Exploratory and Predictive Data Analysis Report on INN Hotel Group Booking Cancellations

Arthur Xia
2023/07
GitHub Link

Contents:

Data Overview and Cleaning

A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests, but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. Such losses are particularly high on last-minute cancellations.

The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

The cancellation of bookings impacts a hotel on various fronts:

Loss of resources (revenue) when the hotel cannot resell the room. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms. Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin. Human resources to make arrangements for the guests.

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. INN Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. You as a data scientist must analyze the data provided to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

Here's an overview of the dataset:

- Entries: 36,275 rows
- Columns: 19

Here are the columns in the dataset:

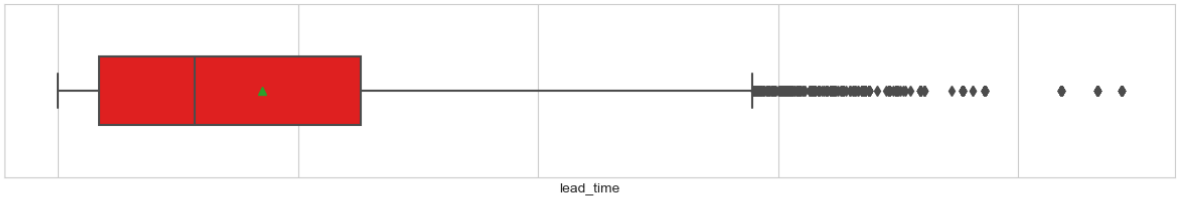
- Booking_ID: Identifier for the booking.
- no_of_adults: Number of adults in the booking.
- no_of_children: Number of children in the booking.
- no_of_weekend_nights: Number of weekend nights in the booking.
- no_of_week_nights: Number of weeknights in the booking.
- type_of_meal_plan: Type of meal plan selected.
- required_car_parking_space: Whether a car parking space is required (0 or 1).
- room_type_reserved: Type of room reserved.
- lead_time: Lead time in days.

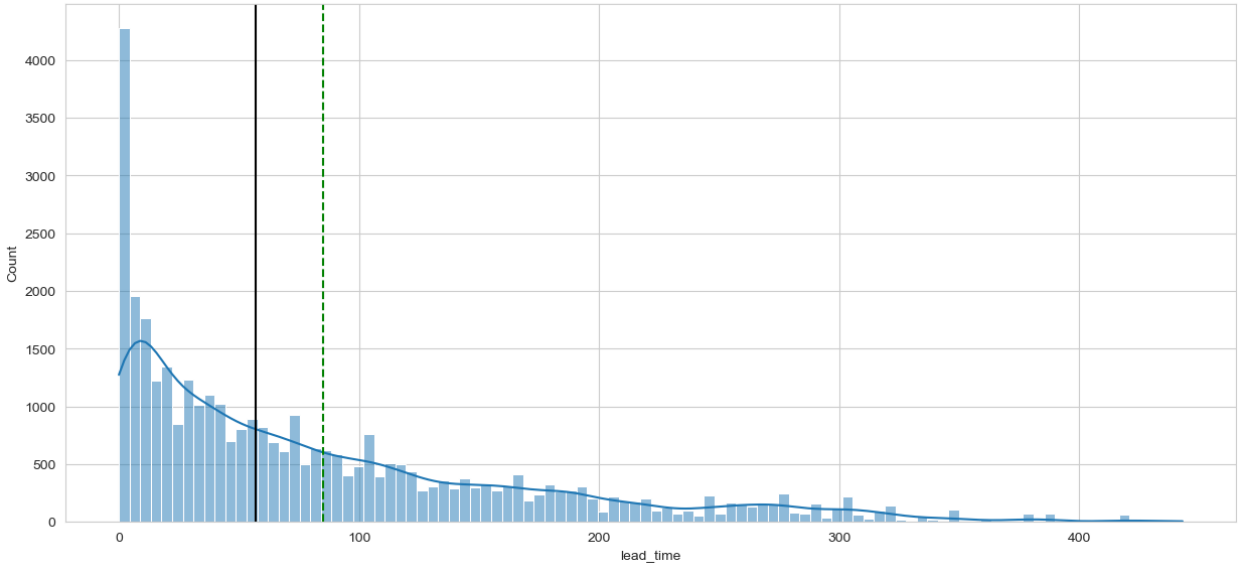
- arrival_year: Year of arrival.
- arrival_month: Month of arrival.
- arrival_date: Date of arrival.
- market_segment_type: Market segment type (e.g., Online, Offline).
- repeated_guest: Whether the guest is a repeated guest (0 or 1).
- no_of_previous_cancellations: Number of previous cancellations by the guest.
- no_of_previous_bookings_not_canceled: Number of previous bookings by the guest that were not canceled.
- avg_price_per_room: Average price per room.
- no_of_special_requests: Number of special requests made by the guest.
- booking_status: Status of the booking (e.g., Canceled, Not_Canceled).

Exploratory Data Analysis

Univariate Analysis

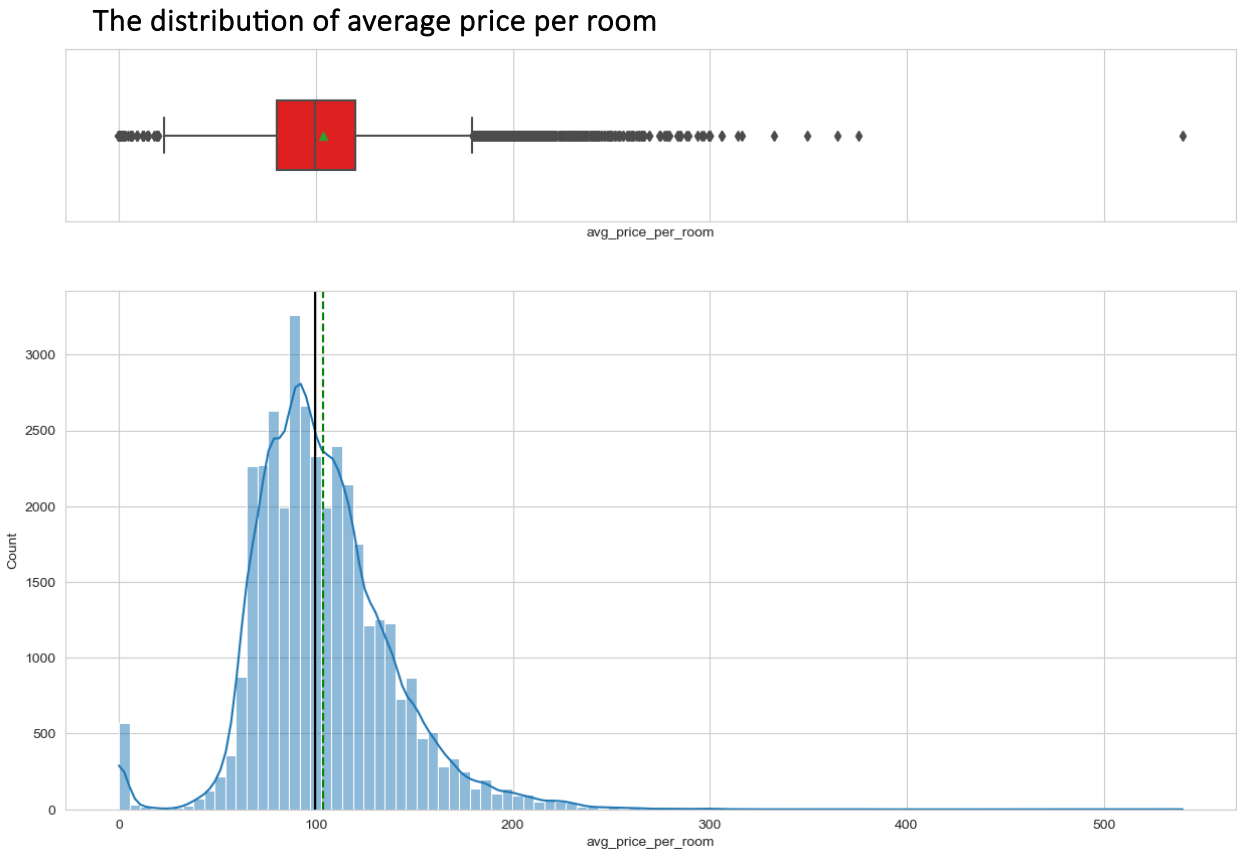
The distribution of lead time





- The distribution of lead time is right-skewed, and there are many outliers.
- Some customers made bookings around 500 days in advance, but most of the customers made bookings around 0-200 days in advance.
- Many customers have made the booking on the same day of arrival as well.

Observations on average price per room



- The distribution of the average price per room is skewed to the right. There are outliers on both sides.
- The average price of a room is around about **100** Euros.
- There is 1 observation where the average price of the room is more than **500** euros. This observation is quite far away from the rest of the values.
- Instead of dropping it, we will clip this to the upper whisker

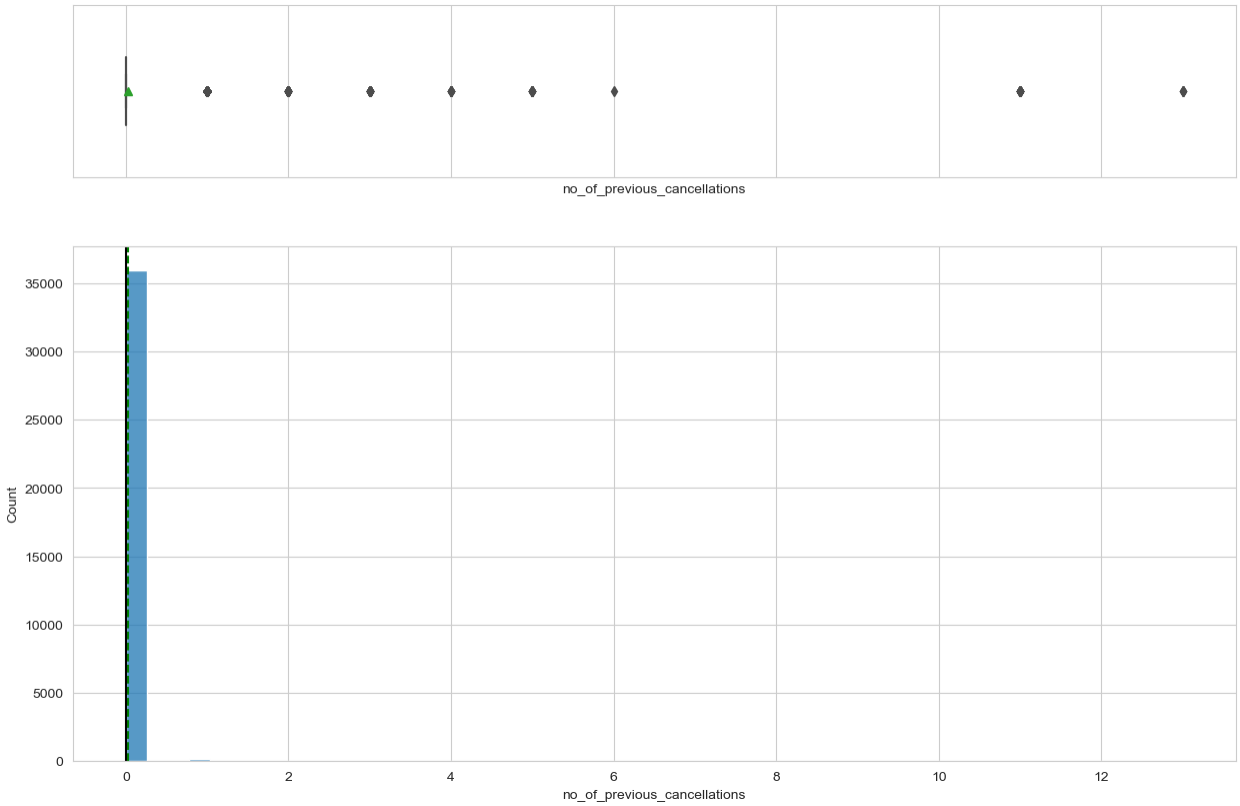
(Q3 + 1.5- IQR) = 179.55

- Interestingly some rooms have a price equal to 0. Let's check them.
- And we find that there are 545 observations where the average price of the room is 0.
- In the market segment column, it looks like many of the observations are complementary.

	market_segment_type
Complementary	354
Online	191

Observations on the number of adults

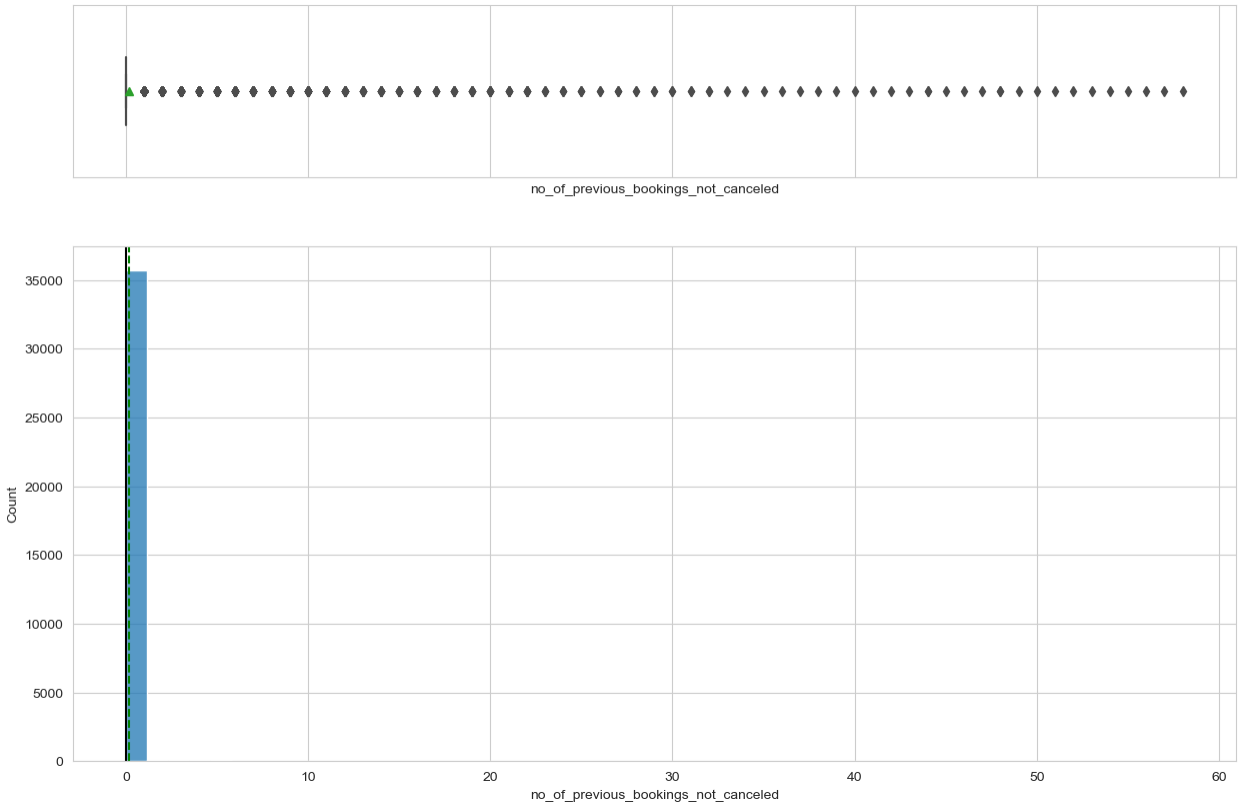
The distribution of the number of adults



- Very few customers have more than **one** cancellation.
- Some customers canceled more than **12** times.

Observations on the number of previous bookings not canceled.

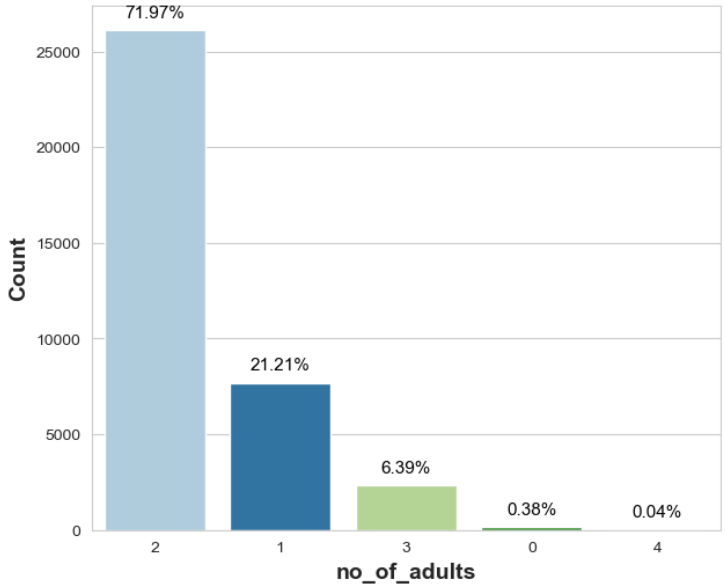
The distribution of the number of previous bookings not canceled.



- Very few customers have more than **1** booking not canceled previously.
- Some customers have not canceled their bookings around **60** times.

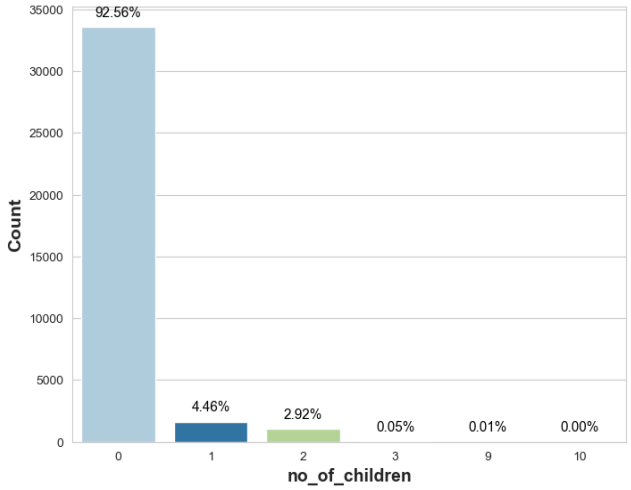
Observations on the number of adults

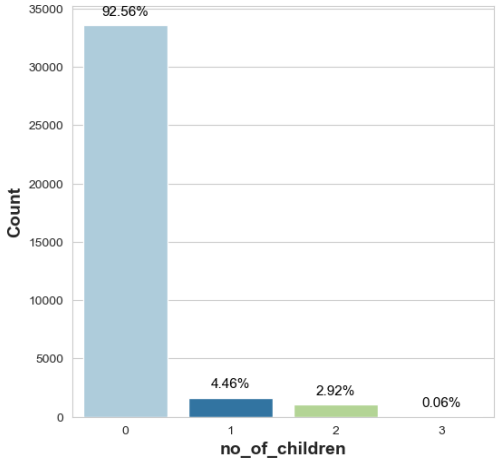
The distribution of the number of adults



- **72%** of the bookings were made for **2** adults.

replacing the values of the number of adults greater than 4.

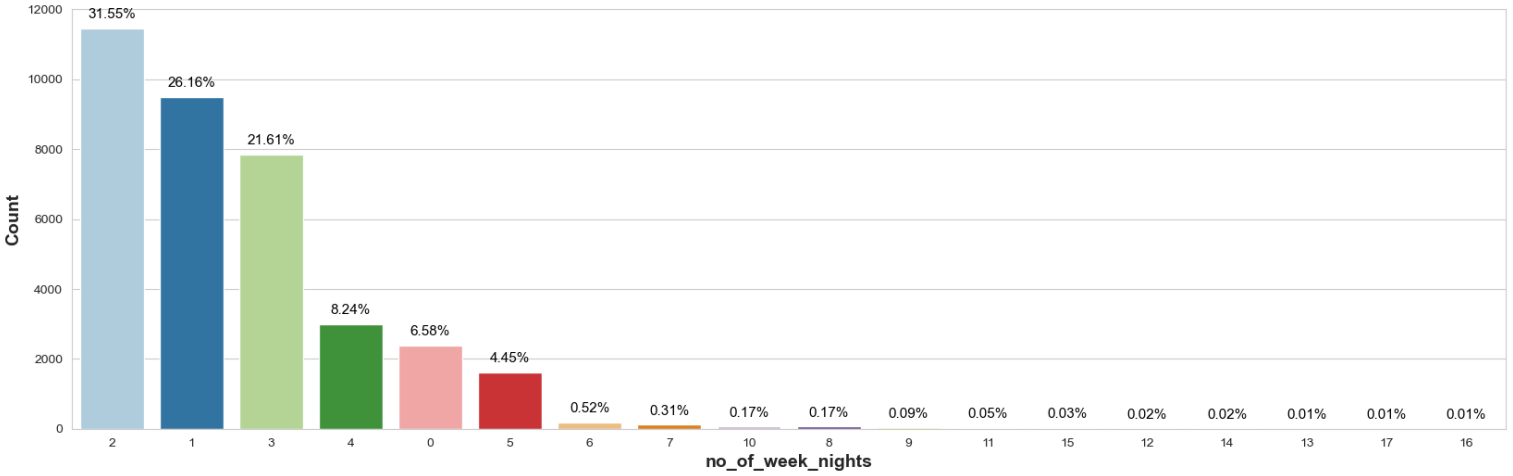




- 92% of the bookings were made for 0 children.
- 6% of the bookings were made for 1 or more children.
- There are some values in the data where the number of children is 9 or 10, which is highly unlikely.
- We will replace these values with the maximum value of 3 children.

Observations on the number of weeknights

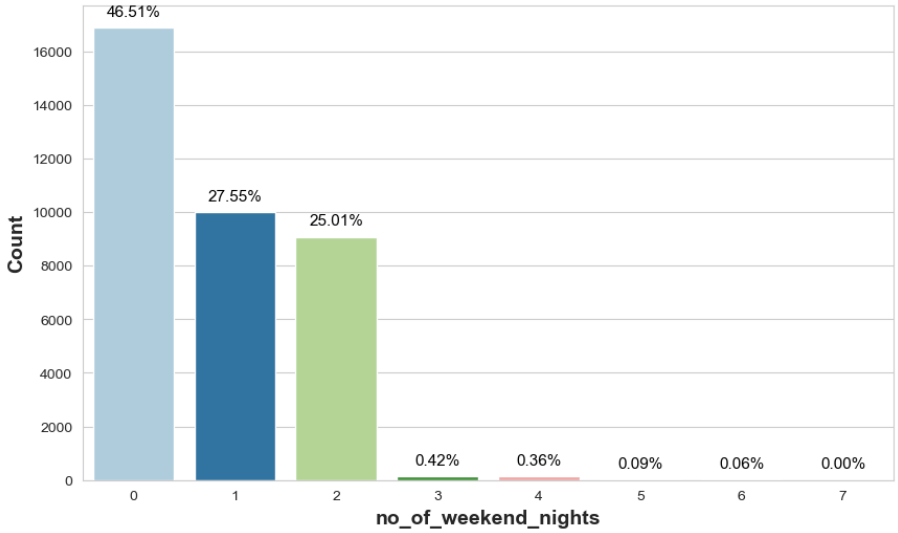
The distribution of the number of weeknights



- Most bookings are made for 2 nights (31.5%) followed by 1 night (26.2%).
- A very small proportion of customers made the booking for more than 10 days.

Observations on the number of weekend nights

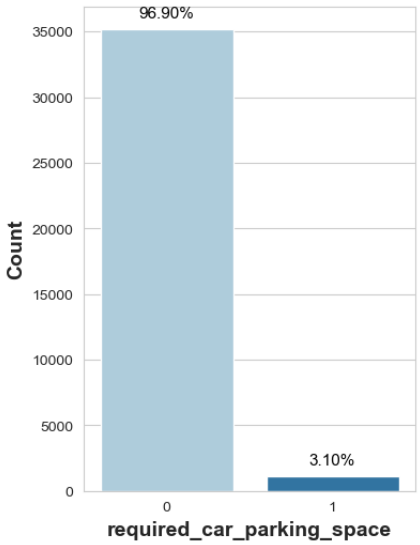
The distribution of the number of weekend nights



- 46.5% of the customers do not plan to spend the weekend in the hotel.
- The percentage of customers planning to spend 1 or 2 weekends in the hotel is almost the same.

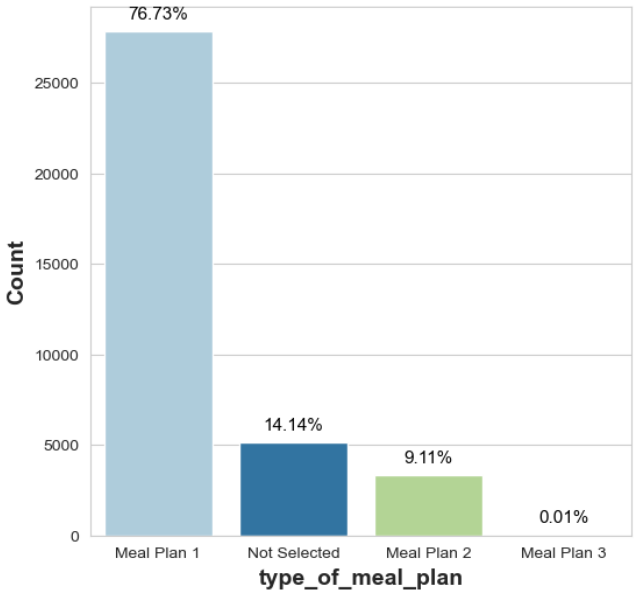
Observations on required car parking space

The distribution of required car parking space



- 96.90% of the customers do not require car parking space.

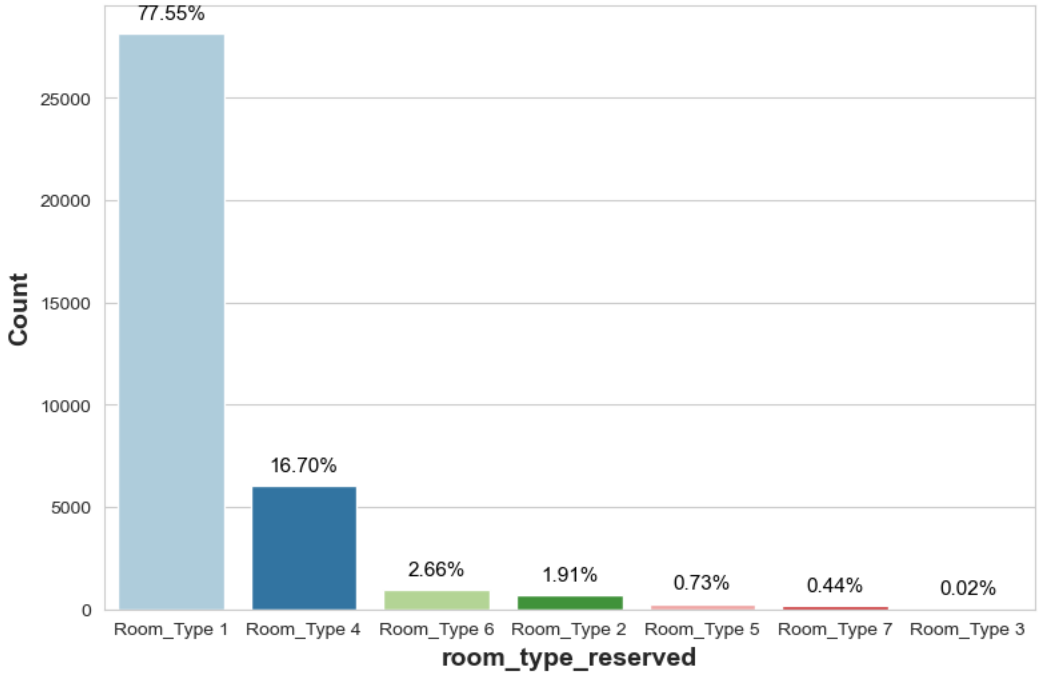
Observations on the type of meal plan



- Most of the customers prefer meal plan 1 which is only breakfast.
- 14.1% of the customers didn't select a meal plan.

Observations on room type reserved.

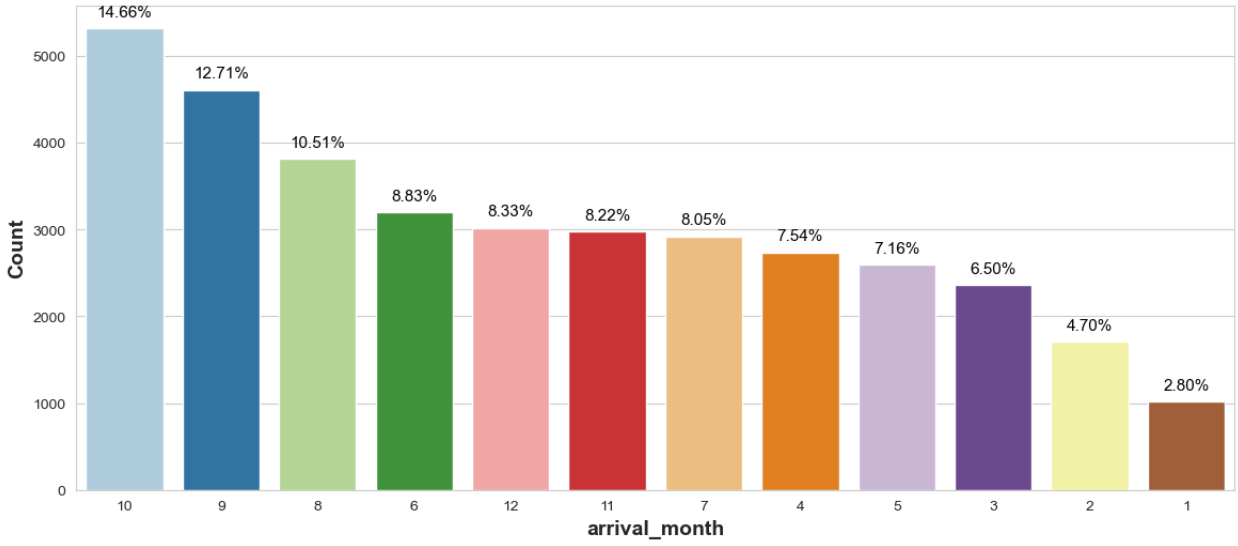
The distribution of room types reserved



- Around 77% of the customers booked Room_Type 1 followed by 17% of the customers booking Room_Type 4.

Observations on arrival month

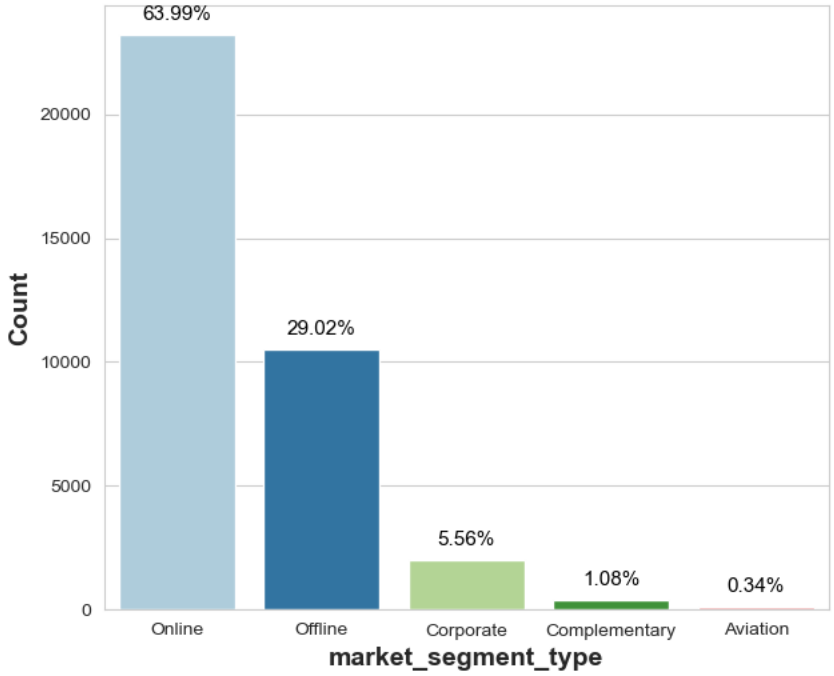
The distribution of arrival month



- October is the busiest month for the hotel followed by September.
- 14.7% of the bookings were made in October.

Observations on market segment type

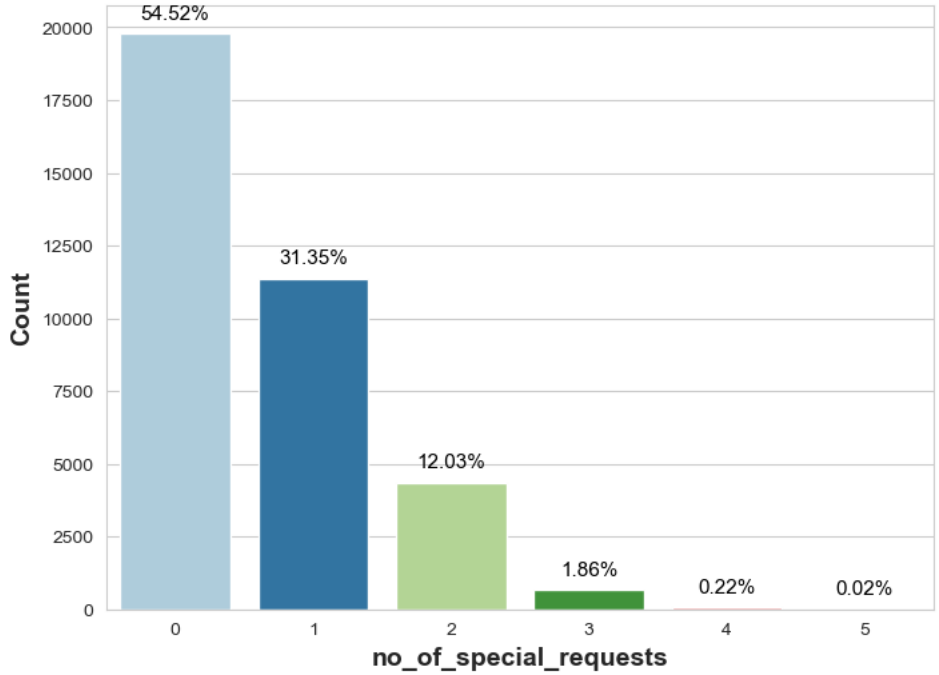
The distribution of market segment type



- 64% of the bookings were made online.

Observations on the number of special requests

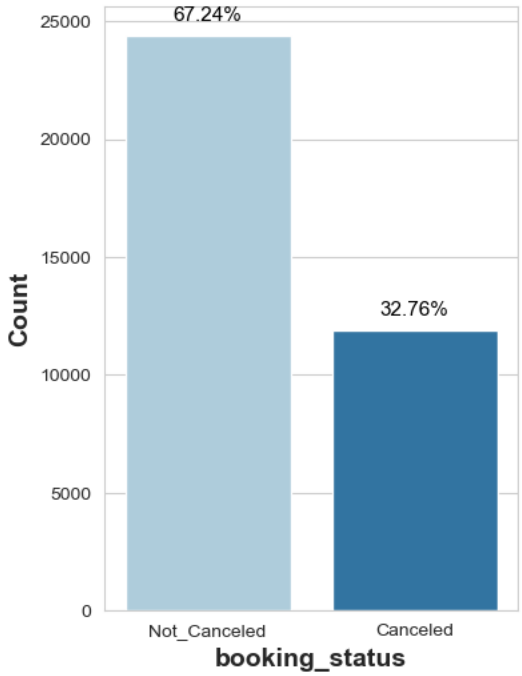
The distribution of the number of special requests



- 54.5% of the customers generally do not make any requests while booking a hotel room.

Observations on booking status

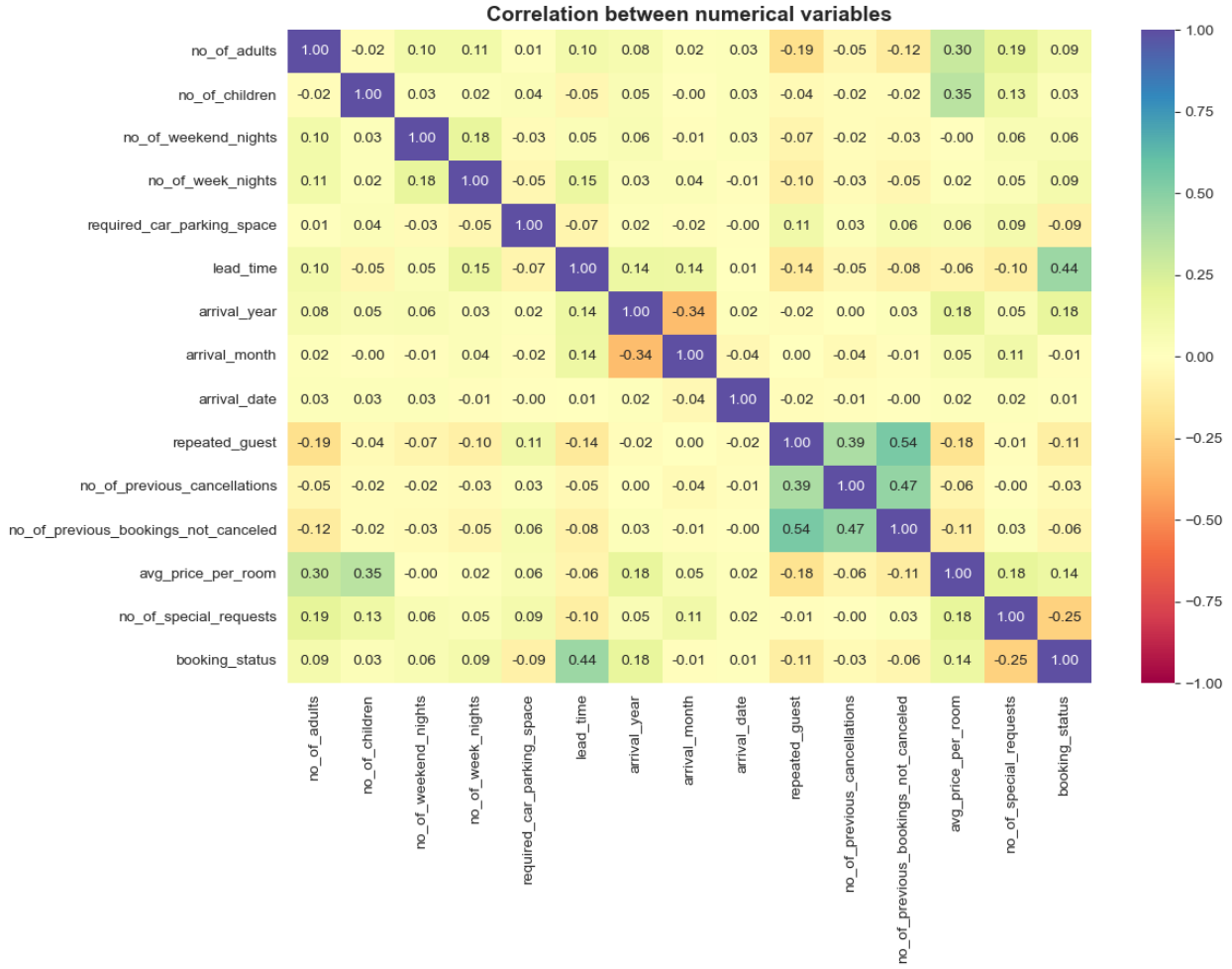
The distribution of booking status



- 32.8% of the bookings were canceled by the customers.

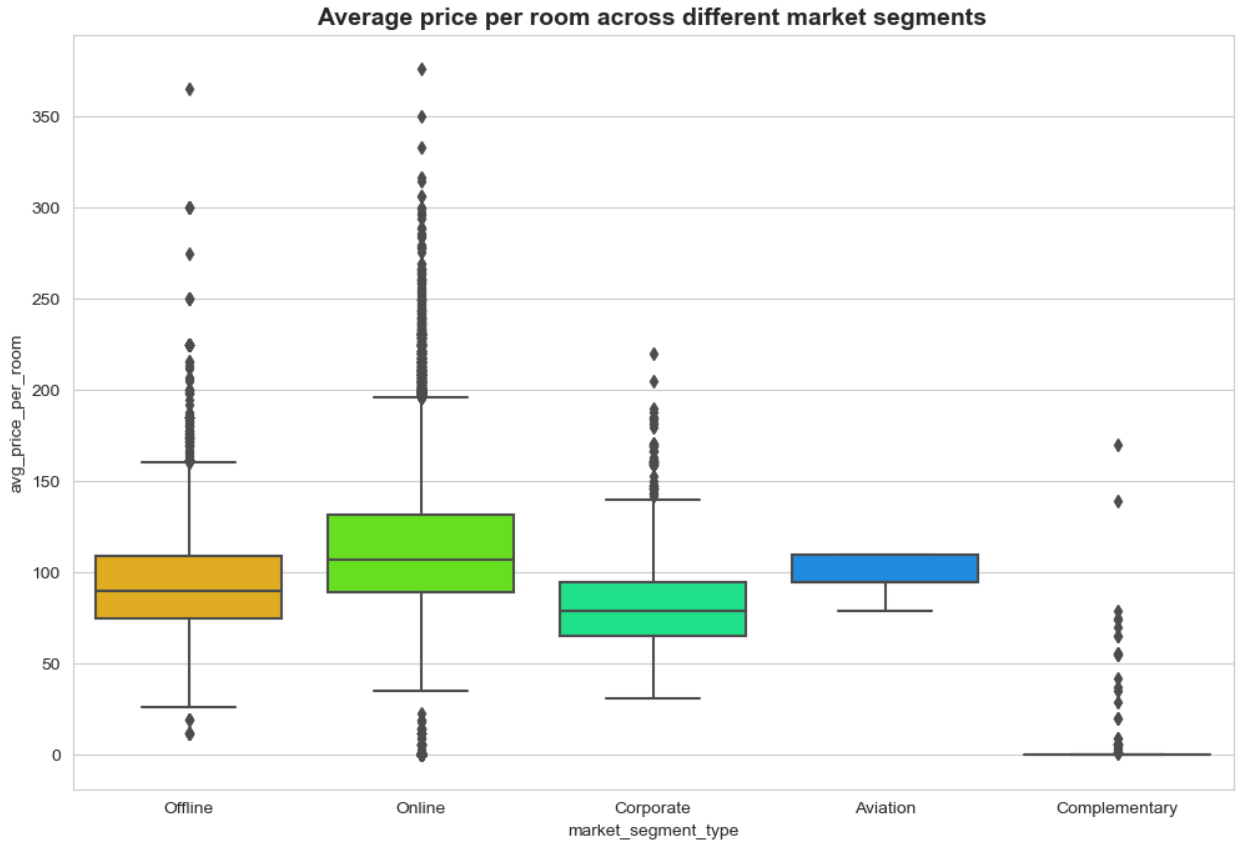
Bivariate Analysis

The correlation between numerical variables



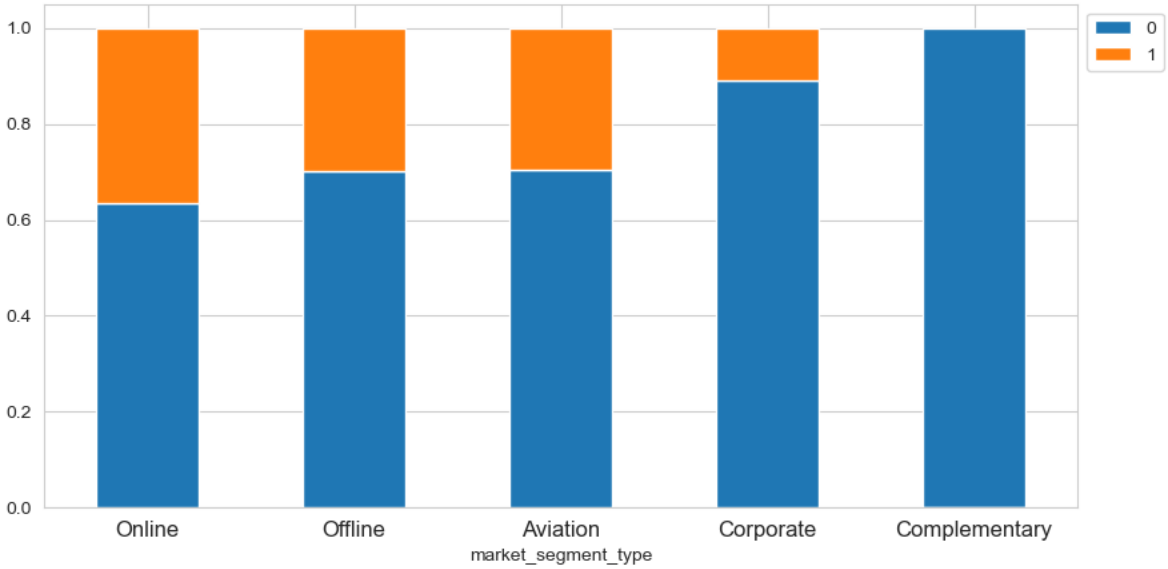
- There's a positive correlation between the number of customers (adults and children) and the average price per room.
- This makes sense as the more the number of customers more rooms they will require thus increasing the cost.
- There's a negative correlation between average room price and repeated guests. The hotel might be giving some loyalty benefits to the customers.
- There's a positive correlation between the number of previous bookings canceled and previous bookings not canceled by a customer and repeated guest.
- There's a positive correlation between lead time and the number of weeknights a customer is planning to stay in the hotel.
- There's a positive correlation between booking status and lead time, indicating higher the lead time higher are the chances of cancellation. We will analyze it further.
- There's a negative correlation between the number of special requests from the customer and the booking status, indicating if a customer has made some special requests the chances of cancellation might decrease. We will analyze it further.

Hotel rates are dynamic and change according to demand and customer demographics. Let's see how prices vary across different market segments.



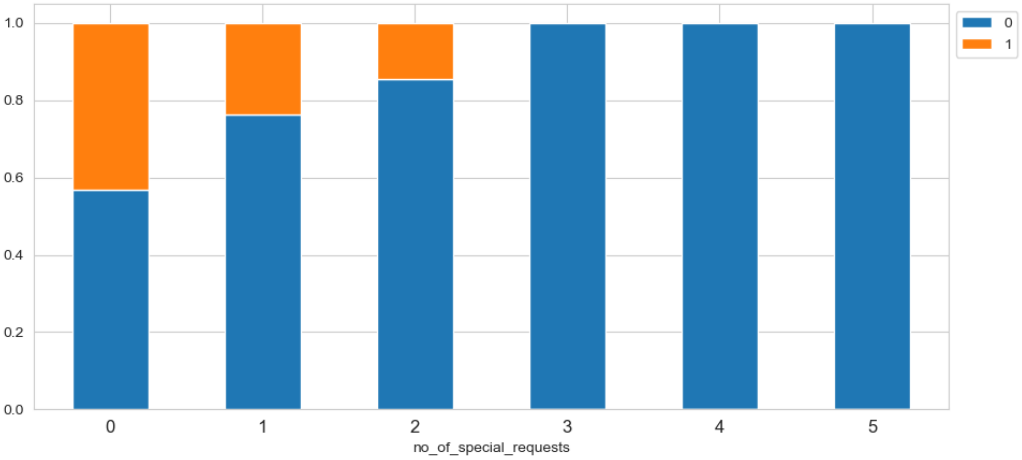
- Rooms booked online have high variations in prices.
- The offline and corporate room prices are almost similar.
- The complementary market segment gets the rooms at very low prices, which makes sense.

Let's see how booking status varies across different market segments. Also, how average price per room impacts booking status?



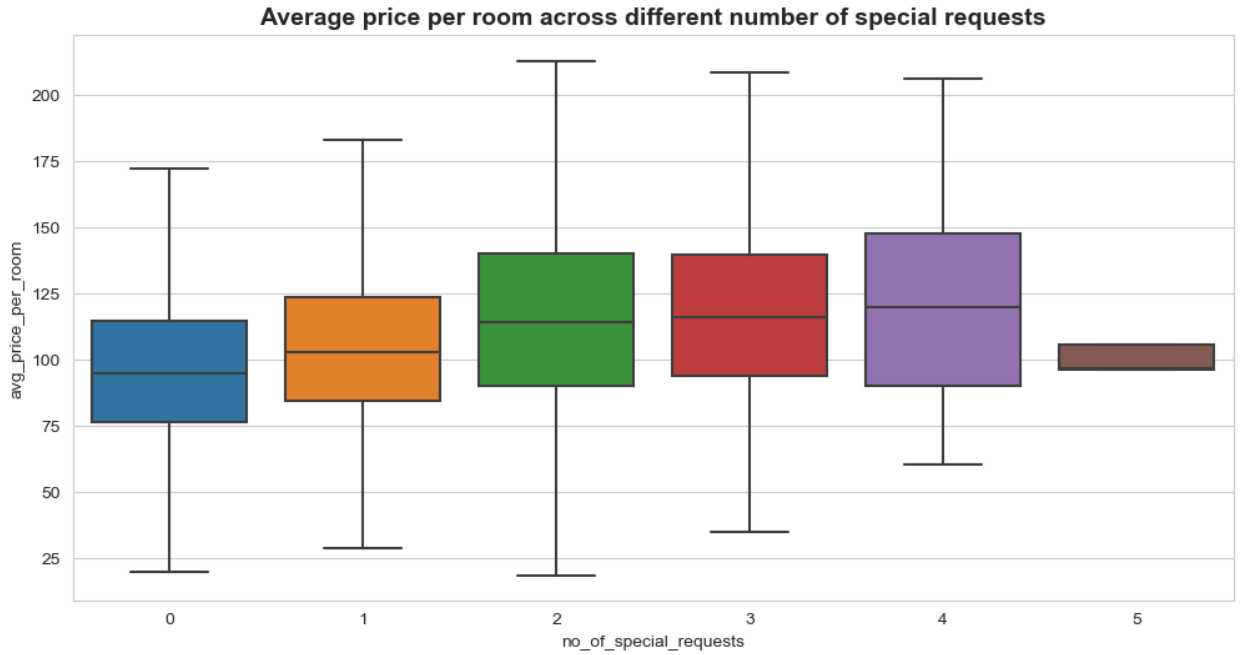
- Around 40% of the online booking were canceled.
- Bookings made offline are less prone to cancellations.
- Corporate segment shows very low cancellations.

Many guests have special requirements when booking a hotel room. Let's see how it impacts cancellations



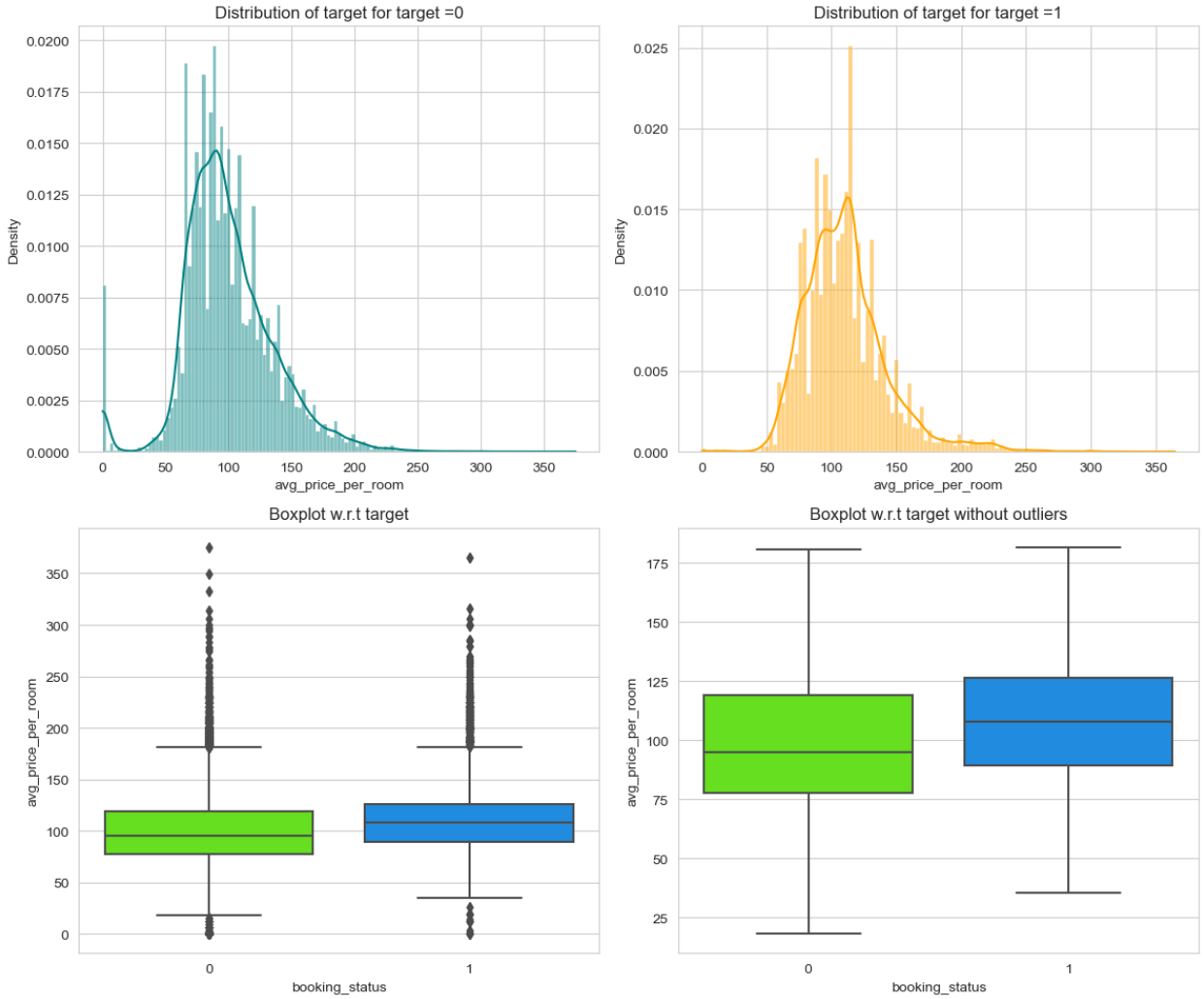
- If a customer has made more than 2 requests there's a very high chance that the booking will not be canceled.

Let's see if the special requirements made by the customer impacts the price per room



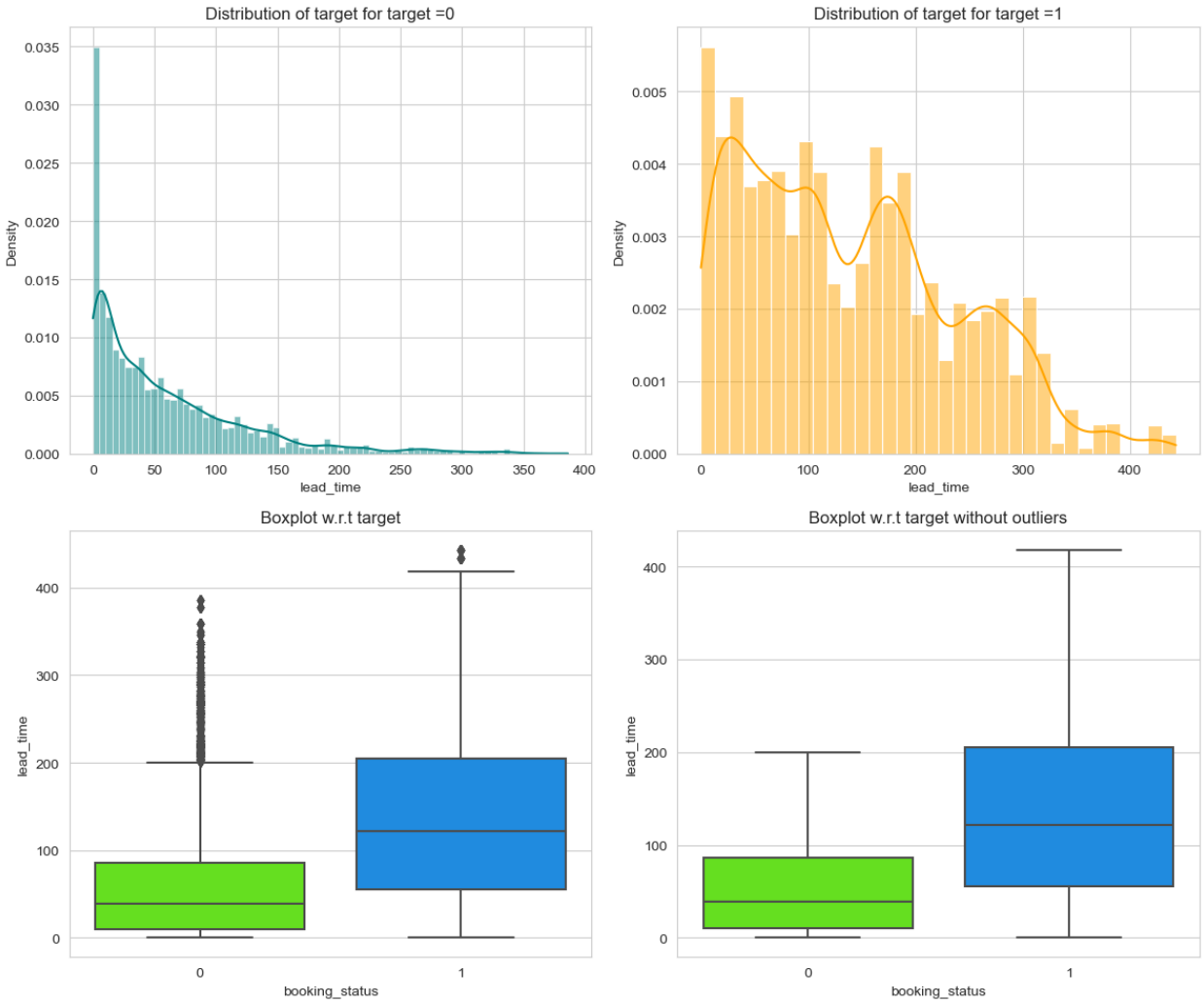
- The median prices of the rooms where some special requests were made by the customers are slightly higher than the rooms where customer didn't make any requests.

We saw earlier that there is a positive correlation between booking status and average price per room. Let's analyze it.



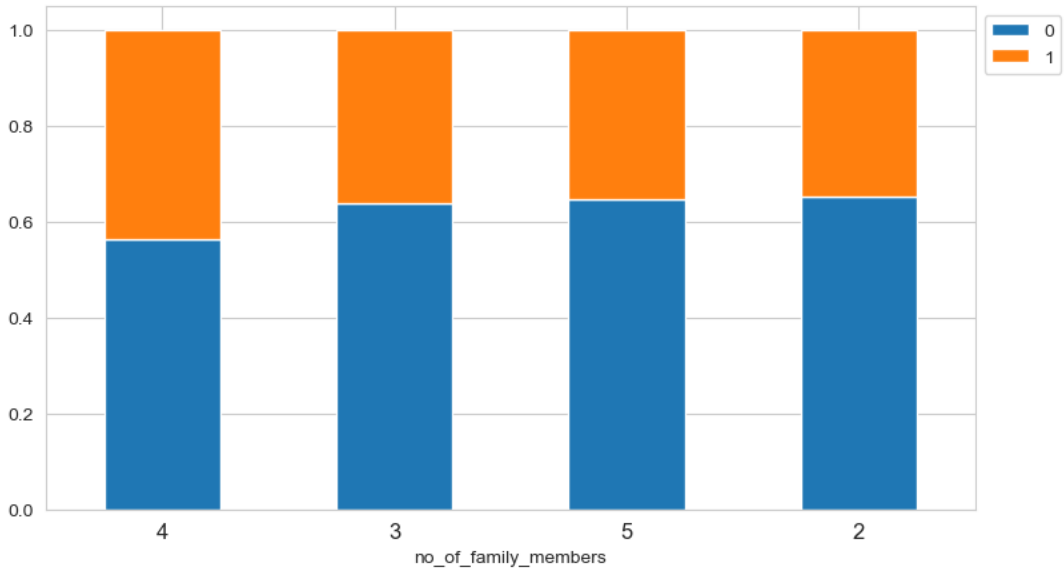
- The distribution of price for canceled bookings and not canceled bookings is quite similar.
- The prices for the canceled bookings are slightly higher than the bookings which were not canceled.

There is a positive correlation between booking status and lead time also. Let's analyze it further.



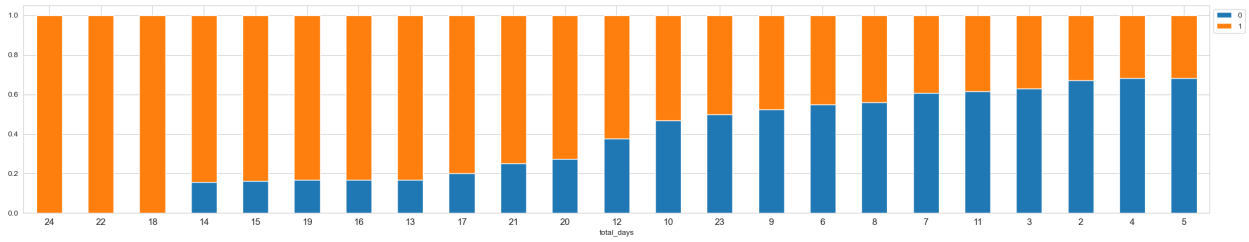
- There's a big difference in the median value of lead time for bookings that were canceled and bookings that were not canceled.
- Higher the lead time higher are the chances of a booking being canceled.

Generally, people travel with their spouse and children for vacations or other activities. Let's create a new dataframe of the customers who traveled with their families and analyze the impact on booking status.



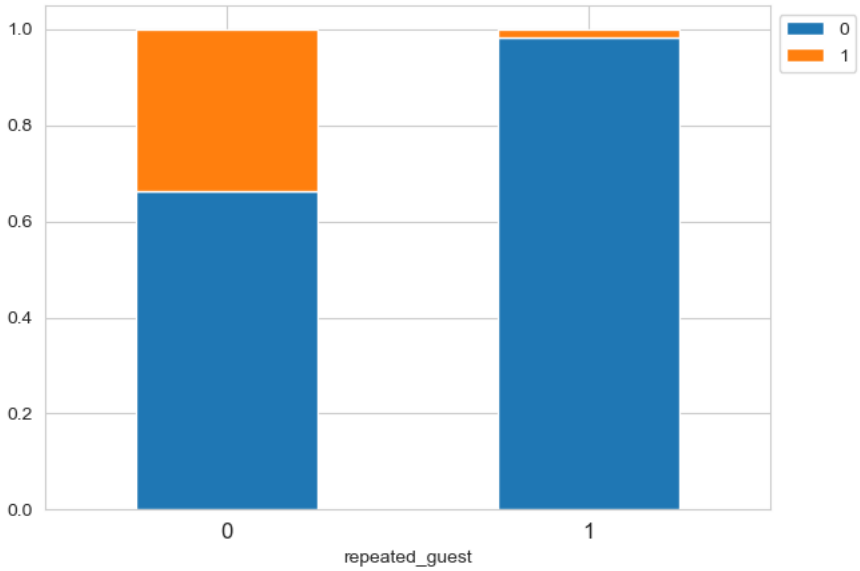
- There's about a 40% chance of a booking getting canceled if the booking is made for 4 family members.

Let's do a similar analysis for the customer who stays for at least a day at the hotel.



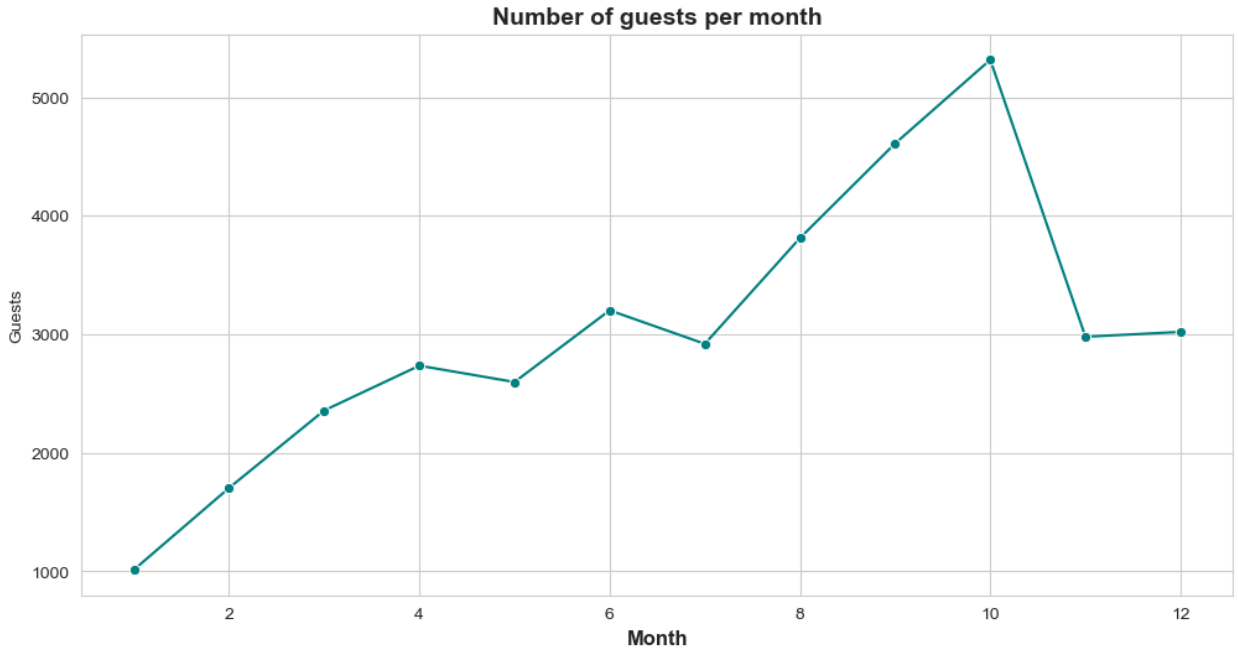
- The general trend is that the chances of cancellation increase as the number of days the customer planned to stay at the hotel increases.

Repeating guests are the guests who stay in the hotel often and are important to brand equity. Let's see what percentage of repeating guests cancel.



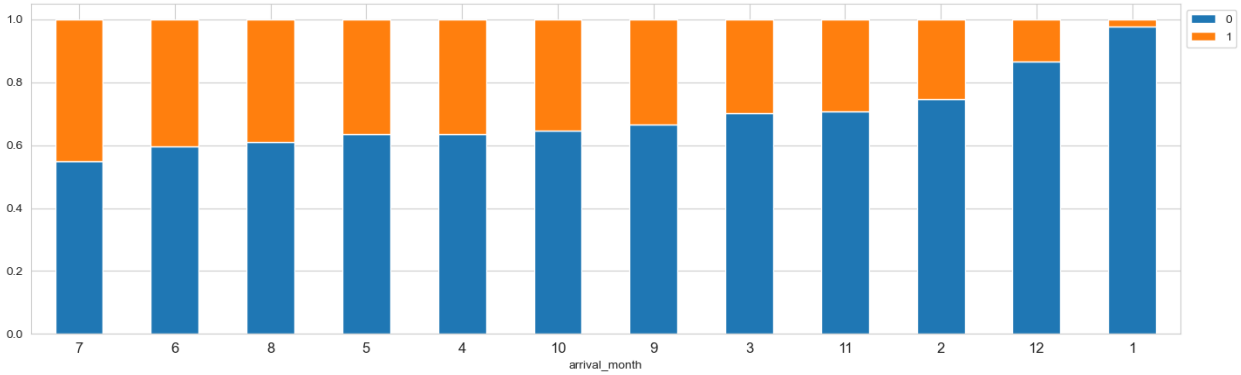
- There are very few repeat customers but the cancellation among them is very less.
- This is a good indication as repeat customers are important for the hospitality industry as they can help in spreading the word of mouth.
- A loyal guest is usually more profitable for the business because they are more familiar with what is on offer at a hotel they have visited before.
- Attracting new customers is tedious and costs more as compared to a repeated guest.

Let's find out what are the busiest months in the hotel.



- The trend shows the number of bookings remains consistent from April to July and the hotel sees around 3000 to 3500 guests.
- Most bookings were made in October- more than 5000 bookings.
- Least bookings were made in January- around 1000 bookings.

Let's check the parentage of bookings canceled each month.

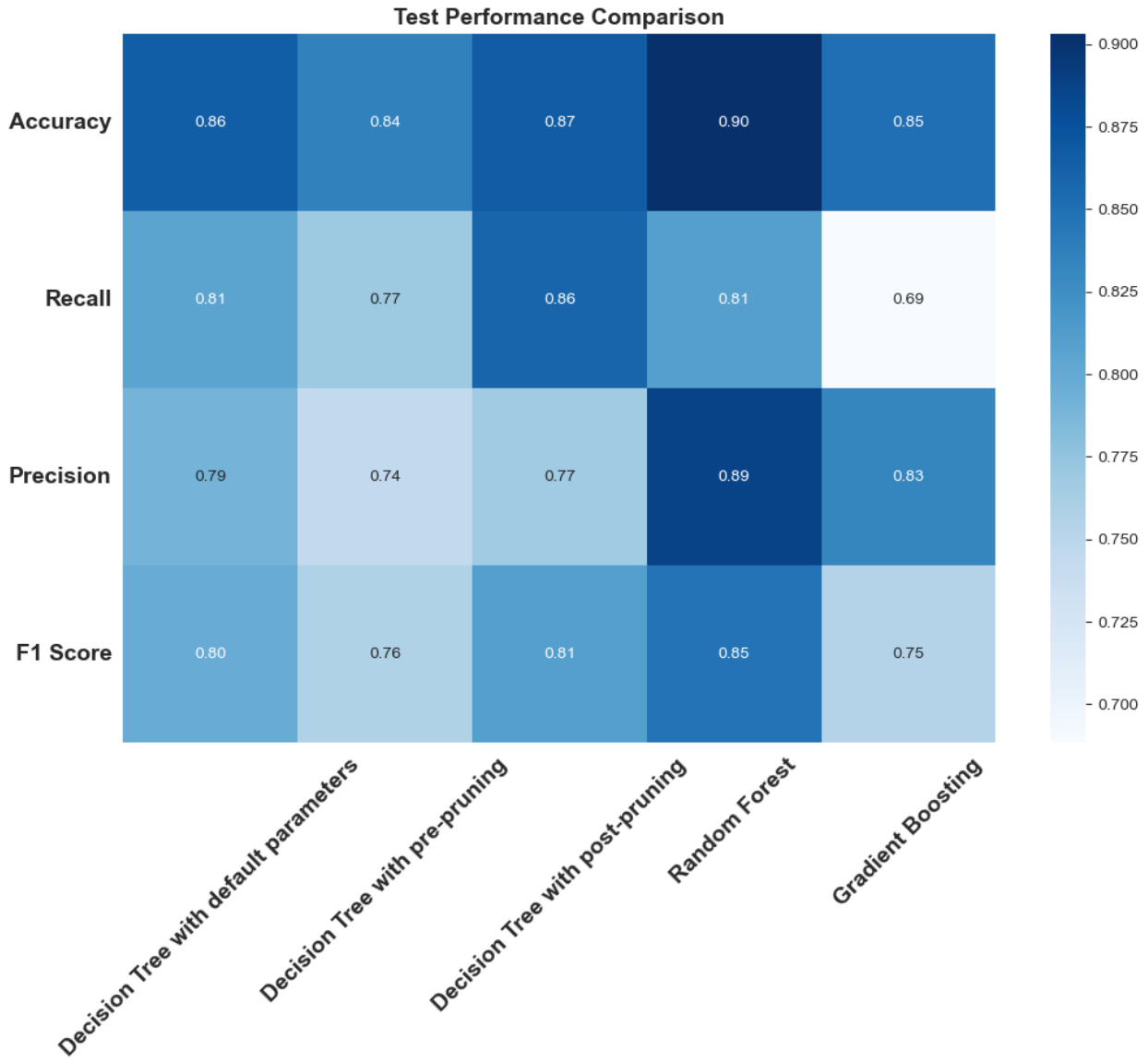


- We see that even though the highest number of bookings were made in September and October- around 40% of these bookings got canceled.
- Least bookings were canceled in December and January- customers might have traveled to celebrate Christmas and New Year.

As hotel room prices are dynamic, Let's see how the prices vary across different months.

- The price of rooms is highest from May to September- around 115 euros per room.

Models & Business Recommendations



- Decision tree model with default parameters is overfitting the training data and is not able to generalize well.
- Pre-pruned tree has given a generalized performance with balanced values of precision and recall.
- Post-pruned tree is giving a high F1 score as compared to other models but the difference between precision and recall is high.
- The hotel will be able to maintain a balance between resources and brand equity using the pre-pruned decision tree model.
- Random forest and Gradient boosting models are giving a generalized performance on training and test set.

Actionable Insights and Recommendations

Insights

- Overall we can see that the Decision Tree model performs better on the dataset.
- Looking at important variables based on p-values in Logistic regression and feature importance in the Decision Tree model
- Lead Time, Number of special requests, Average price per room are important in both model
- From the Logistic Regression model we observe that Lead Time, and Average price per room have a positive relation with bookings getting canclled. And the number of special requests has negative relation with bookings getting cancelled.

Business Recommendations

1. The lead time and the number of special requests made by the customer play a key role in identifying if a booking will be cancelled or not. Bookings where a customer has made a special request and the booking was done under 151 days to the date of arrival are less likely to be canceled.

- Using this information, the hotel can take the following actions:
- Set up a system that can send a prompt like an automated email to the customers before the arrival date asking for a re-confirmation of their booking and any changes they would like to make in their bookings.
- Remind guests about imminent deadlines.

The response given by the customer will give the hotel ample time to re-sell the room or make preparations for the customers' requests.

2. Stricter cancellation policies can be adopted by the hotel.
- The bookings where the average price per room is high, and there were special requests associated should not get a full refund as the loss of resources will be high in these cases.
- Ideally the cancellation policies should be consistent across all market segments but as noticed in our analysis high percentage of bookings done online are cancelled. The booking cancelled online should yield less percentage of refund to the customers.

The refunds, cancellation fee, etc should be highlighted on the website/app before a customer confirms their booking to safeguard guests' interest.

3. The length of stay at the hotel can be restricted.
- We saw in our analysis that bookings, where the total length of stay was more than 5 days, had higher chances of getting cancelled.
- Hotel can allow bookings up to 5 days only and then customers should be asked to re-book if

they wish to stay longer. These policies can be relaxed for corporate and Aviation market segments. For other market segments, the process should be fairly easy to not hamper their experience with the hotel.

Such restrictions can be strategized by the hotel to generate additional revenue.

4. In the months of December and January cancellation to non-cancellation ratio is low. Customers might travel to celebrate Christmas and New Year. The hotel should ensure that enough human resources are available to cater to the needs of the guests.

5. October and September saw the highest number of bookings but also high number of cancellations. This should be investigated further by the hotel.

6. Post-booking interactions can be initiated with the customers.

- Post-booking interactions will show the guests the level of attention and care they would receive at the hotel.
- To give guests a personalized experience, information about local events, nearby places to explore, etc. can be shared from time to time.

7. Improving the experience of repeat customers.

- Our analysis shows that there are very few repeat customers and the cancellation among them is very less which is a good indication as repeat customers are important for the hospitality industry as they can help in spreading the word of mouth.
- A loyal guest is usually more profitable for the business because they are more familiar with offerings from the hotel they have visited before.
- Attracting new customers is tedious and costs more as compared to a repeated guest.
- A loyalty program that offers- special discounts, access to services in hotels, etc for these customers can help in improving their experience.