Capstone Project

Hotel Booking Analysis

Ashok Kondhalkar

Table of Contents

Objective

This hotel booking dataset contains booking information about city and resort hotels. Both datasets share the same structure, with Database having 119390 Rows and 32 columns. All personally identifying information has been removed from the data. We are going to analyse hotel bookings dataset for 3 years i.e. 2015 - 2017.

We will be discussing following steps in upcoming slides.

- Data loading and exploration.
- Data Wrangling
- Data analysis and visualization.

Data Summary

The data table consists of 119,390 rows and 32 columns. So, Our analysis starts with understand feature description of each column mentioned below:

- **'hotel':** Hotel(Resort Hotel or City Hotel)
- **'hotel1':**Copy of hotel
- **'is_canceled**': Booking was cancelled or not.
- **'lead_time'**:Time difference between booking date and date of arrival
- **'arrival_date_year'**:Year of arrival
- **'arrival_date_month'**:Month of arrival
- **'arrival_date_week_number'**:Week of arrival
- **'arrival_date_day_of_month'**:Day of arrival
- **'stays_in_weekend_nights'**:Total Stay on weekend
- **'stays_in_week_nights'**:Total Stay on weekday
- **'adults'**:No.of adults in the room
- **'children'**:No. of children in the room
- **'babies'**:No.of babies in the room
- **'meal'**:Type of meal
- **'country'**:Country of origin
- **'market_segment'**:Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

- **'distribution_channel'**:The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- **'is_repeated_guest'**: Repeated guest or not.
- **'previous_cancellations'**:Customer previously cancelled
- **'previous_bookings_not_canceled'**:Customer previous did not cancel
- **'reserved_room_type'**:Type of room type
- **'assigned_room_type'**:Type of room assigned
- **'booking_changes'**:Any changes in booking
- **'deposit_type'**:Type of deposit for booking
- **'agent'**: Agent used for booking
- **'company'**:Company of booking
- **'days_in_waiting_list'**:Waiting list days
- **'customer_type':**Type of customer based on stay duration
- **'adr'**:average daily rate
- **'required_car_parking_spaces'**:parking required
- **'total_of_special_requests'**:No. of special guests
- **'reservation_status'**:Status of reservation
- **'reservation_status_date’**:Date of status of reservation

We had added two columns for our own convenient analysis.

- **'total_stay'**:'stays_in_week_nights'+'stays_in_weekend_nights'
- **'total_peoples'**='babies'+'adults'+'children'

Data Wrangling

- Data wrangling-also called data cleaning, unifying messy and complex data sets to a meaningful format for easy access and analysis..There are various processes designed to transform raw data into more readily used formats.

*It includes following steps:

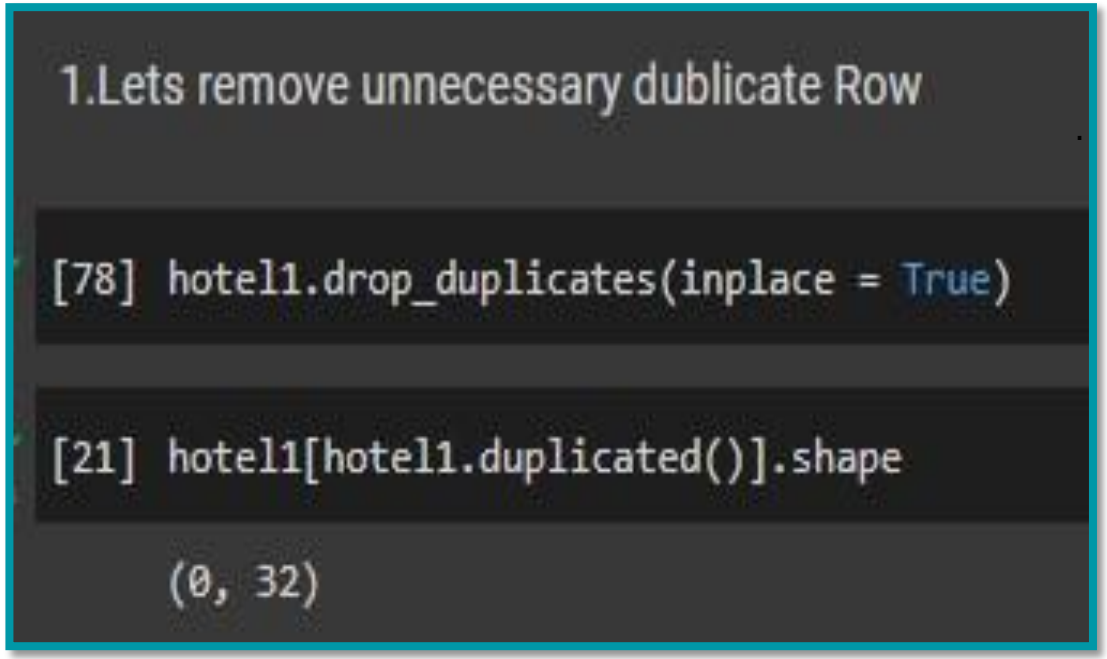
1. Finding unique values.
2. Removing duplicates data.
3. Handling missing values
4. Converting columns to proper data type format.
5. Adding or removing columns for analysis.

1.Finding unique values

```
[14] #Lets make a copy of original DataFrame(Avoid any changes in original DataFrame)
     hotel1=hotel_df.copy()

[15] #lets find unique values if any
     hotel1['hotel'].unique()

     array(['Resort Hotel', 'City Hotel'], dtype=object)

[16] hotel1['is_canceled'].unique()

     array([0, 1])

     hotel1['adults'].unique()

     array([ 2,  1,  3,  4, 40, 26, 50, 27, 55,  0, 20,  6,  5, 10])

[18] #column childern having 0 as well as null values
     hotel1['children'].unique()

     array([ 0.,  1.,  2., 10.,  3., nan])
```

.unique command is used
For finding unique values

2.Removing duplicate data



.drop_duplicates(in place=True) command is used For finding remove duplicate

3.Handling missing values

```
#For better calculation need to replace null values by 0
hotel1['company'] = hotel1['company'].fillna(0)
hotel1['agent'] = hotel1['agent'].fillna(0)

#For better calculation need to replace nan values by its mean
hotel1['children'].fillna(hotel1['children'].mean(),inplace=True)
```

```
children          4
country         452
agent         12193
company       82137
dtype: int64
```

There were 4 columns like company,agent, country and children with missing values

.fillna(0) command used in python to replace null values by zero

4.Convert column to proper data type

```
[29] #convert float datatype to interger
     hotel1['children']=hotel1['children'].astype('int64')
     hotel1['agent']=hotel1['agent'].astype('int64')
     hotel1['company']=hotel1['company'].astype('int64')
```

.astype('int64')
command used in
python for type change

E.g.float to int

5.Adding and removing column for analysis

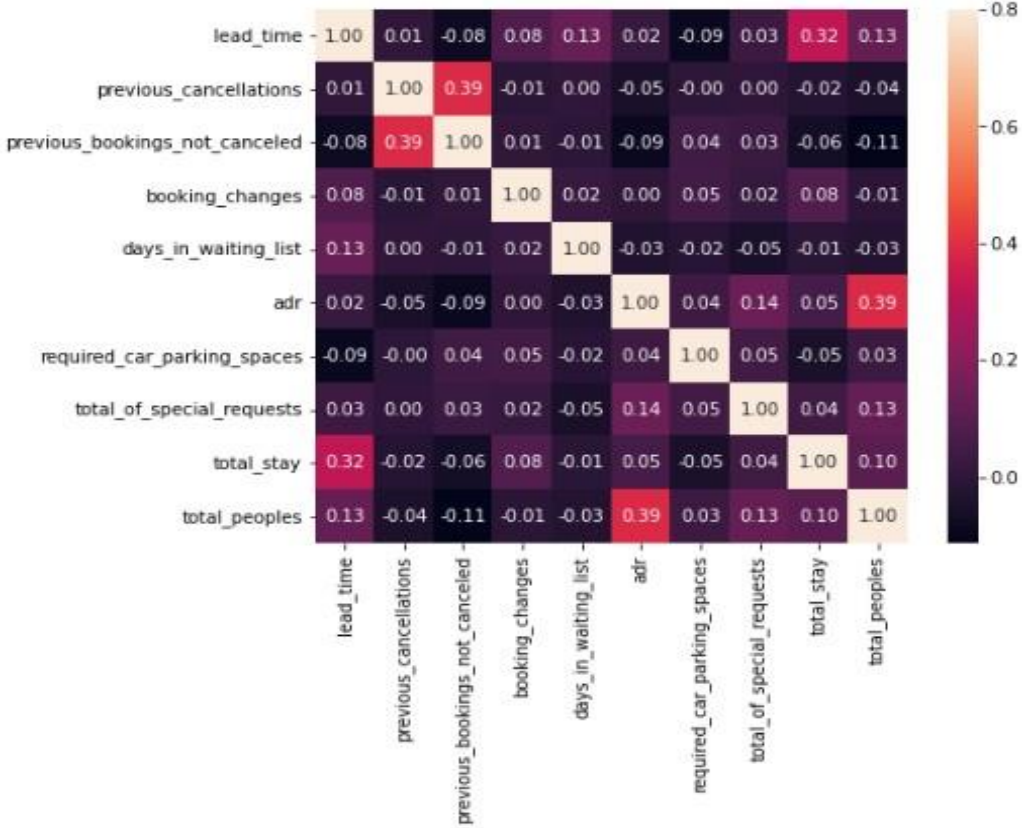
```
[34] #Addition of two columns
     hotel1['total_stay'] =hotel1['stays_in_week_nights']+hotel1['stays_in_weekend_nights']
```

```
#making new column totle_peoples by addition of babies,adults,children
print(hotel1['babies'])
print(hotel1['adults'])
print(hotel1['children'])
```

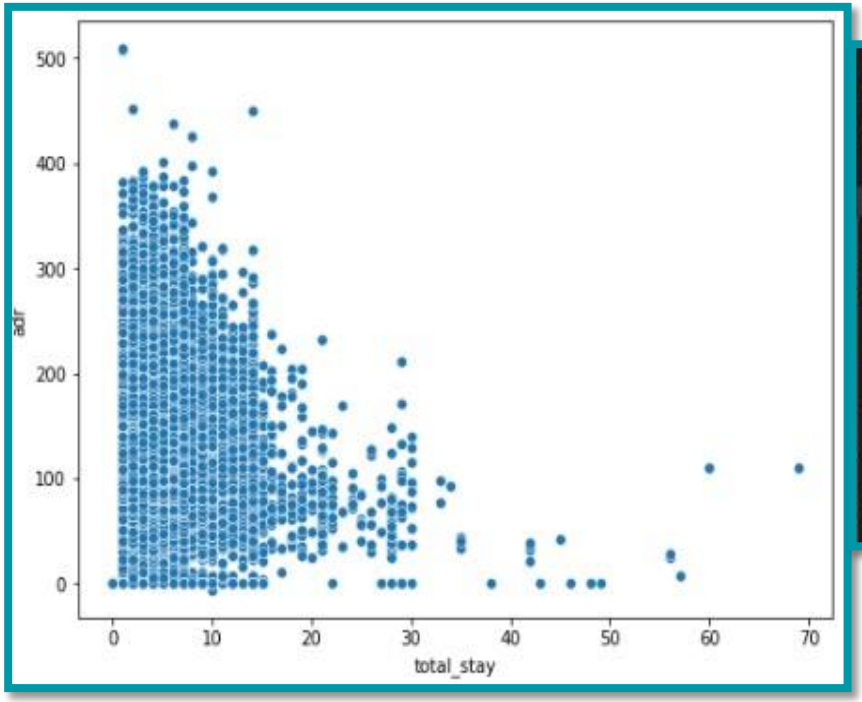
```
[36] #Addition of three columns
     hotel1['total_peoples']=hotel1['babies']+hotel1['adults']+hotel1['children']
```

Correlation matrix analysis

- Here ‘total_stay’ and 'lead_time‘ Have slight correlation.which means means customers are Plan reservation before arrival
- Adr(Average Daily Rate) is correlated with ‘total_peoples’which means no of peoples increases revenue increase.
- Also,’previous_cancellations’ and ‘previous_bookings_not_canceled’ are corelated to one another.which means repeated guest are those who do not cancelled there previous Bookings
- *Here ‘total_stay’ is also correlated with ‘lead_tim
And so more..



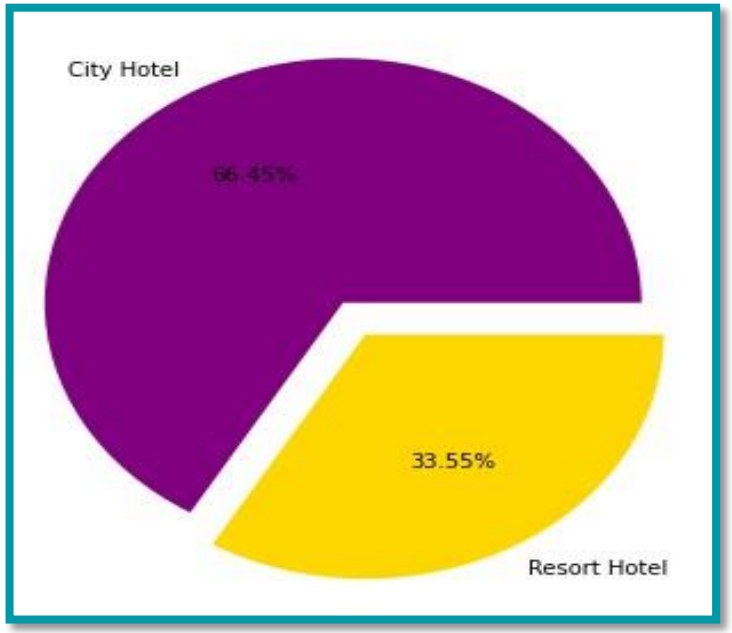
Scatter plot analysis




```
#Here we are just apply condition for better scatterplot
hotel1.drop(hotel1[hotel1['adr']>2000].index,inplace=True)


#Describe size for figure
plt.figure(figsize = (8,6))

#Analysis by using scatterplot
sns.scatterplot(y = 'adr', x = 'total_stay', data = hotel1)
plt.show()
```

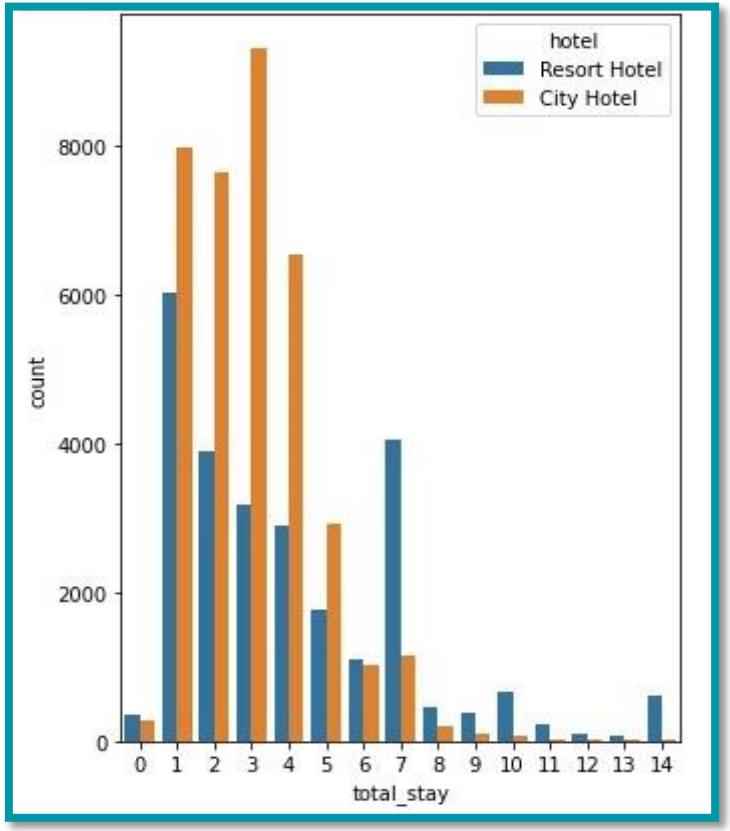
Hotel wise analysis

- **Type of hotel in the market and percentage in each Type?**



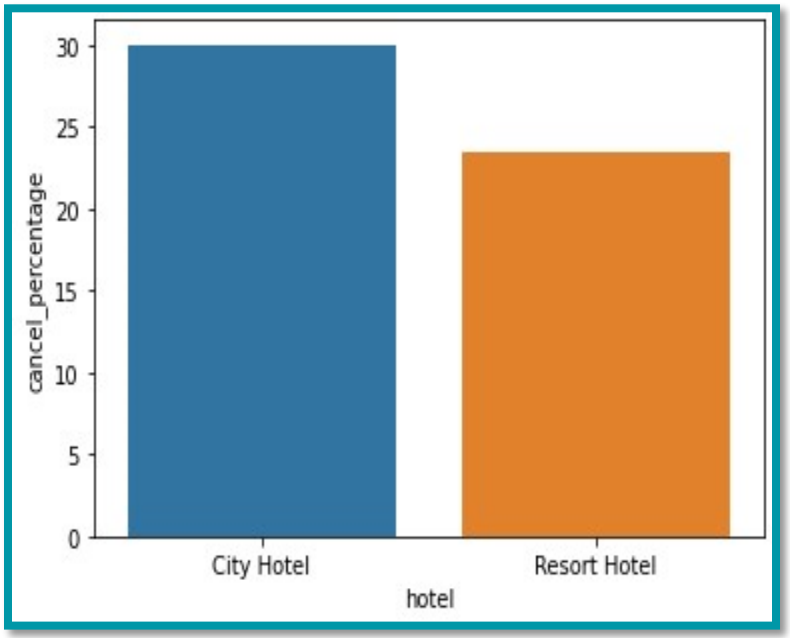
- According to Analysis city hotels are comparatively more expensive than resort hotels.
- city hotel having 66.45% and Resort hotel having 33.55% percentage in hotel market
- Resort hotel should me more expensive so, peoples are stick to city hotel

Which Hotel has higher lead time and preferred stay in each hotel?



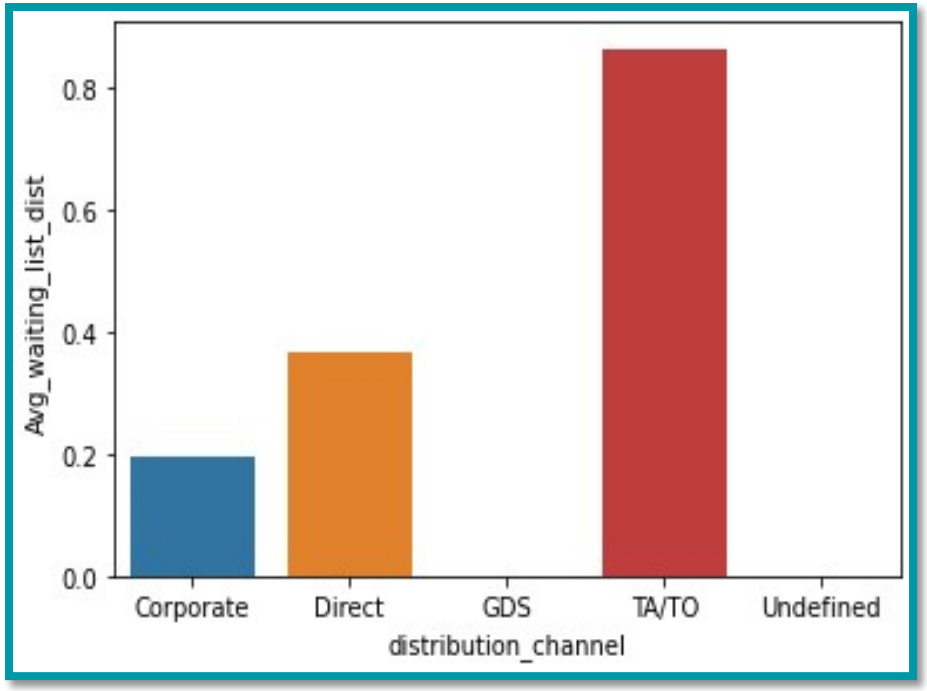
- Lead time of city hotel is more than Resort Hotel
- Resort hotel having longer stay compared to city hotel.i.e.For short stay peoples are choose City Hotel

Which hotel has a higher bookings cancellation rate?



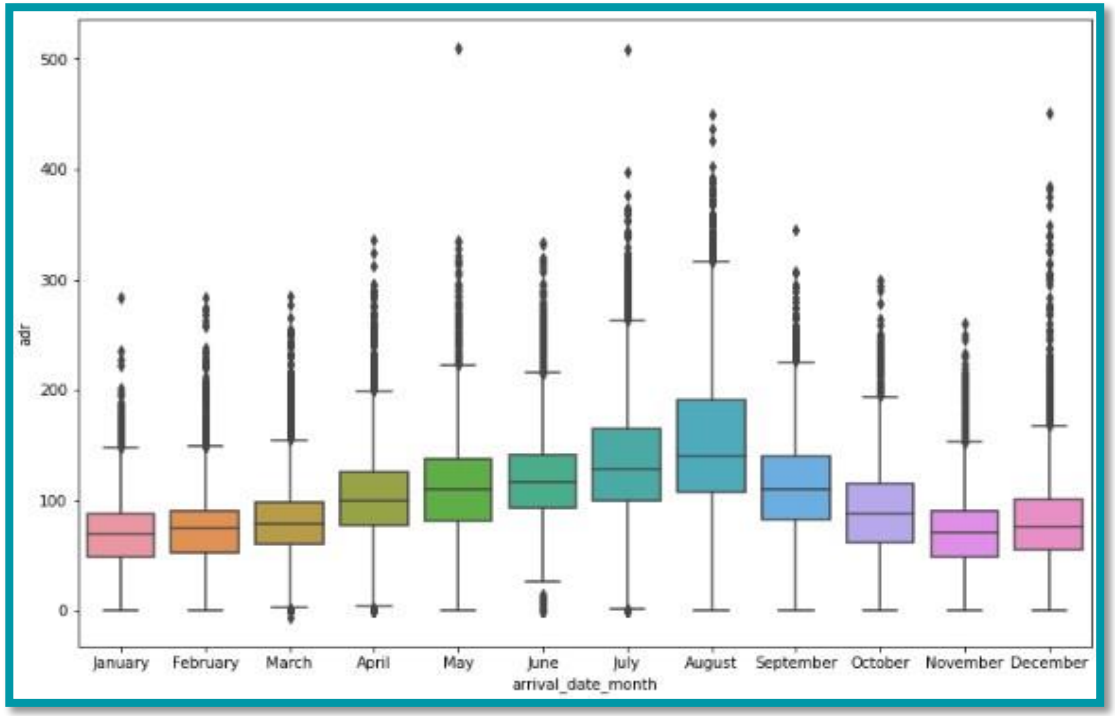
- City Hotel bookings are more cancelled than Resort hotel

Which distribution channel has a longer average waiting time?



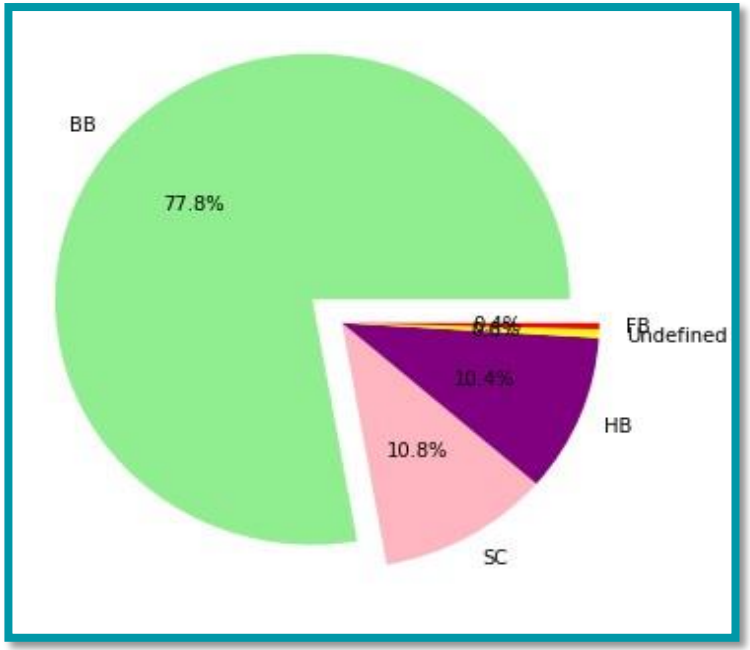
- TA/TO distribution channel having longer waiting time compared to Corporate and Direct
- The term “TA” means “Travel Agents” and “TO” means “Tour Operators”

Which month hotels have high revenue?



- Less peoples are visited to Hotel in January month.so,revenue having huge cut off
- But, hotels have large revenue in August due to more people visiting.

Which type of meal is booked?

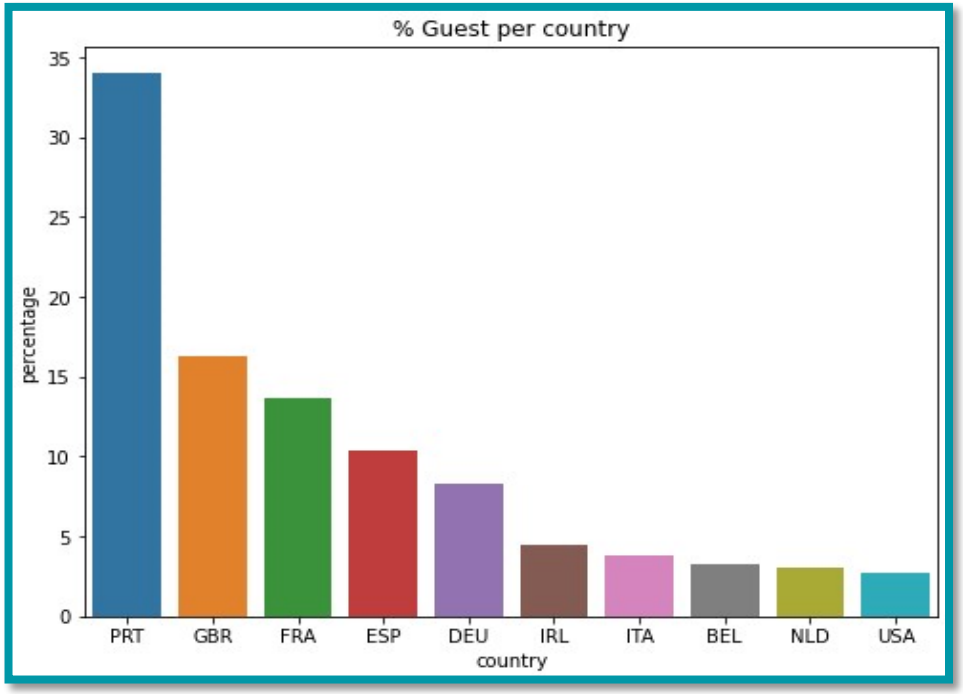


- **BB**.(i.e. Bed & Breakfast) is most preferable type of meal for 77.8% guest

Categories as per standard hospitality meal package

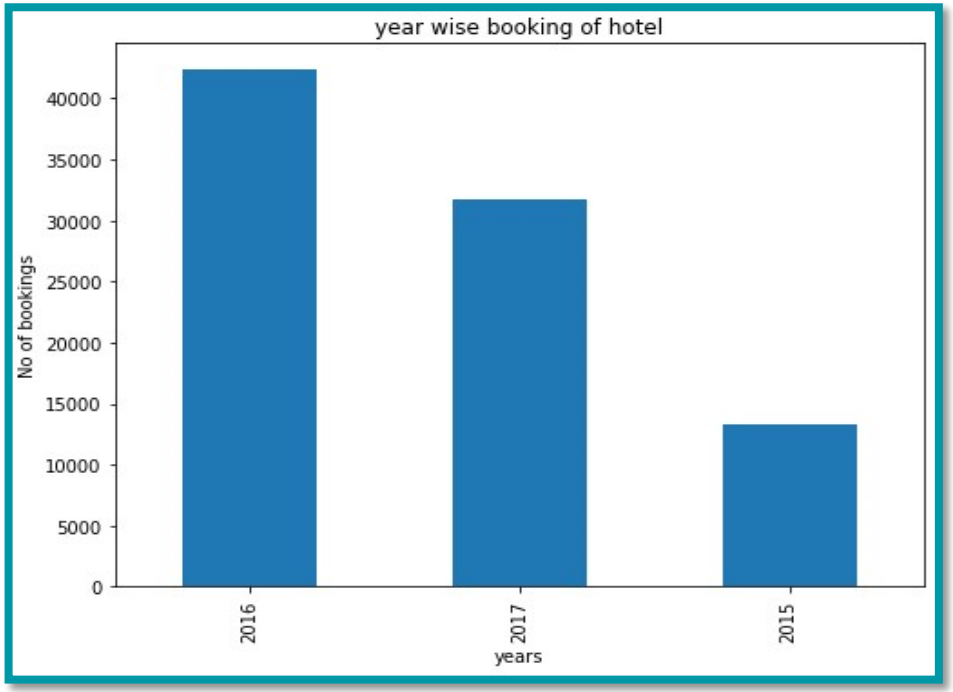
- **Undefined/SC** — no meal package
- **BB** — Bed & Breakfast
- **HB** —
  Half board (breakfast and one other m eal — usually dinner)
- **FB** —
  Full board (breakfast, lunch and dinne r)

Find from which country most guests come from?



- Most of the peoples are come from PRT (i.e. Portugal)

Year wise booking of hotel?



- From Graph it is clear that 2016 had higher bookings compared to 2017 and 2015.
- so,according to given data there is increment of booking with alternate years

Challenges

- Data set having huge data need to segregate.Also,dataset having lots of duplicate value.
- Lots of null values in the dataset.
- Handling null values and replace them with Zero or ‘text'. Do with care so that it doesn’t affect analysis.
- Choosing visualization for different analysis.

Conclusion

- Average daily rate (adr) is directly proportional to totle_peoples.No peoples increases then revenue must be increased.
- The percentage of city hotel is 66.45%.while the percentage of resort hotel is 33.55% is use to stay.So,City hotel connect more no of peoples and having higher lead time
- For longer stay peoples are choose Resort hotel and for short stay choose city hotel
- City Hotel bookings are more cancelled
- Here,TA/TO distribution channel having longer waiting time
- More people visit hotels in August and less people visit in January.
- BB.(i.e. Bed & Breakfast) is most preferable type of meal for 77.8% guest
- Most of the peoples are come from PRT (i.e. Portugal)
- Year 2016 having higher bookings compared to year 2017 and 2015.

Reference

- Almabetter
- Geeksforgeeks
- Stackoverflow
- w3schools

Thank you