Hotel Booking Analysis

Neeraja C, Nabeela P B, Suneel

Abstract

Hotel industry is a very volatile industry and the booking depends on many factors. The main objective behind this project is to explore and analyse data to discover important factors that govern the bookings and give insights to hotel management, which can perform various campaigns

to boost the business and performance.

1.Problem statement

We are provided with data set contains booking information for a city hotel and a resort hotel.

We had analysis data on following questions

- Which is the busiest month?
- Which distribution channel is mostly preferred?
- From which country most of guests come?
- How long people stay in hotel?
- Which is the most booked accommodation type (Single, Couple, Family)?
- Which Market segment is mostly used?
- Which is the busiest year?
- How many repeated guests are coming?
- Cancellation of booking
-

2.Data description

The data set consists of 119390 rows and 32 columns. We can

start our analysis by defining each column.

hotel: The names of the hotel are City Hotel and Resort Hotel
is_cancelled: Cancellation type, if the booking was cancelled or not.

Which takes 2 values 0 and 1.

0 indicates not cancelled.

lead_time: Time between reservation and actual arrival.
arrival_date_year: Year of arrival date
arrival_date_month: Month name of arrival date.
arrival_date_week_number: Week number of arrival date
arrival_date _day_of_month: Day of the month of arrival date
stays_in weekend_nights: Number of weekend nights the guest stayed or booked to stay at the hotel
stays_in_week_nights: Number of week nights 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

country: Country of origin of customer

market_segment: Market segment designation.

In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”.

distribution_channel: The medium through booking was made

is_repeated_guests: Value indicating if the booking name was from a repeated guest (1) or not (0)

previous_cancellations: (0 or 1) Indicates whether or not the guest has previous cancellations

previous_bookings_not_canceled: (0 or 1) Indicates whether or not the guest has previous bookings which are not cancelled.

reserved_room_type: Code of room type reserved.

assigned_room_type: Code of room type assigned.

booking_changes: Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation

deposit_type: Whether refundable/non-refundable/no-deposit made

agent: ID of the travel agency that made the booking

company: The ID of company that made the booking

days_in_waiting_list: number of days the booking was in the waiting list before it was confirmed

customer_type: Type of customers (Transient, group, etc.)

adr: Average daily rate is the average revenue that a hotel receives for each occupied guest room per day

required_car_parking_spaces: Number of car parking spaces required

total_of_special_requests: Number of special requests made

reservation_status: Reservation last status

reservation_status_date: Date of last reservation status

3. Cleaning Data

Before analysing the data we have to remove the ambiguous data that can affect the outcome of EDA(Exploratory Data Analysis).

While cleaning data we will perform following steps:

1) Removing the duplicate rows and values
2) Handling missing values and null values
3) Convert columns to appropriate data types.
4) importing the imprtant columns

4.Steps involved:

4.1. Data wrangling:

After loading the dataset, it is essential to get the right data to be organized for analysis. Data Wrangling is the process of removing errors and combining complex data sets to make them more accessible and easier to analyse. Data Wrangling enable us to tackle more complex data in less time, produce more accurate results, and make better decisions.

4.2. Null Value Treatment:

Our data set contains a some number of null values. Here we are replacing null values by appropriate values in order to produce more accurate and best results.

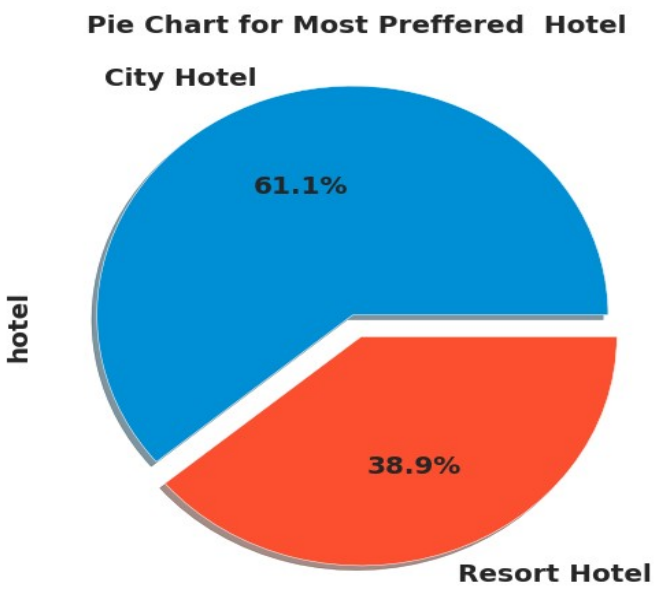
4.3. Exploratory data analysis:

After cleaning the dataset, we performed this method by comparing our target variable with other independent variables. This process helped us figuring various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

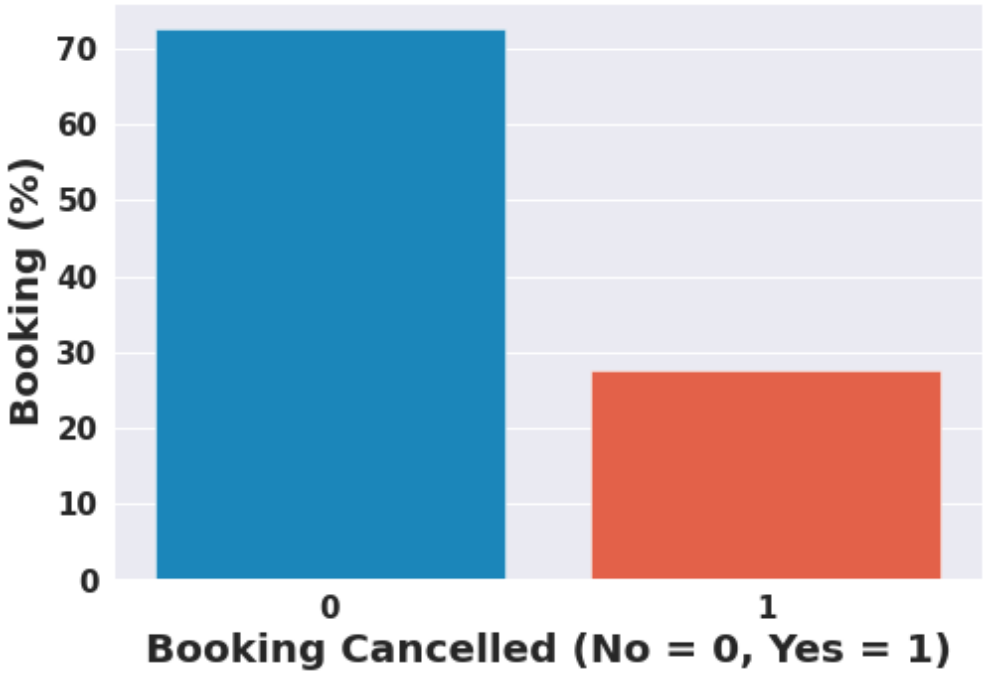
Here we are using Matplotlib and Seaborn libraries in the aspects of visualizing the following graphs and plots had been used to understanding and analysing the data.

- Bar Plot
- Histogram
- Pie Chart
- Line plot
- HeatMap
- Box Plot

Hotel Types

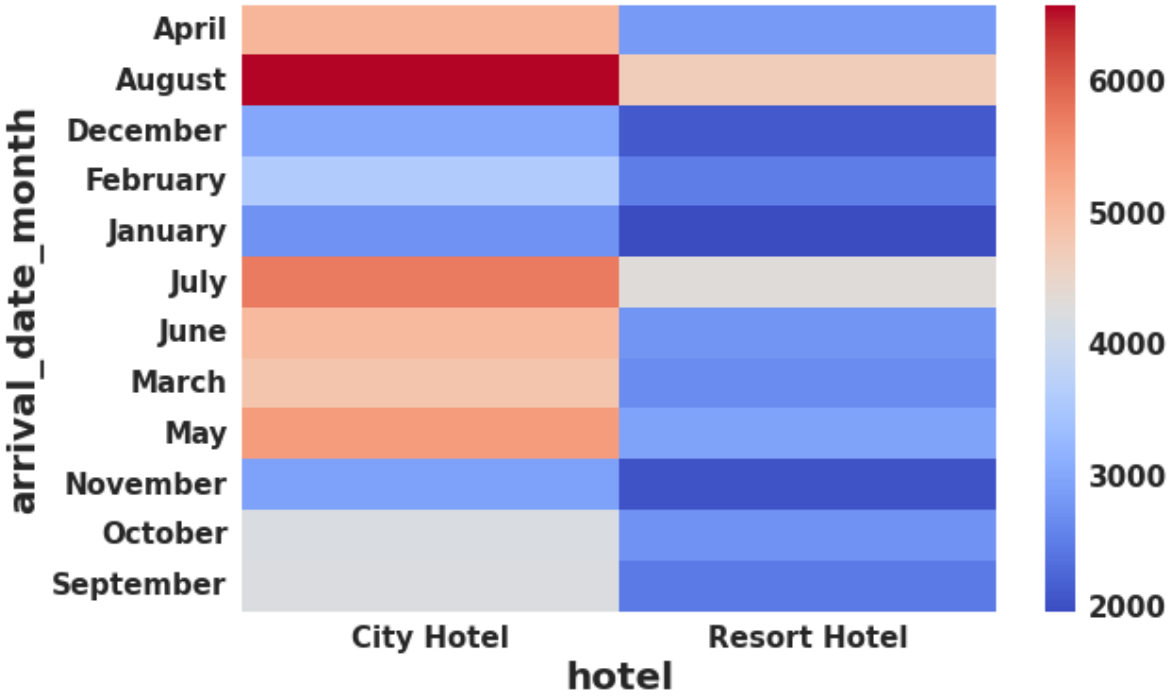


Cancelled Bookings

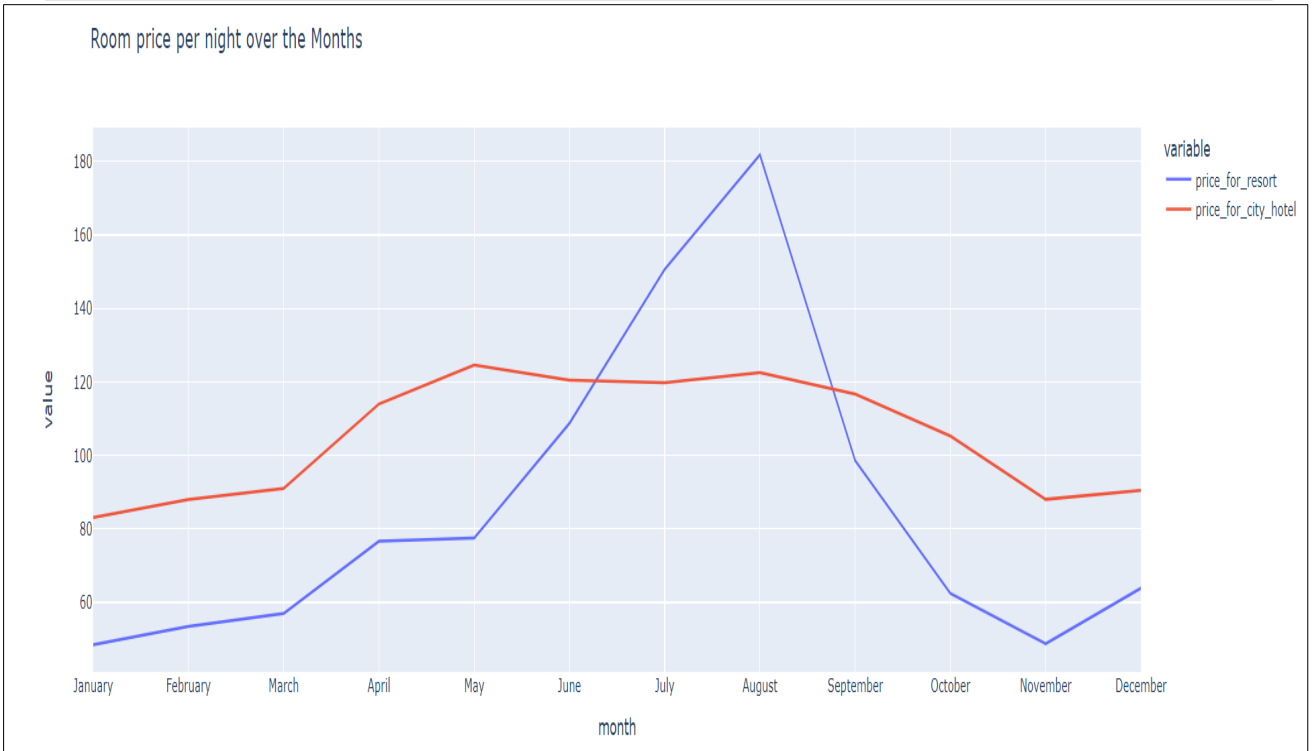


More than 70% of the people did not cancel the booking. City hotel has more number of cancellation compared to Resort hotel

Busiest Months

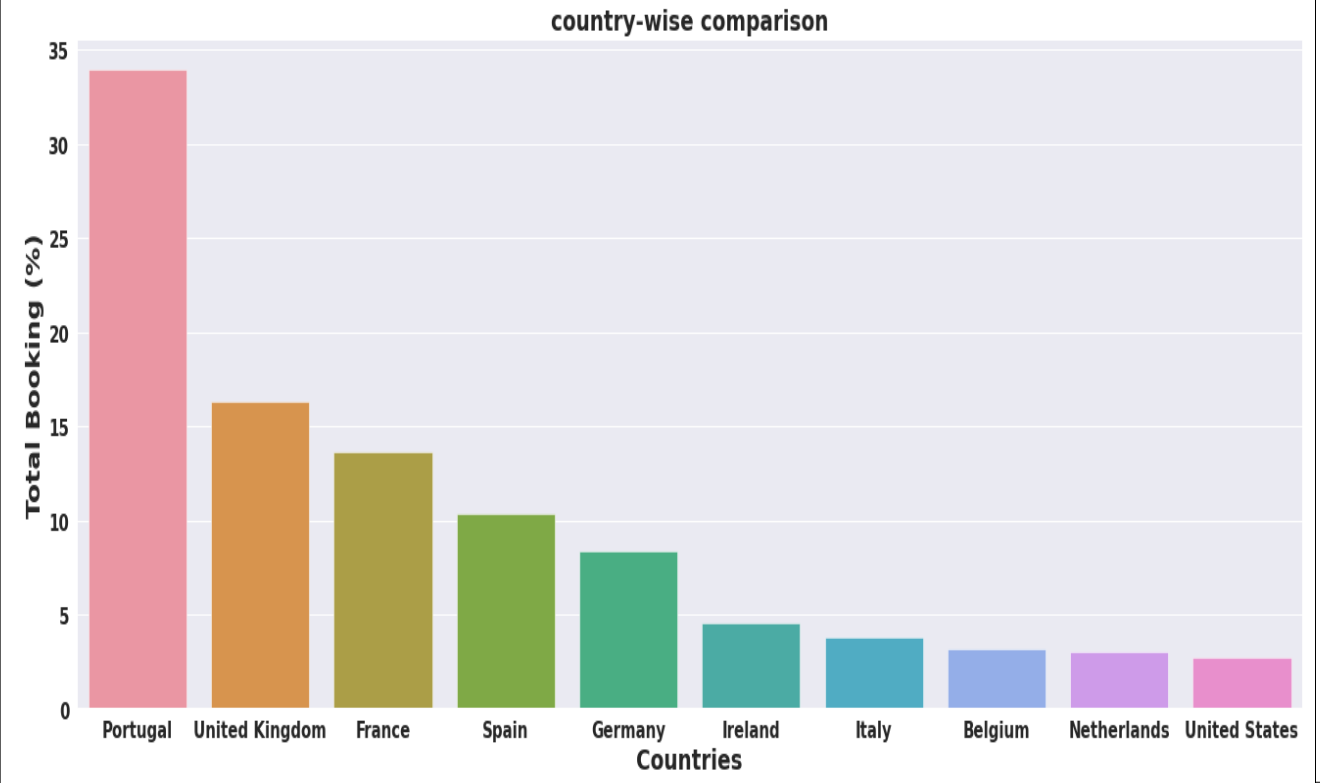


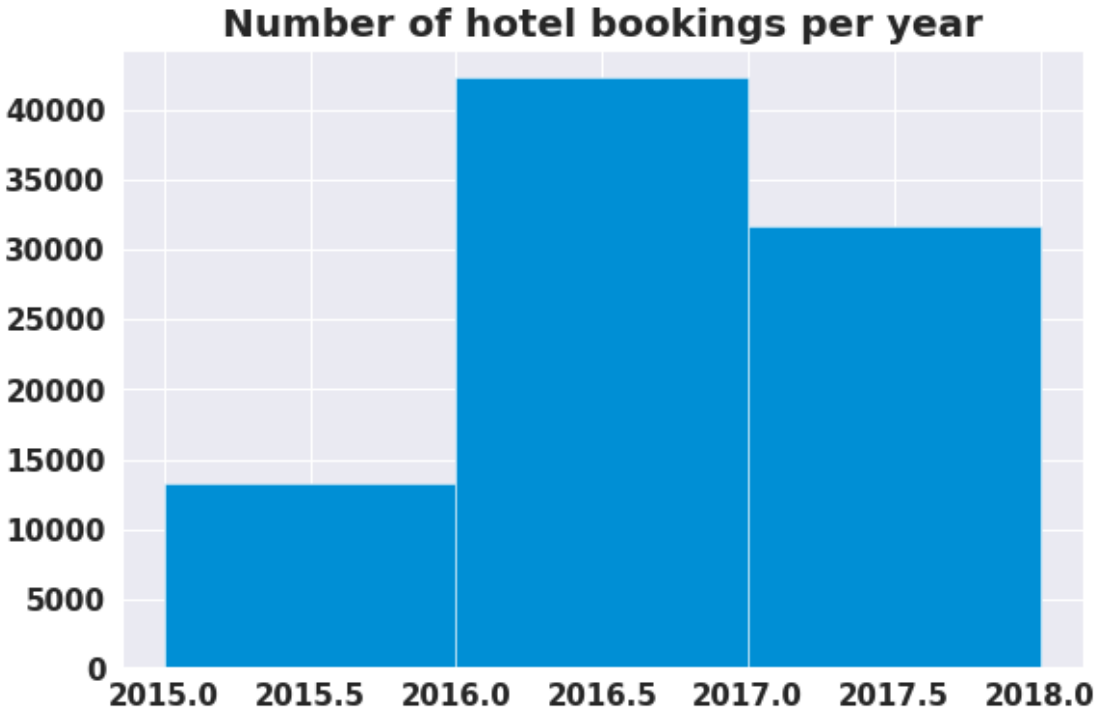
August is the most demanded month of the year.



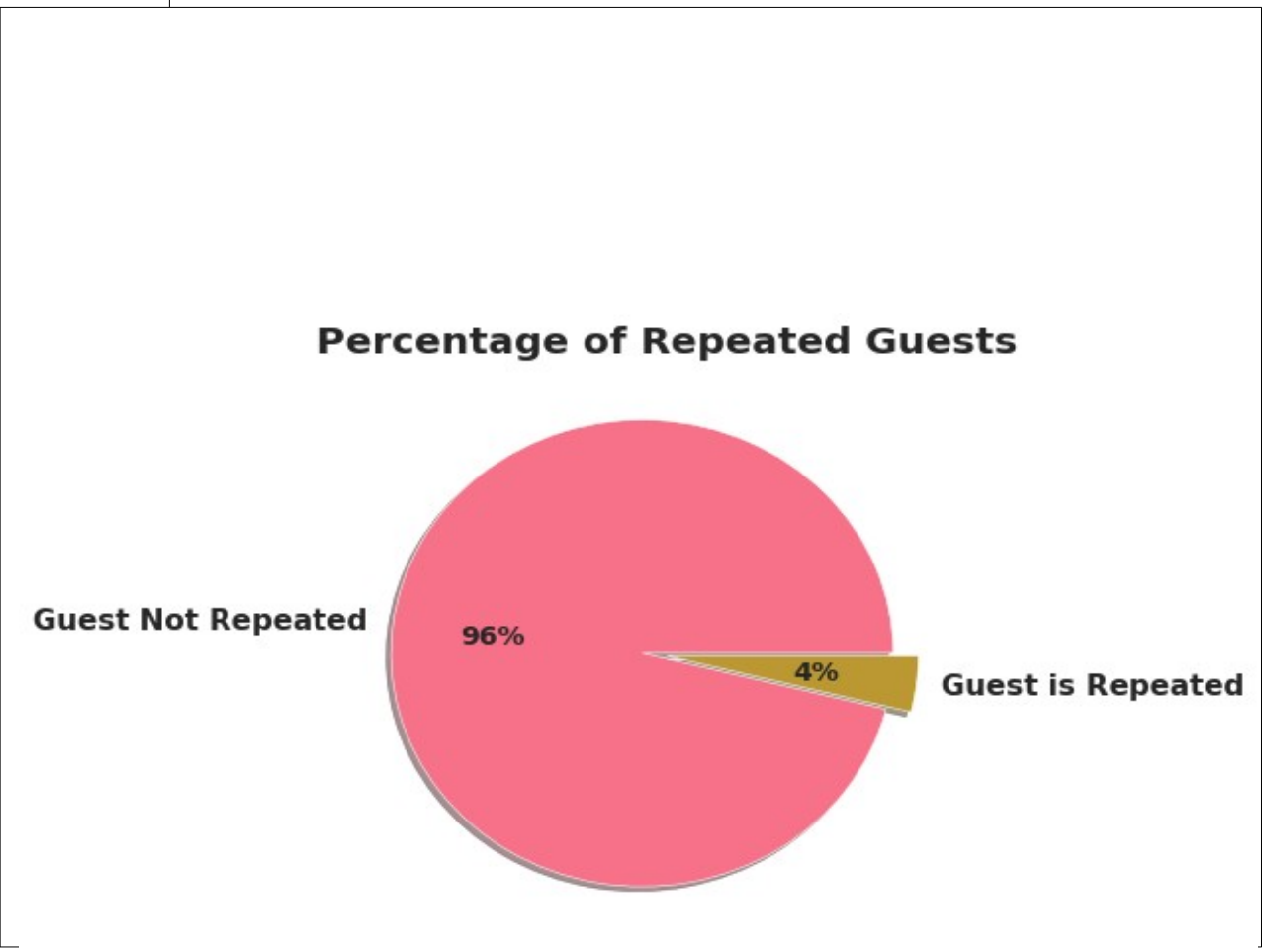
1. The City hotel has more guests during spring and autumn, when the prices are also highest, In July and August there are less visitors, although prices are lower.

2. Guest numbers for the Resort hotel go down slightly from June to September, which is also when the prices are highest. Both hotels have the fewest guests during the winter.



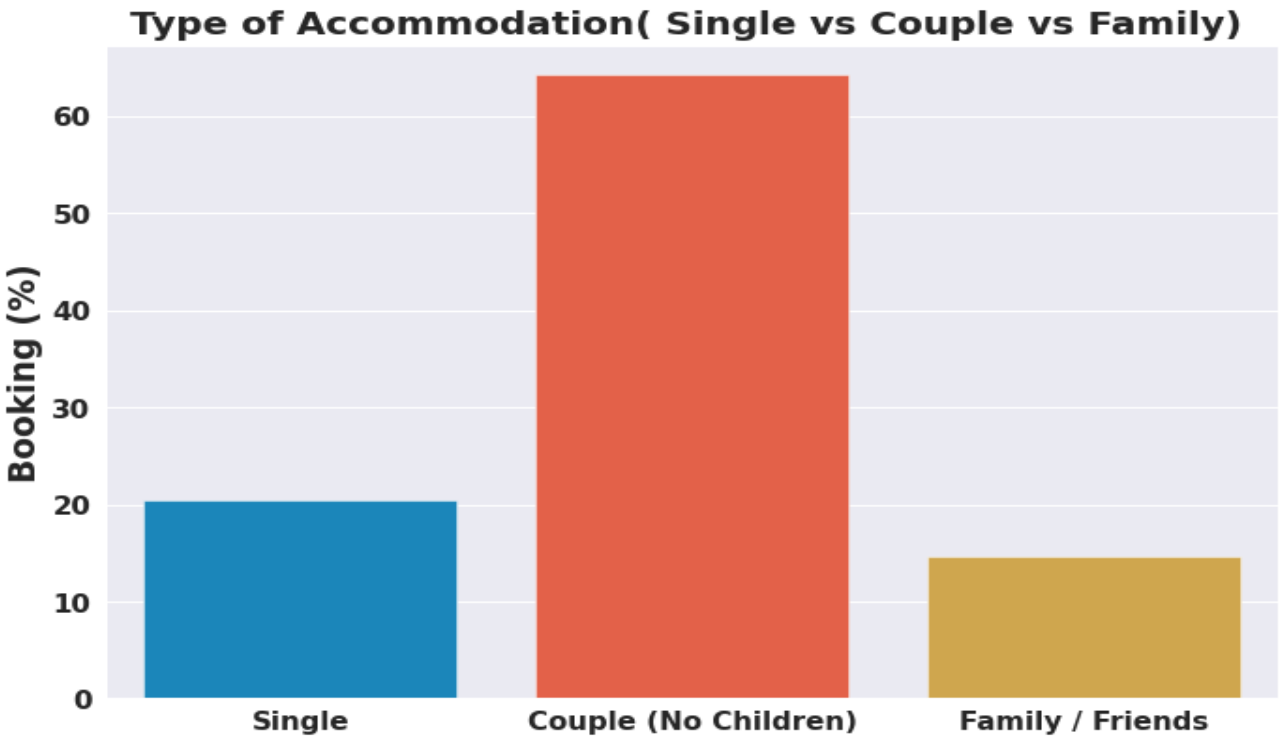


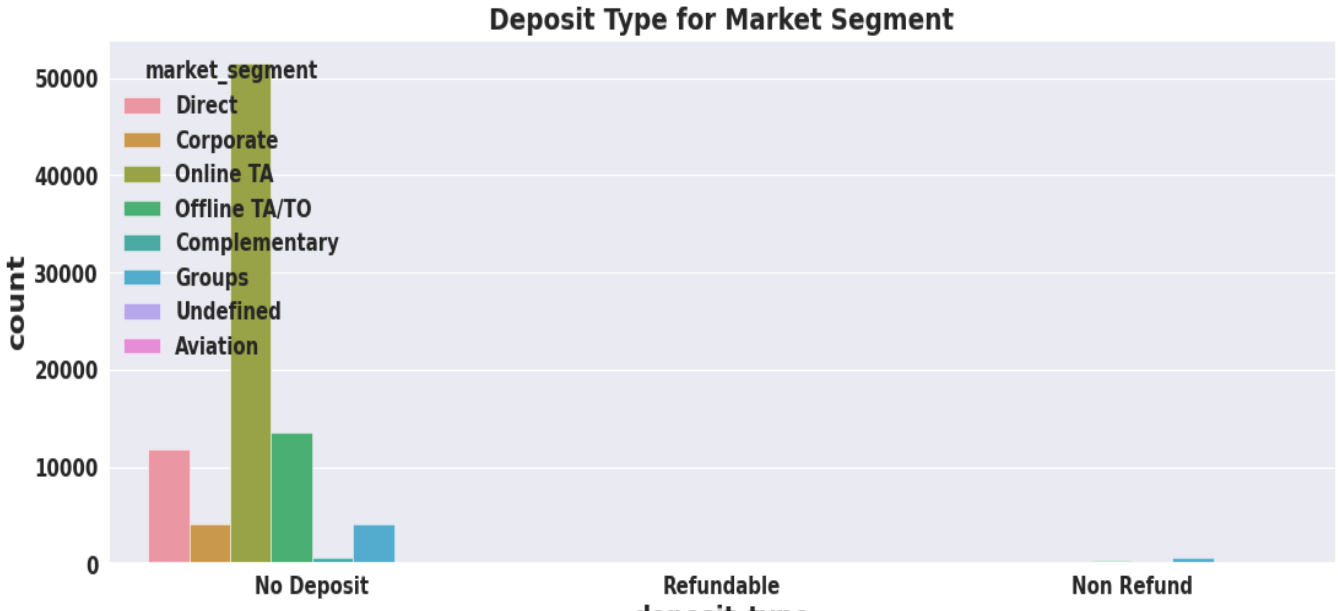
There have been many arrivals in the year 2016 than the remaining years. We can also say that there has been increase in the arrivals as years passes.

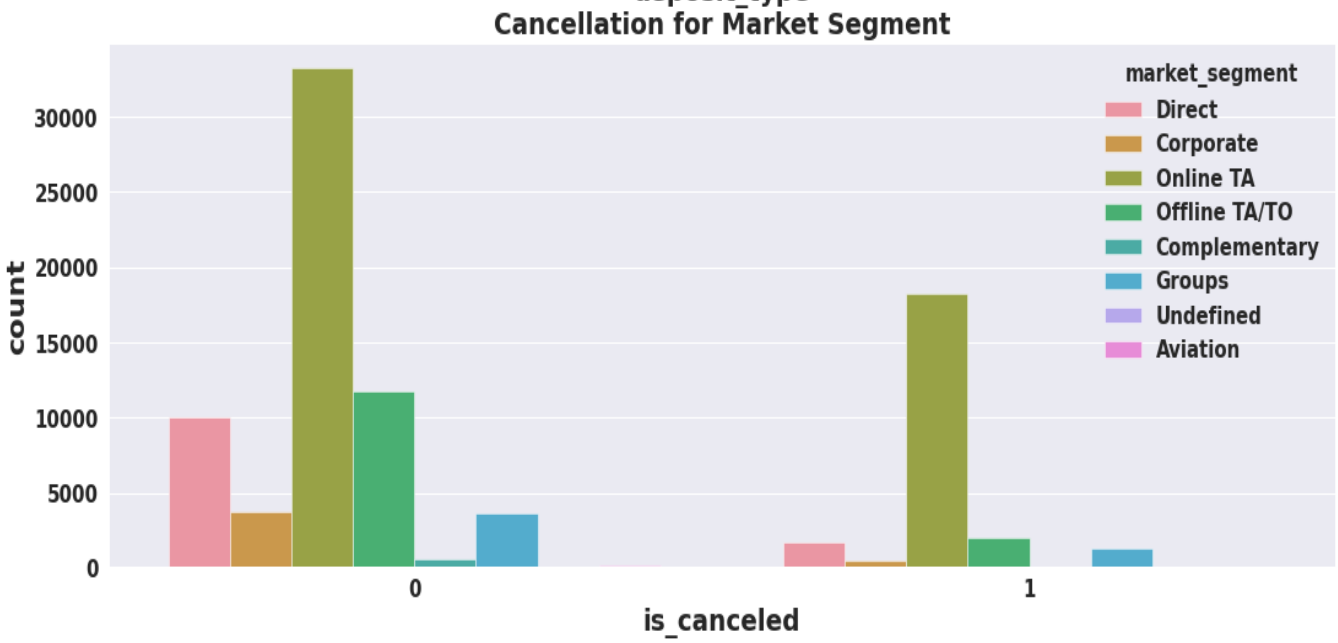


The chart shows only 4% Guest is Repeated in Hotel. It means 3359 guests out of 83981 is repeated.

In resort hotel people like to stay 1 day and in city hotel people like to stay 2-3 days

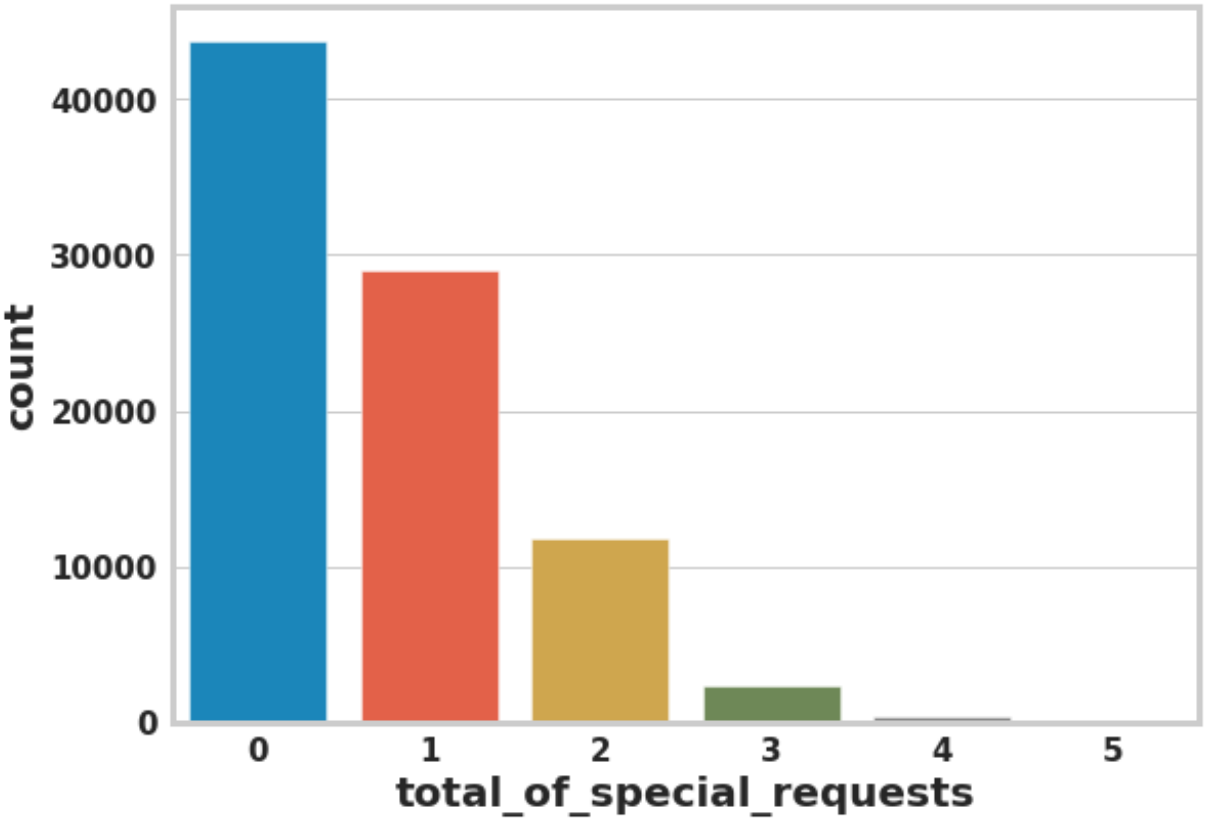


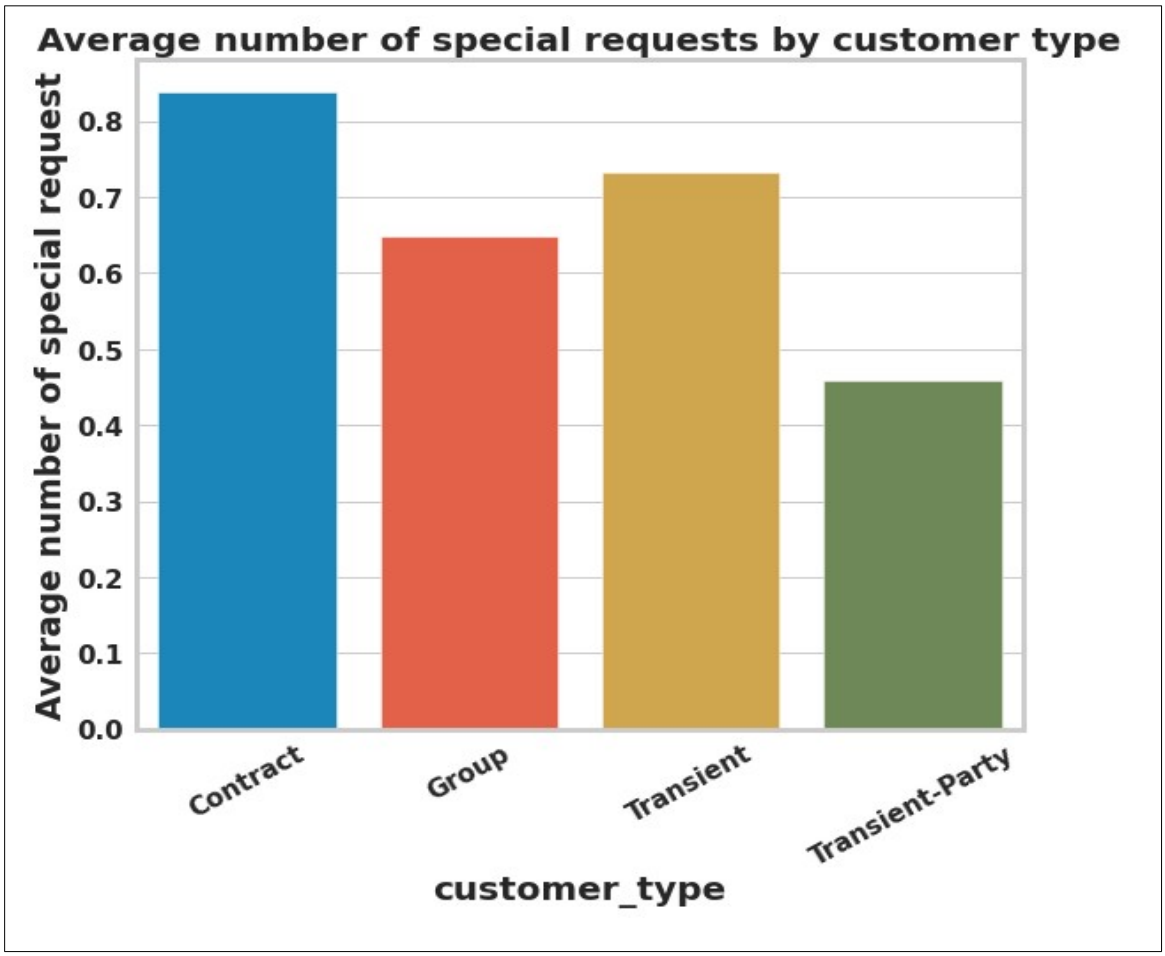




we can see from above graph 1 that most the bookings are done through Online TA segment and from above graph 2 that most cancelation is also done through online TA segment only.

Special Requests

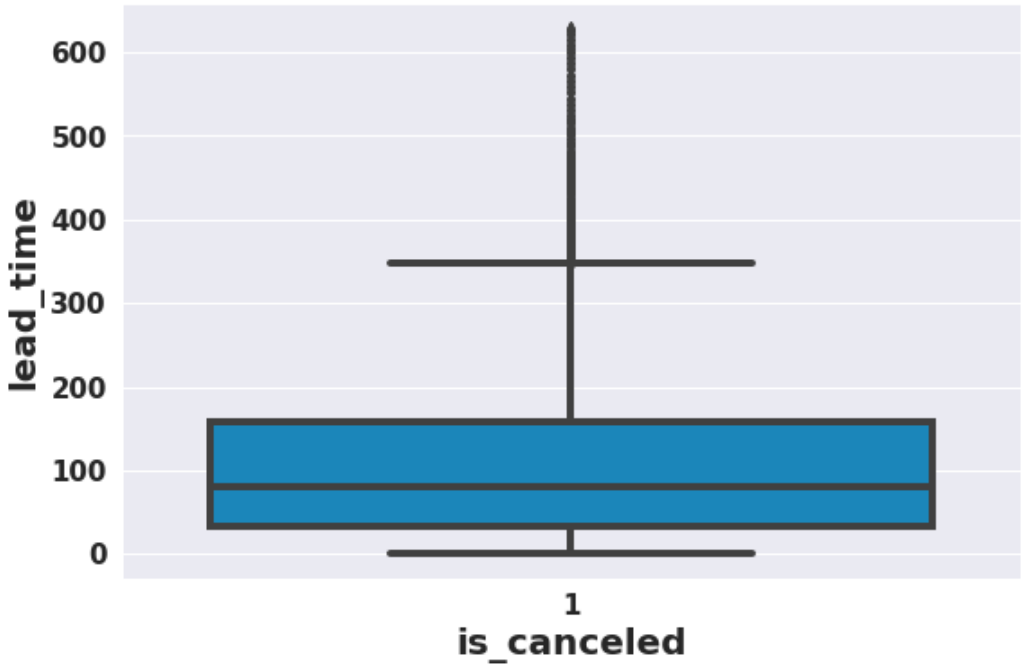




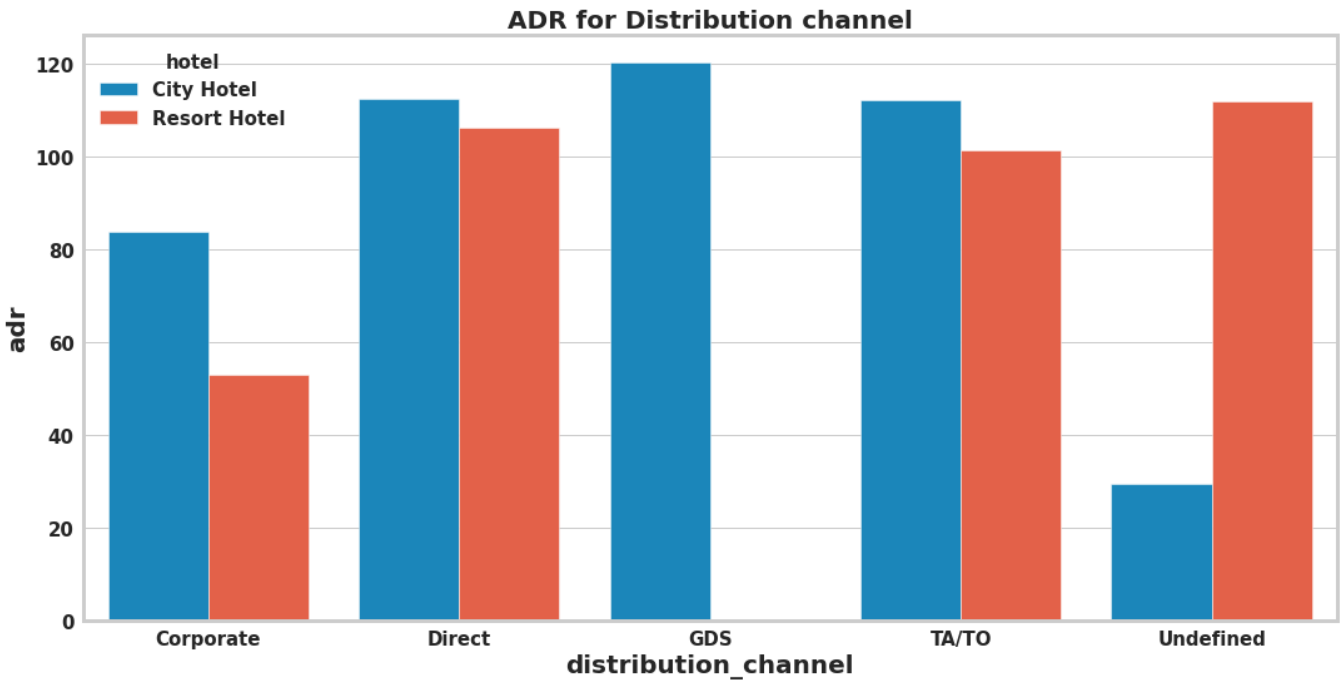
They are not entertaining much special requests. Contract people have more special requests

Distribution channel

The most preferred distribution channel is TA/TO.



When lead time increases the chances for cancelation increases.



Conclusions:

Based on the exploration of Data we can say that:

1. Almost 28% of bookings were cancelled and the remaining are interested to booking hotels.

2. More than 60% of the population booked the City Hotel compared to Resort Hotel.

3. Most bookings were made from July to August. i.e. During Summer Season the hotels are expected to get more no of bookings than any other seasons and the least bookings were made at the start and end of the year.

4. Majority of the guests are from Western Europe Countries to stay in the hotels.

5. We gonna clearly say the average stay in the hotels between 1-3 days.

6. Couples (or 2 adults) are the most popular accommodation type.

7. The most preferred distribution channel is TA/TO in both hotel type.

8. Most of the bookings are done through online TA segment

9. They are not entertaining much special requests. Contract people have more special requests and transient people have lower special requests.

10.When lead time increases the chances of cancellation increases.