

EDA Capstone Project-1

Hotel Booking Analysis

Individual Contributor

Nargis Nasreen

Content

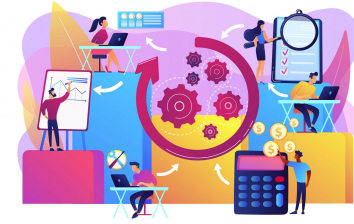
- **Problem Statement**
- **Workflow**
- **Data Summary**
- **Data Cleaning**
- **Data Manipulation**
- **Exploratory Data Analysis**
- **Observation**
- **Conclusion**

Problem Statement

- Travelling is the big part of our life now. We are connected to the whole world and we need to travel sometimes for work, sometimes to explore, sometimes to spend good time with family members, etc.
- So, have you ever wondered 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? What if you wanted to predict 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!

Let's explore and analyze the data to discover important factors that govern the bookings.

Workflow



**Data collection
and
understanding**

Data Cleaning

**Data
Manipulation**

**Exploratory
Data Analysis**

**Analyzing
the result**

Data Summary



Dataset Name: Hotel Bookings.csv

Shape:

Rows – 119390

Columns – 32

Features:

Hotel, Is_cancelled, Arrival_date_year, Arrival_date_month, Arrival_date_week_number, Day_of_the_month, Stays_in_weekend_nights, Stays_in_week_nights, Adults, Children, Babies, Meals, Country, Market_segment, Distribution Channel, Is_repeated_guest, Previous_cancellations, Previous_bookings_not_canceled, Reserved_room_type, Assigned_room_type, Booking_changes, Deposit_type, Agent, Lead_time, Days_in_waiting_list, Customer_type, ADR, Required_car_parking_spaces, Booking changes, Reservation_status, reservation_status_date

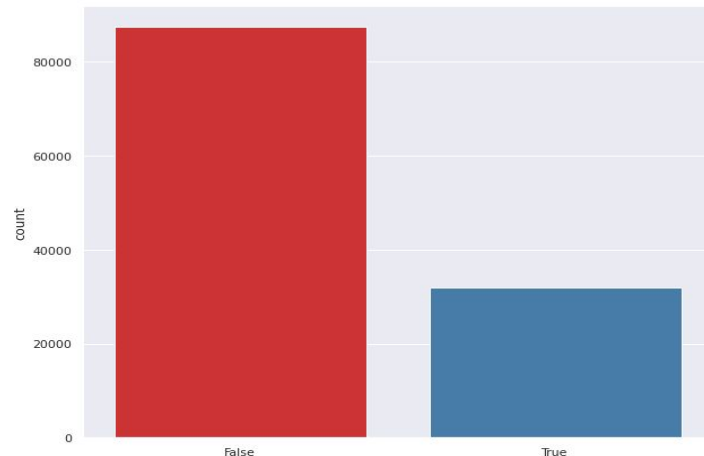
Data Cleaning

The dataset contains **31994** duplicate rows.

It is very important to remove the duplicate as it will take unnecessary space and affect the visualization as well.

To handle the duplicate rows:

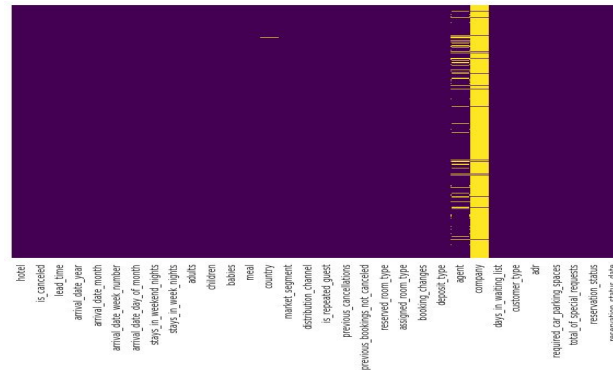
- I used `drop_duplicates()` method to remove all the duplicates.



Data Cleaning

The dataset contains NaN values in few columns like:

- company
- agent
- country
- Children



To handle the null values:

- I replaced company and agent missing values with 0.
- country missing values with 'Others'.
- Children Column has only 4 missing values, so replace with 0 considering no children.

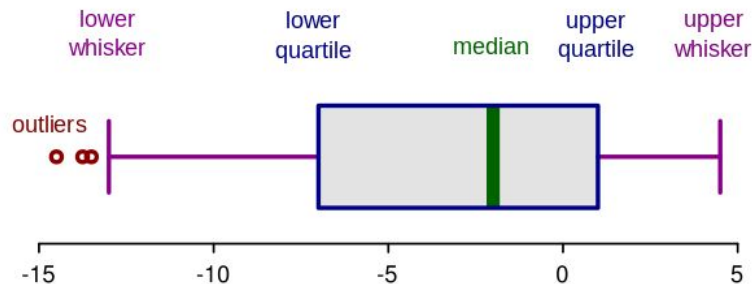
Data Cleaning

Outliers: An Outlier is a datapoint in a dataset that is distant from all the observations.

To handle outliers I used IQR i.e, I consider all the datapoint ,that lies below 1.5 times of IQR from lower quartile(q_1) and above 1.5 times of IQR from upper quartile(q_3) and replaced them with lower bound and upper bound.

Lower bound: $(Q_1 - 1.5 * IQR)$

Upper bound: $(Q_1 + 1.5 * IQR)$



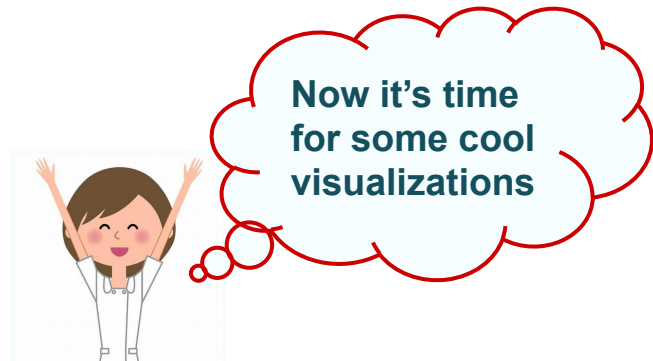
Data Manipulation

I combined few columns together as it will give better results and visualizations.

Combining columns for an effective study:

→ `total_people = adults+babies+children`

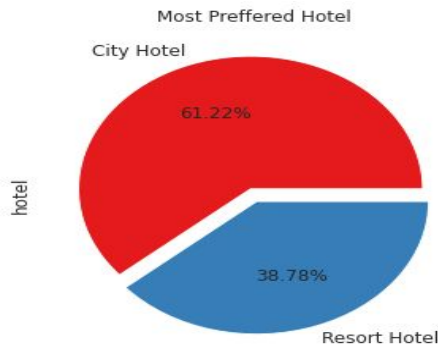
→ `total_stays = stays_in_weekend_nights+stays_in_week_nights`



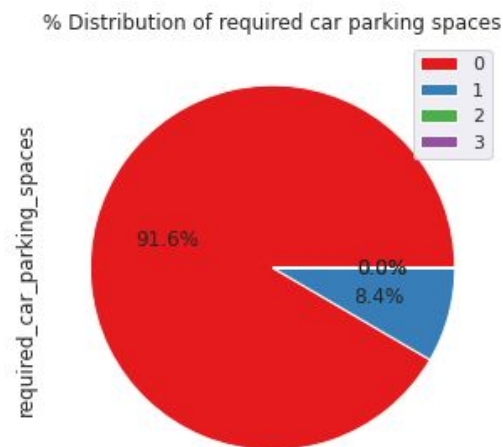
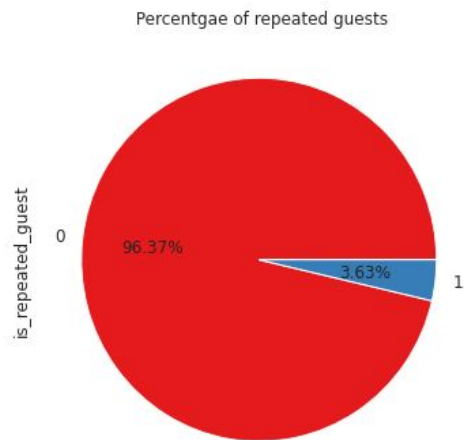
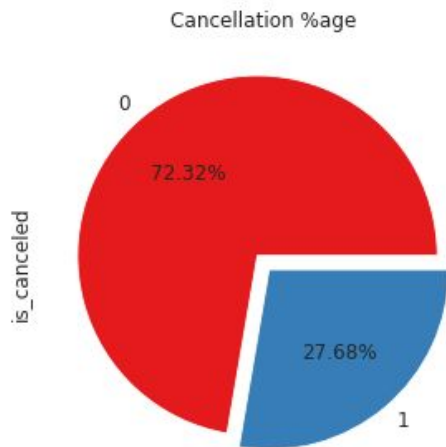
Exploratory Data Analysis

- 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.
- So we I have analyzed both city hotel and resort hotel together as well as separately.

People prefer city hotel more than resort hotel.

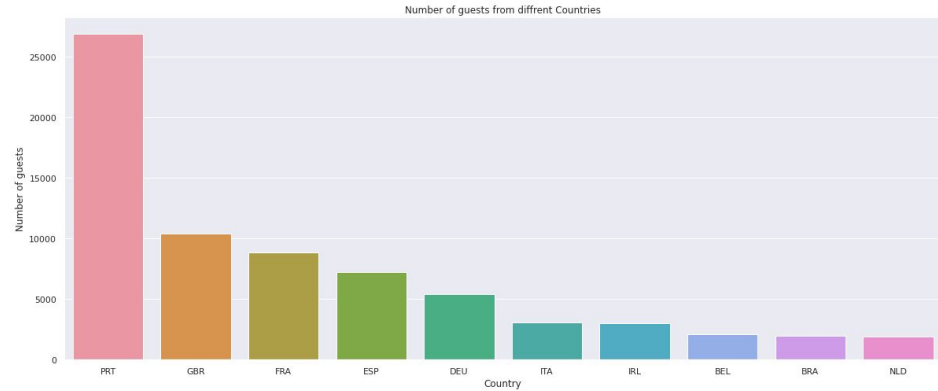
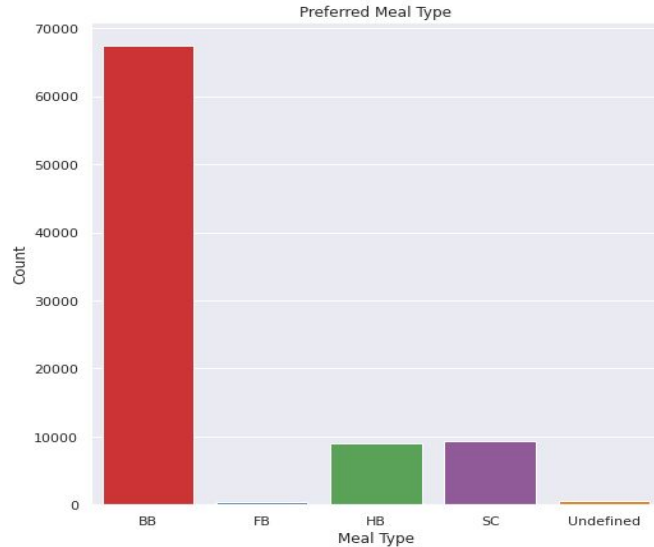


Exploratory Data Analysis



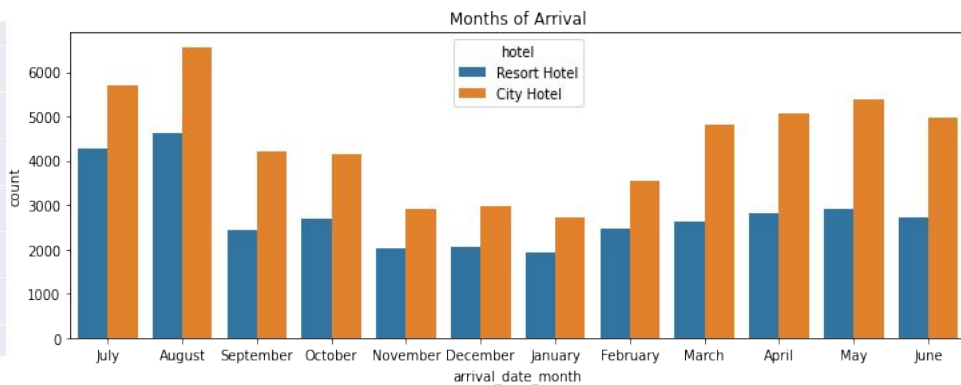
- Out of total bookings 27.68% got cancelled.
- Only 3.63% of guests are repeated guests.
- 91.6% of guests do not need parking space, only 8.4% of guests need 1 parking space.

Exploratory Data Analysis



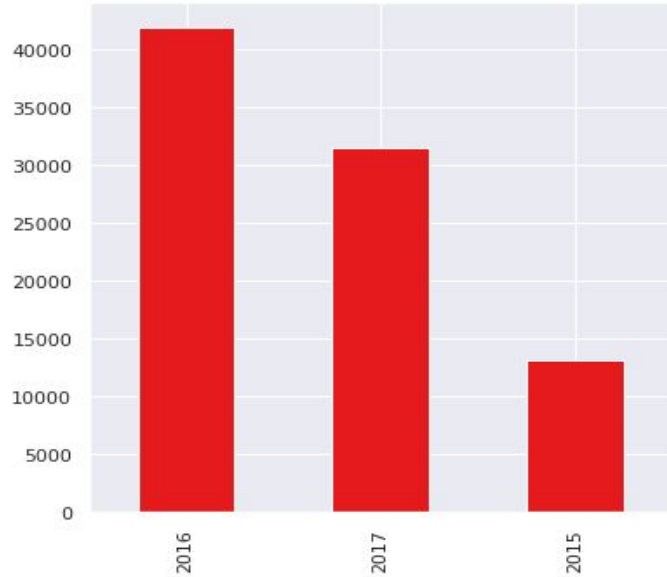
- So the most preferred meal type by the guests is BB(Bed and Breakfast)
- Most of the guests are from Portugal followed by United Kingdom and France.

Exploratory Data Analysis



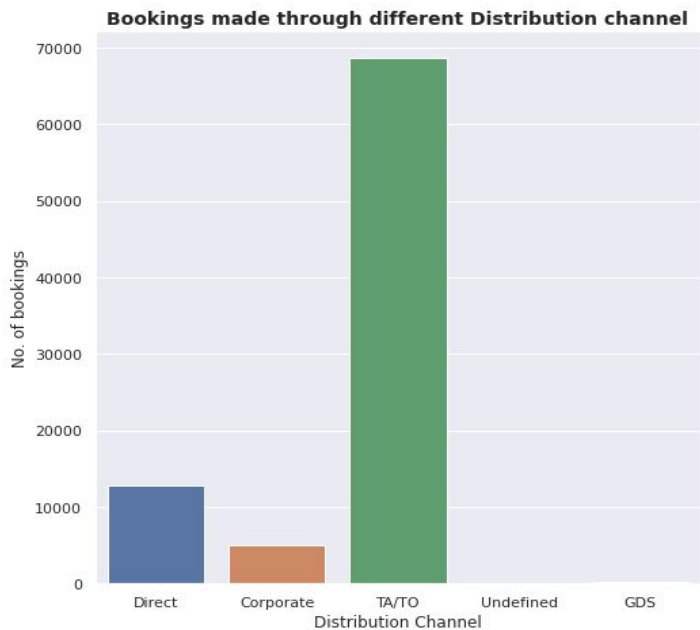
- July and August have the maximum number of Bookings.
- And in each month booking in city hotel is higher than resort hotel.

Exploratory Data Analysis

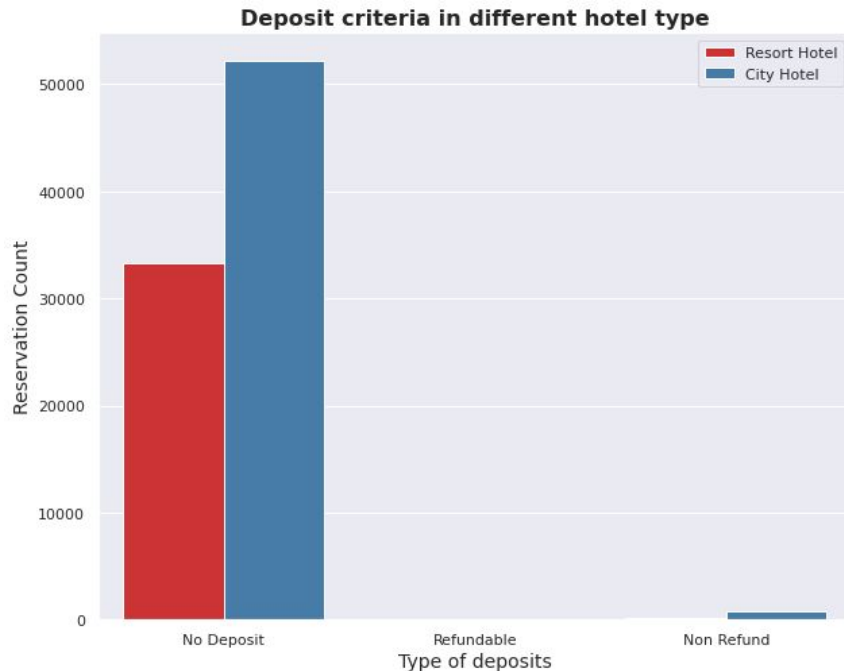


- 2016 has maximum number of booking.
- Again in each year bookings made in city hotel is higher than resort hotel but in 2015 bookings made in resort hotel is slightly higher than city hotel.

Exploratory Data Analysis

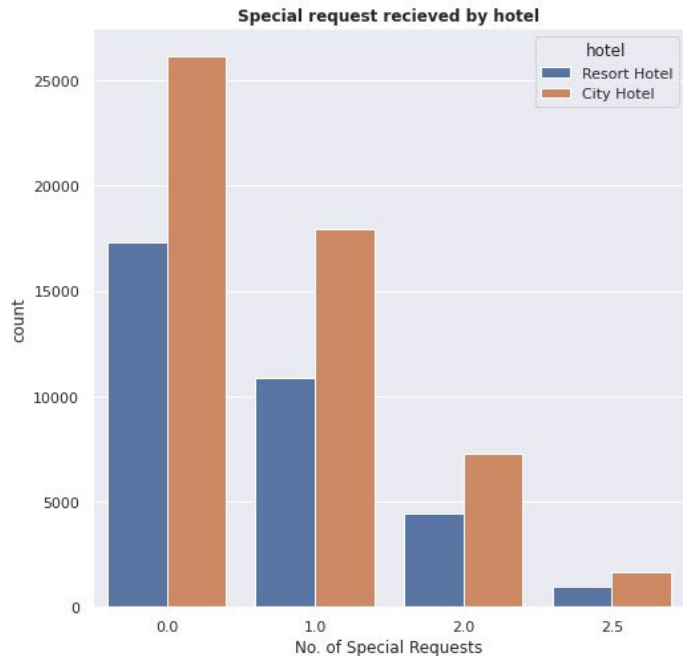


- TA/TO: Travel agents/Tour operators
- TA/TO made maximum number of bookings

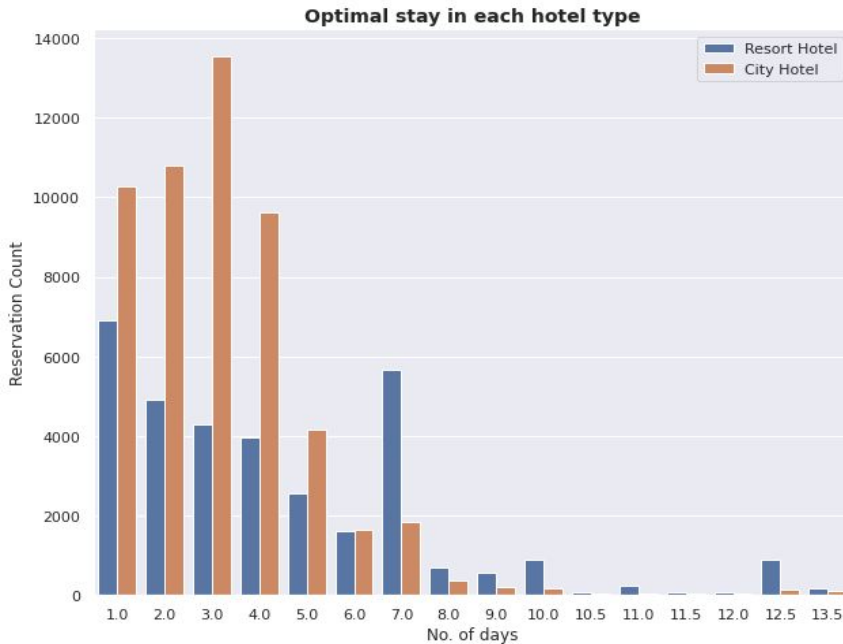


No deposit criteria is followed by both type of hotels.

Exploratory Data Analysis



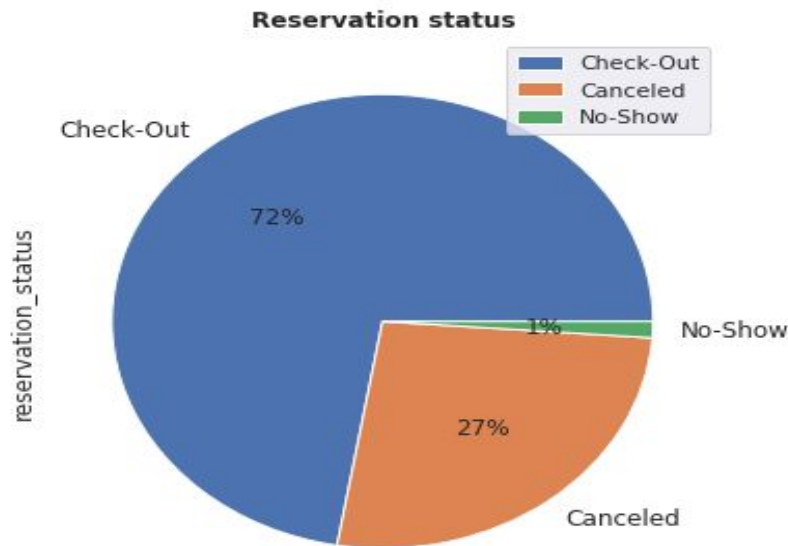
City hotels receives more special requests than Resort hotels.



For days \leq 5 people prefer to stay in city hotel.

For days $>$ 5 people prefer to stay in resort hotel.

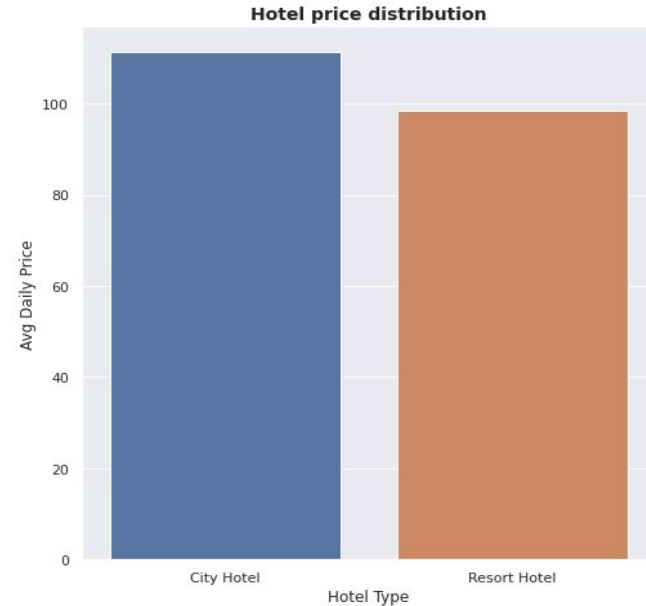
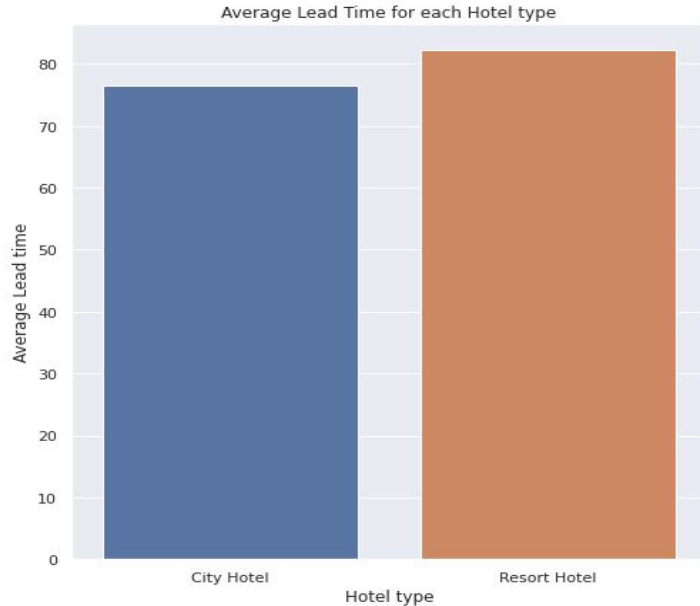
Exploratory Data Analysis



This analysis is done on both city hotel and resort hotel together:

- Out of total no of reservations 72% actually show up
- 27% got canceled
- 1% reservation got No show.

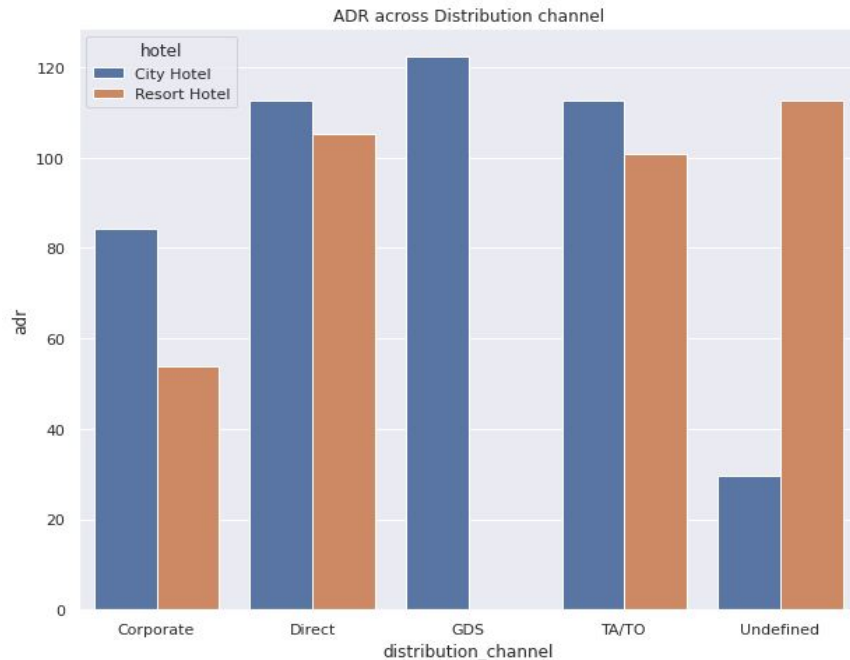
Exploratory Data Analysis



- Average lead time of resort hotel is higher than city hotel, i.e, people prefer to stay in resort hotel for maximum number of days.
- ADR of city hotel is higher than resort hotel.

Exploratory Data Analysis

- Direct and TA/TO has almost equal ADR in both types of hotels which is high among other channels.
- GDS has high ADR in city hotel type but almost negligible for resort hotel. GDS needs to increase resort hotel bookings.
- From this, it's clearly visible that Direct and TA/TO are generating more revenues than other distribution channel.



Observation

1. City hotels are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.
2. 27.5 % bookings were got cancelled out of all the bookings.
3. Only 3.9 % people were revisited the hotels. Rest 96.1 % were new guests. So hotel needs should use different offers and ways to increase the %age of revisited guests.
4. Most of the customers (91.6%) do not require car parking spaces.
5. 79.1 % bookings were made through TA/TO (travel agents/Tour operators).
6. BB(Bed & Breakfast) is the most preferred type of meal by the guests.
7. Maximum number of guests were from Portugal, i.e. more than 25000 guests.
8. Most of the bookings for City hotels and Resort hotel were happened in 2016.
9. Average ADR for city hotel is high as compared to resort hotels. These City hotels are generating more revenue than the resort hotels.
10. Average lead time for resort hotel is high.
11. If guests have to stay more than 5 days they prefer resort hotel.

Conclusion

- Chances of cancellation is high when there are no deposits taken by hotels, so hotels should take minimum deposits to minimise the rate of cancellation.
- Maximum number of bookings are in the month May to August, so hotels should provide exciting deal to customers to increase their booking in off season.
- As hotels are getting less repeated customers so management should take customer's feedback and improve the hotel facilities to increase the count of their repeated guests.

Thank you