EDA

Hotel Booking Analysis

# Summary:

The main aim of this project is to perform an Explaratory Data Analysis on hotel booking analysis which investigates the different booking patterns and the customer behaviour and suggests measures that can be implemented to increase the revenue.

The project contains booking information for a city hotel and resort hotel, such as the booking date, month,year,customer information such as number of adults, children and babies.As the first step I explored the data and inspected over the duplicate values and the missing values.Then I cleaned the dataset by handling the null values and outliers and dropping the duplicate values and irrelevant data.Then it was visualized using the graphical techniques.

EDA is divided into univariate, bivariate and multi variate analysis. In univariate, data is analysed on only one variable. For example, the most preferred hotel by the customers, the busiest month, the year with most number of bookings etc.

In bivariate and mutivariate two or more than two variables are compared with each other.For example, the hotel with the highest number of cancellations are taking place, the month in which most cancellations are taking place, the number of repeated guests for each hotel etc.

From these analysis we can take measures to reduce the number of cancellations and to improve the customer service. By taking these measures we can make a revenue increase.

# 1. Problem Statement

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 been removed from the data. Explore and analyse the data to discover important factors that govern the bookings.

# Introduction

The purpose of this exploratory data analysis was to explore the hotel booking data set and identify potential relationships between key variables.The project aims to gain the interesting insight into the different factors that affect the booking patterns and cancellations and to take measures to enhance the quality and customer satisfaction, thus by increasing the revenue.

# 3.Procedure

Importing Libraries

Relevant libraries like NumPy for numerical operations, Pandas for data manipulation, matplotlib and seaborn for data visualization were loaded.

Reading Data

After drive was mounted, data from csv file was read and store in a pandas data frame.

Data inspection

We examine the dataset's shape and columns after importing it. The dataset contains 119390 rows and  32 columns. We check variables and their datatypes for null values using the info() function. We know the basic description of each variable by using the describe() function.

Variables description

Hotel --> H1= Resort Hotel ,H2=City Hotel

is\_cancelled --> If the booking was cancelled(1) or not(O)

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 for arrival date

arrival\_dat\_day -->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\_week\_nights--> Number of week nights (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-->Kind of meal opted for

country--> Country code

market\_segment-->Which segment the customer belongs to

Distribution \_channel--> How the customer accessed the stay- corporate booking/ Direct/TA.TO

is\_repeated\_guest-->Guest coming for first time or not

previous\_cancellation--> Was there a cancellation before

previous\_bookings -->Count of previous bookings

reserved\_room\_type--> Type of room reserved

assigned\_room\_type--> Type of room assigned

booking\_changes--> Count of changes made to booking

deposit\_type--> Deposit type

agent -->Booked through agent

days\_in\_waiting\_list--> Number of days in waiting list

customer\_type-->Type of customer

required\_car\_parking-->If car parking is required

total\_of\_special\_req--> Number of additional special requirements

reservation\_status-->Reservation of status

reservation\_status\_date-->Date of the specific status

Handling duplicate values

The dataset contains 31994 duplicate values. After dropping the duplicate values the columns are reduced to 87396.

Handling missing values and outliers

Total number of null values in the dataset is  87396.The columns children contains 4, country contains 452, agent contains 12193 and company contains 82137 null values.All the missing values are removed using the values '0' and 'others'. An outlier in the adr column has dropped.

The astype() function is used to change some variables into the correct datatype because they are not in the correct datatype by default.Thus the dataset is cleaned and ready for wrangling.

Explaratory Data Analysis

EDA is divided into univariate, bivariate and multi variate analysis. In univariate, data is analysed on only one variable.

In bivariate and multivariate, two or more than two variables are compared with each other.

Data analysis is performed to answer the following questions:

* Which month is the busiest month for hotels?
* Which hotel is more preferred among travelers?
* In which year most bookings happened for each type of hotels?
* Which agent made the most number of bookings?
* Which type of meal is the most preferred  by the customers?
* What is the booking cancellation percentage?
* What type of customers are coming to each hotel?
* From which country the most customers are coming?
* Whether the customers are repeatedly coming or not?
* Which is the most preferred room type for different types of customers?
* Whether the customer got the reserved room type?
* Which is the most used distribution channel for booking?
* Which Hotel has the highest revenue?
* Which hotel has the highest lead time?
* Which hotel has the history of high number of booking cancellation?
* Which hotel has the history of most repeated customers?
* Which distribution channel generates more income?
* How long people stay in hotels?
* Which hotel has longer waiting time?
* How is average daily rate related to total number of people?
* How many solo travellers, couple or family travelled  per month?

Data Visualization

It involves converting tabular data into graphic images so that users may quickly and easily understand it.

The graphs used are :

* Pie chart
* Histogram
* Count plot
* Bar plot
* Line plot
* Scatter plot
* Heat map
* Pair plot

# 4.Conclusion:

The following conclusions were drawn from analysis:

* 61% customers preferred city hotels. 39% customers preferred resort hotels.
* August and July are the most income generating months. January, November and December are the slowest months for both hotels
* Highest number of bookings are made in the year 2016.Least number of bookings are done in 2015.
* Most income generating agent is 9 followed by 240 and 0.
* Most preferred meal type is BB(Bed & Breakfast)
* 72% customers not cancelled the booking.28% customers cancelled the bookings.
* Majority of the customers are Transient, followed by Transient-Party for both city and resort hotels.Minority is the customers in Group.
* Majority of the customers from Portugal.
* 96% customers new customers. Only 4% is the repeated customers.
* Most customers preferred the 'A' type room.
* Most customers got the 'A' type room.
* Most number of bookings are done through the GDS distribution channel.So the most income generated distribution channel is GDS.
* Higher revenue generated hotel is city, compared to resort hotel.
* Most profittable month for Resort hotel is August and for city hotel is May and August.
* Resort hotel has higher lead time comapared to city hotel.
* Most number of booking cancellations are in city hotels.
* Most number of cancellations take place in the month of August.
* Repeated customers mostly prefer the resort hotels.
* Most revenue generated market segment is Online TA.
* City hotel is busier than resort hotel.
* Most people stayed not more than 3 days.
* When the number of people increases the adr also increases.
* In the month of August most customers are couples. In July and August most of the customers are family.
* There is a positive corelation between is\_canceled and days\_in\_waitinglist.
* There is a strong negative relation between totalof\_special\_requests and is\_repeated guest.
* Negative corelation between is\_repeated guest and adr

**Project done by**

**SAJEENA T R**