Exploratory Data Analysis-

HOTEL BOOKING ANALYSIS

****

* **By Sanjay Verma**

**DataScience Trainee At AlmaBetter**

* **INTRODUCTION-**

Hotel Booking is the deal or agreement between the guests and Hotelier. Hotel Booking means the arrangement for people to stay for vacations or holidays by paying the rent of a room, food and other things.This specifies the bookings of hotel rooms as per the guests request and their period of time for staying at the place to enjoy the holidays.

For a guest, advance booking increases the chances of getting a nice deal from the customers and also availability of room on arrival time.

Hotel Booking process starts with the booking inquiry, checking the room availability, price per night, cancellations, lead time, meal preference,and among others.

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

* **OBJECTIVE-**
* To analyse the Suitable time for Hotel Bookings in a year.
* To study about stays and price variation over a year .
* To predict the special requests received from customers in proportionality.
* To analyse the results of Hotel Bookings from different countries.
* To study meal preference by customers 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**

In this EDA ,an analysis is done on a dataset which helps us to visualize the different factors in data which help us to conclude the different results as hotel booking times,destinations,rates ,etc.and also it will help the customers to interrelate the factors using graphs for better understability.

By this analysis we will conclude which hotels are more likely preferred by customers on account of their stays,meal,room rates and special requests.

* **DATA PREPPING-**

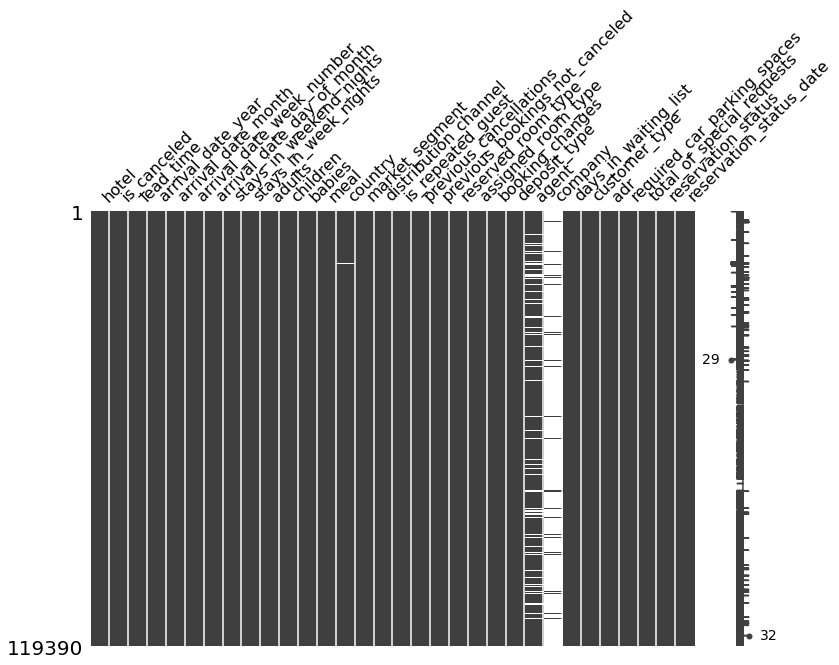
The dataset describes the data of hotel demand. There are two types of hotels in the dataset namely city hotel and resort hotel. The dataset contains numerical and categorical variables,32 variables describing the 119390 observations. In this dataset the observations represent the features of hotel booking,with respect to availability of resources.

The hotel bookings are seen from 1st july of 2015 till 31st august 2017 with the customers effectively arriving and cancellation of bookings.

As the dataset is real business data, it can be used for scientific and educational purposes and many 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-**

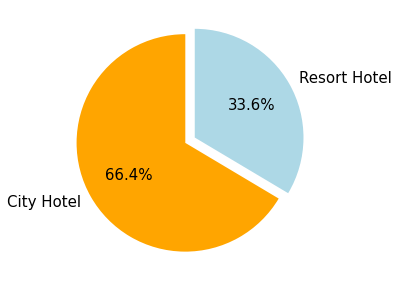
As the data includes 32 columns which are both numerical and categorical.To perform the analysis and different visualizations of factors associated with the hotel booking,we performed the pre-processing of data which could be removing NAN values or creating a new dataframe itself.

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

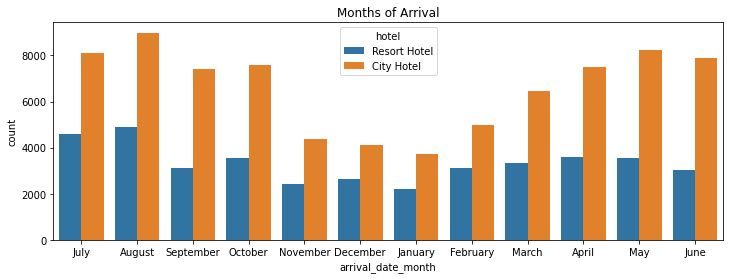
* **TOOLS USED-**
* **NumPy :-** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
* **Pandas :-** pandas is a [software library](https://en.wikipedia.org/wiki/Software_library) written for the [Python programming language](https://en.wikipedia.org/wiki/Python_(programming_language)) for data manipulation and [analysis](https://en.wikipedia.org/wiki/Data_analysis). In particular, it offers [data structures](https://en.wikipedia.org/wiki/Data_structure) and operations for manipulating numerical tables and [time series](https://en.wikipedia.org/wiki/Time_series).The name is derived from the term "[panel data](https://en.wikipedia.org/wiki/Panel_data)", an [econometrics](https://en.wikipedia.org/wiki/Econometrics) term for [data sets](https://en.wikipedia.org/wiki/Data_set) that include observations over multiple time periods for the same individuals.Its name is a play on the phrase "Python data analysis" itself.
* **Matplotlib :-**Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.
* **Seaborn :-** Seaborn is a library for making statistical graphics in Python. It builds on top of [matplotlib](https://matplotlib.org/) and integrates closely with [pandas](https://pandas.pydata.org/) data structures.Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.
* **Plotly :-** Plotly's Python graphing library makes interactive, publication-quality graphs. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, and bubble charts.
* **Missingno :-**Python has a library named missingno which provides a few graphs that let us visualize missing data from a different perspective. This can help us a lot in the handling of missing data. The missingno library is based on matplotlib hence all graphs generated by it'll be static. 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-

****

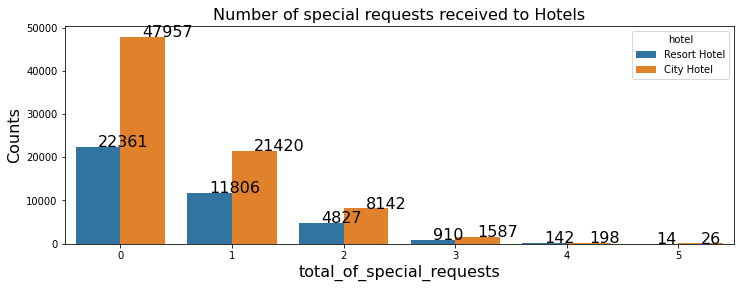
2] Best time to book a hotel room in a year-



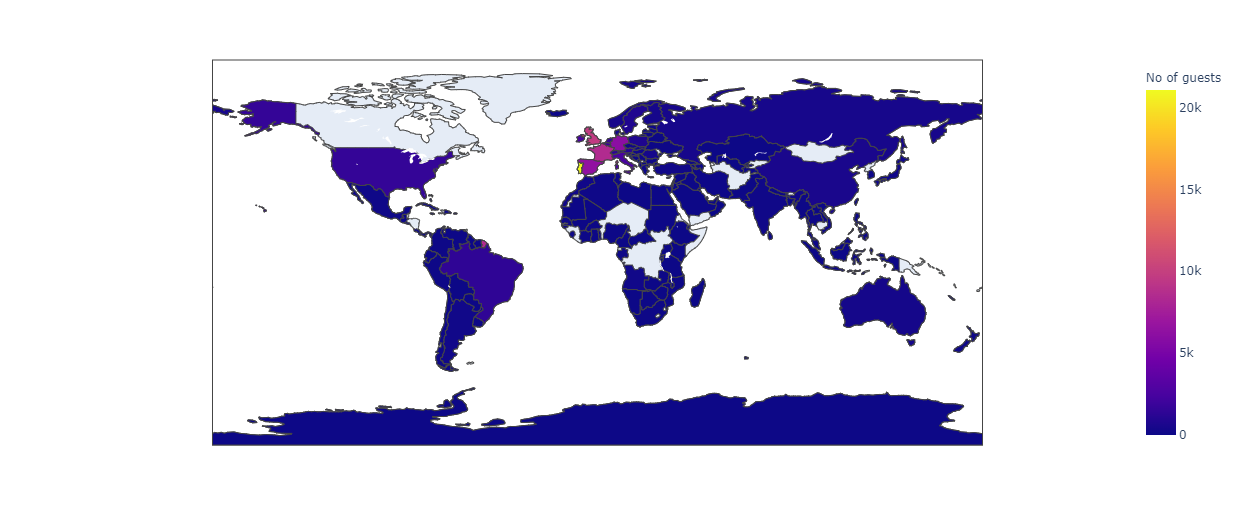
3] Price of a hotel room per night-

|  |
| --- |

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.

### the most busy months?

|  | **month** | **no of guests in resort** | **no of guest in city hotel** |
| --- | --- | --- | --- |
| **0** | January | 1868 | 2254 |
| **1** | February | 2308 | 3064 |
| **2** | March | 2573 | 4072 |
| **3** | April | 2550 | 4015 |
| **4** | May | 2535 | 4579 |
| **5** | June | 2038 | 4366 |
| **6** | July | 3137 | 4782 |
| **7** | August | 3257 | 5381 |
| **8** | September | 2102 | 4290 |
| **9** | October | 2577 | 4337 |
| **10** | November | 1976 | 2696 |
| **11** | December | 2017 | 2392 |

## Checking the price vary per night over the year**-**

|  | **arrival\_date\_month** | **adr** |
| --- | --- | --- |
| **0** | April | 75.867816 |
| **1** | August | 181.205892 |
| **2** | December | 68.322236 |
| **3** | February | 54.147478 |
| **4** | January | 48.708919 |
| **5** | July | 150.122528 |
| **6** | June | 107.921869 |
| **7** | March | 57.012487 |
| **8** | May | 76.657558 |
| **9** | November | 48.681640 |
| **10** | October | 61.727505 |
| **11** | September | 96.416860 |

* 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 Hoteliers aim is to increase the bookings and revenue for the business which directly focus on the features to understand and get the improvement done for necessities.

Exploratory Data Analysis showed us following results-

* City Hotel has more than 60% of hotel bookings compared to Resort Hotel .This shows that , by using proper plan market strategies revenue may be increase for Resort Hotel.
* The busiest month has been July and August. To increase the bookings in other months, consumers can prepare for innovative ideas to approach more customers.
* The average price daily rate of hotel room per person is maximum for City Hotel whereas minimum for Resort Hotel.As bookings are less in Resort hotel,the rate is preferred as minimum with respect to rate of City Hotels.
* When the customers' special requests are fulfilled ,the guests are more likely to stay in the hotel.Cancellation rate decreases as the special requests are increases (fulfilled).
* The most preferred meal type has been Bed & Breakfast.
* We also realise that the high rate of cancellations can be due to high no deposit policies.
* Most guests arrived from PRT.
* Guests stay at a hotel for an average of 5 nights. Hoteliers may increase the guests' stay by accomplishing the necessities and amenities of the hotel as demanded.

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