Hotel Booking Analysis

Rakesh Sahoo , Lubna Zarin
Mahesh Patki

ABSTRACT

The Hotel booking is one of the largest and most popular Android apps in stores. It has anenormous amount of data that can be used to make an optimal model. We have used hotel booking data from the team capstone project dashboard. Thisdata set contains 32 different features that can be used for predicting key factors responsible for Analysis engagement & success stories.

INTRODUCTION

Booking cancellations have a substantial impact in demand management decisions in the hospitality industry. Cancellations limit the ability to make accurate forecasts which is a critical tool in terms of revenue management. To overcome the problems caused by booking cancellations, hotels implement rigid cancellation policies & overbooking strategies, which can also have a negative influence on revenue & goodwill. Using data of world's leading chain of hotels, homes & spaces and addressing booking cancellation prediction as a classification problem in the scope of data science, we in this model try to predict with higher accuracy whether a booking will be cancelled. Using supervised machine learning techniques, a viable model is being created to predict hotel cancellations that further allows organization to look into its cancellation policies and overbooking strategies. It will have a positive impact on the revenue and profitability of the business.

INTEGRAL METHODOLOGY

The entire Analysis is divided into the following phases: Dataset Description, Breakdown of Datasets, Examining the null values & missing values, Data Cleaning, followed by Exploratory Data Analysis by and applying different models. First, we collect the data from Alma's better dashboard. Thereafter we did basic data cleaning and data visualization. After visualizing the data set, we removed some unnecessary features andmade it ready for analyzing the data set using different plots. Next, we conduct data modelling by using Bar plot graphs, violin plots, density plots, etc. Finally, we narrate the analysis results to provide a clear vision of the relationship among the areas of interest

DATASET DESCRIPTION

Let's take a look at the data which is provided in Dataset Hotel booking.csv file

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.

BREAKDOWN OF DATASETS

In order to go ahead for data visualization upon key factors we need to go for certain extra steps before proceeding to the main segment. In this part we are going with the following steps:

1. Importing Analytical necessary library classes for future analysis.

   Reading the csv data file from Googledrive.
2. Setting figure size for futurevisualization.
3. Removing future warnings in seaborneplots.
4. Visualizing all the columns of the respective Data frame.
5. Viewing all data information
6. Checking the Unique values in the column (if any)
7. Converting the data types to similarobjects as the Analysis Demands.
8. Removing the duplicate values, removing null values and refining data

EXAMINING NULL VALUES

The most critical thing from which we can draw some observations is Dataset, however data comes with unexpected values too i.e. sometimes it may be Null or missing in other words the space might be blank. Thus, at thetime of analyzing the first thing which we will do . is to examine the null or missing values onthe Dataset.

It is the first step that will make the results "more" accurate & should be handled before it affects the performance of the models that predict the outcome. By getting information Using info method it can be seen that missing values are more we have seen from above there are 32 columns and 119390 rows are present.

But there are columns like children, country, agent, company contains null values. Hence, several methods to eradicate those null values.

DATA CLEANING

Cleaning data is crucial step before EDA as it will remove the ambiguous data that can affect the outcome of EDA.

While cleaning data we will perform following steps:

1)Remove duplicate rows:

- In the dataset we first need to remove duplicate
- Data by using drop_duplicate() method

2) Handling missing values:

- We have identified four columns having null values children, country , agent , company.
- 'agent' and 'company' contains Nan values for the bookings for which the booking was done directly by customer. Hence we can replace them with 0 as datatype is float.
- 'country' contains Nan values and can be replaced with 'Others' as the datatype is string.
- 'children' contains only 5 Nan values, so it might happen the hotel booking entries are not recorded properly due to human error. Hence we will consider the mean value of 'children' to replace the Nan.

3) Convert columns to appropriate Datatypes:

- Converting datatype of columns children, company
  and agent from float to int.
- changing datatype of column 'reservation status_date to data_type

4) Adding important columns:

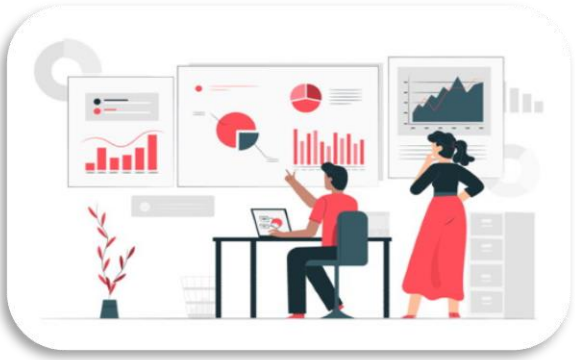
- Adding total staying days in hotels
- Adding total people num as column, i.e. total people

  num = num of adults +children+ babies

DATA VISUALIZATIONS

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.

Data visualization can be utilized for a variety of purposes, and it's important to note that is not only reserved for use by data teams. Management also leverages it to convey organizational structure and hierarchy while data analysts and data scientists use it to discover and explain patterns and trends.

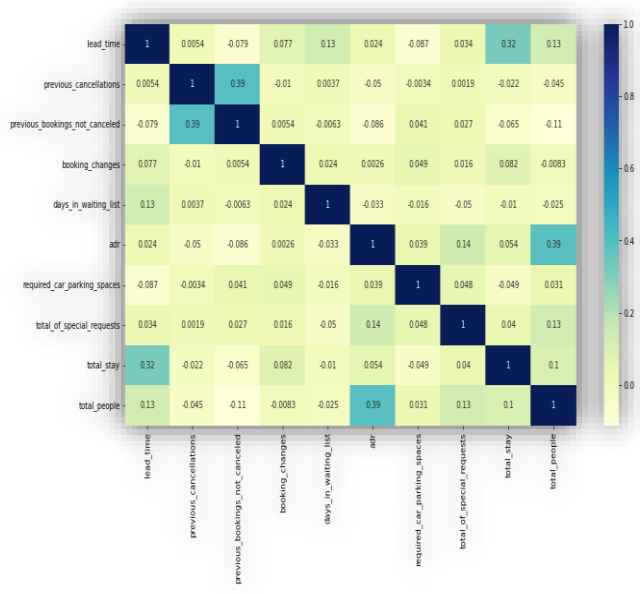


Observation-1

The best time of year to book a hotel room :

1)From corelation we can say when total peoples is greater then the price of average daily price is high

So this period of time is not good for booking



1 Fig of Relations

First the correlation between the numerical data

1-'total_people' and 'adr' has slight correlation. It means more number of people means, more adr which in turns more revenue.

2-'lead_time' and 'total_stay' has slight correlation, which indicates usually people with longer stay plans earlier than the actual arrival.

Observation-2

Lets see does length of stay affects the adr.

1)The pricing of hotel is also depending of day of stay for example- If adr for one day stay 2000rs if we stay 2 days in hotel then its 1600 and similarly if we stay 3 days then the adr is lowest means 1500rs

So this is best length of staying duration

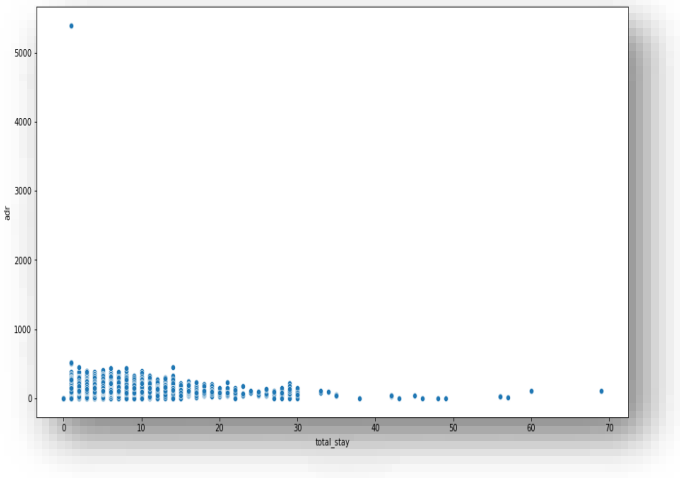
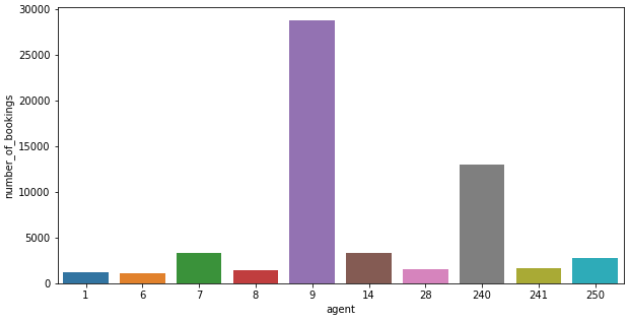


Fig of length of staying and adr

From the scatter plot we can see that as length of 'total_stay' increases the 'adr' decreases. This means for longer stay, the better deal for customer can be finalised.

Observation-3:

Univariate Analysis



Which agent makes most no. of bookings

AS shown in fig number of booking and agent are plotted here we can see agent no. 9 makes maximum bookings.

Observation-4

room type is in most demand and which room type generates highest adr

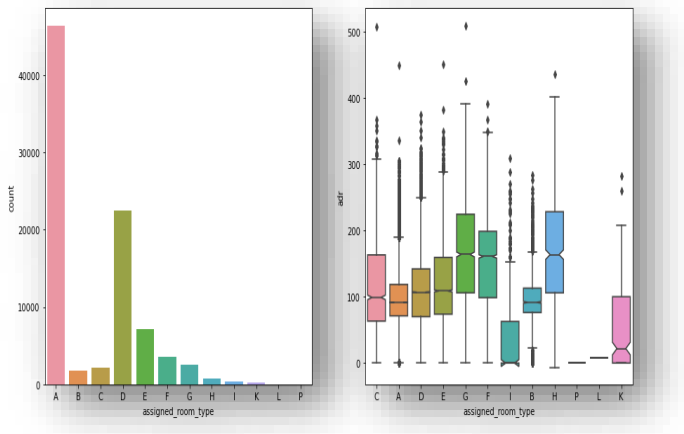


Fig of Analysis Room Demand

As shown in fig Most **demanded room type is A, but better adr rooms are of type H, G and F also. Hotels should increase the no. of room types A and H to maximize revenue**

Observation-5

meal type is most preferred meal of customers:

Here we will use count plot to plot the chart

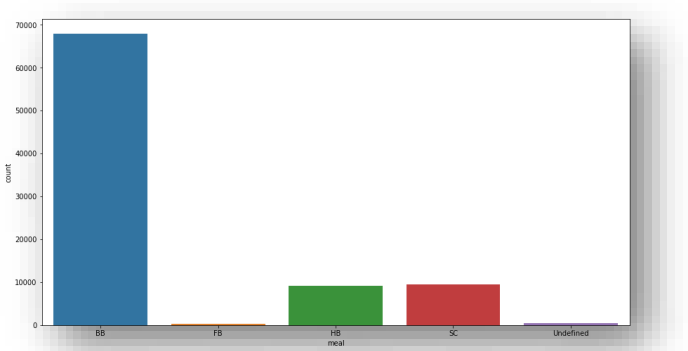


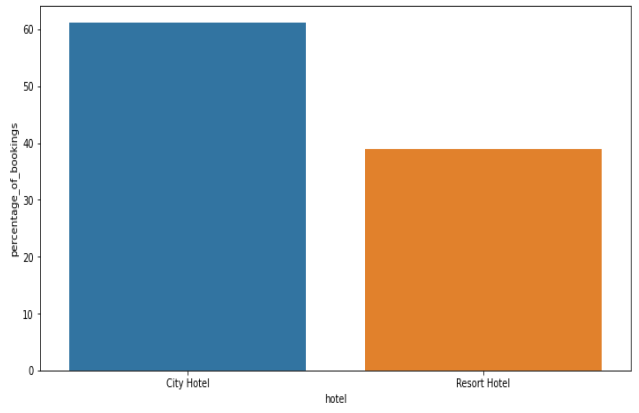
Fig. of meal preferred

As shown in fig Most preferred meal type is BB (Bed and breakfast).

Observation-6

Hotel wise analysis

What is percentage of bookings in each hotel?



As we seen in fig most of the peoples choose city hotel then Resort around 60% bookings are done in 'City Hotel' nd 40 % bookings are done in 'Resort Hotel'.

Form this we can also say 'City Hotel' has more avg adr as compared to 'Resort Hotel', hence 'City Hotel' is having more revenue than 'Resort Hotel'

Observation-7

What is preferred stay length in each hotel?

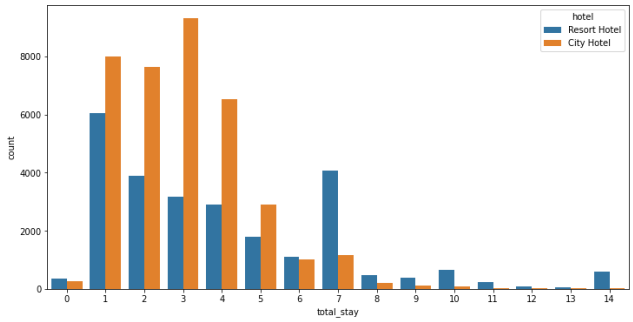
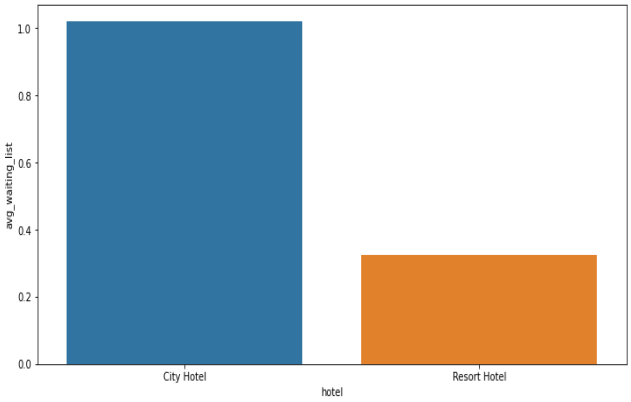


Fig. total stay and count

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.

Observation-8

Which hotel has longer waiting time?



City hotel has significantly longer waiting time, hence City Hotel is much busier than Resort Hotel.

Observation-9:

Which hotel has higher bookings cancellation rate.

Selecting and counting number of cancelled bookings for each hotel.

Counting total number of bookings for each type of hotel

Calculating cancel percentage

The after plotting this in table here we can seen that

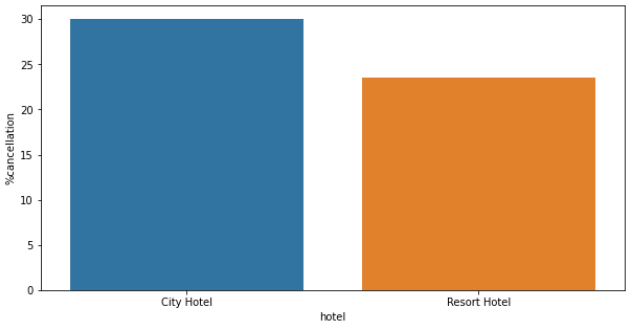


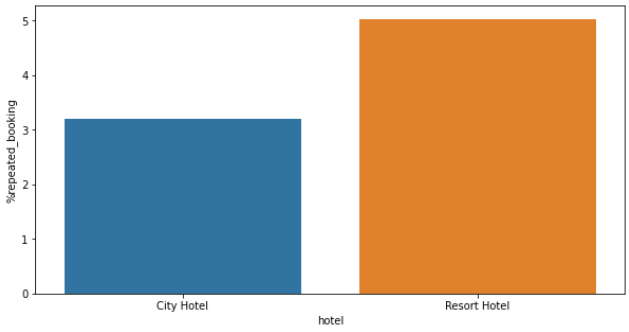
Fig Hotel cancelation rate

Almost 30 % of City Hotel bookings got canceled.

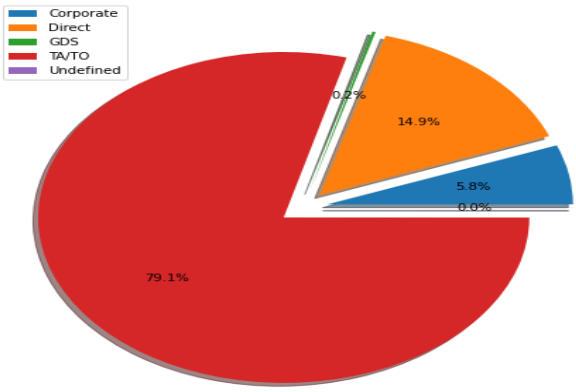
Observation-10:

Distribution Channel wise Analysis

Both hotels have very small percentage that customer will repeat, but Resort hotel has slightly higher repeat % than City Hotel.



the most common channel for booking hotels



As shown in fig the TA/TO has most commonly booking channel

CONCLUSION AND FUTURE WORK

The dataset contains immense possibilities to improve business values and have a positive impact. It is not limited to the problem taken into consideration for this project.

If it has user-friendly front-end user interface then on The base of data set is one of the use full
Data analysis system

ACKNOWLEDGEMENT

This project was completed by Rakesh Sahoo, Mahesh Patki, Lubna Zarin. We are extremely grateful to the celebrated authors whose precious works have been consulted and referred in our project work. We also wish to convey our appreciation to our peers who provided encouragement and timely support in the hour of need