Capstone Project-1

Hotel Booking Analysis

Team Members

Saurabh Daund

Kamya Malhotra

Vaishnavi

Table of contents:

Problem Statement

Hotel industry is a very volatile industry and the bookings depend on variety of factors such as type of hotels, seasonality, days of week and many more which makes analyzing the patterns available in the past data more important to help the hotels plan better for the future bookings.

Using the historical data, hotels can perform various campaigns to boost the business. We can use data visualisation to predict various insights to grow the hotel industry.

We will be using the data available to analyze the factors affecting the hotel bookings. These factors can be used for reporting the trends and predict the future bookings.

Examining the dataset

- Our Hotel Booking Analysis Dataset has total 1,19,390 rows and 32 columns where we observed 4 null values in children column, 488 in country column, 16,340 in agent column and 112,593 in company column, resp.

- We have the following columns for which the further analysis is done:

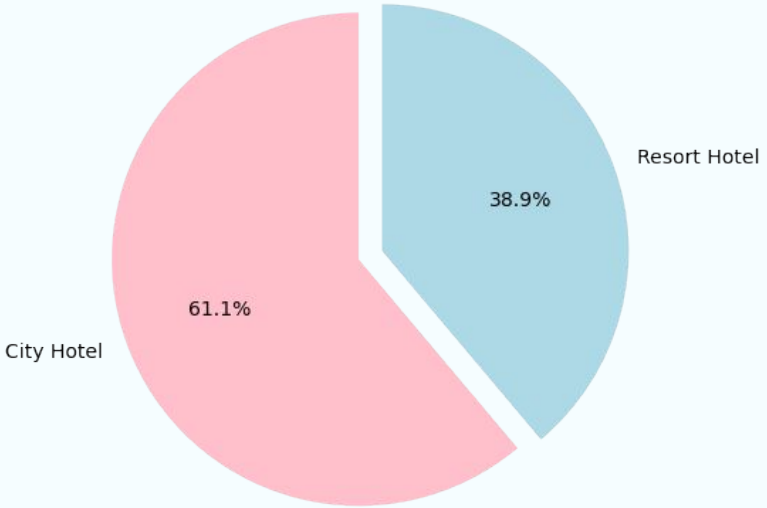
  hotel, is_canceled, lead_time, arrival_date_year, arrival_date_month, arrival_date_week_number,arrival_date_day_of_month, stays_in_weekend_nights, stays_in_week_nights,adults, children, babies, meal, country, market_segment, distribution_channel, is_repeated_guest, previous_cancellations, previous_bookings_not_canceled, 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.

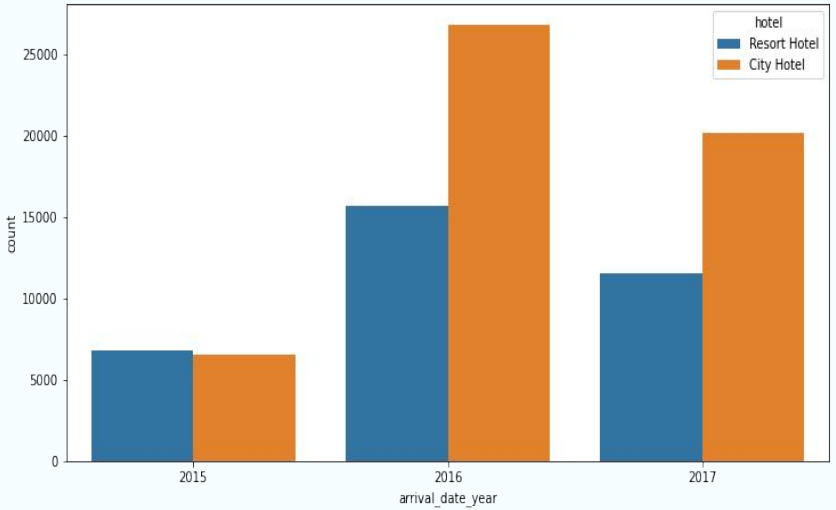
Preparing the Dataset

- We import various Python inbuilt libraries - NumPy, Pandas, Matplotlib and Seaborn to work on the dataset.
- Since most of the null values are present in columns - agent and company therefore, dropping these columns won't affect our dataset.
- We drop 31,994 duplicate values from our dataset for clear understanding.
- Next, we deal with outliers using Boxplot where we get a clear picture about the outliers in numeric columns.
- Replace the outliers with some specific values to prepare our dataset for further visualisation.

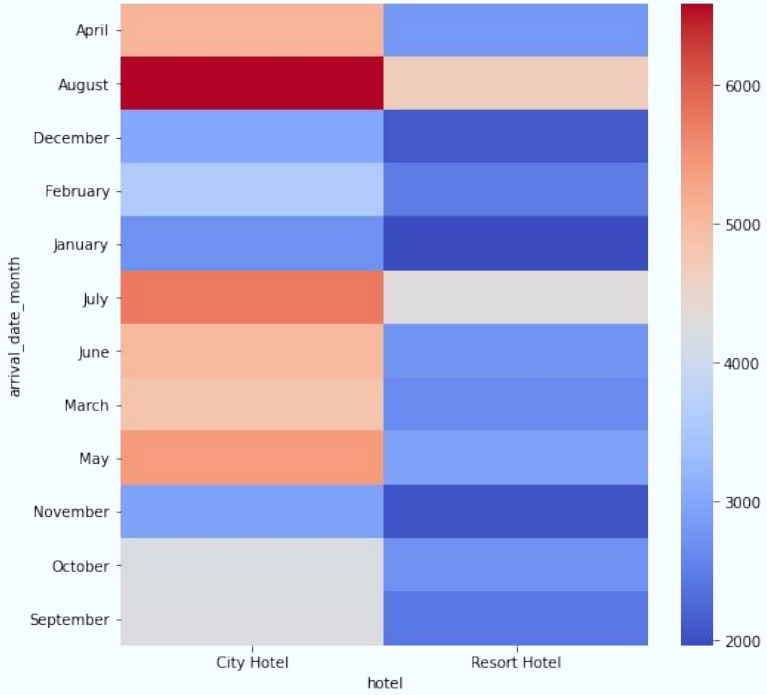
Exploratory Data Analysis

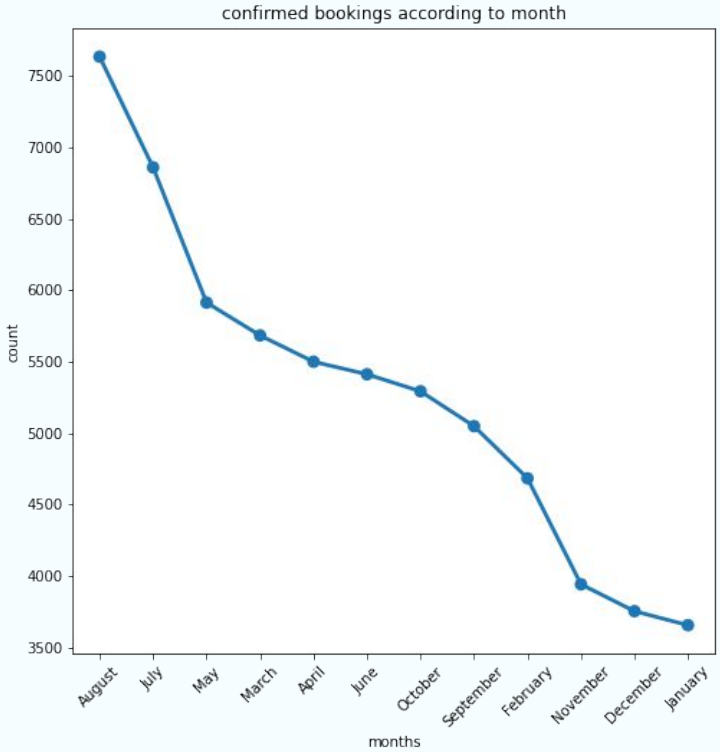
Let's do some data visualisation to see the trends!!





- **61.1% bookings is done for City Hotel and 38.9% for Resort Hotel with most hotels booked in year 2016 and least number of hotels booked in year 2015.**

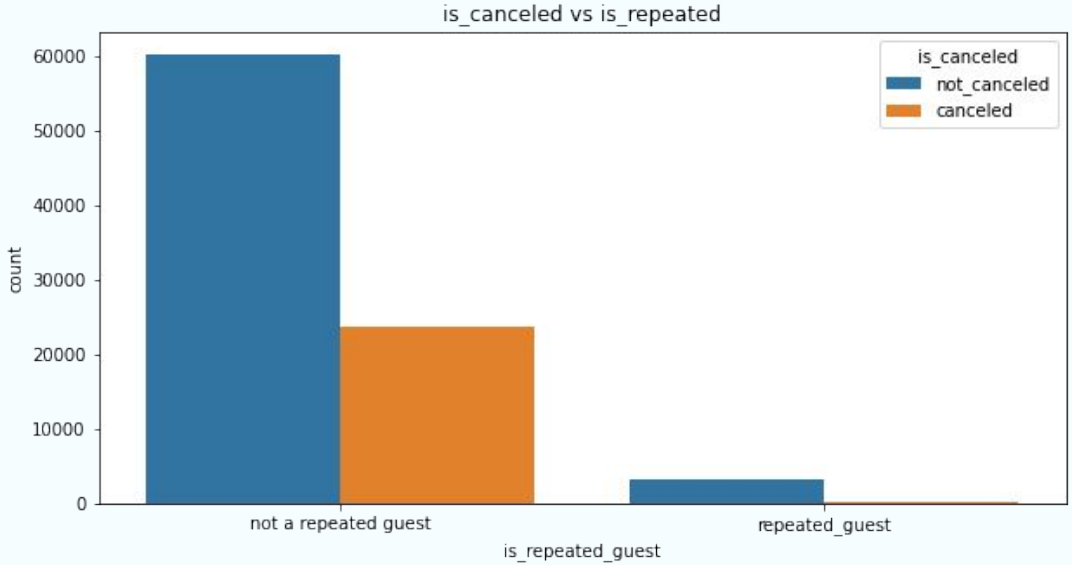




- **The above figures show that City Hotel and Resort Hotel has the highest number of confirmed bookings in the month of August and least number of confirmed bookings in the month of January.**

Some Hypothesis Questions...

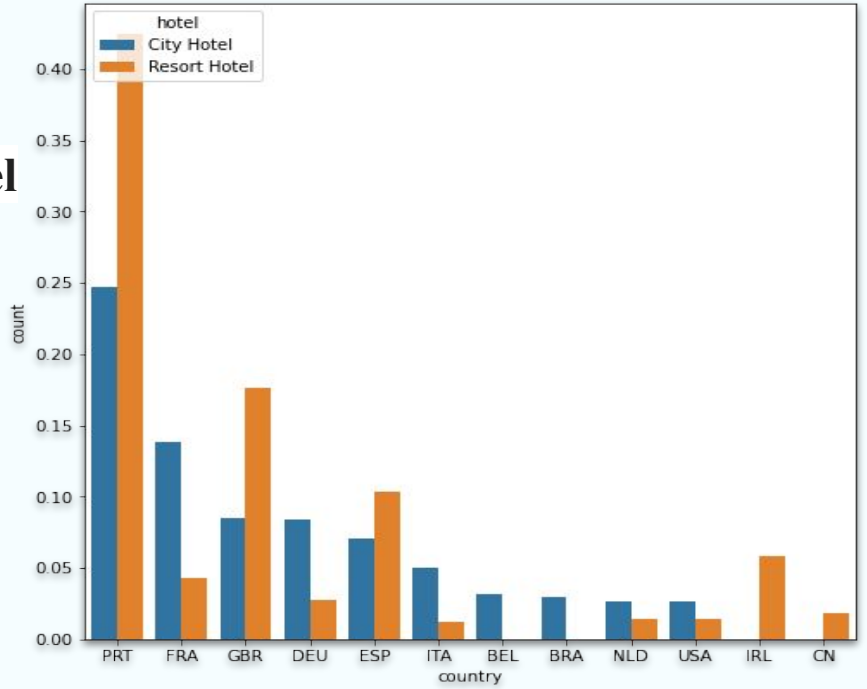
Q1. Whether the booking is canceled by a repeated guest or a non-repeated guest?



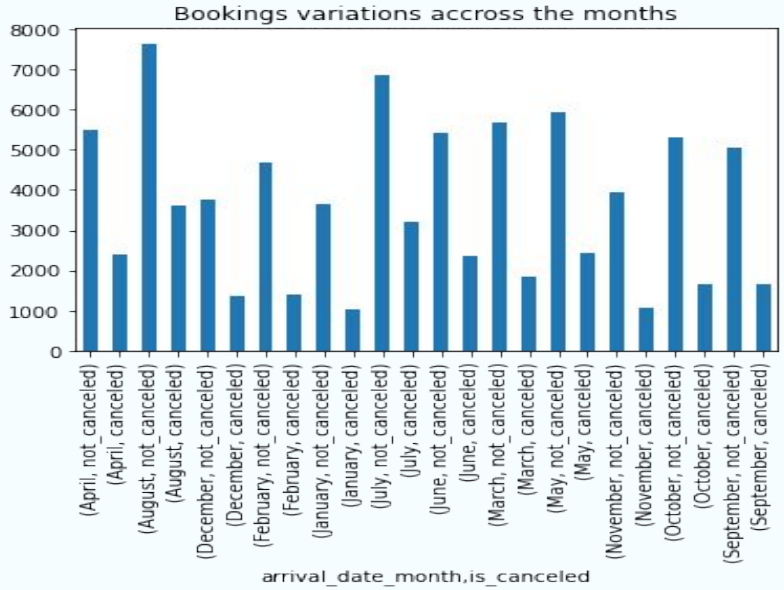
- **We can clearly see that maximum number of bookings are canceled by the guests who booked the hotel for the first time, percentage of canceled by repeated guest is almost zero.**
- **That means, the hotel booking that is canceled is very less likely to be canceled by a repeated guest**.

Q2.Which type of hotel do other country people prefer, we look according to top 10 countries.

- **Despite maximum bookings are done for city type hotel but we see that, resort hotel type is maximum booked by the top 10 countries with Portugal having highest guests.**

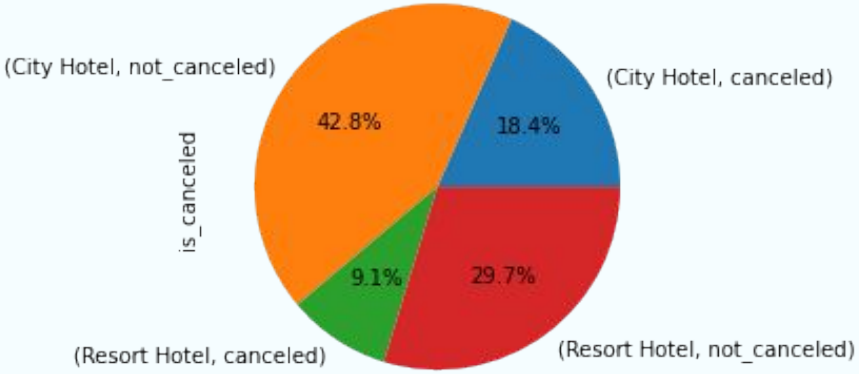


Q3. How cancellations vary according to month?



- **August is the month with maximum cancellations followed by July, May,June and April with lowest cancellations in January.**

Q4. How cancellations vary according to hotel type?

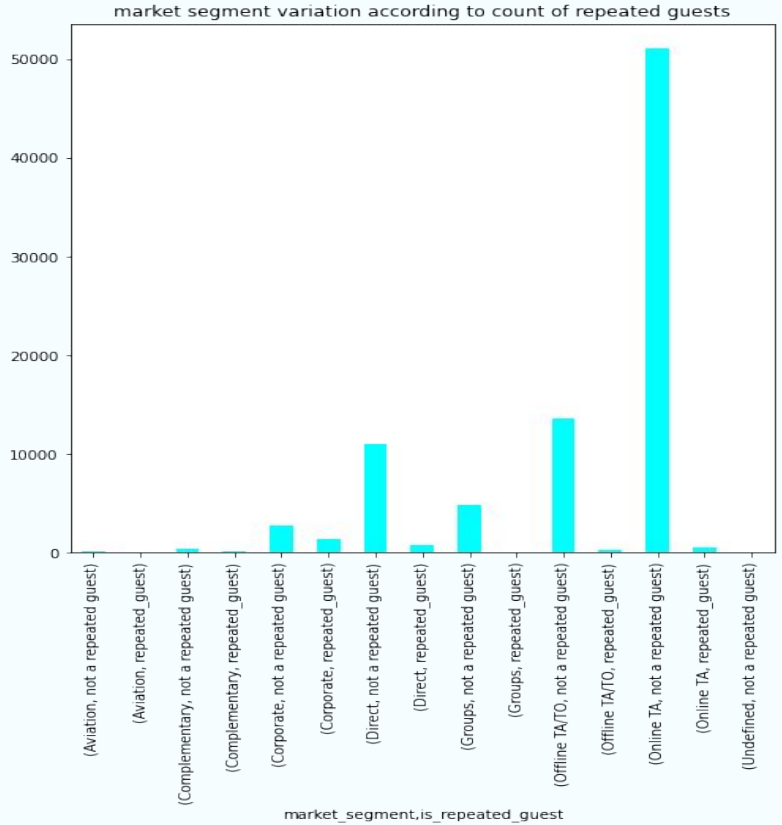


From the above plot we can conclude the following:

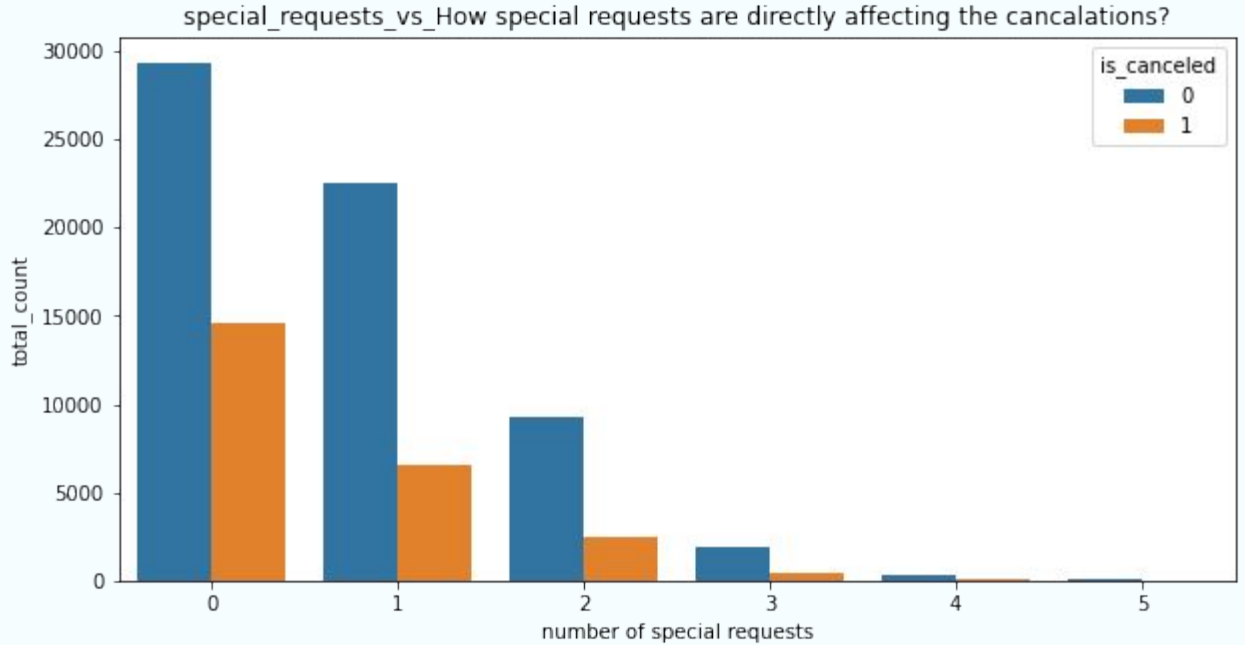
- **The cancellation percentage of City hotel is almost twice the cancellation percentage of Resort Hotel.**
- **The booking percentage of City Hotel is quite large than that of Resort hotel.**

Q5.Do market segment has any impact on repeated guests?

- **It can be seen from the graph that, (online TA/TO) is a segment through which maximum bookings are done, followed by (offline TA/TO).**
- **Online (TA/TO) is a segment with less than 1% repeated guests out of the total bookings of online (TA/TO) and maximum bookings are by new guests from which we conclude that repeated guests don't prefer online(TA/TO).**
- **Thus, from here we conclude that market segment do have an impact on repeated guests, though the segment with maximum bookings have barely any repeated guests.**

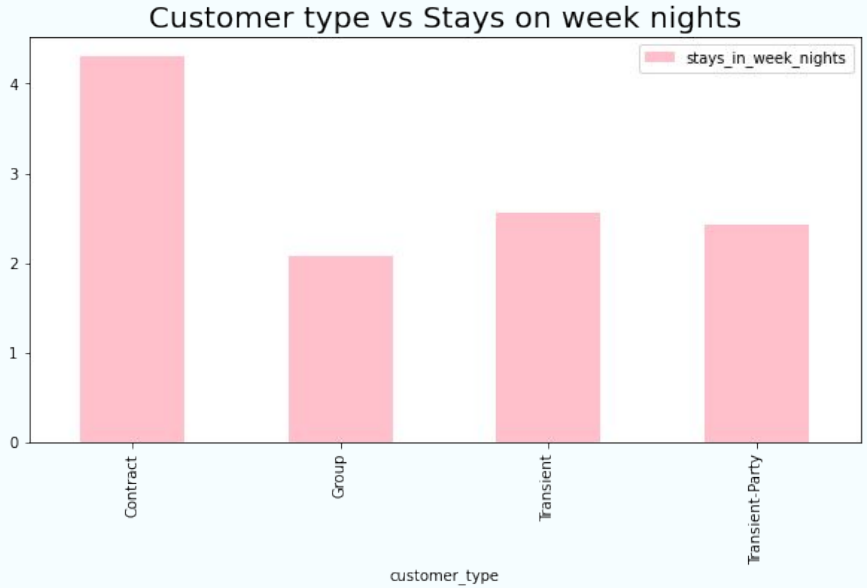


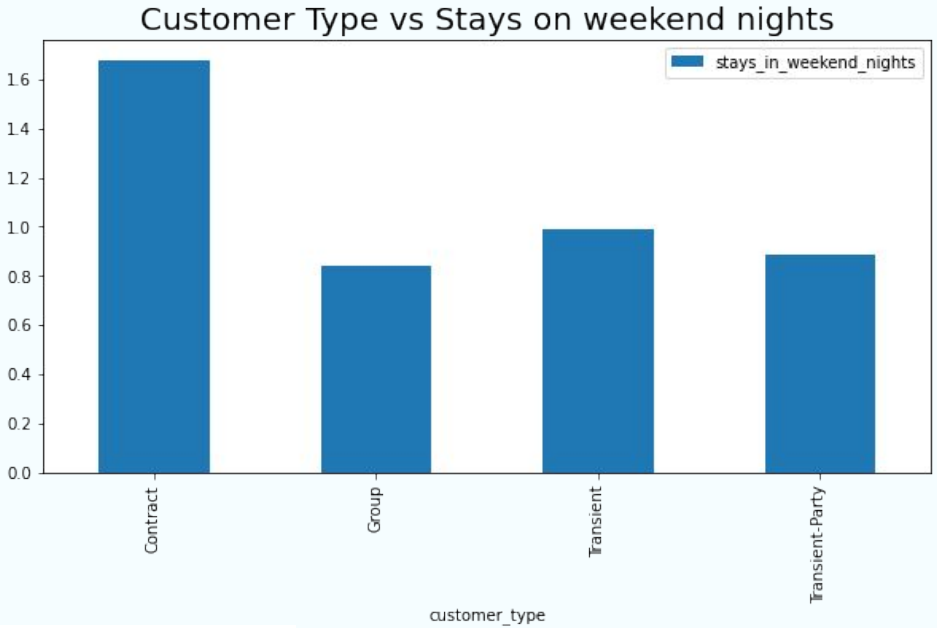
Q6.How special requests are directly affecting the cancellations?



- **Guest with 0 special requests are more likely to cancel the reservation.**
- **Special requests with 2,3,4 numbers have almost zero cancelations.**
- **Therefore, more the number of special requests, less are the number of cancellations.**

Q7.Customers prefer, week or weekend nights for the stay?

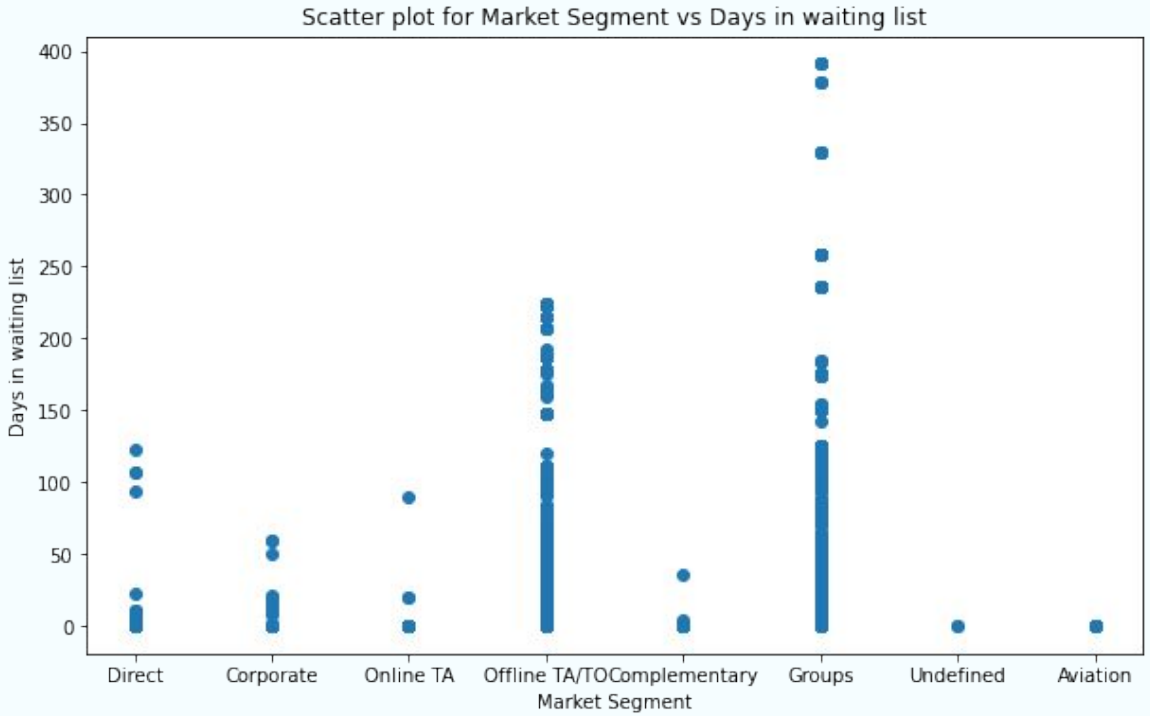




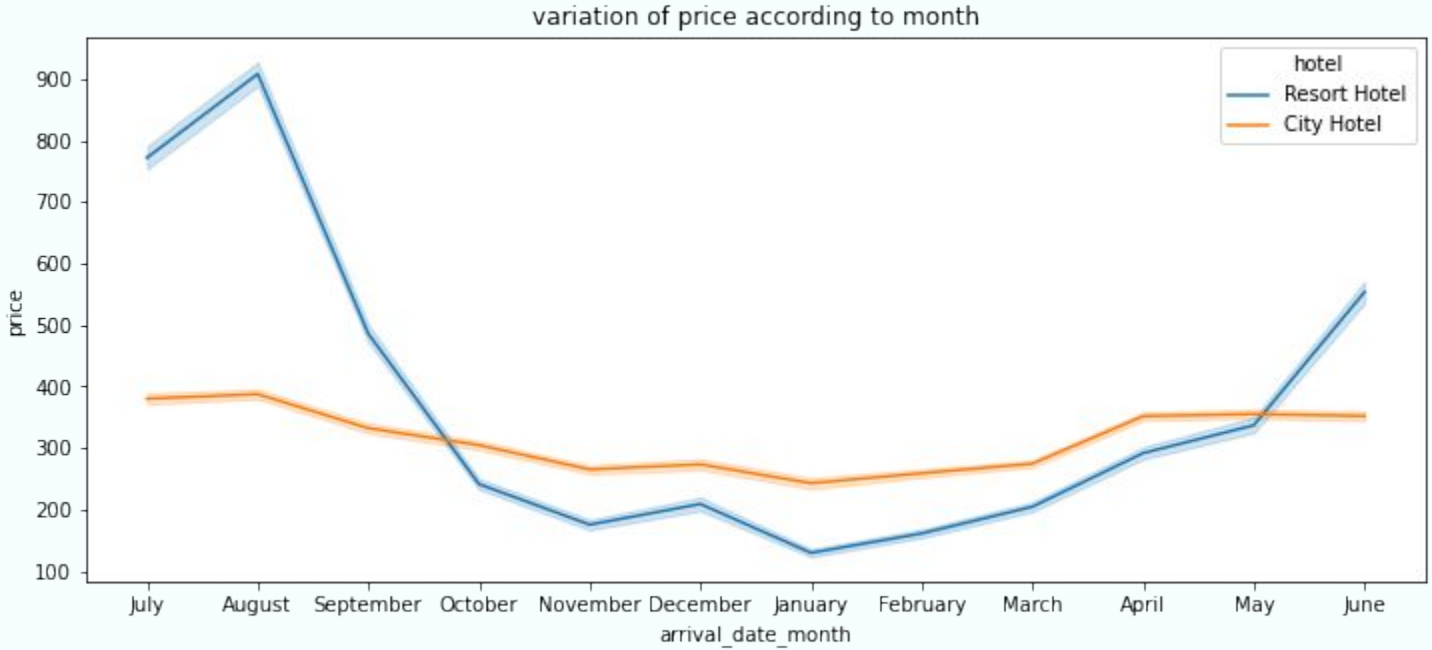
- **Maximum customers prefer stay in week nights rather than weekend nights.**
- **Maximum bookings are done for Contract type of customers followed by Transient.**
- **From the above graphs we can interpret that hotels need to focus on Group customers specially because they are very less in numbers and they also stays minimum number of nights w.r.t. other types of customers.**

Q8.How waiting time vary according to different market segment?

- From the graph, we can see that Aviation segment has the minimum waiting time as compared to other market segments.
- Whereas ,Group segment has the maximum waiting time of around 400 days.



Q9. How prices vary according to month?

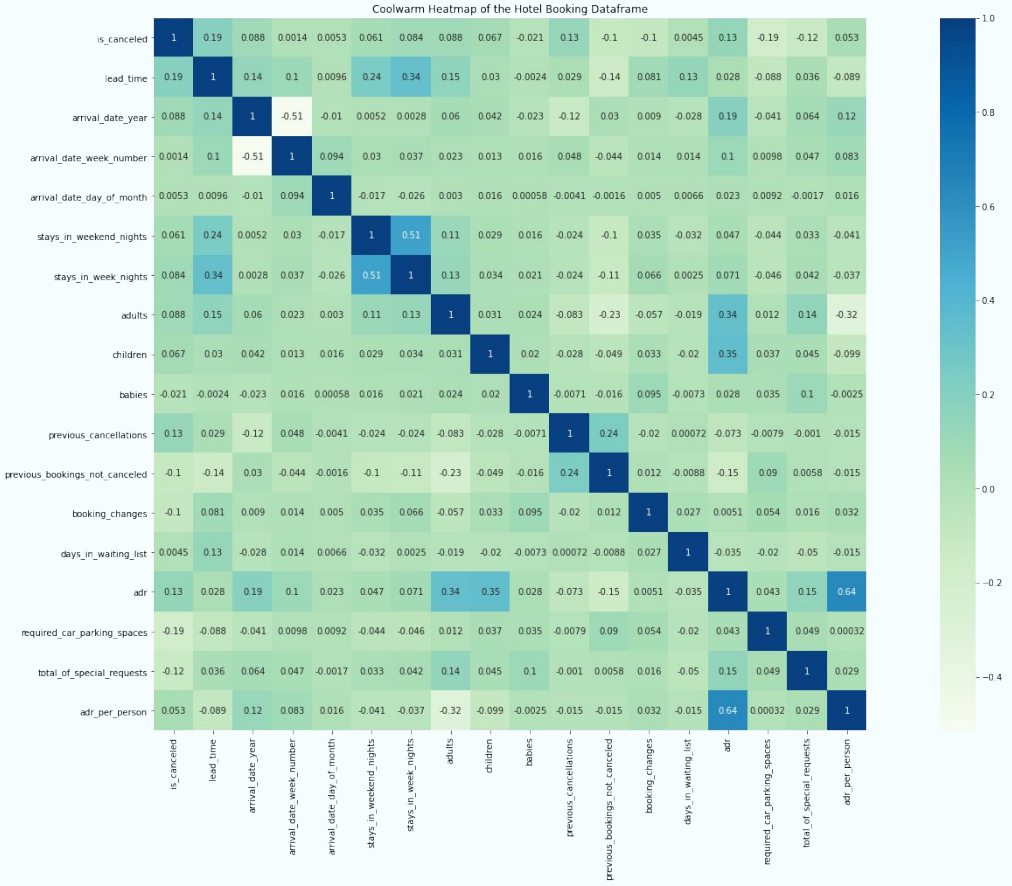


- **The basic conclusion we can make is that City hotel has almost constant price throughout the year whereas, Resort hotel has lot of fluctuations in the prices per month.**
- **Both, Resort hotel and City hotel has maximum price in August.**

Extracting the correlation of the whole dataset using Heatmap.

We get the following conclusions from the heatmap:

- lead_time is the only feature having highest positive correlation of 18% with is_canceled feature.
- Most of the features have insignificant level of correlation with is_canceled feature.
- is_canceled feature have significant level of negative correlation of 19% and 12% with required_car_parking_spaces and total_of_special_requests respectively.



Some of the predictions to increase the sales of Hotel industry

- Cancellation policies should be improved, by applying non refundable charges or some penalty.
- Percentage of online booking cancellation is more, encourage Direct Bookings by offering special discounts.
- Foriegn countries like Portugal, Great Britain, France have more visitor base, apply a marketing team to handle that customer base.
- Hotels should consider the maximum number of special requests from guests to reduce the possibility of cancellations which will eventually help in better customer experience.
- Keep an eye on which market segment cancellations are coming more and why.
- Months between May and August have maximum bookings, these months are peak in business expansion and look after more customer satisfaction.


THANK YOU