Business Presentation

by Narges Shahmohammadi

Contents

In this presentation, we are going to get deep into the Star Hotels and elements which are effective on cancelation of booking the room

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

Business Problem Overview and Solution Approach

- The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled.

- We as data scientists must analyze the data provided to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

Data Overview

- We have a data with 56926 rows and 18 columns
- There is no missing values.
- Also, we have outliers in our data

Some columns need to be dummy like type_of_meal_plan or room_type_reserved

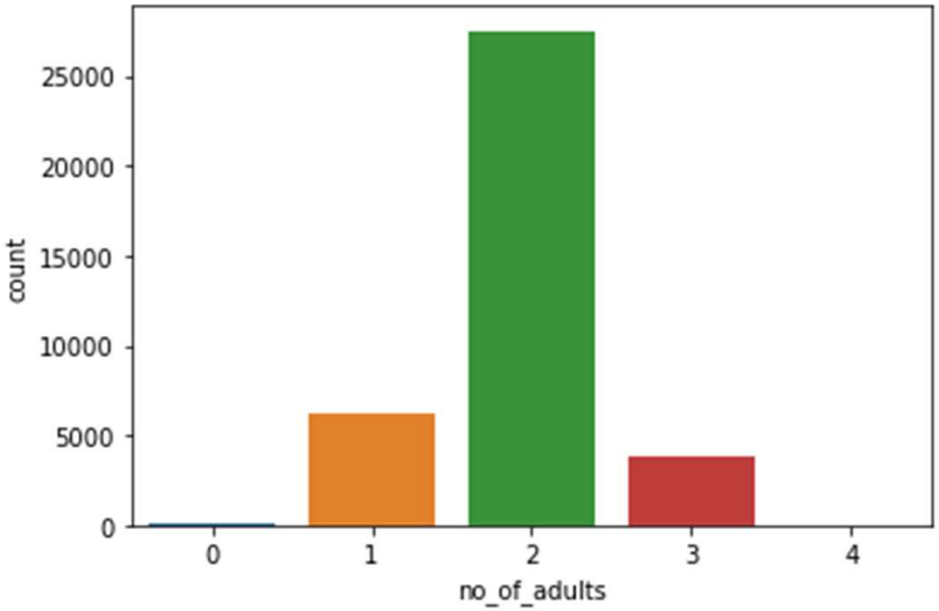
no_of_adults: Number of adults
no_of_children: Number of Children
no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
type_of_meal_plan: 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 Star Hotels Group
lead_time: Number of days between the date of booking and the arrival date
arrival_year: Year of arrival date
arrival_month: Month of arrival date
arrival_date: Date of the month
market_segment_type: Market segment designation.
repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
booking_status: Flag indicating if the booking was canceled or not.

We have 14350 duplicates value which need to delete.

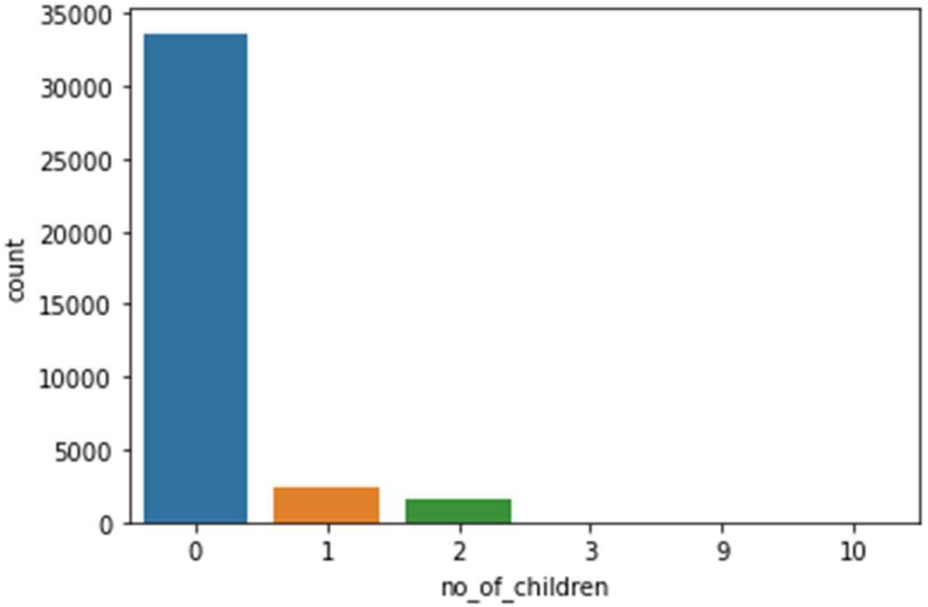
13 columns are integer.
4 columns are object: (booking_status)(market_segment_type)(room_type_reserved)(type_of_meal_plan)
1 column is float: (avg_price_per_room)

EDA

- The reservation is mostly for two adults.

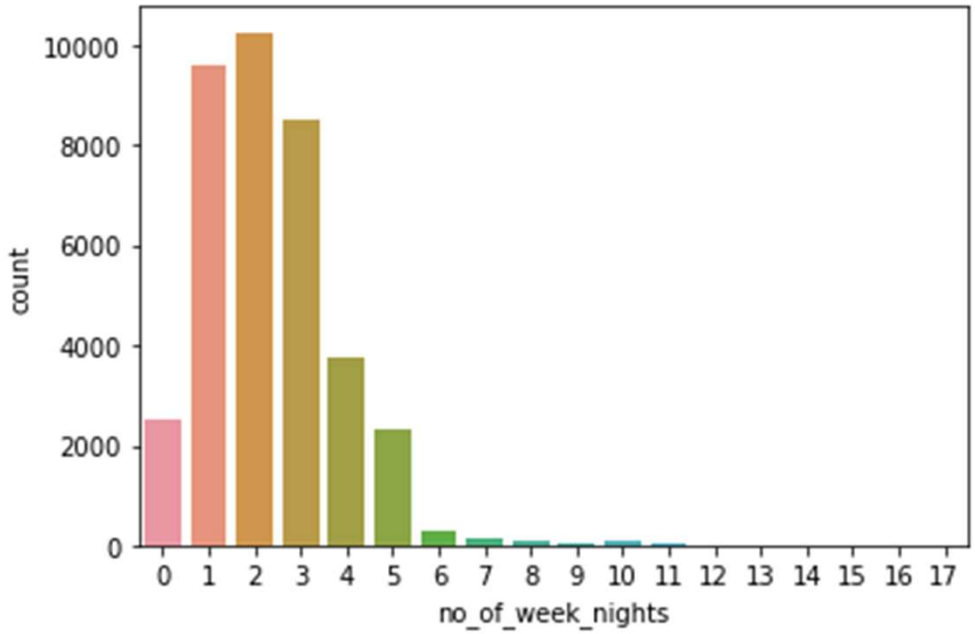


- most of reservation is with no children



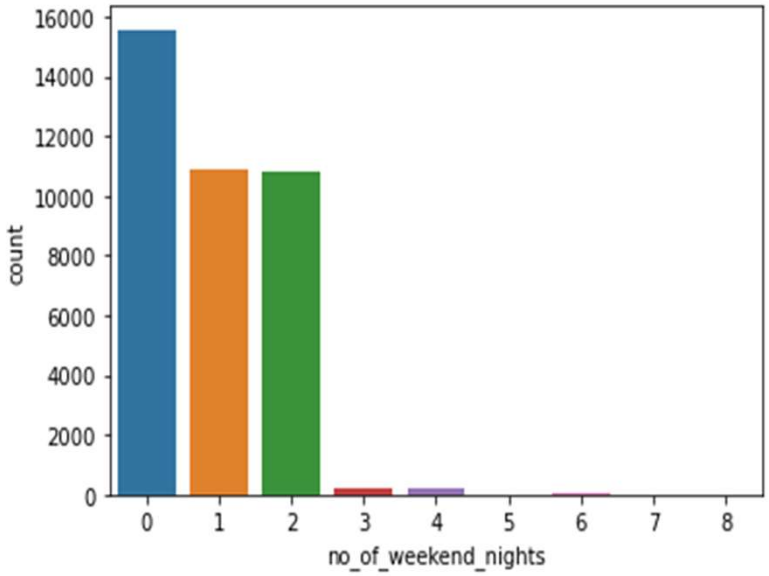
EDA

- Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel. as you can see, 2 , 1 and then 3 weeknights are on the top

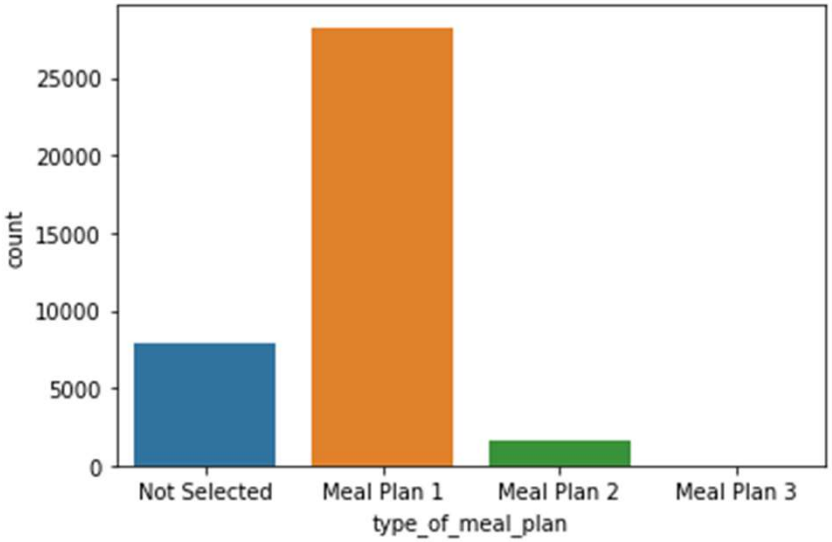


EDA

- Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel. As you see, most of the guests tend to do not book or stay at hotel at the weekends.

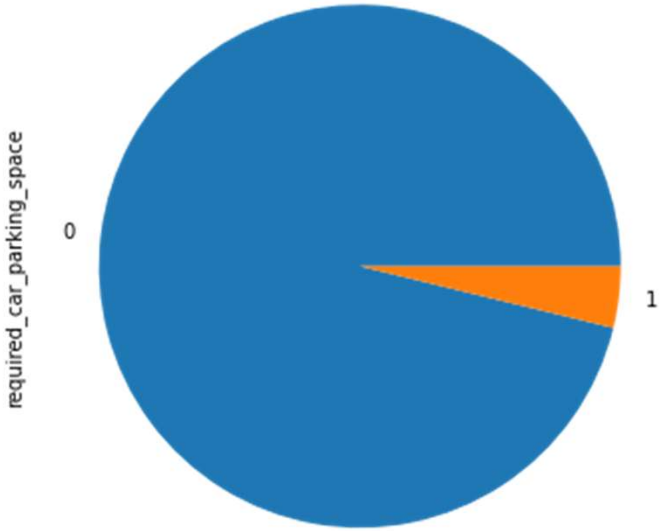


plot of type of meal plan booked by the customer shows us that most of them just booked for Breakfast (Meal plan 1)

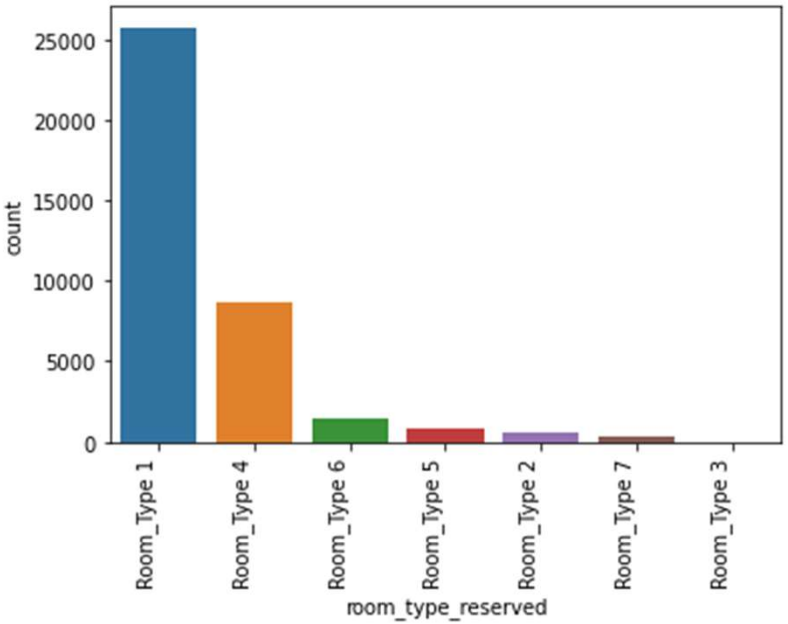


EDA

- Only a few customers require car parking space

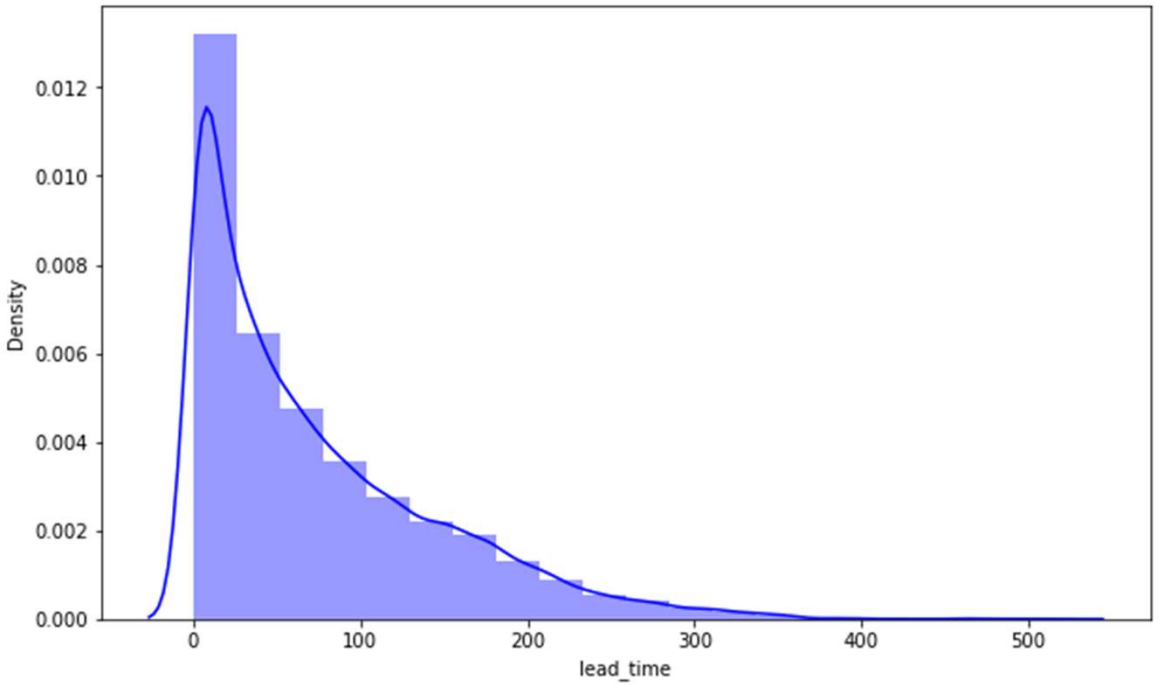


bar plot of type of room reserved by the customer shows that room type 1 is the most popular.



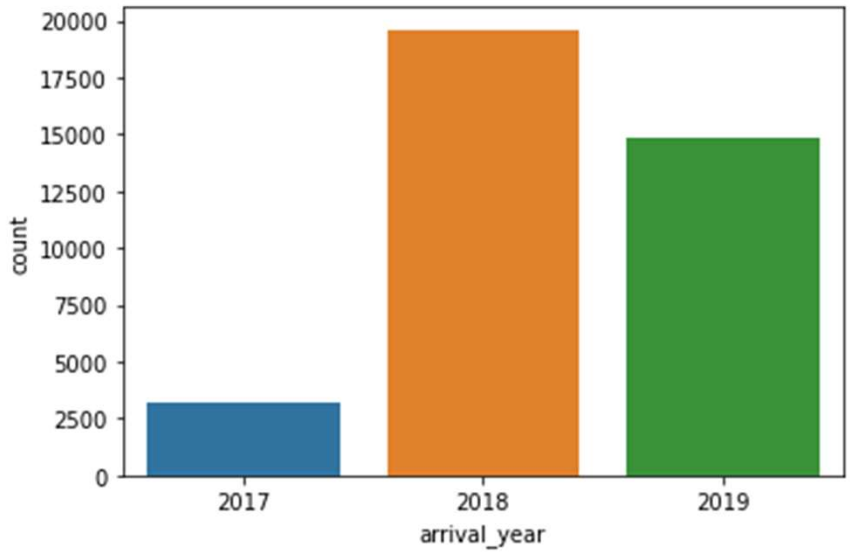
EDA

- Number of days between the date of booking and the guest arrival date have a highly skewed to the right distribution. It also shows that guests tend to book hotel rooms mostly between 0 to 100 days earlier.

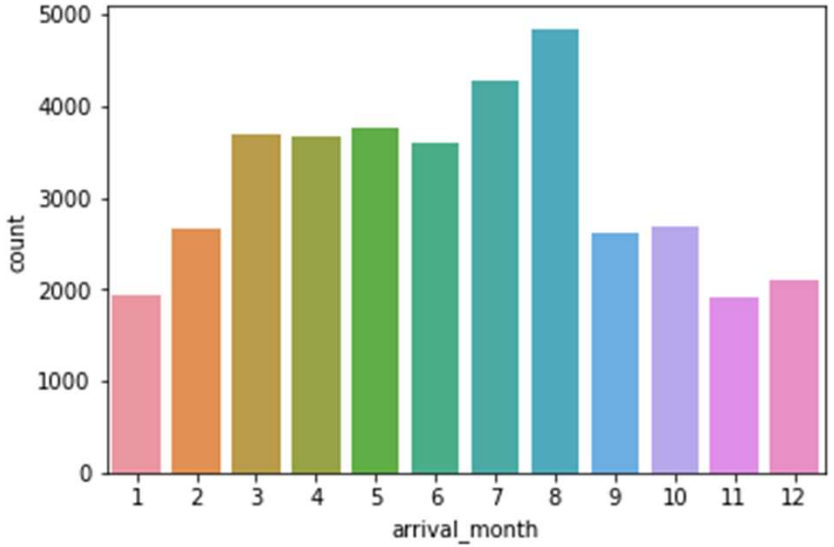


EDA

arrival date for 2018 is on the top

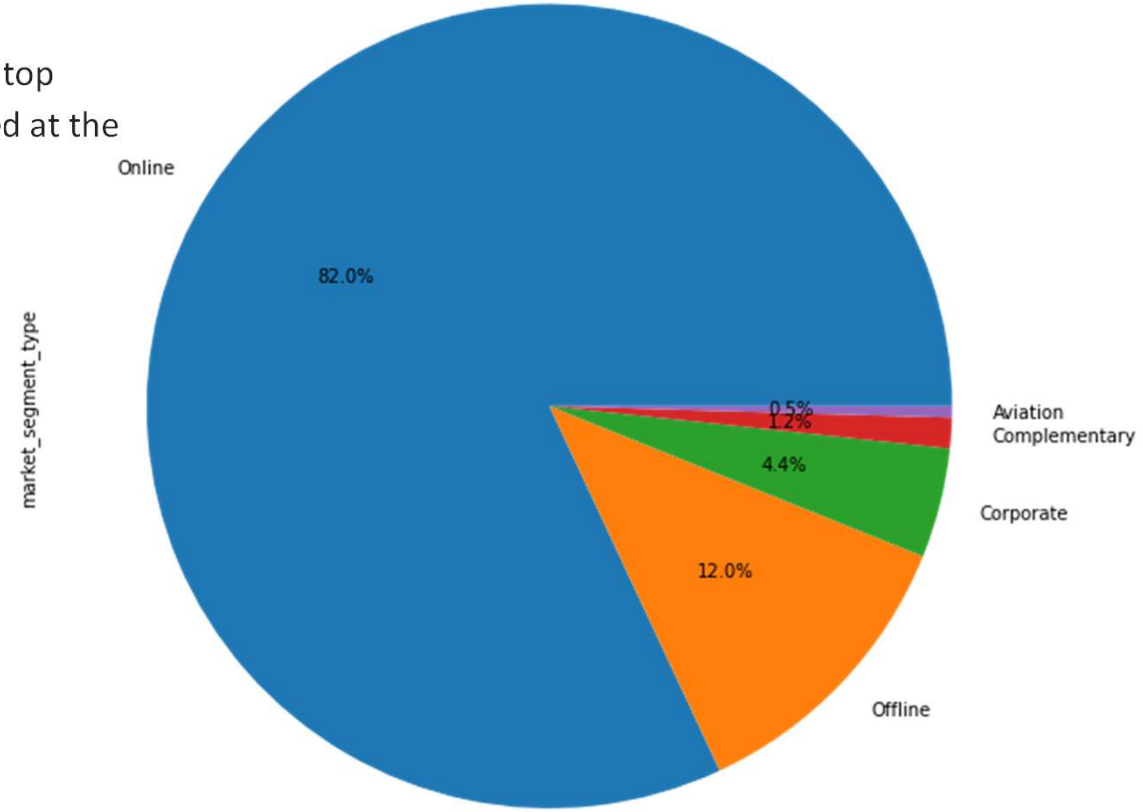


most of the reservations of hotel rooms are for the month of July to August. Also, first and the last two months have the lowest reservation.



EDA

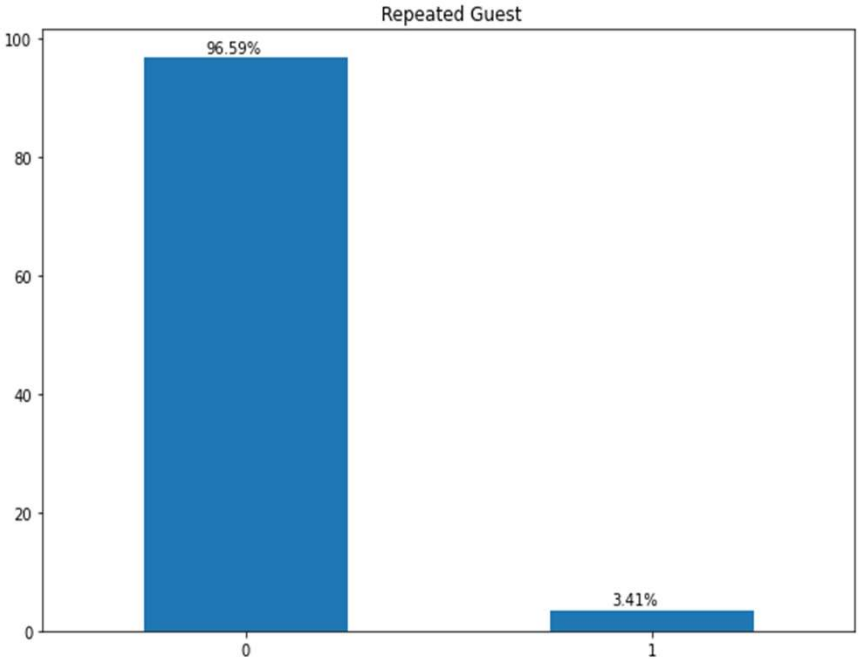
online market segment with 82% is on the top after that offline with 12% has been located at the second place

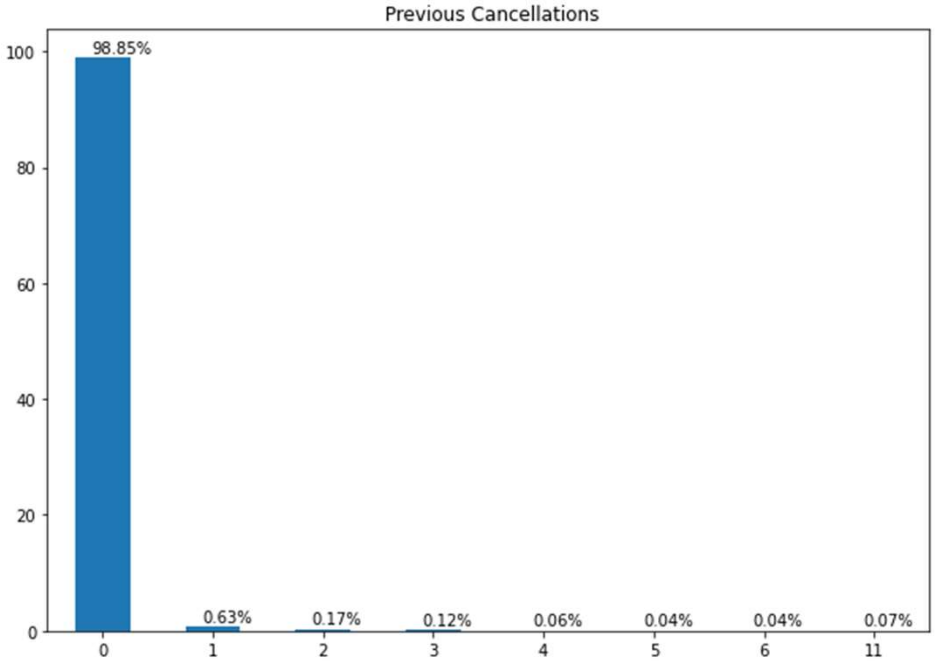


EDA

only 3.41% of the guests are repeated guest

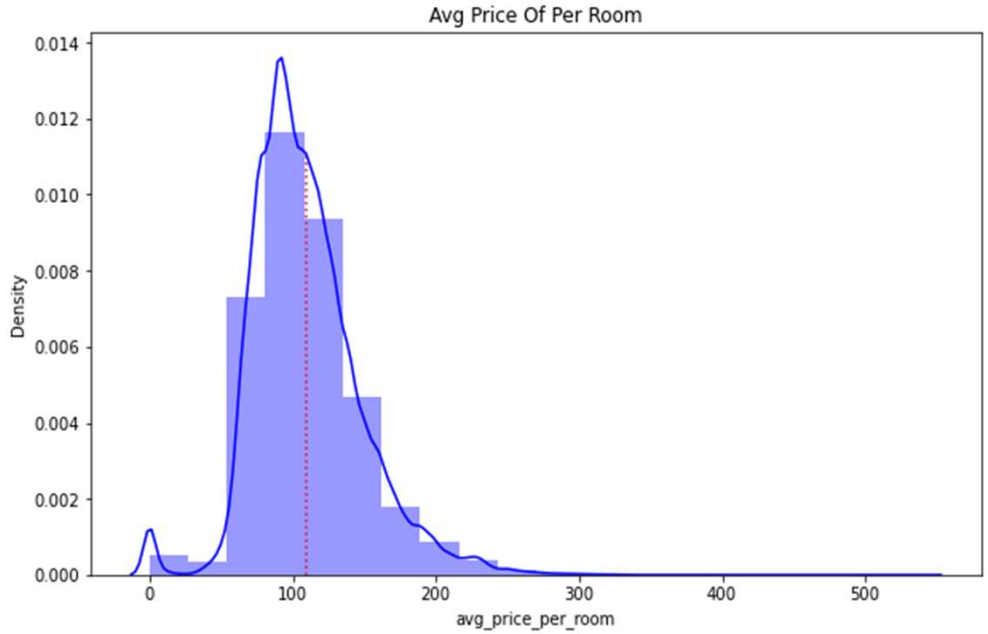
almost all of the guests (99%) have no previous bookings that were canceled





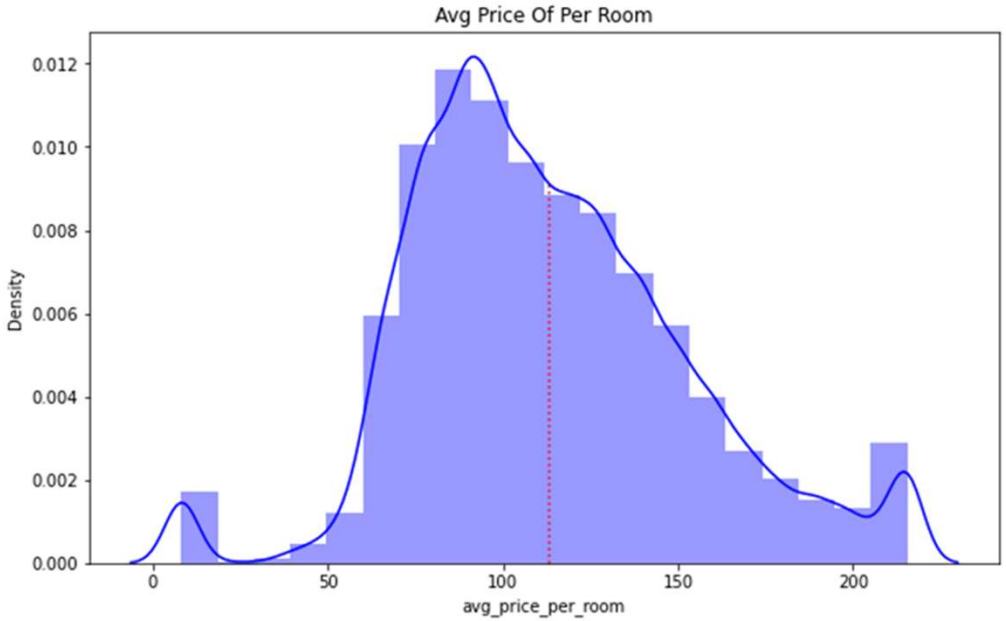
EDA

avg_price_per_room has a distribution skewed to the right. I think it is look like a normal distribution. its average is around 109 Euros.



EDA

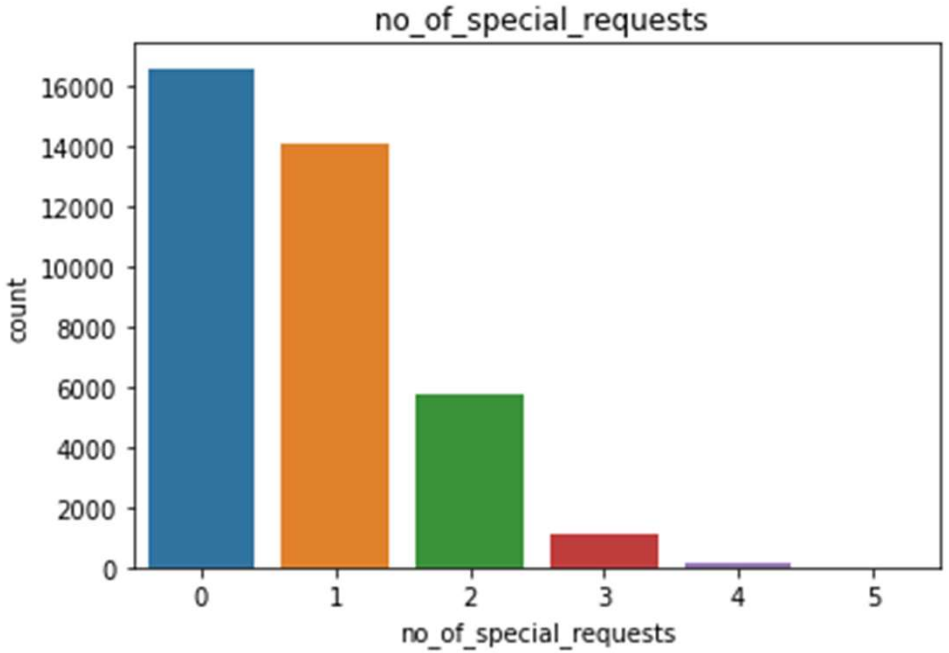
After dealing with outliers its skewness is better and still its mean is around 113 Euro.



EDA

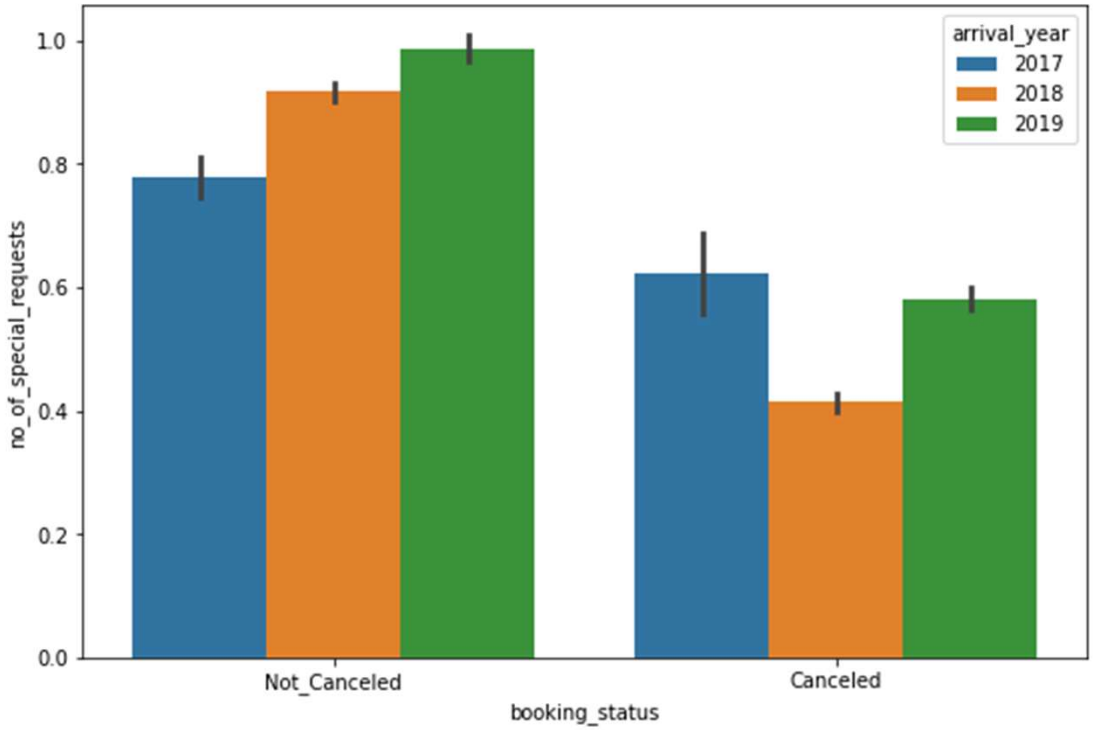
Most of the customers had no special request (16581)
After that there is 1 request with number of 14117

```
0    16581
1    14117
2     5755
3     1120
4      138
5       14
Name: no_of_special_requests
```



EDA

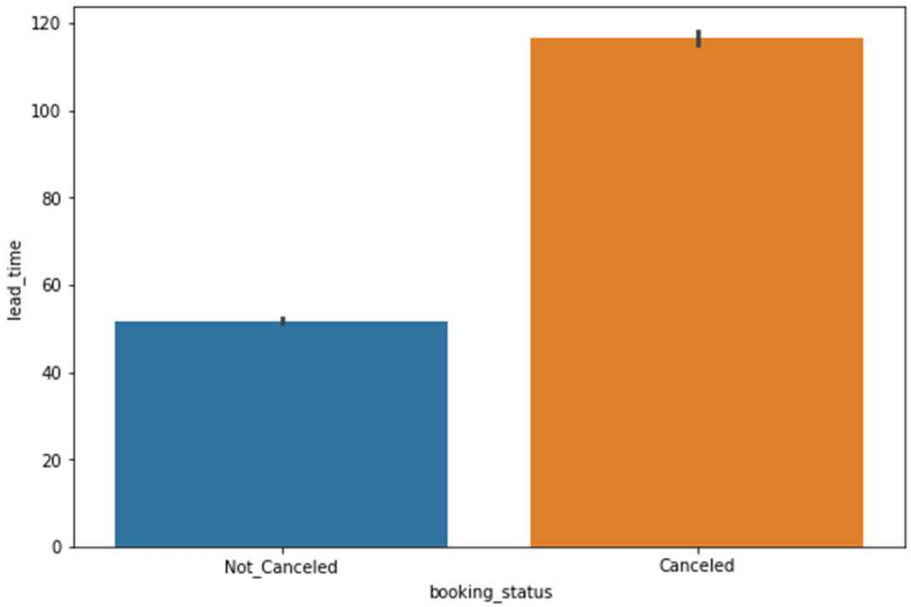
- difference between booking status and the number of special requests in different years. Totally, in 2019 booking the rooms with no cancelation and a greater number of requests are higher than others. in 2017 booking the rooms and its cancelation (with high special request) are higher than others.

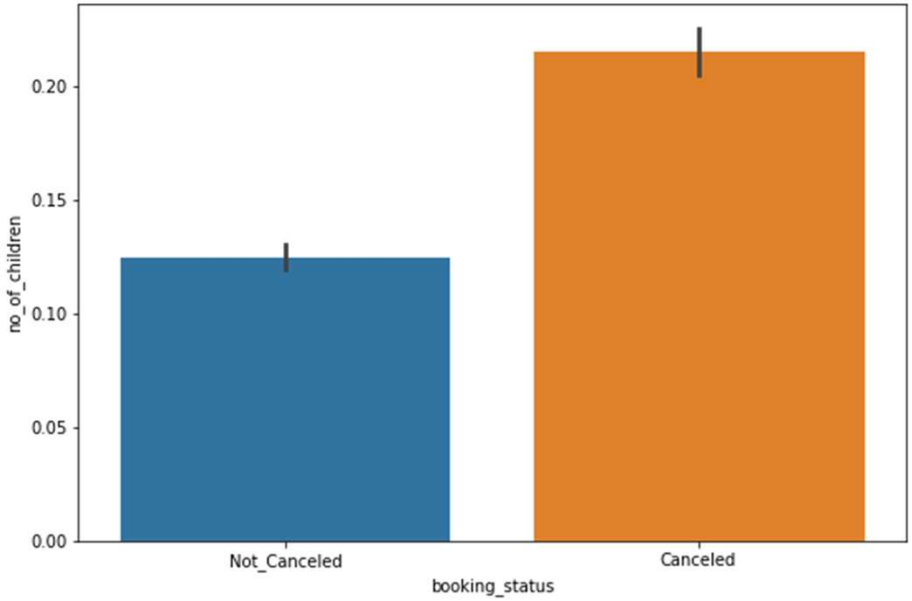


EDA

- The more the lead time, the more cancelation it is.

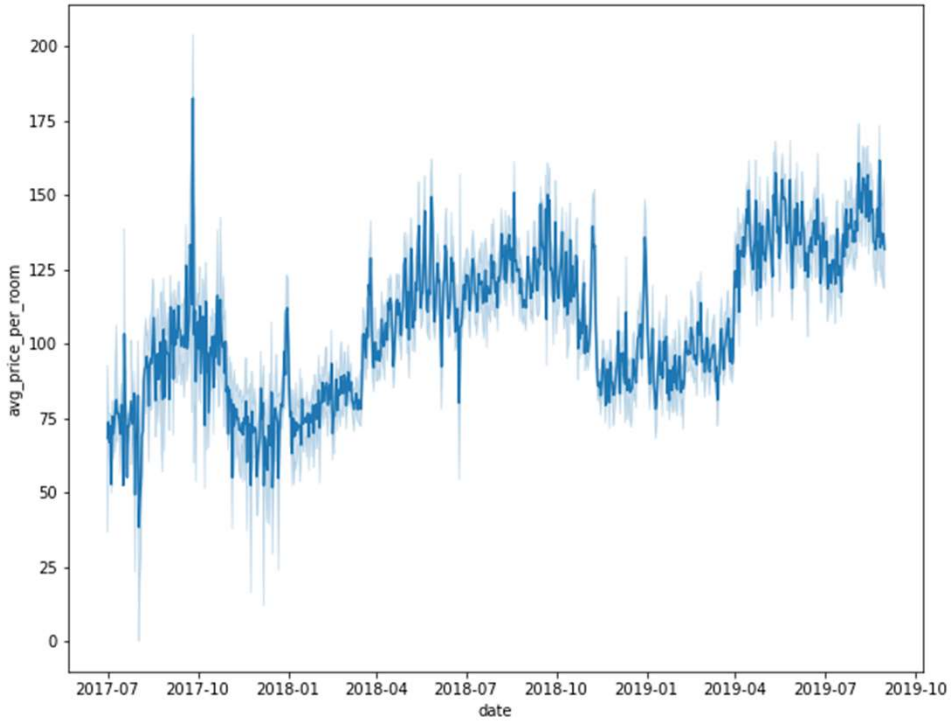
Guests who have more children did cancel their reservations more than others.





EDA

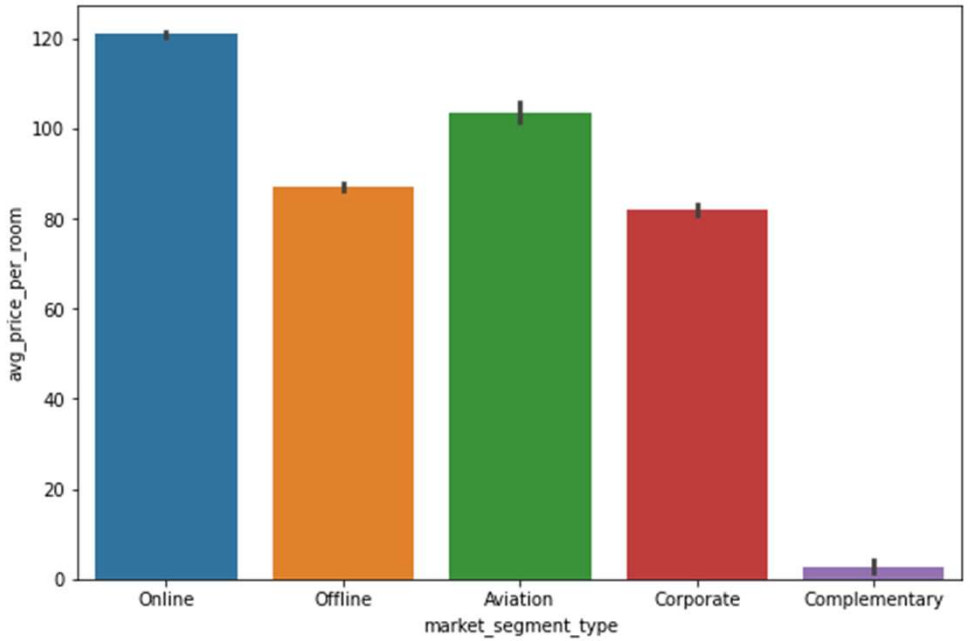
- There is a increasing in price of per room between 2017 to 2019 with a mild slope



EDA

Hotel rates are dynamic and change according to demand and customer demographics. What are the differences in room prices in different market segments?

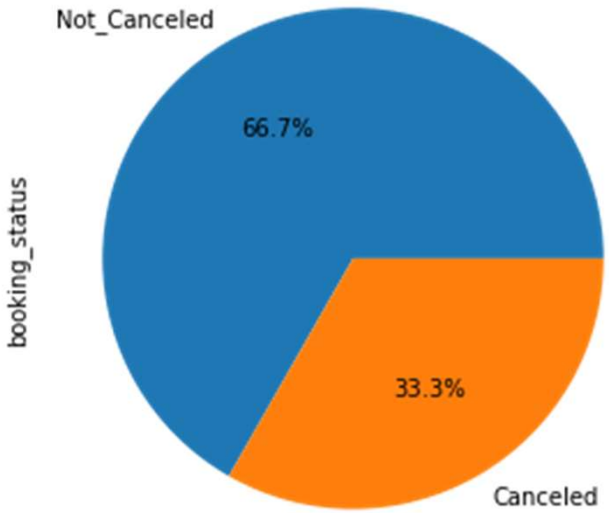
The online and complementary market segments have the highest and lowest average price, respectively.



EDA

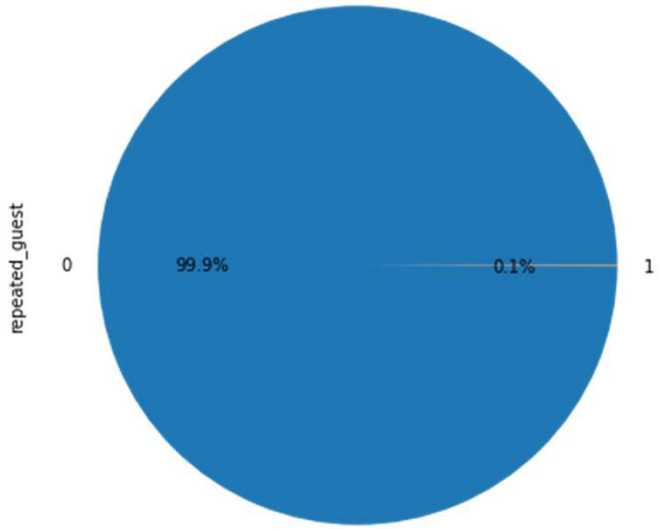
What percentage of bookings are canceled?

33.3% of guests did cancel their books.



Repeating guests are the guests who stay in the hotel often and are important to brand equity. What percentage of repeating guests cancel?

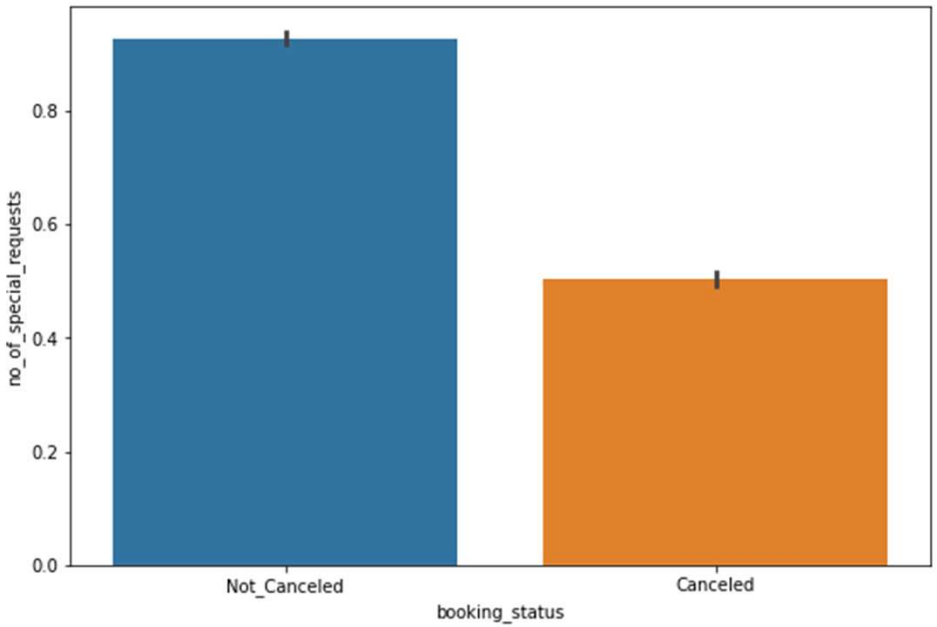
less than 1% of repeated guests cancel their reservations



EDA

Many guests have special requirements when booking a hotel room. Do these requirements affect booking cancellation?

It seems that guests who have high special requirements tend to do not to cancel their books.



Model Performance Summary

- Overview of the model and its parameters

1. Create Dummy Variables for 3 columns:[type_of_meal_plan, room_type_reserved, market_segment_type]
2. Convert column [booking_status] values to 1 & 0
3. Then I split the data:

independent variables
x = all my data but: booking_status, market_segment_type_Aviation, room_type_reserved_Room_Type 7, type_of_meal_plan_Meal Plan 2

dependent variable
y = booking_status

Model Performance Summary

- In order to make statistical inferences from a logistic regression model, it was important to ensure that there is no multicollinearity present in the data.
- I checked its multicollinearity and figured out that some columns have high multicollinearity like [some other columns exhibit high multicollinearity] and so on. I removed them one by one to see which variable has a significant impact on the model's performance. The final list is:

```
Series before feature selection:

no_of_children                          2.101811
no_of_weekend_nights                    2.084844
no_of_week_nights                       3.598011
required_car_parking_space              1.072003
lead_time                               2.325369
arrival_month                           4.115755
arrival_date                            3.291890
repeated_guest                          2.151598
no_of_previous_cancellations            1.560862
no_of_previous_bookings_not_canceled    1.943311
no_of_special_requests                  2.011545
type_of_meal_plan_Meal Plan 3           1.017532
type_of_meal_plan_Not Selected          1.399367
room_type_reserved_Room_Type 2          1.100057
room_type_reserved_Room_Type 3          1.001036
room_type_reserved_Room_Type 4          1.483210
room_type_reserved_Room_Type 5          1.044264
room_type_reserved_Room_Type 6          1.902410
market_segment_type_Complementary       1.162966
market_segment_type_Corporate           1.723579
market_segment_type_Offline             1.248853
dtype: float64
```

Coefficient interpretations

1. no_of_children: Holding all other features constant a 1 unit change in no_of_adults will increase the odds of a cancelation by 1.2 times or a 24.0% increase in odds of cancel the book.
2. no_of_weekend_nights: Holding all other features constant a 1 unit change in the no_of_weekend_nights will decrease the odds of a guest cancel the book by 0.95 times or a decrease of 4.49% decrease in odds of a guest cancel the book.
3. no_of_week_nights: Holding all other features constant a 1 unit change in no_of_week_nights will decrease the odds of a cancelation by 0.97 times or a 2.68% decrease in odds of cancel the book.
4. required_car_parking_space: Holding all other features constant a 1 unit change in required_car_parking_space will decrease the odds of a cancelation by 0.23 times or a 76.3% decrease in odds of cancel the book.
5. lead_time: Holding all other features constant a 1 unit change in lead_time will increase the odds of a cancelation by 0.93 times or a 6.5% increase in odds of cancel the book.
6. arrival_month: Holding all other features constant a 1 unit change in arrival_month will decrease the odds of a cancelation by 0.94 times or a 5.0% decrease in odds of cancel the book.
7. no_of_previous_cancellations: Holding all other features constant a 1 unit change in no_of_previous_cancellations will decrease the odds of a cancelation by 0.117 times or a 88.0% decrease in odds of cancel the book

Since the amount of them are much here I just wrote some of them

Final Model Performance Summary {test & train set}

- All the logistic regression models have given a generalized performance on the training and test set.
- Recall score shows the best result on logistic regression 0.36 threshold (0.74)

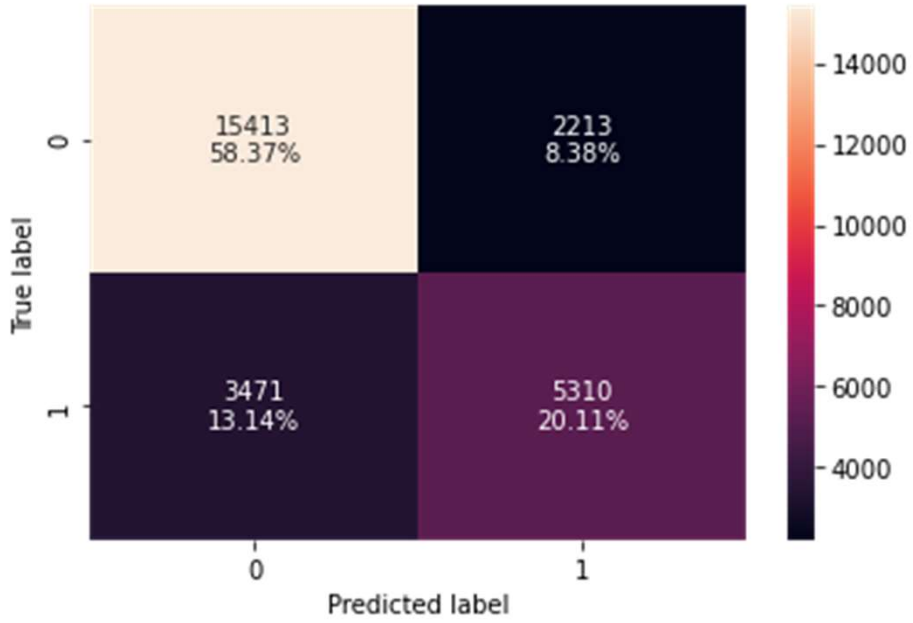
	Logistic Regression statsmodels	Logistic Regression 0.36 Threshold	Logistic Regression 0.51 Threshold
Accuracy	0.785	0.763	0.785
Recall	0.605	0.751	0.595
Precision	0.706	0.619	0.712
F1	0.651	0.679	0.649

test performance comparison:

	Logistic Regression statsmodels	Logistic Regression 0.36 Threshold	Logistic Regression 0.51 Threshold
Accuracy	0.783	0.758	0.784
Recall	0.606	0.746	0.594
Precision	0.701	0.612	0.709
F1	0.650	0.672	0.646

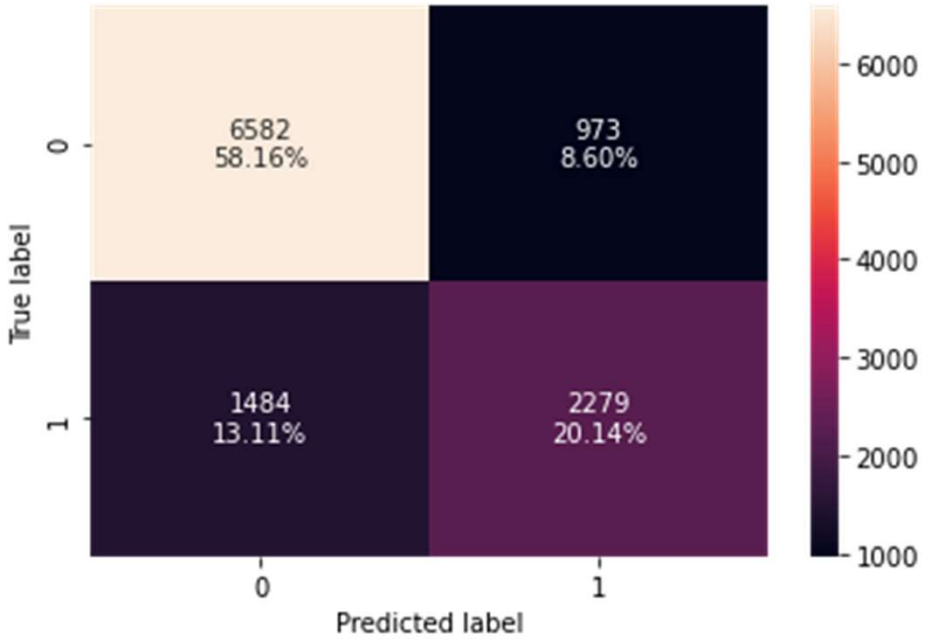
Confusion Matrix {Train set}

- True Positives (TP): we correctly predicted that they do NOT CANCEL 58.37%
- True Negatives (TN): we correctly predicted that they do CANCEL 20.11%
- False Positives (FP): we incorrectly predicted that they do CANCEL (a "Type I error") 2213 Falsely predict positive Type I error (8.38%)
- False Negatives (FN): we incorrectly predicted that they don't CANCEL (a "Type II error") 3471 Falsely predict negative Type II error(13.14%)

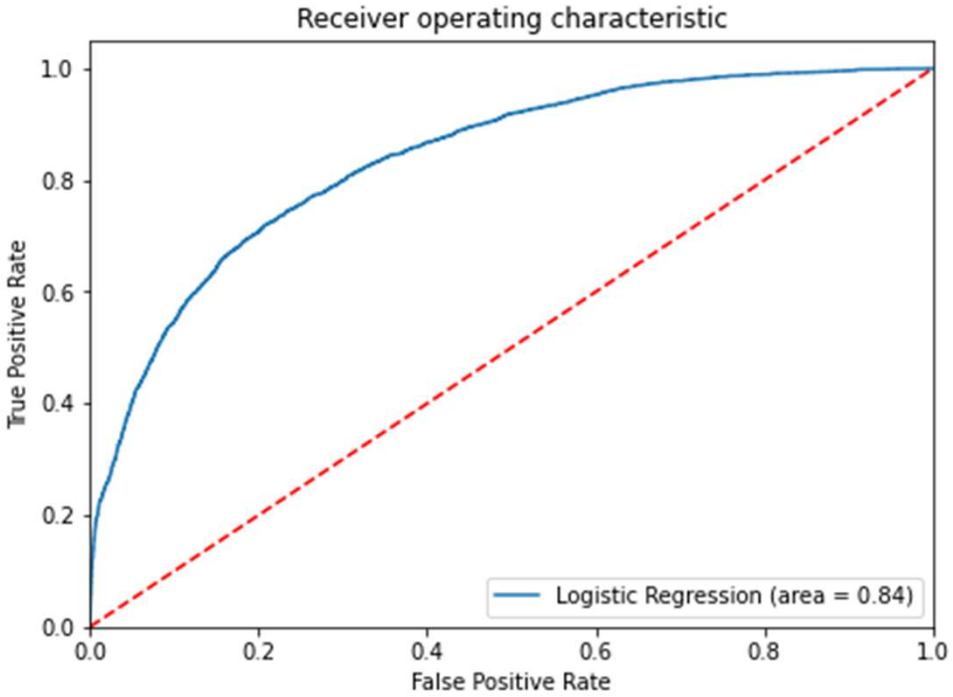


Confusion Matrix {Test set}

- True Positives (TP): we correctly predicted that they do NOT CANCEL 58.16%
- True Negatives (TN): we correctly predicted that they do CANCEL 20.14%
- False Positives (FP): we incorrectly predicted that they do CANCEL (a "Type I error") 937 Falsely predict positive Type I error (8.60%)
- False Negatives (FN): we incorrectly predicted that they don't CANCEL (a "Type II error") 1484 Falsely predict negative Type II error(13.11%)



ROC curve on test set



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

BEST DECISION TREE

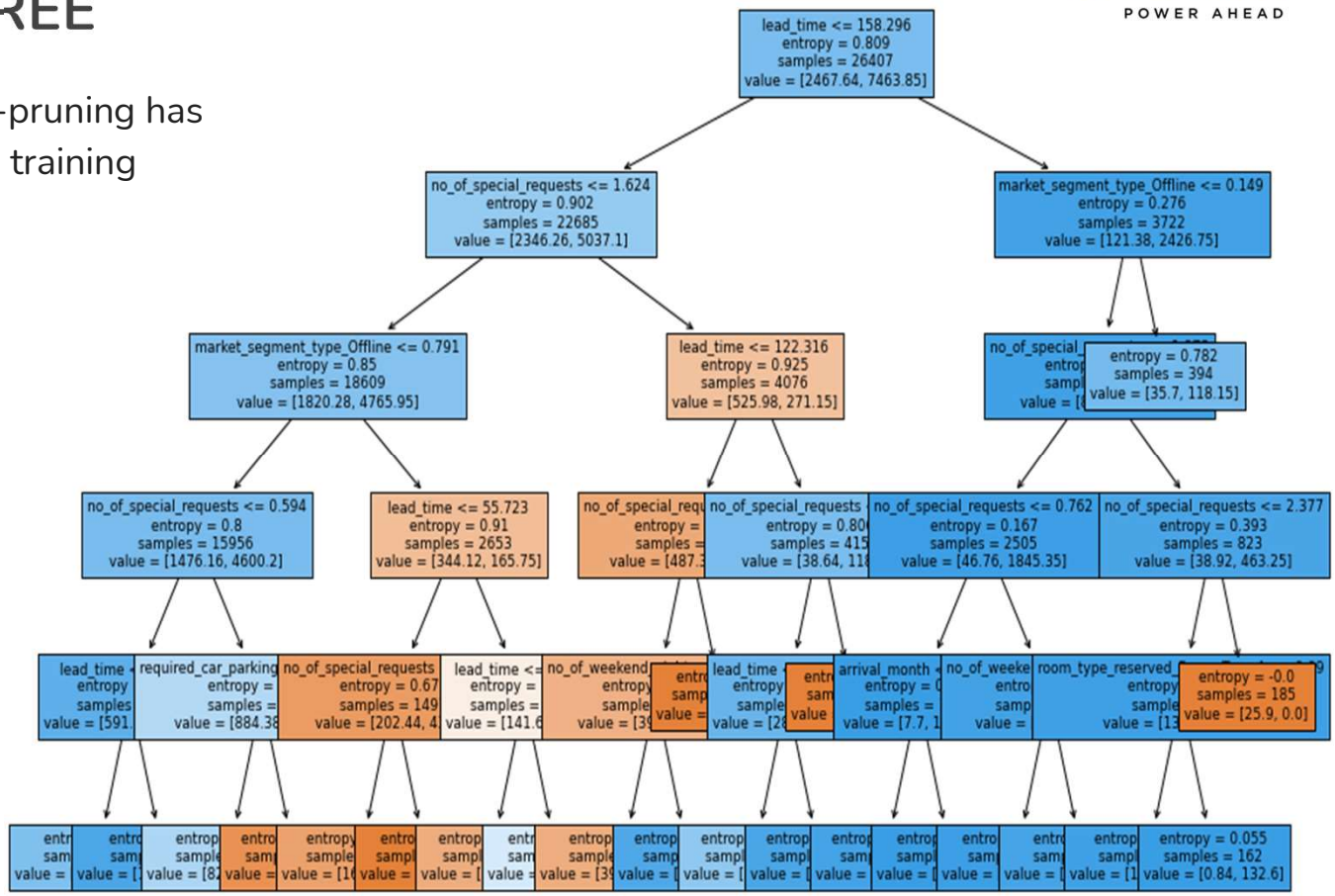
- Decision tree model with pre-pruning has given the best recall score on training data.

testing performance comparison:

	Recall on test set
0	0.622110
1	0.969705
2	0.957215

Training performance comparison:

	Recall on training set
0	1.000000
1	0.971188
2	0.956839



Business Insights and Recommendations

- a) If a guest booking a room with lead_time less than or equal to 158.30 there's a very high chance the guest will not cancel his or her reservation.

- b) If a guest booking a room with lead_time less than or equal to 158.30 and no_of_special_requests is less than or equal to 1.62 then there is a very high chance that the guest will not cancel his or her reservation.

- c)If a guest booking a room with lead_time less than or equal to 158.30 and no_of_special_requests is less than or equal to 1.62 and market_segment_type_Offline is less than or equal 0.79 then there is a very high chance that the guest will not cancel his or her reservation.

- July and August saw the highest number of bookings but also the highest number of cancellations. This should be investigated further by the hotel

- It is observed that the less lead time in the reservation, a smaller number of special requests, and fewer children lead to not cancelation guests booking. By present a good deal or service to people who have a high number of children or have a high special request it may lead to getting a lower chance of canceling between them.

- It is observed that repeating guests, who stay in the hotel often will cancel their booking with a very small percentage (near to zero). so, I think a loyalty program that offers - special discounts, access to services in hotels, etc for these customers can help in improving their experience.

Business Insights and Recommendations

- The online market segment with 82% is on the top of Market segment designation. I believe that managers should invest more in this segment.
- offering a good deal for new guests for second booking can help the hotel to get a high number of repeat guests.

greatlearning
Power Ahead

Happy Learning !