Hotel Booking Pattern & Analysis

Prasham Bhuta

30/03/2020

Analysis of data received from Room Booking Platform

Online platforms such as trivago, goibibo, makemytrip etc. are used for booking hotel rooms, here is a dataset of rooms booked from such platform.

Tasks

- To understand the pattern of bookings, and the general trends followed by users.
- Create a report, as being a part of the platform, for the marketing team.

Let's get started

Importing necessary libraries

```
library(tidyverse)
library(ggplot2)
```

Import data from the csv

```
datas <- read.csv("dataset/hotel_bookings.csv")
```

Get the str of the data

```
str(datas)
```

```
## 'data.frame':    119390 obs. of  32 variables:
##  $ hotel                         : Factor w/ 2 levels "City Hotel","Resort Hotel": 2 2 2 2 2 2 2 2 2 2
##  $ is_canceled                   : int  0 0 0 0 0 0 0 0 1 1 ...
##  $ lead_time                     : int  342 737 7 13 14 14 0 9 85 75 ...
##  $ arrival_date_year             : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ arrival_date_month            : Factor w/ 12 levels "April","August",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ arrival_date_week_number      : int  27 27 27 27 27 27 27 27 27 27 ...
##  $ arrival_date_day_of_month     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ stays_in_weekend_nights       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ stays_in_week_nights          : int  0 0 1 1 2 2 2 2 3 3 ...
##  $ adults                        : int  2 2 1 1 2 2 2 2 2 2 ...
##  $ children                      : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ babies                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ meal                          : Factor w/ 5 levels "BB","FB","HB",..: 1 1 1 1 1 1 1 2 1 3 ...
##  $ country                       : Factor w/ 178 levels "ABW","AGO","AIA",..: 137 137 60 60 60 60 137 1
##  $ market_segment                : Factor w/ 8 levels "Aviation","Complementary",..: 4 4 4 3 7 7 4 4 7
##  $ distribution_channel          : Factor w/ 5 levels "Corporate","Direct",..: 2 2 2 1 4 4 2 2 4 4 ...
##  $ is_repeated_guest             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_cancellations        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ reserved_room_type            : Factor w/ 10 levels "A","B","C","D",..: 3 3 1 1 1 1 3 3 1 4 ...
##  $ assigned_room_type            : Factor w/ 12 levels "A","B","C","D",..: 3 3 3 1 1 1 3 3 1 4 ...
##  $ booking_changes               : int  3 4 0 0 0 0 0 0 0 0 ...
##  $ deposit_type                  : Factor w/ 3 levels "No Deposit","Non Refund",..: 1 1 1 1 1 1 1 1 1 1
##  $ agent                         : Factor w/ 334 levels "1","10","103",..: 334 334 334 157 103 103 334
##  $ company                       : Factor w/ 353 levels "10","100","101",..: 353 353 353 353 353 353 35
##  $ days_in_waiting_list          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ customer_type                 : Factor w/ 4 levels "Contract","Group",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ adr                           : num  0 0 75 75 98 ...
##  $ required_car_parking_spaces   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_of_special_requests     : int  0 0 0 0 1 1 0 1 1 0 ...
##  $ reservation_status            : Factor w/ 3 levels "Canceled","Check-Out",..: 2 2 2 2 2 2 2 2 1 1 ..
##  $ reservation_status_date       : Factor w/ 926 levels "2014-10-17","2014-11-18",..: 122 122 123 123 1
```

Understanding the columns

The data looks clean enough, with proper column headers, as well

- Hotel has two types:
  - Resort hotel
  - City hotel
- Is_cancelled:
  - "1" if the booking is cancelled
- lead_time
  - No of days between booking and booked date
- Arrival : year, month, week_number, day
- Stay
  - No of weekend nights
  - No of week nights (because of price difference during the weekends)
- No of people: adults, children, babies
- Meal booked has 5 types
  - BB - Bed & Breakfast
  - FB - Full Board (Breakfast, Lunch & Dinner)
  - HB - Half Board (Breakfast + 1 other (dinner or lunch, mostly dinner))
  - SC - No Meal package
  - Undefined - No Meal package
- Country (self - explanatory)
- market_segment (group of people who share common characteristic)

- distribution_channel (intermediaries between users and hotel booking eg. websites, travel agents, tour operators)
- is_repeated_guest (has previous booking)
- previous_cancellations (has previously cancelled a booking)
- reserved_room_type (type of room reserved)
- assigned_room_type (type of room assigned, due to high volume this can differ from reserved_room_type)
- booking_changes (no of times changes have been made to the booking)
- deposit_type
  - No Deposit
  - Non Refund - deposit of value equals total cost
  - Refundable - value under the total cost of stay
- agent
  - ID of travel agency that made the booking
- Company
  - ID of the company responsible for booking or payment
- days_in_waiting_list (no of days before the booking was confirmed)
- customer_type
  - Contract
  - Group
  - Transient
  - Transient-party
- adr (average daily rate)
  - adr = (sum_of_all_expenses)/(total_nights_of_stay)
- required_car_parking_spaces
- total_of_special_requests
- reservation_status
  - Canceled
  - Check-Out
  - No Show - Customer did not show up
- reservation_status_date
  - Date when the final changes to the entry was made.

Calculating the NA, values

```
colSums(is.na(datas))
```

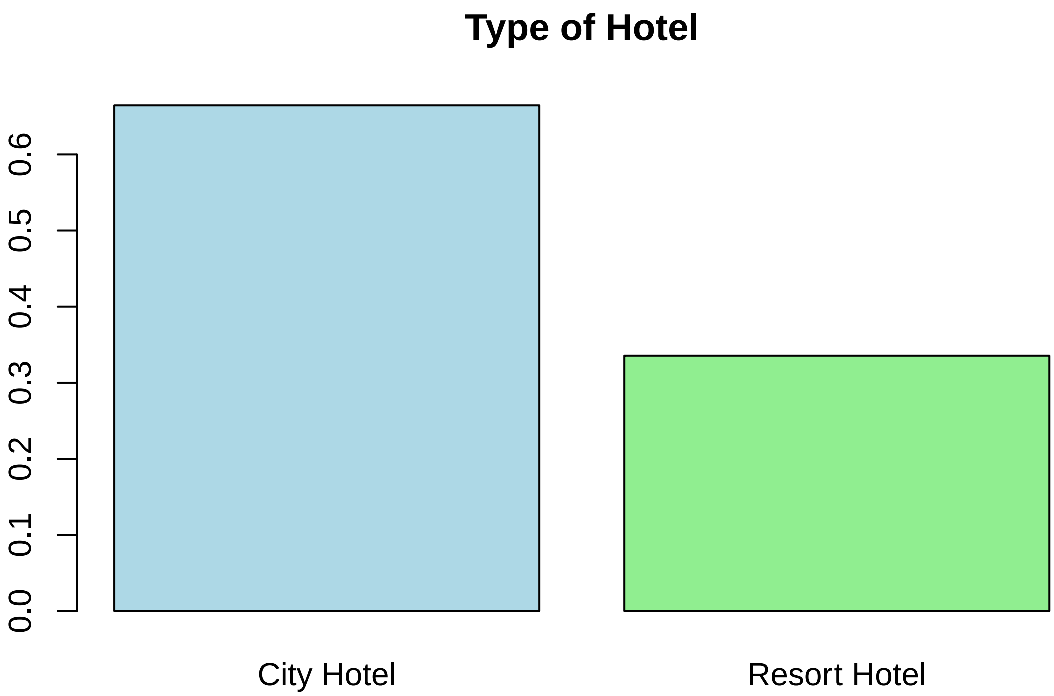
```
##                            hotel                    is_canceled
##                                0                              0
##                        lead_time              arrival_date_year
##                                0                              0
##               arrival_date_month       arrival_date_week_number
##                                0                              0
##        arrival_date_day_of_month        stays_in_weekend_nights
##                                0                              0
##             stays_in_week_nights                         adults
##                                0                              0
##                         children                         babies
##                                4                              0
##                             meal                        country
##                                0                              0
##                   market_segment           distribution_channel
##                                0                              0
##                is_repeated_guest         previous_cancellations
##                                0                              0
## previous_bookings_not_canceled            reserved_room_type
##                                0                              0
##               assigned_room_type                booking_changes
##                                0                              0
##                     deposit_type                          agent
##                                0                              0
##                          company           days_in_waiting_list
##                                0                              0
##                    customer_type                            adr
##                                0                              0
##      required_car_parking_spaces      total_of_special_requests
##                                0                              0
##               reservation_status        reservation_status_date
##                                0                              0
```

There are no such columns with na values that need to be removed or edited

Types of hotel

There are just two types of hotel, Resort & City, so a basic barplot would give the idea of the percentage of booking.

```
counts <- prop.table(table(datas$hotel))
barplot(counts, col = c('lightblue','lightgreen'), main = "Type of Hotel")
```



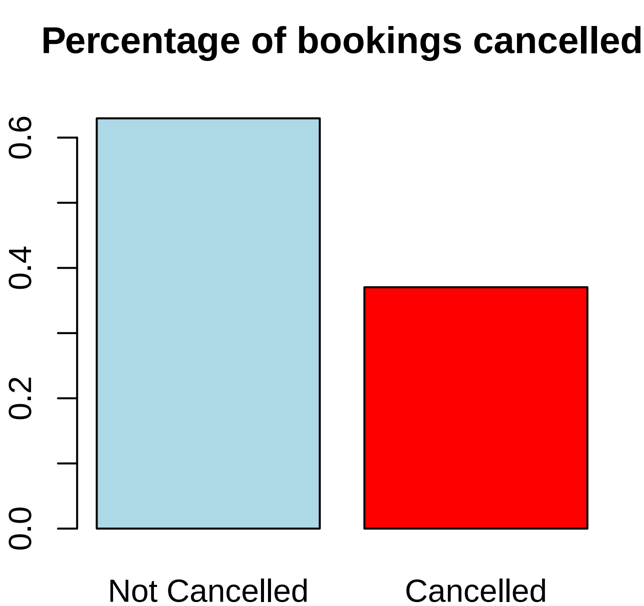
Analysis

- City hotels are twiced as much booked compared to Resort hotels, following reasons can be derived for that.
  - City hotels are better options for corporate bookings, and business purposes
  - Resort hotels can be a good option or larger parties.

Cancelled bookings

To understand what percentage of bookings are cancelled.

```
cancelled <- prop.table(table(datas$is_canceled))
barplot(cancelled, main = "Percentage of bookings cancelled",
        names.arg = c("Not Cancelled", "Cancelled"), col = c("lightblue", "red"))
```



Analysis

- Around 40% of the bookings were cancelled.

Cancellation among types of hotel

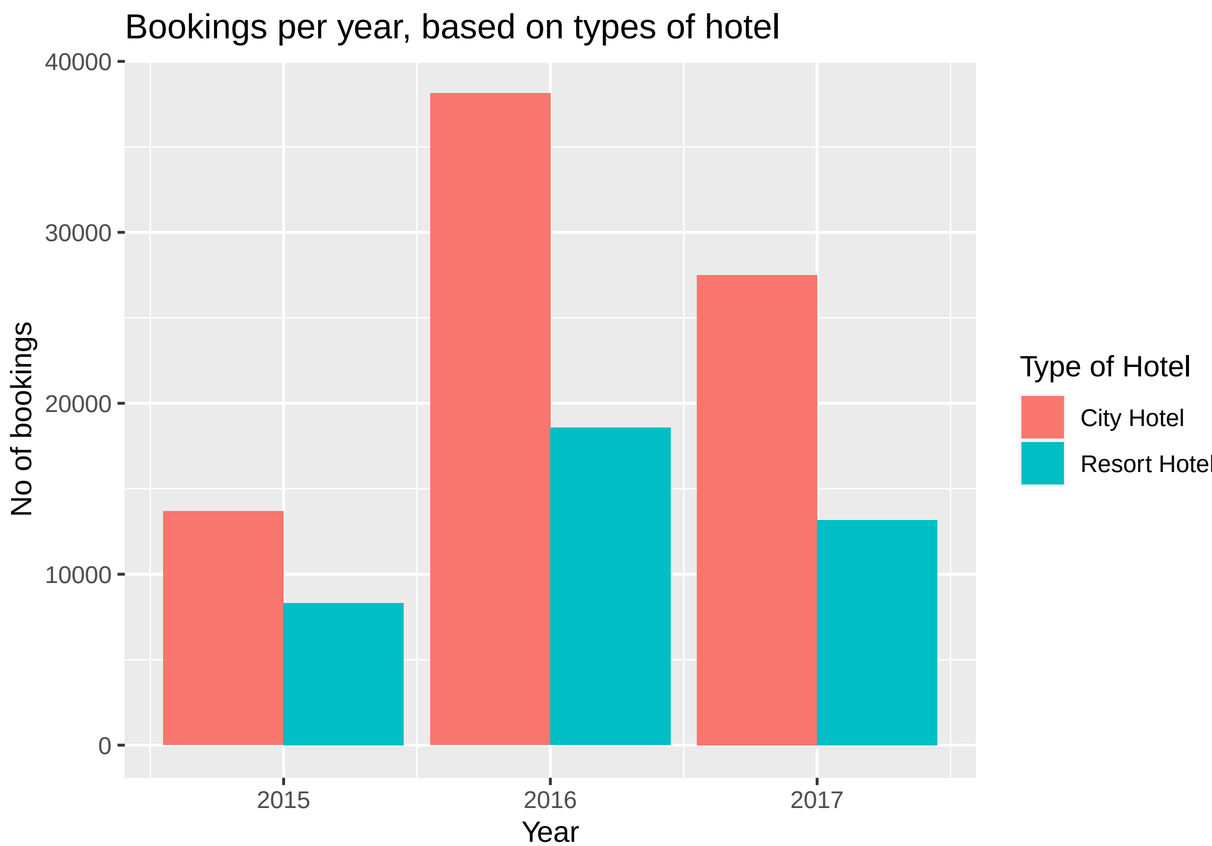
```
p <- datas %>% ggplot(aes(x=is_canceled, fill=hotel))
p <- p + geom_bar()
p <-p + xlab("Is the booking cancelled? 0 = FALSE, 1 = TRUE") + ylab("No of cancellations") + ggtitle("Can
p
```

Analysis

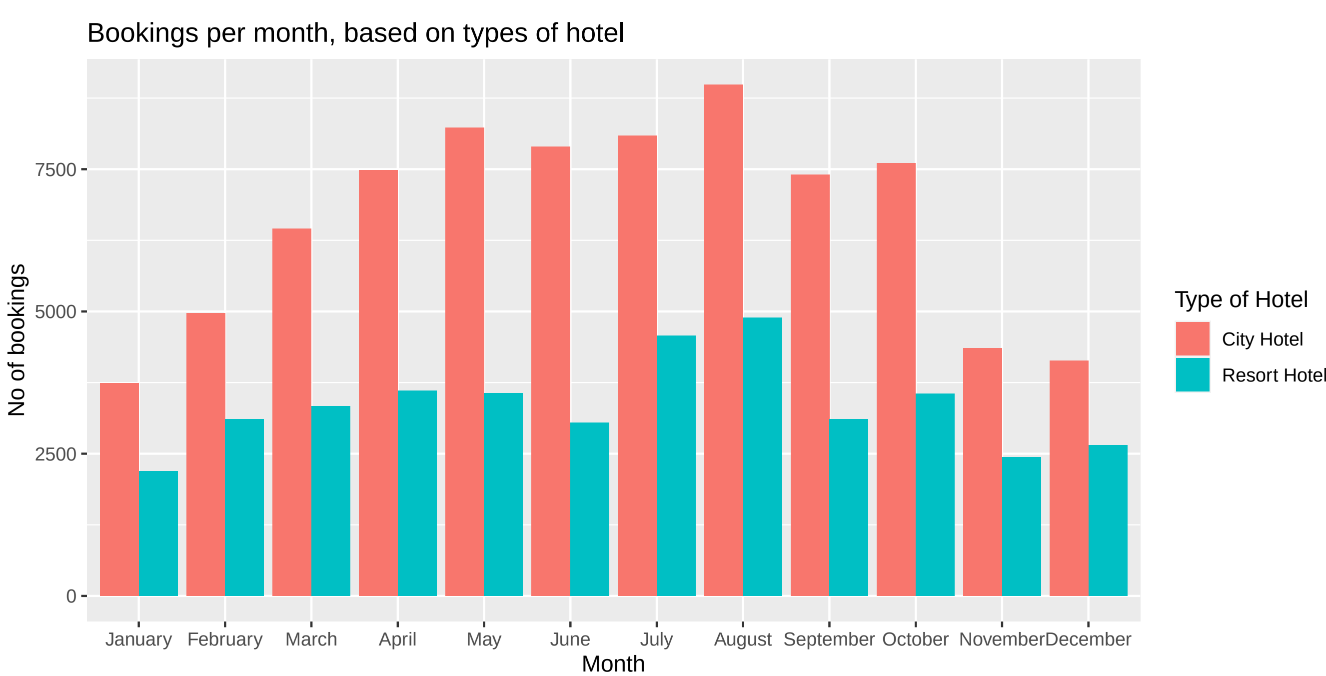
- City hotels are more likely to get cancelled in comparison to resorts.

Overview of Arrival Period

```
p <- datas %>% ggplot(aes(arrival_date_year, fill=hotel, label=hotel))
p <- p +geom_bar(position = "dodge")
p <- p + xlab("Year") + ylab("No of bookings") + ggtitle("Bookings per year, based on types of hotel") + s
p
```




```
datas$arrival_date_month <- factor(datas$arrival_date_month, levels = c("January", "February", "March", "A
p <- datas %>% ggplot(aes(arrival_date_month, fill=hotel, label=hotel))
p <- p +geom_bar(position = "dodge")
p <- p + xlab("Month") + ylab("No of bookings") + ggtitle("Bookings per month, based on types of hotel") +
p
```



Analysis

- 2016 was a good year for hotels.
- More number of hotels were booked during the summer season of `June, July and August`.