Exploratory Data Analysis on Hotel Booking Analysis

Chetan Jadhav,
Pooja Parsana,
Robin Rego
Data Science Trainees,
Alma Better, Bangalore

Abstract:

We had Hotel Booking Dataset of City Hotel as well as Resort Hotel which was to be analyzed and proper insights should be taken out which can be useful to provider in future for making important decisions.

The Dataset which was provided include columns like cancelled Booking, arrival Day/Date/month/year of customer, market segment type, number of family members, parking space, type of meal, type of room etc. The Dataset is of size (119390, 32).

Keywords: Data Science, Data Analysis, EDA, Data Visualization.

1. Problem Statement

Hotel Bookings depends on various factors and if those aren't properly managed can lead to fall of hotel. Factors which affect bookings include food type, prices, month of year, country etc.

Our main objective is to perform Data Analysis of Hotel bookings and to give in-sights to hotel management which will boost their performance.

- **Objective:**
- Exploring and Cleaning the Dataset.
- To establish relationship between various features of the Dataset.
- Present these relationships using various Data Visualization Techniques.
- Draw the useful insights from it

2. Introduction

This project contains real world data record of hotel bookings of city and resort hotel containing details like bookings, cancellations, guest details etc. from 2015 to 2017. Main aim of project is to understand and visualize dataset from hotel and customer point of view.

- **What is EDA?**

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

EDA is primarily wont to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a far better understanding of knowledge set variables and therefore the relationships

between them. It also can help determine if the statistical techniques you're considering for data analysis are appropriate. Originally developed by American mathematician John Tukey within the 1970s, EDA techniques still be a widely used method within the data discovery process today.

- **Why is EDA important in Data Science?**

The main purpose of EDA is to assist check out data before making any assumptions. It can help identify obvious errors, also as better understand patterns within the info, detect outliers or anomalous events, and find interesting relations among the variables.

Data scientists can use exploratory analysis to make sure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they're asking the proper questions. EDA can help answer questions on standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

- **Types of EDA**

There are four primary sorts of EDA:

Univariate Non-Graphical: This is often simplest sort of data analysis, where the info being analyzed consists of only one variable. Since its one variable, it doesn't affect causes or relationships. The most purpose of univariate analysis is to explain the info and find patterns that exist within it.

Univariate Graphical: Non-graphical methods don't provide a full picture of the info. Graphical methods are therefore required. Common sorts of univariate graphics include:

- **Stem-and-leaf plots:** This shows all data values and therefore the shape of the distribution.
- **Histograms:** This shows a bar plot during where each bar represents the frequency (count) or proportion (count/total count) of cases for a variety of values.
- **Box plots:** This shows graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.

Multivariate Non-graphical: Multivariate data arises from quite one variable. Multivariate non-graphical EDA techniques generally show the connection between two or more variables of the info through cross-tabulation or statistics.

Multivariate graphical: Multivariate data uses graphics to display relationships between two or more sets of knowledge. The foremost used graphic may be a grouped bar plot or bar graph with each group representing one level of 1 of the variables and every bar within a gaggle representing the amount of the opposite variable.

Other common sorts of multivariate graphics include:

- **Scatter plot:** This is employed to plot data points on a horizontal and a vertical axis to point out what proportion one variable is suffering from another.

- **Multivariate chart:** This is a graphical representation of the relationships between factors and a response.
- **Run chart:** This is a line graph of knowledge plotted over time.
- **Bubble chart**: This may be a data visualization that displays multiple circles (bubbles) during a two-dimensional plot.
- **Heat map:** This may be a graphical representation of knowledge where values are depicted by colour.

3. Steps involved

The operations on the dataset are done by Python scripts on Google Colab.

A. Importing Libraries:

Following Libraries are used in this analysis

i. **NumPy:** NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of Array.

ii. **Pandas:** Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.

iii. **Matplotlib:** Matplotlib is a visualization library in Python for 2D plots of arrays.

iv. **Seaborn:** Seaborn is a library for making statistical graphics in Python.

B. Data Cleaning:

The dataset contains many columns having huge number of null values; hence we dropped those columns and kept only those columns which were important in our Data Exploration journey. The Data frame before deleting the null values was 119390 rows × 32 columns.

After eliminating the null values the final Data Frame consists of 118898 rows × 30 columns.

C. Creating the visualizations:

Using the various functions of above mentioned libraries, we have created various types of charts like Count Plot, Pie Chart, Line Plot to establish meaningful relationship between the variables of the Data set. You can see those charts in attached code file.

4. Observations

1. The confirmed bookings goes from their lower value (4068) in January to their highest value (8618) in August.
2. Cancelations of bookings is 37.13%.
3. The relation of prices between hotels for resort hotels are higher and fluctuate more than city hotels.
4. For resort hotels, the average daily rate is more expensive during august, July and September.
5. For city hotels, the average daily rate is more expensive during august, July, June and May.
6. The month of highest occupation is August with 11.65% of the reservations. The month of least occupation is January with 4.94% of the reservations.
7. There were 25,829 registered changes in the bookings during this period.
8. 7,342 car parking spaces have been used.

5. Measures that can be taken

Very Less amount of customers revisit the hotel, so hotel staff should consider taking valuable feedback of their services.

- As Portugal, Britain, and France have most customers, marketing team can target those countries as well.
- As July to October has most number of visitors, hotel should focus on trained staff during this time.

6. Conclusion

A. Higher lead time has higher chance of cancellation.

B. The city hotel has more guest during spring and autumn, when the price is also highest.

C. Number of guests for resort hotel go down from June to September, which is also when the price is highest. Thus, these months should be avoided for bookings.

D. April to august is the peak season for the bookings.

E. Cancellation is high when done through agent compared to direct bookings. Hotel need to do marketing and give special incentives for direct booking.

References-

1. Pandas user guide
2. Matplotlib user guide
3. Seaborn user guide
4. Sergio Alves