HOTEL BOOKING ANALYSIS (EDA)

Kaushal Kumar Jha,
Prabhat Pradip Manna,Asif PA,
Shambhu Nath Jha,Priya gupta
Data Science Trainees,
AlmaBetter, Bangalore

Abstract

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 We will be using the data available to analyse the factors affecting the hotel bookings. These insights can help the hotel management to provide better service options to their future guests from understanding the past data.

We have a given dataset and as analyst we have to get the important insights from this data which will help the management in improving their future serviceability by understanding of past data.

1. Introduction

The Hotel Industry like any other business opens up socio-economic opportunities for both owner and customer. It has the function of providing hospitality services to customers.

Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week,adr, and many more. This makes analyzing the patterns available in the past data more important to help the hotels plan better. Using the historical data, hotels can perform various campaigns to boost the business. By analysing the past data hotel managements can boost their business and profit in future.

The hotel customer management can understand the different parameters that significantly affect the hotel booking statistics and take into account what improvements can be made.

We will be using the data available to analyze the factors affecting the hotel bookings. These factors can be used for reporting the trends and predict the future bookings.

2.Problem Statement

We have been provided with the hotel booking dataset,that contains all the important attributes and records of the hotel bookings. As an analyst our job is to explore and analyzethe data and extract the important information that can prove to be crucial for the hotel business owners.

We will analyze only those parameters which will be going to help the stakeholders in making hike in their business and we not remove the useless data that neither provides us with any insights, nor does it fulfill its basic functionality to work as the crucial deciding feature of the dataset.

To extract the most useful information from our data set and make it easy to understand by stakeholders and decision maker we will perform visual analysis on our dataset.

Finally, we have to come up with important conclusions that we extract from the overall analysis. These include:

- Which year having most of the booking of hotels?
- Which was the busiest month?
- Which month has maximum average daily rate(adr)?
- Which type of meal was mostly preferred?
- From which country most of the booking has been made?
- Etc

3. Feature Description

- hotel (Categorical): Type of Hotel (City Hotel / Resort Hotel)
- is_canceled (Numerical): Whether the booking is canceled (1) or not canceled (2)
- lead_time (Numerical): The time taken between when a customer makes a reservation and their actual arrival.
- arrival_date_year (Numerical): Year of arrival
- arrival_date_month (Categorical): Month of arrival
- arrival_date_week_number (Numerical): Week number of arrival
- arrival_date_day_of_the_month (Numerical): Day of month of arrival
- stays_in_weekend_nights (Numerical): Number of stays in weekend nights
- stays_in_week_nights (Numerical): Number of stays in week nights
- adults (Numerical): Number of adult guests
- children (Numerical): Number of children with the guests
- babies (Numerical): Number of babies with the guests
- meal (Categorical): Type of meal booked by guest
- country (Categorical): Country of origin of guests
- market_segment (Categorical): Purpose of and way of booking
- distribution channel (Categorical): Mode of reservation
- is_repeated_guest (Numerical): Whether the guest is repeated (1) or not (0)
- previous_cancellation (Numerical): Whether the guest had canceled previously (1) or not (0)
- previous_bookings_not_canceled (Numerical): Whether the previous booking canceled
- reserved_room_type (Categorical): Type of room reserved by guests
- assigned_room_type (Categorical): Type of room assigned to the guests
- booking_changes (Numerical): Number of changes made to the booking
- deposite_type (Categorical)e: Type of deposit made
- agent (Numerical): ID of agent that booked the hotel
- company (Numerical): ID of company from which the booking was made
- days_in_waiting_list (Numerical): Number of days in waiting list

- customer_type (Categorical): Type of customers
- adr (Numerical): Average Daily Rate (the average revenue earned for an occupied room on a given day.)
- required_car_parking_spaces (Numerical): Number of car parking spaces required
- total_of_special_requests (Numerical): Number of special requests made by guest
- reservation_status (Categorical): Status of reservation (Canceled/Check-Out/No-Show)
- reservation_status_date (Date): Date of latest reservation status

4. Exploratory Data Analysis

We have started with the exploration and analysis of ourHotel Booking dataset. The main environment for the working of our project being used is Python 3 (Google Colaboratory).
We will be using different python modules and libraries such as follows:

- For Data-Analysis: **Numpy** and **Pandas.**
- For Data Visualization: **Seaborn** and **Matplotlib**

Important steps that involve in EDA:

- Data Initialization
- Data Preprocessing
- Data Analysis

Let's discuss these three most important steps of EDA individually.

I.Data Initialization

In this step first of all we have mounted the drive and imported the dataset into the Python notebook environment.

Importing libraries.

In this step we have imported all the pythonlibraries that are essential for our EDA. These include:
NumPy, Pandas, Seaborn and Matplotlib.

Loading the Dataset

In this step, we have used Pandaslibrary to load our dataset that is initially in csv format, into our python notebook environment.

I. Data Preprocessing

We have successfully loaded our data in our python notebook environment so we will understand our data set in this part and than we will proceed with analysis of our dataset. Understanding of data is most important before commencing analysis on this.

Understanding of Dataset

By checking the data, we have come the following specifications:

- Shape of the data frame:
  Our data frame consists of 11930 rows and 32 columns initially.

- Columns present in the data frame:(Refer 3. Feature Description)
- Data Information:
  This gives us the information of range index which basically represents total number of row of data set and the column with their parameters like labels, Non-Null count which indicate total non null value of a respective column, Data type, and Memory used.

  Column Description:

  This gives us the aggregate values in all the numeric columns. That is min,max,mean,count,25%,50% etc.
- Checking for null values:
  This will let us know about the total number of null values present in every columns of dataset.
- Checking for duplicates values:
  This will gives us information about duplicate values present in our data set.

Data Cleaning

1.Null values treatment:
Data cleaning is most important thing from the point of view of analysis because if our dataset contain any column with null value than it will create a lot of problem during visualization of data and it will affect our decision making.

First of all we have checked all columns which contains null values. In our dataset these columns was company,agent,country and children.

In next step we filled all null values of these columns with appropriate values.

2. Dealing with duplicate data:
Having duplicate data in dataset is a problematic things because its presence is directly going to affect our final result.

In the next step we have checked the presence of dulicates value in our data set which given us a shape of (31994, 32).

Before dropping these duplicates from our dataset we have made a copy of our original daraframe for future requirment.

In the next step, we have dropped all duplicate data and got our actual dataset with the shape of (87396,32).

Finally cleaning part of our dataset has been done now we can further proceed with data analysis.

III.Data Analysis

Now we have cleaned dataset so we can further proceed with data analysis.

In this part we have done most of the analysis using data visualization which is easier to understand .

- **Most booked hotel types:**
  We have used countplot to analyze most preferred hotel.
- **Yearly wise Analysis:**
  In this we have analyzed using countplot to get the maximum number of bookings in a year.
- **Monthly wise Analysis:**
  Using countplot we have analyzed the busiest month of the year.
- **Special request analysis:**
  Which month has the most special request made by the guest .
- **Most preferred room type:**
  Which type of rooms are mostly preferred in different year.
  And so on….

Challenges

As our data set was too large so as an analyst we have faced some challenges while working on our Hotel Booking dataset which include:

- Null values: These nullvalues make the statistical analysis harder. As null value is biggest enemy for our dataframe, we have to carefully deal with it.
- Duplicate Value: In our dataset duplicate values was present, and as presence of duplicate values can dramatically affect the final result so we have to be careful with this.
- Other common challenges include correct visualization selection, so that it is easily understood by every stakeholder. Also, coming up with non-obvious insights that can affect the overall business, is a big responsibility of a data analyst.

Conclusion

- In 2015 , Resorts Hotels Were mostly Preferred.
- In 2016 and 2017, City Hotels were preferred mostly.
- Most Customers Booked Hotels in the year 2016 followed by 2017 and then 2015.
- Maximum bookings were made in the month of August.
- In 2015, Most bookings were made in the month of September.
- In 2016, Most bookings were made in the month of August.
- In 2017, Most bookings were made in the month of May.
- A is the most Reserved room Type in all the years.

These conclusions derived from the Exploratory Data Analysis of the Hotel Booking Dataset can help stakeholders better understand the trend of hotel booking, and customer behavior with all the parameters, so as to make better business decisions.