Project Name - Hotel Booking Analysis

(Name: Kumar Mhaske)

Summery –

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests?

This project aims to create meaningful estimators from the data set we have and to perform Exploratory Data Analysis. The data set used for this analysis was provide by Alma better. This data set contains a single file that compares various booking information between two hotels: a city hotel and a resort hotel.

The tools for data analysis used in this project are the packages Numpy and Pandas, and to visualize and explore the data: Matplotlib and Seaborn.

Content of exploratory data analysis.

- Where do the guests come from?
- How much do guests pay for a room per night?
- How does the price per night vary over the year?
- Which are the busiest month?

- How long do people stay at the hotels?
- Bookings by market segment
- How many bookings were canceled?
- Which month has the highest number of cancelations?
- Repeated guest effect on cancellations.
- The number of nights spent at hotels.
- Hotel type with more time spent.
- Effects of deposit on cancellations by segments.
- Relationship of lead time with cancellation.
- Monthly customers and cancellations.

- Data preparation is the process of preparing the data by cleaning and transforming raw data before processing and analysis. It is an important step before processing and often involves reformatting data, making corrections to data, and combining data sets to enrich data.

- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.

We check the number of rows and columns in our data frame by the following command:

▾ Dataset Rows & Columns count

```
[ ]  # Dataset Rows & Columns
     df.shape

     (119390, 32)

[ ]  pd.DataFrame([df.shape],columns=['number_of_rows','number_of_columns'],index=['#'])
```

	number_of_rows	number_of_columns
#	119390	32

Exploratory Analysis and Visualization

Before we ask questions about the hotels, it would help to understand the guests' demographics, i.e., country, arrival date, month, customer type, etc. It's essential to explore these variables to understand the data better. A survey of this scale generally tends to have some selection bias. Let's begin by importing matplotlib.pyplot and seaborn.

Which countries have the most passengers ?

```
df=pd.read_csv(path)
df['total_passengers'] = df['adults'] + df['children'] + df['babies'] - df['is_canceled']
df[['country','adults','children','babies','is_canceled','total_passengers']].groupby('country').sum().nlargest(5,'total_passengers').reset_index()
```

	country	adults	children	babies	is_canceled	total_passengers
0	PRT	86131	3468.0	437	27519	62512.0
1	GBR	23223	1253.0	92	2453	22115.0
2	FRA	20291	1211.0	77	1934	19645.0
3	ESP	16615	1412.0	126	2177	15976.0
4	DEU	13703	477.0	18	1218	12980.0

This shows us the top 10 countries from where the bookings are made.

PRT — Portugal, GBR — United Kingdom, FRA — France, ESP — Spain, DEU — Germany

Missing Values/Null Values

Missing Values/Null Values

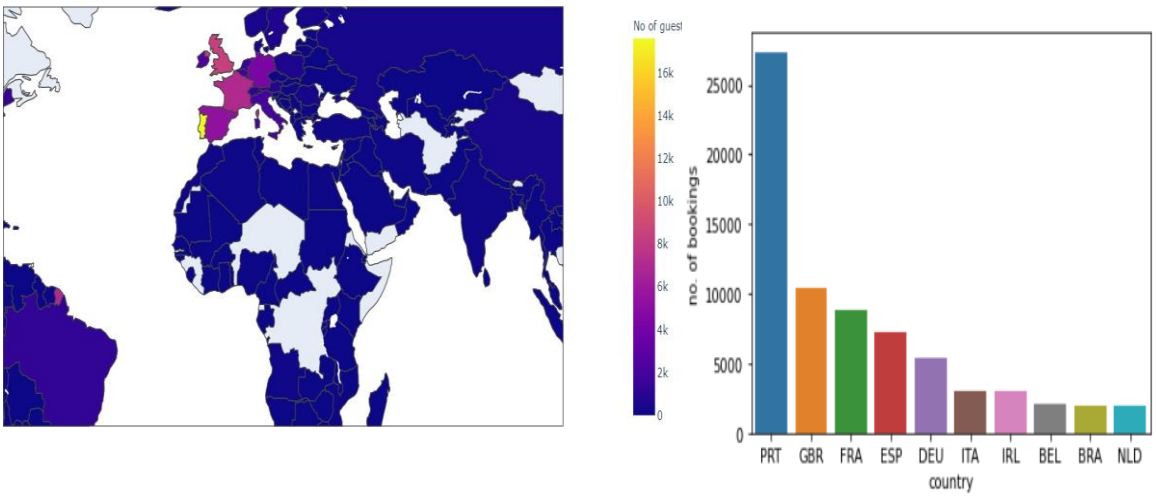
```
[ ] # Missing Values/Null Values Count
    print(df.isnull().sum())
```

```
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                            4
babies                              0
meal                                0
country                           488
market_segment                      0
distribution_channel                0
is_repeated_guest                   0
previous_cancellations              0
previous_bookings_not_canceled      0
reserved_room_type                  0
assigned_room_type                  0
booking_changes                     0
deposit_type                        0
agent                           16340
company                        112593
days_in_waiting_list                0
customer_type                       0
adr                                 0
required_car_parking_spaces         0
total_of_special_requests           0
```

We can visualize this information using a bar chart.

- Most of the customers from European countries like Portugal, Great Britain, France and Spain.

It appears that a disproportionately high number of bookings are from Portugal, probably because the hotel is located in Portugal itself. The second country is the United Kingdom which is approx. 75% behind.



the people country whom reserved a hotel with most number of babies and children.

```
tal_babies_and_children'] = df['babies'] + df['children']
ountry','distribution_channel','deposit_type','reservation_status_date','hotel','total_babies_and_children']].sort_values('total_babies_and_children',ascending=False)
```

	country	distribution_channel	deposit_type	reservation_status_date	hotel	total_babies_and_children
328	PRT	TA/TO	No Deposit	2015-07-12	Resort Hotel	10.0
46619	PRT	TA/TO	No Deposit	2016-01-14	City Hotel	10.0
78656	GBR	Corporate	No Deposit	2015-10-14	City Hotel	9.0
19718	PRT	Direct	No Deposit	2016-01-02	Resort Hotel	3.0
107837	PRT	Direct	No Deposit	2017-03-19	City Hotel	3.0

We find that Agent 9 has made most of the bookings.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3abea7c1c0>
```

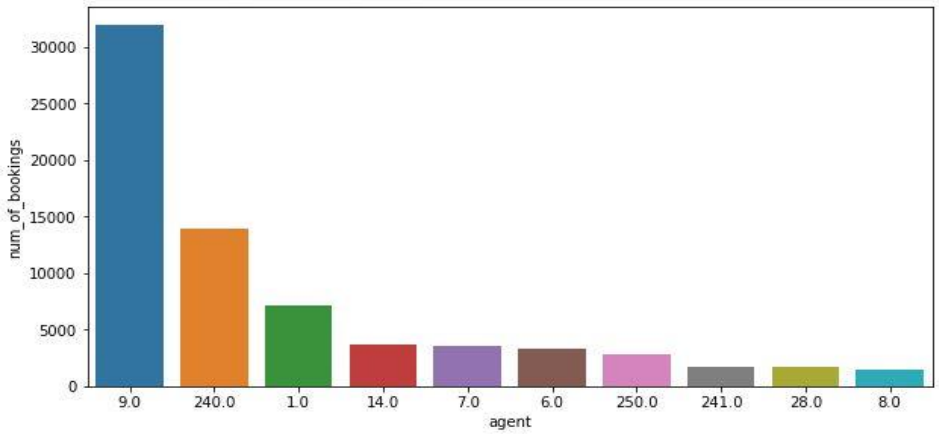
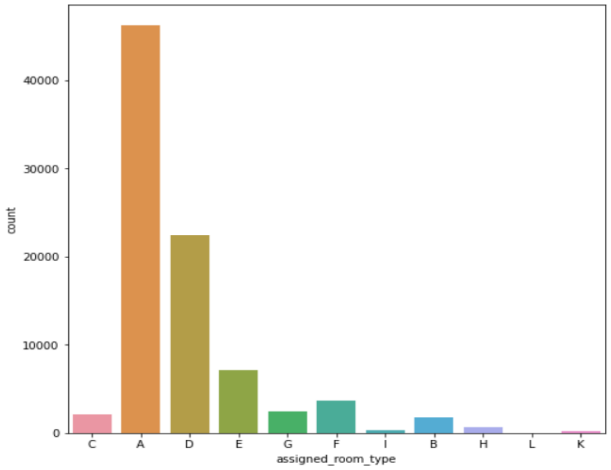
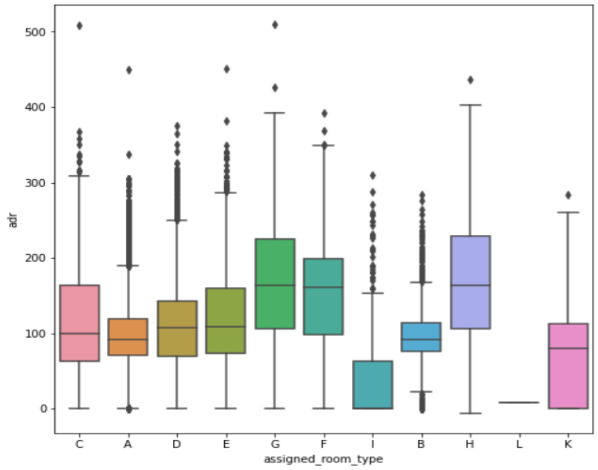


Chart analysis : **Agent no. 9 has made most no. of bookings.**

- Type A room is most demanded by customers.
- Room types C, G and H are some of the highest
- adr(average daily rate) generating rooms.

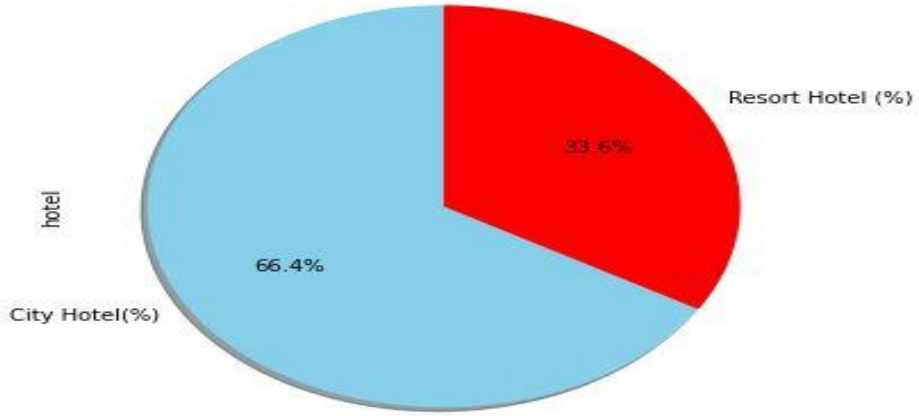




there are 66.4% city Hotels and 33.6% Resort Hotels

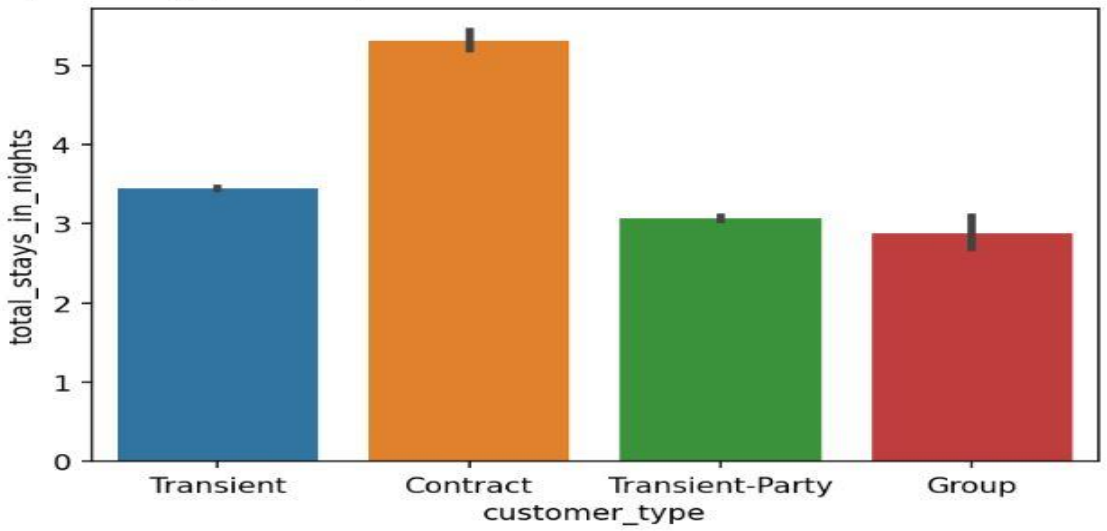
```
[ ]  City Hotel       79330
     Resort Hotel     40060
     Name: hotel, dtype: int64

     <matplotlib.axes._subplots.AxesSubplot at 0x7f3abed1eee0>
```



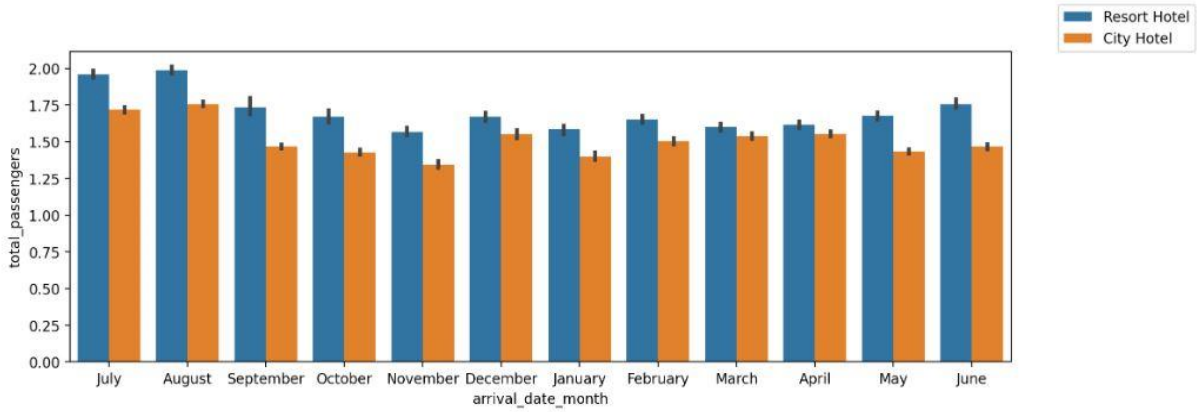
- Contract ' of Customer Types has the most stay duration.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1e3fee7490>
```



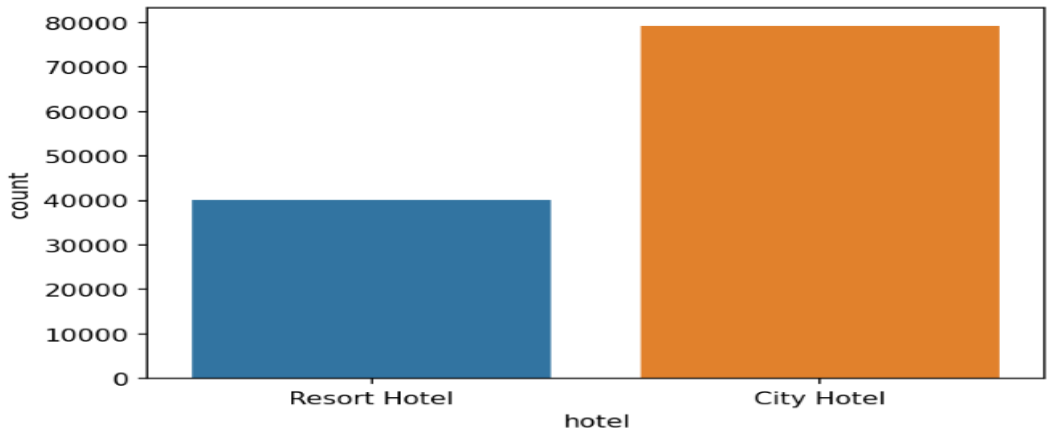
- In both hotels most bookings were made from July to August . In all months, the most passengers are for Resort Hotel ..

```
<matplotlib.legend.Legend at 0x7f1e3d688130>
```



- City Hotel has the most visitors.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f1e3febdd30>
```



Conclusion

- There are 66.4% city Hotels and 33.6% Resort Hotels.
- In month of january and february lead time is low.
- Contract ' of Customer Types has the most stay duration.
- City Hotel has the most visitors.
- The people from PRT country reserved a hotel with most number of babies and children.
- PRT Country has most customers.
- City Hotel has the maximum ADR (Average Daily Rate) which is 5400.0
- Agent no. 9 has made most no. of bookings.
- Most demanded room type is A, but better adr rooms are of type H, G and C also. Hotels should increase the no. of room types A and H to maximise revenue.
- The Most of the customers are from PRT, GBR,FRA and ESP Countries
- In both hotels most bookings were made from July to August . In all months, the most passengers are for Resort Hotel .