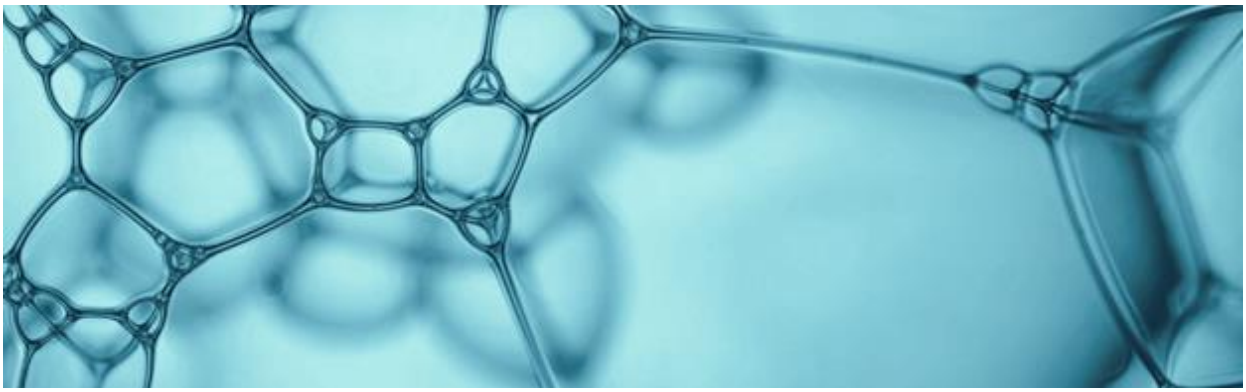

MACHINE LEARNING 1

Business Project Report (Coded)

08.09.2024

Nabankur Ray

PGP-DSBA

Contents

List of figures

List of Tables

Problem Statement - Coded Project

Business Context

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

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

The cancellation of bookings impact a hotel on various fronts:

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

Objective

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

Data Description

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

Data Dictionary:

Booking_ID: the unique identifier of each booking

no_of_adults: Number of adults

no_of_children: Number of Children

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

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

type_of_meal_plan: Type of meal plan booked by the customer:

- Not Selected – No meal plan selected
- Meal Plan 1 – Breakfast
- Meal Plan 2 – Half board (breakfast and one other meal)
- Meal Plan 3 – Full board (breakfast, lunch, and dinner)

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

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

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

arrival_year: Year of arrival date

arrival_month: Month of arrival date

arrival_date: Date of the month

market_segment_type: Market segment designation.

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

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

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

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

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

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

Executive Summary:

INN Hotels Group faces significant revenue loss due to a high rate of booking cancellations, particularly from online channels and bookings made well in advance. This issue affects the hotel chain's ability to optimize room occupancy and profit margins. Through data analysis of the hotel booking patterns, several key factors influencing cancellations have been identified, including lead time, guest type, room pricing, and the market segment. Based on these insights, we recommend strategic actions like dynamic cancellation fees, targeted customer incentives, and better management of online booking channels to mitigate losses. A predictive machine learning model will also be developed to help forecast cancellations, enabling proactive management of room occupancy and pricing.

Deliverables:

- Detailed analysis of key variables influencing booking cancellations.
- Visualizations and insights on univariate and bivariate relationships.
- A machine learning model capable of predicting which bookings are likely to be canceled.
- Evaluation of model performance with key metrics (e.g., accuracy, precision, recall)
- Actionable insights & Business Recommendations on reducing booking cancellations.

Understanding the Data

Dataset Sample

Displaying the first 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	lead_time	arrival_year	arriva
0	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	2017	
1	INN00002	2	0	2	3	Not Selected	0	Room_Type 1	5	2018	
2	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	2018	
3	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	2018	
4	INN00005	2	0	1	1	Not Selected	0	Room_Type 1	48	2018	

Table 1: First 5 rows of the given dataset

Displaying the last 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	lead_time	arrival_year	arr
36270	INN36271	3	0	2	6	Meal Plan 1	0	Room_Type 4	85	2018	
36271	INN36272	2	0	1	3	Meal Plan 1	0	Room_Type 1	228	2018	
36272	INN36273	2	0	2	6	Meal Plan 1	0	Room_Type 1	148	2018	
36273	INN36274	2	0	0	3	Not Selected	0	Room_Type 1	63	2018	
36274	INN36275	2	0	1	2	Meal Plan 1	0	Room_Type 1	207	2018	

Table 2: Last 5 rows of the given dataset

Structure and Types of Data

```
<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 3: Checking the structure and type of data

Observations:

- There are **36275 rows** and **19 Columns** are present in the given datasets.
- It can be observed that no columns have less entries (less than 36275 rows) which indicates that there are **no missing values** in the given dataset.

- There are attributes of different types (int, float, object) in the data.
- There are **14 numerical columns** in the data and **5 categorical columns**.
- **Dependent variable** is the **booking_status** which is of categorical type.

Statistical summary of the Data

Numerical Data

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.00000	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000	4.00000
no_of_children	36275.00000	0.10528	0.40265	0.00000	0.00000	0.00000	0.00000	10.00000
no_of_weekend_nights	36275.00000	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000	7.00000
no_of_week_nights	36275.00000	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000	17.00000
required_car_parking_space	36275.00000	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000	1.00000
lead_time	36275.00000	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000	443.00000
arrival_year	36275.00000	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000	12.00000
arrival_date	36275.00000	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000	31.00000
repeated_guest	36275.00000	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000	1.00000
no_of_previous_cancellations	36275.00000	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000	13.00000
no_of_previous_bookings_not_canceled	36275.00000	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000	58.00000
avg_price_per_room	36275.00000	103.42354	35.08942	0.00000	80.30000	99.45000	120.00000	540.00000
no_of_special_requests	36275.00000	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000	5.00000

Table 4: Statistical summary of the numerical data

Observations:

- Maximum no. of adults is 4 with 75% times the no. of adult is 2. There is chances that mostly couples are like to book the hotel room.
- Maximum no. of children is 10.
- Average price per room is 103.42 euro with maximum price of 540 euro.
- Maximum no. of special request made by the guest is 5.
- The average lead time for bookings is around 85 days, with a wide range (minimum 0 days and maximum 443 days). This large standard deviation (85.93) indicates high variability in how far in advance bookings are made.
- The average number of previous cancellations is very low (mean = 0.0234), with some guests having canceled up to 13 times.
- On average, bookings include around 0.81 weekend nights and 2.20 weeknights, suggesting that many stays span both weekdays and weekends.

Categorical Data

	count	unique	top	freq
Booking_ID	36275	36275	INN00001	1
type_of_meal_plan	36275	4	Meal Plan 1	27835
room_type_reserved	36275	7	Room_Type 1	28130
market_segment_type	36275	5	Online	23214
booking_status	36275	2	Not_Canceled	24390

Table 5: Statistical summary of the categorical data

```
Booking_ID: ['INN00001' 'INN00002' 'INN00003' ... 'INN36273' 'INN36274' 'INN36275']
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']
```

Table 6: Unique value of the categorical data

Observations:

- There are 4 unique values in type_of_meal_plan - 'Meal Plan 1' 'Not Selected' 'Meal Plan 2' 'Meal Plan 3'
- There are 4 unique values in room_type_reserved - Room_Type 1' 'Room_Type 4' 'Room_Type 2' 'Room_Type 6' 'Room_Type 5' 'Room_Type 7' 'Room_Type 3'
- There are 4 unique values in market_segment_type - 'Offline' 'Online' 'Corporate' 'Aviation' 'Complementary'
- There are 4 unique values in booking_status - 'Not_Canceled' 'Canceled'
- From the above observation, it can be infer that there is no discrepancies in categorical column.
- Booking_ID column has 36275 unique values which is equal to count. And there is no significant impact on the analysis, so the column can be dropped.

Checking for missing values

	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: No. of Missing Values in each column

INFERENCE:

- There is no missing values.
- There is no duplicate values.
- There is no null values.

Exploratory Data Analysis

Univariate Analysis

Observation on avg_price_per_room

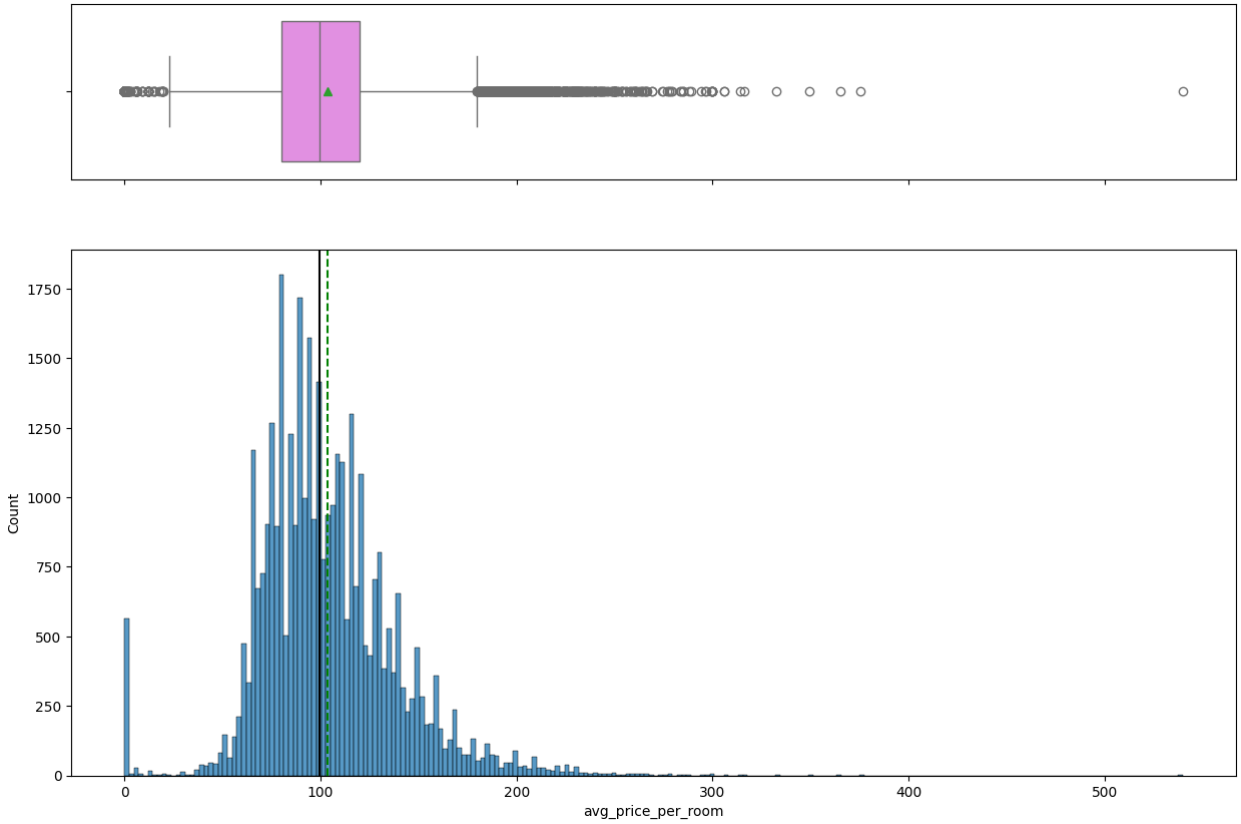

Fig 1: Histogram - boxplot of avg_price_per_room

Observations

- The histogram shows that the avg_price_per_room is heavily right-skewed.
- From the boxplot, it can be seen that there are numerous outliers on the higher end of the distribution, indicating that while most of the data is concentrated within a certain range. These outliers might represent luxury or premium offerings and should not be removed.

Observation on lead_time

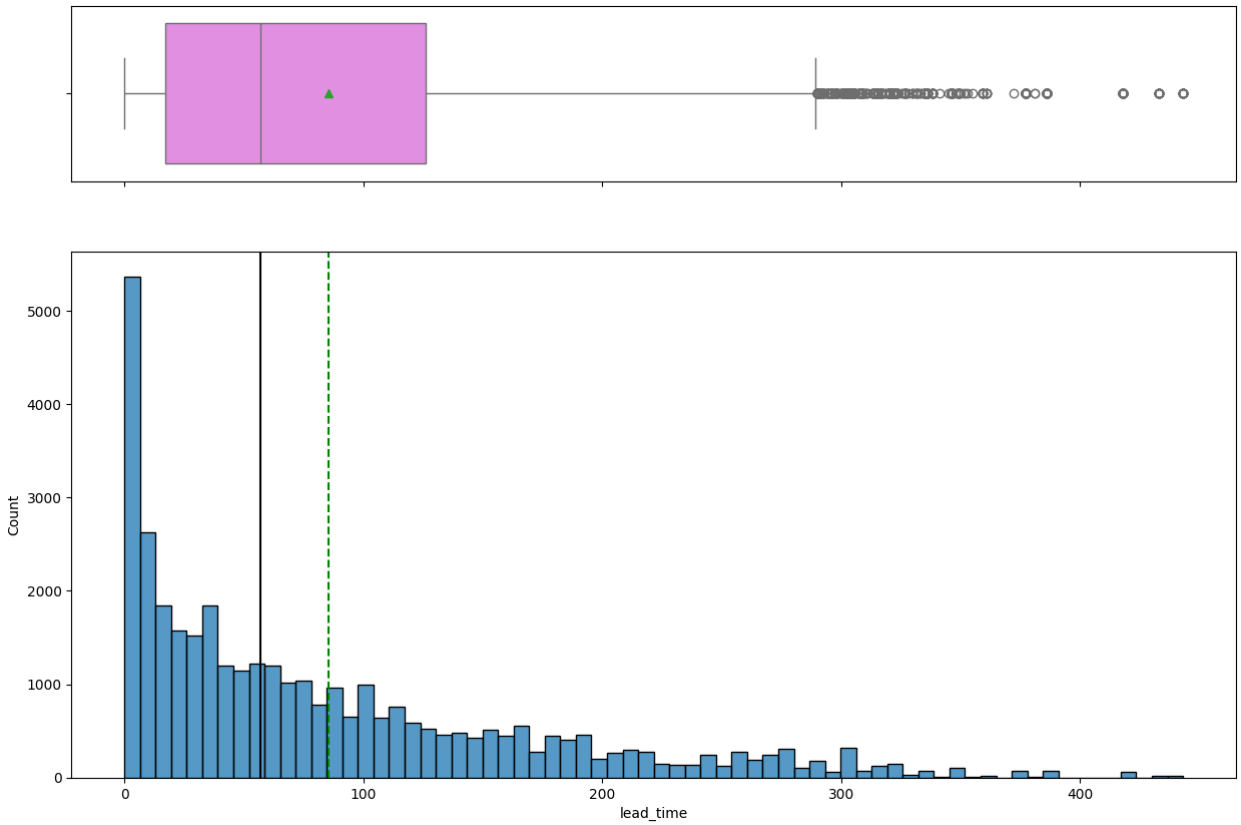

Fig 2: Histogram - boxplot of lead_time

Observations

- The histogram shows that the lead_time is also heavily right-skewed.
- From the boxplot, it can be seen that there are numerous outliers on the higher end of the distribution. These outliers represent that the some customers might booked long back from the arrival date and hence, cannot not be removed.

Observation on no_of_adults

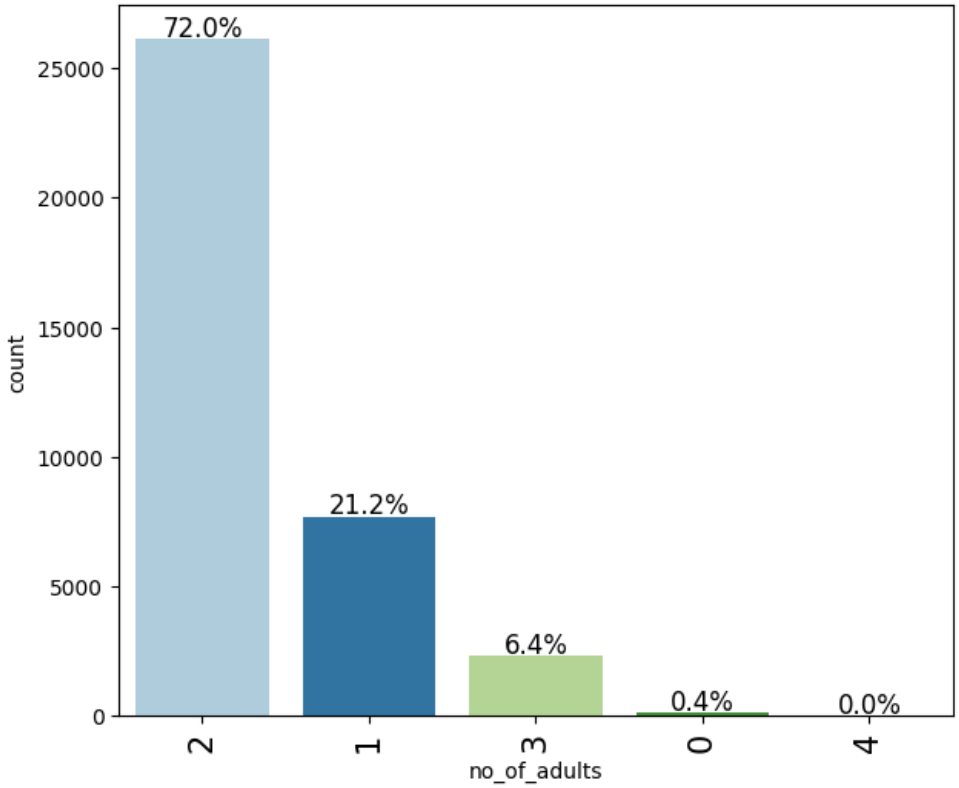

Fig 3: Barplot of no_of_adults

Observations

- The most common group size is two adults, accounting for 72.0% of the total observations. This indicates that the majority of bookings are likely made by couples or two adults traveling together.
- The second most frequent category is bookings made by a single adult, representing 21.2% of the total. This suggests a significant portion of solo travelers or individuals booking rooms.
- Only 6.4% of the bookings are made by groups of three adults.
- Bookings with four adults is neglible(0.0%)
- Bookings with no adults (only child booking) is extremely rare, virtually nonexistent. It might be good to drop those values.

Observation on no_of_children

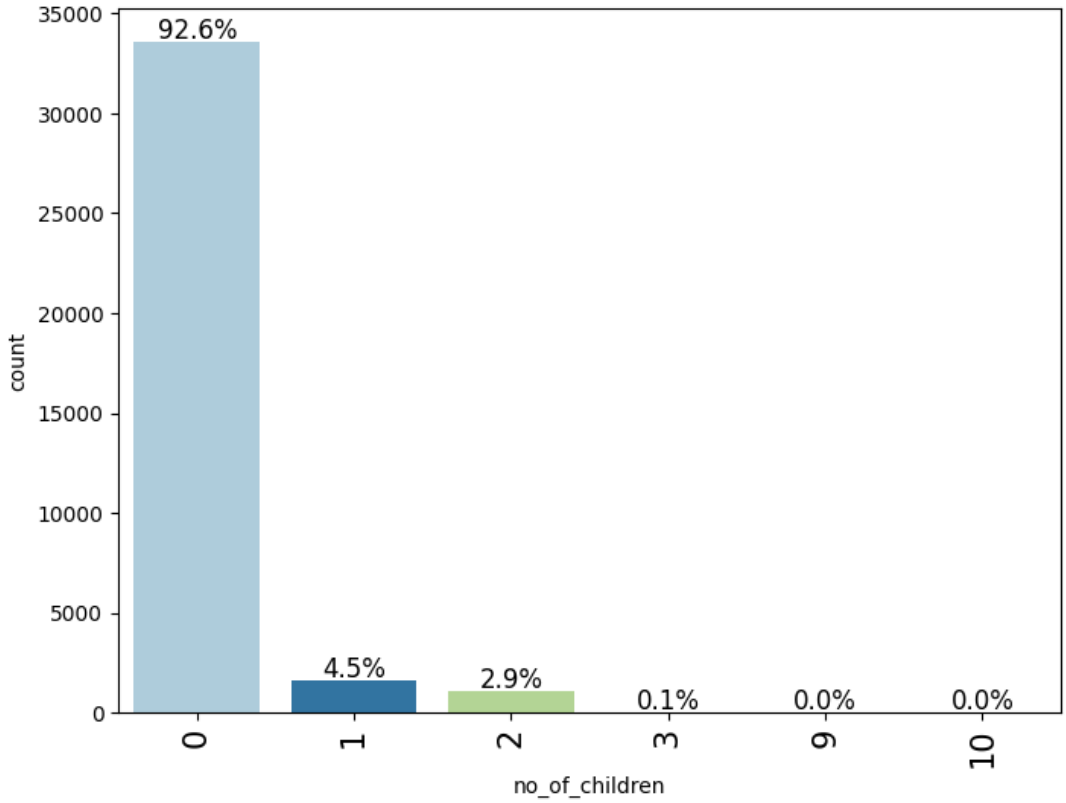

Fig 4: Barplot of no_of_children

Observations

- Approximately 92.6% of the guests, does not have any children. This suggests that most focusing demographic is younger individuals, couples without children.
- The proportion of guests with one child is 4.5%, and those with two children is 2.9%. Guests with three or more children make up a negligible percentage (0.1% or less).

Observation on no_of_weekend_nights

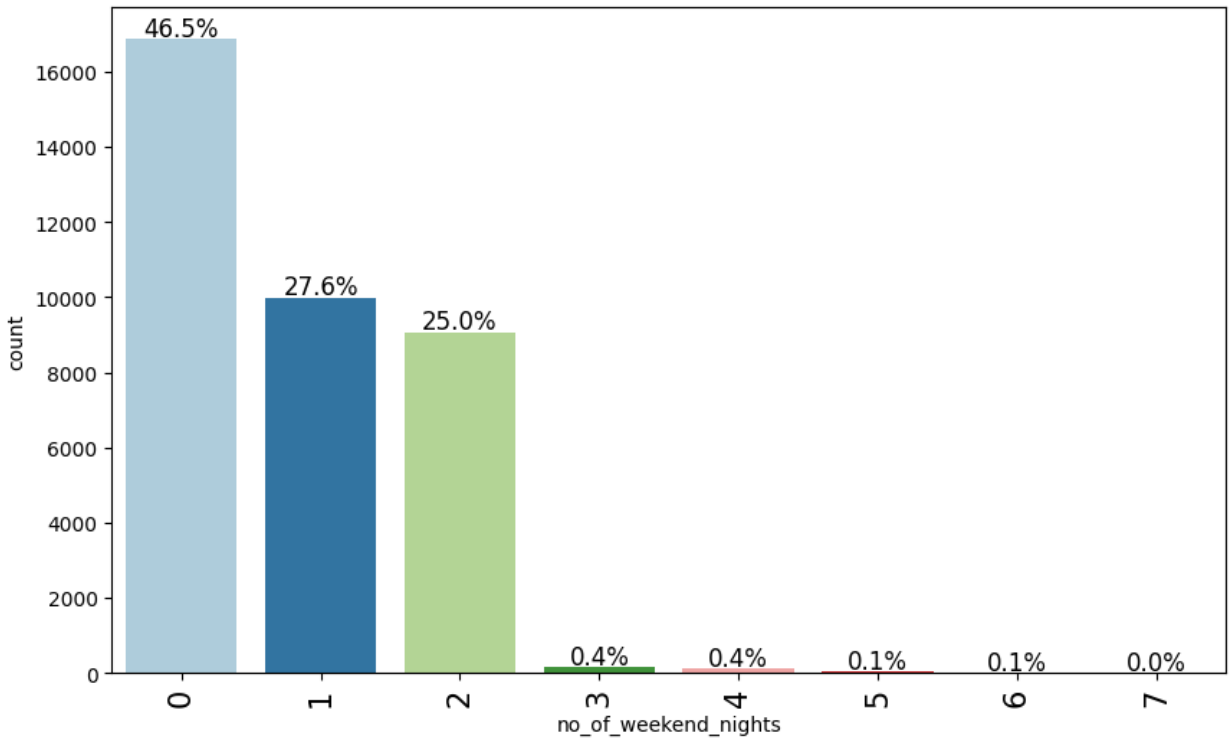

Fig 5: Barplot of no_of_weekend_nights

Observations

- Nearly half of the dataset (46.5%) represents guests that did not spend any weekend nights. This suggests a significant portion of the customers is not utilizing services or facilities during weekends.
- The majority of weekend stays are distributed between 1 or 2 nights, with 27.6% of the population staying for one night and 25.0% staying for two nights. This suggests that short weekend getaways are common among the customers.

Observation on no_of_week_nights

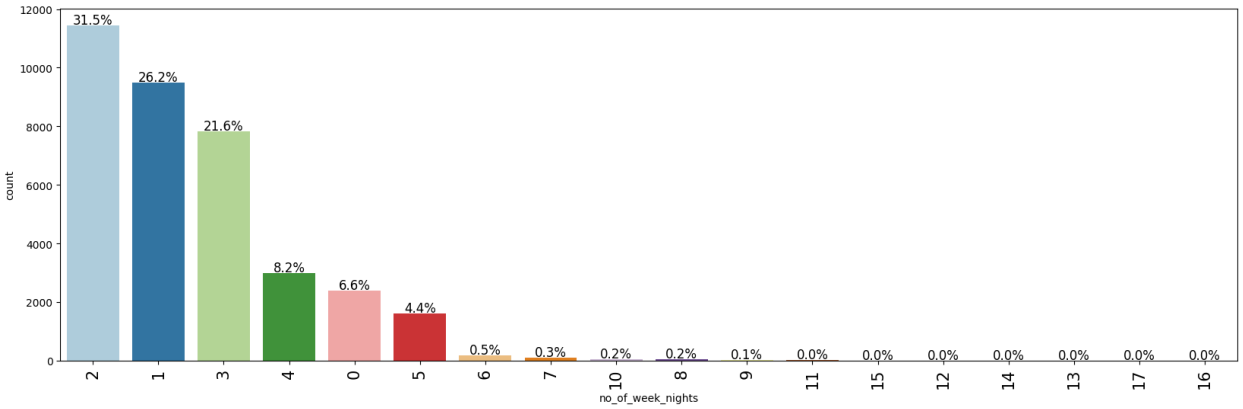

Fig 6: Barplot of no_of_week_nights

Observations

- The highest proportion, 31.5%, of the dataset corresponds to stays of 2 week nights. This suggests that a significant number of customers prefer to spend two nights during the week.
- A considerable percentage of stays are for 1 night (26.2%) and 3 nights (21.6%), indicating that these are also popular choices among the population. The combination of 1, 2, and 3-night stays accounts for nearly 80% of the total, suggesting a strong preference for short to medium-length stays during the week.
- Stays of 4 or 5 nights are less frequent, at 8.2% and 4.4%, respectively. This suggests that extended stays during the week are less common.
- 6.6% of the data corresponds to zero week nights.
- Very few instances (less than 1% each) are observed for stays longer than 5 nights. This suggests that long-term accommodations are rare and likely do not represent the primary market.

Observation on required_car_parking_space

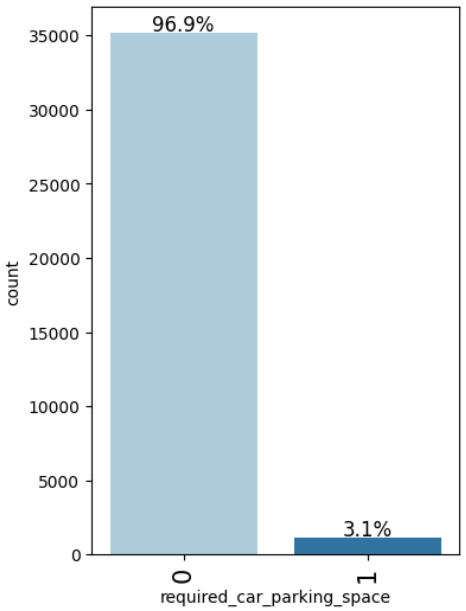

Fig 7: Barplot of required_car_parking_space

Observations

- 96.9% of the guests do not require parking space and only 3.1% of the guests require parking space, which implies that the majority of the customers don't travel with their own car. And another possibility is Guests are coming from far distances.

Observation on arrival_year

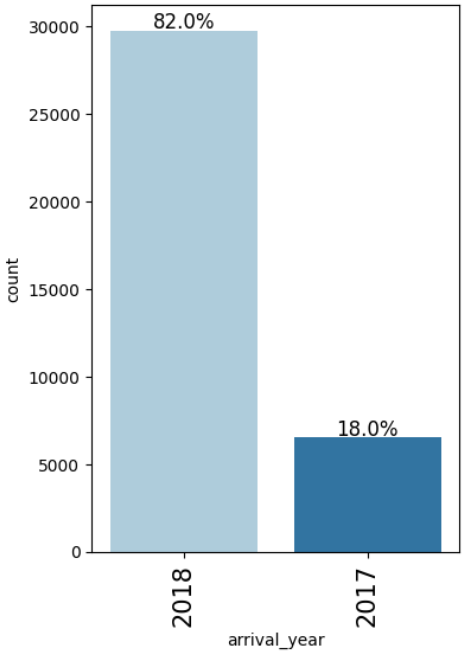

Fig 8: Barplot of arrival_year

Observations

- About 18% of the guest booked in the year 2017 and 82% of the guest booked in the year 2018 which is a huge jump. This shows that the business is doing good.

Observation on arrival_month

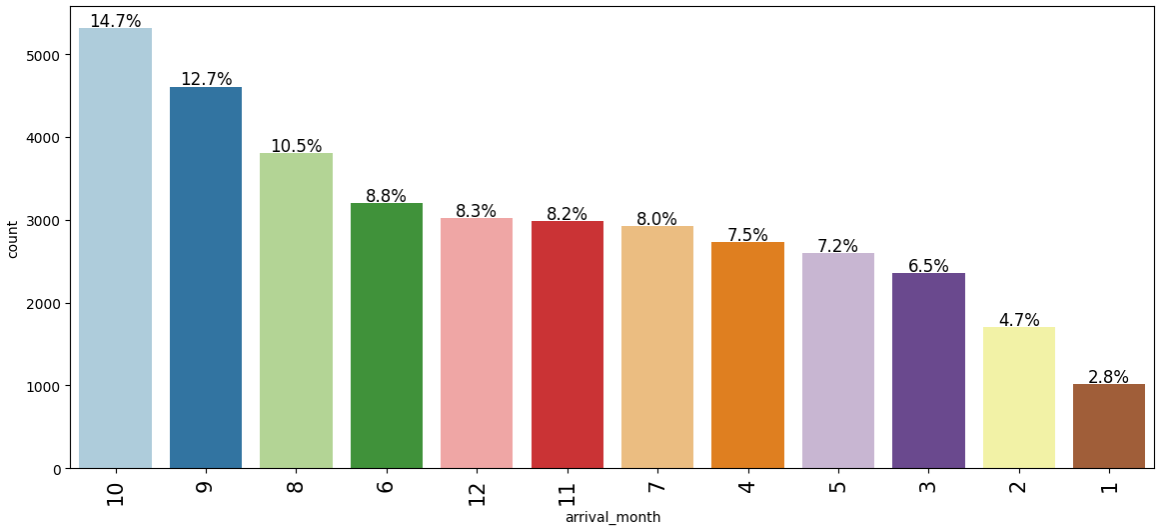

Fig 9: Barplot of arrival_month

Q: What are the busiest months in the hotel?

Ans: October is the busiest Month accounting for 14.7% of the total bookings.

Observations

- The highest count of arrivals occurs in October, accounting for 14.7% of the total. This suggests that October is a **peak & busiest month**.
- September (12.7%) and August (10.5%) also show high arrival counts, indicating a strong third quarter.
- June, December, and November have moderate counts (8.8%, 8.3%, and 8.2% respectively), suggesting relatively steady activity during the summer and early winter months.
- The lowest counts are observed in January (2.8%) and February (4.7%), suggesting a significant dip in activity during the early part of the year.
- There is a gradual increase in arrivals starting from March (6.5%), peaking around mid-year. This gradual upward trend may indicate improving conditions or preparations for busier periods later in the year.
- April and May show similar arrival rates (7.5% and 7.2% respectively), suggesting stable activity during this transition period between spring and summer.

Observation on arrival_date

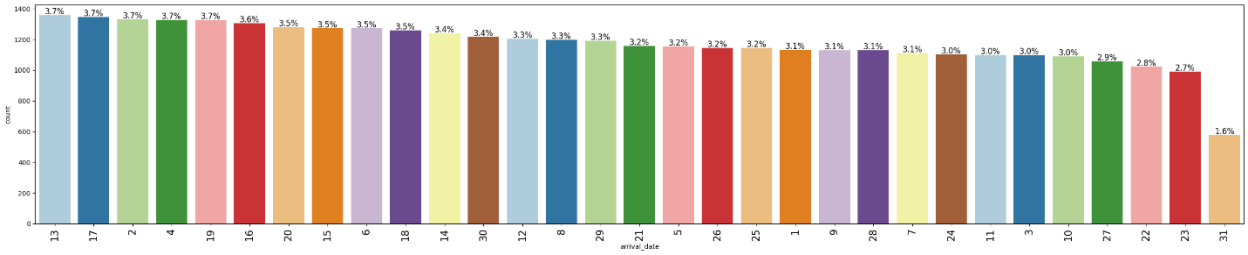

Fig 10: Barplot of arrival_date

Observations

- The arrival counts are relatively evenly distributed across most dates, with no extreme peaks or troughs.
- Dates like the 13th, 17th, 2nd, 4th, and 19th each have the highest count of arrivals at around 3.7%, indicating that there isn't a significant preference for specific dates within a month.
- The 13th, 17th, 2nd, 4th, and 19th are the most common arrival dates, but their frequencies are only marginally higher than the other dates, suggesting these days may have minor factors influencing increased activity.
- The 31st has the lowest count, with only 1.6% of arrivals, likely because not every month has a 31st day. This irregularity can cause a lower arrival frequency for this date.
- The last few days of the month (27th to 31st) and the first few days of the month (1st to 5th) tend to have slightly lower arrival rates, ranging from 2.7% to 3.2%, which could be associated with end-of-month or start-of-month cycles, like payroll schedules or billing periods.
- The overall variation between the highest and lowest counts is minimal, indicating a fairly balanced arrival pattern throughout the month with no significant clustering on specific dates.

Observation on repeated_guest

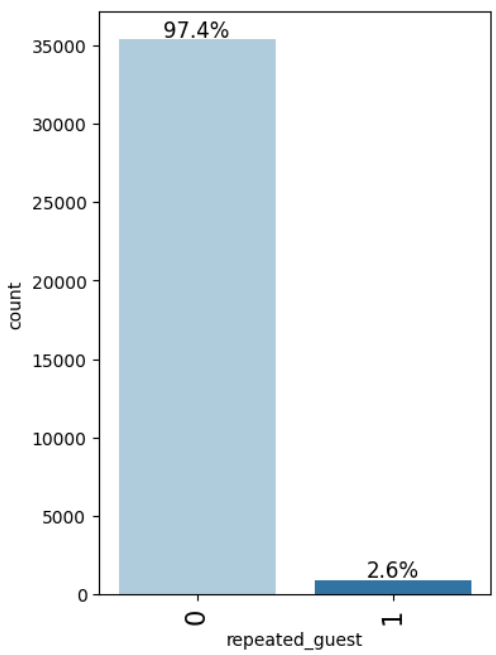

Fig 11: Barplot of repeated_guest

Observations

- Approximately, 97.4% of the guests are not repeated.
- And only 2.6% of the guests are repeated, which is significantly low in terms of business context. Strategies and planning are required to tackle this issue.

Observation on no_of_previous_cancellations

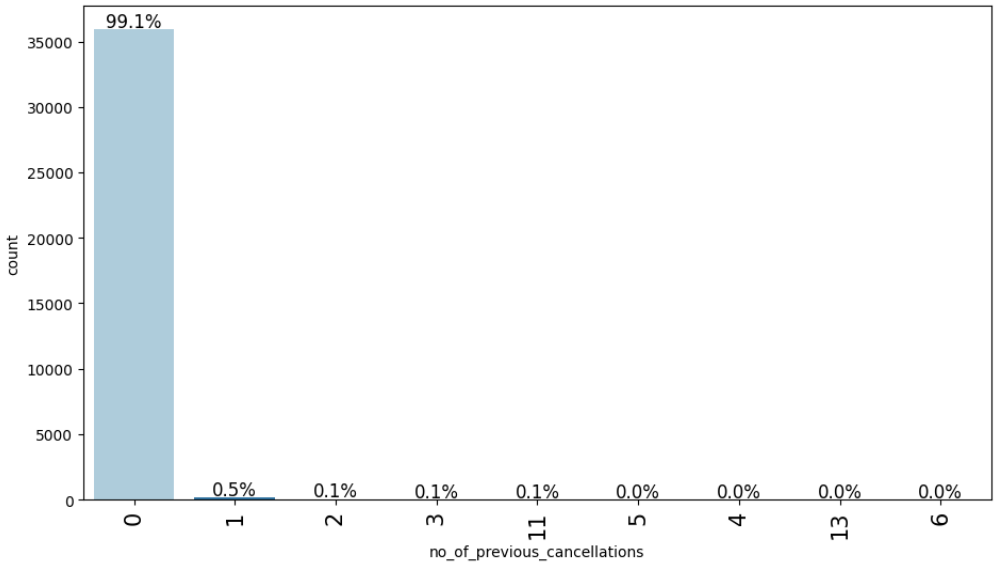

Fig 12: Barplot of no_of_previous_cancellations

Observations

- A majority of customers (99.1%) have no history of previous cancellations, indicating strong booking commitment among most customers. This suggests that the majority of bookings proceed as planned without cancellations, reflecting positively on the customer base's reliability.
- Customers with multiple cancellations (1 or more) constitute an extremely small percentage of the overall bookings.

Observation on no_of_previous_bookings_not_canceled

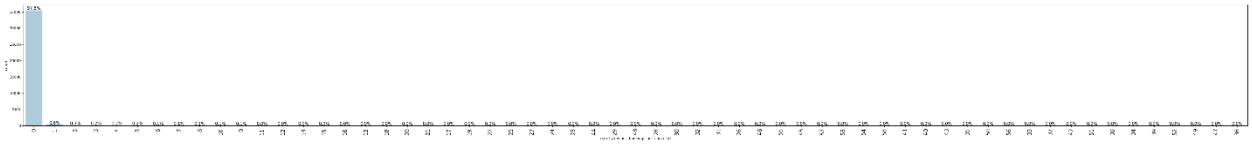

Fig 13: Barplot of no_of_previous_booking_not_canceled

Observations

- The overwhelming majority (97.8%) of customers have no previous bookings that were not canceled. This suggests that most bookings are from new or first-time customers.
- A very small percentage (2.2%) of customers have previous bookings that were not canceled, with the proportions decreasing as the number of such bookings increases. Only 0.6% of customers have one previous non-canceled booking, and the percentages continue to diminish further beyond this point.
- The data indicates extremely low levels of engagement in terms of repeat non-cancelled bookings, with the percentages dropping to 0.0% for customers having more than five previous non-cancelled bookings.

Observation on no_of_special_requests

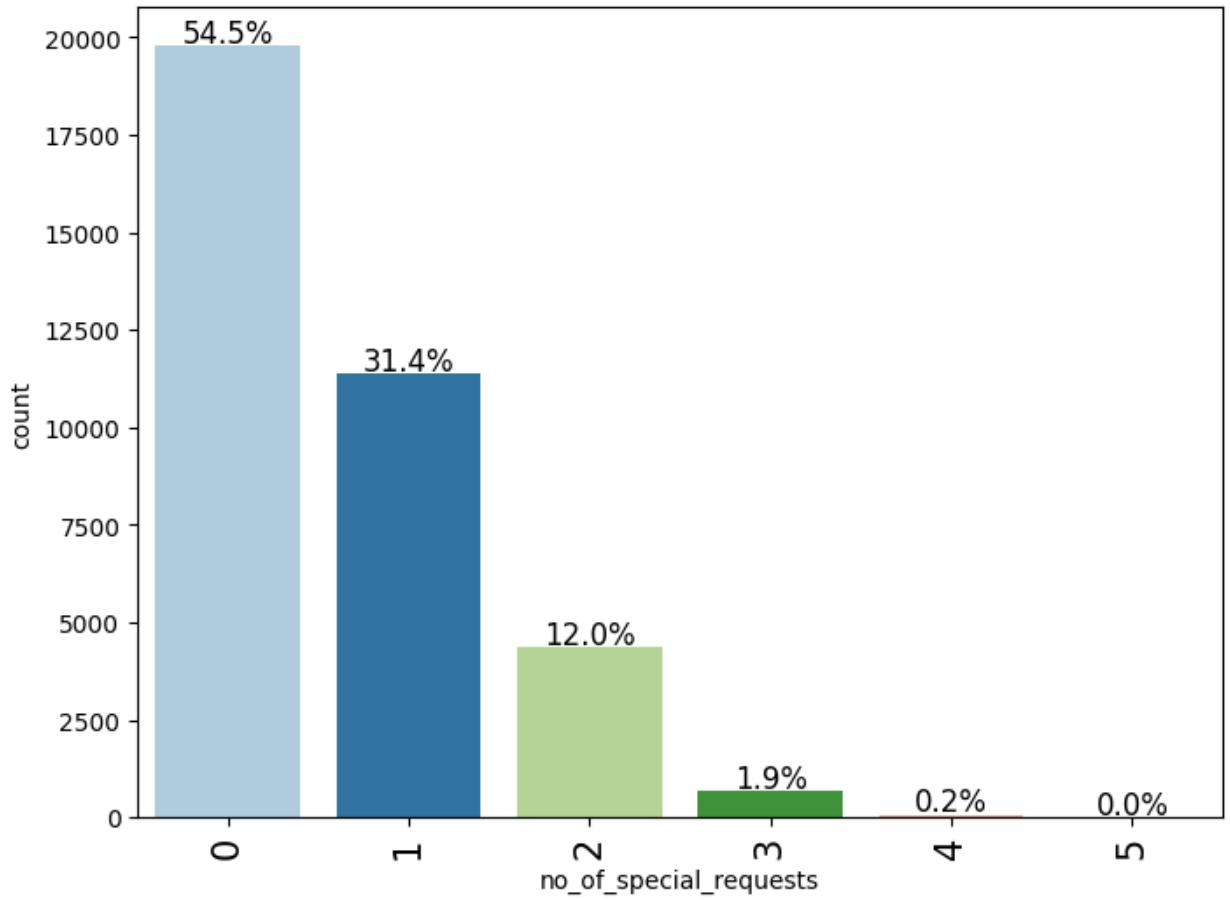

Fig 14: Barplot of no_of_special requests

Observations

- The majority of customers (54.5%) did not make any special requests during their booking, highlighting that while most customers do not require special request.
- A significant portion (31.4%) of customers made exactly one special request.
- The number of customers decreases sharply as the number of special requests increases. Only 12.0% of customers made two requests, 1.9% made three, and negligible proportions made four (0.2%) or five (0.0%) special requests.

Observation on type_of_meal_plan

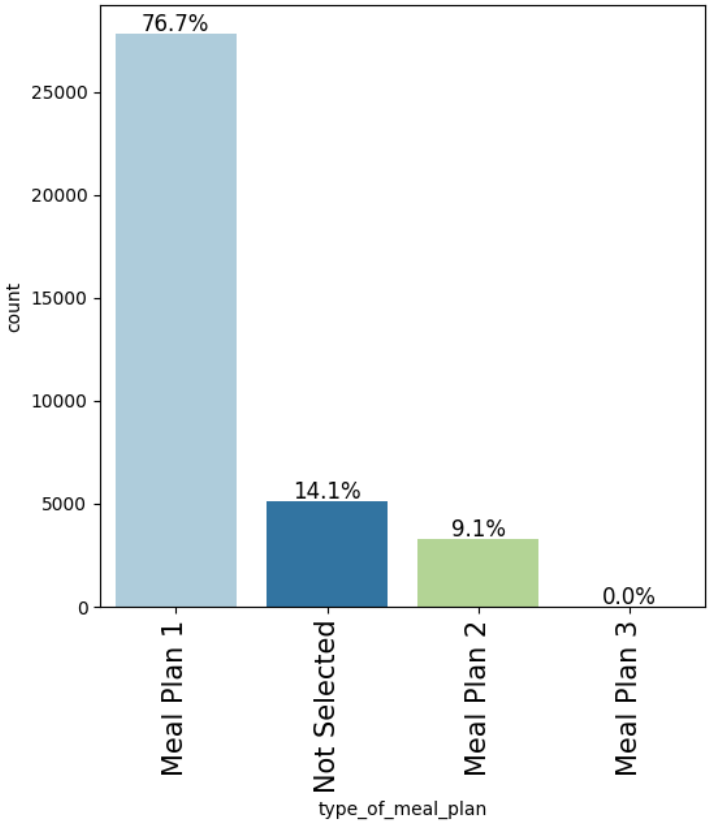

Fig 15: Barplot of type_of_meal_plan

Observations

- 76.7% of the guest preferred Meal Plan 1, making it by far the most preferrable meal type.
- The second most preferred Meal is Meal Plan 2 with 9.1%.
- 14.1% of the guests not selected any type of meal.
- Meal Plan 3 is effectively non-existent with 0.0% guests.

Observation on room_type_reserved

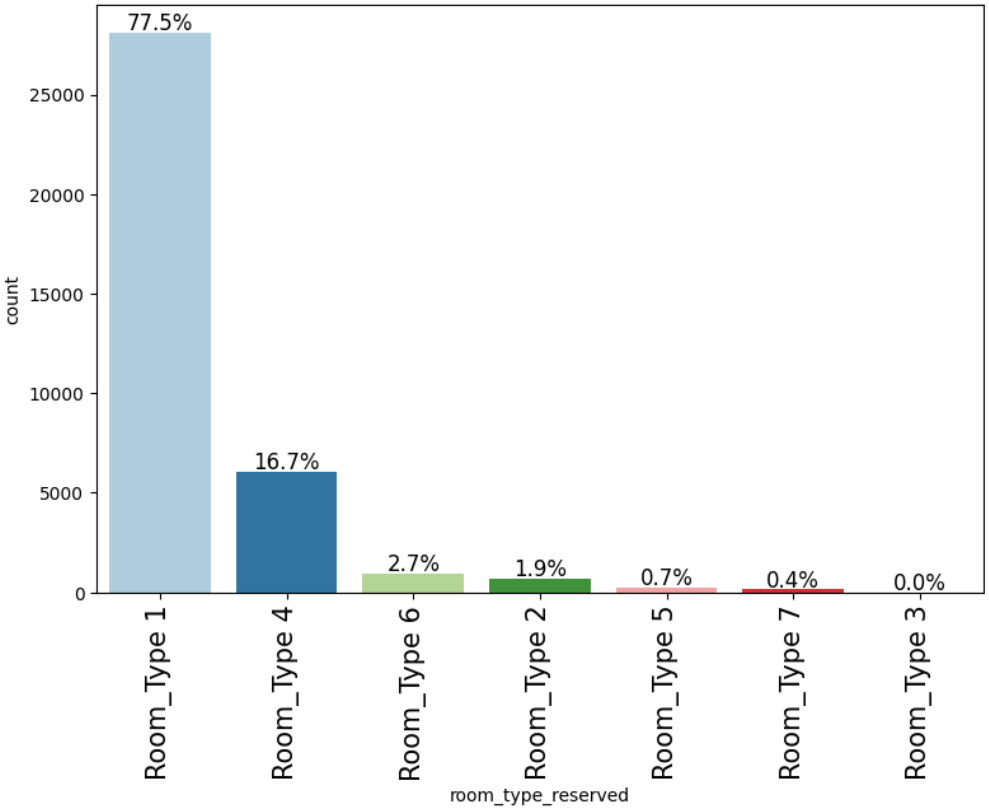

Fig 16: Barplot of room-type_reserved

Observations

- 77.5% of the bookings are for Room Type 1, making it by far the most popular room type.
- 16.7% of bookings are for Room Type 4, which is the second most popular option.
- Room Type 6 and Room Type 2 follow with 2.7% and 1.9% of bookings, respectively. These room types are chosen by a much smaller segment of customers.
- Room Type 5 and Room Type 7 have even lower percentages, with 0.7% and 0.4% of the total bookings, respectively.
- Room Type 3 is effectively non-existent with 0.0% bookings.

Observation on market_segment_type

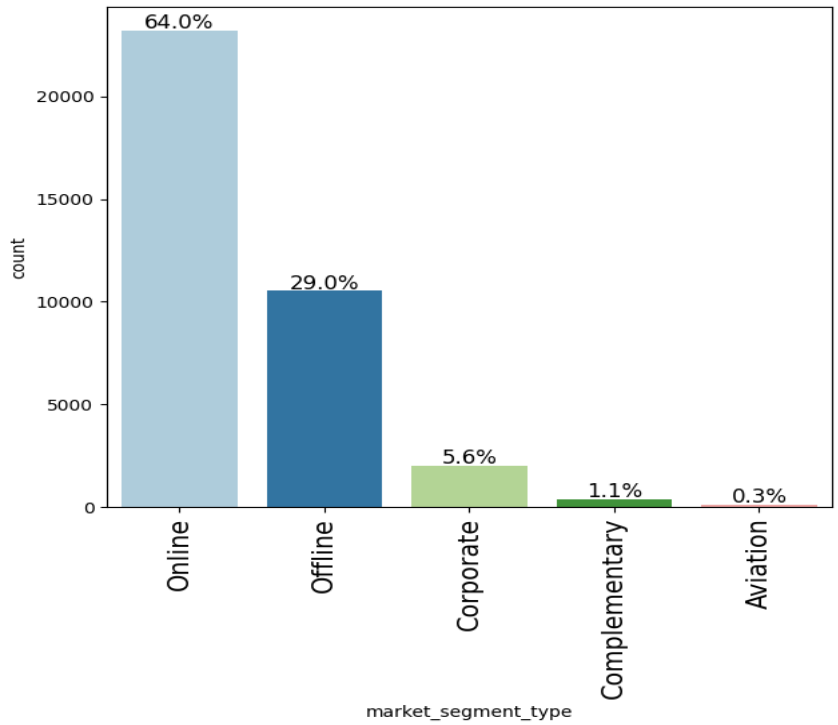

Fig 17: Barplot of market_segment_type

Q: Which market segment do most of the guests come from?

Ans: Most of the guest come from online market with 64%.

Observations

- **64.0% of the bookings come from the Online market segment**, making it the largest contributor by a significant margin. The Hotel should give more focus on online booking as it will help the business to increase booking.
- The Offline segment accounts for 29.0% of bookings, indicating that a substantial portion of customers still prefer or rely on traditional booking methods.
- Other market segment like Corporate, Complementary, Aviation contribute 5.6%, 1.1%, 0.3% which is very small in numbers compared to online & offline chanels.

Observation on booking_status

Q: What percentage of bookings are canceled?

Ans: 32.8% of the bookings were canceled.

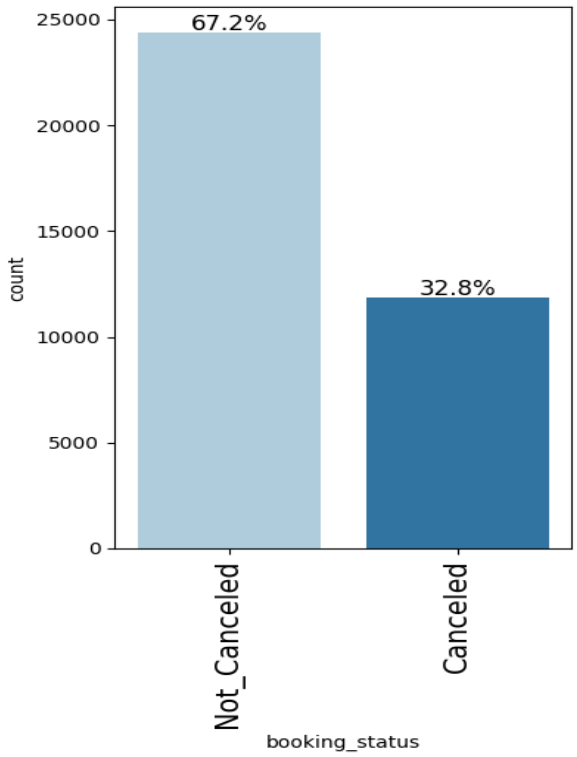

Fig 18: Barplot of booking_status

Observations

- Approximately, 67.2% of the bookings were not canceled, implying that the majority of bookings proceeded as planned.
- Conversely 32.8% of the bookings were canceled. This indicates that nearly one-third of all bookings result in cancellations.

Bivariate Analysis

Correlation check

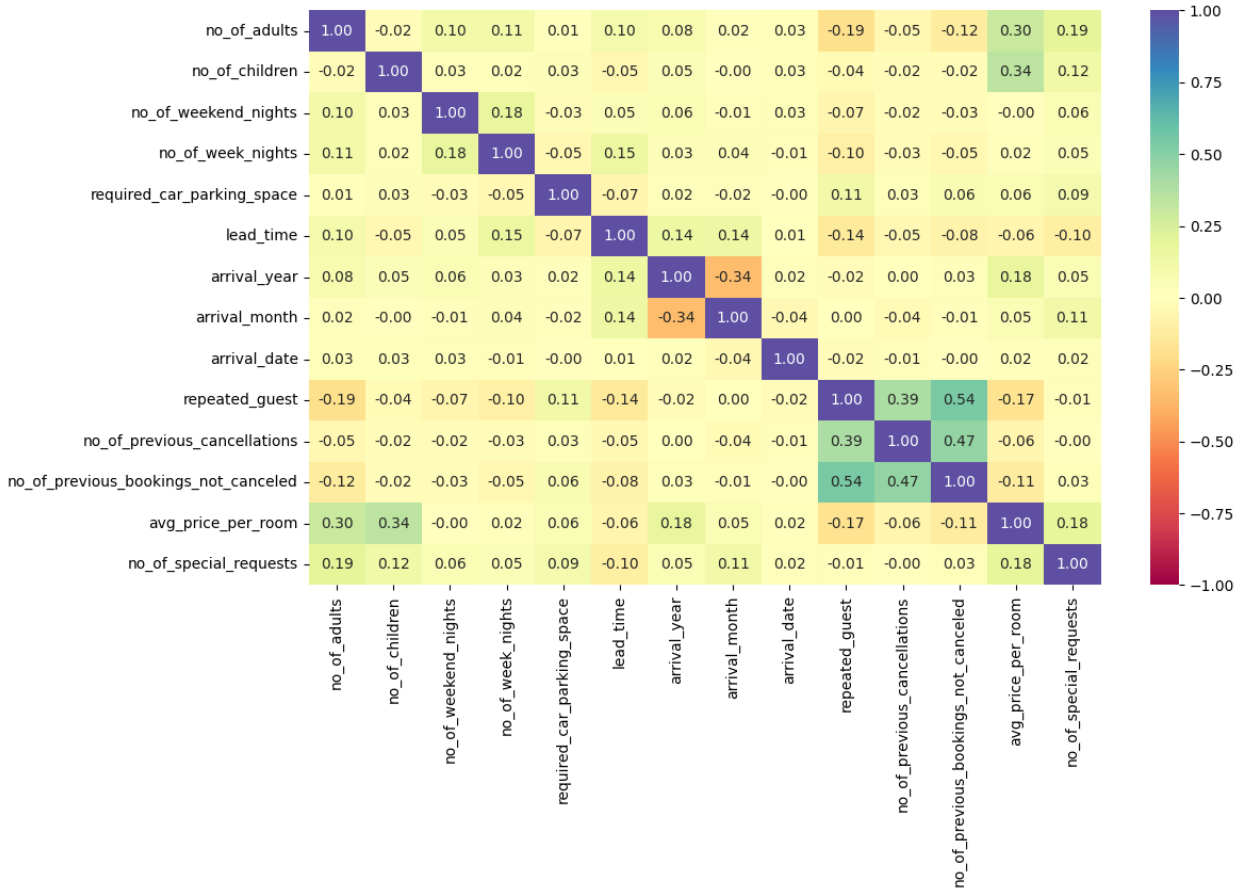

Fig 19: Heatmap of Numerical Variables

Observations

- no_of_previous_cancellations and no_of_previous_bookings_not_canceled (0.54): This indicates that customers with a history of cancellations also have a record of non-canceled bookings, suggesting variability in booking behavior among these individuals.
- avg_price_per_room with no_of_adults (0.30), no_of_children (0.34), and no_of_weekend_nights (0.20): Higher room prices tend to be associated with bookings involving more adults, children, and extended weekend stays, which may reflect larger group bookings or premium accommodation choices.
- no_of_special_requests and repeated_guest (0.11): Repeated guests are slightly more likely to make special requests, potentially due to familiarity with the service or specific needs.
- lead_time and arrival_year (0.14): A slight increase in lead time with more recent years might indicate changing booking patterns, such as planning further in advance.

- no_of_special_requests and no_of_previous_bookings_not_canceled (0.18): Guests with a history of non-canceled bookings are somewhat more likely to have special requests.
- arrival_year and arrival_month (-0.34): Indicates that arrivals in later years are associated with different seasonal patterns or distribution compared to earlier years.
- repeated_guest and no_of_adults (-0.19): Repeat guests tend to book for smaller groups, possibly indicating a preference for individual or couple travel rather than family or group bookings.
- lead_time and repeated_guest (-0.14): Repeat guests generally book with shorter lead times, likely due to increased familiarity and confidence in the booking process.
- Most other variable pairs show very low or near-zero correlations, suggesting little to no direct relationship. These include relationships between variables such as arrival_date, required_car_parking_space, and various other features, indicating independent booking behaviors that are not strongly influenced by these factors.

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

Below observations suggest that Hotel rates are closely tied to the market segment, with online customers paying the highest average prices and corporate clients benefiting from the most discounts.

Observation on market_segment_type vs avg_price_per_room

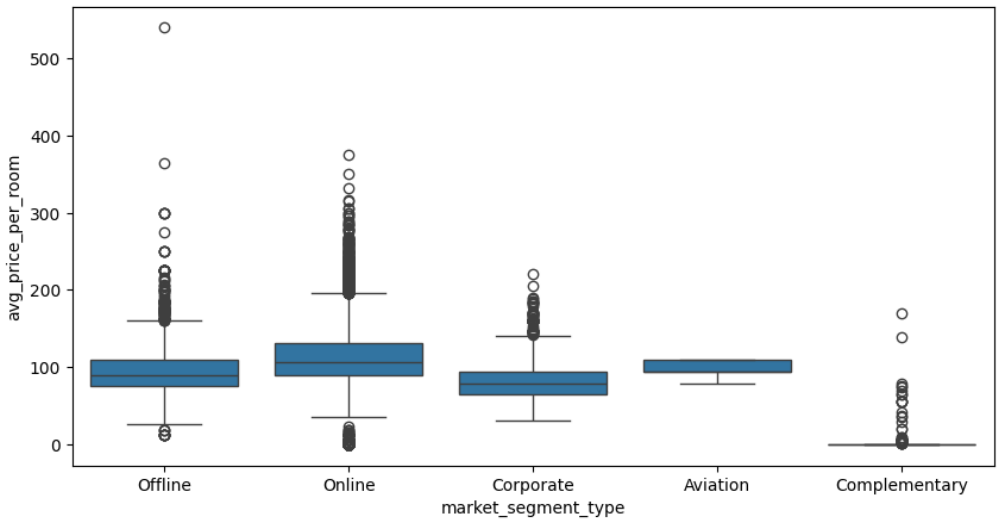

Fig 20: Boxplot of market_segment_type vs avg_price_per_room

market_segment_type	count	mean	std	min	25%	50%	75%	max
Aviation	125.00000	100.70400	8.53836	79.00000	95.00000	95.00000	110.00000	110.00000
Complementary	391.00000	3.14176	15.51297	0.00000	0.00000	0.00000	0.00000	170.00000
Corporate	2017.00000	82.91174	23.69000	31.00000	65.00000	79.00000	95.00000	220.00000
Offline	10528.00000	91.63268	24.99560	12.00000	75.00000	90.00000	109.00000	540.00000
Online	23214.00000	112.25685	35.22032	0.00000	89.00000	107.10000	131.75000	375.50000

Table 8: Statistical summary of market_segment_type vs avg_price_per_room

Observations

- **Online Market Segment** has the highest average price per room, with a mean of **112.26** euro, indicating that online customers might be more willing to pay premium prices.
- **Complementary Segment** shows the lowest average price at 3.14 euro. The high standard deviation suggests some instances where prices can go up to 170 euro, possibly for premium or executive-level complimentary offerings.
- **Offline Segment** follows with an average price of **91.63** euro, which a relatively high average price suggesting that customers who book offline prefer cheap rates compared to online segment.
- **Corporate Segment** has an average price of **82.91** euro, likely reflecting negotiated corporate rates, which are generally lower due to bulk or repeat bookings.

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

Only 1.72% of the repeating guests cancel their booking. This low cancellation rate underscores the importance of cultivating customer loyalty as part of the hotel's overall strategy.

repeated_guest	booking_status	proportion
0	Not_Canceled	66.41958
	Canceled	33.58042
1	Not_Canceled	98.27957
	Canceled	1.72043

Table 9: Proportion table of repeated_guest vs booking_status

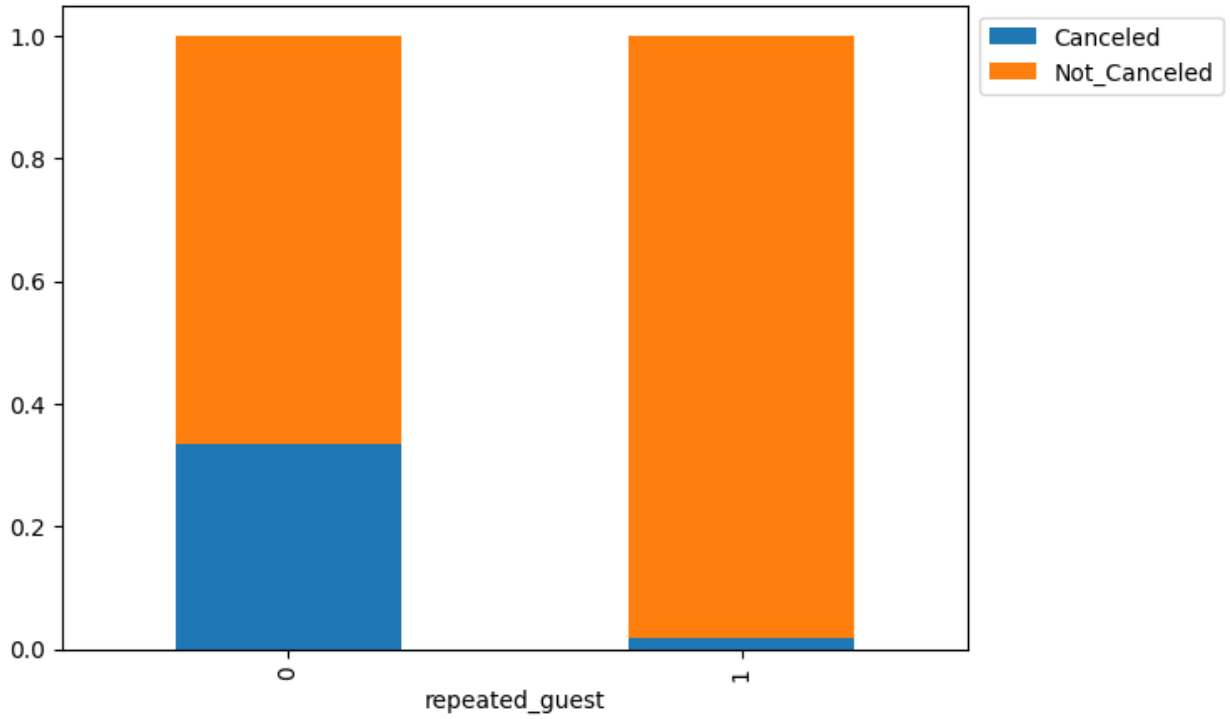

Fig 21: Barplot of repeating_guest vs booking_status

Observations

- The data suggests that repeated guests are much more likely to complete their bookings compared to first-time guests. The almost negligible cancellation rate among repeated guests (1.72%) indicates strong brand loyalty and satisfaction.

- First-time guests (not repeated) have a significant cancellation rate. Specifically, out of 35,345 bookings made by new guests, 11,869 were canceled.

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

Yes, according to the datasets no. of cancellations decreases with no. of special request made by the guests implying that customer service is key for reducing booking cancellations.

Observation on no_of_special_requests vs booking_status

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

Table 10: Count table of no_of_special_requests vs booking_status

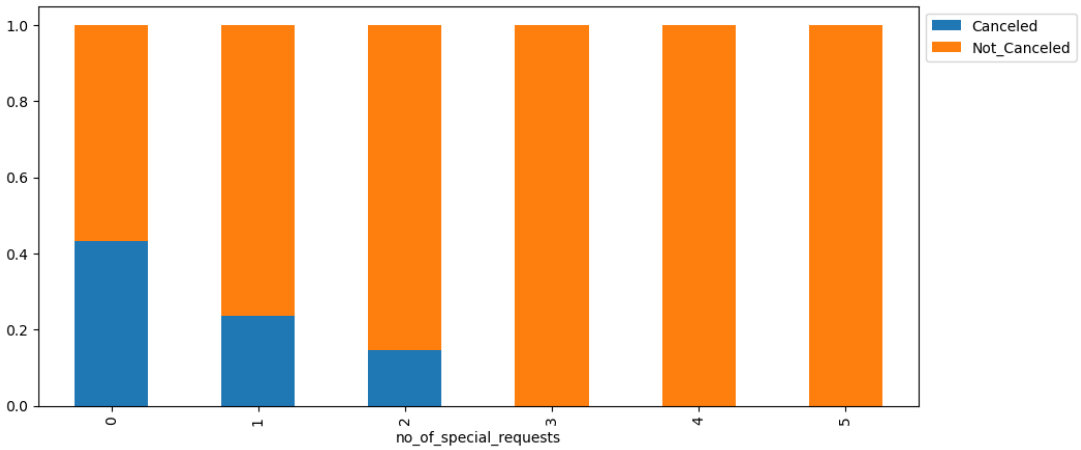

Fig 22: Barplot of no_of_special_request vs booking_status

Observations

- As the number of special requests increases from 0 to 5, the proportion of cancellations (blue bars) consistently decreases. This suggests that guests with more special requests are less likely to cancel their bookings.
- The highest cancellation rate is observed when there are no special requests (more than 50%). This indicates that guests without specific needs are more prone to cancel their bookings.

Lets analyse the booking status with other variable, i.e. the impact of other variable on cancellations.

repeated_guest vs booking_status

booking_status	repeated_guest	proportion
Canceled	0	99.86538
	1	0.13462
Not_Canceled	0	96.25256
	1	3.74744

Table 11: Proportion table of repeated_guest vs booking_status

Observations

- Out of total no. of booking canceled 99.86% of booking canceled by new customer.

arrival_month vs booking_status

Observations

- October has more no. Of cancellation with 1880 cancellation in one month.

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

Table 12: Count table of arrival_month vs booking_status

arrival_year vs booking_status

booking_status	Canceled	Not_Canceled	All
arrival_year			
All	11885	24390	36275
2018	10924	18837	29761
2017	961	5553	6514

Table 13: Count table of arrival_year vs booking_status

Observations

- 2018 has more no. of cancellations compare to previous year.

required_car_parking_space vs booking_status

Observations

- 33.48% of Customer who does not require car parking canceled the booking & 66.51% Not Canceled.

required_car_parking_space	booking_status	proportion
0	Not_Canceled	66.51304
	Canceled	33.48696
1	Not_Canceled	89.85765
	Canceled	10.14235

Table 14: Proportion table of required_car_parking_space vs booking_status

type_of_meal_plan vs booking_status

booking_status	Canceled	Not_Canceled	All
type_of_meal_plan			
All	11885	24390	36275
Meal Plan 1	8679	19156	27835
Not Selected	1699	3431	5130
Meal Plan 2	1506	1799	3305
Meal Plan 3	1	4	5

Table 15: Count table of type_of_meal_plan vs booking_status

Observations

- Customer who prefer Meal Plan 1 are more likely to cancel the booking.

room_type_reserved vs booking_status

Observations

- Customer with room type 1 preferrence has more no. of cancellation followed by room type 4.

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

Table 16: Count table of room_type_reserved vs booking_status

market_segment_type vs booking_status

booking_status	market_segment_type	proportion
Canceled	Online	71.30837
	Offline	26.52924
	Corporate	1.85107
	Aviation	0.31132
Not_Canceled	Online	60.43050
	Offline	30.23780
	Corporate	7.36777
	Complementary	1.60312
	Aviation	0.36080

Table 17: Proportion table of market_segment_type vs booking_status

Observations

- Online cancellation of booking is most with 71.3% cancellations follwed by offline with 26.52% cancellation.

no_of_previous_cancellations vs booking_status

booking_status	Canceled	Not_Canceled	All
no_of_previous_cancellations			
All	11885	24390	36275
0	11869	24068	35937
1	11	187	198
13	4	0	4
3	1	42	43
2	0	46	46
4	0	10	10
5	0	11	11
6	0	1	1
11	0	25	25

Table 18: Count table of no_of_previous_cancellations vs booking_status

Observations

- No. of previous cancellations more than 11 tend to cancel the booking. This implies that customer with more no. of booking likely to cancel the booking.

no_of_week_nights vs booking_status

Observations

- More no. of cancellations were done by the guests who book for 2 week nights.

booking_status	Canceled	Not_Canceled	All
no_of_week_nights			
All	11885	24390	36275
2	3997	7447	11444
3	2574	5265	7839
1	2572	6916	9488
4	1143	1847	2990
0	679	1708	2387
5	632	982	1614
6	88	101	189
10	53	9	62
7	52	61	113
8	32	30	62
9	21	13	34
11	14	3	17
15	8	2	10
12	7	2	9
13	5	0	5
14	4	3	7
16	2	0	2
17	2	1	3

Table 19: Count table of no_of_week_nights vs booking_status

no_of_weekend_nights vs booking_status

Observations

- It can be clearly observed that no. of weekend nights inceases, no. of booking increases and no. of cancellation also increases.

booking_status	Canceled	Not_Canceled	All
no_of_weekend_nights			
All	11885	24390	36275
0	5093	11779	16872
1	3432	6563	9995
2	3157	5914	9071
4	83	46	129
3	74	79	153
5	29	5	34
6	16	4	20
7	1	0	1

Table 20: Count table of no_of_weekend_nights vs booking_status

no_of_children vs booking_status

booking_status	Canceled	Not_Canceled	All
no_of_children			
All	11885	24390	36275
0	10882	22695	33577
1	540	1078	1618
2	457	601	1058
3	5	14	19
9	1	1	2
10	0	1	1

Table 21: Count table of no_of_children vs booking_status

Observations

- Guests with no children cancel more bookings then guests booking with children.

no_of_adults vs booking_status

Observations

- Bookings with 2 adults have the highest cancellation.

booking_status	Canceled	Not_Canceled	All
no_of_adults			
All	11885	24390	36275
2	9119	16989	26108
1	1856	5839	7695
3	863	1454	2317
0	44	95	139
4	3	13	16

Table 22: Count table of no_of_adults vs booking_status

avg_price_per_room vs booking_status

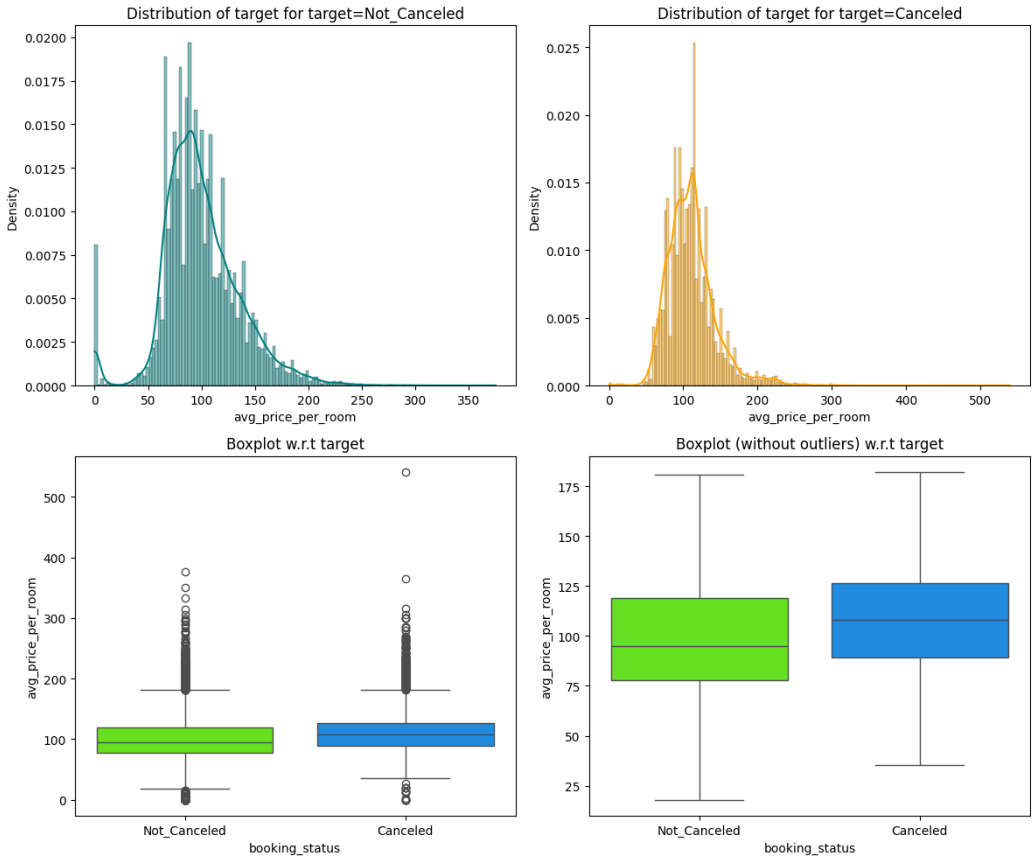

Fig 23: Histogram-boxplot of avg_price_per_room vs booking_status

Observations

- The distribution for both "Canceled" and “Not_Canceled” is right-skewed, but the price for "Canceled" distribution range appears broader, with some higher price outliers extending up to 500. There is a slightly higher concentration of cancellations around the price range of 100-125, and the density decreases quickly beyond this range.
- The tail of the distribution for canceled bookings shows a longer and more dispersed range, indicating that cancellations are more common at higher price points compared to non-canceled bookings.

- For the **"Not_Canceled"** category, the interquartile range (IQR) lies between 50 and 100, with fewer outliers above 150.
- For the **"Canceled"** category, the IQR is wider, with values extending from about 60 to 140, and more outliers in the higher price ranges (above 200).
- The second boxplot (without outliers) shows that even when outliers are removed, the median price for canceled bookings is noticeably higher than that of non-canceled bookings. The median value for canceled bookings is higher (~100) than for non-canceled (~80).
- From the distributions and boxplots, it's evident that bookings with a higher average price per room are more likely to be canceled. The median and the IQR for canceled bookings both show higher price ranges compared to non-canceled bookings.
- The price sensitivity among customers seems to play a significant role. As the price increases beyond 100, the likelihood of cancellation appears to rise. The broader distribution and more frequent outliers in the canceled group suggest that customers may feel uncertain or reconsider their bookings as room prices increase.

lead_time vs booking_status

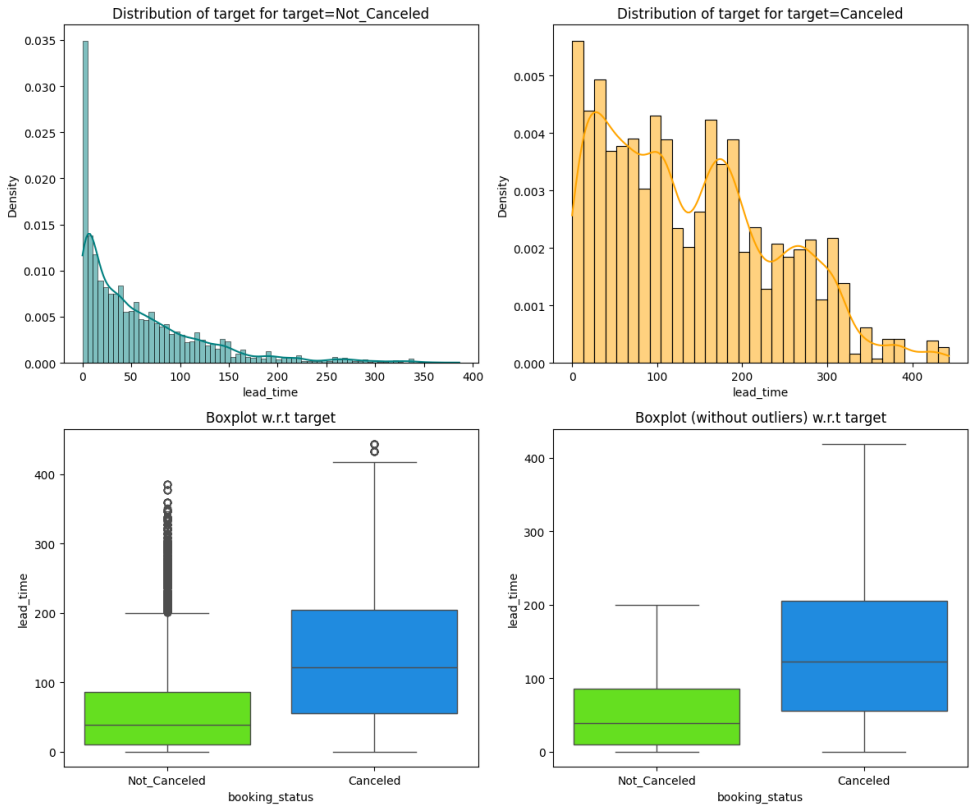

Fig 24: Histogram-boxplot of lead_time vs booking_status

Observations

- The left density plot shows that the lead time for non-canceled bookings is heavily skewed to the right, with most bookings made close to the arrival date. The highest density is for lead times between 0 and 50 days.

- The right density plot shows that canceled bookings exhibit a more uniform distribution over lead times, with a much longer tail. Canceled bookings have a higher probability of occurring at longer lead times compared to non-canceled bookings.
- The boxplot (bottom-left) shows that the median lead time for non-canceled bookings is significantly lower than for canceled bookings. The interquartile range (IQR) for non-canceled bookings is between 20 and 100 days, while the IQR for canceled bookings is much broader, from 60 to around 250 days.
- The boxplot (bottom-right), which excludes outliers, confirms this observation. Even without extreme values, canceled bookings tend to have higher lead times compared to non-canceled bookings.
- From the distributions and boxplots, it's clear that bookings with longer lead times have a higher probability of being canceled. Cancellations are more frequent when there is a large gap between the booking date and the stay date. This could be attributed to changes in customers' plans or reconsideration over time.
- Non-canceled bookings are primarily concentrated in the short lead-time range, meaning customers are more likely to follow through with their reservations if they book closer to their stay date.

EDA INSIGHTS

- Cancellations are influenced by factors like lead time, guest type (repeated vs new), and booking channel.
- The higher cancellation rate among first-time guests suggests a potential area for improvement. The hotel could investigate the reasons behind the cancellations and address these issues, perhaps through flexible booking options or improved communication.
- Given that repeated guests rarely cancel their bookings, it would be strategic for the hotel to focus on increasing the number of loyal, repeat customers. This can be achieved through targeted loyalty programs, personalized marketing, and excellent customer service.
- Businesses may want to pay special attention to bookings with higher room prices as these are at higher risk of cancellation. Offering flexible cancellation policies or incentives for high-priced rooms might help in reducing the cancellation rate.
- To mitigate cancellations, hotels may want to consider implementing stricter cancellation policies for long lead-time bookings. Alternatively, offering incentives or reminders to customers who book far in advance could reduce the likelihood of cancellations.

- October is busiest month at the same time more no. of cancellation also happened in that month.

- The overwhelming preference for Room Type 1 suggests that this room type should be prioritized in terms of inventory, maintenance, and promotional efforts. Ensuring the availability and quality of Room Type 1 could significantly enhance customer satisfaction and occupancy rates.
- Customers with more special requests tend to cancel less often. Tailoring services to accommodate these unique preferences not only enhances customer satisfaction but also increases booking reliability.
- Different room types may have varying cancellation rates, suggesting possible influence of pricing tiers or guest preference.

Data Preprocessing

Outlier Check

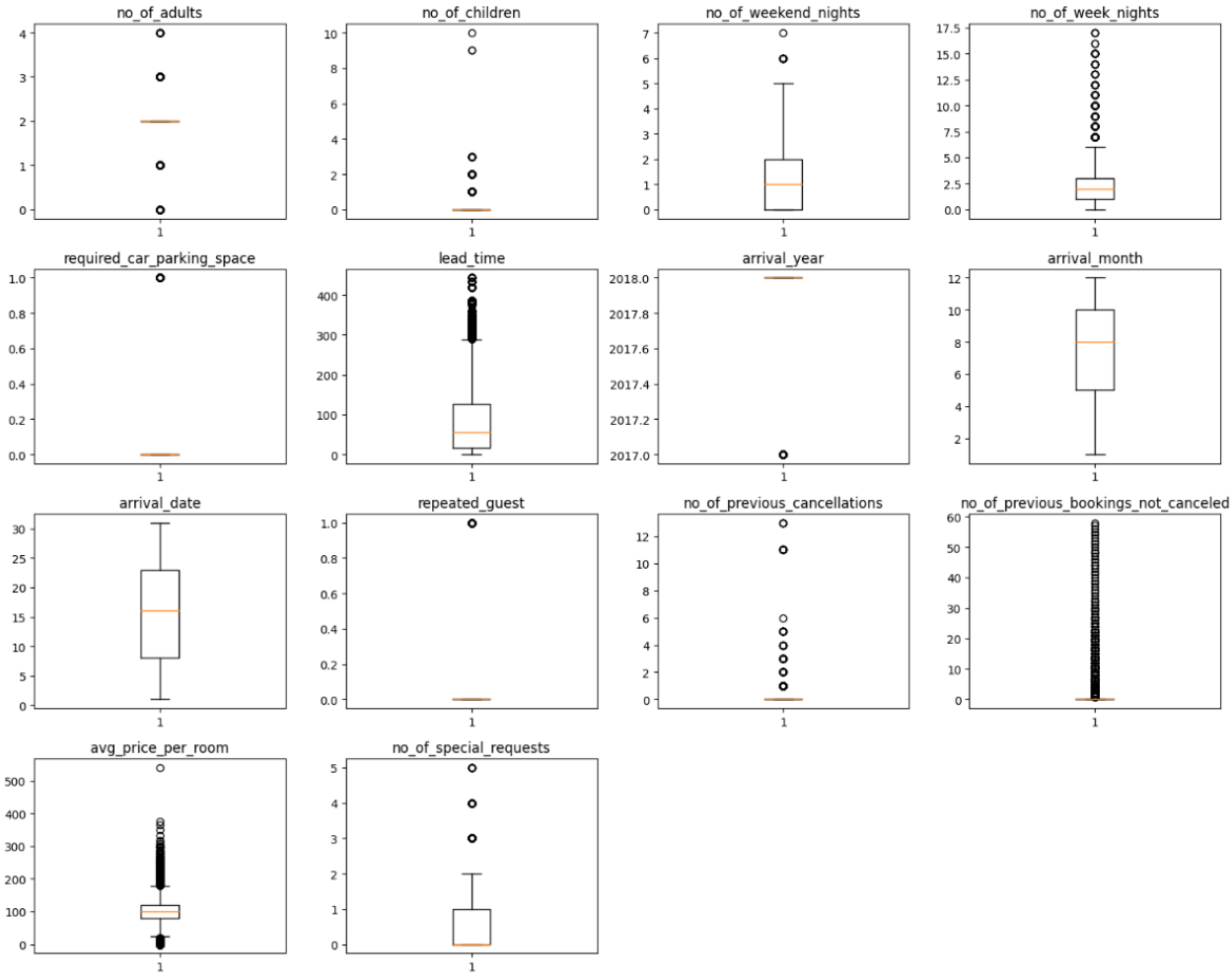

Fig 25: Barplot of all numerical variable

Observations

- **lead_time:** These outliers represent that the some customers might booked long back from the arrival date and hence, cannot not be removed.
- **avg_price-per_room:** The outliers might represent luxury or premium offerings and should not be removed.
- Other Numerical Variable doesnot have continuous values and the ouliers are also normal. Therefore, no need to outlier treatment in this case.

INFERENCE

- There are no outliers that require treatment.
- As We are predicting which booking is likely to be canceled. So, Booking_Status" column converted from categorical to numerical (int) for building models with Not_Canceled as 0 & Canceled as 1
- Booking_Status is defined as dependent variable.
- The data is splitted into train and test sets in a 70:30 ratio for further evaluation.-

```
Shape of Training set :  (25392, 27)
Shape of test set :  (10883, 27)
Shape of Training set :  (25392,)
Shape of test set :  (10883,)
Percentage of classes in training set:
Not_Canceled    0.67238
Canceled        0.32762
Name: booking_status, dtype: float64
Percentage of classes in test set:
Not_Canceled    0.67233
Canceled        0.32767
Name: booking_status, dtype: float64
```

Table 23: Training and test set size

Next, I will build a logistic regression using statsmodels, KNN-Cluster with k=2, Naives-Bayes Cluster and Decision Tree Classifier

Model Building

I will build a logistic regression using statsmodels, KNN-Cluster, Naives-Bayes Cluster, Decision Tree Classifier.

Let’s build and check the model performance.

Logistic Regression using statsmodel –

	coef	std err	z	P>|z|	[0.025	0.975]
const	−1.8366	5650.480	−0.000	1.000	−1.11e+04	1.11e+04
no_of_adults	0.0172	0.020	0.883	0.377	−0.021	0.056
no_of_children	0.0336	0.025	1.371	0.170	−0.014	0.082
no_of_weekend_nights	0.1272	0.017	7.368	0.000	0.093	0.161
no_of_week_nights	0.0497	0.017	2.881	0.004	0.016	0.084
required_car_parking_space	−0.2841	0.024	−11.769	0.000	−0.331	−0.237
lead_time	1.3580	0.023	58.946	0.000	1.313	1.403
arrival_year	0.1682	0.023	7.288	0.000	0.123	0.213
arrival_month	−0.1462	0.020	−7.315	0.000	−0.185	−0.107
arrival_date	0.0260	0.017	1.528	0.127	−0.007	0.059
repeated_guest	−0.3032	0.121	−2.502	0.012	−0.541	−0.066
no_of_previous_cancellations	0.1211	0.035	3.413	0.001	0.052	0.191
no_of_previous_bookings_not_canceled	−2.3837	1.562	−1.526	0.127	−5.445	0.677
avg_price_per_room	0.6480	0.026	24.903	0.000	0.597	0.699
no_of_special_requests	−1.1710	0.024	−48.963	0.000	−1.218	−1.124
type_of_meal_plan_Meal Plan 2	0.0502	0.019	2.607	0.009	0.012	0.088
type_of_meal_plan_Meal Plan 3	0.2878	4117.565	6.99e−05	1.000	−8069.992	8070.567
type_of_meal_plan_Not Selected	0.0695	0.019	3.741	0.000	0.033	0.106
room_type_reserved_Room_Type 2	−0.0560	0.018	−3.129	0.002	−0.091	−0.021
room_type_reserved_Room_Type 3	0.0167	0.027	0.628	0.530	−0.035	0.069
room_type_reserved_Room_Type 4	−0.1000	0.020	−5.019	0.000	−0.139	−0.061
room_type_reserved_Room_Type 5	−0.0566	0.018	−3.173	0.002	−0.092	−0.022
room_type_reserved_Room_Type 6	−0.1356	0.025	−5.534	0.000	−0.184	−0.088
room_type_reserved_Room_Type 7	−0.0897	0.020	−4.566	0.000	−0.128	−0.051
market_segment_type_Complementary	−4.7541	5.49e+04	−8.66e−05	1.000	−1.08e+05	1.08e+05
market_segment_type_Corporate	−0.1951	0.063	−3.087	0.002	−0.319	−0.071
market_segment_type_Offline	−0.7992	0.119	−6.688	0.000	−1.033	−0.565
market_segment_type_Online	0.0034	0.125	0.027	0.979	−0.242	0.248

Table 24: Logistic Regression summary

Observations

- Negative values of the coefficient show that the probability of a customer cancel the booking decreases with the increase of the corresponding attribute value.
- Positive values of the coefficient show that the probability of a customer cancel the booking increases with the increase of the corresponding attribute value.

- p-value of a variable indicates if the variable is significant or not. If we consider the significance level to be 0.05 (5%), then any variable with a p-value less than 0.05 would be considered significant.
- There are some attributes which have high p-values such as type_of_meal_plan_Meal Plan 3 (1.0), room_type_reserved_Room_Type

3(0.53), market_segment_type_Complementary (1.0), market_segment_type_Online(0.979). But these variables might contain multicollinearity, which will affect the p-values.

Logistic Regression Model Performance

Train sets

	Accuracy	Recall	Precision	F1
0	0.80687	0.63313	0.73985	0.68234

Table 25: Logistic Regression model performance on training set

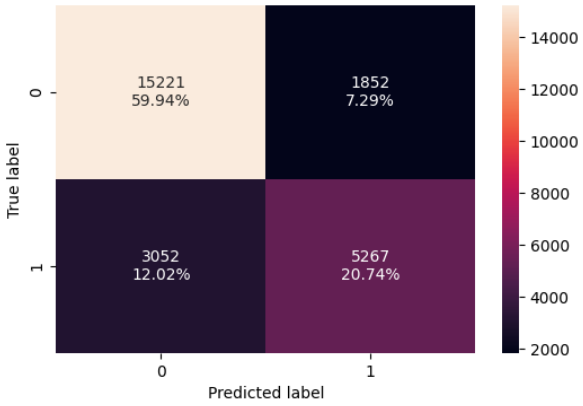

Fig 26: Logistic Regression Confusion matrix on training set

Test sets

	Accuracy	Recall	Precision	F1
0	0.80391	0.62703	0.73553	0.67696

Table 26: Logistic Regression model performance on test set

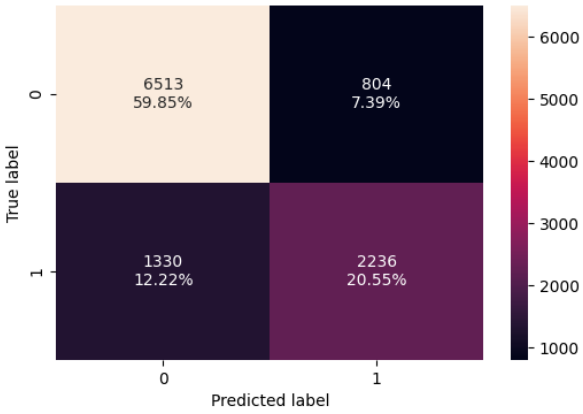

Fig 27: Logistic Regression Confusion matrix on test set

- The f1_score of Logistic Regression model performance on both training and test set is very low with ~0.68 & ~0.67 respectively.

Naive-Bayes Classifier -

Naive-Bayes Classifier Model Performance

Train sets	Test sets

	Accuracy	Recall	Precision	F1
0	0.40903	0.96490	0.35297	0.51687

Table 27: Naive-Bayes Classifier model performance on training set

	Accuracy	Recall	Precision	F1
0	0.40742	0.96411	0.35229	0.51602

Table 28: Naive-Bayes Classifier model performance on test set

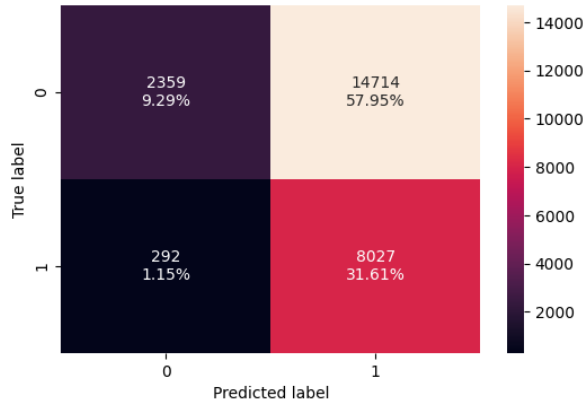

Fig 28: Naive-Bayes Classifier Confusion matrix on training set

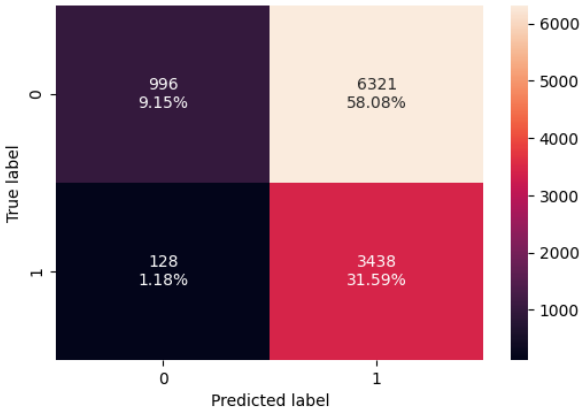

Fig 29: Naive-Bayes Classifier Confusion matrix on test set

- Naive Bayes achieves very high recall (0.96490) on the training set but at the cost of very low precision (0.35297), resulting in a poor F1 score (0.51687)

- The test set performance is consistent, with high recall (0.96411) but low precision (0.35229) and F1 score (0.51602). This indicates the model predicts most positive cases correctly but makes many false positive predictions.

KNN Classifier (K=2) -

Model evaluation criterion

Model can make wrong predictions as:

- **False Positive(FP):** Predicting a customer canceled the booking but in reality the customer did not canceled the booking.
- **False Negative(FN):** Predicting a customer did not canceled the booking but in reality the customer canceled the booking.

To reduce the losses which might occur due false negative the recall need to be maximized, greater the recall score higher are the chances of minimizing the False Negatives.

In order to optimize our model, it's essential to experiment with different values of k to find the most suitable fit for our data. We can commence this process by setting k equal to 2 and gradually exploring other values to assess their impact on the model's performance.

KNN Classifier (K=2) Model Performance

Train sets

	Accuracy	Recall	Precision	F1
0	0.92198	0.76536	0.99547	0.86538

Table 29: KNN Classifier (K=2) model performance on training set

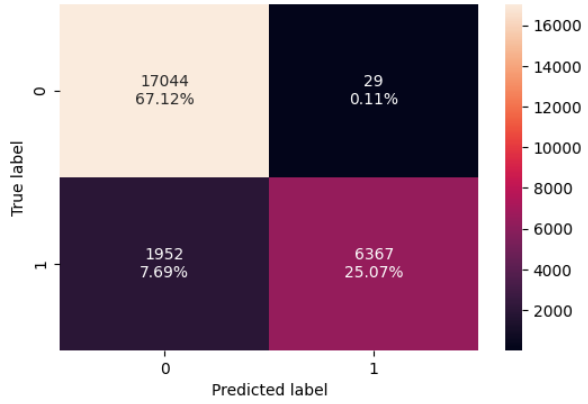

Fig 30: KNN Classifier (K=2) Confusion matrix on training set

Test sets

	Accuracy	Recall	Precision	F1
0	0.84434	0.63208	0.85508	0.72686

Table 30: KNN Classifier (K=2) model performance on test set

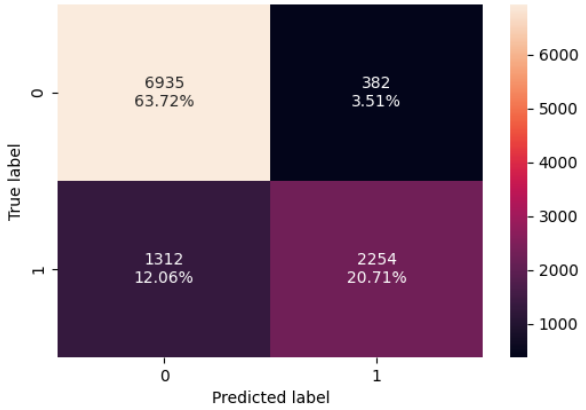

Fig 31: KNN Classifier (K=2) Confusion matrix on test set

- KNN Classifier (K=2) achieves high Accuracy (0.92198) and high precision on the training set but with low recall (0.76536).
- KNN Classifier (K=2) on test set achieves even more poor values in all parameters.

Decision Tree Classifier-

Decision Tree Classifier Model Perfromance

Train sets

	Accuracy	Recall	Precision	F1
0	0.99437	0.98570	0.99708	0.99136

Table 31: Decision Tree Classifier model performance on training set

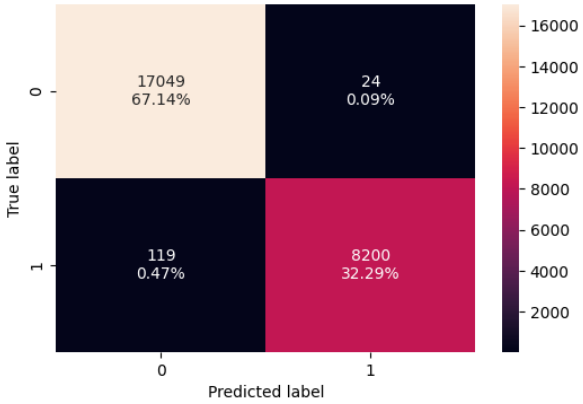

Fig 32: Decision Tree Classifier Confusion matrix on training set

Test sets

	Accuracy	Recall	Precision	F1
0	0.86539	0.79501	0.79434	0.79467

Table 32: Decision Tree Classifier model performance on test set

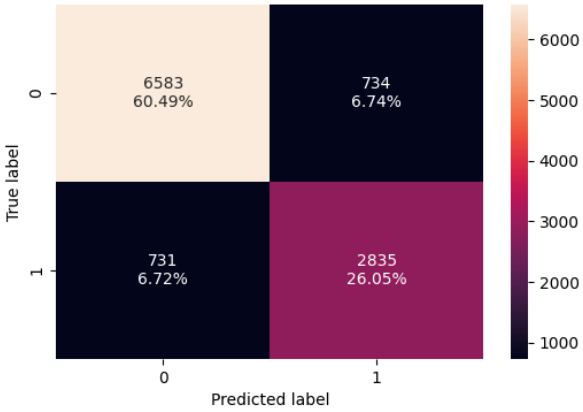

Fig 33: Decision Tree Classifier Confusion matrix on test set

- The model is giving a generalized result now since the recall scores on the training set is 0.98 which shows that the model is able to generalize well on unseen data, but it is of indicative of overfitting as well.

- The recall scores on the test set is 0.79 which is low.

Model Performance Improvement

Logistic Regression (deal with multicollinearity, remove high p-value variables, determine optimal threshold using ROC curve) -

There are some attributes which have high p-values. But these variables might contain multicollinearity, which will affect the p-values.

We will have to remove multicollinearity from the data to get reliable coefficients and p-values, for which we are using Variance Inflation Factors (VIF).

Logistic Regression - Dealing with Multicollinearity

Variance Inflation Factors:

	Variable	VIF
0	const	1.00000
1	no_of_adults	1.34506
2	no_of_children	2.00721
3	no_of_weekend_nights	1.06725
4	no_of_week_nights	1.09437
5	required_car_parking_space	1.03494
6	lead_time	1.40192
7	arrival_year	1.43326
8	arrival_month	1.27740
9	arrival_date	1.00763
10	repeated_guest	1.75019
11	no_of_previous_cancellations	1.32201
12	no_of_previous_bookings_not_canceled	1.57086
13	avg_price_per_room	2.03263
14	no_of_special_requests	1.24723
15	type_of_meal_plan_Meal Plan 2	1.26182
16	type_of_meal_plan_Meal Plan 3	1.00796
17	type_of_meal_plan_Not Selected	1.27921
18	room_type_reserved_Room_Type 2	1.09458
19	room_type_reserved_Room_Type 3	1.00390
20	room_type_reserved_Room_Type 4	1.35578
21	room_type_reserved_Room_Type 5	1.03090
22	room_type_reserved_Room_Type 6	1.99011
23	room_type_reserved_Room_Type 7	1.09207
24	market_segment_type_Complementary	4.35005
25	market_segment_type_Corporate	16.63442
26	market_segment_type_Offline	62.51301
27	market_segment_type_Online	69.47183

Table 33: VIF of Logistic Regression model

Observations

- Most of the variables have VIF values below 5, indicating low to moderate multicollinearity. This suggests that these variables are not highly correlated with each other, which is generally acceptable for a logistic regression model.

- Market_segment_type_Online and market_segment_type_Offline exhibit very high VIF values (69.47 and 62.51, respectively). This indicates severe multicollinearity, meaning these variables are highly correlated with other predictors in the model. Such high VIFs could lead to instability in the regression coefficients and unreliable statistical inferences.
- Market_segment_type_Corporate also has a high VIF (16.63), indicating a significant degree of multicollinearity that should be addressed.
- The high VIF values for certain market segments suggest that the model may benefit from refinement.
- Variables such as no_of_adults, no_of_children, lead_time, avg_price_per_room, and no_of_special_requests have VIFs around or below 2, indicating that multicollinearity is not a significant issue for these predictors.
- The presence of high VIF values in some predictors means that careful consideration is required when interpreting the model coefficients. Multicollinearity can distort the estimated effect sizes and make the model more sensitive to changes in the data.

Dropping market_segment_type_Online due to high VIF. And lets check the VIF score again.

Variance Inflation Factors:

	Variable	VIF
0	const	1.00000
1	no_of_adults	1.32868
2	no_of_children	2.00631
3	no_of_weekend_nights	1.06689
4	no_of_week_nights	1.09377
5	required_car_parking_space	1.03486
6	lead_time	1.39751
7	arrival_year	1.43053
8	arrival_month	1.27629
9	arrival_date	1.00760
10	repeated_guest	1.74705
11	no_of_previous_cancellations	1.32192
12	no_of_previous_bookings_not_canceled	1.57053
13	avg_price_per_room	2.03183
14	no_of_special_requests	1.24244
15	type_of_meal_plan_Meal Plan 2	1.26149
16	type_of_meal_plan_Meal Plan 3	1.00796
17	type_of_meal_plan_Not Selected	1.27711
18	room_type_reserved_Room_Type 2	1.09440
19	room_type_reserved_Room_Type 3	1.00390
20	room_type_reserved_Room_Type 4	1.35157
21	room_type_reserved_Room_Type 5	1.03090
22	room_type_reserved_Room_Type 6	1.98976
23	room_type_reserved_Room_Type 7	1.09194
24	market_segment_type_Complementary	1.32469
25	market_segment_type_Corporate	1.53067
26	market_segment_type_Offline	1.60207

Table 34: VIF of Logistic Regression model after dropping market_segment_type_Online

Now, none of the variables exhibit high multicollinearity, so the values in the summary are reliable.

Let's remove the insignificant features (p-value>0.05).

Remove high p-value variables

Sometimes p-values change after dropping a variable. So, we'll not drop all variables at once.

Instead, we will do the following repeatedly using a loop:

- Build a model, check the p-values of the variables, and drop the column with the highest p-value.
- Create a new model without the dropped feature, check the p-values of the variables, and drop the column with the highest p-value.
- Repeat the above two steps till there are no columns with p-value > 0.05.

Training the Logistic Regression model again with only the significant features

	coef	std err	z	P>|z|	[0.025	0.975]
const	−1.1787	0.024	−50.100	0.000	−1.225	−1.133
no_of_weekend_nights	0.1304	0.017	7.568	0.000	0.097	0.164
no_of_week_nights	0.0509	0.017	2.953	0.003	0.017	0.085
required_car_parking_space	−0.2841	0.024	−11.780	0.000	−0.331	−0.237
lead_time	1.3664	0.023	59.908	0.000	1.322	1.411
arrival_year	0.1650	0.023	7.172	0.000	0.120	0.210
arrival_month	−0.1513	0.020	−7.595	0.000	−0.190	−0.112
repeated_guest	−0.4860	0.094	−5.152	0.000	−0.671	−0.301
no_of_previous_cancellations	0.1007	0.027	3.722	0.000	0.048	0.154
avg_price_per_room	0.6689	0.025	26.645	0.000	0.620	0.718
no_of_special_requests	−1.1672	0.024	−49.272	0.000	−1.214	−1.121
type_of_meal_plan_Meal Plan 2	0.0480	0.019	2.492	0.013	0.010	0.086
type_of_meal_plan_Not Selected	0.0731	0.018	3.974	0.000	0.037	0.109
room_type_reserved_Room_Type 2	−0.0501	0.017	−2.891	0.004	−0.084	−0.016
room_type_reserved_Room_Type 4	−0.1000	0.019	−5.189	0.000	−0.138	−0.062
room_type_reserved_Room_Type 5	−0.0571	0.018	−3.211	0.001	−0.092	−0.022
room_type_reserved_Room_Type 6	−0.1189	0.019	−6.186	0.000	−0.157	−0.081
room_type_reserved_Room_Type 7	−0.0868	0.019	−4.504	0.000	−0.125	−0.049
market_segment_type_Corporate	−0.1996	0.024	−8.445	0.000	−0.246	−0.153
market_segment_type_Offline	−0.8028	0.023	−34.288	0.000	−0.849	−0.757

Table 35: Logistic Regression with significant features summary

Now no feature has p-value greater than 0.05, so we'll consider the features in LogisticReg_tuned as the final trained model.

Determining the optimal threshold using ROC Curve

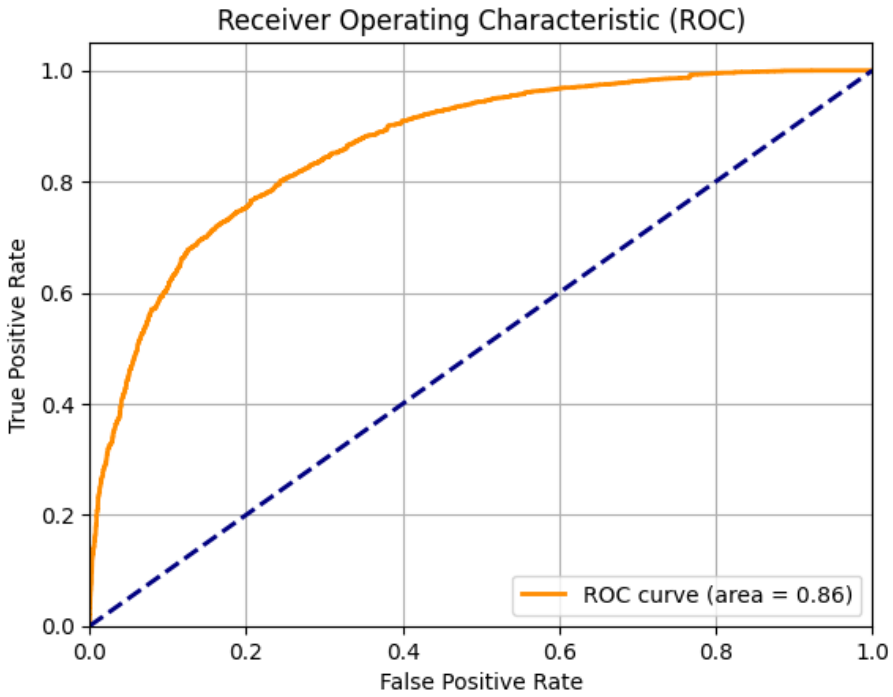

Fig 34: ROC Curve

- The **ROC-AUC Score is 0.86** which is fairly good.
- Optimal Threeshold is **0.333.**

Tunes Logistic Regression model performance

Train sets

	Accuracy	Recall	Precision	F1
0	0.78434	0.76608	0.64354	0.69948

Table 36: Tuned Logistic Regression model performance on training set

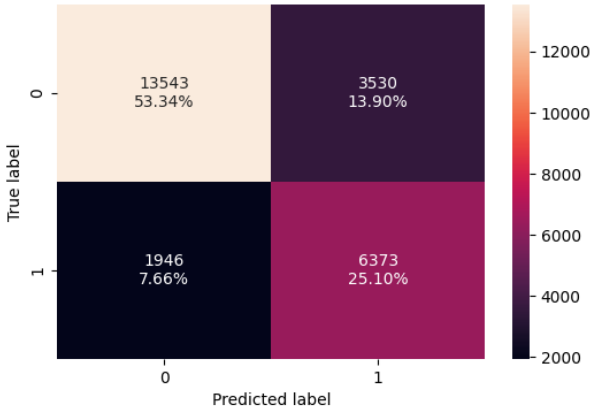

Fig 35: Tuned Logistic Regression confusion matrix on training set

- Accuracy has decreased from 0.80 to 0.78.
- Precision has decreased from 0.73 to 0.64.
- Model is giving a recall of 0.76 as compared to initial model which was giving a recall of 0.63.
- F1 score has improve from 0.68 to 0.69.

Test sets

	Accuracy	Recall	Precision	F1
0	0.77846	0.76612	0.63402	0.69384

Table 37: Tuned Logistic Regression model performance on test set

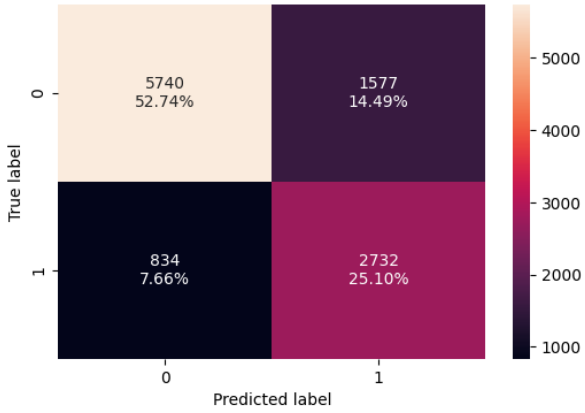

Fig 36: Tuned Logistic Regression confusion matrix on test set

- Accuracy has decreased from 0.80 to 0.77.
- Precision has decreased from 0.73 to 0.63.
- Model is giving a recall of 0.76 as compared to initial model which was giving a recall of 0.62.
- F1 score has improve from 0.67 to 0.69.
- The changes in both training sets and test sets is moreover similar.

KNN Classifier Performance Improvement using different k values -

Let's run the KNN with no of neighbours to be 2 to 20 and finding the optimal number of neighbours from the above list using the recall score.

The **best value of k we get is 3** with a **recall of: 0.7372406057206955**

Building KNN model with number of neighbors as best_k

KNN Classifier Model Performance with best k value

Train sets

	Accuracy	Recall	Precision	F1
0	0.91513	0.84806	0.88787	0.86751

Table 38: Tuned KNN Classifier model performance on training set

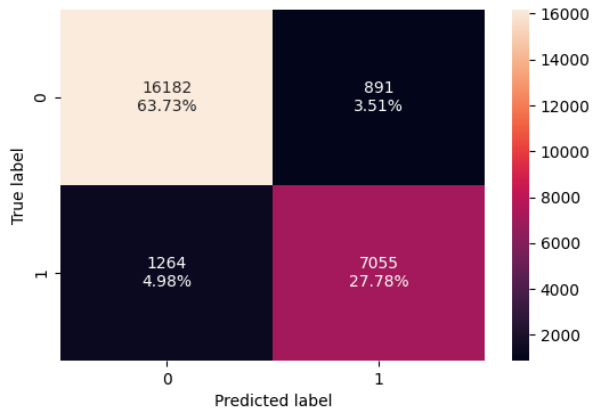

Fig 37: Tuned KNN Classifier confusion matrix on training set

Test sets

	Accuracy	Recall	Precision	F1
0	0.84416	0.73724	0.77597	0.75611

Table 39: Tuned KNN Classifier model performance on test set

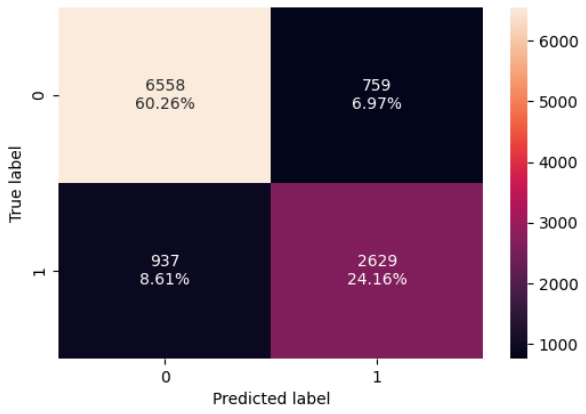

Fig 38: Tuned KNN Classifier confusion matrix on test set

- The tuned KNN model shows excellent performance with high accuracy (0.92198), precision (0.99547), recall (0.76536), and F1 score (0.86538). The base model is far behind.

- On the test set, the tuned KNN maintains high accuracy (0.84434) and precision (0.85508), but recall drops (0.63208), leading to a lower F1 score (0.72686).

Decision Tree Classifier (pre-pruning) -

The decision tree is quite deep and complex, indicating that the model is likely fitting the training data with many decisions. This level of complexity may result in overfitting, where the model captures noise in the data rather than the general patterns. **Pruning** could be considered to simplify the model and enhance generalization.

Decision Tree Classifier (pre-pruning) Model Performance

Train sets

	Accuracy	Recall	Precision	F1
0	0.86354	0.84000	0.76606	0.80133

Table 40: Tuned Decision Tree model performance on training set

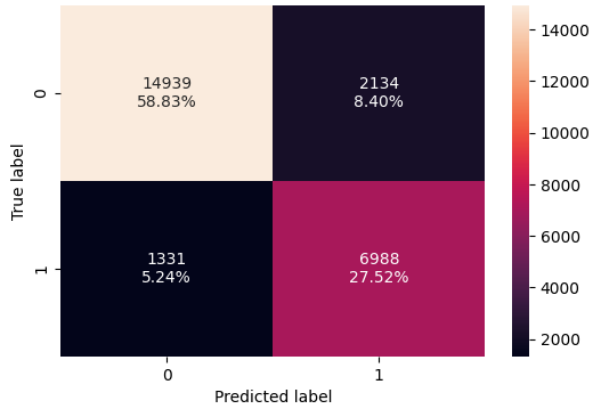

Fig 39: Tuned Decision Tree confusion matrix on training set

Test sets

	Accuracy	Recall	Precision	F1
0	0.85427	0.82025	0.75581	0.78671

Table 41: Tuned Decision Tree model performance on test set

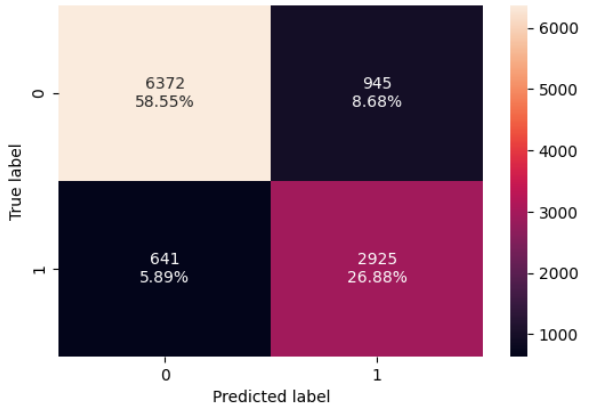

Fig 40: Tuned Decision Tree confusion matrix on test set

- On the training set, the tuned decision tree significantly reduces overfitting, balancing out accuracy (0.86354), recall (0.84000), and precision (0.76606), with a solid F1 score (0.80133).
- On the test set, the tuned decision tree has high accuracy (0.85427), recall (0.82025), and F1 score (0.78671), outperforming most other models.

Decision Tree Visualization.

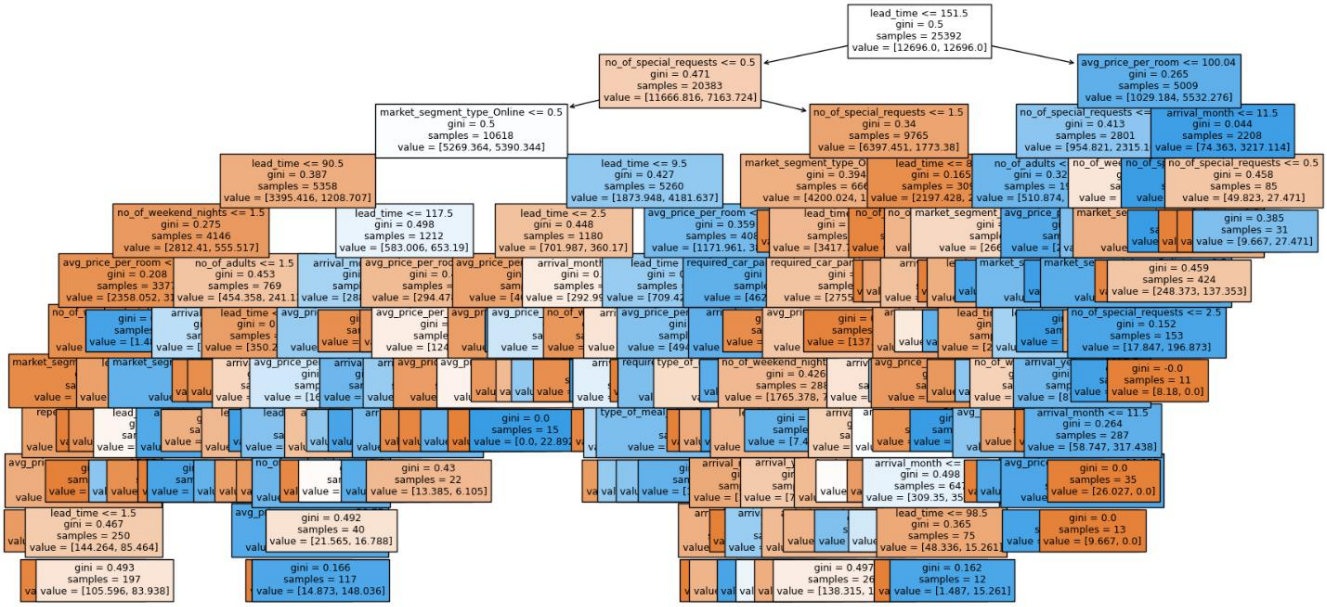

Fig 41: Decision Tree

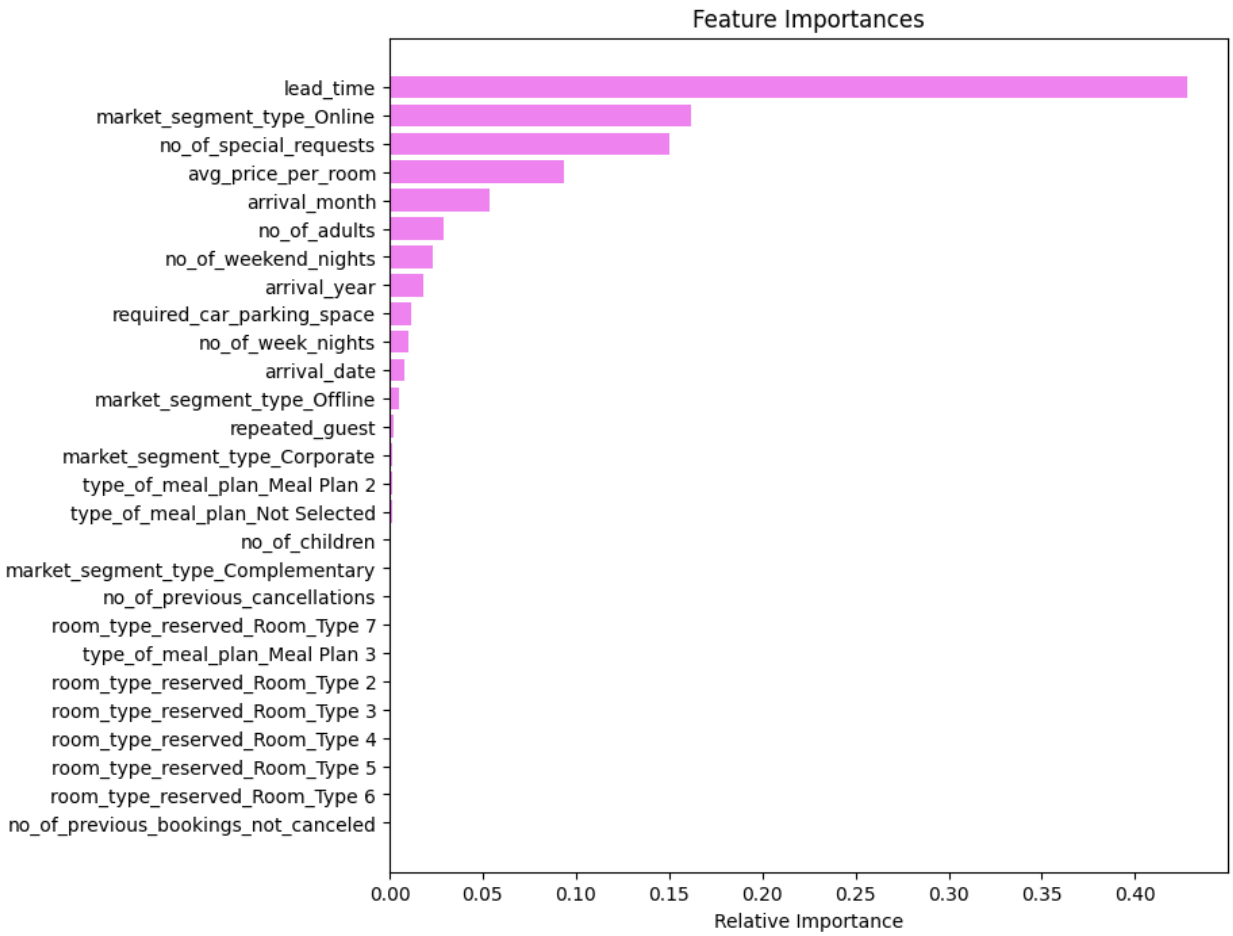

Fig 42: Feature importance

Observations

- The first split of the decision tree is based on lead_time <= 151.5 days. This indicates that **lead time** (the time between booking and check-in) is the most significant factor influencing the target outcome (likely cancellation or booking status).
- As seen from multiple splits, lower lead times (<= 90.5, <= 9.5, <= 2.5) significantly impact the model's decision-making process, suggesting customers who book closer to their stay are less likely to cancel.
- A high number of special requests reduces the likelihood of cancellation (fewer cancellations when no_of_special_requests > 0.5). This could indicate that customers who invest time in making specific requests are more committed to their booking.
- The feature avg_price_per_room also appears in multiple splits. Higher prices tend to result in more cancellations (avg_price_per_room > 100), which could indicate that customers are price-sensitive and more likely to cancel expensive bookings.
- Customers booking through online channels (market_segment_type_Online) seem to exhibit different cancellation behaviors than those who book via other means.
- **Key Features Contributing to Splits**:
  After lead_time, the features with the highest influence on subsequent splits include:
  - **no_of_special_requests** (whether the customer made any special requests).
  - **avg_price_per_room** (the average price per room).
  - **market_segment_type_Online** (whether the booking was made online or through another channel).
  - Other important factors include arrival_month, no_of_weekend_nights, and no_of_adults.

Model Performance Comparison and Final Model Selection

Training Model Performance comparison

Training performance comparison:

	Logistic Regression Base	Logistic Regression Tuned	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Tuned
Accuracy	0.80687	0.78434	0.40903	0.92198	0.91513	0.99437	0.86354
Recall	0.63313	0.76608	0.96490	0.76536	0.84806	0.98570	0.84000
Precision	0.73985	0.64354	0.35297	0.99547	0.88787	0.99708	0.76606
F1	0.68234	0.69948	0.51687	0.86538	0.86751	0.99136	0.80133

Table 42: Training Model Performance comparison table

Test Model comparison

Test set performance comparison:

	Logistic Regression Base	Logistic Regression Tuned	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Tuned
Accuracy	0.80687	0.78434	0.40742	0.84434	0.84416	0.86539	0.85427
Recall	0.63313	0.76608	0.96411	0.63208	0.73724	0.79501	0.82025
Precision	0.73985	0.64354	0.35229	0.85508	0.77597	0.79434	0.75581
F1	0.68234	0.69948	0.51602	0.72686	0.75611	0.79467	0.78671

Table 43: Test Model Performance comparison table

Observations

- The tuned logistic regression model improves recall, indicating better handling of imbalanced classes, but sacrifices some accuracy and precision.
- Naive Bayes may not be the best choice due to its extremely low precision despite high recall.
- The KNN tuned model performs very well on the training set and test set but suffers from lower recall on the test set. This suggests some overfitting and a potential issue with generalization.
- The decision tree model, after tuning, provides a good balance of all metrics across both training and test sets. It is less overfitted compared to the base model and performs consistently well.

Final model selection

- The **Tuned Decision Tree** seems to be the best model overall. It provides a good balance between precision, recall, and F1 score, making it robust and reliable. It avoids overfitting, which was evident in the base decision tree model.

Actionable Insights & Recommendations

- The analysis highlights **lead time**, **booking channel**, **special requests** and **average price**, **arrival month**, **no. of adults** as primary drivers of booking cancellations.
- **Canceled bookings tend to have significantly higher lead times** compared to those that are not canceled. This suggests that the longer the lead time, the higher the likelihood of cancellation, possibly due to changes in travel plans over time.
- **The overwhelming dominance of the Online segment** suggests that any business strategies should prioritize enhancing and optimizing the online booking experience. This could include improving website usability, mobile app functionality, and digital marketing efforts.
- **October, the busiest month with the most cancellations,** hotels could offer incentives like discounts and flexible booking options to encourage guests to maintain their reservations.
- **Repeated guests rarely cancel their bookings,** it would be beneficial for hotels to focus on increasing the number of loyal customers. This can be achieved by developing loyalty programs that offer rewards for returning guests.
- **Customers with specific requests are less likely to cancel.** By customizing services to meet these unique needs, hotels can enhance customer satisfaction and booking reliability.
- Among the models evaluated, **Tuned Decision Tree** seems to be the best model having higher recall, precision, accuracy scores for the training set and in test sets have moderate good scores. Higher recall indicates more effective at correctly identifying likelihood of booking cancellation compared to other models.
- **Good recall performance indicates its effectiveness in minimizing false negatives,** which is critical for reducing booking cancellations.

- **INN Hotel Group should integrate the Tuned Decision Tree model into their operational workflow** as part of a comprehensive predictive maintenance system. By leveraging this model, the company can proactively detect the chances of a booking to be canceled which will help in minimizing the company's loss oc cured due to booking cancellation.

Conclusion

The analysis highlights several opportunities to **enhance no. of booking** and **reduce booking cancellation**. By focusing on **online market segment with repeated customer**, **low lead time bookings, improving customer services, and aligning deals and discount during festive season or vacation time**, the business can achieve sustainable growth and improved customer satisfaction which will ultimately reduce cancellations. These strategies, coupled with personalized customer engagement and leveraging seasonal trends, will enable the company to maximize its revenue potential and reduce cancellation booking.

The above data-driven insights, will help the INN Hotels Group to develop more targeted strategies to manage cancellations and improve customer retention, thereby improving revenue stability and operational efficiency.

END