INN Hotels Project 4

WALTER EFIRD

Table of Contents

Overview

- INN Hotels Group is looking for a solution to the high number of booking cancellations, as new technologies regarding online bookings have led to an increase in customer cancellations.
- Booking cancellations negatively impact the group in several ways:
  - Loss of resources
  - Additional costs of distribution channels by increases in commissions
  - Having to lower prices last minute in order to resell a room
  - Increase in Human Resources to make arrangements for guests

Problem Statement

- Due to the increasing number of cancellations, INN Hotels group has requested a Machine Learning based solution be built to help predict which bookings are most likely to be canceled.
- <u>Objectives</u>
  - To analyze the data provided to discover which factors have a high influence on booking cancellations
  - To build a predictive model that can predict which bookings are going to be cancelled in advance
  - To help in creating profitable policies for cancellations and refunds

Data Dictionary

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

Observations and Statistical Summary

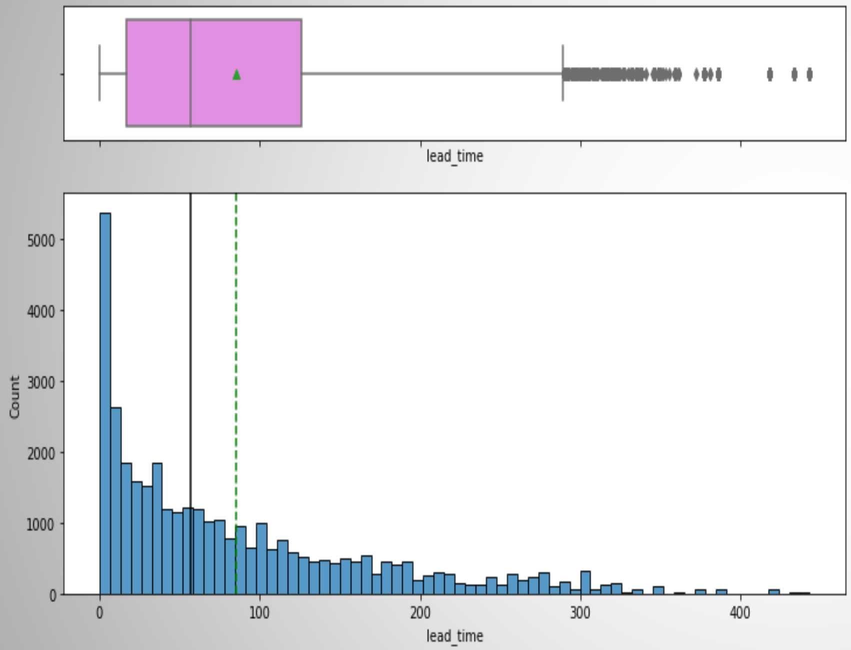
- 36,275 rows and 19 columns
- Datatypes
  - 13 integer datatypes
  - 5 object datatypes
  - 1 float64 datatype
- No missing values
- No duplicated values
- Mean number of adults is 1.85
- Mean number of children is 0.10
- Mean number of weekend nights is 0.81
- Mean number of week nights is 2.20
- Mean number of required parking space is 0.03
- Average lead time is 85.23 days
- The data is from the years 2017 and 2018
- The average month is the middle of July
- The average date is approximately the 15^{th} of each month
- The average repeated guests are 0.02
- Average number of previous cancellations is 0.02
- The mean no of previous bookings not cancelled is 0.15
- Average price per room is $103.42
- Average number of special requests is 0.62

Statistical Summary cont.

- The max number of adults is 4
- The highest number of children is 10 with the lowest being 0
- Longest stay was 17 days
- The longest lead time was 443 days
- The highest room price was $540.00
- The highest number of special requests was 5

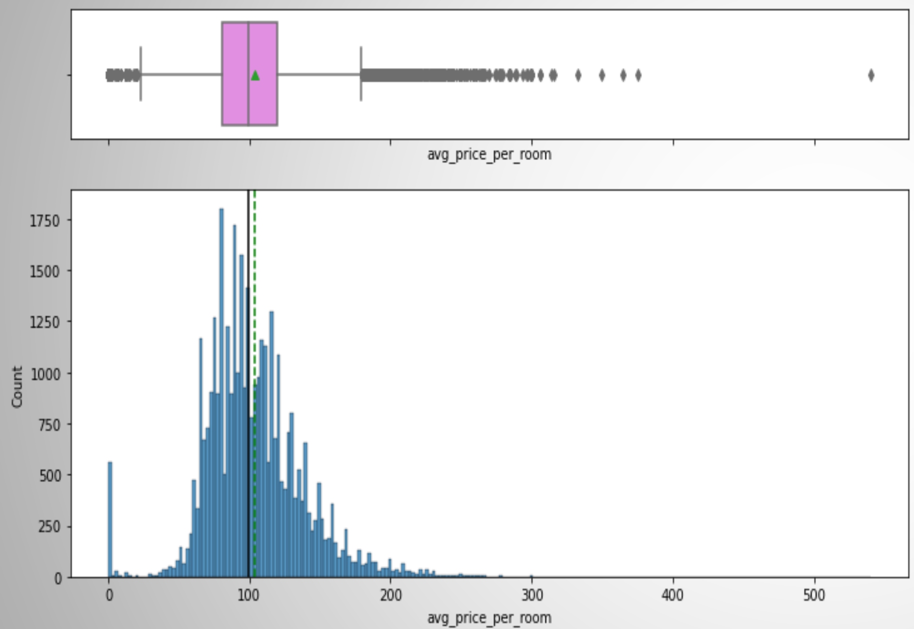
EDA

Lead Time



- The avg lead time was about 85 days
- The median was approximately 60 days
- The feature is heavily right-skewed
- All outliers lie passed the upper whisker

Average price per room



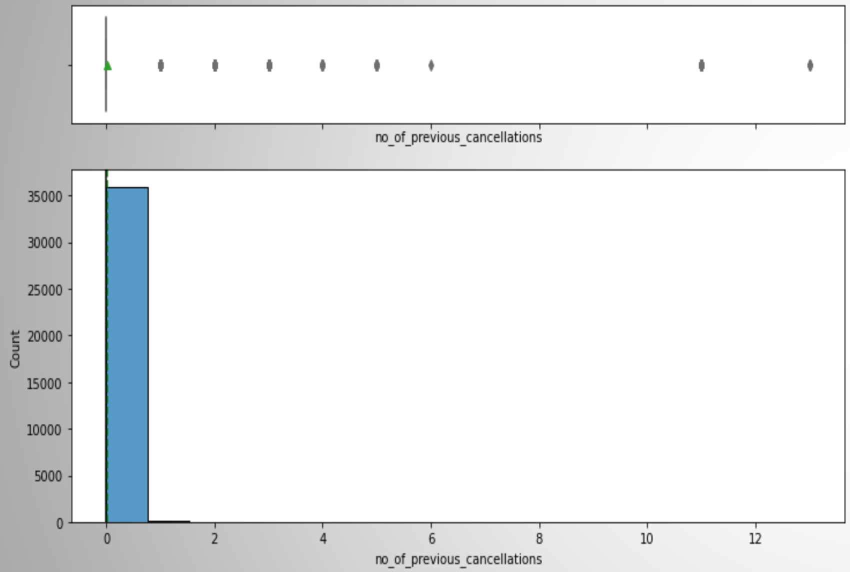
- Average price of a room was $103.00
- The most expensive room was $540.00
- The feature is right-skewed
- Outliers exist on both ends of the tails
- Several rooms had a price of $0, which were not null values
- These were either rooms booked online or were complimentary
- The top 25% of avg price of room ranges from $120.00 - $179.55
- All outliers for the bottom 25% fall below $80.30 while all outliers for the top 25% fall above $179.55

$0 rooms by Market Segment

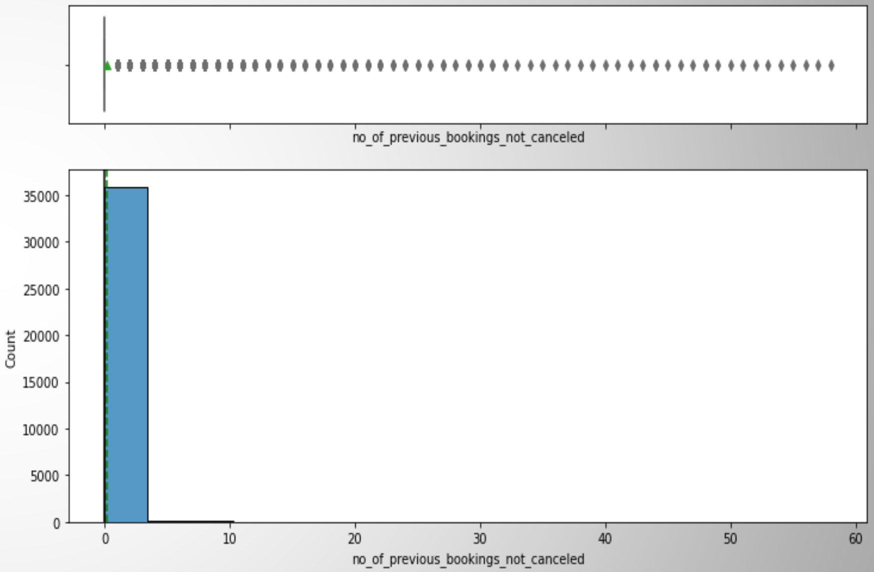
Complementary **354**

Online **191**

Booking Cancellations



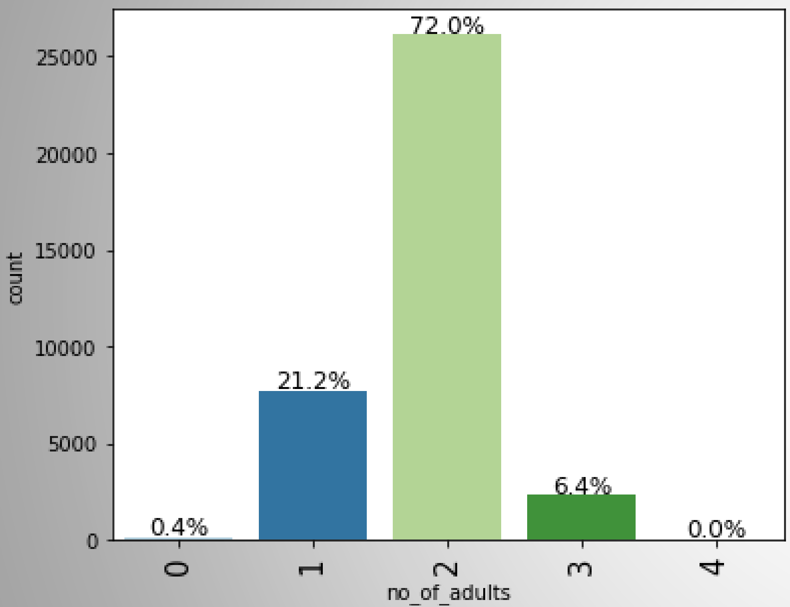
Number of previous cancellations



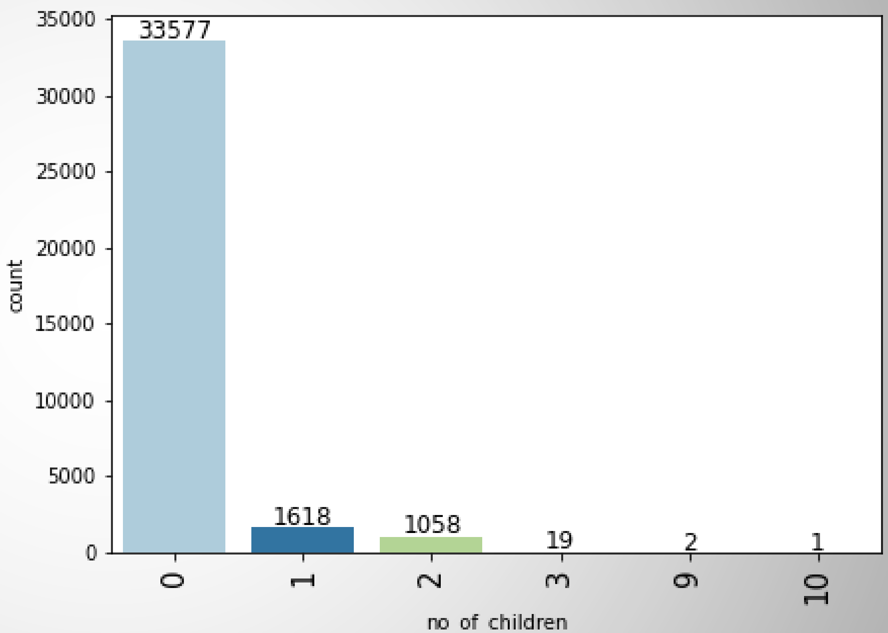
Number of previous bookings not cancelled

Number of Adults and Children

- No. of Adults

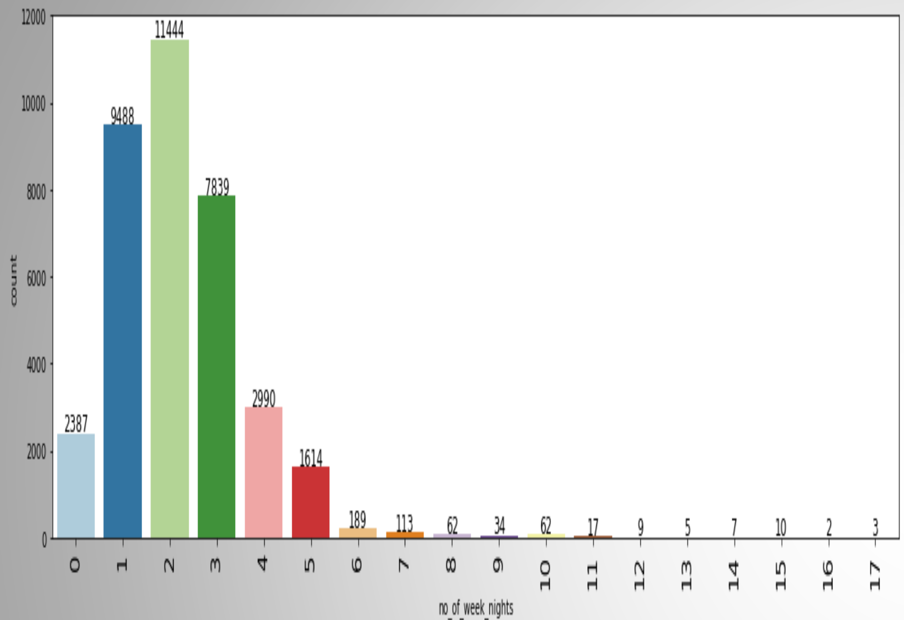


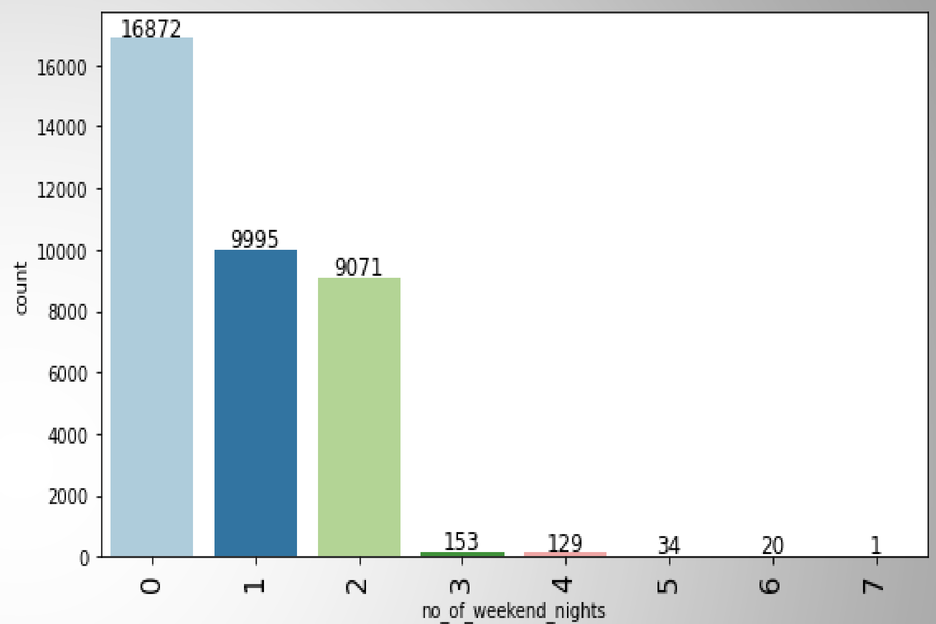
- No. of Children



- Most of the guests were two adults sharing a room.
- 21.2% of guests were by themselves.
- Children hardly frequented any of the hotels with
- There were 2 instances where 9 children were together and 1 instance with 10 children
- Almost all guests were adults staying by themselves or with another adult

Weeknights and Weekend nights

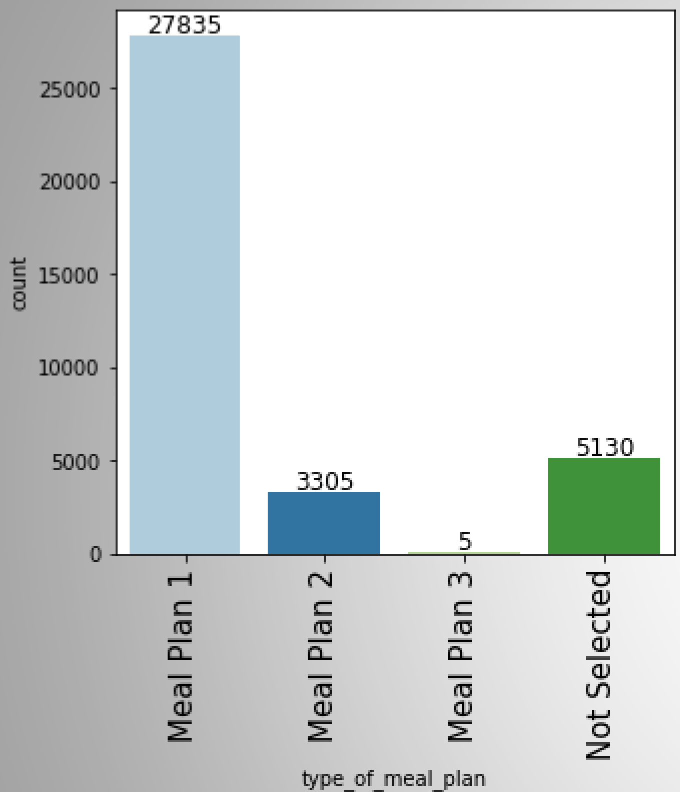


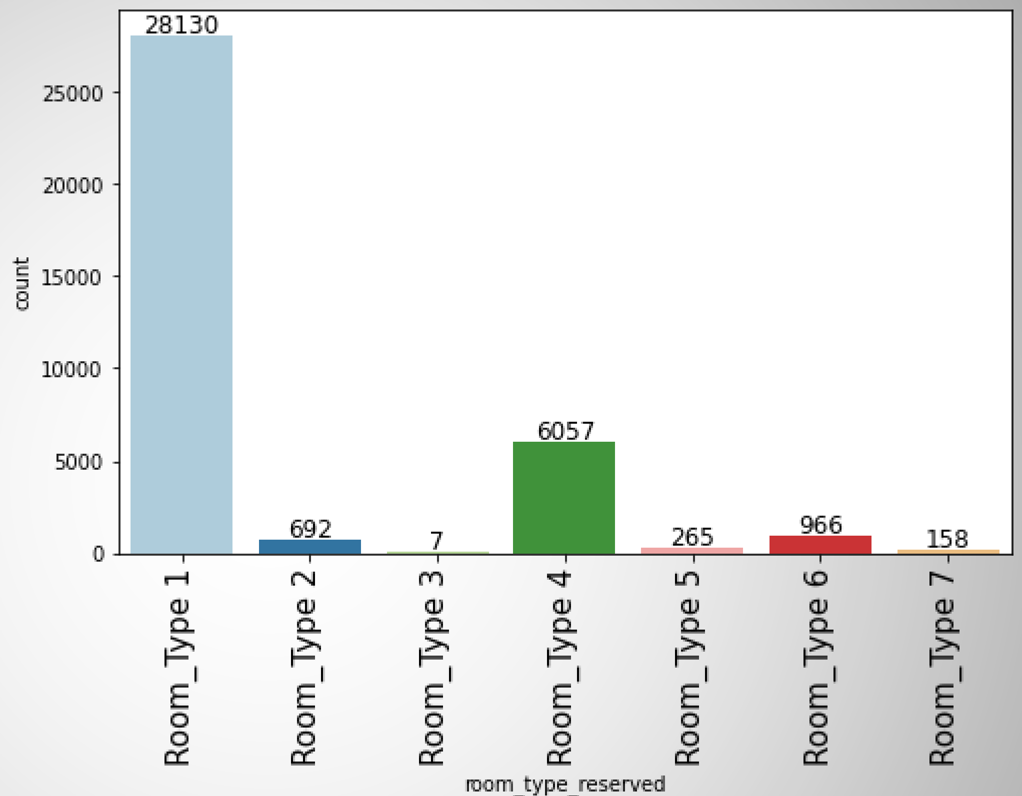


- During the weeknights, most bookings are for 2 days.
- 2-day and 3-day bookings are also booked often during the week.
- There were 3 stays that were 17 days long
- 0 day stays were those less than 24 hours

- 0-day stays were the most frequent during the weekend, which makes sense, as people travel often during the weekend
- 1 and 2-day stays were also very common.

Meal Plan and Room Type



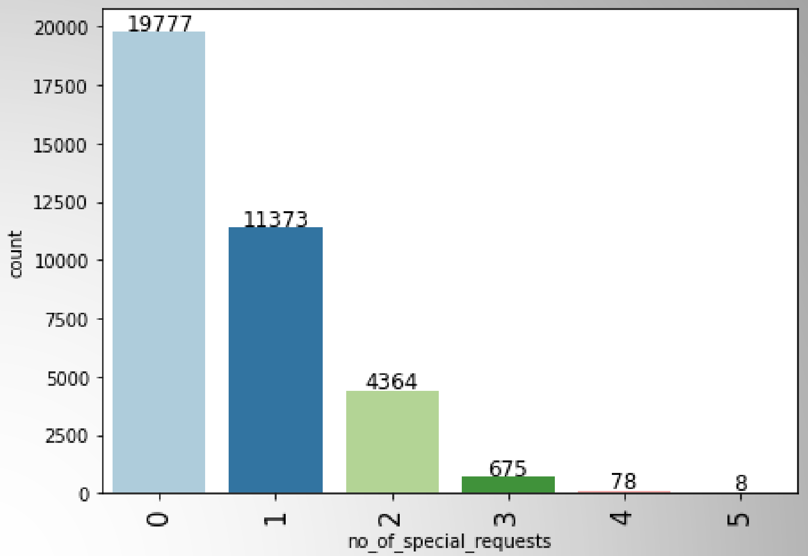


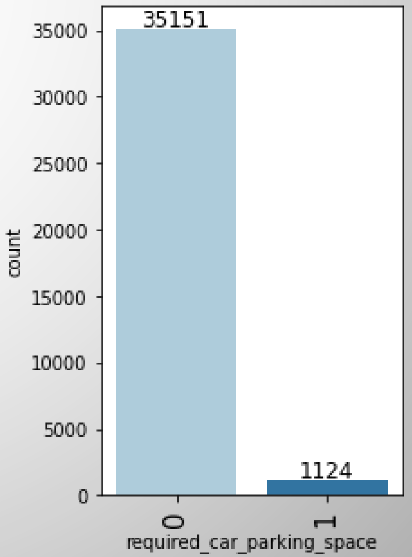
- Meal Plan 1 is the most popular plan, which is just breakfast
- If these do not have much an impact on determining if customer books or not, costs can be cut by discontinuing some meal plans

- Room Type 1 is by far the most popular
- Management should see about discontinuing some of the room types, especially if they increase costs.
- Room Type 4 is the only other room type that is booked or requested.

Special Requests and Parking Spaces

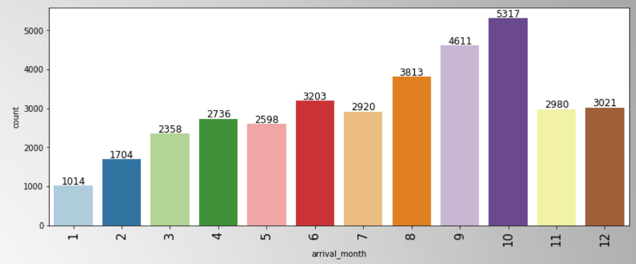
- Special requests and parking spaces are together because, requiring a parking space can be a request.
- The data does not specify if parking space requests are also included in special requests
- Most guests do not have a special requests; however, there is a large enough group that does
- The Hotel Group should make sure they can accommodate these customers in order to retain them

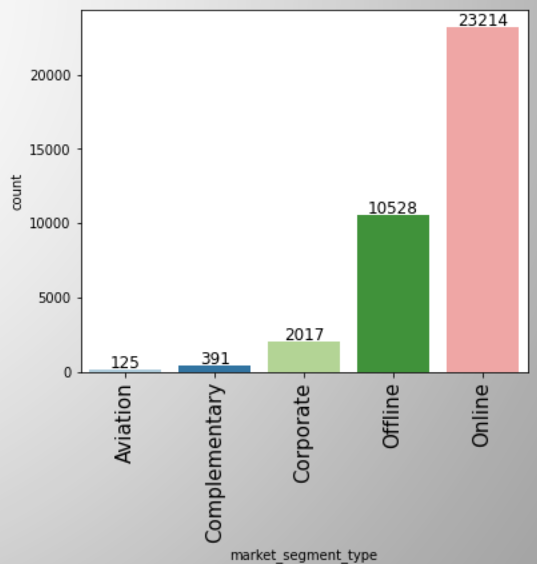




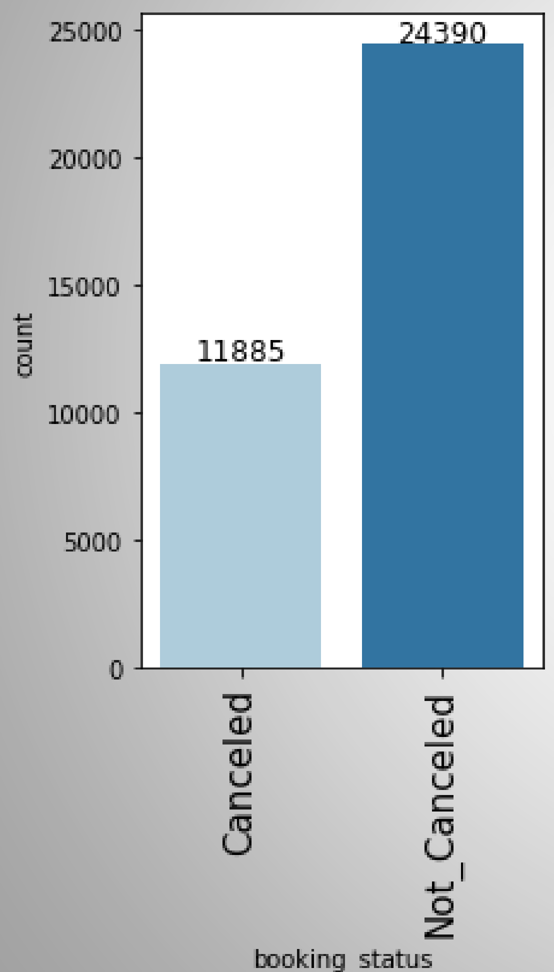
Arrival Month and Market Segment

- The busiest season appears to be end of summer into fall, with October, September and August having the most volume.
- Over half of the guest booked their reservation online. Management should focus on their website to make sure it is user-friendly for bookings, special requests, etc.





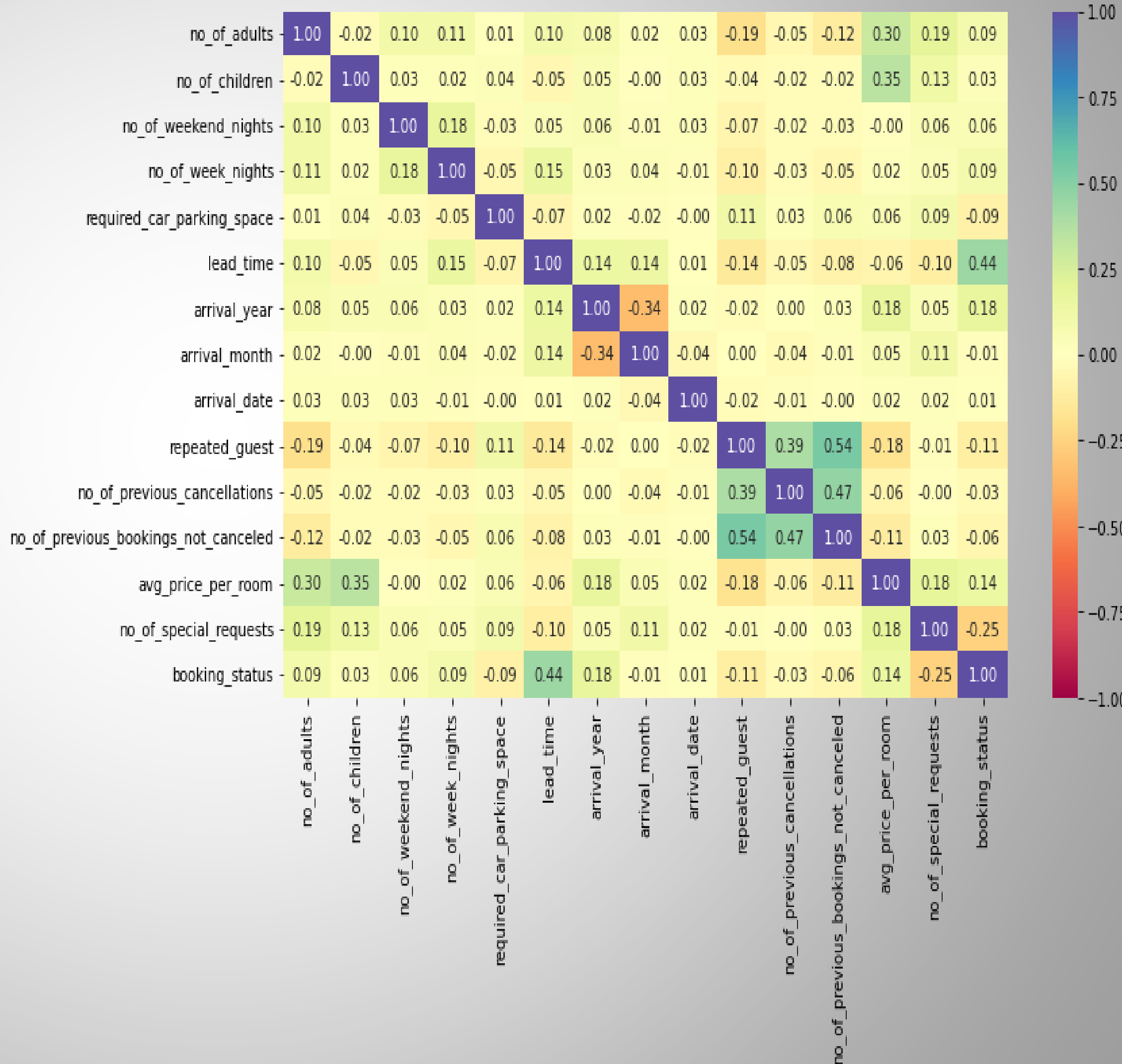
Booking Status



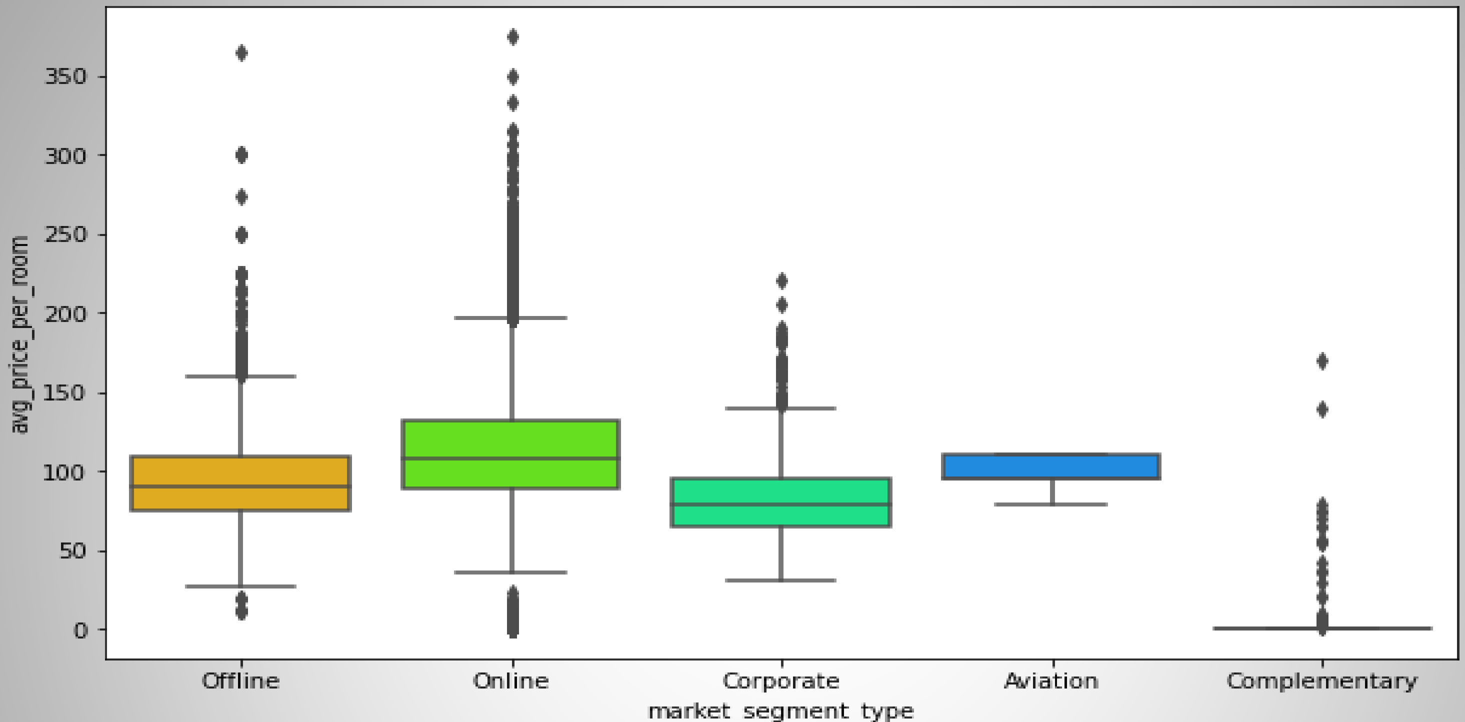
- 33% or 11,885 bookings were cancelled.
- Throughout the analysis, recommendations and observations will be made to help bring this number down.
- INN Hotels may need to restructure their cancellation policy also; however, this will be addressed at the end of the study.

Bivariate Analysis – Heatmap

- The highest positive correlation is among repeated guest number of previous bookings not cancelled.
- Previous cancellations, bookings not cancelled, and repeated guests all have a correlation among them.
- Lead time and booking status also have a moderately strong, positive correlation.
- A negative correlation exists among booking status and number of special requests.

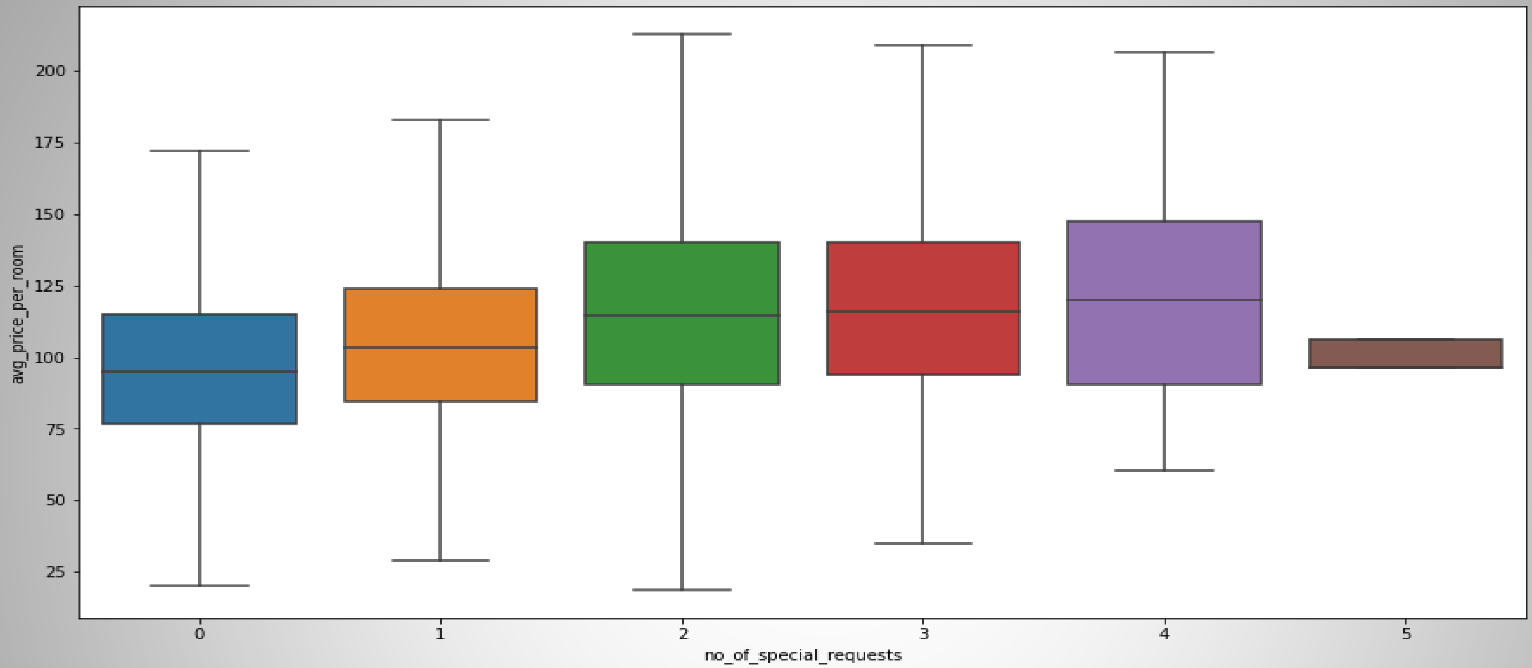


Average Price Per Room and Market Segment



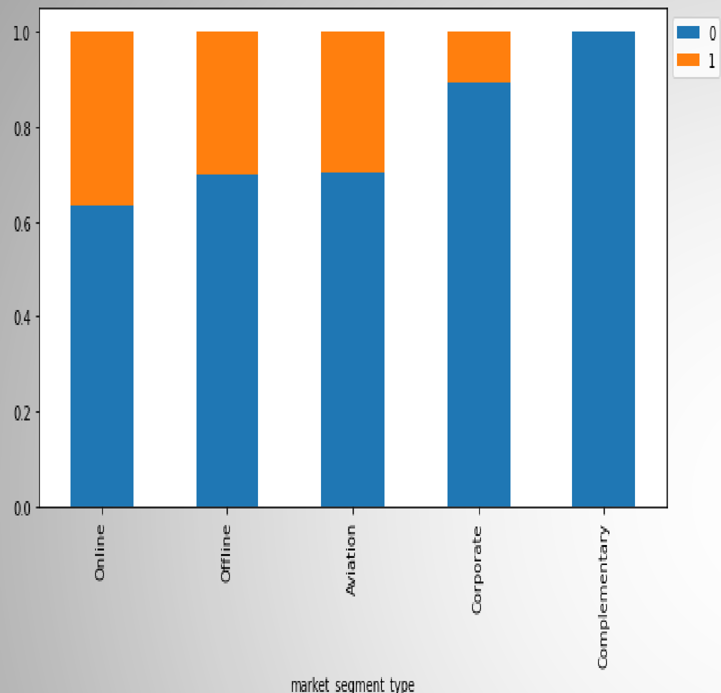
- The price per room varies according to market segment, with the online segment being the highest.
- Besides the “Complimentary” segment, “Corporate” has the lowest average price per room.
- Management should consider increasing prices for “Offline” bookings and possibly decreasing “Online” prices; however, this will be addressed in more depth at the end of the study.

Number of Special Requests and Average Price per Room



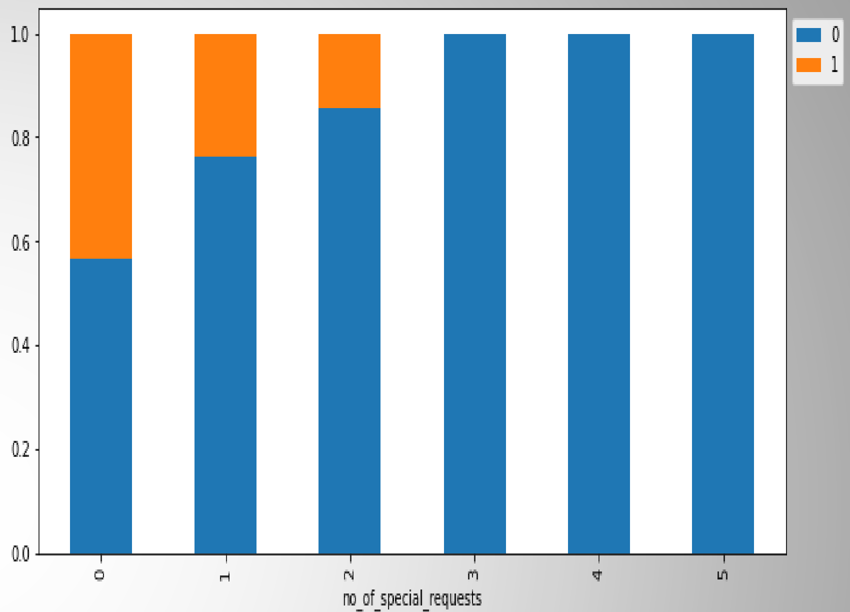
- More special requests increase as the average price of room increases.
- The more a customer pays for a room, the more accommodations they require and/or expect

Market segment type and booking status



- "Online" segment has the highest number of booking cancellations followed by "Offline"
- It may be a good idea to put in place cancellation policies specific to market segment type

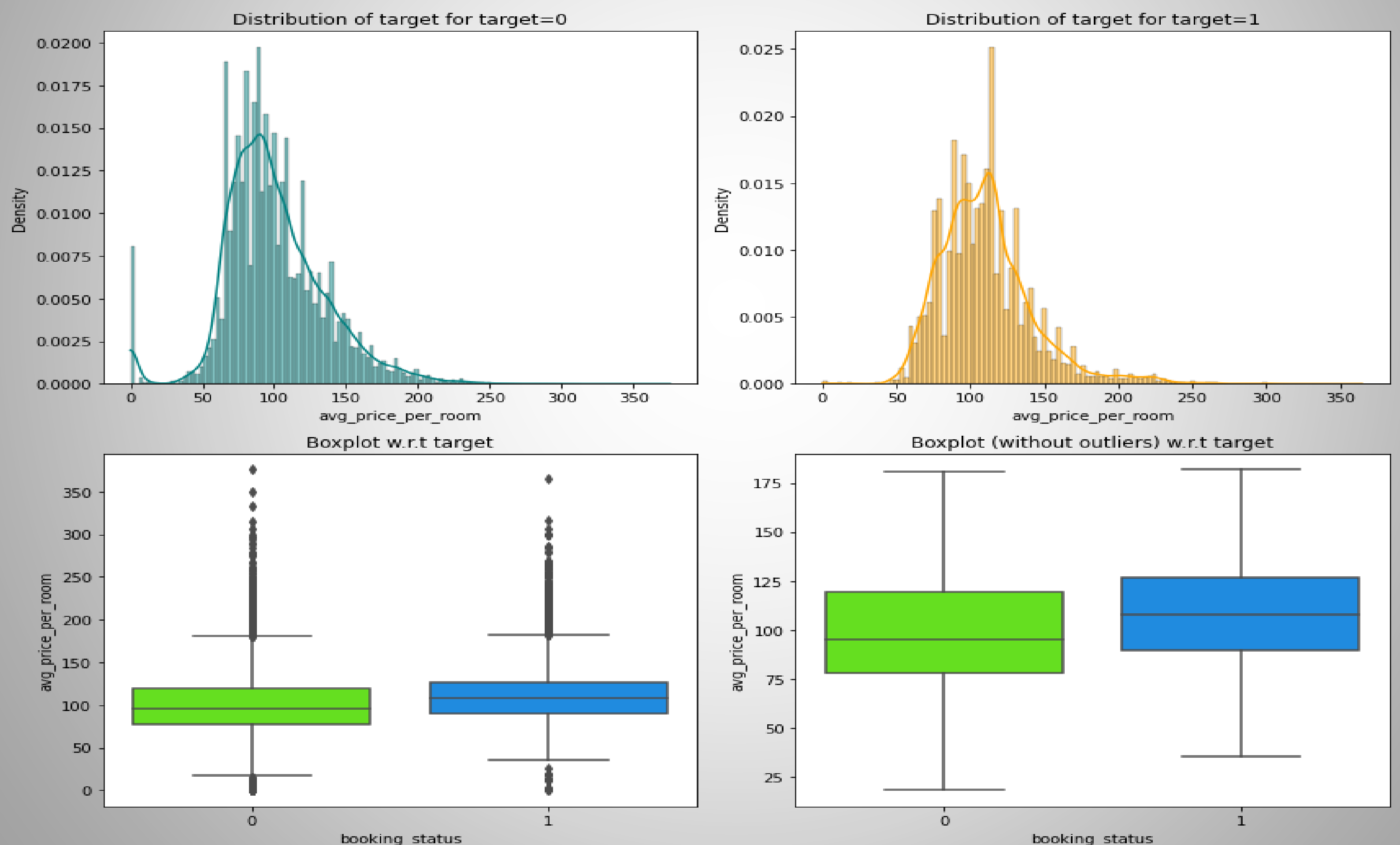
No. of special requests and booking status



- Guests that have 3 or more special requests do not usually cancel.
- Knowing that a guest is not going to cancel is valuable information for day-to-day operations to the company
- Most guests that cancel have no special requests.

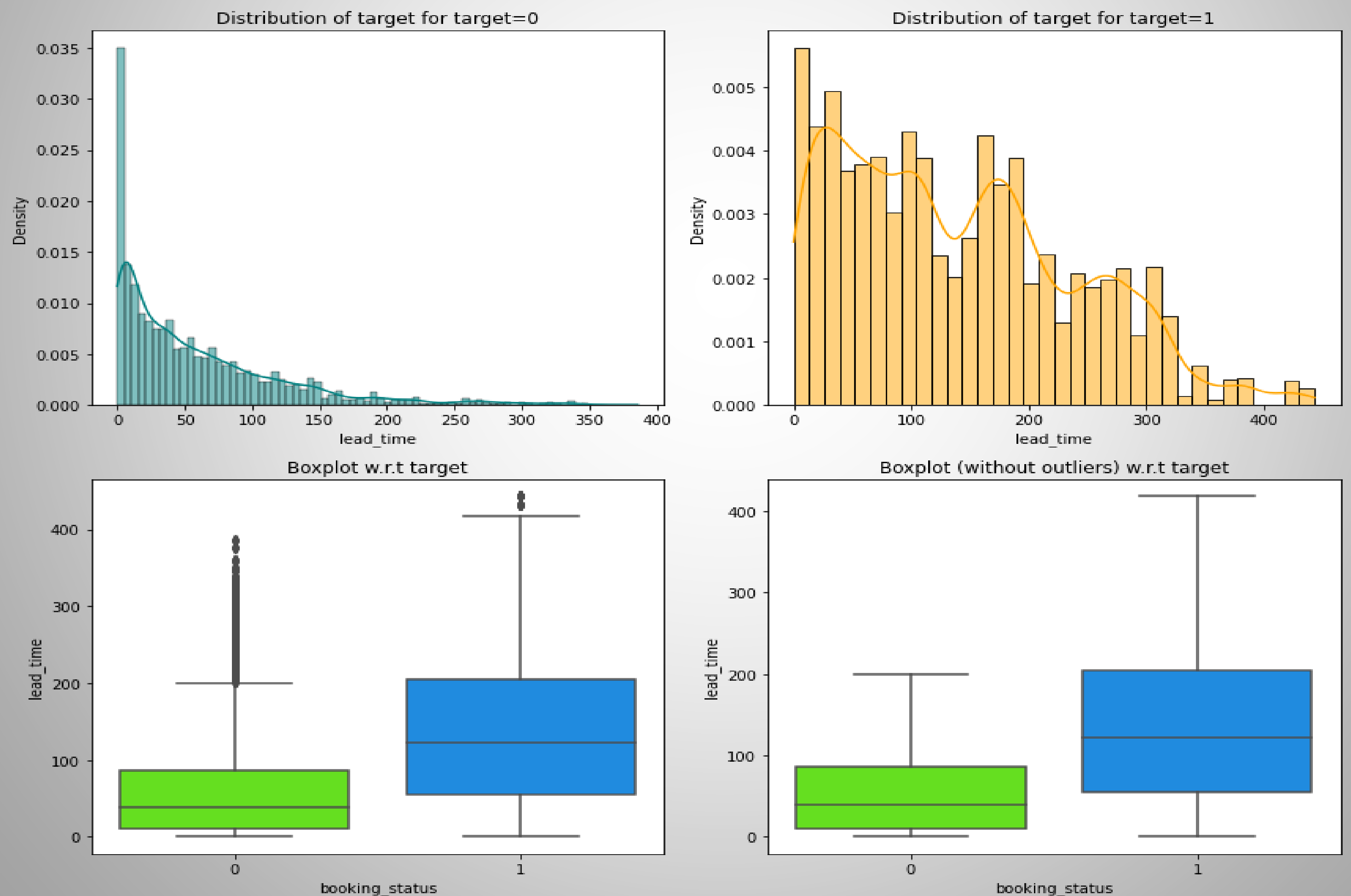
Booking status vs avg price per room

- There is a positive correlation between average price per room and booking status.
- Rooms that were cancelled have a higher price than those that didn't.

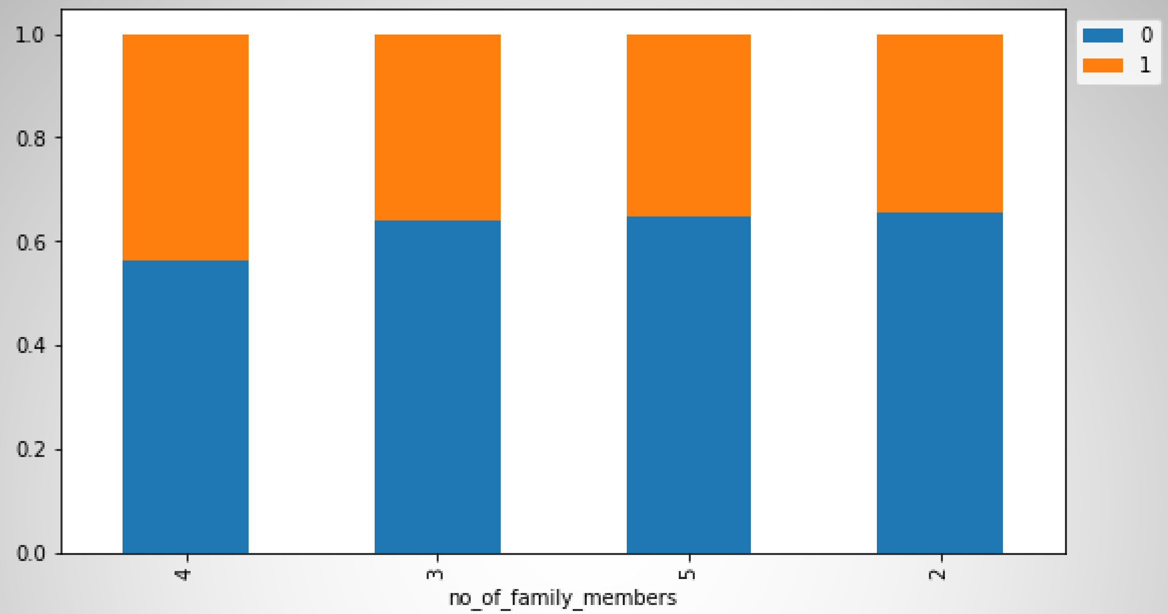


Booking status vs lead time

- A positive correlation exists between lead time and booking status.
- Bookings that were cancelled seem to have longer lead time days than those that weren't cancelled.



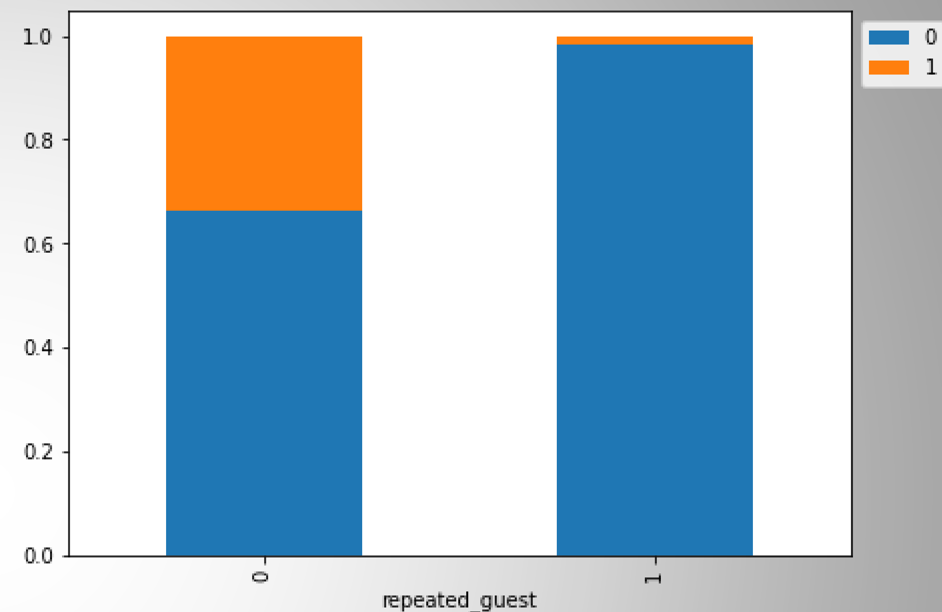
No. of family members vs booking status

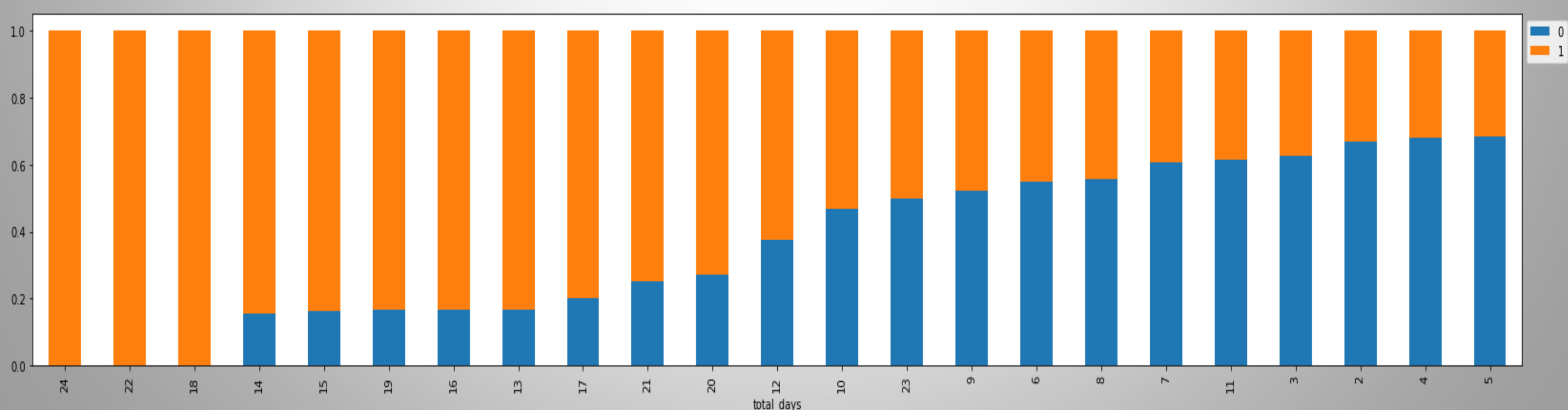


- Cancelled and not cancelled bookings seem to be stable among number of family members.
- Bookings with 4 family members has the greater chance of being cancelled compared to 2, 3 and 5 family member bookings.

Total Days and Booking Status

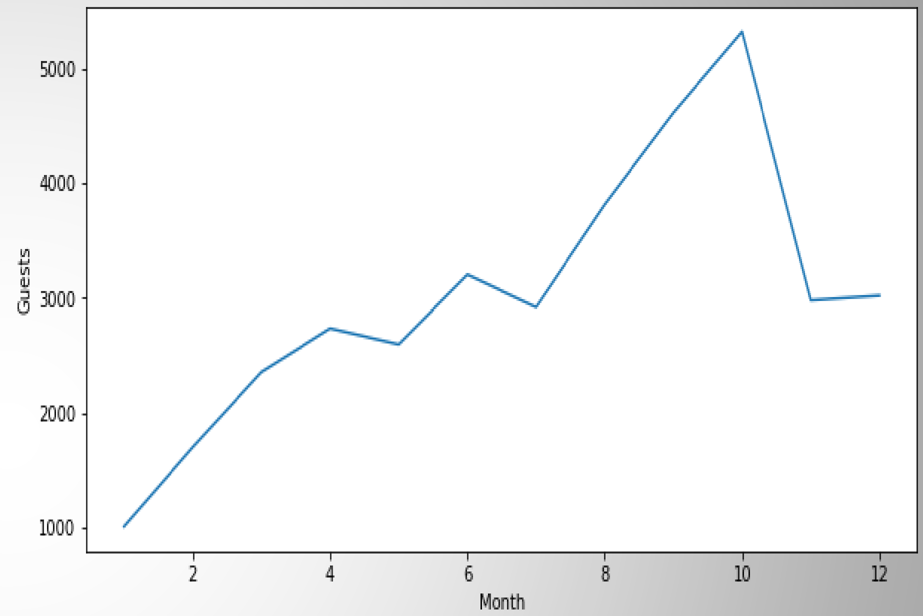
- The chart below represent customers who stay for at least 1 day or longer.
- From the given data, cancellations are more likely to occur when the lead days are larger.
- When it comes to repeated guests, they are less likely to cancel their reservations.

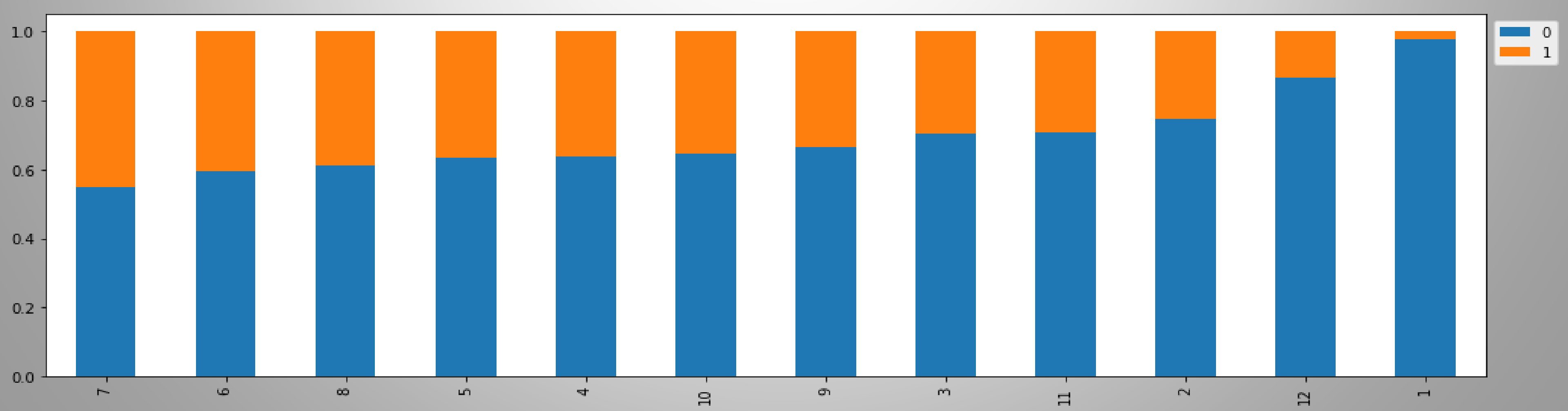




Arrival month, number of guests vs booking status

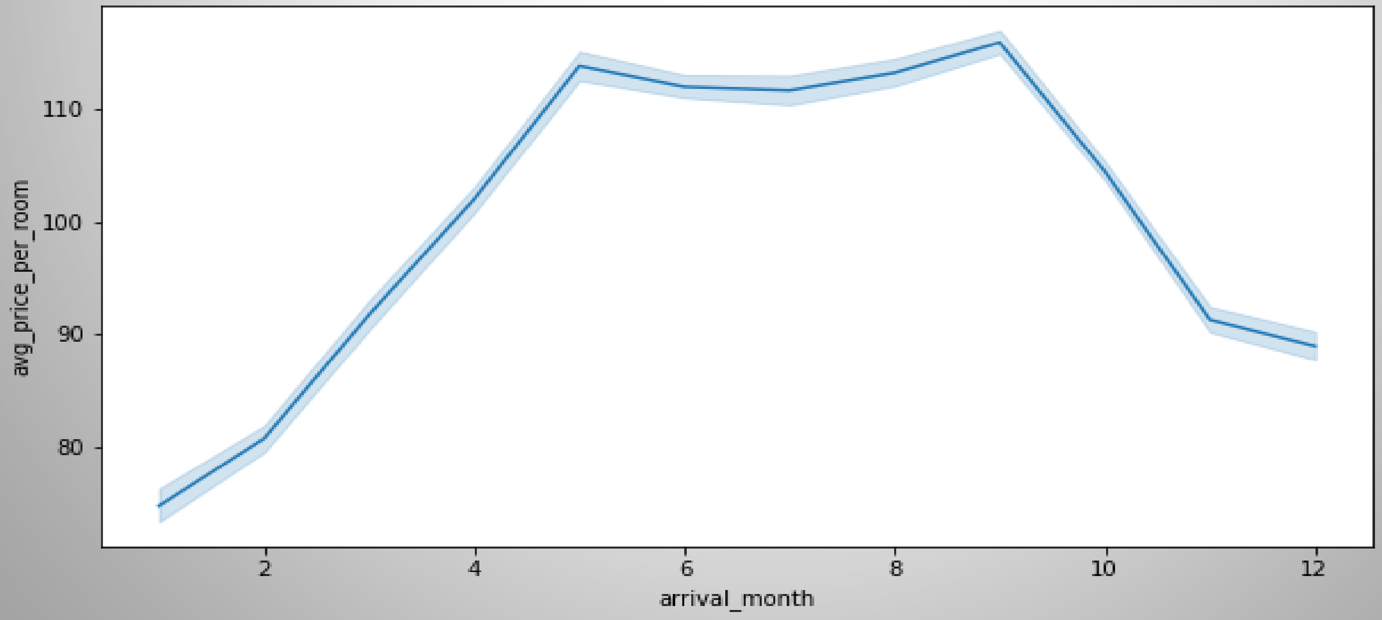
- The busiest month was October followed by September, August, and November.
- INN Hotels has a large number of cancellations in one of their busiest months, August.
- Finding solutions to cut cancellations in the busier months could lead to a vast improvement in profit and cost reduction.
- June and July had the highest number of cancellations followed by August.
- Arrival month gradually increases in the summer and peaks in October.



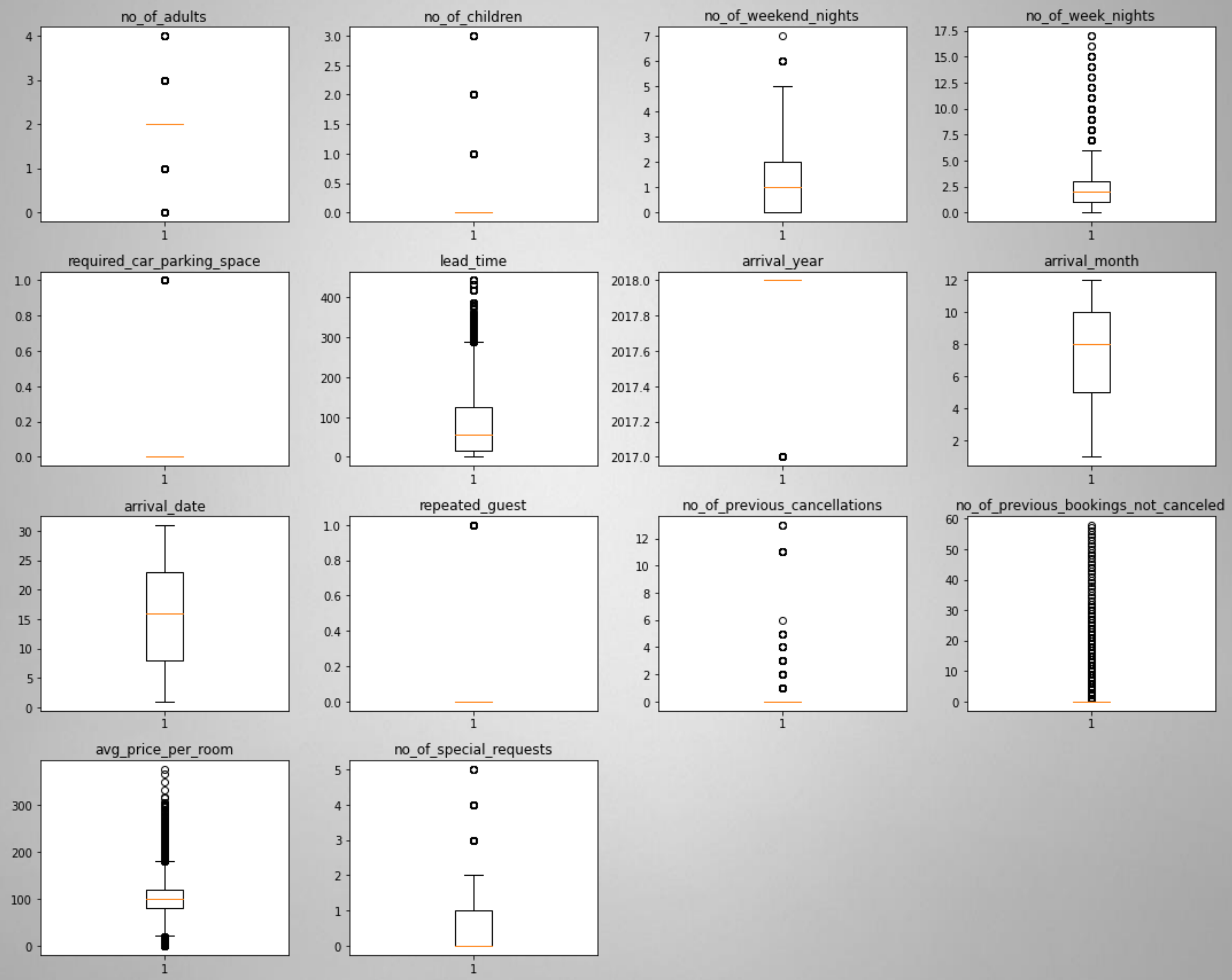


Arrival month and average price per room

- Price per room is the highest during the summer and fall months, which is good as these are the busier months of the year for the company.
- Price drops significantly from approximately $115.00 during busy season, to as low as approximately $70 at the end and very beginning of the year.
- Management may want to consider adjusting prices during both the busier and the slower months of the year.
- Both arrival month and total number of guests peak in October

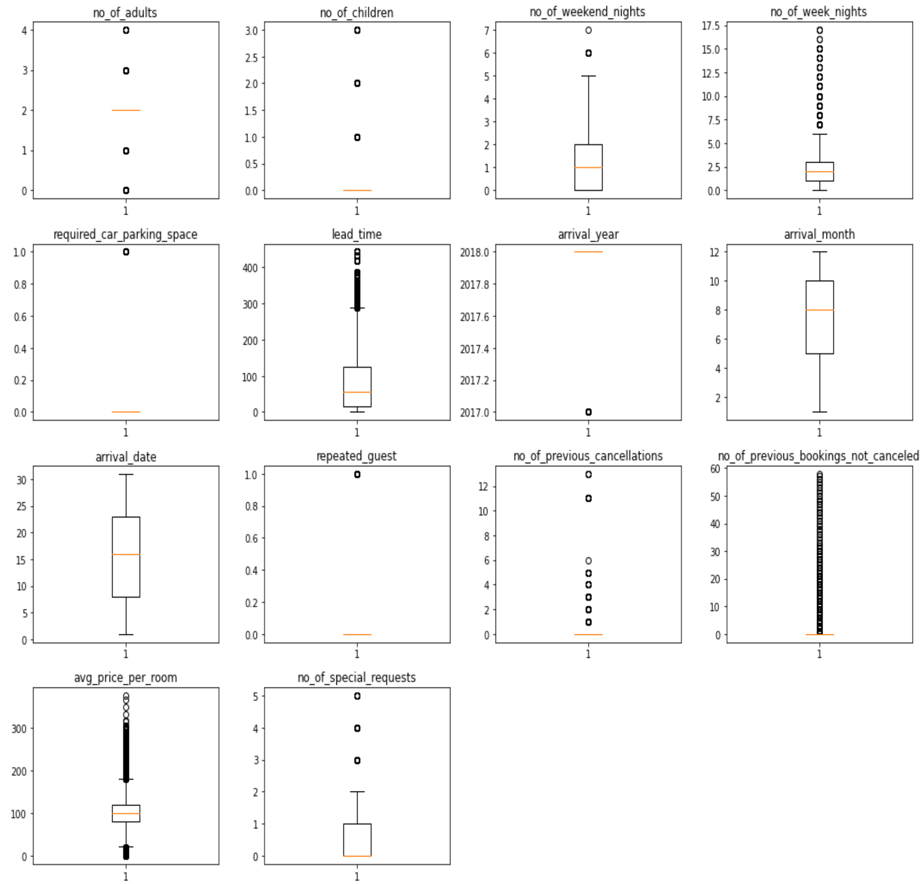


Outlier Detection



Outlier Detection

- All features have outliers except for arrival month.
- Though outliers exist in many features, we will treat not remove them as they are proper values.
- No. of previous bookings not cancelled, lead time, no of week nights, avg price per month all have a large number of outliers



Logistic Regression Model

Default-threshold

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80600	0.63410	0.73971	0.68285

Observations

- p-values of a variable tell us whether it is significant or not, when considering significance level of 0.05
- High p-values exist and will need to be dropped before making interpretations on the model.
- Before dropping p-values though, we must check for multicollinearity as these affect p-values as well
- We will drop the predictor variables that have VIF score greater than 5, and then drop p-values greater than 0.05

Initial Model

Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25364
Method:	MLE	Df Model:	27
Date:	Thu, 17 Nov 2022	Pseudo R-squ.:	0.3292
Time:	16:52:47	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	3987.836	0.004	0.997	-7798.656	7833.373
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	-40.5975	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

Revised Model – Logistic Assumptions Addressed

- This is the model used after removing high p-values and VIF scores.
- Positive coefficients of predictor variables show that an increase in each one will increase the chance of a booking to get cancelled.
- For example, no of previous cancellations has a coefficient of 0.2288, which means it is a feature that has a strong influence with whether a booking will get cancelled or not.
- F-1 score needs to be maximized as the higher the score the higher the chances are of reducing False Negative and False Positives

Training performance:

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

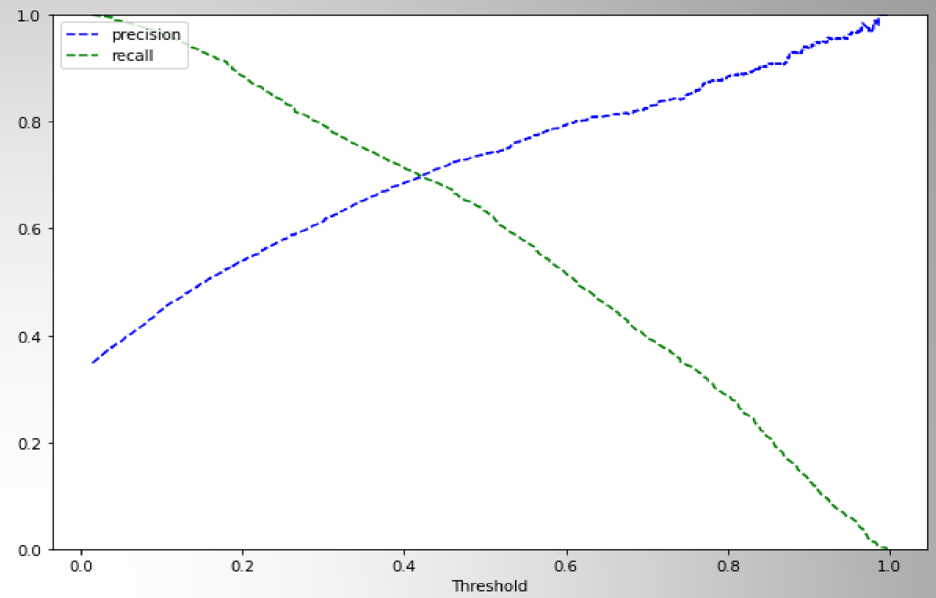
```
                           Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25370
Method:                           MLE   Df Model:                           21
Date:                Mon, 14 Nov 2022   Pseudo R-squ.:                  0.3282
Time:                        12:25:13   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
==================================================================================================
<IPython.core.display.Javascript object>
```

Probability and Odds – Logistic Model

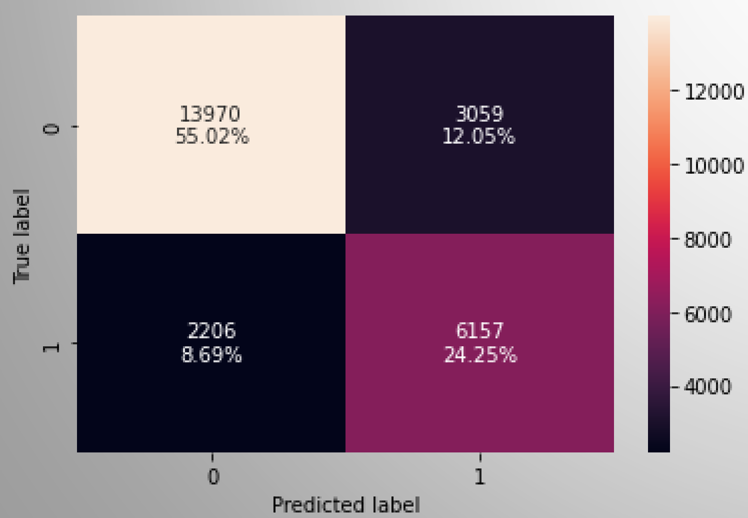
required_car_parking_space	lead_time	arrival_year	arrival_month	repeated_guest	no_of_previous_cancellations	av
0.20296	1.01583	1.57195	0.95839	0.06478	1.25712	
-79.70395	1.58331	57.19508	-4.16120	-93.52180	25.71181	

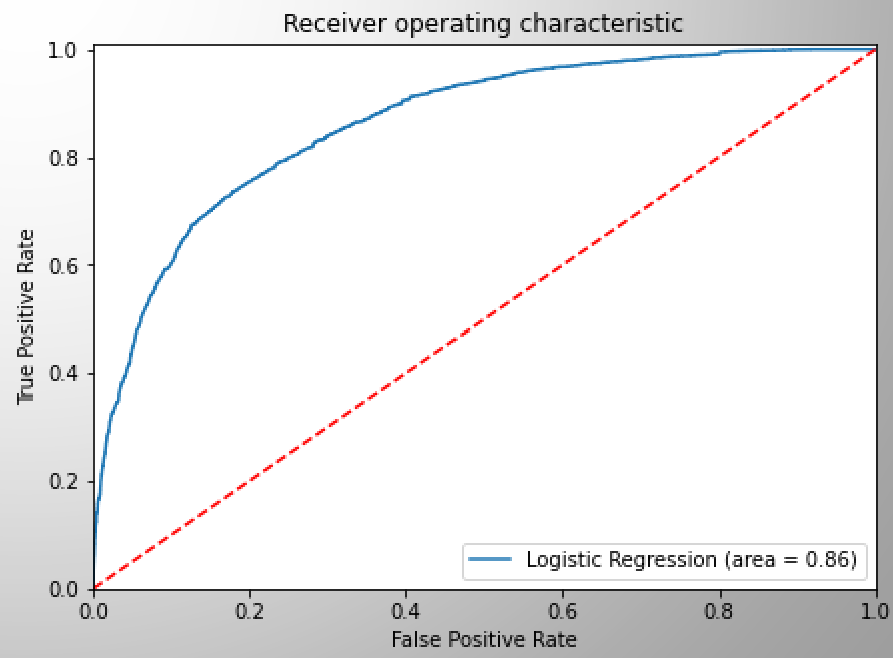
- Lead_time: holding all other features constant a 1 unit change in lead_time will increase the chance of a booking being cancelled 1.01 times or 1.58% increase in odds of being cancelled.
- Arrival_month: Holding all other features constant, a 1 unit change in arrival month will decrease the odds a booking will be canceled ~0.95 times or a ~4.16% decrease in odds of being cancelled.
- Repeated_guest: holding all other features constant, a 1 unit change in repeated_guest, will decrease the chance of a booking being cancelled by 0.06 times or 93.5% decrease of being cancelled.

- In order to increase the F1 score (and increase TP and FN rate) we used the precision-recall curve and the AUC-ROC curve to find the optimal threshold for the model.
- Below is the confusion matrix from the initial model using the training set
- On the top right is the precision-recall curve and the bottom right is the auc-roc curve.
- The optimal threshold using the auc-roc curve came out to be 0.37
- Below is the confusion matrix on training data using the auc-roc curve as optimal threshold.
- Using the precision-recall curve, the optimal threshold came out to 0.42



Optimal threshold—auc-roc curve





Logistic Regression – Model Performance

Observations

- The model is giving an f1 score of 0.70 on the train and test set respectively.
- The threshold using the auc-roc curve with a threshold of 0.37 gave the best results.
- Since the train and test sets are comparable, the model is not overfitting.

Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73622	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868

Testing performance comparison:

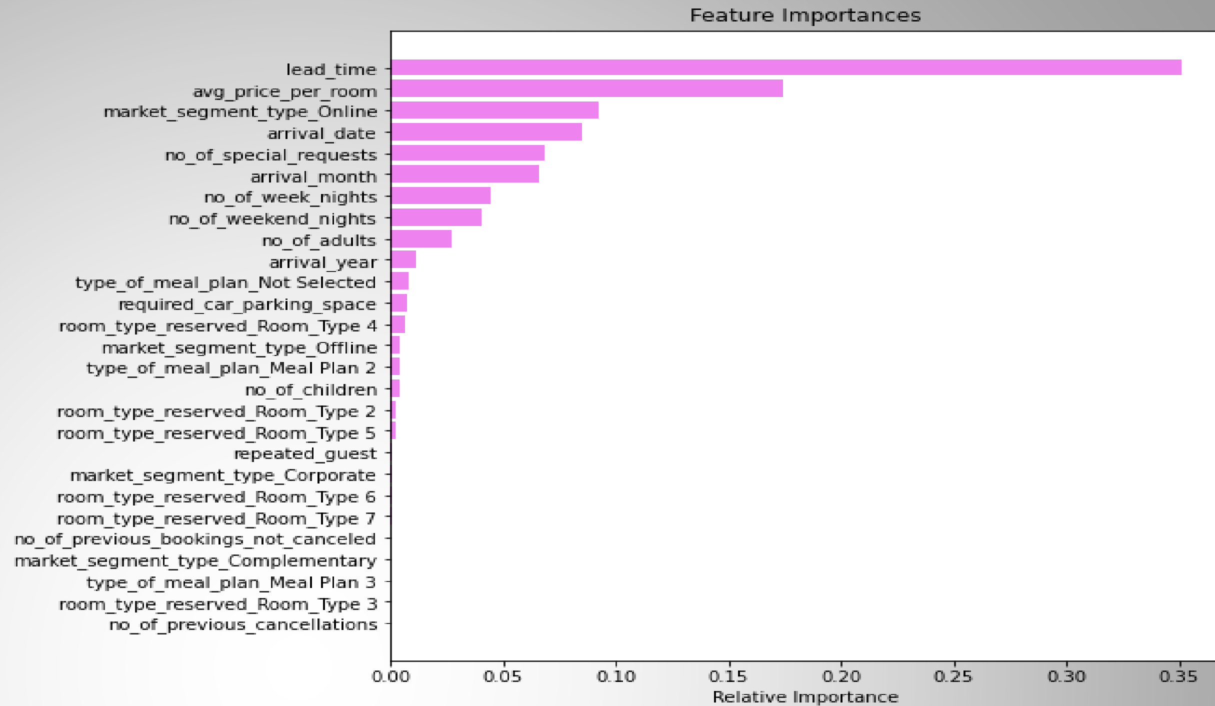
	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80465	0.79555	0.80345
Recall	0.63089	0.73964	0.70358
Precision	0.72900	0.66573	0.69353
F1	0.67641	0.70074	0.69852

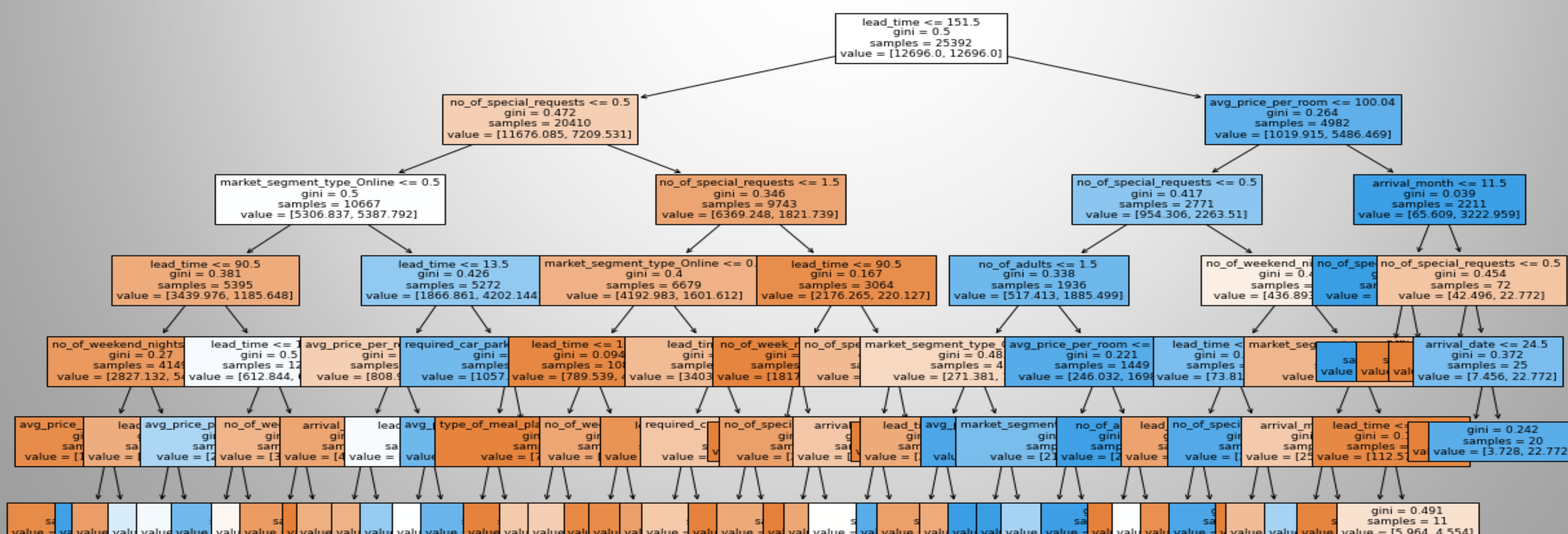
Train-Pre-pruning

	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

Test-Pre-pruning

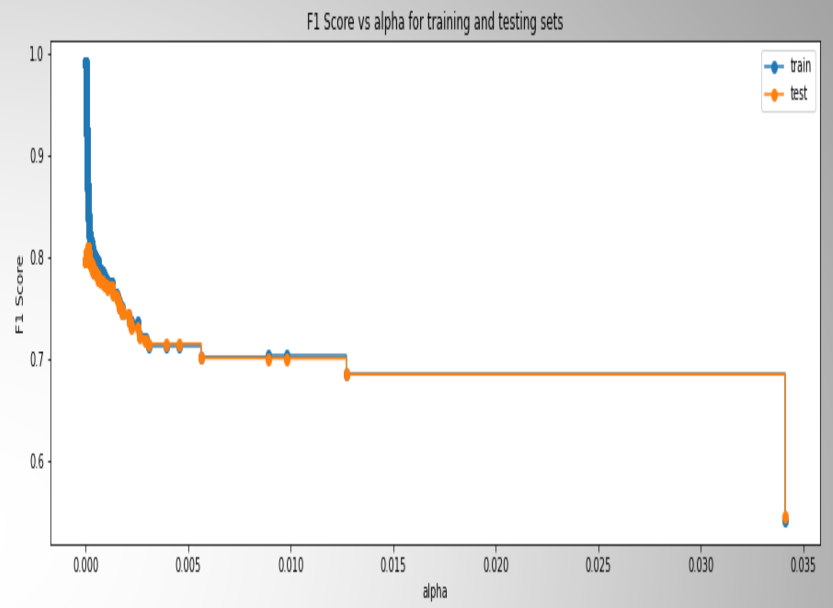
	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309

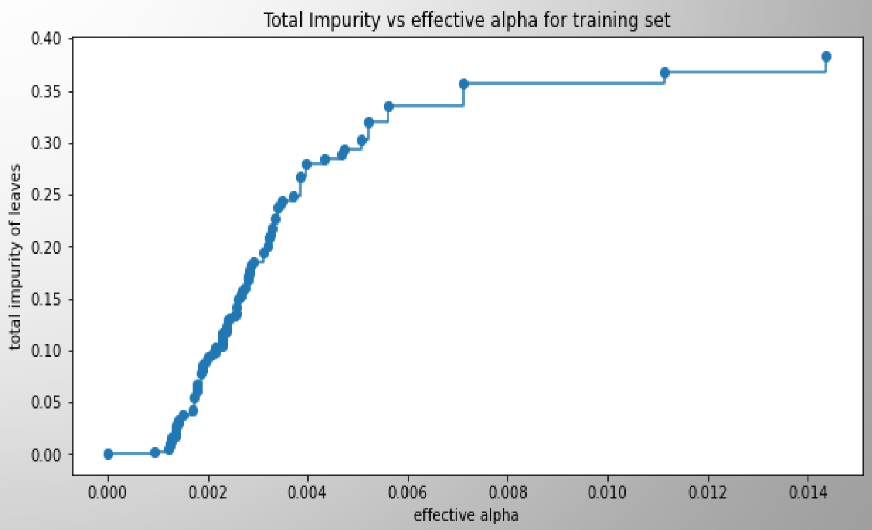




Model Improvement and Evaluation

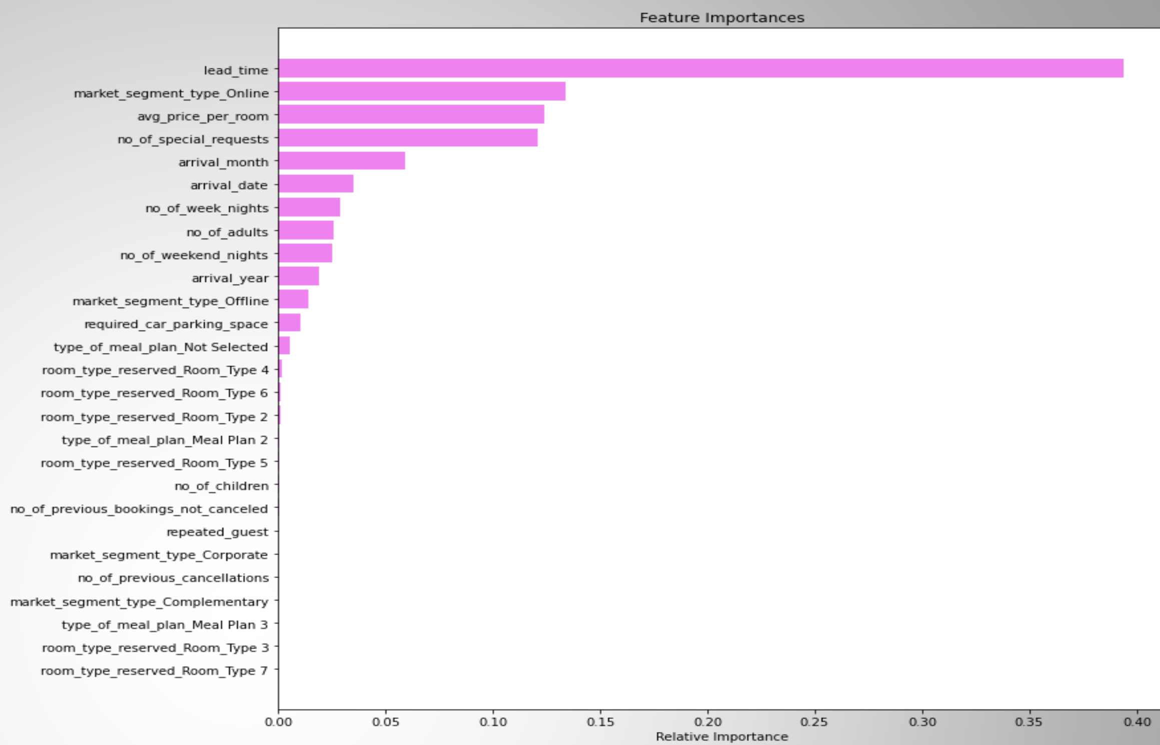
- The farther the tree model grows the more complex it becomes, leading to overfitting.
- Based off the initial train and test performance, overfitting is present
- To improve this, we prune the tree while also choosing the most important features
- The most important features are know as the Gini Index.
- INN Hotels most important feature is lead time, followed by, average price per room and market segment type online.
- We use cost complexity when pruning the tree, which in turn, uses the effective alphas—its goal is to find the weakest link in the tree.





Final Model Performance

- Lead time was the most important feature pre-pruning and post-pruning.
- Post-pruning market segment online jumped average price per room in importance.
- Decision tree post-pruning is giving the highest accuracy on the test set at 0.86
- Decision tree post-pruning saw the most improvement on Recall, jumping from 0.81 pre-pruning to 0.85 post-pruning.
- Decision Tree post-pruning for train and test are comparable.
- Both have reduced overfitting.



Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.99421	0.89954
Recall	0.98661	0.98661	0.90303
Precision	0.99578	0.99578	0.81274
F1	0.99117	0.99117	0.85551

Testing performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.87118	0.86879
Recall	0.81175	0.81175	0.85576
Precision	0.79461	0.79461	0.76614
F1	0.80309	0.80309	0.80848

Conclusion and Recommendations

- The higher the number of special requests, the less likely a booking will be canceled. Specifically, bookings with 3 or more requests are least likely to be cancelled. This should help management to be able to accurately know what their vacancy number will be and will help with future bookings.
- Customers tend to cancel more often when lead days exceed a little over 115 days. Management should possibly follow up with the customer after a certain amount of time before their scheduled stay or develop some type of cancellation policy or deposit policy during the initial booking.
- Management may want to consider raising prices slightly during the slow months as the price per room drops to below $70 in January.
- The people who booked online have the highest probability of cancellation, so it would be beneficial to consider having a customer pay a deposit when booking or if a person cancels within a certain timeframe leading up to their stay, they would have to pay for the entire reservation.