Hotel Booking Data Analysis

By- Aamir Sohail

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. The main aim of the project is to understand and visualize datasets from hotel and customer points 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 the 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. How many booking were cancelled in both type of hotels ?

Q2. From which country most guests are coming?

Q3. Which month is the most occupied and which is the least occupied?

Q4. What are the number of weekend vs weekdays night bookings for resort hotels?

Q5. Which months have cheaper booking rates?

Q6. How many number of customers repeated their bookings?

Q7. What are the number of bookings made by different market segment?

Q8. Which is the most booked accomodation type?

Q9. How does deposit type affects cancelation?

Q10. How does ADR affect cancelation?

Q11. Which type of hotel has longer waiting time?

Q12. How does lead time affect cancellation?

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 Replaced null values of column Agent and Company with 0.

Country columns contains 488 null values. We replaced these null values with mode. Only four children columns contain null values. We replaced these null values with mean. There are many rows that have zero guests including adults, children and babies, we removed those rows because they do not make any sense.

After that our dataset is free from null values. 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:

1. Bar Plot.
2. Count Plot.
3. Pie Chart.
4. Line Plot.
5. Heatmap.

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 and Conclusions:

Observations:

1.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.

2. More visitors are from Portugal, France, Great Britain and Spain being the highest.

3. August is the most occupied month with 11.65% bookings and January is the least occupied month with 4.94% bookings.

4. Number of week nights stay is more as compared to weekend nights stay.

5. Both city and resort hotels, November to January have cheaper average daily rates.

6. If talk about Repeated Guests, the number of repeating guests are very low.

7. The Maximum number of bookings made by Online TA, followed by Offline TA and Direct.

8. Percentage of booking is high in case of 'Couple', which means that maximum number of booking made by couples in both type of hotels say city hotels and resort hotels.

9. 'Non-Refund' policy increases the chances of booking cancellations.

10. he correlation between ADR and cancellations is positive which means as ADR increases number of booking cancellations will also be increases.

11.City hotel has the higher waiting time.

12. As the lead time increases, percentage of booking cancellation also increases.

Conclusion:

1. Target months between May to Aug because those are peak months due to the summer period.

2. Book hotels in month from November to January as they have cheaper average daily rates.

3. Mostly guests are coming from European countries, so target those countries for advertisements.

4. Mostly bookings are done by Online TA market segment and also couples do the bookings mostly.

5. Increase in lead time and ADR impacts more cancellations and also Non-Refund policy made customers to cancel their bookings.

6. Mostly cancellations are done in case of city hotel, so customers prefer resort hotel.

7. Since there are very few repeated guests, focus should be on retaining customers after their first visit by fulfil their more special requests