

Capstone Project 1

Hotel Booking Analysis

By Shreya Ranjan

Hotel Booking Analysis

Data Set

The dataset contains information on bookings for two hotels a resort and a city hotel scheduled to arrive in a period between July 1, 2015 and August 31, 2017.

For both hotels, the same information was collected: 31 variables describing 40,060 observations for the resort and 79,330 observations for the city hotel. That is, the dataset contains information on 119,390 hotel reservations, including those that were canceled. This is real information, so all elements that could identify hotels or customers were removed.

DATA PIPELINE

Data processing

In this first part we've find out null values. Since there were nearly many columns with all null values. I have filled the column with zeros. And I manually go through each features selected from First part .And encoded the categorial features.

EDA

In this part ,we do some exploratory data analysis on the features selected in Data processing to see the trend.

Insights

Finally in this last but not the last part ,we create insights. Creating a insights is not a easy task .It is also a integrative process .We show how to start with a simple model , then slowly add complexity for better performance.

Data Summary

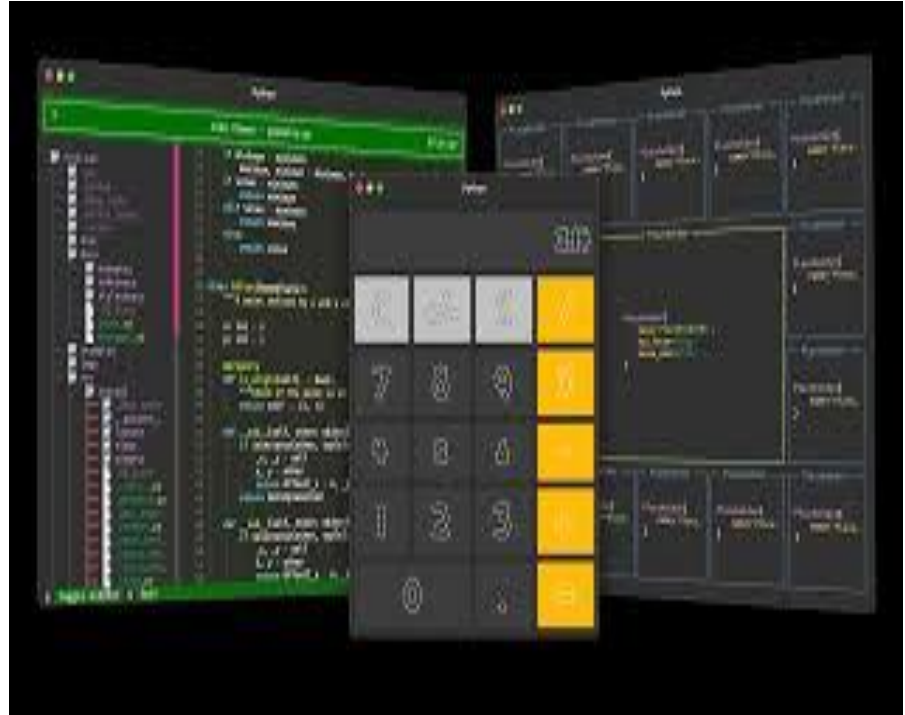
- **Hotel**-The category of resort ,which is Resort and City hotel
- **Is_cancelled**-The value of column show the cancellation type.If the booking was cancelled or not. Values[0,1],where 0 indicates not cancelled.
- **Stayed_in_weekend_nights**: The number of weekend nights stay per reservation.
- **Stayed_in_weekday_nights**: The number of weekday nights stay per reservation.
- **Meal**:Meal preferences per reservation.[BB, FB, HB, SC, Undefined]
- **Country**: The origin country of guest.
- .

Data Summary(contd..)

- **Market_segment:** This column shows how reservation was made and what is the purpose of reservation. Eg. corporate means corporate trip, TA for travel agency
- **Distribution_channel:** The median through which booking is made.
[Direct, TA/TO, corporate, undefined, GDS]
- **Is_repeated_channel:** Shows if the guest is who has arrived earlier or not. Values[0,1] → 0 indicates no and 1 indicates yes person is repeated guest.
- **Days_in_waiting_list:** Number of days between actual booking and transact.
- **Customer_type:** Type of customers (Transient, group, etc)

Libraries Used

- 1}Numpy
- 2}Pandas
- 3}Seaborn
- 4}Matplotlib
- 5}Plotly
- 6}Missingno



Data Wrangling

Shape of dataset-

```
df.shape
(119390, 32)
```

Data set information-

Data columns (total 32 columns):

#	Column	Non-Null Count	Dtype
0	hotel	119390 non-null	object
1	is_canceled	119390 non-null	int64
2	lead_time	119390 non-null	int64
3	arrival_date_year	119390 non-null	int64
4	arrival_date_month	119390 non-null	object
5	arrival_date_week_number	119390 non-null	int64
6	arrival_date_day_of_month	119390 non-null	int64
7	stays_in_weekend_nights	119390 non-null	int64
8	stays_in_week_nights	119390 non-null	int64
9	adults	119390 non-null	int64
10	children	119386 non-null	float64
11	babies	119390 non-null	int64
12	meal	119390 non-null	object
13	country	118902 non-null	object
14	market_segment	119390 non-null	object
15	distribution_channel	119390 non-null	object
16	is_repeated_guest	119390 non-null	int64
17	previous_cancellations	119390 non-null	int64
18	previous_bookings_not_canceled	119390 non-null	int64
19	reserved_room_type	119390 non-null	object
20	assigned_room_type	119390 non-null	object
21	booking_changes	119390 non-null	int64
22	deposit_type	119390 non-null	object
23	agent	103050 non-null	float64
24	company	6797 non-null	float64
25	days_in_waiting_list	119390 non-null	int64
26	customer_type	119390 non-null	object
27	adr	119390 non-null	float64
28	required_car_parking_spaces	119390 non-null	int64

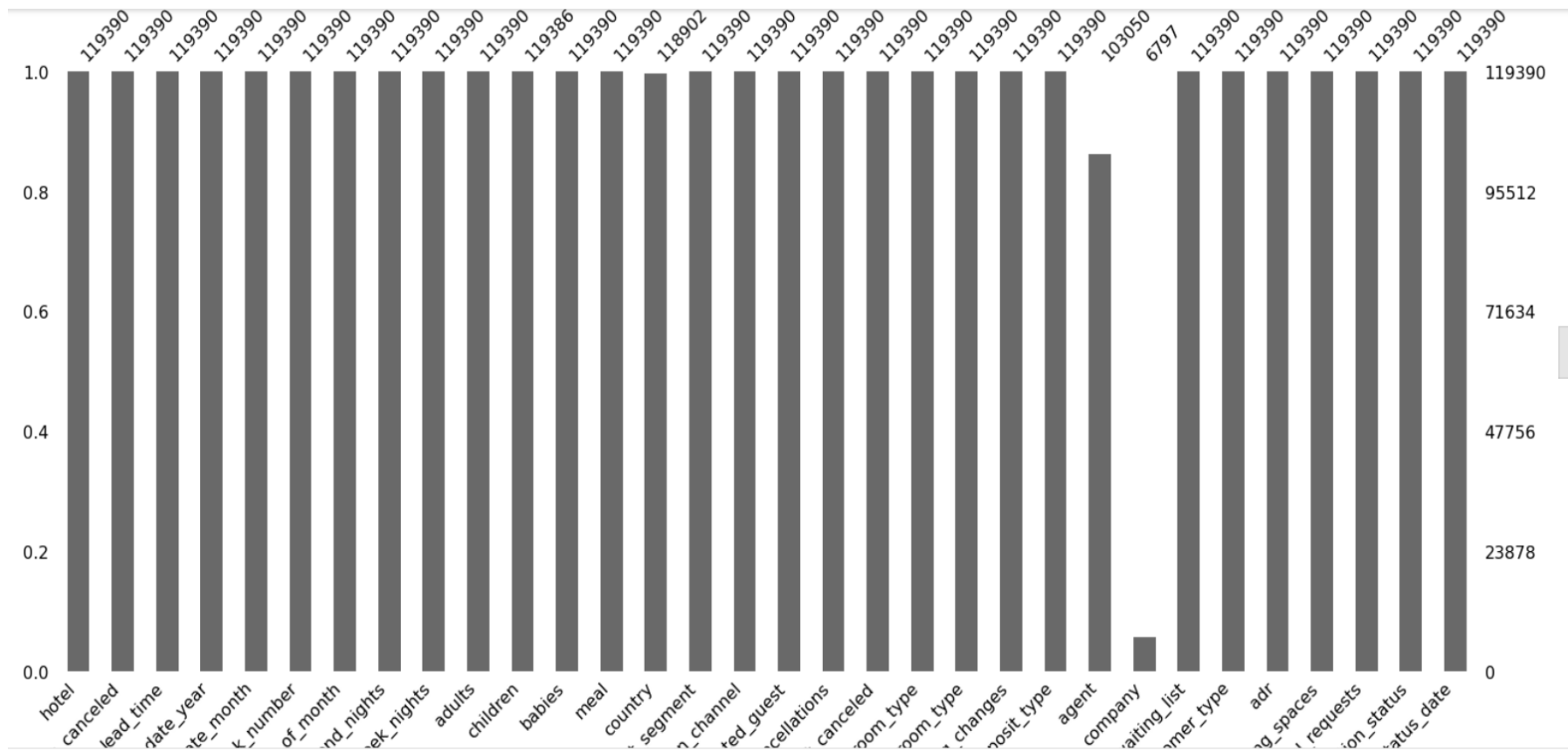
Data Wrangling(contd..)

Finding the null value-



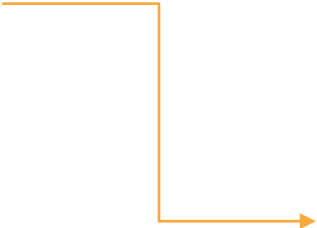
hotel	0
is_canceled	0
lead_time	0
arrival_date_year	0
arrival_date_month	0
arrival_date_week_number	0
arrival_date_day_of_month	0
stays_in_weekend_nights	0
stays_in_week_nights	0
adults	0
children	4
babies	0
meal	0
country	488
market_segment	0
distribution_channel	0
is_repeated_guest	0
previous_cancellations	0
previous_bookings_not_canceled	0
reserved_room_type	0
assigned_room_type	0
booking_changes	0
deposit_type	0
agent	16340
company	112593
days_in_waiting_list	0
customer_type	0
adr	0
required_car_parking_spaces	0
total_of_special_requests	0
reservation_status	0
reservation_status_date	0

Null value visualization



Replacing Null values

```
df2=df.fillna(value=0)
```

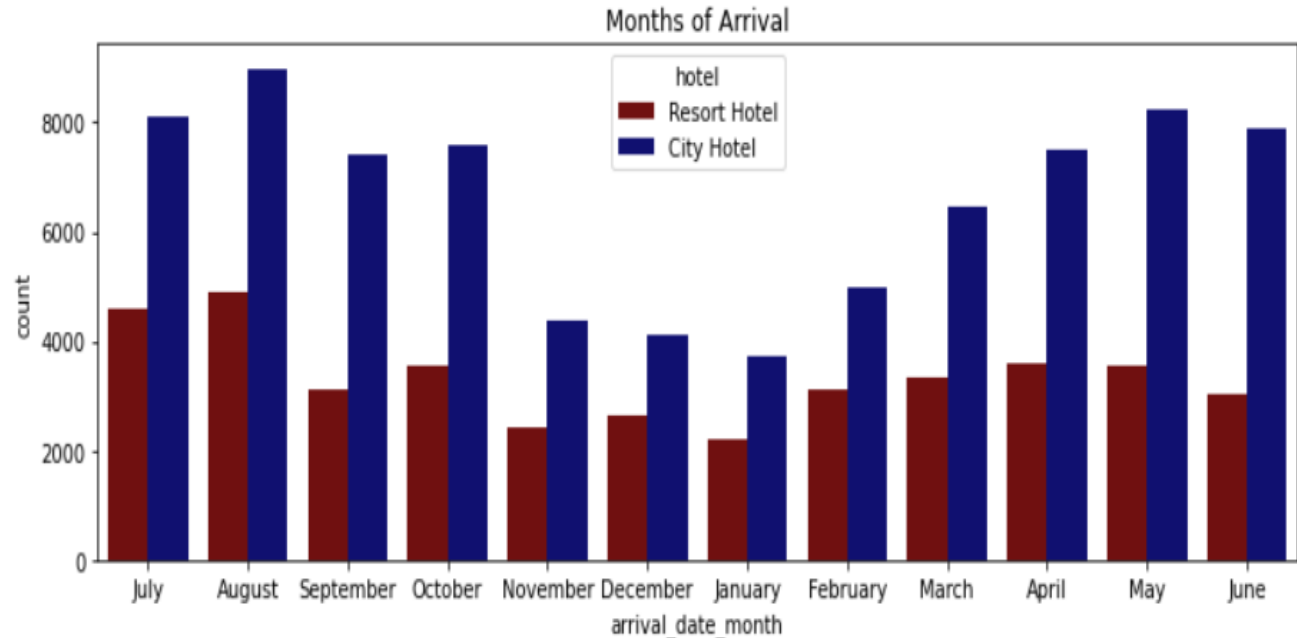


hotel	0
is_canceled	0
lead_time	0
arrival_date_year	0
arrival_date_month	0
arrival_date_week_number	0
arrival_date_day_of_month	0
stays_in_weekend_nights	0
stays_in_week_nights	0
adults	0
children	0
babies	0
meal	0
country	0
market_segment	0
distribution_channel	0
is_repeated_guest	0
previous_cancellations	0
previous_bookings_not_canceled	0
reserved_room_type	0
assigned_room_type	0
booking_changes	0
deposit_type	0
agent	0
company	0
days_in_waiting_list	0
customer_type	0
adr	0
required_car_parking_spaces	0
total_of_special_requests	0
reservation_status	0
reservation_status_date	0

Exploratory data analysis

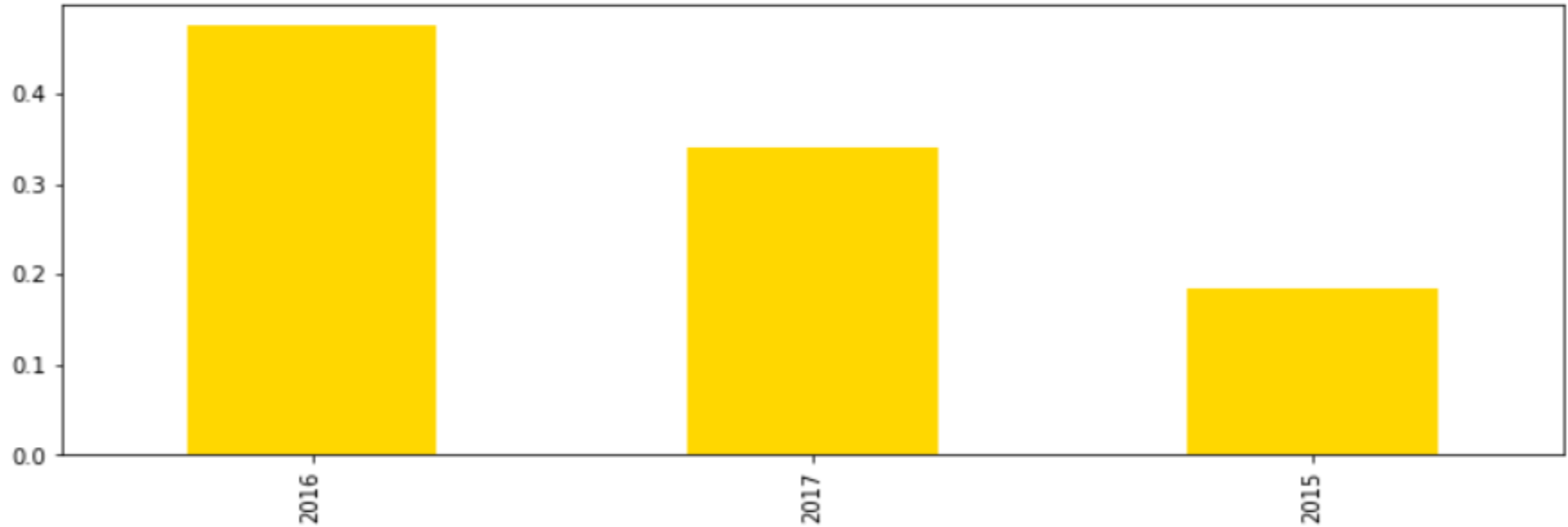
*What is the month with the most guest arrivals?

August	13877
July	12661
May	11791
October	11160
April	11089
June	10939
September	10508
March	9794
February	8068
November	6794
December	6780
January	5929



August and July are the months with the highest number of arrivals for both hotels.

*What is the year with the most guest arrivals?



2016 is the year in which most of the guest have visited

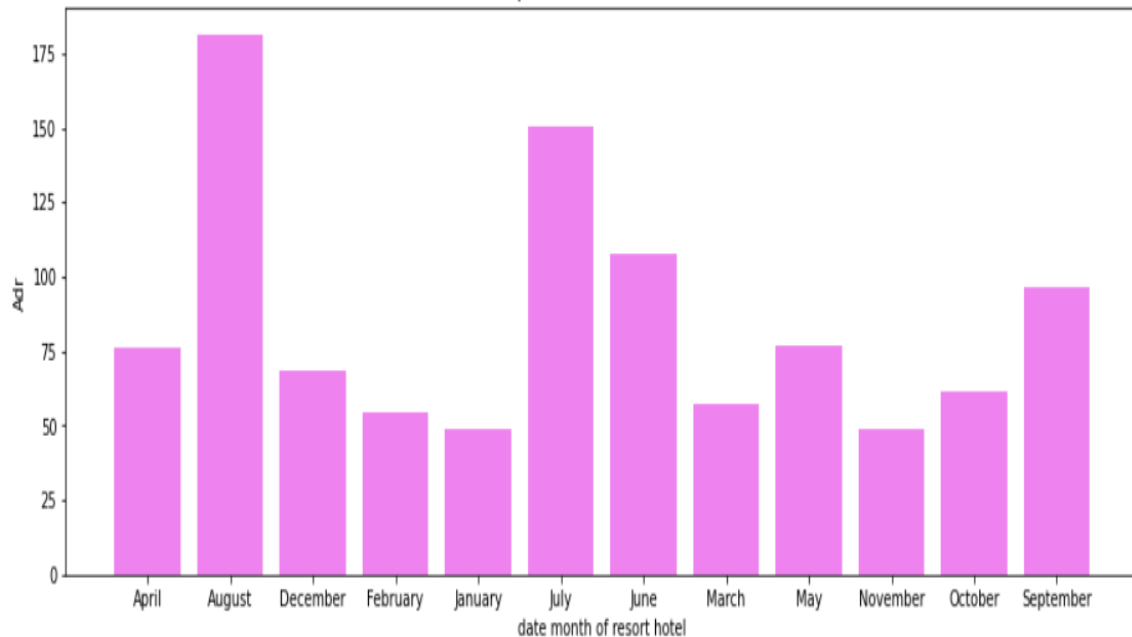
2016	56707
2017	40687
2015	21996

*How does the price vary per night over the year?

Resort hotel analysis



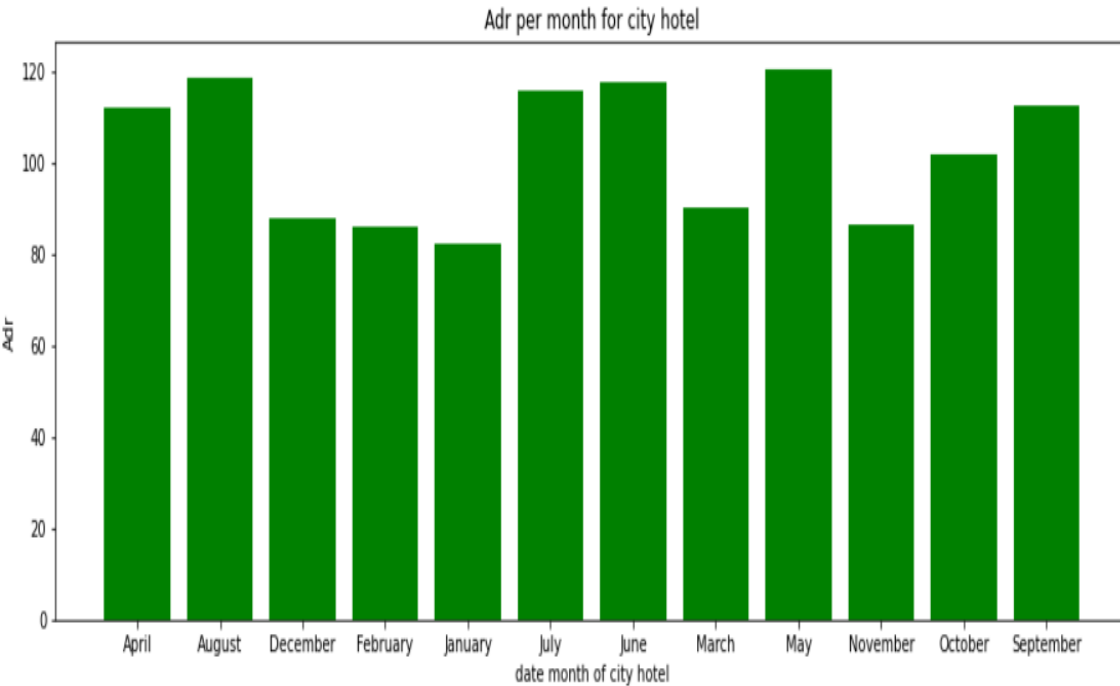
Adr per month for resort hotel



	arrival_date_month	adr
0	April	75.867816
1	August	181.205892
2	December	68.322236
3	February	54.147478
4	January	48.708919
5	July	150.122528
6	June	107.921869
7	March	57.012487
8	May	76.657558
9	November	48.681640
10	October	61.727505
11	September	96.416860

August is having a highest average distributed rate followed by July and June.

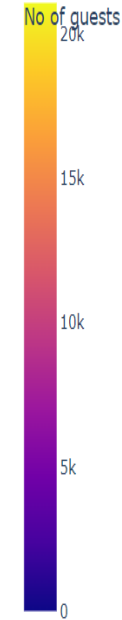
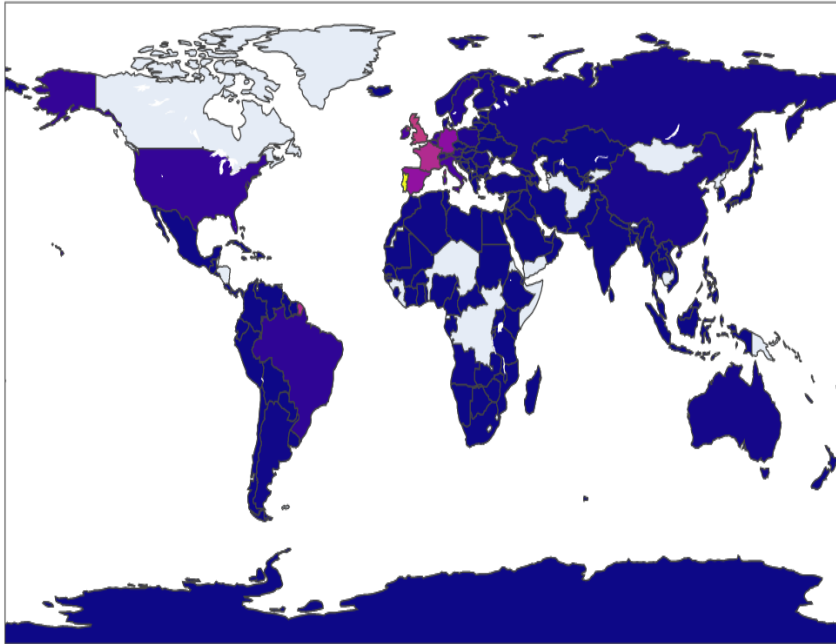
City Hotel analysis



	arrival_date_month	adr
0	April	111.856824
1	August	118.412083
2	December	87.856764
3	February	86.183025
4	January	82.160634
5	July	115.563810
6	June	117.702075
7	March	90.170722
8	May	120.445842
9	November	86.500456
10	October	101.745956
11	September	112.598452

August is having a highest average distributed range followed by june and July.

*Which countries do customers come from?



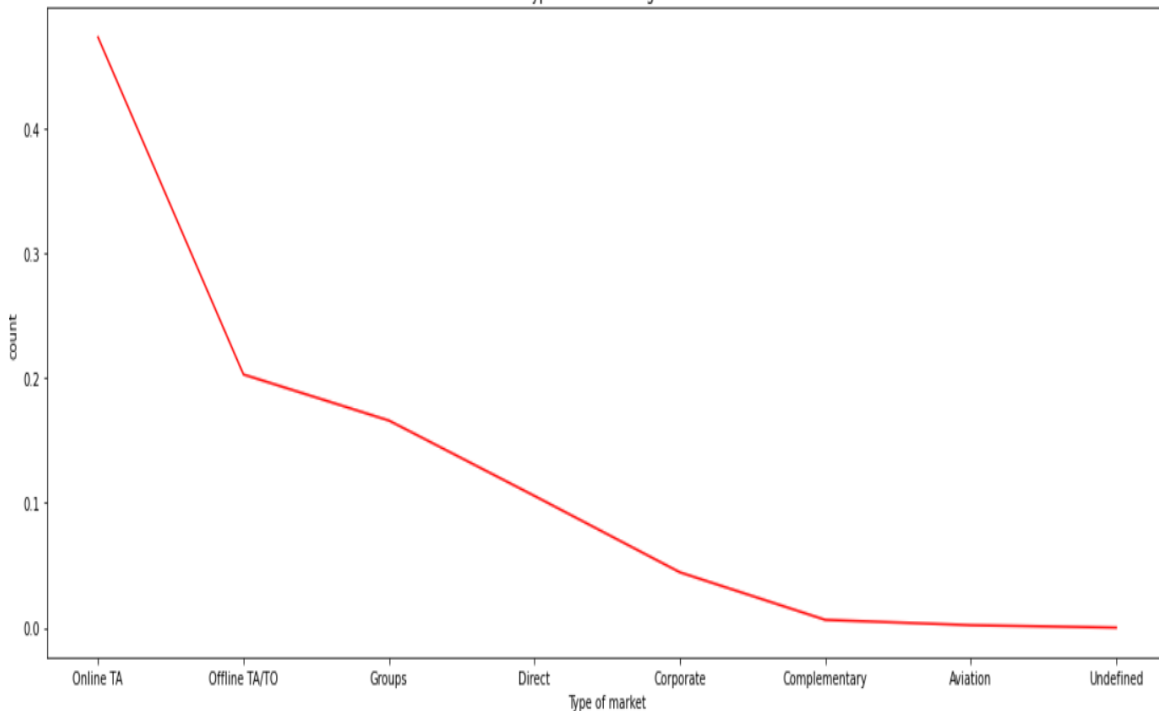
	country	No of guests
0	PRT	21071
1	GBR	9676
2	FRA	8481
3	ESP	6391
4	DEU	6069
...
161	BHR	1
162	DJI	1
163	MLI	1
164	NPL	1
165	FRO	1

Most customers come from Europe, mainly from Portugal and neighboring countries

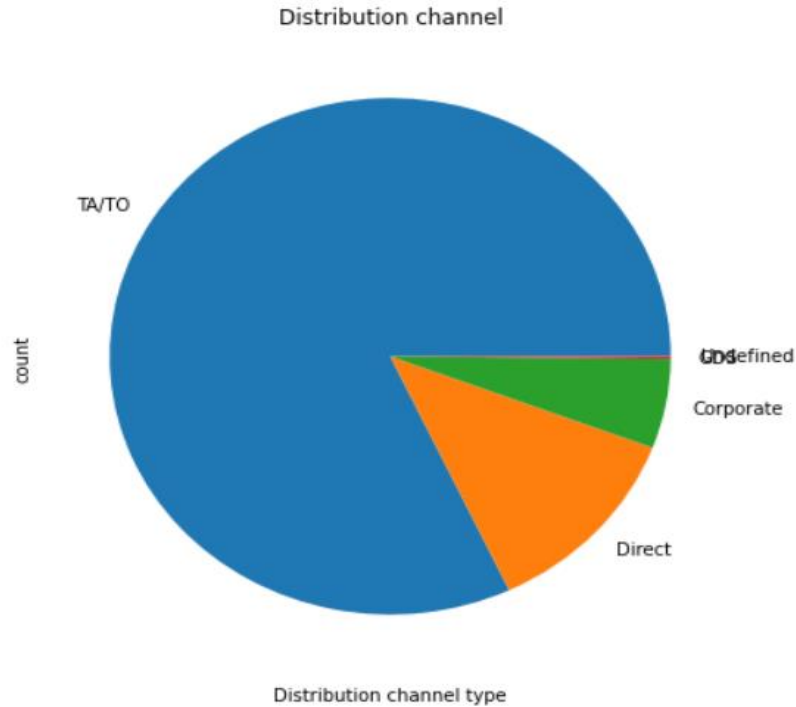
*What is the strongest market segment and distribution channel?



types of market segment



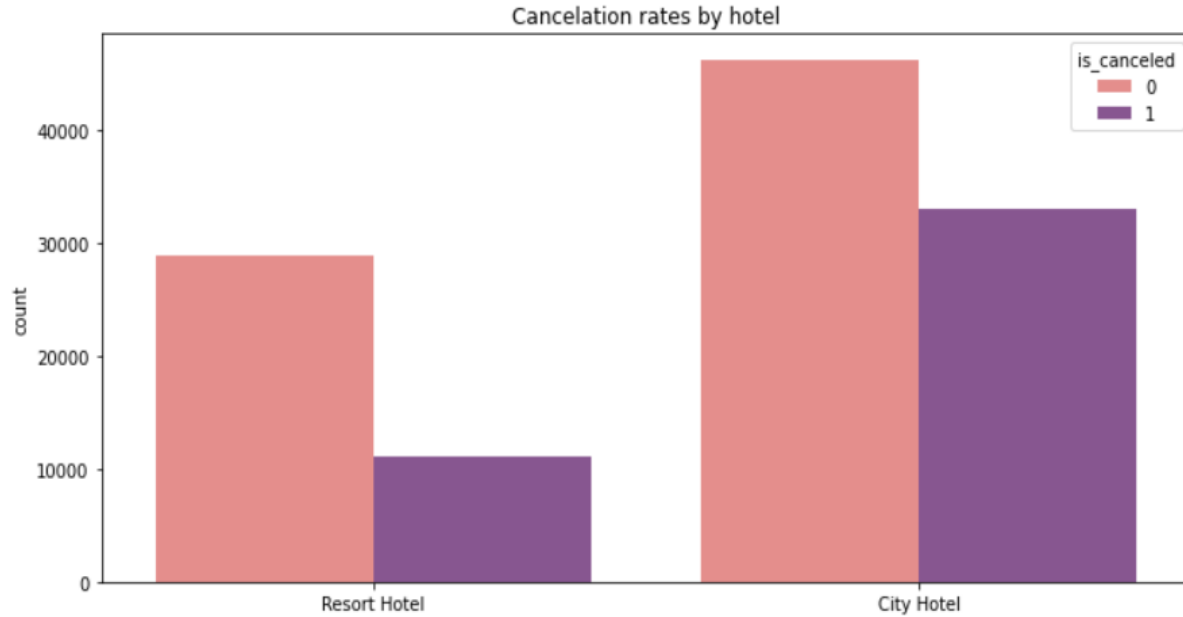
	index	market_segment	
0	Online TA	56477	
1	Offline TA/TO	24219	
2	Groups	19811	
3	Direct	12606	
4	Corporate	5295	
5	Complementary	743	
6	Aviation	237	
7	Undefined	2	



	index	distribution_channel
0	TA/TO	97870
1	Direct	14645
2	Corporate	6677
3	GDS	193
4	Undefined	5

Majority distribution channels and market segment were Travel agencies either offline/online.

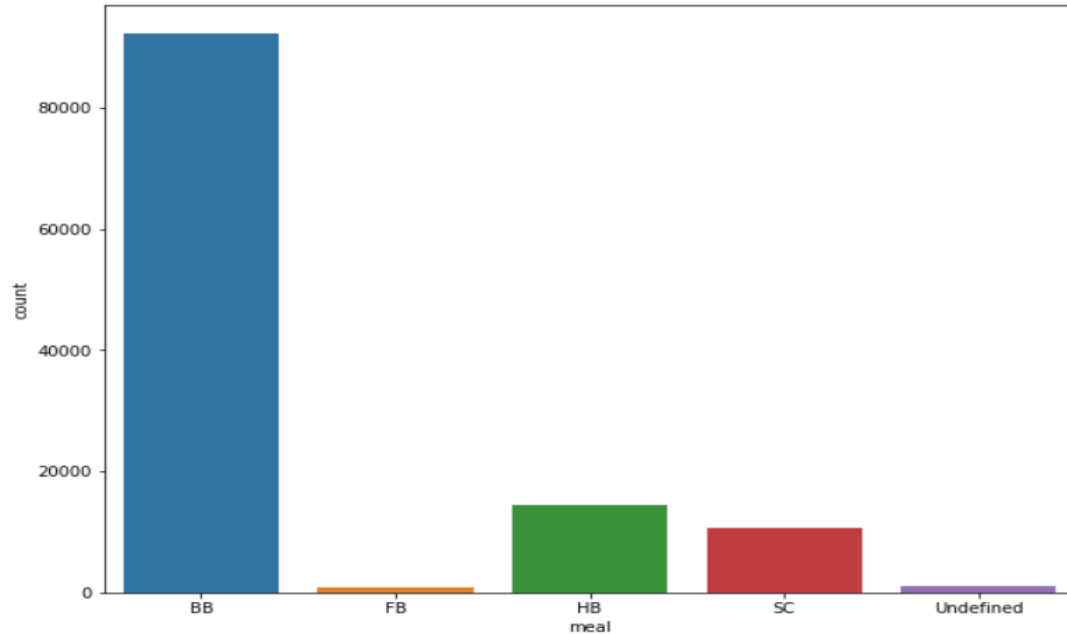
*How many reservations were cancelled out of total?



Here zero shows that much of booking is not cancelled and 1 represent the cancelled booking.

By the graph we can come to the conclusion city hotel booking are cancelled more than the resort hotel booking*

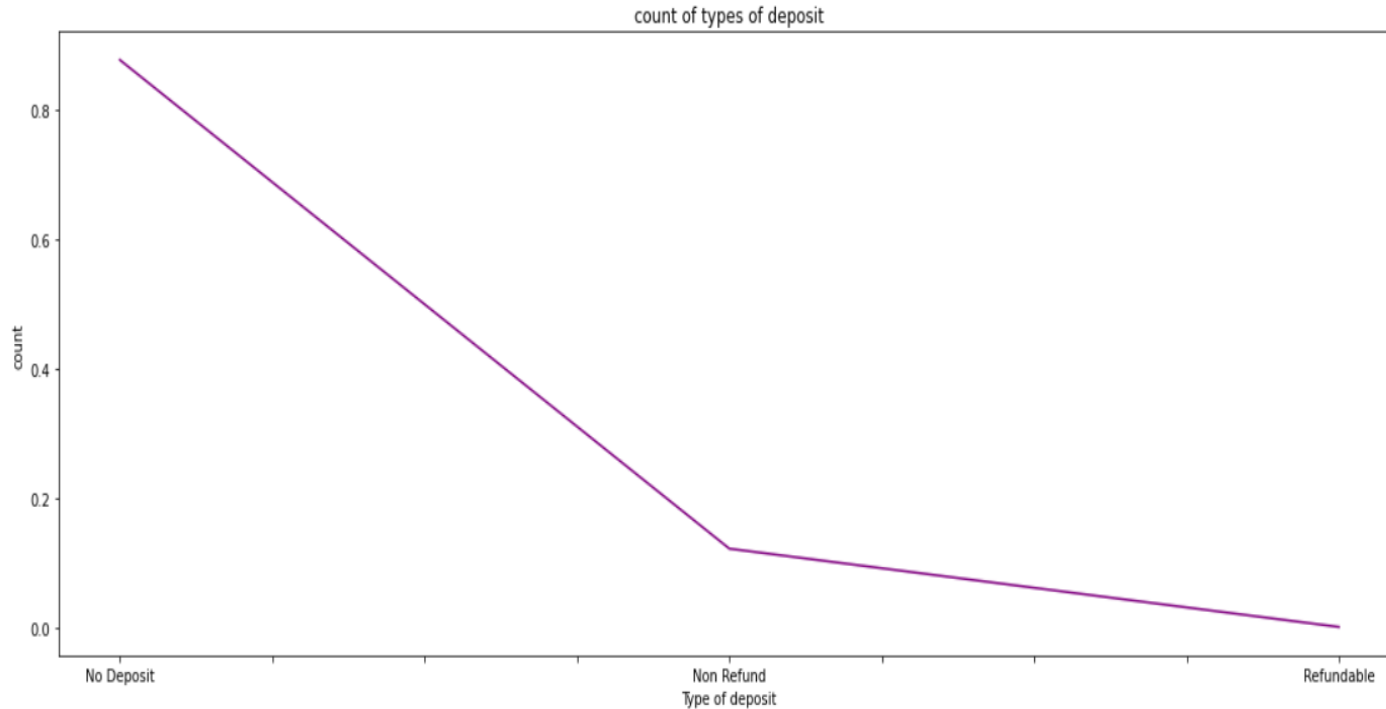
*What is their preferred meal plan?



	index	meal
0	BB	92310
1	HB	14463
2	SC	10650
3	Undefined	1169
4	FB	798

The preferred meal plan is bed and breakfast, as almost 92310 of bookings were made on this type of plan. half board (HB) and full board (FB) are less frequent options, but together they represent 25113 of reservations. It is important to note that practically no clients stayed with the self-catering option.

*What is the most frequent deposit type?

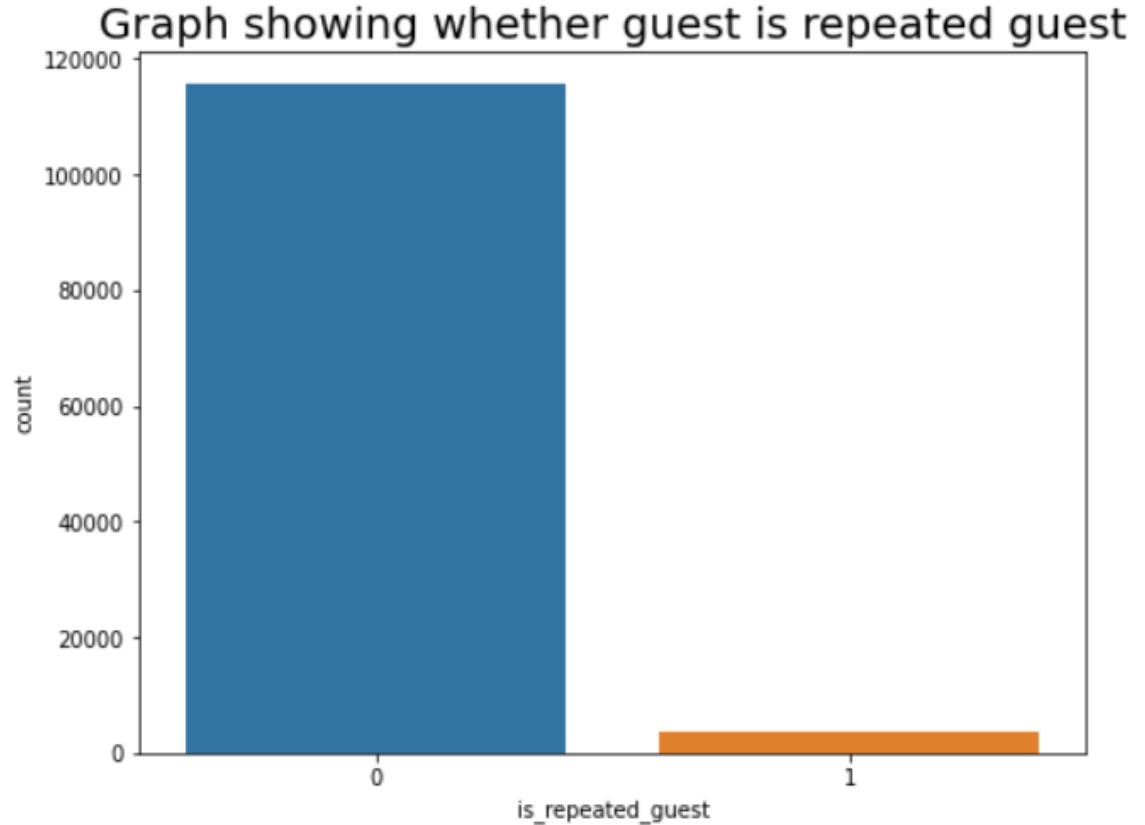


No Deposit	104641
Non Refund	14587
Refundable	162

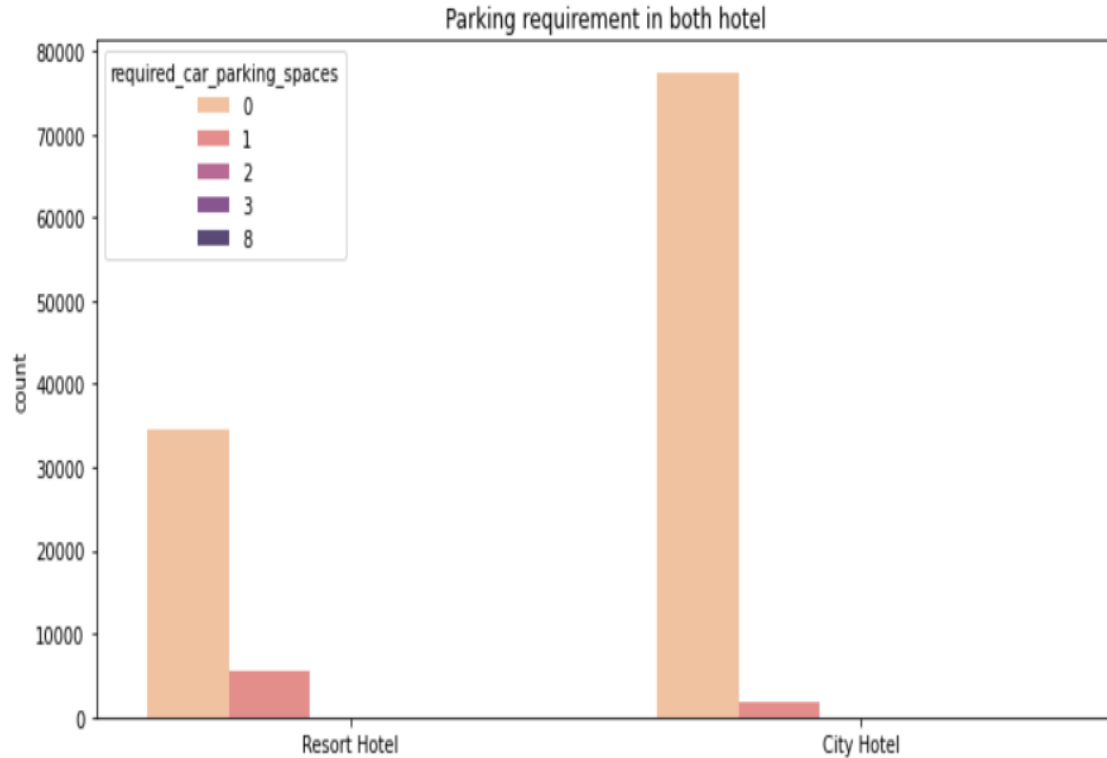
From the above data and graph it is clearly seen at more people are preferring the no deposit as compared to non refund and refundable. And the data set also show that people are preferring the non refund as compared to refundable

*How many reservations were made by repeated guests?

Low number of repeated guests.
A need to target Repeated guests since they have booked before.



*How many people require Parking space?



hotel	required_car_parking_spaces	
City Hotel	0	77404
	1	1921
	2	3
	3	2
Resort Hotel	0	34570
	1	5462
	2	25
	8	2
	3	1

The maximum no of people not require parking but the people who require need 1 parking.
Very less people require more than 2 parking space.

Conclusion

- Majority of the hotels booked are city hotel. Definitely need to spend the most targeting fund on those hotel.
- We also realize that the high rate of cancellations can be due high no deposit policies.
- We should also target months between May to Aug. Those are peak months due to the summer period.
- Majority of the guests are from Western Europe. We should spend a significant amount of our budget on those area.
- Given that we do not have repeated guests, we should target our advertisement on guests to increase returning guests.
- In terms of market segments and distribution channels, TA and TO have shown to be the strongest, followed by the direct channel with the hotel. In the last case, the use of this channel could be incentivized by means of a special offer.

Thank you