Capstone Project - I

EDA on Hotel Booking Data

BY

Yashwant B. Raul
(yashwantraul24@gmail.com)

Mayur S. Marathe
(marathemayu1990@gmail.com

Sanket Gawali
(sanketgawali23@gmail.com)

Project Overview

- We will perform exploratory data analysis with **Hotel Booking Dataset**.
- 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.

- 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.

Process Overview

Data collection and Understanding

Data cleaning and Preprocessing

Exploratory Data Analysis(EDA)

Conclusion

Data Collection And Understanding

After collecting data 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 Description

hotel :Resort Hotel or City Hotel
is_canceled : Value indicating if the booking was canceled (1) or not (0)
lead_time : Number of days that elapsed between the entering date of the booking and the arrival date
arrival_date_year : Year of arrival date
arrival_date_month : Month of arrival date
arrival_date_week_number : Week number of year for arrival date
arrival_date_day_of_month : Day of arrival date
.

Data Collection And Understanding

stays_in_weekend_nights : Number of weekend nights
stays_in_week_nights : Number of week nights

adults : Number of adults

children : Number of children

babies : Number of babies

meal : Type of meal booked.

country : Country of origin.

market_segment : Market segment designation. (TA/TO)

distribution_channel : Booking distribution channel.(T/A/TO)

is_repeated_guest : is a repeated guest (1) or not (0)

previous_cancellations : Number of previous bookings that were cancelled by the customer prior to the current booking

previous_bookings_not_canceled : Number of previous bookings not cancelled by the customer prior to the current booking

reserved_room_type : Code of room type reserved

Data Collection And Understanding

assigned_room_type : Code for the type of room assigned to the booking.
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 : No Deposit, Non Refund , Refundable.
agent : ID of the travel agency that made the booking
company : ID of the company/entity that made the booking .
days_in_waiting_list : Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type : Type of customer. Contract,Group,transient,Transient party.
adr : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
required_car_parking_spaces : Number of car parking spaces required by the customer
total_of_special_requests : Number of special requests made by the customer (e.g. twin bed or high floor)
reservation_status : Reservation last status.

Data Cleaning & Preprocessing

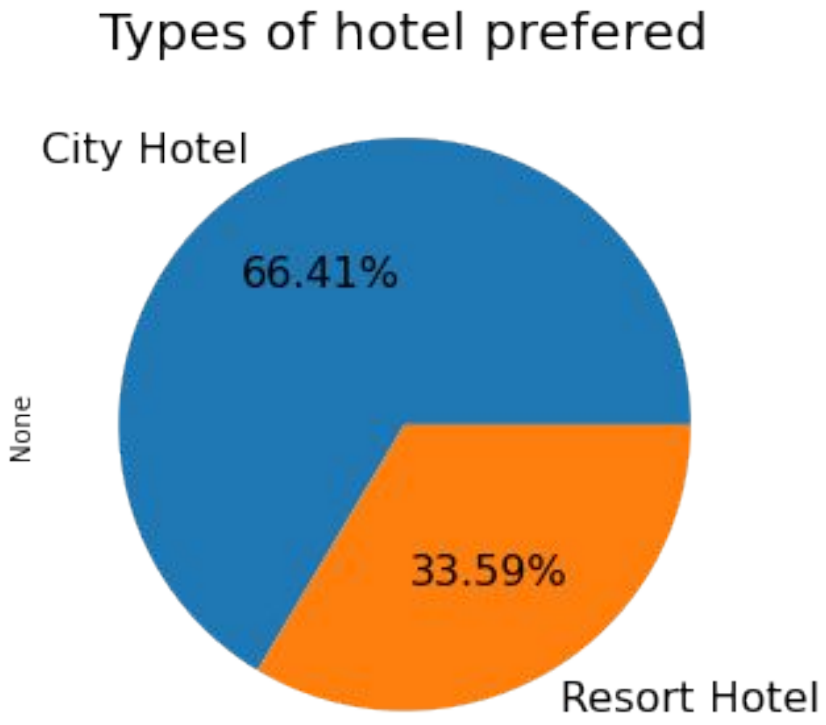
Drop the columns(Company, Agent).
Drop the rows(No guest available).
Filling all null/missing values by 0.
Convert data type into required format.
Replace negative values with 0.
Copy the data frame so that there is not any changes in original data frame.
Create data frame with specific column.

Data Cleaning & Preprocessing

```
[ ] # Drop the column company and agent from the dataframe
    df.drop(['company','agent'],inplace=True,axis=1)

[ ] # filling all null values with 0 in column children
    df['children'] = df['children'].fillna(0.0)

    # convert datatype from float to int in column children
    df['children'] = df['children'].astype(int)

[ ] # Replace the null values of country column with mode value
    df['country'].fillna(df['country'].mode().to_string(), inplace=True)

[ ] # Drop the Rows where guests not available
    df = df.drop(df[(df.adults+df.babies+df.children)==0].index)

[ ] # replace all negative values in adr by 0
    df.loc[df.adr <= 0, 'adr'] =df['adr'].fillna(0, inplace=True)
```

Exploratory Data Analysis(EDA)

Types Of Hotels

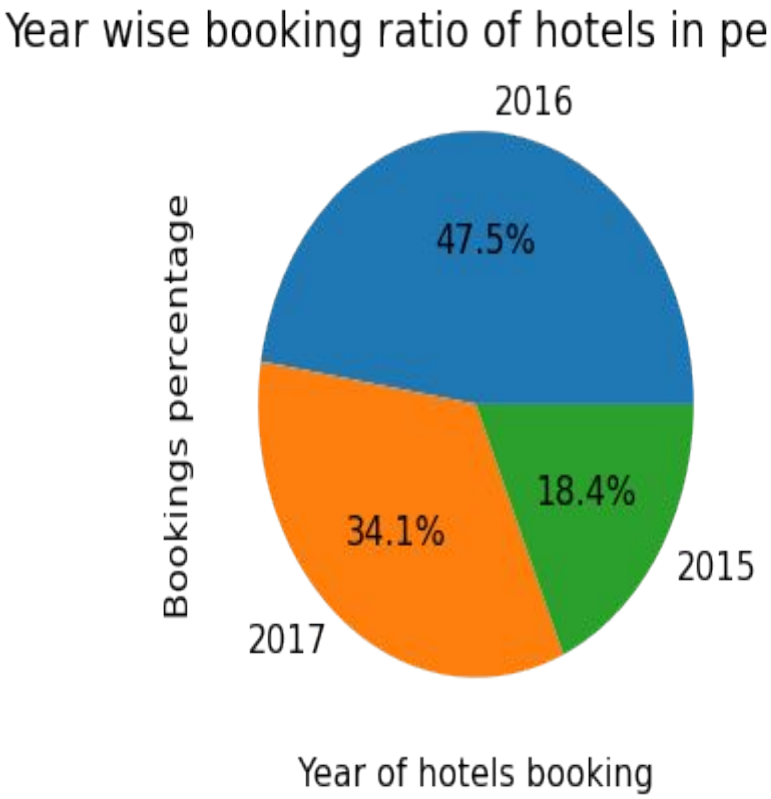


- We have two types of Hotels in the dataset

  Resort Hotel

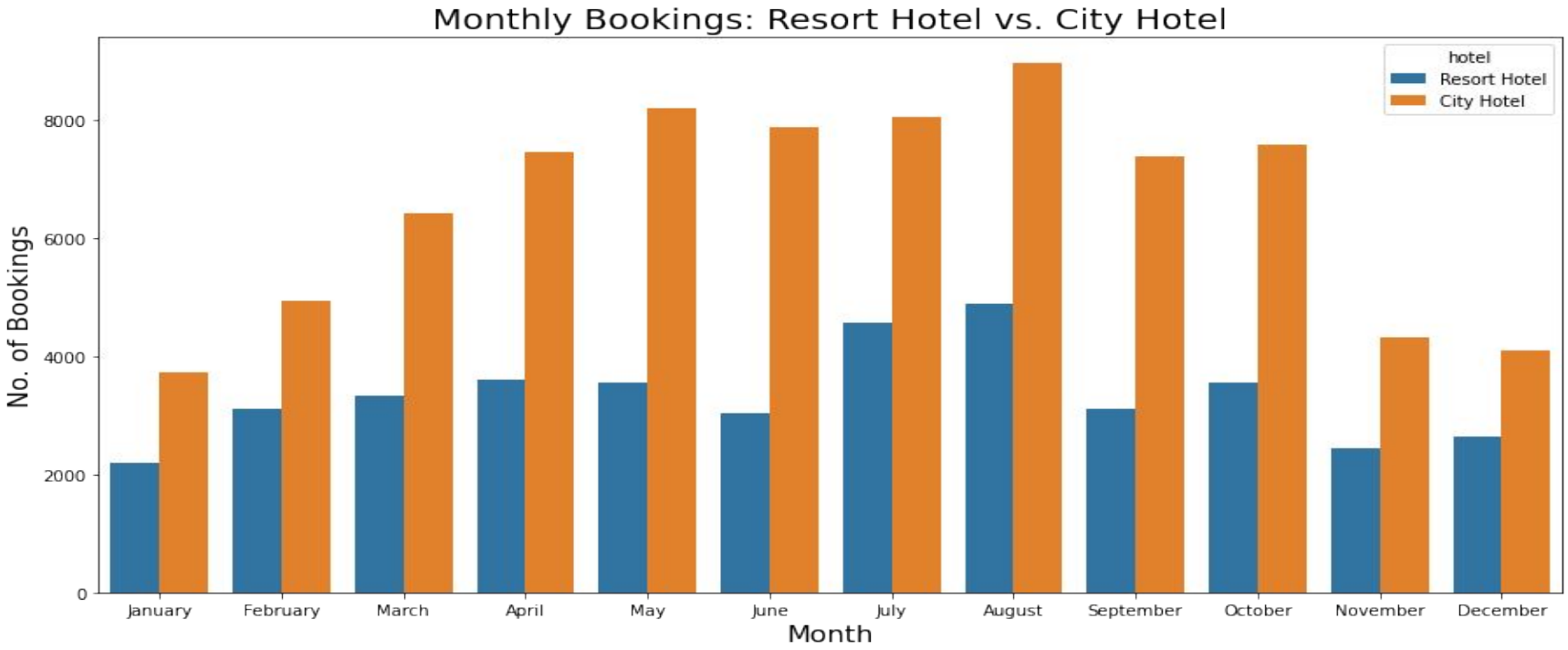
  City Hotel
- We can see that the City Hotels are most preferred by the guest.

Yearly Booking Percentage



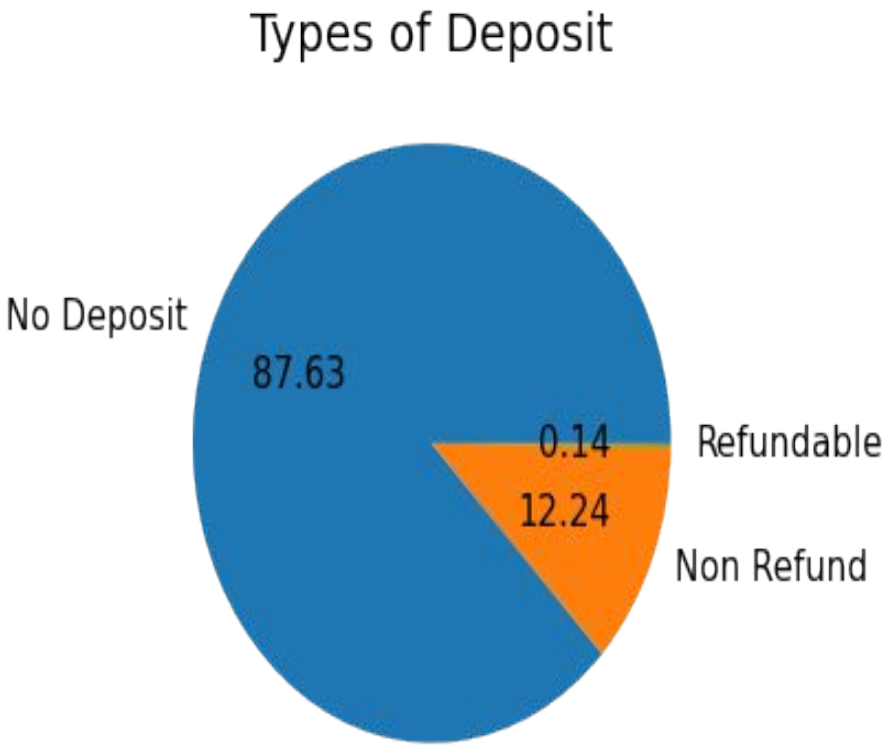
Bookings for 2016 were highest with 48%. Bookings were increasing on yearly basis. For 2016 increment was too high so there was steep fall in bookings for 2017.

Monthly Bookings



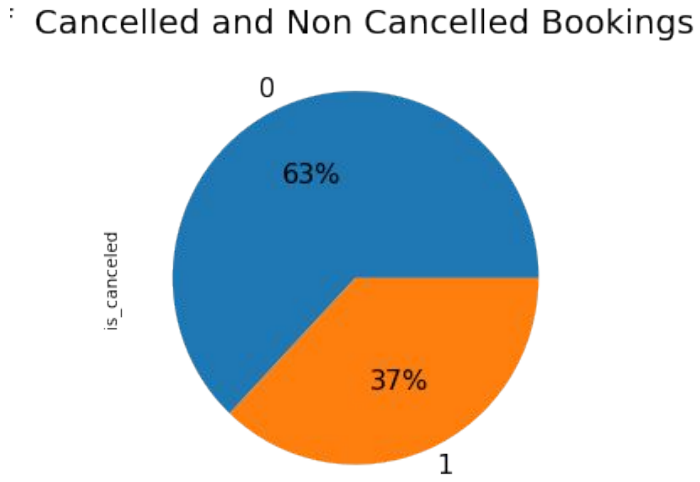
In month of August saw the highest number of hotel bookings for both type of hotels. The lowest hotel bookings were in January.

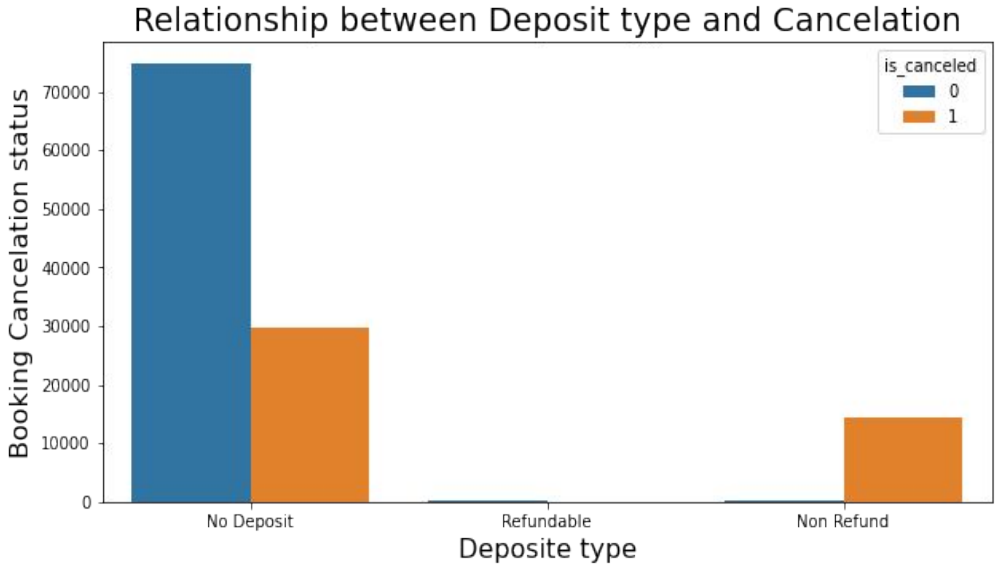
Type Of Deposit



- About 87.63% bookings are without any deposit. 12.24% bookings deposit are non refundable. Guests preferred no deposit type bookings because flexibility to cancel or change the bookings without losing any money.

Deposit Type Vs Cancelation



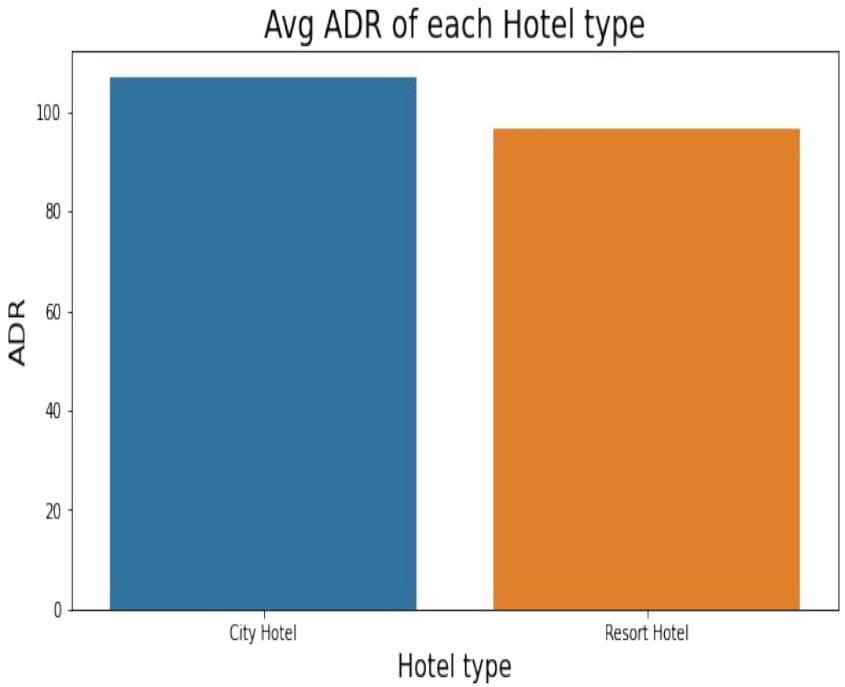


0= not cancled,1= canceled.

37 % of the bookings were cancelled.

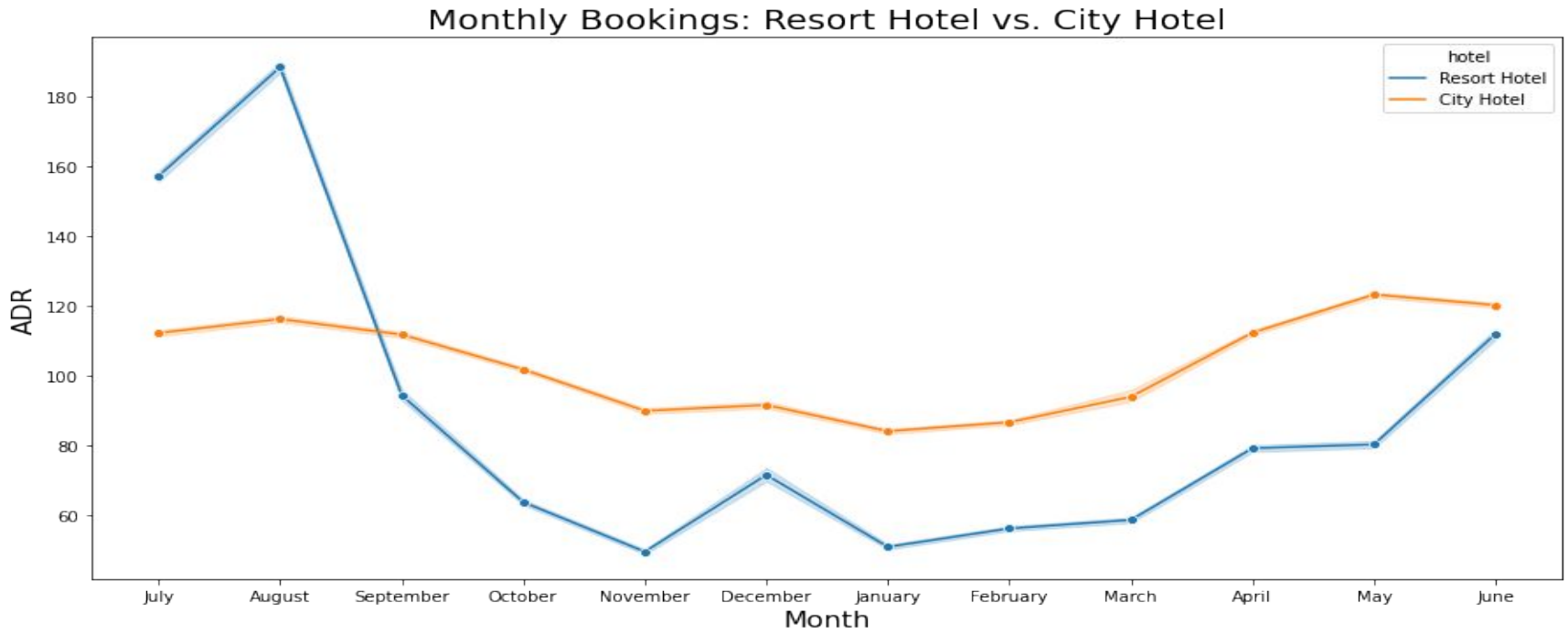
Cancelation of bookings in city hotel were higher than resort hotel. That could explain why the booking with no deposit was higher. So we can say that guests who do not pay any deposit while booking are likely to cancel more reservations

ADR COMPARISON



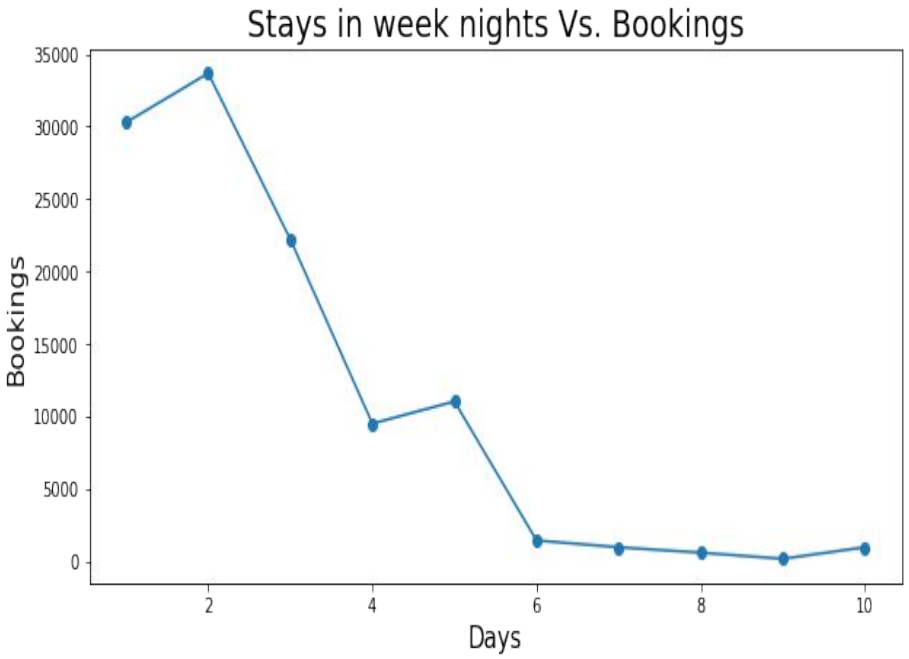
City hotel has the highest ADR. That means city hotels are generating more revenues than the resort hotels

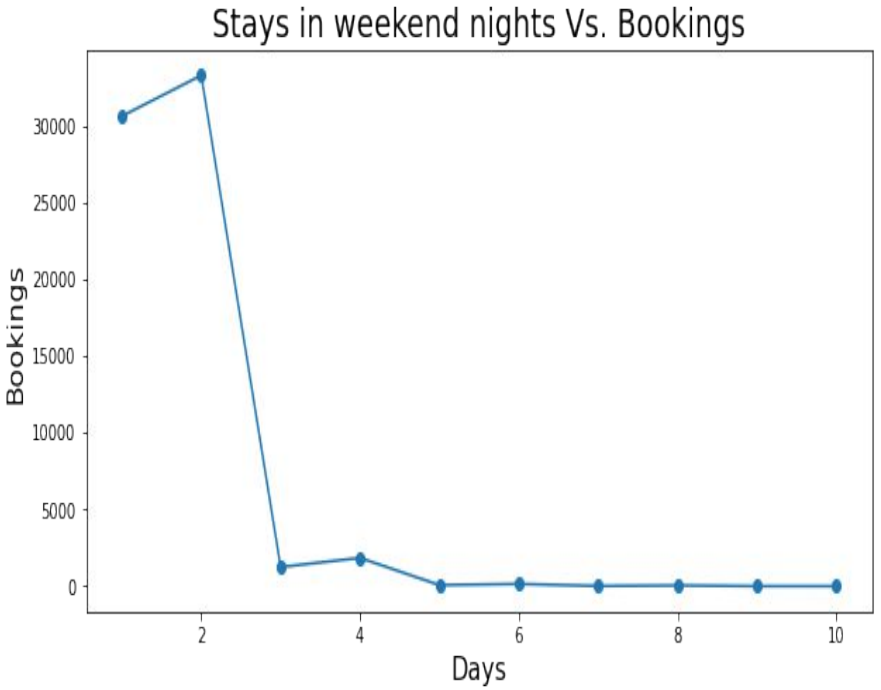
Average Daily Rate Vs. Month



- For both city and resort hotels, Nov to Jan have cheaper average daily rates so this is the best time of year to book a hotel . Also we can see prices for resort hotels are higher and fluctuate more than city hotels.

Stays in Weekend Nights Vs. Week Nights



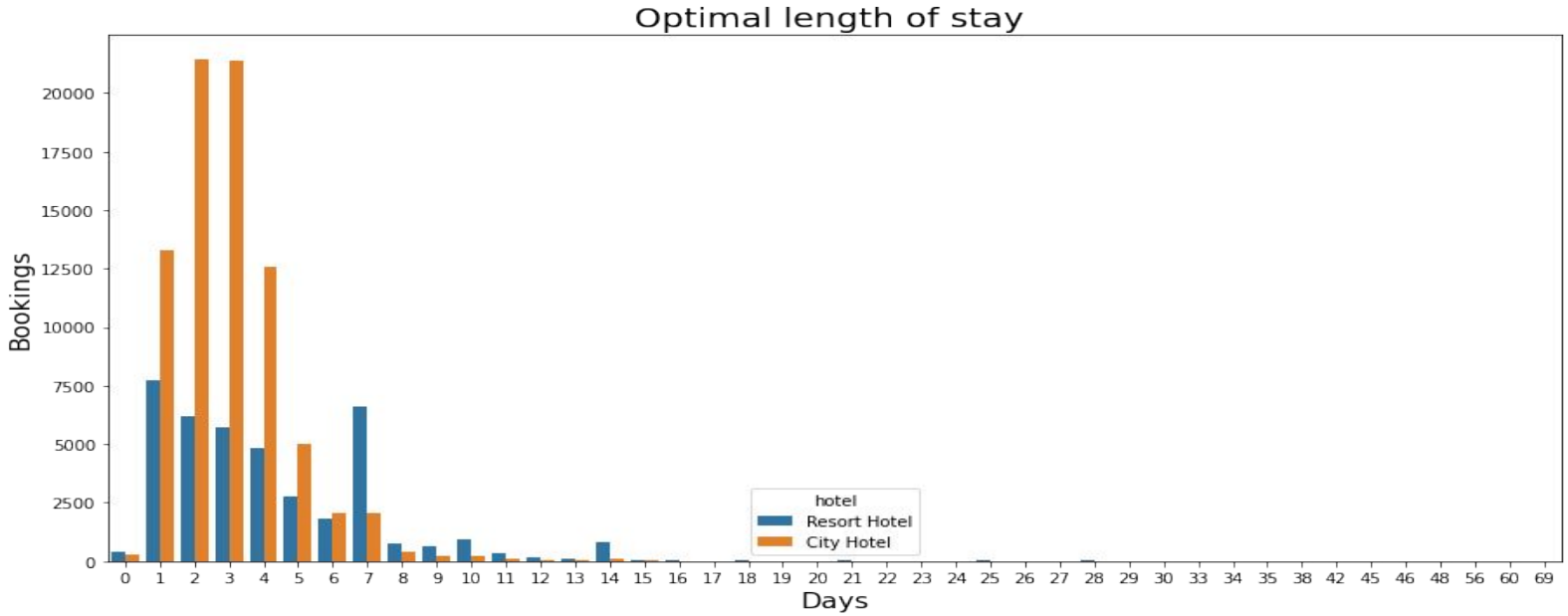


This chart helps us visualize the variation of stays in weekend nights and stays in week nights.

We can see that the majority stays by the guests is during weeknights.

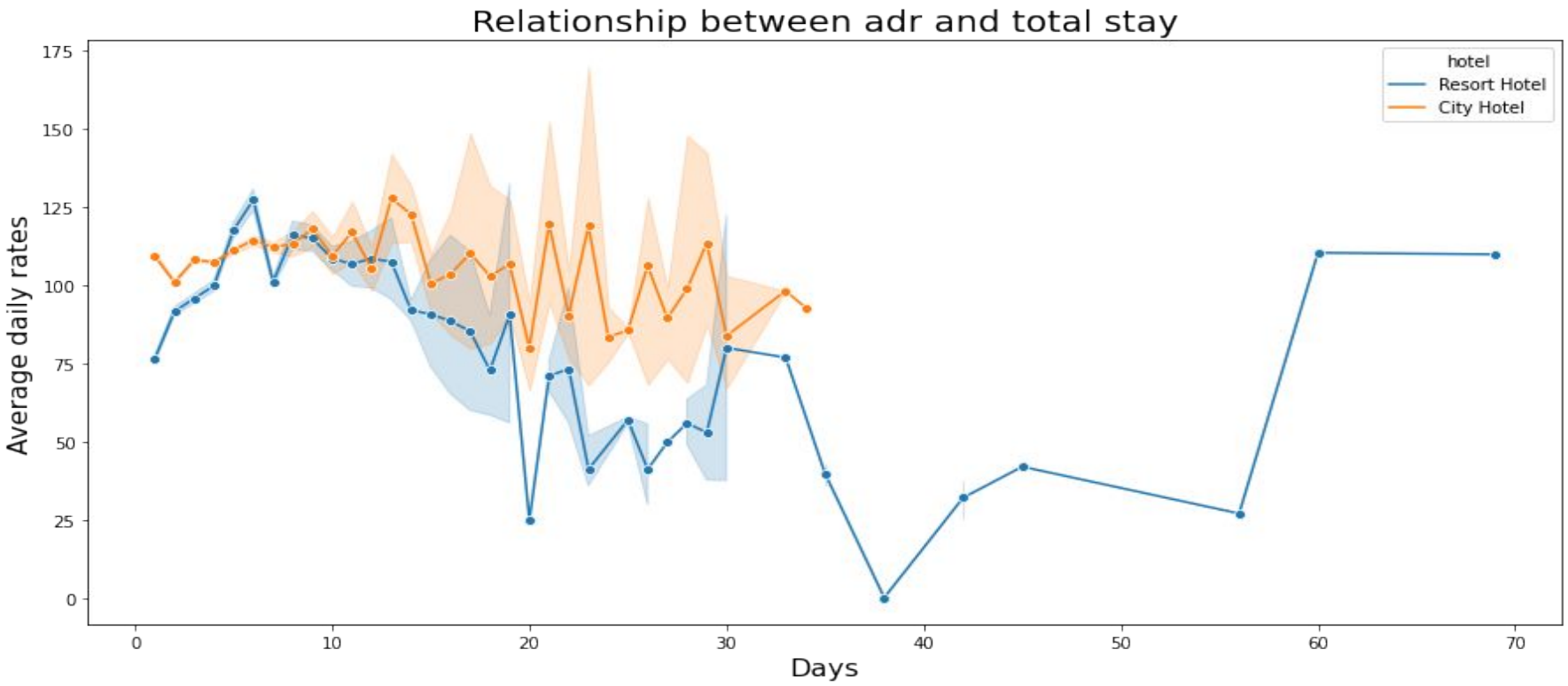
Most of the bookings were for 2 night stays in either in week nights or weekend night

Length Of Stay



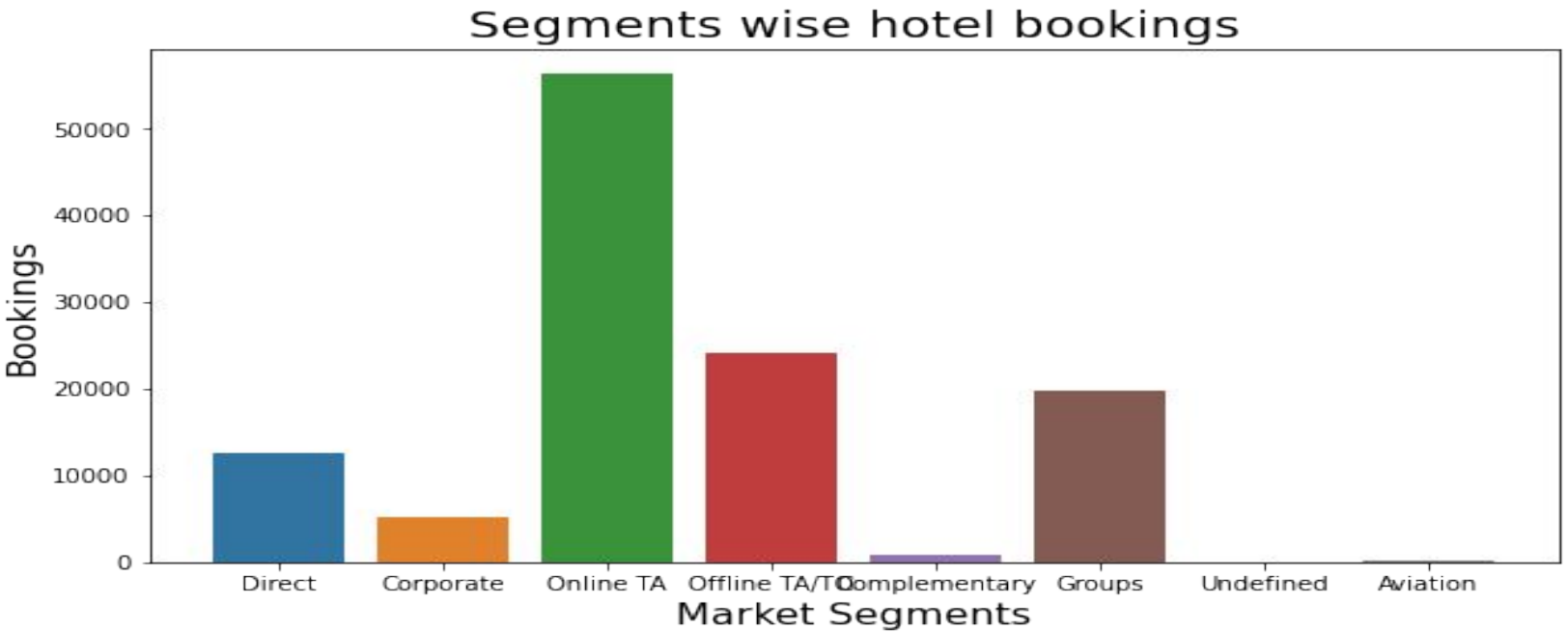
Optimal stay in both the type hotel is less than 7 days but For long stay guest prefer Resort hotel.

ADR Vs. Total Length Of Stay



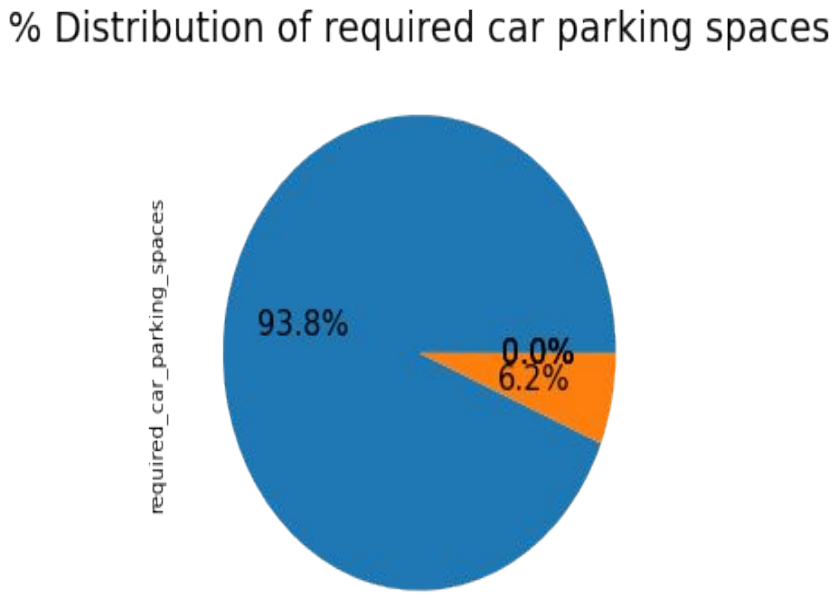
- As increase in length of stay, adr gets decreases. To get best adr need to stay more than Month.

Market Segment



More involvement in bookings with Online TA and Offline TA/TO that means maximum hotel are booked with agents only.

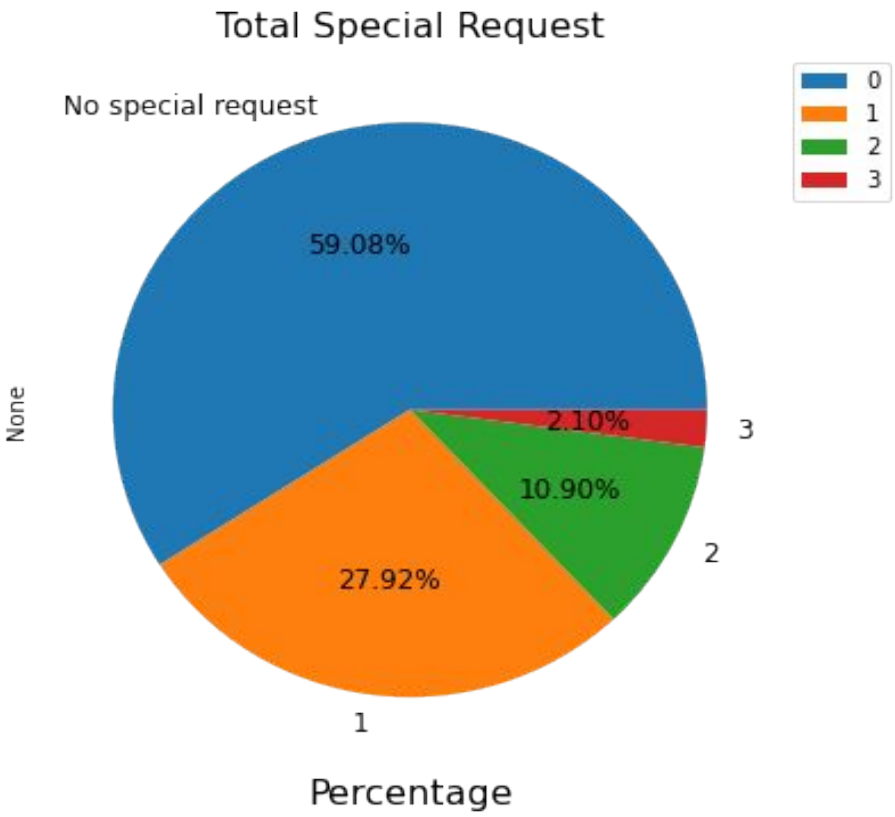
Car Parking Requirements



Almost 94% of bookings, there is no requirement for car parking spaces.

Almost 6.2% of bookings, there is requirement of one car parking space.

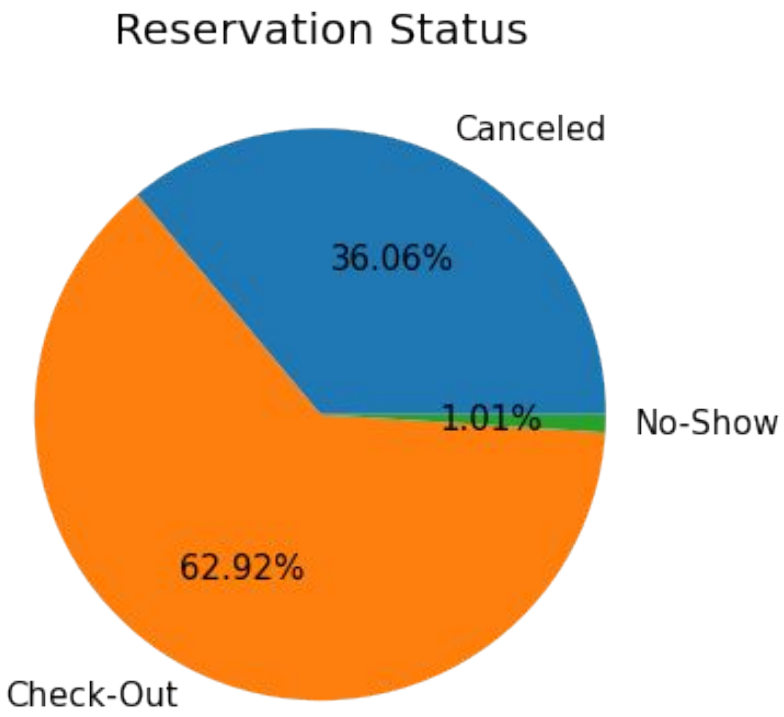
Special Request



About 59% of bookings there are no special requests.

27.92% of bookings there is one special request.

Reservation Status



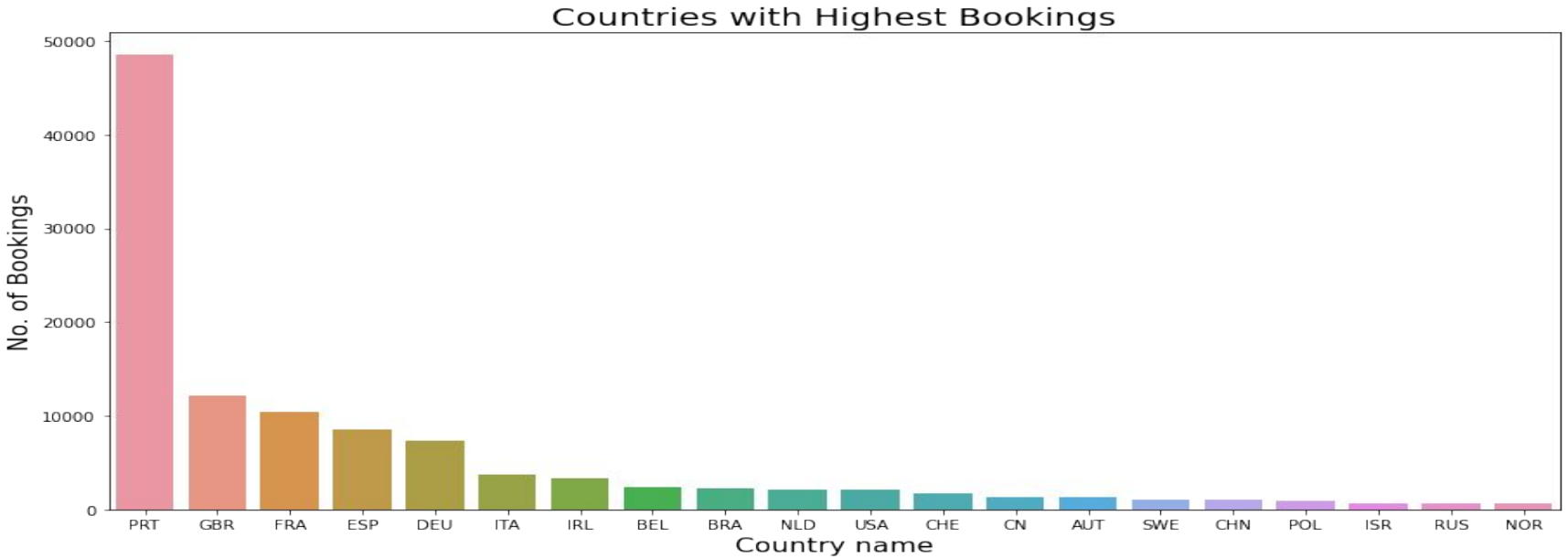
Here we can see the bifurcation of the reservation status of the guests in the hotels.

We can see that 62.92% of guests go forward with their booking

36% of bookings got canceled.

1% of bookings never arrived at hotel without canceling bookings

Countries

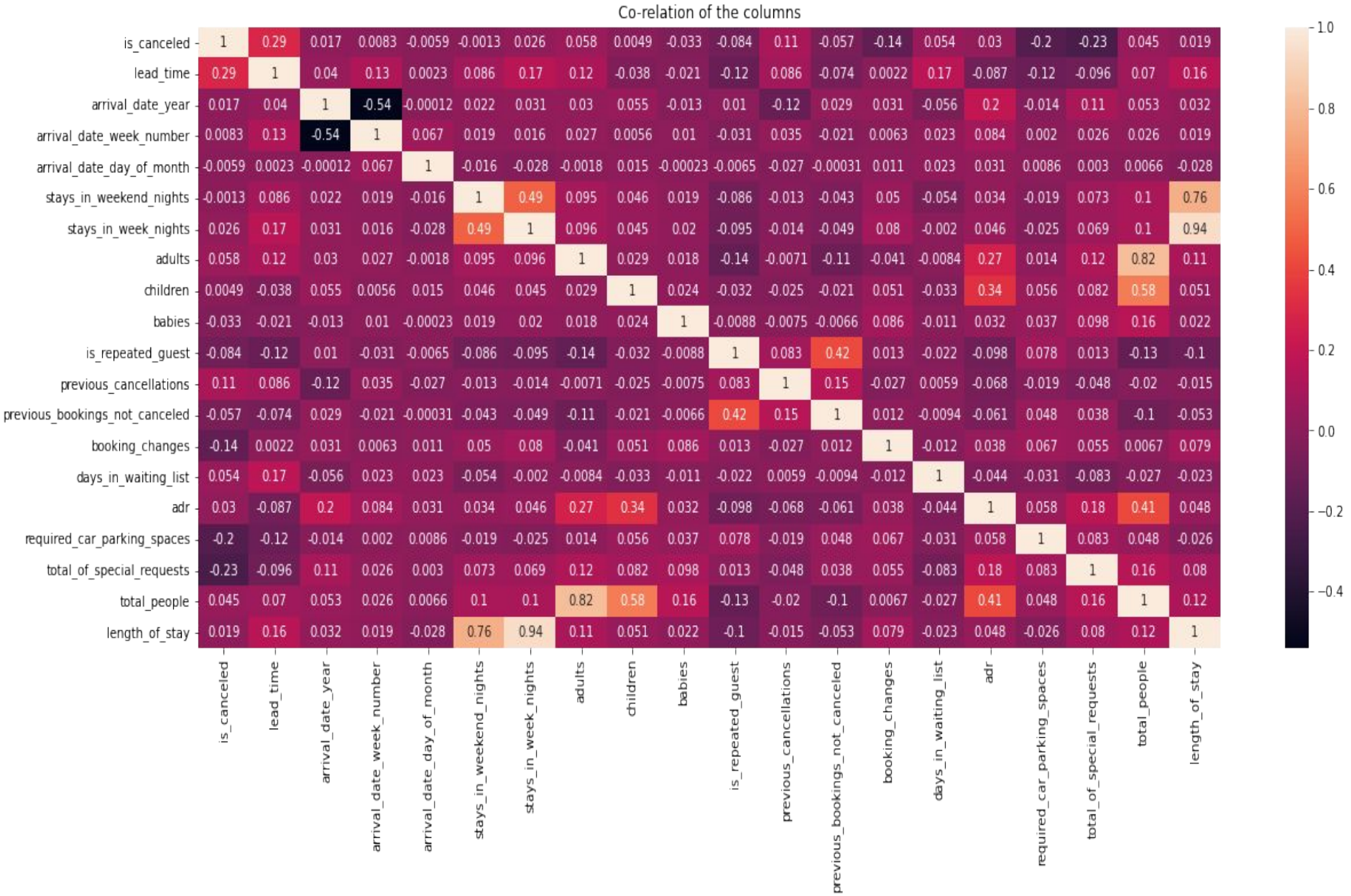


Our dataset implies that the guests arrive from all over the world.

With the help of this visualization we can see the top 5 countries from where the hotels receive most number of guests.

Most of the guest are coming from Portugal(PRT),Great Britain(GRB), France(FRA),Spain(ESP) and Germany(DEU).

Correlation Of Data



Lead-time and total length of stay is positively correlated. That means more is the stay of customer more will be the lead time.

Total people are positively correlated to each other. As increase in people ADR will be more.

is repeated guest and previous bookings not canceled has strong correlation. may be repeated guests are not more likely to cancel their bookings.

As increase in special request ADR will be more.

Special request are more for an adult.

Conclusion

City hotels having maximum booking, which are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.

Considering the years, Bookings for 2016 were highest with 48%. Bookings were increasing on a yearly basis. For 2016 increment was too high so there was a steep fall in bookings for 2017.

The month of August saw the highest number of hotel bookings for both types of hotels. The lowest hotel bookings were in January. Daily rates are also high in the month of August Where low in January so January is the best time to book a hotel with cheaper rate & confirmed booking.

Guests preferred no deposit type bookings because of the flexibility to cancel or change bookings without losing any money. About 87.63% bookings are without any deposit. 12.24% bookings deposit are non refundable.

Conclusion

Out of total booking 37% booking were canceled. With high no. of booking, cancellation is also higher in city hotels and out of that canceled booking, hotels with no deposit have higher cancellation rate.

Repeated guests are not more likely to cancel their bookings.

Considering the rates City hotel has the highest ADR. That means city hotels are generating more revenues than the resort hotels.

Most people prefer a 2 night stay. Optimal length of stay is 7 days. As an increase in length of stay ADR decreases. To get the best adr you need to stay more than a month.

For long stay guests prefer resort hotels.

More involvement in bookings with Online TA and Offline TA/TO that means maximum hotels are booked with agents only.

Conclusion

Special requests are more for an adult. As an increase in special requests and increasing in people ADR will be more.

Most of the guests are coming from Portugal(PRT),Great Britain(GRB), France(FRA),Spain(ESP) and Germany(DEU).

Finally checkout the reservation status, 36% of bookings got canceled.1% of bookings never arrived at hotel without canceling bookings.