Exploratory data analysis - Hotel Booking Analysis

Technical documentation

—

Mohammed Arifuddin Atif
arifuddinatif63@gmail.com

Introduction

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.

Problem statement

We are tasked with performing exploratory data analysis on the given dataset to get relevant insights from the data and understanding the key factors responsible for hotel bookings given particular parameters.

Overview of data

We are given the following columns in our data:

1. **hotel**
2. **is_canceled**
3. **lead_time**
4. **arrival_date_year**
5. **arrival_date_month**
6. **arrival_date_week_number**
7. **arrival_date_day_of_month**
8. **stays_in_weekend_nights**
9. **stays_in_week_nights**
10. **adults**

11. **children**
12. **Babies**
13. **Meal**
14. **Country**
15. **Market_segment**
16. **Distribution_channel**
17. **Is_repeated_guest**
18. **previous _cancellations**
19. **Previous_bookings_not_canceled**
20. **Reserved_room_type**
21. **Assigned_room_type**
22. **Booking_changes**
23. **Deposit_type**
24. **Agent**
25. **Company**
26. **Days_in_waiting_list**
27. **Customer_type**
28. **Adr**
29. **Required_car_parking_space**
30. **Total_of_special_requests**
31. **Reservation_status**
32. **reservation_status_date**

Performing EDA (exploratory data analysis)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to see assumptions with the assistance of summary statistics and graphical representations.

1. Extracting head and tail of the dataset.

2. Extracting info of the dataset which tells the type of data present in different columns.
3. Producing description of data.
4. Checking for null values.
5. Filling the null values with mean and mode of the relevant columns.
6. Creating dummies for non-numeric types of data.
7. Dropping irrelevant columns or columns with null values more than 50%.
8. Detecting and removing outliers from all the columns.
9. Manipulating and concatenating data according to the requirement.
10. Plotting relevant graphs to extract information from them.

Plots used

Plots are a way to describe the data given to us in a visual manner which is more understandable and convenient to draw insights from , compared to raw data.In this project the plots we used are:

1. Countplot - provides the count of required instances from the data.
2. Boxplot - provides information on the outliers present in the data.
3. Histplot - also provides count of instances (seaborn based).
4. Pieplot - provides info in the form of a pie chart.
5. Pairplot - Pairplot visualizes given data to find the relationship between them where the variables can be continuous or categorical.

Conclusions from data

1. There were more bookings in city hotels than resort hotels.
2. Arrival date year were more in the year 2016 compared to 2015 and 2017.
3. Bookings were maximum in the month of august compared to other months.
4. Bookings were minimum in the month of january compared to other months.
5. There were a total of 37.04% of cancelled bookings in the data.
6. Top 3 countries with most bookings are portugal,france,great britain.
7. Online travel agent has the most market segment share in terms of bookings.
8. Aviation has the least market segment share.
9. Most customers preferred BB-bed and breakfast as the meal type during booking.
10. Full board type of meal was the least selected from all the meal types.
11. The customer type with most bookings is the Transient type.
12. The least count of customer type from the data was group type.
13. The type of room most customers preferred was ‘A’ type.
14. The least preferred room type is ‘L’ type.
15. The percent of repeated guests from the dataset is 96.8% which is quite high.
16. Most customers made a no deposit type of booking compared to other booking types.
17. Most customers booked hotels with no parking space.
18. The count of total members per reservation of ‘2’ was higher than other numbers and ‘5’ was the least.

Challenges faced

1. Pre-processing the data was one of the challenges we faced , such as detecting outliers and removing them.
2. Manipulating the data was difficult as it could affect and alter important information provided by the dataset.
3. As the dataset has more columns it was quite difficult to extract relevant information from all the columns.

Final conclusion

We are finally at the conclusion of our project!

We performed exploratory data analysis from the given dataset and were able to draw important and relevant information which explains why in a particular instance there are more hotel bookings compared to other instances.

We also used different types of plots to better visualize and present the data in a more appealing manner.