# Hotel Bookings Exploratory Data Analysis

**Objective**

Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting nonobvious and useful patterns from large data sets. Businesses which generate tremendous quantities of data have to not only leverage it to recognize their modern-day performance (i.e. enterprise intelligence)

Herein we are provided with the data of hotel booking. In which our main objective is to perform Exploratory data analysis on the given dataset and draw useful conclusions about general trends in hotel bookings and how factors governing hotel bookings interact with each other. The project gives various insight of hotel bookings which can be leveraged to improve business performance and customer service

**Dataset**

We are been given with the hotel bookings data. This data 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.

## Data Cleaning

### (1) Removing Duplicate rows

All duplicate rows were removed and dropped.

### (2) Handling null values

* Null values in columns childrens, company and agent were replaced by 0.
* Null values in column country were replaced by 'others'.

**(3) We have Converted the columns to appropriate data types so that we would be able to extract usefull insights**

* Changed data type of children, company, agent are converted to integer type.
* Changed data type of reservation\_status\_date to date 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.

## Exploratory Data Analysis

## Performed Exploratory Data Analysis and tried answering the following questions:

Q1Which meal type is most preffered meal of customers?

Ans. BB is the most preferred meal of customers as represented into the data approximately 78%peoples preffered for it and HB and HC places on the same instance and in the last FB and UNDIFINED are the least.

Q2 Which agent makes most no. of bookings?

Ans. Agent no. 9 has made most no. of bookings then comes 240 and 14 then 7 , 250 , 241 , 28, 8, 1, 6 in the decreasing order of the sales.

Q3 Which room\_type is in most demand and which room\_type generates highest adr?

Ans. A types of rooms are most demanded then comes D,E,F,G, etc.

Q4 What is Booking Percentage in each hotel?

Ans. Around 60% bookings are for City hotel and 40% bookings are for Resort hotel, therefore City Hotel is busier than Resort hotel.

Q5  Which amoung the hotel has higher which has lower lead time?

Ans. City hotel has slightly higher median lead time. Also median lead time is significantly higher for both hotels, this means customers generally plan their hotel visits way early.

Q6  which hotel seems to make more revenue?

Ans. Overall adr of City hotel is slightly higher than Resort hotel and no. of bookings of City hotel is also higher than Resort hotel. Hence, City hotel is makes more revenue.

Q7  Which hotel has longer waiting time and which has less waiting time ?

Ans. City hotel has higher waiting time and resort hotel has less waiting time as the no of bookings in the city hotel is more compared with the resort hotels

Q8 Arrival of booking from diff segments?

Ans. The maximum bookings are been done through online mode in city hotel as well resort hotel

Q9 Deposite whether it is refundable nonrefundable or no depoist?

Ans. Approx 87.6% persons prefferes to book those hotel who don’t charge any deposit, 1%of the persons have booked

Through refundable deposit and 12.2% persons have booked through non refundable deposits

Mainly performed using Matplotlib and Seaborn library and the following graph and plots had been used:

* Heatmap.—Correlation Matrix
* Scatter Plot.--- Scatter Plot Data Analysis
* Bar Chart --- Bar plot
* Pie Chart. --- Pie plot
* Line Plot.
* Box Plot

## Challenges

(1) There was a lot of duplicate data.

(2) Data was present in wrong datatype format.

(3) Choosing appropriate visualization techniques to use was difficult.

(4) A lot of null values were there in the dataset.