

Machine Learning -1

Coded Project

Business Report

DSBA – Course

Created by – Rishabh Gupta

Foreword

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 behaviour. 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 impacts a hotel on various fronts:

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

Contents

Sr. No	Topics	Pages
1	Objective	6
2	Data Overview	7
3	Statistical summary of data	9
4	EDA - Univariate Analysis	11
5	Bivariate Analysis	21
6	Comments to EDA Questions	31
7	Data Preprocessing	35
8	Model Building and analysis	38
9	Decision Tree and performance improvement	54
10	Model Performance Comparison and Final Model Selection	63
11	Actionable Insights and Recommendations	67

List of Tables

Sr. No	Name of Tables	Pages
1	Top 5 rows of dataset	7
2	Basic info of dataset	8
3	Statistical summary of dataset	9
4	Summary of datatypes	11
5	Training and Test set summary	36
6	Logistic regression	38
7	Test performance on Training model	40
8	Test performance on Test model	41
9	VIF values	42
10	Model results after eliminating high p-values	43
11	Test performance for LG1	44

12	Odd ratios and percentage	45
13	Training model results	46
14	Training performance on Training model	48
15	Training performance on Test model	49
16	Training model performance	53
17	Test model performance	53
18	Pruning stats	57
19	Training model results for Pre and Post pruning	65
20	Test model results for Pre and Post pruning	65

List of Figures

Sr. No	Name of Figures	Pages
1	Histogram for numerical variables set 1	12
2	Histogram for numerical variables set 2	13
3	Histogram for numerical variables set 3	14
4	Boxplot for numerical variables set 1	15
5	Boxplot for numerical variables set 2	15
6	Boxplot for numerical variables set 3	16
7	Boxplot for numerical variables set 4	17
8	Boxplot for numerical variables set 5	17
9	Barplot for Type of meal plan and car parking space	18
10	Barplot for Arrival year and Room type	19
11	Barplot for Arrival month and mkt seg type	19
12	Barplot for Booking Status and Repeated numbers	20
13	Scatter plot for set 1	21
14	Scatter plot for Adults vs avg price per room	22
15	Scatter plot for mkt segment vs avg price per room	23
16	Cat plot for No pf special requests with hue as Booking status	24

17	Correlation map	25
18	Stacked barplot for % of cancellations vs mkt seg type	27
19	Stacked barplot for Booking status vs Repeated guest	28
20	Pair Plot	29
21	Line plot for arrival month vs avg price per room	30
22	Question 1 Plot	31
23	Question 1 Plot	32
24	Question 4 Plot	33
25	Question 5 Plot	33
26	Question 6 Plot	34
27	Outlier summary	35
28	Confusion matrix – training model	39
30	Confusion matrix – test model	41
31	Confusion matrix – training model 2	46
32	ROC-AUC on training set	47
33	Confusion matrix – training model with ROC-AUC	48
34	Confusion matrix – test model with ROC-AUC	49
35	ROC-AUC on test set	49
36	Plot for precision recall curve	50
37	Final model performance on training set	51
38	Final model performance on test set	52
39	Decision tree-training set matrix	54
40	Decision tree-test set matrix	55
41	Prepruning examination	56
42	Pruning Training vs Test results	57
43	Decision Tree	58
44	Important variables after Decision tree	58
45	Total impurity vs alpha	59
46	Depth vs Alpha and Alpha vs No of nodes	61

47	F1 score vs Alpha for Training and Test set	62
48	After Pruning – test results Training vs Test	63
49	Decision tree – post pruning	64
50	Important variables 2	64

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 Analysis of problem statement –

The dataset contains data corresponding to polished and unpolished stones.

Sheet name – *INNHotelsGroup.csv*

Data Dictionary –

1. **Booking_ID**: the unique identifier of each booking
2. **no_of_adults**: Number of adults
3. **no_of_children**: Number of Children
4. **no_of_weekend_nights**: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
5. **no_of_week_nights**: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
6. **type_of_meal_plan**: Type of meal plan booked by the customer:
 - a. Not Selected – No meal plan selected
 - b. Meal Plan 1 – Breakfast
 - c. Meal Plan 2 – Half board (breakfast and one other meal)
 - d. Meal Plan 3 – Full board (breakfast, lunch, and dinner)
7. **required_car_parking_space**: Does the customer require a car parking space? (0 - No, 1- Yes)
8. **room_type_reserved**: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
9. **lead_time**: Number of days between the date of booking and the arrival date
10. **arrival_year**: Year of arrival date
11. **arrival_month**: Month of arrival date
12. **arrival_date**: Date of the month
13. **market_segment_type**: Market segment designation.
14. **repeated_guest**: Is the customer a repeated guest? (0 - No, 1- Yes)
15. **no_of_previous_cancellations**: Number of previous bookings that were canceled by the customer prior to the current booking
16. **no_of_previous_bookings_not_canceled**: Number of previous bookings not canceled by the customer prior to the current booking
17. **avg_price_per_room**: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
18. **no_of_special_requests**: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
19. **booking_status**: Flag indicating if the booking was canceled or not.

Categorization of variables –

Continuous Variables - no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, lead_time, arrival_date, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, avg_price_per_room, no_of_special_requests

Categorical Variables - Booking_ID, type_of_meal_plan, required_car_parking_space, room_type_reserved, arrival_year, arrival_month, market_segment_type, repeated_guest, booking_status

Data Overview –

To start the data analysis, we need to import the necessary libraries, specify the working directory, and load the dataset. Then, we will view the first five rows using head () function to get a preliminary understanding of the data. The Dataset has 36275 number of rows with 19 columns.

	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	a
0	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	
1	INN00002	2	0	2	3	Not Selected	0	Room_Type 1	5	
2	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	
3	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	
4	INN00005	2	0	1	1	Not Selected	0	Room_Type 1	48	




TABLE 1 - TOP 5 ROWS OF DATASET


```

<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 - BASIC INFO. OF DATASET

Regarding the datatype of the columns in the dataset, we can confirm that-

- We can observe that there around 5 object datatypes columns and 13 numerical datatypes
- There are no missing values in the dataset

Missing value treatment and Analysis-

- On analysis, we can observe there are no null values in the dataset.
- Also, there are no duplicate entries.
- We have dropped column 'Booking_ID' as it will have no effect on predictive analysis.

Statistical Summary –

Using Describe () function, we can analyse the summary statistics of the dataset –

	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 3 - STATISTICAL SUMMARY OF DATASET

Observations-

Booking Information

- Booking ID: Unique for each of the 36,275 bookings.
- Guests: Most bookings are for 1-2 adults; children are rare.
- Stay Duration: Average 2.2 nights, with more weekday than weekend stays.
- Meal Plan: Meal Plan 1 is the most popular.

Booking Preferences

- Car Parking: Requested in only 3% of bookings.
- Room Type: Room Type 1 is the most booked.
- Lead Time: Bookings are made around 85 days in advance.
- Arrival Year: Primarily from 2018.

Guest Information

- Repeat Guests: Only 2.6% are repeat customers.
- Cancellations: Low rates, with minimal prior cancellations.
- Price per Room: Average is around 103 currency units.

Special Requests

- Special Requests: About 17% of bookings include special requests.
- Trailer Views: The average number of trailer views is 66.91559, with a standard deviation of 35.00108. The maximum number of trailer views is 199.92, which is considerably higher than the average.
- Content Views: The average number of content views is relatively low at 0.4734, with a standard deviation of 0.105914.
- Online Bookings: Online bookings are the most common.

Exploratory Data Analysis

Let's explore what are the values in object data types-

```
type_of_meal_plan
Meal Plan 1      27835
Not Selected     5130
Meal Plan 2      3305
Meal Plan 3         5
Name: count, dtype: int64
*****

room_type_reserved
Room_Type 1     28130
Room_Type 4      6057
Room_Type 6       966
Room_Type 2       692
Room_Type 5       265
Room_Type 7       158
Room_Type 3         7
Name: count, dtype: int64
*****

market_segment_type
Online          23214
Offline         10528
Corporate        2017
Complementary     391
Aviation         125
Name: count, dtype: int64
*****

booking_status
Not_Canceled    24390
Canceled        11885
Name: count, dtype: int64
*****
```

TABLE 4 - SUMMARY OF DATATYPES

Meal Plans:

- The most popular meal plan is "Meal Plan 1," followed by "Not Selected" and "Meal Plan 2."
- Meal Plans 3 and 4 have very low counts, suggesting they are less frequently chosen.

Room Types:

- "Room_Type 1" is the most reserved, followed by "Room_Type 4" and "Room_Type 6."
- Room Types 3, 5, 7 have significantly lower counts, indicating they might be less available or less popular.

Market Segments:

- The majority of bookings are made "Online," followed by "Offline" and "Corporate."
- "Complementary" and "Aviation" have much lower counts, suggesting they are niche or specific booking channels.

Booking Status:

- The majority of bookings are "Not Canceled," while a significant number are "Canceled."

- This indicates that a portion of bookings are not fulfilled, potentially due to cancellations or other reasons.
- **Numerical Variables** - A histogram for visualizing the distribution of numerical variables like no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights.

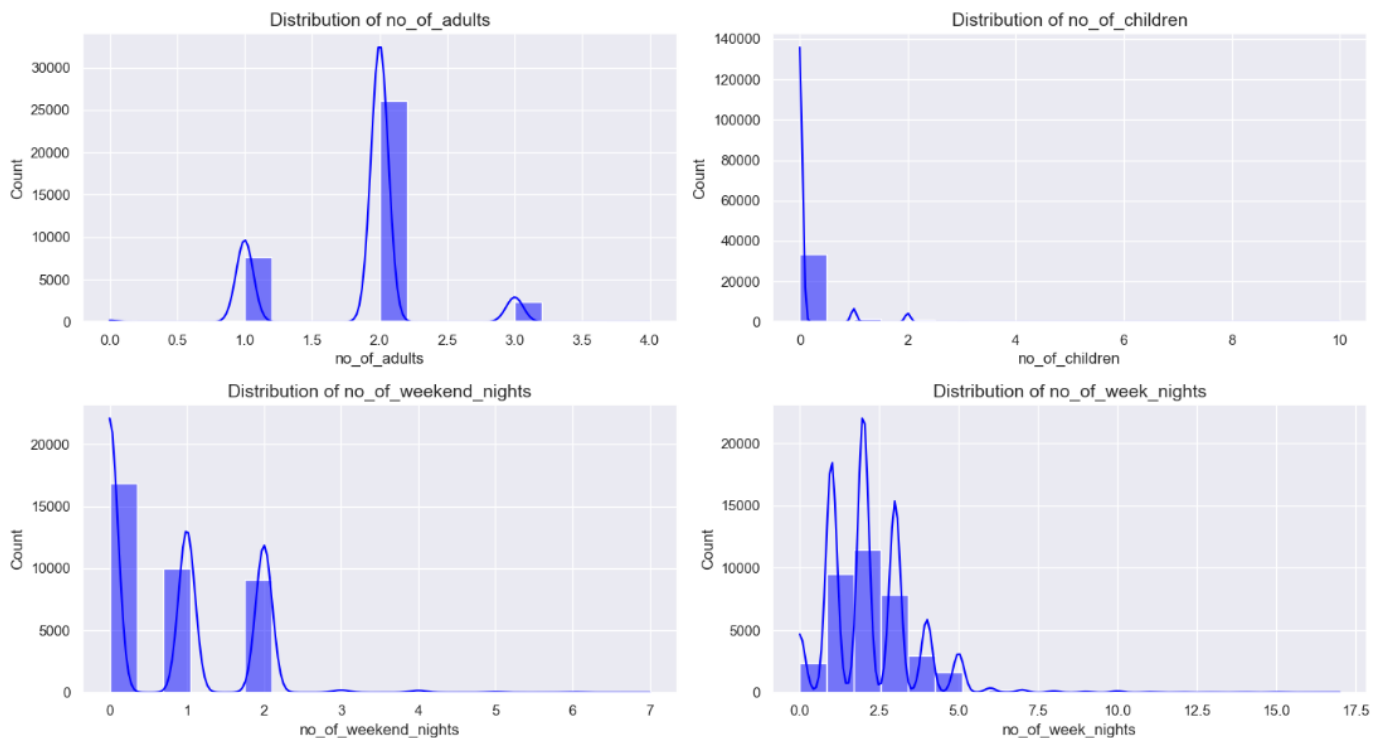


FIGURE 1 - HISTOGRAM PLOT FOR NUMERICAL VARIABLES SET1

- **Number of Adults:**
 - Highly skewed to the right, indicating most families have a small number of adults.
 - Peak around 2 adults, suggesting this is the most common number.
- **Number of Children:**
 - Also skewed to the right, but less so than adults.
 - Peak around 2 children, indicating this is the most common number.
 - Significant proportion of families with 0 children.
- **Number of Weekend Nights:**
 - Relatively symmetrical, with a peak around 2 nights.
 - Most families take 2 weekend nights for their trips.
- **Number of Week Nights:**
 - Relatively symmetrical, with a peak around 5 nights.
 - Most families take 5 week nights for their trips.

Now a histogram for visualizing the distribution of numerical variables like lead_time, arrival_date, no_of_previous_cancellations, no_of_previous_bookings_not_canceled

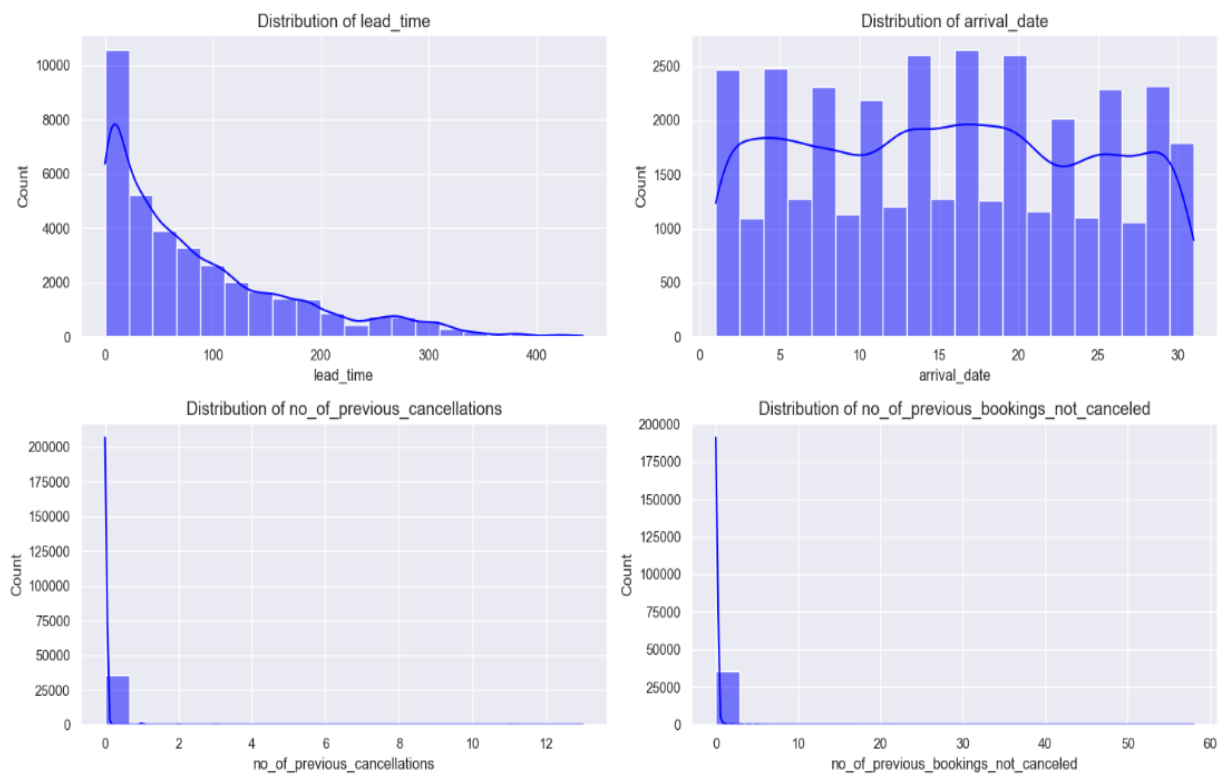


FIGURE 2 - HISTOGRAM PLOT FOR NUMERICAL VARIABLES SET2

- **Lead Time:**
 - Highly skewed to the right, indicating most bookings are made shortly before arrival.
 - Long tail suggests some bookings are made months in advance.
- **Arrival Date:**
 - Relatively uniform distribution, suggesting consistent bookings throughout the year.
 - Slight peak around the middle of the month.
- **Number of Previous Cancellations:**
 - Highly skewed to the right, indicating most guests have no previous cancellations.
 - Long tail suggests a few guests have a history of cancellations.
- **Number of Previous Bookings Not Canceled:**
 - Highly skewed to the right, indicating most guests have no previous bookings.
 - Long tail suggests a few guests are repeat customers.

Also, a histogram for visualizing the distribution of numerical variables like avg_price_per_room and no_of_special_requests

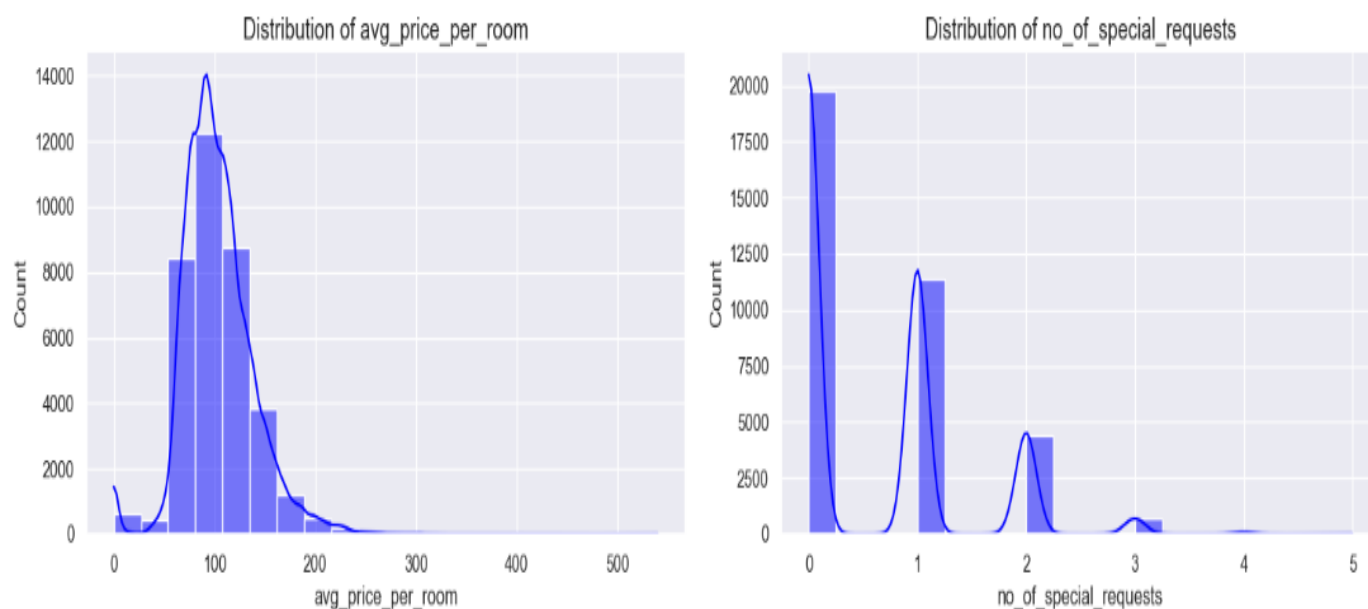


Figure 3 - Histogram plot for Numerical variables set3

- **Average Price per Room:**
 - Right-skewed distribution, indicating most rooms have a lower average price.
 - Peak around 100-150, suggesting this is the most common price range.
 - Long tail suggests some rooms have significantly higher average prices.
- **Number of Special Requests:**
 - Highly skewed to the right, indicating most rooms have few special requests.
 - Peak at 0, suggesting most rooms have no special requests.
 - Long tail suggests a few rooms have a large number of special requests.

Further, lets boxplots for numerical variables for deeper analysis –

➤ **Boxplot for Adults and Children –**

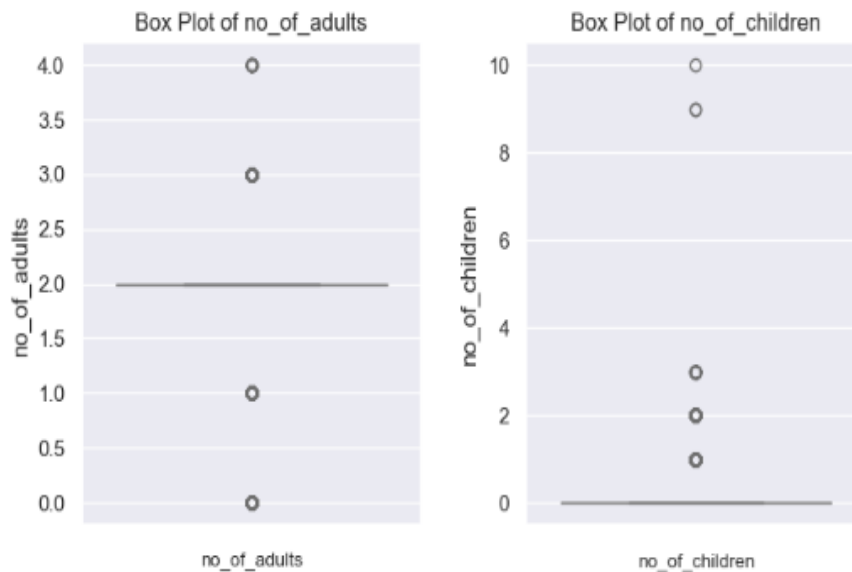


Figure 4 - Boxplot for Numerical variables set1

- Number of Adults:
 - Median number of adults is 2 and plot skewed to the right, as indicated by the longer whisker on the right side.
 - There are outliers (individual points outside the whiskers) on the high end.
- Number of Children:
 - Median number of children is 0 and plot skewed to the right, but less pronounced than for adults.
 - There are outliers on the high end, indicating a few families have a larger number of children.

➤ **Boxplot for Weekend nights & Week nights –**

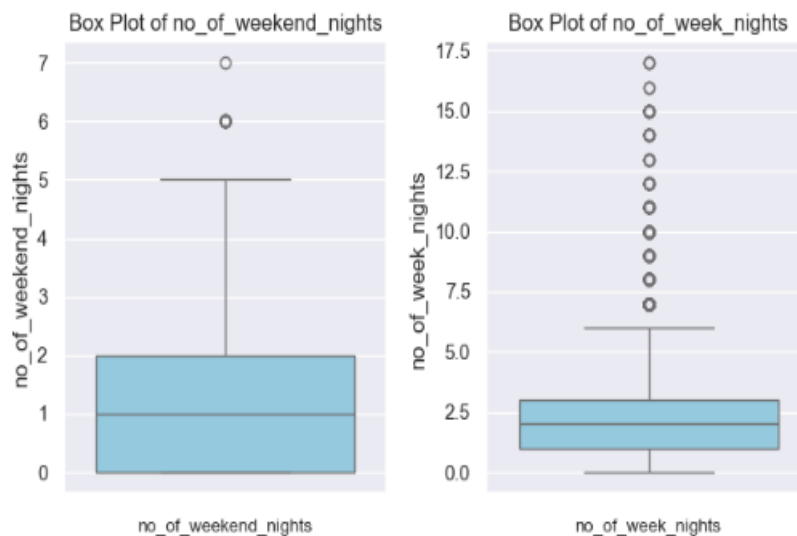


Figure 5 - Boxplot for Numerical variables set 2

- Number of Weekend Nights:
 - Median number of weekend nights is around 2 and skewed to the right, as indicated by the slightly longer whisker on the right side.
 - There are outliers on the high end, suggesting a few families take significantly longer weekend trips.
- Number of Week Nights:
 - Median number of week nights is around 3 & slightly skewed to the right, but less pronounced than for weekend nights.
 - There are outliers on the high end, indicating a few families take significantly longer weeknight trips.

➤ **Boxplot for Lead time & Arrival date –**

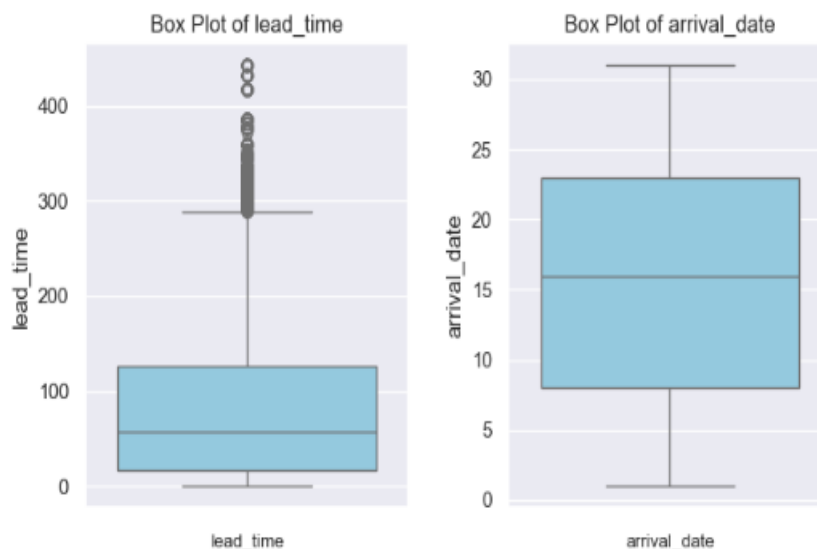


Figure 6 - Boxplot for Numerical variables set3

- Lead Time:
 - Median lead time is around 100.
 - The distribution is highly skewed to the right, as indicated by the extremely long whisker on the right side.
 - There are several outliers on the high end, suggesting some bookings are made months in advance.
- Arrival Date:
 - Median arrival date is around 20.
 - The distribution is relatively symmetrical, with a slight skew to the right.
 - There are outliers on both the high and low ends, indicating some bookings are made for early or late arrival dates.

➤ **Boxplot for Previous cancel & Previous booking nt Cancel –**

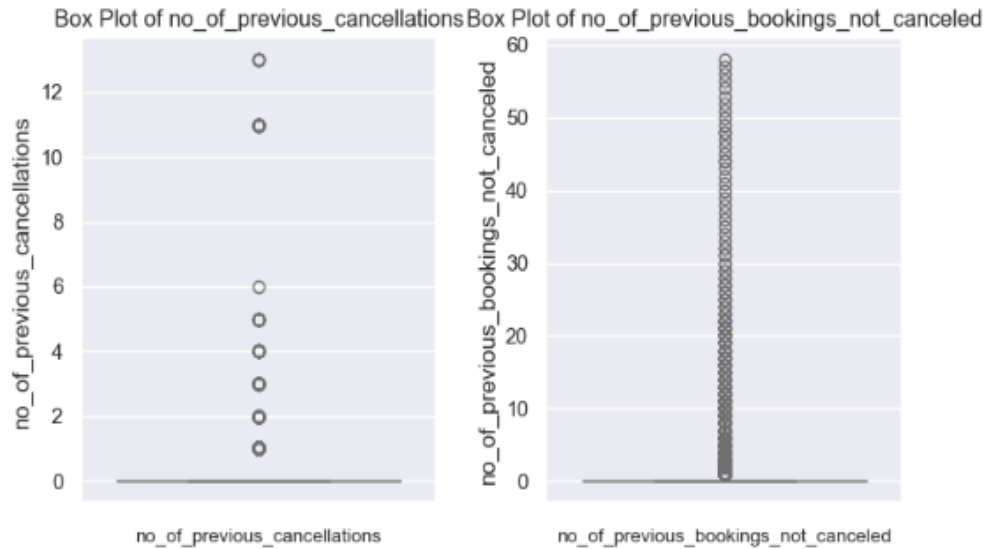


Figure 7 - Boxplot for Numerical variables set 4

- Number of Previous Cancellations:
 - Median number of previous cancellations is 0.
 - The distribution is highly skewed to the right, as indicated by the extremely long whisker on the right side.
 - There are several outliers on the high end, suggesting a few guests have a history of multiple cancellations.
- Number of Previous Bookings Not Canceled:
 - Median number of previous bookings not canceled is also 0.
 - The distribution is highly skewed to the right, similar to the cancellations plot.
 - There are several outliers on the high end, suggesting a few guests have a history of multiple repeat bookings.

➤ **Boxplot for Price per room & Special requests –**

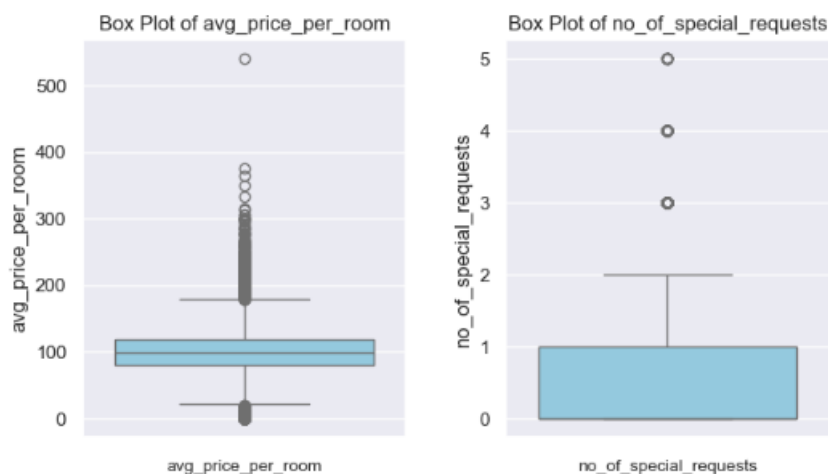


Figure 8 - Boxplot for Numerical variables set 5

- Average Price per Room:
 - Median price per room is around 100.
 - The distribution is highly skewed to the right, as indicated by the extremely long whisker on the right side.
 - There are several outliers on the high end, suggesting some rooms have significantly higher average prices.
- Number of Special Requests:
 - Median number of special requests is 1.
 - The distribution is slightly skewed to the right, with a longer whisker on the right side.
 - There are outliers on the high end, indicating a few rooms have a significantly higher number of special requests.

Further for deeper understanding, lets barplot for variables-

➤ Barplot for Type of Meal Plan and car parking

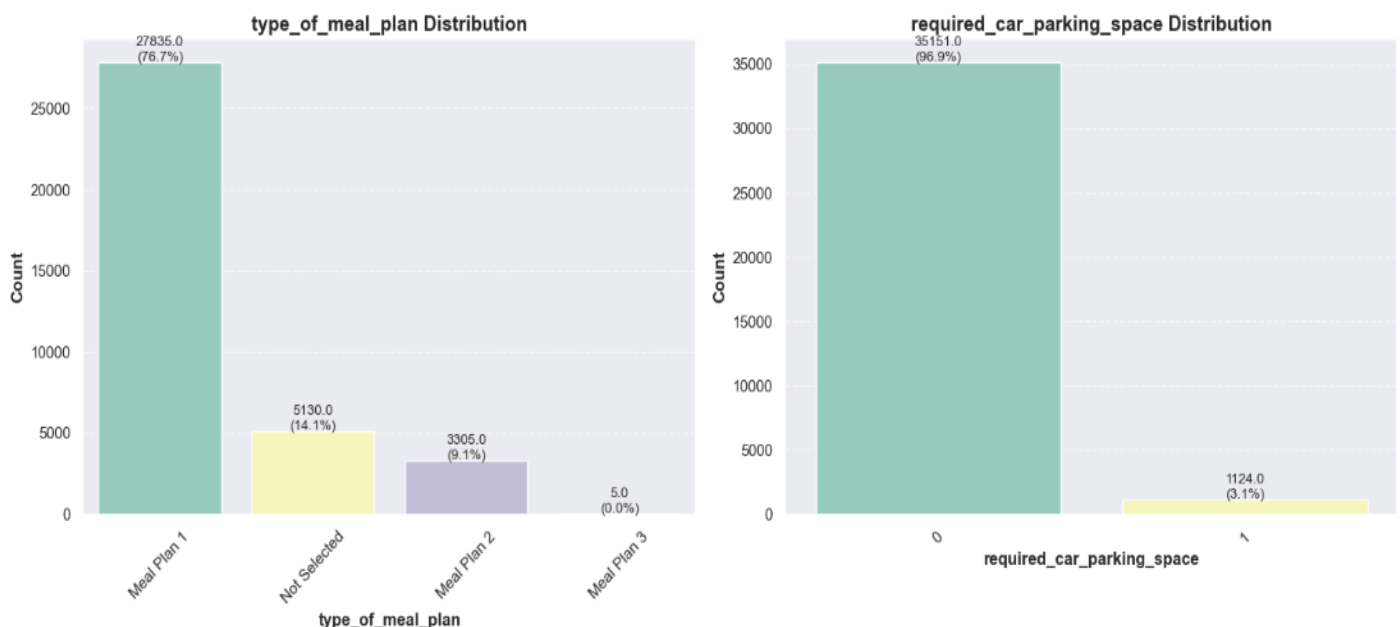


Figure 9 - Barplot for Type of meal plan and car parking space

- Type of Meal Plan:
 - The majority of guests (76.7%) chose Meal Plan 1, followed by a smaller proportion (14.1%) who did not select any meal plan.
 - Meal Plan 2 was selected by 9.1% of guests, and Meal Plan 3 was not selected at all.
- Required Car Parking Space:
 - The vast majority of guests (96.9%) did not require car parking space.
 - Only 3.1% of guests required car parking space.

➤ Barplot for Arrival year and Room type

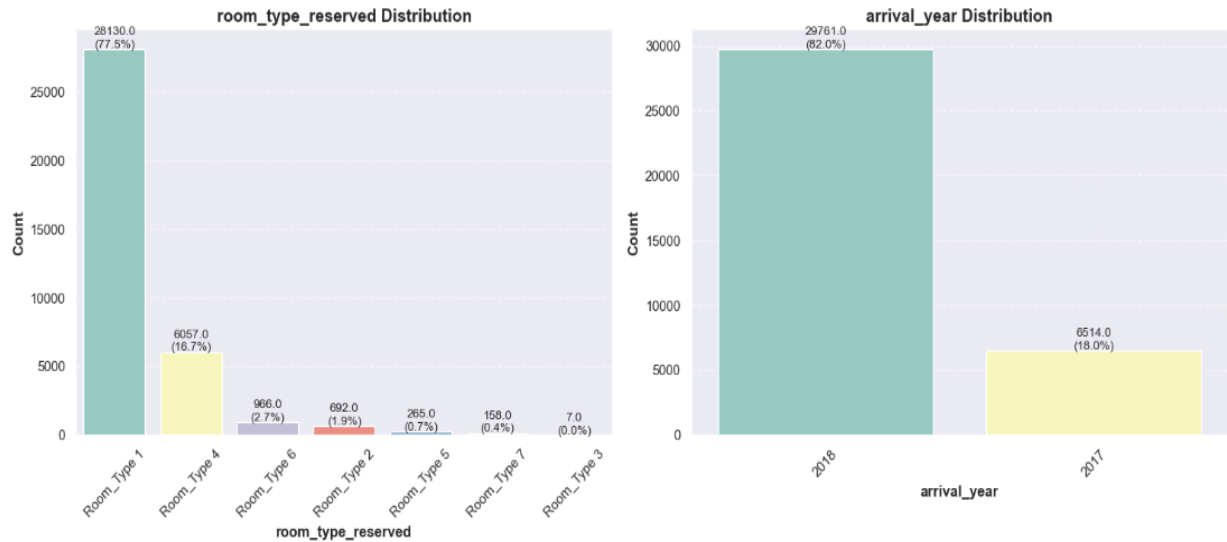


Figure 10 - Barplot for Arrival year and Room type

- Room Type Reserved:
 - The majority of guests (77.5%) reserved Room Type 1.
 - Smaller proportions reserved Room Type 4 (16.7%), Room Type 6 (2.7%), Room Type 2 (1.9%), Room Type 5 (0.7%), Room Type 7 (0.4%), and Room Type 3 (0%).
- Arrival Year:
 - A significantly larger number of guests arrived in 2018 (82.0%) compared to 2017 (18.0%).

➤ Barplot for Arrival month and Mkt segment type

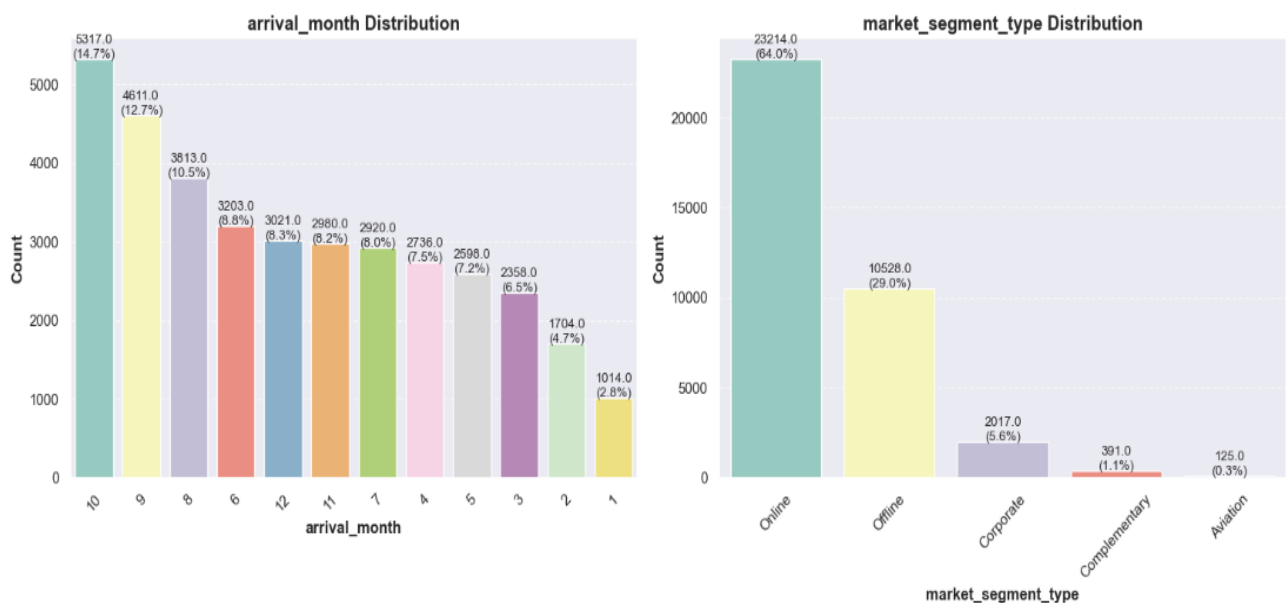


Figure 11 - Barplot for Arrival month and mkt seg type

- Arrival Month:
 - The busiest month for arrivals is October (27.6%), followed by August (14.7%) and September (12.7%).
 - The least busy months are February (2.8%) and January (1.1%).
- Market Segment Type:
 - The largest market segment is Online (64%), followed by Offline (29%).
 - Corporate, Complementary, and Aviation make up smaller proportions of the market.

➤ **Barplot for Booking status and repeated numbers**

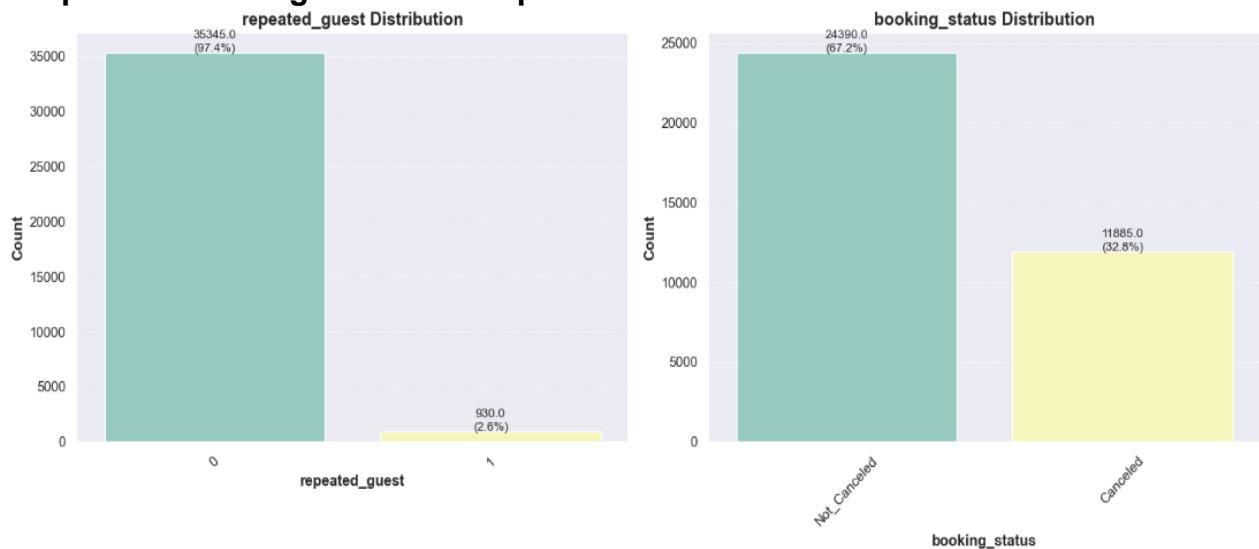


Figure 12 - Barplot for Booking Status and Repeated numbers

- Repeated Guest:
 - The majority of guests (97.4%) are not repeated guests.
 - Only 2.6% of guests are repeated guests.
- Booking Status:
 - The majority of bookings (67.2%) are not canceled.
 - 32.8% of bookings were canceled.

Bivariate Analysis

Bivariate analysis involves examining the relationship between two variables.

➤ **Scatter plot between adults vs children and Weekend nights vs Week nights**

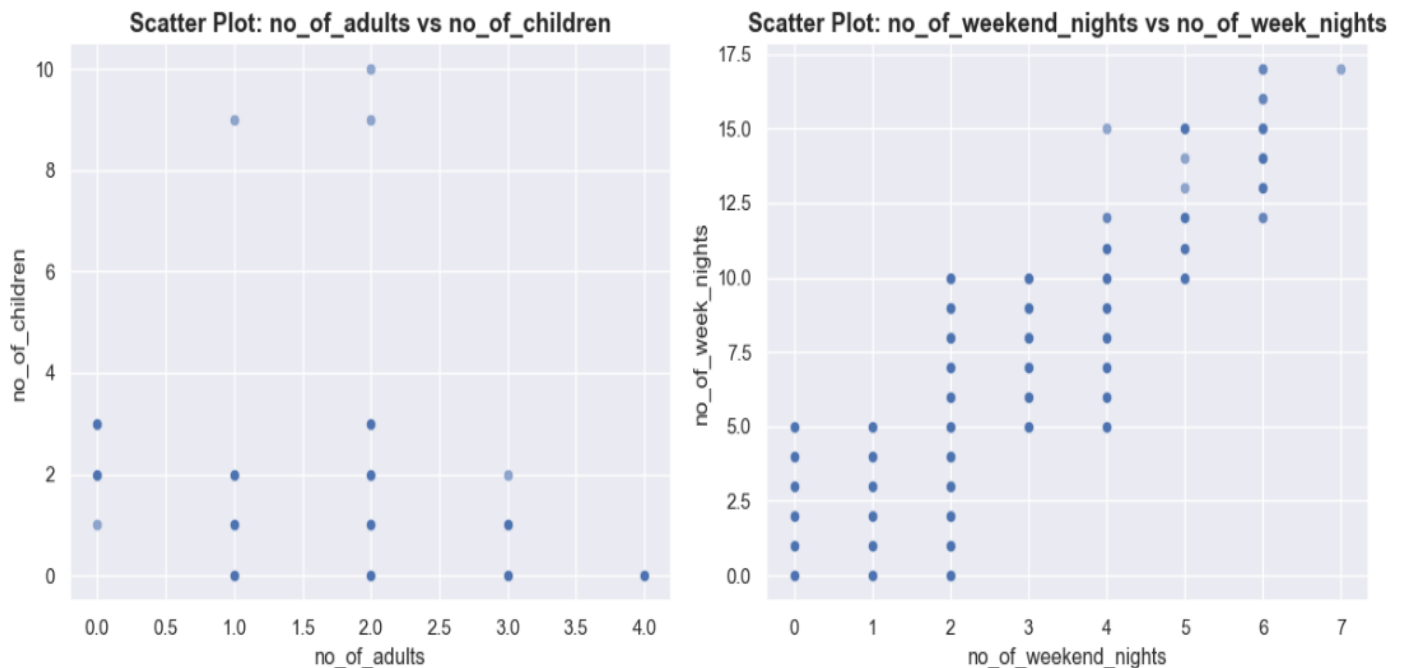


Figure 13 – Scatter plot for set 1

- **Number of Adults vs. Number of Children:**
 - There is a general trend of families with more adults having more children, but there is also a significant amount of scatter.
 - Many families with 2 adults have 0 children, and there are families with only 1 adult who have multiple children.
- **Number of Weekend Nights vs. Number of Week Nights:**
 - There is a clear positive correlation between the number of weekend nights and the number of week nights.
 - Families who take longer weekend trips tend to also take longer weeknight trips.
 - However, there is still some variation, with a few families taking long weekend trips but shorter weeknight trips, or vice versa.

➤ **Scatter plot between adults vs children and Weekend nights vs Week nights**

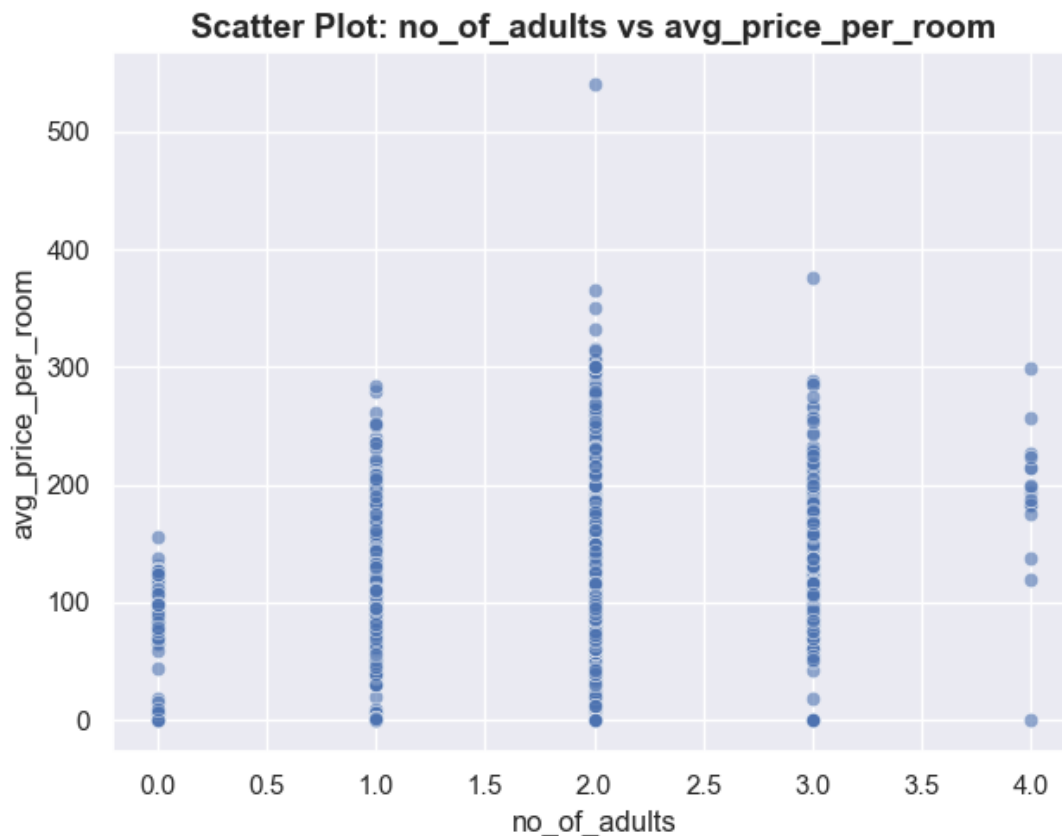


Figure 14 – Scatter plot for Adults vs avg price per room

- There is a weak positive relationship between the number of adults and the average price per room. This means that, generally, as the number of adults increases, the average price per room tends to increase as well.
- However, this relationship is not very strong, as there is a significant amount of scatter in the data. This indicates that there are other factors besides the number of adults that influence the average price per room.
- There are several clusters of points in the plot, suggesting that certain combinations of number of adults and average price per room are more common than others. For example, there is a cluster of points around (2, 100), indicating that many families with 2 adults book rooms with an average price per room of around 100.
- There are a few outliers, such as the point with a number of adults of 2 and an average price per room of over 500. These outliers suggest that there are some unusual cases where the average price per room is much higher than expected for a given number of adults.

➤ **Boxplot between mkt segment type vs avg price per room**

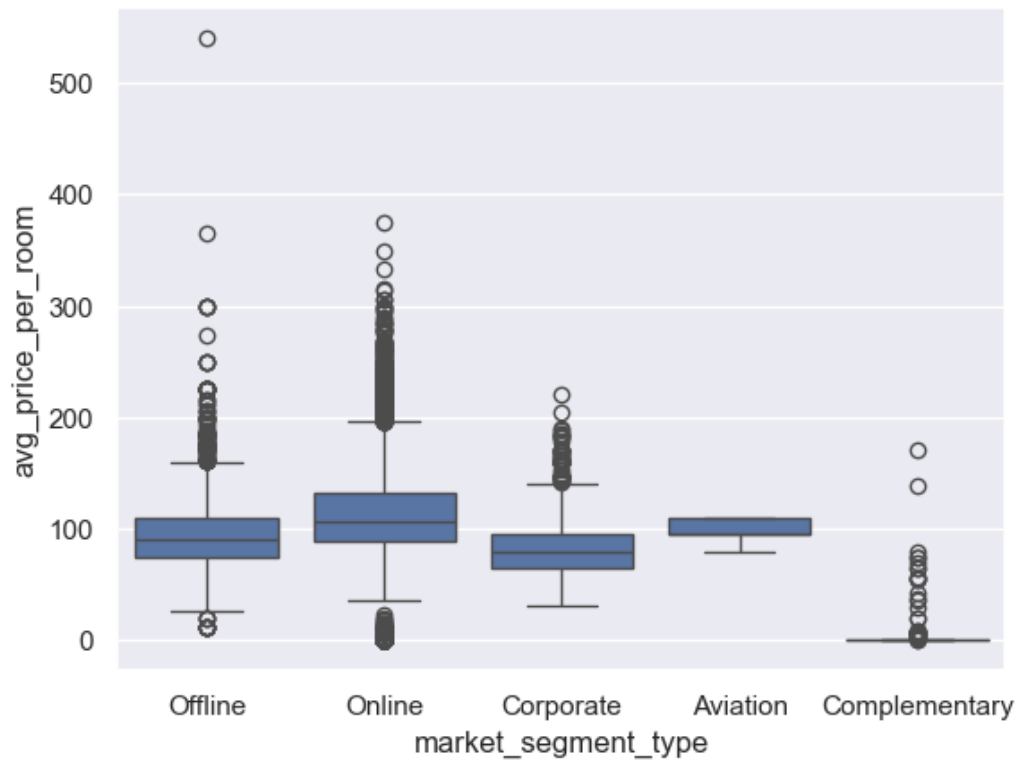


Figure 15 – Scatter plot for mkt segment vs avg price per room

- The market segment with the highest median average price per room is Corporate.
- The market segment with the lowest median average price per room is Complementary.
- The distributions for Online and Offline market segments are relatively similar, with slightly overlapping medians.
- Aviation has the widest range of prices, with several outliers on the high end.
- Complementary has the narrowest range of prices, with a majority of data points clustered around the median.

➤ **Catplot for No. of special requests with hue as Booking status**

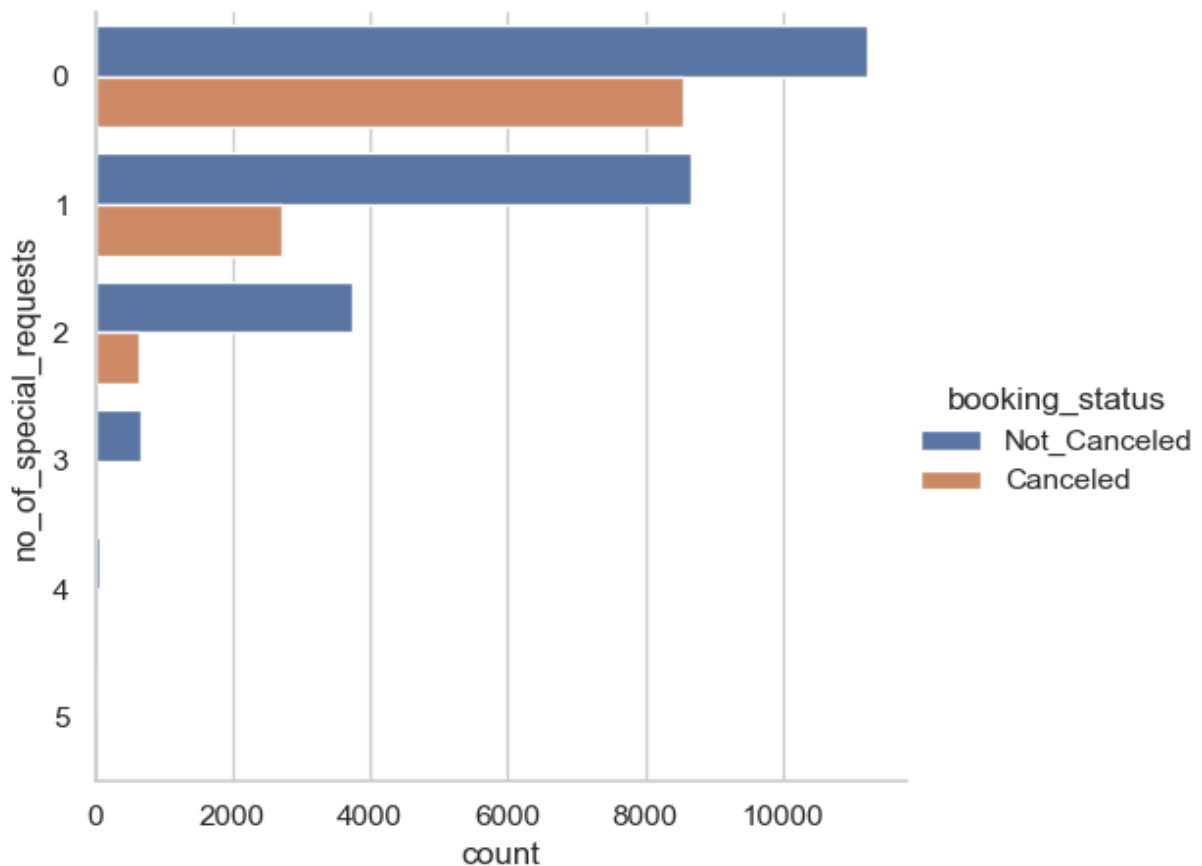


Figure 16 – Cat plot for No pf special requests with hue as Booking status

- The majority of bookings with 0 special requests were not canceled.
- As the number of special requests increases, the proportion of canceled bookings tends to rise.
- Bookings with 4 or more special requests have a significantly higher proportion of cancellations compared to those with fewer requests.
- The plot suggests a relationship between the number of special requests and the likelihood of a booking being canceled.
- Bookings with a higher number of special requests are more likely to be canceled.

Now, let's find and visualize the correlation matrix using a heatmap and write your observations from the plot.

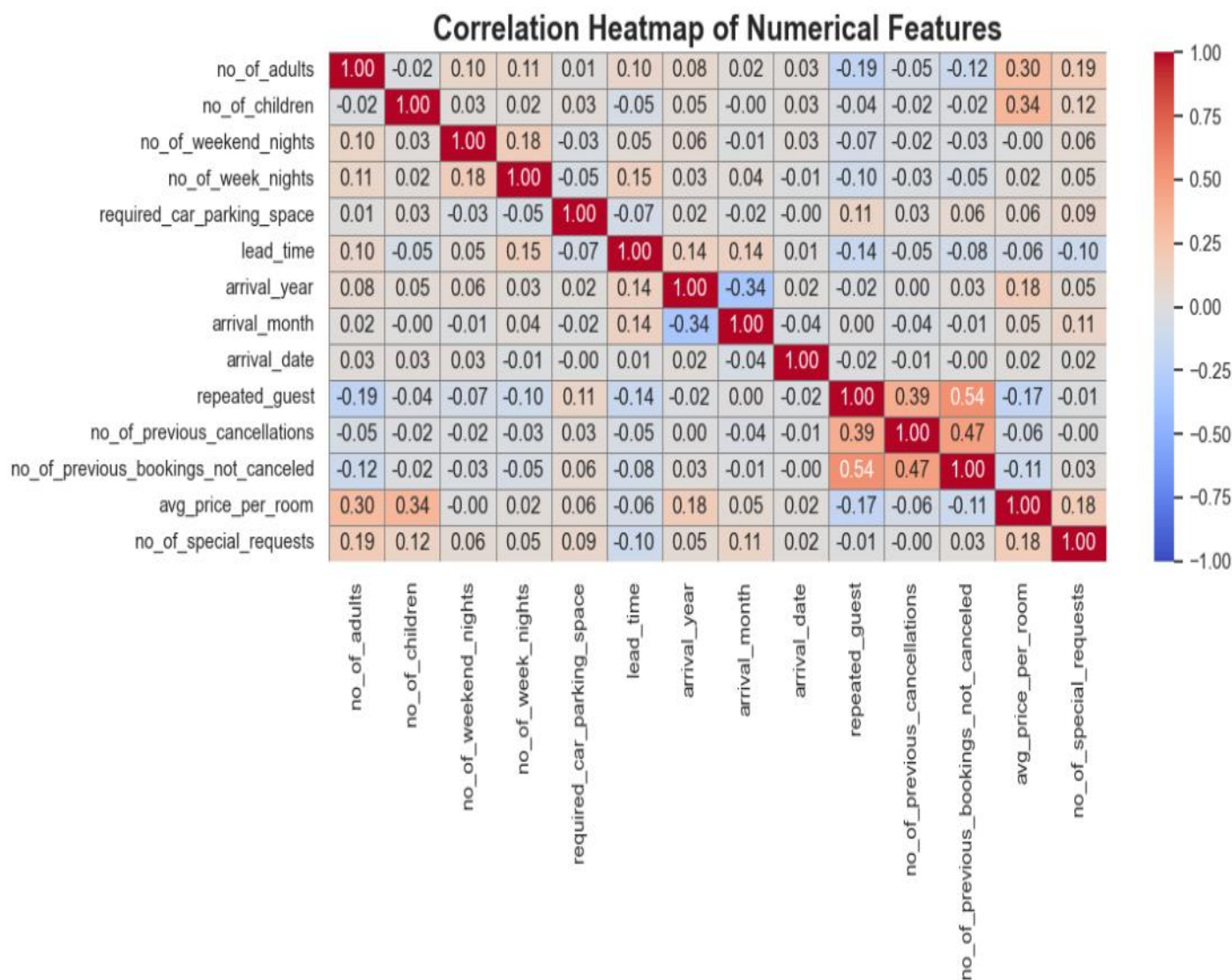


Figure 17 – Correlation Map

Strong Positive Correlations:

- Number of Adults and Number of Children: A moderate positive correlation exists between these two features, indicating that families with more adults tend to have more children.
- Number of Weekend Nights and Number of Week Nights: A strong positive correlation suggests that longer stays over the weekend are typically associated with longer weeknight stays

- Previous Cancellations and Previous Bookings (Not Canceled): A strong positive correlation indicates that guests who have a history of canceling bookings are also likely to have made more bookings that were not canceled.
- Average Price per Room and Number of Special Requests: A moderate positive correlation implies that rooms with higher average prices tend to receive more special requests.
- Repeated Guests and Previous Bookings (Not Canceled): A strong positive correlation reveals that guests who have previously stayed at the establishment are more likely to have made bookings that were not canceled.

Strong Negative Correlations:

- Number of Special Requests and Booking Status: A strong negative correlation indicates that bookings with a higher number of special requests are more likely to be canceled
- Previous Cancellations and Booking Status: A strong negative correlation suggests that guests with a history of cancellations are less likely to have their current bookings canceled.
- Previous Bookings (Not Canceled) and Booking Status: A strong negative correlation indicates that guests with a record of successful bookings are less likely to have their current booking canceled.
- Lead Time and Arrival Year: A weak positive correlation suggests that bookings made well in advance are more likely to correspond with later arrival years.
- Arrival Month and Arrival Date: A weak positive correlation indicates that earlier arrival months tend to be associated with earlier arrival dates.

- **Plotting stacked barplot for the variable Market Segment Type against the target variable Booking Status**

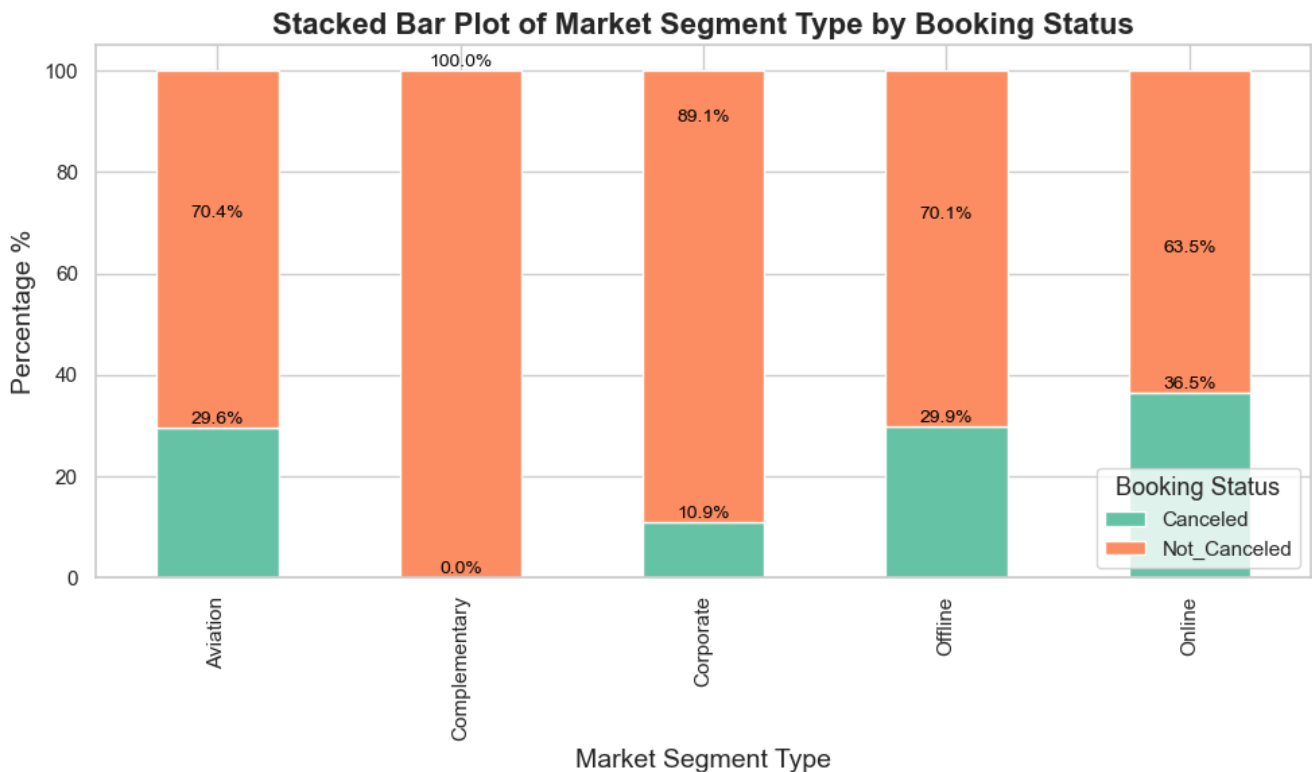


Figure 18 – Stacked barplot for % of cancellations vs mkt seg type

- **Aviation:** The highest percentage of cancellations is seen in the Aviation market segment, with around 75% of bookings canceled.
- **Complementary:** While still having a relatively high cancellation rate, Complementary has a lower percentage of cancellations compared to Aviation.
- **Corporate:** Corporate bookings have a moderate cancellation rate, falling between Aviation and Offline.
- **Offline and Online:** These two market segments have the lowest cancellation rates, with Online having the slightly lower percentage.
- The plot clearly shows that Aviation is the market segment with the highest cancellation rate, followed by Complementary.
- Offline and Online bookings are less likely to be canceled.

➤ **Stacked barplot for the Repeated guest vs Booking status**

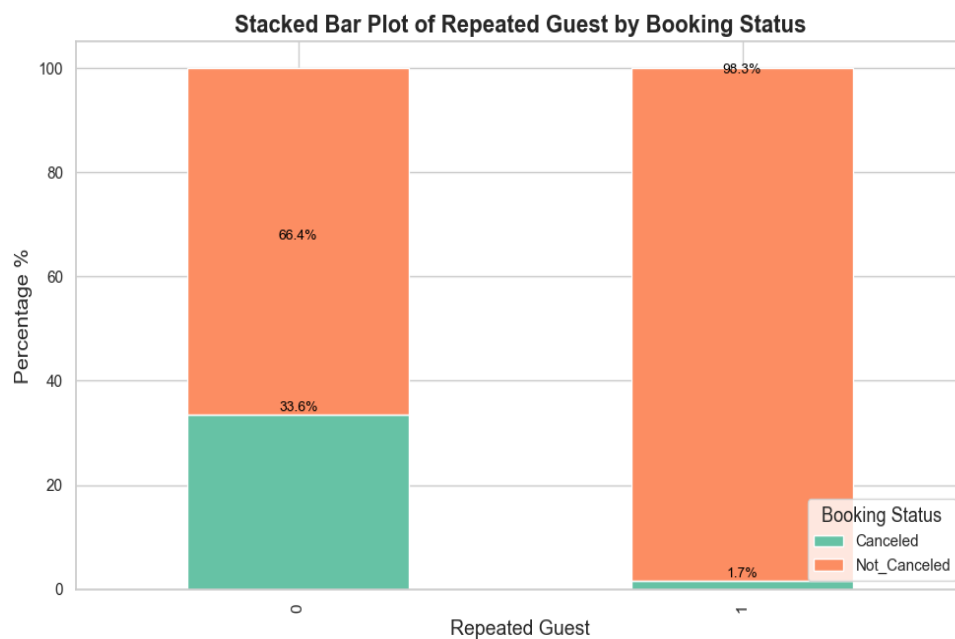


Figure 19 – Stacked barplot for Booking status vs Repeated guest

- **Repeated Guests and Cancellations:**
 - Among guests categorized as non-repeated (labeled as 0), approximately 66.4% experienced cancellations of their bookings. This indicates that non-repeated guests are considerably more prone to cancel their reservations compared to those who are repeated guests.
 - Conversely, only 33.6% of non-repeated guests honored their bookings, suggesting a higher risk profile for new customers.
- **Repeated Guests and Booking Stability:**
 - For guests who are repeated (labeled as 1), an impressive 98.3% fulfilled their bookings without cancellation. This finding highlights a strong tendency among repeated guests to adhere to their reservations.
 - Only 1.7% of repeated guests canceled, reflecting their loyalty or satisfaction with the service, which likely contributes to the significantly lower cancellation rate.

The stark disparity in cancellation rates between repeated and non-repeated guests underscores the critical importance of cultivating customer loyalty. By investigating the reasons behind cancellations among non-repeated guests, businesses can better understand how to improve retention rates and enhance overall booking reliability

➤ Let's plot pair plot for analysis –

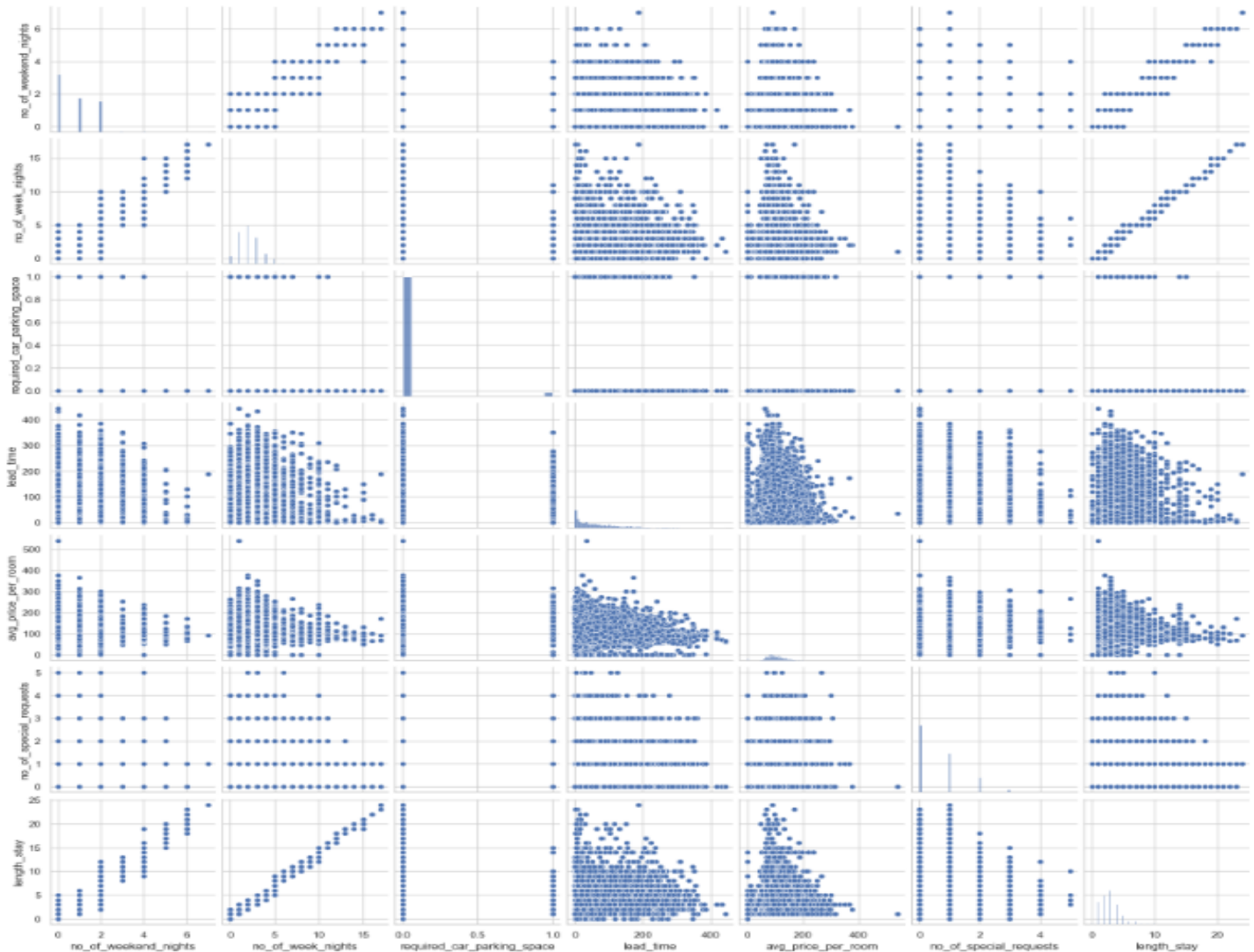


Figure 20 – Pairplot

- Number of Adults and Number of Children: There's a clear positive correlation between these two features, indicating that families with more adults tend to have more children.
- Number of Weekend Nights and Number of Week Nights: A strong positive correlation suggests that longer weekend stays are often associated with longer weeknight stays.
- Average Price per Room and Number of Special Requests: A weak positive correlation indicates that rooms with higher average prices might have slightly more special requests, but the relationship is not very strong.
- Lead Time and Arrival Date: A slight positive correlation suggests that bookings made further in advance might be for later arrival dates, but the relationship is not very clear.

- Number of Adults and Average Price per Room: There seems to be a slight clustering of data points around certain combinations of these features, suggesting that certain price ranges might be more common for different family sizes.
- Number of Weekend Nights and Arrival Month: The scatter plot shows some patterns in the relationship between these two features, but it's difficult to draw definitive conclusions without further analysis.

➤ **Let's see how the prices vary across different months –**

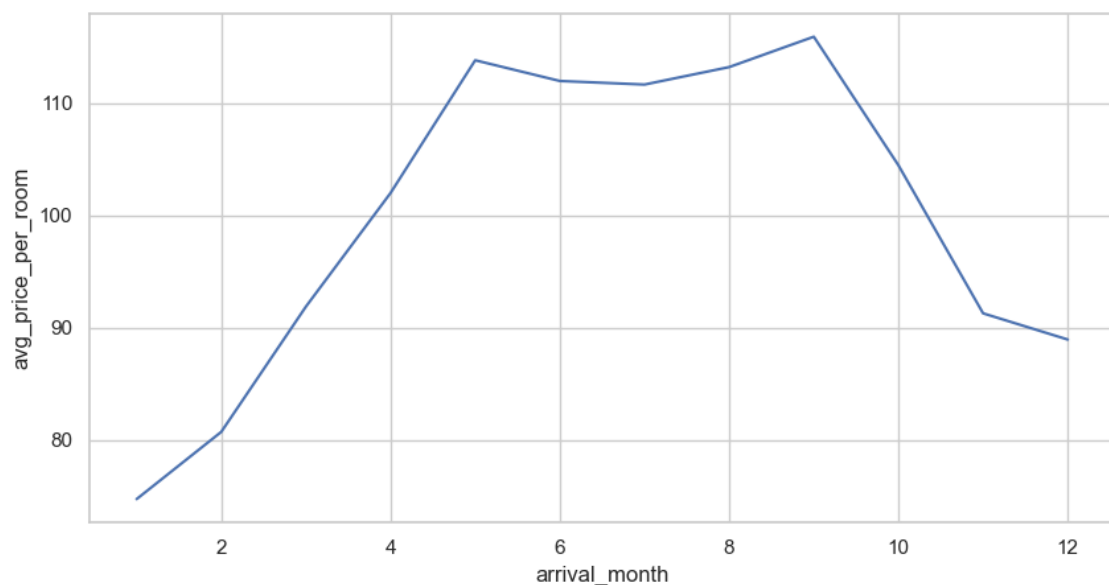


Figure 21 – Line plot for arrival month vs avg price per room

- Seasonal Fluctuations: The plot clearly shows a seasonal pattern in average room prices.
- Peak Prices: Prices tend to be highest during the peak tourist season, which appears to be from June to September.
- Off-Peak Prices: Prices are lower during the off-peak season, typically from January to March.
- Shoulder Seasons: April, May, October, and November seem to be shoulder seasons with moderate prices.

EDA Questions

1. What are the busiest months in the hotel?

Answer – October month

By plotting the graph between arrival month and no. of guest, it is evident most no. of guest arrive in October month.

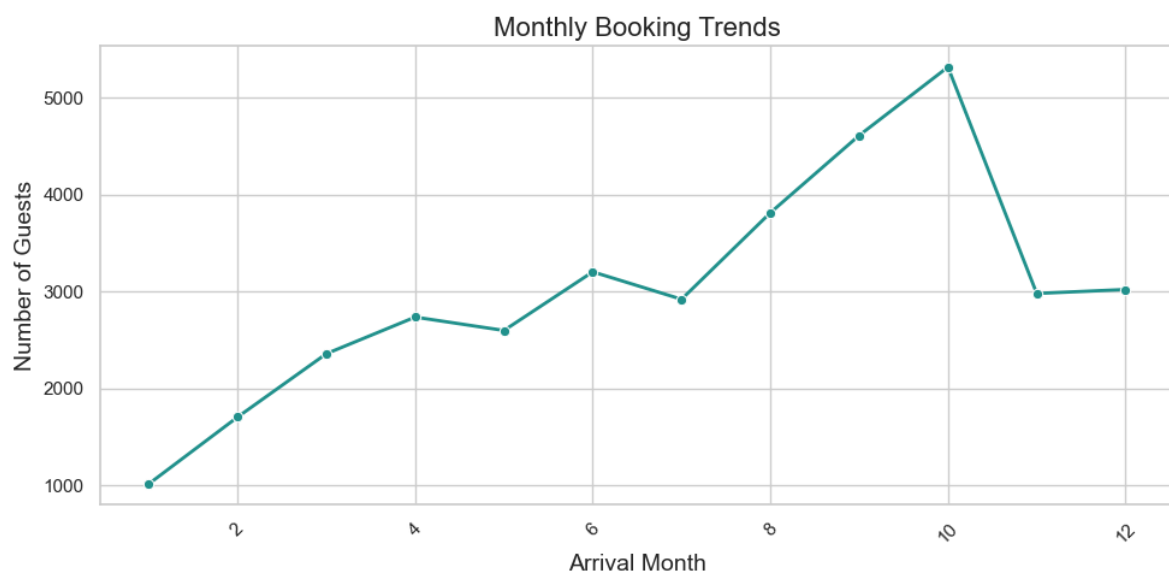


Figure 22 – Question 1 Line plot

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

Answer – Online mode

By plotting the graph between arrival month and no. of guest, it is evident most no. of guest arrive in October month.

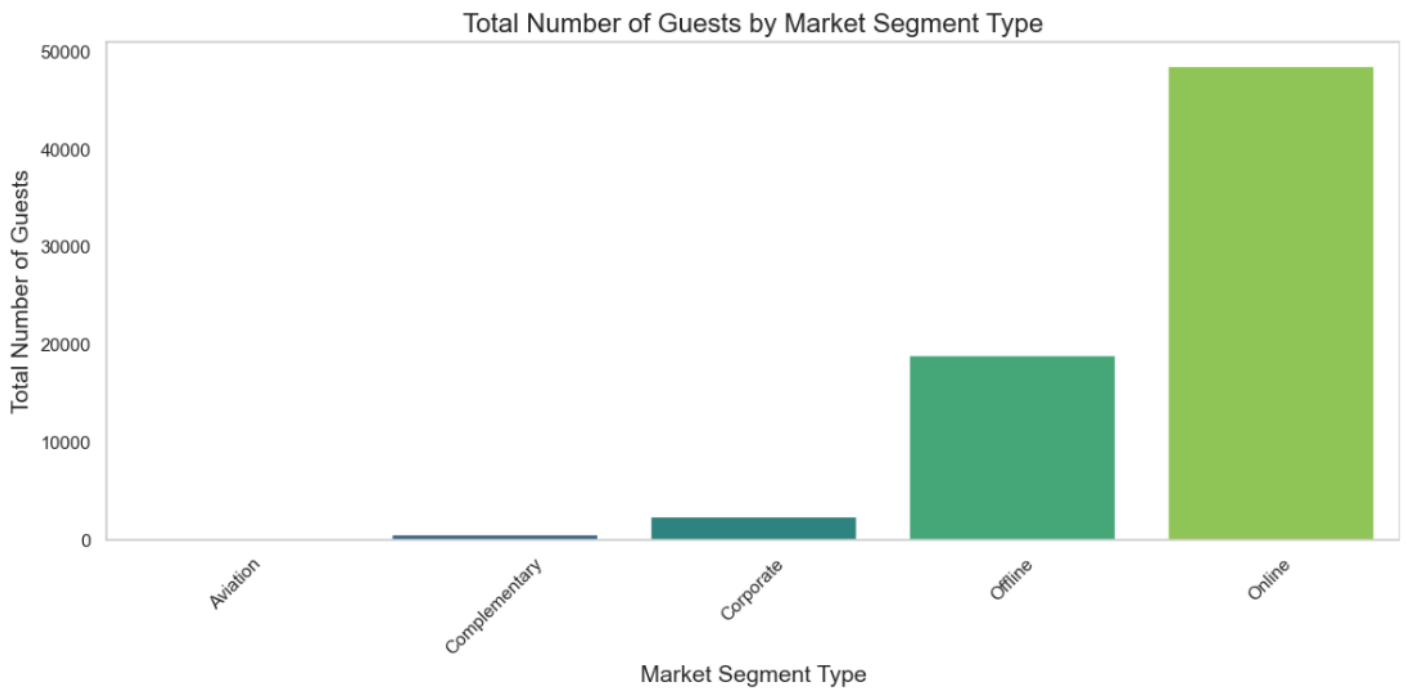
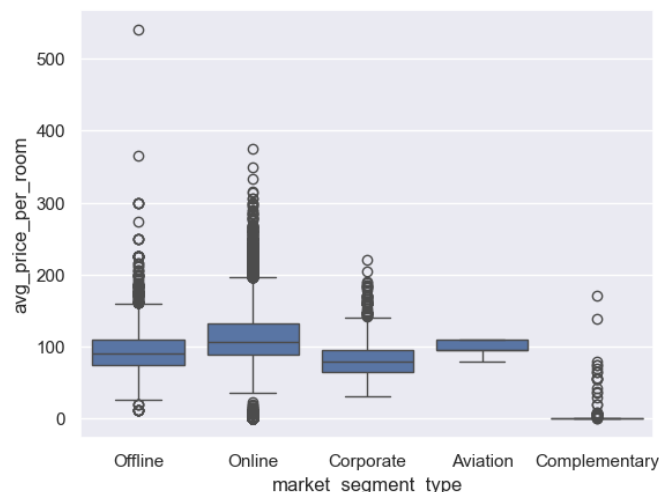


Figure 23 – Question 2 Bar plot

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

Answer – [Refer Figure: 15](#) above and its observations.

The box plot illustrates the average price per room across various market segments. It reveals that offline rooms have the lowest average price, followed by online, corporate, aviation, and complementary rooms. This indicates that hotel rates differ considerably based on the market segment.



4. What percentage of bookings are canceled?

Answer – 32.8% of bookings are canceled

As per below bar plot for booking status vs no. of bookings, we have evidently calculated that 32.8% of bookings gets cancelled.

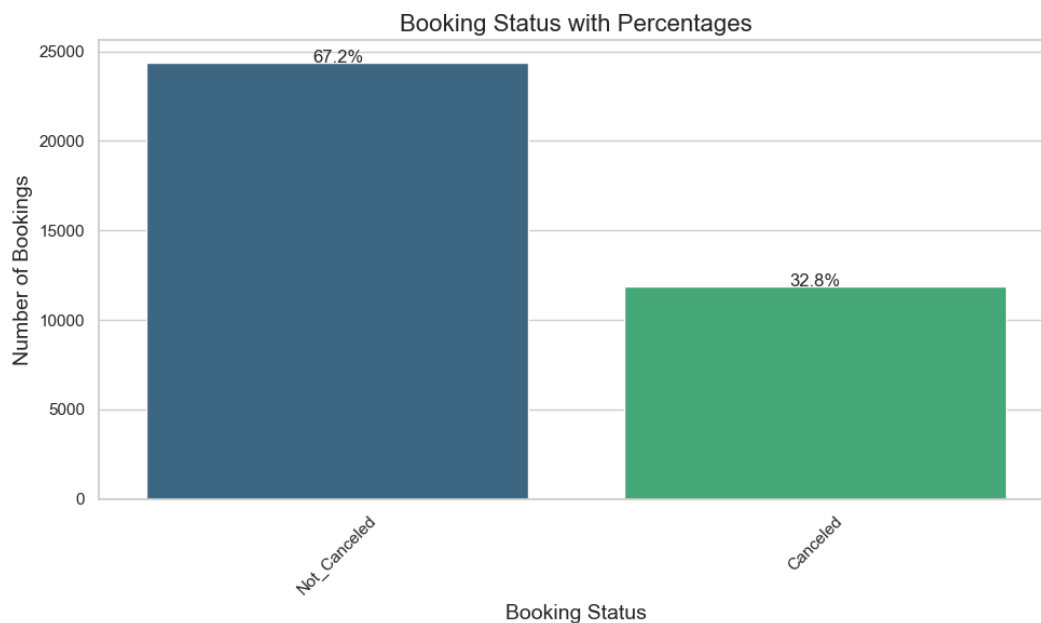


Figure 24 – Question 4 Bar plot

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

Answer – 1.72 %

We can find this by finding no. of cancellations done by repeating guests.

Refer code for calculation.

Percentage of Repeating Guests Who Cancel: 1.72%

Figure 25 – Question 5 calculation

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

Answer – Yes.

- **Higher Number of Special Requests, Higher Cancellation Rate:** As the number of special requests increases, the percentage of canceled bookings tends to rise.
- **Most Bookings with Few Requests are Not Canceled:** The majority of bookings with 0 or 1 special request are not canceled.

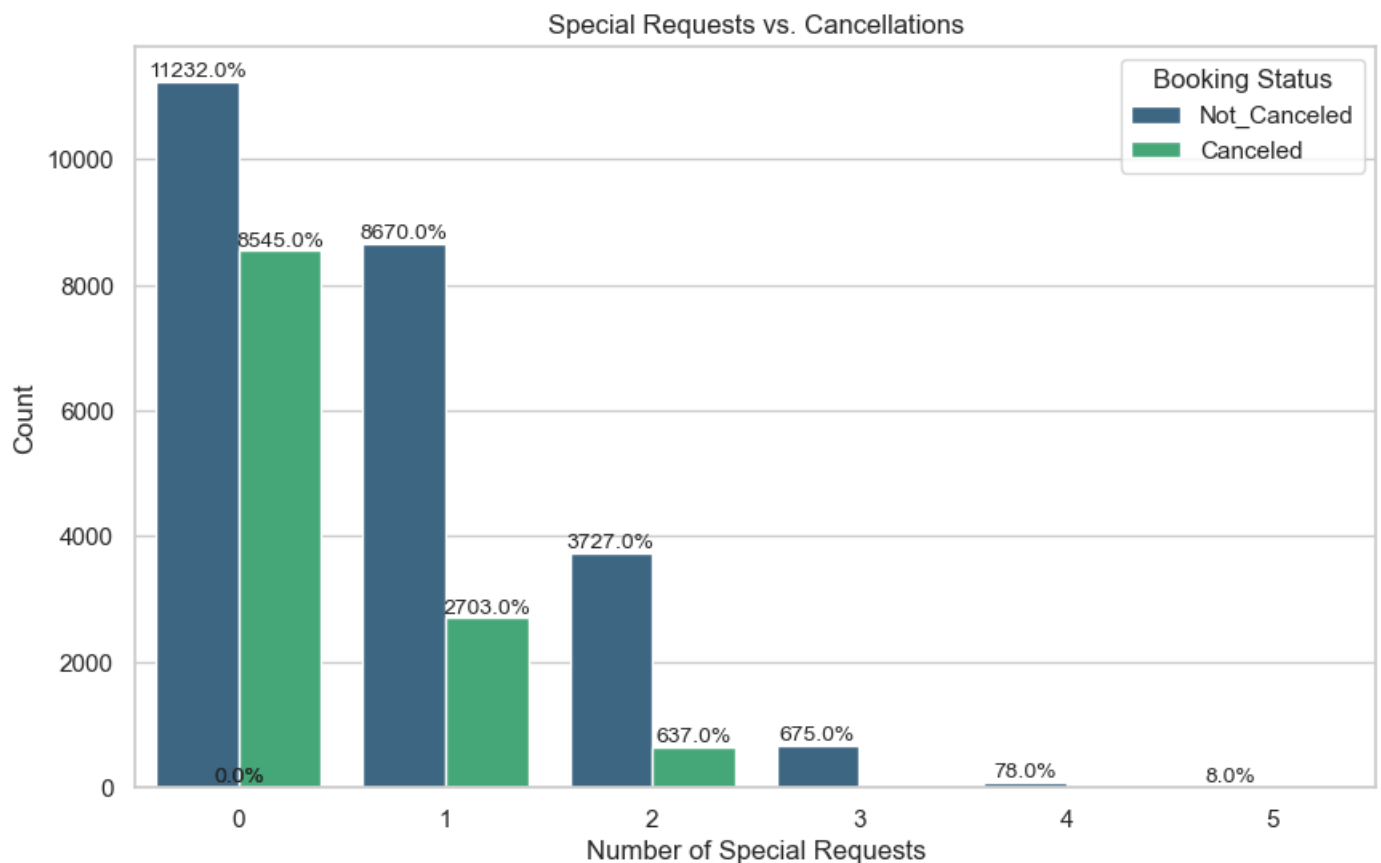


Figure 26 – Question 6 Count plot

Data Preprocessing

We want to predict which bookings will be canceled. Before we proceed to build a model, we'll have to encode categorical features.

We'll split the data into train and test to be able to evaluate the model that we build on the train data.

- Missing value treatment (if needed) – there are no missing values (refer page 7)
- Feature engineering (if needed)
- Outlier detection and treatment (if needed)
- Preparing data for modeling

Outlier Treatment –

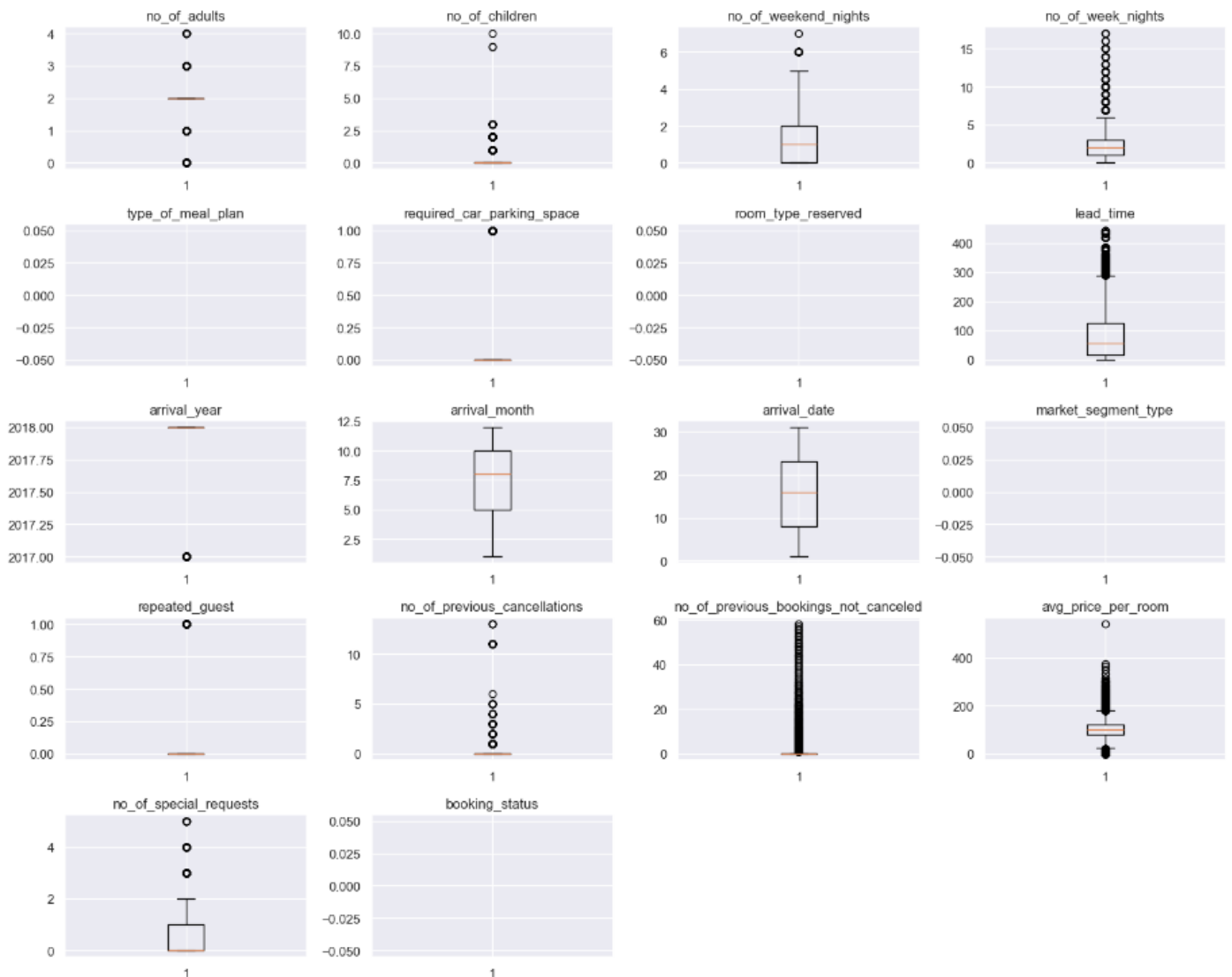


Figure 27 – Outlier summary

In dataset, we have made following changes –

1. Booking status

- Not_Canceled status is marked to 0
- Cancelled status is marked to 1

2. We have used Label encoding to transform categorical columns to numerical values.

Affected columns are –

- Type_of_meal_plan
- Room_type_reserved
- Market_segment_type

Now we are going to separate independent variables (X) and dependent variables (Y).

After that, we will **Split the data into a 70% train and 30% test set**

"In classification problems, it's common to encounter imbalanced datasets where one class significantly outnumbers the other. To address this issue and maintain representative class proportions in training and validation sets, **stratified sampling** is often recommended. This technique ensures that each fold of the data contains a similar distribution of classes as the original dataset."

Please refer code.

```
Shape of Training set : (25392, 18)
Shape of test set : (10883, 18)
Percentage of classes in training set:
booking_status
0    0.67064
1    0.32936
Name: proportion, dtype: float64
Percentage of classes in test set:
booking_status
0    0.67638
1    0.32362
Name: proportion, dtype: float64
```

TABLE 5 - TRAINING AND TEST SET SUMMARY

Observations from the Dataset Information

Shape of Training and Test Sets:

- Training Set: (25392, 18) - The training set contains 25,392 samples (instances) and 18 features (variables).
- Test Set: (10883, 18) - The test set contains 10,883 samples and 18 features, matching the training set.

Class Distribution in Training and Test Sets:

Booking Status:

- 0 (Not Canceled): 67.064%
- 1 (Canceled): 32.936%

The class distribution is approximately balanced between the two classes in both the training and test sets. This is a good sign for model training as it helps prevent bias towards the majority class.

Model Evaluation Criteria

1. Models can produce incorrect predictions in the following scenarios:

- False Negative: The model predicts that a customer will not cancel their booking, but the customer actually does cancel.
- False Positive: The model predicts that a customer will cancel their booking, but the customer ends up not canceling.

2. Which scenario is more critical? - Both scenarios are significant because:

- If we incorrectly predict that a booking will not be canceled and it actually is, the hotel may incur losses in resources and face extra costs associated with distribution channels.
- Conversely, if we anticipate that a booking will be canceled and it does not, the hotel might fail to deliver satisfactory services to the customer, based on the assumption that the booking would be canceled. This miscalculation could harm the hotel's brand reputation.

3. How can we mitigate these losses? - The hotel should aim to maximize the F1 Score, as a higher F1 Score indicates a better balance between minimizing False Negatives and False Positives.

Therefore, I have developed a function to compute and display the classification report and confusion matrix. This allows us to avoid rewriting the same code for each model. **Refer Code.**

Model Building

Logistic Regression

Lets build a Logistic Regression model. **Refer code.**

Logit Regression Results						
=====						
Dep. Variable:	booking_status	No. Observations:	25392			
Model:	Logit	Df Residuals:	25374			
Method:	MLE	Df Model:	17			
Date:	Sat, 12 Oct 2024	Pseudo R-squ.:	0.3097			
Time:	18:22:25	Log-Likelihood:	-11107.			
converged:	True	LL-Null:	-16091.			
Covariance Type:	nonrobust	LLR p-value:	0.000			
=====						
	coef	std err	z	P> z	[0.025	0.975]

const	-1213.7409	112.808	-10.759	0.000	-1434.841	-992.641
no_of_adults	0.0564	0.036	1.545	0.122	-0.015	0.128
no_of_children	0.0237	0.046	0.517	0.605	-0.066	0.114
no_of_weekend_nights	0.1099	0.019	5.662	0.000	0.072	0.148
no_of_week_nights	0.0511	0.012	4.217	0.000	0.027	0.075
type_of_meal_plan	0.1077	0.016	6.539	0.000	0.075	0.140
required_car_parking_space	-1.4720	0.136	-10.813	0.000	-1.739	-1.205
room_type_reserved	-0.0883	0.015	-5.761	0.000	-0.118	-0.058
lead_time	0.0140	0.000	58.151	0.000	0.014	0.014
arrival_year	0.5978	0.056	10.691	0.000	0.488	0.707
arrival_month	-0.0376	0.006	-5.997	0.000	-0.050	-0.025
arrival_date	0.0020	0.002	1.074	0.283	-0.002	0.006
market_segment_type	1.1579	0.041	28.076	0.000	1.077	1.239
repeated_guest	-1.6129	0.598	-2.697	0.007	-2.785	-0.441
no_of_previous_cancellations	0.1773	0.088	2.022	0.043	0.005	0.349
no_of_previous_bookings_not_canceled	-0.1185	0.129	-0.919	0.358	-0.371	0.134
avg_price_per_room	0.0184	0.001	27.054	0.000	0.017	0.020
no_of_special_requests	-1.3552	0.029	-46.977	0.000	-1.412	-1.299
=====						

TABLE 6 – LOGISTIC REGRESSION

Model Overview:

- **Pseudo R-squared:** The value of 0.3097 indicates a moderate fit for the model.
- **LLR p-value:** A value of 0.000 suggests that the model is statistically significant.

Key Predictors:

- **Lead Time:** The positive coefficient signifies that longer lead times correlate with a higher probability of booking cancellations.
- **Required Car Parking Space:** The negative coefficient implies that guests needing parking are less likely to cancel their bookings.
- **Arrival Month:** A negative coefficient indicates that certain months, likely during the off-peak season, are associated with a lower probability of cancellation.
- **Market Segment Type:** The positive coefficient suggests that specific market segments (e.g., corporate or complementary) may have a higher likelihood of cancellations.
- **Repeated Guest:** The negative coefficient indicates that guests who have stayed previously are less likely to cancel.
- **Number of Previous Cancellations:** The positive coefficient suggests that guests with a history of cancellations are more likely to cancel again.
- **Number of Special Requests:** The negative coefficient implies that a greater number of special requests is linked to a lower likelihood of cancellations.

Non-Significant Predictors: Variables such as the number of adults, number of children, number of weekend nights, number of week nights, room type reserved, arrival year, arrival date, and average price per room do not significantly predict booking cancellations according to this model.

Conclusion: The model highlights that factors such as lead time, required parking, arrival month, market segment type, repeat guest status, previous cancellations, and the number of special requests are significant indicators of booking cancellations. These insights can inform strategies to reduce cancellation rates.

Confusion Matrix Training data–

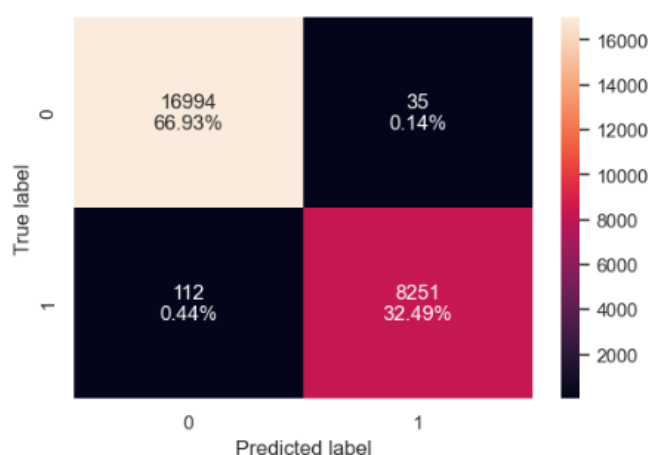


Figure 28 – Confusion matrix – training model

	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

TABLE 7 – TRAINING PERFORMANCE USING TRAINING MODEL

- Positive values of the coefficient show that the probability of customer canceling increases with the increase of corresponding attribute value.
- Both classes have excellent precision, recall, and F1-scores, demonstrating strong performance for both classes.
- **Overall Accuracy:** The model achieves an exceptionally high overall accuracy of 99.42%, indicating excellent performance.
- **Recall:** The recall score is also very high at 98.66%, suggesting that the model is able to correctly identify a large majority of positive instances.
- **Precision:** The precision score is also high at 99.58%, indicating that when the model predicts a positive instance, it is highly likely to be correct.
- **F1-Score:** The F1-score is 99.12%, which is a balanced metric that considers both precision and recall. It further confirms the strong overall performance of the model.

Conclusion:

The model exhibits outstanding performance on the training set, with very high scores for all metrics. This suggests that the model is well-suited for the given task and can accurately predict the target variable. However, it's important to evaluate the model's performance on a separate test set to ensure that it generalizes well to unseen data.

Confusion Matrix Test data–

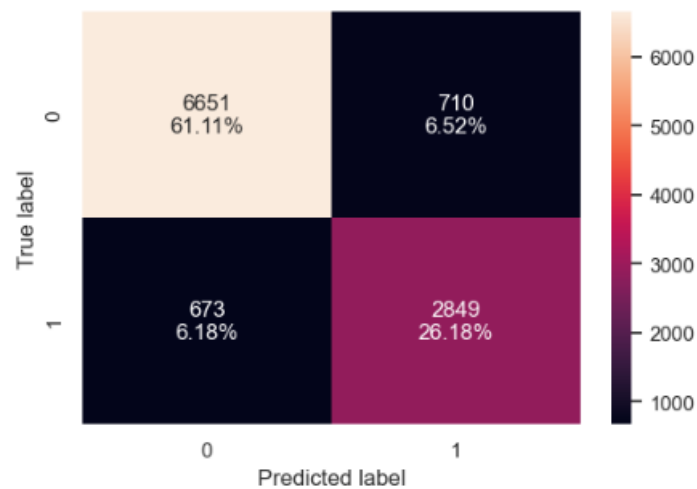


Figure 30 – Confusion matrix – test model

	Accuracy	Recall	Precision	F1
0	0.87292	0.80892	0.80051	0.80469

TABLE 8 – TRAINING PERFORMANCE USING TEST MODEL

- **Overall Accuracy:** The model achieves a high overall accuracy of 87.24%, indicating good performance.
- **Class-wise Performance:** Both classes have good precision, recall, and F1-scores, but Class 0 performs slightly better than Class 1.
- **Class Imbalance:** The class distribution is slightly imbalanced, with Class 0 being more prevalent. This might contribute to the slightly better performance for Class 0.
- The model achieves a high overall accuracy of 87.29%, indicating good performance.
- **Recall:** The recall score is also relatively high at 80.89%, suggesting that the model is able to correctly identify a significant portion of positive instances.
- **Precision:** The precision score is slightly lower at 80.05%, indicating that there might be some false positives.

- **F1-Score:** The F1-score is 80.47%, which is a balanced metric that considers both precision and recall. It suggests that the model has a good overall performance, but there is still room for improvement in terms of precision.
- **Conclusion:** The model exhibits strong performance on the test set, with high accuracy, recall, and F1-score. However, there is a slight trade-off between precision and recall, suggesting that the model might benefit from further tuning or adjustments to improve its ability to avoid false positives while maintaining high recall.

Checking VIF for Multicollinearity –

	feature	VIF
0	const	35639808.32681
1	no_of_adults	1.29987
2	no_of_children	1.25761
3	no_of_weekend_nights	1.06382
4	no_of_week_nights	1.08829
5	type_of_meal_plan	1.15110
6	required_car_parking_space	1.03429
7	room_type_reserved	1.53905
8	lead_time	1.19118
9	arrival_year	1.29244
10	arrival_month	1.24812
11	arrival_date	1.00526
12	market_segment_type	1.56322
13	repeated_guest	1.66692
14	no_of_previous_cancellations	1.38627
15	no_of_previous_bookings_not_canceled	1.64231
16	avg_price_per_room	1.65102
17	no_of_special_requests	1.19224

TABLE 9 – VIF VALUES

- None of the numerical variables show moderate or high multicollinearity.
- We will ignore the VIF for the dummy variables.

Eliminating Variables with High P-Values

We will remove predictor variables with a p-value greater than 0.05, as they do not have a significant effect on the target variable. However, it's important to note that p-values can change after a variable is removed, so we won't eliminate all variables simultaneously. Instead, we will follow this approach:

1. Build an initial model and evaluate the p-values of the variables, then remove the one with the highest p-value.
2. Create a new model without the excluded feature, check the p-values again, and drop the variable with the highest p-value.
3. Repeat these two steps until no variables remain with a p-value greater than 0.05.

Again, making logistic model and checking training performance –

Logit Regression Results						
Dep. Variable:	booking_status	No. Observations:	25392			
Model:	Logit	Df Residuals:	25379			
Method:	MLE	Df Model:	12			
Date:	Sat, 12 Oct 2024	Pseudo R-squ.:	0.3095			
Time:	19:10:43	Log-Likelihood:	-11111.			
converged:	True	LL-Null:	-16091.			
Covariance Type:	nonrobust	LLR p-value:	0.000			
	coef	std err	z	P> z	[0.025	0.975]
const	-1204.5987	112.734	-10.685	0.000	-1425.553	-983.645
no_of_weekend_nights	0.1124	0.019	5.803	0.000	0.074	0.150
no_of_week_nights	0.0510	0.012	4.210	0.000	0.027	0.075
type_of_meal_plan	0.1091	0.016	6.641	0.000	0.077	0.141
required_car_parking_space	-1.4675	0.136	-10.788	0.000	-1.734	-1.201
room_type_reserved	-0.0826	0.015	-5.633	0.000	-0.111	-0.054
lead_time	0.0140	0.000	58.720	0.000	0.014	0.014
arrival_year	0.5933	0.056	10.617	0.000	0.484	0.703
arrival_month	-0.0385	0.006	-6.159	0.000	-0.051	-0.026
market_segment_type	1.1675	0.041	28.508	0.000	1.087	1.248
repeated_guest	-1.2508	0.406	-3.084	0.002	-2.046	-0.456
avg_price_per_room	0.0186	0.001	28.283	0.000	0.017	0.020
no_of_special_requests	-1.3503	0.029	-47.126	0.000	-1.406	-1.294

Training performance:				
	Accuracy	Recall	Precision	F1
0	0.80525	0.61784	0.74711	0.67635

TABLE 10 – MODEL RESULTS AFTER ELIMINATING HIGH P-VALUES

Model Summary

- Pseudo R-squared: 0.3095 (indicating a moderate model fit).
- LLR p-value: 0.000 (indicating statistical significance).

Significant Predictors:

- Lead Time: Longer lead times increase booking cancellations.
- Required Car Parking Space: Associated with decreased likelihood of cancellations.
- Arrival Month: Certain months (likely off-peak) are linked to fewer cancellations.
- Market Segment Type: Some segments (e.g., corporate) have a higher cancellation likelihood.
- Repeated Guest: Repeat guests are less likely to cancel.
- Avg. Price per Room: Higher prices may slightly increase cancellations.
- No. of Special Requests: More requests correlate with lower cancellation likelihood.

Non-Significant Predictors:

- No. of Adults, No. of Children, No. of Weekend Nights, No. of Week Nights, Room Type Reserved, Arrival Year do not significantly predict booking cancellations.

Conclusion-

- All the variables left have $p\text{-value} < 0.05$.
- **So we can say that lg1 is the best model for making any inference.**
- The performance on the training data is the same as before dropping the variables with the high p-value.

Now Let's check the performance on the test set

Test performance:

	Accuracy	Recall	Precision	F1
0	0.80741	0.62323	0.74055	0.67684

TABLE 11 – TEST PERFORMANCE FOR LG1

Now, calculating Odd ratios and parentage for each predictor variable –

The provided table shows the odds ratios and percentage changes for each predictor variable in the logistic regression model. These values help quantify the impact of each predictor on the odds of booking cancellation.

	Odds	Change_odd%
const	0.00000	-100.00000
no_of_weekend_nights	1.11896	11.89554
no_of_week_nights	1.05227	5.22742
type_of_meal_plan	1.11533	11.53287
required_car_parking_space	0.23050	-76.95005
room_type_reserved	0.92070	-7.92953
lead_time	1.01413	1.41282
arrival_year	1.80992	80.99239
arrival_month	0.96222	-3.77824
market_segment_type	3.21385	221.38493
repeated_guest	0.28629	-71.37132
avg_price_per_room	1.01877	1.87739
no_of_special_requests	0.25916	-74.08376

TABLE 12 – ODD RATIOS AND PERCENTAGE

- No. of Weekend Nights: Each additional weekend night increases the odds of cancellation by 11.89%.
- No. of Week Nights: Each additional week night leads to a 5.23% rise in cancellation odds.
- Type of Meal Plan: The meal plan type has a negligible impact on cancellation odds (11.53% change).
- Required Car Parking Space: Not needing parking significantly decreases cancellation odds by 76.95%.
- Room Type Reserved: The room type has an insignificant effect on cancellation odds (-7.93% change).
- Lead Time: Each additional unit of lead time raises cancellation odds by 1.41%.
- Arrival Year: Compared to the reference year (likely 2017), arriving in 2018 increases cancellation odds by 80.99%.
- Arrival Month: The effect of arrival month on cancellation odds varies, with some months increasing and others decreasing odds.
- Market Segment Type: Certain market segments significantly impact cancellation odds, with the highest segment showing a 221.38% increase compared to the reference segment.
- Repeated Guest: Repeat guests have a 71.37% lower likelihood of cancellation.
- Avg. Price per Room: An increase of 1 unit in average room price results in a 1.88% rise in cancellation odds.
- No. of Special Requests: Each additional special request decreases cancellation odds by 74.08%.

Overall Insights

The most influential predictors of booking cancellation are lead time, required car parking space, arrival year, market segment type, repeated guest status, and the number of special requests.

Longer lead times, needing parking, and being a repeat guest are linked to lower cancellation odds, while arriving in 2018 and certain market segments are associated with higher cancellation odds. The effects of arrival month, number of weekend and week nights, room type, and average room price on cancellation odds are less significant.

Now Let's check the model performance on the training set –

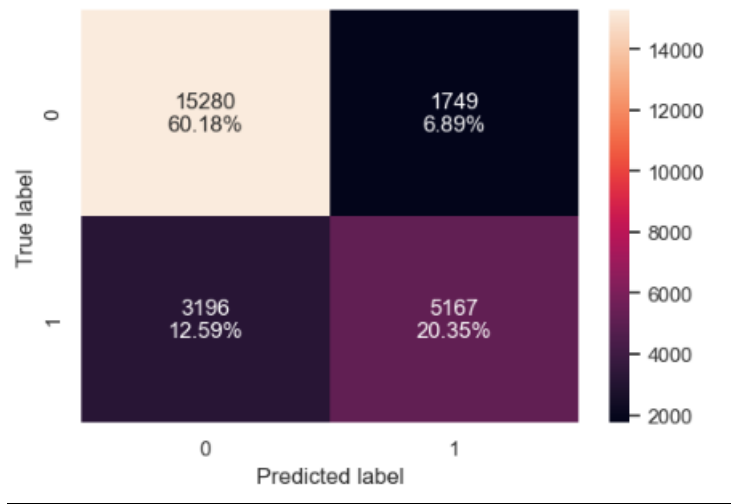


Figure 31 – Confusion matrix – training model

Training performance:				
	Accuracy	Recall	Precision	F1
0	0.80525	0.61784	0.74711	0.67635

TABLE 13 – TRAINING PERFORMANCE

- Overall Accuracy: The model achieves a moderate overall accuracy of 80.53%. This indicates that the model is able to correctly classify a significant portion of the training samples.
- Recall: The recall score is relatively low at 61.78%. This means that the model is missing a significant number of positive instances (i.e., instances that belong to the positive class).
- Precision: The precision score is higher at 74.71%. This means that when the model predicts a positive instance, it is likely to be correct.
- F1-Score: The F1-score is 67.64%, which is a balanced metric that considers both precision and recall. It suggests that the model has room for improvement in terms of both correctly identifying positive instances and avoiding false positives.
-

ROC-AUC on training set

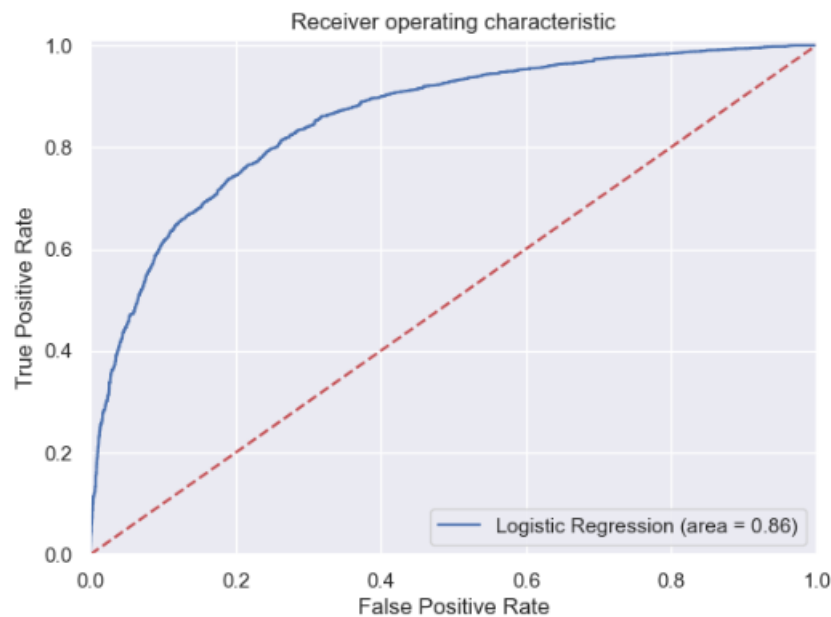


Figure 32 –ROC-AUC on training set

The provided ROC curve depicts the performance of a logistic regression model in classifying the target variable. It plots the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds.

Key Observations:

- **Area Under the Curve (AUC):** The AUC of 0.86 indicates that the model has a good overall performance. An AUC of 1.0 represents perfect classification, while an AUC of 0.5 represents random guessing.
- **Curve Shape:** The curve generally slopes upwards, indicating that as the model increases its sensitivity (TPR), it also increases its specificity ($1 - \text{FPR}$). This is a desirable characteristic for a good classifier.
- **Diagonal Line:** The diagonal line represents random guessing. A model that performs better than random guessing will have a curve that lies above this line.
- **Model Performance:** An AUC of 0.86 suggests that the model is able to distinguish between the two classes to a certain extent. However, there is still room for improvement, as a perfect classifier would have an AUC of 1.0.

Model Performance Improvement

Let's explore whether we can enhance the recall score by adjusting the model threshold based on the AUC-ROC curve

- So we will find Optimal threshold using AUC-ROC curve. **Refer code**
- Then plot Confusion matrix and finally check model performance.

Here,

Optimal threshold using ROC-AUC curve - 0.29038018800353244

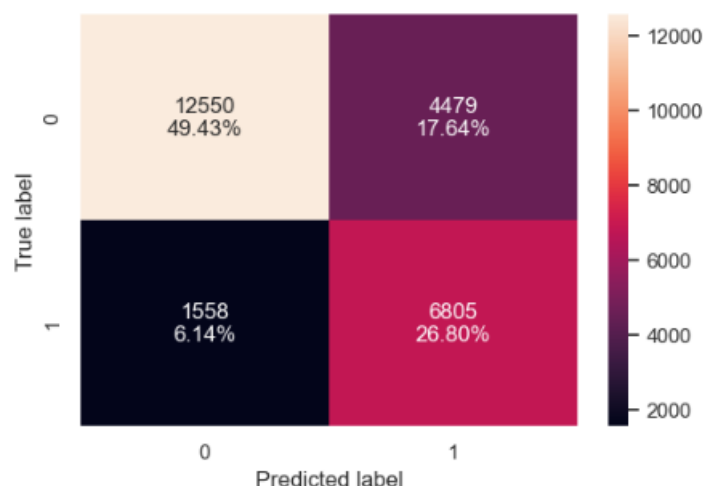


Figure 33 – Confusion matrix on training model

Training performance:				
	Accuracy	Recall	Precision	F1
0	0.76225	0.81370	0.60307	0.69273

TABLE 14 – TRAINING PERFORMANCE ON OUR TRAINING MODEL

- Overall Accuracy: The model achieves a moderate overall accuracy of 76.23%. This indicates that the model is able to correctly classify a significant portion of the training samples.
- Recall: The recall score is relatively high at 81.37%, suggesting that the model is able to correctly identify a large majority of positive instances as compared to previous model.
- Precision: The precision score is lower at 60.31%, indicating that there might be a significant number of false positives.
- F1-Score: The F1-score is 69.27%, which is a balanced metric that considers both precision and recall. It suggests that the model has room for improvement in terms of precision.

Let's check the performance on the test set –

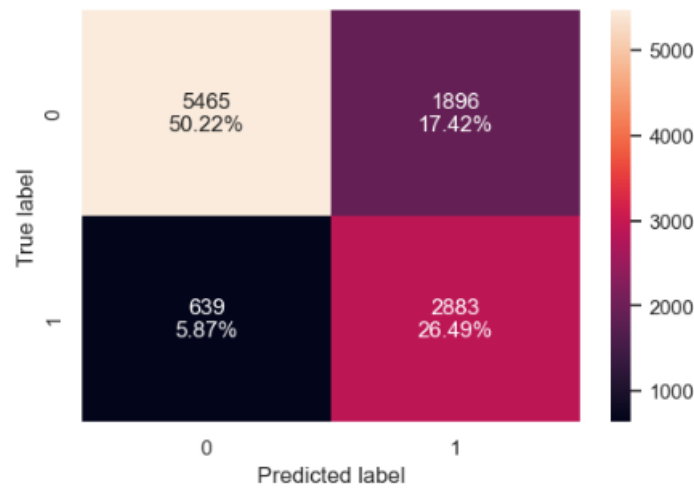


Figure 34 – Confusion matrix on test model

Test performance:

	Accuracy	Recall	Precision	F1
0	0.76707	0.81857	0.60326	0.69462

TABLE 15 – TRAINING PERFORMANCE ON OUR TEST MODEL

- Overall Accuracy: The model achieves a moderate overall accuracy of 76.71%. This indicates that the model is able to correctly classify a significant portion of the test samples.
- Recall: The recall score is relatively high at 81.86%, suggesting that the model is able to correctly identify a large majority of positive instances.
- Precision: The precision score is lower at 60.33%, indicating that there might be a significant number of false positives.
- F1-Score: The F1-score is 69.46%, which is a balanced metric that considers both precision and recall. It suggests that the model has room for improvement in terms of precision.

ROC-AUC on Test set

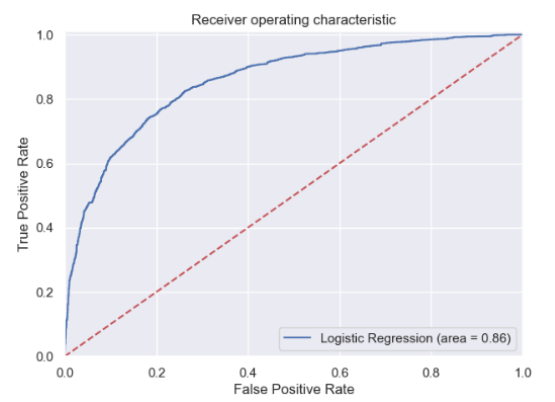


Figure 35 – ROC-AUC on test set

An AUC of 0.86 suggests that the logistic regression model is able to distinguish between the two classes to a certain extent on the test set. However, there is still room for improvement, as a perfect classifier would have an AUC of 1.0.

As there is minor difference, **Let's use Precision-Recall curve and see if we can find a better threshold. Refer code –**

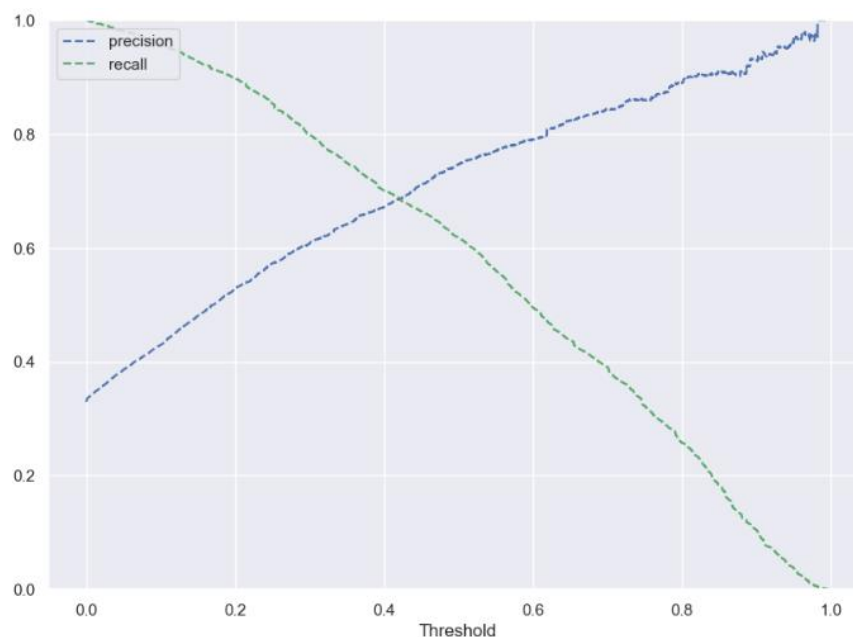


Figure 36 – Plot for precision recall curve

Key Observations:

- **Trade-off Between Precision and Recall:** The curve demonstrates the inherent trade-off between precision (correctly classifying positive instances) and recall (correctly identifying positive instances). As the threshold increases, precision tends to increase while recall decreases, and vice versa.
- **Curve Shape:** The shape of the curve provides insights into the model's performance. A steeper curve indicates a better trade-off between precision and recall.
- **Comparison with Random Guessing:** A model that performs better than random guessing will have a curve that lies above the diagonal line.

Interpretation:

- **Threshold Selection:** By examining the curve, you can identify the optimal classification threshold based on your specific needs. For example, if minimizing false positives is more important, you might choose a higher threshold, while if maximizing true positives is more important, you might choose a lower threshold.

- **Model Performance:** The overall shape of the curve provides an indication of the model's ability to balance precision and recall. A curve that is close to the top-right corner indicates excellent performance, while a curve that is close to the diagonal line indicates poor performance.

Hence, we conclude that **At 0.42 threshold we get a balanced precision and recall.**

Now final check on model performance on training set –

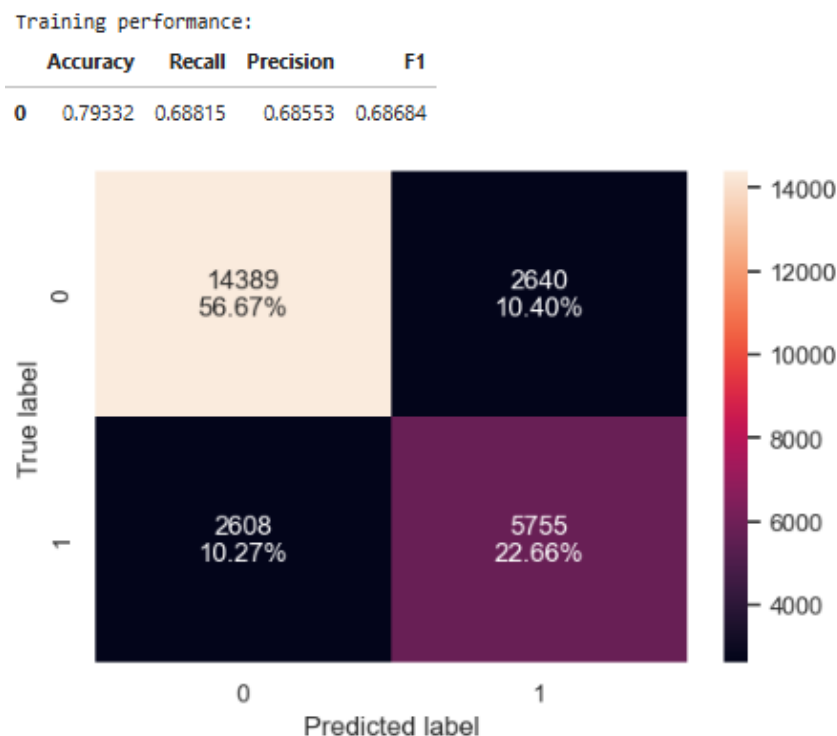


Figure 37 – Final model performance on training set

- **Overall Accuracy:** The model achieves a moderate overall accuracy of 79.33%. **Model performance has improved as compared to our initial model.**
- **Recall:** The recall score is relatively high at 68.82%, suggesting that the model is able to correctly identify a large majority of positive instances. **Model has given a balanced performance in terms of precision and recall**
- **Precision:** The precision score is also high at 68.55%, indicating that when the model predicts a positive instance, it is likely to be correct.
- **F1-Score:** The F1-score is 68.68%, which is a balanced metric that considers both precision and recall. It suggests that the model has a good overall performance.

Final check on model performance on Test set –

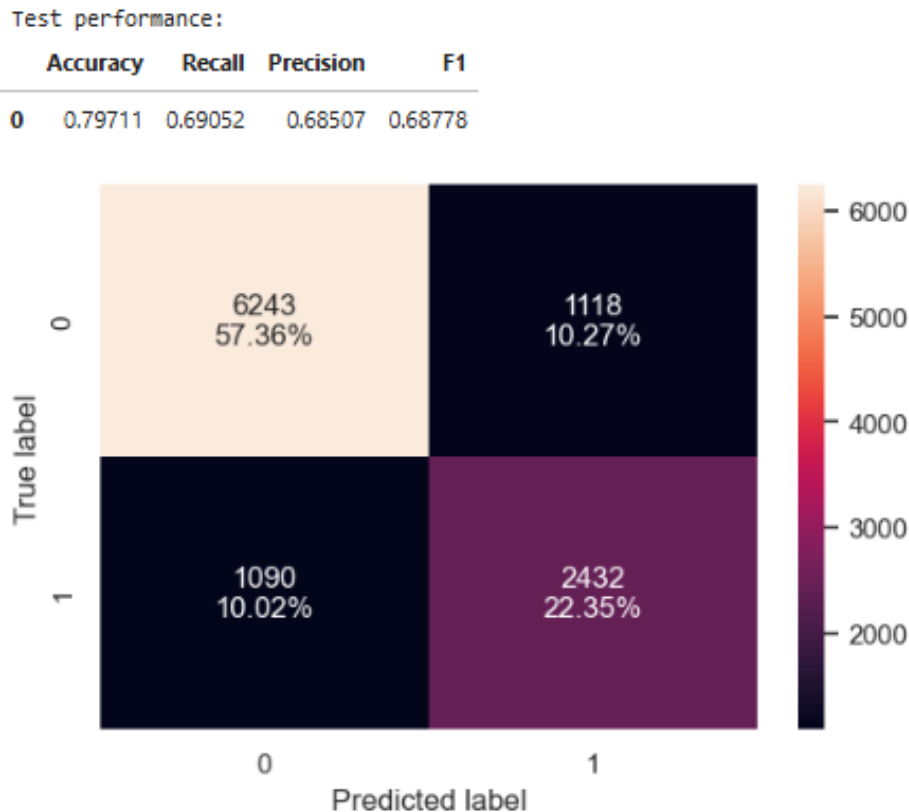


Figure 38 – Final model performance on test set

- Overall Accuracy: The model achieves a moderate overall accuracy of 79.71%. This indicates that the model is able to correctly classify a significant portion of the test samples.
- Recall: The recall score is relatively high at 69.05%, suggesting that the model is able to correctly identify a large majority of positive instances.
- Precision: The precision score is also high at 68.51%, indicating that when the model predicts a positive instance, it is likely to be correct.
- F1-Score: The F1-score is 68.78%, which is a balanced metric that considers both precision and recall. It suggests that the model has a good overall performance.
- Class Imbalance: The class distribution is slightly imbalanced, with Class 0 being more prevalent. This might contribute to the slightly better performance for Class 0.

Model Performance Summary –

Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80525	0.76225	0.79332
Recall	0.61784	0.81370	0.68815
Precision	0.74711	0.60307	0.68553
F1	0.67635	0.69273	0.68684

TABLE 16 – TRAINING PERFORMANCE COMPARISON

Test performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80741	0.76707	0.79711
Recall	0.62323	0.81857	0.69052
Precision	0.74055	0.60326	0.68507
F1	0.67684	0.69462	0.68778

TABLE 17 – TEST PERFORMANCE COMPARISON

Conclusion –

- We have successfully developed a predictive model that the hotel can utilize to forecast which bookings are likely to be canceled, achieving an F1 score of 0.69 on the training set. This model can help formulate targeted marketing strategies.
- The logistic regression models demonstrate consistent performance across both the training and test sets
- When using the model with the default threshold, it yields a low recall but a high precision score. This means the hotel can accurately predict which bookings are not likely to be canceled, enabling them to provide satisfactory service to these customers and maintain brand equity, albeit at the cost of resource allocation.

- With a threshold of 0.37, the model achieves a high recall but low precision score. This allows the hotel to save resources by effectively identifying bookings likely to be canceled, but it may risk damaging brand equity.
- Setting the threshold at 0.42 strikes a balance between recall and precision, allowing the hotel to effectively manage both resources and brand equity.
- The coefficients for variables such as `required_car_parking_space`, `arrival_month`, `repeated_guest`, and `no_of_special_requests` are negative, indicating that an increase in these factors decreases the likelihood of a customer canceling their booking.
- Conversely, the coefficients for variables like `no_of_adults`, `no_of_children`, `no_of_weekend_nights`, `no_of_week_nights`, `lead_time`, `avg_price_per_room`, and `type_of_meal_plan_Not Selected` are positive, suggesting that an increase in these factors raises the chances of booking cancellations.

Decision Tree –

Let's start by creating functions to calculate various metrics and generate a confusion matrix, so we can avoid repeating code for each model.

The `model_performance_classification_statsmodel` function will evaluate model performance, while the `confusion_matrix_statsmodel` function will plot the confusion matrix.

Checking model performance on training set –

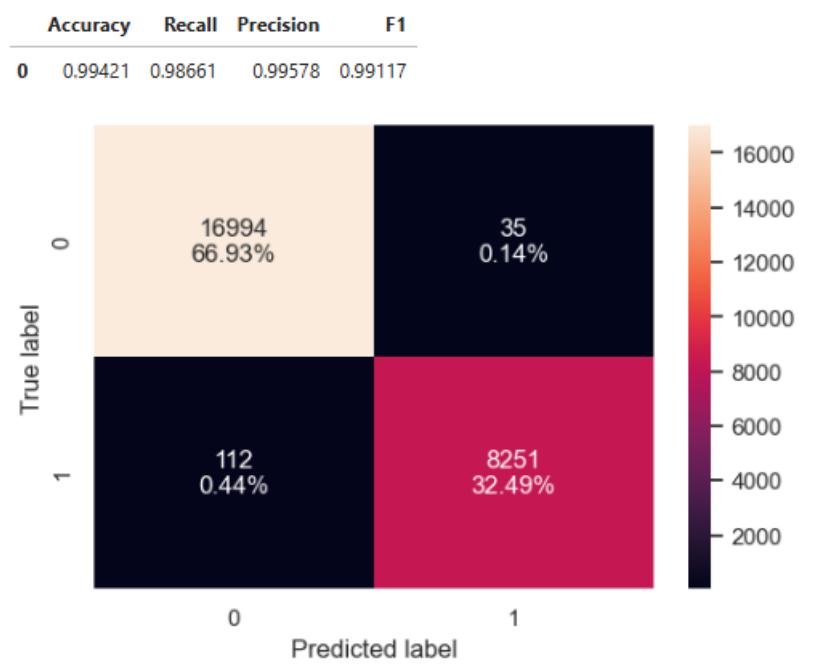


Figure 39 – Decision tree-training set

- There are almost no errors on the training set, with each sample classified correctly.
- Both classes have excellent precision, recall, and F1-scores, demonstrating strong performance for both classes.
- The model has performed exceptionally well on the training data. As we know, a decision tree will continue to grow and perfectly classify each data point if no constraints are applied, as it learns all patterns in the training set.

Now, let's evaluate its performance on the test data to check for potential overfitting.

Checking model performance on Test set –

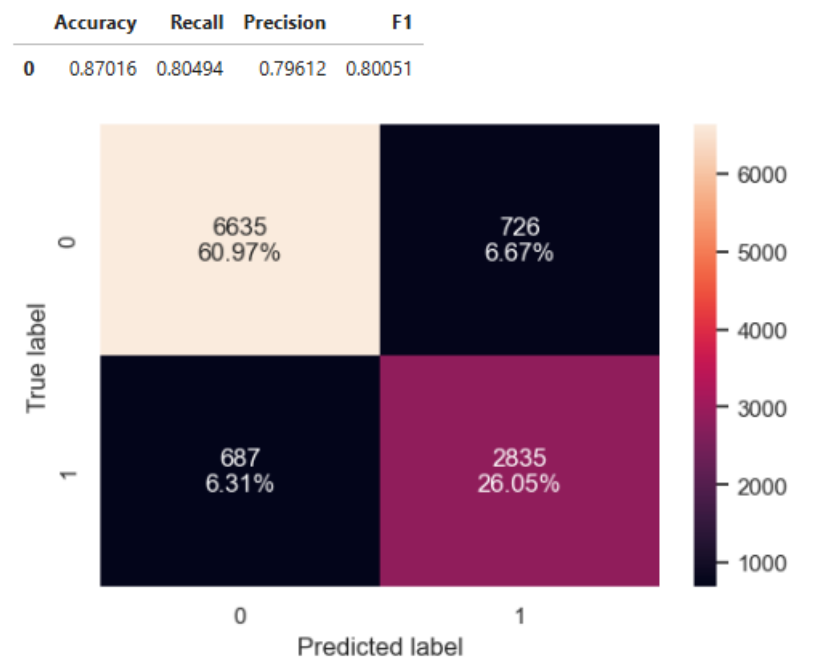


Figure 40 – Decision tree-Test set

- The decision tree model is overfitting the data as expected and not able to generalize well on the test set.
- We will have to prune the decision tree.

Pruning of Tree –

Before pruning the tree, let's examine the most important features. Lead time is the top feature, followed by the average price per room. Now, let's proceed with pruning to see if we can reduce the model's complexity.

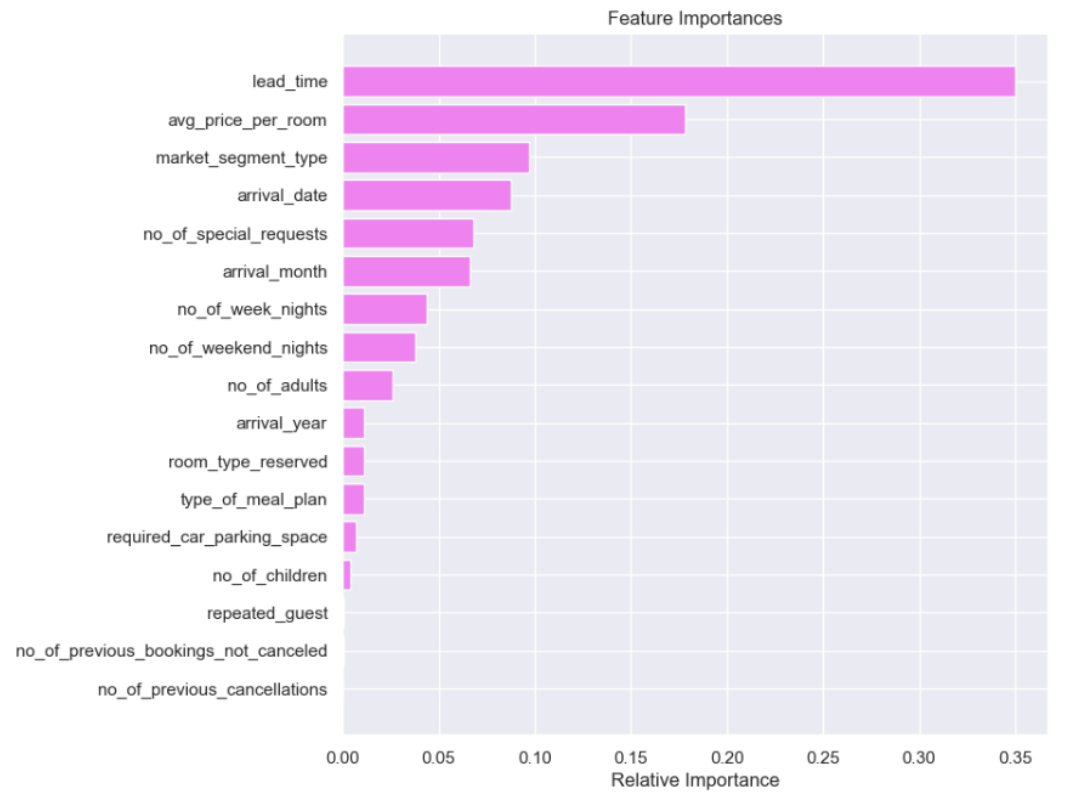


Figure 41 – Prepruning examination

Most Important Features:

- lead_time emerges as the most important feature, indicating that the length of time between booking and arrival significantly influences the target variable.
- avg_price_per_room is also a highly influential feature, suggesting that the price of the room plays a crucial role in the prediction.
- market_segment_type and arrival_date are moderately important, indicating that these factors also contribute to the model's predictions.

Less Important Features:

- Features like no_of_children, repeated_guest, no_of_previous_bookings_not_canceled, and no_of_previous_cancellations have relatively low feature importances, suggesting that they have minimal impact on the model's predictions.

```
DecisionTreeClassifier
DecisionTreeClassifier(class_weight='balanced', max_depth=6, max_leaf_nodes=50,
min_samples_split=10, random_state=1)
```

Table 18 – Pruning stats

Checking model performance on Training set vs Test set–

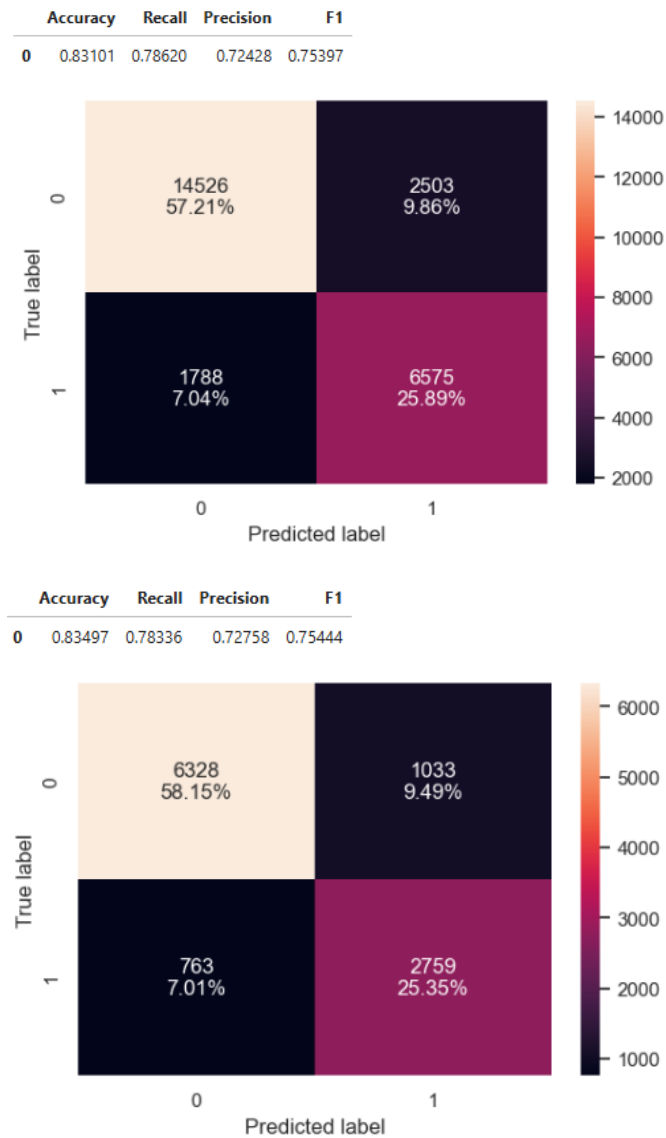


Figure 42 – Pruning – test results Training vs Test

- We can see almost same performance on both sets. This is good sign.

Visualization of Tree –

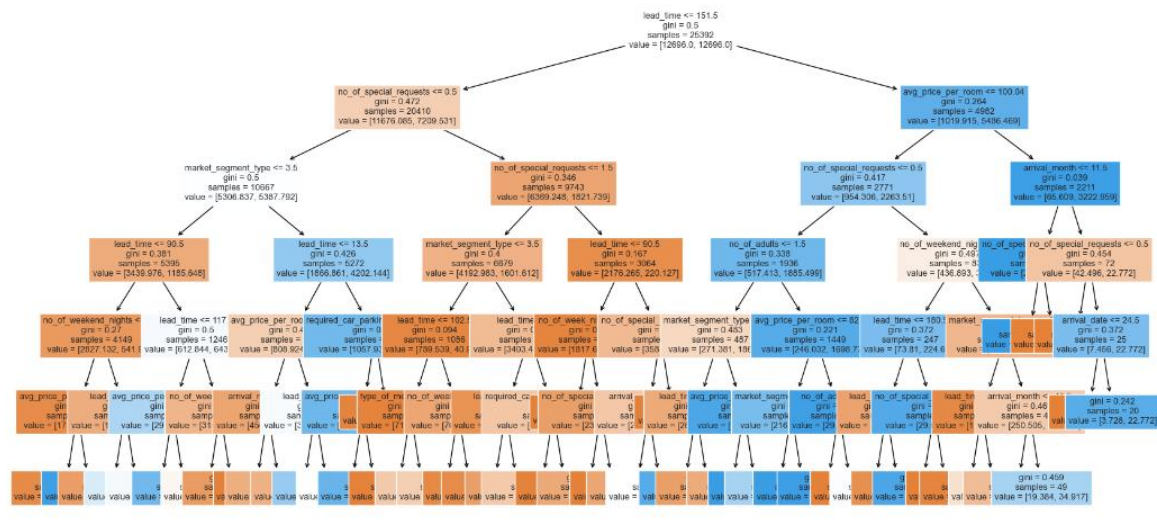


Figure 43 – Decision Tree

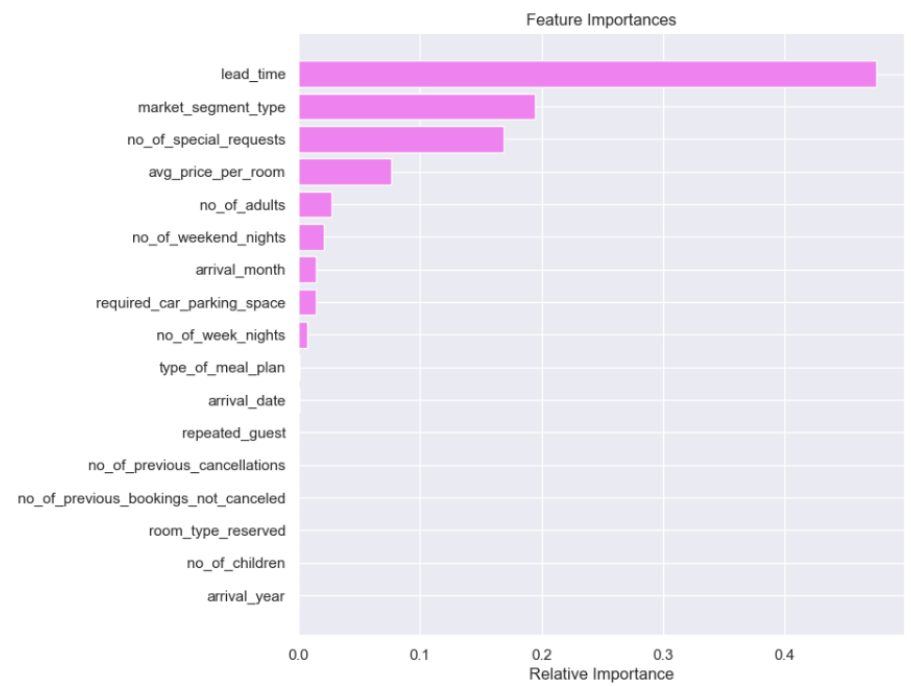


Figure 44– Important variables after Decision tree

Observations

- The pruned tree is now simpler, with more readable rules.
- The model's performance has become more generalized.

Key Features as per figure- 43:

- Lead Time
- Market Segment - Online
- Number of Special Requests
- Average Price per Room

Interpretation of Rules:

- Lead Time is crucial for predicting cancellations, with the model using 151 days as a threshold for the first split.

For bookings made more than 151 days before arrival:

- If the average room price exceeds 100 euros and the arrival month is December, the booking is less likely to be canceled.
- If the average room price is 100 euros or less and there are no special requests, the booking is more likely to be canceled.

For bookings made within 151 days of arrival:

- If the customer has at least one special request, the booking is less likely to be canceled.
- If no special requests are made and the booking was done online, it is more likely to be canceled; otherwise, it is less likely to be canceled.

For further complexity, we could explore deeper levels of the tree.

Cost Complexity Pruning

Cost Complexity Pruning is a technique used to reduce the complexity of a decision tree by pruning its branches based on a cost-complexity criterion. This helps prevent overfitting, making the model more generalized and interpretable.

Cost Complexity = Misclassification Cost + $\alpha \times$ Number of Terminal Nodes

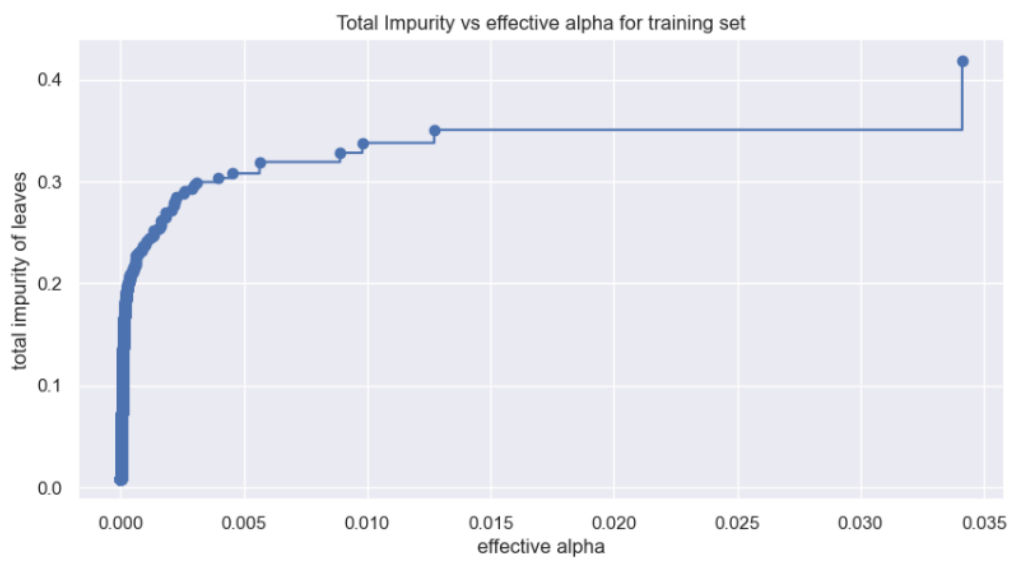


Figure 45– Total impurity vs alpha

Key Observations:

- **Decreasing Total Impurity:** As the alpha value increases, the total impurity of the leaves generally decreases. This suggests that setting a higher alpha value can help simplify the model and reduce overfitting.
- **Elbow Point:** There is an elbow point in the curve where the rate of impurity reduction slows down. This point often indicates a balanced trade-off between model complexity and performance.
- **Optimal Alpha Value:** The optimal alpha value can be identified by locating this elbow point or by assessing model performance on a validation set.

Interpretation:

- **Cost Complexity Pruning:** The plot shows how cost complexity pruning affects the decision tree model. By adjusting the alpha value, you can control tree complexity by pruning branches that add minimal performance value.
- **Trade-off Between Complexity and Performance:** A higher alpha value results in a simpler model with fewer leaves, reducing the risk of overfitting. However, too much pruning could lead to underfitting and lower performance.
- **Finding the Optimal Alpha:** The elbow point on the curve is a useful indicator of the optimal alpha. It represents where further pruning has diminishing returns in impurity reduction, suggesting minimal performance gains from additional pruning.

Next, we train a decision tree using various effective alpha values. The final value in `ccp_alphas` represents the alpha that prunes the entire tree, resulting in `clfs[-1]`, a tree with only one node.

We get –

Number of nodes in the last tree is: 1 with `ccp_alpha: 0.08117914389137065`

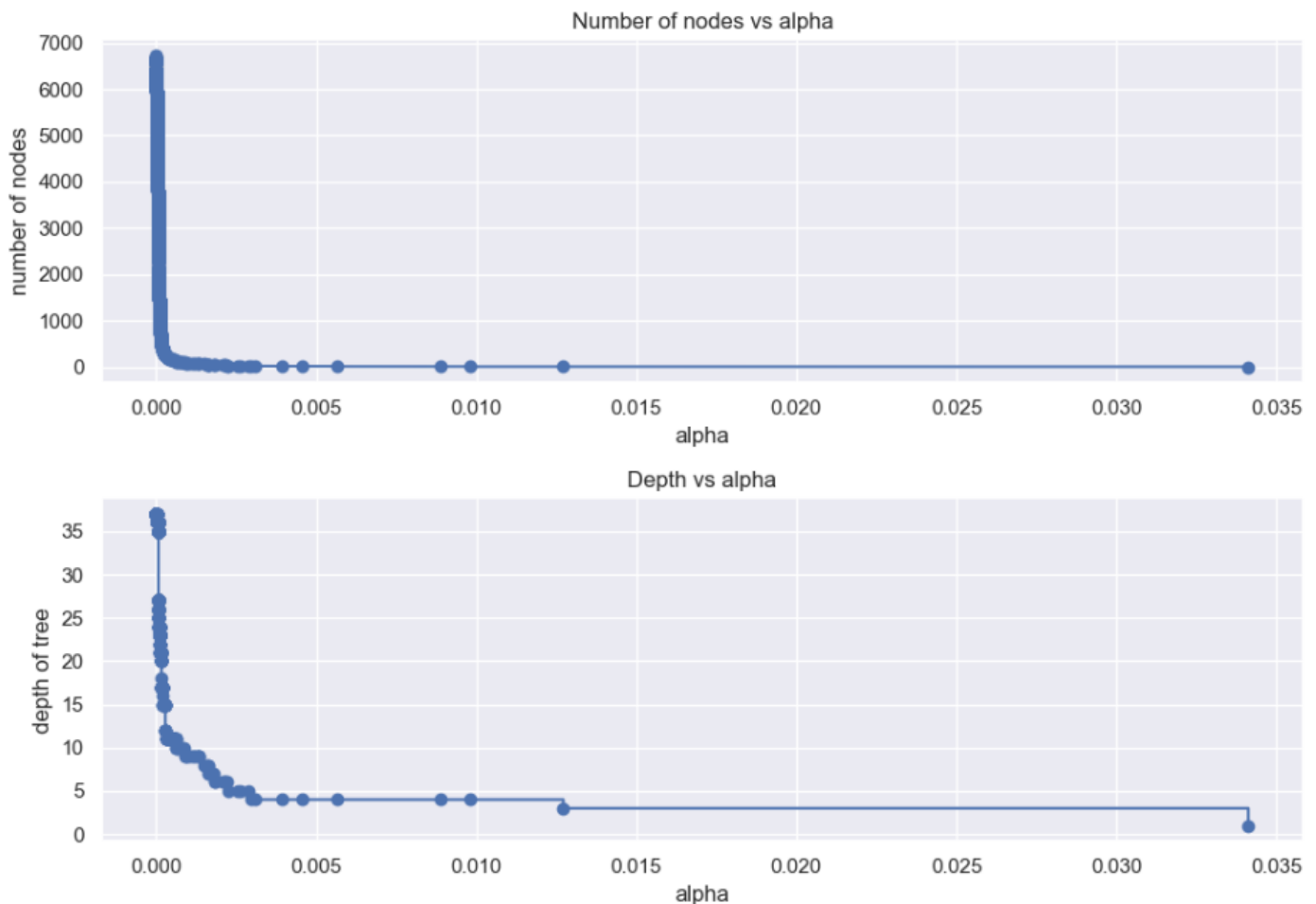


Figure 46– Depth vs Alpha and Nodes vs Alpha

Summary of Pruning Effects

- **Nodes vs. Alpha:**
 - **Decreasing Nodes:** As alpha increases, the number of nodes reduces, simplifying the model.
 - **Elbow Point:** A noticeable elbow point suggests an optimal trade-off between complexity and performance.
- **Depth vs. Alpha:**
 - **Decreasing Depth:** Tree depth also decreases with higher alpha, further simplifying the model.
 - **Elbow Point:** This elbow point helps in identifying the optimal alpha value.

Interpretation:

- **Cost Complexity Pruning:** Adjusting alpha controls tree complexity by pruning branches with minimal impact.
- **Complexity vs. Performance Trade-off:** Higher alpha values simplify the tree, reducing overfitting risk but with a risk of underfitting if over-pruned.
- **Optimal Alpha Selection:** The elbow points indicate where further pruning adds little value.

Further Analysis:

- **Evaluate Performance:** Test different alpha values on a validation set for optimal performance.
- **Visualize Pruned Trees:** Inspect pruned trees across alpha values to see structural changes.

- **Explore Other Pruning Techniques:** Consider alternative methods like pessimistic or error-based pruning for comparison.

So now, we will see **F1 Score vs alpha for training and testing sets respectively.**

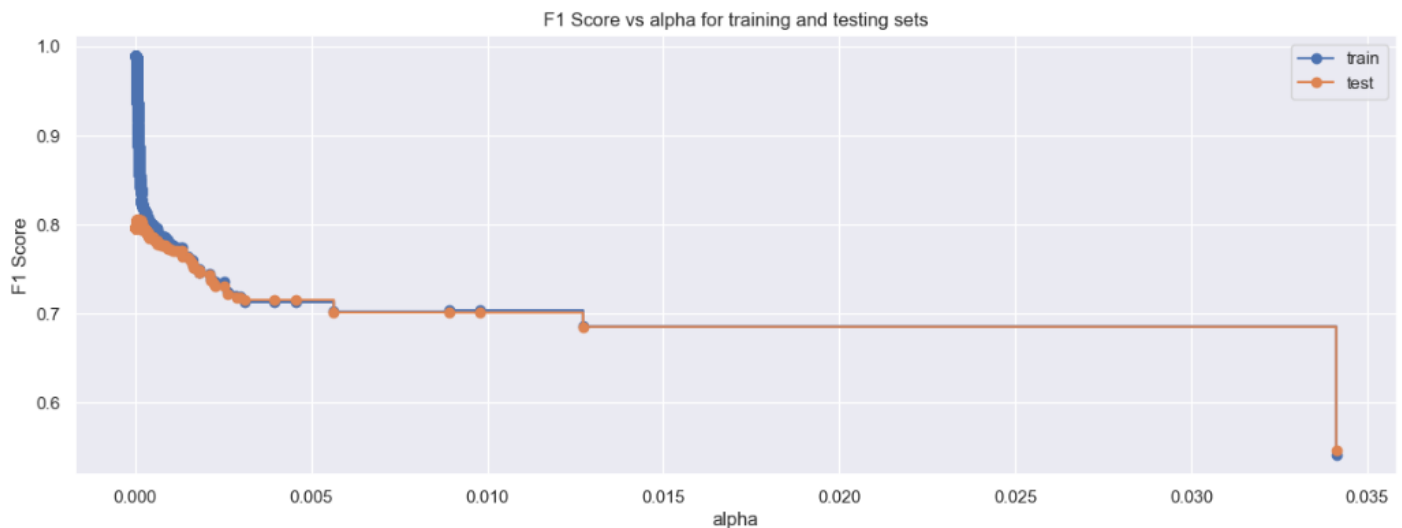


Figure 47– F1 score vs Alpha for Training and Test set

Key Observations:

- **Training vs. Test Performance:** The F1-scores for both training and test sets are plotted against increasing alpha values.
- **Overfitting:** The training set F1-score typically decreases with higher alpha, indicating improved performance on training data as the model simplifies through pruning.
- **Generalization Gap:** As alpha increases, the gap between training and test F1-scores widens, suggesting potential overfitting to the training data and diminished test performance.
- **Optimal Alpha Value:** The optimal alpha likely lies in the range where the training and test F1-scores are closest, reflecting a balance between model complexity and generalization.

Interpretation:

- **Cost Complexity Pruning:** The plot highlights how adjusting alpha affects the decision tree's complexity by pruning less significant branches.
- **Trade-off Between Overfitting and Underfitting:** A lower alpha can lead to overfitting, while a higher alpha risks underfitting, both adversely affecting test performance.
- **Finding the Optimal Alpha:** The goal is to identify the alpha that minimizes the gap between training and test F1-scores, achieving an optimal balance between complexity and generalization.

We get best model as –

```
DecisionTreeClassifier(ccp_alpha=0.0001439672571203496, class_weight='balanced', random_state=1)
```

Checking model performance on Training set vs Test set–

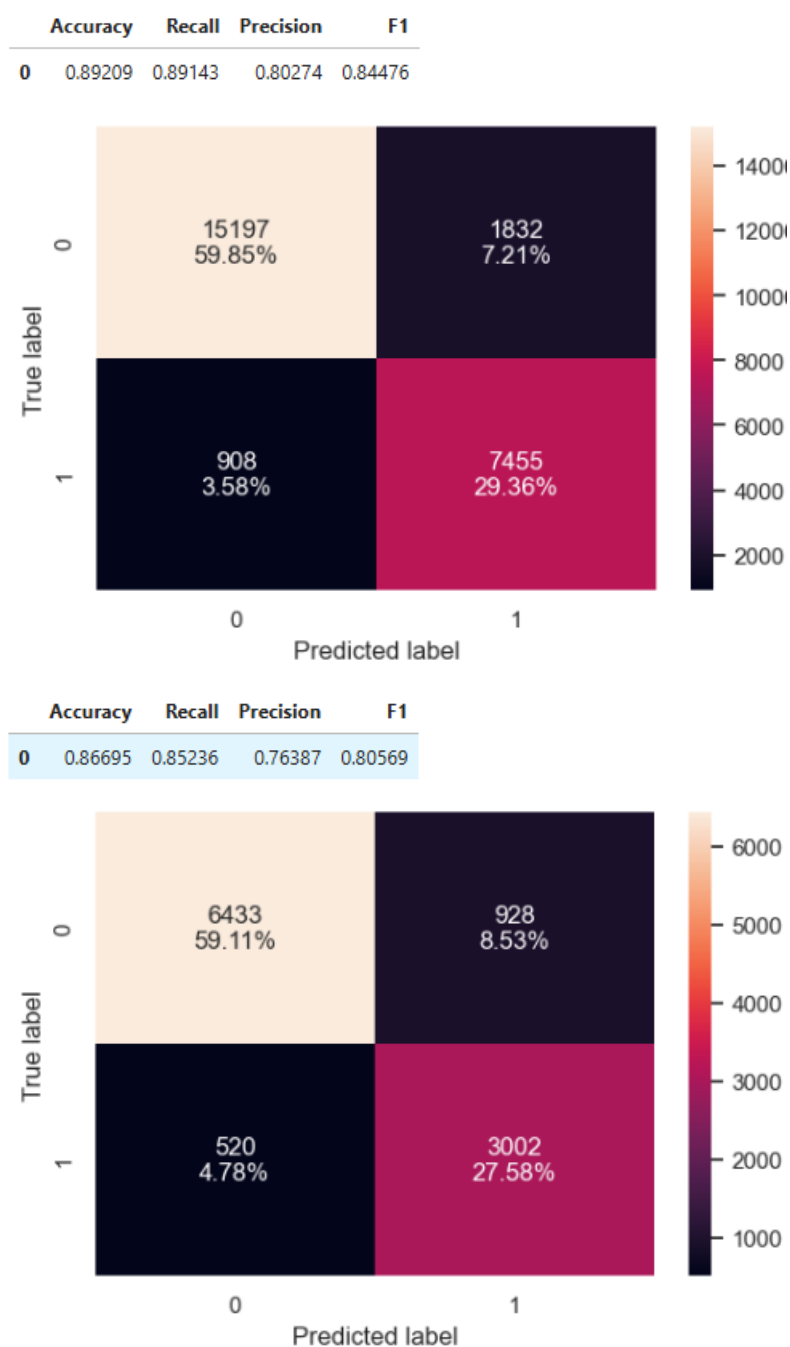


Figure 48 – After Pruning – test results Training vs Test

- After applying post-pruning to the decision tree, we have observed an improvement in the model's performance, which is now more generalized across both the training and test sets.
- While this model is achieving a high recall rate, it is important to note that the **gap between recall and precision has widened**.
- This suggests that, although the model is effectively identifying true positive cases, it may also be misclassifying a significant number of false positives, leading to a decrease in precision.

Now if we observe Decision tree and plot for important variables after post punning–

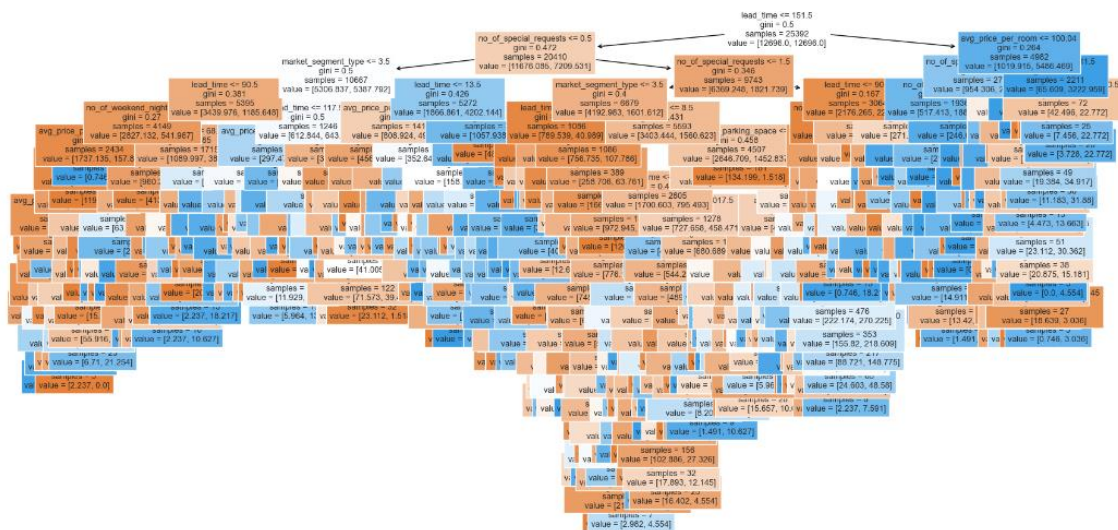


Figure 49 – Decision Tree post prune

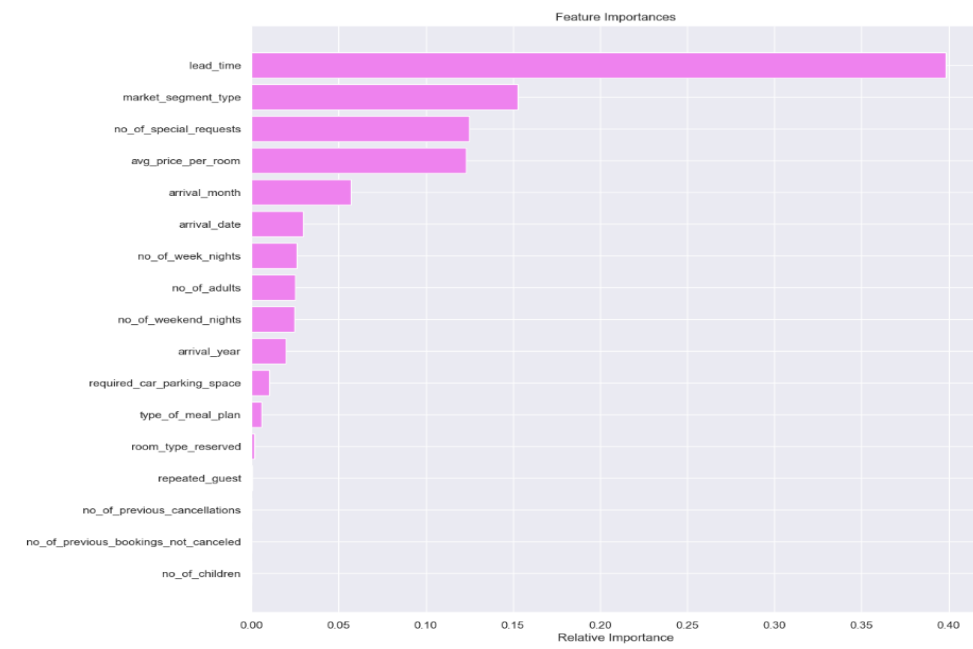


Figure 50 – Important variables 2

- The decision tree we have **now is considerably more complex than the pre-pruned version**. This complexity indicates that the model has many branches and nodes, which can make it harder to interpret and may lead to overfitting.
- However, it's noteworthy that the feature importance values **we obtained remain unchanged** from those identified in the pre-pruned tree.
- This consistency suggests that, despite the increased complexity, the same features are driving the model's predictions, highlighting their continued relevance in influencing the outcomes.

Now, lets compare Training and Test Decision tress -

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83101	0.89209
Recall	0.98661	0.78620	0.89143
Precision	0.99578	0.72428	0.80274
F1	0.99117	0.75397	0.84476

Table 19 Training results for Pre and Post Pruning

Test set performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87016	0.83497	0.86695
Recall	0.80494	0.78336	0.85236
Precision	0.79612	0.72758	0.76387
F1	0.80051	0.75444	0.80569

Table 20 –Test results for Pre and Post Pruning

Analysis of Decision Tree Performance Comparison

Training Set Performance:

- Decision Tree (sklearn): The original decision tree model exhibits outstanding performance on the training set, achieving high accuracy, recall, precision, and F1-score. This indicates an excellent fit to the training data.
- Decision Tree (Pre-Pruning): The pre-pruned model shows a significant decline in performance compared to the original, suggesting that the original model may be overfitting the training data.
- Decision Tree (Post-Pruning): The post-pruned model demonstrates improved performance relative to the pre-pruned version, yet it still does not match the performance of the original model. This indicates that the pre-pruning may have been overly aggressive, resulting in some loss of important information.

Test Set Performance:

- Decision Tree (sklearn): The original model retains good performance on the test set, showing reasonable accuracy, recall, precision, and F1-score.
- Decision Tree (Pre-Pruning): The pre-pruned model exhibits a slight improvement in performance on the test set compared to the original, suggesting that pruning can help mitigate overfitting and enhance generalization to unseen data.
- Decision Tree (Post-Pruning): The post-pruned model also reveals improved performance on the test set compared to the original model, with slightly higher accuracy, recall, and F1-score. This indicates that post-pruning effectively reduces overfitting and enhances the model's generalization ability.

The decision tree model with default parameters is overfitting the training data and struggles to generalize effectively. In contrast, the pre-pruned tree delivers a more generalized performance, achieving a balance between precision and recall. The post-pruned tree demonstrates a higher F1 score compared to the other models; however, there is a significant disparity between precision and recall. By utilizing the pre-pruned decision tree model, the hotel can effectively maintain a balance between resource allocation and brand equity.

Actionable Insights and Recommendations

Insights

- The Decision Tree model outperforms other models on the dataset.
- Key variables, such as Lead Time, Number of Special Requests, and Average Price per Room, are important in both Logistic Regression and Decision Tree models.
- In Logistic Regression, Lead Time and Average Price per Room positively correlate with cancellations, while Number of Special Requests negatively correlates.

Business Recommendations

- Monitor Lead Time and Special Requests: Bookings made under 151 days with special requests are less likely to be canceled. The hotel should:

Implications for Business Strategy

- Enhancing Non-Repeated Guest Experience: The high cancellation rate among non-repeated guests suggests a need for targeted marketing campaigns or special promotions to foster loyalty and reduce cancellations.
- Valuing Repeated Guests: Recognizing repeated guests as a stable customer segment, the business should consider loyalty programs or incentives to encourage their continued patronage.
- Implement an automated email system for booking confirmations and request changes.
- Remind guests of deadlines to allow time for room re-selling or preparations.

Stricter Cancellation Policies:

- High-priced bookings with special requests should not receive a full refund.
- Ensure consistent cancellation policies across all market segments, with reduced refunds for online cancellations.
- Clearly communicate refunds and cancellation fees on the website/app.

Restrict Length of Stay:

- Limit bookings to a maximum of 5 days, with an option to re-book for longer stays, particularly for non-corporate segments.
- This policy can help mitigate cancellations while potentially increasing revenue.

Focus on High-Volume Months:

- December and January see low cancellation rates; ensure adequate staffing.
- Investigate high cancellations in October and September to identify underlying issues.

Enhance Post-Booking Interactions:

- Engage guests post-booking to showcase the level of attention and care they can expect.
- Share information about local events and attractions to personalize their experience.

Improving Experience for Repeated Customers

- Increasing Loyalty: There are few repeated customers, but those who return show low cancellation rates, which is positive for the hospitality industry.
- Cost-Effective Marketing: Repeat customers are more profitable due to familiarity with the hotel. Attracting new customers is more resource-intensive.
- Loyalty Program Development: Implement a program offering special discounts and exclusive services to enhance repeated customer experiences and encourage loyalty. The significant difference in cancellation rates between repeated and non-repeated guests highlights the importance of customer loyalty. Understanding the reasons for cancellations among non-repeated guests can aid in improving retention rates and booking reliability.