EDA - Hotel Booking Analysis

Arvind kale

Priyanka Jain

Mohd Sakib Quraishi

Nitish Kumar

ALMA BETTER TRAINEE

**Abstract:**

Businesses which generate tremendous quantities of data have to not only leverage it to recognize their modern-day performance (i.e. enterprise intelligence) however additionally – and most importantly – to generate prescriptive analytics to orient their strategy towards Data science, which can also additionally at the beginning appear overly ‘techy’ and ‘expensive’ is genuinely pretty feasible. The data analytics performs a pivotal position in hotel industry as it is key to advertising strategy, facilitates in building customer loyalty, and all in all amplify their customer base. The 5 approaches wherein data analytics makes an effective effect at the hotel industry are Customer Data Analysis & Market Segmentation, Real-Time Data and Hotel Pricing Strategies, Managing Hotel Booking Channels, Inventory Management and Demand Forecasting. This is best a partial synopsis of the feasible strategic makes use of data science. However, the primary message holds actual: do not be frightened of facts. Data is everywhere and can enhance the competitiveness of your enterprise. Data also can guide product development, advertising and operations whilst additionally helping in the transformation towards “smart” businesses. This is true for any single expert within the travel-value chain, from hotelier and destination manager to service provider.

The project gives various insight of hotel bookings which can be leveraged to improve business performance and customer service

***Keywords: Exploratory data analysis, pandas, numpy, seaborn, matplotlib,***

***Python, Google Collaboratory***

1.Introduction

The data generated every year is increasing day by day. Bussinesses are trying get insights of the data and use it for the decision making . It also helps in improving performance of the business.

**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. Main aim of the project is to understand and visualize dataset from hotel and customer point of view.**

**1 Reasons for booking cancellations across various parameters.**

**2 best time to book hotel .**

**3 peak season.**

2. Problem statement

The data is of hotel bookings done in various months of the year. It can help us to discover insight like which type of meal is preferred by the customer? From which country the most customers arrive from? Which month has the highest bookings or which month has highest number of cancelation? The data has various columns like 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. Python packages such as pandas, matplotlib and seaborn provide functionalities to analyze the data

3. Column Description

hotel : It has information about type of the Hotel booked. (H1 = Resort Hotel or H2 = City Hotel).

is\_canceled : It indicates if the booking was cancelled (1) or not (0).

lead\_time : It tells about number of days that elapsed between the entering date of the booking into the PMS and the arrival date.

arrival\_date\_year : Year of arrival date.

arrival\_date\_month : Month of arrival date.

arrival\_date\_week : Week number of year for arrival date.

arrival\_date\_day : Day of arrival date.

stays\_in\_week\_nights : Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.

adults : Number of adults.

children : Number of children.

babies : Number of babies

.meal : It indicates type of meal booked. It includes the following categories. 1. Undefined/SC – no meal package; 2. BB – Bed & Breakfast; 3. HB – Half board (breakfast and one other meal – usually dinner); 4. FB – Full board (breakfast, lunch and dinner)

country : It tells about country of origin. Categories are represented in the ISO 3155–3:2013 format

market\_segment : It is Market segment designation. In categories, the term TA means “Travel Agents” and TO means “Tour Operators”.

distribution\_channel : It includes Booking distribution channel. The term TA means “Travel Agents” and TO means “Tour Operators”.

is\_repeated\_guest : Value indicating if the booking name was from a repeated guest (1) or not (0).

previous\_cancellations : It indicates number of previous bookings that were cancelled by the guest prior to the current booking.

previous\_bookings\_not\_canceled : It indicates number of previous bookings not cancelled by the guest prior to the current booking.

reserved\_room\_type : It contains code of room type reserved.

assigned\_room\_type : It contains code for the type of room assigned to the booking which may differ from the reserved room type.

booking\_changes : It indicates number of changes made to the booking.

deposit\_type : It indicates that if the guest has made a deposit to guarantee the booking. This feature can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.

agent : ID of the travel agency that made the booking.

company : ID of the company that made the booking.

days\_in\_waiting\_list : It indicates number of days the booking was in the waiting list.

customer\_type : Type of booking, assuming one of four categories: 1. Contract - when the booking has an allotment or other type of contract associated to it; 2. Group – when the booking is associated to a group; 3. Transient – when the booking isn't part of a group/contract, and isn't associated to other transient booking. 4. Transient-party – when the booking is transient, but is associated to at least other transient booking

adr : Average daily rate is the average revenue that a hotel receives for each occupied guest room per day.

required\_car\_parking\_spaces : Number of car parking spaces required by the customer.

total\_of\_special\_requests : Number of special requests made by the customer.reservation\_status : The last status of reservation, assuming one of three categories: 1. Canceled – booking was canceled by the customer; 2. Check-Out – customer has checked in but already departed; 3. No-Show – customer did not check-in and did inform the hotel of the reason why

reservation\_status\_date : Date at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when was the booking canceled or when did the customer checked-out of the hotel.

4. Exploraotory Data Analysis

EDA stands for exploratory data analysis. Developed by American mathematician John Tukey in the 1970s, it describes the critical process of first examining data. It is used by data scientists to summarize the main characteristics of datasets using various data visualization methods such as:, graphic representations. It allows us to manipulate raw data sets to get the answers needed when spotting patterns, spotting anomalies, testing a hypothesis, or verifying assumptions.   
EDA is mainly used to find out what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of the properties of the dataset and the correlation between them. EDA allows us to determine whether or not the statistical techniques used for data analysis are appropriate.

The objectives of EDA are to:

* Enable unexpected discoveries in the data
* Suggest hypotheses about the [causes](https://en.wikipedia.org/wiki/Causality) of observed [phenomena](https://en.wikipedia.org/wiki/Phenomenon)
* Assess assumptions on which [statistical inference](https://en.wikipedia.org/wiki/Statistical_inference) will be based
* Support the selection of appropriate statistical tools and techniques
* Provide a basis for further data collection through [surveys](https://en.wikipedia.org/wiki/Survey_sampling) or [experiments](https://en.wikipedia.org/wiki/Design_of_experiments)

**Tools Required For Exploratory Data Analysis:**

Some of the most common tools used to create an EDA are:

**2. Python:** An interpreted, object-oriented programming language with dynamic semantics. Its high level, built-in data structures, combined with dynamic binding, make it very attractive for rapid application development, also as to be used as a scripting or glue language to attach existing components together. Python and EDA are often used together to spot missing values in the data set, which is vital so you’ll decide the way to handle missing values for machine learning.

**Role of data analytics in hospitality industry**:

**1) Customer Data Analysis & Market Segmentation**

* It helps to understand the preferences and purposes of a customer.
* According to which customers could be categorized.
* It helps to target right audience for marketing and sales.

**2) Real-Time Data and Hotel Pricing Strategies**

* The analyses of booking patterns shows the demand trends which can be used to implement dynamic pricing.
* It can be used to create tailor-made packages.

**3) Managing Hotel Booking Channels**

* Hotels receive bookings from various channels such as online travel agencies (OTA), direct bookings, and website bookings.
* By analysing the data from different mediums, it can be understood which channel yields the most bookings and which channel needs to be worked on. It also helps to identify the different customer base on these booking channels.
* This data is important to formulate the right marketing strategy.

**4) Inventory Management**

* Data Analytics in hotels is useful for inventory management by maintaining a balance of inventory – you neither run out of inventory nor have an excess of it at any point.

**5) Demand Forecasting**

* By analyzing booking patterns in real-time, the hotels can perform demand forecasting which is useful for revenue management, inventory management, and implementing dynamic pricing strategies. Data Analytics and machine learning help define the optimal room rate in real-time to maximize profit.

**Steps involved**:

* Importing the Libraries: The required libraries such as pandas, numpy, seaborn and matplotlib were imported.
* Initial observation of data: The head of the data is printed and its shape is observed which is found to be (119390, 32)
* Descriptive statistics: The descriptive stats was, then, analysed using the function describe 0. It includes mean, median, mode, frequency, standard deviation, etc.
* Null values and Data types: This step includes the identification of data types for all the variables and the features containing null values.
* Data preprocessing: It is the process of transforming raw data into an understandable format. Preprocessing of data is mainly to

check the data quality which includes data accuracy, completeness, consistency and interpretability. In this step, the null values of the categorical variable (country) was filled with the mode while the null values of numerical variables were filled with their respective means. Further, the rows with zero number of guests were dropped as such data is of no use. In last step of pre-processing, the variables with ‘string’ data type were changed to ‘int’ type.

**Libraries used:**

Matplotlib: This library is responsible for plotting numerical data. And that’s why it is used in data analysis. It is also an open-source library and plots high-defined figures like pie charts, histograms, scatterplots, graphs, etc.

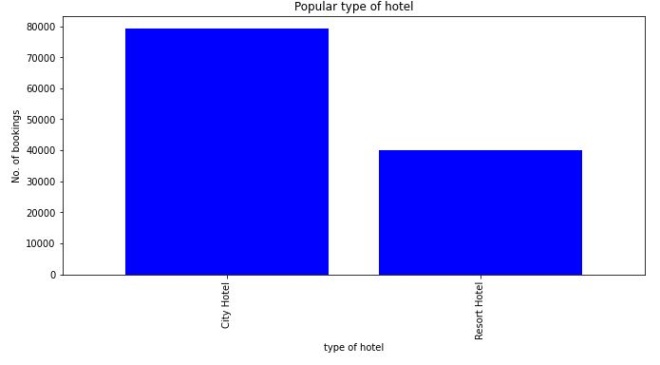
Pandas: Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.

Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. For a brief introduction to the ideas behind the library, you can read the introductory notes or the paper.

Exploratory Data Analysis:

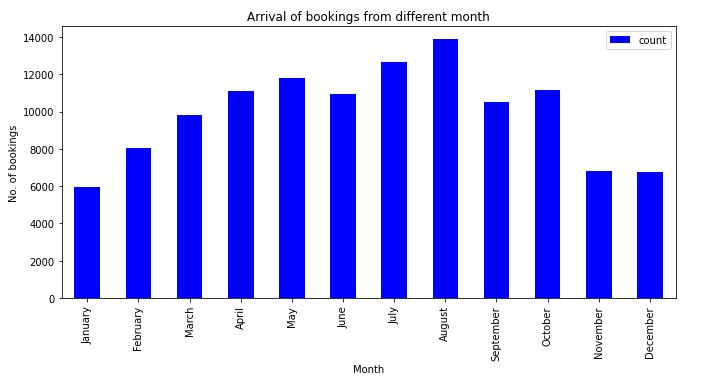
1. Type of hotel prefered:

With pandas and matplotlib visualization we can clearly see from the graph that the city hotel is prefered by more customers than the resort hotel



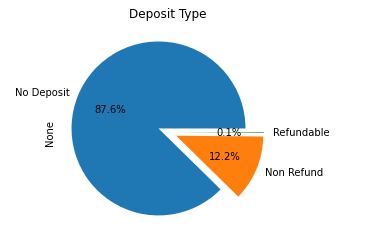
2.Booking Acoording to the monthly data:

The monthly booking data shows that the booking were highest in the month of august while booking were lowest in the month of november, december and january



3.Type of deposit:

The following pie chart shows that almost 88% percent guests does not prefer deposit . 12% of the guests did the non refundable deposit means these type of guests are unlikely to cancel their bookings.



8.Challenges faced:

First of all ,data contains a lot of null values and missing value which have to be clean in order to perform EDA.

There were outlier in the dataset which we have to treat.

The Process is time consuming and required patience and team effort to run in a proper way.

9. **Conclusion**:

1. BB is the most preferred type of meal

2. .Although November and December has almost same bookings ,cancelation rate in December is more

3. Portugal is the country from which most number of guests come from

4. Almost 88 percent of the guets prefer to keep no deposit in advace.

5..41.7% of the total bookings were cancelled for City hotel and 21.7% for the Resort hotel.

6. Guest from Portugal and great Britain prefer economical rooms...

7. Almost 35% of bookings were canceled

9.References:

a. hospitalityinsights.ehl.edu

b.www.researchgate.net/publication/329286

c. becominghuman.ai

d., www.rtinsights.com