Capstone Project-1

Hotel Booking Analysis-EDA

Team Members

1. Abhilasha.M
2. Jatin

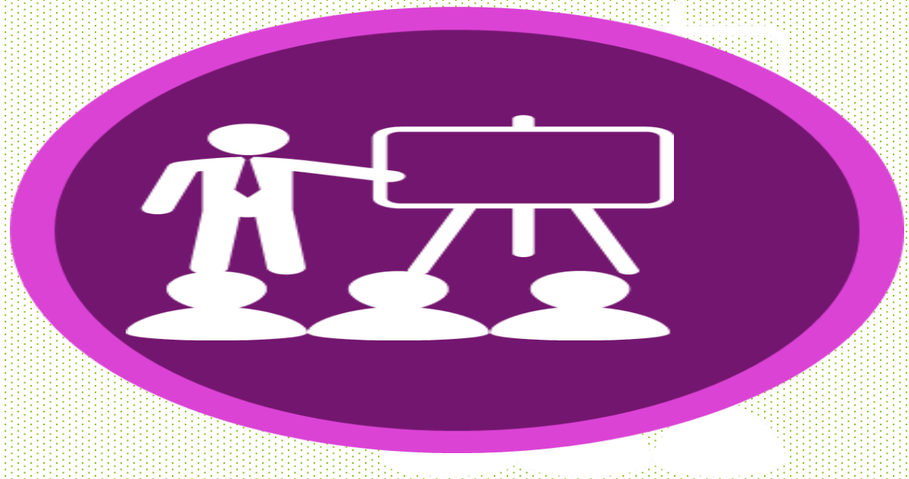


Table of Contents

Problem Statement

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests?

This hotel booking dataset can help you explore those questions!

- 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.
- All personally identifying information has been removed from the data.
- Explore and analyse the data to discover important factors that govern the bookings.

Data Summary

Here we will be doing Exploratory Data Analysis (EDA) on the data set i.e. Hotel Booking analysis.

In this dataset we could see there were totally :

119390 rows of data

32 columns of data

These are the mixture between categorical and numerical data

Our objective is to explore and analyse the data in order to get some insights towards few queries provided to us but not limited to it.

- **In this segment we will understand our dataset variables. Like what does a particular feature means ,what type of data it is etc.**
- **Hotel Booking data set has 32 columns in total. Few columns which are not found significant are dropped.**
- **Let's look at the other columns in detail.**

Understanding the Features..

hotel - Name of hotel (City or Resort)

is_cancelled - Whether the booking is cancelled or not (0 for not cancelled and 1 for cancelled)

lead_time - time (in days) between booking transaction and actual arrival.

arrival_date_year - Year of arrival

arrival_date_month - month of arrival

adults - No. of adults in single booking record.

children - No. of children in single booking record.

babies - No. of babies in single booking record

country - Country of origin of customers (as mentioned by them)

market_segment - What segment via booking was made and for what purpose.(Aviation, Complimentary , Direct, etc.)

distribution_channel - Via which medium booking was made.(Corporate, TA/TO, GDS, Direct)

is_repeated_guest - Whether the customer has made any booking before(0 for No and 1 for Yes)

Understanding the Features..

previous_cancellations - No. of previous cancelled bookings.

previous_bookings_not_cancelled - No. of previous non-cancelled bookings.

reserved_room_type - Room type reserved by a customer

assigned_room_type - Room type assigned to the customer.

booking_changes - Number of booking changes done by customers

deposit_type - Type of deposit at the time of making a booking (No deposit/ Refundable/ No refund)

agent - Id of agent for booking

company - Id of the company making a booking.

days_in_waiting_list **-** No. of days on waiting list.

customer_type - Type of customer(Transient, Group, Transient_Party, Contract.)

adr - Average Daily rate.

total_of_special_requests - total no. of special request.

reservation_status - Whether a customer has checked out or cancelled, or not showed

reservation_status_date - Date of making reservation status

❖ Data Summary continued….

Hotel Booking Analysis

Numerical Data	Categorical Data
Days in waiting List	Previous Bookings
Arrival Day of Month	Previous Cancellations
Babies	Hotel
Children	Is Cancelled
Adults	Distribution Channel
ADR	Agent
Arrival Date – Year	Company
Lead Time	Country
	Market Segment
	Is Repeated Guest
	Reservation Status
	Reserved Room Type

Data Processing

Data Processing and cleaning is necessary in order to get highest quality information in decision making.

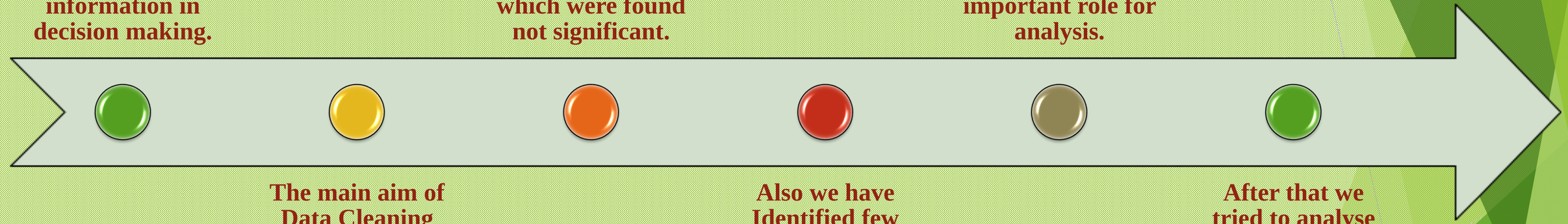
The main aim of Data Cleaning is to identify and remove errors & duplicate data, in order to create a reliable dataset.

In our Hotel Booking dataset we have identified few variables which were found not significant.

Also we have Identified few variables which have Null values and replaced it with appropriate values.

Then we tried to explore all the variables which can play an important role for analysis.

After that we tried to analyse the insights by different types of graphical representation.



EDA

AI

What is EDA?

EDA stands for **"Exploratory Data Analysis "** EDA is applied to **investigate** the data and **summarize** the key **insights**. It will give you the basic understanding of your data, it's distribution, null values and much more.

The following steps are involved in the **process of EDA:**

- **Acquire and loading data**
- **Cleaning dataset**
- **Exploring and Visualizing Data**
- **Analysing relationships between variables**

Approach used for EDA

Acquire and load data : In this segment we have loaded the Hotel Booking ‘csv’ file into the collab notebook and read the csv file.

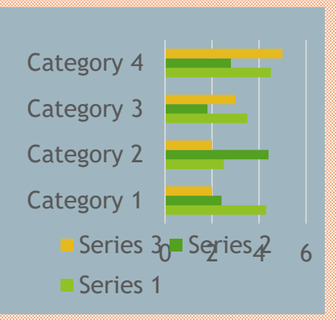
Data Cleaning : In this segment we have removed unwanted columns. Also we have replaced the NULL values to appropriate vales for better understanding of the dataset.

Analysing and visualising data : In this segment we have explored the data and understood the variables which play an important role and the relationship between them. In the coming segment we have tried to answer some hypothetical questions for better understanding of data.

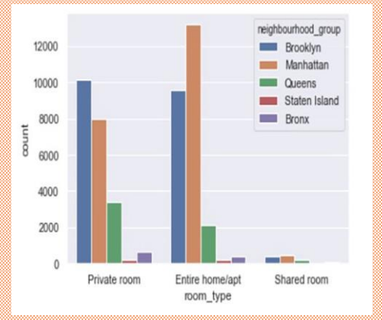
Future Scope : We can see that there are many rooms that are not booked frequently for both the hotels. It might be because of high room rents. If the room rents are managed to be moderate then there can be chances of more number of bookings. There are various other variables that can play an important role in further analysis such as ‘Type of meal’ , ‘booking_changes’, ‘reserved_room type’ etc. and it’s relation with other variables. If we dig deeply various facts can be realized which we were not able to cover during this short duration efficiently.

Graphs used for Data Visualization

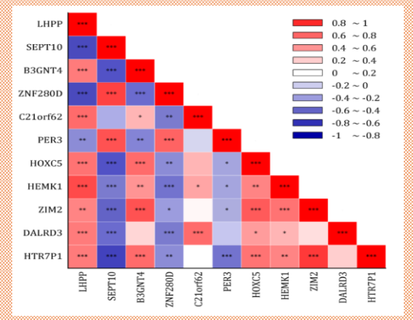
Bar Plot



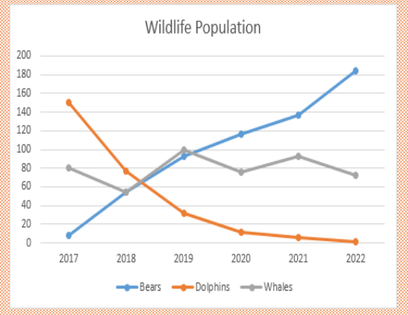
Count Plot



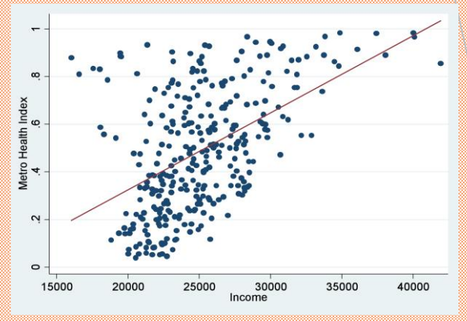
Heat Map Plot



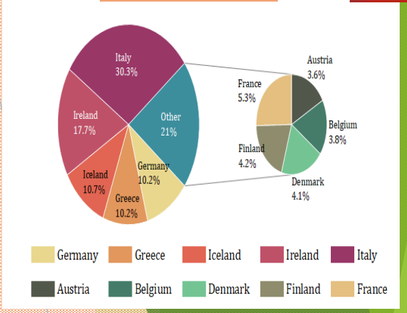
Line Plot



Scatter Plot



Pie Plot



Libraries used for Data Visualization

NumPy

pandas

Matplotlib

klib

seaborn

Exploratory Data Analysis on Hotel Booking Data set

1) From which country maximum number of customers are coming?

2) Give a table of all the countries and their repeated customers showing the highest and the lowest country with repeated customers.

3) In which hotel there is maximum chances of cancellation?

4)What is the data for the repetition of guests for both of the hotels?

5) Which hotel has longer waiting time?

6) Find the number of customers who booked Resort hotel and City and not cancelled booking further.

7) Find the first three months with maximum number of bookings and average rent across all the months for both Resort Hotel and City hotel.

8) Find out the average rent and waiting time for different types of customers for City hotel and Resort Hotel.

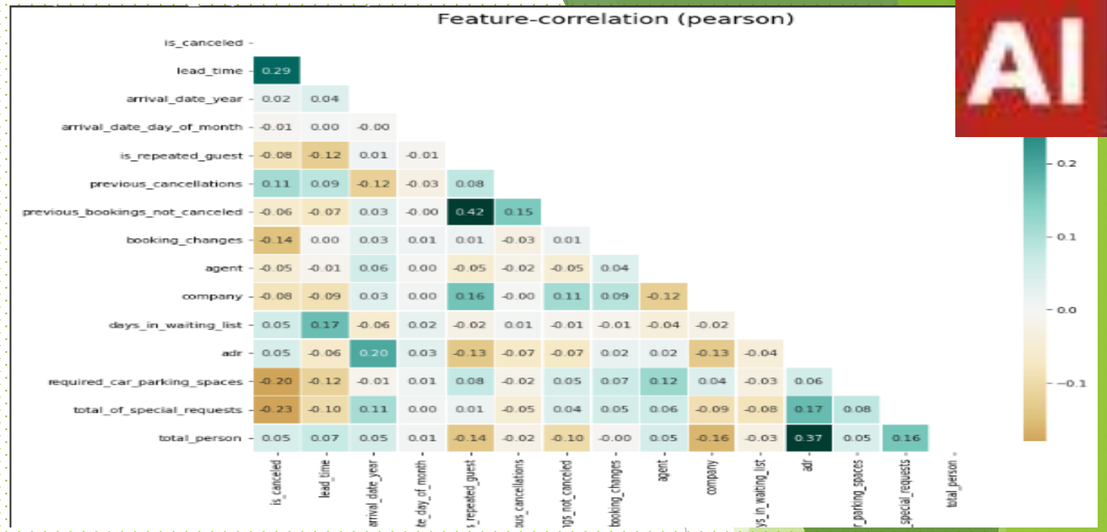
9) Find the Agent who has done most number of bookings for Resort hotel and City Hotel.

10) Find out the count of customers who booked tickets through various modes and through which mode highest booking was made.

11) Find the most popular Rooms booked and their respective rents.

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

EDA



Now let us look at some variables with null values:

- We can see from the data that variables **"company"** and **"agent"** have the highest null values of each **94.3%** and **13.6%** respectively.
- The other columns namely **"child"** and **"country"** have comparatively less number of **Null values.** These have been replaced with appropriate values.

Now, Let's check the correlation between the Variables:

1. Columns 'adr' and 'total_person' are correlated
2. There is a high correlation between 'is_repeated_guests' and ' previous_booking_not_cancelled'
3. There is also a high correlation between columns is_cancelled' and 'lead_time'.

EDA.....

	is_canceled	lead_time	arrival_date_year	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	agent	company	days_in_waiting_list	adr	total_of_special_re	pers
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.0000
mean	0.370416	104.011416	2016.156554	0.031912	0.087118	0.137097	0.221124	74.828319	10.775157	2.321149	101.831122	0.571363	1.9682
std	0.482918	106.863097	0.707476	0.175767	0.844336	1.497437	0.652306	107.141953	53.943884	17.594721	50.535790	0.792798	0.7223
min	0.000000	0.000000	2015.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-6.380000	0.000000	0.0000
25%	0.000000	18.000000	2016.000000	0.000000	0.000000	0.000000	0.000000	7.000000	0.000000	0.000000	69.290000	0.000000	2.0000
50%	0.000000	69.000000	2016.000000	0.000000	0.000000	0.000000	0.000000	9.000000	0.000000	0.000000	94.575000	0.000000	2.0000
75%	1.000000	160.000000	2017.000000	0.000000	0.000000	0.000000	0.000000	152.000000	0.000000	0.000000	126.000000	1.000000	2.0000
max	1.000000	737.000000	2017.000000	1.000000	26.000000	72.000000	21.000000	535.000000	543.000000	391.000000	5400.000000	5.000000	55.0000

The above figure shows the summary of all the columns.

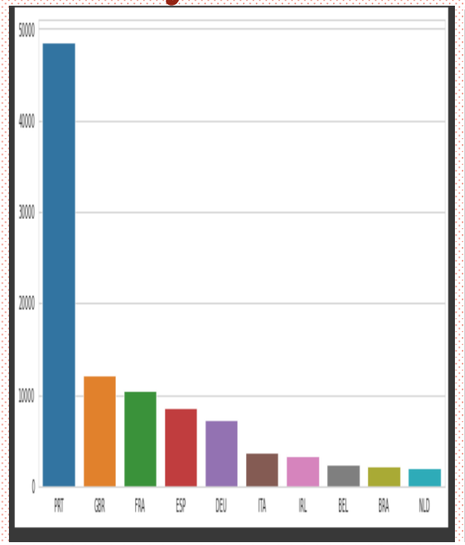
We can see that the minimum value for the variable **“adr”** which is average daily rental for a particular room **is negative** which is a bit strange.

We can also observe that for the variable **“arrival_date_month” the min value is 2015 and max is 2017** which means the bookings are distributed between the years from 2015 to 2017.

We can see that **75%** of bookings have lead time i.e the time (in days) between booking transaction and actual arrival are within the value **160.**

EDA.....

1. Maximum customers are from which country?

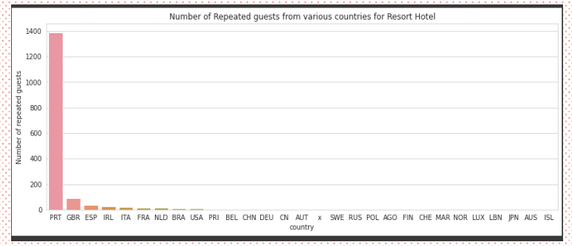


	country
PRT	48590
GBR	12129
FRA	10415
ESP	8568
DEU	7287
ITA	3766
IRL	3375
BEL	2342
BRA	2224
NLD	2104

The above bar plot shows the data of various countries and the number of people of each country coming to hotel. Also on the right hand side there is data of various countries and respective number of bookings.

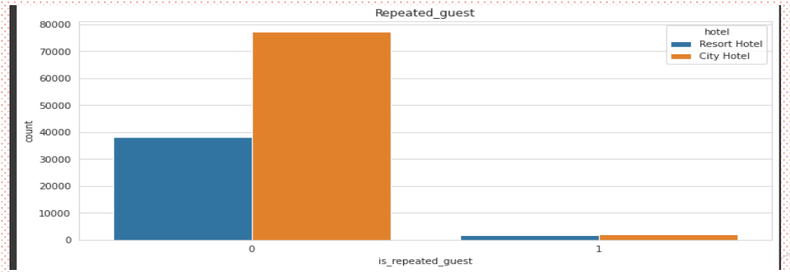
From the data we can see that maximum number of customers are coming from the country Portugal, Great Britain France and Spain..

2. Which countries have the most repeated customers?



From the above bar plot we can see that maximum number of repeated guests are coming from the country **Portugal for both City and Resort hotels.**

3. What % of customers have repeatedly visited the hotels?



Calculating percentage of repeated guest

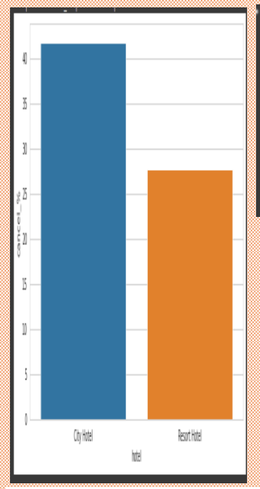
```
((df['is_repeated_guest']!=0).sum()/df['is_repeated_guest'].value_counts().sum())*100
```

```
3.191222045397437
```

From the data we can observe that around 3.2% guests are repeatedly visiting both the hotels

EDA..

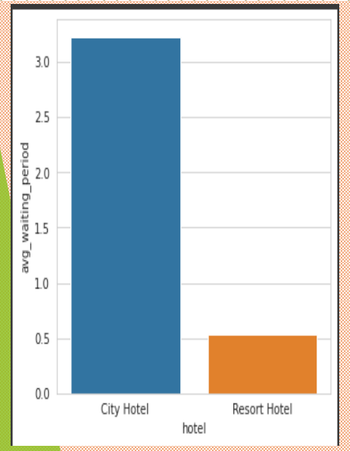
3. In which hotel there is maximum chances of cancellation?



hotel	total_cancelled_bookings	total_bookings	cancel_%
City Hotel	33102	79330	41.73
Resort Hotel	11122	40060	27.76

Here we can see that 41% of City Hotel and 27% of Resort Hotel were cancelled. Hence City Hotel has maximum cancellations compared to Resort hotel.

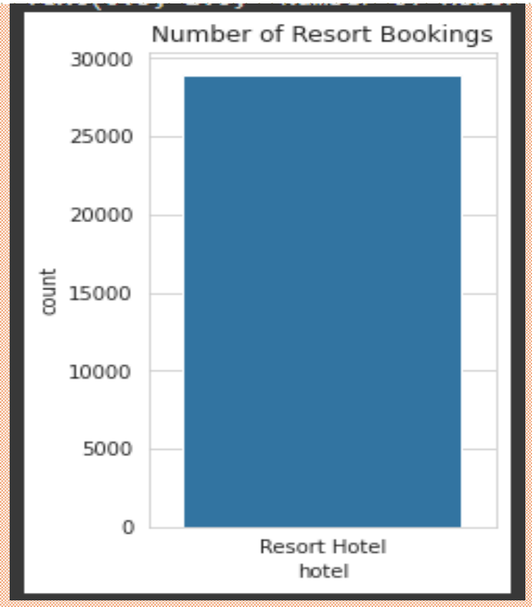
4. Which Hotel has longer waiting time?

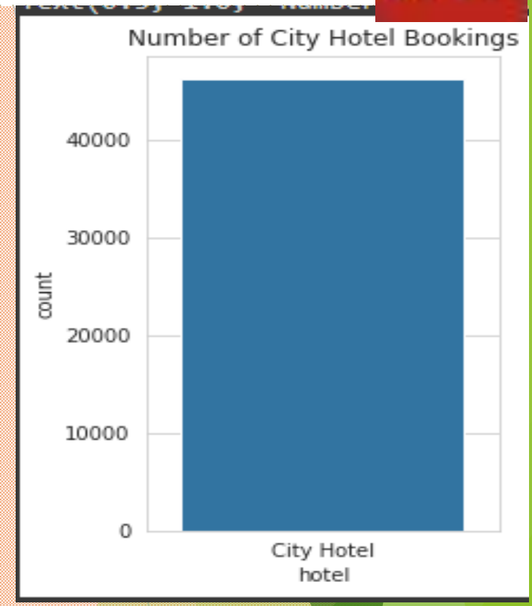


	hotel	avg_waiting_period
0	City Hotel	3.226774
1	Resort Hotel	0.527758

City hotel has significantly longer waiting time, hence City Hotel is much busier than Resort Hotel

5. Number of customers who booked Resort hotel and City and not cancelled booking further.



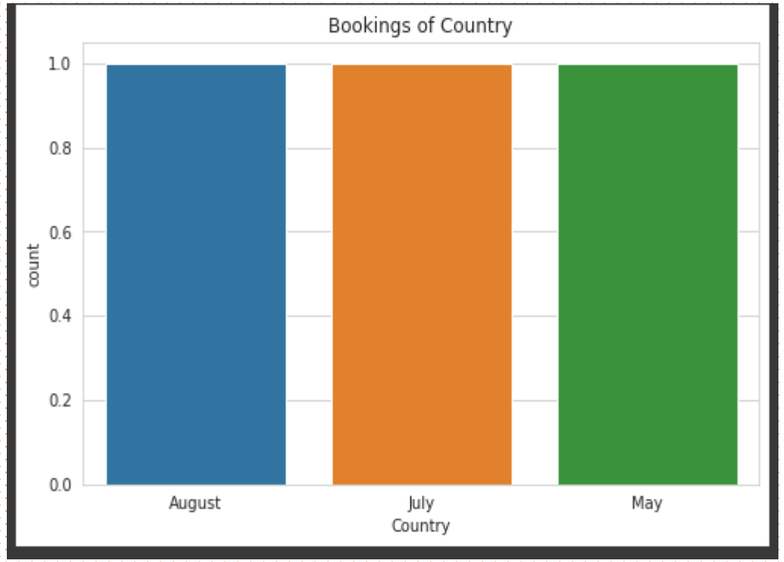


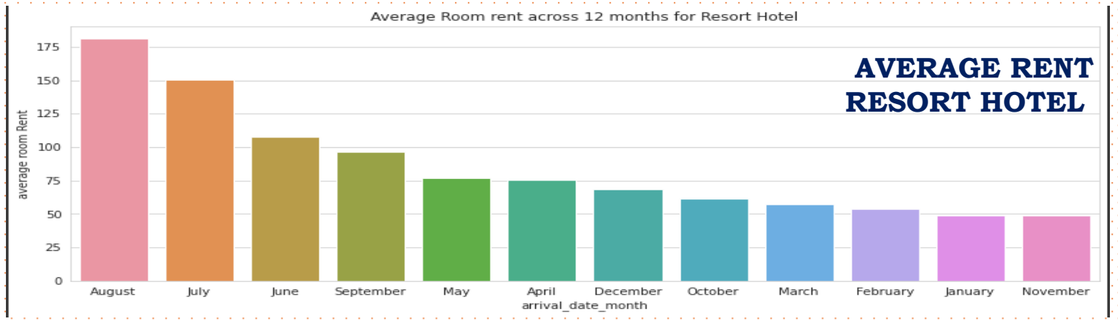
There are 46228 Customers who booked Resort hotel and not cancelled it further also there are 28938 Customers who booked Resort hotel and not cancelled it further. Hence we can say that City Hotel has more number of bookings than Resort Hotel.

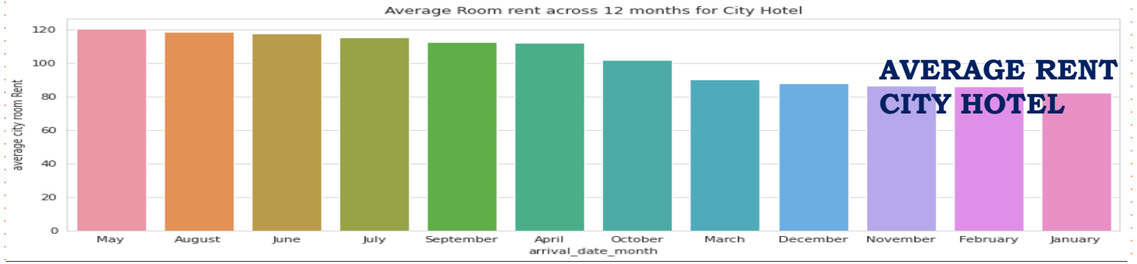
EDA.....

6. Which are the top three months with maximum number of bookings?

7. What is the average rent across all the months for both Resort Hotel and City hotel?



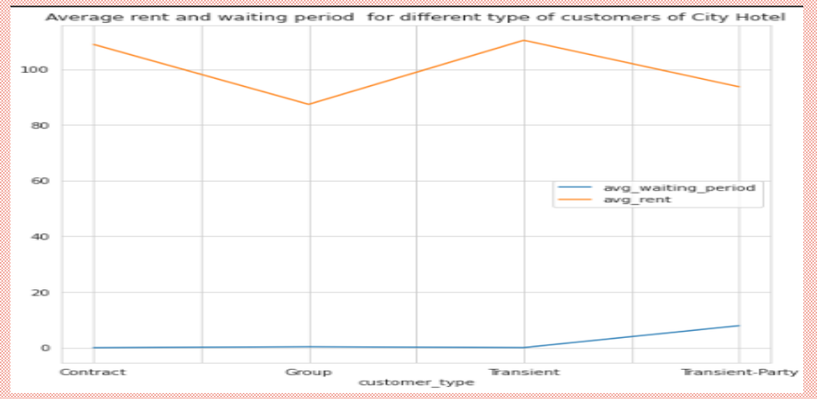




For both City Hotel and Resort Hotel the most busiest months are July, May and August. Also we can see that average rent is maximum in the month of August for Resort Hotel. For City Hotel the average rent is maximum in the month of May. Also the average total revenue for Resort Hotel is 1026.79 whereas for City Hotel the average total revenue for a year is 1231.1. Hence we can clearly infer that City Hotel is making more profit than Resort Hotel.

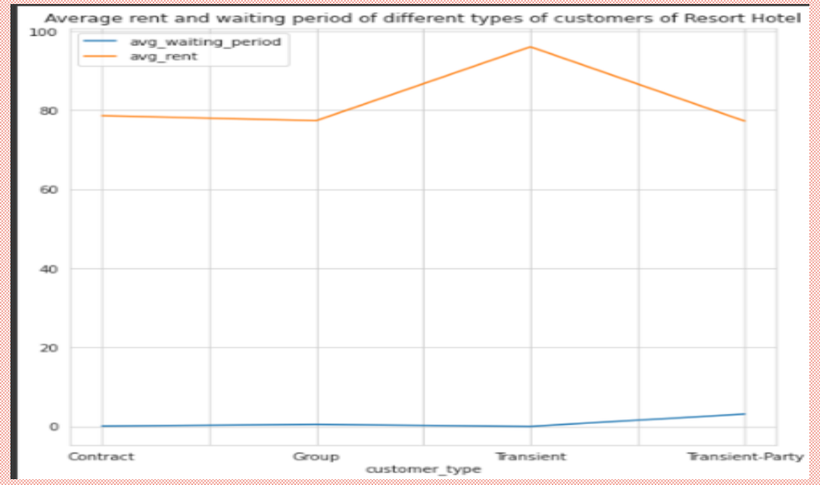
EDA.....

8. What is the average rent and waiting time for different types of customers for City Hotel & Resort Hotel?



For City Hotel customers, For Contract ,Transient and Group type customers the waiting period is zero whereas for Transient-Party it is 7.88 So it is clear that if the booking was made as 'Contract' or 'Group' or 'Transient' the rooms will be booked immediately without waiting time.

The average rent is maximum for 'Transient' type of customers and minimum for 'Transient-Party' type of customers.

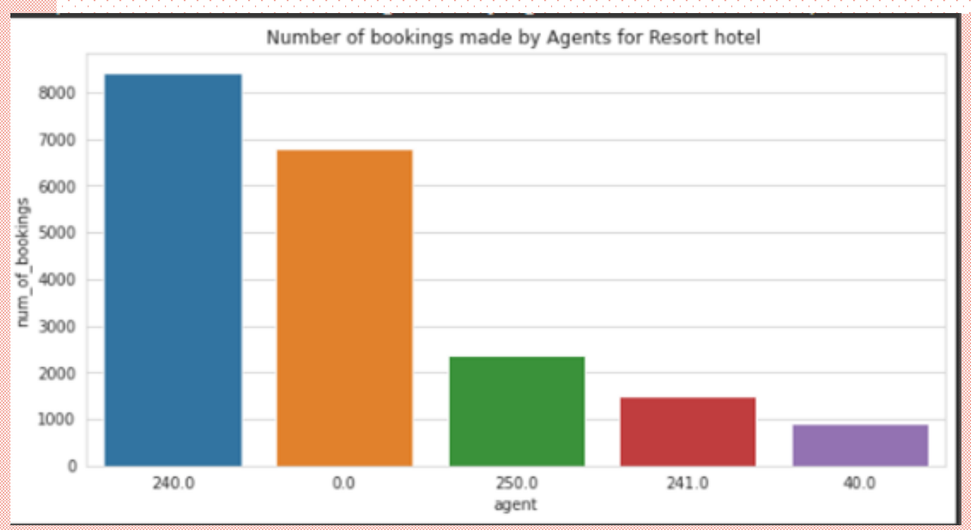


For Resort Hotel customers for Contract ,Transient and Group type customers the waiting period is zero where as for Transient-Party it is 3.12, So it is clear that if the booking was made as 'Contract' or 'Group' or 'Transient' the rooms will be booked immediately without waiting time.

The average rent is maximum for 'Transient' type of customers and minimum for 'Transient-Party' type of customers.

EDA.....

9. Agent who have done most number of bookings for Resort hotel and City Hotel…

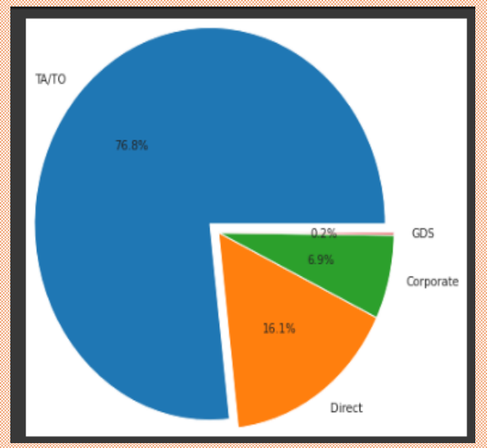


Among the many agents who have booked tickets for Resort Hotel and City Hotel we have shown top five agents.

Agent numbered 240 has done maximum number of bookings for Resort Hotel. Agent numbered 9 has done maximum number of bookings for City Hotel.

EDA.....

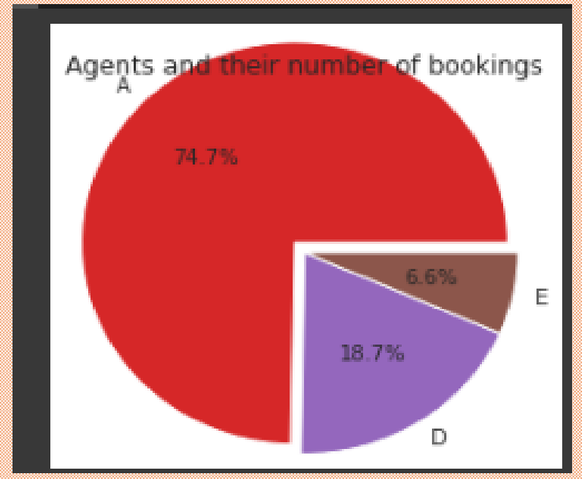
10. What is the count of customers who booked tickets through various modes and through which mode highest booking was made?



	distribution_channel	Number of booking
3	TA/TO	57718
1	Direct	12088
0	Corporate	5203
2	GDS	156
4	Undefined	1

There are various modes in which bookings are made ‘GDS’ , ‘TA/TO’ , ‘Corporate’ , ‘Direct’. In these the maximum number of bookings are made by the mode ‘TA/TO’ whereas very less bookings were made by the mode ‘GDS’. Hence TA/TO is very popular mode to book tickets for hotels.

11. Which are the most popular type of R booked and their respective rents?

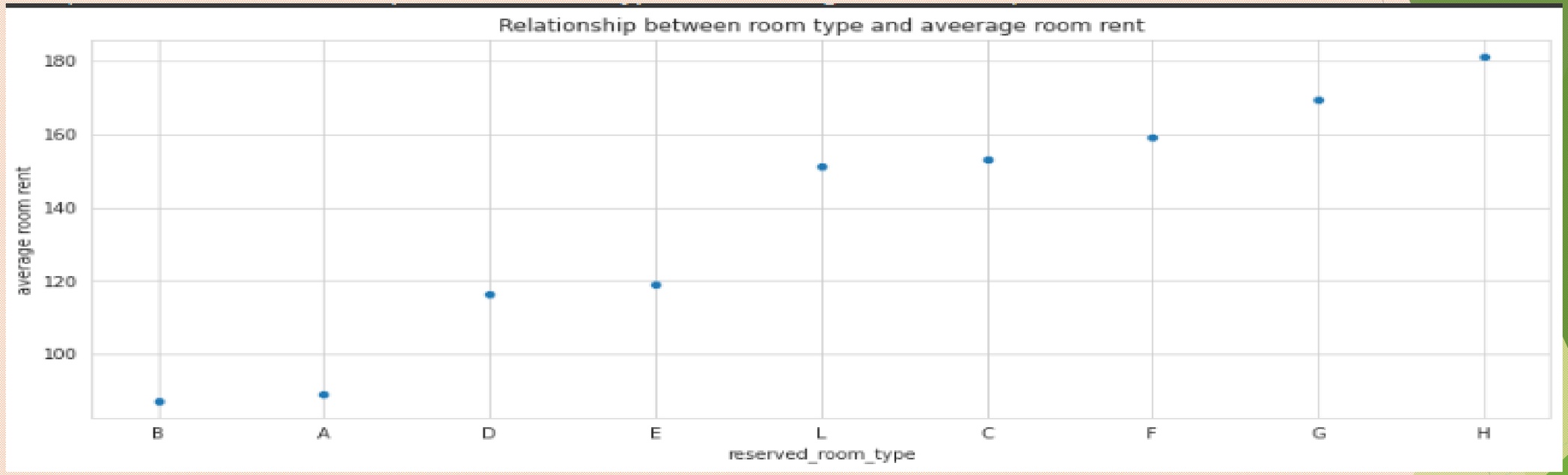


	room type	number_of_bookings
0	A	52364
1	D	13099
2	E	4621

There are numerous types of rooms for both the hotels. However here we are trying to find rooms with maximum number of bookings. Here room type ‘A’ , ‘D’ and ‘E’ have highest number of bookings.

EDA.....

12. How are the types of the room booked, dependent on average rent?



It is observed that the most common rooms booked room types 'A' , 'D' and 'E' have minimal or moderate room rents whereas rarely booked rooms have high room rent. Hence if the room rent is minimal or moderate more number of bookings can be expected for both the hotels.

Conclusion

Most of the customers come from Portugal, Great Britain, France and Spain.

The maximum of repeated customers have come from the country Portugal represented as 'PRT' for both of the hotels.

Almost 41% of City Hotel and 27% of Resort Hotel were cancelled. So, we can clearly see that maximum percentage of people cancelled City Hotel.

Repeated customers is approx. 3.2% for both of the hotels.

City hotel has significantly longer waiting time, hence City Hotel is much busier than Resort Hotel.

There are 46228 Customers who booked Resort hotel and not cancelled it further also there are 28938 Customers who booked City hotel and not cancelled it further.

From the data we can see that August, July and May are the most busiest moths for Resort hotel and City hotel.

The average rent for Resort hotel for all the 12 months is highest in the month of August and lowest in the month of November.

a. Also for City hotel the average rent is at its peak in the month of May and very low in the month of January.

b. Also we have seen that City hotel is making more revenue than Resort Hotel

Conclusion

Regarding Resort Hotel for 'Contract' , 'Transient' and 'Group type customers the waiting period is zero where as for Transient-Party it is 3.12. So it is clear that if the booking was made as 'Contract' or 'Group' or 'Transient' the rooms will be booked immediately without waiting time.

The average rent is maximum for 'Transient' type of customers and minimum for 'Transient-Party' type of customers.

Regarding City Hotel for 'Contract' , 'Transient' and 'Group type' customers the waiting period is zero where as for 'Transient-Party' it is 7.88 So it is clear that if the booking was made as 'Contract' or 'Group' or 'Transient' the rooms will be booked immediately without waiting time.

The average rent is maximum for 'Transient' type of customers and minimum for 'Group' type of customers.

The above data shows the top five agents who have made maximum number of bookings in Resort hotel. We can see that Agent number 240 has made maximum number of booking.

Agent numbered '9'has booked highest number of rooms for City Hotel.

We can see from the above data that the maximum bookings were done by the Agent with agent numbered '9'. The minimum booking was made by the agent numbered '6' for City Hotel.

We can see from the above data maximum number of bookings by customers were made through the channel TA/TO. Whereas very few booked through the channel GDS. Hence TA/TO is very popular channel to book tickets for hotel.

Out of all the agents who were booking rooms ,we can see that agent numbered '240' has booked highest number of rooms for Resort Hotel. Agent numbered '9'has booked highest number of rooms for City Hotel.

The above data shows how the types of room booked and the rent of the rooms are related. It is observed that the most common rooms booked namely room types 'A' , 'D' and 'E' have minimal or moderate room rents whereas rarely booked rooms have high room rent. Hence if the room rent is minimal or moderate more number of bookings can be expected

Challenges Faced

- The data set was pretty huge. Hence analyzing and exploring the data was time consuming.
- The data set had many futuristic variables. Many columns in the data set like ‘children’ , ‘company’ etc. have null values. Hence for better understanding we had to replace it with appropriate values.
- Since the data is pretty huge, it took pretty more time to compute the results.

“Thank You”