Business Report - 4

PG Program in Data Science and Business Analytics

submitted by

Sangram Keshari Patro

BATCH:PGPDSBA.O.AUG24.B

Contents

List of Figures

1 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.

2 Data Description

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

2.1 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.

3 Data Overview

3.1 Importing necessary libraries and the dataset

The dataframe is printed. It has 36275 rows & 19 columns.

3.2 Structure and type of data

Data is explored further. The dataset is free from duplicate rows and contains no null values.

```
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Booking_ID                            36275 non-null  object
 1   no_of_adults                          36275 non-null  int64
 2   no_of_children                        36275 non-null  int64
 3   no_of_weekend_nights                  36275 non-null  int64
 4   no_of_week_nights                     36275 non-null  int64
 5   type_of_meal_plan                     36275 non-null  object
 6   required_car_parking_space            36275 non-null  int64
 7   room_type_reserved                    36275 non-null  object
 8   lead_time                             36275 non-null  int64
 9   arrival_year                          36275 non-null  int64
 10  arrival_month                         36275 non-null  int64
 11  arrival_date                          36275 non-null  int64
 12  market_segment_type                   36275 non-null  object
 13  repeated_guest                        36275 non-null  int64
 14  no_of_previous_cancellations          36275 non-null  int64
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64
 18  booking_status                        36275 non-null  object
```

Figure 1: Table depicting the datatype and Non-Null values in each column.

3.3 Statistical summary

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.0	1.844962	0.518715	0.0	2.0	2.00	2.0	4.0
no_of_children	36275.0	0.105279	0.402648	0.0	0.0	0.00	0.0	10.0
no_of_weekend_nights	36275.0	0.810724	0.870644	0.0	0.0	1.00	2.0	7.0
no_of_week_nights	36275.0	2.204300	1.410905	0.0	1.0	2.00	3.0	17.0
required_car_parking_space	36275.0	0.030986	0.173281	0.0	0.0	0.00	0.0	1.0
lead_time	36275.0	85.232557	85.930817	0.0	17.0	57.00	126.0	443.0
arrival_year	36275.0	2017.820427	0.383836	2017.0	2018.0	2018.00	2018.0	2018.0
arrival_month	36275.0	7.423653	3.069894	1.0	5.0	8.00	10.0	12.0
arrival_date	36275.0	15.596995	8.740447	1.0	8.0	16.00	23.0	31.0
repeated_guest	36275.0	0.025637	0.158053	0.0	0.0	0.00	0.0	1.0
no_of_previous_cancellations	36275.0	0.023349	0.368331	0.0	0.0	0.00	0.0	13.0
no_of_previous_bookings_not_canceled	36275.0	0.153411	1.754171	0.0	0.0	0.00	0.0	58.0
avg_price_per_room	36275.0	103.423539	35.089424	0.0	80.3	99.45	120.0	540.0
no_of_special_requests	36275.0	0.619655	0.786236	0.0	0.0	0.00	1.0	5.0

Figure 2: Statistical summary of the data

4 Exploratory Data Analysis

4.1 Univariate Analysis

4.1.1 Numerical columns

- **'no_of_adults'**

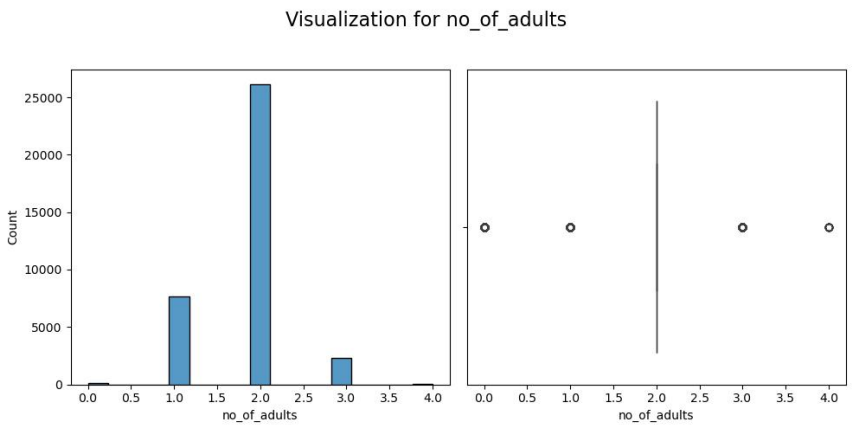


Figure 3: Histogram and boxplot of '**no_of_adults**' column

Observations

- **Bar Plot**: Most bookings are for 2 adults, likely due to couples, which warrants further investigation, followed by 1 adult and 3 adults. Very few bookings are for 0 and 4 adults.
- **Box Plot**: Median is 2 adults, with outliers at 0, 3, and 4; mean and median are overlapping.

Business Recommendations

- Optimize Services: Focus on services for bookings with 1-3 adults.
- Promote Family Stays: Create incentives for bookings with more adults.
- Analyze Outliers: Investigate bookings with 0 or 4 adults.
- Customer Feedback: Collect feedback to improve adult-related services.
- **'no_of_children'**

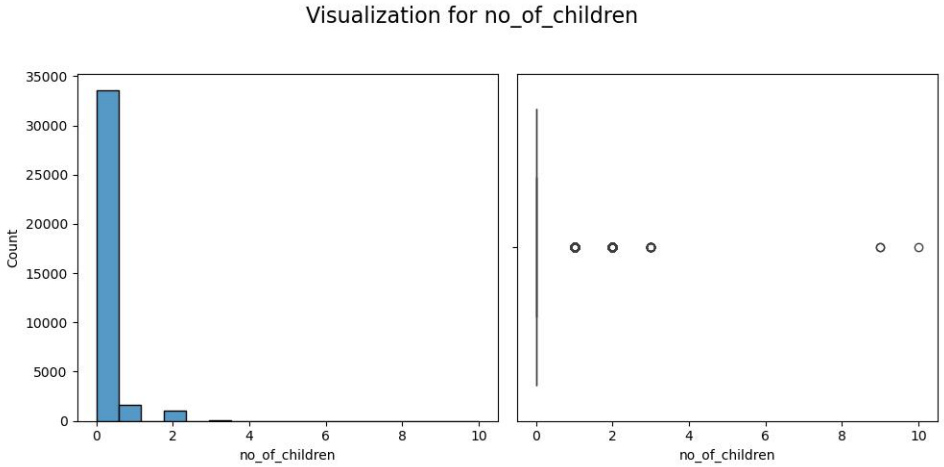


Figure 4: Histogram and boxplot of '**no_of_children**' column

Observations

- **Histogram**: Most individuals have 0 children, with fewer as the number of children increases.
- **Box Plot**: Median is 0 children; mean and median are overlapping, with outliers up to 10 children.

Business Recommendations

- Family-Oriented Services: Enhance services targeting families with children.
- Promote Family Packages: Create attractive packages for families with multiple children.
- Analyze Outliers: Understand needs of families with many children.
- Customer Feedback: Gather insights to improve family-related offerings.
- **no_of_weekend_nights**

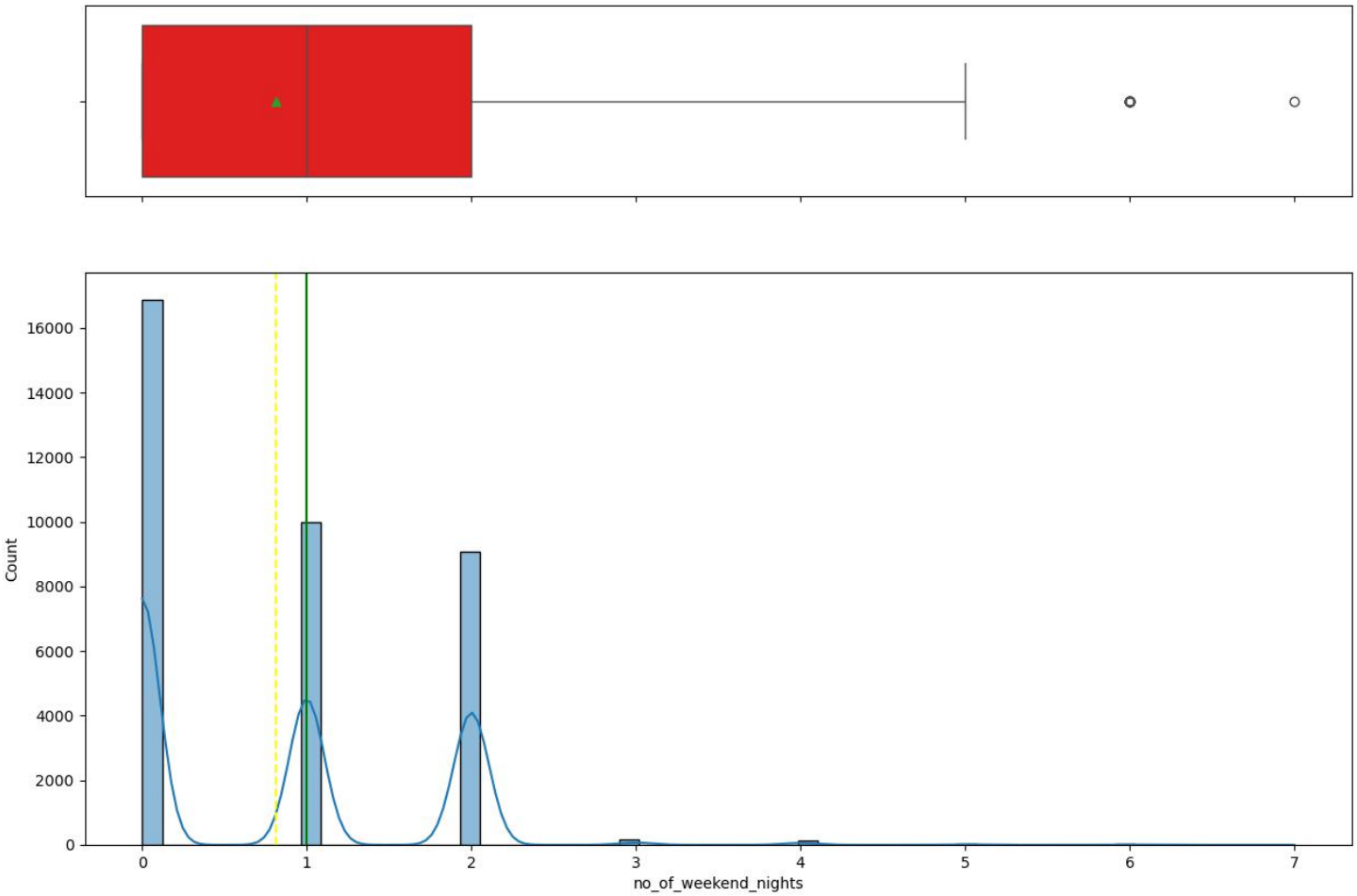


Figure 5: Histogram and boxplot of "**no_of_weekend_nights**" column

Observations

- **Histogram**: Most bookings have 0-1 weekend nights, with counts decreasing as nights increase.
- **Box Plot**: Median is 1 weekend night; mean and median are overlapping, with outliers at 6 and 7.

Business Recommendations

- Optimize Offers: Tailor offers for 0-1 weekend nights.
- Promote Longer Stays: Create incentives for longer weekend stays.
- Analyze Outliers: Investigate reasons for longer stays.
- Customer Feedback: Understand preferences to improve services.
- **no_of_week_nights**

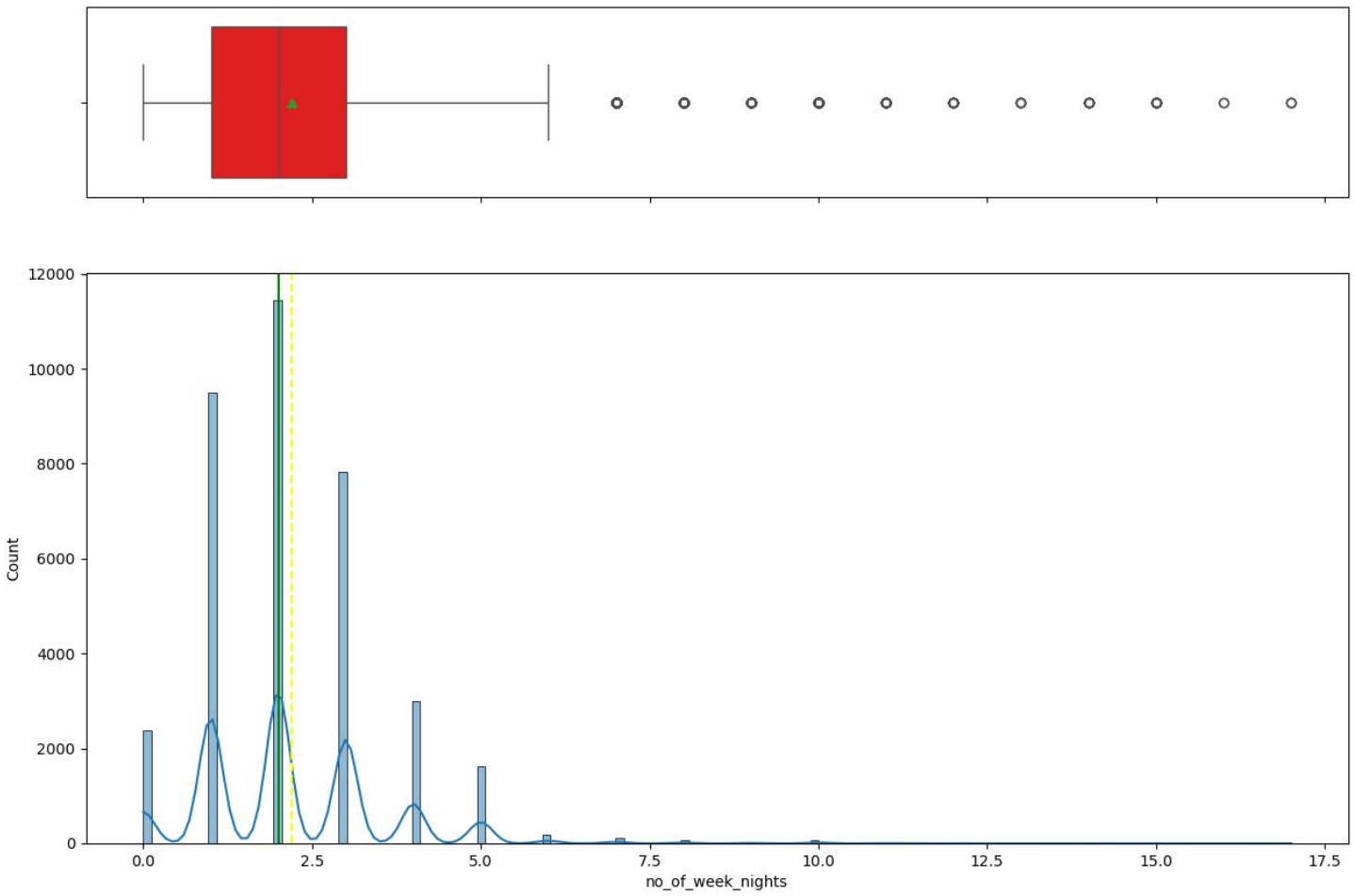


Figure 6: Histogram and boxplot of **'no_of_week_nights'** column

Observations

- **Histogram**: Most bookings have 1-3 week nights.
- **Box Plot**: Median is around 2 week nights; mean and median are overlapping, with several outliers.

Business Recommendations

- Optimize Offers: Tailor offers for stays of 1-3 week nights.
- Promote Longer Stays: Create incentives for longer week night stays.
- Analyze Outliers: Investigate reasons behind longer or shorter stays.
- Customer Feedback: Understand preferences to improve services.
- **required_car_parking_space**

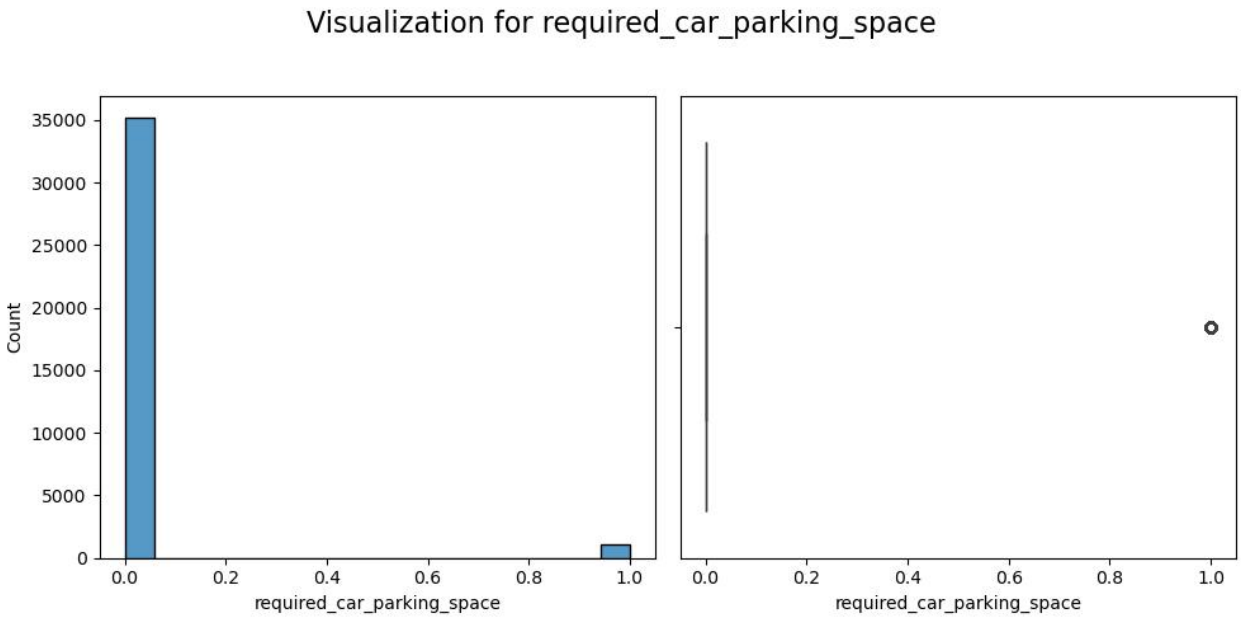


Figure 7: Histogram and boxplot of **'required_car_parking_space'** column

Observations

- **Histogram**: Majority have 0 required car parking spaces.
- **Box Plot**: Median is 0, with an outlier at 1; mean and median are overlapping.

Business Recommendations

- Optimize Parking: Maintain minimal parking space allocation.
- Address Outliers: Develop strategies for bookings needing parking.
- Customer Feedback: Collect feedback on parking needs.
- Promote Services: Highlight available parking facilities.
- **lead_time**

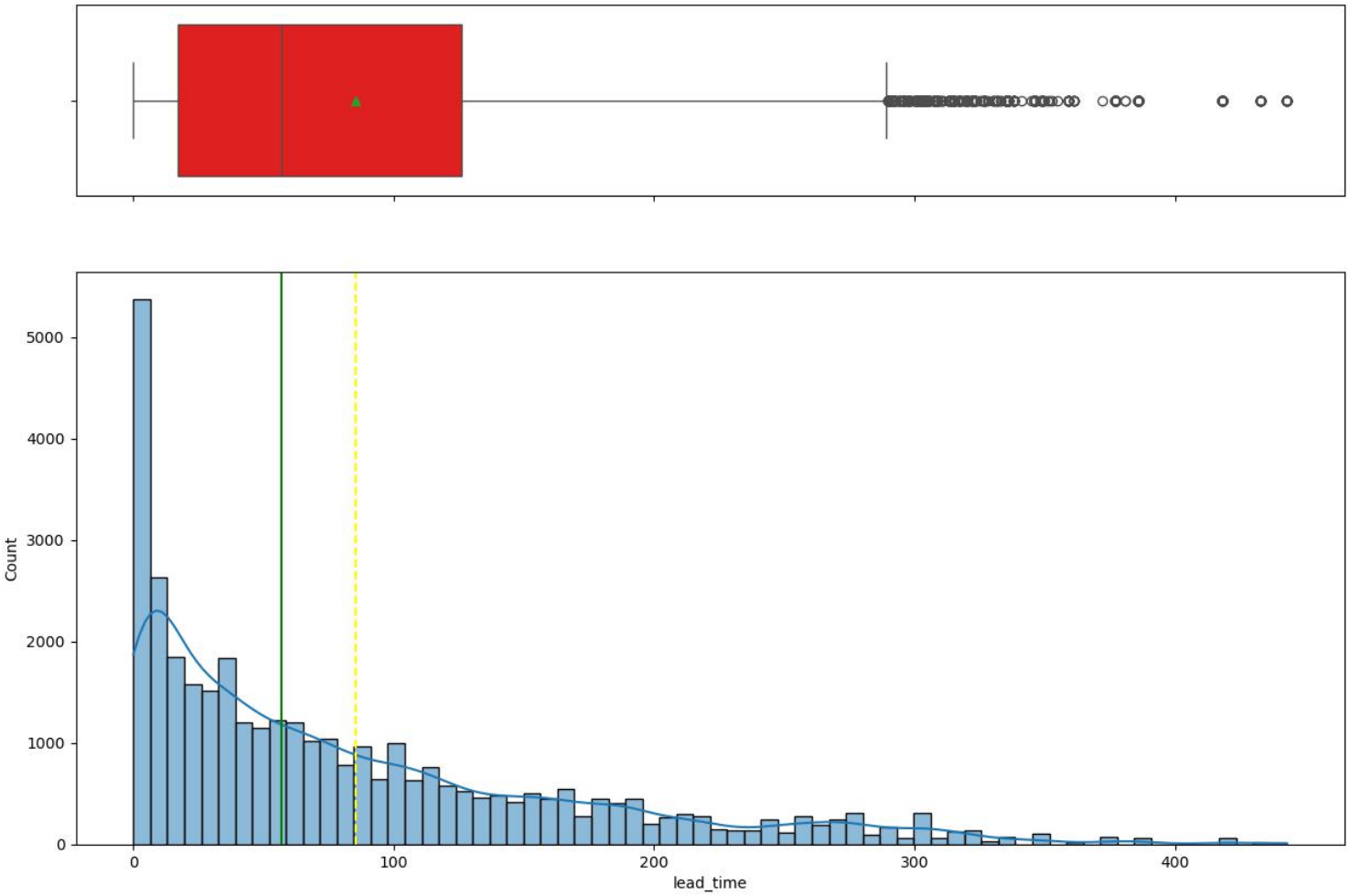


Figure 8: Histogram and boxplot of **'lead_time'** column

Observations

- **Histogram**: Lead times are mostly concentrated in the lower range, with frequency decreasing as lead time increases.
- **Box Plot**: Median lead time is low, with several high-value outliers.

Business Recommendations

- Promote Early Bookings: Encourage customers to book earlier to benefit from longer lead times.
- Analyze Outliers: Investigate reasons for unusually high lead times and address potential issues.

- **arrival_month**

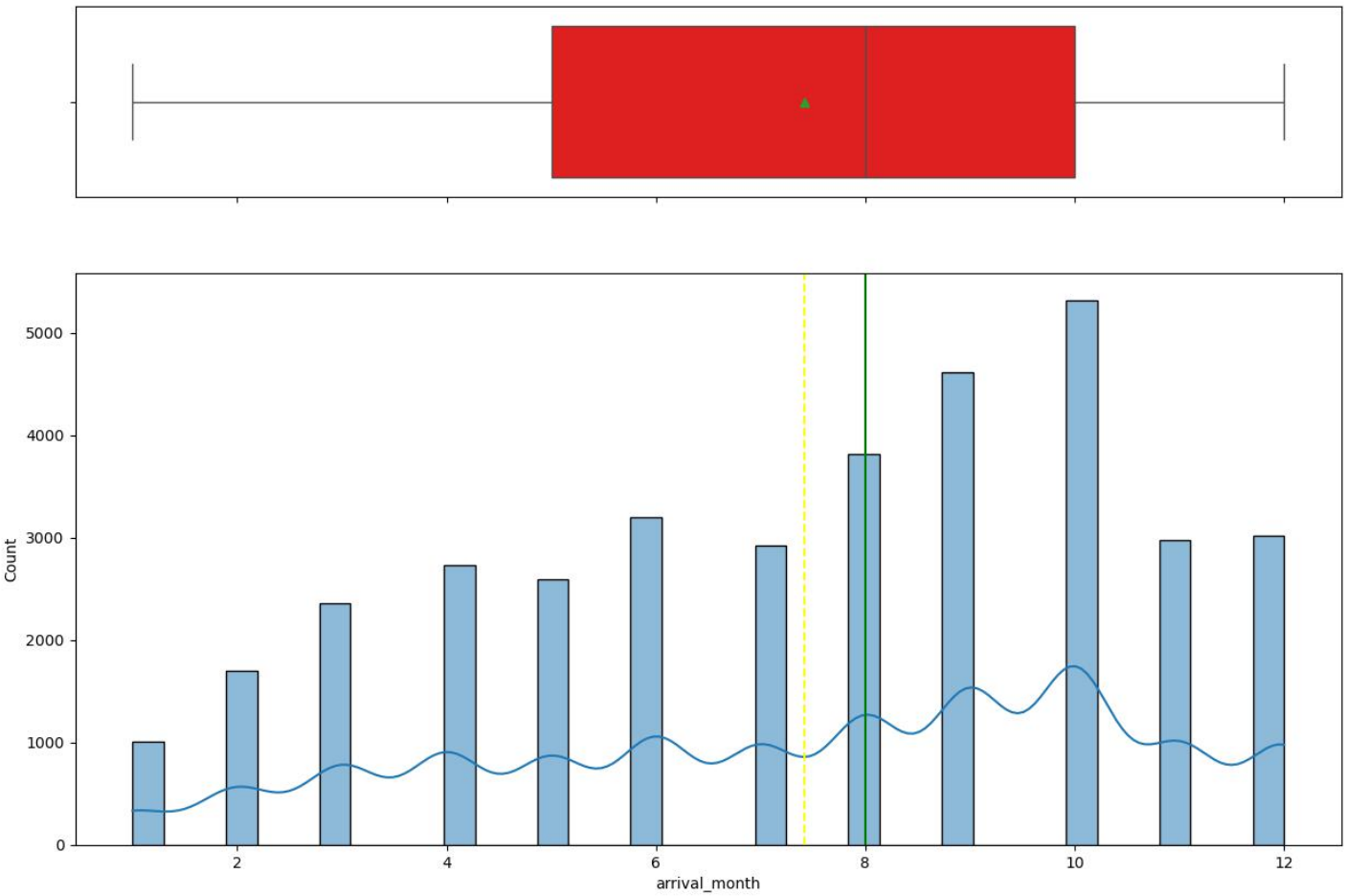


Figure 9: Histogram and boxplot of **'arrival_month'** column

Observations

- **Histogram**: Arrivals peak in months 8 and 10, with lower counts in other months.
- **Box Plot**: Median arrival month is mid-year; mean and median are overlapping, with outliers.

Business Recommendations

- Optimize Promotions: Target promotions for peak months to maximize bookings.
- Balance Off-Peak Months: Develop strategies to increase arrivals during off-peak months.

- Analyze Trends: Understand reasons for high arrivals in specific months.
- Customer Feedback: Gather feedback to improve services throughout the year.
- **repeated_guest**

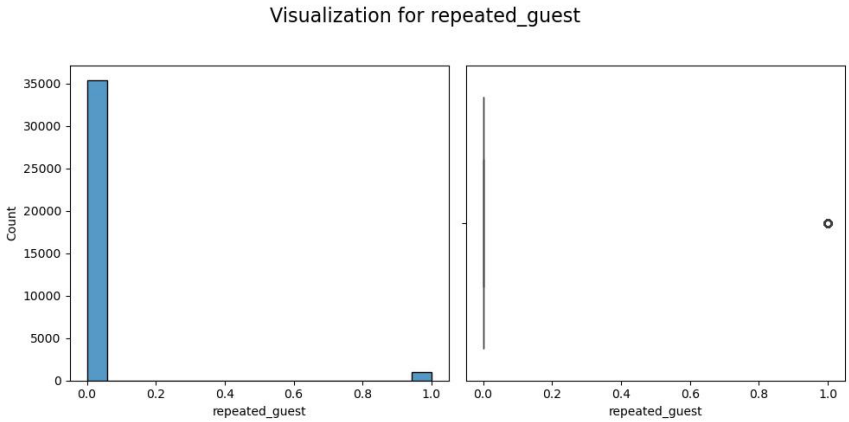


Figure 10: Histogram and boxplot of **'repeated_guest'** column

Observations

- **Histogram**: Most entries have 0 repeated guests.
- **Box Plot**: Median is 0, with one outlier at 1; mean and median are overlapping.

Business Recommendations

- Increase Repeat Guests: Develop loyalty programs to encourage repeat stays.
- Focus on Non-Repeat Customers: Analyze why most guests don't return and improve services.
- Promote Loyalty: Highlight benefits for repeated guests in marketing campaigns.
- Collect Feedback: Understand reasons behind the lack of repeat visits.
- **no_of_previous_cancellations**

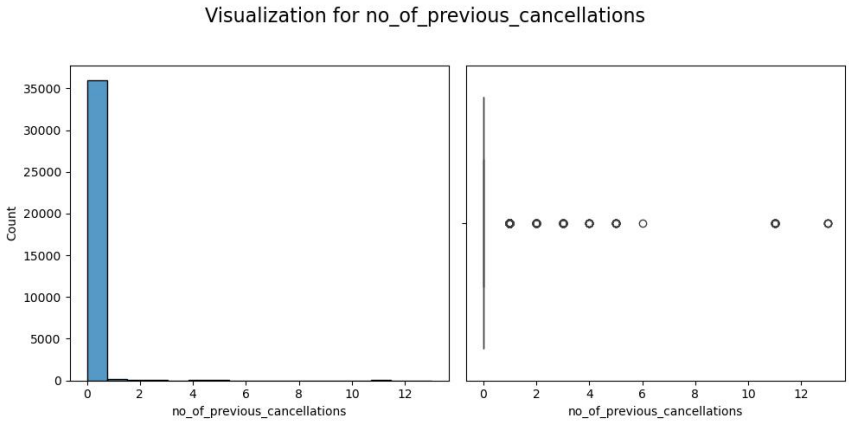


Figure 11: Histogram and boxplot of **'no_of_previous_cancellations'** column

Observations

- **Histogram**: Most entries have zero previous cancellations, with fewer as cancellations increase.
- **Box Plot**: The median is zero with several high-value outliers; mean and median are overlapping.

Business Recommendations

- Reduce Cancellations: Implement policies to minimize cancellations.
- Target Zero-Cancellation Customers: Focus marketing on customers with no previous cancellations.
- Analyze Outliers: Investigate reasons for high cancellation rates among outliers.
- Customer Feedback: Understand and address reasons for cancellations.
- **avg_price_per_room**

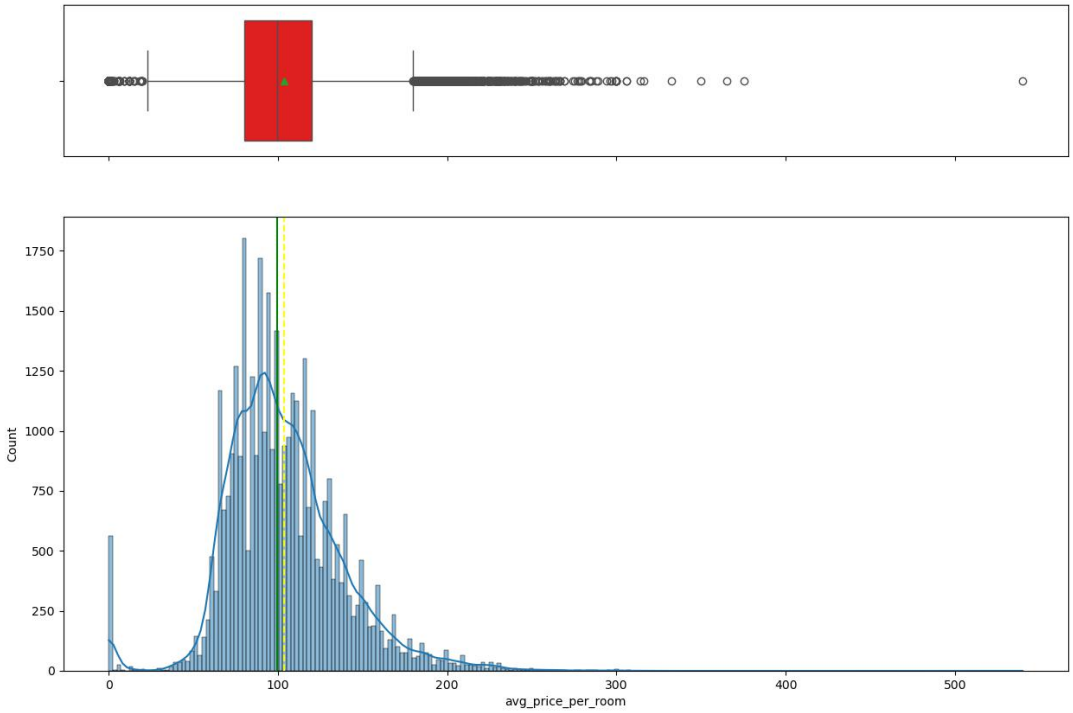


Figure 12: Histogram and boxplot of **'avg_price_per_room'** column

Observations

- **Histogram**: Most rooms have average prices in the lower range, with fewer as price increases.
- **Box Plot**: Median price is low with high-value outliers; mean and median are overlapping.

Business Recommendations

- Optimize Prices: Adjust strategy to target mid-range prices.
- Promote High-Value Rooms: Market unique features to attract premium customers.
- Segment Customers: Tailor promotions for different price segments.
- Gather Feedback: Ensure pricing is competitive and perceived well.
- **no_of_special_requests**

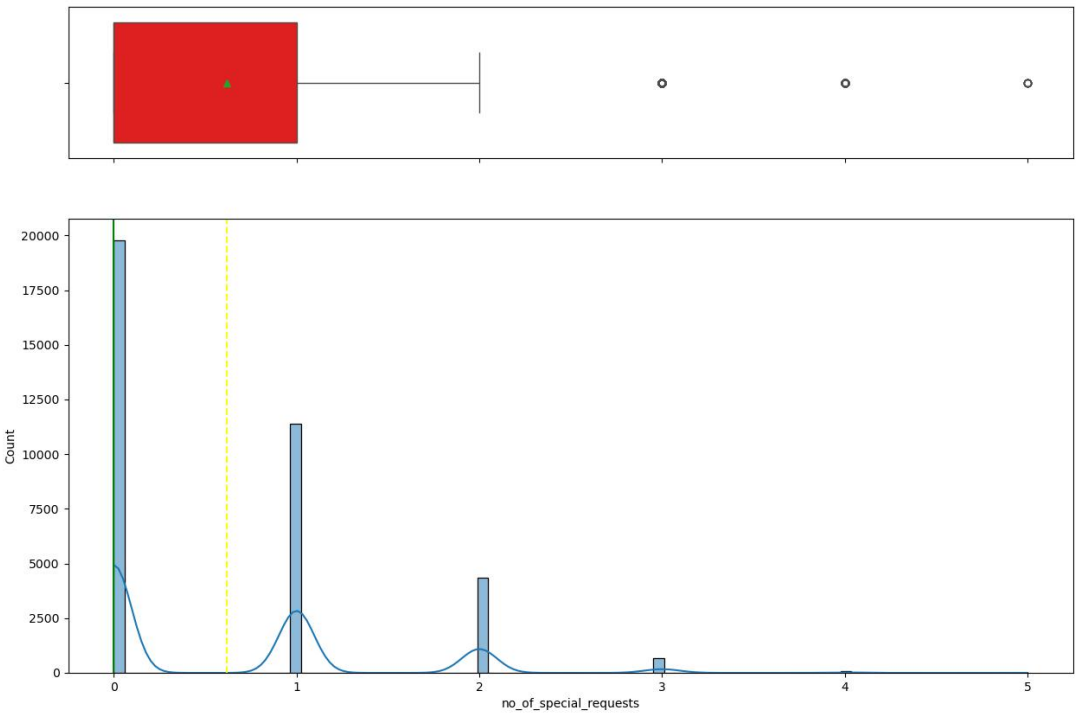


Figure 13: Histogram and boxplot of **'no_of_special_requests'** column

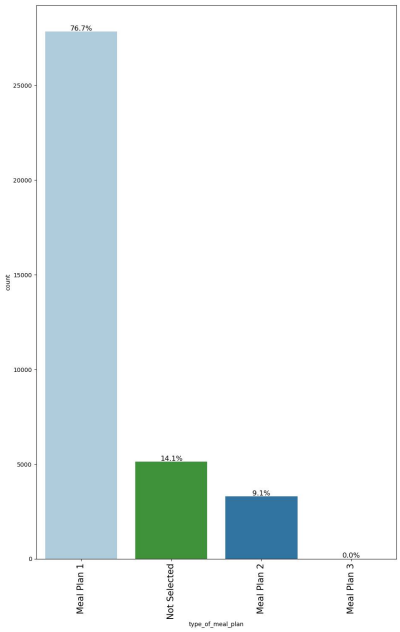
Observations

- **Histogram**: Most bookings have 0-1 special requests.
- **Box Plot**: Median is 1 special request, with few outliers.

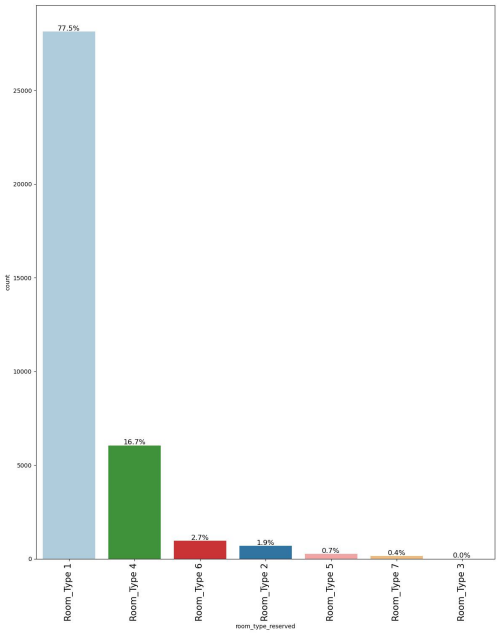
Business Recommendations

- Streamline Processes: Efficiently handle 0-1 special requests.
- Manage Outliers: Develop strategies for higher request bookings.
- Collect Feedback: Understand needs of customers with requests.
- Promote Special Services: Encourage more special requests.

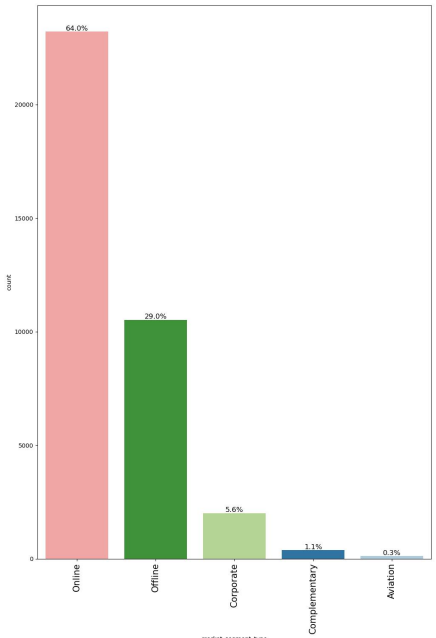
4.1.2 Categorical columns



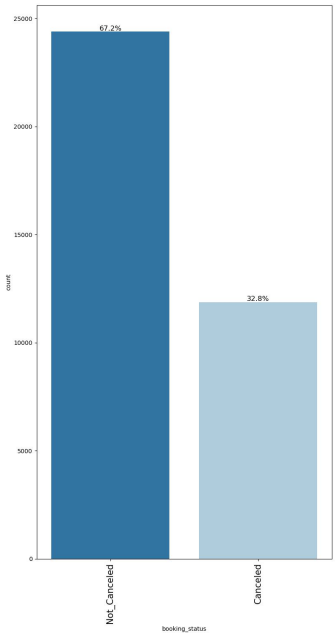
(a) '**type_of_meal_plan**'



(b) '**room_type_reserved**'



(c) '**market_segment_type**'



(d) **'booking_status'**

Figure 14: Barchart of '**type_of_meal_plan**', **'room_type_reserved'**,**'market_segment_type'** and **'booking_status'** column

Observations

- **Room Type Reservations**:
  - Room_Type 1 has the highest number of reservations, accounting for 77.5% of the total.
  - Room_Type 4 follows with 16.7%.
  - Other room types (Room_Type 6, Room_Type 2, Room_Type 5, Room_Type 7) have significantly fewer reservations, and Room_Type 3 has 0% reservations.
- **Booking Status**:
  - Not_Canceled bookings make up 67.2% of the total.
  - Canceled bookings account for 32.8%.
- **Market Segment**:
  - Online segment dominates with 64.0%.
  - Offline follows with 29.0%.
  - Corporate segment is at 5.6%.
  - Complementary and Aviation are minimal, with 1.1% and 0.3% respectively.
- **Meal Plan Selection**:
  - Meal Plan 1 is the most popular, selected by 76.7% of customers.
  - Not Selected is at 14.1%.
  - Meal Plan 2 has a share of 9.1%.
  - Meal Plan 3 has 0% selections.

Business Recommendations

- **Room Type Reservations**:
  - Focus on Room_Type 1 and 4: Ensure these popular rooms are well-maintained and available.
  - Investigate Room_Type 3: Analyze the lack of reservations and consider adjustments to pricing, location, or amenities.
  - Promote Less Reserved Rooms: Develop marketing strategies to boost reservations for less popular room types.
  - Customer Feedback: Collect and analyze feedback to improve the features of less popular room types.
- **Booking Status**:
  - Enhance Customer Experience: Analyze and address factors leading to cancellations.
  - Flexible Policies: Introduce flexible booking and cancellation policies.
  - Targeted Marketing: Promote the benefits of confirmed bookings and offer incentives.
  - Customer Feedback: Gather feedback to understand and address pain points.
- **Market Segment**:
  - Strengthen Online Presence: Invest in online marketing and digital engagement.
  - Boost Offline Engagement: Enhance the offline experience with promotions and loyalty programs.
  - Corporate Partnerships: Develop strategies to attract corporate clients with tailored packages and discounts.
  - Promote Niche Segments: Create specialized marketing campaigns for complementary and aviation segments.
- **Meal Plan Selection**:

- Enhance Meal Plan 1: Maintain quality and add appealing options.
- Understand Non-Selection Reasons: Investigate why some customers did not choose any meal plan.
- Promote Meal Plan 2 and 3: Develop strategies to make these plans more attractive.
- Customer Feedback and Customization: Offer customizable meal plan options based on customer feedback.

4.2 Bivariate Analysis

4.2.1 Numerical variables

- **Heatmap**

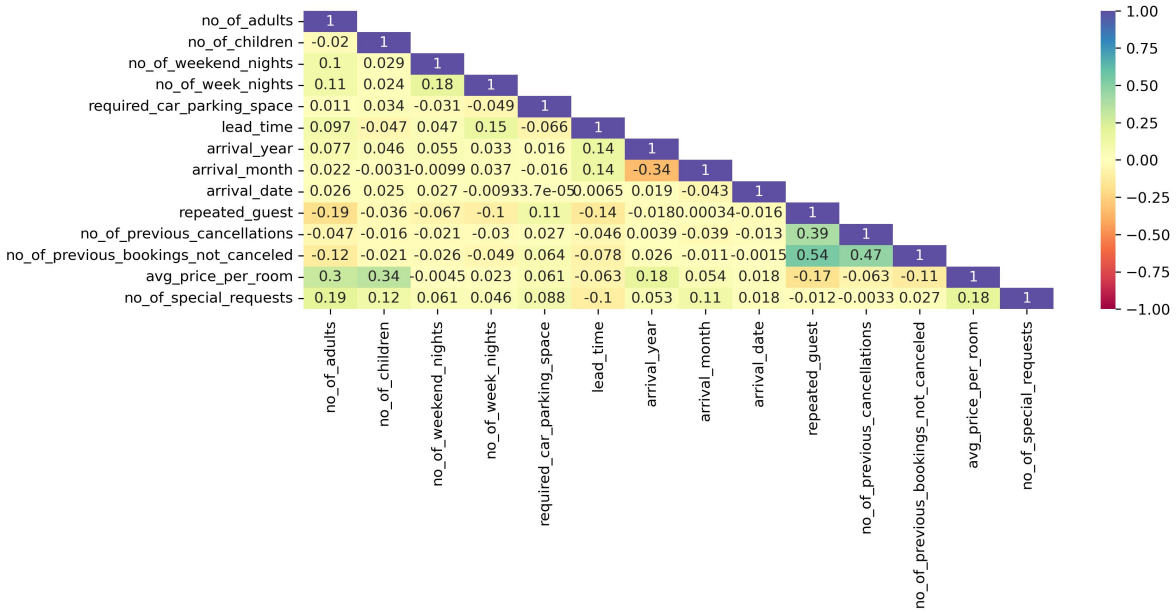


Figure 15: Heatmap of all numerical variables

Correlation Heatmap

Observations

- Most features show low correlations, suggesting minimal multicollinearity.
- Notable correlations: moderate positive correlation (0.54) between `no_of_previous_bookings_not_canceled` and `no_of_previous_cancellations`, weak correlations between `avg_price_per_room` and `no_of_adults` (0.30), `no_of_children` (0.34).
- Negative correlation of `lead_time` with `repeated_guest` suggests longer lead times are linked to non-repeated guests.

Business Recommendations

- Target repeated guests by analyzing their booking behaviors and preferences.
- Optimize lead time strategies to encourage repeat bookings, as longer lead times may decrease loyalty.
- Investigate the relationship between `no_of_special_requests` and `avg_price_per_room` to improve cancellation management and guest satisfaction.
- **Pairplot**

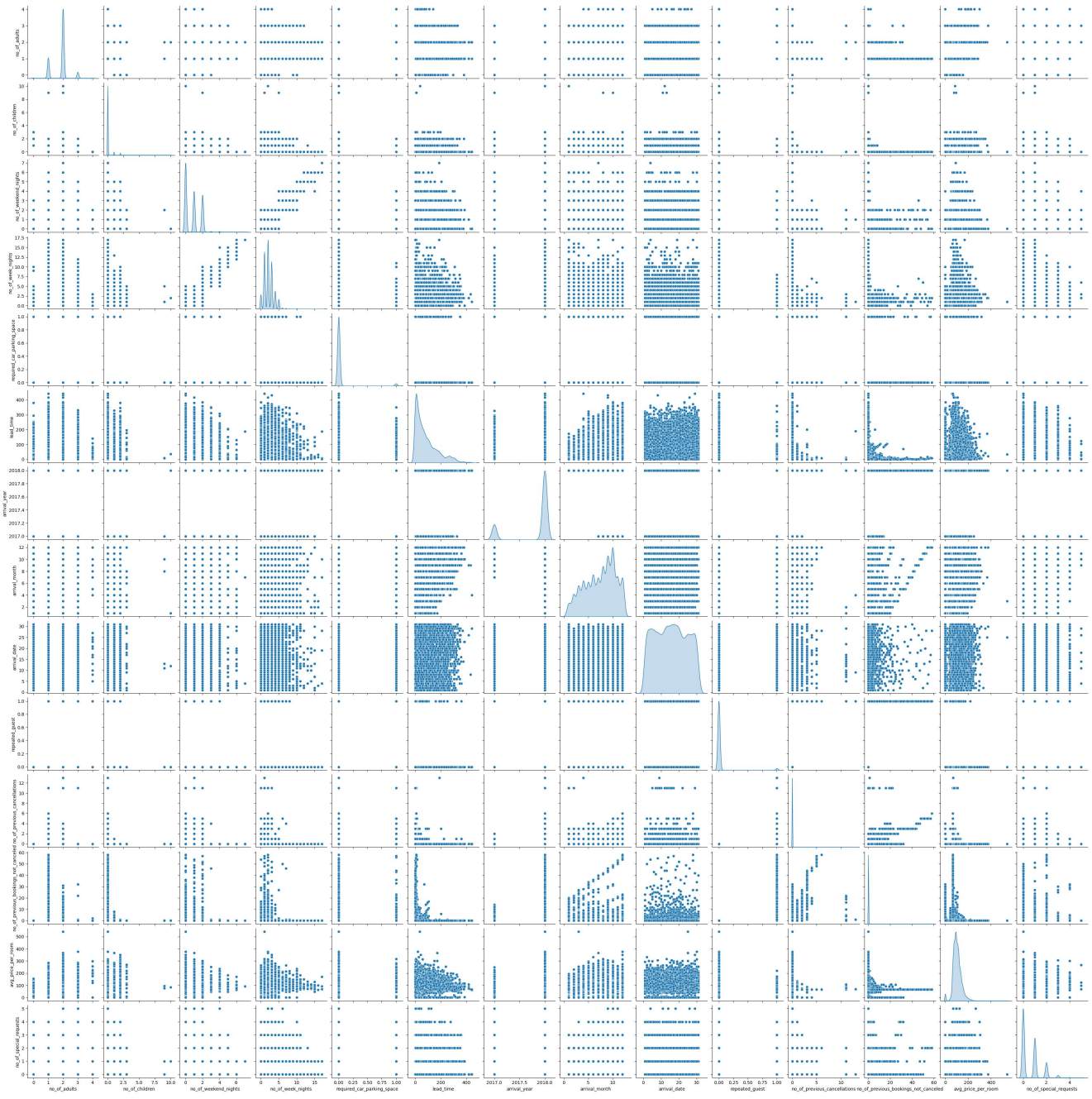


Figure 16: Pairplot of all numerical variables

Pair Plot

Observations

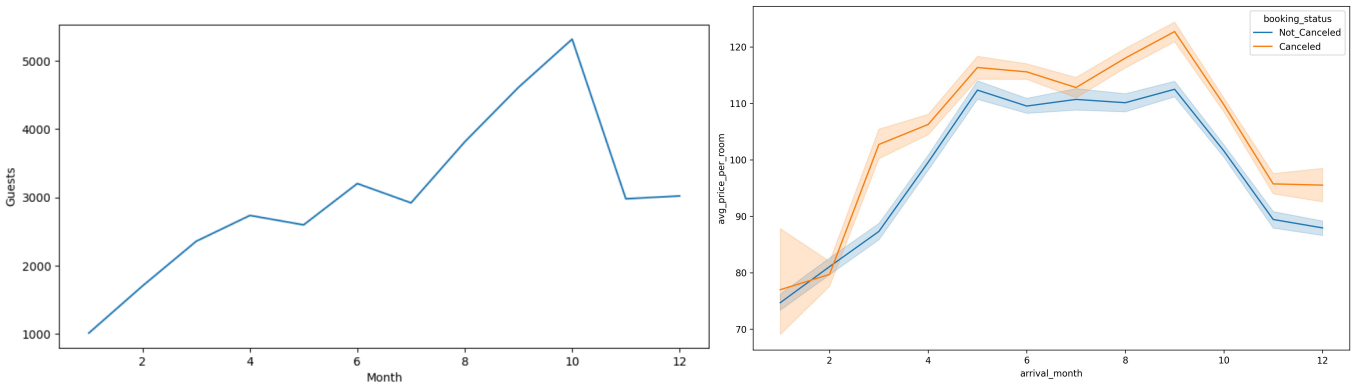
- Non-linear distributions and clustering observed in features like `no_of_special_requests`, `no_of_weekend_nights`, and `no_of_previous_bookings`.
- Most numerical features are highly skewed as seen in density plots.

Business Recommendations

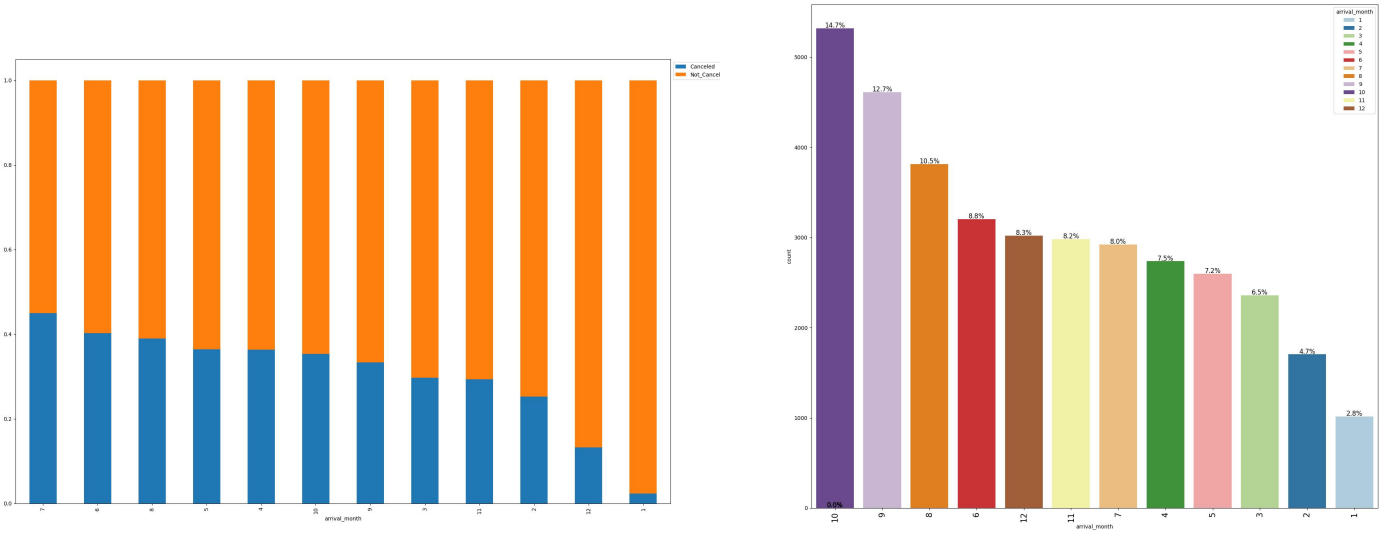
- Investigate clustering in features like `no_of_weekend_nights` and customize marketing efforts based on distinct customer groups.
- Utilize clustering in `avg_price_per_room` for targeted pricing strategies.

4.2.2 Categorical vs numerical variables

- What are the busiest months in the hotel?



(a) No. of guests vs month(**Left**), Avg. price per room vs month(**Right**)



(b) Canceled vs not canceled(**Left**) and Total counts for each month

booking_status	Canceled	Not_Canceled	All
arrival_month			
All	11885	24390	36275
10	1880	3437	5317
9	1538	3073	4611
8	1488	2325	3813
7	1314	1606	2920
6	1291	1912	3203
4	995	1741	2736
5	948	1650	2598
11	875	2105	2980
3	700	1658	2358
2	430	1274	1704
12	402	2619	3021
1	24	990	1014

(c) Table

Figure 17: Avg. price per room and guests vs months

Observations and Recommendations

Observations:

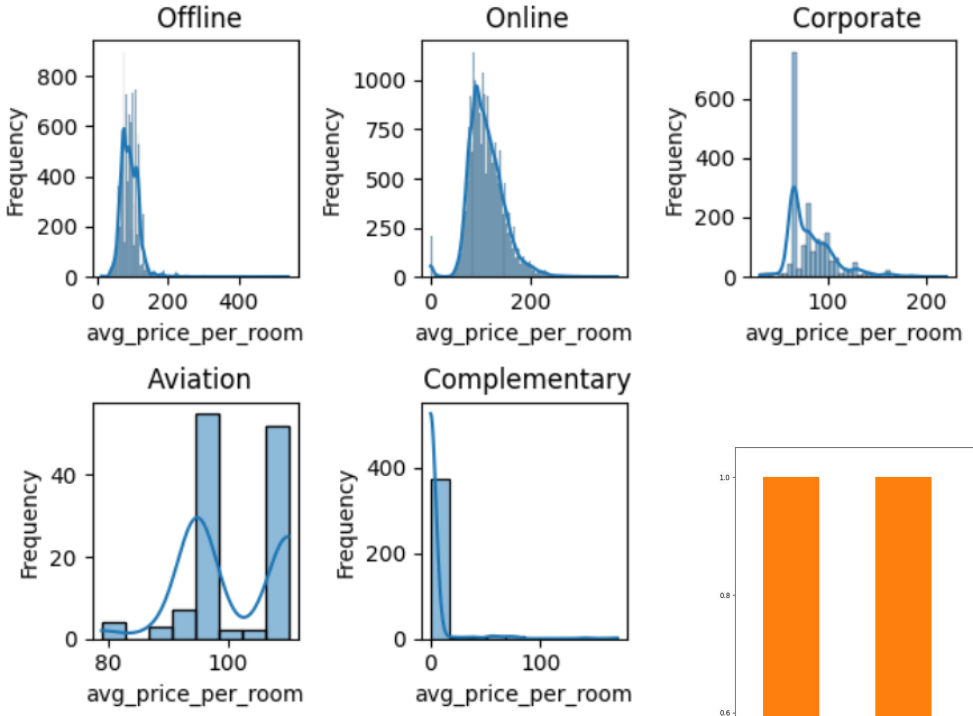
- **Busiest Months:** August, September, and October show peak guest activity.
- **Cancellations:** Higher cancellation rates during peak months; lower in January, February, and December.
- **Average Price:** Peaks in August–October; canceled bookings show slightly higher prices.
- **Off-Peak:** Minimal bookings in January, February, and December.

Recommendations:

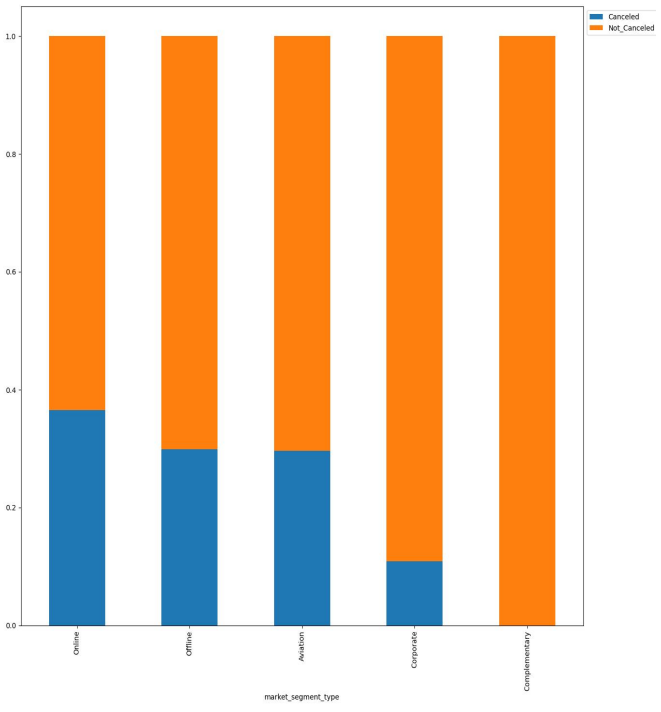
- **Peak Months:** Implement dynamic pricing, minimum stay policies, and deposit-based confirmations.
- **Off-Peak:** Launch discounts, bundles, and targeted campaigns.
- **Price Sensitivity:** Offer flexible pricing strategies to reduce cancellations.
- **Capacity Planning:** Optimize resources for peak months and reduce costs during off-peak seasons.
- **'market_segment_type'**

Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?

Which market segment do most of the guests come from?



booking_status	Canceled	Not_Canceled	All
market_segment_type			
All	11885	24390	36275
Online	8475	14739	23214
Offline	3153	7375	10528
Corporate	220	1797	2017
Aviation	37	88	125
Complementary	0	391	391



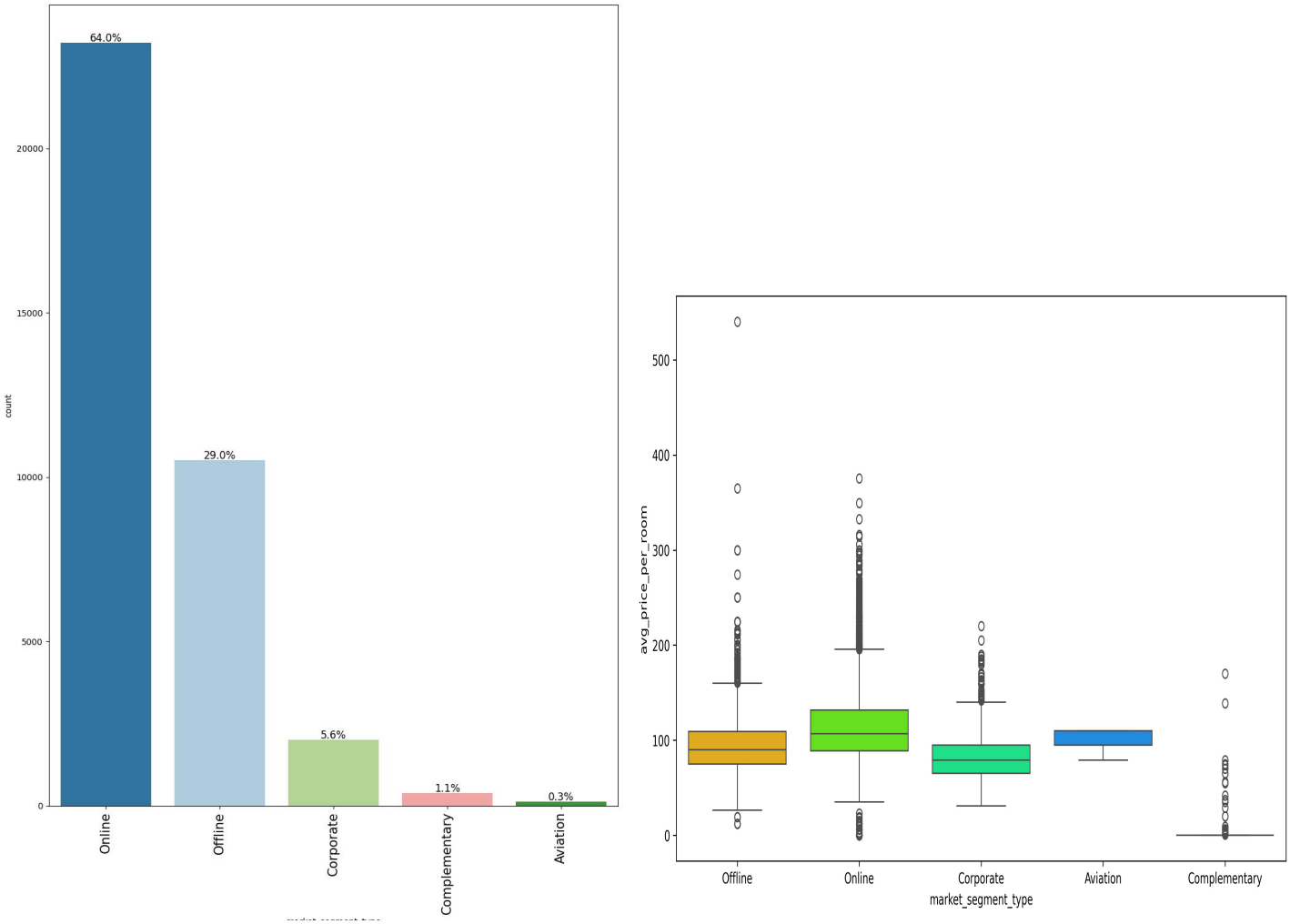


Figure 18: Boxplot,barplot and histplot for various aspects across different **market_segment_type**

Observations

Most of the guests come from online followed by offline and other market segments.

- **Price Distributions:** Online, Offline, and Corporate segments show right-skewed distributions, peaking around 100 INR. Aviation and Online segments have the highest average price per room, while Complementary has the lowest.
- **Cancellation Trends:** Online bookings have the highest cancellation rate due to flexible policies. Offline and Corporate segments have fewer cancellations, with no cancellations in Complementary bookings. Canceled rooms have higher average price per room.
- **Outliers:** Outliers are present in all segments, especially in Online and Offline bookings, indicating some high-price anomalies.

Insights

- **Online Segment:** High cancellations suggest stricter refund policies or incentivizing non-refundable bookings.
- **Corporate and Complementary Segments:** Stable bookings indicate opportunities for long-term contracts or promotions.
- **Operational Strategy:** Predict cancellations in advance to adjust pricing, reallocate resources, and minimize revenue loss.
- **'repeated_guest' vs 'booking_status'**

1. What percentage of bookings are canceled?
2. Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?

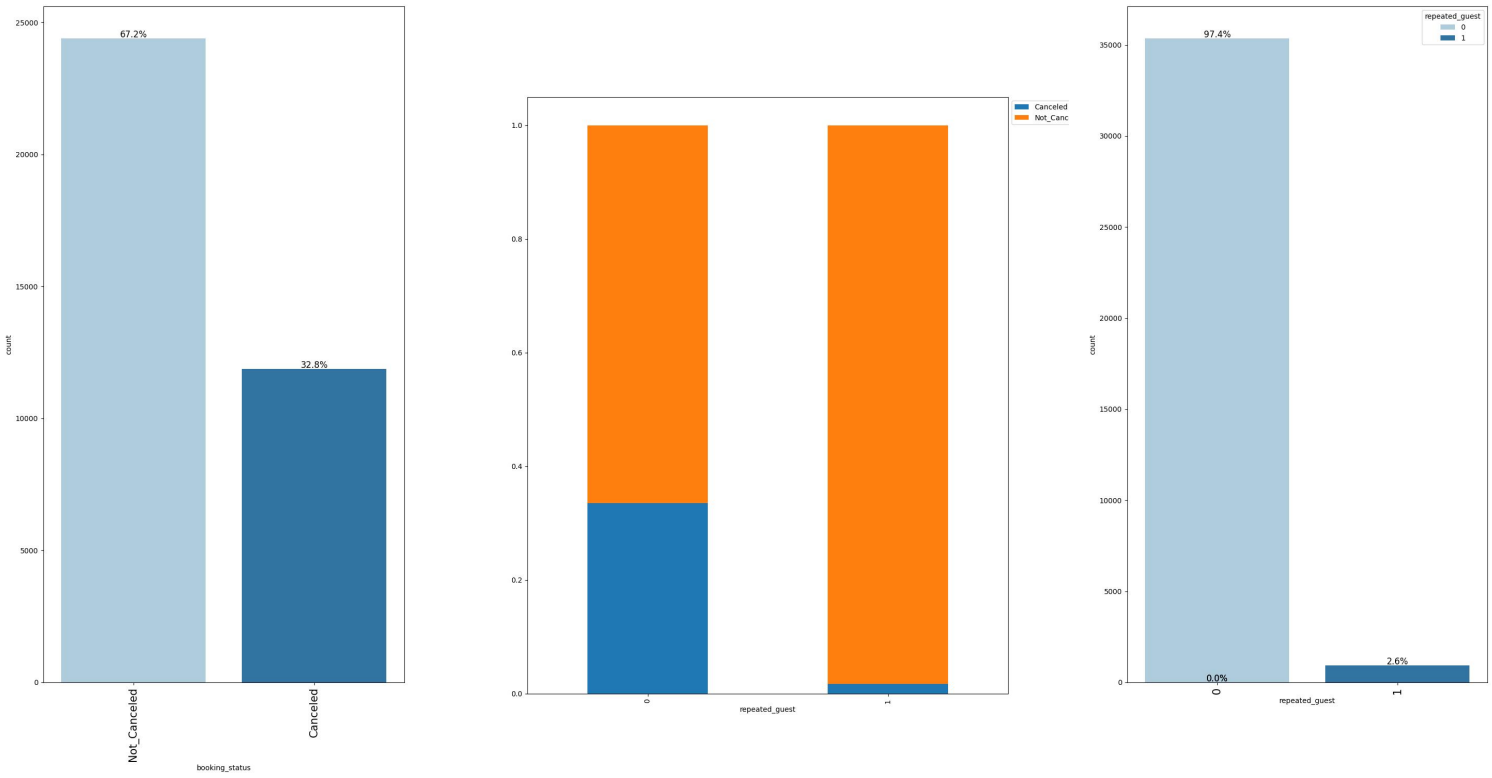


Figure 19: Barchart for various aspects of **repeated guest**

Observations and Recommendations

Observations

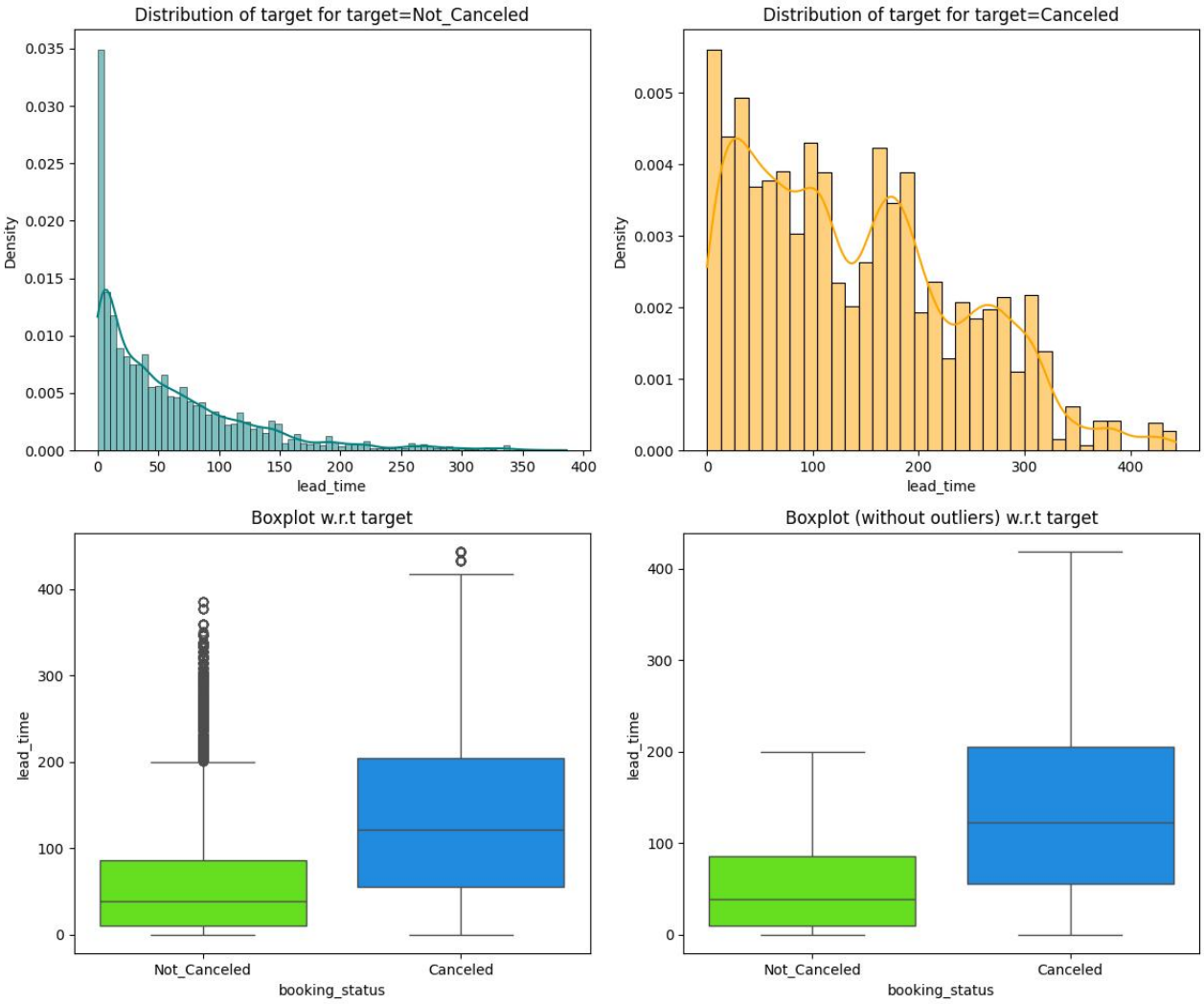
- **High Cancellation Rate:** 32.8% of bookings are canceled, impacting revenue and operations.
- **Repeating Guests Are Rare:** Only 2.6% of bookings come from repeat guests, who show low cancellation rates.

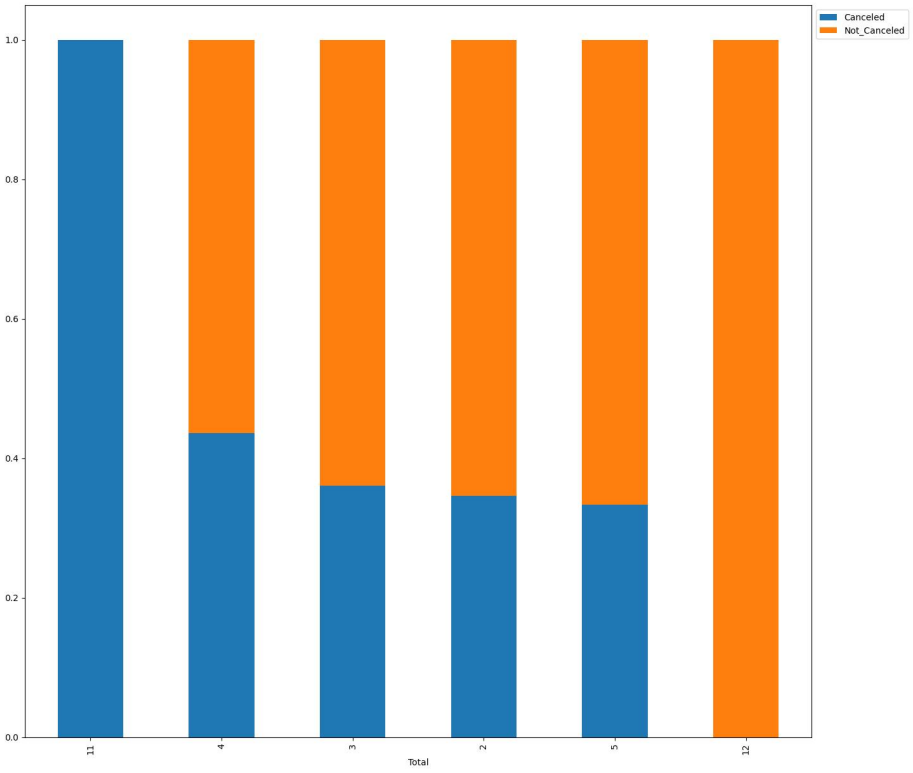
- **First-Time Guests Drive Cancellations:** Most cancellations arise from first-time guests, showing low reliability.

Business Recommendations

- **Customer Retention:** Implement loyalty programs, discounts, and personalized offers to convert first-time guests into repeat customers.
- **Reduce Cancellations:**
  - Introduce stricter policies (e.g., non-refundable deposits, tiered cancellation fees).
  - Provide small discounts for confirmed, non-cancelable bookings.
- **Enhance Guest Experience:** Identify reasons for cancellations and address them with flexible booking options or better pre-arrival communication.
- **Leverage Data Insights:** Segment guests to predict cancellations and optimize pricing strategies.
- **Promote Loyalty Programs:** Offer perks like room upgrades, priority booking, and exclusive deals to encourage repeat guests.
- **'lead_time' vs 'booking_status'**

Let's also study the customers who traveled with their families and analyze the impact on booking status.¶





booking_status	Canceled	Not_Canceled	All
Total			
All	9985	18456	28441
2	8213	15506	23719
3	1368	2425	3793
4	398	514	912
5	5	10	15
11	1	0	1
12	0	1	1

Figure 20: Boxplot and histplot for various aspects across different **lead_time** (Total represents total family members)

Observations and Recommendations

1. Lead Time Distribution

- **Not_Canceled:** Most bookings have lead_time < 50 days (right-skewed).
- **Canceled:** Wider spread; high lead_time (> 150 days) correlates with cancellations.

Recommendations:

- Incentivize early decisions (e.g., discounts, flexible terms).
- Enforce stricter cancellation policies for high lead_time bookings.

2. Boxplot Analysis

- **Not_Canceled:** Median lead_time ~ 30 days.
- **Canceled:** Median ~ 150 days; significant outliers (> 400 days).

Recommendations:

- Engage early bookers proactively (reminders, offers).
- Use predictive models to flag risky bookings.

3. Family Size and Cancellations

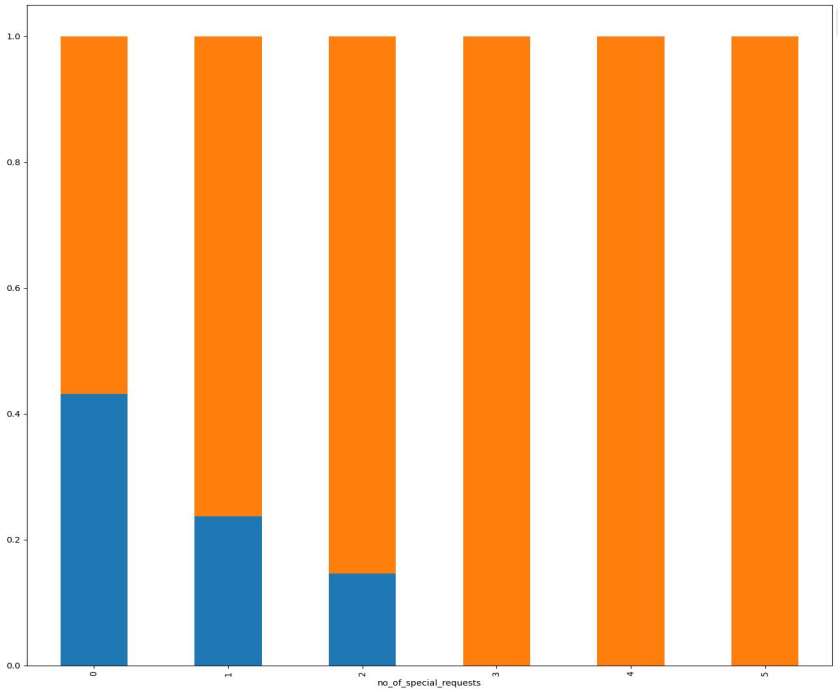
11 and 12 family sizes are outliers as they have only 1 case each so we can safely ignore them for our analysis.

- **Large Families (3-4):** Higher cancellations.
- **Small Families (1-2):** More reliable bookings.

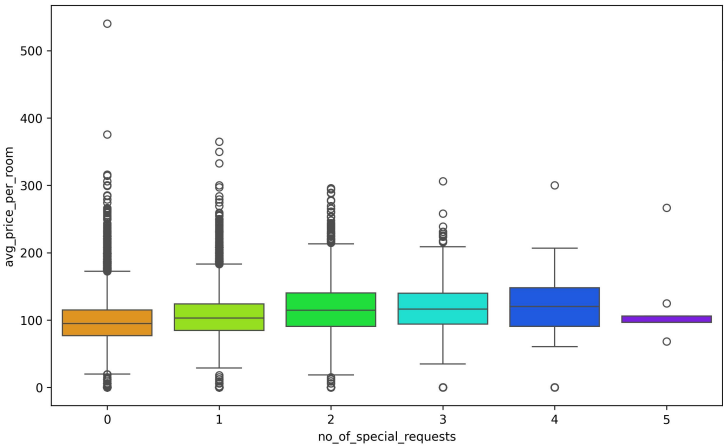
Recommendations:

- Personalized offers for large families.
- Promote group travel benefits to smaller families.

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



booking_status	Canceled	Not_Canceled	All
no_of_special_requests			
All	11885	24390	36275
0	8545	11232	19777
1	2703	8670	11373
2	637	3727	4364
3	0	675	675
4	0	78	78
5	0	8	8



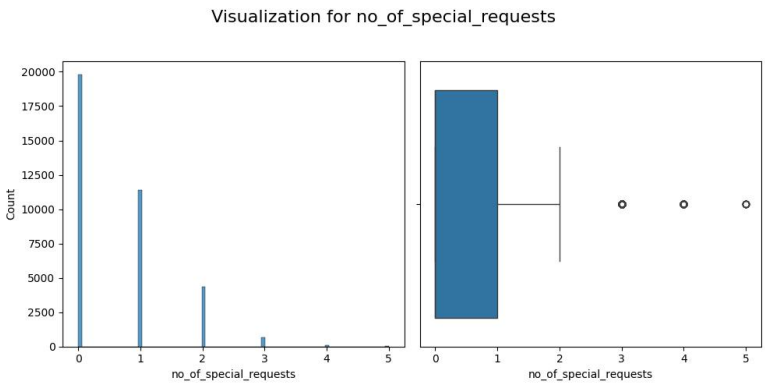


Figure 21: Boxplot and histplot for various aspects across **no. of special requests**

Observations and Recommendations

1. Cancellations and Number of Special Requests

- **0–1 Requests:** High cancellation rates, particularly for bookings with **0 requests** (40%).
- **3+ Requests:** Very low cancellation rates, indicating higher commitment to bookings.

Recommendations:

- Provide **incentives** (e.g., discounts or flexible policies) for bookings with 0–1 requests to reduce cancellations.
- Encourage upfront deposits to ensure serious bookings for guests with fewer requests.

2. Average Price per Room and Special Requests

- **0–1 Requests:** Lower average price with greater variability.
- **3+ Requests:** Higher and more stable room prices with minimal variation.

Recommendations:

- Personalize services and offers for high-value guests with **3+ requests** to ensure satisfaction.
- Optimize strategies to **increase value** from low-request bookings.

3. Distribution of Special Requests

- **Majority Group:** Most bookings have **0–1 special requests**.
- **Outliers:** Bookings with **4–5 requests** are fewer but demonstrate strong reliability and higher value.

Recommendations:

- Address cancellation factors for the majority group (0–1 requests) through targeted solutions.
- Enhance services for guests with **4–5 requests** to improve retention and revenue.

5 Data preprocessing

The dataset contains no missing or duplicate values. The outliers are significant for the data so we don't require to remove them. The data is scaled with standard scaler function.

The `StandardScaler` function standardizes features by removing the mean and scaling to unit variance. The formula used is:

$$z = \frac{(x - \mu)}{\sigma}$$

where x is the original feature, μ is the mean, and σ is the standard deviation.

We have splited the data for training and testing purposes. The model summary is as follows.

6 Model building and Model Performance Improvement

6.1 Logistic Regression

6.1.1 Model 1

Terms and Examples

- **True Positive (TP)**: Correctly predicted cancellation (1).
- **False Positive (FP)**: Predicted cancellation (1), but not canceled (0).
- **False Negative (FN)**: Predicted not canceled (0), but canceled (1).
- **True Negative (TN)**: Correctly predicted not canceled (0).

Implications

- **False Positive (FP)**: Predicts cancellation wrongly, causing unnecessary actions.
- **False Negative (FN)**: Misses actual cancellation, leading to missed opportunities.

Conclusion

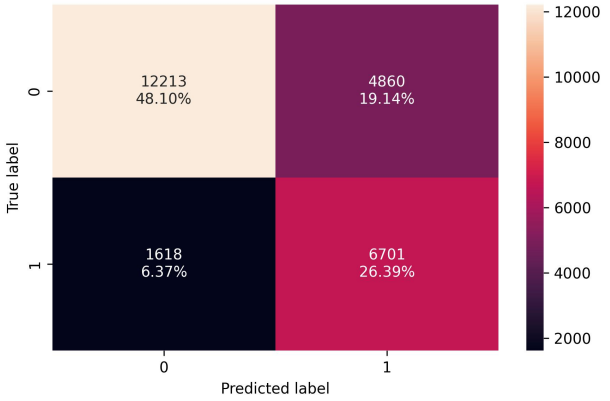
- **Minimize Missed Cancellations**: Focus on reducing FN, increasing Recall.
- **Avoid Unnecessary Actions**: Focus on reducing FP, increasing Precision.

```
VIF values:

const                                    NaN
no_of_adults                        1.338514
no_of_children                      2.020921
no_of_weekend_nights                1.066945
no_of_week_nights                   1.095423
required_car_parking_space          1.035690
lead_time                           1.387228
arrival_year                        1.428618
arrival_month                       1.276298
arrival_date                        1.006743
repeated_guest                      1.784037
no_of_previous_cancellations        1.367466
no_of_previous_bookings_not_canceled 1.637137
avg_price_per_room                  2.030696
no_of_special_requests              1.250267
type_of_meal_plan_Meal Plan 2       1.263369
type_of_meal_plan_Meal Plan 3       1.025579
type_of_meal_plan_Not Selected      1.273438
room_type_reserved_Room_Type 2      1.096867
room_type_reserved_Room_Type 3      1.003266
room_type_reserved_Room_Type 4      1.362944
room_type_reserved_Room_Type 5      1.030519
room_type_reserved_Room_Type 6      1.992228
room_type_reserved_Room_Type 7      1.119514
market_segment_type_Complementary   4.656644
market_segment_type_Corporate      17.156828
market_segment_type_Offline        66.169703
market_segment_type_Online         73.248558
dtype: float64
```

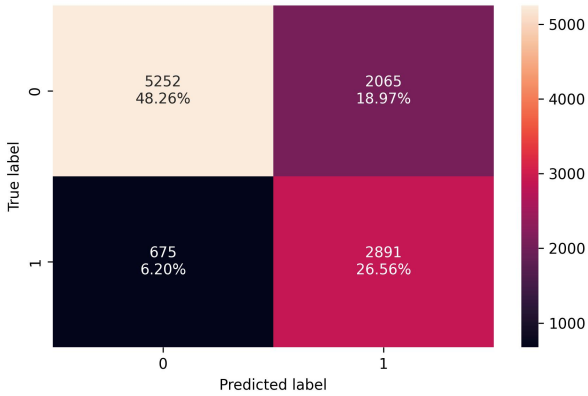
```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                         Logit   Df Residuals:                    25376
Method:                          MLE   Df Model:                           15
Date:               Sun, 15 Dec 2024   Pseudo R-squ.:                  0.1554
Time:                       05:25:06   Log-Likelihood:                -13564.
converged:                     True   LL-Null:                       -16060.
Covariance Type:           nonrobust   LLR p-value:                     0.000
================================================================================
                                   coef    std err          z      P>|z|
--------------------------------------------------------------------------------
no_of_adults                     0.0512      0.016      3.172      0.002
no_of_children                   0.0829      0.020      4.096      0.000
no_of_weekend_nights             0.1249      0.015      8.269      0.000
no_of_week_nights                0.0763      0.015      4.969      0.000
required_car_parking_space      -0.1469      0.016     -8.943      0.000
lead_time                        1.1526      0.021     55.918      0.000
arrival_year                     0.1503      0.018      8.336      0.000
arrival_month                   -0.0795      0.016     -4.838      0.000
avg_price_per_room               0.6196      0.021     29.873      0.000
no_of_special_requests          -0.7004      0.017    -40.157      0.000
type_of_meal_plan_Meal Plan 2   -0.0928      0.017     -5.506      0.000
type_of_meal_plan_Not Selected   0.2081      0.015     13.886      0.000
room_type_reserved_Room_Type 5  -0.0538      0.015     -3.657      0.000
room_type_reserved_Room_Type 6  -0.1126      0.021     -5.481      0.000
room_type_reserved_Room_Type 7  -0.0737      0.017     -4.361      0.000
market_segment_type_Complementary  0.1450   0.020      7.315      0.000
--------------------------------------------------------------------------------
```

Figure 22: Model summary




```
Model Performance (Significant Features):
   Accuracy    Recall  Precision        F1
0   0.74488  0.805505   0.579621  0.674145
```

(a) Model Performance on train data




```
Model Performance (Significant Features - Test Set)
   Accuracy    Recall  Precision        F1
0  0.748231  0.810712   0.583333  0.678479
```

(b) Model Performance on test data

Figure 23: Model 1

- Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.
- When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

1. **Null hypothesis -** Contribution of the column i.e. the coefficient is zero.
2. **Alternate hypothesis -** Contribution of the column i.e. the coefficient is not zero.

If p-value is > 0.05 then we accept our null hypothesis.

6.1.2 Determine optimal threshold using ROC curve

ROC AUC Curve

The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds.

Definitions

- **True Positive Rate (TPR)**: Sensitivity or recall.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **False Positive Rate (FPR)**:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

The AUC measures the area under this curve, indicating model performance; higher AUC is better.

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

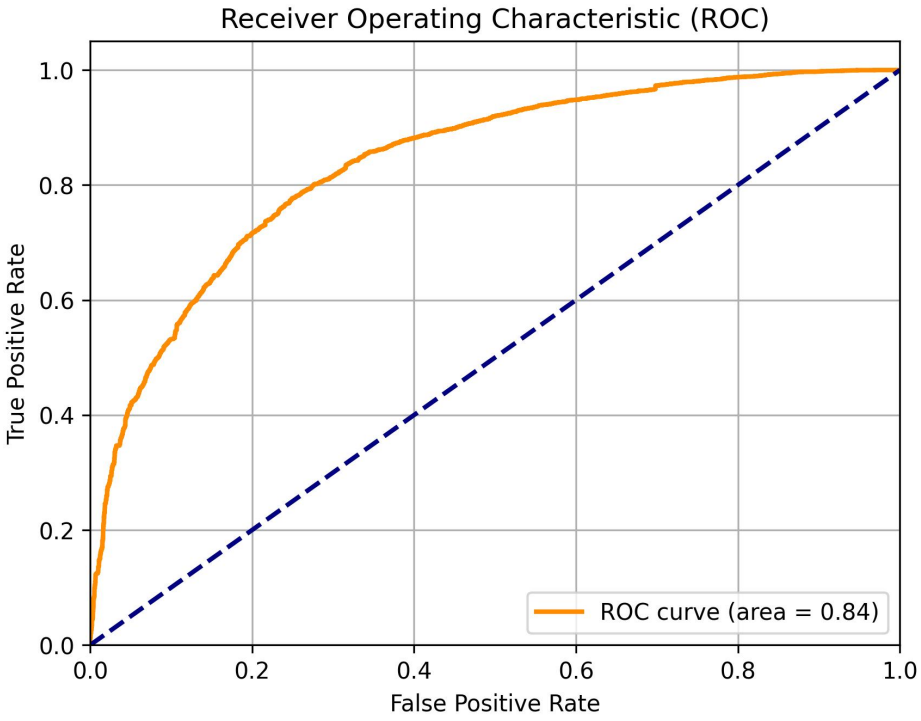


Figure 24: ROC-AUC curve

6.1.3 Model 2

Youden's J Statistic

The Youden's J statistic is a criterion used to identify the optimal threshold for a classification model. It maximizes the balance between sensitivity (True Positive Rate, TPR) and specificity (1 - False Positive Rate, FPR).

Definition

The Youden's J statistic is defined as:

$$J = \text{TPR} - \text{FPR}$$

where:

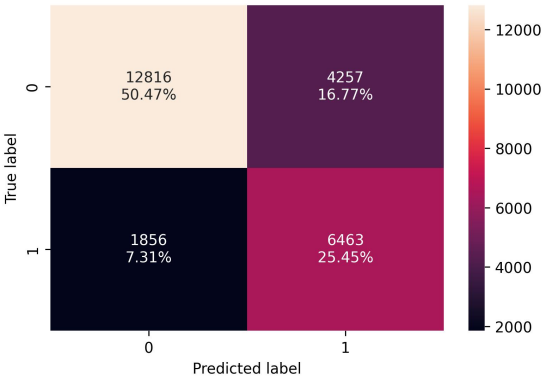
$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{and} \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

Alternatively, it can be expressed in terms of sensitivity and specificity as:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

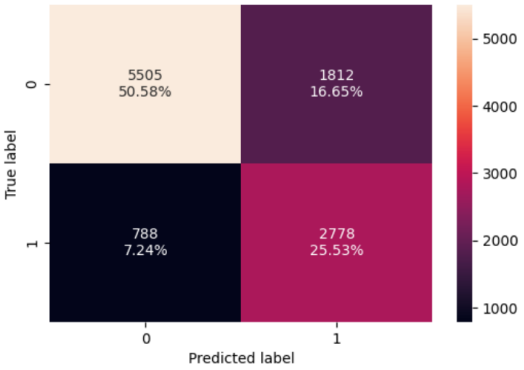
Optimal Threshold

The threshold that maximizes J is considered the optimal threshold, as it represents the best trade-off between sensitivity and specificity. The obtained threshold for this model is $J_{\max}$=0.535



	Accuracy	Recall	Precision	F1
0	0.759255	0.776896	0.602892	0.678922

(a) Model Performance on train data



	Accuracy	Recall	Precision	F1
0	0.761095	0.779024	0.605229	0.681216

(b) Model Performance on test data

Figure 25: Model 2

6.1.4 Finding a better threshold using the Precision-Recall Curve

This method is intuitive and visually interpretable when plotting both Precision and Recall against the thresholds. The intersection point provides a straightforward threshold for balanced performance. Threshold value at the meeting point is 0.6254

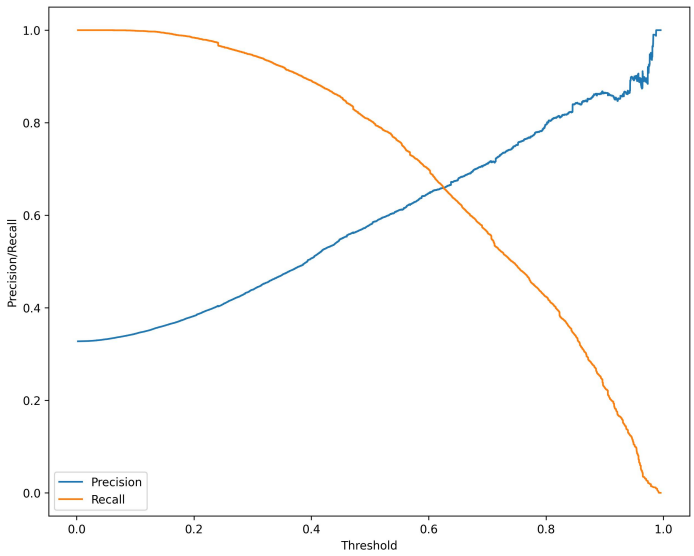
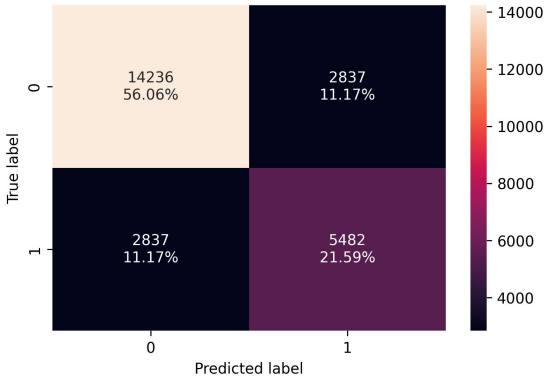
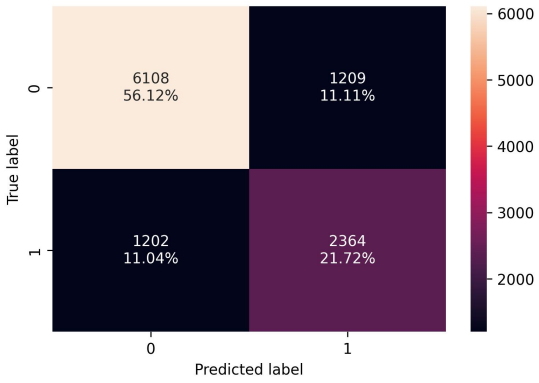


Figure 26: Precision-Recall Curve



	Accuracy	Recall	Precision	F1
0	0.776544	0.658973	0.658973	0.658973

(a) Model Performance on train data



	Accuracy	Recall	Precision	F1
0	0.778462	0.662928	0.661629	0.662278

(b) Model Performance on test data

Figure 27: Model 3

6.1.5 Model 3

Model 3 is built using that threshold of 0.6254.

6.1.6 Final model and performance evaluation

	Logistic Regression-default Threshold	Logistic Regression-0.5351470185002022 Threshold	Logistic Regression-0.6253739096335333 Threshold
Accuracy	0.744880	0.761095	0.776544
Recall	0.805505	0.779024	0.658973
Precision	0.579621	0.605229	0.658973
F1	0.674145	0.681216	0.658973

Figure 28: Training performance comparison

	Logistic Regression-default Threshold	Logistic Regression-0.5351470185002022 Threshold	Logistic Regression-0.6253739096335333 Threshold
Accuracy	0.744880	0.761095	0.776544
Recall	0.805505	0.779024	0.658973
Precision	0.579621	0.605229	0.658973
F1	0.674145	0.681216	0.658973

Figure 29: Testing performance comparison

All models are not overfitting nor are underfitting with both the training data and test data. As avoiding unnecessary actions is important (e.g., sending false cancellation alerts) we focus on reducing False Positives (FP) and increasing Precision. So we construct the model with high precision alongwith high accuracy. Hence we choose Logistic Regression model with threhold being 0.625

Coefficient interpretations

index	Odds	Percentage Change Odds
no_of_adults	1.0525725116114242	5.25725116142418
no_of_children	1.0864234320250807	8.642343202508073
no_of_weekend_nights	1.1330653895913667	13.30653895913667
no_of_week_nights	1.0793384259834875	7.93384259834875
required_car_parking_space	0.8633930048954181	-13.660699510458185
lead_time	3.166356185690746	216.63561856907458
arrival_year	1.1621614578170172	16.21614578170172
arrival_month	0.9236170329368973	-7.638296706310266
avg_price_per_room	1.85824018674228	85.824018674228
no_of_special_requests	0.4963730924017 4526	-50.36269075982547
type_of_meal_plan_Meal Plan 2	0.9114130324575037	-8.858696754249628
type_of_meal_plan_Not Selected	1.2313594814355355	23.135948143553552
room_type_reserved_Room_Type 5	0.9476626914069619	-5.233730859303809
room_type_reserved_Room_Type 6	0.8935067002746477	-10.649329972535227
room_type_reserved_Room_Type 7	0.9289165669940076	-7.10834330059924
market_segment_type_Complementary	1.1560718701734183	15.607187017341829

Figure 30: The coefficients of the logistic regression model and their corresponding percentage change in odds

- Coefficient of some of the variables are positive. An increase in these will lead to increase in chances of canceled bookings.
- The coefficients of the logistic regression model are in terms of log(odd), to find the odds we have to take the exponential of the coefficients.
- Therefore, $odds = \exp(b)$ The percentage change in odds is given as $odds = (exp(b) - 1) \times 100$
- In logistic regression, the coefficients represent the change in the log-odds of the outcome (in this case, booking cancellation) for a one-unit change in the predictor variable, holding all other variables constant.

Guest Demographics and Booking Patterns

Guest Demographics

- **Number of Adults:** An increase in the number of adults increases the odds of cancelling a booking by approximately 5.26%.
- **Number of Children:** Bookings with children are more likely to be canceled. The odds increase by about 8.64% with each additional child.
- **Weekend Stays:** Bookings that include weekend nights have a higher chance of cancellation. The odds of cancellation increase by around 13.31% for each additional weekend night.
- **Weekday Stays:** While not as impactful as weekends, longer weekday stays also slightly increase the odds of cancellation by approximately 7.93%.

Booking Characteristics

- **Car Parking:** Bookings where guests request car parking spaces are significantly less likely to be canceled. The odds of cancellation decrease by about 80%.
- **Lead Time:** Longer lead times dramatically increase the odds of a booking being canceled, with a 216.64% percentage change in odds.
- **Arrival Year:** A later arrival year increases the likelihood of cancellations by 16.22%.

- **Arrival Month:** As the arrival month gets later in the year, the likelihood of cancellation decreases, with a 4.2% decrease in odds.
- **Average Room Price:** Higher average room prices are associated with increased odds of cancellation, approximately 85.82%. This could suggest price sensitivity among some guests.
- **Special Requests:** Bookings with special requests are much less likely to be canceled. Each special request decreases the odds of cancellation by approximately 77%, indicating that guests with specific needs are more committed to their reservations.

Meal Plans

- **Meal Plan 2 (Half Board):** Selecting Meal Plan 2 decreases the odds of cancellation by roughly 8.86%.
- **No Meal Plan:** Not selecting any meal plan increases the odds of cancellation, with a percentage change in odds of about 23.14%.

Room Types

- **Room Type 5:** Choosing Room Type 5 slightly decreases the likelihood of cancellation by about 5.23%.
- **Room Type 6:** Opting for Room Type 6 reduces the chances of cancellation by roughly 10.65%.
- **Room Type 7:** Room Type 7 shows a similar pattern with a slight decrease in odds of cancellation by about 7.11%.

Market Segment

- **Complementary:** Bookings from the 'Complementary' market segment increase the likelihood of cancellation by about 15.61%.

6.2 Naive - Bayes Classifier

Naive Bayes Classifier

The **Naive Bayes Classifier** is a probabilistic machine learning algorithm based on **Bayes' Theorem**. It assumes that features are conditionally independent given the class label. Despite its simplicity, it performs well in tasks such as spam detection and sentiment analysis.

Bayes' Theorem

The algorithm relies on Bayes' theorem, which is expressed as:

$$P(C_k|X) = \frac{P(X|C_k) \cdot P(C_k)}{P(X)},$$

where:

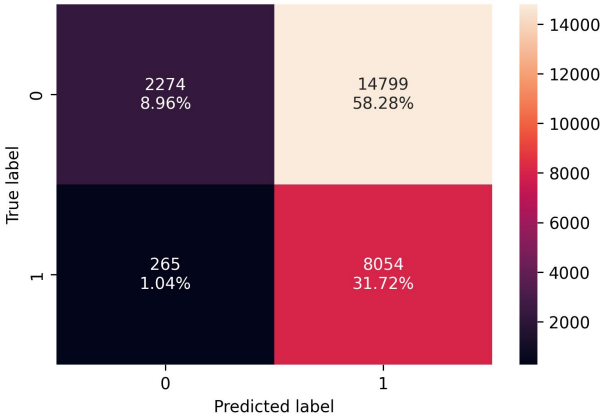
- $P(C_k|X)$: Posterior probability of class C_k given the feature set X,
- $P(X|C_k)$: Likelihood of features given class C_k,
- $P(C_k)$: Prior probability of class C_k,
- $P(X)$: Marginal probability of the features.

Assumption of Independence

The conditional independence assumption simplifies the computation of $P(X|C_k)$:

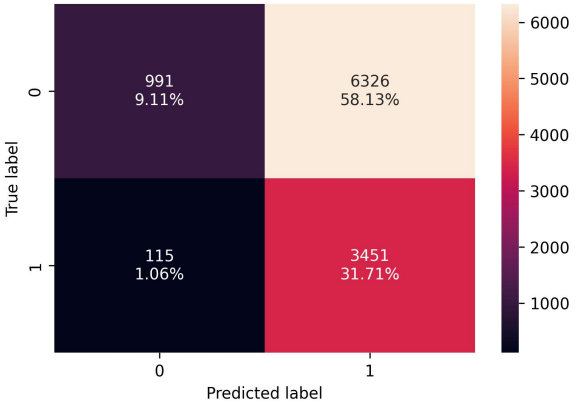
$$P(X|C_k) = \prod_{i=1}^{n} P(x_i|C_k),$$

where x_i are the individual features.



	Accuracy	Recall	Precision	F1
0	0.406742	0.968145	0.352426	0.516746

(a) Model Performance on train data



	Accuracy	Recall	Precision	F1
0	0.40816	0.967751	0.352971	0.517275

(b) Model Performance on test data

Figure 31: Naive Bayes Model

Algorithm Steps

1. Calculate prior probabilities $P(C_k)$ for each class C_k.
2. Compute the likelihood $P(x_i|C_k)$ for all features x_i.
3. Use Bayes' theorem to compute the posterior probability for each class:

$$P(C_k|X) \propto P(C_k) \cdot \prod_{i=1}^{n} P(x_i|C_k).$$

4. Assign the class with the highest posterior probability:

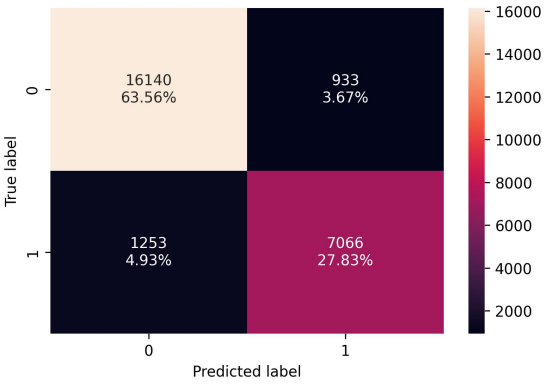
$$\hat{C} = \arg\max_{C_k} P(C_k|X).$$

6.2.1 Checking Naive - Bayes Classifier performance on training and testing set

This model has great recall value but the precision is very less. When minimizing missed cancellations is crucial for resource optimization (such as in reallocating hotel rooms or managing staff), selecting a model that effectively reduces False Negatives (FN) and maximizes Recall becomes paramount. In such scenarios, the Naive Bayes classifier could be a viable option due to its ability to prioritize the identification of potential cancellations, even at the risk of some false positives. This characteristic makes it well-suited for applications where the cost of a missed cancellation outweighs the cost of a false alarm.

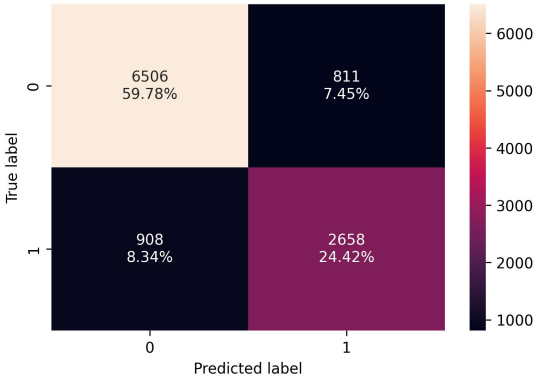
6.3 KNN Classifier (K = 3)

The **KNN Classifier** assigns a class label based on the majority class among the K nearest neighbors. For $K = 3$, the classification steps are:



	Accuracy	Recall	Precision	F1
0	0.91391	0.849381	0.88336	0.866038

(a) Model Performance on train data



	Accuracy	Recall	Precision	F1
0	0.842047	0.745373	0.766215	0.75565

(b) Model Performance on test data

Figure 32: KNN Classifier Model

1. Compute the distance (e.g., Euclidean) between the test point and all training points.
2. Identify the 3 nearest neighbors.
3. Assign the test point the class most frequent among these neighbors.

KNN is simple, non-parametric, and effective for small datasets but computationally expensive for large ones. The performance of the model is shown below.

6.3.1 Checking KNN Classifier performance on training and testing set

KNN classifier performance is improved using different k values. Now we try to do it in two different ways.

- **Method-1(To maximize recall)**

The best k-value turned out to be 3 so the model remains same for this.

- **Method-2(To maximize f1-score)**

The best k-value turned out to be 3 so the model remains same for this too.

6.4 Decision Tree Classifier

Decision Tree Classifier

The **Decision Tree Classifier** uses a tree structure to classify data by splitting it into subsets based on feature values. At each node, a feature is selected to split the data, optimizing a metric such as:

- **Gini Impurity:**

$$G = 1 - \sum_{i=1}^{n} p_i^2,$$

where p_i is the proportion of samples of class i at a node.

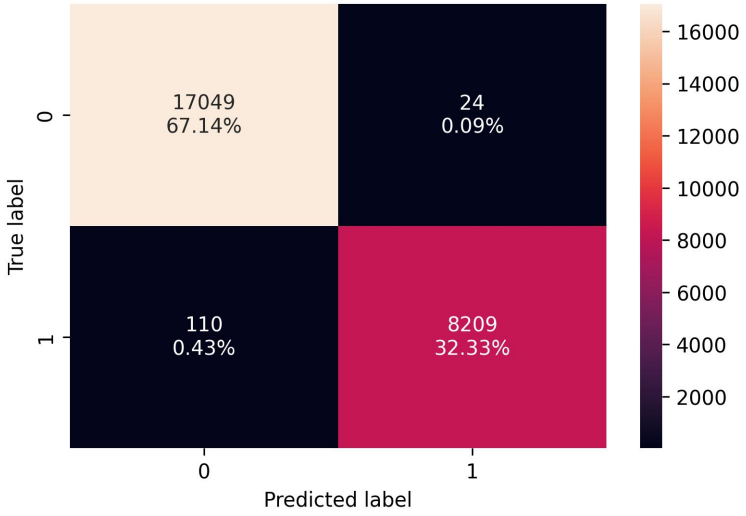
- **Entropy (Information Gain):**

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

Steps:

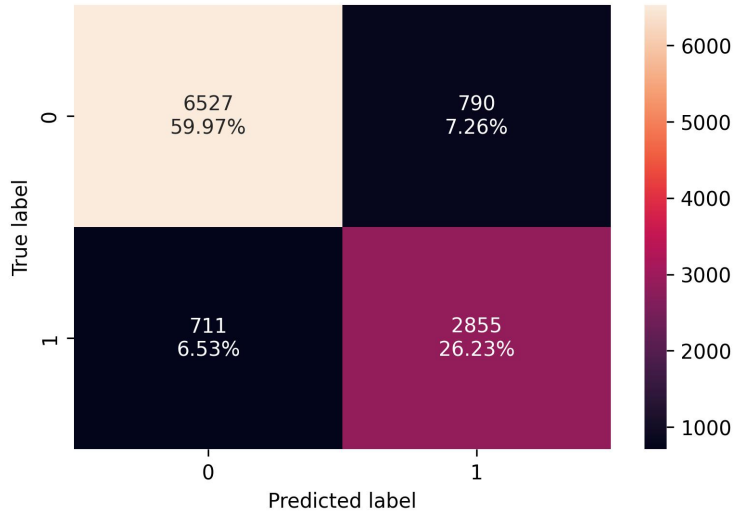
1. Select the best feature to split the data using a splitting criterion.
2. Recursively split the subsets until reaching leaf nodes.
3. Assign class labels to leaf nodes based on the majority class or probabilities.

Decision Trees are easy to interpret but may overfit without proper pruning.



	Accuracy	Recall	Precision	F1
0	0.994723	0.986777	0.997085	0.991904

Figure 33: Model Performance on train data



	Accuracy	Recall	Precision	F1
0	0.860976	0.793887	0.784428	0.789129

Figure 34: Model Performance on test data

Model hasn't performed very well on test dataset. Recall of both training data and testing data is not similar as expected because of overfitting. We have to prune the decision tree.

6.4.1 Decision Tree Classifier (pre-pruning and post-pruning)

- **Pre-pruning**

Pre-Pruning in Decision Trees

Pre-pruning stops the growth of a decision tree early by applying constraints during the tree-building process. This prevents overfitting and improves generalization.

Constraints Used in Pre-Pruning:

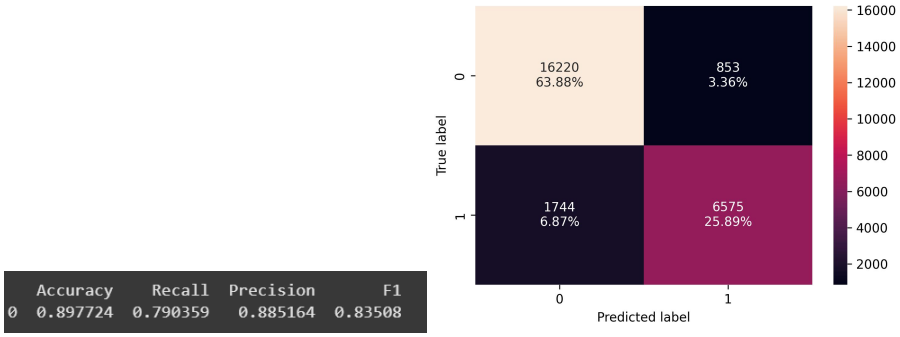
- Maximum depth of the tree.
- Minimum number of samples required to split a node.
- Minimum number of samples required at a leaf node.

Steps:

1. Set constraints, such as maximum depth or minimum samples per split.
2. Grow the tree but halt splits if any constraint is violated.
3. Finalize the tree with fewer nodes, reducing overfitting.

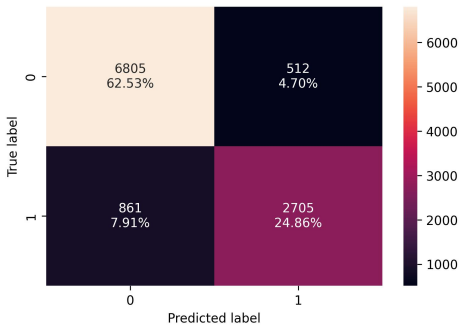
Pre-pruning simplifies the model, making it faster and less prone to noise in the data. The best model after preprunning comes out to be a tree with parameters as

$$\text{max\_depth} = 16, \quad \text{max\_leaf\_nodes} = 250, \quad \text{min\_impurity\_decrease} = 1 \times 10^{-6}$$



	Accuracy	Recall	Precision	F1
0	0.897724	0.790359	0.885164	0.83508

(a) Model Performance on train data



	Accuracy	Recall	Precision	F1
0	0.87384	0.758553	0.840846	0.797582

(b) Model Performance on test data

Figure 35: Model Pre-prunning

The visualization of the tree and important features of the tree are as follows:

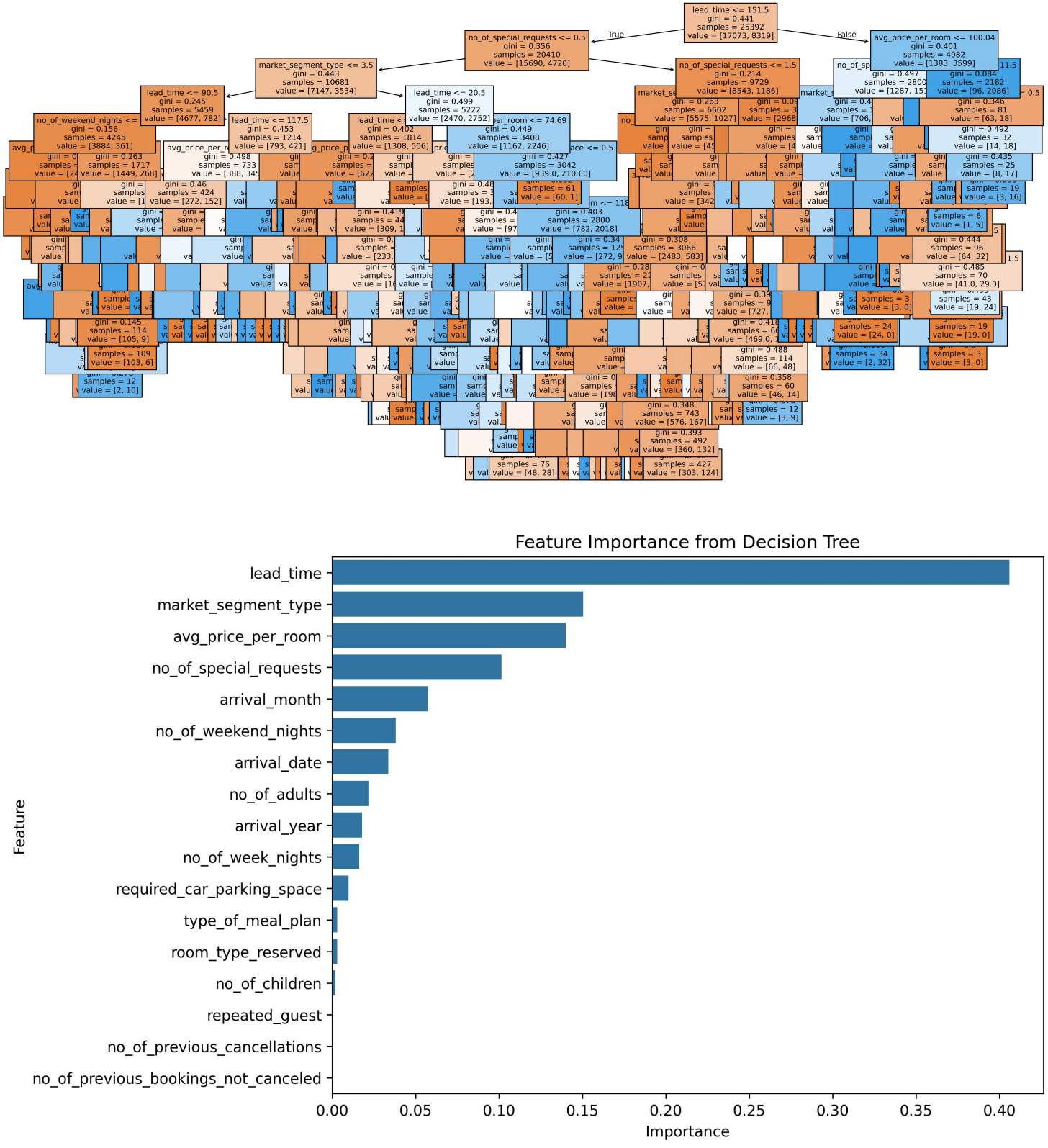


Figure 36: Pre-prunned Tree and its importance features

- **Post-prunning**

Post-Pruning in Decision Trees

Post-pruning simplifies a fully grown decision tree by removing branches that provide little predictive value. It evaluates subtrees and prunes those that do not significantly improve performance, reducing overfitting.

Steps:

1. Grow the decision tree to its full depth.
2. Use a validation set or a pruning criterion to evaluate subtrees.
3. Prune branches that do not significantly improve accuracy.
4. Finalize the pruned tree for better generalization.

Cost Complexity Pruning

A common post-pruning method, **Cost Complexity Pruning**, balances model complexity with prediction accuracy by minimizing:

$$R_\alpha(T) = R(T) + \alpha \cdot |T|,$$

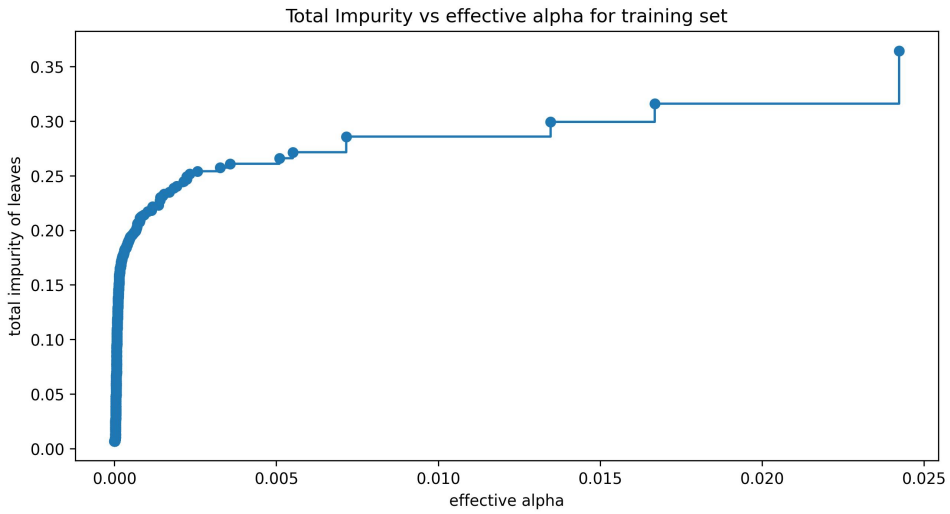
where:

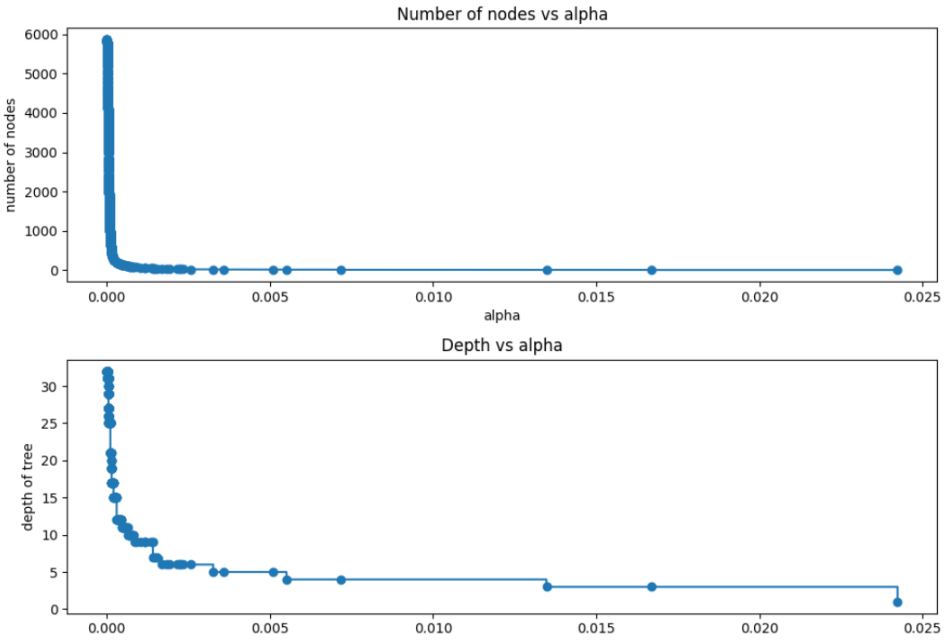
- $R(T)$: Misclassification error of the tree T,
- $|T|$: Number of terminal nodes in T,
- α: Complexity parameter controlling the trade-off between tree size and accuracy.

Steps for Cost Complexity Pruning:

1. Compute $R_\alpha(T)$ for different subtrees.
2. Select the subtree with the smallest $R_\alpha(T)$.
3. Repeat for various values of α to find the optimal pruned tree.

Cost complexity pruning ensures a balance between underfitting and overfitting. With different alphas I have got the impurities of the leaves. Different plots are plotted by changing alpha.





	ccp_alphas	impurities
0	0.000000e+00	0.006842
1	0.000000e+00	0.006842
2	6.352013e-07	0.006843
3	8.078458e-07	0.006843
4	1.125214e-06	0.006844
...	...	...
1308	7.158311e-03	0.285860
1309	1.347162e-02	0.299331
1310	1.668747e-02	0.316019
1311	2.423451e-02	0.364488
1312	7.608448e-02	0.440572

1313 rows × 2 columns

Figure 37: Relation of alpha with other parameters.

Here we show that the number of nodes and tree depth decreases as alpha increases. As alpha increases depth of the tree and number of notes decreases.

6.4.2 Recall vs alpha for training and test sets

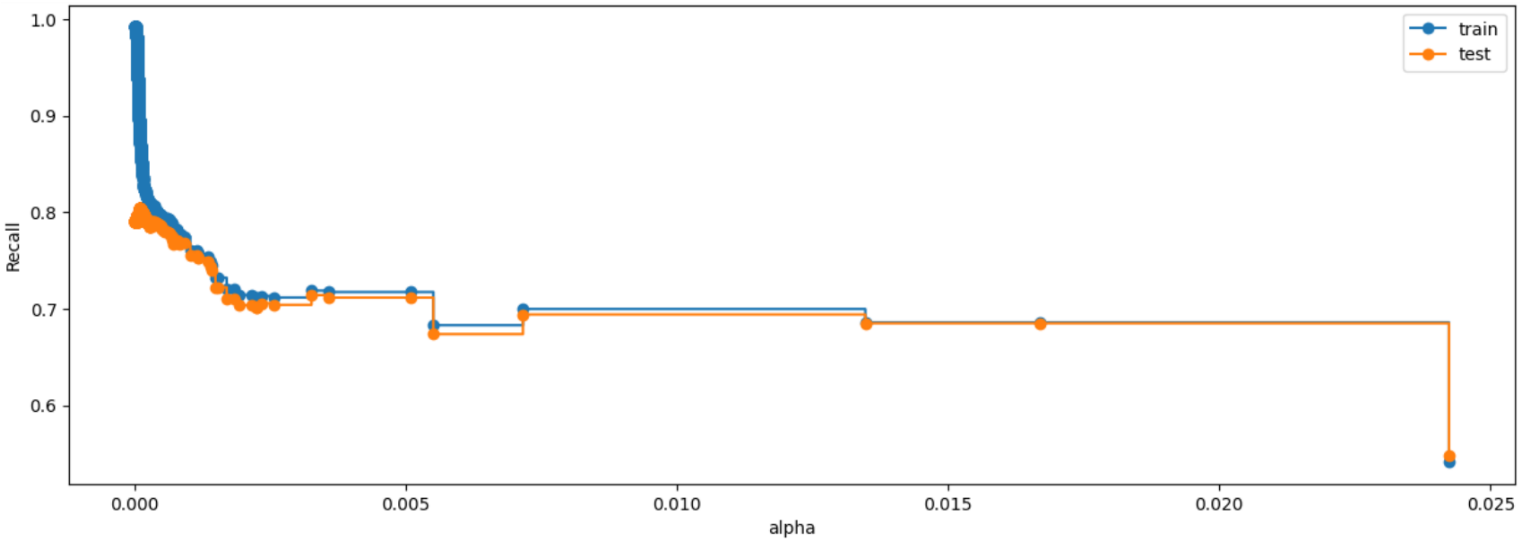
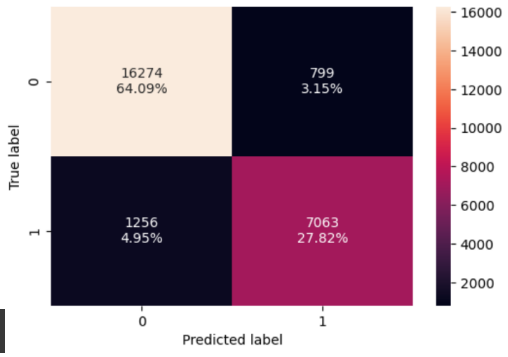


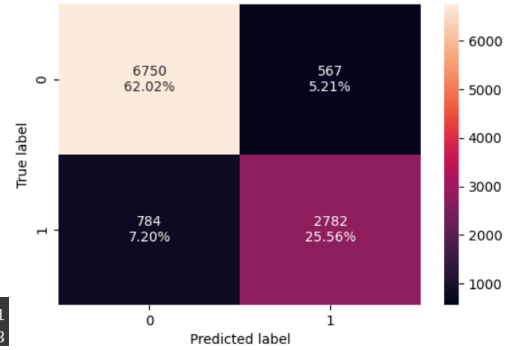
Figure 38: Recall vs alpha for training and test sets

From this plot the best alpha is found out to be 9.76e-05 which is obtained by finding the f1-score$_{\max}$.Now let's check the model performance on training and testing data.



	Accuracy	Recall	Precision	F1
0	0.919069	0.84902	0.898372	0.872999

(a) Model Performance on train data

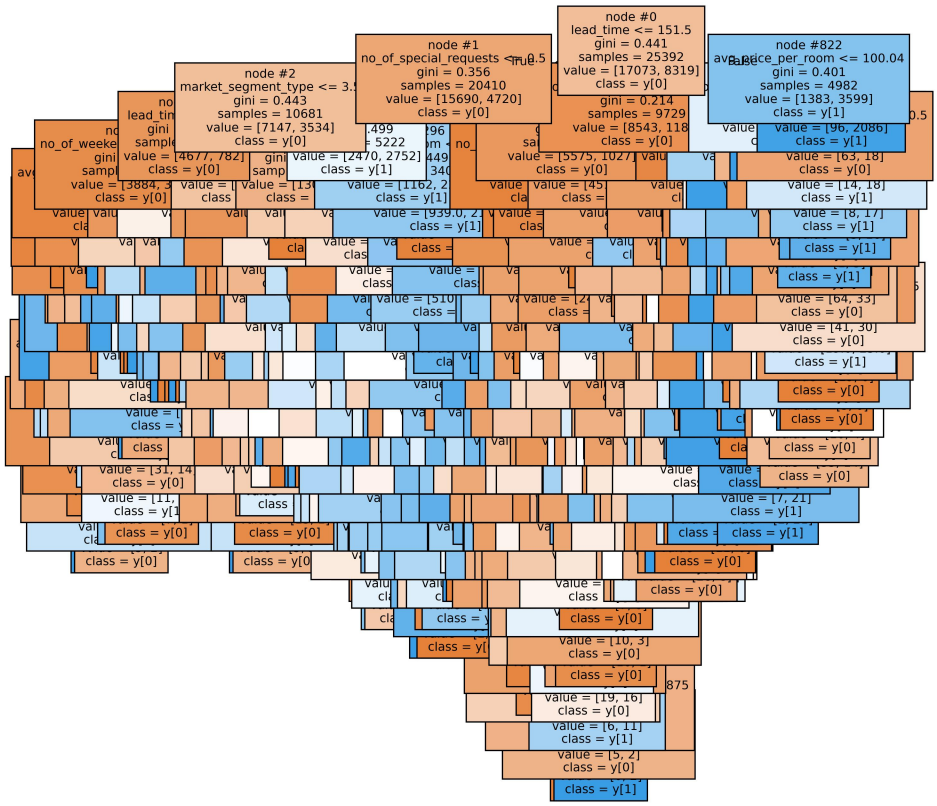


	Accuracy	Recall	Precision	F1
0	0.875861	0.780146	0.830696	0.804628

(b) Model Performance on test data

Figure 39: Model Post-prunning

The visualization of the tree and important features of the tree are as follows:



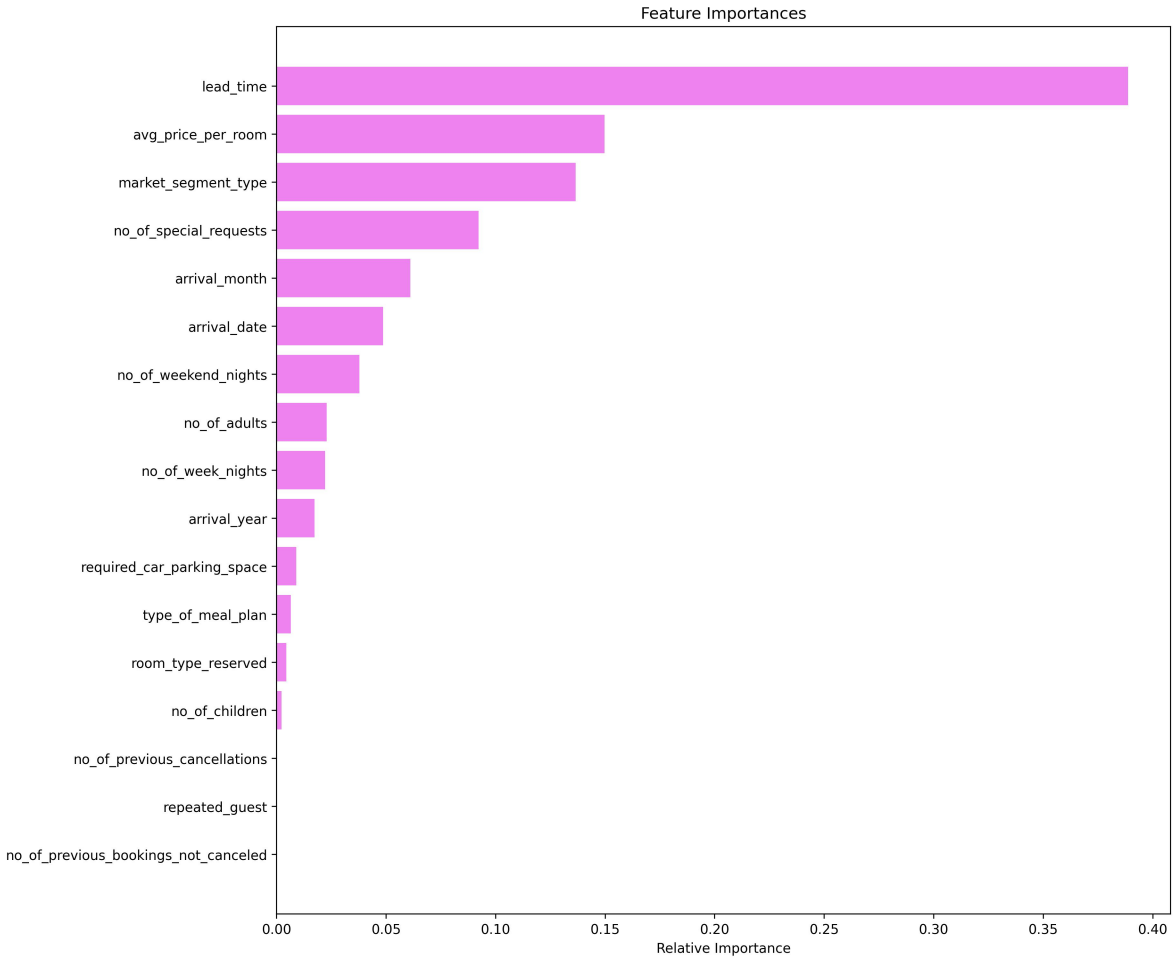


Figure 40: Post-prunned tree and its importance features

6.5 Comparison of Models and Final Model Selection

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.994723	0.897724	0.919069
Recall	0.986777	0.790359	0.849020
Precision	0.997085	0.885164	0.898372
F1	0.991904	0.835080	0.872999

Test set performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.864284	0.873840	0.875861
Recall	0.801739	0.758553	0.780146
Precision	0.787820	0.840846	0.830696
F1	0.794719	0.797582	0.804628

Figure 41: Training and testing performance comparison

Let's review the F1 scores for each model. Out of all the models we have constructed the **Decision Tree Post-Pruning** model has the highest test F1 score (0.804628), making it the best model for maximizing F1 score.

7 Actionable Insights & Recommendations

1. Reduce Cancellations

- **Incentivize Bookings**: Offer raffles or discounts for customers with zero cancellations.
- **Flexible Policies**: Provide clear, flexible cancellation terms.

2. Improve Special Request Handling

- **Streamline Requests**: Simplify handling of special requests to enhance satisfaction.
- **Promote Requests**: Encourage guests to make requests during booking for better preparation.

3. Pricing Strategies

- **Dynamic Pricing**: Lower prices during high-demand months to attract more customers.
- **Early Bird Discounts**: Offer discounts for early bookings.
- **Off-Peak Offers**: Create promotions for off-peak months to boost bookings.

4. Loyalty Program for Repeated Guests

- **Reward Programs**: Offer benefits like reward points or complimentary upgrades for repeat guests.
- **Personalized Experience**: Collect guest preferences for a tailored experience.

5. Optimize Family Services

- **Kid-Friendly Rooms**: Introduce family-oriented packages and kid's rooms.
- **Promote Family Offers**: Offer discounts and services for families with children.

6. Parking Optimization

- **Efficient Allocation**: Allocate limited parking for guests who need it.
- **Clear Communication**: Highlight parking availability during booking.

7. Encourage Early Bookings

- **Early Discounts**: Offer promotions for bookings made well in advance.
- **Investigate Outliers**: Analyze high lead times to improve booking strategies.

8. Dynamic Room Pricing

- **Target Mid-Range**: Adjust pricing for mid-range customers, and market premium rooms for high-paying guests.
- **Special Packages**: Offer personalized packages for customers with special requests.

9. Targeted Promotions & Retention

- **Focus on New Guests**: Offer promotions to non-repeat guests and address why they don't return.
- **Customer Insights**: Use data to understand trends and target high-value customers.