Date: 21:10:2022

**Technical document**

**Capstone Project 1: Hotel Bookings Analysis**

**Team Members:**

**1) Deepanshu Kishore**

**2) Gaurav**

**3) Abhishek**

**4) Shaifali**

**Objective**

The key objective of this project to analyze and explore the given data to conclude the meaningful important factors which can help the hotel management to improve both revenue and quality. Also mainly root cause analysis for the cancellation cases can be scrutinize to take necessary preventive actions

**Dataset**

|  |  |  |
| --- | --- | --- |
| **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 |

**Prerequisites**

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

**Basic Analysis**

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

**Data Cleaning and manipulation** .

**Dataset size:** **119390 rows × 32 columns**

* Identify duplicated rows and remove the same.

**Dataset size:** **87396 rows × 32 columns**

* Calculate percentage values of null values of each column.
* Replace NaN values with 0 for heading Agent & company
* Replace NaN values with their mean values for heading children
* Replace NaN values with 'others' for heading Country

**Exploratory Data Analysis and visualization**

With EDA we have analysed below questions

1. Hotel type analysis
2. Finding cancellation rate on each type of hotel
3. Find peak business season of hotel booking
4. Type of rooms in most demand
5. Distributor Channel Analysis
6. Agent wise bookings Analysis
7. Company wise bookings Analysis
8. Market Segment -Booking Analysis
9. Hotel booking cancellation on basis of days\_in\_waiting\_list
10. Analysis of is\_repeated\_guest column

**Key Conclusions:**

1. City hotel [61.1%] having more booking as compared to resort hotel [38.9%]."
2. Overall, 2017 is most busiest year
3. City Hotel having overall more waiting time which interpret it is more crowded than Resort
4. Here we can see the maximum number of customers are from transient category which is near about 75.1%.
5. Distributor channel TA/TO is giving the most booking business
6. Agent ID 9 is giving maximum hotel bookings so this data can be utilized to decide commission % for agent
7. There are high chances of cancellation when waiting period is high.

Also, with Matplotlib and Seaborn library the following graphs had been plotted

* Bar Plot
* Pie Chart.
* Line Plot.
* Box Plot

**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.