**CAPSTONE PROJECT**

**EDA ON HOTEL BOOKING DATA**

**Shubham Kishorrao Joshi,**

**Akshay Fasale,**

**Kanika Kakra,**

**Rishikesh Dhamale.**

**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 use the patterns to predict the future bookings using time series or decision trees.

We have booking data for two types of hotels:

🡪City Hotel

🡪Resort Hotel

In the data we have different features provided such as (Repeated guests or not, Average daily rate, Etc.).

The main objective of the analysis is to obtain the meaning full information and facts from the given huge datasets as shown above, by cleaning the datasets, doing a proper analysis and visualization and plotting the useful information into different graph and charts so that the trend and relationship between the various indicators on which the analysis is done can be understand easily.

***Keywords: Exploratory data analysis.***

**1.Problem Statement:**

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to know that a hotel has disproportionately high number of special requests? This hotel booking dataset can help you explore those questions!

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.

The main objective is to Explore and analyze the data to discover important factors that govern the bookings.

**The Data features** present in the given data are as follows:

* **Hotel**: hotel type (H1 = Resort Hotel or H2 = City Hotel)
* **Is canceled**: Value indicating if the booking was canceled (1) or not (0)
* **Lead time**: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
* **Arrival date year**: Year of arrival date
* **Arrival date month**: Month of arrival date
* **Arrival date week number**: Week number of year for arrival date
* **Arrival date day of month**: Day of arrival date
* **Stays in weekend nights**: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
* **Stays in weeknights**: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
* **adults**: Number of adults
* **children**: Number of children
* **babies**: Number of babies
* **meal**: Type of meal booked.
* **country**: Country of origin. Categories are represented in the ISO 3155–3:2013 format.
* **Market segment**: Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
* **Distribution channel**: Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
* **Is repeated guest**: Value indicating if the booking name was from a repeated guest (1) or not (0)
* **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.
* **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.
* **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. ID is presented instead of designation for anonymity reasons.
* **Days in waiting list**: Number of days the booking was in the waiting list before it was confirmed to the customer.
* **Customer type**: Type of booking, assuming one of four categories: 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.
* **Average daily rate**: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights.
* **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 (EX. twin bed or high floor)
* **reservation status**: Reservation last status, assuming one of three categories: Canceled, Check-Out, No-Show.
* **Reservation status date**: Date at which the last status was set.

**Steps involved:**

**1.Importing The essential libraries:**

We use python environment for our analysis, and it has some essential libraries which helps in analysis.

We are importing following libraries

For their respective applications:

* **Pandas**: - Pandas is used to analyze data. It has functions for analyzing, cleaning, exploring, and manipulating data.
* **Matplotlib**: - Matplotlib is a graph plotting library in python that serves as a visualization utility. Most of the Matplotlib utilities lies under the py-plot submodule.
* **Seaborn**: - Seaborn is a library for making statistical graphics in Python. It builds on top of Matplotlib and integrates closely with Panda’s data structures.

**2.Loading the Data:**

Now, we need to load our data from the external source, which in this case is uploaded to the drive. Also, we would try to spot the nature and properties of the data that we have. We will use read\_cvs(‘Path of the file ) function to load the data.

**3.Data Cleaning:**

Data cleaning is an important step in the data analytics process in which you either remove or update information that is incomplete or improperly formatted.

There are null values present in 4 column (Features) **Country, Children, Agent, Company**. We can’t analyze the data with null values present. To make data correct we replaced all the null values from agent and company column with zero. Null values in children column are replaced by mean of children in the data. We replaced null values in country column with ‘others.

We changed data type of children, company, agent to int type.

added some new columns dropped some existing columns to make analysis easy.

Created new column Full\_stay by adding stays\_in\_weekend\_nights + stays\_in\_week\_nights.

Created new column total\_members by adding adults+children+babies

* + **Removing outliers:**

One outlier was found in the Average daily rate column. Simply dropped it.

**4. Selecting Indicators:**

From the above list of data features present in the data we are selecting some crucial features which are necessary for analysis part.

we are considering all the features except ‘Company’.

**5.Data Wrangling:**

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. With the amount of data and data sources rapidly growing and expanding, it is getting increasingly essential for large amounts of available data to be organized for analysis.

**6.Data Visualization:**

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

* **Plots we used in visualization:**

Mainly performed using Matplotlib and Seaborn library and the following graph and plots had been used:

* **Bar Plot:** A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.
* **Scatter Plot:** A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data
* **Pie Chart:** A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.
* **Line Plot**: A line chart or line plot or line graph or curve chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.
* **Heatmap:** A heat map is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.
* **Box Plot:** In descriptive statistics, a box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.

**7.Exploring key indicators:**

**Overall Bookings:** It representshow much booking was made in the given data.

City hotel has a greater number of bookings in every year than resort hotel. For both the hotels may, June, July, august were the busiest months over the years.

**Stay Length:** Most common stay length is less than 4 days and generally people prefer City hotel for short stay, but for long stays, Resort Hotel is preferred.

**Average daily rate:** For resort hotels, the average daily rate is more expensive during august, July and September. For city hotels, the average daily rate is more expensive during august, July, June and may.

**Cancellations**: Cancelation rate is higher in city hotel.  most cancellations are also in 'no deposit' type bookings. With more lead time the chance of cancellation also increases

**Is repeated guest:** Both the hotels combined only have 4% of repeated guests. Both the hotels have very low customer retention rate.

**Country**: Most of the customers come from Portugal and western European countries. The hotel should target this region for their marketing campaigns

1. **. Challenges:**

* There was a lot of duplicate data.
* Data was present in wrong datatype format.
* Choosing appropriate visualization techniques to use was difficult.
* A lot of null values were there in the dataset.
* Needed to remove outliers.

**9. Conclusion:**

That's it! We reached the end of our exercise.

Starting with loading the data so far, we have done EDA, null values treatment, encoding of categorical columns, feature selection and Visualization. From the analysis we can conclude:

* Most guests come from western Europe countries. We should spend a significant amount of our budget on those area.
* most reservations are for city hotels.
* Majority of the hotels booked are city hotel. Need to spend the most targeting fund on those hotels.
* We also realise that the high rate of cancellations can be due high no deposit policies.
* We should also target months between May to Aug. Those are peak months due to the summer period.
* Given that we do not have repeated guests, we should target our advertisement on guests to increase returning guests.
* 80% distribution channel is TA/TO
* Most common stay length is less than 4 days and generally people prefer City hotel for short stay, but for long stays, Resort Hotel is preferred.
* November, December, February and January are the months which has less booking so in this period you can get rooms with less average daily rate. And avoid busiest months for hotels (May, June, July, August).

**References-**

* **Python Pandas Documentation** <https://pandas.pydata.org/pandas-docs/stable>
* **Python Matplotlib Documentation** <https://matplotlib.org/stable/index.html>
* **Our EDA Project Documentation**