Hotel Bookings Analysis

Sahil Jagtap, Shweta Jagtap, Swapnil Wankhede

Data Science Trainees,

AlmaBetter, Bangalore

Abstract

Exploratory Data Analysis in the hotel industry is key to marketing strategy, building customer loyalty and enhancing productivity. By analysing the patterns available in the given dataset which has 119390 observations with 32 variables, it is helpful to make the hotel's plan better so that hotel can perform various campaigns to boost the business. Pandas, NumPy, matplotlib, seaborn are the libraries are used to explore, analyse and visualize the data. It is observed that in most of the cases, city hotel is preferable than resort hotel.

Keywords: Hotel, Average daily rate, Length of stay, Bookings, etc.

Problem Statement

This dataset contains the information for a city hotel and a resort hotel, and includes several variables such as when the bookings were made, length of stay, the number of adults, children, and babies etc. All personally identifying information has been removed from the data.

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

Introduction

The objective of this project is to deliver insights to get 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? Which type of distribution channel are mostly used for which hotel type etc.

Many more things have been explored and analysed with the help of given dataset which contains 119390 observations with 32 variables which are mix of float, integer and object datatype including some missing values and duplicates that effect the dataset. There was lots of duplicates data and lots of null values in some variable like country, company, agent which was the tough challenges.

Pandas, NumPy, matplotlib, seaborn are the libraries are used to explore and analyse the data. Pie chart, bar (stacked) chart, line plot, heatmap, etc used to visualize the data. Loading the data into the data frame, Data Exploration and Data Cleaning, Exploratory data analysis are the roadmap was decided to come into proper conclusion.

Data Exploration and Data Cleaning

1. Loading the dataset and exploration of data

A directorial path for the dataset is created using the Pandas read function. Dataset has a shape (119390, 32) which means it has

119390 observation (row labels) and 32 features or variable or column labels. To explore the data, head, tail, info, describe, value_counts, drop_duplicates, loc and iloc, isnull, crosstab etc are used to understand the data. It is found that data contain lots of duplicate's data and lots of null values. It is necessary to do the data cleaning which is process of removing undesired features, values, or any anything which can produce an exception. Some action has been done to overcome this situation these are

2. Removing duplicates rows

It is observed tha there are 31994 duplicates row which has been dropped by using drop function.

3. Handling null values

It is observed that some variable contain the null values which are company(82135), agent(12193), country(452), children(4) with their missing values counts. Null values in columns is replaced by 0.0 except country columns Null values in country is replaced by other.

4. Converting columns to appropriate data types.

It has been changed datatype of children, agent, company into int types as they are not in proper datatypes.

5. Creating new columns

To make the better analysis, new coloum is created by the name new stays_in_weekend_and_week_nights by adding stays_in_weekend_nights and stays_in_week_nights

6. Updated dataset.

Updated dataset has nearly 87228 observations with 33 variables. It is a mix of numerical and categorial variables. By basic inspection, it is found that hotel, arrival date months, meal, market segment, distribution channel, deposit_type etc are categorical variables. Following are variables which are frequently used within dataset have been elaborated with basic meanings

1. **hotel**: resort and city type hotels.

2. **is_canceled**: booking is 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 of 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 – Bed & Breakfast
HB – only two meals including breakfast meal
FB – breakfast, lunch, and dinner etc.

Exploratory Data Analysis

EDA is an approach of analysing data sets to summarize their main characteristics. Pandas, NumPy, matplotlib, seaborn are the libraries use to explore and analyse the data. Pie chart, bar (stacked) chart, line plot, heatmap etc use to visualize the data. Following are the some observations which is analysed and visualized. These are

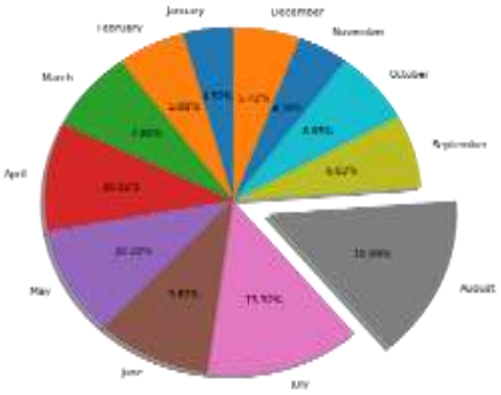
1. Percentages of bookings cancelled and not cancelled month wise

It is found that is_canceled variable takes binary values o and 1

0 indicates bookings not cancelled.

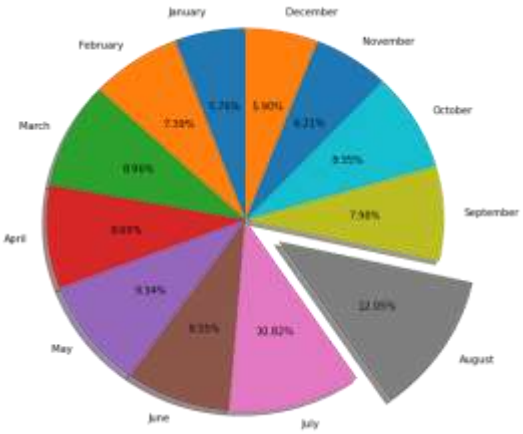
1 indicates bookings cancelled.

Cancelled and not cancelled dataframe is made. It is visualized by pie chart. With this pie chart we got to understand that 15.09% bookings have been cancelled in august month which is maximum among all months as shown below



Cancelled Bookings Pie Chart

12.05% bookings have not been cancelled in august month which is maximum among all months as shown in below.

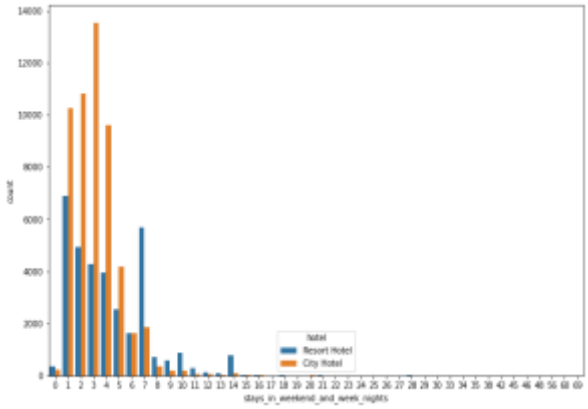


Not Cancelled Bookings Pie Chart

As a result, most hotel rooms are booked in august month which is the best time of year to book a hotel room.

2. Mostly booked hotel in weekend nights and week_nights

To compute a simple cross tabulation of hotel and stays_in_weekend_nights or stays_in_week_nights, crosstab pandas' function used. To visualize the data, grouped bar chart used.



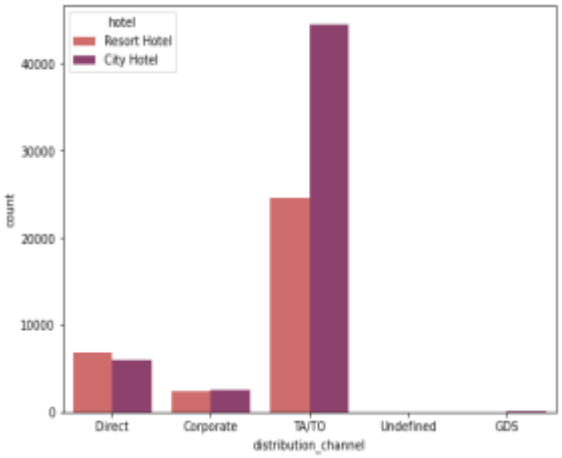
It is observed that City hotels are mostly booked in weekend nights and week nights. It is concluded that for a short period span customer refer the city hotel where as for the longer period span customer refer the resort hotel.

3. Which type of hotel get maximum number of cancelled of bookings?

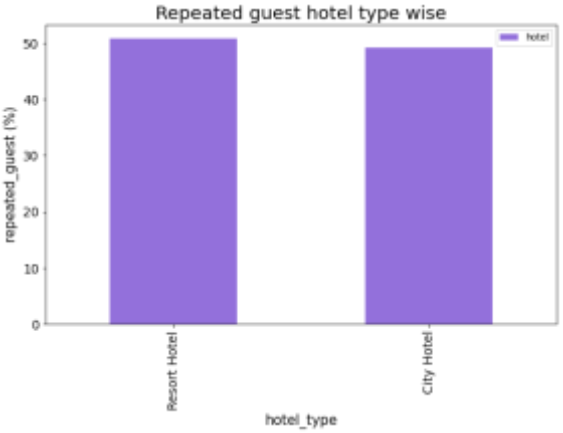
For this analysis, bar chart used. It is found that city hotel has maximum number cancellation of bookings as compared to resort hotel.

4. Which type of distribution channel are mostly used for which hotel type?

Grouped bar chart is used where distribution channel are categorised into direct booking, booking through corporate, online travel agencies, undefined and global distribution system. From the graph we conclude that Most of the rooms of city hotels are sold by TA/TO distribution channel as compared to resort hotel.



5. Which type of hotels have maximum repeated guests?



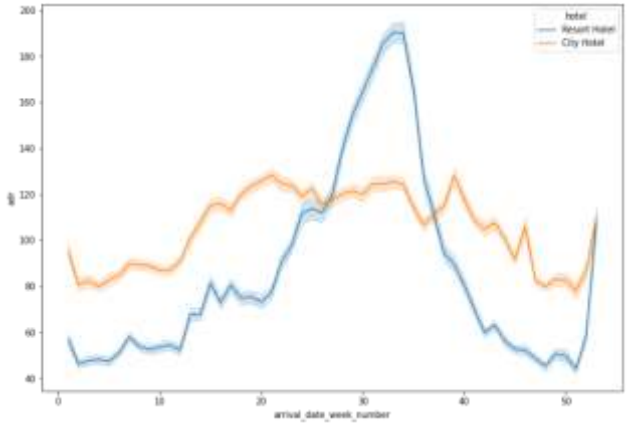
It is observed that resort hotel type has maximum repeated guest as compared to city hotel.

6. Bookings trend of hotels (city hotel and resort hotel) month wise

To visualize the booking trend of hotel line graph is used. It is observed that trend of city and resort hotels is kind of similar with some difference. For the both hotels, more booking is in august month followed by July month and in october month, percentage of booking for both hotels is same.

7. Comparison of Hotel's ADR (Average Daily Rate) week wise.

To visualize hotel's adr week wise as adr is a performance indicator used in hospitality sector to measure the strength of revenue generated.

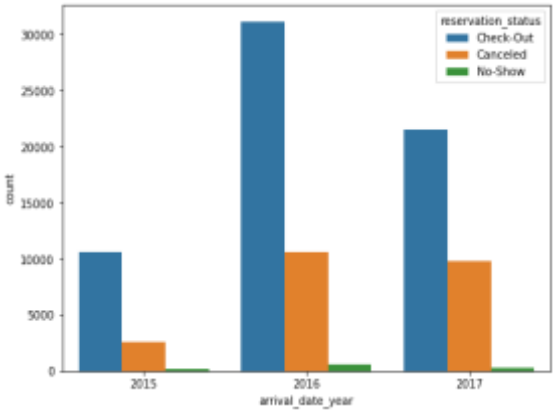


It is observed that difference of adr in both hotel types is high at the starting week of year i.e January month. At the end of week, adr for both hotel is same. Adr for the week

28 to 35 approx. for months of July and August are more for the resort hotel than city hotel.

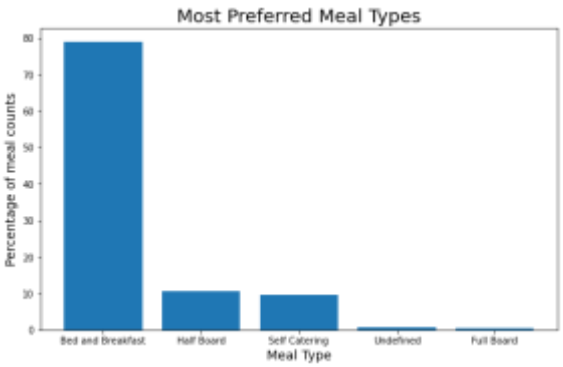
8. In which year maximum number of reservation status is checked out?

It is observed that In 2016, maximum number of reservation_status is checked_out as compared to year 2015 and 2016.



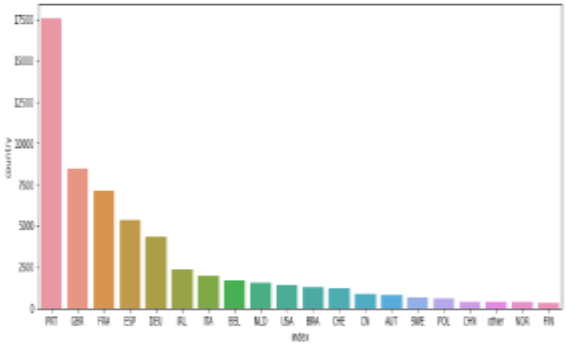
9. Which is the most preferable meal Types?

Meal is categorized into BB (bed and breakfast), HB (Half Board), SC(Self Catering), undefined and Full Board. It is observed that customer prefer BB more as compare to other meal types.



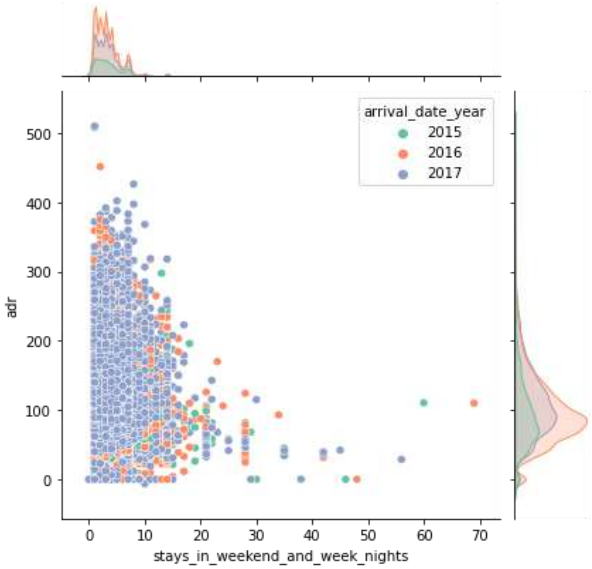
10. Maximum Bookings From Top Four Countries

It is observed that country variable is taken countries. PRT (Portugal), GBR (United Kingdom), FRA (France) and ESP (Spain) are the top four countries where the bookings are maximum.



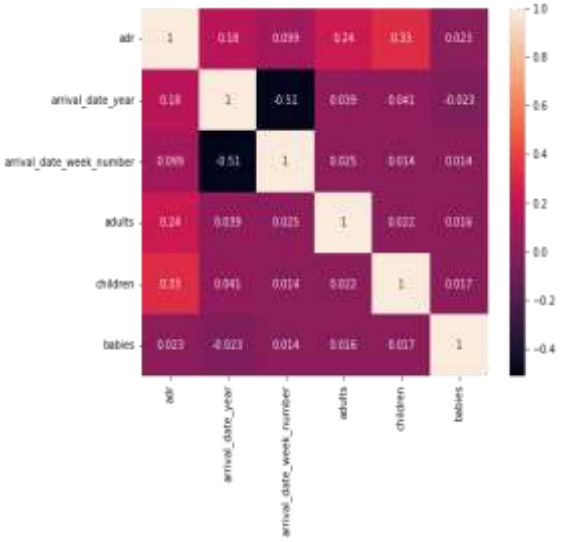
11. The Best Optimal Length To Get The Best Daily Rate

Joint plot is used to visualize adr and stays_in_weekend_and_week_nights. It is observed that most of stays data points for years 2015, 2016 and 2017 are clustered between 1 to 10 week and weeknights. This is the optimal length of stay. As length of stays increases the adr decreses. This means for longer stay, the better deal for customer can be finalised.



12. Corelation between features using Heatmap

To visualize the corelation between features heatmap is used which gives corelation value which lies between -1 to +1. -1 indicates two features are negatively corelated whereas +1 indicates two features are positively corelated. First all the numeric features are analysed then picked some important features those are shown strong relation. The Heatmap is as shown below.



It is observed that adr is positively correlated to children, adults, and babies with 33%, 25% and 2% respectively. It means that maximum and minimum revenue is generated by occupied rooms with children and babies respectively.arrival date week number and arrival date year are 51% negatively correlated.

Conclusion

It was wonderful learning experience while working project. This project gave real insight. We learned that

1. August is the best time of year to book a hotel room.
2. City hotels are mostly booked in weekend nights and week nights and maximum number of cancellations of booking as compare to resort hotel.
3. Most of the rooms of city hotels are sold by TA/TO distribution channel as compared to resort hotel.
4. The city hotel got the higher number of special requests.
6. More booking is in august month followed by july month. In october month, percentage of booking for both hotels is same.
7. ADR for the week 28 to 35 are more for the resort Hotel than city hotel.
8. In 2016, maximum number of reseervation_status is checked_out.
9. For every customer, BB (Bed and Breakfast) is most preferable meal type.
10. Portugal), United Kingdom, France and (Spain) are the top four countries where the bookings are maximum.
11. The best optimal length of stay is 1 to 10 week and weekends nights in order to get the best daily rate.
12. Average daily rate(ADR) is positively correlated to children, adults, and babies with 33%, 25% and 2% respectively.

Challenges

There were lots of duplicate data, Null values. To replace these values and to use particular graph while visualizing the data were tough challenge.

Future Work

This dataset contains huge possibilities to boost hospitality sector. It is not limited to the problem taken into consideration. Future work may be use of regression and classification algorithms on the several features.

References:

1. medium.com
2. towards.datascience.com