**HOTEL BOOKING EDA CAPSTONE PROJECT**

**Completed by:**

**Ankush Kumar**

**Manoj K B**

**Nayan Kumar Jha**

**Data Science Trainees @ AlmaBetter, Bangalore**

**Abstract:**

This project consists of hotel booking datasets of two different types of hotels those are 1. City hotels and 2. Resort hotel from year 2015 till 2017, Including details such as bookings, cancellations, and guest information. The project's main goal is to comprehend and visualize data from the hotel and customer perspectives to get the proper insights from the dataset and to make proper data driven decisions

**Problem Statement / Analysis Direction:**

Our project consisted of hotel booking dataset and after we read the data in several aspects first taking from the hotel manager point of view there were several features, we have seen will affect the cancellation of the booked ticket (So we first took is canceled feature as our target variable) and also from the customers point of view there was certain delay after they booked their room (Sometime there was huge waiting time). And there were many other features which also affect the booking of the hotel rooms.

**Introduction**

This data set contains booking information for a city hotel and a resort hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has from the data.

We will perform exploratory data analysis with python to get insight from the data.

**Steps For Exploratory Data Analysis (EDA)**

1. **Data information:**

Collected various information about the data by reading the data from csv file name as “Hotel Booking”.

We get to know there are total 32 columns, in which 12 columns are numeric and remaining 10 are categorical columns. And our targeg/label class column is “is\_canceled” which contains 0 , 1 binary values in which 1 -> Cancelled Bookin and 0 -> Booking not Cancelled.

There are total 119390 enteries for each columns but also it some columns have missing data i.e. Null/NaN values.

1. **Data Cleaning and Removal of Duplicacy:**

Detected various columns with missing values ['children', 'country', 'agent', 'company']

Out of Total 119390 rows ,'agent' and ''company' column contains very large amout of Null/NaN values in it i.e. 16340 and 112593 and this much of Null/NaN values will impact our analysis.We have Dropped these columns as we haven’t used these columns for our analysis.

And 488 in 'country' and 4 in 'children' column contains Null/NaN values and which is very less so we have just deleted these missing value rows from our dataset.

After that we have checked for data duplicacy in which we have found that there are 31984 duplicate enteries.and we have just deleted these duplicate enteries.

1. **Data Visualization**

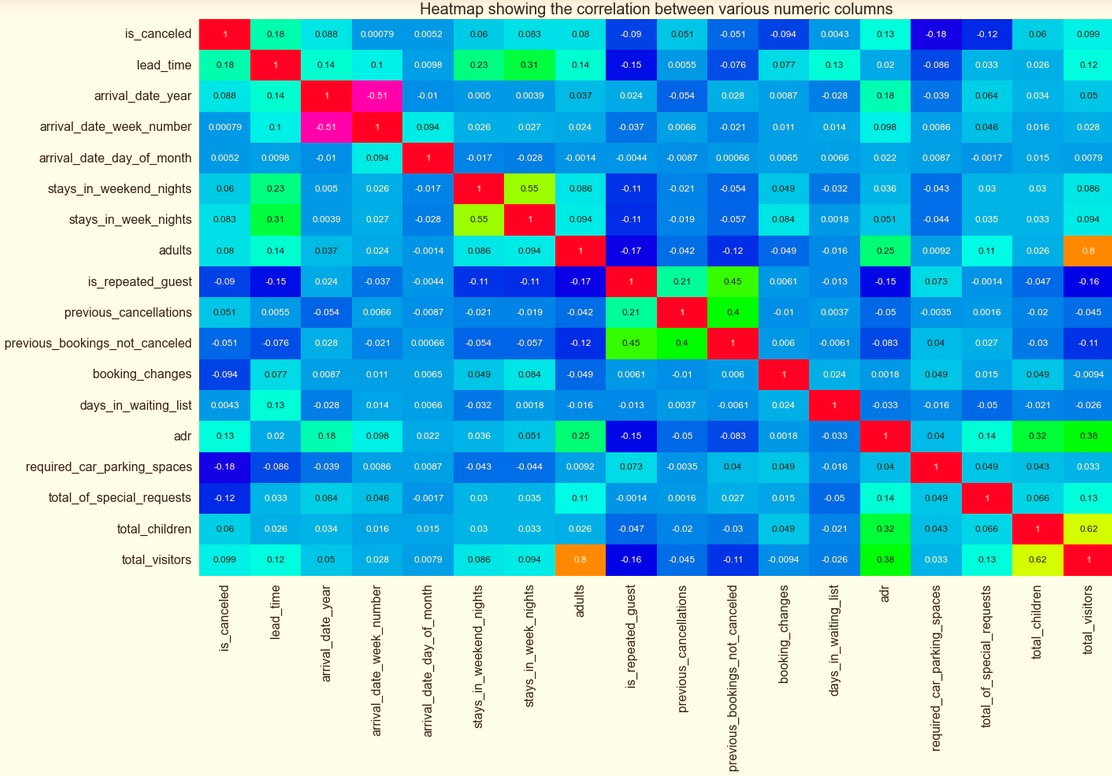
We have performed different types of visualization on the basis of Univariate, Bivariate and multivariate analysis.

Firstly we have conducted Univariate analysis as we also need to understand various features/columns individually that what kind of importance and insights they brings for our analysis.

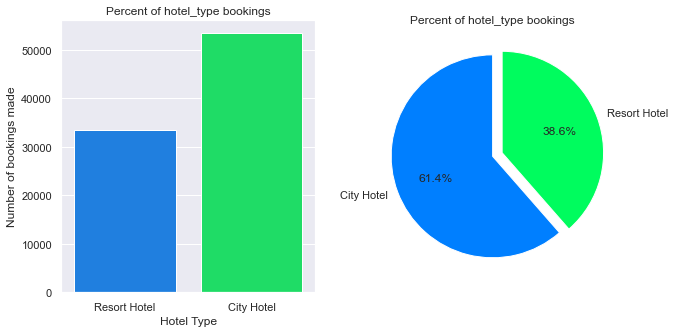
Secondly we have performed Bivariate analysis so that we can analyse the impact of one column/feature to another feature and where these insights leads us.

At last we have performed Multivariate analysis in which we came to know the impact of multiple features to the booking occurrence of hotels.

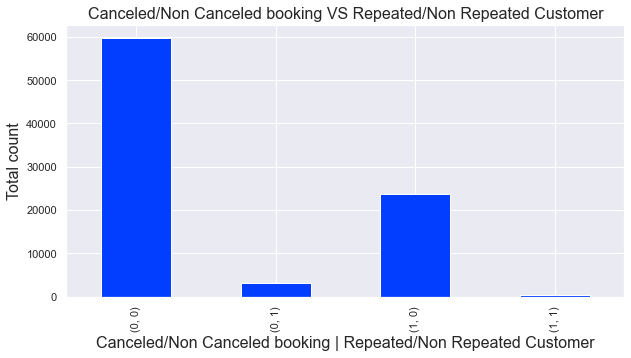
1. **Correlation**

****As we know that correlation is a very important observation task to get the proper idea that how our features are idependent from each other and if they are not independent and are related to each other then there will be lots of abiguity for our model creations that misleads our analysis.

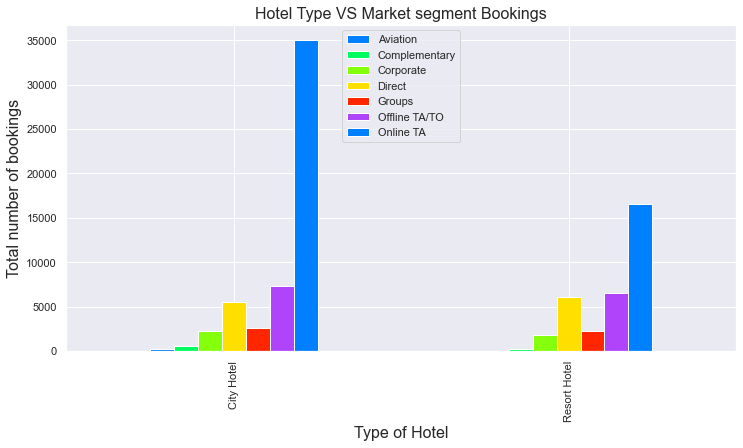
* **Some glimpses we have gathered from our various visualtization analysis:**

****

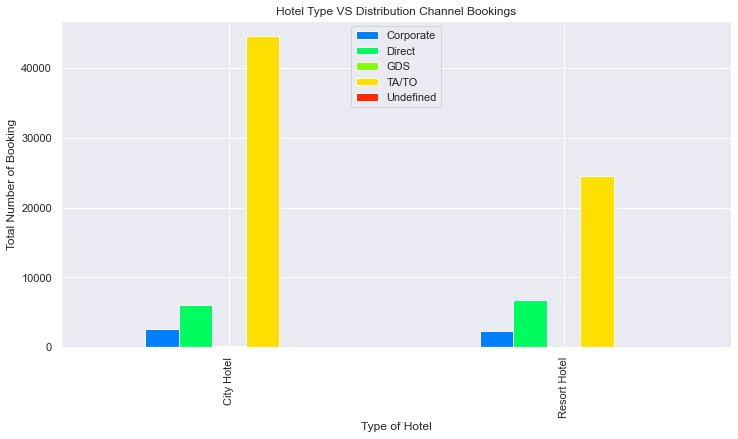
* City hotel has bookings above 50000 which is around 61.4% and Resort hotel has less than 35000 which is around 38.6% bookings.

****

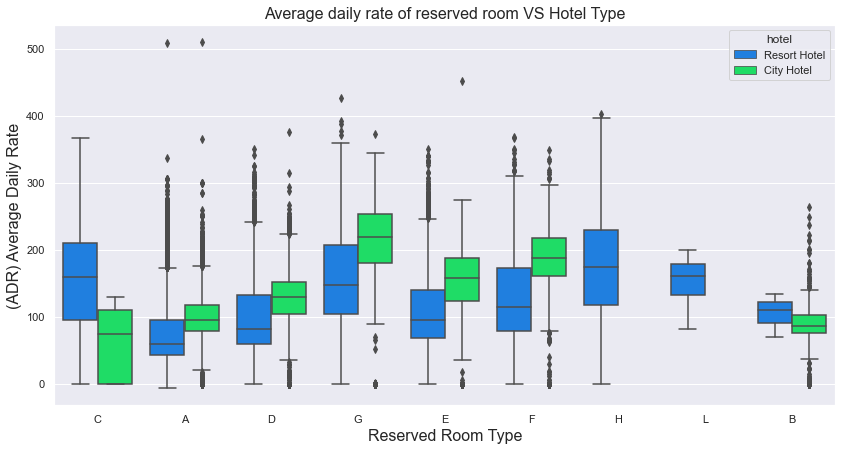
* The lowest cancellation occurs when the customer is repeated.
* That means the existing customers are very less likely to cancel the booking



* Maximum number of booking is done by the Online TA(Travel Agent) for the City Hotels as well as for the Resort Hotels.



* The most number of booking is done by the TA/TO Distribution Channels for both City and Resort Hotel.



* Room type 'G' has the highest and Room type 'C' has the lowest Average Daily Rates (ADR) in City Hotels.
* Room type 'C' and 'H' has the highest and Room type 'A' has the lowest Average Daily Rates (ADR) in Resort Hotels.

**List of Question to help achieving the goal:**

1. Do customers have any preference with or without children's
2. How daily average rate is impacting the reserved room type in hotels
3. Does total stay per month impacted by customers with and without children.
4. Do reservation status impacted by type of customers
5. Do customers cancel their booking if they are allotted with different room type.

**We got very intuitive results by performing analysis on the above questions:**

* The count for stays with and without children is pretty similar in both circumstances.
* Customers who stayed with and without children for 1-7 days have the same count.
* If we consider the usual stay, there will be three to four days during which people will have their children with them, otherwise only adults will be there.
* The majority of guests have no need for parking.
* The Online TA customer group has the highest number of hotel reservations in the city and at resort hotels, the number of customers who booked directly is the highest, however, it is clear that the majority of the appointments are made by consumers of the Online TA group.
* The impact of room type allocation on clients is minimal.

**Conclusion and Summary**

* The vast majority of reservations are for hotels in cities. Resort hotels have fewer cancellations than city hotels. The aviation industry has the shortest wait time.
* The months of August, July, and May saw the most hotel bookings in the city. The lowest cancellation rate occurs when a consumer is repeated. The lead time for July is quite long, whereas the lead time for January and February is extremely short. In July, August, and September, the average daily rate for Resort Hotels is higher.
* In June, July, and May, the average daily rate for City Hotels is higher. Customers travelling with children have little preference for the type of hotel they stay in.
* In both cases, the number of stays with and without children is rather equal. Customers who stayed for 1-7 days with or without children had the same count.
* If we take the average stay, three to four days will be spent with youngsters, while the rest of the time will be spent with adults solely.
* Online TA customer group has the highest number of hotel reservations in the city and at resort hotels. As a result, the impact of room type allocation on clients is minimal. The majority of appointments are made by consumers of the Online TA group who book directly through the website

**Future work recommendation:**

In future we are planning to apply some ML algorithm for doing prediction of waiting time in particular area and hotel as well as to apply the reviews sentiment analysis for booking the hotel as our comfort. We are planning to use several classifiers to evaluate our model and give customers best experience while booking hotel tickets.

**References:**

* Towardsdatascience
* Analyticsvidya
* Becominghumanai
* Siteminder.com
* Tmstudies.net