Hotel Booking Analysis

Rohun Rajvanshi, Pravin Singh, Shubham Jindal,
Shubham Shahu, Keshaw Agarwal

ABSTRACT

Hotel booking is common these days as traveling and tourism is not a concern anymore. This article describes the dataset of two types of hotel where bookings were done in 3 years. The structure of the data set is the same for both types of hotel with 32 variables describing multiple observations for both. Each observation represents the booking, cancellation, special request, market segment, distribution channel, average daily rate, optimal length of stay etc. Since this is hotel real data, all personally identifying information has been removed. This data set is useful in identifying the best time of a year to book a hotel room, best daily rate given to the customer, special requests done for rooms, bookings and cancellation ratio and also which hotel is preferred mostly by the customers.

INTRODUCTION

Hotel industry is a probable industry whose business is never constant. It can change suddenly and unexpectedly. The booking of a room in any hotel depends on multiple and important factors such as type of hotel, seasonality, weekdays, weekends, and many more. The given set of historical data helps the hotels to govern their bookings. With this they can plan their business and growth strategy and can also predict their future bookings.

We will be performing various methods to analyze the given set of data in order to understand the market segment of hotels, their performance, demand, special requests done by customers, country wise bookings, types of room preferred by customers, bookings for weekdays or weekends, bookings done by different channels etc. Analysis of the given set of data will help the hotels in decision making for the future.

INTEGRAL METHODOLOGY

The entire Analysis is divided into the following phases: Breakdown of Datasets, Dataset Description, Examining the null , unique & missing values, Data Cleaning, followed by Exploratory Data Analysis by and applying different models. First, we collected the dataset from the Almabetter dashboard, then we imported the dataset into google colab notebook. Thereafter we did basic data cleaning and data visualization. After visualizing the data set, we removed some unnecessary features and made it ready for analyzing the data set using different charts and plots. Next, we conduct data modeling by using Bar plot graphs, scatter plots, line plots etc. Finally, we provide the analysis results to give out a detailed vision of the relationship among the areas of interest.

DATASET DESCRIPTION

- This data set contains a single file which compares various booking information between two hotels: a city hotel and a resort hotel.
- 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. A total of 32 variables.
- All personally identifying information has been removed from the data.
- Both hotels are assumed to be located in Portugal, however their exact location and name are unknown.
- The dataset contains a total of 119390 entries.

List of original variables:

- hotel - Hotel (H1 = Resort Hotel or H2 = City Hotel).
- is_canceled - Value indicating if the booking was canceled (1) or not (0).
- lead_time - Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
- arrival_date_year - Year of arrival date.
- arrival_date_month - Month of arrival date
- arrival_date_week_number - Week number of year for arrival date.
- arrival_date_day_of_month - Day of arrival date.
- stays_in_weekend_nights - Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- stays_in_week_nights - Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.
- adults - Number of adults.
- children - Number of children.
- babies - Number of babies.
- meal - Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner).
- country - Country of origin. Categories are represented in the ISO 3155–3:2013 format.
- market_segment - Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
- distribution_channel - Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
- is_repeated_guest - Value indicating if the booking name was from a repeated guest (1) or not (0).
- previous_cancellations - Number of previous bookings that were canceled by the customer prior to the current booking.
- previous_bookings_not_canceled - Number of previous bookings not canceled by the customer prior to the current booking.
- reserved_room_type - Code of room type reserved. Code is presented instead of designation for anonymity reasons.
- assigned_room_type - Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
- 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.
- deposit_type - Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the

total stay cost; Refundable – a deposit was made with a value under the total cost of stay.

- agent - ID of the travel agency that made the booking.
- company - ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons.
- days_in_waiting_list - Number of days the booking was in the waiting list before it was confirmed to the customer.
- customer_type - Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking.
- adr - Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights.
- required_car_parking_spaces - Number of car parking spaces required by the customer.
- total_of_special_requests - Number of special requests made by the customer (e.g. twin bed or high floor).
- reservation_status - Reservation last status, assuming one of three categories: Canceled – 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.
- reservation_status_date - Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when the booking was canceled or when the customer checked-out of the hotel.

BREAKDOWN OF DATASETS

In order to go ahead for data visualization upon key factors we need to go for certain extra steps before proceeding to the main segment. In this part we are going with the following steps:

1. Importing Analytical necessary library classes for future analysis.
2. Reading the csv data file from Google drive.
3. Setting figure size for future visualization.
4. Removing future warnings in seaborn plots.
5. Visualizing all the columns of the respective Data frame.
6. Viewing all data information
7. Checking the Unique values in the column (if any).
8. Converting the data types to similar objects as the Analysis Demands.
9. Formatting the “size” column into a single column in the dataset.
10. Eradicating special characters from the dataset columns.

EXAMINING NULL VALUES

The most critical thing from which we can draw some observations is the Dataset, however data comes with unexpected values too i.e. sometimes it may be Null or missing in other words the space might be blank. Thus, at the time of analyzing the first thing which we will do is to examine the null or missing values on the Dataset. It is the

first step that will make the results "more" accurate & should be handled before it affects the performance of the models that predict the outcome. By using isnull() and sum() functions it can be seen that there are null values in 'company', 'agent', 'country' and 'children' columns. Hence, we have used several methods to eradicate those null values.

DATA CLEANING

Data cleaning is one of the most important subtasks of any data science project. Although it can be a very tedious and long process, it's worth should never be underestimated.

we have divided this process in four steps:-

1. Remove duplicate rows
2. Handling missing values
3. Convert columns to appropriate data types
4. Adding important columns

Hence, we now proceed to remove duplicate rows. First we find out the number of duplicate rows by using the duplicated() function and then remove them by using the drop() method.

For the missing values, we replace them by '0', 'Other' or any other data according to the column. We also use mean and mode to replace some missing values depending on the rest of the data in the column.

We then convert all columns to their relevant data types. For example, we convert the data type of "children" to "int64" using astype() function.

We also add any additional columns that we feel are necessary for the analysis. In this project, we added the column "total stay" to display the total stay duration of customers.

DATA VISUALIZATIONS

Here we will perform some initial analysis and visualizations and then multiple graphs, charts and maps will be displayed to get the best conclusion out of it.

Observation 1:

Fig-1

Fig-1, shows that City hotel has 61.13% of bookings and resort hotel has 38.87% of bookings in total 3 years. It is observed that customers prefer city hotels more than resort hotels.

Observation 2:

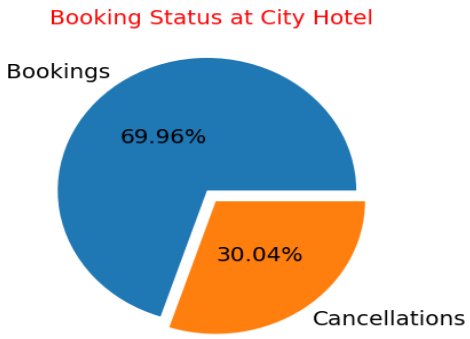

Fig-2 (a)

In Fig-2 (a), it is observed that 1/3rd of the bookings of city hotels were canceled.

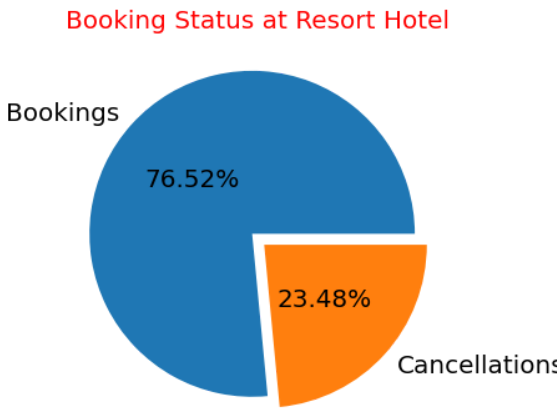

Fig-2 (b)

And Fig-2 (b) shows that at the resort hotel, 1/4th of the bookings were canceled.

So, the conclusion here is that cancellations are more in the case of city hotels.

Observation 3:

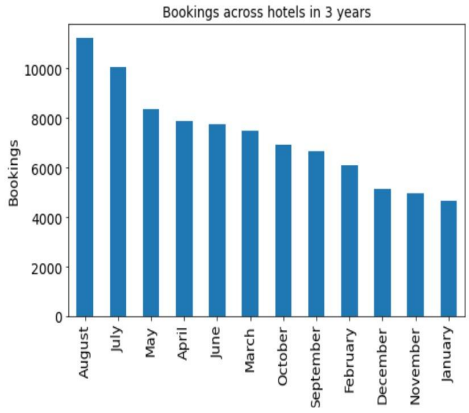

Fig-3

Fig-3 shows that most bookings are in the month of august across 3 years. While analyzing individually, it was observed that in the year 2016, August has the most bookings but in the year 2015 and 2017 the highest number of bookings are in the months of September and May.

Observation 4:

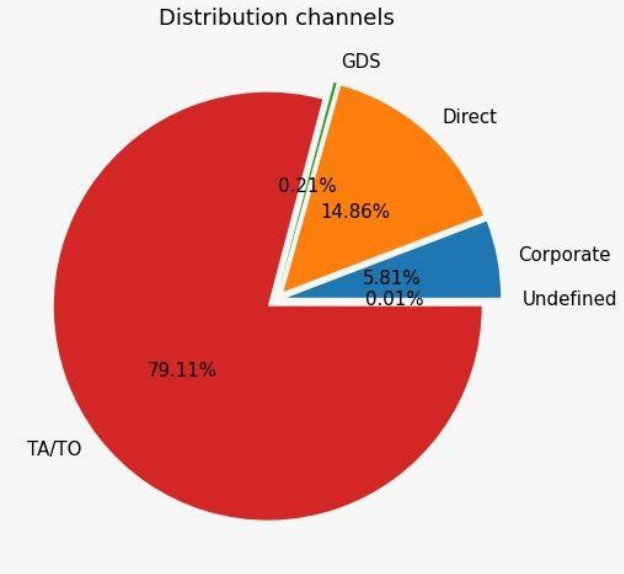

Fig-4

Fig-4 shows that most of the bookings are done through the TA/TO (Travel agents/Tour operator) distribution channel and least through undefined and GDS channel.

Observation 5:

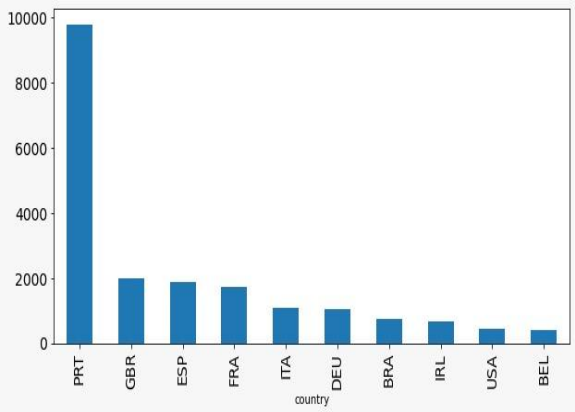

Fig-5

Fig-5 shows that Portugal has the highest number of bookings and cancellations among all countries.

Observation 6:

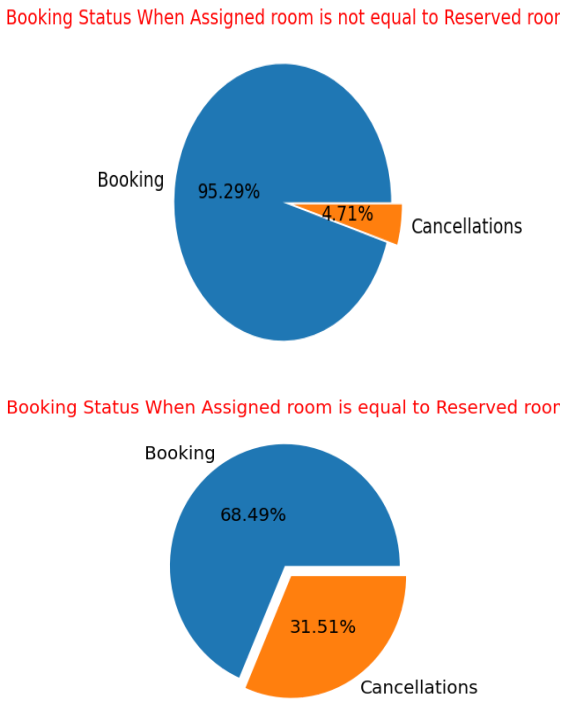

Fig-6

Fig-6 shows that cancellations are less even when the room assigned to the customer is not equal to the reserved room and vice versa.

Observation 7:

Fig-7 represents the comparison of average daily rate between both the hotels. It is concluded that the best daily rate is provided by resort hotels when the optimal length of stay increases. The variation in daily rate of city hotels is not too much.

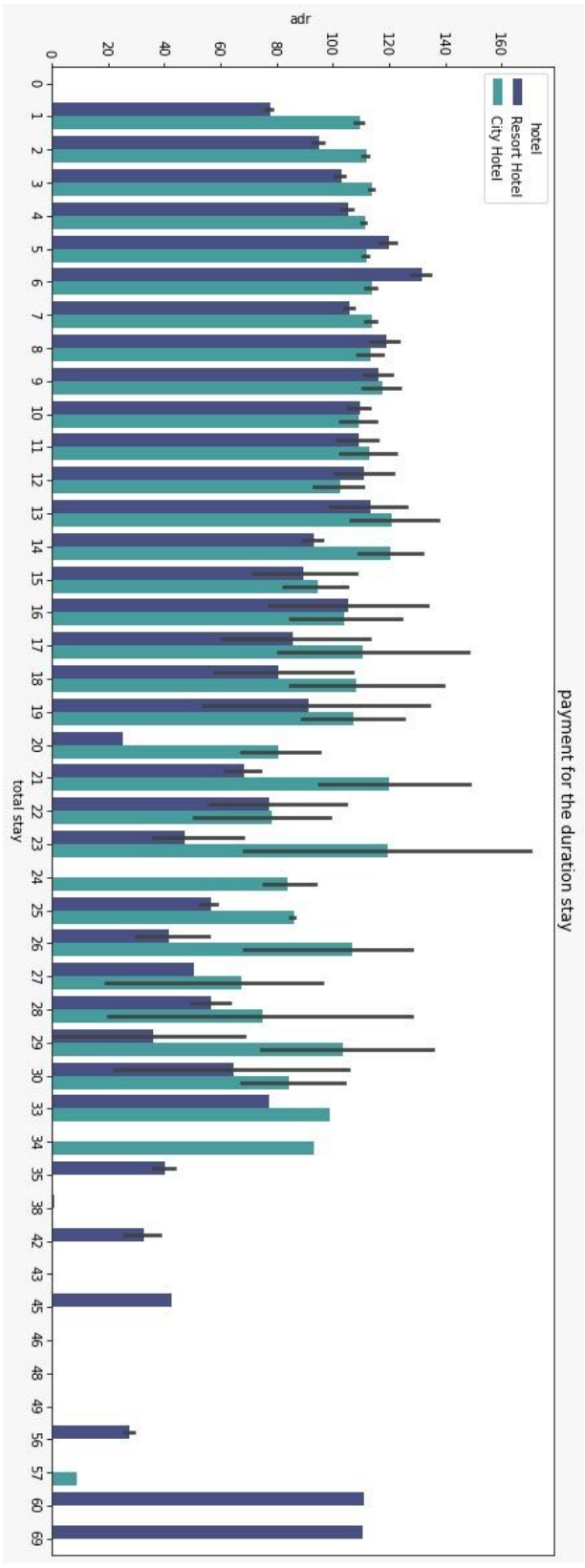

Fig-7

Observation 8:

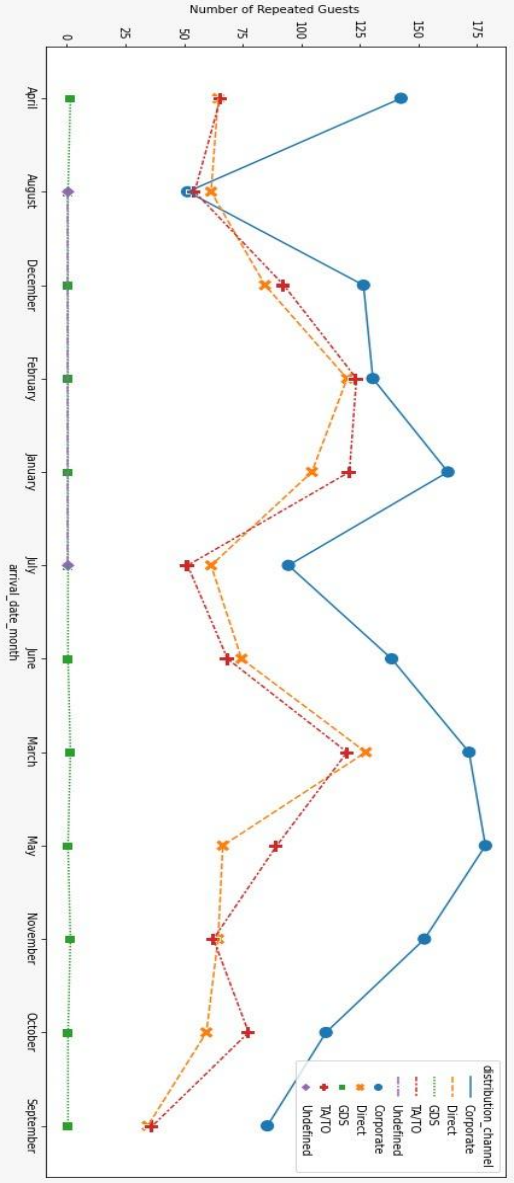

Fig-8

As per the analysis report, Fig-8 shows that most of the guests are from corporate distribution channels and least from GDS and undefined channels.

Observation 9:

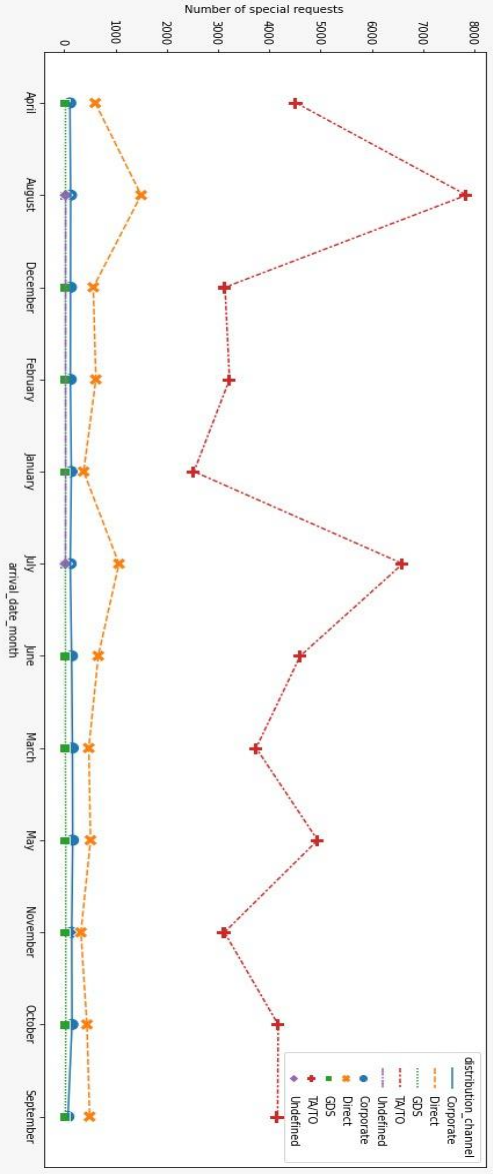

Fig-9

Fig-9 shows that guests delivered through agents and operators consistently make the highest special request across all months.

Observation 10:

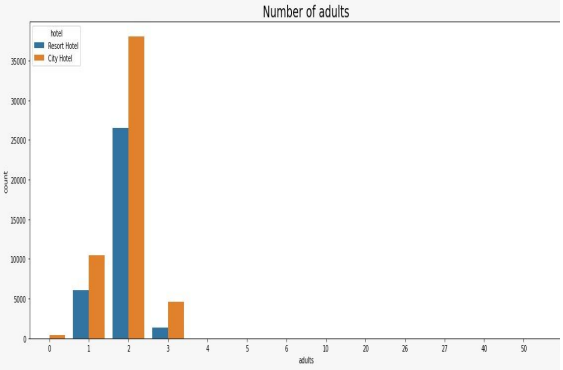

Fig-10

Fig-10 shows that most of the bookings are done by 2 adults.

CONCLUSION

Through the above analysis, we conclude the following points.

- Both of the Hotels Received Highest no. of Bookings in the Month of August in Each Year
- According to the 3 year data, the most preferred month to book a hotel room is August for all customers
- Highest No. of Customers are Coming from Europe and Mainly from Portugal
- Highest No. of Bookings are Coming from TA/TO Channel
- The best average daily rate is observed in resort hotels rather than city hotels when the optimal length of stay increases and also it is observed that the daily rate does not vary much in city hotels.
- Cancellation is highest when the customer does not deposit any amount for the booking
- Most people prefer to stay at the hotels between 1 to 6 days.

ACKNOWLEDGEMENT

This project was completed by Pravin Singh, Shubham Shahu, Rohun Rajvanshi, Shubham Jindal and Keshaw Agarwal. We are extremely grateful to our mentors and seniors who helped us and guided us throughout the project. We are blessed to have multiple social media channels through which we get great insights. We also wish to convey our appreciation to our peers who provided encouragement and timely support in the hour of need.