Capstone Project- 1

Hotel Booking Analysis (EDA)

Individual Project

Tanmaya Kumar Pattanaik

Introduction

When we think of travelling, Hotel booking is one of the major factor which comes to our mind firstly. But as a Data scientist, the questions which comes to mind are:

- **When is the best time of year to book a hotel room?**
- **What is the optimal length of stay in order to get the best daily rate?**
- **Which type of hotels are more in demand?**
- **From which places guests are from mostly?**

So, lets get our hand dirty in data and try to answer these questions.

Hotel booking Dataset

- 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.

Column Information

- **Hotel:** We have two types of hotel in the dataset, namely, City hotel and resort hotel
- **is_cancelled**: If the booking is cancelled or not. 1 indicates cancelled and 0 indicates not cancelled
- **lead_time**: Numbet of days that elapsed between entering date of booking into property management system and arrival date
- **arrival_date_year**: Year of arrival date (2015-2017)
- **arrival_date_month:** Month of arrival date (Jan - Dec)
- **arrival_date_week_number:** Week number of year for arrival date (1-53)
- **arrival_date_day_of_month:** Day of arrival date
- **stays_in_weekend_nights**: No of weekend nights (Sat/Sun) the guest stayed to stay at the hotel.
- **stays_in_week_nights**: No of week nights (Mon - Fri) the guest stayed or booked to stay at the hotel
- **Adults**
- **Children**
- **Babies**
- **Meal:** Type of meal booked. Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)

Column Information(Contd.)

- **Country:** From which places people are travelling mostly
- **Market_segment**: a group of people who share one or more common characteristics, lumped together for marketing purposes. (TA- Travel agents and TO- Tour operator)
- **distribution_channel**: A distribution channel is a chain of businesses or intermediaries through which a good or service passes until it reaches the final buyer or the end consumer
- **is_repeated_guest**: value indicating if the booking name was from repeated guest. 1 indicates Yes and 0 indicates No
- **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. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request.

Column Information(Contd.)

- **booking_changes:** Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
- **deposit_type:** Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
- **Agent**: ID of the travel agency that made the booking
- **Company**: ID of the company/entity that made the booking or responsible for paying the booking.
- **day_in_waiting_list**: Number of days the booking was in the waiting list before it was confirmed to the customer
- **customer_type:** Contract, Group, Transient, Transient Party
- **adr (average daily rate)**
- **required_car_parking_space:** 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: Cancelled, check- out or no-show**
- **reservation_status_date:** Date at which the last status was set

Problem Statement

- We have almost everything in the dataset that is required to analyze it.
- We need to explore and analyze the data to discover important factors that govern the bookings of a Hotel.
- Lets get started in our EDA of the dataset and try to learn more about the data and get to some conclusion.

Required packages for EDA

- Pandas
- Numpy
- Matplotlib
- Seaborn

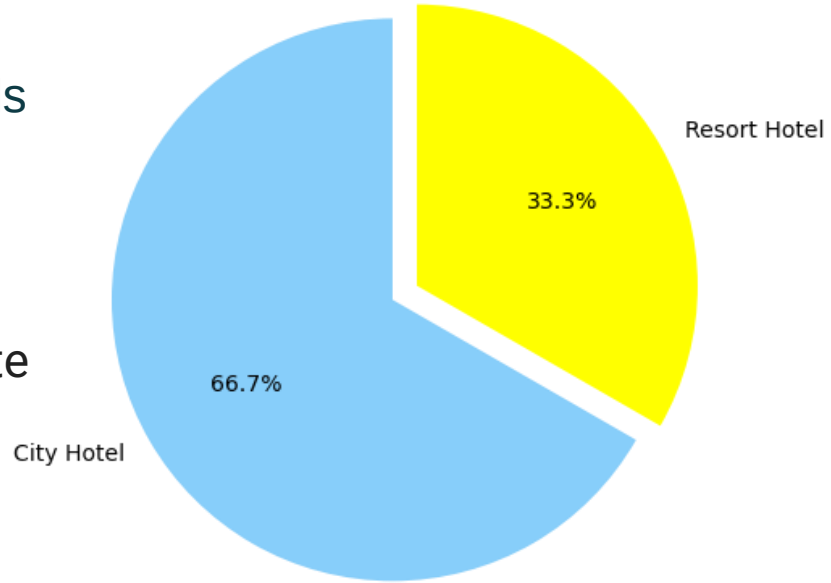
Data and NaN Values

- There are 119390 rows and 32 Columns, which is huge.
- NaN or Null Values: After inspecting the Nan Values of Dataset, we have quite a lot of NaN values in company and agent (Agent- 16340, Company- 112593). But as per our analysis, these 2 has very less impact.
- So decided to remove these columns and move on with our analysis.
- The reason why we can't remove rows with NaN value is because that will mean we are removing 112593 rows out of 119390 rows. So removing columns will be a better idea since those 2 attributes (agents and companies) are unimportant.
- We also have one more column with NaN Values i.e., country column with 488 NaN values. 488 rows out of 119390 is negligible hence I just removed.

Overview of the type of hotel

What do we see here?

- It Seems that a huge proportion of hotels was city hotels. 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.



Overview of the number of people who booked the hotel

- It seems that mean values for adults and children are higher.
- This means that resort hotels are better choice for large families.

Looking into adults

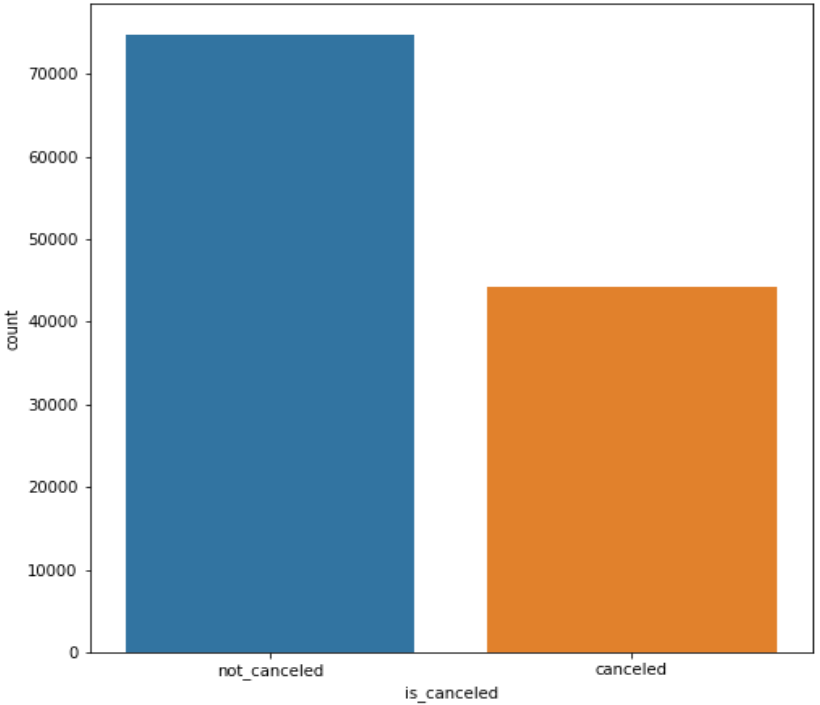
hotel	count	mean	std	min	25%	50%	75%	max
City Hotel	79302.0	1.851126	0.509013	0.0	2.0	2.0	2.0	4.0
Resort Hotel	39596.0	1.872942	0.697112	0.0	2.0	2.0	2.0	55.0

Looking into children

hotel	count	mean	std	min	25%	50%	75%	max
City Hotel	79302.0	0.091397	0.372230	0.0	0.0	0.0	0.0	3.0
Resort Hotel	39596.0	0.129862	0.447192	0.0	0.0	0.0	0.0	10.0

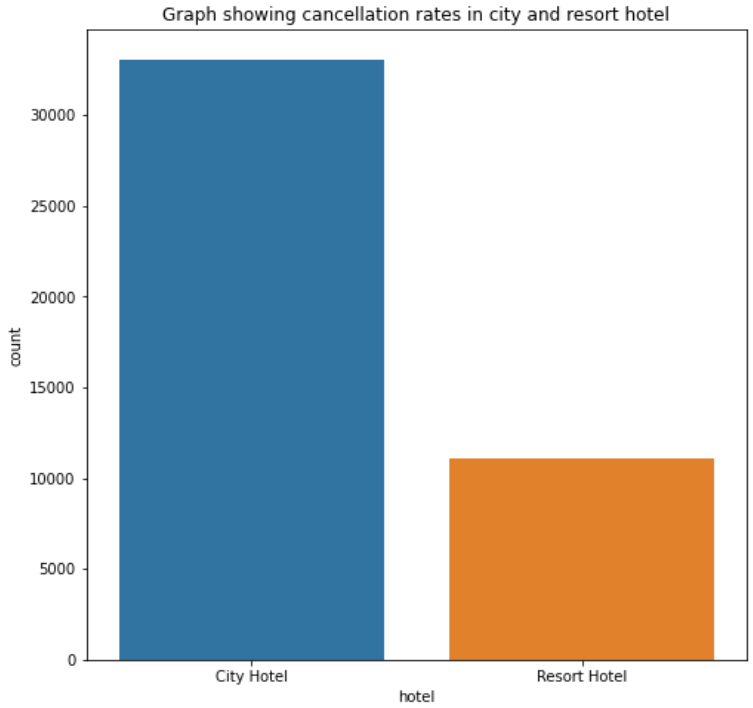
Overview of cancelled bookings

- It seems that majority of the bookings were not canceled.

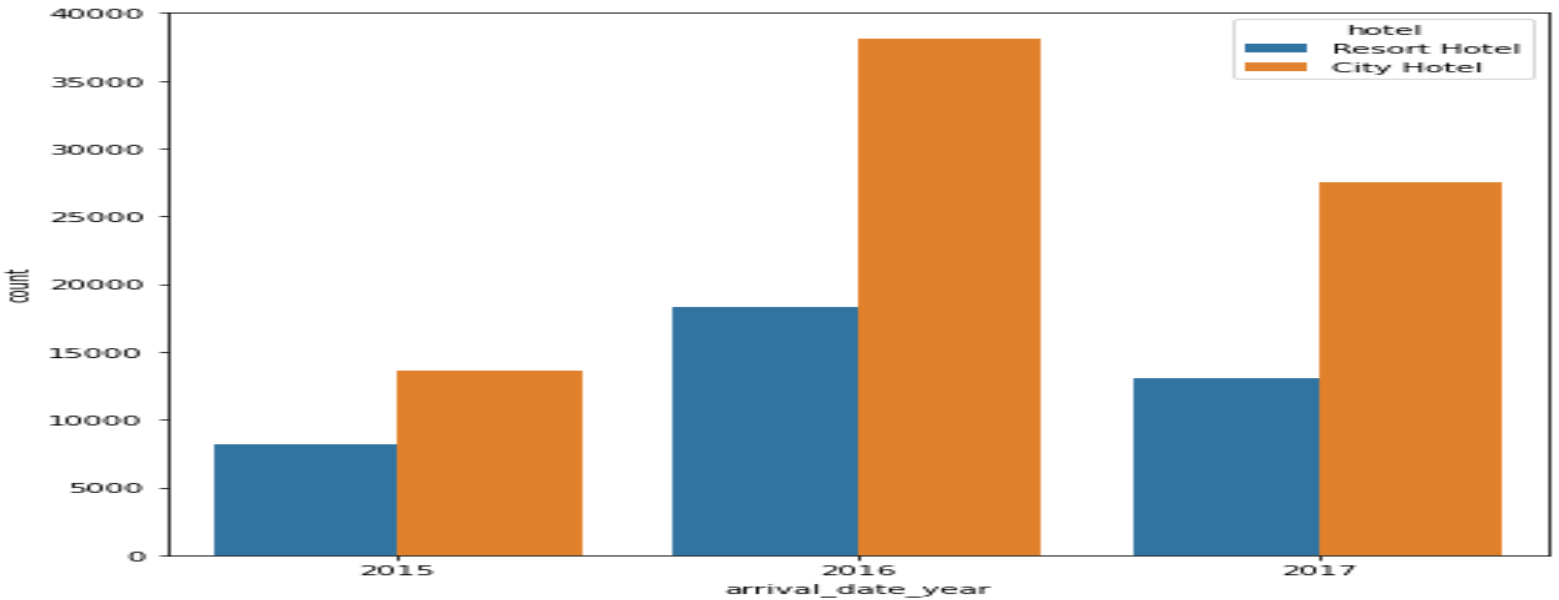


Cancellation rate among different type of hotel

- Huge proportion of cancellation from city hotel.
- This was expected since 3/4 of the hotel bookings belong to city hotels.

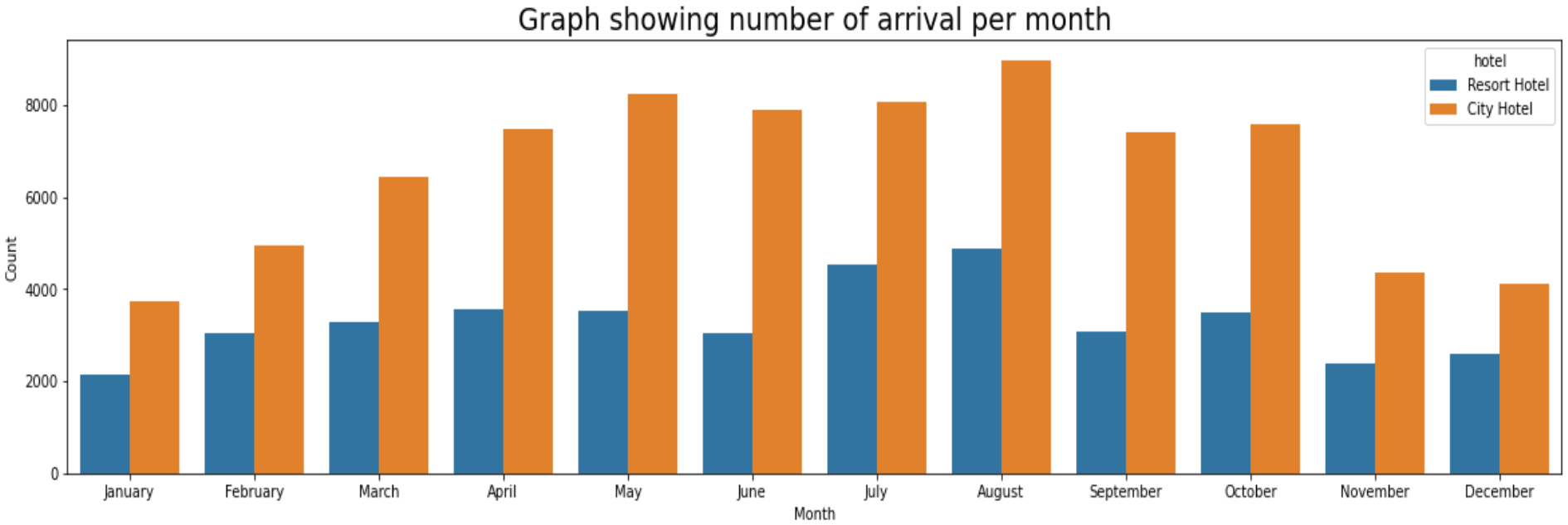


Overview of arrival period



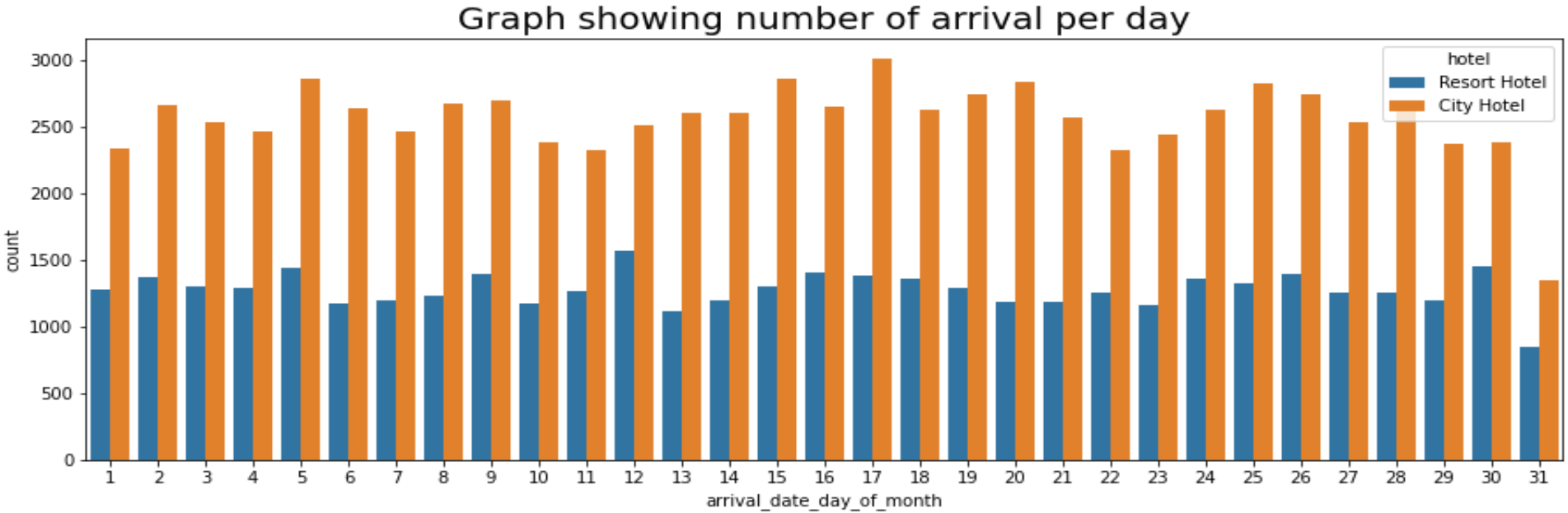
We can see that 2016 seems to be the year where hotel booking is at its highest.

Overview of arrival period(Contd.)



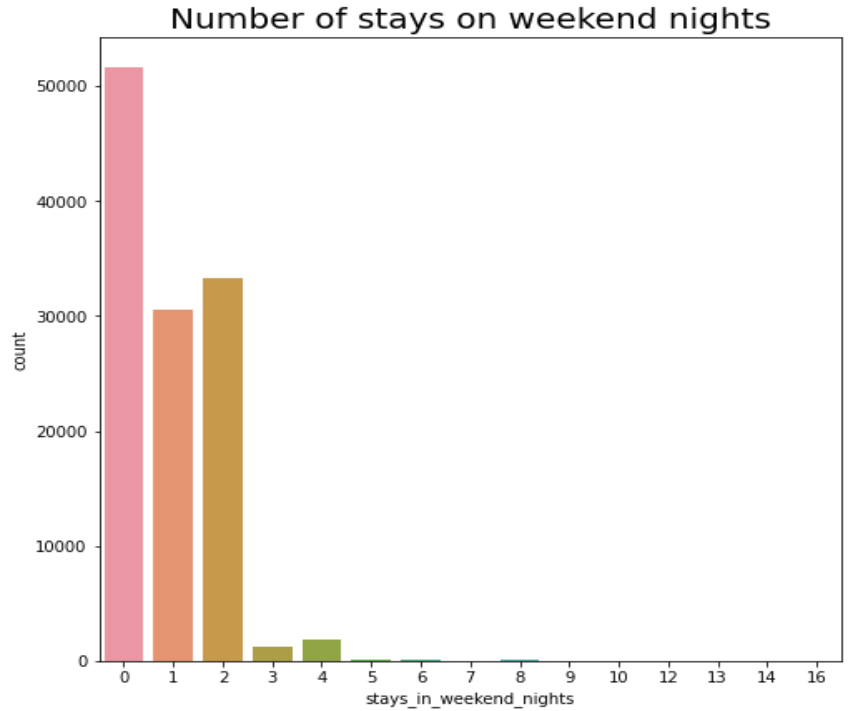
We also see an increasing trend in booking around the middle of the year, with August being the highest. Summer ends around August, followed straight by autumn. It seems that summer period is a peak period for hotel booking.

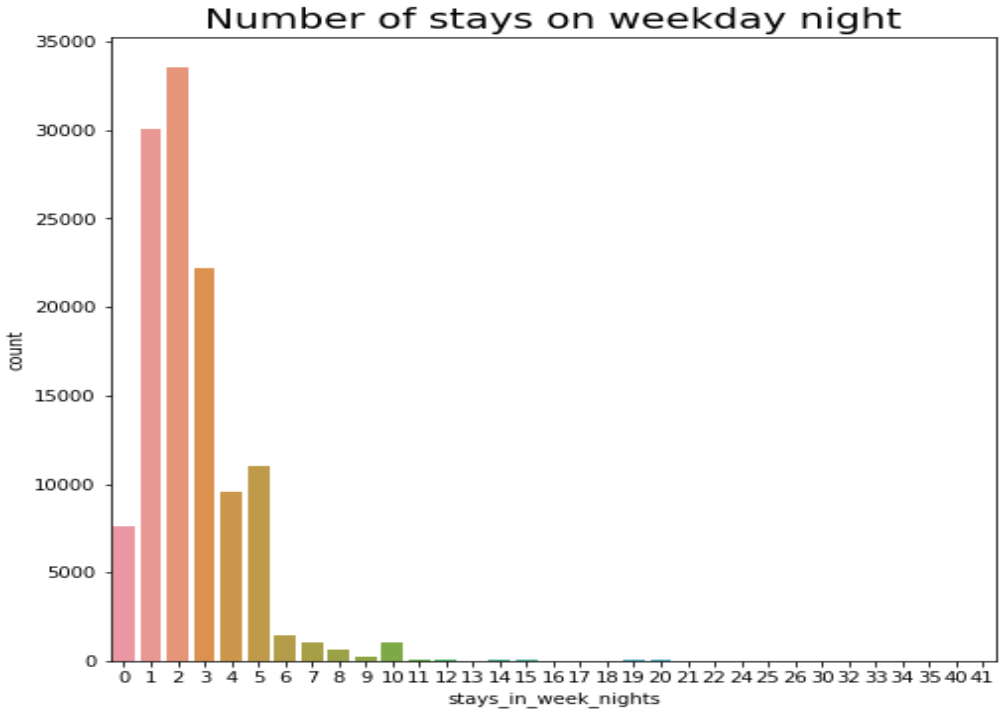
Overview of arrival period(Contd.)



We do notice a roller coaster trend for the arrival day of month. Could the peaks belong to a weekend? (i.e Will people tend to book over the weekends?)

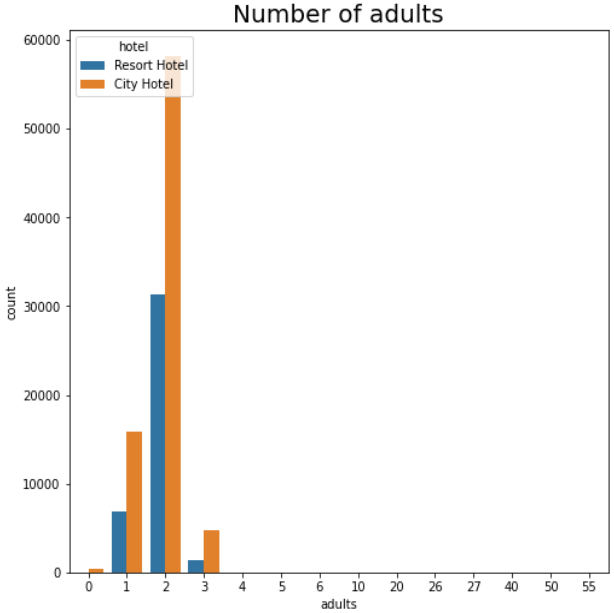
Let's dig deeper into whether the stay is over a weekend or weekday

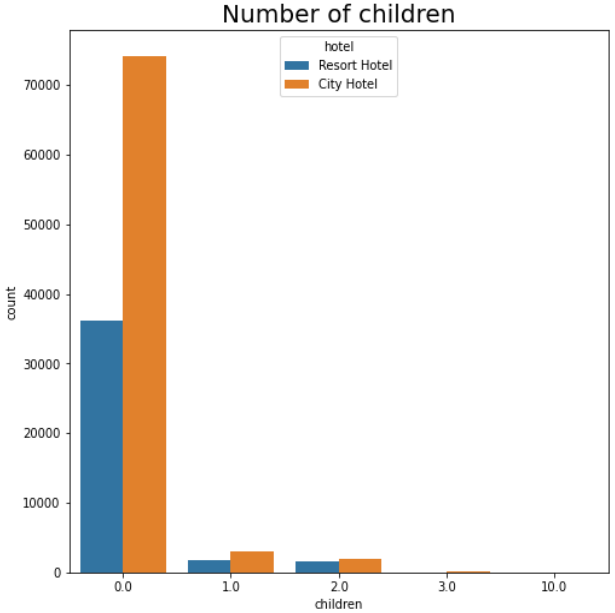


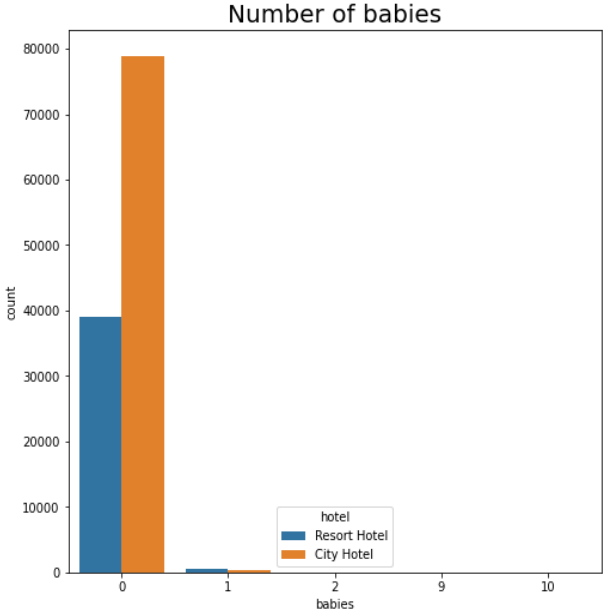


It seems that majority of the stays are over the weekday's night. Hence, it seems that whatever we saw for the chart on day of the month was random.

Type of visitors

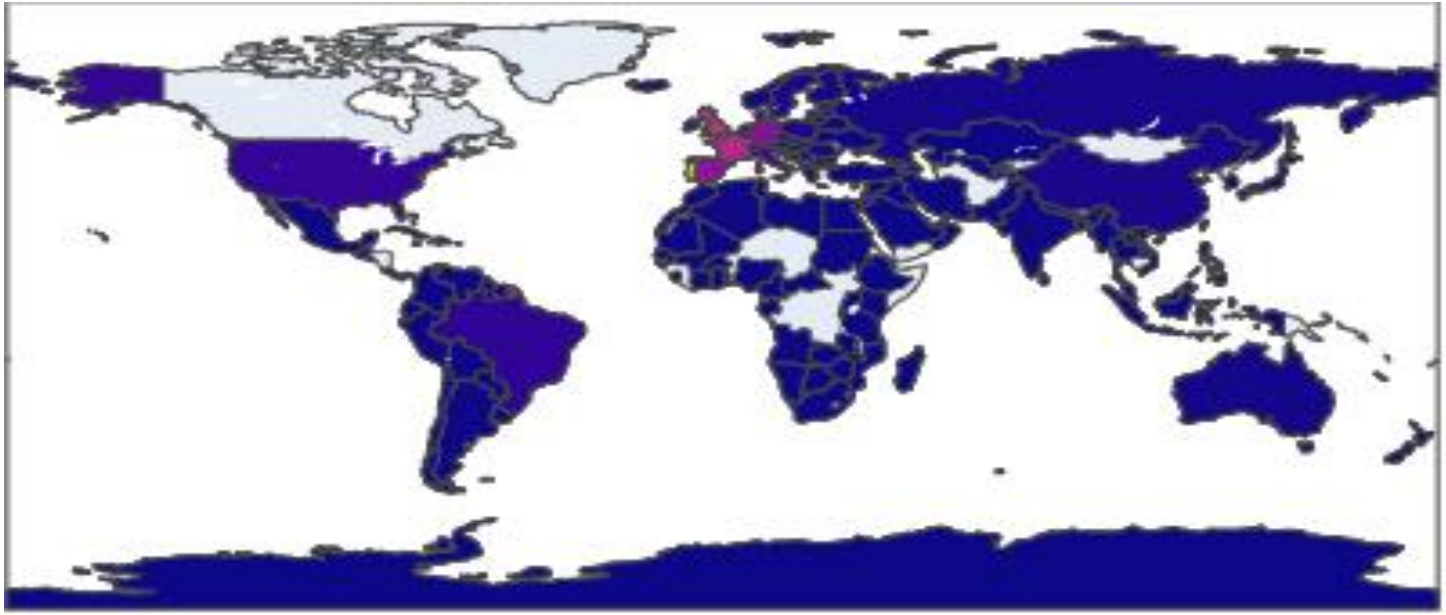






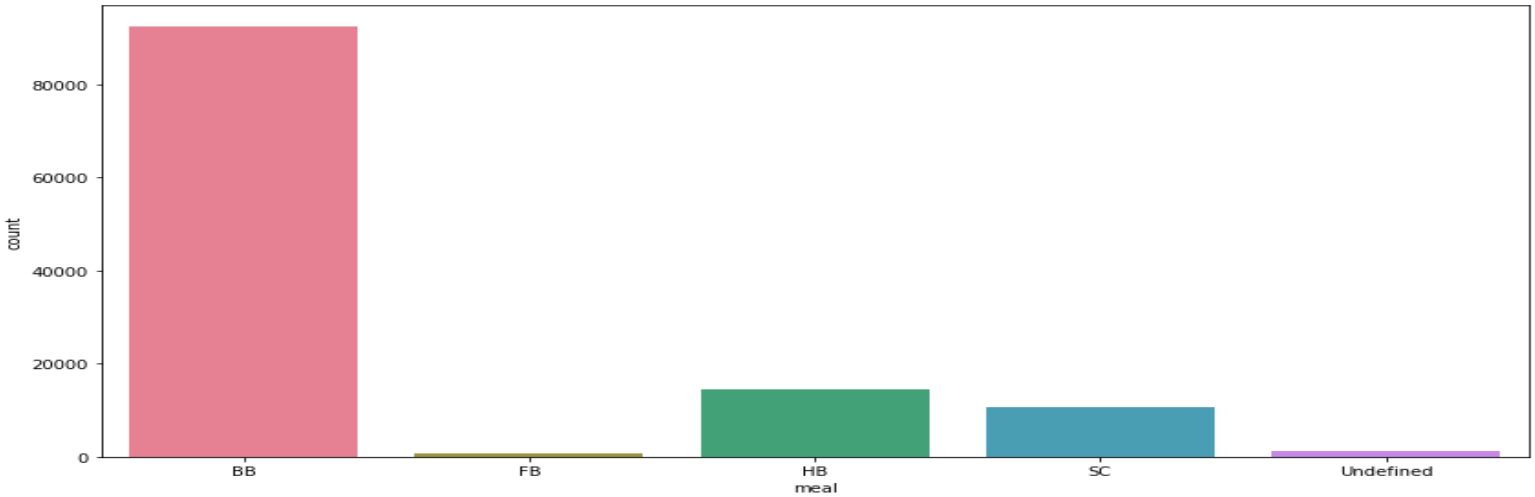
It seems that majority of the visitors travel in pair. Those that travel with children or babies have no specific preference for the type of hotel. We do see that those bringing babies along prefer resort hotels.

Looking into which countries the visitors are from



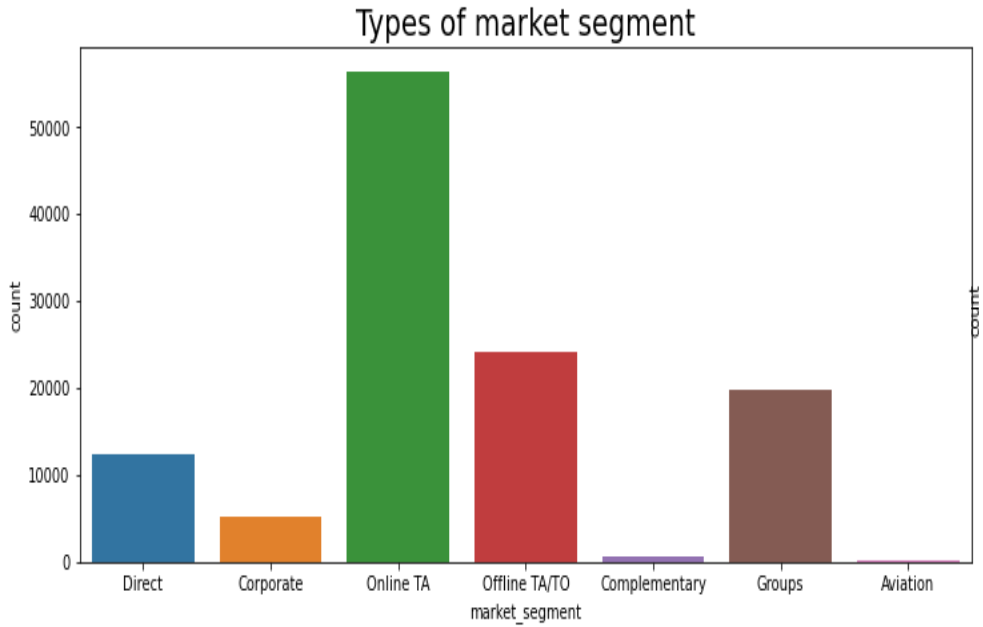
- **We have a huge number of visitors from western europe, namely France,UK and Portugal being the highest.**
- **We can instruct the marketing team to target people of this region**

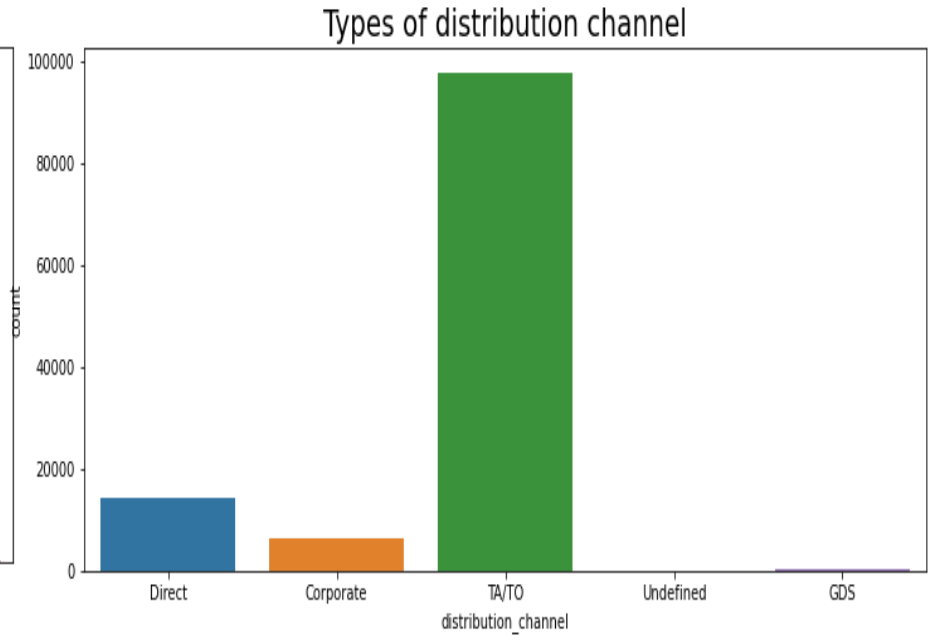
Overview of meal prefered



- **Majority of customers prefer Bed and Breakfast meal type.**
- **HB — Half board (breakfast and one other meal — usually dinne)is somewhat less.**
- **FB — Full board (breakfast, lunch and dinner) and Undefined — no meal package are not prefered by customers.**

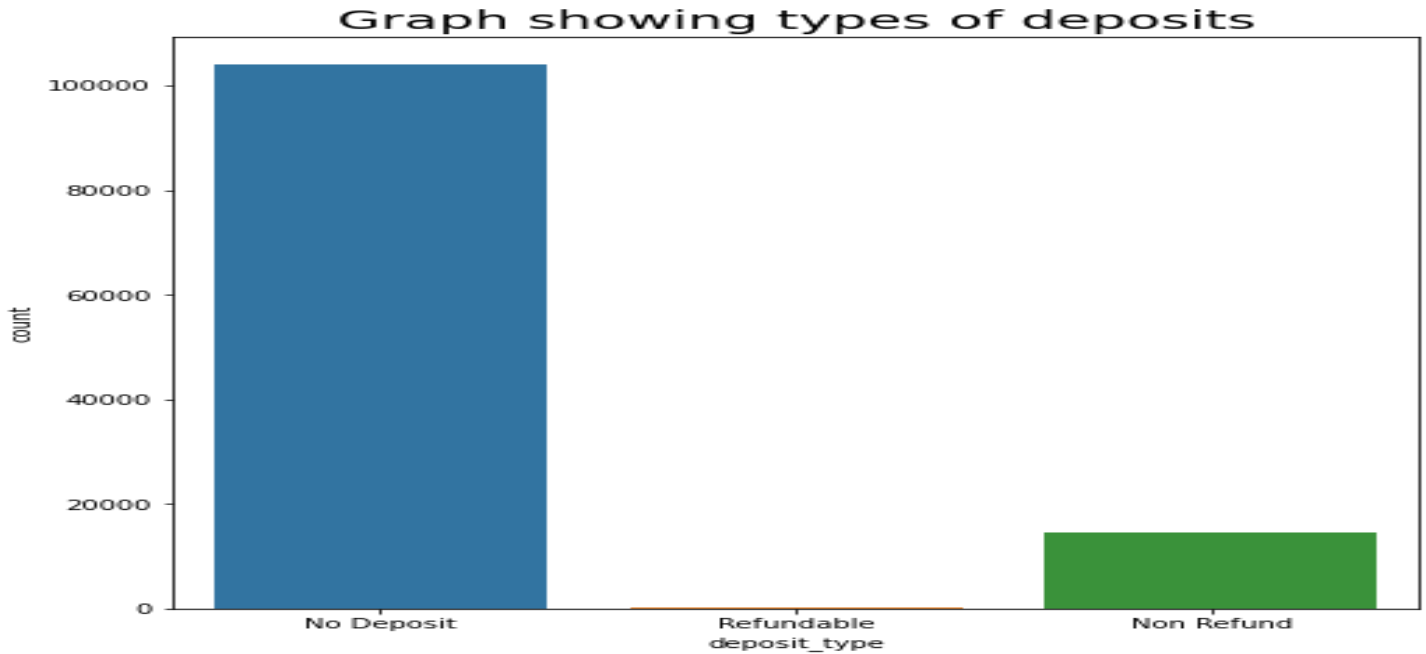
Looking into market segments and distribution channel





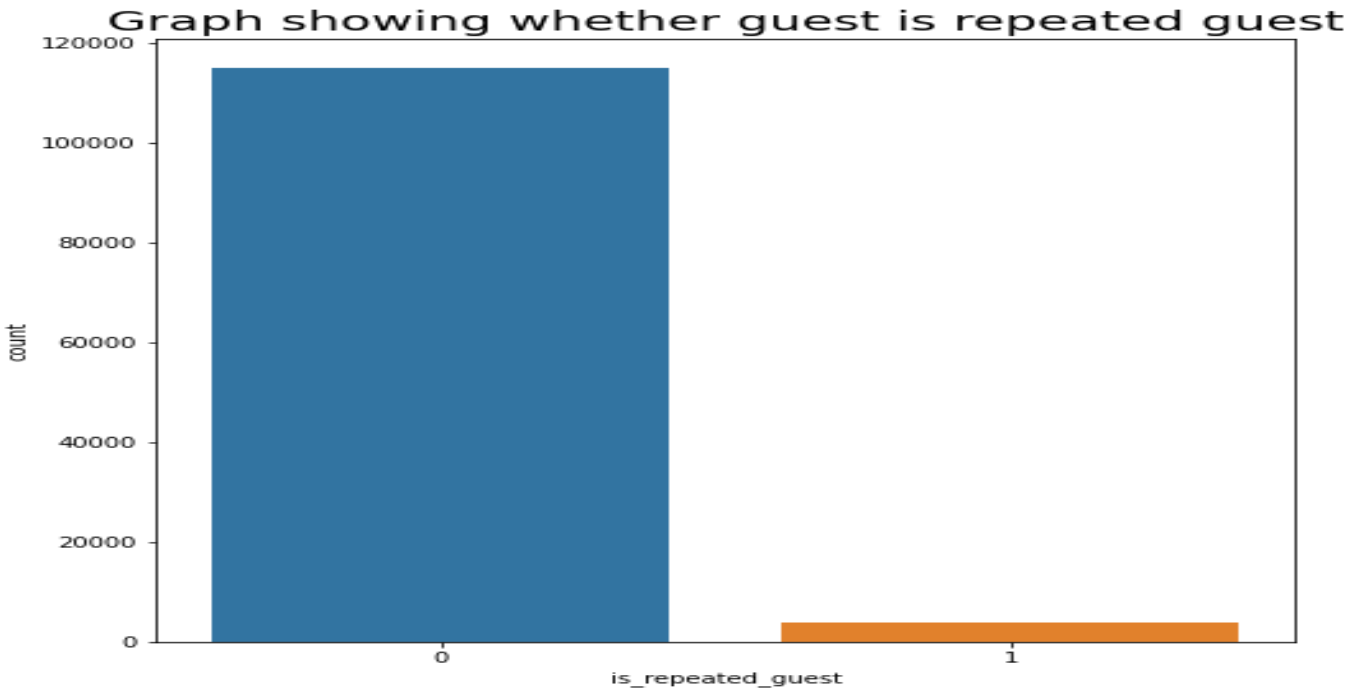
- **Majority of the distribution channels and market segments involve travel agencies (online or offline).**
- **We can target our marketing area to be on these travel agencies website and work with them since majority of the visitors tend to reach out to them**

Looking into deposit types



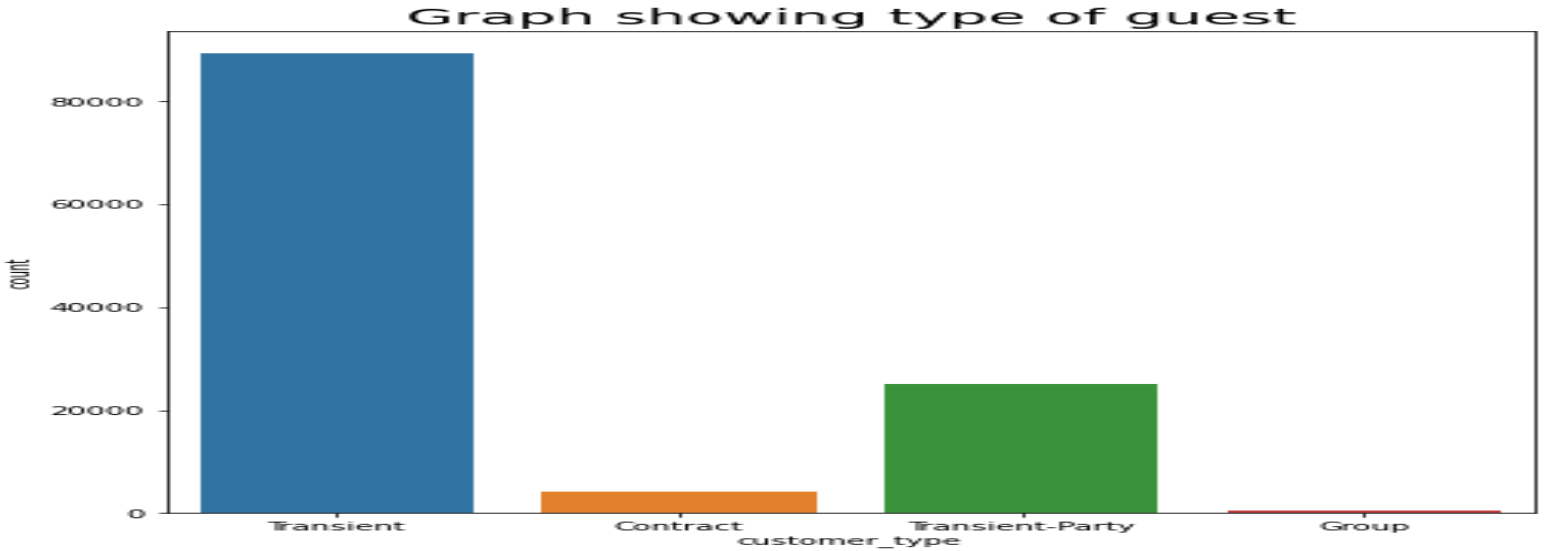
Majority of the booking does not require deposit. That could explain why cancellation rate was actually 50% of non-cancellation rate.

Overview of repeated guests



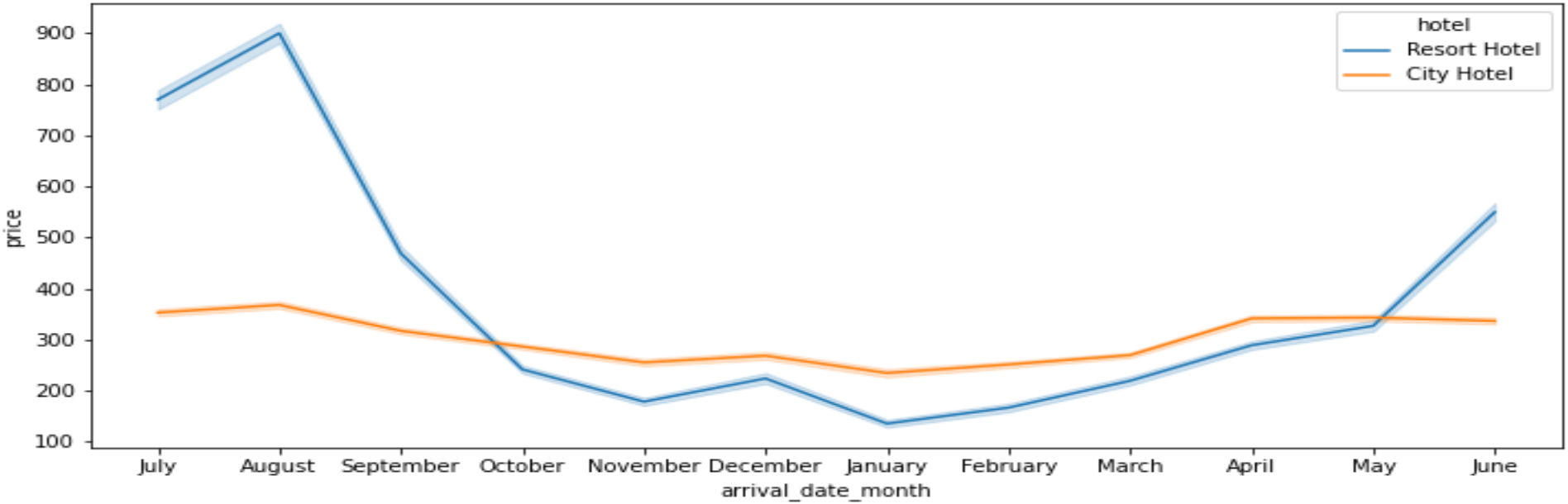
- **Low number of repeated guests.**
- **A need to target repeated guests since they have booked before.**

Looking at types of guests



- **Majority of the bookings are transient.**
- **This means that the booking is not part of a group or contract.**
- **With the ease of booking directly from the website, most people tend to skip the middleman to ensure quick response from their booking.**

Looking into prices per month per hotel



- **Prices of resort hotel are much higher. It seems that that is definitely the case since resort hotels specialize in that.**

- **Prices of city hotel do not fluctuate that much.**

Conclusion

- 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 August. 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.

Thank You