HOTEL BOOKING ANALYSIS

Kapil Narayan singh, Sudhanshu Chouhan,
Nimisha Nooti
Data science trainees,
Alma Better, Bangalore

Abstract:

This data article describes a dataset with hotel demand data. One of the hotels is a resort hotel and the other is a city hotel. The data set share the structure, with 31 variables describing the 119390 observations. Each observation represents a hotel booking. The data set comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were cancelled. Since this is hotel real data, all data elements pertaining hotel or customer identification were deleted. Due to the scarcity of real business data for scientific and educational purposes, these dataset can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

A dataset contains information that can be used to train a machine to learn and predict future outcomes based on the patterns found in the dataset. This is done through algorithms or models. Most of data sets use past events to predict future ones. We will perform exploratory data analysis with python to explore and analyze the data to discover important factors that govern the bookings.

Keywords: Exploratory data analysis, hotel booking, special requests, best time for booking, algorithms or models.

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. All personally identifying information has been removed from the data.

1.1. Road map to explore data:

- **Basic cleaning**: Separating the required data from the dataset to use it for analyzing.
- **Understanding and analyzing the factors affecting the booking**: Factors like seasons, festivals and holidays affect the data.
- **Data processing**: We’ll go through each and every feature and encoded the categorical features. Changed the columns according to requirement of analysis.
- **Visualizing the test assumptions** : We'll check if our data meets the

Assumptions required by most multivariate techniques and represent them in the understandable-form of visualization using bar, pie, line, box etc plots.

1.2. Describing Variables:

1. **Hotel**: Two categories of hotels are Resort Hotel or City Hotel.
2. **is_canceled**: Value indicating if the booking was canceled (1) or not (0)
3. **lead_time**: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
4. **arrival_date_year**: Year of booking arrival date.
5. **arrival_date_month**: Month of booking arrival date.
6. **arrival_date_week_number**: Week number of the booking arrival date.
7. **arrival_date_day_of_month**: Day of booking arrival date.
8. **stays_in_weekend_nights**: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
9. **stays_in_week_nights**: Number of week nights (Monday or Friday) the guest stayed or booked to stay at the hotel.
10. **adults**: Number of adults reserved the hotel stay.
11. **children**: Number of children.
12. **babies**: Number Of babies.
13. **meal**: kind of meal opted for.
14. **country**: Country code.
15. **market_segment**: Market segment designation.
16. **distribution_channel**: Booking distribution channel.
17. **is_repeated_guest**: Is a repeated guest is binary info (1) yes or (0) no.
18. **previous_cancellations**: Number of previous cancellation not cancelled by current booking.
19. **previous_bookings_not_canceled**: Number of previous booking not cancelled by current booking.
20. **reserved_room_type**: Code of room type reserved.
21. **assigned_room_type**: Code for the type of room assigned.
22. **deposit_type**: No deposit, Non Refund, Refundable
23. **booking_changes**: Number of changes made to the booking from the moment PMS until the moment of check in/cancellation.
24. **Agent**: ID of the travel agency that made the booking.
25. **Company**: ID of the company/entity that made the booking.
26. **days_in_waiting_list**: Number of days the booking was in the waiting list before it was confirmed to the customer.
27. **customer_type**: Type of customer. Contact, group, transient and transient party.
28. **adr**: Average daily rate as defined by dividing the sum of all lodging transaction by the total number of staying nights.
29. **required_car_parking_spaces**: Is parking required.
30. **total_of_special_requests**: Number of additional special requirement.
31. **reservation status**: status of reservation
32. **reservation_status_date**: Date of the specific status.

2. Data Description:

From the variables we will evaluate the dependency of hotel booking as follows:

- Hotel wise analysis
- Booking cancellation analysis
- Bookings based on Meals, Countries, Room Type
- Distribution channel wise analysis
- Lead time analysis
- Revenue analysis
- Bookings over time analysis

3. Data Cleaning and Manipulation:

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. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

➢ **Finding the null values in column:**

To Identify the null values in the columns of data set we are using 'isna ()' method. We can see that the columns – [company, agent, country and children] have null values in decreasing order respectively rest of the data don't have any missing value.

➢ **Removing the duplicate values in the data set:**

Now we will find out duplicate values in the dataset using 'duplicate ()' method and remove the duplicate values from the dataset using 'drop ()' method. There are 31994 duplicate values found in the data set.

➢ **Removing the null values in the columns of data set:**

We can notice that maximum columns has non-null values by using 'info ()' method. There are 4 columns ('country', 'children', 'agent, 'company') with null values. Hence, sort the non-null values in rows of that columns. Whenever we are working with above 4 columns of data set we will use 'no_null_df' to eliminate the null values.

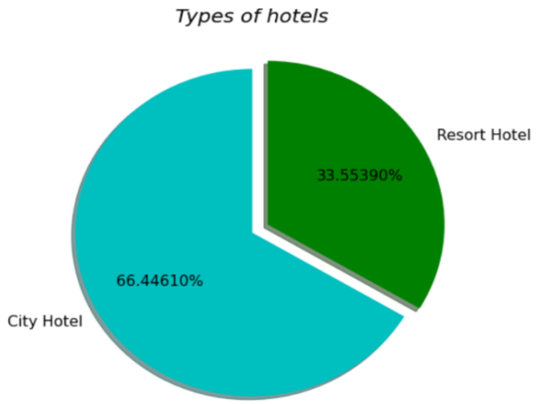
4. Data Visualization:

After data cleaning and manipulation we will visualize our data set through various visualization techniques and discuss about various conclusion we will get from mentioned techniques.

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations.

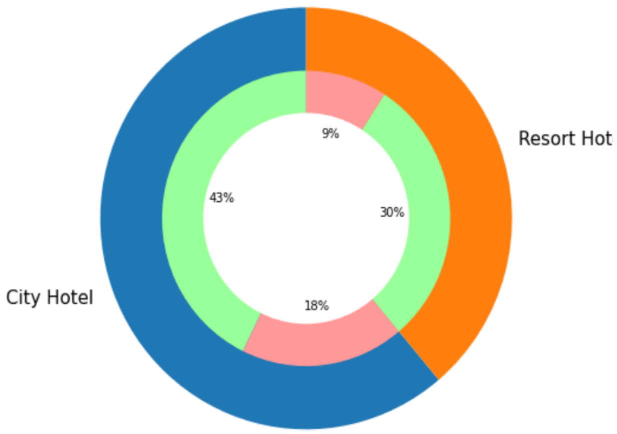
1. Which hotels are guests booking?

The database is divided into two types of hotels: "City" hotels and "Resort" hotels.



Here City Hotel booking is 53428 out of total booking and Resort booking is 33968. There are 66.4% of City Hotels and 33.5% of Resort Hotels were booked. Therefore City Hotels are more preferred by guests compared to Resort hotels. Almost one third of the bookings are for city hotels.

2. Cancelled Booking:



We can observe from the above pie visualization that the max bookings and cancellations are happening in city hotel.

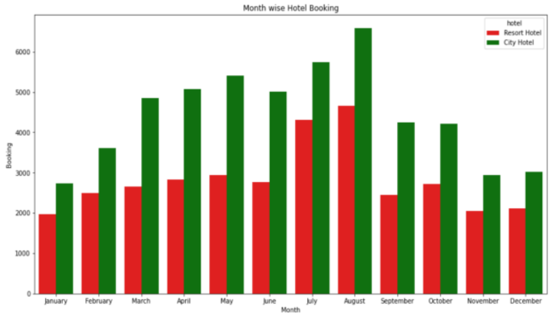
Total bookings that are cancelled = (18%+9%) = 27%

(66.6% of cancellation happening in city hotel)

Total bookings that are not cancelled = (43%+30%) = 73%

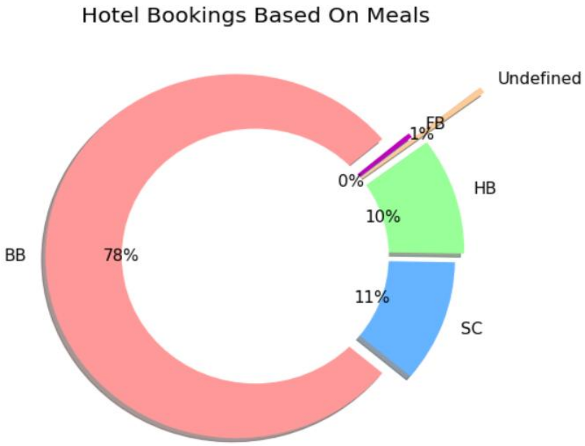
- green = bookings not cancelled
- pink = bookings cancelled

3. Month Wise Booking:



Most of the city and resort bookings are happening in the month of **August** followed by July. Least bookings are happening in the month of January, November and December.

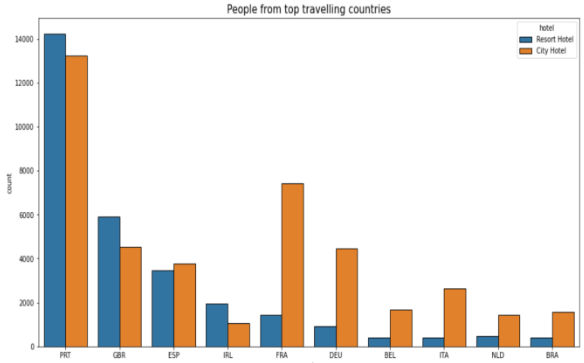
4. Hotel Bookings based on Meals:



- RO: Room only
- BB: Bed & Breakfast
- HB: Half Board (Breakfast and Dinner normally)
- FB: Full Board (Breakfast, Lunch and Dinner)
- AI: All Inclusive (all services of full board plus any others specified in each case)

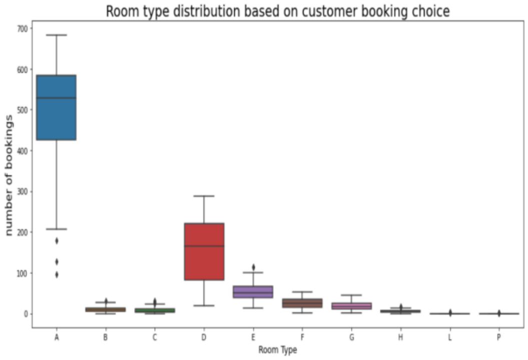
From the above pie visualization we can conclude that 78% of Hotel Bookings are happening on 'BB' meal type i.e., 'BB: Bed & Breakfast‘.

5. Booking analysis based on countries:



From the above bar chart visualization we can notice that most of the hotel bookings are happening in "PTR (Portugal)" country. We can also observe that the maximum people are preferring city hotels compared to Resort Hotels.

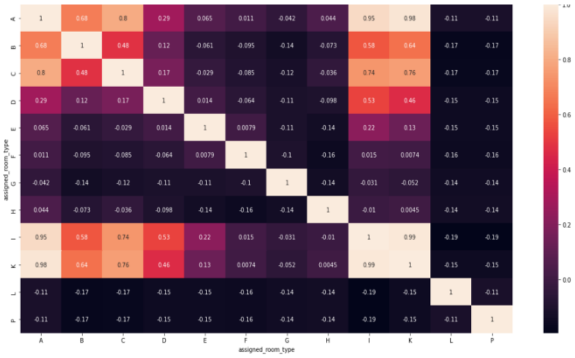
6. Demand of Room Types with respect to weeks of years(2015-17):



Most demanded room types are A next comes D and least bookings are done for room type P and L. we have also analyzed the lead time with weeks in years(2015-2017) and the conclusion is Maximum bookings were happened in 9th and 12th week of every year.

7. The possibility of getting the reserved room type:

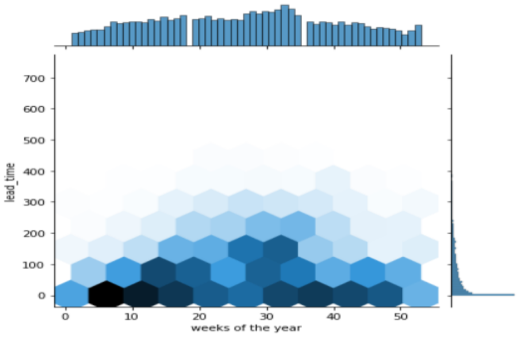
Probability of room allocation for the custo mer choice reserved type A is in the order - A, K, I, C, B.



The lighter color indicates the more probability of getting the reserved type of room and the darker color indicates the less/no probability of getting the room of customer choice.

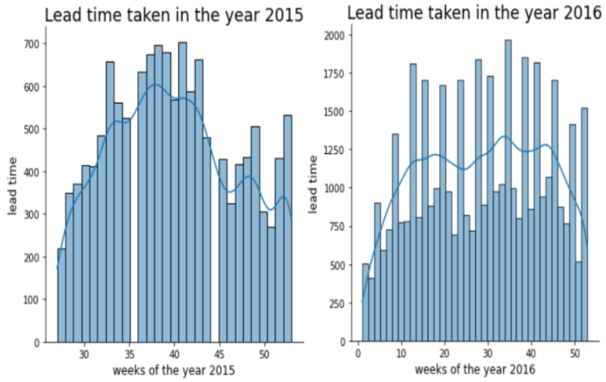
8. Analyzing Lead Time of Bookings:

The bar plots represents max bookings done in the weeks in X-axis i.e., (max bookings done b/w week 30 to 35) and max lead-time taken for booking in Y-axis i.e., (max lead time taken is 0-immediate booking).



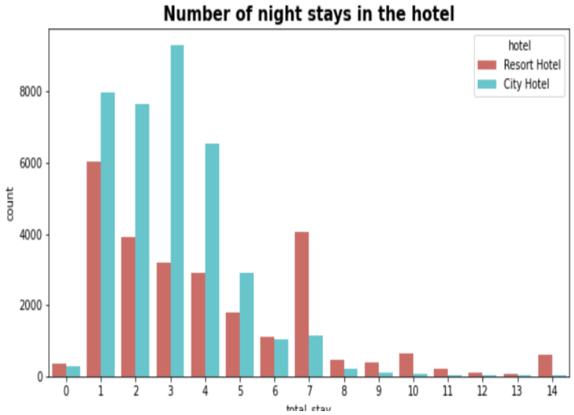
Hex plot represents max lead time taken in the b/w 4th - 18th weeks of the year.

9. Analysis of lead time year by year:

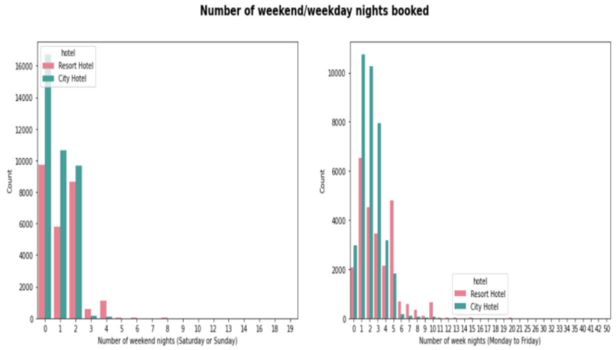


From the above displot analysis we can conclude that maximum lead time taken in bookings is in the year 2016. ‘UEFA Euro 2016 Final’ held in France in which Portugal won the match - may be one of the reason of prior booking of hotels happened with high Lead Time.

10. Number of days people stay in the hotel:

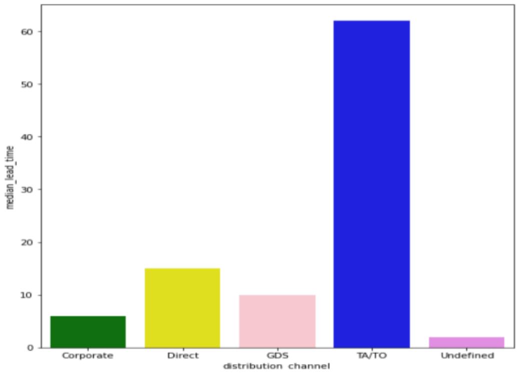


- We can notice that majority of people stay or do a booking of ‘7’ or less than ‘7’ days.
- Maximum night bookings are happening in the city hotels and the max length of stay is ‘3’ days.
- Maximum night bookings length of stay happening in the Resort Hotel is for one night stay.
- We can also observe that if the stay is longer than 7 days then guests prefer to book Resort Hotels only.



We can see that majority of people stay or do a booking of 5 or less than 5 days. Now, we can say the optimal length of stay to get best daily rate is '5' for week nights and '2' for weekend nights. Max night bookings are happening in the city hotels in weekdays and the max length of stay is 1 to 2 days.

11. Analyzing on the basis of distribution channel:



Distribution channel v/s median lead time:

Distribution channel is the costumer accessed by corporate booking/Direct/Travel agent (TA). Travel operator (TO) and Median lead time is the median of number of days that elapsed between the entering date of the booking into the PMS and arriving date.

Through TA/TO distribution channels, bookings were with high lead time i.e., they are booking early compare to other distribution channels.

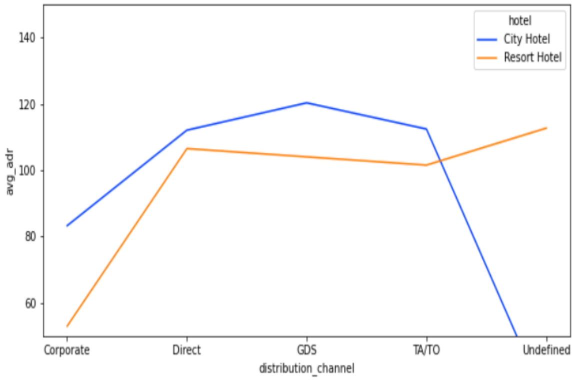
12. ADR (Average daily rate) generated through various distribution channels:

Average daily rate (ADR), one of the three key hotel performance indicators (along with occupancy and Revenue per available room (RevPAR)), is the measure of the average paid for rooms sold in a given time period. The metric covers only revenue-generating guestrooms.

How to calculate ADR:

ADR is calculated by dividing room revenue by rooms sold. The metric is of course applicable for any currency.

ADR = Room Revenue/Rooms Sold



As ADR is the revenue determining factor 'GDS distribution channel' of city hotel bookings are achieving high adr (revenue).

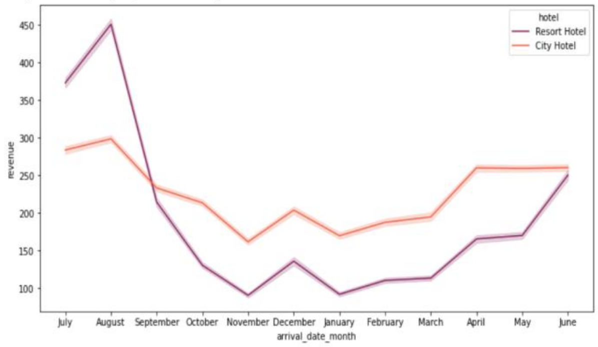
13. Average Revenue of the Hotel:

ADR = Room Revenue/Rooms Sold

Room Revenue = ADR * Rooms Sold

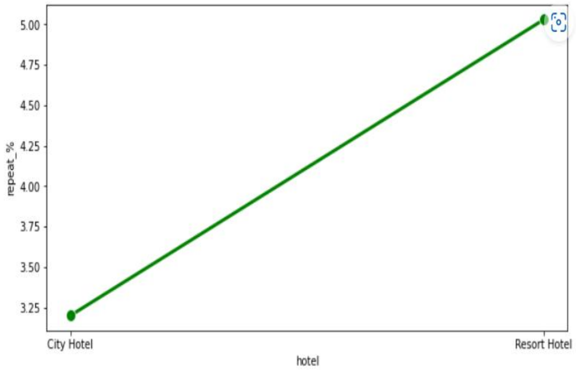
Avg Room Revenue = mean ADR * mean Rooms Sold

Rooms sold are calculated based on no. of booking i.e., no. of adult bookings+ no. of children bookings



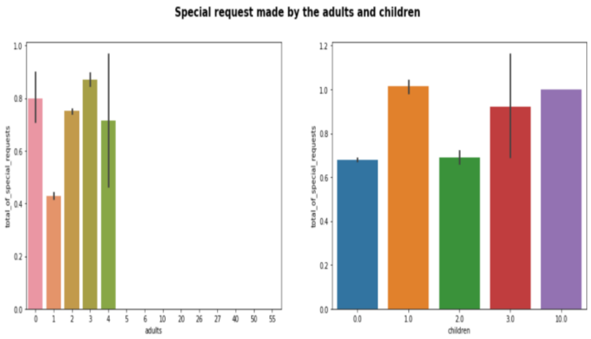
From the above analysis we can notice that the Resort hotels are getting highest revenue in the month of 'august', 'July' and then decreases drastically. City hotel's revenue is almost constant all over year.

14. Chances of customer will return:



There is a very less probability that the customer will repeat. But the return percentage of resort is slightly greater than that of city hotel.

15.Special requests of bookings:



We can see that if the adults are more than 2, there are high chances that the hotel receives more special requests and the no. of special requests for children has no much variation.

5. Conclusion:

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

- Around 61% bookings are of City hotel and 39% bookings are of Resort hotel, therefore City hotels are busier than the Resort Hotels.
- Around 27% of total bookings are cancelled, in that 66.6% cancellations are happening in City hotels.
- In both resort and city hotels most of the bookings are happening in "PTR (Portugal)" country.
- Most of the city and resort bookings are happening in the month of August. Followed by July.
- The Resort hotels are getting highest revenue in the month of 'august', 'July' and then decreasing drastically. City hotel's revenue is almost constant all over year.
- 78% of Hotel Bookings are happening on 'BB' meal type i.e., 'BB: Bed & Breakfast'.
- Most demanded room type is A.
- High probability of lead time taken is '0' i.e., immediate booking are happening but high lead time taken is in the b/w 4th - 18th weeks of the year. Hence, this time is the busiest time of the year.
- Majority of people stay or do a booking of 5 or less than 5 days. Now, we can say the optimal length of stay to get best daily rate is '5' for

week nights and '2' for weekend nights.

- Max night bookings are happening in the city hotels in weekdays and the max length of stay is 1 to 2 days.
- Through TA/TO distribution channels, bookings happened with high lead time i.e., they are booking early compare to other distribution channels.
- As (Average daily rate) ADR is the revenue determining factor 'GDS distribution channel' of city hotel bookings are achieving high adr (revenue).

References-

1. Alma better - www.almabetter.com
2. Tableau - https://www.tableau.com
3. Kaggle - https://www.kaggle.com
4. Geeks for Geeks - www.geeksforgeeks.org