Exploratory Data Analysis (EDA) on Hotel Bookings

Arun Prasath R,
Shabnam Bano,
Roshan Patil,
Rupali Auti

Data science trainees,
AlmaBetter, Bangalore

Abstract:

The hotel industry contributes $3.4 trillion to the worldwide economy. The global hotel industry makes up a key portion of the travel and tourism industry. People rely on hotels for places to stay while vacationing, traveling for work, or other purposes. A dataset of such hotel bookings was provided with various variables.

This project will help to get the insights of performance of hotels and their services, further helps to make decisions for hotel management by data analysis to increase their revenue and customer satisfaction.

Keywords: data analytics, pricing, variables

Introduction:

Data was provided with more than 1 Lakh booking details of the hotel types. It 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 is from the data.

The main objective is to get complete insights of the dataset, which could help hotels to get efficient business. This would in turn help customers to get the right hotels with the right cost and service they required.

Problem Statement:

According to all the information that is personally identified from the above data, following questions were created and tried data analysis on the dataset to get the answers.

1. Number of booking in each hotel type
2. Length of stay in each hotel type
3. Number of Booking month-wise
4. Average daily rates (adr) for both hotels each year
5. Average daily rates(adr) for both hotels in every month
6. Stays in weekend and weekdays in hotels versus bookings
7. Count of adults, children, babies in booking
8. Preference of the meal by customer
9. Top 10 country of origin of customer
10. Market_segment and bookings
11. Number of Weekdays booked versus market segment
12. Number of Weekend nights booked versus market segment
13. Total special requests in each type of market segment
14. Preference of Room types by customer
15. Rooms assigned to customer vs Rooms Reserved by the customer
16. Total previously canceled and not canceled bookings in each hotel type
17. Waiting time versus cancellation
18. Total cancellations for each hotel type
19. Required Parking spaces versus hotel type
20. Total parking space required according to customer type

Hotel:

- Resort hotel
- City hotel

Two types of hotels are provided here. City hotels, which are assumed to be the hotels that are placed inside the city. On the other hand, Resort hotels meant for vacation staying and also used for some professional staying.

Average daily rates (adr):

The Average Daily Rate (ADR) is the price to be paid by a customer for staying per day/night in the room. Which includes the deposit type of the customer paid.

deposit_type:

- The purpose of the advance deposit is to guarantee a reservation
- There are 3 types of deposit:
  1. 'No Deposit' - No deposit needs by the hotel on booking
  2. 'Refundable' - Deposit that can be refunded while vacating the room, which is included in total Room-stay cost.

3. 'Non-Refundable' - Amount paid that can't be refunded once paid Which included in total Room-stay cost.

Customer's activity after booking and before checking in:

Data provides information about whether the customer is new or old. Previous cancellations history, where Number of bookings previously canceled by the customer, Before current booking. Also provided with the Number of bookings previously not canceled by the customer, Before current booking.

How Business works and Market segment of the hotel Bookings:

1. market_segment

Market segment distinction Provides source of information through which customer booked

- Term "TA" - "Travel Agent"
- Term "TO" - "Tour operators"
- Both "TA" and "TO" are considered the same kind of market segment.

2. distribution_channel
   - It is also called "marketing channel"
   - It is the Network through which customer booked

3. customer_type
   - 'Transient' - Simply individual guests requiring a short stay at the hotel
   - 'Contract' - Agreement between hotel authority and customer to require volume room bookings on contract basis.
   - 'Transient-Party' - Booking is Transient and associated with other transient booking
   - 'Group' - Multiple rooms are booked under single customer responsibility

Categories of the customer booking:

The types of persons staying in the room are as follows.

- Number of adults stayed or booked to stay
- Number of children stayed or booked to stay
- Number of babies stayed or booked to stay

Meal preferred: Type of meals Booked.

- BB: Bed & Breakfast
- HB: Half Board (Breakfast and Dinner normally)
- FB: Full Board (Breakfast, Lunch or Dinner)
- Undefined/SC: Rooms only packages without meals.

Room details:

reserved_room_type

- Type of room reserved stored in alphabet codes.

assigned_room_type

- Type of room reserved stored in alphabet codes.

Demand for hotels:

Information regarding the timeline such as year, week, month used to get the insights of the usage of hotels in particular periods which helps to find demand for hotels.

arrival_date_year

- Year of arrival of the Customer.

arrival_date_week_number

- week number of arrival of the Customer.

arrival_date_day_of_month

- Month of arrival of the Customer.

Cancellations of booking:

Information available to get the reason behind the cancellations are,

1. Canceled or not
2. lead_time
3. days_in_waiting_list

If the value in the "canceled or not" column is 1 that means the booking is canceled and if it is 0 that means the booking is not canceled. The Booking Lead Time is the number of days between the time a guest books their room and the time they are scheduled to arrive at the hotel. Number of days the booking was in the waiting list before it was

confirmed to the customer is sorted in days_in_waiting_list.

By analysis it is found that, Increase in the waiting and lead time that leads to increase in number of cancellations

Steps involved:

1. Understand the data.
2. Univariable study.
3. Multivariate study.
4. Basic cleaning.
5. Test assumptions

Understand the data

Data understanding focuses on the comprehension of the information available in the project. In this step we basically check on the kind of variables provided with the dataset, dtype of the columns, shape of the data frame.

Null values Treatment

Our dataset contains numbers of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

Pandas **isnull()** and **notnull()** methods are used to check and manage NULL values in a data frame
The percentage of null values in each variable is found using the following formula.

$$\text{Percentage} = \frac{\text{Number of null values}}{\text{Total number of values}}$$

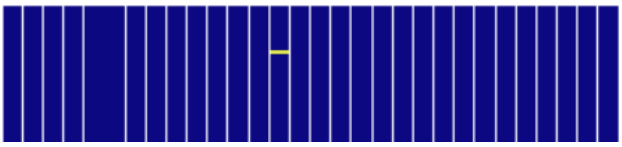
Encoding and dropping of categorical columns

We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

As well as columns containing the large number of the null values shall be removed to increase the accuracy of the analysis.

The **drop() method** removes the specified row or column. By specifying the column/Row axis (axis='columns'), the drop() method removes the specified column/Row.

Heatmap further used to find the null values in any columns.

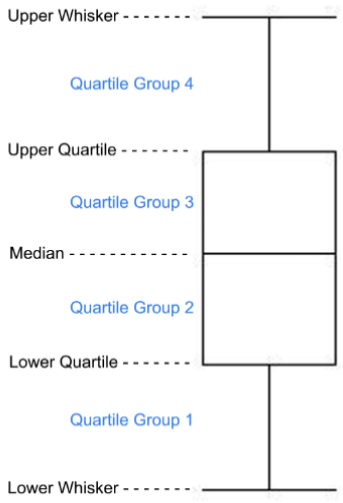


Above analysis helps to find that there are a minimum number of null values in variables "children" and "country".
Hence, filling those null values with appropriate values i.e., filling the null values in the children column as "0"
Filling the null values in the country with the country name which has maximum count in the data.

Standardization of features

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it. The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

Here, outliers are used to get the desired output. This helps to clean the data further by observing the upper and lower limit.

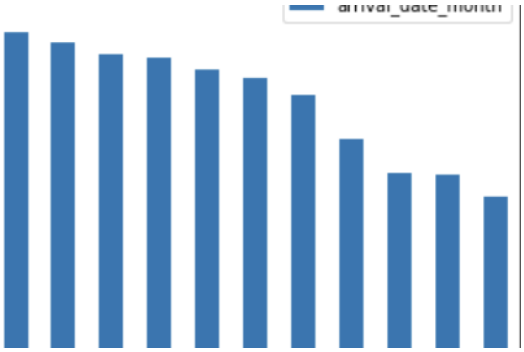


Also, it used to find the outliers in the number of nights customers used to stay in the hotels. If we ignore the outliers in this scenario, the maximum length of stay is more in resort type as resort is mostly used for vacation purposes.
Notably, Median of both the hotels are approximately equal.

Univariable study

Univariate analysis is a basic kind of analysis technique for statistical data. Here the data contains just one variable and does not have to deal with the relationship of a cause and effect.

Information such as number of bookings, Hotel types, monthly booking, stays in week-nights and stays in weekend-nights, Top 10 country of origin of customer, Preference of the meal by customer are used by this method.
For most of the variables a simple bar graph is used to visualize the data in it.

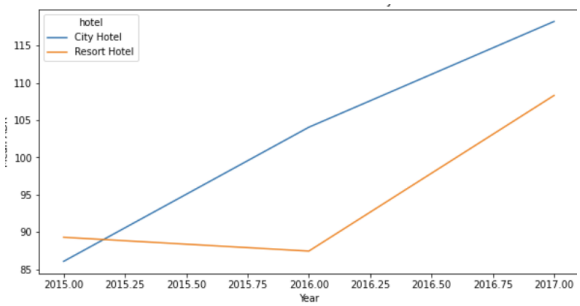


Multivariate study

Multivariate analysis can help determine to what extent it becomes easier to know and predict a value for one variable (possibly a dependent variable) if we know the value of the other variable (possibly the independent variable)

As the method indicates, average daily rates (adr) for both hotels in every month, Stays in weekend and weekdays in hotels vs Bookings, Count of adults, children, babies in booking, Market segment and bookings, Rooms assigned to customer vs Rooms Reserved by the customer.

ADR for each hotel type according to year using Line Plot used in analysis.



Observation helps to find that, City hotels are always higher (adr) than resort hotels. If the trend continues like that, resort hotels (adr) showing maximum increased inclination which means in a few years resort hotel adr will cross the city hotel.

Conclusion:

Some conclusions drawn from the analysis are as follows.

- Customers preferred City Hotel more than Resort Hotel.
- The maximum length of stay is higher (than city hotel) in resort type as the resort is mostly used for vacation purposes.
- Median value of staying days of both the hotels are approximately equal.
- Bookings in the month of August are highest and January found lowest number of bookings.

- In the month of July and till the last week of August Resort hotels received more "adr" than City hotel
- City hotel although dominating in 'adr' in remaining months of the year.
- Maximum Booking done by customer for "2" weeknights stay and "0" weekend nights
- Bookings are mostly made for 2 adults with 1 children in combination
- Customers of any type prefer "BB" Bed and breakfast type meals.
- It should be noted that a greater number of bookings are done by customers from country PRT(Portugal).
- Travel agencies (TA) or Tour operators (TO) play a vital role in hotel booking.
- Except "Direct bookings", all market-segments have a greater number of bookings in city hotel types.
- Room type of "A" preferred mostly by customers
- Also, it should be note the maximum numbers of booking done in the rooms type of “A”,”D”,”E” than others.
- From crosstab analysis, relationship of reserved and assigned rooms found.
- Customer booking in the room type “G” and “H” getting 97.0% of the reserved room.
- Lowest possibility of getting the same room type as reserved in room type "L".
- As we know, the maximum number of bookings done for room type "A" ensured only 85.0% of the same room as reserved by the customer.
- Offline TA/TO and Group market segment has some deviation over customer’s stays in week-nights between Resort and City hotels.
- Undefined and Aviation market segment customers had not shown much interest in the Resort Hotel.
- Direct market segment customers prefer to stay more weekend nights in the Resort Hotel type.
- Online TA customers equally prefer between Resort and City hotels.
- Increase in days in waiting list increasing the cancellation of the booking
- Most customers are not demanding parking space.
- The parking space required by customers is high in resort hotels.
- Percentage cancellation in city hotels is 41.09%.
- Percentage cancellation in resort hotels is 27.66%.
- When the customer type is Transient that means the stay is for a few days so it is possible that customer is bringing his/her own vehicle that's why the parking space required is high for Transient.
- In contract and group booking customers will probably take a hired vehicle form a hotel or from somewhere else that's why they do not need parking space.

References-

1. https://pandas.pydata.org/
2. https://stackoverflow.com/Analytics
3. https://www.wikipedia.org/