EDA-Hotel Booking Analysis

Dinesh Shivaji Jagtap
Data science trainee
Alma Better, Bangalore

Abstract:

The hotel sector is very unpredictable, and bookings depend on a wide range of variables, including hotel type, seasonality, day of the week, meal type, hotel type and many more, Also Most hotel owners like a hotel that is operating at full capacity and bringing in business while running a profitable and difficult hospitality business. To assist the hotels in making better plans, and also to assist the guest for booking the hotel in optimized value with safe stay. It is now even more crucial to analyze the patterns present in the historical data. In order to increase revenue, hotels can run a number of campaigns using past data. Also, in-order to provide proper details to guests. With the use of bar chart and line chart, we can use the patterns to predict the future reservations.

1. Problem Statement

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order 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 hotel booking dataset can help you explore those questions!

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, among other things.

2. Introduction

People look at a variety factor while looking for a hotel to stay at going for vacation, including price, location, availability, and parking, meal and also safe stay.

Price of hotel also varying according to the season, number of guests, month, location of hotel, number of days to stay, etc.

The objective of this project is to deliver insights to understand to best time of place, with optimal price the tourist to choose the right hotel, right price, Proper and safe stay also it helps Hotel Management to take right decision for making any changes in service level.

Following roadmap, I have decided to word and try to get the decided insights from given data,

1. Loading the dataset in colab-Notebook
2. Check the first view of the dataset
3. Check the information of dataset
4. Checking the null values and cleaning of null values
5. Understanding the variables
6. Perform some data analysis using data visualization
7. Conclusion

3. Keywords

There are some keywords we will be using.

- **hotel**: type of hotels
- **is_canceled**: canceled or not
- **lead_time**: no. of days before actual arrival in the hotel
- **arrival_date_year**: year of booking
- **arrival_date_month**: month of booking
- **arrival_date_week_number**: week number of the year in which booking
- **arrival_date_day_of_month**: arrival month date
- **stays_in_weekend_nights**: no. of weekends guest stayed
- **stays_in_week_nights**: no. of weekdays guest stayed
- **meal**: BB – Bed & Breakfast HB – only two meals including breakfast meal FB – breakfast, lunch, and dinner
- **market_segment**: TA: Travel agents TO: Tour operators
- **previous_cancellations**: cancellation in past
- **previous_bookings_not_canceled**: not canceled in the past.

4. Step Involved

- **Loading the dataset in Notebook**

  I have initially uploaded the given data in google drive and given the path of Hotel Booking analysis into the colab using python function and code.

- **First view of data**

  After loading the data set, I have checked the information using Python function. Data contains the shape of 119390 rows and 32 columns. Also, I have checked and lookout the variables

- **Cleaning the data**

  Provided data is not clean and proper to do the analysis, hence we need to clean the data properly to get the desired expected result. Data cleaning is the process to remove the unwanted values, undesired features, etc.

- **Removing of Null values**

  Initially I have checked to data and applied some Python function and it is found that column ‘company’ and ‘agent’ have huge amount of null data which might get the wrong result and ‘children’ and ‘country’ have minimum null data. So we will drop the ‘company’ column. And for remaining column we replaced the Null value to 0.

5. 1: Exploratory Data Analysis

Following are the observation and result using Data analysis and visualization

5.1.1 Hotel Booking Percentage by using Pie chart:

For below analysis, I have considered the whole data of booking, and it is found that city hotel has 66% preference and Resort hotel have 34% of preference. City hotel have more preferred by customers rather than the resort hotel.

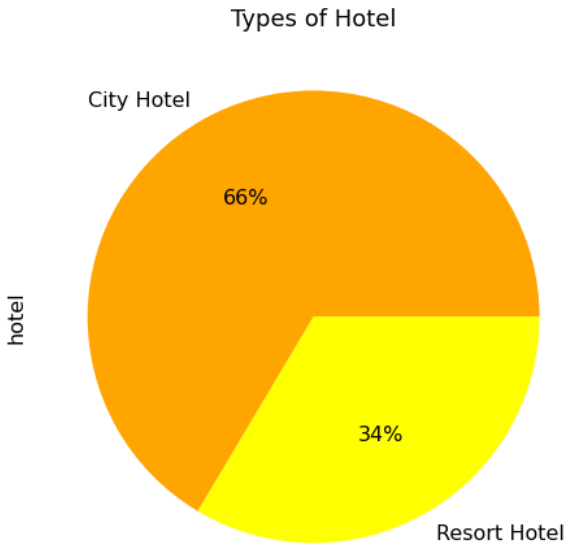

Fig 1: Hotel Booking Percentage

5.1.2 Hotel Cancellation Data

For the below analysis, I have considered the overall data and it is found that over 37% of booking were cancelled by the customers. Which is slightly high

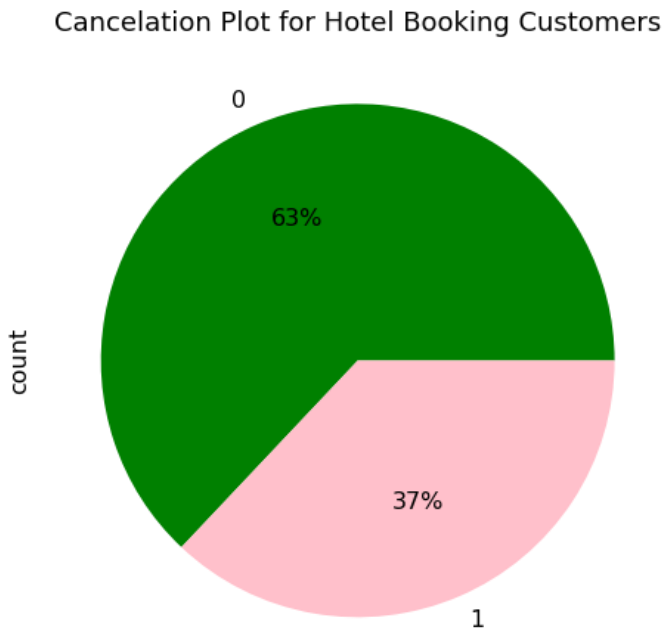

Fig 2 Hotel Cancellation Percentage

5.1.3 Customer Arrival Data Per Year

Count plot: A count plot is kind of like a histogram or a bar graph for some categorical area. It simply shows the number of occurrences of an item based on a certain type of category. I used the data 'arrival_date_year' column from the hotel booking dataset with arrival per year.

We can observe that number of arrivals seems to be high in year 2016 while the bookings seem to be less in 2015 and 2017. But number of arrivals in year 2017 has been drastically down hence hotel management need to check for the same.

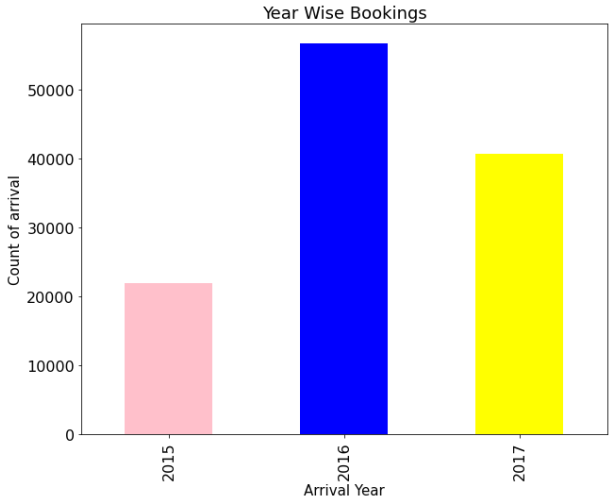

Fig 3: Year Wise Arrival data

5.1.4 Customer Arrival Data Per Year for Each Hotel

For below plot we used the dataset for hotel type and arrival data per year.

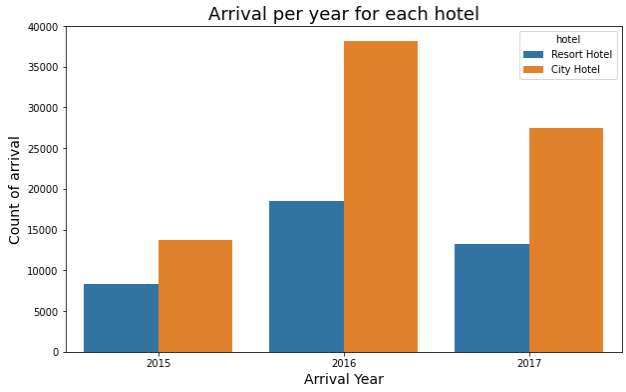

Fig 4: Year Wise Arrival data for Hotel Types

We can observe that number of arrivals seems to be high in year 2016 for city hotel compared to Resort Hotel, while the bookings seem to be less in 2015 and 2017 for both resort hotel and city hotel.

5.1.5 Customer Arrival Data Per Year for Each Hotel

For below plot we used the dataset for hotel type and arrival data per month.

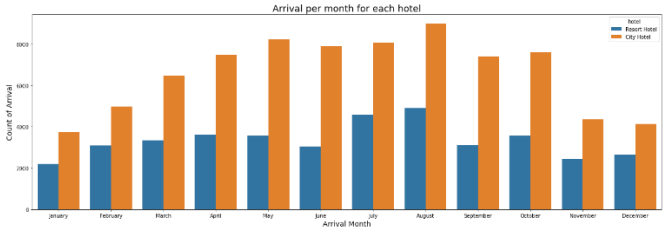

Fig 5: Month Wise Arrival data for Hotel Types

We can observe that number of arrivals seems to be high in month of May, June, July and August for city hotel.
But number of arrivals for Resort Hotel is seems high only for July and August.
For initial month number of arrival is less compared to mid-months.

5.1.6 Number of Bookings vs Country

We used data for top 10 countries with number of total bookings.

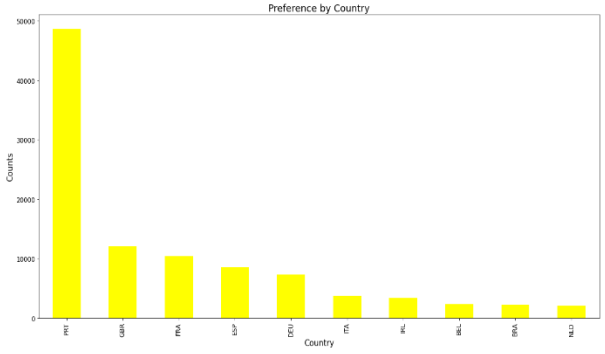

Fig 6: Booking Preference by Country

We found that PRT have more preference followed by GBR.

5.1.7 Number of Cancelled Bookings vs Hotel Types

We used data of booking cancelled and not cancelled by customers along with type of hotel. (0= Not cancelled and 1=Cancelled)

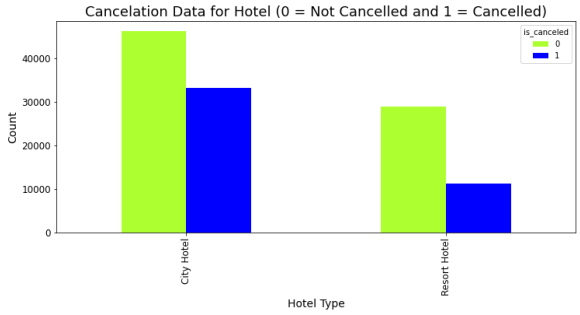

Fig 7: Cancelled Booking vs Hotel Type

We can see that number of cancellations of city Hotel seems very higher than the Resort Hotel.

5.1.8 Average Daily Rate vs Hotel Types

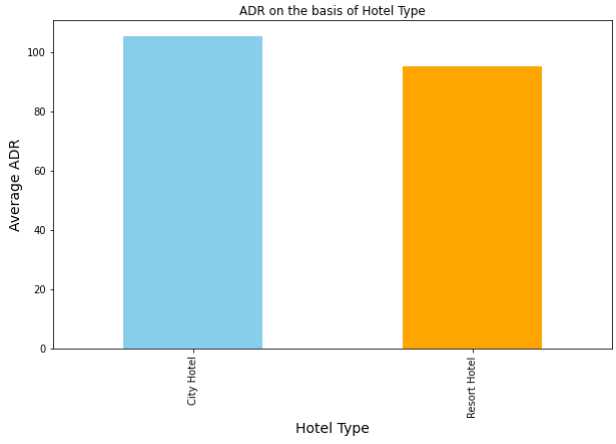

Fig 8: Average ADR vs Hotel Type

ADR for city hotel is slightly more than the Resort hotel, not much difference found in this.

5.1.9 Average Daily Rate vs Hotel Types

Line plot:

The line plot is used to plot the average daily rate (ADR) per month and year. The year is shown in different colors,

Blue indicates year 2015

Red indicates year 2016

Green indicates year 2017

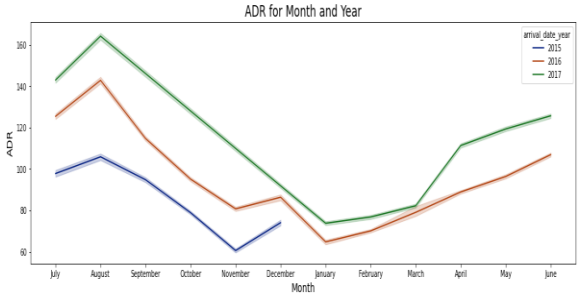

Fig 9: ADR per Month and Year

Above graph clearly shown that hotel business is growing every year.

5.1.10 Price paid for stay vs Month

I have added one extra column in dataset to calculate the price paid for each stay, I have used only non-cancelled booking data. We used the line plot to check the variation in price paid for each stay per month for both the types of hotels.

In the graph I have used two types of colors to show the hotels.

Blue color indicates Resort Hotel

Orange color indicates City Hotel.

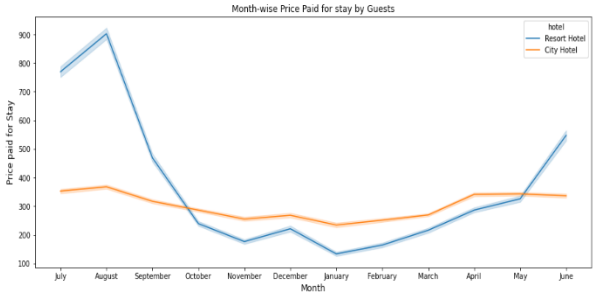

Fig 10: Price paid for stay each Month

As per above graph it is clearly seen that price paid by city hotel for month of July, August and September was less than the Resort hotels.

But for rest of the month price of City hotels were consistently higher than the Resort hotels.

5.1.11 Correlation Map

Seaborn Heatmap: is defined as a graphical representation of data using colors to visualize the value of the matrix. In this, to represent more common values or higher activities brighter colors basically reddish colors are used and to represent less common or activity values, darker colors are preferred.

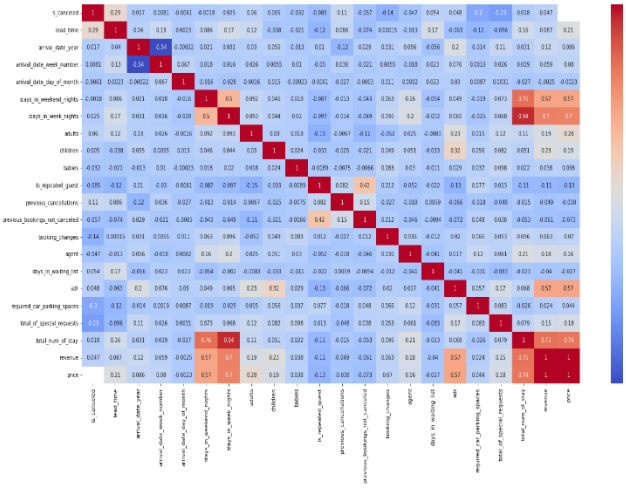

Fig 11: Correlation Map

In the above correlation map we can clearly observe that, stay in week nights was more than the weekend nights.
Also focus on revenue, stay in week nights, total number of stay we can say that revenue was almost same for both the parameters.

6. Conclusion:

That's it! We reached the end of our exercise. Starting with loading the data so far, we have done EDA.

1. City hotel has 66% preference and Resort hotel have 34% of preference. City hotel have more preferred by customers rather than the resort hotel.

2. In overall data and it is found that over 37% of booking were cancelled by the customers. Which is slightly high.

3. Number of cancellations of **City Hotel** seems very higher than the **Resort Hotel**.

4. Number of arrivals seems to be high in year 2016 while the bookings seem to be less in 2015 and 2017.
   But number of arrivals in year 2017 has been drastically down.

5. Number of arrivals seems to be high in year 2016 for **City hotel** compared to **Resort Hotel**, while the bookings seem to be less in 2015 and 2017 for both resort hotel and city hotel.

6. Number of arrivals seems to be high in month of May, June, July and

August for **City hotel**. But number of arrivals for **Resort Hotel** is seems high only for July and August.

7. PRT have more preference followed by GBR.
8. ADR for city hotel is slightly more than the **Resort hotel**
9. As per line graph of ADR per month for three years, clearly seen that hotel business is growing every year.
10. As per line graph Price paid for stay for each month it is clearly seen that price paid by **City hotel** for month of July, August and September was less than the **Resort hotels**.
11. In the correlation map we can clearly observe that, stay in week nights was more than the weekend nights.

References-

1. GeeksforGeeks