PGP-DSBa project report

ML-1 – Coded Project

**BY**

**Ishaan Shakti Jayaraman**

**PGPDSBA.O.JULY24.A**

Contents

[LIST OF FIGURES 4](#_Toc183326477)

[INTRODUCTION 6](#_Toc183326478)

[1.1 Objective 6](#_Toc183326479)

[EXPLORATORY DATA ANALYSIS 7](#_Toc183326480)

[2.1 Problem Definition 7](#_Toc183326481)

[2.2 Key Questions 7](#_Toc183326482)

[2.3 Data Contents 8](#_Toc183326483)

[2.4 Data Dictionary 8](#_Toc183326484)

[2.5 Statistical Summary 10](#_Toc183326485)

[2.6 Univariate Analysis 12](#_Toc183326486)

[2.6.1 No of Adults 12](#_Toc183326487)

[2.6.2 No of Children 13](#_Toc183326488)

[2.6.3 No of Weekend Nights 14](#_Toc183326489)

[2.6.4 No of Week Nights 14](#_Toc183326490)

[2.6.5 Type of Meal Plan 15](#_Toc183326491)

[2.6.6 Required Car Parking Space 16](#_Toc183326492)

[2.6.7 Room Type Reserved 17](#_Toc183326493)

[2.6.8 Lead Time 18](#_Toc183326494)

[2.6.9 Arrival Year 19](#_Toc183326495)

[2.7.0 Arrival Month 19](#_Toc183326496)

[2.7.1 Arrival Date 20](#_Toc183326497)

[2.7.2 Market Segment Type 20](#_Toc183326498)

[2.7.3 Repeated Guests 21](#_Toc183326499)

[2.7.4 No of Previous Cancellations 22](#_Toc183326500)

[2.7.5 No of Previous Bookings Not Cancelled 23](#_Toc183326501)

[2.7.6 Average Price per Room 23](#_Toc183326502)

[2.7.7 No of Special Requests 24](#_Toc183326503)

[2.7.8 Booking Status 25](#_Toc183326504)

[2.7 Bivariate Analysis 26](#_Toc183326505)

[2.7.1 Heatmap 26](#_Toc183326506)

[2.7.2 Average Price per Room VS Booking Status 27](#_Toc183326507)

[2.7.3 Booking Status VS No of Adults + Children 28](#_Toc183326508)

[2.7.4 Average Price per Room VS No of Adults 29](#_Toc183326509)

[2.7.5 Average Price per Room VS Market Segment 30](#_Toc183326510)

[2.7.6 Booking Status VS Market Segment 31](#_Toc183326511)

[2.7.7 Repeated Guests VS Booking Status 32](#_Toc183326512)

[2.7.8 Average Price per Room vs Arrival Month 33](#_Toc183326513)

[2.7.9 No of Special Requests VS Booking Status 34](#_Toc183326514)

[2.7.10 Lead time VS Booking Status 35](#_Toc183326515)

[2.8 Answers to the Key Questions 36](#_Toc183326516)

[2.8.1 What are the busiest months in the hotel? 36](#_Toc183326517)

[2.8.2 Which market segment do most of the guests come from? 37](#_Toc183326518)

[2.8.3 Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments? 38](#_Toc183326519)

[2.8.4 What percentage of bookings are canceled? 39](#_Toc183326520)

[2.8.5 Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel? 40](#_Toc183326521)

[2.8.6 Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation? 41](#_Toc183326522)

[DATA PREPROCESSING 42](#_Toc183326523)

[3.1 Missing Value Check 42](#_Toc183326524)

[3.2 Duplicate Value Check 42](#_Toc183326525)

[3.3 Outlier Check 42](#_Toc183326526)

[3.4 Feature Engineering 42](#_Toc183326527)

[3.5 Data Preparation for Modeling 43](#_Toc183326528)

[MODEL BUILDING 43](#_Toc183326529)

[4.1 Model Evaluation Criteria 43](#_Toc183326530)

[4.2 Logistic Regression Model 43](#_Toc183326531)

[4.3 Naive-Bayes Classifier Model 46](#_Toc183326532)

[4.4 KNN Classifier Model 47](#_Toc183326533)

[4.5 Decision Tree Classifier Model 49](#_Toc183326534)

[MODEL PERFORMANCE IMPROVEMENT 51](#_Toc183326535)

[5.1 Improving Logistic Regression Model 51](#_Toc183326536)

[5.1.1 Check for Multicollinearity 51](#_Toc183326537)

[5.1.2 Removing High P-value Variables 53](#_Toc183326538)

[5.1.3 Determining Optimal Threshold using ROC Curve 54](#_Toc183326539)

[5.2 Improving KNN Classifier 56](#_Toc183326540)

[5.3 Improving the Decision Tree Model 57](#_Toc183326541)

[5.3.1 Pre-Pruning 57](#_Toc183326542)

[5.3.2 Post-Pruning 60](#_Toc183326543)

[FINAL MODEL COMPARISON AND SELECTION 66](#_Toc183326544)

[ACTIONABLE INSIGHTS & RECOMMENDATIONS 67](#_Toc183326545)

# LIST OF FIGURES

[Figure 1 - Statistical Summary of the dataset 10](#_Toc183325995)

[Figure 2 - Analysis of No of Adults 12](#_Toc183325996)

[Figure 3 - Analysis of No of Children 13](#_Toc183325997)

[Figure 4 - Analysis of No of Weekend Nights 14](#_Toc183325998)

[Figure 5 - Analysis of No of Week Nights 14](#_Toc183325999)

[Figure 6 - Analysis of Type of Meal Plan 15](#_Toc183326000)

[Figure 7 - Analysis of Required Car Parking Space 16](#_Toc183326001)

[Figure 8 - Analysis of Room Type Reserved 17](#_Toc183326002)

[Figure 9 - Analysis of Lead Time 18](#_Toc183326003)

[Figure 10 - Analysis of Arrival Year 19](#_Toc183326004)

[Figure 11 - Analysis of Arrival Month 19](#_Toc183326005)

[Figure 12 - Analysis of Arrival Date 20](#_Toc183326006)

[Figure 13 - Analysis of Market Segment Type 20](#_Toc183326007)

[Figure 14 - Analysis of Repeated Guests 21](#_Toc183326008)

[Figure 15 - Analysis of No of Previous Cancellations 22](#_Toc183326009)

[Figure 16 - Analysis of No of Previous Bookings not Cancelled 23](#_Toc183326010)

[Figure 17 - Analysis of Average Price per Room 23](#_Toc183326011)

[Figure 18 - Analysis of No of Special Requests 24](#_Toc183326012)

[Figure 19 - Analysis of Booking Status 25](#_Toc183326013)

[Figure 20 - Heatmap of all Numerical Variables 26](#_Toc183326014)

[Figure 21 - Avg price against Booking Status 27](#_Toc183326015)

[Figure 22 - Booking Status against Total Family 28](#_Toc183326016)

[Figure 23 - Avg Price against No of Adults 29](#_Toc183326017)

[Figure 24 - Avg Price against Market Segment 30](#_Toc183326018)

[Figure 25 - Booking Status against Market Segment 31](#_Toc183326019)

[Figure 26 - Repeated Guests against Booking Status 32](#_Toc183326020)

[Figure 27 - Avg Price against Arrival Month 33](#_Toc183326021)

[Figure 28 - Special Requests against Booking Status 33](#_Toc183326022)

[Figure 29 – Lead time against Booking Status 34](#_Toc183326023)

[Figure 30 - Total Guests per Month 35](#_Toc183326024)

[Figure 31 – Guests per Market Segment 36](#_Toc183326025)

[Figure 32 - Avg Price per Room against Market Segment 37](#_Toc183326026)

[Figure 33 - Booking Status 38](#_Toc183326027)

[Figure 34 - Repeated Guests against Booking status 39](#_Toc183326028)

[Figure 35 - Special Requests against Booking Status 40](#_Toc183326029)

[Figure 36 - Outlier Detection using Boxplots 41](#_Toc183326030)

[Figure 37 - Logistic Regression Model Results 43](#_Toc183326031)

[Figure 38 - Confusion Matrix on Training Data (LR) 44](#_Toc183326032)

[Figure 39 - Confusion Matrix on Test Data (LR) 44](#_Toc183326033)

[Figure 40 - Confusion Matrix on Training Data (NB) 45](#_Toc183326034)

[Figure 41 - Confusion Matrix on Test Data (NB) 46](#_Toc183326035)

[Figure 42 - Confusion Matrix on Training Data (KNN) 47](#_Toc183326036)

[Figure 43 - Confusion Matrix on Test Data (KNN) 47](#_Toc183326037)

[Figure 44 - Confusion Matrix on Training Data (DT) 48](#_Toc183326038)

[Figure 45 - Confusion Matrix on Test Data (DT) 49](#_Toc183326039)

[Figure 46 - VIF Variables on Original Dataset 50](#_Toc183326040)

[Figure 47 - VIF Variables after dropping Online Market 51](#_Toc183326041)

[Figure 48 - Logistic Regression Model Post Muticollinearity 52](#_Toc183326042)

[Figure 49 - ROC Training Test 53](#_Toc183326043)

[Figure 50 - Confusion Matrix on Training Data (Tuned LR) 54](#_Toc183326044)

[Figure 51 - Confusion Matrix on Test Data (Tuned LR) 54](#_Toc183326045)

[Figure 52 - Confusion Matrix on Training Data (Tuned KNN) 55](#_Toc183326046)

[Figure 53 - Confusion Matrix on Test Data (Tuned KNN) 56](#_Toc183326047)

[Figure 54 - Confusion Matrix on Training Data (Pre-pruned DT) 57](#_Toc183326048)

[Figure 55 - Confusion Matrix on Test Data (Pre-pruned DT) 57](#_Toc183326049)

[Figure 56 - Pre-pruned Decision Tree 58](#_Toc183326050)

[Figure 57 - Pre-pruned Freature Importances 59](#_Toc183326051)

[Figure 58 - Total impurity against Effective Alpha 60](#_Toc183326052)

[Figure 59 - Number of nodes against Effective Alpha 61](#_Toc183326053)

[Figure 60 - Depth against Effective Alpha 61](#_Toc183326054)

[Figure 61 - F1 Score against Training and Testing Data 61](#_Toc183326055)

[Figure 62 - Confusion Matrix on Training Data (Post-pruned DT) 62](#_Toc183326056)

[Figure 63 - Confusion Matrix on Test Data (Post-pruned DT) 62](#_Toc183326057)

[Figure 64 - Post-pruned Decision Tree 63](#_Toc183326058)

[Figure 65 - Post Pruned Feature Importances 64](#_Toc183326059)

[Figure 66 - Training & Testing Data Model Comparison 65](#_Toc183326060)

# INTRODUCTION

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 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.

## 1.1 Objective

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

# EXPLORATORY DATA ANALYSIS

## 2.1 Problem Definition

INN Hotels Group is facing a significant challenge with the high number of booking cancellations and no-shows, leading to various operational and financial setbacks. The cancellations result in lost revenue, increased costs associated with distribution channels and last-minute marketing efforts, and additional administrative burden on staff. To mitigate these issues, the hotel group seeks a data-driven solution that can predict which bookings are likely to be canceled in advance. This will enable the hotel chain to optimize their cancellation policies, minimize losses, and improve operational efficiency.

The objective is to construct a Machine Learning model that predicts the likelihood of a hotel booking cancellation. This model will analyze guest data to identify key factors such as guest characteristics, booking behaviors, and timing to predict cancellations. By accurately predicting cancellations INN Hotels Group can implement strategies to reduce cancellation. Additionally, the model will help the hotel chain avoid by providing insights into potential recommendations to reduce cancellations, leading to more efficient resource management and reduced financial losses.

## 2.2 Key Questions

These are the key questions to be answered during exploratory data analysis

1. What are the busiest months in the hotel?
2. Which market segment do most of the guests come from?
3. Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?
4. What percentage of bookings are canceled?
5. Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?
6. Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?

## 2.3 Data Contents

The dataset (INNHotelsGroup.csv) consists of various customers booking details, here is a summary

* There are 36275 unique observations.
* There are 19 columns of various booking details pertaining to each customer.
* There are 5 numerical datatypes.
* There are 13 integer datatypes and 1 float datatype.
* “required\_car\_parking\_space” and “repeated\_guest” consists of “0” and “1” which can be considered as “Yes” and “No” values.
* “Booking\_ID” entirely consists of unique values, so this column will be dropped from the dataset.
* There are no null values.
* There are no duplicate values.

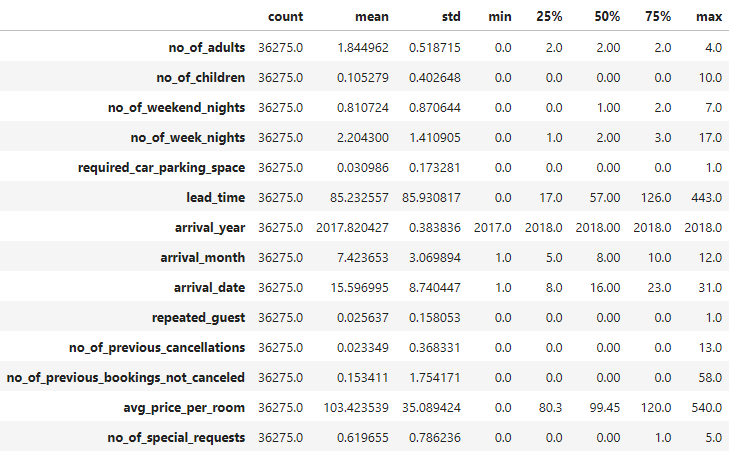
## 2.4 Data Dictionary

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

* 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.

## 2.5 Statistical Summary

Figure - Statistical Summary of the dataset



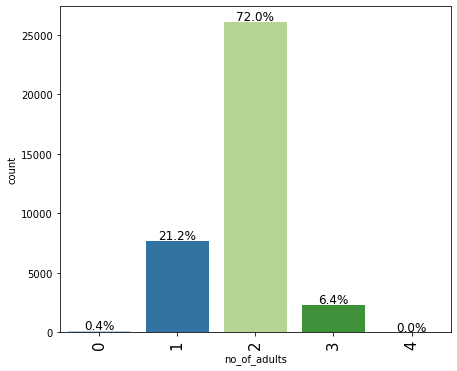
**Insights:**

* **no\_of\_adults:** The average number of adults per booking is 1.84. 50% of bookings have 2 adults or fewer. There are some bookings with a high number of adults (max of 4).
* **no\_of\_children:** The average number of children per booking is 0.10. 50% of bookings have no children. There are some bookings with a high number of children (max of 10).
* **no\_of\_weekend\_nights:** The average number of weekend nights per booking is 0.81. 50% of bookings have 1 weekend night or fewer. There are some bookings with a high number of weekend nights (7).
* **no\_of\_week\_nights:** The average number of week nights per booking is 2.20. 50% of bookings have 2 week nights or fewer. There are some bookings with a high number of week nights (17).
* **required\_car\_parking\_space:** Most bookings (75%) do not require car parking. Only a small percentage of bookings require car parking.
* **lead\_time:** The average lead time for bookings is 85 days. 50% of bookings have a lead time of 57 days or fewer. There are some bookings with very long lead times (443 days).
* **arrival\_year:** The majority of bookings in the dataset were made in the year 2018 and a minority in the year 2017.
* **arrival\_month:** The average arrival months are between July and August, Overall arrivals are balanced throughout the year.
* **arrival\_date:** The average arrival date is 15.59 which is the middle of the month. The minimum is 1 and the maximum is 31 which indicates that arrivals are spread-out throughout the month
* **repeated\_guest:** Most bookings are from new guests. A small percentage of bookings are from repeat guests.
* **no\_of\_previous\_cancellations:** Most guests have no previous cancellations. There are a few guests with a high number of previous cancellations.
* **no\_of\_previous\_bookings\_not\_canceled:** Majority of the data shows that guests with previous bookings did not cancel. A small percentage of the data shows a high number of cancellations (58).
* **avg\_price\_per\_room:** The average price per room is €103.42. 50% of bookings have a price of €99.45 or lower. There are some bookings with maximum price of €540.
* **no\_of\_special\_requests:** The average number of special requests per booking is 0.62. 50% of bookings have no special requests. There are some bookings with a high number of special requests (5).

## 2.6 Univariate Analysis

### 2.6.1 No of Adults

Figure - Analysis of No of Adults

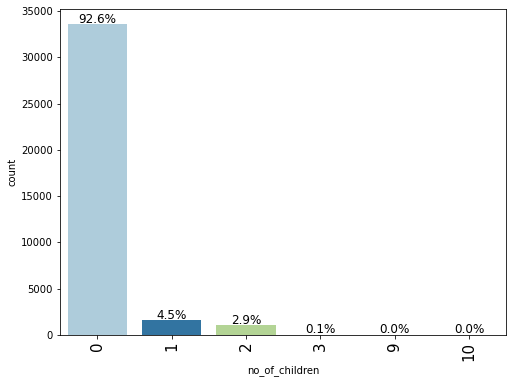


**Insights:**

* Majority of the bookings (72%) include 2 adults.
* Bookings with 1 adult account for 21.2%.
* 6.4% of the bookings include 3 adults.

### 2.6.2 No of Children

Figure - Analysis of No of Children

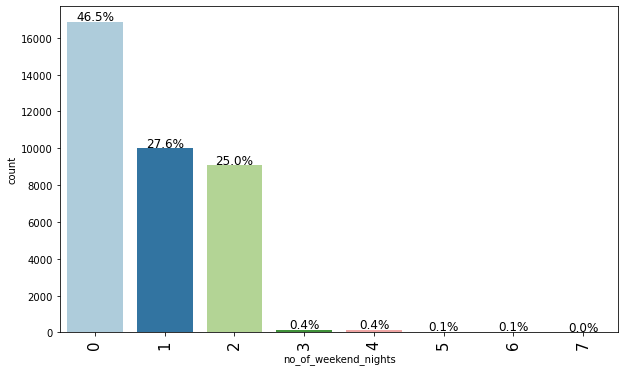


**Insights:**

* Majority of bookings (92.6%) do not include any children.
* Booking that include 1 child are 4.5% of the data and bookings that include 2 children are 2.9% of the data.
* Minor number of bookings include 9 or 10 children, this will be replaced with the maximum amount of 3.

### 2.6.3 No of Weekend Nights

Figure - Analysis of No of Weekend Nights

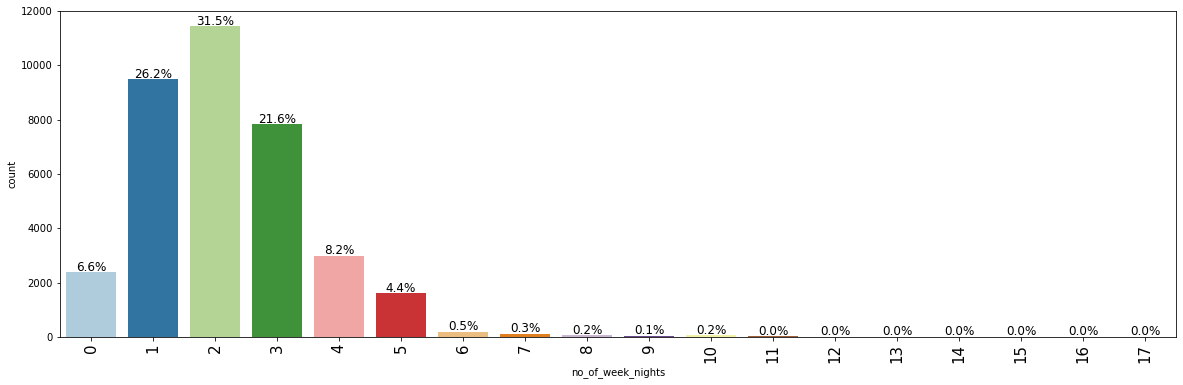


**Insights:**

* 46.5% of total bookings do not include weekend nights.
* Bookings with 1 weekend night account for 27.6% of the data.
* Bookings with 2 weekend nights account for 25% of the data.

### 2.6.4 No of Week Nights

Figure - Analysis of No of Week Nights

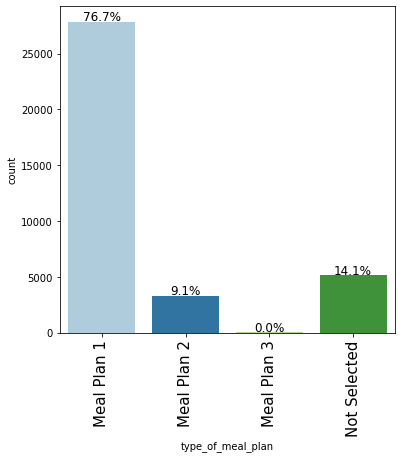


**Insights:**

* 31.5% of the total bookings include two week nights.
* 26.2% of the total bookings include 1 week night.
* 21.6% of the total bookings include 3 week nights.

### 2.6.5 Type of Meal Plan

Figure - Analysis of Type of Meal Plan

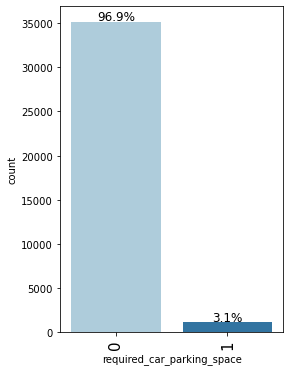


**Insights:**

* In 76.7% of the total bookings, Meal Plan 1 was chosen which includes just breakfast.
* 14.1% have not selected a meal plan.
* 9.1% have selected Meal Plan 2, which includes Half Board.

### 2.6.6 Required Car Parking Space

Figure - Analysis of Required Car Parking Space

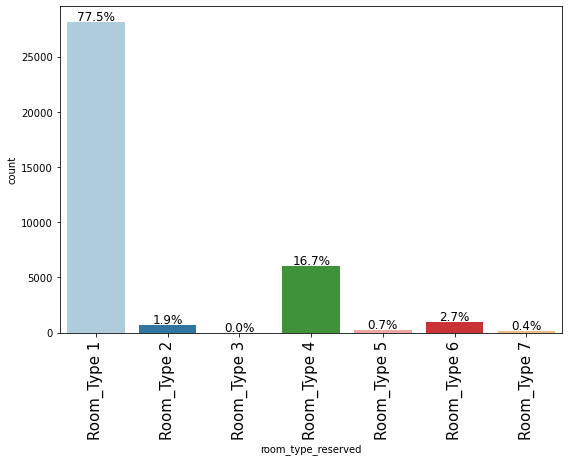


**Insights:**

* 96.9% of the Bookings do not require car parking space.

### 2.6.7 Room Type Reserved

Figure - Analysis of Room Type Reserved

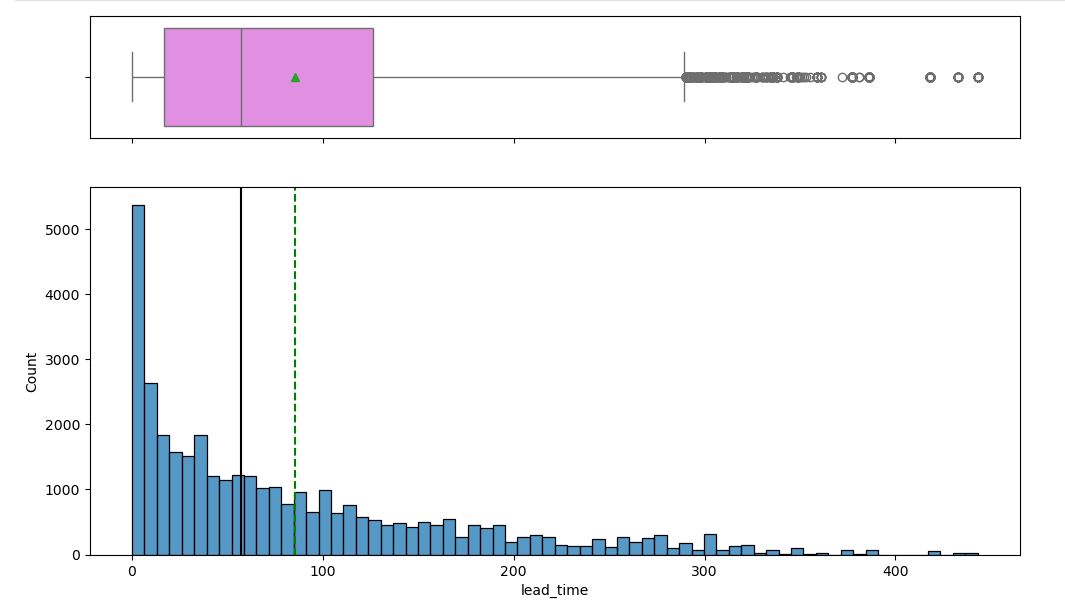


**Insights:**

* Room Type 1 account for 77.5% of the total bookings making it the most popular choice.
* In 16.7% of the total bookings, Room Type 4 was chosen.

### 2.6.8 Lead Time

Figure - Analysis of Lead Time

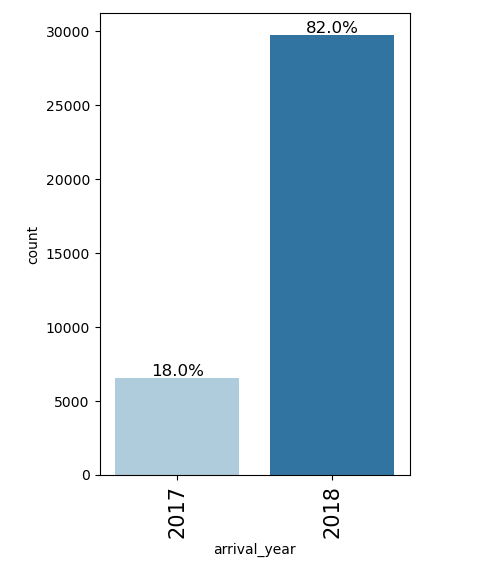


**Insights:**

* The average time is around 85 days between the booking and arrival date which is approximately around 3 months.
* 75% of bookings have a lead time of around 120 days, which is approximately 4 months.
* The distribution is heavily skewed right.

### 2.6.9 Arrival Year

Figure - Analysis of Arrival Year

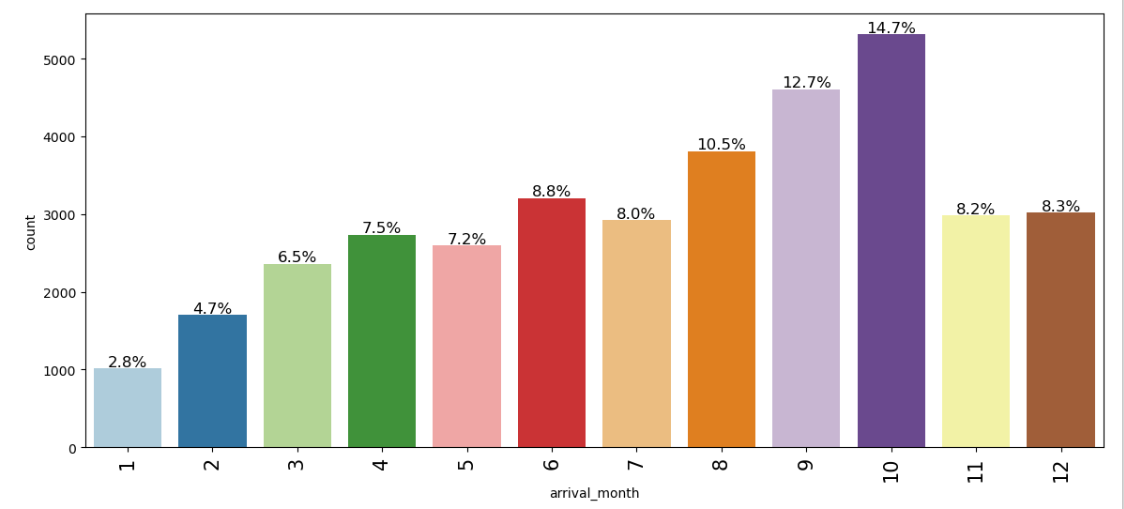


**Insights:**

* 82% of the bookings are in the year 2018.
* 18% of the bookings are in the year 2017.

### 2.7.0 Arrival Month

Figure - Analysis of Arrival Month

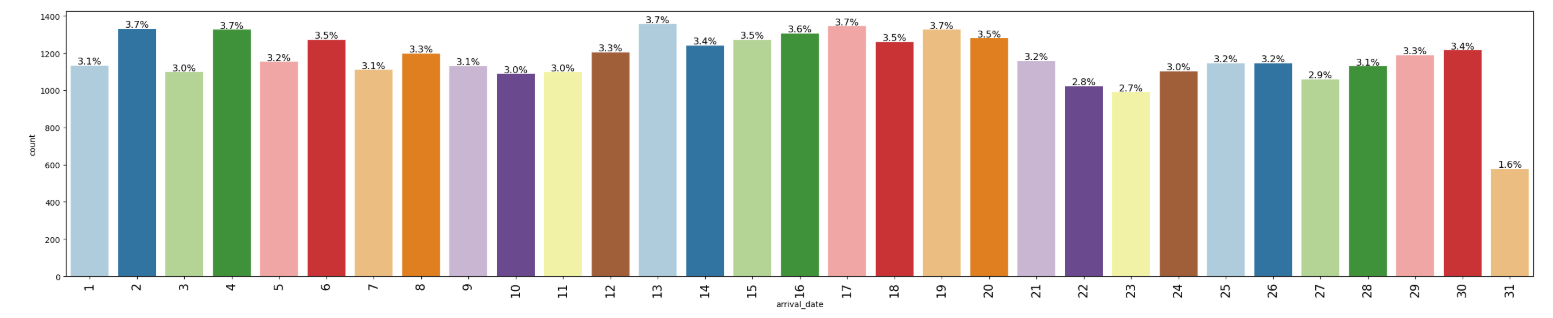


**Insights:**

* The chart indicates that there is seasonal increase in bookings beginning in spring and decreasing during fall season.

### 2.7.1 Arrival Date

Figure - Analysis of Arrival Date

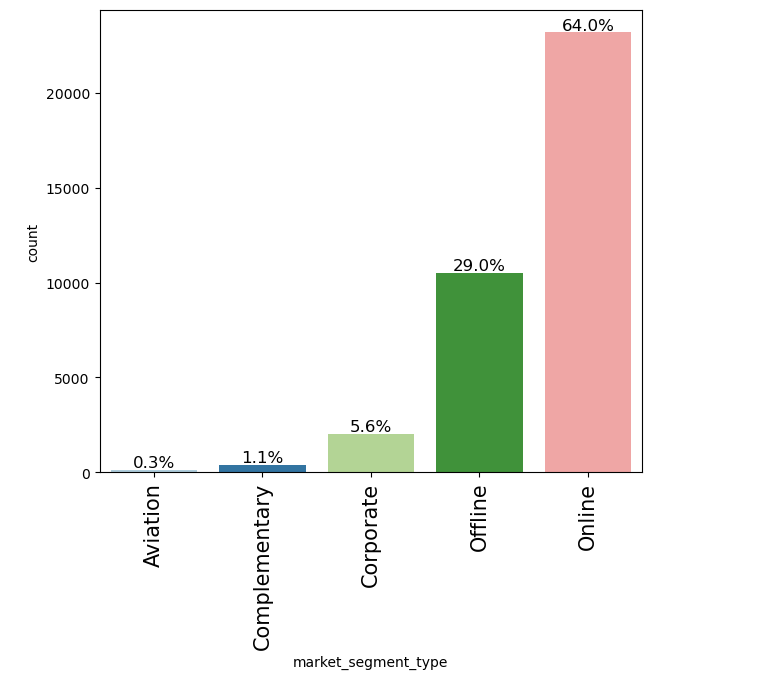


**Insights:**

* There is an almost uniform distribution which indicates that arrival dates are more or less spread out among all dates.

### 2.7.2 Market Segment Type

Figure - Analysis of Market Segment Type

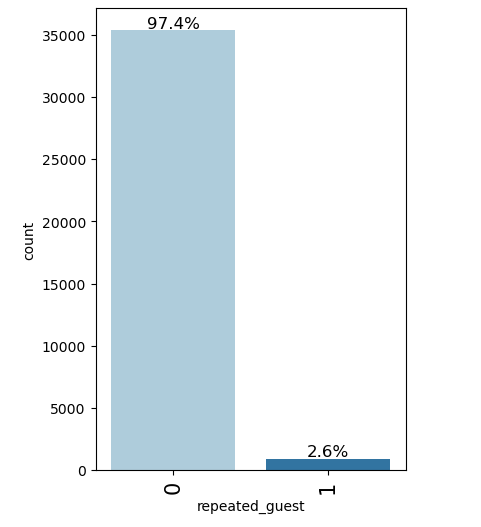


**Insights:**

* Majority of bookings (64%) are done through the “Online” Segment
* “Offline” segment accounts for 29% of the total bookings
* “Corporate” segment accounts for 5.6% of the bookings which indicates small amount of business clientele.
* “Aviation” and “Complimentary” segment form the least number of bookings.

### 2.7.3 Repeated Guests

Figure - Analysis of Repeated Guests

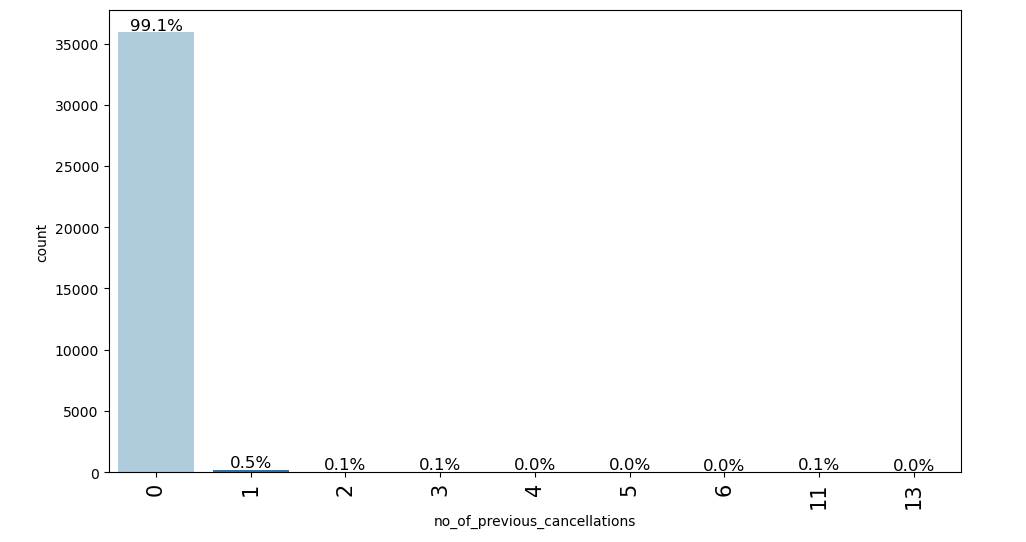


**Insights:**

* Around 97.4% of the total bookings are made by customers who have not previously booked before.
* 2.6% of the bookings are from repeat customers.

### 2.7.4 No of Previous Cancellations

Figure - Analysis of No of Previous Cancellations



**Insights:**

* 99.1% of the bookings are made by customers who have not previously cancelled any bookings.
* 0.5% of the customers have cancelled a booking previously.

### 2.7.5 No of Previous Bookings Not Cancelled

Figure - Analysis of No of Previous Bookings not Cancelled

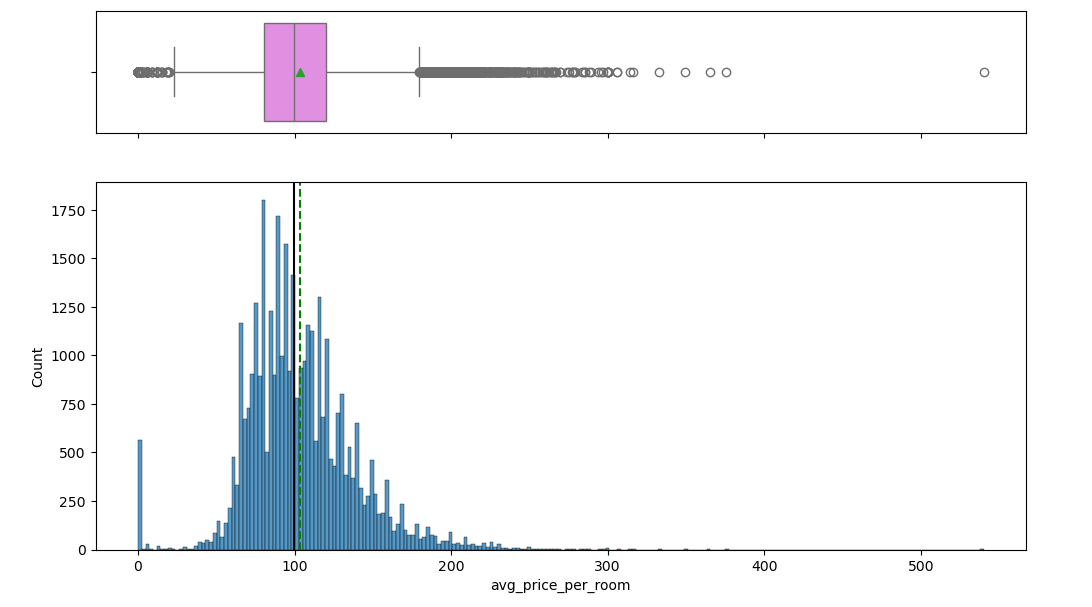


**Insights:**

* Majority of the bookings have no previous bookings which were cancelled.
* Few customers have previous bookings which are not cancelled.

### 2.7.6 Average Price per Room

Figure - Analysis of Average Price per Room

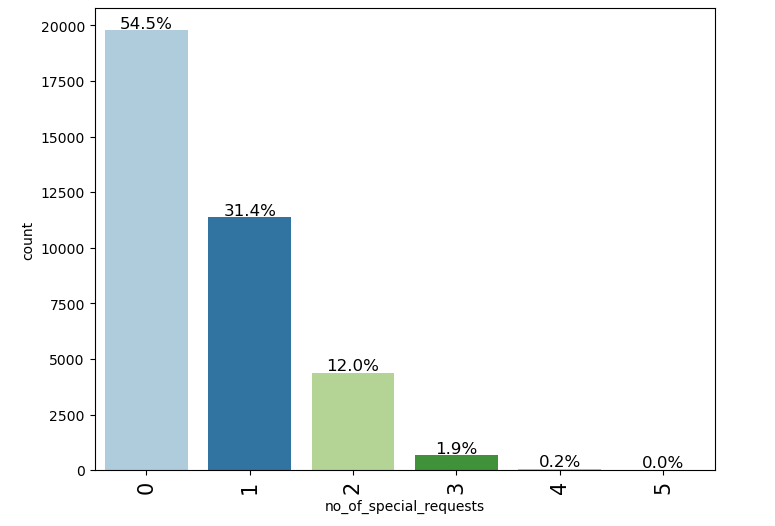


**Insights:**

* The Average price per room is a little €100 approximately.
* The distribution is right skewed indicating fewer amount of bookings as the price increases.

### 2.7.7 No of Special Requests

Figure - Analysis of No of Special Requests

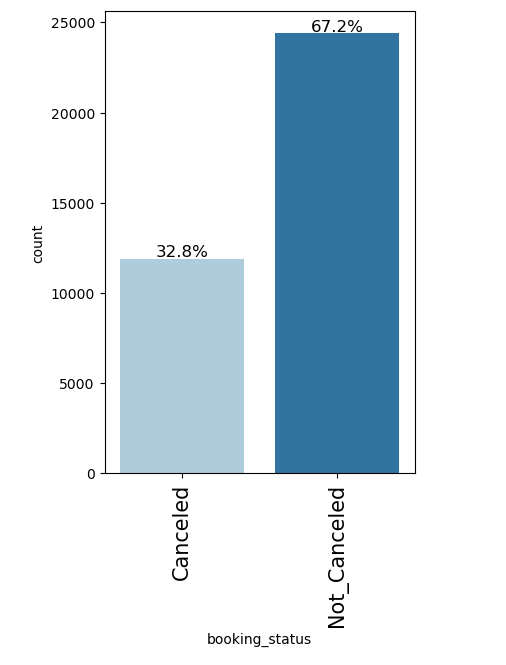


**Insights:**

* 54.5% of the bookings were made by customers who did not have any special requests.
* 31.4% of the bookings included 1 special request made by the customer.
* 12% of the bookings include 2 special requests.
* Less than 2% of the bookings include 3 and more special requests.

### 2.7.8 Booking Status

Figure - Analysis of Booking Status



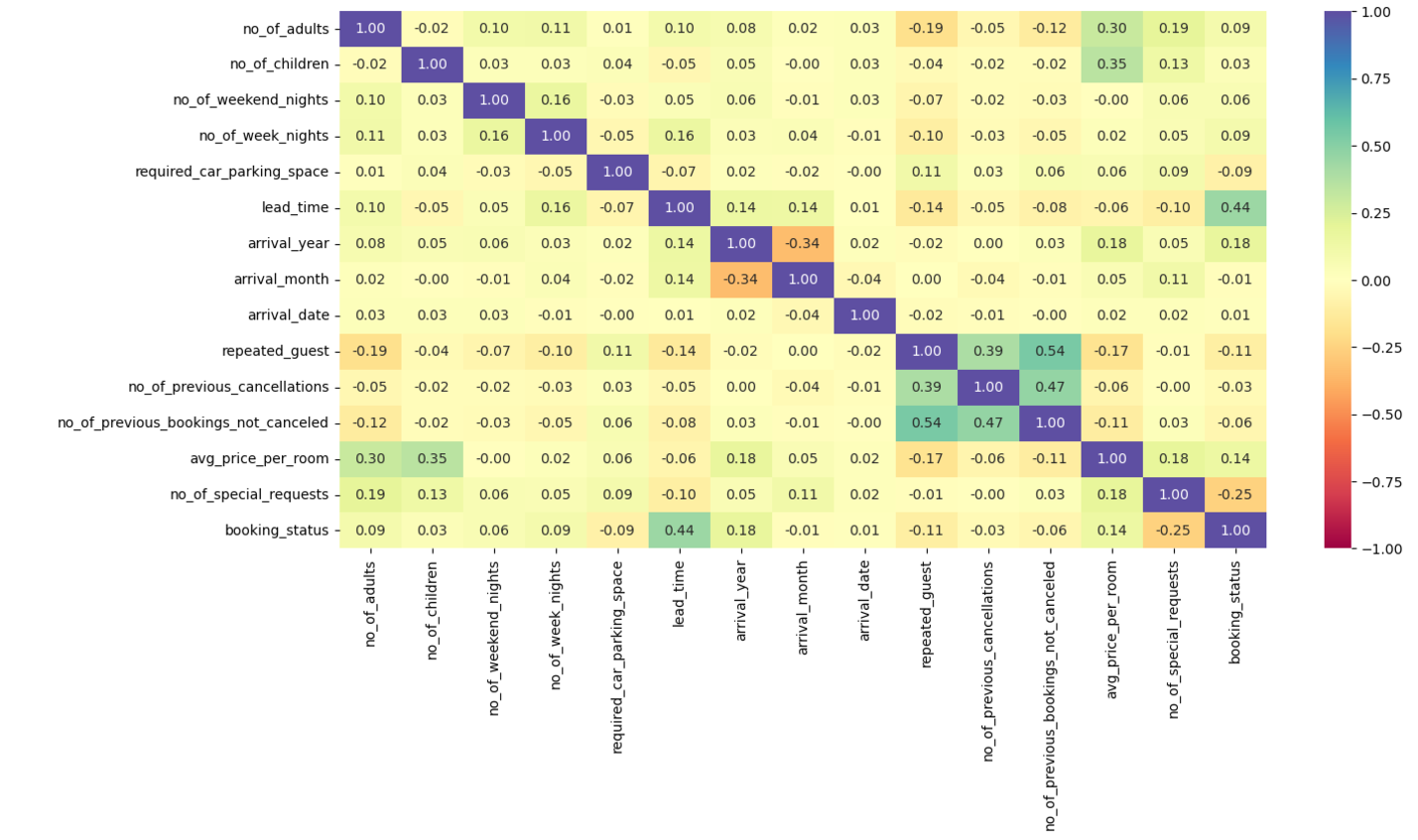
**Insights:**

* 32.8% of the bookings were cancelled.
* 67.2% of the bookings were not cancelled.
* The values of “Canceled” and “Not\_Canceled” will be converted to integer type as “0” and “1” respectively.

## 2.7 Bivariate Analysis

### 2.7.1 Heatmap

Figure - Heatmap of all Numerical Variables

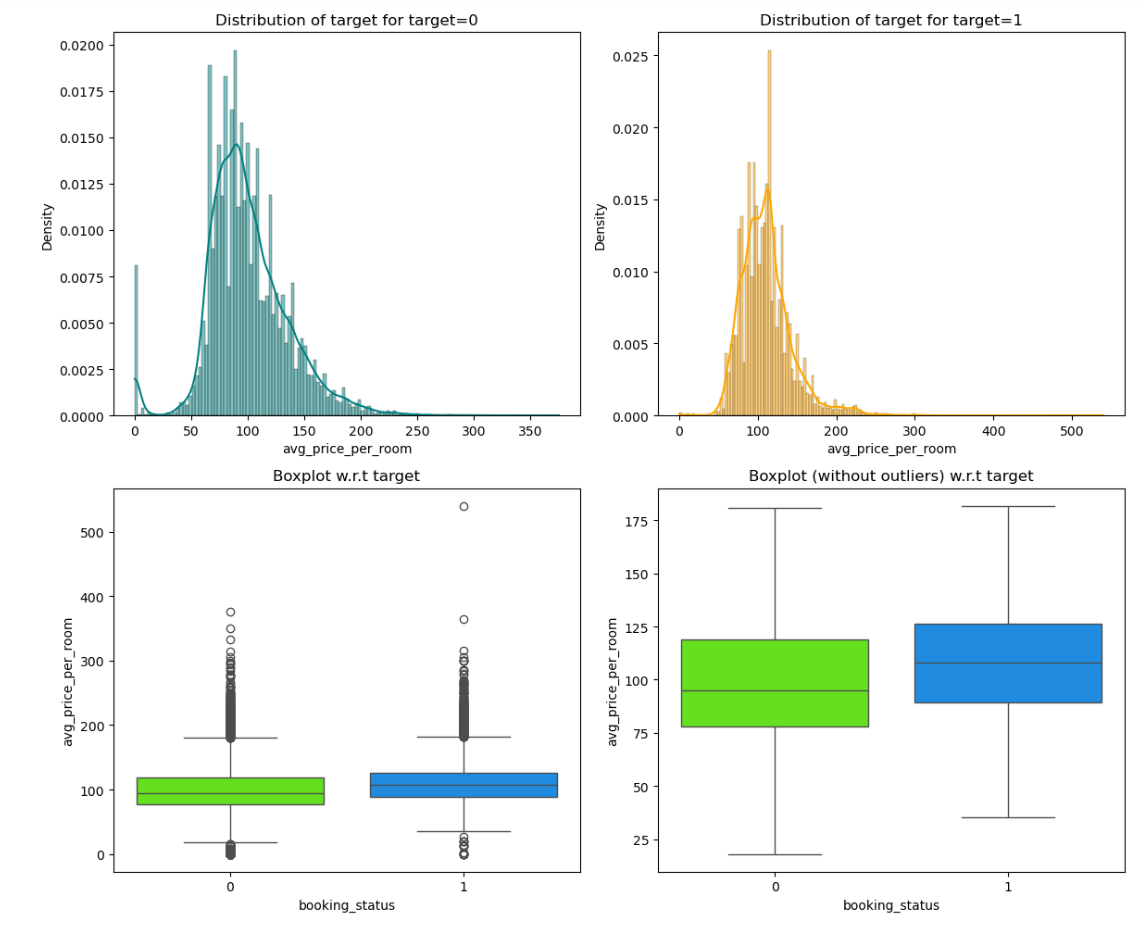


**Insights:**

* “no\_of\_adults” and “avg\_price\_per\_room” show a moderate positive correlation of about 0.30, suggesting that as the number of adults increases, the average price per room tends to rise as well.
* There is a similar moderate positive correlation of 0.35 between “no\_of\_children” and “avg\_price\_per\_room”, indicating that bookings with a higher number of children are linked to higher average room prices.
* “repeated\_guest” is positively correlated with “no\_of\_previous\_bookings\_not\_cancelled” which suggests that repeat customers have a high chance of not cancelling bookings.
* “booking\_status” shows a positive correlation with “lead\_time” indicating that bookings with longer lead times might have a slightly higher likelihood of being canceled.
* There is a negative correlation between “no\_of\_special\_requests” and “booking\_status” indicating less chances of the booking getting canceled when the customer has more special requests.

### 2.7.2 Average Price per Room VS Booking Status

Figure - Avg price against Booking Status

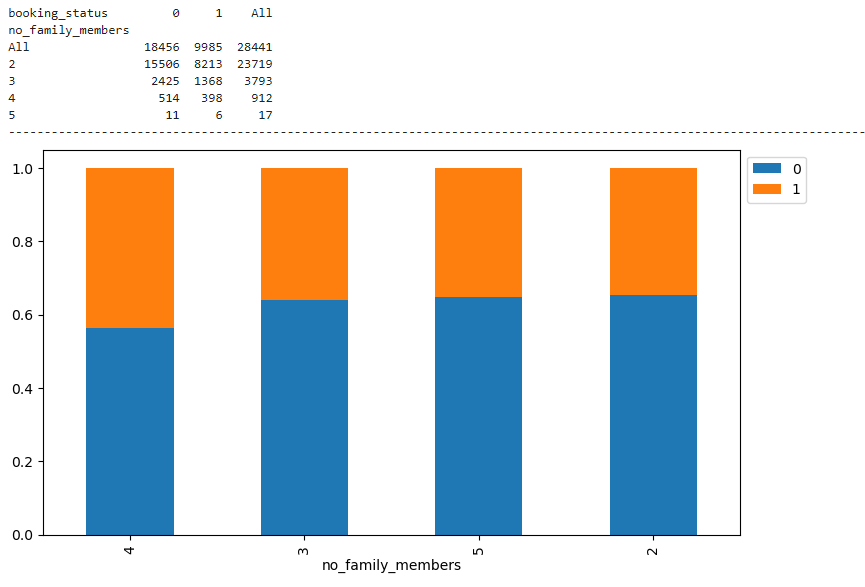


**Insights:**

* Both distributions are somewhat similar
* The median price of cancelled bookings is slightly higher than the median price of confirmed bookings.

### 2.7.3 Booking Status VS No of Adults + Children

Figure - Booking Status against Total Family

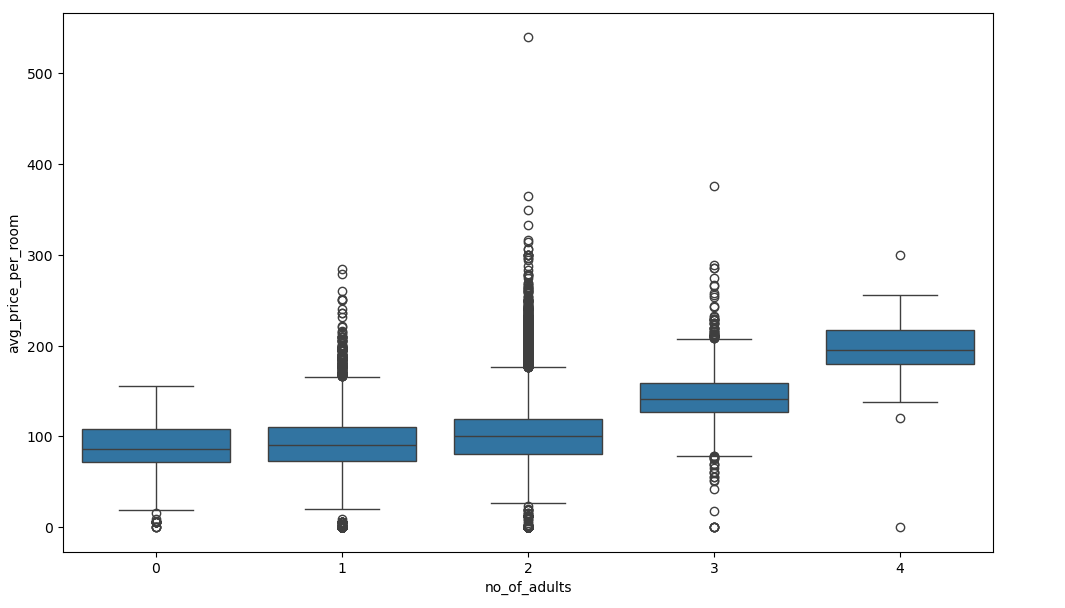


**Insights:**

* Based on the Chart there is approximately 40% chance the cancellations happen if there are more than 2 children

### 2.7.4 Average Price per Room VS No of Adults

Figure - Avg Price against No of Adults

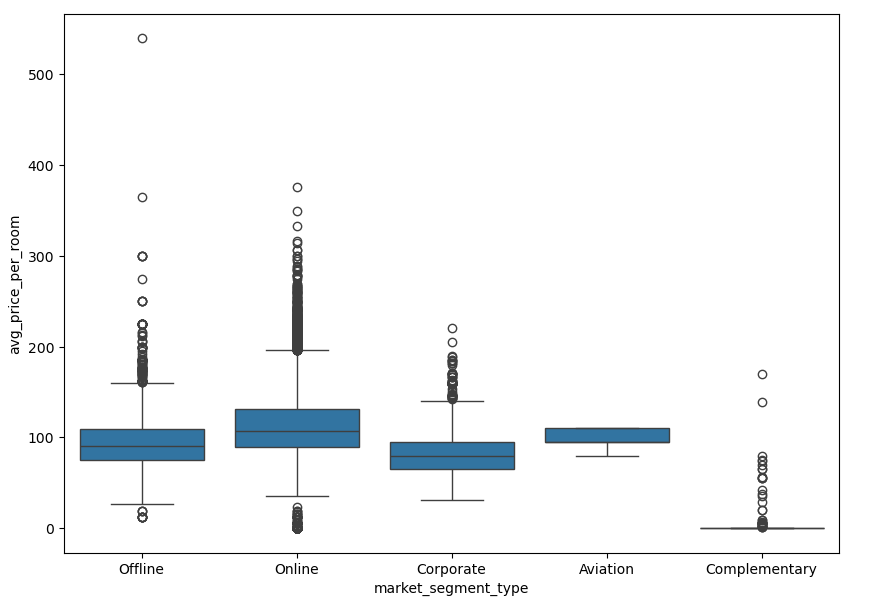


**Insights:**

* There is high correlation between average price and no of adults, based on the boxplot we can see that as no of adults in a booking increase so does the price of the room.

### 2.7.5 Average Price per Room VS Market Segment

Figure - Avg Price against Market Segment

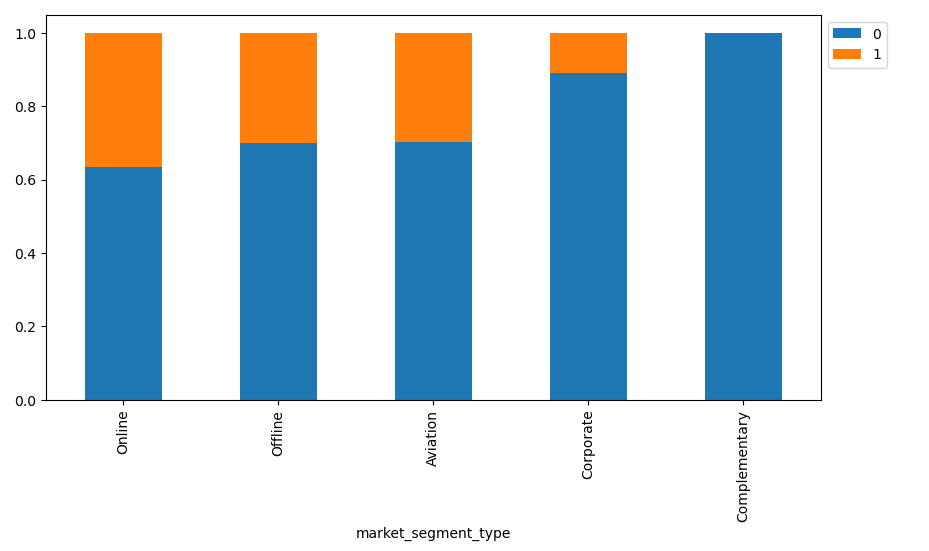


**Insights:**

* The online segment accounts for the highest prices with an average room rate of around €110 in the total bookings
* With the exception of Complimentary, Corporate segment offers the lowest prices of rooms on average.

### 2.7.6 Booking Status VS Market Segment

Figure - Booking Status against Market Segment

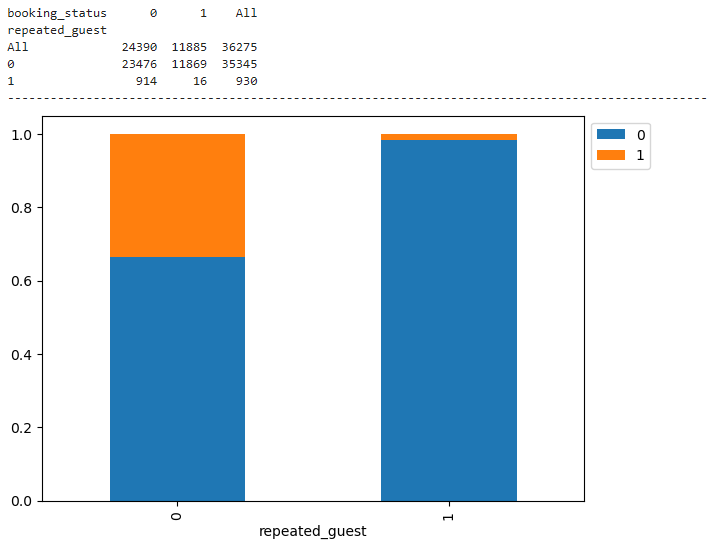


**Insights:**

* No Bookings which were complementary were cancelled by the guests.
* Approximately 38% of the Online bookings segment were cancelled.
* Offline and Aviation segments have a similar level of cancellations.
* Corporate segment bookings have approximately less than 10% chance of being cancelled.

### 2.7.7 Repeated Guests VS Booking Status

Figure - Repeated Guests against Booking Status

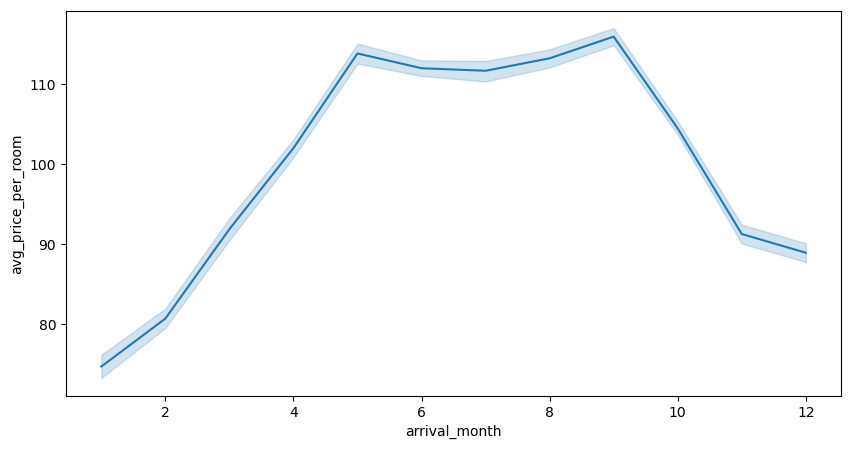


**Insights:**

* Repeat customers account for a small percentage of the total bookings, however the chances of cancellation from repeat customers are very low.

### 2.7.8 Average Price per Room vs Arrival Month

Figure - Avg Price against Arrival Month

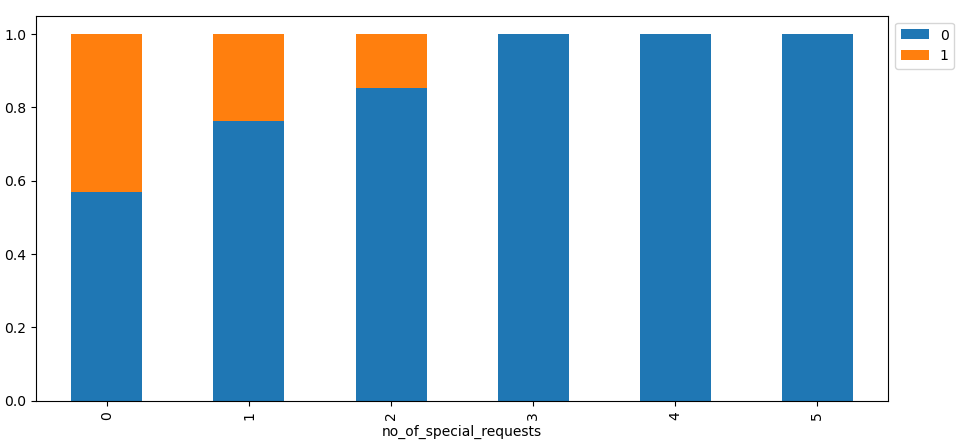


**Insights:**

* The average price of the rooms increases during spring-summer season which is the high season where majority of bookings happen are expected to happen.

### 2.7.9 No of Special Requests VS Booking Status

Figure - Special Requests against Booking Status

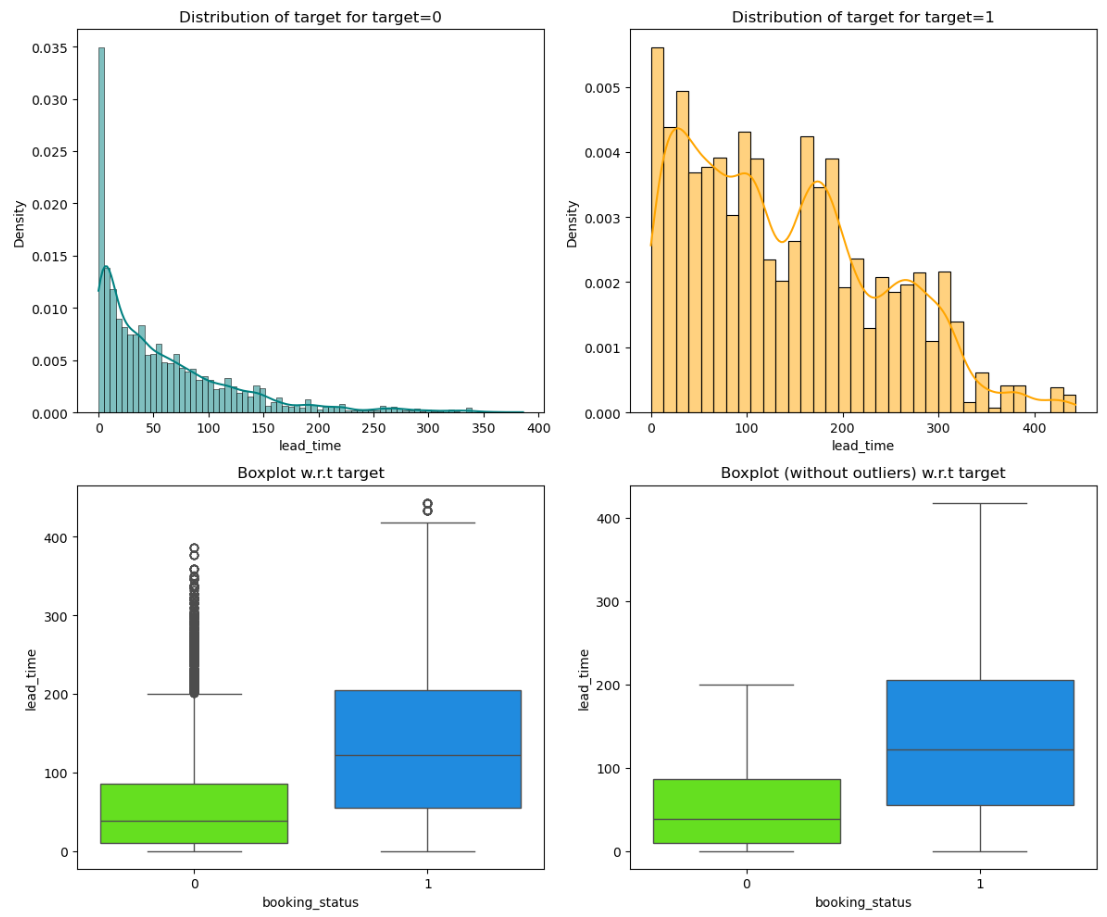


**Insights:**

* The chances of a customer cancelling a booking decrease if they have made more than 2 special requests.

### 2.7.10 Lead time VS Booking Status

Figure – Lead time against Booking Status



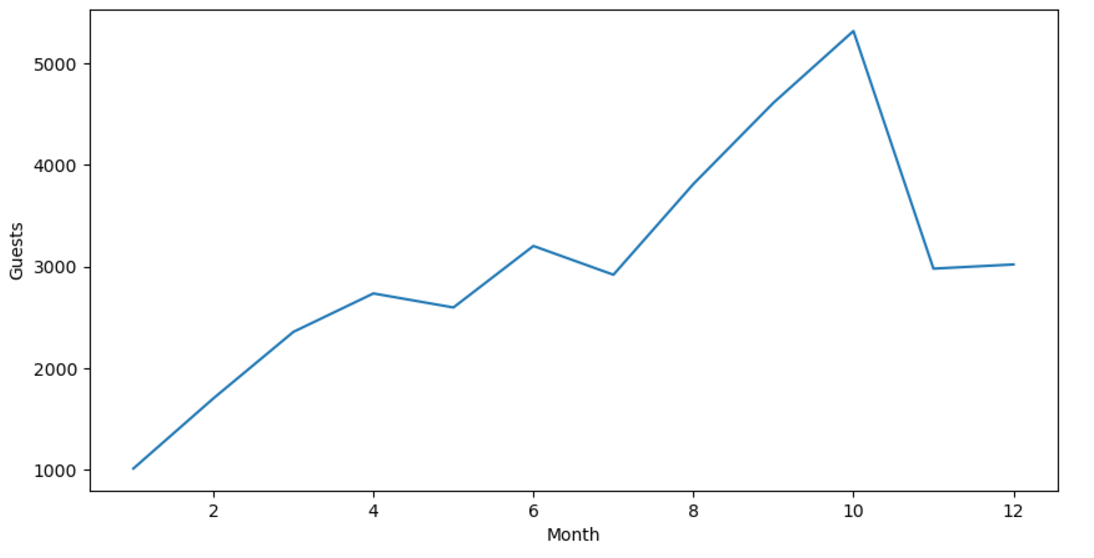
**Insights:**

* The longer the lead the more chances the booking has to get cancelled by the customer.

## 2.8 Answers to the Key Questions

### 2.8.1 What are the busiest months in the hotel?

Figure - Total Guests per Month

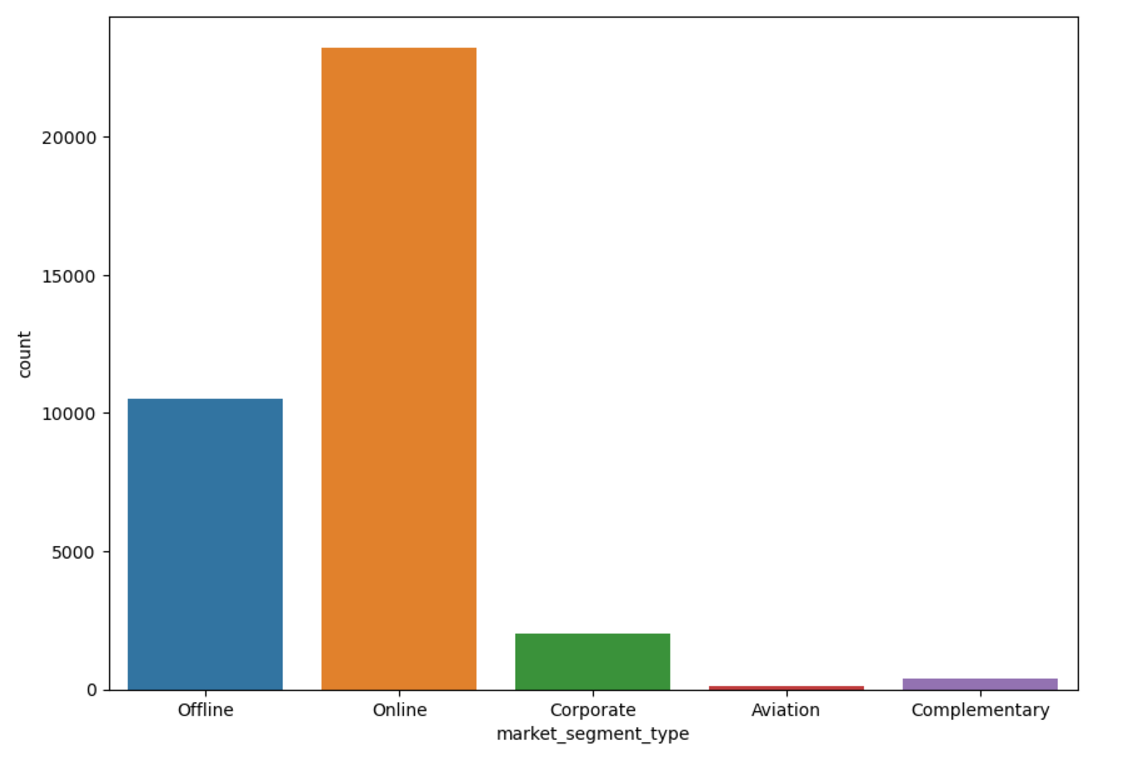


The busiest months in the hotel are August, September and October. The highest number of guests arrive in the month of October.

Summer season is considered as high season for most hotel bookings due to factors such as vacations, holidays and weather conditions.

### 2.8.2 Which market segment do most of the guests come from?

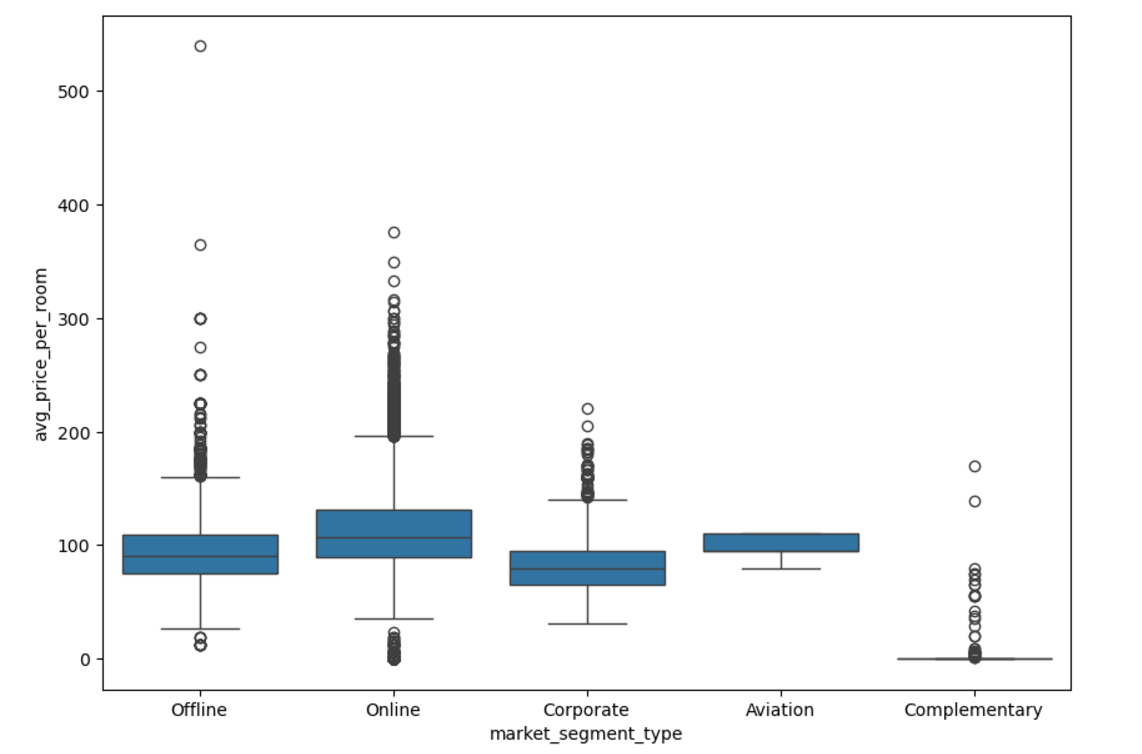
Figure – Guests per Market Segment



The vast majority of the customers come from the “Online” segment. In today’s society online channels such as websites and online agencies remain the dominant market for hotel bookings.

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

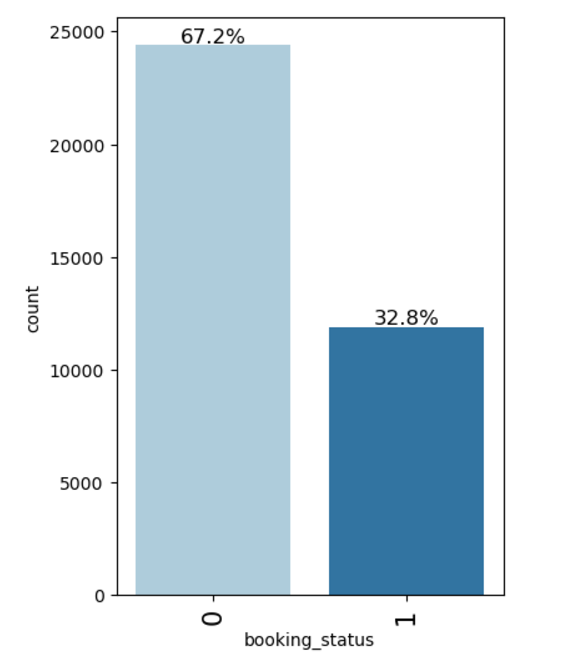
Figure - Avg Price per Room against Market Segment



The online market segment has the highest average prices which indicates that bookings made through various online channels tend to have higher rates compared to other segments. This could be attributed to various factors such as the convenience of online platforms. The offline market segment tends to have relatively lower room prices compared to the online segment. This might be because offline bookings may involve direct negotiation with the hotel or traditional travel agencies, which could result in lower negotiated rates or special discounts. The room prices vary across different market segments.

### 2.8.4 What percentage of bookings are canceled?

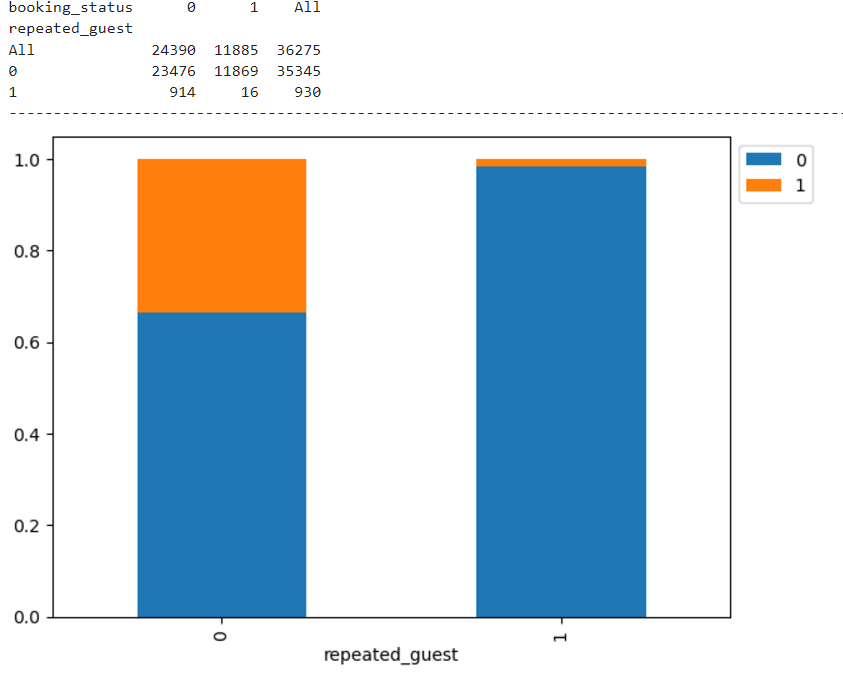
Figure - Booking Status



From the above barplot we can see that around 32.8% of the total bookings have been cancelled.

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

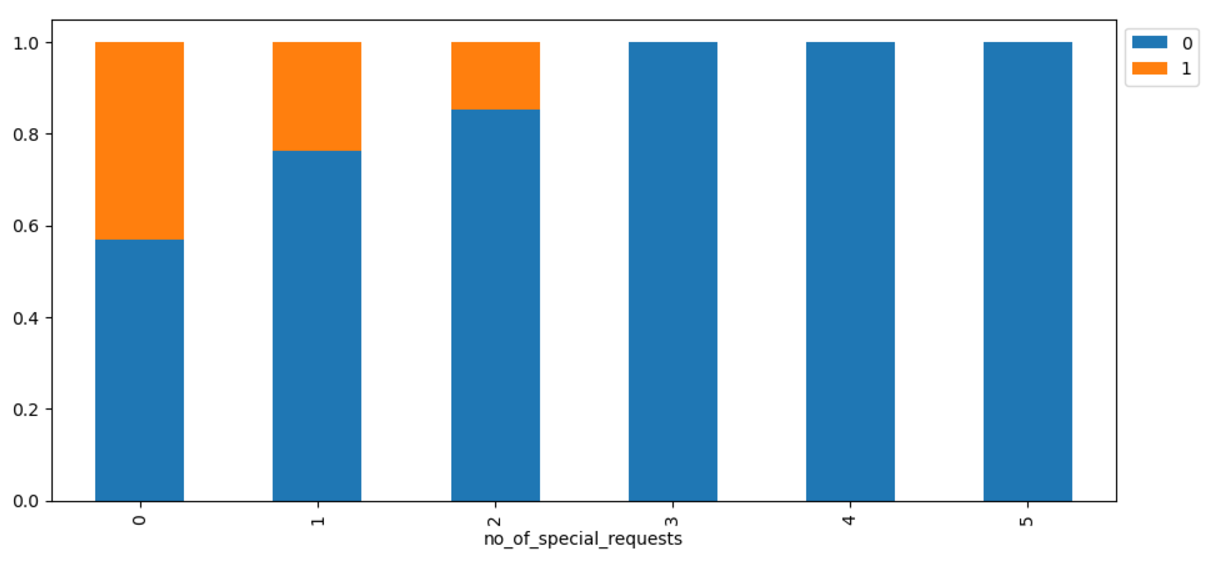
Figure - Repeated Guests against Booking status



From the overall bookings, 930 customers have made previous bookings before and only 16 of them have cancelled. Only 1.7% of repeated guests cancel the booking. 98.3% do not cancel their booking.

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

Figure - Special Requests against Booking Status



As the number of special requests increases, the cancellation rate decreases. Bookings with no special requests had an approximate 43% cancellation rate, while those with 1 special request had a lower rate of approximately 24%. The cancellation rate further dropped to approximately 15% for bookings with 2 special requests. Bookings with 3 to 5 special requests showed no cancellations at all. This indicates that guests with special requests are more committed to their bookings and less likely to cancel.

# DATA PREPROCESSING

## 3.1 Missing Value Check

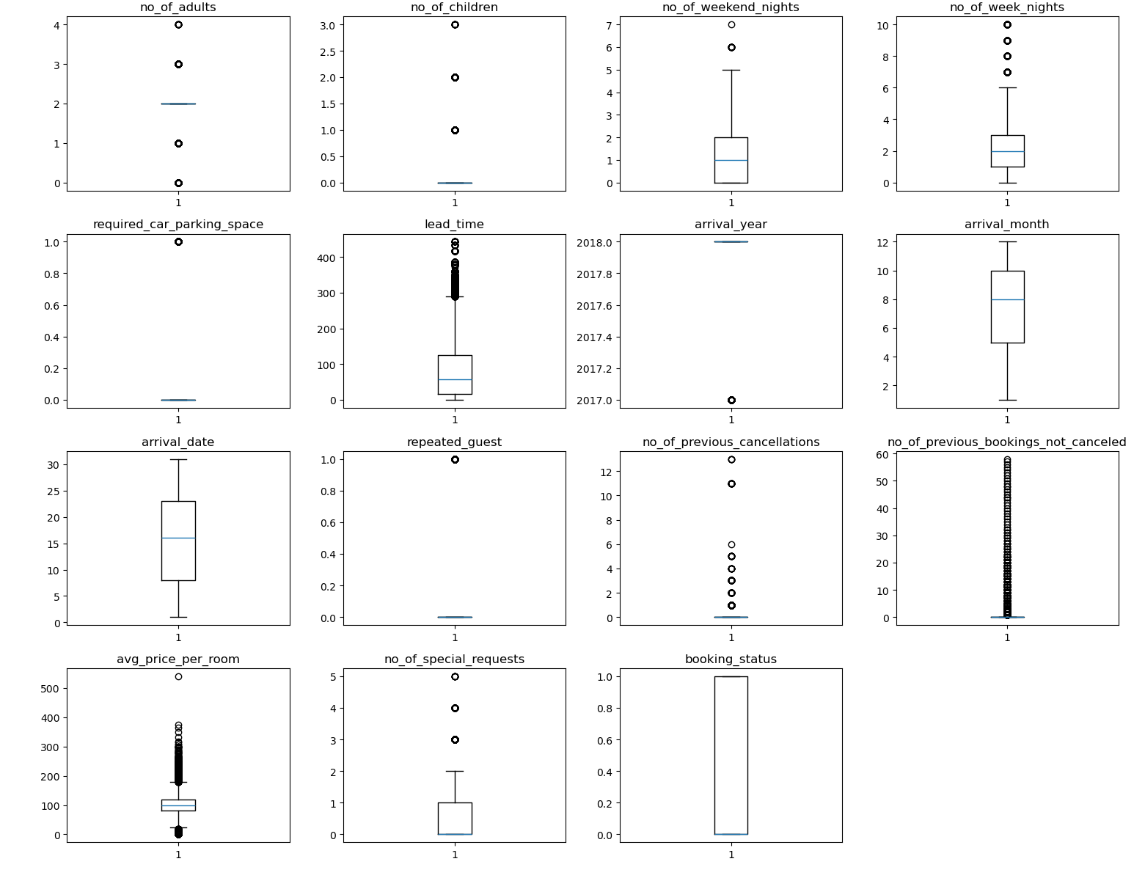
There are no missing values in the dataset therefore no treatment is required.

## 3.2 Duplicate Value Check

There are no duplicate values in the dataset therefore no treatment is required.

## 3.3 Outlier Check

Figure - Outlier Detection using Boxplots



The extreme outliers in “avg\_price\_per\_room” are treated based on the IQR method. The rest of the outliers will be considered as genuine values as they can contribute to providing insights so these outliers will not be needing treatment.

## 3.4 Feature Engineering

The “not canceled” and “canceled” values of the “booking\_status” column in the dataset has been encoded as “0” and “1” respectively. This is needed to build an efficient model.

## 3.5 Data Preparation for Modeling

We want to determine the factors for booking cancellations. Before building the model all categorical features will be encoded. The data will split into training and test data. The model performance will be built on the training data and evaluated on the test data. We will split the data in 70:30 ratio. 70% of the data will be used as training data, 30% of the data will be used as testing data.

# MODEL BUILDING

## 4.1 Model Evaluation Criteria

There are chances the model can make wrong predictions such as

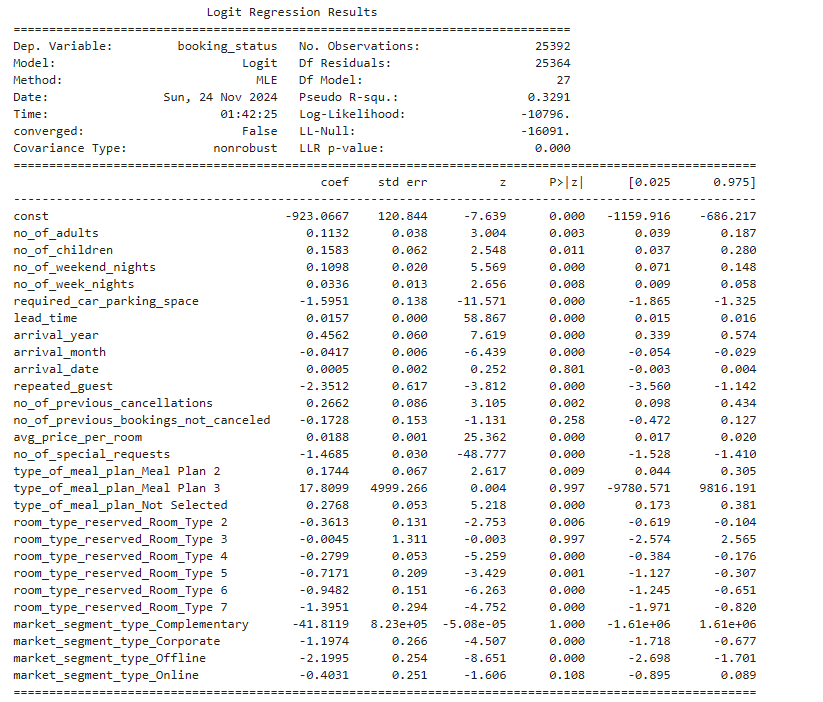
1. Predicting a customer will cancel their booking when they will not which will lead to the hotel potentially overbooking and cause problems during customer arrival.
2. Predicting a customer will not cancel their booking but they will cancel it which will lead to loss of revenue for the hotel.

It is important for the hotel to reduce both false positives and false negatives as they both can negatively impact it. When creating a model, we will optimize the F1 score so that the chances both false positives and false negatives will be reduced.

## 4.2 Logistic Regression Model

Logistic Regression is a statistical model used for binary classification problems, where the goal is to predict one of two possible outcomes (such as 0 or 1, true or false, yes or no). Logistic regression works by modeling the probability of the default class (usually "1" or the positive class) given a set of input features. The model applies a logistic function (also known as the sigmoid function) to the output of a linear equation. This transforms the linear output into a probability that ranges between 0 and 1.

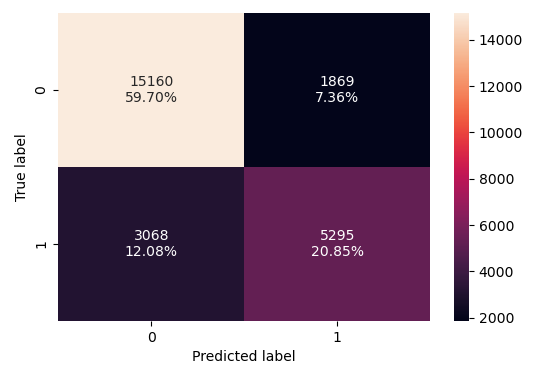
Figure - Logistic Regression Model Results



**Observations:**

* Negative coefficients indicate that as the value of the corresponding attribute increases, the likelihood of booking cancellation decreases.
* Positive coefficients suggest that an increase in the value of the corresponding attribute raises the likelihood of booking cancellation.
* The p-value of a variable helps determine its significance. If the significance level is set at 0.05, variables with a p-value below 0.05 are considered statistically significant.

Figure - Confusion Matrix on Training Data (LR)



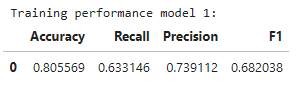
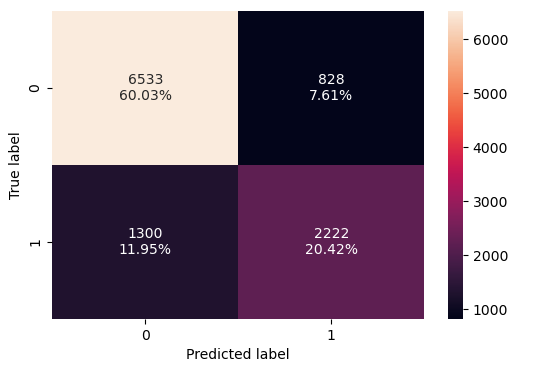
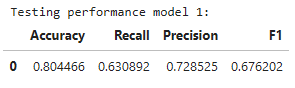


Figure - Confusion Matrix on Test Data (LR)



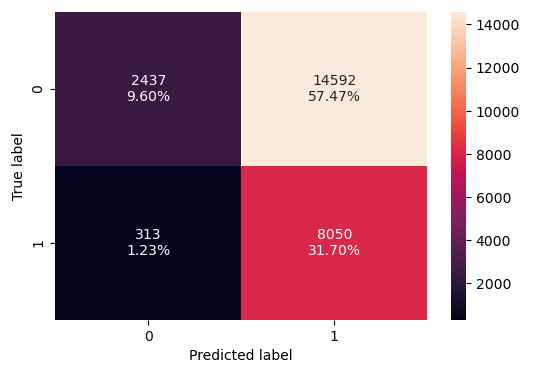


Based on the training and testing results, the F1 score of the model is approximately 0.68, the model will need to be optimized.

## 4.3 Naive-Bayes Classifier Model

The Naive Bayes Classifier is a probabilistic machine learning model used for classification tasks, based on Bayes' Theorem. It assumes that the features are conditionally independent given the class label, which simplifies the computation of probabilities. The model calculates the likelihood of a class based on the given input features and assigns the class with the highest posterior probability. It's efficient, especially for high-dimensional data like text, and performs well even with small datasets.

Figure - Confusion Matrix on Training Data (NB)



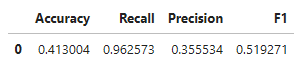
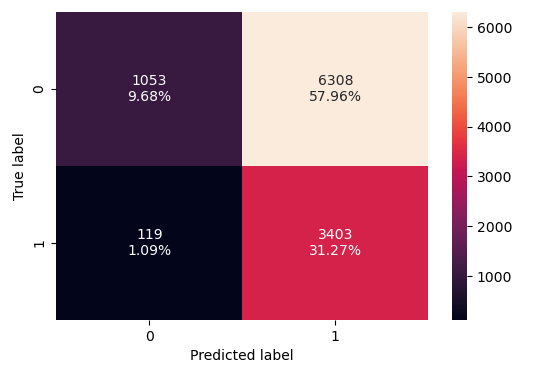
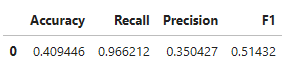


Figure - Confusion Matrix on Test Data (NB)



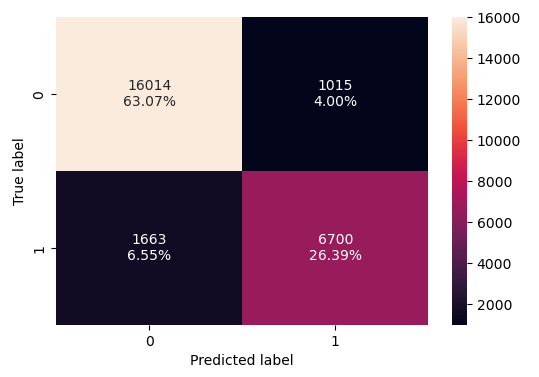


Based on the training and testing results, The F1 score has significantly reduced to around 0.51. This model is not optimal.

## 4.4 KNN Classifier Model

The K-Nearest Neighbors (KNN) classifier is a simple, non-parametric algorithm used for classification tasks. It classifies a data point based on the majority class of its K nearest neighbors in the feature space. The distance between data points is typically measured using Euclidean distance or other distance metrics. KNN makes predictions by looking at the closest training examples to a test point and assigning the most common class among them.

Figure - Confusion Matrix on Training Data (KNN)



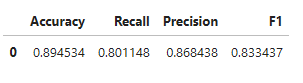
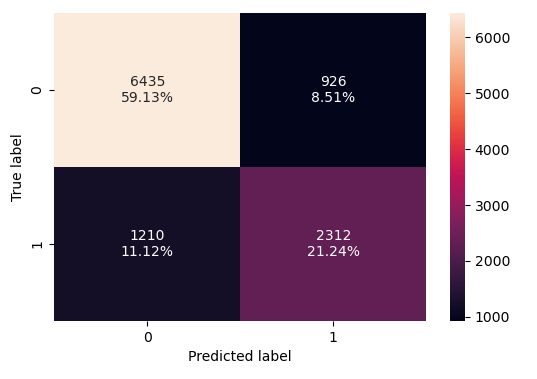
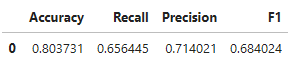


Figure - Confusion Matrix on Test Data (KNN)



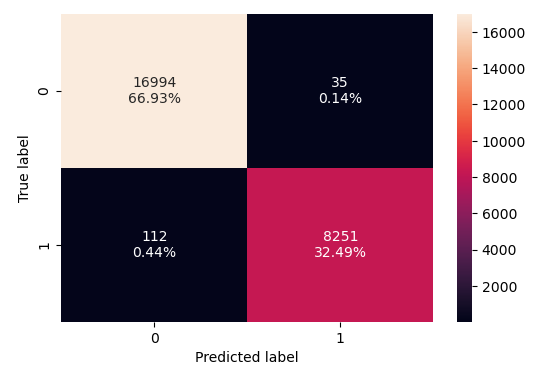


Based on the results, the model performs well on the training data with an F1 score of around 0.83 however performs poorly on the testing data with the F1 score of around 0.68.

## 4.5 Decision Tree Classifier Model

The Decision Tree Classifier is a supervised machine learning algorithm that splits data into subsets based on feature values, creating a tree-like structure. Each internal node of the tree represents a feature, and each branch represents a decision rule, with leaf nodes representing the class labels. It recursively splits the data to maximize information gain or minimize impurity. Decision trees are easy to interpret and visualize, making them highly interpretable models. However, they can be prone to overfitting, especially with complex trees, which can be mitigated with pruning techniques.

Figure - Confusion Matrix on Training Data (DT)



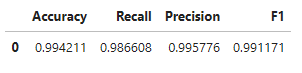
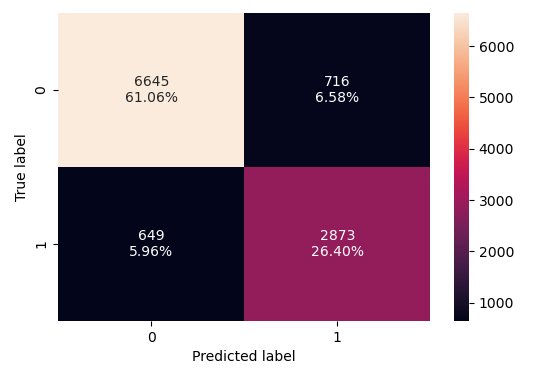
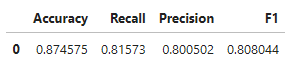


Figure - Confusion Matrix on Test Data (DT)





The decision tree classifier model performs really well on the training data with an F1 score of around 0.99 and the model performance is reduced while testing the test data with an F1 score of around 0.80. This model may be the most optimal so far.

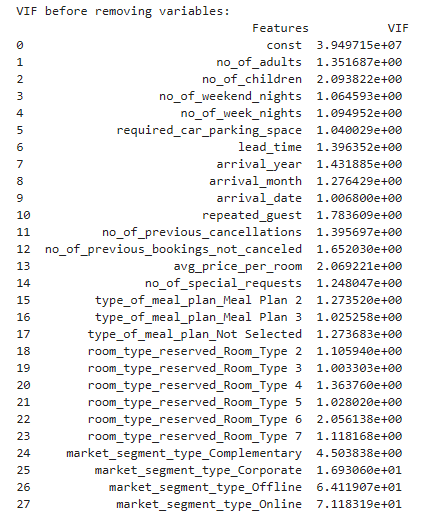
# MODEL PERFORMANCE IMPROVEMENT

## 5.1 Improving Logistic Regression Model

To improve the logistic regression model, we will deal with multicollinearity, remove high p-value variables, determine optimal threshold using ROC curve.

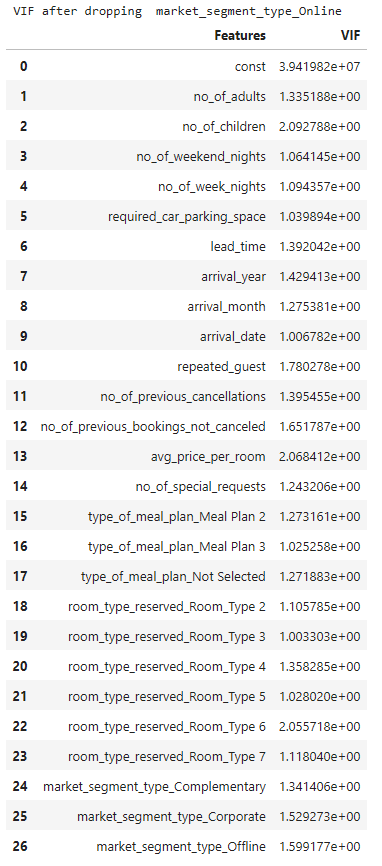
### 5.1.1 Check for Multicollinearity

Figure - VIF Variables on Original Dataset



Based on the VIF values, there are few variables that have a VIF >=5, We will remove each variable one by one and check once more. “market\_segment\_type\_Online” will be dropped first.

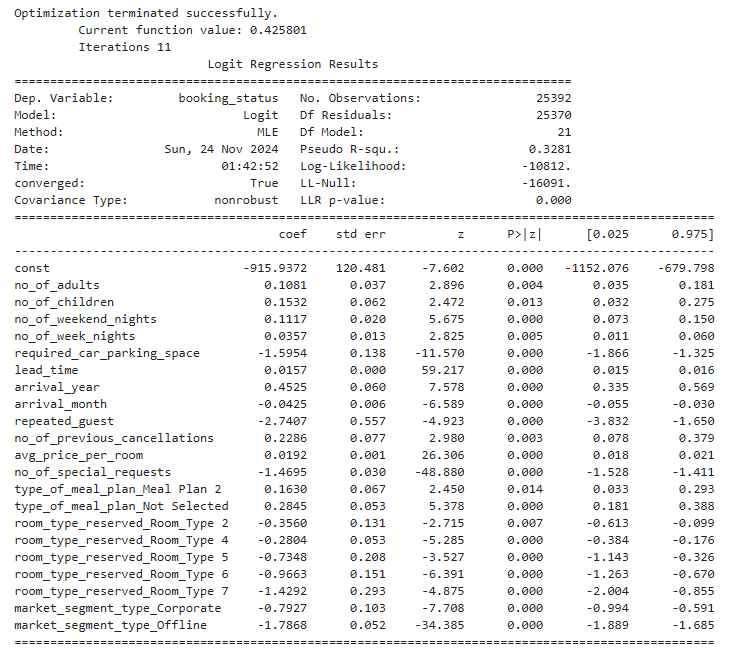
Figure - VIF Variables after dropping Online Market



After dropping “market\_segment\_type\_Online” we can see that all variables have a VIF value of < 5. There is no multicollinearity present in the data.

### 5.1.2 Removing High P-value Variables

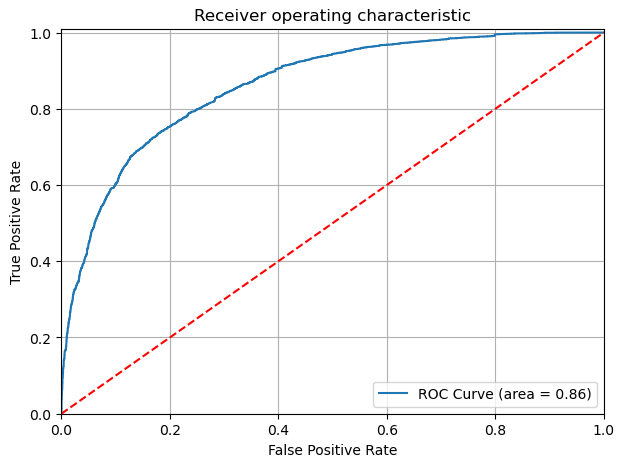
Figure - Logistic Regression Model Post Muticollinearity



All variables have a P-value of < 0.05. There is no need to remove any variables so the dataset will remain as it is.

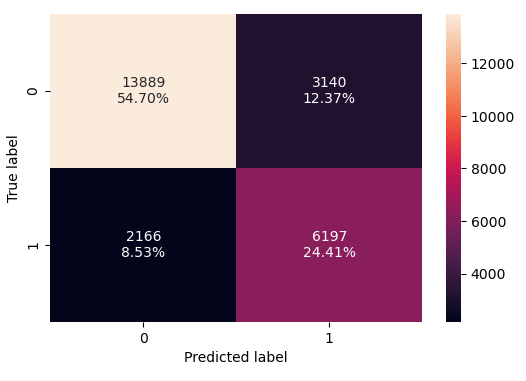
### 5.1.3 Determining Optimal Threshold using ROC Curve

Figure - ROC Training Test



The ROC-AUC Curve has a score of 0.86, The model has a good ability to distinguish between the two classes, as the AUC value is closer to 1 than to 0.5. We obtain a threshold of around 0.36 and we can rebuild the model based on the new corrections.

Figure - Confusion Matrix on Training Data (Tuned LR)



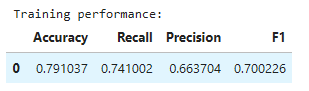
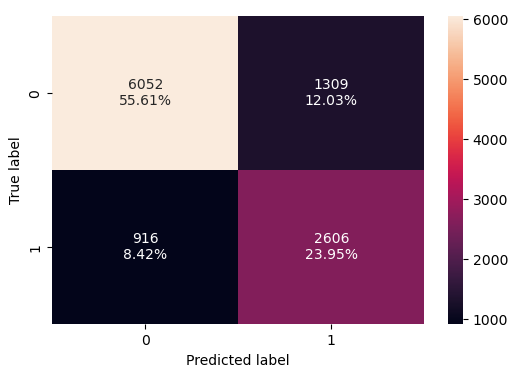
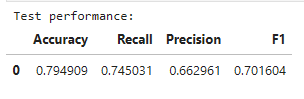


Figure - Confusion Matrix on Test Data (Tuned LR)

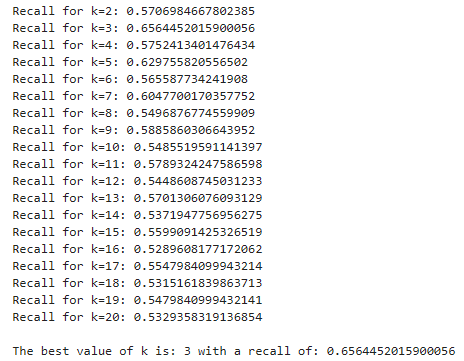




Rebuilding the model based on the improved dataset shows only a slight increase of the F1 score to around 0.70 on both the training and testing data.

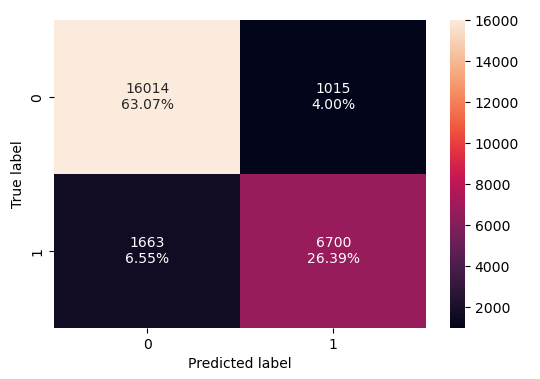
## 5.2 Improving KNN Classifier

The KNN Classifier model can be improved by using different k values, finding the largest k value will help us to reduce overfitting but it cannot be too large as it will lead to underfitting, we will use a threshold of 1-12 and find the largest recall value among them.



From this we can see the best value of k is 3 with a recall of approximately 0.65.

Figure - Confusion Matrix on Training Data (Tuned KNN)



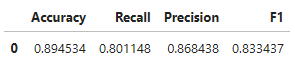
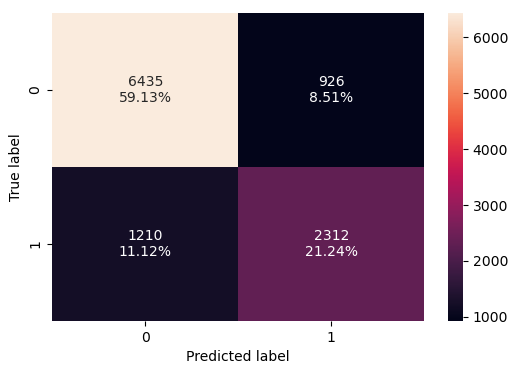
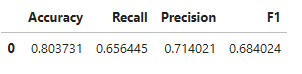


Figure - Confusion Matrix on Test Data (Tuned KNN)





Checking the new tuned KNN Model, we can see that the F1 value performs well on the training data but performs worse on the training data.

## 5.3 Improving the Decision Tree Model

### 5.3.1 Pre-Pruning

Pre-pruning in decision trees refers to the practice of limiting the growth of the tree during its construction to prevent overfitting. This is done by setting constraints such as maximum depth, minimum samples required to split a node, and minimum samples required in a leaf node. By imposing these limits, the tree is prevented from becoming too complex and capturing noise in the training data. The goal is to create a simpler model that generalizes better to unseen data.

Figure - Confusion Matrix on Training Data (Pre-pruned DT)

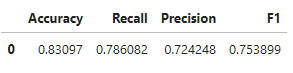
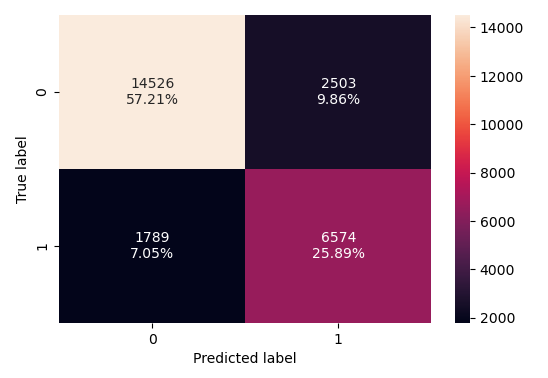
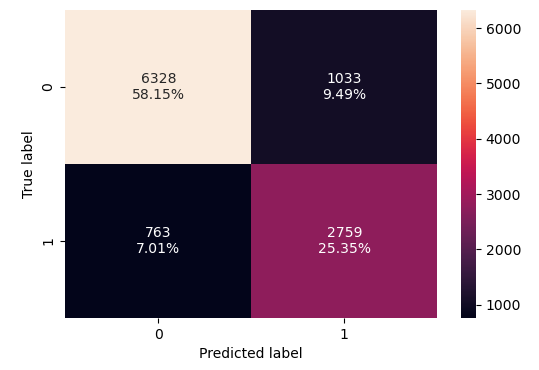
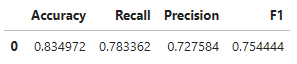


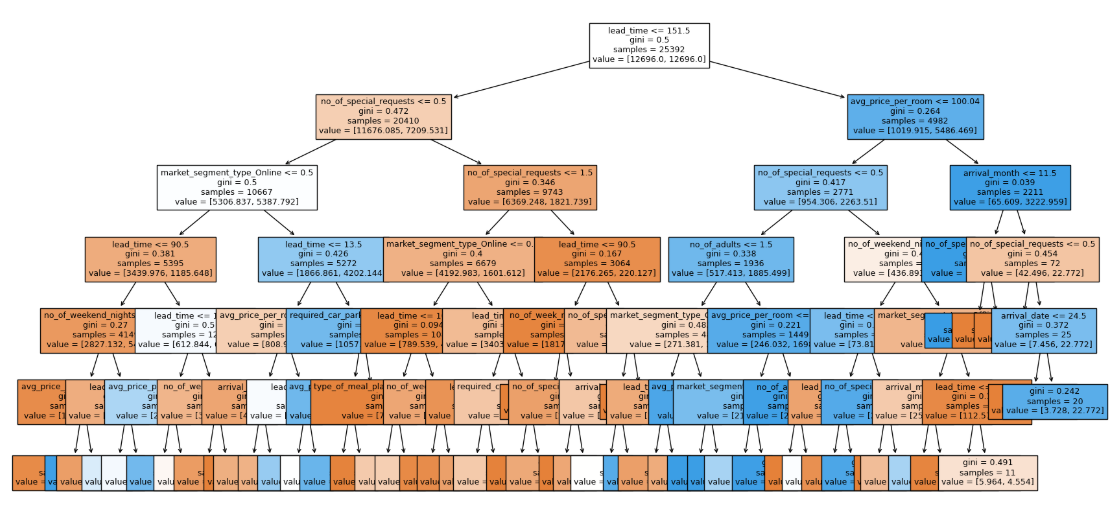
Figure - Confusion Matrix on Test Data (Pre-pruned DT)





Pre-pruning we can see the F1 values are at approximately 0.75 for both the testing and training data which is balanced but overall reduced in its precision to predict cancellations

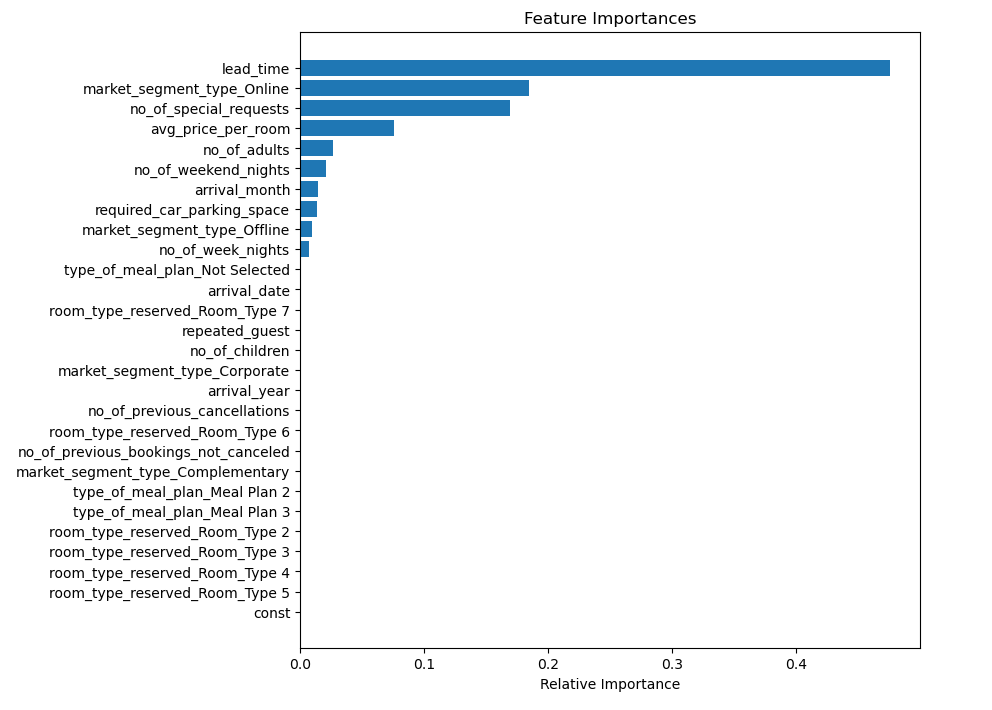
Figure - Pre-pruned Decision Tree



The decision tree is simpler. Lead time plays a major role in determining if the booking is canceled or not, 151.5 days is considered as the minimum threshold the model uses to make the first split. The tree can be interpreted as

* If lead time is less the 151.5 days and if there is at least 1 special request, the booking is less likely to be cancelled. If there were no requests and the booking was done online, there is more chance the booking will get cancelled.
* If lead time is more than 151.5 days and the average room price is less than 100.04 euros and the customer hasn’t made a special request the booking is likely to get cancelled. If the average price is greater than 100.04 euros and the arrival month is November-December, the booking is more likely to not get cancelled, and so on

Figure - Pre-pruned Freature Importances

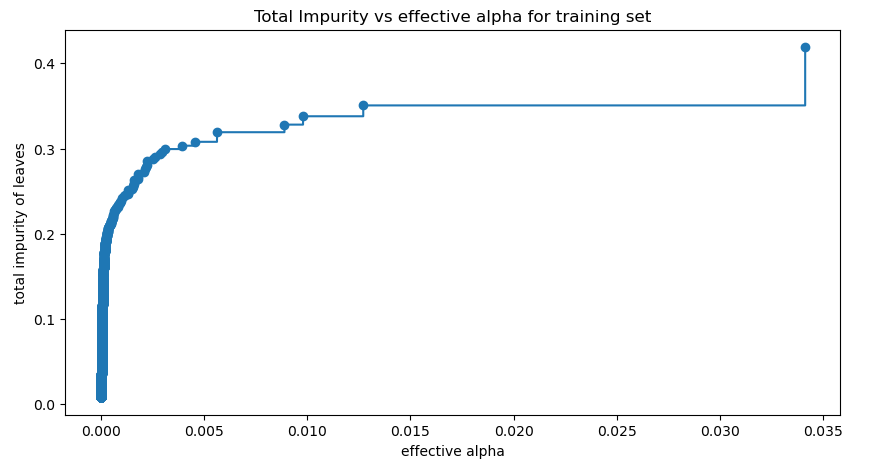


“lead\_time” is the most important feature in the tree building model followed by “market\_segment\_type\_Online”, ‘no\_of\_special\_requests” and “avg\_price\_per\_room”.

### 5.3.2 Post-Pruning

Post-pruning in decision trees involves growing a tree fully and then trimming back branches that don't provide significant predictive power, reducing overfitting. After the tree is built, nodes or subtrees that add little value to model accuracy are removed based on criteria like cross-validation performance. Common methods include cost-complexity pruning, which balances tree size with error reduction. Post-pruning helps to simplify the model and improve generalization to unseen data. The main goal is to avoid overly complex models while retaining essential patterns from the data.

Figure - Total impurity against Effective Alpha



The decision tree will be trained using effective alphas The last value in ccp\_alphas is the alpha value that prunes the whole tree, leaving the tree, clfs[-1], with one node. Number of nodes in the last tree is: 1 with ccp\_alpha: 0.0811.

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

Figure - Number of nodes against Effective Alpha

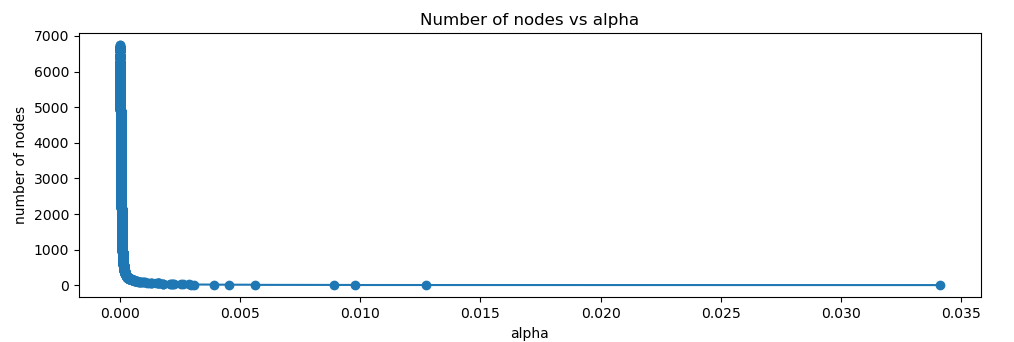


Figure - Depth against Effective Alpha

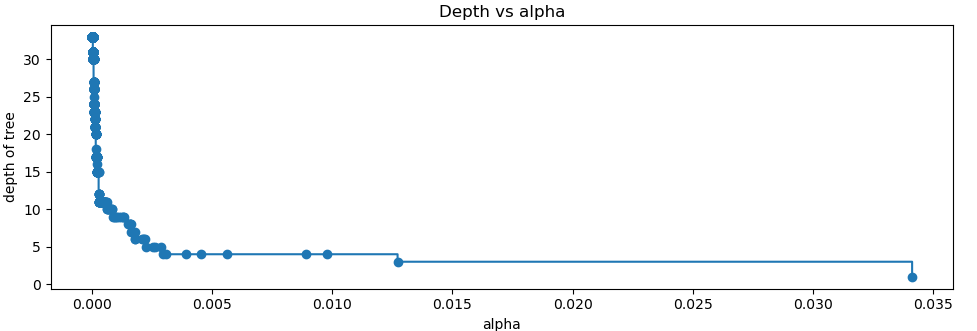
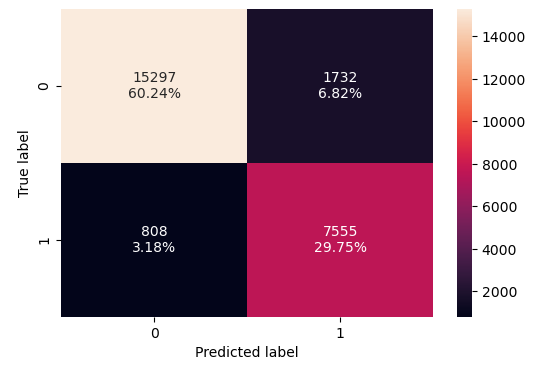


Figure - F1 Score against Training and Testing Data



Figure - Confusion Matrix on Training Data (Post-pruned DT)



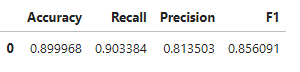
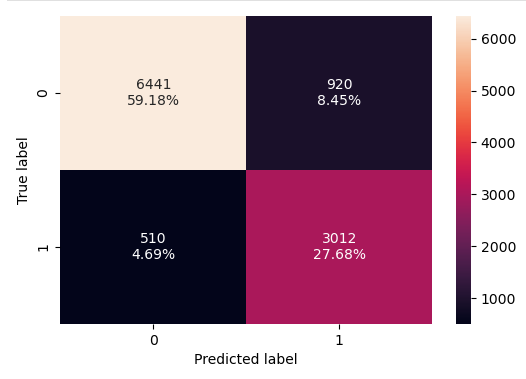
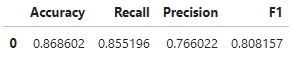


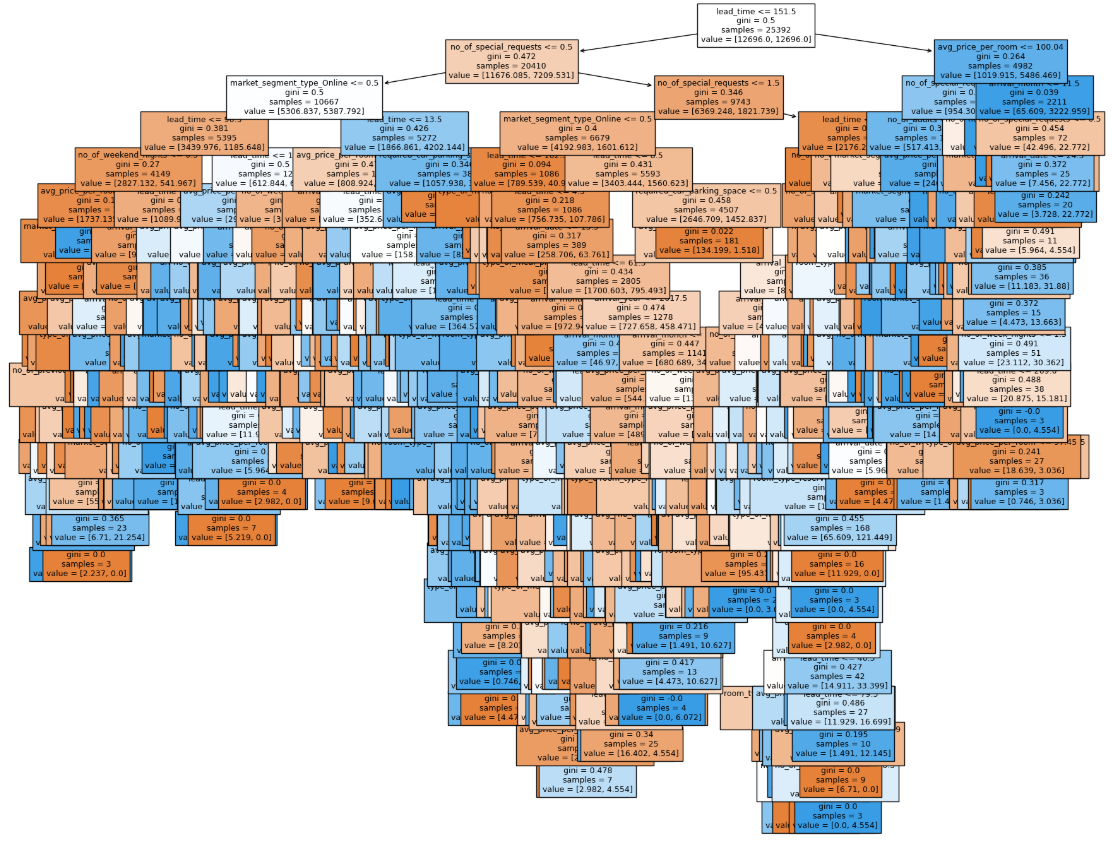
Figure - Confusion Matrix on Test Data (Post-pruned DT)





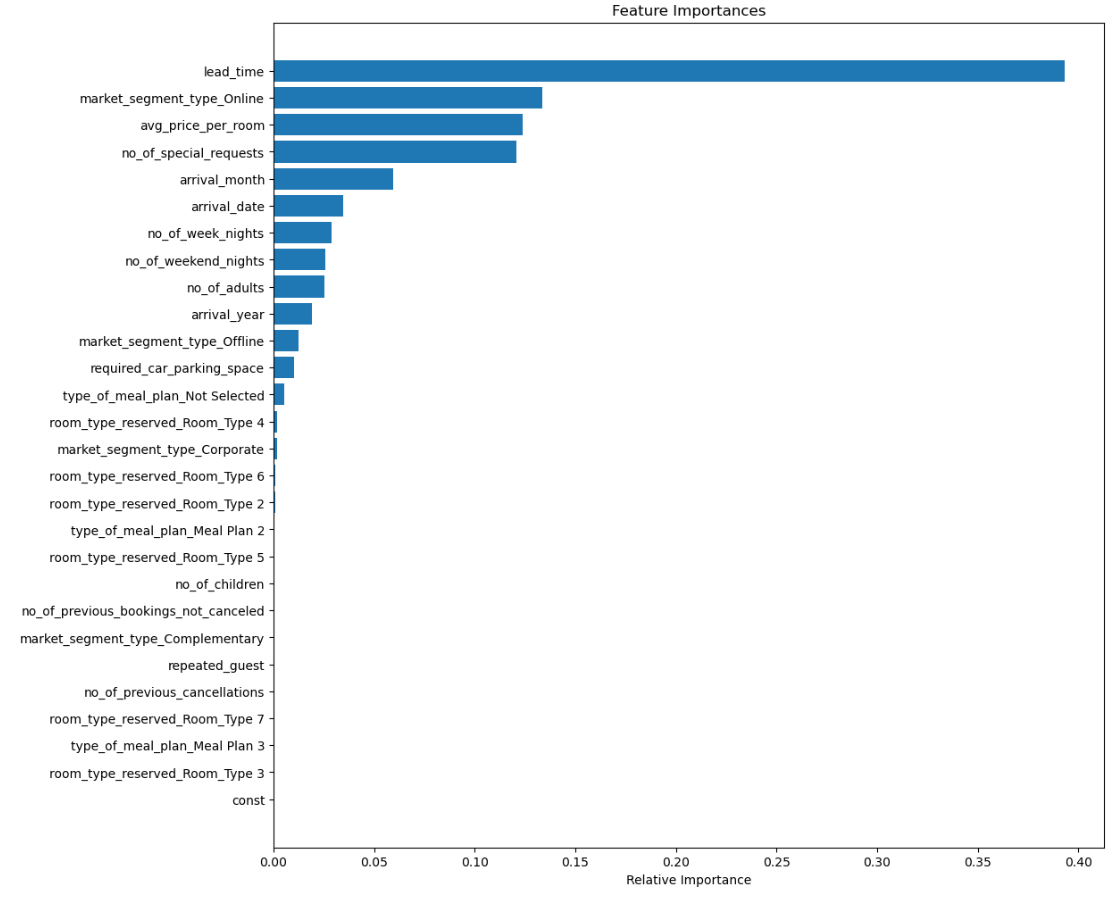
After pruning the F1 scores as well as recall and precision are more generalized and balanced. The model gives an F1 score of 0.85 in the training data and 0.80 in the testing data.

Figure - Post-pruned Decision Tree



The decision tree is more complex when compared to the decision tree from pre-pruning. The root node of the tree begins similar to the pre-pruned decision tree root node by considering lead time of 151.5 days as the threshold to begin the first split. All the decisions made at each node of the decision tree which determines if a booking is cancelled or not can be viewed by examining the rules of the decision tree.

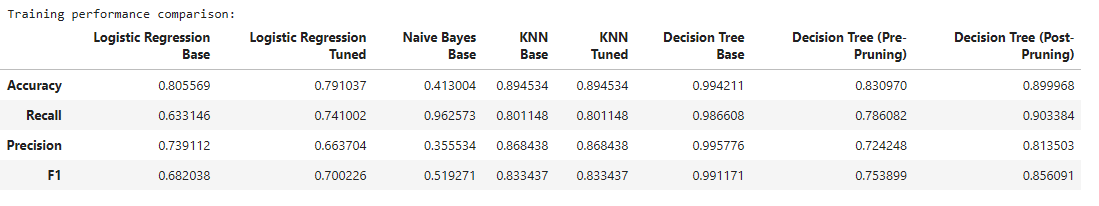
Figure - Post Pruned Feature Importances

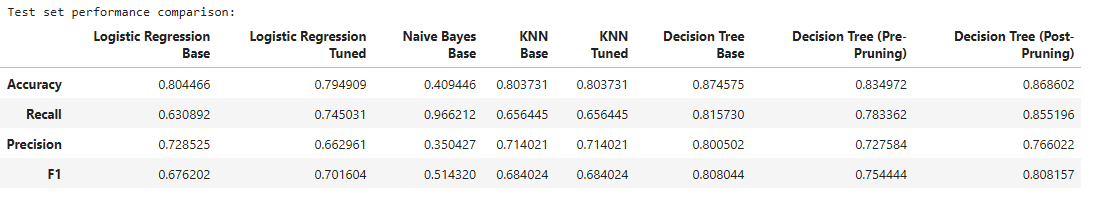


The feature importances are the same as before “lead\_time”, “market\_segment\_type\_Online”, ‘no\_of\_special\_requests” and “avg\_price\_per\_room” are the most important. These have not changed significantly from the important features of pre-pruning model.

# FINAL MODEL COMPARISON AND SELECTION

Figure – Training & Testing Data Model Comparison





Comparing all the models we can infer the following

* Logistic regression model gives a lower F1 value at around 0.68 making the model not a reliable predictor of booking cancellations.
* The Tuned logistic regression only performs slightly better at 0.70 still making the model not reliable.
* Naïve bayes model performs the worst on both training and testing data with the F1 score at around 0.51
* The KNN model performs well on the training data with an F1 score of 0.83 but performs worse on the test data, making it not a reliable model.
* The tuned KNN model performs almost exactly the same the base KNN model making it unreliable.
* The decision tree model performs really well in the training data with an F1 score of 0.99 but performs worse on the test data with a significant difference.
* The pre-pruned decision tree model has reduced overfitting but a lower F1 score with around 0.78 and a lower Recall and Precision value as well.
* The post-pruned decision model has the best overall performance with a high F1 score of 0.85 and a balanced Recall and Precision value.

We can conclude that the Decision Tree post pruned model is the best model to predict whether a booking will be cancelled by the customer.

# ACTIONABLE INSIGHTS & RECOMMENDATIONS

* The most important factors to consider are lead time and average price per room, a lead time of over 151 days and a higher average price for the room shows that there is a likely chance of the booking being cancelled.
* The hotel could keep a limited timer on allowing for free cancellations, if the cancellation occurs after the timer, then a fee will be imposed on the customer for cancelling the booking.
* The hotel can also offer a room at a cheaper rate with the compromise that it is non-refundable, this will incentive customers to not cancel their bookings with the benefit that they get the room at a cheaper rate.
* Repeat customers tend to have a lesser chance of cancelling bookings, the hotel can work to familiarize and interact with previous customers who left on a satisfied note preferable through online promotion materials to entice them to rebook with the hotel.
* The data shows that there is seasonal pattern in the booking behavior, during spring – summer time the bookings increase until October following which there is a decline until the next year, The hotel can concentrate on hiring more staff during the peak seasons to ensure there is maximum guest satisfaction encouraging them to become repeat customers.
* The hotel can send reminder emails to customers with tailored offers for bookings with long lead times, such as special offers or added amenities such as “free breakfast” as it is the most picked option, to motivate customers to stay committed.
* Corporate segments tend to have the least cancellations, offering special rates to this segment along with extra amenities or services to convert them into repeat customers.
* Customers with more special requests tend to not cancel their bookings, the hotel can push to offer these requests to be fulfilled on arrival for all customers so that they are satisfied and are less likely to cancel the booking.
* Higher average price per room contributes to increased chances of cancellation, This could be due to rooms available at a cheaper rate in another place, The hotel can study the market and ensure their prices are kept competitive to reduce chances of cancellation due to price.
* Room Type 1 tends to be the most popular choice among the customers, bundling the room with other offers such as breakfast plan can incentivize the customer to not cancel their booking.