Hotel Booking Analysis

Shubham Shankar Panaskar

Data science trainee,
Alma Better, Bangalore.

Abstract:

A hotel is an establishment that provides paid lodging on a short-term basis. Small, lower-priced hotels may offer only the most basic guest services and facilities. Larger, higher-priced hotels may provide additional guest facilities.

A resort is a self-contained commercial establishment that tries to provide most of a vacationer's wants, such as food, drink, lodging, sports, entertainment, and shopping, on the premises. The term resort may be used for a hotel property that provides an array of amenities, typically including entertainment and recreational activities.

1. Problem Statement

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

2. Introduction

When people search for a hotel to stay on vacation they consider various factors such as price, location, availability, parking space, food, accommodation options, etc.

Prices of the Hotels can also vary according to the month of booking, the number of guests, days of stay, hotel locations, hotel ratings, any special request, etc.

The objective of this project is to deliver insights to understand when the best time of year to book a hotel room is? Or the optimal length of stay to get the best daily rate? 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!

Following is the roadmap we decided to work on before deep diving straight into the solution.

1. Loading the data into the data frame
2. Cleaning the data
3. Statistical extraction of data

4. Exploratory data analysis
5. Conclusion

3. Some Keywords

There are some keywords we will be using.

1. **hotel**: type of hotels
2. **is_canceled**: canceled or not
3. **lead_time**: no. of days before actual arrival in the hotel
4. **arrival_date_year**: year of booking
5. **arrival_date_month**: month of booking
6. **arrival_date_week_number**: week number of the year in which booking
7.**arrival_date_day_of_month**: arrival month date
8. **stays_in_weekend_nights**: no. of weekends guest stayed
9. **stays_in_week_nights**: no. of weekdays guest stayed
10. **meal**: BB – Bread and Breakfast
HB – only two meals including breakfast meal
FB – breakfast, lunch, and dinner
11. **market_segment**: TA: Travel agents
TO: Tour operators
12. **previous_cancellations**: cancellation in past
13. **previous_bookings_not_canceled**: not canceled in the past.

4. Steps involved:

- **Loading the dataset**

We created a directorial path for the Hotel Booking dataset, using the Pandas read function we read it. It has a shape (119390, 32) which means it has 119390-row labels and 32 features or column labels.
After reading it we found which are the dependent variables and which are the independent variables.

- **Cleaning and Transforming Data**

Cleaning is the process of removing undesired features, values, or any suffix, prefix, or anything which can produce an exception.
Transforming is completely a different process, transforming is required to ensure the consistent data type of features because inconsistent data type will generate an obstacle during the execution of the program.
These two processes have specific subprocesses as follows.

- **Unwanted Data Removal**

In this step we ensured to make a data type of feature consistent by removing characteristics from the values of features, to make them usable. Such as Agent and children are the columns with float datatype, but their values are in integer. So we will convert them into integers.

- **Null values Treatment**

The company and agent column has 94 % of Null values so it’s not

feasible to fill that many null values so we drop this column.
The country column contains the country codes of the guests, it is a categorical feature so we will replace null values with the mode value.

5.1: Exploratory Data Analysis

Following are the observations using Exploratory Data Analysis and visualization.

5.1.1 Hotel Booking Percentage with Pie Chart:

With this Pie Chart we got to understand that out of total bookings 66.4% is City Hotel and 33.6% is Resort Hotel.

City Hotels are most preferred hotel by guests. Thus city hotels has maximum bookings.

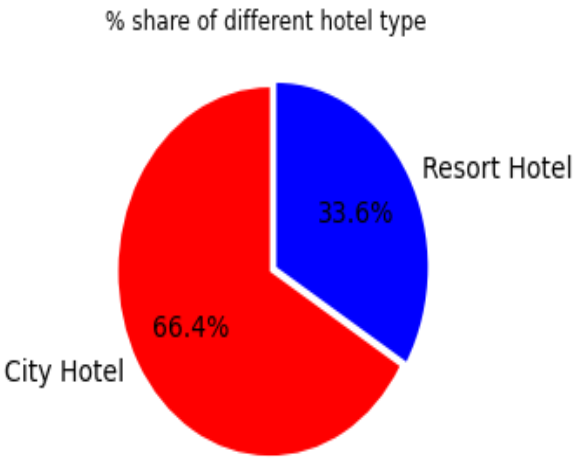

Fig 1: Hotel Booking Percentage

5.1.2 Correlation Between features Seaborn Heatmap:

A heat map (or heatmap) is a data visualization technique that shows the magnitude of a phenomenon as color in two dimensions.

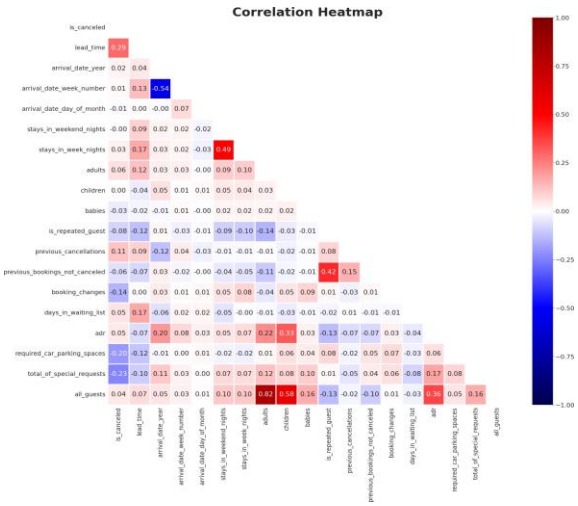

Fig 2: Correlation Matrix

A **positive correlation** is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases or one variable decreases while the other decreases.

A **negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other.

A **zero correlation** exists when there is no relationship between two variables.
By plotting seaborn heatmap correlation we got to know that there are three features in the given hotel booking data set which are highly correlated with each other, It can be observed arrival_date_week_number and arrival_date_year is 54% negatively correlated,

previous_bookings_not_cancelled and is_repeated_guest are 42% positively correlated and adr and children are positively correlated by 33%

5.1.3 Number of Bookings vs Months

Count plot:

A count plot is kind of like a histogram or a bar graph for some categorical area. It simply shows the number of occurrences of an item based on a certain type of category. We use the arrival_date_month column available in the hotel booking data set to plot the number of bookings per month.

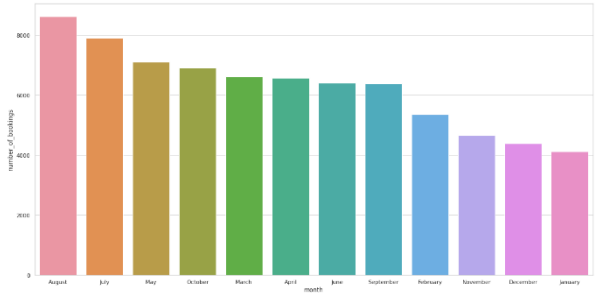

Fig 3: Number of Bookings vs Months

The number of bookings in August and July month is more as compared to other months.

5.1.4 Lead Time vs Cancelation[%]

Scatter Plot:

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded, one additional variable can be displayed.

We used the scatter plot to understand the relation between the lead time and cancellation.

The scatter plot we implied gives the result that when the lead time increases there are higher chances of booking cancellation.

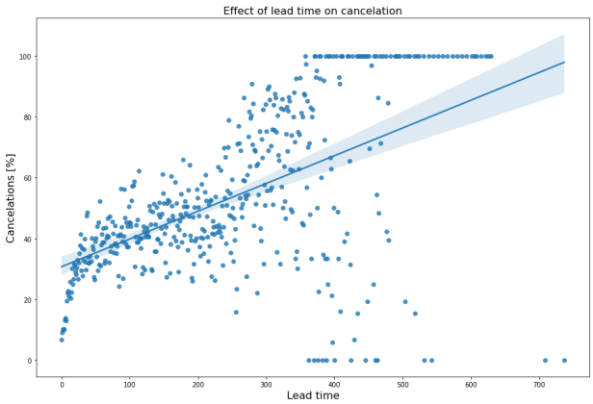

Fig 4: Lead Time vs Cancelation[%]

5.1.5 Number of bookings vs Country

Plotly bar plot:

We used bar plots to find out the country with the most guest's origin.

With the help of the bar plot, we found that Portugal is the country with the most guest's origin.

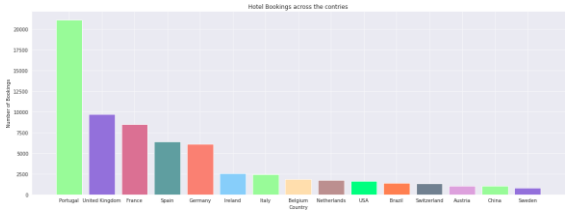

Fig 5: Number of Booking vs Country

5.1.6 ADR vs Arrival_Date_ Month

Line Plot: Using this line plot we find Which months have cheaper booking rates? For both city and resort hotels, November to January have cheaper average daily rates.

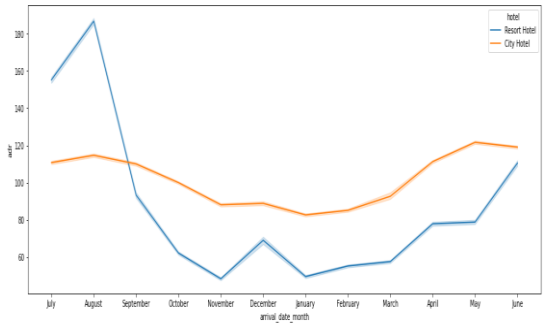

Fig 6: ADR vs arrival_date_month

5.1.7 Meal Type Bar plot:

Meal type is the categorical column available in the hotel booking dataset.

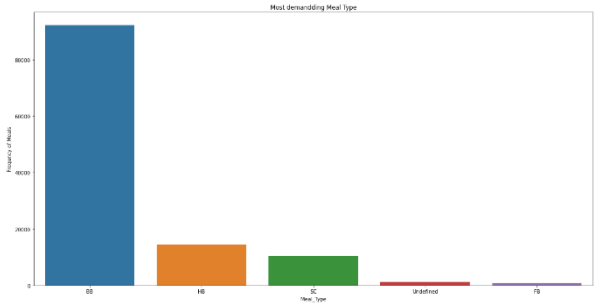

Fig 7: Meal Type

The above Bar Graph plot shows that ‘Breadand Breakfast’ was the most ordered meal type.

Conclusion

1. More than 37% of bookings were canceled.
2. Online Travel Agents followed by Offline Travel Agents brings in most of the bookings.
3. Portugal is the top country from where most hotel bookings are coming.
4. Bed and Breakfast is most preferred meal
5. Month of August is the most trending month for Hotel Booking.
6. Couple (or 2 adults) is the most popular accommodation type. So hotels can make plans accordingly.
7. Plotting the heatmap
    - ADR and children are positively correlated by 33%.
    - It can be observed that arrival_date_week_number and arrival_date_year are 54% negatively correlated.
    - Previous_bookings_not_cancelled and is_repeated_guest are 42% positively correlated.
8. Average Daily Rate (ADR) for the months of July and August are strikingly more for the Resort Hotel than the City Hotel.
9. It is observed that lead time has a positive correlation with cancellation.

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. Many other interesting possibilities can be explored using this dataset.

Future work can include

1. Live interactive dashboard can be built using tableau.

References

1. Geeksforgeeks