HOTEL BOOKING ANALYSIS

Samadrita Purakayastha
Shivam Pandey
Dinesh Kumar Nayak
Data science trainees,
AlmaBetter, Bangalore

Abstract:

In this a, we will discuss exploratory data analysis and data visualization of the hotel booking data set. In this project, we need to find the average fees of the hotel and explore the data.

The data set contains the column names like the hotel, is_canceled, lead_time, etc.

Keywords:Hotels,analysis,data

1.Problem Statement

- ❑ For this project we will be analyzing Hotel Booking data. 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 and many more.
- ❑ Hotel industry is a very volatile industry and the bookings depends on above factors and many more.
- ❑ The main objective behind this project is to explore and analyze data to discover important factors that governs the bookings and give insights to hotel management ,which can perform various campaigns to boost the business and performance.

2. Introduction

Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week and many more. This makes analyzing the patterns available in the past data more important to help the hotels plan better. Using the historical data, hotels can perform various campaigns to boost the business. We can use the patterns to predict the future bookings using time series or decision trees.

We will be using the data available to analyze the factors affecting the hotel bookings. These factors can be used for reporting the trends and predict the future bookings.

3. Types of key metrics for hotel bookings

- The number of cancellations
- Number of bookings on weekday vs weekends
- Most preferred meal types

- Country wise bookings
- New customers acquired
- Customer lifetime value of the existing customers
- Type of rooms preferred by customers
- Booking types,
- Hotels available for booking
- The revenue of the hotels

We will be using various lenses to look through the data to analyze patterns associated with each segment such as:

- The type of hotel
- Day of week
- Type of customers
- Type of rooms

4. Data Sets

hotel :Resort Hotel or City Hotel
is_canceled : Value indicating if the booking was canceled (1) or not (0)
lead_time : Number of days that elapsed between the entering date of the booking and the arrival date
arrival_date_year : Year of arrival date
arrival_date_month : Month of arrival date
arrival_date_week_number : Week number of year for arrival date
arrival_date_day_of_month : Day of arrival date
stays_in_weekend_nights : Number of weekend nights
stays_in_week_nights : Number of week nights.
adults : Number of adults
children : Number of children
babies : Number of babies
meal : Type of meal booked.
country : Country of origin
market_segment : Market segment designation.(TA/TO)
distribution_channel : Booking distribution channel.(T/A/TO)
is_repeated_guest : is a repeated guest (1) or not (0)
previous_cancellations : Number of previous bookings that were cancelled by the customer prior to the current booking
previous_bookings_not_canceled : Number of previous bookings not cancelled by the customer prior to the current booking
reserved_room_type : Code of room type reserved.
assigned_room_type : Code for the type of room assigned to the booking.
booking_changes : Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
deposit_type : No Deposit, Non Refund , Refundable.
agent : ID of the travel agency that made the booking
company : ID of the company/entity that made the booking .
days_in_waiting_list : Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type : type of customer. Contract,Group,transient,Transient party.
adr : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
required_car_parking_spaces : Number of car parking spaces required by the customer
total_of_special_requests : Number of special requests made by the customer (e.g. twin bed or high floor)
reservation_status : Reservation last status.

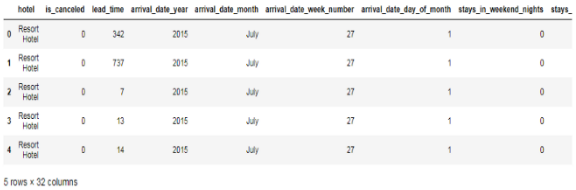
6. Steps involved:

- Import the libraries need in the project.

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

- Read and view the data of hotel booking demand with the help of pandas read_csv method.

```
data=pd.read_csv('hotel_bookings.csv')
data.head()
```



	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_
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	

5 rows × 32 columns

- Checking number of data types in the train dataset:

```
data.dtypes.value_counts()
```

We see that there are 4 float64 columns, 16 int64 columns, and 12 object columns. We have to find the hotel booking demand based on the data. First, we need to find the type of hotel people are booking more.

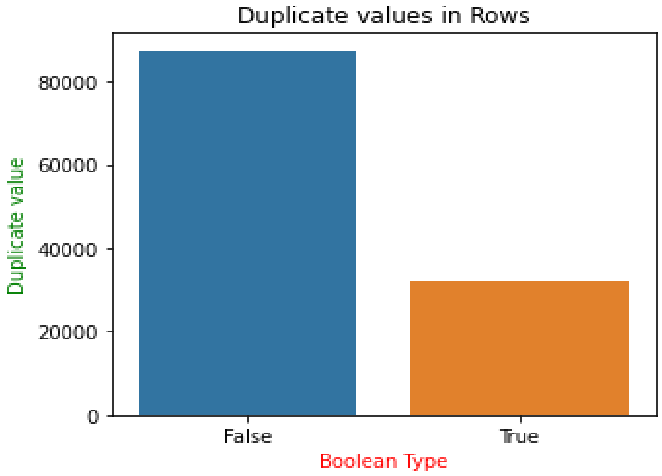
```
df1 = df.copy()
df1['hotel'].value_counts()
#output:
City Hotel       79330
Resort Hotel     40060
Name: hotel, dtype: int64
```

- let’s see duplicate value in rows using Boolean Type

```
data.duplicated().value_counts()
```

- Plotting a graph of duplicate value

```
plt.figure(figure=(5,4))
sns.countplot(x=data.duplicated())
plt.title(‘Duplicatedvalues in Rows’,colour=’black’)
ply.ylabel(‘duplicate value’,color=’green’)
plt.xlabel(‘boolean type’,color=’red’)
```



- Filling up the Null values

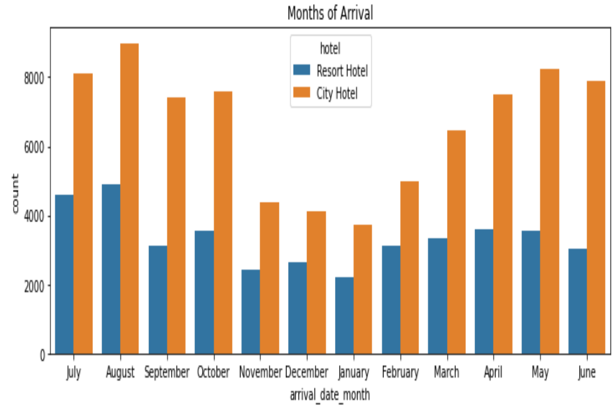
```
data['agent'].fillna(0,inplace=True)
data['company'].fillna(0,inplace=True)
data['country'].fillna('others',inplace=True)
data['children'].fillna(0,inplace=True)
```

- People are booking city hotels more than Resort hotels. Now we want to know in which month the people book the hotel.

```
df1['arrival_date_month'].value
_counts()#output:
August       13877
July         12661
May          11791
October      11160
April        11089
June         10939
September    10508
March         9794
February      8068
November      6794
December      6780
January       5929
Name: arrival_date_month,
dtype: int64
```

- Plotting the months against hotel type with seaborn library.

```
plt.figure(figsize=(12,4))
sns.countplot(x='arrival_date_m
onth', hue = 'hotel', data=
dfdf1)
plt.title('Months of Arrival')
plt.show()
```

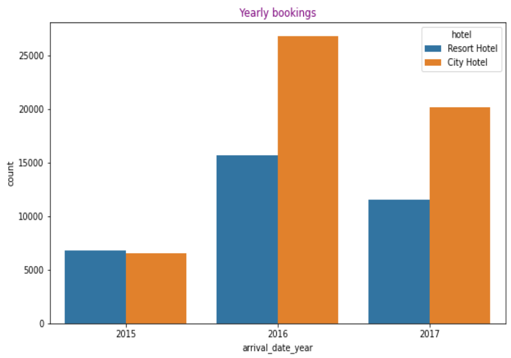


- Now, find the year in which most booked hotel.

```
df1['arrival_date_year'].value_
counts()
```

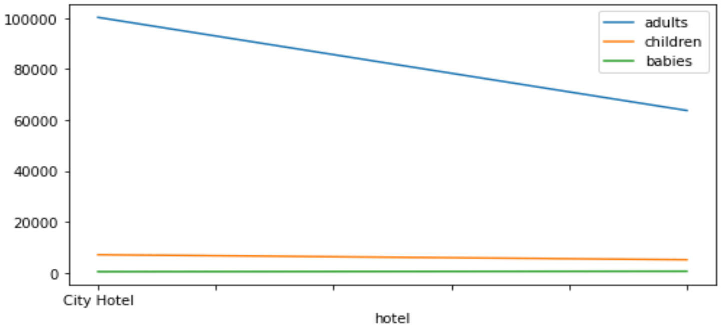
- Plotting the year column with bar chart.

```
Plt=figure(figsize(10,6)
Sns.countplot(x=df[‘arriv
al_date_year),hue=df[‘hot
el’]
Plt.title(“Yearly
Bookings”,colour purple)
```



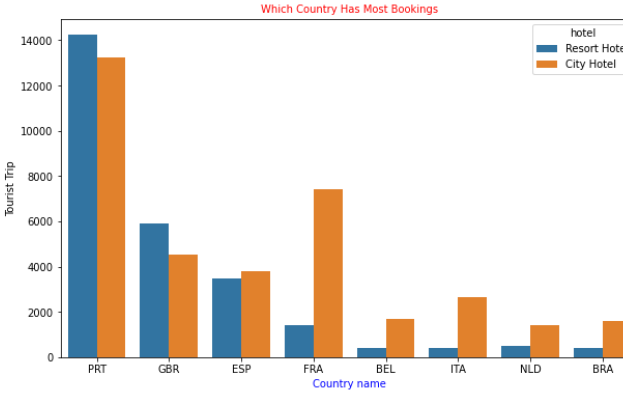
- Visited no of adults children and babies

```
df1=data.groupby(['hotel'])['adults','chil
dren','babies'].sum()
plt.rcParams['figure.figsize
']=(8,4)
plt.plot(df1)
plt.title("Visited no of Adults,Children a
nd babies in each hotel",color='black')
plt.ylabel('Number of Visiters',color='red
',fontsize=6)
plt.xlabel('hotel',color='gr
een',fontsize=6)
df1.plot()
```



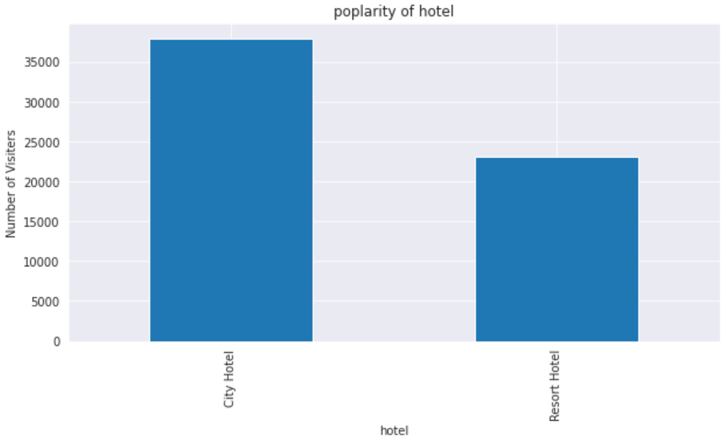
- Let’s check which Country is visited most

```
top_countries=('PRT','GBR','FRA','ESP','DUE','ITA','TRL','BEL','BRA','NLD')
data_1=data.loc[data['country'].isin(top_countries)]['country']
plt.rcParams['figure.figsize'] = (10, 6)
ax = sns.countplot(data_1, hue = data['hotel'])
ax.set_xlabel('Country name',color= 'blue',fontsize =10)
ax.set_ylabel('Tourist Trip',color='black',fontsize =10)
ax.set_title('Which Country Has Most Bookings',color='red', fontsize =10)
plt.show()
```



- Let’s check which Hotel gets more special requests

```
df2=data.groupby('hotel')['total_of_special_requests'].sum()
px=df2.plot.bar(figsize=(10,5))
plt.title("poplarity of hotel")
plt.ylabel('Number of Visiters')
plt.xlabel('hotel')
```

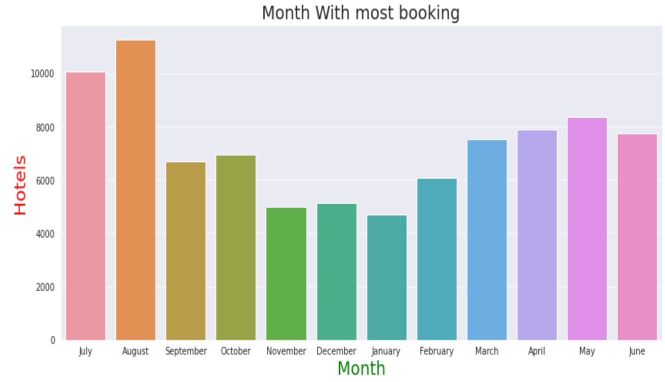


- Let’s check Room price analysis

```
city_hotel=city.groupby(['arrival_date_month'])['adr'].mean().reset_index()
city_hotel
final=resort_hotel.merge(city_hotel,on='arrival_date_month')
final.columns=['month','price_for_resort','price_for_city_hotel']
```

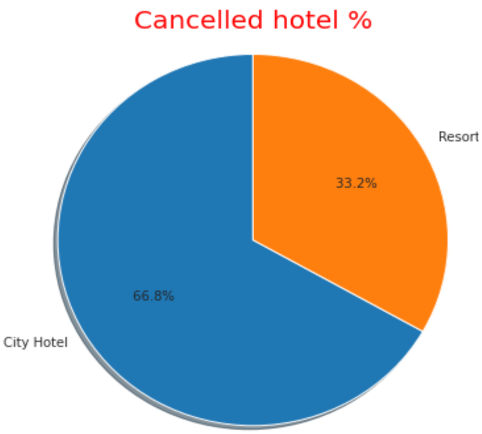
- Let’s check Month with max bookings

```
plt.figure(figsize = (14,6))
sns.set_style("darkgrid")
ax = sns.countplot(x = data['arrival_date_month'], data = data)
ax.set_xlabel('Month',color='green', fontsize = 20)
ax.set_ylabel('Hotels',color='red', fontsize = 20)
ax.set_title('Month With most booking',fontsize =20)
plt.show()
```



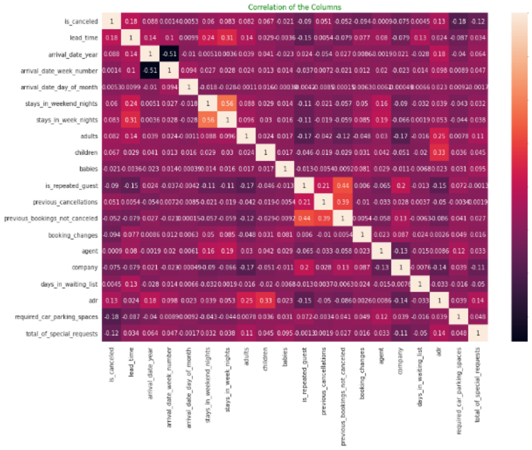
- Let’s check which Hotel gets more Cancellation

```
hotel_index=canceled_hotel.index
hotel_value=canceled_hotel.values
plt.rcParams['figure.figsize']=(8,6)
plt.pie(hotel_value,labels=hotel_index,autopct='% 1.1f%%',shadow=True,startangle=90)
plt.axis('equal')
plt.title("Cancelled hotel %",color='red',fontsize=20)
plt.show()
```



- Let’s check Correlation with each columns

```
fig, ax = plt.subplots(figsize=(15,10))
sns.heatmap(data.corr(),annt=True)
plt.title("Correlation of the Columns",color='green')
```



8. Conclusion:

- Around 60% bookings are for City hotel and 40% bookings are for Resort hotel, therefore City Hotel is busier than Resort hotel. Also the overall adr of City hotel is slightly higher than Resort hotel.
- Mostly guests stay for less than 5 days in hotel and for longer stays Resort hotel is preferred.
- Both hotels have significantly higher booking cancellation rates and very few guests less than 3 % return for another booking in City hotel. 5% guests return for stay in Resort hotel.
- Most of the guests came from european countries, with most of guests coming from Portugal.
- Guests use different channels for making bookings out of which most preferred way is TA/TO.
- For hotels higher adr deals come via GDS channel, so hotels should increase their popularity on this channel.
- Almost 30% of bookings via TA/TO are cancelled.
- Not getting same room as reserved, longer lead time and waiting time do not affect cancellation of bookings. Although different room allotment do lowers the adr.
- July- August are the most busier and profitable months for both of hotels.
- Within a month, adr gradually increases as month ends, with small sudden rise on weekends.
- Couples are the most common guests for hotels, hence hotels can plan services according to couples needs to increase revenue.
- More number of people in guests results in more number of special requests.
- Bookings made via complementary market segment and adults have on average high no. of special request.
- For customers, generally the longer stays (more than 15 days) can result in better deals in terms of low adr.

And many more conclusions.

References-

1. GeeksforGeeks
2. Almabetter