INN Hotel Project

Supervised Learning – Classification

A client cancellations prediction analysis

25 – 03 – 2022

Contents / Agenda

- Business Problem Overview and Solution Approach
- Executive Summary
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Business Problem Overview and Solution Approach

A significant number of hotel bookings are called off due to cancellations or no-shows. The typical reasons for cancellations include the 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.

1. The cancellation of bookings impact a hotel on various fronts:
2. Loss of resources (revenue) when the hotel cannot resell the room.
3. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
4. Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
5. Human resources to make arrangements for the guests.

For the development, of a solution for this problem a consideration was realized to de develop either a regression solution or a decision tree. As both approaches can help to realize a classification of the result for the cancellation analysis. The model can give us an input of the expect cancelations based on diferent factors of the guests that are registering for a stay.

Executive Summary

- One of the most important opportunities for the revenue of the INN otel Group is through a system of penalty fees based on the days before the cancellation can be done, as one of the segments with the highest rate of cancellation is the online one.

- For the attention brought to the guest the process of adding a special request in the process of the reservation can help the clients to complete the reservation and reduce the total cancellations done.

- The meal plans can have upscale sales with the reduction of the price between the first and second type of plan.

EDA Results

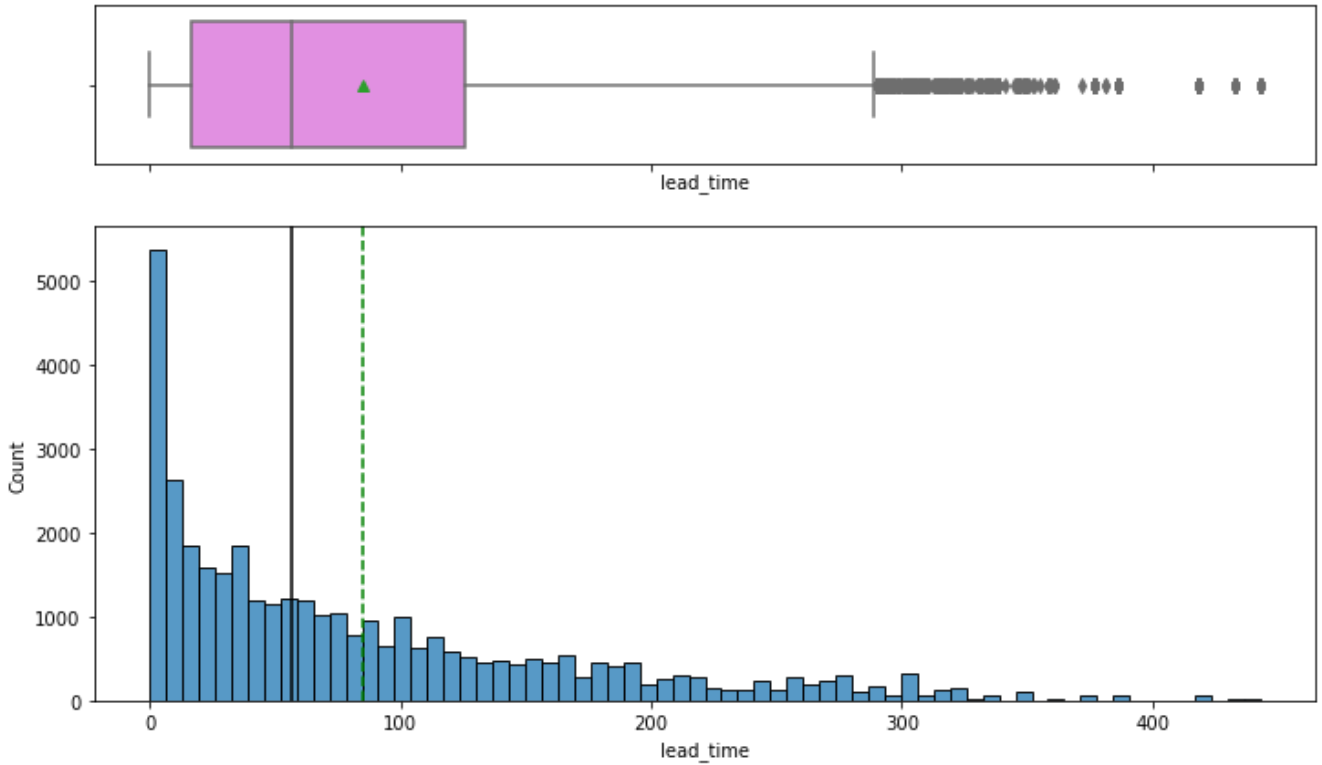
- The dataset provided contains 36275 rows and 19 columns.
- The following table shows a statistical summary of the numerical variables:

	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.00000	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000	4.00000
no_of_children	36275.00000	0.10528	0.40265	0.00000	0.00000	0.00000	0.00000	10.00000
no_of_weekend_nights	36275.00000	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000	7.00000
no_of_week_nights	36275.00000	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000	17.00000
required_car_parking_space	36275.00000	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000	1.00000
lead_time	36275.00000	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000	443.00000
arrival_year	36275.00000	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000	12.00000
arrival_date	36275.00000	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000	31.00000
repeated_guest	36275.00000	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000	1.00000
no_of_previous_cancellations	36275.00000	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000	13.00000
no_of_previous_bookings_not_canceled	36275.00000	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000	58.00000
avg_price_per_room	36275.00000	103.42354	35.08942	0.00000	80.30000	99.45000	120.00000	540.00000
no_of_special_requests	36275.00000	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000	5.00000

Link to Appendix slide on data background check

EDA Results – lead_time

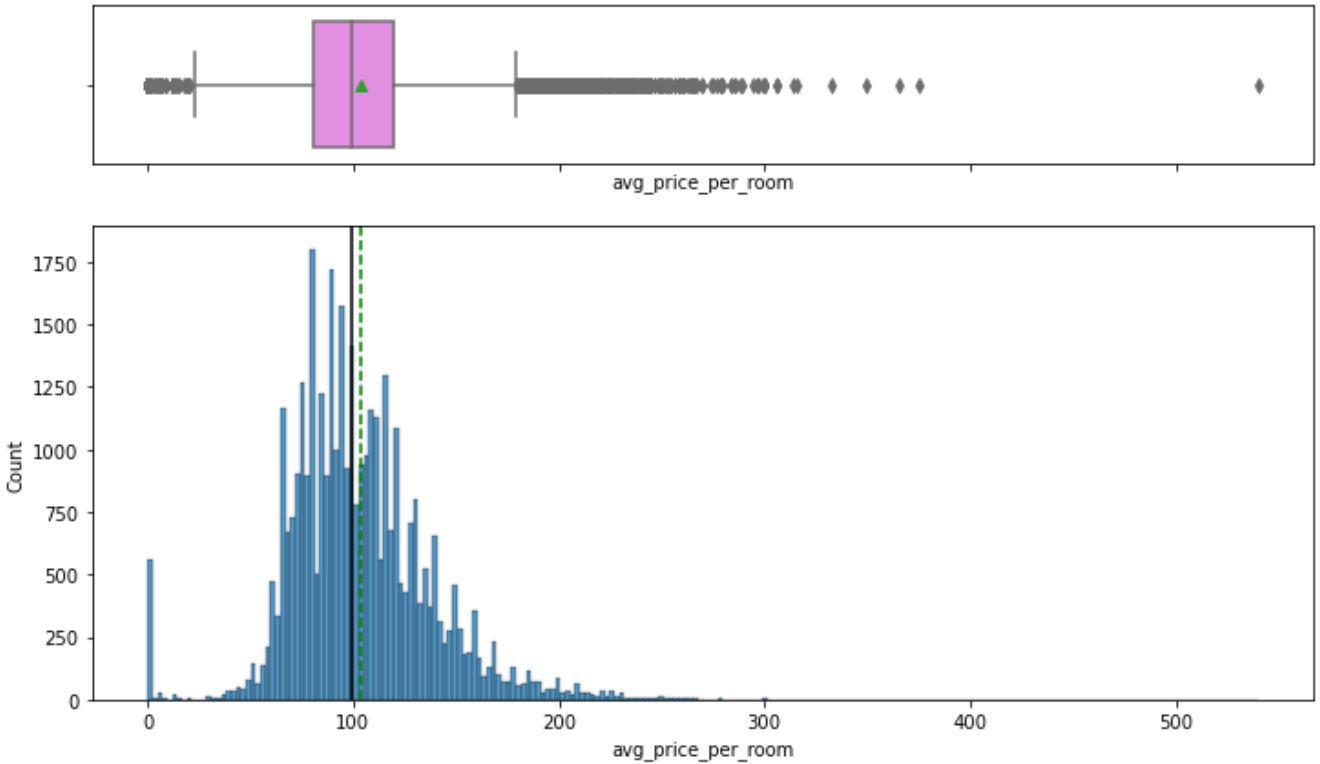
The data of the lead time is right-skewed. Most of the guests are requesting a room the same day they are staying. The average guest, request a room with 85 days of anticipation, but 50% of the request are requested under 57 days. Even though the mode of the lead is the same day, the rest of the requests are made with enough time in advance.



Link to Appendix slide on data background check

EDA Results – avg_price_per_room

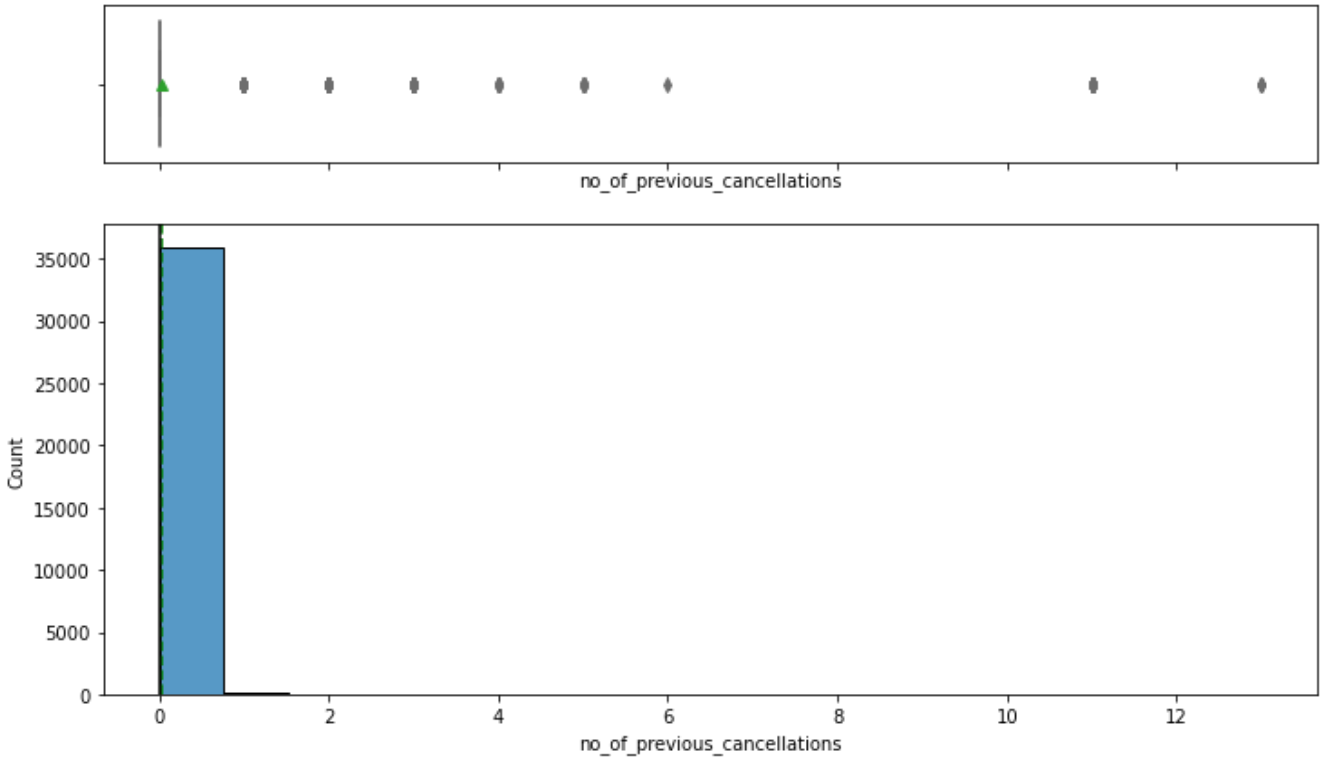
The average price per room is $ 103, and the mean price per room is $ 99. This variable is right-skewed as there are some outlier values that cause the skewness of this variable. Based on the data the skewed data is caused by the prices almost over $ 180.00



Link to Appendix slide on data background check

EDA Results – no_of_previous_cancellations

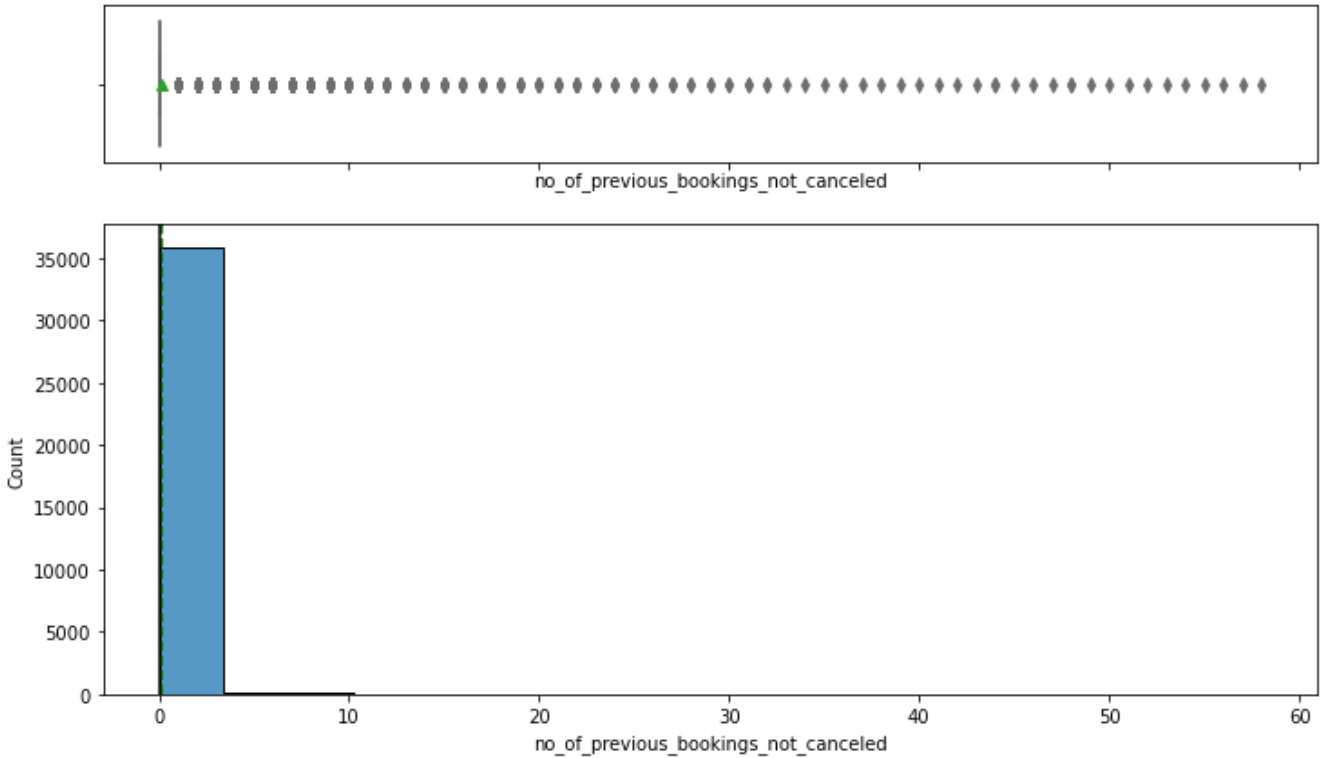
The number of previous cancellations tends to be 0 as the majority of the guest stay for the first time in the hotel. Therefore, it is expected to have little previous cancellation, and this would limit the predictions of the model with this variable.



Link to Appendix slide on data background check

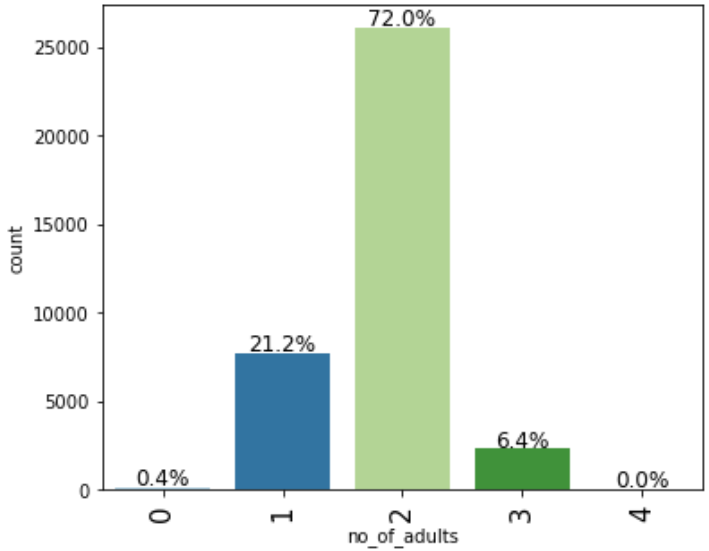
EDA Results - no_of_previous_bookings_not_canceled

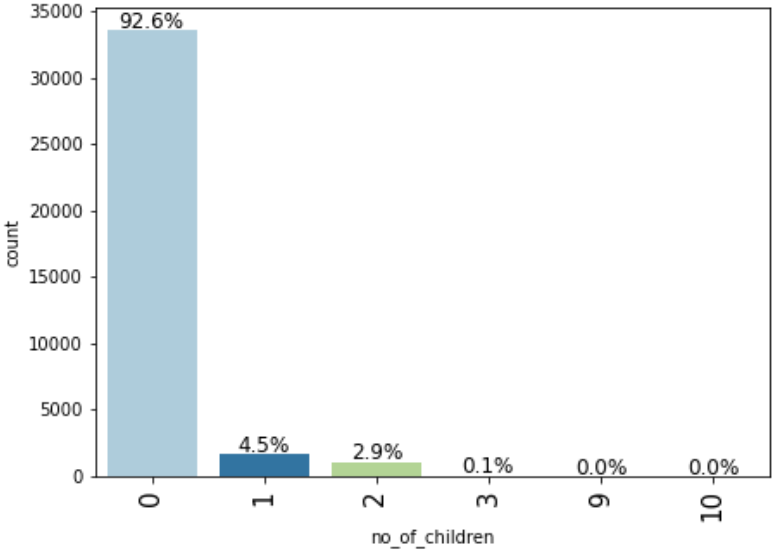
The number of previous bookings not cancelled tends to be 0 as the majority of the guest stay for the first time in the hotel. Therefore, it is expected to none previous not canceled reservations, and this would limit the predictions of the model with this variable.



Link to Appendix slide on data background check

EDA Results – Adults and Childs



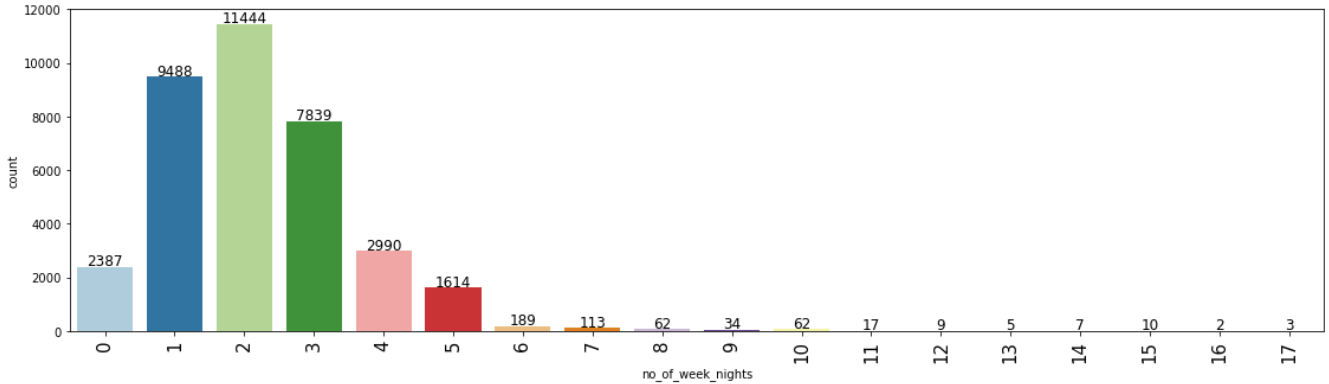


The number of adults tends to be a couple of guests per room. The number of children tends to be 0 and only in 7% of the cases child stay in the hotel.

Link to Appendix slide on data background check

EDA Results

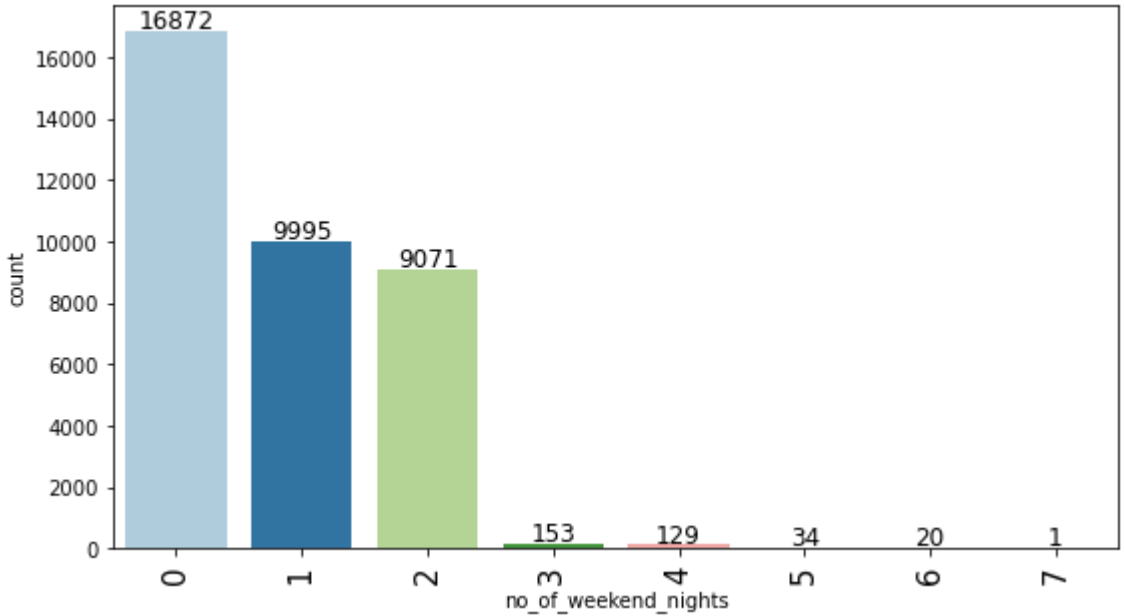
Most of the guest stay between 1 and 3 days at the hotel. The number 0 can be understood as the client that have never stayed in a weekend night.



Link to Appendix slide on data background check

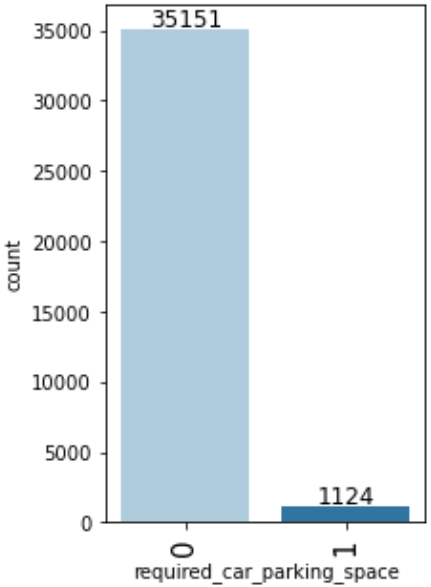
EDA Results

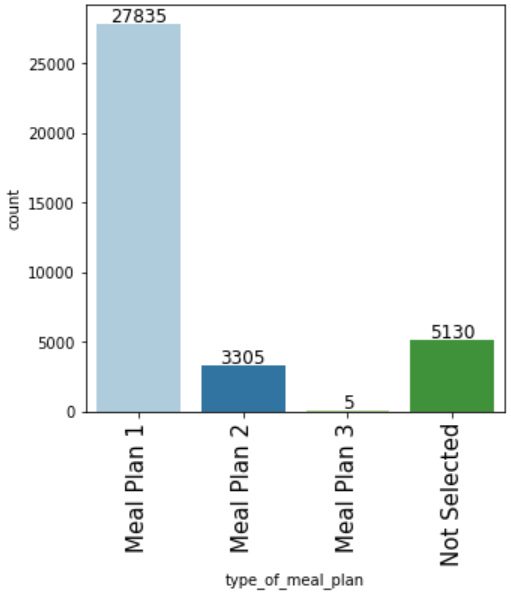
The client that stay on the weekend nights usually spend one or the two days. There are some cases when they stay on multiple weekends on a single reservation.



Link to Appendix slide on data background check

EDA Results – Parking Spaces and Meal plan

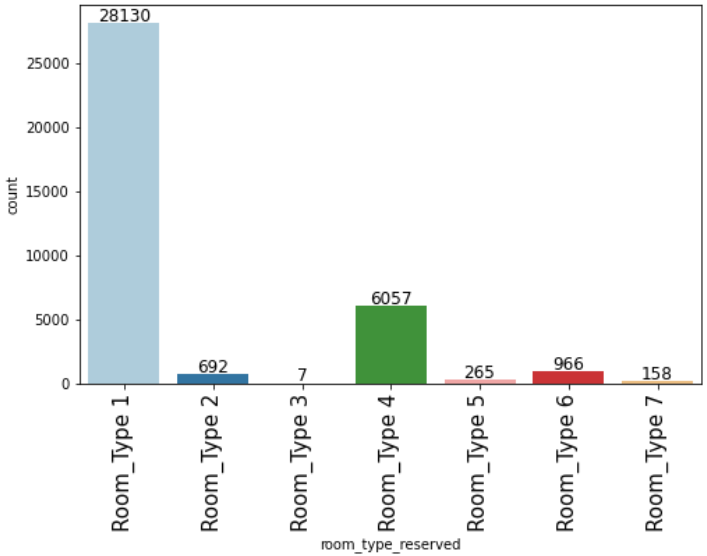


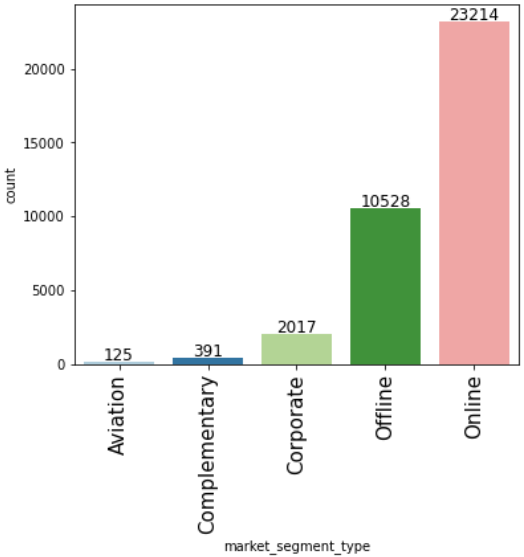


The parking space isn't a defining factor as in little cases is requested. Then the case for the meal plan over the 85% of the guests at least will have breakfast in the hotel. The rest of the meals tend to be less desired by the clients.

Link to Appendix slide on data background check

EDA Results – Type of room and Market Segment



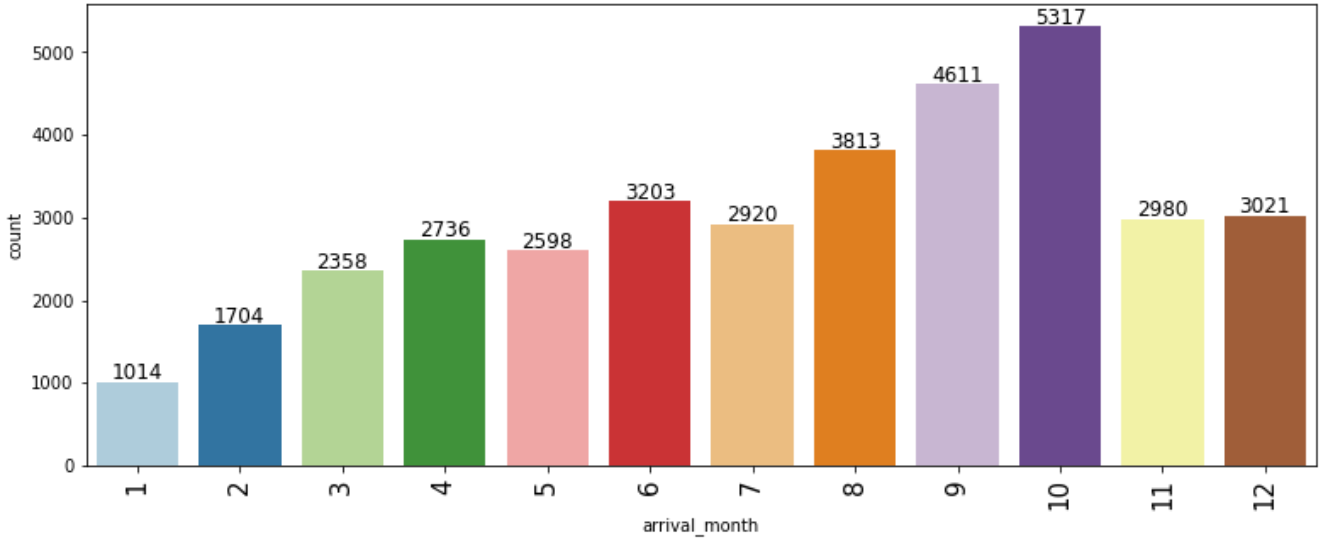


The most requested type of room is room 1, then followed by room type 4. The most common market that requests to stay is the guests who go online and offline to make a reservation.

Link to Appendix slide on data background check

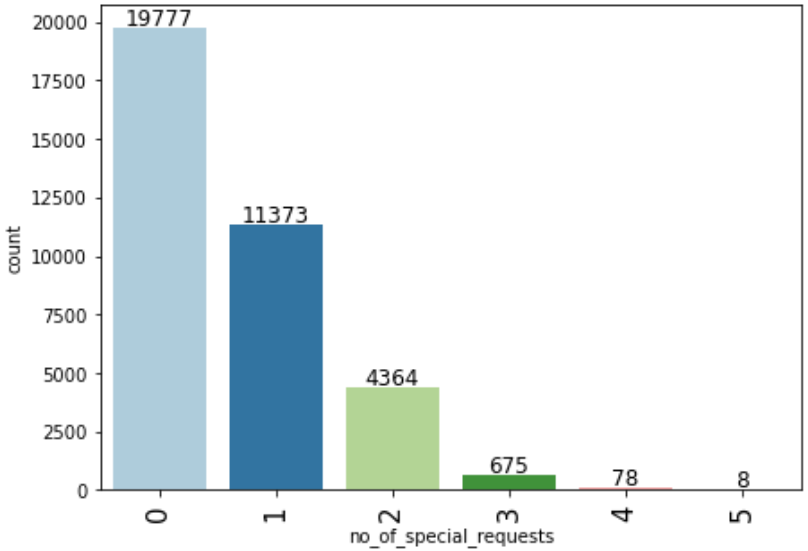
EDA Results – Arrival Month

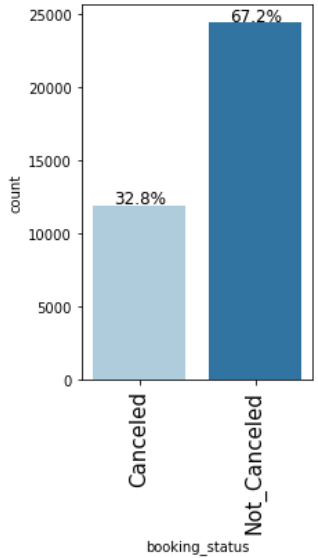
The most come month to stay in the hotel is in October. The second semester on the Hotel group is the one where they usually have the most demand for the services. January is the month with the least flow of guests of the year.



Link to Appendix slide on data background check

EDA Results – Special Request and Booking Status



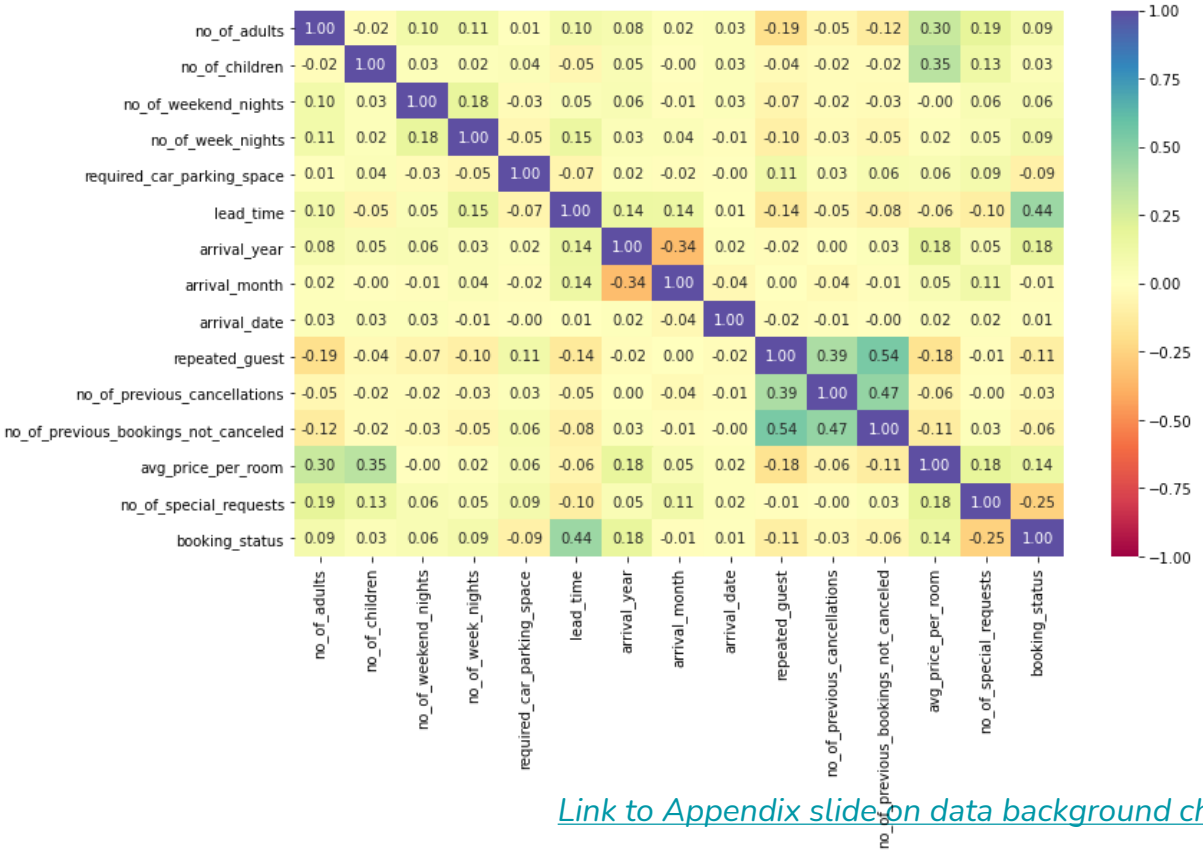


The client usually has at the must 1 special request for their rooms. The cases with over 3 special request are the exception. The cancelled reservations consist of almost 33% of all the reservations. This seems to be common for the hotel to expect a cancellation of the reservation made.

Link to Appendix slide on data background check

EDA Results – Heatmap

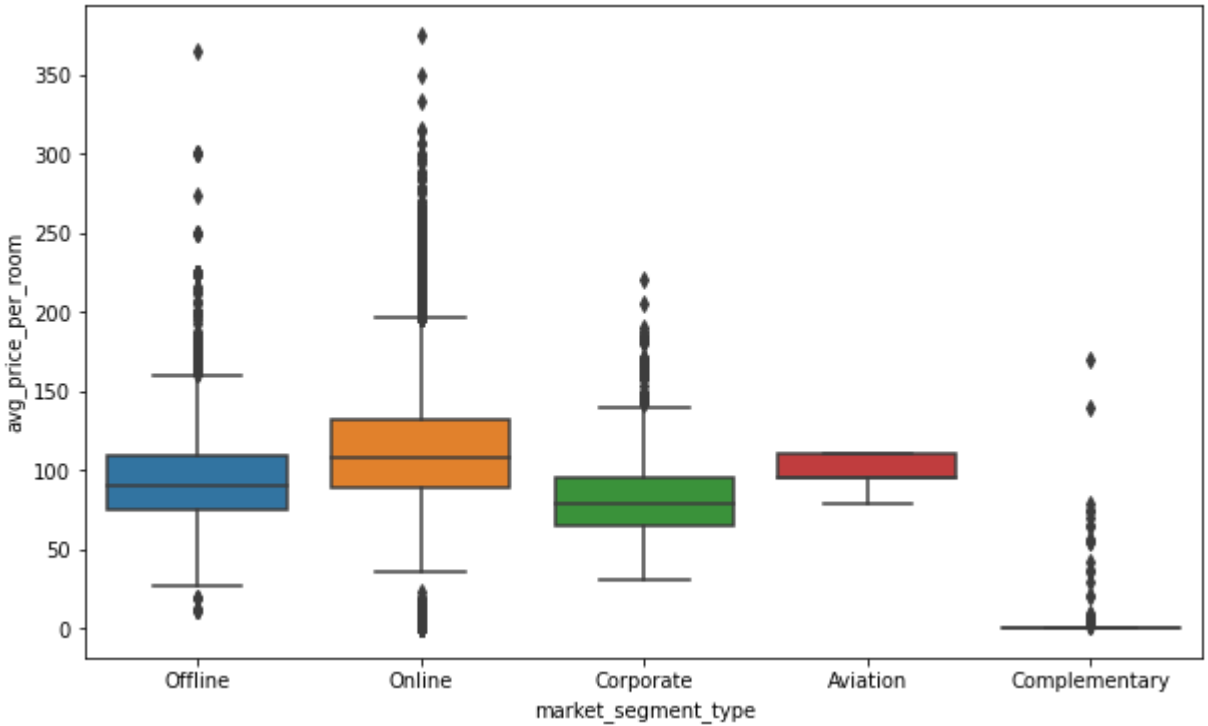
The price of the rooms shows growth with the total number of children. This can be expected as more rooms are requested with the children that come with the adults. Also, if the leading time is high, it shows a positive correlation with the stay of the guest being fulfilled.



Link to Appendix slide on data background check

EDA Results – Price VS Segment

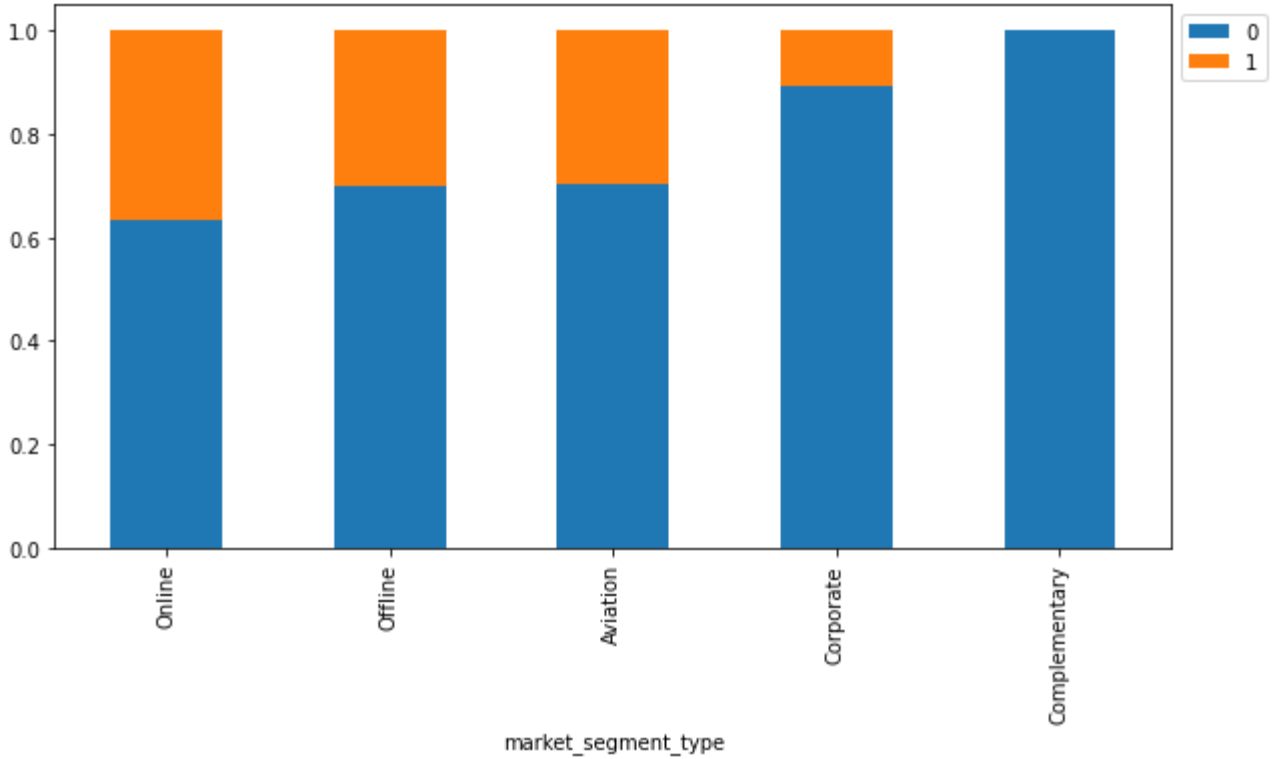
The prices for the different type of guest shows the prices stay in a similar range, on average near $90 - $105. The ones for the aviation companies show that negotiation was done to establish the price of the rooms.



Link to Appendix slide on data background check

EDA Results – Segment vs Cancelation

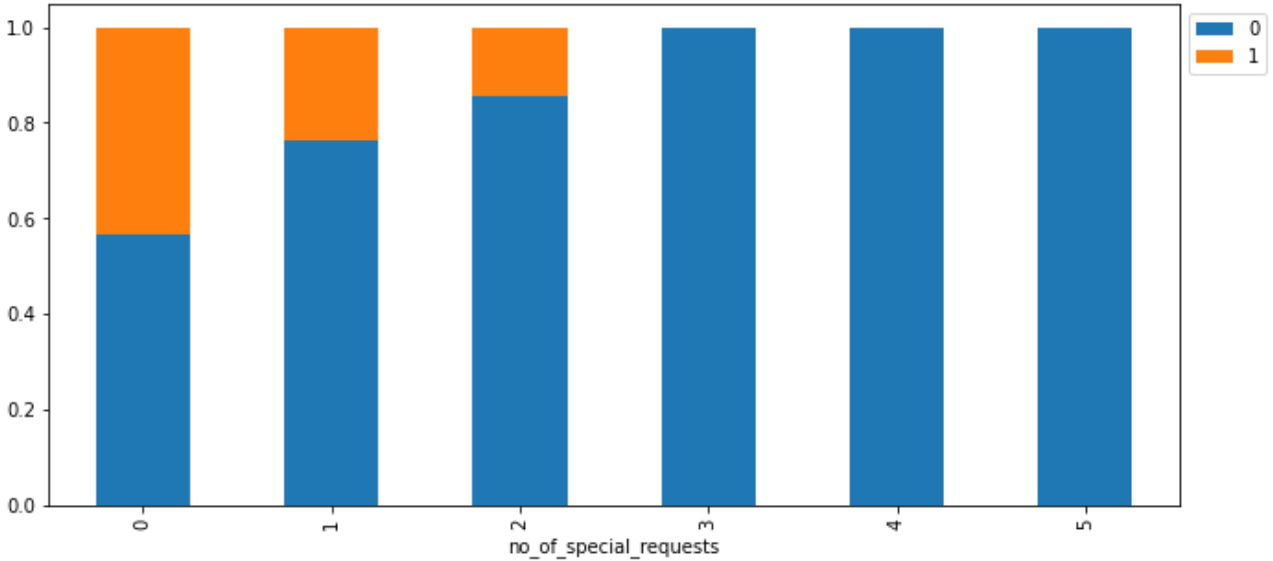
The most common cancellations occur in the online, offline and aviation segments. These segments have a similar level of cancellations compared to the corporate ones. The complementary segment has 0 since these are for the guests that are staying at the hotel.



Link to Appendix slide on data background check

EDA Results – Segment vs Cancelation

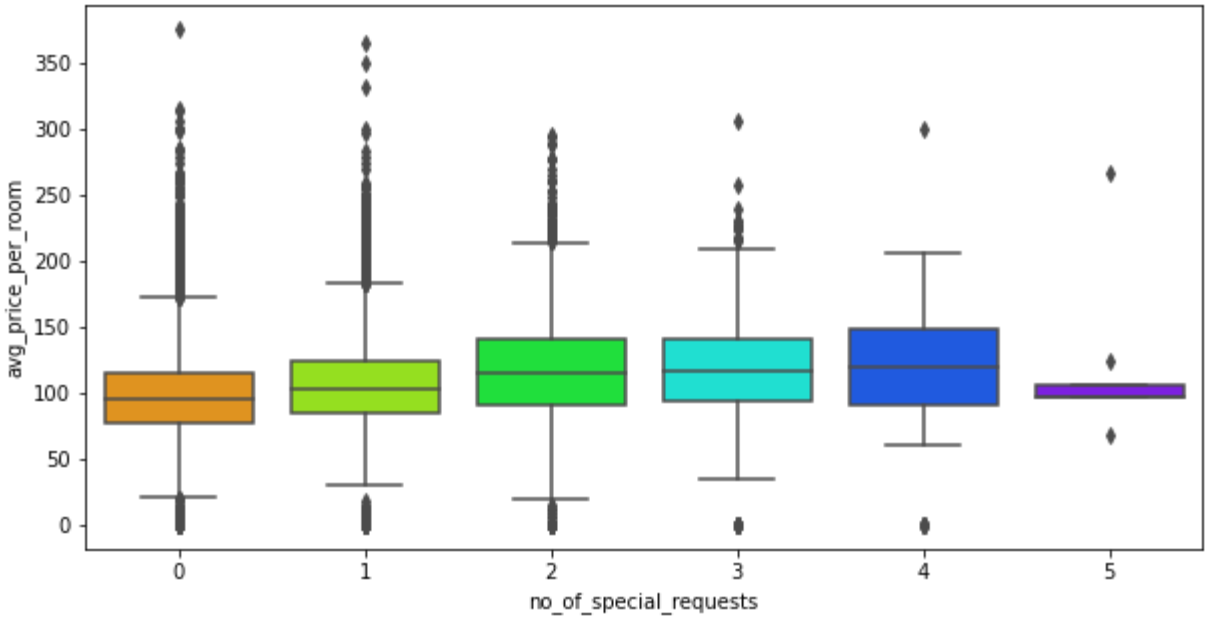
Based on the total number of special requests it shows a correlation between a cancellation and a successful reservation. Almost 43% of the rooms with no special request are cancelled. But realizing a special request is an indication of a probable successful reservation.



Link to Appendix slide on data background check

EDA Results – Segment vs Cancelation

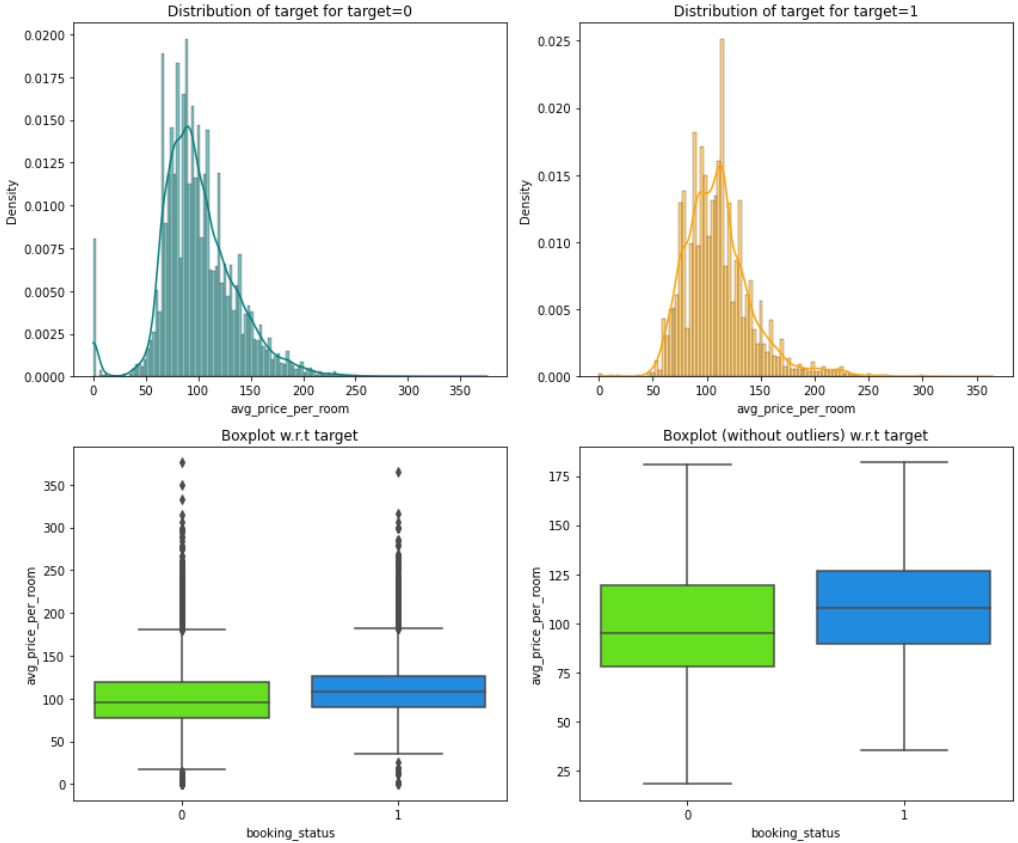
The price of the room with the special request doesn't show a significant variance. With the second request, the rest of the request shows the same price. There is no significant change of value on the room with the subsequent request.



Link to Appendix slide on data background check

EDA Results – Price vs Status

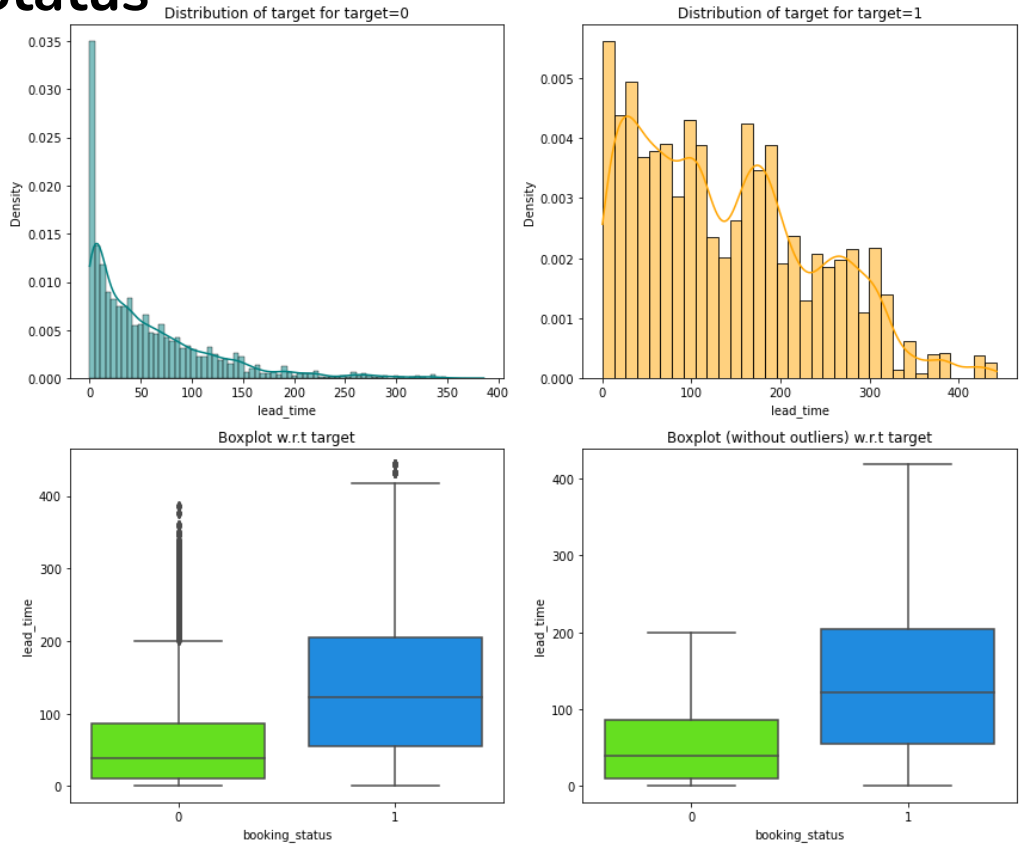
The distribution of cancelled rooms and successful reservations doesn't show a significant difference between the groups. In the boxplot without the outliers, the cancelled reservations show a slightly higher mean price for the room.



Link to Appendix slide on data background check

EDA Results – Lead time vs Status

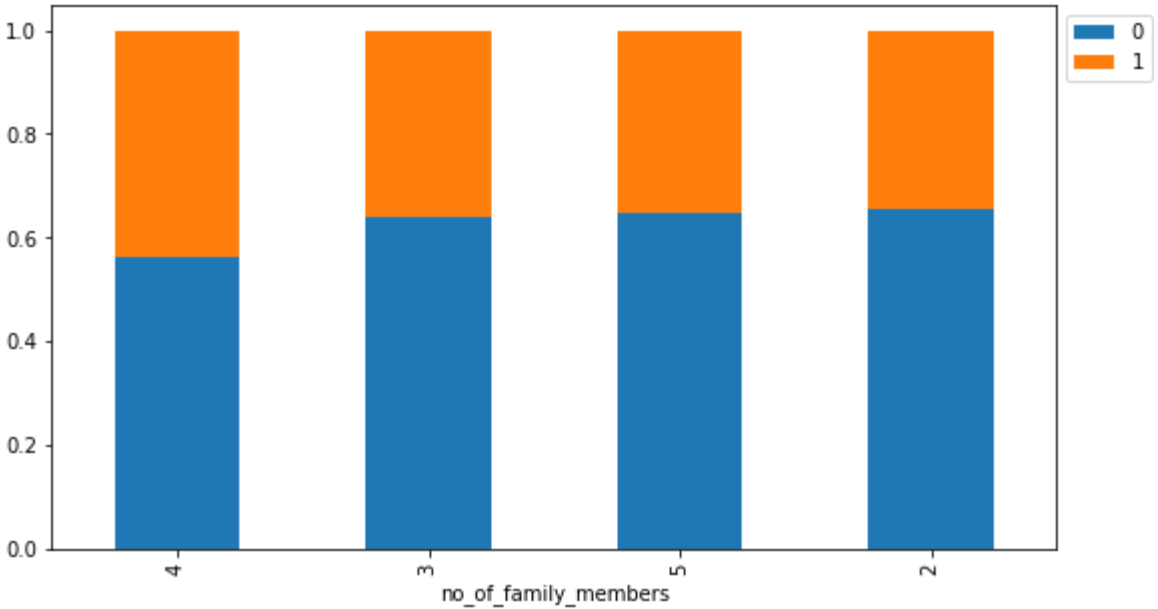
The distribution of cancelled rooms and successful reservations doesn't show a significant difference between the groups. In the boxplot without the outliers, the cancelled reservations show a slightly higher mean price for the room.



Link to Appendix slide on data background check

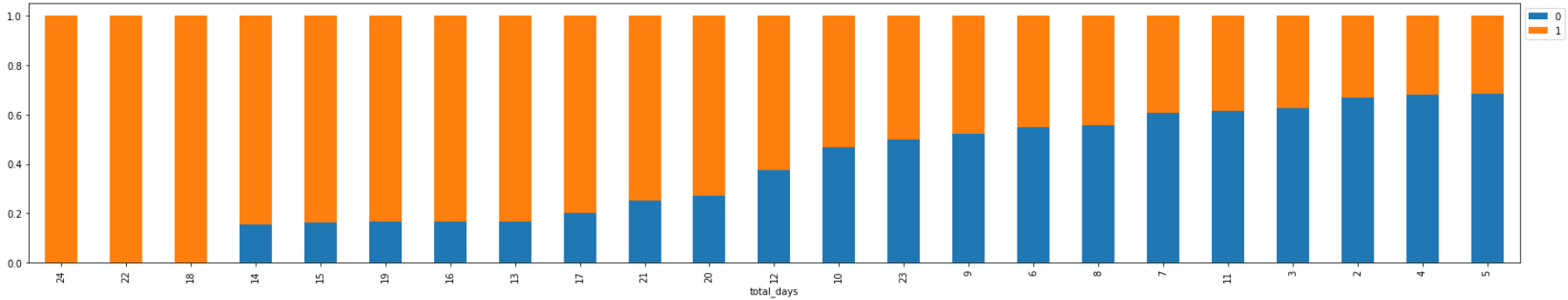
EDA Results – Segment vs Cancelation

Based on the total number of special requests it shows a correlation between a cancellation and a successful reservation. Almost 43% of the rooms with no special request are cancelled. But realizing a special request is an indication of a probable successful reservation.



Link to Appendix slide on data background check

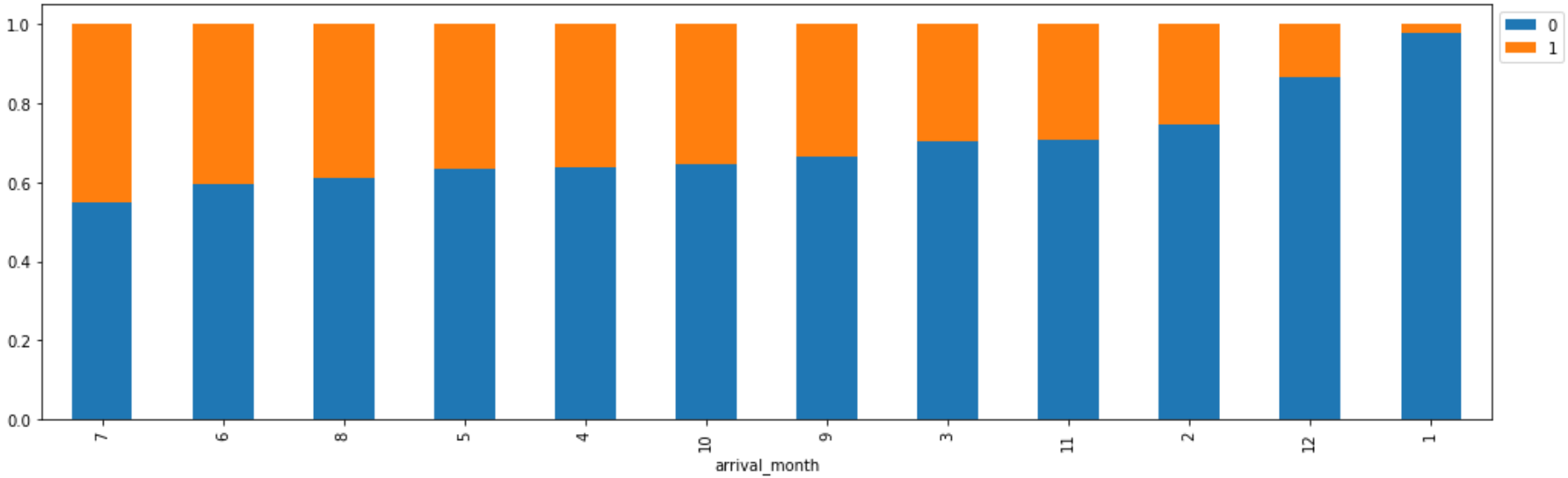
EDA Results – Total days vs Cancelation



Based on the information if a guest stays in a range between 2 – 5 days shows the guest tends to cancel the least number of reservations.

Link to Appendix slide on data background check

EDA Results – Month vs Cancelation

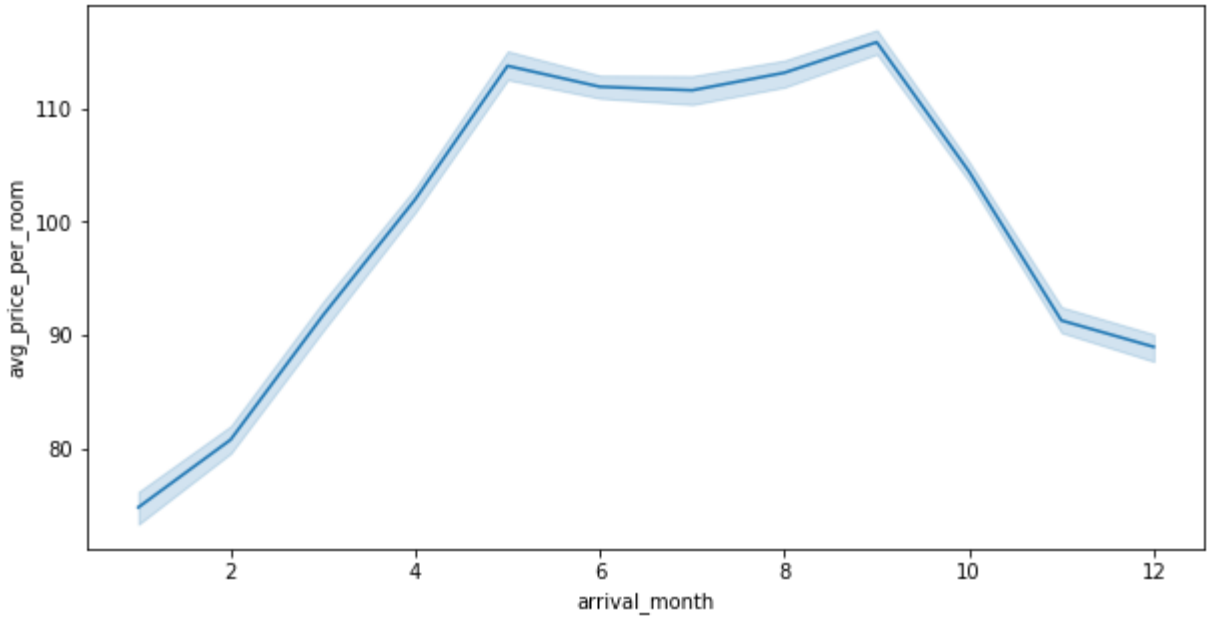


The month with fewer cancellations are January and December, both months have under 20 % of cancellations over the reservations. Julia is the worst month as it is near the 50% of cancellation over the reservations done. Finally, October being the month with more guests has an almost 40% of cancellations over the total reservations.

Link to Appendix slide on data background check

EDA Results – Segment vs Cancelation

The pricier months are from May to September. During October, the highest request month, prices began to fall in comparison to the previous month, around a 7%.



Link to Appendix slide on data background check

Data Preprocessing

- A Duplicate value check was realized, and the dataset shows no duplicated values.
- No Missing value treatment was realized as no variable had a missing value.
- A Outlier check was realized, but none of the variables was treated.
- Data preparation for modelling, the principal variable that was changed was avg_price_per_room, which was changed the value of the values over 500 to $ 179 .55 which is the result of 75th percentile plus 1.5 times the interquartile range.

Model Performance Summary – Considerations

- As part of the considerations realized for this model it was considered that either the false positive, guests that are predicted to stay but cancel and the false negative, the guests that stay but were considered to cancel their reservation. With this consideration, the final objective of the models is to reduce the f1 scores in order to reduce both the false positives and false negatives as both impacts the development of the business from a profit and reputational view.
- The month with the higher rate of cancellation also has the higher prices per room. The proposed policy of cancellation is a way to ensure revenue with the number of expected guests.

Link to Appendix slide on model assumptions

Model Performance Summary – Logistic Regression

The most important values for the logistic regression model are:

- Repeated guest
- Market segment is offline
- Requests a parking space
- The total special request realized

These variables even though are important during the evaluation of the variables it was shown that only a few guests needed a parking space or the number of special requests of 3 or more which were important factors to determine if a guest would fulfill the reservation.
The repeated guest is not common for the hotel. As was shown on the data the most common type of guest was a new one. Also, with the parking space, the vast majority didn't request any parking space.

```
Optimization terminated successfully.
        Current function value: 0.425731
        Iterations 11
                          Logit Regression Results
==============================================================================
Dep. Variable:         booking_status   No. Observations:                25392
Model:                          Logit   Df Residuals:                    25370
Method:                           MLE   Df Model:                           21
Date:                Thu, 24 Mar 2022   Pseudo R-squ.:                 0.3282
Time:                        17:12:10   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

For the results of the model, due to the classified nature of the problem, the different results of the model were compared with the use of the Accuracy, Recall, Precision and F1 parameters. The results are the following for the training and test data:

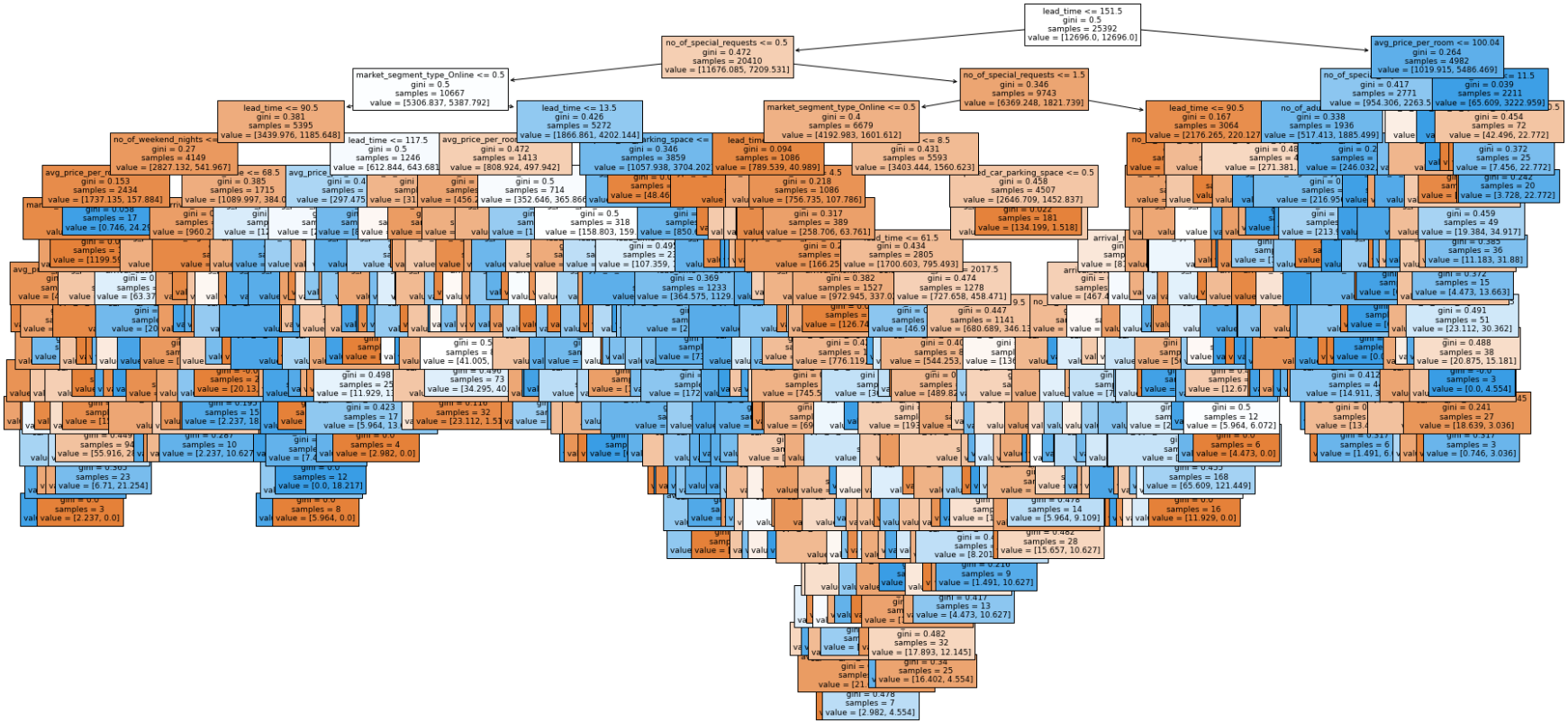
	0.5 Threshold	0.37 Threshold	0.42 Threshold
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73622	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868

	0.5 Threshold	0.37 Threshold	0.42 Threshold
Accuracy	0.80545	0.79555	0.80345
Recall	0.63267	0.73964	0.70358
Precision	0.73907	0.66573	0.69353
F1	0.68174	0.70074	0.69852

With the results obtained the Threshold was 0.37 as it has the highest F1 score in both the training and test data. Also, both F1 scores are almost very similar showing that the model was not overfitted.

Link to Appendix slide on model assumptions

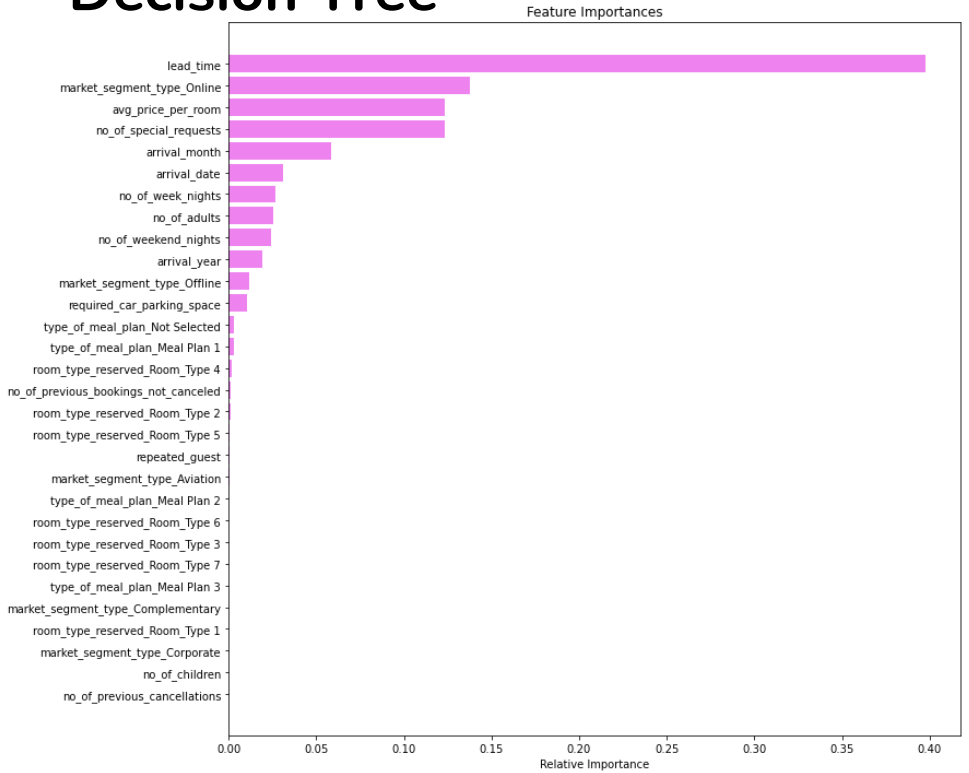
Model Performance Summary – Decision Tree



Link to Appendix slide on model assumptions

Model Performance Summary – Decision Tree

For the decision tree selected the most important factor is the lead time. As in the decision tree, this factor has the most relative importance. Then based on the segment if it is online, it has a higher consideration along with the number of special requests. As was seen the more special request the customer has the higher chance of fulfilling the reservation. Also, the average price per room was considered with a significant amount of importance. The arrival month has a significant impact on the result of the model as was seen some month has a higher probability of cancellation.



Link to Appendix slide on model assumptions

Model Performance Summary – Decision Tree

For the results of the model, due to the classification nature of the problem, the different results of the model were compared with the use of the Accuracy, Recall, Precision and F1 parameters. The results are the following for the training and test data:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83109	0.89438
Recall	0.98661	0.78608	0.89705
Precision	0.99578	0.72449	0.80468
F1	0.99117	0.75403	0.84835

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87182	0.83488	0.86778
Recall	0.80522	0.78308	0.85434
Precision	0.80000	0.72751	0.76468
F1	0.80260	0.75427	0.80703

With the results obtained the Post – Pruning the one with the highest F1 score in both the training and test data. This result will help us to have a more precise model

Link to Appendix slide on model assumptions

Business Recomendations

- Based on the results of the models provided by the ML analysis has shown that the variable of the number of special requests is a good indicator if a guest will fulfill its reservation. As was shown in the analysis of that variable one single special request showed a 20% increase in reservations completed versus no special requests. The Hotel can introduce in its processes of reservation the option of special requests that guests might require. This special treatment may want the guest to stay.

- As stated by the problem of the INN Hotel group one of their issues is to lose based on the request of the client. For this, it is recommended to establish policies of fees based on time before the day of arrival. The fees can be scaled based on the remaining days. This is particularly effective with the online market, as is most of the guests and can help to reduce the costs of the cancellations the guests realize.

Link to Appendix slide on model assumptions

Business Recomendations

- Based on the results of the models provided by the ML analysis has shown that the variable of the number of special requests is a good indicator if a guest will fulfill its reservation. As was shown in the analysis of that variable one single special request showed a 20% increase in reservations completed versus no special requests. The Hotel can introduce in its processes of reservation the option of special requests that guests might require. This special treatment may want the guest to stay.

- As stated by the problem of the INN Hotel group one of their issues is to lose based on the request of the client. For this, it is recommended to establish policies of fees based on time before the day of arrival. The fees can be scaled based on the remaining days. This is particularly effective with the online market, as is most of the guests and can help to reduce the costs of the cancellations the guests realize.

Link to Appendix slide on model assumptions

Business Recomendations

- The meal plans seem like an opportunity to boost sales. As most of the guests that select a meal plan is the type 1, which is breakfast. A price strategy can be made to make the rest of the meals cost less in comparison to the breakfast.

Link to Appendix slide on model assumptions

APPENDIX

Data Background and Contents - Dictionary

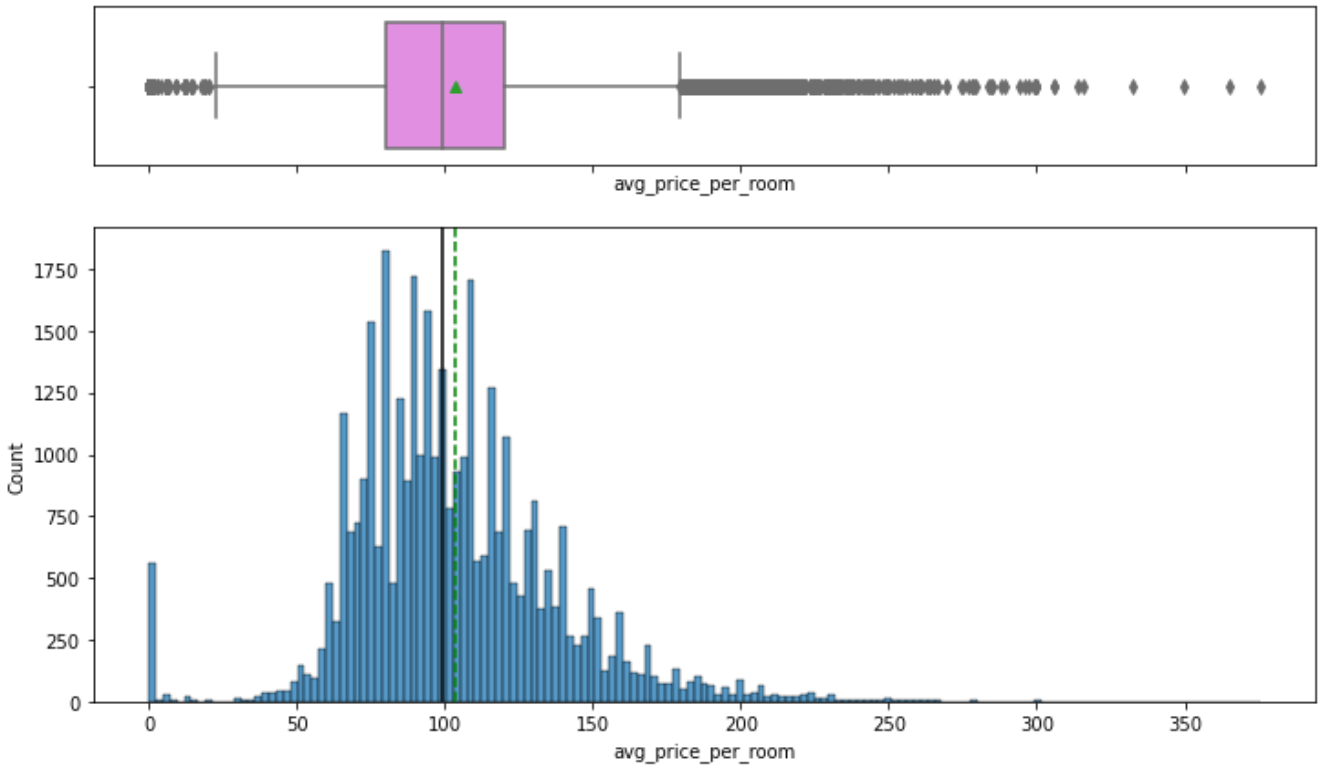
- Booking_ID: the unique identifier of each booking
- no_of_adults: Number of adults
- no_of_children: Number of Children
- no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- no_of_week_nights: Number of weeknights (Monday to Friday)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

Data Background and Contents

Object	Float64	int64
• Booking_ID • type_of_meal_plan • room_type_reserved • market_segment_type • booking_status	• avg_price_per_room	• 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_not_canceled • no_of_special_requests

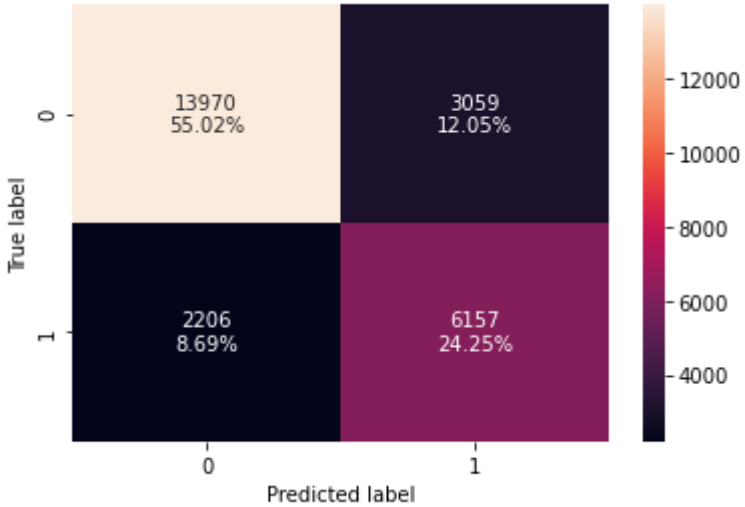
Model Assumptions

An update was realized to the avg_price_per_room to change the values higher of $ 500 to the upper whisker results as 1.5 times of the inter quartile range plus the 75th quantile of the data. The new value is $ 179.55.

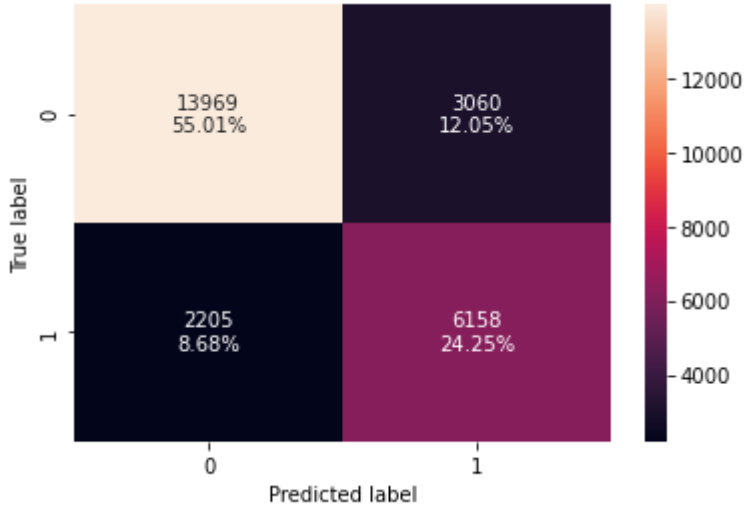


Performance of Logistic Model – 0.37 Threshold

- Training Set

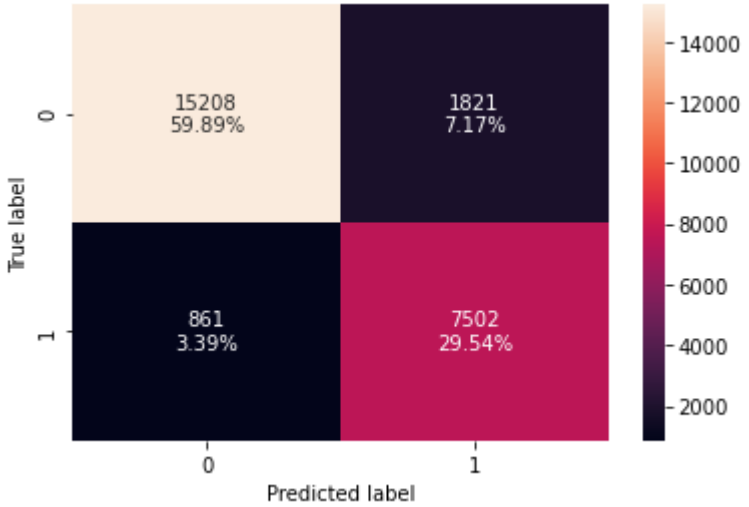


- Test Set

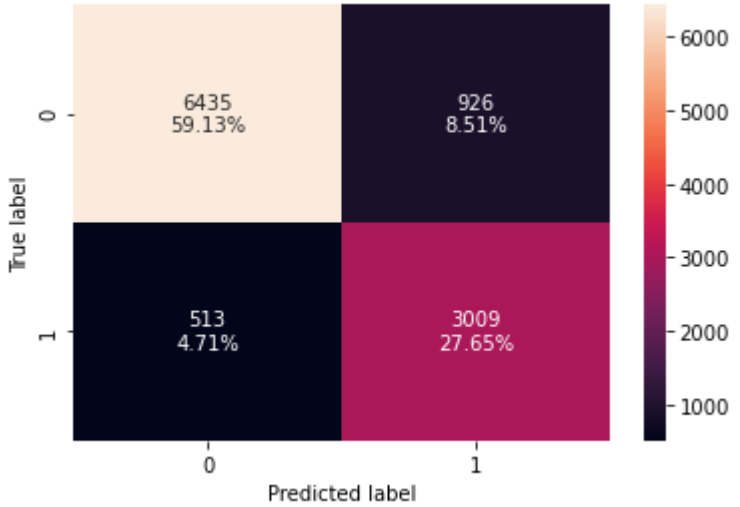


Performance of Decision Tree - Post Prunning

- Training Set



- Test Set



Great Learning

Happy Learning !