Capstone Project 1

Hotel Booking Analysis

Kiran Ahire

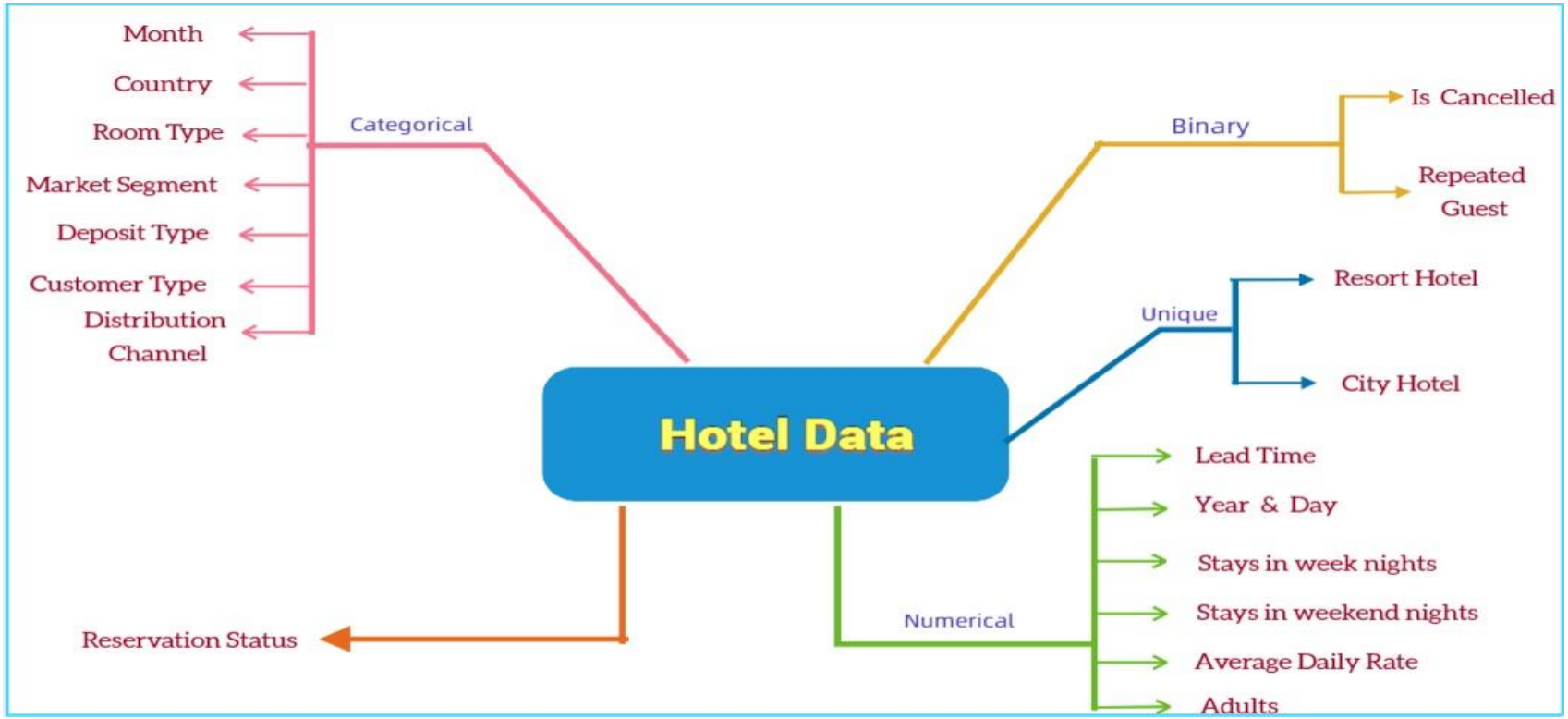
Points for Discussion

- Business Objective
- Data Summary
- Descriptive Analysis
- Corelation between Data
- Outliers in Data
- Univariate Analysis
- Bivariate Analysis
- Conclusion

Business Objective

- For this project we will be analysing hotel booking data. This dataset contains booking information about City hotel and 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.
- Hotel industry is very volatile industry and the bookings depends on above factors and many more.
- The main objective behind this project is to explore and analyse data to discover important factors that govern the bookings and give insights to the hotel management which can perform various campaigns to boost the business and performance.

Data Summary



Descriptive Analysis

info() -

```
hotel_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
```

dtypes()-

```
[50] #Getting datatype
     hotel_df.dtypes
```

```
hotel                           object
is_canceled                      int64
lead_time                        int64
arrival_date_year                int64
arrival_date_month              object
```

shape –

```
[88] #Getting number of rows and columns
     hotel_df.shape
```

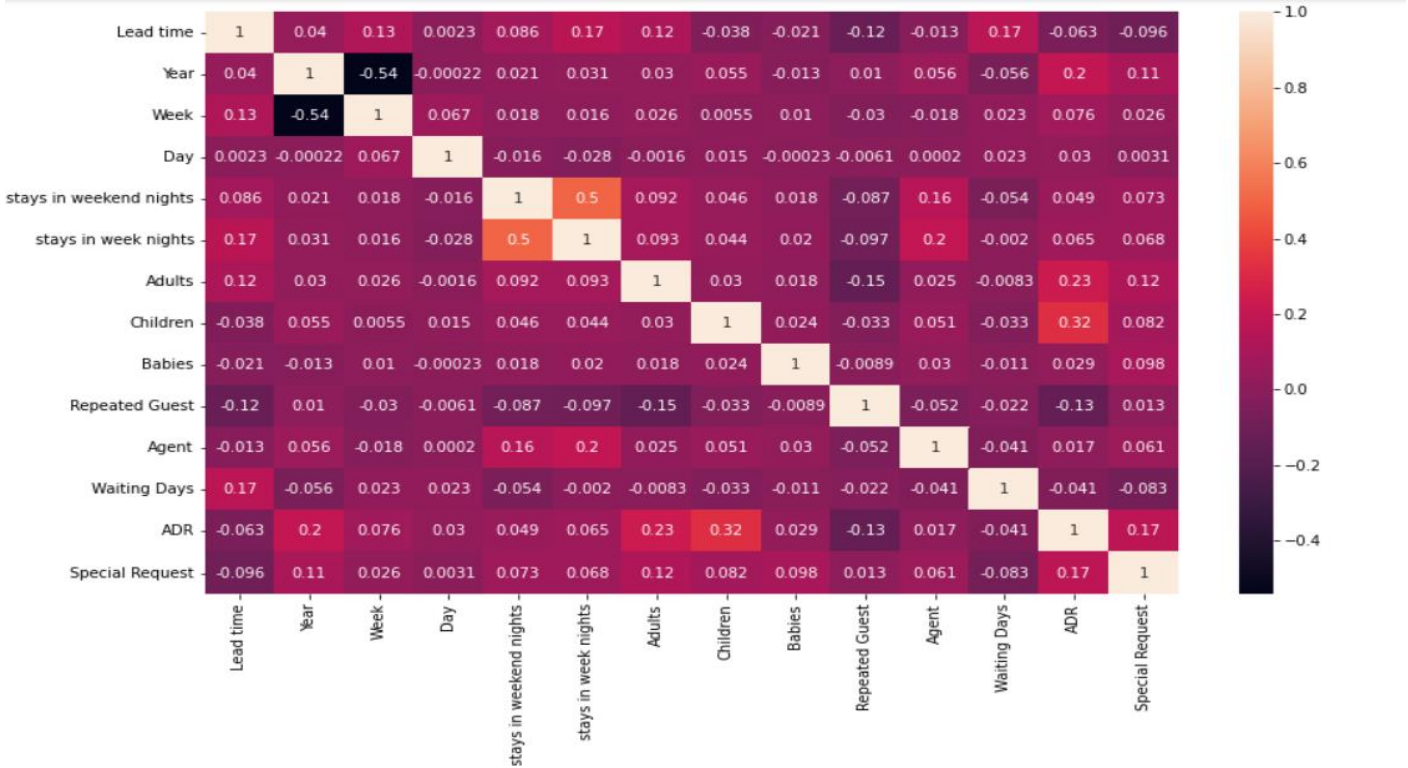
```
(119390, 32)
```

describe() -

```
hotel_df.describe()
```

	is_canceled	lead_time	arrival_date_year
count	119390.000000	119390.000000	119390.000000
mean	0.370416	104.011416	2016.156554
std	0.482918	106.863097	0.707476
min	0.000000	0.000000	2015.000000
25%	0.000000	18.000000	2016.000000
50%	0.000000	69.000000	2016.000000
75%	1.000000	160.000000	2017.000000
max	1.000000	737.000000	2017.000000

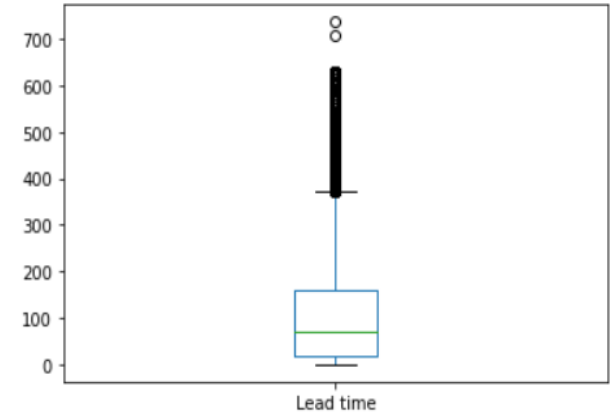
Corelation

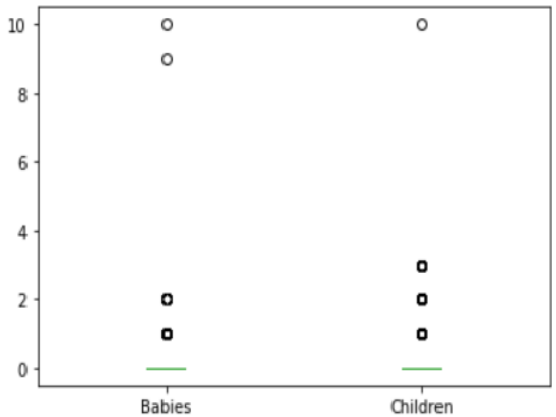


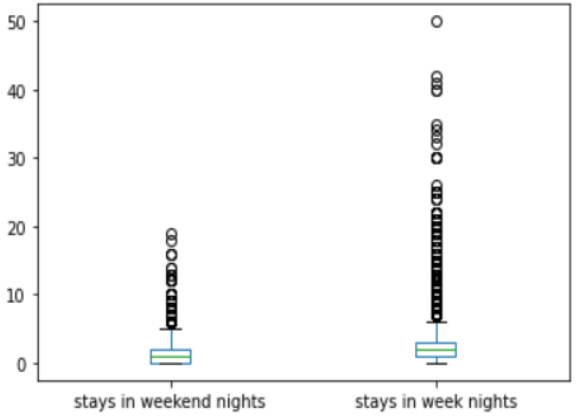
- From the above Corelation matrix we can conclude that Children and ADR are corelated.

Outliers

- An outlier is **an observation that lies an abnormal distance from other values in a random sample from a population**.
- In our Dataset features like Lead time, Babies , Children , Stays in Week nights , Stays in Weekend nights have outliers or extreme values.







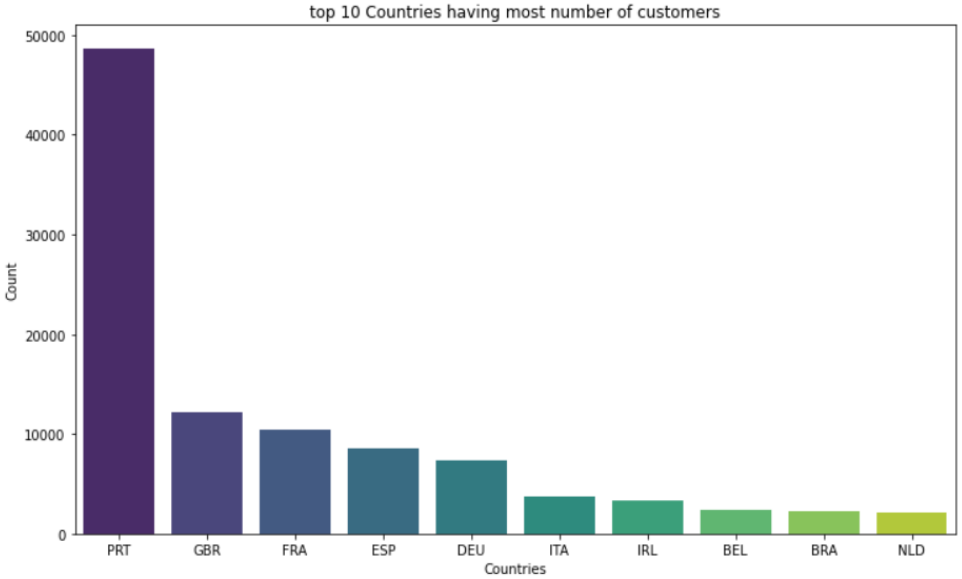
Univariate Analysis

- Univariate analysis is **the simplest form of analyzing data**.
- In Univariate analysis data has only one variable. It doesn't deal with causes or relationships and it's major purpose is to describe.
- It takes data, summarizes that data and finds patterns in the data.
- In Univariate analysis we have
    1. Top 10 Countries having most number of customers
    2. Most Customer Type
    3. Top Distribution channel
    4. Maximum Customers year

Top 10 Countries having most number of customers

- The below information shows the order of top 10 most frequent Country in the dataset.
- A quite far gap can be seen between Country PTR as top rank with other Countries.

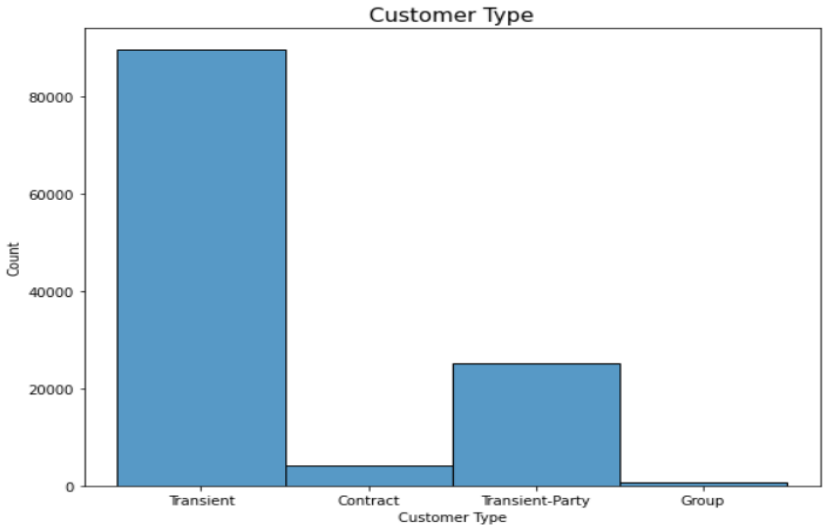
- **Conclusion :** Country Portugal h number of customers with the count of 48590 followed by Great Britain and then Spain.



Most Customer Type

- In this we are trying to find most visited customers are of which type.
- In this we have four type of customers namely Transient, Contract , Transient Party and Group.

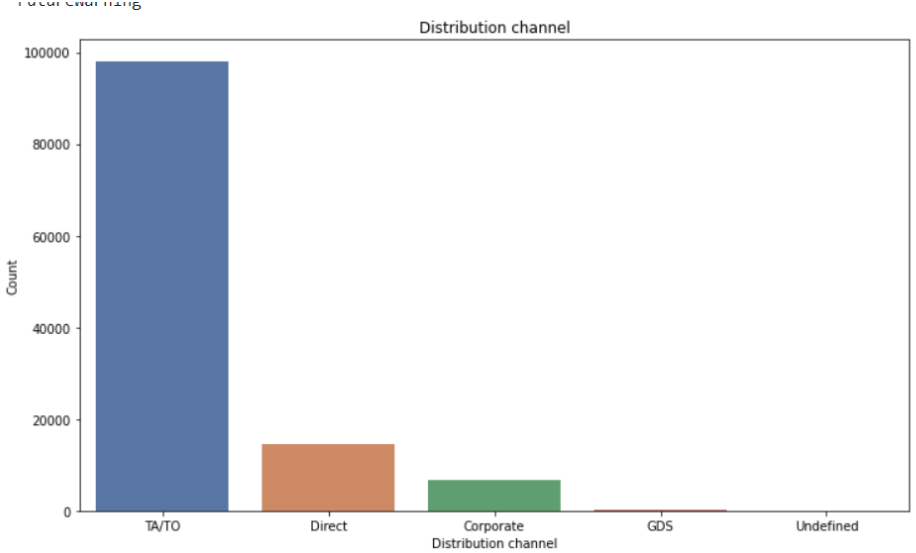
- **Conclusion :** From this graph we car that most number of customers stay for only short period of time i.e. of Transient type with the count of 89613.



Top Distribution channel

- In this we are trying to find top Distribution Channel which customers used for Reservation.
- We have four types of Distribution Channel such as TA , Direct , Corporate , GDS and Undefined.

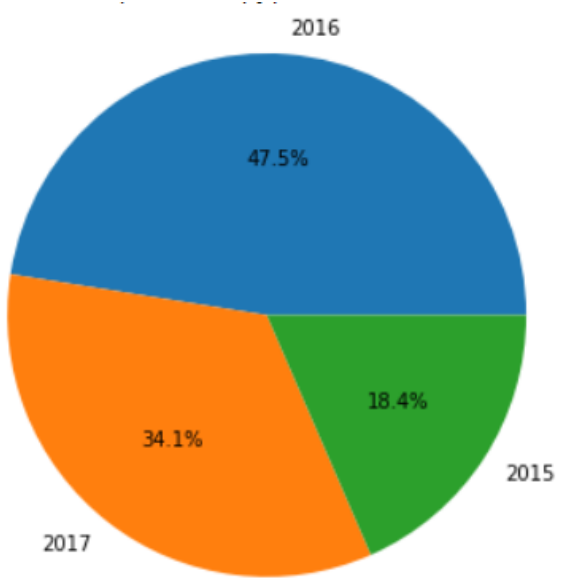
- **Conclusion :** The graph clearly sho Number of customers used Trave Agent To booked their room with the count Of 97870 customers.



Year having most number of customers

- In this we are trying to find most number of customers visited in which year.
- In our Dataset we have given the data of 3 years i.e 2015, 2016 and 2017.

- **Conclusion :** From our pie chart we can conclude that in 2016 percentage of customers visited is 47.5% which is larger than years 2015 and 2017.



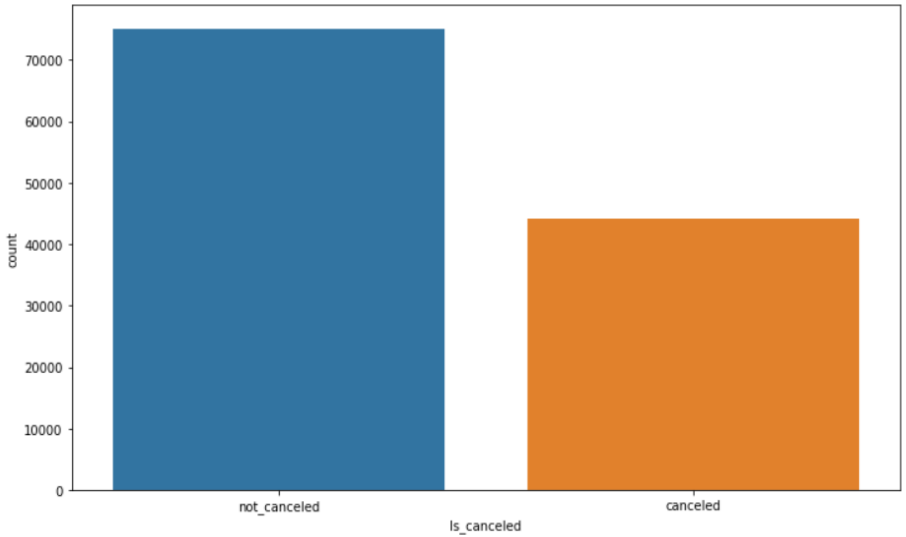
Bivariate Analysis

- Bivariate analysis is one **of the simplest forms of quantitative** (statistical) analysis.
- It is the analysis of two variables for the purpose of determining the empirical relationship between them.
- It is the analysis of the relationship between the two variables.
- In Bivariate Analysis we have
    1. Analyzing cancelled bookings data
    2. Find out which type of hotel has most number of customers
    3. Best time to book the room
    4. Number of arrivals per year

Analyzing cancelled bookings data

- In this we are trying to analyse the cancelled bookings.
- In our Dataset we have an attribute Is cancelled which is in binary form if 1 i.e. booking is cancelled and 0 for not cancelled.

- **Conclusion:** Using a countplot, we were able to graph the total amount of canceled vs non- cancelled data. It appears the majority of bookings were not cancelled and About 50% of all bookings are cancelled.



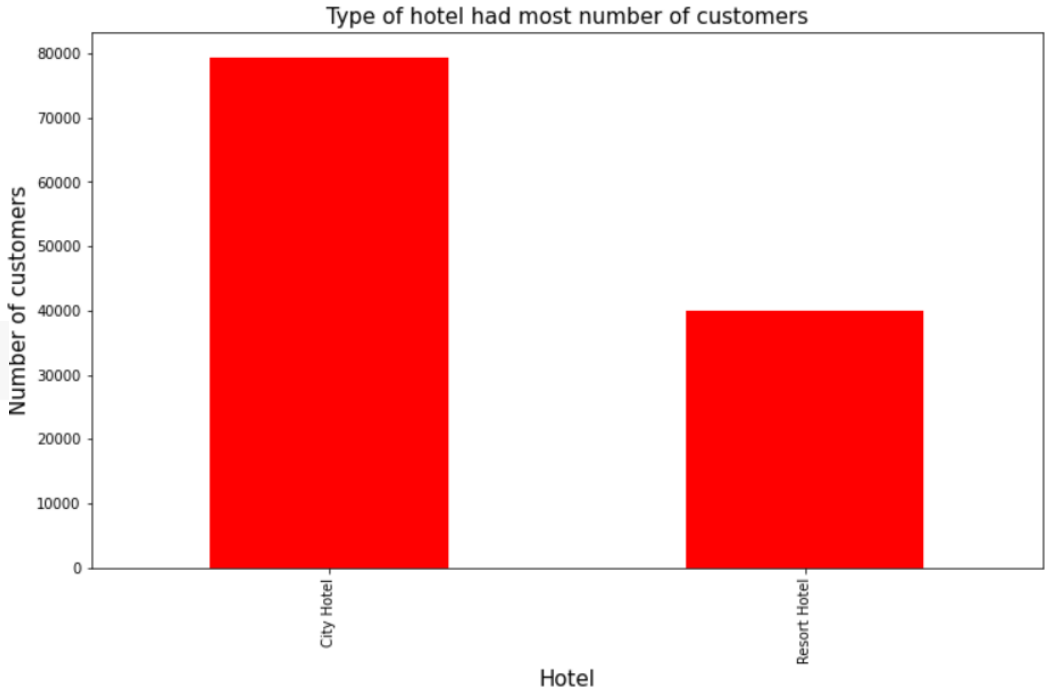
Type of hotel having most number of customer

- In our dataset there are two types of hotel viz. City Hotel and Resort Hotel.
- We are trying to find out which type of hotel has most number of customers.

- **Conclusion :** From this we can Conclude that City Hotel has more customers than Resort hotel.

```
hotel_df2['Hotel'].value_counts()
```

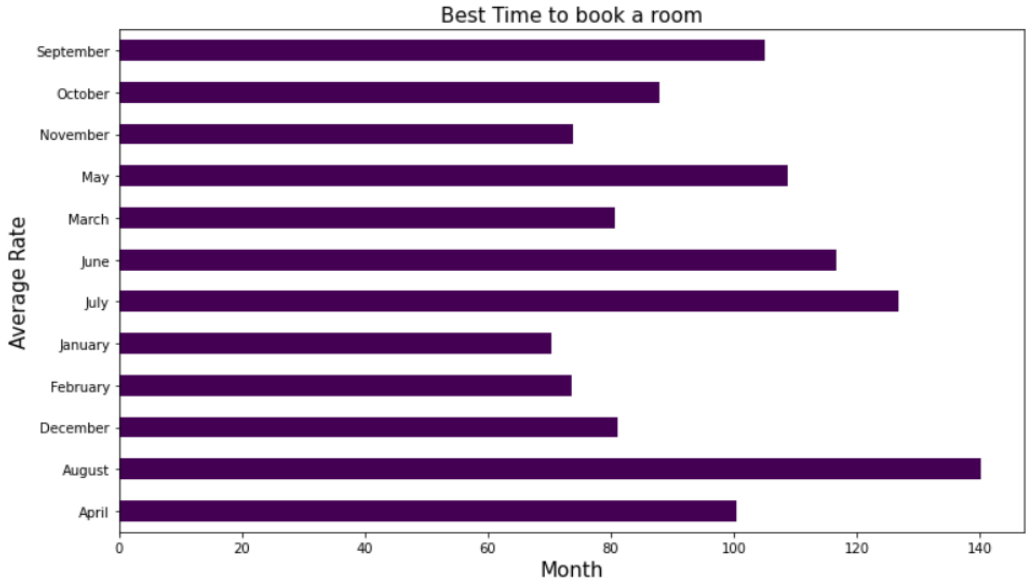
```
City Hotel      79330
Resort Hotel    40060
Name: Hotel, dtype: int64
```



Best time to book the room

- In this we are trying find out best time to book the room.
- In our dataset we have Attribute Month And ADR (Average Daily Rate) using which we are trying to get best time to book a room.

- **Conclusion :** From this we can conclude that January has to lowest ADR which means January is the best time to book the room. Also Highest daily rates occurred in the summer (June, July, August)

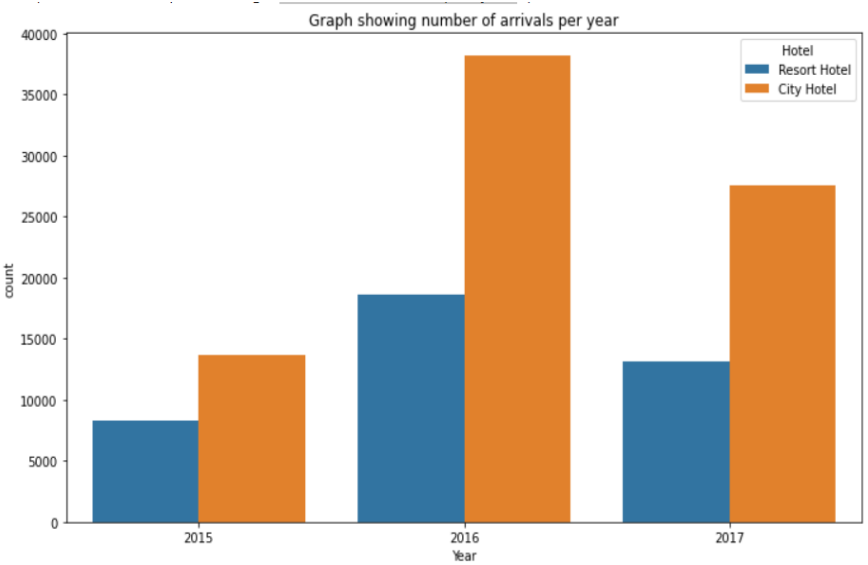


Number of arrivals per year

- In this we are trying to find number of customers arrived per year.
- We have number of customers of City Hotel and Resort Hotel w.r.t Year.

- **Conclusion :** City hotels had the most bookings consistently each year with the largest amount of bookings in 2016.

  Resort Hotel had less number of bookings each year as compared to City Hotel.



Conclusion

After Performing Exploratory Data Analysis we get following insights from the data :

- City Hotel had most number of customers than Resort Hotel.
- January month has lowest Rate comparing to other months. Hence January is the best time to book the room.
- Most number of customer used travel agents to book their room.
- Maximum Hotel Bookings are during the year 2016 followed by 2017.
- Most number of customers are of Transient Type.
- Most number of customers are from Portugal followed by Great Britain and then Spain.
- About 50% of all bookings are cancelled.
- Highest daily rates occurred in the summer (June, July, August)