**Hotel Booking Analysis**

**Shashank Mishra**

**Y Ishwar Rao**

**Ganesh Patil**

**Pradeep Kumar Yadav**

**Data science trainees,**

**Almabetter, Bangalore**

**Abstract:**

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.

***Keywords: Exploratory data analysis, Hotel booking analysis***

**Introduction:**

Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week and many more. This makes analyzing the patterns available in the past data more important to help the hotels plan better. Using the historical data, hotels can perform various campaigns to boost the business. We can do EDA to predict the future bookings, most engaged months of coming year, additional facilities which can attract more customers and based upon the data, we can raise the revenue.

**Problem Statement:**

**1.Percentage of bookings in each hotels?**

* Percentage of bookings in each hotels-

Two Types of hotels:

1.City Hotel

2.Resort Hotel

Around 60% bookings are for City hotel and 40% bookings are for Resort hotel.

**2.Which hotel makes more revenue?**

Avg adr of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue.

* Now in the hotel Revenue we created The Bar Graph to show the which hotel makes more revenue.

**3.Which hotel has higher lead time?**

City hotel has slightly higher median lead time. Also median lead time is significantly higher in each case, this means customers generally plan their hotel visits way to early

* In This we calculate the average of waiting period and show in bar graph.

**4.Which significant distribution channel has highest cancellation percentage?**

TA/TO has highest booking cancellation %. Therefore, a booking via TA/TO is 30% likely to get cancelled

Not getting same room as demanded is not the case of cancellation of rooms. A significant percentage of bookings are not cancelled even after getting different room as demanded.

**5.Which channel is mostly used for early booking of hotels?**

Channel which is mostly used for early booking of hotels is also TA/TO.

GDS channel brings higher revenue generating deals for City hotel, in contrast to that most bookings come via TA/TO. City Hotel can work to increase outreach on GDS channels to get more higher revenue generating deals

**6.What are the most busy months for hotels?**

From the month of July to August the number of bookings increased and in August, City Hotel got most number of guests

* Now We use the Scatter Plot to show the number of bookings increased and in August

**7.In which months hotels charges higher adr?**

The revenue aspect looks different, the Resort Hotels receives more revenue with respect to City Hotel. From May to August there was rapid increase in adr. August recorded the highest.

* Here also we used scatter plot to determine there was rapid increase

in adr. August record the highest

**8.How does booking numbers and adr changes within a month?**

We can see that graph Arrival\_num has small peaks at regular interval of days. This can be due to increase in arrival weekend.

Also, the avg adr tends to go up as month ends. Therefore charges are more at the end of month.

**9.How does bookings varies along year for different types of customers.**

Mostly bookings are done by couples. It is clear from graph that there is a sudden surge in arrival num of couples and family in months of July and August.

So better plans can be planned accordingly at that time for these type of customers.

**Steps involved:**

* **Exploratory Data Analysis**

After loading the dataset we performed this method by comparing our target variable that is Hotel booking analysis with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

* **Null values Treatment**

Our dataset contains a large number of null values which might tend to disturb our accuracy hence we replaced them at the beginning of our project in order to get a better result.

## **Check for the wrong entry**

Our dataset contains a there are 3 categories of guests which are children, adults and babies**.**

**If hotel is booked than there should be any entries in children, adults or babies’ columns, it should not zero at a time. Hence we drop them at the beginning if our project in order to get a better result.**

* **Data analysis and Data visualisation**

### Here we do data analysis and data visualization by using matplotlib and seaborn to the graph for better understanding of the insights.

**Conclusion:**

That's all there is to it! We have reached the end of our exercise. The data has been loaded, null values have been treated, categorical columns encoded, and major reasons that govern hotel bookings have been identified, along with steps to increase them.

* Around 60% bookings are for city hotels and around 40% bookings are for resort hotels therefore city hotel is busier than resort hotel ,also overall adr of city hotel is slightly higher than resort hotel.
* The majority of reservations are for city hotels.
* The number of repeated guests is too low.
* Most of the bookings either in the canceled or checkout done by online TA.
* City hotels and resort hotels maximum number of bookings by online TA.
* That aviation industry has the minimum number of days on the waiting list.
* More visitors are from western europe, namely Portugal,France, Great Britain, and Spain being the highest.
* Families with children have no particular preference for the hotel type.

## **Challenges:**

**(1) There was a lot of duplicate data.**

**(2) Data was present in wrong datatype format.**

**(3) Choosing appropriate visualization techniques to use was difficult.**

**(4) A lot of null values were there in the dataset.**

**References-**

* Pandas user guide: <https://pandas.pydata.org/docs/user_guide/index.html>
* Matplotlib user guide: <https://matplotlib.org/3.3.1/users/index.html>
* Seaborn user guide & tutorial: <https://seaborn.pydata.org/tutorial.html>
* Tutorials: 1) [https://stackoverflow.com](https://stackoverflow.com/) , 2) [https://www.w3schools.com](https://www.w3schools.com/)