InnHotels Project

By: Nihal Kala

Problem Statement

- Many hotel bookings are cancelled due to change of plans and scheduling conflicts. While it may help guests as they aren't charged very much, it serves to decrease the revenue profits of hotels.
- With online booking channels, they have changed customer behavior.
- Goal: With a machine learning solution, this will predict which booking will be cancelled and what factors highly influence booking cancellations by using a predictive model for figuring out which booking will be cancelled in advance.

Head and Tail of Data

	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
0	INN00001	2	0	1	2	Meal Plan 1	0	Room_Type 1	
1	INN00002	2	0	2	3	Not Selected	0	Room_Type 1	
2	INN00003	1	0	2	1	Meal Plan 1	0	Room_Type 1	
3	INN00004	2	0	0	2	Meal Plan 1	0	Room_Type 1	
4	INN00005	2	0	1	1	Not Selected	0	Room_Type 1	

	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
36270	INN36271	3	0	2	6	Meal Plan 1	0	Room_Type 4
36271	INN36272	2	0	1	3	Meal Plan 1	0	Room_Type 1
36272	INN36273	2	0	2	6	Meal Plan 1	0	Room_Type 1
36273	INN36274	2	0	0	3	Not Selected	0	Room_Type 1
36274	INN36275	2	0	1	2	Meal Plan 1	0	Room_Type 1

Shape of Data: 36275 rows, 18 columns
No duplicated values

Data Info

```
<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
```

Data head without booking ID

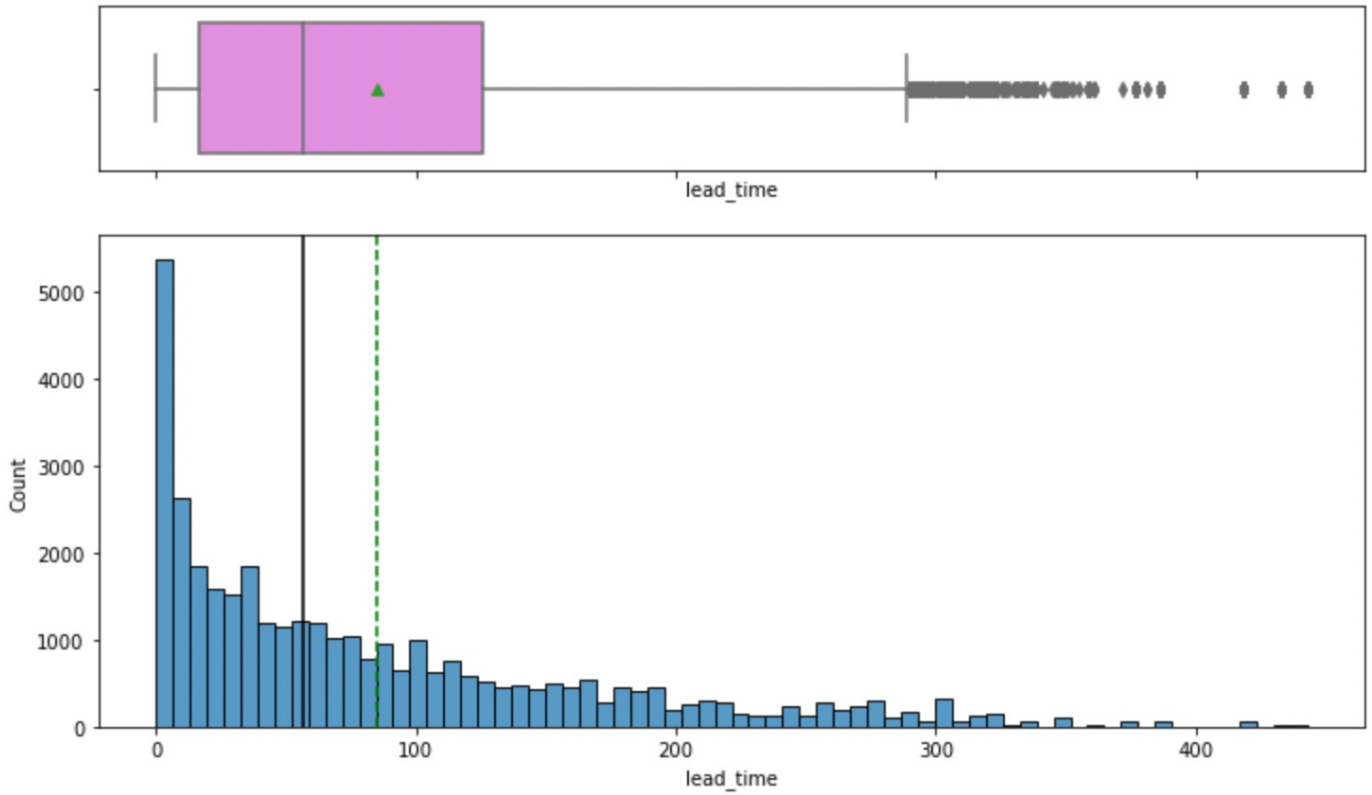
	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space	room_type_reserved	lead_time	arrival_year	arrival_mor
0	2	0	1	2	Meal Plan 1	0	Room_Type 1	224	2017	
1	2	0	2	3	Not Selected	0	Room_Type 1	5	2018	
2	1	0	2	1	Meal Plan 1	0	Room_Type 1	1	2018	
3	2	0	0	2	Meal Plan 1	0	Room_Type 1	211	2018	
4	2	0	1	1	Not Selected	0	Room_Type 1	48	2018	

Description of Data

	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
count	36275.00000	36275.00000	36275.00000	36275.00000	36275.00000	36275.00000	36275.00000	36275.00000	36275.00000	36275.00000
mean	1.84496	0.10528	0.81072	2.20430	0.03099	85.23256	2017.82043	7.42365	15.59700	0.02564
std	0.51871	0.40265	0.87064	1.41090	0.17328	85.93082	0.38384	3.06989	8.74045	0.15805
min	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	2017.00000	1.00000	1.00000	0.00000
25%	2.00000	0.00000	0.00000	1.00000	0.00000	17.00000	2018.00000	5.00000	8.00000	0.00000
50%	2.00000	0.00000	1.00000	2.00000	0.00000	57.00000	2018.00000	8.00000	16.00000	0.00000
75%	2.00000	0.00000	2.00000	3.00000	0.00000	126.00000	2018.00000	10.00000	23.00000	0.00000
max	4.00000	10.00000	7.00000	17.00000	1.00000	443.00000	2018.00000	12.00000	31.00000	1.00000

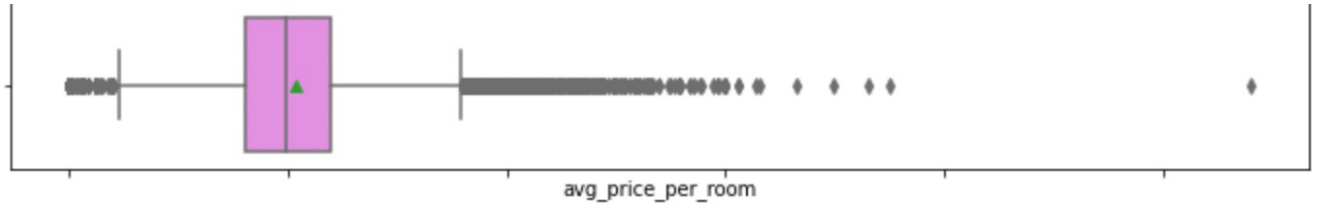
no_of_previous_cancellations	no_of_previous_bookings_not_canceled	avg_price_per_room	no_of_special_requests
36275.00000	36275.00000	36275.00000	36275.00000
0.02335	0.15341	103.42354	0.61966
0.36833	1.75417	35.08942	0.78624
0.00000	0.00000	0.00000	0.00000
0.00000	0.00000	80.30000	0.00000
0.00000	0.00000	99.45000	0.00000
0.00000	0.00000	120.00000	1.00000
13.00000	58.00000	540.00000	5.00000

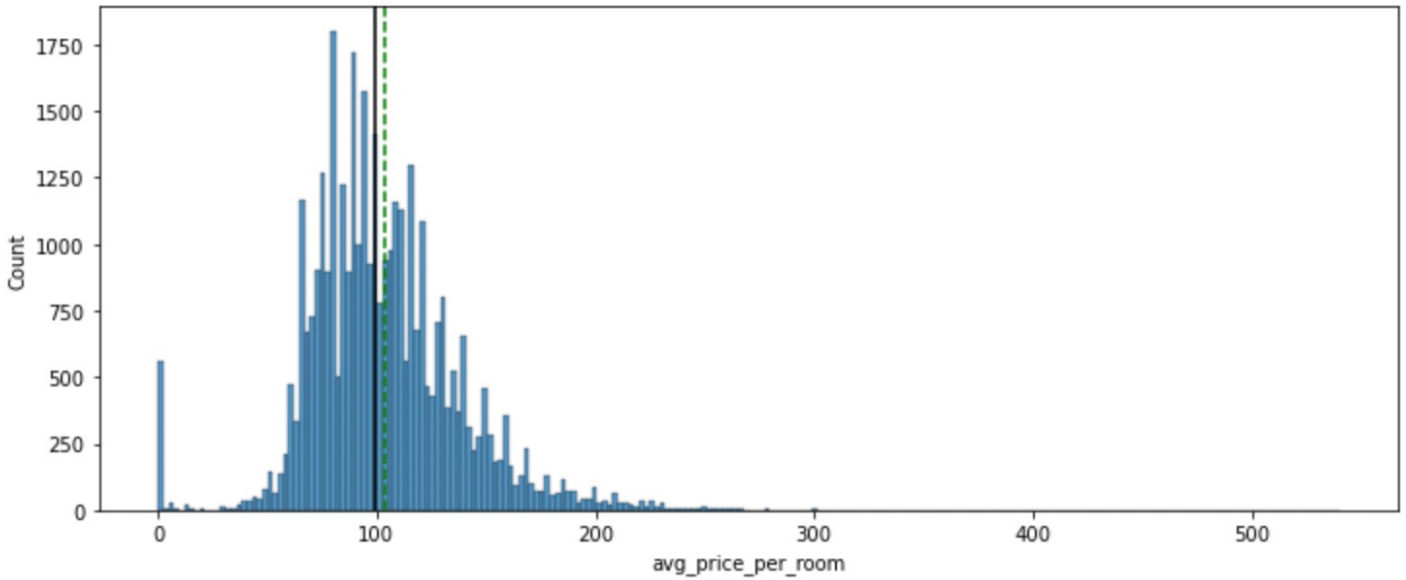
Lead Time Histogram/Box Plot



Distribution of lead time is right-skewed.
Boxplot shows outliers on right end.

Average price per room Histogram/Box Plot





Mean average price per room: 103.42
Lower and Upper Quartiles: 80.3 and 120

Average Price Per Room by Market Segment Type

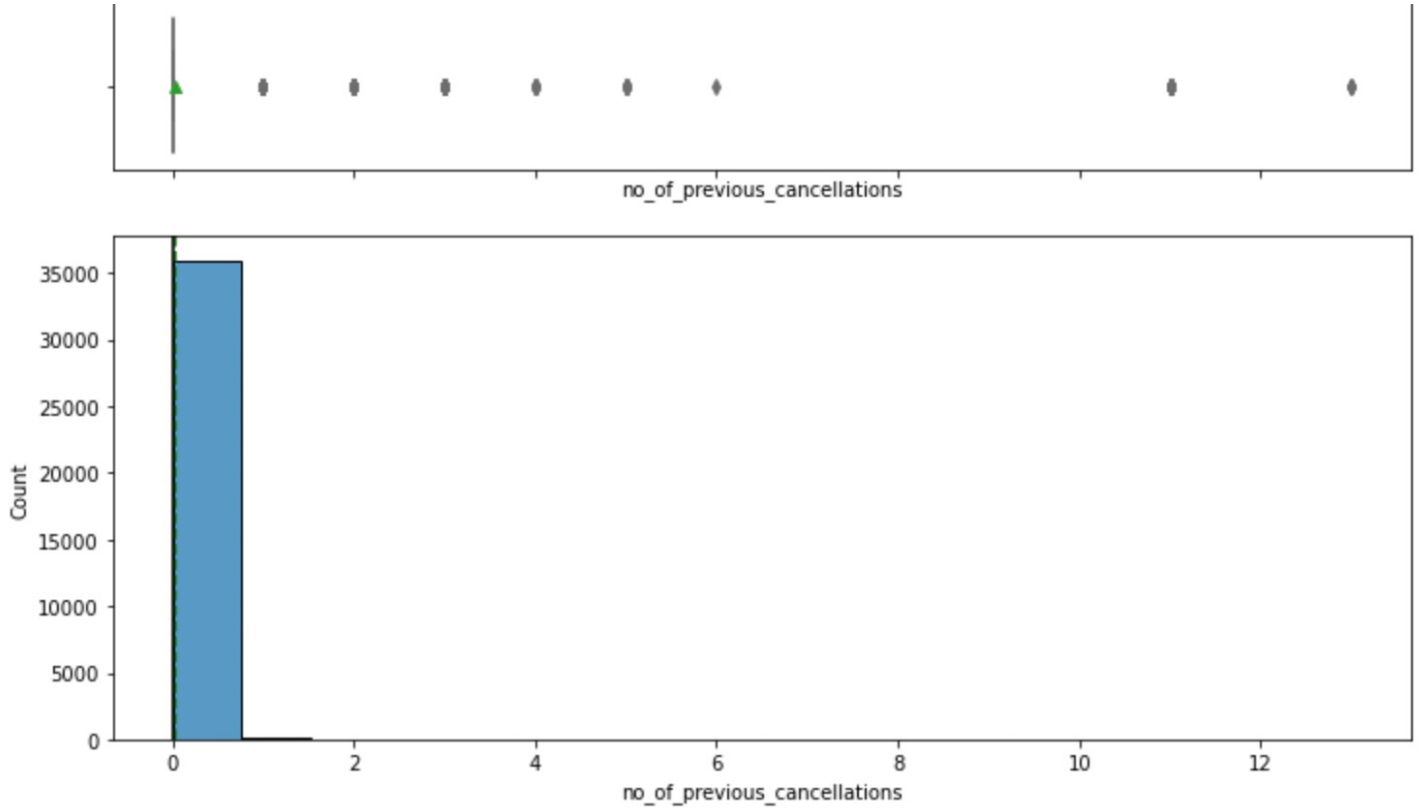
```
Complementary     354
Online            191
Name: market_segment_type, dtype: int64
```

More people prefer complementary over online type.

Interquartile Range Upper Whisker

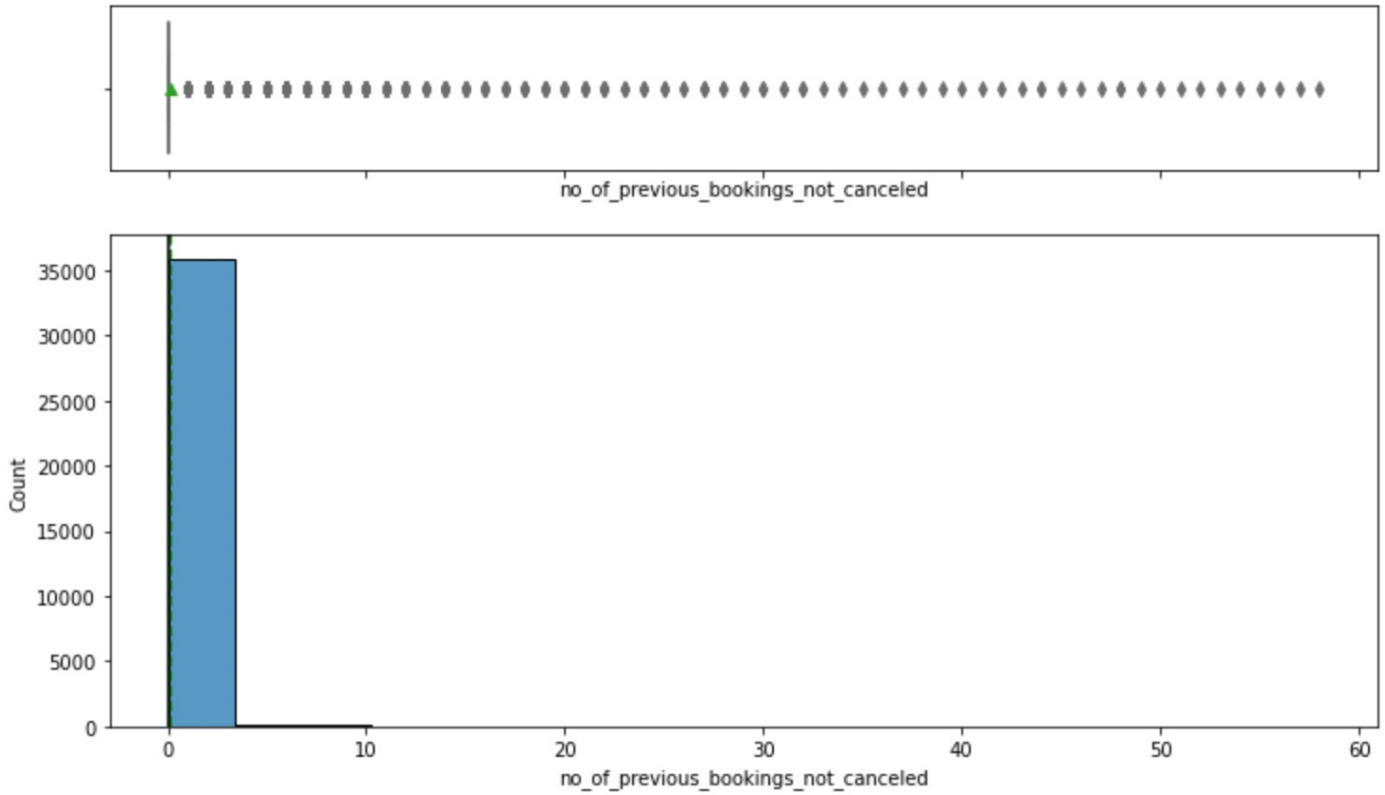
- IQR = Q3-Q1
- 120-80.3=39.7
- (39.7)(1.5)+120= 179.55

Histogram/Box Plot: Previous cancellations



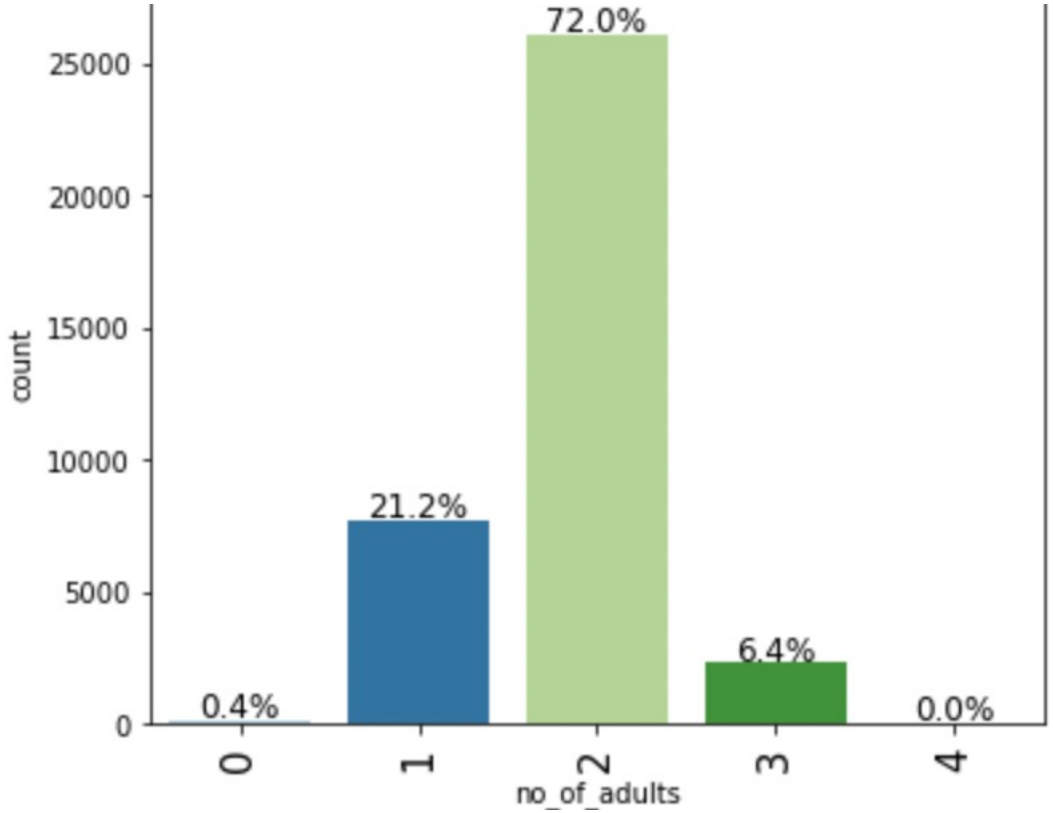
Distribution is right skewed.
Boxplot shows outliers to the right end.
Very few customers had previous cancellations.

Histogram/Boxplot: Previous Bookings Not Cancelled



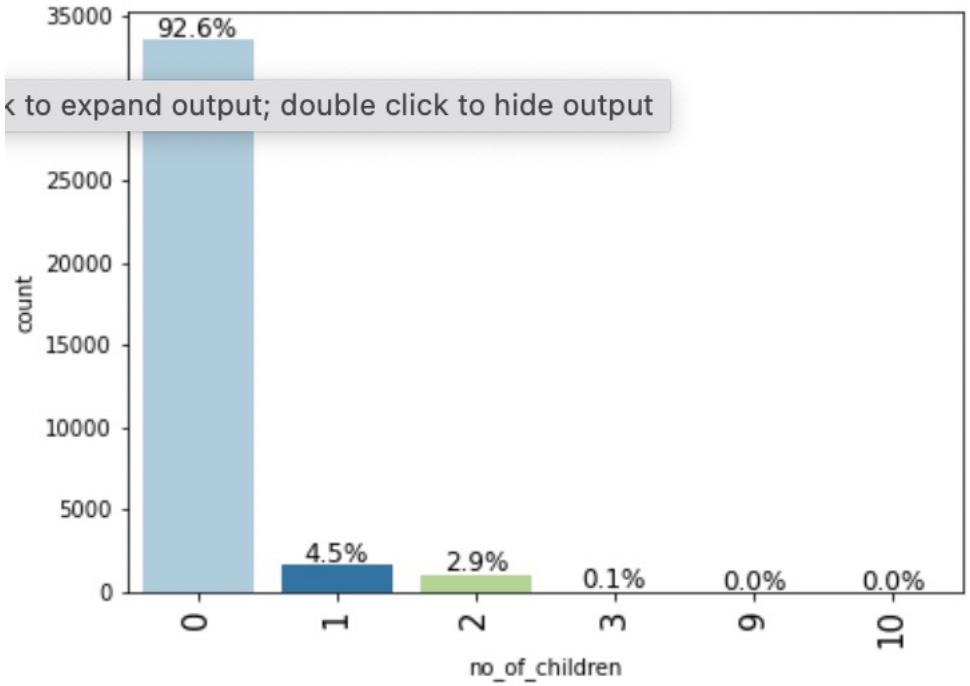
Distribution is right skewed.
Boxplot shows outliers to the right end.
Very few customers had previous bookings not cancelled.

Bar Graph: Number of adults



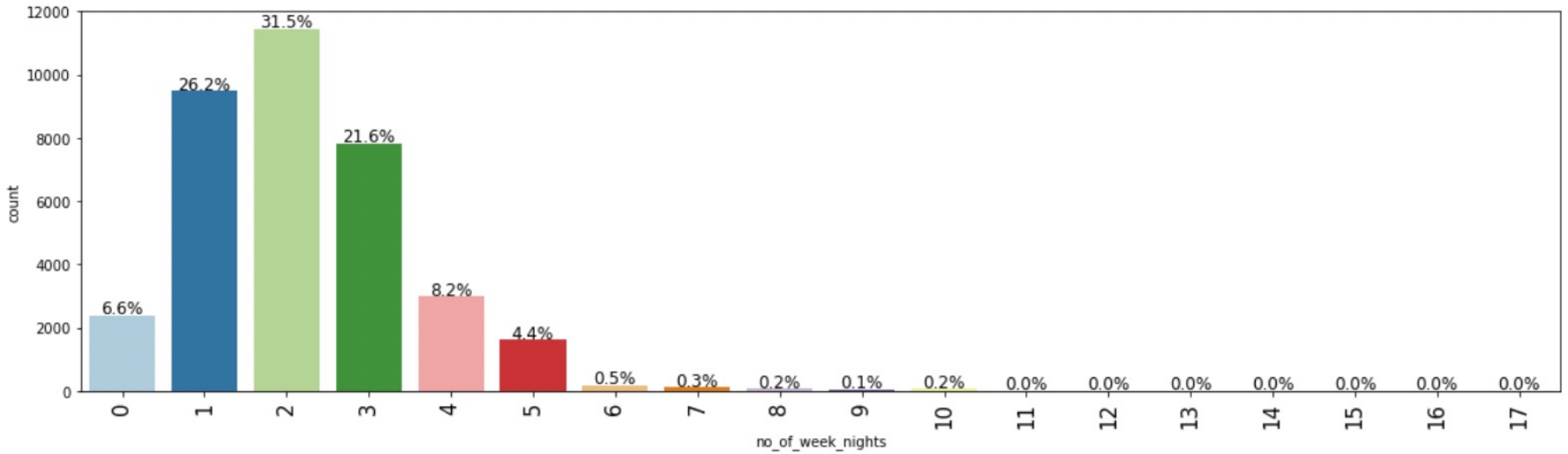
Bookings were most for two adults and least for 0 and 4 adults.

Bar graph: Number of children



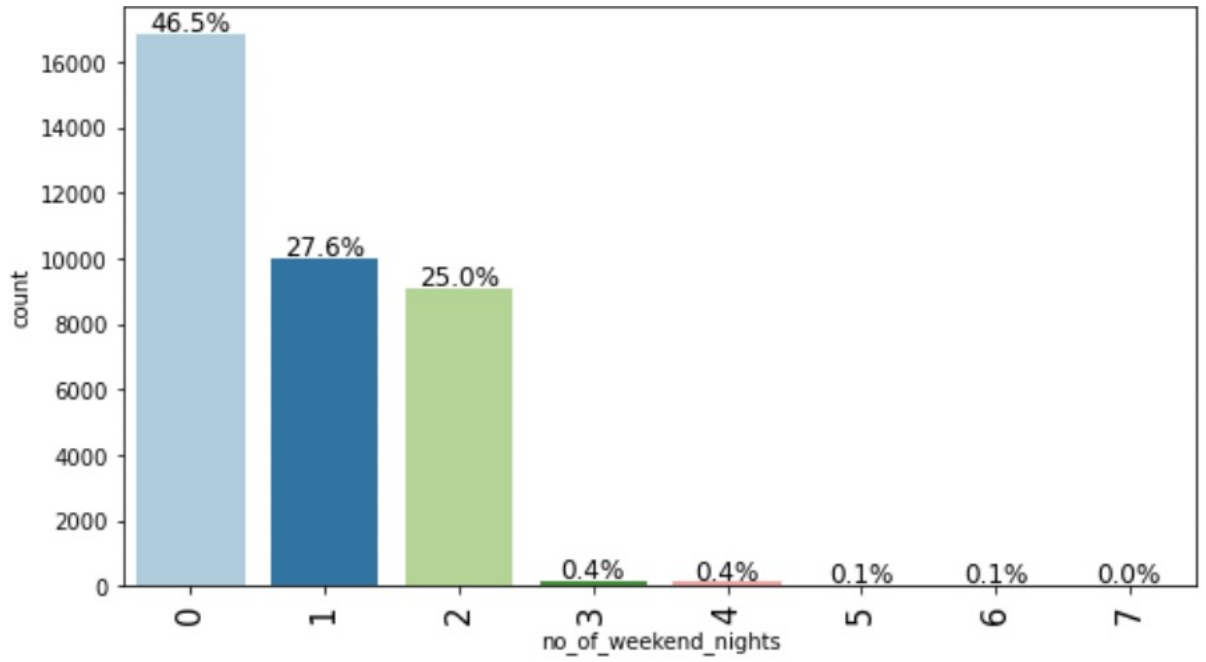
Bookings were most for 0 children and least for 3 and more children.

Bar graph: Number of weeknights



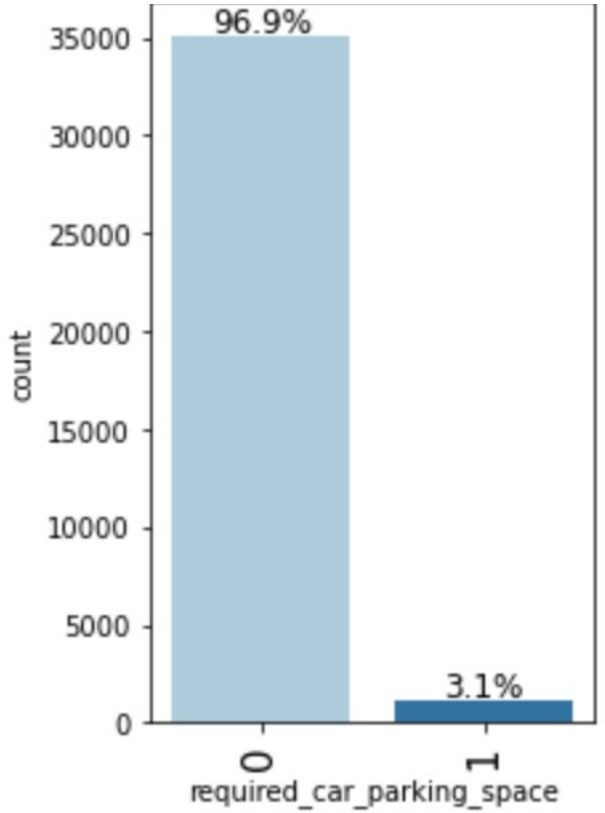
Most guests stayed 2 weeknights at the hotel.

Bar graph: Number of Weekend Nights



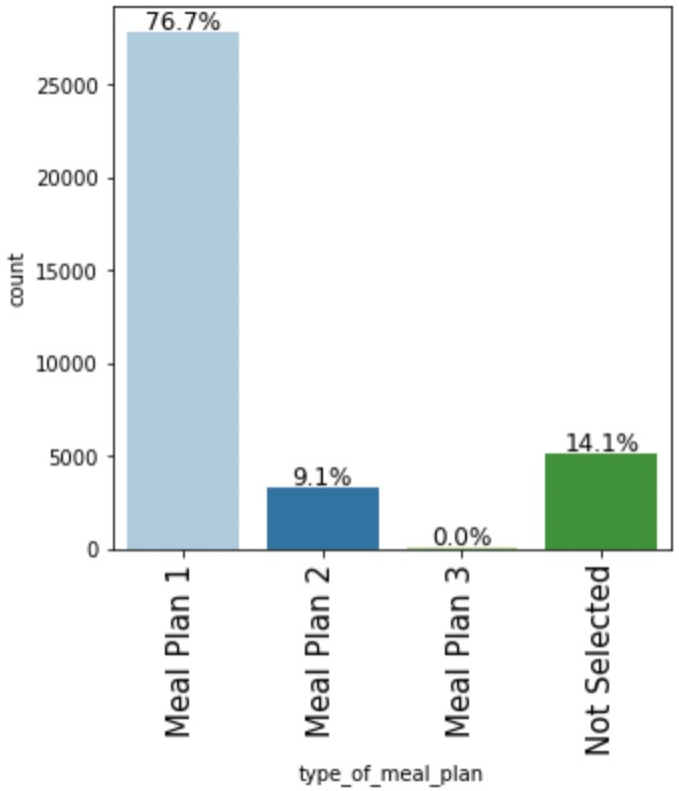
Most guests stayed 0 weekend nights at the hotel, while the least stayed for 3 or more weekend nights.

Bar graph: Required car parking space



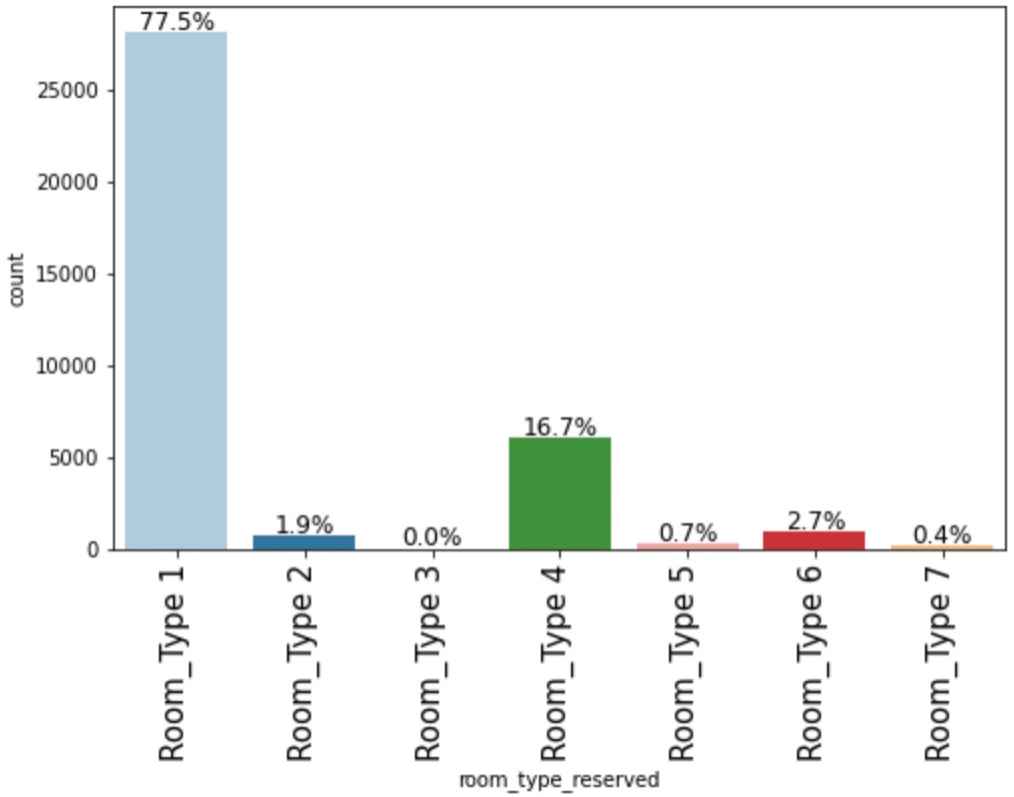
Most customers don’t require car parking space.

Bar graph: Meal plan



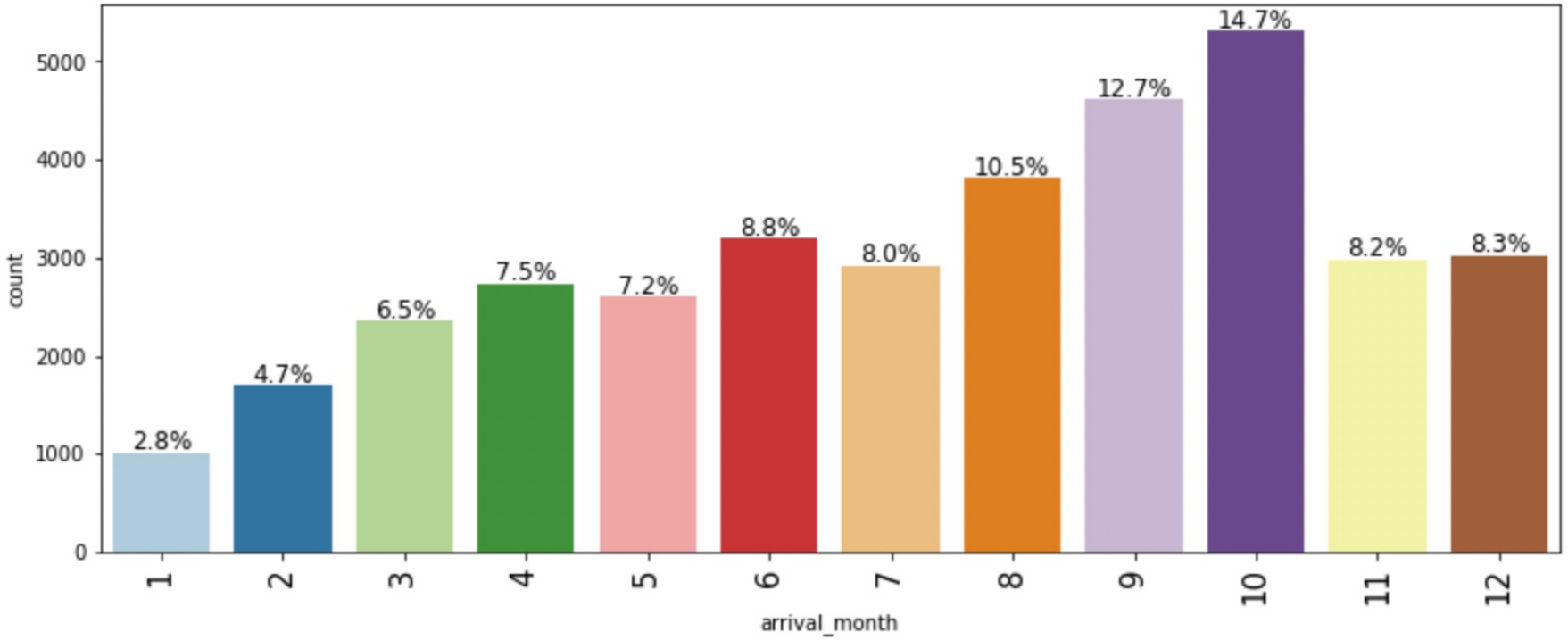
Most customers had a meal plan of breakfast, while no one had a full board meal plan.

Bar graph: Room type reserved



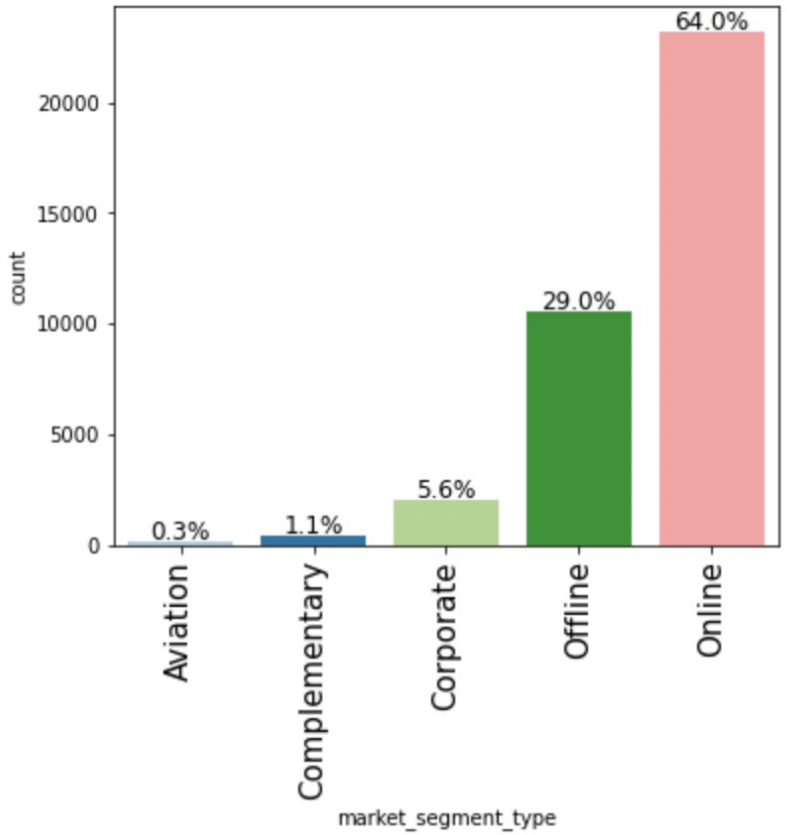
Most customers reserved Room Type 1, while the least reserved were Room type 3 and Room type 7.

Bar graph: Arrival Month



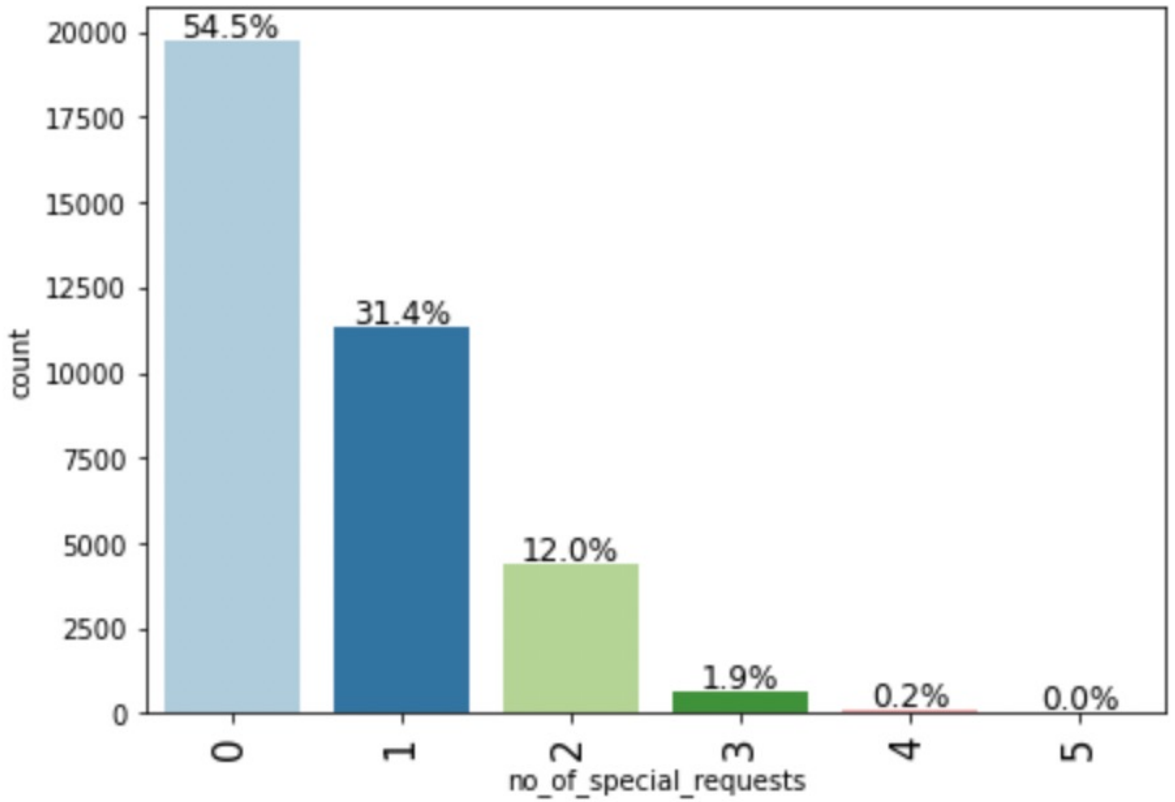
Most people arrived in Month 10 and least people arrived in Month 1.

Bar graph: Market Segment Type



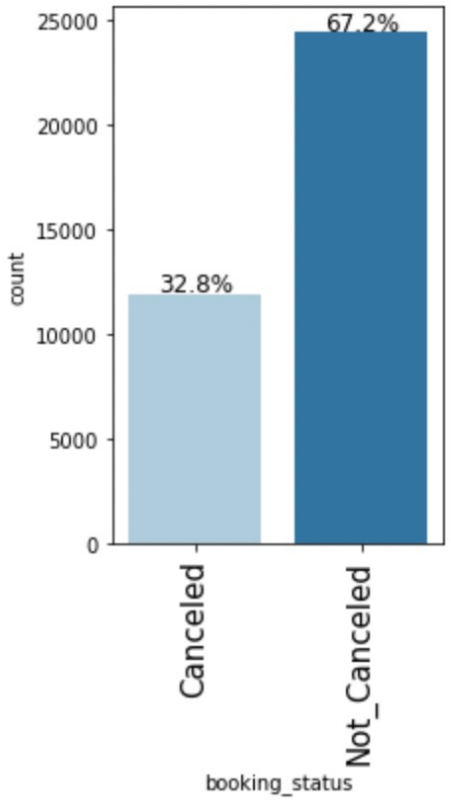
Most preferred market segment type: Online
Least preferred market segment type: Aviation

Bar graph: Number of special requests



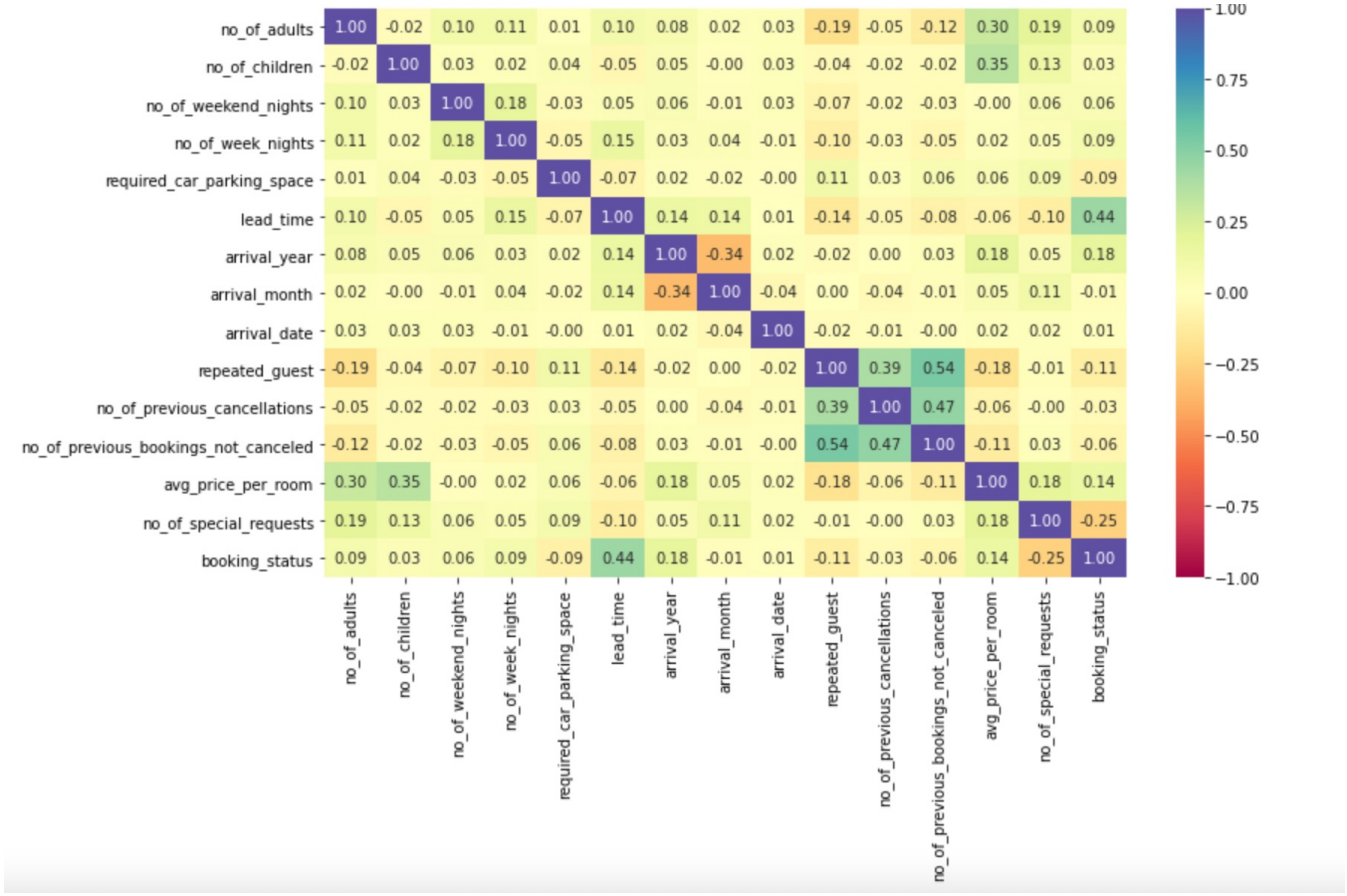
Most frequency had bookings with 0 special request, while it was significantly the least with 3 and 4 special requests.

Bar graph: Cancellation status



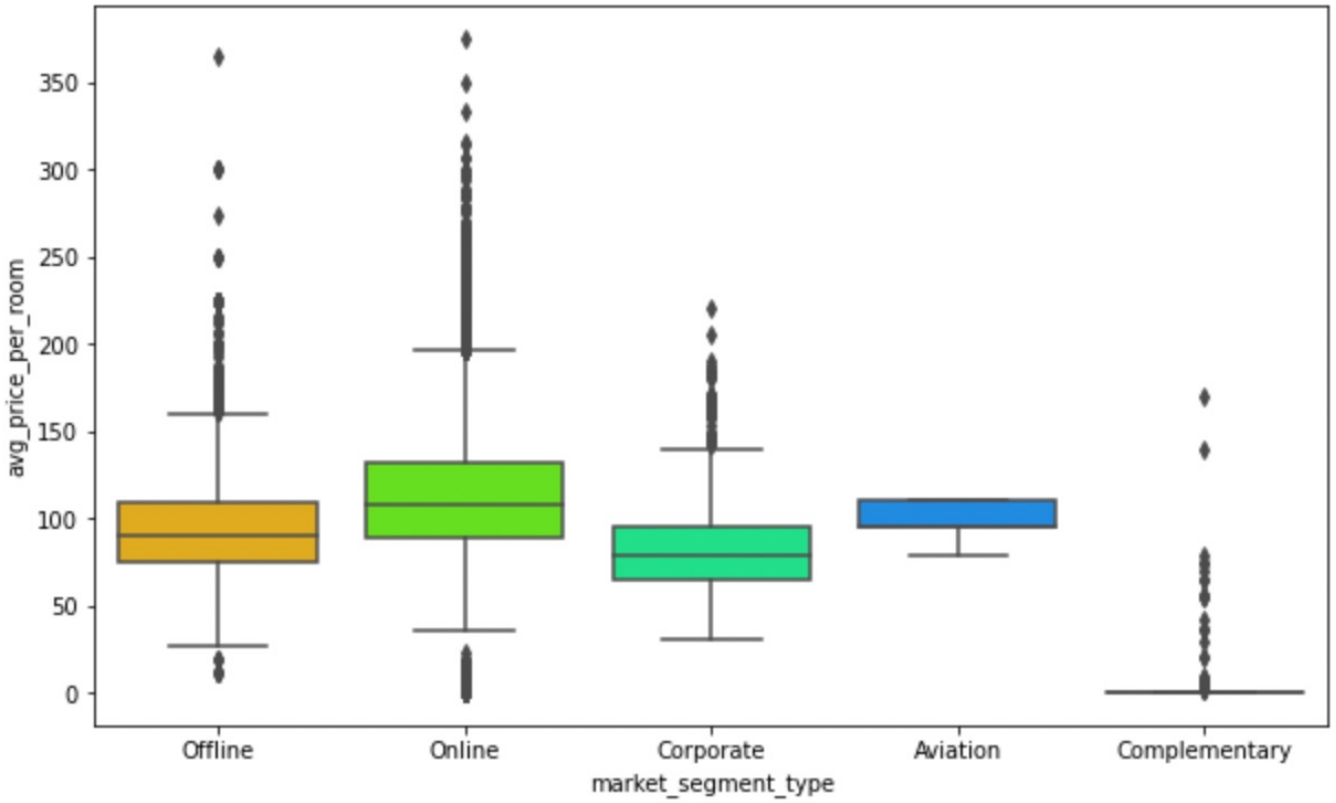
More people have a noncancelled booking status than a cancelled booking status.

Heat Map



Not much significant correlation, with most significant one being between repeated guest and number of previous bookings not cancelled.

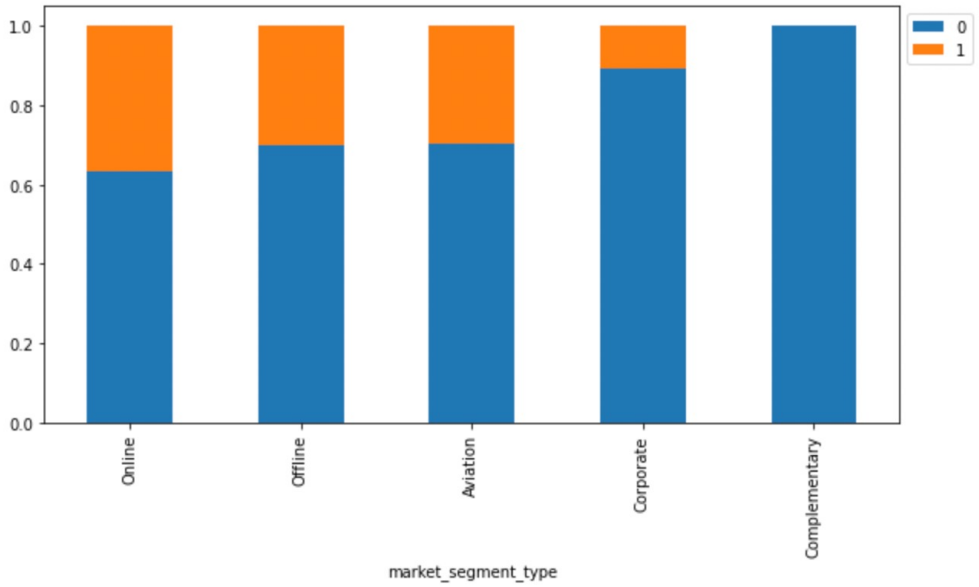
Box plots by Market Segment Type and Average Price per Room



Online has the highest average price per room with the most outliers to the right, while complementary has average price of 0 as it is right skewed with many outliers to right.

Stacked Bar Graph: Market Segment Type

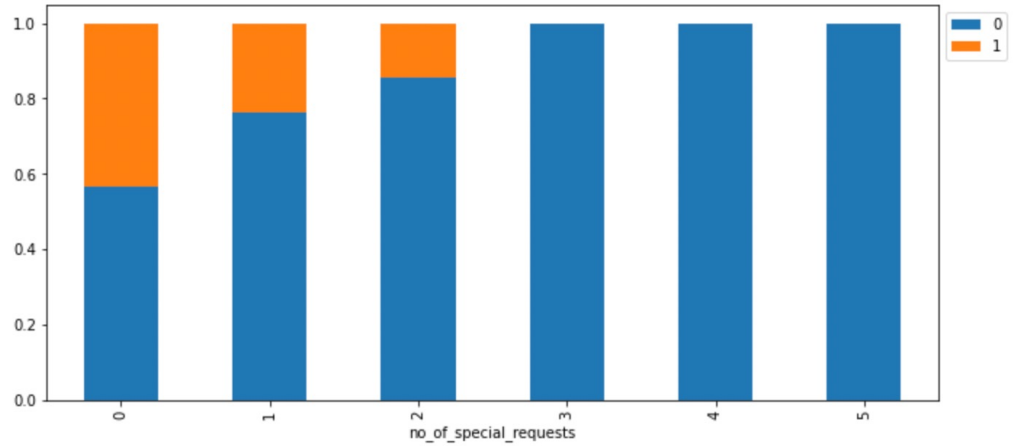
booking_status	0	1	All
market_segment_type			
All	24390	11885	36275
Online	14739	8475	23214
Offline	7375	3153	10528
Corporate	1797	220	2017
Aviation	88	37	125
Complementary	391	0	391



Most people have booked online market segment type, followed by offline, corporate, and aviation and finally no one with complementary market segment type.

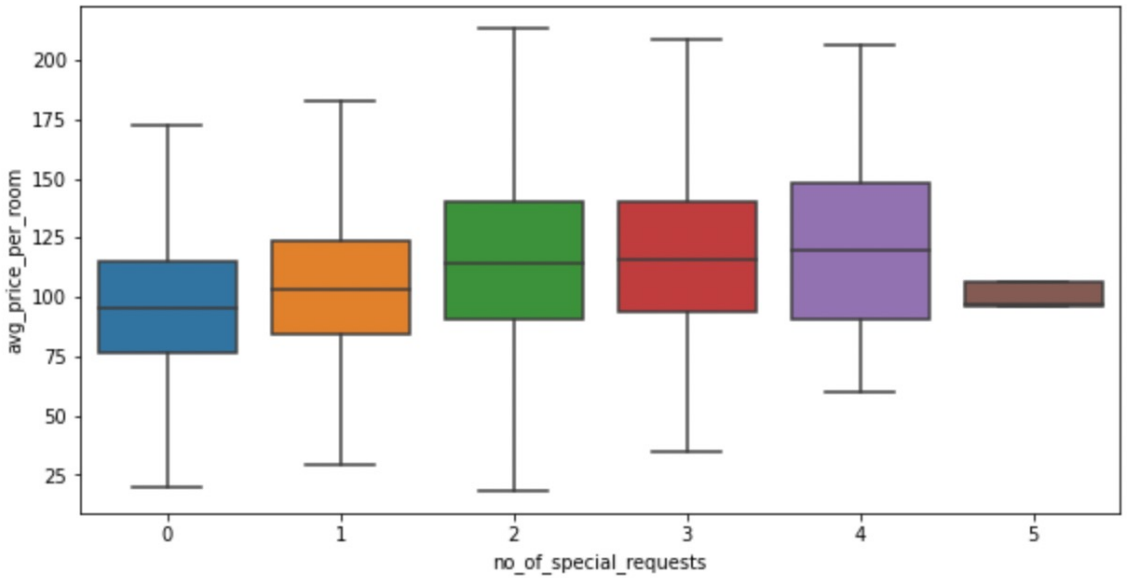
Stacked Bar Graph: Special Requests

booking_status	0	1	All
no_of_special_requests			
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



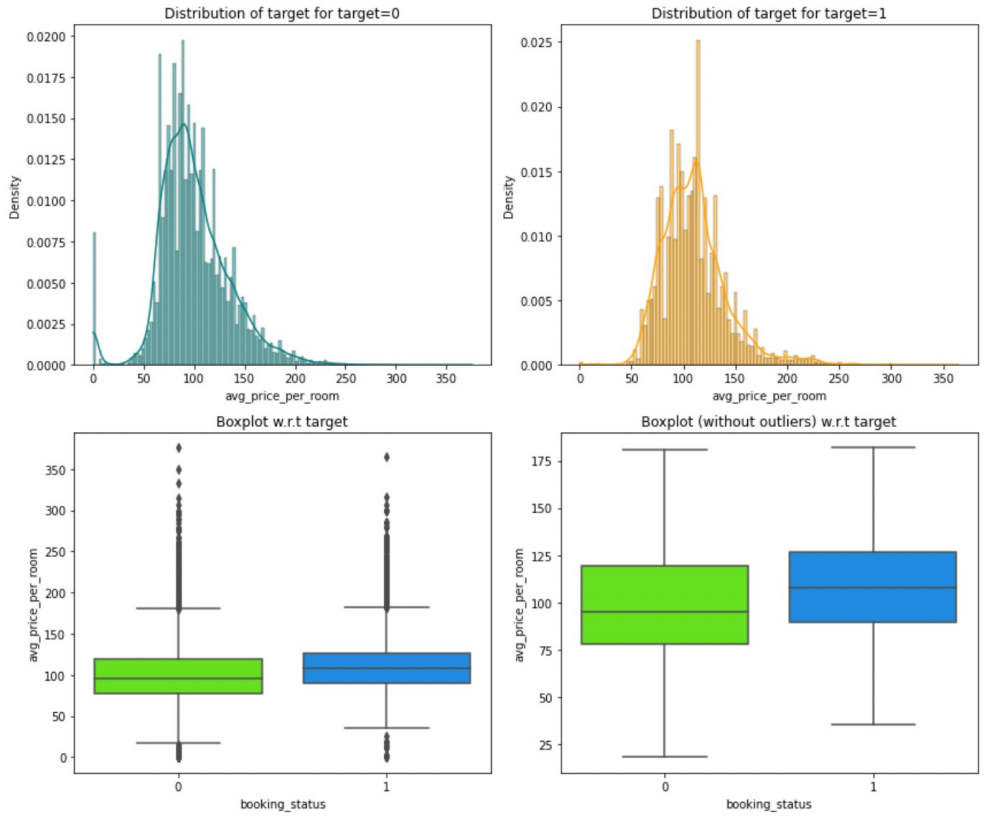
Most people who booked have 0 special requests, followed by 1 and 2 special requests.

Boxplot (Special requests vs. avg price of room)



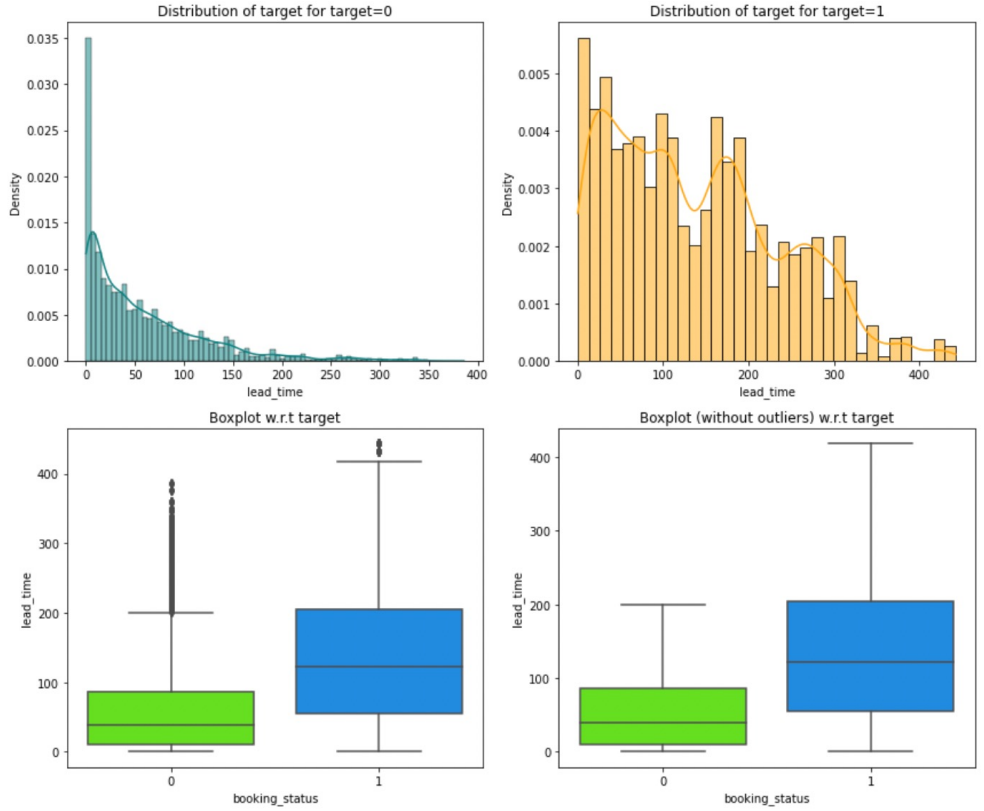
As the number of special requests increase, the average price per room increases, until it hits 5 requests.

Distribution plot (Average price per room vs. Booking Status)



Cancelled booking status rooms had a higher average price per room than noncancelled booking status rooms. A positive correlation exists between booking status and average price per room.

Distribution plot (Lead time vs. Booking Status)

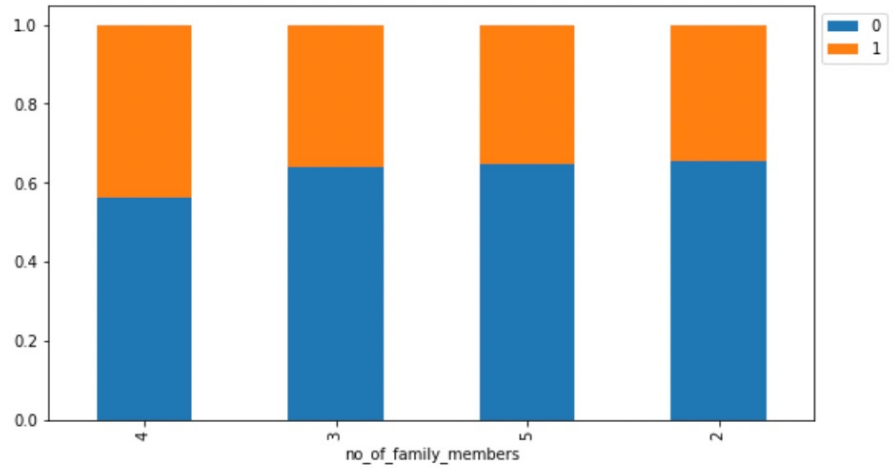


Average Lead time was greater on average for cancelled booking status rooms than noncancelled booking status rooms. A positive correlation exists between booking status and lead time.

Family Data Shape/Stacked Bar Graph

- 28,441 rows, 18 columns

booking_status	0	1	All
no_of_family_members			
All	18456	9985	28441
2	15506	8213	23719
3	2425	1368	3793
4	514	398	912
5	11	6	17

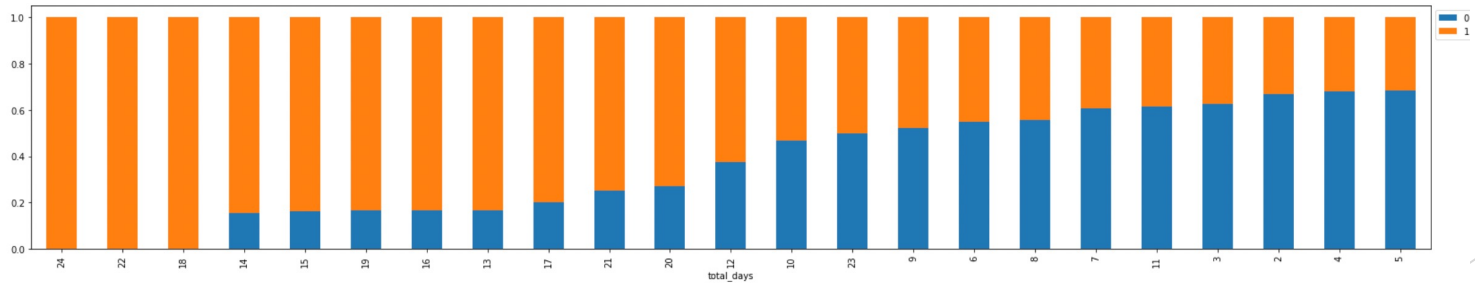


Most people who booked rooms that had cancelled status came in family members of 2, followed by 3, 4, and 5 family members.

Stay Data Shape/Stacked Bar Graph

- 17094 rows, 18 columns

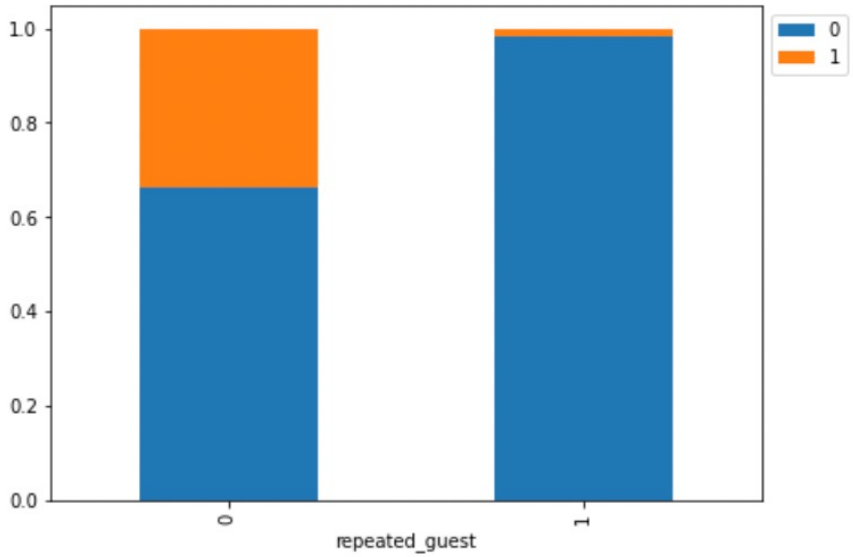
booking_status	0	1	All
total_days			
All	10979	6115	17094
3	3689	2183	5872
4	2977	1387	4364
5	1593	738	2331
2	1301	639	1940
6	566	465	1031
7	590	383	973
8	100	79	179
10	51	58	109
9	58	53	111
14	5	27	32
15	5	26	31
13	3	15	18
12	9	15	24
11	24	15	39
20	3	8	11
19	1	5	6
16	1	5	6
17	1	4	5
18	0	3	3
21	1	3	4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



The majority of people who stayed at the hotel for 3 days with cancelled booking status.

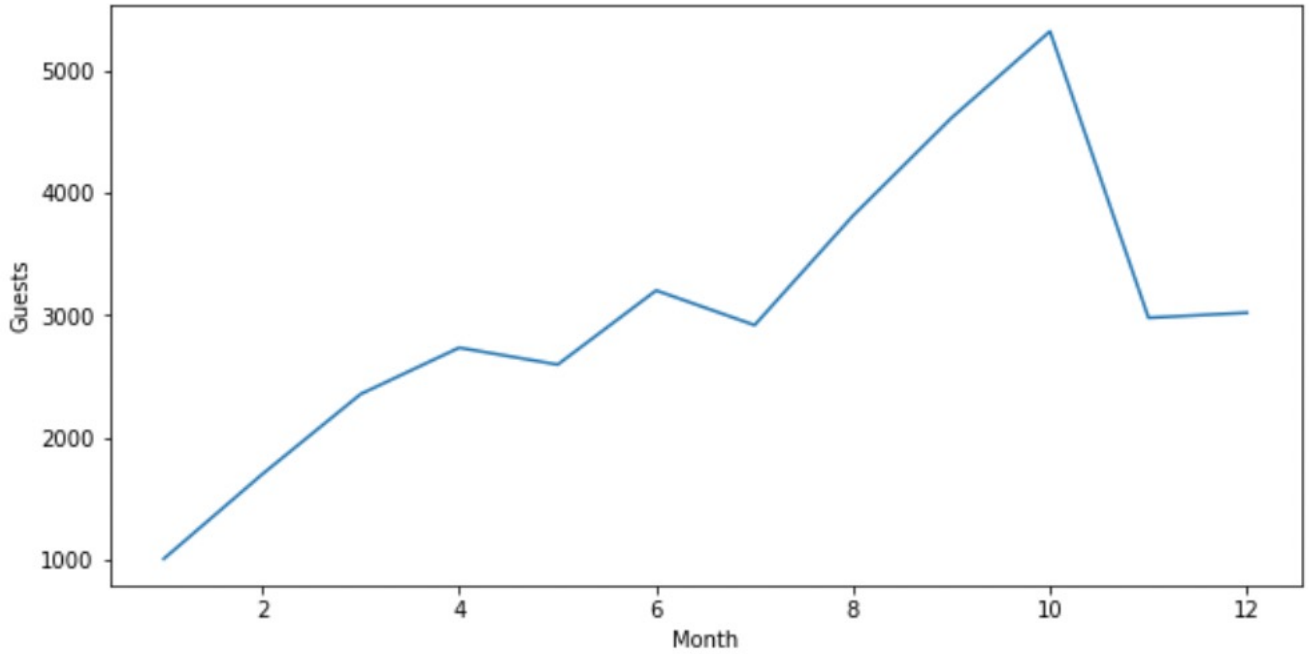
Stacked Bar Graph: Repeat Guest

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



People who weren't repeated guests had a greater cancelling booking status than people who were repeated guests. Only about 1.72% repeating guests cancelled their stay at the hotel.

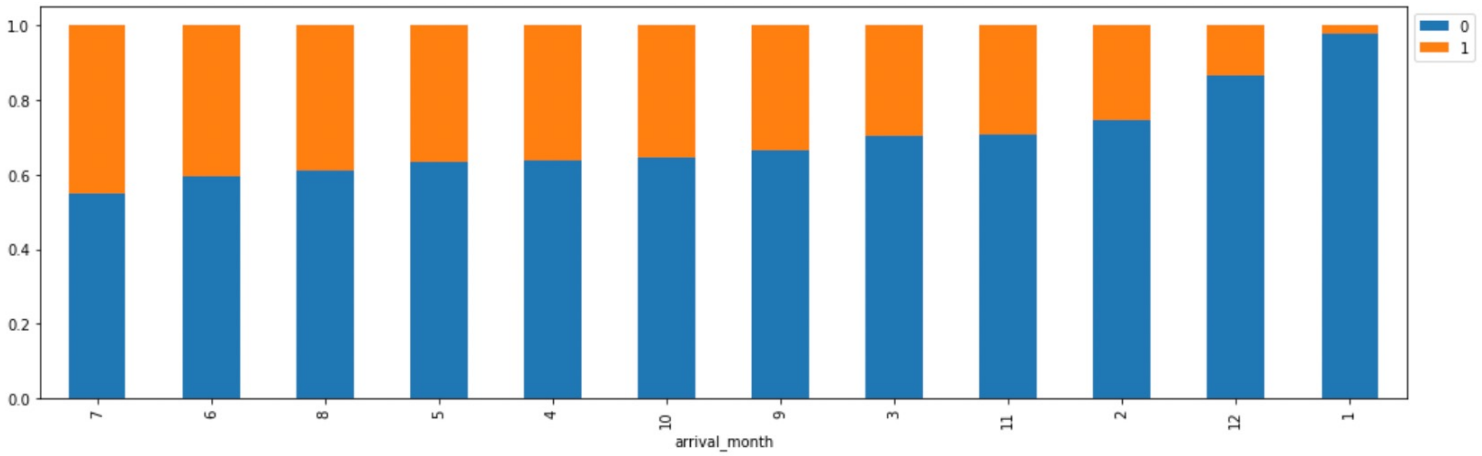
Busiest Month in Hotel



The busiest month for staying in the hotel is October as the number of guests for the month is over 5000.

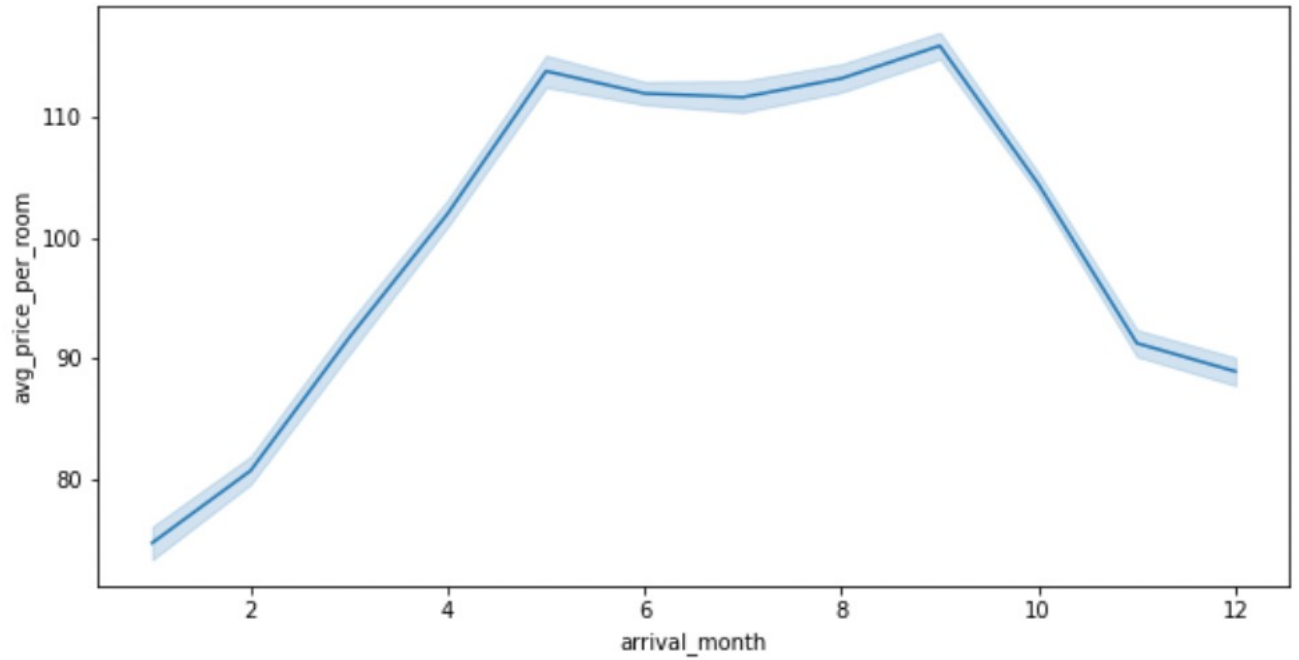
Stacked Barplot: Cancelled bookings in each month

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



October has the most cancelled bookings of 32.76%, while January has the least cancelled bookings of 2.37%.

Price Variation in Each Month



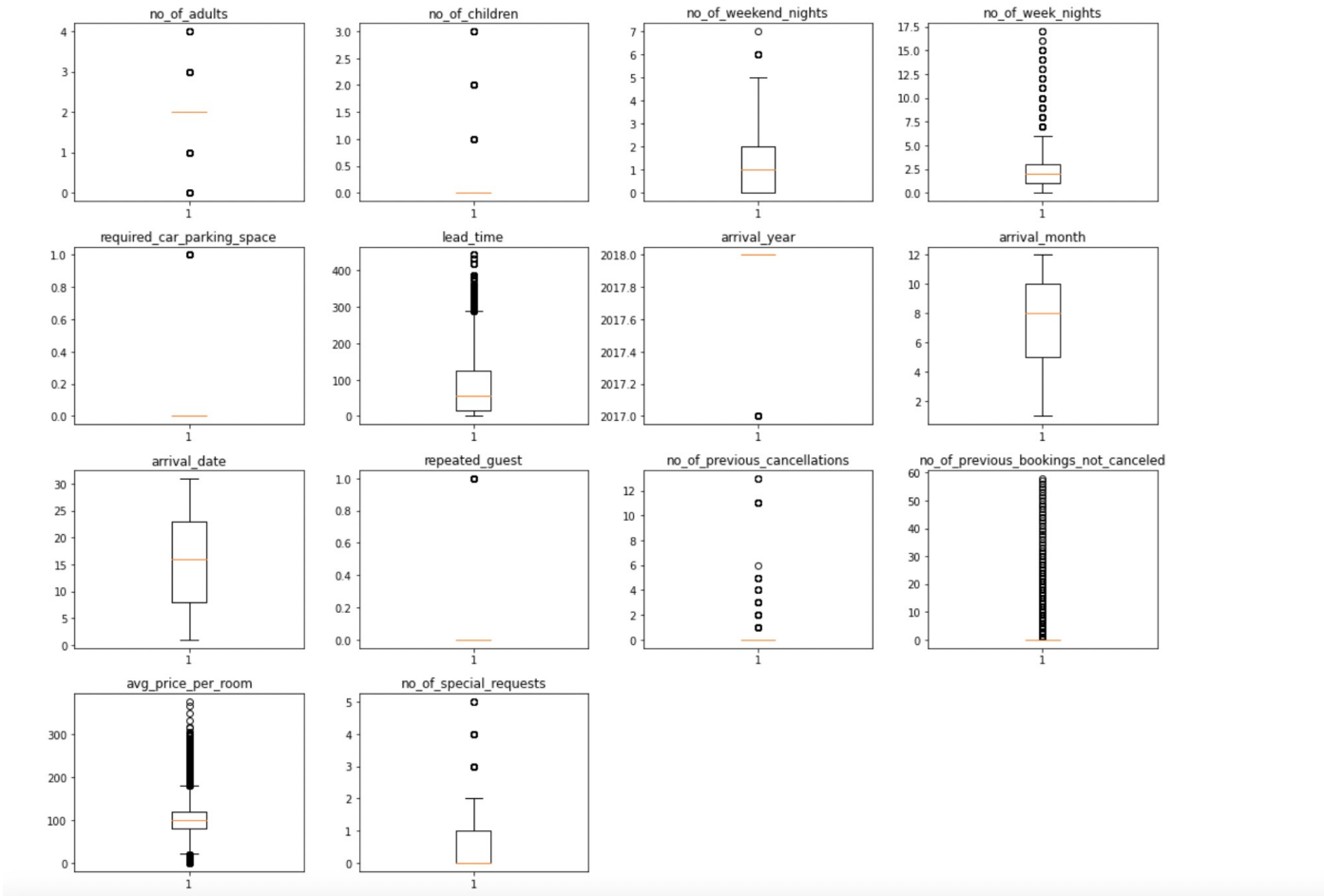
The average price per room is highest in September and lowest average price per room is in January.

Booking Status

0	24390
1	11885

More people have a non-cancelled booking status than a cancelled booking status.

Outlier Detection



Data Preparation for Modeling

```
Shape of Training set :  (25392, 27)
Shape of test set :  (10883, 27)
Percentage of classes in training set:
0    0.67064
1    0.32936
Name: booking_status, dtype: float64
Percentage of classes in test set:
0    0.67638
1    0.32362
Name: booking_status, dtype: float64
```

Logistic Regression Summary

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25364
Method:	MLE	Df Model:	27
Date:	Thu, 24 Mar 2022	Pseudo R-squ.:	0.3321
Time:	13:44:29	Log-Likelihood:	-10726.
converged:	False	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-886.9564	121.335	-7.310	0.000	-1124.769	-649.144
no_of_adults	0.0323	0.038	0.856	0.392	-0.042	0.106
no_of_children	0.0680	0.063	1.075	0.282	-0.056	0.192
no_of_weekend_nights	0.1461	0.020	7.367	0.000	0.107	0.185
no_of_week_nights	0.0354	0.012	2.881	0.004	0.011	0.059
required_car_parking_space	-1.6149	0.137	-11.772	0.000	-1.884	-1.346
lead_time	0.0158	0.000	58.942	0.000	0.015	0.016
arrival_year	0.4382	0.060	7.288	0.000	0.320	0.556
arrival_month	-0.0476	0.006	-7.336	0.000	-0.060	-0.035
arrival_date	0.0030	0.002	1.543	0.123	-0.001	0.007
repeated_guest	-1.9184	0.767	-2.502	0.012	-3.421	-0.416
no_of_previous_cancellations	0.3476	0.102	3.413	0.001	0.148	0.547
no_of_previous_bookings_not_canceled	-1.3824	0.906	-1.527	0.127	-3.157	0.393
avg_price_per_room	0.0185	0.001	24.946	0.000	0.017	0.020
no_of_special_requests	-1.4899	0.030	-48.954	0.000	-1.550	-1.430
type_of_meal_plan_Meal Plan 2	0.1732	0.067	2.586	0.010	0.042	0.305
type_of_meal_plan_Meal Plan 3	14.0425	757.204	0.019	0.985	-1470.050	1498.135
type_of_meal_plan_Not Selected	0.1986	0.053	3.724	0.000	0.094	0.303
room_type_reserved_Room_Type 2	-0.4087	0.134	-3.061	0.002	-0.670	-0.147
room_type_reserved_Room_Type 3	1.1880	1.891	0.628	0.530	-2.519	4.895
room_type_reserved_Room_Type 4	-0.2697	0.053	-5.050	0.000	-0.374	-0.165
room_type_reserved_Room_Type 5	-0.6814	0.215	-3.170	0.002	-1.103	-0.260
room_type_reserved_Room_Type 6	-0.8243	0.155	-5.317	0.000	-1.128	-0.520
room_type_reserved_Room_Type 7	-1.3507	0.298	-4.536	0.000	-1.934	-0.767
market_segment_type_Complementary	-22.7337	1556.734	-0.015	0.988	-3073.877	3028.410
market_segment_type_Corporate	-0.8518	0.276	-3.088	0.002	-1.392	-0.311
market_segment_type_Offline	-1.7631	0.264	-6.683	0.000	-2.280	-1.246
market_segment_type_Online	0.0082	0.261	0.031	0.975	-0.503	0.520

Negative values of the coefficient shows that probability of a customer booking a cancellation decreases with the increase of corresponding attribute value.

Positive values of the coefficient show that that probability of customer booking a cancellation increases with the increase of corresponding attribute value.

Some variables might contain multicollinearity.

Training Performance Metrics

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80679	0.63277	0.73985	0.68213

Multicollinearity Check (Training Performance)

	feature	VIF
0	const	39624634.31385
1	no_of_adults	1.34705
2	no_of_children	2.08820
3	no_of_weekend_nights	1.06732
4	no_of_week_nights	1.09435
5	required_car_parking_space	1.03499
6	lead_time	1.40219
7	arrival_year	1.43432
8	arrival_month	1.27809
9	arrival_date	1.00771
10	repeated_guest	1.75027
11	no_of_previous_cancellations	1.32201
12	no_of_previous_bookings_not_canceled	1.57089
13	avg_price_per_room	2.05023
14	no_of_special_requests	1.24765
15	type_of_meal_plan_Meal Plan 2	1.26322
16	type_of_meal_plan_Meal Plan 3	1.00797
17	type_of_meal_plan_Not Selected	1.27995
18	room_type_reserved_Room_Type 2	1.09653
19	room_type_reserved_Room_Type 3	1.00390
20	room_type_reserved_Room_Type 4	1.35779
21	room_type_reserved_Room_Type 5	1.03107
22	room_type_reserved_Room_Type 6	2.04832
23	room_type_reserved_Room_Type 7	1.09399
24	market_segment_type_Complementary	4.35263
25	market_segment_type_Corporate	16.63449
26	market_segment_type_Offline	62.51418
27	market_segment_type_Online	69.47432

Market segment types have high VIFS. displaying multicollinearity so we must remove them.

Dropping High P-values

['const', '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_Corporate', 'market_segment_type_Offline']

Logistic Regression Results after Removing High P-Values

Training performance accuracy, recall, precision, and F1 have decreased slightly.

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25372
Method:	MLE	Df Model:	19
Date:	Thu, 24 Mar 2022	Pseudo R-squ.:	0.3312
Time:	13:59:28	Log-Likelihood:	-10742.
converged:	True	LL-Null:	-16060.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-869.4077	120.911	-7.191	0.000	-1106.388	-632.427
no_of_weekend_nights	0.1497	0.020	7.564	0.000	0.111	0.189
no_of_week_nights	0.0362	0.012	2.950	0.003	0.012	0.060
required_car_parking_space	-1.6151	0.137	-11.783	0.000	-1.884	-1.346
lead_time	0.0159	0.000	59.906	0.000	0.015	0.016
arrival_year	0.4295	0.060	7.168	0.000	0.312	0.547
arrival_month	-0.0493	0.006	-7.614	0.000	-0.062	-0.037
repeated_guest	-3.0741	0.597	-5.151	0.000	-4.244	-1.904
no_of_previous_cancellations	0.2891	0.078	3.721	0.000	0.137	0.441
avg_price_per_room	0.0191	0.001	26.685	0.000	0.018	0.020
no_of_special_requests	-1.4855	0.030	-49.274	0.000	-1.545	-1.426
type_of_meal_plan_Meal Plan 2	0.1657	0.067	2.475	0.013	0.035	0.297
type_of_meal_plan_Not Selected	0.2100	0.053	3.977	0.000	0.106	0.313
room_type_reserved_Room_Type 2	-0.3733	0.129	-2.890	0.004	-0.627	-0.120
room_type_reserved_Room_Type 4	-0.2685	0.052	-5.203	0.000	-0.370	-0.167
room_type_reserved_Room_Type 5	-0.6889	0.214	-3.216	0.001	-1.109	-0.269
room_type_reserved_Room_Type 6	-0.7440	0.120	-6.203	0.000	-0.979	-0.509
room_type_reserved_Room_Type 7	-1.3177	0.292	-4.512	0.000	-1.890	-0.745
market_segment_type_Corporate	-0.8710	0.103	-8.444	0.000	-1.073	-0.669
market_segment_type_Offline	-1.7713	0.052	-34.275	0.000	-1.873	-1.670

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80663	0.63265	0.73950	0.68191

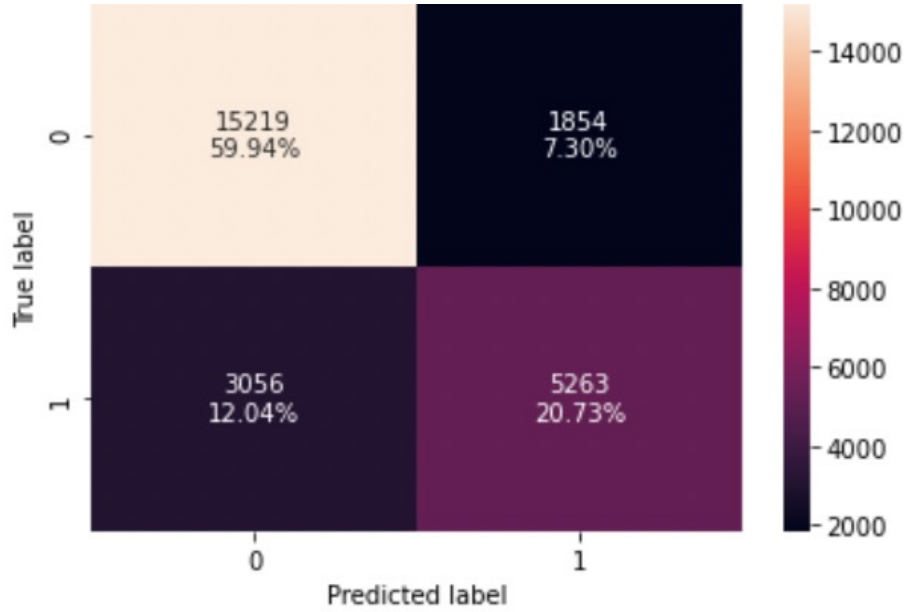
Coefficients to Odds

	const	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
Odds	0.00000	1.16153	1.03682	0.19886	1.01601	1.53653	0.95191	0.04623	1.33523	1.01925	0.22638
Change_odd%	-100.00000	16.15325	3.68173	-80.11371	1.60134	53.65324	-4.80895	-95.37690	33.52334	1.92531	-77.36171

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_Corporate
1.18026	1.23366	0.68845	0.76451	0.50213	0.47521	0.26776	0.41854
18.02564	23.36580	-31.15506	-23.54886	-49.78672	-52.47870	-73.22410	-58.14571

market_segment_type_Offline
0.17012
-82.98809

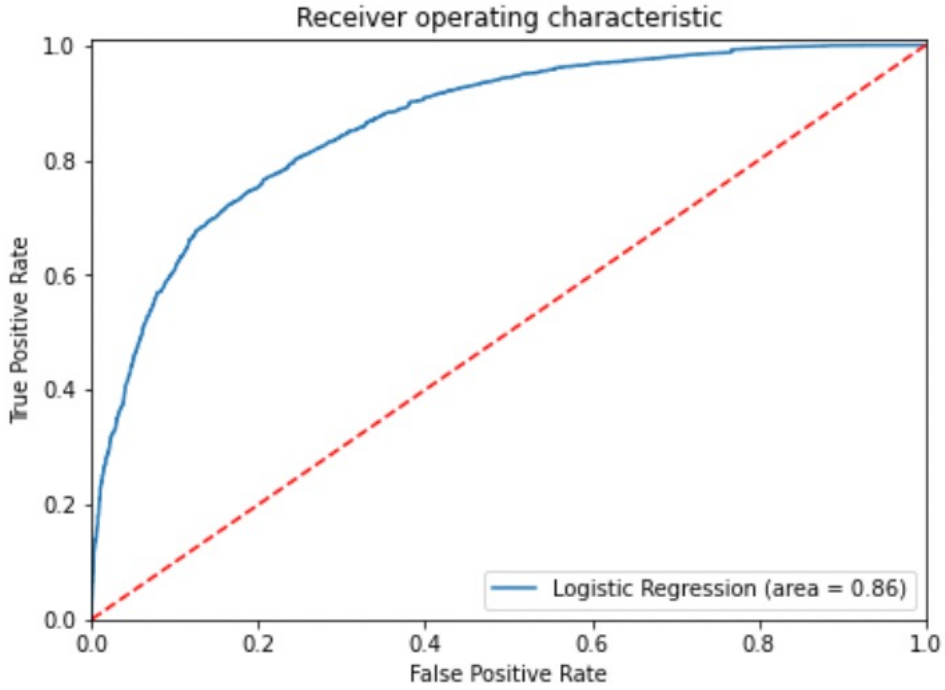
Checking Model Performance on Training Set



Training performance:

	Accuracy	Recall	Precision	F1
0	0.80663	0.63265	0.73950	0.68191

ROC-AUC on training set

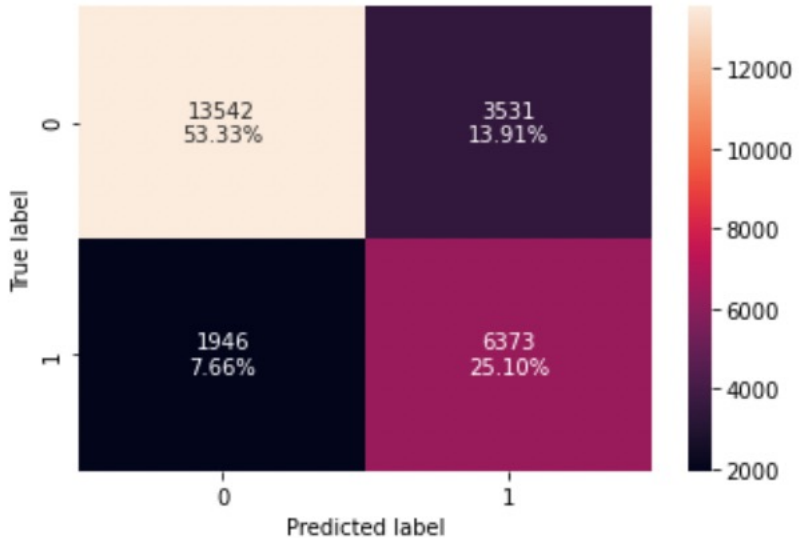


Logistic Regression model is giving a good performance on training set.

Optimal Threshold using ROC-AUC curve

- 0.3330753336833494

Model Performance Improvement on Training Set

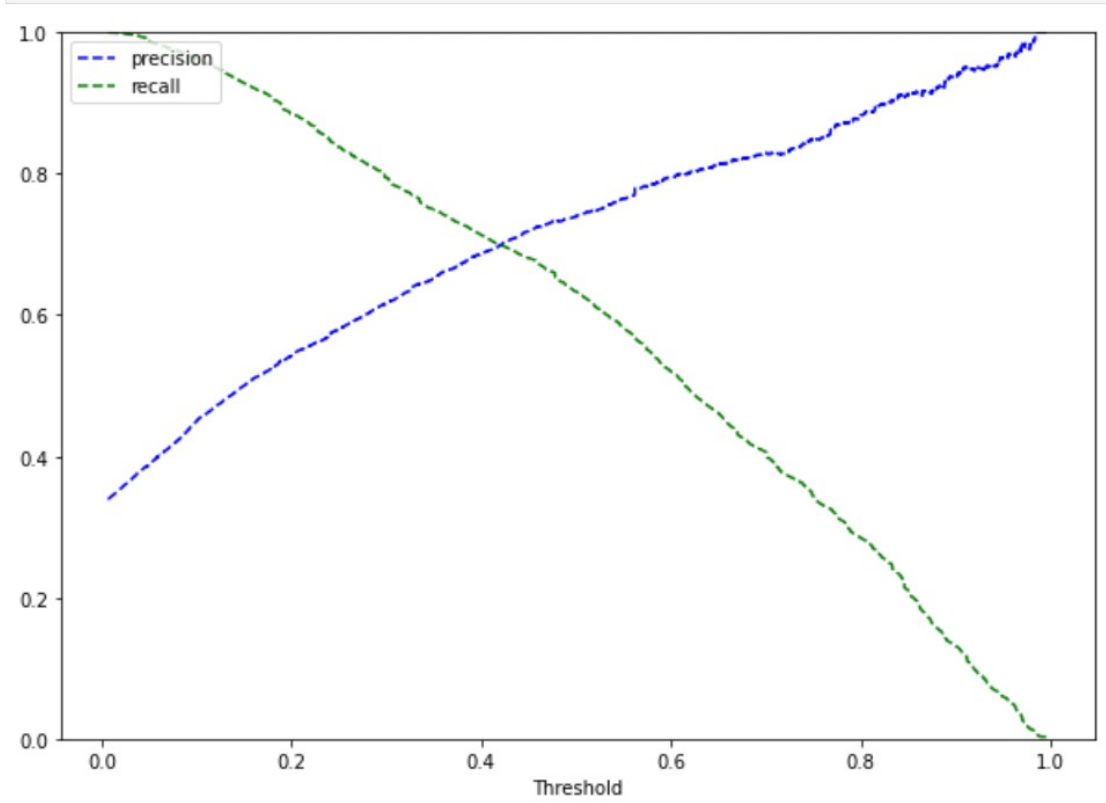


Training performance:

	Accuracy	Recall	Precision	F1
0	0.78430	0.76608	0.64348	0.69945

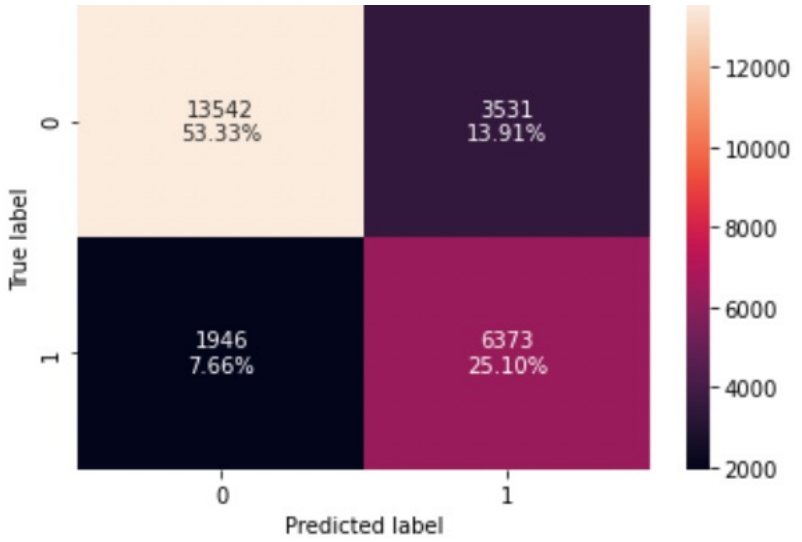
Model performance improved significantly. It's giving a higher recall of 0.76608 compared to 0.63265. Precision decreased from 0.7395 to 0.64348.

Precision-Recall Curve for Better Threshold



At threshold of 0.42, we get balanced recall and precision.

Checking Model Performance on Training Set

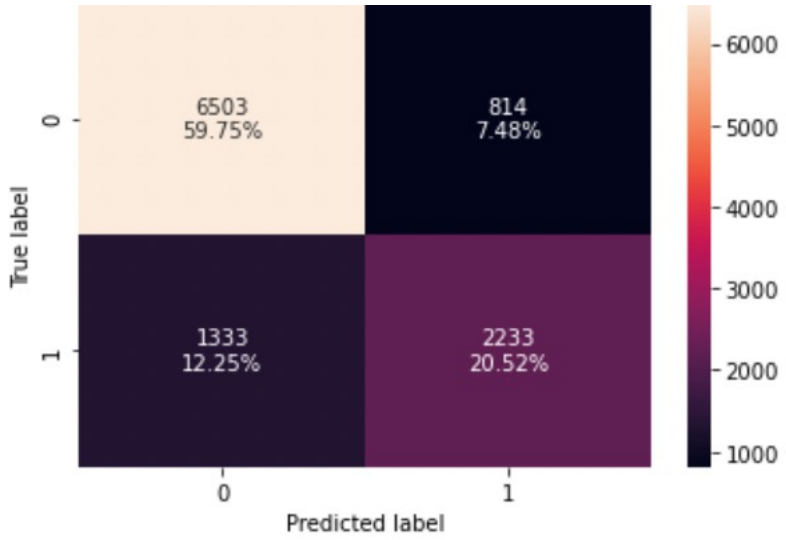


Training performance:

	Accuracy	Recall	Precision	F1
0	0.80265	0.69876	0.69885	0.69880

Model is performing well on training set from an accuracy standpoint.

Checking Model Performance on Test Set

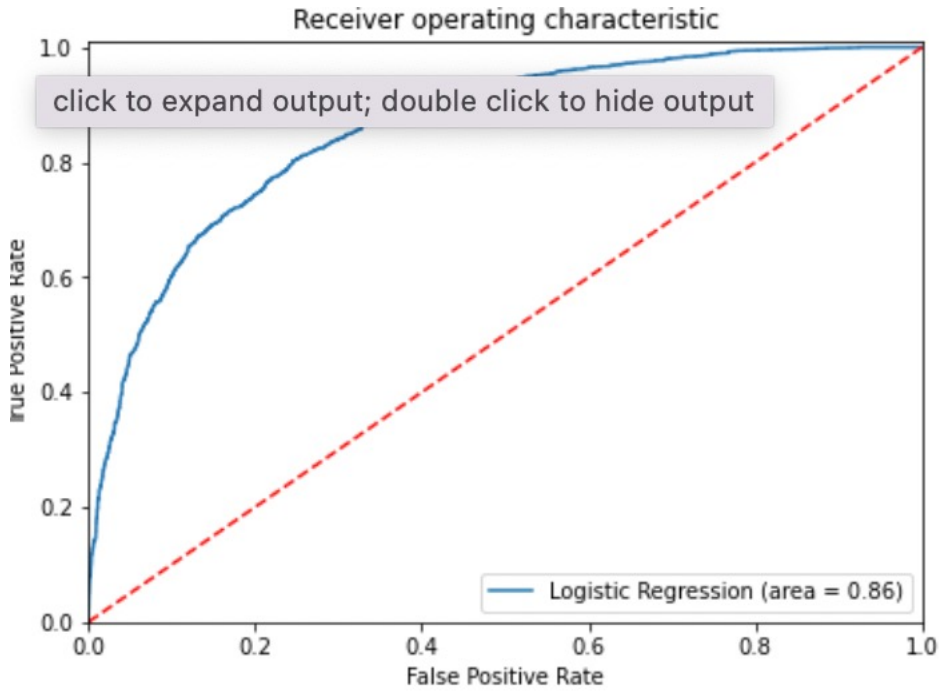


Test performance:

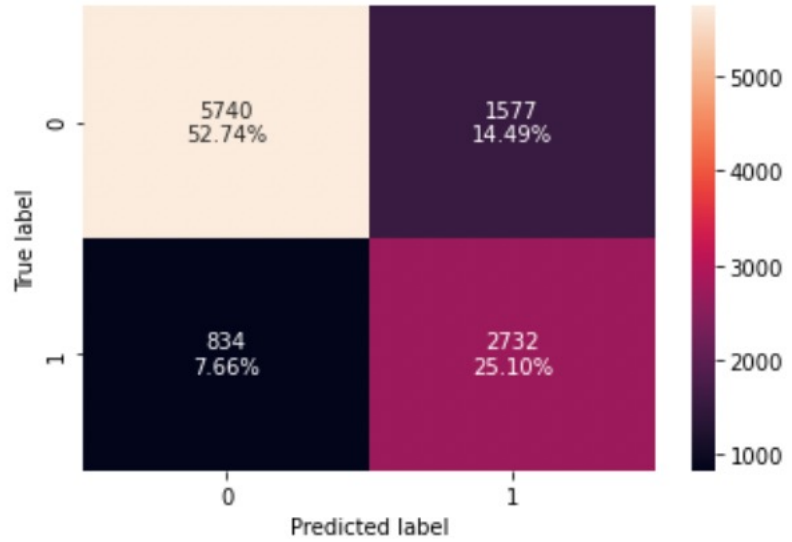
	Accuracy	Recall	Precision	F1
0	0.80272	0.62619	0.73285	0.67534

Model with default threshold

ROC Curve on Test Set



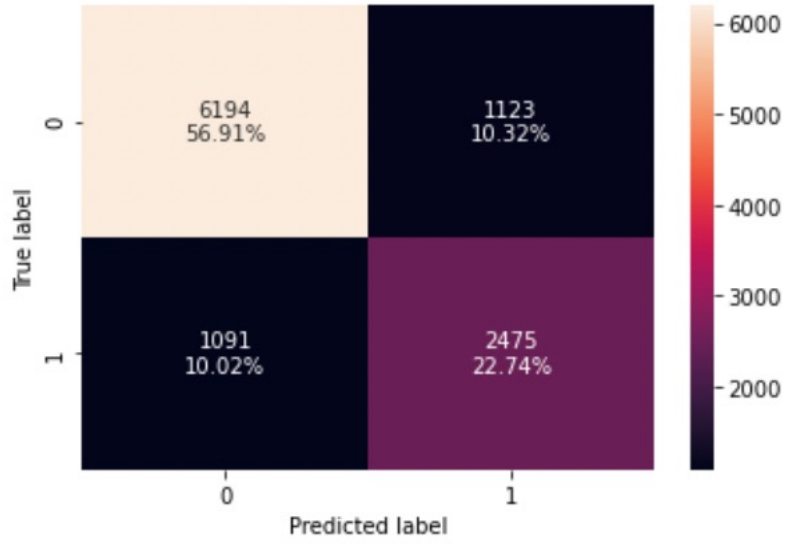
Model with Threshold 0.37



Test performance:

	Accuracy	Recall	Precision	F1
0	0.77846	0.76612	0.63402	0.69384

Model with Threshold 0.42



Test performance:

	Accuracy	Recall	Precision	F1
0	0.79656	0.69405	0.68788	0.69095

Model Performance Summary

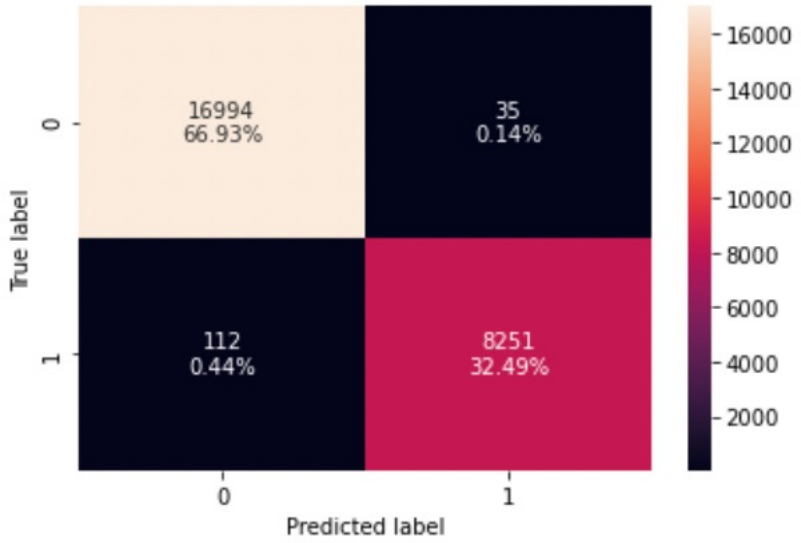
Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80663	0.78430	0.80265
Recall	0.63265	0.76608	0.69876
Precision	0.73950	0.64348	0.69885
F1	0.68191	0.69945	0.69880

Test set performance comparison:

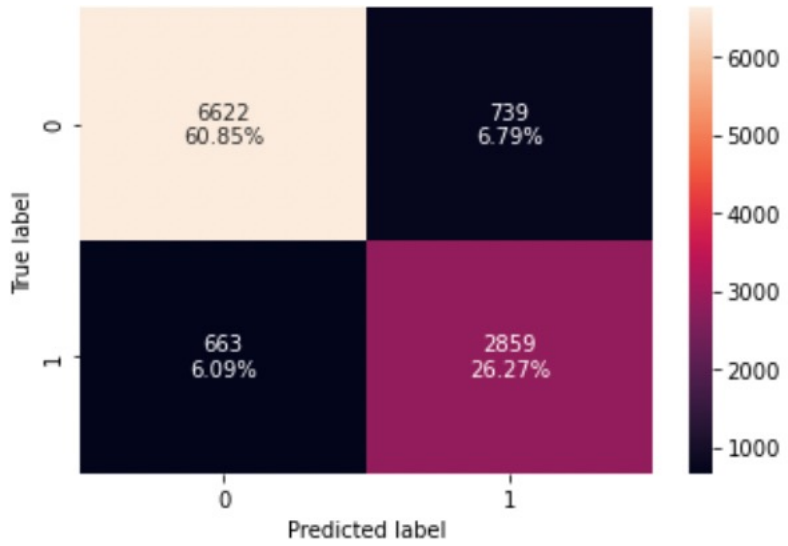
	Logistic Regression statsmodel	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80272	0.77846	0.79656
Recall	0.62619	0.76612	0.69405
Precision	0.73285	0.63402	0.68788
F1	0.67534	0.69384	0.69095

Decision Tree: Checking Model Performance on Training Set



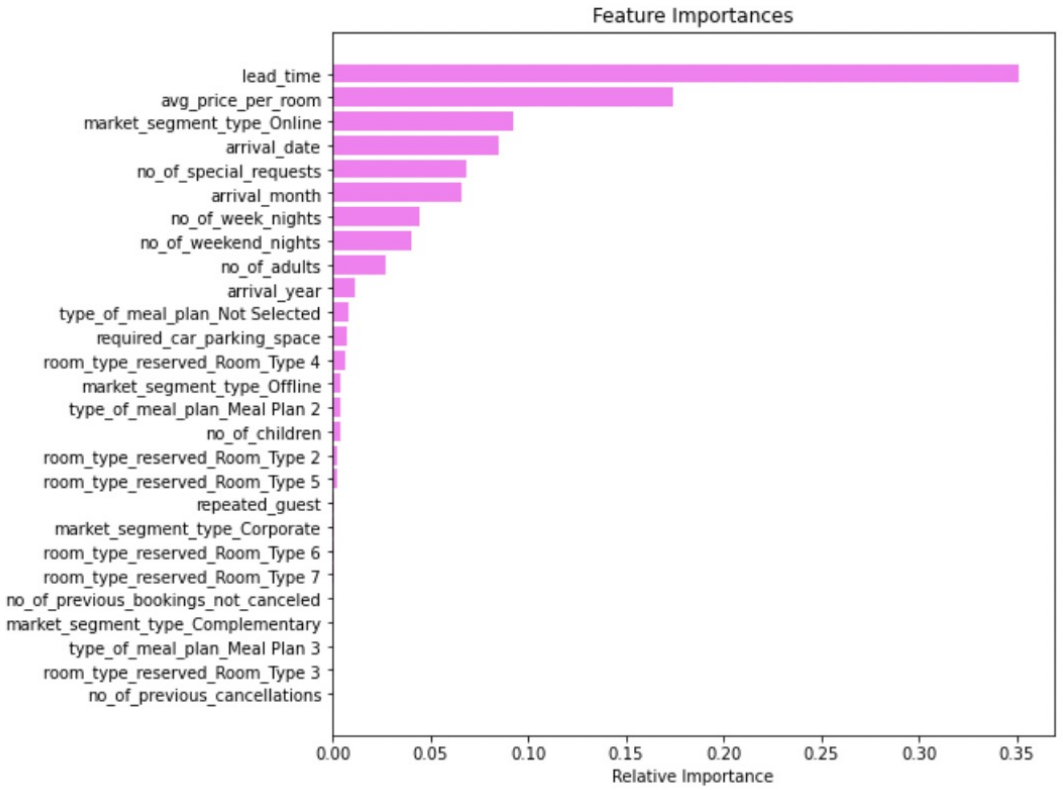
	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

Decision Tree: Checking Model Performance on Test Set



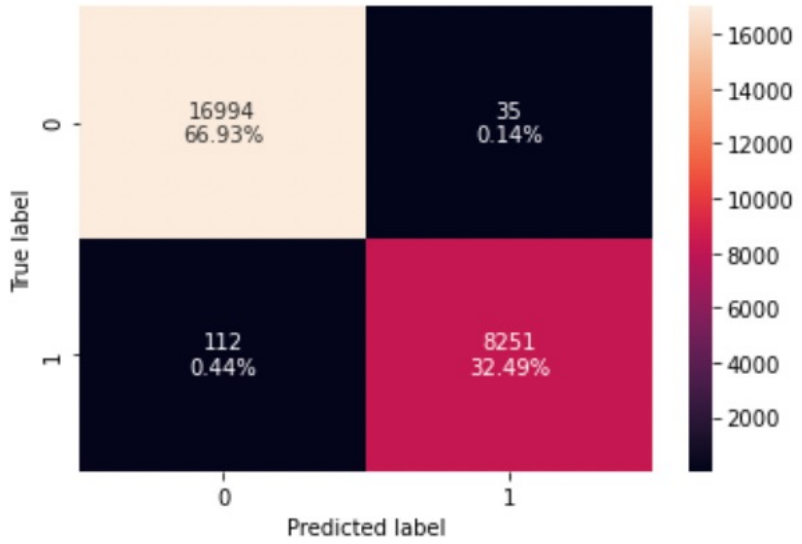
	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

Feature Importances



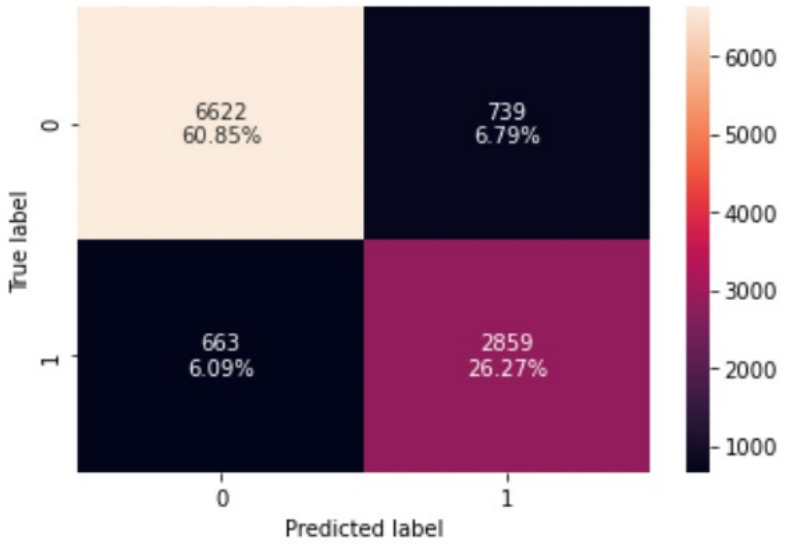
According to decision tree model, lead time is most important variable for predicting booking status.

Pre-Pruning: Checking Model Performance on Training Set



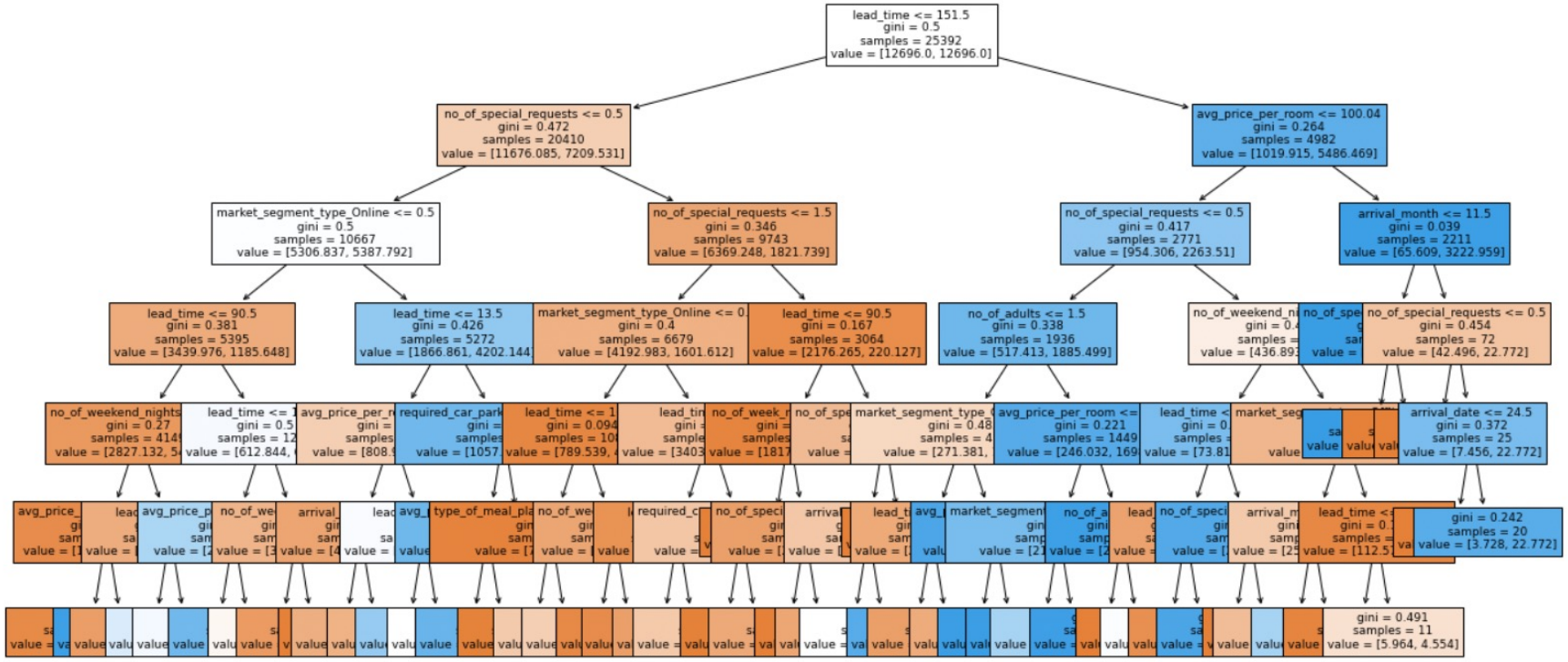
	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

Pre-Pruning: Checking Model Performance on Test Set

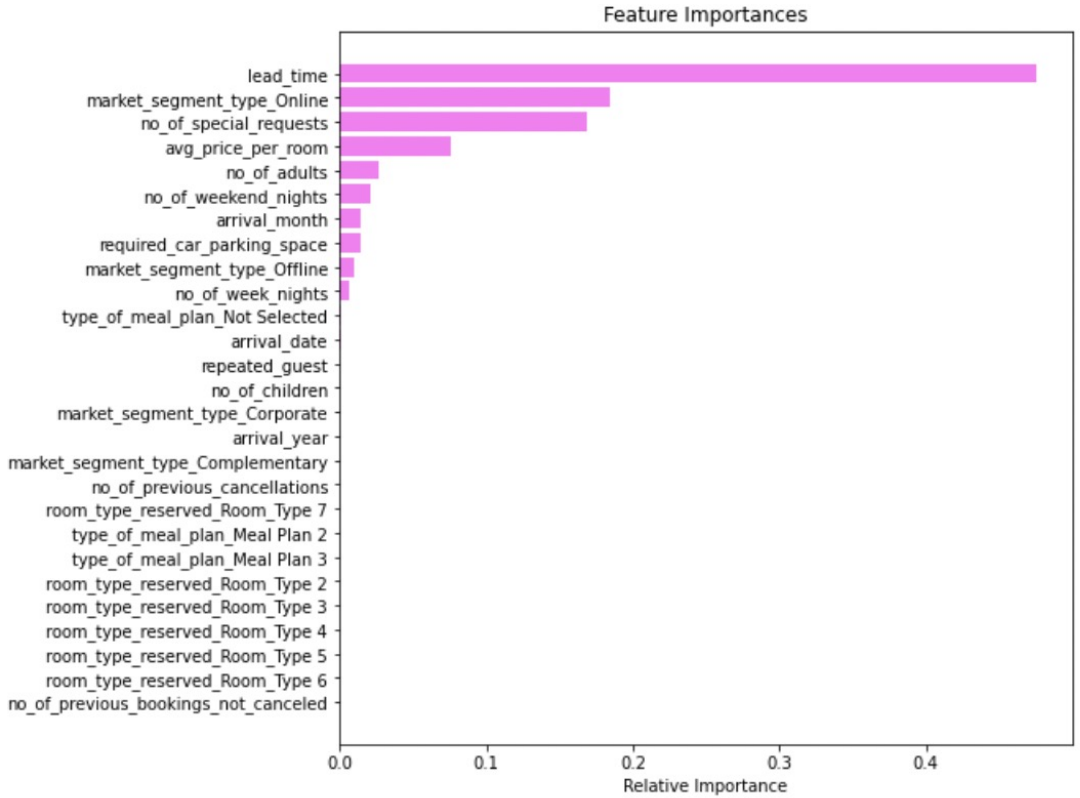


	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

Visualizing Decision Tree



Importance of Features in Tree Building



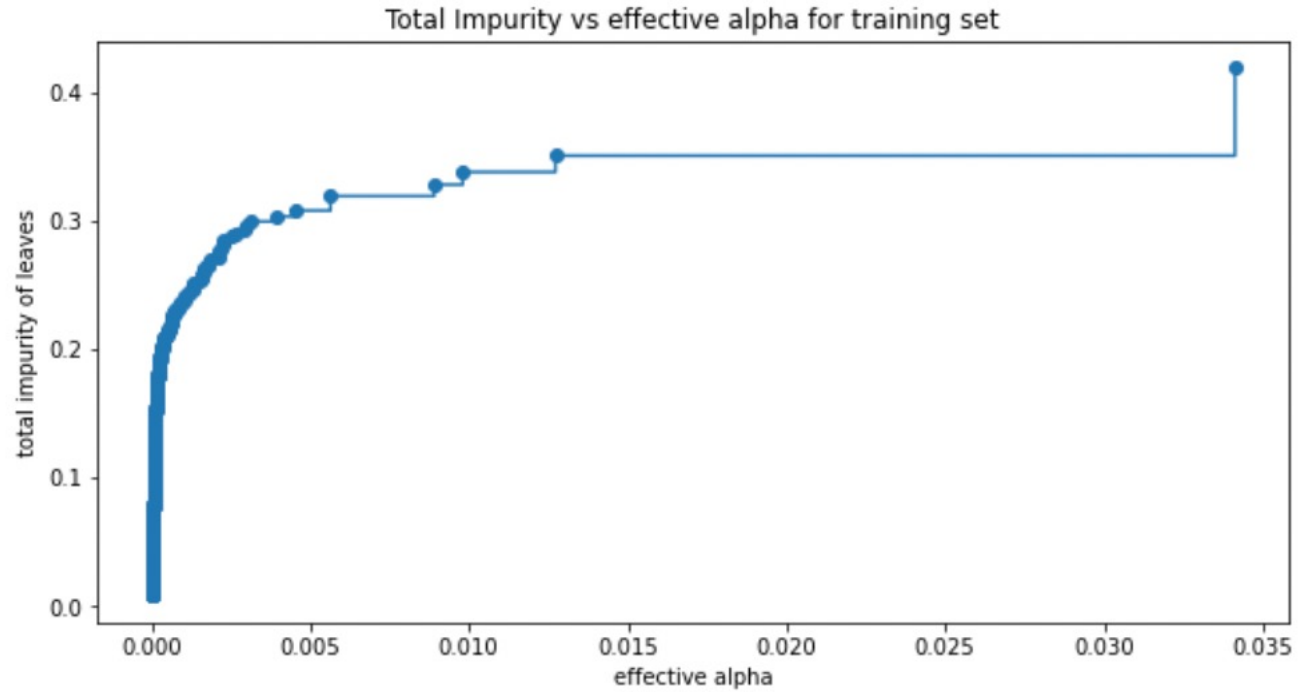
Some important features were lost, bust post-pruning might do a better job.

Cost Complexity Pruning

	ccp_alphas	impurities
0	0.00000	0.00838
1	0.00000	0.00838
2	0.00000	0.00838
3	0.00000	0.00838
4	0.00000	0.00838
...	...	...
1889	0.00890	0.32806
1890	0.00980	0.33786
1891	0.01272	0.35058
1892	0.03412	0.41882
1893	0.08118	0.50000

1894 rows × 2 columns

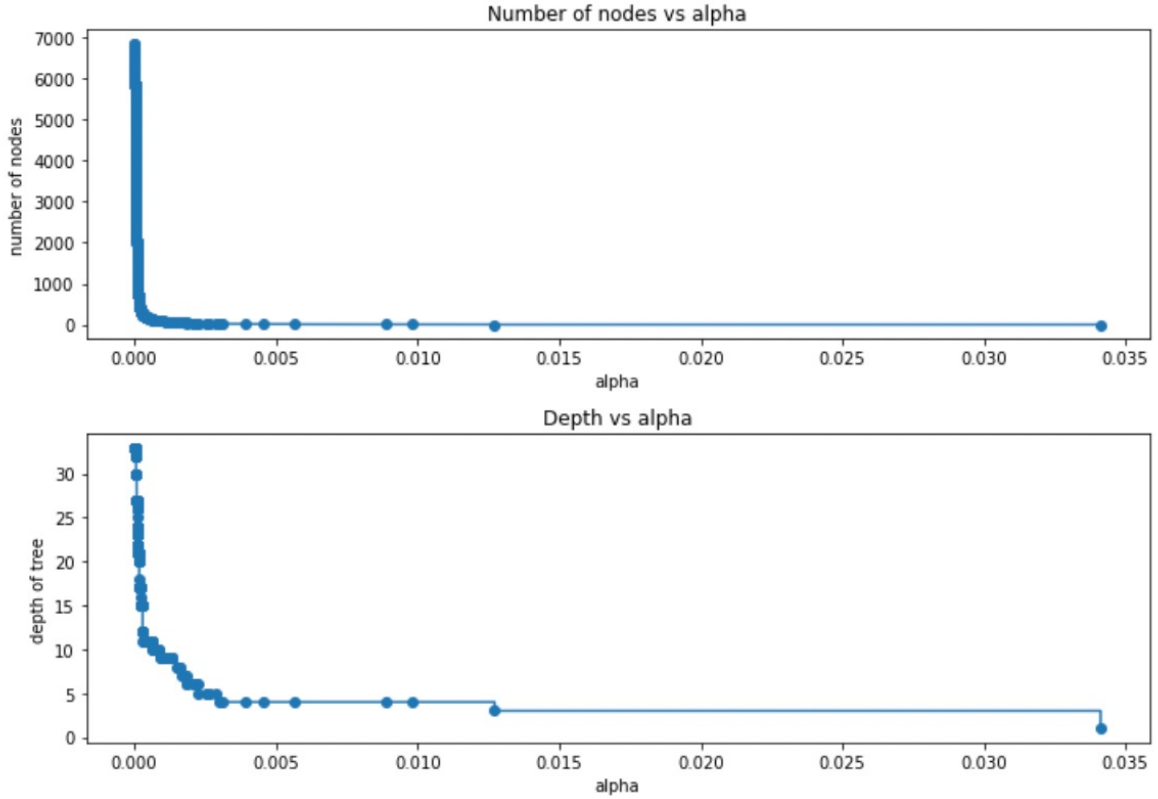
Total Impurity vs. Effective Alpha for Training Set



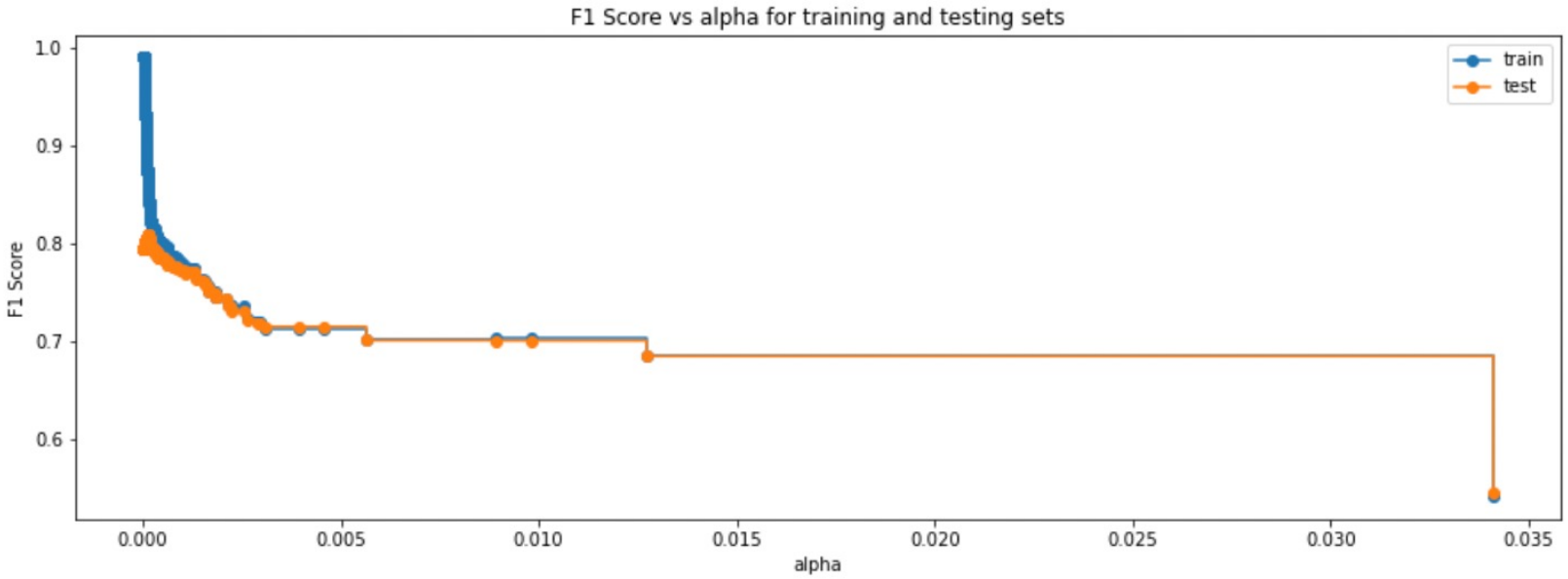
Training Decision Tree with Effective Alphas

```
Number of nodes in the last tree is: 1 with ccp_alpha: 0.08117914389136954
```

Number of nodes and depth vs. alpha

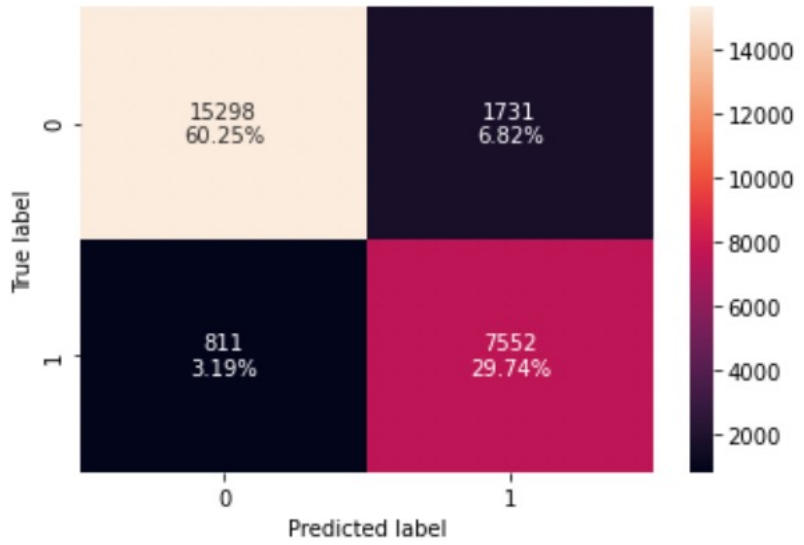


F1 Score vs alpha for training and testing sets



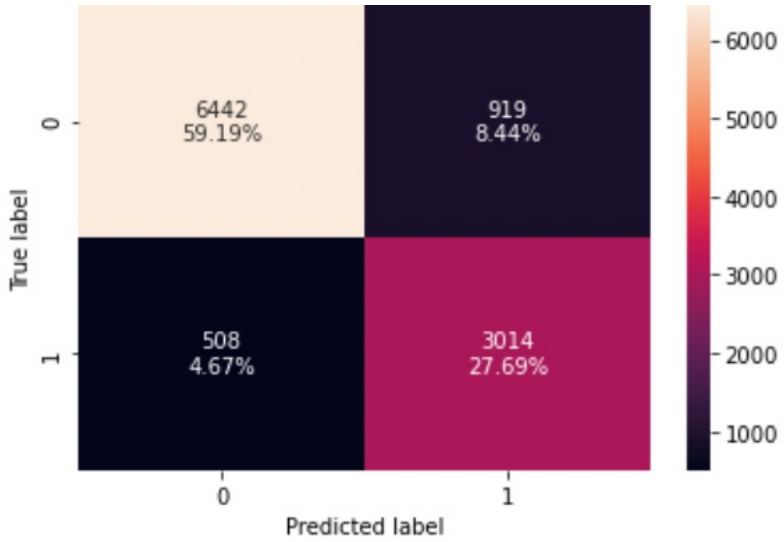

```
DecisionTreeClassifier(ccp_alpha=0.0001226763315516706, class_weight='balanced',
                       random_state=1)
```

Checking Performance on Training Set



	Accuracy	Recall	Precision	F1
0	0.89989	0.90303	0.81353	0.85594

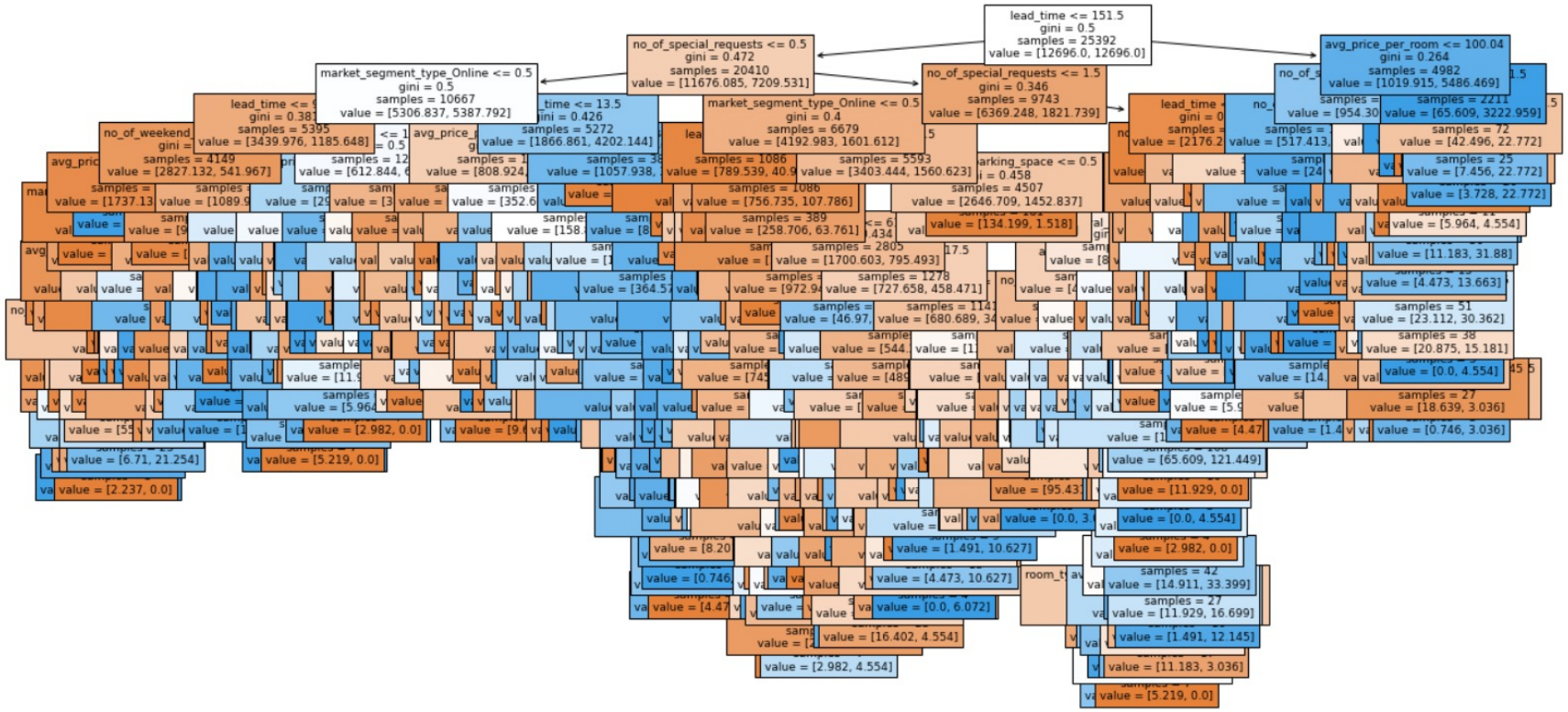
Checking Performance on Test Set



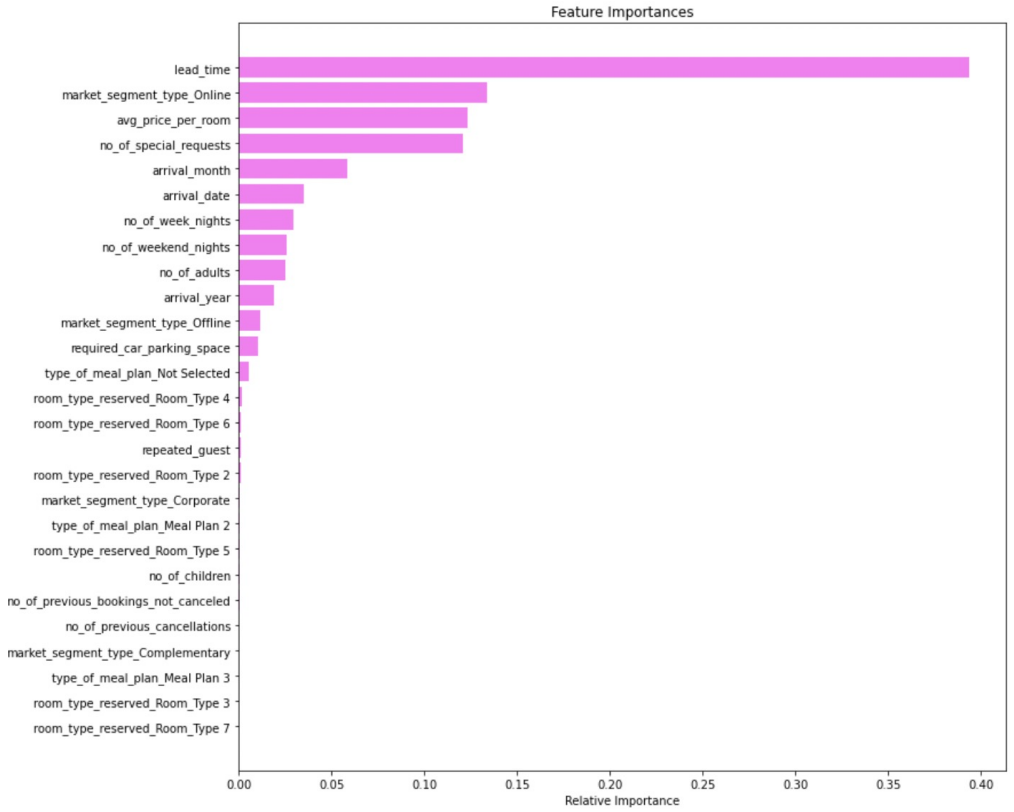
	Accuracy	Recall	Precision	F1
0	0.86888	0.85576	0.76634	0.80858

With test set, we get highest recall.

Visualizing Decision Tree Post-Pruning



Features Importance Post-Pruning



Comparing All Decision Tree Models

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.99421	0.89989
Recall	0.98661	0.98661	0.90303
Precision	0.99578	0.99578	0.81353
F1	0.99117	0.99117	0.85594

Testing performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.87118	0.86888
Recall	0.81175	0.81175	0.85576
Precision	0.79461	0.79461	0.76634
F1	0.80309	0.80309	0.80858

Conclusions

- The models are giving a generalized performance on the training and test sets.
- The models built with machine learning can be used to predict if a customer will book a hotel cancellation.
- Highest recall is 0.98661 on training set.
- Model with default threshold gives a lower recall, but higher precision, suggesting the resources it will save, but will not be able to retain customers in hotel bookings.
- When using decision trees, pre-pruning showed a lower recall, but higher precision than post-pruning.
- The importance of using pruning was to reduce overfitting.
- Lead time is shown to be the most significant variable in determining if a customer will book a hotel cancellation.
- With decision trees, easy interpretation is conducted through visualizing decision trees and their confusion matrices.

Recommendations

- Have price deals/offers based on booking date and arrival date for special occasions such as holiday events, which could influence customer not cancelling hotel stays.
- Conduct advertisements to promote hotel room characteristics and benefits.
- Differentiate price deals and activities planning for weeknights and weekend nights.
- Have most hotel booking transactions performed online as it is most dominant market segment type.