Hotel Booking Analysis

**By: Mani Raj**

**Alma Better Capstone Project**

# Abstract:

The hotel industry has a very peculiar set of booking arrangements with different distribution

Channels . I have worked on the hotel industry dataset to get a comprehensive image of the hospitality industry. From basic data visualizations to complex multivariate analysis, I have performed data analysis at every level of the dataset to draw out key insights about the industry.

I was given this project on Wednesday and my mentor suggested me, so i divide the work by choosing 5-6 variables for the analysis. So I started brainstorming on the questions

and assumptions that I wanted to solve using analysis. I arranged all the required documents for the analysis. I am grateful to Alma Better for giving me the opportunity to work on such an exciting project.

# Problem Statement:

I have been put into the shoes of a marketing team for a new hotel startup that wants to know about the hotel industry. I have been provided a dataset of the hotel industry of the region in which our startup aims to begin its initial operations. My primary goal is to understand the market of this particular region and try to use the data analysis techniques to draw out key features of the market. My secondary goal is to draw out actionable insights from our analysis and give conclusions about key aspects of the market such as cancellation rate, distribution channels, and alike.

# Data Summary:

Based upon the initial assessment we found that the data was pretty much clean except for some missing values in a few columns. Upon using the info () method, we draw out the following key insights about the data:-

1. The dataset has a shape of (119390, 32) which means that it contains approximately 1.2 lakh rows and 32 columns.
2. Our Dataset has 4 columns with float64 dtype, 16 columns with int64 dtype, and 12 columns with object dtype.
3. In our Dataset, we observed null values in the following columns:
   * 4 null values in the children column
   * 488 null values in the country column
   * 16,340 null values in the agent column
   * 112,593 null values in the company column

We have the following column names provided to us in the dataset,

* hotel
* is\_canceled
* lead\_time
* arrival\_date\_year
* arrival\_date\_month
* arrival\_date\_week\_number
* arrival\_date\_day\_of\_month
* stays\_in\_weekend\_nights
* stays\_in\_week\_nights
* adults
* children
* babies
* meal
* country
* market\_segment
* distribution\_channel
* is\_repeated\_guest
* previous\_cancellations
* previous\_bookings\_not\_canceled
* reserved\_room\_type
* assigned\_room\_type
* booking\_changes
* deposit\_type
* agent
* company
* days\_in\_waiting\_list
* Customer\_type
* adr
* required\_car\_parking\_spaces
* total\_of\_special\_requests
* reservation\_status
* reservation\_status\_date

# Steps involved in the Data Analysis: -

1. **Framing the questions:** Before any form of analysis, it is important to frame the questions that we want to know from the data. For this, I performed many creative thinking techniques such as brainstorming, discuss with our friends, to take out all the questions that could be asked related to the dataset.
2. **Filtering out the ideas:** After getting a long list of questions and

Assumptions that I want to solve from the dataset. As per the guidance of our mentor, we started filtering out my question list. I approach to filter questions was to think from a customer and business perspective and look for answers to only those questions that mattered to any one of them.

1. **Cleaning the Data:** As mentioned before, the dataset provided to me by the Alma Better faculty was pretty much clean from the start. By clean I mean that it didn’t have any nested lists or dictionaries as row elements or wrong dtypes in data frame columns. However, it had four variables with null values, so I had to take them into account before proceeding further with our analysis.
2. **EDA analysis:** By EDA I mean exploratory data analysis. In this, I looked at the data frame and decided the target variables (Important Columns) based upon which I was going to conduct further analysis. I started

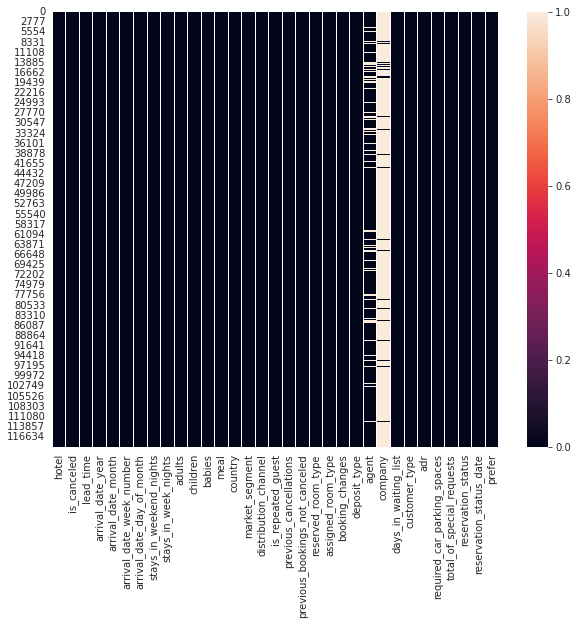
comparing our target variable with other independent variables (leftover columns) to get insights into the relationship between them. This also gave us a better picture of how the different variables affect the target variable.

1. **Visualization of Insights:** After I completed the analysis of the data, I used matplotlib and seaborn libraries of python to present analysis graphically. I used Pie charts, bar charts, heatmaps, displots, and much more to give complex insights in an eye-catching manner. I learned a lot about the different visualization tools that are available for data analysis.

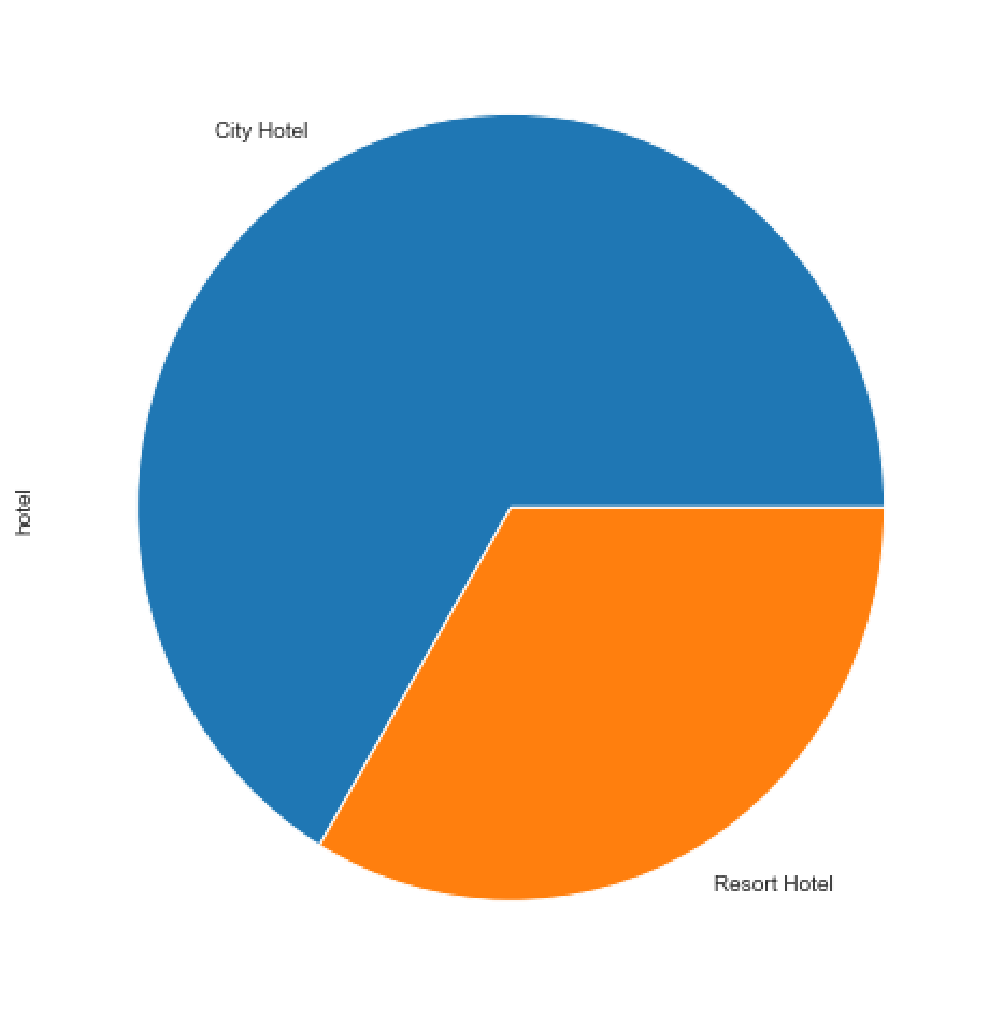
**Drawing Conclusions and Finding Answers:** Finally I warped up each analysis by drawing out conclusions from them.

**Let's find insights from the Dataset**

**Finding Null value using Heatmap**

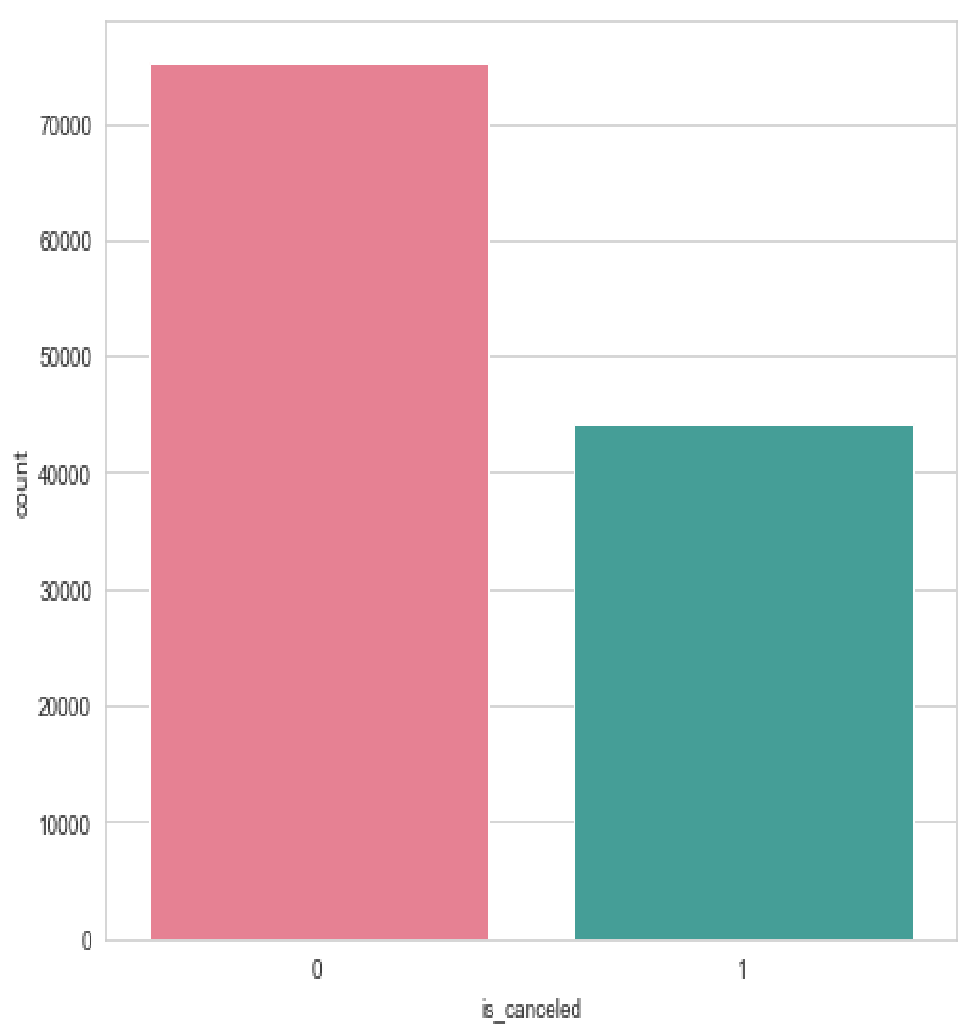
****As it is clearly seen that ‘agent’ and ‘company’ column has large number of null values.

**Types of Hotels**



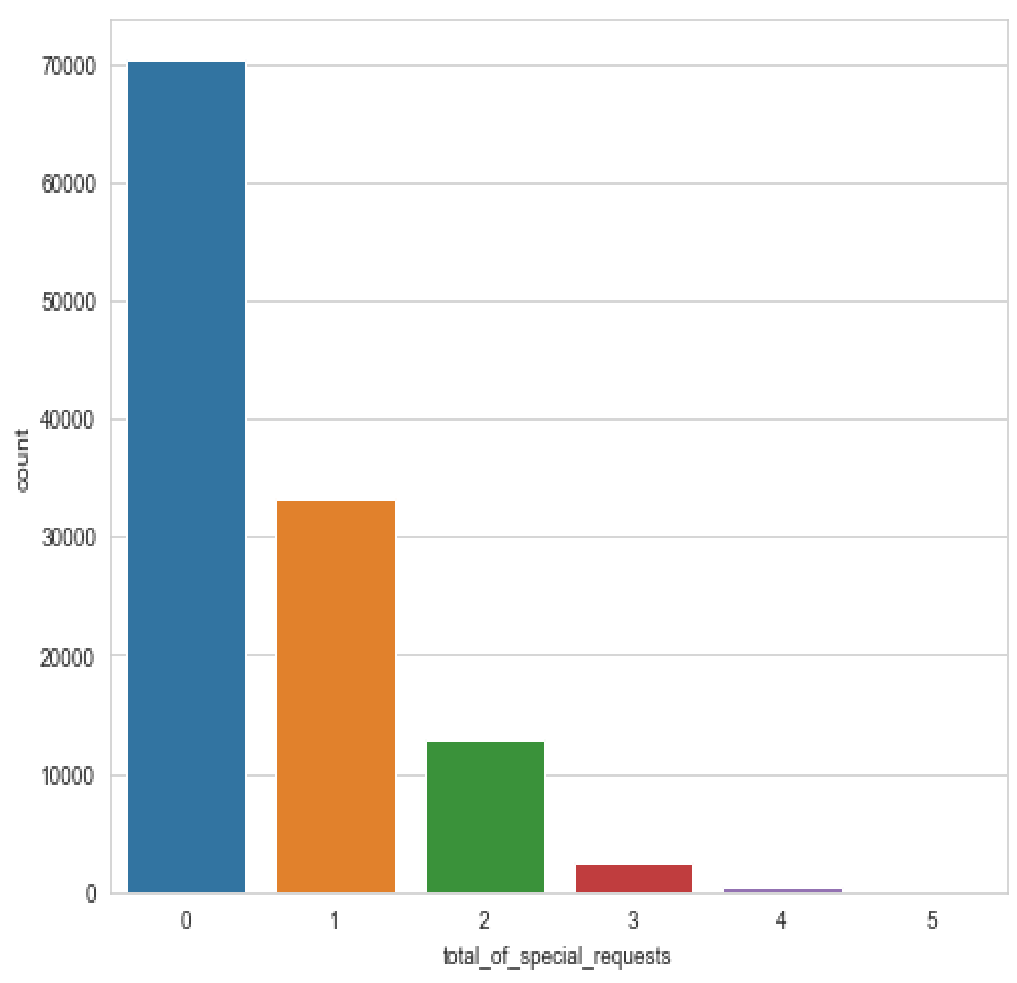
* Majority of the booked Hotel is City Hotel than Resort Hotel

**HOW MANY BOOKING WERE CANCELLED?**

****

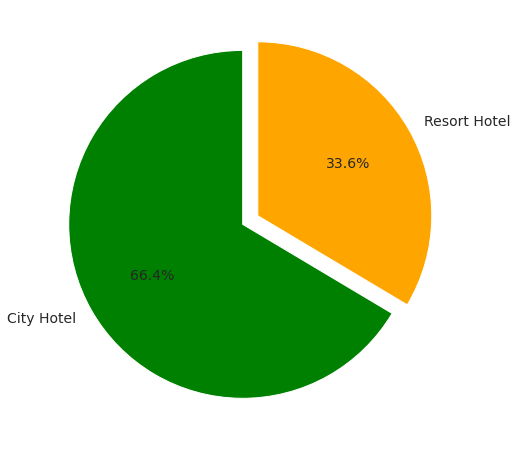
Conclusion: During the year I have 37% of cancelations

**SPECIAL REQUESTS**

****

As I can see here among all one special booking request were made almost 27% of total bookings. Two special request were made nearly 10% among all and 3 special request is nearly 2%.

**BOOKING RATIO BETWEEN RESORT HOTEL AND CITY HOTEL**

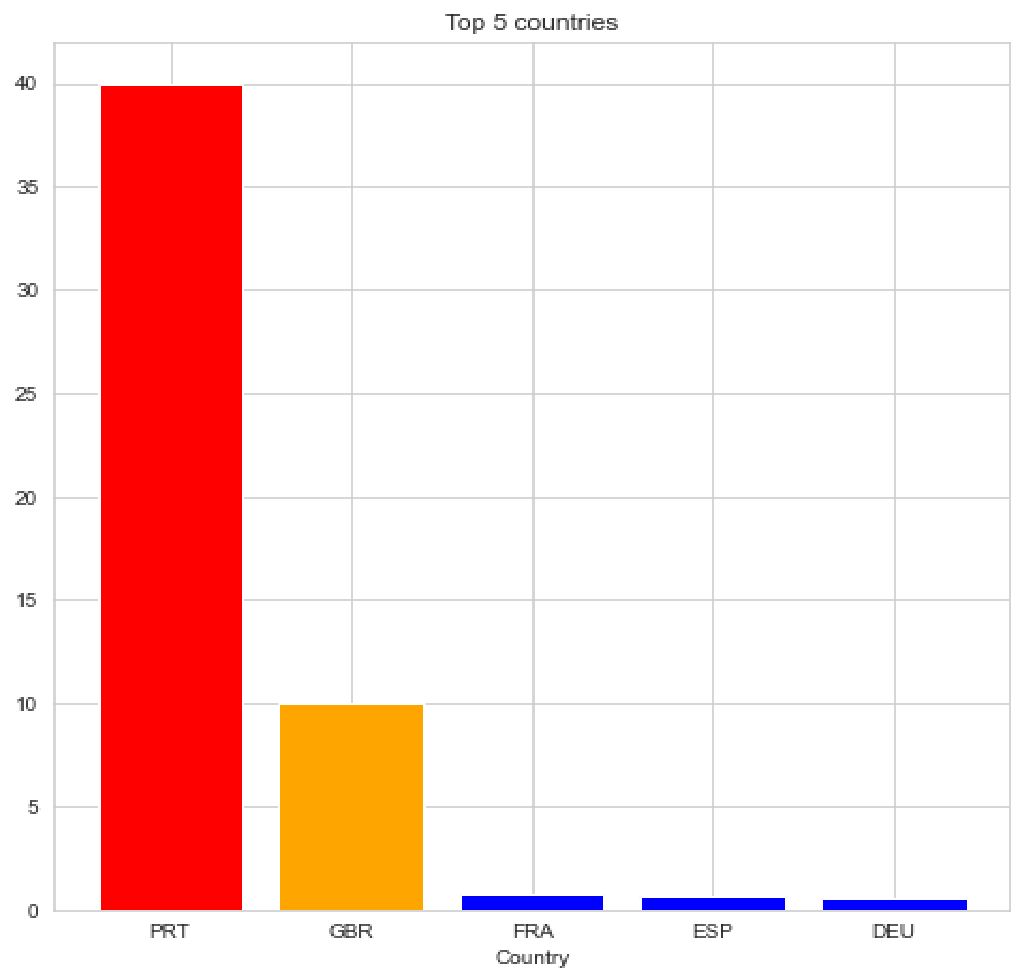
****

What do I see here?

It seems that a huge proportion of hotels were city hotel. Resort hotel tend to be on the expensive side and most people will just

Stick with city hotel. Also, resort hotels tend to be appropriate for larger group of people

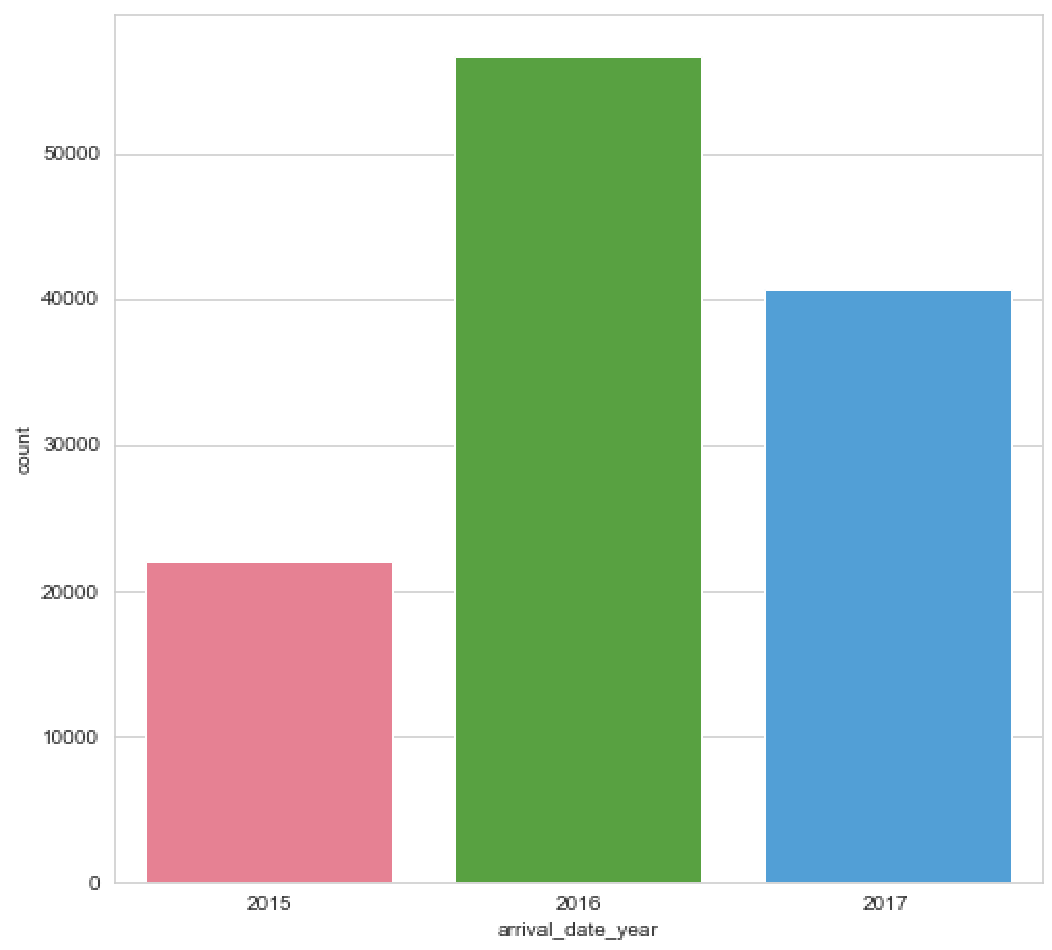
**FROM WHICH COUNTRY MOST GUEST COME?**

****

As I can see, Portugal tops the list with 40.69% of the guest, followed by – Great Britain with 10.15%, France, Spain and Germany.

Majority of the guests are from European Countries.

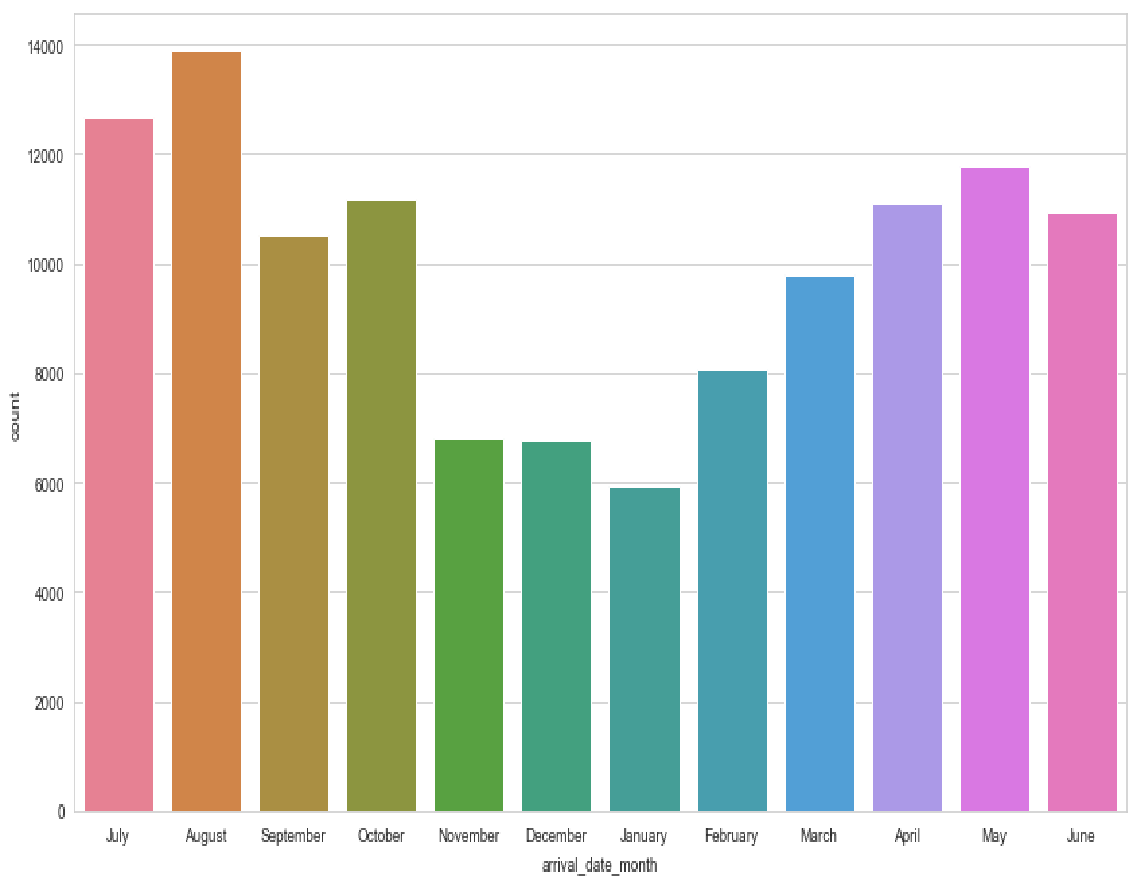
**BOOKING PER YEAR**

******

There has been many arrivals in the year 2016 than the remaning years

We can also say that there has been increase in the arrivals as years passes.

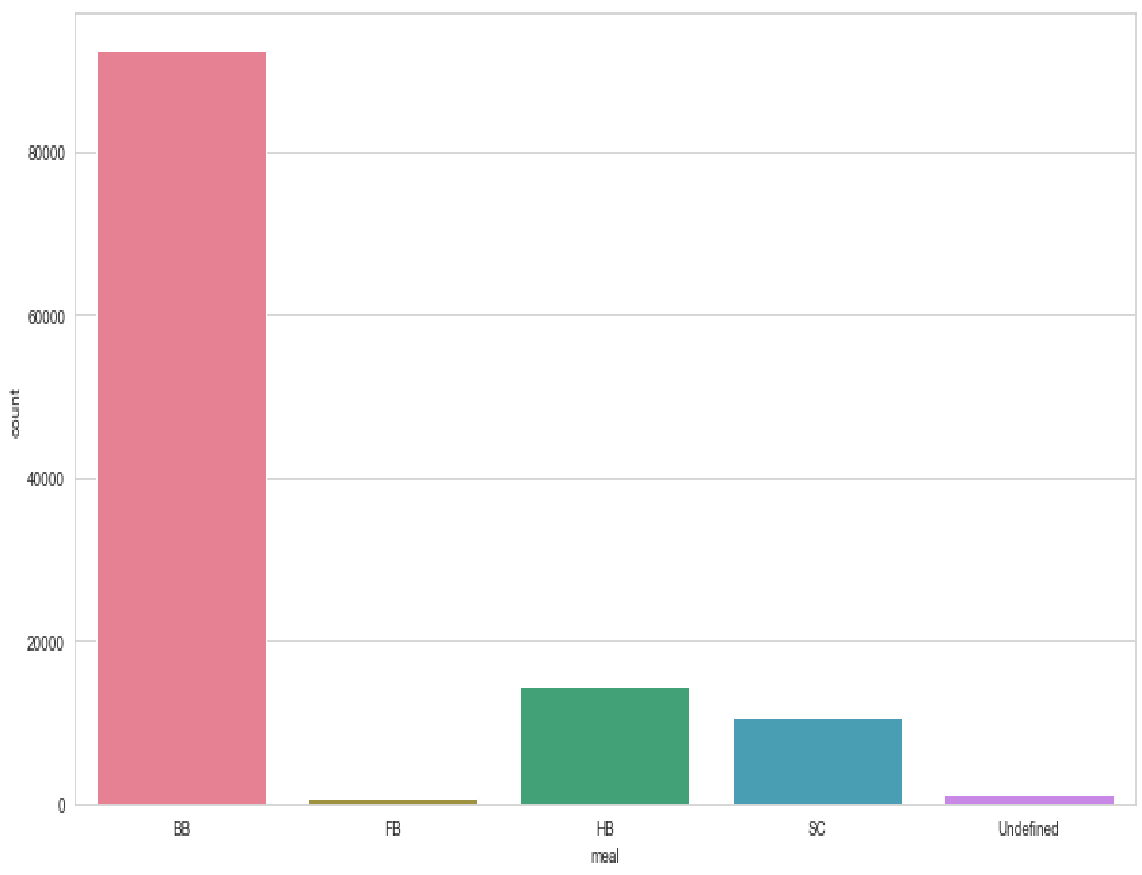
**BUSIEST MONTH FOR HOTELS**



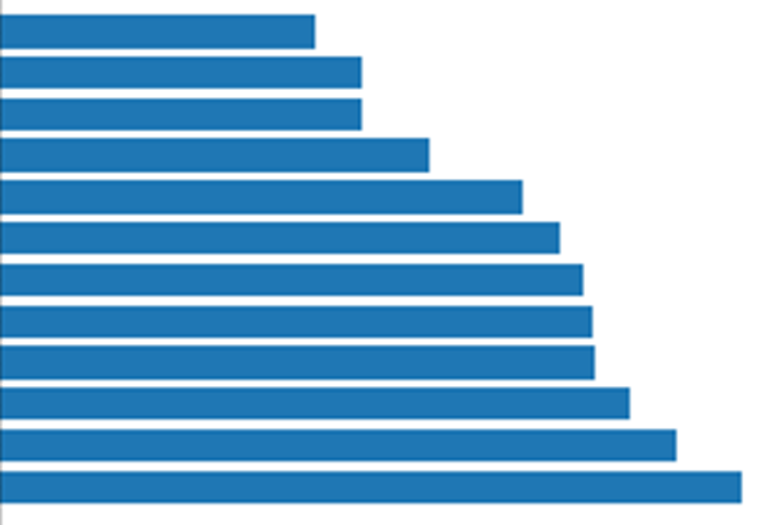
The month of highest occupation is august with 11.62% of the reservations. The month of least occupation is January with 4.96% of the reservations.

The Busiest month for hotel is August with 11.62% of the reservations.

**MEAL TYPE**

******

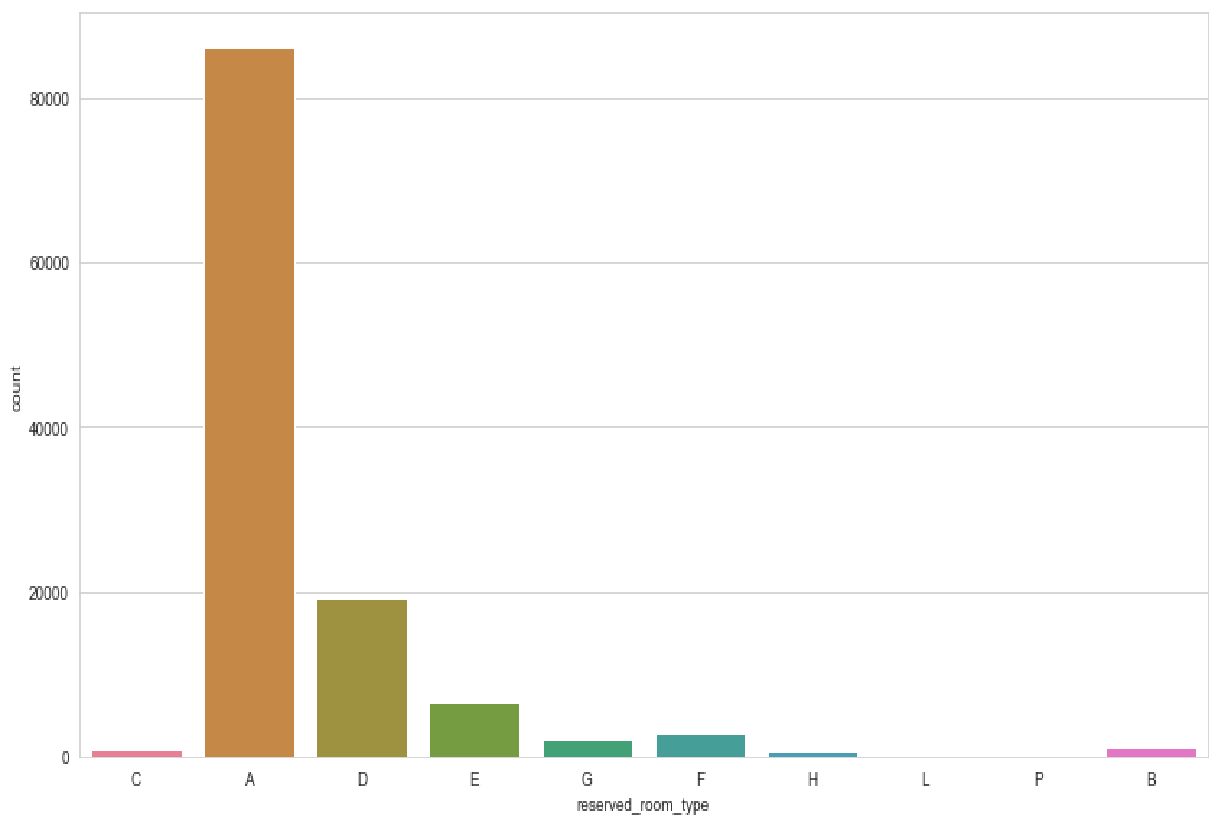
**No of travellers in Various Months**

****

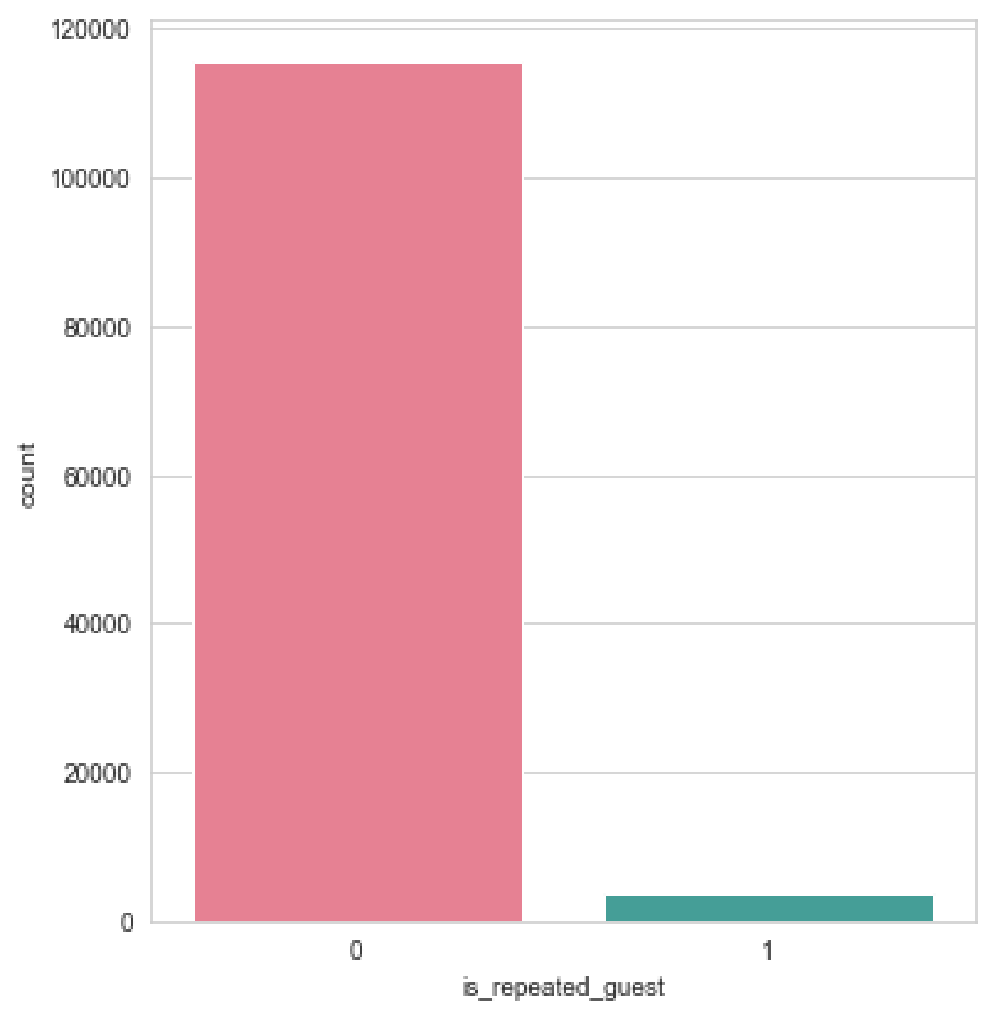


So, I can say that peak months are: August, July, May (Summer Time)

**ROOM TYPE**

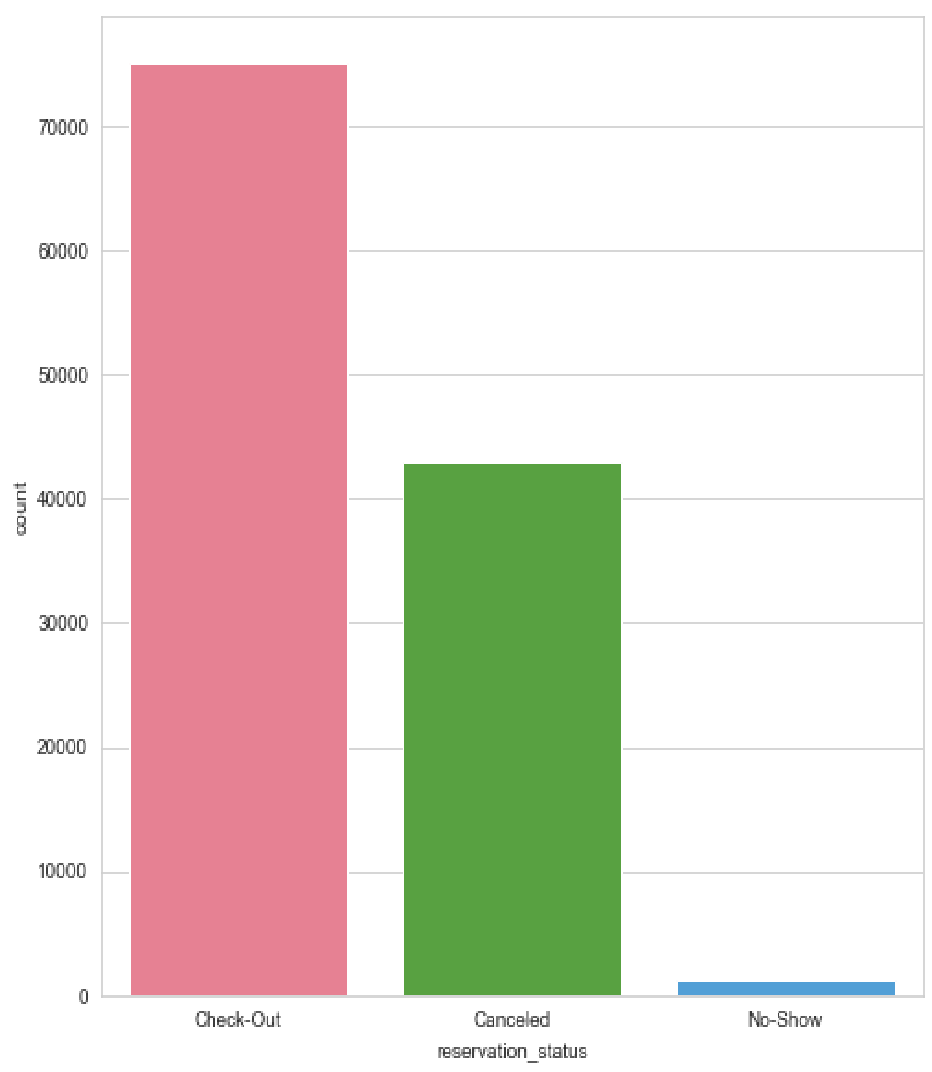
******

**REPEATED GUEST**

******

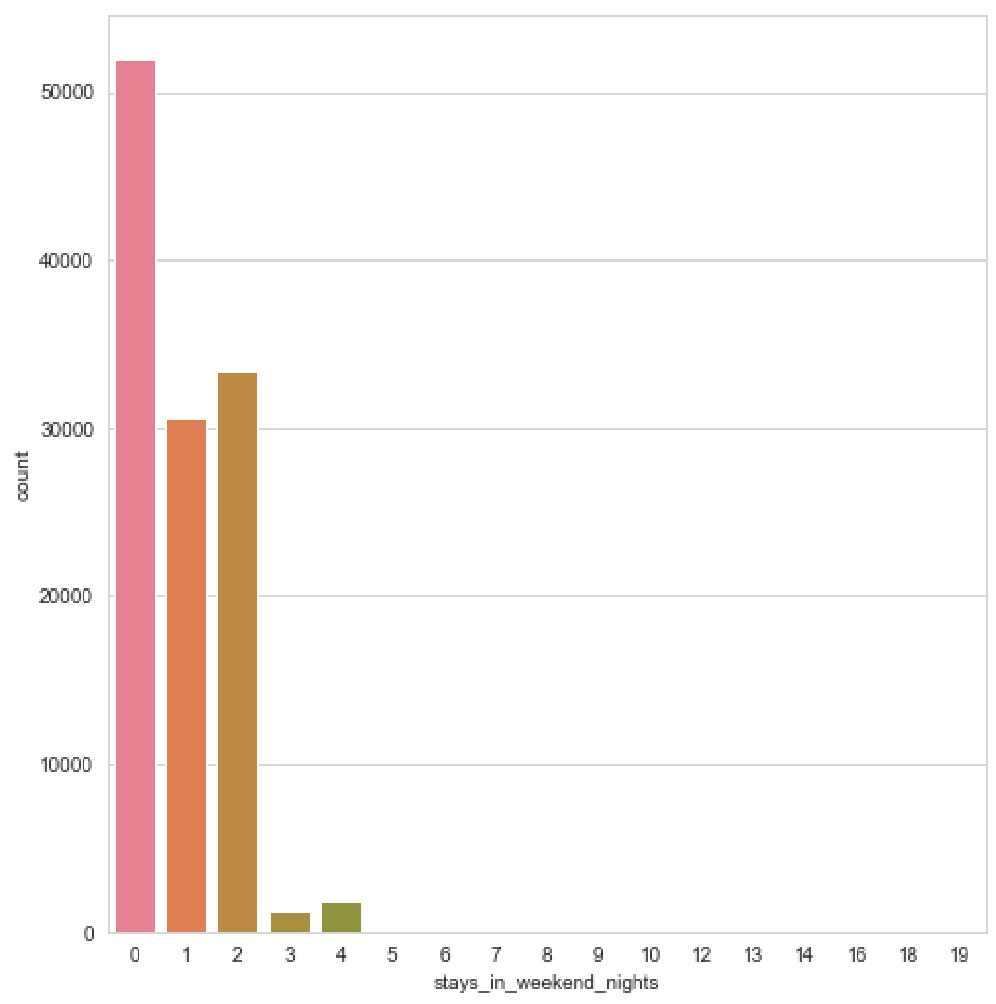
* Repetition of the customer is very rare

**Reservation Status**

****

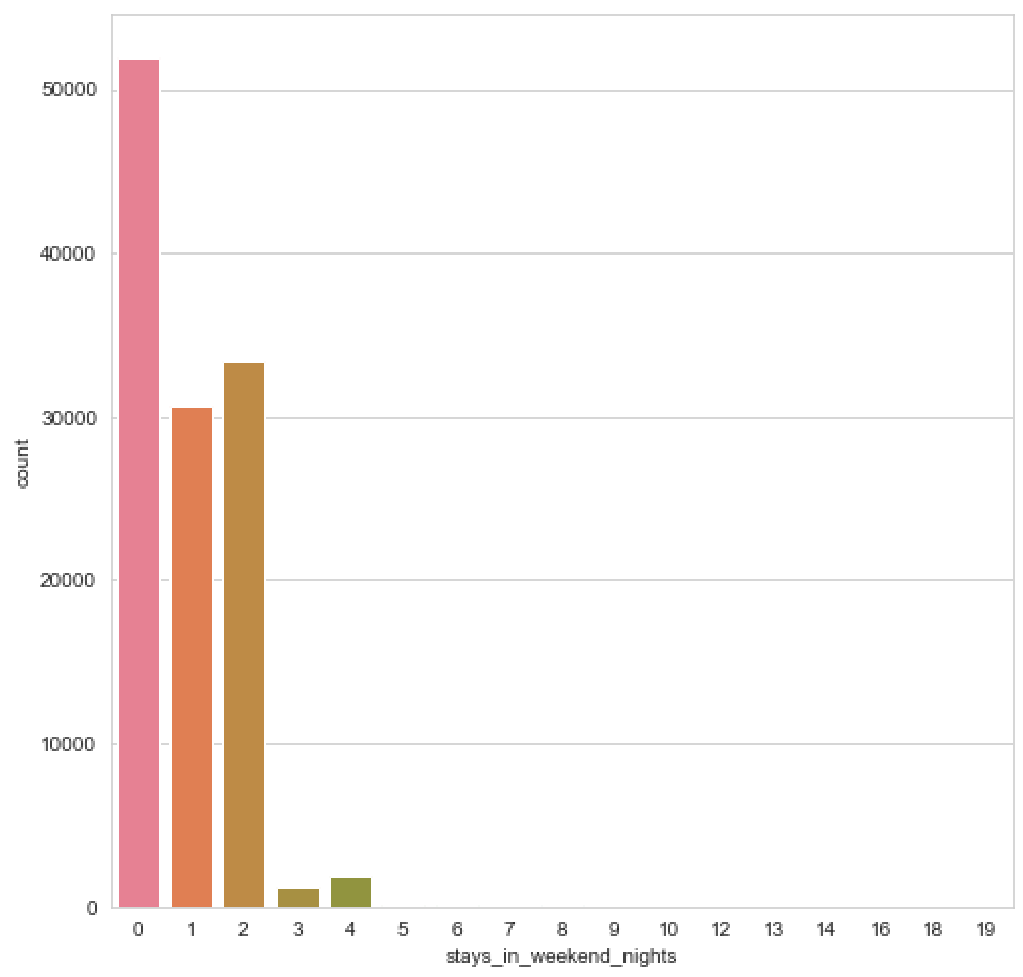
* Check out is most preferred by the customers

**AVERAGE STAY WEEKENDS VS WEEKDAYS**

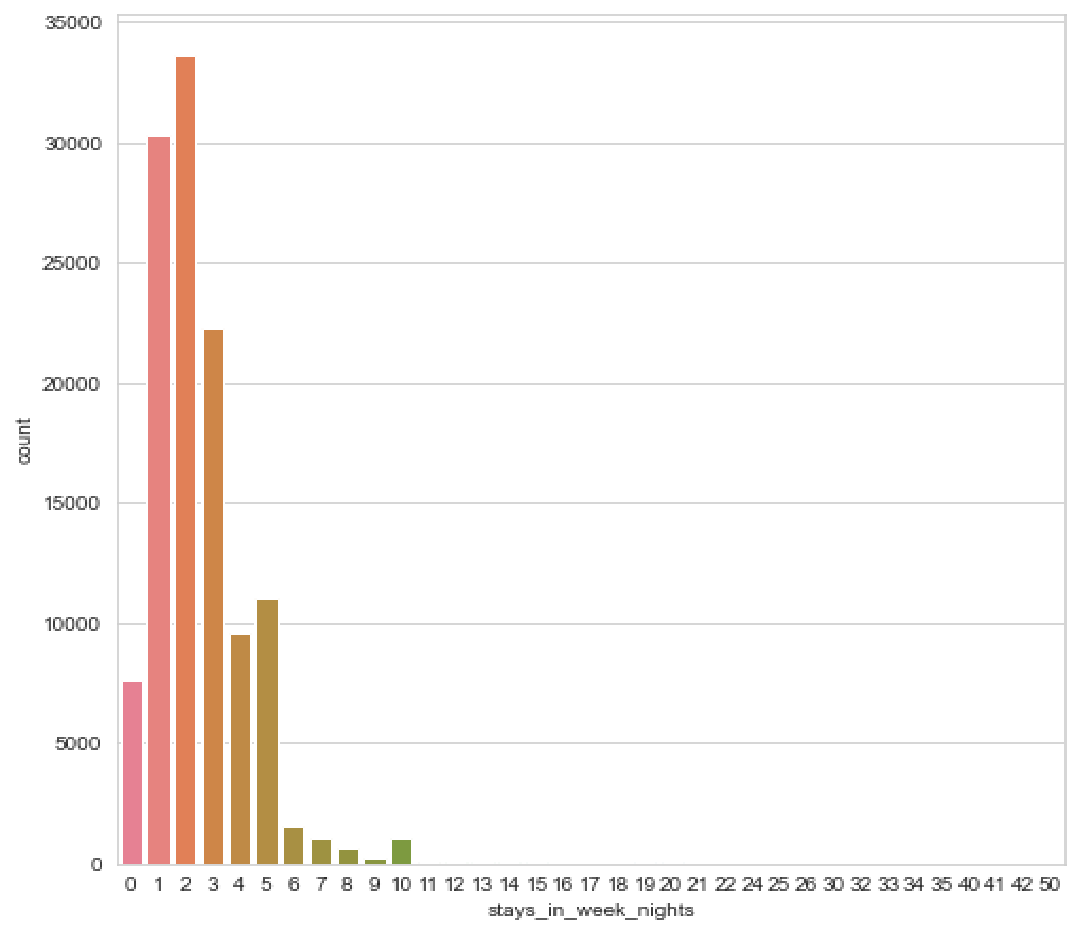
******

* PEOPLE GENERALLY PREFER LONG STAYS ON WEEKDAYS RATHER THAN WEEKENDS

**AVERAGE STAYS WEEKENDS**

******

**AVERAGE STAYS WEEKDAYS**

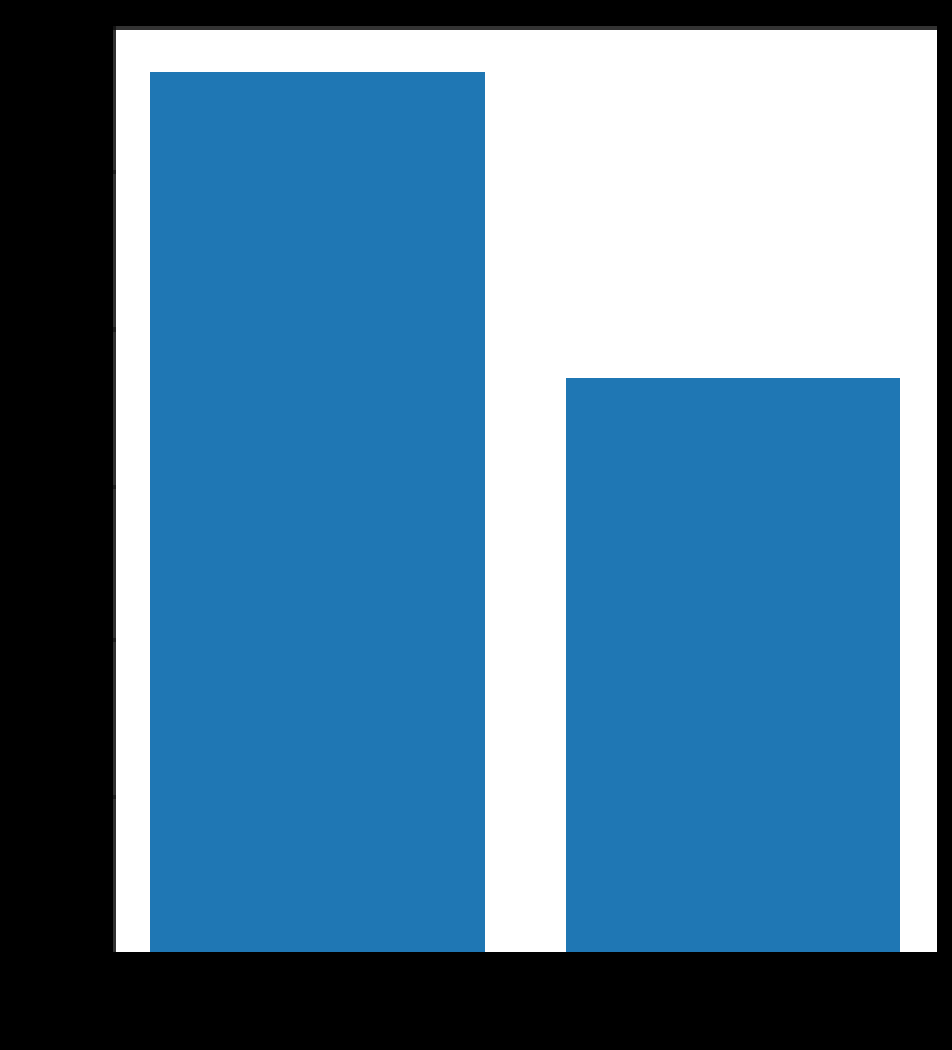
******

* There are around 23,027 rooms booked as single room

* There are around 89,680 rooms booked as couples room

**Online Booking v/s Offline Booking**

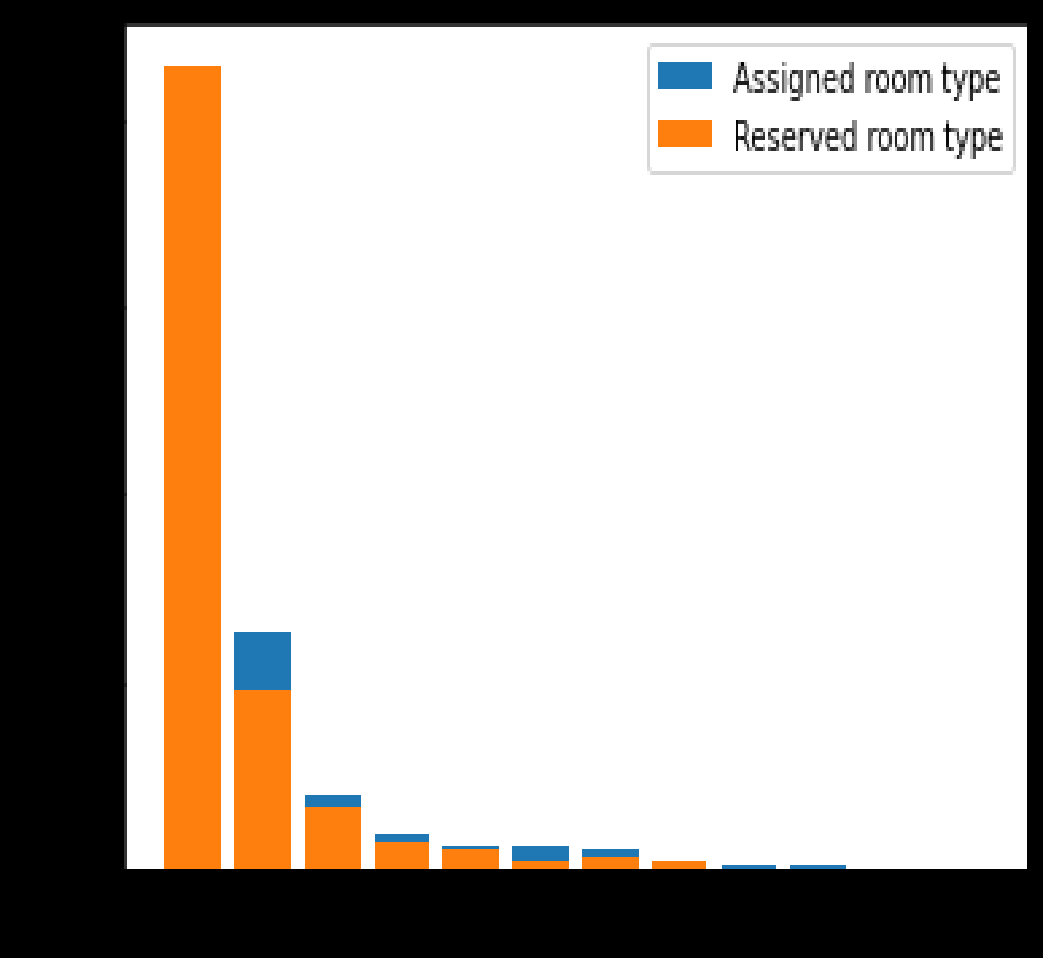




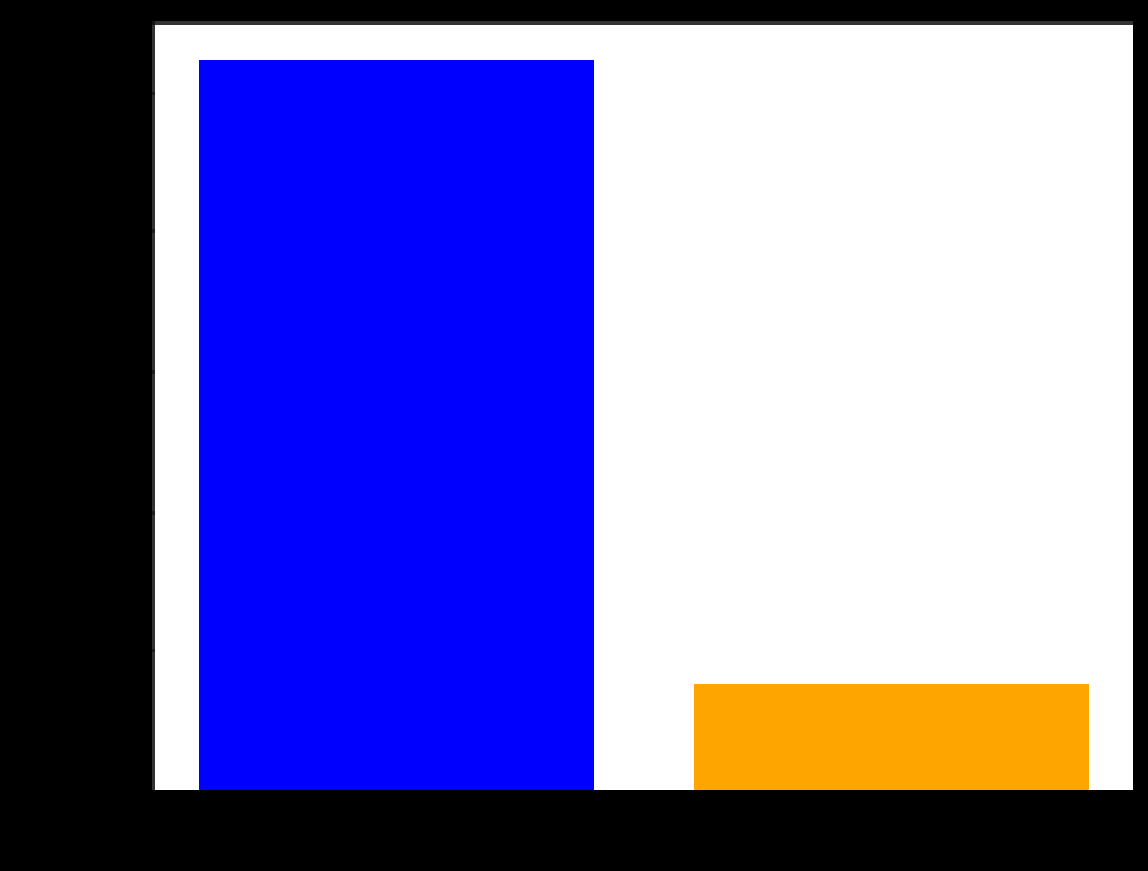
Online Booking Offline Booking

* So, Majority of the customers book through online rather than offline booking

**Satisfactory of customers**

****

**A D E F G C B H I K**

******

Satisfied Change of Room

* Majority of the customers has got the room type they have booked

**Conclusion**

* Majority of the hotels booked are city hotel. Definitely need to spend the most targeting fund on those hotels.
* We also realize that the high rate of cancellations can be due high no deposit policies.
* We should also target months between May to Aug. Those are peak months due to the summer period.
* Majority of the guests are from Western Europe. We should spend a significant amount of our budget on those areas.
* Given that we do not have repeated guests, we should target our advertisement on guests to increase returning guests