Project 4

- Machine Learning -

(Predicting Hotel Booking Cancellations)

1.Introduction

1.1.Problem Definition

The hospitality industry faces an ever-evolving landscape with changing customer preferences and technological advancements. Online booking platforms have made cancellations more frequent, leading to operational and revenue challenges for hotels. Cancellations and no-shows not only disrupt revenue streams but also affect resource management and profit margins, particularly with last-minute cancellations that leave little opportunity for reselling rooms.

Hotels like INN Hotels Group must optimize their operations to mitigate these impacts while maintaining customer satisfaction. Effective cancellation management strategies rely heavily on understanding customer behavior and predicting cancellations, which can enable proactive actions to minimize losses.

Business Problem

High booking cancellations are causing INN Hotels Group significant challenges, including:

- **Revenue Loss**: Rooms left unsold due to cancellations result in direct revenue losses.
- **Operational Inefficiency**: Last-minute cancellations demand increased human resources and distribution efforts to resell rooms.
- **Lower Profit Margins**: Discounts or reduced pricing to resell canceled rooms cut into profits.
- **Additional Costs**: Increased spending on promotions and third-party commissions to resell rooms exacerbates the issue.

These challenges hinder the hotel’s ability to maintain profitability and compete effectively in the market.

Objective

To address the high number of booking cancellations, a machine learning-based solution is required to predict cancellations in advance and identify factors influencing this behavior. Specifically, the objectives are:

- **Analysis**: Understand the key factors driving booking cancellations through exploratory data analysis (EDA).
- **Prediction**: Develop a predictive model to forecast which bookings are likely to be canceled.
- **Policy Recommendations**: Provide actionable insights to the INN Hotels Group to reduce cancellations and optimize resource allocation.

1.2.Data Background and Contents

Dataset Overview

The dataset consists of **36,275 rows and 19 columns**, representing various attributes related to hotel bookings. Each row corresponds to a single booking, and the columns capture details such as customer demographics, booking specifics, and cancellations. The dataset has no missing values and comprises a mix of numerical and categorical variables.

Key points:

- **Data Shape**: (36275, 19)
- **Types of Variables**:
  - Numerical: 14 columns
  - Categorical: 5 columns
- **Column Overview**:
  - Unique identifier (Booking_ID): Dropped due to all unique values.

1.2.2.Statistical Summary

A detailed statistical summary of the numerical columns reveals insights into the distribution of values:

- **Numerical Columns**:
  - **no_of_adults**: Average = 1.84, Min = 0, Max = 4.
  - **no_of_children**: Average = 0.11, Min = 0, Max = 10.
  - **lead_time**: Average = 85.23 days, Max = 443 days.
  - **avg_price_per_room**: Average = 103.42 EUR, Min = 0, Max = 540 EUR.
  - **no_of_previous_cancellations**: Most bookings have zero cancellations, but the maximum observed is 13.
- **Categorical Columns**:
  - **type_of_meal_plan**: 4 unique levels, with "Meal Plan 1" being the most common (77%).
  - **market_segment_type**: 5 unique levels, with "Online" accounting for 64%.
  - **booking_status**: Binary variable with two values, "Not_Canceled" (67%) and "Canceled" (33%).

Target Variable

- no_of_previous_cancellations:
  - Indicates the number of cancellations made by the customer before the current booking.
  - Most values are zero, but there is a small proportion of customers with multiple cancellations.
  - This variable serves as the dependent variable for the bivariate and predictive analyses.

Independent Variables (Features)

The remaining columns serve as predictors or independent variables, capturing various aspects of bookings:

1. Demographics:
   - no_of_adults, no_of_children.
2. Booking Details:
   - no_of_weekend_nights, no_of_week_nights, type_of_meal_plan, required_car_parking_space, room_type_reserved.
3. Booking Behavior:
   - lead_time, market_segment_type, repeated_guest, no_of_special_requests.
4. Temporal Variables:
   - arrival_year, arrival_month, arrival_date.
5. Economic Variable:
   - avg_price_per_room.

1.3.Univariate Analysis :

1. Number of Adults in Bookings:

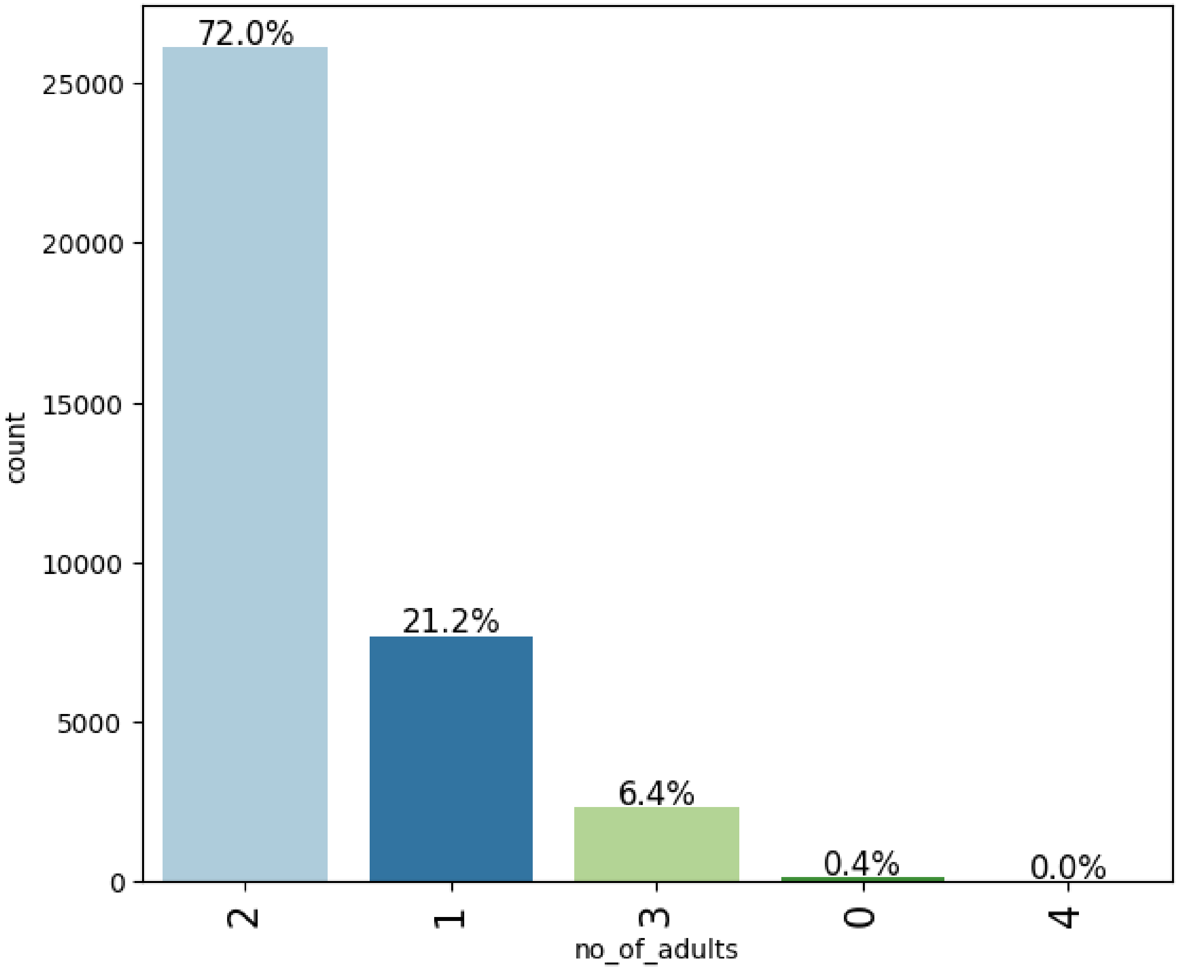

Figure 1. Number of Adults in Bookings

Interpretation:

Most bookings involve 2 adults, reflecting a dominant demographic of couples or pairs. Single or group bookings are less common, likely due to pricing structures, room capacities, or specific travel purposes like family vacations or solo business trips.

2. Number of Children in Bookings:

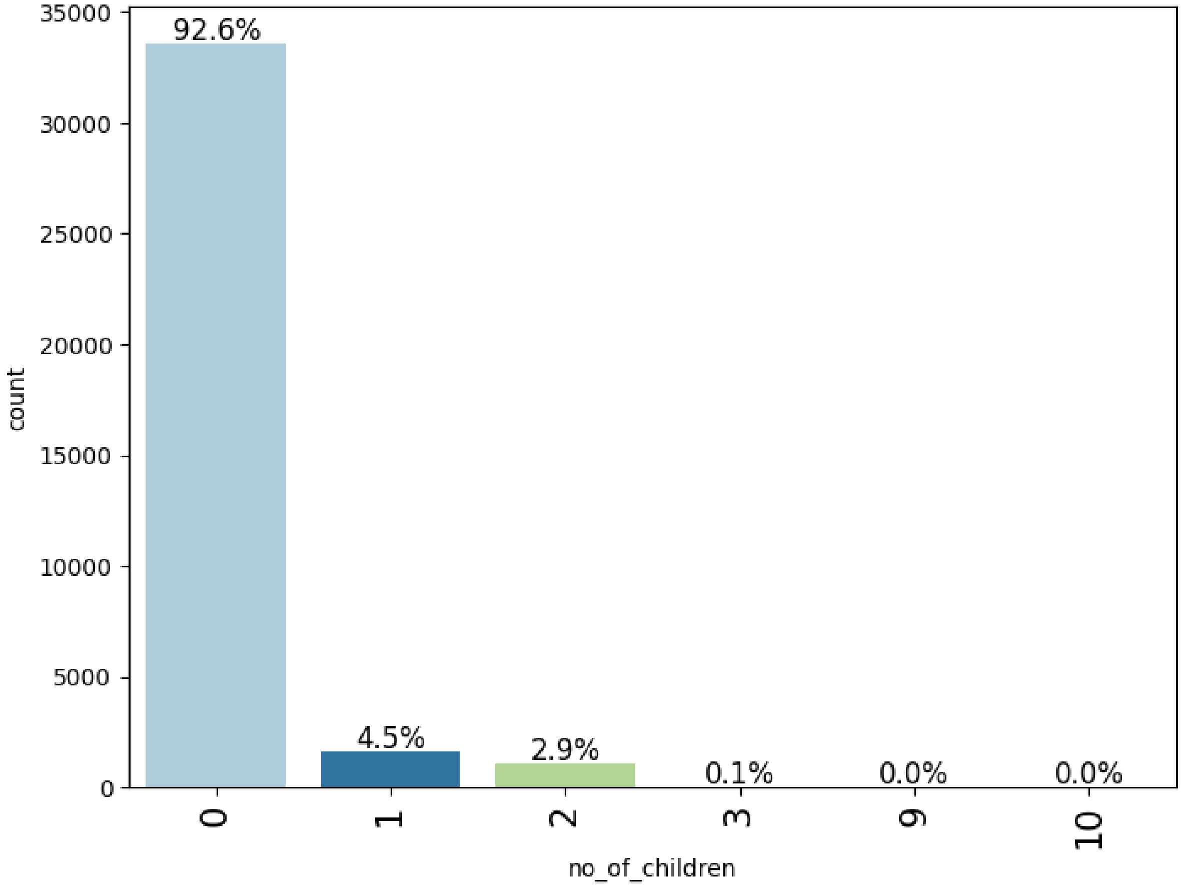

Figure 2. Number of Children in Bookings

Interpretation:

The majority of bookings are child-free, indicating families form a smaller segment of the customer base. Bookings involving 1 or 2 children are more frequent, but larger groups are rare, likely constrained by room capacities or pricing considerations.

3. Weekend Nights Included in Bookings:

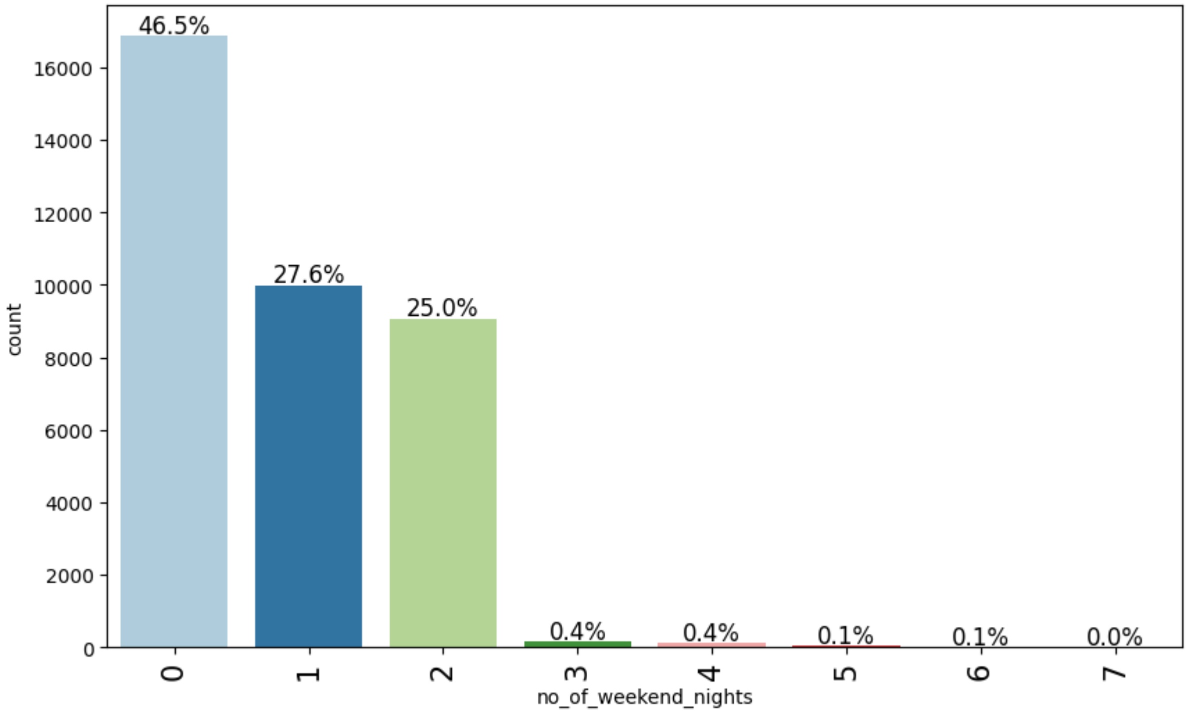

Figure 3. Weekend Nights Included in Bookings

Interpretation:

Most bookings include 0–2 weekend nights, suggesting shorter leisure stays. Extended stays over the weekend are uncommon, likely due to guest preferences for brief getaways or weekend-only travel.

4. Weekday Nights Included in Bookings:

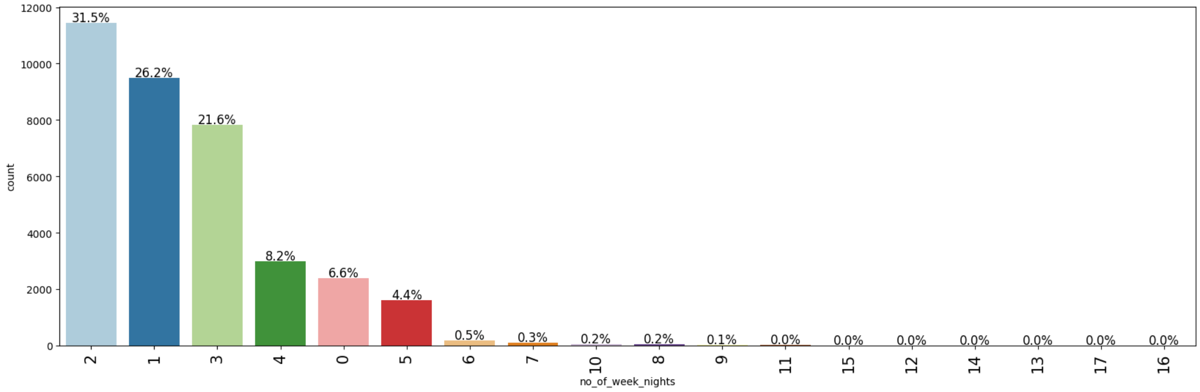

Figure 4. Weekday Nights Included in Bookings

Interpretation:

Stays of 1–3 weekday nights dominate, indicating frequent short visits, possibly for business or quick travel plans. Stays exceeding 5 nights are rare, likely linked to special events, conferences, or extended vacations.

5. Distribution of Meal Plan Preferences Among Guests:

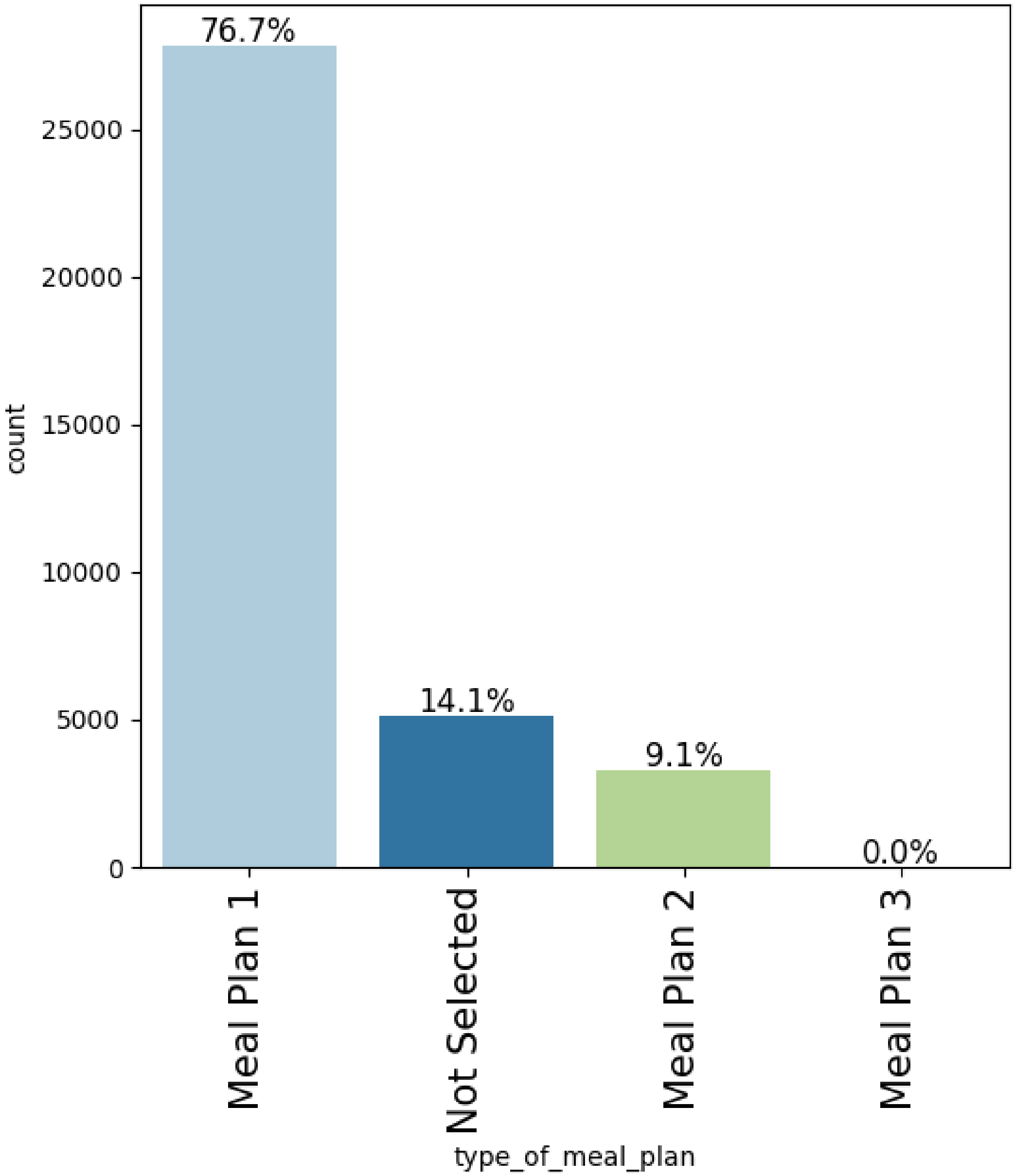

Figure 5. Distribution of Meal Plan Preferences Among Guests

Interpretation:

Meal Plan 1 is favored by 76.7% of guests, indicating it meets customer preferences effectively. A small percentage opt for Meal Plan 2, while very few choose Meal Plan 3. Additionally, 14.1% of guests select no meal plan, which may suggest they prefer dining off-site or have shorter stays.

6. Car Parking Space Requirement Across Bookings:

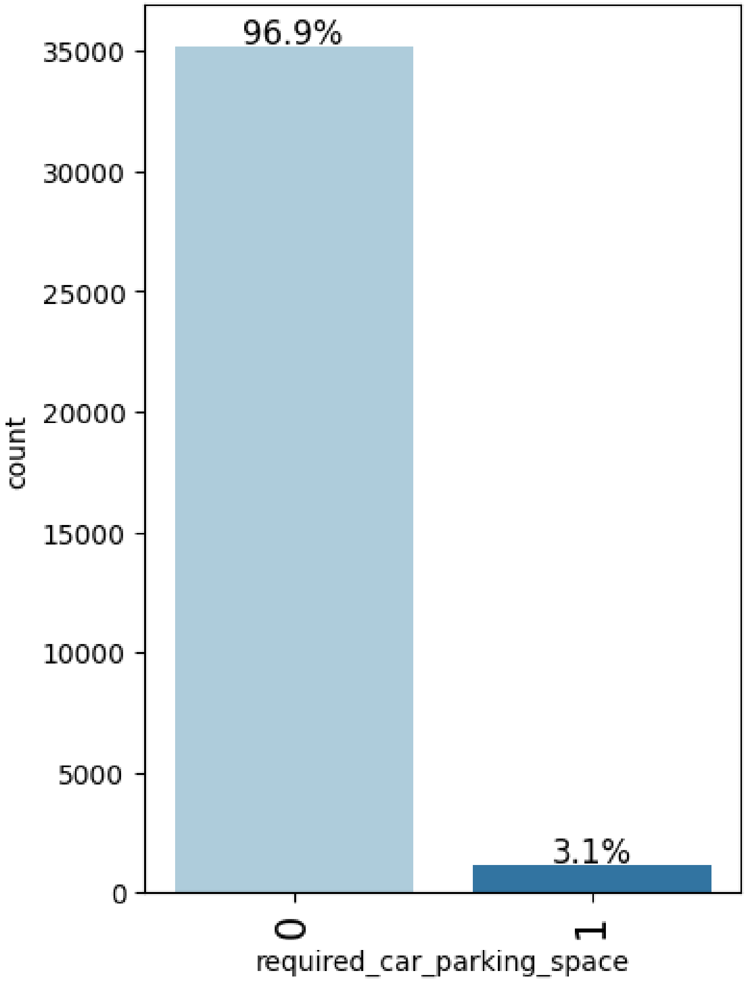

Figure 6: Car Parking Space Requirement Across Bookings

Interpretation:

Only 3.1% of bookings include car parking requests, suggesting most guests rely on public transportation, taxis, or other travel modes. This low demand for parking may reflect the urban nature of hotel locations or preferences of international tourists.

7. Popularity of Reserved Room Types:

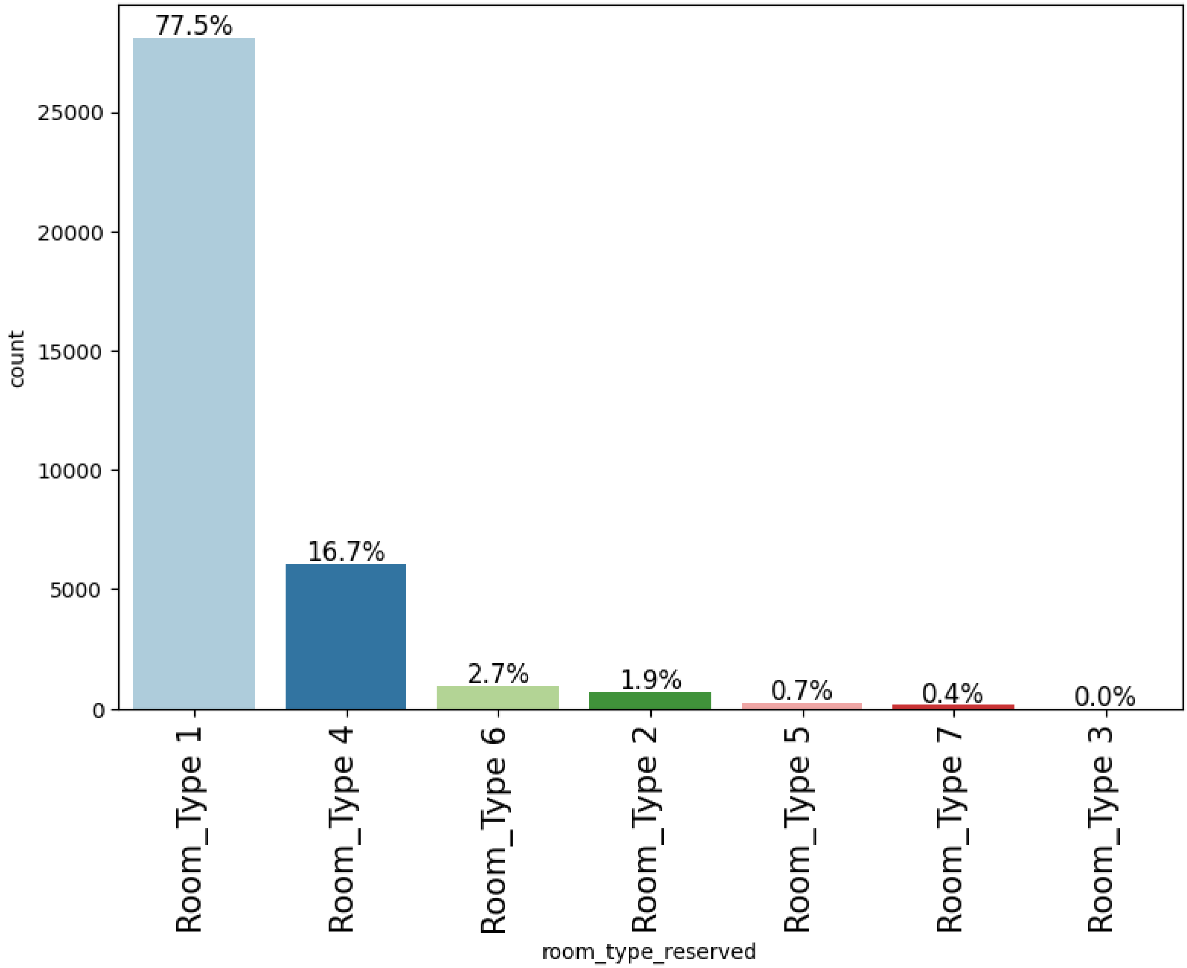

Figure 7. Popularity of Reserved Room Types

Interpretation:

Room_Type 1 accounts for 77.5% of reservations, implying it is the most suitable in terms of pricing, amenities, or availability. Other room types are less popular, potentially due to higher costs or lower inventory, impacting customer choices.

8. Lead Time Distribution for Hotel Bookings:

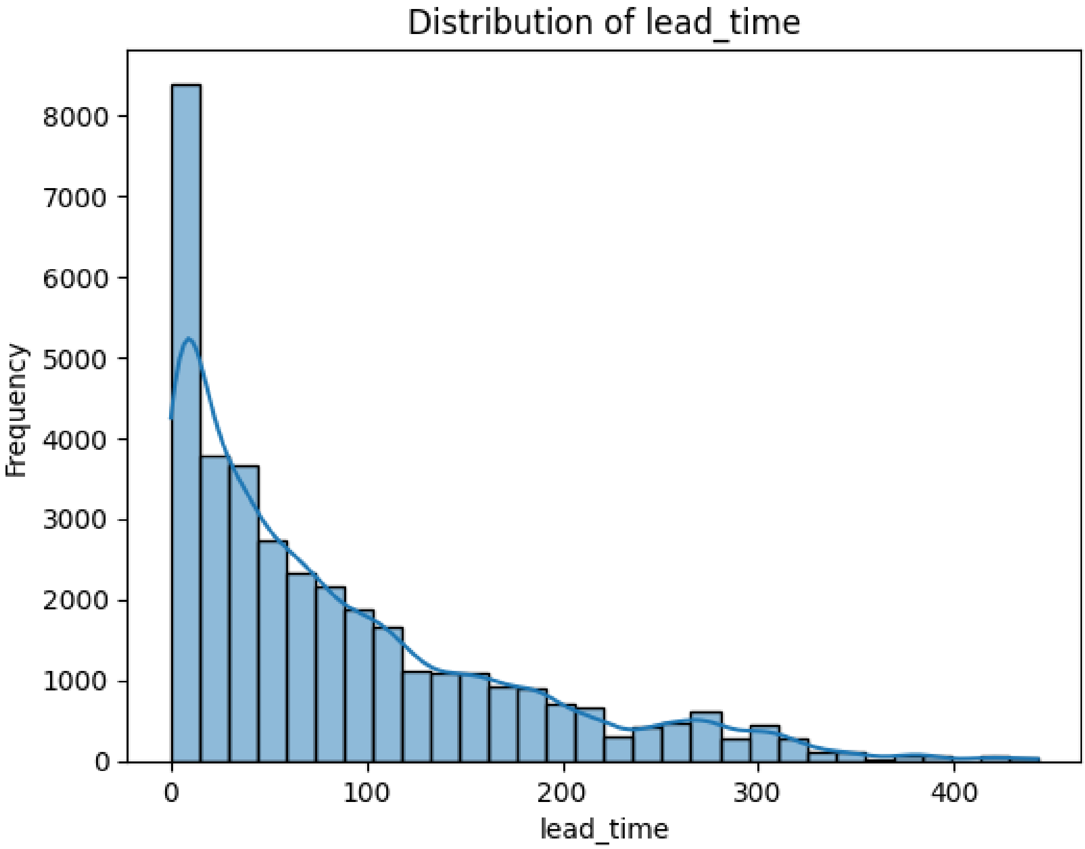

Figure 8. Lead Time Distribution for Hotel Bookings

Interpretation:

The distribution of lead time shows a right-skewed pattern, with most bookings made closer to the check-in date. A significant number of guests book within a short time frame (under 50 days), while fewer reservations are made with a lead time exceeding 200 days, indicating last-minute bookings dominate.

9. Yearly Distribution of Guest Arrivals:

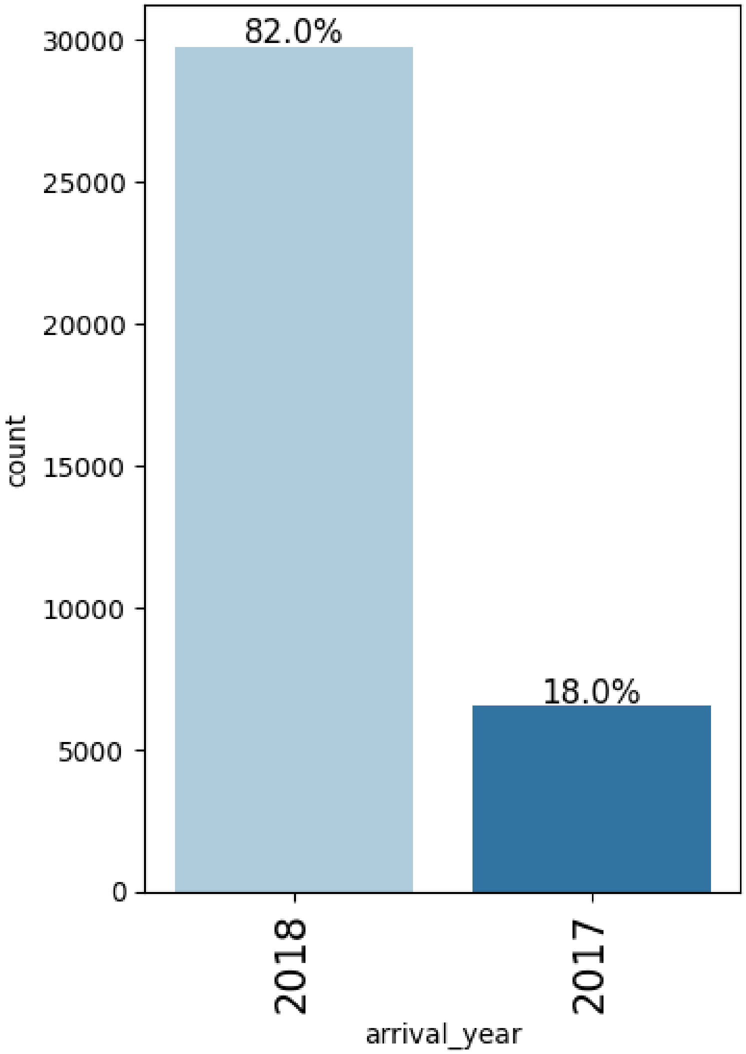

Figure 9. Yearly Distribution of Guest Arrivals

Interpretation:

The bar chart illustrates the distribution of arrivals across two years. The majority of arrivals, accounting for 82%, occurred in 2018. In contrast, only 18% of arrivals took place in 2017. This indicates a significant increase in arrivals from 2017 to 2018.

10. Seasonal Trends in Guest Arrivals (By Month):

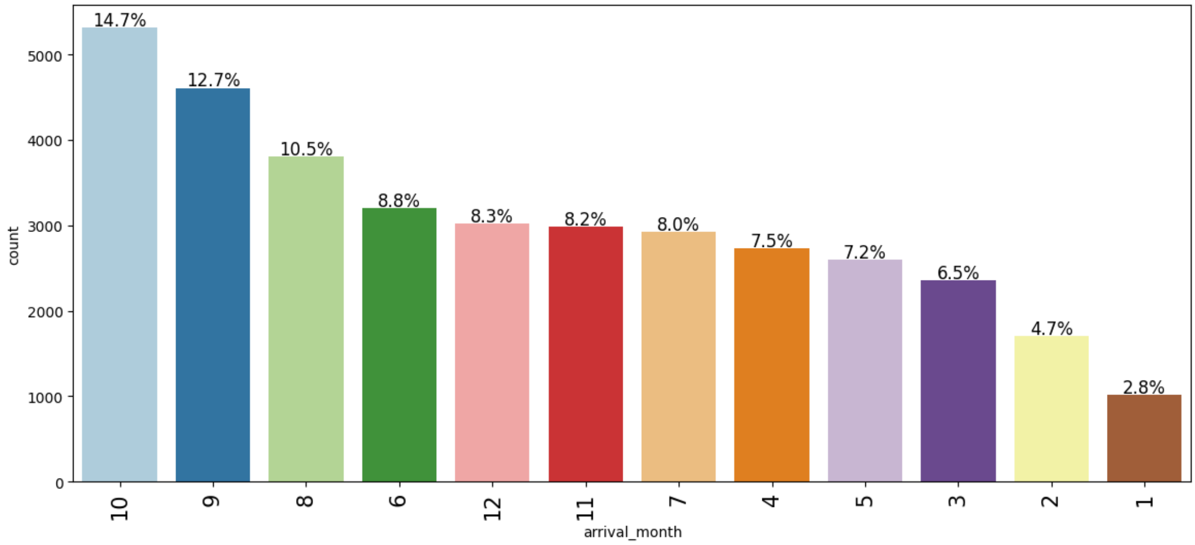

Figure 10. Seasonal Trends in Guest Arrivals (By Month)

Interpretation:

The bar chart shows the distribution of arrivals across different months. The month of October saw the highest number of arrivals, accounting for 14.7% of the total. This is followed by September with 12.7% and August with 10.5%. The number of arrivals gradually decreases from October to January, with January having the lowest number of arrivals at 2.8%.

11. Distribution of Arrival Dates Within Each Month:

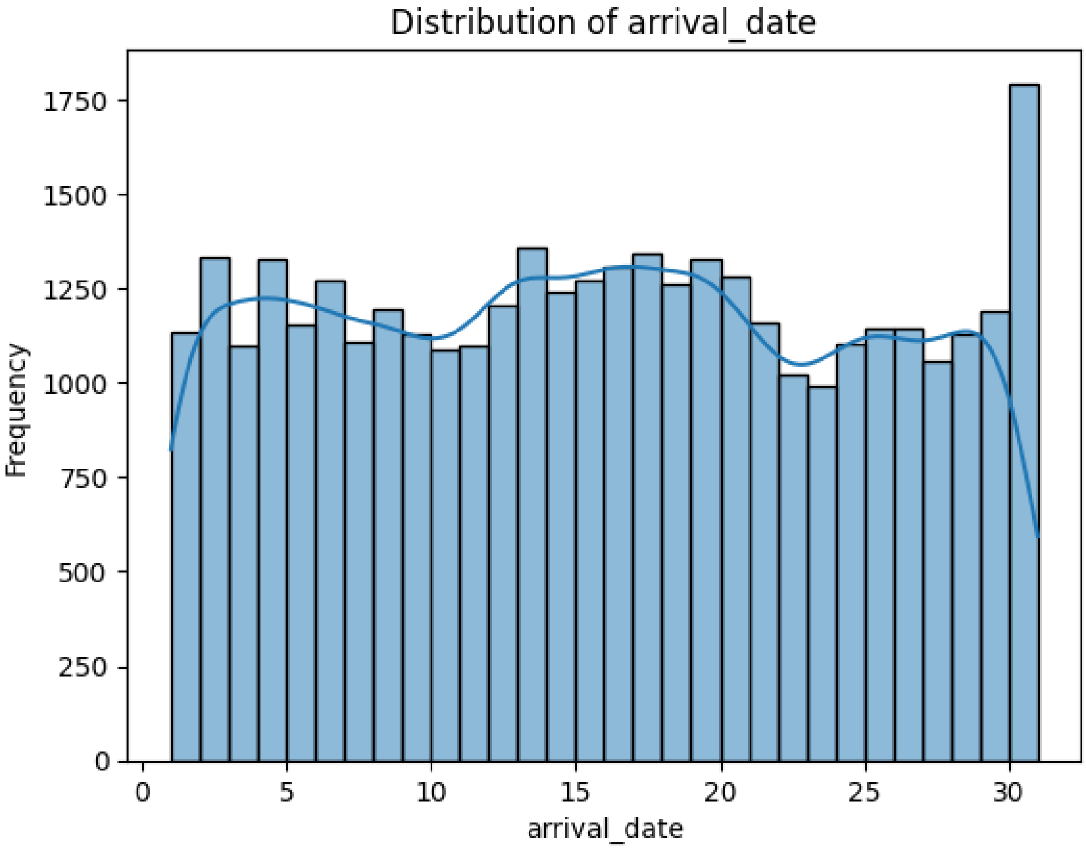

Figure 11. Distribution of Arrival Dates Within Each Month

Interpretation:

The distribution appears to be roughly uniform, with some minor fluctuations. There is a slight peak around the 15th day of the month and another smaller peak around the 30th day. Overall, the data suggests that arrivals are relatively evenly spread throughout the month, with no significant patterns or trends.

12. Booking Distribution Across Market Segments:

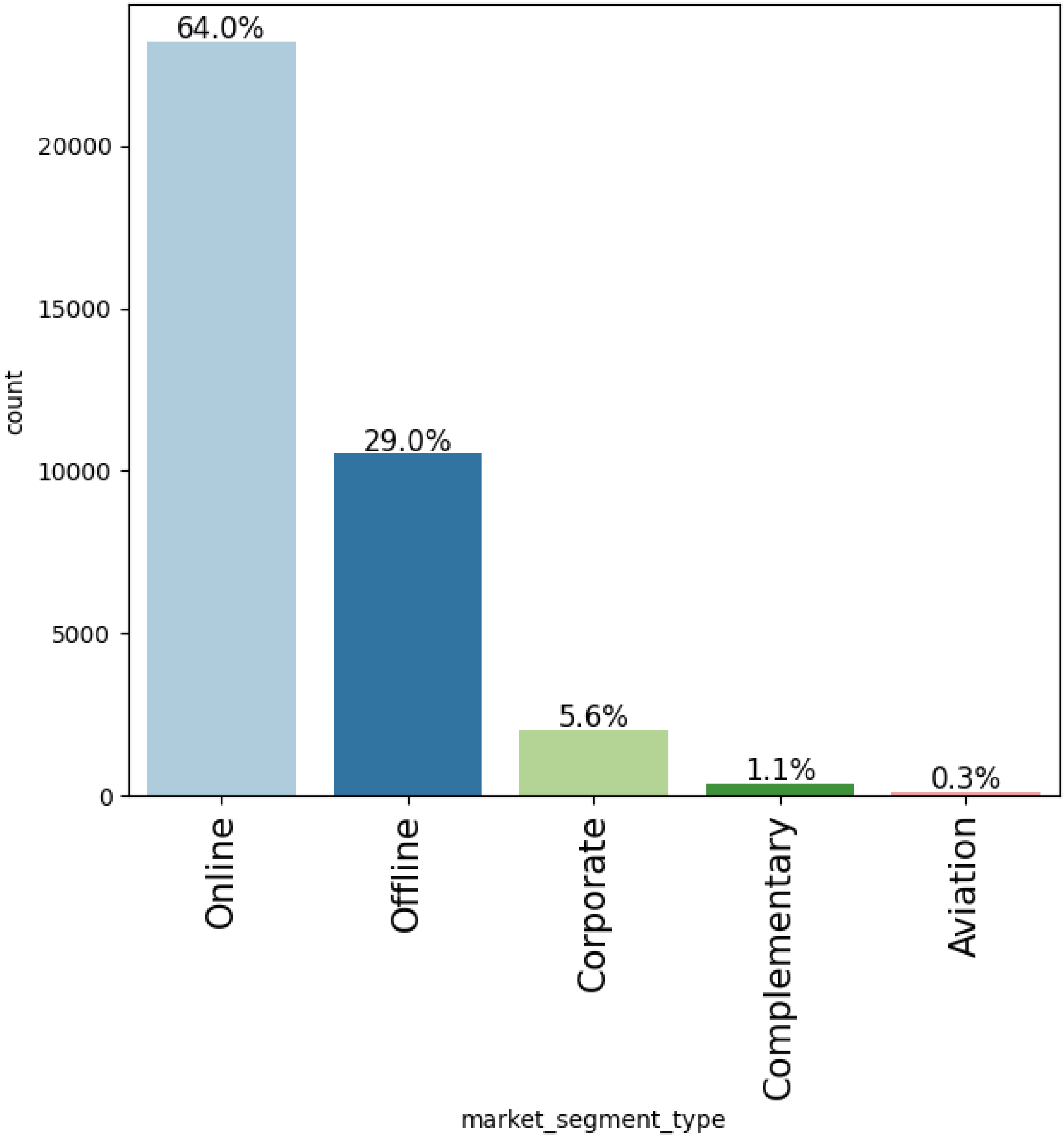

Figure 12. Booking Distribution Across Market Segments

Interpretation:

Online channels dominate with 64% of bookings, reflecting the growing reliance on digital platforms for travel planning. Offline and corporate bookings hold smaller shares of 29% and 5.6%, underscoring a shift toward convenience and online offers.

13. Percentage of Repeated Guests in Bookings:

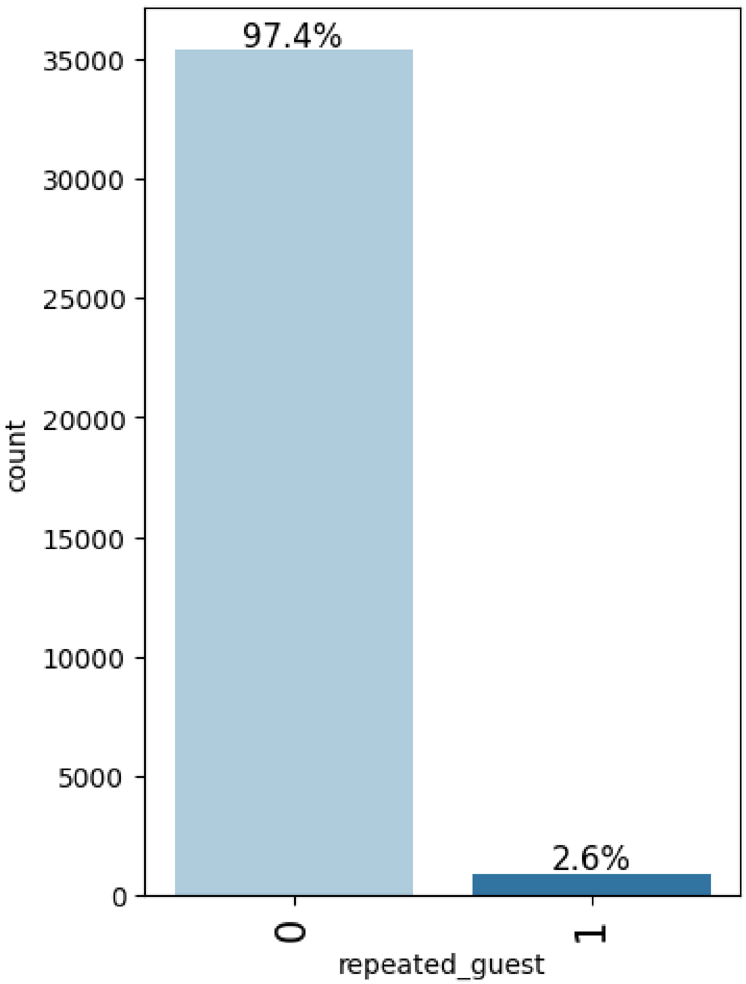

Figure 13. Percentage of Repeated Guests in Bookings

Interpretation:

Repeated guests represent just 2.6% of total bookings, indicating a low retention rate. This suggests a need for loyalty programs or initiatives to encourage return visits, especially in an industry where repeat customers can drive long-term profitability.

14. Count of Previous Cancellations by Guests:

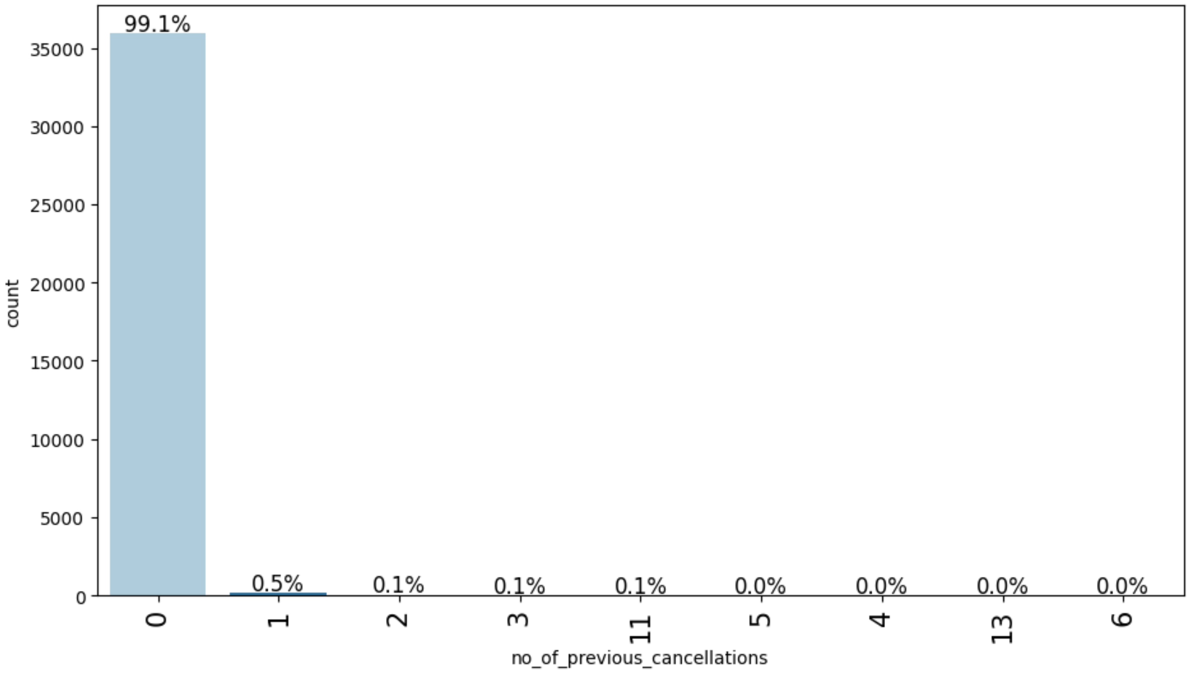

Figure 14: Count of Previous Cancellations by Guests

Interpretation:

The bar chart illustrates the distribution of previous cancellations among customers. The majority, 99.1%, have had no cancellations. A small percentage, 0.5%, have experienced one cancellation, while the remaining 0.3% account for more than one cancellation.

15. Popularity of Reserved Room Types:

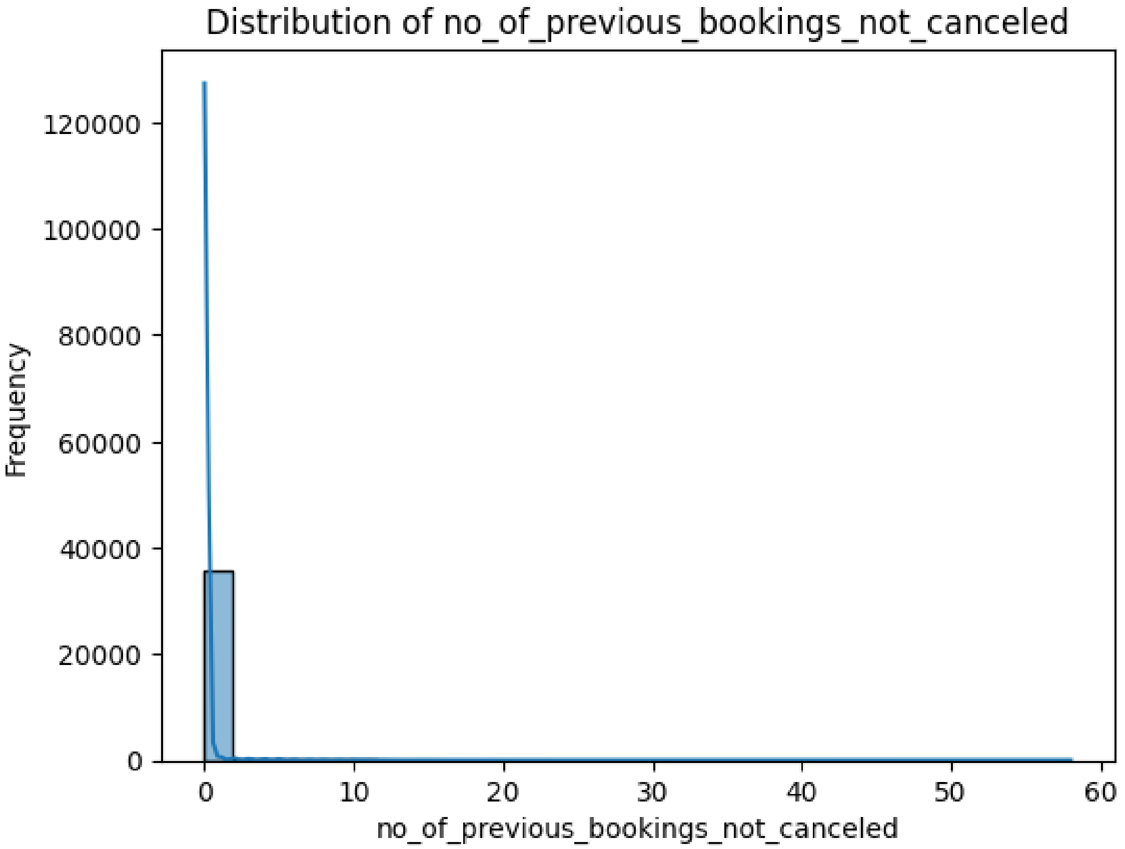

Figure 15. Popularity of Reserved Room Types

Interpretation:

Most guests are first-time visitors with no previous non-canceled bookings. A small group of loyal customers highlights potential opportunities for targeted engagement strategies to increase repeat bookings and strengthen customer relationships.

16. Distribution of Average Room Prices (in Euros):

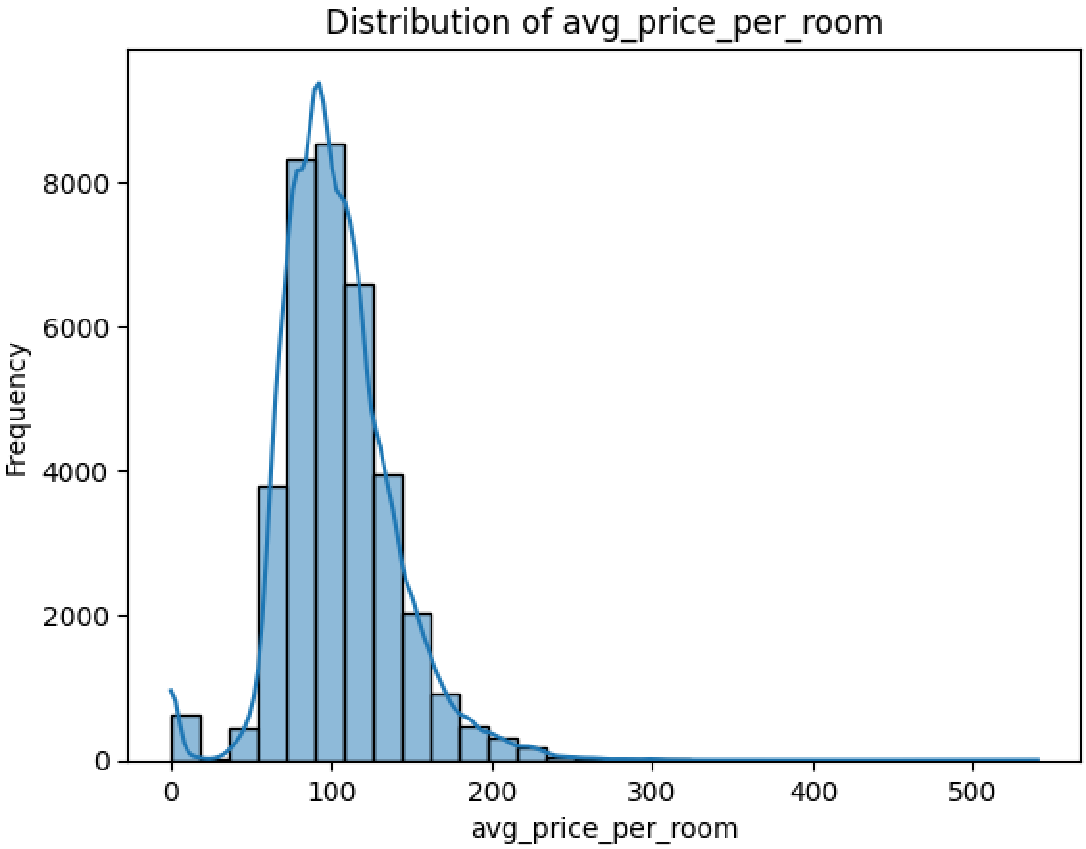

Figure 16. Distribution of Average Room Prices (in Euros)

Interpretation:

Room prices are concentrated between €50–€150, reflecting a mid-range pricing strategy. Premium rooms priced above €300 are rare, likely catering to niche segments. Price variations may be influenced by room types, demand, and booking channels.

17. Number of Special Requests Made by Guests:

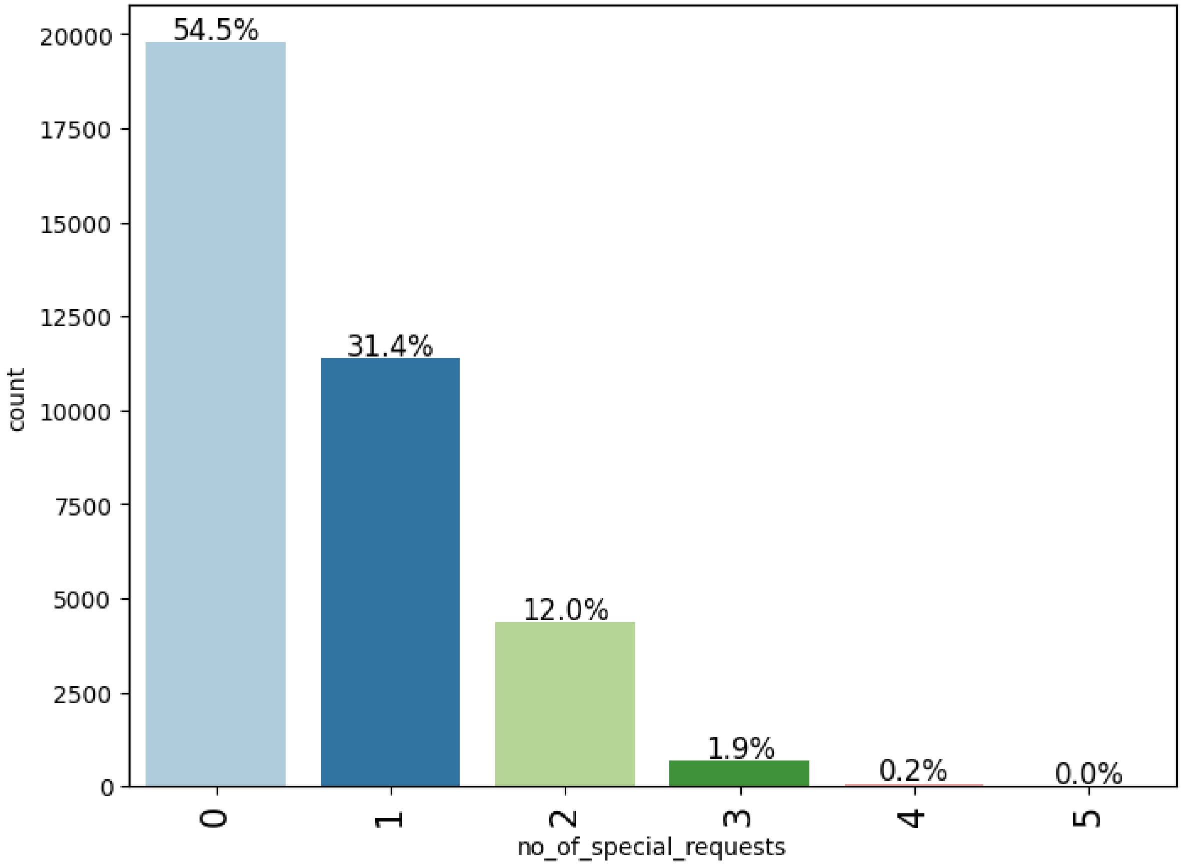

Figure 17. Number of Special Requests Made by Guests

Interpretation:

Around 54.5% of bookings have no special requests, 31.4% include one special request, 12% include two, 1.9% have three, and 0.2% feature four special requests. This distribution highlights varying customer preferences, with the majority opting for minimal or no additional requests during booking.

18. Proportion of Canceled vs. Non-Canceled Bookings:

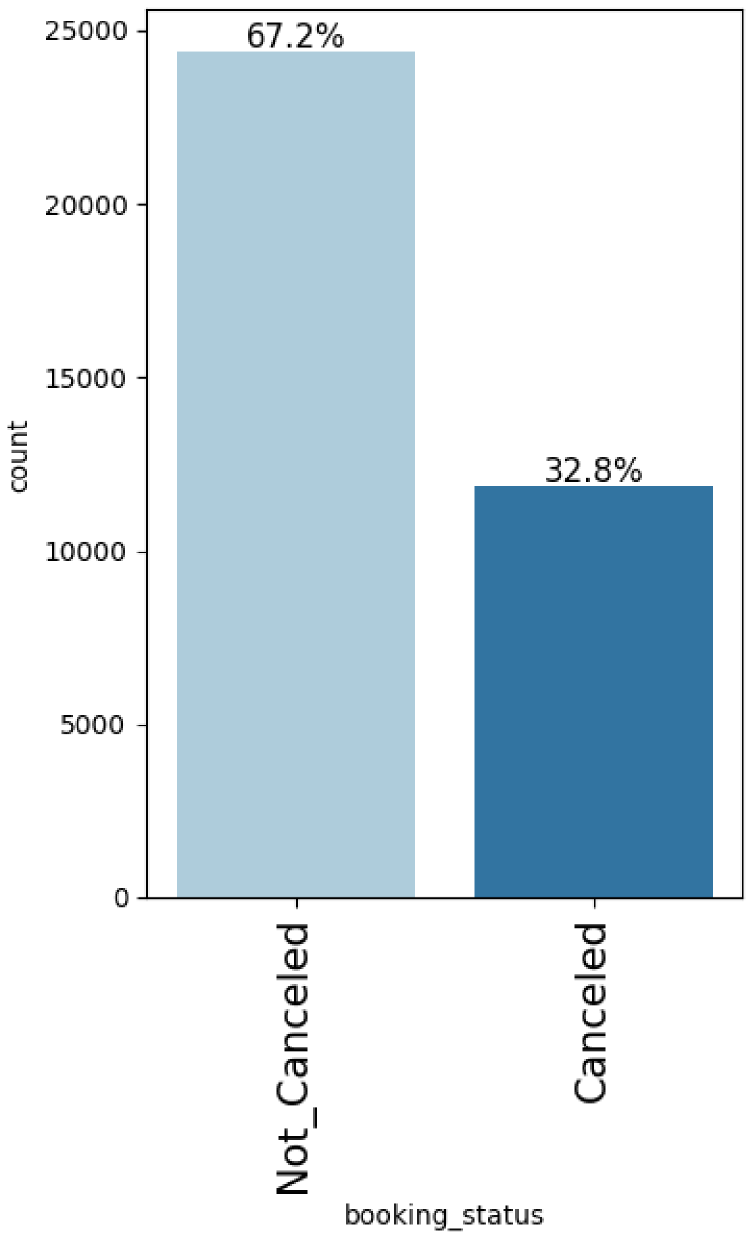

Figure 18. Proportion of Canceled vs. Non-Canceled Bookings

Interpretation:

The bar chart shows the distribution of booking statuses. The majority of bookings, 67.2%, were not canceled. In contrast, 32.8% of bookings were canceled. This indicates that a significant proportion of bookings were completed without cancellation.

1.4.Bivariate Analysis:

19. Relationship between No. of Adults vs. booking status:

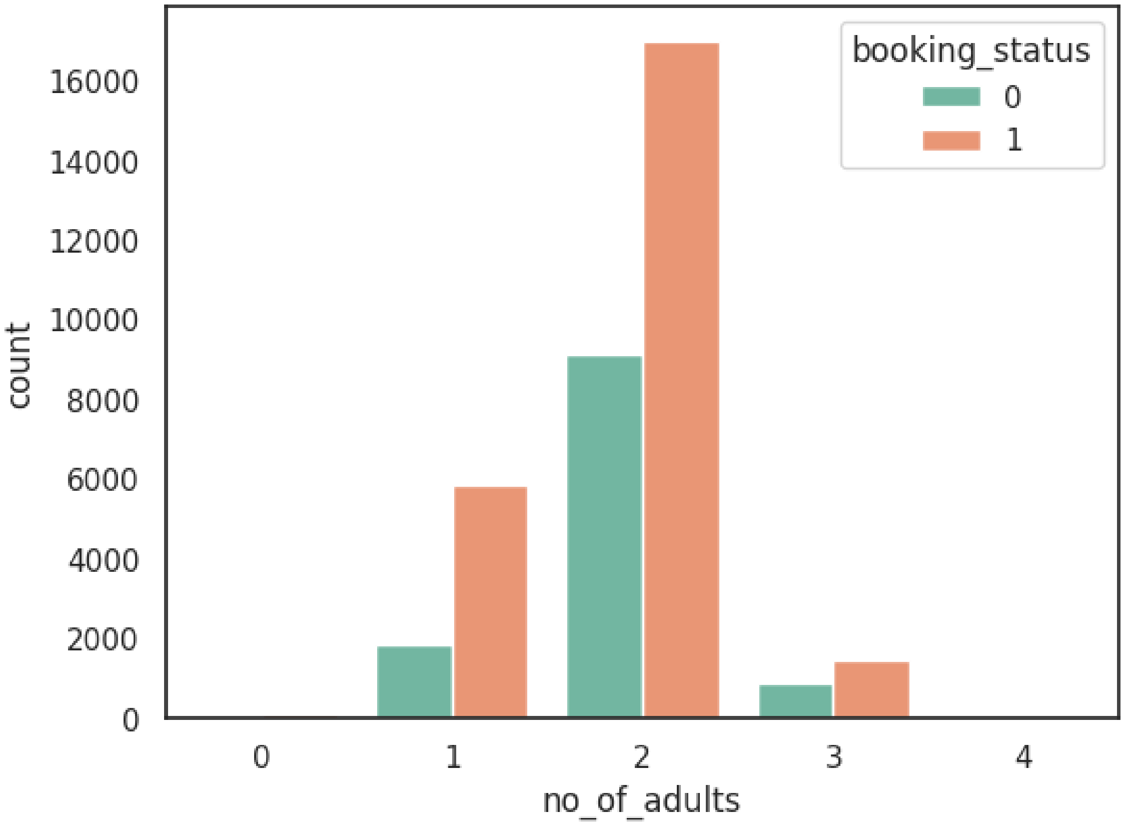

Figure 19.Relationship between No. of Adults vs. booking status

Interpretation:

The plot shows the count of bookings (booking_status) based on the number of adults (no_of_adults). For 1 adult, around 2,000 bookings are canceled and 6,000 confirmed. For 2 adults, approximately 9,000 bookings are canceled, while above 16,000 are confirmed, making it the most common group. For 3 adults, about 400 are canceled and 800 confirmed, while bookings for 4 adults are minimal, with both categories below 200. Confirmed bookings dominate overall, especially for 2 adults.

20. Relationship between No. of Children vs. booking_status :

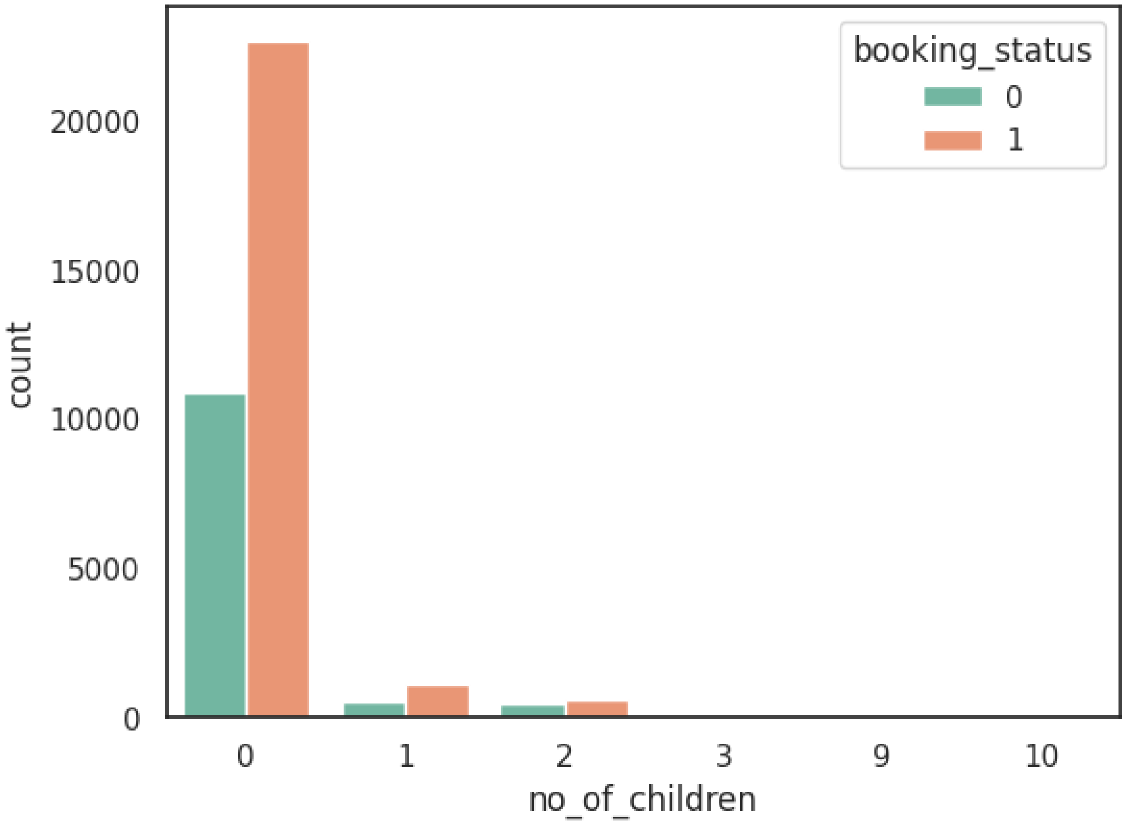

Figure 20. Relationship between No. of Children vs booking status

Interpretation:

The plot shows booking counts by the number of children (no_of_children). For 0 children, around 10,000 bookings are canceled (0), and over 22,000 are confirmed (1). For 1 child, about 1,000 bookings are canceled and 1,500 confirmed. For 2 children, cancellations and confirmations are each below 500. Bookings for 3 or more children are negligible. Confirmed bookings dominate, particularly for families with no children.

21. Relationship between No. of Weekend Nights vs. booking status:

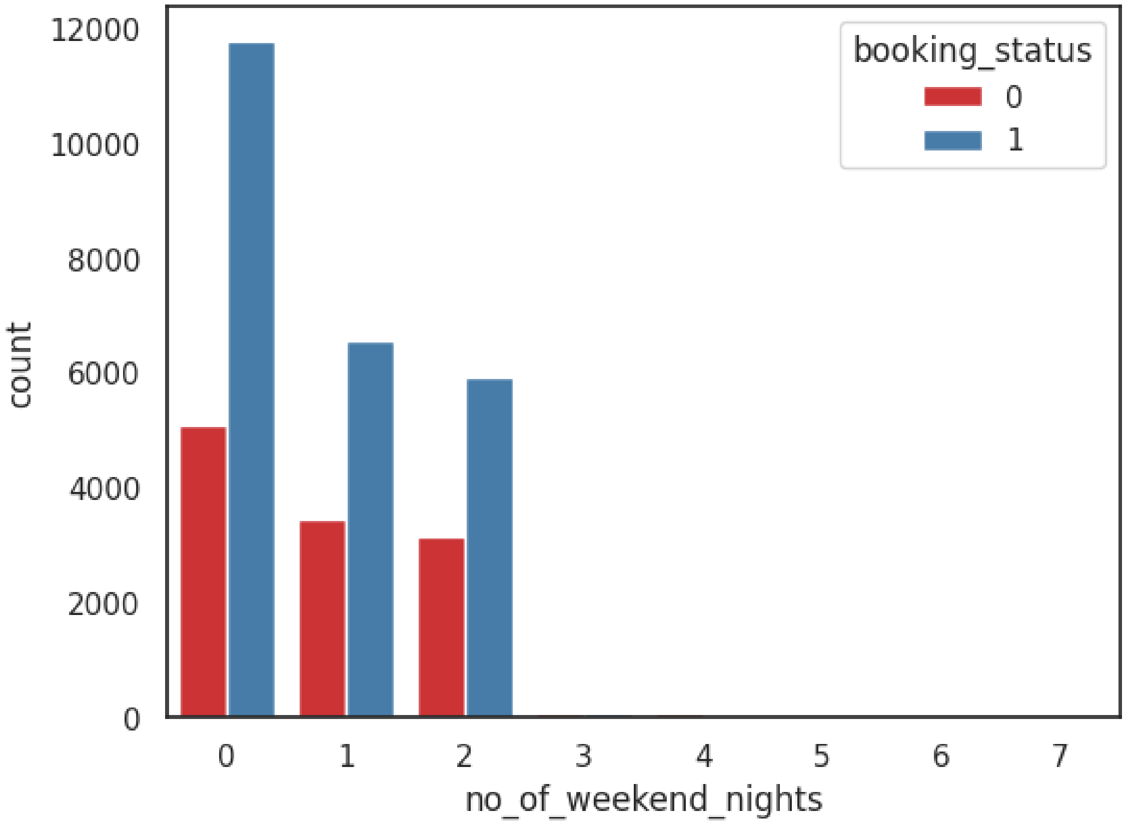

Figure 21. Relationship between No. of Weekend Nights vs. booking status

Interpretation:

The plot shows most bookings are for 0 weekend nights, with 5,000 cancellations and over 11,500 confirmations. For 1 weekend night, 6,500 were canceled and 3,500 confirmed, while for 2 weekend nights, cancellations are 3,000 and confirmations 5,800. Bookings for 3 weekend nights are minimal. Bookings without weekend nights are most common, with higher completion but notable cancellations. Analyzing cancellations could offer management insights.

22. Relationship between No. of Week Nights vs. booking status

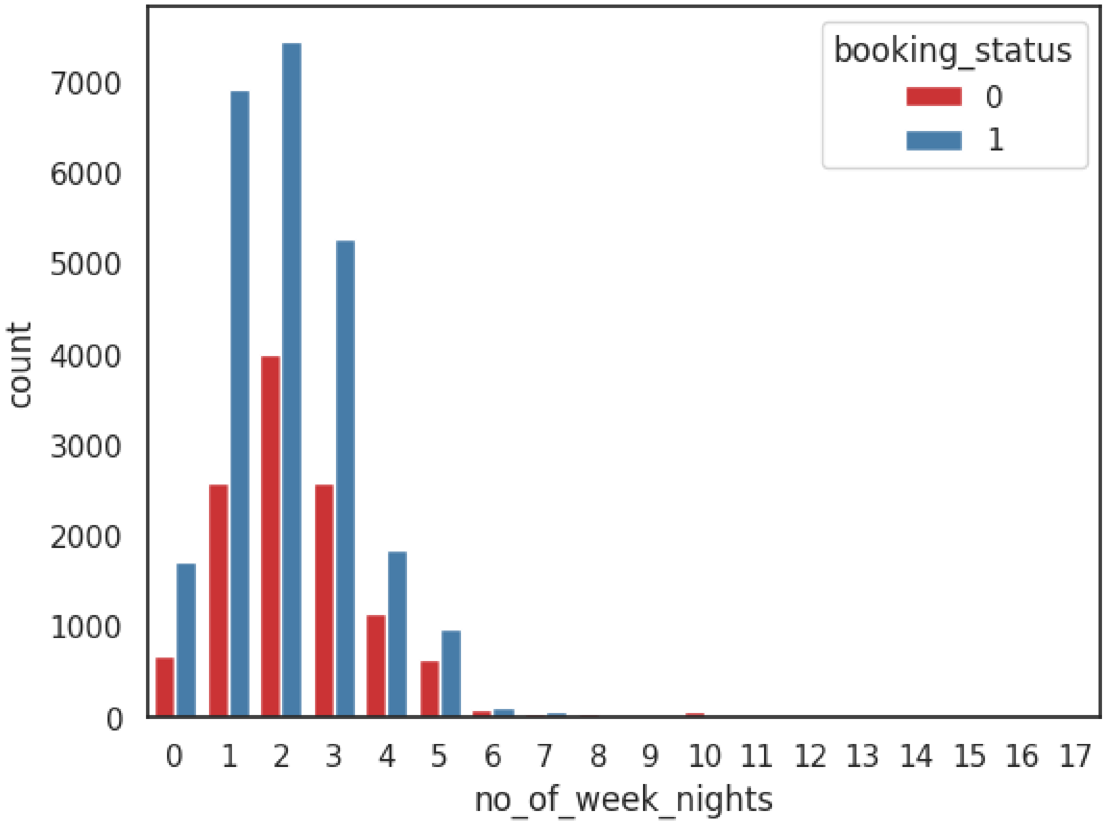

Figure 22. Relationship between No. of Week Nights vs. booking status

Interpretation:

Bookings for 1–3 weeknights have the highest counts, with confirmed bookings significantly outnumbering cancellations. For 0, 4, and 5 weeknights, bookings decrease, and the gap between confirmed and canceled bookings narrows. For 6–10 weeknights, bookings are fewer, with cancellations slightly exceeding confirmations. Beyond 11 weeknights, bookings are minimal and mostly canceled.

23. Relationship between Type of Meal Plan vs. booking status :

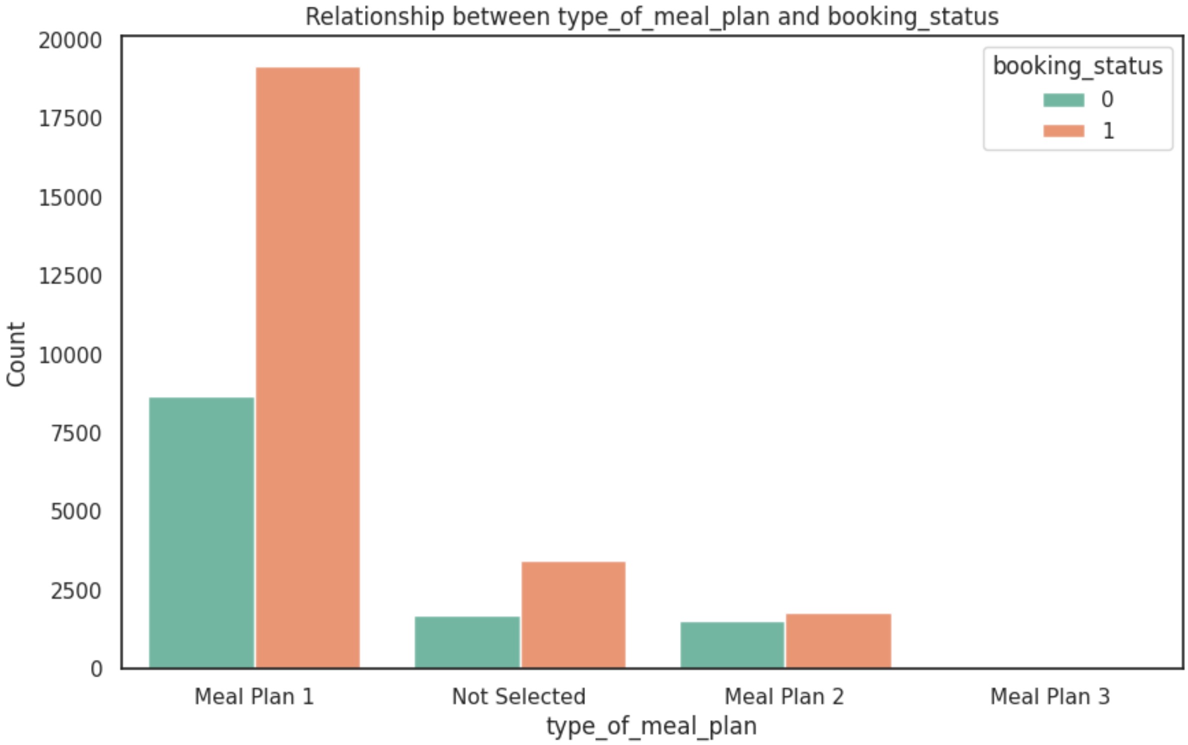

Figure 23. Relationship between Type of Meal Plan vs. booking status

Interpretation:

For Meal Plan 1, around 9,000 bookings were canceled and 19,000 confirmed, making it the most popular. Meal Plan 2 had 1,500 confirmations and 1,000 cancellations, while Meal Plan 3 had the fewest bookings in both categories. With no meal plan selected, about 1,800 bookings were canceled and 3,500 confirmed. Meal Plan 1 dominates, with significantly higher confirmations.

24. Relationship between Required Car Parking Space vs. booking status:

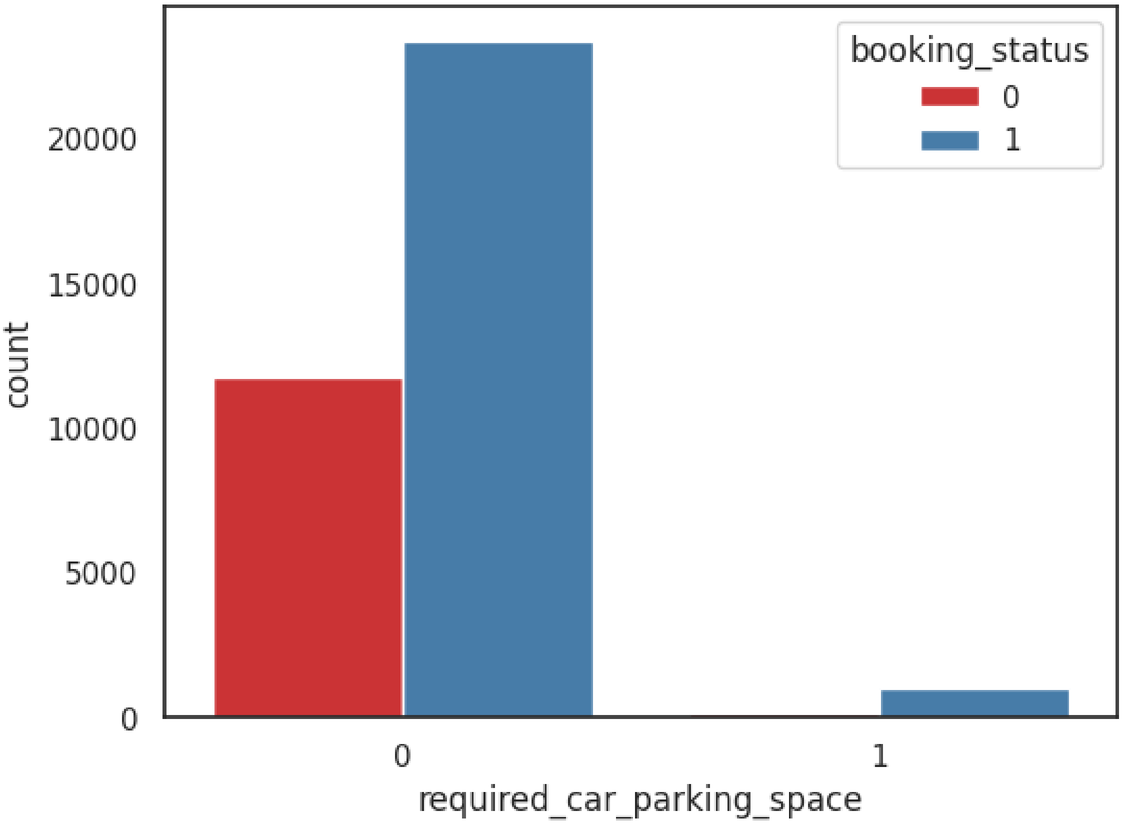

Figure 24. Relationship between Required Car Parking Space vs. booking status

Interpretation:

For bookings with no car parking space required, around 12,000 were canceled and over 23,000 confirmed, indicating a higher completion rate. For bookings with car parking space required, very few were confirmed, with about 1,000 cancellations. The data shows that bookings without car parking are far more common and have a significantly higher confirmation rate.

25. Relationship between Room Type Reserved vs. booking status:

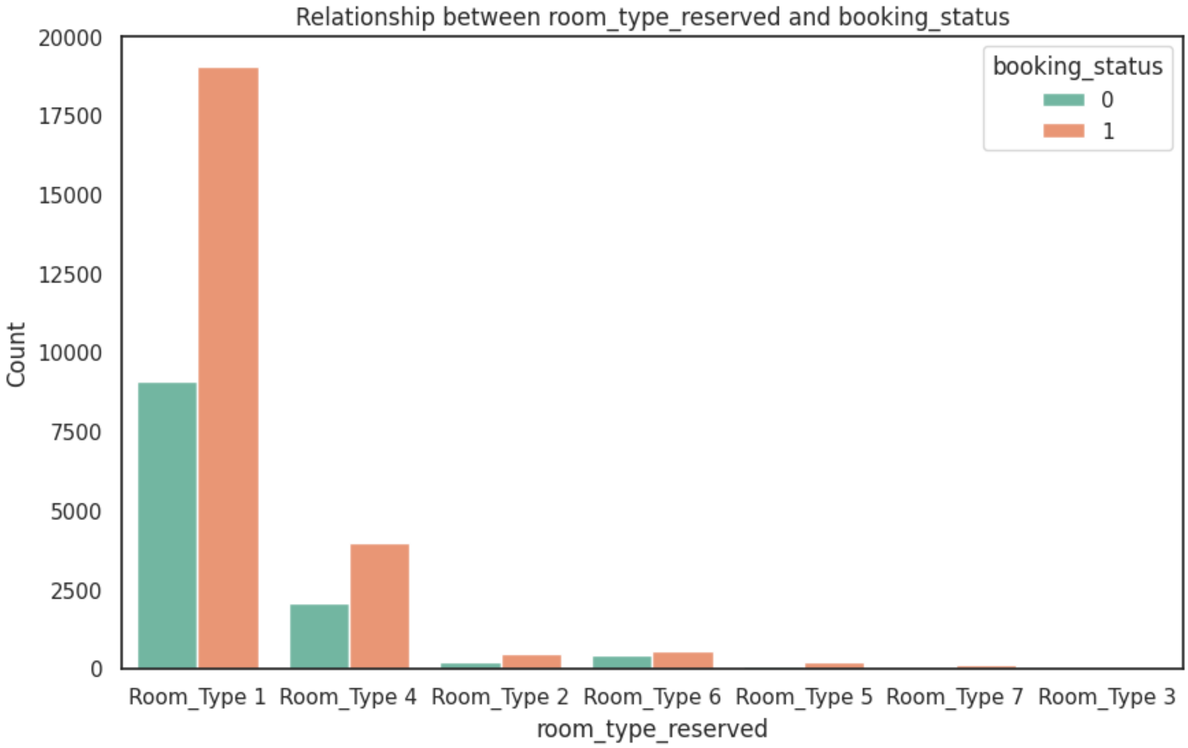

Figure 25. Relationship between Room Type Reserved vs. booking status:

Interpretation:

Room Type 1 has the highest bookings, with 9,000 canceled and 19,000 confirmed. Room Type 4 follows with 3,500 confirmed and 1,500 canceled. Room Types 2, 6, and 5, 7, 3 have fewer bookings, with Room Type 2 showing 1,500 confirmed and 500 canceled, and Room Type 6 having 1,000 confirmed and 800 canceled. Bookings for Room Types 5, 7, and 3 are minimal.

26. Relationship between Lead Time vs. booking status:

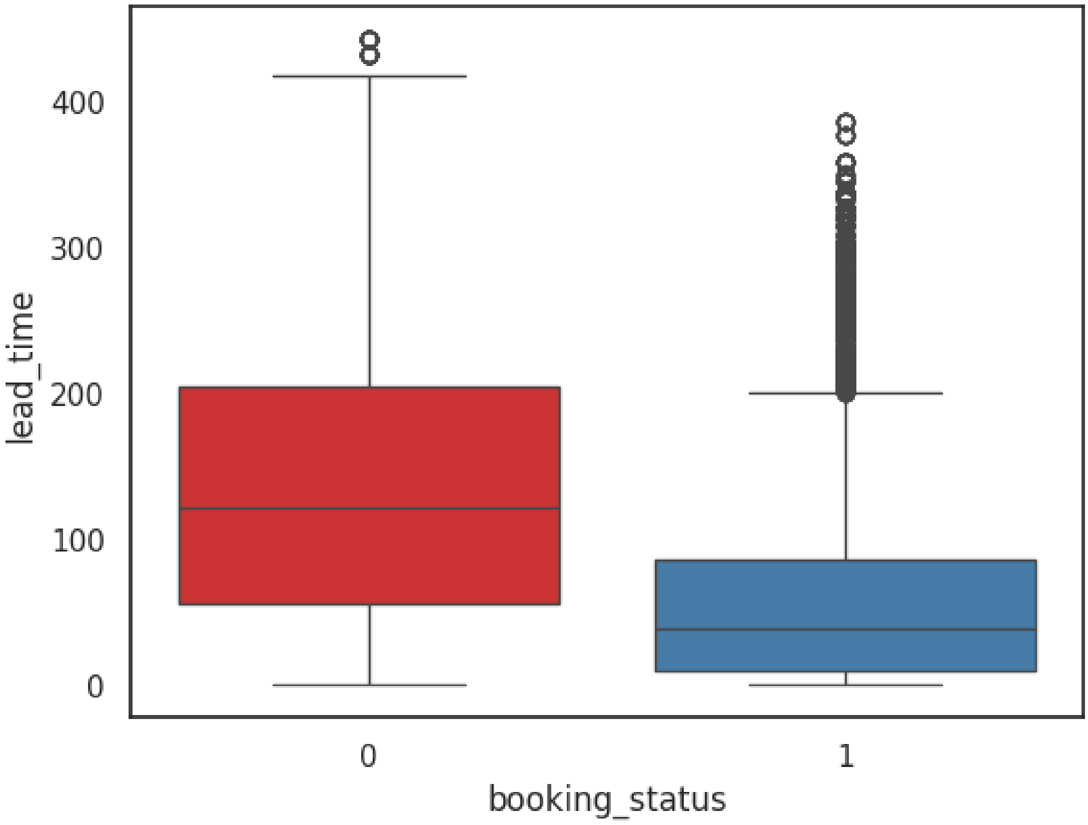

Figure 26. Relationship between Lead Time vs. booking status

Interpretation:

The box plot shows that bookings with longer lead times are more likely to be canceled. Canceled bookings have a median lead time of around 120 days, with a larger interquartile range (IQR) and several long outliers. Confirmed bookings have a median of 50 days and a smaller IQR, with fewer outliers. Analyzing cancellation reasons and considering other factors like room type and customer demographics could provide deeper insights.

27. Relationship between Arrival Month vs. booking status:

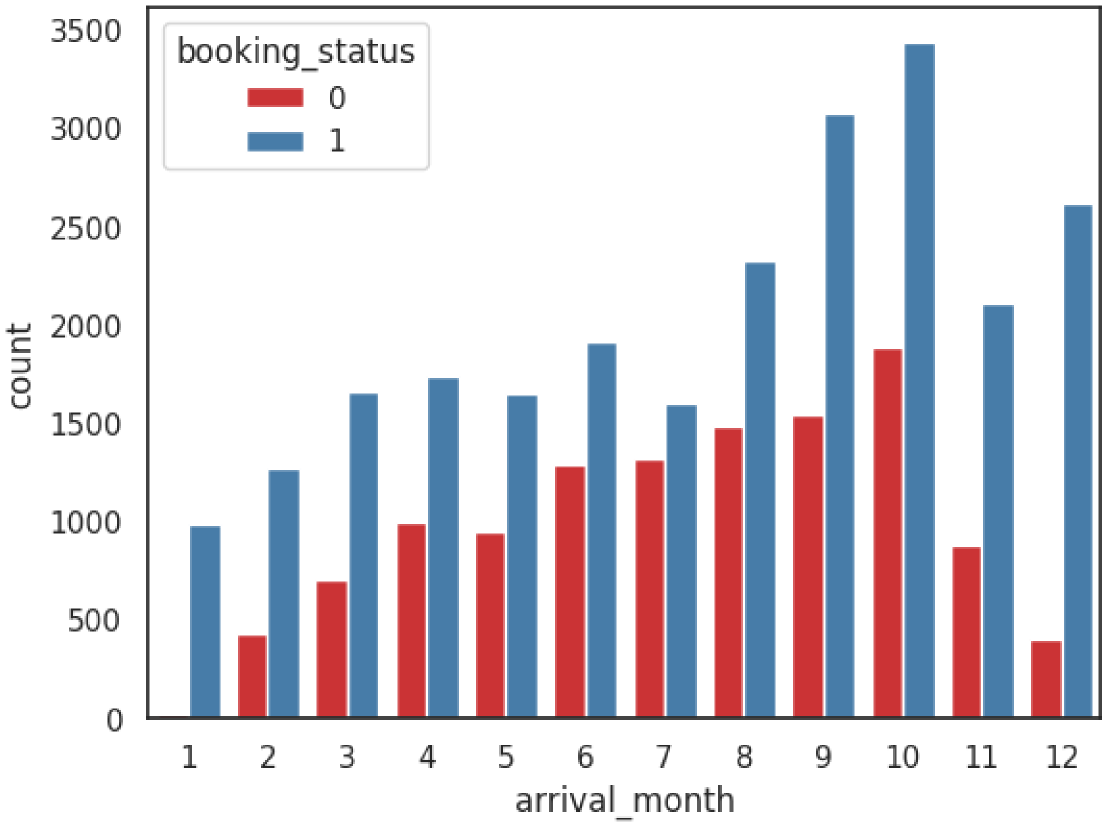

Figure 27. Relationship between Arrival Month vs. booking status

Interpretation:

October has the highest bookings, with around 3,400 confirmed and 1,800 canceled, followed by September with 3,000 confirmed and 1,500 canceled. January has the fewest cancellations, and overall, bookings gradually increased from January to October before dipping in November and December. This trend indicates a peak in bookings during autumn, followed by a decline towards the year-end.

28. Relationship between Market Segment Type vs. booking status:

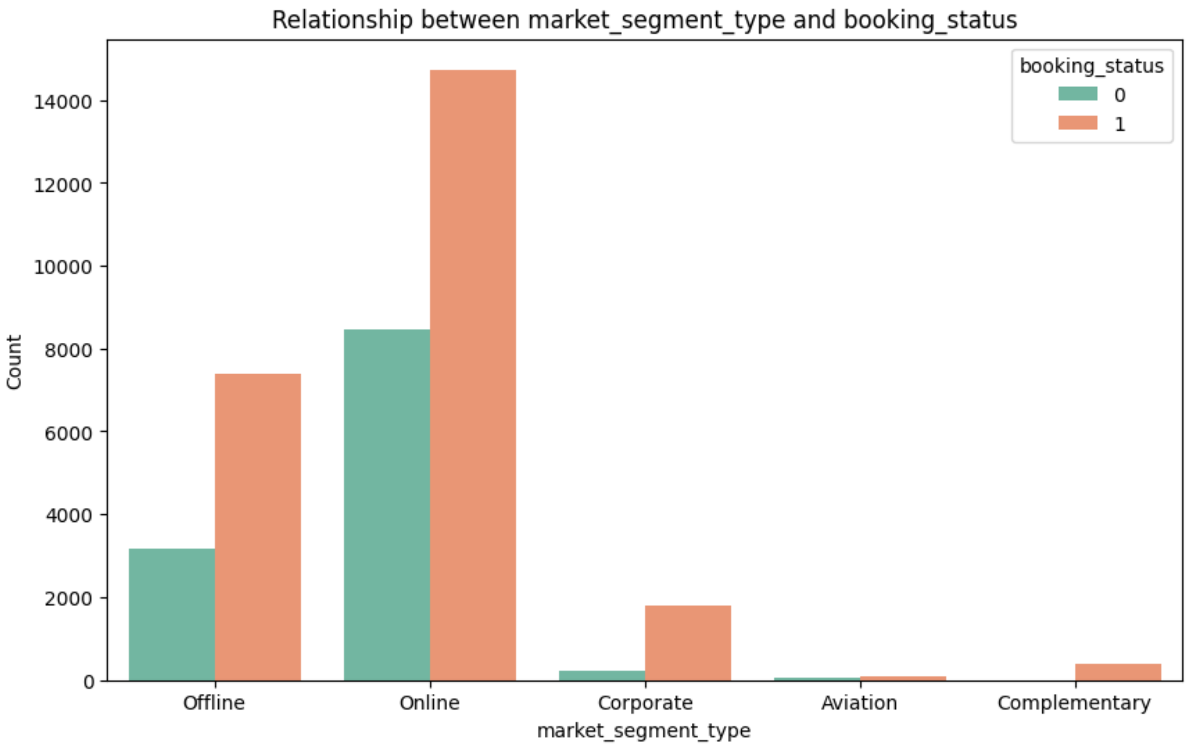

Figure 28. Relationship between Market Segment Type vs. booking status

Interpretation:

Interpretation:

The Online segment has the highest bookings, with confirmed significantly outnumbering cancellations. The Offline segment has fewer bookings, with confirmed still dominant withh a considerable gap. Other segments, such as Corporate, Aviation, and Complimentary, have minimal bookings and relatively higher cancellation rates, showing the prominence of online bookings in overall trends.

29. Relationship between Repeated Guest vs. booking status:

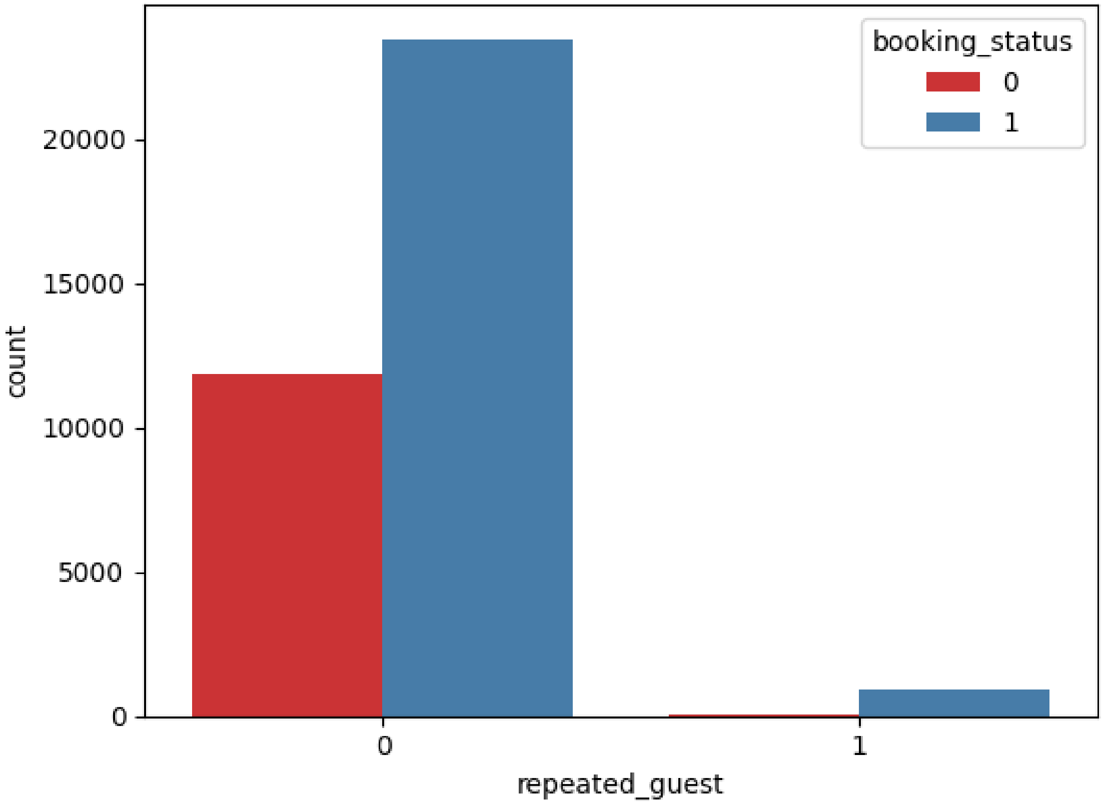

Figure 29. Relationship between Repeated Guest vs. booking status

Interpretation:

New guests have the highest bookings, with confirmed significantly outnumbering cancellations. Repeat guests account for far fewer bookings, with confirmed still dominant but a narrower gap between confirmed and canceled bookings. This highlights the prominence of new guests in overall booking trends.

30. Relationship between Avg. Price per Room vs. booking status:

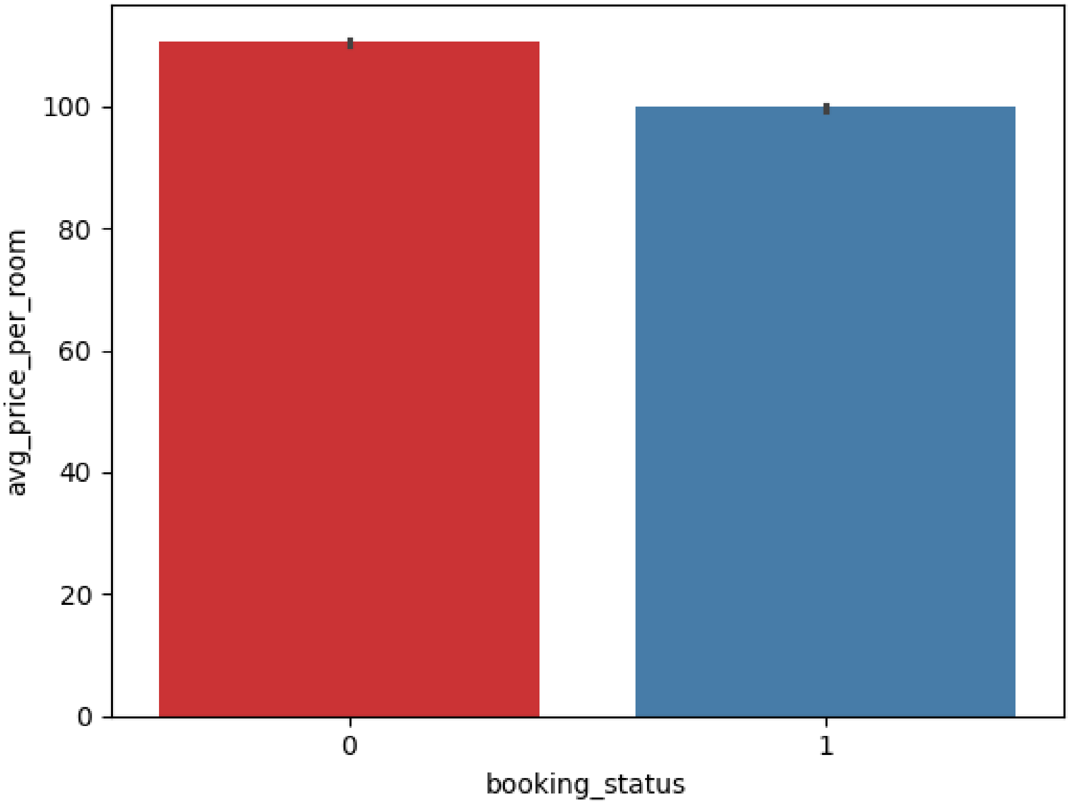

Figure 30. Relationship between Avg. Price per Room vs. booking status

Interpretation:

Canceled bookings have a higher average price per room, with greater variability indicated by larger error bars, showing some significantly high prices. Confirmed bookings have a lower average price per room and smaller error bars, indicating less variability. This suggests price stability for confirmed bookings and a possible link between higher prices and cancellations.

31. Relationship between No. of Special Requests vs. booking status:

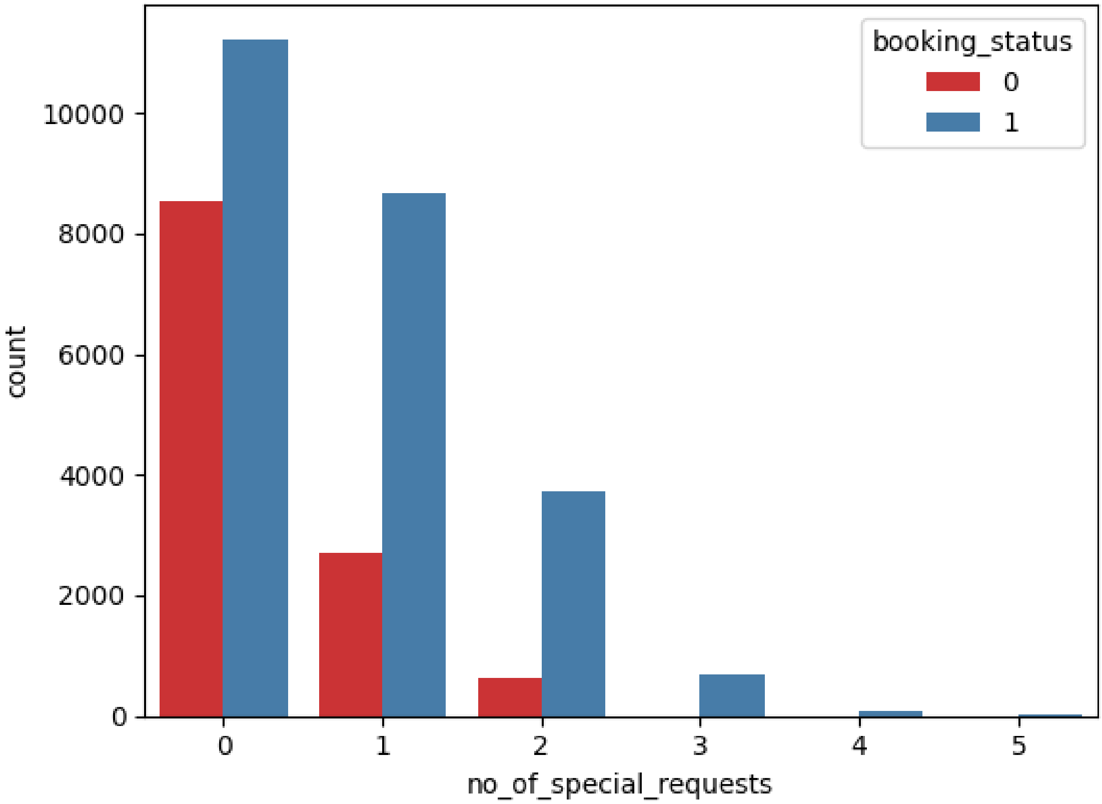

Figure 31. Relationship between No. of Special Requests vs. booking status

Interpretation:

Bookings with no special requests are the highest, with confirmed significantly outnumbering canceled bookings. For 1–3 special requests, bookings decrease as requests increase, but confirmed bookings still dominate with a wider gap. Bookings with 4–5 special requests are minimal, and the cancellation rate is lower, suggesting a link between fewer requests and higher booking volumes.

1.5. Key Answers to the key questions:

1. What are the busiest months in the hotel:

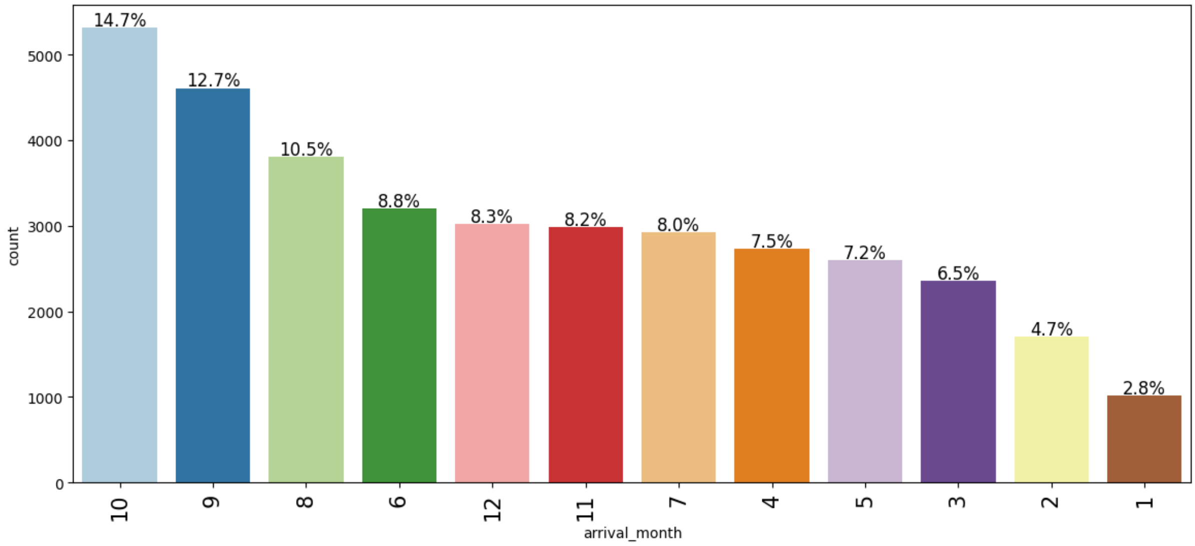

Figure 10. Seasonal Trends in Guest Arrivals (By Month)

Answer 1:

The bar chart shows the distribution of arrivals across different months. The month of October saw the highest number of arrivals (busiest month), accounting for 14.7% of the total. This is followed by September with 12.7% and August with 10.5%. The number of arrivals gradually decreases from October to January, with January having the lowest number of arrivals at 2.8%.

2. Which market segment do most of the guests book from?

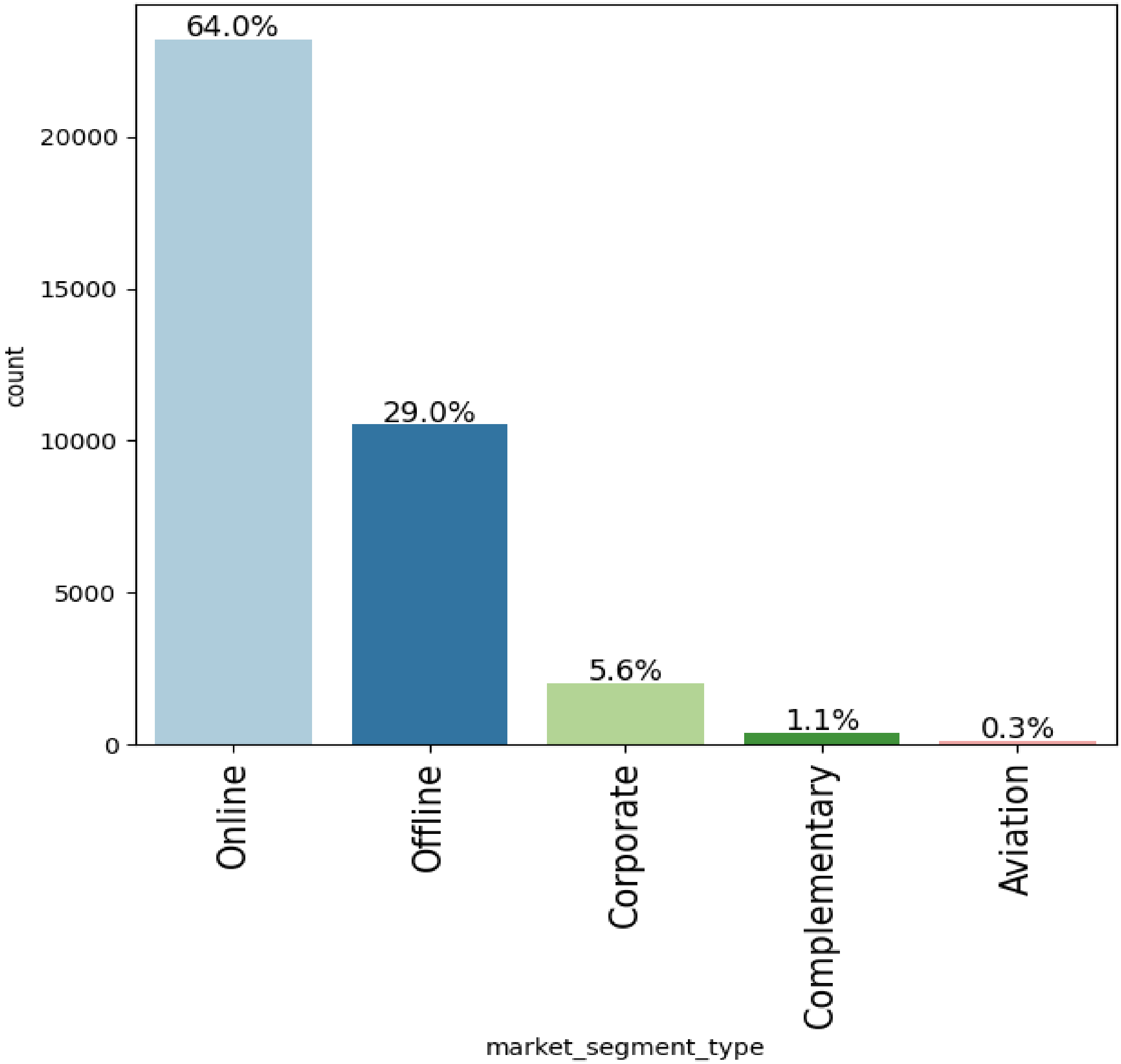

Figure 12. Booking Distribution Across Market Segments

Answer 2:

Online channels now account for 64% of bookings, highlighting the increasing reliance on digital platforms for travel planning. In contrast, offline and corporate bookings represent smaller shares at 29% and 5.6%, respectively, signaling a shift towards the convenience and variety of online options. The complementary segment makes up 1.1%, while the aviation segment contributes just 0.3%, reflecting their limited roles in the overall booking landscape.

3. What are the differences in room prices in different market segments?

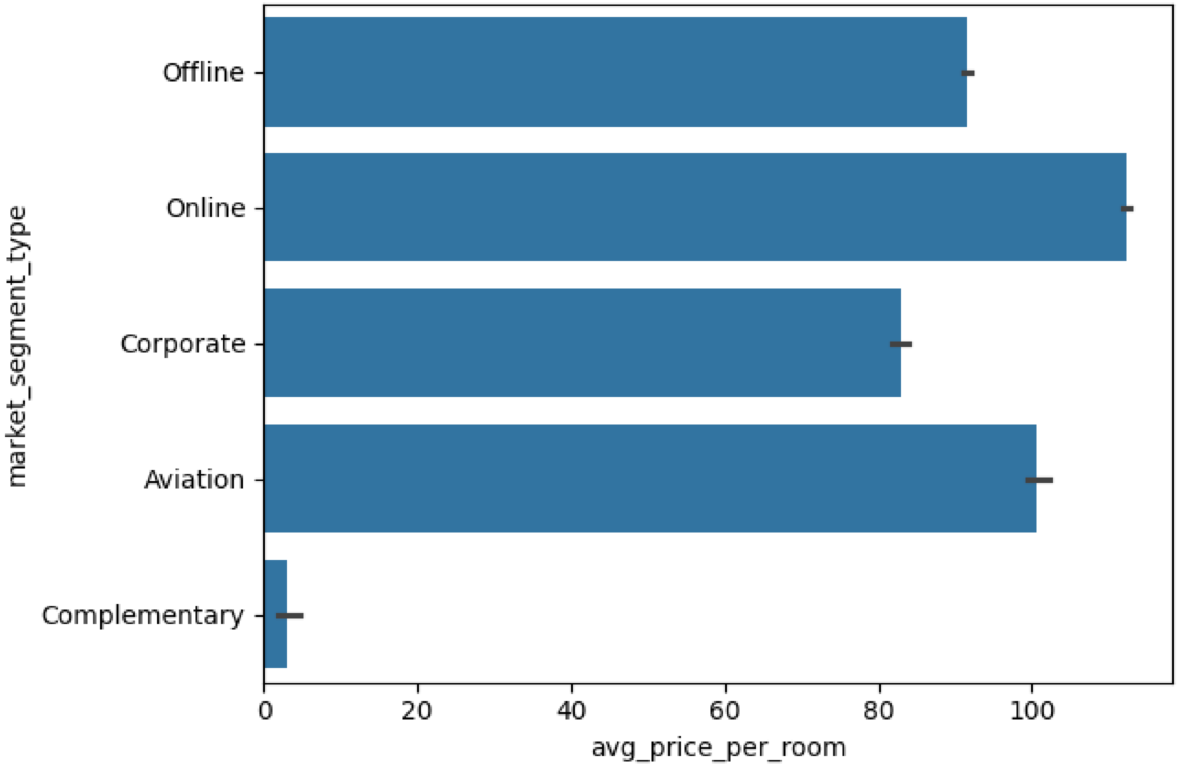

Figure 32. Average First-Day Views by Day of the Week for ShowTime Content

Answer 3:

The bar plot illustrates the average price per room in euros for different market segments. Online bookings have the highest average price, exceeding 100 euros per room. In contrast, complementary bookings have the lowest average price, with the bar barely visible. Corporate and aviation bookings fall in the middle range, with average prices around 80 and 95 euros per room, respectively. This suggests that market segment is a significant factor influencing room pricing.

4. What percentage of bookings are canceled?

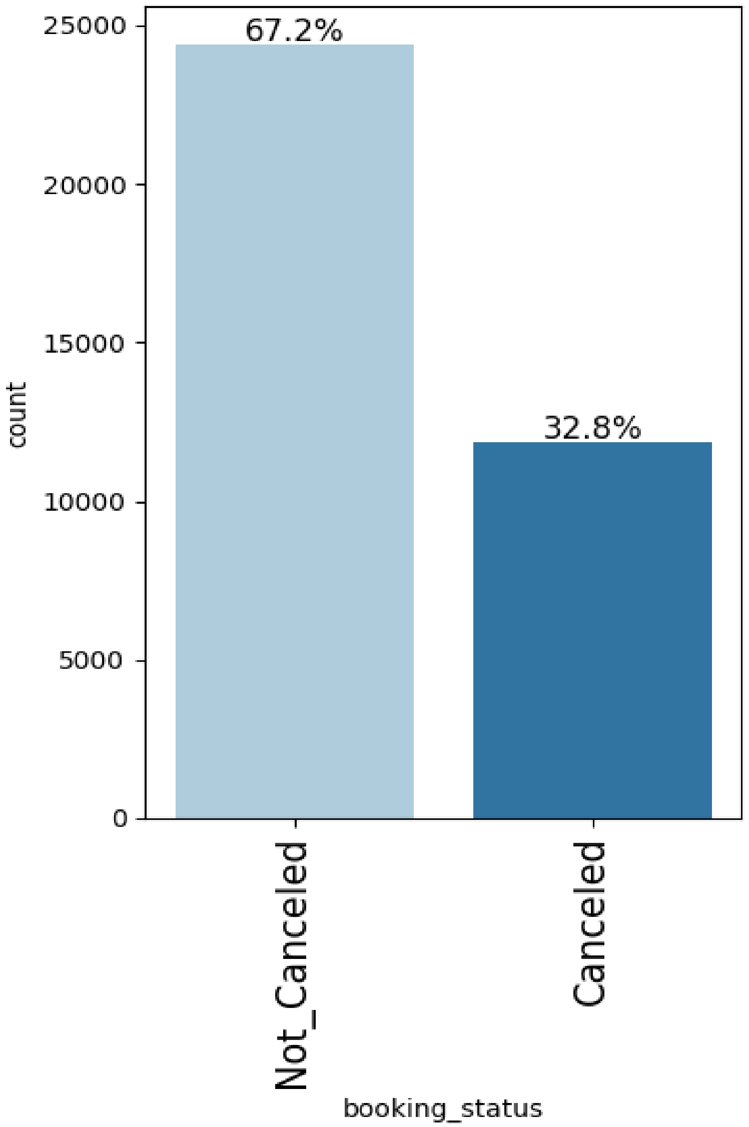

Figure 18. Proportion of Canceled vs. Non-Canceled Bookings

Answer 4:

The bar plot shows the distribution of booking statuses, with the height of each bar representing the count of bookings in each category. From the plot, we can see that the majority of bookings are not canceled. The taller bar for "Not_Canceled" indicates that this category has a higher count. The percentage displayed on top of this bar shows that approximately 67.2% of bookings are not canceled. Therefore, the percentage of canceled bookings are approximately 32.8%.

5.Repeating guests are the guests who stay in the hotel often and are important to brand equity?

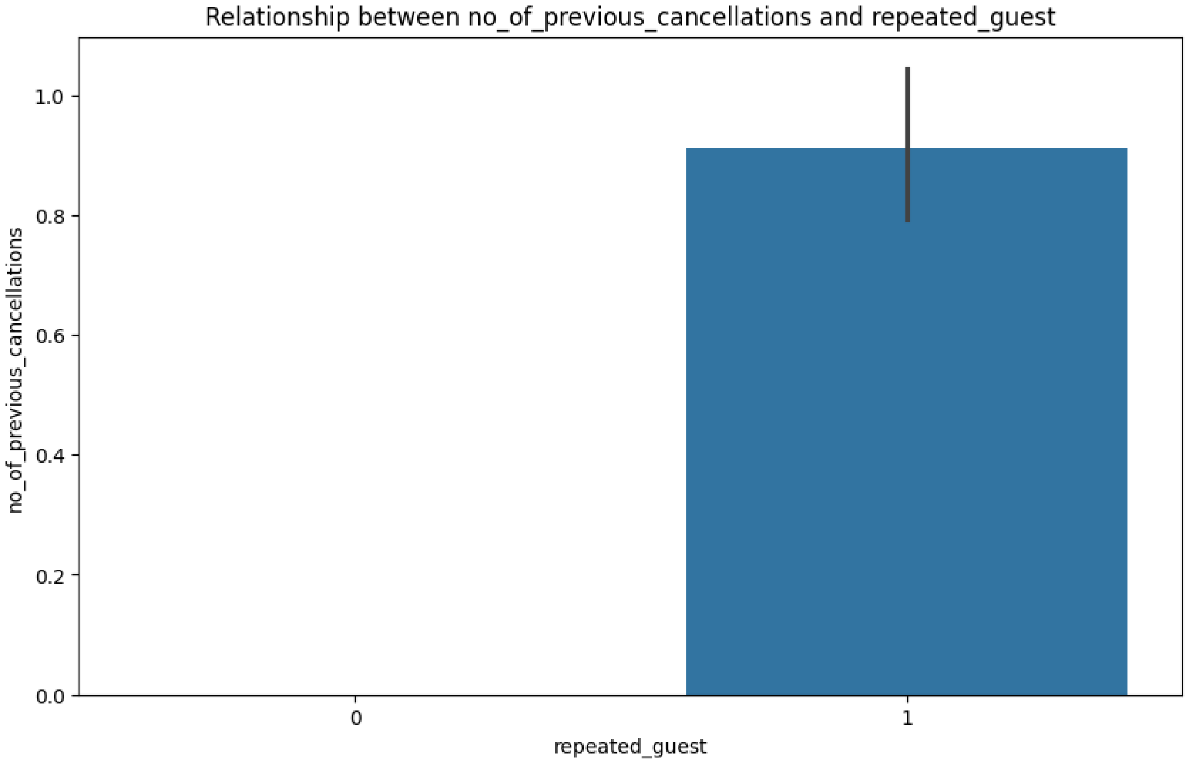

Figure 33. Relationship between Repeated Guest vs. No. of Previous Cancellations

Answer 5:

The plot shows the number of previous cancellations for repeated and non-repeated guests. The bar for repeated guests (represented by 1) is significantly higher than the bar for non-repeated guests (represented by 0). This indicates that, on average, repeated guests have a higher number of previous cancellations compared to non-repeated guests. Repeated guests are the most valuable assets. Implementing targeted loyalty programs and data-driven strategies to retain them and reduce cancellations must be an important step.

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

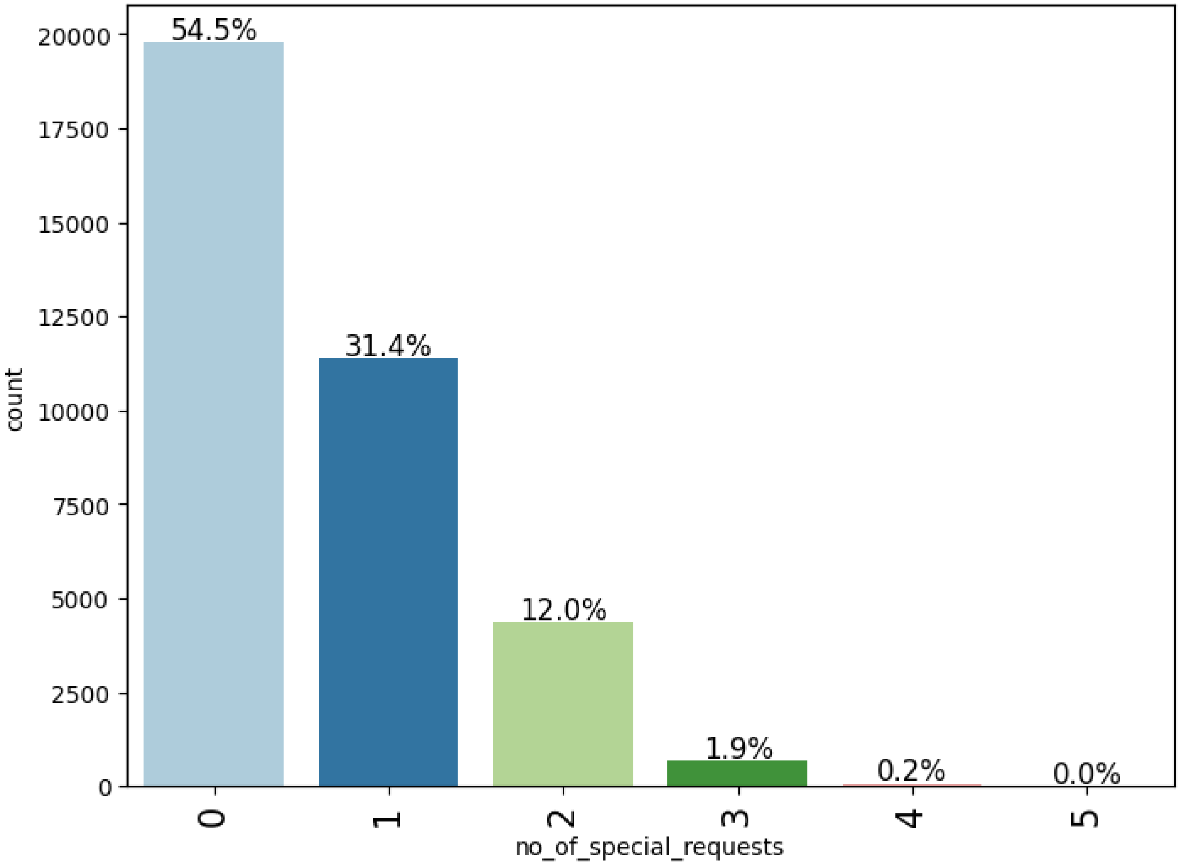

Figure 17. Number of Special Requests Made by Guests

Answer 6:

The bar plot illustrates that a majority of guests (54.5%) do not have any special requests. As the number of special requests increases, the number of guests with those requests decreases significantly. This suggests that guests with more specific requirements might be more committed to their stay. They may have a higher investment in their trip and are likely to be more satisfied with a personalized experience. Therefore, accommodating these special requests could lead to higher guest satisfaction and lower cancellation rates.

7. How does lead time (time between booking and check-in) affect booking cancellations?

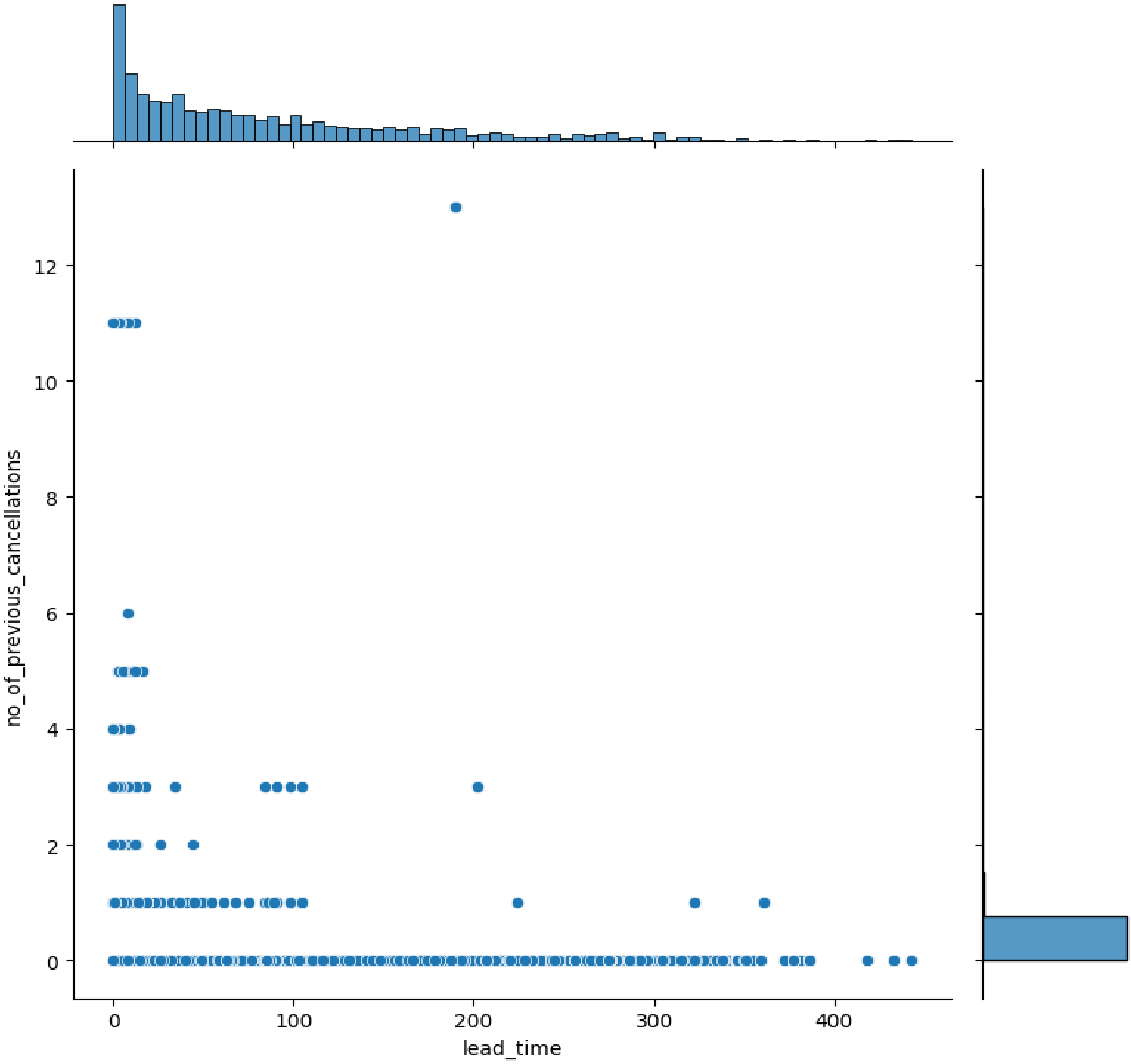

Figure 34. Relationship between Lead Time vs. No. of Previous Cancellations

Answer 7:

The plot reveals that most bookings (over 50%) have a lead time of less than 100 days. As the lead time increases, the number of bookings significantly decreases. For instance, there are very few bookings with a lead time exceeding 300 days. Interestingly, the plot also suggests a correlation between lead time and previous cancellations. Bookings with shorter lead times (less than 100 days) tend to have a higher number of previous cancellations compared to those with longer lead times. This implies that guests with longer booking horizons are more likely to fulfill their reservations.

8. What is the relationship between room types, booking status, and the need for car parking space?

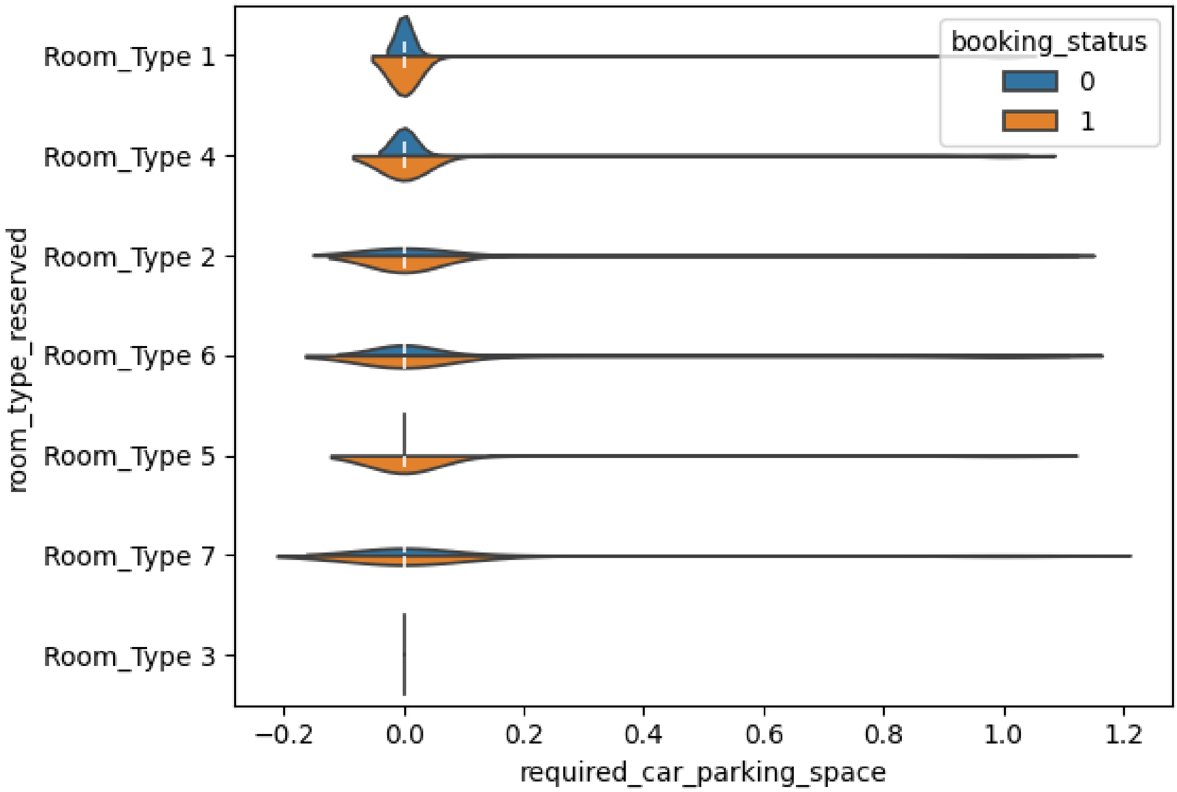

Figure 35. Distribution of Required Car Parking Space by Room Type and Booking Status

Answer 8:

The plot presents a Violin Plot comparing the distribution of required car parking space for different room types reserved, split by booking status (0 for canceled, 1 for confirmed). For Room Type 1, the canceled bookings (booking_status 0) tend to require no parking space, while confirmed bookings (booking_status 1) show a slightly higher proportion requiring parking. Room Types 4, 2, and 6 show similar distributions, with confirmed bookings requiring more parking spaces on average. Room Type 3 has the widest distribution of car parking space required, with a higher density of canceled bookings needing parking space.

1.6.Insights Based on EDA

Key Observations from Univariate Analysis:

1. **Lead Time**: The majority of bookings are made with short lead times, typically under 50 days. However, there is a long tail, indicating some bookings are made well in advance.

2. **Market Segment Type**: Online bookings dominate, followed by corporate and offline bookings. Complementary bookings represent a smaller portion but show unique cancellation and request trends.

3. **Room Type Distribution**: Certain room types are in higher demand, indicating customer preferences based on affordability or amenities offered.

4. **Number of Special Requests**: Most guests have either no special requests or a minimal number, but a small subset of guests tends to make multiple requests.

5. **Cancellations**: A small proportion of guests have prior cancellations, which might indicate recurring patterns or issues in specific demographics.

Key Observations from Bivariate Analysis:

1. **Number of Adults vs. Booking Status:**
    - Bookings with 1 or 2 adults are the most common, with confirmed bookings significantly outnumbering cancellations, especially for 2 adults. Bookings for 3+ adults are less frequent, with cancellations showing a higher proportion.

2. **Number of Children vs. Booking Status:**
    - Families with no children dominate the bookings, with confirmed bookings significantly higher than cancellations. As the number of children increases, cancellations become more prominent, and the overall number of bookings decreases.

3. **Number of Weekend Nights vs. Booking Status:**
    - Most bookings are for 0 weekend nights, with a notable drop in cancellations for these bookings. For 1 weekend night, cancellations surpass confirmations.

The overall trend shows a higher confirmation rate for bookings without weekend nights.

4. **Number of Week Nights vs. Booking Status:**
    - Confirmed bookings dominate for 1-3 weeknights, while cancellations rise for bookings with 4-5 weeknights. For 6 or more weeknights, cancellations slightly exceed confirmations.
5. **Type of Meal Plan vs. Booking Status:**
    - Meal Plan 1 is the most popular, with significantly more confirmed bookings compared to cancellations. Meal Plan 3 has fewer bookings, and cancellations are minimal for those with meal plans.
6. **Required Car Parking Space vs. Booking Status:**
    - Bookings with no car parking required are more common, showing a higher confirmation rate. When parking is required, cancellations are more frequent, particularly for Room Type 3, suggesting a potential issue with parking availability or guest preferences.
7. **Room Type Reserved vs. Booking Status:**
    - Room Type 1 has the highest booking volumes with a higher confirmation rate. Room Type 4 also shows good confirmation rates, while other room types like Room Type 2 show moderate bookings but with higher cancellations. Room Type 3 is the least popular and has the highest cancellation rate.
8. **Lead Time vs. Booking Status:**
    - Longer lead times are associated with higher cancellation rates. Canceled bookings have a higher median lead time (around 120 days), whereas confirmed bookings have a lower median lead time (around 50 days), indicating that longer-term bookings are more prone to cancellations.

9. **Arrival Month vs. Booking Status:**

   - October and September experience the highest number of bookings with more confirmations than cancellations. The months after October show a decline in bookings and cancellations, with January seeing the fewest cancellations.

10. **Market Segment Type vs. Booking Status:**

    - Online bookings dominate with a higher number of confirmed bookings. The Offline segment also has a high confirmation rate but fewer bookings overall. Corporate and complementary segments show higher cancellation rates and account for a smaller proportion of bookings.

11. **Repeated Guest vs. Booking Status:**

    - Repeated guests have a higher proportion of confirmed bookings compared to cancellations. New guests represent a larger portion of total bookings, but they are more likely to cancel compared to repeated guests.

12. **Average Price per Room vs. Booking Status:**

    - Canceled bookings have higher average room prices, especially in cases of high outliers, while confirmed bookings show lower price variability. This suggests that higher prices might be linked to cancellations, possibly due to changes in guest plans.

13. **Number of Special Requests vs. Booking Status:**

    - Guests who make no special requests have the highest number of confirmed bookings. As the number of special requests increases, cancellations slightly decrease, suggesting that guests with more specific requests are more committed to their bookings.

General Insights:

1. **Seasonality:**

   - There is a noticeable peak in bookings during the months of October and September, with October being the busiest month. This seasonal spike indicates that these months may require additional resources, staff, and targeted marketing efforts. A decline in bookings occurs towards the year-end, with January seeing the lowest number of arrivals.

2. **Booking Channels:**

   - Online bookings dominate the booking landscape, accounting for a substantial share (64% of total bookings). However, they also exhibit a higher cancellation rate, indicating that while the volume is high, there may be issues with booking commitment. Offline bookings and corporate bookings, though smaller in number, have higher confirmation rates and are more reliable, suggesting a more stable customer base for these channels.

3. **Customer Behavior:**

   - Guests with higher lead times are more likely to cancel their bookings, as longer lead times are associated with higher cancellation rates. Additionally, complementary bookings show the highest cancellation rates, indicating that guests in this category are less committed. Offering targeted incentives or cancellation policies may help reduce cancellations, especially for these segments.

4. **Guest Retention:**

   - Repeated guests form a critical part of the hotel's customer base. They are significantly less likely to cancel their bookings compared to new guests, demonstrating the value of customer loyalty. By focusing on retaining repeated guests through loyalty programs and personalized experiences, the hotel can reduce cancellations and improve revenue stability.

5. **Room Demand:**

   - Room Type 1 is in the highest demand, with confirmed bookings significantly outnumbering cancellations. Other room types, like Room Type 4, show good demand as well but with higher cancellation rates. Certain market segments, such as online bookings, show consistent demand for specific room types. This insight suggests the potential for dynamic pricing strategies that can optimize room availability and pricing across various market segments. Tailored promotions could also help boost demand for underperforming room types.

2.Data preprocessing:

2.1.Missing Value Treatment

Observation:

No columns in the dataset have missing values. All features have a count of zero for missing values.

Rationale for Missing Value Treatment:

1. **No Missing Data**:
   - Since no missing values exist, there is no need for imputation or removal of records at this stage.
2. **Verification**:
   - It is essential to validate the absence of missing data to ensure data completeness and accuracy. This ensures no errors arise from unobserved data gaps during subsequent analysis.
3. **Future Considerations**:
   - If any missing values are introduced during feature engineering or derived computations, appropriate handling techniques such as mean/median imputation, mode imputation, or advanced methods like KNN imputation may be applied based on the nature of the data.

No further action is required for missing value treatment at this stage since the dataset is already complete.

2.2.Outlier Treatment:

Outliers of numerical columns:

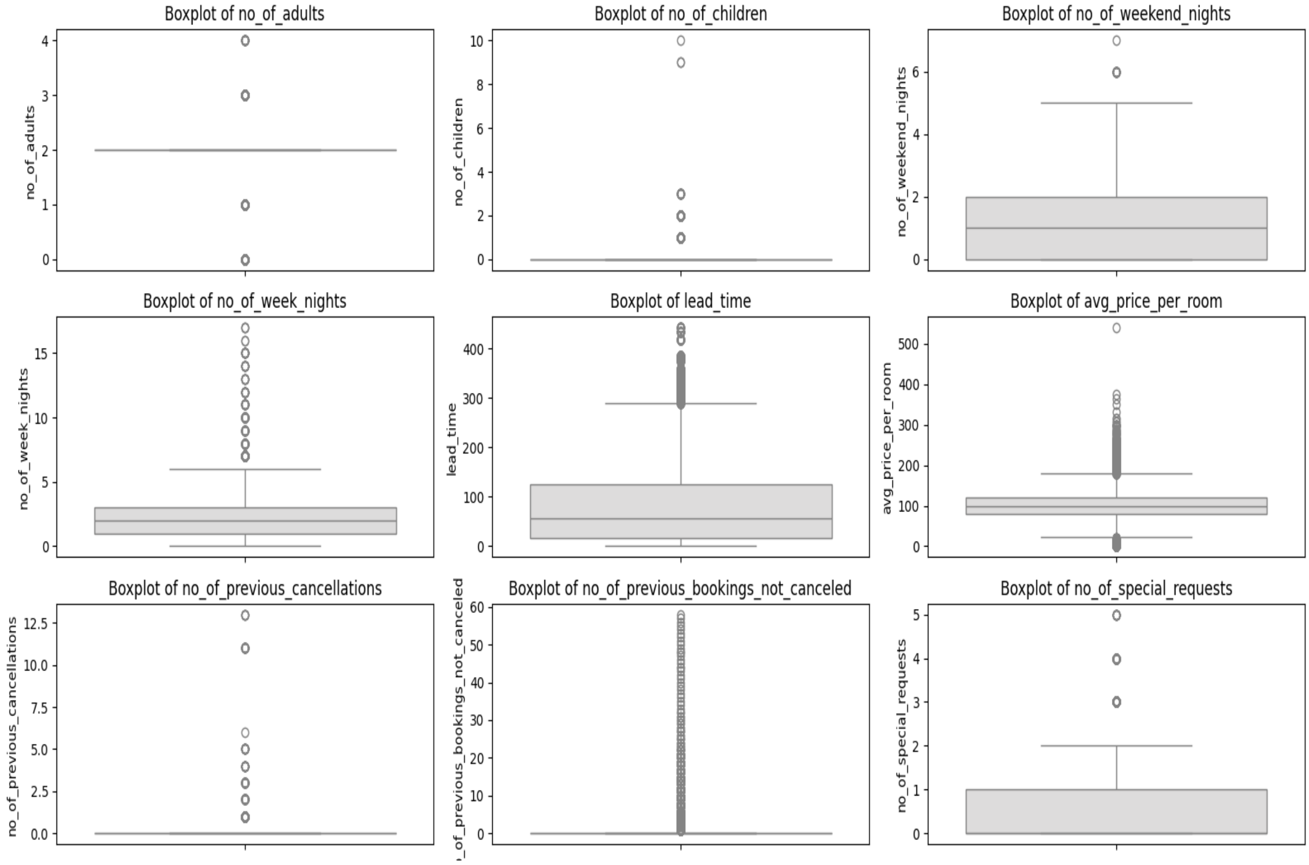

Figure 36. Outliers of numerical columns

In the dataset, outliers were initially identified using the interquartile range (IQR) method, which helps detect values that lie significantly outside the typical range of data. Outliers were found in several key columns:

Column	Outliers Before Treatment
no_of_adults	10,167
no_of_children	2,698
no_of_weekend_nights	21
no_of_week_nights	324
lead_time	1,331
no_of_previous_cancellations	338
no_of_previous_bookings_not_canceled	812
avg_price_per_room	1,696
no_of_special_requests	761

Table 1. Outlier Counts Before Treatment for Hotel Booking Data

These outliers were either much smaller or larger than expected, falling outside the range of 1.5 times the IQR from the first and third quartiles. Because outliers can distort statistical analysis and influence the results of a linear regression model, it was essential to address them appropriately.

The outliers were treated using an IQR-based capping method, where values exceeding the upper and lower whiskers were adjusted to the respective limits. This approach ensures that extreme values are brought within a reasonable range without entirely removing them, preserving the overall structure of the data.

After performing this outlier treatment, the dataset was reevaluated, and no outliers remained in any of the columns. This thorough approach ensures the data is now clean, allowing for more reliable analysis and predictive modeling without skewing results due to extreme values.

2.3.Feature Engineering

Feature engineering transforms raw data into a format suitable for machine learning models by creating meaningful features. In this step, categorical variables were converted into numerical representations using One-Hot Encoding, and boolean columns were handled appropriately. One-hot encoding was applied to the variables: type_of_meal_plan, room_type_reserved, market_segment_type, and booking_status. This process created binary columns for each category within these variables, except for the first category, which was excluded to avoid multicollinearity. For example, type_of_meal_plan generated columns such as type_of_meal_plan_Meal Plan 2, type_of_meal_plan_Meal Plan 3, and type_of_meal_plan_Not Selected. This transformation ensures that the model interprets these categories accurately without introducing bias from numerical ranking.

Additionally, boolean columns were converted to integer values (0 and 1) for better compatibility with machine learning algorithms. As a result, the dataset now contains numerical columns for features like the number of adults, children, weekend nights, and weekday nights, alongside binary columns for categorical variables. This ensures that the data is clean, standardized, and ready for modeling. With these transformations, the dataset not only retains its interpretability but also provides a structured input that enhances the accuracy and efficiency of downstream predictive models.

2.4. Data Scaling

In the Data Scaling process, we focused on scaling the numerical data to ensure uniformity in the dataset. Scaling is a critical step because numerical features often have different ranges and magnitudes, which can negatively affect machine learning models, especially those sensitive to the scale of input features, such as gradient descent-based algorithms and distance-based models like k-Nearest Neighbors.

We applied Min-Max Scaling to all numerical columns using the MinMaxScaler, transforming their values to lie within the range [0, 1]. This approach preserves the distribution and relationships between features while bringing them to a comparable scale. The numerical columns subjected to scaling include no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, lead_time, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, avg_price_per_room, and no_of_special_requests.

Rationale:

1. **Uniform Range:** Min-Max Scaling standardizes the feature values to a consistent range, preventing features with larger magnitudes from dominating the learning process.
2. **Model Efficiency:** Models like logistic regression and neural networks converge faster and perform better when input features are scaled.
3. **Preserving Relationships:** Unlike standardization, Min-Max Scaling does not distort the original distribution or relationships among features, which is advantageous when maintaining interpretability.
4. **Optimal Performance:** Scaling ensures optimal model performance by balancing the influence of all numerical features during training.

This scaled dataset is now ready for robust and efficient model training, enabling a fair representation of all numerical features in predictive modeling.

2.5. Train-Test Split

The train-test split is a fundamental step in data preprocessing that involves dividing the dataset into two subsets: one for training the machine learning model and another for testing its performance. In this case, a 30-70 split is used, where 30% of the data is allocated for testing and 70% is reserved for training.

The training set is used to teach the model, enabling it to learn the underlying patterns and relationships between the features and the target variable. The target variable in this case is booking_status, which represents whether a booking is confirmed (1) or canceled (0). The remaining 30% of the data is kept aside as the testing set, which is used to evaluate the model's performance on unseen data.

This 30-70 split strikes a balance between providing enough data for the model to learn effectively while still leaving a substantial portion for testing. The primary goal is to ensure that the model is generalizable and not overfitted to the training data. Testing the model on data it has not seen during training helps to simulate real-world performance and provides a reliable estimate of how the model will behave when applied to new data. By using this approach, we can assess key performance metrics such as accuracy, precision, recall, and F1-score, ensuring the model is ready for deployment in practical applications.

3.Model building

3.1. Defining the Optimal Evaluation Metric for Model Performance:

we identified key metrics to assess model performance, ensuring a comprehensive evaluation of their suitability for predicting the target variable, booking_status. Metrics such as accuracy, precision, recall, and F1-score were used to gain a detailed understanding of each model's strengths and weaknesses. Additionally, the confusion matrix was analyzed to observe false positives and false negatives, particularly critical for understanding cancellation patterns. For instance, recall was emphasized for its ability to measure the proportion of canceled bookings that were correctly predicted. This decision ensured a focus on reducing missed cancellations, which are critical for operational planning in the hospitality industry. By using a balanced set of metrics, we ensured a thorough evaluation of the models’ ability to generalize and meet the project’s requirements.

3.2. Logistic Regression (statsmodels):

Logistic Regression was the first model implemented, utilizing statsmodels for building and evaluating its performance. The model achieved an accuracy of 80.26%, reflecting a solid performance in classifying bookings. Metrics such as precision (83.23%) and recall (88.68%) indicated a slight bias towards predicting confirmed bookings more effectively than cancellations. This was further corroborated by the F1-score of 85.87%, which balanced the trade-off between precision and recall. The confusion matrix revealed a reasonably good balance in predictions, although there were some notable misclassifications, particularly for canceled bookings. This model’s performance indicated its capability to serve as a reliable baseline for comparison against more complex algorithms. However, it highlighted room for improvement, especially in accurately predicting cancellations, which is a key business goal.

3.3. KNN Classifier (sklearn):

The K-Nearest Neighbors (KNN) classifier was implemented using scikit-learn with the default hyperparameter value of k=5k=5k=5. Post-evaluation, the model displayed strong performance metrics, with an accuracy of 80.43%, precision of 83.35%, recall of 88.81%, and F1-score of 85.99%. These metrics slightly outperformed Logistic Regression, showcasing KNN’s ability to capture non-linear relationships in the dataset. The confusion matrix indicated a notable improvement in classifying canceled bookings, with fewer false negatives compared to Logistic Regression. However, KNN's computational intensity and sensitivity to data scaling were noted as potential drawbacks, which might necessitate further hyperparameter tuning or optimization. Despite these limitations, KNN proved to be a competitive model, offering insights into data relationships that were not as apparent with linear models.

3.4. Naive-Bayes Classifier (sklearn)

The Naive-Bayes model, built using scikit-learn, struggled with the dataset, resulting in poor performance. The accuracy was 39.82%, with a precision of 91.68%, but a dismally low recall of 12.13% and F1-score of 21.43%. These results were due to the model's reliance on the assumption of feature independence, which was not valid in this dataset. The confusion matrix revealed significant misclassifications, particularly for canceled bookings, as the model failed to effectively differentiate between the two classes. While Naive-Bayes is often effective for simpler datasets, its inability to handle complex interactions between features made it unsuitable for this task. This outcome emphasized the importance of understanding dataset characteristics before selecting a model and highlighted the need to explore more sophisticated algorithms.

3.5. Decision Tree Classifier (sklearn)

The Decision Tree Classifier, implemented using scikit-learn, emerged as the top-performing model, with an accuracy of 86.58%, precision of 90.60%, recall of 89.43%, and F1-score of 90.01%. The confusion matrix demonstrated significantly fewer misclassifications compared to other models, particularly for canceled bookings. Decision Trees excelled due to their ability to capture non-linear patterns and interactions between features, making them highly effective for this dataset. Additionally, the interpretability of the model allowed for insights into key factors influencing booking status, such as lead time, market segment, and special requests. This strong performance, coupled with its flexibility and clarity, highlighted the Decision Tree Classifier as the most suitable model for predicting booking_status at this stage.

3.6. Analyzing and Interpreting Model Performance Metrics

Overall, the Decision Tree Classifier stood out as the best-performing model, with superior metrics across accuracy, precision, recall, and F1-score. Its ability to handle non-linear interactions and provide interpretable results made it highly effective for this dataset. The KNN Classifier performed competitively, offering an alternative approach that captured data relationships effectively. Logistic Regression served as a reliable baseline, delivering reasonable results while being simpler and faster to implement. Conversely, the Naive-Bayes Classifier demonstrated poor performance due to its inability to handle feature dependencies. These findings underscored the importance of model selection and highlighted Decision Trees as the optimal choice for this problem. Further steps could include hyperparameter tuning and ensemble techniques to enhance model robustness and performance.

4. Model Performance Improvement

4.1. Optimizing Model Performance

The initial focus was on identifying areas for improvement in each model. We prioritized handling issues like overfitting, feature selection, and hyperparameter tuning to achieve better generalization. The overarching goal was to refine models to balance precision, recall, and overall accuracy, ensuring robust predictions in practical scenarios.

4.2. Tune Logistic Regression:

Logistic Regression is a fundamental algorithm for binary classification problems, and its performance was systematically improved by addressing multicollinearity, removing statistically insignificant features, and optimizing the decision threshold. To handle multicollinearity, we conducted a Variance Inflation Factor (VIF) analysis, removing features with VIF > 10 to reduce redundancy, stabilize the model, and improve the interpretability of the coefficients. This helped ensure that the model's predictors were meaningful and not highly correlated with each other. To further streamline the model, we also identified and removed features with p-values greater than 0.05, ensuring that only statistically significant variables contributed to the model. This process not only improved the model’s efficiency but also minimized overfitting, making it more generalizable.

A key improvement step was threshold optimization, where we used the Receiver Operating Characteristic (ROC) curve to assess how the model performed at different thresholds. The ROC curve plotted the True Positive Rate (TPR or Recall) against the False Positive Rate (FPR) at various threshold levels. The curve showed the trade-off between sensitivity and specificity, providing insights into the model’s discriminatory ability. The Area Under the Curve (AUC) score was calculated to be 0.87, indicating strong performance. By using Youden's J statistic, the optimal threshold was determined to be 0.6813, as this point maximized the difference between TPR and FPR.

Below is the ROC curve, illustrating this trade-off and highlighting the performance of the model at different thresholds:

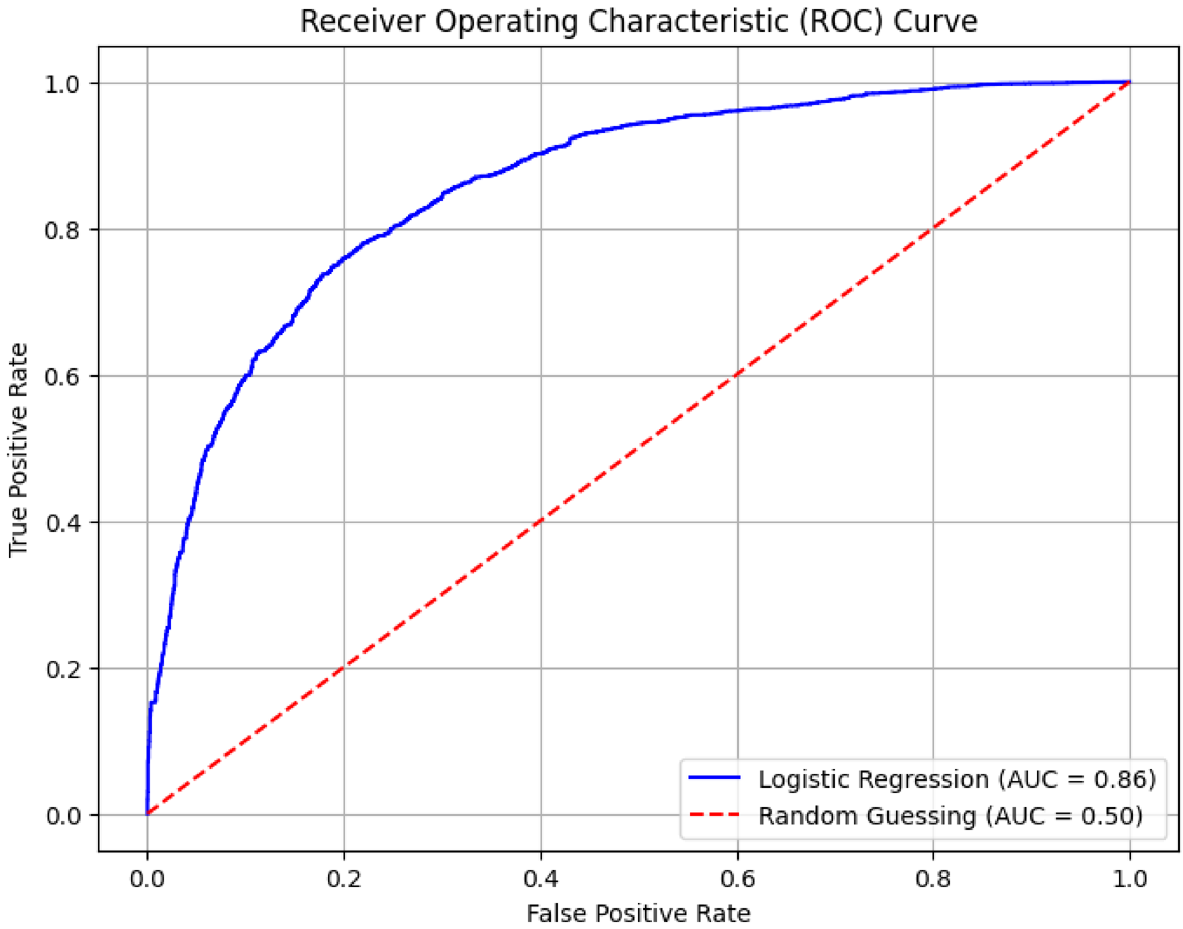

Figure 37. Receiver Operating Characteristic (ROC) Curve

While the optimal threshold improved precision (88.15%) and reduced false positives, it did so at the expense of recall (77.88%). In contrast, using the default threshold of 0.5 resulted in a better balance between precision (83.16%) and recall (88.57%). Based on the specific business needs, we chose to retain the default threshold, as it offered a better trade-off between correctly identifying canceled bookings (high recall) and minimizing false positives (good precision).

After these refinements, the Logistic Regression model achieved an accuracy of 80.14%, an F1-score of 85.78%, and a confusion matrix that reflected a balanced performance between false positives and false negatives. Addressing multicollinearity, selecting statistically significant features, and optimizing the threshold helped make the model more reliable and robust for predicting booking_status. The improvements ensured the model struck the right balance between precision and recall, making it a strong choice for the given application.

4.3.Tune KNN Classifier:

The K-Nearest Neighbors (KNN) classifier was tuned to enhance its performance by optimizing the number of neighbors (k). Initially, the model used a default value of k, but to improve its accuracy, the number of neighbors was adjusted between 3 and 20 using cross-validation. After extensive testing, the optimal value of k was found to be 9, which significantly improved the model's ability to classify both confirmed and canceled bookings.

Before tuning, the KNN model yielded an accuracy of 80.43%, precision of 83.35%, recall of 88.81%, and an F1-score of 85.99%. The confusion matrix at this stage showed that the model had a reasonable balance between false positives and false negatives, but there was still room for improvement. After tuning, the model's performance improved substantially, reaching an accuracy of 85.32%, precision of 87.65%, recall of 91.13%, and an F1-score of 89.36%. The confusion matrix after tuning indicated a reduction in false negatives, highlighting that the model became more reliable in identifying canceled bookings.

The performance improvement suggests that the optimal choice of k allowed the model to better capture the underlying patterns in the data, especially in distinguishing between confirmed and canceled bookings. This enhancement makes the KNN classifier a more effective tool for predicting booking status, ensuring it minimizes misclassifications, especially in the critical recall aspect. By reducing false negatives, the tuned KNN model becomes more suitable for applications where the goal is to accurately identify canceled bookings.

4.4.Tune Decision Tree Classifier:

The Decision Tree Classifier was tuned using pre-pruning techniques to improve generalization and avoid overfitting. The maximum depth of the tree was limited to 5, based on initial testing, which prevented the model from becoming overly complex. By restricting the depth, the model focused on the most critical decision points, reducing its tendency to overfit the training data and ensuring it generalized better to unseen data. Various tree depths were tested, and a balanced depth of 5 was determined to provide optimal performance, effectively capturing key patterns in the data while maintaining interpretability.

Before tuning, the Decision Tree model showed solid performance with an accuracy of 86.58%, precision of 90.60%, recall of 89.43%, and an F1-score of 90.01%. The confusion matrix at this stage revealed a strong model performance with relatively few misclassifications. After applying pre-pruning, the model's accuracy decreased slightly to 83.07%, but its precision improved to 85.28%, and recall increased to 90.60%. The F1-score also rose to 87.86%, indicating the model’s ability to balance false positives and false negatives better.

While pre-pruning resulted in a slight decrease in accuracy, it significantly improved the model’s generalization and interpretability. By reducing complexity, the Decision Tree became a more reliable tool for making predictions, especially in terms of recall. The pruned model's improvements demonstrate its effectiveness in handling the trade-off between model simplicity and predictive power, offering a good balance for real-world applications where both precision and recall are crucial.

4.5.Evaluation of Tuned Model Performance Across Metrics:

A comparative evaluation of the performance metrics across the tuned models—Logistic Regression, KNN Classifier, and Decision Tree Classifier—revealed distinct strengths for each model. The KNN Classifier emerged as the top performer with the highest accuracy (85.32%), precision (87.65%), recall (91.13%), and F1-score (89.36%). These results indicate that KNN was highly effective in reducing false negatives, making it particularly suitable for applications where capturing positive instances is critical. The model’s ability to maintain a high recall rate while achieving competitive precision shows its strength in handling imbalanced datasets.

The Decision Tree Classifier, which was pre-pruned to prevent overfitting, also demonstrated strong performance with accuracy of 83.07%, precision of 85.28%, recall of 90.60%, and an F1-score of 87.86%. While it didn’t outperform KNN in all metrics, it provided an excellent balance of predictive power and model interpretability, making it a reliable choice in situations where understanding decision-making is as important as performance.

The Logistic Regression model, while slightly behind in most metrics, still provided reliable results, particularly excelling in precision with 83.16%. This makes it a strong option for cases where minimizing false positives is a priority, such as when incorrect positive predictions could have significant consequences.

In conclusion, the KNN Classifier stands out as the best performing model overall, but the Decision Tree offers valuable interpretability, and Logistic Regression remains a solid option for tasks requiring high precision. Each model’s performance reflects the trade-offs between accuracy, recall, and interpretability, ensuring that the best model can be selected based on the specific needs of the deployment scenario.

5. Model Performance Comparison and Final Model Selection

5.1. Comparative Analysis of Model Performance:

Evaluating model performance involves analyzing metrics such as accuracy, precision, recall, and F1-score. Each metric reflects a different aspect of the model's ability to handle data and make predictions. For this business scenario, the balance between false positives and false negatives is critical, making F1-score the deciding factor. Below, the performance of three models—Logistic Regression, KNN Classifier, and Decision Tree Classifier—was assessed both before and after tuning.

The Logistic Regression model displayed stable performance and high precision, excelling in minimizing false positives. The KNN Classifier exhibited improved results post-tuning, with notable gains in recall and F1-score. However, the Decision Tree Classifier (Default Hyperparameters) surpassed the other models, achieving the highest F1-score, indicating the best balance between precision and recall. This section details the comparative analysis for each model, highlighting their strengths and trade-offs.

Logistic Regression

Pre-Tuning (Default Threshold 0.5)

` Before tuning, Logistic Regression demonstrated satisfactory predictive performance, achieving an accuracy of 80.14%, precision of 83.16%, recall of 88.57%, and an F1-score of 85.78%. The confusion matrix showed 841 false negatives and 1,320 false positives, indicating high recall but highlighting room for improvement in balance between precision and recall.

Post-Tuning (Optimal Threshold 0.6813)

After optimizing the threshold, the model's accuracy dropped to 77.96%, with precision improving to 88.15%. However, recall declined to 77.88%, leading to a lower F1-score of 82.70%. While precision improved by reducing false positives, this came at the cost of more false negatives, making the model less balanced.

KNN Classifier

Before Tuning

The initial KNN model performed well, achieving an accuracy of 80.43%, precision of 83.35%, recall of 88.81%, and an F1-score of 85.99%. The confusion matrix revealed a significant number of false negatives, emphasizing the need for parameter optimization to improve recall and overall balance.

After Tuning (k = 9)

Tuning the k parameter to 9 substantially enhanced the model's performance. Accuracy increased to 85.32%, precision to 87.65%, recall to 91.13%, and F1-score to 89.36%. This improvement was reflected in fewer false negatives, making the model more robust and reliable for the business scenario.

Decision Tree Classifier

Default Hyperparameters

Using default hyperparameters, the Decision Tree Classifier achieved excellent metrics, including an accuracy of 86.57%, precision of 90.49%, recall of 89.53%, and an F1-score of 90.01%. The confusion matrix indicated a balanced trade-off between false positives and false negatives, making this the best-performing model overall.

Pre-Pruned

Pre-pruning reduced the tree depth, slightly trading off performance for simplicity. The pruned model had an accuracy of 83.07%, precision of 85.28%, recall of 90.60%, and an F1-score of 87.86%. While easier to interpret, this version sacrificed predictive strength compared to the default configuration.

5.2. Rationale for Selecting the Best Model

The Decision Tree Classifier with default hyperparameters emerged as the final model for this business problem due to its superior performance and interpretability. With the highest F1-score (90.01%), it demonstrated the best balance between precision and recall, making it ideal for addressing the comparable costs of false positives and false negatives.

In this business scenario, both error types carry significant consequences. False Negatives, where a canceled booking is incorrectly predicted as not canceled, can lead to unutilized resources due to halted bookings. Conversely, False Positives, predicting a booking will be canceled when it is not, may result in overbooking, operational inefficiencies, and customer dissatisfaction. Given the comparable impact of these errors, achieving a balanced trade-off is crucial, and the F1-score, which accounts for both precision and recall, is the most relevant metric for model evaluation.

While the Logistic Regression model was reliable, its post-tuning F1-score (82.70%) fell significantly due to the trade-off between precision and recall during threshold optimization. The KNN Classifier, although demonstrating strong post-tuning performance, still lagged slightly behind the Decision Tree in overall balance and metrics.

Beyond metrics, the Decision Tree’s interpretability enhances its appeal. It provides clear insights into the factors influencing predictions, aligning with the business's requirement for transparent and explainable AI solutions. By combining strong performance with actionable insights, the Decision Tree Classifier is the optimal choice for deployment.

6.Actionable Insights & Recommendations

6.1. Actionable Strategies for Mitigating Booking Cancellations and Enhancing Business Efficiency:

Based on the analysis and model evaluation conducted in this project, the following actionable insights and recommendations can guide business decision-making and future strategies:

1. Focus on the Impact of Booking Cancellations

Both false positives (predicting cancellations that don’t occur) and false negatives (failing to predict cancellations that do occur) have substantial financial and operational consequences. To mitigate these risks:

- Implement **dynamic booking strategies** such as overbooking policies, informed by predictive models, to minimize resource wastage from false negatives.
- For false positives, adopt measures like **customer confirmation campaigns** or incentivized booking retention to reduce cancellations.

2. Leverage the Decision Tree Classifier for Deployment

The Decision Tree Classifier, with its high F1-score (90.01%), is the most reliable and interpretable model for predicting booking cancellations. Its balance between precision and recall makes it ideal for this scenario, ensuring minimal operational disruptions while maintaining customer satisfaction.

- **Deploy the Decision Tree model** as part of a real-time booking management system to predict potential cancellations accurately.
- Use the model’s insights to identify key predictors (e.g., booking timing, customer type) and design targeted interventions.

3. Optimize Data Collection and Preprocessing

Data quality and preparation played a critical role in achieving strong model performance. Moving forward:

- Enhance the dataset by collecting more granular features (e.g., customer booking behavior trends, weather conditions).
- Address missing or inconsistent data to maintain high data integrity for future modeling efforts.

4. Continuously Monitor Model Performance

Model performance can degrade over time due to changes in booking behavior, seasonality, or external factors. Therefore:

- Establish a **regular model retraining schedule** using updated data to ensure continued accuracy and relevance.
- Monitor metrics such as precision, recall, and F1-score post-deployment to assess the model's effectiveness.

5. Integrate Insights into Business Strategy

Use the model's predictions and feature importance to inform broader business strategies:

- Target customers more likely to cancel with **personalized retention campaigns** or flexible cancellation policies.
- Identify segments with low cancellation risk and offer tailored upselling opportunities to maximize revenue.

6. Extend Analysis to Related Applications

The predictive framework developed in this project can be adapted to other areas of the business:

- Use similar models for **customer churn prediction**, helping to retain valuable customers.
- Implement predictive analytics for **resource allocation and demand forecasting**, improving operational efficiency.

By deploying the optimized Decision Tree Classifier and acting on these recommendations, the business can significantly reduce the impact of booking cancellations while enhancing overall operational and customer experience outcomes.