Capstone Project-1

EDA On Hotel Booking Analysis

BY

Mohmmad Owes khan

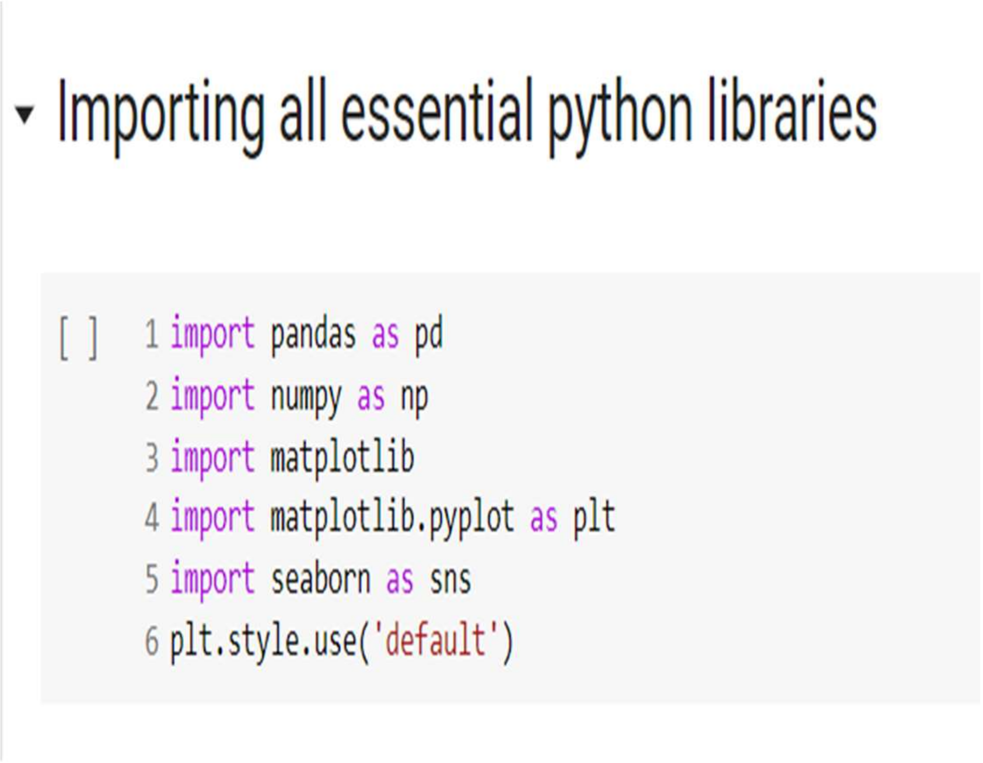
(Cohort Zurich)

Points to Discuss:

- Data collection
- Data summary
- Data cleaning
- EDA
- Univariate analysis
- Some important questions related to hotel industry
- Bivariate and Multivariate Analysis
- Final Ratings obtained from data
- Correlation heatmap
- Suggestions to hotel industry from analysing and visualization of data.
- Conclusion

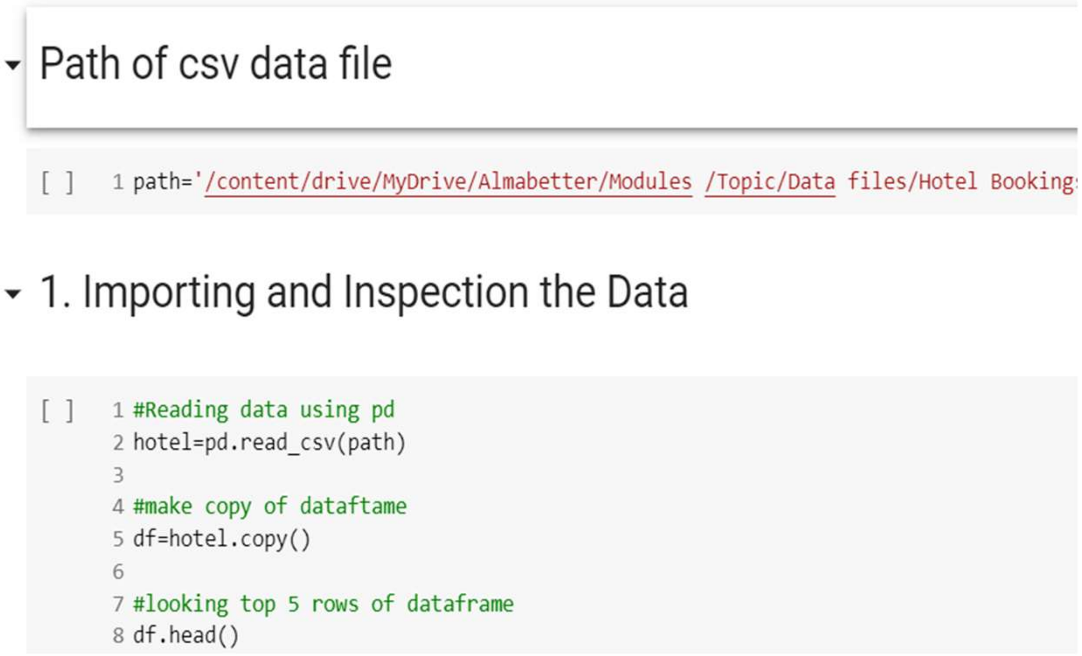
Data collection:-

Importing the data in the python and converting data into more convenient form with the help of different libraries in the python for better analyzing and visualization.



Importing all essential python libraries

```
[ ]
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('default')
```



Path of csv data file

```
[ ]
path='/content/drive/MyDrive/Almabetter/Modules /Topic/Data files/Hotel Booking
```

1. Importing and Inspection the Data

```
[ ]
#Reading data using pd
hotel=pd.read_csv(path)

#make copy of dataftame
df=hotel.copy()

#looking top 5 rows of dataframe
df.head()
```

Data Summary

Given data set has 119390 rows and 32 columns of variables which are crucial for hotel bookings.Let’s understand the few columns others are self explanatory.

hotel: The category of hotels, which are two resort hotel and city hotel.

is_cancelled : The value of column show the cancellation type. If the booking was cancelled or not. Values[0,1], where 0 indicates not cancelled.

lead_time : The time between reservation and actual arrival.

stayed_in_weekend_nights: The number of weekend nights stay per reservation

stayed_in_weekday_nights: The number of weekday nights stay per reservation.

meal: Meal preferences per reservation.[BB,FB,HB,SC,Undefined]

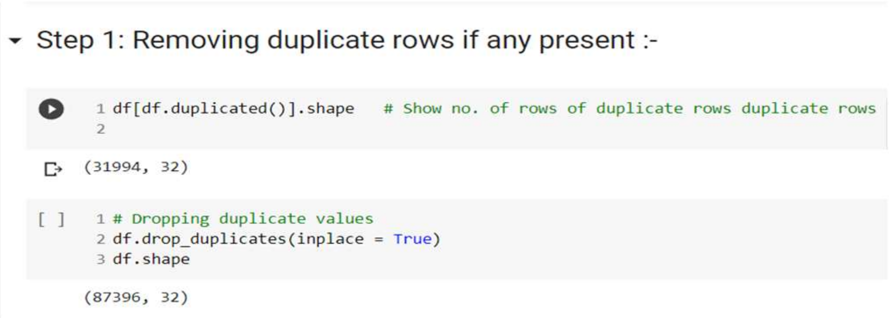
Country: The origin country of guest.

Cleaning the data:-

Cleaning data is important step before EDA as it will remove the unwanted data that can affect the outcome of our EDA.

While cleaning data we will perform following steps:

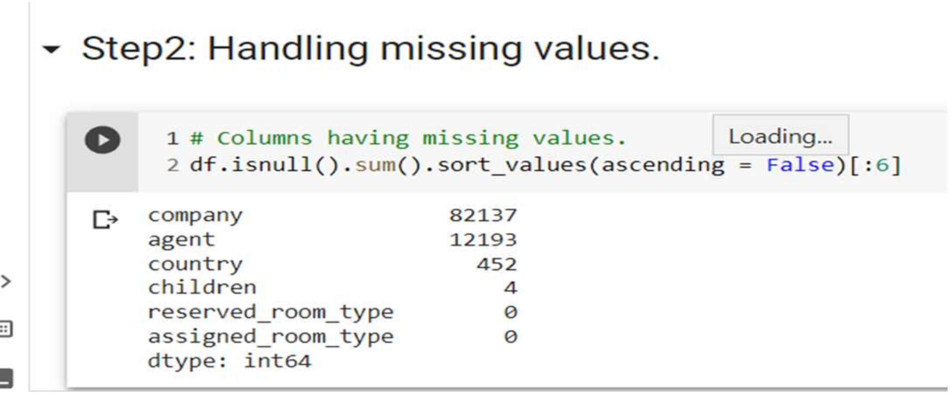
1) Deleting duplicate rows
2) Handling missing values.
3) Convert columns to appropriate datatypes.
4) Adding important columns



Step 1: Removing duplicate rows if any present :-

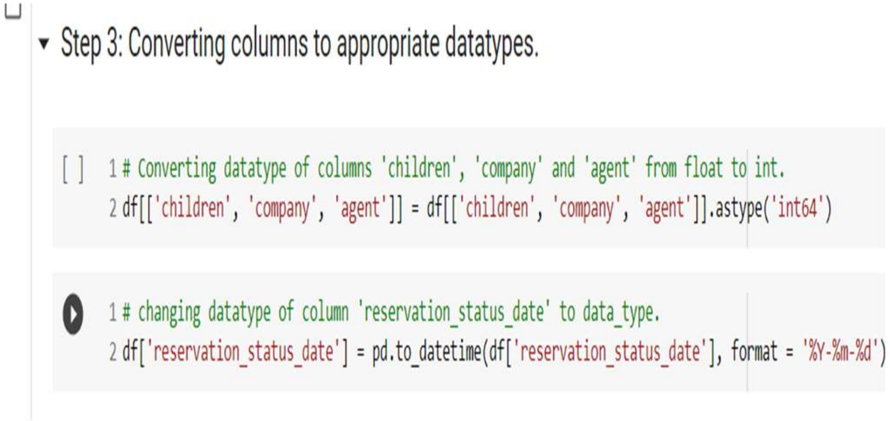
```
df[df.duplicated()].shape    # Show no. of rows of duplicate rows duplicate rows

(31994, 32)
```

```
# Dropping duplicate values
df.drop_duplicates(inplace = True)
df.shape

(87396, 32)
```



Step2: Handling missing values.

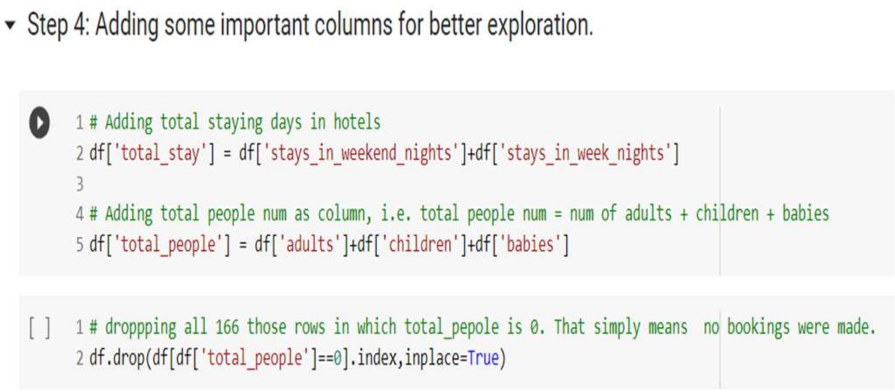
```
# Columns having missing values.
df.isnull().sum().sort_values(ascending = False)[:6]

company               82137
agent                 12193
country                 452
children                  4
reserved_room_type        0
assigned_room_type        0
dtype: int64
```



Step 3: Converting columns to appropriate datatypes.

```
# Converting datatype of columns 'children', 'company' and 'agent' from float to int.
df[['children', 'company', 'agent']] = df[['children', 'company', 'agent']].astype('int64')
```

```
# changing datatype of column 'reservation_status_date' to data_type.
df['reservation_status_date'] = pd.to_datetime(df['reservation_status_date'], format = '%Y-%m-%d')
```



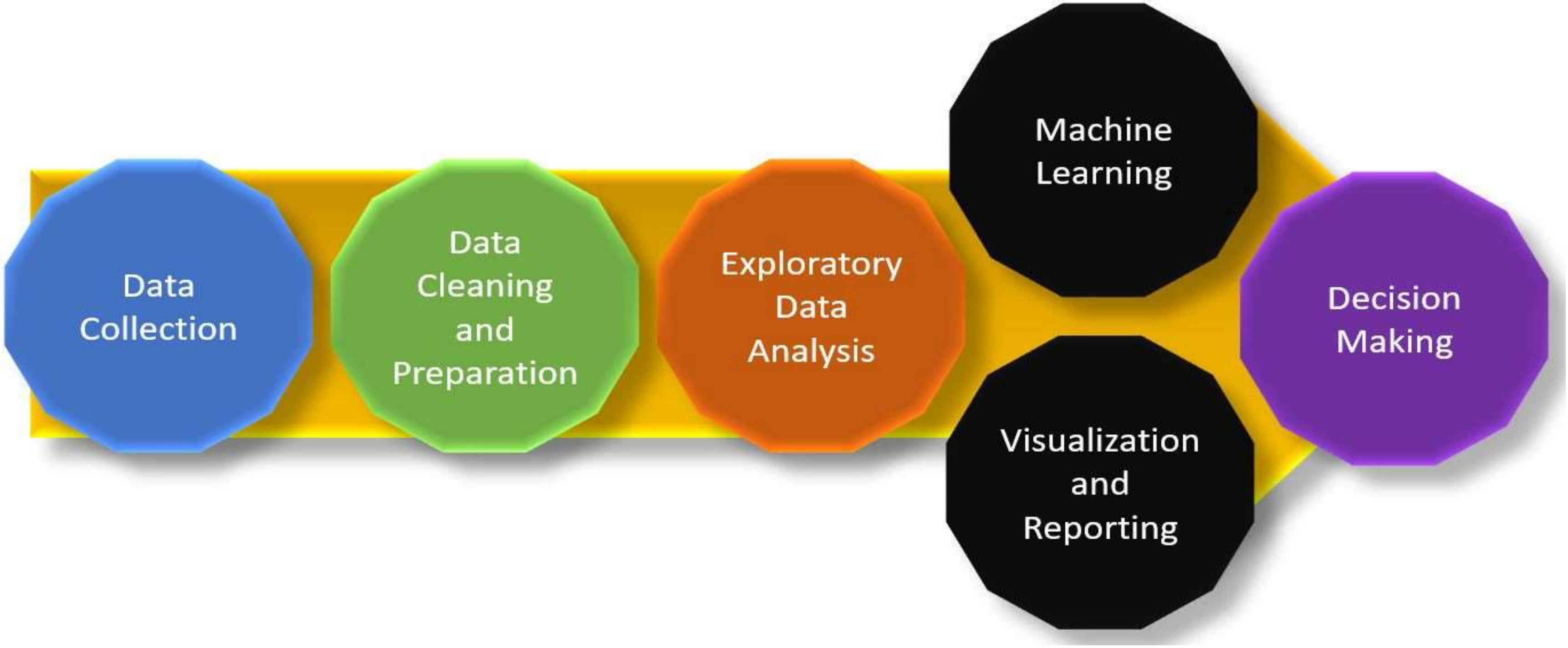
Step 4: Adding some important columns for better exploration.

```
# Adding total staying days in hotels
df['total_stay'] = df['stays_in_weekend_nights']+df['stays_in_week_nights']

# Adding total people num as column, i.e. total people num = num of adults + children + babies
df['total_people'] = df['adults']+df['children']+df['babies']
```

```
# droppping all 166 those rows in which total_pepole is 0. That simply means  no bookings were made.
df.drop(df[df['total_people']==0].index,inplace=True)
```

EDA:-

- **Exploratory Data Analysis** is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.


Data Collection
Data Cleaning and Preparation
Exploratory Data Analysis
Machine Learning
Visualization and Reporting
Decision Making

Univariate Analysis:-

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable

Let’s find out the answer of some questions with the help of univariate analysis.
we answered following questions:

Some important questions:-

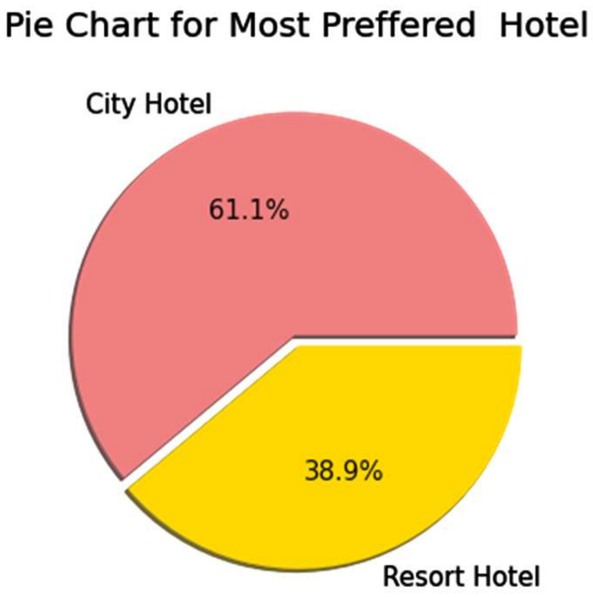
Which type of hotels is mostly preferred by the guest?

What is the Percentage of repeated guests?

What is the percentage of booking cancelation?

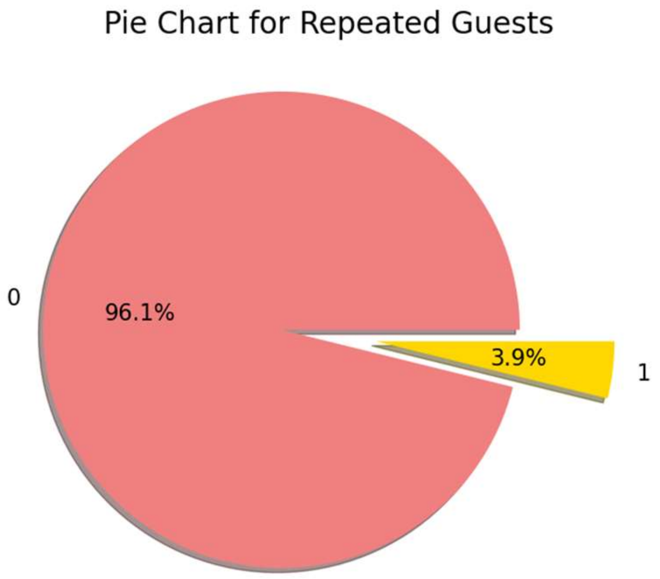
And many more.

Which type of hotels is mostly preferred by the guest?



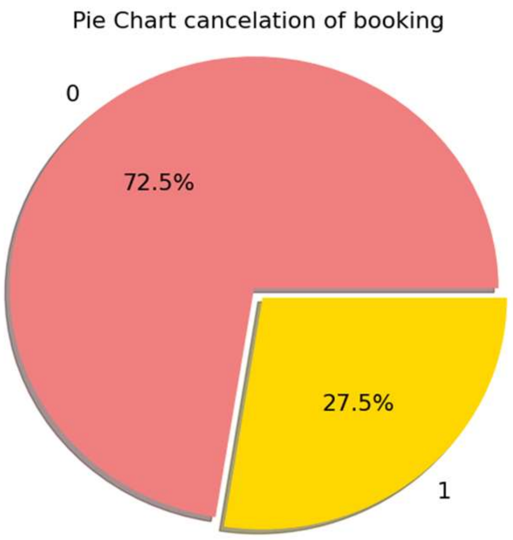
City Hotel is most preferred hotel by guests. Thus city hotels has maximum bookings which is 61.1%.

What is the Percentage of repeated guests?



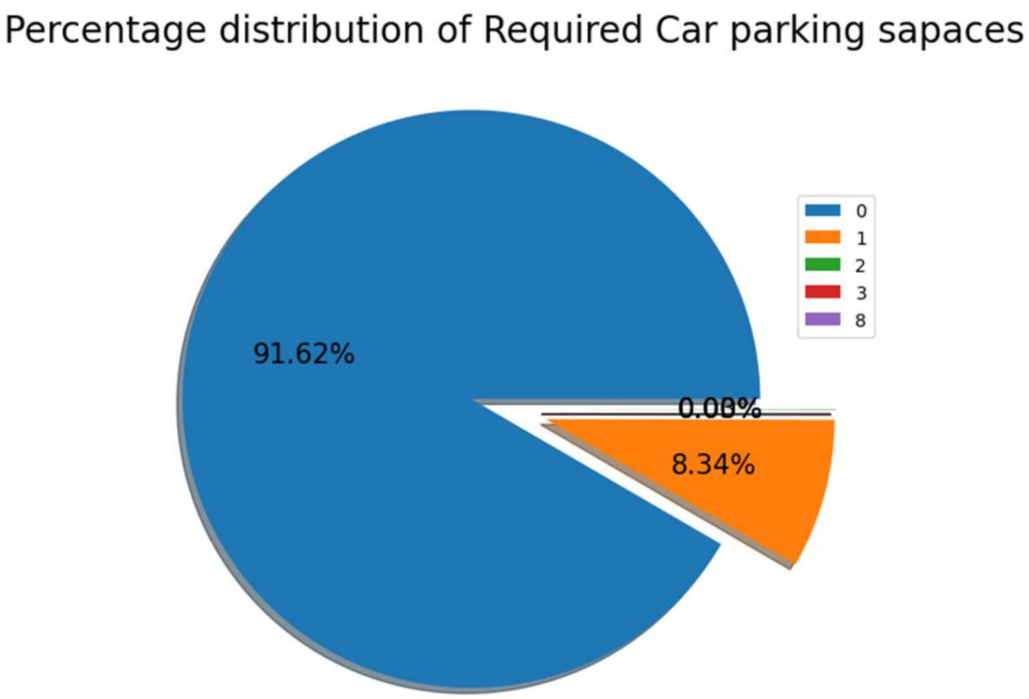
Only 3.9% guests are repeated and remaining all are new guests.

What is the percentage of booking cancelation?



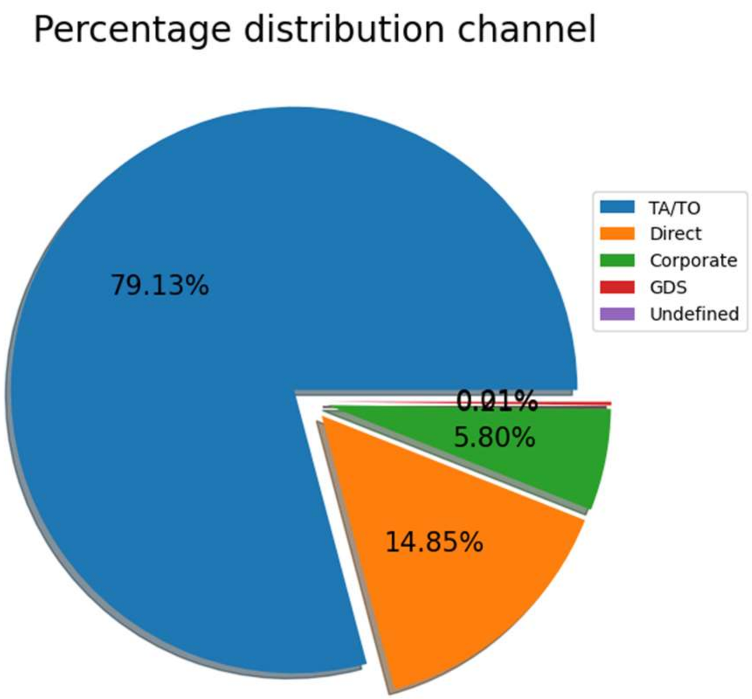
0 means no cancelation of booking ,It is 72.5%.
1 means cancelation of booking, It is 27.5%.

What is the percentage distribution of required car parking spaces?



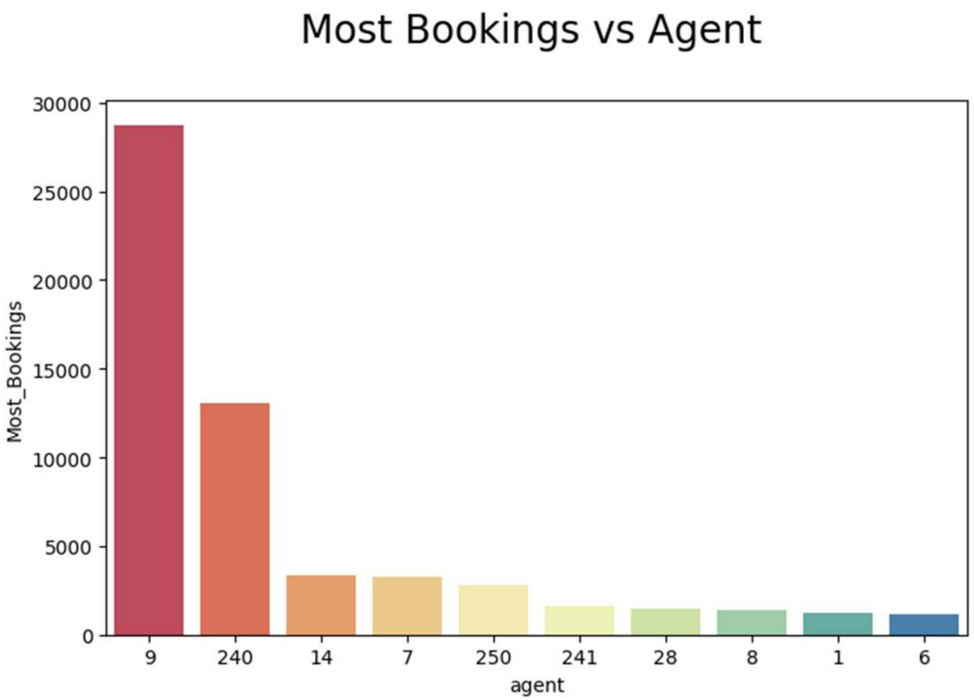
91.6 % guests did not required the parking space. only 8.3 % guests required only 1 parking space.

What is the percentage distribution channel?



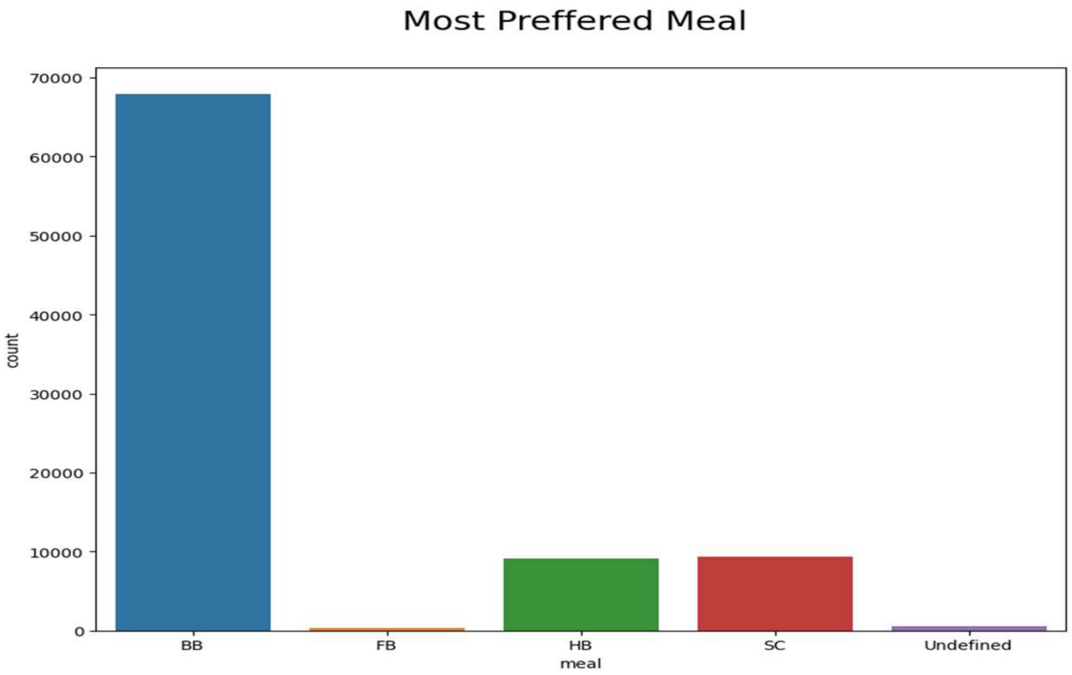
79% distribution channel is i.e. Travel agency or travel operators.

Which Agent made the most bookings?



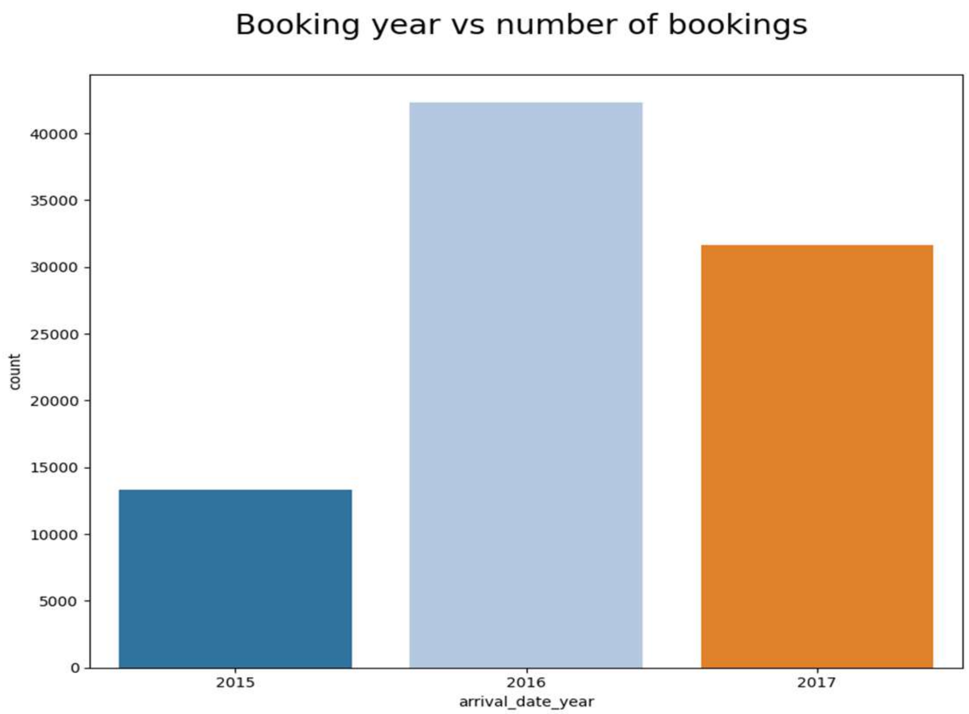
Agent number 9 made the most booking more than 25k.
Then agent number 240 and 14.

Which meal type is most preferred meal of customers?



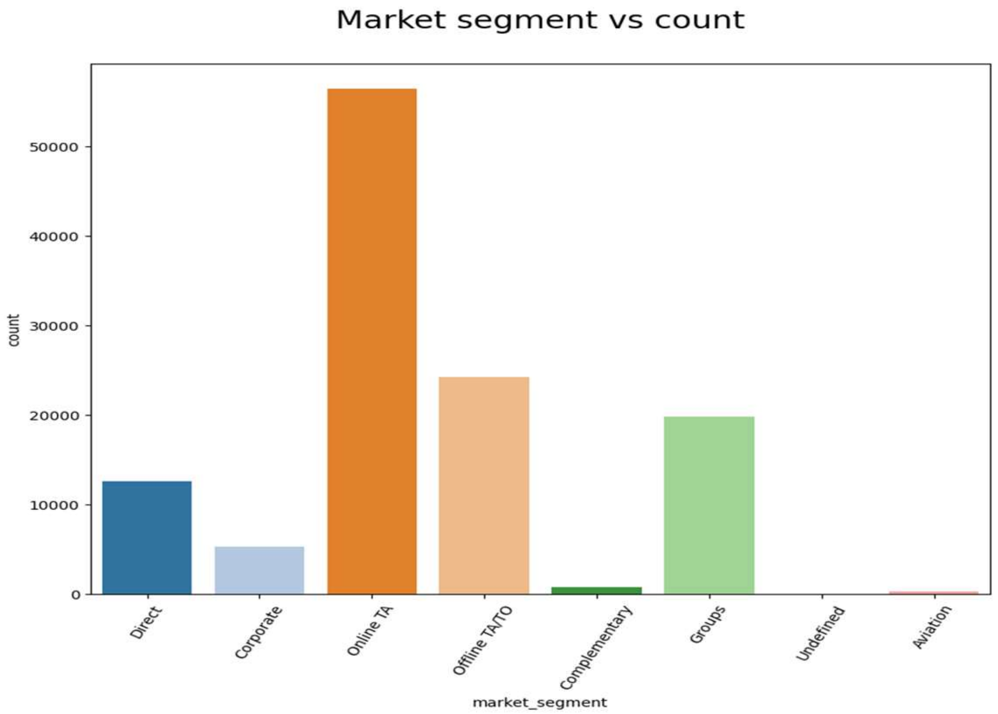
Most preferred meal type is BB (Bed and breakfast).
Types of meal in hotels:
BB - (Bed and Breakfast)
HB- (Half Board)
FB- (Full Board)
SC- (Self Catering)

Which year has the most bookings done?



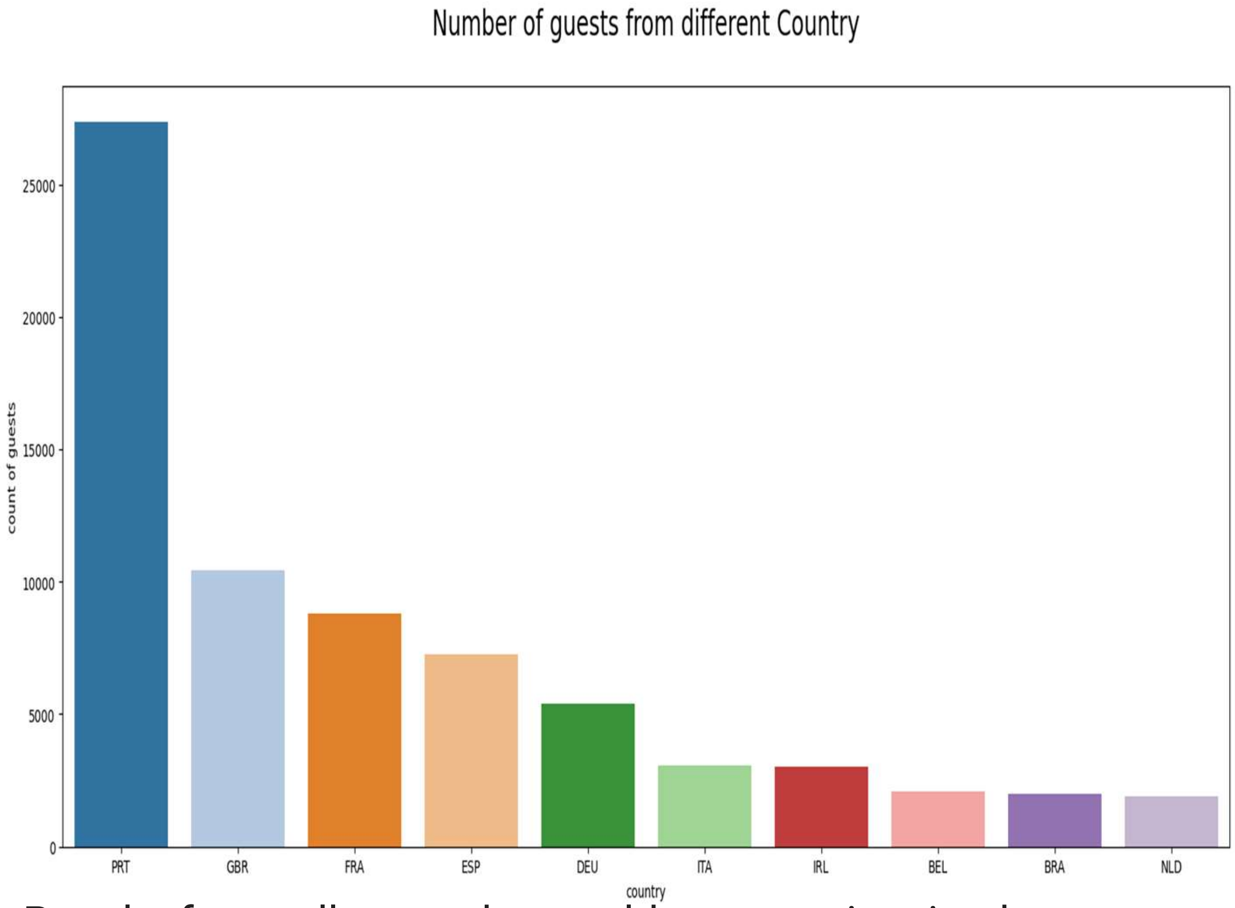
2016 Year has the most bookings 40k
2015 has less than 15k bookings

Which is the most preferred market segment?



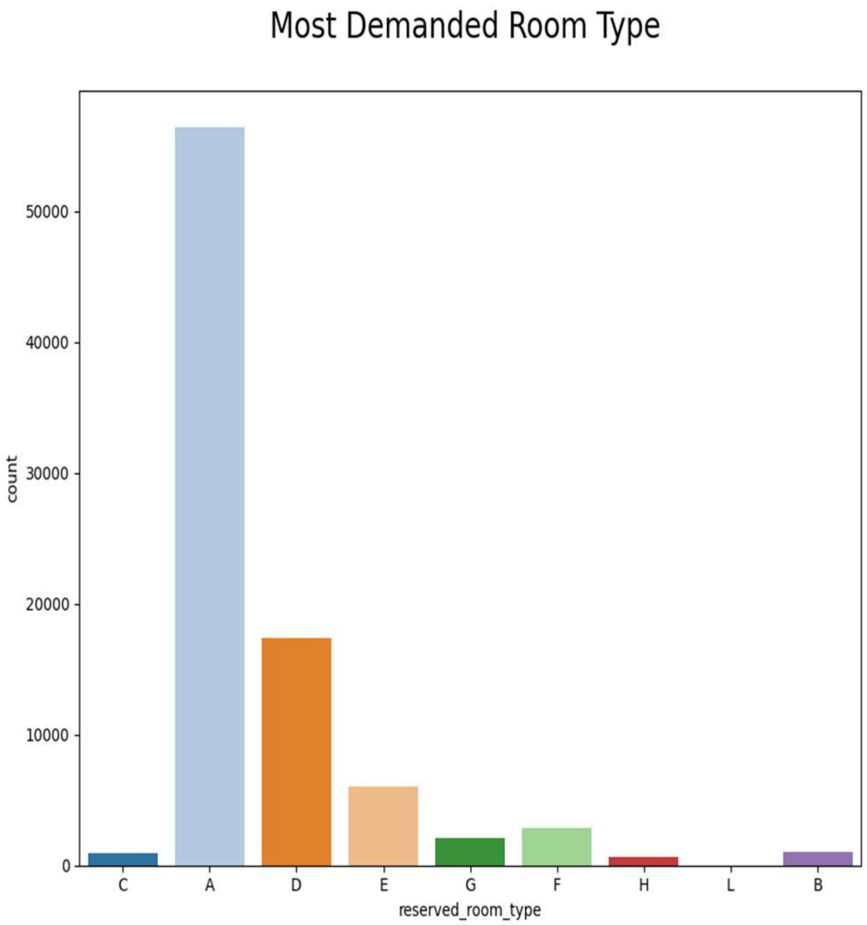
Most of the market segment used Online TA
i.e. online travel agency .

From which country the most guests are coming?



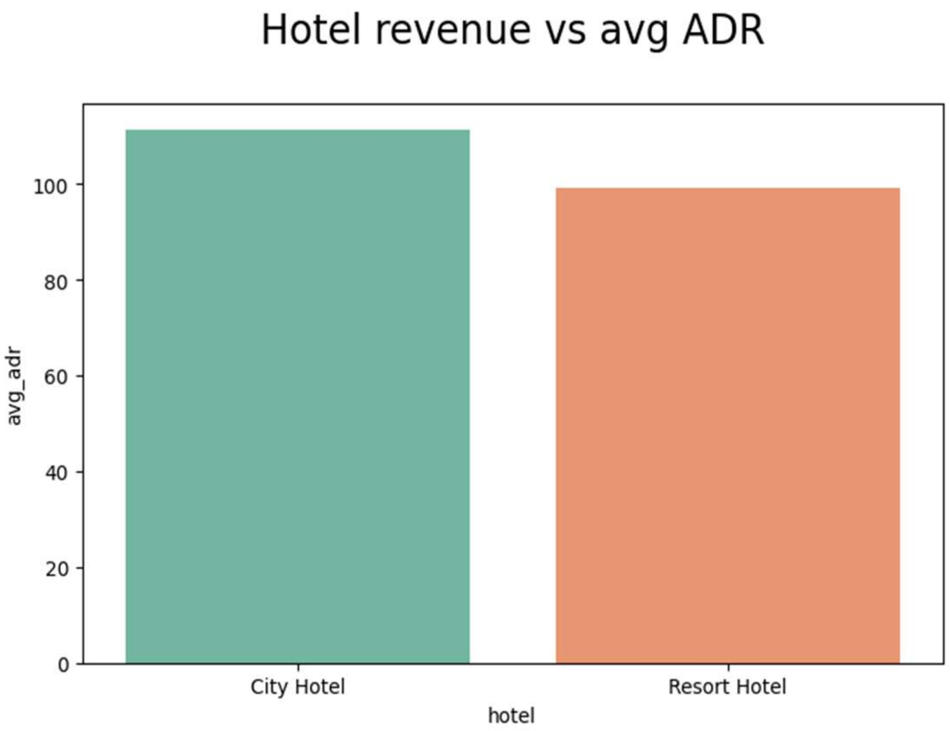
People from all over the world are staying in these two hotels. Most guests are from Portugal and other countries in Europe.

Which room type is in most demand?



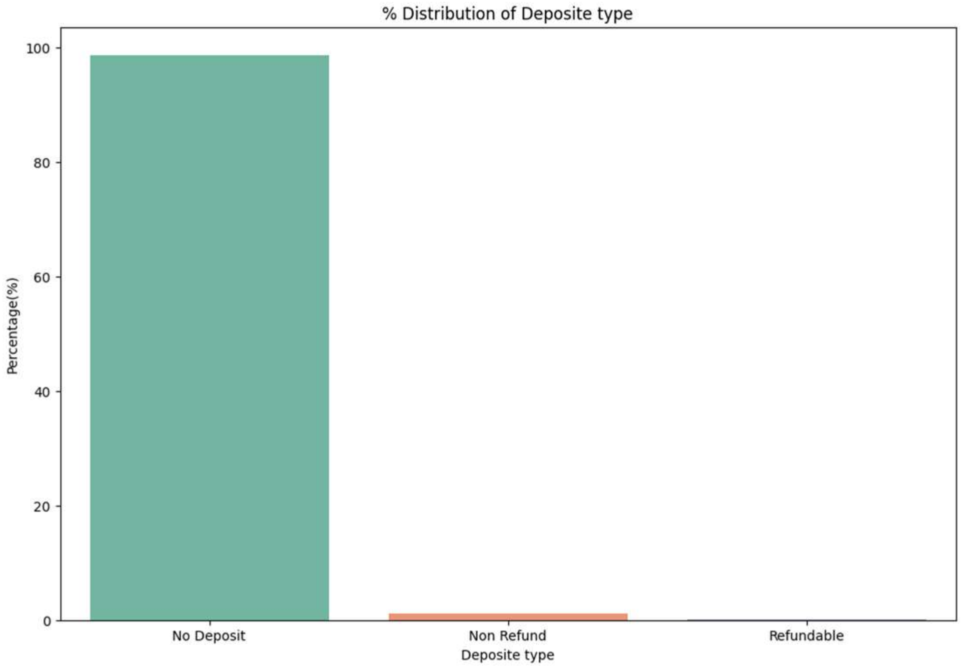
Most demanded room type is A.

Which hotel seems to make more revenue?



Avg ADR of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue

what is the percentage distribution by deposit type?

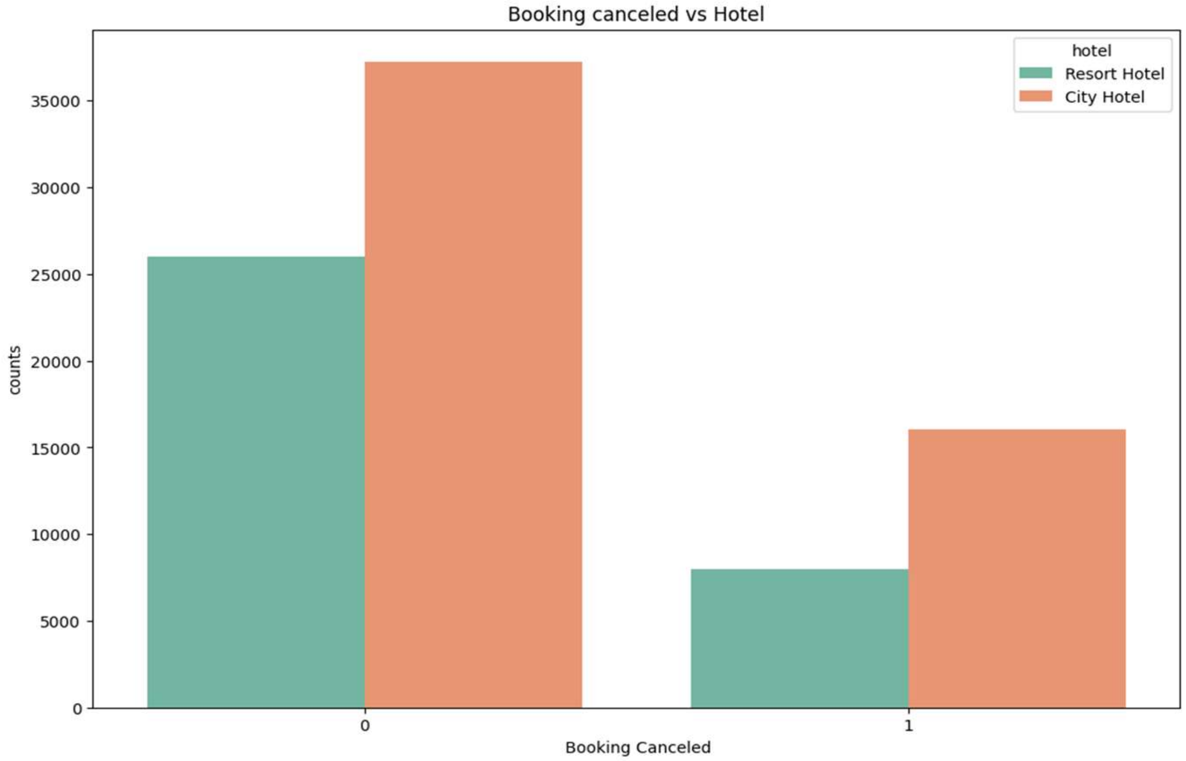


Almost 98 % of the guests prefer "No deposit" type of deposit so more chances of booking cancelation.

Bivariate and Multivariate Analysis:-

A Bivariate analysis is will measure the correlations between the two variables. Multivariate analysis is will measure the correlations between more than two variables.

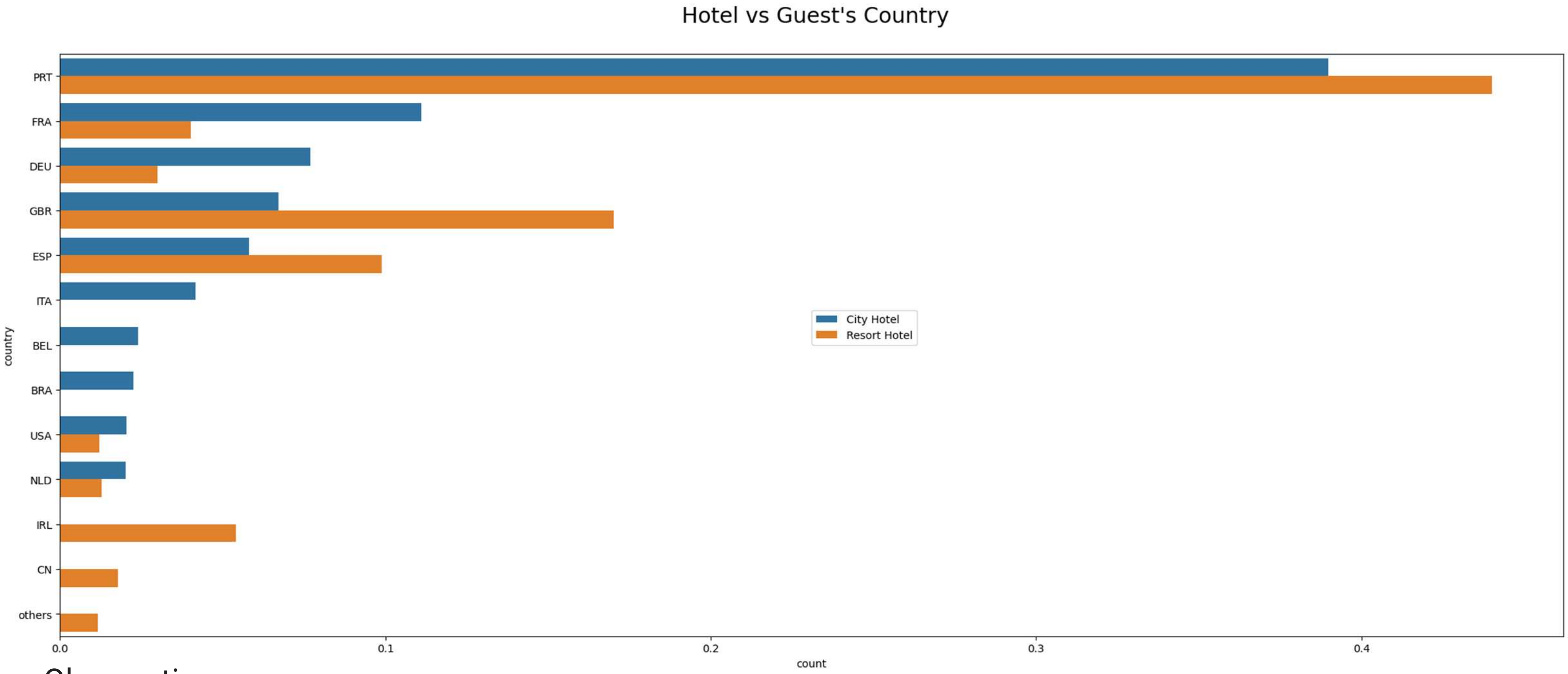
what is the booking cancelation numbers with respect to hotel?



Observation:-

Booking cancelation of city hotel is more as compared to resort hotel.
Almost 15k city hotel booking is canceled.

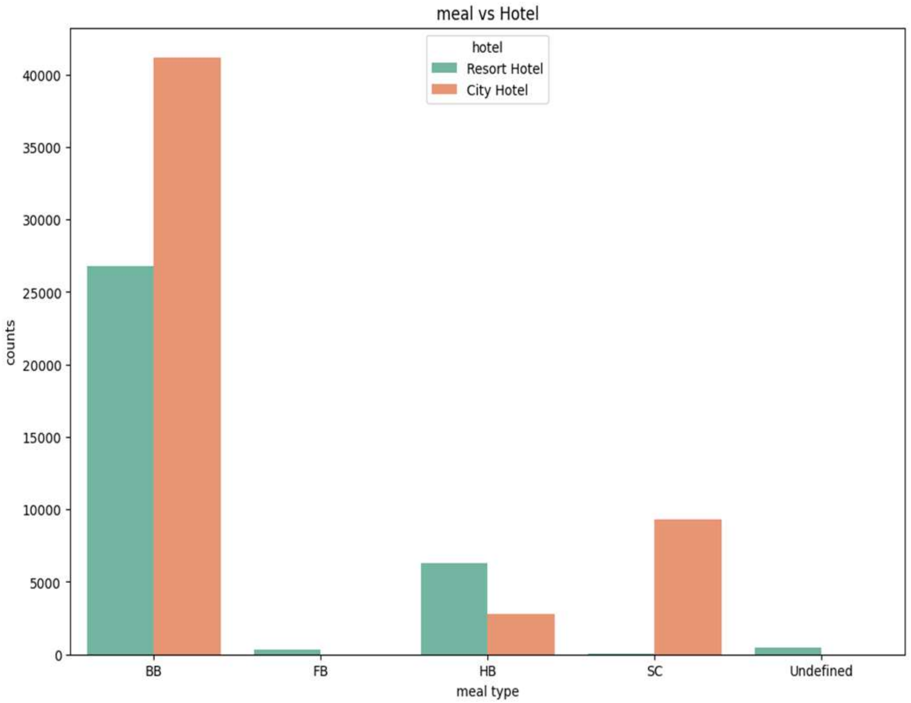
which is the most preferred hotel type by different country guests?



Observations:-

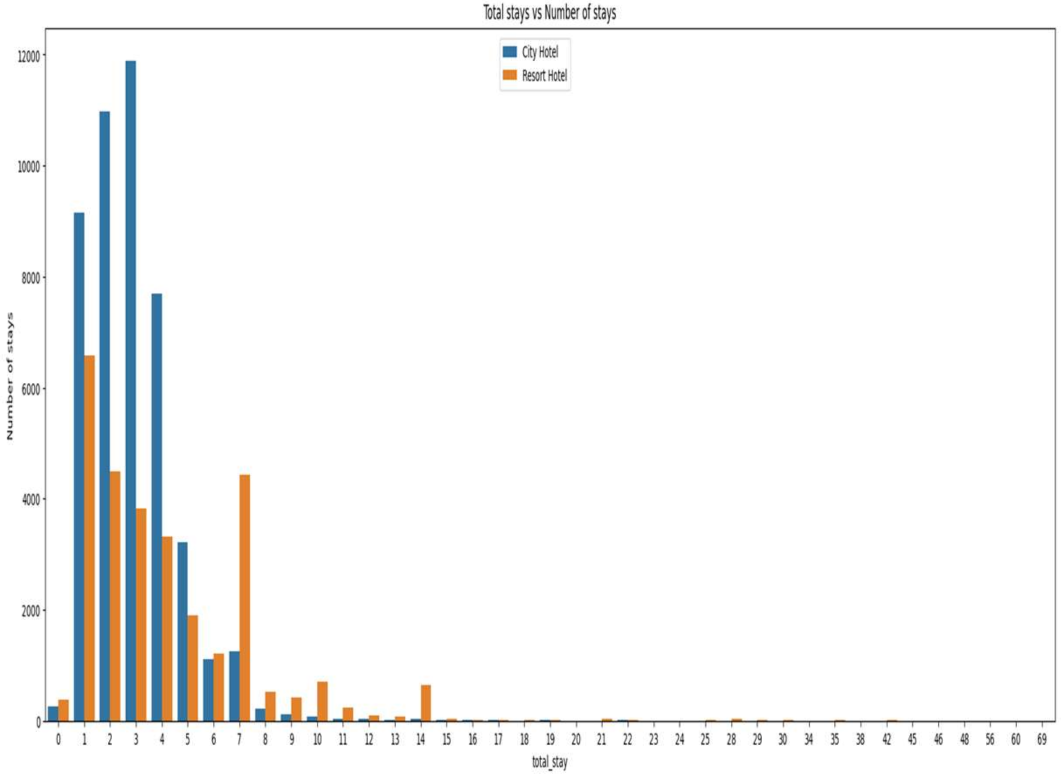
Most of the guests are from PRT and they happen to choose Resort Hotel more than City Hotel

What is the meal type provided by different hotel type?



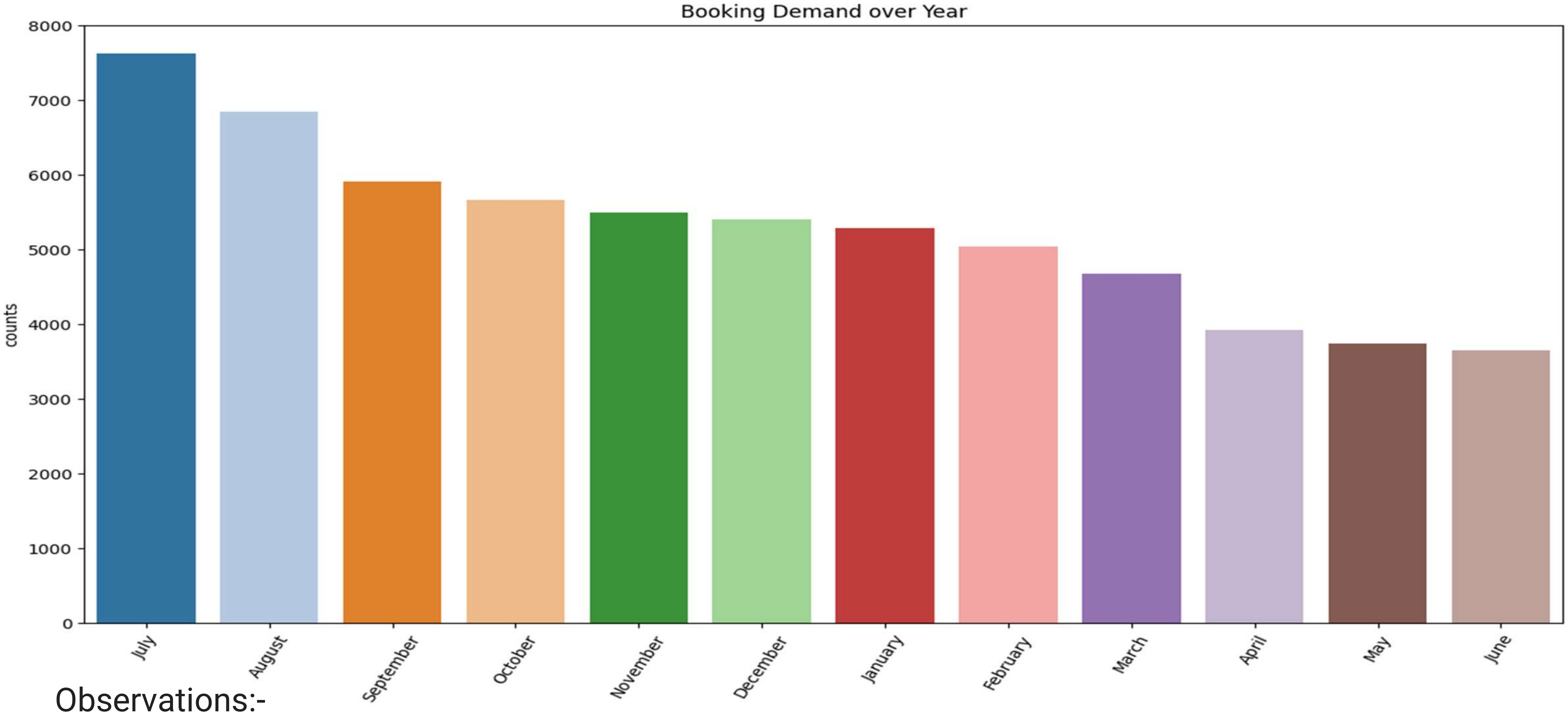
Observation:-
From meal we can observe that City Hotel provide 'BB' and Resort Hotel provide 'HB'

How long do people stay at the hotels?



Observations:-
Most people prefer to stay at the city hotels of for maximum 7 days
For more than 7 days stay people prefer resort hotels.

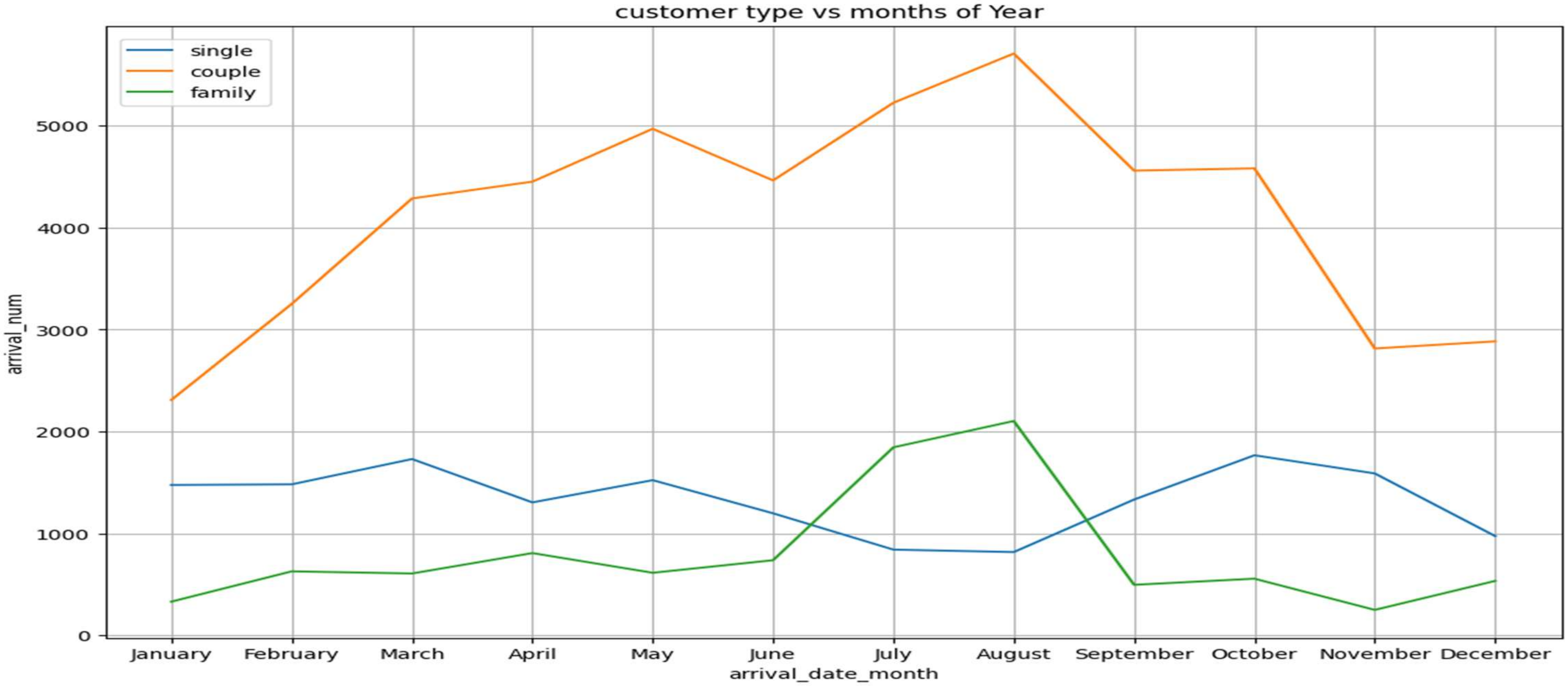
How is the Booking demand over year?



Observations:-

July is the most demanded month of the year. Interestingly, number of bookings decreased over year and June, the month before July, is the least demanded one.

Type of customers arriving in different months of year

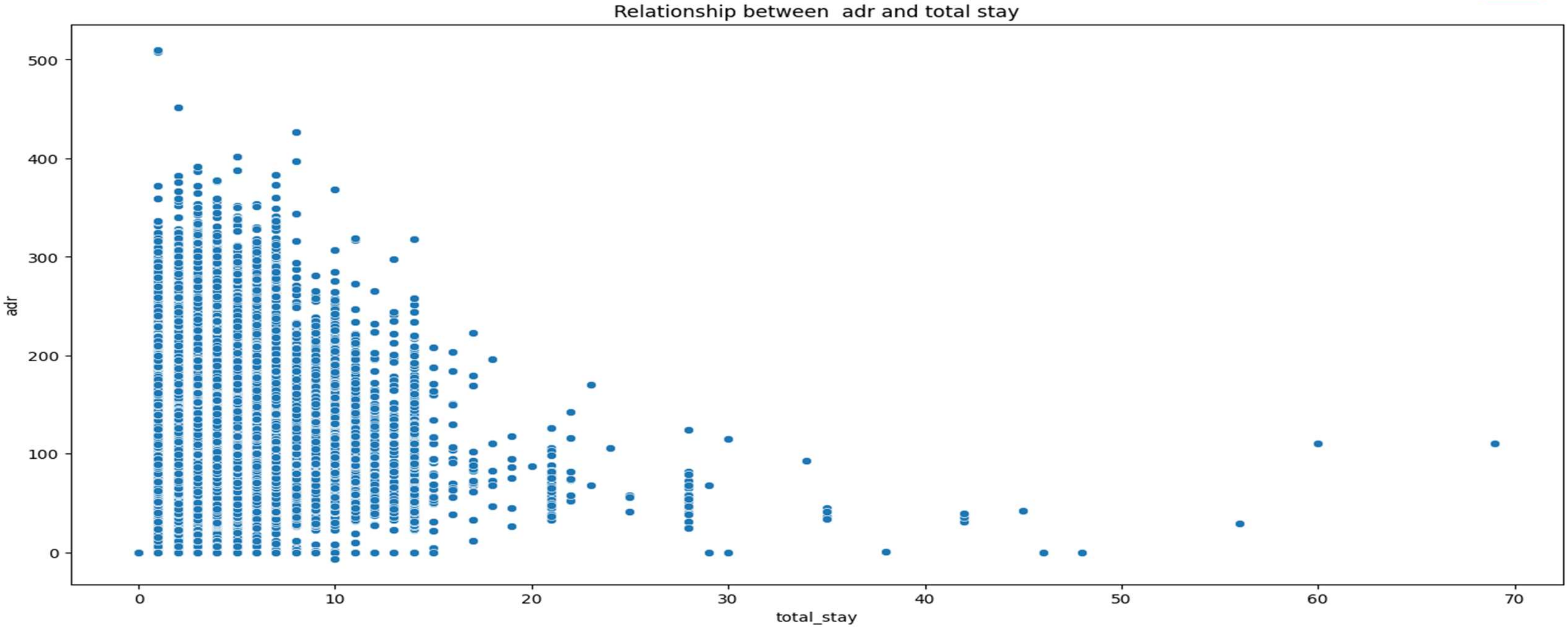


Observations:-

Mostly bookings are done by couples(although we are not sure that they are couple as there is no gender column in data).

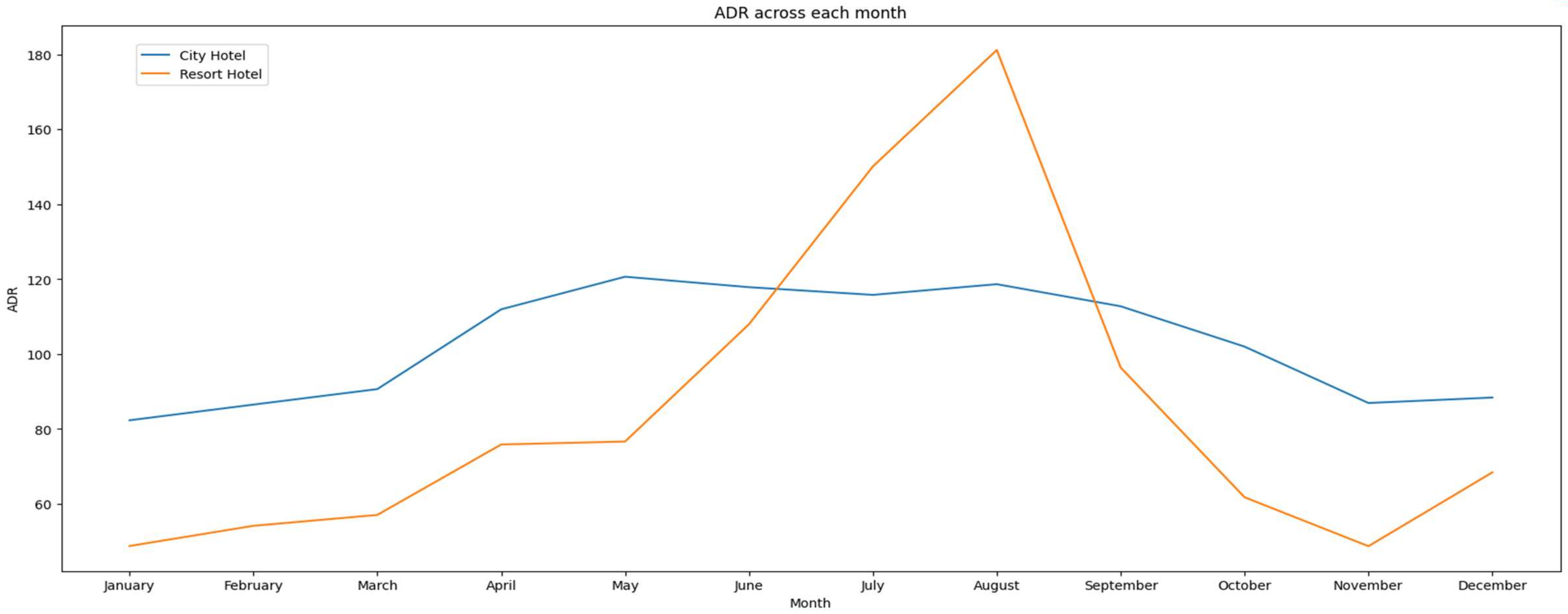
It is clear from graph that their is a sudden surge in arrival number of couples and family in months of July and August. So better plans can be planned accordingly at that time for these type of customers

Relation between ADR and Total stay



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

How is ADR across the different months?



Observation:-

For Resort hotel ADR is high in the month June, July, August as compared to City Hotels. May be Customers/People wants to spend their Summer vacation in Resorts Hotels.

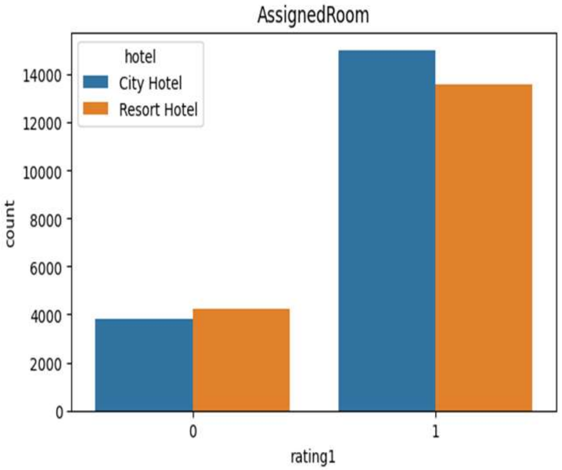
The best time for guests to visit Resort or City hotels is January, February, March, April, October, November and December as the average daily rate in this month is very low.

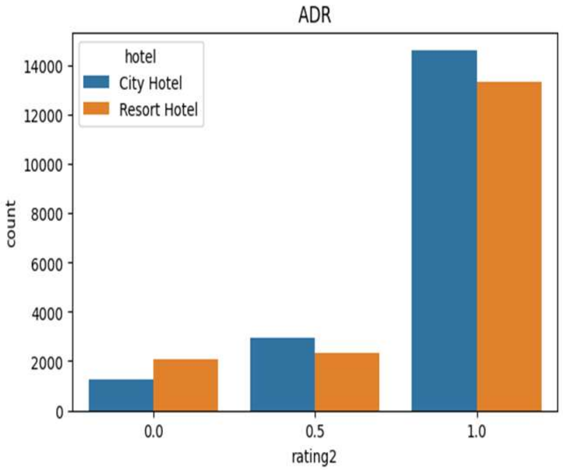
Let's find Rating provided by customers, which help us to find out the reasons for booking cancelation of Hotels.

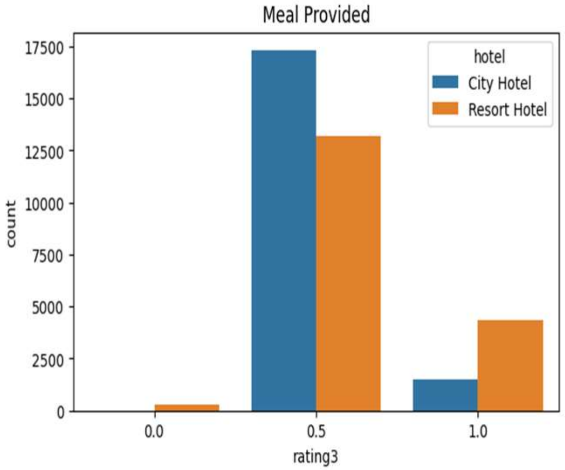
Rating1:-If reserved room type and assigned room
type is same then the customer will give rating1=1 else rating1=0 to hotel.

Rating2:-If ADR is less than 150 then the customer will give rating2=1 and .5 if ADR is in between 150 to 200 else rating2=0 to hotel.

Rating3:-If meal type is FB or HB then the customer will give rating3=1 and rating3=0.5 for BB,SC else rating3=0to hotel.



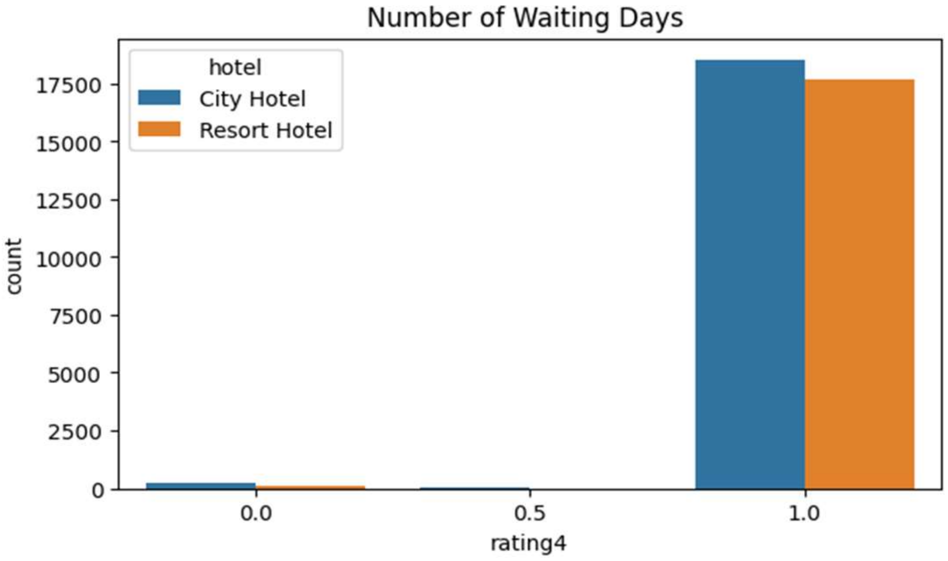


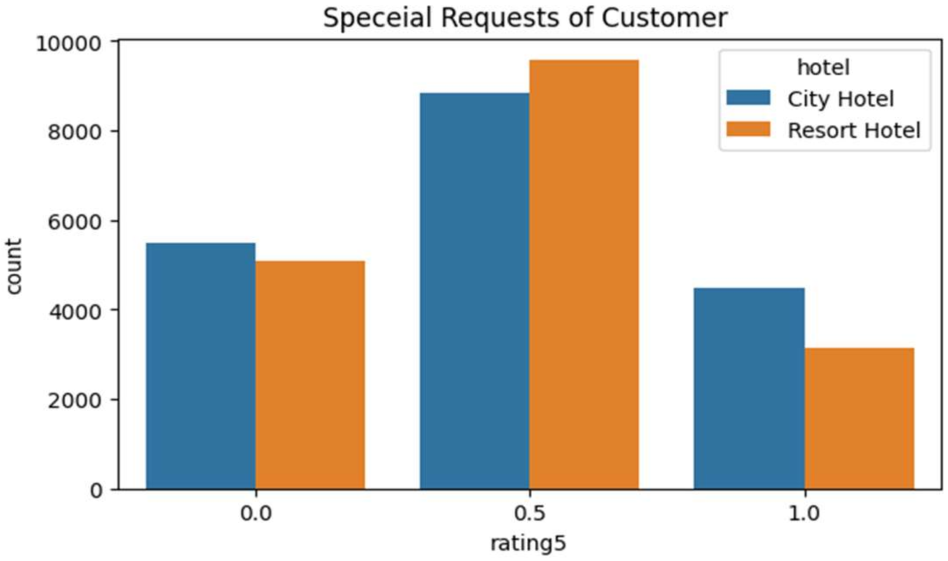


Observations:-
1]Both the hotels will provide the reserved room to the guests.
2]City hotels are less expensive as compared to resort hotels.
3]Resort hotels are providing meal type (FB or HB) where as maximum city hotels are providing meal type (BB or SC).

Rating4:-If waiting days in list is less than 15 then the customer will give rating4=1 andrating4=0.5 if in between 15 to 30 days else rating4=0 to hotel.

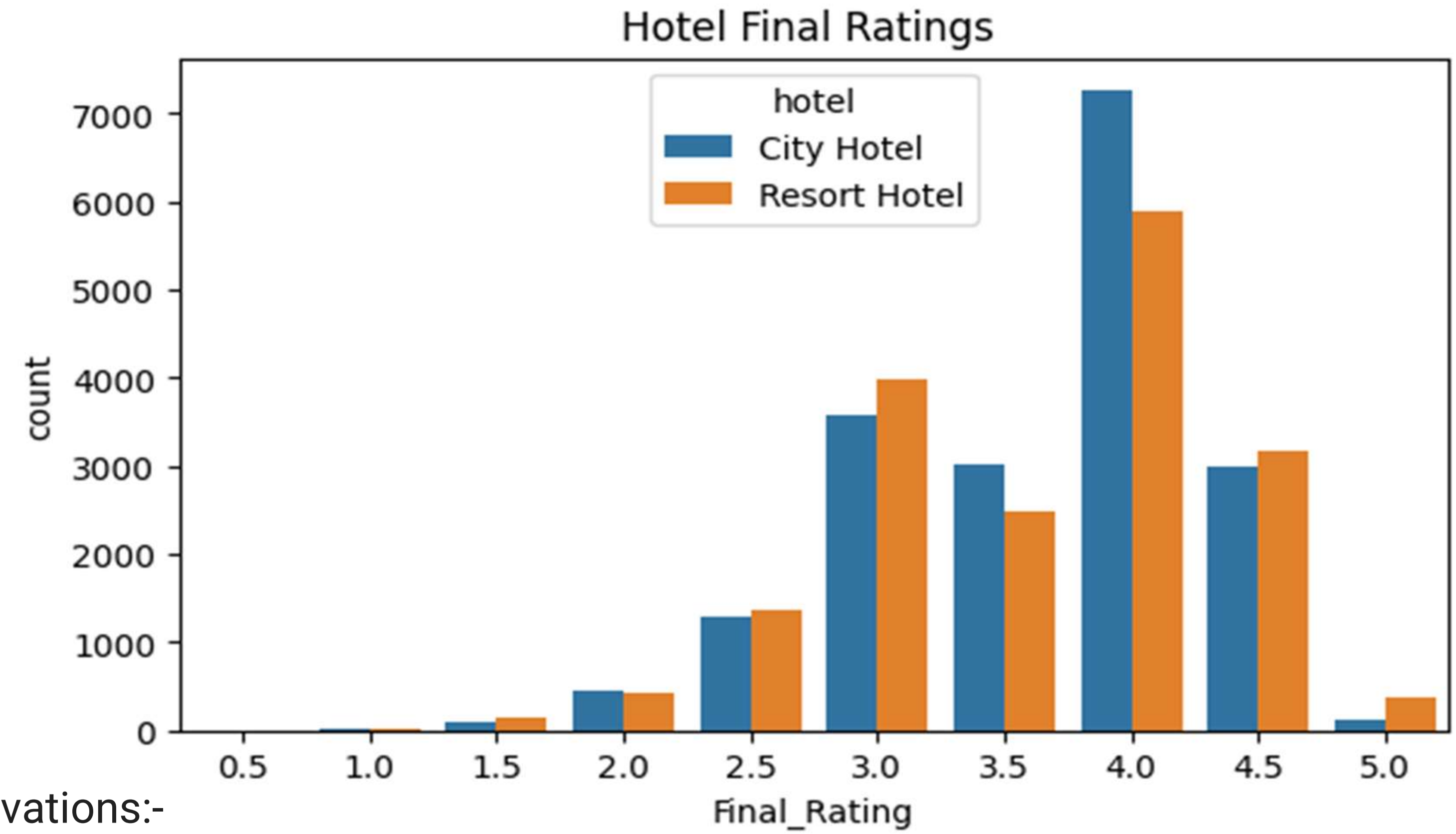
Rating5:If customer get adr less than 200 with more than 1 special requests(for ex:- Free Wifi, Room Service, Seaview, Complimentary Toiletries.) then the customer will give rating5=1 and 0.5 if ADR less than 120 and 1sp.req.else rating5=0 to hotel.





Observations:-
Most of the City and Resort hotels have waiting time less than 15 days.
Maximum customer giving 0.5 ratings according to there special requests and very few hotels providing reserved room with special requests of customer because of that rating is less than 1.

Final Rating



Observations:-

There are around 7K City hotels and 6k resort hotels with rating 4.0 out of 36K hotel rating given by the customer.
Resort hotels have 4.5 above ratings as compared to city hotels but less number of such resort hotels.

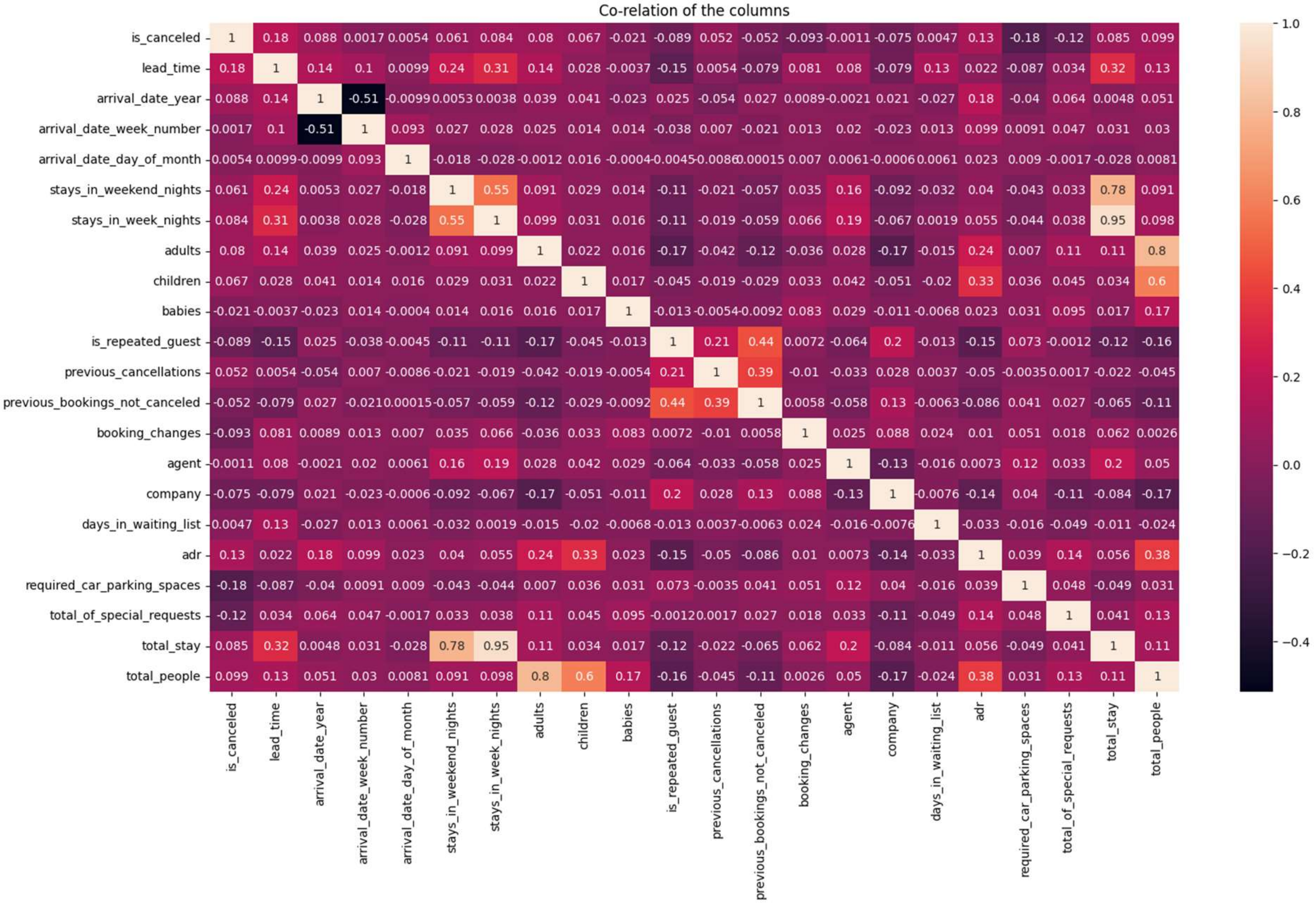
Correlation Heat map

Observations:-

1) lead-time and total stay is positively corelated.

2)adr and total people are positively corelated .

3) is repeated guest and previous bookings not canceled has strong correlation.



	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	agent	company	days_in_waiting_list	adr	required_car_parking_spaces	total_of_special_requests	total_stay	total_people
is_canceled	1	0.18	0.088	0.0017	0.0054	0.061	0.084	0.08	0.067	-0.021	-0.089	0.052	-0.052	-0.093	-0.0011	-0.075	0.0047	0.13	-0.18	-0.12	0.085	0.099
lead_time	0.18	1	0.14	0.1	0.0099	0.24	0.31	0.14	0.028	-0.0037	-0.15	0.0054	-0.079	0.081	0.08	-0.079	0.13	0.022	-0.087	0.034	0.32	0.13
arrival_date_year	0.088	0.14	1	-0.51	-0.0099	0.0053	0.0038	0.039	0.041	-0.023	0.025	-0.054	0.027	0.0089	-0.0021	0.021	-0.027	0.18	-0.04	0.064	0.0048	0.051
arrival_date_week_number	0.0017	0.1	-0.51	1	0.093	0.027	0.028	0.025	0.014	0.014	-0.038	0.007	-0.021	0.013	0.02	-0.023	0.013	0.099	0.0091	0.047	0.031	0.03
arrival_date_day_of_month	0.0054	0.0099	-0.0099	0.093	1	-0.018	-0.028	-0.0012	0.016	-0.0004	-0.0045	-0.0086	0.00015	0.007	0.0061	-0.0006	0.0061	0.023	0.009	-0.0017	-0.028	0.0081
stays_in_weekend_nights	0.061	0.24	0.0053	0.027	-0.018	1	0.55	0.091	0.029	0.014	-0.11	-0.021	-0.057	0.035	0.16	-0.092	-0.032	0.04	-0.043	0.033	0.78	0.091
stays_in_week_nights	0.084	0.31	0.0038	0.028	-0.028	0.55	1	0.099	0.031	0.016	-0.11	-0.019	-0.059	0.066	0.19	-0.067	0.0019	0.055	-0.044	0.038	0.95	0.098
adults	0.08	0.14	0.039	0.025	-0.0012	0.091	0.099	1	0.022	0.016	-0.17	-0.042	-0.12	-0.036	0.028	-0.17	-0.015	0.24	0.007	0.11	0.11	0.8
children	0.067	0.028	0.041	0.014	0.016	0.029	0.031	0.022	1	0.017	-0.045	-0.019	-0.029	0.033	0.042	-0.051	-0.02	0.33	0.036	0.045	0.034	0.6
babies	-0.021	-0.0037	-0.023	0.014	-0.0004	0.014	0.016	0.016	0.017	1	-0.013	-0.0054	-0.0092	0.083	0.029	-0.011	-0.0068	0.023	0.031	0.095	0.017	0.17
is_repeated_guest	-0.089	-0.15	0.025	-0.038	-0.0045	-0.11	-0.11	-0.17	-0.045	-0.013	1	0.21	0.44	0.0072	-0.064	0.2	-0.013	-0.15	0.073	-0.0012	-0.12	-0.16
previous_cancellations	0.052	0.0054	-0.054	0.007	-0.0086	-0.021	-0.019	-0.042	-0.019	-0.0054	0.21	1	0.39	-0.01	-0.033	0.028	0.0037	-0.05	-0.0035	0.0017	-0.022	-0.045
previous_bookings_not_canceled	-0.052	-0.079	0.027	-0.021	0.00015	-0.057	-0.059	-0.12	-0.029	-0.0092	0.44	0.39	1	0.0058	-0.058	0.13	-0.0063	-0.086	0.041	0.027	-0.065	-0.11
booking_changes	-0.093	0.081	0.0089	0.013	0.007	0.035	0.066	-0.036	0.033	0.083	0.0072	-0.01	0.0058	1	0.025	0.088	0.024	0.01	0.051	0.018	0.062	0.0026
agent	-0.0011	0.08	-0.0021	0.02	0.0061	0.16	0.19	0.028	0.042	0.029	-0.064	-0.033	-0.058	0.025	1	-0.13	-0.016	0.0073	0.12	0.033	0.2	0.05
company	-0.075	-0.079	0.021	-0.023	-0.0006	-0.092	-0.067	-0.17	-0.051	-0.011	0.2	0.028	0.13	0.088	-0.13	1	-0.0076	-0.14	0.04	-0.11	-0.084	-0.17
days_in_waiting_list	0.0047	0.13	-0.027	0.013	0.0061	-0.032	0.0019	-0.015	-0.02	-0.0068	-0.013	0.0037	-0.0063	0.024	-0.016	-0.0076	1	-0.033	-0.016	-0.049	-0.011	-0.024
adr	0.13	0.022	0.18	0.099	0.023	0.04	0.055	0.24	0.33	0.023	-0.15	-0.05	-0.086	0.01	0.0073	-0.14	-0.033	1	0.039	0.14	0.056	0.38
required_car_parking_spaces	-0.18	-0.087	-0.04	0.0091	0.009	-0.043	-0.044	0.007	0.036	0.031	0.073	-0.0035	0.041	0.051	0.12	0.04	-0.016	0.039	1	0.048	-0.049	0.031
total_of_special_requests	-0.12	0.034	0.064	0.047	-0.0017	0.033	0.038	0.11	0.045	0.095	-0.0012	0.0017	0.027	0.018	0.033	-0.11	-0.049	0.14	0.048	1	0.041	0.13
total_stay	0.085	0.32	0.0048	0.031	-0.028	0.78	0.95	0.11	0.034	0.017	-0.12	-0.022	-0.065	0.062	0.2	-0.084	-0.011	0.056	-0.049	0.041	1	0.11
total_people	0.099	0.13	0.051	0.03	0.0081	0.091	0.098	0.8	0.6	0.17	-0.16	-0.045	-0.11	0.0026	0.05	-0.17	-0.024	0.38	0.031	0.13	0.11	1

Suggestions To Hotel Industry :-

Only few percent guests are repeated and remaining all are new guests so reward repeat guests.

Only single agent made the most booking try to give bonus and encourage other.

Most guests are from Portugal and other countries in Europe so attract your guests with advertisements , flight bookings discounts and special offers in different countries.

Almost all the guests prefer "No deposit" type of deposit so more chances of booking cancelation so make sure you have a solid cancellation policy in place.
You can confirm the booking with your guest and then offer them a discount once done.

Understand which tours or packages generate the most no show hotel reservation and take immediate action accordingly such as special discounts to couple and family in summer vacation(July ,Aug & Sep) .

Get in touch with guests through feedbacks try to improve your Final Ratings.

Conclusion

- Complete Conclusion of data:(Summary)
- 1.61% is City Hotel is and 39% is Resort Hotel
- 2. Most of the market segment used is Online Travel Agency and Offline Tour Agent.
- 3. 27.5% booking has been canceled.
- 4. 80% distribution channel is TA/TO
- 5. maximum guest book room type A .
- 6. Only 3.9% guests are repeated and remaining all are new guests.
- 7. 91.6 % guests did not required the parking space. only 8.3 % guests required only 1 parking space.
- 8.Agent number 9 made the most booking more than 25k.
- 9. Most preferred meal type is BB (Bed and breakfast).
- 10. 2016 Year has the most bookings 40k, 2015 has less than 15k bookings
- 11. People from all over the world are staying in these two hotels. Most guests are from Portugal and other countries in Europe guests are from PRT are happen to choose Resort Hotel more than City Hotel.
- 12. Almost 98 % of the guests prefer "No deposit" type of deposit so more chances of booking cancelation.
- 13. Booking cancelation of city hotel is more as compared to resort hotel.
- 14. Most people prefer to stay at the city hotels of for maximum 7 days For more than 7 days stay people prefer resort hotels.
- 15. July is the most demanded month of the year. Interestingly, number of bookings decreased over year and June, the month before July, is the least demanded one.
- 16. stay increases ADR is decreasing and For Resort hotel ADR is high in the month June, July, August as compared to City Hotels.
- 17.Very less number hotels with rating above 4.0 so that can be reason for the booking cancelations.

Thank You