EDA - Hotel Booking Analysis

**RAHUL Y M**

ALMABETTER TRAINEE

**Abstract:**

Businesses that generate tremendous quantities of data have to not only leverage it to recognize their modern-day performance (i.e. enterprise intelligence) but additionally – and most importantly – to generate prescriptive analytics to orient their strategy toward Data science, which can also add at the beginning appear overly ‘techy’ and ‘expensive’ is genuinely pretty feasible. Data analytics performs a pivotal position in the 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 on 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 strategy that 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 hoteliers and destination managers to service providers.

The project gives various insights into 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. Businesses are trying to get insights into the data and use it for decision-making. It also helps in improving the 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. The main aim of the project is to understand and visualize the datasets from hotel and customer points of view.**

**1. Reason for booking cancelations across various parameters.**

2. **Best time to book a hotel**

**3. Peak season.**

2. Problem statement

The data is from 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? Which month has the highest bookings or which month has the highest number of cancelations? 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 the type of the Hotel booked. (H1 = Resort Hotel or H2 = City Hotel).

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

lead\_time: It tells about a 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 weeknights (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 the 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 the 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 too means “Tour Operators”.

distribution\_channel : It includes the 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 a number of previous bookings that were canceled by the guest before the current booking.

previous\_bookings\_not\_canceled : It indicates a number of previous bookings not canceled by the guest before the current booking.

reserved\_room\_type : It contains the 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 a number of changes made to the booking.

deposit\_type : It indicates that if the guest has deposited 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 the 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 a number of days the booking was on 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 with it; 2. Group – when the booking is associated with a group; 3. Transient – when the booking isn't part of a group/contract and isn't associated with another transient booking. 4. Transient-party – when the booking is transient but is associated with at least another 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 – the 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 the customer checked out of the hotel.

4. Exploratory 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:

**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, analyzed 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.

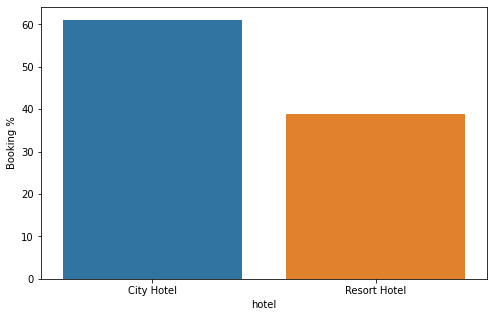
Pandas: Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing,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

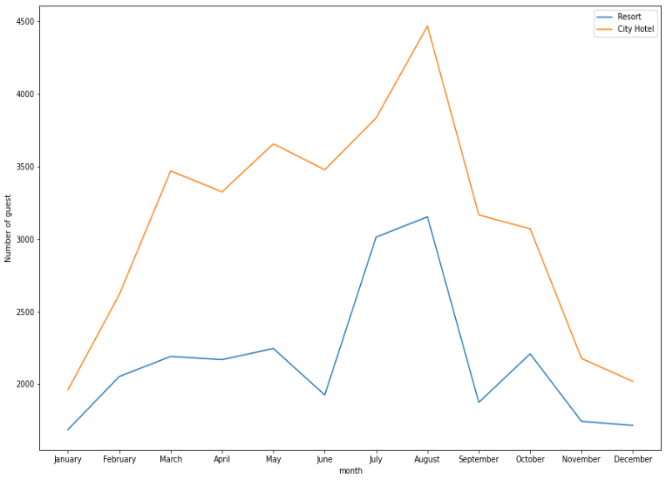
1. Type of hotel prefered:

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



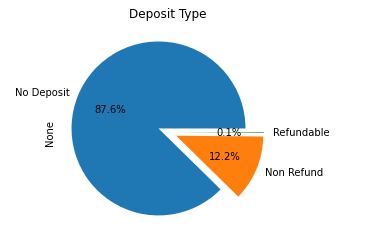
2.Booking According to the monthly data:

The monthly booking data shows that the booking was highest in the month of august while booking were lowest in the month of November, December and January

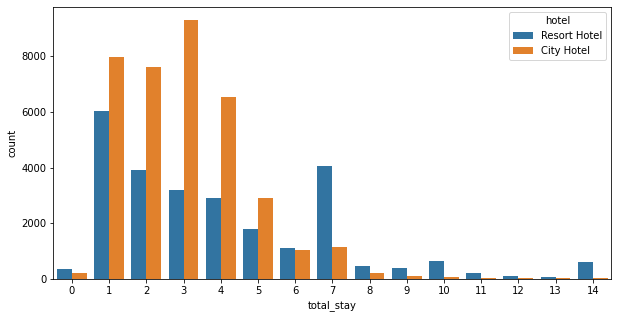


3. Type of deposit:

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.

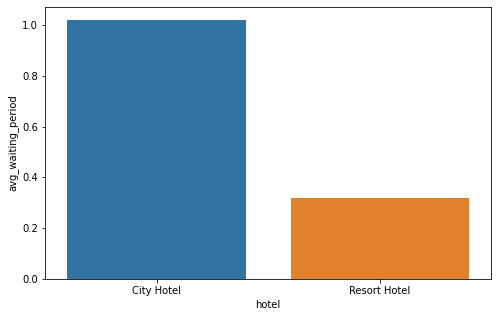


4. What is preferred stay length in each hotel?



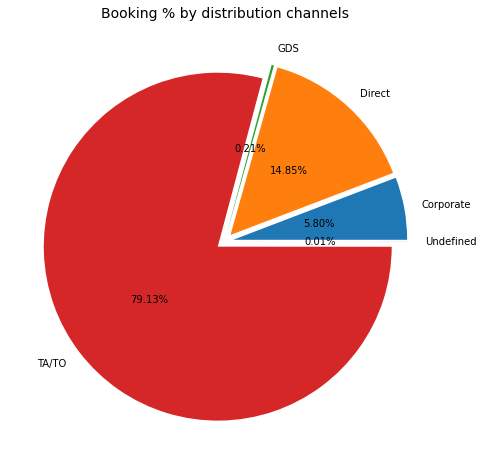
Most common stay length is less than 4 days and generally, people prefer a City hotel for a short stay, but for longer stays, Resort Hotel is preferred.

5. Which hotel has longer waiting time?



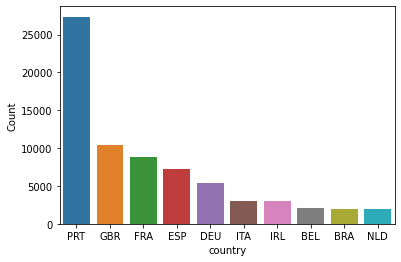
City hotel has significantly longer waiting time, hence City Hotel is much busier than Resort Hotel.

6. Which channel is mostly used for early booking of hotels?



TA/TO is mostly used for planning Hotel visits ahead of time. But for sudden visits other mediums are most preferred.

7.From where are the most guests coming?



8. Challenges faced:

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

There were outliers in the dataset that we have to treat.

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

9. **Conclusion**:

1. Bed and Breakfast is the most preferred type of meal

2. Most of the bookings are done through TA/TO and Direct channel.

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

4. Almost 88 percent of the guests prefer to keep no deposit in advance.

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

6. Both hotels have a very small percentage that customers will repeat, but the Resort hotel has a slightly higher repeat % than City Hotel.

7. Guests from Portugal and Great Britain prefer economical rooms...

8. Almost 35% of bookings were canceled.

9. Mostly bookings are done by couples (although we are not sure that they are a couple as data doesn't tell about that)

It is clear from the graph that there is a sudden surge in the arrival num of couples and families in the months of July and August. So better plans can be planned accordingly at that time for these types of customers.

9. References:

a. hospitalityinsights.ehl.edu

b. geekforgeeks

c. kaggle

d.analyticsvidhya