**Capstone project**

# EDA on Hotel Booking

# Analysis by

**Contributor Role**

**Suraj Kad (**[**suraj.kad.90@gmail.com**](mailto:suraj.kad.90@gmail.com)**)**

* **Shubham Sawant (**[**shubhamsawant248@gmail.com**](mailto:shubhamsawant248@gmail.com)**)**
* **TanvirPatel**[**tanviripatel1998@gmail.com**](mailto:tanviripatel1998@gmail.com)

**DataScienceTrainees**

**Almabetter**

1. **Defining Problem statement**
2. **Reading and Inspection**
3. **Cleaning the Dataset**
4. **Exploratory Data Analysis**
5. **Data Visualization**

**DATA PIPELINE**



## Data Pipeline :

A data pipeline is a set of tools and processes used to automate the movement and transformation of data between a source system and a target repository.

## Data Reading:

read and write tabular data using pandas functions.



Data, when initially obtained, must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format (*known as* structured data) for further analysis, often through the use of spreadsheet or statistical software.



Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.



Once the datasets are cleaned, they can then be analyzed. Analysts may apply a variety of techniques, referred to as exploratory data analysis, to begin understanding the messages contained within the obtained data.



Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

* 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 you 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.
* All personally identifying information has been removed from the data. Explore and analyze the data to discover important factors that govern the bookings

* + 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
    - 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.
    - distribution\_channel : Booking distribution channel.
    - 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.
* 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.

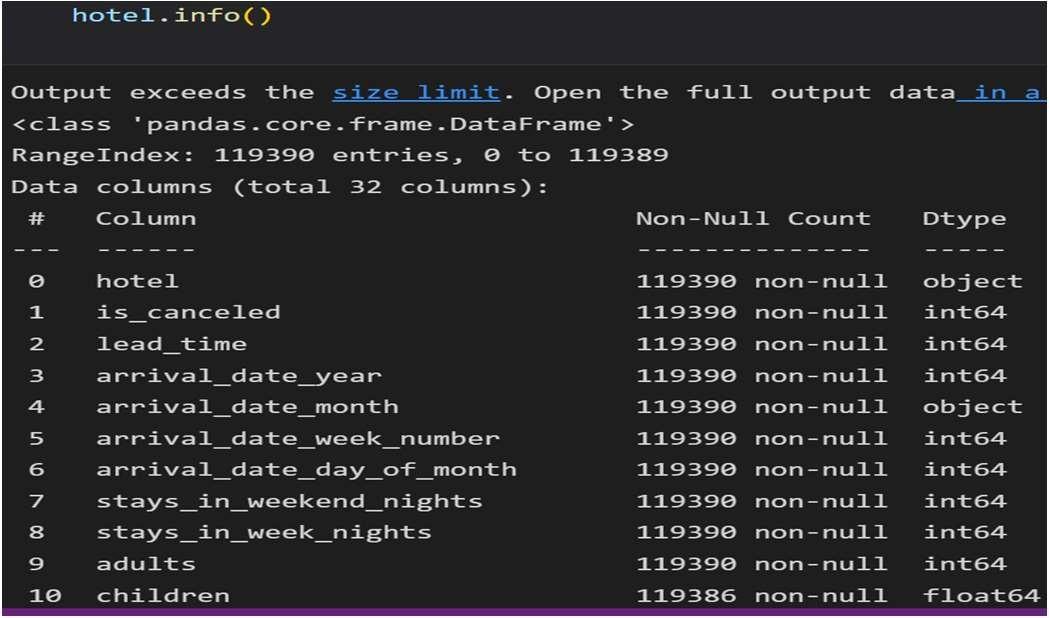
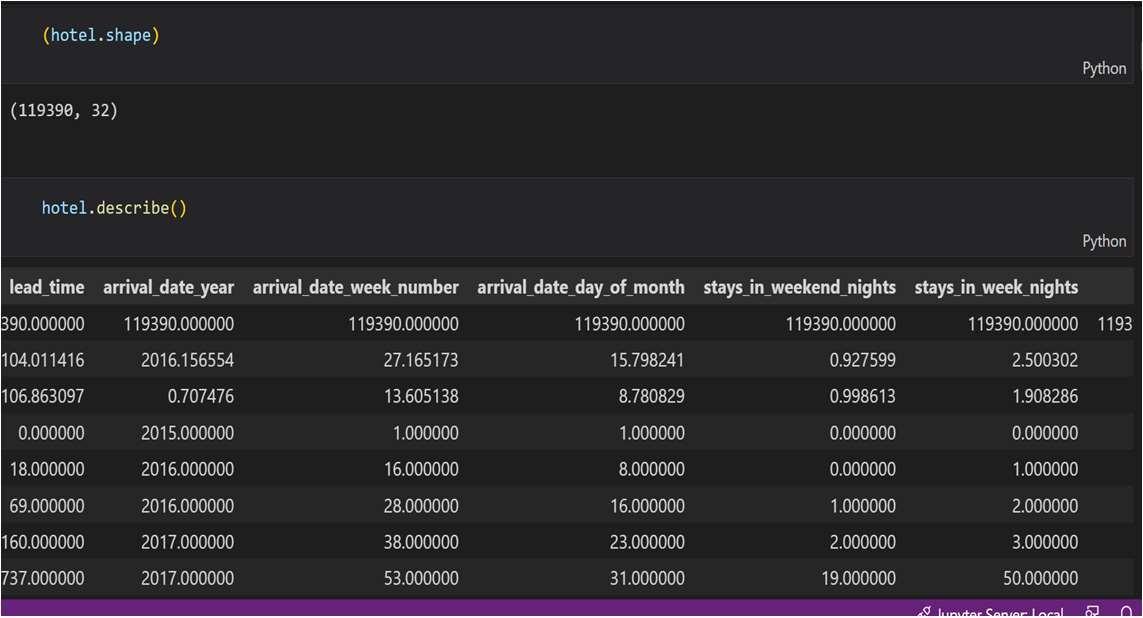
 



* + Read and write Hotel Booking (**tabular)** data using pandas functions

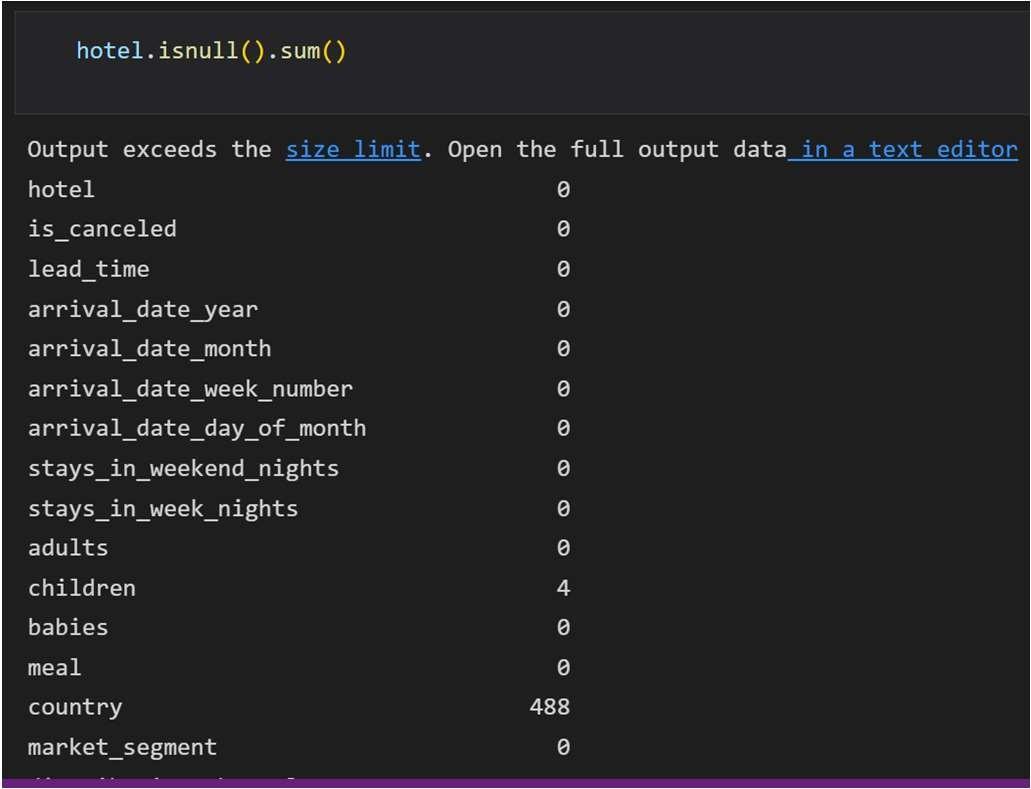
Pandas **head()** method is used to return top n (5 by default) rows of a data frame for hotel booking dataFrame.

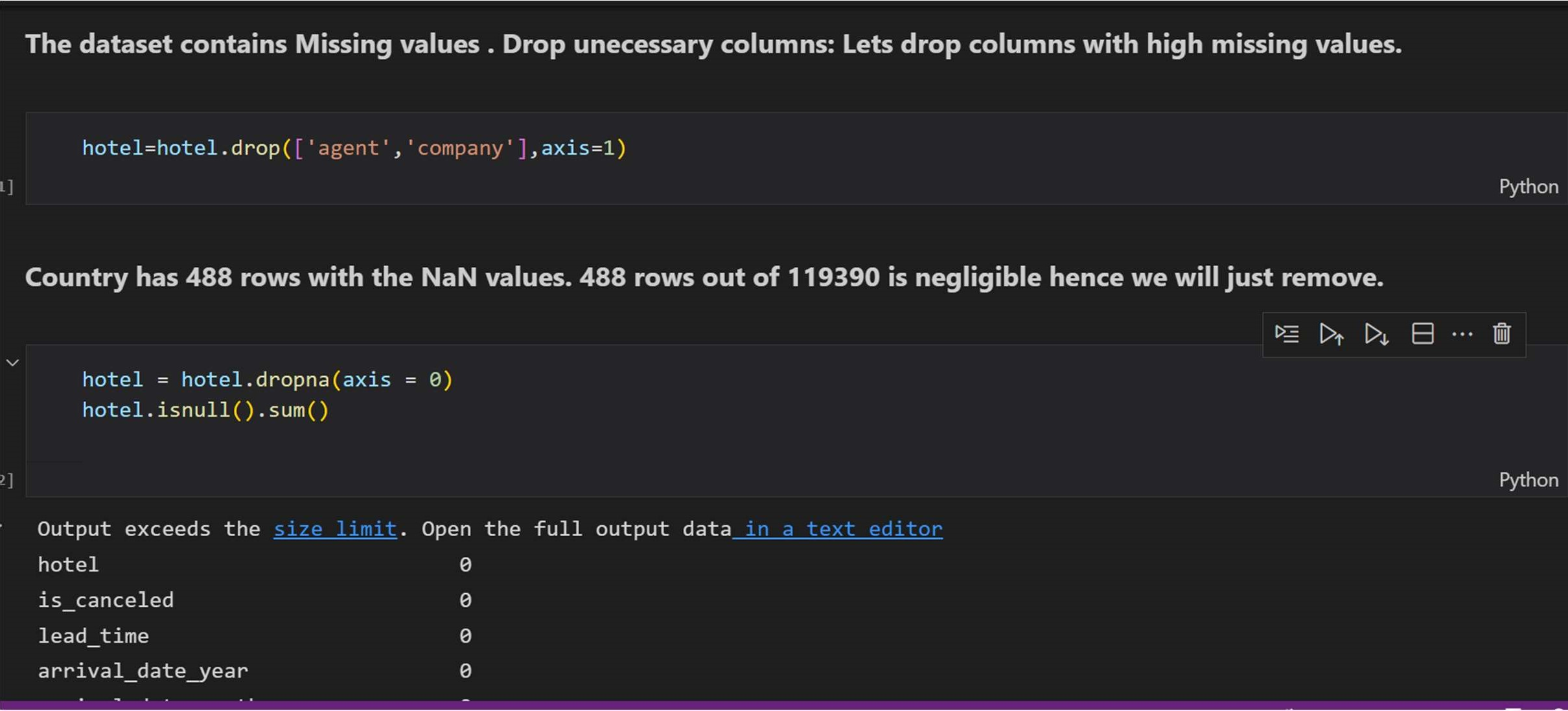
 

* The info() method **prints information about the hotel booking DataFrame**. The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).
* Shape attribute in Pandas **enables us to obtain the shape of a DataFrame**. In this DataFrame has a shape of (119390, 32) , this implies that the DataFrame is made up of 119390 rows and 32 columns of data.
* The describe() method **returns description of the data in the DataFrame.** If the DataFrame contains numerical data, the description contains these information for each column

* + We will count total number of NaN data present in hotel booking dataset and find out the number of NaN or missing values in each columns.
  + As we can see in the output**, hotel.isnull().sum()**

Function has return **country** columns **488 NaN** data are present.



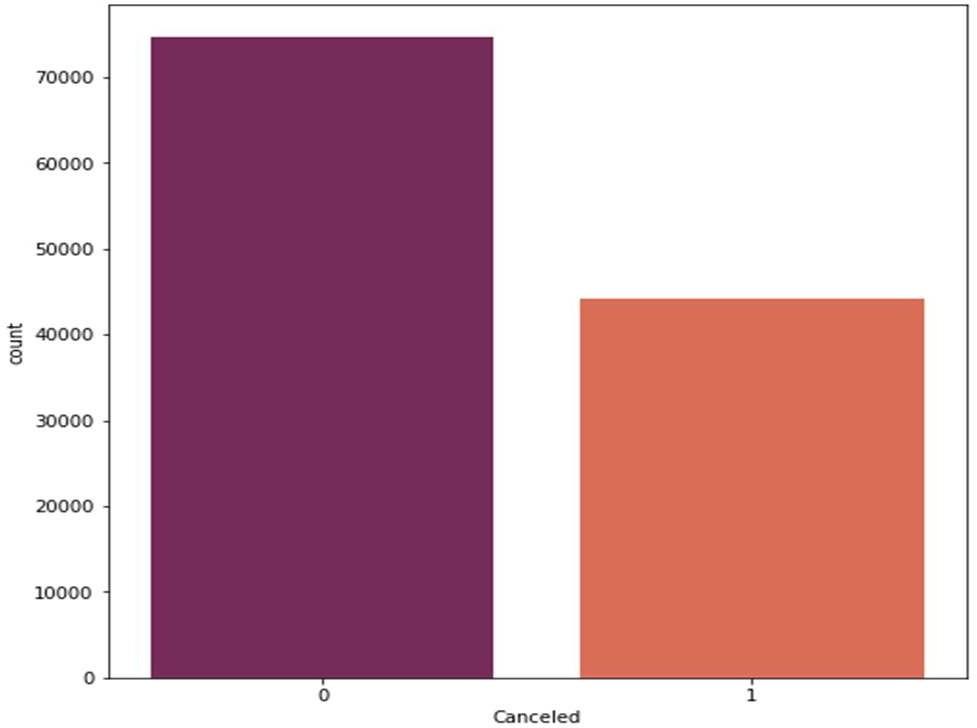
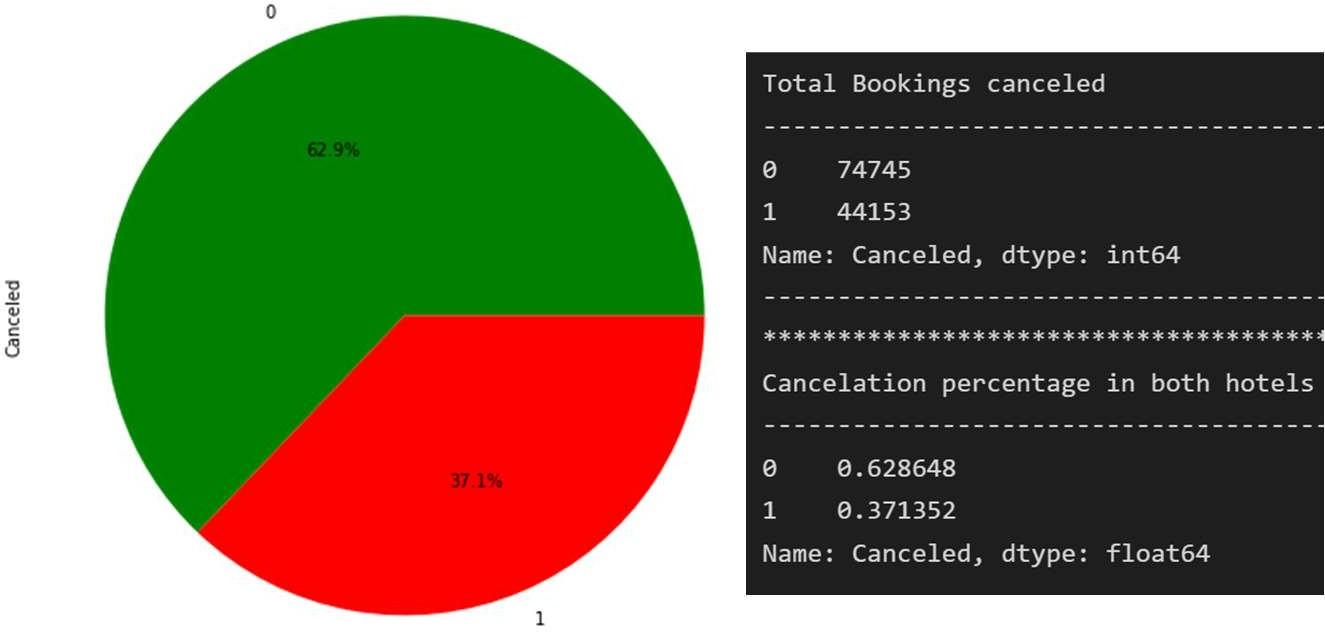
 

* + - This method is quite useful when we need to rename some selected columns because we need to specify information only for the columns which are to be renamed
    - we'll convert each value of a column of integers to string using the **astype(str)** function

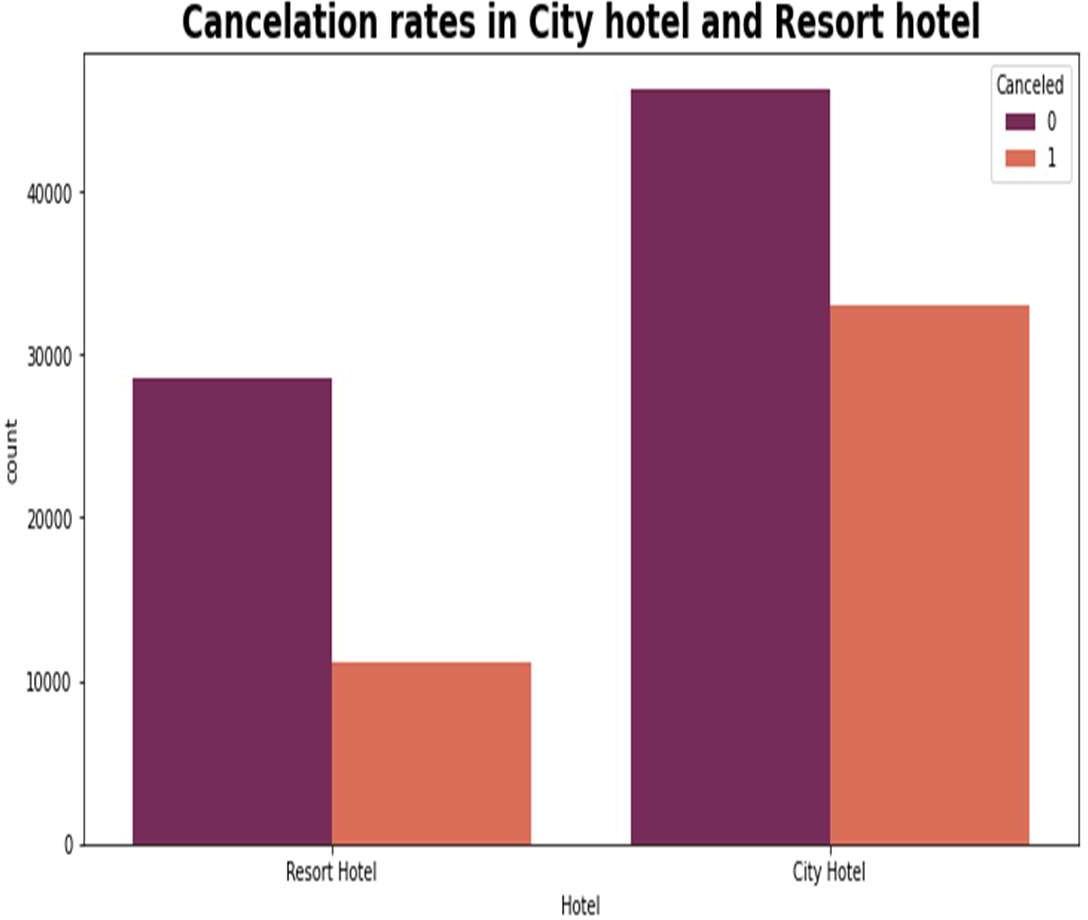
* Around 66.4% bookings are for City hotel and 33.6% bookings are for Resort hotel.
* Avg adr of Resort hotel is slightly lower than that of City hotel. Hence, City hotel seems to be making slightly more revenue.
  + Total stay length and lead time are slightly correlated. This may means that for longer hotel stays, people generally plan little before the actual arrival.
  + adr is slightly correlated with total\_people, which makes sense as more no. of people means more service to deliver, therefore more adr.

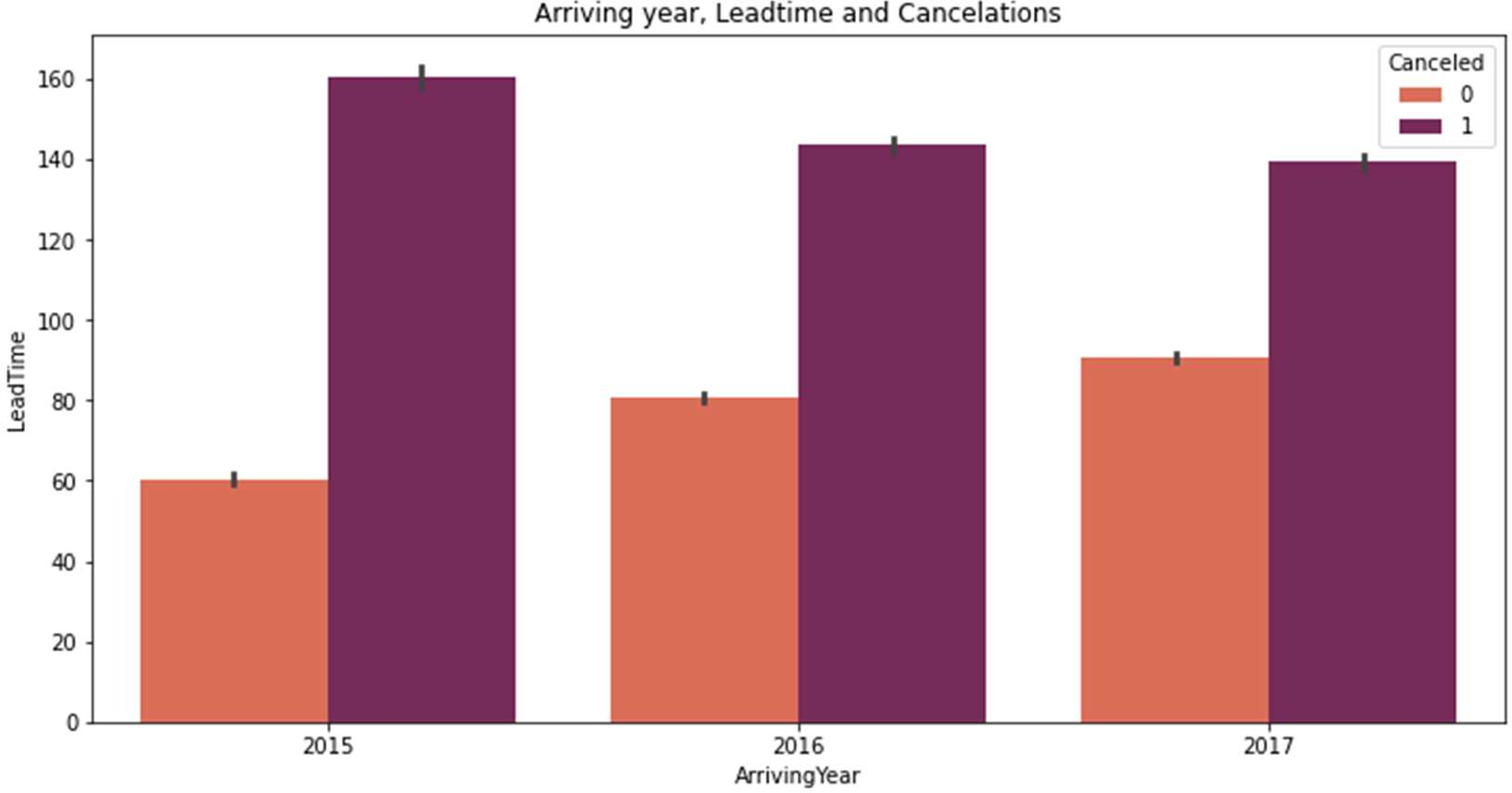
 

* Majority of bookings were not canceled, still some half of the bookings were canceled
  + According to the pie chart, 63% of bookings were not canceled and 37% of the bookings were canceled at the Hotel.

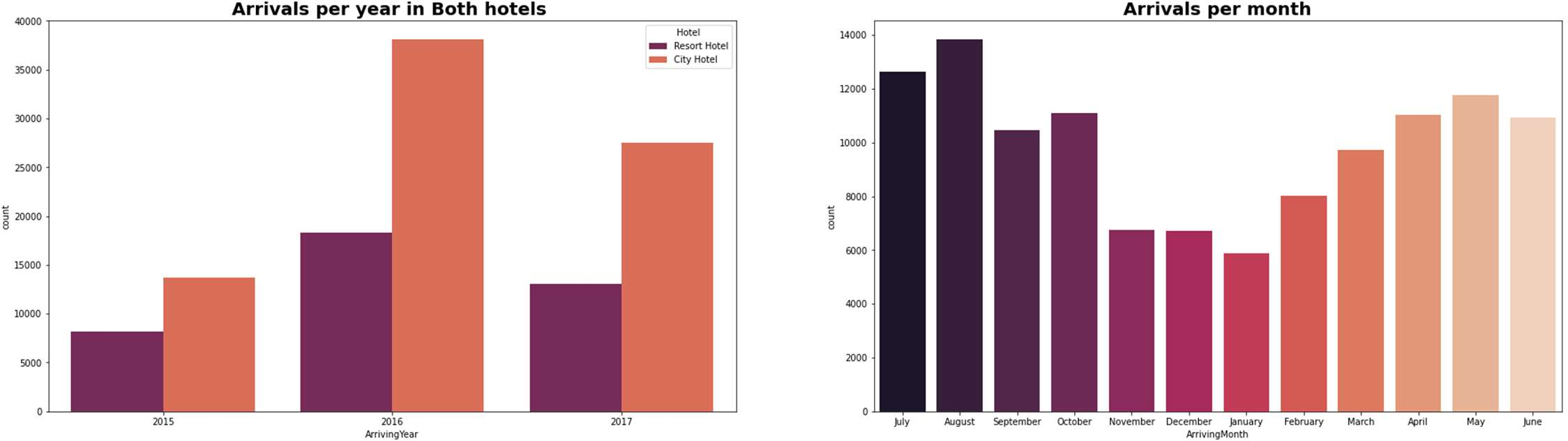
* + - Most bookings were in city hotel
    - Cancelations in Resort hotel is less compared to city hotel



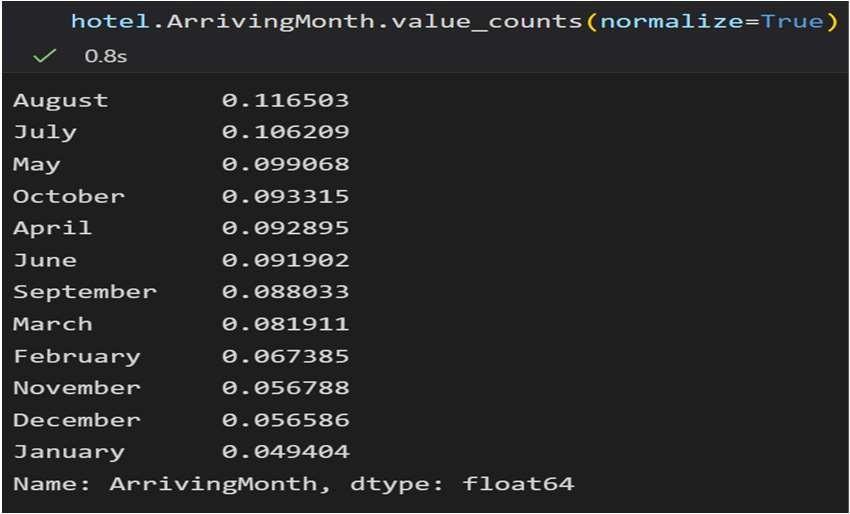
For all the 3 years, bookings with a lead time less than 100 days have fewer chances of getting cancelled, and lead time more than 100 days have more chances of getting cancelled.

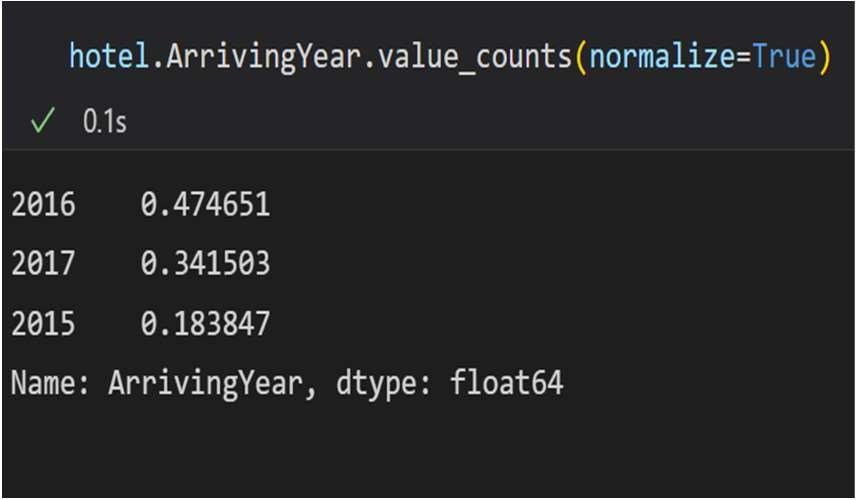
 



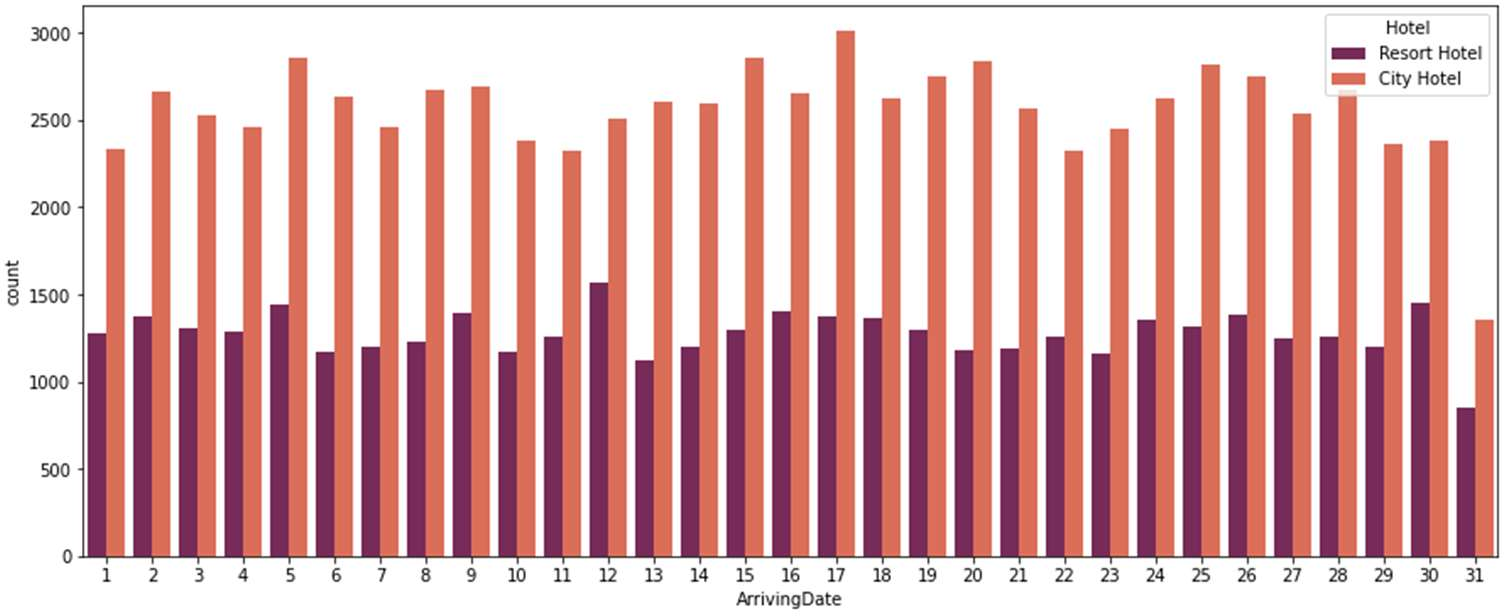
* We can see most of the bookings were in the year 2016 and bookings were done in City hotel
* Most bookings were done in the month of May , June, July, August add Codded Markdown
* From the month of July to August the number of bookings increased and in August, City Hotel got most number of guests.

47% bookings were done in 2016, 34% in 2017 and 18 percent in 2015. We can see increasing tendency in bookings year wise

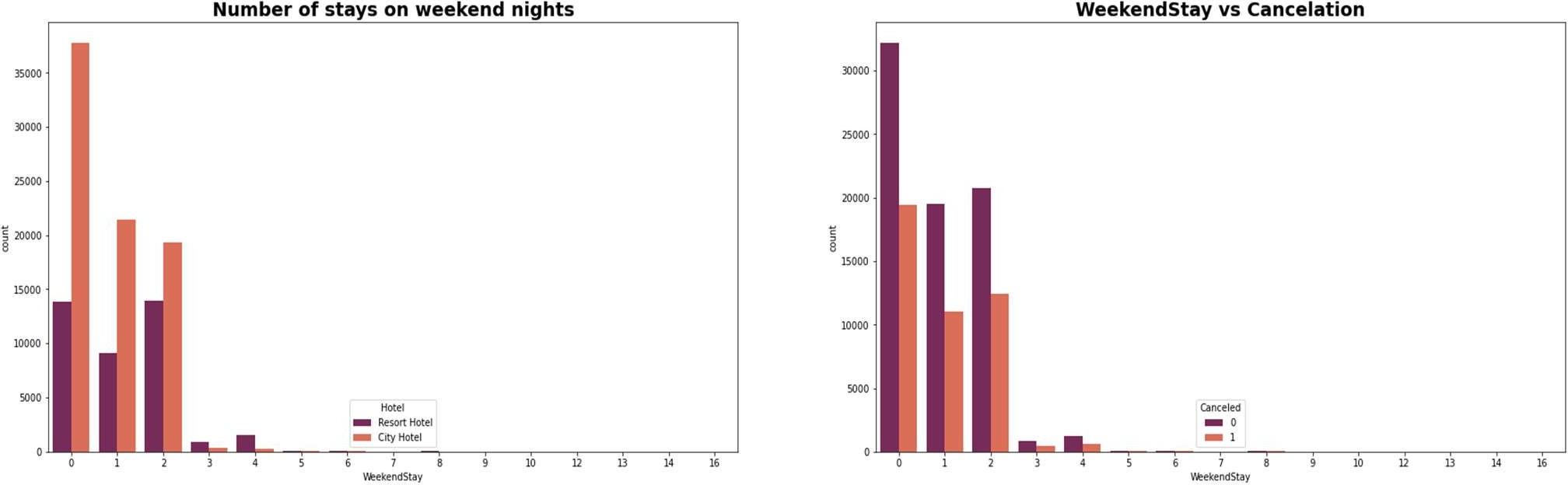
August is the most occupied (busiest) month with 11.6% bookings and January is the most unoccupied month with 4.9% bookings.



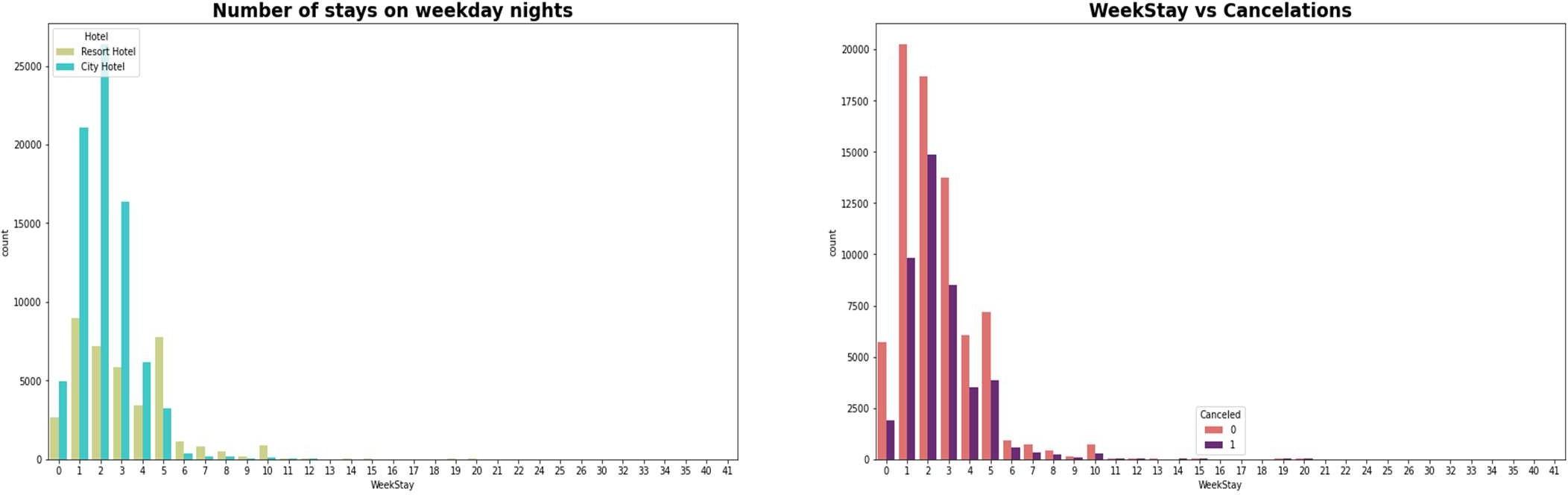
* + Month end day has very less arrivals
  + Bookings are more in City hotel



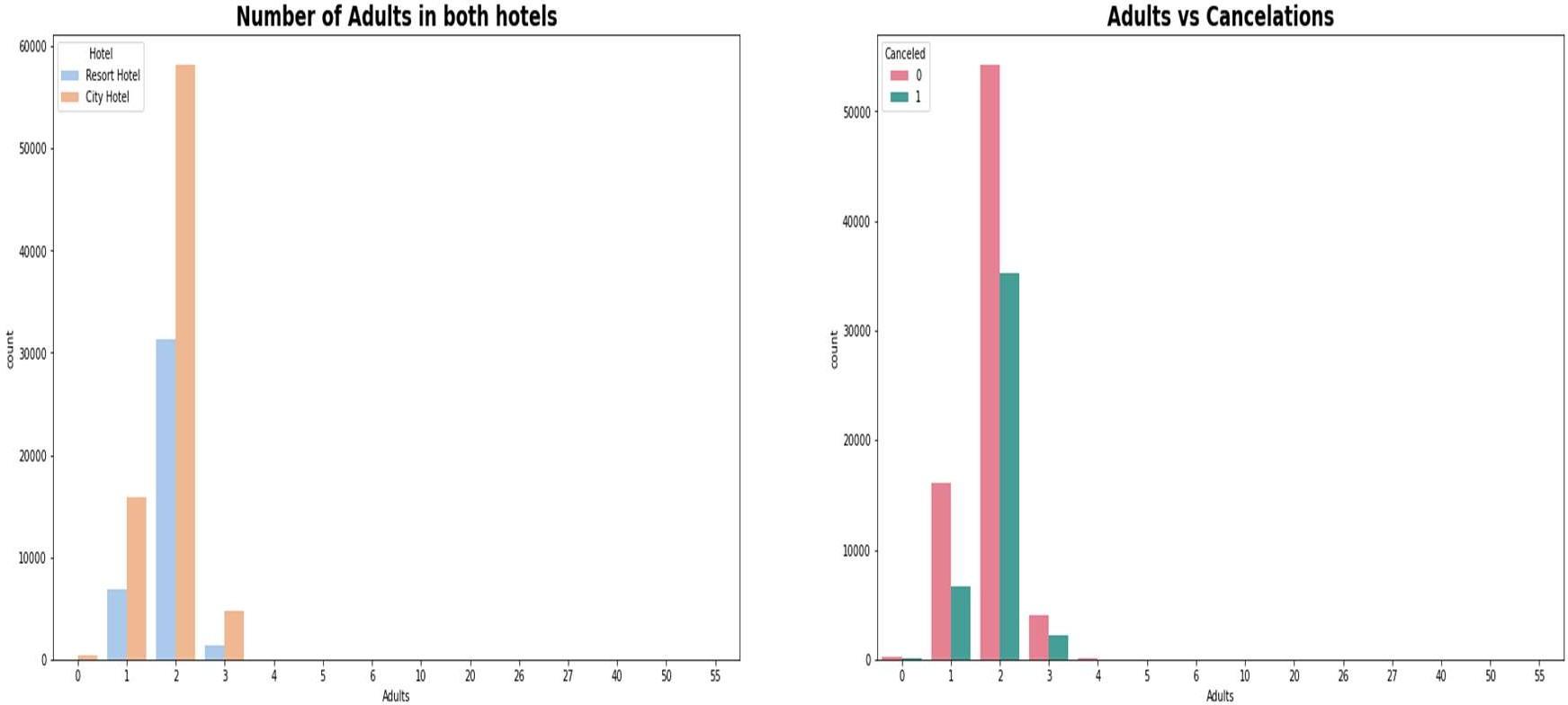
* In the first graph we can see that most of the weekend nights were booked in City Hotel
* Second plot shows most of weekend nights which were booked were not canceled

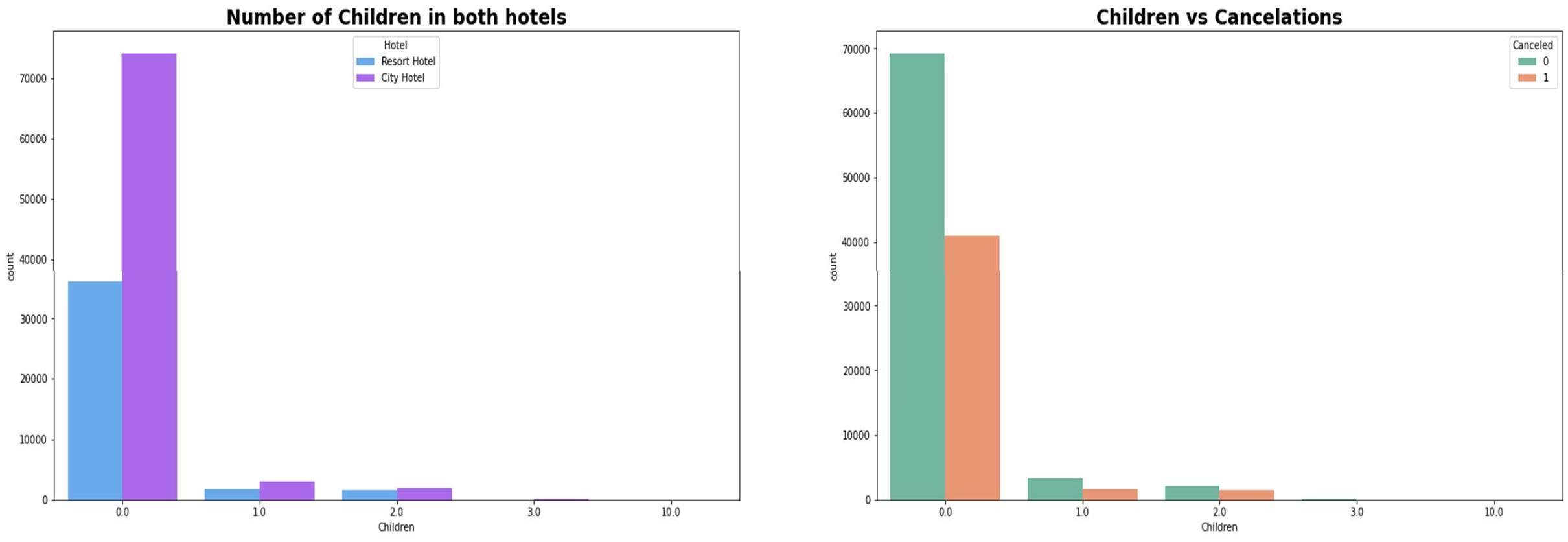


* + Weekday night stays were more in City Hotel
  + Less cancelations were observed

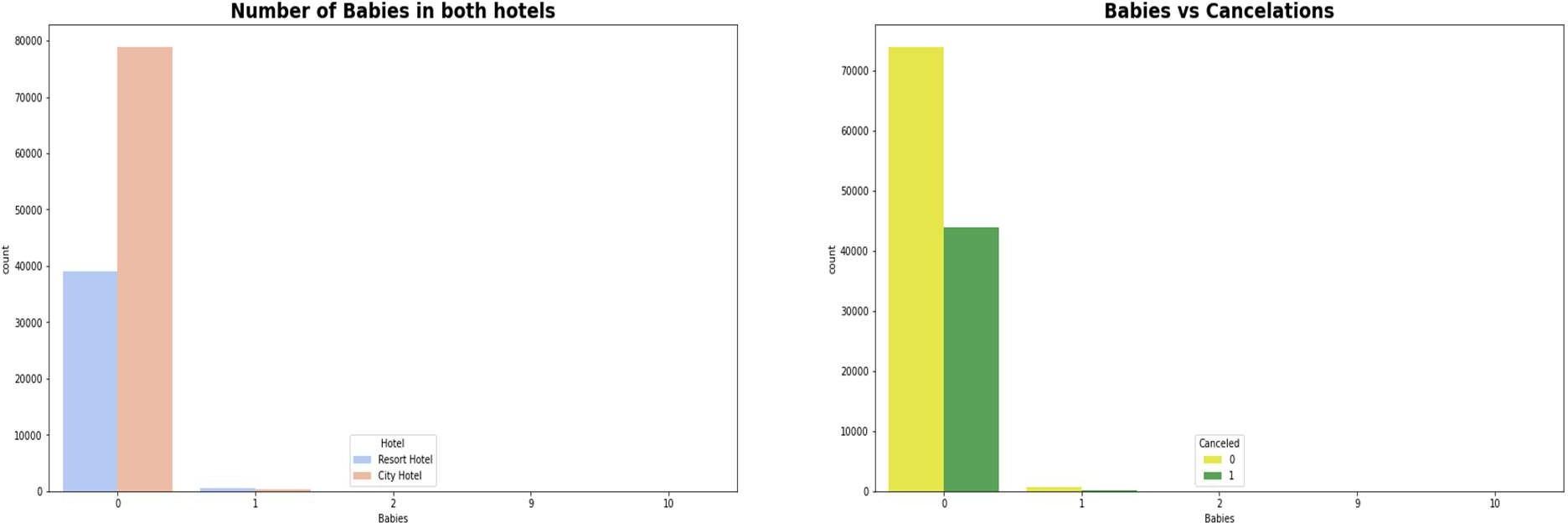


Adults who were 2 in number are more and preferred city hotel rather than resort hotel, infarct more than half the visitors even canceled the bookings



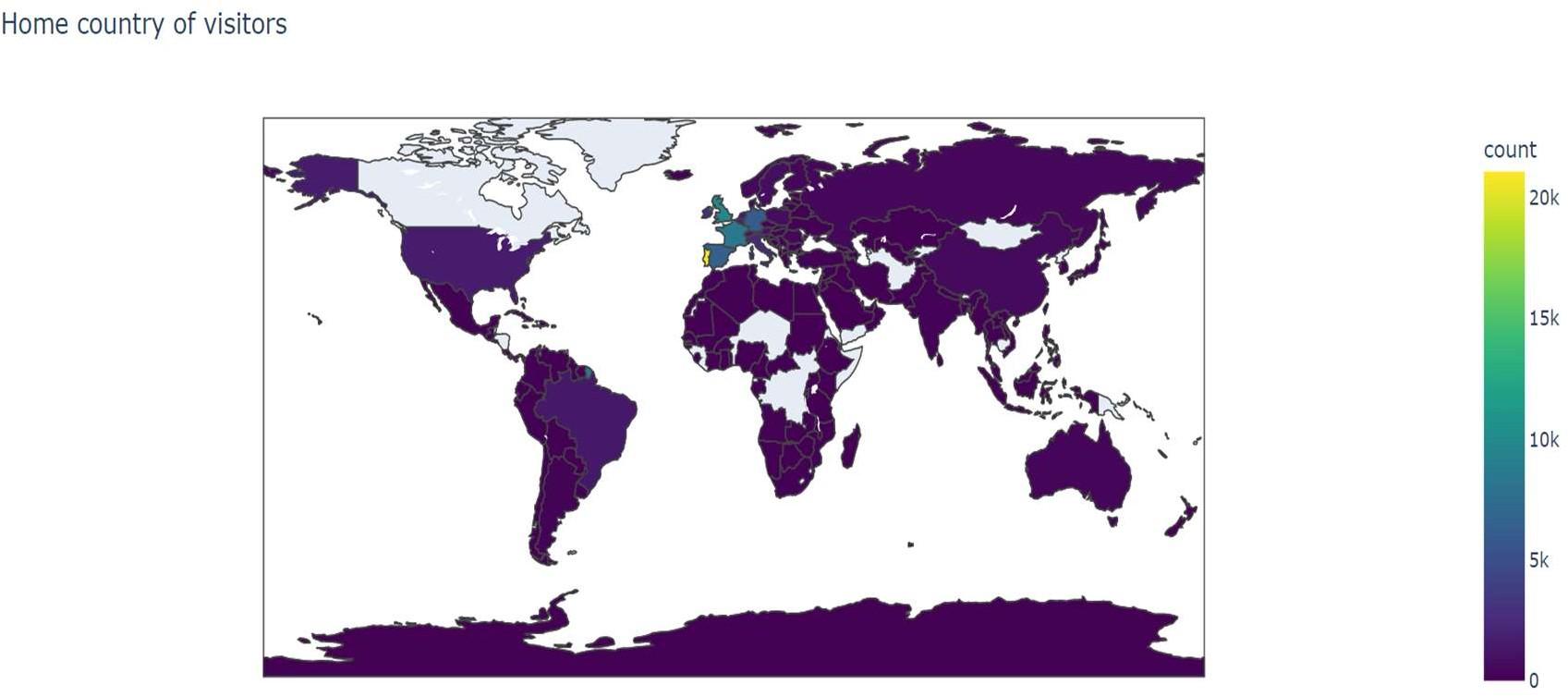
 

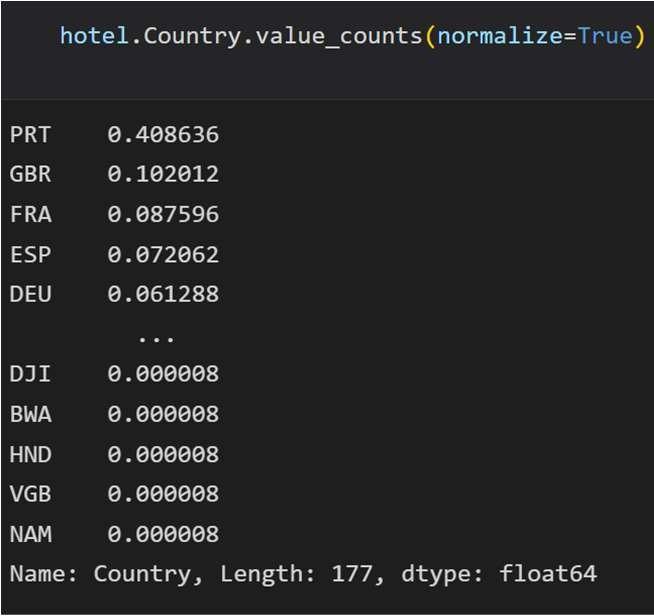
 



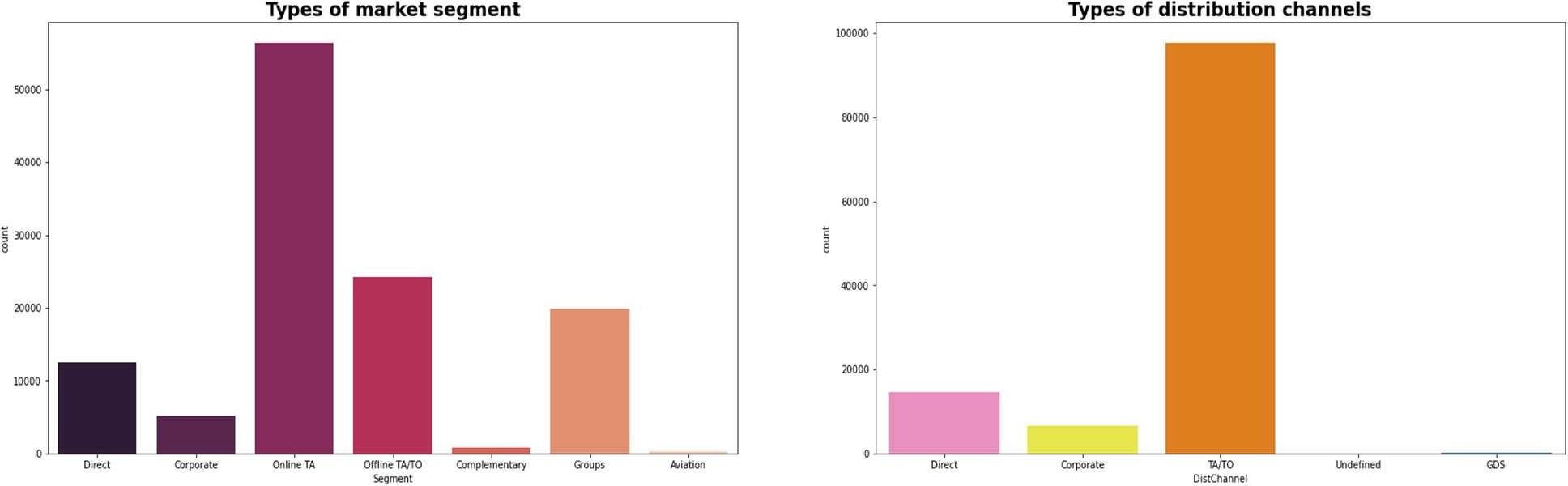
* Most visitors were arrived in pair with no children/ Babies and preferred City hotel over resort hotel
* visitors who had 1 or 2 children also preferred city hotel



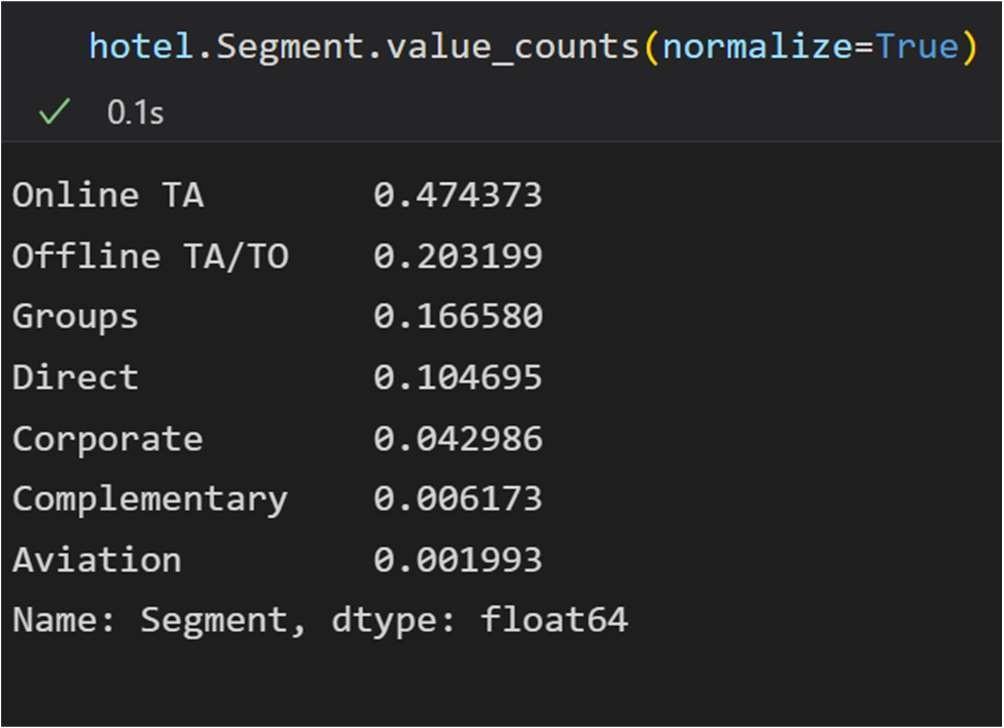
Most of the visitors are from western Europe, namely France, UK and Portugal being the highest.

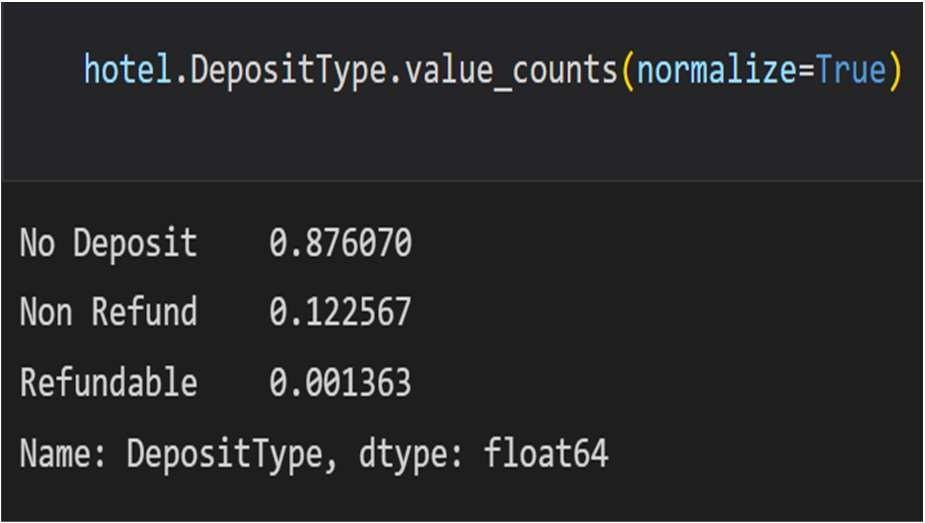
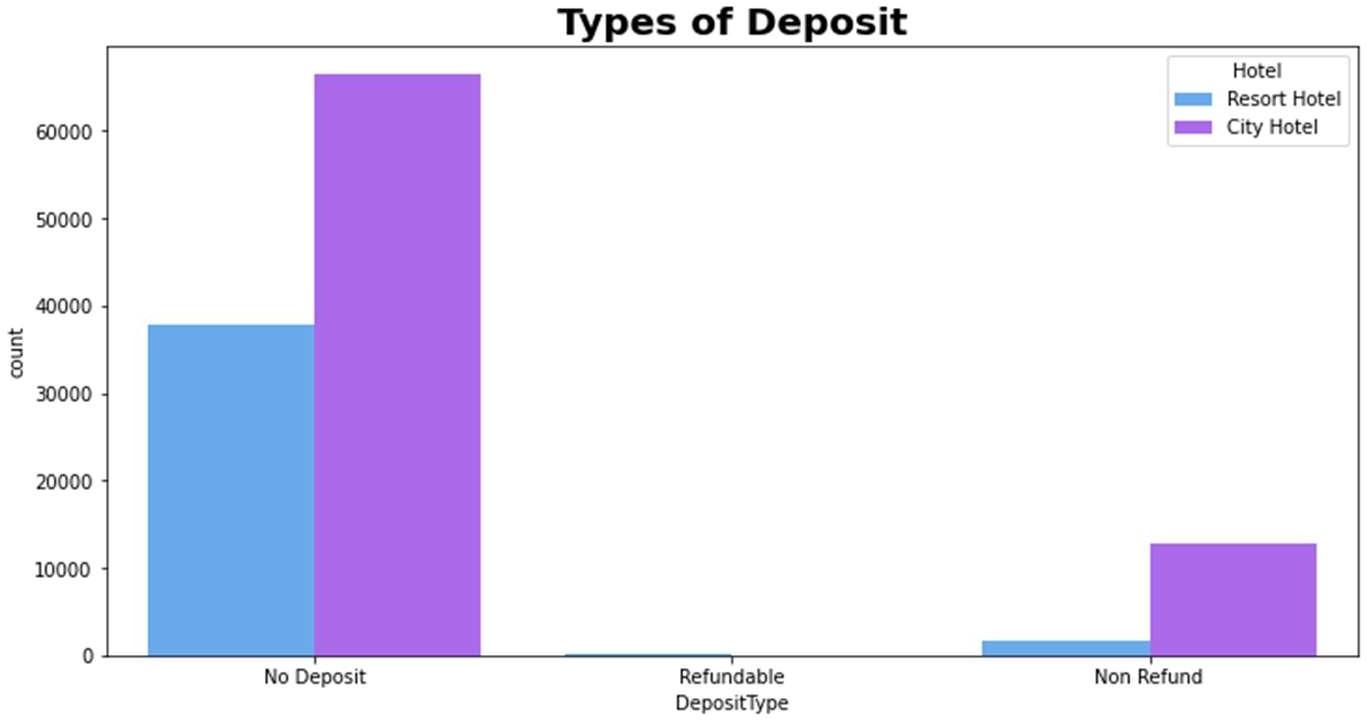


* Majority Distribution channels and Market segment were Travel agencies wither offline/online. So better focus more on this
* Here we can see that the most of guest are making reservation through TA/TO channels which is travel agency and tour operator.
* Than the second most used channel is direct.
* Channel which is mostly used for early booking of hotels is also TA/TO.

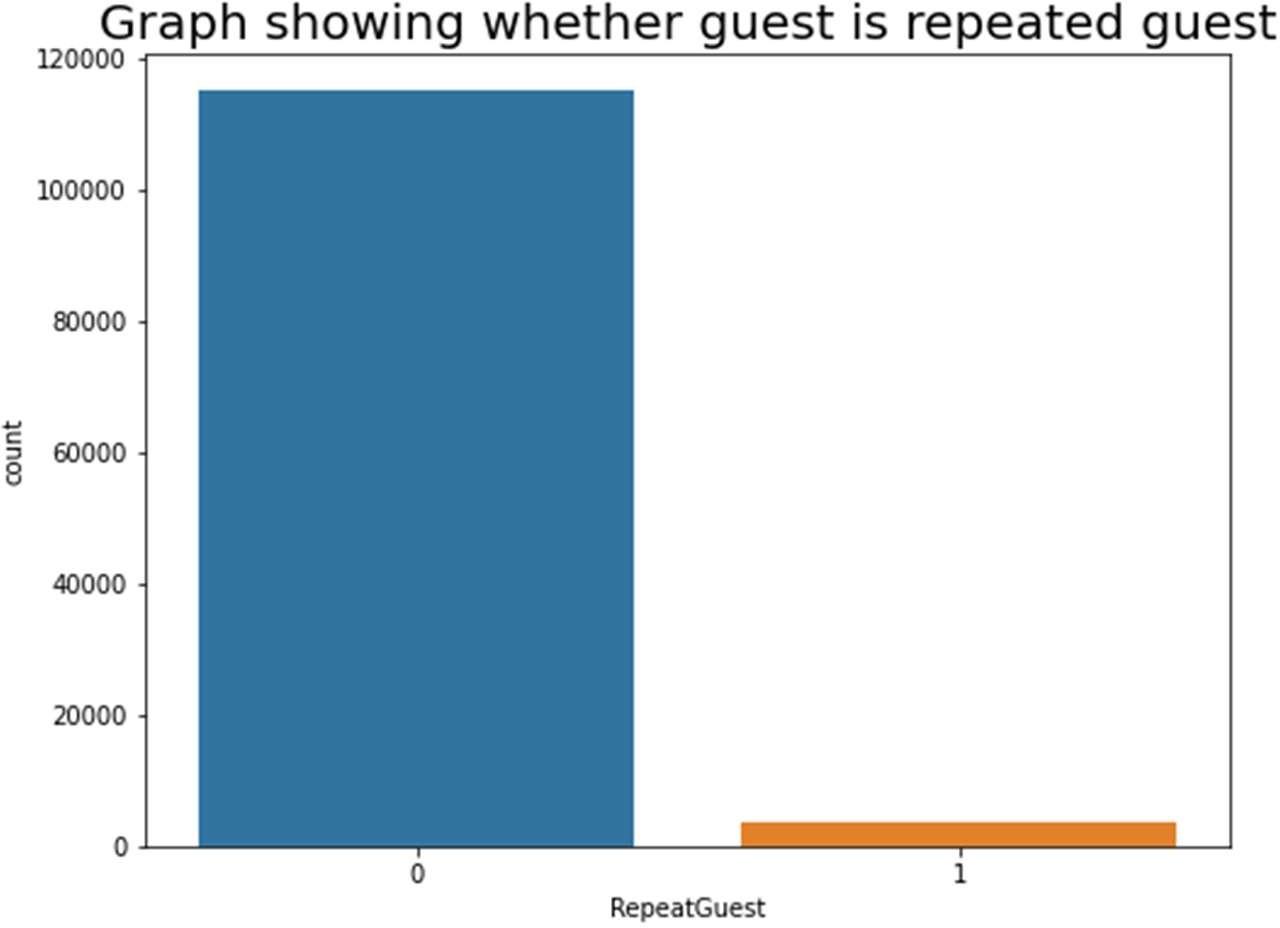
Around 47% of bookings are made via Online Travel Agents, almost 20% of bookings are made via Offline Travel Agents and less than 20% are Direct bookings without any other agents.



87.6% of the bookings are done without any Deposit12.2% of the bookings are done with Non- Refundable Deposit and 1.3% are Refundable Deposits

* Low number of repeated guests.
* A need to target repeated guests since they have booked before.
* Both hotels have very small percentage that customer will repeat, but Resort hotel has slightly higher repeat % than City Hotel.

# Price per Month for Resort Hotel and City Hotel

Prices of Resort Hotel are much higher and fluctuates timely during the months.

Prices of City Hotel do not fluctuate that much.

# ADR: Average Daily Rate

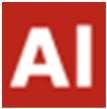
For Resort Hotel, ADR is more expensive during July, August & September and for City Hotel, ADR is slightly more during March, April & May.

* Firstly, higher lead time has higher chance of cancellation. Also, history of previous cancellations increases chances of cancellation.
* Secondly, The City hotel has more guests during spring and autumn, when the prices are also highest, In July and August there are less visitors, although prices are lower. Thus, customers can get good deal on bookings in July and August in city hotel.
* Guest numbers for the Resort hotel go down slightly from June to September, which is also when the prices are highest. Thus, these months should be avoided for bookings.
* Thirdly, May to August is the peak season of bookings. Both hotels have the fewest guests during the winter.
* Fourthly, No deposit cancellations are high compared to other categories but these should not be discouraged per se as bookings in this category are also very high compared to non refundable type bookings.

# Data Summary

* Majority of the hotels booked are city hotel. Definitely need to spend the most targeting fund on those hotel.
* We also realise 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 area.
* Given that we do not have repeated guests, we should target our advertisement on guests to increase returning guests.
* Set Non-refundable Rates, Collect deposits, and implement more rigid cancellation policies.
* Encourage Direct bookings by offering special discounts
* Monitor where the cancellations are coming from such as Market Segment, distribution channels, etc.



**Thank You**