**Hotel Bookings Exploratory Data Analysis**

## **Objective**

Main objective is to 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.

## **Dataset**

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

|  |  |  |
| --- | --- | --- |
| **Sr.No.** | **Data Input** | **Description** |
| 1 | hotel | City and Resort hotel |
| 2 | is\_canceled | indicating booking cancelled (1) or not cancelled (0) |
| 3 | lead-time | the time difference between booking date and actual check in in the hotel |
| 4 | arrival\_date\_year | Year of arrival date |
| 5 | arrival\_date\_month | Month of arrival date |
| 6 | arrival\_date\_week\_number | Week no of year for arrival date |
| 7 | arrival\_date\_day\_of\_month | day of arrival date |
| 8 | stays\_in\_weekend\_nights | no of weekends night |
| 9 | stays\_in\_week\_nights | no of week nights |
| 10 | adults | no of adults |
| 11 | children | no of children |
| 12 | babies | no of babies |
| 13 | meal | type of meal   1. **BB** : Bed and Breakfast 2. **HB** : Half Board (Breakfast and Dinner normally) 3. **FB** : Full Board (Breakfast, Lunch and Dinner) 4. **SC** : Self-catering 5. **Undefined**: Rooms only packages without meals |
| 14 | country | customers country of origin |
| 15 | market\_segment | Market segment type -This provides source of information through which customer booked  **1.TA** - "Travel Agent"  **2. TO** - "Tour operators"  **3.Direct** -"Direct booking |
| 16 | distribution\_channel | booking description channel is the source of information through which customer booked hotel  **1.TA/TO** - "Travel Agent"/"Tour operators"  **2.Direct** -"Direct booking"  **3.Corporate**- "Corporate booking" |
| 17 | is\_repeated\_guest | if repeated guest (1) or no(0) |
| 18 | previous\_cancellations | no of previous bookings those are cancelled by the customer before the current booking |
| 19 | previous\_bookings\_not\_canceled | no of previous bookings not cancelled by the customer before the current booking |
| 20 | reserved\_room\_type | Type of reserved room  C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'P' 'B' |
| 21 | assigned\_room\_type | Type of assigned room  C' 'A' 'D' 'E' 'G' 'F' 'I' 'B' 'H' 'P' 'L' 'K' |
| 22 | booking\_changes | no of changes made in the booking from the moment the booking was entered till check in or cancellation  3 4 0 1 2 5 17 6 8 7 10 16 9 13 12 20 14 15 11 21 18 |
| 23 | deposit\_type | no deposit or refundable or non-refundable  No Deposit' 'Refundable' 'Non-Refund' |
| 24 | agent | ID of travel agent |
| 25 | company | ID of the company that made the booking |
| 26 | days\_in\_waiting\_list | no of days the booking was in waiting list |
| 27 | customer\_type | type of customer contract, group   1. **Transient** 2. **Transient-Party** 3. **Group** 4. **Contract** |
| 28 | adr | Average daily rate |
| 29 | required\_car\_parking\_spaces | required car parking spaces |
| 30 | total\_of\_special\_requests | no of special request |
| 31 | reservation\_status | reservation last status  'Check-Out' 'Cancelled' 'No-Show' |
| 32 | reservation\_status\_date | check out date |

* Total number of rows in data: 119390
* Total number of columns: 32

## **Prerequisites**

* Import Python libraries.
* Mount google drive to google colab
* Authorize notebook to access google drive files

## **Understand dataset input**

* Find out the total columns and rows of dataset
* Find the data type of each column.
* Find the continuous and categorical data
* Find individual distribution for some of the columns
* Also check the correlation between dependent columns

## **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 0.
* Null values in column country were replaced by 'others'.
* Null values in column children were replaced by the mean of the column.

## **Exploratory Data Analysis**

Performed EDA and tried answering the following questions:

1. Find the type of hotel people are booking?
2. Find in which month the people book the hotel?
3. Find in which year the people book the hotel?
4. Which are the most busiest months?
5. Which hotel has higher bookings cancellation rate?
6. What is the average price per room?
7. Correlation Heatmap

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

* Bar Plot.
* Seaborn plot.
* Line Plot.
* Box Plot.
* Heatmap.

**Challenges**

1. Dataset contains a lot of duplications.
2. Against few columns having a lot of Null values.
3. Few dataset columns with wrong datatype format.

## **Conclusions:**

1. City hotel bookings has twice the amount of reservations as compared to Resort hotel booking over same period year
2. People are booking city hotels more than Resort hotels. In which month the people book the hotel
3. More than double bookings were made in the year 2016, compared to the previous year. But the bookings decreased in next year 2017
4. Both city and resort hotel bookings are peak in August and lowest in January. From March to October it seems to be busiest for the year. Bookings falls down from November to January.
5. The average price per room depends on its type and the standard deviation
6. Correlation Heatmap