greatlearning
Learning for Life

Machine Learning – 1 Graded Project on INN Hotels

Contents

Background

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 impact a hotel on various fronts:

1. Loss of resources (revenue) when the hotel cannot resell the room.

2. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.

3. Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.

4. Human resources to make arrangements for the guests.

Objective

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 have to 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.

Data Description

The data contains the different attributes of customers' booking details. The detailed data dictionary is given below.

Data Dictionary:

Booking_ID: the unique identifier of each booking

no_of_adults: Number of adults

no_of_children: Number of Children

no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

no_of_week_nights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel

type_of_meal_plan: Type of meal plan booked by the customer:

Not Selected – No meal plan selected

Meal Plan 1 – Breakfast

Meal Plan 2 – Half board (breakfast and one other meal)

Meal Plan 3 – Full board (breakfast, lunch, and dinner)

required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)

room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group

lead_time: Number of days between the date of booking and the arrival date

arrival_year: Year of arrival date

arrival_month: Month of arrival date

arrival_date: Date of the month

market_segment_type: Market segment designation.

repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)

no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking

no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking

avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)

no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)

booking_status: Flag indicating if the booking was canceled or not.

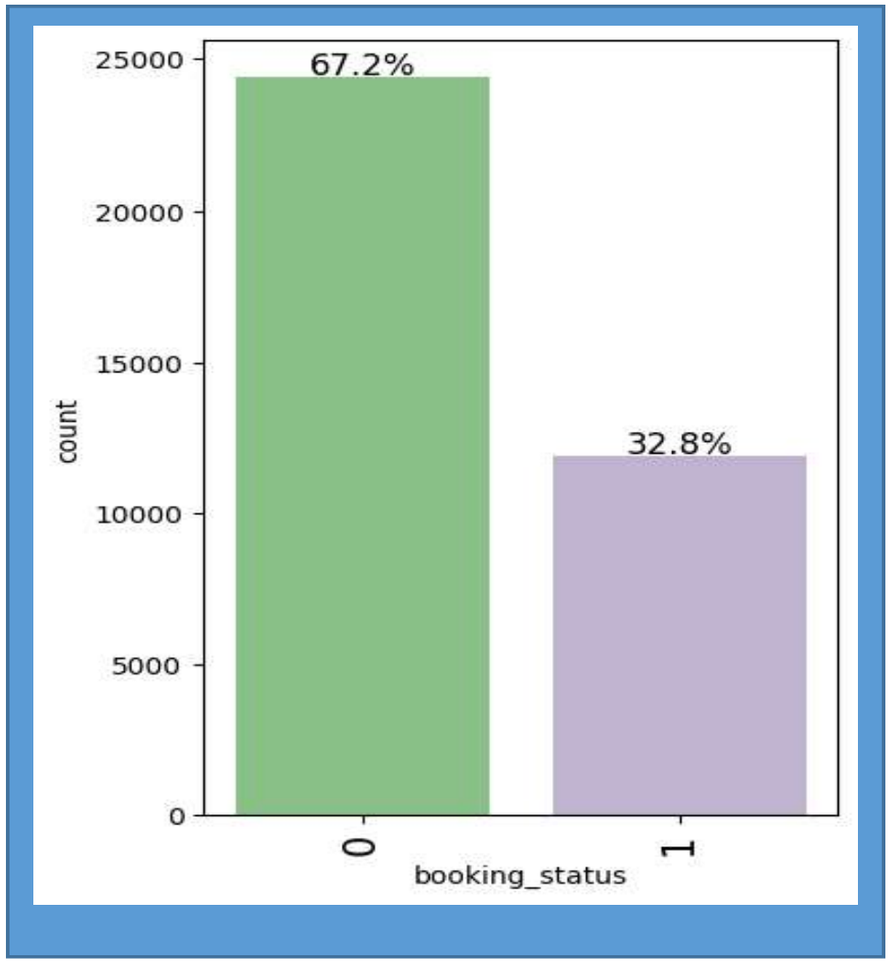
Data Overview

The data contains 36,275 bookings and 19 customers' booking details

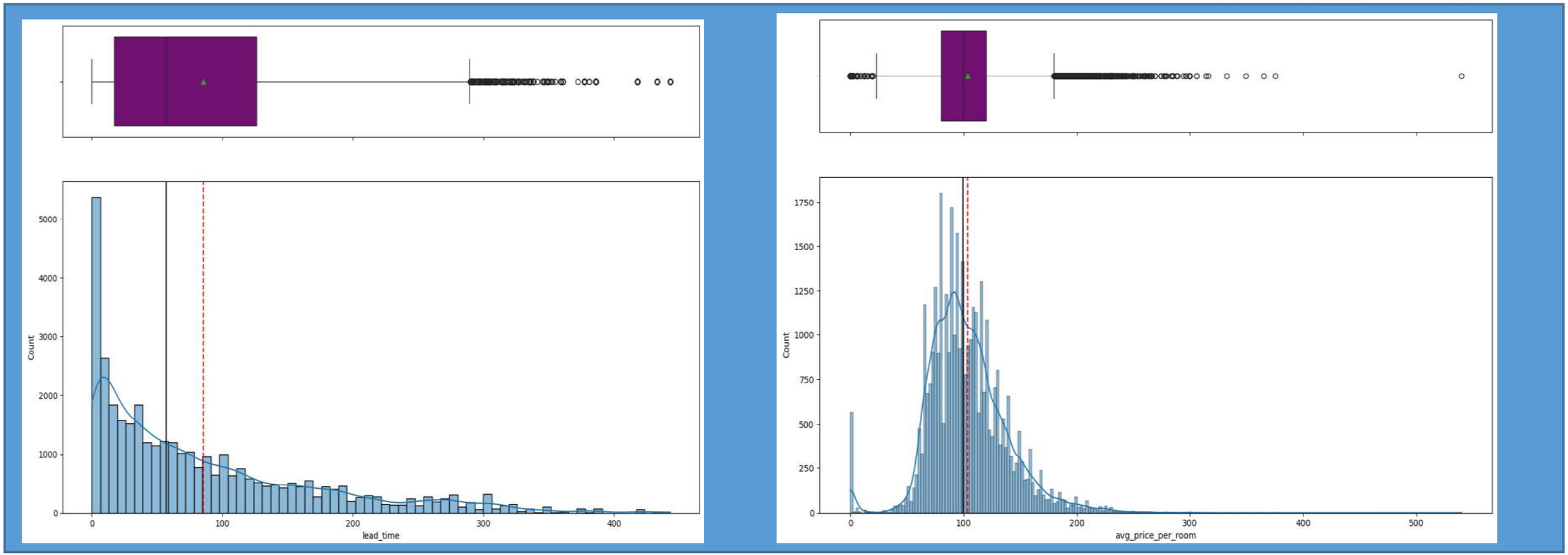
We will use booking status (canceled and not cancel) as target.

There is missing value.

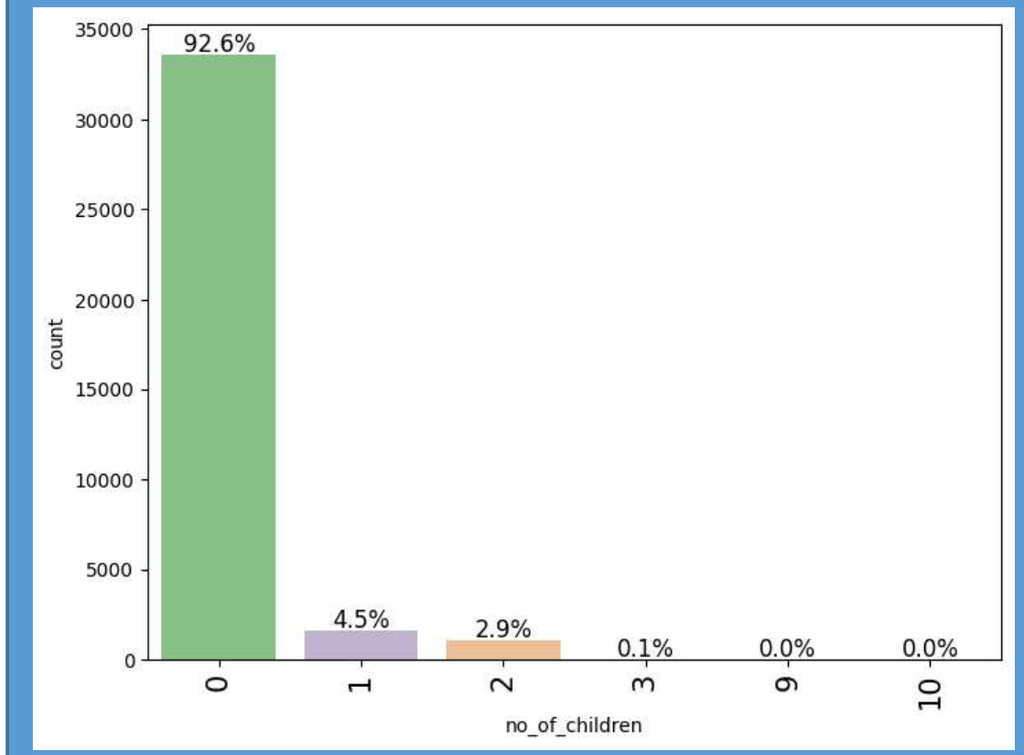
Exploratory Data Analysis (EDA)

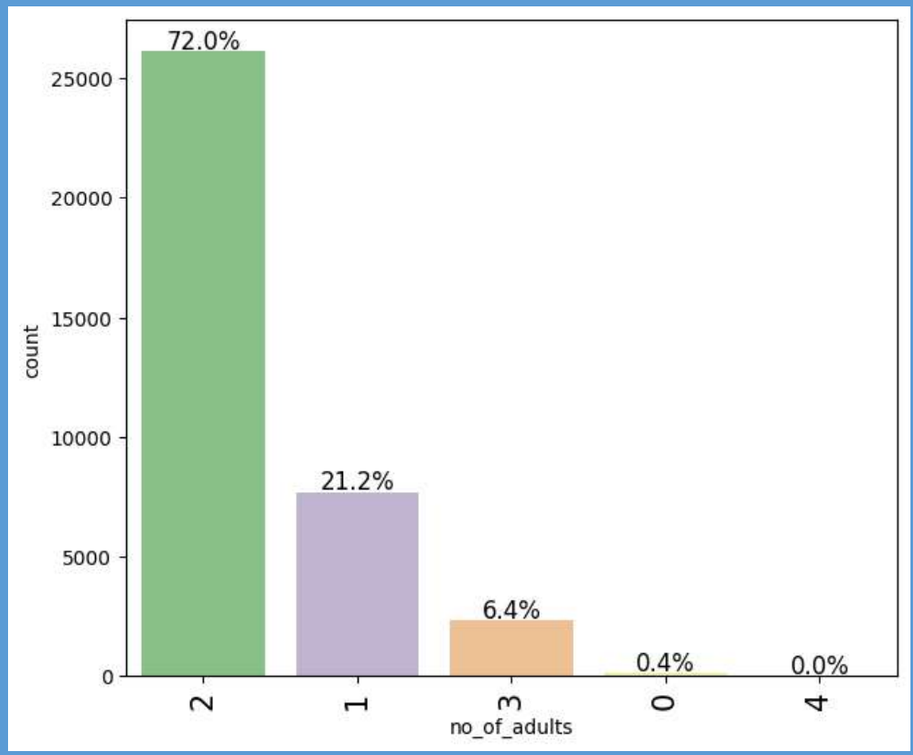


- Booking Status: Not cancel bookingas 0 and canceled as 1
- There are 24,390 bookings (67.2%)that did not cancel and 11,885 bookings (32.8%) which has been canceled.
- It's a very significant number of cancellations that we will need to analyze to reduce resources and toimprove revenue and profits of thehotels.

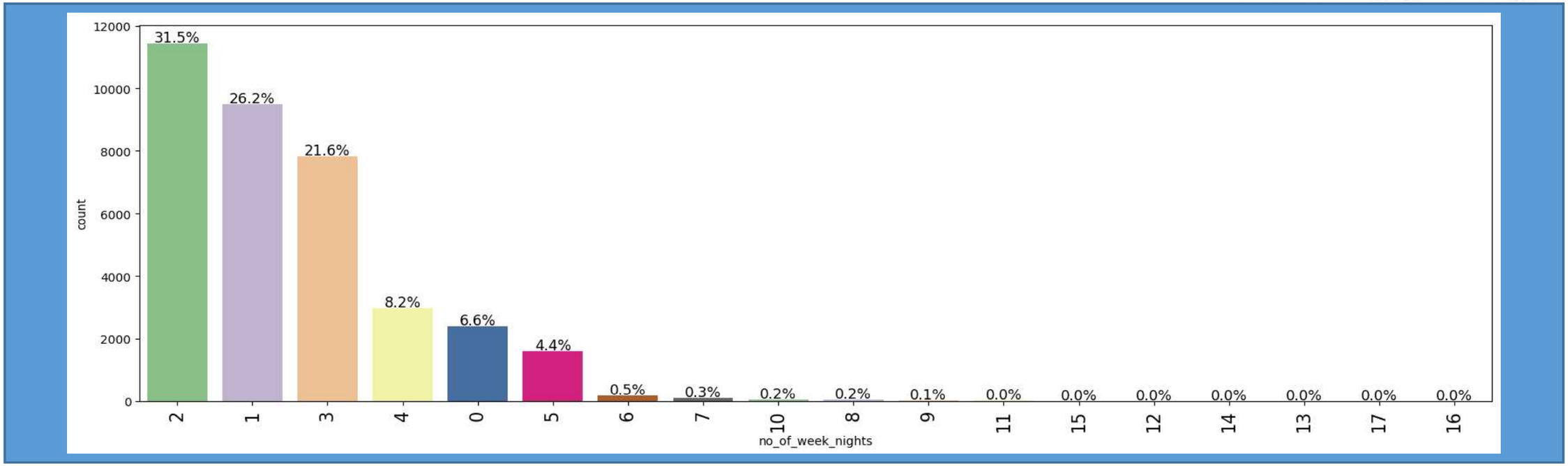


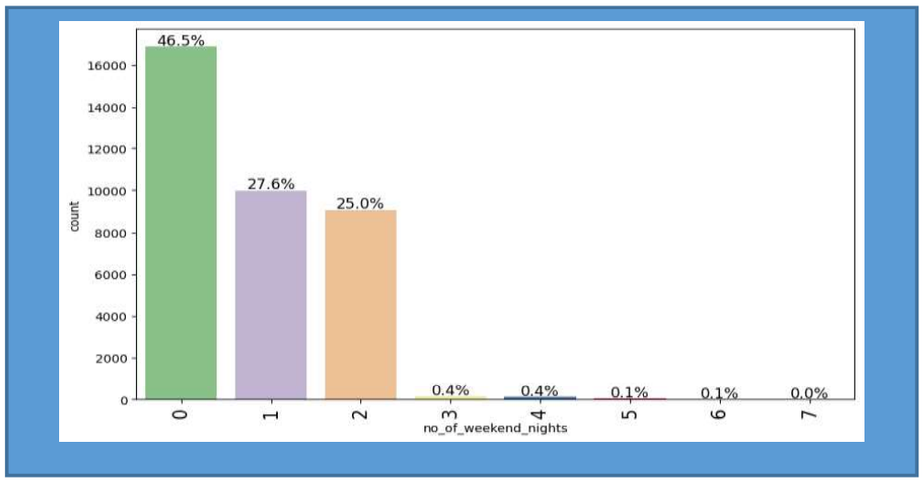
- ***Lead time***, number of day from booked to arrival date, has a heavily right-skewed with lot of outliers. Morethan 5,000 booking were made on the day of or a few days before an arrival date. Maximum lead time is 443 days and median is 57 days.
- ***Average price per room*** is a right-skewed distribution with lots of outliers. Mean and Median are around100 dollars. There are more than 500 booking counts that 0 dollars. Its upper whisker is 179.55 dollars.



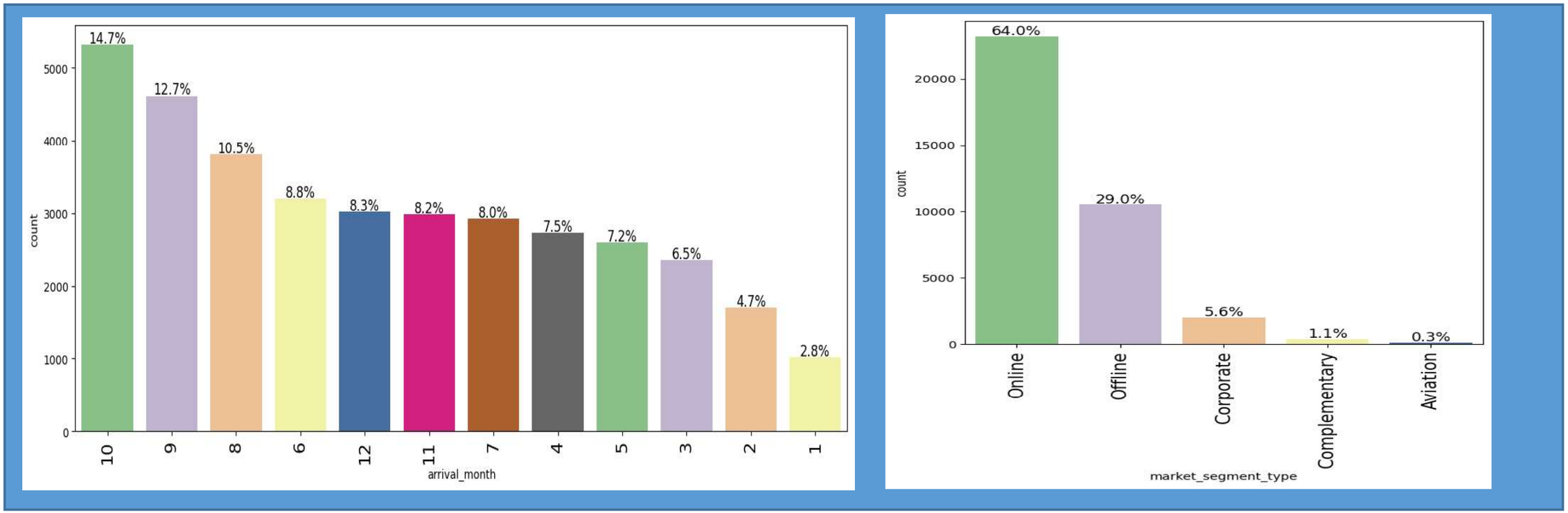


- 72% of Bookings was made for 2 adults and 21% was for 1 adults.
- Very small amount of bookings included children

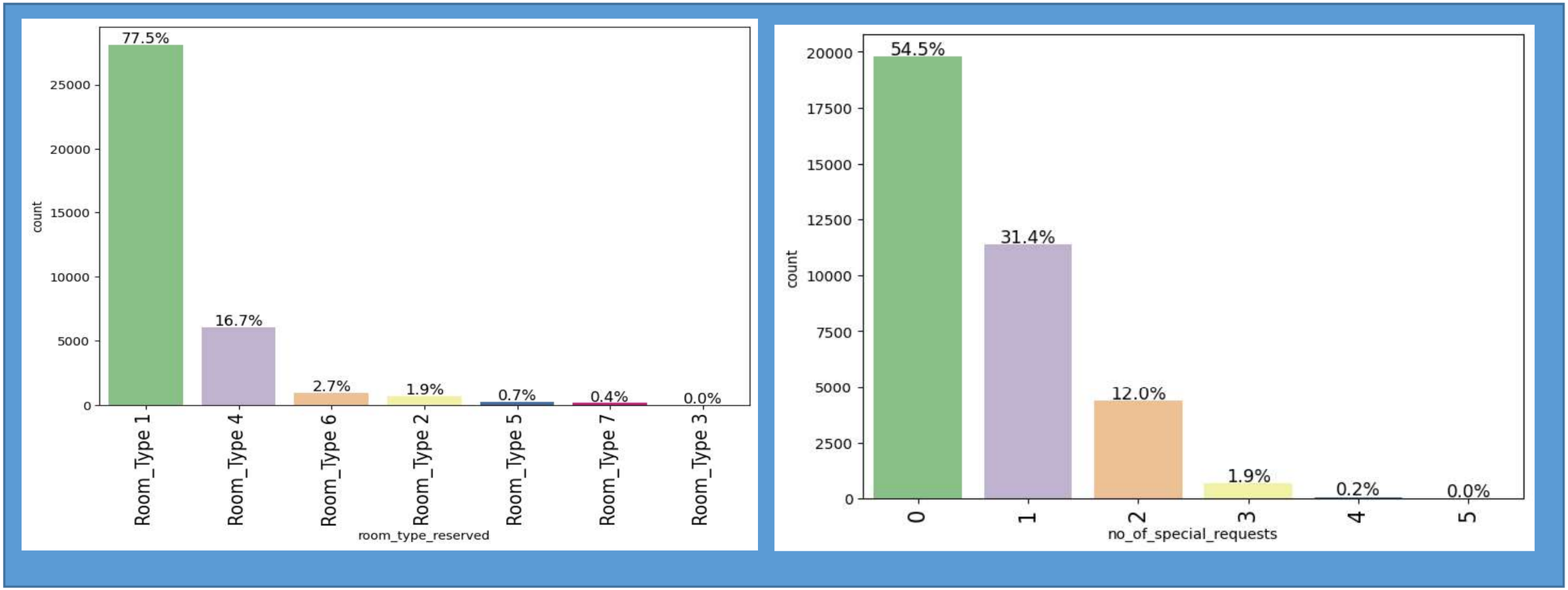




- ***No of weekday nights*** has a range from 0 to 17 days with 2 nights as the most frequent bookings, following by 1 days. Thebookings that have 0 weekday night, assuming that they are leisure travels.
- ***No of weekend nights*** has a range from 0 to 7 days. The mostfrequent bookings is 0 nights, assuming customers booked rooms for business. Following by 1 nights and 2 nights.
- A customers will most likely book a room for business.



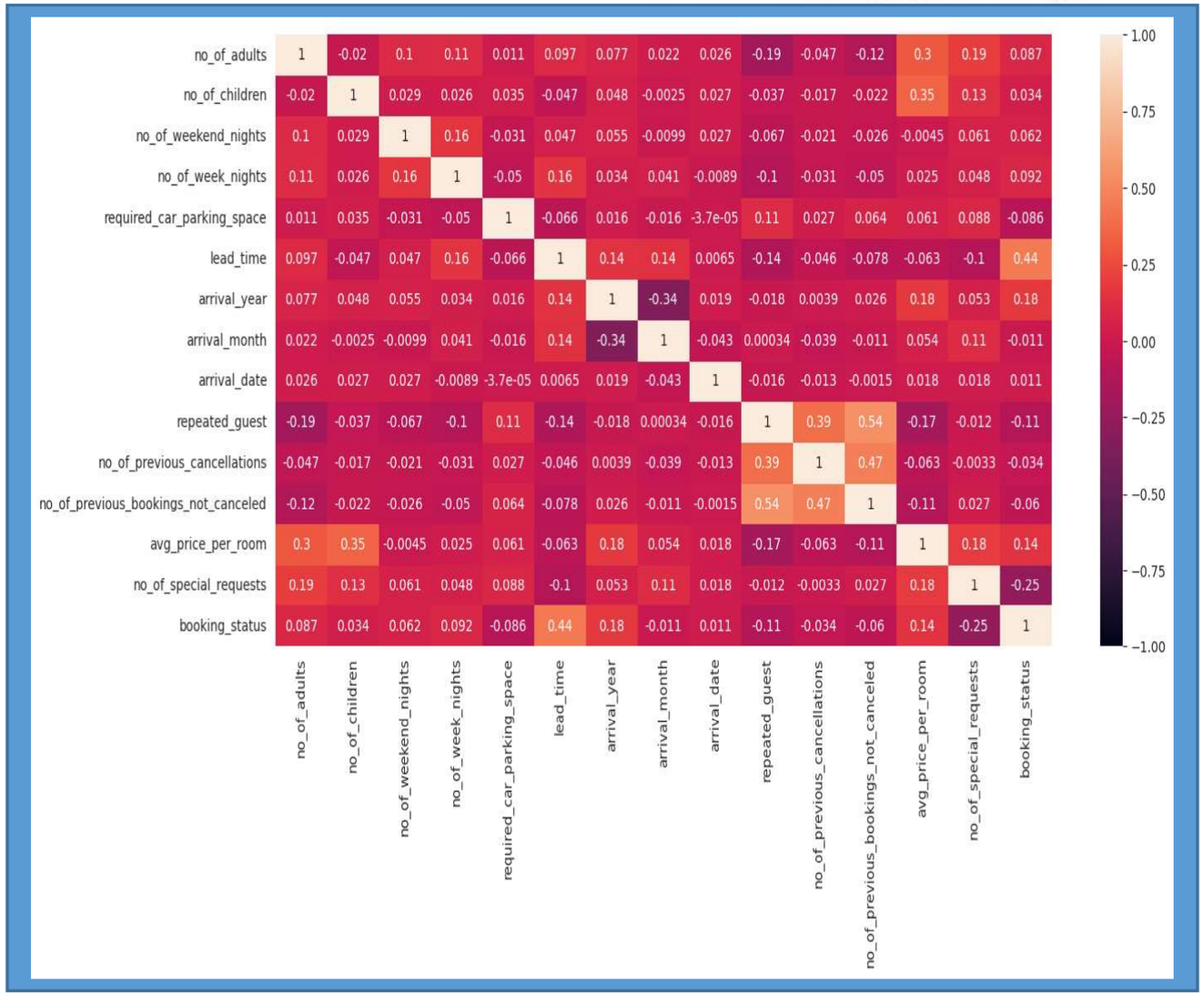
- The top 3 ***arrival months*** are October at 12.7% bookings, September at 12.7% bookings, and August 10.5% bookings. Holidays season in November and December have similar booking counts
- There are 5 type of ***market segments***. Customers made room reservations via online, most convenience.Following by offline, corporate, complementary and aviation.

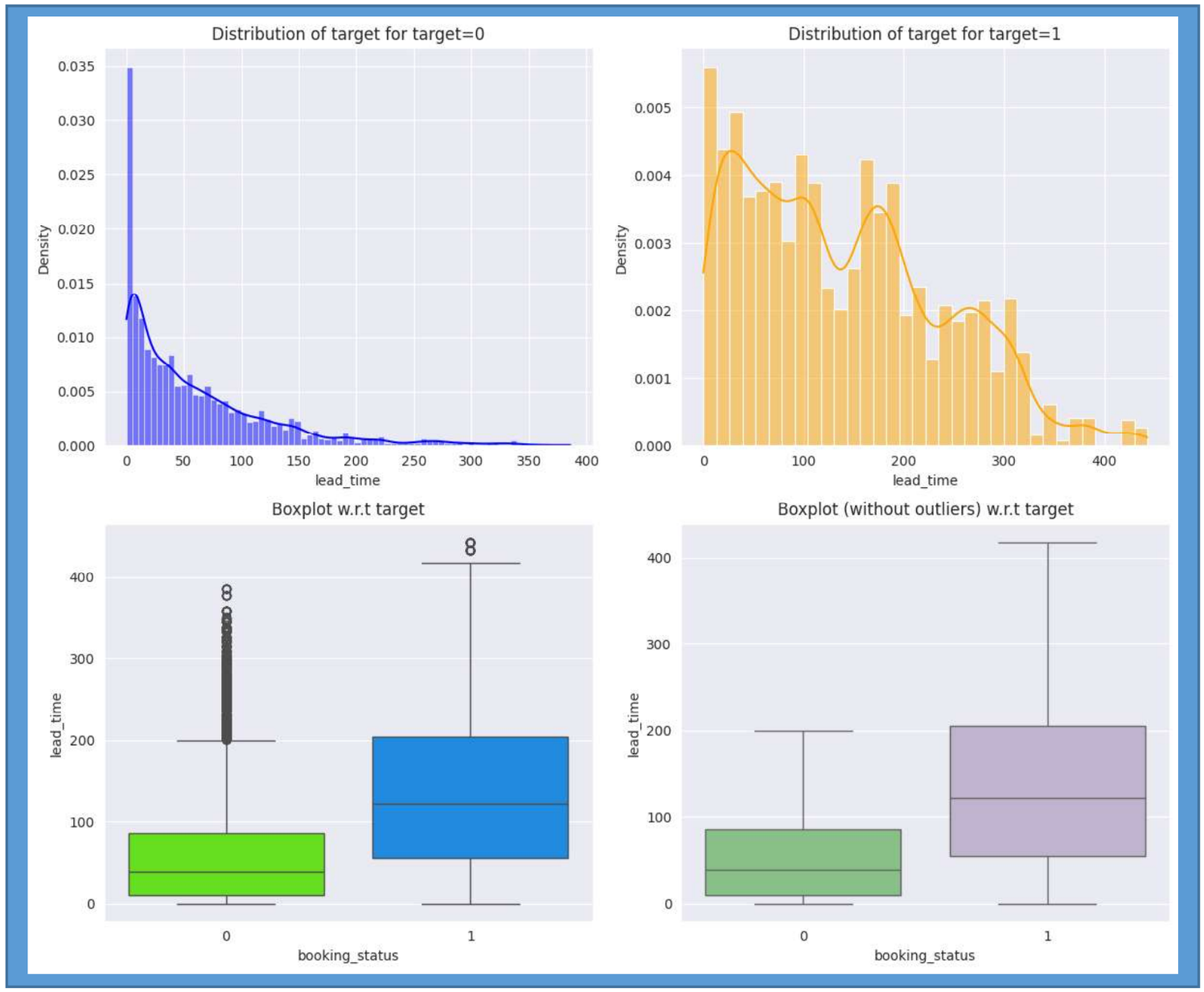


- ***No of special requests*** has a range from 0-5 requests. Most bookings, 54.5% bookings, have no specialrequest, following by 1 request, 2 requests, and the rest.
- ***Room type reserved*** has 7 types of rooms. Most of customers choose room type 1 as 77.5%bookings.

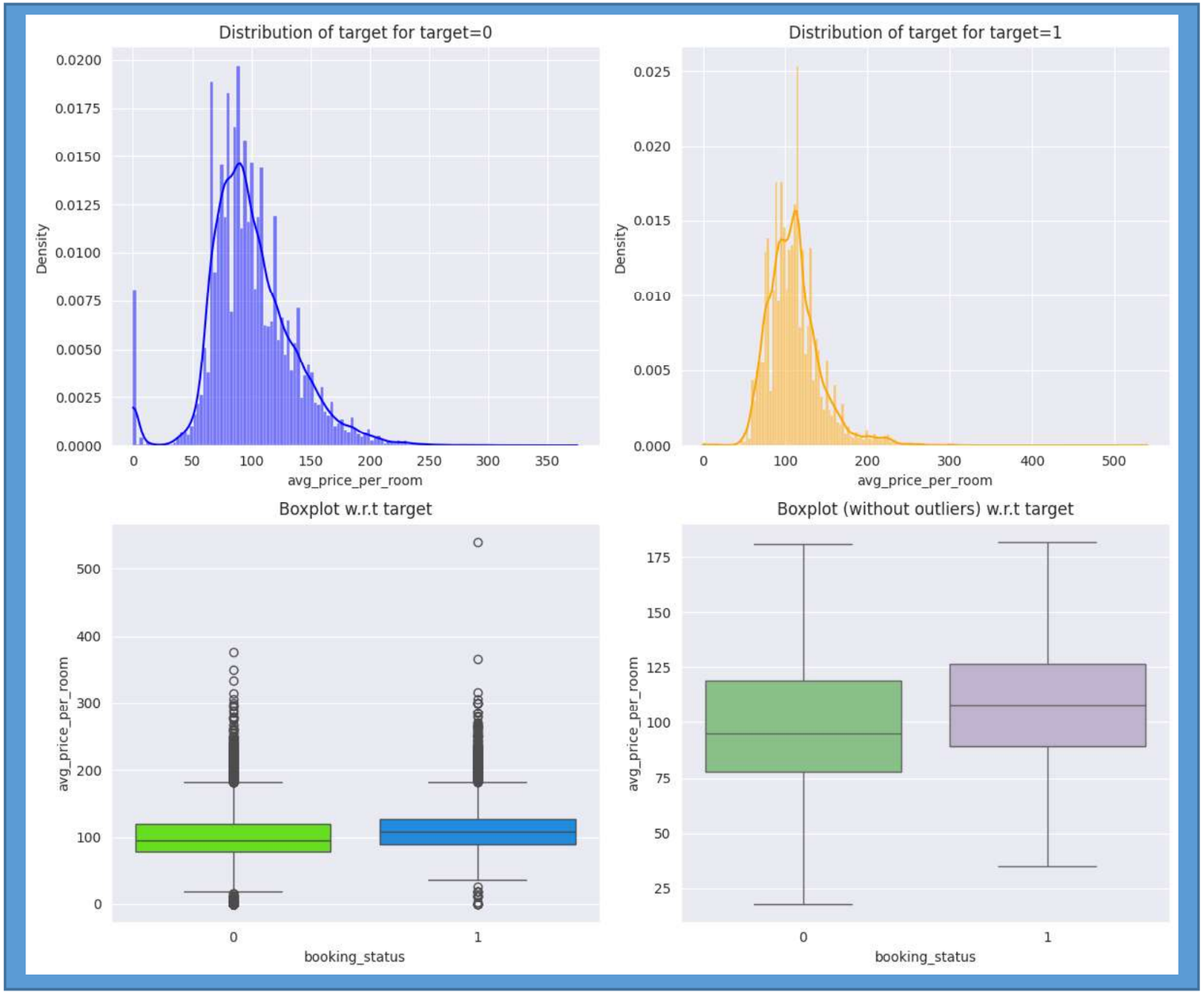
Correlation

- ***Lead_time*** has the highest correlation with booking_status at 0.44. Following by no_of_special_request at -0.25.
- ***repeated_guest*** has 0.54 correlation with no_of_previous_bookings_not_canceled.
- ***ave_price_per_room*** has 0.30 correlation with no_of_adult and 0.35 correlation withnumber of children.
- ***booking_status*** has 0.14 correlation with ave_price_per_room and -0.11 with repeated-guest
- We will investigate more of the relationship between lead_time and booking_status and more.

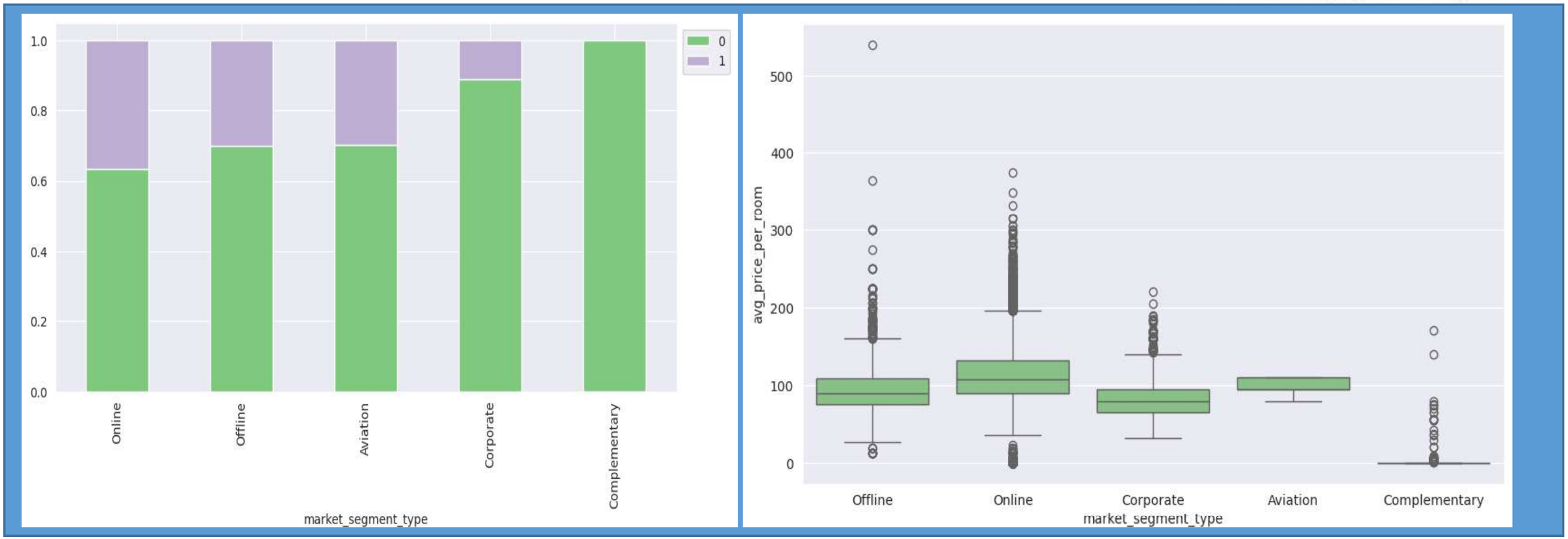




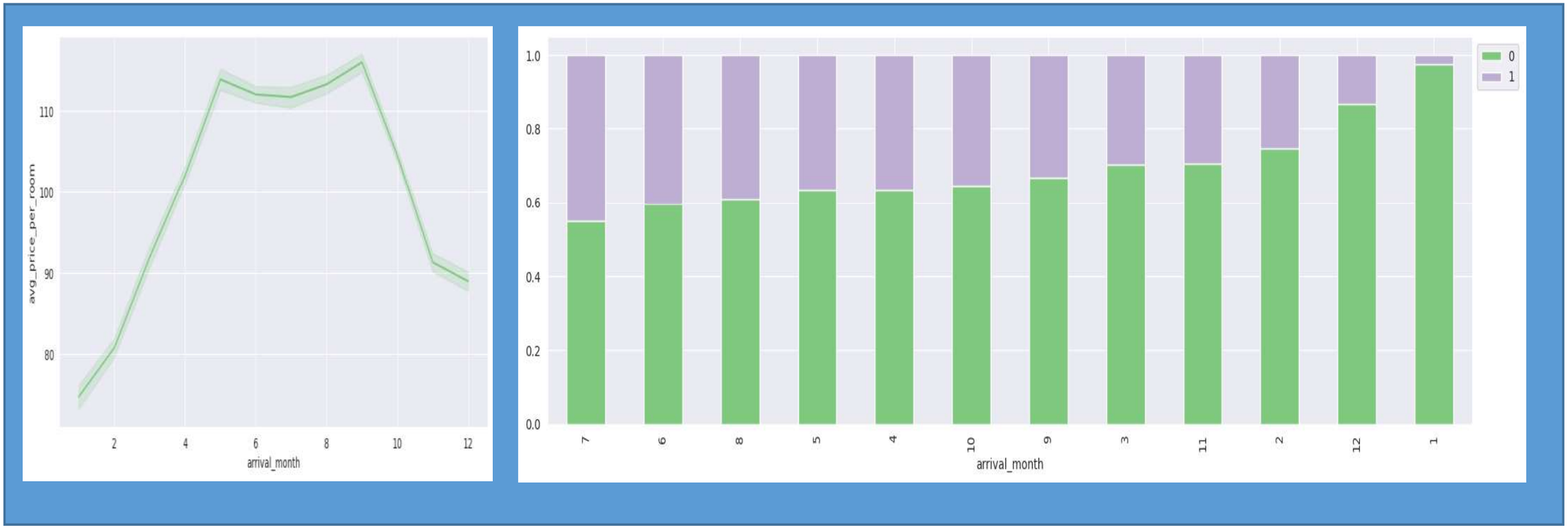
- The median of ***lead_time*** of not cancel booking is around 50 days.
- The median of lead_time of canceled booking is around 120days.
- Without outliers, the range for lead_time of not cancel booking is0-200 days and for lead_time of canceled booking is 0-400 days.
- From this observation, the longerthe lead_time of booking, the higher chance for cancellations.



- With and without outliers, the ***average price per room*** for canceled booking is ~110 dollars and for not canceled bookings is ~95 dollars.
- Customers who canceled theirbooking may find a more affordable room from differenthotels.
- Customers who did not cancel their booking are satisfied withtheir room prices and see it as good prices.

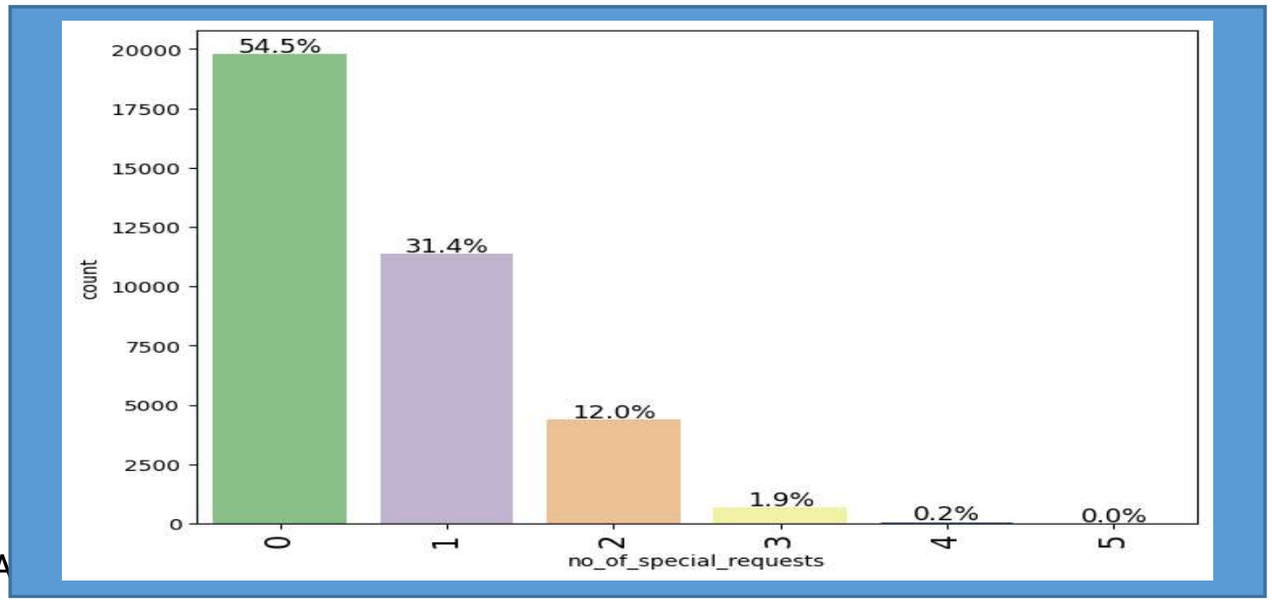


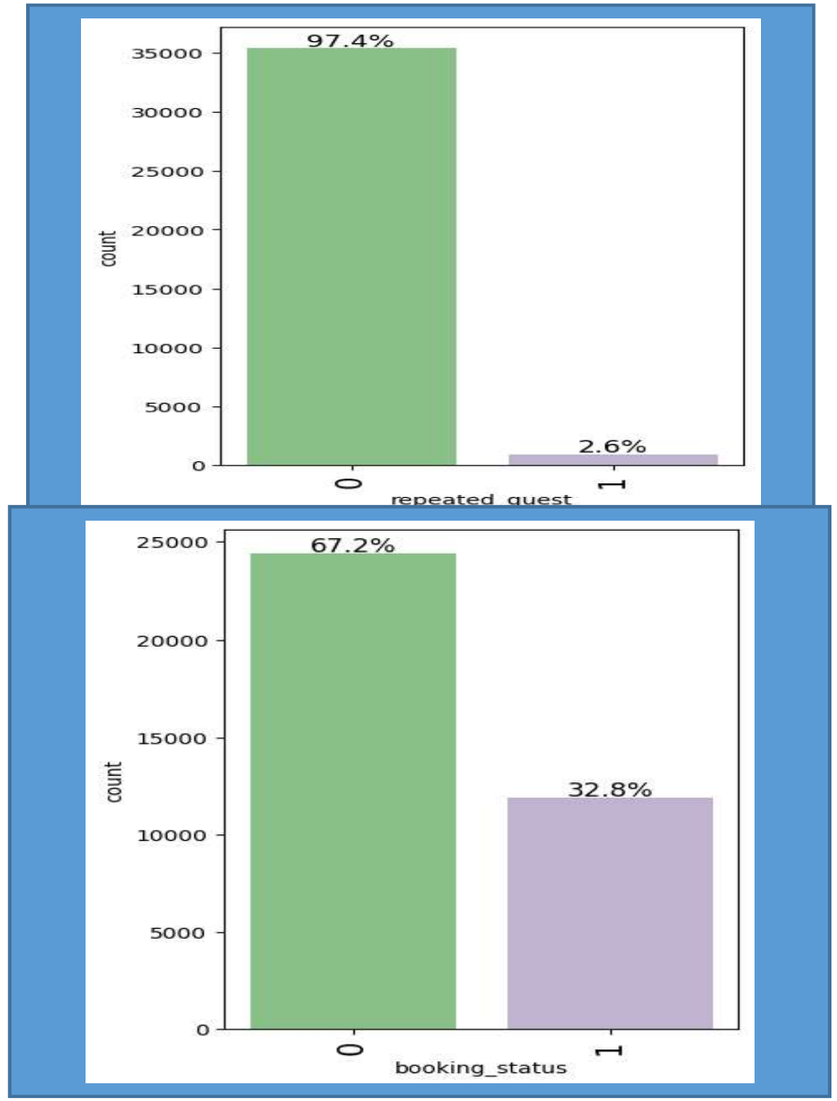
- Observation on ***market segment type***, Online booking has the highest cancellations (~40%), following by Offline, Aviation, Corporate and Complementary (~0%). Aviation cancellations mightdue to flight delays or flight canceled.
- ***Average price per room*** for online bookings has the highest price (over 100 dollars), followingaviation bookings, offline bookings and complementary (data shown that a large number of bookings is free).



- The most expensive months are September, April, October, May, August, June and July . The average price inthese months are around 110 dollars or higher. Correlating to the most cancellations months.
- The least expensive months are January, February, March, December, and November. Those pricesare around 90 dollars or less. Correlating to the less cancellations months.

- Bookings of ***repeated customers*** is 97.4% that bookings did not cancel and for not repeated customers has ~65% cancellations. For
- Top 3 of median price of Bookings with ***No of special requests*** is 2, 3, and 4, which is around 120 dollars. For 0request bookings price is around 90 dollars. All of them have outliers.





1. **What are the busiest months in the hotel?**
    - The busiest months in the hotel are October, September, and August.
        - October with approximately 14.7% of the total bookings, is the busiest month.
        - September follows closely with 12.7% of the bookings and August comes next with 10.5% of the bookings. These months experience higher demand and occupancy rates, likely due to various factors such as holidays, seasonal events, or favorable weather conditions. The hotel may need to plan accordingly and ensure sufficient resources and staffing during these peak months to provide a satisfactory experience for guests.
2. **Which market segment do most of the guests come from?**

- The majority of guests come from the online market segment. This market segment has approximately 64% of the total guests. It indicates that a significant number of bookings are made through online channels such as hotel booking websites, online travel agencies, or direct online reservations. The online market segment is followed by offline sources, which has about 29% of the total guests. These offline sources may include direct bookings through phone calls, walk-ins, or bookings made through traditional travel agencies. From this we can see a clear dominant market segment, which could help the hotel focus on marketing strategies and efforts on channels that generate the most bookings.

3. **Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?**
    - The room prices vary across different market segments. The online market segment appears to have the highest prices, indicating that bookings made through online channels tend to have higher rates compared to other segments. This could be attributed to various factors such as the convenience of online plattaforms.
    - On the other hand, the offline market segment tends to have relatively lower room prices compared to the online segment. This might be because offline bookings may involve direct negotiation with the hotel or traditional travel agencies, which could result in lower negotiated rates or special discounts.

- It's important to note that the differences in room prices among market segments can also be influenced by factors like customer demographics, booking patterns, and demand fluctuations. Therefore, the hotel should always adjust its pricing strategy to optimize revenue based on market dynamics and customer preferences for each market segment.

4. **What percentage of bookings are canceled?**
    - We have 32.8% of the bookings canceled

5. **Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?**

- From repeating guests approximately 2.6% of them have made repeated bookings, which is around 930 bookings. Out of these 930 bookings, only 16 reservations have been canceled, resulting in 1.7%. This indicates that the majority of repeating guests has a high level of commitment, as they rarely cancel their bookings.

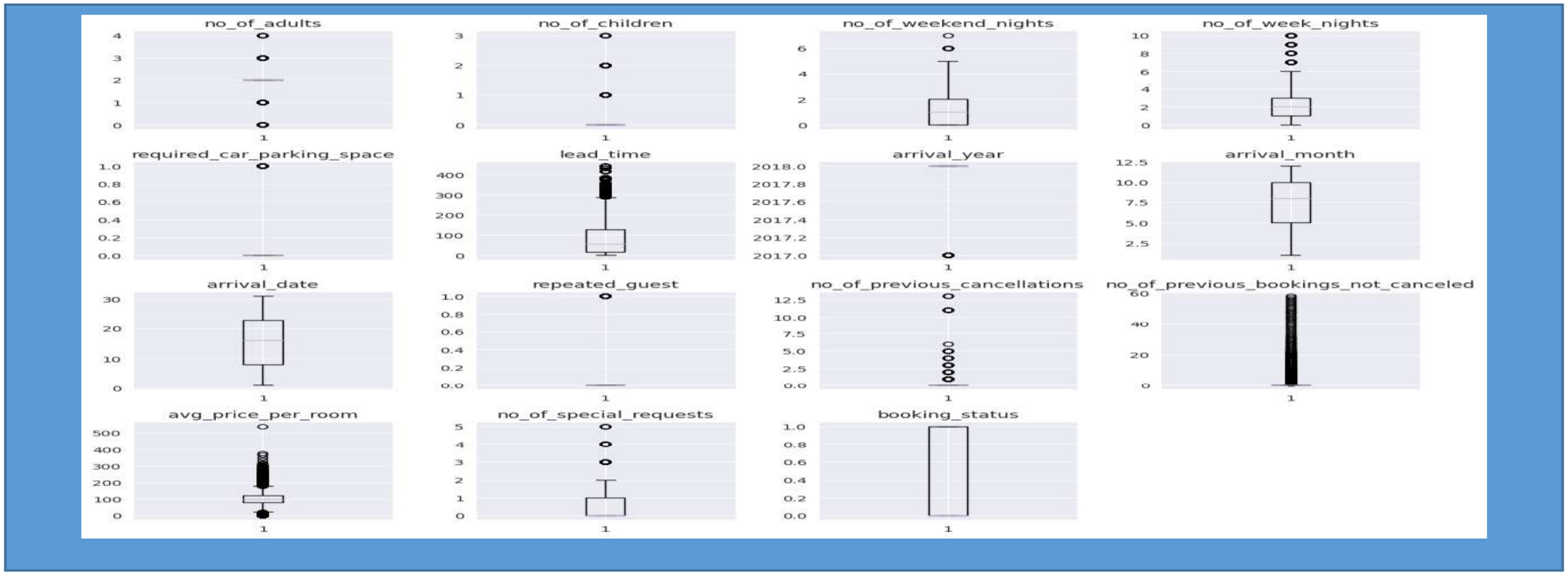
6. **Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?**

- As the number of special requests increases, the proportion of cancellations tends to decrease. This suggests that guests who make more special requests are less likely to cancel their bookings.

- Regarding bookings with no special requests, 43.22% were canceled.
- For bookings with 1 special request, the cancellation rate decreased to 23.79%.
- Bookings with 2 special requests had a lower cancellation rate of 14.61%.
- Bookings with 3, 4, or 5 special requests had no cancellations at all.
- This trend suggests that when guests have specific preferences or special requirements, they tend to be more committed to their bookings and less likely to cancel.

Data Preprocessing

- We have decided to include the outliers in our analysis because they contain valuable information. They can provide unique insights and contribute to a more comprehensive understanding of the data



Data Preparation

Before we proceed to build a model, we will:

- Drop Booking_ID column. We decided to group booking ID because they are unique numbers.We cannot use it for pattern recognitions.
- Treat outliers
  - Avg_price_per_room: There are only outliers above upper whisker.
    - Calculated upper whisker which is 179.55 dollars
    - Assigned 179.55 dollars to outliers greater or equal to 500 dollars.
  - No_of_children: We used 3 children to replacing bookings with 9 or 10 children
- Encode categorical features: Type_of_meal_plan, Room_type_reserved,Market_segment_type, and Booking_status
- Split the data into train (70%) and test (30%) to evaluate the model that we build on the traindata.

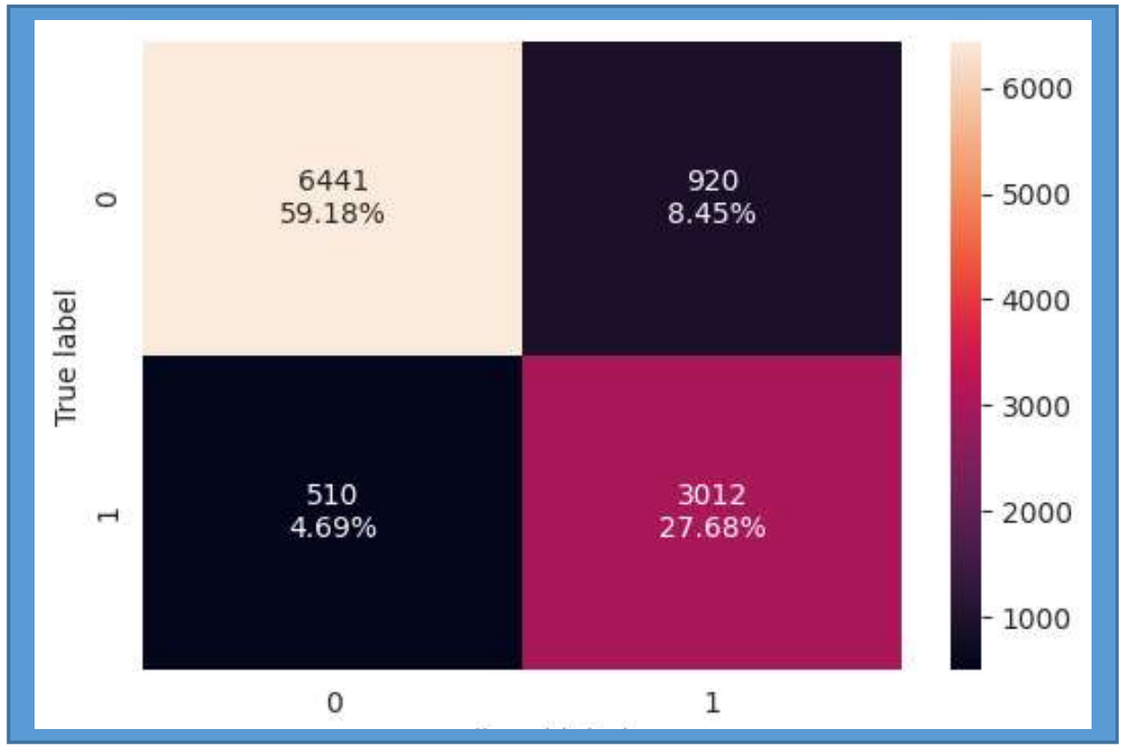
Model Performance Summary

- We want to predict which bookings will be canceled.
- Model can make wrong prediction as ***false negative*** (predicting a booking to be canceledwhen it does not) and ***false positive*** (predicting a booking to not cancel when it does).
- We decided that both false negative and false positive are important.
  - ***False negative***: the hotel might not be able to provide satisfactory services to the customer by assuming that this booking will be canceled. This might damage the brandequity.
  - ***False positive***: the hotel will lose resources and will have to bear additional costs ofdistribution channels try to resell the room.
- We want to reduce the losses by want ***F1 Score*** to be maximized for higher the chances ofminimizing False Negatives and False Positives.

We will use ***Logistic Regression Model*** and ***Decision Tree Model*** for prediction

Confusion Matrix

- We want F1 Score to be maximized,minimizing False Negatives and False Positives.
- False Negatives: 510 bookings(4.69%)
- False Positives: 920 bookings(8.45%)



Logistic Regression model :

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80549	0.79104	0.80029
Recall	0.63207	0.74112	0.69939
Precision	0.73951	0.66367	0.69581
F1	0.68158	0.70026	0.69760

Decision Tree Model

	Decision Tree with class_weight	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87081	0.83497	0.86860
Recall	0.80835	0.78336	0.85520
Precision	0.79570	0.72758	0.76602
F1	0.80197	0.75444	0.80816

- All three models demonstrate strong performance on both the training and test data, indicating that they do not suffer from overfitting.
- Based on these results, we can conclude that the logistic regression model with a threshold of 0.37 has the best performance

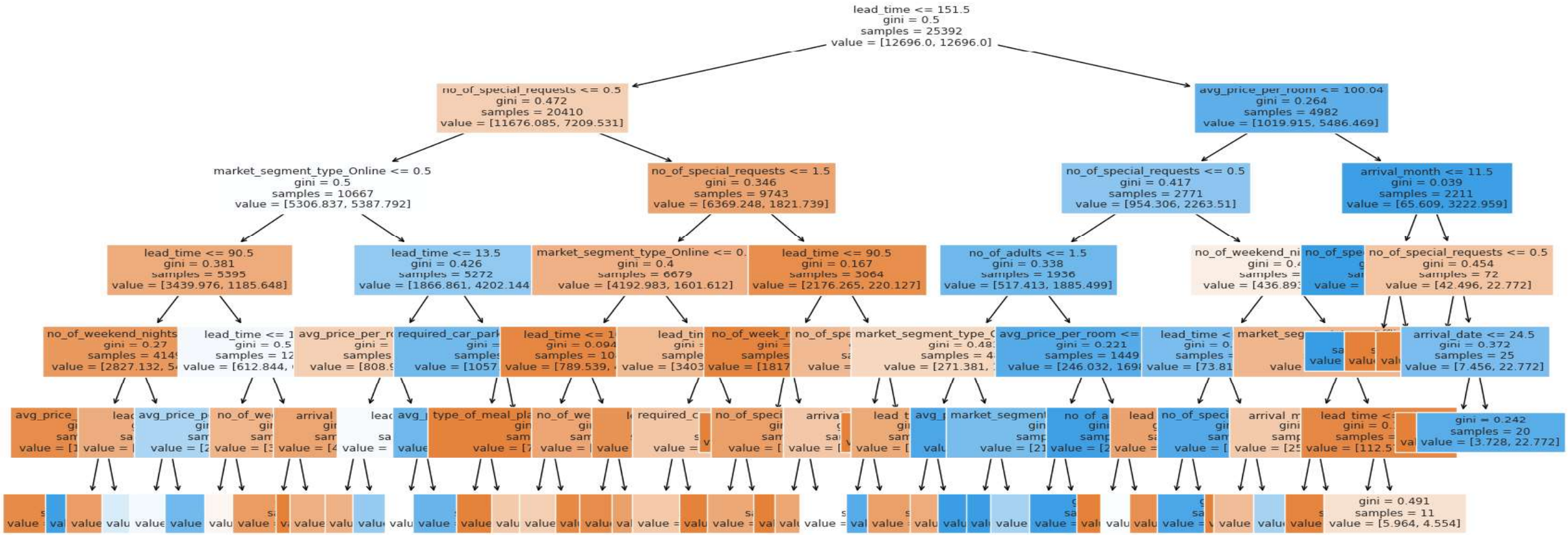
across various metrics, including accuracy, recall, precision, and F1 score. Therefore, it is recommended as the optimal model for predicting booking cancellations in this scenario.

- This means that if the predicted probability of a booking being canceled exceeds 37%, we will classify it as a cancellation. If the predicted probability falls below 37%, we will classify it as a non-cancellation. This threshold selection allows us to strike a balance between accurately identifying cancellations (high recall) and minimizing false positives (high precision).

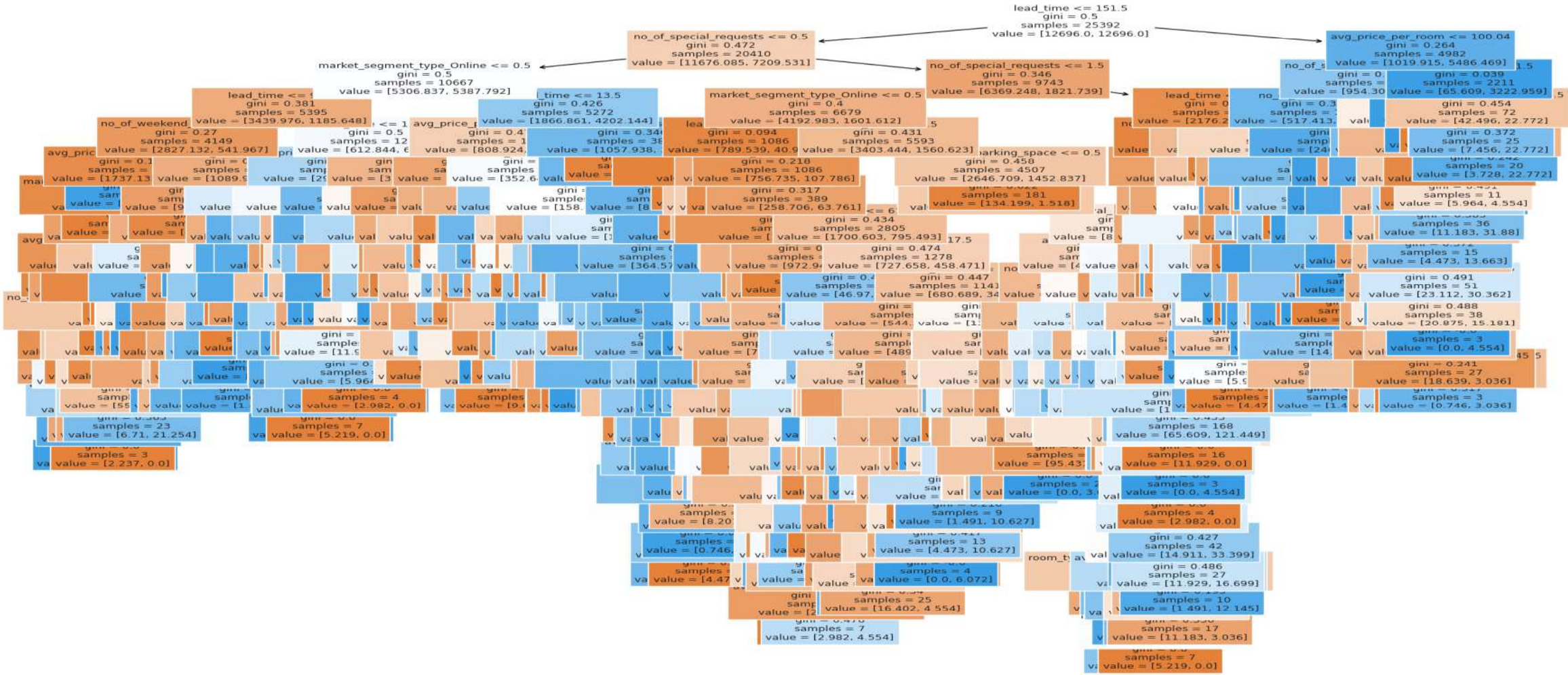
Decision Tree Post – Purning

- **The best model**: alpha =0.0001226763315516706, class_weight='balanced', andrandom_state = 1
- **F1 Score**: the Training set and the Test set are very close. This model is a greatmodel for that this data set.
- The best **alpha** is very small that indicates a very complex model.

Visualizing the Decision Tree



Decision tree Post – Purning



Business Insights and Recommendations

1. The model with a threshold of 0.37 performs better than the other models across multiple evaluation metrics, such as accuracy, precision, recall, and F1 score. Therefore, it is recommended as the optimal choice for accurate cancellation predictions.
2. Among the different decision tree models tested, the one with post-pruning has the highest recall on the test set. Recall measures the model's ability to correctly identify actual booking cancellations. A higher recall suggests that the model is more effective at capturing real cancellations. The recommended model is the decision tree with post-pruning for maximizing the detection of booking cancellations.

- What profitable policies for cancellations and refunds can the hotel adopt?
    - The hotel can offer a refund policy where the amount decreases as the check-in date approaches. This may encourage guests to cancel well in advance, giving the hotel an opportunity to rebook the room.
    - Also, it could include an option to modify the booking that allows the customer to make changes up to a certain day before check-in, this may prevent cancellations when they just want changes in the reservation.
        - What other recommendations would you suggest to the hotel?
    - The hotel may consider paying special attention to Lead time, Online market segment and average price per room because according to the decision tree model thore are the most important variables in determining if a booking will be cancelled.
    - The prices of rooms in different market segments can vary significantly. In particular, the online market segment, which makes up the majority of our guests at 64%, generally has higher prices compared to other segments. This could happen for the dynamics of online booking platforms who offer convenience, an extensive range of options, and easy price comparison. By recognizing this trend, the marketing team can take advantage of the opportunity to attract more customers.

- From the analysis of the data, it is clear that there is a seasonal pattern in the booking behavior. The number of bookings starts to increase from August, reaching its peak in October with 5317 bookings. This period likely corresponds to a popular travel season or event. After October, the number of bookings gradually decreased, with November still showing a relatively high number of bookings (2980). However, there is a significant decline in December, January, and February, with January having the lowest number of bookings (1014).
    - The marketing team can take advantage of this trend by designing a strong campaign to increase bookings during the winter season. They can achieve this by offering appealing deals that attract more customers, resulting in higher occupancy rates.
    - Also by aligning business decisions with booking patterns, hotels can effectively allocate resources and optimize operations year-round. During busy times, hiring extra staff is important to provide great customer service and handle the increased demand efficiently. In contrast, during slower periods, adjusting staffing levels and optimizing operations can help manage costs while still delivering a high level of service.
- According to the analysis repeated guests are less likely to cancel a booking, implement a loyalty program that offers benefits and rewards to these customers can encourage guests to choose your hotel for future stays and reduce the likelihood of cancellations.

- INN Hotels should keep getting data and making further analysis about the reasons customers have to cancel and also .