Hotel Booking Demand – An Exploratory Data Analysis using Python

Data Science Trainees:

Almabetter

Vineeth Kumar V
Aman Jain
Lalan Jaiswal
Susmita Sardar
Sanjay Khatri

Abstract

Hotel booking is an arrangement that we make to get a particular room/space in hotel at a particular time in future. Data that we used is taken from Almabetter Team Capstone project dashboard. We figured out 10 relevant questions and performed data cleaning and exploratory data analysis. Our experiment could help to understand the relations between these variables and how it can help to take better business decisions by the stake holders.

Acknowledgement

This project was completed by Susmita Sardar, Vineeth Kumar V, Aman Jain, Sanjay Khatri, Lalan Jaiswal. We are extremely to authors of projects which we referred for this project. We also like to thank those mentors who thought us various concepts of python that we used to complete this project.

Introductions

Hotel booking system are application that allow guest to book rooms etc. online. These application be run by the hotel chains or the booking websites which have tie up with the hotel. These details are then passed to the backend which would be managed by the hotel staff, system admin. This helps in improved efficiency, time management, competitive edge, reduced human error, and easier data collection.

Data Set

Following are the Column parameters used available in the data.

hotel : Resort Hotel / City Hotel

is_canceled:if the booking was canceled- 1 / not – 0

lead_time : No of days between the entering date of the booking into the PMS and the arrival date

arrival_date_year : Year of arrival date

arrival_date_month : Month of arrival date

arrival_date_week_number : Week no of year for arrival date

arrival_date_day_of_month : Day of arrival date

stays_in_weekend_nights : No of weekend nigh

stays_in_weekend_nights : No of weekend nights (Saturday / Sunday) the guest stayed / booked to stay at the hotel

stays_in_week_nights : No of week nights (Monday to Friday) the guest stayed / booked to stay at the hotel

adults : No of adults

children : No of children

babies : No of babies

meal : Type of meal booked.

country : Country of origin. market_segment : Market segment designation. In categories, “TA” - “Travel Agents” and “TO” - “Tour Operators”

distribution_channel : Booking distribution channel. “TA” - “Travel Agents” and “TO” - “Tour Operators”

is_repeated_guest : tells whether the booking name was repeated guest -1 / not- 0

previous_cancellations : No of previous bookings that were cancelled by customer prior to the current booking

previous_bookings_not_canceled : No of previous bookings not cancelled by customer prior to the current booking

reserved_room_type : Code of room type reserved.

assigned_room_type : Code for the type of room assigned.

booking_changes : No of changes made between booking till moment of check-in or cancellation

deposit_type: Tells customer made a deposit to guarantee the booking.

agent : ID of the travel agency that made the booking company : ID of the company that made the booking

days_in_waiting_list : No of days the booking was in the waiting list before it got confirmed to customer

customer_type : Type of customer

adr : Average Daily Rate = sum of all lodging transactions/total number of staying nights

required_car_parking_spaces : No of car parking spaces required

total_of_special_requests : No of special requests made by customer

reservation_status : Reservation last status, assuming the below categories

Canceled – customer cancelled the booking

Check-Out – customer has checked in and departed also

No-Show – customer did not check-in and did inform the hotel as well

reservation_status_date : Date at which the last status was set. Can be combined with Reservation Status to get booking cancelled or customer checked-out

Objective

To perform EDA on above data set and get meaningful insights, trends, conclusion in the hotel booking and how each factor correlate with other.

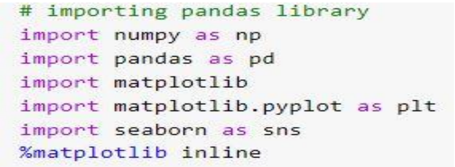
We use exploratory analysis using python to get insights.

We will try to observe the following questions in these analysis

1. Which hotel have the maximum number of bookings?
2. What is the percentage of cancellation?
3. Calculate the ADR with respect to distribution channel?
4. Find out the booking trends on total stays by the customers?
5. Which has the average ARD between hotels?
6. Which is the most preferred room type by the customers?
7. Market segment that has highest cancellation rate?
8. Which meal type is most preferred meal of customers?
9. Which Hotel makes More revenue?
10. Percentage of car parking space required?

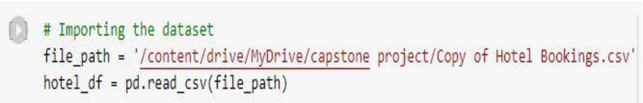
and Correlation of the columns

Importing the required Libraries



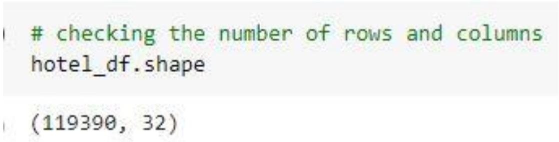
```
# importing pandas library
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Importing Data Set



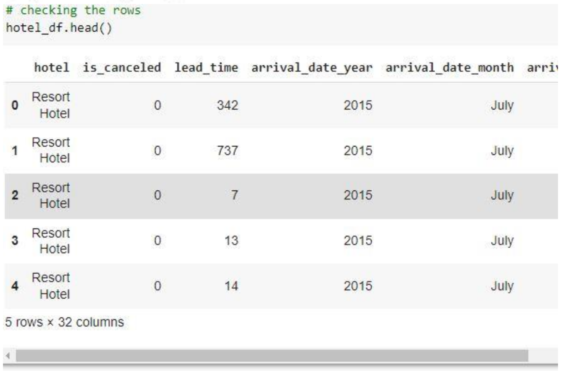
```
# Importing the dataset
file_path = '/content/drive/MyDrive/capstone project/Copy of Hotel Bookings.csv'
hotel_df = pd.read_csv(file_path)
```

Data Exploration



```
# checking the number of rows and columns
hotel_df.shape
```

```
(119390, 32)
```



```
# checking the rows
hotel_df.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arri
0	Resort Hotel	0	342	2015	July	
1	Resort Hotel	0	737	2015	July	
2	Resort Hotel	0	7	2015	July	
3	Resort Hotel	0	13	2015	July	
4	Resort Hotel	0	14	2015	July	

5 rows × 32 columns

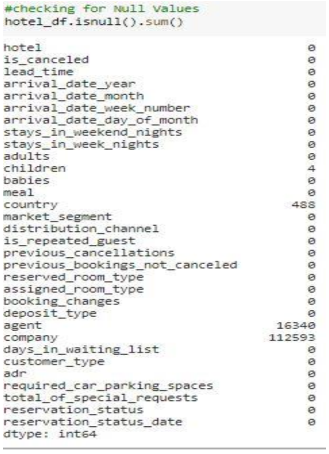
Data Cleaning

Before doing the EDA we need to clean our data by removing all the null values so that we can get correct outcome after doing EDA. While cleaning the data we will follow few steps:

* Removing the null and NaN values.

*   Dropping the duplicate rows.

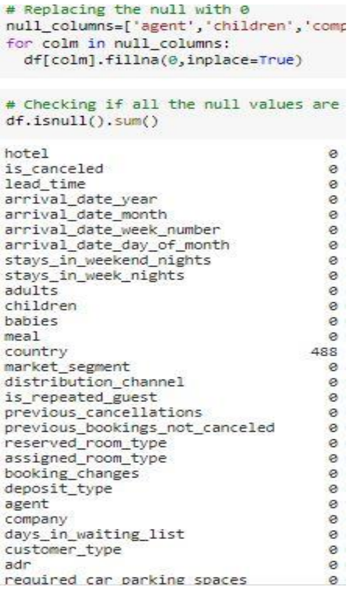
a) Checking null values



```
#checking for Null Values
hotel_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
reservation_status                     0
reservation_status_date                0
dtype: int64
```

b) Replacing Null value with zero

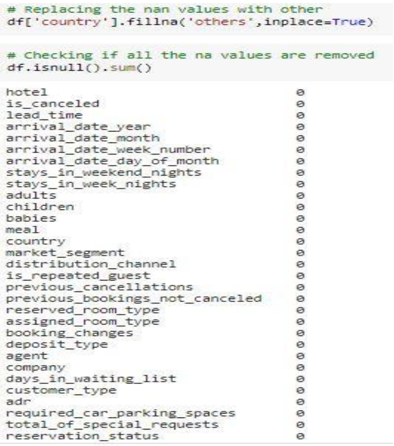


```
# Replacing the null with 0
null_columns=['agent','children','comp
for colm in null_columns:
  df[colm].fillna(0,inplace=True)
```

```
# Checking if all the null values are
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                           0
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                              0
company                            0
days_in_waiting_list               0
customer_type                      0
adr                                0
required_car_parking_spaces        0
```

c) Replacing null values of country with others

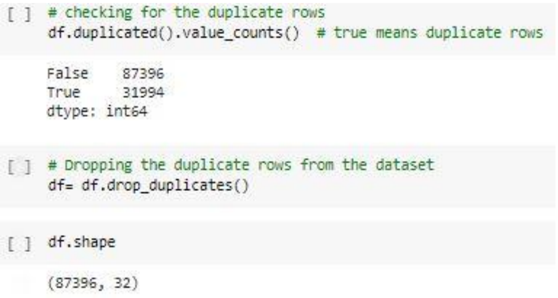


```
# Replacing the nan values with other
df['country'].fillna('others',inplace=True)
```

```
# Checking if all the na values are removed
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                           0
babies                             0
meal                               0
country                            0
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                              0
company                            0
days_in_waiting_list               0
customer_type                      0
adr                                0
required_car_parking_spaces        0
total_of_special_requests          0
reservation_status                 0
```

d) Removing Duplicates



```
[ ] # checking for the duplicate rows
    df.duplicated().value_counts()  # true means duplicate rows

    False    87396
    True     31994
    dtype: int64
```

```
[ ] # Dropping the duplicate rows from the dataset
    df= df.drop_duplicates()
```

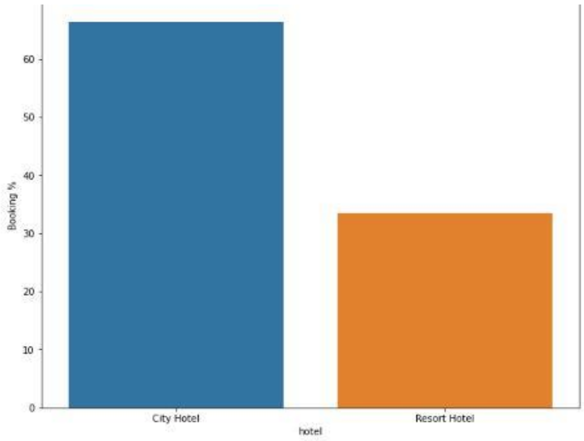
```
[ ] df.shape

    (87396, 32)
```

Exploratory Data Analysis

EDA is applied to investigate the data and summarize the key insights. It will give us the basic understanding of our data, it's distribution, null values and much more. We can either explore data using graphs or through some python functions.

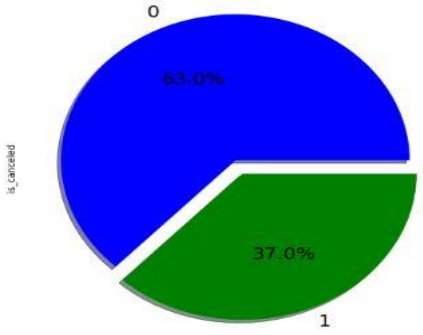
1.Which hotel have the maximum number of bookings?



Observation

From above graph we can see that city hotel have maximum number of booking than Resort hotel

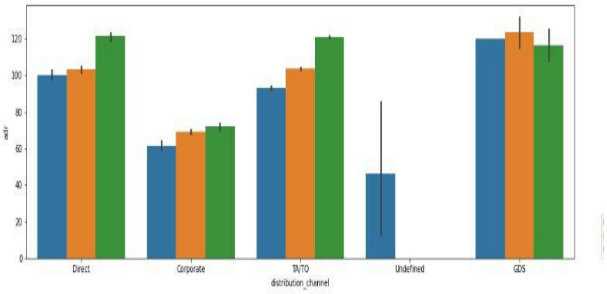
2.What is the pecentage of cancellation?



Observation

From the above piechart we see that:- 0 = Non cancelled 1 = cancelled. 27.5 % is cancelled

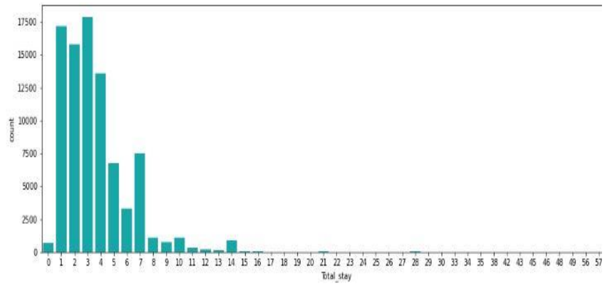
3.Calculate the ADR with respect to distribution channel?



Observation

Max ADR comes from GDS(overall) except for year 2017. Min ADR comes from Corporate and undefined channel in the 3 years

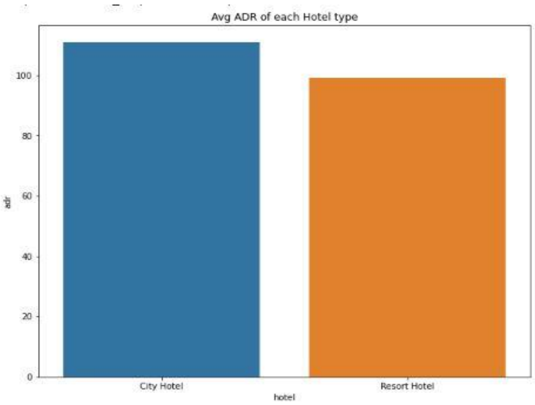
4.Find out the booking trends on total stays by the customers?



Observation

Most of the people stayed less than 5 days in the hotel and only few people. stayed beyond 10 days.

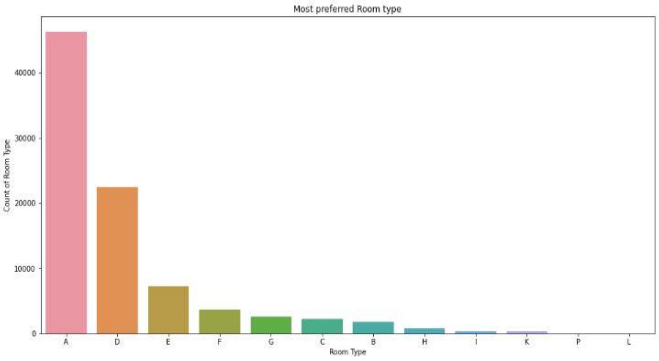
5.Which has the average ADR between hotels?



Observation

City hotel has the highest ADR. That means city hotels are generating more revenues than the resort hotels. More the ADR More is the revenue.

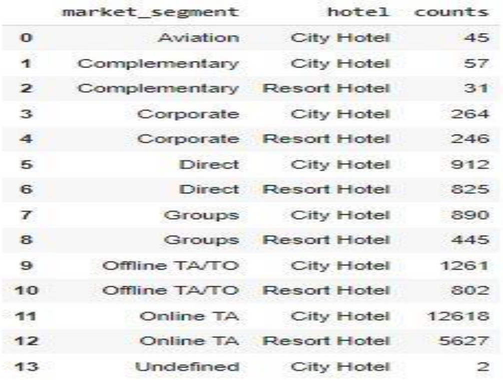
6.Which is the most preferred room type by the customers?



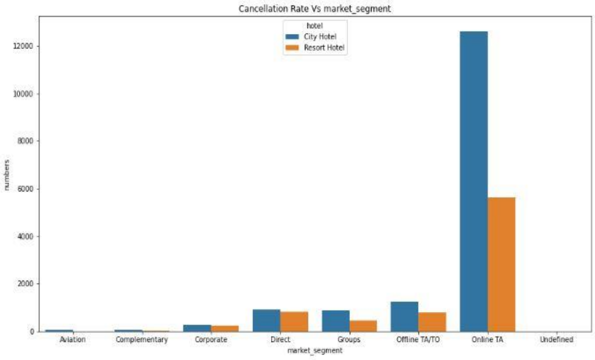
Observation

The most preferred Room type is "A".

7.Market segment that has highest cancellation rate?



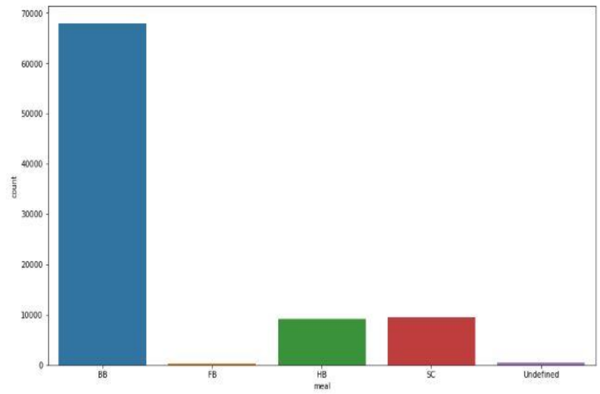
	market_segment	hotel	counts
0	Aviation	City Hotel	45
1	Complementary	City Hotel	57
2	Complementary	Resort Hotel	31
3	Corporate	City Hotel	264
4	Corporate	Resort Hotel	246
5	Direct	City Hotel	912
6	Direct	Resort Hotel	825
7	Groups	City Hotel	890
8	Groups	Resort Hotel	445
9	Offline TA/TO	City Hotel	1261
10	Offline TA/TO	Resort Hotel	802
11	Online TA	City Hotel	12618
12	Online TA	Resort Hotel	5627
13	Undefined	City Hotel	2



Observation

Online TA/TO has highest cancellation rate

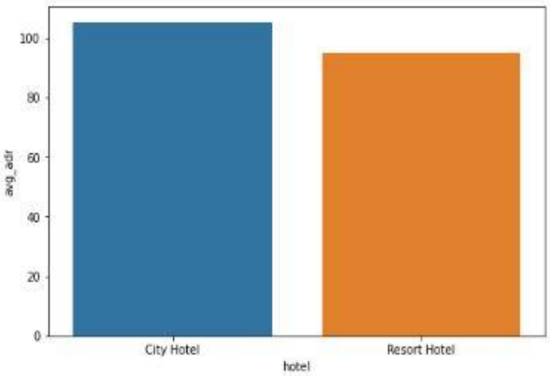
8.Which meal type is most preferred meal of customers?



Observation

BB is the most preferred meal by the customers.

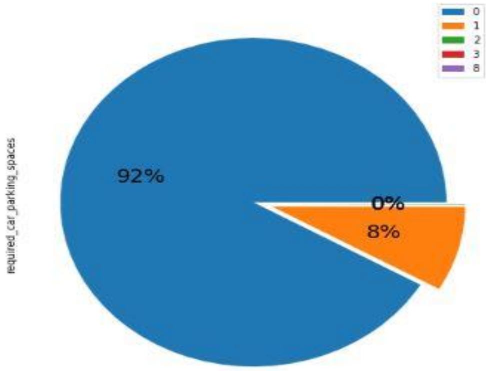
9.Which Hotel makes More revenue?



Observation

City hotel makes more revenue

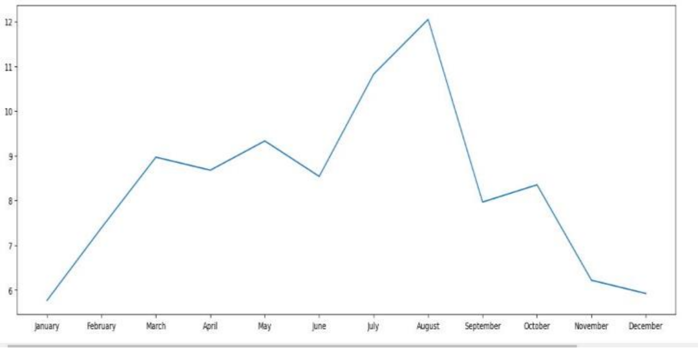
10.Percentage of car parking space required?



Observation

91.6 % guests did not required the parking space. only 8.3 % guests required only 1 parking space.

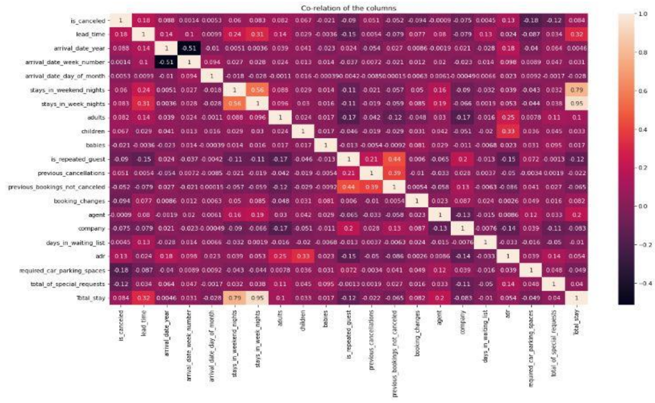
11.What are the **busiest booking month of a calendar year?**



Observation

Max Booking are done from the period of July to Mid-August in a calendar Year. Min Booking are done from the period of January - February and November – December. Booking are low in the start and end of a calendar year and it gradually increase and peaks at middle of a calendar year and then drops down.

Correlation of the columns



Observation

lead_time and total_stay is positively corelated.it means stay of cutsomer is proportional to lead time. is_canceled and same _room_alloted_or_not are negatively corelated. That means customer is unlikely to cancel his bookings if he doesn’t get the same room as per reserved room.

Adults, childrens and babies are corelated to each other. more customer more the ADR. Total stay length and lead time have slight correlation. This may means that for longer hotel stays people generally plan little before the the actual arrival.

Conclusion

* Maximum guests are booking city hotel over resort hotel.

* The percentage of cancellation is 27.5% and the percentage of non-cancellation is 72.5%.

* Since GDS gives better ADR, it would be the best among distribution channel as it gives higher revenue even with lower occupancy.

* Most of the people stayed less than 5 days in the hotel.

* City hotel has the highest ADR. That means city hotels are generating more revenues than the resort hotels. More the ADR More is the revenue.

* After analysing we see that the most preferred Room type is "A".

* Online TA/TO has highest cancellation rate.

* BB is the most preferred meal by the guests.

* After analysing we see that city hotel makes more revenue than resort hotel.

* 91.6 % guests did not require the parking space.

* Max. Booking happens July to Mid-August. Marketing and booking offer provided in these months can yield higher booking.

Challenges

1. Figuring out the right questions from raw data
2. Data cleaning and identifying the right tool and visualization technique.
3. Time restrictions of meeting since most of the team members where working professionals.

References

- *Github*
- *Kaggle*
- *Analytics Reports*