HOTEL BOOKING ANALYSIS

Aditya Saw and Parijat Krishna
Data science trainees,
AlmaBetter, Bangalore

1.Objective

We received a dataset of hotel reservations.

Our primary goal is to conduct EDA on the provided dataset and derive valuable findings regarding broad hotel booking trends and how various factors interact to affect hotel bookings.

2.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.

3. Dataset

We are given a hotel bookings dataset. This dataset contains booking information for a city hotel and a resort hotel. It contains the following features.

1. hotel : Hotel(Resort Hotel or City Hotel)

2. is_canceled : Value indicating if the booking was canceled (1) or not (0)

3. lead_time : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date

4. arrival_date_year : Year of arrival date

5. arrival_date_month : Month of arrival date

6. arrival_date_week_number : Week number of year for arrival date

7. arrival_date_day_of_month : Day of arrival date

8. stays_in_weekend_nights : Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

9. stays_in_week_nights : Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

10. adults : Number of adults

11. children : Number of children

12. babies : Number of babies

13. meal : Type of meal booked. Categories are presented in standard hospitality meal packages:

14. country : Country of origin.`

15. market_segment : Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

16. distribution_channel : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”

17. is_repeated_guest : Value indicating if the booking name was from a repeated guest (1) or not (0)

18. previous_cancellations : Number of previous bookings that were cancelled by the customer prior to the current booking

19. previous_bookings_not_canceled : Number of previous bookings not cancelled by the customer prior to the current booking

20. reserved_room_type : Code of room type reserved. Code is presented instead of designation for anonymity reasons.

21. assigned_room_type : Code for the type of room assigned to the booking.

22. 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

23. deposit_type : Indication on if the customer made a deposit to guarantee the booking.

24. agent : ID of the travel agency that made the booking

25. company : ID of the company/entity that made the booking or responsible for paying the booking.

26. days_in_waiting_list : Number of days the booking was in the waiting list before it was confirmed to the customer

27. customer_type : Type of booking, assuming one of four categories

a) adr : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights

b) required_car_parking_spaces : Number of car parking spaces required by the customer

c) total_of_special_requests : Number of special requests made by the customer (e.g. twin bed or high floor)

d) reservation_status : Reservation last status, assuming one of three categories

Cancelled **–** booking was canceled by the customer

Check-Out **–** customer has checked in but already departed

No-Show **–** customer did not check-in and did inform the hotel of the reason why

1.reservation_status_date : Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel.

There are total **119390 rows** and **32 columns** in a given data set.

4.Steps involved in Data Cleaning:

1.Removing duplicate rows:

All redundant rows were removed.

2.Handling null values:

- Null values in columns company and agent were replaced by 0.
- Null values in column country were replaced by 'others'.
- Null values in column children were replaced by the mean of the column.

3. Converting columns to appropriate data types:

- Changed data type of children, company, agent to int type.
- Changed data type of reservation_status_date to date type.

4. Removing outliers:

- One outlier was found in the adr column. Simply dropped it.

5. Creating new columns:

- Created new column total_stay by adding stays_in_weekend_nights+stays_in_week_nights.
- Created new column total_people by adding adults+children+babies.

5.Research Methodology:

Univariate Analysis-

Univariate Analysis is a quantitative-statistical method of evaluation. With this approach of analysis, each variable in a data set is examined separately and the results are each summarised separately.

Therefore, unlike bivariate and multivariate analysis, which look at interactions between several variables, univariate data only serves to describe one component of a piece of research. Although different forms can be utilised, a frequency distribution table or bar graph is the simplest way to combine the data for a single variable (e.g. pie chart, histogram etc.). This indicates that one of these selected modes of presentation is used to analyse the number of examples in a given category (variable).

Bivariate analysis-

Bivariate analysis is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y. So, here we analyse the changes occured between the two variables and to what extent. Apart from bivariate, there are other two statistical analyses, which are Univariate (for one variable) and Multivariate (for multiple variables).

In statistics, we usually interpret the given set of data and make statements and predictions about it. During the research, an analysis attempts to determine the impact and cause in order to conclude the given variables.

Bivariate analysis is stated to be an analysis of any concurrent relation between two variables or attributes. This study explores the relationship of two variables as well as the depth of this relationship to figure out if there are any discrepancies between two variables and any causes of this difference. Some of the examples are percentage table, scatter plot, etc.

6.Libraries and Tools used in Data Visualisation:

Library used in Data Visualisation are-

1.Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication quality plots.
- Make interactive figures that can zoom, pan, update.

- Customize visual style and layout.
- Export to many file formats.
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib.

2.Seaborn- Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

Graphical tools used in Data Visualization-

- Bar Plot.
- Histogram.
- Scatter Plot.
- Pie Chart.
- Line Plot.
- Heatmap.
- Box Plot

7. EDA (Exploratory Data Analysis):

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

7.(i) Data Analysis:

1.Which type of hotels is mostly preferred by the guests?

2. Which agent made the most bookings?

3. What is the percentage of cancellation of booked hotels?

4. What is the percentage of repeated guests?

5. What is the percentage distribution of *Customer Type*?

6. What is the most percentage of booking changes made by the customer?

7. Most preferred type of food by guests?

8. From which country the most guests are coming?

9. The most preferred room type by the customers?

10. Which year has the most hotel bookings?

11. Which hotels have the most repeated guests?

12. Which distribution channel contributed more to ADR in order to increase the the income?

13. What is the percentage distribution of required car parking spaces?

14.What is Percentage distribution of Deposit type?

15. What is most preferred stay length in each hotel?

16. Which hotel makes more revenue?

17. For which hotel, does people have to wait longer to get a booking confirmed?

18. Whether stay is over a weekend or weekday?

19. Which channel is contributing most for early booking of the hotel?

20. Which distribution channel brings better revenue generating deals for hotels?

7. (ii) Conclusion:

1. City Hotel is more preferred by guest as compared to Resort Hotel. Thus, City hotel has maximum bookings as compared to Resort Hotel.

2. The Agent ID-9 had done most of the bookings.

3. 27.5 % of the bookings is cancelled.

4. Repeated guest are less in numbers that is 3.9% only.

In order to retained the guests/customers, management should take feedbacks from guests and try to imporve the services.

5. Transient customer type is most which is 82.4 % while percentage of bookings associated by other groups is vey low.

6. Almost 82% of the bookings was not changed by guests while approx. 10% bookings was changed by guests.

7.

a) The most preferred food by the customers is BB (Bed and Breakfast).

b)HB (Half Board) and SC (Self catering) are equally preferred.

8. Most of the guests are coming from Portugal (PRT) that is more than 25000 guests.

9. The most preferred room type is A.

10. July and August months had the most Bookings. Summer vaccation can be the reason for bookings.

11.

a) 2016 Year has the most bookings.

b) 2015 has less than 10,000 bookings

c) While overall city hotels had the most of the bookings.

12. City Hotels have the highest percentage of booking cancellation as compared to Resort Hotels.

13. 93.8 % guests did not required the parking space. Only 6.2 % guests required only 1 parking space.

14. 98.7 % of the guests prefer No deposit type of stay.

15. The optimal stay in both types of hotels is less than 7 days.

16.

a) City hotels has slightly high average lead time than resort hotel.

b) Hence, city hotel makes slightly more revenue then resort hotel.

17.

a) City hotel has significantly longer waiting time then resort hotel.

b) Hence, City Hotel is much busier than Resort Hotel.

18. Majority of the stays are over the weekday's night. Whatever we saw for the chart on day of the month was random.

19. a) We concluded that the most of the bookings we have received from TA/TO.

b) Other than undefined the least booking we have received from Corporate.

20.

a) In terms of revenue GDS is the most revenue generating Channel but its only for City hotel. For Resort Hotel its contribution is negligible as compared to other channels distribution.

b) Undefined can be associated to multiple channel distribution channels whose data is not provided so after undefined bookings from TA/TO are generating most revenue for the Resort Hotel.

c) Apart from other ditribution channel Direct bookings are also playing the crucial role in terms of revenue generation but we need to focus more on other less revenue generating mediums in order to increase the overall revenue.

8. Challenges:

1. Huge chunk of the data was to be handled keeping in mind not to miss anything which is even of little reference.
2. Feature selection was quiet challenging as our data set had many futuristic features which had no relevance for initial detection.
3. Computation time