# Capstone Project

## EDA on Hotel Booking Analysis


# TEAM MEMBERS

- Zunaid
- Sukesh Shetty
- Bhushan
- Biru Solankar


Introduction to Exploratory Data Analysis

Problem *statement*


Importing Python libraries & Usage

Data Preparation & Data Wrangling

Data visualization using Python & Finding solutions


Some important questions


Results & Conclusions

# Introduction to Exploratory Data Analysis(EDA)

Exploratory Data Analysis (EDA) is the process of visualizing and analyzing data to extract insights from it. In other words, EDA is the process of summarizing important characteristics of data in order to gain better understanding of the dataset.

EDA is very essential because it is a good practice to first understand the problem statement and the various relationships between the data features before getting your hands dirty.

Understanding EDA To understand the steps involved in EDA, we will use Python as the programming language and Colab Notebooks because it's open-source, and not only it's an excellent IDE but also very good for visualization and presentation.


**Technically, The primary motive of EDA is to**

- Examine the data distribution
- Handling missing values of the dataset(a most common issue with every dataset)
- Handling the outliers
- Removing duplicate data
- Encoding the categorical variables
- Normalizing and Scaling

# Problem Statement

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal length of stay in order to get the best daily rate? What if you wanted to predict whether or not a hotel was likely to receive a disproportionately high number of special requests? This hotel booking dataset can help us explore those questions!

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.

- This dataset has around 119390 observations(rows) and 32 columns and it is mix with numeric and categorical values.

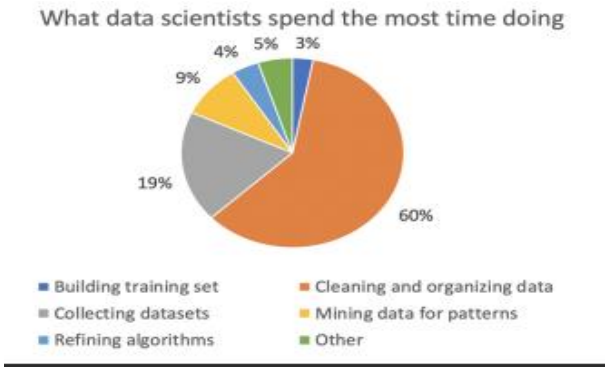
- We will find out the distribution of every important column to answer the problem statement.

# Importing Python library and Usage

- A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer.
- • Python have more than 137,000 libraries, Python can be used to create applications and models in a variety of fields, for instance, machine learning, data science, data visualization, image and data manipulation, and many more…
- • In this Project we are mainly Using some libraries..(i.e., Numpy, Pandas, matplotlib, seaborn,etc..)
- • After importing libraries we will come up with an imaginary story to explain the data
- • Data visualization also helps to tell stories by curating data into a form easier to understand, highlighting the trends and outliers.
- • A good visualization tells a story, removing the noise from data and highlighting the useful information.

# Data Preparation & Data Wrangling

- Data preprocessing is an important task. It is a data mining technique that transforms raw data into a more understandable, useful and efficient format.
- Data has a better idea. This idea will be cleared and understandable after performing data preprocessing. The data cleaning process detects and removes the errors and inconsistencies present in the data and improves its quality.
- Data quality problems occur due to misspellings during data entry, missing values or any other invalid data. Basically, “dirty” data is transformed into clean data. “Dirty” data does not produce the accurate and good results. Garbage data gives garbage out. So it becomes very important to handle this data. Professionals spend a lot of their time on this step.
- What to do to clean data?
- 1. Handle Missing Values
- 2. Handle Noise and Outliers
- 3. Remove Unwanted data



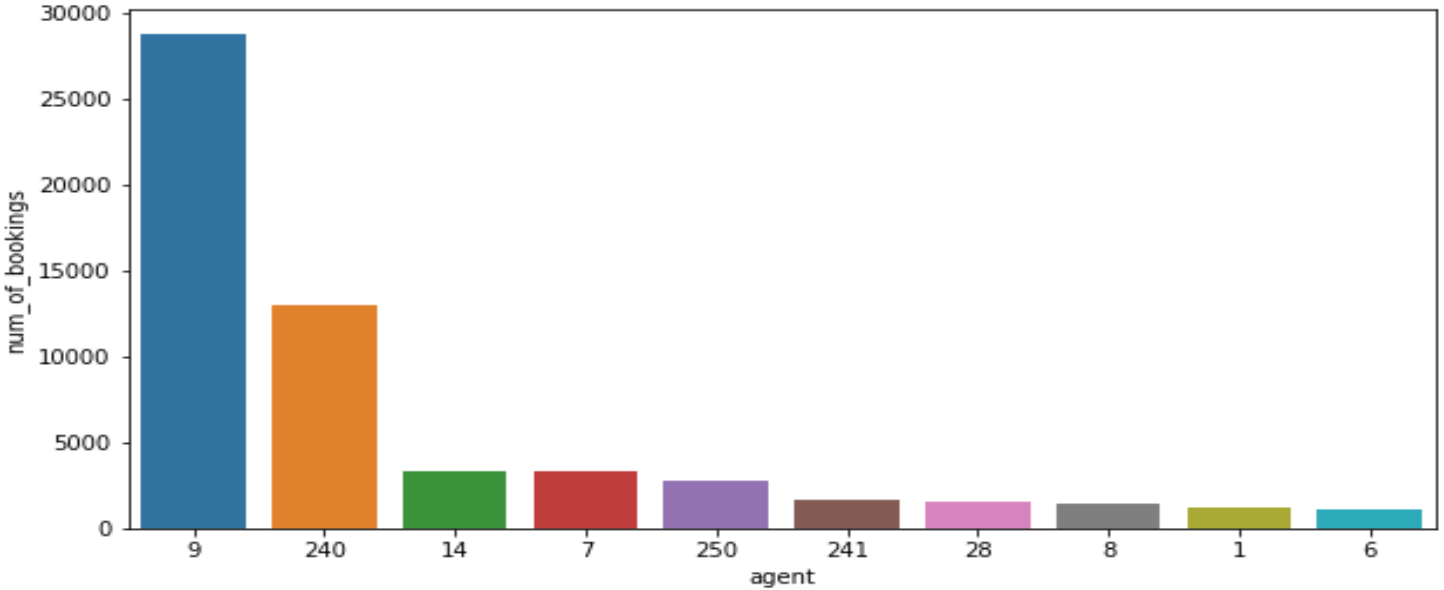
# Time for some QnA?

# Data Visualization divided into four parts.

- 1. Univariate Analysis
- 2. Hotel wise Analysis
- 3. Distribution wise Analysis
- 4. Time wise Analysis

# Univariate Analysis

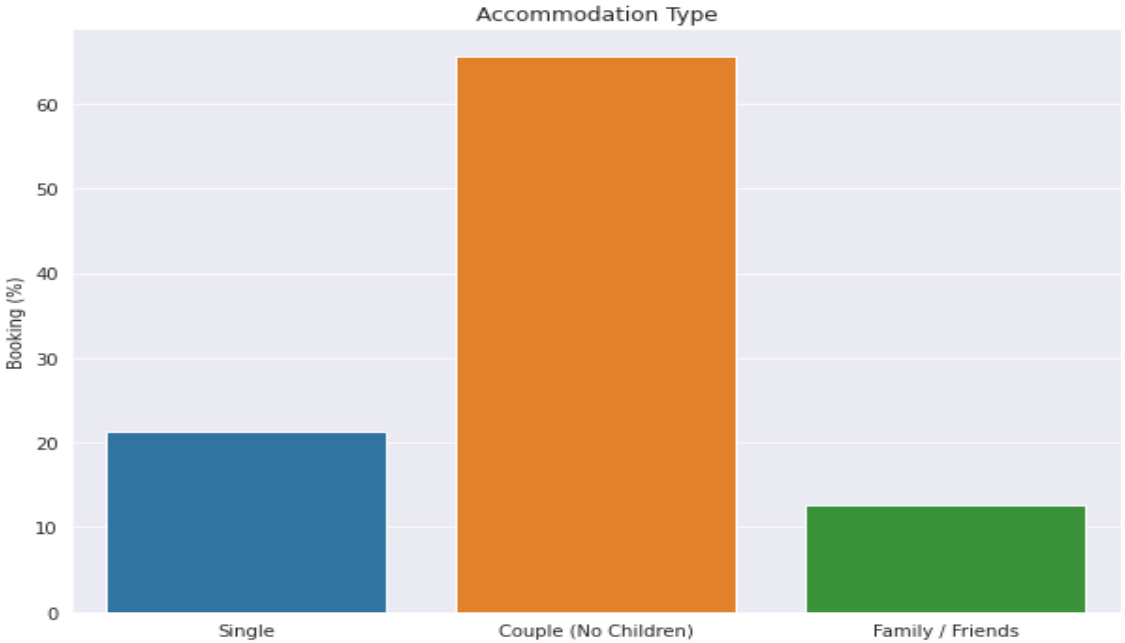
## Q1) Which agent makes most no. of bookings?




- Agent with id no. 9 has made most no. of bookings, around more than 25000.

# Univariate Analysis

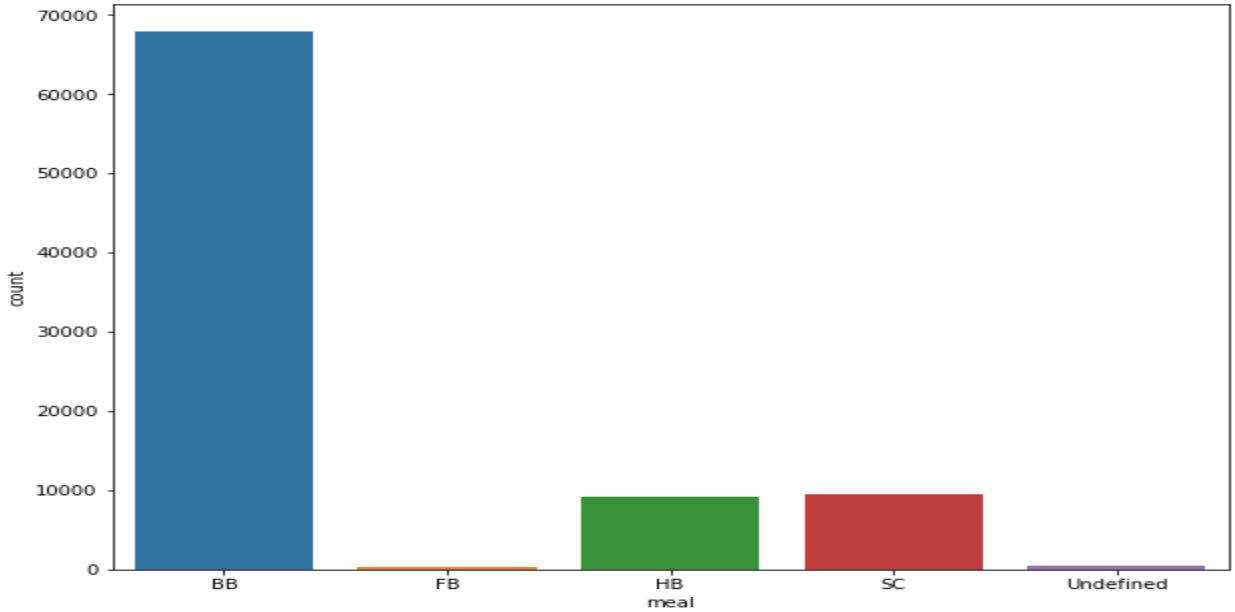
## 2. Which was the most booked accommodation type




- Accommodation type ‘Couple’ is most in demand. So, Hotels can make arrangement plans accordingly.

# Univariate Analysis

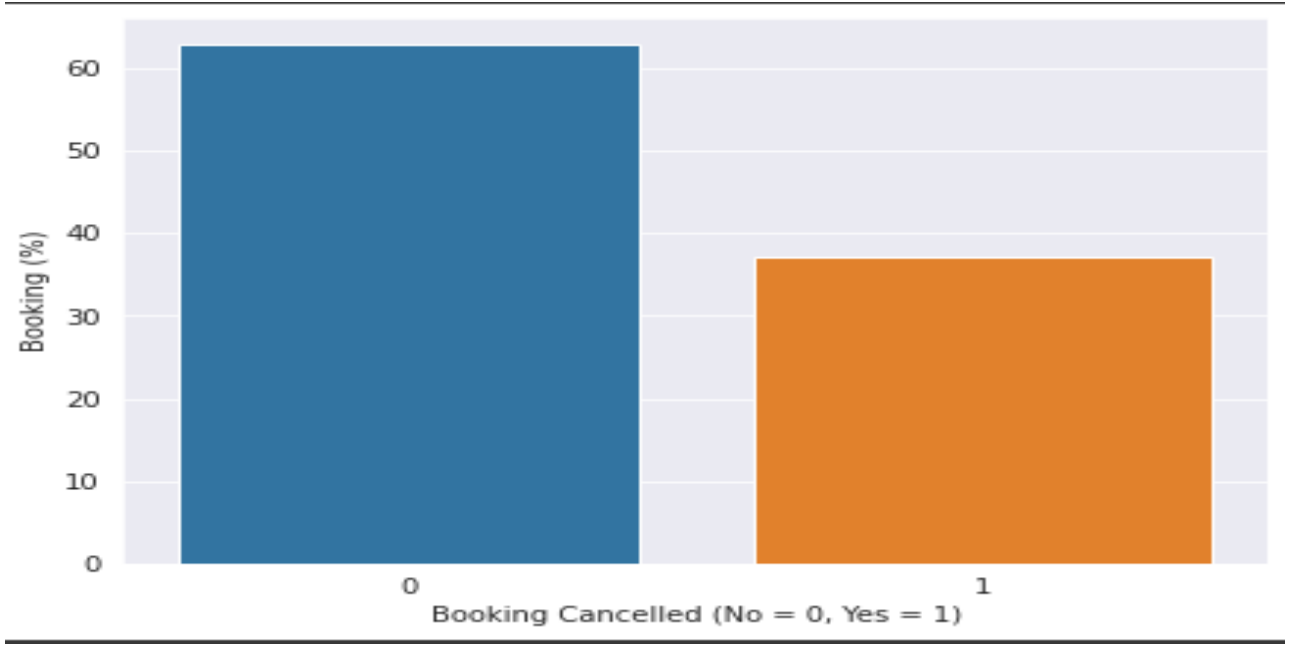
## Q3) Which meal type is most preferred meal of customers?




- Most preferred meal type is BB (Bed and breakfast).

# Hotel wise Analysis

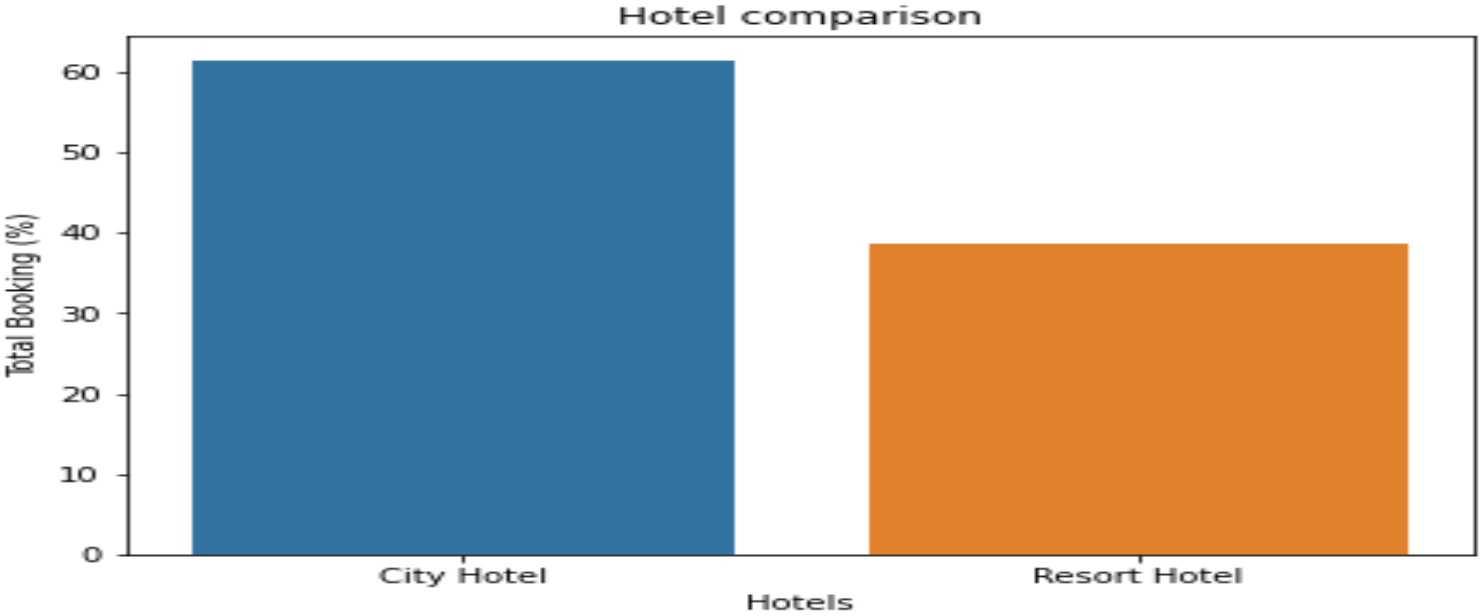
## 1. How Many Booking Were Cancelled?




We can see that more than 35% of the bookings are cancelled.

# Hotel wise Analysis

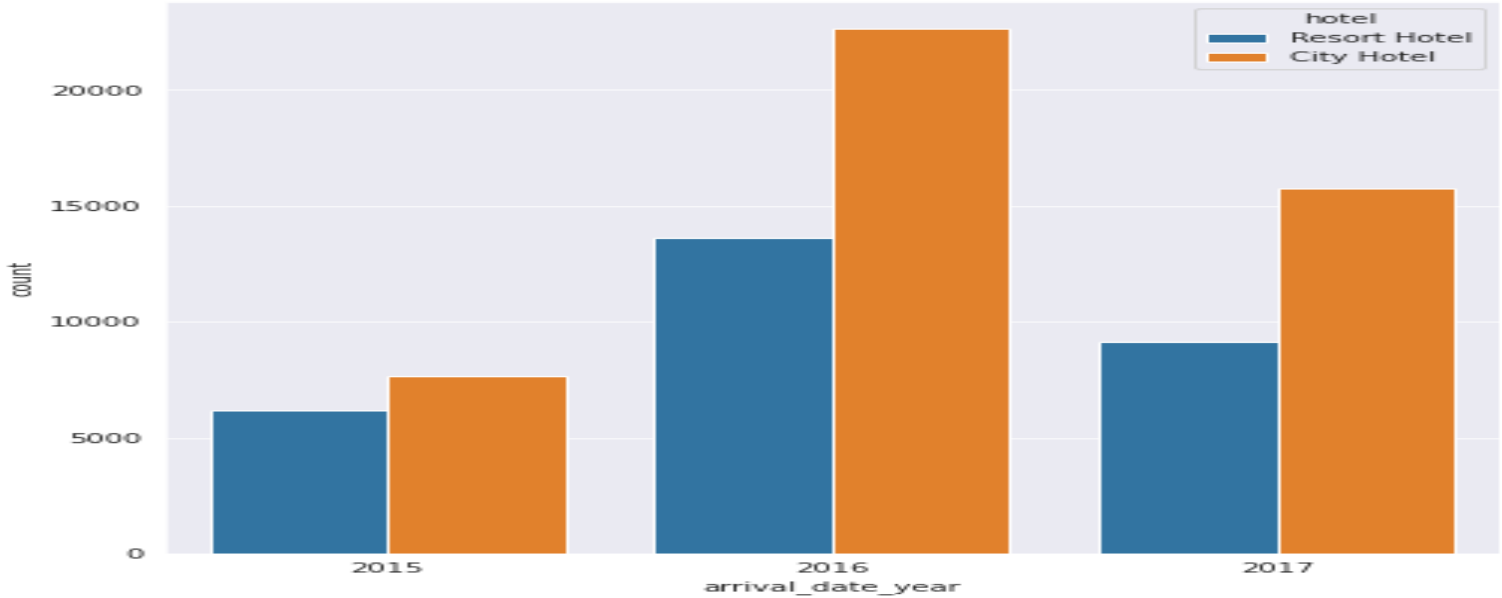
## 2. What is the booking ratio between Resort Hotel and City Hotel?




- It is shown that majority of the booking is done in city area which is 60% from this we can assume that majority of the booking is done for work related purposes for eg business related trips or to meet a client.

# Hotel wise Analysis

## 3. What is the percentage of booking for each year?




- We can clearly see that year 2016 is the most busiest year of both City and Resort hotel (nearly 50%) of hotel booking were done this year. And 2015 is the year when least booking were done.

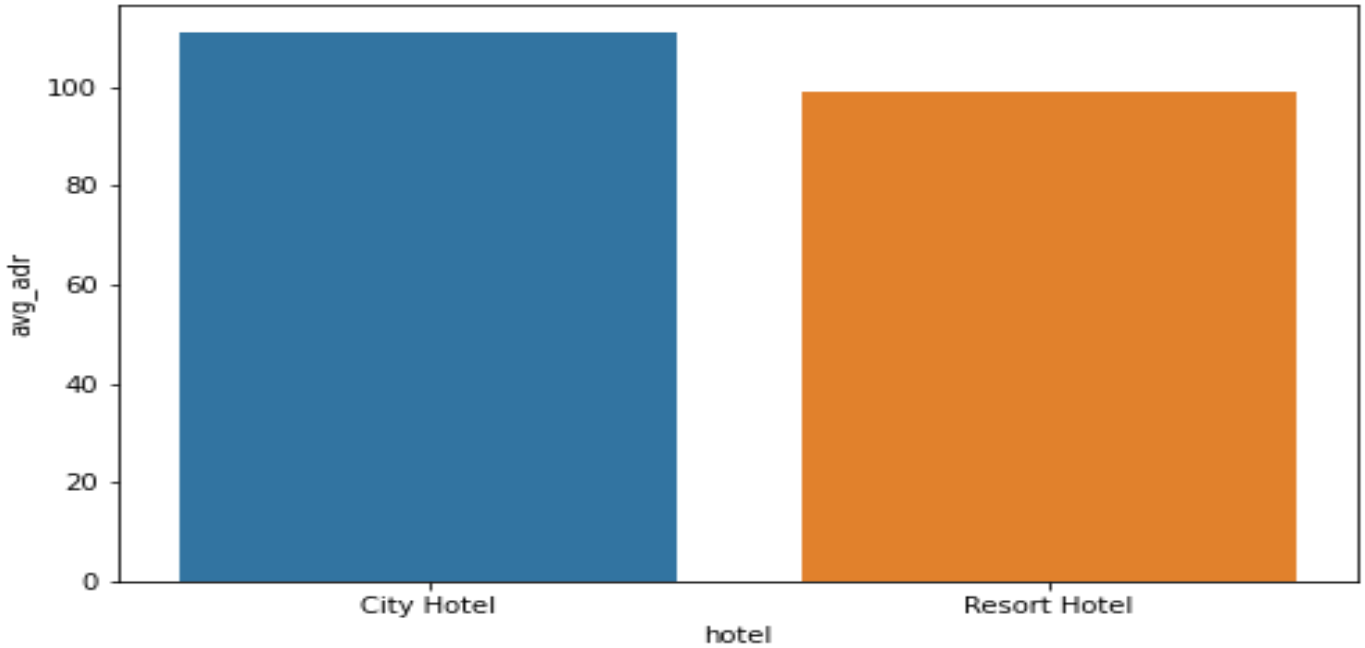
# Hotel wise Analysis

## 4. How Long People Stay in the hotel?


- Most people stay for one, two, or three. More than 60% of guests come under these three options.

# Hotel wise Analysis

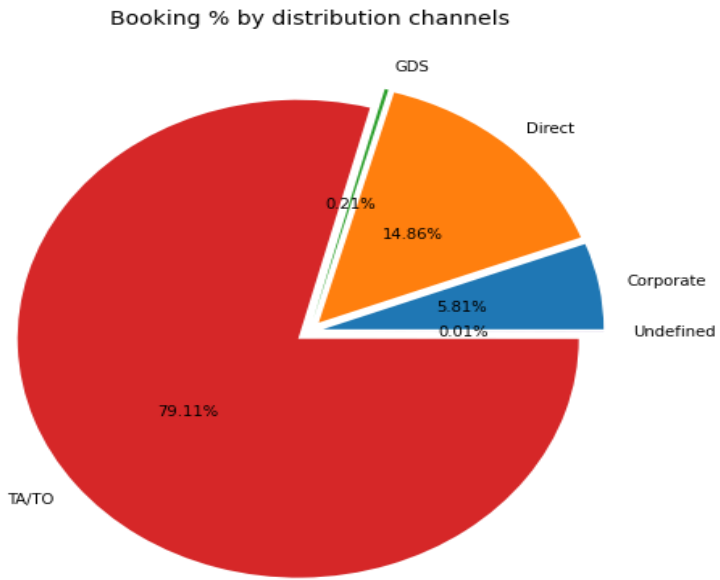
## 5. which hotel seems to make more revenue?




Average adr(avg. daily rate) of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue

# Distribution wise Analysis

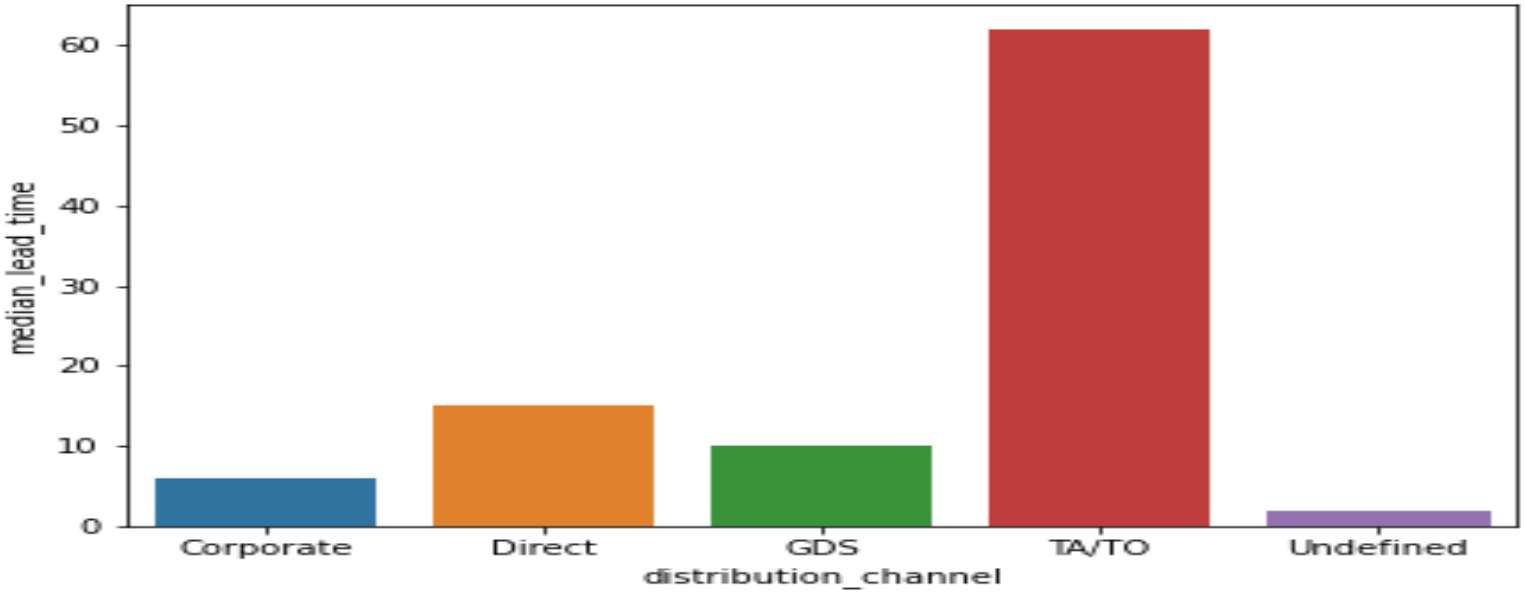
## Q1) Which is the most common channel for booking hotels?




- TA/TO is the most common channel for booking hotels, about 79.11 % booking were done by the TA/TO channels.

# Distribution wise Analysis

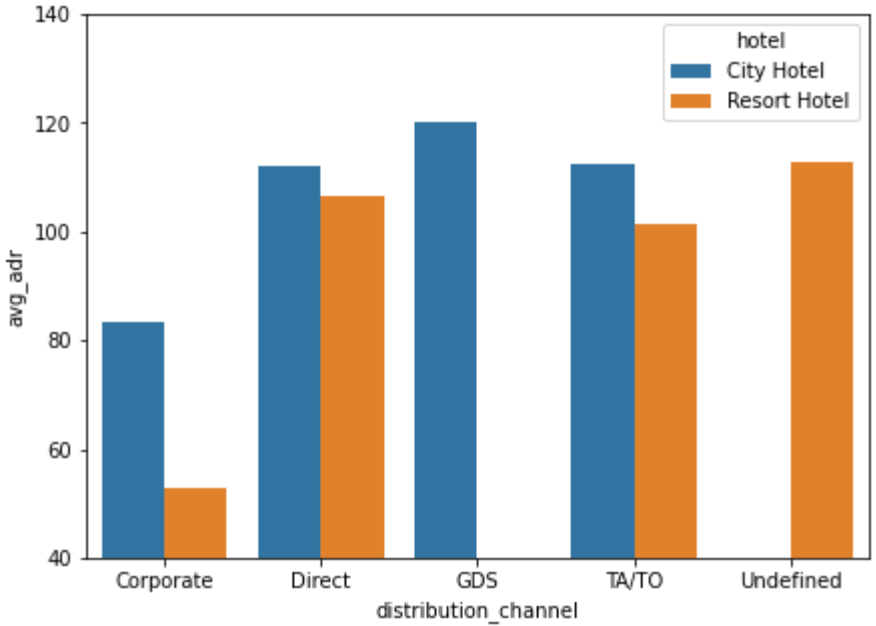
## Q2) Which channel is mostly used for early booking of hotels?




- TA/TO is mostly used for early booking of hotels.

# Distribution wise Analysis

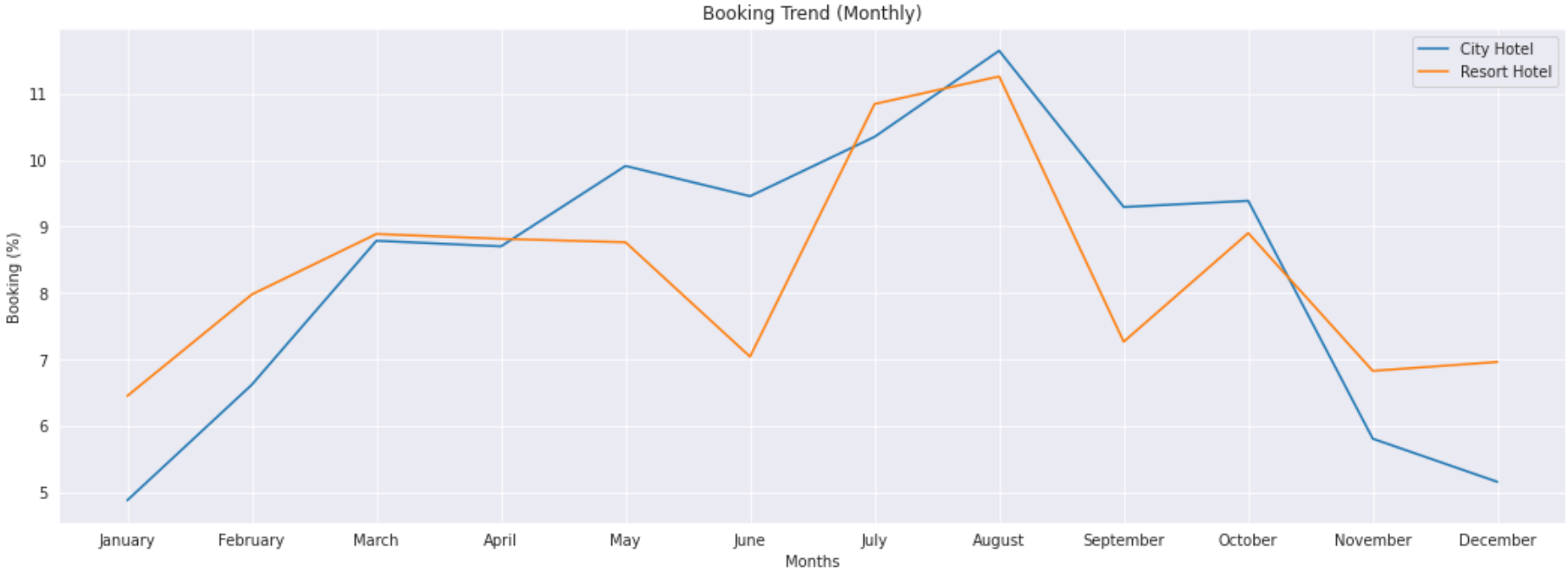
## Q3) Which distribution channel brings better revenue generating deals for hotels?




GDS channel brings higher revenue generating deals for City Hotel , in contrast to the most booking come via TA/TO. City hotel can work to increase out reach on GDS channels to get more higher revenue generating deals.

# Time wise Analysis

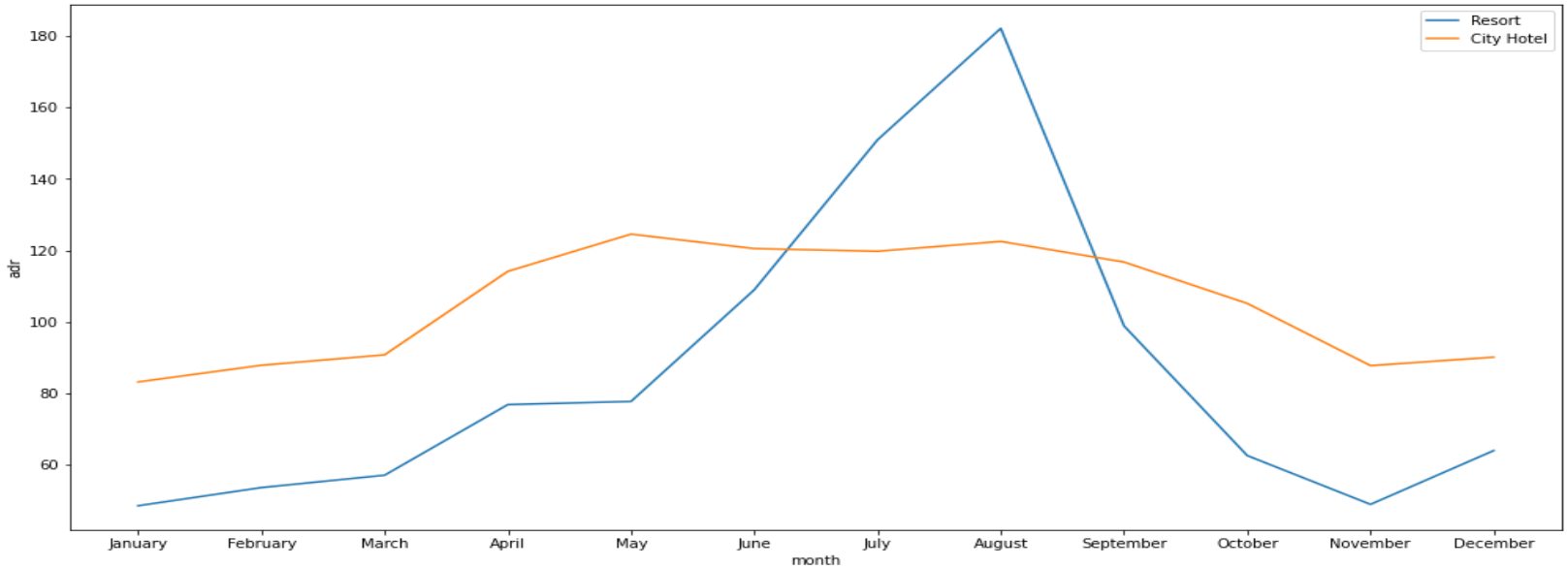
## 1. Which is the most busy month for hotel?




From month June to August the number of booking were increased rapidly for both City and Resort Hotel. And in August most of the booking were done .

# Time wise Analysis

## 2. Which month results in high revenue.



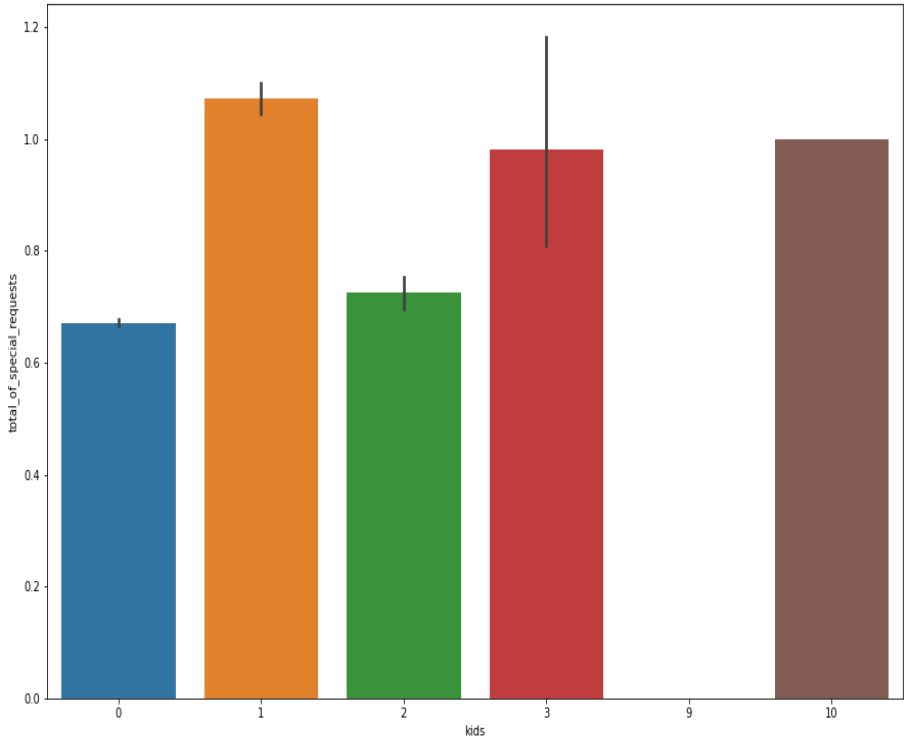

- The revenue aspect looks different, the Resort Hotels receive more revenue with respect to city hotel. From May to August their was rapid increase in adr. In August recorded the highest revenue.

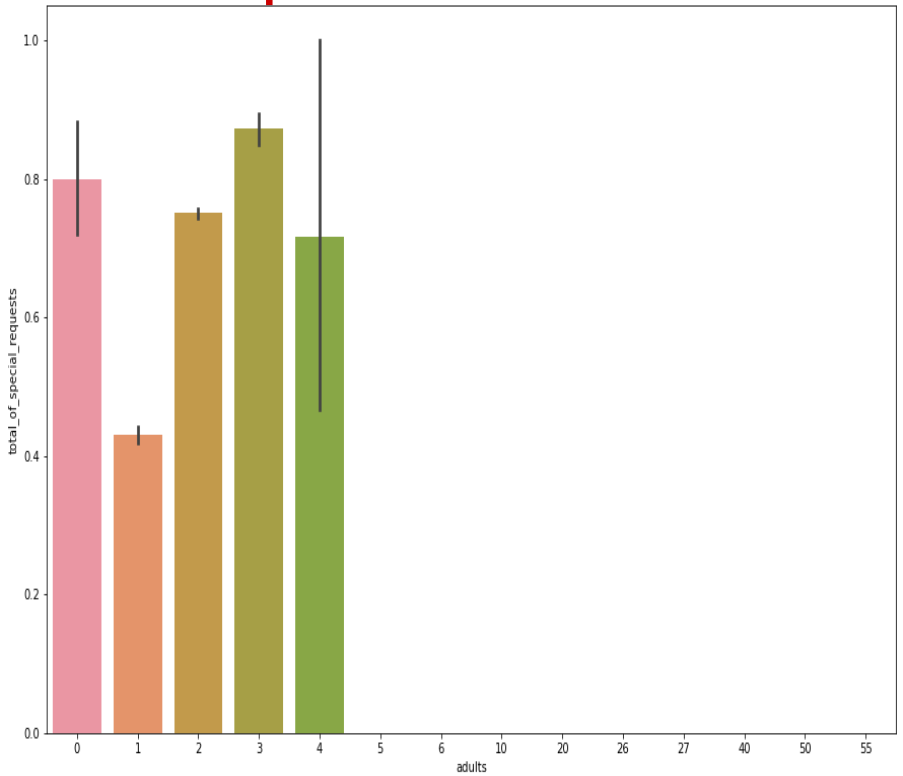
# Some important questions

Some other analysis are also done, which are as follows

- 1. What are the different reason for special request.
- 2. What is the optimal stay length for better deal for customers.
- 3. From which country most guests come?

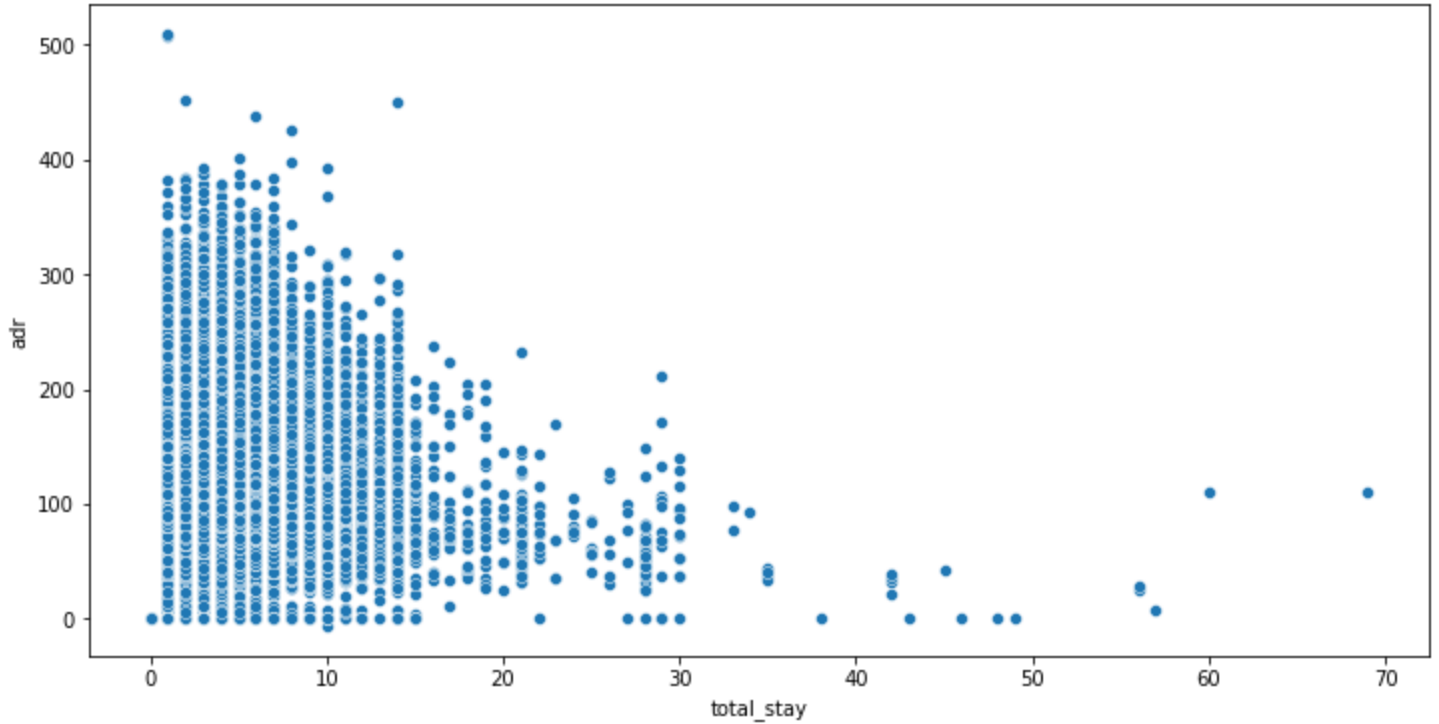
# Reason for special request.



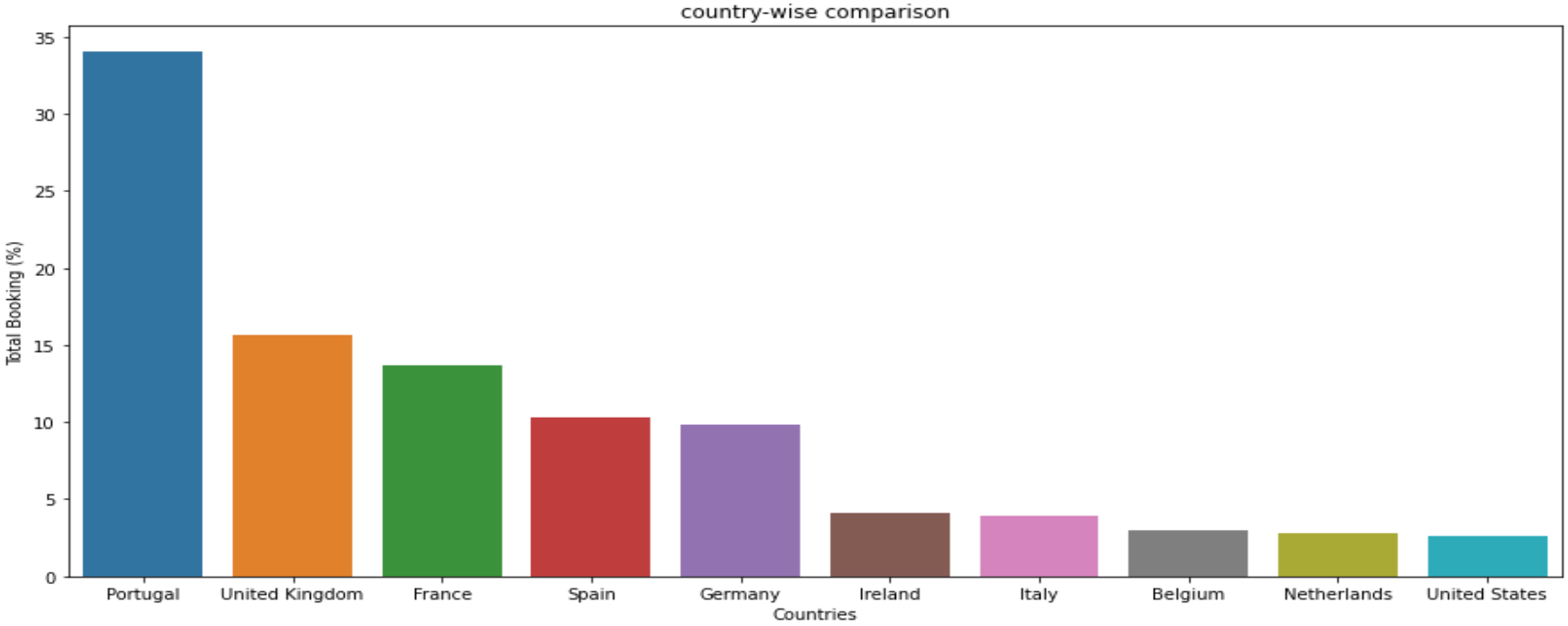

The number of special request are almost same in the kids section. But we can see that if the adults are more than 2 there are more chances that hotels will receive more special request.

# Optimal stay length for better deal for customers.




For shorter stays the adr very high but for longer stays greater than 15 days adr is comparaitively very less. Therefore customer can get better deal for longer stays in terms of lower adr.

# From which country most guests come?




- Portugal, UK and France, Spain and Germany are the top 5 countries from most guests come, more than 80% guests come from these 5 countries.

# Conclusion

- Almost 35% of bookings were canceled.
- More than 60% of the population booked the City hotel.
- More than double bookings were made in 2016, compared to the previous year. But the bookings decreased by almost 15% next year.
- Most bookings were made from July to August. And the least bookings were made at the start and end of the year.
- Mostly guests stay for less than 5 days in hotels and for longer stays Resort hotel is preferred.
- Guests use different channels for making booking out of which most preferred way is TA/TO.
- Portugal, UK and France, Spain and Germany are the 5 top countries from most guests come, more than 80% of guests come from these 5 countries
- For Resort hotel, the most popular stay duration is three, two, one, and seven days respectively.
- For City hotel, most popular stay duration is three, two and one respectively.
- Couple (or 2 adults) is the most popular accommodation type. So hotels can make arrangement plans accordingly
- The best time of booking hotel are the starting and last month of the year in order to get the best the deal.
- For customers, generally the longer stays (more than 15 days) can result in better deal in terms of low adr.
- More number of people in guests results in more number of special request.

# Thank you