HOTEL BOOKING ANALYSIS (EDA)

Shadab Shaikh,
Isha Jadhav
Data Science Trainees,
AlmaBetter, Bangalore

Abstract

Around the world, a large number of people book a hotel room for whether they are traveling, holidays and other purposes. While it looks like an easy job from a customer's perspective, the hotel business runners have to maintain a large amount of record and data so that they can understand the historic booking patterns from the customers and have to find ways to scale their businesses even higher.

One such dataset is available with us, and as proficient analysts, it is our job to get as many insights from the given data. These insights can help the hotel management to provide better service options to their future guests from understanding the past data.

1. Introduction

The hospitality infrastructure is an important aspect of a thriving tourism industry. It is the body that generates the maximum revenue apart from traveling and other important institutions. One such part of the total hospitality infrastructure is the hotel industry. Technicalities apart, these are the places where the customers spend most of their time in, be it resting or having meals. Hence, the hotels have a big impact on the overall tourism industry as it is one of the most profit generating parts.

Having the past customer records, here, the booking data, is crucial for the hotel businesses to thrive and also scale their business overall. 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.

As data analysts, we can help the stakeholders to easily understand the underlying trends of the historic data, and come up with innovative solutions, so that they can use them to better their revenues and businesses overall.

2.Problem Statement

We have been provided with the hotel booking dataset, that contains all the important attributes and records of the hotel bookings. Our job is to explore and analyze the data and come up with important insights that can prove to be crucial for the hotel business owners.

We will identify the parameters that are most important from the business perspective and get rid of 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. We will also prepare visual analysis that could be easily understood by the stakeholders and decision makers.

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

- Which are the months of Highest and Lowest Customer arrival count?
- Which is the most reserved room type and least reserved room type?
- How many total customers have registered in the hotel?
- Which is the most common and least common customer type?
- Which are the most common countries of origin?
- 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): Time difference between reservation and arrival of guests
- 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
- 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 (refundable/non-refundable/no deposit)
- 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 (Average revenue made by the hotel per room per 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 begin the exploration and analysis of our Hotel Booking dataset. The main environment for the working of our project being used is Python 3 (Google Colaboratory). We will be incorporating different python modules and libraries such as follows:

- **NumPy** and **Pandas** for data analysis
- **Seaborn** and **Matplotlib** for data visualization

The EDA is done using the following steps:

- Data Initialization
- Data Preprocessing
- Data Analysis

These three main steps are further elaborated in the same documentation.

Let us discuss these steps in detail as follows:

I. Data Initialization

This initial stage includes setting up the work environment, importing the dataset into the Python notebook environment.

Importing libraries.

In this step we set up the python notebook environment by installing different python libraries that are essential for our EDA. These include:
NumPy, Pandas, Seaborn and Matplotlib.

Loading the Dataset

Here, we load our dataset that is initially in csv format, into our python notebook environment. For this, we use Pandas library.

II. Data Preprocessing

Now that we have the work environment set up and our dataset loaded, we can start working on the analysis of our data. This requires an initial overview of the dataset as to thoroughly understand the dataset.

Data Overview

By checking the data, we have the following specifications of the data available:

- 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 different parameters of the columns present in the data. These include number of columns, column labels, column data types, memory usage, range index, non-null values.
- Column Description:
  This gives us the aggregate values in all the numeric columns
- Checking for null values:
  This gives us the number of null values present in different columns in our data. We have missing data in the country, agent and company column.

Data Cleaning

Clean data is essential for analysts to get a clearer picture and sense of the data and make important decisions.

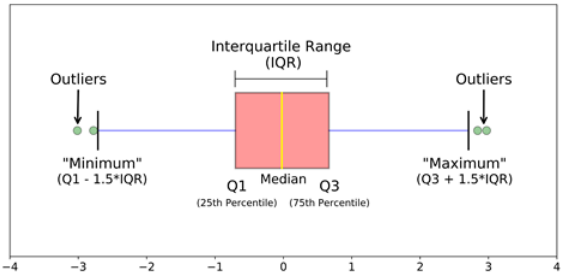
Initially, we remove the features that are of low relevance and that contain so many null values, they are basically unworkable.

After this we take care of the outliers and continue with further data processing.

There are different techniques that can be used to achieve a clean workable dataset.

- Visualization: by boxplot or histogram plot
- Skewness: The skewness value should be within the range of -1 to 1 for a normal distribution, any major changes from this value may indicate the presence of outliers.
- Interquartile Range: IQR
- Standard Deviation: It shows the variability distribution of the data.
- Flooring or capping
- Trimming

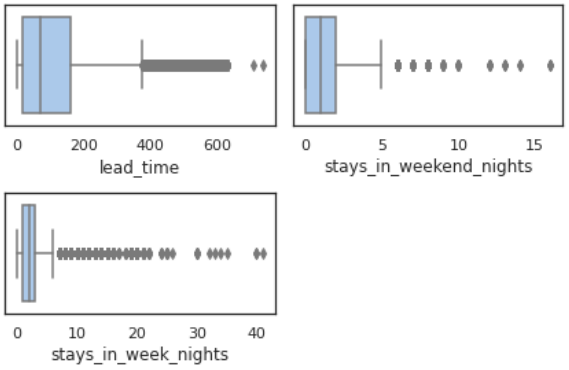
Firstly, we demonstrate and remove the outlier based on our own understanding by setting up the threshold limit. And in terms of outlier, we used IQR. In descriptive statistics, the interquartile range (IQR) is a measure of statistical dispersion, which is the spread of the data. The IQR may also be called the mid spread, middle 50%, or Hspread. It is defined as the difference between the 75th and 25th percentiles of the data.



And lastly, we used quantile-based technique to treat the outlier, Capping is replacing all higher side values exceeding a certain theoretical maximum or upper control limit (UCL) by the UCL value.

- **Outlier detection**

  Using the above strategy, we analyze the features that contain outliers. Some of these are listed below.



III. Data Analysis

Now that we have a clean dataset, we are ready for our analysis. We can now easily evaluate different features alone as well as in combination with other features.

Univariate Analysis:

In this type of analysis, we take a look at one feature at a time and try to get the basic and advanced insights from that feature alone. We use this analysis to get the following insights:

- Understand the trends and patterns of data.
- Analyze the frequency and other such characteristics of data.
- Know the distribution of the variables in the data.
- Visualize the relationship that may exist between different variables

Bivariate Analysis:

In a Bivariate Analysis, we try to analyze two features instead of one, and finally determine the classification of output we are looking for. It is a methodical statistical technique applied to a pair of variables (features/ attributes) of data to determine the empirical relationship between them. In order words, it is meant to determine any concurrent relations. There are three main types of bivariate analysis. They are as follows:

- Scatter Plots - It makes use of dots to represent the values for two different numeric variables.
- Regression Analysis- This involves a wide range of tools that can be utilized to determine just how the data points might be related. It tends to provide us with an equation for the curve/line along with giving us the correlation coefficient.
- Correlation Coefficients - This shows how one particular variable moves about with relation to another.

Multivariate Analysis:

Multivariate analysis deals with such a complex set of data with more than two features and variables. There are two types of multivariate analysis techniques:

- Dependence techniques, which look at cause-and-effect relationships between variables
- Interdependence techniques, which explore the structure of a dataset.

Correlation Matrix

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to

summarize a large dataset and to identify and visualize patterns in the given data.

Challenges

While working with such a large dataset, a data analyst may face many challenges that make their job difficult. Some of the difficulties faced while dealing with the Hotel Booking dataset include:

- Missing/Null values: These missing/null values make the statistical analysis harder. So, one must carefully choose between eliminating the entire feature or removing the null values, depending upon the importance of that feature and the number of null values.
- Unusual Data Type: This is a common problem with many large datasets. The solution for it is typecasting such features into the desired data types, so that it is easier to work with.
- Outlier Values: These values are the ones that lie way outside the range of average values. The problem with having outliers is they make the data unstable, raising issues such as skewness of data, etc. Hence, these must be carefully removed.
- 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

- Room type 'A' is the most reserved room with 71.99% and Room type 'B' is the least reserved room type.
- 'August' is the month of Highest and 'January' is the month of Lowest Customer Arrival count with 11.6% and 4.9% respectively.
- 'Transient' customers are the most common and 'Group' is the least common customer type with 75% and 0.47% respectively.
- Around 84% customers have not made any changes and around 16% customers made booking changes.
- More than 68% guests have only 2 members, around 18% are traveling alone. Some are traveling in groups more than 30.
- Around 37% of customers canceled their bookings.
- 'Resort' average daily price is very fluctuating compared to 'Hotel' which is stable.

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.