Hotel Booking Analysis

Vidit Ghelani,
Mahesh Lakum
Data Science Trainees,
AlmaBetter, Bangalore

Abstract:

When it comes to hotel business, there are a number of parameters that affect the business. A lot can be determined from the bookings and arrivals for a specific hotel or a group of hotels in some region.

Our experiment may help understand the performance of a hotel based on the demographics of people arriving, hotel facility and services, etc.

Keywords: EDA (Exploratory Data Analysis)

1. Problem Statement

This is an EDA problem, hence we had to generate a problem statement by ourselves. Here, after viewing the data, we came up with about twenty questions and tried to answer them. The questions are selected carefully because they act as a viewpoint. Viewpoint indicated the different way in which we look at the data. The entire output of the analysis depends on this. These are called KPIs. (Key Performance Indicators)

2. Introduction

The hotels keep a record of all the booking information. Based on their basic observations like customer traffic, seasons, events, customer demands, they alter the price of their services and hence make profit. These decisions could be made in a better way with nice and timely analysis of the data and then applying accurate predictions to the same.

3. Understanding the features

The raw data contains various features mentioned in columns. Those features are as below:

1. hotel : contains the type of Hotel in the data:
   Resort hotel and City hotel
2. is_canceled: contains the Cancellation Status of Booking
   1 mean Canceled and 0 means Not Canceled
3. lead_time: This is the difference between the booking date and arrival date.
4. arrival_date_year : This is the year in which the visitor arrived
   2015, 2016, 2017
5. Arrival_date_month: This is the month in which the visitor arrived
   January to December
6. arrival_date_week_number: This is the week number of year in which the visitor arrived
   1 to 53
7. arrival_date_day_of_month: This is the day number of month when the visitor arrived
   1 to 31
8. stays_in_weekend_nights: This is the number of weekend nights, i.e. Saturday and Sunday
9. stays_in_week_nights: This is the number of week nights, i.e. Monday to Friday
10. adults: This contains the number of adults per booking
11. children: This contains the number of children per booking
12. babies: This contains the number of

babies per booking

13. meal: This mentions the type of meal preferred by the customers.
    Undefined/SC means no meal package,
    BB means Bed & Breakfast,
    HB means Half board (i.e., breakfast & one other meal – usually dinner),
    FB means Full board (i.e., breakfast, lunch & dinner)
14. country: This gives the country of origin of the visitor
15. market_segment: This gives the group of people based on market
    Direct, Corporate, Online TA, Offline TA/TO, Complementary, Groups, Aviation
    Where, TA: Travel Agents, TO: Tour Operators
16. distribution_channel: This mentions the type of distribution channel
    Direct, Corporate, TA/TO, Undefined, GDS
17. is_repeated_guest: This says if a customer is repeated
    1 means repeated customer, 0 means not repeated
18. previous_cancellations: This gives the number of previous bookings that were canceled by the customer prior to the current booking
19. previous_bookings_not_canceled: This gives the number of previous bookings not canceled by the customer prior to the current booking
20. reserved_room_type: This gives the type of room reserved
    'C', 'A', 'D', 'E', 'G', 'F', 'H', 'L', 'P', 'B'
21. assigned_room_type: This gives the type of room whose possession is given at the time of arrival.
    'C', 'A', 'D', 'E', 'G', 'F', 'H', 'L', 'P', 'B'
22. booking_changes: This gives the number of bookings changed
23. deposit_type: This gives the types of deposit
    No Deposit, Non Refund, Refundable
24. agent: Agent Id
25. company: Company Id
26. day_in_waiting_list: Number of days the booking was in the waiting list before confirmation
27. customer_type: Type of customer
    Contract, Group, Transient, Transient-party
28. adr: means average daily rate
29. required_car_parking_spaces: Number of car parking spaces required by the customer
30. total_of_special_requests: Number of special requests made by the customer
31. reservation_status: Status of reservation
    Canceled, Check-Out, No-Show
32. reservation_status_date: Date at which the last status was updated

4. Approaching the Problem

The entire process to analyze the data and to draw out a useful conclusion was done in three steps:

4.1 Pre- Processing

4.1.0. View the data

4.1.0.1. Understand the Features

4.1.1. Inspecting the data

Here first viewed the information of data with the '.info()' command. However, isnull().sum() serves to be more appealing. Hence, using it to spot the null values.

1. Inspecting the data for duplicate values and null values
2. Get the basics statistics for each feature. (i.e. use df.describe())

4.1.2. Cleaning the data

We will create a new dataframe and deal with the null values by replacing it with appropriate values or may drop them. This way we do not make any changes to the raw data.

1. Remove duplicate values
2. Replace null values
3. Remove irrelevant data (i.e. drop canceled bookings data)

4. Dealing with the Outliers
5. Changing the data types of features where necessary.
6. Add derived features (added columns kids and total visitors)
7. Dropping the rows that contain zero total visitors.
8. Creating Functions.

4.1.3. Formulating the Questions

This is also called defining the KPIs or viewpoints to analyze the data.

We have formulated the following questions:

1. What are the types of Hotels in the data?
2. What is the percentage of booking for each hotel?
3. What is the year wise trend of bookings for each hotel?
4. Which agent made the most number of bookings?
5. Enlist the country of origin of the majority of visitors.
6. What is the busiest time for hotels?
7. What is the proportion of weekend and weekday nights? Is there any difference between them?
8. How many bookings were previously canceled?
9. Which market segment does most visitors come from?
10. Which distribution channel does most visitors come from?
11. How many visitors are repeating?
12. Which is the most preferred meal?
13. Which is the most preferred deposit type?
14. How many visitors asked for car parking space?
15. Which month has the highest average daily rate per person?
16. What is the trend of ADR?
17. Which room Type is high in demand?
18. How likely is the hotel to receive a disproportionately high number of special requests?
19. Which hotel type has a longer waiting time for booking?
20. Which hotel type has a higher lead time for booking?

4.2 Performing exploratory data analysis (EDA)

Here, the data was rigorously analyzed to answer the above questions. Based on the analysis we have answered the questions.

4.3 Answer the questions based on analysis and draw out the conclusions

1. The types of hotels are Resort Hotel and City Hotel
2. Percentage of booking for Resort Hotel is 41.1 % and for City Hotel is 58.9%
3. From the graph we can say that:
    1. Year 2016 had the most number of bookings. This trend remains the same for both the hotel types. However, this conclusion is less relevant because we have data of only six months for the year of 2015 and 2017 which is half as compared to the data of the year 2016.
    2. Considering the above view point, we can say that 2017 performed relatively better with just the data of six months, it has achieved 75% booking as compared to the previous year (i.e. 2016).
4. Agent with Id number 9 made the most number of bookings.
5. Majority of the visitors are from the country 'Portugal' with Country Code 'PRT'. The second one on the list is the 'Great Britain', i.e., 'GBR'.
6. From the analysis we can say that:
   For the year 2015 - September and October are busiest.
   For the year 2016 - August is the busiest followed by Jully, September and October
   For the year 2017 - May and July are the busiest.

For the year 2015 - Week number 39, 41 are busiest
For the year 2016 - Week number 27,38 are busiest
For the year 2017- Week number 20,28 are busiest

Analysis for arrivals on days of month, seems to follow a cyclical pattern. However, not much could be concluded from this. So, we divided the plot yearwise. It is observed that generally bookings peaked on the following dates: 2,13,21,28.

7. From this we can say that the number of stays on weekends for ‘2’ person occupancy is more compared to that on weekdays. And on weekdays the number of occupancy ‘1’ peaks. So, as per our hypothesis, this could possibly be due to the contribution by visitors under business tours.
8. 538
9. It is observed from the graph that the majority of the visitors arrived via Online Travel Agents.
10. It is observed that for both the hotels the distribution channel 'TA/TO' (Travel Agents/ Tour Operators)contributed the maximum.
11. From the plot we find that a very small number of visitors are repeating. When we check the hotewise data, they have nearly the same amount of repeat visitors. Overall percentage of repeat visitors is as low as 4.9%.
12. From the data we see that BB is the most preferred meal type for the majority of visitors.
13. Majority of the visitors preferred to give no deposit.
14. About 88.44% of the visitors did not require space for car parking. Of the rest, the second major, i.e. 11.5% visitors needed parking space for a single car.
15. August has the highest average daily rate per person, followed by July and June respectively.
16. From the graphs we can say that the ADR for resort hotel types is quite fluctuating compared to that of city hotels. When checked yearly for months, the ADR forms a bell shaped curve with August at the center. The month of January has the least ADR value.
17. From this we can see that the Room Type A is high in demand.
    The room types with highest adr range are C, G, F, H & K, however effective for profit are G, H, F and E.
    Looking at individual hotel types, the beneficial Room Types for hotel profit are:
    City Hotel - G, F and E
    Resort Hotel - H, G, F and C
18. From the data we see that the maximum number of special requests received from a single booking is ‘5’; and it amounts to only 0.1% of the total bookings received. Considering ‘3’ requests as a threshold, we can say that the chances to have a disproportionately high number of special requests are less than 3.6%.
19. City Hotel takes longer to confirm booking status.
20. City Hotel has slightly higher lead time compared to the resort hotel.

5. Observations:

From this analysis we observed the following:

1. Majority of the booking came for City Hotel, i.e. 58.9%, which City Hotel is more preferred.
2. Booking trend for both the hotels is nearly the same. However, talking about the volume of bookings, it is the same in 2015, but for 2016, City Hotel received more bookings.
3. Agent with Id number 9 made the most number of bookings.
4. Majority of the visitors arrived from the country of Portugal.
5. Occupancy of hotels:
   For the year 2015 - September and October are the busiest
   For the year 2016 - August followed by July, September and October are

busiest
For the year 2017 - May and July are the busiest.

6. Single visitors preferred weekday stays, while visitors traveling in pairs preferred weekend stays more. Possibly, they are couples.
7. Majority of the visitors arrived from online travel agents (TA) market segment. The same applies to distribution channels.
8. Majority of the time booking for visitors traveling in pairs arrived via online travel agents (TA) .
9. Majority of the visitors preferred meal type BB (Bed & Breakfast).
10. Only 4.9% of customers are repeated.
11. Of the arriving customers, a total 538 bookings were previously canceled.
12. Majority of the customers do not prefer to pay a deposit amount.
13. About 88.4% of visitors did not require car parking space. Which means the remaining ones are traveling by car or are renting a car in the state.
14. August has the highest average daily rate per person.
15. ADR for resort hotel types is quite fluctuating compared to that of city hotels. When checked yearly for months, the ADR forms a bell shaped curve with August at the center. The month of January has the least ADR value.
16. Room Type A is high in demand.
17. 46.2% of visitors do not make any special request. 35.5% of visitors have 1 special request.
18. City Hotel takes longer to confirm booking status.
19. City Hotel has slightly higher lead time compared to the resort hotel.

6. Challenges Faced:

The first challenge was to identify the KPIs. As it is very crucial for analysis.
Next was the data wrangling step. Which demanded observation in many different aspects. Looking at the data, plotting it, choosing the plot type etc., points were something that kept on evolving. For certain questions we had to plot data in more than one type or plot and also we had to alter the parameters in the plot to draw out a useful conclusion.

7. Conclusion:

The overall data to analyze was reduced considerably in the data cleaning stage. Hence, the hotel must provide cleaner data. The period for which the data is available must be the same. Presently, it is from July 2015 to August 2017. To increase the profit, hotels must increase the types of rooms in high demand. Hotels must try to reduce the booking time. Very few customers are repeated. There are a number of factors behind this. Hotels must try to add more features to track those factors. Maybe altering them could help to retain the customers.

References:

1. GeekforGeeks
2. Stackoverflow
3. Python guide
4. Matplotlib guide
5. Seaborn guide
6. Github
7. Kaggle