INN Hotels Booking Cancellation Prediction

PGP-DSBA _ INN Hotels Project

December 15, 2023

Contents / Agenda

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

Executive Summary (1/3): Conclusions on the Logistic Models

- A predictive logistic model has been built that permits the INN Hotel Management to determine, with an f1_score of 0.70 on the training set, whether a given client will cancel their booking or not and, thus, better inform management of bookings
- Among the three models assessed, on both the training and test data sets, the model with default threshold was the most accurate while the model with a threshold of 0.37 had the highest f1 score
- Increase in parking space requested, month of arrival (from January to December), and number of special requests tends to reduce the likelihood of a client cancelling their booking
- The likelihood of cancellation is increased by increase in the number of people (adults and children) involved, the number of days booked, and the lead time

Executive Summary (2/3): Conclusions on the Tree Classifier Models

- A predictive model has been built that permits the Hotel Management to determine, with an f1 score of 0.99 on the training set, the likelihood of a client cancelling their booking

- Among the three models assessed: On the training data set, the untuned model was the most accurate and had the highest f1 score; On the test data set, the untuned model was the most accurate whereas the post-tuned model had the highest f1 score

- We observed, from the decision tree, that when the lead time is between than 150 and 163 days, the average price per room is less than 100 euros, there are no special requests, and the market segment type is not online, then bookings made after May will most likely be cancelled

Executive Summary (3/3): General Conclusions and Recommendations

- Clearly, the tree classifier performs much better than the logistic regressor
- The Hotel Management should pay attention to the number of people, the number of days booked, the lead time, the average price per room, the number of special requests, the market segment type, and the month of arrival since these are important parameters that determine the likelihood of a client cancelling their bookings
- Management should consider exploring several other parameters that may help provide a better prediction of the likelihood of a client cancelling their booking
- On the part of the data science team, we will consider testing several other models and hyper parameter tuning in an effort to improve prediction performance

Business Problem Overview

- INN Hotels Group registered a significantly high number of reservation cancellations within its chain of hotels in Portugal
- In order to better manage registered bookings, my services, as a data scientist, was requested to analyze provided data on a number of bookings and help predict likelihood of any given booking to be cancelled
- My task consisted of building a predictive model that will help determine, with a satisfactory level of probability, the cancellation status of bookings made and help establish cost-effective policies for cancellations and refunds

Solution Approach

- The following steps were taken to accomplish the above-stated task:
  1. An exploratory data analysis was carried out on the data provided to uncover pre-modelling patterns
  2. Logistic regression models were then built with different probability thresholds and their performances compared
  3. Next, untuned, pre-tuned, and post-tuned tree classification models were built and their respective performances also compared
  4. Finally, the performances of the logistic regression and tree classification models were compared and general conclusions on some significant parameters established

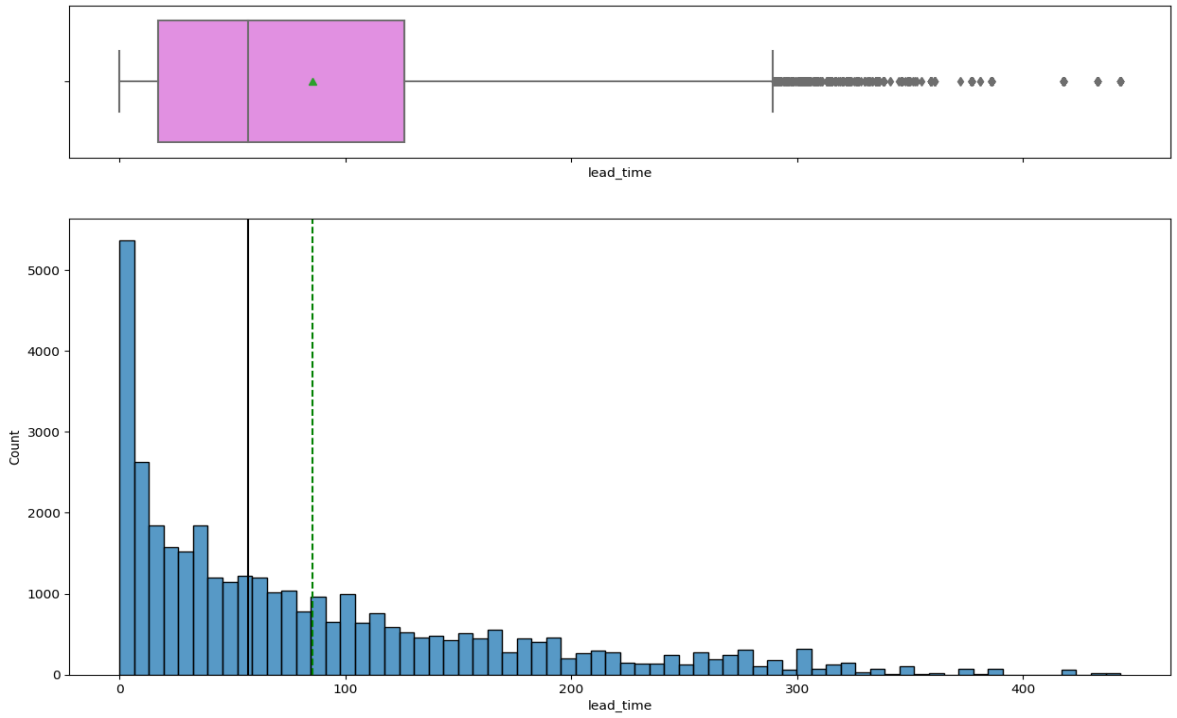
EDA Results: Statistical Summary of the Data

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
no_of_adults	36275.00000	NaN	NaN	NaN	1.84496	0.51871	0.00000	2.00000	2.00000	2.00000	4.00000
no_of_children	36275.00000	NaN	NaN	NaN	0.10528	0.40265	0.00000	0.00000	0.00000	0.00000	10.00000
no_of_weekend_nights	36275.00000	NaN	NaN	NaN	0.81072	0.87064	0.00000	0.00000	1.00000	2.00000	7.00000
no_of_week_nights	36275.00000	NaN	NaN	NaN	2.20430	1.41090	0.00000	1.00000	2.00000	3.00000	17.00000
type_of_meal_plan	36275	4	Meal Plan 1	27835	NaN	NaN	NaN	NaN	NaN	NaN	NaN
required_car_parking_space	36275.00000	NaN	NaN	NaN	0.03099	0.17328	0.00000	0.00000	0.00000	0.00000	1.00000
room_type_reserved	36275	7	Room_Type 1	28130	NaN	NaN	NaN	NaN	NaN	NaN	NaN
lead_time	36275.00000	NaN	NaN	NaN	85.23256	85.93082	0.00000	17.00000	57.00000	126.00000	443.00000
arrival_year	36275.00000	NaN	NaN	NaN	2017.82043	0.38384	2017.00000	2018.00000	2018.00000	2018.00000	2018.00000
arrival_month	36275.00000	NaN	NaN	NaN	7.42365	3.06989	1.00000	5.00000	8.00000	10.00000	12.00000
arrival_date	36275.00000	NaN	NaN	NaN	15.59700	8.74045	1.00000	8.00000	16.00000	23.00000	31.00000
market_segment_type	36275	5	Online	23214	NaN	NaN	NaN	NaN	NaN	NaN	NaN
repeated_guest	36275.00000	NaN	NaN	NaN	0.02564	0.15805	0.00000	0.00000	0.00000	0.00000	1.00000
no_of_previous_cancellations	36275.00000	NaN	NaN	NaN	0.02335	0.36833	0.00000	0.00000	0.00000	0.00000	13.00000
no_of_previous_bookings_not_canceled	36275.00000	NaN	NaN	NaN	0.15341	1.75417	0.00000	0.00000	0.00000	0.00000	58.00000
avg_price_per_room	36275.00000	NaN	NaN	NaN	103.42354	35.08942	0.00000	80.30000	99.45000	120.00000	540.00000
no_of_special_requests	36275.00000	NaN	NaN	NaN	0.61966	0.78624	0.00000	0.00000	0.00000	1.00000	5.00000
booking_status	36275	2	Not_Canceled	24390	NaN	NaN	NaN	NaN	NaN	NaN	NaN

- On average, about 2 adults are registered per booking suggesting that most of the bookings are made by couples
- More than 3/4 of the bookings register no children though some have had up to 10 children; it might be interesting exploring the details of bookings registering such large amount of children
- 3/4 of bookings register at most 2 weekend nights and 3 week nights; the maximum numbers of weekend and week nights being 7 and 17 respectively suggests that some bookings might be non-contiguous
- Most bookings register only breakfast as meal plan
- Most bookings include no request for car parking space
- Room_Type 1 is the most booked room type
- On average, rooms are booked 85 days before arrival; it might be interesting to explore bookings made over a year before arrival (443 days)
- The bookings were registered for arrivals in the years 2017 and 2018; and possibly every month and every day of the month
- Most bookings were made online and by new clients
- Most of the clients had never cancelled their bookings before; it might be interesting investigating previous cancellations as high as 13 and those having previous bookings as high as 58 - some client qualification and customer loyalty insights might be uncovered
- Room prices range between 0 and 540 Euros; interesting exploring cases of unusually low prices or rooms give for free; Might they be cases of loyalty offers, humanitarian gestures by the Chain Hotel Group, etc. ?
- Most clients made at most one special request
- Close to 70% of the bookings were not cancelled

Link to Appendix slide on data background check

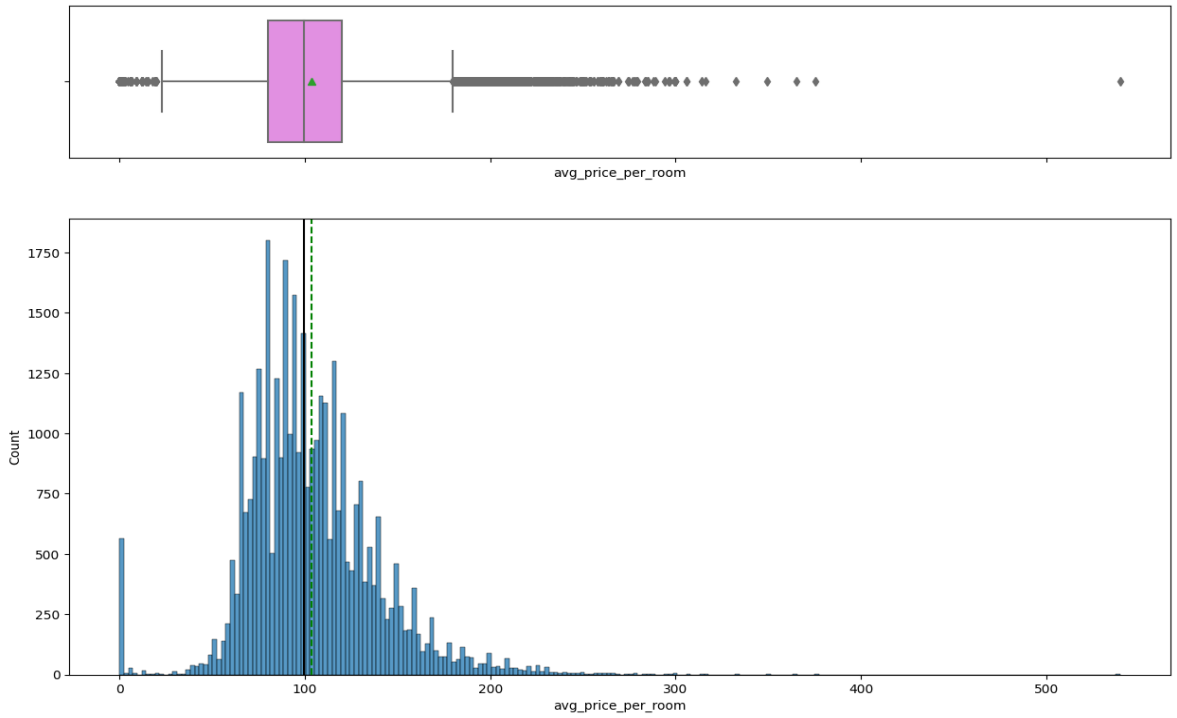
EDA Results: Univariate Analysis: Lead Time



- The lead time is right-skewed with several upper outliers
- The median and mean are approximately 60 and 80 days respectively

Link to Appendix slide on data background check

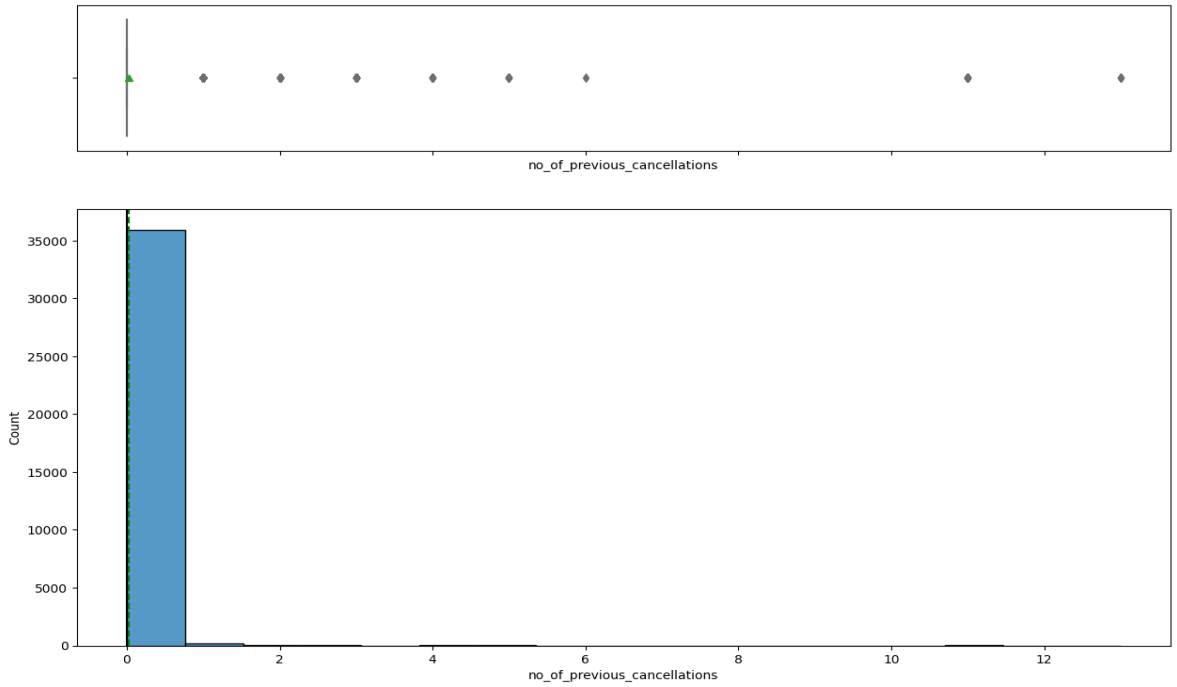
EDA Results: Univariate Analysis: Price Per Room



- The average price per room is right-skewed with several outliers on both side of the distribution; a significant number (over 500) of free or very low-priced bookings
- The distribution is quasi-normal for the most part and the median is about 100 euros, slightly lower than the average

Link to Appendix slide on data background check

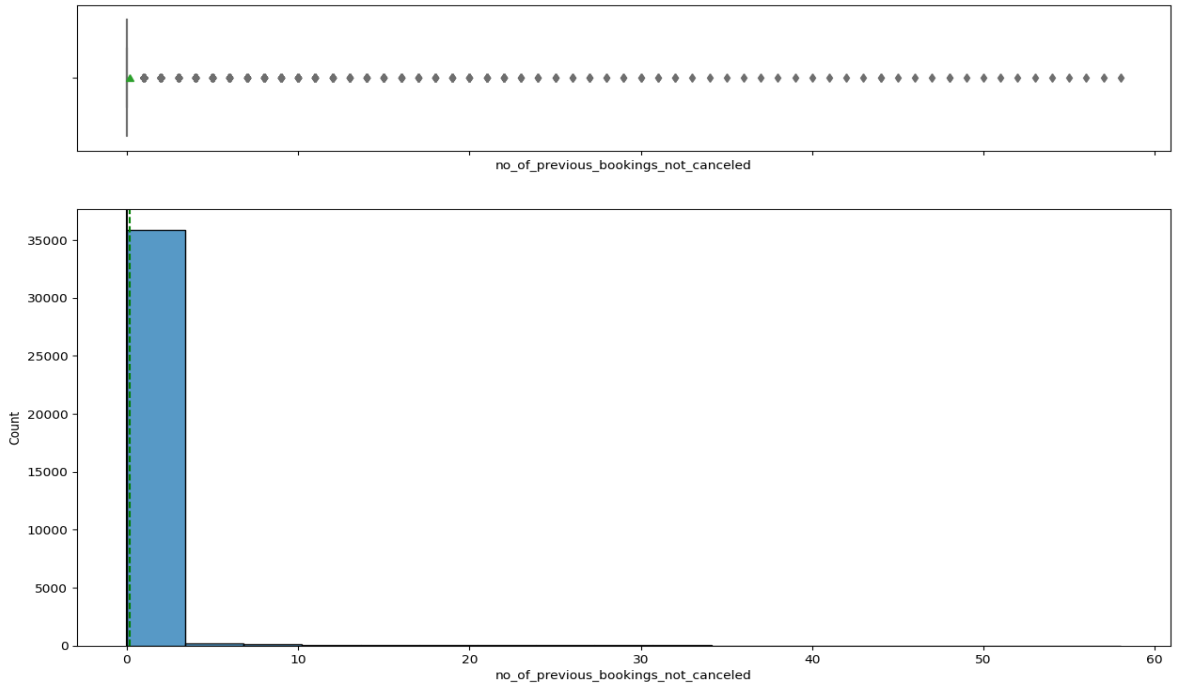
EDA Results: Univariate Analysis: Previous Booking Cancellations



- Almost all the clients surveyed had little or no previous cancellations
- The number of previous cancellations is right-skewed

Link to Appendix slide on data background check

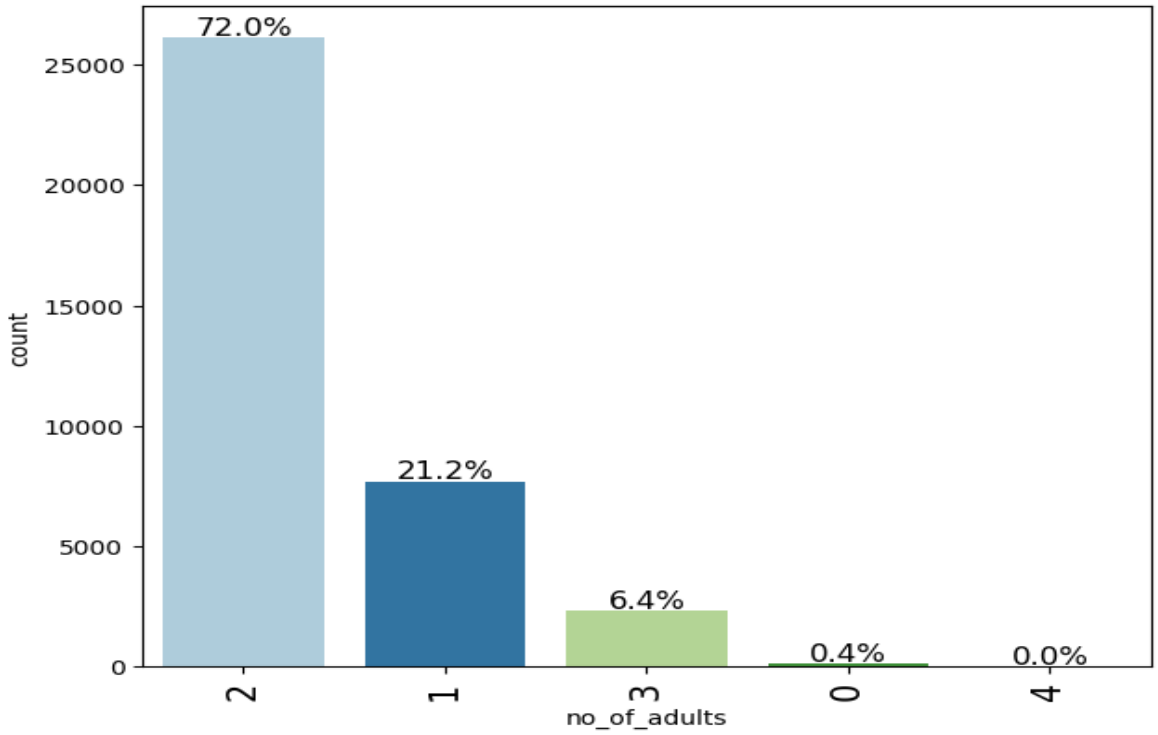
EDA Results: Univariate Analysis: Previous Booking Not Cancelled



- The large majority of clients also have little or no previous bookings not cancelled, indicating that most of the bookings were made by new clients
- The number of previous bookings not cancelled is also right-skewed

Link to Appendix slide on data background check

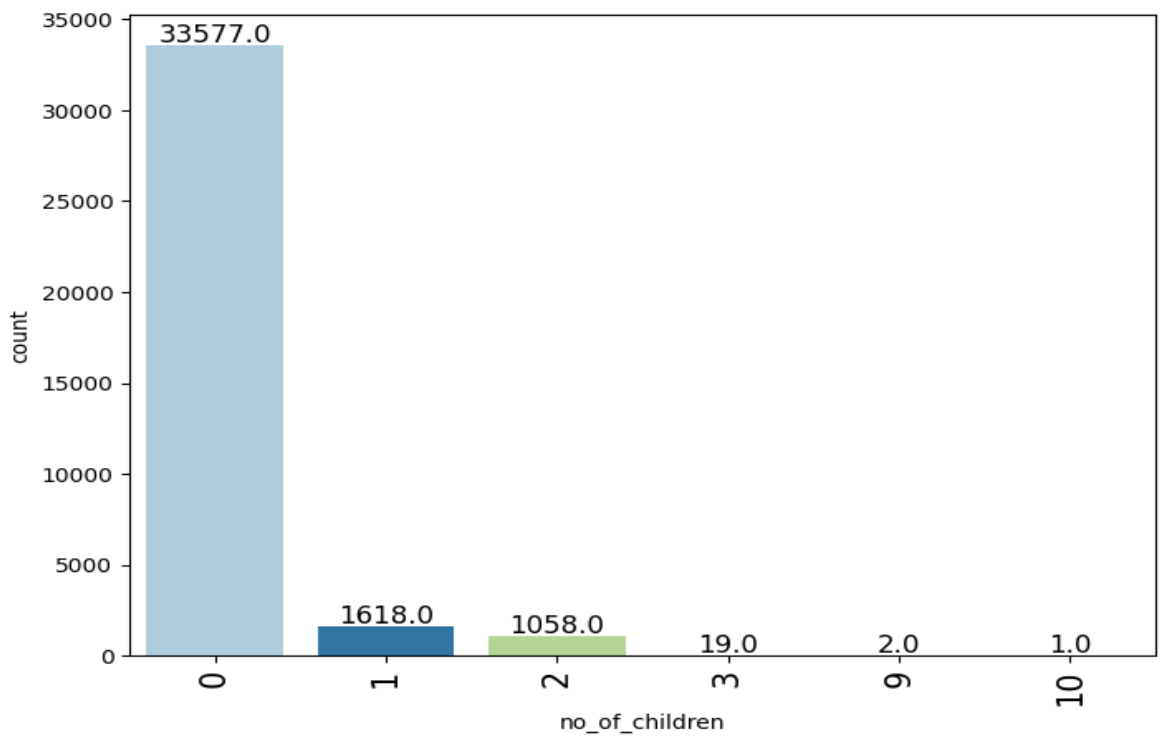
EDA Results: Univariate Analysis: Number of Adults



- We have confirmation that the majority (72%) of bookings were made by adult couples
- For about 21% of the bookings, single occupants were declared
- A very small percentage (0.4%) of the bookings were made by non-adults (logically children) and a much smaller percentage registered 4 adults per booking

Link to Appendix slide on data background check

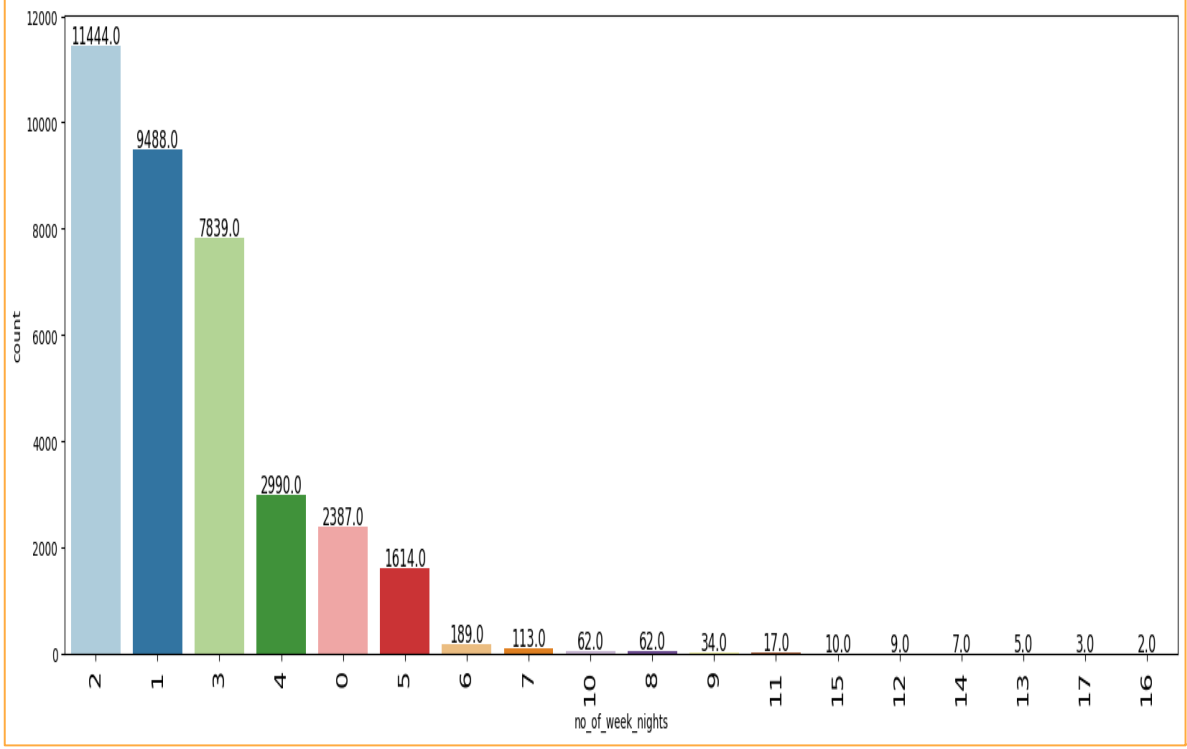
EDA Results: Univariate Analysis: Number of Children



- About 93% of the bookings registered no children
- 3 bookings (Close to 0%) registered 9 to 10 children

Link to Appendix slide on data background check

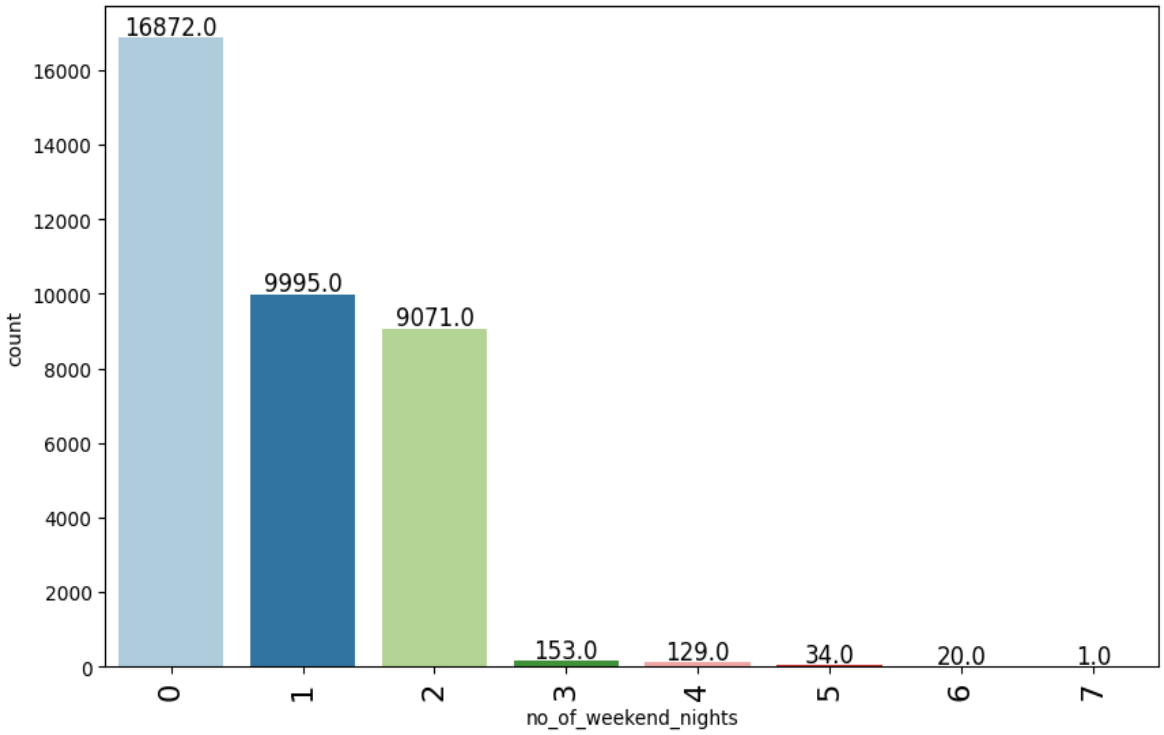
EDA Results: Univariate Analysis: Number of Week Nights



- The greatest (32%) of week nights registered per booking is 2, followed by 1 (26%), suggesting that most of the bookings were made for short stays
- About 0% of the bookings registered above 10 week nights

Link to Appendix slide on data background check

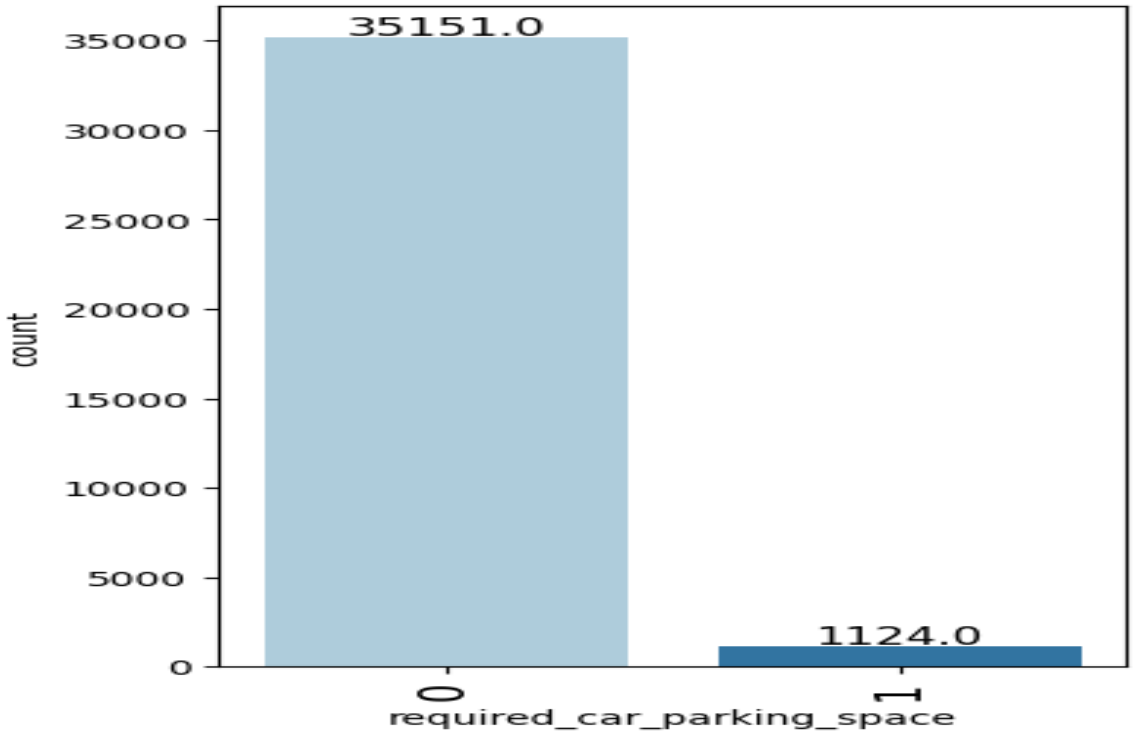
EDA Results: Univariate Analysis: Number of Weekend Nights



- Approximately 47% of the bookings included no weekend nights, 28% 1, and 25% 2 weekend nights
- The remaining less than 1% of bookings registered 3 to 7 weekend nights

Link to Appendix slide on data background check

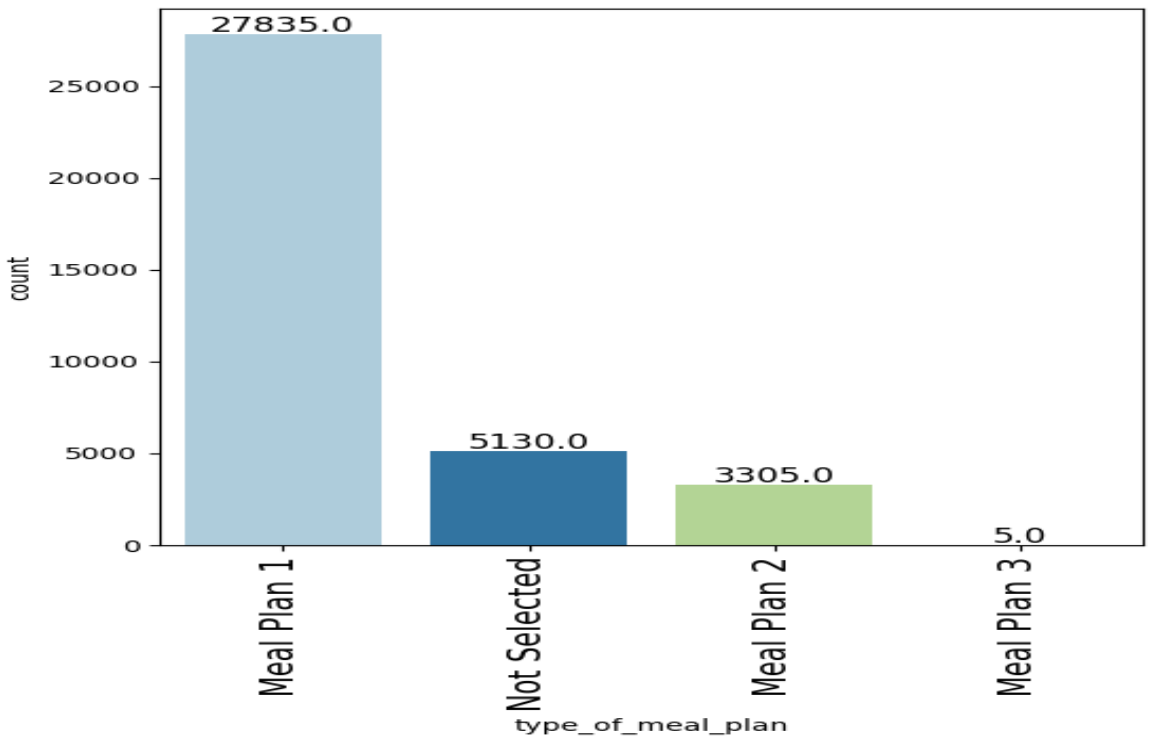
EDA Results: Univariate Analysis: Required Car Parking Space



- 97% of bookings were made by clients who required no car parking space and no client required more than 1 car parking space

Link to Appendix slide on data background check

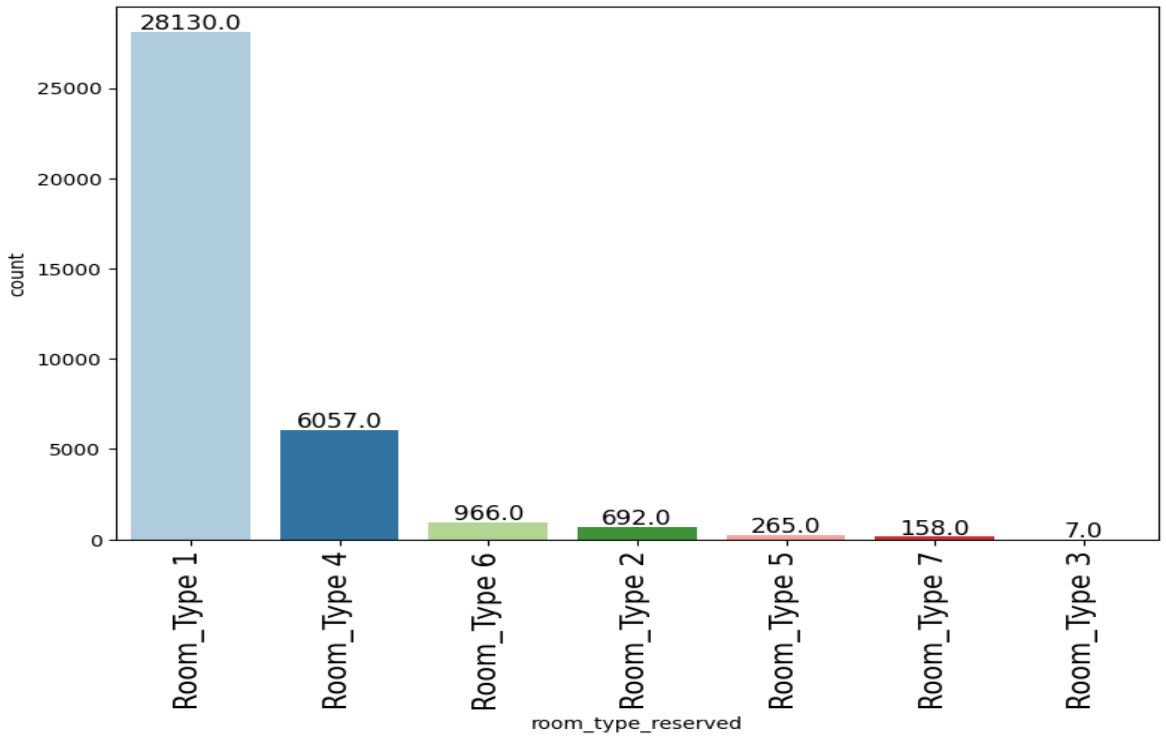
EDA Results: Univariate Analysis: Type of Meal Plan



- 77% of the bookings were made by clients who chose breakfast only as meal plan
- 14% selected no meal plan and close to 0% opted for breakfast, lunch, and dinner

Link to Appendix slide on data background check

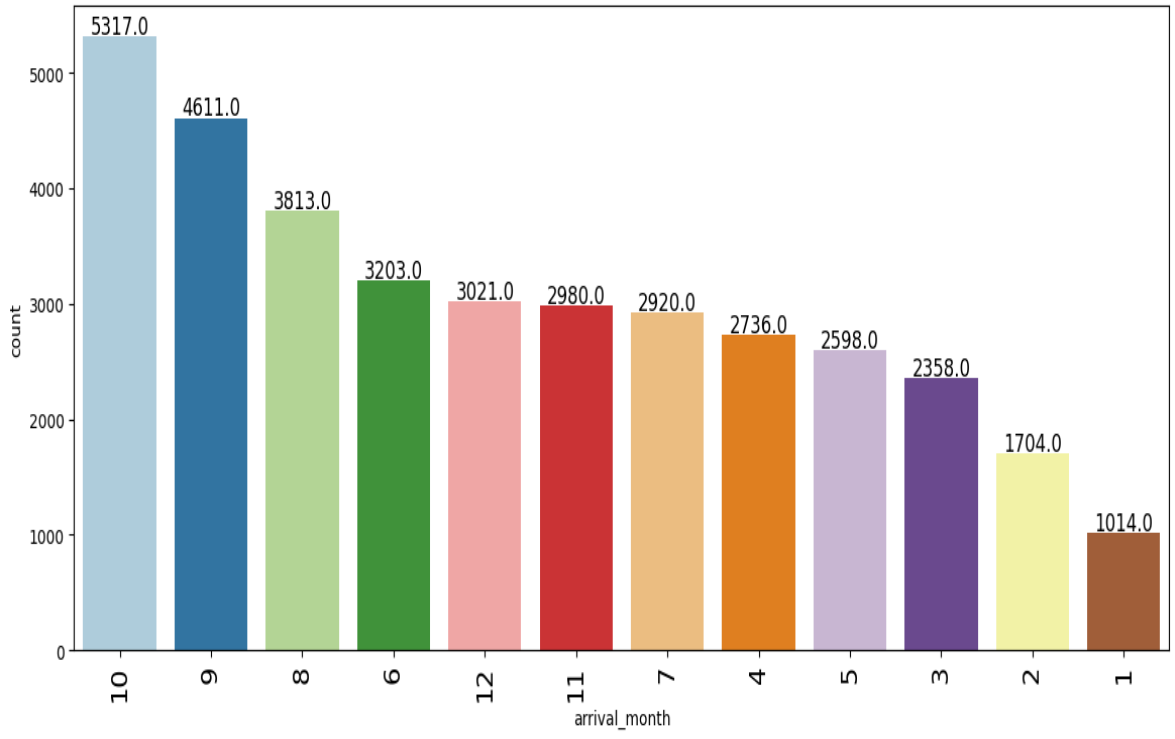
EDA Results: Univariate Analysis: Room Type Reserved



- 78% of the bookings registered Room Type 1
- Close to 0% registered Room Type 3

Link to Appendix slide on data background check

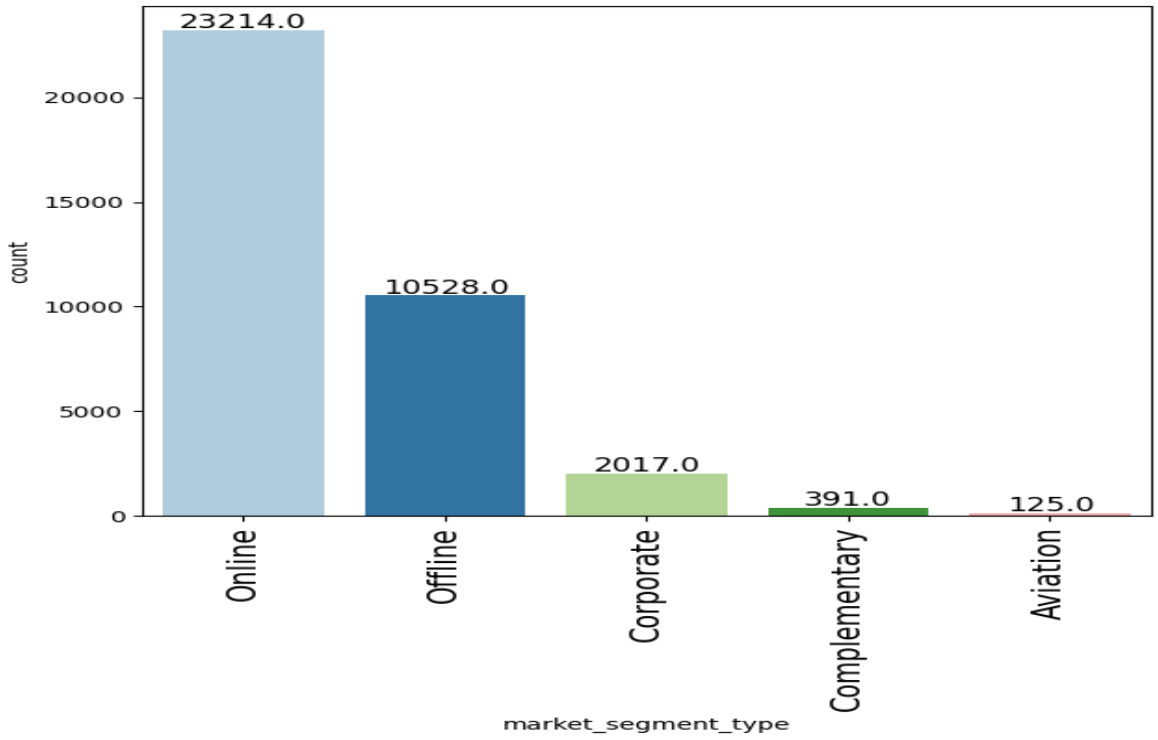
EDA Results: Univariate Analysis: Arrival Month



- Most frequently registered arrival month is October (15%)
- The least frequently registered month is January (3%)

Link to Appendix slide on data background check

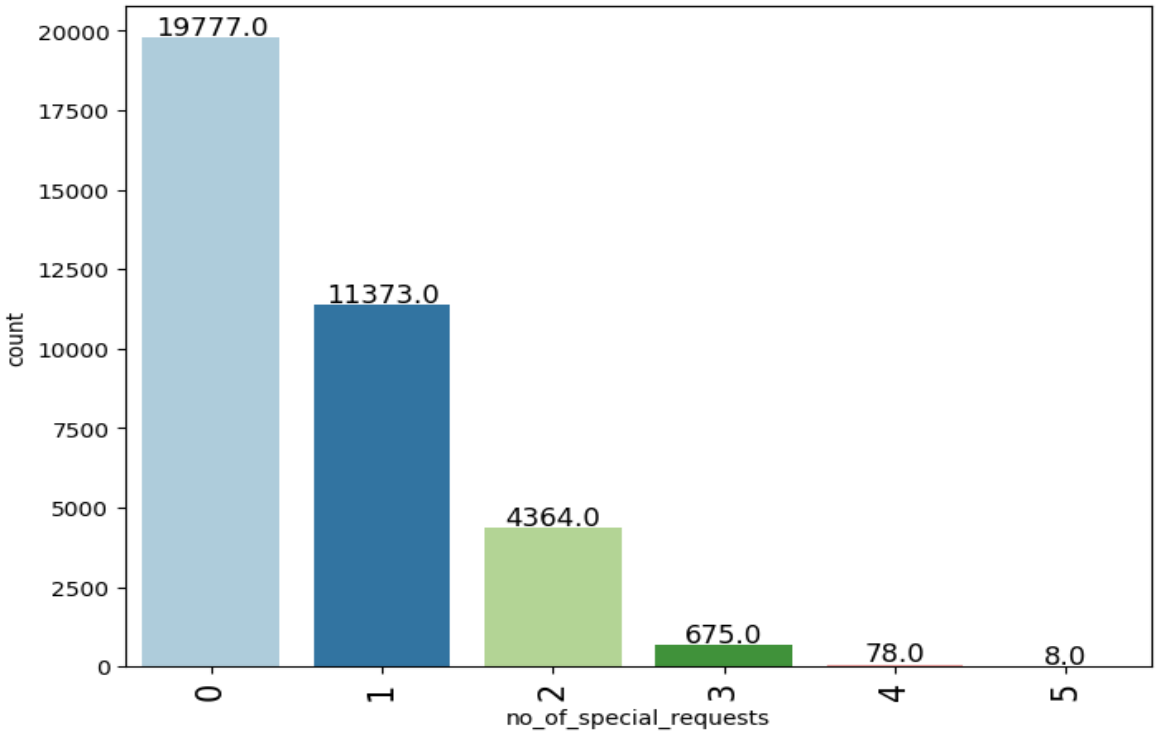
EDA Results: Univariate Analysis: Market Segment Type



- 64% of the bookings were registered from the online market segment
- The least number of bookings were registered from the aviation market segment (0.3%)

Link to Appendix slide on data background check

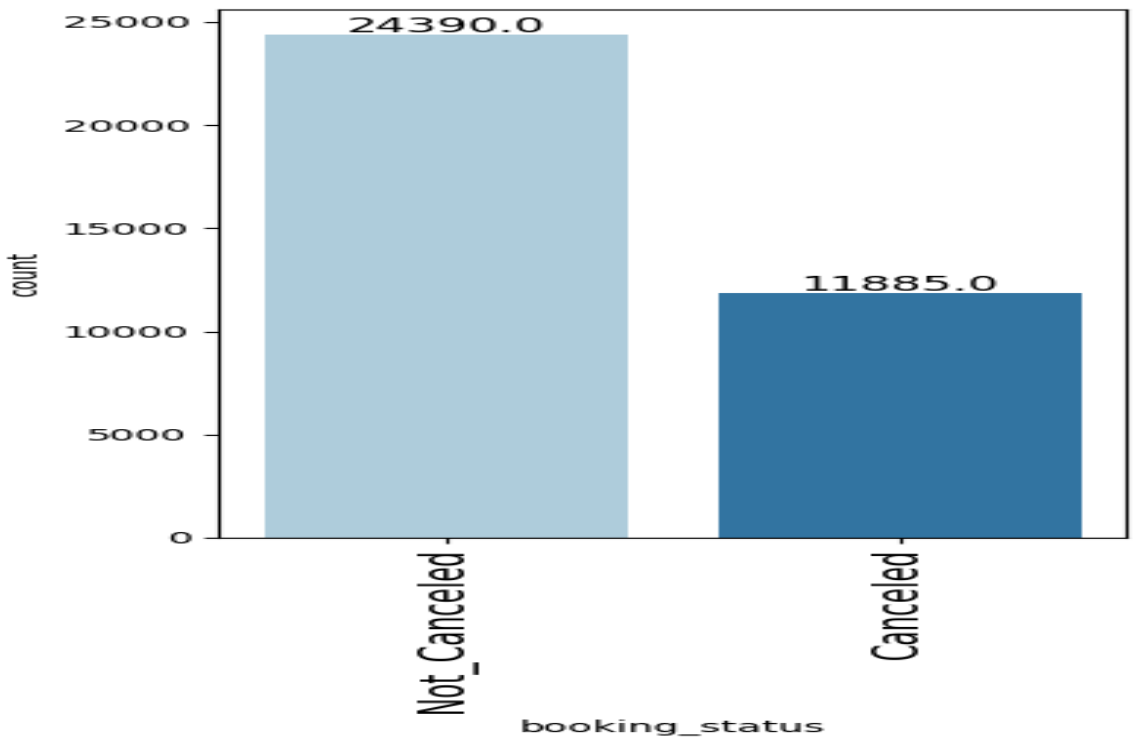
EDA Results: Univariate Analysis: Number of Special Requests



- 54% of the bookings came from clients who made no special requests
- 31% made 1 special request and 0.2% made 4 to 5 special requests

Link to Appendix slide on data background check

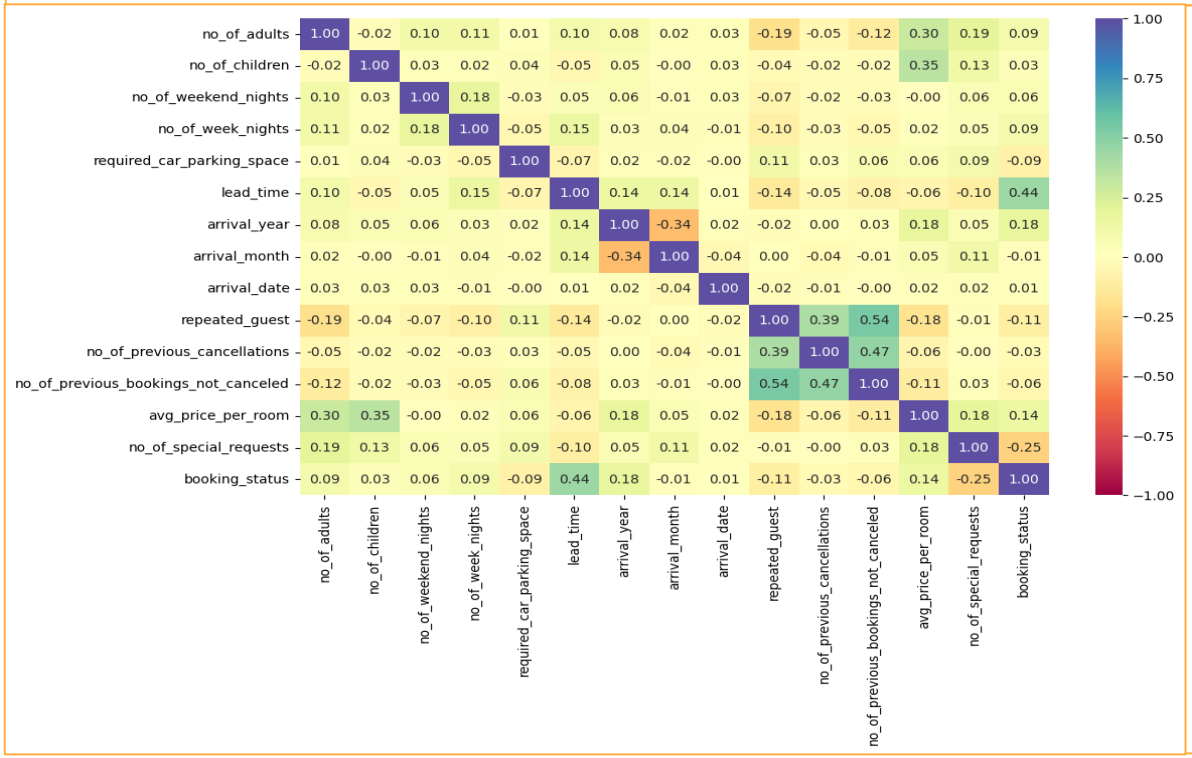
EDA Results: Univariate Analysis: Booking Status



- 67% of the bookings were not cancelled

Link to Appendix slide on data background check

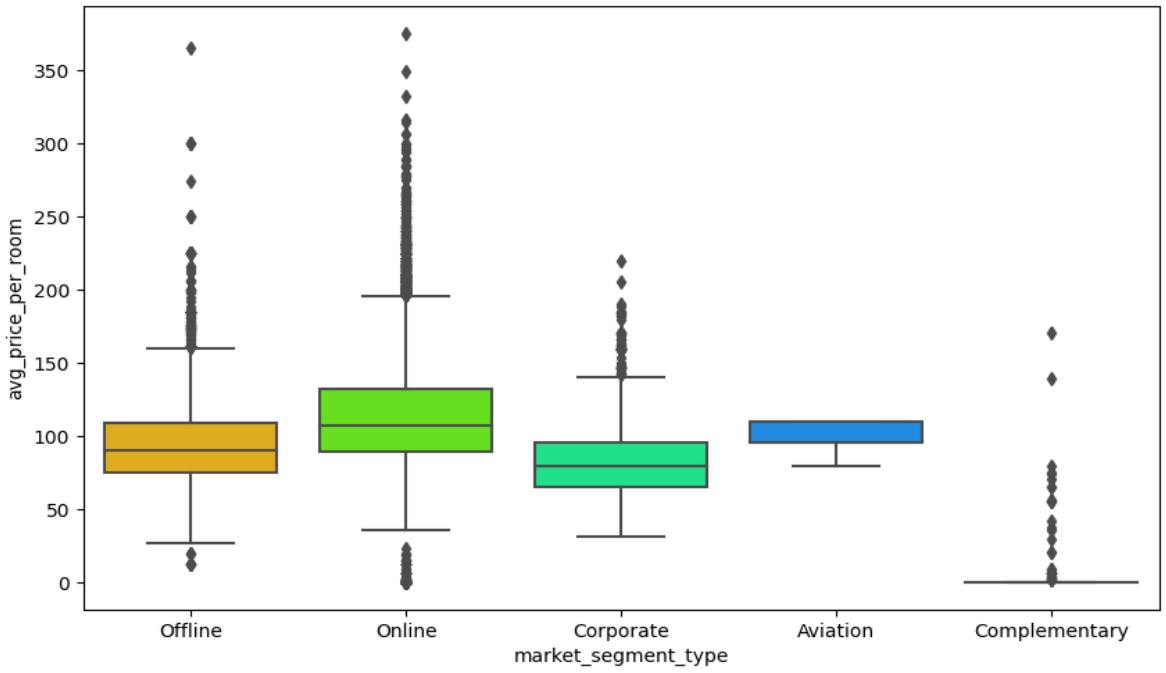
EDA Results: Bivariate Analysis: Correlation (Heatmap)



- The most significant correlation (0.54) registered is between the repeated guest status and the number of previous bookings not cancelled
- The booking status is most correlated (0.44) with the lead time, suggesting that there is a higher probability that a booking is cancelled for greater lead times

Link to Appendix slide on data background check

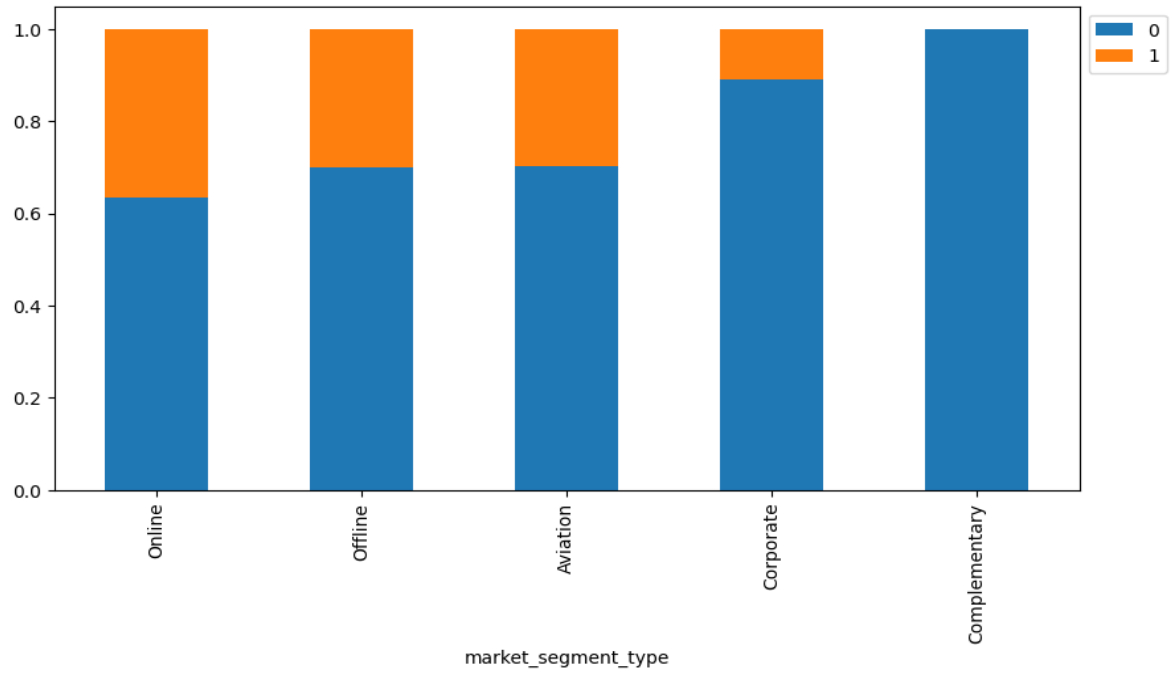
EDA Results: Bivariate Analysis: Average Price Per Room vs Market Segment Type



- The online market segment registered the most varied and tended to register the greatest average price per room
- The complementary market segment registered the least varied and tended to register the least average price per room

Link to Appendix slide on data background check

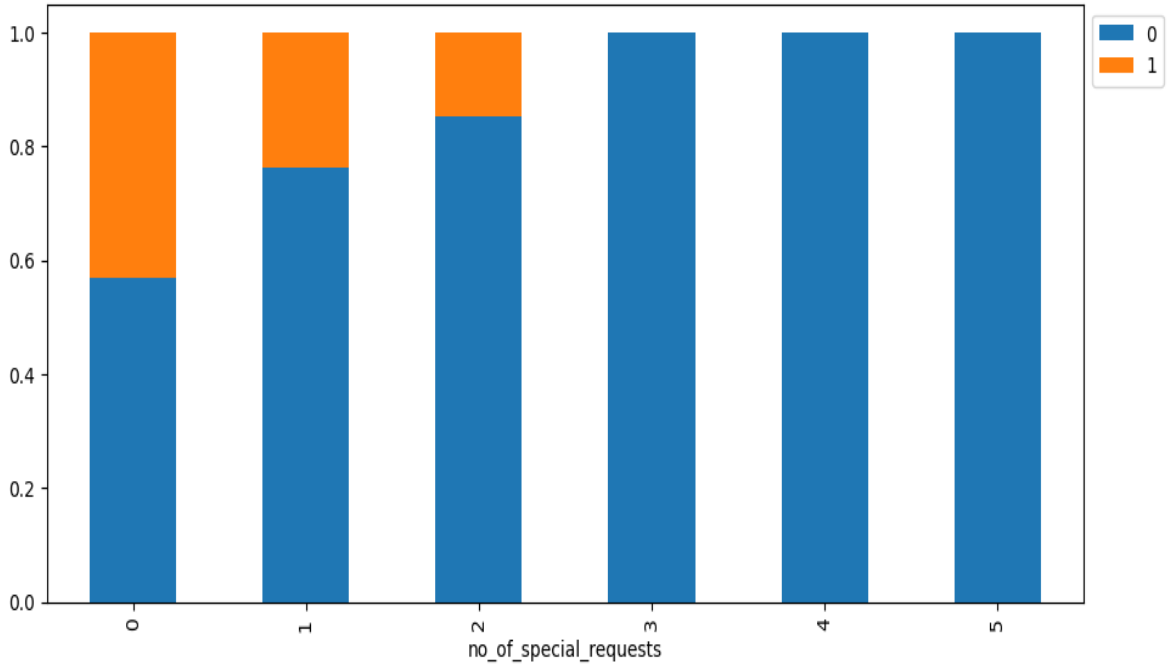
EDA Results: Bivariate Analysis: Booking Status vs Market Segment Type



- The online market segment registered the greatest percentage of cancellations meanwhile the complementary market segment had no cancellations

Link to Appendix slide on data background check

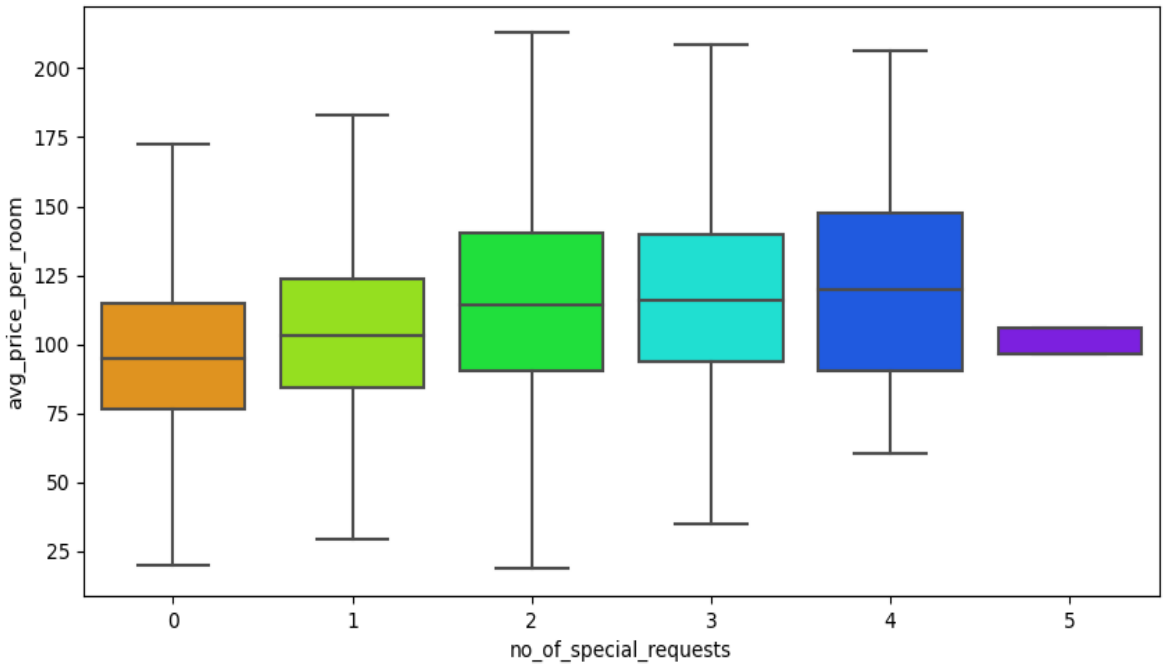
EDA Results: Bivariate Analysis: Booking Status vs Number of Special Requests



- The greatest percentage of cancellations was registered for bookings with no associated special requests whereas no cancellations were made for bookings with more than 2 special requests

Link to Appendix slide on data background check

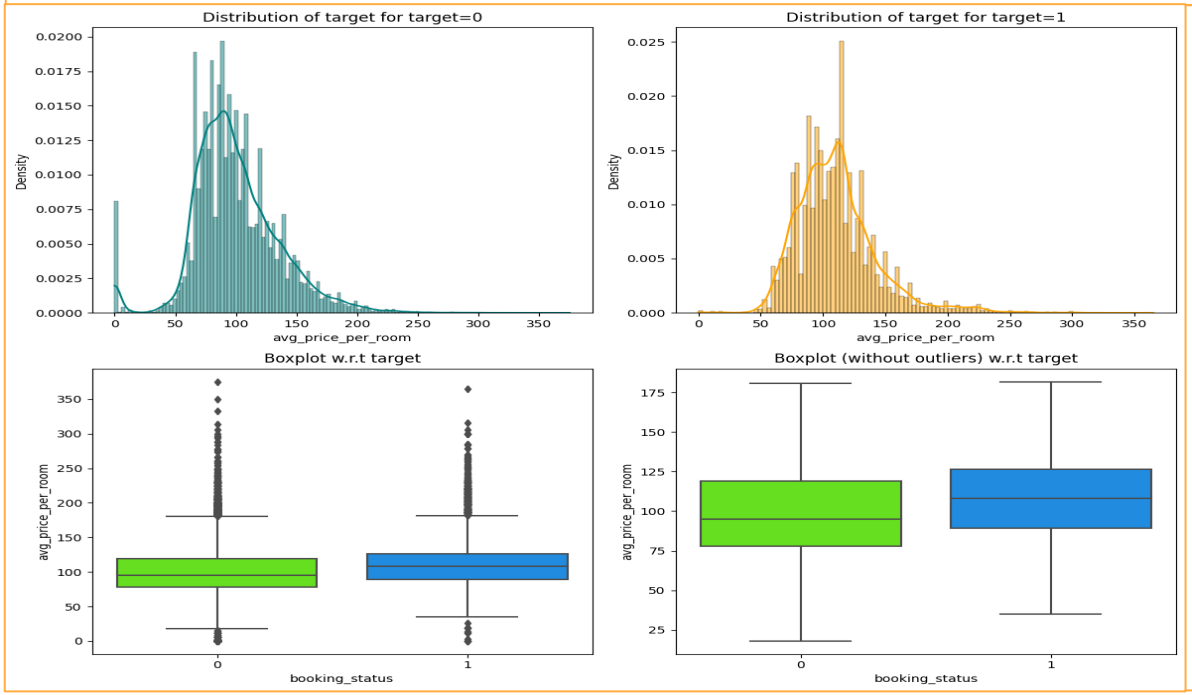
EDA Results: Bivariate Analysis: Average Price Per Room vs Number of Special Requests



- Bookings with 4 special requests tended to have the highest and most varied average price per room
- Bookings with 5 special requests had the least varied and, together with bookings having no special request, tended to have the least average price per room

Link to Appendix slide on data background check

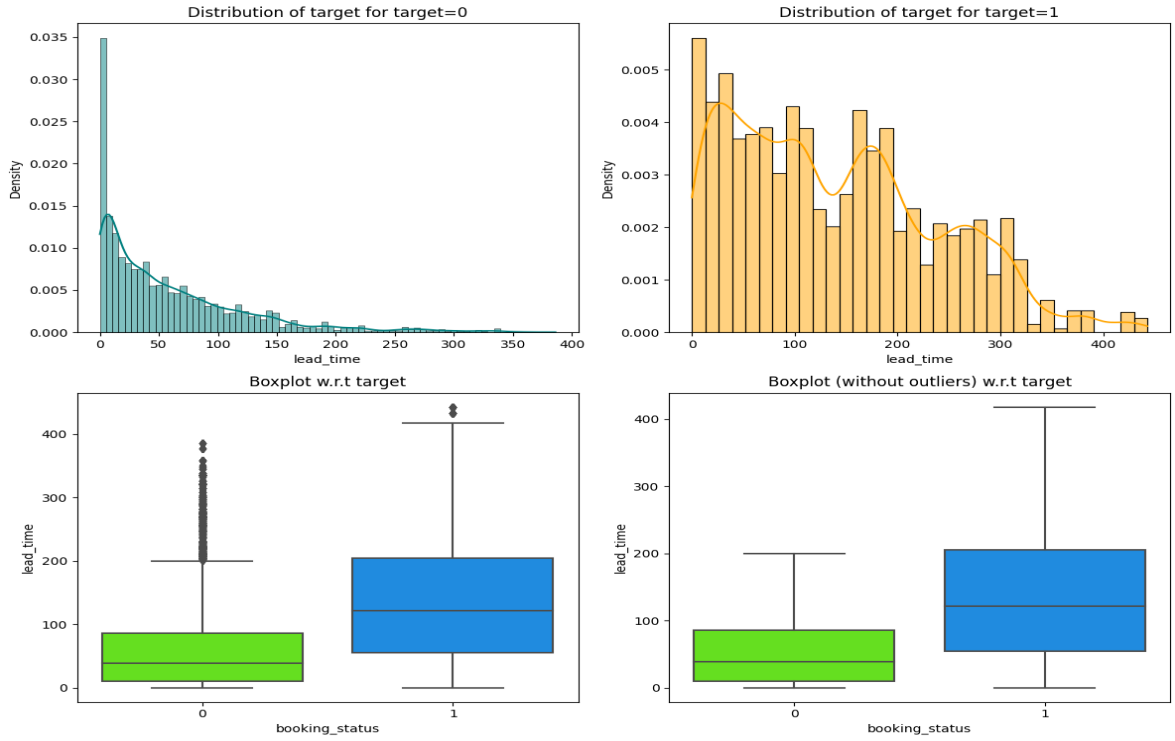
EDA Results: Bivariate Analysis: Average Price Per Room vs Booking Status



- The average price per room is right skewed for both cancelled and not cancelled bookings
- The average price per room tends to be about 15 euros higher for cancelled than not cancelled bookings

Link to Appendix slide on data background check

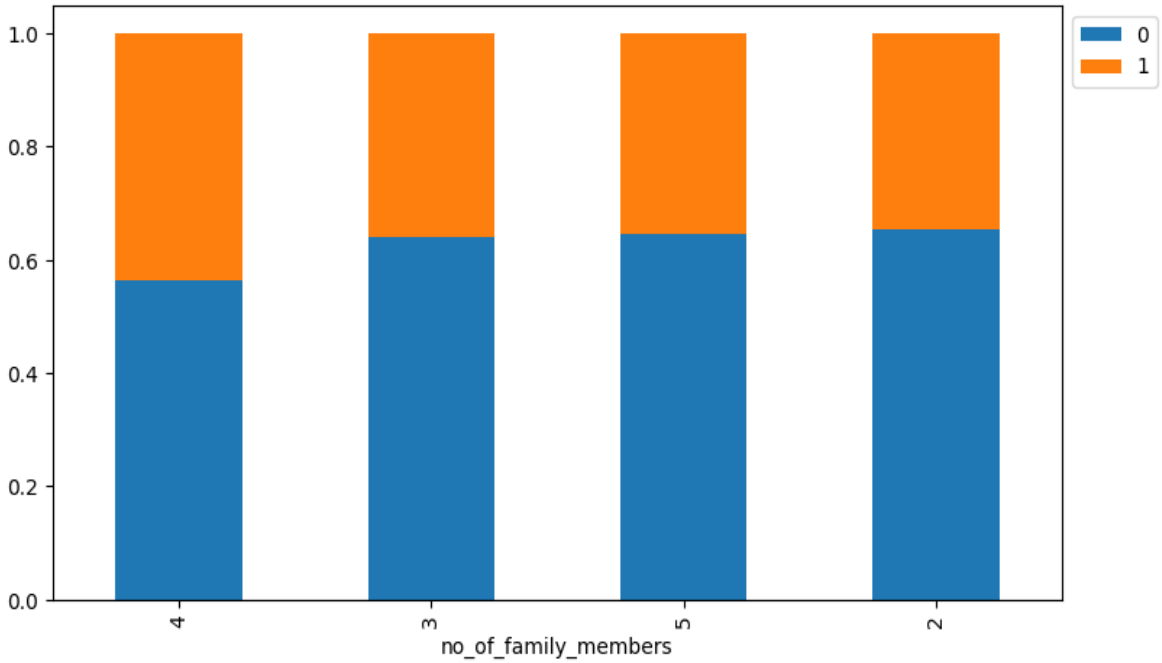
EDA Results: Bivariate Analysis: Lead Time vs Booking Status



- The lead time of both cancelled and not cancelled bookings is right-skewed
- The lead time tends to be about 70 days longer and is significantly more varied for cancelled than not cancelled bookings

Link to Appendix slide on data background check

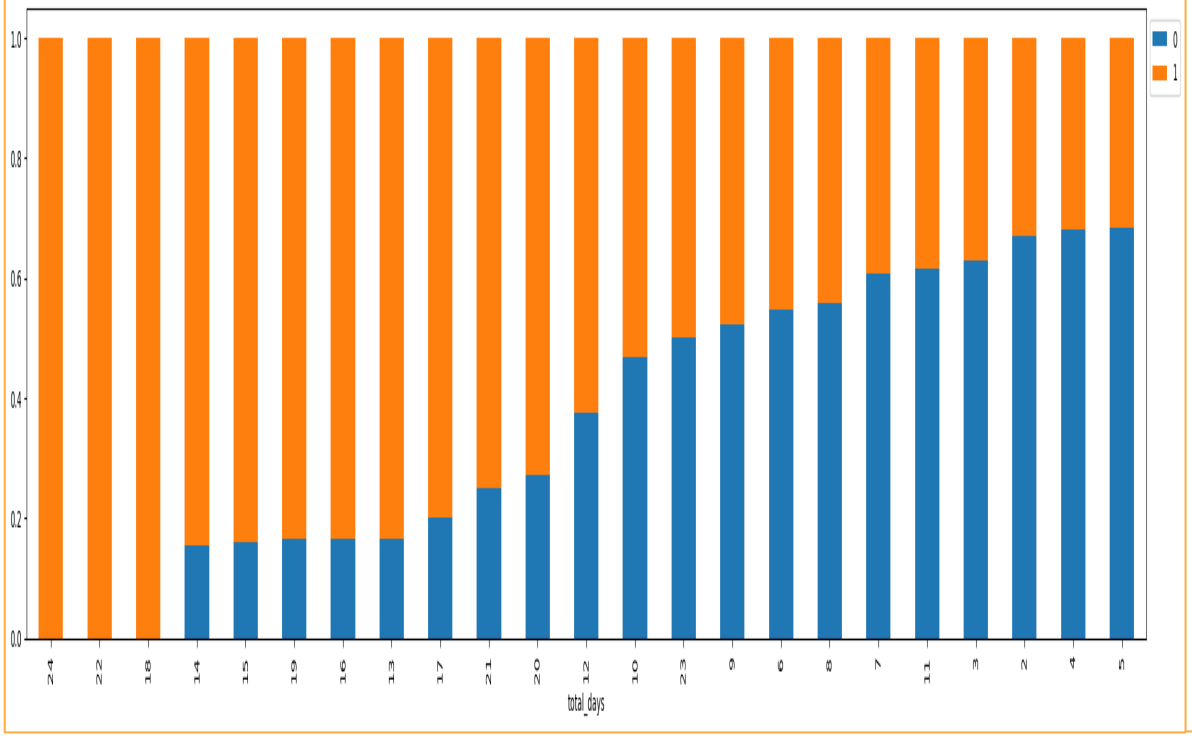
EDA Results: Bivariate Analysis: Number of Family Members vs Booking Status



- The greatest percentage of cancellations were registered for bookings with 4 family members
- Bookings registering 3, 5, and 2 family members had approximately the same percentage of cancellation

Link to Appendix slide on data background check

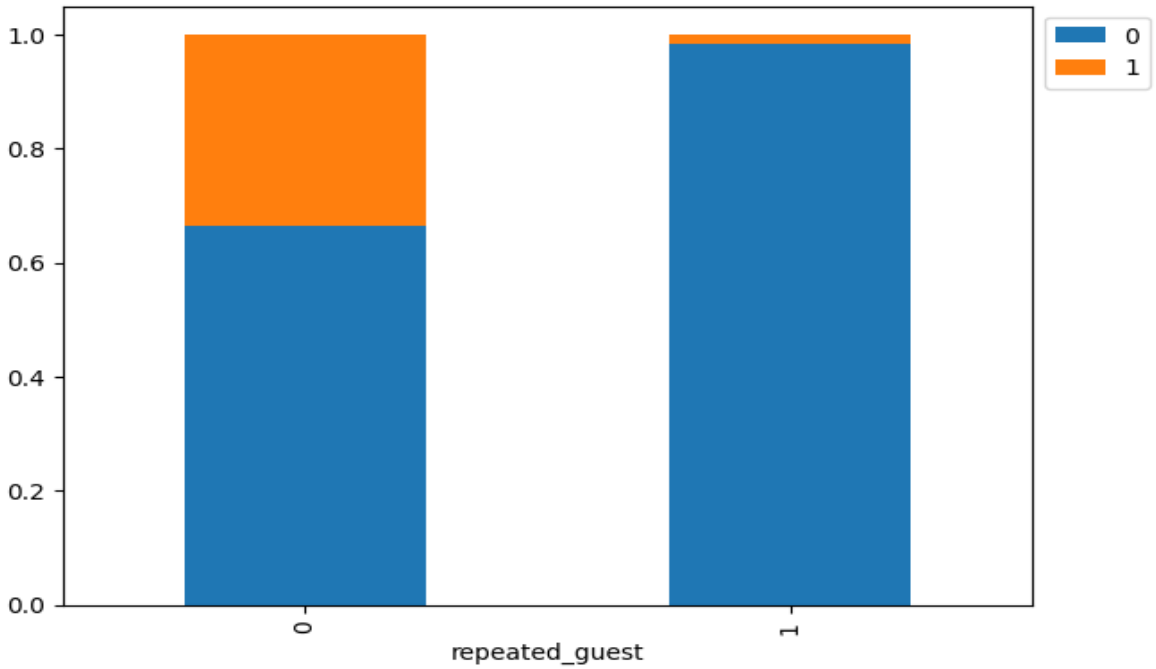
EDA Results: Bivariate Analysis: Total Days vs Booking Status



- For bookings registering both week and weekend nights, clients who booked 18, 22, and 24 days cancelled their bookings
- For bookings registering both week and weekend nights, bookings for 2, 4, and 5 days registered the least percentage of cancellations

Link to Appendix slide on data background check

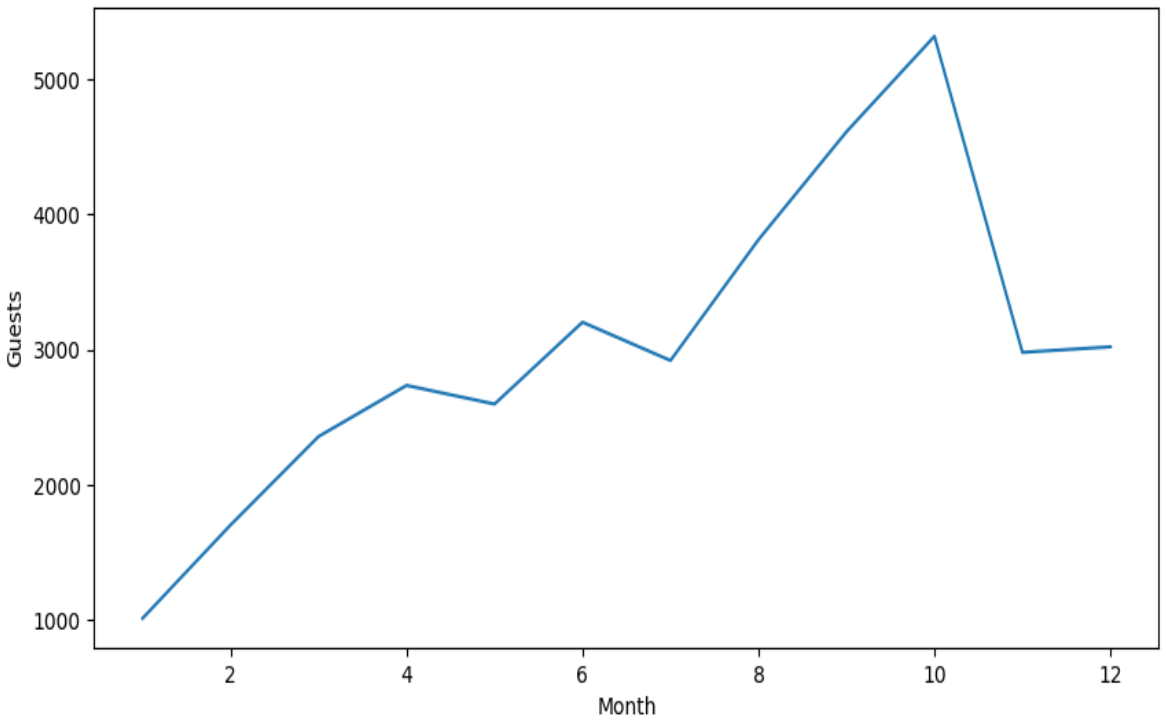
EDA Results: Bivariate Analysis: Repeated Guests vs Booking Status



- New clients are much more likely to cancel their bookings
- Old clients almost never cancel

Link to Appendix slide on data background check

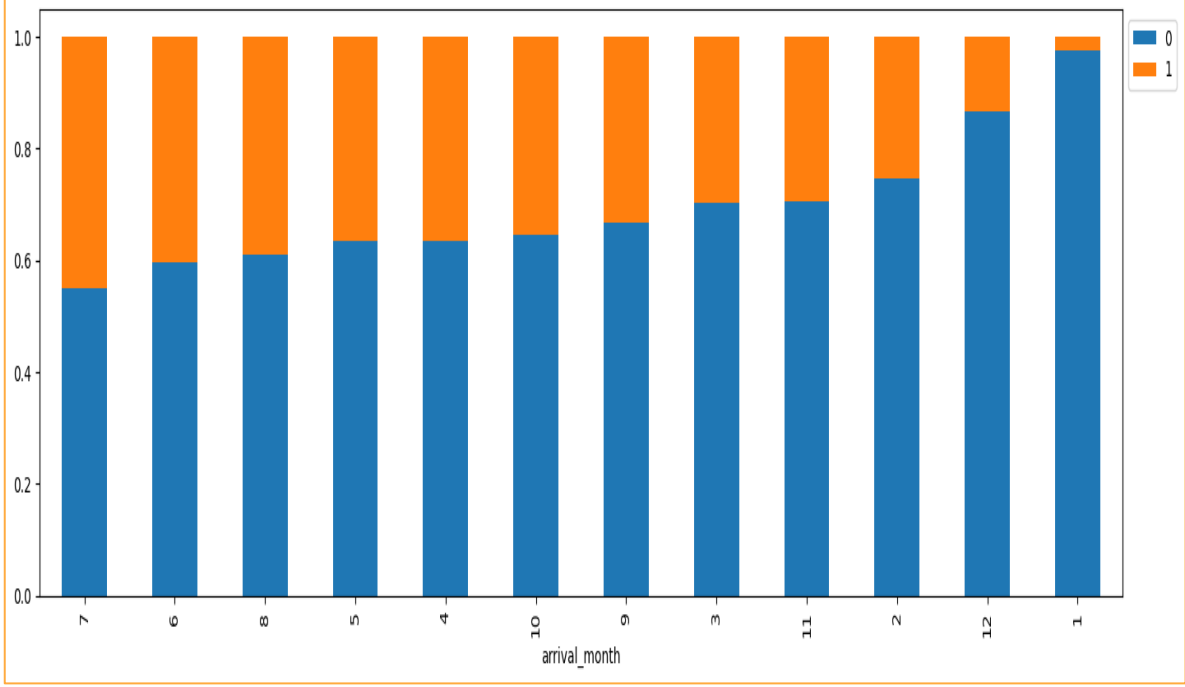
EDA Results: Bivariate Analysis: Month vs Guests



- We have confirmation that October registered the greatest number of arrivals whereas January had the least

Link to Appendix slide on data background check

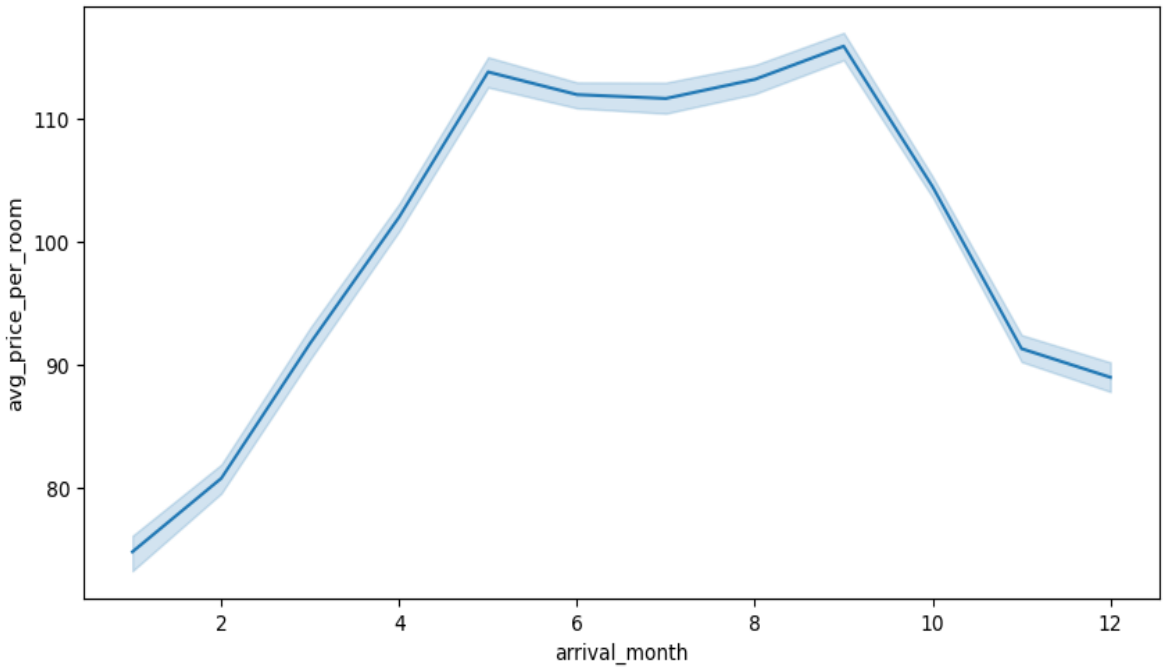
EDA Results: Bivariate Analysis: Arrival Month vs Booking Status



- The greatest percentage of cancellations was registered for bookings by clients arriving in July and the least for those arriving in January

Link to Appendix slide on data background check

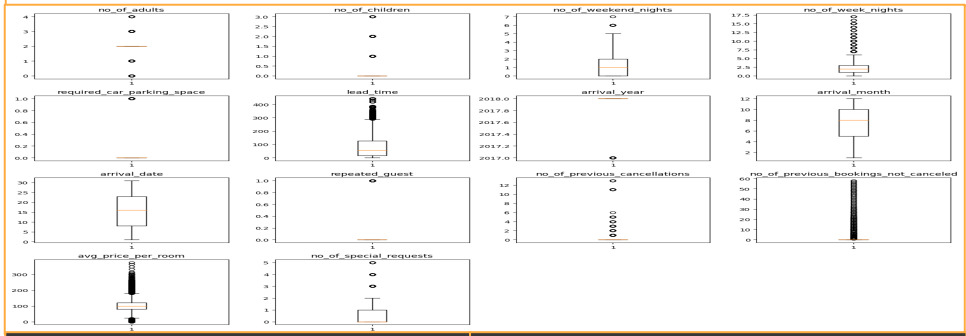
EDA Results: Bivariate Analysis: Arrival Month vs Average Price Per Room



- The most expensive bookings were registered for arrivals between May and September; correlates with trend of arrivals - demand was clearly highest during summer

Link to Appendix slide on data background check

Data Preprocessing



Logit Regression Results

Dep. Variable:	booking_status	No. Observations:	25392
Model:	Logit	Df Residuals:	25370
Method:	MLE	Df Model:	21
Date:	Wed, 13 Dec 2023	Pseudo R-squ.:	0.3282
Time:	02:36:37	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

```
Shape of Training set :  (25392, 28)
Shape of test set :  (10883, 28)
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
```

- The dataset contains no duplicates and no missing values
- Most of the attributes are right-skewed
- The arrival year is left-skewed
- The number of adults has roughly equal upper and lower outliers
- The arrival month and arrival date have no outliers
- A special form of winsorization has been used to handle outliers of the average price per room: Values greater than 500 have been assigned the value of the upper whisker, which is about 180
- Both the logistic and tree models benefitted of similar encoding:
  - Splitting the data into predictor and target variables
    Encoding the target variable (Booking Status): 1 for cancellation and 0 for non-cancellation
  - Dummy transformation of categorical predictor variables
    Adding a constant to the predictor variables solely for the logistic regression model
    Finally splitting the transformed variables into training and rest data respectively in the ratio 7:3
- The class distribution in the training and test data set is approximately the same (1 cancellation to 2 non-cancellations), confirming our previous observation that about 2/3 of the bookings were not cancelled
- None of the non-dummy variables has a variance inflation factor greater than 5; so we can conclude that no multicollinearity exists among the predictor variables
- After excluding relatively insignificant (high p-value) variables, the performance of the logistic model remains practically the same on the training data set: 81% accuracy, 63% recall, 74% precision, and 68% F1 score

Data Preprocessing

- Duplicate value check
- Missing value treatment
- Outlier check (treatment if needed)
- Feature engineering
- Data preparation for modeling

***Note**: You can use more than one slide if needed*

Model Performance Summary

Training Performance - Logit

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

Training Performance - Tree

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.99421	0.83097	0.89954
Recall	0.98661	0.78608	0.90303
Precision	0.99578	0.72425	0.81274
F1	0.99117	0.75390	0.85551

Test Performance - Logit

Test set performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80465	0.79555	0.80345
Recall	0.63089	0.73364	0.70338
Precision	0.72900	0.66573	0.69353
F1	0.67641	0.70074	0.69842

Test Performance - Tree

Test performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86679
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76614
F1	0.80309	0.75444	0.80848

- The performance of the logistic model on the training data set for the three thresholds is as follows:
  - Accuracy is highest for the default threshold and lowest for the threshold of 37%
  - The F1 score is highest for the threshold of 37% and least for default threshold
- The performance of the logistic model on the test data set for the three thresholds is as follows:
  - Accuracy is highest for model with the default threshold and lowest for the model with a threshold of 0.37
  - The F1 score is highest for the model with a threshold of 0.37 and lowest for the model with the default threshold
- The performance of the untuned, pre-tuned, and post-tuned tree classifier models on the training data is as follows:
  - Both the accuracy and f1 score are highest for the untuned model and lowest for the pre-tuned model
- The performance of the untuned, pre-tuned, and post-tuned tree classifier models on the test data is as follows:
  - The accuracy is highest for the untuned model and lowest for the pre-tuned model
  - The f1 score is highest for the post-tuned model and lowest for the pre-tuned model

APPENDIX

Data Background and Contents

```
<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
```

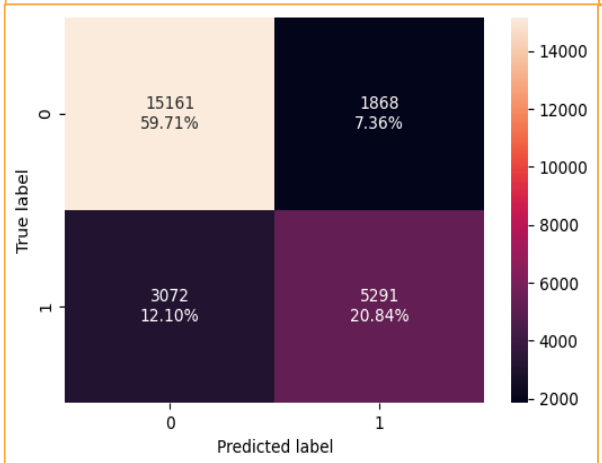
- The data has 5 categorical columns including Booking_ID which will be dropped because it has distinct values for each row and is simply used for identification
- The dataset has 19 columns (attributes) and 36275 rows (records)
- The dataset has 5 categorical columns and 14 numerical columns (13 int and 1 float); the categorical columns need to be encoded before running any model on the data
- The target column/variable (booking_status) is categorical and needs to be encoded before running any model on the data

Model Building - Logistic Regression

- The two principal verifications carried out to validate the use of logistic regression are multicollinearity and significance of each parameter assessed respectively by evaluation of the respective variance inflation factors and p-values
- The initial logistic model performance without parameter tuning registered an accuracy of 81%, recall of 63%, precision of 74%, and F1 score of 68% on the training set
- None of the non-dummy variables has a variance inflation factor greater than 5; so we can conclude that no multicolinearity exists among the predictor variables
- After excluding relatively insignificant (high p-value) variables, the performance of the logistic model remains practically the same on the training data set: 81% accuracy, 63% recall, 74% precision, and 68% F1 score
- For the logistic model trained on significant parameters, all else respectively held constant, an increase of one adult, child, required car parking space, arrival year, number of previous cancellations, average price per room, and number of special requests will result in corresponding 11% increase, 17% increase, 80% decrease, 57% increase, 26% increase, 2% increase, and 77% decrease of the odds of the booking being cancelled
- For the logistic model trained on significant parameters, the odds of an old client cancelling his/her booking is 94% less than that of a new client

Model Performance Evaluation and Improvement - Logistic Regression: Default Threshold

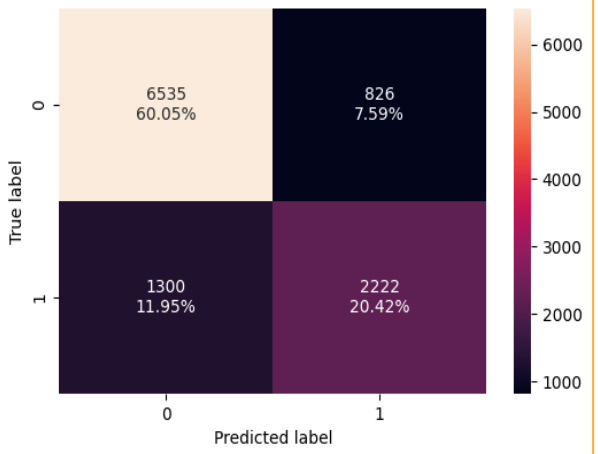
Training Performance



Training performance:

	Accuracy	Recall	Precision	F1
0	0.80545	0.63267	0.73907	0.68174

Test Performance



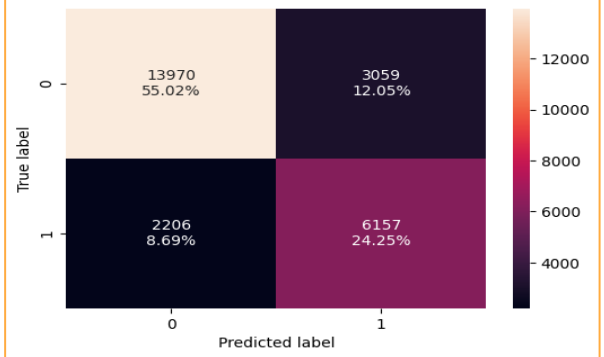
Test performance:

	Accuracy	Recall	Precision	F1
0	0.80465	0.63089	0.72900	0.67641

- Performance of the logistic model with the default threshold on the training data set:
  - 60% TN, 7% FP, 12% FN, and 21% TP
  - 81% accuracy, 63% recall, 74% precision, and 68% F1 score
- Performance of the logistic model with the default threshold on the test data set:
  - 60% TN, 8% FP, 12% FN, and 20% TP
  - 80% accuracy, 63% recall, 73% precision, and 68% f1 score

Model Performance Evaluation and Improvement - Logistic Regression: Optimal AUC-ROC Threshold (0.37)

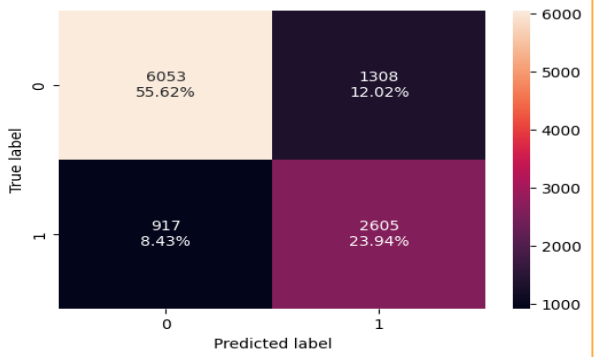
Training Performance



Training performance:

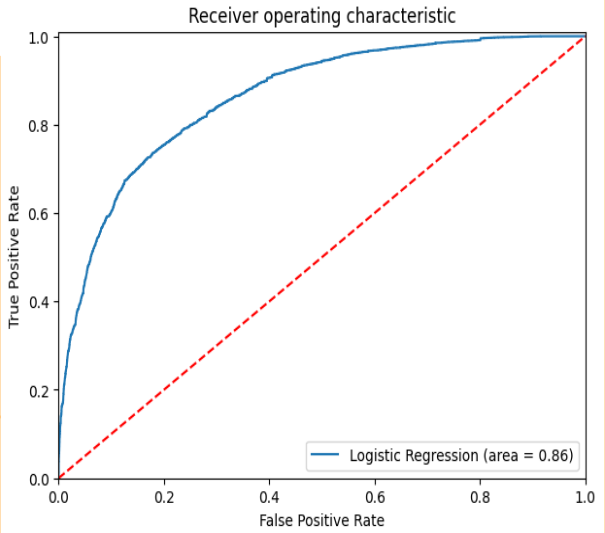
	Accuracy	Recall	Precision	F1
0	0.79265	0.73622	0.66808	0.70049

Test Performance



Test performance:

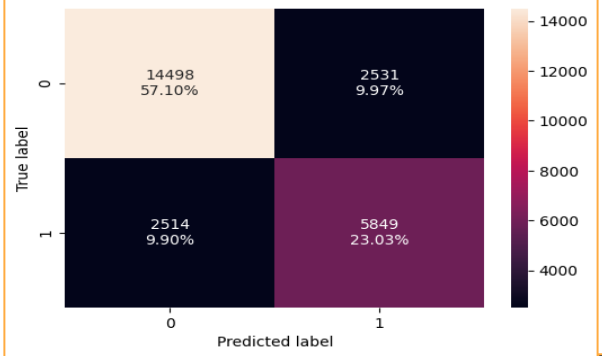
	Accuracy	Recall	Precision	F1
0	0.79555	0.73964	0.66573	0.70074



- Performance of the logistic model with optimal auc-roc threshold on the training data set:
  - 55% TN, 12% FP, 9% FN, and 24% TP
  - 79% accuracy, 74% recall, 67% precision, and 70% F1 score
- Performance of the logistic model with optimal auc-roc threshold on the test data set:
  - 56% TN, 12% FP, 8% FN, and 24% TP
  - 80% accuracy, 74% recall, 67% precision, and 70% f1 score

Model Performance Evaluation and Improvement - Logistic Regression: Optimal Precision-Recall Threshold (0.42)

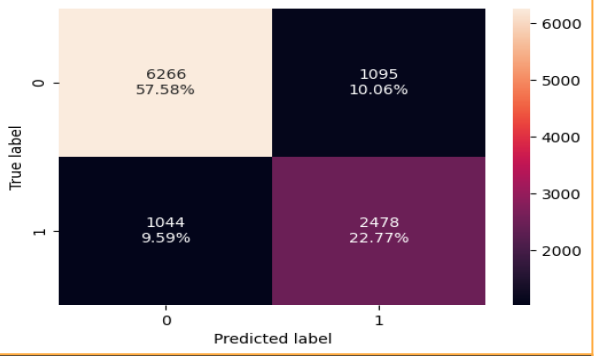
Training Performance



Training performance:

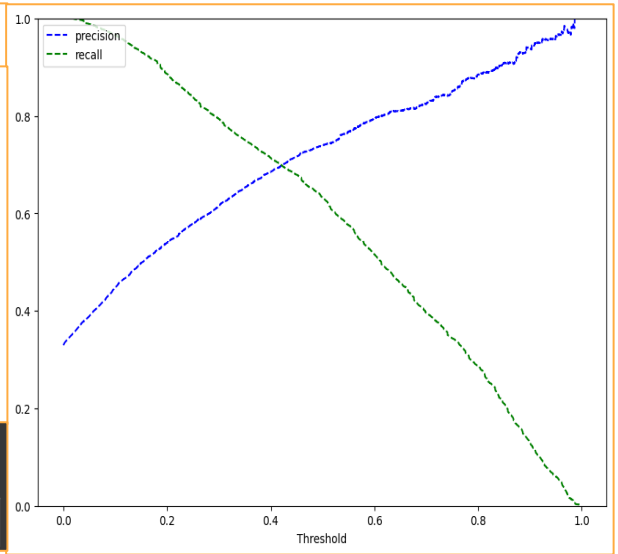
	Accuracy	Recall	Precision	F1
0	0.80132	0.69939	0.69797	0.69868

Test Performance



Test performance:

	Accuracy	Recall	Precision	F1
0	0.80345	0.70358	0.69353	0.69852



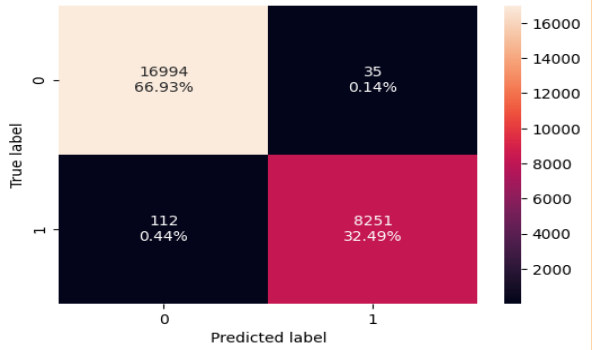
- Performance of the logistic model with optimal precision-recall threshold on the training data set:
  - 57% TN, 10% FP, 10% FN, and 23% TP
  - 80% accuracy, 70% recall, 70% precision, and 70% F1 score
- Performance of the logistic model with optimal precision-recall threshold on the test data set:
  - 58% TN, 10% FP, 10% FN, and 23% TP
  - 80% accuracy, 70% recall, 69% precision, and 70% f1 score

Model Building - Decision Tree

- Just as in the case of logistic regression, we will encode both the prediction and target variables, transforming categorical variables into dummy variables and splitting the data into training and test data sets in a ratio of 7:3 (No need adding a constant to the predictor variables since the target model is non-linear)
- We will then build the Decision Tree Classifier model, fitting it to the train data (one advantage of the tree classifier model is the relatively limited feature engineering required before building the model)
- The model registered varying performance for the training and test sets and depending on whether hyper-parameters were untuned, pre-tuned, or post-tuned

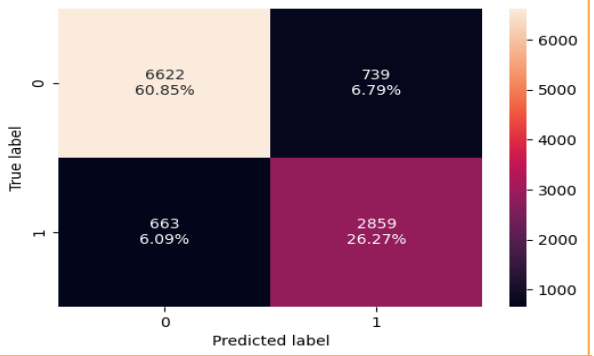
Model Performance Evaluation and Improvement - Decision Tree: Untuned Hyperparameters

Training Performance

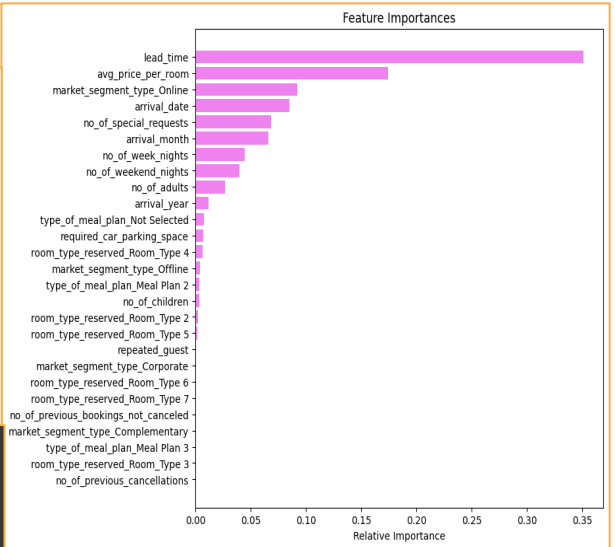


	Accuracy	Recall	Precision	F1
0	0.99421	0.98661	0.99578	0.99117

Test Performance



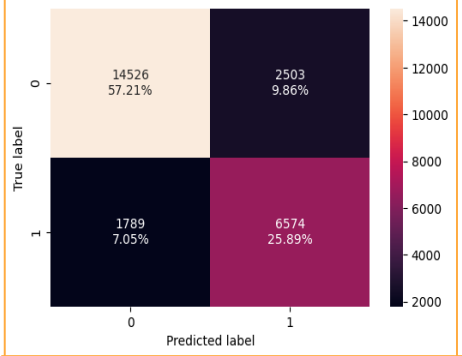
	Accuracy	Recall	Precision	F1
0	0.87118	0.81175	0.79461	0.80309



- Performance of the tree classifier model with untuned hyperparameters on the training data set:
  - 67% TN, 0.1% FP, 0.4% FN, and 32% TP
  - 99% accuracy, 99% recall, 100% precision, and 99% f1 score
- Performance of the tree classifier model with untuned hyperparameters on the test data set:
  - 61% TN, 7% FP, 6% FN, and 26% TP
  - 87% accuracy, 81% recall, 79% precision, and 80% f1 score
- The lead time is the most important feature of the tree classifier model followed by the average price per room

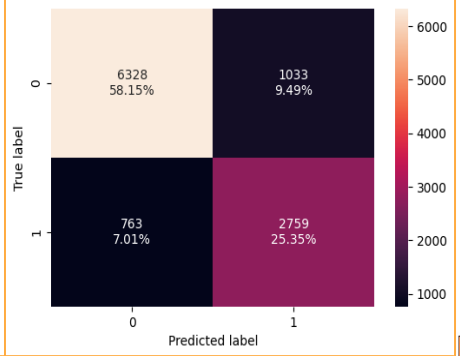
Model Performance Evaluation and Improvement - Decision Tree: Pre-Tuned Hyperparameters

Training Performance

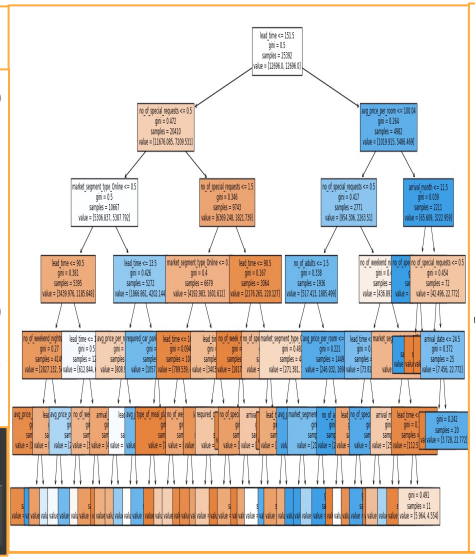


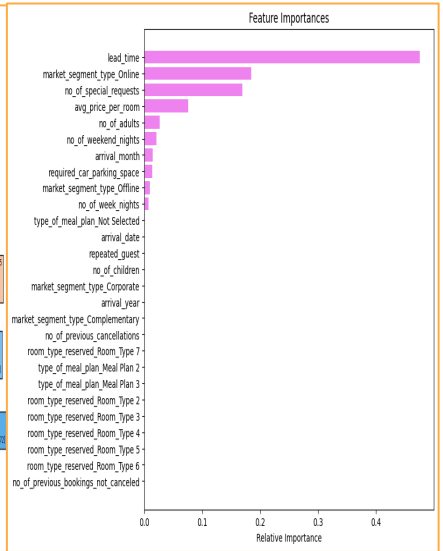
	Accuracy	Recall	Precision	F1
0	0.83097	0.78608	0.72425	0.75390

Test Performance



	Accuracy	Recall	Precision	F1
0	0.83497	0.78336	0.72758	0.75444

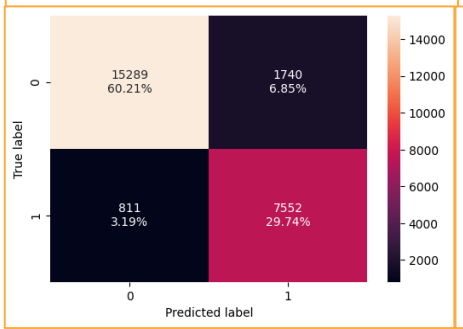




- The pre-tuned tree classifier model uses the following hyperparameters: balanced class weight, max depth of 6, max leaf nodes of 50, and minimum samples per split of 10
- Performance of the tree classifier model with pre-tuned hyperparameters on the training data set:
  - 57% TN, 10% FP, 7% FN, and 26% TP
  - 83% accuracy,79% recall, 72% precision, and 75% f1 score
- Performance of the tree classifier model with pre-tuned hyperparameters on the test data set:
  - 58% TN, 9% FP, 7% FN, and 25% TP
  - 83% accuracy, 78% recall, 73% precision, and 75% f1 score
- The lead time is the most important feature of the pre-tuned tree classifier model followed by the online market segment type

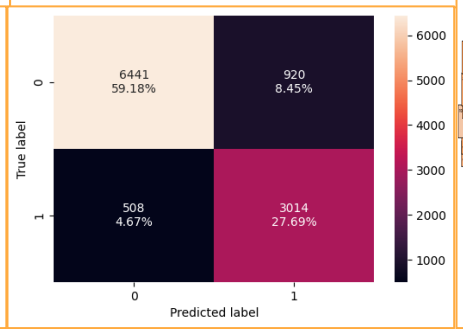
Model Performance Evaluation and Improvement - Decision Tree: Post-Tuned Hyperparameters

Training Performance

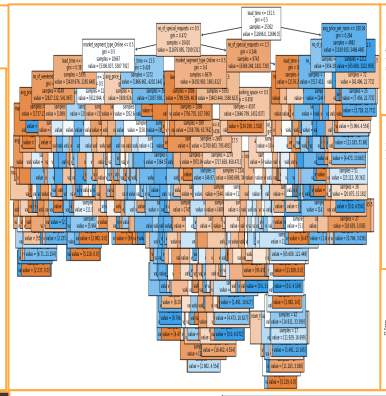


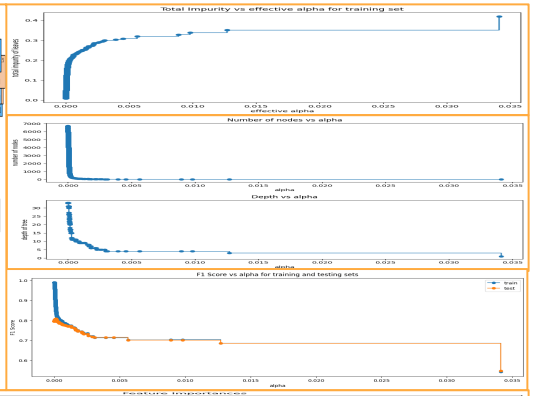
	Accuracy	Recall	Precision	F1
0	0.89954	0.90303	0.81274	0.85551

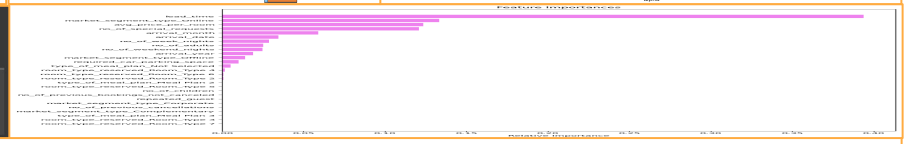
Test Performance



	Accuracy	Recall	Precision	F1
0	0.86879	0.85576	0.76614	0.80848







- The total impurity of leaves increases with increase in the effective alpha but the relationship is non-linear
- The number of nodes and depth of tree decreases with increase in alpha but the relationships are non-linear
- The f1 score vs alpha curve is very similar for training and test data
- The best (highest f1 score for test data) post-tuned tree classifier model has an alpha value of 0.00012 and the class weight is balanced
- Performance of the tree classifier model with post-tuned hyperparameters on the training data set:
  - 60% TN, 7% FP, 3% FN, and 30% TP
  - 90% accuracy,90% recall, 81% precision, and 86% f1 score
- Performance of the tree classifier model with post-tuned hyperparameters on the test data set:
  - 59% TN, 8% FP, 5% FN, and 28% TP
  - 87% accuracy, 86% recall, 77% precision, and 81% f1 score
- The lead time is the most important feature of the post-tuned tree classifier model followed by the online market segment type

Great Learning

Happy Learning !