Capstone Project-I

Hotel Booking Analysis

Team Members

BANKA KAVYA SREE
SHANTANU PAWAR
MOHAMMAD KASHIF

Agenda

Overview of Capstone Project- Hotel Booking Analysis.

- About the Project
- Life Cycle
- Data-Set Explanation
- EDA
- Data Visualization
- Summary
- Q & A

About the Project

Hotel Booking Analysis

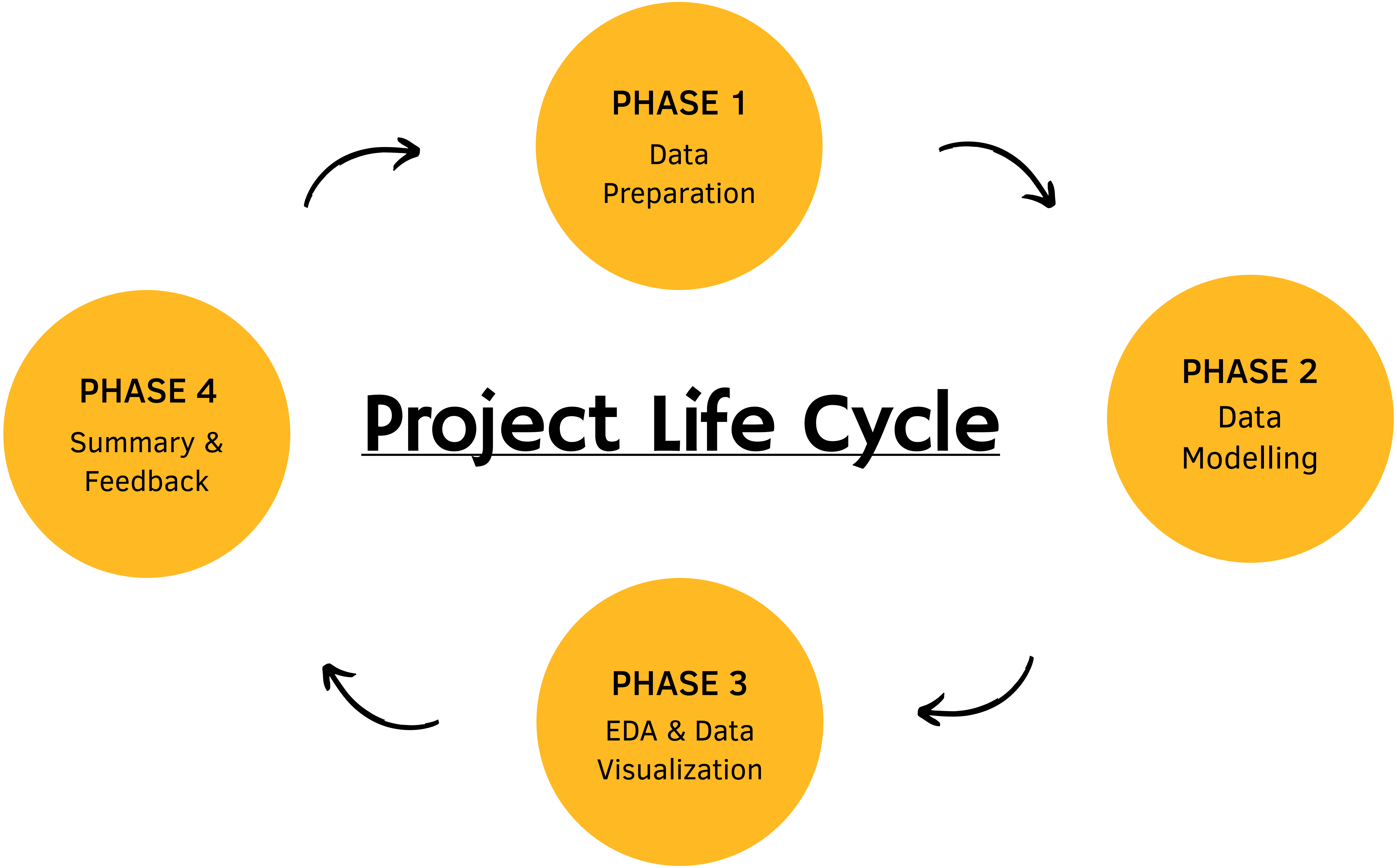
OBJECTIVES:

- There is hotel booking dataset which contains booking information for a city hotel and a resort hotel
- Explore and analyze the data to discover important factors that govern the bookings.

GOALS:

- Data Preparation & Cleaning - For efficient & accurate analysis.
- Exploratory Data Analysis(EDA) & Data Visualization
- To find the revenue for the hotel.
- To find out maximum bookings according to months, year and types of channel.
- To find out highest cancellation %
- To find most preferred Hotel, i.e 'City Hotel' or 'Resort Hotel'.
- Many more.....

Project Life Cycle



PHASE 1

Data Preparation

PHASE 2

Data Modelling

PHASE 3

EDA & Data Visualization

PHASE 4

Summary & Feedback

About the Data-Set

We will be the first look out our data set for further collaboration feature.

INTRODUCTION:

- This data set (hotel_booking.csv) contains booking information for a city hotel and a resort hotel. It includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of required parking spaces, among other things. All personally identifying information has been removed from the data.
- This data set consist of 119390 Rows & 32 Columns.
- There are so many null values are present in data set, which will affect our analysis, so for better results we must have to clean our data set or remove all null & duplicates value from it.

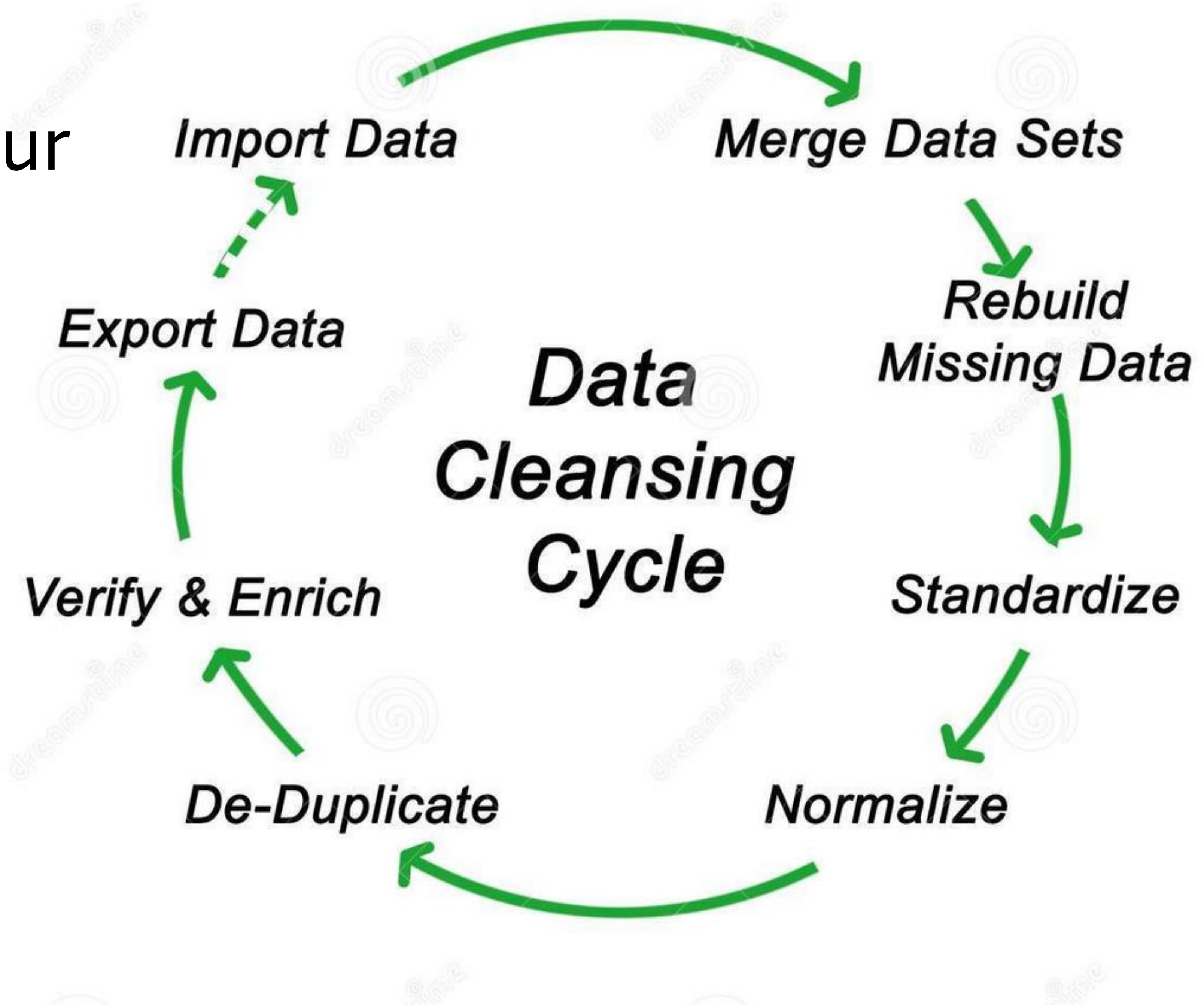
Data Cleaning

DATA CLEANING:

Data cleaning means fixing bad data in your data set.

Bad data could be:

- Empty cells or null values.
- Data in wrong format.
- Wrong data.
- Duplicates.

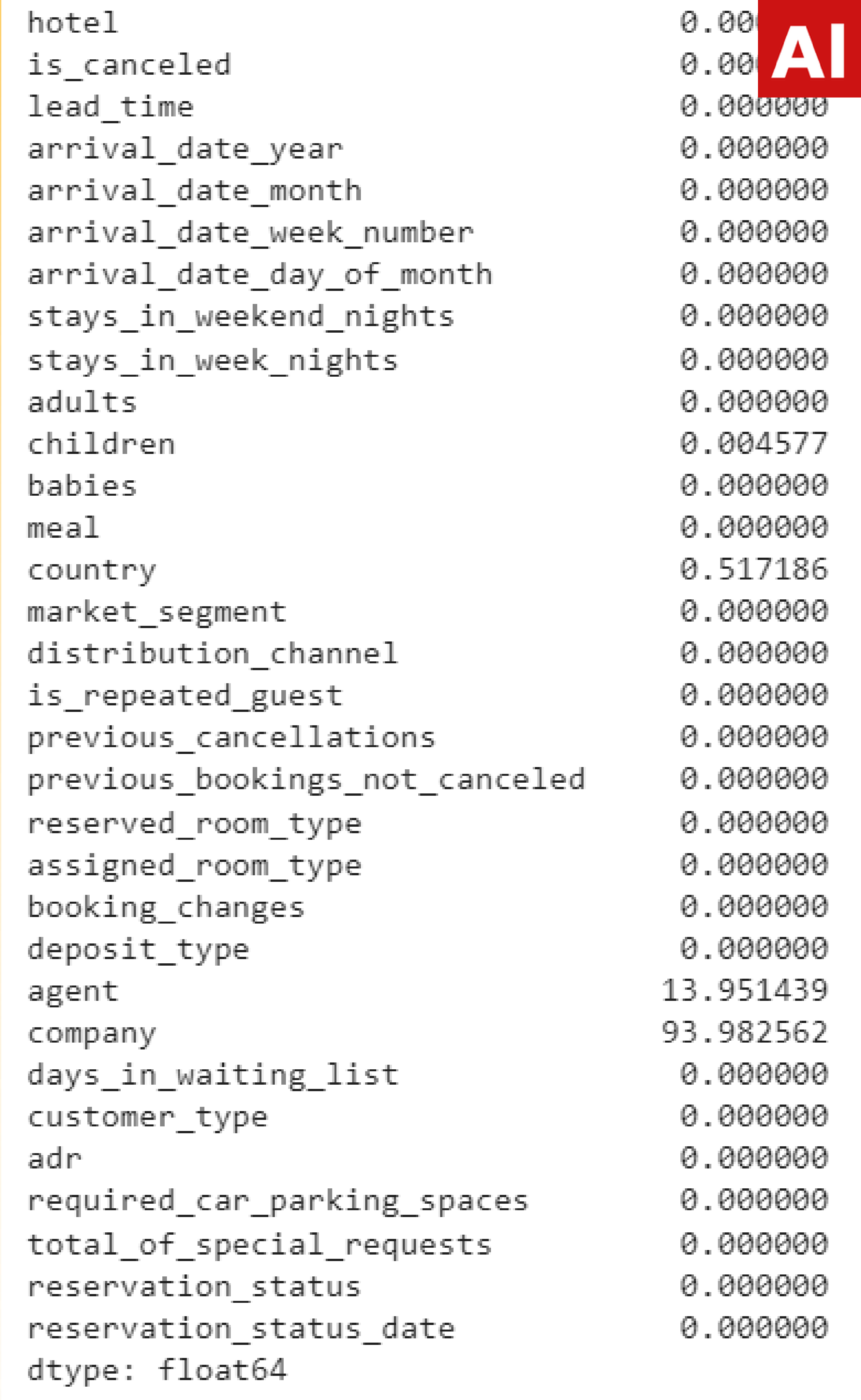


Data Cleaning(Contd..)

We will be the first clean all the null values & duplicates from data set.

OBJECTIVES:

- As we can see, there is about 94% of the data null in the column 'Company'. This is a large percentage of the data and thus cannot be used for analysis and visualization. So, we drop the column ‘Company’.

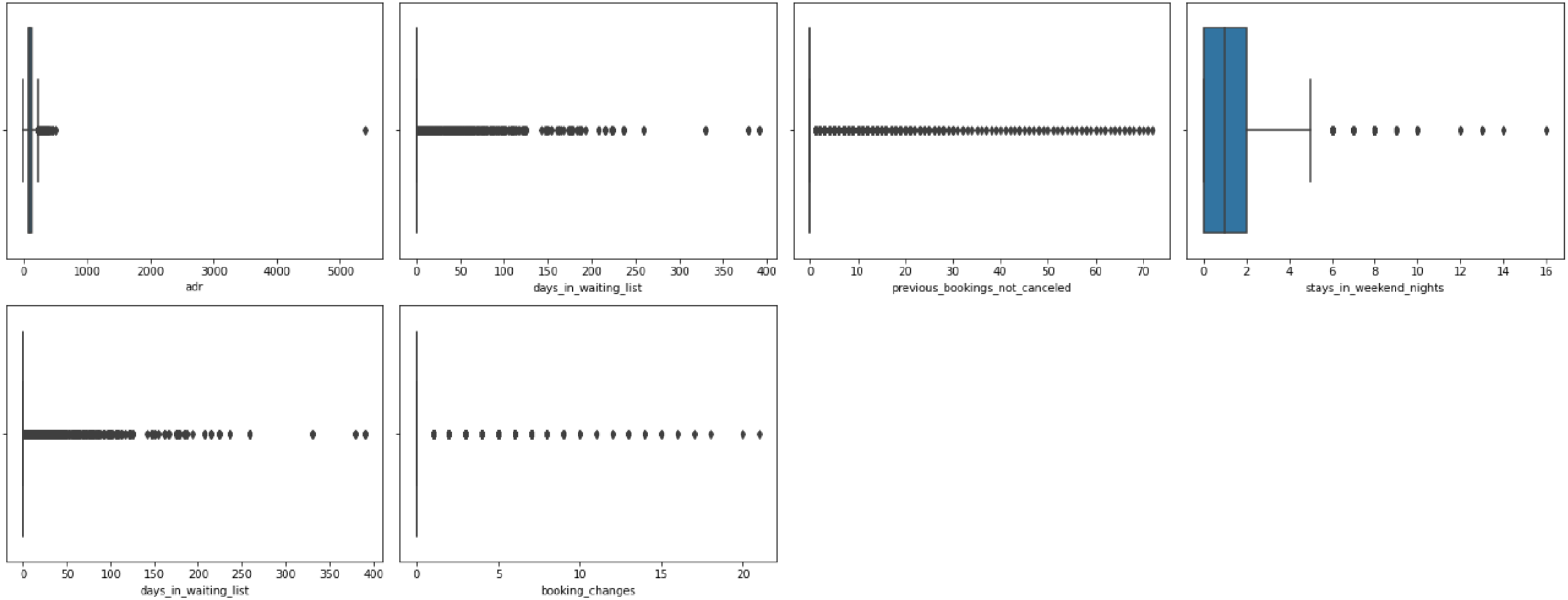


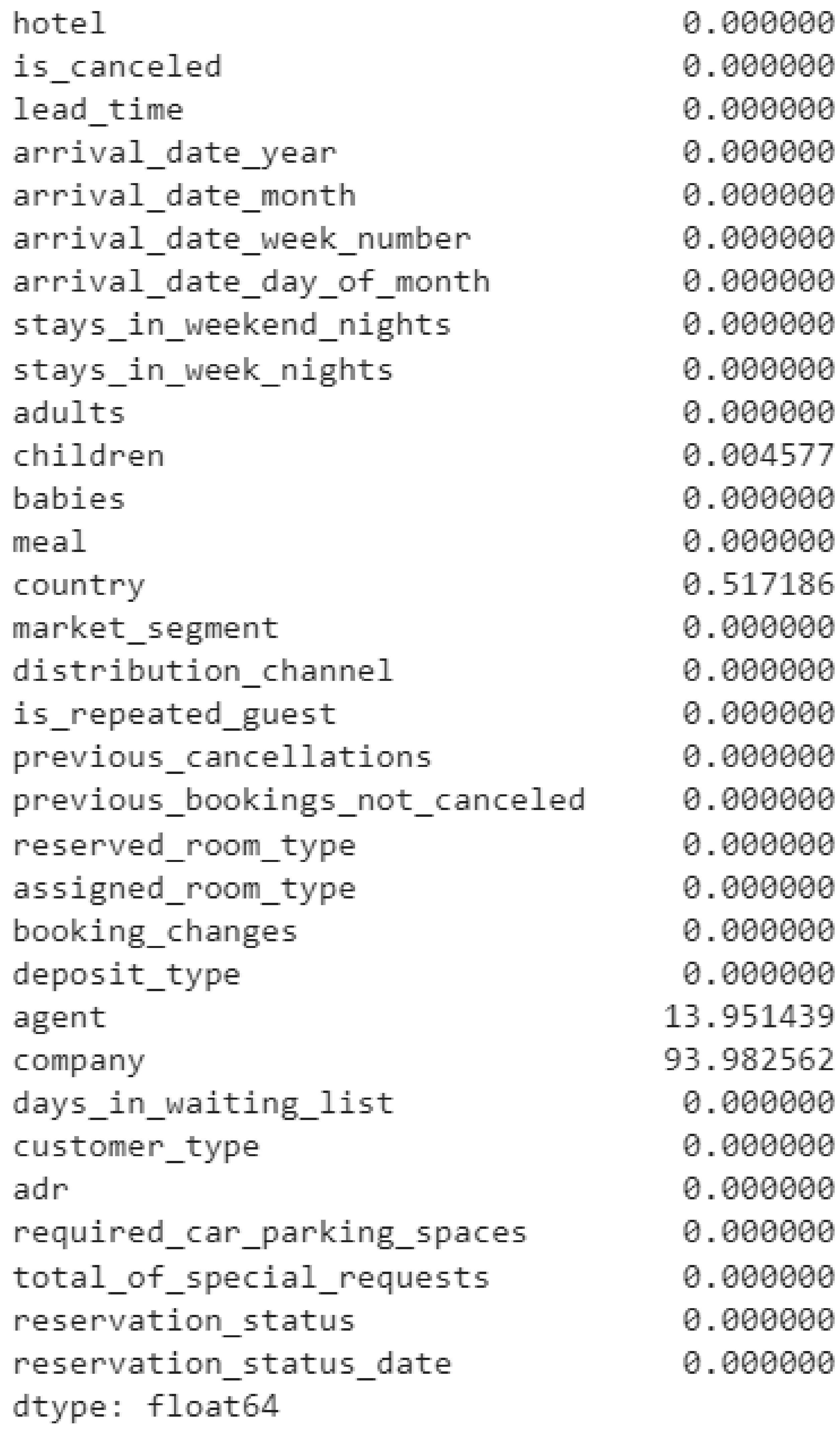
```
hotel                              0.00
is_canceled                        0.00
lead_time                          0.000000
arrival_date_year                  0.000000
arrival_date_month                 0.000000
arrival_date_week_number           0.000000
arrival_date_day_of_month          0.000000
stays_in_weekend_nights            0.000000
stays_in_week_nights               0.000000
adults                             0.000000
children                           0.004577
babies                             0.000000
meal                               0.000000
country                            0.517186
market_segment                     0.000000
distribution_channel               0.000000
is_repeated_guest                  0.000000
previous_cancellations             0.000000
previous_bookings_not_canceled     0.000000
reserved_room_type                 0.000000
assigned_room_type                 0.000000
booking_changes                    0.000000
deposit_type                       0.000000
agent                             13.951439
company                           93.982562
days_in_waiting_list               0.000000
customer_type                      0.000000
adr                                0.000000
required_car_parking_spaces        0.000000
total_of_special_requests          0.000000
reservation_status                 0.000000
reservation_status_date            0.000000
dtype: float64
```

Data Cleaning(Contd..)

Outliers

An outlier is an extremely high or extremely low value in the dataset. It could affect our analysis. In our data set, there is an outliers present as shown in figure. So, as we can see ‘ADR' outlier could affect our analysis, so we will remove the outliers from our data set.

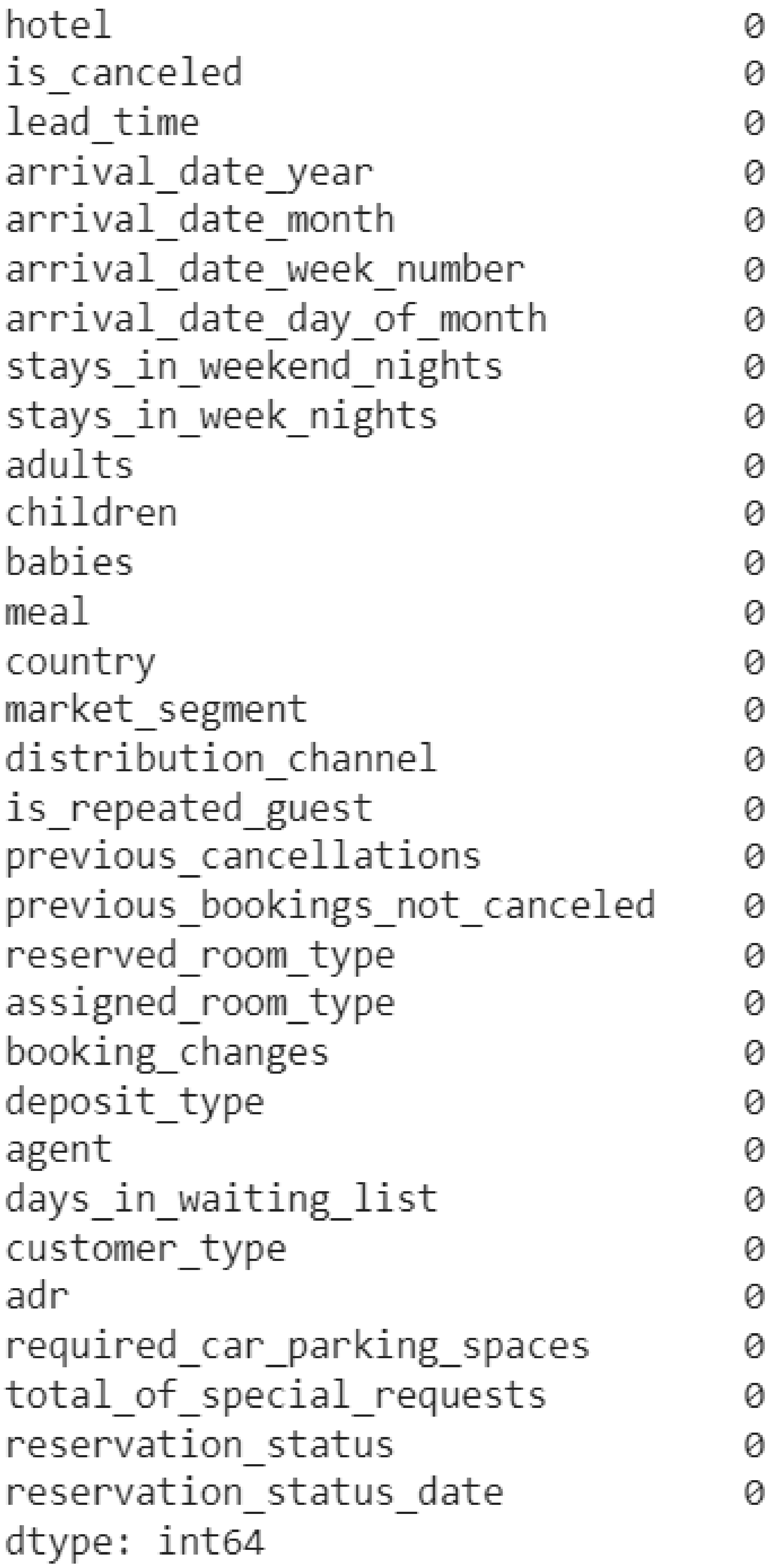




```
hotel                              0.000000
is_canceled                        0.000000
lead_time                          0.000000
arrival_date_year                  0.000000
arrival_date_month                 0.000000
arrival_date_week_number           0.000000
arrival_date_day_of_month          0.000000
stays_in_weekend_nights            0.000000
stays_in_week_nights               0.000000
adults                             0.000000
children                           0.004577
babies                             0.000000
meal                               0.000000
country                            0.517186
market_segment                     0.000000
distribution_channel               0.000000
is_repeated_guest                  0.000000
previous_cancellations             0.000000
previous_bookings_not_canceled     0.000000
reserved_room_type                 0.000000
assigned_room_type                 0.000000
booking_changes                    0.000000
deposit_type                       0.000000
agent                             13.951439
company                           93.982562
days_in_waiting_list               0.000000
customer_type                      0.000000
adr                                0.000000
required_car_parking_spaces        0.000000
total_of_special_requests          0.000000
reservation_status                 0.000000
reservation_status_date            0.000000
dtype: float64
```

Data Cleaning(Contd..)

After cleaning of data by using Pandas library methods we can see , there are no more null values present in our data set.



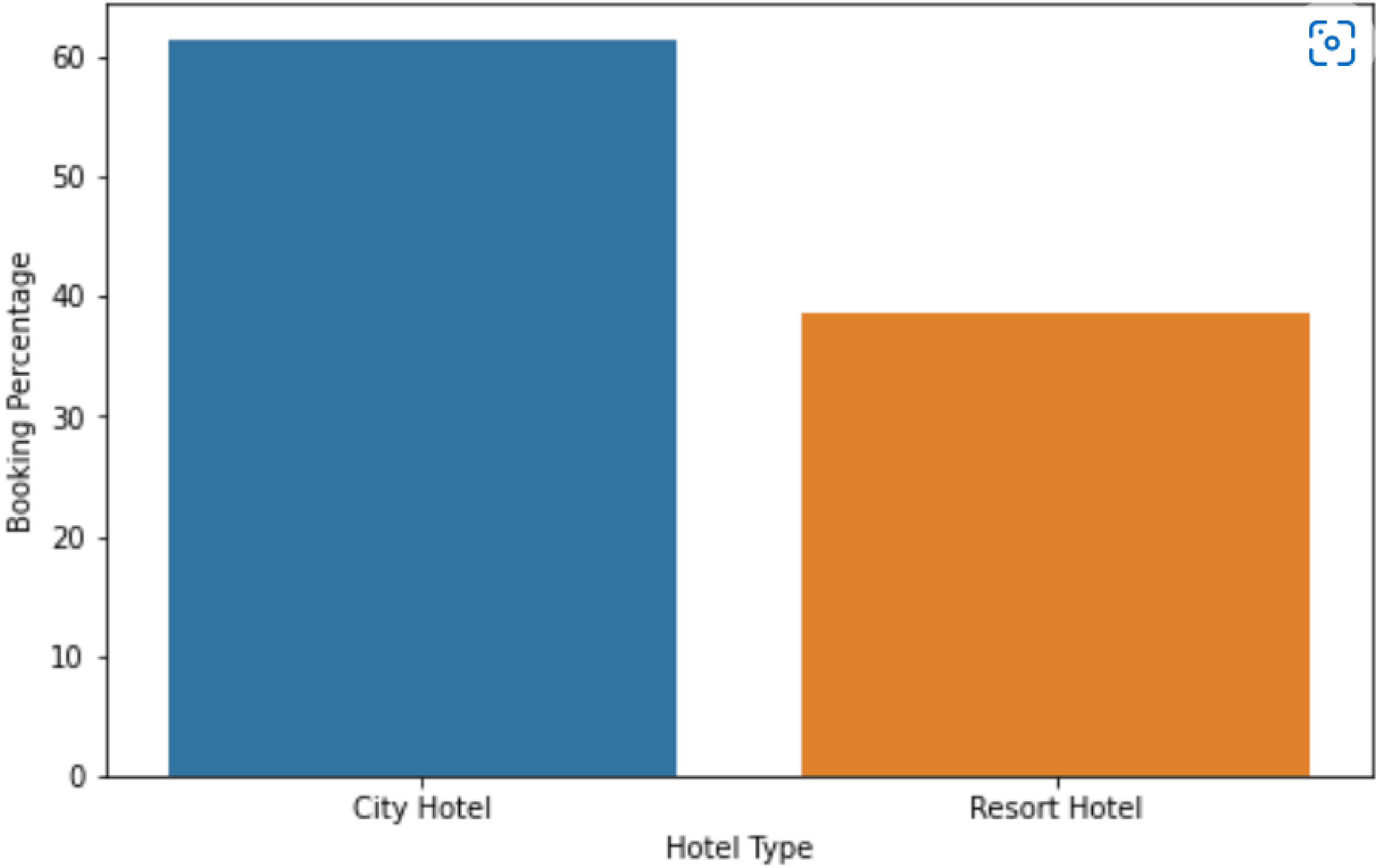
```
hotel                             0
is_canceled                       0
lead_time                         0
arrival_date_year                 0
arrival_date_month                0
arrival_date_week_number          0
arrival_date_day_of_month         0
stays_in_weekend_nights           0
stays_in_week_nights              0
adults                            0
children                          0
babies                            0
meal                              0
country                           0
market_segment                    0
distribution_channel              0
is_repeated_guest                 0
previous_cancellations            0
previous_bookings_not_canceled    0
reserved_room_type                0
assigned_room_type                0
booking_changes                   0
deposit_type                      0
agent                             0
days_in_waiting_list              0
customer_type                     0
adr                               0
required_car_parking_spaces       0
total_of_special_requests         0
reservation_status                0
reservation_status_date           0
dtype: int64
```

EDA & Data Visualization

- **Which hotel has highest percentage of guests?**

City Hotel has 61.44 percentage of guests and Resort hotel has 38.55 percentage of guests. More guests showed interest to reside on City Hotel.

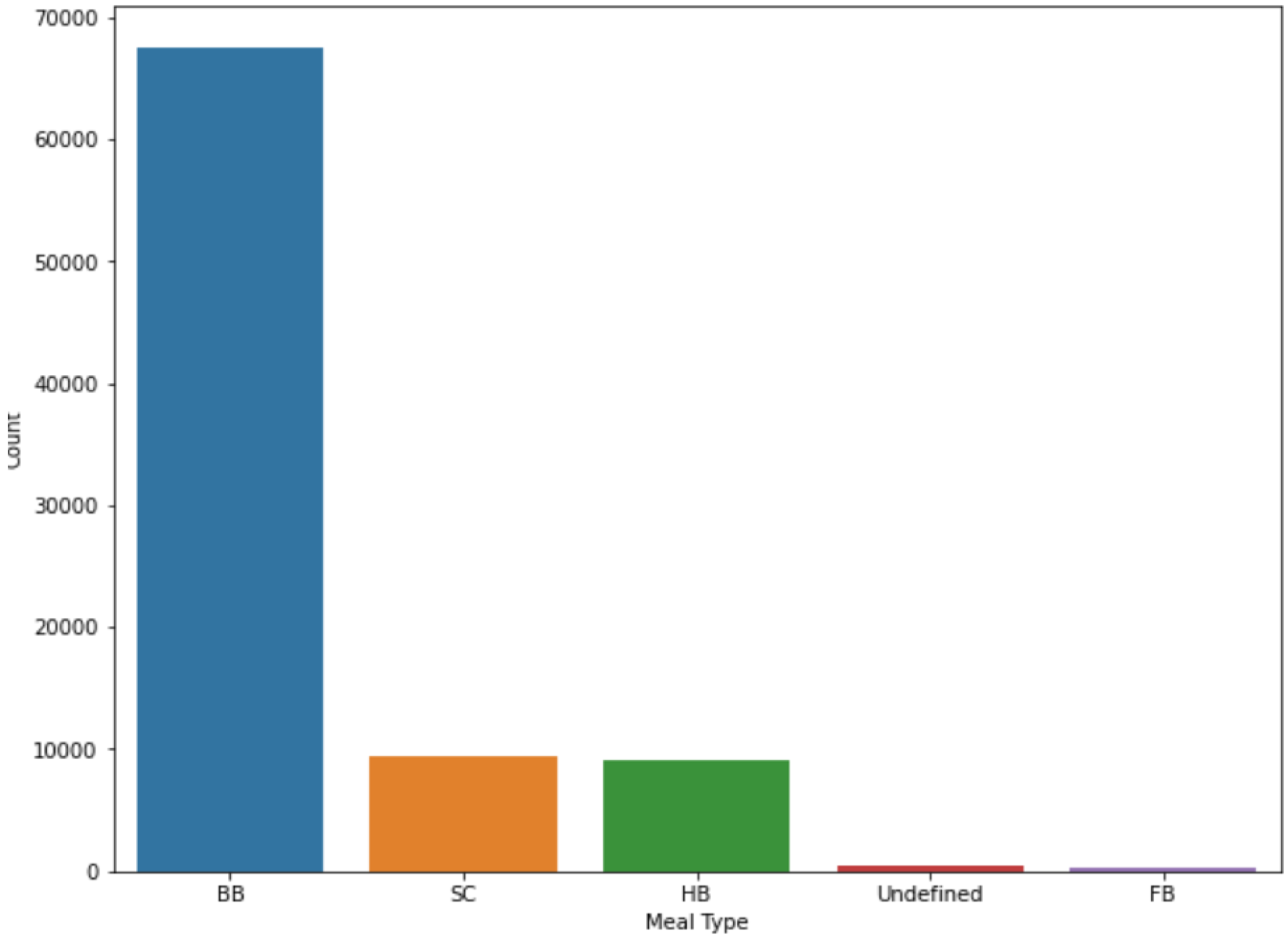
```
     Hotel Type  Booking Percentage
0    City Hotel           61.441931
1  Resort Hotel           38.558069
```



EDA & Data Visualization(contd..)

- Which type of meal is ordered the most?

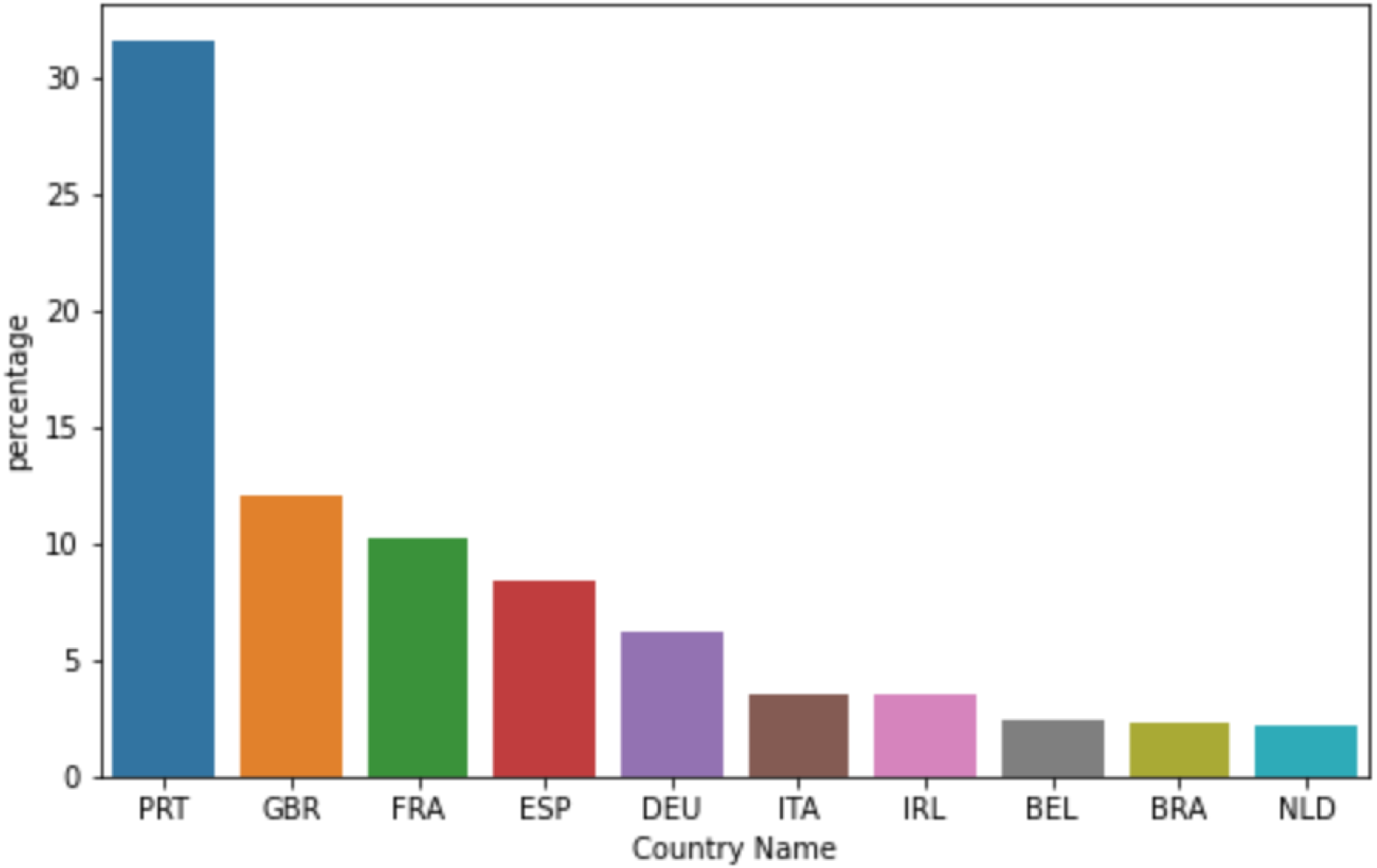
According to analysis, BB Type of meal was ordered by most of the guests.



EDA & Data Visualization(continued..)

- **Which Country has the most number of guests?**

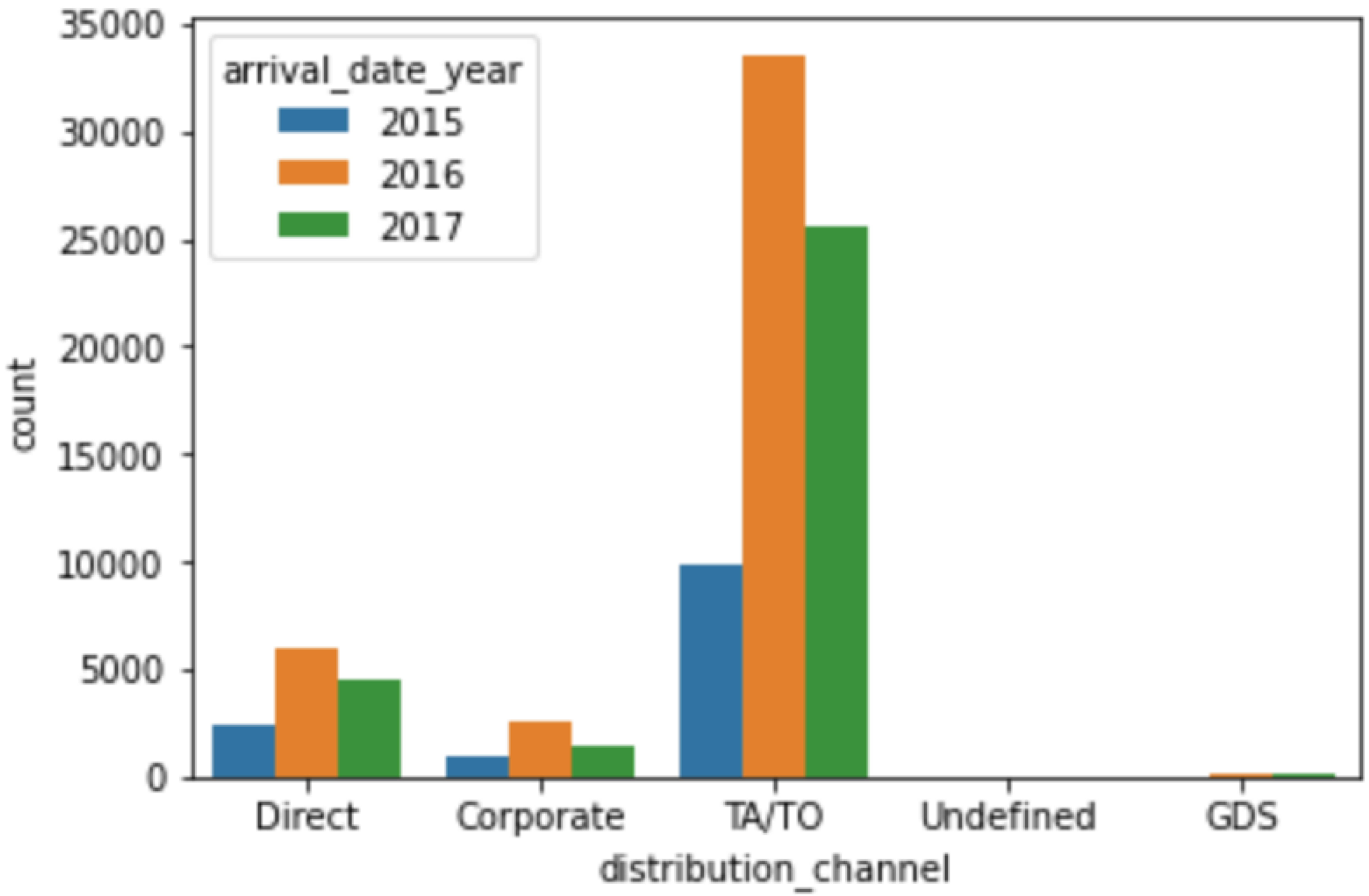
Portugal has the most guests, followed by Great Britain.



EDA & Data Visualization(continued..)

- Which Distribution Channel is most common for hotel bookings, year wise?

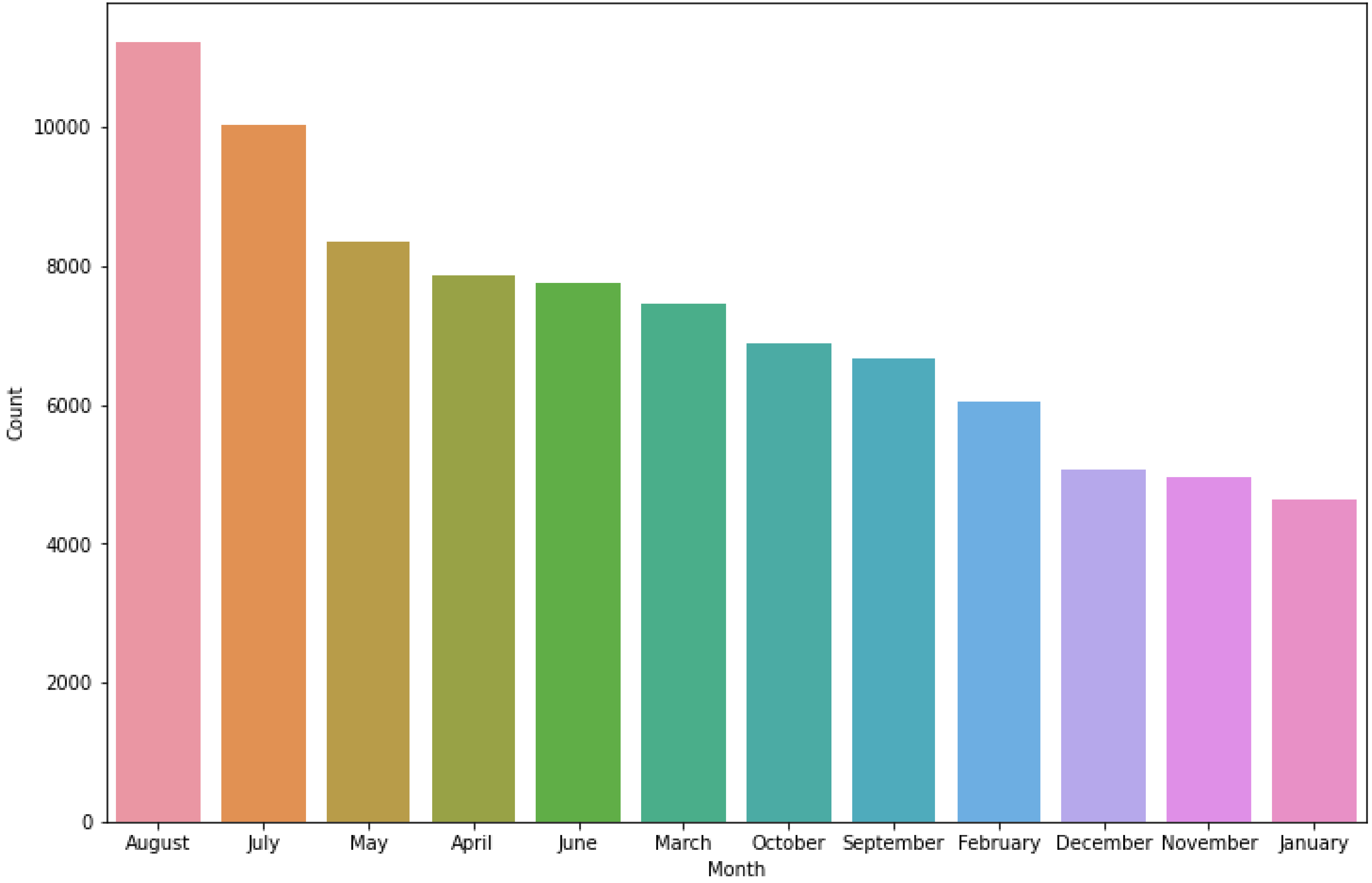
We can see that TA/TO is most used channel for hotel booking in all the years.



- Monthly booking analysis for hotels & in which month most of the bookings are happened?

As we can see maximum bookings were made in the month of August. And the least bookings were made at the start of the year.

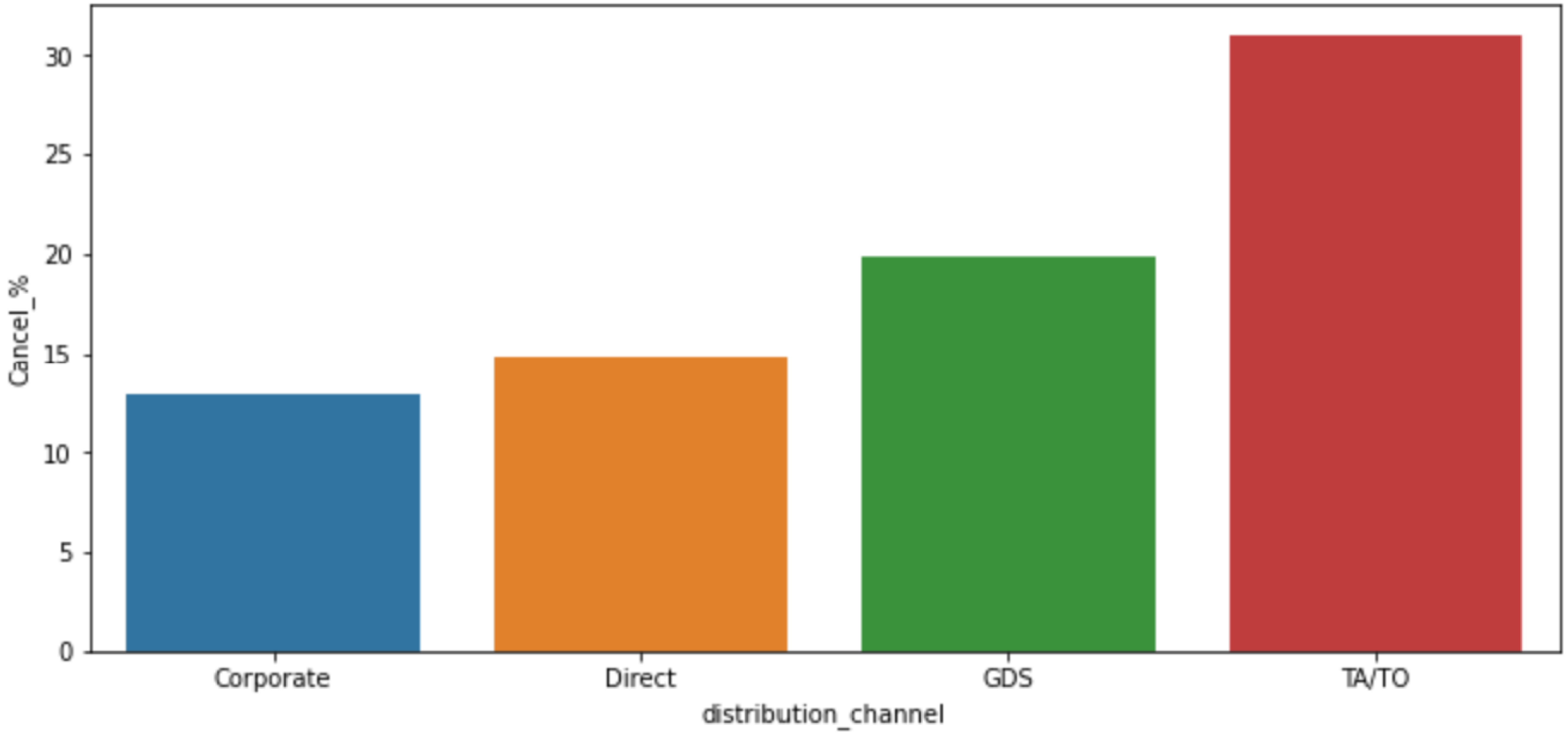
EDA & Data Visualization(continued..)



- Which has highest cancellation percentage according to distribution channel?

Hence, we can see here that highest cancellation percentage is for TA/TO which is around 30% cancellation percentage.

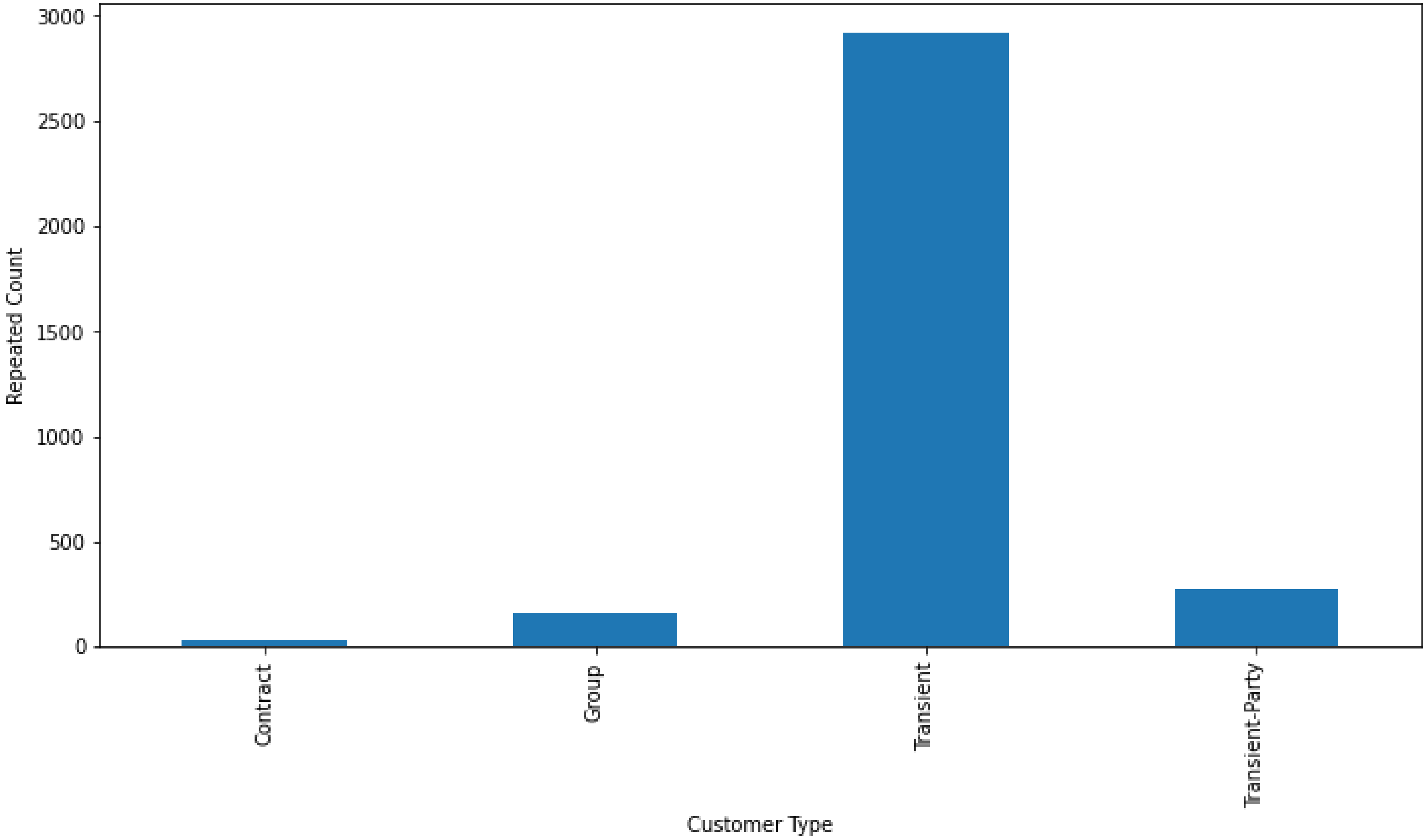
EDA & Data Visualization(continued..)



- Type of customers which are mostly repeated in each hotel?

As we can see here that the maximum number of repeated customers are the "Transient type" i.e., the "Short-time customers" are mostly repeated.

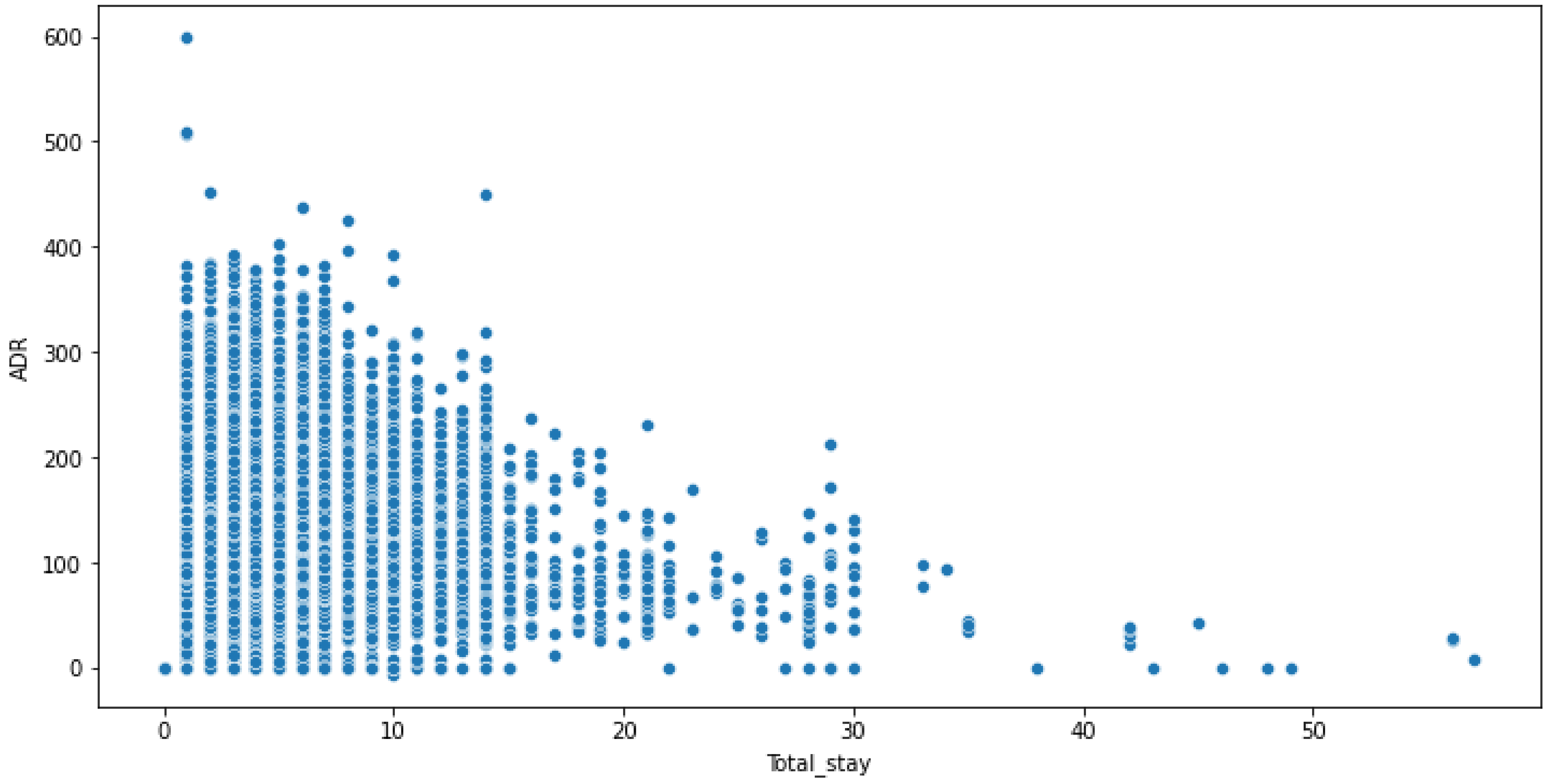
EDA & Data Visualization(continued..)



- What is the optimal stay duration for best daily rate?

As the duration increases the ADR slightly decreases and with the duration more than 2 weeks a noticeable difference can be seen. Thus we can say that the optimal duration of stay for better Average daily rate is roughly around 2 weeks.

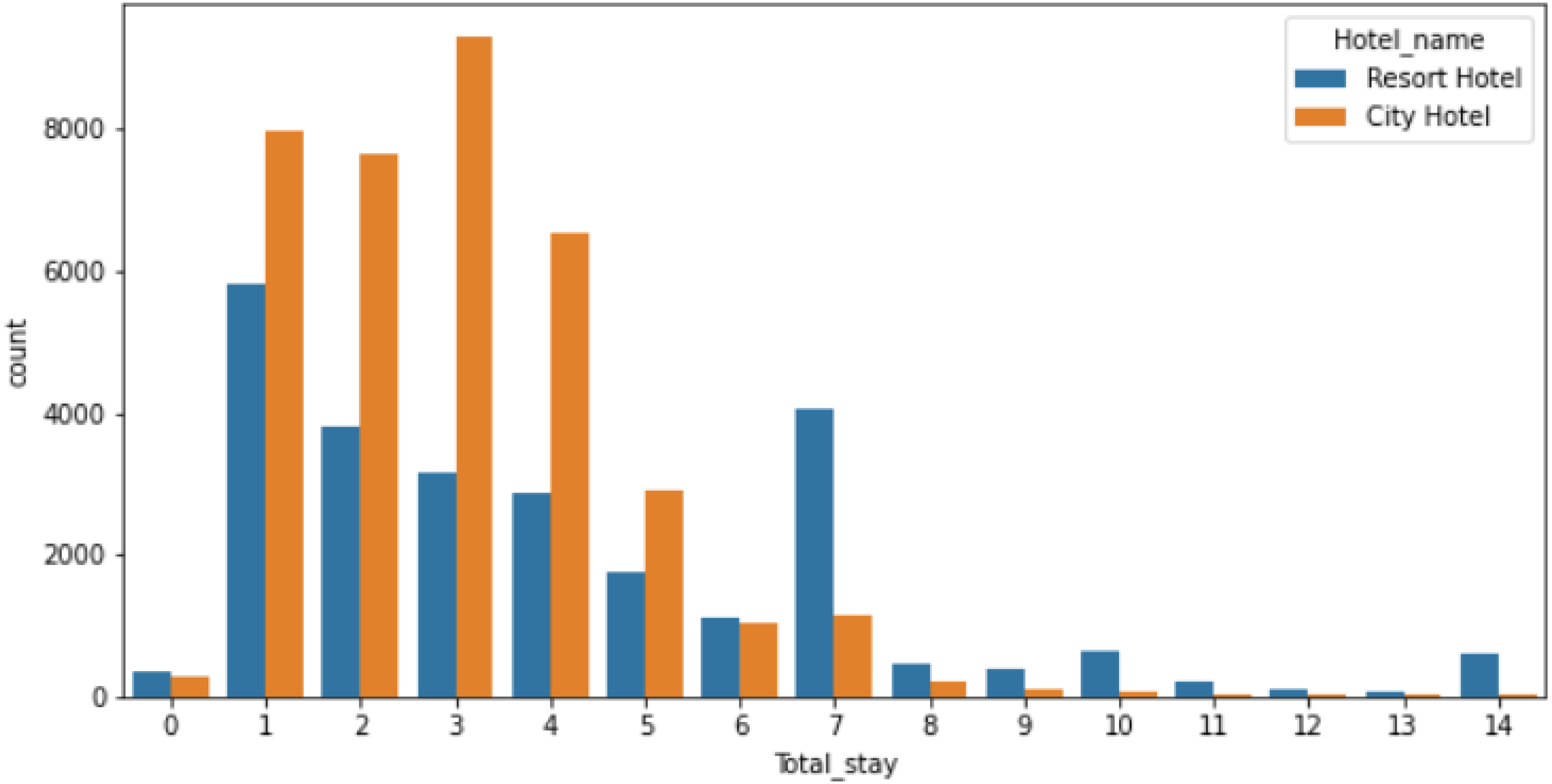
EDA & Data Visualization(continued..)



- What is the preferred duration of stay for the guests?

The preferred duration of stay for most of the guests is 1-4 days.
We can also see that after a period of 7 days the bookings for Resort hotel are more than City hotel.

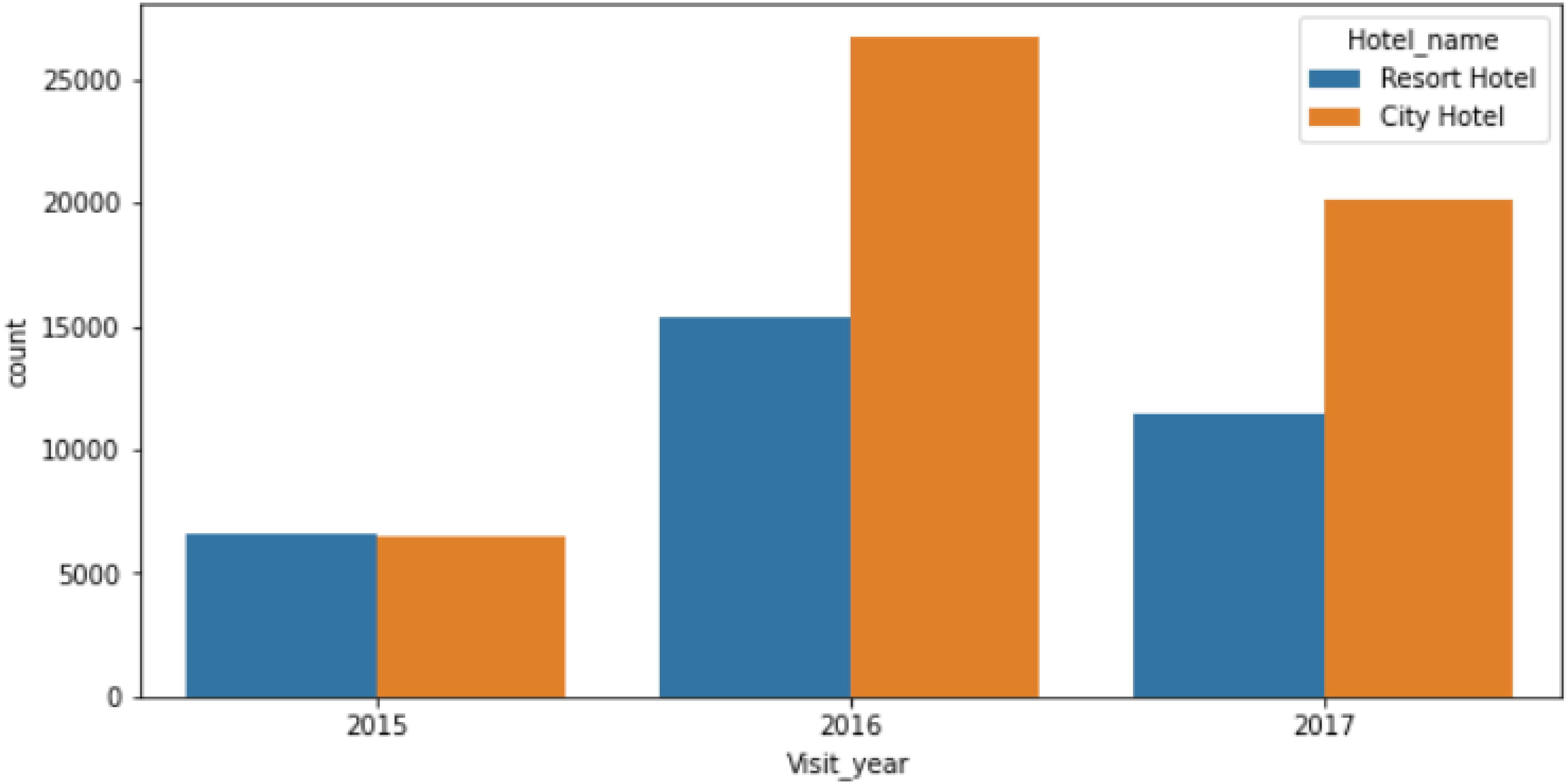
EDA & Data Visualization(continued..)



EDA & Data Visualization(continued..)

- What are the bookings for the hotels over the years?

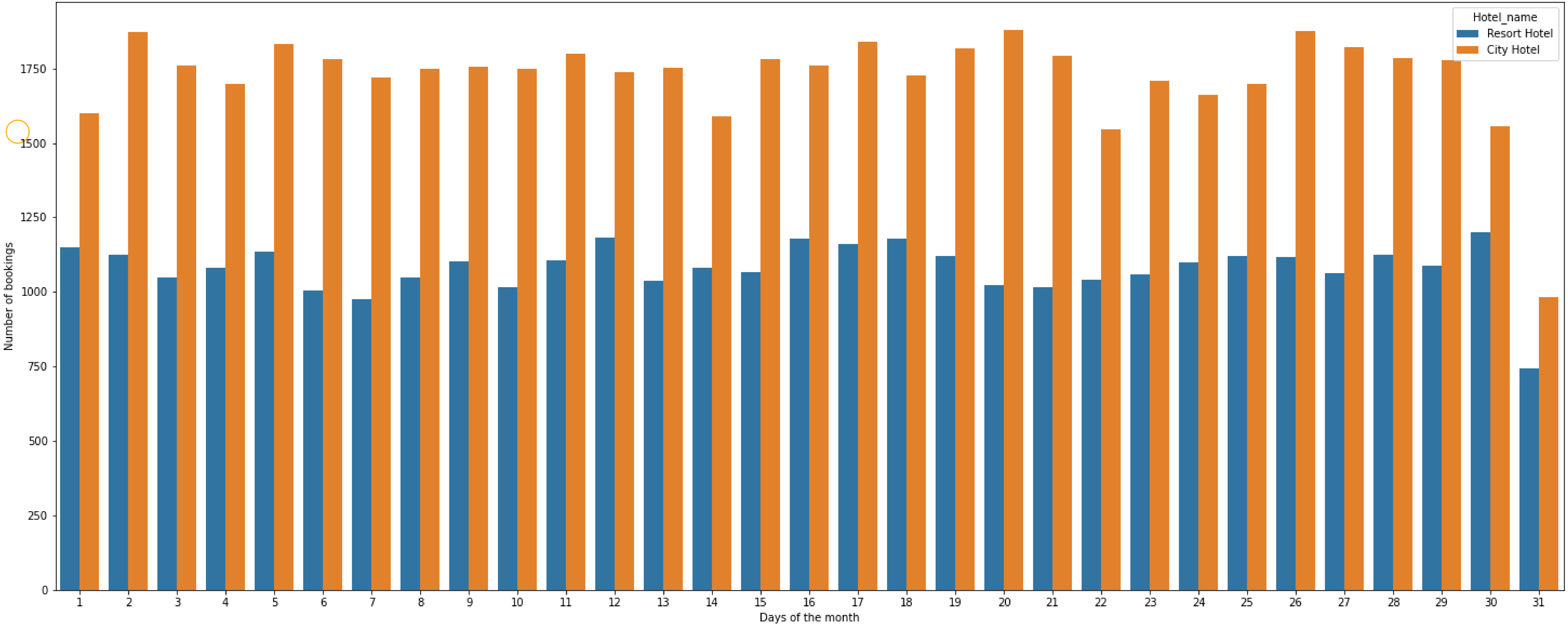
We can observe that the bookings for the year 2015 for both the hotels were almost the same. Whereas in 2016 and 2017 the overall bookings for City hotel are more than Resort hotel.



EDA & Data Visualization(continued..)

Bookings observed throughout the month for the hotels

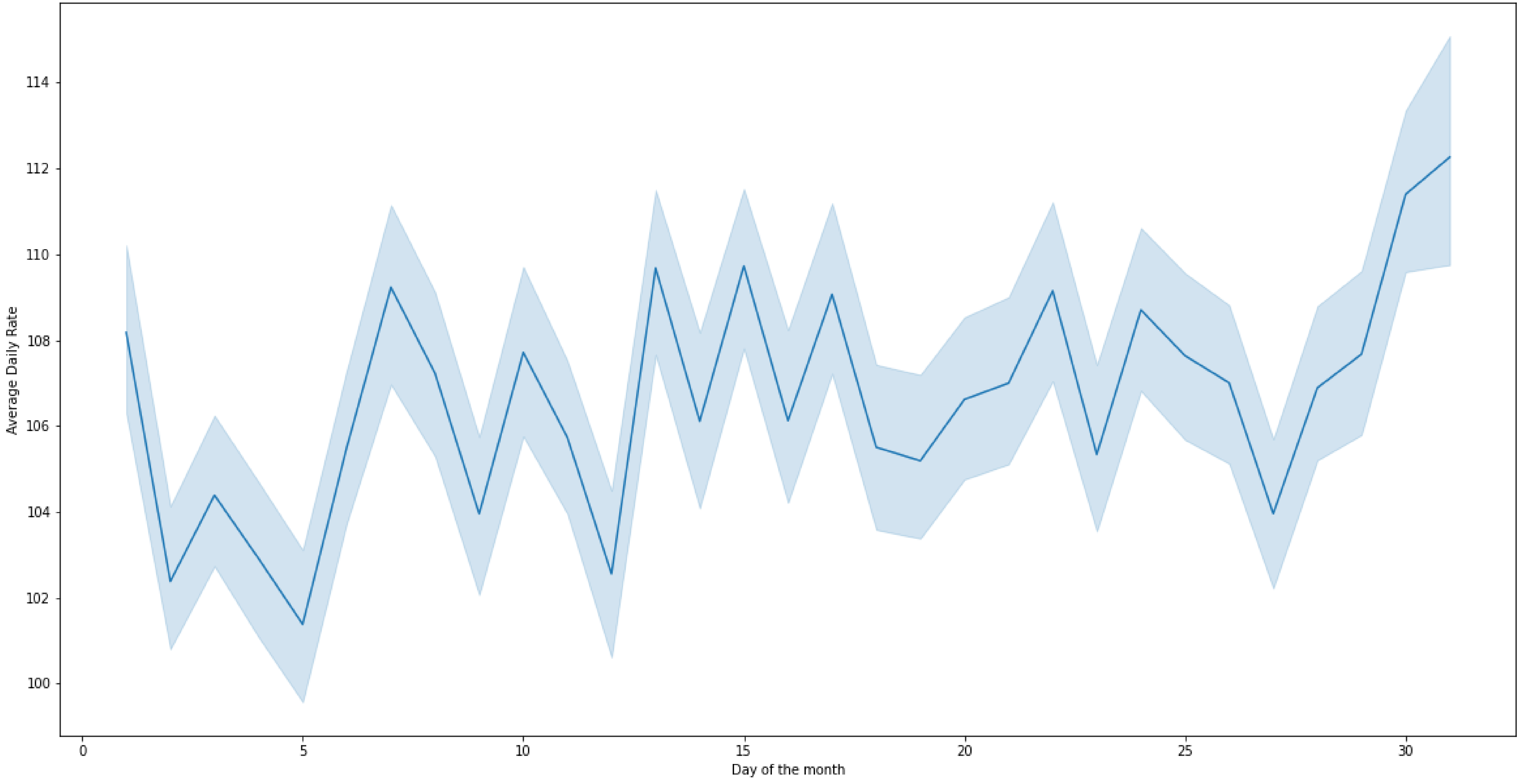
We can see that the bookings throughout the month are fairly the same, with a considerable decrease in the bookings observed by the end of the month.



- **Average daily rate based on the days of the month**

We can infer that the Average daily rate varies a lot throughout the year with increase in daily rate being observed on days lower bookings to obtain a good revenue throughout the month.

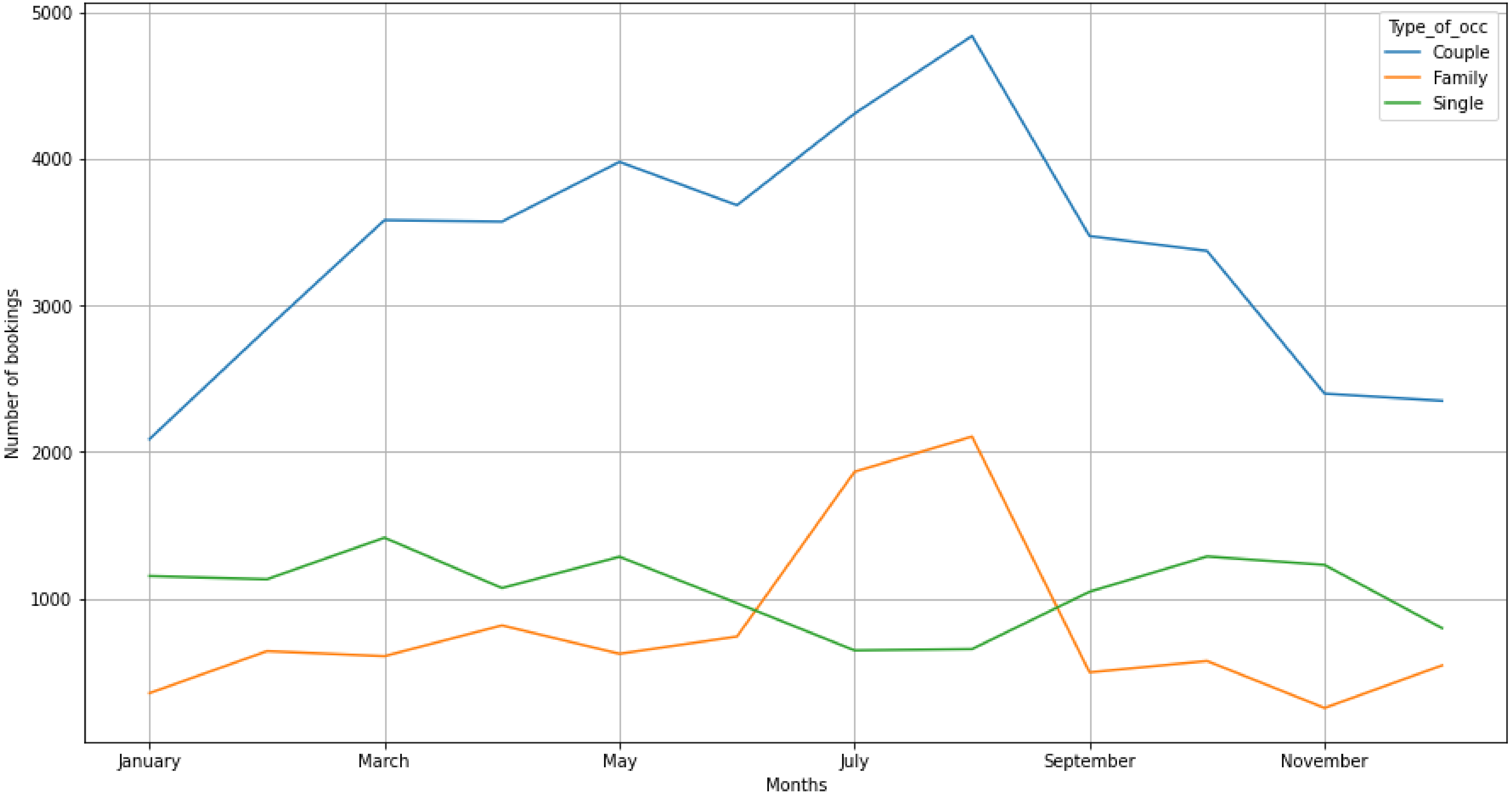
EDA & Data Visualization(continued..)



EDA & Data Visualization(continued..)

- Comparison of bookings across various categories

We can see that the booking for Single occupants stay low throughout the year. The bookings for families visiting the hotels have an increase in the months of July and August which were previously observed to be the busiest months. We can see that the most number of bookings in the hotels are made by couples throughout the year for both hotels.



Project Conclusion

- ✓ More guests showed interest to reside on City Hotel.
- ✓ BB Type of meal was ordered by most of the guests.
- ✓ Country Portugal has most percentage of guests.
- ✓ TA/TO is most commonly used channel for hotel booking in all the years.
- ✓ Maximum bookings were made in the month of August and the least bookings were made at the start of the year.
- ✓ Highest cancellation % is for TA/TO distribution channel.
- ✓ Maximum number of repeated customers are the "Transient type" i.e the "Short-time customers"
- ✓ Booking for Single occupants stay fairly low throughout the year. The bookings for families visiting the hotels have a increase in the months of July and August.
  Most number of bookings in the hotels are made by couples throughout the year for both hotels.

THANK YOU

Q&A