Hotel Booking Data Analysis

By- Neeraj Sachan

Data Science Enthusiast at AlmaBetter

Cohort- Seattle

Abstract

This project contains the real world data record of hotel bookings of a city and a resort hotel containing details like bookings, cancellations, guest details etc. from 2015 to 2017. Main aim of the project is to understand and visualize dataset from hotel and customer point of view i.e.

- ✓ reasons for booking cancellations across various parameters
- ✓ best time to book hotel
- ✓ peak season etc.

and give suggestions to reduce these cancellations and increase revenue of hotels.

This project is part of my Data Analysis with Python.

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 helps in exploring those questions!.

2. Introduction

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.

3. Data Summary

The data contains following variables:

- hotel: Name of hotel (City or Resort)

- is_canceled: Whether the booking is canceled or not (0 for no canceled and 1 for canceled)
-lead_time: time (in days) between booking transaction and actual arrival.

- arrival_date_year: Year of arrival
- arrival_date_month: month of arrival
- arrival_date_week_number: week number of arrival date.
- arrival_date_day_of_month: Day of month of arrival date
- stays_in_weekend_nights: No. of weekend nights spent in a hotel
- stays_in_week_nights: No. of weeknights spent in a hotel
- adults: No. of adults in single booking record.
- children: No. of children in single booking record.
- babies: No. of babies in single booking record.
- meal: Type of meal chosen
- country: Country of origin of customers (as mentioned by them)
- market_segment: What segment via booking was made and for what purpose.
- distribution_channel: Via which medium booking was made.
- is_repeated_guest: Whether the customer has made any booking before(0 for No and 1 for Yes)
- previous_cancellations: No. of previous canceled bookings.
- previous_bookings_not_canceled: No. of previous non-canceled bookings.
- reserved_room_type: Room type reserved by a customer.
- assigned_room_type: Room type assigned to the customer.
- booking_changes: No. of booking changes done by customers
- deposit_type: Type of deposit at the time of making a booking (No deposit/ Refundable/ No refund)
- agent: Id of agent for booking
- company: Id of the company making a booking
- days_in_waiting_list: No. of days on waiting list.
- customer_type: Type of customer (Transient, Group, etc.)
- adr: Average Daily rate.
- required_car_parking_spaces: No. of car parking asked in booking
- total_of_special_requests: total no. of special request.
- reservation_status: Whether a customer has checked out or canceled, or not showed
- reservation_status_date: Date of making reservation status.

4. Steps Involved:

4.1. Creating Questions:

We created following questions for our analysis:

Q1. Which type of customers do more bookings?

Q2. Let’s have an overview which type of hotel generally people prefer to book?

Q3. Let’s have an overview which type of deposit is more preferred by customers?

Q4. What kind of food is mostly preferred by the guest?

Q5. Let’s have an overview how much guests pay for a room per night?

Q6. What is the most preferred room type?

Q7. Looking into which countries mostly visitors are coming from?

Q8. What is the percentage of repeating guests?

Q9. Which hotel has high chance that its customer will return for another stay?

Q10.Which hotels generating more ADR?

Q11.What is the relationship between total number of guests and ADR?

Q12.What is the relationship between total stay and ADR?

Q13.Which year had the highest Booking?

Q14.In which month most of the bookings happened?

Q15.How does the price vary over the year?

Q16.What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests?

Q17.How long people stays in hotel?

Q18.Is customer canceled their bookings if they are not allotted with the same room type which was reserved by them?

Q19.Which distribution channel contributed more to generate high ADR?

Q20.Which distribution channel has the highest Cancellation rate?

4.2 Importing all important Libraries

For our EDA process firstly, I imported all the important libraries like Pandas, NumPy, Matplotlib, Seaborn etc.

4.3 Observing Data and Cleaning Data (Data Preparation)

This hotel dataset contains 32 features and 119390 observations. Each observation represents the complete detail about the booking. Only For columns (company, agent, country and children) in our dataset contain null values. But the "company" and "agent" columns contain very large number of null values i.e. 112593 and 16340 respectively. So, we dropped these columns. Country columns contains 488 null values. We replaced these null values with ABC. Only four children columns contain null values. We replaced these null values with zero. After that our dataset is free from null values. After that for reducing columns, we merged the adults, children and babies column into a single column namely total guest. Now, the data is fully prepared for EDA.

4.4. Exploratory Data Analysis

After above steps, I have done the Exploratory Data Analysis of our data for answering all the questions, which we made earlier in the 1st Step. Mostly *pandas* library is used for performing operations. For visualizing are data, I used the following graphs and plots using Seaborn and Matplotlib Libraries:

- Bar Plot.
- Scatter Plot.
- Pie Chart.
- Line Plot.
- Heatmap.
- Box Plot

By plotting different graphs and plots, we can visualize the different aspects how they are performing. We can also see different correlation between different variable how they are affecting each other and finally affecting the business.

5. Observations (What do we see here?) and Conclusions:

❖ Observations

1) Transient customers are making maximum number of bookings, followed by Transient Party Customers.
2) People generally prefer to do their bookings in City Hotels as compared to the Resort Hotels. The Resort Hotels are generally more costlier than the City Hotels, that's why people prefer more City Hotels.
3) Maximum of customer prefer No-Deposit method for their booking payments. Very less people done their bookings using No-Refund payment deposit method.
4) Customers preferred BB (Breakfast and Bed) food options mostly.
5) Mostly guests prefer Apartment type hotel, because these are very cheap for the both City and Resort hotels. D type of hotels are also cheap therefore, people also books D type hotels frequently. G Type city hotels are most expensive.
6) The most preferred room is A (room type).
7) Most are the guests are coming from the Portugal, followed by Great Britain and France.
8) 37.08% bookings was cancelled by the guest. It was found that the repeated guest cancelled their bookings very rarely. The percentage of repeated guest is 4.27%. In order to retained the guests management should take feedbacks from guests and try to improve the services.
9) Both hotels have small percentage that customer will repeat, but Resort hotel has slightly higher repeat % than city hotel.
10) Resort Hotel has the lowest ADR. That means Resort hotels are generating less revenues than the city hotels. More the ADR more is the revenue.
11) As the number of guests increases ADR also gets increases.

12) As the total stay increase ADR is decreasing. Thus, for longer stays customers can get good ADR (price).
13) 2016 has the highest bookings. And 2015 had less bookings.
14) July and August had themost number of bookings for both hotels.
15) In winter hotels had minimum bookings, that's why price is lower winter .Therefore, we can say that the winter is best for planning any trip.
16) Guests booked hotels mostly for 0 to 3 nights. Very few guests booked hotels for more than 8 days. ADR is also higher for less nights stay. As the number of nights increases, ADR is also gets increases. Thus, for longer stay guested can good price.
17) Majority of customer do not canceled their booking, when they don't get the desired room. so hotels need to take little care about this. Guests very rarely made special request along with their bookings.
18) 'Direct' and 'TA/TO' has almost equally contributed in ADR in both types of hotels. GDS has highly contributed in ADR in 'City Hotel' type. GDS need to increase Resort Hotel Bookings, for increasing its ADR. Resorts made high ADR by Undefined mode of booking also.
19) In TO/TA, City hotels has the high cancellation rate compared to resort hotels .In 'direct' the hotels has the almost same cancellation rate.

❖ Conclusion:

1. Best months for planning a trip are October to February, because prices for both the hotels are lessor as compare to other months due to less bookings.

2. Guest numbers for the Resort hotel go down slightly from June to September, which is also when the prices are highest. Thus, these months should be avoided for booking.

3. Very large number of customers are cancelling their bookings, so hotels need to make strict cancellation policy, like they can use non-refund options.

4. Cancellations are high when done through agents compared to direct booking. Hotels need to do marketing and give special incentives for direct booking as these may establish personal one to one relationship promoting customer loyalty.

5. The number of repeated guest is very low, in order to retained the guests management should take feedbacks from guests and try to improve the services.