MACHINE LEARNING - I
BUSINESS REPORT

Submitted by

Dr. JEMIMAH J P P

BUSINESS REPORT

Contents

List of figures

List of Tables

1. EXPLORATORY DATA ANALYSIS (EDA)

1.1 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 impact 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.

1.2 Objective

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be cancelled. 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 analyse 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 cancelled in advance, and help in formulating profitable policies for cancellations and refunds.

1.3 Data description and information

The data set from INN Hotels Group informs about the booking information, details about the customers, and their preferences of rooms. Due to change of plans, scheduling conflicts etc., the Hotels Group incur last-minute cancellations which leads to loss in revenue, additional expenses on distribution, reduction in profit margin and so on. Thus a Machine Learning based solution is desirable to predict the bookings that are likely to be cancelled. From a data scientist view, the provided dataset can be used to analyse various factors that influence booking cancellations, predict cancellations in advance, and formulate policies for cancellations and refunds. The information about the different variables mentioned in the data set is elaborated in Table 1.

Information

Predictor Variables	**Description**
Booking_ID	the unique identifier of each booking
no_of_adults	Number of adults
no_of_children	Number of Children

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

Table 1: Variables and its description

1.4 Data overview

The necessary packages need to be imported, the working directory is set and the data file is loaded to understand and describe the overview of the provided dataset.

Displaying the first few rows and last few columns of the dataset

The dataset consists of 36275 rows and 19 columns. The 36275 rows represents the booking status of customers who visit the INN Hotels. The 18 columns that give the details on

various driving factors are no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, type_of_meal_plan, required_car_parking_space, room_type_reserved, lead_time, arrival_year, arrival_month, arrival_date, market_segment_type, repeated_guest, no_of_previous_cancellations, no_of_previous_bookings_not_cancelled, avg_price_per_room, no_of_special_requests. These 18 columns drive the target variable, the booking_status. The "Booking_ID" column shows the unique identification number given to every customer who books rooms at INN Group Hotels and this column has no role to play in model prediction, so is not considered as a driving factor.

Tables 1and 2 show the details of the list of first and last five contents available in the dataset of the ShowTime service provider respectively.

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arrival_year	arrival_month	arrival_date	market_segment_type	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
0	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	2017	10	2	Offline	0	0	0	65.00000	0	Not_Canceled
1	INN00002	2	0	2	3	Not Selected	0	Room_Type 1	5	2018	11	6	Online	0	0	0	106.68000	1	Not_Canceled
2	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	2018	2	28	Online	0	0	0	60.00000	0	Canceled
3	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	2018	5	20	Online	0	0	0	100.00000	0	Canceled
4	INN00005	2	0	1	1	Not Selected	0	Room_Type 1	48	2018	4	11	Online	0	0	0	94.50000	0	Canceled

Table 2: Top five rows of the dataset

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arrival_year	arrival_month	arrival_date	market_segment_type	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
36270	INN36271	3	0	2	6	Meal Plan 1	0	Room_Type 4	85	2018	8	3	Online	0	0	0	167.80000	1	Not_Canceled
36271	INN36272	2	0	1	3	Meal Plan 1	0	Room_Type 1	228	2018	10	17	Online	0	0	0	90.95000	2	Canceled
36272	INN36273	2	0	2	6	Meal Plan 1	0	Room_Type 1	148	2018	7	1	Online	0	0	0	98.39000	2	Not_Canceled
36273	INN36274	2	0	0	3	Not Selected	0	Room_Type 1	63	2018	4	21	Online	0	0	0	94.50000	0	Canceled
36274	INN36275	2	0	1	2	Meal Plan 1	0	Room_Type 1	207	2018	12	30	Offline	0	0	0	161.67000	0	Not_Canceled

Table 3: Bottom five rows of the dataset

Checking the data types of the columns for the dataset

The dataset consists of 14 numerical columns and 5 object type columns. The no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, lead_time, arrival_year, arrival_month, arrival_date, no_of_previous_cancellations, avg_price_per_room, no_of_special_requests, no_of_previous_bookings_not_cancelled are the numerical columns of the dataset.

The type_of_meal_plan, room_type_reserved, market_segment_type, and booking_status are the object type columns in the dataset. The repeated_guest and required_car_parking_space columns describe the details if any guests have booked rooms repeatedly and if any guest require car parking facility respectively, and it is being read as integer type. But based on its description these columns should reveal the category and it can be preferably in categorical format. From the information obtained it is observed that there is no missing values in the 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 4: Information about the columns of the dataset

Checking for missing values

The table 5 shows that the provided dataset does not contain any missing values.

```
Booking_ID                            0
no_of_adults                          0
no_of_children                        0
no_of_weekend_nights                  0
no_of_week_nights                     0
type_of_meal_plan                     0
required_car_parking_space            0
room_type_reserved                    0
lead_time                             0
arrival_year                          0
arrival_month                         0
arrival_date                          0
market_segment_type                   0
repeated_guest                        0
no_of_previous_cancellations          0
no_of_previous_bookings_not_canceled  0
avg_price_per_room                    0
no_of_special_requests                0
booking_status                        0
dtype: int64
```

Table 5: Checking for missing values

Checking for duplicate values

- It is also observed that there are no duplicate entries in the dataset.

Unique values in the attributes of the dataset

no_of_adults	5
no_of_children	6
no_of_weekend_nights	8
no_of_week_nights	18
type_of_meal_plan	4
required_car_parking_space	2
room_type_reserved	7
lead_time	352
arrival_year	2
arrival_month	12
arrival_date	31
market_segment_type	5
repeated_guest	2
no_of_previous_cancellations	9
no_of_previous_bookings_not_canceled	59
avg_price_per_room	3930
no_of_special_requests	6
booking_status	2
dtype: int64	

Table 6: Unique values in the dataset

- Among the variables in the dataset, lead_time, avg_price_per_room, and no_of_previous_bookings_not_canceled have the highest counts of unique values.

Statistical summary of the numerical columns of the dataset

The table 6 shows the statistical summary of the numerical columns present in the data set

	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 7: Description of the numerical columns of the dataset

Observations

- **no_of_adults:** As the mean and median values are close to 2, it is observed that most of the bookings have 1 or 2 adults. The maximum number of adults in a booking is 4.

- **no_of_children:** On an average, only few children accompany the guests, showing a mean of approximately 0.1. The maximum number of children in a booking is 10, although 75% of bookings have no children.

- **no_of_weekend_nights:** The average number of weekend nights booked is approximately 0.8, 50th percentile is one weekend night, 75th percentile is 2 weekend nights, while the maximum is 7.

- **no_of_week_nights:** On an average, guests book for a stay of approximately 2.2 weeknights. The maximum number of weeknights in a booking is 17, which shows that some customers prefer staying for long weeknights.

- **required_car_parking_space:** Looking into the 25th, 50th and 75th percentile values, we see that this predictor variable is of the form of object/ categorical variable read in integer format.

- **lead_time:** The average lead time between booking and arrival is approximately 85 days, with the maximum of 443 days. The mean value is greater than the median (57 days) and this indicates that the distribution of lead_time is right skewed.

- **arrival_year:** As the 25th, 50th, 75th percentile and the maximum value of this column being `2018`, and also the min value being `2017` it can be understood that this dataset was observed between `2017-2018`. It can also be noted that majority of the observations were from `2018`.

- **arrival_month:** From the 25th and the 75th quartile, we see that INN hotel groups have their bookings spread throughout the year, with the mean arrival month being `JULY-AUGUST`, for the mean value shows 7.4.

- **arrival_date:** From the minimum(1) and maximum(31) values, it can be observed that the bookings have been taken place throughout the month while the average arrival date falls around the middle of the month, for the mean value is 15.6.

- **repeated_guest:** Looking into the 25th, 50th and 75th percentile values, we see that this predictor variable is of the form of object/ categorical variable read in integer format.

- **no_of_previous_cancellations:** Most of the bookings have no previous cancellations, as read by the mean and median both being 0. However, the mean number of previous cancellations is 0.023. It is also noted that for a maximum 13 bookings, the customer cancelled in prior to the current booking.

- **no_of_previous_bookings_not_canceled:** On an average approximately 0.15 bookings were not cancelled in prior. For most of the bookings the customer showed up, as read by the mean and median both being 0. It is also noted that for a maximum of 58 bookings, the customer showed up by not cancelling in prior to the current booking.

- **avg_price_per_room:**
  - The average price per room is approximately €103.4, with a standard deviation of €35.1.
  - The minimum price is €0, indicating some bookings are given for free or had special offer pricing.
  - The maximum price is €540, indicating the presence of bookings for high-priced rooms.

- **no_of_special_requests:** On an average, guests make approximately 0.62 special requests per booking, with a maximum of 5 special requests.

The distribution of these predictor variables can be best understood using a box plot and histograms. Their impact against categorical variables is also visualized using bar plots, line plots etc. All the binary variables can be treated as objects/categories.

Statistical summary of the categorical/ object columns of the dataset

	count	unique	top	freq
Booking_ID	36275	36275	INN00001	1
type_of_meal_plan	36275	4	Meal Plan 1	27835
required_car_parking_space	36275	2	0	35151
room_type_reserved	36275	7	Room_Type 1	28130
market_segment_type	36275	5	Online	23214
repeated_guest	36275	2	0	35345
booking_status	36275	2	Not_Canceled	24390

Table 8: Description of the categorical columns of the dataset

Checking for anomalous values in categorical variables

The unique values are determined for each categorical variable to check if any junk/garbage values present in the dataset. This check helps us to identify if any data entry issues are present. From the determined unique values it's concluded that there is no data entry issues present.

```
Booking_ID
INN00001    1
INN24187    1
INN24181    1
INN24182    1
INN24183    1
           ..
INN12086    1
INN12085    1
INN12084    1
INN12083    1
INN36275    1
Name: count, Length: 36275, dtype: int64
----------------------------------------
type_of_meal_plan
Meal Plan 1     27835
Not Selected     5130
Meal Plan 2      3305
Meal Plan 3         5
Name: count, dtype: int64
----------------------------------------
required_car_parking_space
0    35151
1     1124
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
----------------------------------------
repeated_guest
0    35345
1      930
Name: count, dtype: int64
----------------------------------------
booking_status
Not_Canceled    24390
Canceled        11885
Name: count, dtype: int64
----------------------------------------
```

Table 9: Value counts of the categorical variables of the dataset

Observations

- **Booking_ID:** `Booking_ID` can be dropped as it has 36275 unique values, thus would not add any significance to our analysis.

- **type_of_meal_plan:** There are 4 types of meal plans, and from the unique values it is seen that the four types are `Not Selected`, `Meal Plan 1`, `Meal Plan 2` and `Meal Plan 3`. It is noted that nearly 27835 bookings preferred `Meal Plan 1`.

- **required_car_parking_space:** It has two options as `0:No` and `1:Yes`. It is noted that a majority of 35151 bookings did not require a car parking space.

- **room_type_reserved:** There are 7 types of rooms available for bookings. Nearly 28130 bookings preferred for `Room_Type 1`.

- **market_segment_type:** There are 5 types of market segments through which bookings have been done, of which `Online` market segment has attracted about 23214 bookings.

- **repeated_guest:** Most of the guests, nearly 35345 havenot done their bookings repeatedly. We can say them as one time guest to INN Hotels group.

- **booking_status:** Majority of the booking status, nearly 24390 bookings remain not cancelled. Whereas we see nearly 11885 bookings remain cancelled, i.e., about 32.7% of bookings get cancelled.

It is also noted that there is no anomalous values in these categorical variables. It is also seen that the `"Booking_ID"` column can be dropped before proceeding with the Univariate and Bivariate analysis

1.5 Univariate analysis

The univariate analysis is carried out to explore all the variables and their distributions are observed. Generally, histograms, boxplots, countplots, etc. are used for univariate exploration. The categorical variables are explored using countplots and the numerical variables are explored using histograms and boxplots respectively. For the given dataset, it is observed that all numerical columns cannot be studied using histograms and box plots, as they don't have continuous numbers. Numerical columns with distinct integer values can be visualized clearly using a bar plot rather than by a histogram/ box plot. Accordingly this section analyses certain numerical columns using histogram and box plots and some using bar plots.

Numerical variables

- no_of_adults
- no_of_children
- no_of_weekend_nights
- no_of_week_nights
- lead_time
- arrival_year

- arrival_month
- arrival_date
- no_of_previous_cancellations
- no_of_previous_bookings_not_canceled
- avg_price_per_room
- no_of_special_requests

no_of_adults

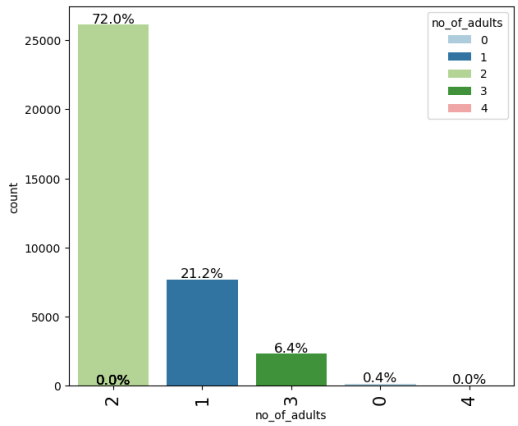
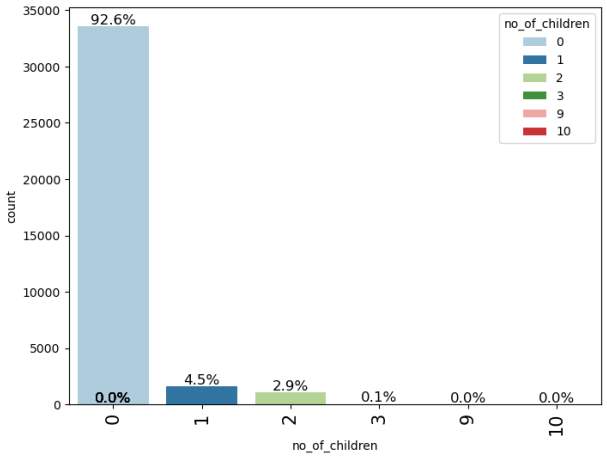


Figure 1: Labelled bar plot for no_of_adults

Observation:

- 72% of the bookings include 2 adults.
- 21.2% of bookings are done by just 1 adult
- 6.4% of bookings include 3 adults
- 0.4% of bookings do not include adults

no_of_children



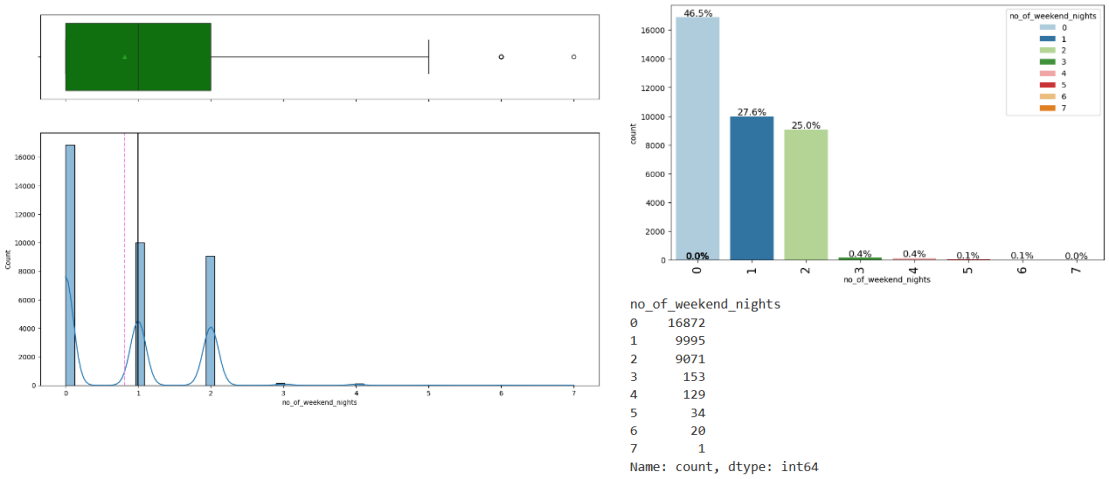

```
no_of_children
0      33577
1       1618
2       1058
3         19
9          2
10         1
Name: count, dtype: int64
```

Figure 2: Labelled bar plot and value counts for no_of_children

Observation:

- 92.6% of the bookings do not include any children,
- 4.5% of bookings include one child
- 2.9% of bookings include two children
- 0.1% of (~19)bookings included three children
- One of the rarest bookings had 10 children

no_of_weekend_nights



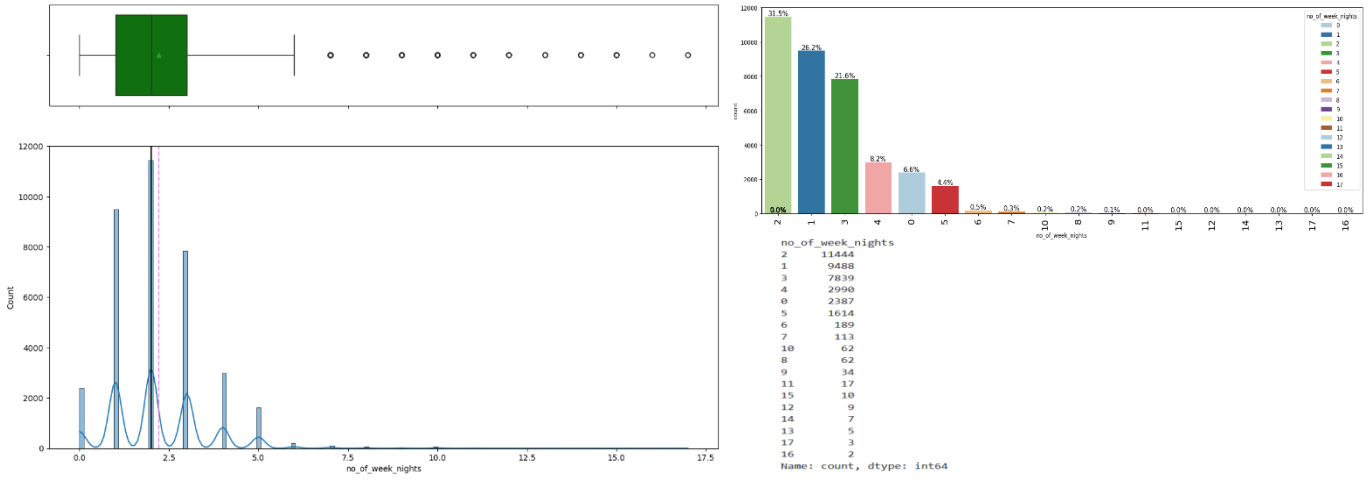

```
no_of_weekend_nights
0    16872
1     9995
2     9071
3      153
4      129
5       34
6       20
7        1
Name: count, dtype: int64
```

Figure 3: Histogram, Box plot, labelled bar plot and value counts for no_of_weekend_nights

Observation:

- 46.5% of the bookings did not include weekend nights.
- 27.6% of bookings included one weekend night.
- 25% of bookings included two weekend nights.
- 0.4% of (~153)bookings included three weekend nights.
- One of the rarest bookings were done for 7 weekend nights.

no_of_week_nights




```
no_of_week_nights
2     11444
1      9488
3      7839
4      2990
0      2387
5      1614
6       189
7       113
10       62
8        62
9        34
11       17
15       10
12        9
14        7
13        5
17        3
16        2
Name: count, dtype: int64
```

Figure 4: Histogram, Box plot, labelled bar plot and value counts for no_of_week_nights

Observation:

- 31.5% of the bookings did not included two week nights
- 26.2% of bookings included one week night.
- 21.6% of bookings included three week nights.
- 8.2% of (~2990) bookings included four week nights.
- 6.6% of (~2387) bookings did not include any week nights.
- Two of the rarest bookings were done for 16 week nights.

lead_time

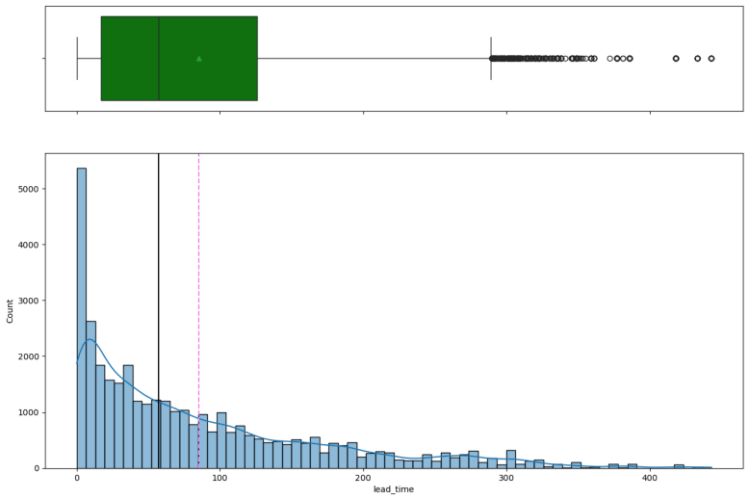


Figure 5: Histogram and Box plot for lead_time

Observation:

- Here it is noticed that the mean greater than the median and the distribution has a longer right tail, hence right skewed.
- Most of the bookings, nearly more than 5000 bookings have a zero lead time and at rare cases there are bookings which has a larger lead time say 400+ days.
- On an average, the number of days between the booking and arrival date is 85. This observation suggests that, on an average, guests make their reservations in prior to three months of their arrival.
- 75% of the bookings have a lead time of 126 or less days, which suggests that majority of customers tend to reserve rooms within four months of their arrival date.

arrival_year

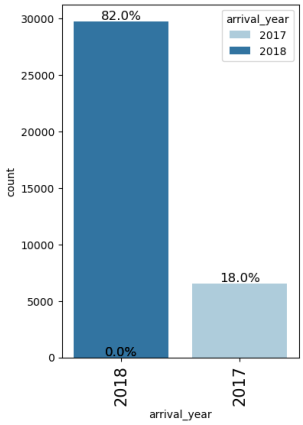
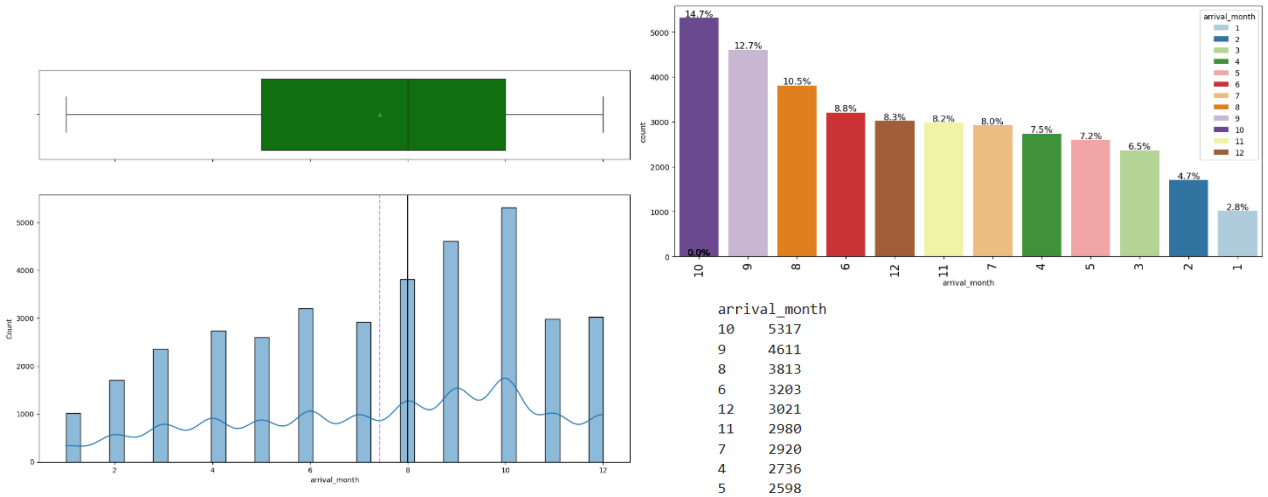


Figure 6: Labelled bar plot for arrival_year

Observation:

- Nearly 82% of the bookings were done during 2018.
- And 18% of the bookings were done during 2017

arrival_month




```
arrival_month
10    5317
9     4611
8     3813
6     3203
12    3021
11    2980
7     2920
4     2736
5     2598
3     2358
2     1704
1     1014
Name: count, dtype: int64
```

Figure 7: Histogram, Box plot, labelled bar plot and value counts for arrival_month

Observation:

- It is observed that the number of bookings starts increasing from August (10.5%) and attains a peak during October (14.7%).
- After October, the number of bookings gradually decreases, but November and December still have considerable bookings of around 8.2% and 8.3% respectively, which is similar as during June & July.
- After December, during January and February the bookings are low, with January being the lowest with 2.8%. This may be related to winter season being unfavourable for tours and visits.

arrival_date

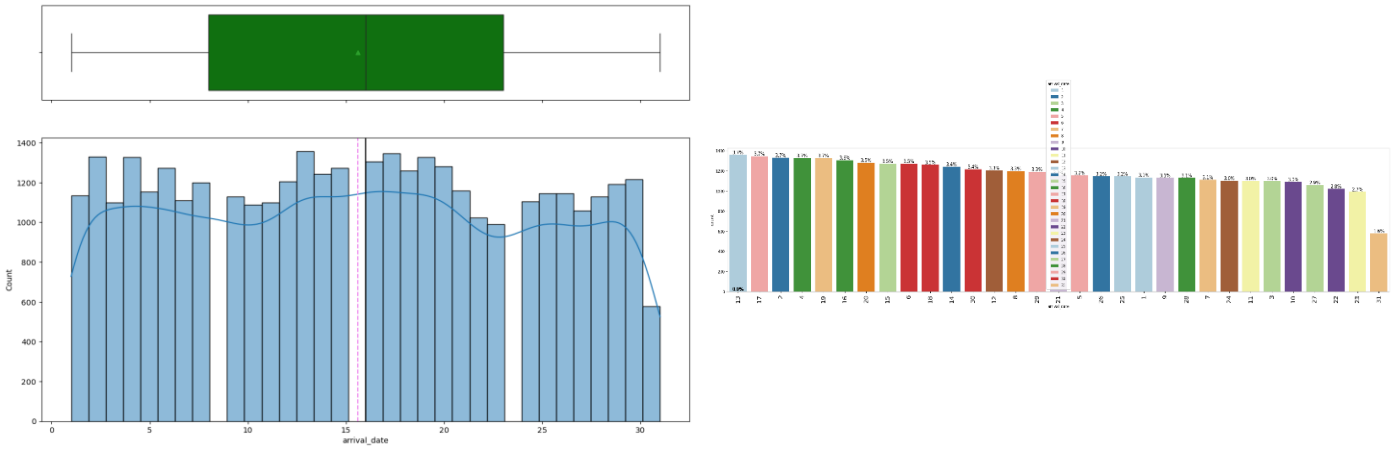
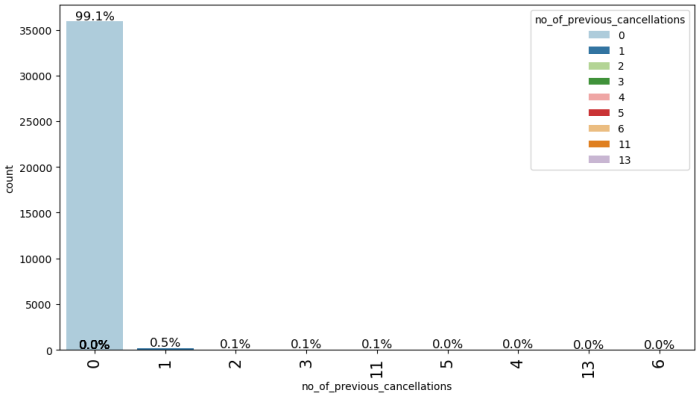


Figure 8: Histogram, Box plot and labelled bar plot for arrival_date

Observation:

- The distribution of arrival dates appears to be fairly uniform, with small differences between them.
- But the bookings on 31st of a month is only 1.6%.
- This lower percentage is due to the fact that there are 31 days in only 7 months of a year, thus showing about half of the percentile compared to other dates in a month.
- Among the arrival dates, the 13th has the highest bookings of about 3.7%. Similarly 17th, 2nd, 4th and 19th, have most of the bookings done.

no_of_previous_cancellations



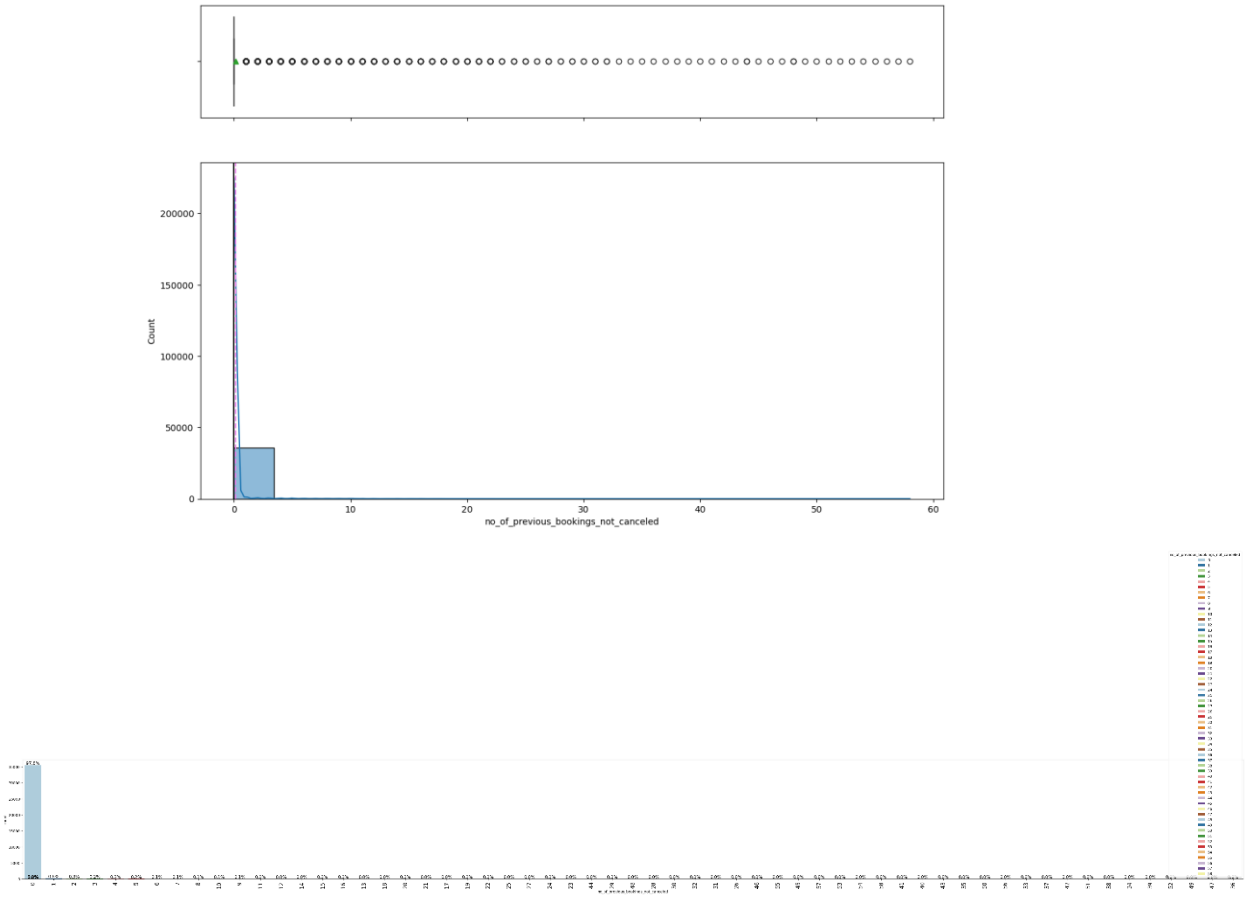

```
no_of_previous_cancellations
0     35937
1       198
2        46
3        43
11       25
5        11
4        10
13        4
6         1
Name: count, dtype: int64
```

Figure 9: Labelled bar plot and value counts for no_of_previous_cancellations

Observation:

- Nearly 99.1% of current booking customers have not cancelled their bookings earlier
- 0.5% of current booking customers (~198) have cancelled their previous bookings once
- Approximately 43 to 46 current booking customers have cancelled their previous bookings twice or thrice
- Very few current booking customers have cancelled their previous bookings more than thrice.

no_of_previous_bookings_not_canceled



no_of_previous_bookings_not_canceled	Total count	Percentage
0	35463	97.8%
1	228	0.6%
2	112	0.3%
3	80	0.2%
4	65	0.2%
5	60	0.2%
6	36	0.1%
7	24	0.1%
8	23	0.1%
10	19	0.1%
9	19	0.1%

Figure 10: Histogram, Box plot, labelled bar plot and value counts for no_of_previous_bookings_not_canceled

Observation:

- Nearly 97.8% of current booking customers had no previous bookings that was not cancelled in prior to this current booking
- There are also few customers who kept their bookings without cancelling them, thus a maximum of 58 bookings were kept not cancelled

avg_price_per_room

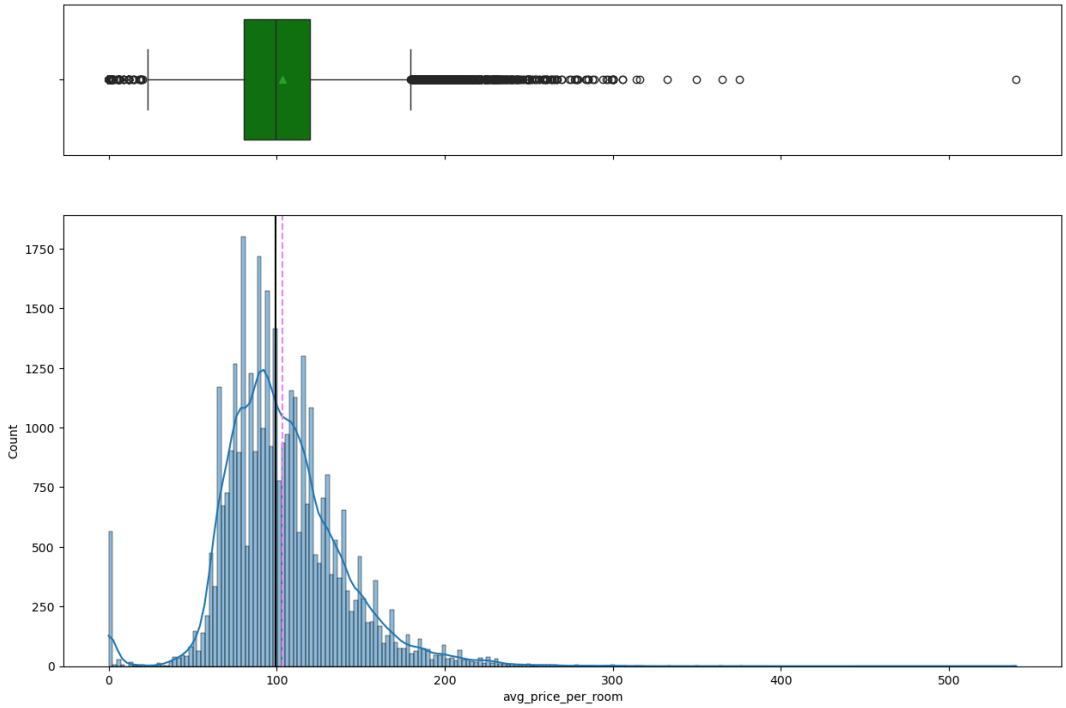
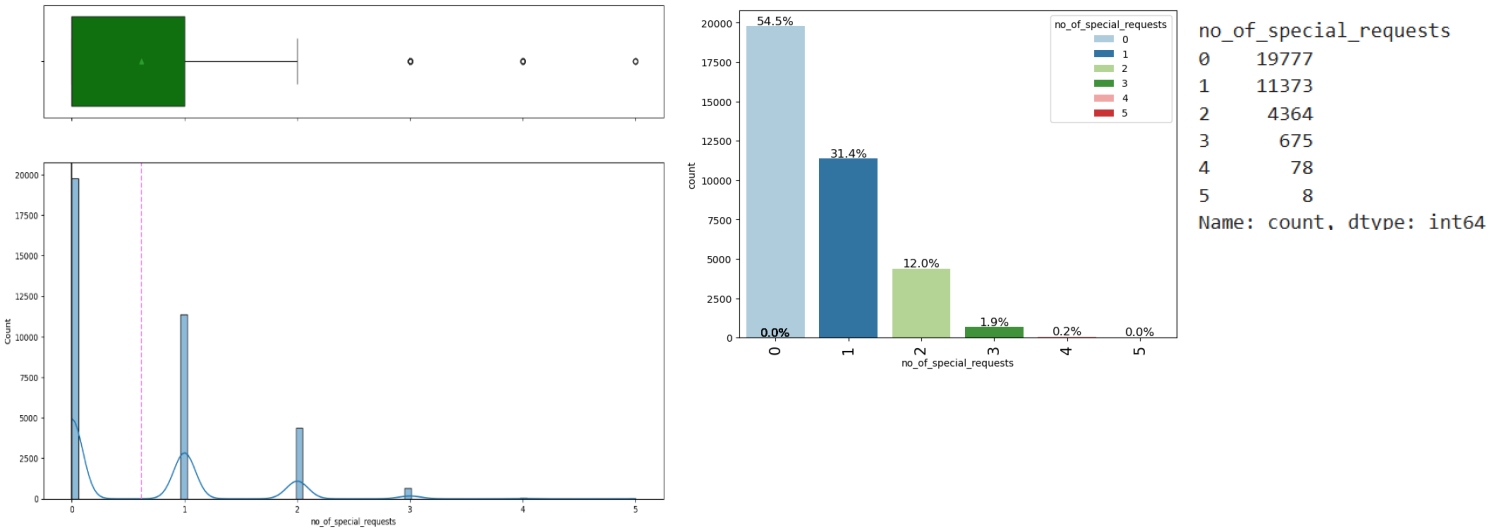


Figure 11: Histogram and Box plot for avg_price_per_room

Observation:

- Here it is noticed that the mean is slightly greater than the median and the distribution has a longer right tail, hence positively skewed
- Most of the bookings are concentrated on lower priced rooms, while still there are bookings for rooms with extreme high price of about €540.

no_of_special_requests




```
no_of_special_requests
0    19777
1    11373
2     4364
3      675
4       78
5        8
Name: count, dtype: int64
```

Figure 12: Histogram, Box plot, labelled bar plot and value counts for no_of_special_requests

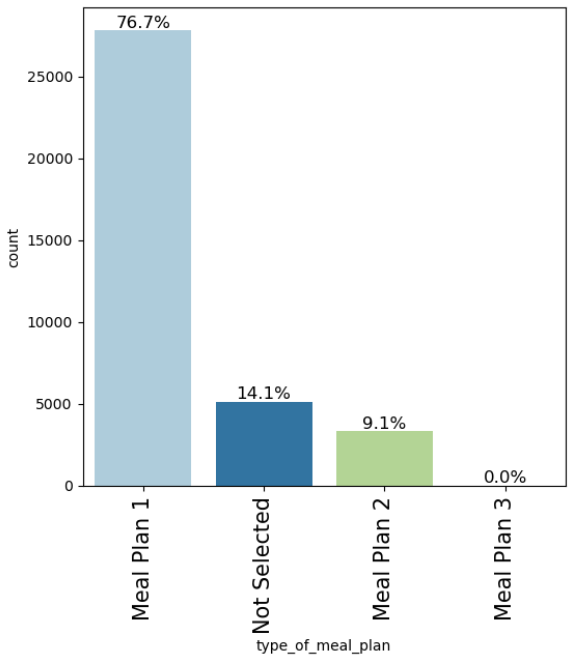
Observation:

- About 54.5% of current bookings have no special requests
- 31.4% of bookings have one special request
- 12.0% of bookings have two special requests
- A rare of 8 bookings have a maximum of 5 special requests

Categorical variables

- type_of_meal_plan
- required_car_parking_space
- room_type_reserved
- market_segment_type
- repeated_guest
- booking_status

type_of_meal_plan



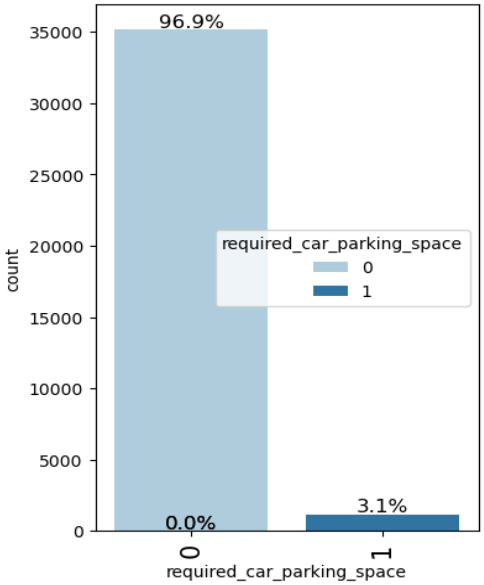

```
type_of_meal_plan
Meal Plan 1     27835
Not Selected     5130
Meal Plan 2      3305
Meal Plan 3         5
Name: count, dtype: int64
```

Figure 13: Labelled bar plot and value counts for type_of_meal_plan

Observation:

- About 76.7% of current bookings prefer having Meal Plan 1, i.e., only Breakfast
- 14.1% of bookings have not selected their meal plans
- 9.1% of bookings have given their option for Meal Plan 2, i.e., one other meal along with Breakfast
- A rare of 5 bookings have gone with an option for Meal Plan 3 which includes breakfast, lunch and dinner

required_car_parking_space



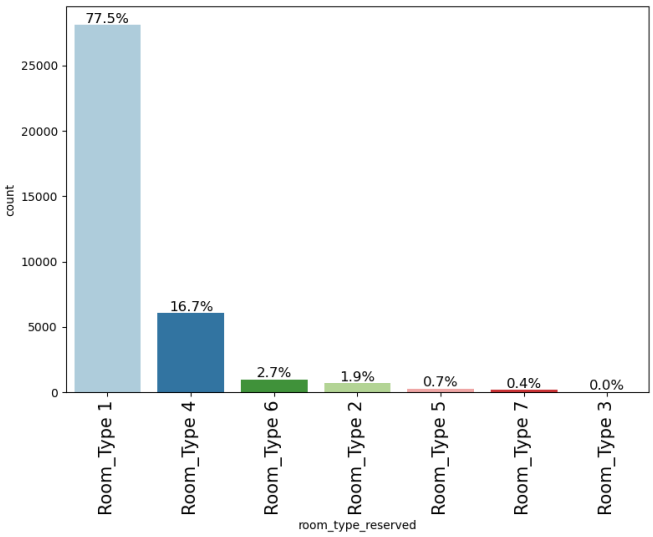

```
required_car_parking_space
0    35151
1     1124
Name: count, dtype: int64
```

Figure 14: Labelled bar plot and value counts for required_car_parking_space

Observation:

- About 96.9% of current bookings do not prefer/require a car parking space.
- Only ~3.1% of the bookings, corresponding to 1124 reservations, required parking space for cars.

room_type_reserved



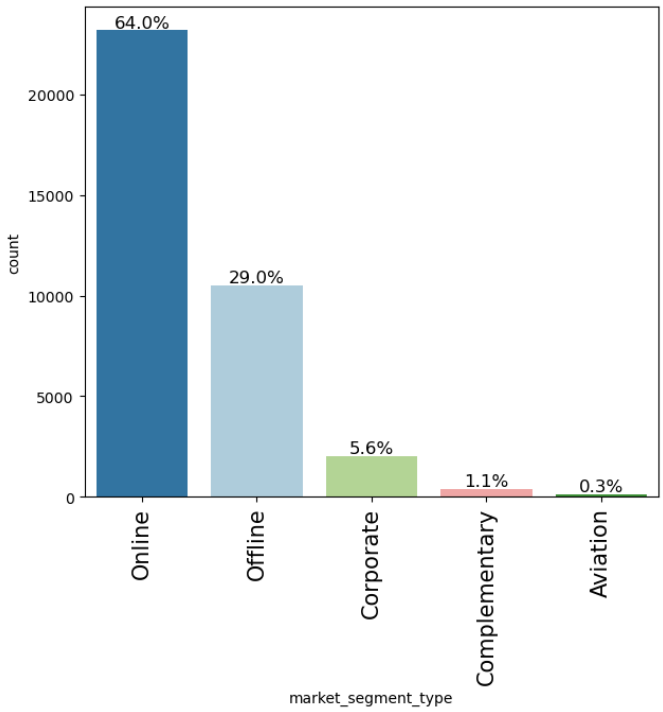

```
room_type_reserved
Room_Type 1    28130
Room_Type 4     6057
Room_Type 6      966
Room_Type 2      692
Room_Type 5      265
Room_Type 7      158
Room_Type 3        7
Name: count, dtype: int64
```

Figure 15: Labelled bar plot and value counts for room_type_reserved

Observation:

- About 77.5% of current bookings prefer/require Room_Type 1
- 16.7% of current bookings booked Room_Type 2
- Only 7 of the total current bookings, preferred Room_Type 3
- There is no much details given regarding these Room_Types as how will these rooms be and what all facilities will be here, etc.,. So it will be difficult to analyze its significance with respect to booking cancellations.

market_segment_type



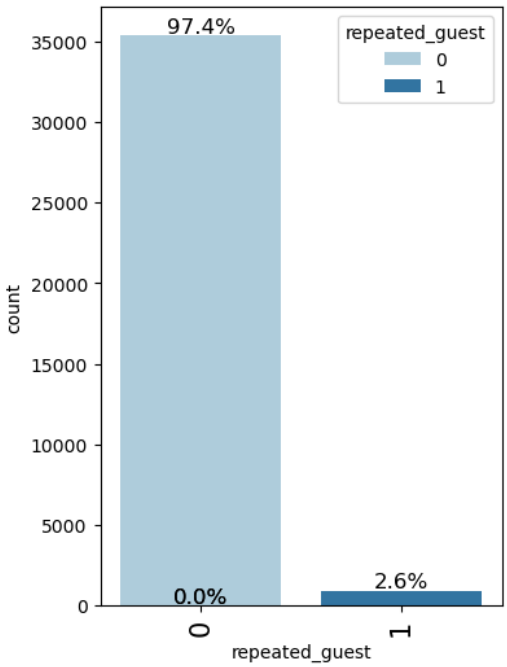

```
market_segment_type
Online           23214
Offline          10528
Corporate         2017
Complementary      391
Aviation           125
Name: count, dtype: int64
```

Figure 16: Labelled bar plot and value counts for market_segment_type

Observation:

- Most of the bookings, ~ 64%, are made through "Online" market segment.
- The next sought after market segment is "Offline," with ~ 29% of the bookings.
- The "Corporate" market segment has around 5.6% of the bookings. This suggests that some bookings were made by corporate clients.
- The "Complementary" market segment has only 1.1% of the bookings.
- Aviation is the lowest market segment, with only 0.3% of the bookings.

repeated_guest



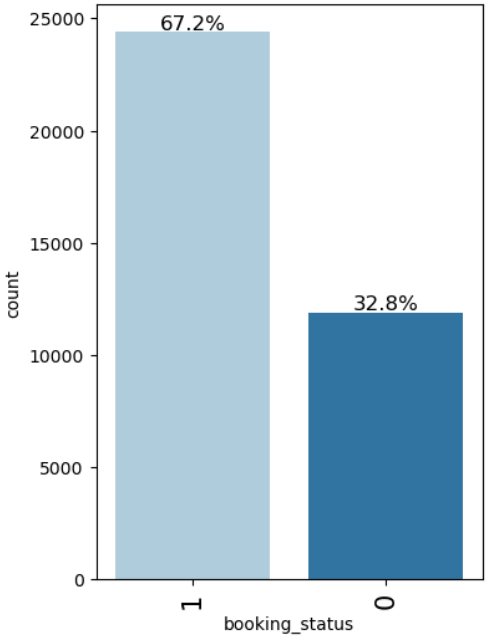

```
repeated_guest
0    35345
1      930
Name: count, dtype: int64
```

Figure 17: Labelled bar plot and value counts for repeated_guest

Observation:

- Nearly 97.44% of customers (35345 bookings), have not been made more than once.
- A very less no. of customers (~930) 2.6%, have booked more than once. This shows their preference for INN Group Hotels

booking_status




```
booking_status
1    24390
0    11885
Name: count, dtype: int64
```

Figure 18: Labelled bar plot and value counts for booking_status

Observation:

- Only 67.2% of customers have not cancelled their bookings whereas 32.8% of customers have cancelled their bookings

Distribution of numerical variables in the data

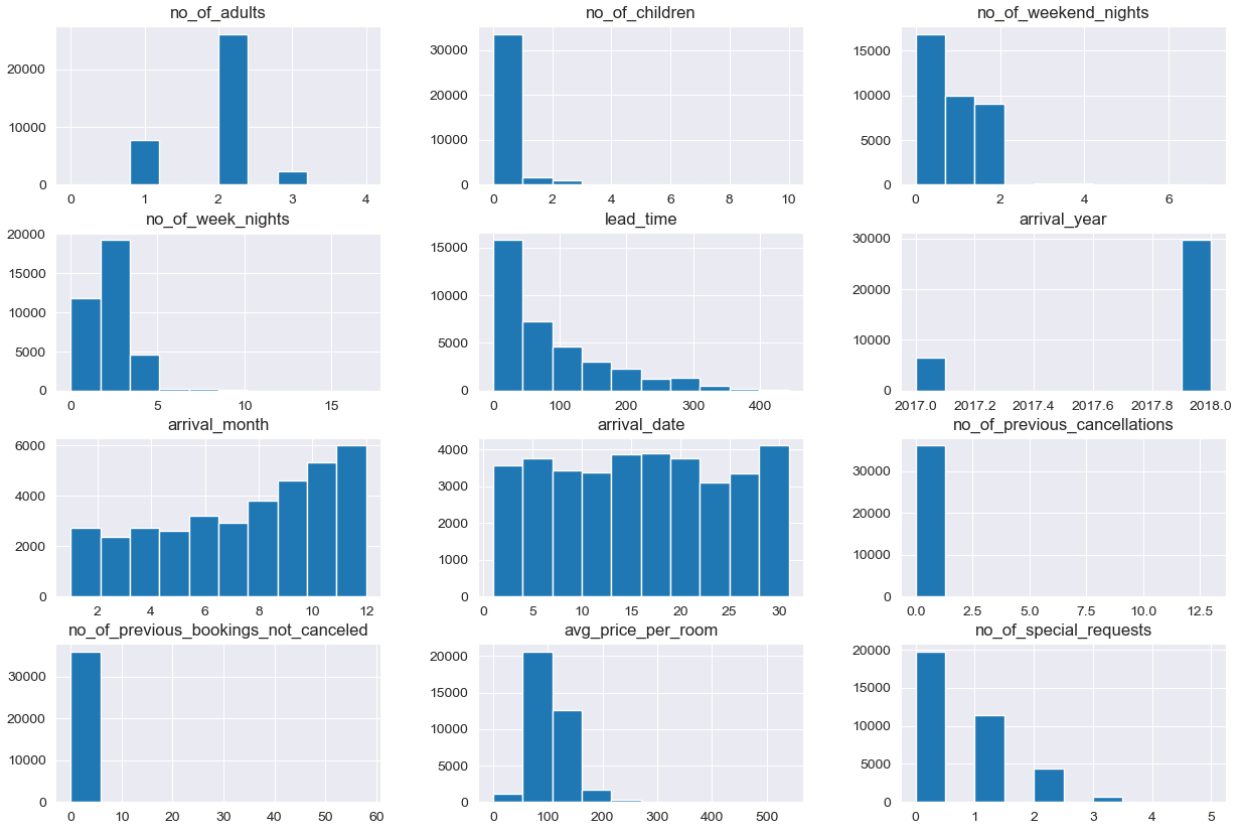


Figure 19: Distribution of numerical variables in the data

1.6 Bivariate analysis

Correlation between numerical variables

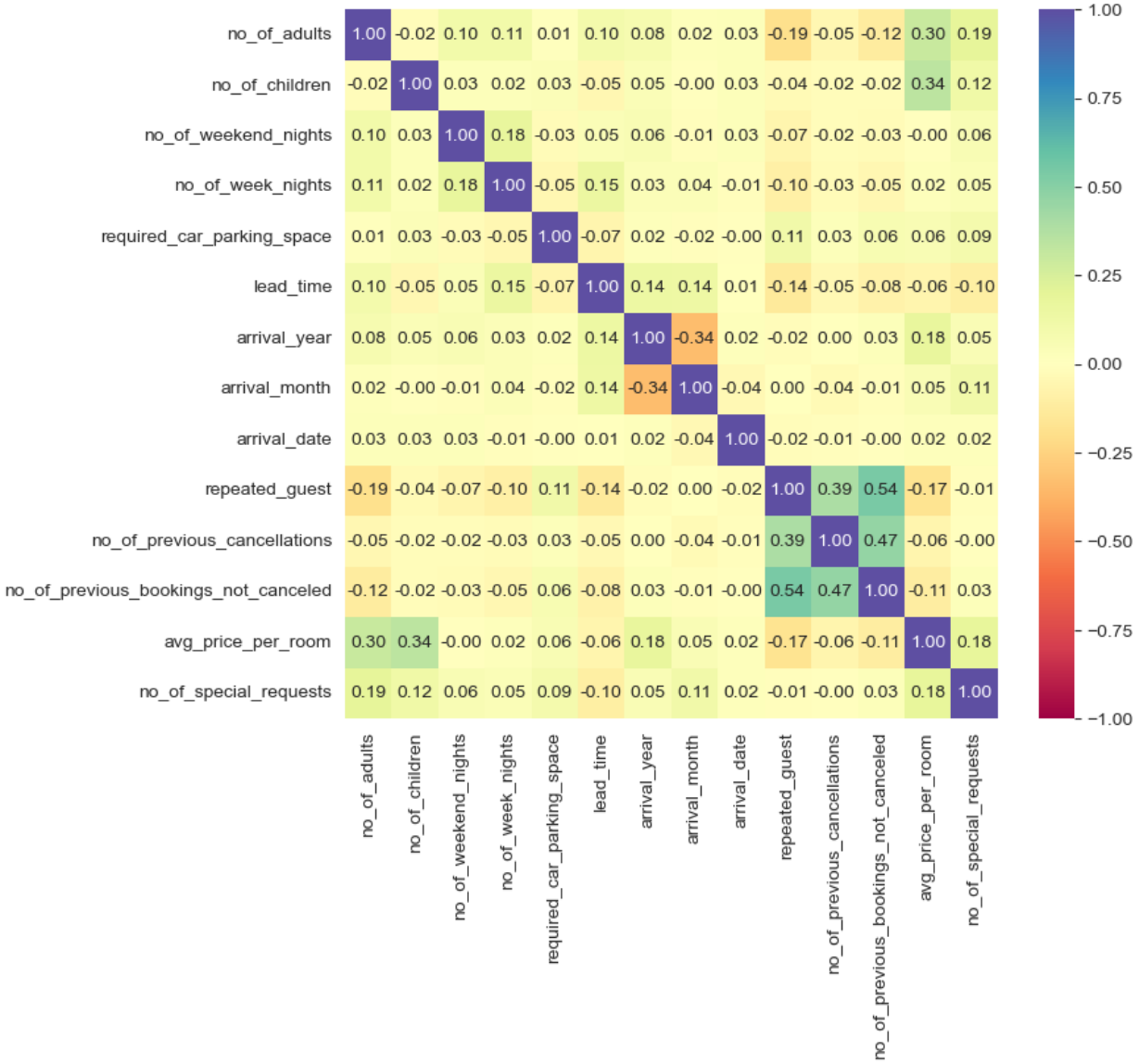


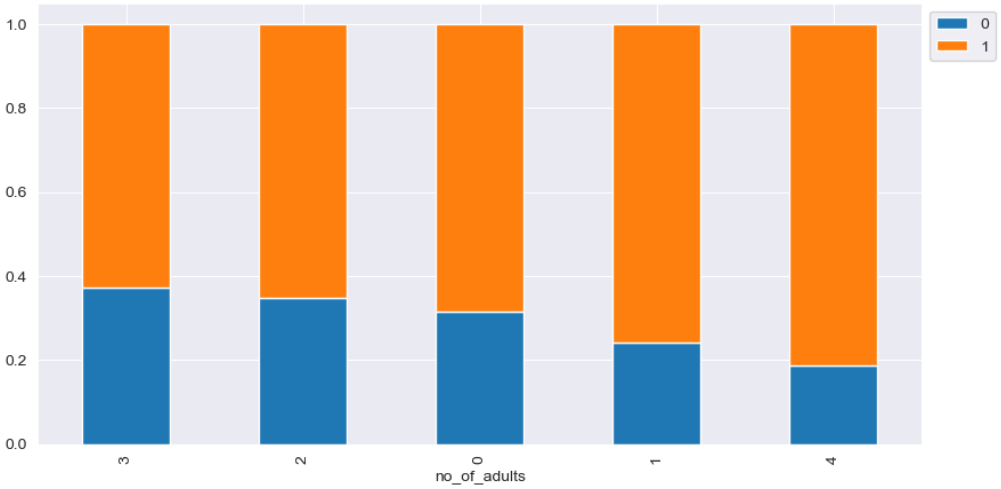
Figure 20: Heatmap between numerical variables

Observation:

- `no_of_adults` is positively correlated with the `avg_price_per_room` with a correlation value of 0.30, thus an increase in `no_of_adults` in a booking tends to increase the `avg_price_per_room`.
- `no_of_children` also is positively correlated with the `avg_price_per_room` with a correlation value of 0.34, potraying a similar visualization as before.
- `no_of_weekend_nights` and `no_of_week_nights` is negatively correlated with `required_car_parking_space` with values -0.03 and -0.05 respectively, whcih implies guests staying for longer weekend and week nights do not prefer a car parking space.

- `required_car_parking_space` is positively correlated with the `avg_price_per_room` with a value of 0.06, which implies that customers who are in need of a car parking space are charged high compared to others who don't.
- `required_car_parking_space` is also positively correlated with the `repeated_guest`, with a value of 0.11, which tells that guests who prefer to stay at INN hotels group repeatedly requires a parking space.
- `lead_time` has a negative correlation with `avg_price_per_room` (~ -0.06), which implies that advance bookings have a cheap/less `avg_price_per_room`
- `repeated_guest` show a negative correlation with `no_of_adults`,`no_of_children`,`no_of_weekend_nights`, `no_of_week_nights` and `no_of_special_requests`. This shows that repeated guests may have less no. of adults, children and special requests compared to one time visitors. And also they tend to stay less at nights.
- `booking_status` has a negative correlation with `lead_time` and `avg_price_per_room`, which shows that bookings done very earlier are likely to be cancelled easily and they are also charged with high room prices.
- It's important to note that correlation does not imply causation.

Observations on no_of_adults vs booking_status

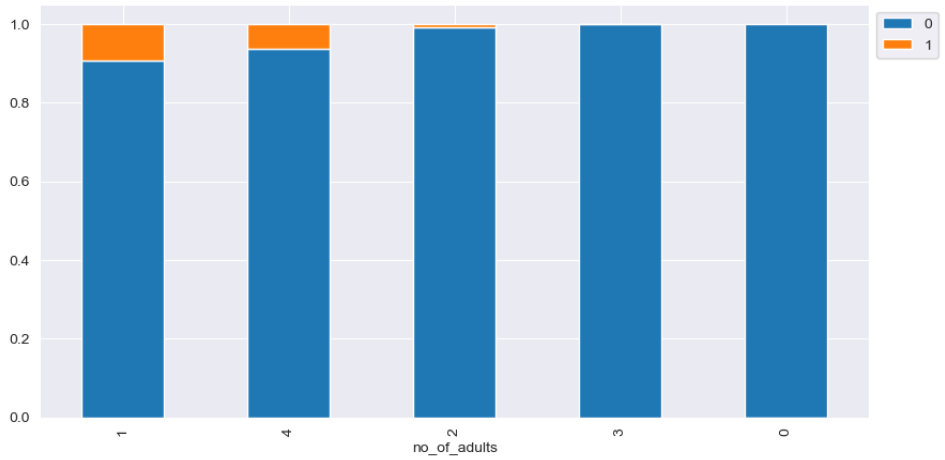


booking_status	0	1	All
no_of_adults			
All	11885	24390	36275
2	9119	16989	26108
1	1856	5839	7695
3	863	1454	2317
0	44	95	139
4	3	13	16

Figure 21: Stacked bar plot for no_of_adults vs booking_status

- It is observed that the bookings with maximum of 4 adults, the cancellation is very less compared to 1, 3, 0 and 2
- The bookings done by single adult also tends not to be cancelled easily

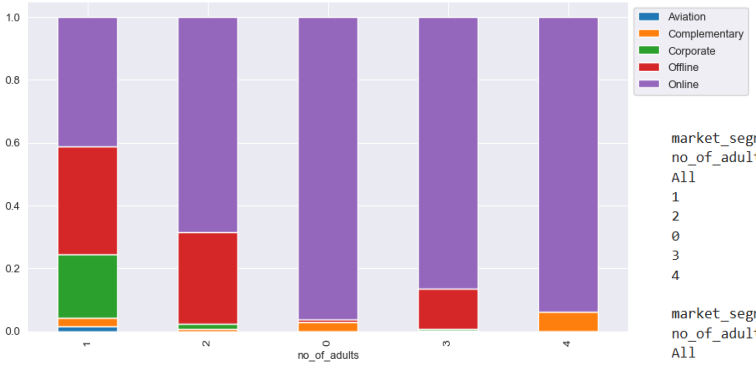
Observations on no_of_adults vs repeated_guest



repeated_guest	0	1	All
no_of_adults			
All	35345	930	36275
1	6974	721	7695
2	25903	205	26108
3	2314	3	2317
4	15	1	16
0	139	0	139

Figure 22: Stacked bar plot for no_of_adults vs repeated_guest

Observations on no_of_adults vs market_segment_type



market_segment_type	Aviation	Complementary	Corporate	Offline	Online	\
no_of_adults						
All	125	391	2017	10528	23214	
1	123	199	1566	2634	3173	
2	2	184	438	7598	17886	
0	0	4	0	1	134	
3	0	3	13	295	2006	
4	0	1	0	0	15	

market_segment_type	All
no_of_adults	
All	36275
1	7695
2	26108
0	139
3	2317
4	16

Figure 23: Stacked bar plot for no_of_adults vs market_segment_type

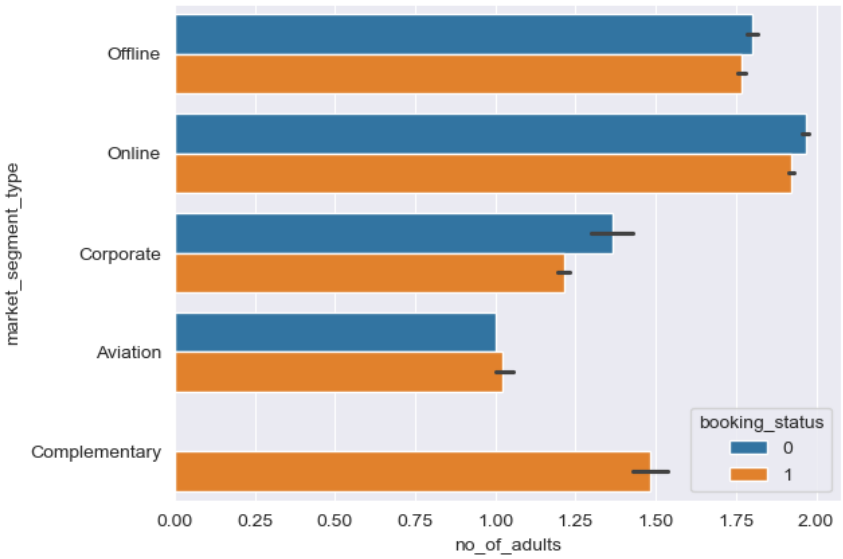


Figure 24: Bar plot for no_of_adults vs market_segment_type

- From the above two visualizations it can be viewed that the most of the repeated guests are single adults, followed by 4 adults. And it is also observed that 4 adults don't accompany children.
- These single adults have done their bookings mostly through online, offline or by corporate. And guest booking with 4 adults have done their bookings widely through online while few were through complementary
- It is also noted that complimentary market segments were not cancelled, while bookings done through online market segments face higher cancellations compared to other market segments

Observations on no_of_adults vs avg_price_per_room

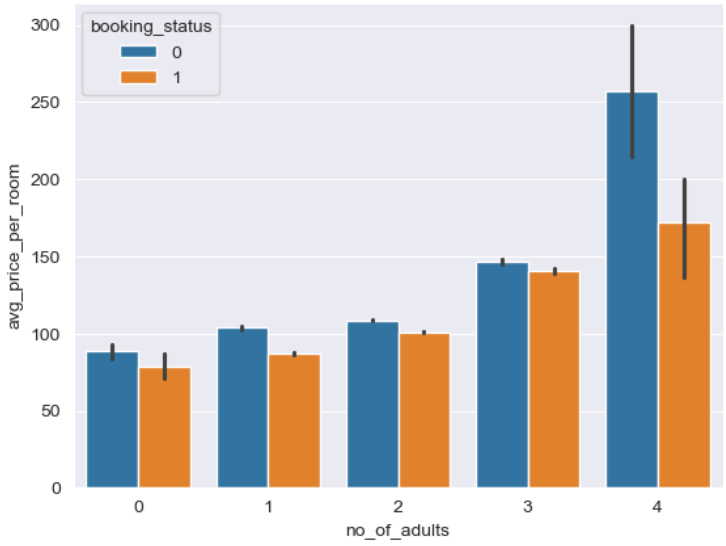


Figure 25: Bar plot for no_of_adults vs avg_price_per_room

- The average price per room increases with the increase in no. of adults in the bookings and it is evident that the cancellations is high compared to no_cancellations irrespective of the no. of adults in the bookings.
- It is noted that when the no. of adults is four the cancellation to no_cancellation ratio is large compared to others. As observed with four adults no children is accompanying. So four adults can be friends or colleagues, so their plans are morelikely to be dissmissed leading to cancellations. And it is also observed that their bookings are either through online/ complimentary market segments indicating not a mandatory plans.

Observations on no_of_adults vs no_of_previous_cancellations

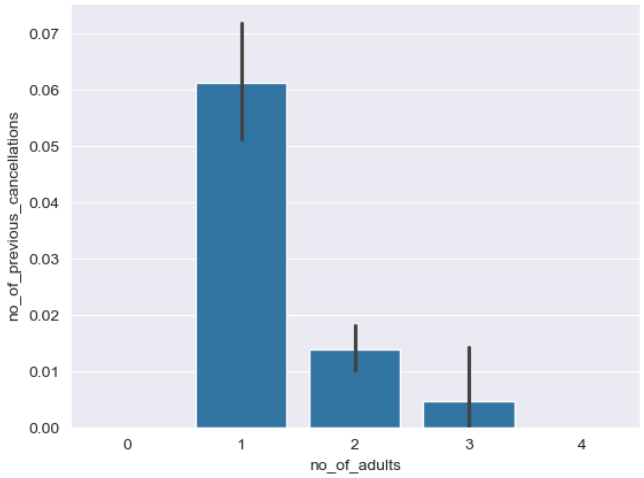


Figure 26: Bar plot for no_of_adults vs no_of_previous_cancellations

- The number of previous cancellations is high when bookings are done by single adults

Observations on no_of_adults vs no_of_previous_bookings_not_canceled

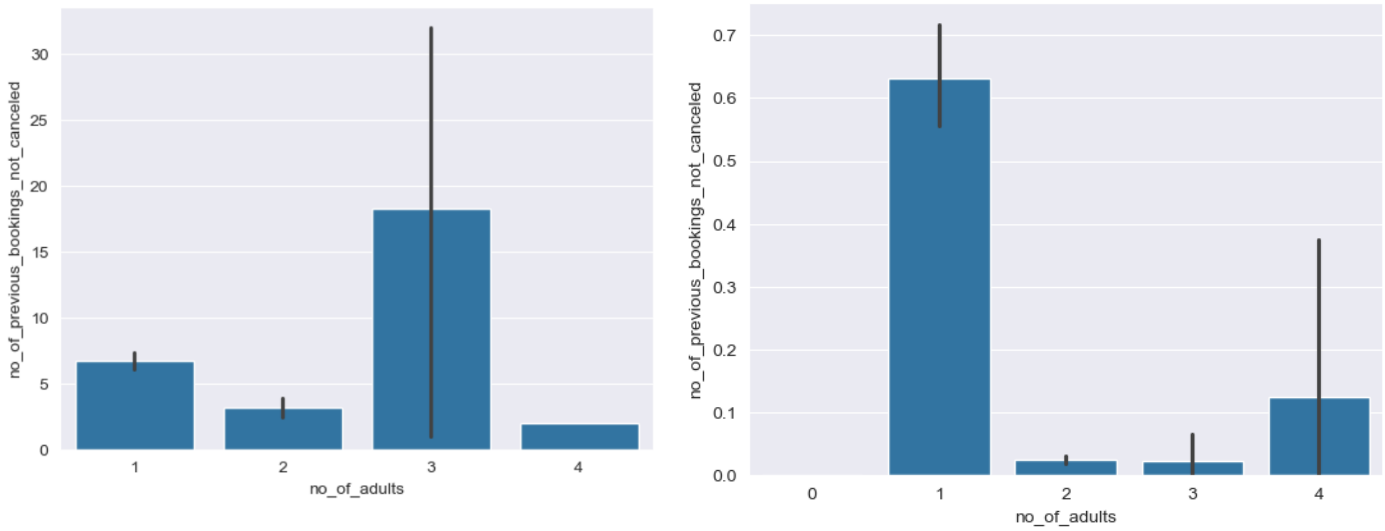
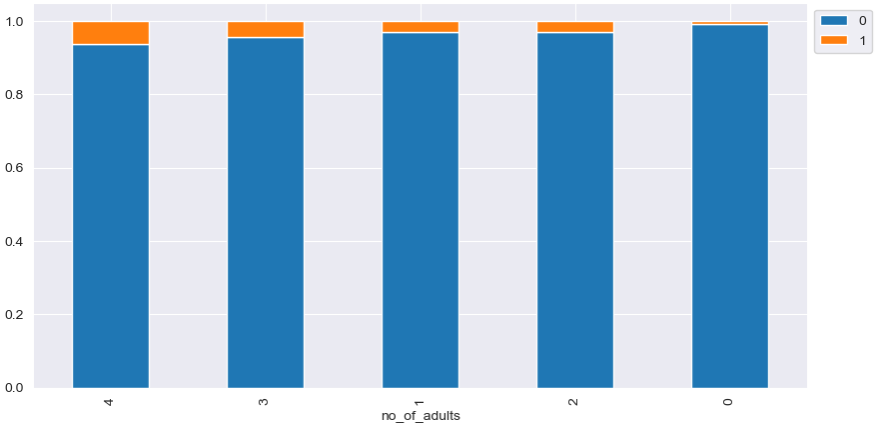


Figure 27: Bar plot for no_of_adults vs no_of_previous_bookings_not_canceled

- The number of previous bookings that are not cancelled in prior to the current booking is mostly done by single adults
- But with repeated guests it is seen that bookings done by 3 adults were not cancelled in prior to the current booking

Observations on no_of_adults vs required_car_parking_space

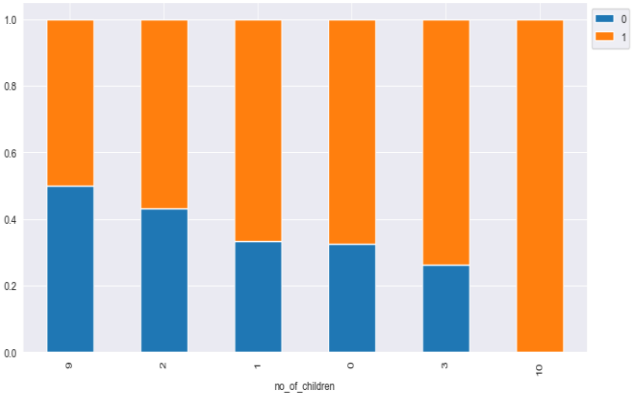


required_car_parking_space	0	1	All
no_of_adults			
All	35151	1124	36275
2	25325	783	26108
1	7457	238	7695
3	2216	101	2317
0	138	1	139
4	15	1	16

Figure 28: Stacked bar plot for no_of_adults vs required_car_parking_space

- It is observed that majority of the customers don't require a parking space while the bookings done by many no. of adults require car parking space compared to the bookings done by few no. of adults

Observations on no_of_children vs booking_status



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

Figure 29: Stacked bar plot for no_of_children vs booking_status

- As with increase in number of children the bookings tends to get cancelled easily.
- There is also a rare single booking done and executed with 10 children

Observations on no_of_children vs avg_price_per_room

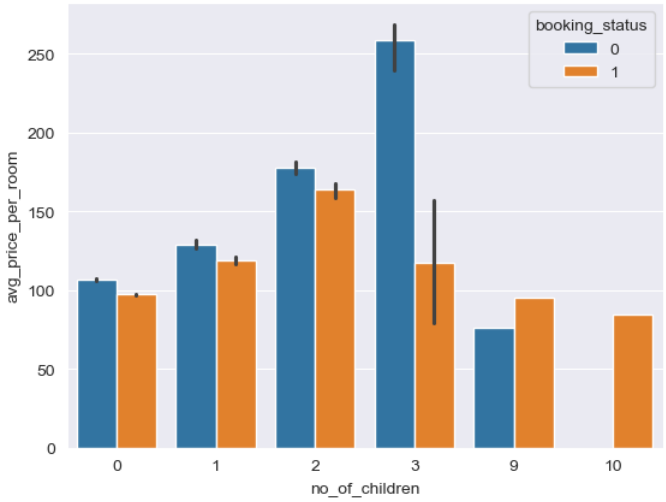
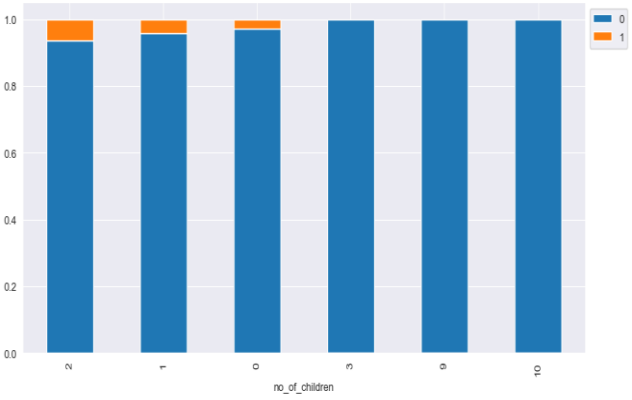


Figure 30: Bar plot for no_of_children vs avg_price_per_room

- As with increase in number of children tends to increase up to 2 children, but any bookings done with children 3 and more than that, the average price per room decreases
- This shows that INN Hotels group encourages with price discounts for the bookings done with many children, but this can be concluded with their requirement of parking space

Observations on no_of_children vs required_car_parking_space

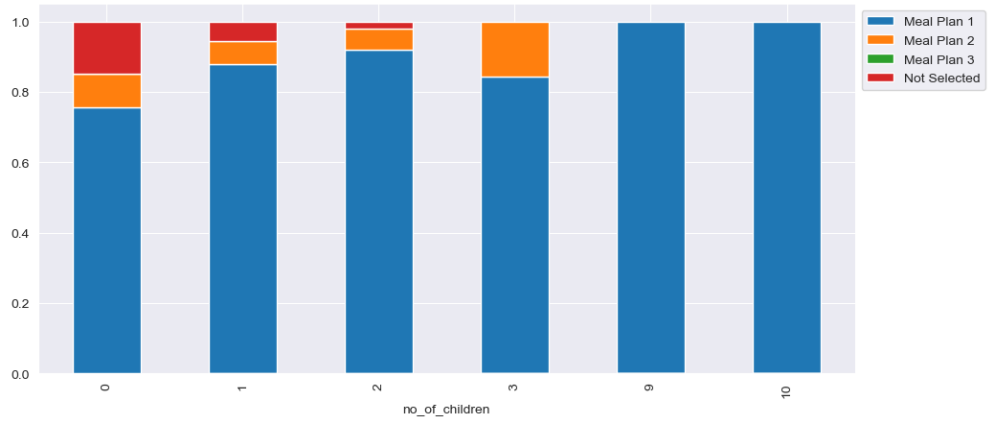


required_car_parking_space	0	1	All
no_of_children			
All	35151	1124	36275
0	32590	987	33577
1	1549	69	1618
2	990	68	1058
3	19	0	19
9	2	0	2
10	1	0	1

Figure 31: Stacked bar plot for no_of_children vs required_car_parking_space

- Bookings done with 0 to 2 children required a car parking space while bookings done with 3 or more children did not require a parking space
- This is also a reason for having a low average price rooms for children with 3 and more.

Observations on no_of_children vs type_of_meal_plan



type_of_meal_plan	Meal Plan 1	Meal Plan 2	Meal Plan 3	Not Selected	All
no_of_children					
0	25420	3134	5	5018	33577
All	27835	3305	5	5130	36275
1	1422	106	0	90	1618
2	974	62	0	22	1058
3	16	3	0	0	19
9	2	0	0	0	2
10	1	0	0	0	1

Figure 32: Stacked bar plot for no_of_children vs type_of_meal_plan

- Few of the bookings with children up to 3 tend to prefer Meal Plan 2 which included another meal along with breakfast
- While bookings with many children go with choosing only the breakfast which remains complimentary in most of the hotels

Observations on required_car_parking_space vs avg_price_per_room

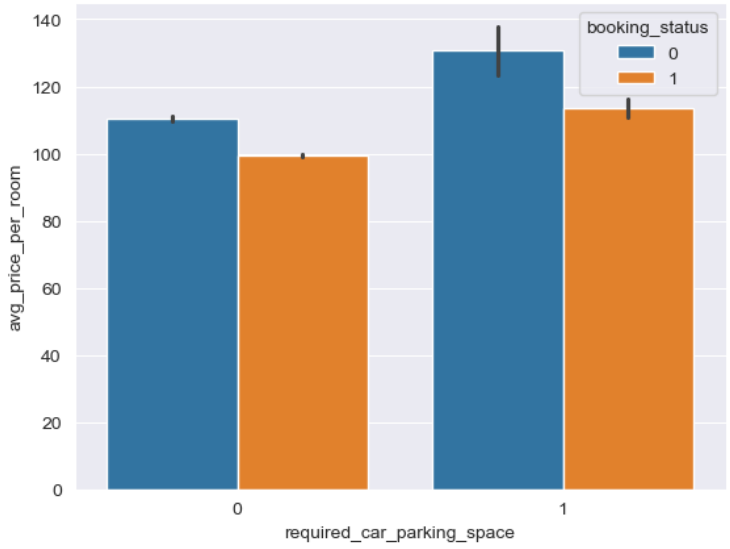


Figure 33: Bar plot for required_car_parking_space vs avg_price_per_room

- Bookings done with the requirement of a car parking space are charged high with average price per room
- It is also noted that bookings requiring a car parking space tend to show up compared to bookings that do not require a parking space
- Moreover, the charge levied for the parking space along with higher average room price also increases the chance of booking cancellations

Observations on no_of_weekend_nights vs avg_price_per_room

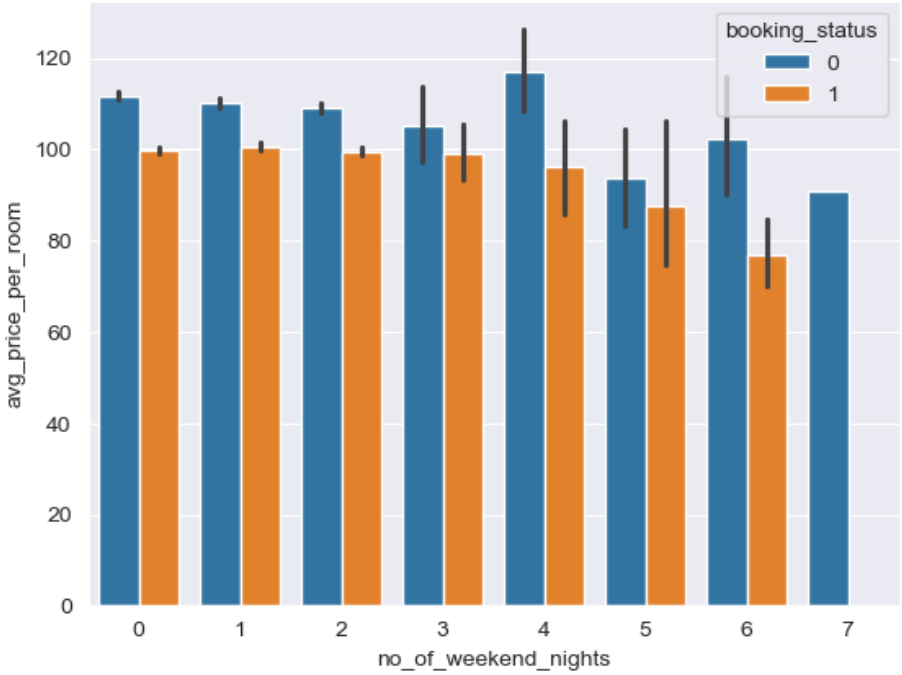


Figure 34: Bar plot for no_of_weekend_nights vs avg_price_per_room

Observations on no_of_week_nights vs avg_price_per_room

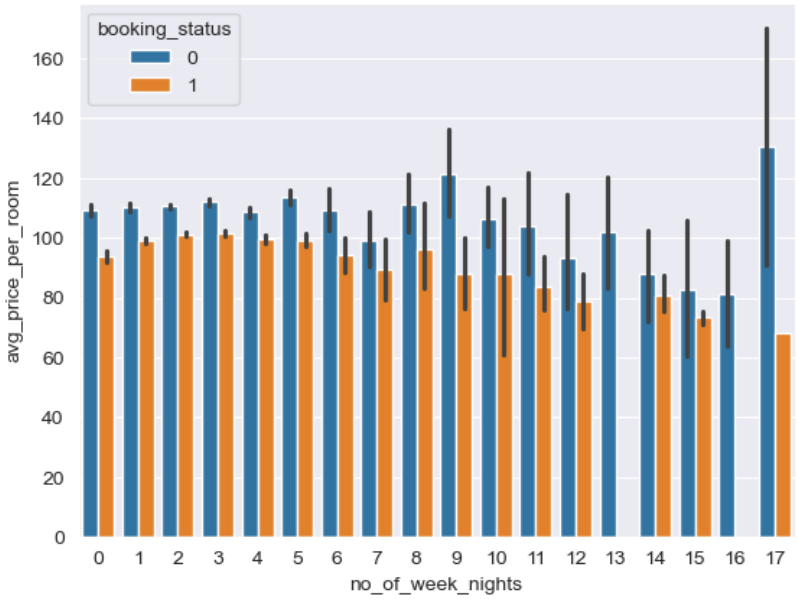


Figure 35: Bar plot for no_of_week_nights vs avg_price_per_room

- With the above two visualizations of bookings done for longer stays, the average price per room decreases
- It is also observed that the bookings were showed up with shorter stays while with longer stays the bookings tend to gets cancelled easily

Observations on type_of_meal_plan vs avg_price_per_room

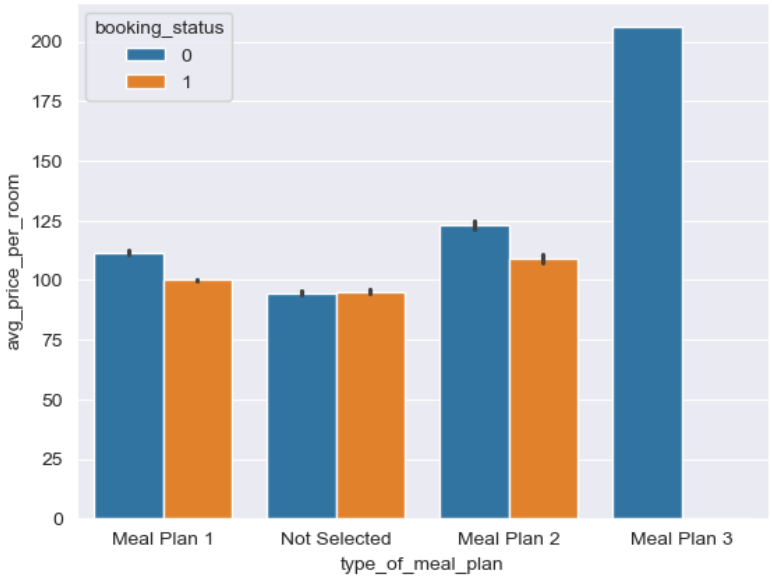


Figure 36: Bar plot for type_of_meal_plan vs avg_price_per_room

- There is a slight variation observed in average price per rooms with respect to the selection of meal plans.
- The room price is bit high for the bookings done with preference for Meal Plan 2 and Meal Plan 3
- It is also observed that bookings done with no selection of meal plans has less cancellations compared to the bookings which showed preference for meal plans

Observations on arrival_month vs avg_price_per_room

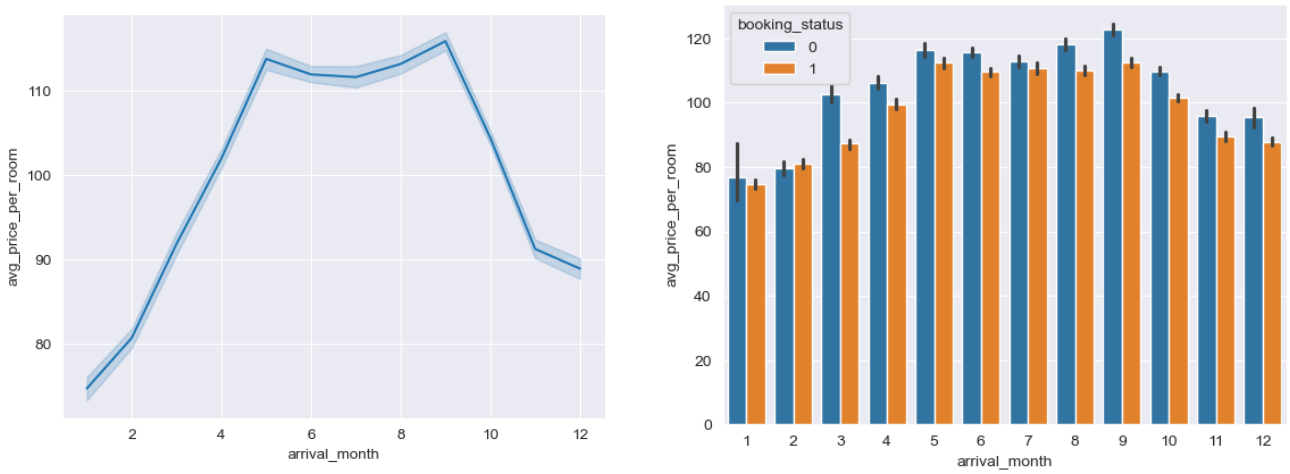


Figure 37: Line plot and bar plot for arrival_month vs avg_price_per_room

- It is noted that the average price per room increases during a season, say from May to September
- While it is at the lowest during January and February, but gradually increases from April and remains high during May till September
- Again it sees a fall after October

Observations on market_segment_type vs avg_price_per_room

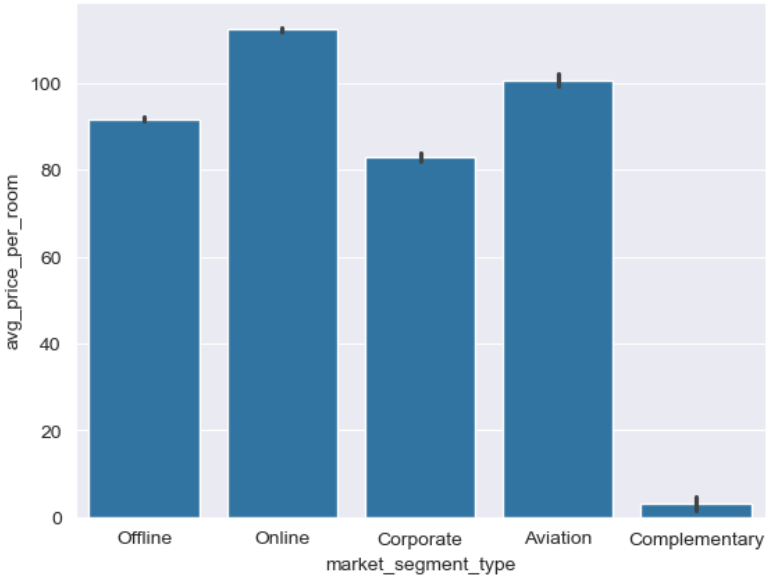


Figure 38: Bar plot for market_segment_type vs avg_price_per_room

- Online bookings tend to be charged with higher average price per room greater than €100.
- While the next higher price is taken by bookings done through Aviation which is ~€100.
- Offline bookings also were charged higher while corporate bookings are charged around €80 to €85.

Observations on market_segment_type vs booking_status

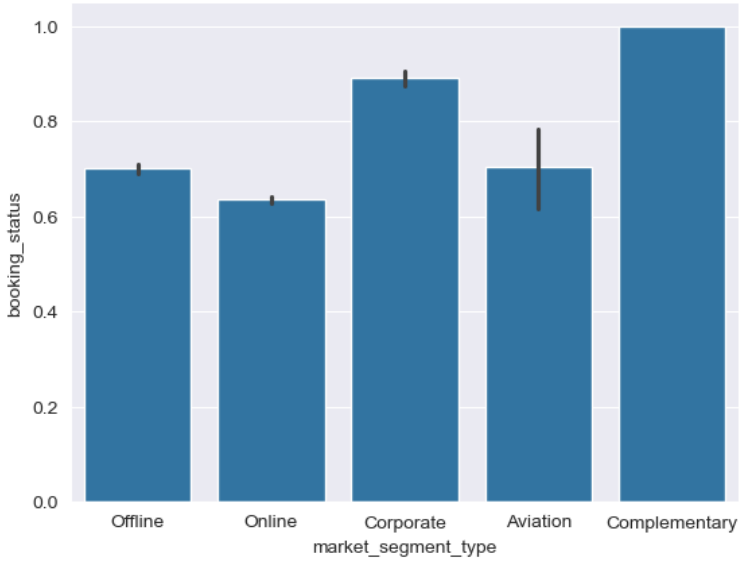
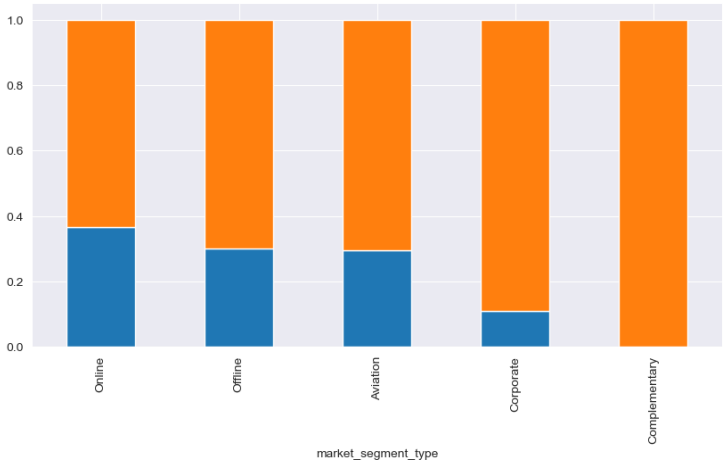


Figure 39: Bar plot for market_segment_type vs booking_status



booking_status	0	1	All
market_segment_type			
All	11885	24390	36275
Online	8475	14739	23214
Offline	3153	7375	10528
Corporate	220	1797	2017
Aviation	37	88	125
Complementary	0	391	391

Figure 40: Stacked bar plot for market_segment_type vs booking_status

- Complementary bookings never seem to face cancellations
- Corporate bookings are less likely to be cancelled, while Aviation comes the next with regard to less cancellations
- Online bookings are more prone to cancellations

Observations on no_of_special_requests vs avg_price_per_room

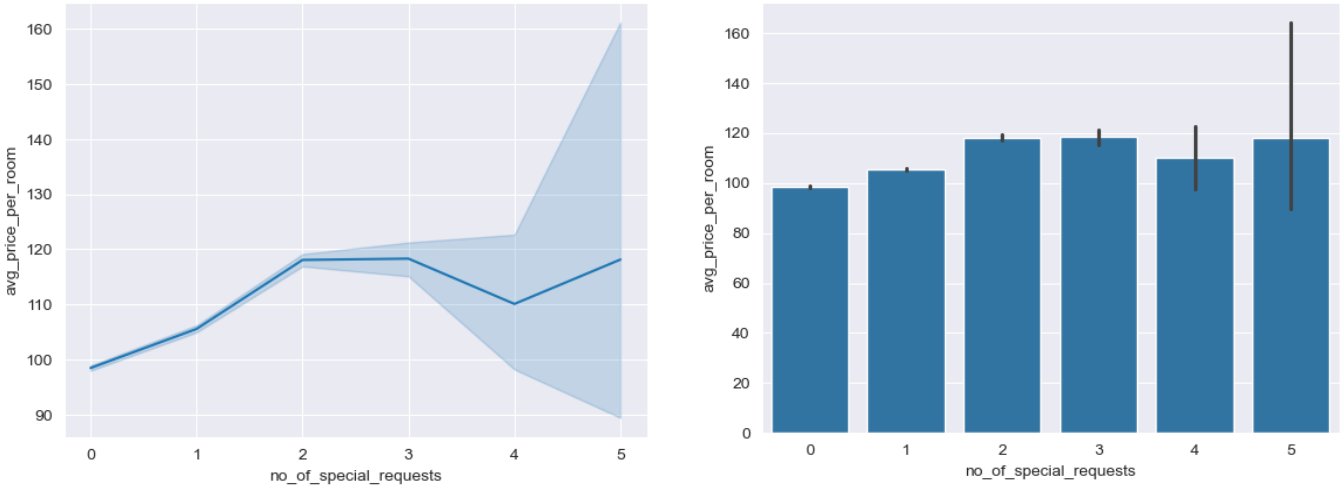
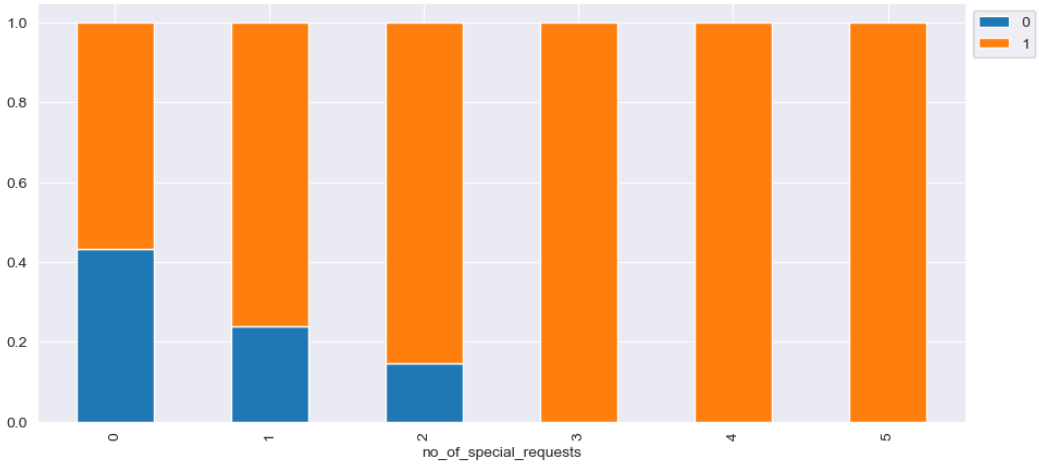


Figure 41: Line plot and bar plot for no_of_special_requests vs avg_price_per_room

- The average price per room increases with the no. of special requests

Observations on no_of_special_requests vs booking_status

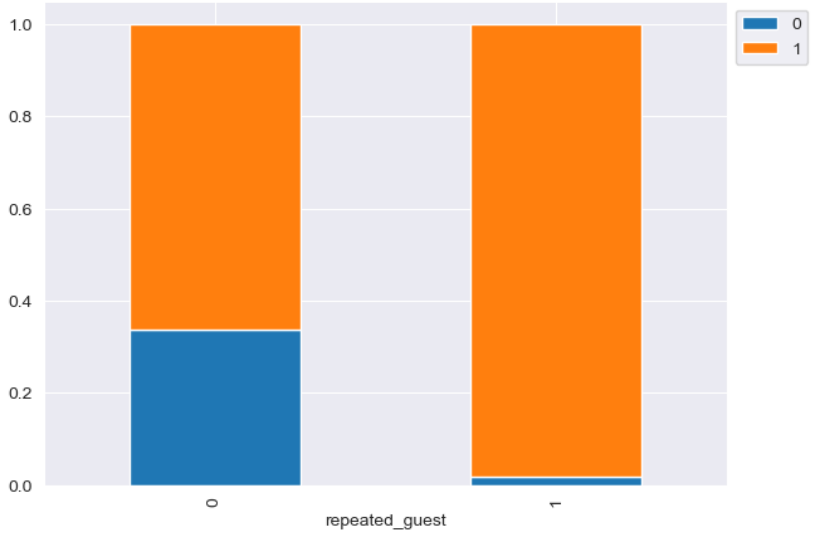


booking_status	0	1	All
no_of_special_requests			
All	11885	24390	36275
0	8545	11232	19777
1	2703	8670	11373
2	637	3727	4364
3	0	675	675
4	0	78	78
5	0	8	8

Figure 42: Stacked bar plot for no_of_special_requests vs booking_status

- Bookings with no special requests:
    - 8545 were canceled
    - 11,232 were not canceled
- Bookings with 1 special request:
    - 2703 were canceled
    - 8670 were not canceled
- Bookings with 2 special requests:
    - 637 were canceled
    - 3727 were not canceled
- There were no cancellations for bookings with 3, 4, or 5 special requests and the respective bookings are 675, 78, and 8.
- It is observed that with the increase in the no. of special request, the proportion of cancellations tends to decrease. This suggests that guests who make more special requests are less likely to cancel their bookings.

Observations on repeated_guests vs booking_status



booking_status	0	1	All
repeated_guest			
All	11885	24390	36275
0	11869	23476	35345
1	16	914	930

Figure 43: Stacked bar plot for repeated_guests vs booking_status

- Among bookings with repeated guests:
    - 914 of the bookings were not cancelled
    - 16 bookings were canceled
- Among bookings with no repeated guests:
    - 23476 bookings were not canceled.
    - 11869 were canceled.

Distribution plot on avg_price_per_room vs booking_status

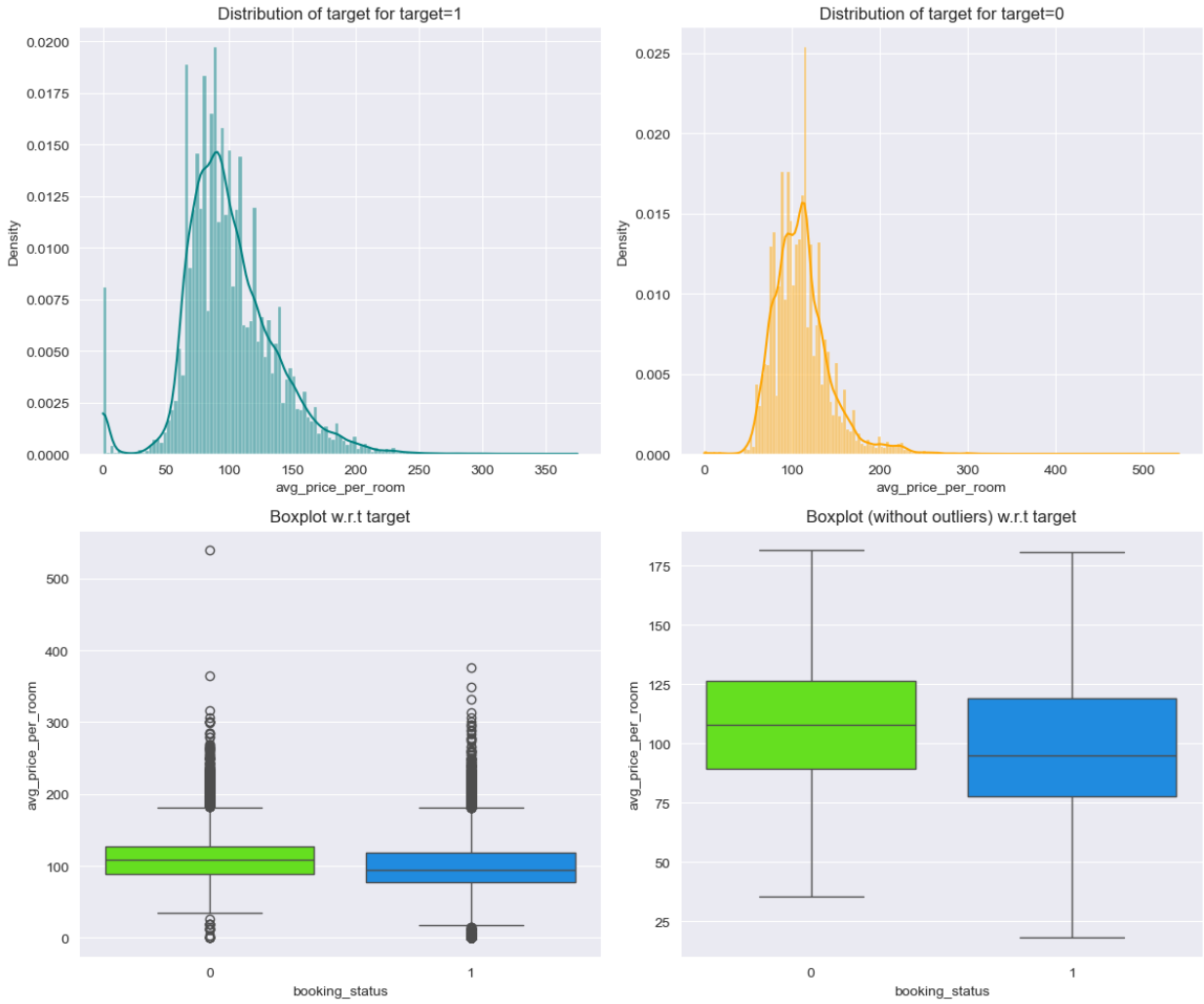


Figure 44: Distribution plot on avg_price_per_room vs booking_status

- Bookings done for a higher average price were prone to get cancelled positively

Distribution plot on lead time vs booking_status

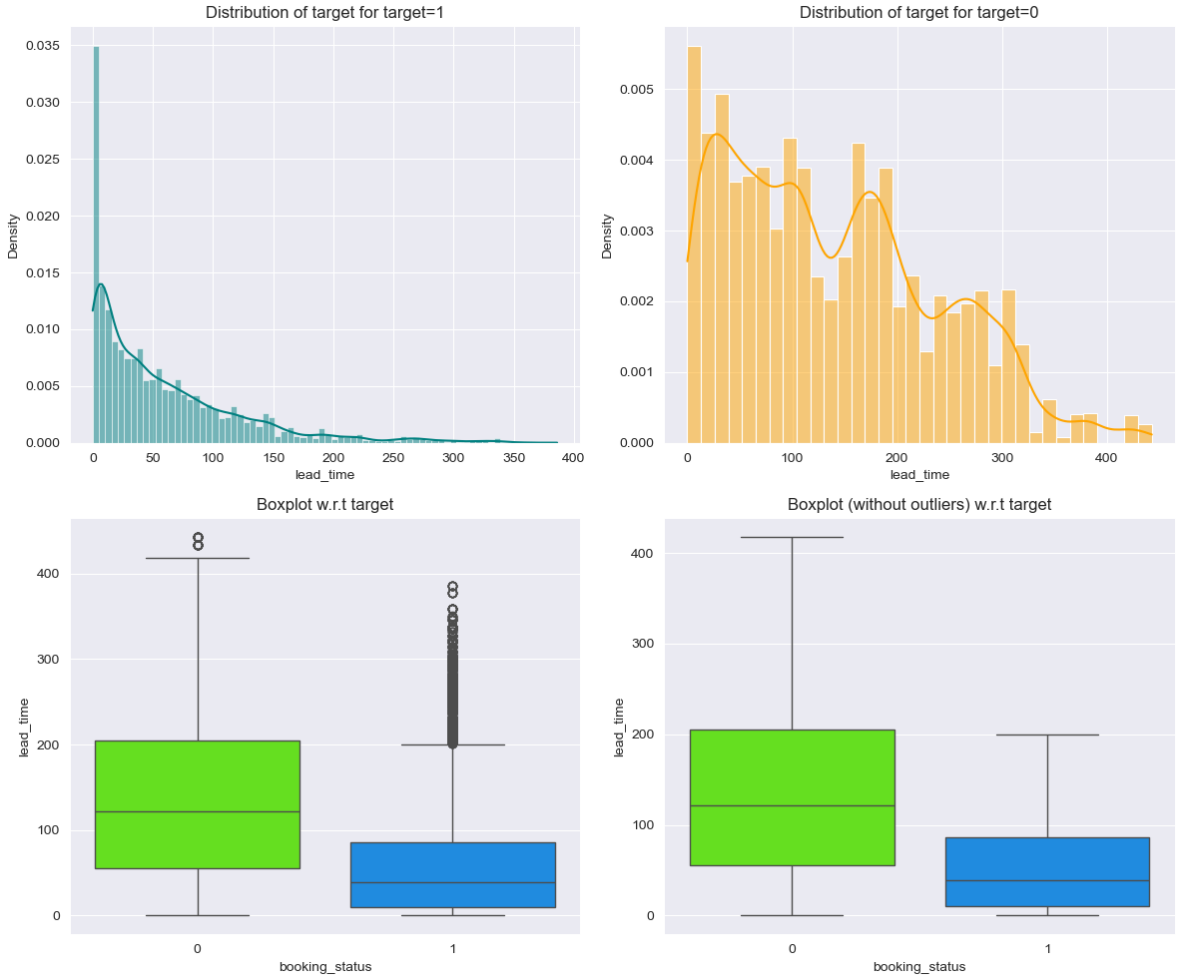


Figure 45: Distribution plot on lead_time vs booking_status

- Bookings done with the larger lead time were prone to get cancelled easily

Distribution plot on arrival_date vs booking_status

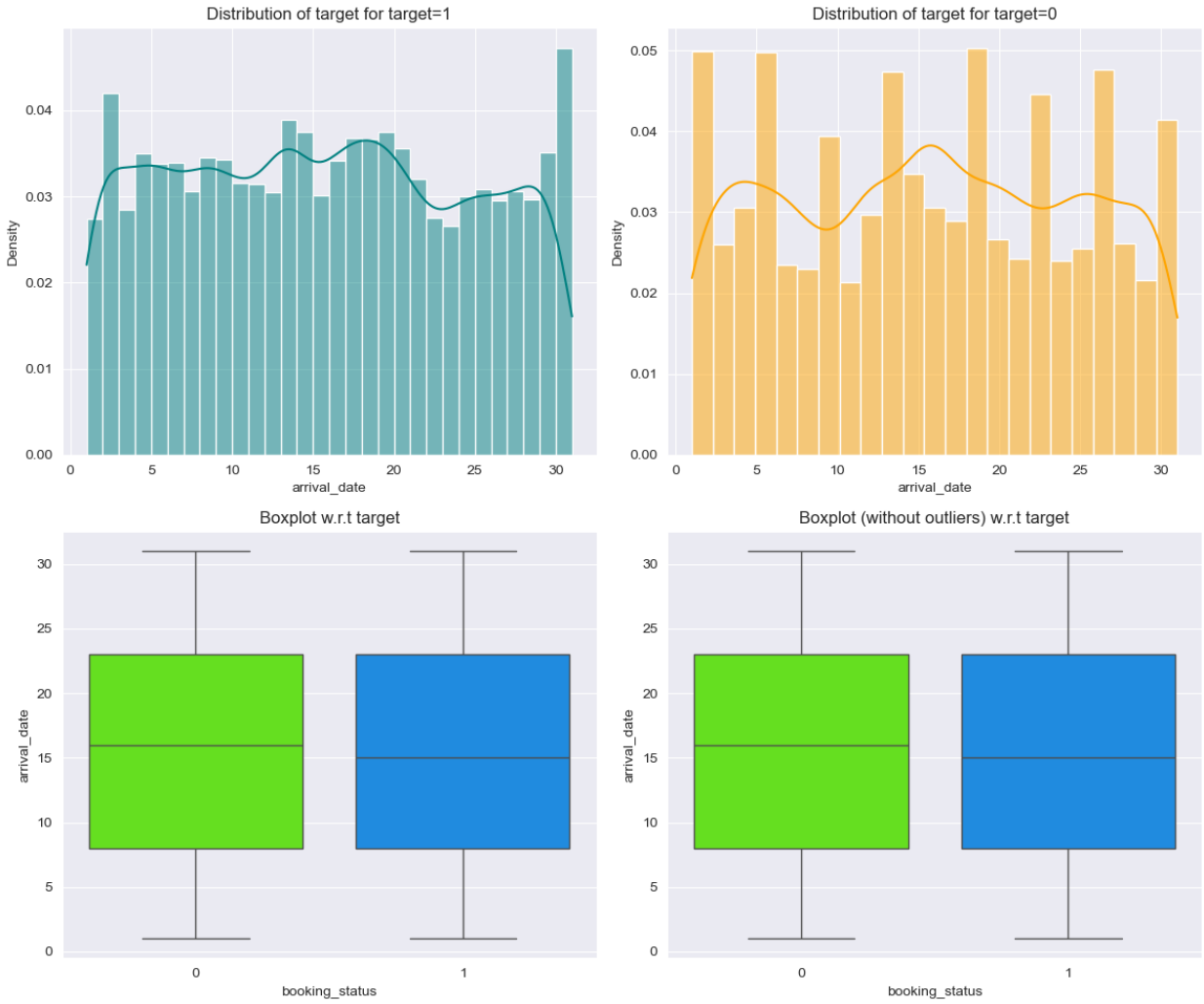
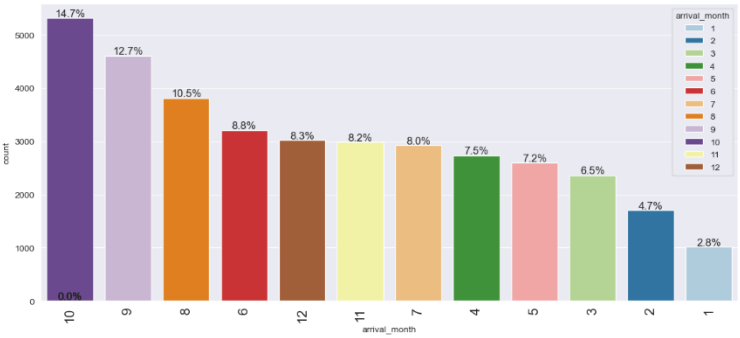


Figure 46: Distribution plot on arrival_date vs booking_status

- Bookings were done evenly throughout the month

KEY QUESTIONS

1.7 What are the busiest months in the hotel?




```
arrival_month
10    5317
9     4611
8     3813
Name: count, dtype: int64
```

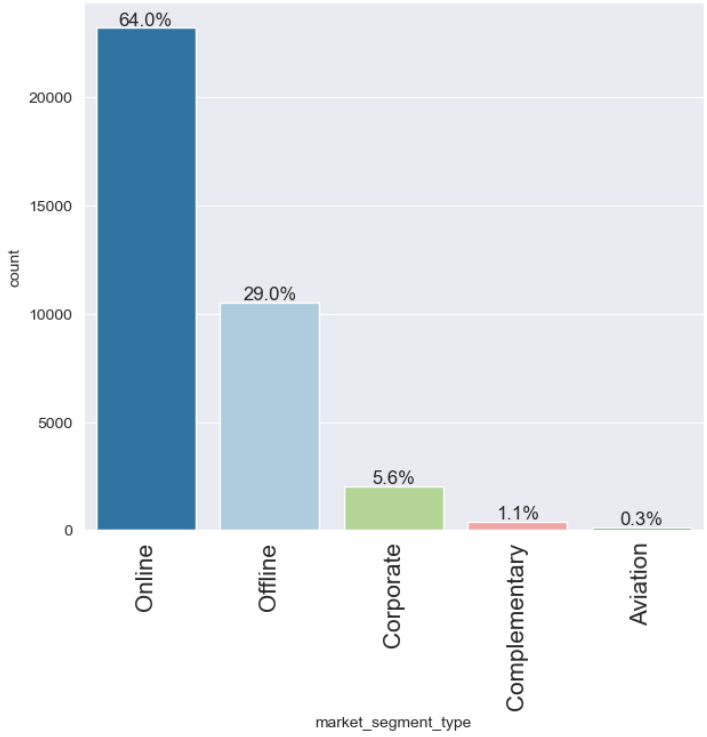
Figure 47: Labelled bar plot for arrival_month in the hotel

Observation:

- It is observed that the number of bookings starts increasing from August (10.5%) and attains a peak during October (14.7%).

October month is the busiest months with a maximum of 5317 bookings in the hotel followed by September and August

1.8 Which market segment do most of the guests come from?




```
market_segment_type
Online        23214
Offline       10528
Corporate      2017
Name: count, dtype: int64
```

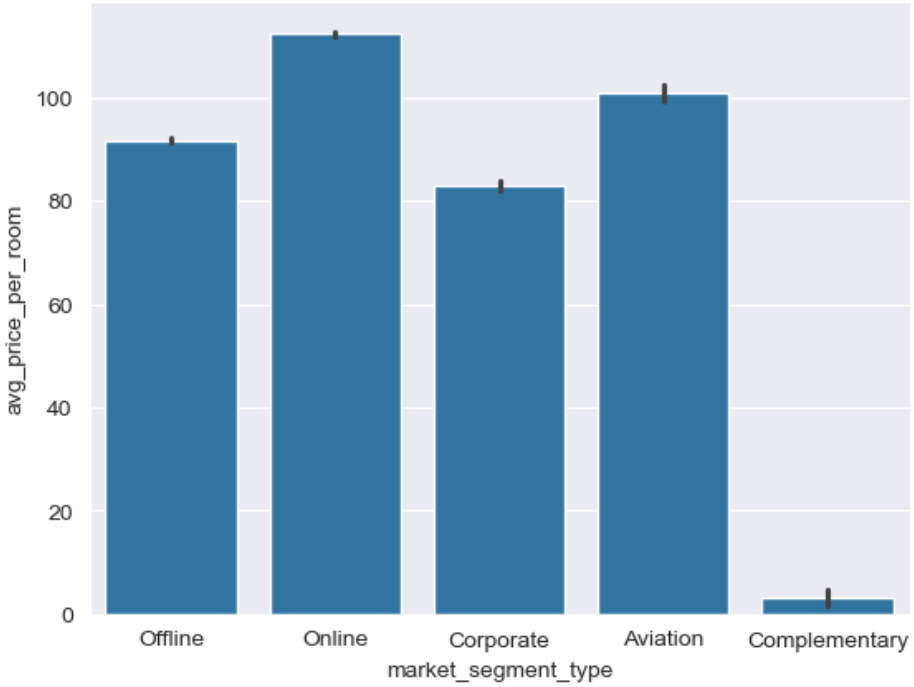
Figure 48: Labelled bar plot for market_segment_type

Observation:

- It is observed that the 64% of the bookings were made through onnline while 29% of bookings were made through offline

Most of the guests nearly 23214 bookings out of 36275 come from ONLINE market segment

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




```
market_segment_type
Aviation        100.70400
Complementary     3.14176
Corporate        82.91174
Offline          91.63268
Online          112.25685
Name: avg_price_per_room, dtype: float64
```

Figure 49: Bar plot for avg_price_per_room in the hotel

Observation:

- It is observed that the average room price of respective market segments are as follows:
    - Online - €112.26
    - Aviation - €100.70
    - Offline - €91.63
    - Corporate - €82.91
    - Complementary - €3.14

ONLINE bookings yield the highest average price per room = €112.26

1.10 What percentage of bookings are canceled?

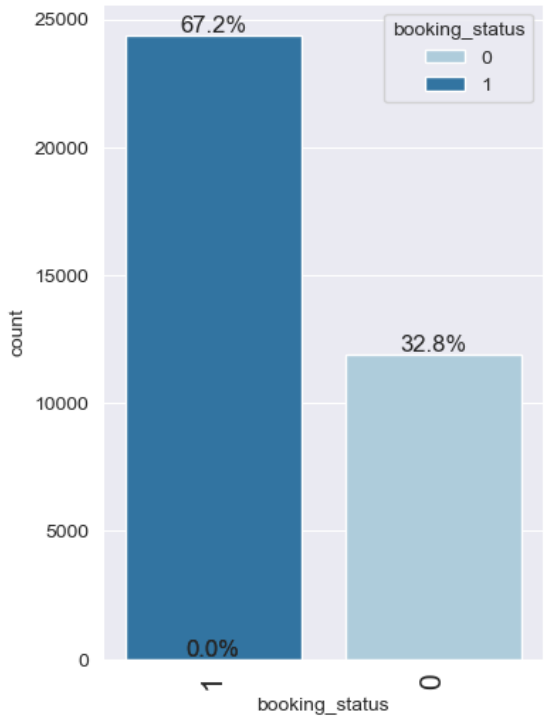


Figure 50: Labelled bar plot for booking_status in the hotel

Observation:

32.8% of the bookings were cancelled

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

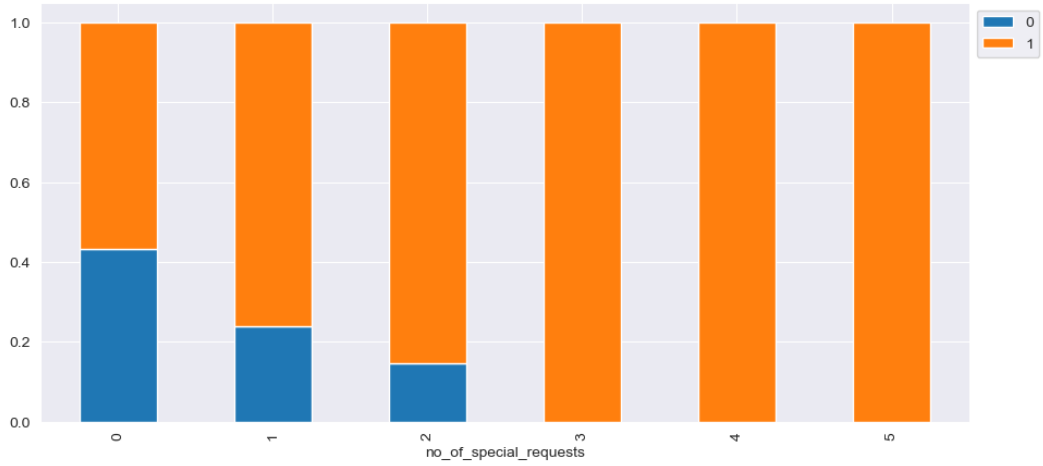
```
repeated_guest  booking_status
0               1                23476
                0                11869
1               1                  914
                0                   16
Name: count, dtype: int64
```

```
repeated_guest  booking_status
0               1                66.4%
                0                33.6%
1               1                98.3%
                0                 1.7%
Name: proportion, dtype: object
```

Table 9: Count and percentage of repeating guests in room cancellation

1.7% of the repeating guests cancel their booking at the hotel

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



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

Figure 51: Stacked bar plot for no_of_special_requests vs booking_status

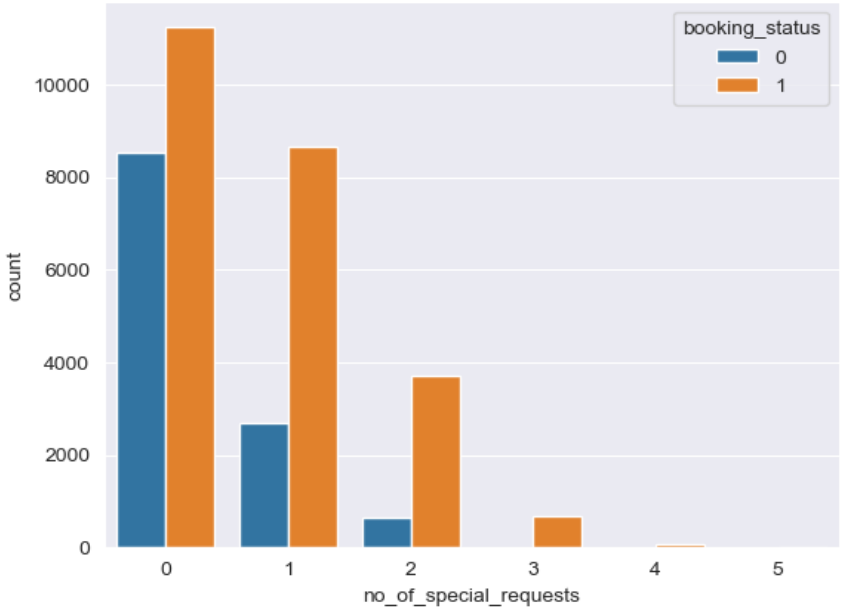


Figure 52: Count plot for no_of_special_requests with hue=' booking_status'

no_of_special_requests	booking_status	
0	1	56.8%
	0	43.2%
1	1	76.2%
	0	23.8%
2	1	85.4%
	0	14.6%
3	1	100.0%
4	1	100.0%
5	1	100.0%

Name: proportion, dtype: object

Table 9: Percentage of special request guests in room cancellation

Observation:

- As the number of special requests increases, the proportion of cancellations tends to decrease. This suggests that guests who make more special requests are less likely to cancel their bookings.

    - Bookings with no special requests, 43.22% were canceled.
    - Bookings with 1 special request, the cancellation rate decreased to 23.8%.
    - Bookings with 2 special requests, the cancellation rate is 14.6%.
    - Bookings with 3, 4, or 5 special requests had no cancellations at all

With the increase in the no. of special request by guests, there is decline in room cancellation

1.13 Insights based on EDA

- It is observed that the bookings with maximum of 4 adults, the cancellation is very less compared to bookings done with 1, 3, 0 and 2 adults
- These single adults have done their bookings mostly through online, offline or by corporate. And guest booking with 4 adults have done their bookings widely through online while few were through complementary
- It is also noted that complimentary market segments were not cancelled, while bookings done through online market segments face higher cancellations compared to other market segments
- When the bookings is done only by adults more than 3, then the chance of cancellation looks evident compared to 2 adults with children or only two adults.
- The bookings done by adults more than 2 are more likely to require a car parking space while majority of the bookings don't require a parking space
- The number of previous cancellations is high when bookings are done by single adults
- Bookings done with 0 to 2 children required a car parking space while bookings done with 3 or more children did not require a parking space

- And also bookings with 3 or more children had a lower average price per rooms
- It is observed that with the increase in the no. of special request, the proportion of cancellations tends to decrease. This suggests that guests who make more special requests are less likely to cancel their bookings.
- Complementary bookings never seem to face cancellations
- Corporate bookings are less likely to be cancelled, while Aviation comes the next with regard to less cancellations
- Online bookings are more prone to cancellations
- It is also noted that bookings requiring a car parking space tend to show up compared to bookings that do not require a parking space
- Moreover, the charge levied for the parking space along with higher average room price also increases the chance of booking cancellations
- Bookings done for a higher average price were prone to get cancelled positively
- Bookings done with the larger lead time were prone to get cancelled easily
- Bookings were done evenly throughout the month

2. DATA PRE-PROCESSING

2.1 Duplicate Value check

In order to build an efficient model it is essential to know that if the data set does not contain any duplicate values from the pre-existing rows. The command to check duplicate entries is duplicated().sum(). This returns the total number of duplicated entries in the data set. The provided INN Hotels Group data set does not contain any duplicated entries.

2.2 Missing value treatment

Another pre-processing step is to check if the provided data set has missed any values in any of the columns by using the isnull().sum() command. This command counts the missing values in each columns and returns the sum of missing values in each of the column respectively. From Table 5, it can be understood that there is no values missing in this data set.

2.3 Outlier treatment

We see there are so many outliers in each of the numerical column in the data set. So it's indeed essential to carefully examine the data set before treating the outliers. Upon observing and examining the dataset, the following conclusion is made with respect to outliers.

- From the above histogram and box plots, we see there are so many outliers in no_of_week_nights, no_of_adults, no_of_children, lead_time, avg_price_per_room, no_previous_cancellations, no_of_previous_bookings_not_canceled.
- These extreme values can be considered for model building, as treating them does not produce an efficient prediction on the model.
- Moreover, each booking is unique and it has wide variety of values depending upon the market segment and seasons and they provide a better understanding about the data.
- Hence such extreme values are considered essential to gain valuable insights on the model, thus leaving them untreated.

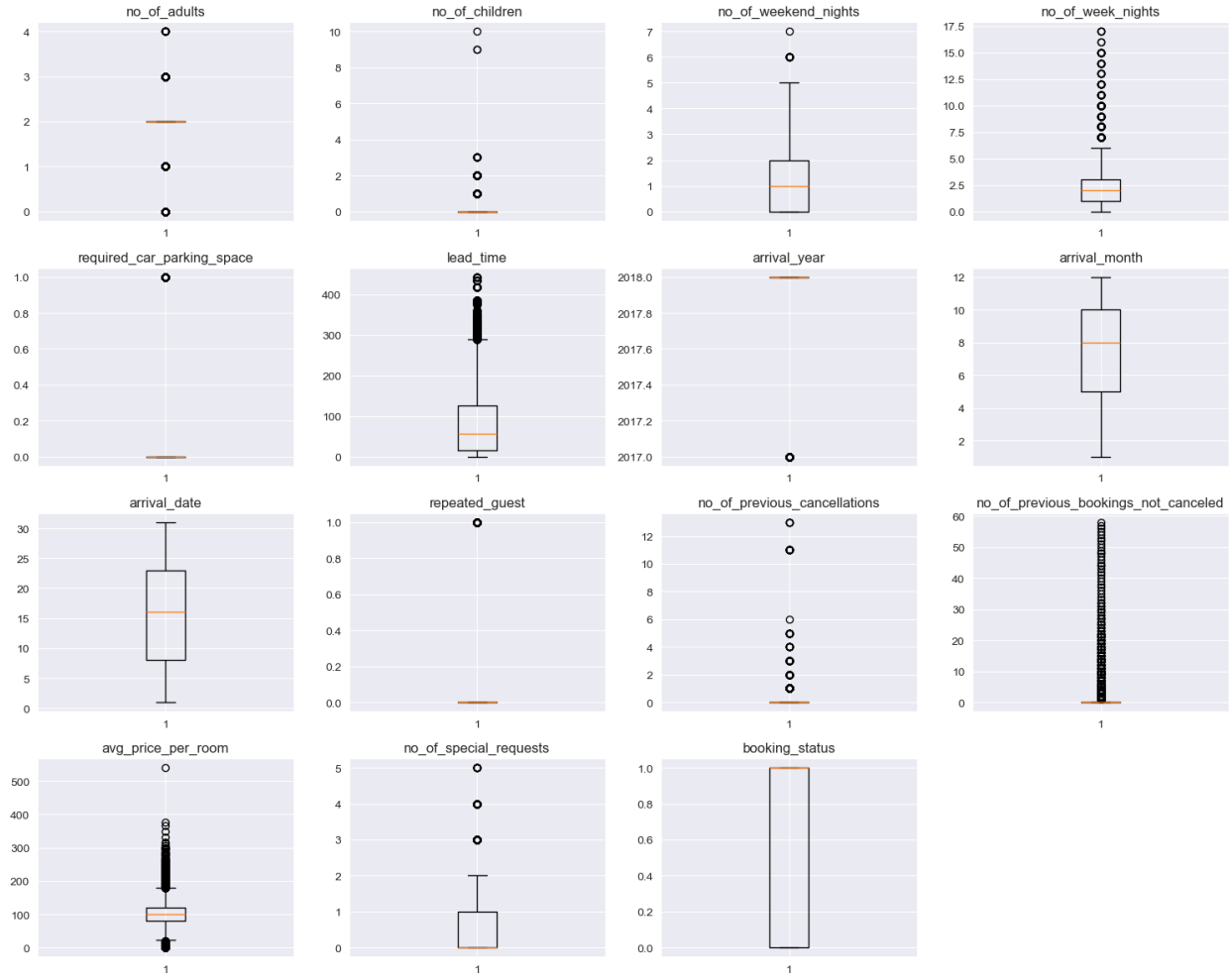


Figure 53: Outlier checks for numerical variables of the dataset

2.4 Feature engineering

- The "Booking_ID" column is unique and has no significance in model prediction. Hence the column is dropped.
- Considering the "booking_status" column from Table 4, it is an object type column. But in order to analyse its impact on various numerical and categorical variables, it was considered as a numerical column, throughout the Univariate and Bivariate analysis. Hence the column needs replacement of its categorical values of "Not_Canceled" with "1" and "Canceled" with "0"

	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arrival_year	arrival_month	arrival_date	market_segment_type	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests	booking_status
0	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	2017	10	2	Offline	0	0	0	65.00000	0	1
1	2	0	2	3	Not Selected	0	Room_Type 1	5	2018	11	6	Online	0	0	0	106.68000	1	1
2	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	2018	2	28	Online	0	0	0	60.00000	0	0
3	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	2018	5	20	Online	0	0	0	100.00000	0	0
4	2	0	1	1	Not Selected	0	Room_Type 1	48	2018	4	11	Online	0	0	0	94.50000	0	0

Table 11: Feature engineering with dropped and replaced columns

2.5 Data preparation for modelling

- We want to predict which factor is more likely cause the booking status to be cancelled.
- 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.

Create dummy variables

Values under categorical columns cannot be read into an equation. So one-hot encoding technique is applied to these categorical columns and it is established using a `get-dummies()` function in the pandas dataframe.

	0	1	2	3	4
const	1.00000	1.00000	1.00000	1.00000	1.00000
no_of_adults	2	2	1	2	2
no_of_children	0	0	0	0	0
no_of_weekend_nights	1	2	2	0	1
no_of_week_nights	2	3	1	2	1
lead_time	224	5	1	211	48
arrival_year	2017	2018	2018	2018	2018
arrival_month	10	11	2	5	4
arrival_date	2	6	28	20	11
no_of_previous_cancellations	0	0	0	0	0
no_of_previous_bookings_not_canceled	0	0	0	0	0
avg_price_per_room	65.00000	106.68000	60.00000	100.00000	94.50000
no_of_special_requests	0	1	0	0	0
type_of_meal_plan_Meal Plan 2	False	False	False	False	False
type_of_meal_plan_Meal Plan 3	False	False	False	False	False
type_of_meal_plan_Not Selected	False	True	False	False	True
required_car_parking_space_1	False	False	False	False	False
room_type_reserved_Room_Type 2	False	False	False	False	False
room_type_reserved_Room_Type 3	False	False	False	False	False
room_type_reserved_Room_Type 4	False	False	False	False	False
room_type_reserved_Room_Type 5	False	False	False	False	False
room_type_reserved_Room_Type 6	False	False	False	False	False
room_type_reserved_Room_Type 7	False	False	False	False	False
market_segment_type_Complementary	False	False	False	False	False
market_segment_type_Corporate	False	False	False	False	False
market_segment_type_Offline	True	False	False	False	False
market_segment_type_Online	False	True	True	True	True
repeated_guest_1	False	False	False	False	False

Table 12: Create dummy variables

Split the data

The entire data set is split into dependent and independent variables and the independent variables together is assigned to a variable and the dependent variable is assigned a variable.

In this data set,

Independent Variables	no_of_adults
	no_of_children
	no_of_weekend_nights
	no_of_week_nights
	lead_time
	arrival_year
	arrival_month
	arrival_date
	no_of_previous_cancellations
	no_of_previous_bookings_not_canceled
	avg_price_per_room
	no_of_special_requests
	type_of_meal_plan_Meal Plan 2
	type_of_meal_plan_Meal Plan 3
	type_of_meal_plan_Not Selected
	required_car_parking_space_1
	room_type_reserved_Room_Type 2
	room_type_reserved_Room_Type 3
	room_type_reserved_Room_Type 4
	room_type_reserved_Room_Type 5
	room_type_reserved_Room_Type 6
	room_type_reserved_Room_Type 7
	market_segment_type_Complementary
	market_segment_type_Corporate
	market_segment_type_Offline
	market_segment_type_Online
	repeated_guest_1
Dependent Variable	**booking_status**

Table 13: Split data- Dependent and independent variables

Train the data

The split data set as independent and dependent variables is further split into train and test datasets in 70:30 ratio

No. of rows in train data=25392

No. of rows in test data=10883

```
Shape of Training set :  (25392, 27)
Shape of test set :  (10883, 27)
Percentage of classes in training set:
booking_status
1.00000   0.67399
0.00000   0.32601
Name: proportion, dtype: float64
Percentage of classes in test set:
booking_status
1.00000   0.66857
0.00000   0.33143
Name: proportion, dtype: float64
```

Table 14: Shape of the training and test data set

Scaling the data

The data is scaled using the StandardScaler() function. The StandardScaler is used to standardize the input data in a way that ensures that the data points have a balanced scale, which is crucial for machine learning algorithms, especially those that are sensitive to differences in feature scales.

Add intercept to the data

After splitting the data, an intercept is added in order to train the data before building the model

3. MODEL BUILDING-LOGISTIC REGRESSION

3.1 Model evaluation criterion

Model can make wrong predictions as:

1. Predicting a person booking a room will cancel, but they do not. - False Negative - Loss of reputation
2. Predicting a person booking a room will not cancel, but they do. - False Positive - Loss of revenue

Which case is more important?

Both are important:

- If we predict the guest will cancel and then they do not, then we will reallocate their room to another guest and not have a room available to them upon their arrival. This costs the hotel a significant amount of money (by offering them a complimentary upgraded room), likely losing a repeated guest(s), and generating negative review(s) for the hotel.

- If we predict the guest will not cancel their booking but then they do, we will lose the revenue generated from their reservation, have to incur the costs of selling the room, and sometimes give the room for rebooking at a discount price.

How to reduce these costs i.e maximize True Positives?

- We need to reduce both False Negatives and False Positives

- **F1_score** should be maximized as the greater the f1_score, the higher the chances of reducing both False Negatives and False Positives and identifying both the classes correctly

- F1_score is computed as

$$f1_{score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

3.2 Build the Logistic Regression (Stats Model)

Adding constant to the data

	0	1	2	3	4
const	1.00000	1.00000	1.00000	1.00000	1.00000
no_of_adults	0.29850	0.29850	0.29850	0.29850	0.29850
no_of_children	-0.26164	-0.26164	-0.26164	-0.26164	2.22659
no_of_weekend_nights	0.21880	0.21880	0.21880	1.37121	-0.93361
no_of_week_nights	0.57188	0.57188	1.28792	-1.57623	1.28792
required_car_parking_space	-0.17990	-0.17990	-0.17990	-0.17990	-0.17990
lead_time	1.33640	-0.07477	-0.08644	-0.28470	1.34806
arrival_year	0.46936	0.46936	0.46936	-2.13056	0.46936
arrival_month	0.18828	-1.44595	-1.11911	0.84197	1.16882
arrival_date	1.53204	0.95931	-1.10252	-1.33161	-1.67525
repeated_guest	-0.16067	-0.16067	-0.16067	-0.16067	-0.16067
no_of_previous_cancellations	-0.06313	-0.06313	-0.06313	-0.06313	-0.06313
no_of_previous_bookings_not_canceled	-0.08587	-0.08587	-0.08587	-0.08587	-0.08587
avg_price_per_room	-0.35749	-0.35749	-0.11509	-0.35606	-0.60473
no_of_special_requests	-0.78611	0.48528	0.48528	-0.78611	3.02807
type_of_meal_plan_Meal Plan 2	-0.31846	-0.31846	-0.31846	-0.31846	-0.31846
type_of_meal_plan_Meal Plan 3	-0.00888	-0.00888	-0.00888	-0.00888	-0.00888
type_of_meal_plan_Not Selected	-0.40381	-0.40381	-0.40381	-0.40381	-0.40381
room_type_reserved_Room_Type 2	-0.14144	-0.14144	-0.14144	-0.14144	-0.14144
room_type_reserved_Room_Type 3	-0.01255	-0.01255	-0.01255	-0.01255	-0.01255
room_type_reserved_Room_Type 4	-0.44785	-0.44785	2.23290	-0.44785	-0.44785
room_type_reserved_Room_Type 5	-0.08590	-0.08590	-0.08590	-0.08590	-0.08590
room_type_reserved_Room_Type 6	-0.16425	-0.16425	-0.16425	-0.16425	-0.16425
room_type_reserved_Room_Type 7	-0.06626	-0.06626	-0.06626	-0.06626	-0.06626
market_segment_type_Complementary	-0.10406	-0.10406	-0.10406	-0.10406	-0.10406
market_segment_type_Corporate	-0.24402	-0.24402	-0.24402	-0.24402	-0.24402
market_segment_type_Offline	-0.64120	-0.64120	-0.64120	1.55958	-0.64120
market_segment_type_Online	0.75262	0.75262	0.75262	-1.32869	0.75262

Table 15: Adding constant to the data

Fit logistic regression model (with Statsmodel)

- We will now perform logistic regression using statsmodels, a Python module that provides functions for the estimation of many statistical models, as well as for conducting statistical tests, and statistical data exploration.

- Using statsmodels, we will be able to check the statistical validity of our model - identify the significant predictors from p-values that we get for each predictor variable.

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.424659
         Iterations: 35
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25364
Method:                           MLE   Df Model:                           27
Date:                Sat, 03 May 2025   Pseudo R-squ.:                 0.3273
Time:                        20:47:47   Log-Likelihood:                -10783.
converged:                      False   LL-Null:                       -16030.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=========================================================================================================
                                            coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     1.5417    229.185      0.007      0.995    -447.653     450.736
no_of_adults                             -0.0631      0.020     -3.231      0.001      -0.101      -0.025
no_of_children                           -0.0429      0.024     -1.799      0.072      -0.090       0.004
no_of_weekend_nights                     -0.1049      0.017     -6.095      0.000      -0.139      -0.071
no_of_week_nights                        -0.0447      0.017     -2.601      0.009      -0.078      -0.011
required_car_parking_space                0.2961      0.025     11.932      0.000       0.247       0.345
lead_time                                -1.3259      0.023    -58.373      0.000      -1.370      -1.281
arrival_year                             -0.1590      0.023     -6.945      0.000      -0.204      -0.114
arrival_month                             0.1310      0.020      6.592      0.000       0.092       0.170
arrival_date                             -0.0224      0.017     -1.326      0.185      -0.056       0.011
repeated_guest                            0.4028      0.103      3.908      0.000       0.201       0.605
no_of_previous_cancellations             -0.1008      0.029     -3.431      0.001      -0.158      -0.043
no_of_previous_bookings_not_canceled      0.1294      0.161      0.803      0.422      -0.186       0.445
avg_price_per_room                       -0.6538      0.026    -25.313      0.000      -0.704      -0.603
no_of_special_requests                    1.1687      0.024     49.096      0.000       1.122       1.215
type_of_meal_plan_Meal Plan 2            -0.0514      0.019     -2.687      0.007      -0.089      -0.014
type_of_meal_plan_Meal Plan 3            -0.1772    134.634     -0.001      0.999    -264.056     263.701
type_of_meal_plan_Not Selected           -0.0667      0.018     -3.615      0.000      -0.103      -0.031
room_type_reserved_Room_Type 2            0.0602      0.018      3.325      0.001       0.025       0.096
room_type_reserved_Room_Type 3           -0.0138      0.025     -0.555      0.579      -0.063       0.035
room_type_reserved_Room_Type 4            0.1065      0.020      5.370      0.000       0.068       0.145
room_type_reserved_Room_Type 5            0.0585      0.018      3.309      0.001       0.024       0.093
room_type_reserved_Room_Type 6            0.1492      0.024      6.205      0.000       0.102       0.196
room_type_reserved_Room_Type 7            0.0816      0.020      4.069      0.000       0.042       0.121
market_segment_type_Complementary         3.4873   2210.577      0.002      0.999   -4329.165    4336.139
market_segment_type_Corporate             0.1971      0.063      3.125      0.002       0.073       0.321
market_segment_type_Offline               0.8349      0.119      6.992      0.000       0.601       1.069
market_segment_type_Online                0.0219      0.125      0.175      0.861      -0.223       0.266
=========================================================================================================
```

Table 16: Logistic regression model summary

Checking Logistic Regression model performance on training set

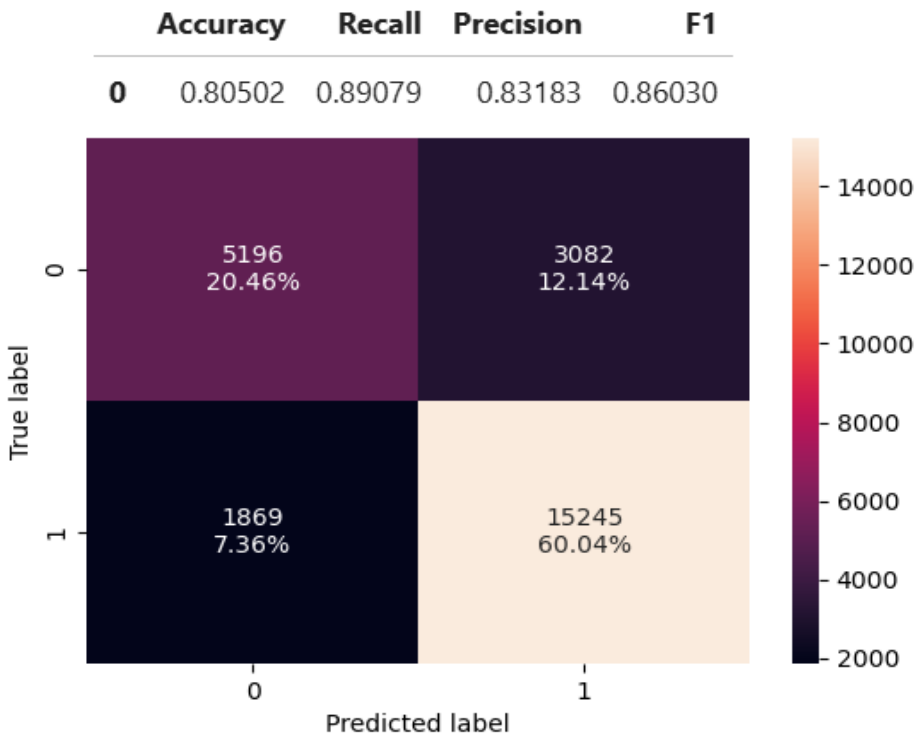


Figure 54: Confusion matrix of LR model performance on training set

Checking Logistic Regression model performance on testing set

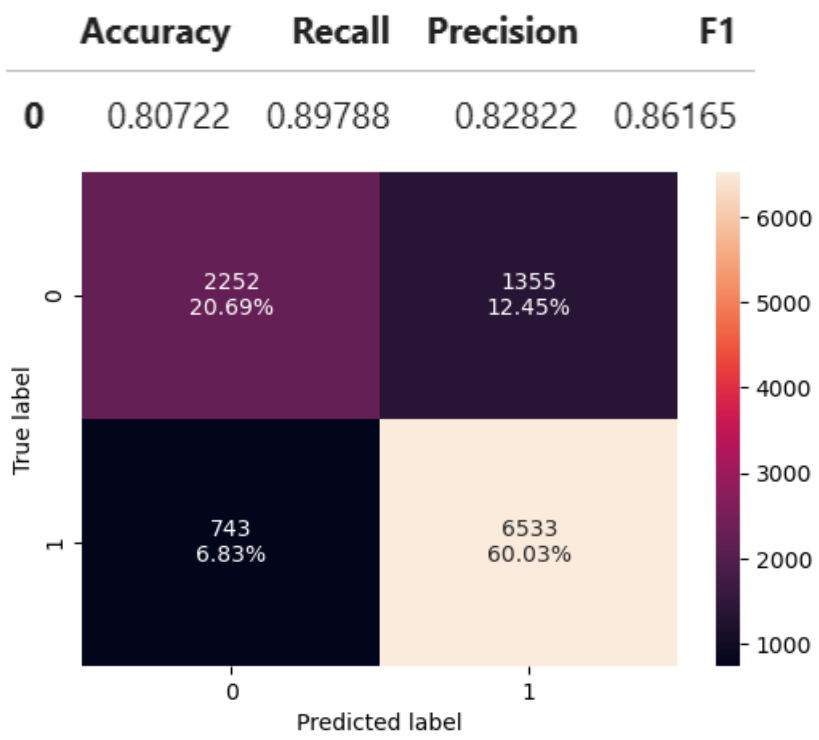


Figure 55: Confusion matrix of LR model performance on testing set

- **The Logistic Regression model has a very good performance on both training and test sets with a F1_score of 0.86**
- **The model shows a strong balance between the precision and recall with 0.83 and 0.89 values respectively on both training and testing tests**
- **The model also shows a good accuracy values of 0.80 on both training and testing tests**

The Logistic Regression model can be further fine-tuned to improve its performance by dealing with multicollinearity, dealing with high P-value variables, determining the optimal threshold using ROC curve.

3.3 Build the Decision Tree Classifier (sklearn)

- We will now build the decision tree model for prediction using DecisionTreeClassifier from the sklearn library.

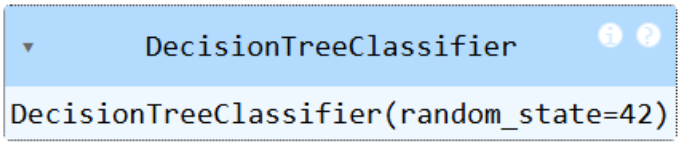


Figure 56: Decision tree model summary

Checking Decision tree model performance on training set

	Accuracy	Recall	Precision	F1
0	0.99429	0.99562	0.99591	0.99576

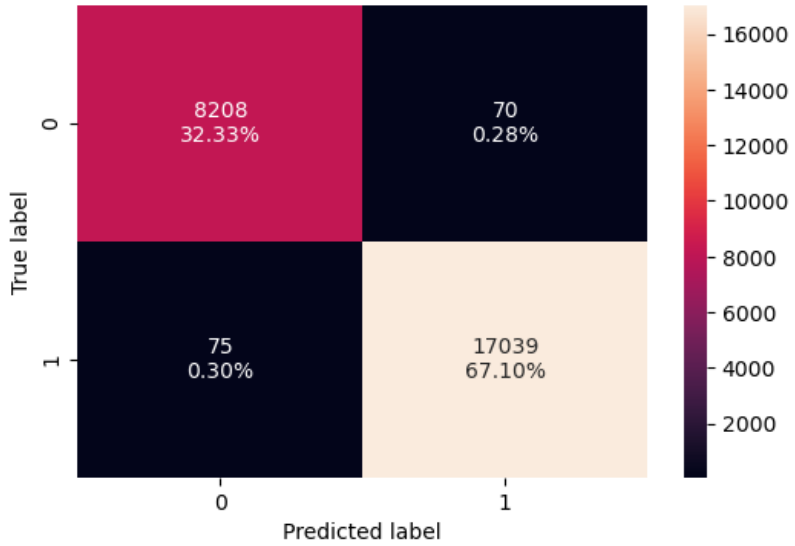


Figure 57: Confusion matrix of DT model performance on training set

- Model has performed very well on the training set.
- As we know a decision tree will continue to grow and classify each data point correctly if no restrictions are applied as the trees will learn all the patterns in the training set.
- Let's check the performance on test data to see if the model is overfitting.

Checking Decision tree model performance on testing set

	Accuracy	Recall	Precision	F1
0	0.86474	0.89335	0.90328	0.89829

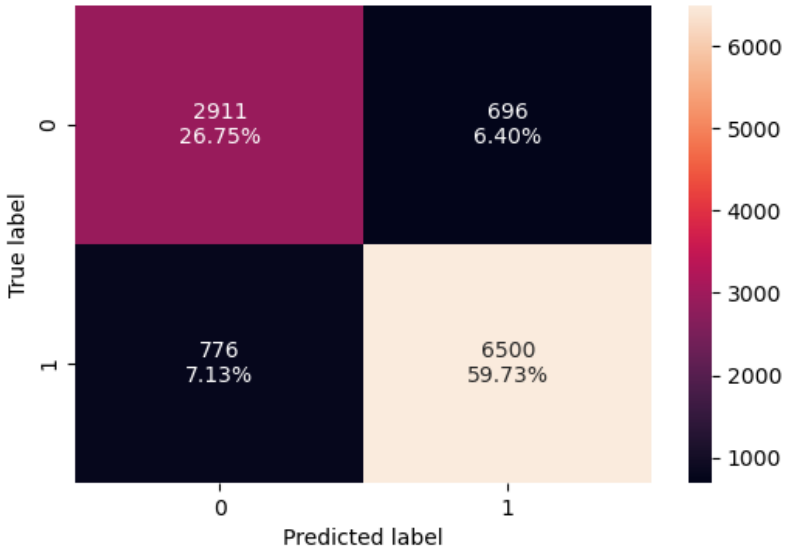


Figure 58: Confusion matrix of DT model performance on testing set

- 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.

Model Performance before tuning

Training and Test performance comparison of Pretuned models:

	LR_Pretune_Train	LR_Pretune_Test	DT_Pretune_Train	DT_Pretune_Test
Accuracy	0.80502	0.80722	0.99429	0.86474
Recall	0.89079	0.89788	0.99562	0.89335
Precision	0.83183	0.82822	0.99591	0.90328
F1	0.86030	0.86165	0.99576	0.89829

Table 17: Training and Test performance of pretuned logistic regression and decision tree models

3.4 Comment on the model performance across different metrics

- The linear regression model was able to classify that were correct, whether positive or negative with an **accuracy** of **0.80** both in its test and training sets
- The **accuracy** values considered in decision tree model shows that it performed well with training set but obtains a value of only **0.86** in test set. This indicates that the model was not able to generalize well on test set
- The linear regression model was able to classify the actual positives correctly as positives, thus showing a **recall** value of **0.89** both in its test and training sets
- The **recall** values obtained in decision tree model portrays that it performed well with training with a score of **0.99**, but was not able to do similarly on the test set, thus shows a value of **0.89**
- The linear regression model's positive predictions were accurate with the **precision** value of **0.83** both in its training and test sets
- The decision tree model's **precision** score on its training set is **0.99**, while the value is **0.90** on its test set, thus proves a less performance on test set compared to the training set
- The balance between the precision and recall portrayed by the **F1_score** for the linear regression model is **0.86** both in its training and test sets
- The **F1_score** for the decision tree model is **0.99** on the training set and **0.89** on the test set, which shows that model seems to be overfitting
- Looking into the performance of the model through different metrics, it can be understood that both the models need to be fine-tuned to understand and improve its performance

4. MODEL PERFORMANCE IMPROVEMENT

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

<u>Logistic Regression - Dealing with Multicollinearity</u>

- Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the correlation between variables is high, it can cause problems when we fit the model and interpret the results. When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

- There are different ways of detecting (or testing) multicollinearity. One such way is by using the Variance Inflation Factor, or VIF.

- **Variance Inflation Factor (VIF):** Variance inflation factors measure the inflation in the variances of the regression parameter estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient β_k is "inflated" by the existence of correlation among the predictor variables in the model.
  - If VIF is 1, then there is no correlation among the k^{th} predictor and the remaining predictor variables, and hence, the variance of β_k is not inflated at all.

- **General Rule of thumb:**
  - If VIF is between 1 and 5, then there is low multicollinearity.
  - If VIF is between 5 and 10, we say there is moderate multicollinearity.
  - If VIF is exceeding 10, it shows signs of high multicollinearity.

Variance Inflation Factors:

	Variable	VIF
0	const	1.00000
1	no_of_adults	1.34299
2	no_of_children	2.00637
3	no_of_weekend_nights	1.06511
4	no_of_week_nights	1.09149
5	required_car_parking_space	1.03707
6	lead_time	1.39281
7	arrival_year	1.43007
8	arrival_month	1.27171
9	arrival_date	1.00666
10	repeated_guest	1.76765
11	no_of_previous_cancellations	1.37023
12	no_of_previous_bookings_not_canceled	1.61806
13	avg_price_per_room	2.04856
14	no_of_special_requests	1.24585
15	type_of_meal_plan_Meal Plan 2	1.26460
16	type_of_meal_plan_Meal Plan 3	1.00610
17	type_of_meal_plan_Not Selected	1.27551
18	room_type_reserved_Room_Type 2	1.09146
19	room_type_reserved_Room_Type 3	1.00490
20	room_type_reserved_Room_Type 4	1.35628
21	room_type_reserved_Room_Type 5	1.03320
22	room_type_reserved_Room_Type 6	1.98810
23	room_type_reserved_Room_Type 7	1.10446
24	market_segment_type_Complementary	4.49253
25	market_segment_type_Corporate	17.21348
26	market_segment_type_Offline	64.41691
27	market_segment_type_Online	71.43903

Table 17: Variables with high VIFs

- **Dropping columns with VIF>5**
  It is noted that market_segment_type variables have a very high VIFs. Hence the columns with VIF>5 were dropped iteratively to get the final dataset with dropped columns

Dropping market_segment_type_Online due to high VIF

Variance Inflation Factors:

	Variable	VIF
0	no_of_adults	1.32591
1	no_of_children	2.00540
2	no_of_weekend_nights	1.06439
3	no_of_week_nights	1.09028
4	required_car_parking_space	1.03704
5	lead_time	1.38817
6	arrival_year	1.42771
7	arrival_month	1.27096
8	arrival_date	1.00664
9	repeated_guest	1.76569
10	no_of_previous_cancellations	1.37006
11	no_of_previous_bookings_not_canceled	1.61793
12	avg_price_per_room	2.04799
13	no_of_special_requests	1.24130
14	type_of_meal_plan_Meal Plan 2	1.26427
15	type_of_meal_plan_Meal Plan 3	1.00610
16	type_of_meal_plan_Not Selected	1.27359
17	room_type_reserved_Room_Type 2	1.09127
18	room_type_reserved_Room_Type 3	1.00490
19	room_type_reserved_Room_Type 4	1.35202
20	room_type_reserved_Room_Type 5	1.03320
21	room_type_reserved_Room_Type 6	1.98782
22	room_type_reserved_Room_Type 7	1.10434
23	market_segment_type_Complementary	1.33887
24	market_segment_type_Corporate	1.53973
25	market_segment_type_Offline	1.60447

Table 18: Variables after dropping market_segment_type_online

- As the dataset had multicollinearity issue, the p-values will also change.
- We dealt with multicollinearity in order to interpret the p-values.

Dealing with high p-value variables

- Some of the dummy variables in the data have p-value > 0.05. So, they are not significant and we'll drop them

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.424659
         Iterations: 35
Dropping column type_of_meal_plan_Meal Plan 3 with p-value: 0.9989497009712687
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.424703
         Iterations: 35
Dropping column market_segment_type_Complementary with p-value: 0.9998497560624925
Optimization terminated successfully.
         Current function value: 0.425167
         Iterations 10
Dropping column room_type_reserved_Room_Type 3 with p-value: 0.6591034670615359
Optimization terminated successfully.
         Current function value: 0.425171
         Iterations 10
Dropping column no_of_previous_bookings_not_canceled with p-value: 0.4190723002504886
Optimization terminated successfully.
         Current function value: 0.425190
         Iterations 9
Dropping column arrival_date with p-value: 0.19830155371616764
Optimization terminated successfully.
         Current function value: 0.425223
         Iterations 9
Dropping column market_segment_type_Corporate with p-value: 0.16086258278486953
Optimization terminated successfully.
         Current function value: 0.425260
         Iterations 9
Dropping column no_of_children with p-value: 0.07881927209324423
Optimization terminated successfully.
         Current function value: 0.425319
         Iterations 9
```

Table 19: Dropping of high P_value columns

- **Selected_features**

```
['const', 'no_of_adults', 'no_of_weekend_nights', 'no_of_week_nights', 'required_car_parking_space', 'lead_time', 'arrival_year', 'arrival_month', 'repeated_guest', 'no_of_previous_cancellations', 'avg_price_per_room', 'no_of_special_requests', 'type_of_meal_plan_Meal Plan 2', 'type_of_meal_plan_Not Selected', 'room_type_reserved_Room_Type 2', 'room_type_reserved_Room_Type 4', 'room_type_reserved_Room_Type 5', 'room_type_reserved_Room_Type 6', 'room_type_reserved_Room_Type 7', 'market_segment_type_Offline', 'market_segment_type_Online']
```

	0	1	2	3	4	5	6	7	8	9
const	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
no_of_adults	0.29850	0.29850	0.29850	0.29850	0.29850	0.29850	0.29850	-1.62816	-1.62816	0.29850
no_of_weekend_nights	0.21880	0.21880	0.21880	1.37121	-0.93361	-0.93361	-0.93361	0.21880	0.21880	0.21880
no_of_week_nights	0.57188	0.57188	1.28792	-1.57623	1.28792	-0.14416	-0.86019	-0.86019	-0.14416	-0.14416
required_car_parking_space	-0.17990	-0.17990	-0.17990	-0.17990	-0.17990	-0.17990	-0.17990	-0.17990	-0.17990	-0.17990
lead_time	1.33640	-0.07477	-0.08644	-0.28470	1.34806	-0.85617	-0.90282	2.51432	-0.26138	-0.33135
arrival_year	0.46936	0.46936	0.46936	-2.13056	0.46936	0.46936	0.46936	0.46936	0.46936	0.46936
arrival_month	0.18828	-1.44595	-1.11911	0.84197	1.16882	-1.77280	-1.77280	-0.13857	-0.79226	-0.46541
repeated_guest	-0.16067	-0.16067	-0.16067	-0.16067	-0.16067	-0.16067	-0.16067	-0.16067	-0.16067	-0.16067
no_of_previous_cancellations	-0.06313	-0.06313	-0.06313	-0.06313	-0.06313	-0.06313	-0.06313	-0.06313	-0.06313	-0.06313
avg_price_per_room	-0.35749	-0.35749	-0.11509	-0.35606	-0.60473	1.92533	-0.52717	-0.38458	0.75612	-0.36547
no_of_special_requests	-0.78611	0.48528	0.48528	-0.78611	3.02807	-0.78611	-0.78611	-0.78611	0.48528	0.48528
type_of_meal_plan_Meal Plan 2	-0.31846	-0.31846	-0.31846	-0.31846	-0.31846	-0.31846	-0.31846	3.14015	-0.31846	-0.31846
type_of_meal_plan_Not Selected	-0.40381	-0.40381	-0.40381	-0.40381	-0.40381	-0.40381	-0.40381	-0.40381	-0.40381	-0.40381
room_type_reserved_Room_Type 2	-0.14144	-0.14144	-0.14144	-0.14144	-0.14144	-0.14144	-0.14144	-0.14144	-0.14144	-0.14144
room_type_reserved_Room_Type 4	-0.44785	-0.44785	2.23290	-0.44785	-0.44785	-0.44785	-0.44785	-0.44785	-0.44785	-0.44785
room_type_reserved_Room_Type 5	-0.08590	-0.08590	-0.08590	-0.08590	-0.08590	-0.08590	-0.08590	-0.08590	-0.08590	-0.08590
room_type_reserved_Room_Type 6	-0.16425	-0.16425	-0.16425	-0.16425	-0.16425	6.08843	-0.16425	-0.16425	-0.16425	-0.16425
room_type_reserved_Room_Type 7	-0.06626	-0.06626	-0.06626	-0.06626	-0.06626	-0.06626	-0.06626	-0.06626	-0.06626	-0.06626
market_segment_type_Offline	-0.64120	-0.64120	-0.64120	1.55958	-0.64120	-0.64120	-0.64120	1.55958	-0.64120	1.55958
market_segment_type_Online	0.75262	0.75262	0.75262	-1.32869	0.75262	0.75262	0.75262	-1.32869	-1.32869	-1.32869

Table 20: Dataset with only significant features

Training the Logistic Regression model again with only the significant features

```
Optimization terminated successfully.
         Current function value: 0.425319
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                         Logit   Df Residuals:                    25371
Method:                          MLE   Df Model:                           20
Date:                Sun, 04 May 2025   Pseudo R-squ.:                  0.3263
Time:                        16:02:12   Log-Likelihood:                -10800.
converged:                       True   LL-Null:                       -16030.
Covariance Type:            nonrobust   LLR p-value:                     0.000
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                            1.1866      0.023     50.623      0.000       1.141       1.233
no_of_adults                    -0.0557      0.019     -2.881      0.004      -0.094      -0.018
no_of_weekend_nights            -0.1087      0.017     -6.330      0.000      -0.142      -0.075
no_of_week_nights               -0.0492      0.017     -2.863      0.004      -0.083      -0.016
required_car_parking_space       0.2957      0.025     11.926      0.000       0.247       0.344
lead_time                       -1.3237      0.023    -58.439      0.000      -1.368      -1.279
arrival_year                    -0.1599      0.023     -7.008      0.000      -0.205      -0.115
arrival_month                    0.1338      0.020      6.748      0.000       0.095       0.173
repeated_guest                   0.4443      0.096      4.623      0.000       0.256       0.633
no_of_previous_cancellations    -0.0971      0.028     -3.411      0.001      -0.153      -0.041
avg_price_per_room              -0.6681      0.025    -26.272      0.000      -0.718      -0.618
no_of_special_requests           1.1664      0.024     49.137      0.000       1.120       1.213
type_of_meal_plan_Meal Plan 2   -0.0496      0.019     -2.596      0.009      -0.087      -0.012
type_of_meal_plan_Not Selected  -0.0663      0.018     -3.605      0.000      -0.102      -0.030
room_type_reserved_Room_Type 2   0.0521      0.018      2.958      0.003       0.018       0.087
room_type_reserved_Room_Type 4   0.1075      0.020      5.450      0.000       0.069       0.146
room_type_reserved_Room_Type 5   0.0605      0.018      3.431      0.001       0.026       0.095
room_type_reserved_Room_Type 6   0.1251      0.019      6.594      0.000       0.088       0.162
room_type_reserved_Room_Type 7   0.0776      0.020      3.937      0.000       0.039       0.116
market_segment_type_Offline      0.4634      0.046     10.139      0.000       0.374       0.553
market_segment_type_Online      -0.3699      0.046     -7.972      0.000      -0.461      -0.279
================================================================================================
```

Table 21: Trained logistic regression model with only significant features

Determining optimal threshold using ROC Curve

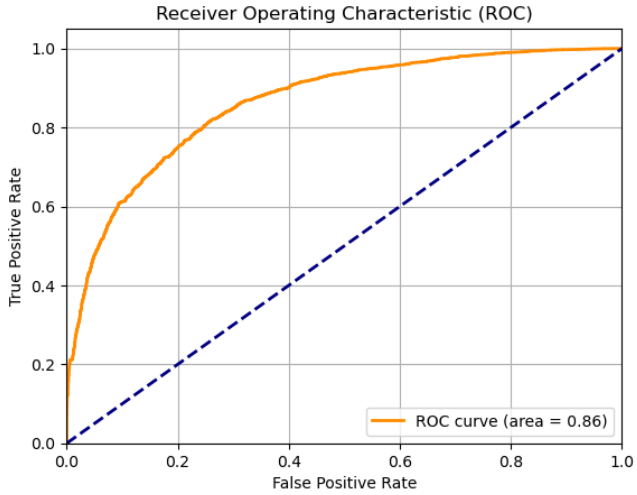


Figure 59: ROC Curve

Optimal Threshold: 0.637

Checking tuned Logistic Regression model performance on training set

	Accuracy	Recall	Precision	F1
0	0.79214	0.81647	0.86735	0.84114

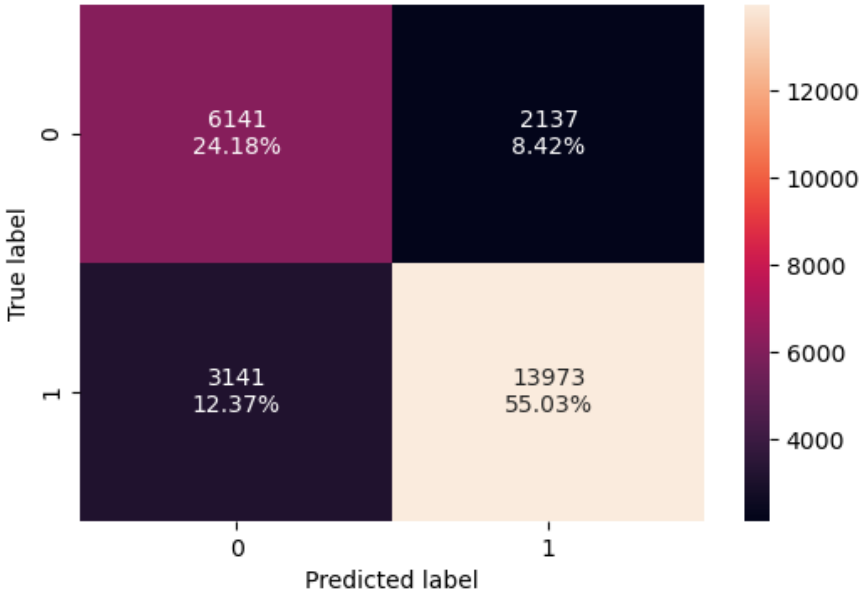


Figure 60: Confusion matrix of tuned LR model performance on training set

Checking tuned Logistic Regression model performance on test set

	Accuracy	Recall	Precision	F1
0	0.79261	0.82614	0.85835	0.84194

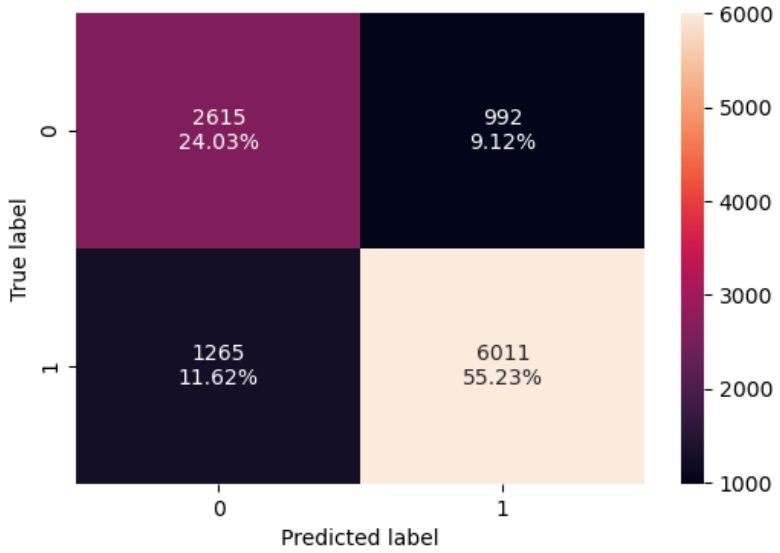


Figure 61: Confusion matrix of tuned LR model performance on testing set

- The performance of the Logistic Regression model has been fine_tuned considerably (on both training and test sets) by dealing with multicollinearity, dropping variables with high p-values and setting an optimal threshold.

- **The Logistic Regression model has a good performance on both training and test sets with a F1_score of 0.84 after tuning by dealing with multicollinearity, dropping variables with high p-values and setting an optimal threshold value of 0.637**
- **The model shows a strong balance between the precision and recall with 0.86 and 0.82 values respectively on both training and testing tests**
- **The model also shows a good accuracy values of 0.80 on both training and testing tests**
- **The accuracy value of the model remains the same even after tuning the model**

Comments

- Based on these results, we can conclude that the logistic regression model with a threshold of 0.5 has the best performance across various metrics, including accuracy, recall, precision, and F1 score. Therefore, it is recommended as the optimal model for predicting booking cancellations in this scenario.

- This means that if the predicted probability of a booking being canceled exceeds 50%, we will classify it as a cancellation. If the predicted probability falls below 50%, we will classify it as a non-cancellation. This threshold selection allows us to strike a balance between accurately identifying cancellations (high recall) and minimizing false positives (high precision).

4.2 Decision Tree Classifier (pre-pruning and post-pruning)

Pre - pruning the tree

- We will now build the pre-pruned decision tree model for prediction using DecisionTreeClassifier from the sklearn library.
- The max_depth of the tree is set to 11, with 100 max_leaf_nodes, 30 min_samples_split.

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=11, max_leaf_nodes=100, min_samples_split=30,
                       random_state=42)
```

Checking pre-pruned Decision Tree Classifier performance on training set

	Accuracy	Recall	Precision	F1
0	0.87571	0.92906	0.89116	0.90972

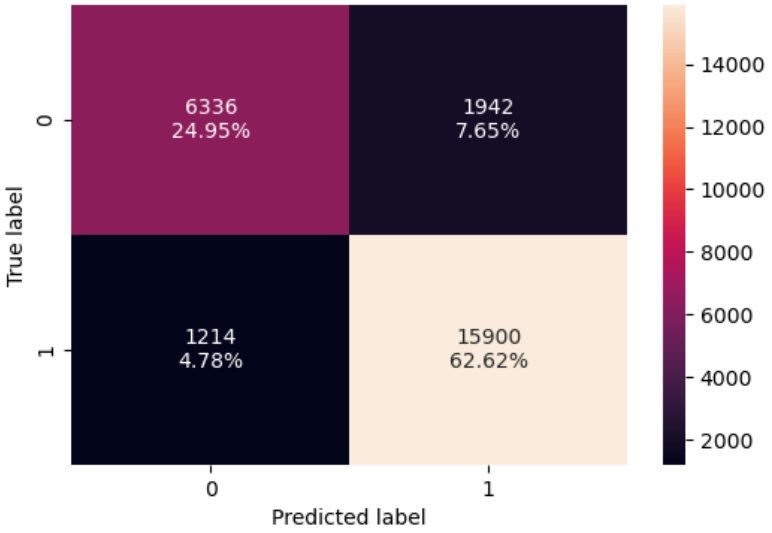


Figure 62: Confusion matrix of pre-pruned DT model performance on training set

Checking pre-pruned Decision Tree Classifier performance on test set

	Accuracy	Recall	Precision	F1
0	0.87384	0.93073	0.88627	0.90796

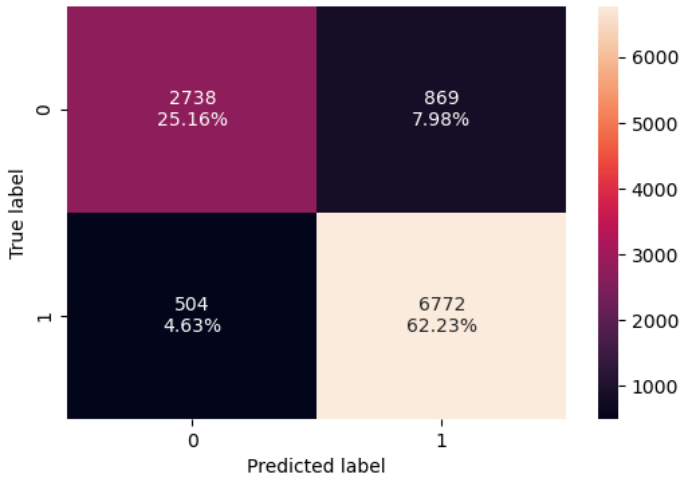


Figure 63: Confusion matrix of pre-pruned DT model performance on testing set

- **The decision tree before pre-pruning was not able to generalize well on the test test and was found overfitting.**
- **The Decision Tree model after pre-pruning has a good performance on both training and test sets with a F1_score of 0.90**
- **The model shows a strong balance between the precision and recall with 0.89 and 0.93 values respectively on both training and testing tests**
- **The model also shows a good accuracy values of 0.87 on both training and testing tests**

Visualizing the Decision Tree

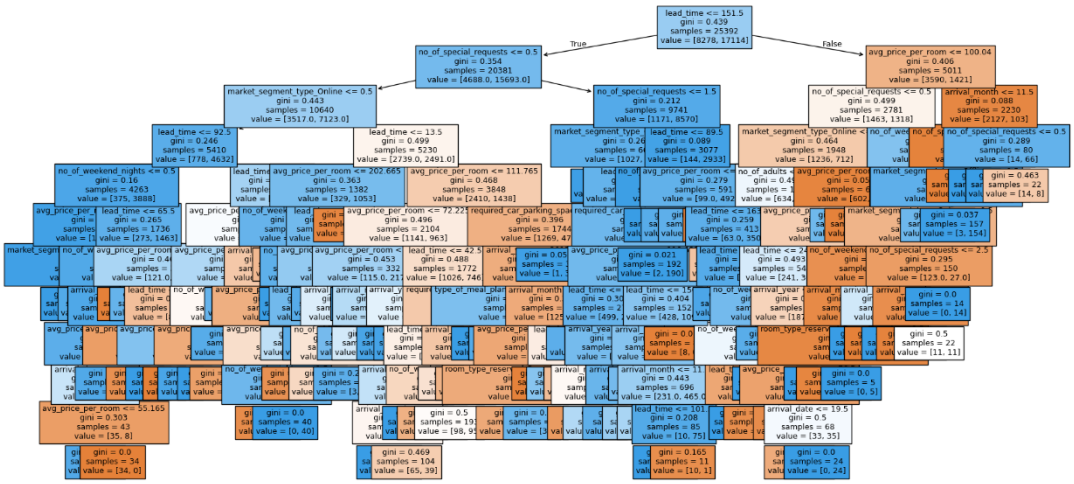


Figure 64: Pre-Pruned decision tree

Analyzing Feature Importance for tuned Decision Tree Classifier

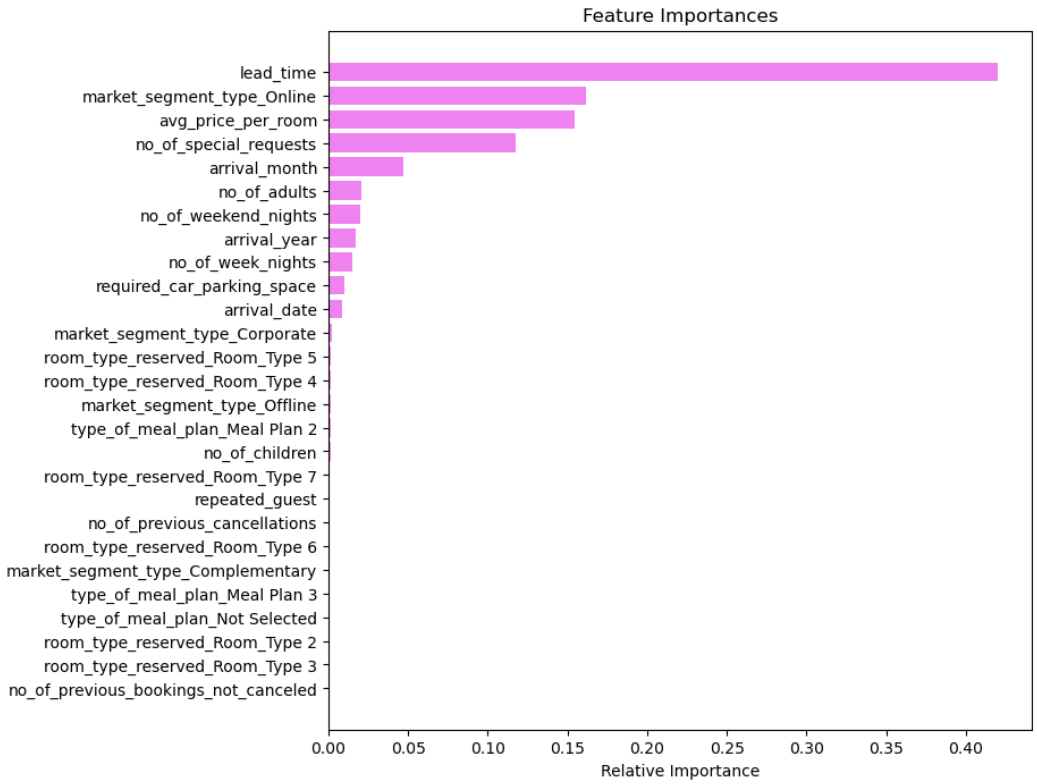


Figure 65: Feature importance of the Pre-Pruned decision tree

According to the pre-tuned decision tree model Lead Time is the most important variable for predicting the booking_status

Post - pruning the tree

- We calculate the ccp_alphas

	ccp_alphas	impurities
0	0.00000	0.00761
1	-0.00000	0.00761
2	0.00000	0.00762
3	0.00000	0.00762
4	0.00000	0.00762
...	...	...
1340	0.00709	0.28665
1341	0.01207	0.29872
1342	0.01784	0.31655
1343	0.02397	0.36450
1344	0.07495	0.43945

Table 22: CCP_alphas

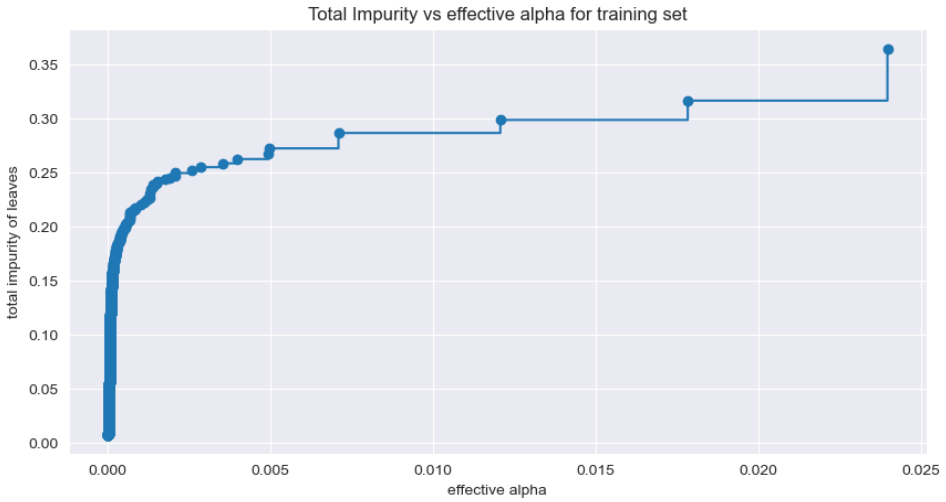


Figure 66: Total impurity vs effective alpha for training set

- Next, we train a decision tree 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: 0.07495203066381129
- 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.

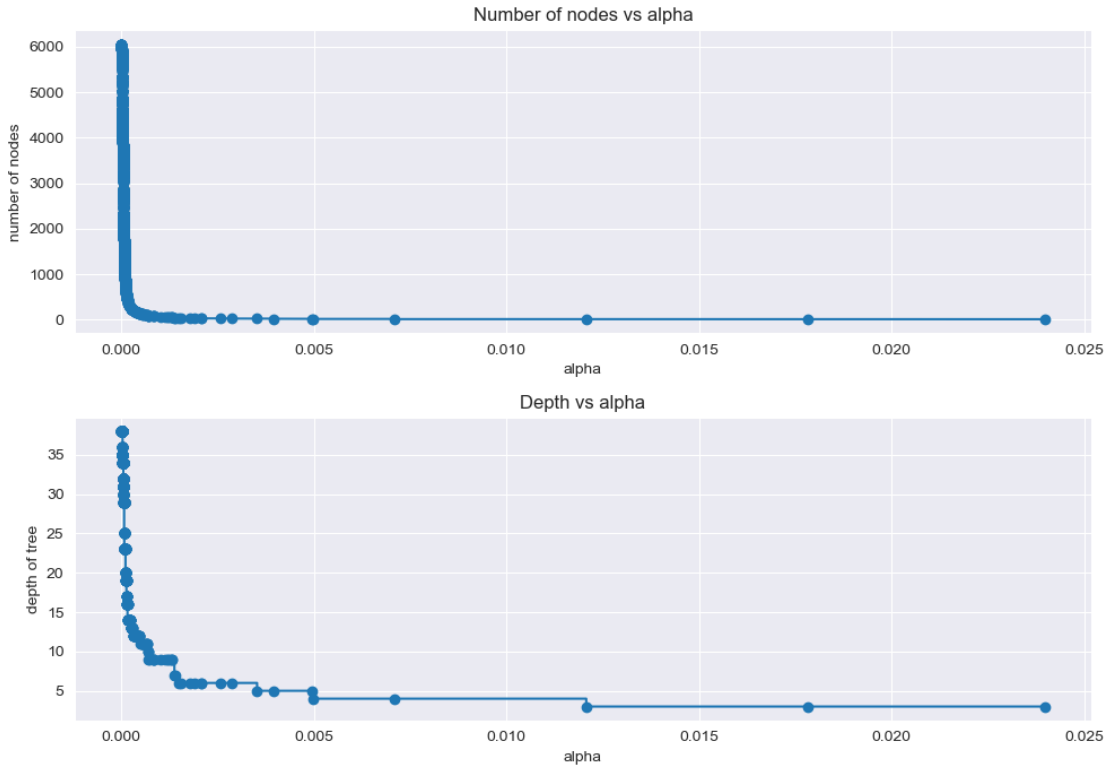


Figure 67: No. of nodes vs alpha and Tree depth vs alpha

Recall vs alpha for training and test sets

Figure 68: Recall vs alpha for training and test sets

Post-pruning using ccp_alpha

```
DecisionTreeClassifier(ccp_alpha=0.0002782421404924967, random_state=1)
```

```
DecisionTreeClassifier(ccp_alpha=0.07495203066381129,
                       class_weight={0: 0.126, 1: 0.874}, random_state=1)
```

- Post-pruning using ccp alpha returns the same model as the initial model (Tree with no pruning).
- As post pruning model is the same as the initial decision tree mode, the performance and feature importance will also be the same.

Checking post-pruned Decision Tree Classifier performance on training set

	Accuracy	Recall	Precision	F1
0	0.67399	1.00000	0.67399	0.80525

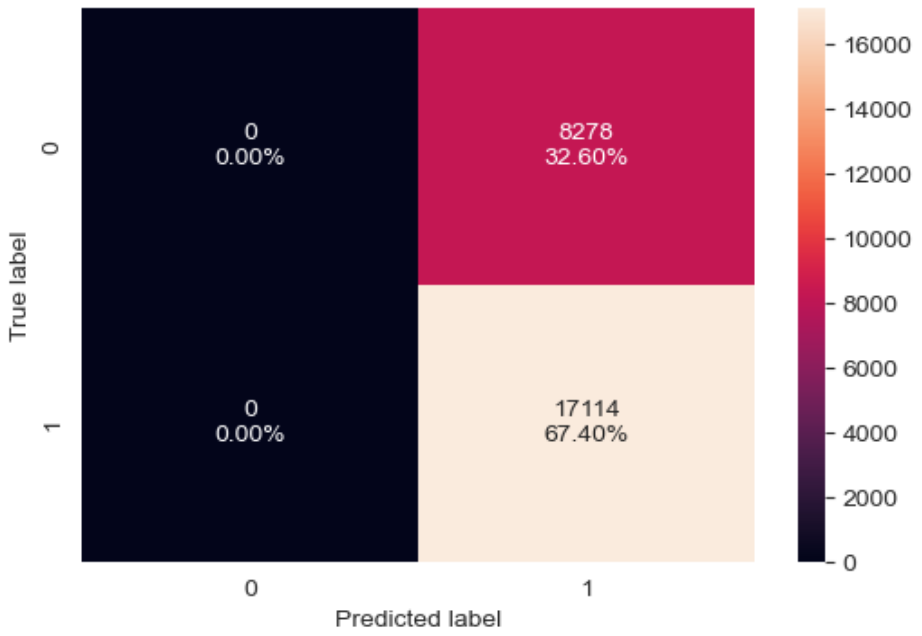


Figure 69: Confusion matrix of post-pruned DT model performance on training set

Checking post-pruned Decision Tree Classifier performance on test set

	Accuracy	Recall	Precision	F1
0	0.66857	1.00000	0.66857	0.80137

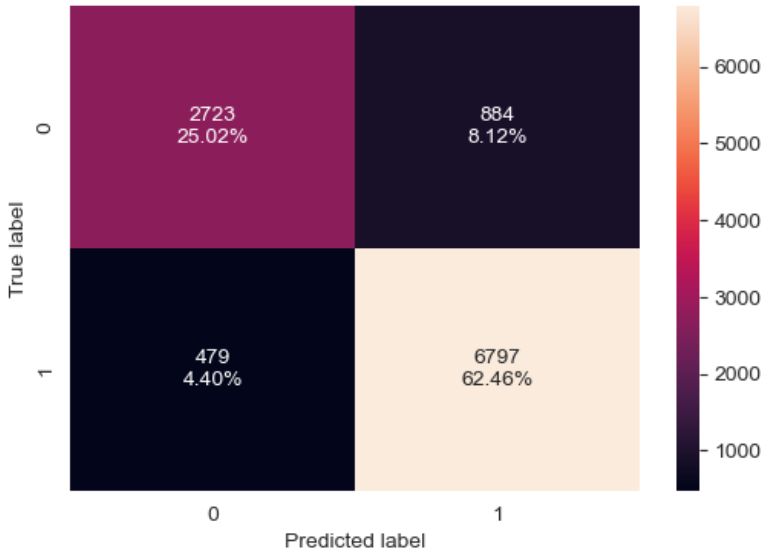
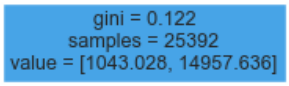


Figure 70: Confusion matrix of post-pruned DT model performance on training set

Visualizing the Decision Tree




```
|--- weights: [1043.03, 14957.64] class: 1.0
```

Figure 71: Post-Pruned decision tree

Analyzing Feature Importance for tuned Decision Tree Classifier

	Imp
no_of_adults	0.00000
type_of_meal_plan_Meal Plan 2	0.00000
market_segment_type_Offline	0.00000
market_segment_type_Corporate	0.00000
market_segment_type_Complementary	0.00000
room_type_reserved_Room_Type 7	0.00000
room_type_reserved_Room_Type 6	0.00000
room_type_reserved_Room_Type 5	0.00000
room_type_reserved_Room_Type 4	0.00000
room_type_reserved_Room_Type 3	0.00000
room_type_reserved_Room_Type 2	0.00000
type_of_meal_plan_Not Selected	0.00000
type_of_meal_plan_Meal Plan 3	0.00000
no_of_special_requests	0.00000
no_of_children	0.00000
avg_price_per_room	0.00000
no_of_previous_bookings_not_canceled	0.00000
no_of_previous_cancellations	0.00000
repeated_guest	0.00000
arrival_date	0.00000
arrival_month	0.00000
arrival_year	0.00000
lead_time	0.00000
required_car_parking_space	0.00000
no_of_week_nights	0.00000
no_of_weekend_nights	0.00000
market_segment_type_Online	0.00000

Table 23: Feature importance for post tuned decision tree

The post_pruned decision tree does not give importance to any of the features, and thus becomes a failure

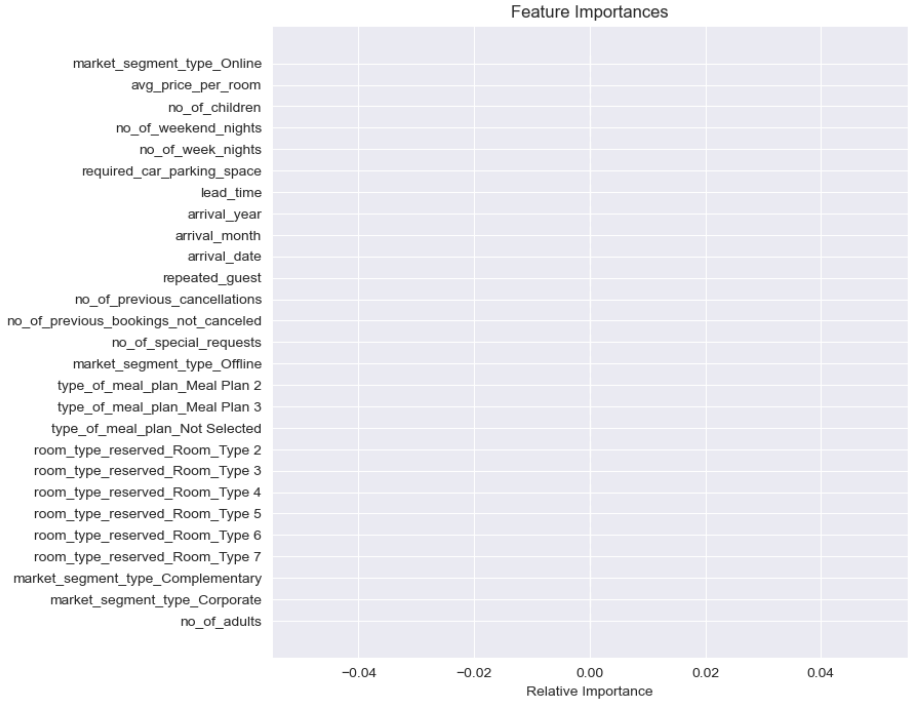


Figure 72: Feature importance for post tuned decision tree

5. MODEL PERFORMANCE COMPARISON AND FINAL MODEL SELECTION

5.1 Training performance comparison

Training performance comparison:

	Logistic Regression Base	Logistic Regression Improved	Decision Tree Base	Decision Tree Pre-Pruned	Decision Tree Post-Pruned
Accuracy	0.80502	0.79214	0.99429	0.87571	0.67399
Recall	0.89079	0.81647	0.99562	0.92906	1.00000
Precision	0.83183	0.86735	0.99591	0.89116	0.67399
F1	0.86030	0.84114	0.99576	0.90972	0.80525

Table 24: Training performance comparison

	Logistic Regression Base	Logistic Regression Tuned	Decision Tree Base	Decision Tree Pre-Pruned	Decision Tree Post-Pruned
Accuracy	0.80722	0.79261	0.86474	0.87384	0.66857
Recall	0.89788	0.82614	0.89335	0.93073	1.00000
Precision	0.82822	0.85835	0.90328	0.88627	0.66857
F1	0.86165	0.84194	0.89829	0.90796	0.80137

Table 25: Testing performance comparison

Observations

- Decision tree model with default parameters is overfitting the training data and is not able to generalize well.
- Pre-pruned tree has given a generalized performance with the recall score of 0.929 and 0.930 on training and test set, respectively.
- The company can predict the interested leads better using the pre-pruned tree.

6. ACTIONABLE INSIGHTS AND RECOMMENDATIONS

Conclusions

- The cancellation decisions of 36,275 bookings through Logistic Regression model and Decision tree classifier was analysed to create a predictive model for the same
- These models can be used by INN Hotels Group to predict the booking_status prior to the date of arrival

Logistic Regression model

- The Logistic Regression model works well on train and test data sets with the default threshold value of 0.5, thus strikes a good balance between the precision and the recall giving a F1_score of 0.86
- From the Logistic Regression model, the important predictor variables for predicting the booking cancellations are the Lead_time, avg_price_per_room and market_segment_type_online. That is one unit increase in the lead_time increases 1.3237 units of cancellation, one unit increase in avg_price_per_room increases 0.6681 units of booking cancellations, one unit increase in market_segment_type_online increases 0.3699 units of booking cancellations.
- Another important predictor variable is the no_of_special_requests which decreases the booking cancellations. One unit increase in no_special_request decreases 1.1644 units of booking cancellations

Decision Tree model

- The Decision Tree model seems to be overfitting with its base model and required pruning
- The pre_pruned decision tree better than the post_pruned across multiple evaluation metrics, such as accuracy, precision, and F1 score.
- The recall value of the post_pruned tree is good compared to pre_pruned tree. But the post_pruned tree does not give importance to any features and thus remains a failure.
- So it is better for the INN Groups to predict the booking cancellations with help of the pre_pruned decision tree which gives a recall value of 0.93, F1_score of 0.90, precision of 0.88 and accuracy of 0.87.
- The important predictor variables suggested by the pre_pruned model to predict the booking status in prior are the Lead_time, avg_price_per_room market_segment_type_online and the no_special requests

ACTIONABLE INSIGHTS

- Through the EDA and predictions from both the models it is understood that guests booking cheaper rooms, with shorter lead times, requiring a parking space, being a repeat guest, with higher number of special requests, from the Corporate and Offline market segments are less likely to cancel bookings. Conversely, guests booking more expensive rooms, with longer lead times, through the Online market segment are more likely to cancel bookings.

Profitable policies for cancellations and refunds

- Upon considering the important predictor variables and features the INN Hotels Group can frame a separate cancellation and refund policies for its frequent guests
- Also various incentives can be provided for bookings done through Corporate/ Offline market segments
- Even in case of overbooking the Hotel management to ensure providing best customer service to the repeated guest and this can be done by hiring additional staff.
- The hotel can offer a refund policy where the amount decreases as the check-in date approaches. This may encourage guests to cancel well in advance, giving the hotel an opportunity to rebook the room.
- Also, it could include an option to modify the booking that allows the customer to make changes up to a certain day before check-in, this may prevent cancellations when they just want changes in the reservation.

RECOMMENDATIONS

- The hotel management had to consider the Lead time, Online market segment and average price per room as they are the most important variables in determining if a booking will be cancelled.
- The hotel management can provide offers for online booking to attract more customers.
- From the EDA, we see that bookings are high during a particular season. The number of bookings starts to increase from August, reaching its peak in October with 5317 bookings and decreases during the winter
- The marketing team can compaign to increase bookings during the winter season. They can achieve this by offering appealing deals that attract more customers, resulting in higher occupancy rates.
- According to the analysis repeated guests are less likely to cancel a booking, so offering rewards to repeated guests can encourage them to choose INN hotel for future stays and reduce the likelihood of cancellations.
- INN Hotels should keep getting data and making further analysis about the reasons customers have to cancel and also .