Hotel Booking Analysis

Sai Narasimha Charan Thulasi
Data science trainees,
AlmaBetter, Bangalore

Abstract:

This Hotel dataset with the two hotel datasets first one is the about Resort hotel and second one is about the City hotel.Both datasets have the same data structure with 32 variables describing the data of the both of the hotels.both of the datasets lies in between 1st of july of 2015 to 31st august 2017 including the bookings that arrived and that of the bookings are cancelled. Due to the scaricity of real business data for scientific and educational purposes,these datasets can have an important role for resesearch and education in revenue management,machine learning or data minining as well asother fields.Descrptive analytics can be employed to further understand patterns,trends and anamolaies in data used platform research in different problems it are bookings,cancellation,reservation
against the results which are already known.researchers can use data sets to benchmark bookings prediction cancellation models against the results which are already known.Educators can use the datasets for machine learning classification or segmentation problem.Educators can use the datasets to obtain either statistics or data mining training.The hotel industries mainly based on the problems related to the predictions and forcasting based on the data which is generate from the aviation.In this format it is known as the PNR(passanger record name).actually it is generated by the aviation industry.in this dataset the data is related to various aspects which shows that complete information about the dataset.Not all the variables coming from the bookings or change log database tables some come from other tables as well and some are engineered from different tables.

2. Introduction

The dataset contains the data of the two hotels that is one of the resort hotel and second one is the city hotel.32 variables are provided into the dataset which is contains about 119390 entries altogether in this project we have done
EDA(exploratory data analysis) and generated some interesting facts about the data which is important for the prediction of the growth with respect to hotels.

3. 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

predict whether or not a hotel was likely to receive a 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. All personally identifying information has been removed from the data.

Explore and analyze the data to discover important factors that govern the bookings.

4. Data summary

The dataset contains the data of the two hotels that is one of the resort hotel and second one is the city hotel.32 variables are provided into the dataset which is contains about 119390 entries which contains the values such as float,integers,objects. The 32 columns hotels,is_cancelled,lead_time,arrival_date_year,arrival_date_month,arrival_date_week_number,arrival_date_day_month,stays_in_weekend_nights,stays_in_week_nights,adults,children,babies,meal,country,market_segment,distribution_channel,is_repeated_guest,previous_cancellations,previous_bookings_not_canceled,reserved_room_type,assigned_room_type,booking_changes,deposit_type,agent,company,days_in_waiting_list,customer_type,adr,required_car_parking_spaces,total_of_special_requests,reservation_status,reservation_status_date.

3. Steps involved

Various steps involved in this process to analyze the data

Importing required libraries

Mounting drive

Data cleaning

EDA(Exploratory Data Analysis)

1)Importing Required libraries

The libraries which are used in this project are numpy and pandas which is imported first and for the data visualization we have used matplotlib as well as seaborn libraries.In this project we have created different pie charts,bar charts,heat maps etc by using the both of libraries.Numpy and Pandas are both very useful while Performing Exploratory Data Analysis.

2)Mounting Drive

Next step is related to mounting the csv file,After mounting the csv file to colab notebook we have able read the all data and get info regarding data that means how much rows and columns etc.

3)Data Cleaning

1) The first step involves the checking and removing of duplicated rows in this step we have used value count method for duplicate rows and removed the values.

2) The next step involves the Handling of missing values in this step we handled values using null function after getting result we found that columns has 0 values means none of children made the transaction then we have replaced all null values under this column with the mean value of children.Next column with missing value is"country".This column represents the country of origin of customer.since,This column has datatype of string we will replace the missing value with the mode of country column.There are some rows with total number of adults,children or babies equal to zero.So we will remove such rows.The last step involves the converting the columns to appropriate data types.

3)Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

The process consists of several steps:

1.Importing a dataset
2.Understanding the big picture Preparation
3.Understanding of variables
4.Study of the relationships between variables
5.Brainstorming

Types Of EDA

Univariate Analysis

Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable

separately. It is possible for two kinds of variables- Categorical and Numerical.

Some patterns that can be easily identified with univariate analysis are Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation.

Univariate data can be described through: Frequency Distribution Tables The frequency distribution table reflects how often an occurrence has taken place in the data. It gives a brief idea of the data and makes it easier to find patterns. Example:

The list of IQ scores is: 118, 139, 124, 125, 127, 128, 129, 130, 130, 133, 136, 138, 141, 142, 149, 130, 154.

Bar Charts

The bar graph is very convenient while comparing categories of data or different groups of data. It helps to track changes over time. It is best for visualizing discrete data.

Pie Charts

Pie charts are mainly used to comprehend how a group is broken down into smaller pieces. The whole pie represents 100 percent, and the slices denote the relative size of that particular category.

Bivariate Analysis of two Numerical Variables (Numerical-Numerical)

Scatter Plot

A scatter plot represents individual pieces of data using dots. These plots make it easier to see if two variables are related to each other.

The resulting pattern indicates the type (linear or non-linear) and strength of the relationship between two variables.

Linear Correlation

Linear Correlation represents the strength of a linear relationship between two numerical variables. If there is no correlation between the two variables, there is no tendency to change along with the values of the second quantity.

Multivariate Analysis

Multivariate analysis is required when more than two variables have to be analyzed simultaneously. It is a tremendously hard task for the human brain to visualize a relationship among 4 variables in a graph and thus multivariate analysis is used to study more complex sets of data. Types of Multivariate Analysis include Cluster Analysis, Factor Analysis, Multiple Regression Analysis, Principal Component Analysis, etc. More than 20 different ways to perform multivariate analysis exist and which one to choose depends upon the type of data and the end goal to achieve. The most common ways are:

Conclusions

In this project we have done Hotel Booking Analysis and get some very

1. City Hotel seems to be more preferred among travellers and it also generates m ore revenue & profit.

2. Most number of bookings are made in July and August as compared rest of the months.

3. Room Type A is the most preferred ro om type among travellers.

4. Most number of bookings are made from Portugal & Great Britain.

5. Most of the guest stays for
1-4 days in the hotels.

6. City Hotel retains more number of guests.

7. Around one-
fourth of the total booking gets cancelled More cancellations are from City Hotel.

8. New guest tends to cancel bookings m ore than repeated customers.

9. Lead time, number of days in waiting list or assignation of reserved room to cu stomer does not affect cancellation of bookings.

10. Corporate has the most percentage of repeated guests while TA/TO has the least whereas in the case of cancelled bookings TA/TO has the most percentage while Corporate has the least.

11. The length of the stay decreases as A DR increases probably to reduce the cost

References

1. **Stack Over Flow**

2. **Geeksforgeeks**

3. **Wikipedia**