Capstone Project - 1

Hotel Booking - Exploratory Data Analysis

Team Members

Nihal Habeeb

Parvez Makandar

Contents

- Introduction to Exploratory Data Analysis
- Hotel Booking Dataset - Insights
- Problem Statement
- Data Cleaning
- Exploratory Data Analysis
- Conclusion

Introduction to Exploratory Data Analysis

What is EDA?

The process of studying and summarizing the data in order to understand it as much as possible to make informed decisions.

It can involve any methods which allow us to get some underlying information about the dataset such as count plots, distribution plots, relationship between two or more variables.

It is very necessary to have a very good idea about the dataset before making any assumption thus making EDA an indispensable part of data science.

Introduction to Exploratory Data Analysis

Importance of EDA:

- Exploratory analysis alone can sometimes contribute to making important decisions.
- Helps understand the data in order to choose an ML model.
- Detecting errors and anomalies in data.
- Selecting important features that help make better predictions.
- Reducing bias while fitting a model.
- Discovering interesting relationships between variables.

Introduction to Exploratory Data Analysis

Even data cleaning (removing null values), converting the data to a readable form etc. requires some basic exploration.

EDA can be an endless exploration of the data leading to potentially interesting observations.

Hotel Booking Dataset

This dataset gives the booking information for a city hotel and a resort hotel. It includes detailed data on time of arrival, number of guests, country of guests, cancellations, rates, distribution channels etc.

These information can be used to find data trends which could be very useful for decision making both for the customers as well as the hotels. This helps the hotels to understand where they have to focus and what are the expectations of their customers.

Hotel Booking Data

hotel: city hotel or resort hotel

is_canceled: whether the booking was cancelled

lead_time: number of days between entering booking info and arrival

arrival_date_year: year of arrival

arrival_date_month: month of arrival

arrival_date_week_number: week number of arrival date

arrival_date_day_of_month: day of arrival

stays_in_weekend_nights: number of weekend nights stayed

stays_in_week_nights: number of week nights stayed

adults: number of adult guests

children: number of children

babies: number of babies guests

country: country of origin of guests

Hotel Booking Data

market_segment : Online TA, Offline TA/TO, Groups, Direct, Corporate etc.
distribution_channel: TA/TO, Direct, Corporate, GDS etc.
is_repeated_guest: if the guest has booked earlier
reserved_room_type: code of reserved room type (actual type anonymous)
assigned_room_type: code of room type assigned
booking_changes: number of changes made in the booking
deposit_type: No Deposit; Non Refund - deposit of total stay cost made; Refundable – a deposit was made with a value under the total cost of stay.
agent: ID of agency that made the booking
company: ID of company responsible for the booking
day_in_waiting_list: Number of days the booking took before confirmation.
adr: average daily rate

Hotel Booking Data

Importance of this Analysis

- It gives us many important information like the popular room types, cancellations, booking activity across hotels etc.
- The hotel businesses can use this to provide better services and meet the popular demands, thus benefiting both the hotels and customers.
- Customers can use the information to get the best possible deals and the best time to book.
- New businesses can make more informed decisions on what kinds of services to provide.
- Hotels can make certain changes based on data trends. For example, they could increase their meal options by considering the cuisines of common countries where guests come from.

Problem Statement

Our objective is to observe the underlying patterns and understand what factors affect the bookings.

1. Which hotel has the most bookings?
2. How many bookings were cancelled?
3. Which is the busiest month across the year 2016?
4. Which is the most sought after room type?
5. From which countries do most guests arrive?
6. Which is the largest market segment and distribution channels?
7. How many nights do the guests stay?
8. How many booking changes did customers make?
9. How many of the customers are repeated guests?
10. Which hotel is more expensive?

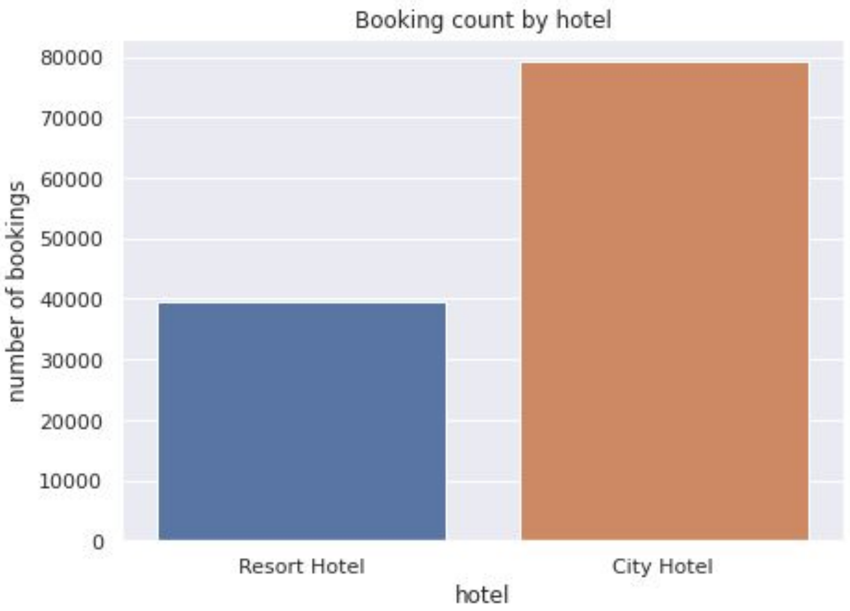
Data Cleaning

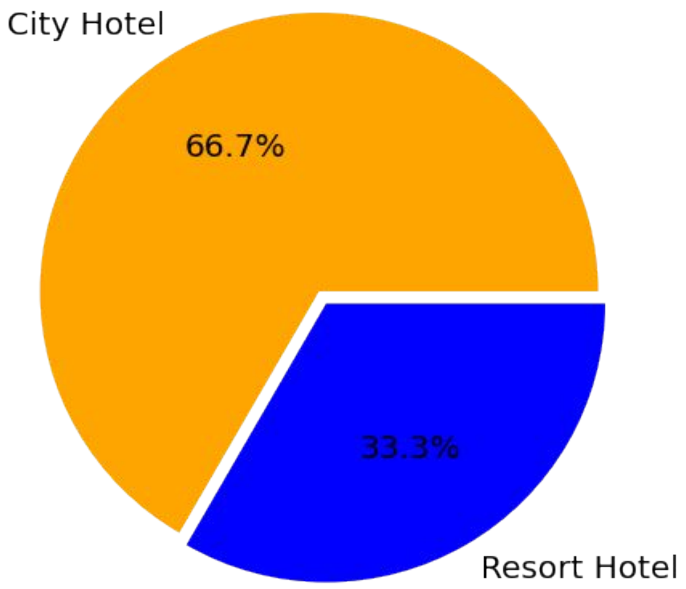
Preparing the data for analysis:

- Removing null values
- info and describe methods are used to get some general idea about the data
- Removing some erroneous data (zero total guests)

Exploratory Data Analysis

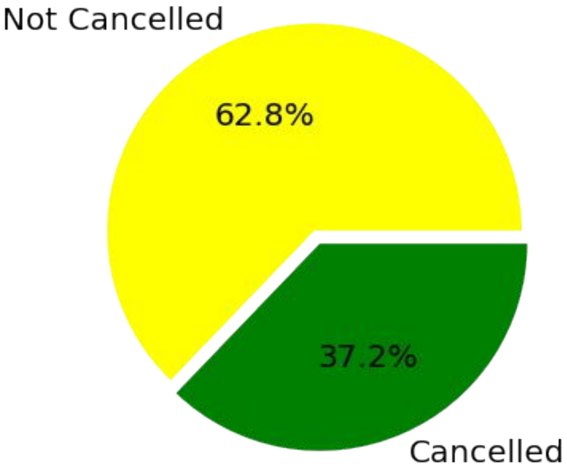
Distribution of bookings across the two hotels.

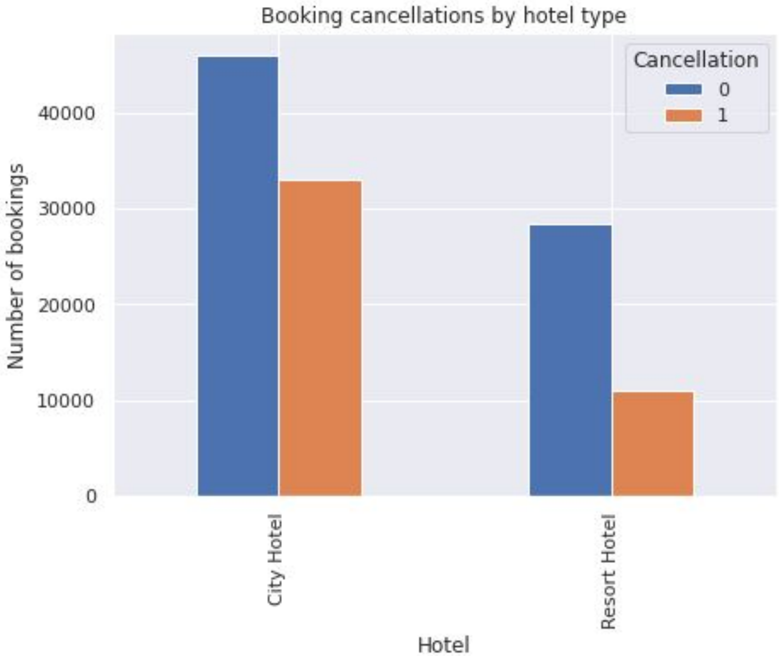




Exploratory Data Analysis

Percentage of booking cancellations (overall and for individual hotels)





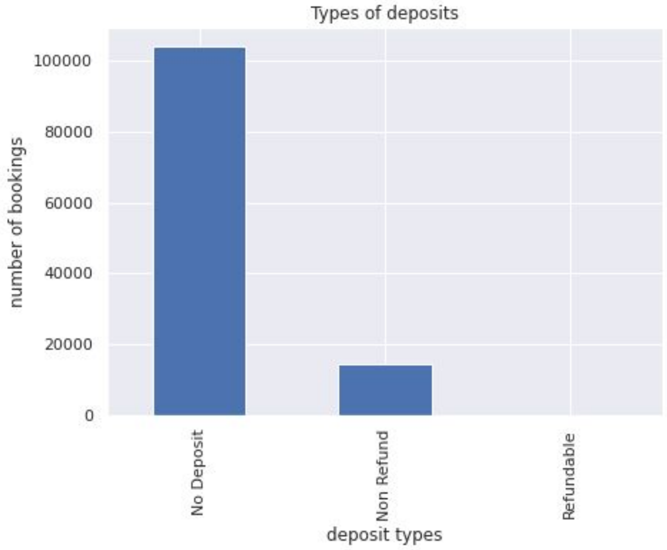
City Hotel	41.77
Resort Hotel	27.98

Exploratory Data Analysis

Factors that can potentially reduce cancellations:

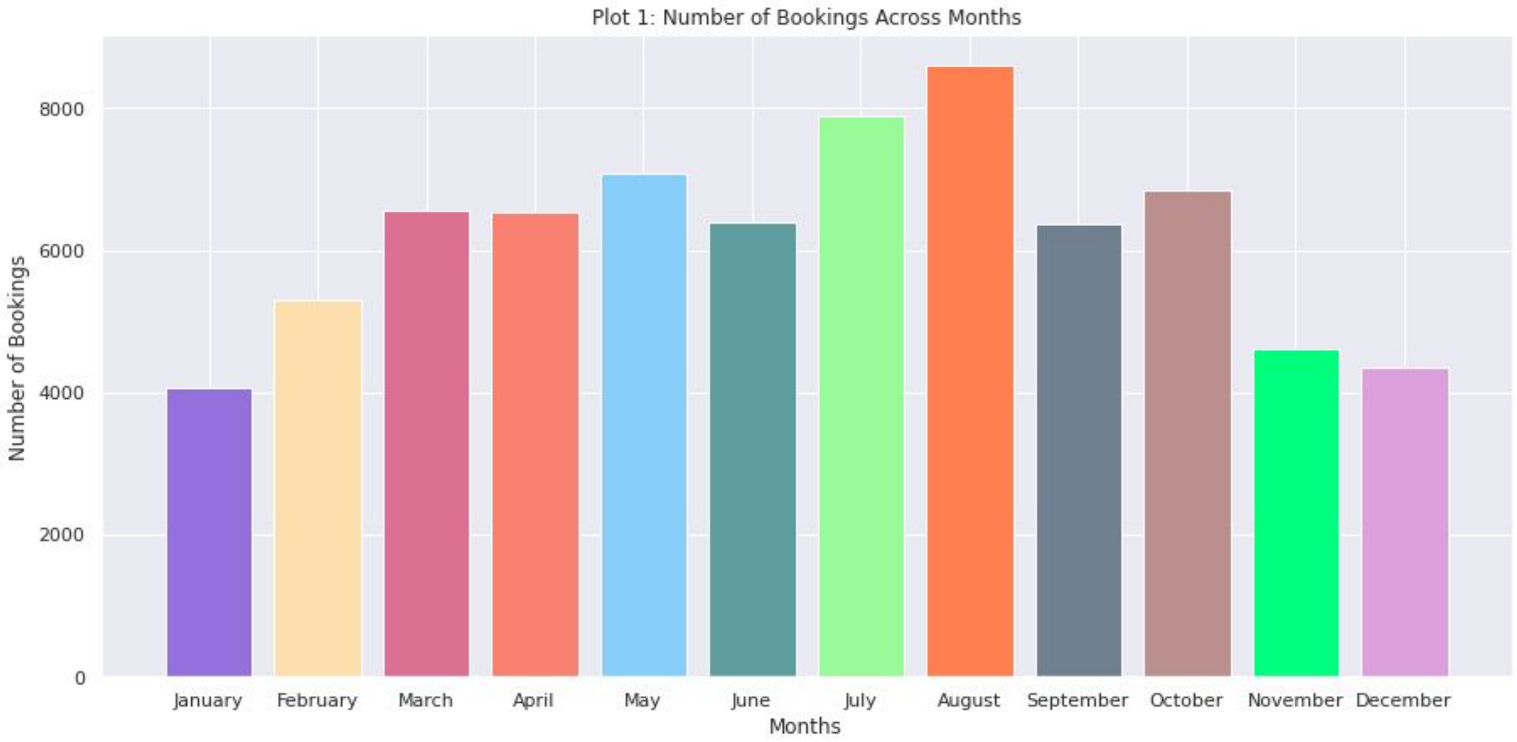
- Deposit requirement while booking
- Reducing waiting list time

Cancellation	**Mean (waiting list days)**
Not cancelled	1.597
Cancelled	3.571



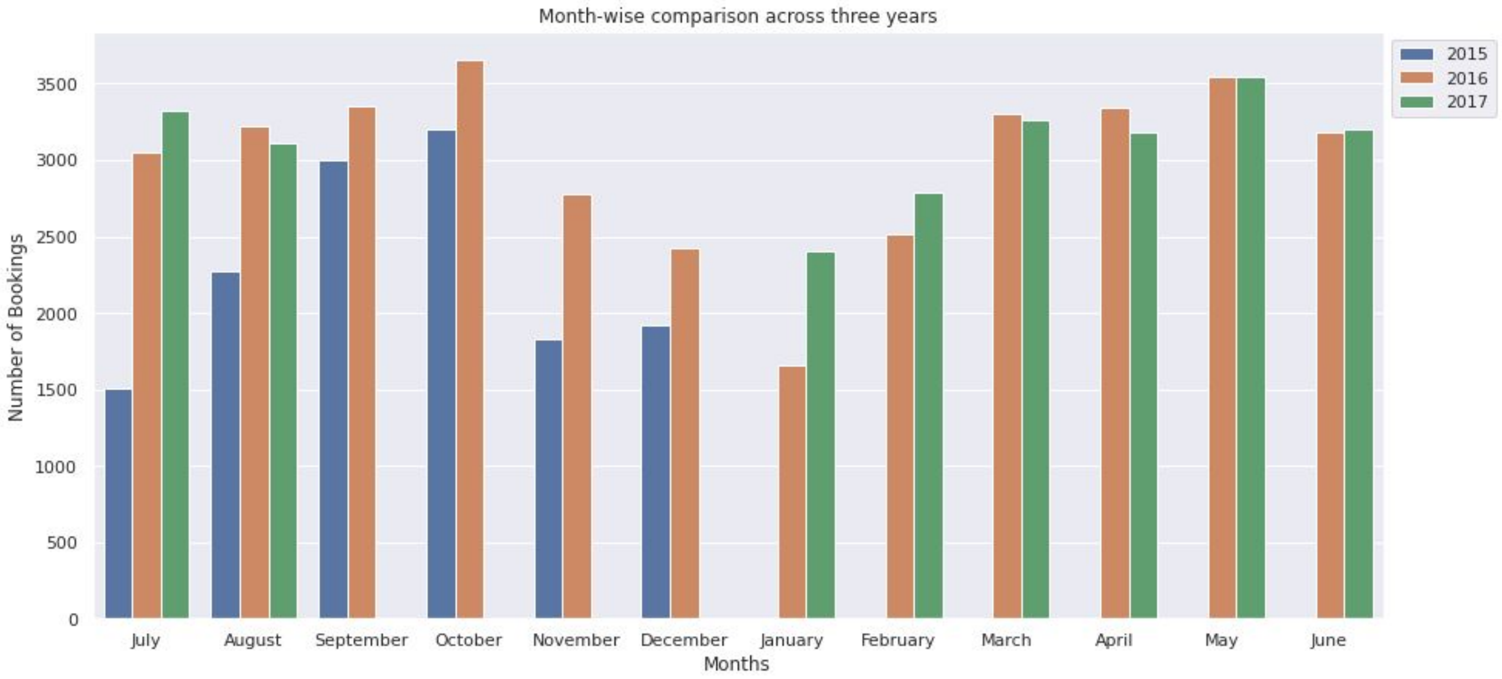
Exploratory Data Analysis

Month-wise distribution of the available data



Exploratory Data Analysis

Month-wise comparison across all the years



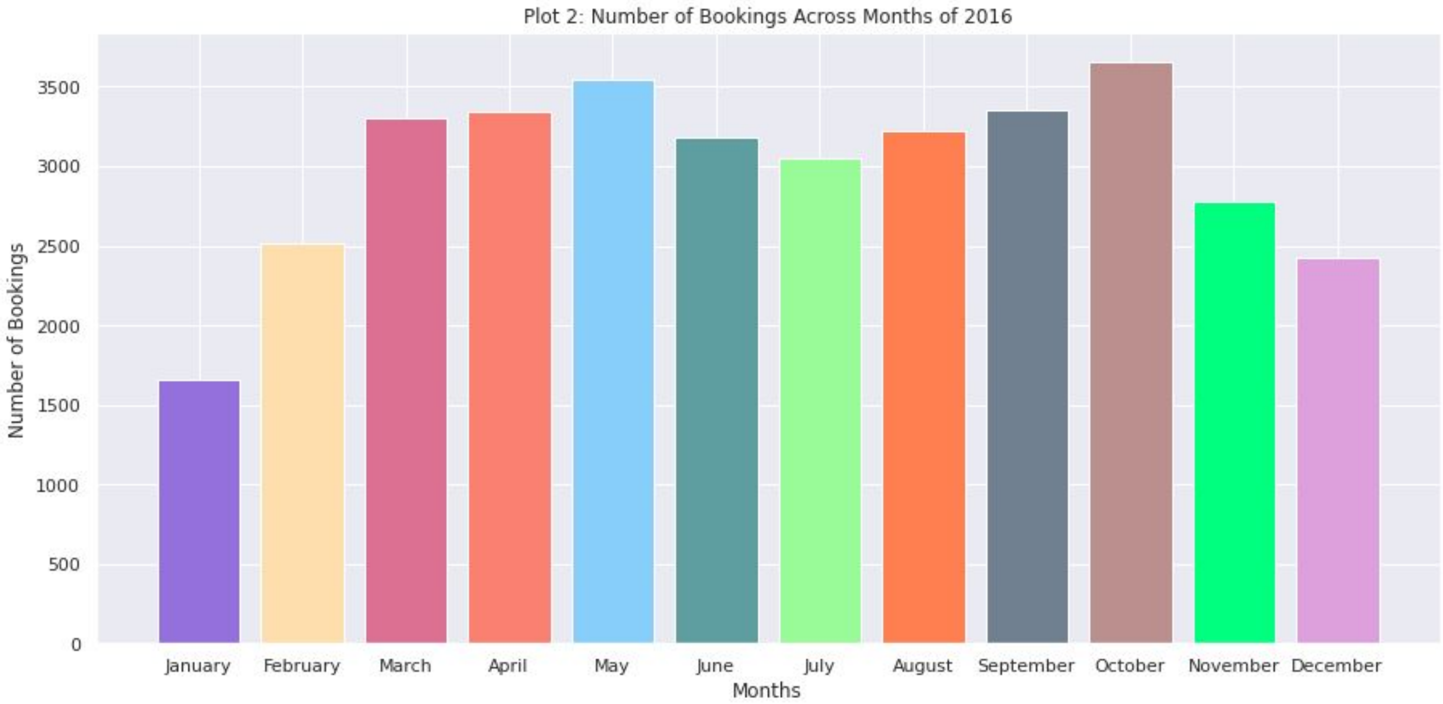
Exploratory Data Analysis

Month-wise comparison across all the years

- The dataset is incomplete!
- We do not have month-wise information for all the three years
- Dataset starts from July 2015 to August 2017
- We cannot make proper year-wise comparison
- We cannot make month-wise comparison across three years

Exploratory Data Analysis

Month-wise distribution across 2016



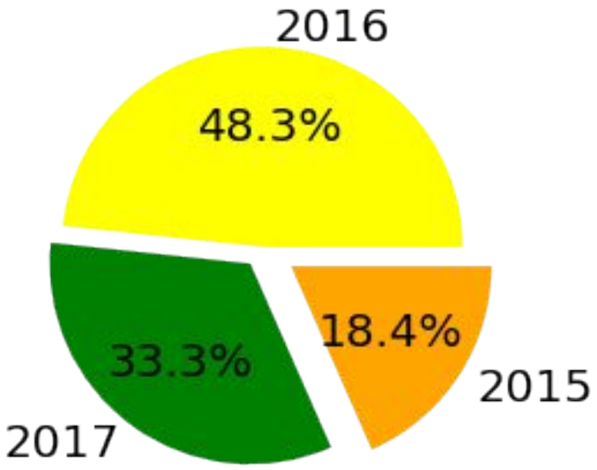
Exploratory Data Analysis

Month-wise distribution across 2016

- The number of bookings are more consistent
- October followed by May has the highest count
- Beginning and end of year has relatively higher count

Exploratory Data Analysis

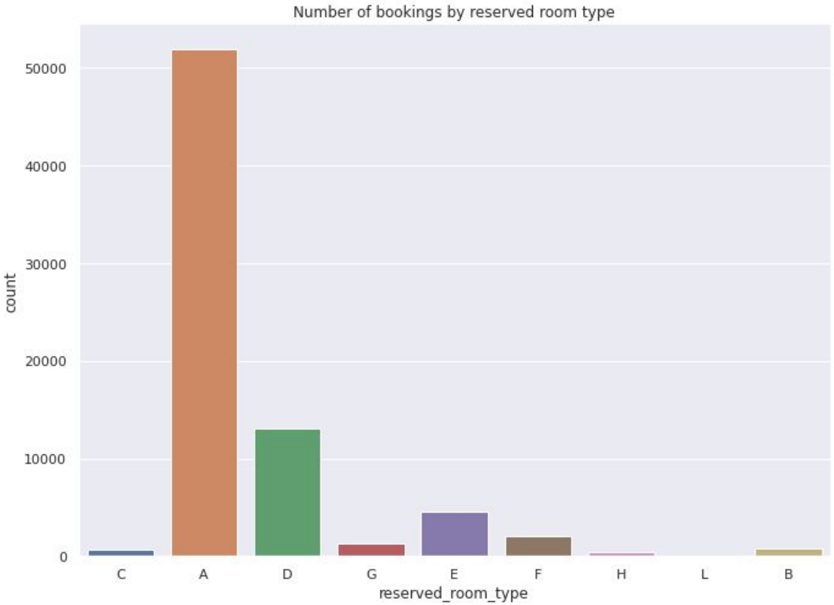
Year-wise distribution of the available data

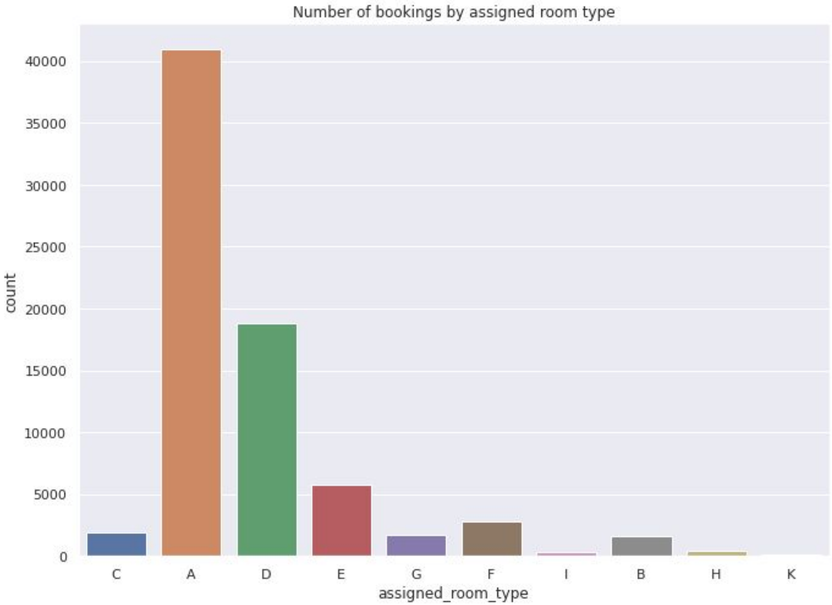


We cannot conclude that 2016 is the busiest year as the dataset do not have the information of every months for the three years

Exploratory Data Analysis

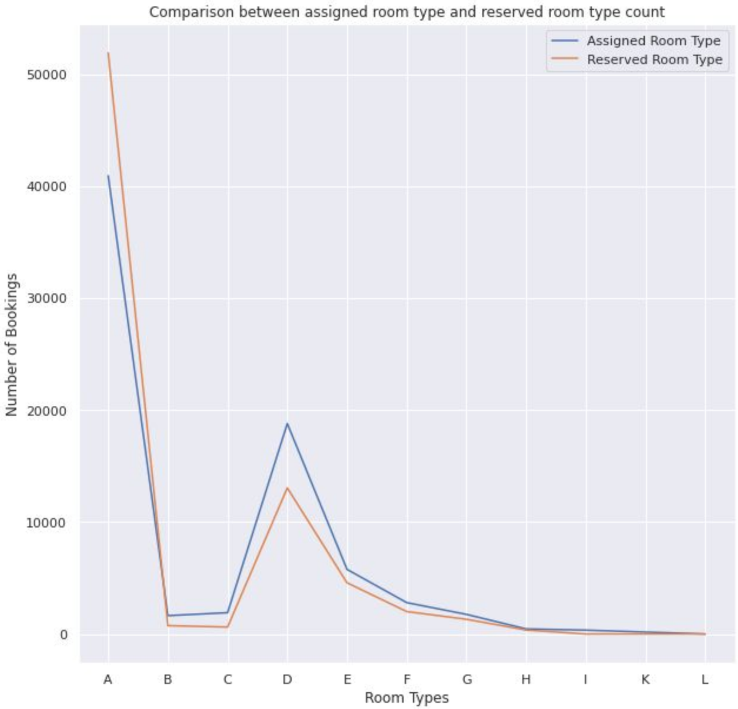
Number of bookings across room types





Exploratory Data Analysis

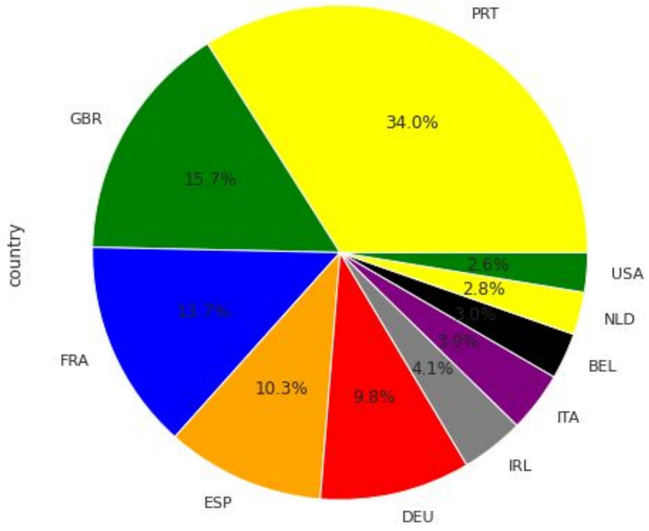
Number of bookings across room types

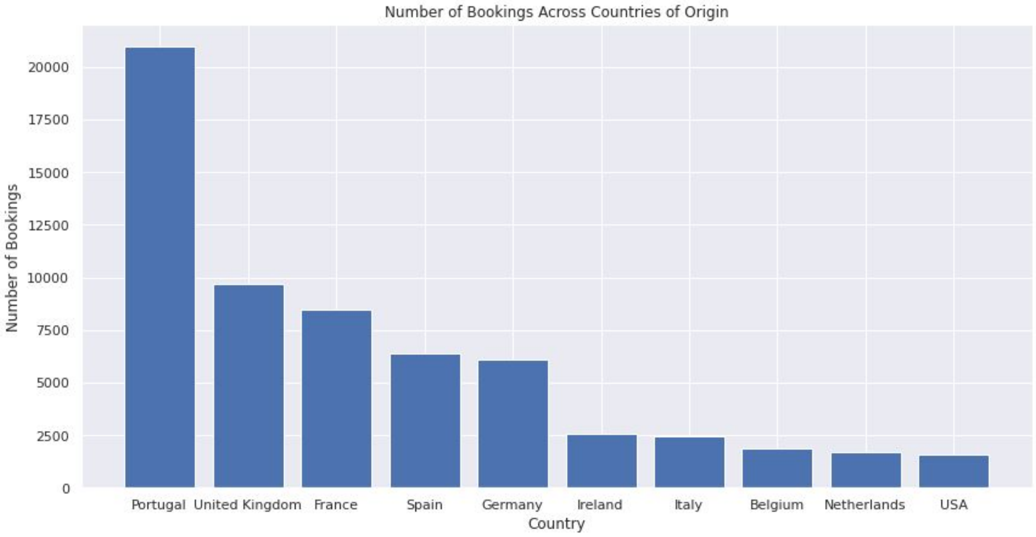


- Type A is the most sought after
- The demand for type A is not met
- It is satisfied by other room types

Exploratory Data Analysis

Country of origin of guests





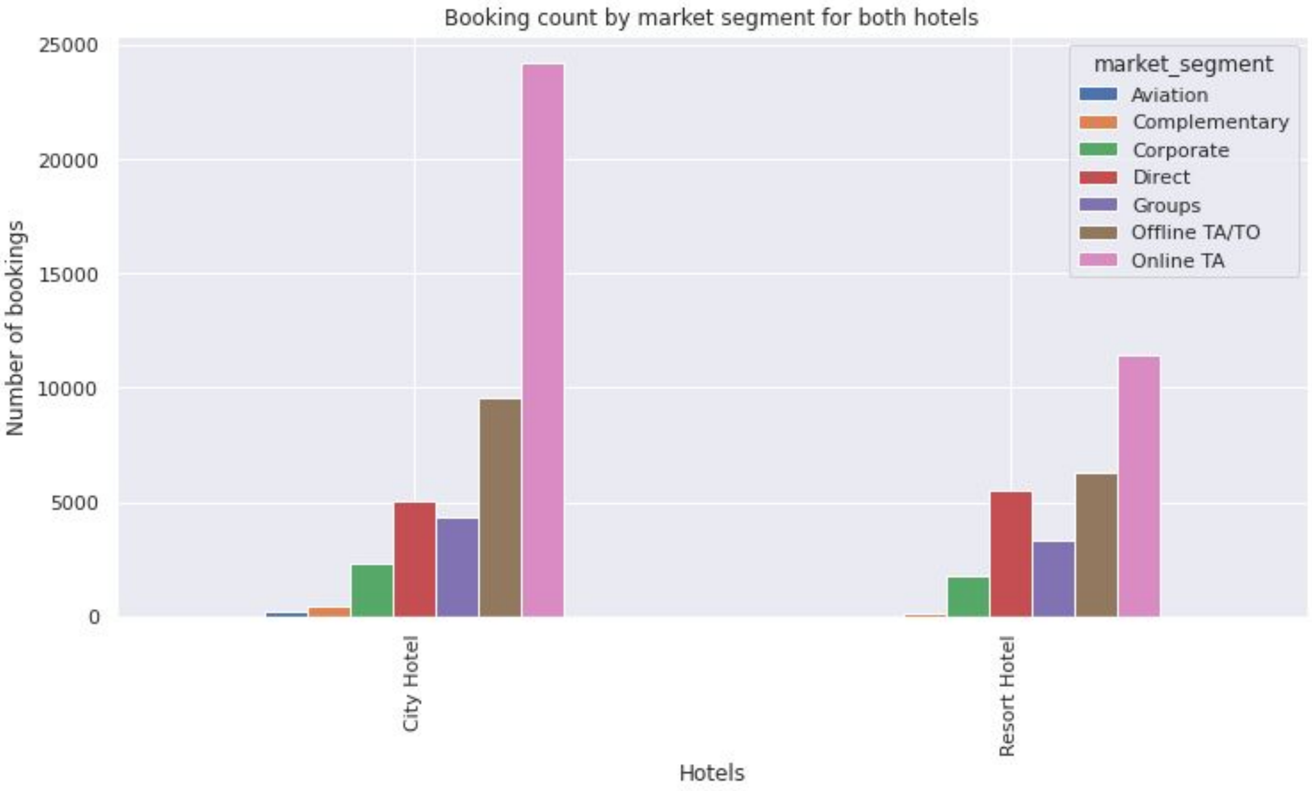
Exploratory Data Analysis

Country of origin of guests

- Portugal is the most common country where guests arrive from
- Followed by UK and France.
- Most guests are from European countries.

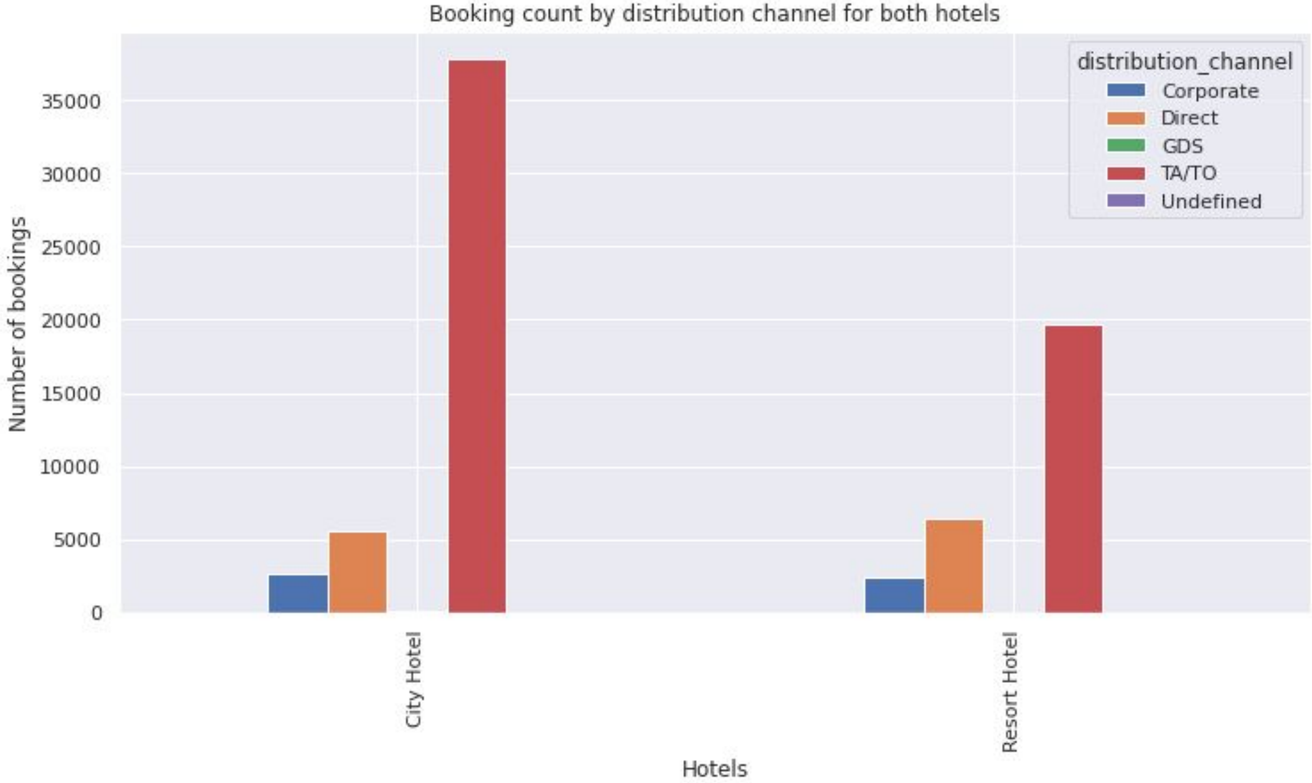
Exploratory Data Analysis

Booking count across market segment



Exploratory Data Analysis

Booking count across distribution channel



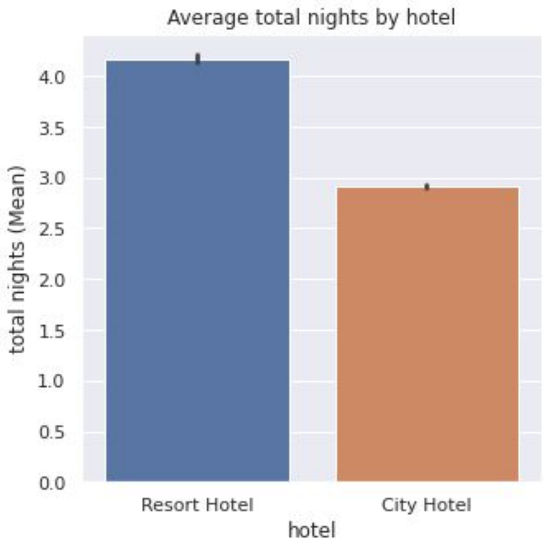
Exploratory Data Analysis

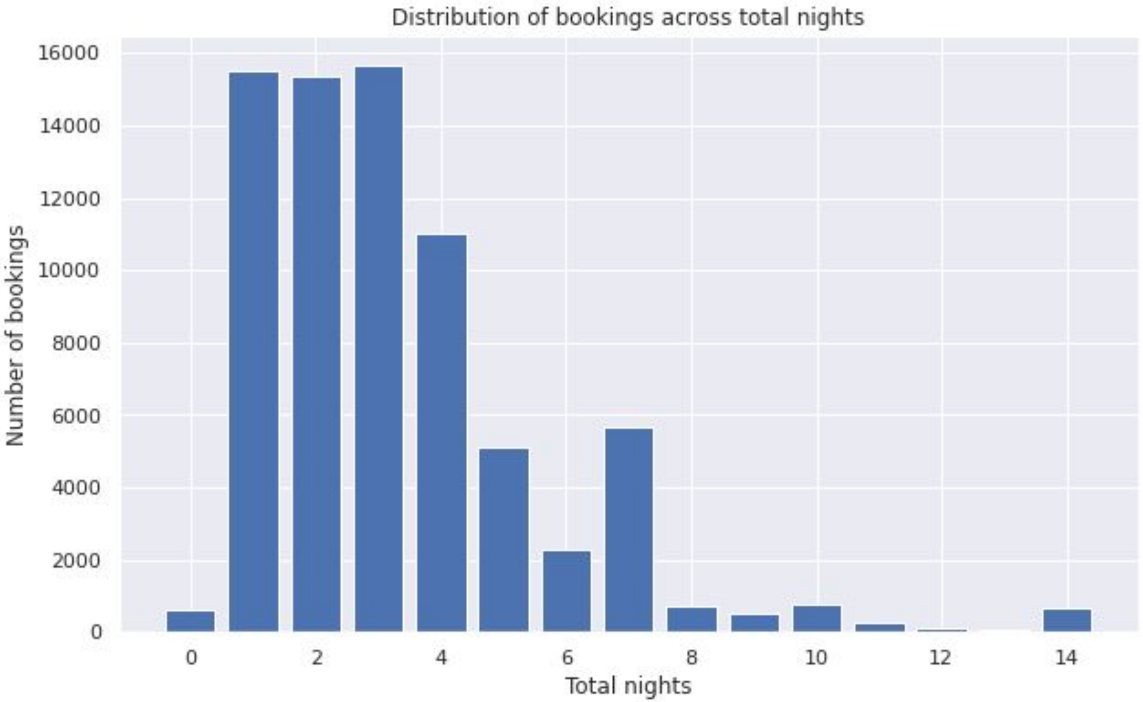
Market segment and distribution channel

TA/TO from market segment	51453
TA/TO from distribution	57507
Direct booking from market segment	10504
Direct booking from distribution channel	11908
Corporate booking from market segment	4121
Corporate booking from distribution channel	5018

Exploratory Data Analysis

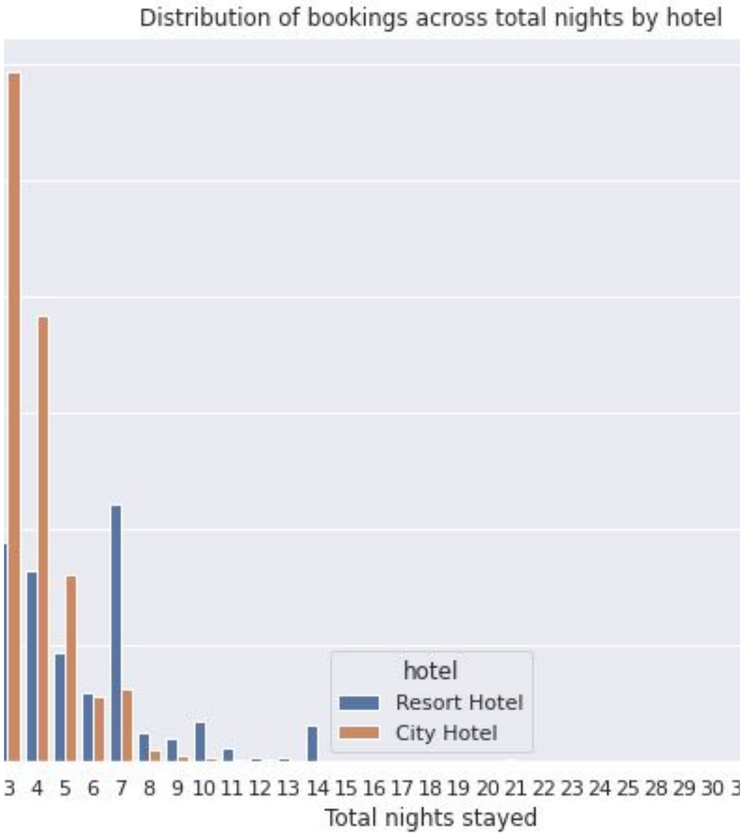
Total nights stayed





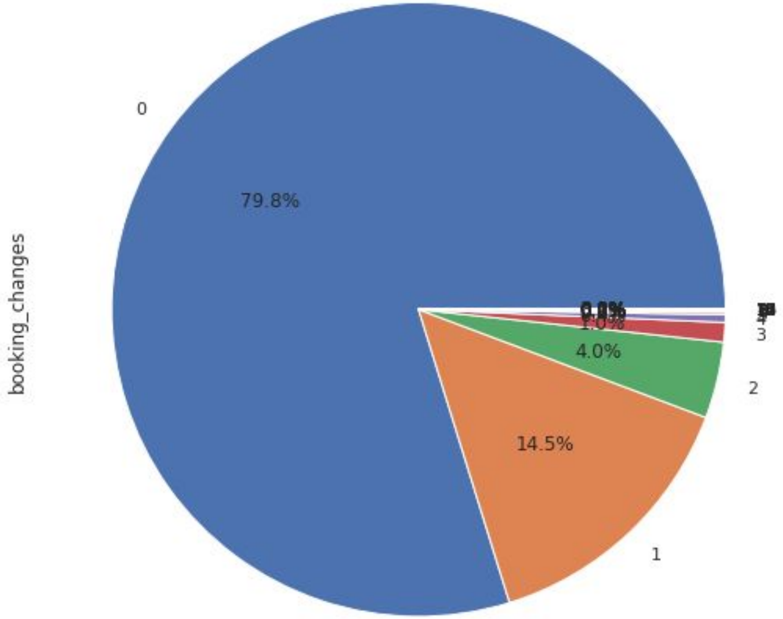
Exploratory Data Analysis

Total nights stayed



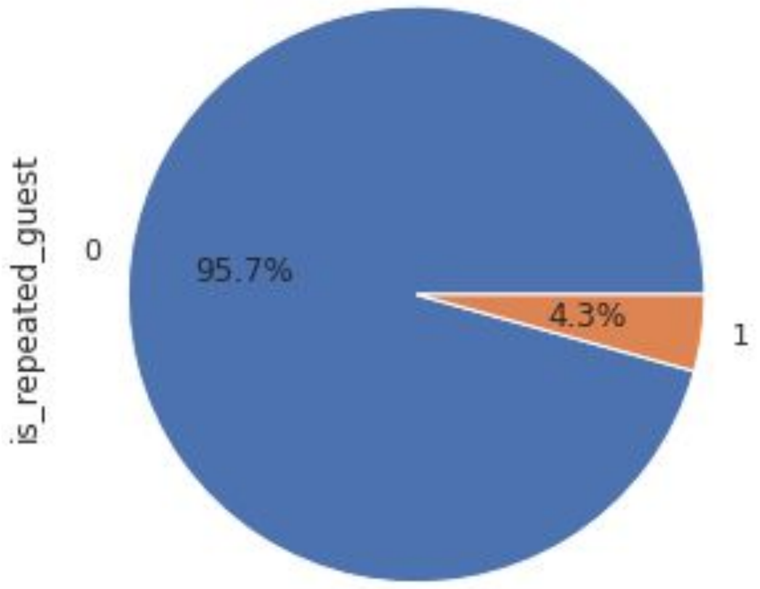
Exploratory Data Analysis

Booking changes



Exploratory Data Analysis

Repeated guests



Exploratory Data Analysis

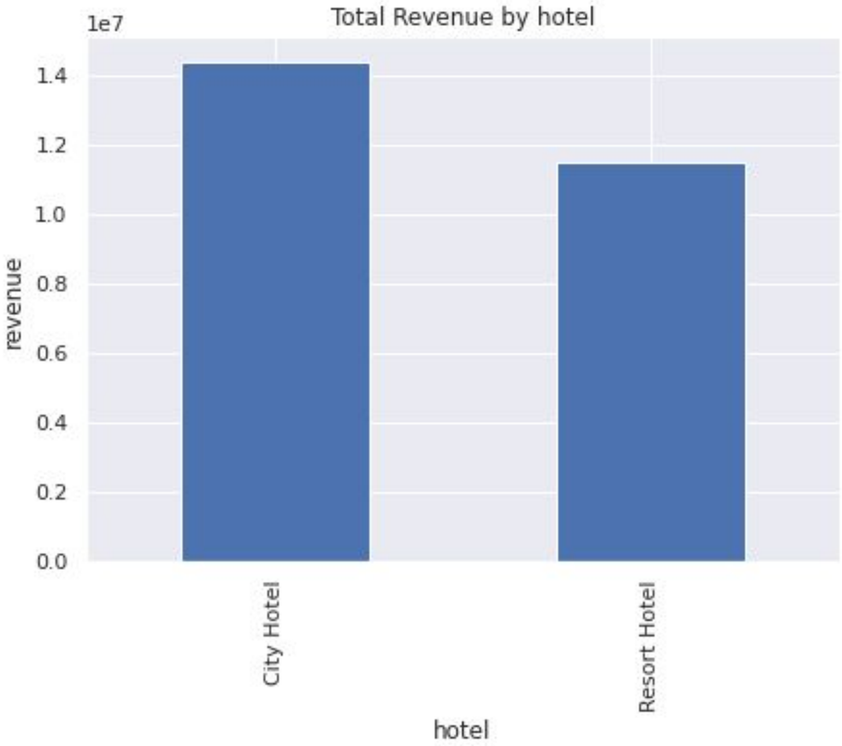
ADR and total rate

Average ADR

City Hotel	106.04
Resort Hotel	91.27

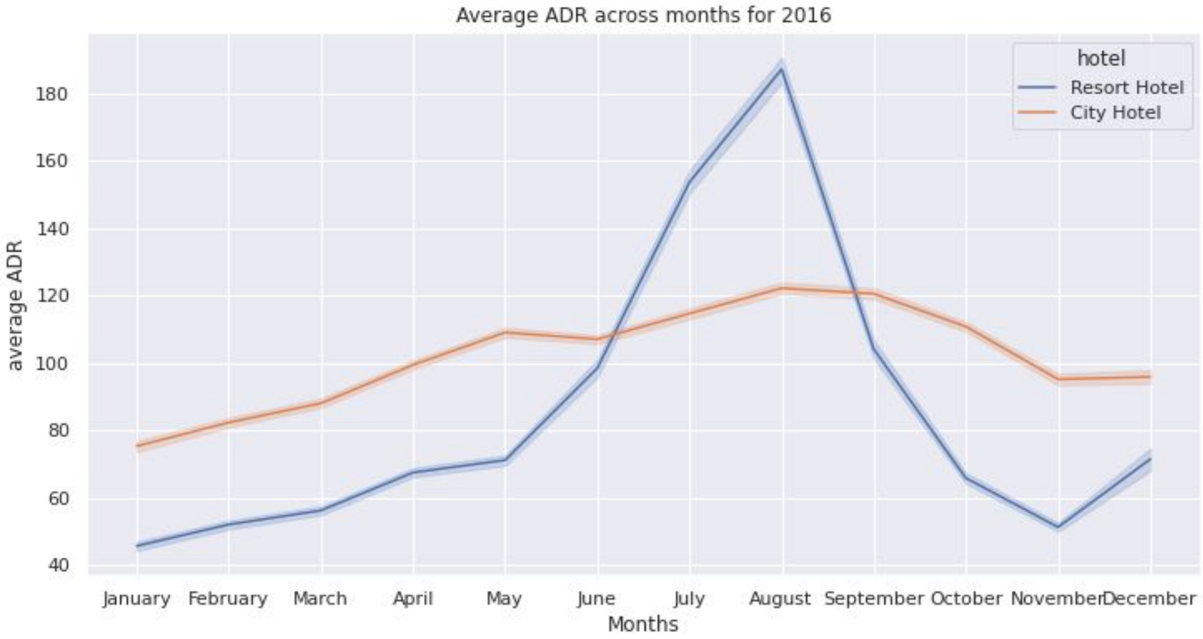
Average Total Rate

City Hotel	312.15
Resort Hotel	403.92



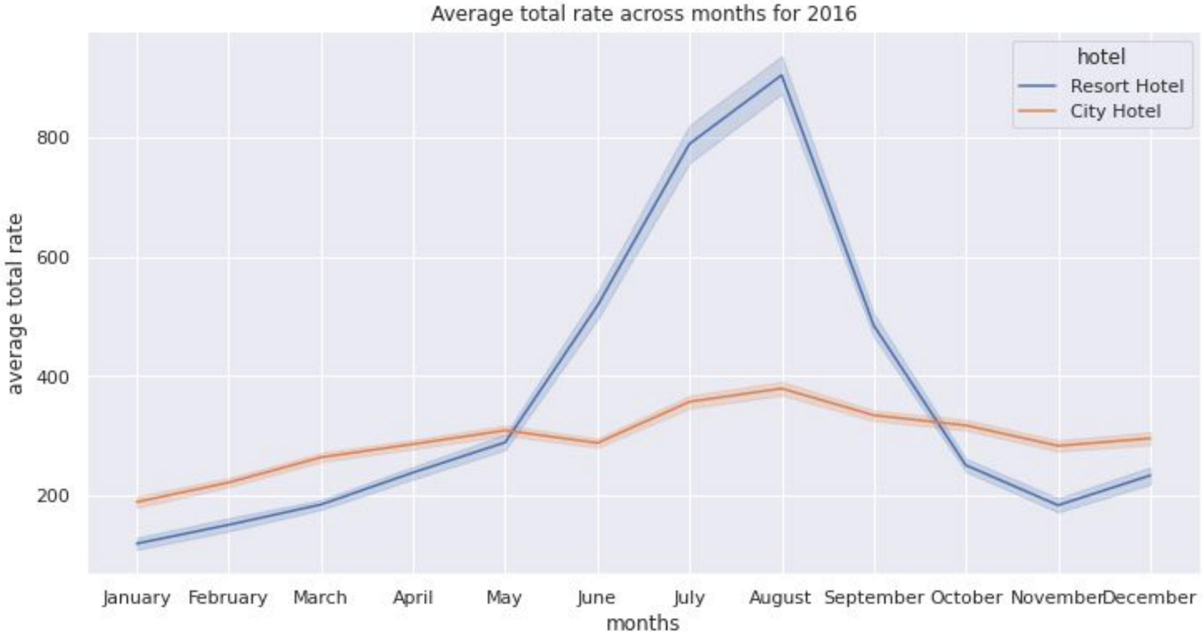
Exploratory Data Analysis

Month-wise comparison of ADR and Total Rate for 2016



Exploratory Data Analysis

Month-wise comparison of ADR and Total Rate for 2016



Conclusion

- City hotel has more bookings than resort hotel and it also makes more revenue in total.
- There are a fairly high percentage of cancellations (approx. 37 %).
- Type A is the most in-demand room type and the demand is not fully satisfied.
- A large portion of the guests are from European countries (most common one being Portugal).
- Travel agencies (especially online TAs) are the common intermediary.
- 1-3 days is the most common period of stay. Resort hotel has more guests that stay longer.
- Most customers make few or no changes in booking.
- There are very few repeated guests.
- City hotel is slightly more expensive generally. But the resort hotel witnesses a huge increase in price in July and August.
- Month-wise or year-wise distributions cannot be accurately made as the dataset does not have information of every month for 2015 and 2017.

THANK YOU