Project
EDA On Hotel Bookings

By:
Ajeet Kumar

Problem Statement

- In this project we will analyze the data of the hotel booking dataset.
- This hotel dataset contain booking information for city and resort hotels with their corresponding variables such as canceled bookings, arrival data per annum, arrival data per month, arrival data per day , types of guests(children , adults , babies) and company etc.
- Hotel booking is a very big field and depends upon the different factors of such as it’s type of booking data , date ,year , month ,types of meal etc.
- Our main purpose behind this analysis is to fetch the important data factor to check in which time we can get hotel booking in minimum/maximum price , types of meal guests does prefer to have, ratio of babies , children , adults who do come in hotel.
- When do guests prefer to come in week.

Process Of Analysis

- We have divide data into three different parts;

- In EDA we have analysis data on single , double and multiple variables.

Data Collection And Understanding of Data

- After data collection, it's very important to understand it and for this we do analysis on data set. After the analysis we get to know that how many columns and rows are present in our data set . We have 119390 rows and 32 columns.
- **Data Set's Columns:**

  Hotel : City or Resort Hotel
  is_canceled : it has two values 0 and 1. 0 does stand for not canceled 1 stands for canceled.
  lead_time : No of days from entering to exit the hotel.
  arrival_date_year : it indicates the year arrival date of guest.
  arrival_date_month: it indicates the arrival date of guest.
  arrival_date_day : it indicates the arrival day of guest.
  stays_in_weakend_nights : No. of week night stayed in hotel.
  stays_in_weak_night : No. of week nights stayed in hotel.
  adults : no of adults
  children: no of children
  babies : no of babies.
  country : name of country

Data Collection And Understanding of Data

market_segment : arial segment has many such as ; Direct, Corporate, Online TA, Offline TA/TO, Complementary, Groups, Undefined, Aviation.

Distribition_channel : distribution channel has many values such as ; Direct, Corporate, TA/TO, Undefined, GDS.

Is_repeated_guest : it contain two values 0 and 1. if guest repeated then this values will 1 else 0.

previous_cancellations : it represent how many times guest has canceled booking.

previous_bookings_not_canceled: it represent the booking that didn't cancel.

reserved_room_type: it does represent hotel type of booking **.**

assigned_room_type : it represent type of assign room after reservation.

booking_changes: it represent who does changes the their booking types.

Agent: it represent agent ID.

company : it represent name of company.

days_in_waiting_list : it represent the waiting list.

customer_type : it represent the customer types .Types are as follows;(Transient, Contract, Transient-Party, Group)

adr : it conation addresses of all guests.

required_car_parking_spaces:it contain information wheither is it required for parking or not.

total_of_special_requests :represent total special guests.

reservation_status : it contain information of booking status

reservation_status_date: contain reservation status date.

Filtering Data And Manipulation

We did rename our data set columns names;

Raw data frame columns before rename

```
[59] df.columns
```

```
Index(['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',
       '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_status_date'],
      dtype='object')
```

Data frame after rename it's columns: Here, we created another data frame with name **data_f** and changed the column name in it.

```
data_f.columns
```

```
Index(['Hotel', 'Canceled', 'LeadTime', 'ArrivingYear', 'ArrivingMonth',
       'ArrivingWeek', 'ArrivingDate', 'WeekendStay', 'WeekStay', 'Adults',
       'Children', 'Babies', 'Meal', 'Country', 'Segment', 'DistChannel',
       'RepeatGuest', 'PrevCancel', 'PrevBook', 'BookRoomType',
       'AssignRoomType', 'ChangeBooking', 'DepositType', 'agent', 'company',
       'WaitingDays', 'CustomerType', 'ADR', 'ParkSpace', 'SpecialRequest',
       'Reservation', 'ReservationDate'],
      dtype='object')
```

Filtering Data And Manipulation

- Finding missing values ;

```
data_f.isnull().sum()
```

there are only 4 columns that contain null values

```
data_f.isnull().sum().sort_values(ascending = False)[0:4]
```

```
company          112593
Arrivingdate      72998
agent             16340
Country             488
dtype: int64
```

- Make changes in null value: Here, we have changed to null values from 0.

```
# Handling the missing values

data_f['company'].fillna(0,inplace=True)
data_f['agent'].fillna(0,inplace=True)
data_f['Country'].fillna(0,inplace=True)
data_f['Children'].fillna(data_f['Children'].mean(),inplace=True)
```

Filtering Data And Manipulation

Finding duplicate values:

```
data_f.duplicated().value_counts()
```

```
False    87230
True     31980
dtype: int64
```

There are 87230 not null values and 31980 null values.

Dopped Duplicate values

```
[33] # i found the 31994 duplicate values so now drop of these duplicate values in the data set
     data_f.drop_duplicates(inplace=True)
```

```
data_f.shape
```

```
(87230, 33)
```

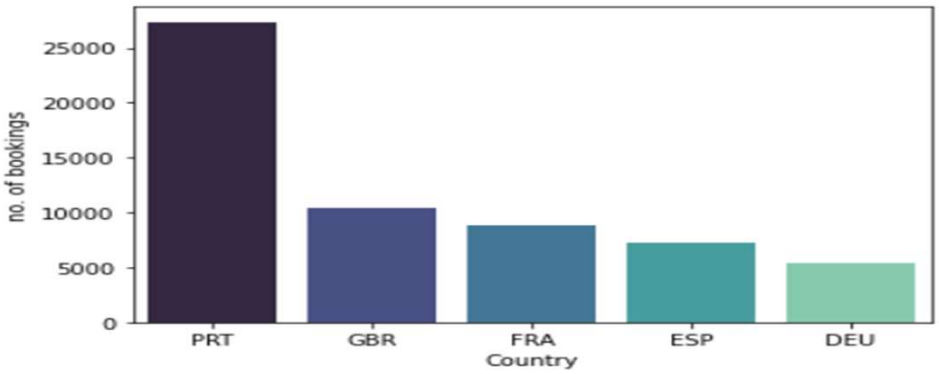
Now, total rows are 87230

EDA And Visualization

Here we have analysis data of country column to see top 5 countries visitor data that have visited most.

```
grp_by_country = data_f.groupby('Country')
data = pd.DataFrame(grp_by_country.size()).rename(columns = {0:'no. of bookings'}).sort_values('no. of bookings', ascending = False)
data = data[:5]
sns.barplot(x = data.index, y = data['no. of bookings'],palette='mako')
plt.show()
```

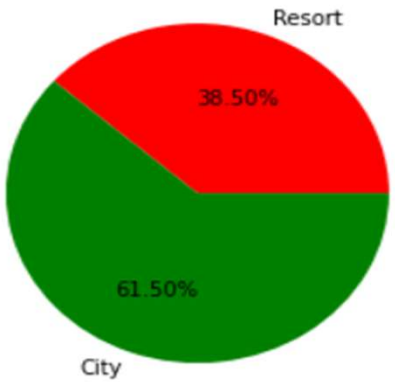
Visualization no most visitor country



'GBR'(Great Britain) has the second highest no of the visitor(10424) then comes the 'FRA'(France) with the third highest no of the visitors(8823)

EDA And Visualization

Which hotel has more cancelation

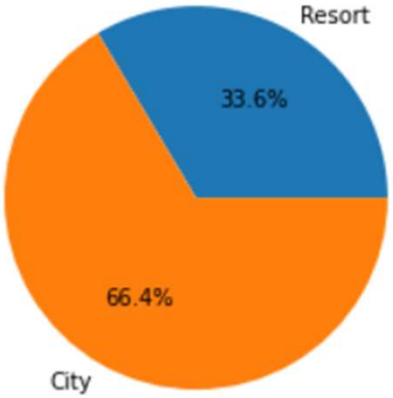


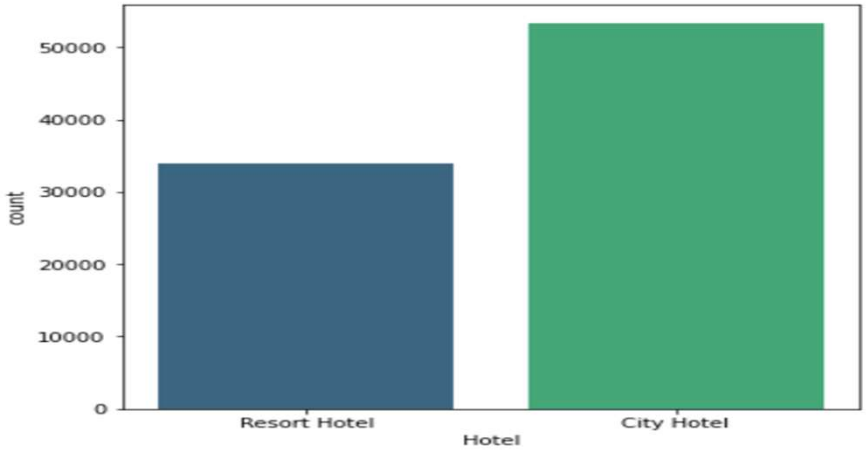
Conclusion:

- Most of the cancellation are in city hotel because it has more reservation .
- City hotel has 61.50% of cancellation.
- Some time if any person does about to come because of some circumstances he/she not able to come and it is cheaper then resort so guest does cancel city hotel most.

EDA And Visualization

Here we are analyzing which type of hotel have Highest reservation



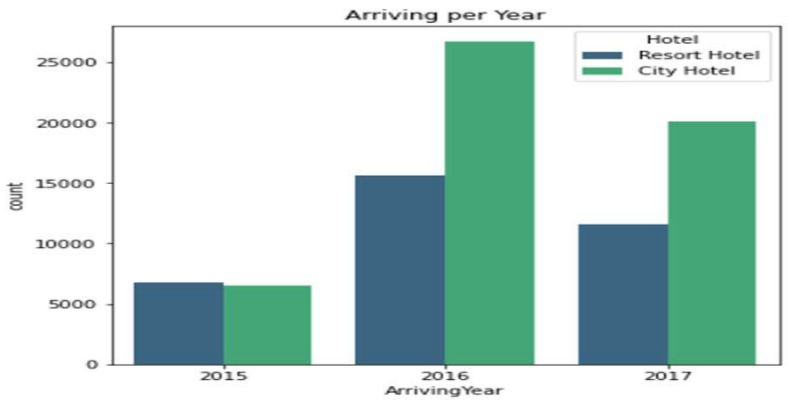


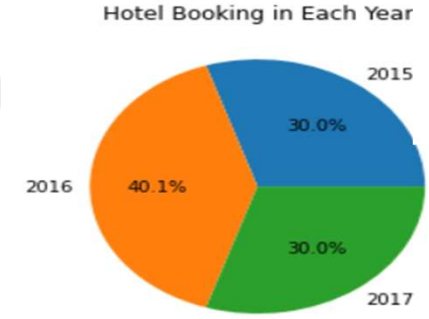
Conclusion:

- City hotel has maximum reservation City hotel 66.4% and Resort Hotel has 33.6% .
- Most of the people who come to visit any place they used to reserve hotel in city that's why here city hotel has more reservation.

EDA And Visualization

Here we are analyzing data on year wise hotel booking



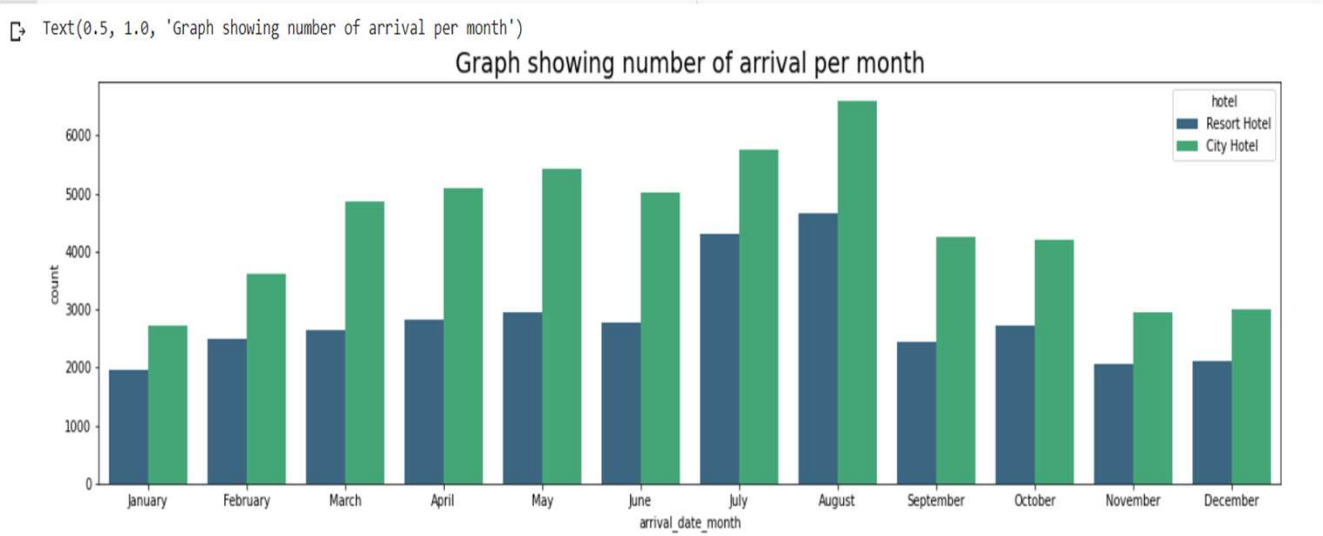


Conclusion:

- Hotel booking in 2015 is 30.0% ,2016 is 40.1% , 2017 is 33.0% include resort and city hotel.
- Resort and city has highest booking in 2016.
- Lowest booking or resort and city hotel 2015.

EDA And Visualization

Here we’re going to analyze and visualize data of busy months in year

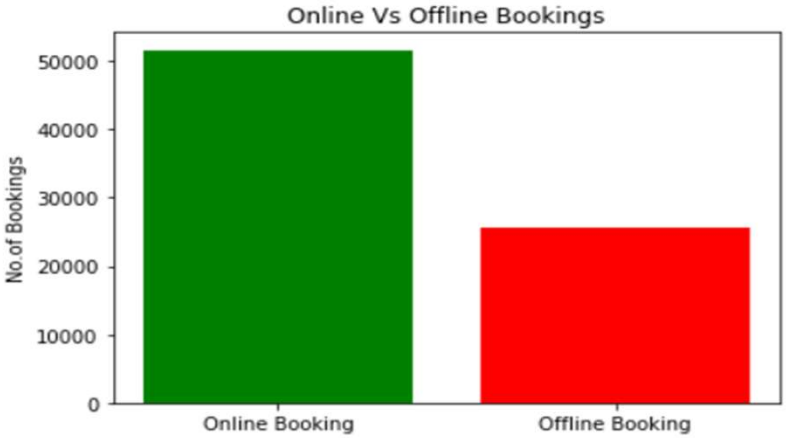


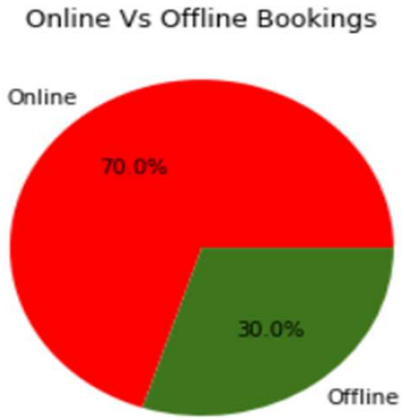
Conclusion:

- In august month hotel booking are on peak .
- All months of years are average booking.
- Maximum bookings in month of August.

EDA And Visualization

Online and offline booking



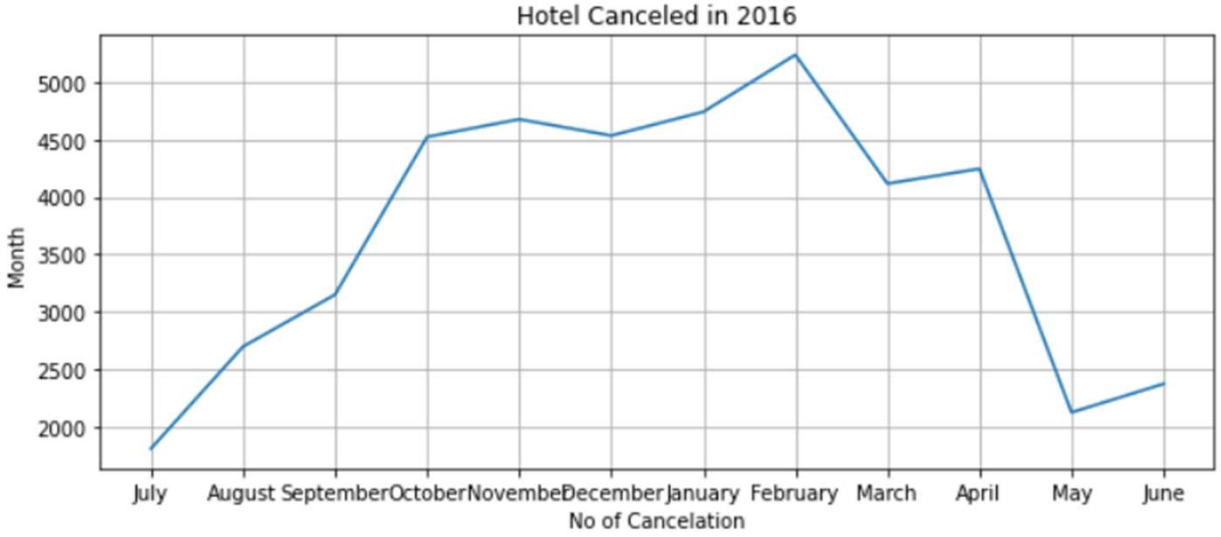


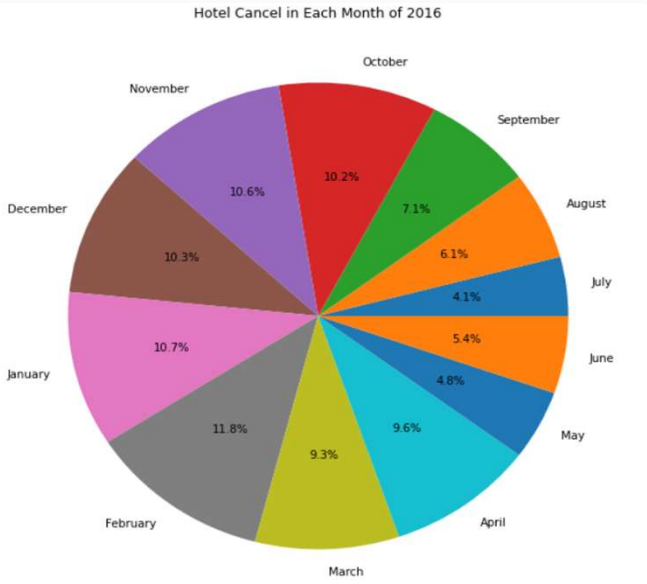
Conclusion:

- Most of the people at this time , in all over the world have smart phone . So they does prefer to book online most.
- People do online booking because they can see the rating of the hotel and without going to the hotel they can book hotel in a single click.

EDA And Visualization

Here we are doing analysis and visualization of 2016 cancelations



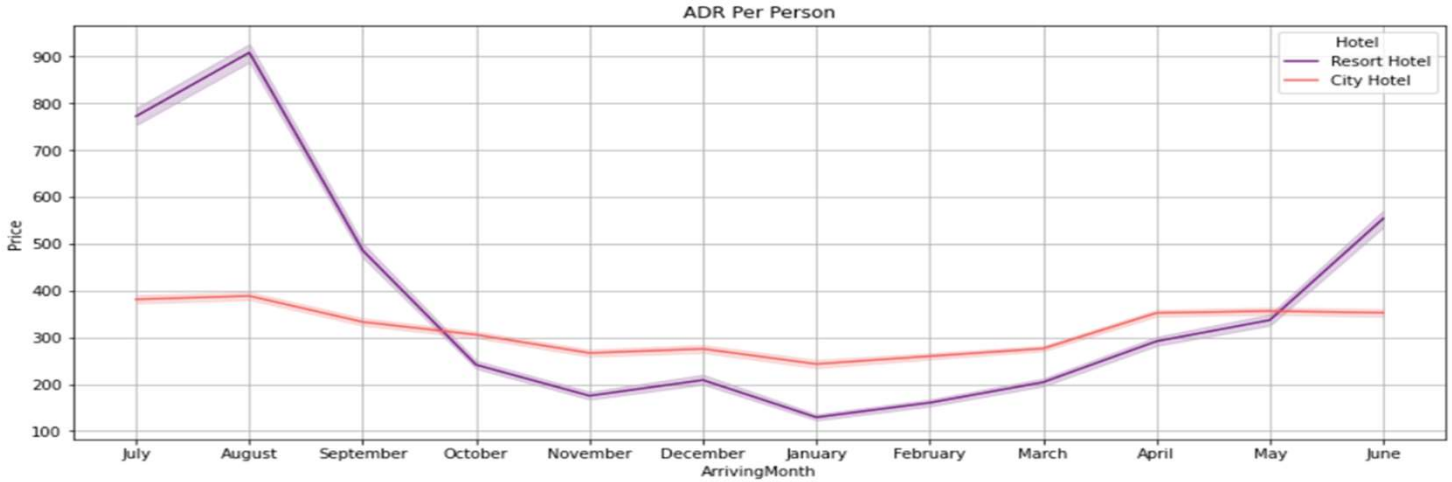


Conclusion:

- In month of February cancelation is maximum with 11.8% and minimum with 4.1% in July.
- In month of February maximum persons do hotel booking that's why cancelation ratio lies on peak.

EDA And Visualization

Here we are calculating average daily price for per person

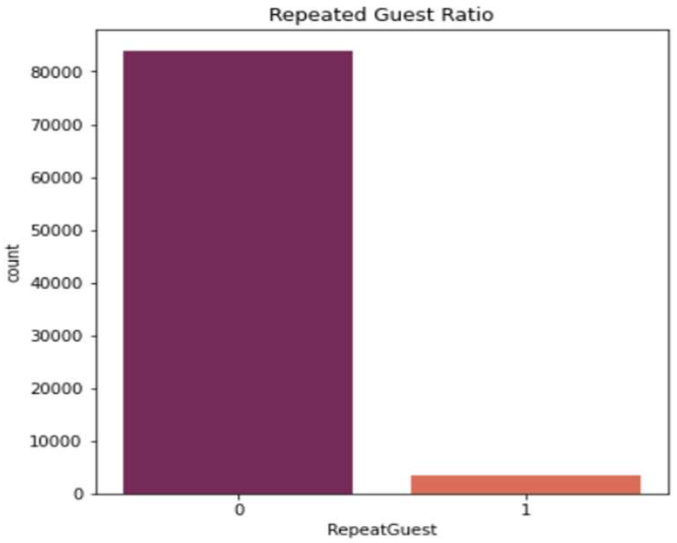


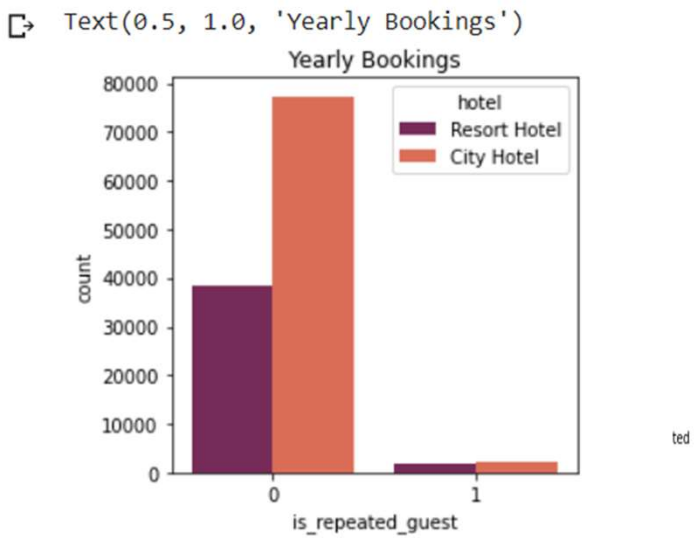
Conclusion:

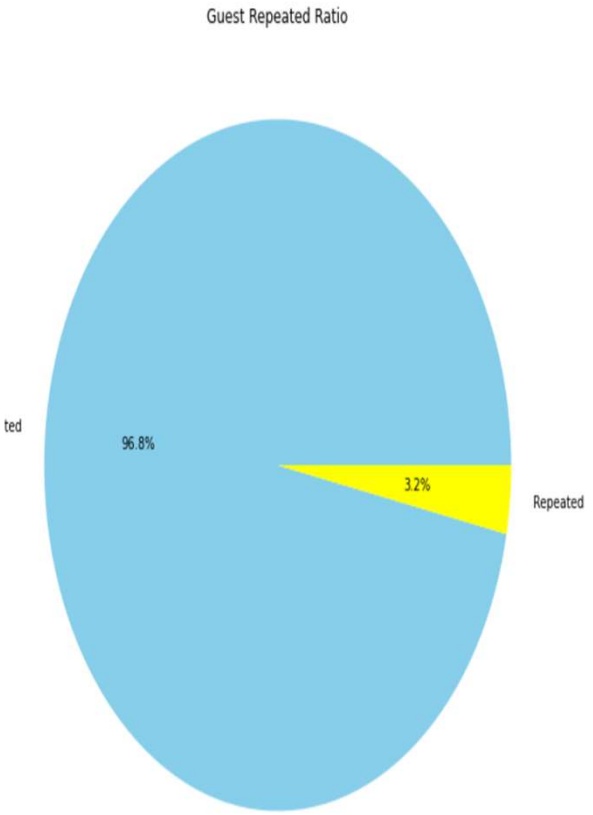
- Average daily price for per person does high in august month.
- Average daily price for per person does minimum in January month.
- Most of the people do prefer to stay in city hotel so it’s price does not increase or decrease like resort hotel.
- In resort hotel people mostly go for vacation so they reserve resort because they required more space to fun with friend and family .

EDA And Visualization

Here we are doing analysis on data of guest repeated ratio.





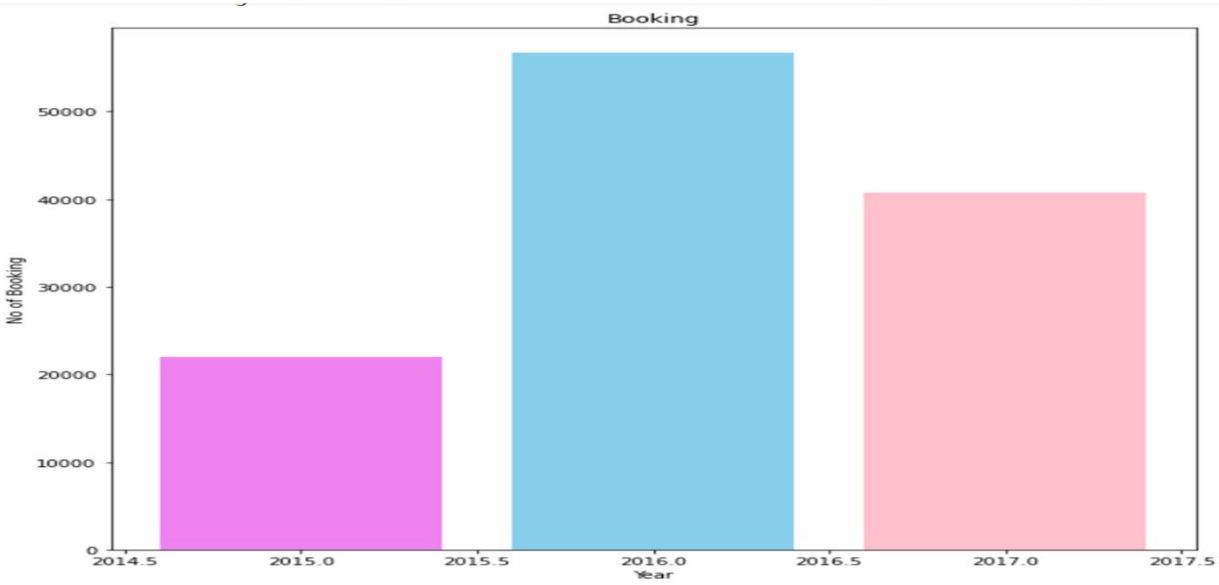


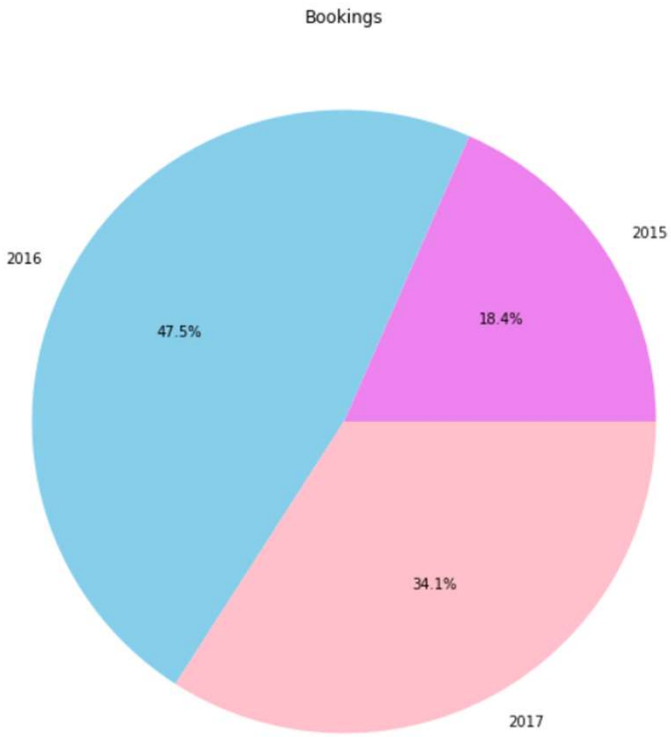
Conclusion:

- According to ratio of analyzing the data we can cay that there is a very few no of guest who does repeat hotel booking .
- Only 3.2% guests has repeated any 96.8% didn’t repeat.

EDA And Visualization

Here we are doing analysis that in which year has highest bookings



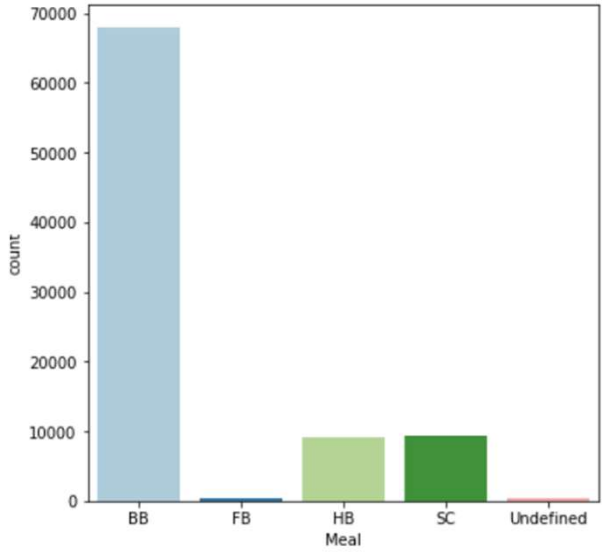


Conclusion:

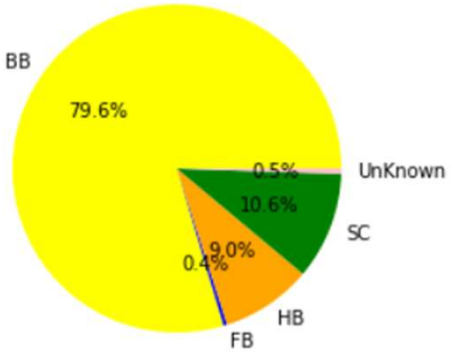
- Here in 2016 has the highest no. of bookings and 2015 has minimum no of bookings.
- We have given only 6 months of data so no of bookings are low due to lesser month than 2016

EDA And Visualization

Which is the most popular meal order by the visitors




```
BB
FB
HB
SC
Undefined
[33679, 150, 3788, 4477, 219]
```



Conclusion:

- BB is the most popular meal ordered by the visitor(67907).
- Bed and breakfast (typically shortened to B&B or BnB) is a small lodging establishment that offers overnight accommodation and breakfast

Thank you