INN Hotels Project

Supervised Learning - Classification

August 9, 2022

Sunny Amirize, MBA

Contents / Agenda

Executive Summary

To analyze the data provided and build a linear regression model to predict help in predicting which booking is likely to be canceled., the follow steps were followed;

- Sanity checks on the dataset.
- Exploratory data analysis was done.
- Data preprocessing was done.
- Model building and performance checks.
- Model assumptions check.
- Final model.
- Decision Tree.

After building the final model to get the OLS Regression Results, the insights and recommendation are as follows;
For the model, we have;

- The lead time was identified as the most important feature; a longer lead time increases the odds of cancellations. Policies need to be introduced to restrict how far in advance bookings can be made before the check-in date.
- Hotel policies must restrict the length of stay as bookings for more extended stay periods also increase the odds of cancellations.
- The repeat guests are identified to have lower odds of cancellations. Hotel policies need to incentivize current & previous guests to increase conversion as repeated guests.
- More bookings and cancellations were found to occur over months (March-August) compared to (September-February)
- Observing market segments, the avg price per room has been higher in instances where bookings have been canceled than in cases in which bookings have not been canceled. More competition information is required to ensure that our pricing is competitive to retain guests.

Business Problem Overview and Solution Approach

Great Learning
POWER AHEAD

• Problem Statement

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

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

The cancellation of bookings impact a hotel on various fronts:

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

• Solution approach/methodology

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

EDA Results

- **Key results from EDA**
- The lead time was identified as the most important feature; a longer lead time increases the odds of cancellations.
- The repeat guests are identified to have lower odds of cancellations.
- More bookings and cancellations were found to occur over months (March-August) compared to (September-February)

- **Please mention answers to the insight-based questions provided**
- The lead time was identified as the most important feature; a longer lead time increases the odds of cancellations. Policies need to be introduced to restrict how far in advance bookings can be made before the check-in date.
- Hotel policies must restrict the length of stay as bookings for more extended stay periods also increase the odds of cancellations.
- The repeat guests are identified to have lower odds of cancellations. Hotel policies need to incentivize current & previous guests to increase conversion as repeated guests.
- More bookings and cancellations were found to occur over months (March-August) compared to (September-February)
- Observing market segments, the avg price per room has been higher in instances where bookings have been canceled than in cases in which bookings have not been canceled. More competition information is required to ensure that our pricing is competitive to retain guests.

Link to Appendix slide on data background check

EDA Results

View the top 5 rows of the dataset

	0	1	2	3	4
Booking_ID	INN00001	INN00002	INN00003	INN00004	INN00005
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
type_of_meal_plan	Meal Plan 1	Not Selected	Meal Plan 1	Meal Plan 1	Not Selected
required_car_parking_space	0	0	0	0	0
room_type_reserved	Room_Type 1	Room_Type 1	Room_Type 1	Room_Type 1	Room_Type 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
market_segment_type	Offline	Online	Online	Online	Online
repeated_guest	0	0	0	0	0
no_of_previous_cancellations	0	0	0	0	0
no_of_previous_bookings_not_canceled	0	0	0	0	0
avg_price_per_room	65.0	106.68	60.0	100.0	94.5
no_of_special_requests	0	1	0	0	0
booking_status	Not_Canceled	Not_Canceled	Canceled	Canceled	Canceled

View the last 5 rows of the dataset

	36270	36271	36272	36273	36274
Booking_ID	INN36271	INN36272	INN36273	INN36274	INN36275
no_of_adults	3	2	2	2	2
no_of_children	0	0	0	0	0
no_of_weekend_nights	2	1	2	0	1
no_of_week_nights	6	3	6	3	2
type_of_meal_plan	Meal Plan 1	Meal Plan 1	Meal Plan 1	Not Selected	Meal Plan 1
required_car_parking_space	0	0	0	0	0
room_type_reserved	Room_Type 4	Room_Type 1	Room_Type 1	Room_Type 1	Room_Type 1
lead_time	85	228	148	63	207
arrival_year	2018	2018	2018	2018	2018
arrival_month	8	10	7	4	12
arrival_date	3	17	1	21	30
market_segment_type	Online	Online	Online	Online	Offline
repeated_guest	0	0	0	0	0
no_of_previous_cancellations	0	0	0	0	0
no_of_previous_bookings_not_canceled	0	0	0	0	0
avg_price_per_room	167.8	90.95	98.39	94.5	161.67
no_of_special_requests	1	2	2	0	0
booking_status	Not_Canceled	Canceled	Not_Canceled	Canceled	Not_Canceled

Link to Appendix slide on data background check

EDA Results

Checking the shape/dimension of the dataset.

The dataset has 36275 rows and 19 columns.

```
(36275, 19)
```

[Link to Appendix slide on data background check]

EDA Results

Checking the data types of the columns for the dataset.

There are 5 columns of the dtype object,
1 column of the dtype float64,
and 13 columns of the dtype int64.

```
<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
```

[Link to Appendix slide on data background check]

EDA Results

Checking for duplicate values.

```
False    36275
dtype: int64
```

[Link to Appendix slide on data background check](#)

EDA Results

Dropping the Booking_ID column from the dataframe

	0	1	2	3	4
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
type_of_meal_plan	Meal Plan 1	Not Selected	Meal Plan 1	Meal Plan 1	Not Selected
required_car_parking_space	0	0	0	0	0
room_type_reserved	Room_Type 1	Room_Type 1	Room_Type 1	Room_Type 1	Room_Type 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
market_segment_type	Offline	Online	Online	Online	Online
repeated_guest	0	0	0	0	0
no_of_previous_cancellations	0	0	0	0	0
no_of_previous_bookings_not_canceled	0	0	0	0	0
avg_price_per_room	65.0	106.68	60.0	100.0	94.5
no_of_special_requests	0	1	0	0	0
booking_status	Not_Canceled	Not_Canceled	Canceled	Canceled	Canceled

Link to Appendix slide on data background check

EDA Results

Statistical summary of the dataset.

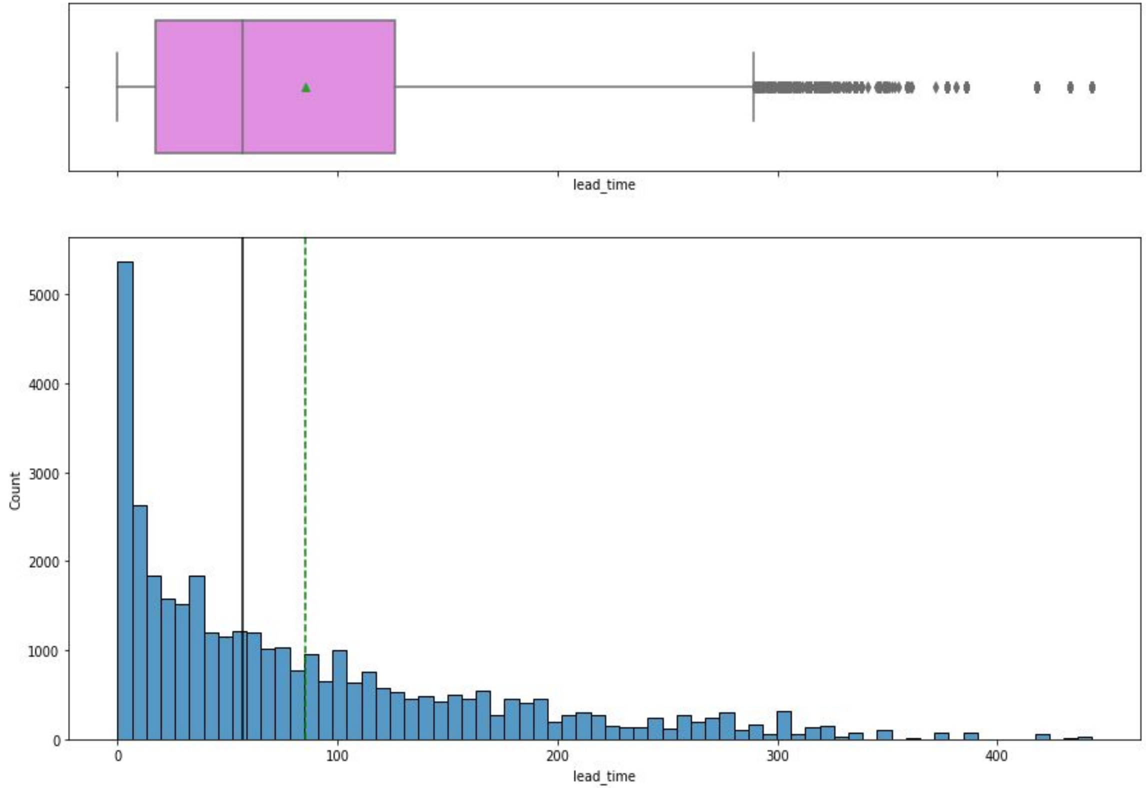
	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	required_car_parking_space	lead_time	arrival_year	arrival_month	arrival_date	repeated_guest	no_of_previous_cancellations	no_of_previous_bookings_no
count	36275.000000	36275.000000	36275.000000	36275.000000	36275.000000	36275.000000	36275.000000	36275.000000	36275.000000	36275.000000	36275.000000	36
mean	1.844962	0.105279	0.810724	2.204300	0.030986	85.232557	2017.820427	7.423653	15.596995	0.025637	0.023349	
std	0.518715	0.402648	0.870644	1.410905	0.173281	85.930817	0.383836	3.069894	8.740447	0.158053	0.368331	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2017.000000	1.000000	1.000000	0.000000	0.000000	
25%	2.000000	0.000000	0.000000	1.000000	0.000000	17.000000	2018.000000	5.000000	8.000000	0.000000	0.000000	
50%	2.000000	0.000000	1.000000	2.000000	0.000000	57.000000	2018.000000	8.000000	16.000000	0.000000	0.000000	
75%	2.000000	0.000000	2.000000	3.000000	0.000000	126.000000	2018.000000	10.000000	23.000000	0.000000	0.000000	
max	4.000000	10.000000	7.000000	17.000000	1.000000	443.000000	2018.000000	12.000000	31.000000	1.000000	13.000000	

Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on lead time

Visual analysis of both distributions shows
- right-skewed.
- the mean is around 90 days (lead time)
- outliers to the right indicate rooms to
the right have high lead times.

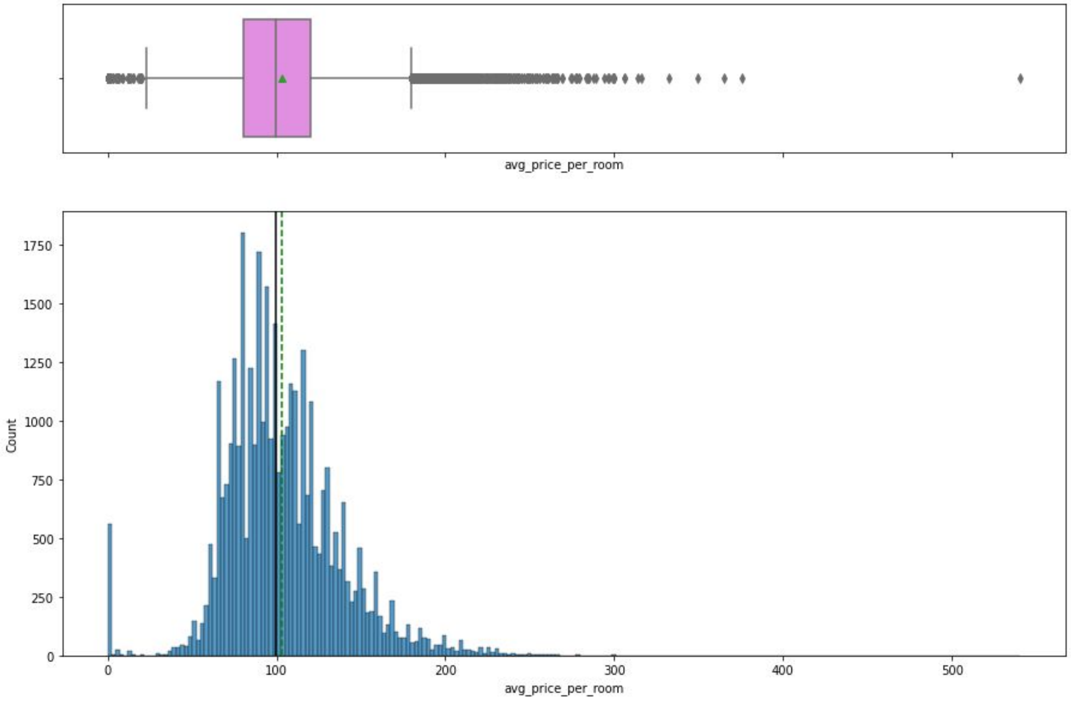


[Link to Appendix slide on data background check]

EDA Results
Univariate Analysis

Observations on average price per room

Visual analysis of both distributions shows
- right-skewed.
- the mean is around 101 euros.
- outliers to the right indicate more expensive rooms.



[*Link to Appendix slide on data background check*]

EDA Results

Average price per room

	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
63	1	0	0	1	Meal Plan 1	0	Room_Type 1	2	2017	9	10	Complementary	0
145	1	0	0	2	Meal Plan 1	0	Room_Type 1	13	2018	6	1	Complementary	1
209	1	0	0	0	Meal Plan 1	0	Room_Type 1	4	2018	2	27	Complementary	0
266	1	0	0	2	Meal Plan 1	0	Room_Type 1	1	2017	8	12	Complementary	1
267	1	0	2	1	Meal Plan 1	0	Room_Type 1	4	2017	8	23	Complementary	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
35983	1	0	0	1	Meal Plan 1	0	Room_Type 7	0	2018	6	7	Complementary	1
36080	1	0	1	1	Meal Plan 1	0	Room_Type 7	0	2018	3	21	Complementary	1
36114	1	0	0	1	Meal Plan 1	0	Room_Type 1	1	2018	3	2	Online	0
36217	2	0	2	1	Meal Plan 1	0	Room_Type 2	3	2017	8	9	Online	0
36250	1	0	0	2	Meal Plan 2	0	Room_Type 1	6	2017	12	10	Online	0

Link to Appendix slide on data background check

EDA Results

The value of the upper whisker

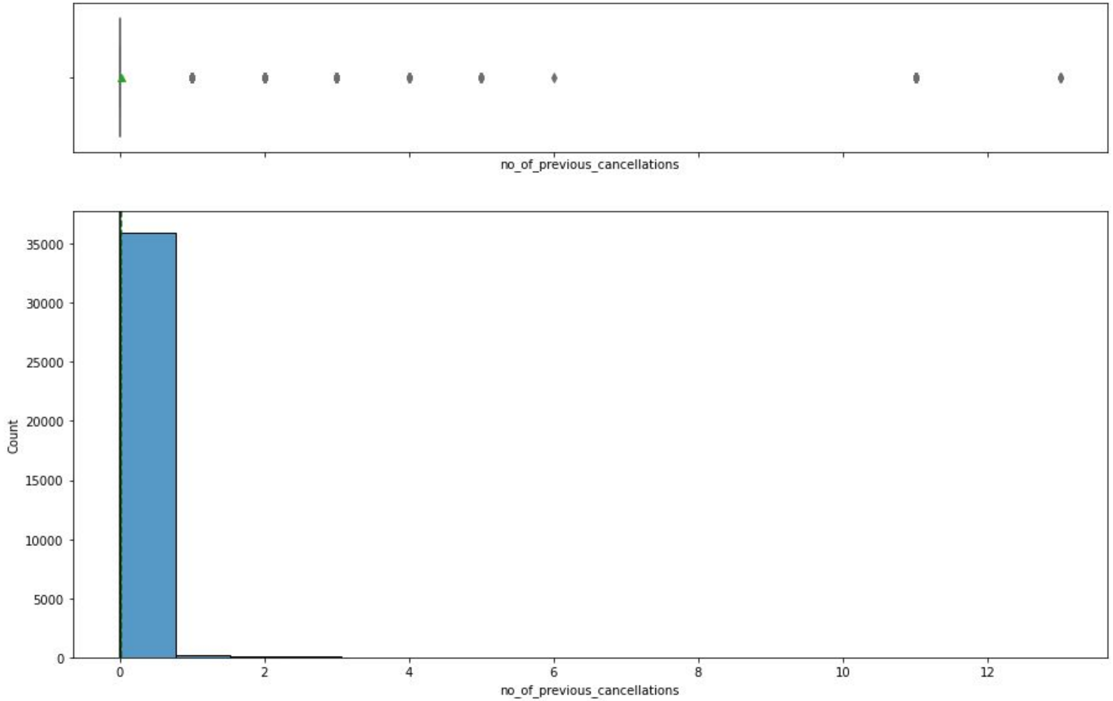
179.55

[Link to Appendix slide on data background check]

EDA Results
Univariate Analysis

Observations on number of previous cancellations

Visual analysis of both distributions shows
- right-skewed.
- the mean is 0
- few outliers to the right

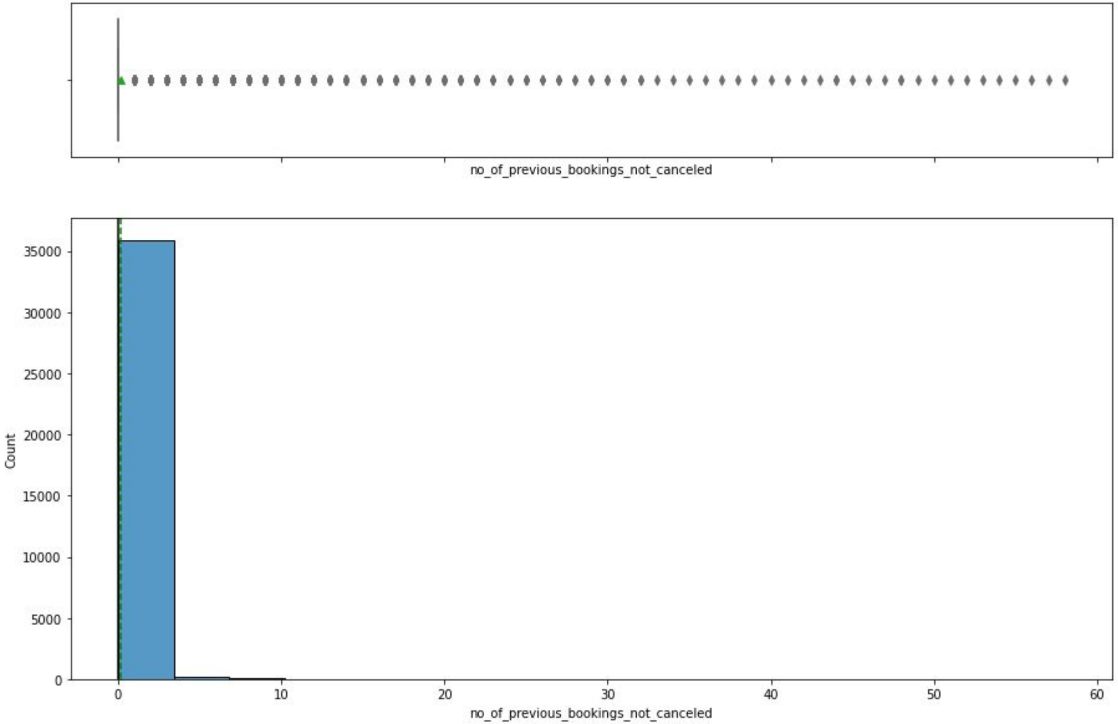


[*Link to Appendix slide on data background check*]

EDA Results
Univariate Analysis

Observations on number of previous not canceled

Visual analysis of both distributions shows
- right-skewed.
- the mean is 0
- consistence outliers to the right

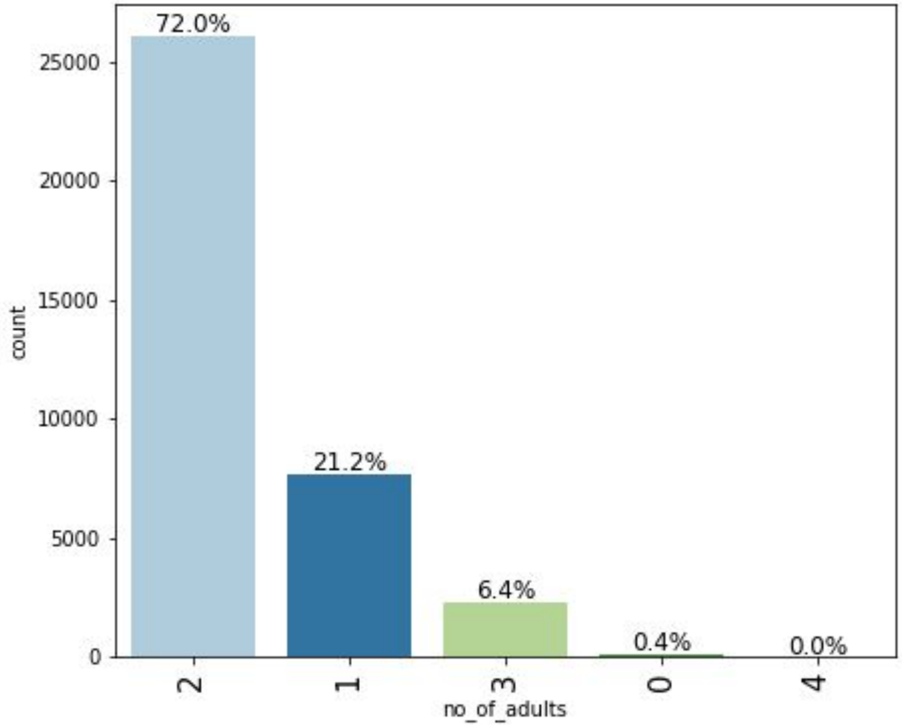


[Link to Appendix slide on data background check]

EDA Results
Univariate Analysis

Observations on number of adults

Visual analysis of the bar graph shows
- 72% of bookings included two adults, which indicates that two adults booked for hotel room.

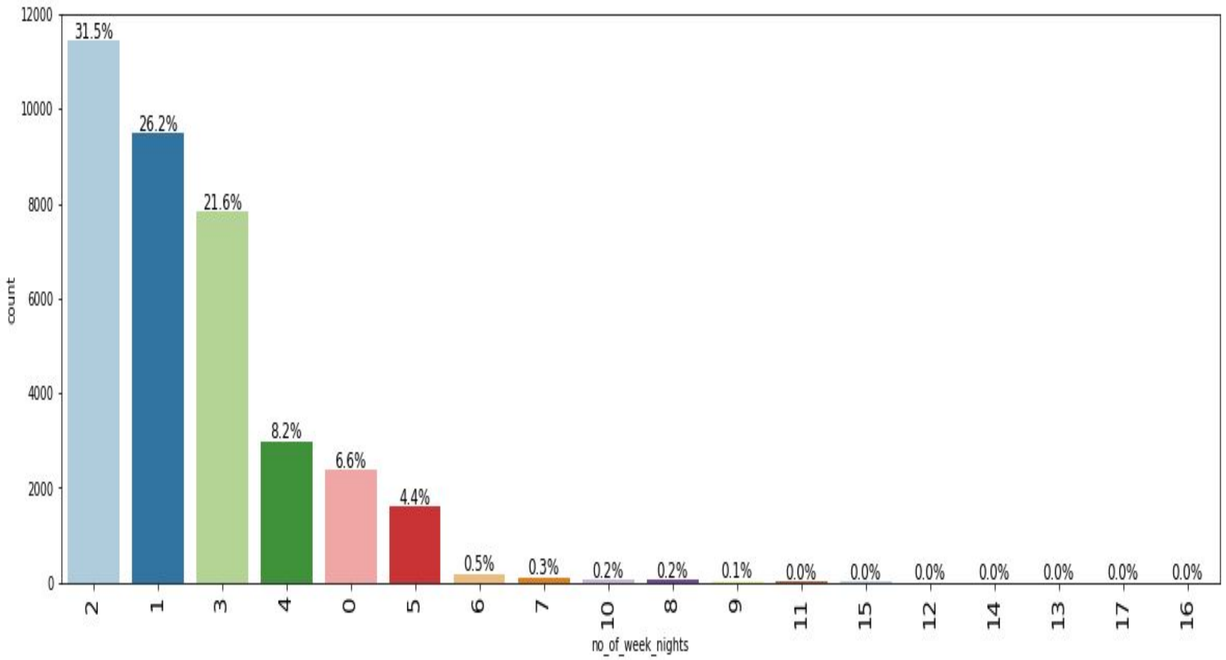


[*Link to Appendix slide on data background check*]

EDA Results
Univariate Analysis

Observations on number of week nights

Visual analysis of the bar graph shows
- 31.5% of bookings which included two adults, occurred week nights (Monday to Friday).

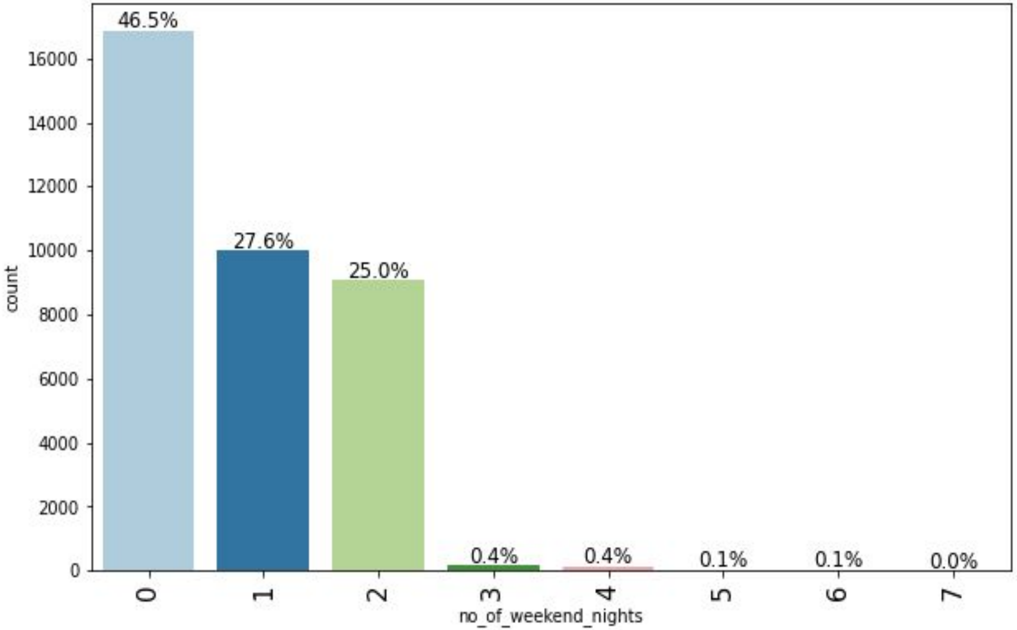


[Link to Appendix slide on data background check]()

EDA Results
Univariate Analysis

Observations on number of week nights

Visual analysis of the bar graph shows
- 46.5% of bookings included no weekend nights.

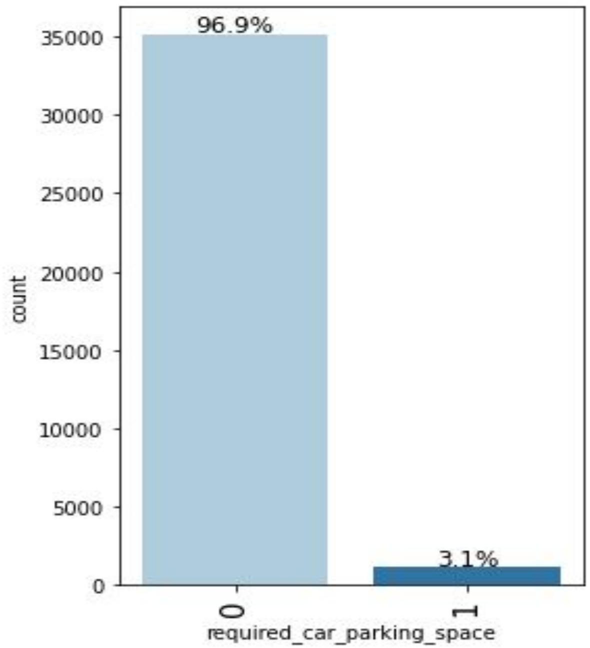


Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on required car packing space

Visual analysis of the bar graph shows
- 96.9% guest who booked hotel rooms did not require car packing space and 3.1% require 1 packing space.

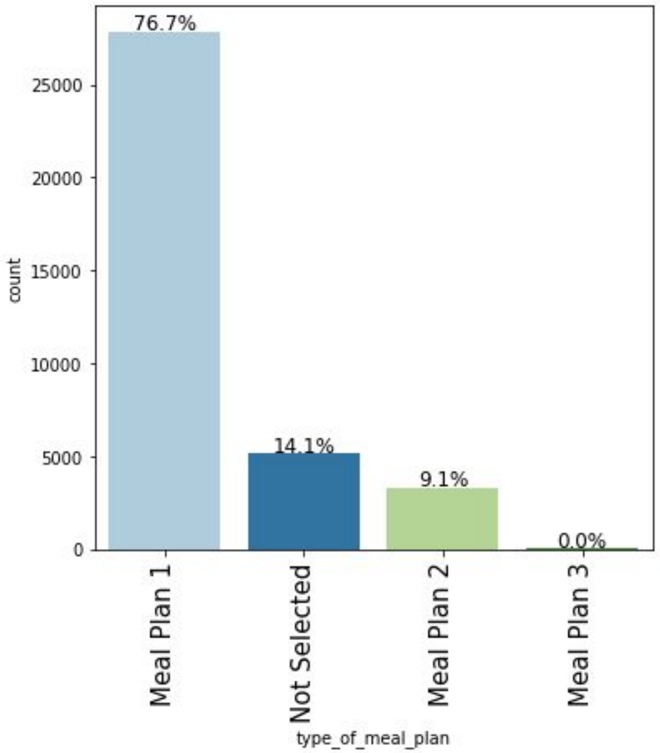


Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on type of meal plan

Visual analysis of the bar graph shows
- 76.7% guest who booked hotel room required Meal Plan 1.

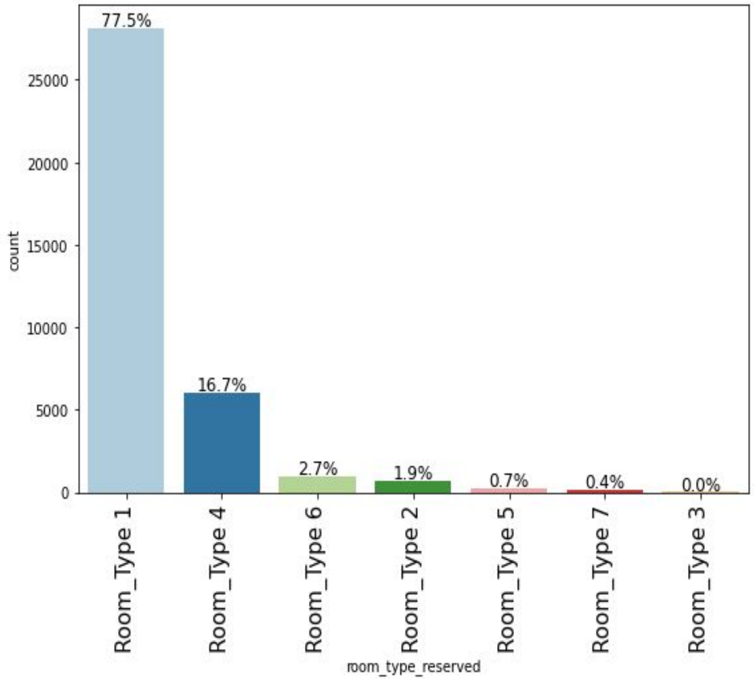


[Link to Appendix slide on data background check]()

EDA Results
Univariate Analysis

Observations on room type reserved

Visual analysis of the bar graph shows
- 77.5% guest who booked, made reservations for room type 1.

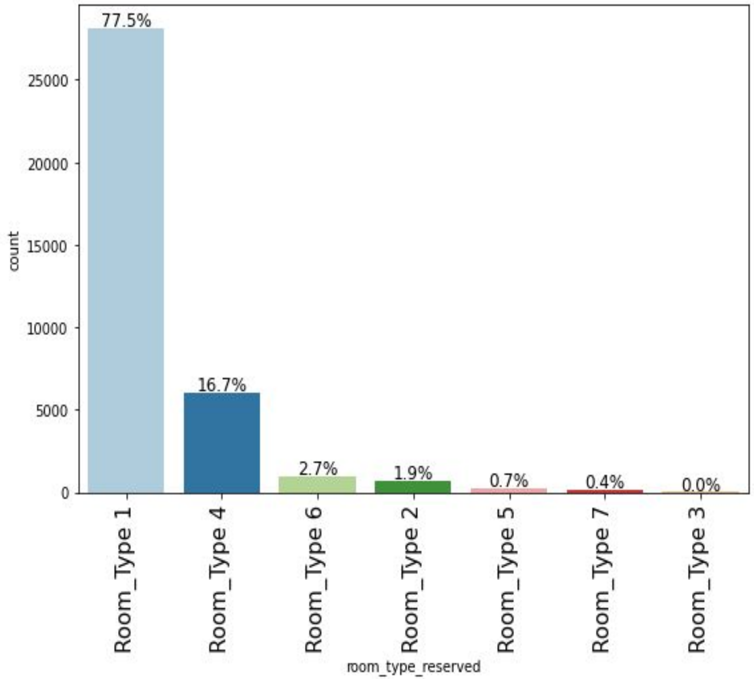


[*Link to Appendix slide on data background check*]

EDA Results
Univariate Analysis

Observations on room type reserved

Visual analysis of the bar graph shows
- 77.5% guest who booked, made reservations for room type 1.

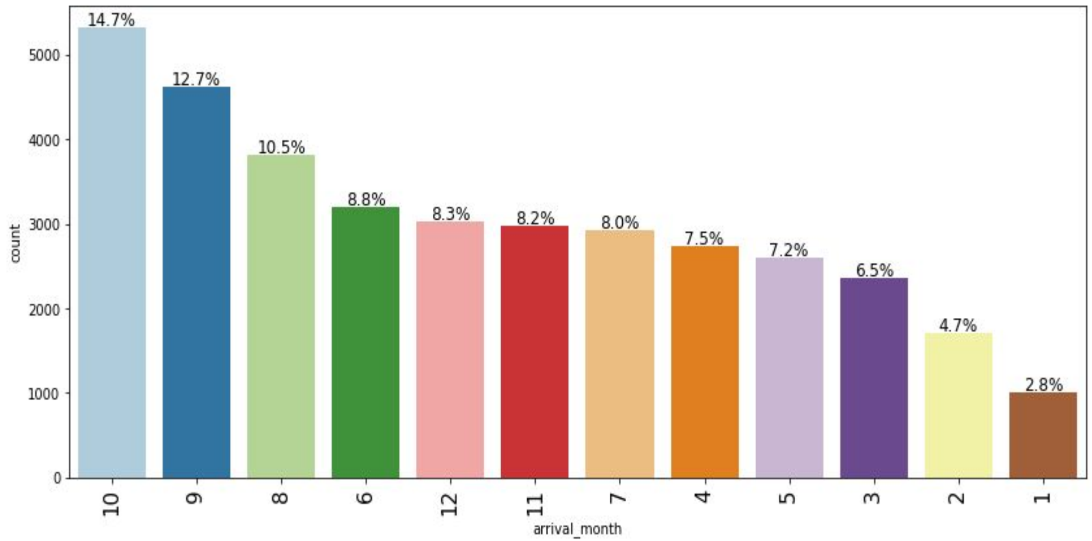


Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on arrival month

Visual analysis of the bar graph shows
- 14.7% of guest who booked arrived the hotel in the 10th month, indicating they arrived in the month of August.

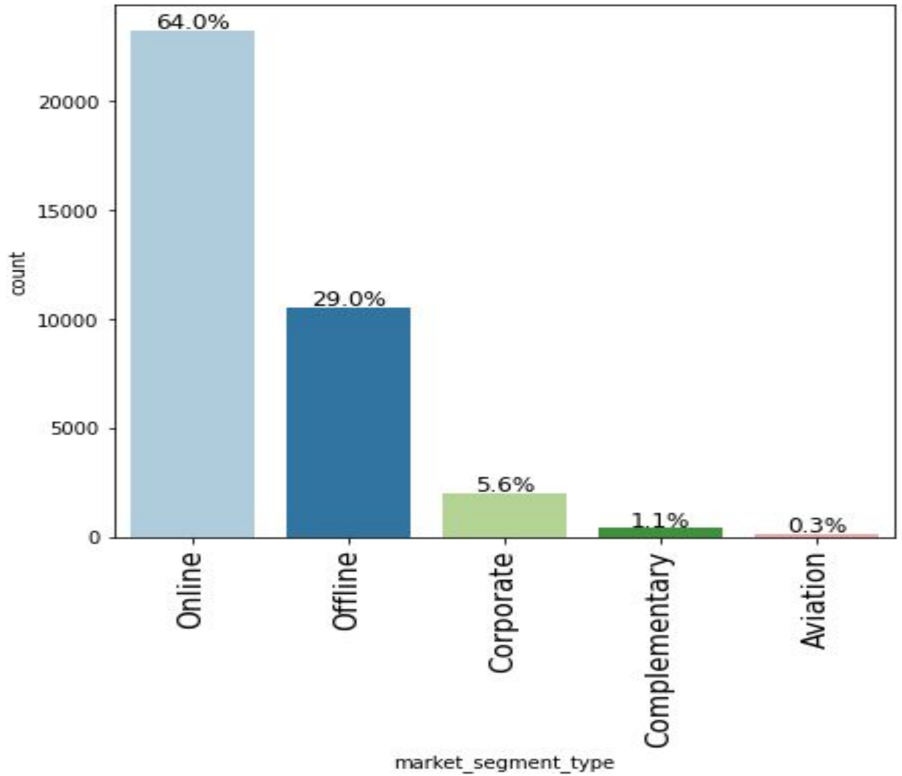


[*Link to Appendix slide on data background check*](#)

EDA Results
Univariate Analysis

Observations on market segment type

Visual analysis of the bar graph shows
- 64% of the bookings occurred online.

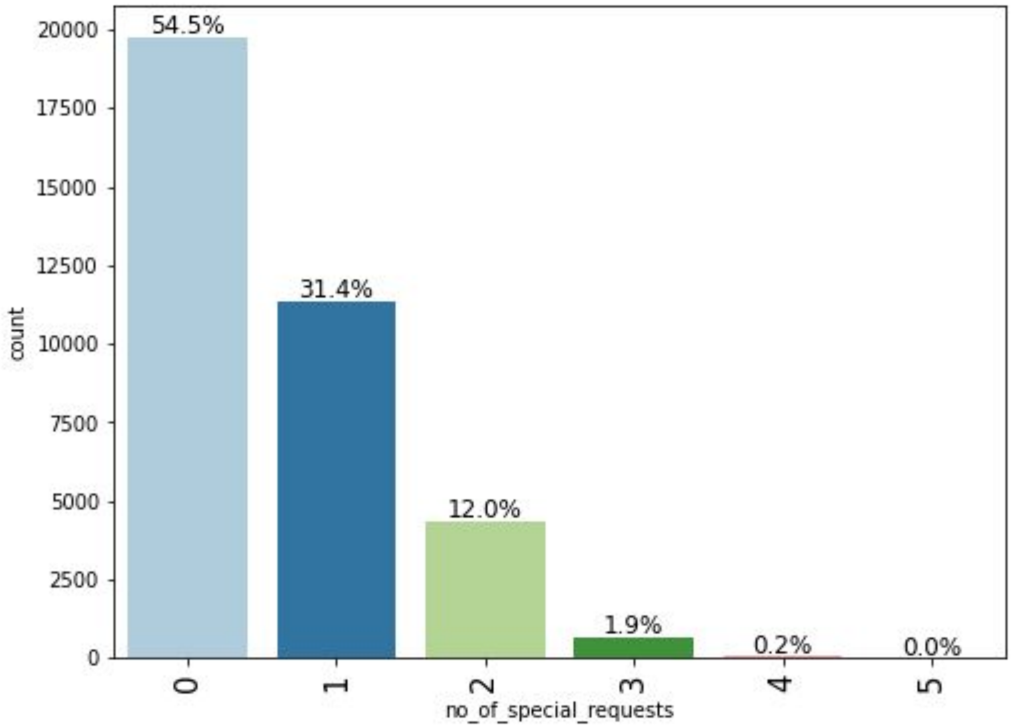


Link to Appendix slide on data background check

EDA Results
Univariate Analysis

Observations on number of special requests

Visual analysis of the bar graph shows
- 54.5% of the bookings has no special guest.

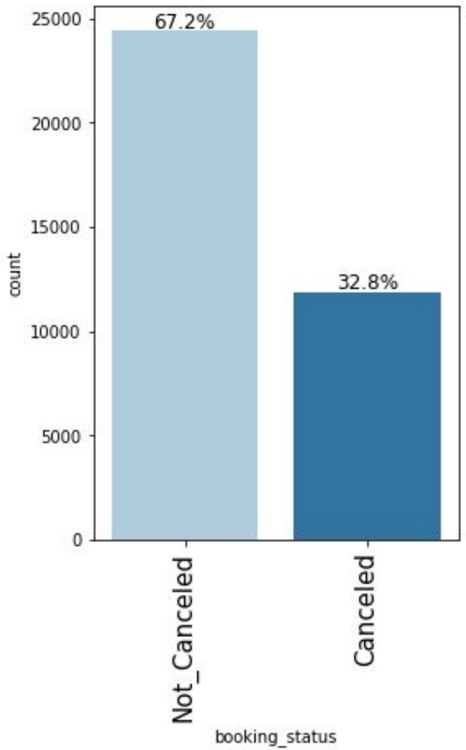


[*Link to Appendix slide on data background check*]

EDA Results
Univariate Analysis

Observations on booking status

Visual analysis of the bar graph shows
- 67.2% of bookings were not cancelled and
32.8% were canceled.

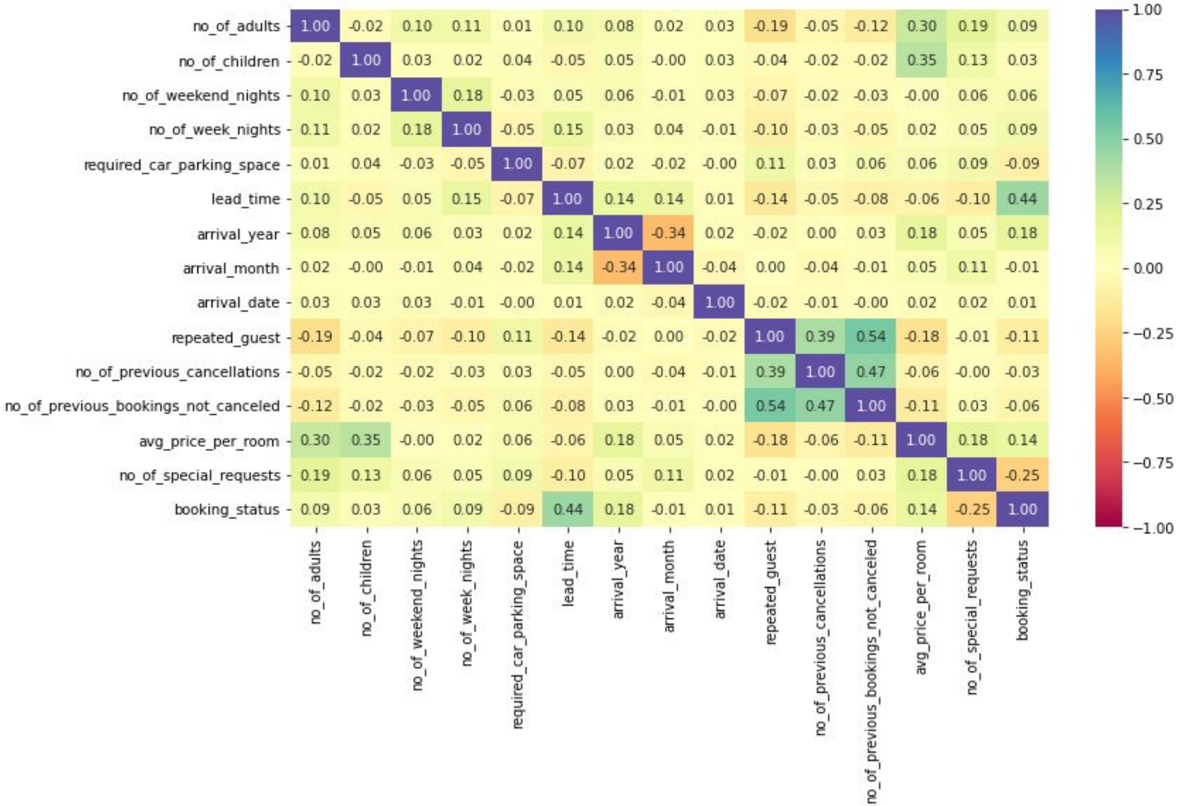


[Link to Appendix slide on data background check]

EDA Results
Bivariate Analysis

Observations on correlation heat map

Visual analysis of the heatmap shows
- Increase in number of adults and children, will increase the average price of a room. Increase in bookings will increase the lead time.
-There are lots of positive correlations.

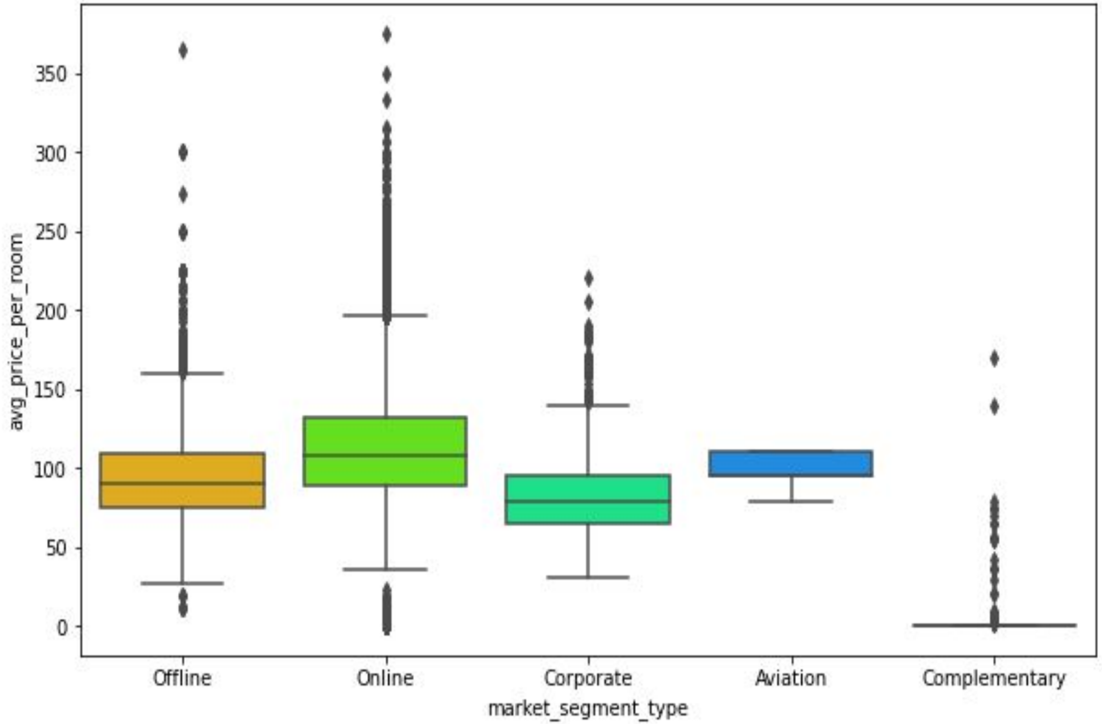


[Link to Appendix slide on data background check]()

EDA Results
Bivariate Analysis

Observations on how prices vary across different market

Visual analysis of the boxplot shows
- Online bookings have increased price compared to other market segment. Corporate bookings have the least average price per room.



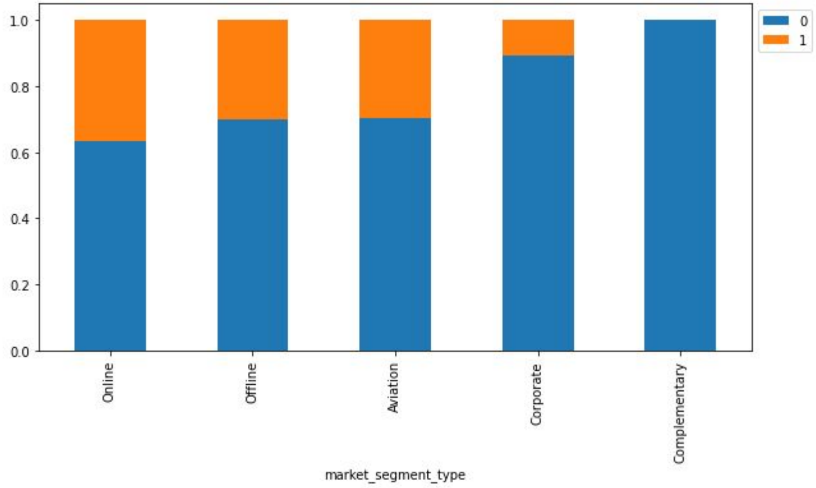
Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on how average price per room impacts booking status

Visual analysis of the stacked barplot shows
- The booking status for online booking has the highest increase compared to other market segment booking status.

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



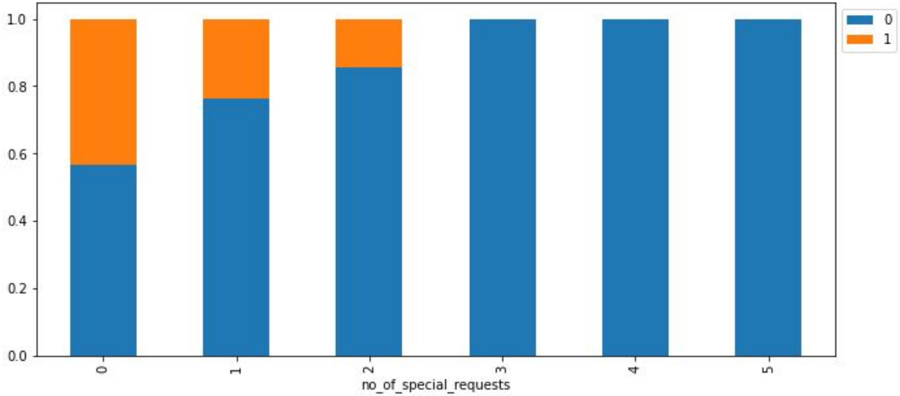
Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on how guest special requirements impacts cancellation

Visual analysis of the stacked barplot shows
- The booking status for guest with no or zero special requirements is the highest.

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

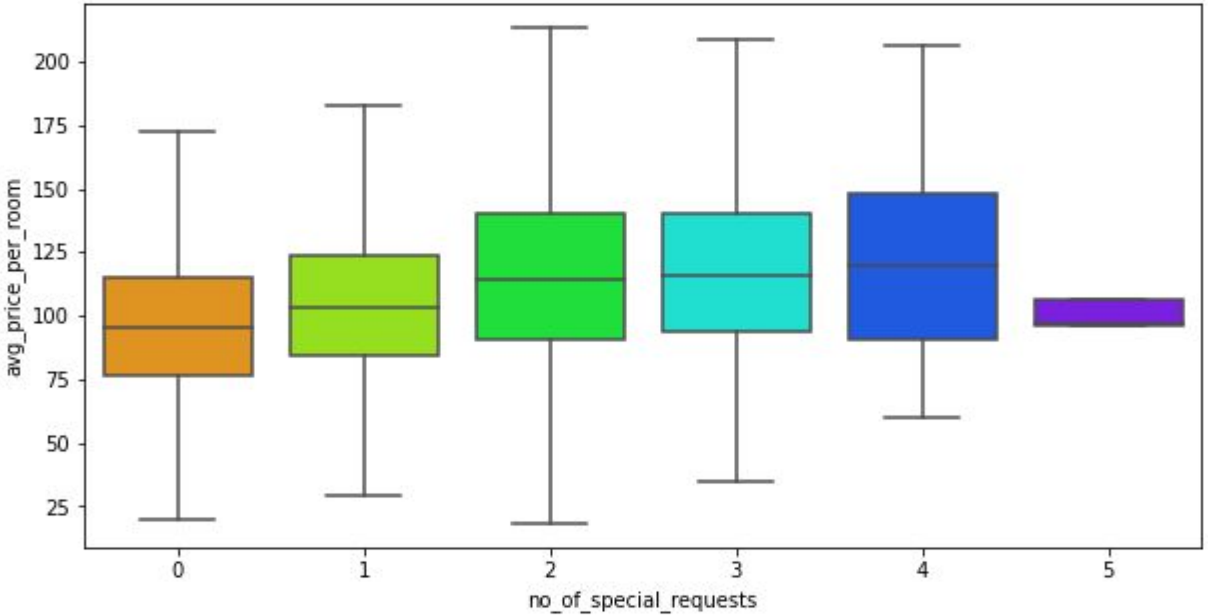


Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on how special requests made by the customers impact the prices of a room

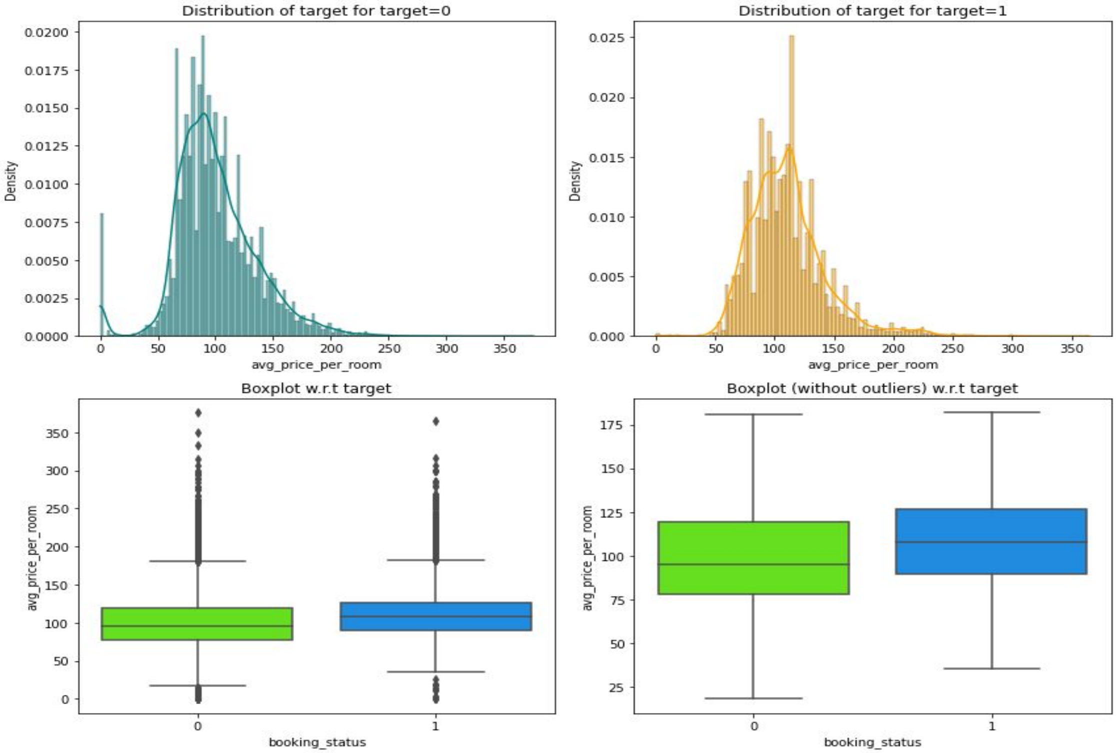
Visual analysis of the boxplot shows
- Bookings with 4 special requests saw increased average price per room.



[Link to Appendix slide on data background check]

EDA Results
Bivariate Analysis

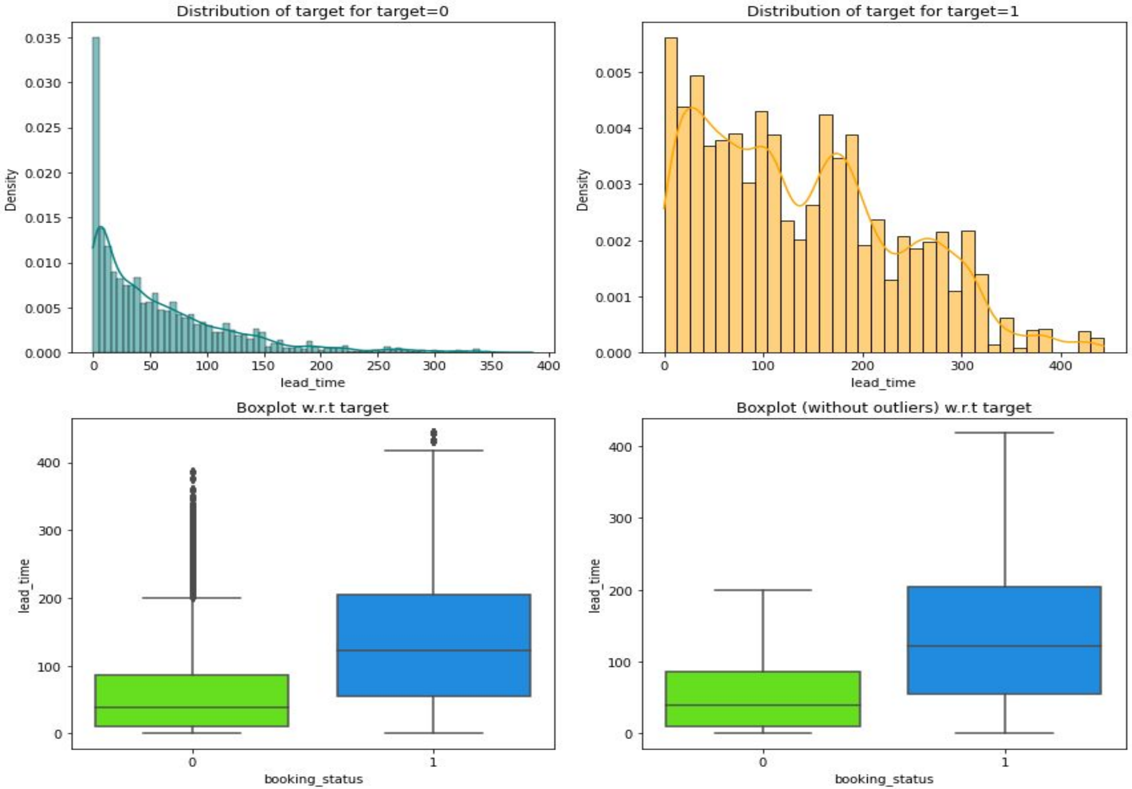
Analyzing the positive correlation between booking status and average price per room



[Link to Appendix slide on data background check]()

EDA Results
Bivariate Analysis

Analyzing the positive correlation between booking status and lead time



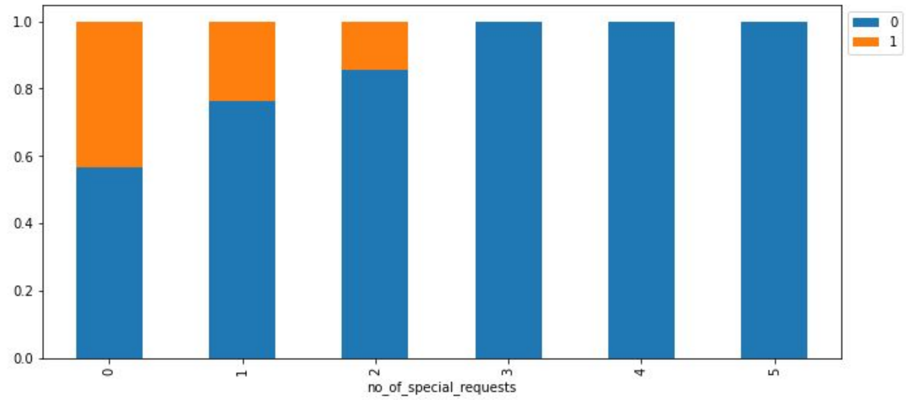
[Link to Appendix slide on data background check]

EDA Results
Bivariate Analysis

Observations on how guest special requirements impacts cancellation

Visual analysis of the stacked barplot shows
- The booking status for guest with no or zero special requirements is the highest.

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



[Link to Appendix slide on data background check](#)

EDA Results

Dataframe of the customers who traveled their families

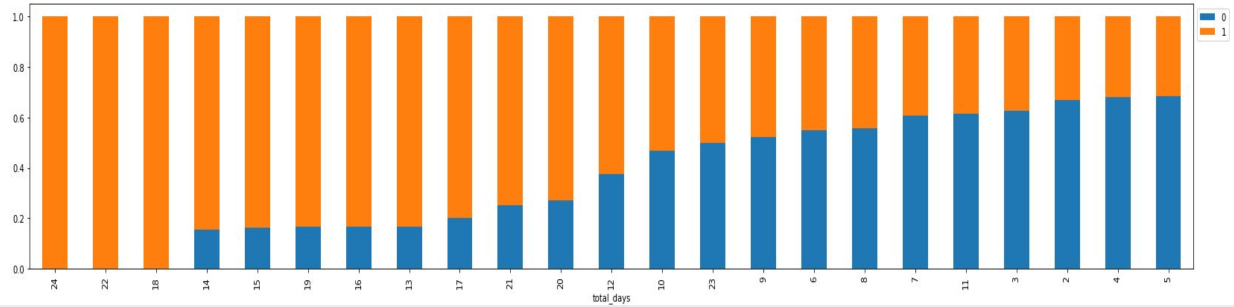
```
(28441, 18)
```

Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Totals days and booking status

booking_status	0	1	All
total_days			
All	10979	6115	17094
3	3689	2183	5872
4	2977	1387	4364
5	1593	738	2331
2	1301	639	1940
6	566	465	1031
7	590	383	973
8	100	79	179
10	51	58	109
9	58	53	111
14	5	27	32
15	5	26	31
13	3	15	18
12	9	15	24
11	24	15	39
20	3	8	11
19	1	5	6
16	1	5	6
17	1	4	5
18	0	3	3
21	1	3	4
22	0	2	2
23	1	1	2
24	0	1	1



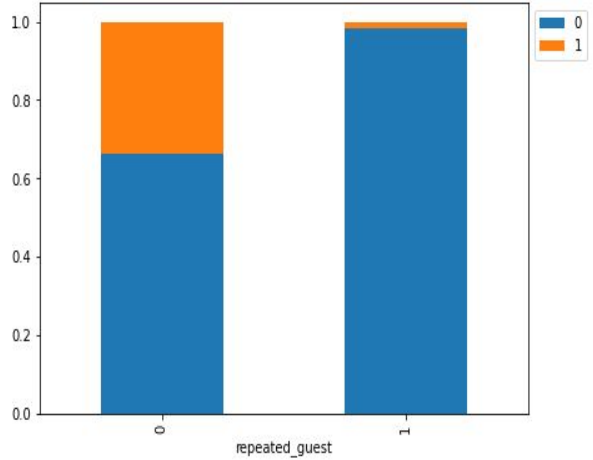
[Link to Appendix slide on data background check]

EDA Results
Bivariate Analysis

Observations on cancellation from repeated guests

Visual analysis of the stacked barplot shows - cancellation are more from

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

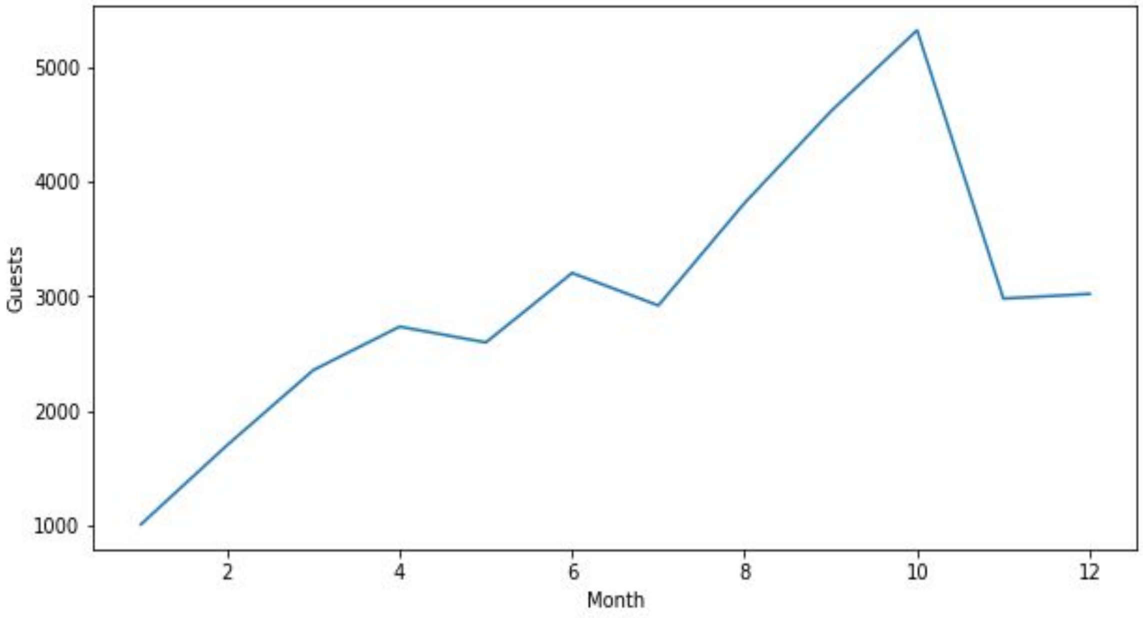


[Link to Appendix slide on data background check]

EDA Results
Bivariate Analysis

Observations on busiest months in the hotel

Visual analysis of the line plot shows
- The busiest months are between the 7th-10th month (July - August) with guest between 2500 - 5500



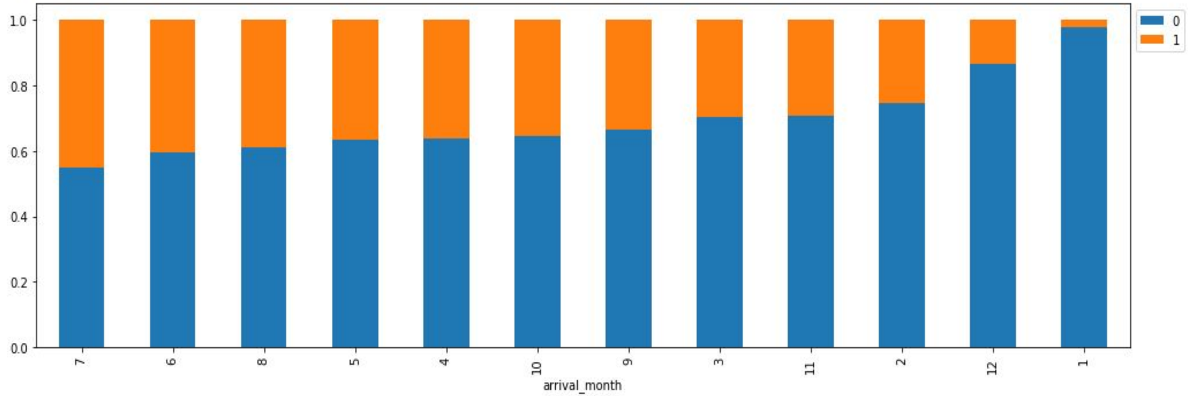
Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on the arrival month and booking status

Visual analysis of the stacked barplot shows
- guest arrival months are high for
October, September, August and July.
(10th, 9th, 8th and 7th month)

booking_status	0	1	All
arrival_month			
All	24390	11885	36275
10	3437	1880	5317
9	3073	1538	4611
8	2325	1488	3813
7	1606	1314	2920
6	1912	1291	3203
4	1741	995	2736
5	1650	948	2598
11	2105	875	2980
3	1658	700	2358
2	1274	430	1704
12	2619	402	3021
1	990	24	1014

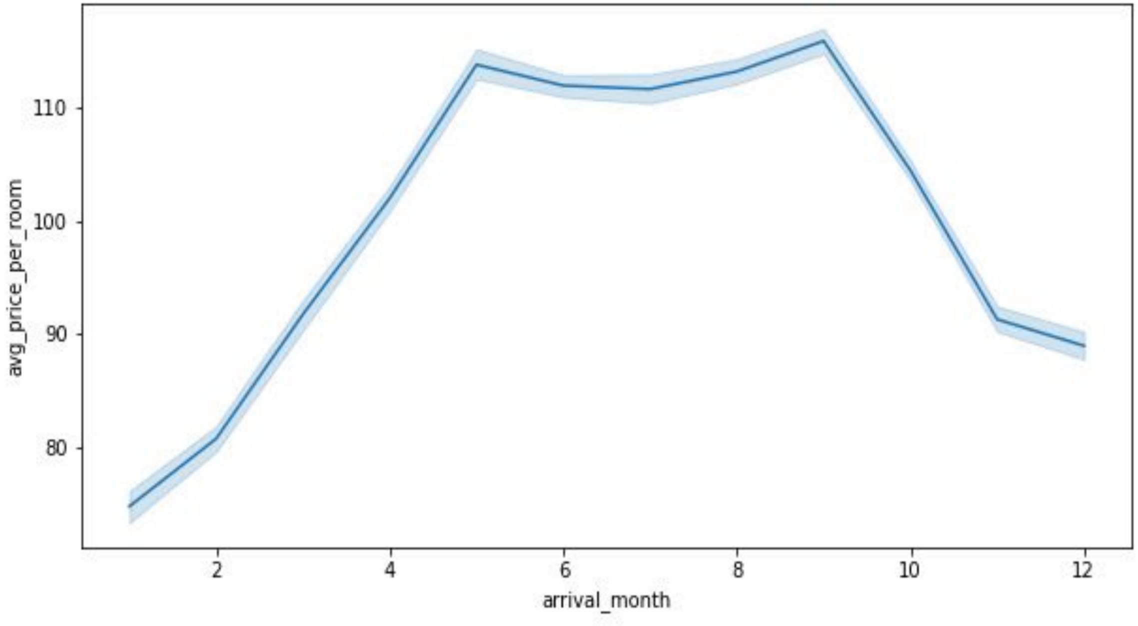


Link to Appendix slide on data background check

EDA Results
Bivariate Analysis

Observations on price varying across different months

Visual analysis of the line plot shows
- hotel rooms prices vary with the highest prices between the 5th - 9th arrival months.

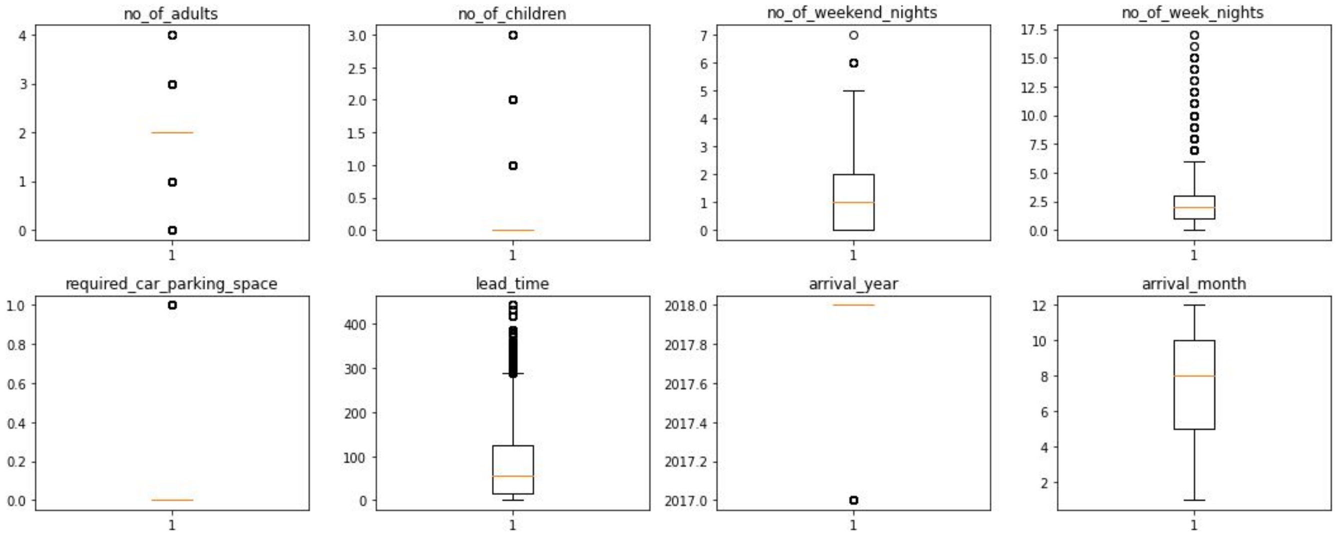


[*Link to Appendix slide on data background check*]

EDA Results
Bivariate Analysis

Check for outliers

Visual analysis of the boxplot shows
- few outliers found in no of adults, no of children, no of weekend nights, required car packing space, and arrival year. Majority of outliers found in no of week nights and lead time.

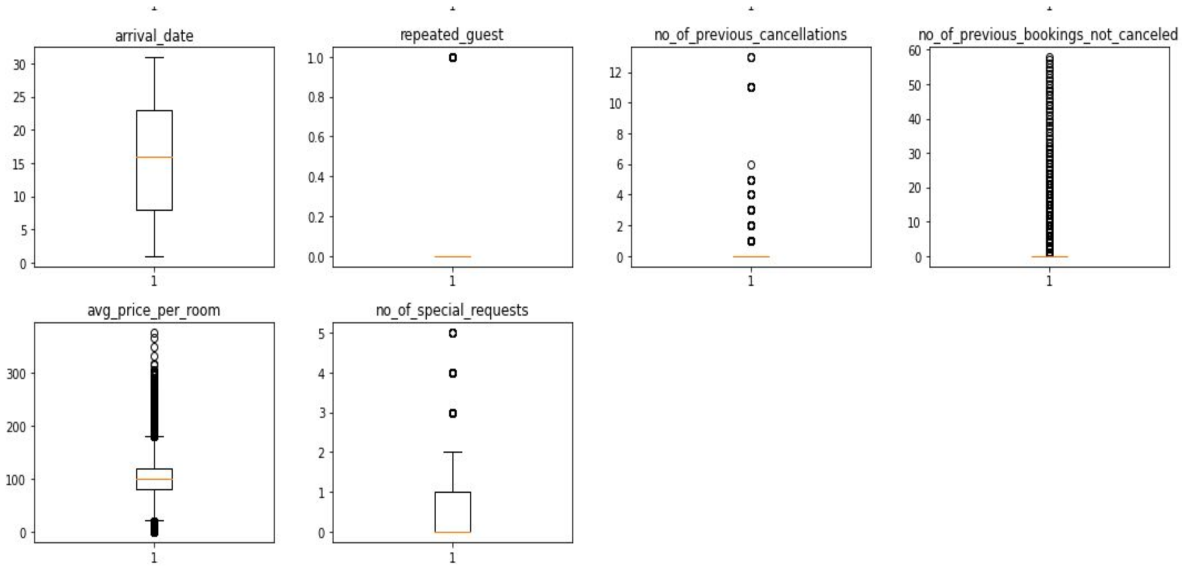


[Link to Appendix slide on data background check]()

EDA Results
Bivariate Analysis

Check for outliers

Visual analysis of the boxplot shows
- few outliers found in repeated guests, no of previous cancellations, no of special requests. Majority of outliers found no of previous bookings not canceled and average price per room.



[*Link to Appendix slide on data background check*]

Data Preprocessing

Data preparation for modeling

To predict which bookings will be canceled.

```
Shape of Training set :  (25392, 30)
Shape of test set :  (10883, 30)
Percentage of classes in training set:
0    0.670644
1    0.329356
Name: booking_status, dtype: float64
Percentage of classes in test set:
0    0.676376
1    0.323624
Name: booking_status, dtype: float64
```

Model Performance Summary

Logistic Regression

Visual analysis of the Logit Regression Results shows
- Negative values of coefficients indicates that booking is more likely to be cancelled with an increase in attribute value and positive values of coefficients.

p-value of a variable indicates if the variable is significant or not. If we consider the significance level to be 0.05 (5%), then any variable with a p-value less than 0.05 would be considered significant.

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25364
Method:	MLE	Df Model:	27
Date:	Sat, 06 Aug 2022	Pseudo R-squ.:	0.3292
Time:	19:24:33	Log-Likelihood:	-10794.
converged:	False	LL-Null:	-16091.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-922.8266	120.832	-7.637	0.000	-1159.653	-686.000
no_of_adults	0.1137	0.038	3.019	0.003	0.040	0.188
no_of_children	0.1580	0.062	2.544	0.011	0.036	0.280
no_of_weekend_nights	0.1067	0.020	5.395	0.000	0.068	0.145
no_of_week_nights	0.0397	0.012	3.235	0.001	0.016	0.064
required_car_parking_space	-1.5943	0.138	-11.565	0.000	-1.865	-1.324
lead_time	0.0157	0.000	58.863	0.000	0.015	0.016
arrival_year	0.4561	0.060	7.617	0.000	0.339	0.573
arrival_month	-0.0417	0.006	-6.441	0.000	-0.054	-0.029
arrival_date	0.0005	0.002	0.259	0.796	-0.003	0.004
repeated_guest	-2.3472	0.617	-3.806	0.000	-3.556	-1.139
no_of_previous_cancellations	0.2664	0.086	3.108	0.002	0.098	0.434
no_of_previous_bookings_not_canceled	-0.1727	0.153	-1.131	0.258	-0.472	0.127
avg_price_per_room	0.0188	0.001	25.396	0.000	0.017	0.020
no_of_special_requests	-1.4689	0.030	-48.782	0.000	-1.528	-1.410
type_of_meal_plan_Meal Plan 2	0.1756	0.067	2.636	0.008	0.045	0.306
type_of_meal_plan_Meal Plan 3	17.3584	3987.873	0.004	0.997	-7798.729	7833.446
type_of_meal_plan_Not Selected	0.2784	0.053	5.247	0.000	0.174	0.382
room_type_reserved_Room_Type 2	-0.3605	0.131	-2.748	0.006	-0.618	-0.103
room_type_reserved_Room_Type 3	-0.0012	1.310	-0.001	0.999	-2.568	2.566
room_type_reserved_Room_Type 4	-0.2823	0.053	-5.304	0.000	-0.387	-0.178
room_type_reserved_Room_Type 5	-0.7189	0.209	-3.438	0.001	-1.129	-0.309
room_type_reserved_Room_Type 6	-0.9501	0.151	-6.274	0.000	-1.247	-0.653
room_type_reserved_Room_Type 7	-1.4003	0.294	-4.770	0.000	-1.976	-0.825
market_segment_type_Complementary	-40.5976	5.65e+05	-7.19e-05	1.000	-1.11e+06	1.11e+06
market_segment_type_Corporate	-1.1924	0.266	-4.483	0.000	-1.714	-0.671
market_segment_type_Offline	-2.1946	0.255	-8.621	0.000	-2.694	-1.696
market_segment_type_Online	-0.3995	0.251	-1.590	0.112	-0.892	0.093

Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

Test performance

Training performance:

	Accuracy	Recall	Precision	F1
0	0.806002	0.634103	0.739713	0.682848

[*Link to Appendix slide on model assumptions*]

Model Performance Summary

Multicollinearity

Checking_vif (X_train)

	feature	VIF
0	const	3.949769e+07
1	no_of_adults	1.351135e+00
2	no_of_children	2.093583e+00
3	no_of_weekend_nights	1.069484e+00
4	no_of_week_nights	1.095711e+00
5	required_car_parking_space	1.039972e+00
6	lead_time	1.395175e+00
7	arrival_year	1.431904e+00
8	arrival_month	1.276334e+00
9	arrival_date	1.006795e+00
10	repeated_guest	1.783576e+00
11	no_of_previous_cancellations	1.395693e+00
12	no_of_previous_bookings_not_canceled	1.652000e+00
13	avg_price_per_room	2.068603e+00
14	no_of_special_requests	1.247981e+00
15	type_of_meal_plan_Meal Plan 2	1.273283e+00
16	type_of_meal_plan_Meal Plan 3	1.025258e+00
17	type_of_meal_plan_Not Selected	1.273060e+00
18	room_type_reserved_Room_Type 2	1.105954e+00
19	room_type_reserved_Room_Type 3	1.003303e+00
20	room_type_reserved_Room_Type 4	1.363606e+00
21	room_type_reserved_Room_Type 5	1.028000e+00
22	room_type_reserved_Room_Type 6	2.056136e+00
23	room_type_reserved_Room_Type 7	1.118156e+00
24	market_segment_type_Complementary	4.502756e+00
25	market_segment_type_Corporate	1.692829e+01
26	market_segment_type_Offline	6.411564e+01
27	market_segment_type_Online	7.118026e+01

Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

Visual analysis of the Logit Regression Results shows
- All p values are now <0.05. We will consider columns in X_train1 as final and lg1 as the final model

```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25370
Method:                           MLE   Df Model:                           21
Date:                Sun, 07 Aug 2022   Pseudo R-squ.:                  0.3282
Time:                        00:45:39   Log-Likelihood:                -10810.
converged:                       True   LL-Null:                       -16091.
Covariance Type:            nonrobust   LLR p-value:                     0.000
```

	coef	std err	z	P>|z|	[0.025	0.975]
const	-915.6391	120.471	-7.600	0.000	-1151.758	-679.520
no_of_adults	0.1088	0.037	2.914	0.004	0.036	0.182
no_of_children	0.1531	0.062	2.470	0.014	0.032	0.275
no_of_weekend_nights	0.1086	0.020	5.498	0.000	0.070	0.147
no_of_week_nights	0.0417	0.012	3.399	0.001	0.018	0.066
required_car_parking_space	-1.5947	0.138	-11.564	0.000	-1.865	-1.324
lead_time	0.0157	0.000	59.213	0.000	0.015	0.016
arrival_year	0.4523	0.060	7.576	0.000	0.335	0.569
arrival_month	-0.0425	0.006	-6.591	0.000	-0.055	-0.030
repeated_guest	-2.7367	0.557	-4.916	0.000	-3.828	-1.646
no_of_previous_cancellations	0.2288	0.077	2.983	0.003	0.078	0.379
avg_price_per_room	0.0192	0.001	26.336	0.000	0.018	0.021
no_of_special_requests	-1.4698	0.030	-48.884	0.000	-1.529	-1.411
type_of_meal_plan_Meal Plan 2	0.1642	0.067	2.469	0.014	0.034	0.295
type_of_meal_plan_Not Selected	0.2860	0.053	5.406	0.000	0.182	0.390
room_type_reserved_Room_Type 2	-0.3552	0.131	-2.709	0.007	-0.612	-0.098
room_type_reserved_Room_Type 4	-0.2828	0.053	-5.330	0.000	-0.387	-0.179
room_type_reserved_Room_Type 5	-0.7364	0.208	-3.535	0.000	-1.145	-0.328
room_type_reserved_Room_Type 6	-0.9682	0.151	-6.403	0.000	-1.265	-0.672
room_type_reserved_Room_Type 7	-1.4343	0.293	-4.892	0.000	-2.009	-0.860
market_segment_type_Corporate	-0.7913	0.103	-7.692	0.000	-0.993	-0.590
market_segment_type_Offline	-1.7854	0.052	-34.363	0.000	-1.887	-1.684

[Link to Appendix slide on model assumptions]

Model Performance Summary

Logistic Regression

Test performance

Training performance:

	Accuracy	Recall	Precision	F1
0	0.805451	0.632668	0.73907	0.681742

[Link to Appendix slide on model assumptions](#)

Model Performance Summary

Logistic Regression

Converting coefficients to odds interpretations;

Attributes contributing to “No Cancellations” -
no_of_adults/children/weekend_nights/week_nights/lead time
arrival year/previous cancellations/ave_price_per_room
type_of_meal Plan 2/type_of_meal_not_selected

no_of_adults: Holding all other features constant, a unit change in no_of_adults will lead to 11.49% increase in (no cancellation) odds.

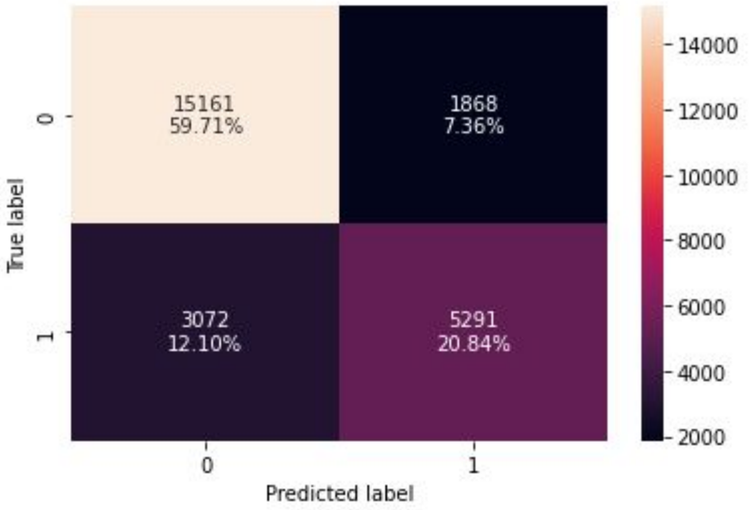
	Odds	Change_odd%
const	0.000000	-100.000000
no_of_adults	1.114910	11.490960
no_of_children	1.165459	16.545927
no_of_weekend_nights	1.114697	11.469662
no_of_week_nights	1.042584	4.258406
required_car_parking_space	0.202961	-79.703947
lead_time	1.015833	1.583312
arrival_year	1.571951	57.195078
arrival_month	0.958388	-4.161197
repeated_guest	0.064782	-93.521802
no_of_previous_cancellations	1.257118	25.711810
avg_price_per_room	1.019368	1.936838
no_of_special_requests	0.229963	-77.003739
type_of_meal_plan_Meal Plan 2	1.178464	17.846408
type_of_meal_plan_Not Selected	1.331095	33.109465
room_type_reserved_Room_Type 2	0.701041	-29.895882
room_type_reserved_Room_Type 4	0.753645	-24.635508
room_type_reserved_Room_Type 5	0.478845	-52.115481
room_type_reserved_Room_Type 6	0.379771	-62.022895
room_type_reserved_Room_Type 7	0.238271	-76.172939
market_segment_type_Corporate	0.453263	-54.673731
market_segment_type_Offline	0.167728	-83.227238

[Link to Appendix slide on model assumptions]

Model Performance Summary

Logistic Regression

Checking model performance on the training set



Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

Test performance

Training performance:

	Accuracy	Recall	Precision	F1
0	0.805451	0.632668	0.73907	0.681742

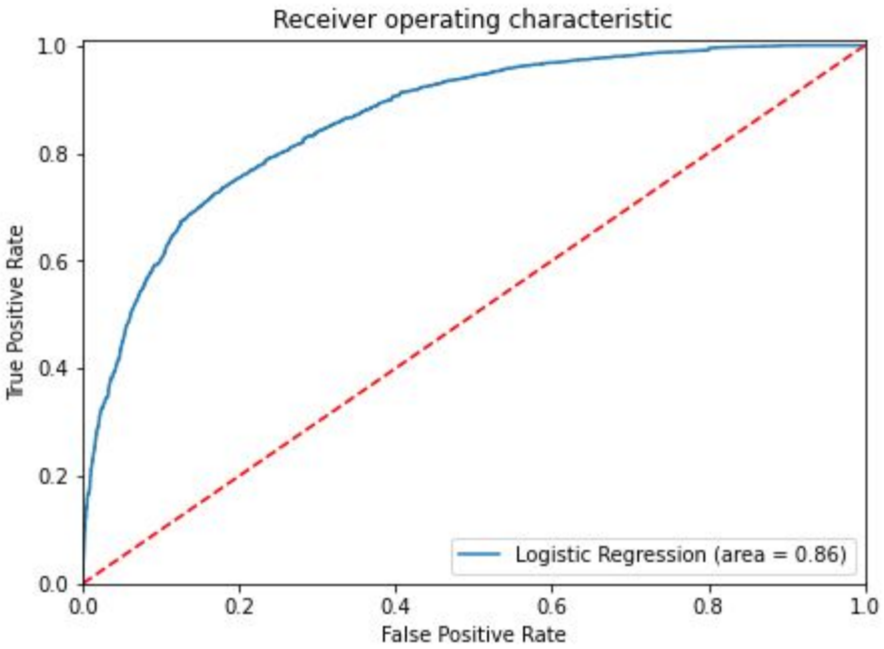
[Link to Appendix slide on model assumptions]

Model Performance Summary

Logistic Regression

ROC - AUC
Training set

Observation: Logistic Regression model
is giving a good performance on training set.

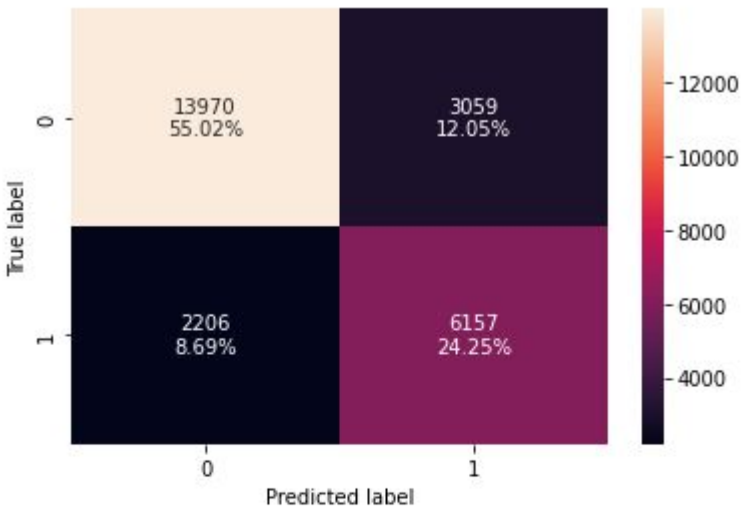


[Link to Appendix slide on model assumptions]()

Model Performance Summary

Logistic Regression

Checking model performance on the training set



[*Link to Appendix slide on model assumptions*]

Model Performance Summary

Logistic Regression

Test performance
Recall improved significantly to 0.73 compared to the former 0.63
Precision decreased from 0.73 to 0.66

Training performance:

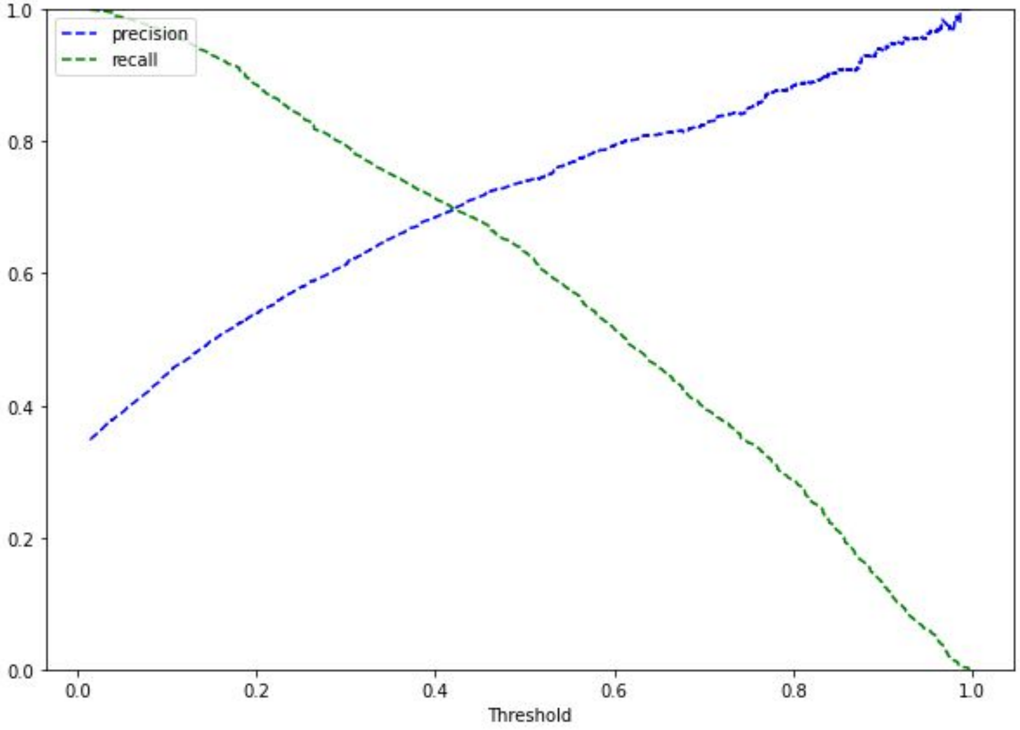
	Accuracy	Recall	Precision	F1
0	0.792651	0.736219	0.668077	0.700495

Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

At the threshold of 0.4, we get balanced recall and precision

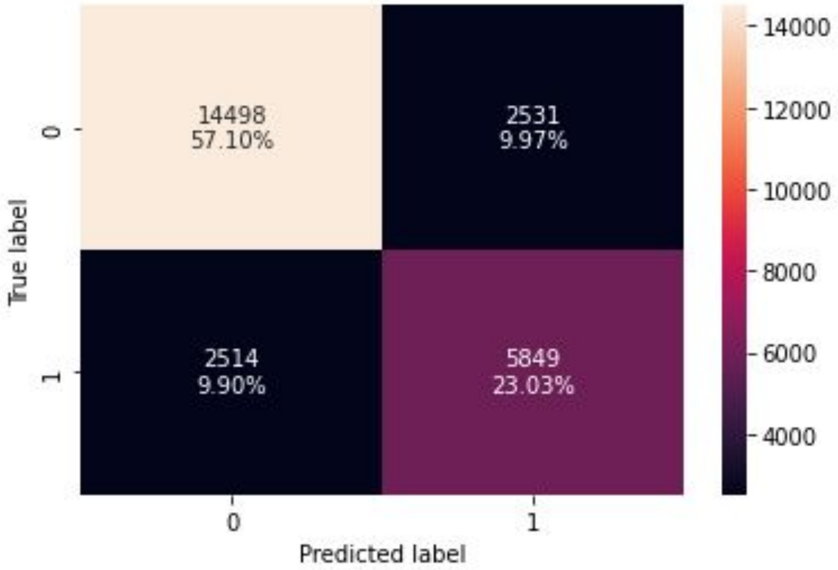


[Link to Appendix slide on model assumptions]

Model Performance Summary

Logistic Regression

Checking model performance on the training set



Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

Test performance
Recall increased
Recall decreased

Training performance:

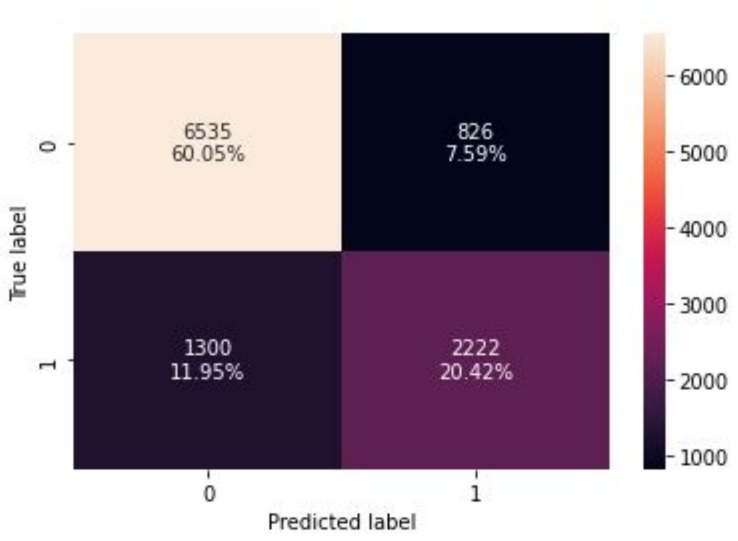
	Accuracy	Recall	Precision	F1
0	0.801315	0.69939	0.697971	0.69868

[*Link to Appendix slide on model assumptions*]

Model Performance Summary

Logistic Regression

Using model with default threshold



[*Link to Appendix slide on model assumptions*]

Model Performance Summary

Logistic Regression

Test performance

Test performance:

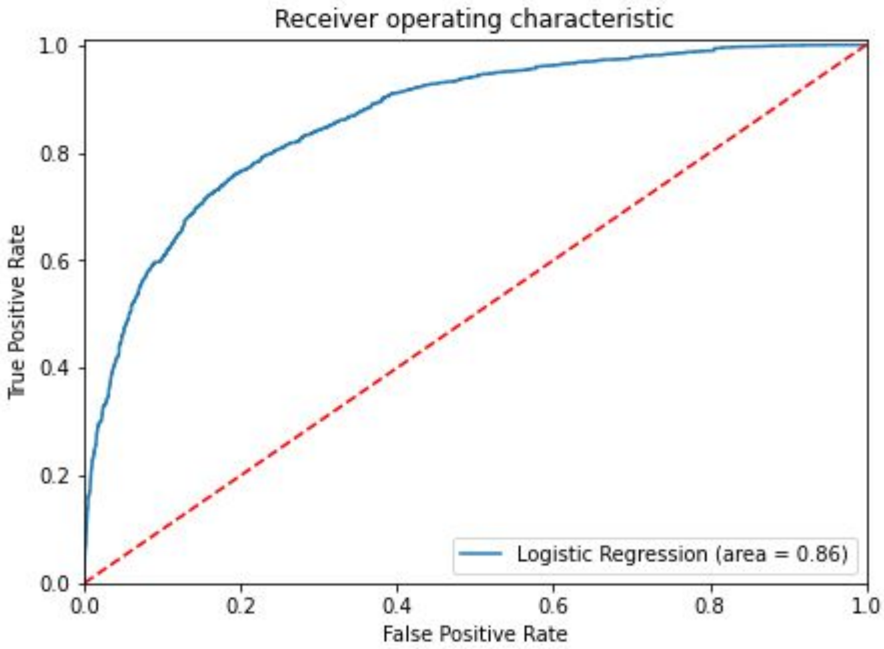
	Accuracy	Recall	Precision	F1
0	0.804649	0.630892	0.729003	0.676408

[Link to Appendix slide on model assumptions]

Model Performance Summary

Logistic Regression

Training performance

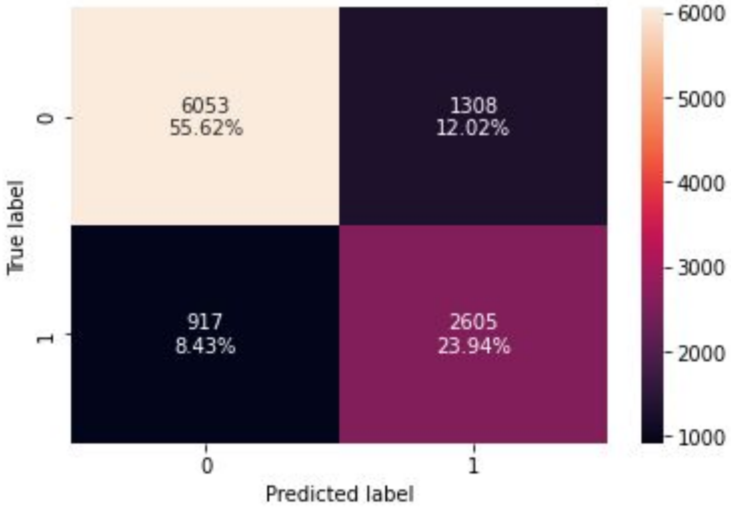


Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

Using model with threshold = 0.37



Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

Test performance

Test performance:

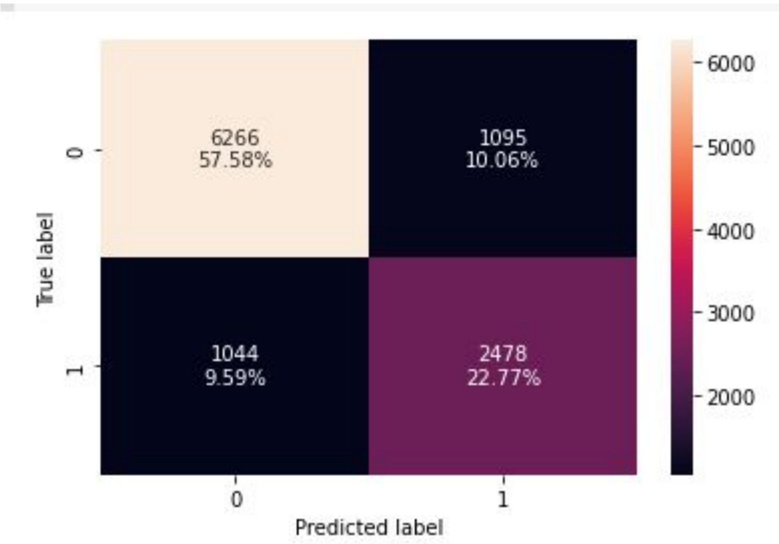
	Accuracy	Recall	Precision	F1
0	0.795553	0.739637	0.66573	0.70074

[Link to Appendix slide on model assumptions]

Model Performance Summary

Logistic Regression

Using model with threshold = 0.42



[Link to Appendix slide on model assumptions]

Model Performance Summary

Logistic Regression

Test performance

Test performance:

	Accuracy	Recall	Precision	F1
0	0.803455	0.703578	0.693535	0.69852

[Link to Appendix slide on model assumptions]

Model Performance Summary

Decision Tree

Training performance comparison

Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.805451	0.792651	0.801315
Recall	0.632668	0.736219	0.699390
Precision	0.739070	0.668077	0.697971
F1	0.681742	0.700495	0.698680

Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

Test set performance comparison

Test set performance comparison:

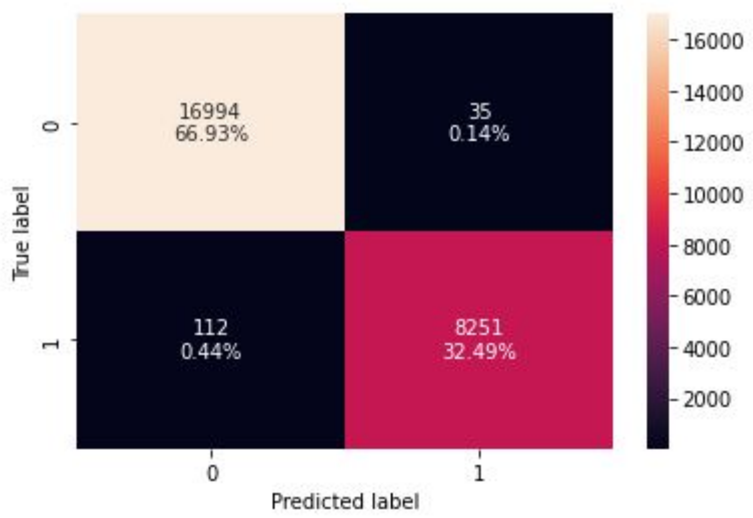
	Logistic Regression statsmodel	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.804649	0.795553	0.803455
Recall	0.630892	0.739637	0.703578
Precision	0.729003	0.665730	0.693535
F1	0.676408	0.700740	0.698520

Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

Checking model performance on training set



[Link to Appendix slide on model assumptions]()

Model Performance Summary

Model Performance Check

Checking model performance on training set

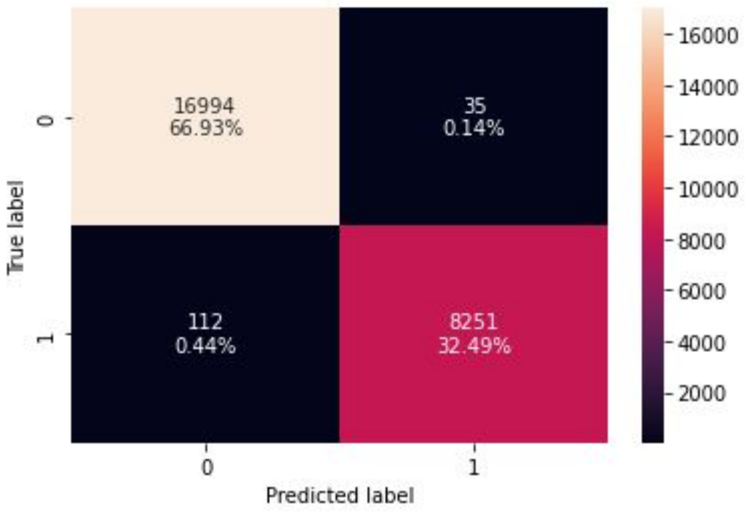
	Accuracy	Recall	Precision	F1
0	0.994211	0.986608	0.995776	0.991171

Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

Checking model performance on test set



Link to Appendix slide on model assumptions

Model Performance Summary

Model Performance Check

Checking model performance on test set

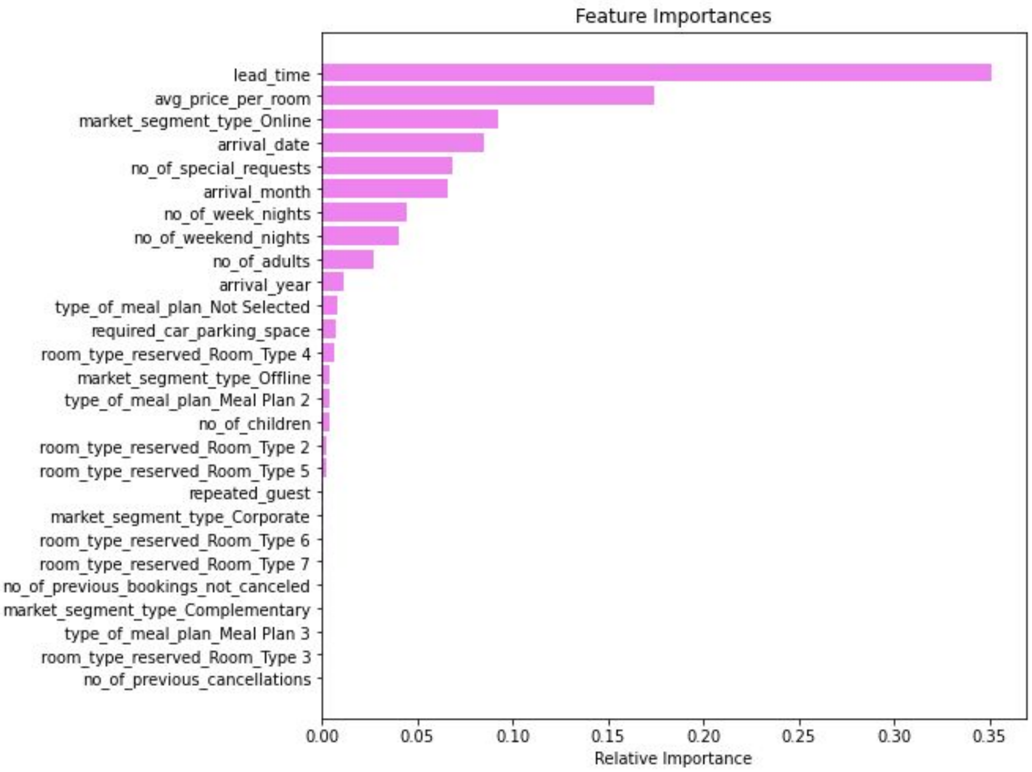
	Accuracy	Recall	Precision	F1
0	0.994211	0.986608	0.995776	0.991171

Link to Appendix slide on model assumptions

Model Performance Summary

Checking important features

In pre-tuned decision tree, lead_time and avg_price_per_room is the most important features.

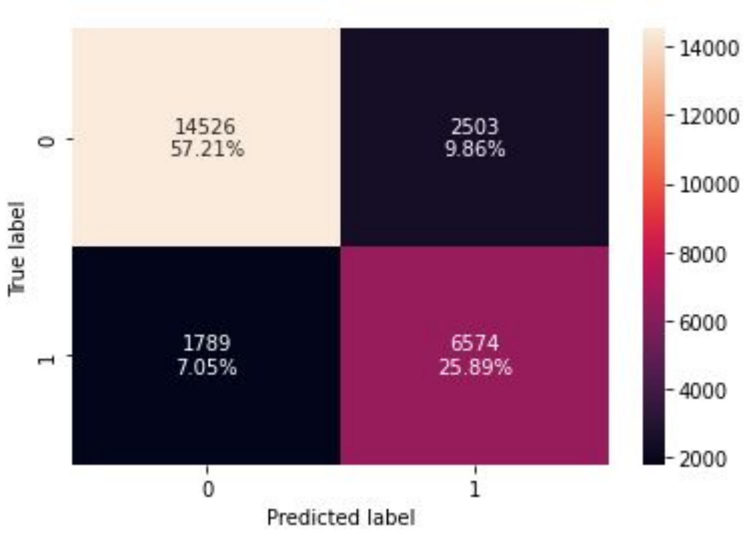


[*Link to Appendix slide on model assumptions*]

Model Performance Summary

Logistic Regression

Checking performance on training set



[*Link to Appendix slide on model assumptions*]()

Model Performance Summary

Model Performance Check

Checking performance on test set

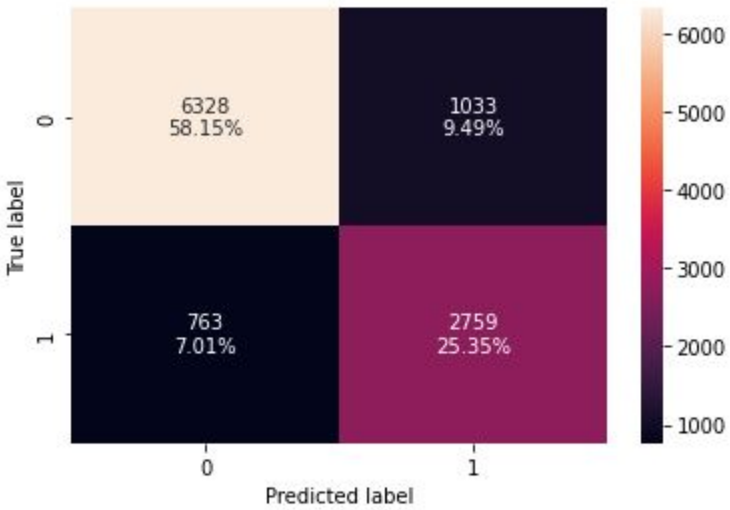
	Accuracy	Recall	Precision	F1
0	0.83097	0.786082	0.724248	0.753899

Link to Appendix slide on model assumptions

Model Performance Summary

Logistic Regression

Checking performance on test set



[Link to Appendix slide on model assumptions]

Model Performance Summary

Model Performance Check

Checking performance on test set

Since the model is giving a generalized Result comparing recall scores on both The train and test data (0.78), this shows The model is able to generalize well on unseen data.

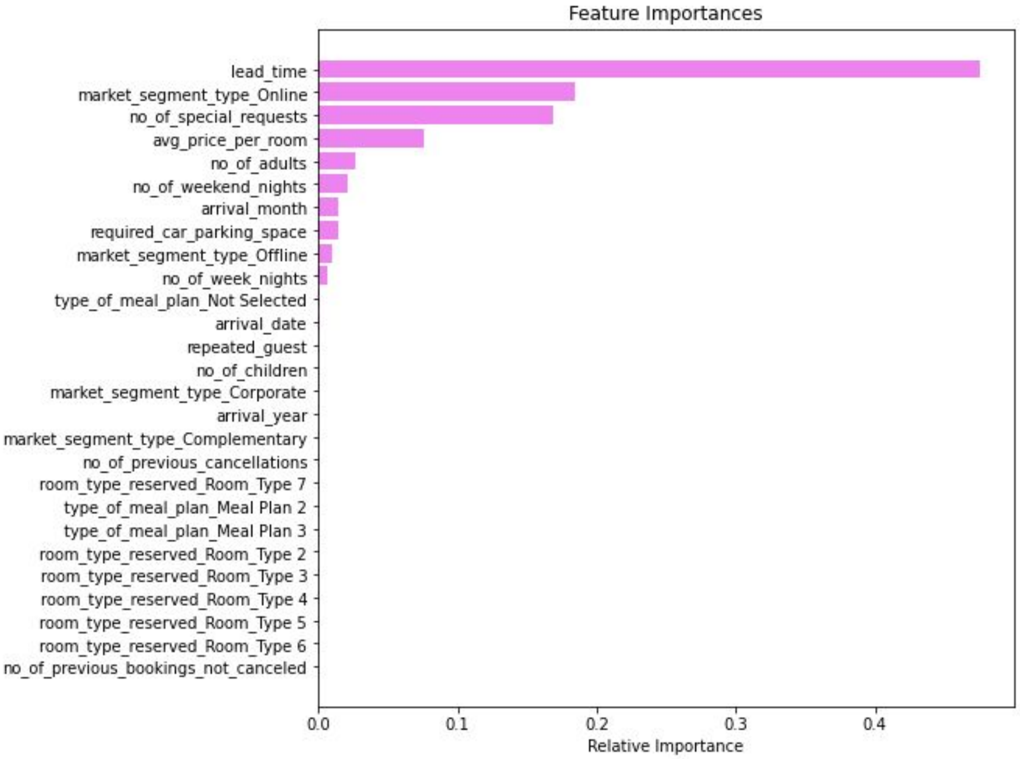
	Accuracy	Recall	Precision	F1
0	0.834972	0.783362	0.727584	0.754444

[Link to Appendix slide on model assumptions]

Model Performance Summary

Visualizing the Decision Tree

In the pretuned decision tree, the lead_time and market_segment_type_Online are the most important features. Avg_price_per_room dropped to third place.



[Link to Appendix slide on model assumptions]

Model Performance Summary

Cost Complexity Pruning

DataFrame (path)

	ccp_alphas	impurities
0	0.000000e+00	0.008376
1	0.000000e+00	0.008376
2	2.933821e-20	0.008376
3	2.933821e-20	0.008376
4	2.933821e-20	0.008376
...	...	...
1839	8.901596e-03	0.328058
1840	9.802243e-03	0.337860
1841	1.271875e-02	0.350579
1842	3.412090e-02	0.418821
1843	8.117914e-02	0.500000

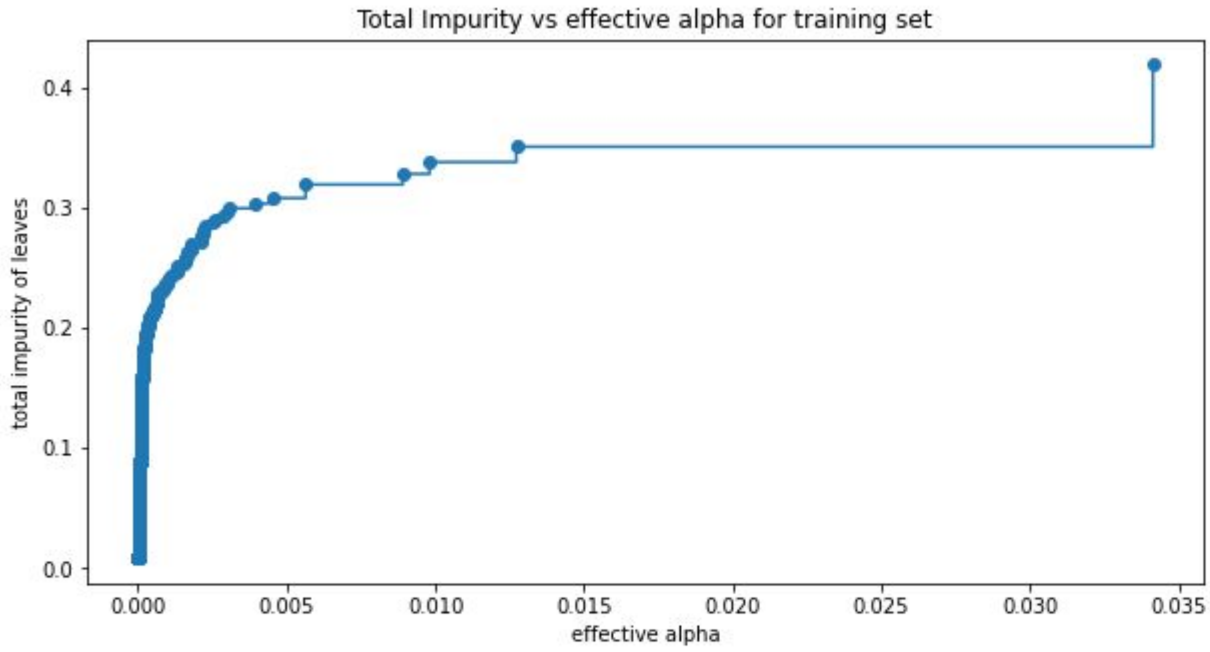
1844 rows × 2 columns

[Link to Appendix slide on model assumptions]

Model Performance Summary

Cost Complexity Pruning

Total Impurity vs effective alpha for training set



[*Link to Appendix slide on model assumptions*]

Model Performance Summary

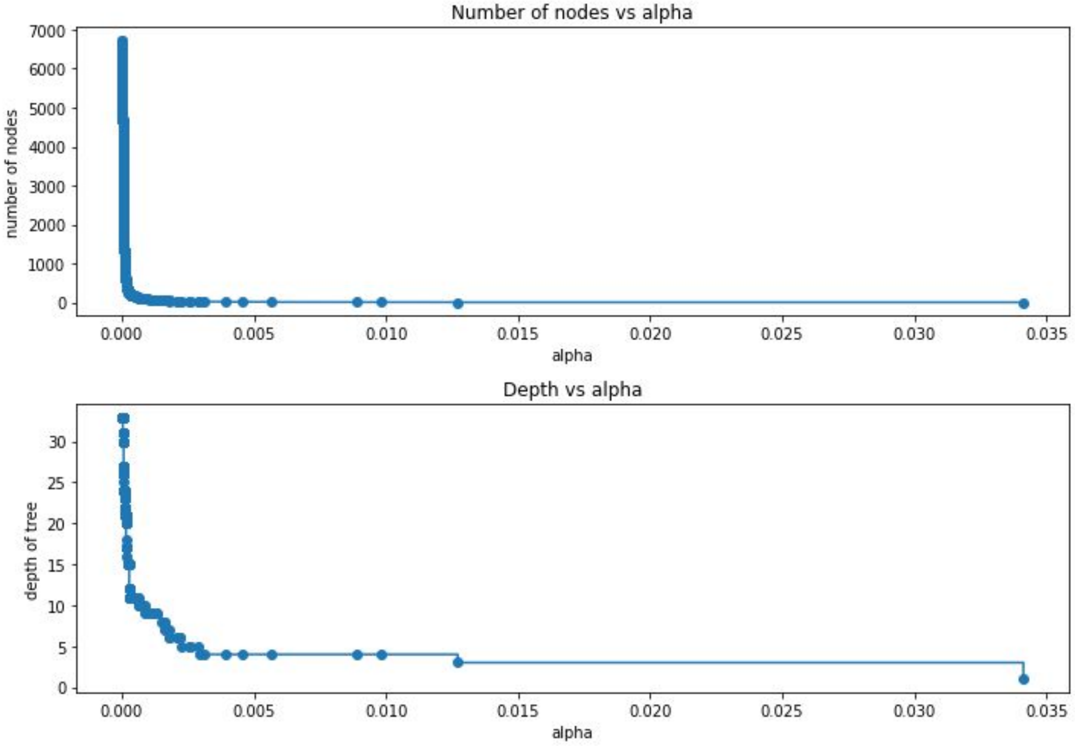
Cost Complexity Pruning

```
Number of nodes in the last tree is: 1 with ccp_alpha: 0.0811791438913696
```

[*Link to Appendix slide on model assumptions*]

Model Performance Summary

Cost Complexity Pruning

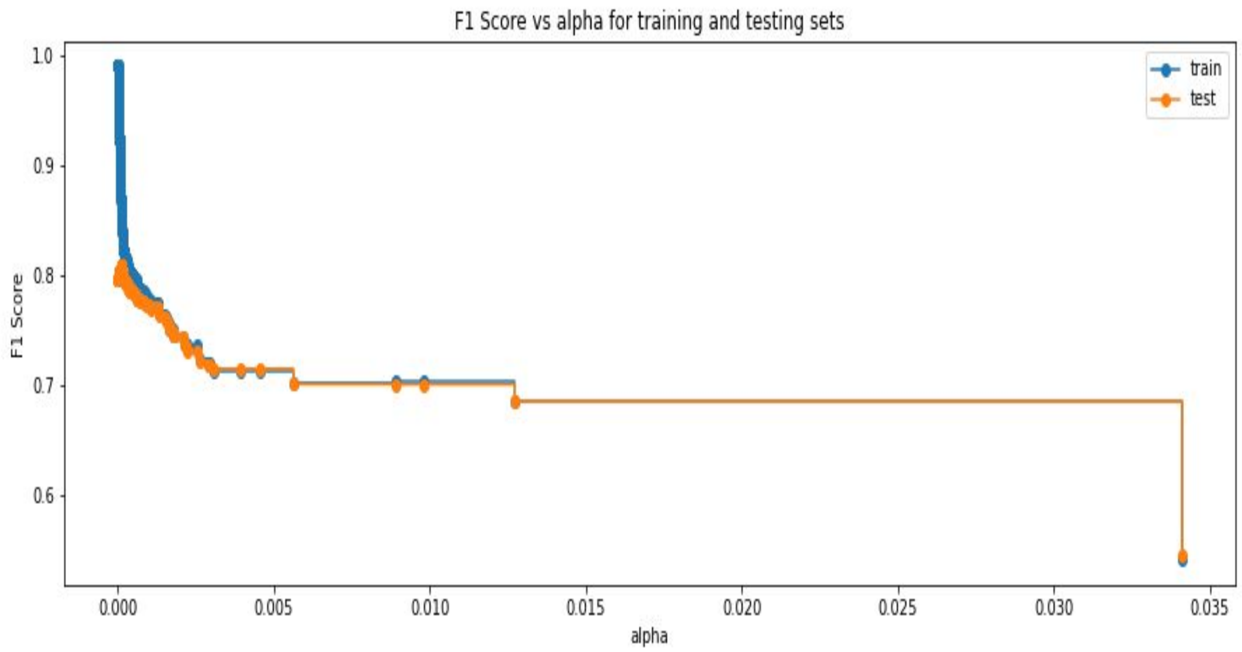


[*Link to Appendix slide on model assumptions*]

Model Performance Summary

Cost Complexity Pruning

F1 Score vs alpha for training and testing sets

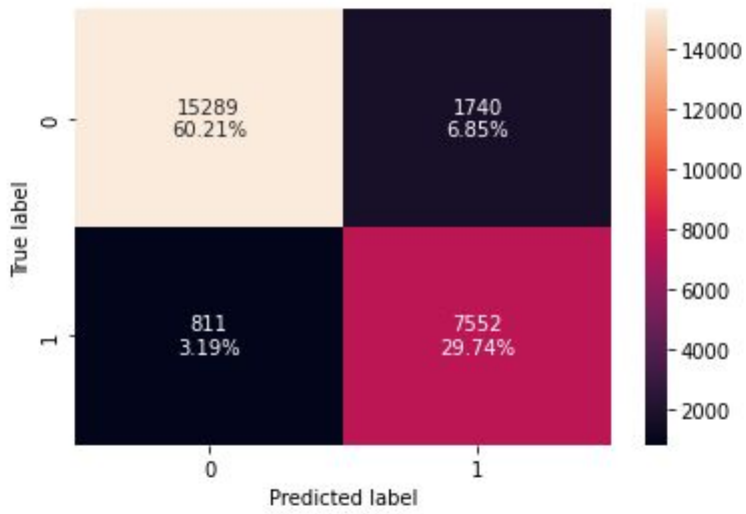


[Link to Appendix slide on model assumptions]

Model Performance Summary

Cost Complexity Pruning

Checking performance on training set



[Link to Appendix slide on model assumptions]()

Model Performance Summary

Cost Complexity Pruning

Checking performance on training set

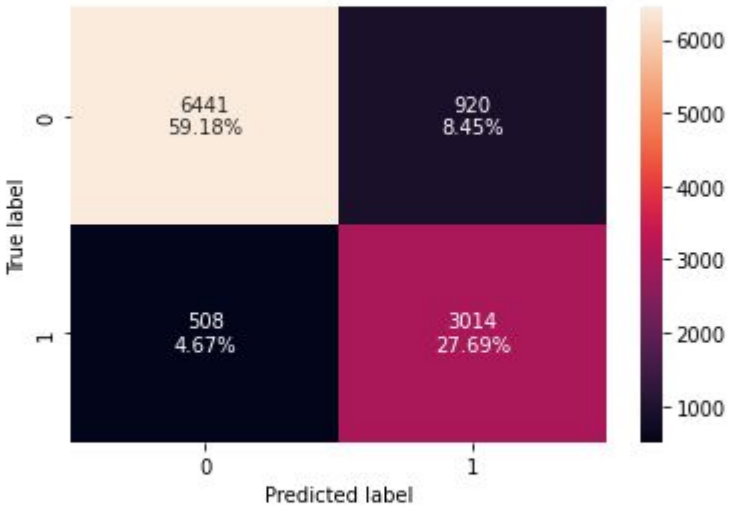
	Accuracy	Recall	Precision	F1
0	0.899535	0.903025	0.812742	0.855508

Link to Appendix slide on model assumptions

Model Performance Summary

Cost Complexity Pruning

Checking performance on test set



[*Link to Appendix slide on model assumptions*]

Model Performance Summary

Cost Complexity Pruning

Checking performance on test set

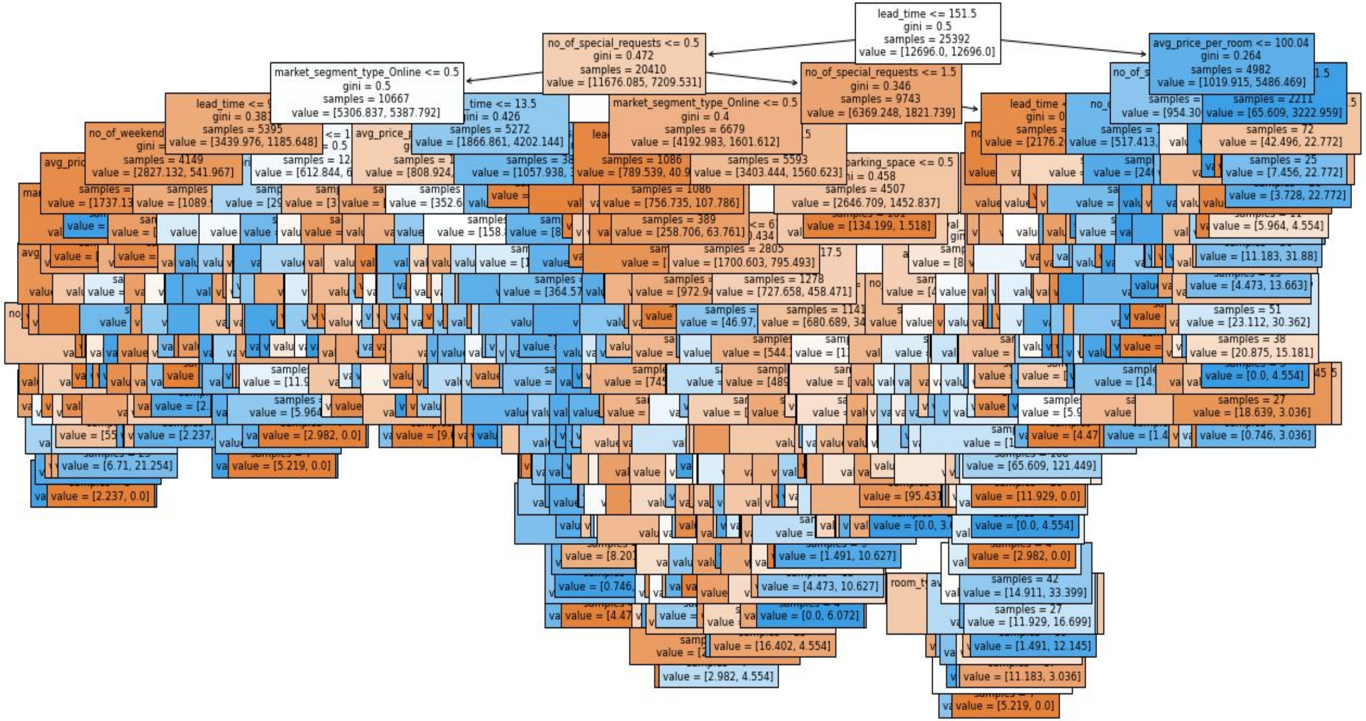
	Accuracy	Recall	Precision	F1
0	0.868786	0.855764	0.766141	0.808476

Link to Appendix slide on model assumptions

Model Performance Summary

Cost Complexity Pruning

Checking performance on test set

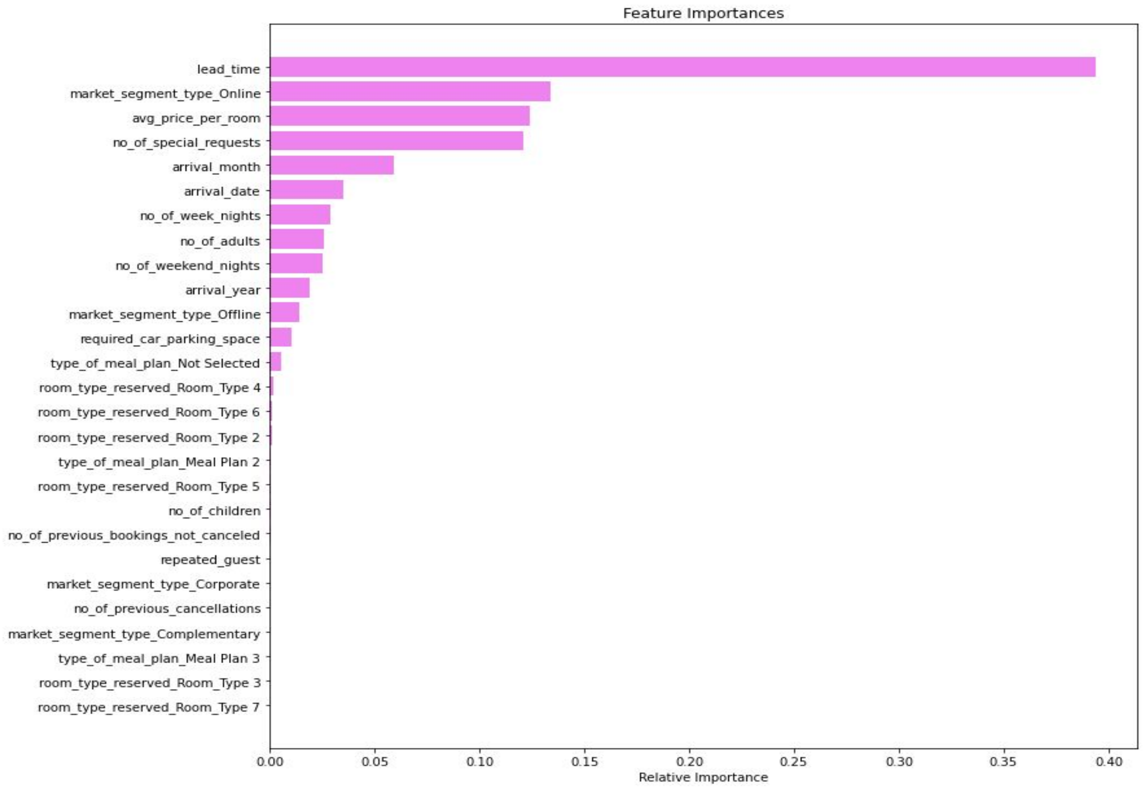


[Link to Appendix slide on model assumptions](#)

Model Performance Summary

Cost Complexity Pruning

The lead_time and market_segment_type_Online are the most important features.



[Link to Appendix slide on model assumptions]()

Model Performance Summary

Decision Tree

Training & Test performance comparison are giving generalized results.

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.994211	0.830970	0.899535
Recall	0.986608	0.786082	0.903025
Precision	0.995776	0.724248	0.812742
F1	0.991171	0.753899	0.855508

Test set performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.994211	0.834972	0.899535
Recall	0.986608	0.783362	0.903025
Precision	0.995776	0.727584	0.812742
F1	0.991171	0.754444	0.855508

Link to Appendix slide on model assumptions

Actionable Insights and Recommendations

- We have built a predictive model that INN Hotels can use to determine which booking will likely be canceled.
- All the logistic regression models have been given a generalized performance on the training and test set.
- The lead time was identified as the most important feature; a longer lead time increases the odds of cancellations. Policies need to be introduced to restrict how far in advance bookings can be made before the check-in date.
- Hotel policies must restrict the length of stay as bookings for more extended stay periods also increase the odds of cancellations.
- The repeat guests are identified to have lower odds of cancellations. Hotel policies need to incentivize current & previous guests to increase conversion as repeated guests.
- More bookings and cancellations were found to occur over months (March-August) compared to (September-February)
- Observing market segments, the avg price per room has been higher in instances where bookings have been canceled than in cases in which bookings have not been canceled. More competition information is required to ensure that our pricing is competitive to retain guests.

Happy Learning !