Hotel Bookings Exploratory Data Analysis

Objective

We are provided with a hotel bookings dataset.
Out main objective is perform EDA on the given dataset and draw useful conclusions about general trends in hotel bookings and how factors governing hotel bookings interact with each other.

Dataset

We are given a hotel bookings dataset. This dataset contains booking information for a city hotel and a resort hotel. It contains the following features.

- hotel: Name of hotel (City or Resort)
- is_canceled: Whether the booking is canceled or not (0 for no canceled and 1 for canceled)
- lead_time: time (in days) between booking transaction and actual arrival.
- arrival_date_year: Year of arrival
- arrival_date_month: month of arrival
- arrival_date_week_number: week number of arrival date.
- arrival_date_day_of_month: Day of month of arrival date
- stays_in_weekend_nights: No. of weekend nights spent in a hotel
- stays_in_week_nights: No. of weeknights spent in a hotel
- adults: No. of adults in single booking record.
- children: No. of children in single booking record.
- babies: No. of babies in single booking record.
- meal: Type of meal chosen
- country: Country of origin of customers (as mentioned by them)
- market_segment: What segment via booking was made and for what purpose.
- distribution_channel: Via which medium booking was made.
- is_repeated_guest: Whether the customer has made any booking before.
- previous_cancellations: No. of previous canceled bookings.
- previous_bookings_not_canceled: No. of previous non-canceled bookings.
- reserved_room_type: Room type reserved by a customer.
- assigned_room_type: Room type assigned to the customer.
- booking_changes: No. of booking changes done by customers
- deposit_type: Type of deposit at the time of making a booking (No deposit Refundable/ No refund)
- agent: Id of agent for booking
- company: Id of the company making a booking
- days_in_waiting_list: No. of days on waiting list.
- customer_type: Type of customer(Transient, Group, etc.)
- adr: Average Daily rate.
- required_car_parking_spaces: No. of car parking asked in booking
- total_of_special_requests: total no. of special request.
- reservation_status: Whether a customer has checked out or canceled,or not showed
- reservation_status_date: Date of making reservation status.
- Total number of rows in data: 119390 - Total number of columns: 32

Data Cleaning and Feature Engineering

(1) Converting columns to appropriate data types

- Changed data type of `children`, `company`, `agent` to int type.
- Changed data type of `reservation_status_date` to date type.

(2) Creating a copy of original DataFrame

- hotel_df = df.copy()

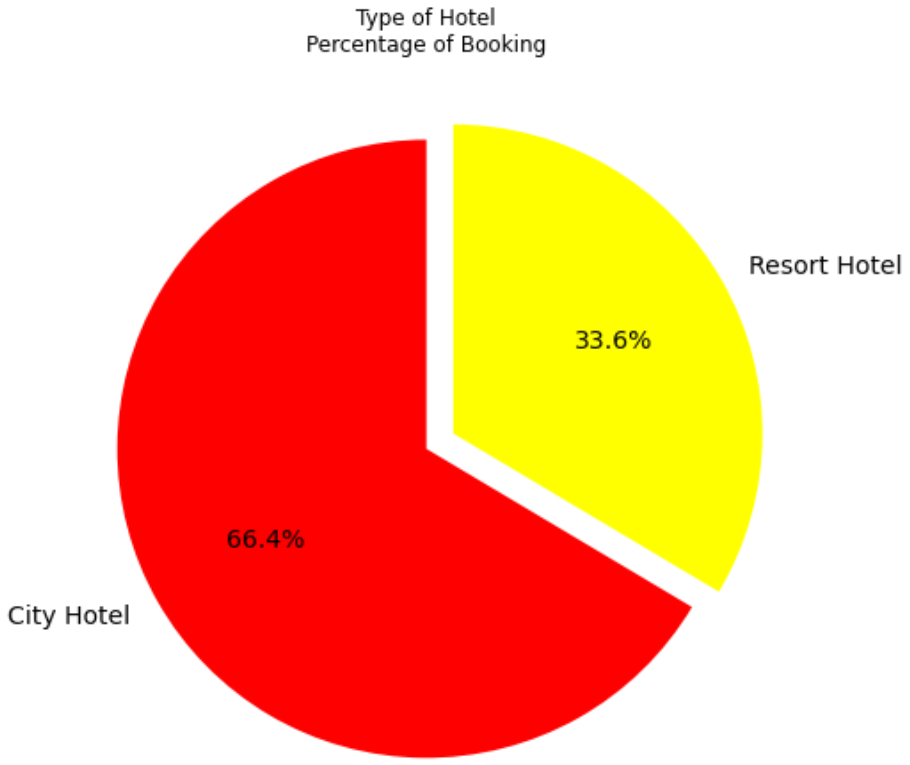
(3) Removing the duplicate rows and cleaning the null value

- hotel_df.drop_duplicates(inplace=True)

Exploratory Data Analysis

Hotel wise analysis

Around 60% bookings are for City hotel 66.40% and 33.60% bookings are for Resort hotel, therefore City Hotel is busier than Resort hotel.



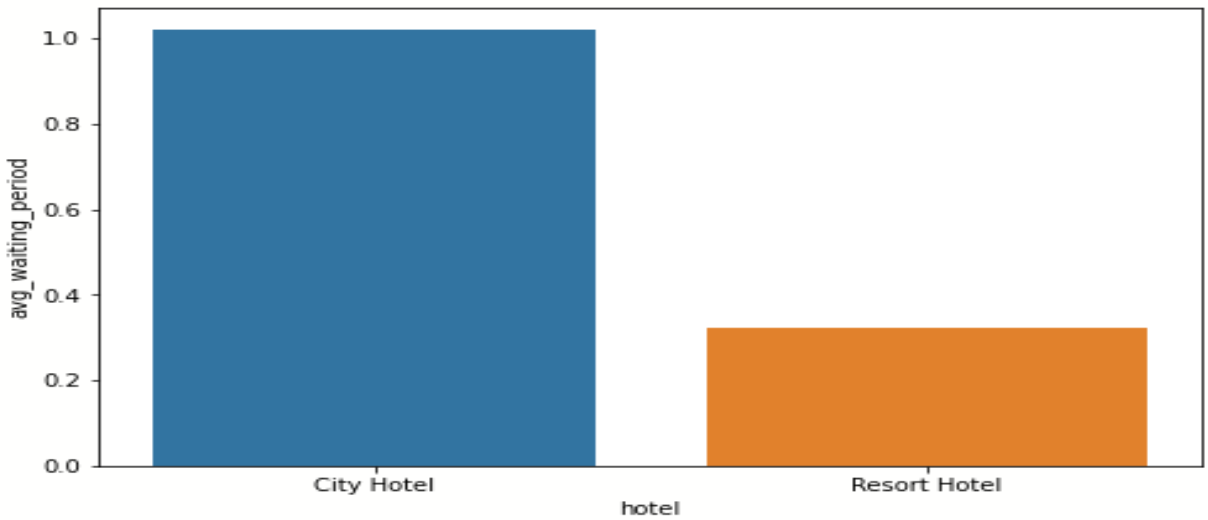
Checked-in and Cancled Booking

which Hotel has higher lead time?

City hotel has slightly higher median lead time. Also median lead time is significantly higher for both hotels, this means customers generally plan their hotel visits way early.

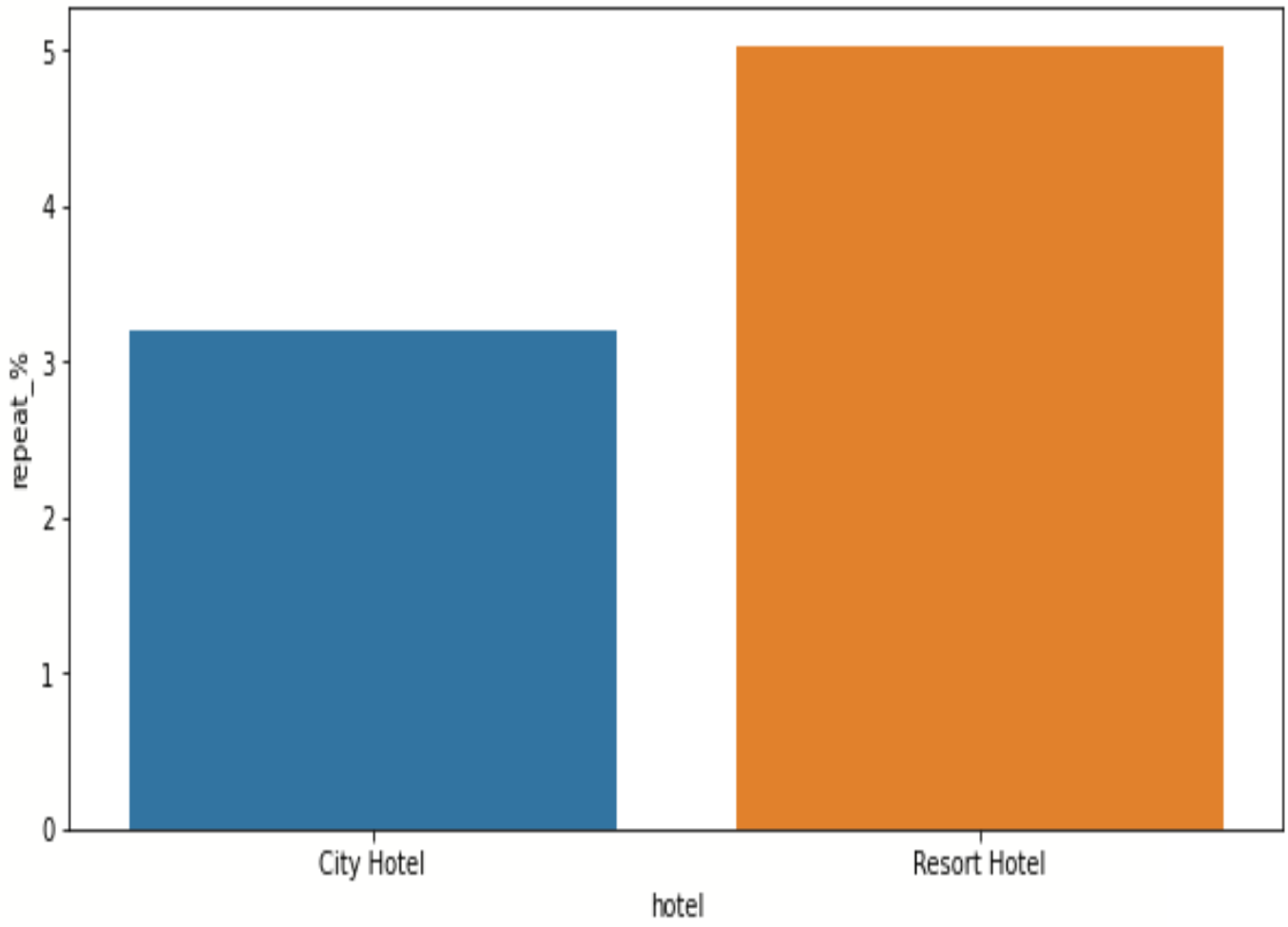
which hotel has longer waiting time?

City hotel has longer waiting time in comparison of resort hotel.



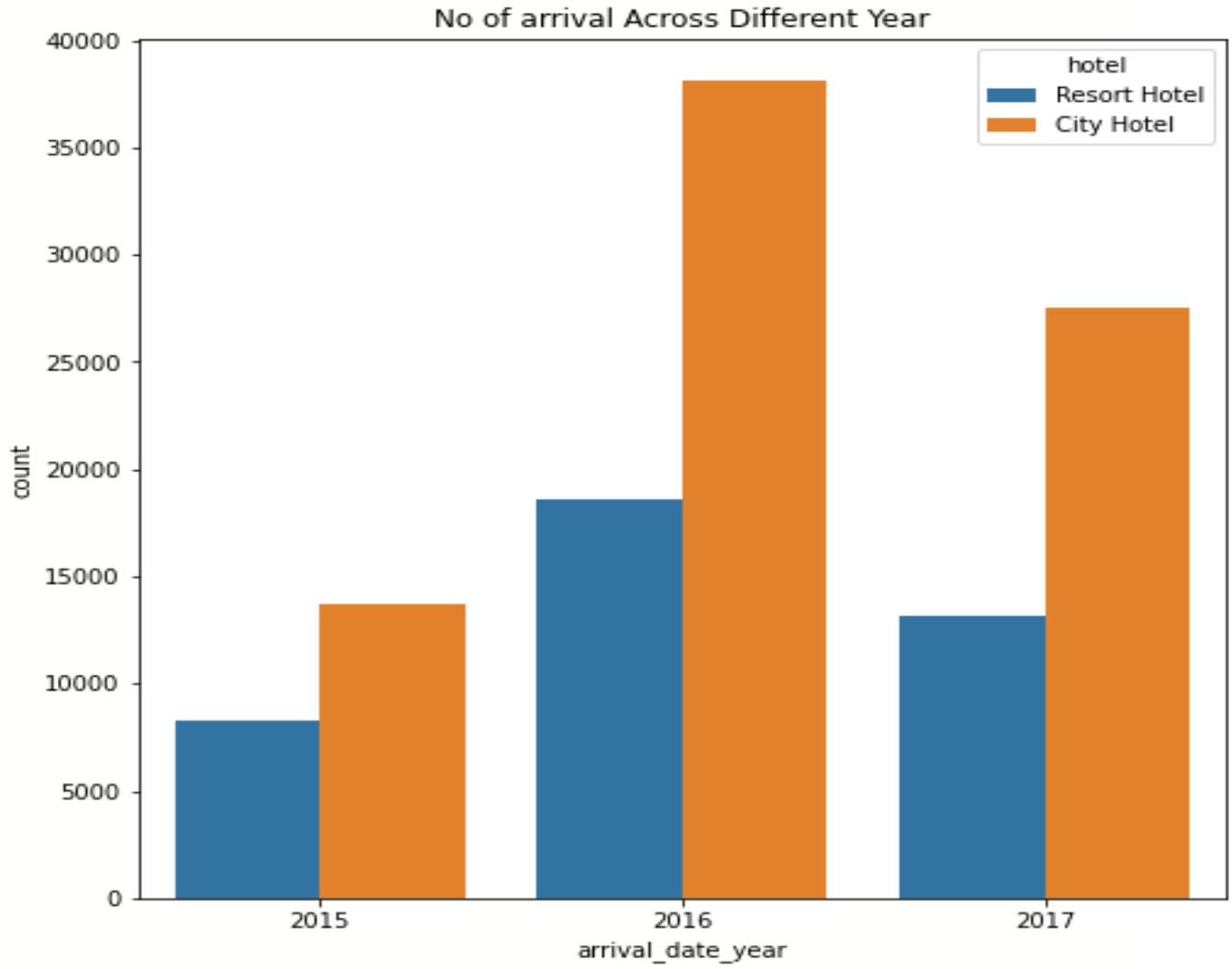
which hotel has more number of repeated guest?

Resort hotel has more number of repeated guest.

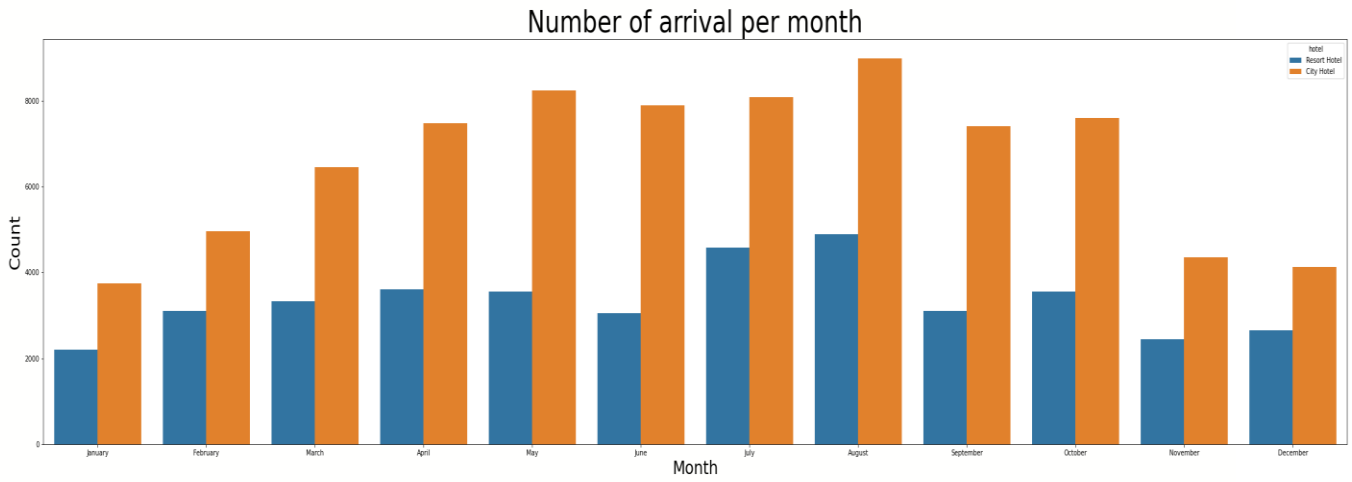


Booking across different Year, Months & Days

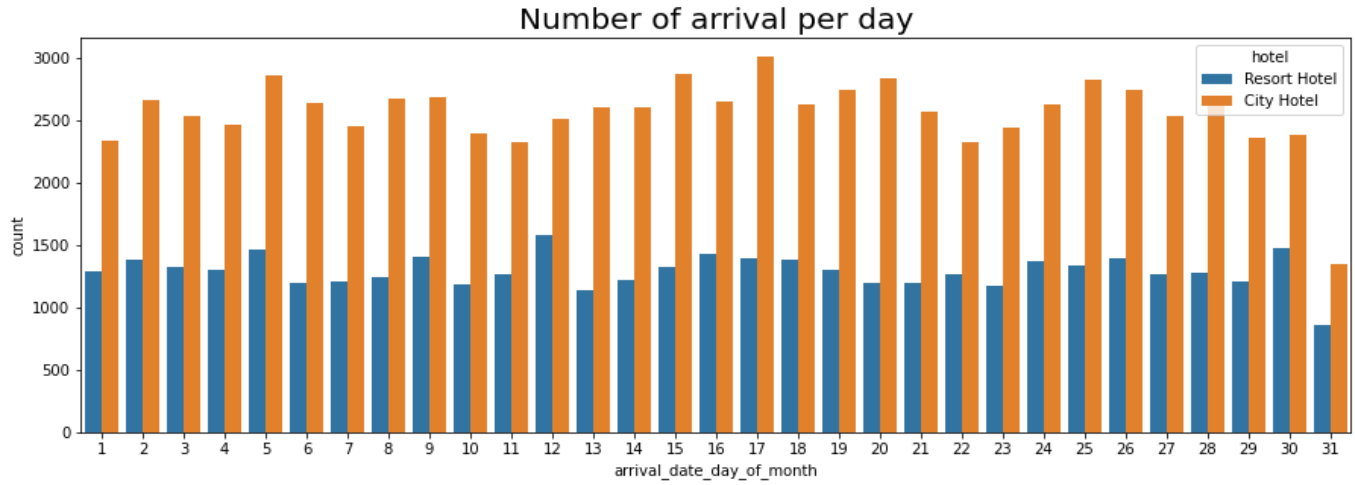
2016-2017 are the most busier and profitable Year for both of hotels.


No of arrival Across Different Year
hotel
Resort Hotel
City Hotel
40000
35000
30000
25000
20000
15000
10000
5000
0
count
2015
2016
2017
arrival_date_year

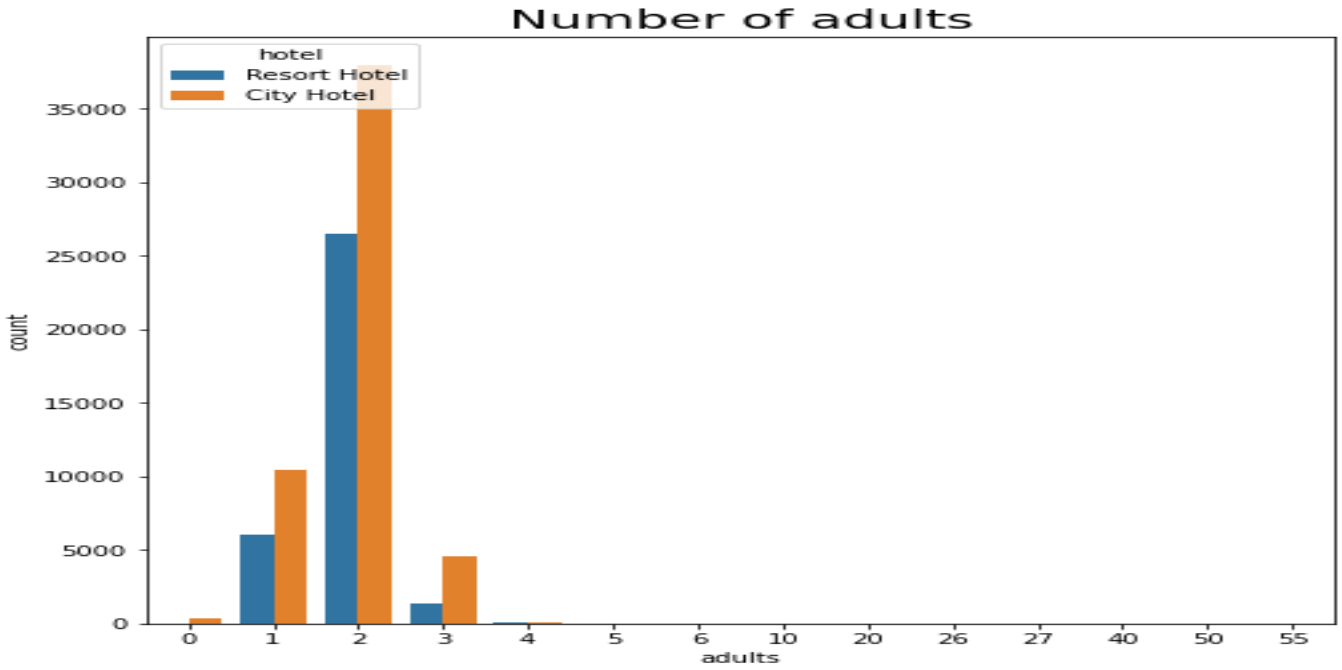
MonthWise

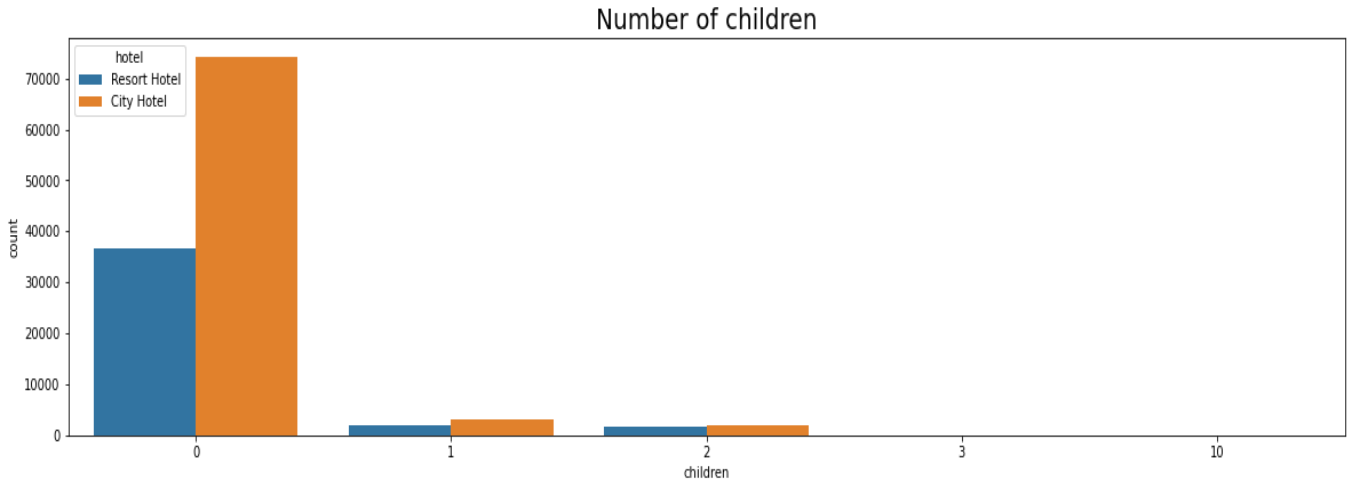

Number of arrival per month
hotel
Resort Hotel
City Hotel
Count
Month
January
February
March
April
May
June
July
August
September
October
November
December

Daywise

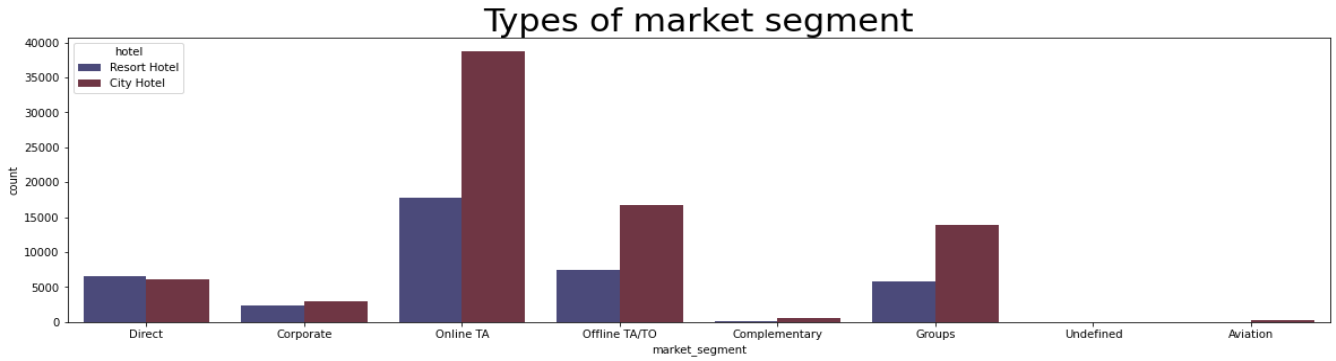


Type of Visitor

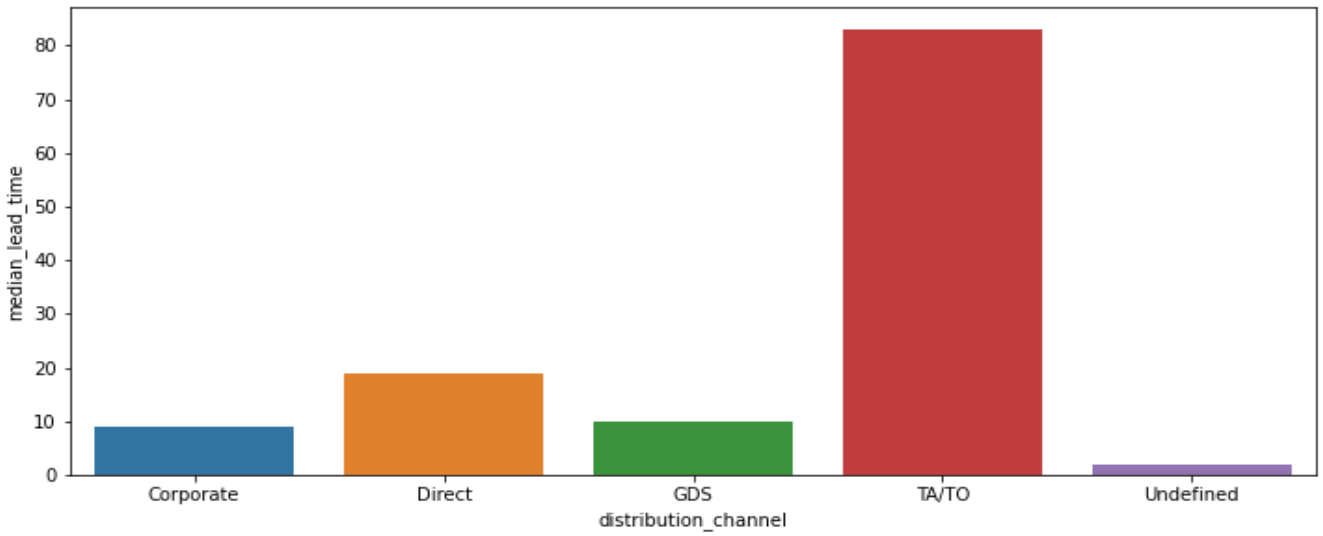




Market Segment

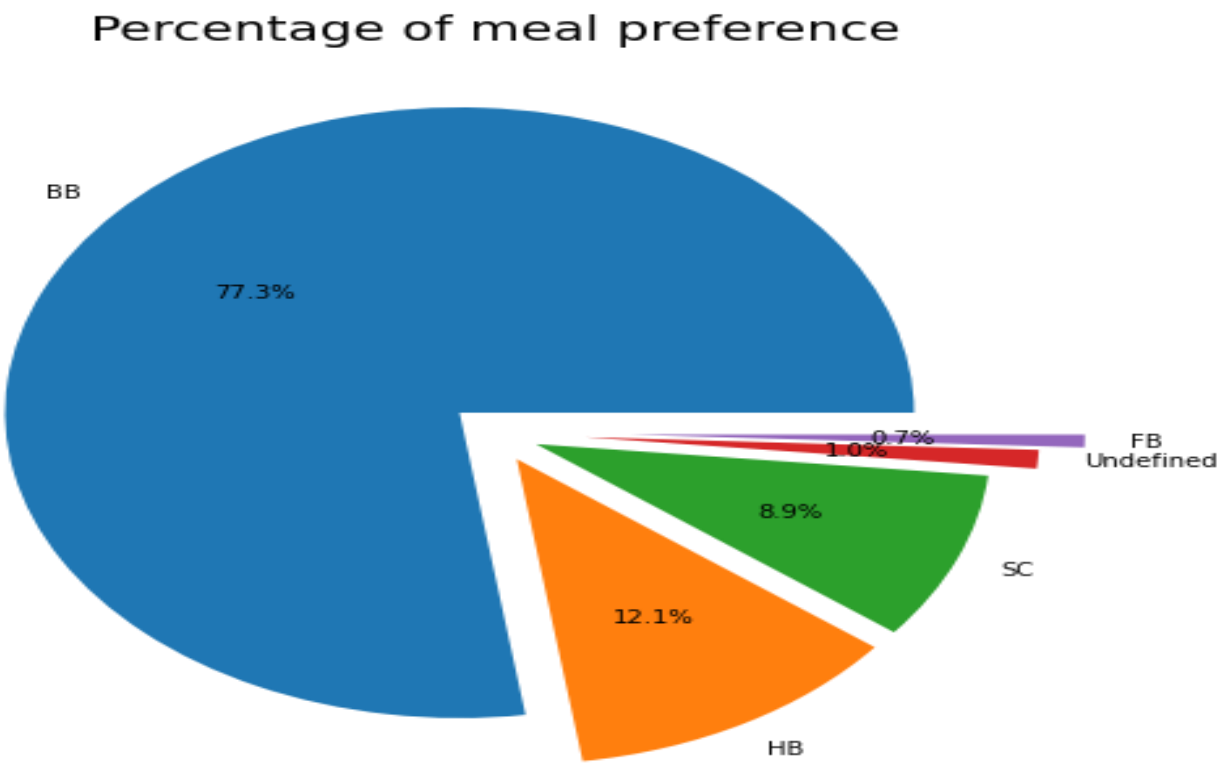


Distribution Channel by Analysis



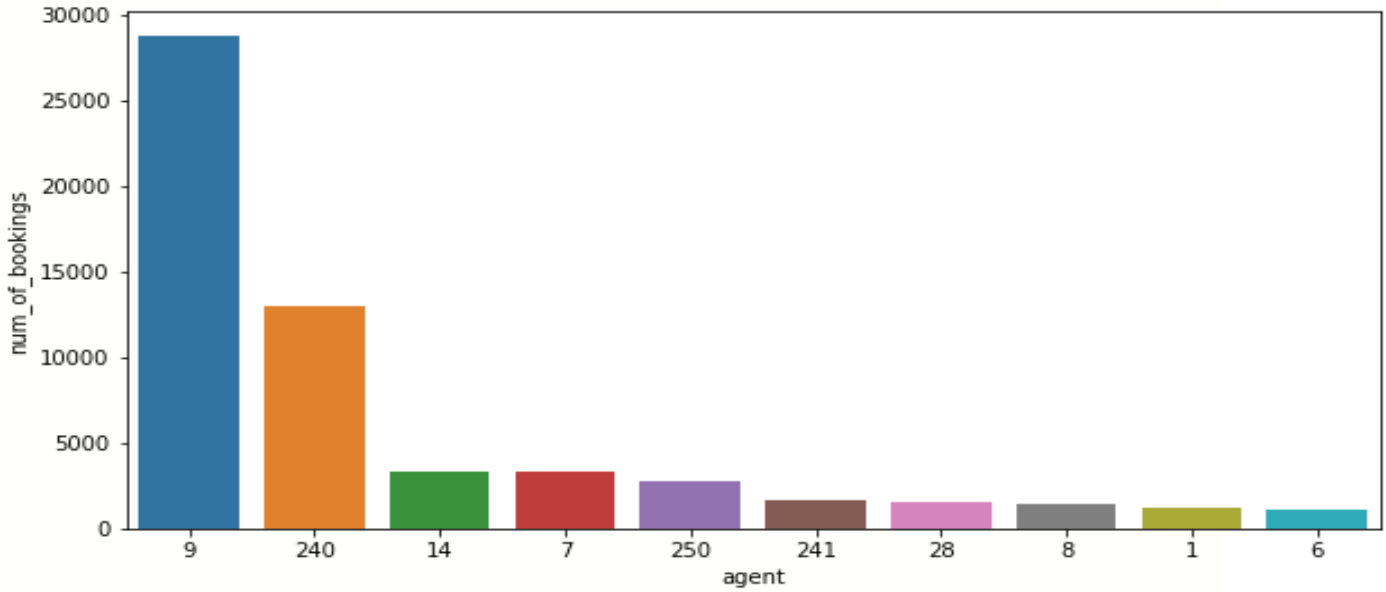
Percentage of meal preference

Most popular meal type is BB.



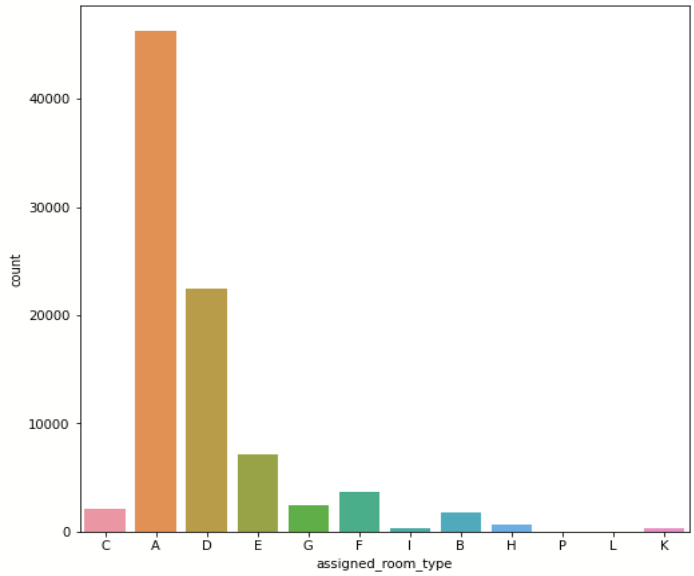
Which agent makes most no. of bookings?

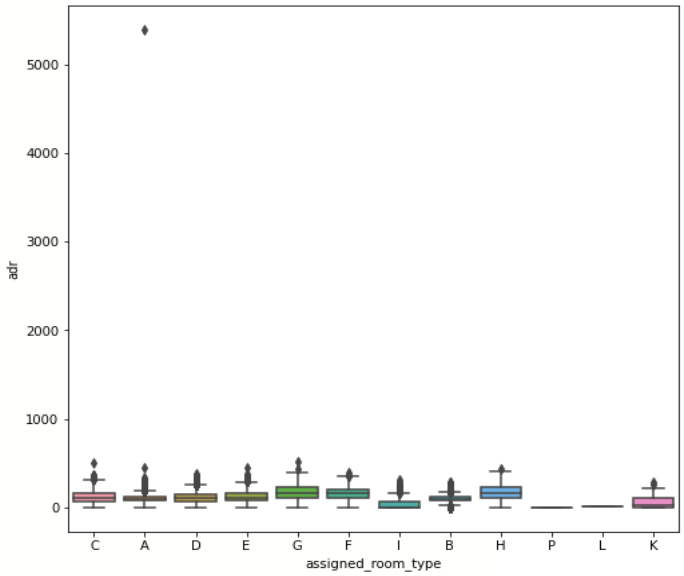
Agent no. 9 has made most no. of bookings.



Which room type is in most demand and which room type generatest highest adr?

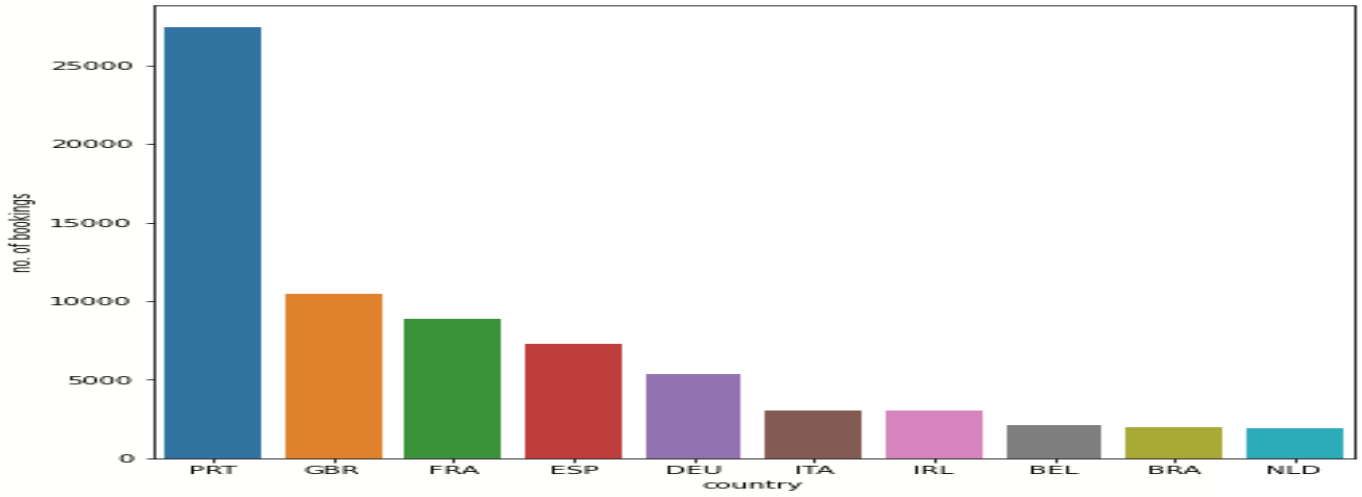
Most demanded room type is A, but better adr generating rooms F, G and E. Hotels should increase the no. of room types A and D to maximise revenue.





which countries most of the customers visit these hotels?

Most of the guests came from European countries, with highest number of guests from Portugal.



Conclusion

(1) Around 66.40% bookings are for City hotel and 33.60% bookings are for Resort hotel, therefore City Hotel is busier than Resort hotel. Also the overall adr of City hotel is slightly higher than Resort hotel.
(2) Mostly guests stay for less than 5 days in hotel and for longer stays Resort hotel is preferred.
(3) Both hotels have significantly higher booking cancellation rates and very few guests less than 3 % return for another booking in City hotel. 5% guests return for stay in Resort hotel.
(4) Most of the guests came from European countries, with most of guests coming from Portugal.
(5) Guests use different channels for making bookings out of which most preferred way is TA/TO.
(6) For hotels higher adr deals come via GDS channel, so hotels should increase their popularity on this channel.
(7) Almost 30% of bookings via TA/TO are cancelled.
(8) Not getting same room as reserved, longer lead time and waiting time do not affect cancellation of bookings. Although different room allotment do lowers the adr.
(9) July- August are the most busier and profitable months for both of hotels.
(10) Within a month, adr gradually increases as month ends, with small sudden rise on weekends.
(11) Couples are the most common guests for hotels, hence hotels can plan services according to couples needs to increase revenue.
(12) More number of people in guests results in more number of special requests.
(13) Bookings made via complementary market segment and adults have on average high no. of special request.
(14) For customers, generally the longer stays (more than 15 days) can result in better deals in terms of low adr.

Challenges

(1) There was a lot of duplicate data.
(2) Data was present in wrong datatype format.
(3) Choosing appropriate visualization techniques to use was difficult.
(4) A lot of null values were there in the dataset.