**Exploratory Data Analysis -**

**HOTEL BOOKING ANALYSIS**

****

* **By Poonam Shevkar**

**Data Science Trainee**

**At AlmaBetter**

* **INTRODUCTION-**

A Hotel Booking is the arrangement where people stay to spend their vacations or holidays by paying for rooms,their meals and many more.

The word Booking is described as reserving or blocking a specific room for a guest for a certain period of time which is made as per the request made by the guest while booking. For a guest, reserving a room in advance increases the chances to a great extent to get a nice deal and they will be assured of room on arrival.

Hotel Booking procedure starts with the reservation inquiry, checking the room availability,reservation source,cancellations,lead time , meal preference and among others.

In this project, we are going to have an overview and predict the factors which are responsible for Hotel Booking parameters.

* **OBJECTIVE-**
* To study and analyse the best time of Hotel Bookings in a year.
* To analyse the optimal stay and price variation over a year .
* To predict whether or not hotels received the special requests from customers in proportionality.
* To analyse the Hotel Bookings from different parts of the countries.
* To study meal preference and Hotel Bookings/Cancellations.
* To study the correlation between the variables.
* **PROBLEM STATEMENT-**

In the proposed system , we are going to use the Hotel Booking Analysis database to analyze the best time of year to book a hotel room , to analyze the optimal stay to get the best rate of hotel room and also, to predict the special request received to hotels from customers .

The data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. From it, we can understand the customer's behavior and it might help us make better decisions.

* **METHODOLOGY- Exploratory Data Analysis**

Using EDA, we will perform analysis to understand the datasets and to show the visualizations of hotel bookings factors with different features and also hotel booking around the worldwide by the customers.With graphs, we will study the factors which are inter-related with each other.

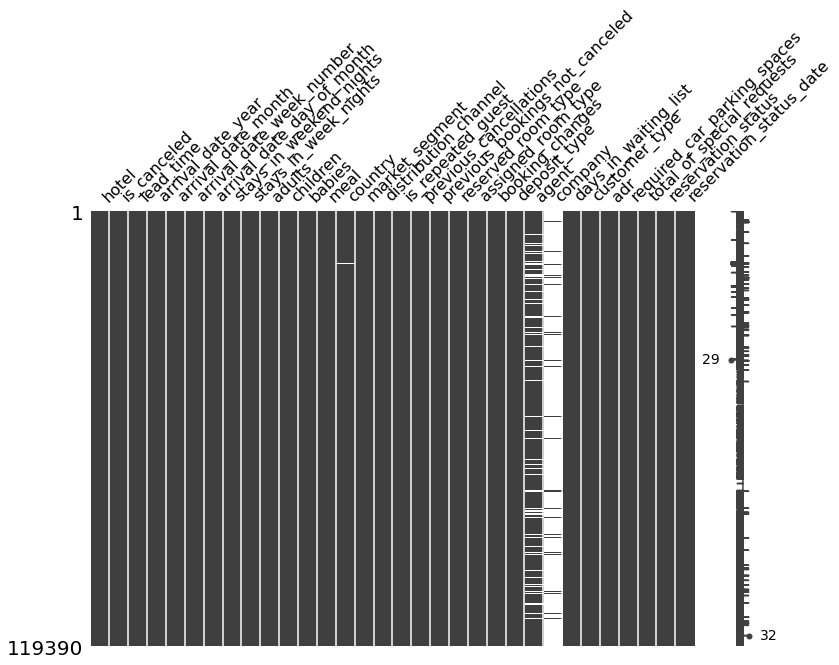
We will predict whether or not a hotel was likely to receive a disproportionately high number of special requests.

* **DATA PREPPING-**

This data describes datasets with hotel demand data. One of the hotels is a resort hotel and the other is a city hotel. The datasets share the systematic structure, with 32 variables describing the 119390 observations. Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. Since this is hotel real data, all data elements pertaining hotel or customer identification were deleted. Due to the scarcity of real business data for scientific and educational purposes, these datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

**Columns used in the analysis-**

* **Hotel**
  + H1: Resort hotel
  + H2: City hotel
* **is\_canceled**
  + 1: Canceled
  + 0: Not canceled
* **lead\_time**
  + No of days that elapsed between entering date of booking into property management system and arrival date
* **arrival\_date\_year**
  + Year of arrival date (2015-2017)
* **arrival\_date\_month**
  + Month of arrival date (Jan - Dec)
* **arrival\_date\_week\_numberr**
  + Week number of year for arrival date (1-53)
* **arrival\_date\_day\_of\_month**
  + Day of arrival date
* **stays\_in\_weekend\_nights**
  + No of weekend nights (Sat/Sun) the guest stayed or booked to stay at the hotel
* **stays\_in\_week\_nights**
  + No of week nights (Mon - Fri) the guest stayed or booked to stay at the hotel
* **Adults**
* **Children**
* **Babies**
* **meal**
  + Type of meal booked. Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
* **country**
* **market\_segment** (a group of people who share one or more common characteristics, lumped together for marketing purposes)
  + TA: Travel agents
  + TO: Tour operators
* **distribution\_channel** (A distribution channel is a chain of businesses or intermediaries through which a good or service passes until it reaches the final buyer or the end consumer)
  + TA: Travel agents
  + TO: Tour operators
* **is\_repeated\_guest** (value indicating if the booking name was from repeated guest)
  + 1: Yes
  + 0: No
* **previous\_cancellations**
  + Number of previous bookings that were cancelled by the customer prior to the current booking
* **previous\_bookings\_not\_canceled**
  + Number of previous bookings not cancelled by the customer prior to the current booking
* **reserved\_room\_type**
  + Code of room type reserved. Code is presented instead of designation for anonymity reasons.
* **assigned\_room\_type**
  + Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
* **booking\_changes**
  + Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
* **deposit\_type**
  + Indication on if the customer made a deposit to guarantee the booking. This variable 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/entity that made the booking or responsible for paying the booking.
* **day\_in\_waiting\_list**
  + Number of days the booking was in the waiting list before it was confirmed to the customer.
* **customer\_type**
  + Contract - when the booking has an allotment or other type of contract associated to it;
  + Group – when the booking is associated to a group;
  + Transient – when the booking is not part of a group or contract, and is not associated to other transient booking;
  + Transient-party – when the booking is transient, but is associated to at least other transient booking
* **adr (average daily rate)**
* **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 (e.g. twin bed or high floor)
* **reservation\_status**
  + Canceled – booking was canceled by the customer;
  + Check-Out – customer has checked in but already departed;
  + 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 ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel
* **CHALLENGES FACED-**
* **Size**: The size of the dataset right away is the strength as well as the most challenging aspect of this dataset. Although there are 32 fields that describe every factor, it’s only natural that such a wide range of information may not be available for every Hotel Bookings. This results in a large number of NAN values.



* **NaN values**: NaN values in the dataset had to be cleaned up accordingly. For instance NaN values in Numerical and categorical fields like agent, company ,country and children were filled with 0. NaN values in textual fields like company for instance were filled with ‘unknown’.
* **Country Map** :- As we have wide data ,to map the country with respect to the number of guests of Hotels was a little tricky task to match the parameters correctly. Finally after some trials , using the reset index command we got the Folium map.
* **APPROACH USED-**

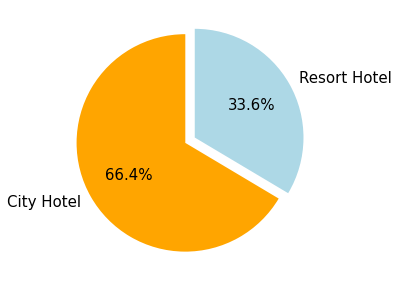
As the data includes 32 columns which are both numerical and categorical.To perform the analysis and different visualizations of factors associated with the hotel booking,we performed the pre-processing of data which could be removing NAN values or creating a new dataframe itself.

And, decided the factors which are important for analysis from the view of EDA.

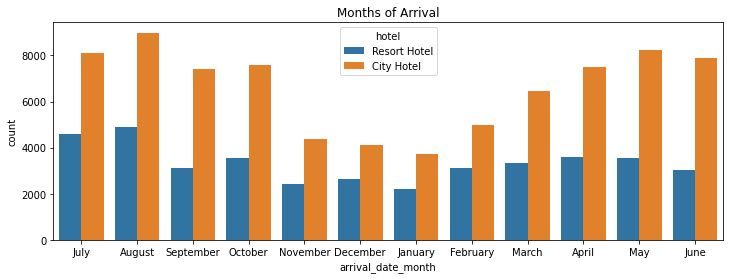
* **TOOLS USED-**
* **NumPy :-** NumPy is an open-source numerical Python library. NumPy contains a multi-dimensional array and matrix data structures. It can be utilized to perform a number of mathematical operations on arrays such as trigonometric, statistical, and algebraic routines.
* **Pandas :-** Pandas is a useful library in data analysis. It can be used to perform data manipulation and analysis. Pandas provide powerful and easy-to-use data structures, as well as the means to quickly perform operations on these structures.
* **Matplotlib :-**Matplotlib is a Python library that helps in visualizing and analyzing the data and helps in better understanding of the data with the help of graphical, pictorial visualizations that can be simulated using the matplotlib library.
* **Seaborn :-**Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with dataframes and the Pandas library. The graphs created can also be customized easily.
* **Plotly :-** The plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.
* **Missingno :-**Missingno is an excellent and simple to use Python library that provides a series of visualisations to understand the presence and distribution of missing data within a pandas dataframe. This can be in the form of either a barplot, matrix plot, heatmap, or a dendrogram.
* **ANALYSIS-**

**Exploratory Data Analysis-**

1] Percentage of Hotel Booking-

****

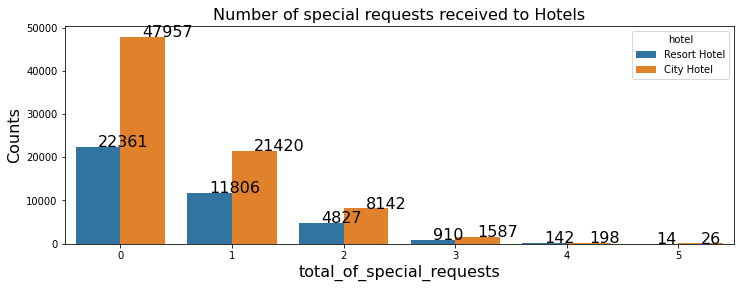
2] Best time to book a hotel room in a year-



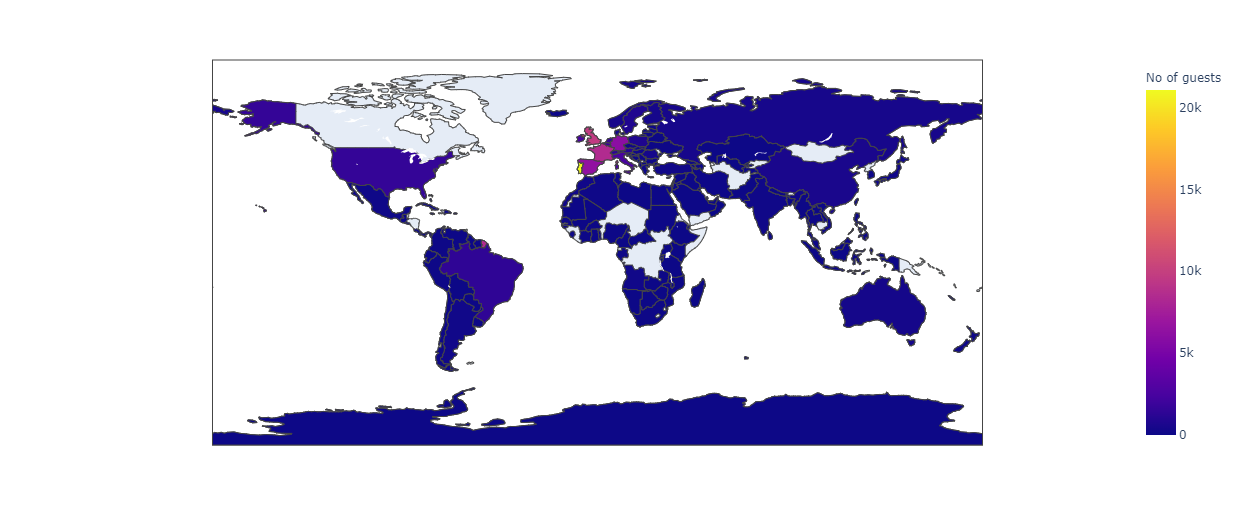
3] Price of a hotel room per night-

|  |
| --- |

4] Number of special requests received from customers-



5] Hotel Bookings from different countries**-**



**Descriptive Analysis-**

Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

* Count for Hotel type

| **City Hotel 79330**  **Resort Hotel 40060**  **Name: hotel, dtype: int6** |
| --- |

* Meal preference

| **BB 0.773180**  **HB 0.121141**  **SC 0.089203**  **Undefined 0.009791**  **FB 0.006684**  **Name: meal, dtype: float64** |
| --- |

* Correlation between the variables-

| **is\_canceled 1.000000**  **lead\_time 0.293123**  **total\_of\_special\_requests 0.234658**  **required\_car\_parking\_spaces 0.195498**  **booking\_changes 0.144381**  **previous\_cancellations 0.110133**  **is\_repeated\_guest 0.084793**  **adults 0.060017**  **previous\_bookings\_not\_canceled 0.057358**  **days\_in\_waiting\_list 0.054186**  **adr 0.047557**  **babies 0.032491**  **stays\_in\_week\_nights 0.024765**  **arrival\_date\_year 0.016660**  **arrival\_date\_week\_number 0.008148**  **arrival\_date\_day\_of\_month 0.006130**  **stays\_in\_weekend\_nights 0.001791**  **Name: is\_canceled, dtype: float64** |
| --- |

* **CONCLUSION-**

The Hotels ultimately want to increase their revenue and they would like to understand and concentrate on each and everything that improves their revenue.

The conclusions from my analysis of the data from 2015 to 2017 are :

From the Exploratory Data Analysis

* Most of the bookings are done for the City Hotel and the major revenue is from that as well. So the consumer can plan appropriate marketing strategies.
* The busiest month has been August. So the consumer can be prepared to expect more guests, make necessary arrangements, and do appropriate marketing. Also this helps them understand which months have the least bookings and can plan to give some deals to attract customers in the off season .We should also target months between May to Aug. Those are peak months due to the summer period.
* The average price daily rate of hotel room per person is high for City Hotel whereas low for Resort Hotel.This shows that hotel bookings are less in Resort Hotel,so they prefer to have a minimum rate.To overcome this , advertisement should be used by the Hotel marketing strategies.
* Cancellation rate decreases as the special requests are accomplished.When the customers special requests are fulfilled ,the guests are more likely to stay in the hotel.
* The most preferred meal type has been Bed & Breakfast. The consumer can be prepared to have more resources for meal time.
* We also realise that the high rate of cancellations can be due to high no deposit policies.
* Most visitors are from PRT. This helps the consumer to know his target customers and can attract them during their holiday seasons.
* To increase the stay at week nights ,hotels should acquire the necessities required by the customers.

**THANK YOU !!!!...**