Capstone Project

Hotel Booking Analysis

Technical Documentation

Tanjul Gohar

Contents

Abstract

The dataset describes two types of hotels, City and Resort hotel which has total **1,19,390 rows** and **32 columns**.

We explored the Datasets and performed some visualizing techniques to analyse them.
Exploratory Data Analysis is the method of exploring the data by generating insights from them and checking all assumptions, as well as extracting underlying hidden patterns from the data with the help of various tools and graphical techniques like bar plot, histogram**,** correlation, matrices, etc. to help the hotel industry take key decisions which will help them improve their performances in market as well as profits, too.

Introduction

The hotel industry is a very volatile industry and it is the important section of the service industry that deals with guest accommodation, catering for customers who require overnight accommodation.
The precise features and services provided to guests can vary quite drastically from one hotel to another, and hotel owners generally aim to attract a particular type of customer through their pricing model and marketing strategy, or via the range of services they offer. The bookings depend on variety of factors such as type of hotels, prices, months, meal, country and many more.

This makes analysing the patterns available in the past data more important to help the hotels plan well. Using the historical data, hotels can perform various campaigns to boost the business.
Resort hotels are usually located in the mountains, on an island, or in some other exotic locations away from cities. These hotels have recreational facilities, scenery, and golf, tennis, and sailing, skiing and swimming. Resort hotels provide enjoyable and memorable guest experiences that encourage guest to repeat to the resort

Problem Statement

We will be using the data available to analyse the factors affecting the hotel bookings. These factors can be used for reporting the trends and predict the future bookings.

We will tackle this problem statement in the following steps:

Step 1: Data Cleaning
Step 2: Dealing with outliers
Step 3: Visualising the numerical and categorical univariate columns
Step 4: Bivariate analysis
Step 5: Gathering useful insights

Steps of Exploratory Data Analysis

We have to follow various procedures such as – Importing packages, data cleaning, viewing univariate variables and bivariate variables and giving useful insights.

1. Importing the required libraries

Now, here we have many inbuilt libraries that are used like NumPy as np, Pandas as pd, Matplotlib as plt, and seaborn as sns. Where NumPy and Pandas are used for data analyses while matplotlib and seaborn used in visualizing the data.

2. Loading the data into the data frame.

After importing the libraries, the next step was loading data into the data frame by the use of Pandas library. Now read the data from a CSV file into a Pandas Data Frame. We can see that the value from the data set were comma-separated.

3. Identification of data types

The data types method to identify the data type of the variables in the dataset. Before cleaning the data, let's check the quality of the data and data types of each column.

4. Drop duplicate values

We check the duplicate values and then, drop the total 31,994 duplicate values from the dataset.

4. Dropping irrelevant columns and null values

Now we have removed all the rows which contain the Null or N/A values. Let's remove some columns that we will not need, so that data processed faster.

5. Data Visualisation

Data visualization, as its name suggests, is to observe the data using various types of plots, graphs etc. Various plots include histogram, scatterplot, boxplot,

heatmap etc. We will use Matplotlib and Seaborn together to visualize a few variables.

Now, we take a look at some insights of the dataset using data visualisation.

Data Summary

After the exploratory data analysis, we found that the data was pretty much clean except for some missing values in a few columns. Upon using the info () method, we found following information: -

1. The dataset has a shape of (119390, 32) which means that it contains approximately 1.2 lakh rows and 32 columns.
2. Our Dataset has 4 columns with float64 dtype, 16 columns with int64 dtype, and 12 columns with object dtype.

3. In our Dataset, we observed null values in the following columns:

- 4 null values in the children column
- 488 null values in the country column
- 16,340 null values in the agent column
- 112,593 null values in the company column

We have the following column names provided to us in the dataset and their description:

- **hotel** - hotel type(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:

    1. Undefined/SC – no meal package
    2. BB – Bed & Breakfast
    3. HB – Half board (breakfast and one other meal – usually dinner)
    4. FB – Full board (breakfast, lunch and dinner)

- **country** - Country of origin. Categories are represented in the ISO 3155–3:2013 format.
- **market_segment** - 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 cancelled by the customer prior to the current booking.
- **previous_bookings_not_canceled**- Number of previous bookings not cancelled 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. (Measures the average rental revenue earned for an occupied room per day. The operating performance of a hotel or other lodging business can be determined by using

the ADR. Multiplying the ADR by the occupancy rate equals the revenue per available room.)

- **required_car_parking_spaces** -Number of car parking spaces required by the customer

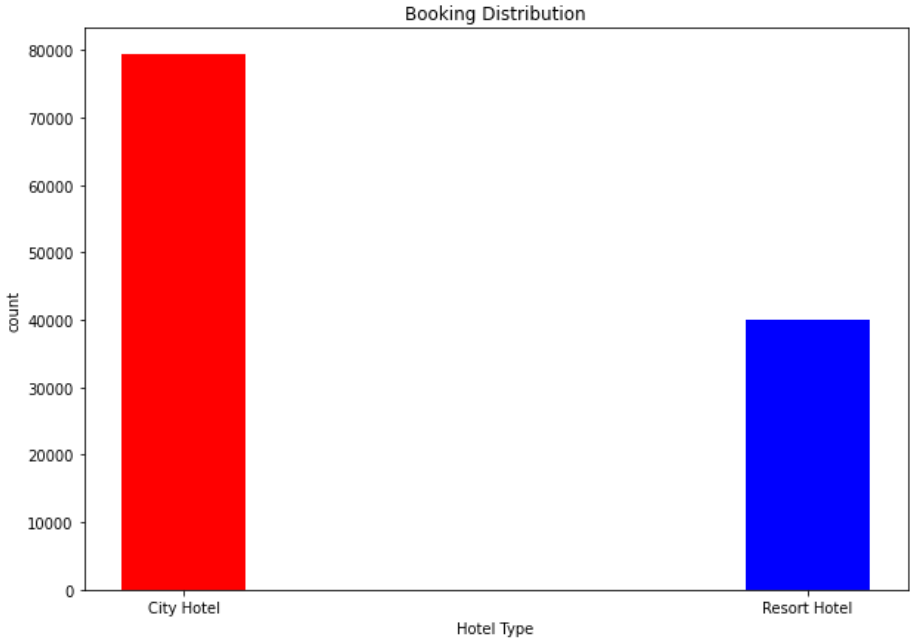
- **total_of_special_requests** -Number of special requests made by the customer (e.g. twin bed or high floor)
- **reservation_status**- Reservation last status, assuming one of three categories:
  - Canceled – booking was cancelled by the customer
  - Check-Out – customer has checked in but already departed
  - No-Show – customer did not check-in and did inform the hotel of the reason why.

- **reservation_status_date**- ate at which the last status was set. This variable can be used in conjunction with the Reservation Status to understand when the booking was cancelled or when did the customer checked-out of the hotel.

- To analyze the market precisely, we have planned to bifurcate the analysis into a set of questions on which we would work up on.

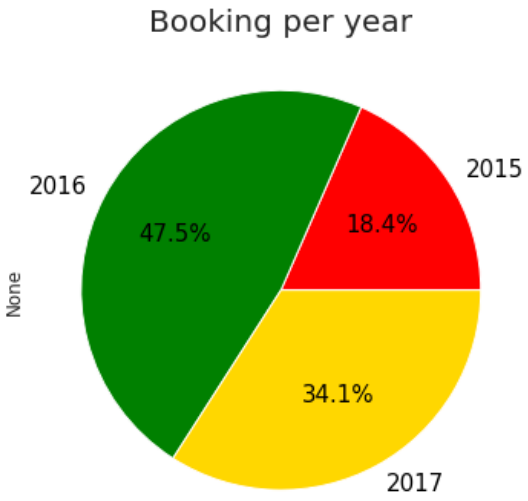
BOOKING ANALYSIS

➢ What is the booking distribution for each hotel?



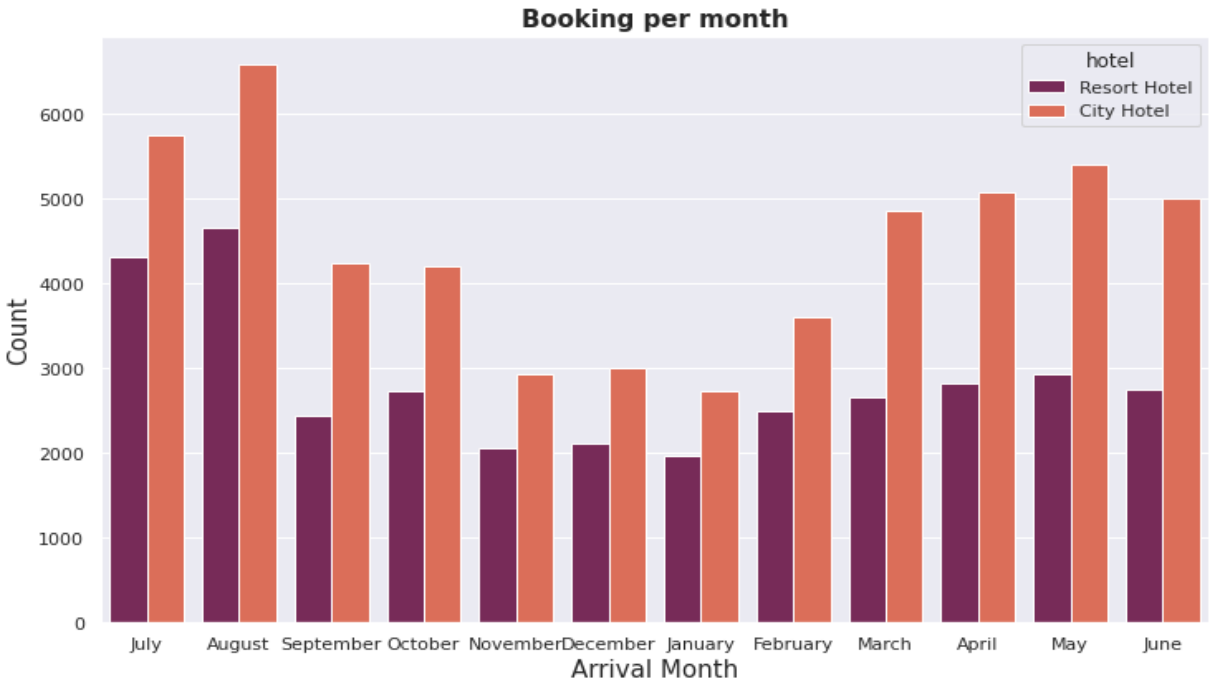
- From the above bar graph, we can conclude that the number of bookings in the City Hotel was around 79330 & for Resort Hotel are around 40060.
- The maximum booking was present in a city hotel that means most of the customers prefer to book city hotels because city hotels would be cheaper than Resort hotels
- Resort hotels need to work on their market strategy to reduce the prices.

➢ Which year had the most bookings?



- The above pie charts show the maximum percentage of different years in which customers preferred to come to both the hotels.
- The number of bookings was higher in 2016 with 47.5%.
- Number of bookings declined after the year 2016, 2015 it was 18.4%, and in 2017 it was only 34.1%.

➢ Which hotel has most number of guests & in which months?

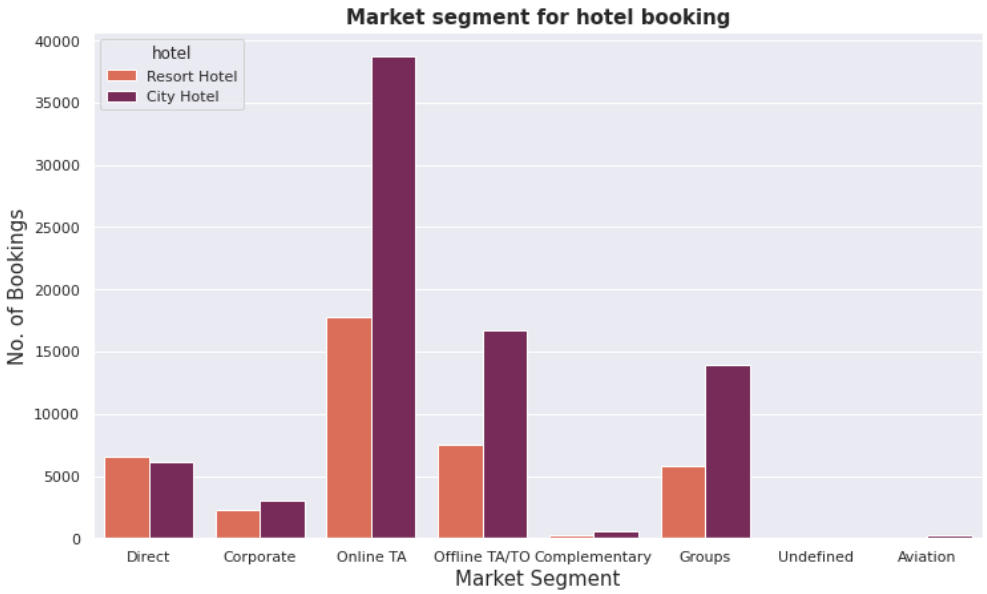


- The above bar charts show the months in which the maximum customers were coming in both the hotels.
- August has maximum bookings with 6,600 reservation counts and, after that July and may have above 5000 counts. It means that the rainy season was preferred, by most of the guests.
- In other seasons try to incorporate some seasonal packages with a discount so that hotel doesn't have to suffer in the rest of the months.

➢ Which Distribution channel contributes the highest to the bookings:

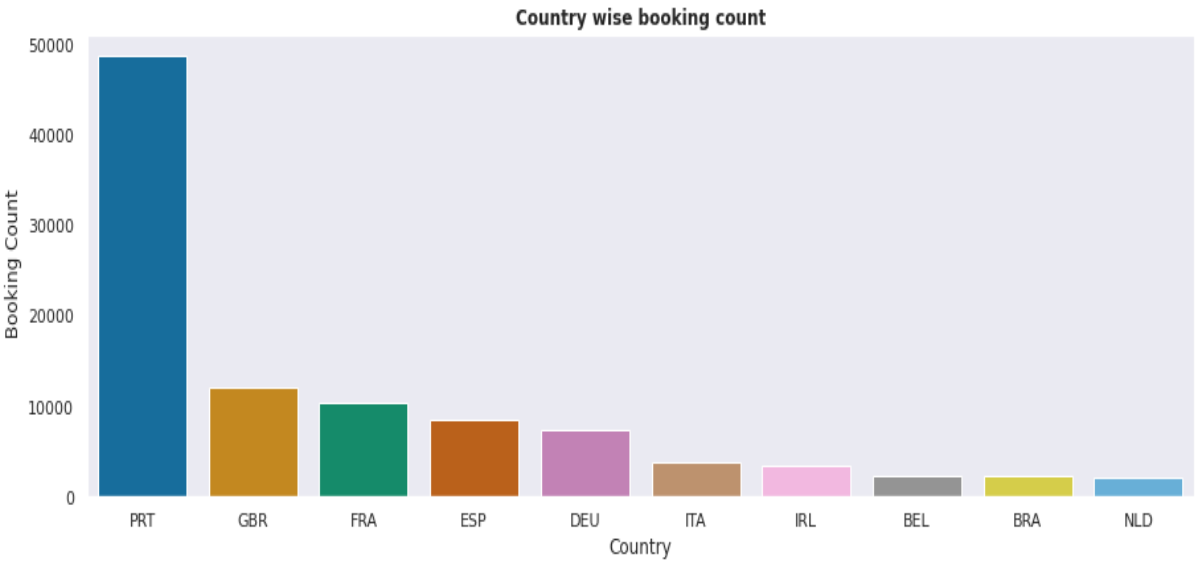
- We can see that TA/TO (Travel Agents/Tour Operators) channels have a huge booking than all other channels with 70000 counts.
- Whereas, **the Direct** distribution channel contributes only under 10000 bookings.
- We can say from the above graph analysis that people prefer bookings by the travel agents and tour operators than a direct booking, this could have happened because they give a discount.
- Hotels should focus on direct bookings by giving some discount or some extra services it helps to catch the more people and high bookings for another next year.

➢ Which Market segment contributes the highest to the bookings:



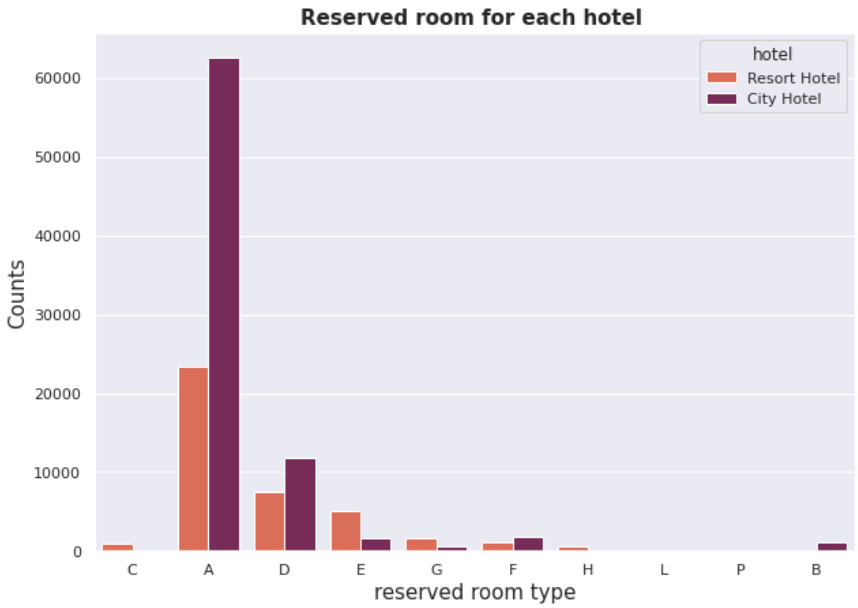
- We saw in the above graph, so here we can say that online TA/TO is the most preferred market segment by the customers. It leads more revenue than others.
- Hoteliers focus on more cost-effective marketing strategies by **Direct** market segment or direct interaction with guest to ensure a more prosperous future for business.
- There are no agents or other distribution partners that must be paid a commission when a guest book directly online.

➢ Top 10 countries with maximum bookings:



- From the above graph, we conclude that Portugal (PRT) was the country with the highest bookings after that we have Great Britain "GBR".
- We should mainly focus on the key aspects of lifestyles of these countries to make our travelers happy and increase repeating customers as most of our travelers were coming from the above countries only.
- Hotels should run tourism business by team up to Tourism Company and drive more traffic by the tourisms.

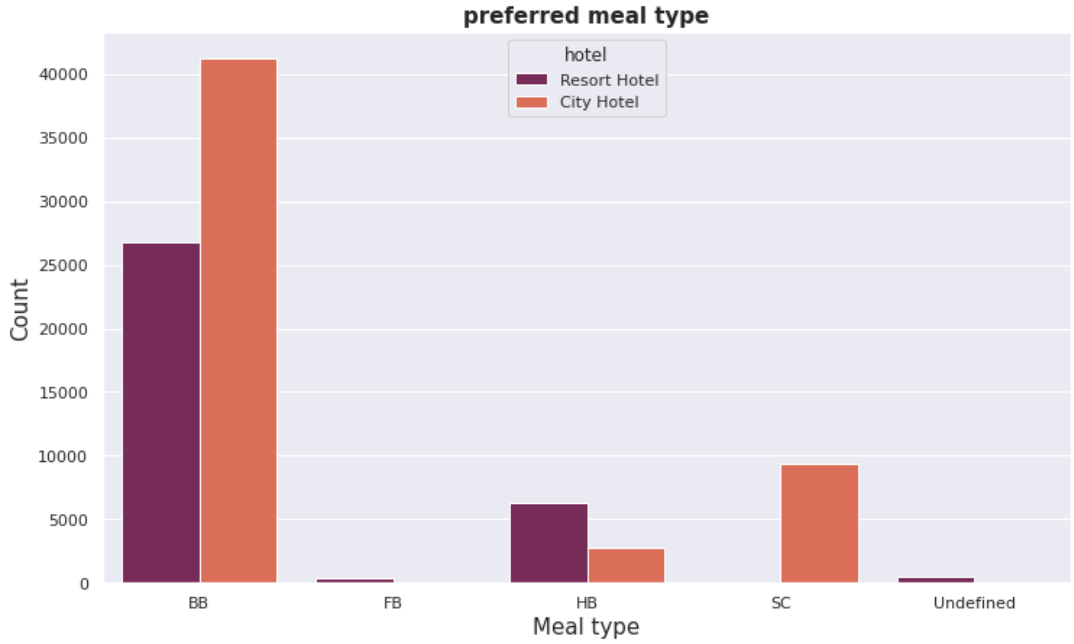
➢ Reserved and assigned rooms for each type of hotels



- Reserved rooms are A, B, C, D, E, F, G, H, L. The most preferred room type was A and after that D was a little more.
- The maximum bookings of around 66,000 were reserved for room type A, followed by 12,000 reserved for room type D.

- The Types of Assigned rooms were A, B, C, D, E, F, G, H, I. The maximum bookings were assigned for the room type A above 50,000 out of 66,000 of the reserved room types for A.
- From the above each graph shows that maximum number of customers get assigned room which they reserved.
- Here we can say room type A is so lower in price with some good facilities, on the other hand, the least booked room has very expensive and luxurious.
- We can make packages with the least booking room and include some type of external activity. Guests want to experience these and will be paying for these amenities.

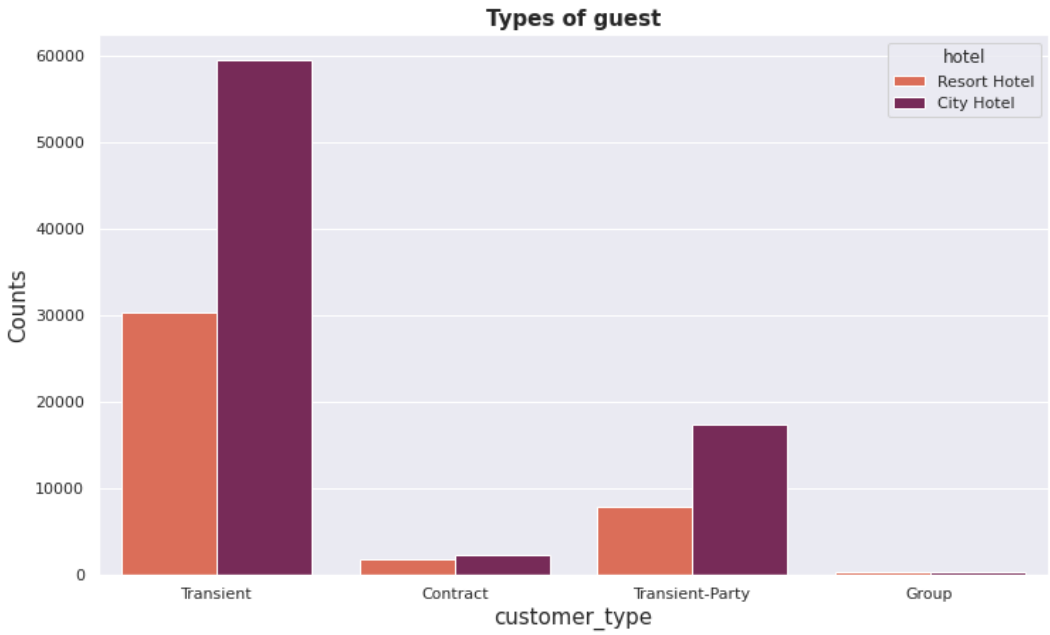
➢ **What were the different kinds of meals that the corporates prefer in their visit?**



- ❖ 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)

- People generally opt Bed and Breakfast instead of full-fledged meals.
- The above graph states that, BB (Bed Breakfast) type meal was most preferred in both types of hotels.

➢ Type of customers in both hotels:

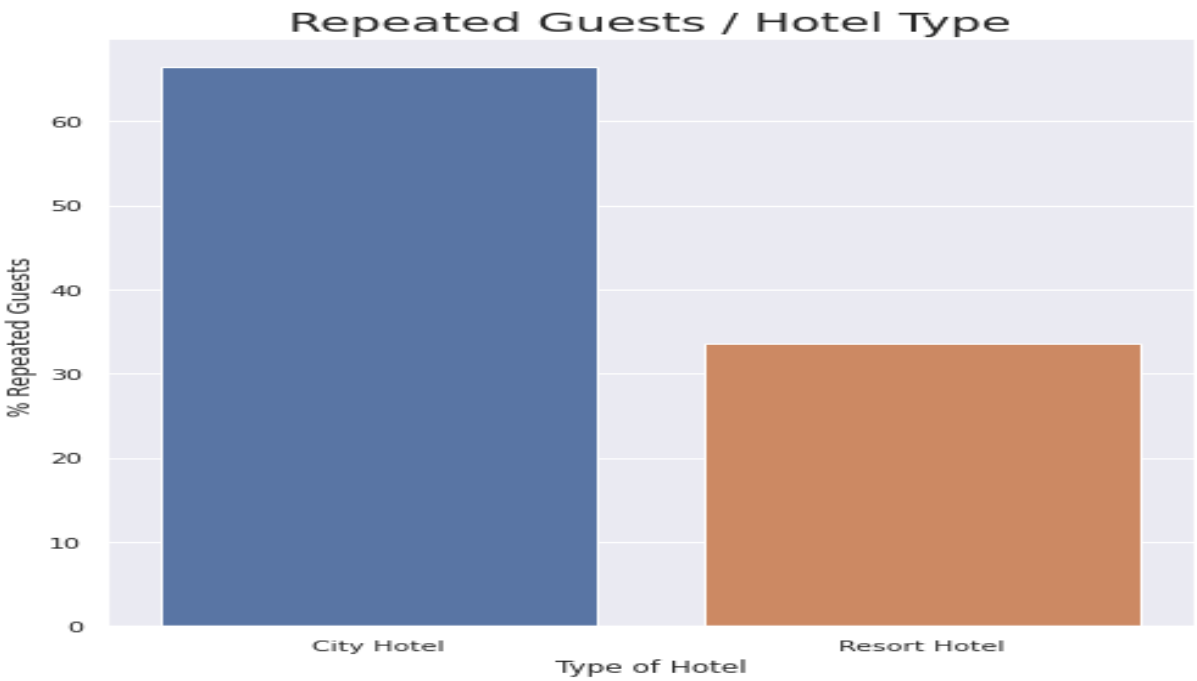


- The graph clearly shows that the Transient type of customers was coming in a maximum amount in both the hotels.

- Transient customer means a customer who stays for a very short period likes business people. Such guests only need basic services like a clean and comfortable room for the night, and a nice breakfast in the morning, laundry and pressing not much more than this.

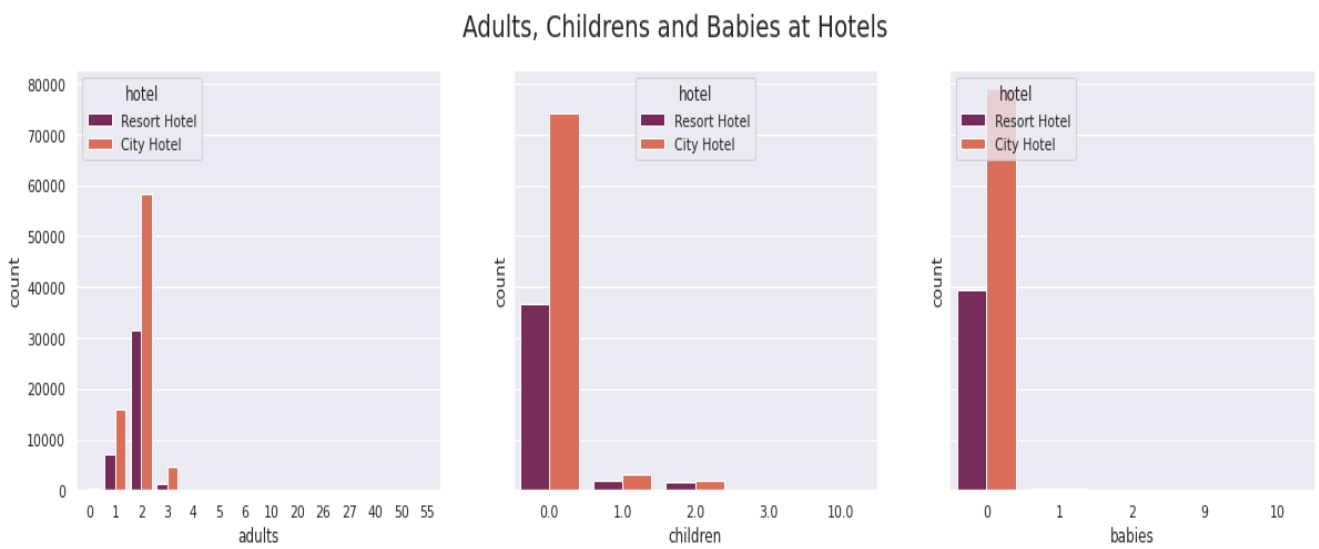
- Group type of customers was least like families and couples. Never ignore group type customers. Hotels need to work on this group of people by occupying kids' activities, pet friendly and can give personalised packages too like couple's packages with some room service discount.

➢ How many guests prefer to visit repetitively?



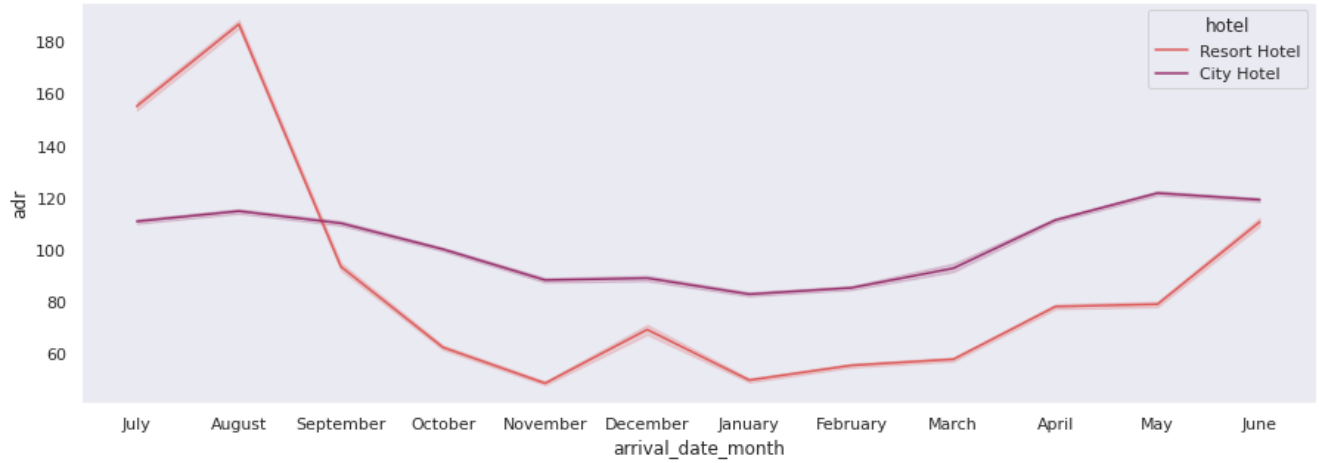
- City hotel has around two times more repeated guests as compared to the resort hotel.
- It was easier and less expensive to target repeat guests than to generate new leads, for this, we can offer some discounts and special package deals.

➢ How many kids do we have the most among the hotels?



- City Hotel has more counts for children, babies, and adults than the resort hotel.
- Family has slightly different needs. Hotels can prepare a list of family-friendly restaurants, child- safe environment, play areas, larger rooms with some discounts.

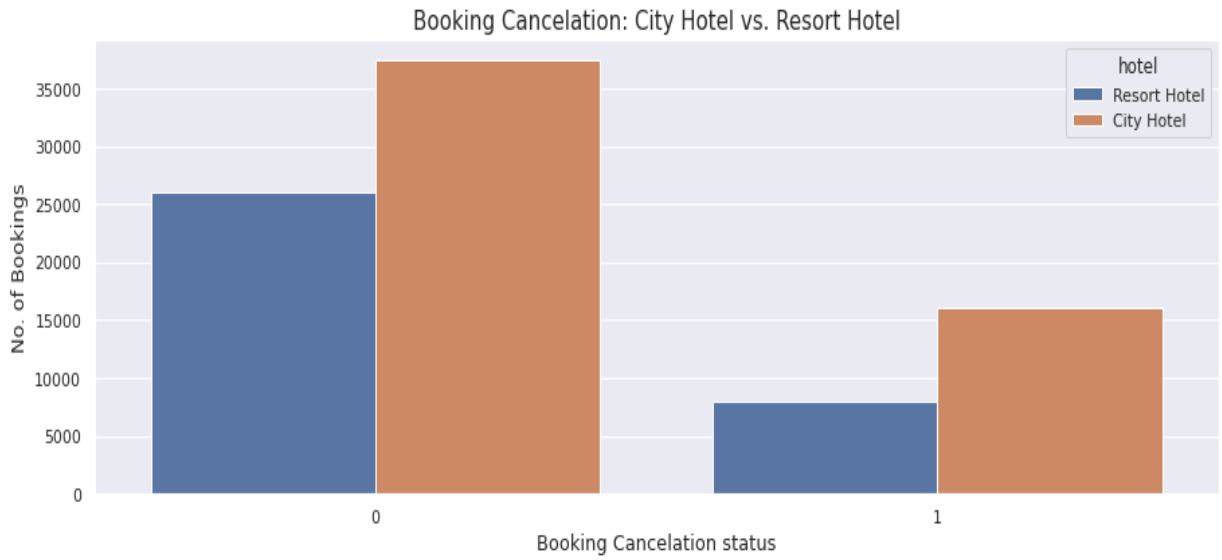
➢ Variation of the ADR (Average Daily Rate) according to month



- The average daily rate (ADR) shows how much revenue was made per room on average.
- The above plot concludes that the ADR was highest during **August** and **July** in the resort hotel. On the other hand, **city hotel** has maximum ADR was present in **May**, **June**, **August**.
- To generate revenue for the rest of the month, the hotel focusses on pricing strategies include complimentary offers.

Key: AM-ADR = Average monthly ADR, ADR = Average daily rate

➢ How high is the cancellation rate between the two hotels monthly?



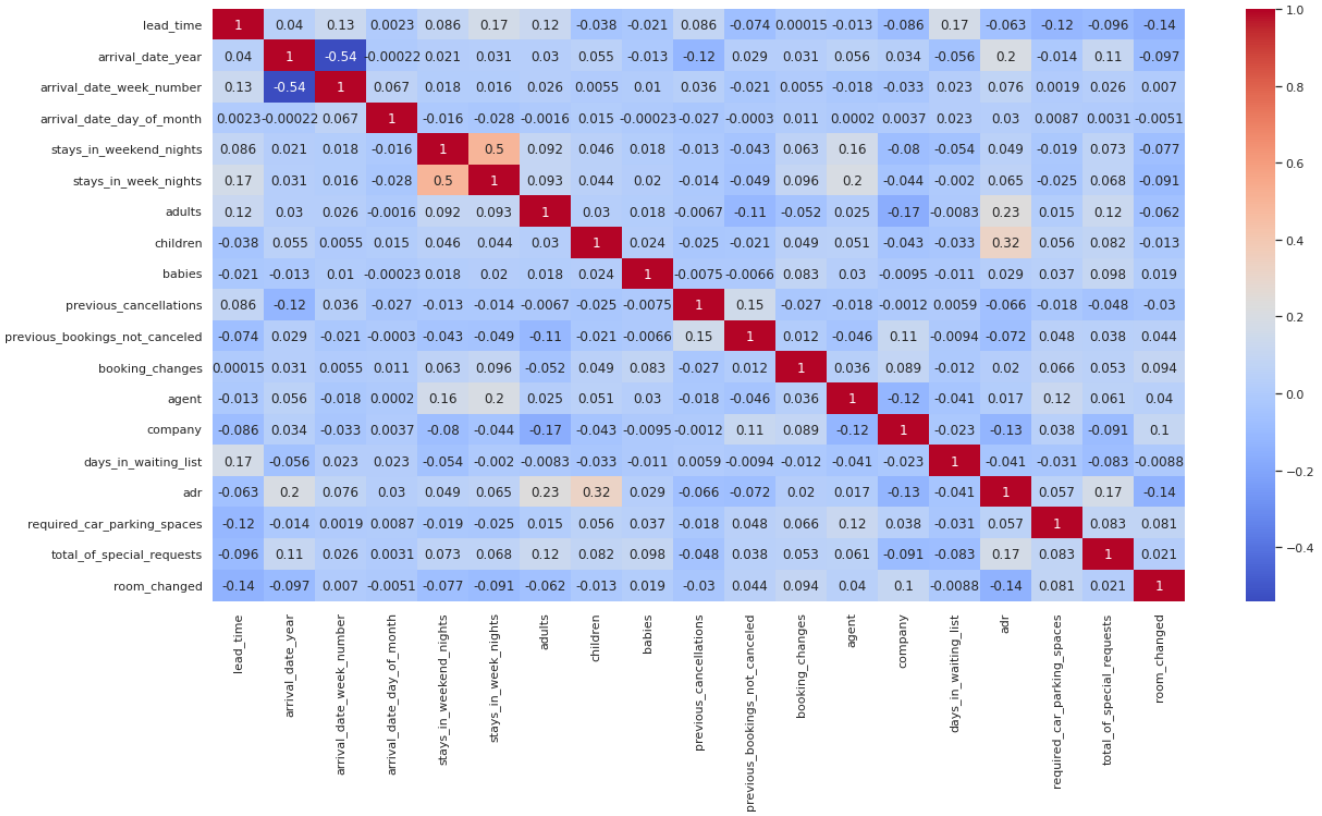
From the above analysis we can concluded that:

- City hotel has a quite large rate of cancellations than that of the resort hotels.
- Although the total bookings were more in city hotels the rate of confirmed bookings was more in resort hotels.

➢ Correlation between the all variables:

At first, I was doing a correlation analysis to understand the level of correlation between all variables. One of the best ways to find the relationship between the features can be done using heat maps.

- Positive correlation was represented by dark shades
- Negative correlation was represented by lighter shades.

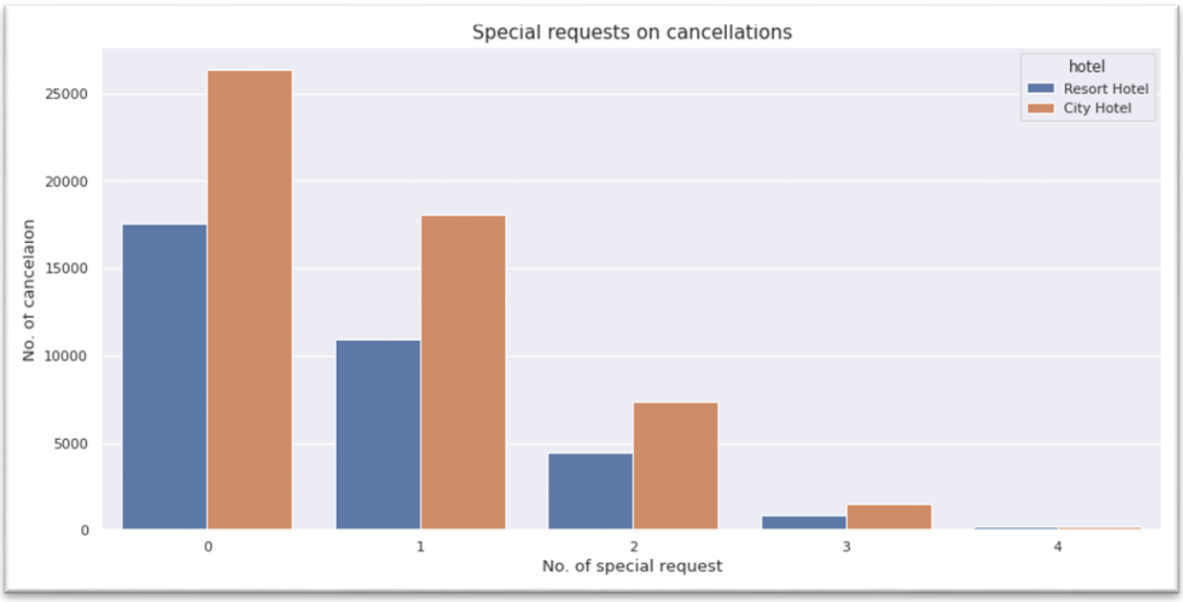


Upon Observing the above heat map, we observe the following things: -

- From heatmap, positive correlation of total number of special requests with adr is highest (0.17), then with adults (0.12) & then with arrival year (0.11).

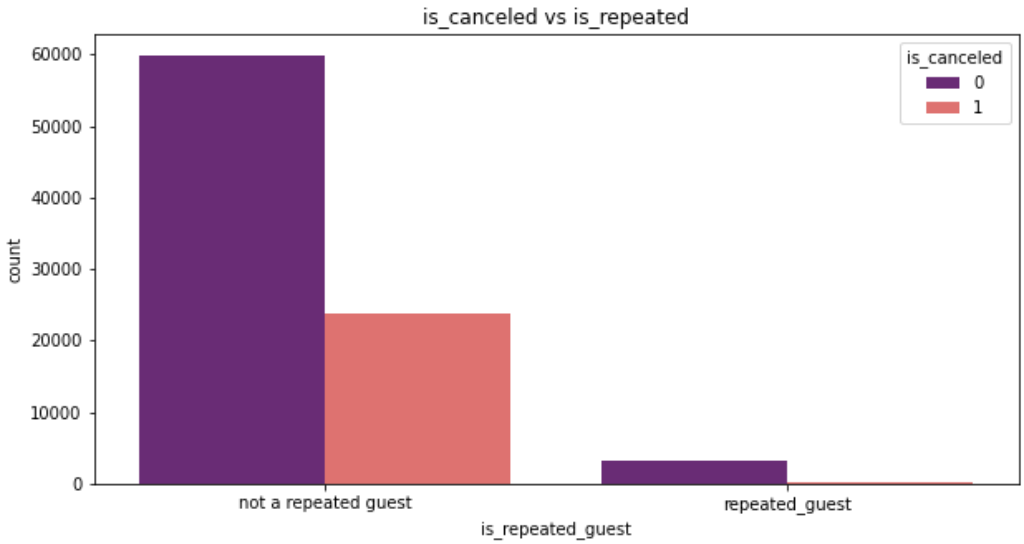
- So, hotel may receive high number of special requests with high adr, adults and arrival year.
- lead_time was the variable having some significant level positive correlation 0.17 with days_in_waiting_list & stays_in_week_nights
  then 0.08 with previous_cancellations.

➢ Impact of special requests on cancellations



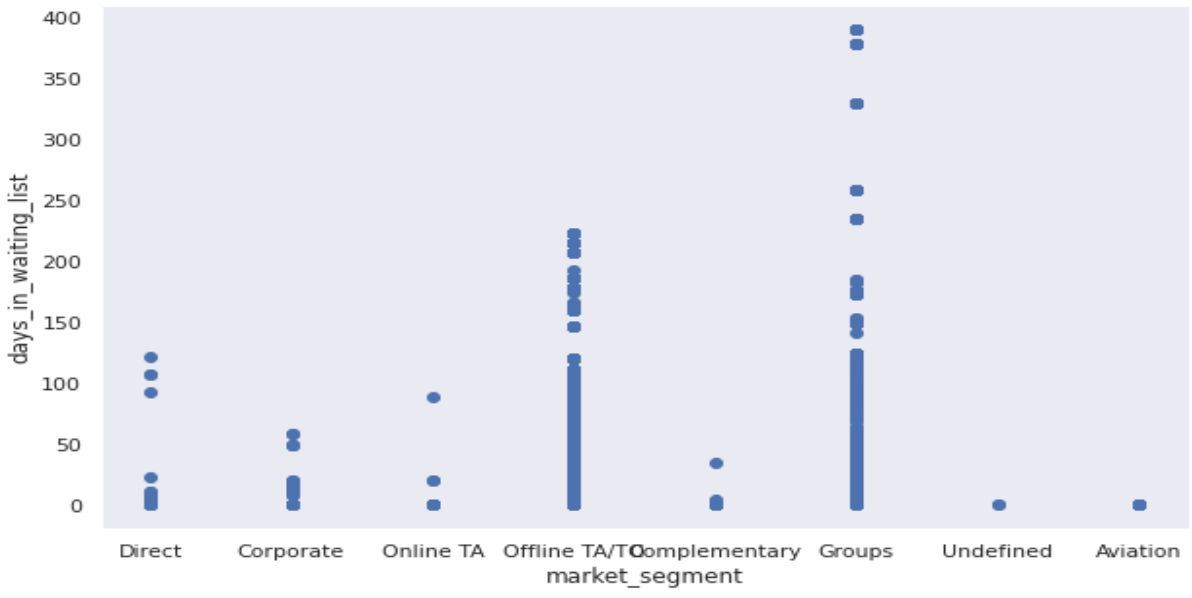
- Guest with zero special requests was more likely to cancel the reservation.
- Special requests with 2, 3, 4 numbers have almost zero cancelations.
- Therefore, more the number of special requests less than the number of cancellations.

➢ Variation of cancellations according to repeated guest



- The above graph shows that the maximum number of bookings were cancelled, by the guests who have booked the hotel for the first time.
- We see that when the hotel booking, was cancelled and the customer was a repeated guest then the entry was almost zero which means that the repeated guest was less likely to cancel his booking.

➢ Market segment having least number of days on the waiting list.



- A customer of the aviation market segment has minimum waiting time as compared to the other market segments.
- On the other hand, market segment group customers have the maximum number of waiting time of nearly about 400 days.
- Aviation industry has the minimum days on waiting list. The reason for this might be that when the flight has to land at a location it should provide immediate accommodation to all of his working staff (like pilot, cabin crew).
- Hotel management make immediate arrangement and provide rooms for the group type of customers with almost zero waiting lists.

Some of the future predictions we can make from the above data visualizations so as to help the hotel industry take key decisions and to improve their efficiency, performance and profits.

- Online booking cancellations was more, encourages direct bookings by offering special discounts.
- Foreign countries like Portugal, Britain, and France have more visitor base; apply marketing team to handle that customer base.
- Hotels should consider the maximum number of special requests from guests to reduce the possibility of cancellations which will eventually help in better customer experiences.
- Months between May and August have maximum bookings; these months were peak in business expansion and look after more customer satisfaction.
- Based on the results of our EDA, hotels can plan on targeting the new customers to increase spends and maintains good relations with the existing customers.
- The insights on the meal and room preferences can help them to price the commodities better.

Closing remarks:

- From the above analysis, we concluded that the hotel business tends to make more profit during the rainy season because most of the bookings were in August, July, and May. So, the hotels can keep experimenting with their packages, interiors, and amenities as there were more new guests than the repeated guests.
- Even though the percentage of total bookings was more in city hotels as compared to resorts but the number of cancellations in city hotels was also maximum in resorts.
- City Hotel welcomes more children than the resort one, so they can add a separate kid's section, which will not only attract more kids but also increase the number of repetitive guests over time.
- Guest of Portugal country, which can help us build more Portuguese friendly dishes for the premium packages.
- The corporates mostly prefer the Bed and Breakfast package, due to which the hotel can add more features to the package, also can profit from the customization feature.
- An overall analysis suggests that the business aspect of the City Hotel venture was much more profitable than that of a Resort one, but if the investor sees potential in the business and was ready to invest more, a Resort Hotel will be a valuable add-on.
- In the end, we can see that the hotel industry was facing an imminent challenge of rising cancellation rates and a steep decline in the number of bookings from the last few years. So, using the actionable insights would help the marketing manager to better steer the start up in the hotel industry market of this region.