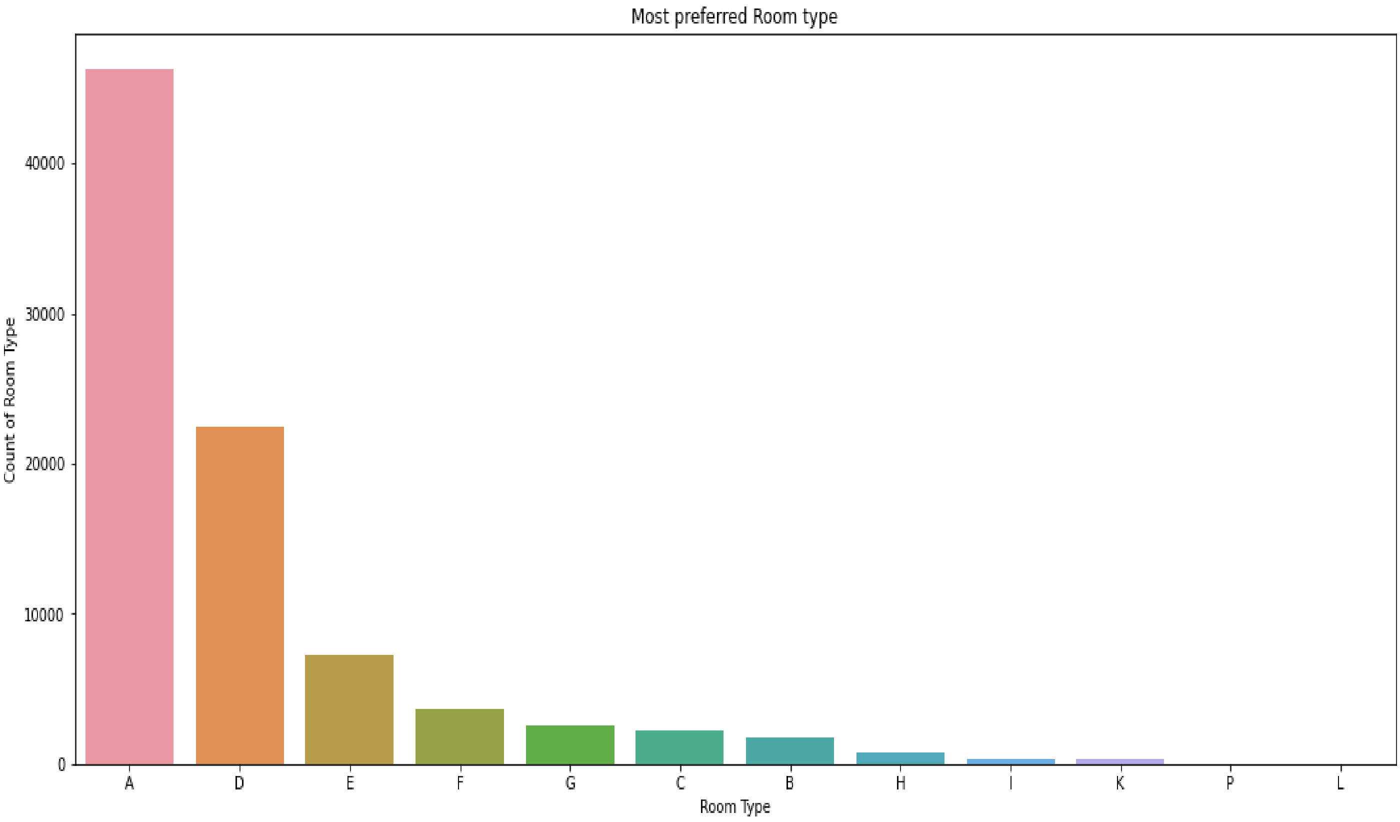
BEL -Belgium

BRA -Brazil

NLD-Netherlands

Conclusions: Most of the guests are coming from portugal .More than 25000 guests are coming from portugal.

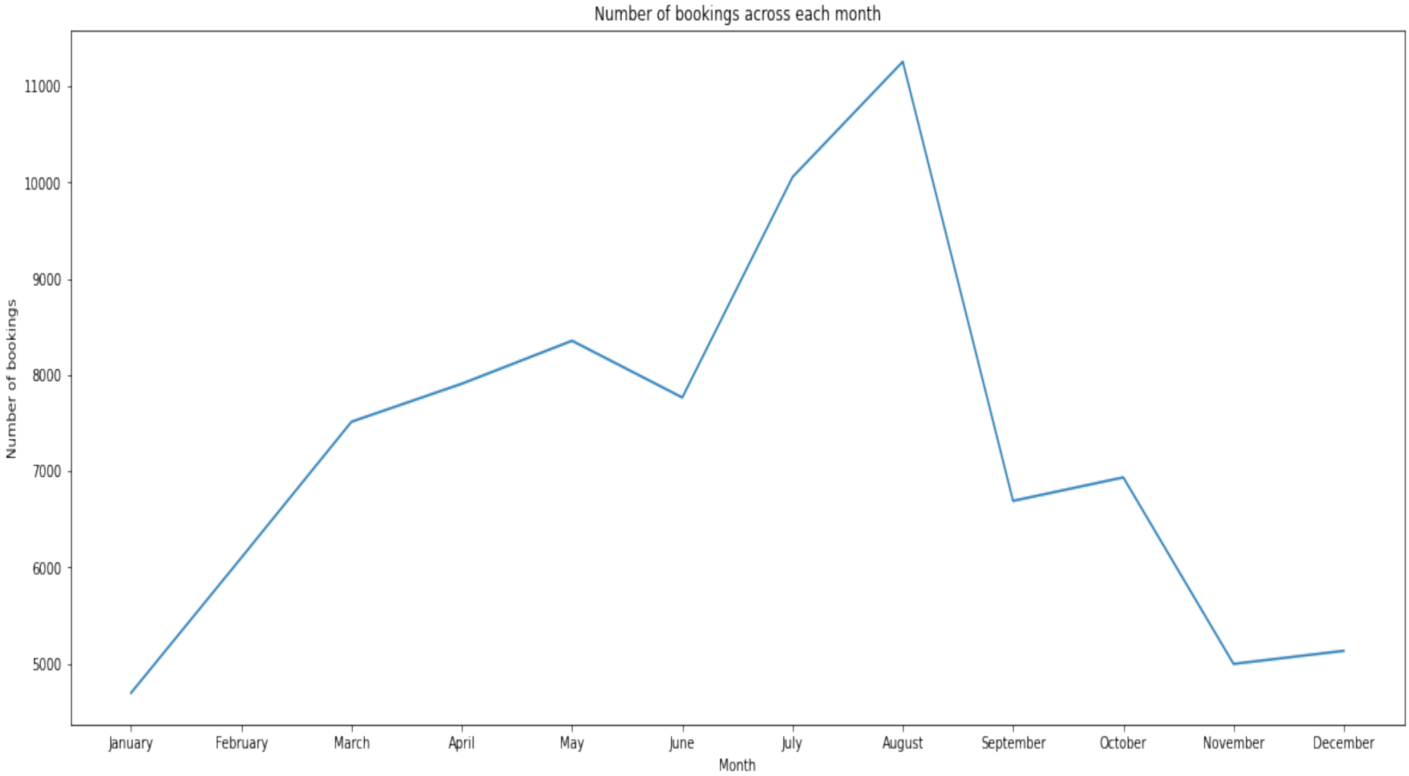
11.Which is the most preferred room type by the guests?



Conclusions:

The most preferred Room type is "A".

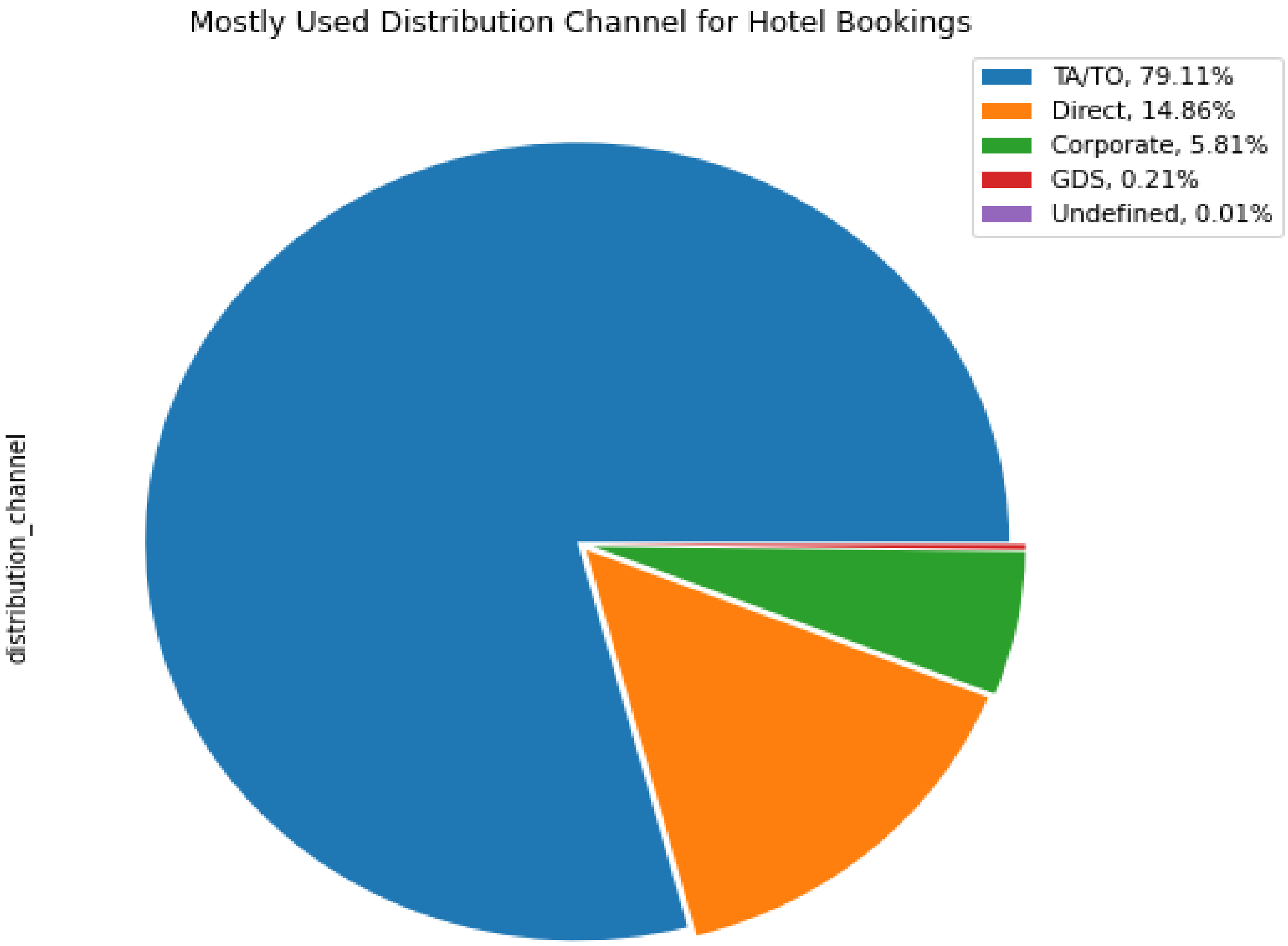
12.In which month most of the bookings happened?



Conclusions:

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

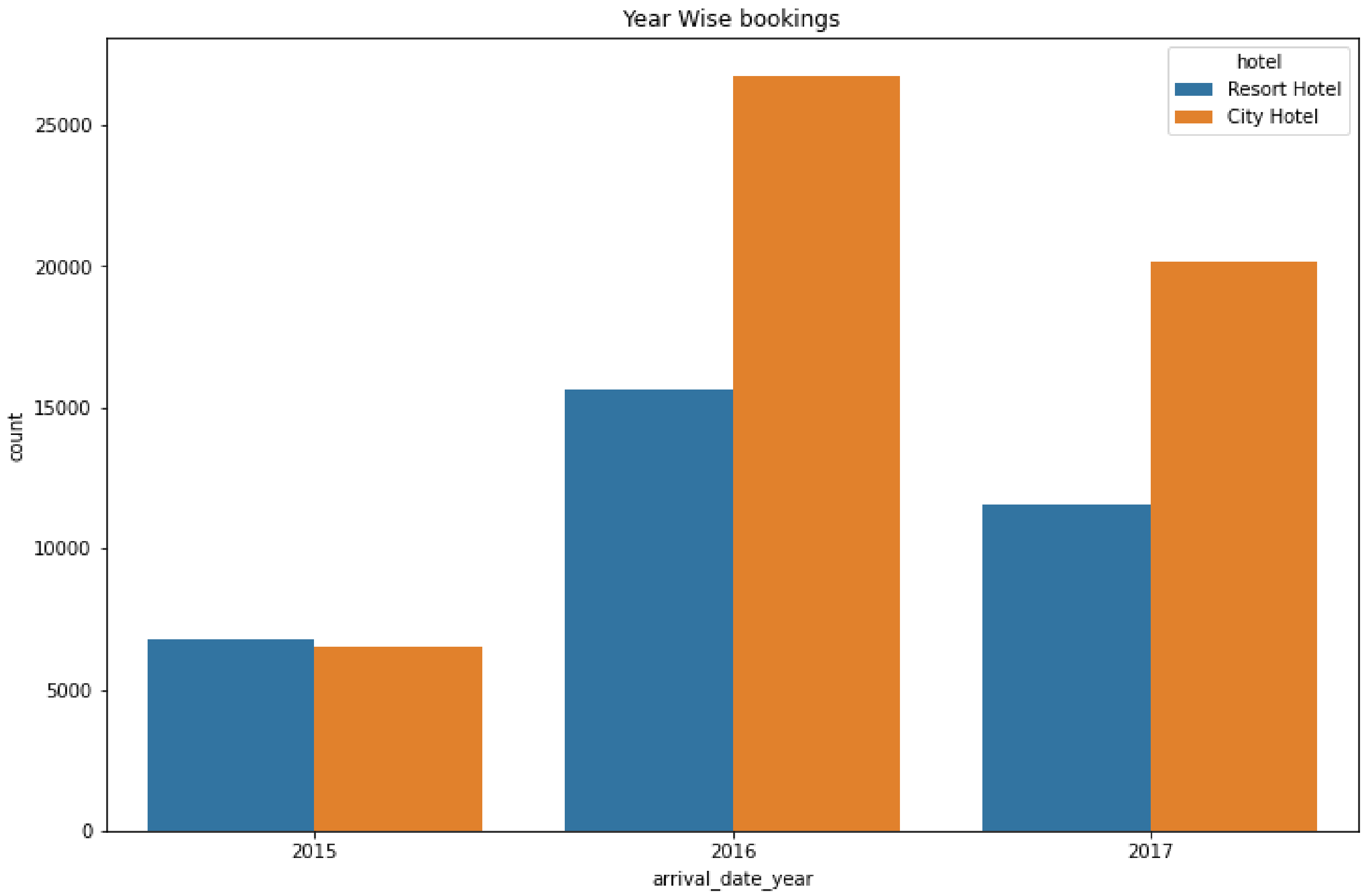
13.Which Distribution channel is mostly used for hotel bookings?



Conclusions:

'TA/TO' is mostly(79.11%) used for booking hoetls

14.Which year had the highest bookings?



Observation

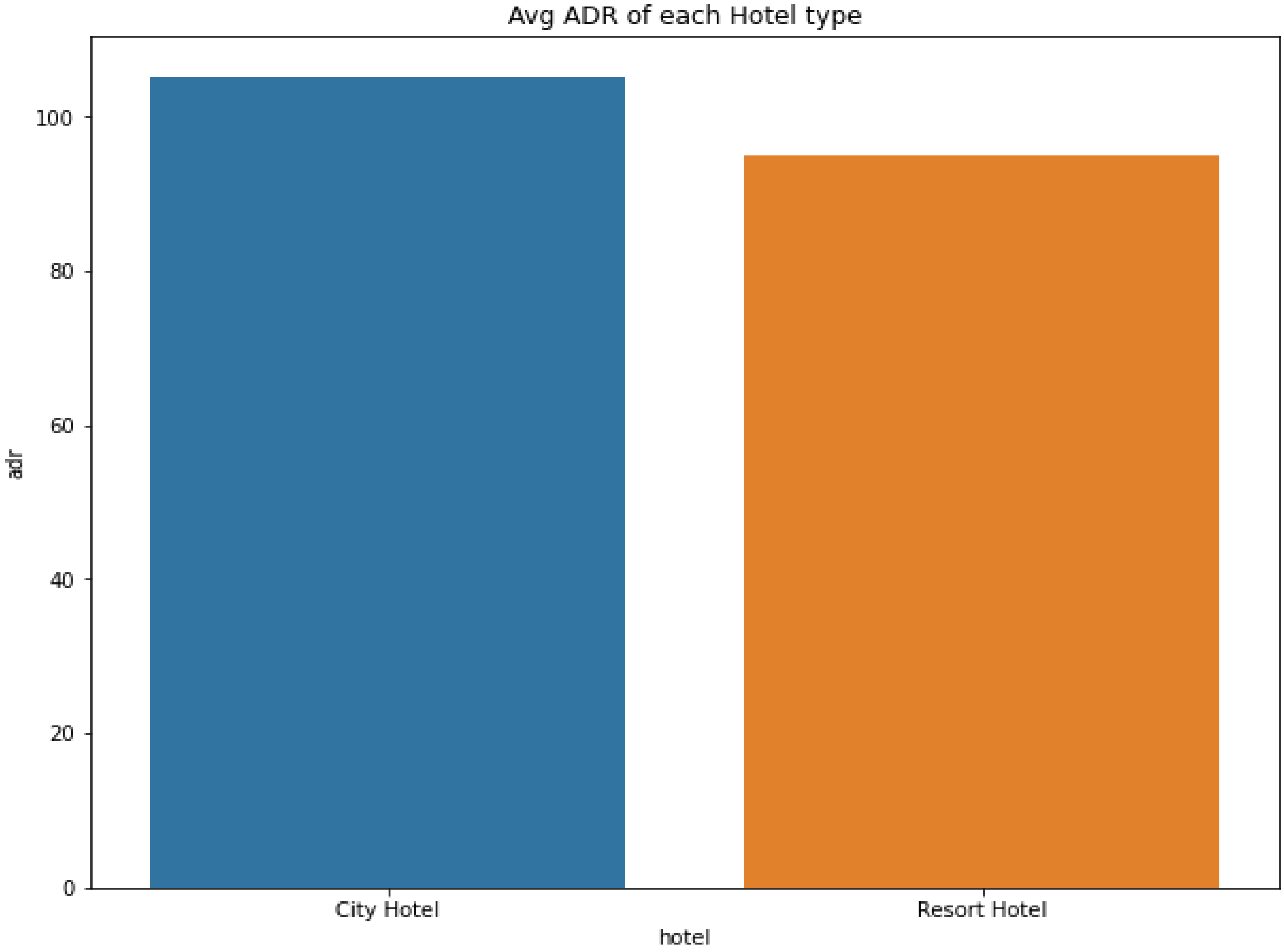
2016 had the higest bookings.

2015 had less 7000 bookings.

Conclusions: City hotels had the most of the bookings.

Bivariate and Multivariate Analysis

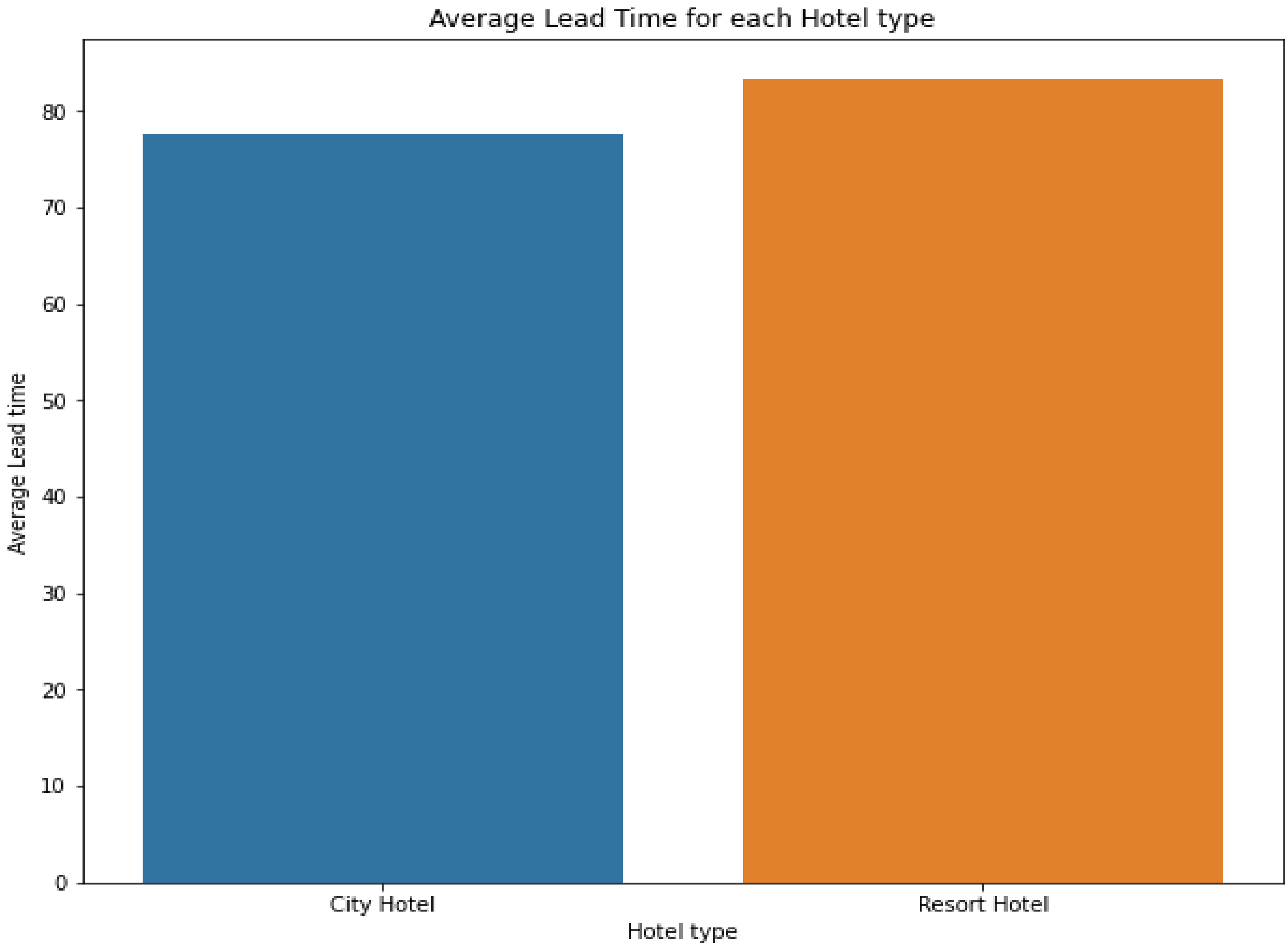
1.Which Hotel type has the highest ADR?



Conclusions:

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.

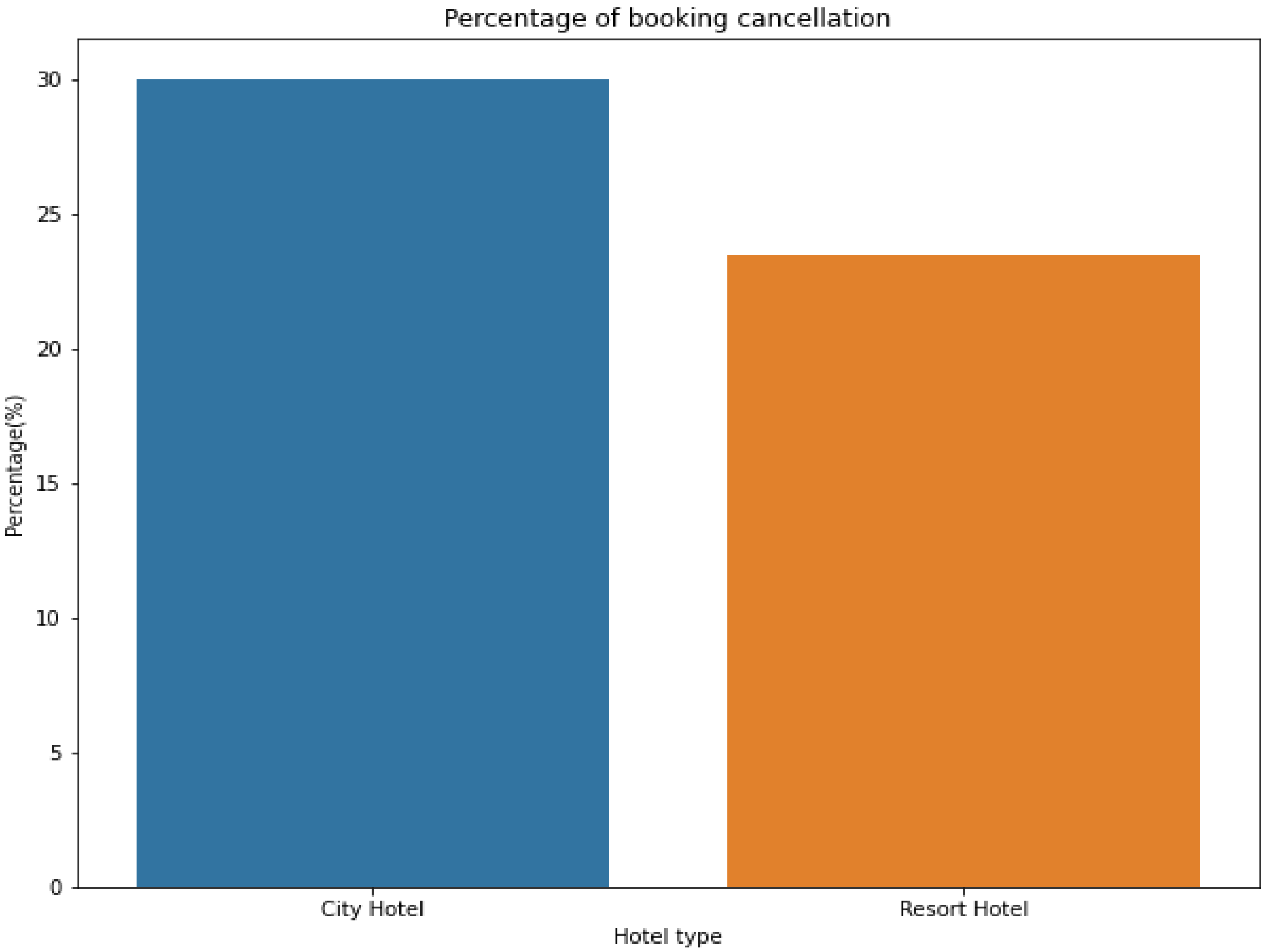
2.Which hotel type has the more lead time?



Conclusions:

Resort hotels has slightly high avg lead time. That means customers plan their trips very early.

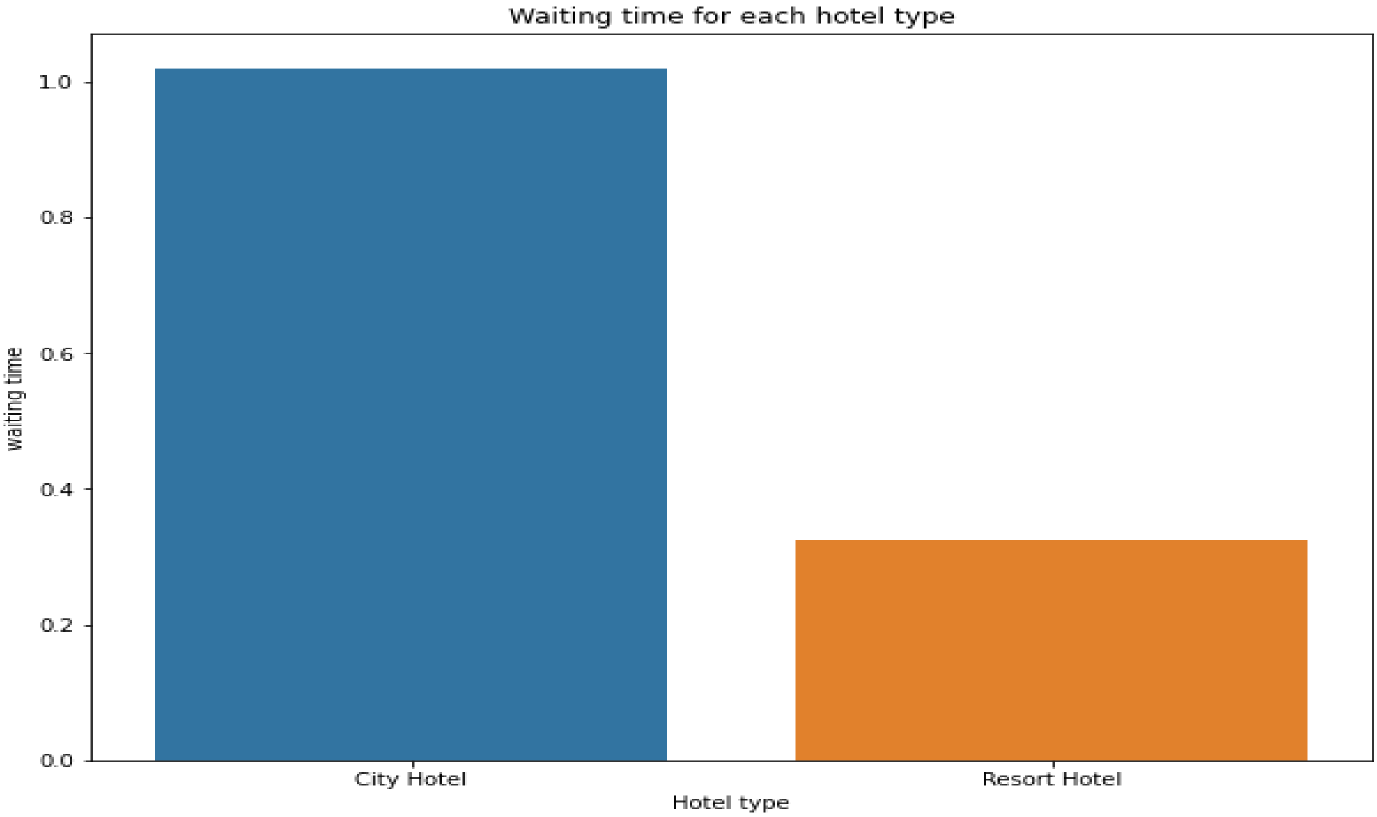
3.Which hotel has highest percentage of booking cancellation?



Conclusions:

City hotel has more booking cancellation than resort hotel.

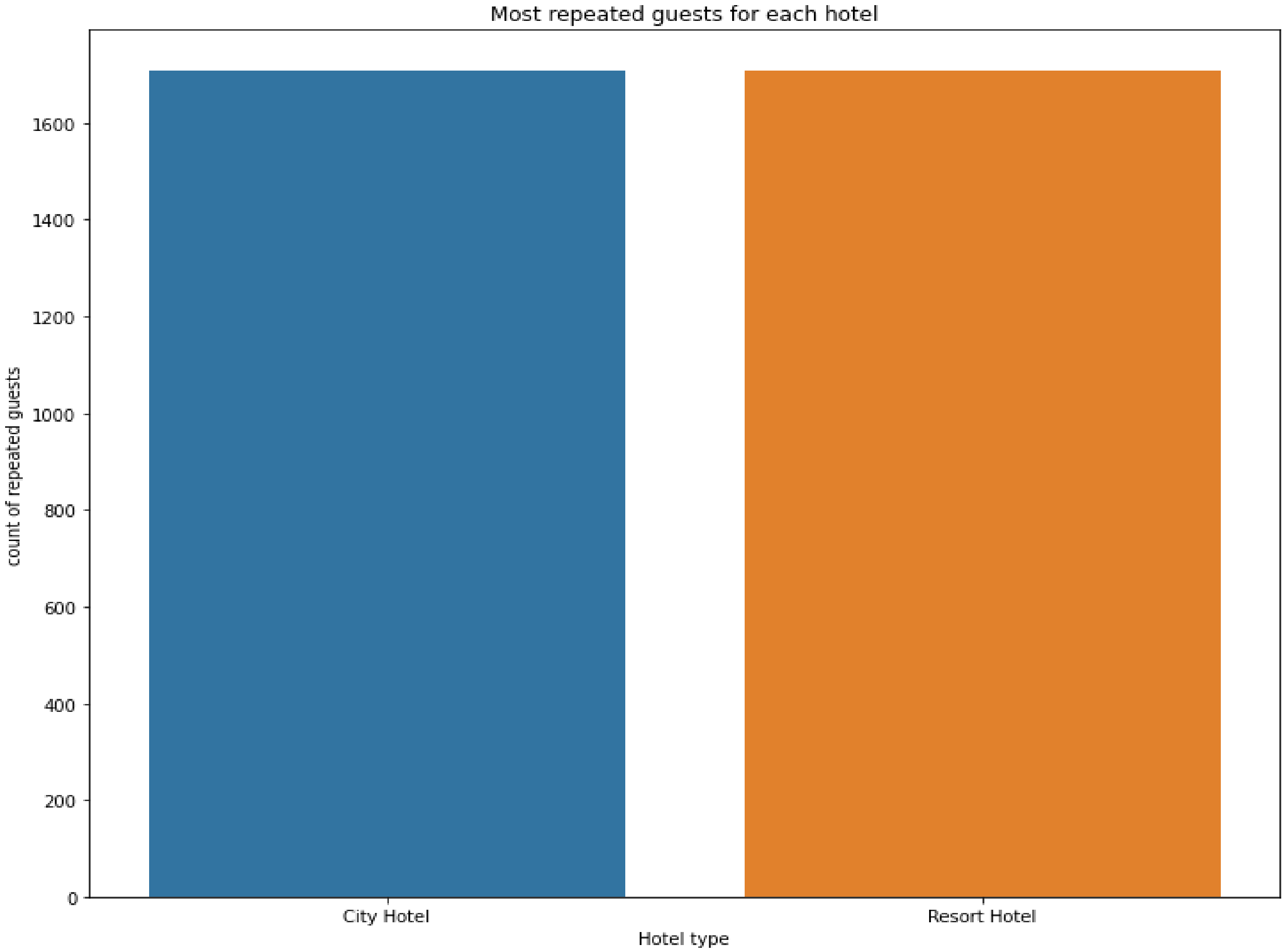
4.Which hotel has longer waiting time?



Conclusions:

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

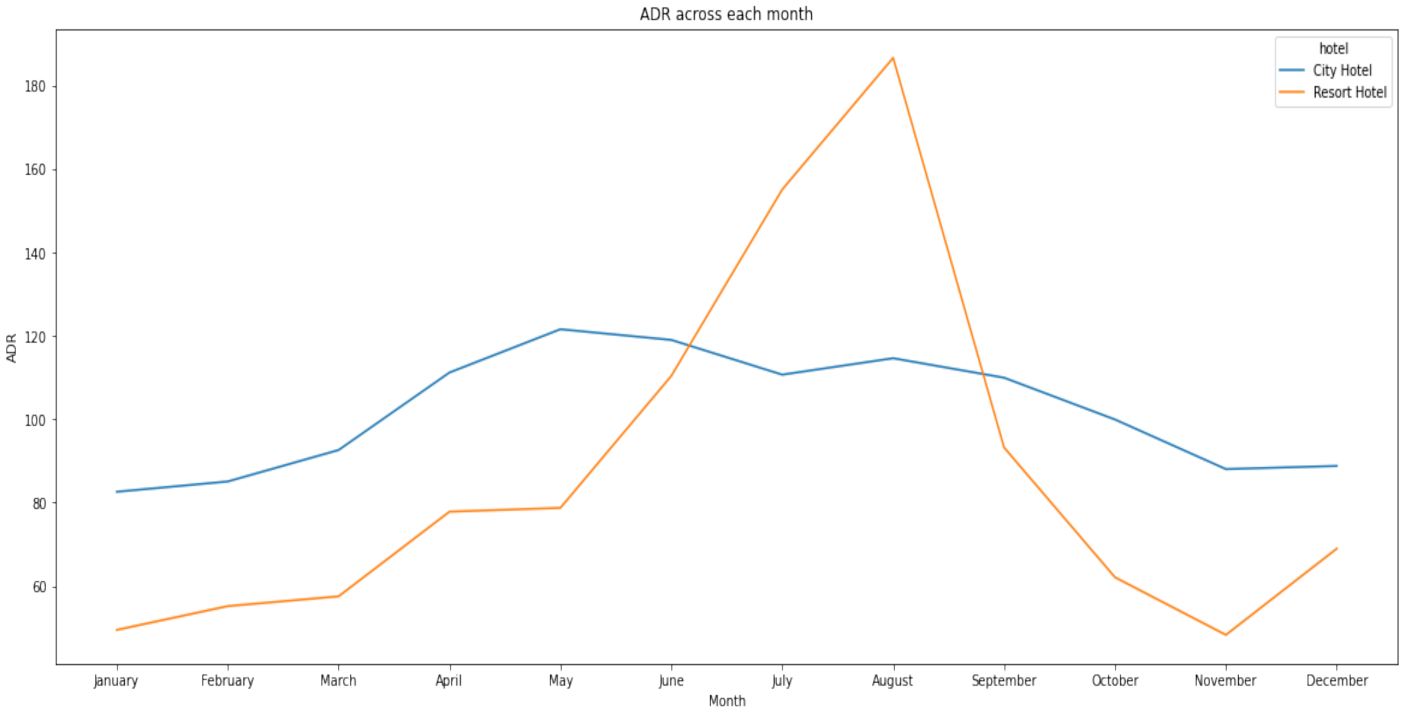
5.Which Hotels has the most repeated guests?



Conclusions:

It is almost similar for both hotels.

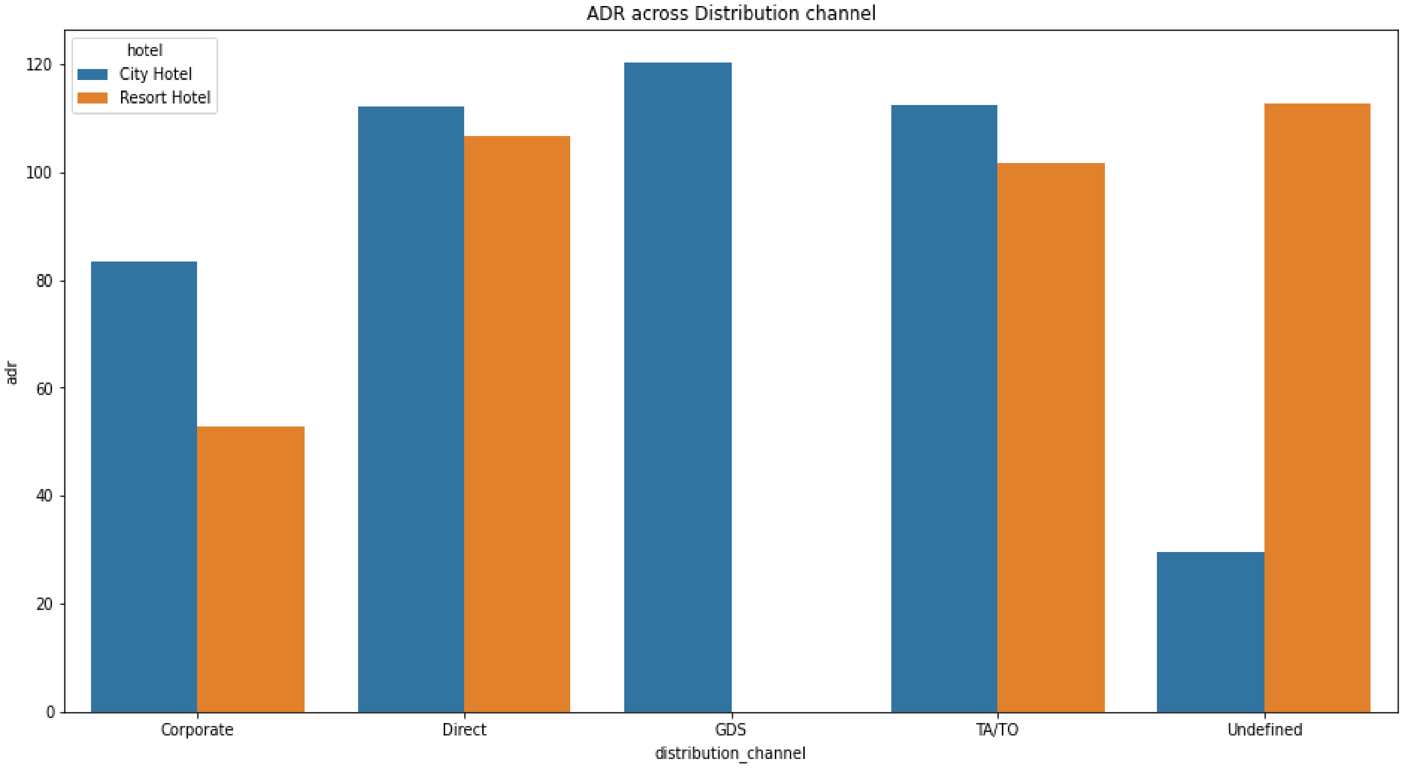
6.Which month has the highest ADR?



Conclusions:

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

7.Which distribution channel contributed more to adr in order to increase the the income?



Observation:

1. Corporate- These are corporate hotel booing companies which makes bookings possible.
2. GDS-A GDS is a worldwide conduit between travel bookers and suppliers, such as hotels and other accommodation providers. It communicates live product, price and availability data to travel agents and online booking engines, and allows for automated transactions.
3. Direct- means that bookings are directly made with the respective hotels
4. TA/TO- means that booings are made through travel agents or travel operators.
5. Undefined- Bookings are undefined. may be customers made their bookings on arrival.

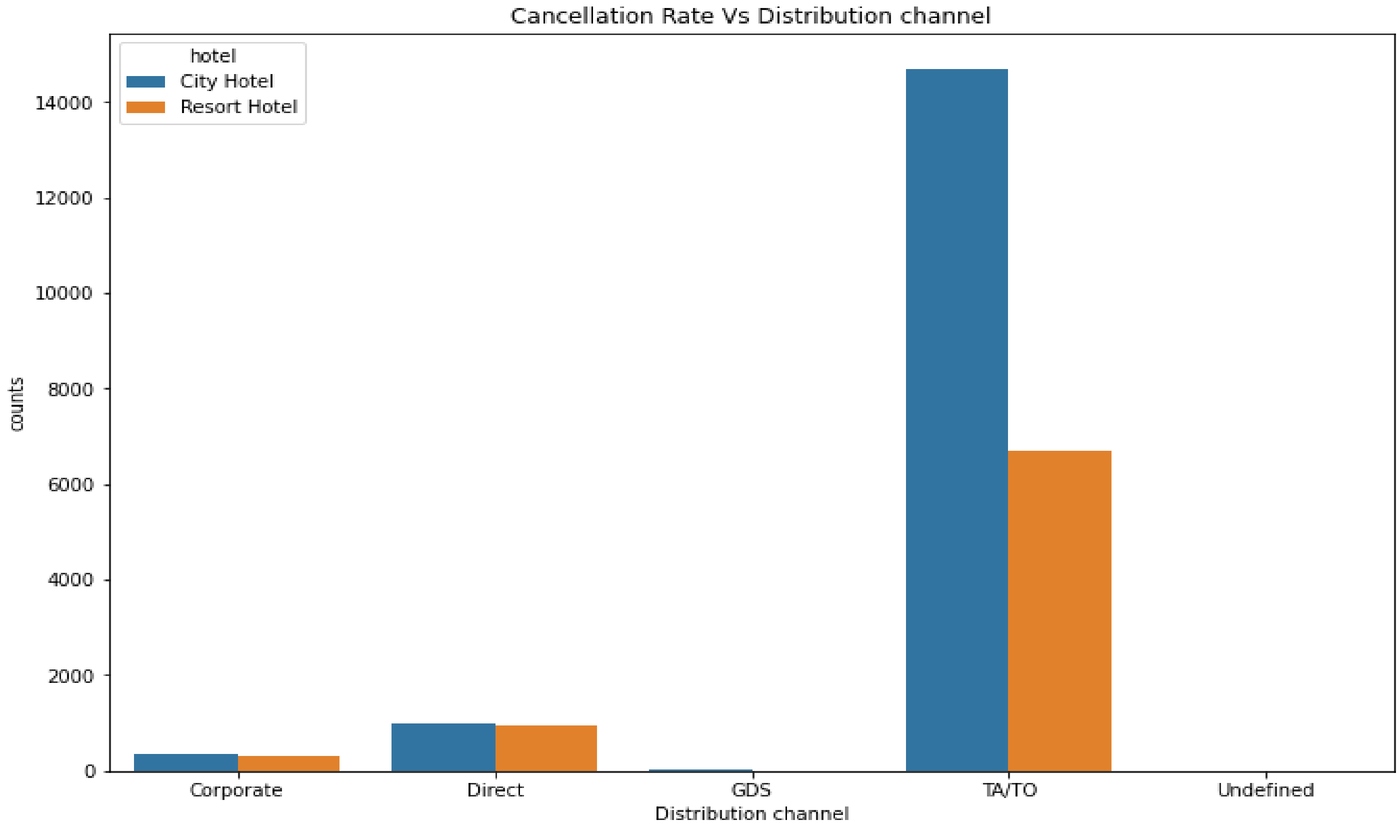
Conclusions: From the plot is clear that -

'Direct' and 'TA/TO' has almost equally contributed in adr in both type of hotels i.e. 'City Hotel' and 'Resort Hotel'.

GDS has highly contributed in adr in 'City Hotel' type.

GDS needs to increase Resort Hotel bookings.

8. Which distribution channel has the higest cancellation rate?

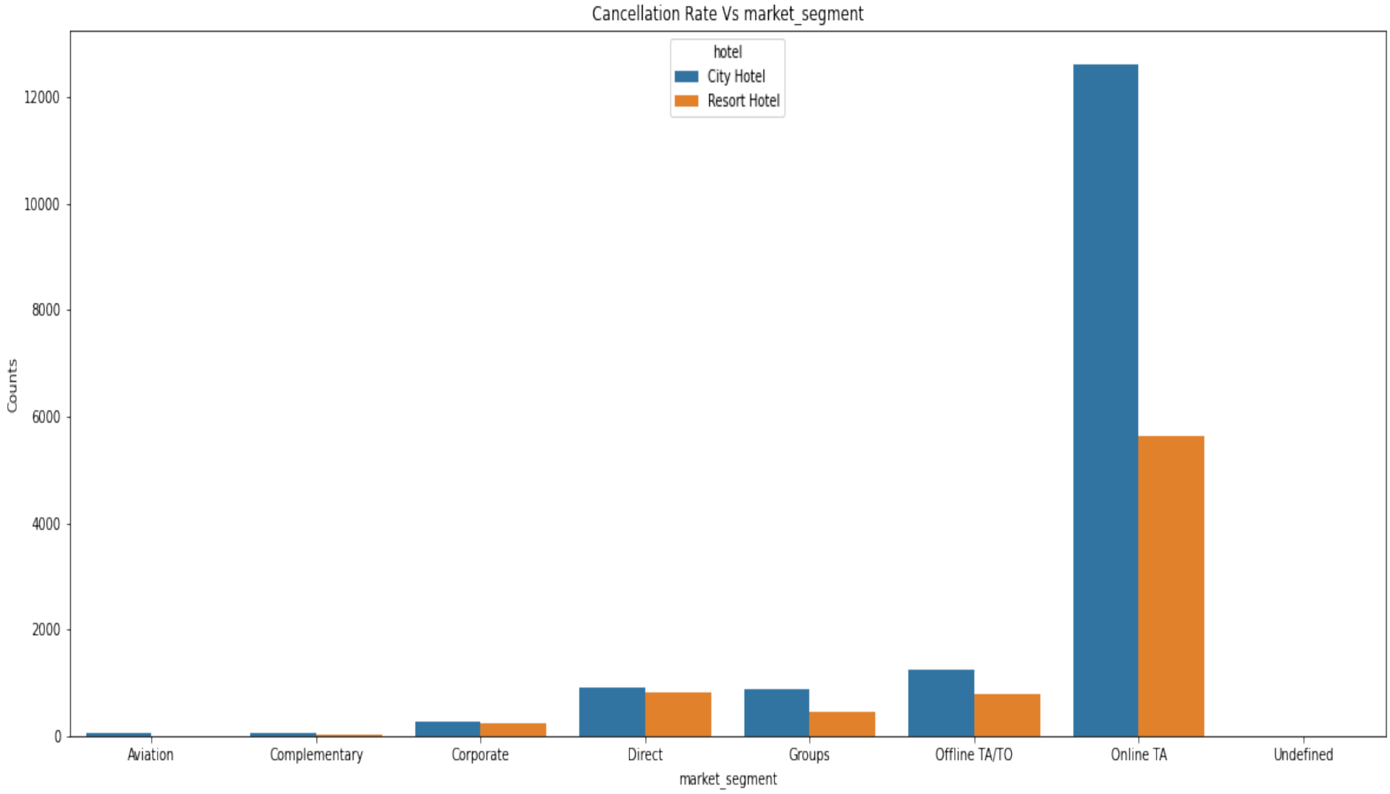


Conclusions:

1. In "TA/TO", City hotels has the high cancellation rate compared to resort hotels.

2. In "direct" both the hotels has almost same cancellation rate.

9.Which Market Segment has the higest cancellation rate?

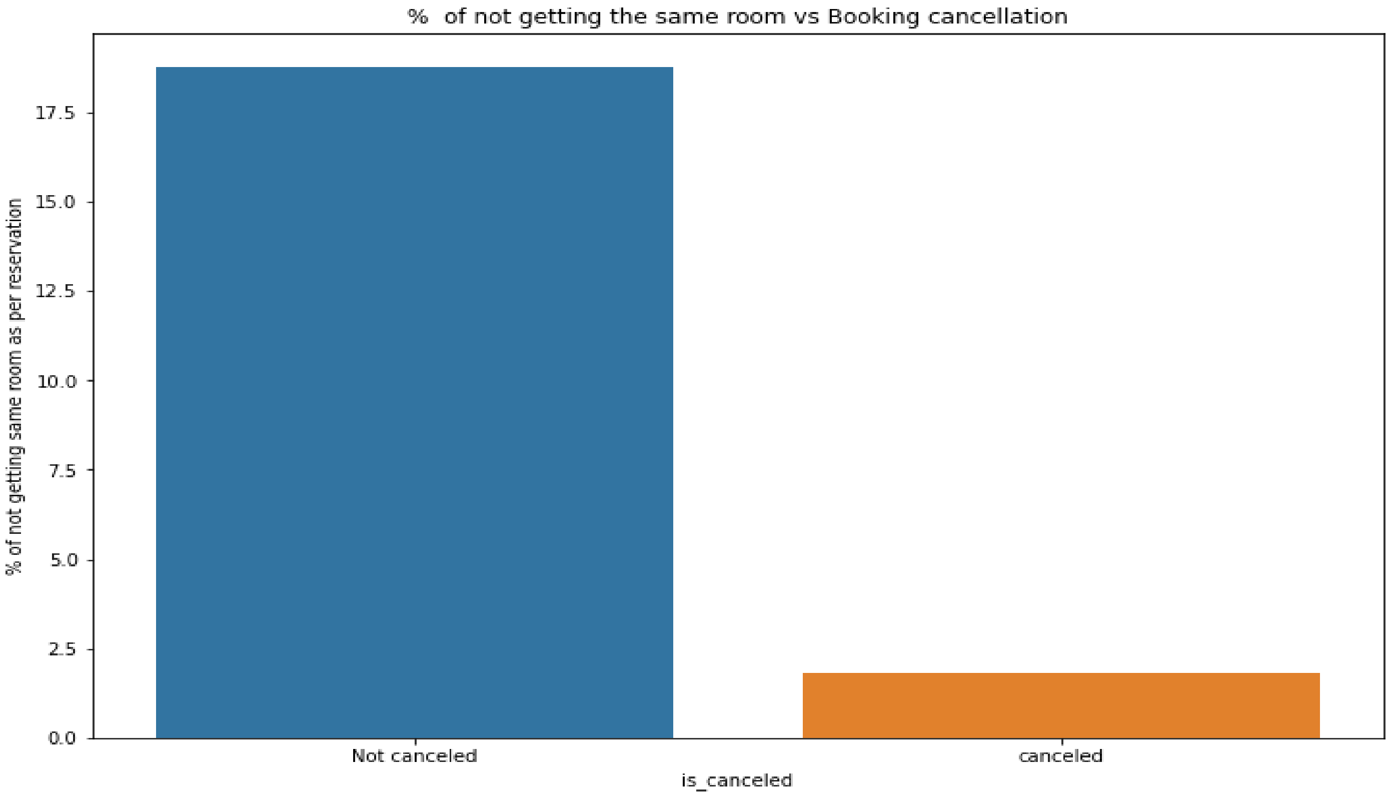


Conclusions:

1. 'Online T/A' has the highest cancellation in both type of cities

2. In order to reduce the booking cancellations hotels need to set the refundable/ no refundable and deposit policies policies

10. Does the guests alloted with the same room type which was reserved by them?

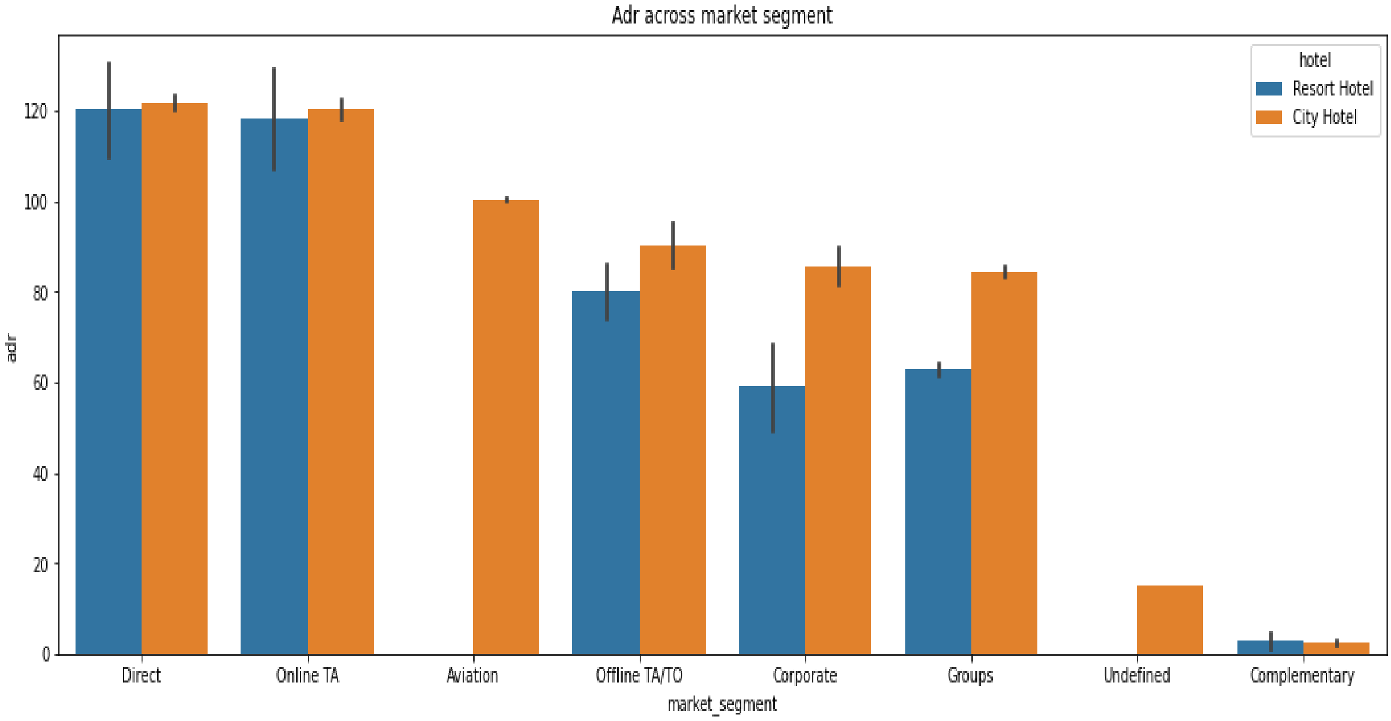


Conclusions:

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.

11. ADR across different market segment.

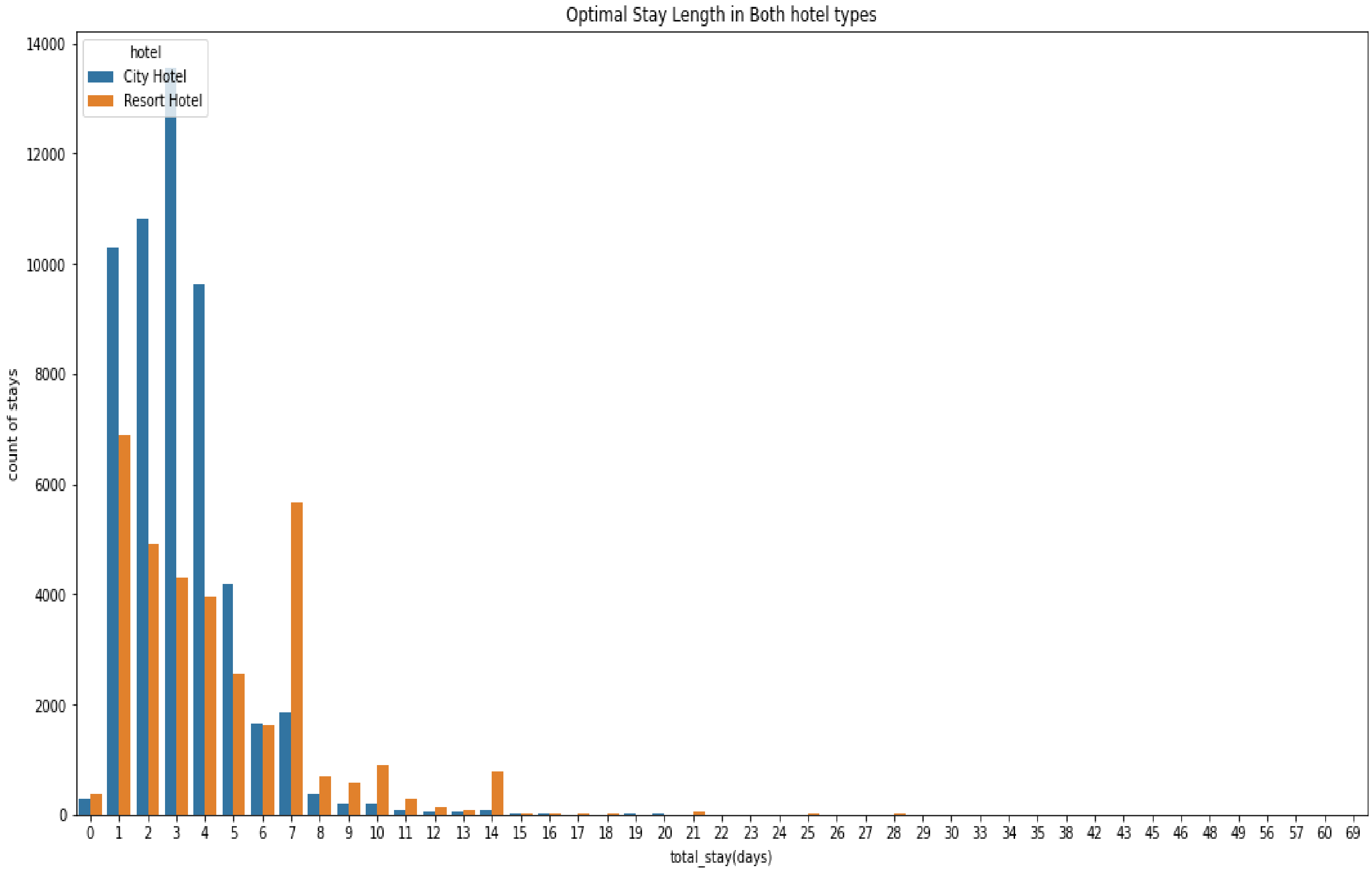
```
Text(0.5, 1.0, 'Adr across market segment')
```



Conclusions:

1. 'Direct' and 'Online TA' are contributing the most in both types of hotels.
2. Aviation segment should focus on increasing the bookings of 'City Hotel'

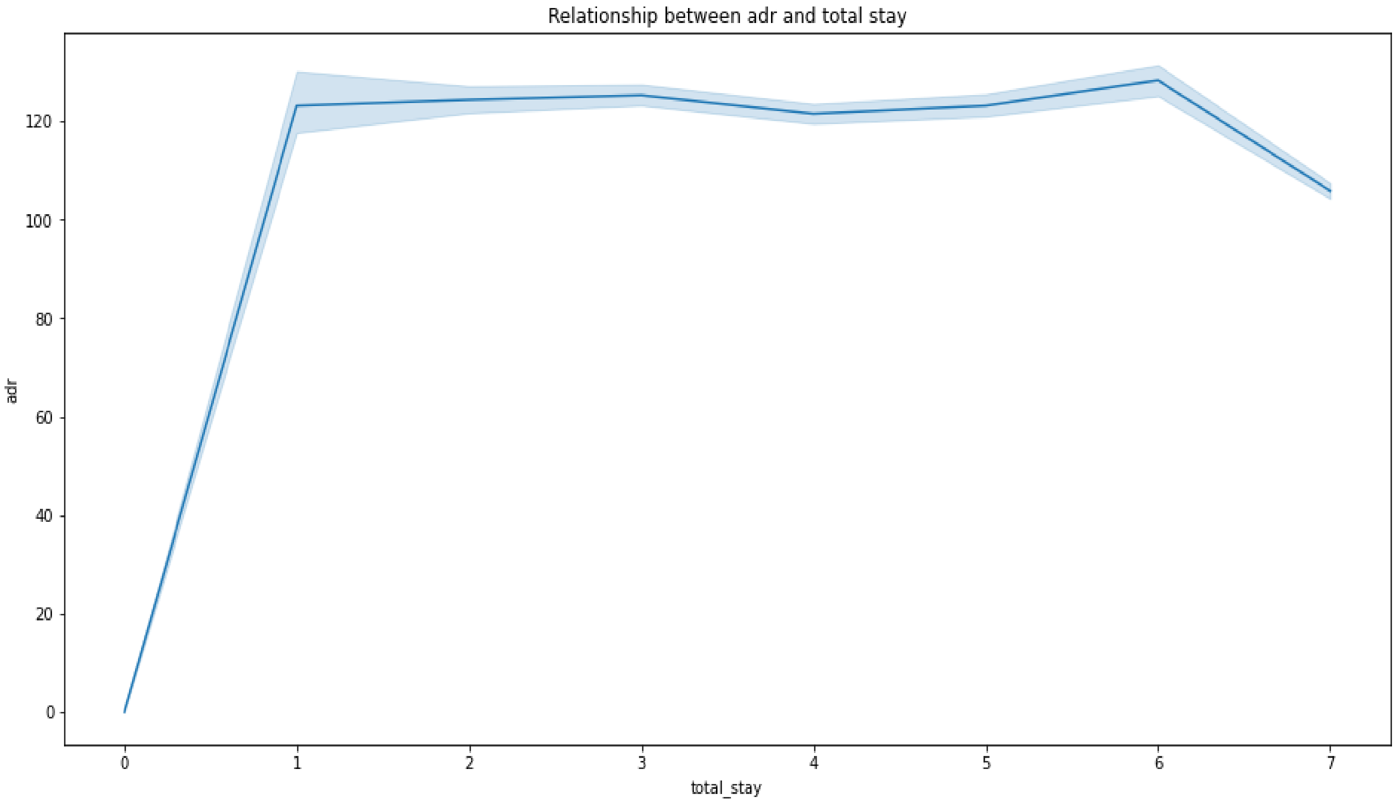
12.What is the Optimal stay length in both types of hotels ?



Conclusions:

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

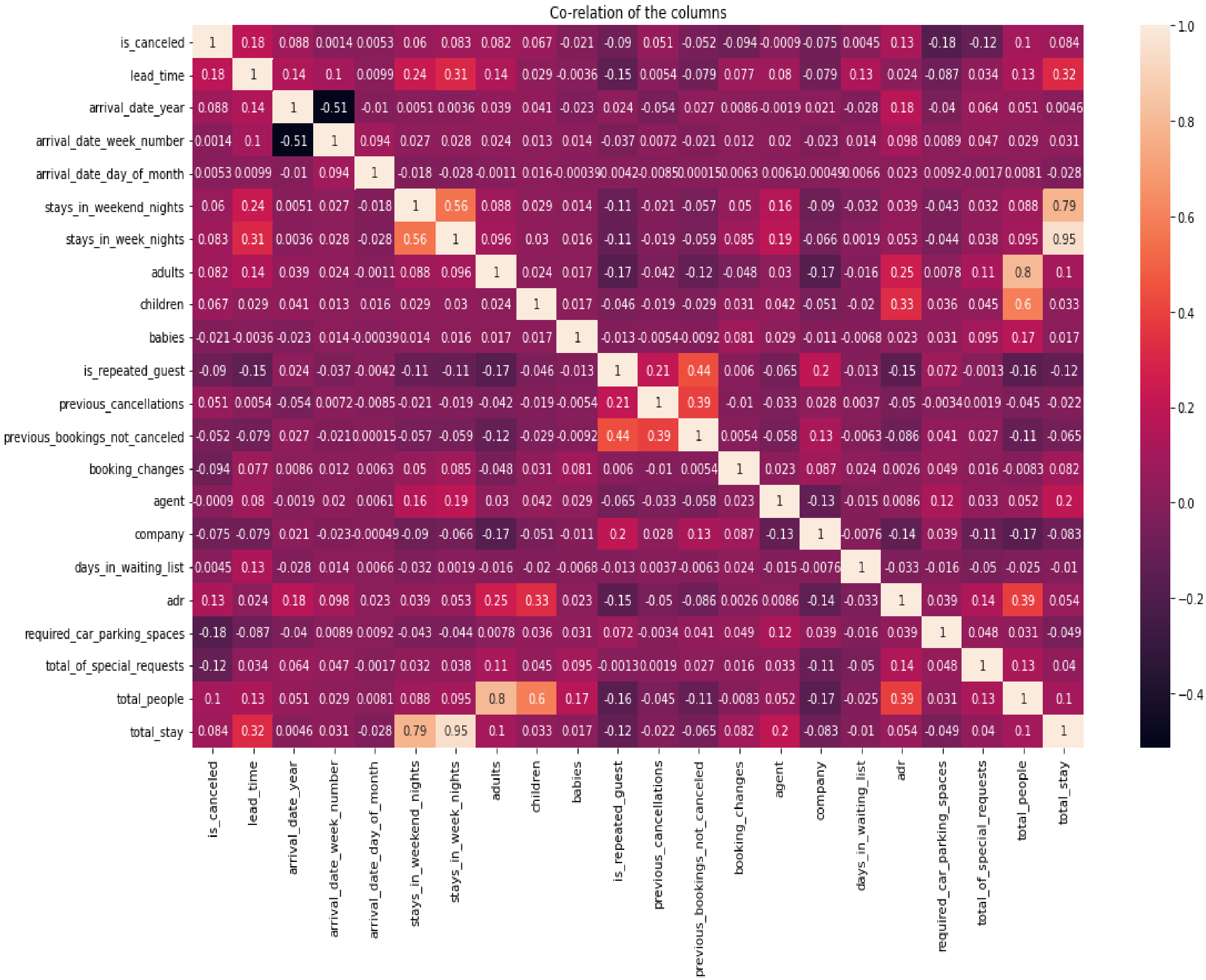
13. Relationship between ADR and total stay.



Conclusions:

As the total stay increases the adr also increases.

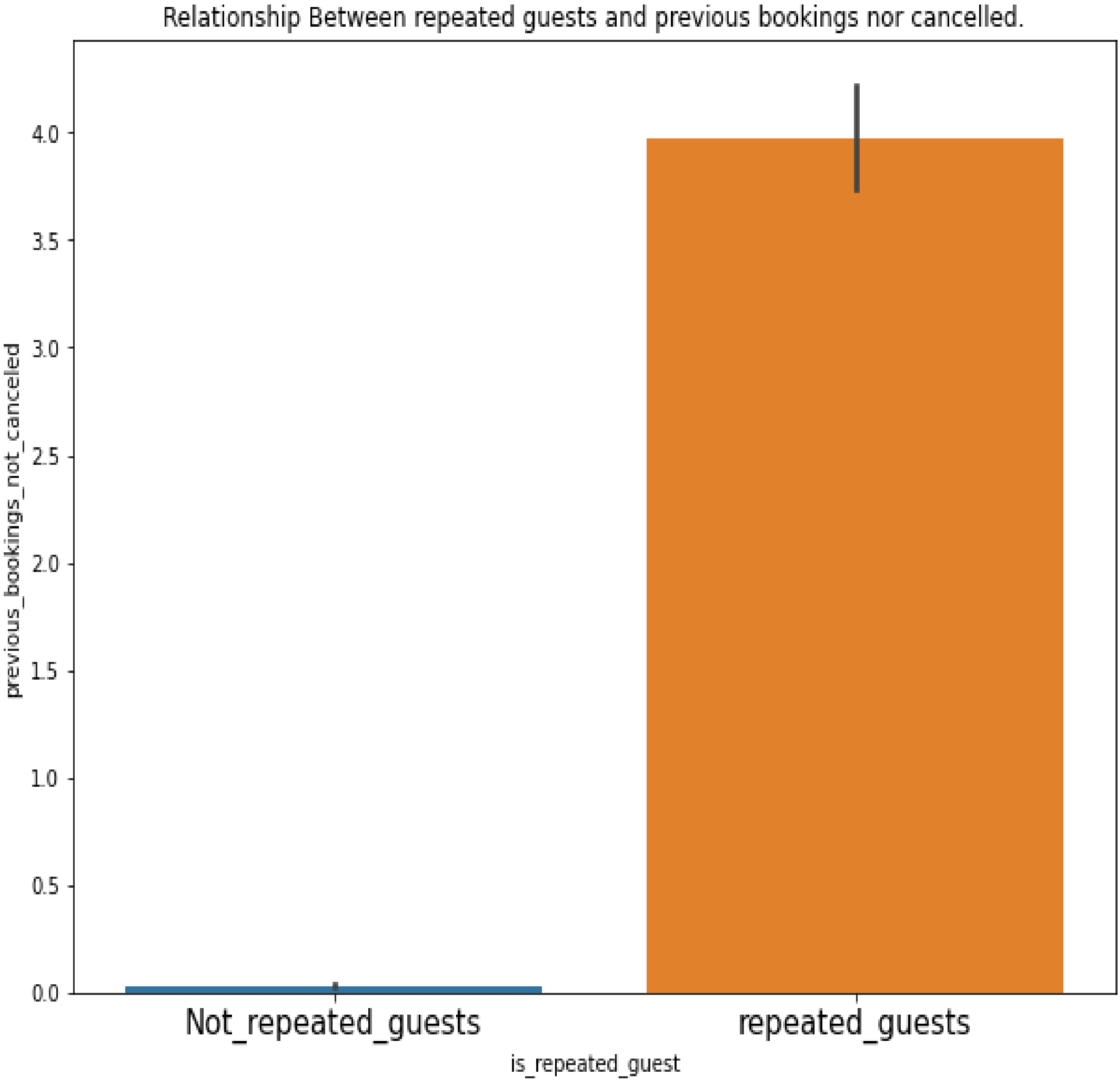
14. Correlation of the columns



Conclusions:

1. 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.
2. lead_time and total_stay is positively corelated.That means more is the stay of cutsomer more will be the lead time.
3. adults,childrens and babies are corelated to each other. That means more the people more will be adr.
4. is_repeated guest and previous bookings not canceled has strong corelation. may be repeated guests are not more likely to cancel their bookings.

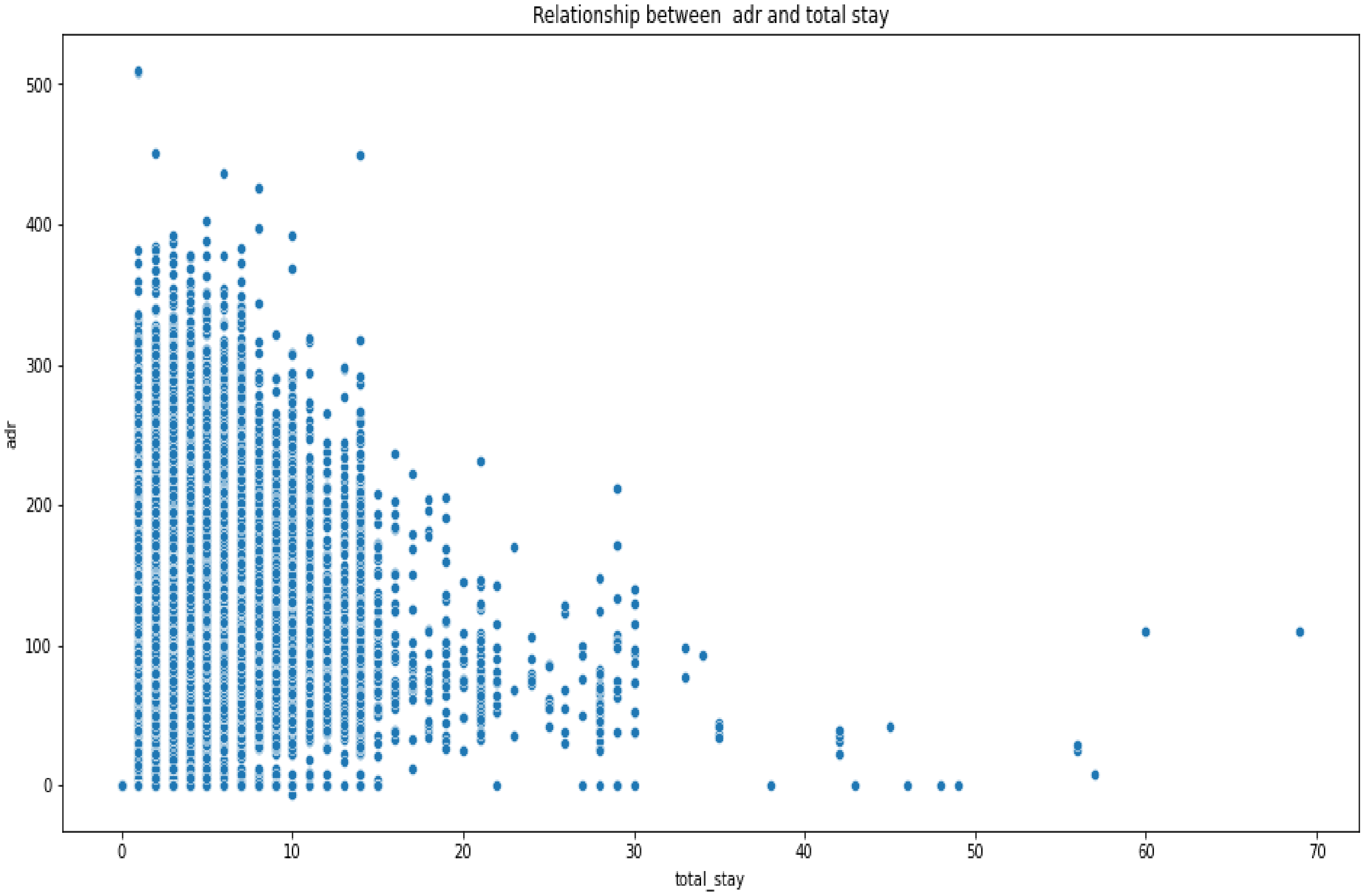
15. Relationship between the repeated guests and previous bookings not canceled?



Conclusions:

Not Repeated guests are more likely to cancel their bookings.

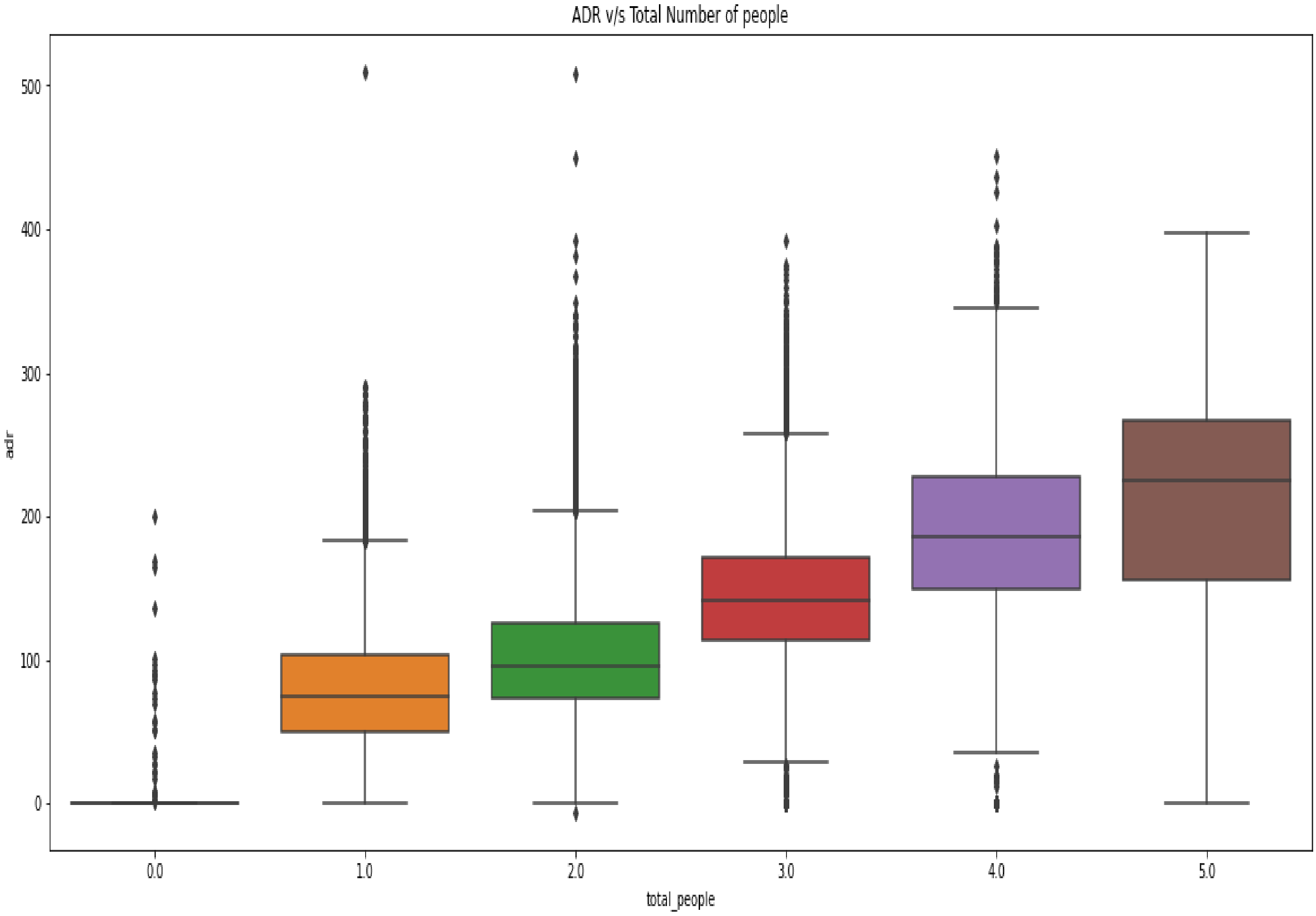
16.Relationship between adr and total stay.



Conclusions:

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

17. ADR relationship with total number of people



Conclusions:

As the total number of people increases adr also increases.

Thus adr and total people are directly proportional to each other.

Final Conclusions:

- ✓ City hotels are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
- ✓ 27.5 % bookings were got cancelled out of all the bookings.
- ✓ Only 3.9 % people were revisited the hotels. Rest 96.1 % were new guests. Thus retention rate is low.
- ✓ The percentage of 0 changes made in the booking was more than 82 %. Percentage of Single changes made was about 10%.
- ✓ Most of the customers (91.6%) do not require car parking spaces.
- ✓ 79.1 % bookings were made through TA/TO (travel agents/Tour operators).
- ✓ BB(Bed & Breakfast) is the most preferred type of meal by the guests.
- ✓ Maximum number of guests were from Portugal, i.e. more than 25000 guests.
- ✓ Most of the bookings for City hotels and Resort hotel were happened in 2016.
- ✓ Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
- ✓ Booking cancellation rate is high for City hotels which almost 30 %.
- ✓ Average lead time for resort hotel is high.
- ✓ Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
- ✓ Resort hotels have the most repeated guests.
- ✓ Optimal stay in both the type hotel is less than 7 days. Usually people stay for a week.
- ✓ Almost 19 % people did not cancel their bookings even after not getting the same room which they reserved while booking hotel. Only 2.5 % people cancelled the booking.

Thank You