Capstone Project - 1

Hotel Booking Analysis

PRESENTED BY
SHUBHAM PANASKAR

Content

Problem Statement

- For this project we will be analyzing Hotel Booking data. 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.

- Hotel industry is a very volatile industry and the bookings depends on above factors and many more.

Problem Statement(contd)

- The main objective behind this project is to explore and analyze 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.

- It’s very important to understand your data. So we had hotel Booking analysis data. Which had 119390 rows and 32 columns. So let’s understand this 32 columns.

Data Summary

The dataset spans over three years - 2015, 2016 and 2017.

1. **‘hotel’ :** Denotes the type of hotel - Resort hotel or city hotel.
2. **‘Is_canceled’:** Denoted by ‘1’ if booking was cancelled or ‘0’ if booking was not cancelled.
3. **‘Lead_time’ :** Period between time of booking and checking in (considered in days here).
4. **‘arrival_date_month’ :** The month when customer arrived at the hotel.
5. **‘country’:** The country of origin of the customer; has 158 countries listed.
6. **‘Days_in_waiting_list’ :** Number of days the booking was in the waiting list before it was confirmed to the customer.
7. **‘Deposit_type’ :** Indication on if the customer made a deposit to guarantee the booking. Three categories, No-deposit, Non-Refund, Refundable.

Data Summary(contd)

8.‘Adr’ : Average Daily rate as defined by the average rental revenue earned for an occupied room per day.

9. ‘Adults, Babies, Children’ : Number of adults, babies and children.

10. ‘Assigned Room Type’ : Code for the type of room assigned to the booking.

11. ‘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.

12. ‘Distribution_channel’ : Booking distribution channel.

13.‘Is_repeated_guest’ : Value indicating if the booking name was from a repeated guest(1) or not (0).

Data Summary(contd)

14.'Company' : ID of the company/entity that made the booking or responsible for paying the booking.

15. '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 bookings is not part of a group or contract, and is not associated to other transient booking; Transient party – when the booking is transient to at least other transient booking.

16. 'Market_segment' : Market segment designation.

17.'Previous_cancellations' : Number of previous bookings that were cancelled by the customer prior to the current booking.

18.'Required_car_parking_spaces' : Number of car parking spaces required by the customer.

Data Summary(contd)

19.'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.

20. 'Reservation_status_date' : Date at which the last status was set. This variable can be used in conjunction with the *ReservationStatus* to understand when was the booking canceled or when did the customer checked-out of the hotel.

21. 'Reserved_room_type' : Code of room type reserved.

22.'Types_of_special_requests' : Number of special requests made by the customer(e.g. Twin bed or high floor)

23.'Stays_in_weekend_nights, Stays_in_week_nights' : Number of weekend nights and week nights the guest stayed or booked to stay at the hotel.

Data Cleaning and Manipulation

- **There were 4 columns company, agent, country and children with missing values.**

  - Columns like Agent and Company have maximum number of null values as compared to other columns and we will replace all null values with 0, because these are not missing values instead they will be considered as "Not Applicable“

  - We will replace all missing values of column 'children' with rounded mean value as it contains the count of children.

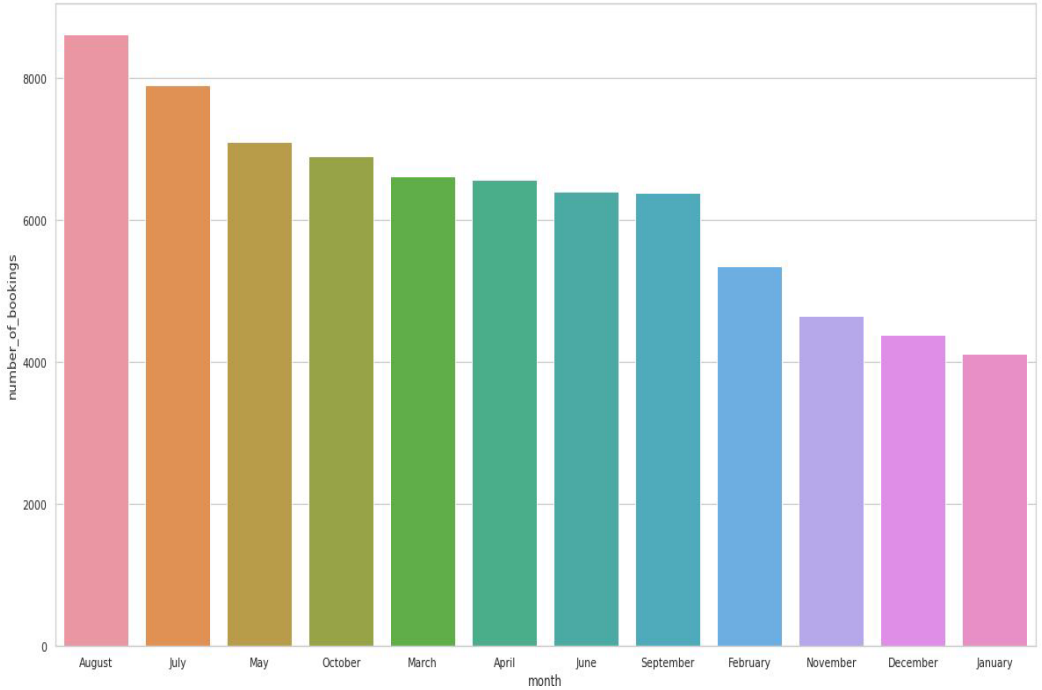
  - We will replace NULL values of 'country' column with mode of that column.

- Handling Duplicates: Data had 31994 duplicates values. So we dropped it from the data.

Analysis of Data

Which month has least traffic?

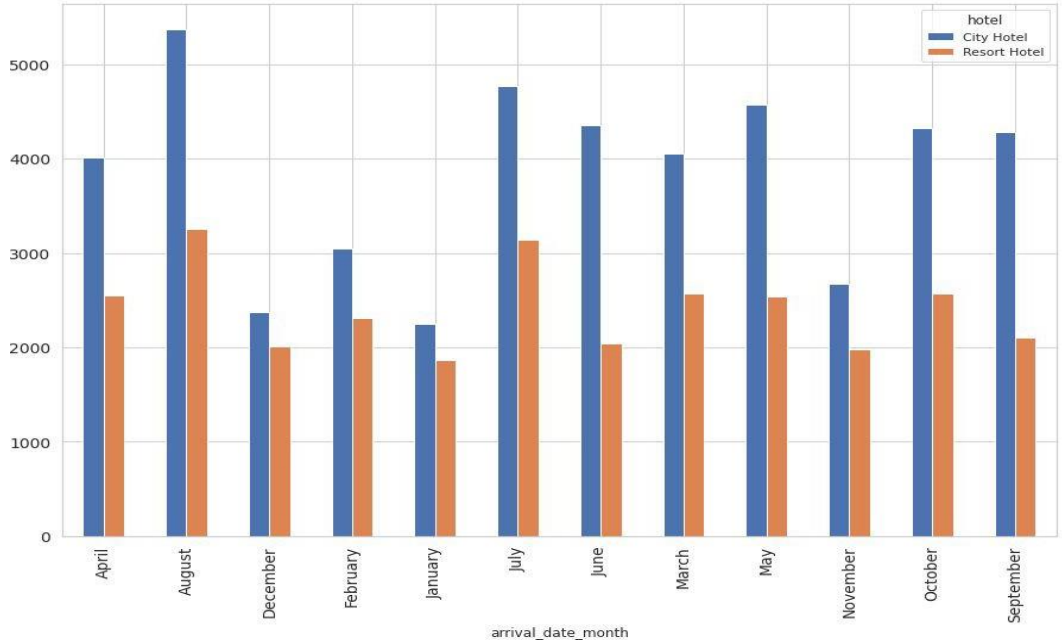
January month has least traffic, which means Customers should do booking in month of January.



Analysis of Data(contd)

Which month has maximum traffic?

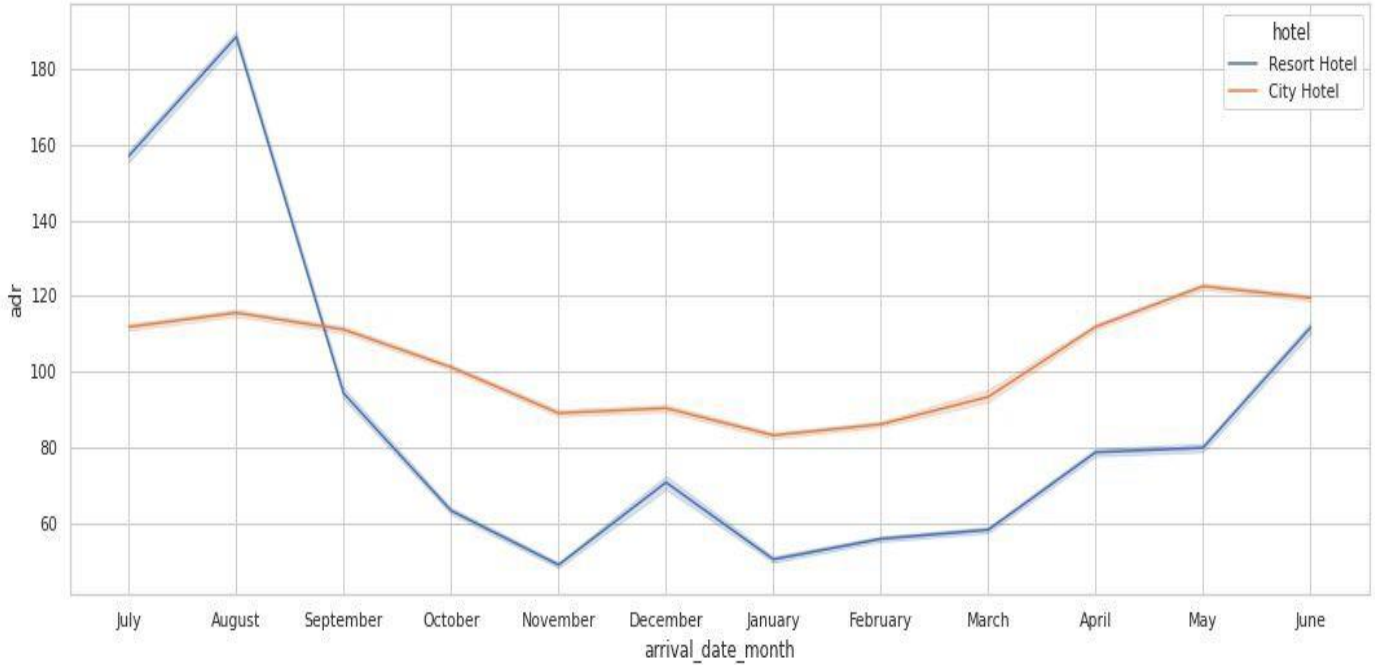
August month has the most traffic which means customer should avoid do bookings in month of August.



Analysis of Data(contd)

Which months have cheaper booking rates?

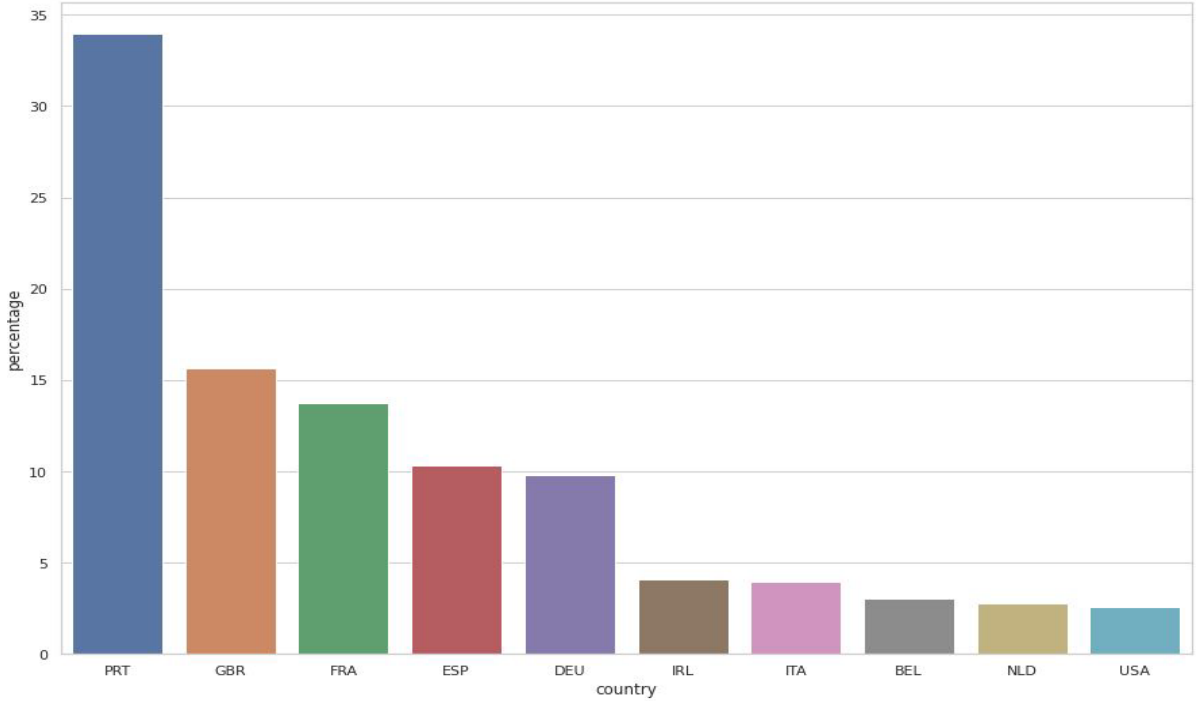
November to January have cheaper average daily rates.



Analysis of Data(contd)

From which country most guests are coming?

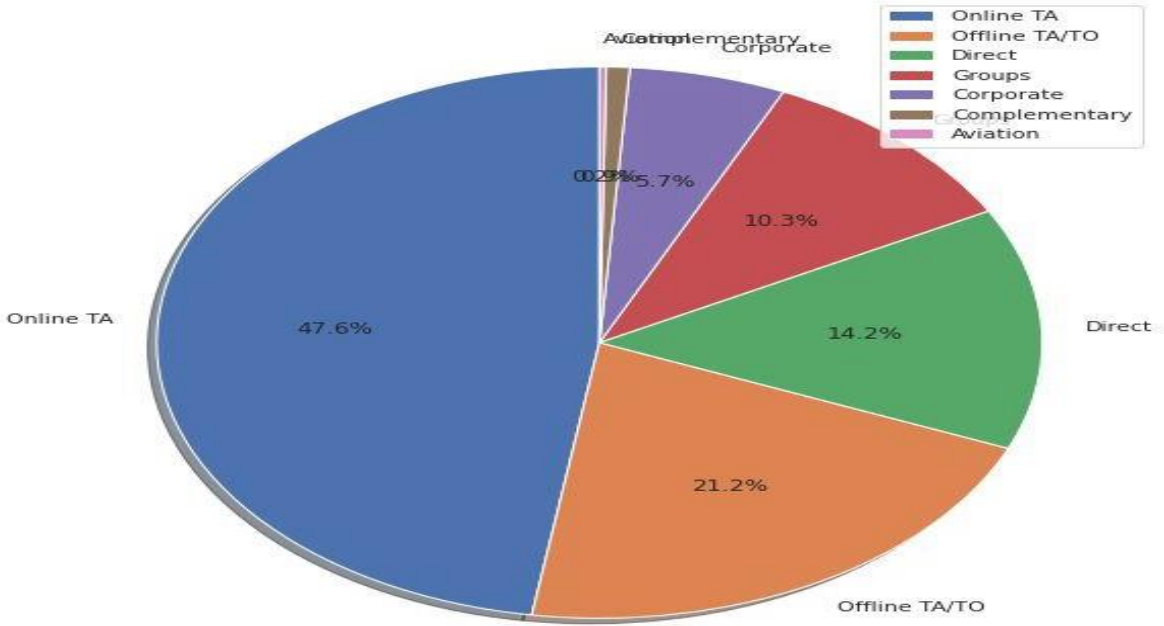
More visitors are from Western Europe namely, portugal , France, Spain.



Analysis of Data(contd)

Who does the booking?

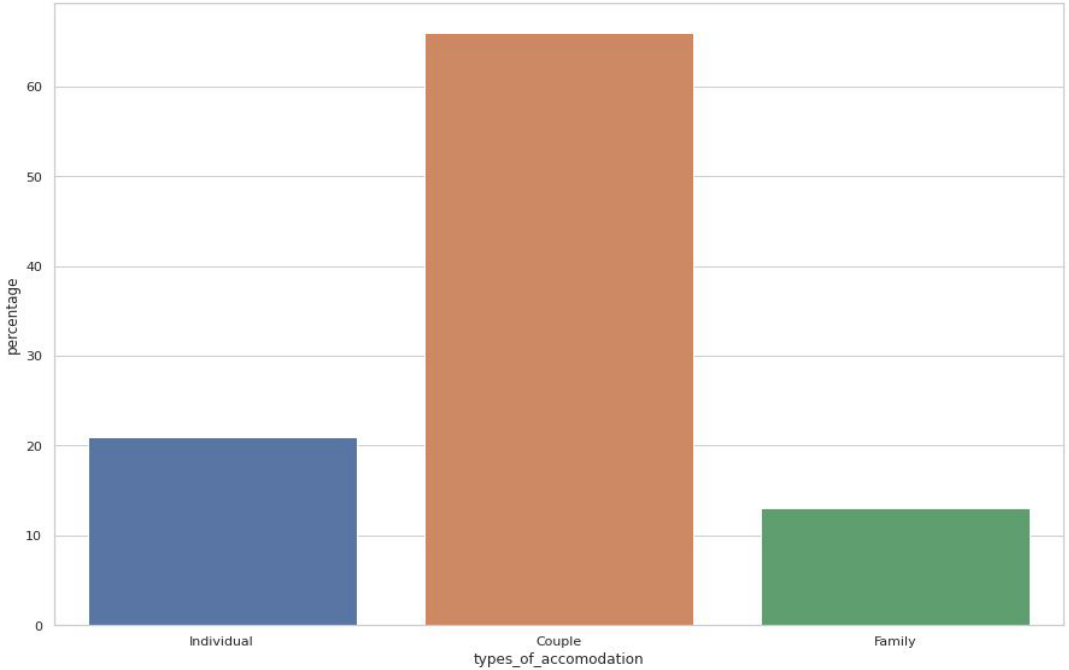
Maximum number of bookings made by Online Travel Agents.



Analysis of Data(contd)

Who does the booking?

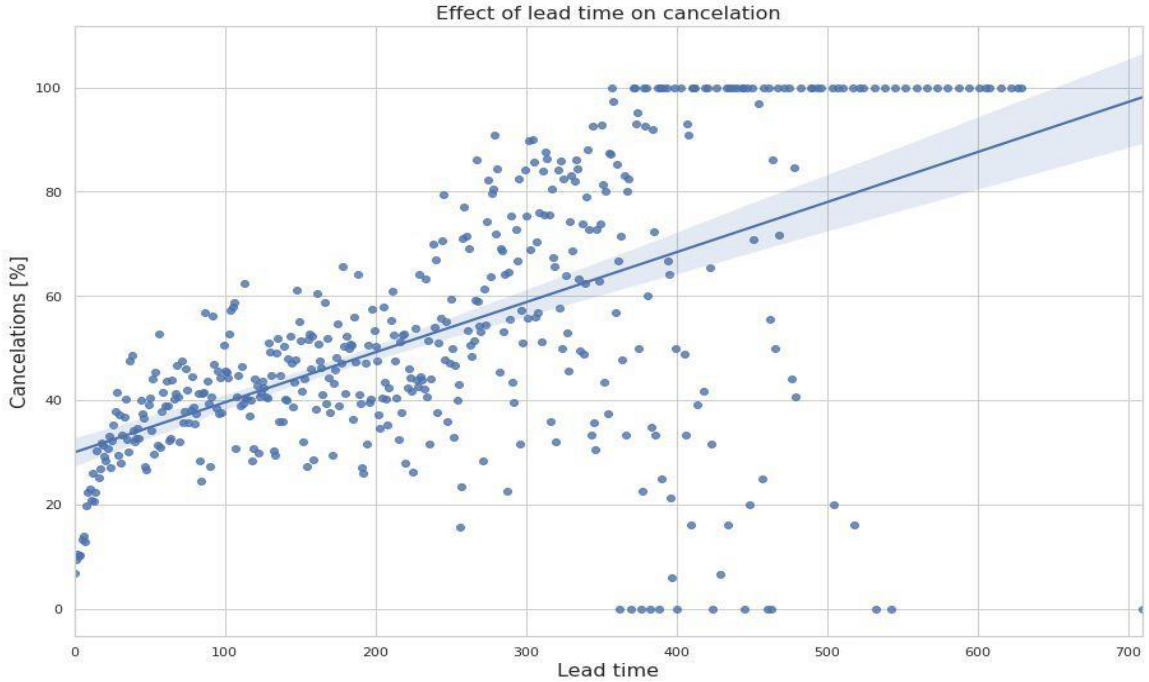
Maximum number of Bookings made by ‘Couples’.



Analysis of Data(contd)

What causes cancellations?

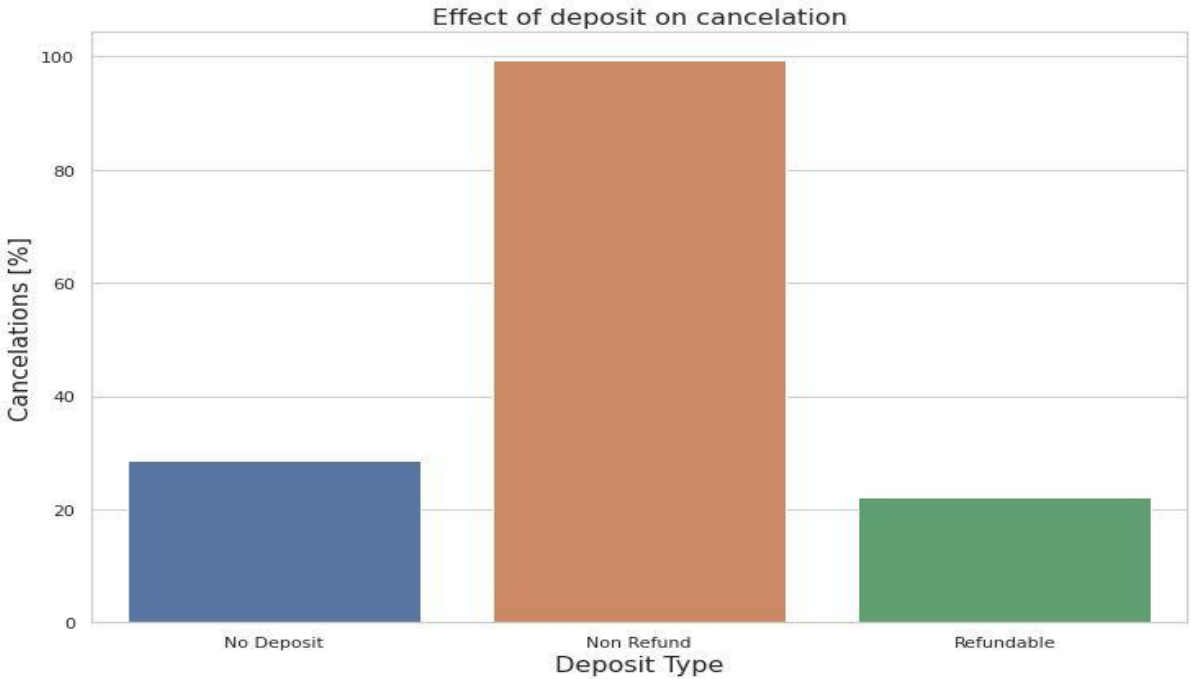
As lead time increases, percentage of booking cancellation also increases



Analysis of Data(contd)

What causes cancellations?

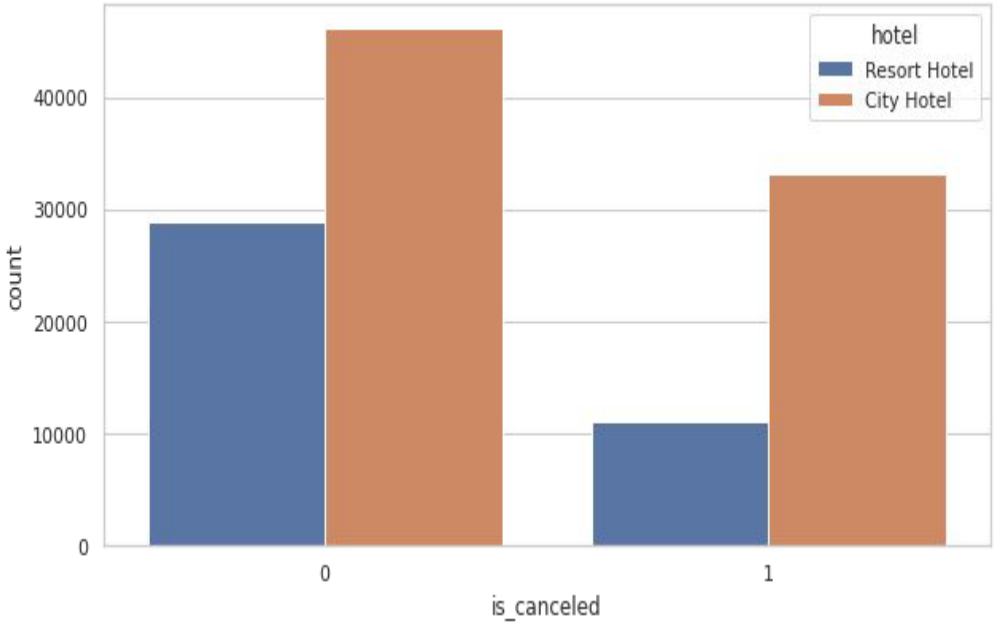
‘Non-Refund’ policy increases the number of booking cancellations



Analysis of Data(contd)

Which type of hotel is mostly preferred by customers?

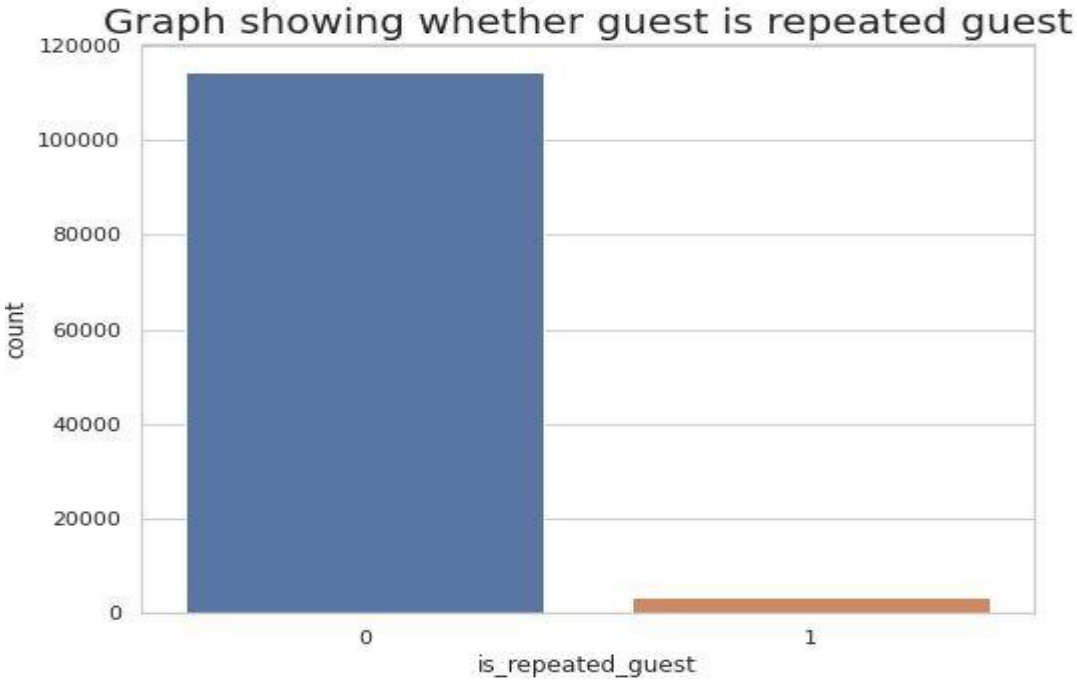
Number of cancelled booking is high in case of city hotel. So city hotel have to focus on advertisement campaign.



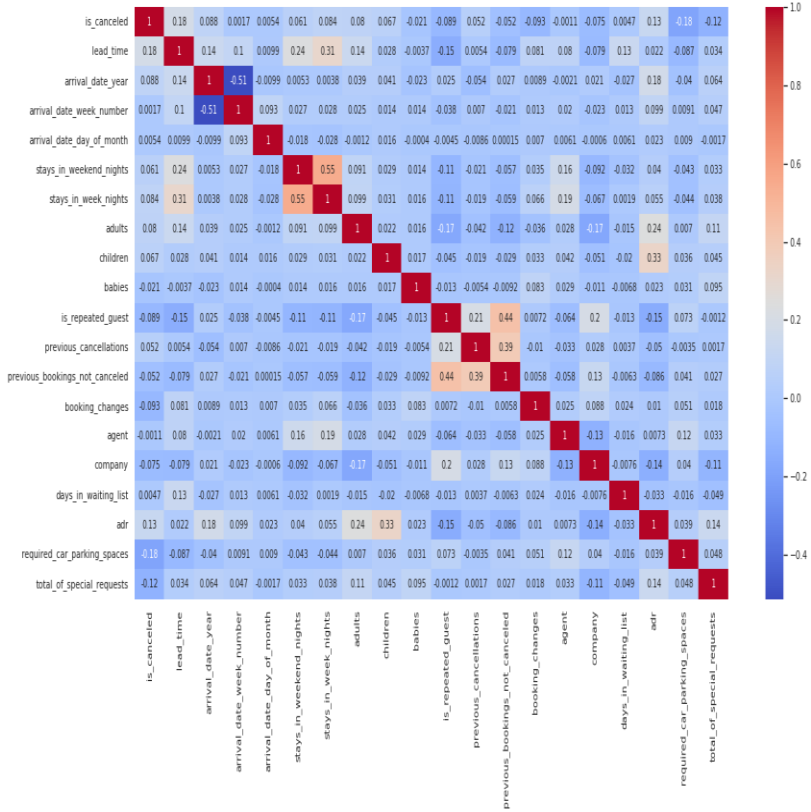
Analysis of Data(contd)

What is the number of repeated bookings?

Number of repeated bookings are very low. So hotel management can come up with new idea that will increase customer retention.



Analysis of Data(contd)



- is canceled and same_room_alloted_or_not are negatively correlated. Not getting the same room as per reserved room is not the reason for booking cancellations.
- lead-time and total stay is positively correlated means more is the stay of customer more will be the lead time.
- ADR and total people are highly correlated. That means more the people more will be adr.High adr means high revenue
- is_repeated_guest and previous_bookings Not_canceled has strong correlation. May be repeated guests are not more likely to cancel their bookings.

Challenges

1. Time taken to articulate the problem statements.
2. Consideration of data subsets.
3. Many rows have zero guests including adults, babies and children.

Conclusions

1. Target months between May to Aug because those are peak months due to the summer period.
2. Book hotels in month from November to January as they have cheaper average daily rates.
3. Mostly guests are coming from European countries, so target those countries for advertisements.
4. Mostly bookings are done by Online TA market segment and also couples do the bookings mostly.
5. Increase in lead time and ADR impacts more cancellations and also Non-Refund policy made customers to cancel their bookings.
6. Mostly cancellations are done in case of city hotel, so customers prefer resort hotel.
7. Since there are very few repeated guests, focus should be on retaining customers after their first visit by fulfil their more special requests.

Q and A