EDA Capstone Project

HOTEL BOOKING ANALYSIS

by

SHUBHAM DANDNAIK

Steps Involved

- Problem Statement
- Data Exploration
- Data Cleaning
- Exploratory Data Analysis
- Visualizing the Data
- Conclusion

Problem Statement

- We have Hotel Booking dataset for year 2015 , 2016 and 2017 which contain bookings of various type of hotels. We have to find out what the various factors affecting the bookings and also what majors to be taken to get more bookings.

- By performing EDA on the given dataset we will find out the the answers for the various type of questions also deriving the meaningful insights from the given dataset which would help us to improve further.

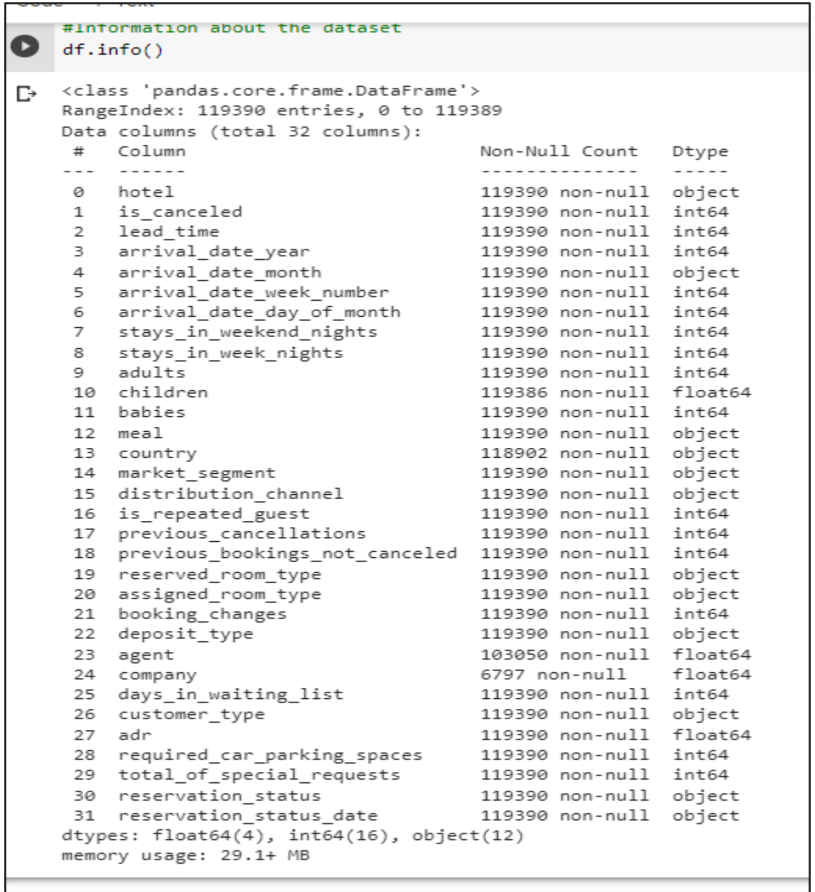
Dataset Information

- The dataset contains 119390 rows and 32 columns

- Hotel
- is_canceled
- lead_time
- arrival_date_year
- arrival_date_month
- arrival_date_week_number
- arrival_date_day_of_month
- stays_in_weekend_nights
- stays_in_week_nights
- adults
- children
- babies
- meal
- country
- market_segment
- distribution_channel_status_date
- is_repeated_guest
- previous_cancellations
- previous_bookings_not_canceled
- reserved_room_type
- assigned_room_type
- booking_changes
- deposit_type
- agent
- company
- days_in_waiting_list
- customer_type
- adr
- required_car_parking_spaces
- total_of_special_requests
- reservation_status
- reservation

Data Exploration

Dataset Info

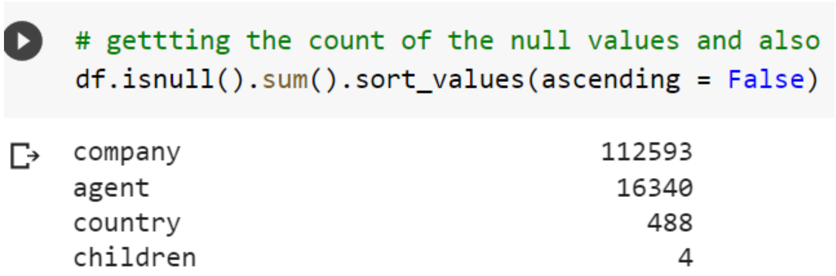


```
#Information about the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
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  
 29  total_of_special_requests       119390 non-null  int64  
 30  reservation_status              119390 non-null  object 
 31  reservation_status_date         119390 non-null  object 
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

Dealing with null values

Finding out the number of columns which contains null values

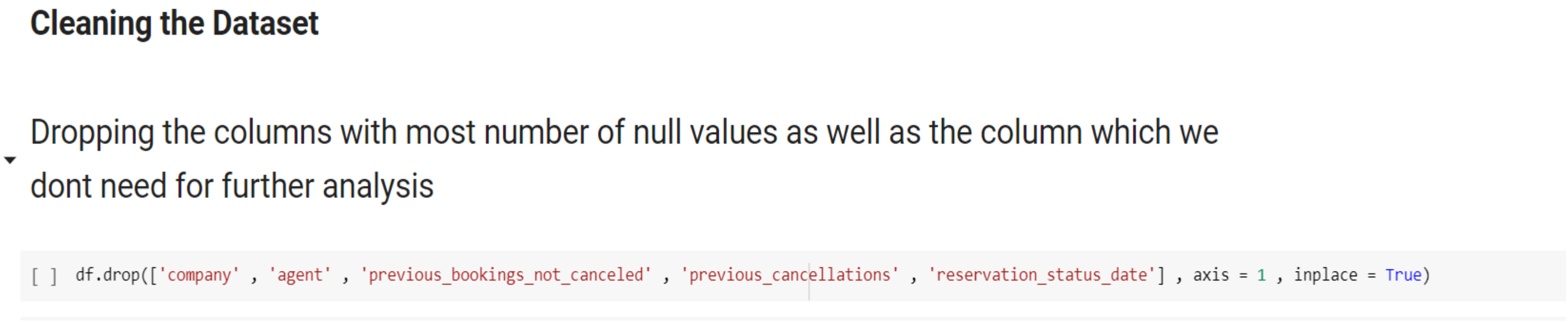


```
# gettting the count of the null values and also
df.isnull().sum().sort_values(ascending = False)
```

```
company                     112593
agent                        16340
country                        488
children                         4
```

Only these 4 columns has the null values

Dropping the columns which contains the most number of null values also droping the columns which is not helpful for our EDA

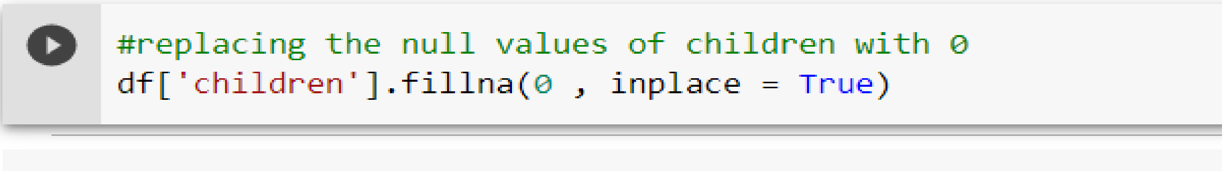


Cleaning the Dataset

Dropping the columns with most number of null values as well as the column which we dont need for further analysis

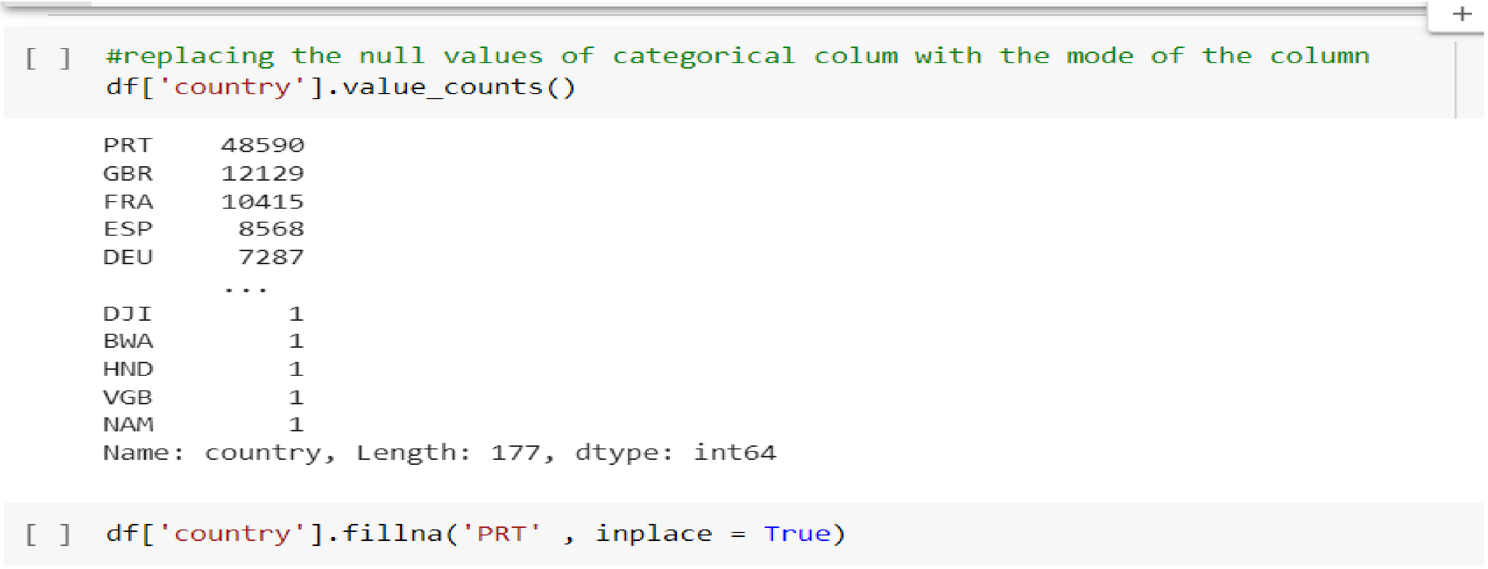
```
[ ] df.drop(['company' , 'agent' , 'previous_bookings_not_canceled' , 'previous_cancellations' , 'reservation_status_date'] , axis = 1 , inplace = True)
```

Filling the null values of numeric column Children with 0



```
#replacing the null values of children with 0
df['children'].fillna(0 , inplace = True)
```

Filling null values of categorical column Country with its mode

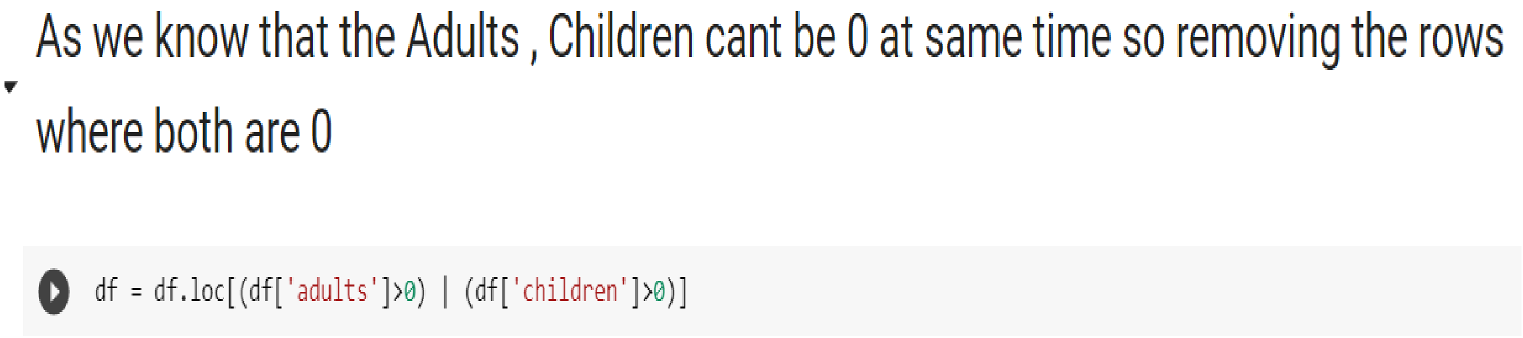


```
[ ]  #replacing the null values of categorical colum with the mode of the column
     df['country'].value_counts()
```

```
PRT     48590
GBR     12129
FRA     10415
ESP      8568
DEU      7287
        ...
DJI         1
BWA         1
HND         1
VGB         1
NAM         1
Name: country, Length: 177, dtype: int64
```

```
[ ]  df['country'].fillna('PRT' , inplace = True)
```

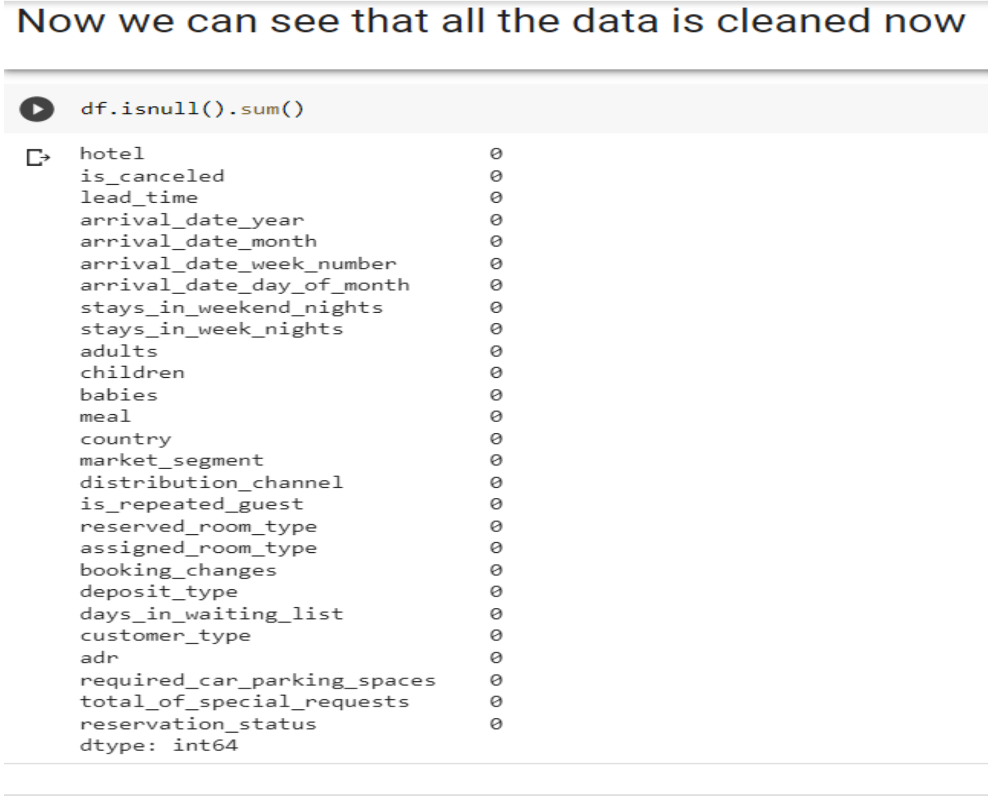
Dropping the rows which contains the no of adults and no of children is equal to 0 at same time



As we know that the Adults , Children cant be 0 at same time so removing the rows where both are 0

```
df = df.loc[(df['adults']>0) | (df['children']>0)]
```

Cleaned Dataset



Now we can see that all the data is cleaned now

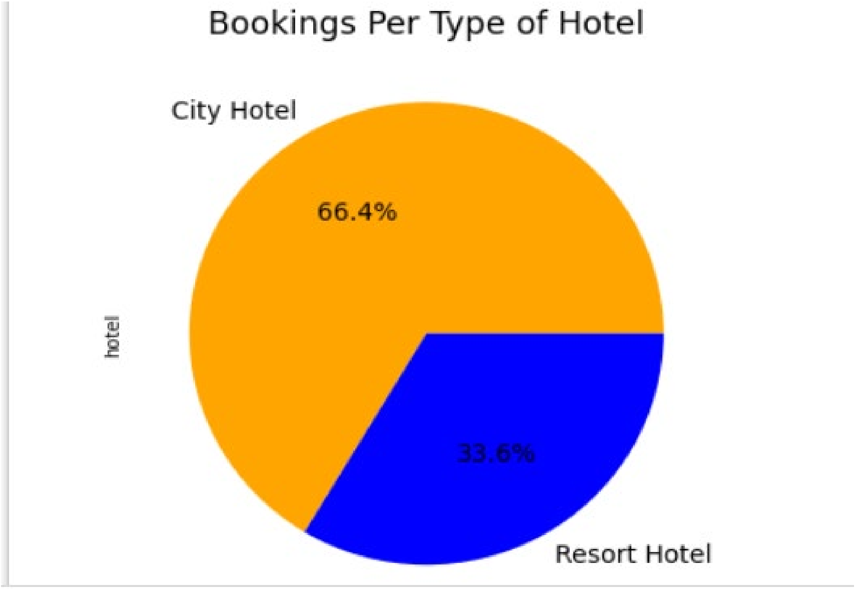
```
df.isnull().sum()
```

```
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
reserved_room_type              0
assigned_room_type              0
booking_changes                 0
deposit_type                    0
days_in_waiting_list            0
customer_type                   0
adr                             0
required_car_parking_spaces     0
total_of_special_requests       0
reservation_status              0
dtype: int64
```

Now we can see that the above dataset is now cleaned and ready for the analysis

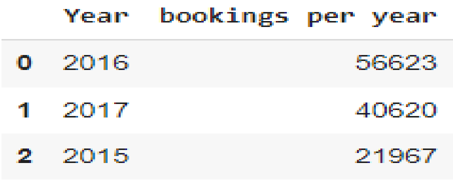
Data Analysis & Data Visulization

Which Hotel is most Prefered by the customers ?

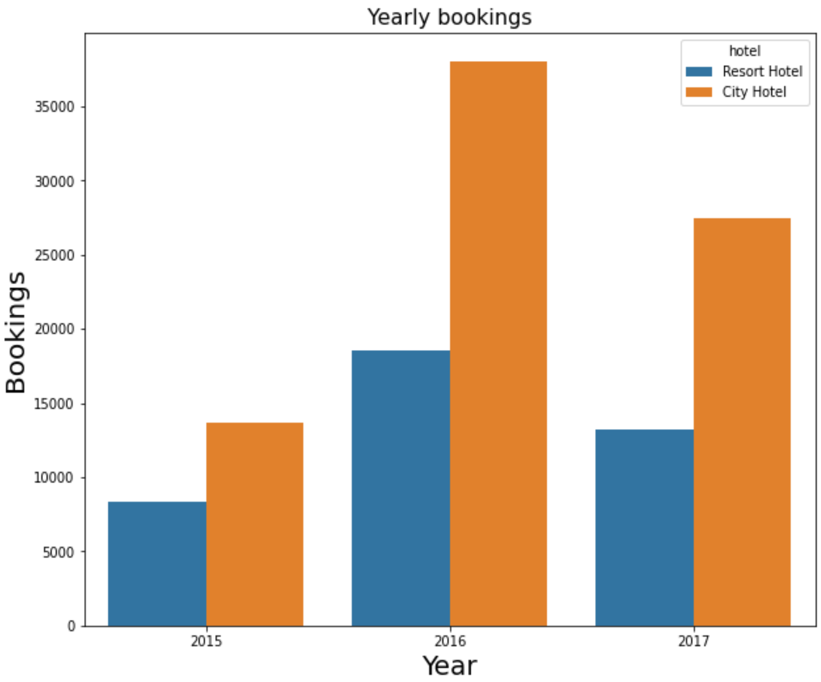


- City Hotel is booked more than Resort Hotel.
- 66.4% bookings made for the City Hotel.
- 33.6% bookings made for the Resort Hotel.

Bookings per Year as per Hotel Type

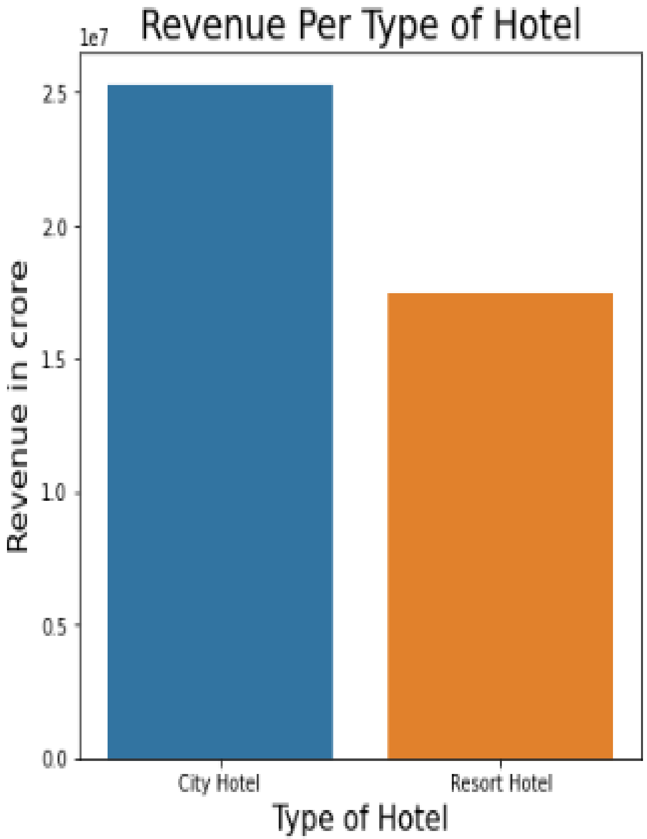


	Year	bookings per year
0	2016	56623
1	2017	40620
2	2015	21967



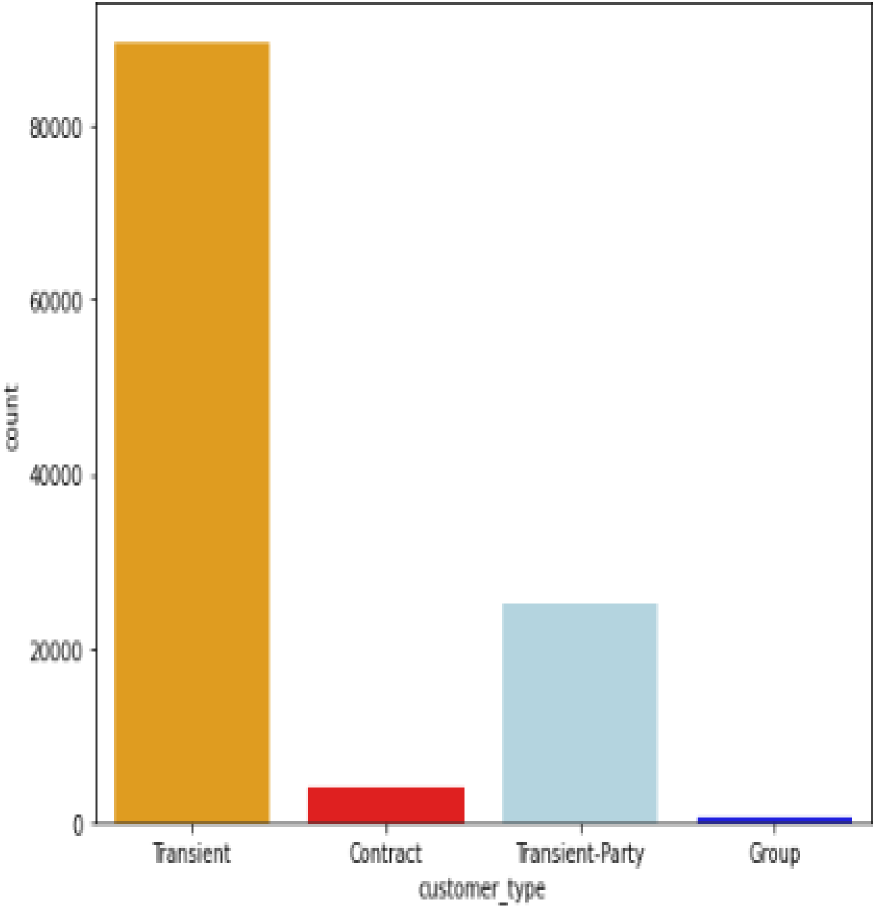
- ➢ Total Bookings made in year 2015 = 21967
- ➢ Total Bookings made in year 2016 = 56623
- ➢ Total Bookings made in year 2017 = 40620
- ➢ Bookings made in year 2016 were more than other years for both type of Hotels

Total revenue Generated by the Hotels



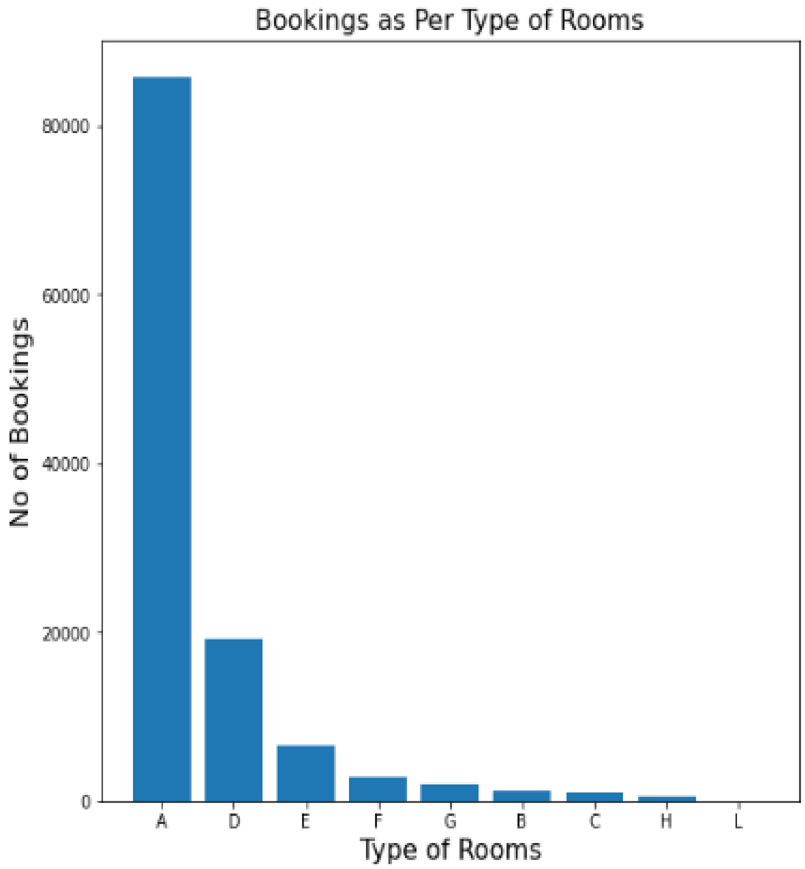
- Revenue generated by City Hotel = 25270401
- Revenue generated by Resort Hotel = 17443747
- City Hotel has generated more revenue than Resort Hotel

Bookings As per Customer Type



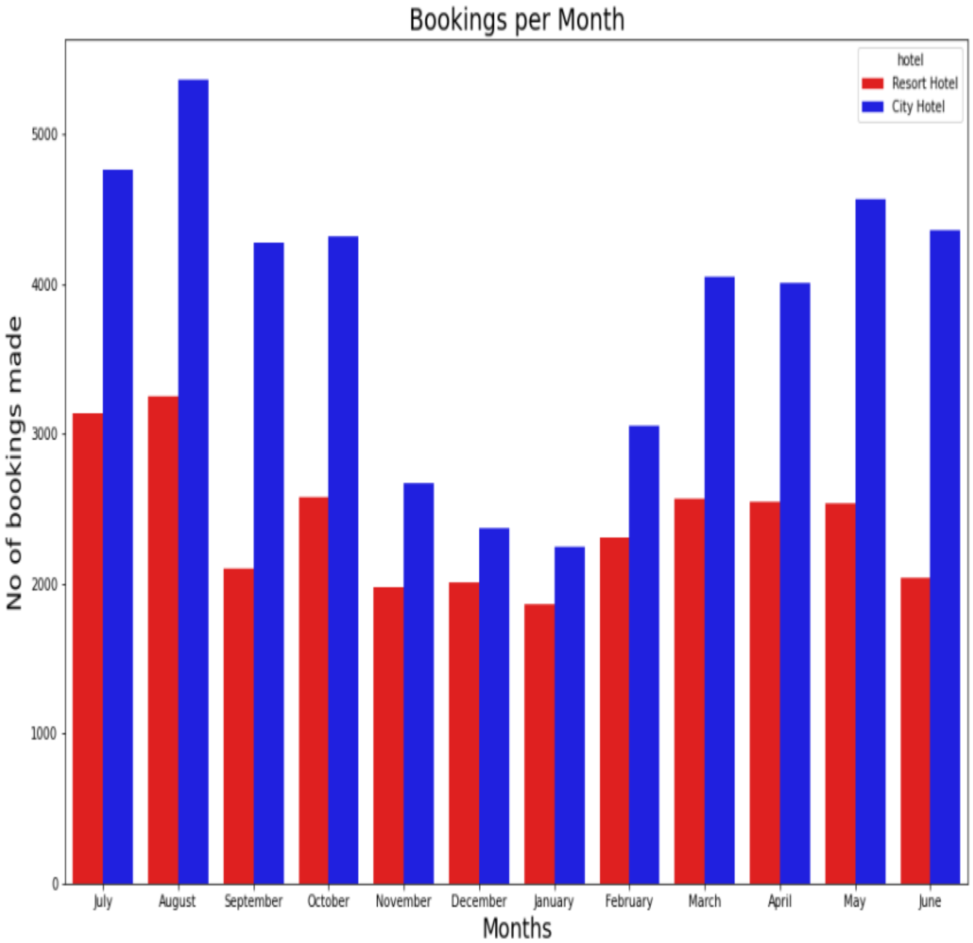
- Booking made by Transient Customer = 89476
- Bookings made by Transient-Party = 25088
- Bookings made by Contract = 4072
- Bookings made by Group = 574
- Bookings by Transient customers were much higher than the others

Room type preferred by the customers



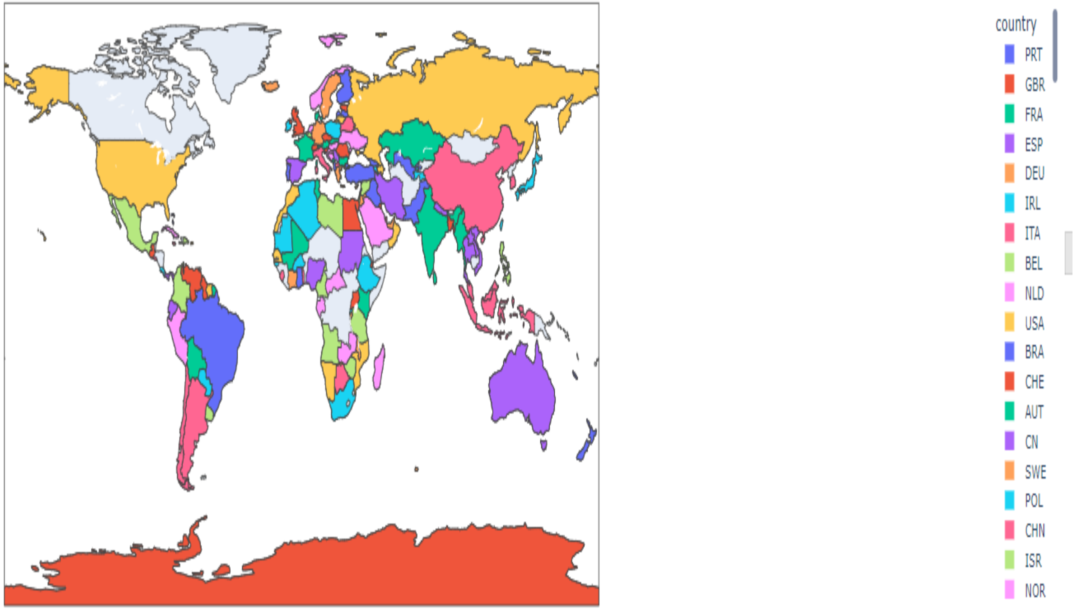
- A type room is most preferred by the customers.
- D type room is the second most preferred by the customers
- E type room is the third most preferred by the customers
- Other type rooms were booked very less no of times

Bookings made per Month

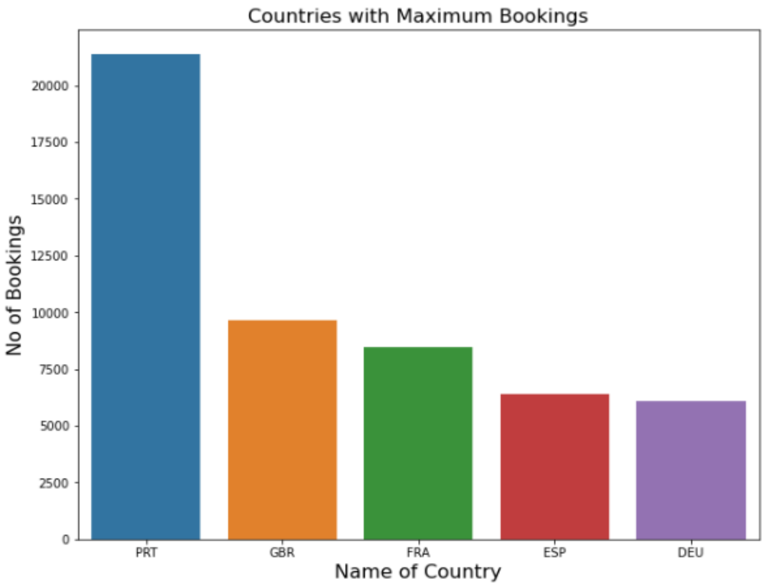


- ➢ We can see that highest bookings were made in month of July , August for both type of Hotels.
- ➢ May , June , September and October has almost same number of bookings for City Hotel.
- ➢ March , April and May has almost same bookings for Resort Hotel.
- ➢ January has the lowest bookings for both type of Hotels

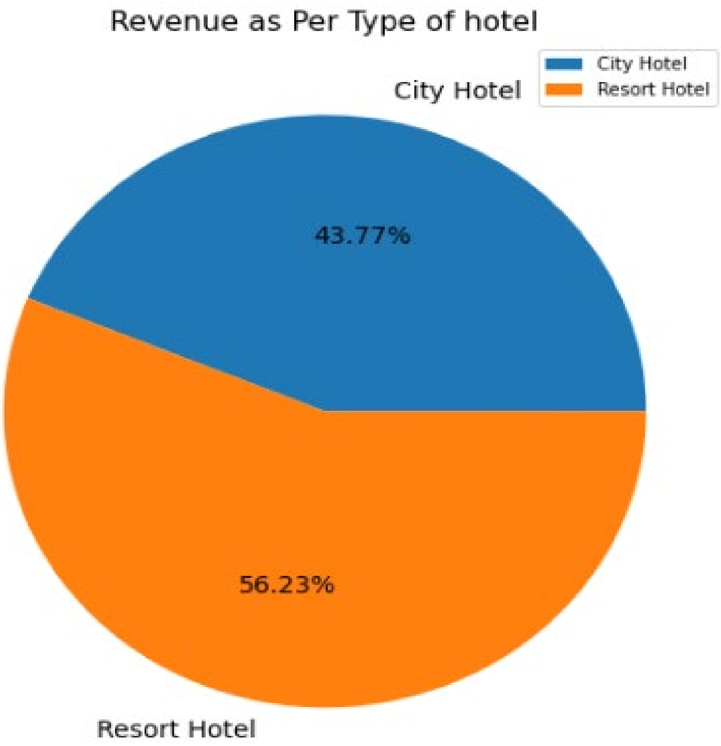
Successive Bookings as per country



Top 5 Countries which has most successive bookings



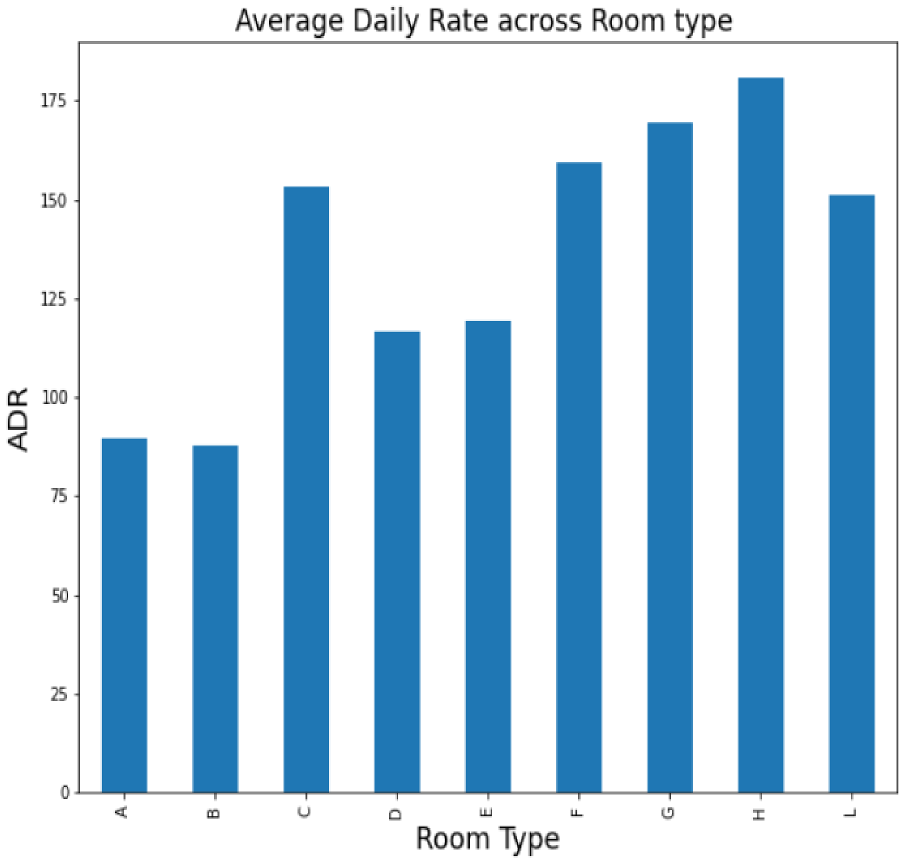
Average Revenue



- Average revenue is Higher for Resort Hotels compared to City Hotel.

- As ADR is high for Resort Hotel average revenue is high for Resort Hotel.

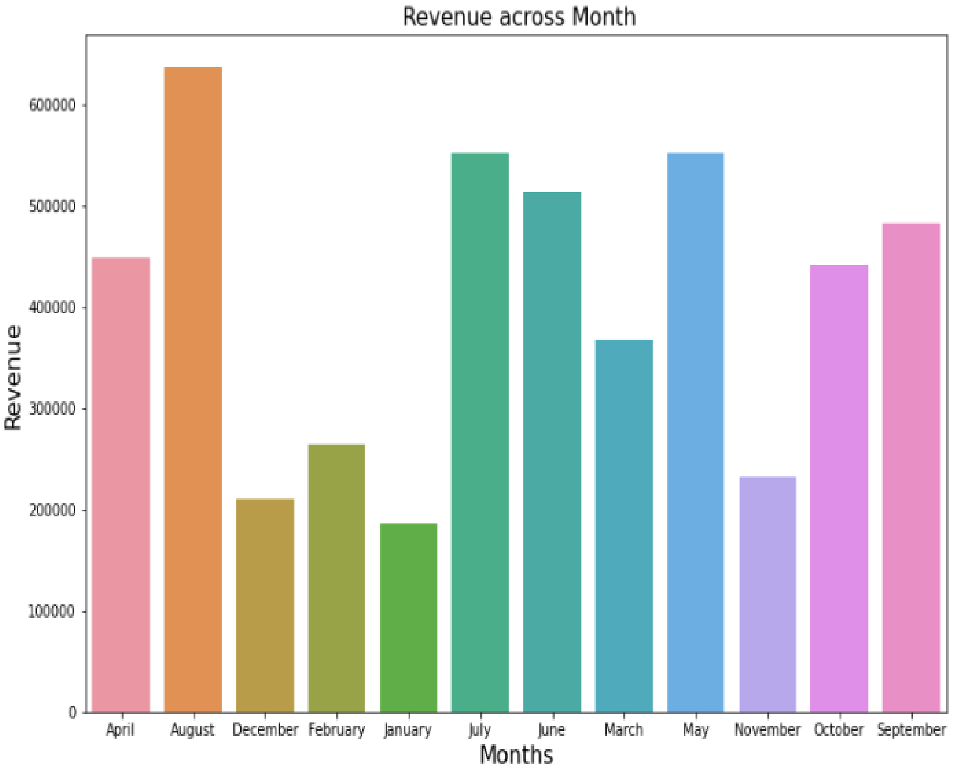
Average Daily Rate Across all Room Type



- G and H type of room has the highest ADR with respect to others

- A and B has the lowest ADR

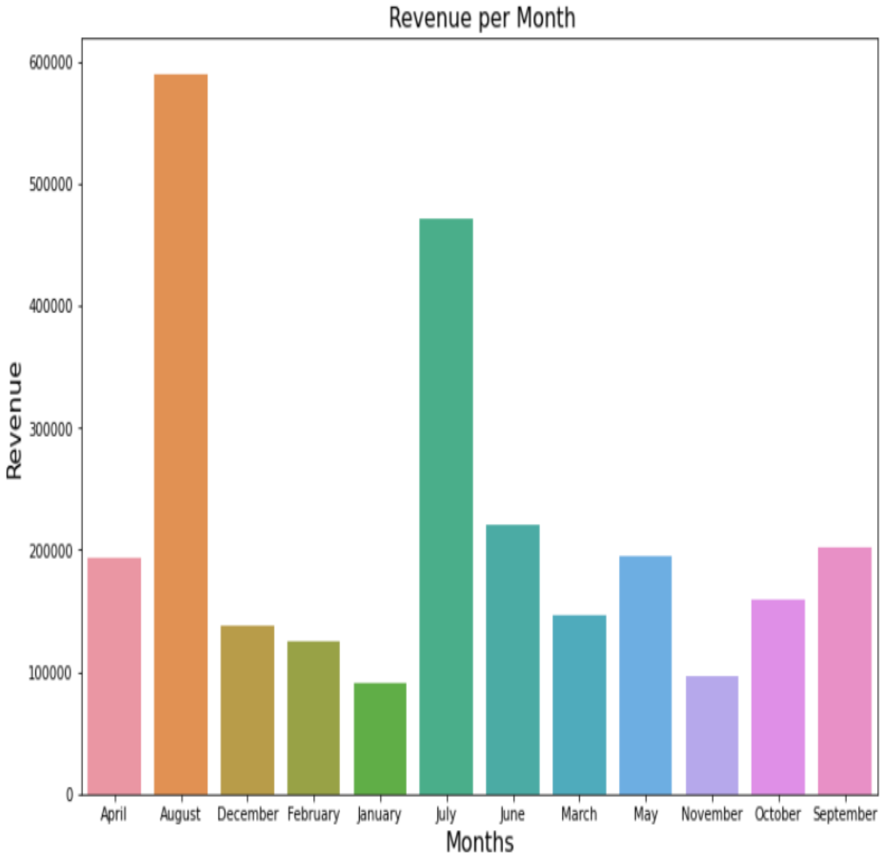
Revenue Across All Months for City Hotel



- ➢ August Month has generated the highest revenue.

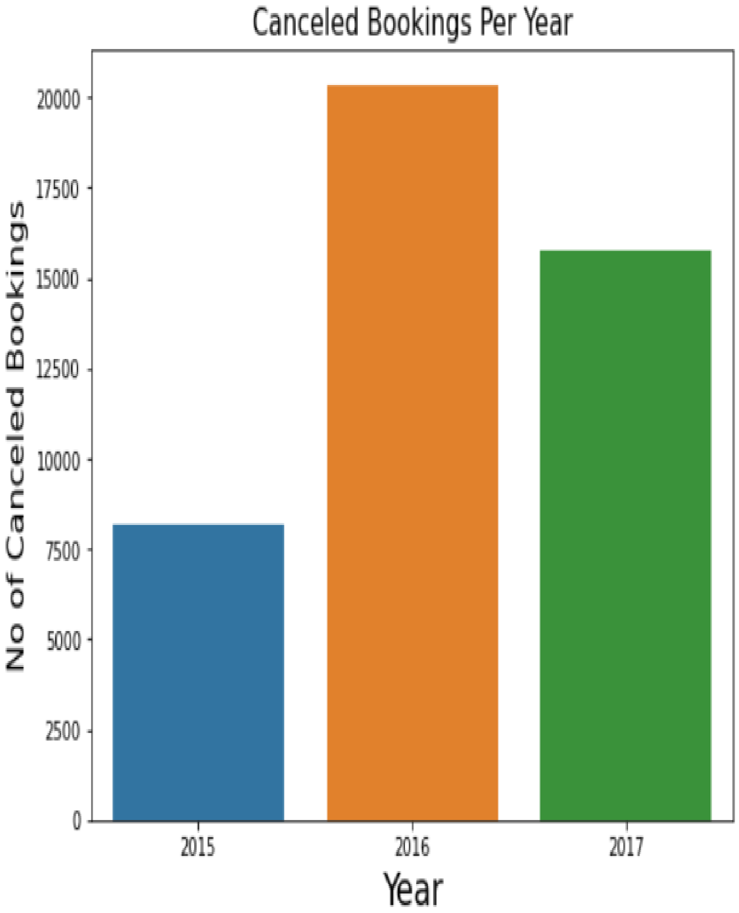
- ➢ May and July has generated almost same revenue.

Revenue across all months for Resort Hotel



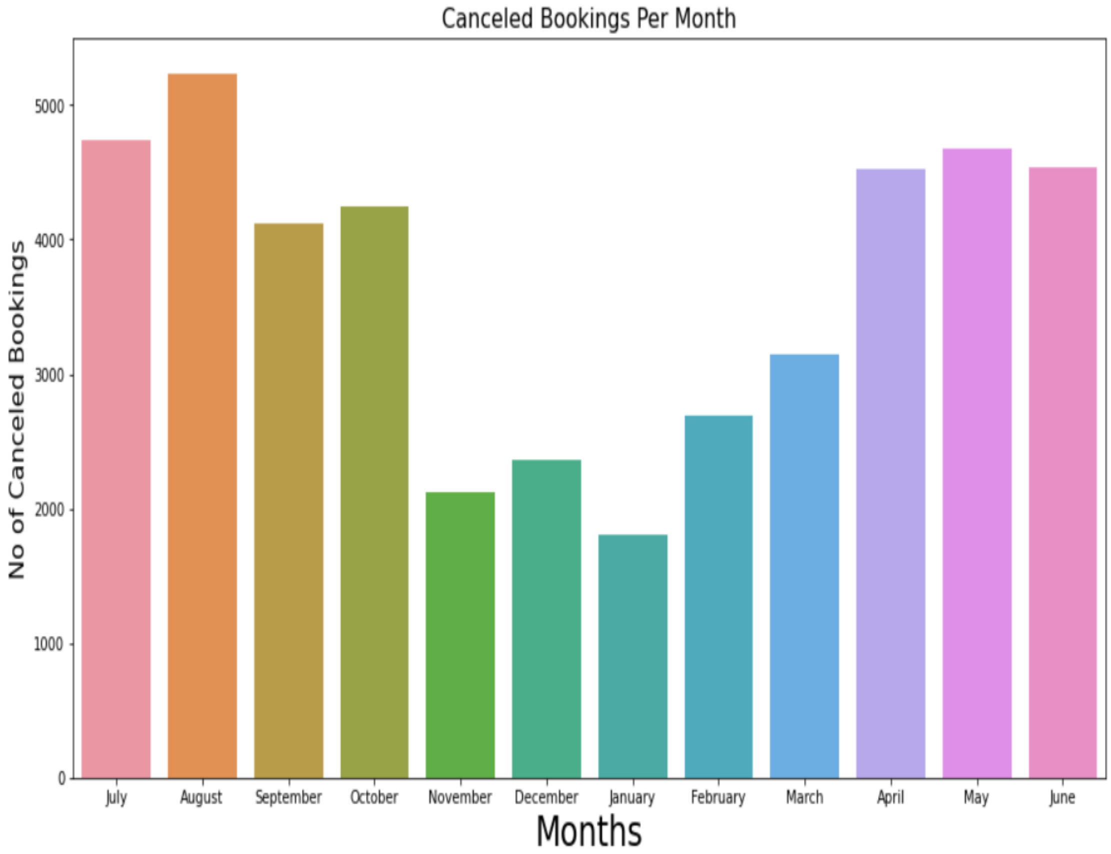
- ➢ August and July Months has generated the major revenue for Resort Hotel
- ➢ January , February , November and December has generated the least Revenue.

Canceled Bookings Per Year



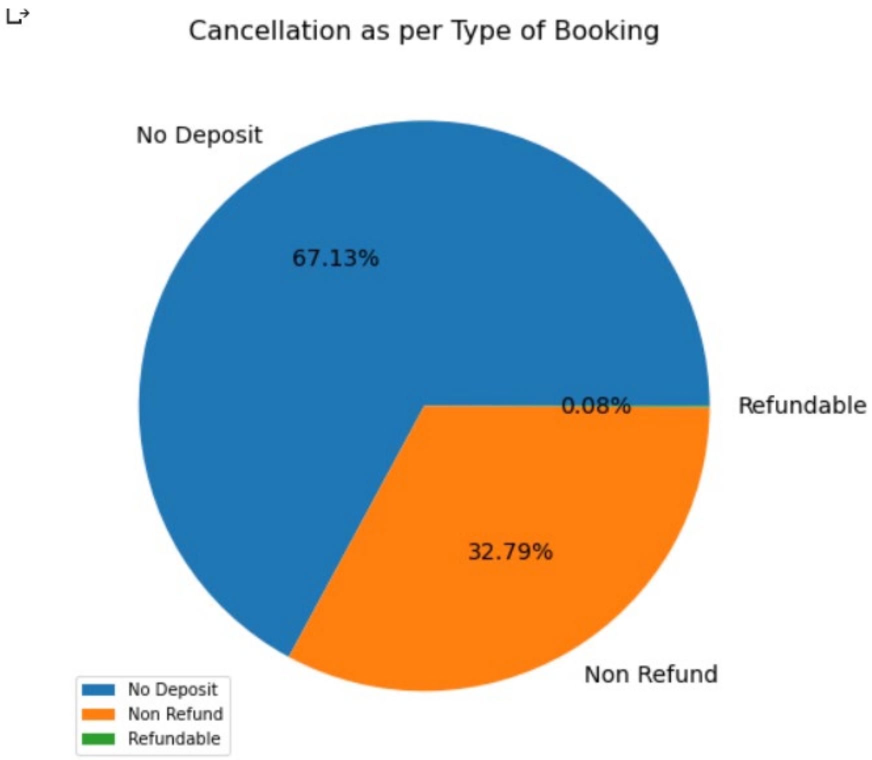
- ➢ Year 2016 has most number of Cancelled Bookings.

Cancelled Bookings Across all Months



- May , June , July and August has the highest number Cancelled Bookings months
- January and November has lowest number Cancelled Bookings

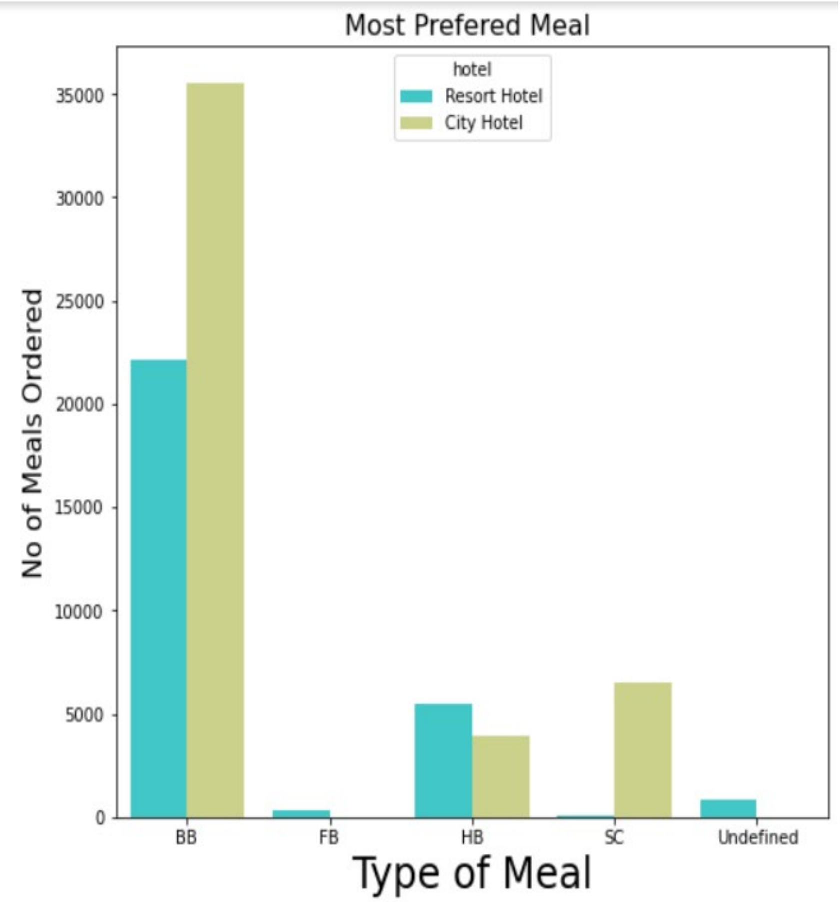
Cancellation made According to type of Deposit made



Cancellation as per Type of Booking

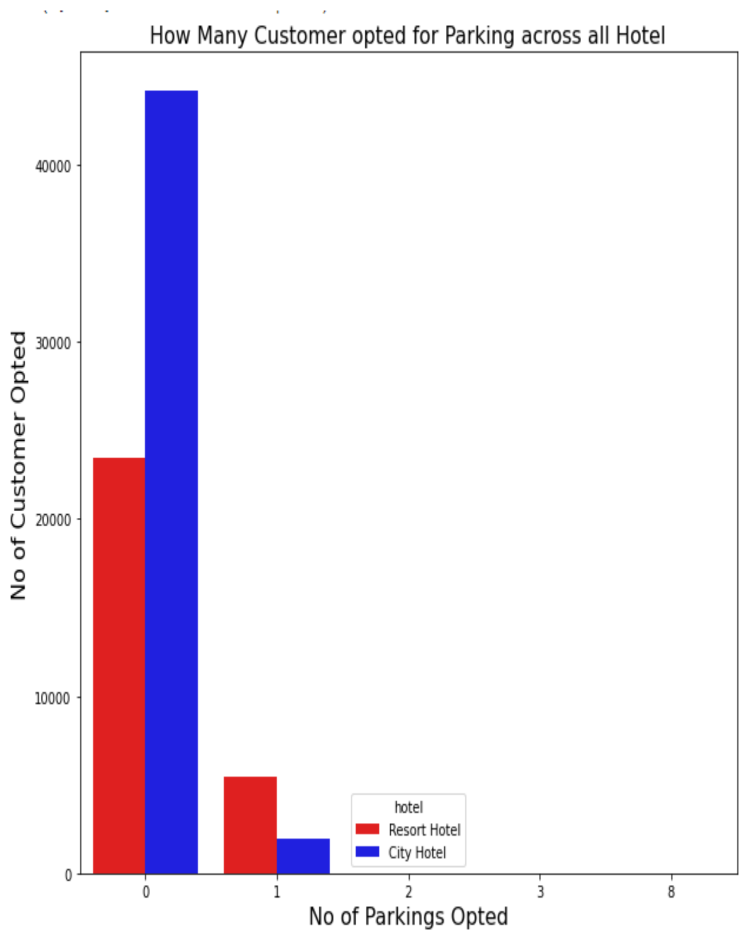
- ➢ Cancellation made across No Deposit is higher than non-refund and refundable

Most Preferred meal



- BB meal is most Preferred meal by Customers for both type of Hotels
- Sc meal is least Preferred Meal for Resort Hotel.
- FB meal is least preferred by the City Hotel

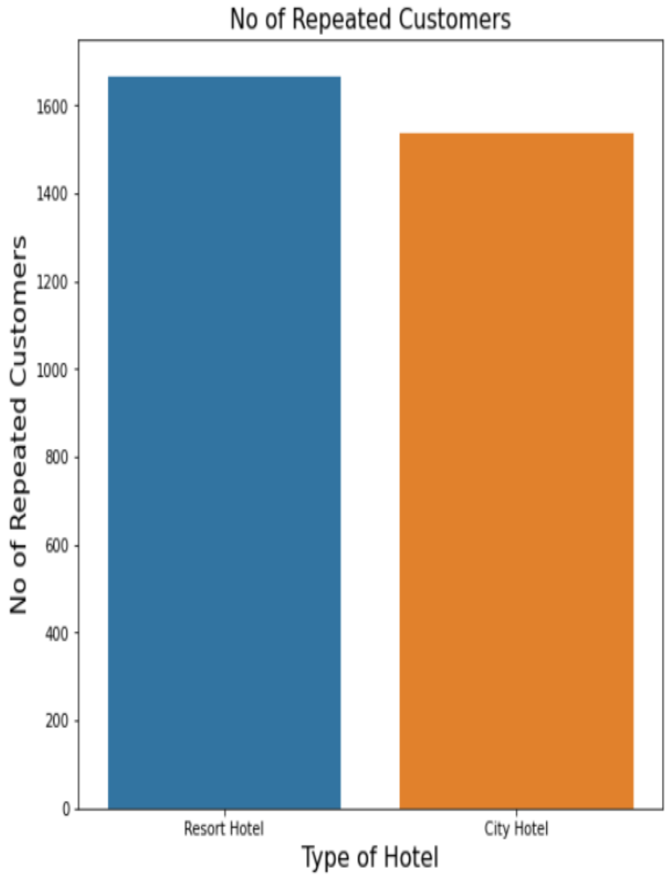
No of People opted for Car Parking



- Maximum no of customers has not opted for parking space.

- Very less number of people has opted for 1 car parking space.

Repeated Customers Across type of Hotels



- Resort Hotel has the Highest number of Repeated Customers

Observations

- Most of the people booked City Hotel
- Most number of booking were in year 2016 as it has data of all months 2015 and 2017
- As the bookings were much more higher in City Hotel we generated most revenue from City Hotel
- As per type of customer Transient customers have booked most of the times
- In year 2015 most bookings were in month of September and October. In Year 2016 most bookings were in month of June and October. In year 2017 most bookings were in months of may and June.
- Transient Customer books more often.
- From the most Bookings per we can see that in the months of June , July and August has highest bookings
- July and August months has the highest rate of Bookings.
- Portugal , Great Britan and France has booked the hotels most number of times.

Observations

- Avg Revenue per Day for Resort Hotel is 401.06 which is about 56.23% and for City Hotel 312.15 which is about 43.77%

- Avg revenue per day was highest in months of May and june for City Hotel and July and August for the Resort Hotel.

- Most Cancelled bookings were in year of 2016

- July and August months has the highest cancellations.

- Most Cancellations were done by City Hotel customers.

- BB meal is the preferred type of meal for both the type of hotel

- Very Less number of Customers opted for parking.

- Resort Hotel has the Highest number of repeated customers.

Conclusion

- Most bookings were made for City Hotel but with much less number of bookings Resort Hotel has generated approximate 41% revenue out of Total so focus more on Resort Hotel customers to generate more revenue

- Most of the bookings were from PRT , GBR , France so advertise more in other countries with some special offers target these top countries more.

- Most booked type of rooms were A,D and E and very less bookings for other type of rooms so increase A,D and E type of rooms.

- Major bookings were cancelled in months of July and August try to send exciting offers for the booked customers in these months.

- Most Cancelled bookings were from the customers which hasn't paid the deposits so try to take deposits from more customers.

- There are very less no of repeated customers try to understand customer needs and try to fulfill maximum of it.

- Focus more on Transient type of Customers they are more likely to book.