INN HOTELS PROJECT

BY: DR. CYNTHIA OFFOHA

AGENDA

- EXECUTIVE SUMMARY
- BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH
- DATA OVERVIEW
- EDA - UNIVARIATE ANALYSIS & KEY QUESTIONS
- EDA – BIVARIATE ANALYSIS & KEY QUESTIONS
- LOGISTIC REGRESSION & DECISION TREE
- RECOMMENDATIONS & CONCLUSIONS

EXECUTIVE SUMMARY

- Some hotels bookings are called off due to cancellations or no-shows with reasons being due to change of plans and scheduling conflicts
- New technologies have been developed in promoting online booking channels which have made it easier for customers
- It is shown that cancellation of bookings impact hotels in various aspects such as:
  - Loss of resources when hotel cannot resell the room
  - Cost of distribution channels by increasing commissions or paying for publicity
  - Decreasing prices last minute in order for hotel to resell room which leads to reduction in profit margin
  - Guest arrangements situated by human resources

BUSINESS PROBLEM OVERVIEW & SOLUTION APPROACH

- The increased number of cancellations have led to the urgent need for Machine Learning expertise to assist in predicting which booking is likely to be cancelled
- INN Hotels Group has a chain of hotels in Portugal which is facing problems with high number of cancellations booking and has reached out for data-driven solutions
- As Data Scientist, the goal is to analyze the data provided to determine which factors are affecting the cancellations on booking
    - Thus, the data scientists must develop a predictive model that can deter which booking is going to be cancelled in advance
    - Also, data scientists will assist in formulation of profitable policies for cancellations and refunds

DATA OVERVIEW

VARIABLE	DESCRIPTIONS
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

VARIABLE	DESCRIPTION
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.

DATA STRUCTURE

	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan	required_car_parking_space
0	INN00001	2	0	1	2	Meal Plan 1	0
1	INN00002	2	0	2	3	Not Selected	0
2	INN00003	1	0	2	1	Meal Plan 1	0
3	INN00004	2	0	0	2	Meal Plan 1	0
4	INN00005	2	0	1	1	Not Selected	0

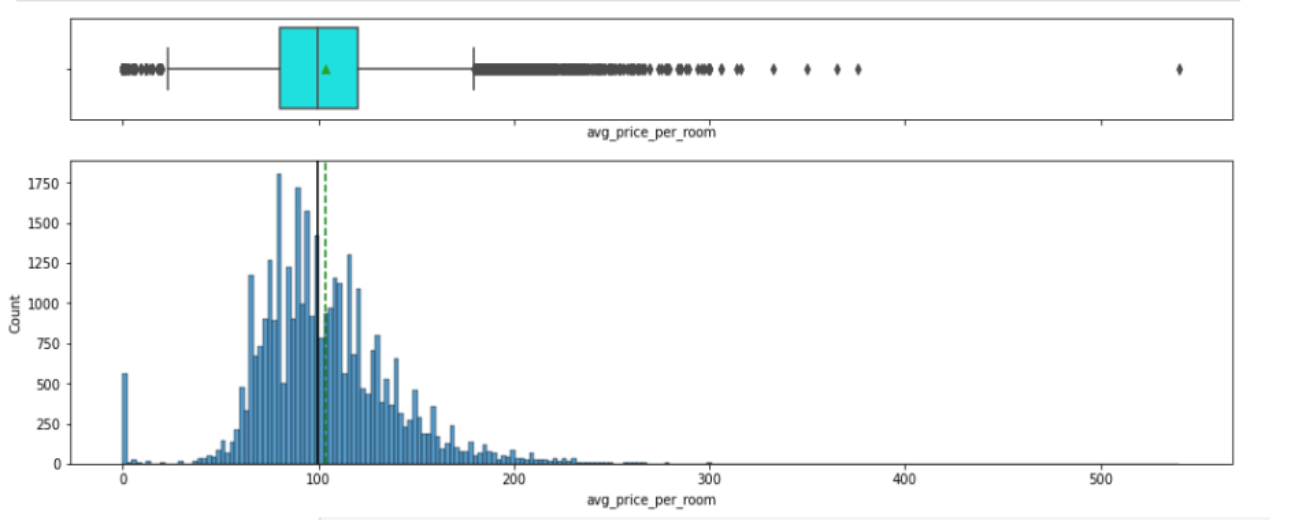
	Booking_ID	no_of_adults	no_of_children	no_of_weekend_nights	no_of_week_nights	type_of_meal_plan
30048	INN30049	2	0	1	2	Not Selected
21003	INN21004	2	0	0	3	Meal Plan 1
23749	INN23750	2	0	0	2	Meal Plan 2
30908	INN30909	1	0	0	3	Not Selected
9276	INN09277	1	0	2	2	Meal Plan 1

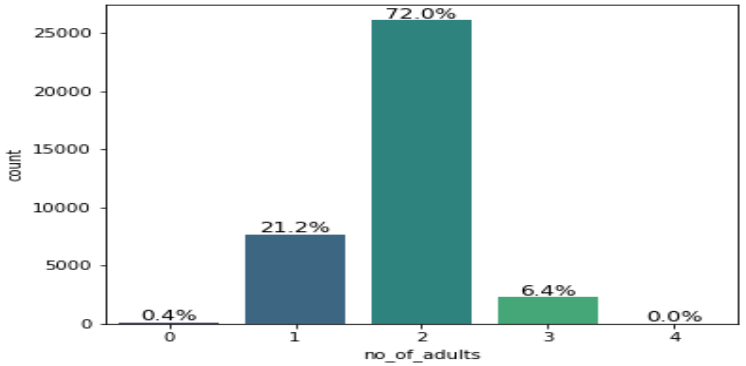
```
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Booking_ID                            36275 non-null  object
 1   no_of_adults                          36275 non-null  int64
 2   no_of_children                        36275 non-null  int64
 3   type_of_meal_plan                     36275 non-null  object
 4   required_car_parking_space            36275 non-null  int64
 5   room_type_reserved                    36275 non-null  object
 6   lead_time                             36275 non-null  int64
 7   arrival_year                          36275 non-null  int64
 8   arrival_month                         36275 non-null  int64
 9   arrival_date                          36275 non-null  int64
 10  market_segment_type                   36275 non-null  object
 11  repeated_guest                        36275 non-null  int64
 12  no_of_previous_cancellations          36275 non-null  int64
 13  no_of_previous_bookings_not_canceled  36275 non-null  int64
 14  avg_price_per_room                    36275 non-null  float64
 15  no_of_special_requests                36275 non-null  int64
 16  booking_status                        36275 non-null  object
 17  total_nights                          36275 non-null  int64
dtypes: float64(1), int64(12), object(5)
```

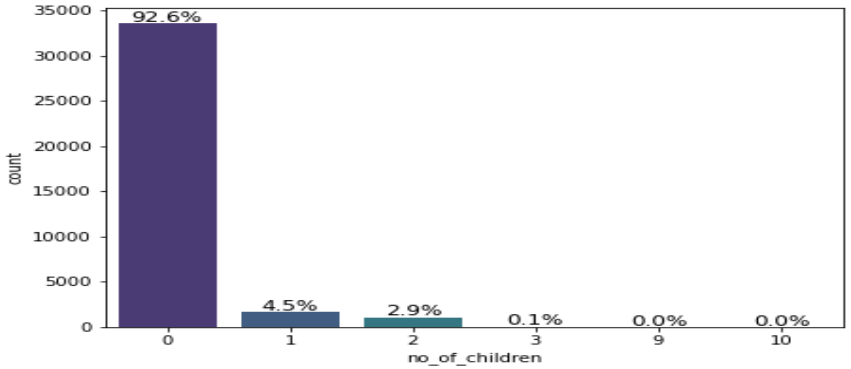
OBSERVATIONS

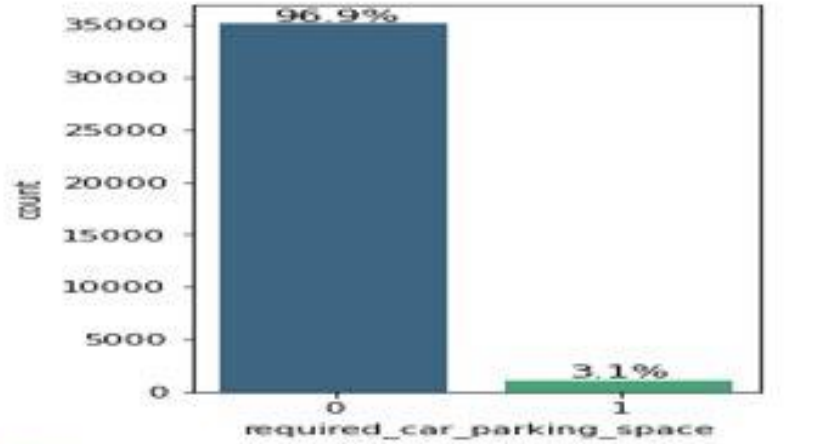
- Dataframe has 36275 rows and 19 columns
- No duplicates values in dataset and no missing data across all columns
- 5 columns of the dtype object, 1 columns of the dtype float64, and 13 columns of the dtype int64.
- Target variable for both the Logistic Regression model and the Decision Tree model will be *booking_status*
- room_type_reserved values are ciphered (encoded) by INN Hotels and not be useful for model building
- Booking_ID values illustrate no material information for model building

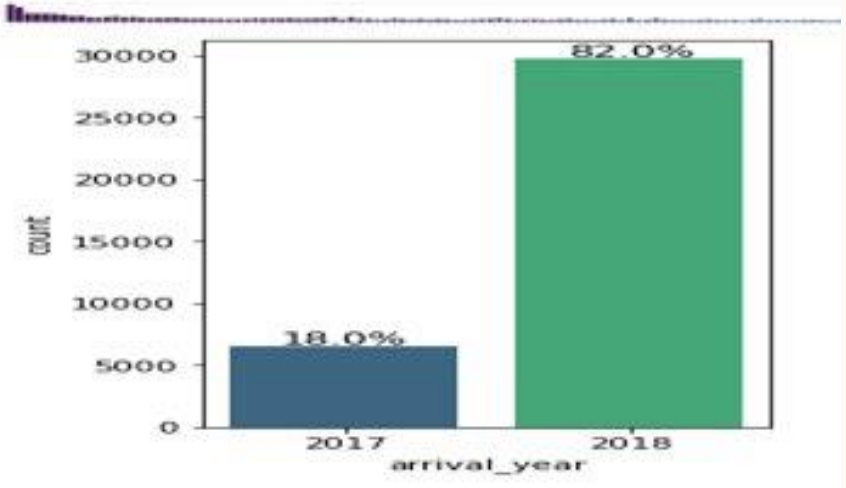
EDA: UNIVARIATE ANALYSIS



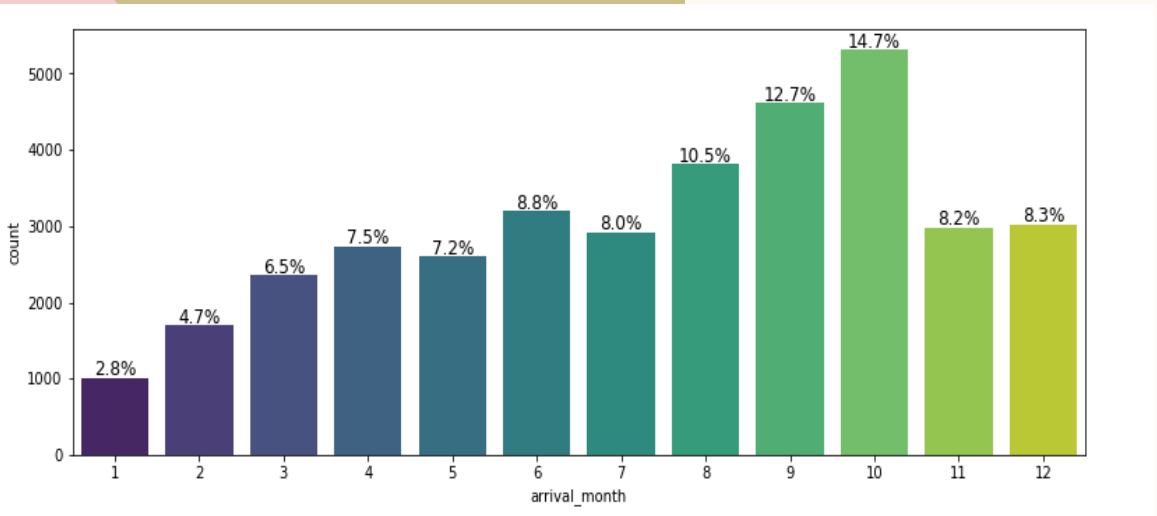


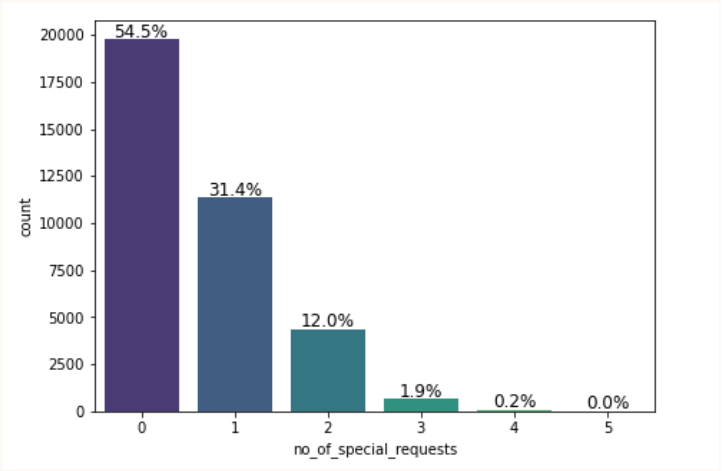


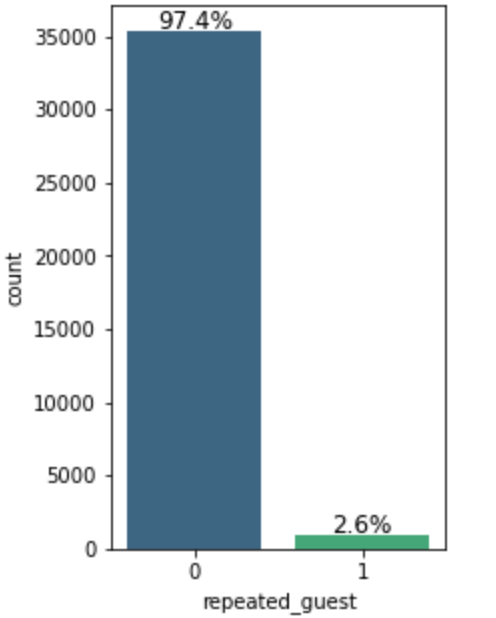


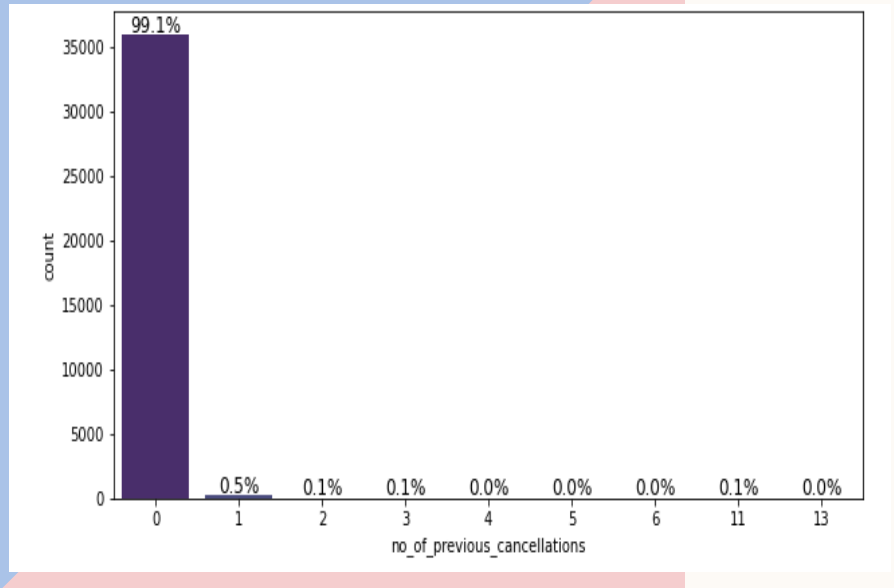


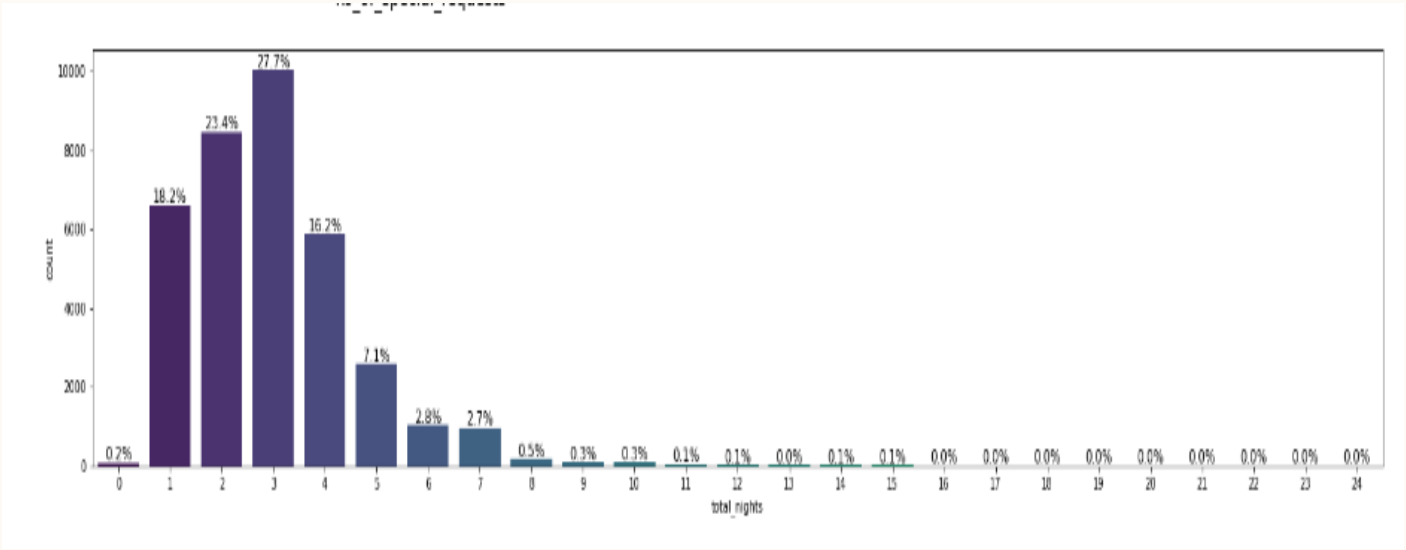
EDA: UNIVARIATE ANALYSIS



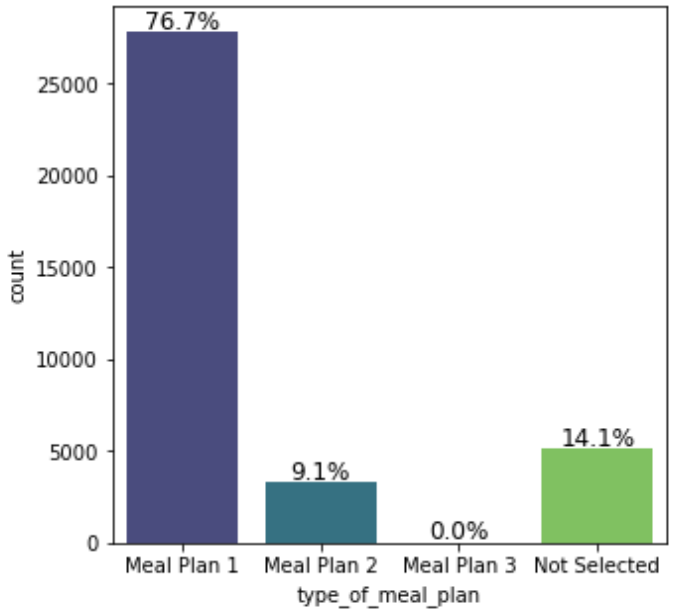


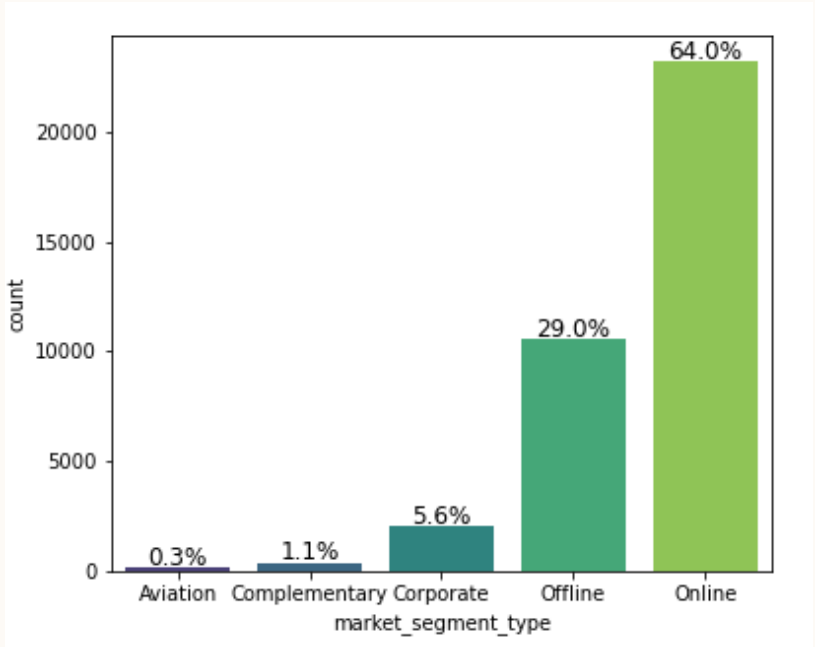


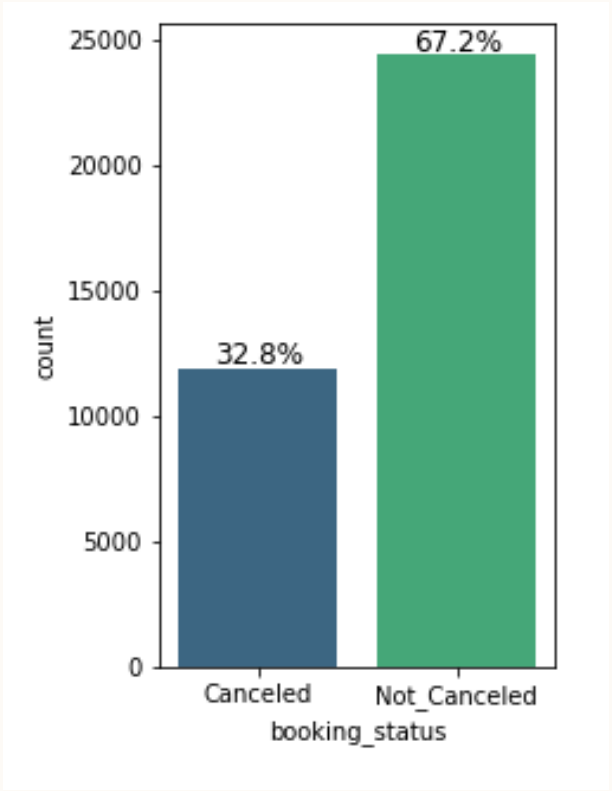


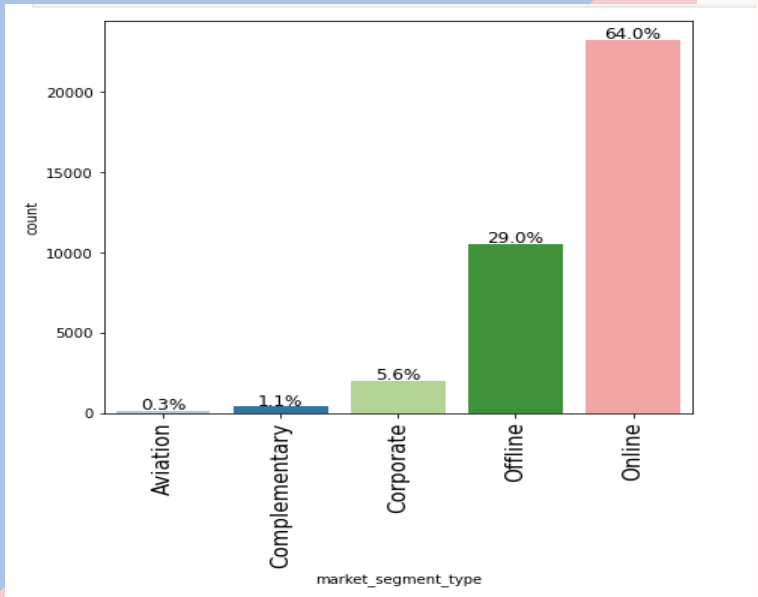


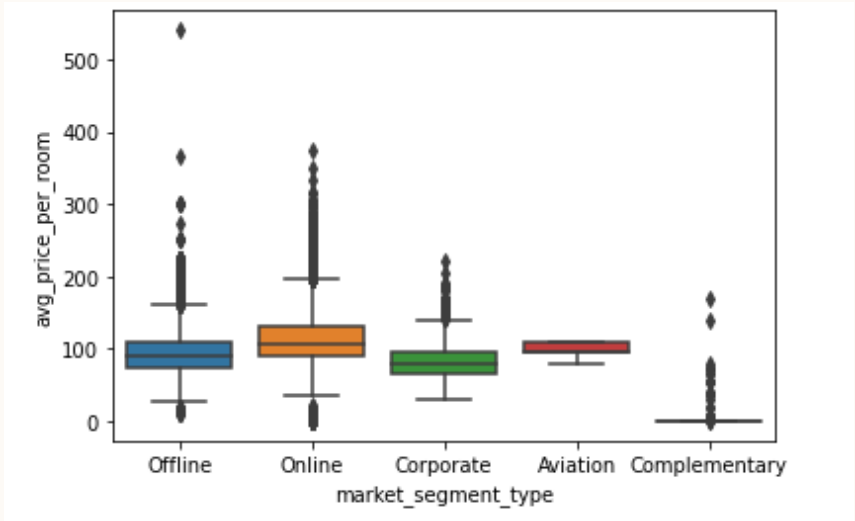
EDA: UNIVARIATE ANALYSIS

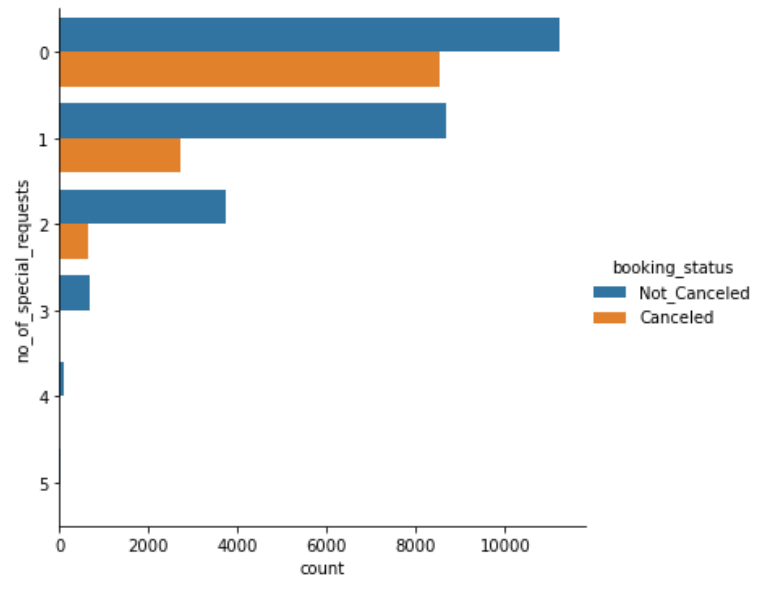


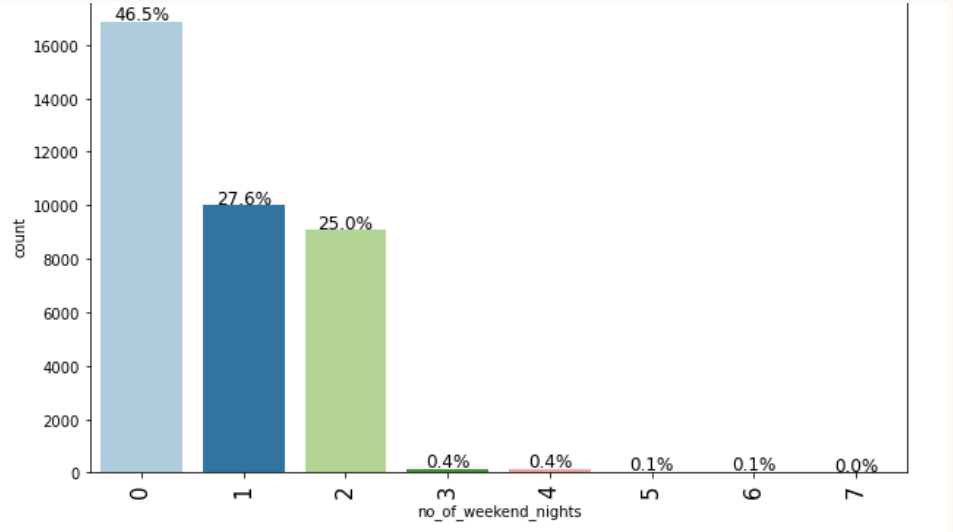


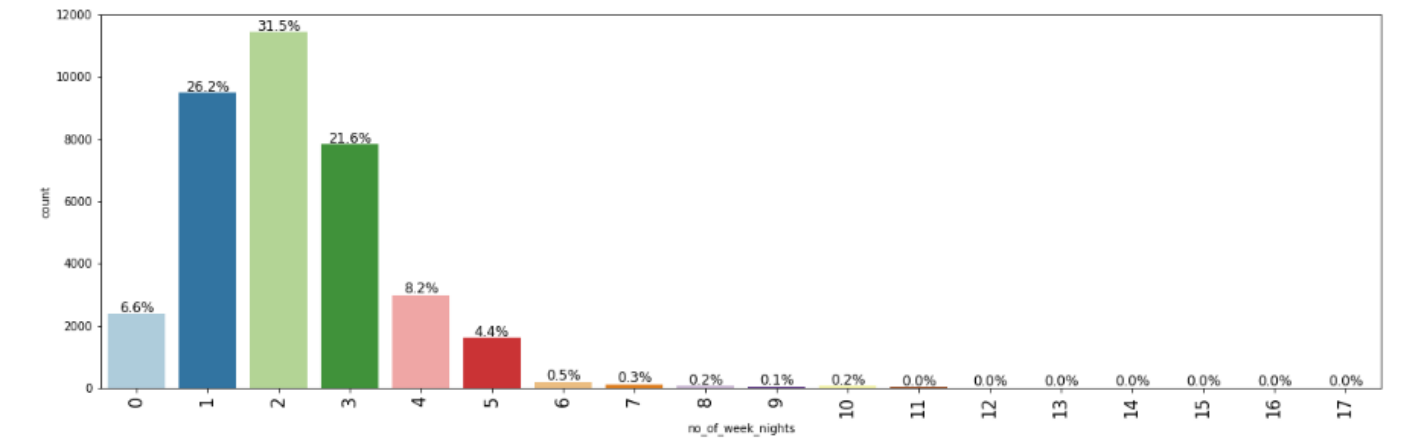


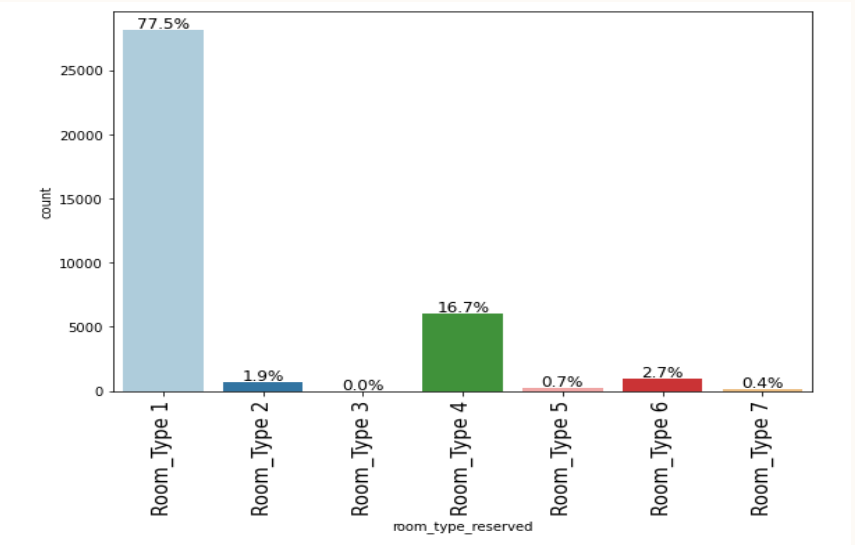





no_of_special_requests
count
booking_status
Not_Canceled
Canceled


count
no_of_weekend_nights
46.5%
27.6%
25.0%
0.4%
0.4%
0.1%
0.1%
0.0%


count
no_of_week_nights
6.6%
26.2%
31.5%
21.6%
8.2%
4.4%
0.5%
0.3%
0.2%
0.1%
0.2%
0.0%
0.0%
0.0%
0.0%
0.0%
0.0%
0.0%

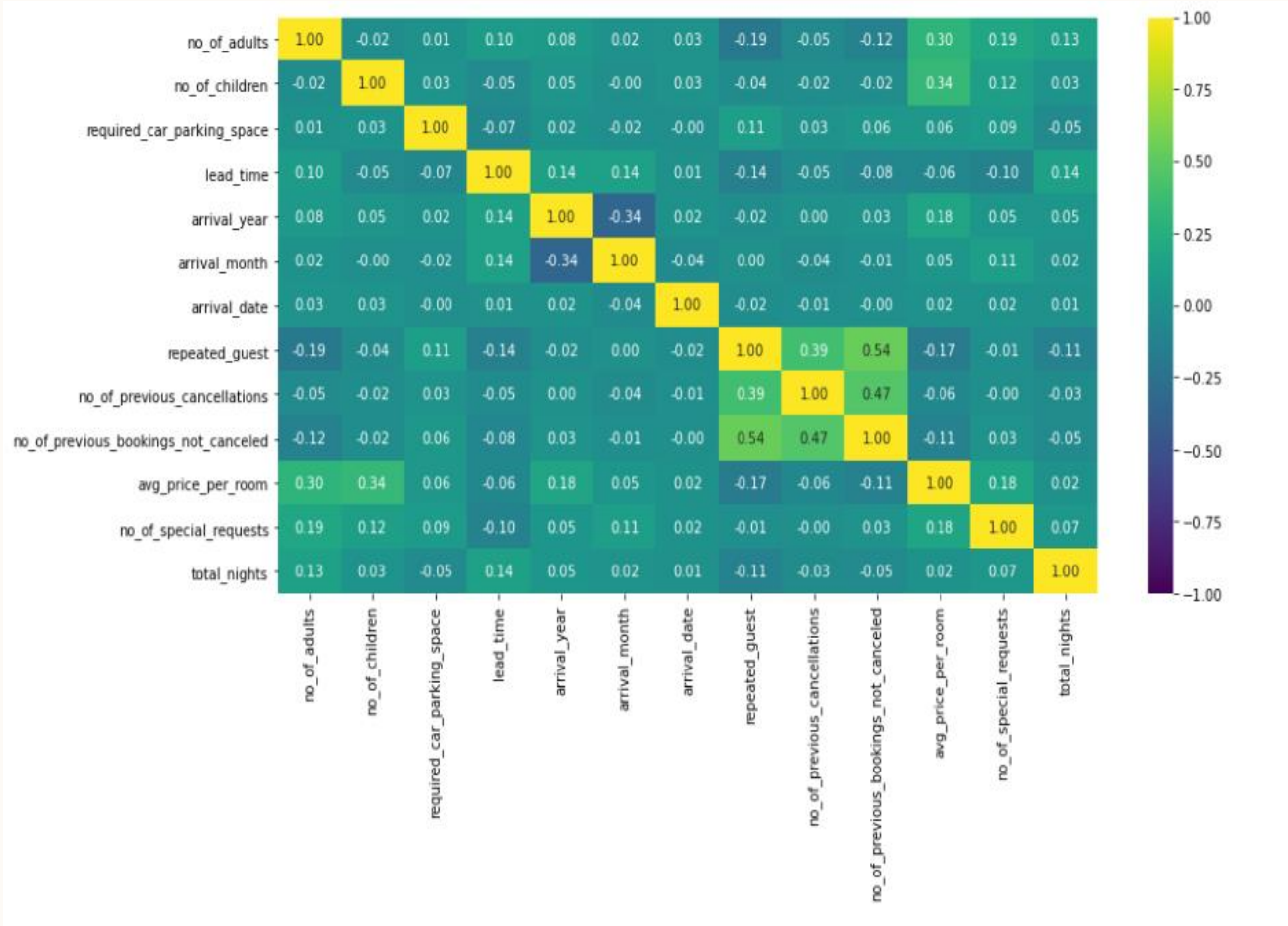

count
room_type_reserved
77.5%
1.9%
0.0%
16.7%
0.7%
2.7%
0.4%
Room_Type 1
Room_Type 2
Room_Type 3
Room_Type 4
Room_Type 5
Room_Type 6
Room_Type 7

OBSERVATIONS: UA

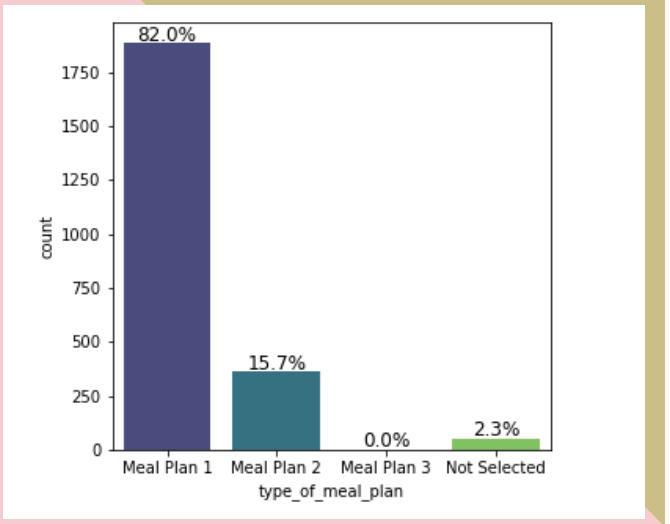
- It is shown that october is the busiest month followed by September and August
- Approx. 64% (23214) bookings are made online which yield the highest average price per room ($112.25), followed by bookings through Aviation and Offline market
- Approx. 33% (11885) bookings were cancelled whereas ~2% (16 out of 930) total bookings made by repeated guest were cancelled
- It is shown that as # of special requests made by customers increases, the proportion of bookings cancelled decreases
- The median avg. Room price ~100 / night
    - Plot showed right skewed distribution depicting significant influence of highly priced rooms
- 78% of the bookings include two or more adults and 21% of bookings were single adult
- ~93% of bookings consist no children
- ~97% of guests required no parking space and not a single booking required more than one space
- 82% of the bookings took place during 2018 and 2.6% of bookings came from repeat guests
- 0.9% of bookings (338 out of 36,275 total bookings) were booked by guest with one or more previous cancellations on record
- >90% of bookings chose only breakfast or no meal

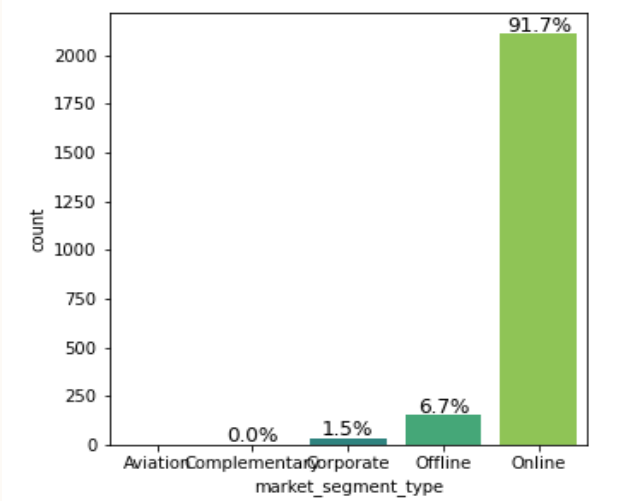
EDA: BIVARIATE ANALYSIS

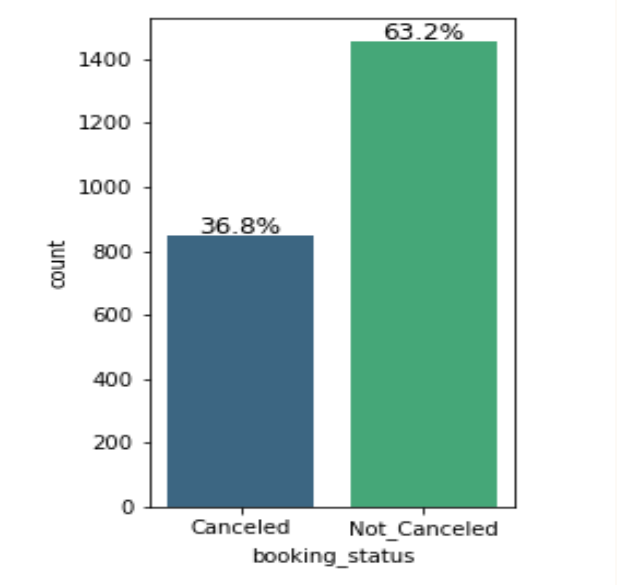
- Most variables showed no correlations
- As the number of adults and children increase, the avg. price of a room can be expected to increase
- A repeat guest are more likely to have both previously cancelled and not cancelled booking:
  - A repeat guest has more opportunities to decide to cancel or not cancel these previous bookings
  - The weak positive correlation between number of previous cancellations and number of previous bookings not cancelled supports this conclusion

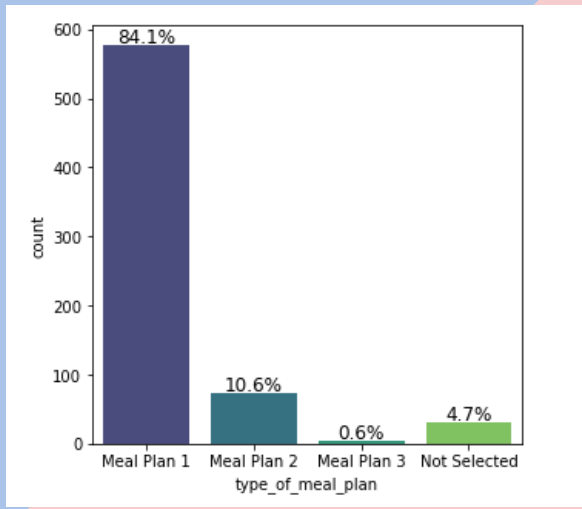


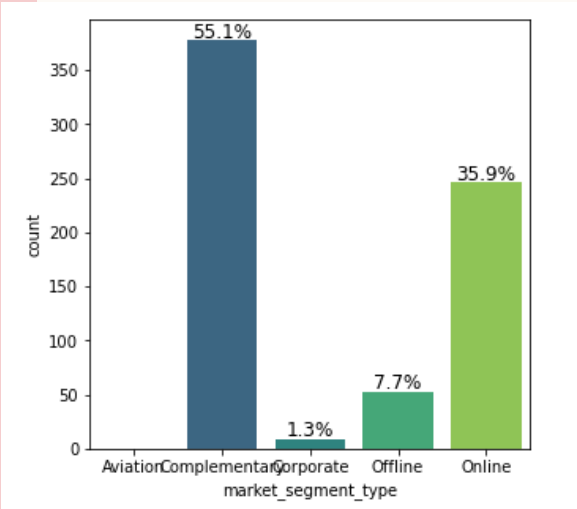
EDA: BIVARIATE ANALYSIS

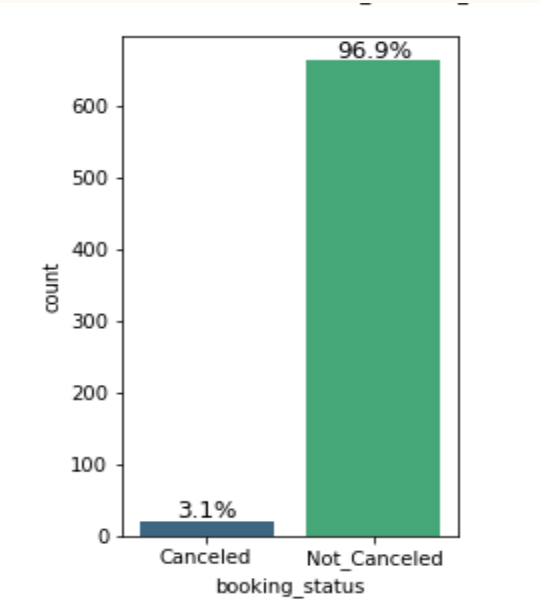


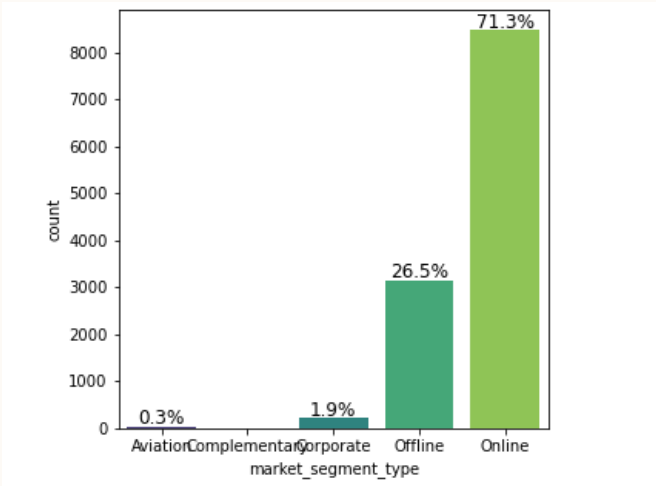












OBSERVATIONS: BA

- 565 rooms sold at no cost to the guests during the period observed
- Of the rooms with low outlier average room prices (<~40 /night):
  - **Only 3% of these guests cancelled reservations**
- Of the rooms with high outlier average room prices (>~160/night):
  - More than a third (~36%) cancelled reservations, 4% higher than the cancelation rate for all bookings
  - Over 91% of these high-priced reservations were booked online
- 71.3% of cancelled bookings were booked online

LOGISTIC REGRESSION & MULTICOLLINEARITY

	coef	std err	z	P>|z|	[0.025	0.975]
const	-947.8054	120.138	-7.889	0.000	-1183.271	-712.340
no_of_adults	0.0698	0.036	1.914	0.056	-0.002	0.141
no_of_children	-0.0450	0.046	-0.986	0.324	-0.134	0.044
required_car_parking_space	-1.5856	0.138	-11.494	0.000	-1.856	-1.315
lead_time	0.0156	0.000	59.066	0.000	0.015	0.016
arrival_year	0.4685	0.060	7.871	0.000	0.352	0.585
arrival_month	-0.0391	0.006	-6.085	0.000	-0.052	-0.027
arrival_date	0.0004	0.002	0.206	0.837	-0.003	0.004
repeated_guest	-2.3829	0.617	-3.862	0.000	-3.592	-1.174
no_of_previous_cancellations	0.2687	0.085	3.166	0.002	0.102	0.435
no_of_previous_bookings_not_canceled	-0.1749	0.154	-1.134	0.257	-0.477	0.127
avg_price_per_room	0.0162	0.001	24.363	0.000	0.015	0.018
no_of_special_requests	-1.4556	0.030	-48.576	0.000	-1.514	-1.397
total_nights	0.0534	0.009	5.656	0.000	0.035	0.072
type_of_meal_plan_Meal Plan 2	0.2485	0.066	3.777	0.000	0.120	0.378
type_of_meal_plan_Meal Plan 3	16.7413	2512.467	0.007	0.995	-4907.604	4941.087
type_of_meal_plan_Not Selected	0.3337	0.051	6.527	0.000	0.234	0.434
market_segment_type_Complementary	-28.9673	2751.609	-0.011	0.992	-5422.022	5364.088
market_segment_type_Corporate	-1.1178	0.264	-4.242	0.000	-1.634	-0.601
market_segment_type_Offline	-2.0655	0.252	-8.201	0.000	-2.559	-1.572
market_segment_type_Online	-0.2948	0.249	-1.183	0.237	-0.783	0.194

	feature	VIF
0	const	39091160.37
20	market_segment_type_Online	70.87
19	market_segment_type_Offline	63.67
18	market_segment_type_Corporate	16.82
17	market_segment_type_Complementary	4.44
8	repeated_guest	1.78
11	avg_price_per_room	1.75
10	no_of_previous_bookings_not_canceled	1.65
5	arrival_year	1.42
9	no_of_previous_cancellations	1.39
4	lead_time	1.38
1	no_of_adults	1.27
6	arrival_month	1.27
14	type_of_meal_plan_Meal Plan 2	1.26
12	no_of_special_requests	1.24
16	type_of_meal_plan_Not Selected	1.20
2	no_of_children	1.19
13	total_nights	1.09
3	required_car_parking_space	1.04
15	type_of_meal_plan_Meal Plan 3	1.01
7	arrival_date	1.01

OBSERVATIONS

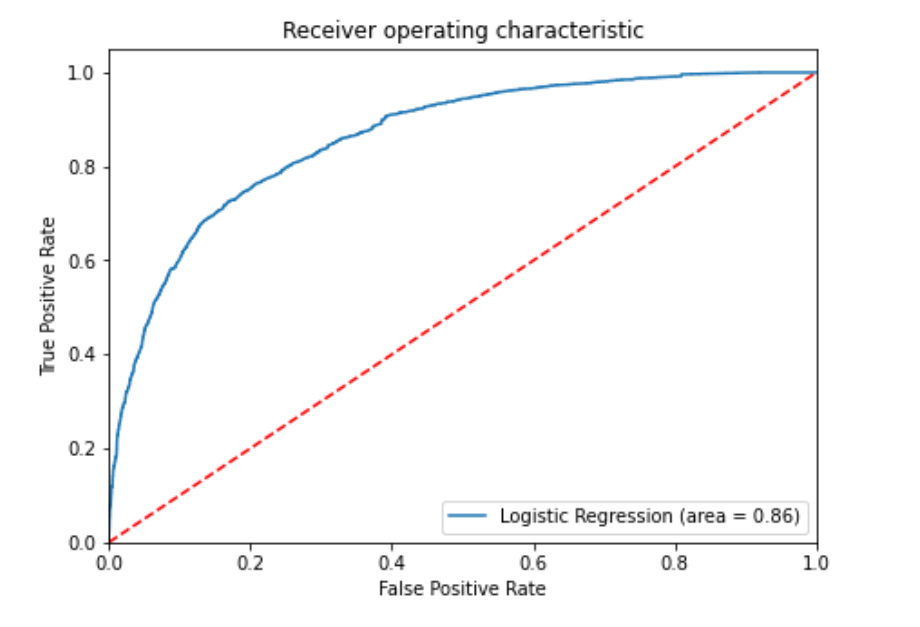
- **VIF standards:**
  - **If VIF is between 1 and 5, then there is low multicollinearity.**
  - **If VIF is between 5 and 10, we say there is moderate multicollinearity.**
  - **If VIF is exceeding 10, it shows signs of high multicollinearity.**

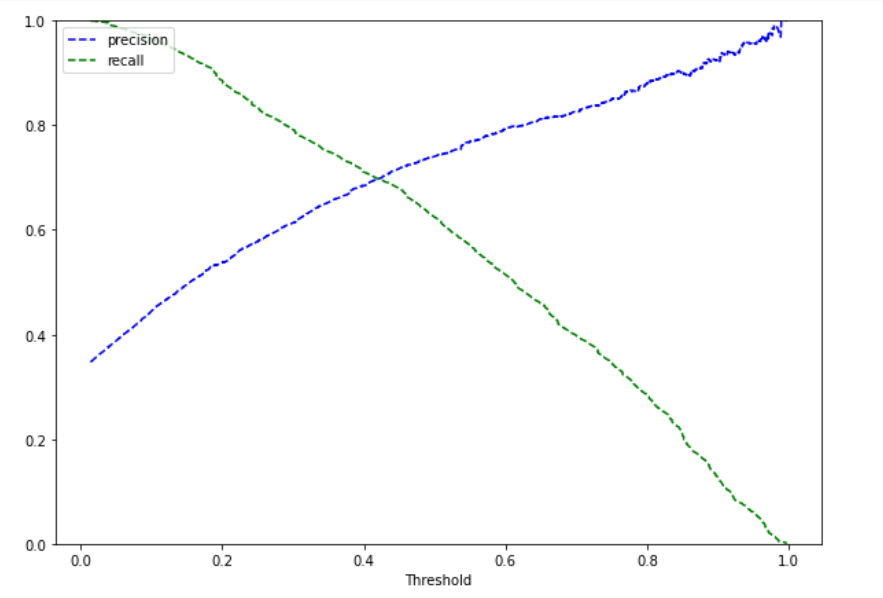
- (-) coefficient indicates the probability of a hotel booking being cancelled decreases with increase of corresponding attribute value
- (+) coefficient indicates probability of a hotel booking being cancelled increases with increase of corresponding attribute value
- p-value of a variable indicates if the variable is significant or not
  - If we consider the significance level to be 0.05 (5%), then any variable with a p-value < 0.05 would be considered significant. Thus, variables might contain multicollinearity, which will affect the p-values
  - Might have to eliminate multicollinearity from the data to get reliable coefficients and p-values
- As there is no multicollinearity, can look at the p-values of predictor variables to verify significance
- Of the 20 variables in this logistic regression model, 7 of predictor variables are shown to be statistically insignificant

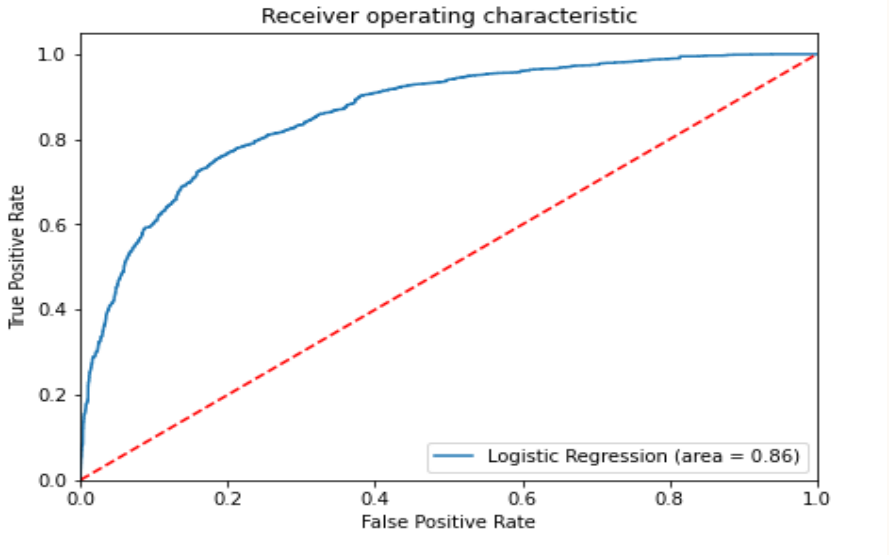
OBSERVATIONS: COEFFICIENT INTERPRETATIONS

- *repeated_guest*: Holding all other features constant, a unit change in repeated_guest decreases the odds that booking being canceled by 0.06 times or a **94%** decrease in odds of a booking being canceled
- *market_segment_type_Offline*: Holding all other features constant, a unit change in market_segment_type_Offline decreases the odds that booking being canceled by 0.17 times or an **83%** decrease in odds of a booking being canceled
- A guest requiring a parking space decreases the odds of that booking being canceled by 0.20 times or an **80%** decrease in odds of a booking being canceled
- A single additional increase in the average cost per night increases the odds of that booking being canceled by 1.02 times or a **2%** increase in the odds of a booking being canceled
- A single additional night booked at the hotel increases the odds of that booking being canceled by 1.96 times or a **6%** increase in the odds of a booking being canceled

PERFORMANCE EVALUATION



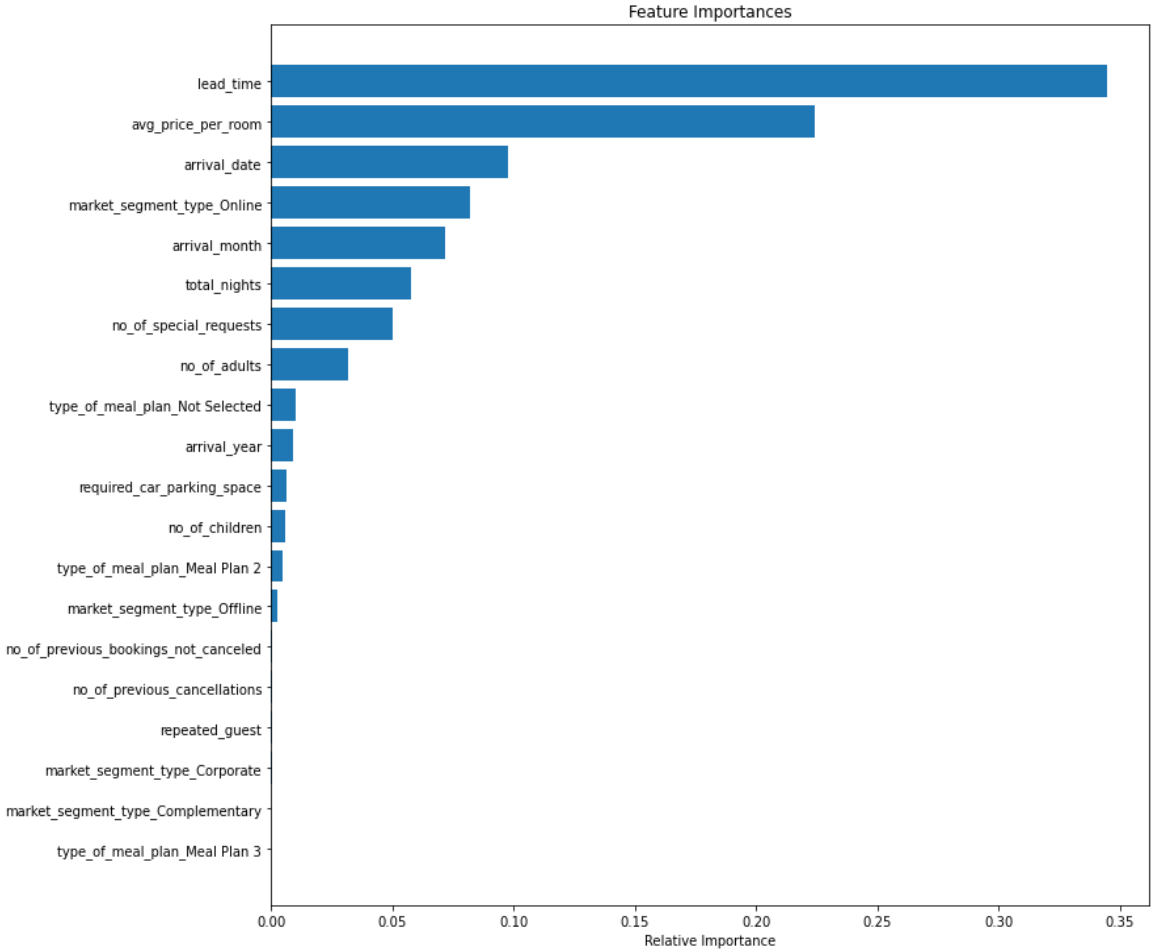




OBSERVATIONS

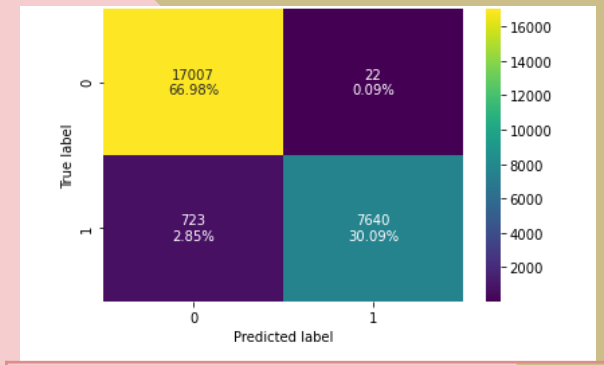
- Predictive model was developed that can be utilized by INN Hotels to predict which bookings can be canceled with an F1 score of 0.70 on the training set
- Logistic regression models gave generalized performance on both the training and test set, illustrating similar model can be generated for INN Hotels in production
- Coefficients for the number of adults, the lead time prior to a booking, the arrival year (i.e., 2018 v. 2017), the number of previous cancellations, the average room price, the total nights booked, selecting Meal Plan 2, and not selecting a meal plan are (+) indicating increase chances of hotel booking being cancelled
- Coefficients for requiring a parking space, arrival month, being a repeat guest, the number of special requests, and the market segments for Corporate and Offline are (-) indicating an increase will lead to decrease in chances of a hotel booking being cancelled

DECISION TREE MODEL

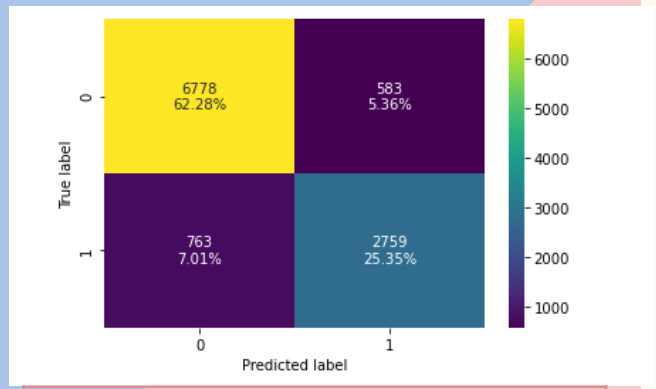


- *lead_time* and *avg_price_per_room* are the two most important variables for predicting a booking's cancellation by a factor of 3 and 2, respectively, over the third most important variable, *arrival_date*.

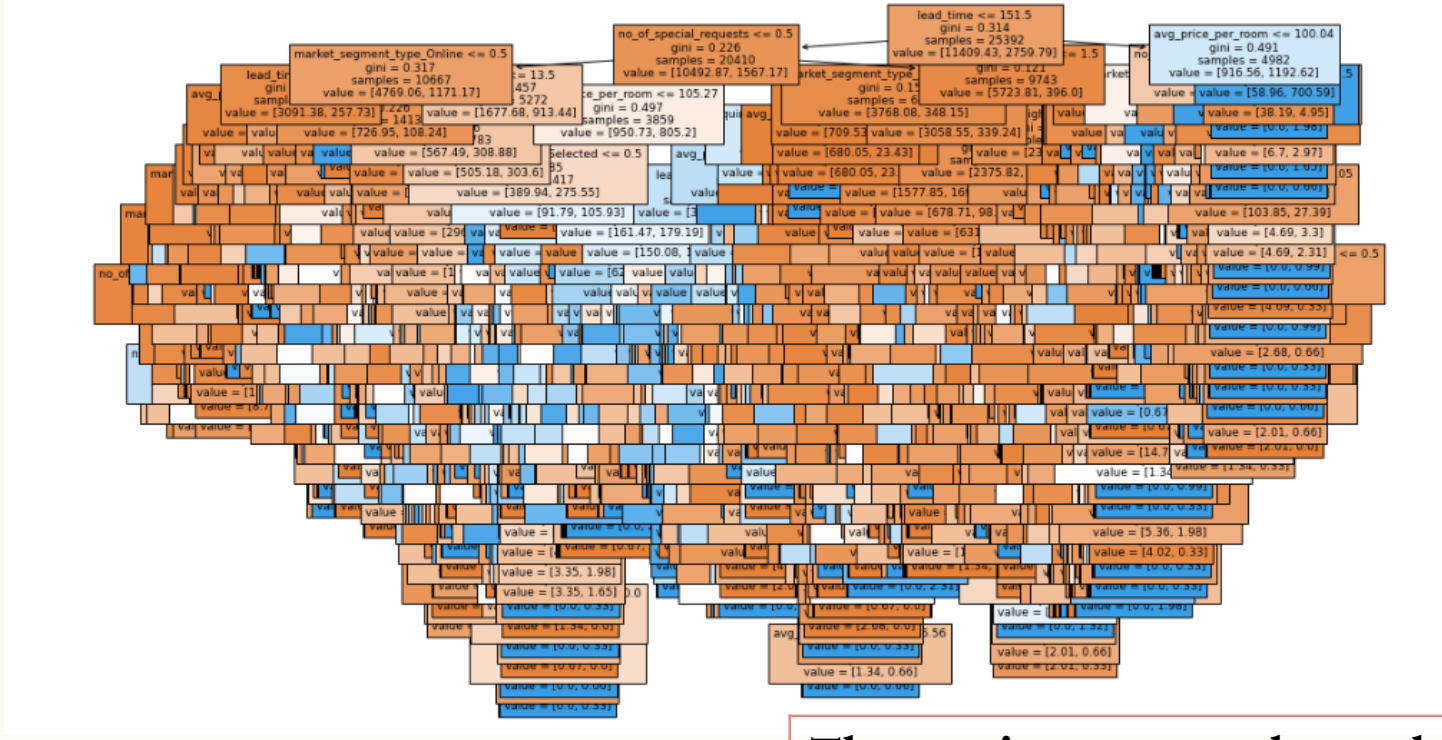
DECISION TREE VISUALIZATION



Performance on training set

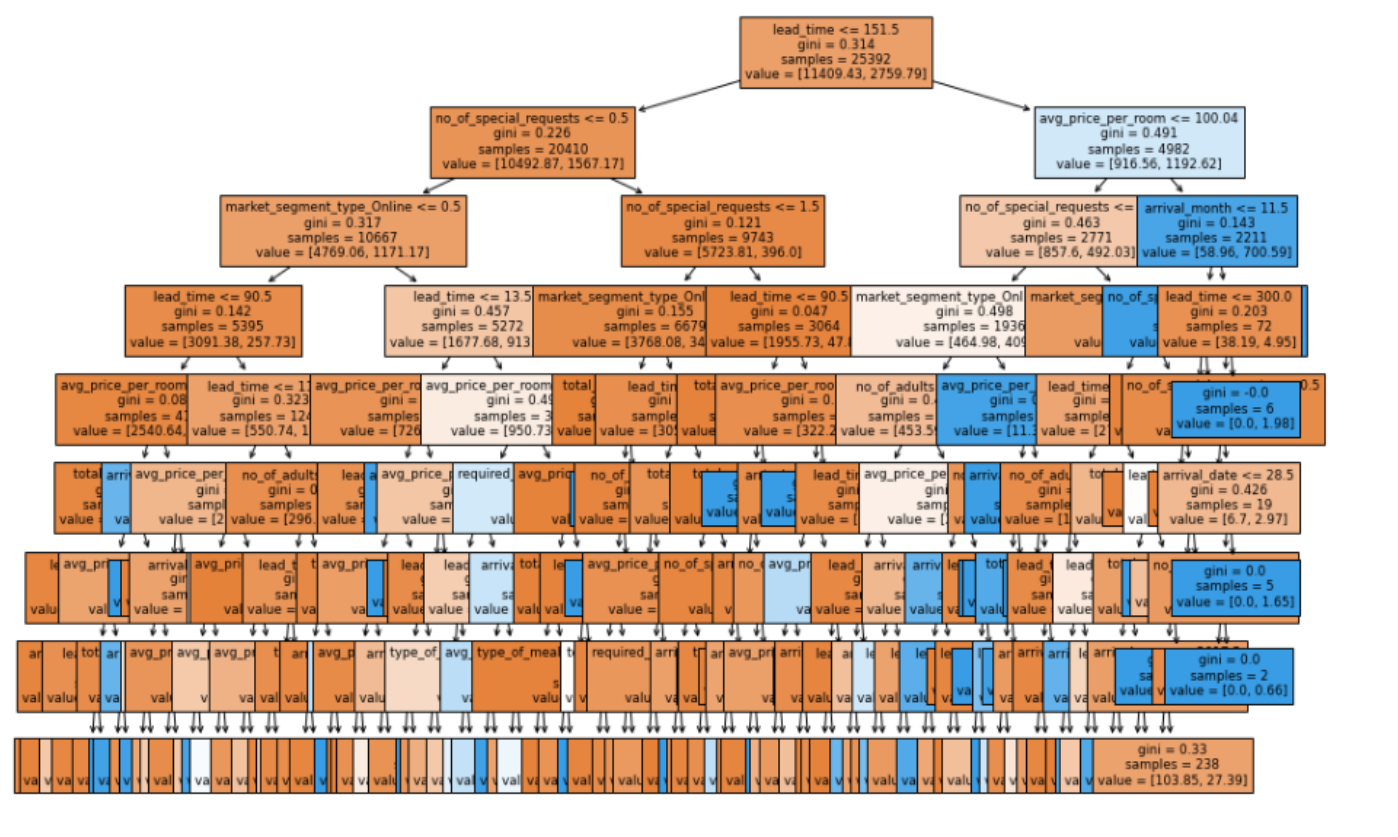


Performance on test set



The tree is very complex and difficult to interpret

VISUALIZING DECISION TREE



Decision tree is less complex and easier to analyze

Training performance comparison:

	Decision-Tree	Pre-Pruned Tree	Post-Pruned Tree	Readible Tree
F1	0.99	0.96	0.95	0.75

Test performance comparison:

	Decision-Tree	Pre-Pruned Tree	Post-Pruned Tree	Readible Tree
F1	0.80	0.81	0.80	0.75

CONCLUSIONS

- Cancelation decisions of 36,275 bookings were analyzed via four different Decision Tree classifiers to generate a predictive model
  - The models can be utilized by INN Hotels Group to validate if a booking will be canceled prior to the check-in date
- All decision-tree models perform better through objective criterion of F1 score than best-performing logistic regression model
- Each model's decision-tree and confusion matrix were investigated to get a better understanding of each model. Thus, the predictions from the original, pre-pruned, and post-pruned decision-tree model would be difficult to interpret by the client
- Pre-pruning and post-pruning methods resulted in minimal reductions to overfitting.
- However, the best performing model, the pre-pruned decision-tree, still shows a significant disparity between its performance on the training dataset and the test dataset.
  - This shows that the model's predictions, as compared to the best logistic regression model, may not be as generalizable
- INN Hotels should consider the tradeoff that exists for decision-tree models with respect to performance, overfitting, and understanding the decision-making criteria of the model
  - If a more understandable prediction model is the objective, then the minimal depth required for this decision-tree model to perform better than the best logistic regression model, is 8.
    - However, if the tree still too complex for the client's use-case and preferences, than potentially the logistic regression model is the ideal prediction tool
  - If INN Hotels is seeking the best performing prediction model then the pre-pruned tree is the best option
- Based on all models, *lead_time* and *avg_price_per_room* were the two most important variables for predicting a booking's cancellation

CONCLUSIONS

- From our analysis, it shows that guests booking cheaper rooms, with shorter lead times, requiring a parking space, being a repeat guest, with higher number of special requests, from the Corporate and Offline market segments are less likely to cancel bookings
    - Thus, guests booking more expensive rooms, with longer lead times, through the Online market segment have greater chance to cancel bookings
- Based on the coefficients in the logistic regression models and the features in the decision-tree models, both prediction models indicate that INN Hotels should consider separate cancellation and refund policies for its guests travelling for business or personal reasons
- The data analysis suggests that introducing a rewards program for business travelers (e.g., requiring more frequent trips, booking on short notice from a corporate sales channel, and directed by corporate travel guidelines to book a room with the lowest available cost) should further incentive these guests to book at INN Hotels and follow-through on their travel plans
- Moreover, when a hotel is at capacity or overbooked, management could utilize the model to ensure all repeat guests or guests travelling for business reasons have rooms available
    - Thus, management can combine predictions from both models to identify the "most likely case" that a booking will be canceled and reallocate that room to a booking for that room category which is the "least likely case"
    - A disclaimer should be in place that this model should not replace the industry experience of its management team regarding managing its hotel's capacity
        - Moreover , the models should provide supplemental evidence in support of its decision-making process

RECOMMENDATIONS

- To further improve the utility of the models, the hotel can provide approximations of the costs related to the outcomes corresponding to true/false positives/negatives
- The team can optimize the models' predictions to achieve the highest expected profits, versus optimizing for F1 score, which we chose for our evaluation criteria based on the client's use-case
- INN hotels should gear towards offering the best room rates before 5 months and after can increase the prices slightly and increase profit
- Implementation of nonrefundable deposit should be in place on all rooms in advance of over 5 months
- The 'Full Board' option should be replaced on the booking with a menu of special requests available such as:
  - VIP a champagne toast at sunset your first night; Room upgrades; WiFi; Laundry Bag; Slippers

THANK YOU