Capstone Project

Hotel Booking Analysis (EDA)

BY

Snehal D. Ramteke

INTRODUCTION

- We are here to explore a hotel booking dataset to discover important factors that govern bookings. This dataset contains booking information for a city hotel and a resort hotel and includes information such as when the booking was made. The number of adults, children, length of stay, available parking spaces & other things.
- Hotel industry is a very volatile industry and the bookings depend on the above factors and many more.

Problem Statement

1) Which are the months of highest and least occupation?
2) What is the most popular meal package?
3) How many booking changes have been done during the studied period?
4) How many people have been registered in the hotel?
5) What is the most common customer type?

Workflow

- we divide our workflow into the following three steps.

➢**EDA** will be divided into 3 following 3 analyses.

▪Univariate analysis:- It is the simplest of the three analyses where the data you are analyzing is only one variable

▪ Bivariate analysis:- It is where you are comparing two variables to study their relationship.

▪Multivariate Analysis:- It is similar to Bivariate analysis but you are comparing more than two variables

- After collecting data it's important to understand your data. So we had Hotel Booking analysis data which had 119390 rows and 32 columns. So let us understand these 32 columns.

Dataset

We are given a hotel booking dataset. This dataset contains booking information for a city hotel and a resort hotel. It contains the following features.

- **hotel**: name of the hotel whether city hotel or resort hotel.
- **is canceled** : (0 or 1) indicates whether the booking was canceled or not.
- **lead_time**: the 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 on arrival date.
- **stayed_in_weekend_nights**: The number of weekend nights stay per reservation.
- **stayed_in_week_nights**: The number of weeknights stays per reservation.
- **adults, Children, Babies**: Number of adults, children & babies arriving.
- **meal:** Types of meals booked.

Dataset(contd..)

- **Country**: The origin country of the guest.
- **market_segment**: Shows how the reservation was made and what is the purpose of the reservation.
- **distribution_channel**: The medium through booking was made.
- **is_repeated_guest**: (0 or 1) indicates whether or not the booking id of a repeated guest.
- **days_in_waiting_list**: Number of days between actual booking and transaction.
- **reserved_room_type**: Type of room booked.
- **assigned_room_type**: Type of room assigned.
- **booking_changes**: Number of changes.
- **deposit_type**: refundable /non-refundable /no-deposit made.
- **Agent**: ID of the travel agency that made the booking.

Dataset(contd..)

- **Company**: The name of the company that made the booking or is responsible paying for the booking.
- **Customer_type**: Types of customers.
- **Adr(average_daily_rate):** 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_request**: Number of special requests made.
- **reservation_status**: self-explanatory.
- **reservation_status_date**: self-explanatory.

Treating Missing Values

```
#Now lets check how many cells are missing from our dataset.
hotel_df.isnull().sum()
```

```
hotel                           0
is_canceled                     0
lead_time                       0
arrival_date_year               0
arrival_date_month              0
arrival_date_week_number        0
arrival_date_day_of_month       0
stays_in_weekend_nights         0
stays_in_week_nights            0
adults                          0
children                        4
babies                          0
meal                            0
country                       488
market_segment                  0
distribution_channel            0
is_repeated_guest               0
```

- After loading the dataset we look what is the data & variables in the dataset and after we will use the isnull function to see whether missing values are present or not.
- If it is a continuous variable and has missing values we will fill the missing values with mean or median according to the variable condition and if it is a categorical variable we will fill missing values with mode values.

We have a Column with a Missing Value

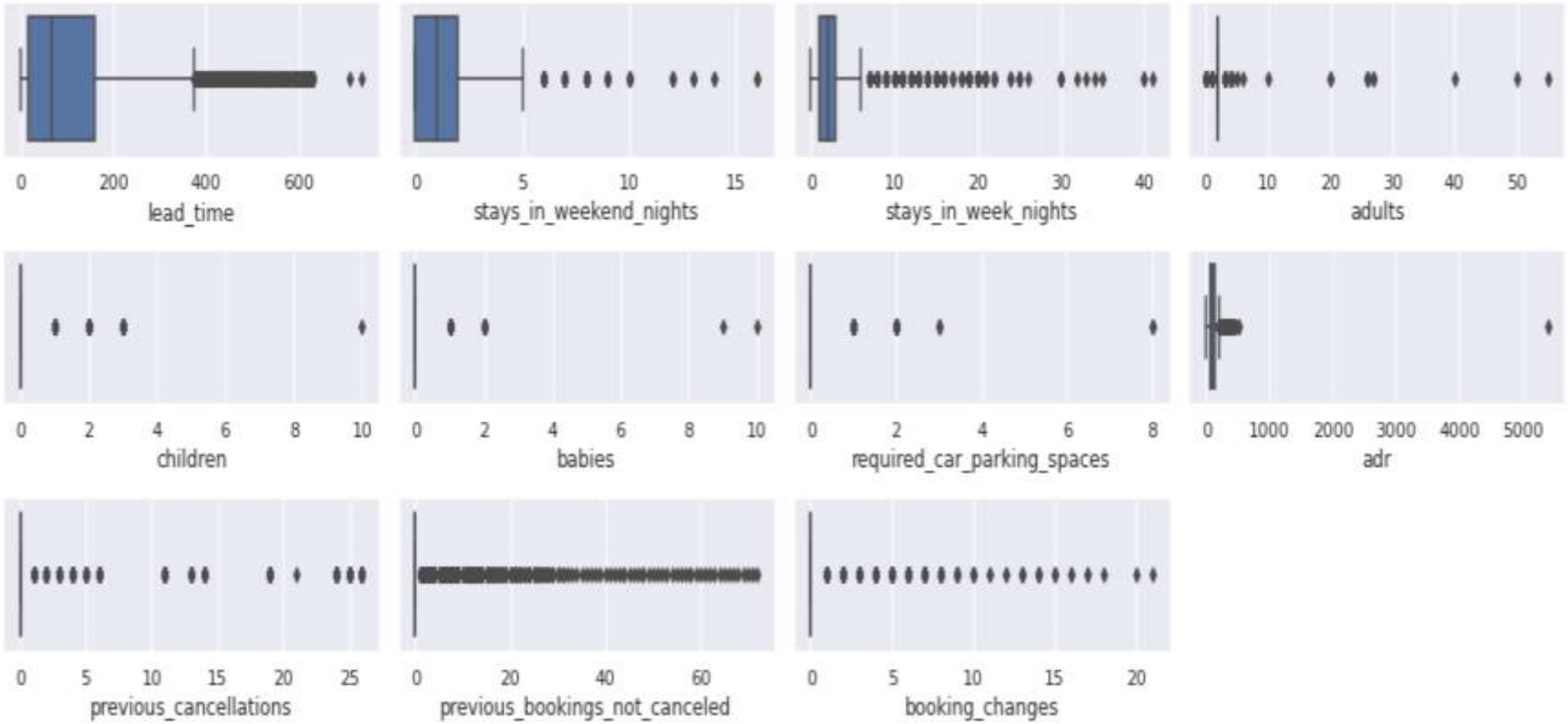
```
[11] # We delete agent and company columns
     hotel_df=hotel_df.drop(['agent','company'],axis=1)
```

- The columns “agent” and “company” have a high percentage of missing values. As these columns won’t be relevant to our analysis, we can delete them.

```
[12] # We delete rows with empty cells
     hotel_df = hotel_df.dropna(axis = 0)
```

Now we will drop the days_in_waiting_list column because we won’t use it for this analysis

We have seen some outliers



Exploratory Data Analysis (EDA)

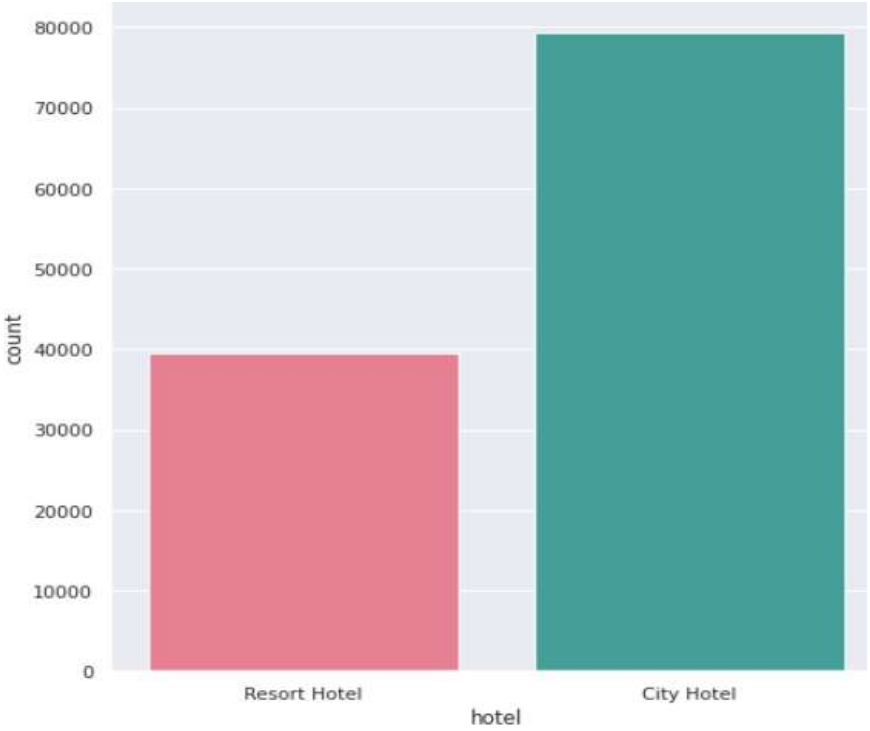
- we will explore the data to get insights about it.
- There are the following types of visualization below –

- Hotel Type
- Reservation Status
- Total Members
- Cancelled Booking
- Distribution Channel
- Deposit Type
- Booking Changes
- Customer Types
- Repeated Guest

Continued

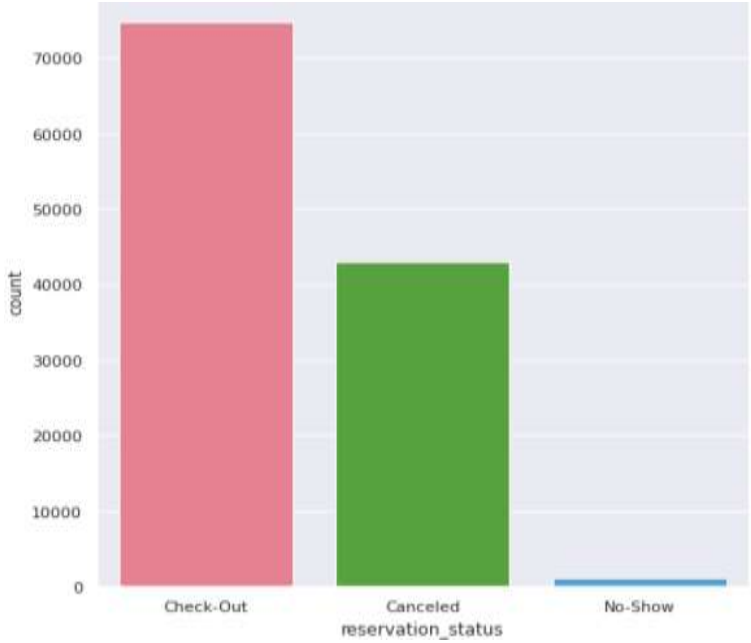
- Reserved Room Types
- Assigned Room Types
- Market Segment
- Meal
- Country
- Year
- Months
- Average Daily Rate (ADR)

- Hotel Types



- Proportion of reservations between hotel types.

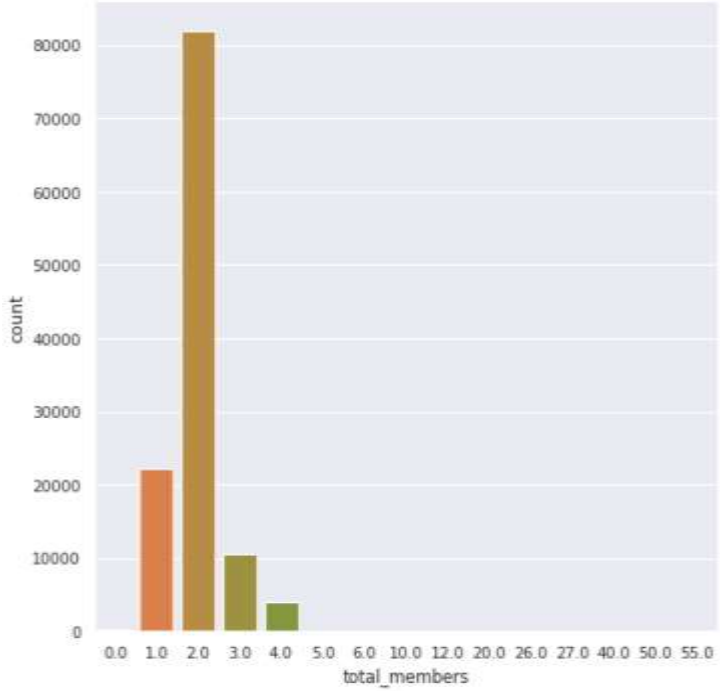
- Reservation Status



Reservation the 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 — the customer did not check in and did inform the hotel of the reason why

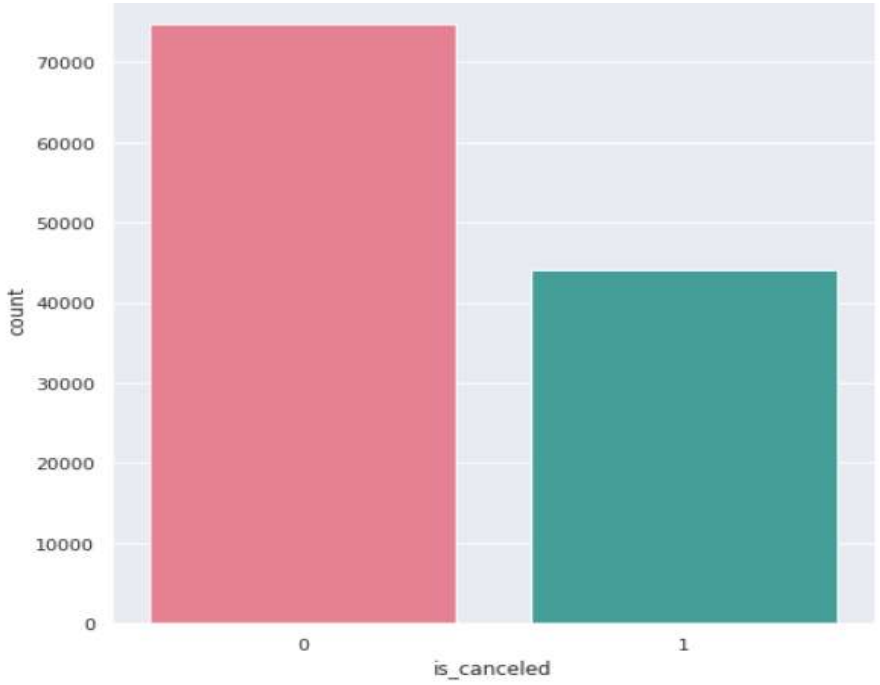
-Total Members

Total members per reservation



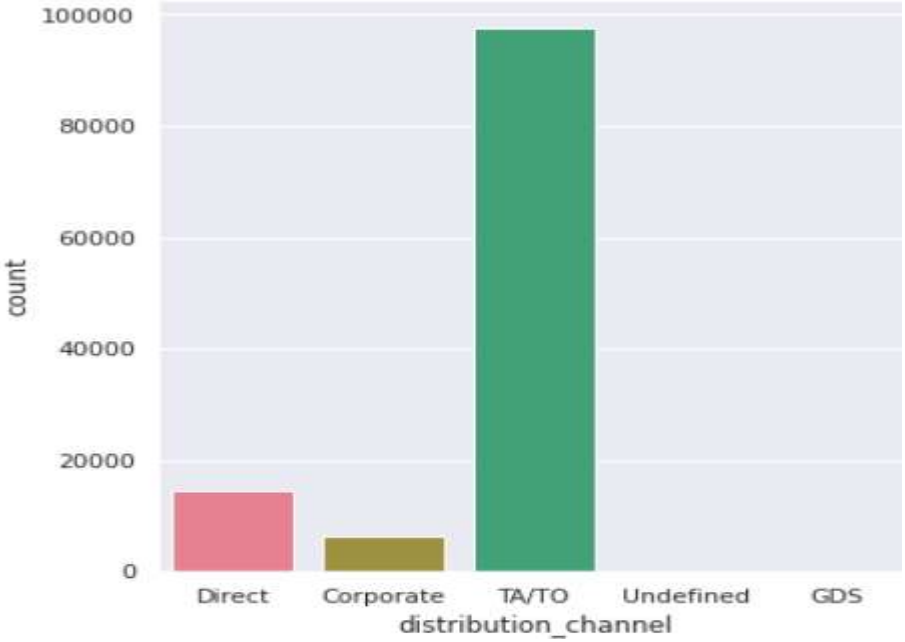
-Cancelled Booking

During the year, we have a 37.13% of cancelations.



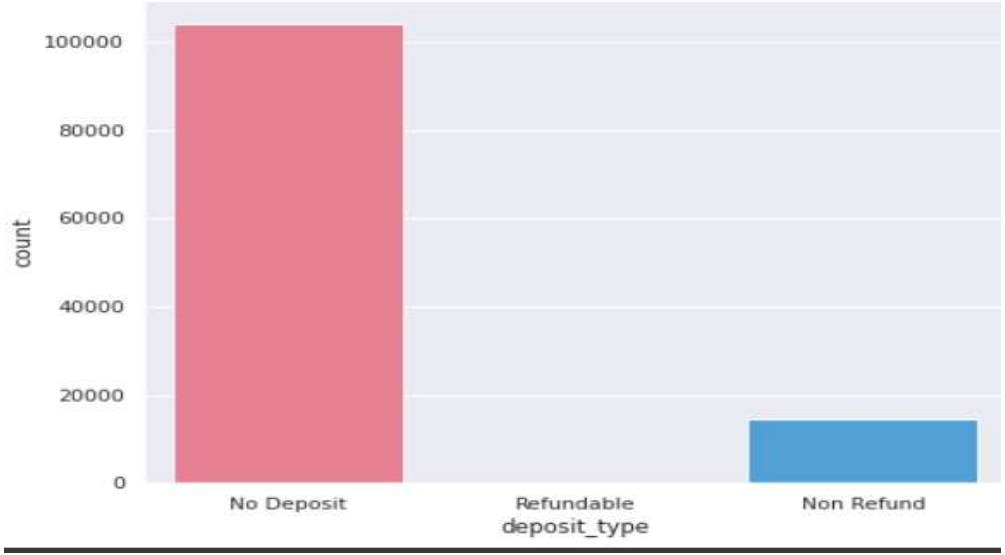
-Distribution channel

Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”



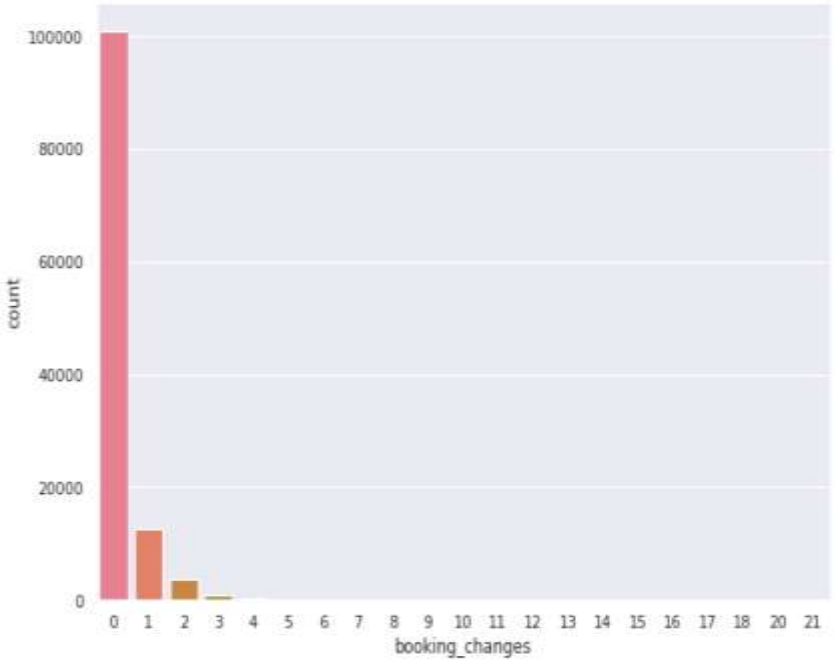
–Deposit Type

Indication 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 the stay.



-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



-Customer Types

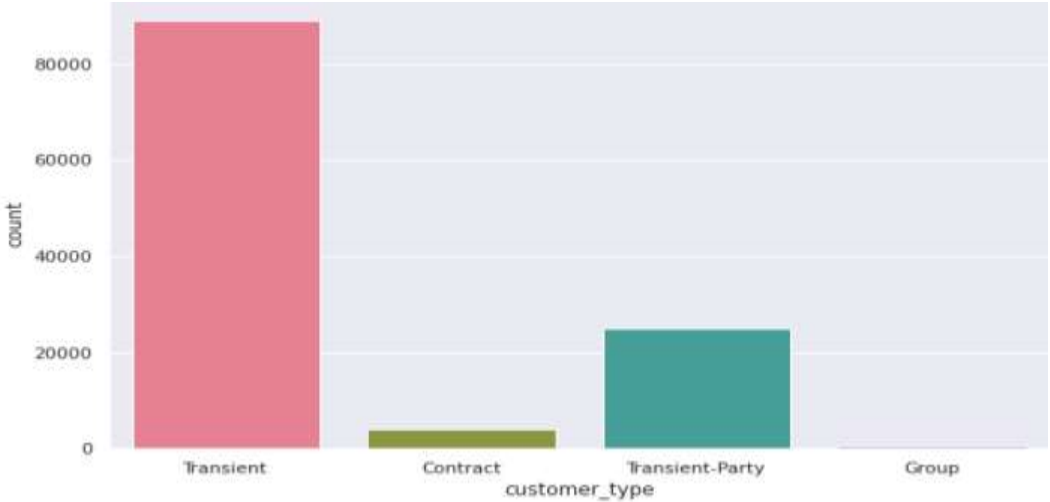
Type of booking, assuming one of four categories:

•Contract — when the booking has an allotment or other type of contract associated with it;

•Group — when the booking is associated with a group;

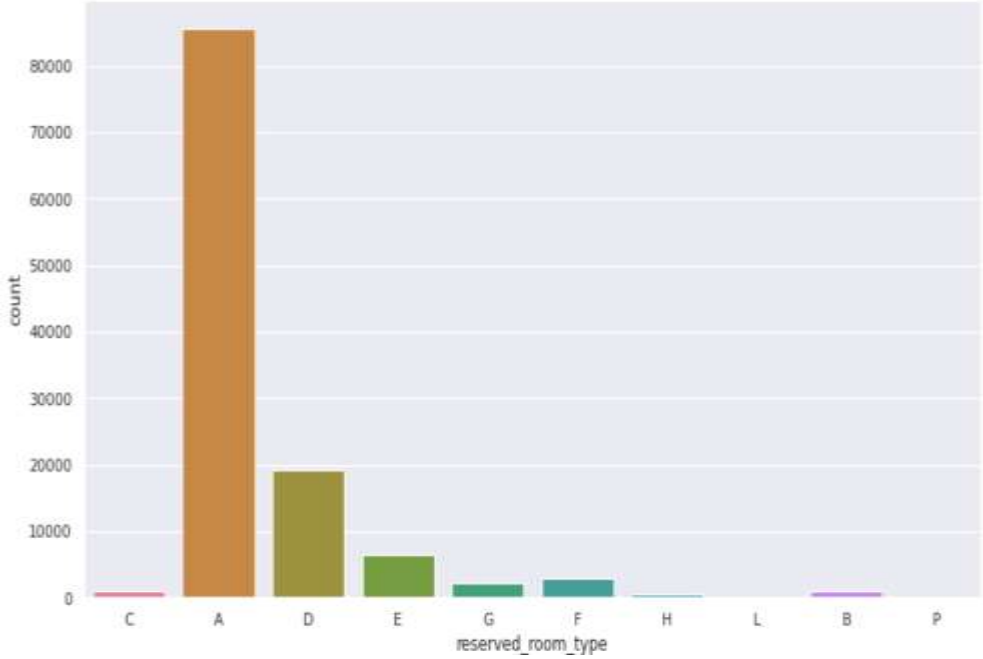
•Transient — when the booking is not part of a group or contract and is not associated with another transient booking;

•Transient-party — when the booking is transient but is associated with at least another



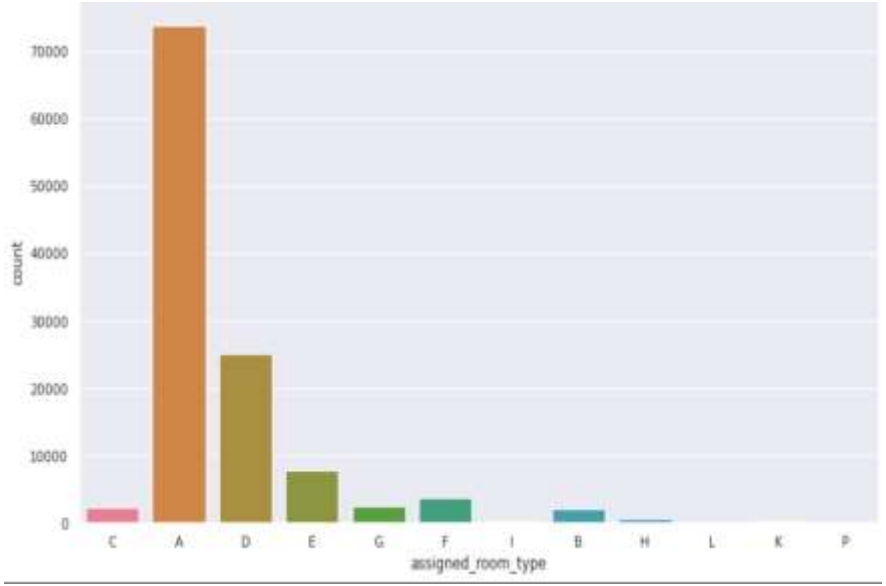
-Reserved Room Types

Code of room type reserved. Code is presented instead of designation for anonymity reasons.



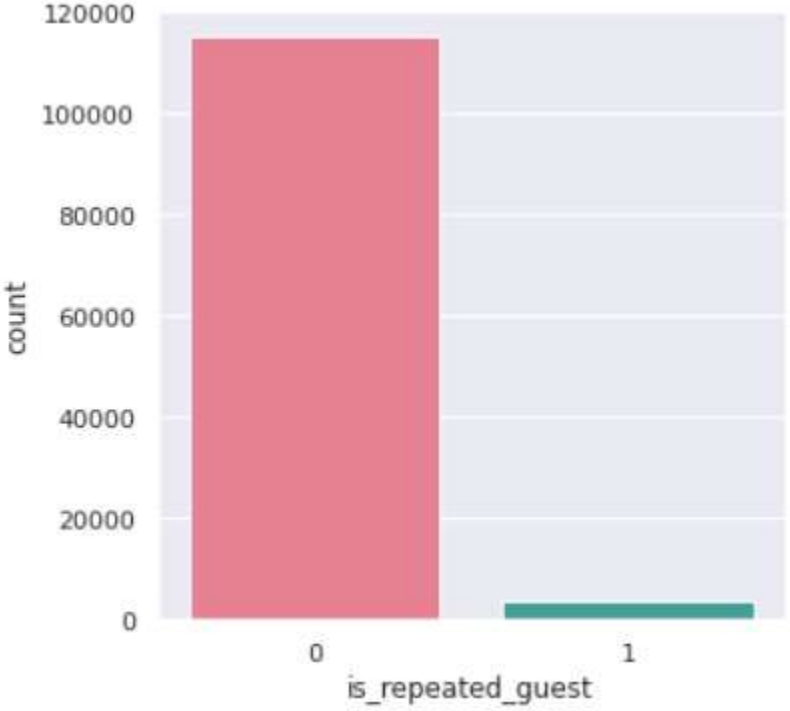
-Assigned Room Types

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.



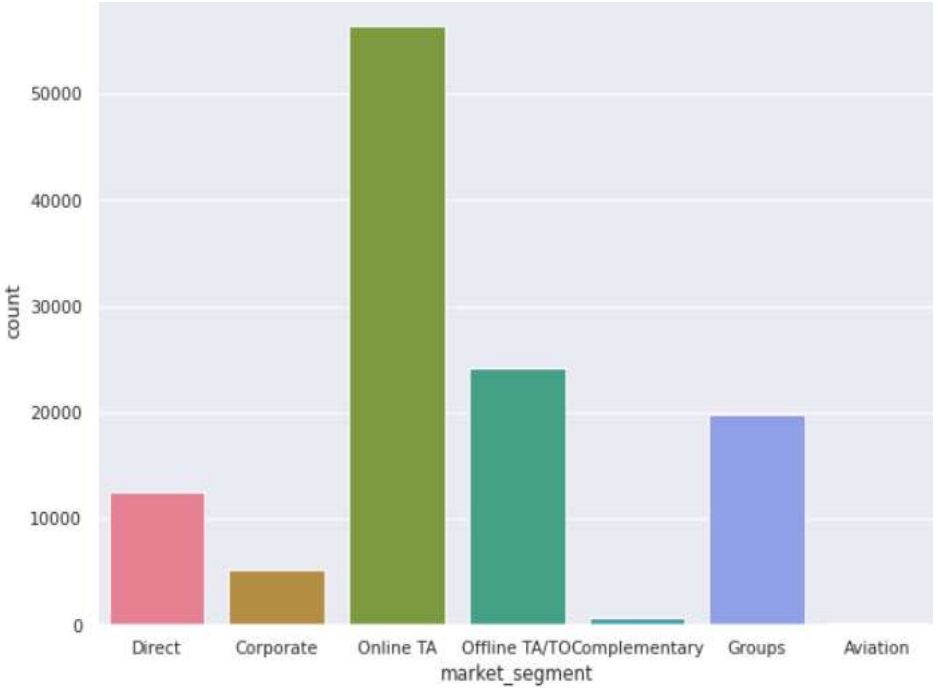
-Repeated Guest

The booking name was from a repeated guest (1) or not (0)

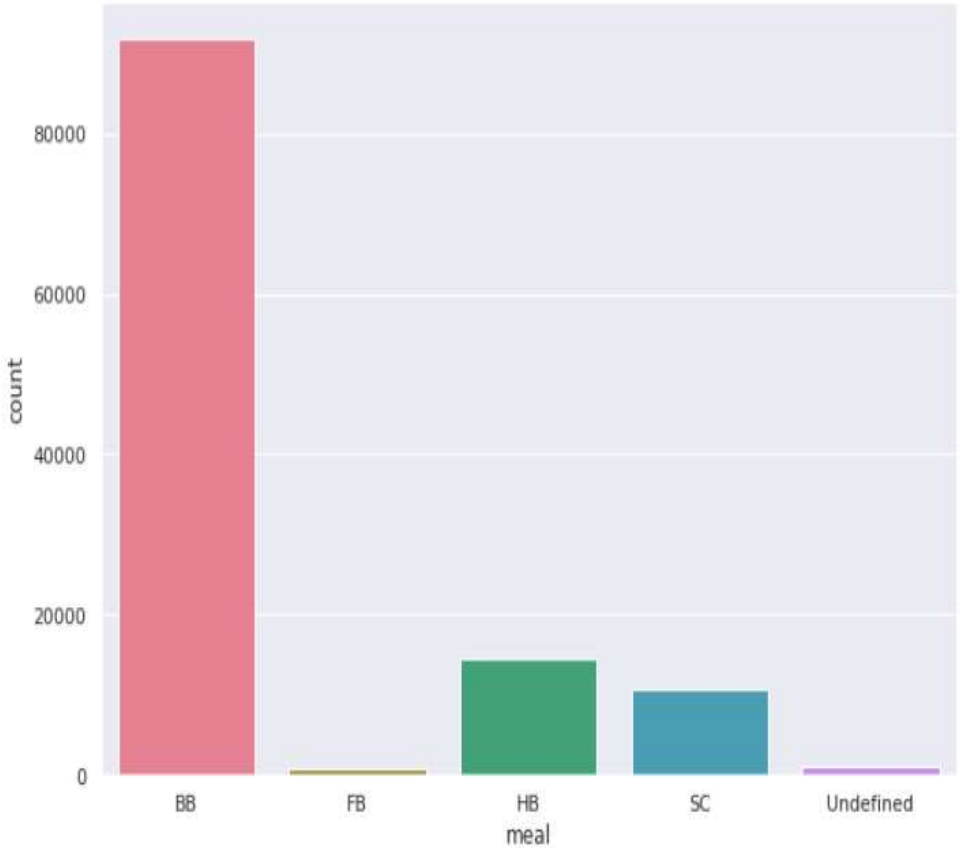


-Market Segment

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



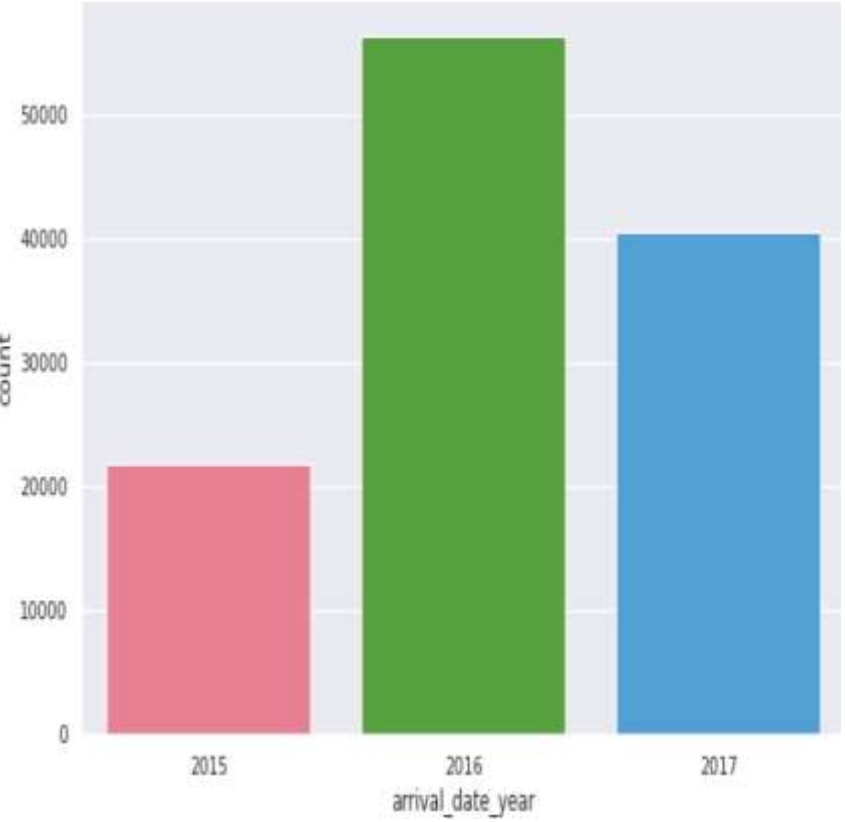
-Meal



- Type of meal booked.
- Categories are presented in standard hospitality meal packages:
- Undefined/SC — no meal package;
- BB — Bed & Breakfast is the most preferred type of meal by the guests.
- Full board (FB) is the least preferred.
- HB — Half board (breakfast and one other meal — usually dinner);
- FB — Full board (breakfast, lunch, and dinner)

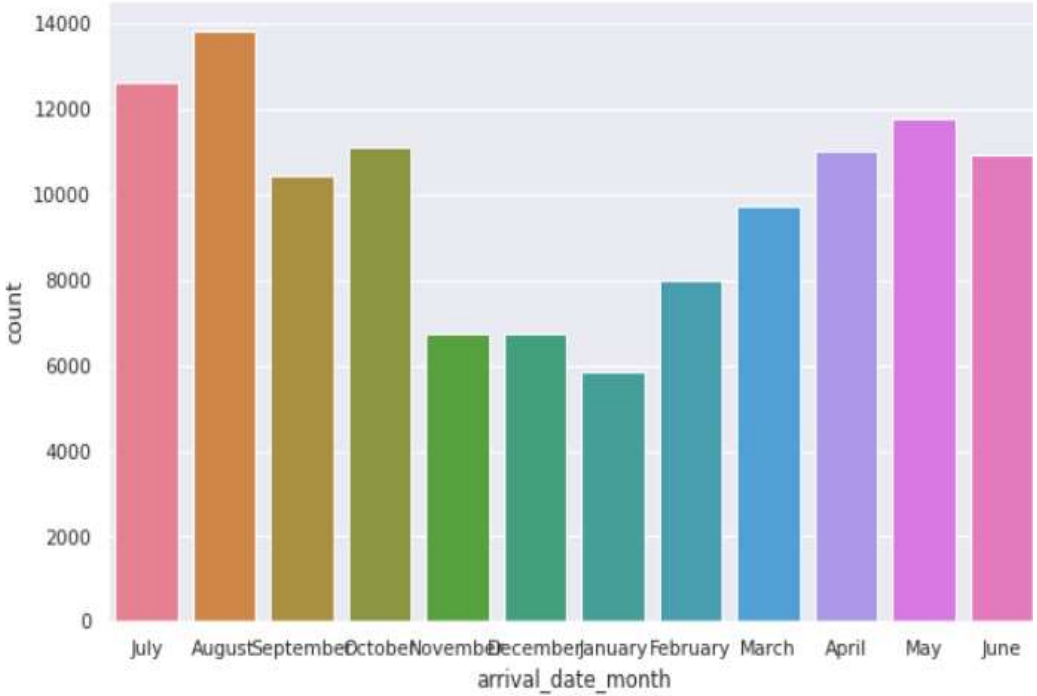
-Year

- Year of arrival date



-Month

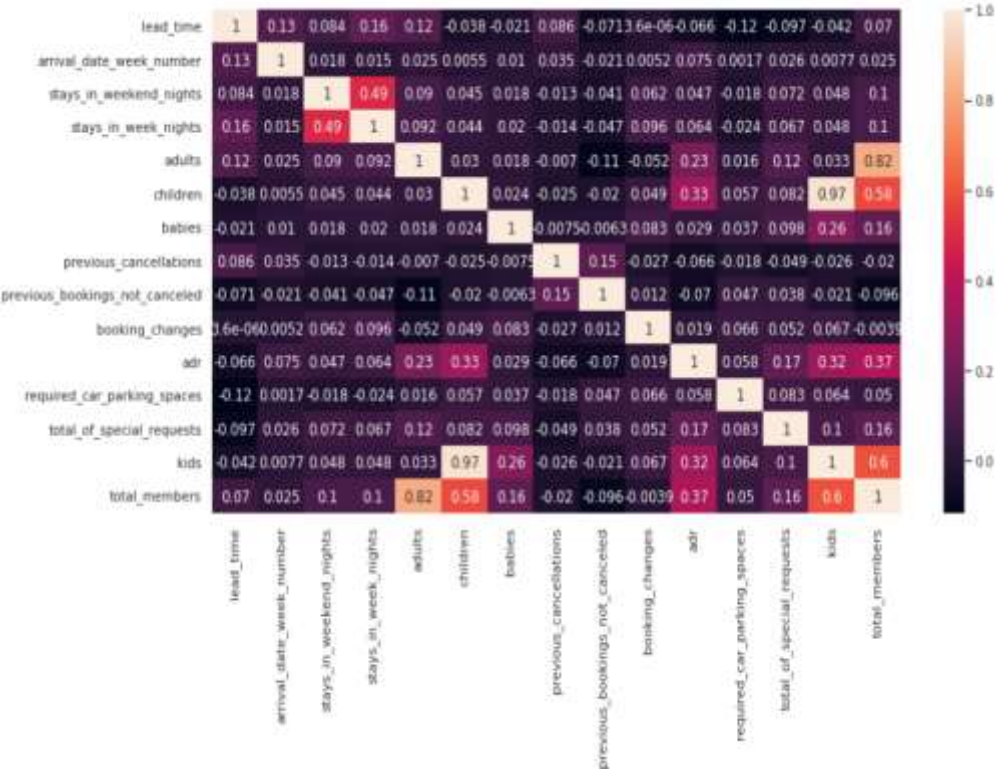
the arrival date by months



- Average Daily Rate

For resort hotels, the average daily rate is more expensive during august, July, and September.
For city hotels, the average daily rate is more expensive during august, July, June, and may.

correlation matrix



- 1) Total stay length and lead time have a slight correlation. This may mean that for longer hotel stays people generally plan little before the actual arrival.
- 2) adr is slightly correlated with total_people, which makes sense as more no. of people means more revenue, therefore more adr.
- Let's see does the length of stay affects the adr.

Conclusion-

- **Lead -time and total stay are positively correlated meaning the more are the stay of the customer more will be the lead time.**
- **ADR(average daily rate) and total people are highly correlated. That means more people will be the adr. High adr means high revenue.**

Conclusion

1) City hotels are the most preferred hotel type by guests. We can say the City hotel is the busiest hotel**.**
2) In this exercise we see if there are any missing values present or not.
3) the majority of reservations are for city hotels.
4) The majority of guests come from western Europe countries.
5) Here ADR(average daily rate) is correlated with no_of_people. It simply indicates that more revenue is generated when the number of people increases.
6) The number of repeated guests is too low.
7) Waiting time period for City hotels is high as compared to resort hotels. That means city hotels are much busier than Resort hotels.
8) The month of highest occupation is august with 11.65% of the reservations. The months of least
9) The majority of reservations convert into successful transactions.
10) Resort hotels have the most repeated guests.

THANK U