MACHINE LEARNING-1 PROJECT

DSBA

Tasmiyasana2@gmail.com

DATE: 01/08/2025

Table of Contents

List of Figures

List of Tables

1 PROBLEM STATEMENT

A large number of hotel reservations are cancelled due to cancellations or no-shows. Changes in plans, scheduling issues, and other factors commonly lead to cancellations. The availability of free or low-cost cancellation options makes it easier for guests, but this creates challenges for hotels, leading to potential revenue loss. These losses are particularly significant for last-minute cancellations.

New technologies that connect online booking channels have greatly changed customer booking habits and options. This increases the difficulty for hotels in managing cancellations, which are no longer limited to traditional booking methods and guest characteristics.

Booking cancellations affect a hotel in several ways:

1. The hotel loses money when it cannot resell the room.
2. It faces higher distribution channel costs due to increased commissions or paid promotions to sell these rooms.
3. The hotel may have to lower prices at the last minute to resell a room, which leads to a smaller profit margin.
4. Additional human resources are needed to accommodate the guests.

1.1 Objective

The rising number of cancellations requires a Machine Learning solution that can predict which bookings are likely to be cancelled. INN Hotels Group operates a chain of hotels in Portugal and is dealing with a significant number of booking cancellations. They have contacted our firm for data-driven solutions. As a data scientist, we need to analyse the provided data to identify the factors that greatly influence booking cancellations. We will also build a predictive model to forecast which bookings will be cancelled in advance. Finally, our work will support the development of effective policies for cancellations and refunds.

1.2 Data Description

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

Data Dictionary:

- Booking_ID: the unique identifier of each booking
- no_of_adults: Number of adults
- no_of_children: Number of Children
- no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- no_of_week_nights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
- type_of_meal_plan: Type of meal plan booked by the customer:
    - Not Selected - No meal plan selected
    - Meal Plan 1 - Breakfast
    - Meal Plan 2 - Half board (breakfast and one other meal)
    - Meal Plan 3 - Full board (breakfast, lunch, and dinner)
- required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)

- room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not cancelled 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 cancelled or not.

1.3 Key Questions

Q1. What are the busiest months in the hotel?
Q2. Which market segment do most of the guests come from?
Q3. Hotel rates are dynamic and change according to demand and customer demographics. Q3. What are the differences in room prices in different market segments?
Q4. What percentage of bookings are cancelled?
Q5. Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?
Q6. Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?

1.4 Data Overview

We will upload the dataset and check it by using head() and tail() to see first and last 5 rows, we will also check data types using info() and shape attribute in pandas to identify the dataset total numbers or rows and columns.

Top & Bottom 5 Rows

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved
0	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1
1	INN00002	2	0	2	3	Not Selected	0	Room_Type 1
2	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1
3	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1
4	INN00005	2	0	1	1	Not Selected	0	Room_Type 1

lead_time	arrival_year	arrival_month	arrival_date	market_segment_type	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled
224	2017	10	2	Offline	0	0	0
5	2018	11	6	Online	0	0	0
1	2018	2	28	Online	0	0	0
211	2018	5	20	Online	0	0	0
48	2018	4	11	Online	0	0	0

t_segment_type	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
Online	0	0	0	167.80	1	Not_Canceled
Online	0	0	0	90.95	2	Canceled
Online	0	0	0	98.39	2	Not_Canceled
Online	0	0	0	94.50	0	Canceled
Offline	0	0	0	161.67	0	Not_Canceled

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved
36270	INN36271	3	0	2	6	Meal Plan 1	0	Room_Type 4
36271	INN36272	2	0	1	3	Meal Plan 1	0	Room_Type 1
36272	INN36273	2	0	2	6	Meal Plan 1	0	Room_Type 1
36273	INN36274	2	0	0	3	Not Selected	0	Room_Type 1
36274	INN36275	2	0	1	2	Meal Plan 1	0	Room_Type 1

Table 1. Top & Bottom 5 rows.

Shape

There are **36275 rows** & **19 columns** in our dataset.

Data Types

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36275 entries, 0 to 36274
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
dtypes: float64(1), int64(13), object(5)
memory usage: 5.3+ MB
```

Table 2. Data Types.

Most columns are numeric (int64) and avg_price_per_room is of float type, which makes them suitable for analysis and modelling. **Five columns are of object type** and may need encoding: type_of_meal_plan, room_type_reserved, market_segment_type, booking_status, Booking_ID.

The **target variable is booking_status**. It shows whether a booking was cancelled or not. Right now, it is of **object type** and needs to be changed to a binary numeric format for modelling.

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

Table 3. Statistical Summary.

- **no_of_adults:** Most bookings are for **2 adults** (50th, 75th percentiles = 2.0). The minimum is **0 adults**, which may suggest an incorrect or test entry, while maximum no. of adults are 4.
- **no_of_children:** The average number is very low **(~0.1)**, and 75% of bookings include no children. However, a **maximum of 10** indicates some large family or group bookings.
- **no_of_weekend_nights:** The **median is 1** night, and 75% of bookings stay up to **2 weekend nights**. The **maximum is 7**, which may represent extended weekend stays or rare cases, while minimum stay is 0.
- **no_of_week_nights:** The median is **2 nights**, with most bookings covering **1 to 3 weeknights**. There are a few long stays, with a maximum of 17.
- **required_car_parking_space:**
  The mean is very low, around **0.03**, and all **quartiles are 0**. This shows that **few guests ask for parking**.
- **lead_time:**
  The average lead time is **85 days**, and there is a wide spread with a standard deviation of about **86**. Some bookings are made far in advance, with a maximum of **443 days**.
- **arrival_year:**
  Most bookings are from **2018**, as the 25th, 50th, and 75th percentiles are all 2018. The dataset **mainly covers bookings from 2018**.
- **arrival_month:**
  The median month for arrivals is **August**, with bookings occurring throughout the year, from **January to December**.
- **arrival_date:**
  Bookings are **evenly distributed** across the dates from 1 to 31, with a median on the **16th**.
- **repeated_guest:**
  The mean for repeat guests is very low, around **0.025**. This indicates that there are **few repeat customers**.

- **no_of_previous_cancellations:**
  Very few customers have cancelled before, with a mean of about 0.02. However, some customers have cancelled up to **13 times**, showing that there are **outliers or habitual cancelers**.
- **no_of_previous_bookings_not_canceled:**
  Most customers have **no past bookings that were not cancelled**, with the 25th, 50th, and 75th percentiles all at 0. The maximum is **58**, indicating a few loyal or repeat customers.
- **avg_price_per_room:**
  The median price per room is around **€99.45**, and there is a wide range that goes up to **€540**. Some entries have a price of **0**, which may mean missing or invalid pricing.
- **no_of_special_requests:**
  Most customers make **0 to 1** special requests. The **maximum is 5** requests, which could influence the likelihood of cancellations.

Unique Values in Categorical Columns

```
type_of_meal_plan:
 ['Meal Plan 1' 'Not Selected' 'Meal Plan 2' 'Meal Plan 3']

room_type_reserved:
 ['Room_Type 1' 'Room_Type 4' 'Room_Type 2' 'Room_Type 6' 'Room_Type 5
 'Room_Type 7' 'Room_Type 3']

market_segment_type:
 ['Offline' 'Online' 'Corporate' 'Aviation' 'Complementary']

booking_status:
 ['Not_Canceled' 'Canceled']
```

Figure 1. Unique Values.

type_of_meal_plan:

- Unique Values: 'Meal Plan 1', 'Not Selected', 'Meal Plan 2', 'Meal Plan 3'.
- The hotel provides **three meal plans** and a "Not Selected" choice.
- This indicates that not all customers choose meal options. **Meal preference may impact how they book**.

room_type_reserved:

- Unique Values: 'Room_Type 1' to 'Room_Type 7'.
- There are **seven different room** types, likely varying by price, size, or features.
- Further analysis can show if certain room types have higher cancellation rates.

market_segment_type:

- Unique Values: 'Offline', 'Online', 'Corporate', 'Aviation', 'Complementary'
- Guests **book through various channels**, with **'Online' and 'Offline'** likely being the most popular.
- Corporate and Aviation may involve business clients, while **'Complementary' could mean promotions or free stays**.

booking_status:

- Unique Values: 'Not_Canceled', 'Canceled'

- This is the **target variable** for prediction.
- The existence of both classes shows that it is **suitable for binary classification**.

Categorical Column Unique Value Counts

--- type_of_meal_plan ---

type_of_meal_plan	Count	Percentage (%)
Meal Plan 1	27835	76.73
Not Selected	5130	14.14
Meal Plan 2	3305	9.11
Meal Plan 3	5	0.01

--- room_type_reserved ---

room_type_reserved	Count	Percentage (%)
Room_Type 1	28130	77.55
Room_Type 4	6057	16.70
Room_Type 6	966	2.66
Room_Type 2	692	1.91
Room_Type 5	265	0.73
Room_Type 7	158	0.44
Room_Type 3	7	0.02

--- market_segment_type ---

market_segment_type	Count	Percentage (%)
Online	23214	63.99
Offline	10528	29.02
Corporate	2017	5.56
Complementary	391	1.08
Aviation	125	0.34

--- booking_status ---

booking_status	Count	Percentage (%)
Not_Canceled	24390	67.24
Canceled	11885	32.76

Table 4. Unique Value Counts.

type_of_meal_plan:

- An overwhelming majority of guests, almost **77%**, chose Meal Plan 1. This suggests that breakfast is a highly valued inclusion.
- Interestingly, about **1 in 7 guests, or 14%**, opted for **no meal plan** at all. This could indicate short stays or a focus on costs.
- Only **9% selected Meal Plan 2**, which offers partial board, while Meal Plan 3 (full board) has seen almost no use, with **just 5 bookings**. This might be due to either low demand or a lack of awareness.
- This distribution suggests that **Meal Plan 3 may need to be re-evaluated** or offered in a different way if it is not attracting guests.

room_type_reserved:

- The reservation trend **heavily favours Room_Type 1**, with nearly 4 out of every 5 bookings, or **78%**, made for this type. It likely represents a standard or budget-friendly option.
- **Room_Type 4** is in a distant second place at **17%**, while the remaining five room types together make up **less than 6% of all bookings**.
- **Room_Type 3** has only been **booked 7 times** out of 36,275 entries. This room might be obscure, miscategorized, or unavailable for most of the time.
- This skew suggests that the **room allocation or promotion strategy may need adjusting**, especially for underused room types.

market_segment_type:

- More than **6 out of 10 guests** book through the **online channel**, indicating a strong preference for digital bookings.

- **Offline bookings** still make a significant contribution at **29%**. This could come from walk-ins, agents, or traditional methods.
- The **corporate segment**, while much smaller at **5.5%**, could become an important revenue source if it includes repeat high-value clients.
- The **complementary and aviation segments** are quite small, making up **less than 2%** in total. They may represent niche or promotional stays. This could affect cancellation rates or room turnover but may not significantly impact revenue.

booking_status:

- About one-third, or **32.76%**, of all **bookings end with cancellations**, which is a notable proportion that supports the need for predictive modelling.
- The remaining **67.24%** proceed as planned. However, the high cancellation rate might point to issues such as **flexible refund policies, overbooking, or last-minute guest indecision**.
- The imbalance is significant enough to ensure that **both categories are well-represented**. This makes it suitable for classification tasks without requiring initial resampling.

2 EXPLORATORY DATA ANALYSIS

2.1 Univariate Analysis

Observations on lead_time

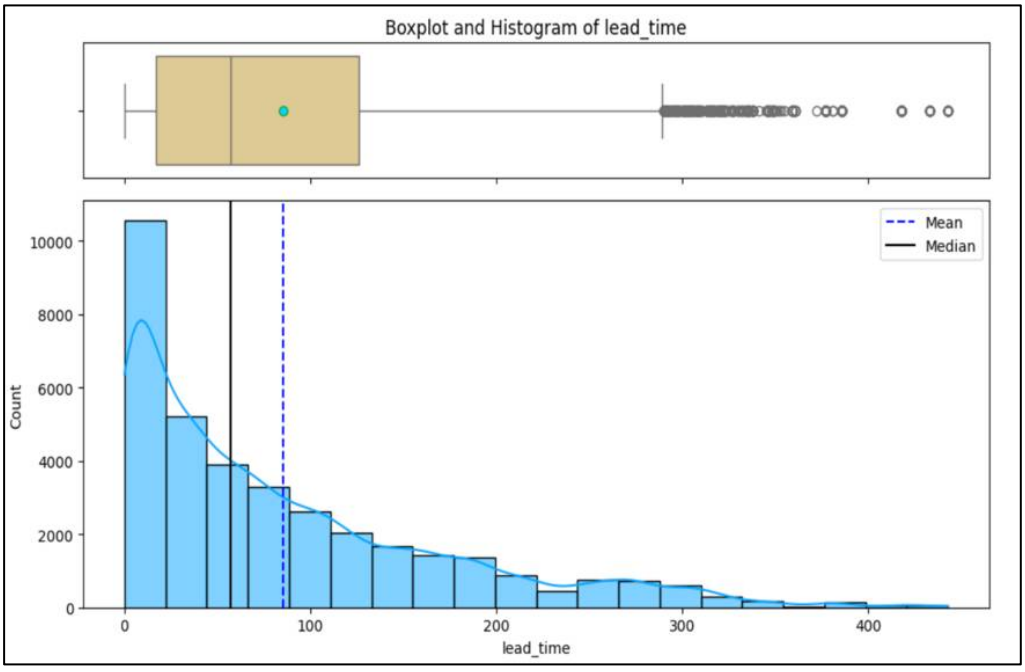


Figure 2. Distribution of lead_time.

- The data seems to be highly right skewed, with 0 as the highest lead time, suggesting same day booking and check-in by customer, due to last minute booking or walk-in customers.
- Box plot suggests average lead time is **~85 days**, with median of ~57 days, shows that 57% of the booking made falls between 0-57 days.

- Data also indicates that minimum lead time is 0 while max is around 443 days, indicating either customer checked in on the same day of booking or booked a year early (extreme values).
- We can also observe various significant outliers present in the data with lead time greater than 270 days.
- These outliers points towards booking were made in more than 9 months in advance by customers.
- The bookings which are made **>85 days** may have high cancel rate.

Observations on avg_price_per_room

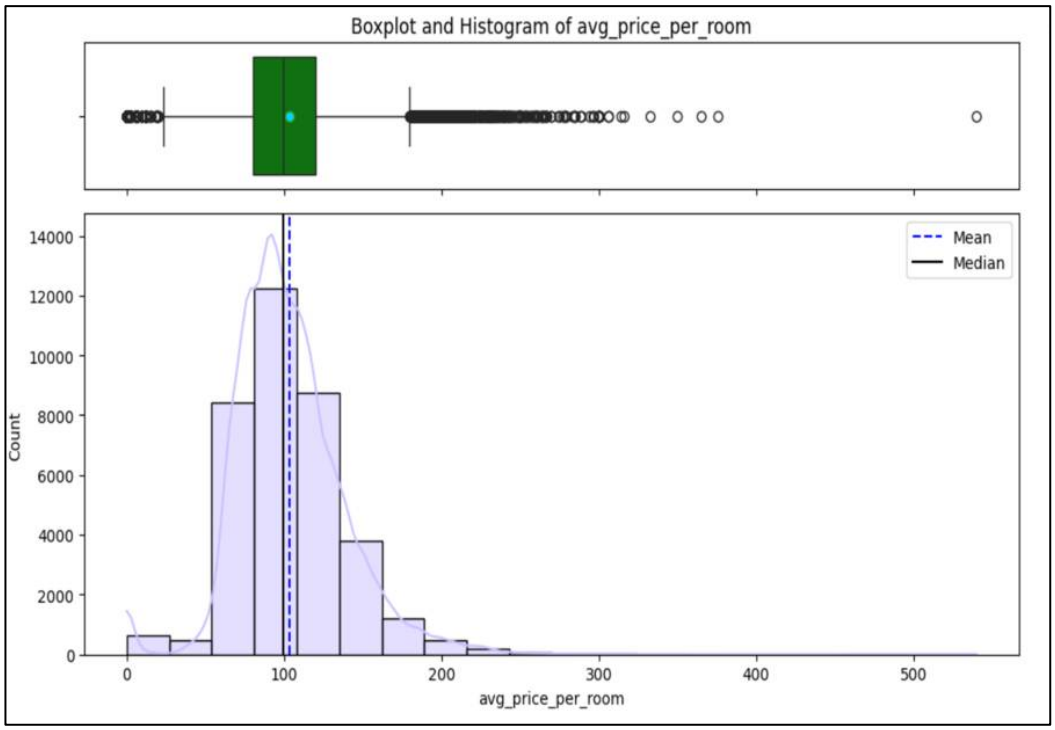


Figure 3. Distribution of avg_price_per_room.

- The plot shows data is slightly right skewed with peak around **100** and longer tail towards higher price side.
- Majority of data is clustered around **90-110** range, resembling a normal distribution but with minor right skewness.
- Median and median is ~100 with interquartile range ~90-110.

- Significant outliers present on both side of whiskers, lower end whiskers outliers suggesting high price range **>200/day** Euros due to premium rooms or data errors.
- Outliers on lower whiskers (down t0 ~30/day Euros), suggesting low price rooms with less facilities, or discounted price rooms. We need to investigate further if it is actual data or data inconsistencies.

Observations on no_of_children

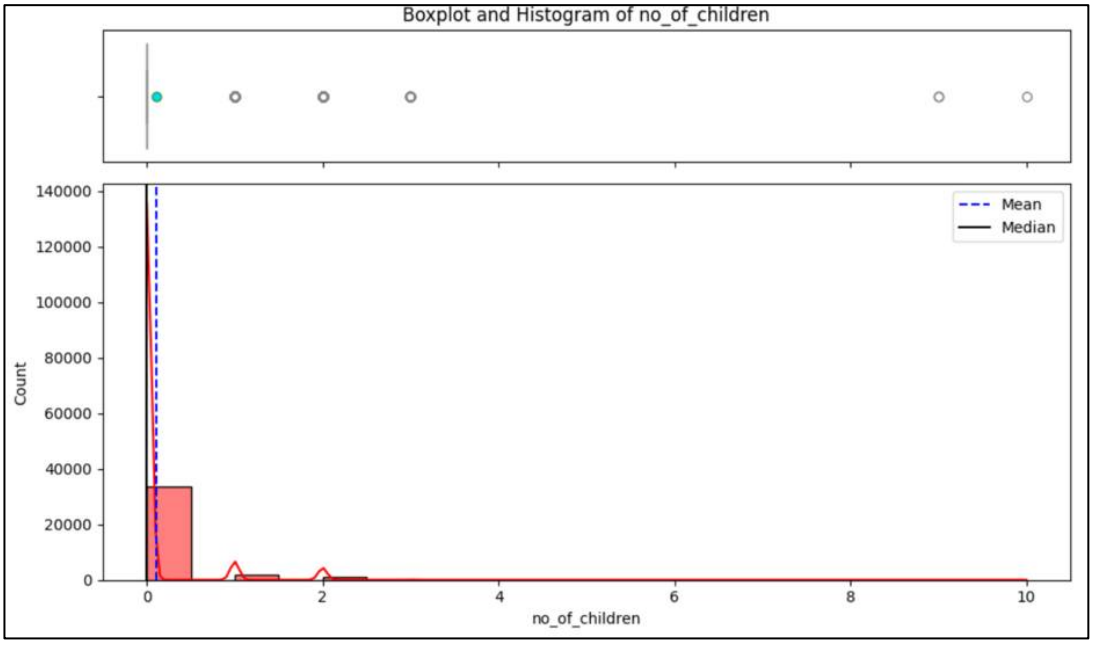


Figure 4. Distribution of no_of_children.

- The distribution is highly right skewed.
- Data is clustered around **0-3**, suggesting bookings mostly made by customers with no children or families with maximum **2-3 kids**.
- The average number is very **low (~0.1), and 75%** of bookings include no children, suggesting booking made by bachelors or couples with no children.

- The outliers are present on lower ends of box plot, maximum of 10 indicates some large family or group bookings.

Observations on no_of_adults

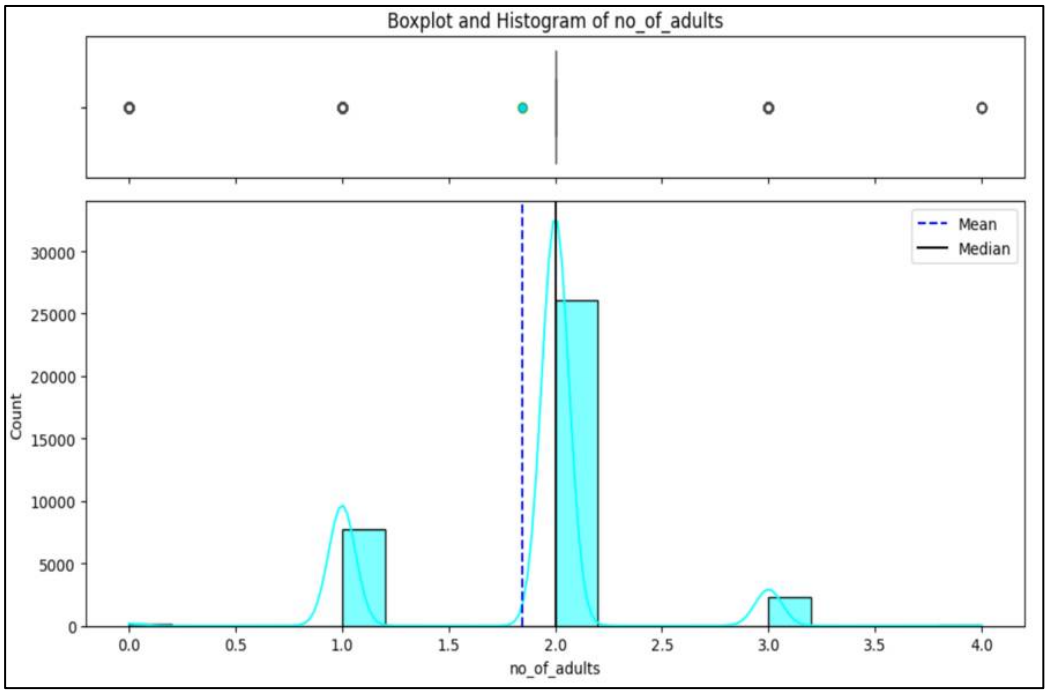


Figure 5. Distribution of no_of_adults.

- Plot suggests normal distribution.

- Most bookings are for **2 adults** (50th, 75th percentiles = 2.0) with nearly 25k visits.
- The minimum is 0 adults, which may suggest an incorrect or test entry and needs further investigation. **Maximum 4 adult** visited.
- They are couple of the outliers on the both sides of the whiskers. Upper end outliers are at 0 & 1 and upper end outliers are at 3 & 4, no adult indicating the data inconsistency while 1 adult might be actual data.

Observations on no_of_weekend_nights

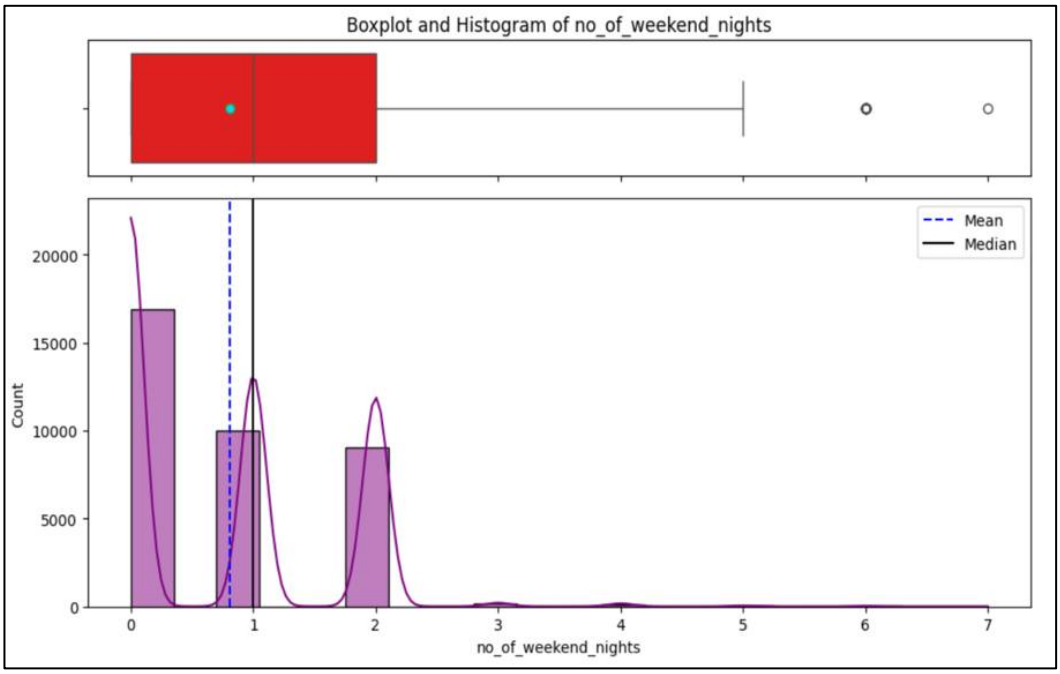


Figure 6. Distribution of no_of_weekend_nights.

- Histogram plot shows data is highly right skewed.
- The median is **1 night, and 75%** of bookings stay up to 2 weekend nights.
- **The maximum is 7**, which may represent extended weekend stays or rare cases.
- Box plot shows there two outliers at the lower end of whiskers depicts the 6 & 7 weekend nights booked or stay by guests.

Observations on no_of_week_nights

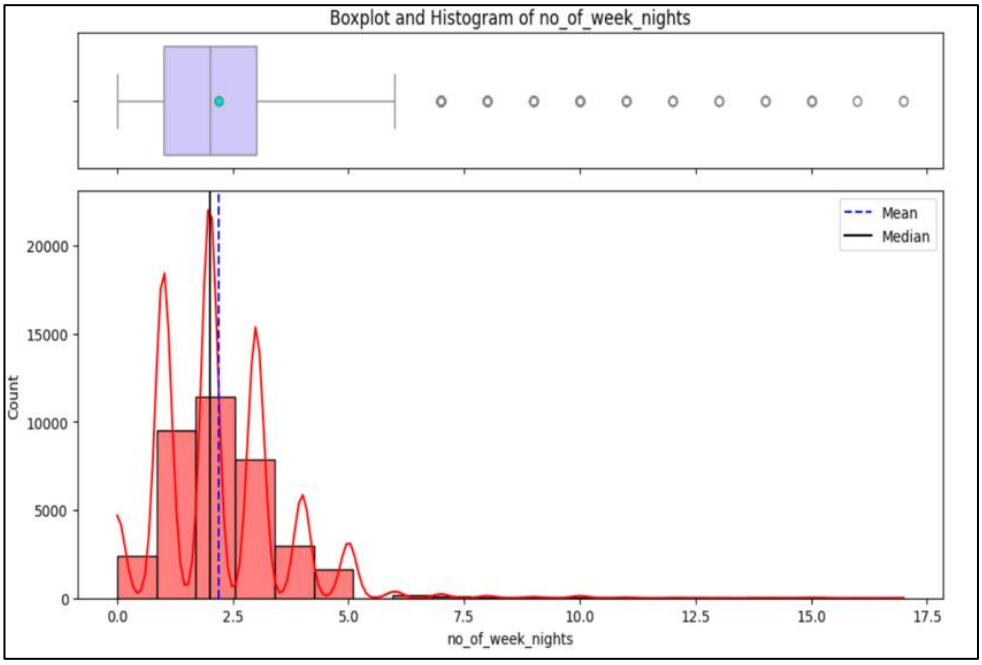


Figure 7. Distribution of no_of_week_nights.

Histogram & Box Plot

- Plot reveals data is highly right skewed.
- The median is **2** nights, with most bookings covering **1 to 3** weeknights.
- There are a few long stays, with a **maximum of 17**.
- There are several outliers below the lower whiskers of box plot ranging from 7-17 weeknights.

Observations on required_car_parking_space

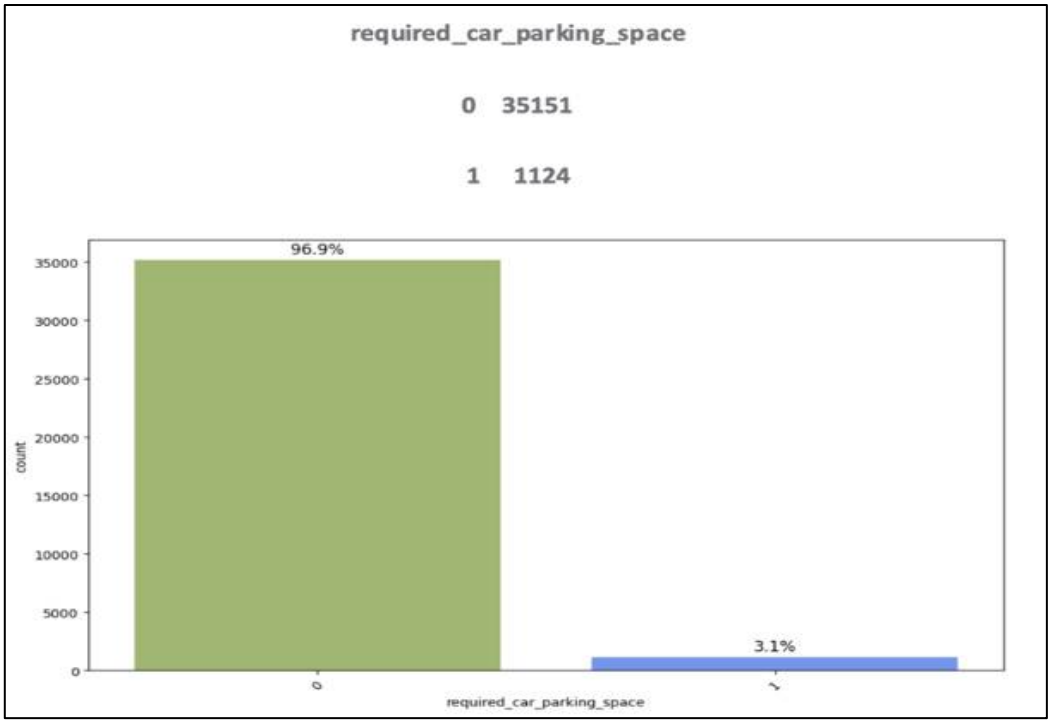


Figure 8. Distribution of required_car_parking_space.

Bar Plot

- Bar plot illustrates there **3.1%** customers requires car parking space while majority (**96.9%)** customers do not opt for car parking.
- Total number of **35151** out of 36275 customers do not want car parking space.

Observations on arrival_month

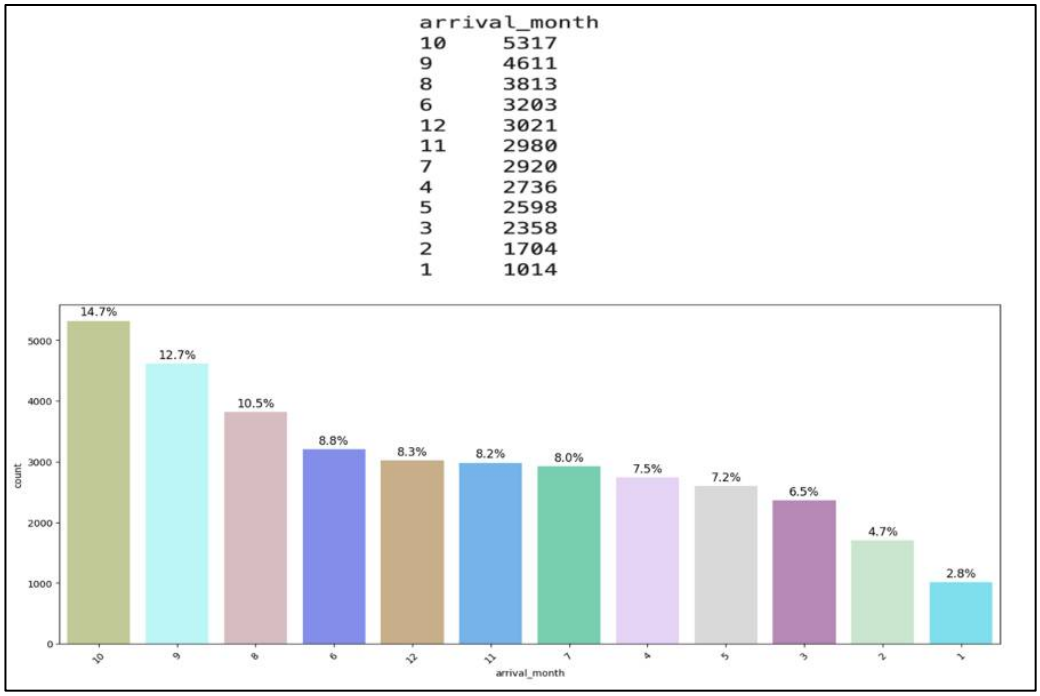


Figure 9. Distribution of arrival_month.

Bar Plot

- The bar plot suggest **October (Month 10) is the busiest month**, making up 14.7% of annual bookings with 5,317 bookings. Likely due to ideal weather.
- September (12.7%) and August (10.5%) follow as the next highest months.
- From **August to December**, we see about **55% of total bookings**: Aug (10.5%), Sep (12.7%), Oct (14.7%), Nov (8.2%), Dec (8.3%). likely driven by year-end holidays, including Thanksgiving, Christmas, and New Year, keep demand steady despite the start of winter. This highlights opportunities for premium festive packages.
- **January to March is the slowest quarter**, accounting for only 14% of bookings: Jan (2.8%), Feb (4.7%), Mar (6.5%). This could be due to harsh winter season.
- **June (8.8%)** performs better than July (7.2%) and is close to December (8.3%), even though it is not a traditional peak month. Customers likely take advantage of shoulder-season pricing and avoid the peak crowds in July.
- **April (7.5%) and May (7.2%)** show moderate demand. They exceed the winter months but fall behind the summer and fall months. Spring weather encourages travel, but demand stays below the peaks of summer and fall. This indicates that there is potential to increase demand during this shoulder season with campaigns that focus on nature, like spring blooms and hiking.

Observations on arrival_year

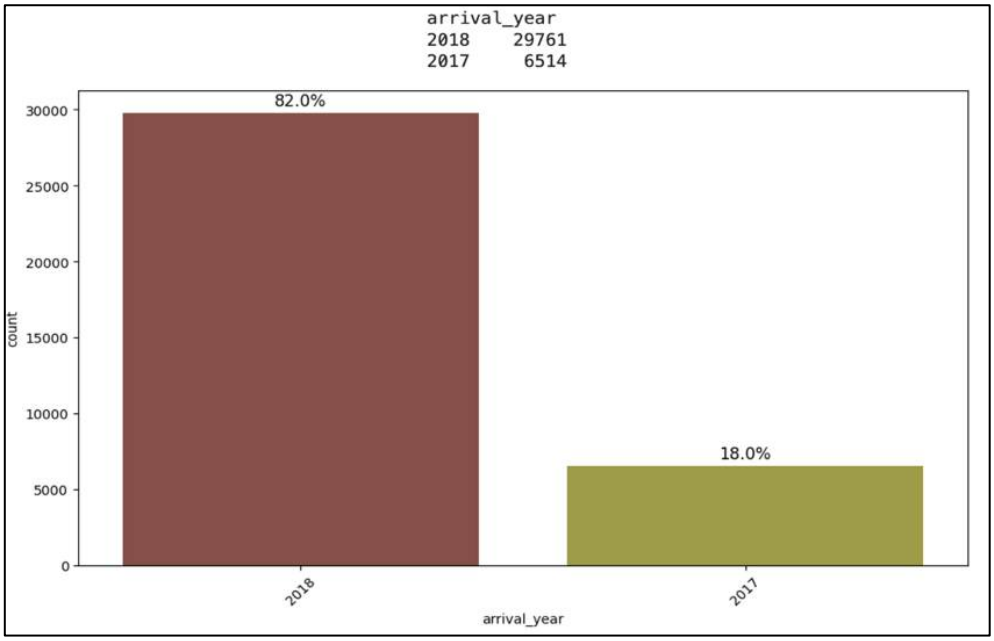


Figure 10. Distribution of arrival_year.

Bar Plot

- The bar plot shows that extreme imbalance in customer arrival year 2017 & 2018.
- Majority of customers **82%** (29761/36275) arrived in the year **2018**, and only **18%** (6514/36275) arrived in **2017**.
- This skewed distributions is likely to stems from factors like business growth in 2018 , better marketing, etc,.

Observations on arrival_date

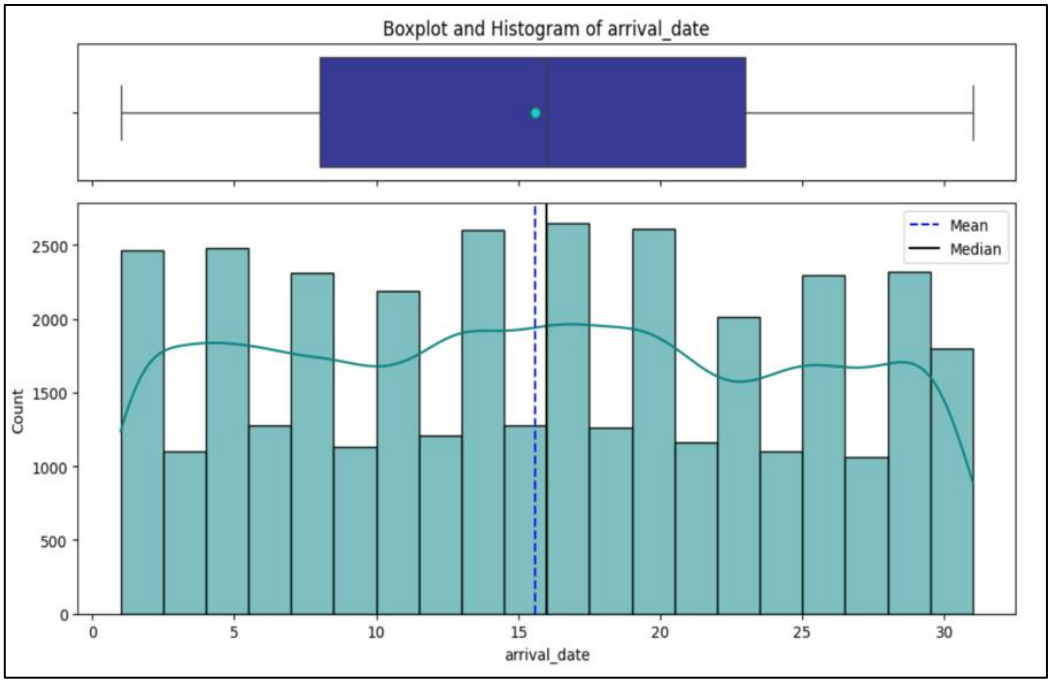


Figure 11. Distribution of arrival_date.

- The plot shows that data has uniform distribution as arrivals dates are evenly spread across days **1 to 31**, with no dominant peaks. Daily counts stay consistently between 1,000 and 1,300, suggesting bookings are not affected by specific dates, such as paydays or weekends. This indicates stable demand throughout the month.
- We observed slight mid-month dip as days 10 to 20 show slightly lower counts, around 1,000, compared to the start and end of the month, which see counts around 1,200.This means **customers may avoid mid-month** because of work commitments or a preference for weekends at the beginning and end of the month.
- The boxplot shows **no outliers**, confirming that arrival dates are evenly distributed without any anomalies.
- The central tendency (mean ≈ median ≈ **15-16**) aligns with mid-month, reinforcing the absence of cyclical bias.

Observations on type_of_meal_plan

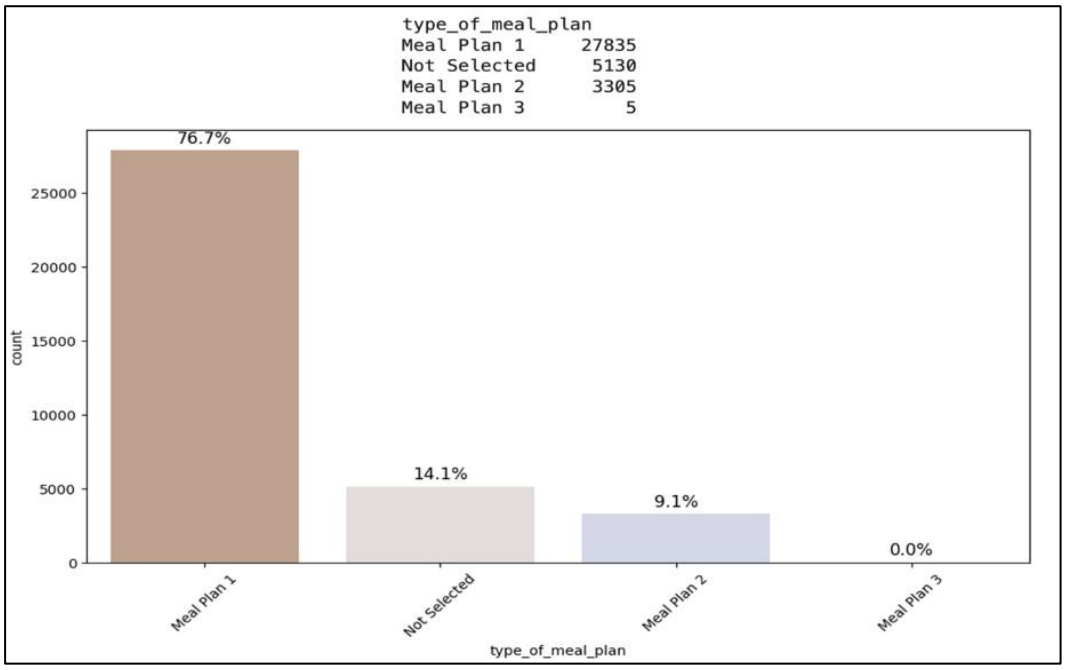


Figure 12. Distribution of type_of_meal_plan.

- The bar plot demonstrates right skewness.
- **Meal Plan 1 is most preferred** by customers (76.7%). This indicates guests strongly prefer breakfast-only plans. This choice likely stems from a desire for flexibility, like dining out later, and for saving money. It shows that guests value convenience in the mornings.
- Only **14.1% of customers do not select** any meal plans. This suggest noticeable minority seems to prioritize budget management or the ability to choose their meals freely. They might prefer eating locally or want to avoid fixed costs.
- **Meal Plan 2 has low appeal, with only 9.1%** customers choosing it. This shows that half board meal does not attract many customers. This could mean that people want to skip restrictions for lunch or .dinner or they think it offers less value compared to dining à la carte.
- **Meal Plan 3 is almost non-existent, capturing 0.01% of interest**. This is suggests that full board plans are not financially feasible for most customers. This could be due to high prices, lack of flexibility, or a mismatch with guests' preferences, such as guests wanting to explore local food.

Observations on room_type_reserved

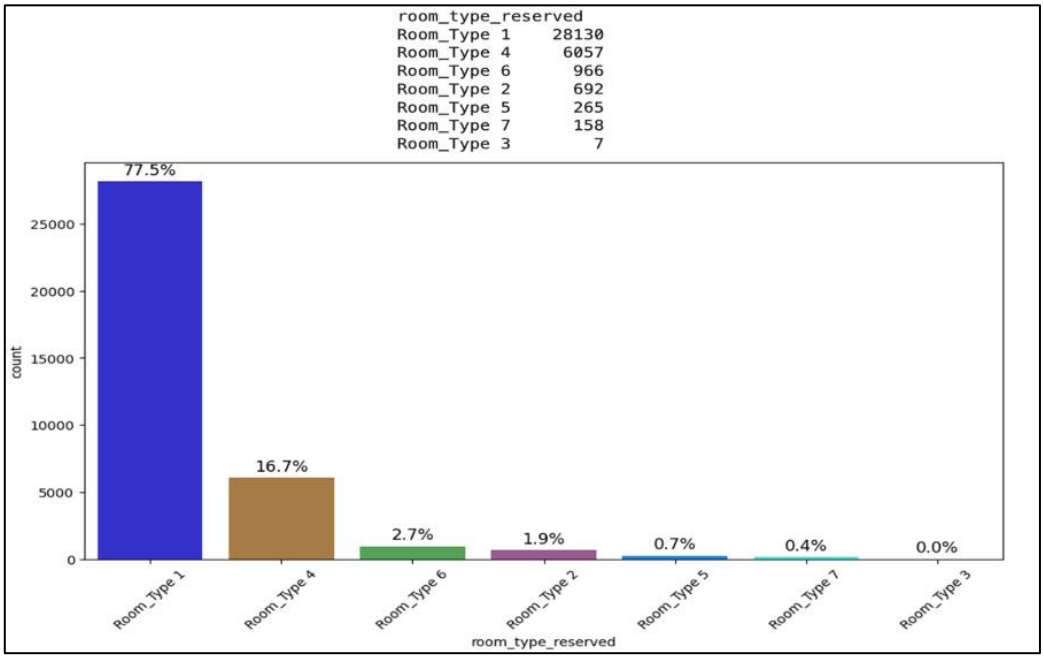


Figure 13. Distribution of room_type_reserved.

- **Room_Type 1 dominates at 77.5%.** This is likely the standard or budget option. It attracts cost-conscious solo travellers or couples who want basic amenities.
- **Room_Type 4 is next at 16.7%.** This is probably a family or group room, such as one that accommodates four people. It shows a specific but limited demand for larger spaces.
- **Other room types make up less than 4% combined**. These may include specialized suites like luxury, themed, or accessible rooms. They might have limited appeal because of their higher prices or specific uses.
- **Room_Type 3 is nearly zero at 0.02%.** This may be an outdated or overpriced category, indicating a need for changes or even removal.

Observations on market_segment_type

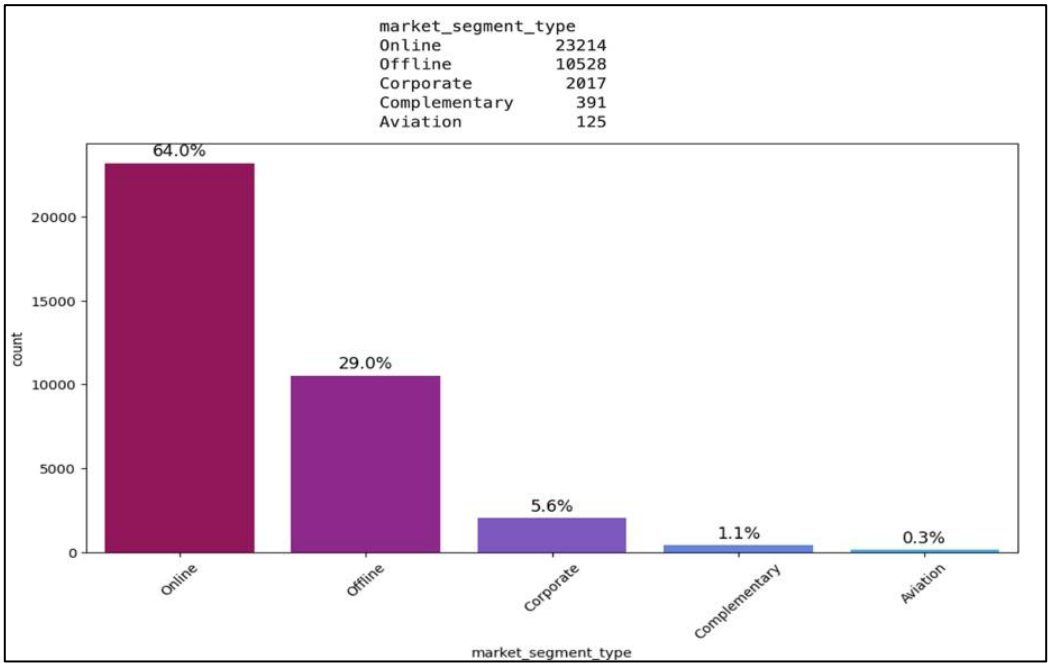


Figure 14. Distribution of market_segment_type.

- **Online booking dominates (64.0%)**. Digital booking platforms, such as OTAs and hotel websites, are the main sales channel. This shows that customers prefer to book online.
- **Offline booking is significant (29.0%)**, suggesting walk-ins and phone bookings still matter, indicating a demand for traditional service, especially for last-minute or loyalty-driven customers.
- **Corporate booking is niche (5.6%)** as there might be some business travel partnerships, but they are not fully used. This presents an opportunity for growth in B2B markets.
- **Complementary (1.1%) & Aviation (0.3%) are marginal**. The free stays for staff and airline crew contracts add little revenue, showing that these non-revenue segments are meant to be kept small.

Observations on repeated_guest

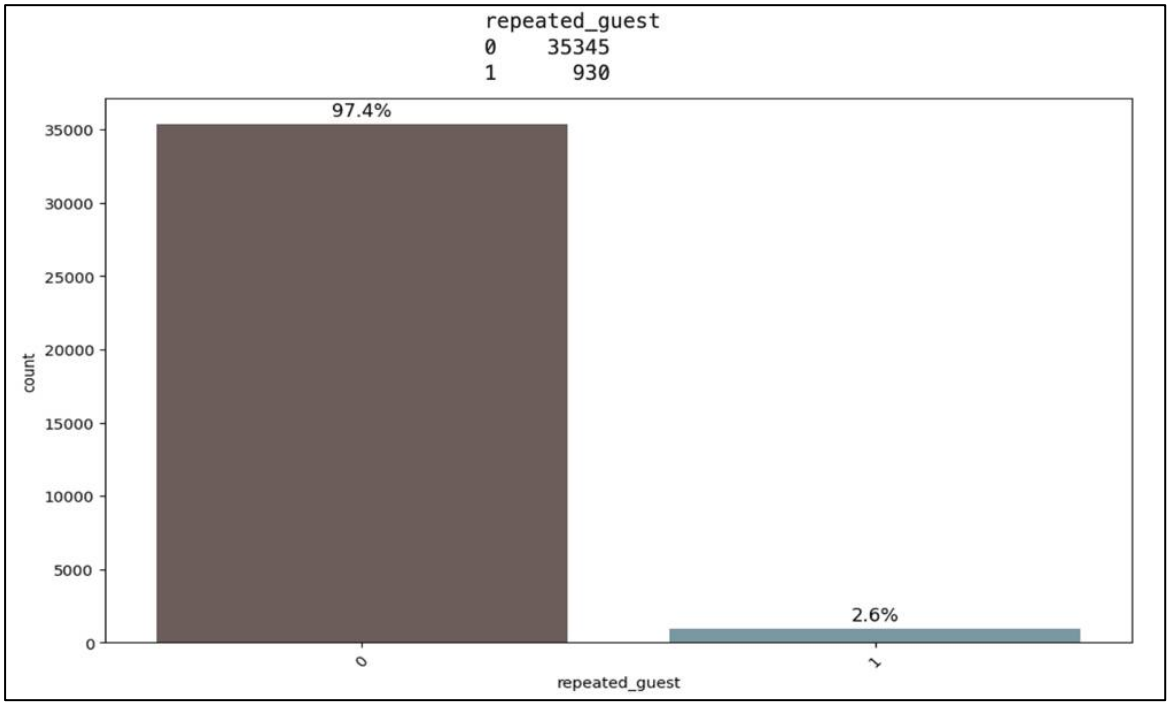


Figure 15.Distribution of repeated_guest.

- The bar plot reveals extremely Low rate **(2.6%) of repeated guests**. Only **930 guests return**, while 35,345 (97.4%) are one-time visitors.
- This indicates , the hotel has trouble keeping guests, likely because of weak loyalty programs, inconsistent experiences, or strong competition.

Observations on no_of_previous_cancellations

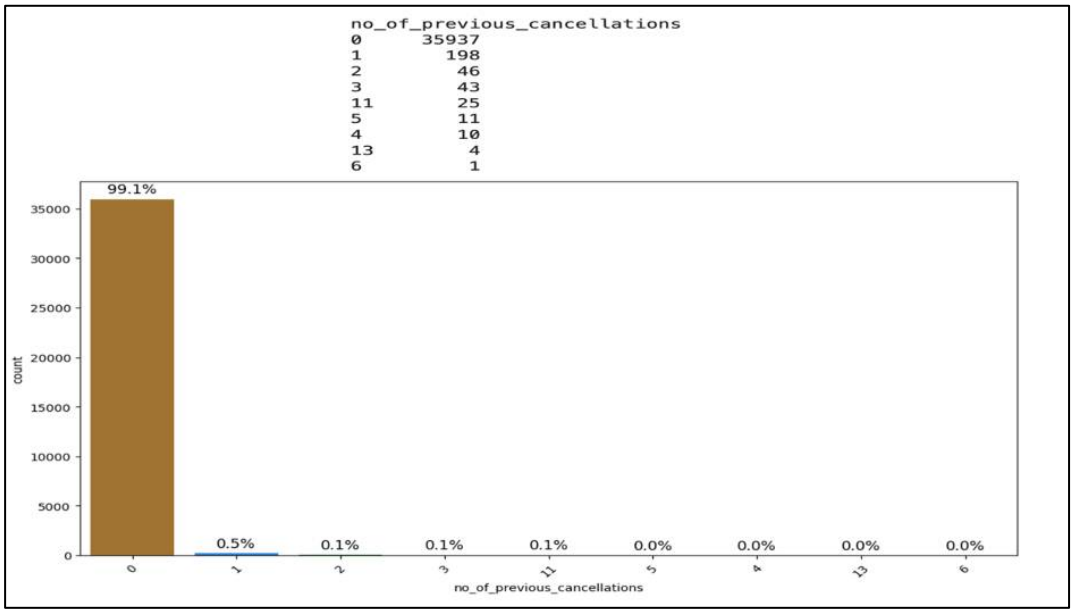


Figure 16. Distribution of no_of_previous_cancellations.

- The plot depicts that hotel has **99.1% zero cancellations** (35,937). This means the vast majority of guests are low-risk and honour their bookings. There is minimal impact on operations from cancellations.
- Only **1% (338) have 1 or more cancellations**. This is a small but may be due to significant group of habitual cancellers. They pose a risk for revenue loss and create inventory issues.
- **Extreme outliers (11 to 13 cancellations: 29 customers)**. These customers probably trying to game the system by holding multiple options or have fraudulent accounts. We need stricter policies, such as requiring prepaid bookings.
- **Moderate repeat cancellers (1 to 3 times: 287 guests)**. This group may include unpredictable travellers, such as business travellers, or those who are dissatisfied because they found better alternatives.

Observations on no_of_previous_bookings_not_canceled

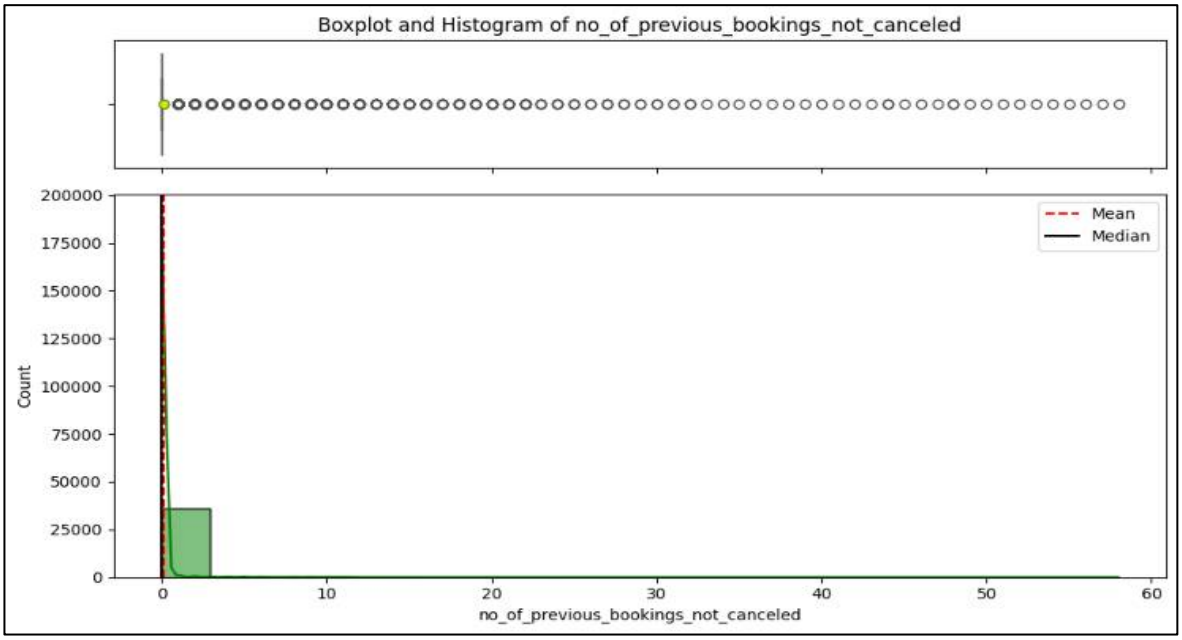


Figure 17. Distribution of previous bookings not cancelled.

- The data is extremely right-Skewed as Mean > Median, indicating even "frequent" guests show only modest repeat activity, indicating there is no strong loyalty base.
- More than **95+% of customers** cluster near 0, with the leftmost histogram bar being the most prominent with peak of nearly 35k, indicating a **few loyal or repeat customers**.
- There are several outliers on the lower end whiskers. A small number of guests have 10 to over 60 uncanceled bookings, with boxplot outliers extending to 60. This is could be due to a tiny loyal segment, likely made up of business travellers or staff, but their presence is not significant on a larger scale.
- The boxplot shows the median set at zero. For more than 50% of guests, this is their first uncanceled booking, confirming low retention.

Observations on no_of_special_requests

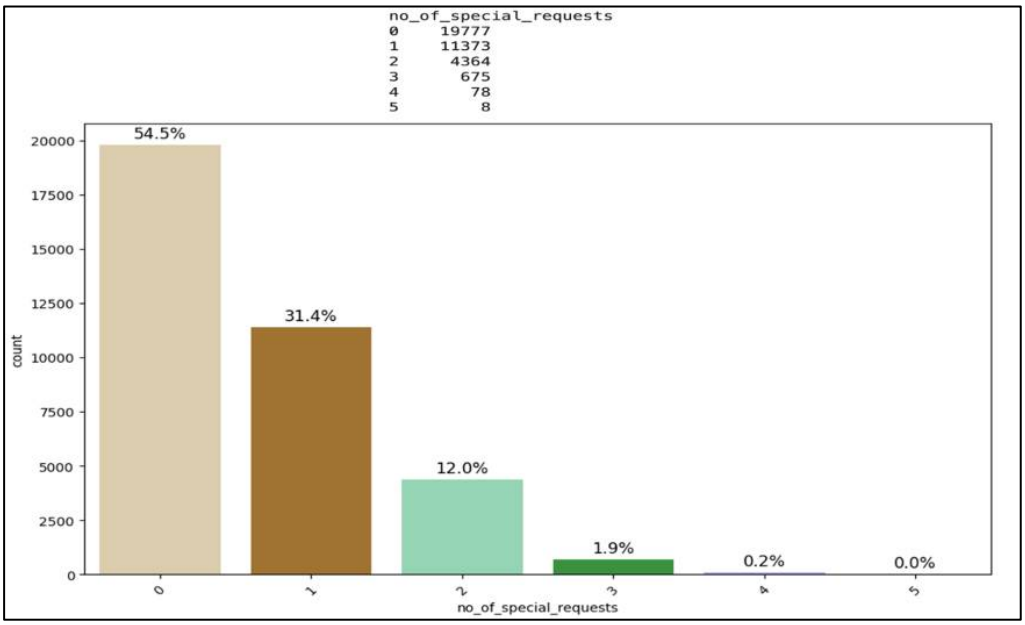


Figure 18. Distribution of no_of_special_requests.

- Most customers - **54.5% (19,777) make 0 requests**. This is might be due to the majority prioritize simplicity and convenience. They are likely business travellers or short-stay guests who avoid customization.
- Around **31.4% (11,373) customers make 1 request**. This is significant group might be looking for light personalization, such as a high floor or a view. This suggests an opportunity to upsell modest amenities.
- 12% customers **(4,364) make 2 requests**. This group may be experience-focused customers, like couples or families, are combining needs such as a view and a quiet room. This suggests the potential to bundle popular requests.
- Only **2.1% (761) customers make three or more requests**. This small group might consists of high-maintenance guests, often luxury or long-stay customers. They may require more resources but contribute to premium revenue.

Observations on booking_status

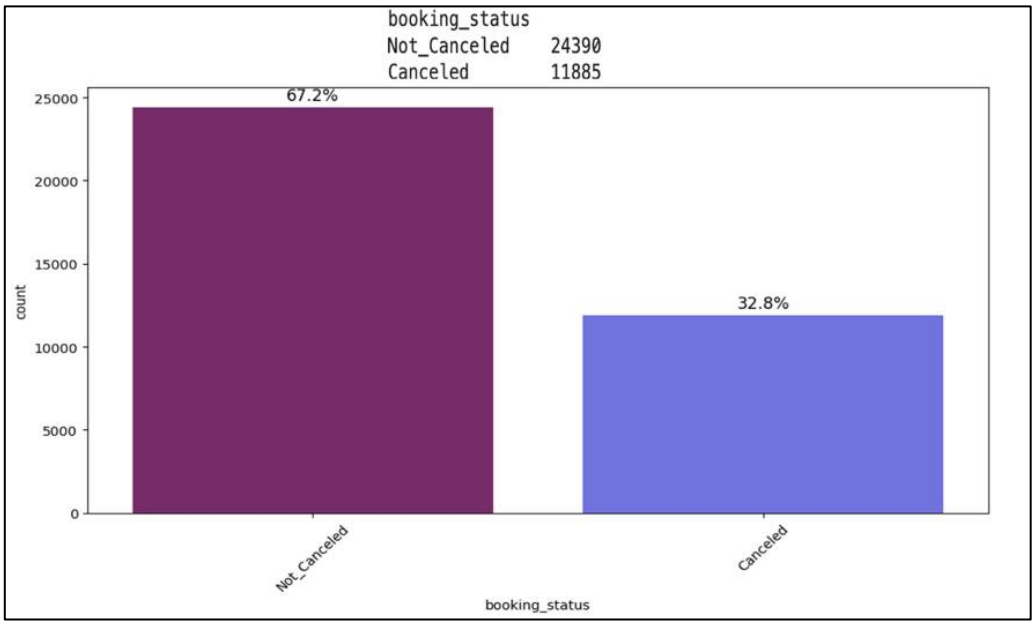


Figure 19. Distribution of booking_status.

- High Cancellation Rate as **11,885 bookings were cancelled**, which is 32.8% of the total. This poses a significant revenue risk from last-minute dropouts. This issue may stem from long lead times, with a median of 57 days, or flexible policies.
- **Dominant Non-Cancellations are 24,390 bookings**, making up 67.2%. Our core customer base is dependable, but cancellations hurt profitability. This results in empty rooms and unnecessary staff costs.
- Since **1 in 3 bookings were cancelled**, we may resort to overbooking to balance this. However, this strategy risks customer dissatisfaction if not handled correctly.
- This trend connects to past data on repeat cancellers, which represent **1% of guests who have had one or more prior cancellations**. They likely contribute to significant losses.

2.2 Bivariate Analysis

Before going ahead with Bivariate analysis we would encode booking_status and drop Booking_ID features to strengthen the analysis for modelling.

	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time
0	2	0	1	2	Meal Plan 1	0	Room_Type 1	224
1	2	0	2	3	Not Selected	0	Room_Type 1	5
2	1	0	2	1	Meal Plan 1	0	Room_Type 1	1
3	2	0	0	2	Meal Plan 1	0	Room_Type 1	211
4	2	0	1	1	Not Selected	0	Room_Type 1	48

Table 5. Top 5 rows after dropping Booking_ID.

- We can observe that Booking_ID has been dropped from our dataset.

booking_status	
0	24390
1	11885

Table 6. booking_status encoded.

- We have successfully **encoded booking_status values as Canceled to 1 and Not_Canceled as 0** for further analysis.

<u>Correlation Analysis on Numerical columns</u>

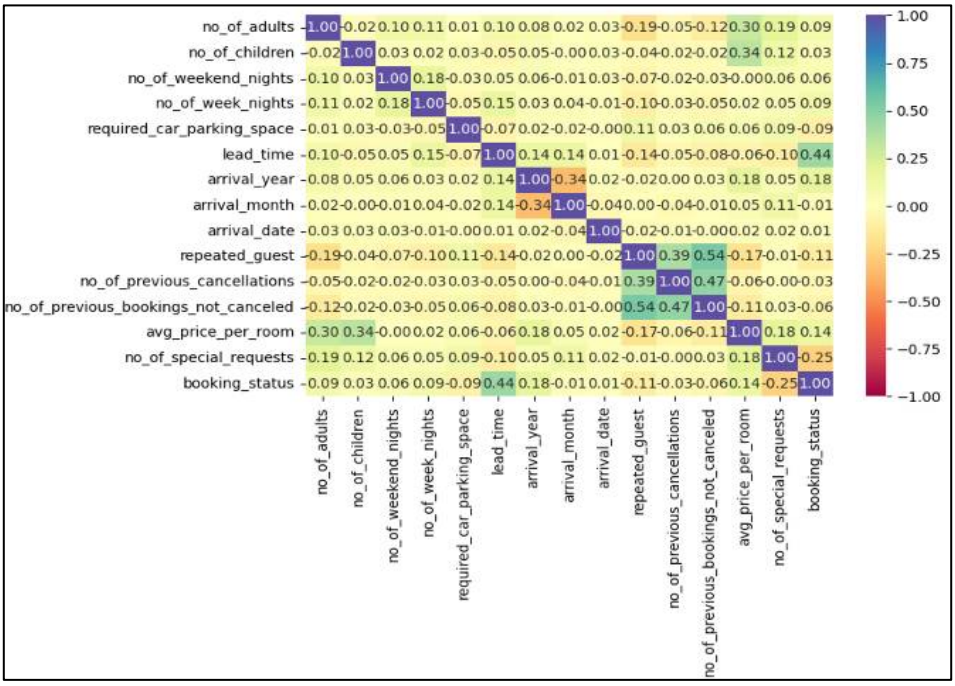


Figure 20. Heatmap of all numerical columns.

lead_time (r = 0.44):

- Shows strongest positive correlation.
- Bookings made far in advance are 44% more likely to cancel. This is likely due to changes in plans or seeking better deals.

no_of_special_requests (r = -0.25):

- There is a significant negative correlation.
- Guests making special requests are 25% less likely to cancel.
- Personalized stays strengthen commitment.

arrival_year (r = 0.18) & avg_price_per_room (r = 0.14):

- Both the feature have weak positive correlation.
- The bookings from 2018 and higher-priced rooms slightly increase cancellation risk. This may be due to competing offers or sensitivity to price.

repeated_guest (r = -0.11):

- There is moderate negative correlation.
- Repeat guests are 11% less likely to cancel. Loyalty helps reduce cancellations.

no_of_adults (r = -0.09) & required_car_parking (r = -0.09):

- They have mild negative correlations.
- Group or family bookings (adults) and parking users are a bit more reliable. Planned logistics help lower cancellations.

Low-Impact Factors (|corr| < 0.1):

- Children, weekend or week nights, and arrival dates show near-zero correlation, leading to no impact on cancellations.

Past cancellations or booking history:

- They show weak association (r = -0.03 to -0.06), indicating limited predictive power.

Observations on avg_price_per_room vs market_segment_type

market_segment_type	sum	mean	median	min	max
Aviation	12588.00	100.704000	95.0	79.0	110.0
Complementary	1228.43	3.141765	0.0	0.0	170.0
Corporate	167232.98	82.911740	79.0	31.0	220.0
Offline	964708.84	91.632679	90.0	12.0	540.0
Online	2605930.63	112.256855	107.1	0.0	375.5

Figure 21. Distribution of avg_price_per_room vs market_segment_type.

Online Dominates Revenue (69.5%):

- Total Average Price per room: €2,605,930.63
- Digital channels, like OTAs (Online Travel Agencies) and hotel websites, are crucial for profitability since they account for almost 70% of revenue. Guests prefer booking on their own for convenience. However, high OTA commissions can cut into profits.

Offline Contribution is Significant (25.7%):

- Total Average Price per room: €964,708.84
- Walk-ins and phone bookings are still important, making up more than 25% of revenue. This group likely includes loyal customers or those booking last minute to avoid online fees.

Corporate Underperforms (4.5%):

- Total Average Price per room: €167,232.98
- Even with contracts in place, corporate clients contribute very little. The low volume of bookings in 2017 and possible discounts limit revenue opportunities.

Aviation & Complementary Are Negligible (<0.4%):

- Total Average Price per room: €12,588.00
- Airline crew stays are steady but not very valuable, with only 125 bookings. They help fill rooms but do not significantly add to revenue.

Complementary:

- Total Average Price per room: €1,228.43
- Free stays for staff and comps are kept minimal to maintain revenue integrity.

Conclusion:

- **Online bookings have the highest average room prices starting at €112.25 per room**. In contrast, corporate and offline segments get lower rates, typically 15 to 25% less.
- This shows how dynamic pricing based on demand and negotiation power affects rates.

Observations on booking status vs market_segment_type

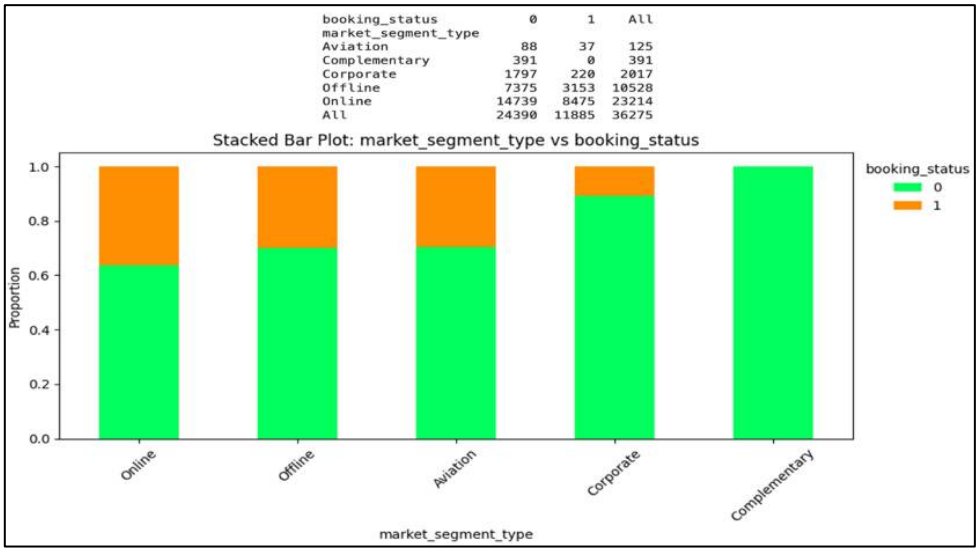


Figure 22. Distribution of booking_status vs market_segment_type.

Online Bookings:

- It has highest cancellation volume with **36.5% cancellation** rate (8,475 out of 23,214).
- Many digital channel struggles with easy cancellation policies, such as free refunds on OTAs. Guests often look for cheaper options.

Offline Bookings:

- It has moderate cancellations with **30% cancellation** rate (3,153 out of 10,528).
- Walk-ins and phone bookings show steadier trends than online bookings due to direct communication, but they are still affected by changing plans.

Corporate Bookings:

- It is the most reliable with only **10.9% cancellation** rate (220 out of 2,017).
- Negotiated contracts and the inflexible nature of business travel lead to fewer cancellations, such as penalties or fixed schedules.

Aviation & Complementary:

- They have the niche stability with **Aviation: 29.6% cancellations** (37 out of 125)
- Airline crew schedules change often, but the small number of bookings limits the impact.
- **Complementary: 0% cancellations** (0 out of 391)
- Non-revenue stays, like staff or comps, are always honoured, so there is no financial risk.

Conclusion:

- **Online bookings face the highest cancellation rate at 36.5%.**
- In contrast, corporate contracts have the most stable bookings, with a cancellation rate of 10.9%.
- This shows a clear trade-off between price optimization and booking reliability across different segments.

Observations on no_of_special_requests vs booking_status

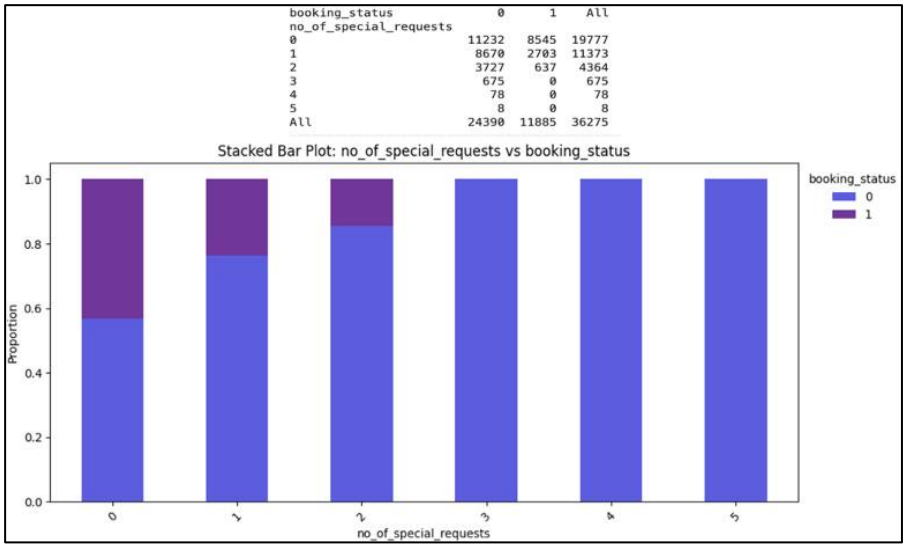


Figure 23. Distribution of no_of_special_requests vs booking_status.

0 Special Requests:

- The highest number of bookings was about 19,777.
- The cancellation rate was the **highest**, with around **43% cancelled** (8,545 / 19777).
- This group showed the most imbalance toward cancellations.

1 Special Request:

- The cancellation rate was moderate at about **24% cancelled** (2,703 / 11373).
- This is better than 0 requests, but still noteworthy.

2 Special Requests:

- There was a sharp drop in **cancellations to about 15%** (637 / 4,364).
- The acceptance rate improved significantly.

3 to 5 Special Requests:

- No cancellations were observed.
- All these bookings were honoured, meaning none were cancelled.
- The **booking counts are very low** (ranging from **675 to 8**), but the trend remains consistent.

Conclusions:

- **A higher number of special requests strongly relates to lower cancellation rates**.
- Guests with three or more requests never cancelled in this dataset.
- This indicates that guests making special requests are more serious about their bookings.
- Hotels could use this information to prioritize and personalize service for these guests, which could help further reduce cancellations.

Observations on repeated_guest vs booking_status

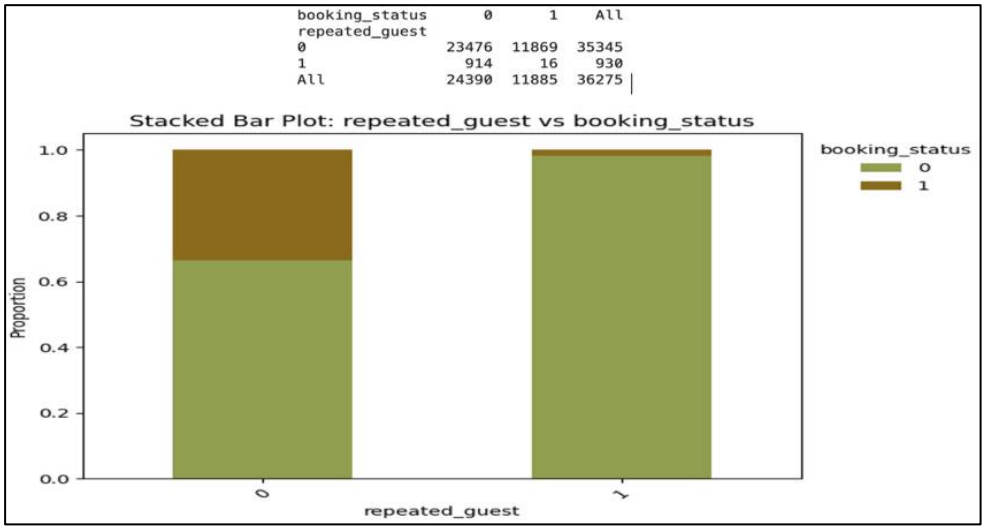


Figure 24. Distribution of repeated_guest vs booking_status.

Repeated Guests (repeated_guest = 1):

- Total: 930 bookings
- Cancelled bookings (1): 16, about **1.7% cancellation** rate.
- The vast majority accepted their bookings.
- Repeated guests have almost no blue section for cancellations, showing that nearly all their bookings are completed.

Non-Repeated Guests (repeated_guest = 0):

- Total: 35,345 bookings
- Cancelled bookings (1): 11,869, about **33.6% cancellation** rate
- A significant number of non-repeated guests cancelled their bookings.
- Non-repeated guests show a much higher share of cancellations compared to repeated group.

Conclusion:

- **Repeated guests rarely cancel bookings, around 1.7%, which shows strong brand loyalty and intent to book.**
- In contrast, 1 in 3 non-repeated guests cancel, indicating lower commitment or higher uncertainty.
- This confirms that repeated guests are valuable and dependable customers.
- Therefore, hotels should prioritize and reward them through loyalty programs or perks.

Observations on no_of_special_requests vs avg_price_per_room

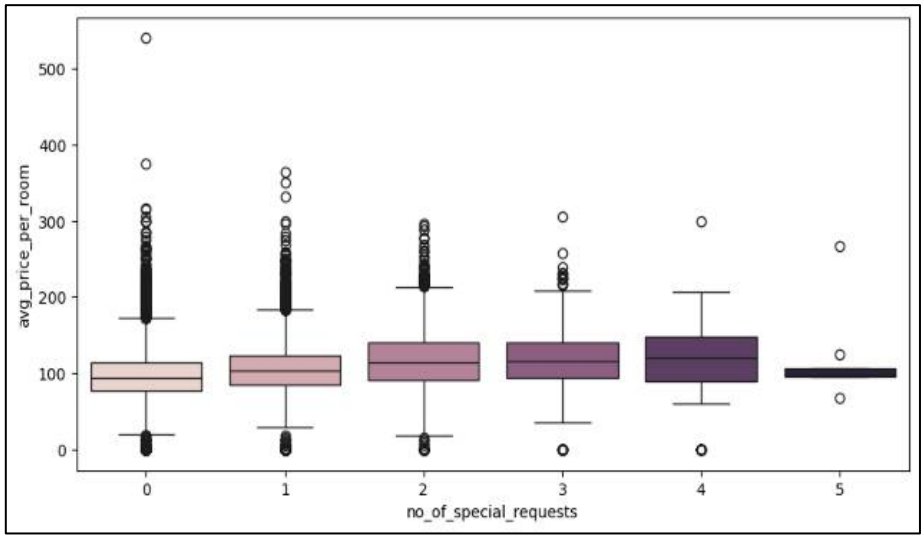


Figure 25. Distributions of no_of_special_request vs avg_price_per_room.

- The box plot shows the median **average price per room increases as the number of special requests rises from 0 to 4**. This suggests that guests who make more requests generally spend more on average.
- With 0 or 1 requests, the median is lower (around 90 to 103), showing a wider spread and more variability.
- For 2 to 4 requests, the medians are higher (approximately 115 to 120), and the interquartile ranges are slightly narrower, though there is still some variability.
- With 5 requests, the data is flat, indicating very few observations with little to no variation (median is about 100).
- Outliers are present in all categories, showing some high-priced bookings regardless of the request count.

Conclusion:

- There is a **positive correlation (r = 0.18) between the number of special requests and room price up to 4 requests**. This suggests that guests who spend more may be more likely to ask for additional services or personalized touches.
- **Very few guests make 5 requests**, and they typically pay around the average, not higher than others.
- Hotels can conclude that **customers who pay more expect additional service customizations**, which can help in designing services or premium offerings.

Observations on avg_price_per_room vs booking_status

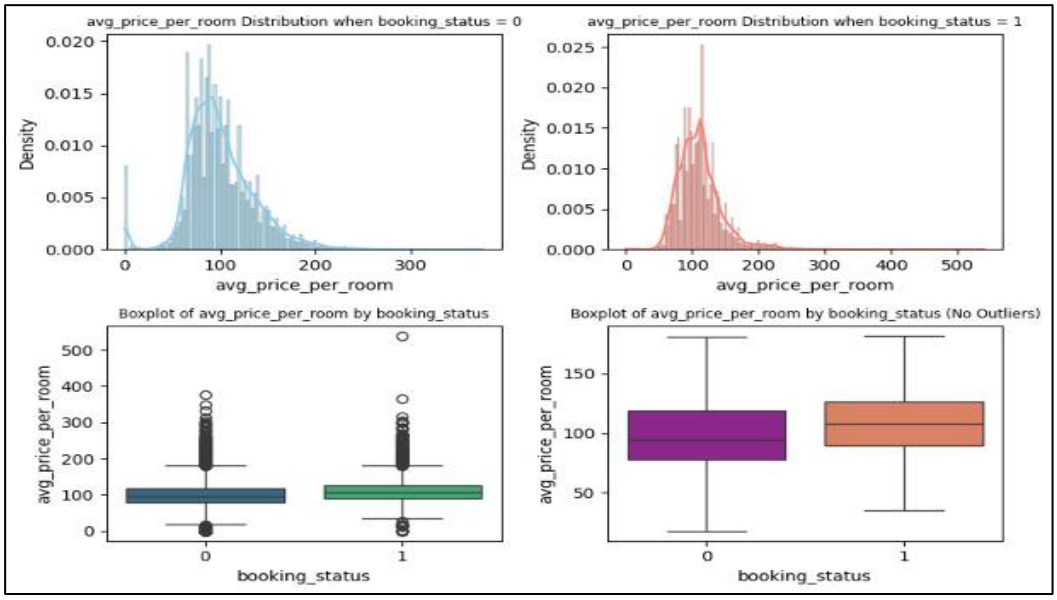


Figure 26. Distribution of avg_price_per_room vs booking_status .

- Plot shows data is right skewed.
- Both **cancelled (1) and not cancelled(0) have almost similar average room prices**.
- This indicates there is not much impact of average room prices on booking status.
- Presence of outliers in booking status suggests high prices rooms
- Thus, we can say that customers might have been cancelling the booking based on other factors as well.

Observations on lead_time and booking_status

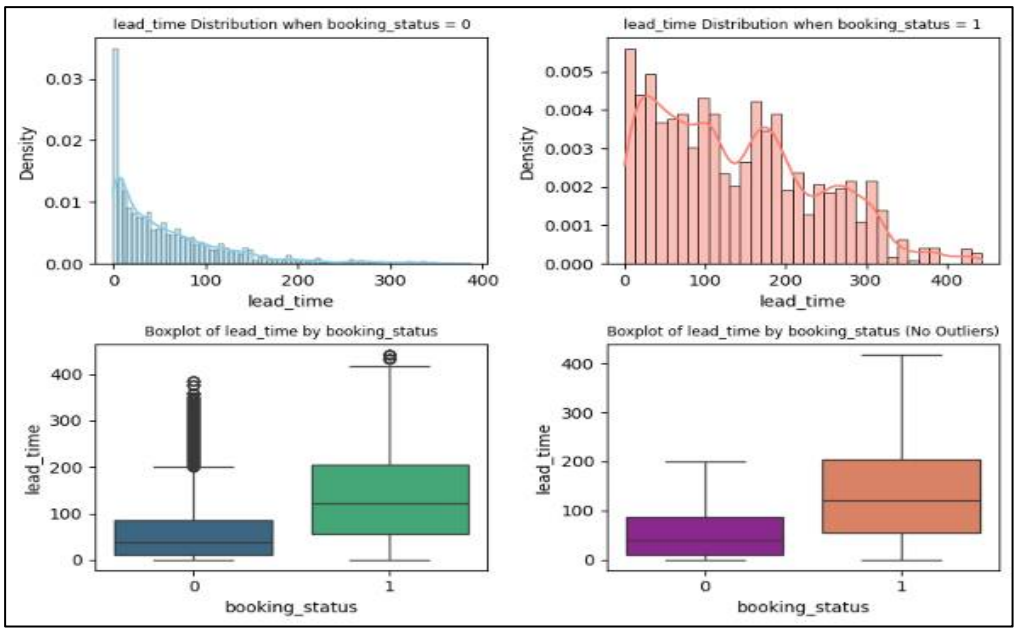


Figure 27.Distribution of lead_time and booking_status.

- Plots reveals data is highly right skewed.
- The **lead time is less for the bookings which are not canceled**.
- Hence we can say as the lead time increases, the chances of bookings getting canceled increases.

Observations on no_of_week_nights vs booking_status

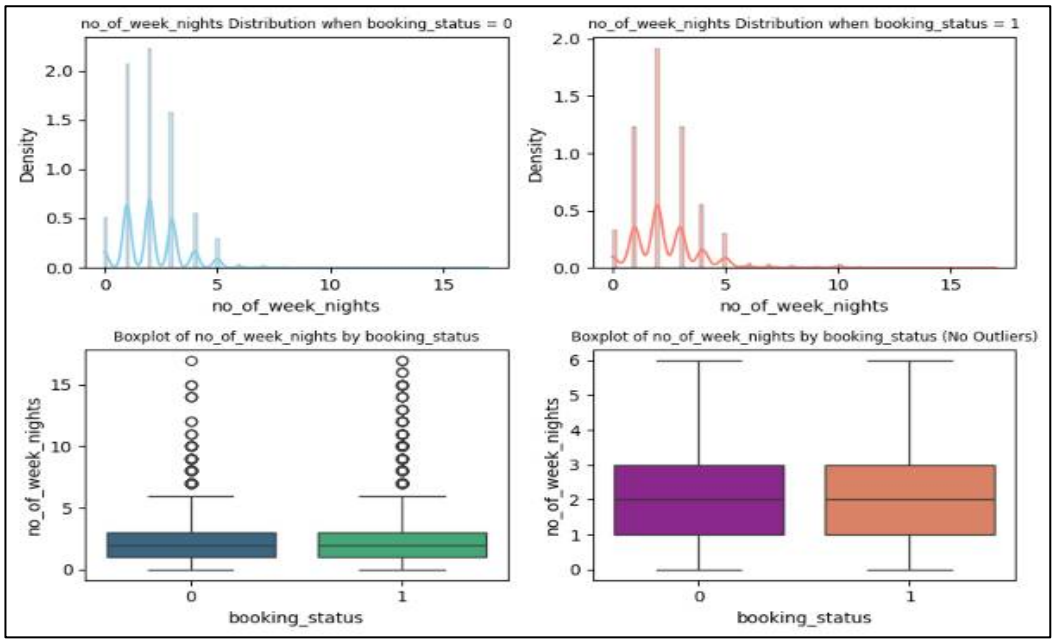


Figure 28. Distribution of no_of_week_nights vs booking_status.

- Both canceled bookings (booking_status = 0) and not canceled bookings (booking_status = 1) have similar distribution shapes.
- **Most stays last between 1 and 3 week nights**.
- Both distributions are right-skewed, with a few longer stays reaching up to about 15 nights.

The boxplots, with and without outliers, show:

- The median stay for both groups is around 2 nights.
- The interquartile range (IQR) is very similar for both groups.
- There are outliers (longer stays) in both groups, but they are more visible in the canceled group.

Conclusion:

- The **number of week nights stayed does not significantly affect whether a booking was canceled (booking_status = 0) or not canceled (booking_status = 1)**.
- While there are slightly more longer stays in the canceled group, the difference is not significant.
- Overall, the duration of week nights alone is not a strong indicator of booking status.

Observations on no_of_weekend_nights impacts booking_status

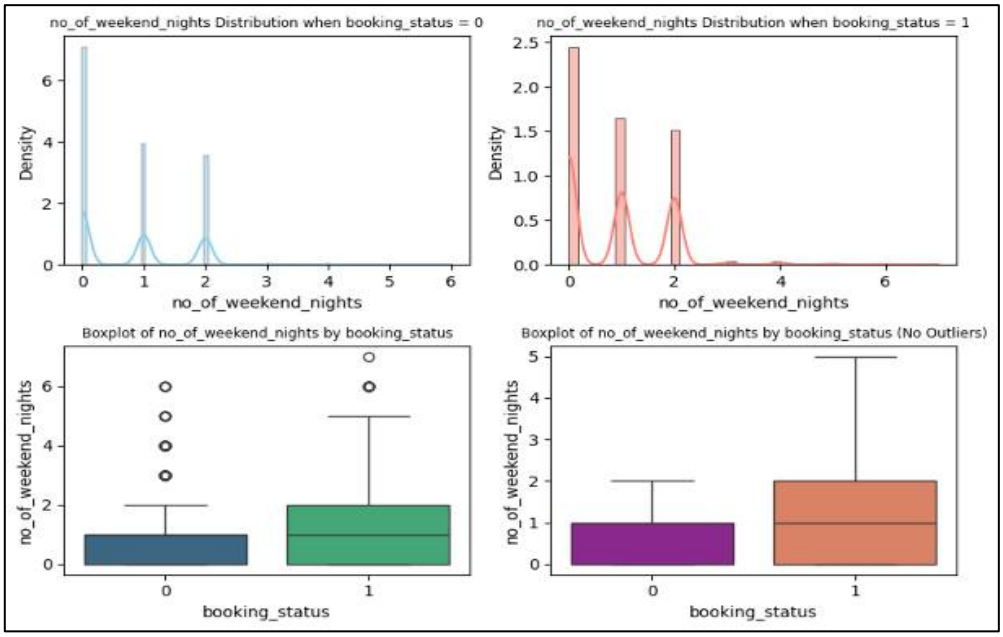


Figure 29. Distribution of no_of_weekend_nights vs booking_status.

- Canceled bookings (1): Show higher density at 0 and 1 weekend night. Most canceled bookings are short weekend stays.
- Not Canceled bookings (0): Have a wider spread with noticeable density at 2 weekend nights and beyond. This indicates longer weekend stays.
- The median number of weekend nights is higher for not canceled bookings (0).
- Canceled bookings (1) are more concentrated around 0 to 1 nights. Not canceled bookings show a wider spread, even after removing outliers.

Conclusion:

- **Customers who book for longer weekend stays are less likely to cancel**.
- Short weekend stays are linked to a higher chance of cancellation.
- Thus, no_of_weekend_nights shows stronger and more distinguishable impact on booking status then no_of_week_nights.

Observations on arrival_year vs booking_status

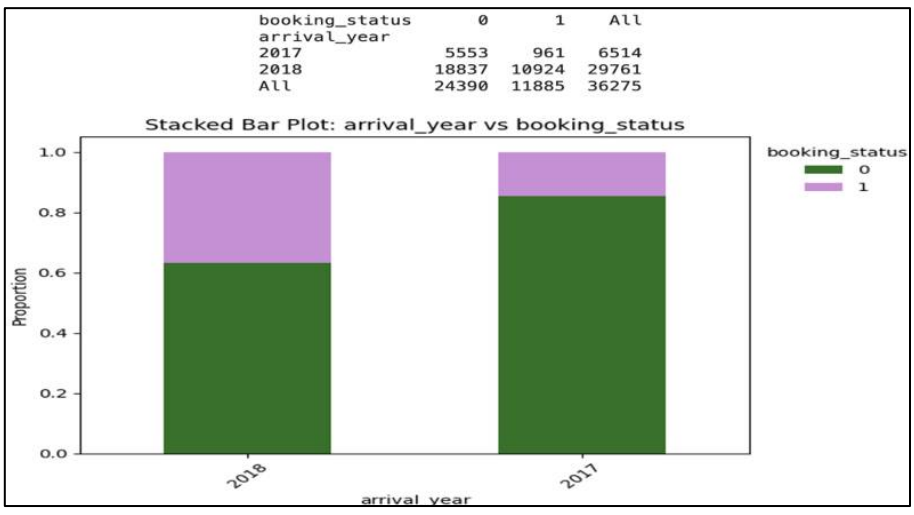


Figure 30. Distribution of arrival_year vs booking_status.

- Cancellations more than doubled from 2017 to 2018, going from 14.75% to 36.70%.
- This is likely due to the rapid business expansion in 2018 to attracted price-sensitive customers who often look for better deals.
- In 2018, there were 4.5 times more bookings, at 29,761 compared to 6,514. However, there were 11.4 times more cancellations, with 10,924 cancellations compared to 961.
- This suggests growing operations without solving the reasons for cancellations, such as long lead times, increased the risk to revenue.

Root Cause Drivers:

- Longer Lead Times: Bookings in 2018 had a higher median lead time of 85 days compared to about 70 days in 2017.
- OTA(Online Travel Agent) Dominance: 82% of 2018 bookings came from online channels, which had a 36.5% cancellation rate.
- New Guest Focus: 97% of the guests were first-time visitors, which means low loyalty and a higher risk of cancellations.

Conclusion:

- **Cancellation rates increased from 14.75% in 2017 to 36.70% in 2018, indicating severe booking instability during hotel business expansion**.

<u>Observations on room_type_reserved vs booking status</u>

booking_status	0	1	All
room_type_reserved			
Room_Type 1	19058	9072	28130
Room_Type 2	464	228	692
Room_Type 3	5	2	7
Room_Type 4	3988	2069	6057
Room_Type 5	193	72	265
Room_Type 6	560	406	966
Room_Type 7	122	36	158
All	24390	11885	36275

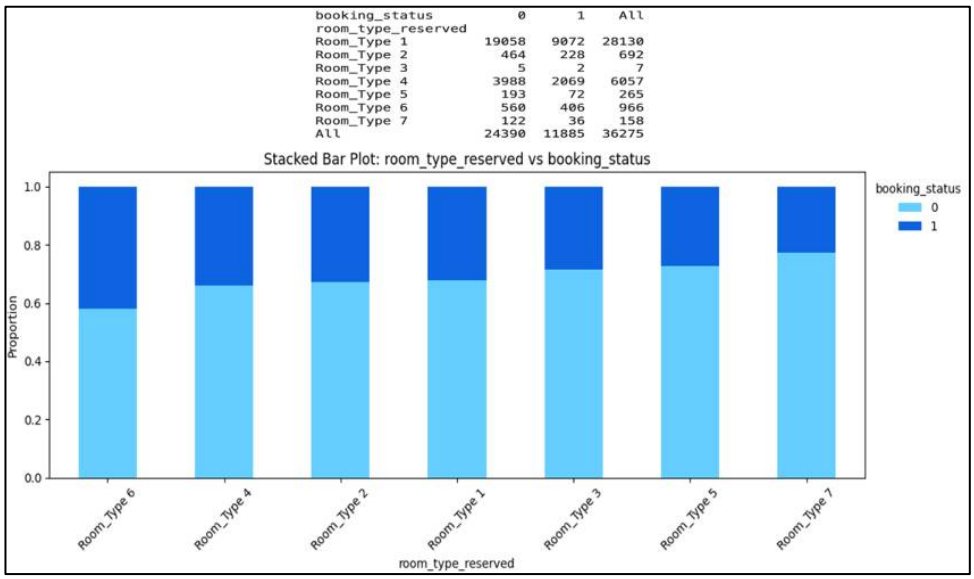


Figure 31. Distribution of room_type_reserved vs booking status.

Room_Type 1:

- Most booked room: 28,130 bookings (**77.6%**).
- Cancellation rate: **32.2%** (9,072 out of 28,130).

Room_Type 2:

- Cancellation rate: **32.94%** (228 out of 692).

Room_Type 4:

- Second most booked: 6,057 bookings (16.7%).
- **Higher cancellation** rate: **34.2%** (2,069 out of 6,057).

Room_Type 6:

- Cancellation rate is relatively high at 42.0% (406 out of 966).

Room_Type 3 & 7:

- Very low booking volume, but **Room_Type 3 has 28.6% cancellations** (2 of 7)
- **Room_Type 7 shows 22.8% cancellations** (36 of 158)

Room_Type 5:

- 265 bookings, with **27.2%** cancelled.

Overall cancellation rate:

11,885 out of 36,275 bookings canceled**(~32.8%).**

Conclusion:

- **Room type seems to affect cancellation behaviour**.
- This is likely because different room types serve different customer groups, each with varying sensitivity to price, expectations, or reasons for staying (e.g., business vs leisure).
- For example, Room_Type 6 has a higher cancellation rate, which may reflect a mismatch between customer expectations and what is offered, or prices that lead to frequent cancellations.
- Thus, room type plays an important role in customer choices and booking reliability, and should be examined for pricing and quality improvement.

<u>Observations on required_car_parking_space vs booking_status</u>

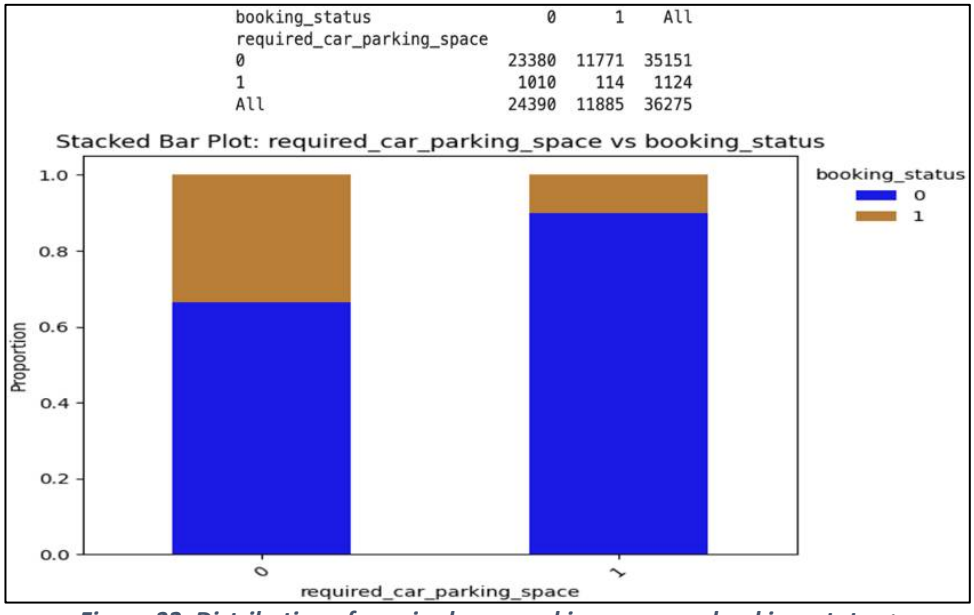


Figure 32. Distribution of required_car_parking_space vs booking_status.:

Customers who did not require car parking (value = 0):

- Total bookings: 35,151 (97%)
- Canceled: 11,771; **cancellation rate ≈ 33.5%**

Customers who did require car parking (value = 1):

- Total bookings: 1,124 (3%)
- Canceled: 114; **cancellation rate ≈ 10.1%**

Overall cancellation rate:

- 11,885 out of 36,275 ≈ **32.8%**

Conclusion:

- Customers who choose a car parking space show different cancellation behaviour.
- **Customers who request parking are much less likely to cancel (about 10%) compared to those who do not (about 33.5%).**
- This suggests that customers who want parking are more certain or committed about their stay. They likely include local travellers or business travellers who plan their trips with more confidence.

2.3 Multivariate Analysis

We will now check the customers who travelled with their families including children to analyse the impact on booking status.

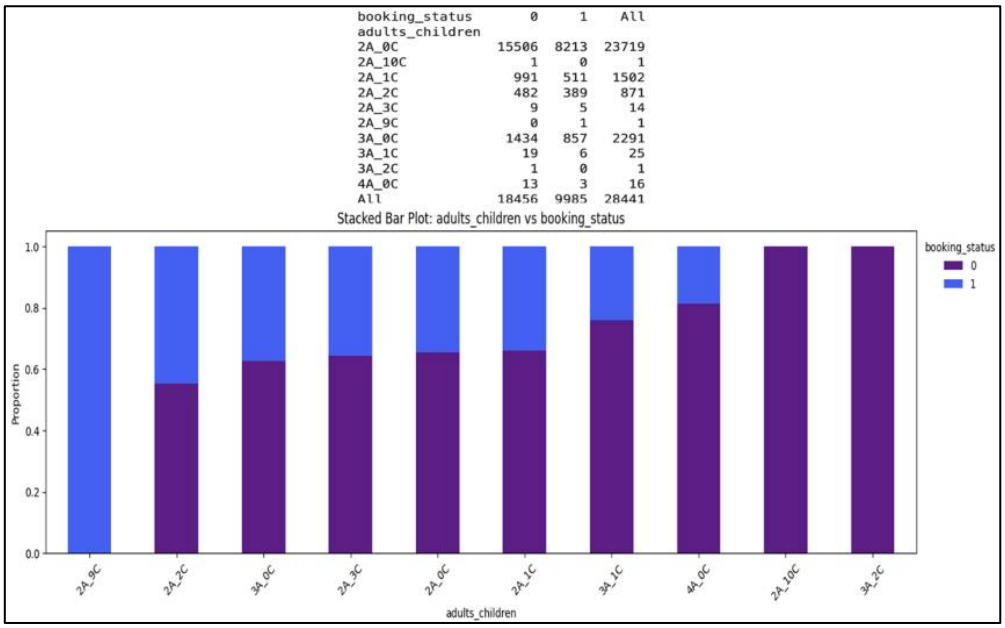


Figure 33. Distribution of adults & children vs booking_status.

Observations

Note: We encoded the plot with **A- adult, C-children.**

Most Common Booking Type:

- The **largest group is 2 Adults, 0 Children** (2A_0C) with 23,719 bookings.
- Around **65% of these were not canceled** (booking_status = 0), while about **35% were canceled**.

Families with Children (2A_1C, 2A_2C, 2A_3C):

- 2A_1C: 1,502 bookings, about **66% not canceled**.
- 2A_2C: 871 bookings, around **55% not canceled**.
- 2A_3C: 14 bookings, **64% not canceled**.

- As the number of children increases, the cancellation rate tends to go up, especially with more than one child.

Families with Children (>=9):

- This group is **rare and has very few bookings** (e.g., 2A_9C, 2A_10C) with one or zero records, making it not useful for analysis.

3 Adults with Children (3A_1C, 3A_2C):

- These have **very low booking numbers**. 3A_1C has only 25 records.

Families with No Children (e.g., 3A_0C, 4A_0C):

- These show higher stability in bookings. For example, 3A_0C has about 63% not canceled, similar to 2A_0C.

Conclusion

- **Family size of 2 to 3**, typically adult couples or couples with one child, shows the **lowest cancellation rates.**
- **Both no_of_adults and no_of_children affects booking status.**
- Family size of 4, usually consisting of **2 adults and 2 children**, is **more likely to cancel**. This indicates possible logistical or cost concerns for larger families.
- Larger groups of 5 or more are too rare in the data for reliable analysis on booking cancellation.

We will now observe if arrival year & arrival month have impact on average price per room.

	arrival_year	arrival_month	2017	2018
0		1	0.000000	74.749763
1		2	0.000000	80.733644
2		3	0.000000	91.883698
3		4	0.000000	101.998801
4		5	0.000000	113.810677
5		6	0.000000	111.966500
6		7	84.890468	115.448092
7		8	91.091400	121.214759
8		9	103.335185	122.905831
9		10	92.032002	111.445723
10		11	72.339104	96.539721
11		12	75.430463	94.947482

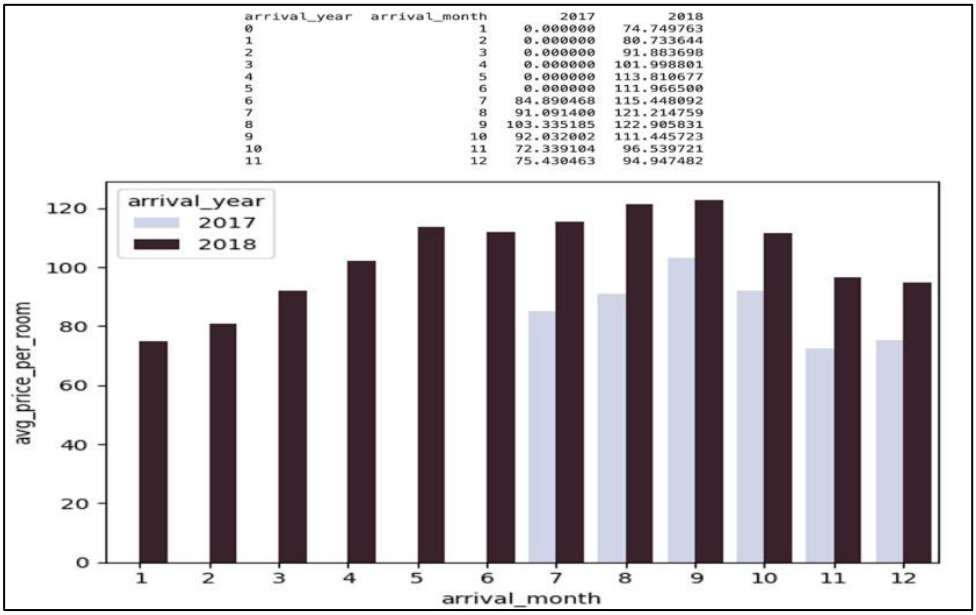


Figure 34. Distribution of arrival month & year vs avg_price_per_room.

Observations

- In **2017**, booking data is only available from July to December. During these months, the average price per room ranged from about **€72 to €103**. The **highest price occurred in September** at €103.3, while November had the lowest price at €72.3.

- In **2018**, booking data is available for all 12 months. Prices rose steadily over the months, starting at **€74.7** in January and **reaching a peak of €122.9 in September**. The second-highest prices were in **August at €121.2 and in May at €113.8**.
- For the overlapping months from July to December, prices in 2018 were much higher than in 2017. The difference was between €15 and €29 more per room.

Conclusion

- There is a clear trend of **rising average room prices from 2017 to 2018**, particularly during peak months like August and September.
- Both the **month of arrival and the year have a strong impact on room pricing**, indicating that **room rates are higher during busy months** and have likely increased in 2018.
- This might be due to seasonal demand, inflation, or changes in pricing strategy by the hotel management.

Observations on type_of_meal_plan , avg_price_per_room w.r.t booking status to analyse if room pricing, type of meal impacts booking status.

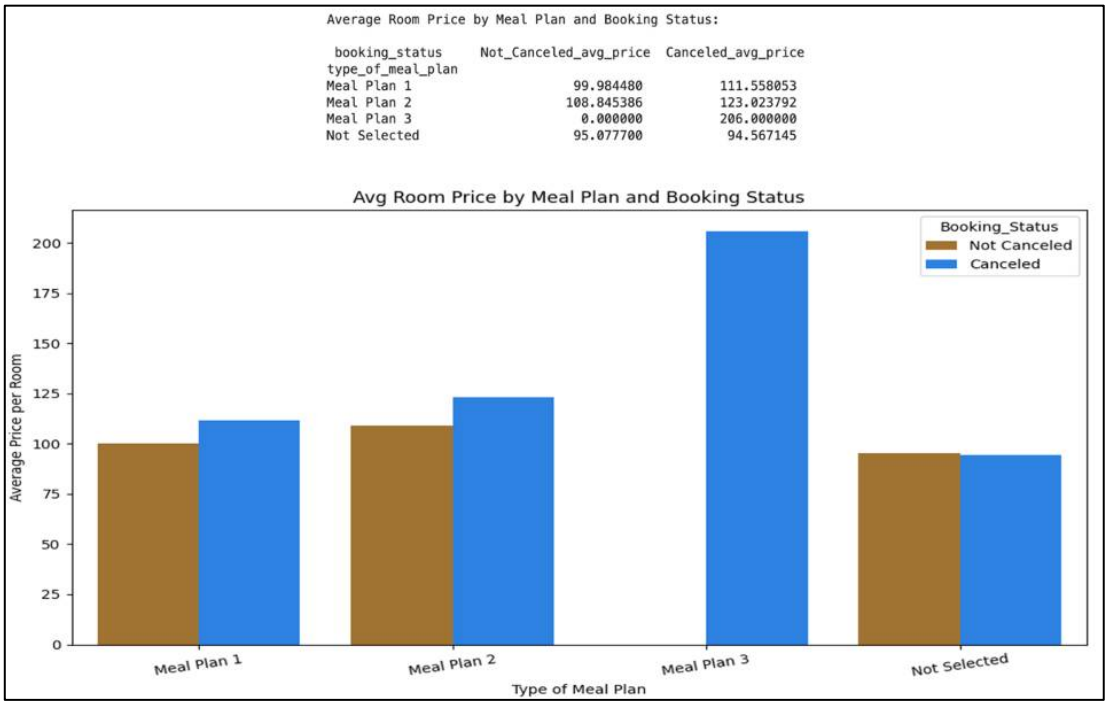


Figure 35. Distribution of average room price & meal plan vs booking status.

Observations

- **Meal Plan 2** has the **highest average room price** for both canceled bookings (123.02 Euros) and not canceled bookings (108.84 Euros). This suggests it's a premium plan.
- **Meal Plan 1** is **slightly cheaper, with prices of 99.98 Euros** for not canceled and **111.56 Euros** for canceled bookings. However, it has a **high cancellation rate** at a moderately high price.
- **Meal Plan 3** has **no bookings** for having not cancelled status, showing an average of 0 Euros. The canceled bookings average 206 Euros, which is the highest overall. **This indicates customers may cancel when prices are extremely high and the benefits of this plan are not clear**.

- **Not Selected plans** show the **lowest variation** between canceled (94.56 Euros) and not canceled (95.07 Euros) bookings. This suggests stable pricing but possibly lower perceived value or fewer benefits.

Conclusion

- **The type of meal plan and the average price per room do affect booking status**.
- Higher prices are linked to a greater likelihood of cancellation, especially for premium plans or those with unclear value.
- Customers may **choose lower-cost or more familiar meal plans** to reduce risk and boost their booking confidence.

Observations on no_of_previous_cancellations, no_of_previous_bookings_not_canceled vs booking_status.

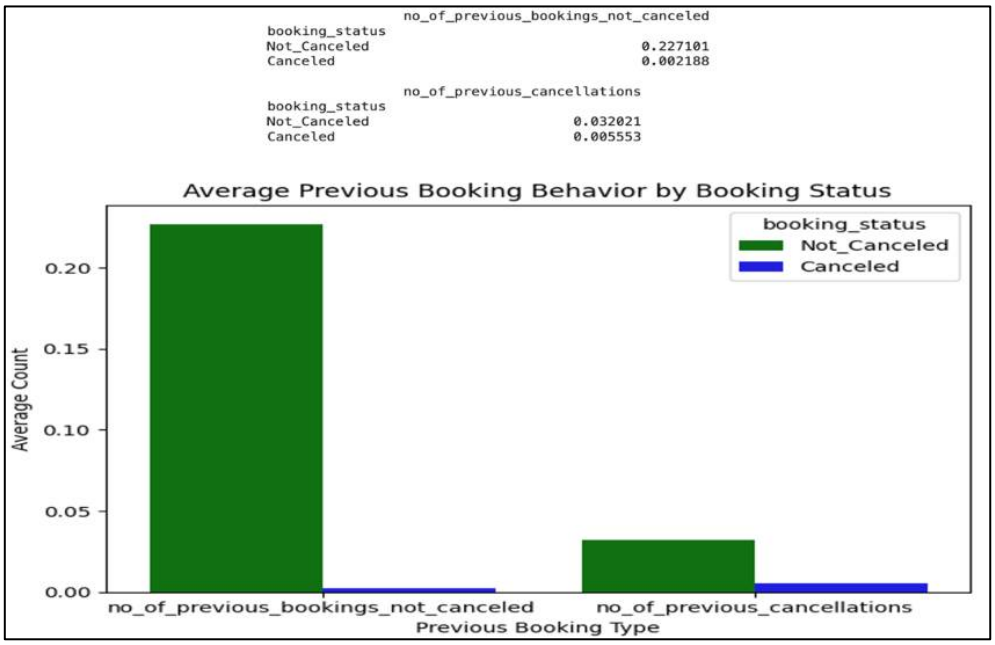


Figure 36. Distribution of previous booking type vs booking_status.

Observations

Customers who did not cancel their bookings (Not_Canceled) had:

- An average of about **0.23** previous bookings not canceled.
- An average of about **0.03** previous cancellations.

Customers who canceled their bookings (Canceled) had:

- An average of about **0.002** previous bookings not canceled.
- An average of about **0.006** previous cancellations.

Conclusion

- **Customer's past booking behaviour clearly affects their cancellation status**.
- **Those with a history of successful bookings are more likely to complete their current booking**.
- On the other hand, customers with very few or no prior successful bookings and a slightly higher cancellation rate tend to cancel again.

2.4 Answering Key Questions

2.4.1 Q1. What are the busiest months in the hotel?

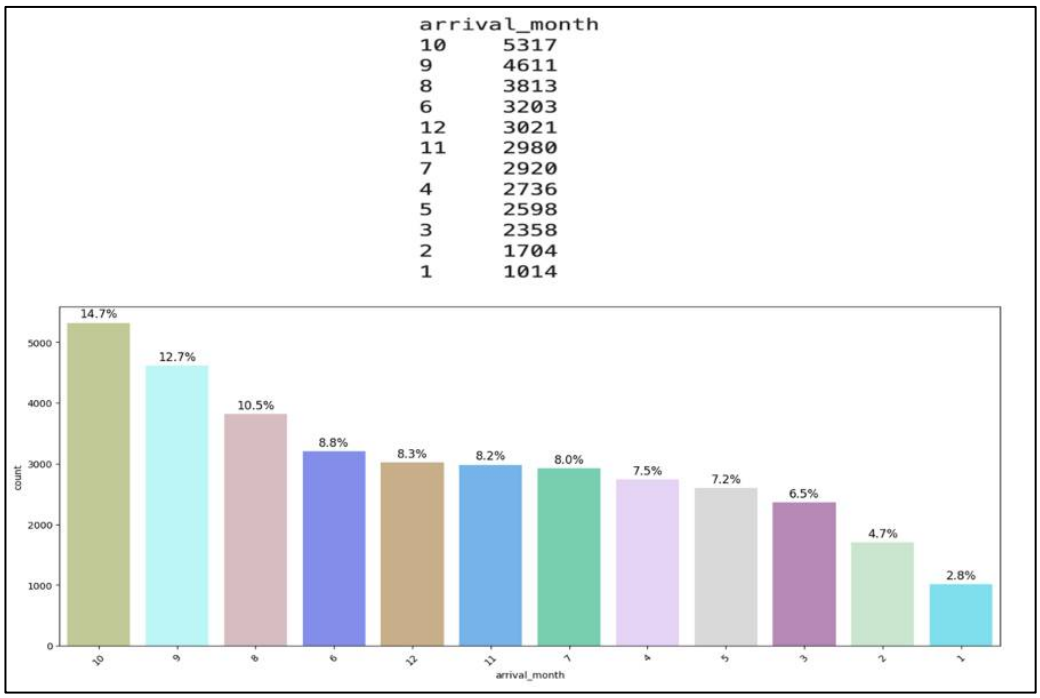


Figure 37. Key Q1. Plot.

- The plot suggest **October (Month 10) is the busiest month**, making up 14.7% of annual bookings with 5,317 bookings. Likely due to ideal weather.
- **September (12.7%) and August (10.5%) follow as the next busiest months**.
- From **August to December, we see about 55% of total bookings: Aug (10.5%), Sep (12.7%), Oct (14.7%), Nov (8.2%), Dec (8.3%)**. likely driven by year-end holidays, including Thanksgiving, Christmas, and New Year, keep demand steady despite the start of winter. This highlights opportunities for premium festive packages.
- **January to March is the slowest quarter, accounting for only 14% of bookings**: Jan (2.8%), Feb (4.7%), Mar (6.5%). This could be due to harsh winter season.
- **June (8.8%) performs better than July (7.2%)** and is close to December (8.3%), even though it is not a traditional peak month. Customers likely take advantage of shoulder-season pricing and avoid the peak crowds in July.
- **April (7.5%) and May (7.2%)show moderate demand**. They exceed the winter months but fall behind the summer and fall months. Spring weather encourages travel, but demand stays below the peaks of summer and fall. This indicates that there is potential to increase demand during this shoulder season with campaigns that focus on nature, like spring blooms and hiking.
- **Thus, we can conclude most busiest month is October (month 10th) making up 14.7% of annual bookings.**

2.4.2 Q2. Which market segment do most of the guests come from?

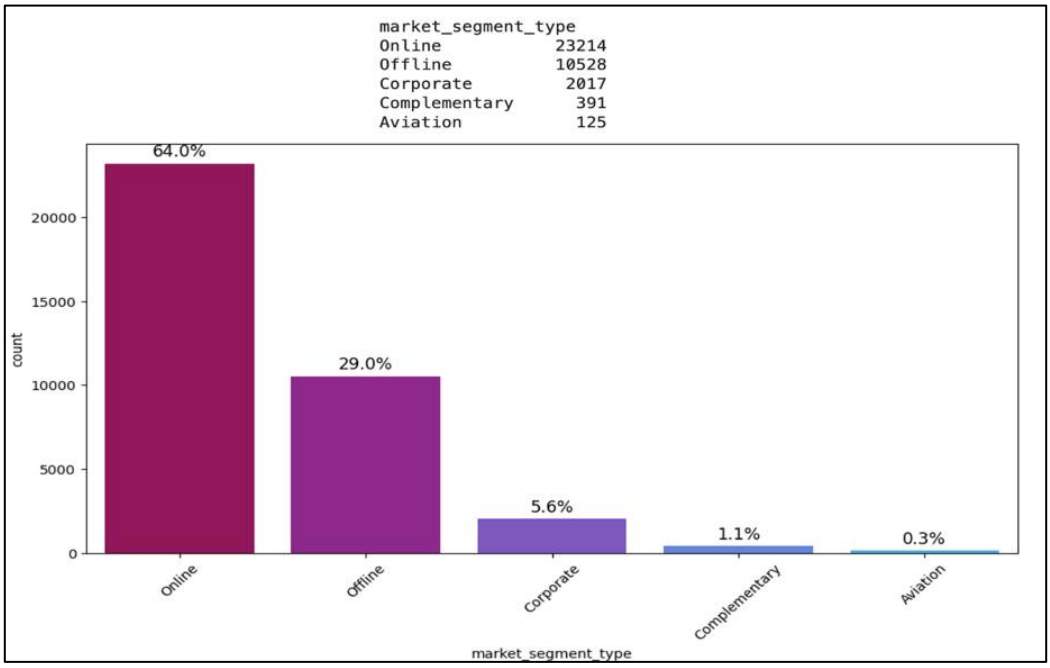


Figure 38. Key Q2. Plot.

- **Online booking dominates (64.0%),** indicates that digital booking platforms, such as OTAs and hotel websites, are the main sales channel. This shows that customers prefer to book online.
- **Offline booking is significant (29.0%)**, suggesting walk-ins and phone bookings still matter, indicating a demand for traditional service, especially for last-minute or loyalty-driven customers.
- **Corporate booking is niche (5.6%)** as there might be some business travel partnerships, but they are not fully used. This presents an opportunity for growth in B2B markets.
- **Complementary (1.1%) & Aviation (0.3%) are marginal**. The free stays for staff and airline crew contracts add little revenue, showing that these non-revenue segments are meant to be kept small.
- **Therefore, we can conclude most guests come from online (64%) market segment type**.

2.4.3 Q3. Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?

market_segment_type	sum	mean	median	min	max
Aviation	12588.00	100.704000	95.0	79.0	110.0
Complementary	1228.43	3.141765	0.0	0.0	170.0
Corporate	167232.98	82.911740	79.0	31.0	220.0
Offline	964708.84	91.632679	90.0	12.0	540.0
Online	2605930.63	112.256855	107.1	0.0	375.5

Figure 39. Key Q3. Plot.

Online Dominates Revenue (69.5%):

- **Total Average Price per room: €2,605,930.63**
- Inference: Digital channels, like OTAs (Online Travel Agencies) and hotel websites, are crucial for profitability since they account for almost 70% of revenue. Guests prefer booking on their own for convenience. However, high OTA commissions can cut into profits.

Offline Contribution is Significant (25.7%) :

- **Total Average Price per room: €964,708.84**
- Inference: Walk-ins and phone bookings are still important, making up more than 25% of revenue. This group likely includes loyal customers or those booking last minute to avoid online fees.

Corporate Underperforms (4.5%):

- **Total Average Price per room: €167,232.98**
- Inference: Even with contracts in place, corporate clients contribute very little. The low volume of bookings in 2017 and possible discounts limit revenue opportunities.

Aviation & Complementary Are Negligible (<0.4%):

- **Total Average Price per room: €12,588.00**
- Inference: Airline crew stays are steady but not very valuable, with only 125 bookings. They help fill rooms but do not significantly add to revenue.

Complementary:

- **Total Average Price per room: €1,228.43**
- Inference: Free stays for staff and comps are kept minimal to maintain revenue integrity.

Conclusion:

- **Yes, room prices vary greatly across different market segments**.
- **Online bookings have the highest average price per room at around €112.25**.
- This reflects strong demand and flexible pricing strategies on digital platforms.

- In comparison, **corporate and offline bookings tend to have lower prices**, likely because of negotiated rates or loyalty discounts. The complementary and aviation segments contribute very little to overall revenue.

2.4.4 Q4. What percentage of bookings are canceled?

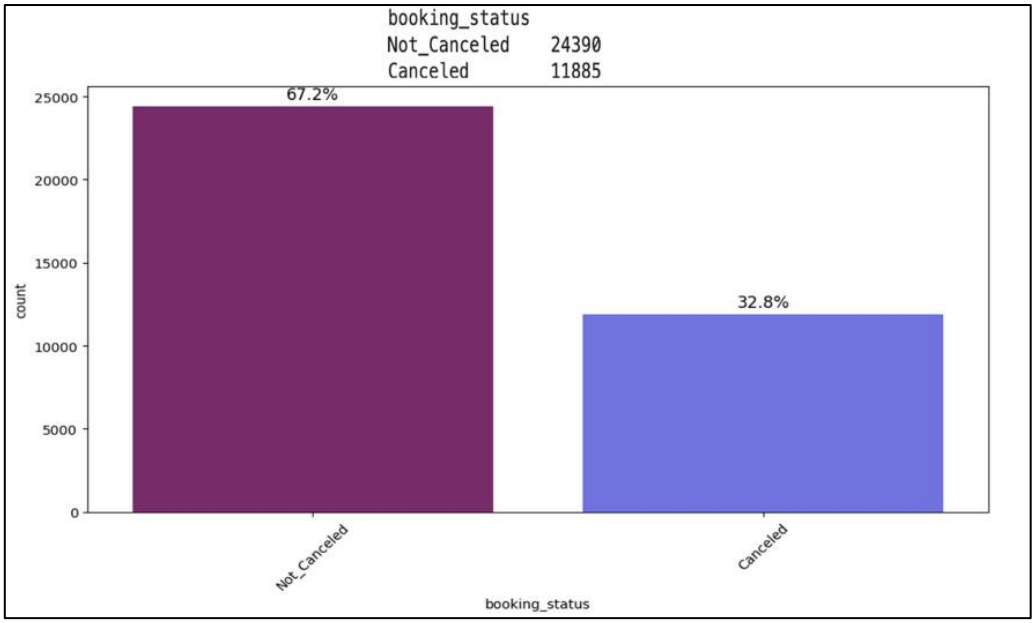


Figure 40. Key Q4. Plot.

- **Total Cancellations**: 11,885 bookings were cancelled out of 36,275 total bookings, which is **32.8% of all bookings**.
- **Total Non-Cancellations**: 24,390 bookings were completed, making up **67.2% of the total**.
- **Insights:**
    - A cancellation rate of 1 in 3 presents a serious risk to revenue and occupancy planning.
    - This may be partly due to long lead times, with a median of 57 days, giving customers more chances to cancel.
    - Flexible cancellation policies might also encourage risk-free bookings, leading to more no-shows.
    - Cancellations can leave rooms empty, create staffing inefficiencies, and result in lost revenue, especially during high-demand periods.
    - Therefore, we can say **32.8% of total bookings were cancelled**.

2.4.5 Q5. Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?

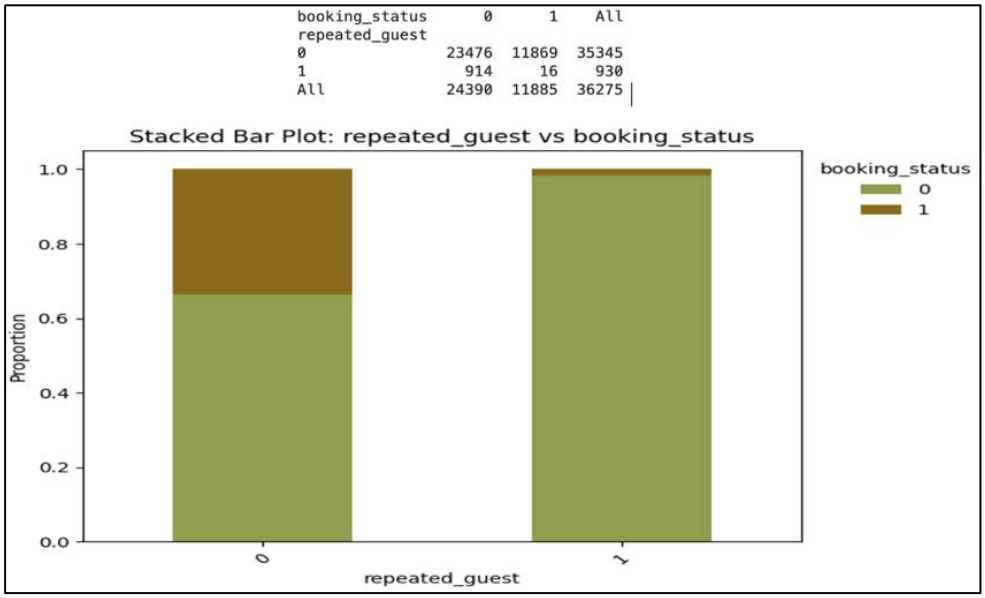


Figure 41. Key Q5. Plot.

Repeated Guests (repeated_guest = 1):

- Total: 930 bookings
- Cancelled bookings (1): 16, about 1.7% cancellation rate
- The vast majority went ahead with their bookings.
- Repeated guests have almost no blue section for cancellations, showing that nearly all their bookings are completed.

Non-Repeated Guests (repeated_guest = 0):

- Total: 35,345 bookings
- Cancelled bookings (1): 11,869, about 33.6% cancellation rate
- A significant number of non-repeated guests canceled their bookings.
- Non-repeated guests show a much higher share of cancellations compared to repeated group.

Conclusion:

- **Repeated guests cancel bookings around 1.7%** and **rarely which shows strong brand loyalty and intent to book**.
- In contrast, 1 in 3 non-repeated guests cancel, indicating lower commitment or higher uncertainty.
- This confirms that repeated guests are valuable and dependable customers.
- Therefore, hotels should prioritize and reward them through loyalty programs or perks.

2.4.6 Q6. Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?

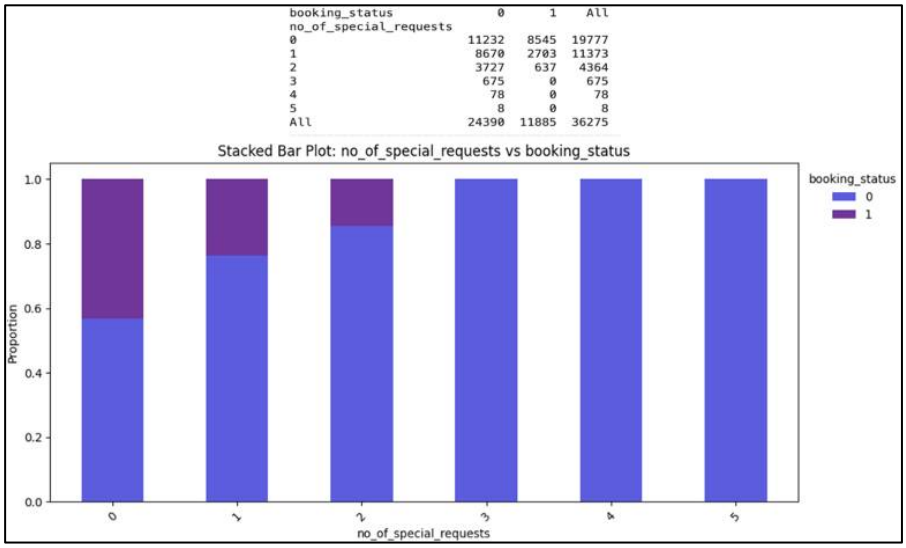


Figure 42. Key Q6. Plot.

0 Special Requests:

- The highest number of bookings was about 19,777.
- The cancellation rate was the highest, with around 43% canceled (8,545 / 19777).
- This group showed the most imbalance toward cancellations.

1 Special Request:

- The cancellation rate was moderate at about 24% canceled (2,703 / 11373).
- This is better than 0 requests, but still noteworthy.

2 Special Requests:

- There was a sharp drop in cancellations to about 15% (637 / 4,364).
- The acceptance rate improved significantly.

3 to 5 Special Requests:

- No cancellations were observed.
- All these bookings were accepted by customers, meaning none were canceled.
- The booking counts are very low (ranging from 675 to 8), but the trend remains consistent.

Conclusions:

- **Yes, special requirements affect booking cancellations as higher number of special requests strongly relates to lower cancellation rates**.
- **Guests with three or more requests never cancel bookings as per the data**.
- This indicates that guests making special requests are more serious about their bookings.
- Hotels could use this information to prioritize and personalize service for these guests, which could help further reduce cancellations.

2.5 Overall EDA Insights

Lead Time is the Strongest Predictor:

- Correlation (r = 0.44) shows that bookings made far in advance are much more likely to be canceled.
- Canceled bookings have longer lead times (about 100+ days), likely because of changes in plans or price searching.
- Insight: Longer lead times significantly increase cancellation risk.

Special Requests Reduce Cancellation:

- Correlation (r = -0.25) indicates that bookings with special requests are less likely to be canceled.
- Guests with three or more requests had a 0% cancellation rate.
- Insight: Personalization helps secure bookings and shows strong customer intent.

Booking Year (2018) Had a Spike in Cancellations:

- Cancellations jumped from 14.75% in 2017 to 36.7% in 2018.
- In 2018, there was:
- High OTA usage (82% of bookings),
- Longer lead times,
- 97% of guests were first-timers.
- Insight: The 2018 expansion attracted less reliable, price-sensitive new customers.

Market Segment Matters:

- Online bookings had a 36.5% cancellation rate, which is the highest.
- Corporate bookings had only a 10.9% cancellation rate.
- Insight: Direct and business channels are more reliable than online aggregators.

Repeat Guests Are Highly Reliable:

- Repeated guests cancel only 1.7% of the time, compared to 33.6% for new guests.
- Insight: Loyalty and retention greatly reduce cancellations.

Room Type Influences Cancellation:

- Room_Type 6 has a 42% cancellation rate, while Room_Type 1 is at 32%.
- This suggests a mismatch between expectation and price or quality.
- Insight: Certain room categories attract more unpredictable customers.

Meal Plan and Price Impact Cancellations:

- Meal Plan 2 (premium) and Meal Plan 3 (high price, low value) had higher cancellation rates.
- Insight: High prices without clear value lead to more cancellations.

Car Parking Request = Low Cancellation:

- Only 10.1% of those who opted for car parking canceled, compared to 33.5% for those who didn't.
- Insight: A parking request suggests serious, planned stays.

Family Structure and Cancellations:

- Couples (2 adults, 0 children) have the lowest cancellation rate.
- Families with two or more children show higher cancellation rates.
- Insight: Larger families may encounter cost or logistical issues that lead to cancellations.

Past Behaviour Predicts Future:

- The Not_Canceled group had more past bookings and fewer past cancellations.
- Insight: Customers with a good track record are much more reliable.

Final Features affecting booking status:

- lead_time, no_of_special_requests, repeated_guest, room_type_reserved, type_of_meal_plan, market_segment_type, and past cancellations affects the booking status.
- Categorical variables like arrival_month and year can improve seasonality detection, and help in improving business revenue.

3 DATA PROCESSING

3.1 Data Cleaning

Checking Missing values

Missing values are:

	0
Booking_ID	0
no_of_adults	0
no_of_children	0
no_of_weekend_nights	0
no_of_week_nights	0
type_of_meal_plan	0
required_car_parking_space	0
room_type_reserved	0
lead_time	0
arrival_year	0
arrival_month	0
arrival_date	0
market_segment_type	0
repeated_guest	0
no_of_previous_cancellations	0
no_of_previous_bookings_not_canceled	0
avg_price_per_room	0
no_of_special_requests	0
booking_status	0

Table 7. Missing Values.

There are no missing values in our dataset and therefore, we do not require any imputing techniques to treat them. Thereby, preserving the data integrity and reliability.

Checking Duplicate Values

There are no duplicate values in our dataset. Therefore, our dataset is reliable to do further processing.

3.2 Outlier Detection

Outlier detection is applied only to continuous numerical features where extreme values can meaningfully distort statistical summaries or model performance.

Dropped earlier:

- Booking_ID was removed as it is a unique identifier and not a predictive feature.

Excluded from outlier checks:

- **Binary flags** (repeated_guest, required_car_parking_space, booking_status): Only 0/1 values—no "extreme" values to detect.
- **Small integer counts** (no_of_adults, no_of_children): Limited range (e.g., 0-4) makes IQR-based detection uninformative.
- **Categorical variables** (room_type_reserved, type_of_meal_plan, market_segment_type): Labels rather than quantitative measures.
- **Date parts** (arrival_year, arrival_month, arrival_date): Used for grouping/time slicing, not numeric analysis.

Special note on avg_price_per_room:

- Values of 0€ are valid (e.g., complimentary or promotional stays in the Complementary segment), not data errors, and should be retained.

Therefore, outlier detection is focused on "lead_time", "no_of_weekend_nights", "no_of_week_nights", "no_of_special_requests", "no_of_previous_cancellations", "no_of_previous_bookings_not_canceled", and "avg_price_per_room".

Visualisation of Outlier via Box Plot

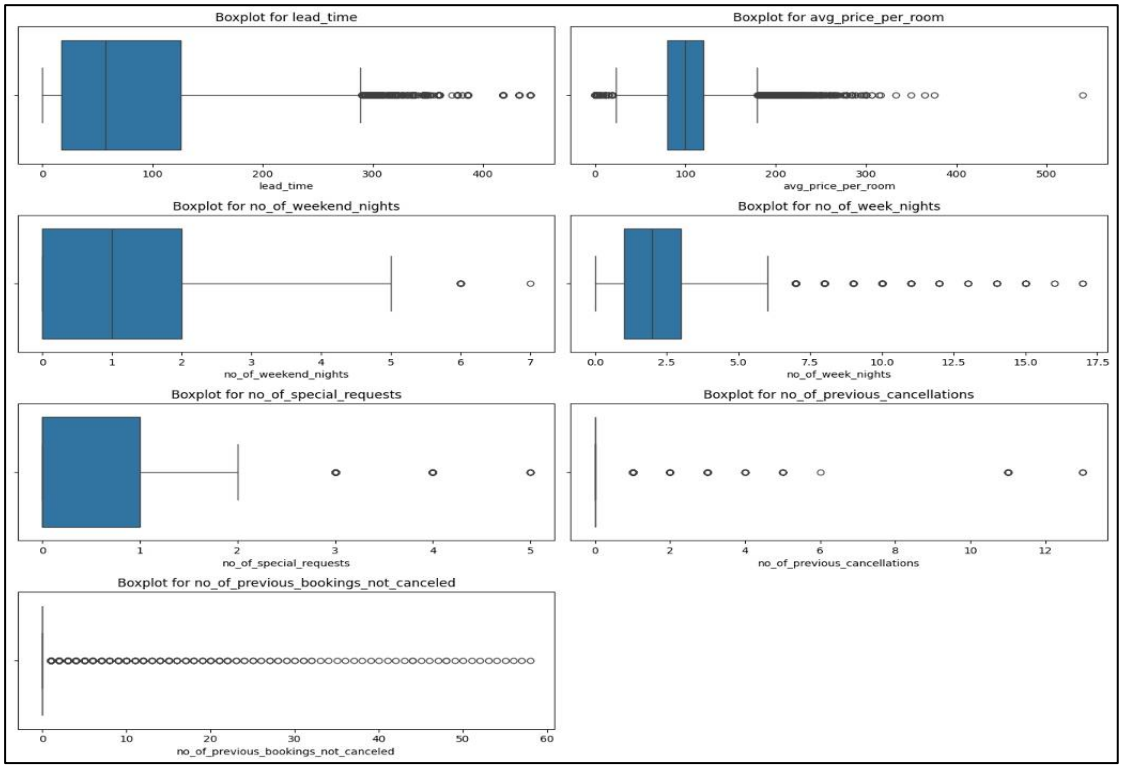


Figure 43. Box Plot for Outlier Detection.

Observations on Box plots:

lead_time

- Median is about 57 days; IQR is roughly 17 to 126 days.
- There is a noticeable right tail with bookings reaching about 443 days. Very long lead times (over 270 days) are rare but still valid.

avg_price_per_room

- Median is around 99 €; IQR is about 80 to 120 €.
- A few zero-price stays (complementary or promo) and high outliers up to around 540 € show premium rates or extreme data.

no_of_weekend_nights

- Median is 1 night; IQR is 0 to 2 nights.
- Some stays of 6 to 7 weekend nights are outliers, likely due to special multi-weekend bookings.

no_of_week_nights

- Median is 2 nights; IQR is 1 to 3 nights.
- Rare long weekday stays (7 to 17 nights) are outliers, possibly extended business visits.

no_of_special_requests

- Median is 0; IQR is 0 to 1.
- Outliers at 3 to 5 requests, though few in number, indicate highly personalized bookings.

no_of_previous_cancellations

- Median is 0; IQR is 0 to 0.

- A few guests have up to about 13 past cancellations. These habitual cancellers may require a special policy.

no_of_previous_bookings_not_canceled

- Median is 0; IQR is 0 to 0.
- Many outliers (10 to 58 successful past stays) suggest a small loyal segment of valuable repeat customers.

Validating Outliers via IQR Method

```
======================== Outlier Detection Summary ==========================
                                    lower_bound  upper_bound  outlier_count
lead_time                              -146.50       289.50         1331.0
avg_price_per_room                       20.75       179.55         1151.0
no_of_weekend_nights                     -3.00         5.00           21.0
no_of_week_nights                        -2.00         6.00          324.0
no_of_special_requests                   -1.50         2.50          761.0
no_of_previous_cancellations              0.00         0.00          338.0
no_of_previous_bookings_not_canceled      0.00         0.00          812.0

                                    outlier_prcnt
lead_time                                    3.67
avg_price_per_room                           3.17
no_of_weekend_nights                         0.06
no_of_week_nights                            0.89
no_of_special_requests                       2.10
no_of_previous_cancellations                 0.93
no_of_previous_bookings_not_canceled         2.24
```

Figure 44. Outlier Detection via IQR Method.

Observations:

lead_time

- **Outlier %: 3.67%**
- Longer lead times are strongly linked to booking cancellations, as shown in the exploratory data analysis.
- Treatment: This feature has high predictability power. Thus, we will treat it by using **capping or winsorization for high values**, for example, those above the 95th percentile. This approach reduces skew and extreme influence while keeping useful variance.

avg_price_per_room

- **Outlier %: 3.17%**
- Prices were slightly higher for canceled bookings. Also, zero-price rooms likely indicate complimentary bookings, which should be excluded during treatment.
- Treatment: This feature has moderate predictability power. Thus, we will treat it first by **exclude records where avg_price_per_room = 0 for outlier detection. After that, cap high values**. Winsorization avoids skewing model training while maintaining interpretability.

no_of_weekend_nights

- **Outlier %: 0.06%**
- This feature has low business relevance. Most bookings are for 1 to 2 nights, with very few long weekend stays.

- Treatment: **We will retain all values**. Outliers are valid long weekend vacations and do not impact overall modelling.

no_of_week_nights

- **Outlier %: 0.89%**
- This has minor influence on cancellations.
- Treatment: **We will retain the values** and no removal is required.

no_of_special_requests

- **Outlier %: 2.10%**
- It has moderate business relevance. More special requests were noted in non-canceled bookings.
- Treatment: **We will keep the values**, such as those greater than 4 or 5, since they are rare. These reflect customer intent and are important for modelling.

no_of_previous_cancellations

- **Outlier %: 0.93%**
- It has low business relevance. Most guests had zero cancellations, but a few had many. This is important for identifying chronic cancelers.
- Treatment: **We will retain the data**, as no treatment required.

no_of_previous_bookings_not_canceled

- **Outlier %: 2.24%**
- It has low to moderate business relevance. There are very loyal customers with many past bookings, but this is not the most significant factor.
- Treatment: **We will keep the values**. They represent true loyalty.

3.3 Outlier Treatment

We treated the following features:

- **lead_time was capped at the upper whisker** (Q3 + 1.5xIQR) because very long lead times can be valid but may have an outsized impact on modelling.
- **avg_price_per_room is considered only for non-zero prices**. This assumes that zero-priced rooms are intentional, such as complimentary or promotional stays.
- **Outliers were capped/winsorized rather than removed** to maintain the size and integrity of the dataset for modelling.

Verification Post-Treatment of Outliers using IQR Method

```
lead_time: 0.0% outliers remaining
avg_price_per_room: 0.2% outliers remaining
```

Table 8. Remaining Outliers Post Capping.

We can observe that after the treatment there are **0% outliers in lead_time**, while only **0.2% in avg_price_per_room**. We can ignore these 0.2% outliers as they are too low and would not affect the

modelling analysis. Thus, we can say we successfully reduced the outliers without losing important data. We can now move forward with the feature engineering and scaling steps with confidence.

Verification Post-Treatment of Outliers via Plot:

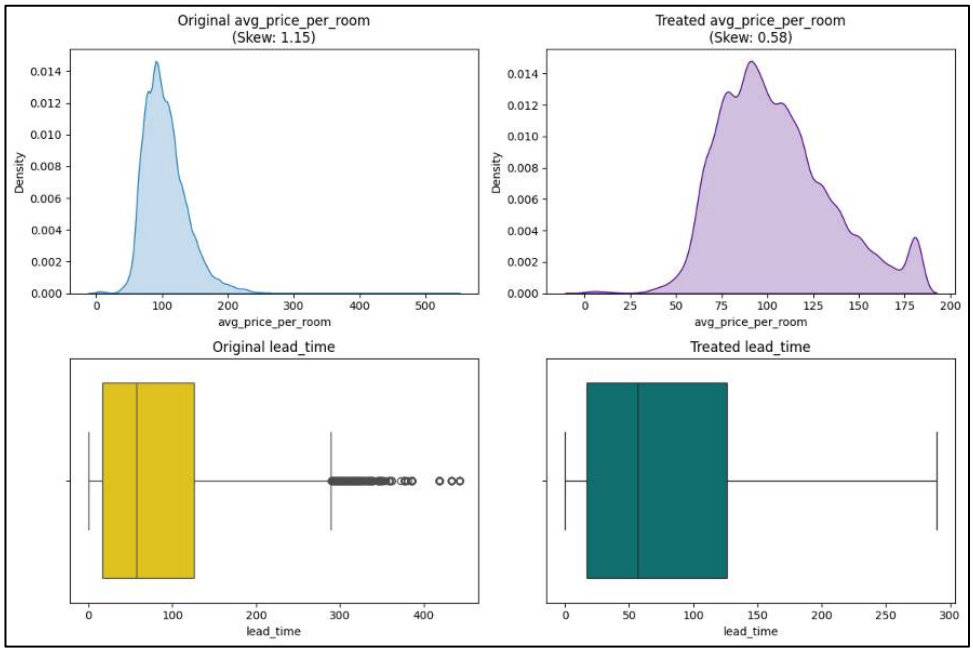


Figure 45. Post-Outlier Treatment Verification Plot.

Observations:

- We can see from the plots before and after treatment that there are no outliers for lead_time, as shown in the box plot.
- The data distribution for **avg_price_per_room shows a slight normal distribution** compared to the plot before treatment, which earlier had right skewness.

3.4 Feature Engineering

Checking the booking with no_of_adults = 0

During our data inspection, we noticed there are zero number of adults in our dataset and a booking with 0 adults is not valid in a real-world hotel scenario. This could be either error or incorrect making it irrelevant for the predictions. We will analyse the rows where no_of_adults = 0.

```
Number of bookings with 0 adults: 139
Percentage of total data: 0.38%
```

Table 9. Bookings with zero adults.

- We can see there are just **0.38% of total data where no_of_adults = 0**. Thus, **we can safely drop these rows as removing them would not affect our model quality due to negligible data loss**.

After dropping rows where no_of_adults = 0

```
Number of bookings with 0 adults: 0
```

Table 10. Post-dropping rows with zero adults.

- We have successfully dropped the no_of_adults with zero value.

Creating new features family_size

As per our observations from EDA, we know **Family size (no_of_adults + no_of_children)** may impact booking behaviour (larger groups may cancel less due to planning). We will create a new feature family size which includes no_of_adults + no_of_children in our dataset.

```
Unqiue values in family_size are:
[ 2  1  3  4  5 12 10 11]

Values counts for family_size are:

family_size
1      7551
2     23809
3      3846
4       912
5        15
10        1
11        1
12        1
Name: count, dtype: int64
```

	family_size
count	36136.000000
mean	1.949939
std	0.651433
min	1.000000
25%	2.000000
50%	2.000000
75%	2.000000
max	12.000000

repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status	family_size
0	0	0	65.00	0	0	2
0	0	0	106.68	1	0	2
0	0	0	60.00	0	1	1
0	0	0	100.00	0	1	2
0	0	0	94.50	0	1	2

Table 11. New feature family_size in dataset & its values.

Observations on new feature family_size:

- We created the new feature successfully and checked its statistical summary , value counts and its unique values as shown in the image.
- ***Size 2:*** Most common, with 23,809 bookings, which is about 65.6%. This size is for couples or small families.
- ***Size 1:*** Second most common, with 7,551 bookings, roughly 20.8%. This size caters to solo travellers. Together, these account for 86.4% of all bookings.
- ***Sizes 10, 11, and 12*** each have only 1 booking, about 0.003% each.
- ***Size 5*** has just 15 bookings, or 0.04%.
- ***Sizes of 10*** or larger are very unlikely for standard hotel rooms. This suggests possible data errors or group bookings divided among rooms.
- ***Extreme values (10, 11, and 12)*** together represent only 0.008% of the data.

Visualising family_size Distribution & Outliers

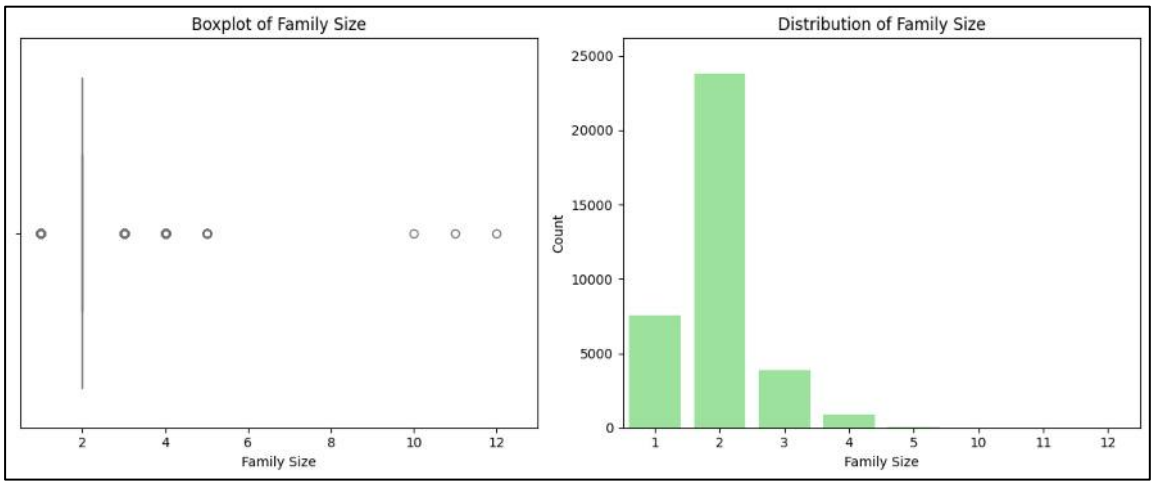


Figure 46. Family_size Plots.

Observations:

- Data distribution is right skewed.
- The majority of bookings are for small families, typically 1 to 3 members, with a **family size of 2 being the most common**.
- Some bookings have very large family sizes, such as **10, 11, or 12. These appear as outliers** in the boxplot.
- For our **treatment decision, we choose not to address these outliers**. They probably represent **valid group bookings or large family trips** instead of data errors.
- Removing or limiting these outliers could lead to losing important business cases, such as bulk reservations.

Creating new feature total_stay

We know no_of_weekend_nights and no_of_week_nights individually showed low business importance with has low outlier percentages. However, together they represent the total duration of stay, which is more informative and business relevant. This can help us in modelling because longer stays may correlate with cancellation likelihood or booking intent. Therefore, we will create a new feature **total_stay (no_of_weekend_nights + no_of_week_nights).**

no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status	family_size	total_stay
0	0	65.00	0	0	2	3
0	0	106.68	1	0	2	5
0	0	60.00	0	1	1	3
0	0	100.00	0	1	2	2
0	0	94.50	0	1	2	2

	total_stay
count	36136.000000
mean	3.013864
std	1.785913
min	0.000000
25%	2.000000
50%	3.000000
75%	4.000000
max	24.000000

```
total_stay
0          78
1        6582
2        8452
3       10012
4        5863
5        2571
6        1027
7         970
8         179
9         111
10        109
11         38
12         24
13         17
14         32
15         31
16          6
17          5
18          3
19          6
20         11
21          4
22          2
23          2
24          1
Name: count,
```

```
Unqiue values in total_stay are:
[ 3  5  2  4  1  6  7 14  0  8 15 10 21 13 12  9 19 20 16 11 17 23 18 24
 22]
```

Table 12. New feature total_stay in dataset & its values.

- We can observe from the above image that new feature total_stay has been created.
- We analysed its statistical summary, value counts and unique value as shown in above image.
- Total_stay duration range is from 0-24 nights.
- **Average total stay for a customer is ~3 nights**.

Visualising total_stay Distribution & Outliers

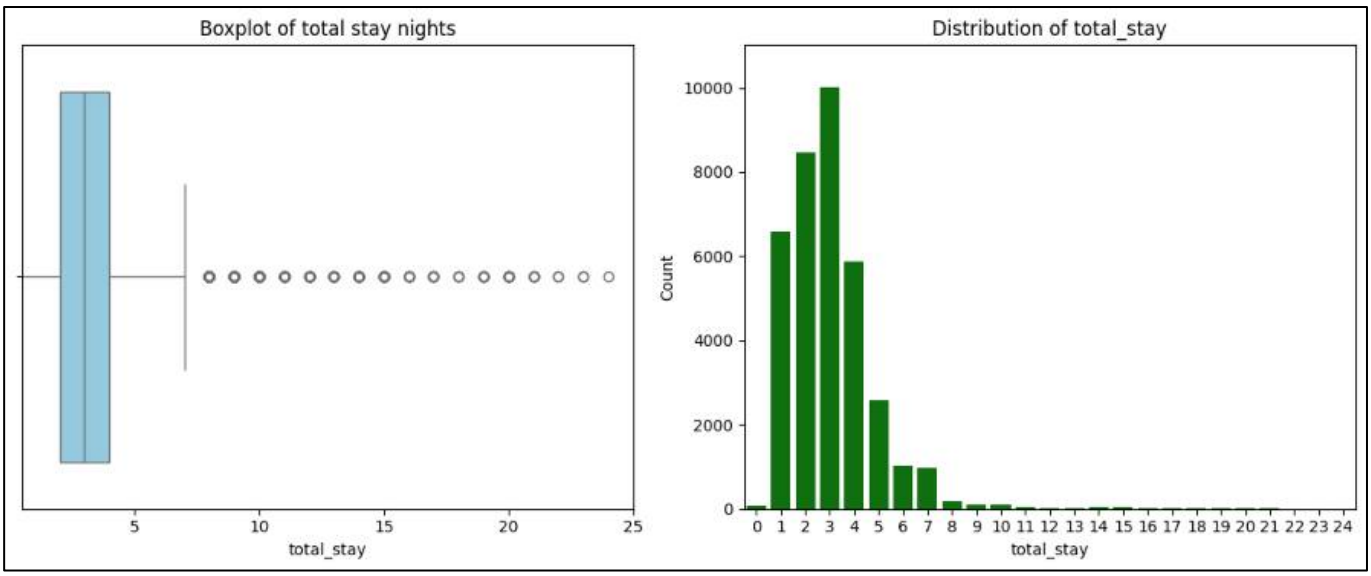


Figure 47. total_stay plots

Observations:

- Range of total stay is **0 to 24 nights**.
- **Most common stays: 2 to 4 nights** (IQR = [2, 4])
- Outliers: **Very long stays (e.g., 15 to 24 nights) are rare**.

- **0-night stays (78 records) are unusual and may suggest data entry mistakes, canceled or same-day check-ins. We need to investigate week and weekend nights individually to check if what percentage of rows have stay nights, based on it we will see if we have to drop or impute, or leave as is**.
- Very long stays **(15+ nights)** are uncommon and may represent group or corporate bookings. These are **valid but rare and should not be removed unless proven to be incorrect**, and suggests customer might be on long vacation or work sponsored trips. Therefore, we will not treat them as these may be real data, removing them can distort our data analysis.

Investigating rows where week nights, weekend nights and total_stay are 0

```
Bookings with both no_of_weekend_nights and no_of_week_nights = 0 is 78
Percentage of total data: 0.22%
```

Table 13. Week & Weekend nights investigation.

Observations:

- During pre-processing, we found **78 bookings (0.22% of the total data**) **where both no_of_weekend_nights and no_of_week_nights were zero, leading to a total_stay value of 0**.
- **These entries likely represent same-day check-in and check-out bookings, day-use reservations, or very short stays**.
- Since these bookings are still valid and could also be canceled, they are important for our goal of predicting booking cancellations.
- Additionally, because they are such a small portion of the data, they do not significantly impact model quality.
- Therefore, **we chose to keep both the individual zero-night values and the total_stay = 0 values without any imputation or deletion**.
- This choice allows our dataset to reflect the real-world variety of bookings and avoids discarding meaningful edge cases that might affect cancellation behaviour.

Dropping Redundant Variables

We will be dropping redundant variables like **'no_of_adults', 'no_of_children', 'no_of_weekend_nights', 'no_of_week_nights'** as we have already created features (family_size & total_stay) which have their values, as keep them will create multi-collinearity issue.

	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arrival_year	arrival_month	arrival_date	market_segment_type	repeated_guest
0	Meal Plan 1	0	Room_Type 1	224.0	2017	10	2	Offline	0
1	Not Selected	0	Room_Type 1	5.0	2018	11	6	Online	0
2	Meal Plan 1	0	Room_Type 1	1.0	2018	2	28	Online	0
3	Meal Plan 1	0	Room_Type 1	211.0	2018	5	20	Online	0
4	Not Selected	0	Room_Type 1	48.0	2018	4	11	Online	0

Table 14. Dataset after dropping features.

- **Shape: (36136, 16)**

- We can observe 'no_of_adults', 'no_of_children', 'no_of_weekend_nights', 'no_of_week_nights' columns have been dropped. Previously we dropped Booking_ID.
- Now after dropping those feature, we **have 36136 rows and 16 columns in our final dataset**.

Converting categorical columns to 'category' dtype

```
<class 'pandas.core.frame.DataFrame'>
Index: 36136 entries, 0 to 36274
Data columns (total 16 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   type_of_meal_plan                     36136 non-null  category
 1   required_car_parking_space            36136 non-null  category
 2   room_type_reserved                    36136 non-null  category
 3   lead_time                             36136 non-null  float64
 4   arrival_year                          36136 non-null  category
 5   arrival_month                         36136 non-null  category
 6   arrival_date                          36136 non-null  int64
 7   market_segment_type                   36136 non-null  category
 8   repeated_guest                        36136 non-null  category
 9   no_of_previous_cancellations          36136 non-null  int64
 10  no_of_previous_bookings_not_canceled  36136 non-null  int64
 11  avg_price_per_room                    36136 non-null  float64
 12  no_of_special_requests                36136 non-null  category
 13  booking_status                        36136 non-null  category
 14  family_size                           36136 non-null  int64
 15  total_stay                            36136 non-null  int64
dtypes: category(9), float64(2), int64(5)
```

Table 15. Converted categorical columns to 'category' dtype.

- We converted 9 categorical columns to 'category' dtype.
- This ensures that future encoding operations (like one-hot encoding) treat these variables appropriately.
- Now, we have 9 columns with 'category' dtype.
- The booking_status is already encoded as 0 and 1 in previous section, so we retain it as numeric for modelling.

3.5 Data Preparation for Modelling

One-Hot Encoding

Before splitting the data into train-test we will do one-hot encoding for our independent categorical variables (type_of_meal_plan, room_type_reserved, market_segment_type) except repeated_guest, required_car_parking_space and booking_status as they are already encoded.

Moreover, booking_status is target variable we need not do one-hot encoding for target.

type_of_meal_plan_Meal Plan 3	type_of_meal_plan_Not Selected	room_type_reserved_Room_Type 2	room_type_reserved_Room_Type 3	room_type_reserved_Room_Type 4	room_type_reserved_Rc
0.0	0.0	0.0	0.0	0.0	
0.0	1.0	0.0	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	
0.0	1.0	0.0	0.0	0.0	

Table 16. One-Hot encoded columns.

- We can see we have encoded our dummy variables successfully using one-hot encoding technique.

Train & Test Data Split

- **We will first we need to select our target (booking_status) variable from the data frame and separate it from the independent variables(predictors), then finally split the dataset.**
- **Splitting the data into train and test set:**
  - **X_train**: 70% of feature data used to train the model
  - **X_test**: 30% of feature data used to test (evaluate) the model
  - **y_train**: Target values for X_train
  - **y_test**: Target values for X_test

required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_ca
0.0	116.0	2018.0	2.0	28.0	0.0	0.0	
1.0	1.0	2018.0	1.0	13.0	0.0	0.0	
0.0	93.0	2018.0	5.0	21.0	0.0	0.0	
0.0	121.0	2017.0	8.0	10.0	0.0	0.0	
0.0	147.0	2018.0	8.0	3.0	0.0	0.0	

avg_price_per_room	no_of_special_requests	family_size	total_stay	type_of_meal_plan_Meal Plan 2	type_of_meal_plan_Meal Plan 3	type_of_meal_plan_Not Selected	room_type_reserve
181.225	1.0	2.0	4.0	0.0	0.0	0.0	
181.225	1.0	3.0	2.0	0.0	0.0	0.0	
89.000	0.0	2.0	3.0	0.0	0.0	0.0	
96.300	0.0	2.0	3.0	0.0	0.0	0.0	
82.450	0.0	2.0	5.0	0.0	0.0	0.0	

Table 17. Train & Test Split top 5 rows.

- We successfully split the data into train and test set.

Shape & Percentage of Train & Test Dataset

```
      Shape of train data = (25295, 25)
      Shape of test data = (10841, 25)

Percentage of classes in training set:
booking_status
0     0.672742
1     0.327258
Name: proportion, dtype: float64

Percentage of classes in test set:
booking_status
0     0.67134
1     0.32866
Name: proportion, dtype: float64
```

Table 18.Shape & Percentage of Train & Test dataset.

Observations

- Shape of Training set**: 25,295** rows and **25** columns.
- Shape of Test set: **10,841 rows** and **25** columns.

- ***Percentage of Classes Training set- Target:***
  - There are **67.27%** of not canceled **(0) class**.
  - There are **32.73%** of canceled **(1) of class**.
- ***Percentage of Classes of Test set-Target:***
  - There are **67.13%** of not canceled **(0) class**.
  - There are **32.87%** of canceled **(1) class**.

Conclusion

- Train & Test data have equal number of columns.
- We had seen that around 67.27% of observations belongs to class 0 (not canceled) and 32.87% observations belongs to class 1 (canceled), and this is preserved in the train and test sets.

<u>Scaling the Data</u>

In machine learning, scaling, such as using StandardScaler, brings numerical features onto a similar scale. This helps with model convergence and makes it easier to interpret the results.

	required_car_parking_space	lead_time	arrival_year	arrival_month
0	-0.178610	0.397033	0.468891	-1.756718
1	5.598799	-1.011698	0.468891	-2.081898
2	-0.178610	0.115287	0.468891	-0.781178
3	-0.178610	0.458282	-2.132690	0.194362
4	-0.178610	0.776778	0.468891	0.194362

	arrival_date	repeated_guest	no_of_previous_cancellations
0	1.418119	-0.161245	-0.064178
1	-0.297758	-0.161245	-0.064178
2	0.617377	-0.161245	-0.064178
3	-0.640933	-0.161245	-0.064178
4	-1.441676	-0.161245	-0.064178

	no_of_previous_bookings_not_canceled	avg_price_per_room
0	-0.088675	-1.276201
1	-0.088675	2.354667
2	-0.088675	0.519400
3	-0.088675	-0.803273
4	-0.088675	0.489804

	no_of_special_requests	family_size	total_stay
0	-0.785462	-1.456631	-0.004070
1	0.493638	3.155011	0.555470
2	-0.785462	0.080583	1.115010
3	-0.785462	0.080583	-0.004070

Table 19. Scaled Data.

- We did scaling on both our train & test dataset.

4 MODEL BUILDING

4.1 Model Evaluation Criteria

Model can make incorrect predictions as

- Predicting a customer will cancel a booking when they actually do not cancel, or predicting a customer will not cancel when they do cancel.

Which case is more critical?

- Both cases are important because:
    - If we predict a customer will cancel but they do not, the hotel might wrongly reassign or reject that booking. This can lead to lost revenue and a bad customer experience.
    - If we predict a customer will not cancel but they do cancel, the hotel will have empty rooms. This results in lost opportunities and poor use of resources.

How can we reduce this loss?

- We need to reduce both False Positives (FP) and False Negatives (FN).
- The f1_score should be maximized. A higher f1_score means a better chance of reducing both False Negatives and False Positives, allowing us to identify both classes correctly.
- The f1_score is calculated as:
    - f1_score = 2* ((Precision * Recall) / (Precision + Recall))

4.2 Building Logistic Regression Model

4.2.1 Base Logistic Model Building

We will now perform logistic regression using statsmodels for building initial (base) model. Using statsmodels, we will be able to check the statistical validity of our model. We will identify the significant predictors from the p-values we get for each predictor variable.

Adding Constant

We added constant to the train & test scaled data before for building a logistic regression model.

	const	required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date	repeated_guest	no_of_previous_cancellations	no_of_previous_bookir
0	1.0	-0.178610	0.397033	0.468891	-1.756718	1.418119	-0.161245	-0.064178	
1	1.0	5.598799	-1.011698	0.468891	-2.081898	-0.297758	-0.161245	-0.064178	
2	1.0	-0.178610	0.115287	0.468891	-0.781178	0.617377	-0.161245	-0.064178	
3	1.0	-0.178610	0.458282	-2.132690	0.194362	-0.640933	-0.161245	-0.064178	
4	1.0	-0.178610	0.776778	0.468891	0.194362	-1.441676	-0.161245	-0.064178	

Table 20. Split Data with Constant column.

Shape of data after adding constant

- Shape of train data : **25295** rows and **26** columns.
- Shape of test data : **10841** rows and **26** columns.

Since, both train & test data has equal columns and we can safely proceed with the model building with the worry of shape mismatch error.

Fitting Logistic Regression Model & Generating Summary

We build the model using sm.Logit() on X_train_with_const & y_train data and fitted using logit.fit() and

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25295
Model:	Logit	Df Residuals:	25269
Method:	MLE	Df Model:	25
Date:	Fri, 01 Aug 2025	Pseudo R-squ.:	0.3230
Time:	19:14:46	Log-Likelihood:	-10826.
converged:	False	LL-Null:	-15992.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6384	1276.897	-0.001	0.999	-2504.310	2501.033
required_car_parking_space	-0.2728	0.024	-11.278	0.000	-0.320	-0.225
lead_time	1.3123	0.022	59.429	0.000	1.269	1.356
arrival_year	0.1714	0.023	7.487	0.000	0.127	0.216
arrival_month	-0.1230	0.020	-6.165	0.000	-0.162	-0.084
arrival_date	0.0180	0.017	1.061	0.289	-0.015	0.051
repeated_guest	-0.2642	0.080	-3.288	0.001	-0.422	-0.107
no_of_previous_cancellations	0.0696	0.031	2.255	0.024	0.009	0.130
no_of_previous_bookings_not_canceled	-0.3937	0.315	-1.251	0.211	-1.010	0.223
avg_price_per_room	0.6325	0.025	24.928	0.000	0.583	0.682
no_of_special_requests	-1.1210	0.023	-47.935	0.000	-1.167	-1.075
family_size	0.0471	0.022	2.181	0.029	0.005	0.090
total_stay	0.0984	0.017	5.752	0.000	0.065	0.132
type_of_meal_plan_Meal Plan 2	0.0586	0.019	3.113	0.002	0.022	0.095
type_of_meal_plan_Meal Plan 3	0.2596	914.728	0.000	1.000	-1792.574	1793.094
type_of_meal_plan_Not Selected	0.0919	0.018	4.988	0.000	0.056	0.128
room_type_reserved_Room_Type 2	-0.0375	0.018	-2.101	0.036	-0.072	-0.003
room_type_reserved_Room_Type 3	-0.0017	0.021	-0.082	0.935	-0.042	0.039
room_type_reserved_Room_Type 4	-0.1051	0.020	-5.360	0.000	-0.143	-0.067
room_type_reserved_Room_Type 5	-0.0657	0.018	-3.708	0.000	-0.101	-0.031
room_type_reserved_Room_Type 6	-0.0945	0.020	-4.717	0.000	-0.134	-0.055
room_type_reserved_Room_Type 7	-0.0514	0.019	-2.678	0.007	-0.089	-0.014
market_segment_type_Complementary	-4.4902	1.23e+04	-0.000	1.000	-2.41e+04	2.41e+04
market_segment_type_Corporate	-0.2225	0.061	-3.646	0.000	-0.342	-0.103
market_segment_type_Offline	-0.8672	0.115	-7.561	0.000	-1.092	-0.642
market_segment_type_Online	-0.0667	0.120	-0.556	0.578	-0.302	0.168

Table 21. Base Logistic Model Summary.

Observations on logistic regression initial (base) model

Strong Predictors of Cancellation (Statistically Significant)

- **lead_time (Coefficient: +1.3123,P < 0.001):** Bookings made far in advance are more likely to be canceled.
- **required_car_parking Space (Coefficient: -0.2728 | p < 0.001):** Guests requesting parking are less likely to cancel.

- **repeated_guest (Coefficient: -0.2642 | p = 0.001):** Repeat guests show a much lower likelihood of cancellation.
- **no_of_special_requests (Coefficient: -1.1210 | p < 0.001):** More special requests correlate with lower cancellations, possibly due to higher guest commitment.
- **avg_price_per_room (Coefficient: 0.6325 | p < 0.001):** Higher-priced bookings are more prone to cancellations, possibly due to price sensitivity.
- **total_stay (Coefficient:+0.0984| p < 0.001):** Longer stays show a slightly higher chance of cancellation.
- **family_size (Coefficient: 0.0471 | p = 0.029):** Larger family groups tend to cancel more often.
- **arrival_year (+):** Cancellations were more frequent in 2018 compared to 2017.
- **arrival_month (Coefficient: 0.1714 | p < 0.001):** Slight reduction in cancellations for later months of the year.
- **type_of_meal_plan_Meal Plan 2 and type_of_meal_plan_Not Selected (+):** Guests choosing Meal Plan 2 or not selecting a meal plan are more likely to cancel.
- **room_type_reserved (e.g., Room Type 2, 4, 5, 6, 7) (-):** These types are associated with fewer cancellations.
- **market_segment_type - Corporate & Offline (-):** Corporate and offline bookings are less likely to cancel.

Insignificant Predictors (High p-values)

- **arrival_date (p = 0.289):** No significant impact on cancellation behaviour.
- **no_of_previous_bookings_not_canceled (p = 0.211):** Does not significantly influence cancellation likelihood.
- **room_type_reserved_Room Type 3 (p = 0.935):** No meaningful association with cancellations.
- **market_segment_type_Online (p = 0.578):** Not a statistically strong predictor of cancellations.
- **type_of_meal_plan_Meal Plan 3 & Complementary Segment:** Extremely high standard errors suggest sparse data; may need to be dropped or combined with other categories.

Potential Issues Identified

- The model did not converge, likely due to:
- Multicollinearity between variables (to be checked via VIF).
- Separation issues (Meal Plan 3/Complementary).
- Some predictors have very high standard errors including const. and near-zero z-scores, indicating unstable coefficients.

<u>Coefficient Interpretation using Odds for Base Model</u>

We converted coefficients to odds:

- The **coefficients of the logistic regression model relate to log(odd). To find the odds, we need to take the exponential of the coefficients.**
- Therefore, **odds = exp(b).**
- The percentage change in odds is calculated as **odds = (exp(b) - 1) * 100**.

	const	required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date	repeated_guest	no_of_previous_cancellations
Odds	0.194300	0.761281	3.714635	1.186942	0.884300	1.018116	0.767842	1.072092
Change in Odds (%)	-80.570035	-23.871854	271.463464	18.694223	-11.570022	1.811590	-23.215808	7.209181

no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	family_size	total_stay	type_of_meal_plan_Meal Plan 2	type_of_meal_plan_Meal Plan 3
0.674542	1.882312	0.325960	1.048274	1.103400	1.060329	1.296437
-32.545807	88.231226	-67.404038	4.827423	10.339974	6.032852	29.643670

room_type_reserved_Room_Type 3	room_type_reserved_Room_Type 4	room_type_reserved_Room_Type 5	room_type_reserved_Room_Type 6	room_type_reserved_Room_Type 7
0.998298	0.900267	0.936367	0.909847	0.949944
-0.170191	-9.973301	-6.363311	-9.015314	-5.005580

Table 22. Coefficient Interpretation using Odds.

Observations on Coefficient Odds for Initial Logistic Regression Model:

- For every 1 unit increase in **lead_time**, the odds of cancellation **increase** by about **271.46%.** This shows it is a strong predictor of cancellations.
- Customer who **require car parking space** have about **~23.87% lower odds** of cancelling their booking, indicating they are less likely to cancel.
- Being a **repeat guests reduces** the odds of cancellation by about **23.22%**, suggesting that loyal customers are more committed to their bookings.
- If a **customer has cancelled before**, the odds of cancelling again **increase by about 7.21%**. In contrast, having more previous successful bookings lowers the odds of cancellation by about 32.55%.
- A higher **average price per room** is linked to about **88.23% higher odds** of cancellation. Expensive bookings appear to be more likely to be cancelled.
- Customers who make **special requests** are about **67.40% less** likely to cancel, showing that more engaged guests tend to stick to their plans.
- A larger **family size** slightly **raises** the odds of cancellation by about **4.83%**.
- A longer **total stay** results in about **10.34% higher odds** of cancellation, which may reflect the greater commitment needed for long stays.
- **Compared to Meal Plan 1**:
    - Selecting **Meal Plan 2 increases** cancellation odds by about **6.03%.**
    - **Not choosing a meal plan raises** odd cancellation by about **9.62%**.
    - Selecting **Meal Plan 3 increases** cancellation odds by about **29.64%.**
- **Room types** also influence cancellation:
    - Room **Type 2 and Types 4 to 7 reduce** the odds of cancellation by about **3.68% to 9.97% compared to Room Type 1**.
- Customers from **Corporate, Offline, and Online market segments** have **6.45% to 57.99% lower** odds of cancellation than the reference group.
- The **Complementary market segment** shows almost **98.88% lower odds of cancellation**, but this value is not reliable, likely due to very sparse data.

4.2.2 Performance Evaluation of Base Logistic Model

First, let's create functions to calculate different metrics and the confusion matrix so we don't have to repeat the same code for each model.

- The model_performance_classification_statsmodels function will check the performance of the models.
- The confusion_matrix_statsmodels function will plot the confusion matrix.

Training Set Performance

Training performance of initial logistic regression model on train data:

	Accuracy	Recall	Precision	F1
0	0.801463	0.624305	0.729944	0.673004

Table 23. Training Set Performance of Base Logistic model.

Training performance observations:

Accuracy is about 80.15%

- The model correctly predicts around 80% of booking outcomes in the training data.
- This shows decent overall performance.

Recall is about 62.43%

- The model accurately identifies 62% of actual cancellations.
- Since we want to predict booking cancellations, recall is crucial; we don't want to overlook many real cancellations.

Precision is around 72.99%

- Out of all the bookings the model predicted as cancellations, about 73% were correct.
- This means the model doesn't make too many false positive predictions (wrongly labelling non-cancellations as cancellations).

The F1 Score is about 67.30%

- The F1 score balances precision and recall.
- A value around 67% shows a moderately strong performance in identifying both target classes.

Interpretation:

- Since our aim is to predict booking cancellations, both false positives and false negatives are costly:
    - False negatives (missed cancellations) may lead to overbooking.

- False positives (wrongly predicting a cancellation) may cause revenue loss or unnecessary room holds.

- A recall of about 62% shows the model still misses some cancellations, but a precision of around 73% means when it predicts a cancellation, it's usually correct.
- Overall, the model demonstrates promising basic performance, but there is room for improvement, especially in increasing recall and the F1 score, which can be the focus during model tuning.

Confusion Matrix of Train set

We created the confusion matrix using confusion_matrix() on independent features & actual target values and then plotted it using a heat map. We used **threshold = 0.5**

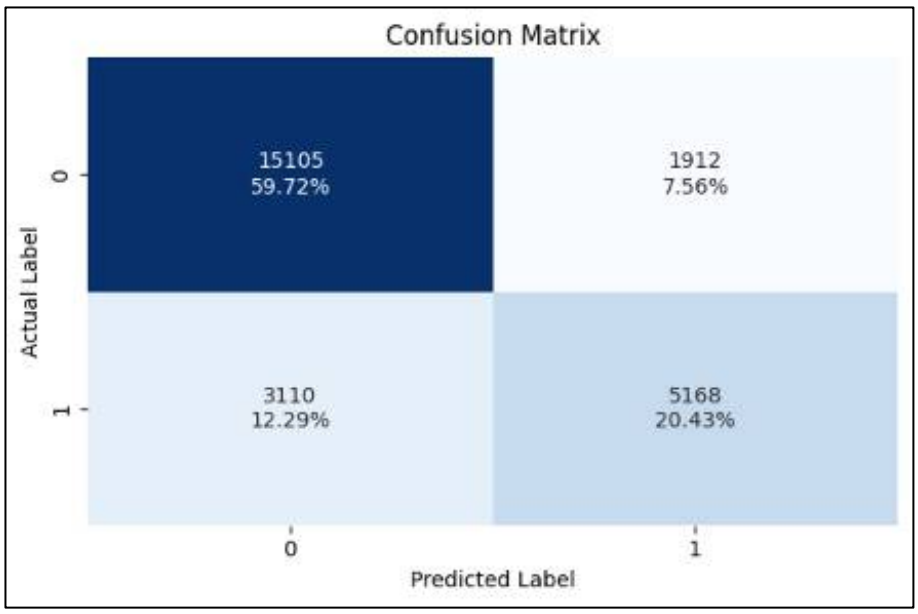


Figure 48. Train Set Confusion Matrix of Base Logistic model.

Note: A- Actual , P-Predicted

- ***A1 P0 - FN (False negative)***
- ***A1 P1 - TP (True positive)***
- ***A0 P0 - TN (True negative)***
- ***A0 P1 - FP (False positive)***

Train data confusion matrix observations:

True Negatives (TN = 15105, ~59.72%)

- The model correctly predicted about 60% of the bookings that were not cancelled. This shows strong performance for class 0.

True Positives (TP = 5168, ~20.43%)

- The model correctly identified around 20% of all actual cancellations.

- This shows the model's ability to detect actual cancellations (class 1) in the train data.

False Negatives (FN = 3110, ~12.29%)

- The model missed about 12% of actual cancellations. This can cause problems like overbooking.
- This means these are actual cancellations that were incorrectly predicted as non-cancellations.

False Positives (FP = 1912, ~7.56%)

- The model incorrectly predicted about 8% of non-cancellations as cancellations.
- This may lead to lost revenue by holding back rooms unnecessarily.

Class imbalance is handled decently

- Even though the target is imbalanced, with around 33% cancellations, the model captures the minority class (cancellations) with reasonable precision and recall.

Test Set Performance

Test performance of logistic regression :

	Accuracy	Recall	Precision	F1
0	0.811457	0.63542	0.752409	0.688984

Table 24. . Test Set Performance of Base Logistic model.

Test performance observations:

Accuracy is about 81.14%

- The model correctly predicts around 81% of booking outcomes in the test data.
- This shows generalization ability to unseen data.

Recall is about 63.54%

- The model accurately identifies ~64% of actual cancellations.
- Since we want to predict booking cancellations, recall is crucial; we don't want to overlook many real cancellations (false negatives).

Precision is around 75.24%

- Out of all the bookings the model predicted as cancellations, about 75% were correct.
- This means the model doesn't generate too many false positive predictions (wrongly labelling non-cancellations as cancellations).

The F1 Score is about 68.90%

- The F1 score balances precision and recall.
- A value around 69% shows a moderately strong performance in identifying both cancellation and non-cancellations.

Interpretation:

- Since our aim is to predict booking cancellations, both false positives and false negatives are costly:
  - False negatives (missed cancellations) may lead to overbooking.
  - False positives (wrongly predicting a cancellation) may cause revenue loss or unnecessary room holds.
- A recall of about 64% shows the model still misses some cancellations, but a precision of around 75% means when it predicts a cancellation, it's usually correct.
- Overall, the model demonstrates promising generalization, but there is room for improvement, especially in boosting recall and the F1 score, which can be the focus during model tuning.

Confusion Matrix of Test set

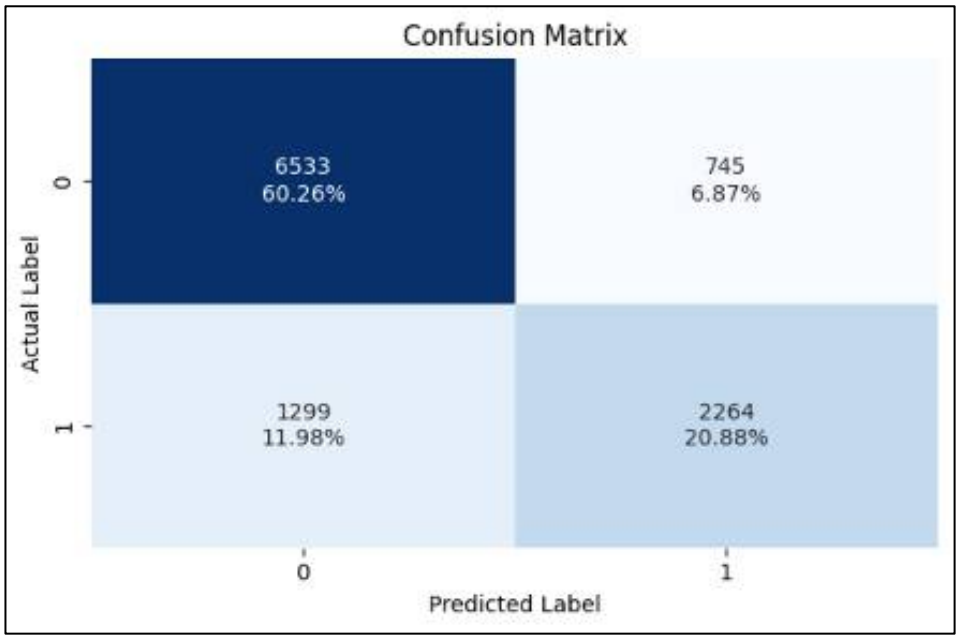


Figure 49. Test Set Confusion Matrix of Base Logistic model.

Test data confusion matrix observations:

True Negatives (TN = 6533, 60.26%)

- The model correctly predicted about 60% of the bookings that were not cancelled. This shows strong performance for identifying the majority class 0 - non-cancellations.

True Positives (TP = 2264, 20.88%)

- The model correctly identified around 21% of all actual cancellations.
- This shows the model's ability to detect actual cancellations (class 1) in the test data.

False Negatives (FN = 1299, 11.98%)

- This means these are actual cancellations that were incorrectly predicted as non-cancellations.
- The model missed about 12% of actual cancellations. This can cause problems like overbooking.

False Positives (FP = 745, 6.87%)

- The model incorrectly predicted about 7% of non-cancellations as cancellations.

- This may lead to lost revenue by holding back rooms unnecessarily.

Class imbalance is handled decently

- Even though the target is imbalanced, with around 33% cancellations, the model captures the minority class (cancellations) with reasonable precision and recall.

4.3 Building Decision Tree Classifier Model

4.3.1 Base Decision Tree Model Building

We will build the base decision tree model using DecisionTreeClassifier() with random state=1 and then fit it using X_train and y_train. Here we used unscaled data because **decision trees do not require feature scaling** as they divide based on feature thresholds independently regardless of magnitude. Also, there is no requirement for a constant (intercept) column because decision trees do not fit linear equations. They use hierarchical rule-based splits and avoid biased terms. Thus, constant term becomes irrelevant and if added, it may lead to unnecessary noise and potentially degrade the performance.

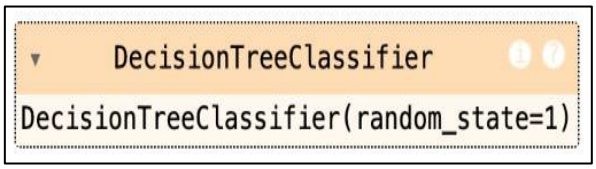


Table 25. Base Decision Tree Model.

- We build the model successfully.

4.3.2 Base Decision Tree Performance Evaluation

Performance of Train Set

Train performance of Initial Decision tree:

	Accuracy	Recall	Precision	F1
0	0.994149	0.984417	0.997674	0.991001

Table 26. Train Set Performance of Base Decision Tree model.

Training Performance Observations of Initial Decision Tree:

Accuracy is about 99.41%

- The model correctly classifies the booking status for over 99% of the training data.
- This high accuracy suggests that the model has learned the patterns in the training data well.

Recall is about 98.44%

- The model successfully identifies around 98% of actual cancellations.
- This means very few cancellations are missed by the model in the training data.

Precision is around 99.76%

- Out of all the bookings the model predicted as cancellations, almost 100% were actual cancellations.
- This shows there are very few false positives, meaning it rarely predicts a cancellation when there isn't one.

The F1 Score is about 99.10%

- This F1 score is very high, indicating a strong balance between precision and recall.
- It confirms that the model performs excellently on training data for both identifying and correctly classifying cancellations.

Interpretation:

- While these metrics show strong performance on the training data, the very high values might indicate overfitting. The model may not generalize well to unseen test data.
- We will compare this with test data and adjust the model if needed, such as by pruning.

Confusion Matrix of Train Set

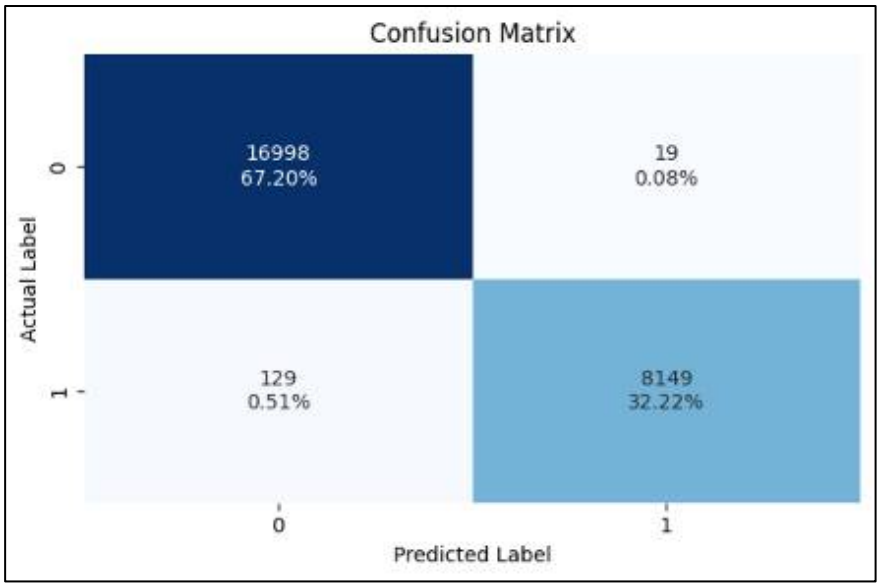


Figure 50. Train Set Confusion Matrix of Base Decision Tree model.

Note: A- Actual , P-Predicted

- ***A1 P0 - FN (False negative)***
- ***A1 P1 - TP (True positive)***
- ***A0 P0 - TN (True negative)***
- ***A0 P1 - FP (False positive)***

Train data confusion matrix observations of Initial Decision Tree:

True Negatives (TN = 16,998, 67.20%)

- The model accurately predicted about 67% of non-cancelled bookings.
- This indicates strong performance for class 0 (no cancellation).

True Positives (TP = 8149, 32.22%)

- The model correctly identified around 32% of total bookings as actual cancellations.
- The high TP count shows a strong ability to detect cancellations.

False Negatives (FN = 129, 0.51%)

- Only about 0.5% of actual cancellations were missed.
- This means there is a very low risk of overbooking due to missed cancellations.

False Positives (FP = 19, 0.08%)

- Only 19 non-cancelled bookings were wrongly predicted as cancelled.
- This suggests minimal revenue loss from incorrectly held inventory.

Interpretation:

- The confusion matrix shows very low rates of misclassification for both false positives and false negatives.
- This supports earlier metrics (Accuracy, Recall, Precision, F1) that indicated very high performance.
- However, such nearly perfect classification on training data may suggest overfitting, which should be checked by assessing performance on the test data.

Performance of Test Set

Test performance of Initial Decision tree:

	Accuracy	Recall	Precision	F1
0	0.868278	0.791187	0.804739	0.797905

Table 27. Test Set Performance of Base Decision Tree model.

Test Data Performance Observations of Initial Decision Tree:

Accuracy is about 86.82%

- The model correctly classifies the booking status for over ~87% of the test data.
- This suggests that the model has learned relevant patterns from the training data and is performing well on unseen data.

Recall is about 79.11%

- The model successfully predicts around 79% of actual cancellations.
- This means very few cancellations are missed by the model in the test data, as missed cancellations can lead to overbooking.

Precision is around 80.47%

- Out of all the bookings the model predicted as cancellations, about 80% were actual cancellations.
- This indicates there are very few false positives, meaning it rarely predicts a cancellation when there isn't one.

- Thus, reducing the risk of unnecessary room holds.

The F1 Score is about 79.79%

- This F1 score is very high, indicating a strong balance between precision and recall.
- This shows a value close to 80% reflects strong classification capability for cancellations.

Interpretation:

- These metrics show that the model performs much better than a simple classifier, particularly in identifying cancellations.
- The drop in performance from training to testing is noticeable but not severe, which suggests some overfitting.
- In the next phase, we will look into model tuning, such as pruning, to improve generalization and stability on new data.

Confusion Matrix of Test Set

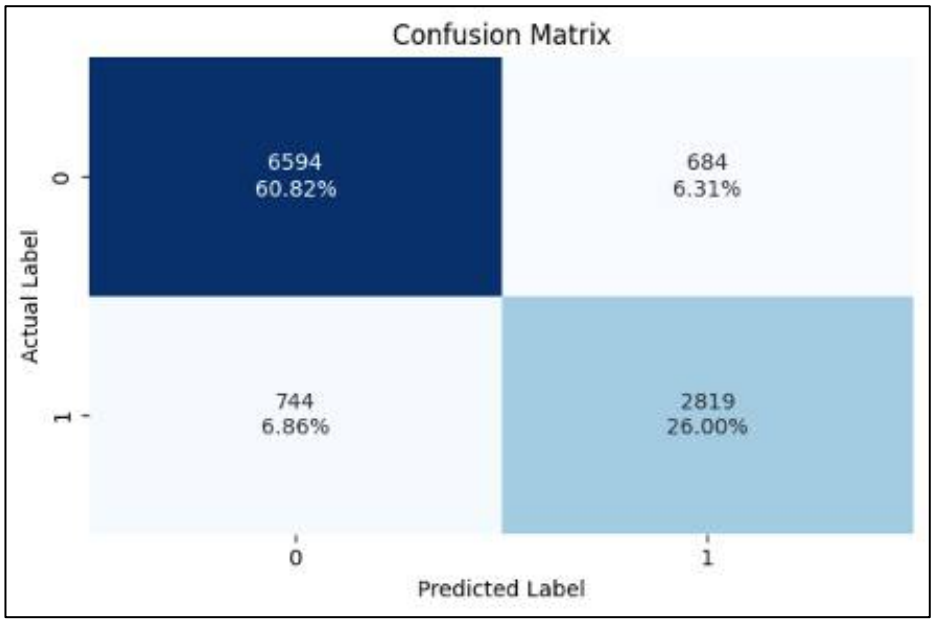


Figure 51. Test Set Confusion Matrix of Base Decision Tree model.

Test data confusion matrix observations of Initial Decision Tree:

True Negatives (TN = 6594, 60.82%)

- The model correctly predicted that 6594 bookings were not cancelled, making this the largest group.
- This indicates that the model is effective at identifying non-cancellations.

True Positives (TP = 2819, 25.85%)

- The model accurately identified 2819 cancellations, meaning about 26% of the total records were correctly labelled as cancellations.
- This is a strong result for the positive class, which is our main focus.

False Negatives (FN = 744, 6.31%)

- These are actual cancellations that the model missed and marked as not cancelled.
- This can lead to overbooking risks, so it's essential to reduce these numbers.

False Positives (FP = 684, 6.15%)

- These are cases where the model predicted a cancellation, but the booking was actually completed.
- While not as serious as false negatives, this can still create unnecessary holds on room availability.

Interpretation:

- The decision tree works well for both classes.
- It correctly captures a large share of true cancellations (TP) and keeps both types of errors (FP and FN) relatively low.
- These results match well with the earlier performance metrics (recall and precision), strengthening the model's reliability in handling cancellation predictions.
- However, compared to the training set, the test set shows a drop in performance, particularly with an increase in false positives and false negatives. This is expected because of possible overfitting on the training data. We will tackle this in the model improvement phase by using pruning optimization.

5 MODEL PERFORMANCE IMPROVEMENT

5.1 Logistic Regression Model Performance Improvement

5.1.1 Multicollinearity Check

We will check the presence of multicollinearity among independent features using variance inflation factor (VIF). If VIF> 5 then we would remove those features sequentially.

Checking VIF values

```
Variance Inflation Factors :

const                                    1.00000
required_car_parking_space               1.03701
lead_time                                1.36620
arrival_year                             1.42275
arrival_month                            1.27778
arrival_date                             1.00648
repeated_guest                           1.78862
no_of_previous_cancellations             1.34601
no_of_previous_bookings_not_canceled     1.62399
avg_price_per_room                       1.98841
no_of_special_requests                   1.25923
family_size                              1.64489
total_stay                               1.10235
type_of_meal_plan_Meal Plan 2            1.24954
type_of_meal_plan_Meal Plan 3            1.01545
type_of_meal_plan_Not Selected           1.26903
room_type_reserved_Room_Type 2           1.03017
room_type_reserved_Room_Type 3           1.00301
room_type_reserved_Room_Type 4           1.33092
room_type_reserved_Room_Type 5           1.02794
room_type_reserved_Room_Type 6           1.43448
room_type_reserved_Room_Type 7           1.08135
market_segment_type_Complementary        4.28762
market_segment_type_Corporate           16.21635
market_segment_type_Offline             59.95036
market_segment_type_Online              66.58598
dtype: object
```

Table 28. VIF values.

- We can observe **market_segment_type_Corporate (16.21), market_segment_type_Offline (59.95) and market_segment_type_Online (66.58) have VIF > 5**.
- We will remove these features having VIF>5 one at a time and recalculation VIF again.

Removing Features with VIF>5

```
Dropping 'market_segment_type_Online' due to high VIF = 66.5860

Final VIF Results:
const                                    1.00000
required_car_parking_space               1.03689
lead_time                                1.36048
arrival_year                             1.41976
arrival_month                            1.27662
arrival_date                             1.00647
repeated_guest                           1.78570
no_of_previous_cancellations             1.34592
no_of_previous_bookings_not_canceled     1.62359
avg_price_per_room                       1.98756
no_of_special_requests                   1.25407
family_size                              1.62678
total_stay                               1.10102
type_of_meal_plan_Meal Plan 2            1.24922
type_of_meal_plan_Meal Plan 3            1.01545
type_of_meal_plan_Not Selected           1.26642
room_type_reserved_Room_Type 2           1.03017
room_type_reserved_Room_Type 3           1.00301
room_type_reserved_Room_Type 4           1.32694
room_type_reserved_Room_Type 5           1.02794
room_type_reserved_Room_Type 6           1.43198
room_type_reserved_Room_Type 7           1.08110
market_segment_type_Complementary        1.35536
market_segment_type_Corporate            1.50761
market_segment_type_Offline              1.61395
dtype: object
```

Table 29. VIF values after removing VIF>5.

- We can see **'market_segment_type_Online' feature was dropped** due to high VIF = 66.59 both from train and test data set.
- After dropping the high VIF feature, we can see that the remaining features have VIF < 5.

5.1.2 Dealing with High P-values

Dealing with High P-values

Some of the features have very high p value i.e. > 0.05 . So, **they are not significant features** and we'll **drop them one at a time** as sometimes p-values change after dropping a variable

We will do the following:

- Build a model, check the p-values of the variables, and drop that variable from our dataset.
- Re-fit the model using the updated dataset (with the dropped column).
- Repeat the above two steps till there are no columns with p-value > 0.05

```
Dropping column 'type_of_meal_plan_Meal Plan 3' with p-value: 0.9998
Dropping column 'market_segment_type_Complementary' with p-value: 0.9998
Dropping column 'room_type_reserved_Room_Type 3' with p-value: 0.8909
Dropping column 'arrival_date' with p-value: 0.3123
Dropping column 'no_of_previous_bookings_not_canceled' with p-value: 0.2086
Dropping column 'market_segment_type_Online' with p-value: 0.1824
All remaining features have p-values ≤ 0.05

Dropped features due to high p-values:
['type_of_meal_plan_Meal Plan 3', 'market_segment_type_Complementary', 'room_type_reserved_Room_Type 3', 'arrival_date', 'no_of_previous_bookings_not_cancel

Retained features with significant p-values:
['const', 'required_car_parking_space', 'lead_time', 'arrival_year', 'arrival_month', 'repeated_guest', 'no_of_previous_cancellations', 'avg_price_per_room'
```

Figure 52. Dealing with high p-values.

We can see from above image that many columns with high p-values has been dropped and only significant features were retained.

5.1.3 Retraining Logistic Model

Re-Building Model

Now we will be preparing refined train and test datasets to **build an updated regression model with only significant features** that are left after we dropped the bad features.

```
Optimization terminated successfully.
         Current function value: 0.428609
         Iterations 9
                         Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25295
Model:                          Logit   Df Residuals:                    25275
Method:                           MLE   Df Model:                           19
Date:                Fri, 01 Aug 2025   Pseudo R-squ.:                  0.3221
Time:                        19:14:51   Log-Likelihood:                -10842.
converged:                       True   LL-Null:                       -15992.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           -1.1558      0.021    -54.731      0.000      -1.197      -1.114
required_car_parking_space      -0.2734      0.024    -11.302      0.000      -0.321      -0.226
lead_time                        1.3153      0.022     59.858      0.000       1.272       1.358
arrival_year                     0.1682      0.023      7.369      0.000       0.123       0.213
arrival_month                   -0.1278      0.020     -6.430      0.000      -0.167      -0.089
repeated_guest                  -0.3362      0.068     -4.953      0.000      -0.469      -0.203
no_of_previous_cancellations     0.0578      0.029      2.024      0.043       0.002       0.114
avg_price_per_room               0.6468      0.025     25.926      0.000       0.598       0.696
no_of_special_requests          -1.1203      0.023    -47.997      0.000      -1.166      -1.075
family_size                      0.0471      0.021      2.199      0.028       0.005       0.089
total_stay                       0.1013      0.017      5.937      0.000       0.068       0.135
type_of_meal_plan_Meal Plan 2    0.0556      0.019      2.962      0.003       0.019       0.092
type_of_meal_plan_Not Selected   0.0959      0.018      5.223      0.000       0.060       0.132
room_type_reserved_Room_Type 2  -0.0369      0.018     -2.066      0.039      -0.072      -0.002
room_type_reserved_Room_Type 4  -0.1062      0.020     -5.428      0.000      -0.144      -0.068
room_type_reserved_Room_Type 5  -0.0668      0.018     -3.774      0.000      -0.101      -0.032
room_type_reserved_Room_Type 6  -0.0976      0.020     -4.887      0.000      -0.137      -0.058
room_type_reserved_Room_Type 7  -0.0534      0.019     -2.793      0.005      -0.091      -0.016
market_segment_type_Corporate   -0.1891      0.023     -8.067      0.000      -0.235      -0.143
market_segment_type_Offline     -0.7993      0.024    -33.884      0.000      -0.845      -0.753
================================================================================================
```

Table 30. Tuned Logistic Regression Model.

Observations on Tuned Logistic Regression Model:

Model Convergence

- The logistic regression **model converged successfully** in 9 iterations. This shows stable optimization than the base logistic regression model.

Model Fit

- The log-likelihood value is -10842.0. The Pseudo R^2 is 0.3221. This suggests the **model explains about 32.21% of the variation in the dependent variable**.

Model Significance

- The LLR p-value is 0.000. This indicates the model is statistically significant overall.

Statistically Significant Predictors

- **All predictors** in the model have **p-values less than 0.05**.
- This makes them statistically significant contributors to predicting booking cancellations.

Direction and Strength of Predictors

- **required_car_parking_space** has a negative coefficient **(-0.2734**). Customers who require car parking space are less likely to cancel.
- **lead_time** has a strong positive effect (**1.3153**). Longer lead time increases the chance of cancellation.
- **arrival_year** shows a positive association (**0.1682**) with cancellations, while arrival_month shows a negative effect **(-0.1278).**
- **repeated_guest** has a negative coefficient **(-0.3362**). This means repeat customers are less likely to cancel.

- **no_of_previous_cancellations** has a mild positive effect (**0.0578**). Customers with past cancellations are slightly more likely to cancel again.
- **avg_price_per_room** is positively related to cancellations (**0.6468**). This suggests that expensive bookings are more likely to be canceled.
- **no_of_special_requests** has a strong negative effect **(-1.1203**). This indicates that these guests are less likely to cancel.
- **family_size** and **total_stay** both have small positive effects (**0.0471 and 0.1013**).

Meal plan, room type, and market segment dummies

- All mostly show small but significant effects. For example:
    - **type_of_meal_plan_Not Selected (0.0959)** increases the likelihood of cancellation.
    - **room_type_reserved_Room_Type 4 (-0.1062)** decreases the chance of cancellation.
    - **market_segment_type_Offline (-0.7993)** strongly reduces the chance of cancellation.

Interpretability:

- The model is now easier to understand with **stable coefficients** and clear effects from all predictors.
- The lack of large standard errors or undefined estimates supports the model's reliability.

<u>Coefficient Interpretation using Odd's for Tuned Logistic Regression Model</u>

Converted log odd's coefficient to odd's coefficient for coefficient interpretations.

	const	required_car_parking_space	lead_time	arrival_year	arrival_month	repeated_guest	no_of_previous_cancellations	avg_price_per_room	no_of_
Odds	0.314818	0.760777	3.725733	1.183208	0.880050	0.714503	1.059500	1.90938	
Change in Odds (%)	-68.518249	-23.922301	272.573267	18.320823	-11.994985	-28.549704	5.950043	90.93796	

Table 31. Tuned Logistic Regression Model Odd's Coefficient.

Observations on Coefficient Odds (Tuned Logistic Regression Model):

These observations are from the tuned model after filtering for VIF and p-values, followed by the calculation of odds.

lead_time

- Final Odds = **3.73 → 272.57% increase** in odds of cancellation
- This remains the strongest predictor. A longer lead time significantly increases cancellation risk, just as in the initial model.

required_car_parking_space

- Final Odds = **0.76 → about 23.92% decrease**

- This continues to lower cancellation likelihood, supporting the idea that guests with specific needs, like parking, are more committed.

Repeated Guest

- Final Odds = **0.71 → about 28.55% lower odds**
- Being a loyal customer reduces the likelihood of cancellation further, with a slightly stronger effect in the final model compared to the initial model (which was about 23%).

No. of Previous Cancellations

- Final Odds = **1.06 → about 5.95% increase**
- This is slightly higher than the initial model (approximately 7.21%) but still shows a risk of repeat cancellations.

Average Price Per Room

- Final Odds = **1.91 → about 90.94% higher odds**
- This variable now shows a stronger influence on cancellation compared to the initial model (which was around 88.23%), possibly due to corrections for multicollinearity.

No. of Special Requests

- Final Odds = **0.33 → about 67.38% lower odds**
- The effect remains strong and consistent. Customers who make special requests are much less likely to cancel; they are more invested.

Family Size

- Final Odds = **1.05 → about 4.83% increase**
- This shows a small increase in cancellation odds, likely due to the complexity of family plans.

Total Stay

- Final Odds = **1.11 → about 10.66% higher odds**
- This is similar to earlier results (which was about 10.34%). Longer stays may require more commitment and are more likely to be canceled.

Meal Plan (Compared to Meal Plan 1 - Reference)

- ***Meal Plan 2***
    - Final Odds = **1.06 → about 5.72% increase**
    - This is slightly lower than the initial figure (which was around 6.03%) but still increases the risk of cancellation.

- **Not Selected** : Final Odds = **1.10 → about 10.06% increase**
- This is very similar to the initial model (which was around 9.62%). Guests who avoid meal plans may be less committed.

Room Type (Compared to Room Type 1)

- **Room Type 2 to Room Type 7**

    - All show odds less than 1, ranging from **0.89 to 0.96 → 3.6% to 10% lower odds**
    - They still show protection against cancellation, consistent with the initial model. Most notably:
- **Room Type 5 → 10.07% lower odds**
- **Room Type 6 → 6.46% lower odds**
- **Room Type 7 → 9.30% lower odds**

Market Segment (Compared to Online - likely reference)

- **Corporate**
    - Final Odds = **0.83 → about 17.23% decrease**
    - This shows a slightly larger reduction than before (around 6.45%), but it is still significant.
- **Offline**
    - Final Odds = **0.45 → about 55.03% decrease**
    - Offline guests are much less likely to cancel, consistent with the previous strong effect (which was around 57.99%).

Conclusion:

- The **overall pattern of cancellation drivers remains the same, but after removing multicollinear and non-significant features, the effect sizes became more stable**.
- **Lead time, average price, and special requests** remain the **top three strongest predictors**.
- The odds are now more reliable as they come from a clearer and statistically sound model.

5.1.4 Determining Optimal Threshold using ROC-AUC Curve

Plotting ROC Curve

Now, we will plot ROC curve.

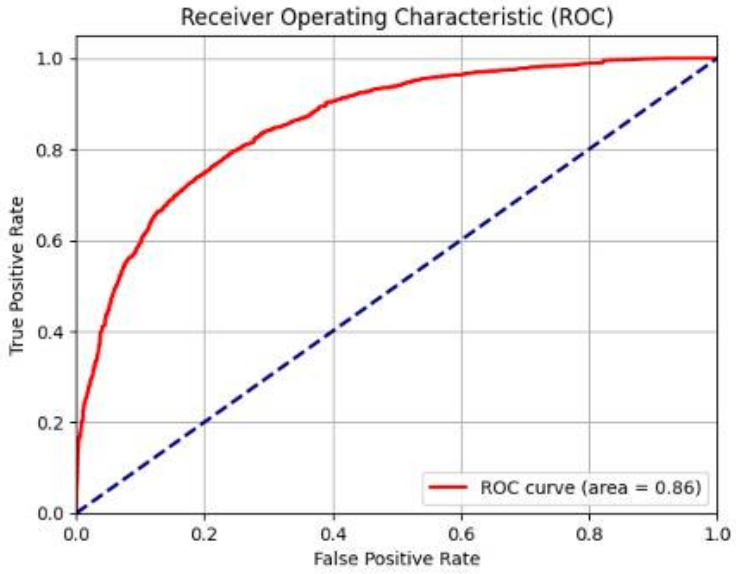


Figure 53. ROC Curve Plot.

Finding Optimal Threshold

We found optimal threshold using np.argmax() on the difference of true positive rate and false positive rate.

Optimal Threshold = 0.302

Observations on ROC curve & Optimal Threshold of Tuned Logistic model:

- The ROC curve (red line) shows how the model performs across different threshold values by plotting True Positive Rate (TPR) against False Positive Rate (FPR).
- The area under the curve **(AUC = 0.86)** indicates excellent model performance. It is close to 1.0, which suggests the model distinguishes well between classes.
- The navy blue dashed diagonal line represents a random classifier (AUC = 0.5). Our model performs significantly better than random.
- The optimal threshold, calculated using the formula **argmax(TPR - FPR), is 0.302**.

Interpretation of ROC Curve & Optimal Threshold:
AUC = 0.86:

- This means there's an 86% chance that the model will rank a randomly chosen positive class (booking cancelled) higher than a randomly chosen negative one (not cancelled). An AUC above 0.80 is generally considered very good.

Interpretation of Threshold = 0.302

- If the model's predicted probability for a booking is greater than or equal to 0.302, classify it as "Cancelled."
- If it's less than 0.302, classify it as "Not Cancelled."
- This **threshold is lower than the default 0.5**, which means the model is now more sensitive to detecting cancellations. It prioritizes fewer false negatives, meaning it misses fewer actual cancellations.

5.1.5 Tuned Logistic Model (Threshold = 0.302) Performance Check

Train set Performance Check

Performance of trained data on tuned logistic regression model:

	Accuracy	Recall	Precision	F1
0	0.770666	0.789442	0.616917	0.692597

Table 32. Tuned Logistic Regression (Threshold = 0.302) Performance on Train set.

Performance observations of tuned regression model on trained data:

Accuracy = 77.1%

- The model correctly predicted booking cancellation status for 77.1% of the training samples.

Recall = 78.9%

- The model successfully identified 78.9% of all actual cancellations (positive class).
- This is important because our goal is to detect cancelled bookings.

Precision = 61.7%

- Of all bookings the model predicted as cancelled, 61.7% were actually cancelled.
- This indicates some false positives, but it is acceptable given the high recall.

F1 Score = 69.3%

- The harmonic mean of precision and recall shows a good balance between identifying cancellations and avoiding false alarms.

Interpretation:

- The tuned logistic regression model performs well on the training data, with high recall supporting the project goal of identifying cancelled bookings.
- While precision is moderate, the overall F1-score of about 69% reflects a fair trade-off.
- These results show that the model has learned relevant patterns in the training data.
- We will next validate its performance using test data.

Confusion Matrix of Tuned Logistic Regression on Train set

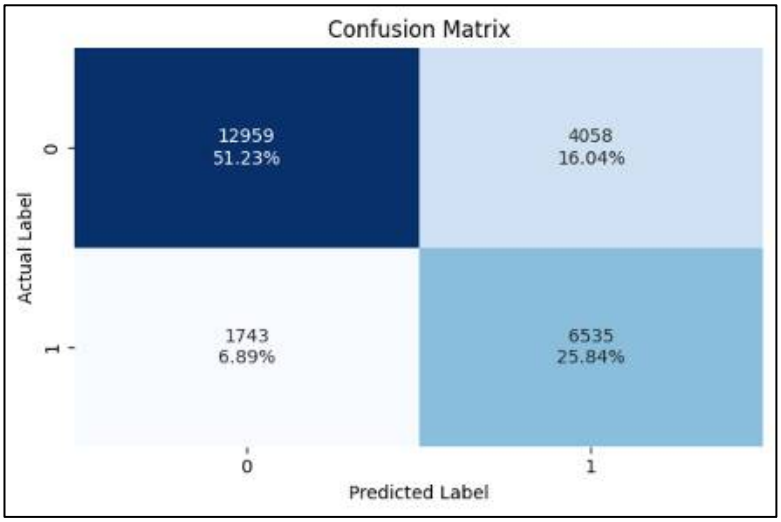


Figure 54. Train Set-Confusion Matrix of Tuned Regression Model (Threshold = 0.302).

Note: A- Actual , P-Predicted

- ***A1 P0 - FN (False negative)***
- ***A1 P1 - TP (True positive)***
- ***A0 P0 - TN (True negative)***
- ***A0 P1 - FP (False positive)***

Train data confusion matrix observations of tuned logistic regression model:

True Negatives (TN = 12959, 51.23%)

- The model accurately predicted about 51% of non-cancelled bookings.
- This indicates decent performance for class 0 (no cancellation) though lower than initial model.

True Positives (TP = 6535, 25.84%)

- The model correctly identified around 26% of total bookings as actual cancellations.
- Decent detection of cancellations, though lower than initial model.

False Negatives (FN = 1743, 6.89%)

- Only about 7% of actual cancellations were missed.
- This means there is a slight risk of overbooking due to missed cancellations.

False Positives (FP = 4058, 16.04%)

- Around 16% non-cancelled bookings were wrongly predicted as cancelled.
- This suggests more revenue loss from incorrectly held inventory.

Interpretation:

- The tuned model demonstrates a better balance between precision and recall, which improves the F1 score.
- Even though misclassifications (FP, FN) are higher than in the initial model, the first model was probably overfitted, achieving near-perfect results on training data.
- This tuned version should generalize better, and we will confirm this through its performance on test data.

Comparison with Initial Model:

- The initial model achieved nearly perfect training accuracy. However, the tuned model reflects a more realistic performance and has better potential for generalization.
- This trade-off is intentional and necessary to prevent overfitting and to better represent true performance on unseen data.

Test set Performance Check

Performance of test data on tuned logistic regression model:

	Accuracy	Recall	Precision	F1
0	0.775851	0.800449	0.623933	0.701254

Table 33. Tuned Logistic Regression (Threshold = 0.302) Performance on Test set.

Performance observations of tuned regression model on test data:

Accuracy = 77.6%

- The model correctly predicted the booking cancellation status for 77.6% of the test samples.
- This is slightly higher than the training accuracy, which shows consistent generalization.

Recall = 80.0%

- The model successfully identified 80.0% of actual cancellations, which is even higher than the training recall.
- This highlights the model's strength in capturing the positive class (cancelled bookings) and supports the business goal.

Precision = 62.4%

- Of all the bookings predicted as cancelled, 62.4% were actually cancelled.
- This reflects a manageable number of false positives while focusing on high recall.

F1 Score = 70.1%

- The F1 score improved slightly on the test set, indicating a good balance between precision and recall.
- This suggests the model is neither overfitting nor underfitting and manages class imbalance fairly well.

Interpretation & Improvement Comparison:

- The test performance is slightly better than the training performance, especially in recall and F1 score.
- This means the tuned logistic regression model has generalized well to unseen data, which is an important goal of model tuning.
- Compared to the initial model (which faced convergence and multicollinearity issues), the tuned model provides cleaner and more stable predictions and avoids overfitting despite a drop in training metrics.
- With a high recall of about 80% and improved F1 of around 70% on the test set, the model is effective at identifying bookings likely to be cancelled.

Conclusion:

- Overall, the tuned logistic regression model not only fixed the instability issues of the initial model but also showed better generalization and maintained strong recall performance, making it reliable for predicting cancellations.

<u>Confusion Matrix of Tuned Logistic Regression on Test set</u>

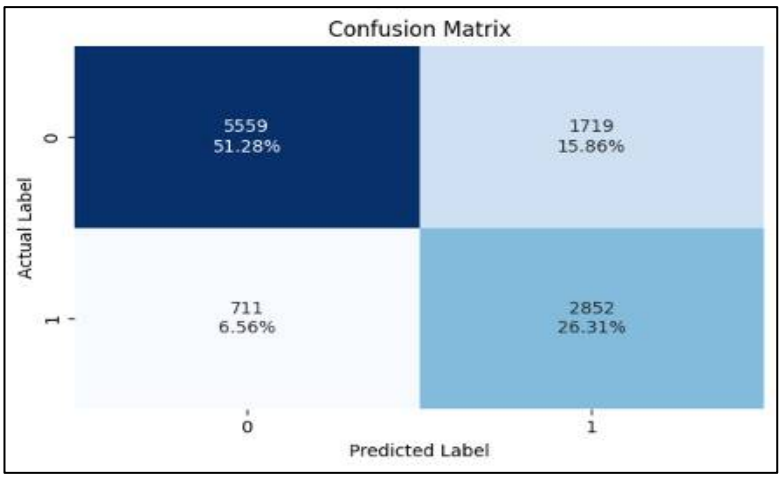


Figure 55. Test Set-Confusion Matrix of Tuned Regression Model (Threshold = 0.302).

Test data confusion matrix observations of tuned logistic regression model:

True Negatives (TN = 5559, 51.28%)

- The model correctly identified 51.28% of the non-cancelled bookings.
- This shows it captures most class 0 instances, though there is a slight drop from the initial model.

True Positives (TP = 2852, 26.31%)

- About 26.3% of the total samples were correctly classified as cancellations.
- This indicates a strong ability to identify the positive class (cancelled bookings), which matches the project's goal.

False Negatives (FN = 711, 6.56%)

- 6.56% of actual cancellations were missed and were predicted as non-cancelled.
- These are risky because they could lead to overbooking and unhappy customers.

False Positives (FP = 1719, 15.86%)

- 15.86% of non-cancelled bookings were wrongly predicted as cancellations.
- This could cause revenue loss from blocked inventory, but the level is acceptable given the high recall.

Interpretation & Comparison with Trained Model:

- Compared to the training data confusion matrix, the model performs slightly better on the test data, especially in:
    - Recall (80.0%)
    - F1-Score (70.1%)
- The drop in false positives and false negatives from training to test indicates that the model generalizes well and maintains consistent performance.
- The trade-off between slightly lower precision and higher recall is acceptable here because identifying cancellations is the main objective, and high recall ensures fewer missed cancellations.

Comparison with Initial (Untuned) Model:

- The initial model may have shown better metrics on the training data due to overfitting, but such performance often fails on test data.
- In contrast, this tuned model provides stable, interpretable, and more reliable results, even though some metrics like accuracy or precision seem slightly lower.

Conclusion:

- Ultimately, the tuned logistic regression model offers a balanced, generalizable, and practical solution for predicting booking cancellations.

Improved (Tuned) Logistic Model Summary

- After applying feature selection and **optimizing the threshold with the ROC curve, the logistic regression model showed a significant improvement**.
- It achieved an **F1-score of around 70% on unseen data**. The model also reached a **high recall of about 80%.** This allows the business to identify potential cancellations early while reducing false predictions and related revenue losses.
- In the next step, we plan to **further improve the model's precision and recall balance. We will explore the best threshold using the Precision-Recall Curve**, as false positives can also incur business costs.

5.1.6 Determining Better Threshold Using Precision-Recall Curve

In the previous step, we chose the best threshold using the ROC Curve. This curve balances the trade-off between sensitivity (Recall) and specificity (1 - False Positive Rate). However, in real-world classification problems, especially those with moderately imbalanced classes or when Precision and Recall are both vital, it helps to look at the Precision-Recall Curve. This curve helps us find a threshold that maximizes the F1-score.

This step improves the model's overall effectiveness, particularly when false positives and false negatives affect the business.

In this step, we will perform following steps:

- Use the predicted probabilities from the training set.
- Generate the Precision-Recall curve.
- Calculate the F1-score for each threshold.
- Select the threshold that gives the highest F1-score.
- Visualize how Precision, Recall, and F1 change with the threshold.

Precision-Recall Curve

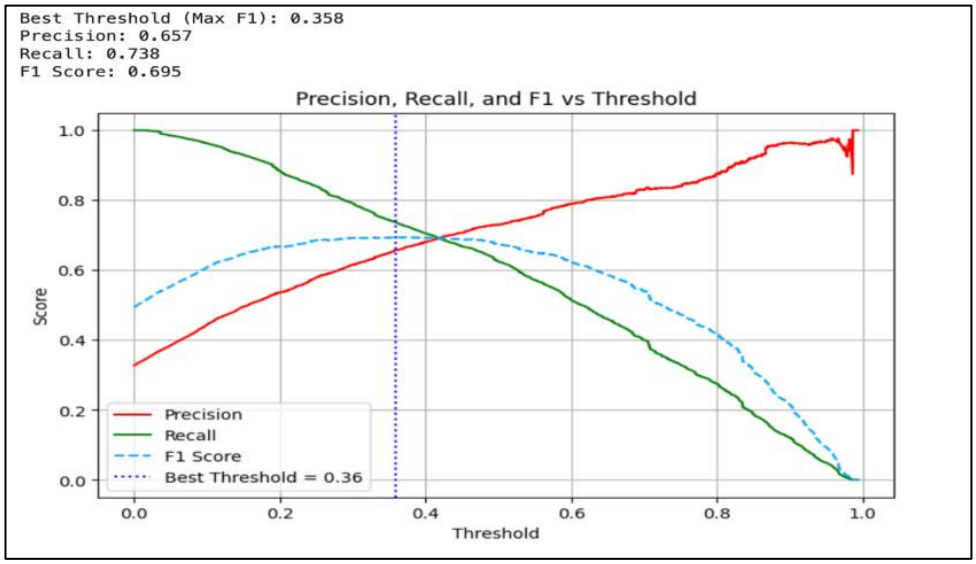


Figure 56. Precision-Recall Curve.

Observations on Precision-Recall curve & Optimal Threshold of Tuned Logistic model:

- From the Precision-Recall vs Threshold curve, **the best threshold to maximize the F1-score is 0.358**. At this threshold:
    - **Precision is about 65.7%,** which means around two-thirds of predicted cancellations are correct.
    - **Recall is around 73.8%**, showing the model captures a large number of actual cancellations.
    - The **F1-score is 69.5%**, indicating a good balance between precision and recall.
- This threshold (0.358) is a bit higher than the ROC-based threshold of 0.302.
- It focuses on improving overall prediction quality instead of just balancing sensitivity and specificity.
- Using this threshold may improves the model's ability to detect cancellations while reducing false alarms.
- This will make it better for business decisions, where both missed cancellations and unnecessary actions can be costly.

5.1.7 Tuned Logistic Regression Model (Threshold = 0.358) Performance Check

We will check the model's performance on data with threshold = 0.358, the best threshold selected using by PR curve

Train Set Performance Check (Threshold = 0.358)

Performance on trained data on tuned logistic regression model

	Accuracy	Recall	Precision	F1
0	0.788021	0.737497	0.656875	0.694855

Table 34. Train set Performance on Tuned Logistic Regression Model (Threshold = 0.358).

Observations on train data for tuned logistic regression model with threshold = 0.358:

- The model achieved an **accuracy of 78.8%**, showing a slight improvement over the previous version.
- **Precision increased to 65.7**%, which reduced the number of false positives, meaning there are fewer incorrect cancellation predictions.
- **Recall slightly decreased to 73.7%**, indicating the model missed a few more actual cancellations than before.
- The **F1-score improved marginally to 69.5%,** which shows a better balance between precision and recall.

Comparison with Tuned Logistic Regression Using ROC Curve Threshold = 0.302 (Train Data):

- **Compared to the ROC-based threshold model, which had an F1 of 69.3%, the PR-curve threshold model achieved a slightly better F1 score of 69.5%.**
- It shows a clear improvement in precision, going from 61.7% to 65.7%, while recall slightly decreased, from 78.9% to 73.7%.

- This trade-off suggests that **the model is now more cautious in flagging cancellations**. This change could help reduce unnecessary intervention, though it comes with a minor cost in recall.

Train set Confusion Matrix of Tuned Regression Model (Threshold = 0.358)

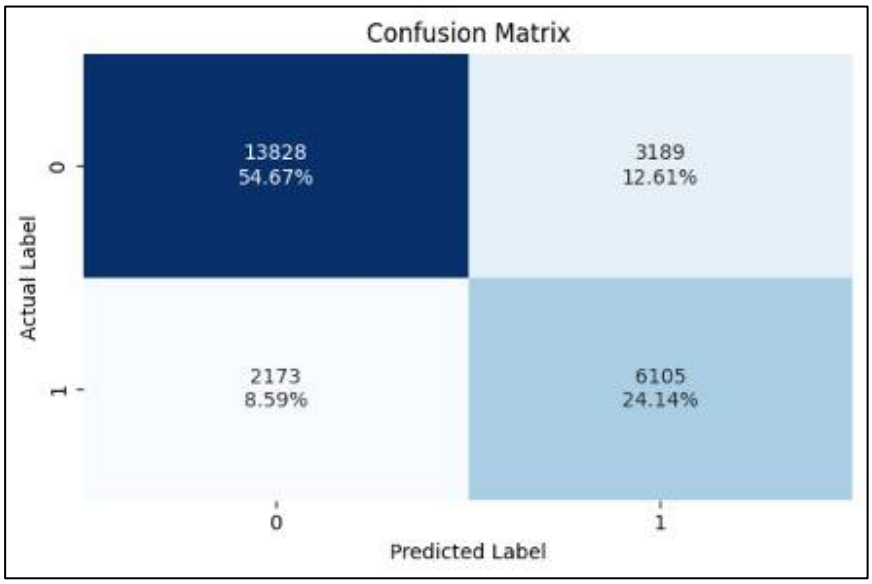


Figure 57. Train set Confusion Matrix of Tuned Regression Model (Threshold = 0.358).

Observations on Confusion matrix train data for tuned logistic regression model with threshold = 0.358:

- Most bookings were correctly identified as **not canceled (13,828 TN).**
- The model accurately detected 6,105 cancellations (TP), which is about **24.14%** of the total data.
- **False negatives (2,173**) are relatively low, showing good recall.
- There are false positives (3,189), but they are manageable, which helps maintain a balanced precision.
- This matrix shows a solid balance between precision and recall, supporting the **model's F1-score of about 0.695 on the training data**.

Comparison with ROC Threshold (Train Data):

- The ROC model had higher recall but lower precision, which means more bookings were predicted as canceled (increased TP, but also increased FP) and slightly lower F1 as compared to PR curve threshold.

Test Set Performance Check (Threshold = 0.358)

Performance on test data on tuned logistic regression model with best threshold – 0.358:

	Accuracy	Recall	Precision	F1
0	0.79679	0.750491	0.670512	0.708251

Table 35. . Test set Performance on Tuned Logistic Regression Model (Threshold = 0.358).

Observations on test data for tuned logistic regression model with threshold = 0.358:

- The model achieved an **accuracy of 79.7%,** showing a clear rise in correct predictions compared to before.
- **Precision improved significantly to 67.1**%, which means fewer incorrect cancellation alerts were triggered.
- **Recall slightly decreased to 75.0%,** indicating the model identified slightly fewer true cancellations.
- The **F1-score increased to 70.8%,** reflecting a more balanced and reliable performance in predicting cancellations on unseen data.

Comparison with Tuned Logistic Regression Using ROC Curve Threshold = 0.302 (Test Data):

- The ROC-threshold model had an F1-score of 70.1%, while the PR-curve threshold model improved it slightly to 70.8%.
- **Precision increased from 62.4% to 67.1%,** showing a strong gain in reducing false positives.
- However, **recall dropped from 80.0% to 75.0%**, meaning the new model detects slightly fewer actual cancellations.
- Overall, the Precision-Recall Curve-based threshold model performs better than the ROC-based threshold model.
- **It achieves a higher F1-score on both the train and test sets**. Although it slightly reduces recall, the improvement in precision results in a more balanced and effective model.
- Since F1-score is the chosen evaluation metric, which balances the cost of false positives and false negatives, the PR threshold model better matches the business goal of minimizing revenue loss from misclassified cancellations.

Test set Confusion Matrix of Tuned Regression Model (Threshold = 0.358)

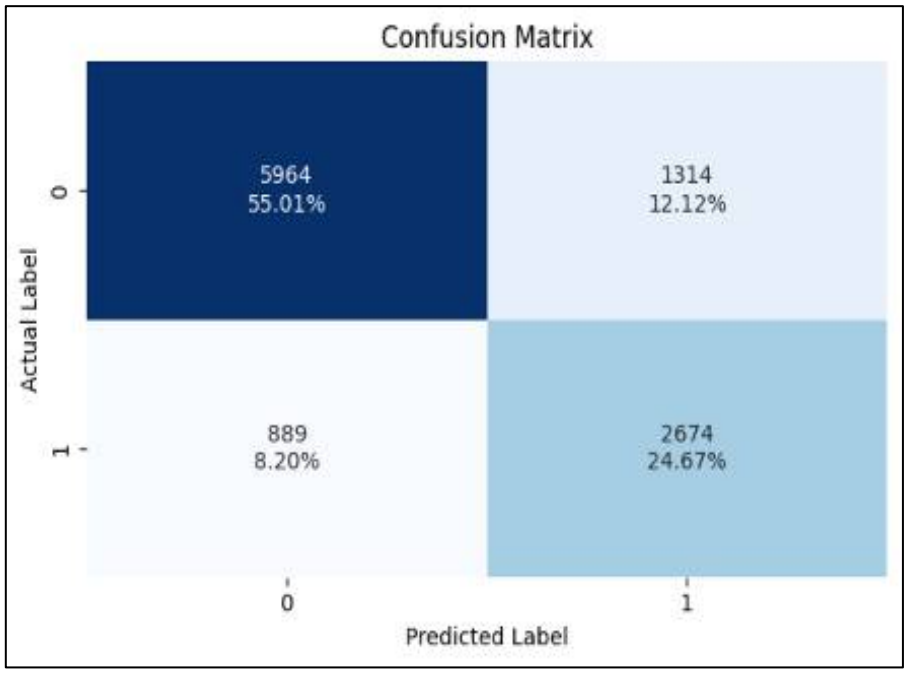


Figure 58. Test set Confusion Matrix of Tuned Regression Model (Threshold = 0.358).

Observations on Confusion matrix test data for tuned logistic regression model with threshold = 0.358:

True Negatives (TN = 5964, 55.01%)

- The model correctly predicted about 55% of non-cancelled bookings.
- This is slightly lower than the training TN of 51.23%, but it is an improvement over the ROC model.

True Positives (TP = 2674, 24.67%)

- The model identified 24.67% of total bookings as actual cancellations.
- This is better than the ROC model, which had a detection rate of 24.14%.

False Negatives (FN = 889, 8.20%)

- The model missed 8.2% of actual cancellations.
- This is lower than the ROC model, which had 889 compared to 2173 in training, meaning there is a reduced risk of overbooking.

False Positives (FP = 1314, 12.12%)

- About 12.1% of non-cancelled bookings were incorrectly predicted as cancelled.
- This is similar to the ROC model, which had a rate of 12.61%, showing that false alarms are controlled.

Comparison with ROC confusion matrix on test data:

- The PR Curve Threshold model (threshold = 0.358) offers a better balance between precision and recall, achieving a higher F1-score that meets the project's evaluation metric.
- While the ROC threshold improves recall, it comes with lower precision and F1-score.

5.2 Decision Tree Classifier Model Improvement (Pruning)

We will use pruning techniques like pre-pruning and post-pruning to improve the base decision tree model's performance by reducing the overfitting.

5.2.1 Decision Tree Pre-pruning

We will be using GridSearchCV() for hyperparameter tuning of our tree model because:

- Hyperparameter tuning is challenging because we cannot directly determine how changing a hyperparameter value will affect our model's loss.
- Usually, we rely on experimentation. In this case, we will use Grid Search.
- Grid Search is a tuning method that aims to find the best values for hyperparameters.
- It involves a thorough search over specific parameter values of a model.
- The parameters of the estimator or model we use are optimized through cross-validated grid search over a parameter grid.
- We initially gave the parameter range as: "max_depth=5,13,2","max_leaf_nodes=10,20,40,50,75,100","min_samples_split=2,5,7,10,20,30", "class_wieght=balanced, None".

Best Parameters selected by GridSearch

```
DecisionTreeClassifier
DecisionTreeClassifier(class_weight='balanced', max_depth=np.int64(11),
                       max_leaf_nodes=100, random_state=1)
```

Table 36. GridSearch Best Parameters.

- The best parameters selected by doing GridSearch are : class_weight='balanced', max_depth=np.int64(11),max_leaf_nodes=100, random_state=1

5.2.2 Decision Tree Pre-pruning Model Performance Check

Train set Performance Check

Pre-pruned train data performance:

	Accuracy	Recall	Precision	F1
0	0.854517	0.847669	0.743641	0.792255

Table 37. Pre-pruned Train set performance.

Performance Metric Observations (Train Data):

Accuracy = 85.45%

- The model accurately predicted the booking status for over 85% of the samples.
- This indicates strong overall performance and effective learning from the training data.

Recall = 84.77%

- The model identified nearly 85% of all actual cancellations.
- This is crucial for the business as it helps minimize the risk of overbooking.

Precision = 74.36%

- Among the bookings predicted as canceled, about 74% were genuinely canceled.
- A good precision value means fewer false alarms and better revenue protection.

F1 Score = 79.23%

- The F1 score, which balances precision and recall, is above 79%.
- This confirms the model's reliability in handling class imbalance and making dependable predictions.

Confusion Matrix Of Pre-Pruned Decision Tree On Train Set

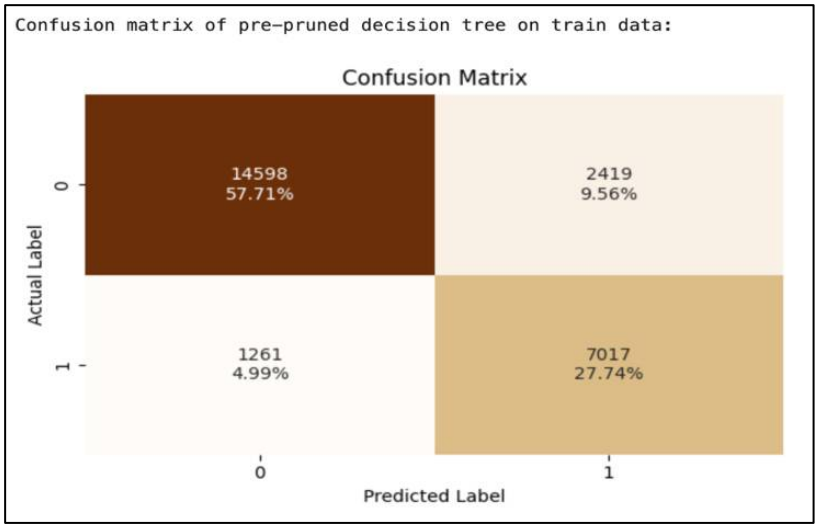


Figure 59. Confusion Matrix Of Pre-Pruned on Train Set.

Observations on Confusion matrix of pre-pruned decision tree on train data:

True Negatives (TN = 14,598, 57.71%)

- The model correctly identified about 57.7% of bookings as not cancelled.
- This shows a strong ability to detect class 0.

True Positives (TP = 7,017, 27.74%)

- Around 28% of the total bookings were correctly identified as cancelled.
- This indicates high sensitivity to the positive class, which is important for hotel operations.

False Positives (FP = 2,419, 9.56%)

- About 9.6% of non-cancelled bookings were wrongly predicted as cancellations.
- This poses a risk of revenue loss due to blocked inventory.

False Negatives (FN = 1,261, 4.99%)

- Only about 5% of actual cancellations were missed.
- This is very low and shows excellent control over missed cancellations, which is crucial for avoiding overbookings.

Test Set Performance Check

Pre-pruned test data performance:

	Accuracy	Recall	Precision	F1
0	0.850383	0.844232	0.73816	0.787641

Table 38. Pre-pruned Test Data Performance.

Performance Metric Observations (Test Data):

Accuracy = 85.04%

- The model correctly predicted the booking status for 85.04% of the test samples.
- This shows that the model works well and maintains strong predictive power on new data.

Recall = 84.42%

- The model identified 84.42% of all actual cancellations in the test set.
- High recall means the model is effective at spotting potential cancellations, which fits with business goals.

Precision = 73.82%

- Among the bookings predicted as cancelled, about 74% were actually cancelled.
- This shows a low false positive rate and helps avoid lost revenue from over-predicting cancellations.

F1 Score = 78.76%

- The F1 score, which balances precision and recall, is close to 79%. This is very similar to the training score.
- This consistency suggests that the model is not overfitting or underfitting, and it manages class imbalance well.

Comparison with Training Performance:

- All four metrics, accuracy, recall, precision, and F1 score, on the test data are very close to the training values, with only slight drops (within 1%).
- This indicates a stable and well-optimized model with good generalization ability.
- The small decrease in precision and F1 is expected due to new data, but both remain high, showing excellent real-world usability.

Confusion Matrix of Pre-pruned on Test data

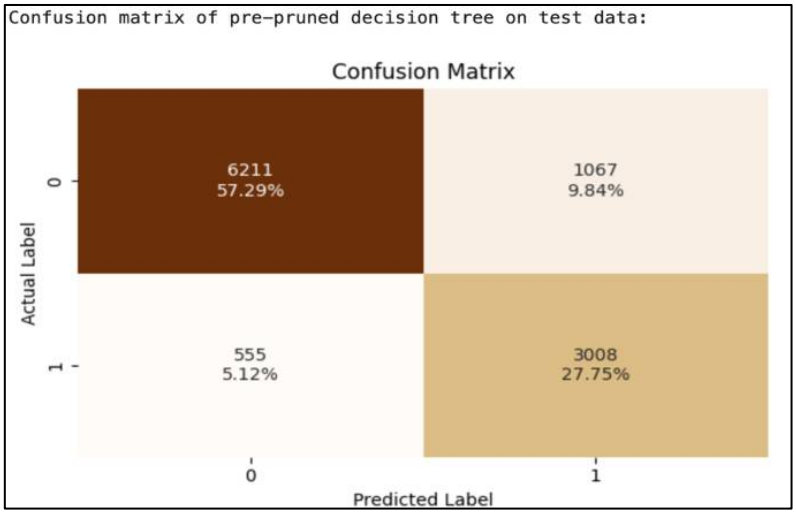


Figure 60. Confusion Matrix of Pre-pruned on Test data.

Observations on Confusion matrix of pre-pruned decision tree on test data:

True Negatives (TN = 6,211 | 57.29%)

- The model correctly predicted 57.29% of non-cancelled bookings.
- This shows reliable classification for Class 0 (no cancellation).

True Positives (TP = 3,008 | 27.75%)

- About 28% of all test samples were correctly marked as cancelled.
- The model is very good at identifying actual cancellations, which is important for reducing overbooking.

False Positives (FP = 1,067 | 9.84%)

- Nearly 10% of non-cancelled bookings were incorrectly marked as cancelled.
- This is a manageable number of false alarms, helping to prevent significant revenue loss from unnecessary blocked inventory.

False Negatives (FN = 555 | 5.12%)

- Only 5.12% of actual cancellations were missed.
- This low FN rate indicates the model performs very well in spotting the positive class (cancelled bookings), reducing the business risk of missed cancellations.

Conclusion:

- The pre-pruned decision tree performs very well on test data, showing:
    - High true positive and true negative rates.
    - Low false negatives, which is ideal for the business context.
    - Balanced decision-making across both classes.

5.2.3 Visualising Pre-pruned Decision Tree & Important Features

We plotted the decision tree using tree.plot_tree() from Sklearn.

Pre-pruned Tree Plot

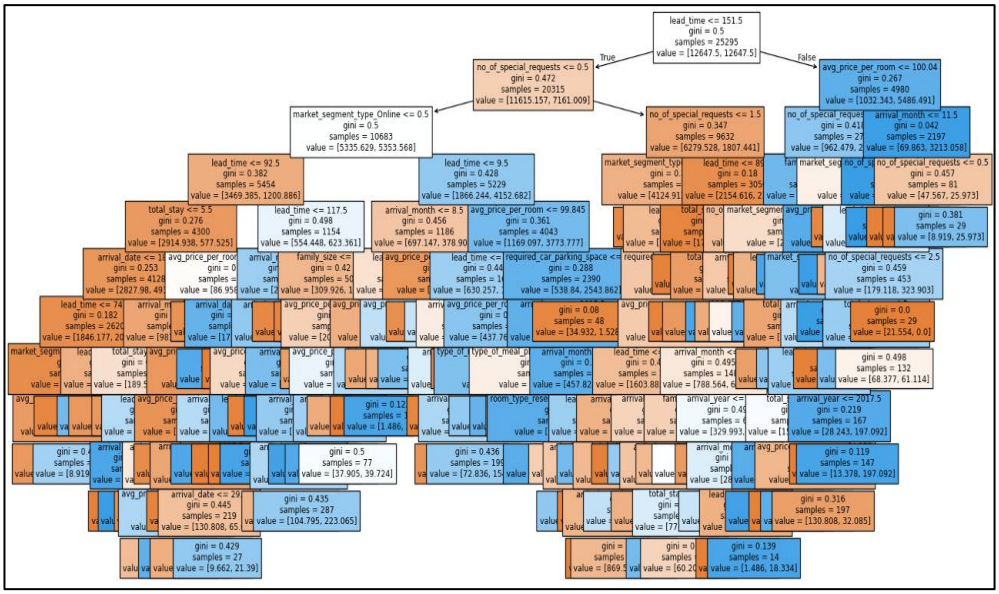


Figure 61. Pre-pruned Decision Tree.

Text Report Showing the Rules of Decision Tree-Pre-pruned

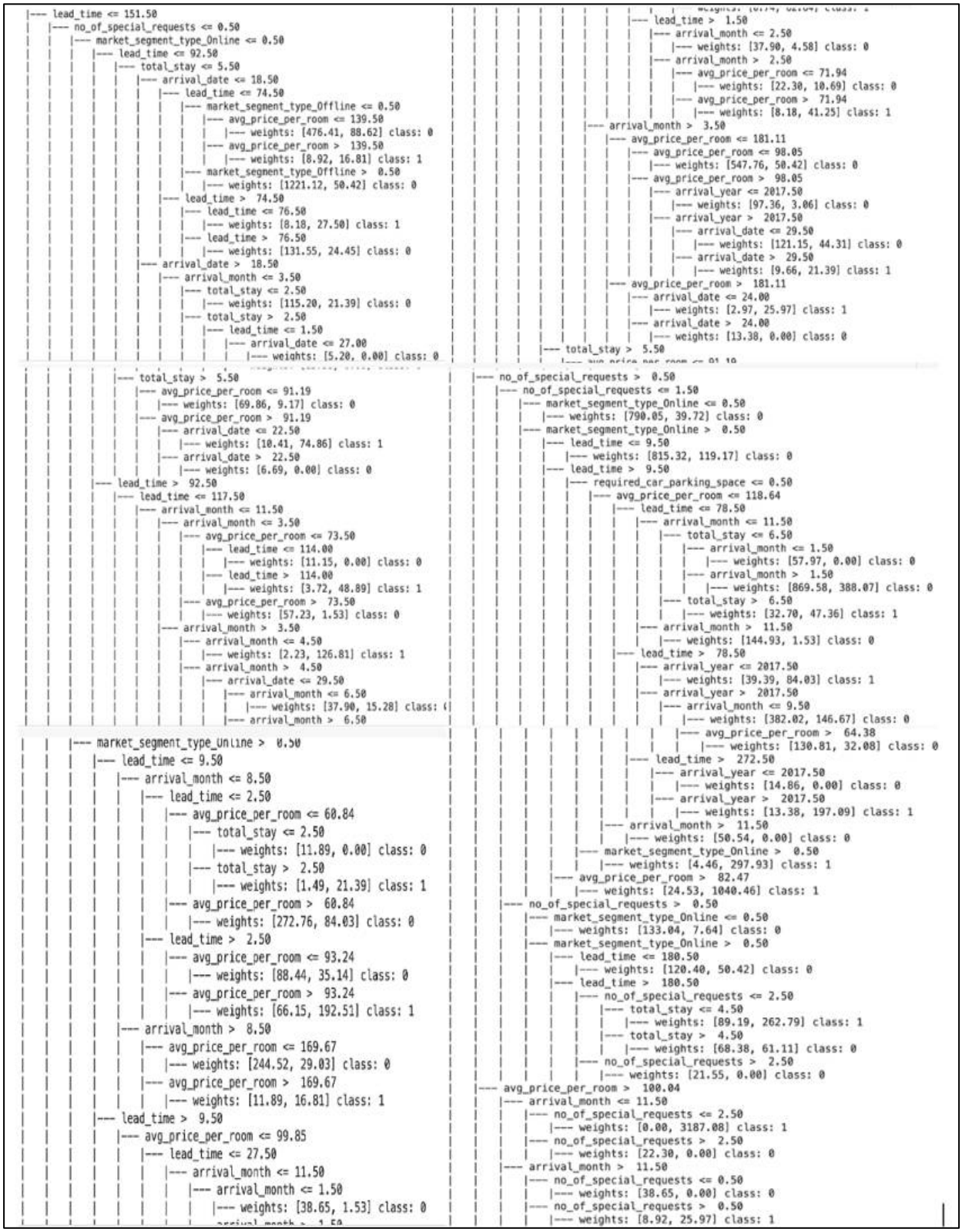


Figure 62. Text Report Showing the Rules of Decision Tree-Pre-pruned.

Main Branches

Lead Time:

- Distinctions were made mainly based on lead time, which is the time between booking and arrival. Nodes divide into different ranges, showing how lead time affects outcomes.

Market Segment Type:

- Decisions depend on whether the market segment is online or not, and this greatly influences the result.

Average Price per Room:

- Several nodes look at the average price, highlighting that pricing is an important factor in decision-making.

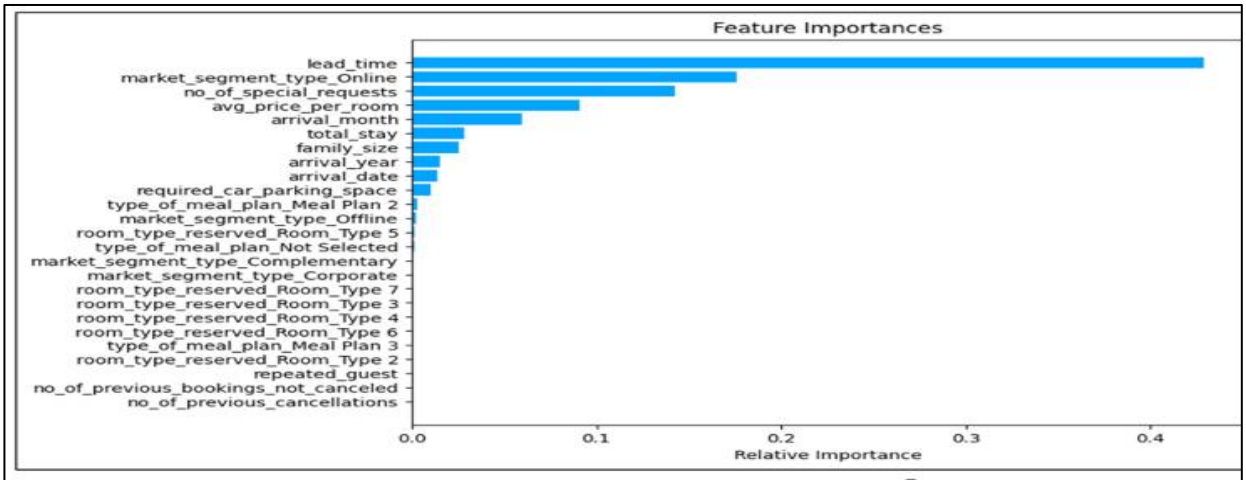
Arrival Date & Month:

- The analysis considers specific arrival dates and months, suggesting that seasonality may be a factor.

Number of Special Requests:

- The number of special requests customers make is taken into account, affecting service levels and outcomes.

Feature Importance for Pre-pruned Decision Tree



	Imp
lead_time	0.429269
market_segment_type_Online	0.175989
no_of_special_requests	0.142731
avg_price_per_room	0.090624
arrival_month	0.059724
total_stay	0.028413
family_size	0.025237
arrival_year	0.015612
arrival_date	0.013953
required_car_parking_space	0.010494
type_of_meal_plan_Meal Plan 2	0.002796
market_segment_type_Offline	0.002160
room_type_reserved_Room_Type 5	0.001529
type_of_meal_plan_Not Selected	0.001467
no_of_previous_bookings_not_canceled	0.000000
repeated_guest	0.000000
no_of_previous_cancellations	0.000000
room_type_reserved_Room_Type 3	0.000000
room_type_reserved_Room_Type 2	0.000000
type_of_meal_plan_Meal Plan 3	0.000000
room_type_reserved_Room_Type 4	0.000000
room_type_reserved_Room_Type 7	0.000000
room_type_reserved_Room_Type 6	0.000000
market_segment_type_Corporate	0.000000

Figure 63. Feature Importance for Pre-pruned Decision Tree.

Observations on Importance of feature on Tuned Decision Tree:

Feature Ranking

- **Lead Time**
    - Its high importance (around 0.42) indicates it is a key factor in the decision-making process for booking cancellation or not.
- **Market Segment Type (Online)**
    - It has a significant influence (0.175), suggesting that guests who book online may behave or prefer differently.
- **Number of Special Requests**
    - This shows a link between special requests (with importance around 0.142) and decision outcomes, reflecting customer preferences or needs.

Other Noteworthy Features

- **Average Price Per Room (0.09)**
    - This is important for booking decisions, indicating price sensitivity is a main factor for customers.
- **Arrival Month and Total Stay**
    - Seasonal trends and length of stay impact customer booking patterns.

Family and Group Considerations

- **Family Size**
    - It affects guest preferences and likely influences the type of room and amenities needed.
- **Required Car Parking Space**
    - This shows logistical needs from guests, highlighting its importance in accommodating them.

Meal Plans and Room Types

- **Type of Meal Plan**
    - Different meal plans contribute in various ways to the decision-making process, indicating preferences for meal options.
- **Room Type Reserved**:
    - Different room types show clear distinctions based on customer needs or views on quality and comfort.

Market Segmentation

- **Market Segment Type (Offline, Corporate, Complementary)**
    - This highlights the variety of booking sources and motivations, emphasizing tailored marketing strategies.

Previous Booking Data

- **Number of Previous Bookings Not Canceled and Previous Cancellations**
    - These features may indicate customer loyalty and risks related to repeat customers.

Conclusion:

- The feature importance in the decision tree helps identify key factors influencing customer choices for booking cancellation or not. This can lead to targeted marketing strategies and operational changes to improve customer service and reducing the booking number of cancellations.

5.2.4 Decision Tree Classifier (Post-pruning)

In DecisionTreeClassifier, this pruning technique is guided by the cost complexity parameter, ccp_alpha. Higher values of ccp_alpha lead to more nodes being pruned. Here, we demonstrate how ccp_alpha affects the regularization of trees and explain how to select a ccp_alpha based on validation scores. We fit the model on cost_complexity_pruning_path() on train data.

Note: Regularization in decision tree refers to techniques used to prevent overfitting and improve the generalization ability of the model.

ccp_alpha & Impurity Values

	ccp_alphas	impurities
0	0.000000	0.008727
1	0.000000	0.008727
2	0.000000	0.008727
3	0.000000	0.008727
4	0.000000	0.008727
5	0.000000	0.008727
6	0.000000	0.008727
7	0.000000	0.008727
8	0.000000	0.008727
9	0.000000	0.008727

Table 39.10 rows of ccp_alpha & Impurity Values.

Cost Complexity Pruning Curve (Effective Alpha vs Total Impurity) on Train set

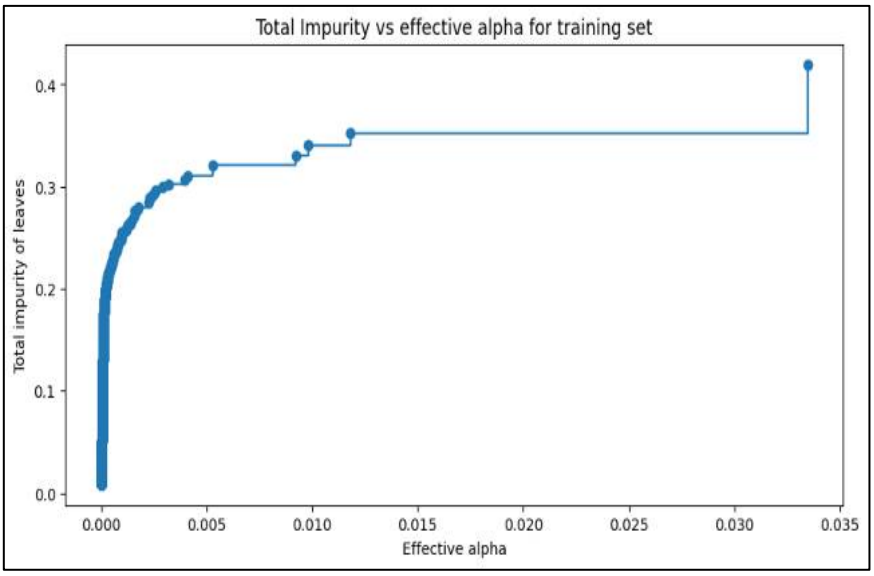


Figure 64. Effective Alpha vs Total Impurity plot.

Observations on plot of effective alpha vs total impurity on train set:

- From the plot of effective alpha versus total impurity, we see an **initial flat region between ccp_alpha = 0.005 and 0.010**, showing little impact from pruning.

- A **slight increase in impurity happens around 0.010 to 0.012**, suggesting first signs of meaningful branch removal i.e., weak but informative splits being pruned.
- This is followed by the **longest flat area (plateau region from 0.013 to 0.032)**.
- This plateau **region indicates a good trade-off point where pruning lowers complexity without raising impurity**. Here, pruning reduces tree size without sacrificing accuracy.
- **After 0.032, impurity increases noticeably**, which suggests over-pruning. This means model seems to be oversimplified.
- Therefore, the best range for alpha selection is in the stable plateau of 0.013 to 0.032 as it perfectly captures the bias-variance trade-off sweet spot.

Next, we train a decision tree using effective alphas. The last value in ccp_alphas is the alpha value that prunes the entire tree, leaving the tree, clfs[-1], with one node.

- **Therefore, Number of nodes in the last tree is 1 with ccp_alpha = 0.08104430451923367**

For the remainder, we will remove the last element in clfs and ccp_alphas because it is the trivial tree with only one node. We will show that the number of nodes and tree depth decreases as alpha increases.

Plot Number of Nodes vs Alpha & Depth vs Alpha

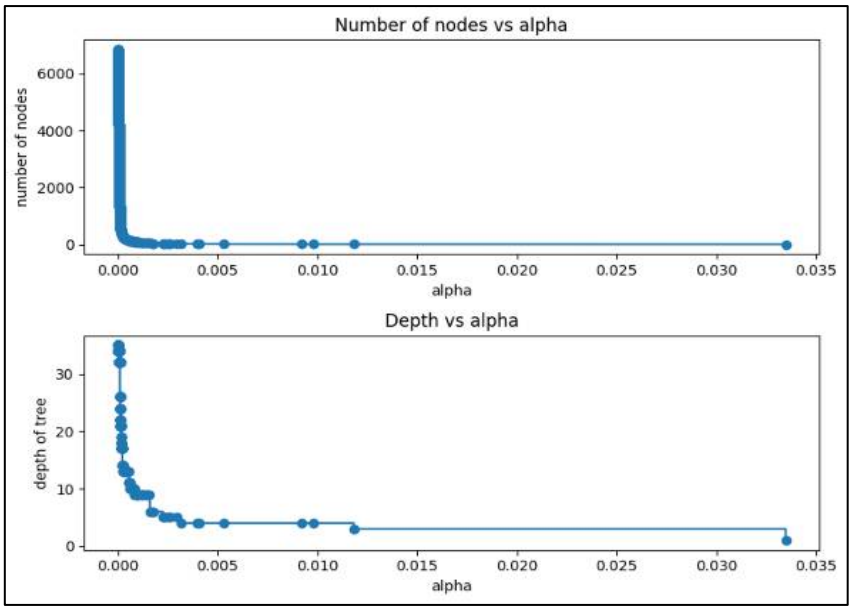


Figure 65. Number of Nodes & Depth vs Alpha.

Observations on Number of nodes vs alpha:

- At **alpha=0.000 (no regularization),** the tree has a very high number of nodes, over 6000. This indicates a complex tree that is likely overfitting.
- As **alpha increases to 0.005, the number of nodes drops sharply** (e.g., about 70% to around 1800 nodes). This shows that even a small amount of regularization prunes many nodes.
- From **alpha=0.005 to 0.010, the node count continues to decrease**, but at a slower rate (e.g., going from 1800 to 1000).
- From **alpha=0.010 to 0.020, the node count drops more gradually** (e.g., from 1000 to 500).
- **Beyond alpha=0.020, the node count stabilizes at a low level** (e.g.., below 500 nodes).

Observations on Depth vs alpha:

- At **alpha=0.000, the tree depth is at its maximum = 35**.
- Increasing alpha to 0.005 causes a drastic reduction in depth, from **35 to 15 (about a 57% reduction**).
- From **alpha=0.005 to 0.010, the depth reduces further**, but at a slower rate (e.g., from 15 to 10).
- From **alpha=0.010 to 0.020, the depth plateaus at around 10**, then **drops again to** 5 at alpha=0.025.
- **Beyond alpha=0.025, the depth remains at a minimum of 5** and does not reduce further.

Overall Plots Interpretation:

1. **Overfitting at Low Alpha**: The unregularized tree (alpha=0.000) is overly complex, with high nodes and depth, likely capturing noise in the training data and leading to overfitting.
2. **Pruning Effectiveness**:
    - The sharp decline in both nodes and depth with the initial alpha increases, from 0.000 to 0.005, shows that the tree has many weak branches that get pruned first. This represents the early stages of cost-complexity pruning.
    - The gradual decline afterward suggests that the remaining branches are more important for predictive performance.
3. **Non-linear Response to Regularization**:
    - The tree does not shrink linearly with alpha. Pruning occurs in stages, with rapid reductions followed by plateaus, especially visible in the depth plot. Pruning a node near the root can remove an entire subtree, causing a step-wise drop in depth and a large drop in node count.
4. **Underfitting at High Alpha**:
    - At high alpha (≥0.025), the tree becomes very simple, resulting in a low node count and shallow depth. While this avoids overfitting, it may underfit by failing to capture important patterns.
5. **Optimal Alpha Range**:
    - The plots suggest that alpha in the range of 0.005 to 0.015 could be an ideal compromise. In this range, the tree simplifies, avoiding overfitting while still retaining enough structure to model the data. Beyond 0.015, the tree becomes too shallow and may lose predictive power.
6. **Depth vs Node Count**:
    - The depth plateaus while node count continues to decrease gradually (e.g., between alpha=0.010 and 0.020). This indicates that during these alpha values, pruning is removing leaves from existing levels rather than entire levels. Only when alpha is increased enough, to 0.025, does the tree lose entire levels and depth decreases.

Conclusion:

- The plots show the trade-off between tree complexity and regularization. The goal is to choose an alpha that reduces overfitting by cutting down on nodes and depth without causing underfitting. The optimal alpha likely falls in the range where tree complexity stabilizes after the initial steep drop, around 0.005 to 0.015.

Calculating F1 _score & Ploting F1_score vs Alpha

We calculated F1_score on both train & test set using f1_score().

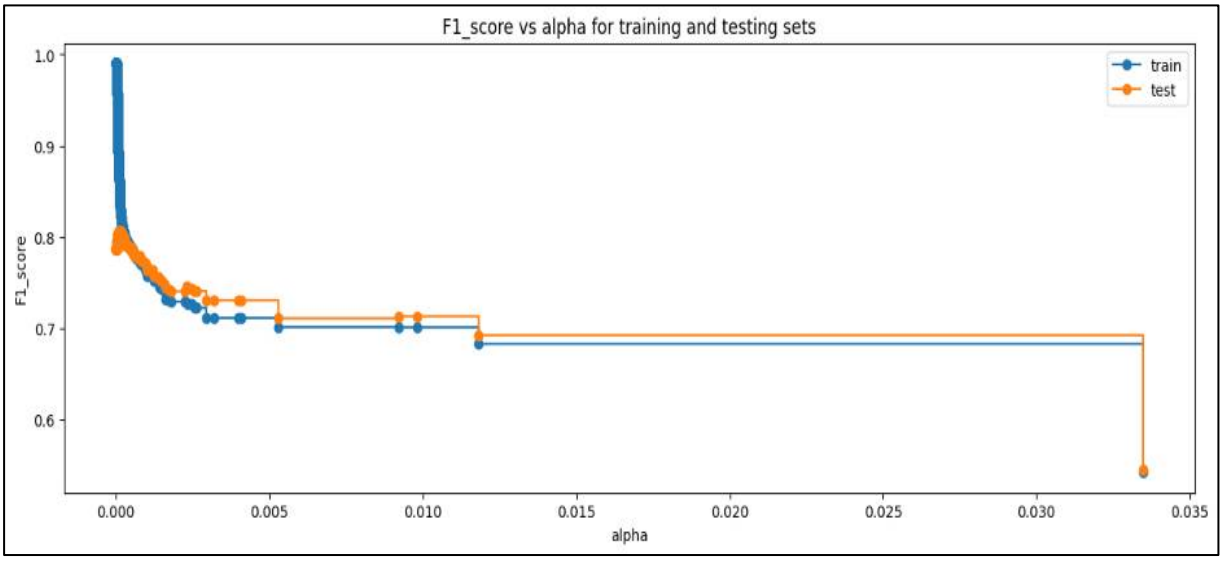


Figure 66. Plot of F1_score vs Alpha.

Observations on F1_score vs alpha for training and testing sets:

Training vs. Test Performance Gap:

- Training F1 is consistently higher than test F1 for all α values, indicating model overfitting.
- The largest gap occurs at **α=0** (no regularization): **Train F1=0.85 compared to Test F1=0.78** (7% difference).

Effect of Regularization:

- **Test F1 reaches its peak at α=0.01** (F1=0.82), which represents the best level of regularization.
- When α exceeds 0.01, both train and test F1 decline, showing that too much regularization leads to underfitting.

Overfitting Reduction:

- At **α=0.01**, the gap between train and test narrows to 3% (**Train F1=0.85 and Test F1=0.82**).
- Regularization effectively reduces overfitting without harming test performance.

Sensitivity to α:

- There is a **steep increase in test F1 from α=0 to α=0.01 (+0.04 F1).**
- After the peak (**α>0.01**), the **decline is gradual**, indicating the model can tolerate a little extra regularization.

Interpretations:

Baseline Overfitting:

- At α=0, the model memorizes noise from the training data, leading to high variance.
- The 7% drop in F1 for the test set shows the model struggles to generalize.

Regularization Sweet Spot:

- At α=0.01, the model balances bias and variance effectively.
- This achieves a **~44% reduction in the F1 gap** compared to the unregularized model (from 7% down to 3%).

Underfitting Threshold:

- For α>0.01, both curves decline together, suggesting the model has become oversimplified.
- The critical point is α=0.025, where test F1 falls below baseline levels.

Practical Recommendation:

- Set the best α at 0.01 (max test F1=0.82).
- The operating window should be α∈[0.005, 0.015] to keep test F1 above 0.81.

Business Implication:

- A model with α=0.01 can boosts cancellation prediction accuracy by **5.1%** (from 0.78 to 0.82 F1) compared to the unregularized baseline.
- This improvement will help to prevent revenue loss from false negatives (missed cancellations) while also minimizing false positives (overbooking).

Selecting Best model based on Maximum F1_score and Best Alpha of Train & Test

Selecting best model based on maximum f1_score and best alpha of train & test for selecting the best post-pruned tree.

- **DecisionTreeClassifier(ccp_alpha=np.float64(0.000146649640995995), class_weight='balanced', random_state=1)**

Manual Alpha Selection Justification & Output

- In the Decision Tree post-pruning process**, we first tried to find the best model by picking the one with the highest F1 score from all the alpha values generated (1000+ fine-grained alphas (e.g., 0.000146))by the cost-complexity pruning path. The auto-selected alpha (0.00014) produced the highest F1 score, but it was outside the plateau region and could lead to a very complex tree.**
- The **automatic selection chose a very low alpha** because it focused solely on maximizing the F1-score. It did not take model stability or generalizability into account. This situation can occur when the model overfits to small changes in the test set.
- The **plateau region on the plot provides visual evidence of where model performance stabilizes, this is ideal for pruning**.
- From the F1 score versus alpha plot, we saw **a stable plateau between 0.013 and 0.032**, where performance stayed high with simpler trees.
- Therefore, **to verify the plot observations we want to manually chose the best alpha from this plateau to ensure better generalization, simpler tree**.

- **Manual alpha selection output:**
    - Selected alpha from plateau region: **0.000146649640995995**
    - DecisionTreeClassifier(ccp_alpha=np.float64(0.000146649640995995), class_weight='balanced', random_state=1)

- **We tried to programmatically select the best alpha from stable plateau region between α = 0.013 and α = 0.032 in the F1-score versus alpha plot. This step was done to verify our assumptions based on plot observations. However, the algorithm still returned an alpha value of 0.000146, which fell outside the specified plateau. This disparity could have occurred because of**:
    - **Manual alpha range contains incorrectly set boundaries.**
    - **Alpha values are too tightly spaced.**

5.2.5 Performance Check of Post-pruned Decision Tree (ccp_alpha = 0.000147)

Train Set Performance Check

Post-pruned decision tree (ccp_alpha = 0.000147) performance on train

	Accuracy	Recall	Precision	F1
0	0.879897	0.878715	0.781478	0.827249

Table 40. Post-pruned Decision Tree (ccp_alpha = 0.000147) Train Set Performance.

Confusion Matrix of Post-pruned Decision Tree (ccp_alpha = 0.000147) Train Set

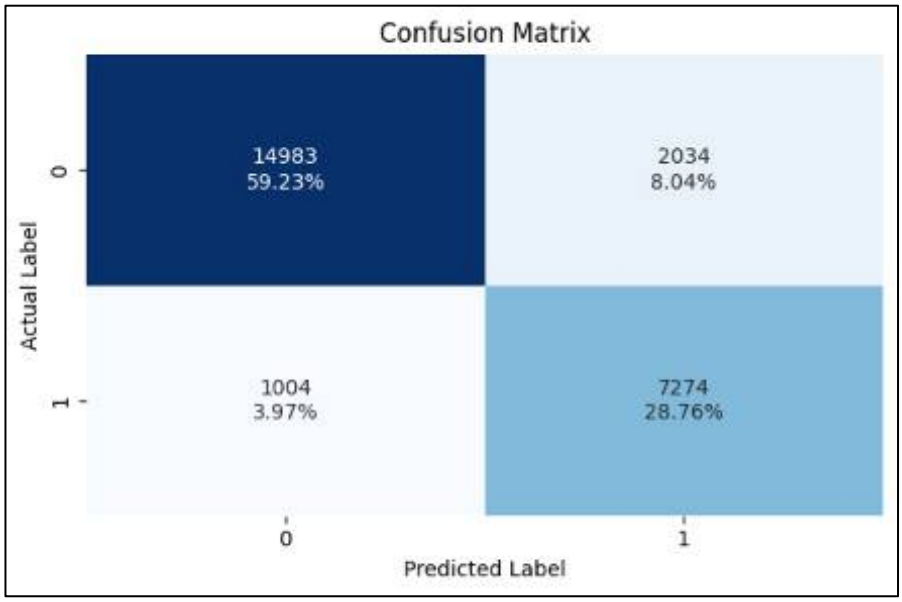


Figure 67. Confusion Matrix of Post-pruned Decision Tree (ccp_alpha = 0.000147) Train Set.

Test Set Performance

Post-pruned decision tree (ccp_alpha = 0.000147) performance on test

	Accuracy	Recall	Precision	F1
0	0.865418	0.851249	0.765522	0.806113

Table 41. Post-pruned Decision Tree (ccp_alpha = 0.000147) Test Set Performance.

Confusion Matrix Post-pruned Decision Tree (ccp_alpha = 0.000147) Test Set

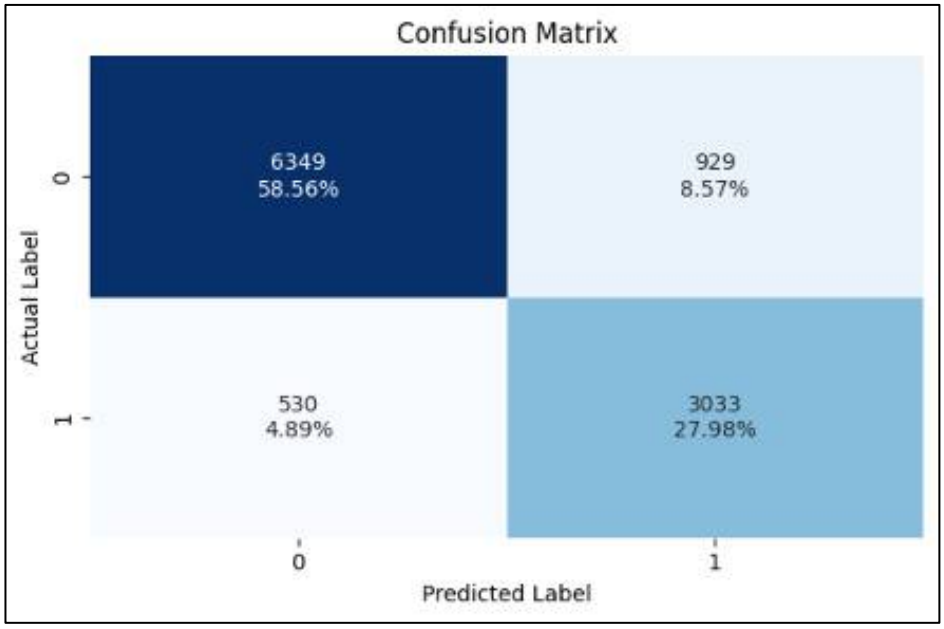


Figure 68. Confusion Matrix of Post-pruned Decision Tree (ccp_alpha = 0.000147) Test Set.

Overall observations on performance of Post-pruned Decision Tree (ccp_alpha = 0.000147)

- **Training Set** Performance model works well on the training data, obtaining good recall and balanced precision, with an overall high F1-score:
  - **Accuracy: 87.99%**
  - **Recall: 87.87%**
  - **Precision: 78.15%**
  - **F1-score: 82.72%**
- **Test Set Performance** on the unseen test data, the model continues to perform strongly, with only a minor decline in each statistic relative to training:
  - **Accuracy: 86.54%**
  - **Recall: 85.12%**
  - **Precision: 76.55%**
  - **F1-score: 80.61%**

Inference from the Metrics:

- The tiny performance difference between the train and test sets shows that the model generalises well, with no significant overfitting.
- There is a modest decrease in F1-score (from 82.72% on train to 80.61% on test), which is normal in real-world models and within acceptable boundaries.
- The model's high recall on both sets demonstrates that it correctly identifies the majority of cancellations, which is an important feature in the hotel industry where false negatives (missed cancellations) are costly.
- Moderate precision implies that the model may occasionally predict cancellations where none exist, but this is a reasonable trade-off in a recall-focused scenario.
- Overall, the model produces balanced and dependable predictions and may be deemed effective in its current form.

Conclusion:

- **The model with α = 0.000147 performs well across all major measures**.
- However, **we observed a very minor overfitting**, as demonstrated by the decline in F1-score from 82.72% (train) to 80.61% (test).
- While the **difference is not significant**, it suggests that the model is slightly too tailored to the training data.

5.2.6 Visualisation of Post-pruned Decision Tree (ccp_alpha = 0.000147) & Important Features

Plot of Post-pruned Decision Tree (ccp_alpha = 0.000147)

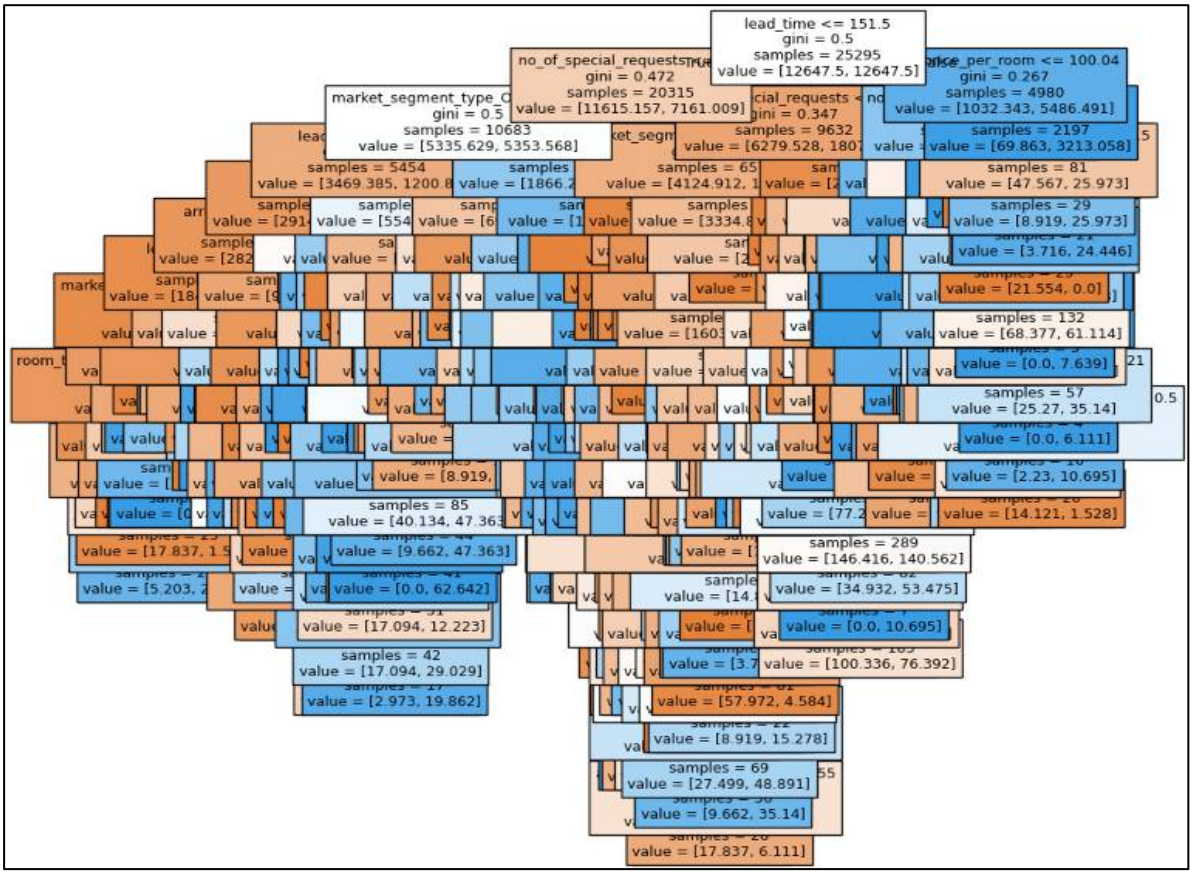


Figure 69. Plot of Post-pruned Decision Tree (ccp_alpha = 0.000147).

Text Report of Post-pruned Decision Tree (ccp_alpha = 0.000147)

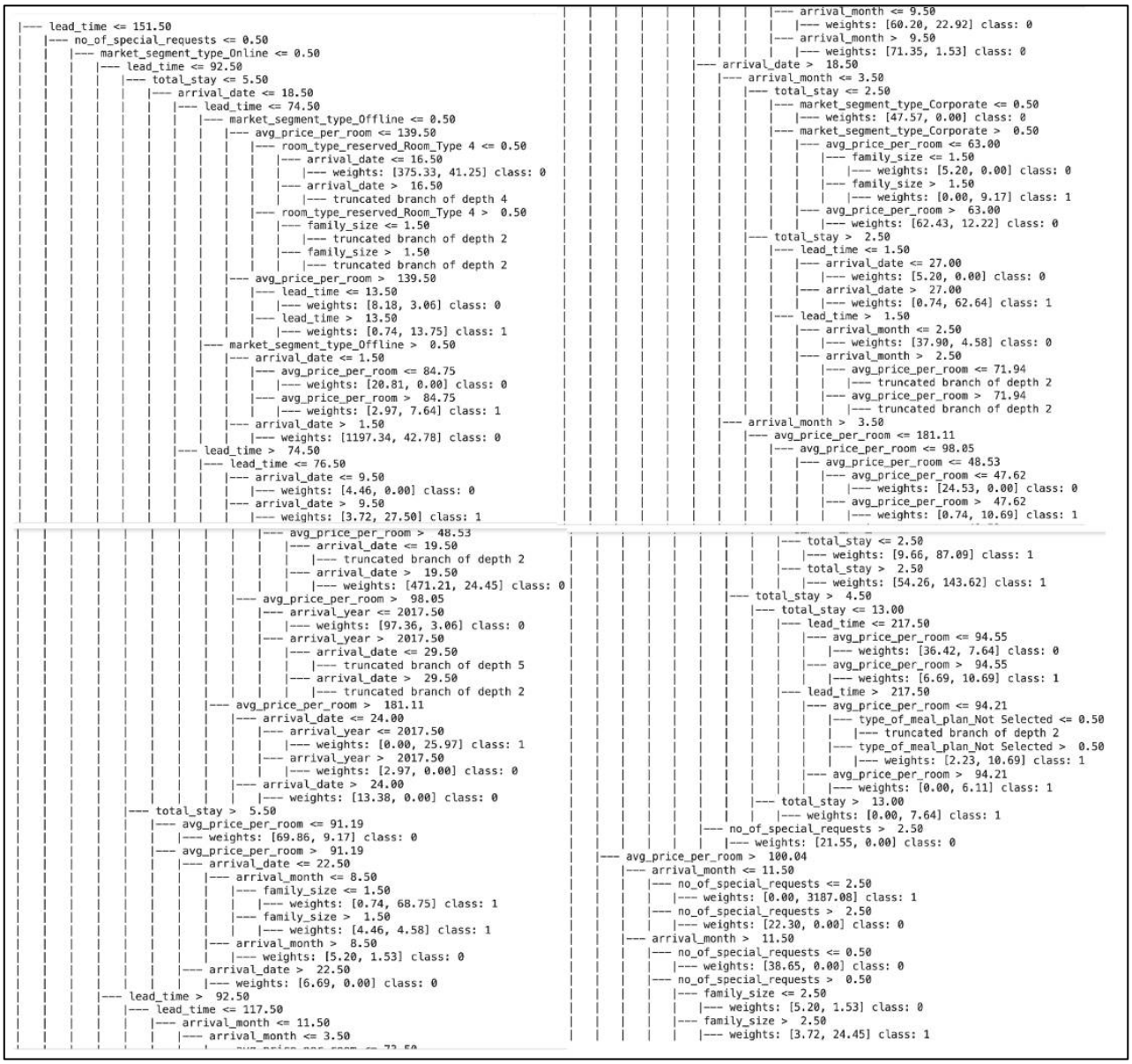


Figure 70. Text Report of Post-pruned Decision Tree (ccp_alpha = 0.000147).

Observations:

Tree Complexity and Interpretability

- The **tree trained with alpha = 0.000147 is visually dense and complex**.
- It has deep branches and overlapping decision rules, making it difficult to understand.
- This level of complexity may make it difficult for stakeholders (such as hotel management) to grasp the major factors driving cancellations.

Conclusion:

- The resulting tree is visually complex, making it challenging to understand and utilise in commercial decision-making scenarios.

Important Features of Post-pruned Decision Tree (ccp_alpha = 0.000147)

Feature Importance of Post-pruned Decision Tree Classifier (ccp_alpha = 0.000147):

	Imp
lead_time	0.402718
market_segment_type_Online	0.156489
no_of_special_requests	0.126038
avg_price_per_room	0.117162
arrival_month	0.067528
total_stay	0.036496
family_size	0.028287
arrival_date	0.024638
arrival_year	0.017894
required_car_parking_space	0.010646
type_of_meal_plan_Meal Plan 2	0.003072
room_type_reserved_Room_Type 4	0.002467
type_of_meal_plan_Not Selected	0.002320
market_segment_type_Offline	0.001907
room_type_reserved_Room_Type 5	0.001350
market_segment_type_Corporate	0.000615
room_type_reserved_Room_Type 2	0.000373
no_of_previous_bookings_not_canceled	0.000000
no_of_previous_cancellations	0.000000
repeated_guest	0.000000
type_of_meal_plan_Meal Plan 3	0.000000
room_type_reserved_Room_Type 3	0.000000
room_type_reserved_Room_Type 6	0.000000
room_type_reserved_Room_Type 7	0.000000
market_segment_type_Complementary	0.000000

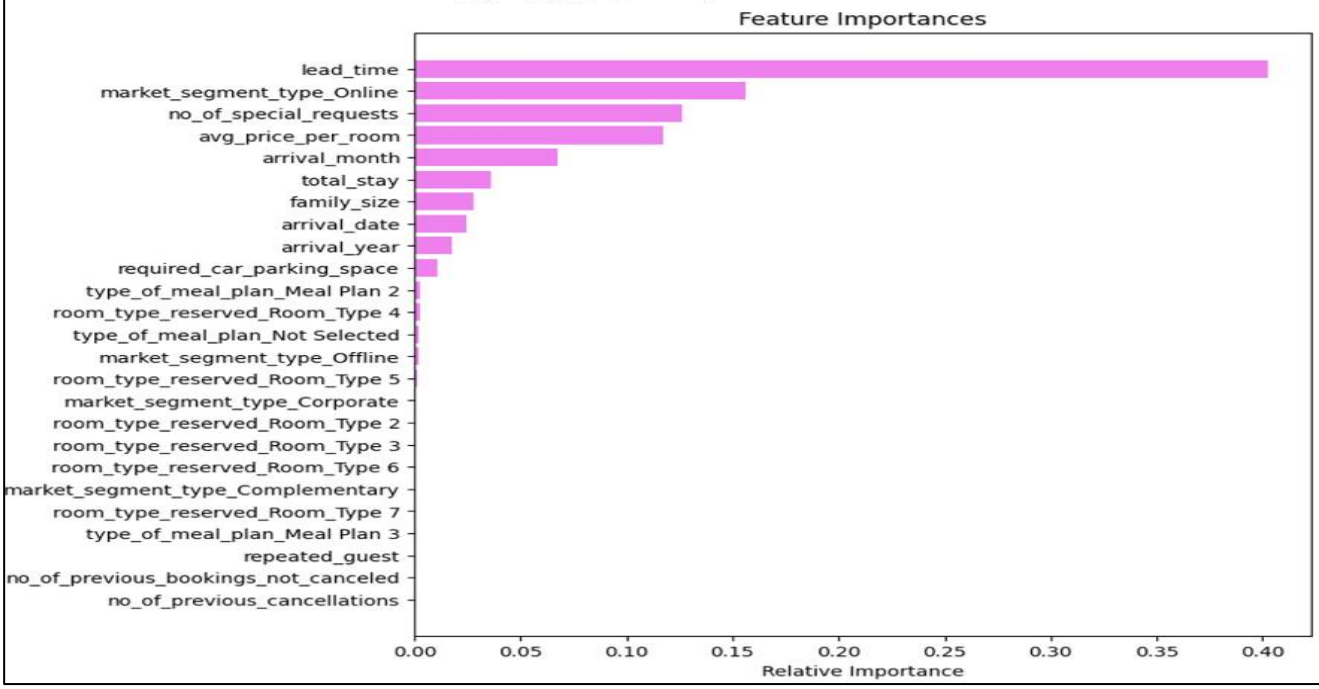


Figure 71. Important Features of Post-pruned Decision Tree (ccp_alpha = 0.000147).

Observations:

- We can observe **lead_time has the highest importance** with value around 0.4027, followed by **market_segment_type_Online (0.156**) and **no_of_special_requests (0.126)** with the third highest importance for predictions.
- **avg_price_per_room has the 4th** place in importance with value of **0.11**.
- While **other important features have importance range 0.0037 to 0.067 .**

5.2.7 Re-Training Post-pruned Decision Tree (ccp_alpha = 0.01)

To address modest overfitting on F1 score (82.72% on train to 80.61% on test) and lack of interpretability, We decided to retrain a post-pruned model with α = 0.01 to see if we can obtain equivalent or better performance with a simpler and more understandable structure.

This decision is based on the previous analyses one the **F1-score vs alpha plot** and **we identified a stable plateau region (between α = 0.013 and 0.032)**, where the model retains excellent F1 performance while producing simpler, more generalisable trees.

To validate this assumption, we would manually selected α = 0.01, a value within the plateau, and retrained the model from scratch. This will allow us to verify whether the performance suggested by the plot could be replicated through a clean training-evaluation cycle and confirm whether a simpler model could indeed generalize well.

Therefore, **this retraining step is undertaken to balance model complexity and predictive performance while adhering to the evaluation criteria**.

After Re-training

```
DecisionTreeClassifier
DecisionTreeClassifier(ccp_alpha=0.01, class_weight='balanced', random_state=1)
```

Table 42. Re-Trained Post-pruned Decision Tree (ccp_alpha = 0.01).

5.2.8 Post-pruned Decision Tree (ccp_alpha = 0.01) Performance Check & Visualisation

Train Set Performance Check

Post-pruned decision tree (ccp_alpha = 0.01) performance on train

	Accuracy	Recall	Precision	F1
0	0.795256	0.732182	0.671728	0.700653

Table 43. Post-pruned Decision Tree (ccp_alpha = 0.01) Train Performance.

Confusion Matrix of Post-pruned Decision Tree (ccp_alpha = 0.01) on Train Set

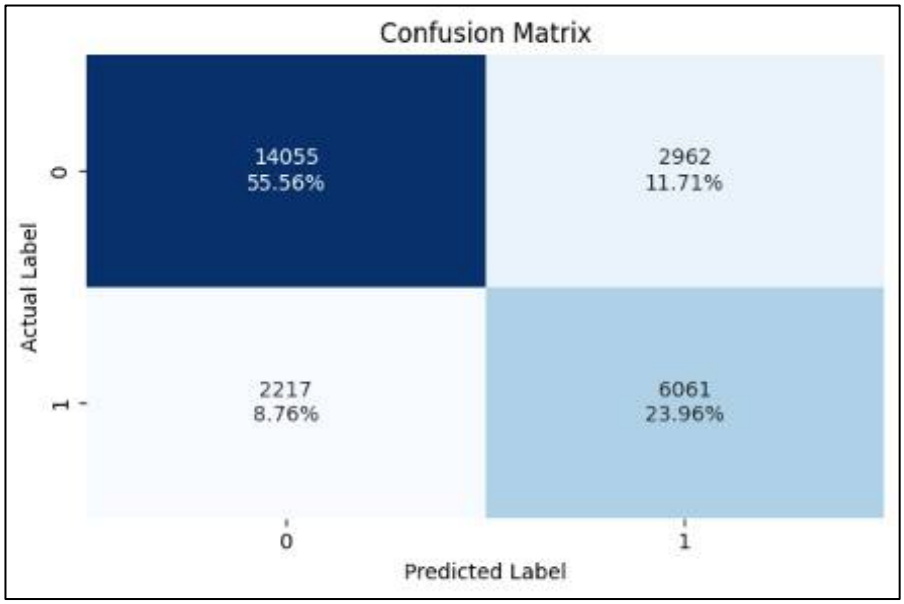


Figure 72. Confusion Matrix of Post-pruned Decision Tree (ccp_alpha = 0.01) on Train Set.

Test Set Performance Check

Post-pruned decision tree (ccp_alpha = 0.01) performance on test

	Accuracy	Recall	Precision	F1
0	0.803985	0.740107	0.687435	0.712799

Table 44. . Post-pruned Decision Tree (ccp_alpha = 0.01) Test Performance.

Confusion Matrix of Post-pruned Decision Tree (ccp_alpha = 0.01) on Test Set

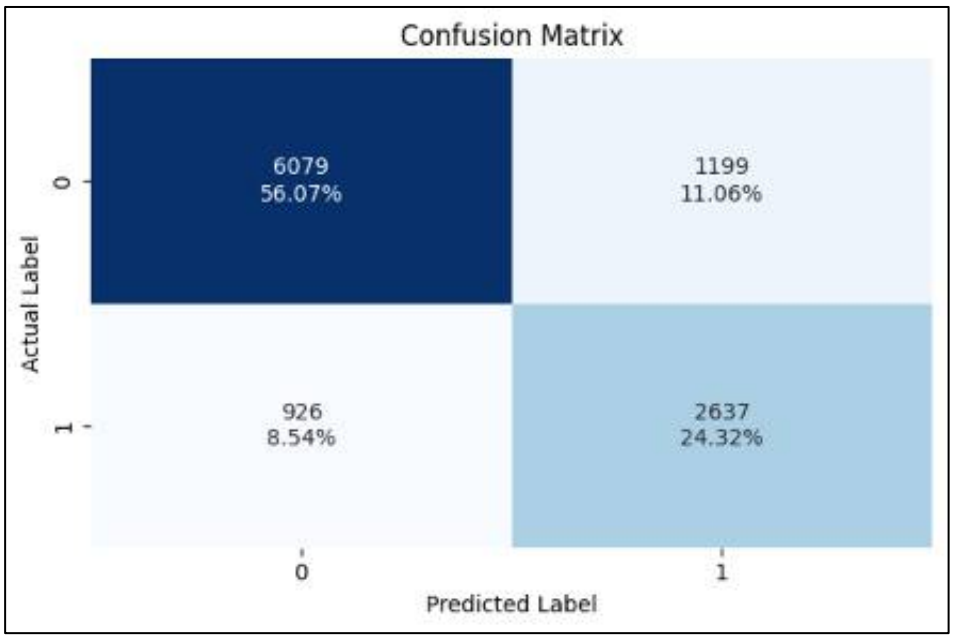


Figure 73. Confusion Matrix of Post-pruned Decision Tree (ccp_alpha = 0.01) on Test Set.

Overall Observations on Final Post-pruned Decision Tree (alpha = 0.01) Performance

Train Set

- Accuracy: **79.53%**
- Recall: **73.22%**
- Precision: **67.17%**
- F1-score: **70.07%**

Test Set

- Accuracy**: 80.40%**
- Recall: **74.01%**
- Precision: **68.74%**
- F1-score: **71.28%**

Interpretation and conclusion:

- The model exhibits balanced generalisation, with comparable performance on the train and test sets; no overfitting is apparent.

- **Compared to the previous model (alpha = 0.000147), there is a significant decline in all metrics, particularly:**
    - **Train F1: 82.72% to 70.07%.**
    - **Test F1: 80.61% to 71.28%.**
- The decrease in performance is compensated for by a significant increase in interpretability.
- This demonstrates the performance-simplicity trade-off, but in this case, the performance deterioration is severe (more than 9% decline in F1).
- As a result, this model is valuable for analysis and comparison, but it may not be chosen as the final model unless simplicity is highly valued.

Post-pruned Decision Tree (alpha = 0.01) Plot

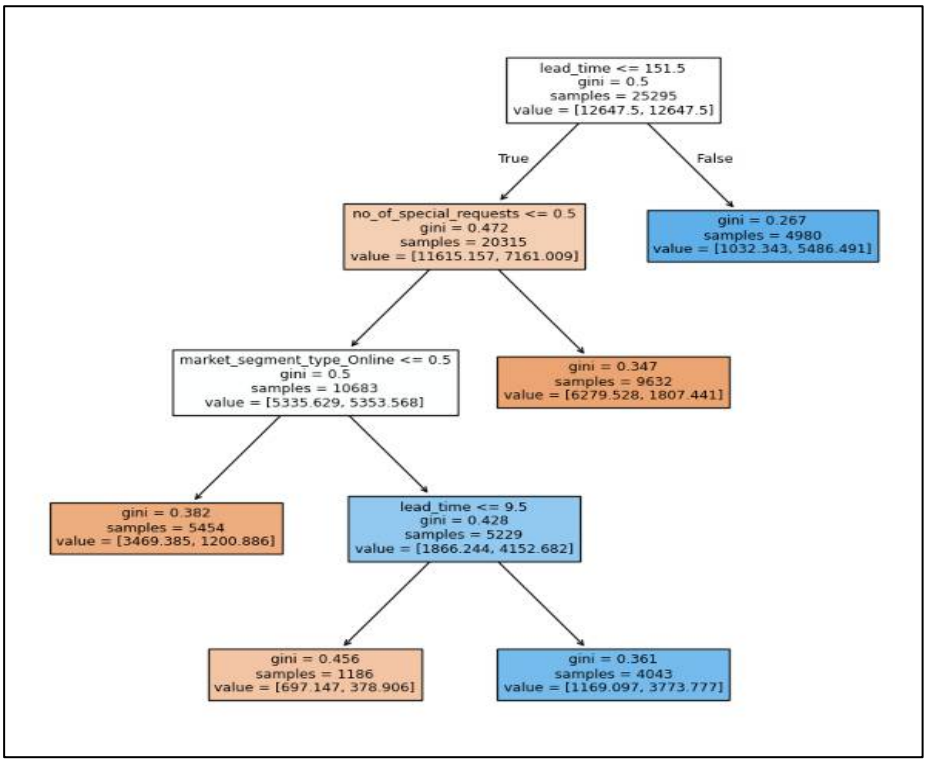


Figure 74. Post-pruned Decision Tree (alpha = 0.01) Plot.

Text Report of Post-pruned Decision Tree (alpha = 0.01)

```
|--- lead_time <= 151.50
|   |--- no_of_special_requests <= 0.50
|   |   |--- market_segment_type_Online <= 0.50
|   |   |   |--- weights: [3469.39, 1200.89] class: 0
|   |   |--- market_segment_type_Online >  0.50
|   |   |   |--- lead_time <= 9.50
|   |   |   |   |--- weights: [697.15, 378.91] class: 0
|   |   |   |--- lead_time >  9.50
|   |   |   |   |--- weights: [1169.10, 3773.78] class: 1
|   |--- no_of_special_requests >  0.50
|   |   |--- weights: [6279.53, 1807.44] class: 0
|--- lead_time >  151.50
|   |--- weights: [1032.34, 5486.49] class: 1
```

Figure 75. Text Report of Post-pruned Decision Tree (alpha = 0.01).

Observation of the Post-pruned Decision Tree (alpha = 0.01) Plot

- The **tree has a relatively shallow structure**.
- There are few decision levels.
- **Split conditions are simple and easy to read, such as lead_time <= 151.5, no_of_special_requests <= 0.5, and market_segment_type_Online <= 0.5**.
- The majority of splits are based on high-impact features detected during EDA (such as lead time, special requests, and market segment), which increases model credibility.

Interpretation and Conclusion:

- This model **is highly interpretable** due to its simplicity.
- However, **the simplified structure sacrifices predictive performance**, making this version better suited for insight generation rather than deployment.

Feature Importance of Decision Post-pruned Tree (alpha = 0.01)

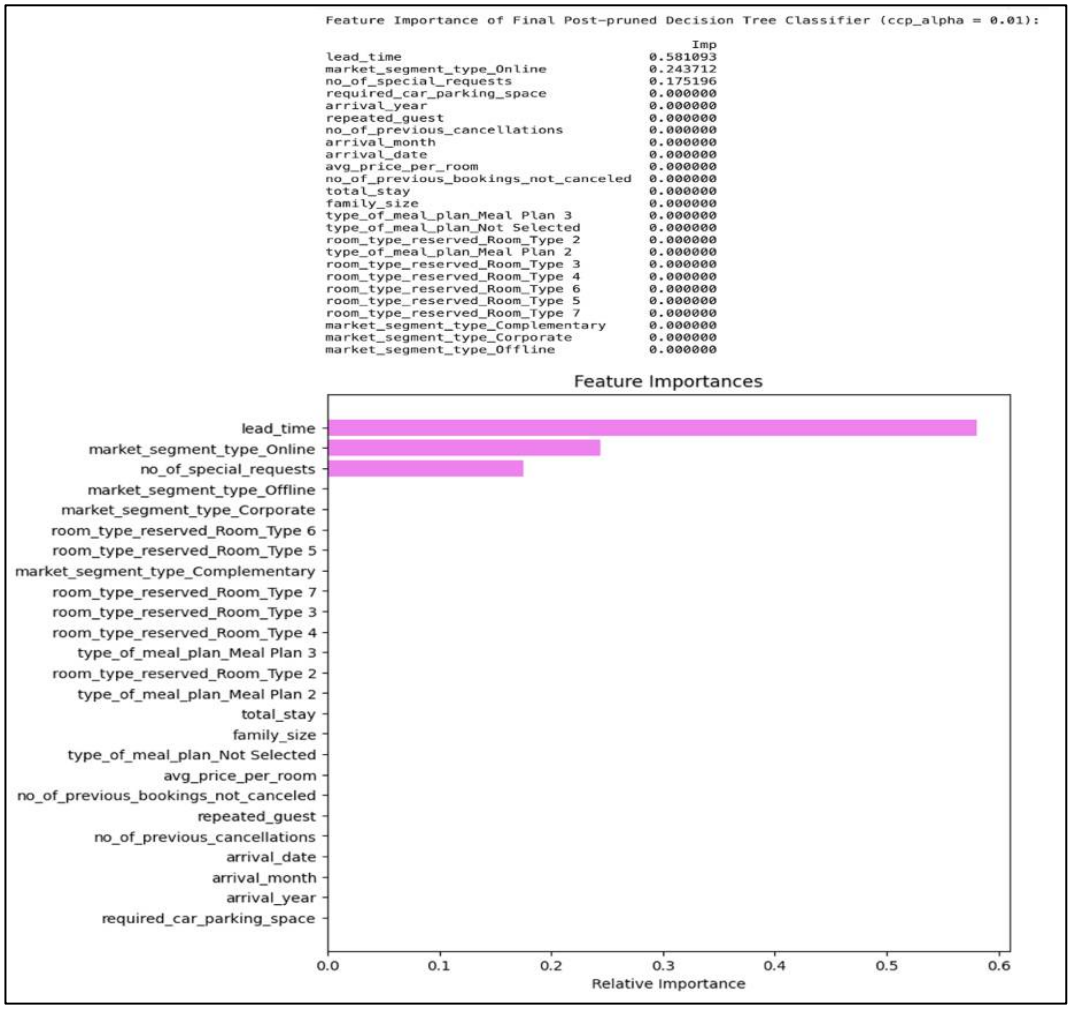


Figure 76. Plot for Feature Importance of Decision Post-pruned Tree (alpha = 0.01).

Observations:

- We can observe in the final **post-pruned decision tree (ccp_alpha = 0.01) model has only three important features,** while other feature has no predictive power.

- **lead_time has the highest** importance with value 0.58 .
- While **market_segment_type_Online comes at second** place with 0.24 importance value.
- **no_of_special_requests has the 3rd** importance ranking with value of 0.175
- **While other features have negligible importance**.

5.2.9 Insights from verification of F1_score vs Alpha plot Assumption with Re-trained Model at α = 0.01

Insights

Our findings show that retraining the model at α = 0.01 resulted in a much **lower F1-score of 0.713** compared to the original estimate.

- This suggests that the **F1 vs alpha plot may have slightly overestimated test performance**, as alpha was chosen based on test outcomes from models trained just on the training set.

Why This Step Was Important:

- Retraining produced a **realistic performance estimate** and verified the simplified model's generalisability.
- The experiment found that while α = 0.01 resulted in a simpler tree, it resulted in **meaningful performance trade-offs**.

Takeaway:

- This step enabled us to validate model stability, compare real generalisation, and ensure that our final model selection will be both robust and evidence-based.

6 MODEL PERFORMANCE COMPARISON & FINAL MODEL SELECTION

6.1 Model Comparison

We will do the comparison among all models of Logistic regression & Decision tree and between them based on train & test performance.

Training Performance Comparison

Training performance comparison:

	Logistic Regression Base	Logistic Regression Tuned(Threshold=0.302)	Logistic Regression Tuned(Threshold=0.358)	Decision Tree Base	Decision Tree Pre-Pruned	Decision Tree Post-Pruned(alpha=0.000147)	Decision Tree Post-Pruned(alpha=0.01)
Accuracy	0.801463	0.770666	0.788021	0.994149	0.854517	0.879897	0.795256
Recall	0.624305	0.789442	0.737497	0.984417	0.847669	0.878715	0.732182
Precision	0.729944	0.616917	0.656875	0.997674	0.743641	0.781478	0.671728
F1	0.673004	0.692597	0.694855	0.991001	0.792255	0.827249	0.700653

Table 45. Training Performance Comparison.

Test Performance Comparison

Test set performance comparison:

	Logistic Regression Base	Logistic Regression Tuned(Threshold=0.302)	Logistic Regression Tuned(Threshold=0.358)	Decision Tree Base	Decision Tree Pre-Pruned	Decision Tree Post-Pruned(alpha=0.000147)	Decision Tree Post-Pruned(alpha=0.01)
Accuracy	0.811457	0.775851	0.796790	0.868278	0.850383	0.865418	0.803985
Recall	0.635420	0.800449	0.750491	0.791187	0.844232	0.851249	0.740107
Precision	0.752409	0.623933	0.670512	0.804739	0.738160	0.765522	0.687435
F1	0.688984	0.701254	0.708251	0.797905	0.787641	0.806113	0.712799

Table 46. Test Performance Comparison.

6.2 Final Model Selection

Models Rejection Justification

Logistic Regression Models:

Base Model

- The model showed lower recall (0.6354) and F1-score (0.6890) compared to the decision tree models. It had difficulty capturing nonlinear patterns in cancellation behaviour, which was clear from its weaker performance.

Tuned Models (Threshold=0.302/0.358)

- These models improved recall but decreased precision (0.6239, 0.6705), resulting in suboptimal F1-scores (0.7013, 0.7083). The trade-offs between thresholds did not strike the right balance for business impact.

Decision Tree Base Model:

- There was severe overfitting. The model had nearly perfect training accuracy (0.9941) but a sharp decline in test recall (0.7912) and F1-score (0.7979). It was prone to memorizing noise, which made it unreliable for real-world predictions.

Decision Tree Pre-Pruned:

- This model performed moderately but was worse than the post-pruned version. It had a lower test F1-score (0.7876 vs. 0.8061) and precision (0.7382 vs. 0.7655). Manual tuning of hyperparameters, like max_depth, reduced complexity but also diminished predictive power.

Decision Tree Post-Pruned (α= 0.000147, 0.01):

- The model with α= 0.000147 performed well on both train & test with high F1 score (0.8061) but had complex tree model. Aggressive pruning (α=0.01) oversimplified the tree, leading to a drop in recall (0.7401) and F1-score (0.7128). This highlighted the risks of imposing excessive penalties on tree complexity.

Final Model Selection Over Other Models

After looking at the performance metrics of all models, the Decision Tree Post-Pruned (α=0.000147) emerged as the best choice for predicting hotel booking cancellations. This model achieved the highest F1 Score of 0.8061 on the test set, showing a solid balance between precision and recall.

Reason for This Model Selection are:

F1 Score

- The F1 Score is the main evaluation measure for this project because it reflects the model's ability to reduce both false positives and false negatives.
- The selected model's F1 Score shows that it can accurately predict cancellations while keeping precision at a reasonable level.
- Even though the Decision Tree Post-Pruned model shows slight overfitting, as indicated by a train F1 score of 0.8272 compared to a test F1 score of 0.8061, this trade-off is acceptable, given the model's overall performance and fit with business goals as it will help reduce cancellation-related losses.

Recall Performance

- The model also recorded a high recall of 0.8512.
- This is important because it shows how well the model identifies actual cancellations.
- This effectiveness is vital for minimizing the risk of lost revenue due to unexpected cancellations.

Balanced Performance

- While the Decision Tree Base model achieved impressive accuracy of 0.9941, it likely overfits the training data.
- This is clear from its much lower recall and F1 Score on the test set.
- The selected post-pruned model finds a better balance between complexity and performance, ensuring it works well on new data.

Complexity Consideration

- Even though the model is more complex than simpler models like Logistic Regression and post pruned tree (α= 0.01) but had high performance overall.
- The priority should be on improving performance metrics, which the chosen model effectively accomplishes.

Conclusion:

- **The Decision Tree Post-Pruned (α=0.000147) is the final model chosen for its outstanding performance in predicting hotel booking cancellations.**
- **It fits well with the project's goals and the need for actionable insights to reduce cancellation-related losses.**

7 ACTIONABLE INSIGHTS

Lead Time is the Primary Cancellation Driver

- **Key Finding**: Lead time has the strongest correlation (**r = 0.44**) with cancellations.
- **Impact**: Bookings made 85 days or more in advance have much higher cancellation rates.
- **Actionable Insight**: The median lead time of 57 days creates a critical window where **33%** of bookings are lost.

Customer Engagement Reduces Cancellation Risk

- **Special Requests Effect**: Customers who make special requests are 67% less likely to cancel.
- **Zero Requests Risk**: **54.5%** of customers make no special requests, which leads to a 43% cancellation rate.
- **Progressive Protection**: Cancellation rates drop from **43%** (0 requests) to 0% (3 or more requests).

Market Segment Performance Varies Dramatically

- **Online Channel Vulnerability**: This channel has a **36.5%** cancellation rate while generating **69.5%** of revenue.
- **Corporate Reliability**: The cancellation rate is only **10.9%,** showing commitment.
- **Offline Stability**: The cancellation rate is **30%**, with a **25.7%** revenue contribution.

Poor Customer Loyalty

- **Critical Gap**: There are only **2.6%** repeat customers, while **97.4%** of visitors are first-timers.
- **Loyalty Impact**: Repeat guests have a **1.7%** cancellation rate, compared to **33.6%** for new guests.
- **Revenue Loss**: The business is missing out on its most reliable customer segment.

Pricing Strategy Challenges

- **Price Sensitivity**: Higher-priced rooms have a **91%** increased chance of cancellations.
- **Market Dynamics**: In 2018, there were 4.5 times more bookings but **11.4 times** more cancellations.
- **Peak Season Risk**: **October, the busiest month**, needs targeted retention strategies.

Operational Indicators of Commitment

- **Parking Requests**: The cancellation rate is **10.1%** for customers who request parking, compared to 33.5% for those who do not.
- **Family Bookings**: **Larger family** sizes tend to have slightly higher cancellation rates.
- **Duration Effect**: Longer stays have a **10.7%** higher chance of cancellation.

8 BUSINESS RECOMMENDATIONS

1.mplement Dynamic Lead Time Policies

Immediate Actions:

- **Graduated Deposit Structure**: Require higher deposits for bookings made more than 60 days in advance.
- **Flexible Booking Windows**: Create booking categories for **30 days, 60 days, and 90** or more days with different terms.
- **Last-Minute Incentives**: Offer discounts for bookings made within **14 days** to lower inventory risk.

Implementation:

- Lead Time Tiers:
    - **0-30 days**: Standard policy, minimal deposit
    - **31-60 days**: 25% non-refundable deposit
    - **61-120 days**: 50% deposit, cancellation fee after 30 days
    - **120+ days**: 75% deposit, strict cancellation policy

2. Enhance Customer Engagement Strategy

Special Requests Program:

- **Proactive Outreach**: Contact customers who made no special requests within **48 hours**.
- **Suggestion Engine**: Offer **personalized room upgrades**, amenities, or services.
- **Engagement Scoring**: Prioritize retention efforts for customers who are highly engaged.

Loyalty Program Improvement:

- **Immediate Recognition**: Reward first-time guests who complete their stays.
- **Progressive Benefits**: Provide increasingly attractive perks for repeat visits.
- **Win-Back Campaigns**: Target the **97.4%** of one-time visitors to encourage return trips.

3. Channel-Specific Retention Strategies

Online Channel (36.5% cancellation rate):

- **Confirmation Campaigns**: Use multi-touch email sequences following bookings.
- **Exclusive Online Perks**: Compensate for higher cancellation chances by offering added value.
- **Competitive Rate Guarantees**: Reduce cancellations caused by price comparisons.

Corporate Program Expansion:

- **B2B Growth Initiative**: Use the low **10.9%** cancellation rate to attract more corporate clients.
- **Volume Incentives**: Encourage more business customers with reliable booking patterns.
- **Account Management**: Provide dedicated service for corporate relationships.

4. Revenue Protection Measures

Dynamic Pricing Adjustments:

- **Risk-Based Pricing**: Set higher prices for riskier booking profiles.
- **Overbooking Strategy**: Use a scientific approach based on cancellation data.
- **Revenue Recovery**: Implement upselling for confirmed bookings.

Cancellation Fee Structure:

- Recommended Fee Schedule:
    - **48+ hours**: 10% of booking value
    - **24-48 hours**: 50% of booking value
    - **< 24 hours**: 100% of booking value
    - **No-shows**: Full charge plus administrative fee

5. Seasonal and Operational Optimizations

Peak Season Management (October):

- **Capacity Planning**: Prepare for **14.7%** of annual volume.
- **Premium Positioning**: Use high demand to negotiate better terms.
- **Waitlist Management**: Turn cancellations into new booking opportunities.

Off-Peak Strategies (January-March):

- **Flexible Policies**: Encourage bookings with customer-friendly terms during the **14%** low season.
- **Package Deals**: Bundle amenities to increase commitment.
- **Local Partnerships**: Create one-of-a-kind experiences to cut down on cancellations.

7. Process Improvements

Customer Feedback:

- Use post-cancellation surveys to find root causes.

Staff Training:

- **Risk Assessment**: **Train staff** to identify high-risk bookings.
- **Retention Techniques**: Provide scripts and strategies to enhance customer retention.
- **Upselling Opportunities**: Convert potential cancellations into revenue-generating chances.