EDA On Hotel Booking Analysis

Mohd Faheem
Data science trainee,
AlmaBetter, Bangalore

Abstract:

The knowledge and understanding of quality standards of guests helps hotel managers improve the quality of hotel services and increase guests' satisfaction with the hotel stay. Different aspects of a hotel offer participate in the guests' evaluation of the hotel experience. The factors that influence guests' satisfaction level are also named "hotel attributes". There exist a large number of factors that influence guests' impressions, but certainly not all of them have the same level of importance for guests. In order to be aware of the strengths and weaknesses of their businesses, hotel management has to identify which improvements in the hotel operations can bring additional value to their guests. One of the most reliable information sources for gaining customer knowledge is undoubtedly the feedback provided directly from customers. If analyzed properly, it can be exploited for the purpose of improving the hotel operations and raising profits.We can analize the data with the help of python libraries like Pandas,Numpy,matplotlib and seaborn . we can see insight from the dataset.

1.Introduction

Hotels plays an important role for any person or traveler who are travelling from one destination to another. Hotel play an important role for tourists whether the tourist is local or international. Hotel provides many best services to the customer such as parking area, food, room service and also it provides services that is offered by customer. By providing these services Hotels take the valuable feedback from the customers. By these feedbacks Hotels maintains their reputation in the city/area. If the services are poor, the bookings of that hotel are low and if the services are awesome then high bookings in that hotel takes place.

Hospitality generates revenue for local economies directly when tourists spend money in hotels, restaurants and entertainment venues. Hospitality industry is growing, with more and more people spending their money for vacation and leisure activities. People may only lodge into a hotel when it’s a holiday season or a special event, thus the demand for staying in room is not equally distributed across the year. Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week 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.

2.Problem Statement

A dataset containing 119390 records across 32 features has been given with information regarding bookings of two hotels from July 2015 to August 2017. These two hotels are City Hotel and Resort Hotel.

The main objective is to explore the given dataset and discover the factors which govern the bookings. The dataset will be analyzed and from the conclusions drawn from it will be used to recognize the missteps taken by the manager. With this information, hotels will be equipped to improve their performance.

We will be analyzing some key metrics for hotel bookings like: Most preferred meal types, country wise bookings, new customers acquired, the number of cancellations, number of bookings on weekday vs weekends, customer lifetime value of the existing customers, type of rooms preferred by customers, booking types, hotels available for booking and the revenue of the hotels

Further, we will be using various angles to look through the data to analyze patterns associated with each segment such as: The type of hotel, day of week, type of customers,type of rooms.

Data analysis is performed to answer the following questions:

- Which hotel is more preferred among guest?
- Which hotel retains more customers?
- Which is the busiest month?
- Which is the most popular room type?
- From which country the greatest number of bookings were made?
- How Long People Stay in the hotel?
- How many bookings were cancelled?

3.Objective

Out main objective is perform EDA on the given dataset and draw useful conclusions about general trends in hotel bookings and how factors governing hotel bookings interact with each other.

4.Steps Involved:

4.1.DataSet

We are given a hotel bookings dataset. This dataset contains booking information for a city hotel and a resort hotel. It contains the following features.

- 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.

4.2. Data Cleaning

(1)Removing Duplicate rows

All duplicate rows were dropped.

(2) Handling null values

- Null values in columns company and agent were replaced by mean of the company column and mode of the agent column.
- Null values in column country were replaced by 'bfill'method.
- Null values in column children were replaced by the 0.

(3) Converting columns to appropriate data types

- Changed data type of children, company, agent to int type.

(4) Removing outliers

- One outlier was found in the adr column. Simply dropped it.

(5) Creating new columns

- Created new column total_stay by adding stays_in_weekend_nights+stays_in_week_nights.
- Created new column total_people by adding adults+children+babies.

4.3. Exploratory Data Analysis

EDA was carried out in 3 steps:

1.Correlation Analysis

It is used to measure the strength of the linear relationship between two variables and compute their association. Correlation analysis calculates the level of change in one variable due to the change in the other. Correlation analysis of the dataset was carried out using a correlation heatmap with the features.

2.Univariate Analysis

Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately.

3.Bivariate Analysis

Bi means two and variate means variable, so here there are two variables. The analysis is related to cause and the relationship between the two variables.

5.Libraries

5.1.Numpy

What is NumPy?

NumPy is a Python library used for working with arrays.It also has functions for working in domain of linear algebra, fourier transform, and matrices.NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.NumPy stands for Numerical Python.

Why Use NumPy?

In Python we have lists that serve the purpose of arrays, but they are slow to process.NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.Arrays are very frequently used in data science, where speed and resources are very important.

Why is NumPy Faster Than Lists?

NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently.This behavior is called locality of reference in computer science.This is the main reason why NumPy is faster than lists. Also it is optimized to work with latest CPU architectures.

Which Language is NumPy written in?

NumPy is a Python library and is written partially in Python, but most of the parts that require fast computation are written in C or C++.

5.2.Pandas

What is Pandas?

Pandas is a Python library used for working with data sets.It has functions for analyzing, cleaning, exploring, and manipulating data.The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Why Use Pandas?

Pandas allows us to analyze big data and make conclusions based on statistical theories.Pandas can clean messy data sets, and make them readable and relevant.Relevant data is very important in data science.

What Can Pandas Do?

Pandas gives you answers about the data. Like:

- Is there a correlation between two or more columns?
- What is average value?
- Max value?
- Min value?

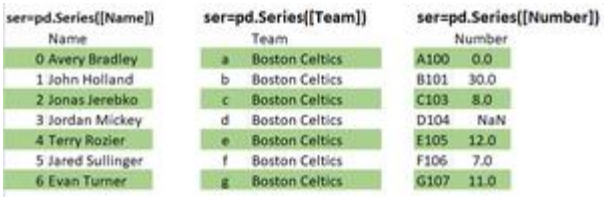
Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called *cleaning* the data.

Advantages

- Fast and efficient for manipulating and analyzing data.
- Data from different file objects can be loaded.
- Easy handling of missing data (represented as NaN) in floating point as well as non-floating point data
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects
- Data set merging and joining.
- Flexible reshaping and pivoting of data sets
- Provides time-series functionality.
- Powerful group by functionality for performing split-apply-combine operations on data sets.

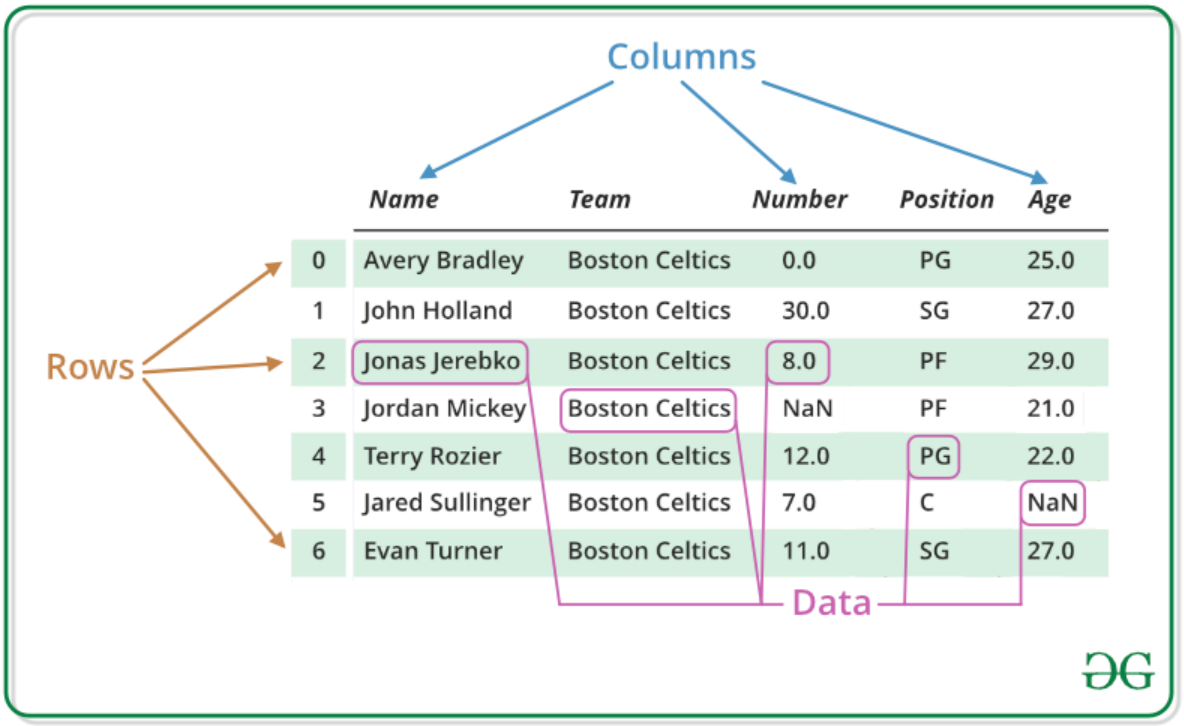
- **Series:** Pandas Series is a one-dimensional labeled array capable of holding data of any type (integer, string, float, python objects, etc.). The axis labels are collectively

called indexes. Pandas Series is nothing but a column in an excel sheet. Labels need not be unique but must be a hashable type. The object supports both integer and label-based indexing and provides a host of methods for performing operations involving the index.

- 

DataFrame

Pandas DataFrame is a two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns). A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas DataFrame consists of three principal components, the data, rows, and columns.



5.3.Matpoltlib

What is Matplotlib?

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.

Matplotlib is a low level graph plotting library in python that serves as a visualization utility.Matplotlib was created by John D. Hunter.Matplotlib is open source and we can use it freely.Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility.

Basic plots in Matplotlib :

Matplotlib comes with a wide variety of plots. Plots helps to understand trends, patterns, and to make correlations. They're typically instruments for reasoning about quantitative information. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Important Types of Plots

- Bar graphs
- Histograms
- Scatter plots
- Line plot
- Pie Plot

5.4.Seaborn

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.

Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

Different categories of plot in Seaborn

Plots are basically used for visualizing the relationship between variables. Those variables can be either be completely numerical or a category like a group, class or division. Seaborn divides plot into the below categories –

- **Relational plots:** This plot is used to understand the relation between two variables.
- **Categorical plots:** This plot deals with categorical variables and how they can be visualized.
- **Distribution plots:** This plot is used for examining univariate and bivariate distributions
- **Regression plots:** The regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyses.
- **Matrix plots:** A matrix plot is an array of scatterplots.
- **Multi-plot grids:** It is an useful approach is to draw multiple instances of the same plot on different subsets of the dataset.

Installation

For python environment :

pip install seaborn

Important Features of Seaborn

Seaborn is built on top of Python's core visualization library Matplotlib. It is meant to serve as a complement, and not a replacement. However, Seaborn comes with some very important features. Let us see a few of them here. The features help in −

- Built in themes for styling matplotlib graphics
- Visualizing univariate and bivariate data
- Fitting in and visualizing linear regression models
- Plotting statistical time series data
- Seaborn works well with NumPy and Pandas data structures
- It comes with built in themes for styling Matplotlib graphics

7.Conclusion

The following conclusions were drawn from analysis:

- City hotels are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
- Most number of bookings are made in July and August.
- agent iD no 9 made most number of booking
- The length of the stay decreases as ADR increases probably to reduce the cost.
- Room Type A is the most preferred room type among guests.
- 27.5 % bookings were got cancelled out of all the bookings.
- Only 3.9 % people were revisited the hotels. Rest 96.1 % were new guests. Thus retention rate is low.
- The percentage of 0 changes made in the booking was more than 82 %. Percentage of Single changes made was about 10%.
- Most of the customers (91.6%) do not require car parking spaces.
- 79.1 % bookings were made through TA/TO (travel agents/Tour operators).
- BB(Bed & Breakfast) is the most preferred type of meal by the guests.
- Maximum number of guests were from Portugal, i.e. more than 25000 guests
- Most of the bookings for City hotels and Resort hotel were happened in 2016.
- Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
- Booking cancellation rate is high for City hotels which almost 30 %.
- Average lead time for resort hotel is high.
- Waiting time period for City hotel is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
- Optimal stay in both the type hotel is less than 5 days.

References:

1. https://www.w3schools.com/python/numpy/numpy_intro.asp
2. https://www.w3schools.com/python/pandas/pandas_intro.asp
3. https://www.geeksforgeeks.org/introduction-to-pandas-in-python/
4. https://www.w3schools.com/python/matplotlib_intro.asp
5. https://www.geeksforgeeks.org/python-introduction-matplotlib/
6. https://www.geeksforgeeks.org/introduction-to-seaborn-python/