STAR HOTELS PROJECT

Submitted By:
Sathya

OverView

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.

Objective

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. Star Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. You as a data scientist have to 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

The following are the observations in the file :

1) *The file contains 56926 rows with 18 columns*

2) *There are four categorical columns in the file type_of_meal_plan , room_type_reserved, market_segment_type, market_segment_type*

3) *The target variable is Object with values Cancelled & Not_Cancelled*

4) *Other columns are numerical*

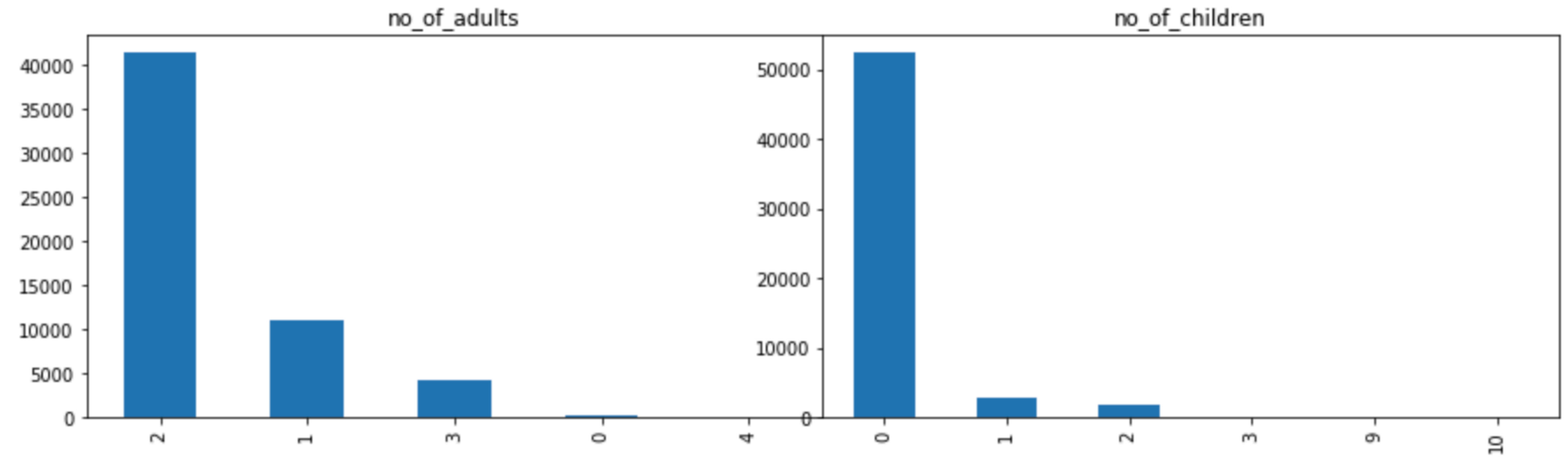
EDA

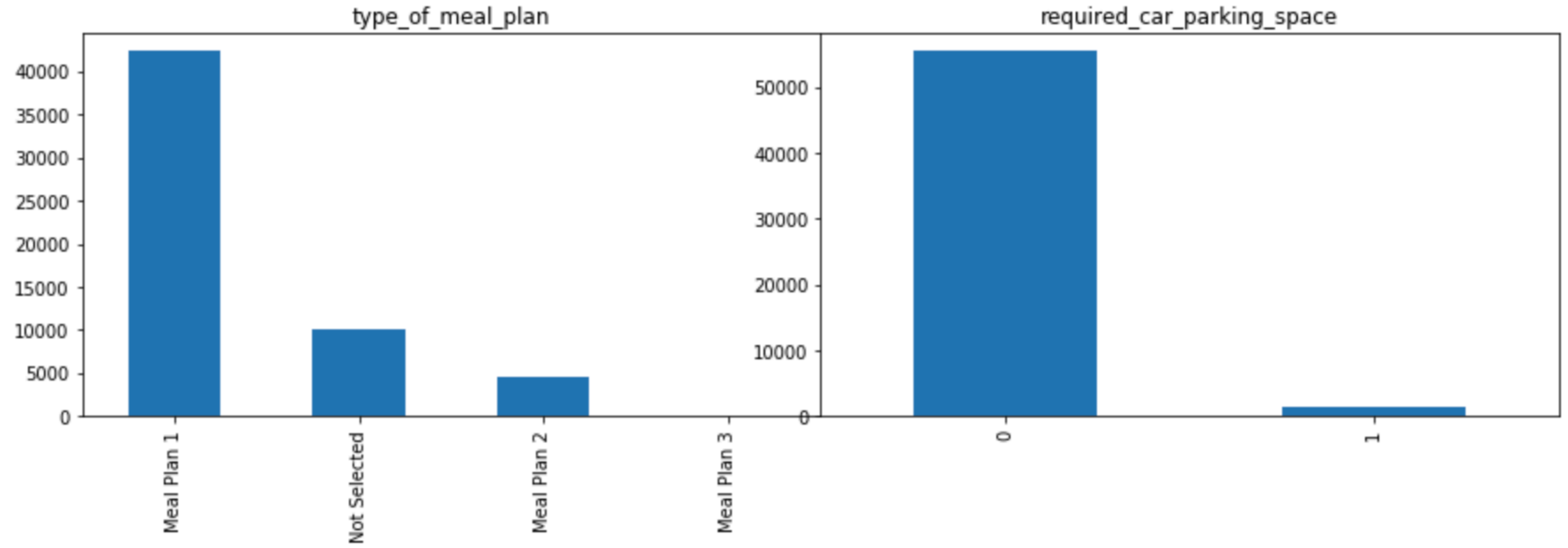
The mean/median values of numerical columns in the dataset are as below

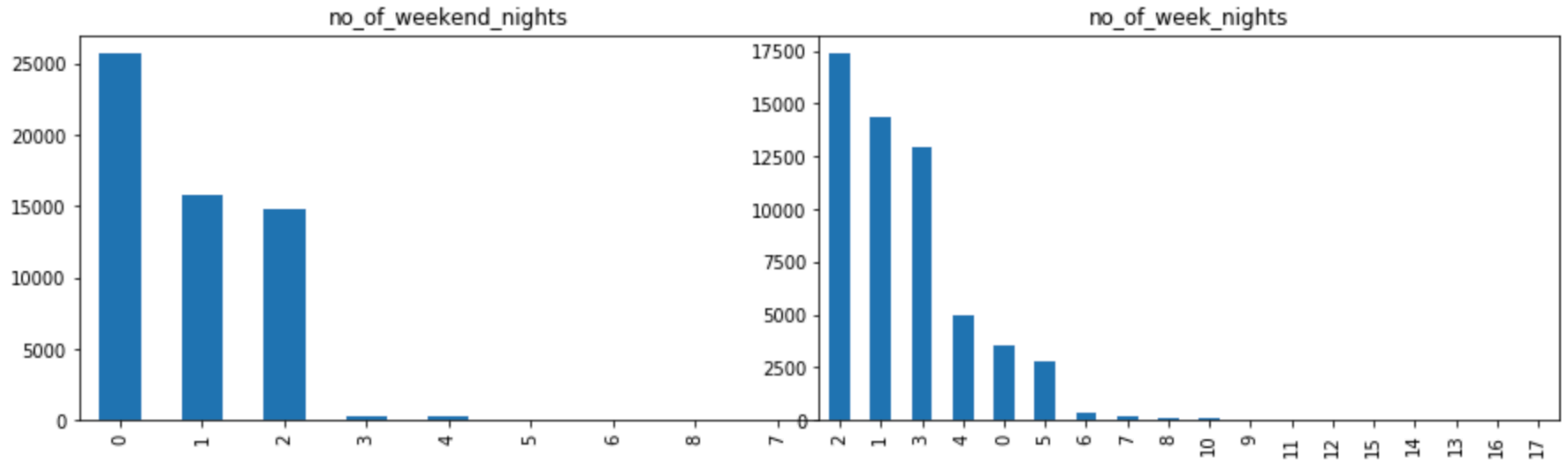
	count	mean	std	min	25%	50%	75%	max
no_of_adults	56926.0	1.875856	0.518667	0.0	2.0	2.0	2.0	4.0
no_of_children	56926.0	0.110723	0.408885	0.0	0.0	0.0	0.0	10.0
no_of_weekend_nights	56926.0	0.835840	0.875900	0.0	0.0	1.0	2.0	8.0
no_of_week_nights	56926.0	2.261901	1.432371	0.0	1.0	2.0	3.0	17.0
required_car_parking_space	56926.0	0.026332	0.160123	0.0	0.0	0.0	0.0	1.0
lead_time	56926.0	93.713909	92.408296	0.0	21.0	65.0	142.0	521.0
arrival_year	56926.0	2018.248340	0.644619	2017.0	2018.0	2018.0	2019.0	2019.0
arrival_month	56926.0	6.490215	3.027185	1.0	4.0	6.0	9.0	12.0
arrival_date	56926.0	15.635913	8.718717	1.0	8.0	16.0	23.0	31.0
repeated_guest	56926.0	0.024664	0.155099	0.0	0.0	0.0	0.0	1.0
no_of_previous_cancellations	56926.0	0.020939	0.326142	0.0	0.0	0.0	0.0	13.0
no_of_previous_bookings_not_canceled	56926.0	0.167902	1.943647	0.0	0.0	0.0	0.0	72.0
avg_price_per_room	56926.0	109.610570	38.256075	0.0	85.0	105.0	129.7	540.0
no_of_special_requests	56926.0	0.666040	0.814257	0.0	0.0	0.0	1.0	5.0

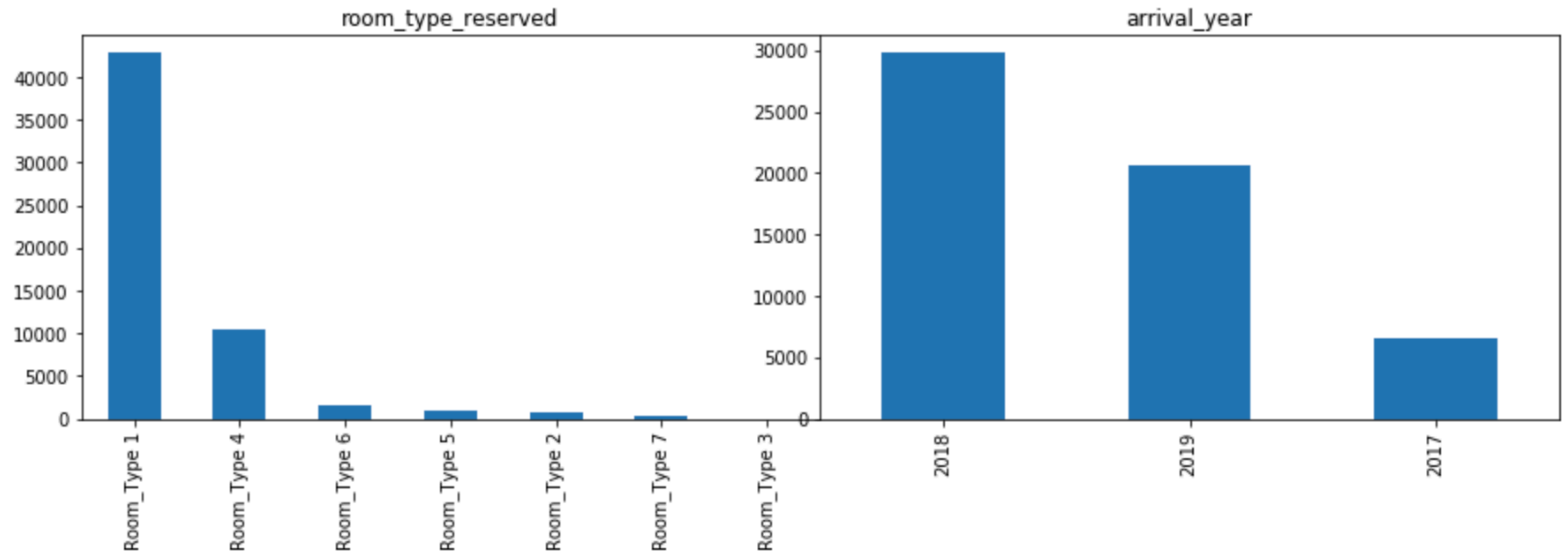
EDA

Value Counts of different fields from the dataset



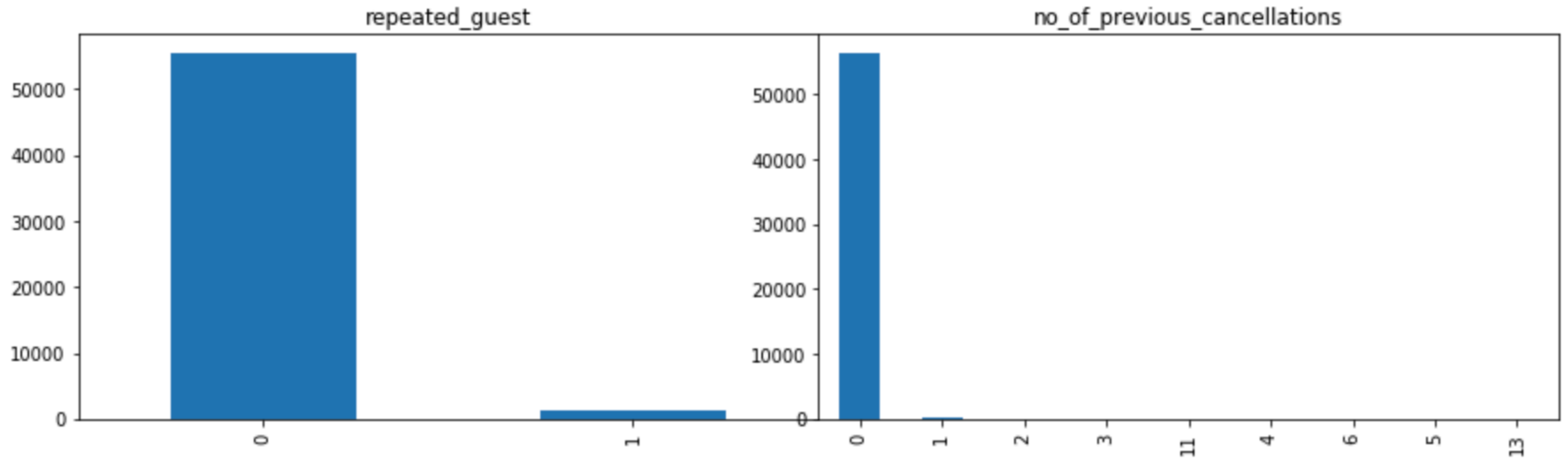


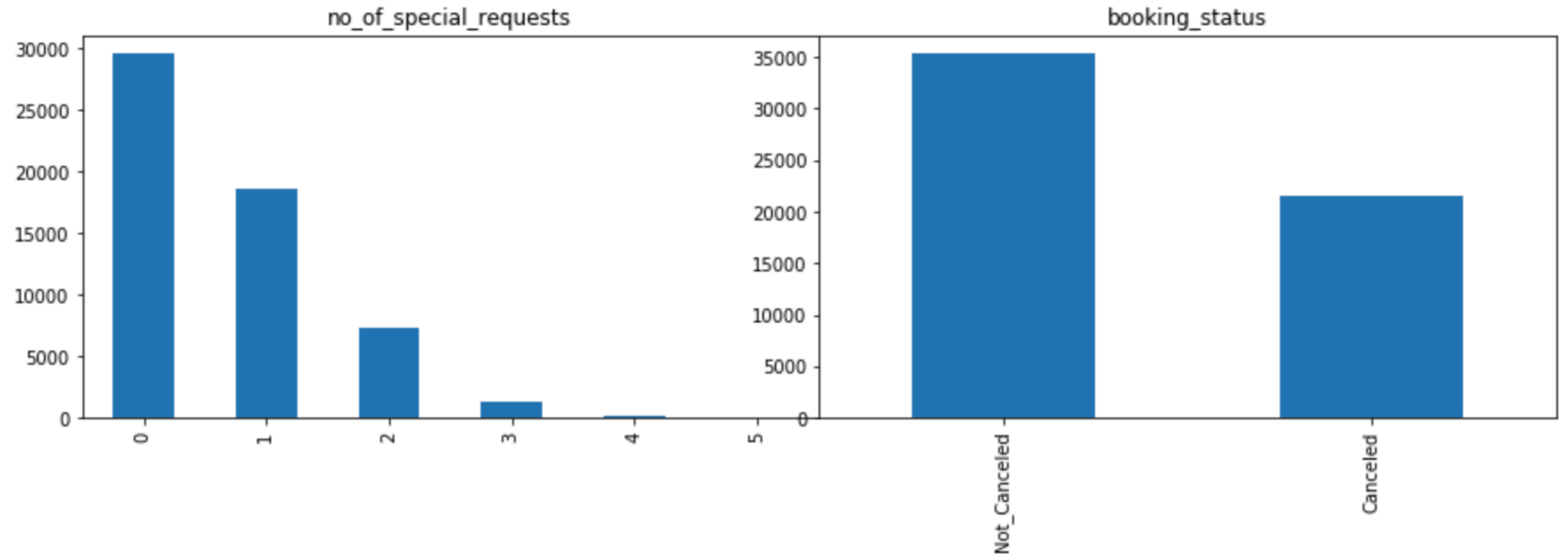




EDA

Value Counts of different fields from the dataset





1. The data spans from the year 2017,2018,2019

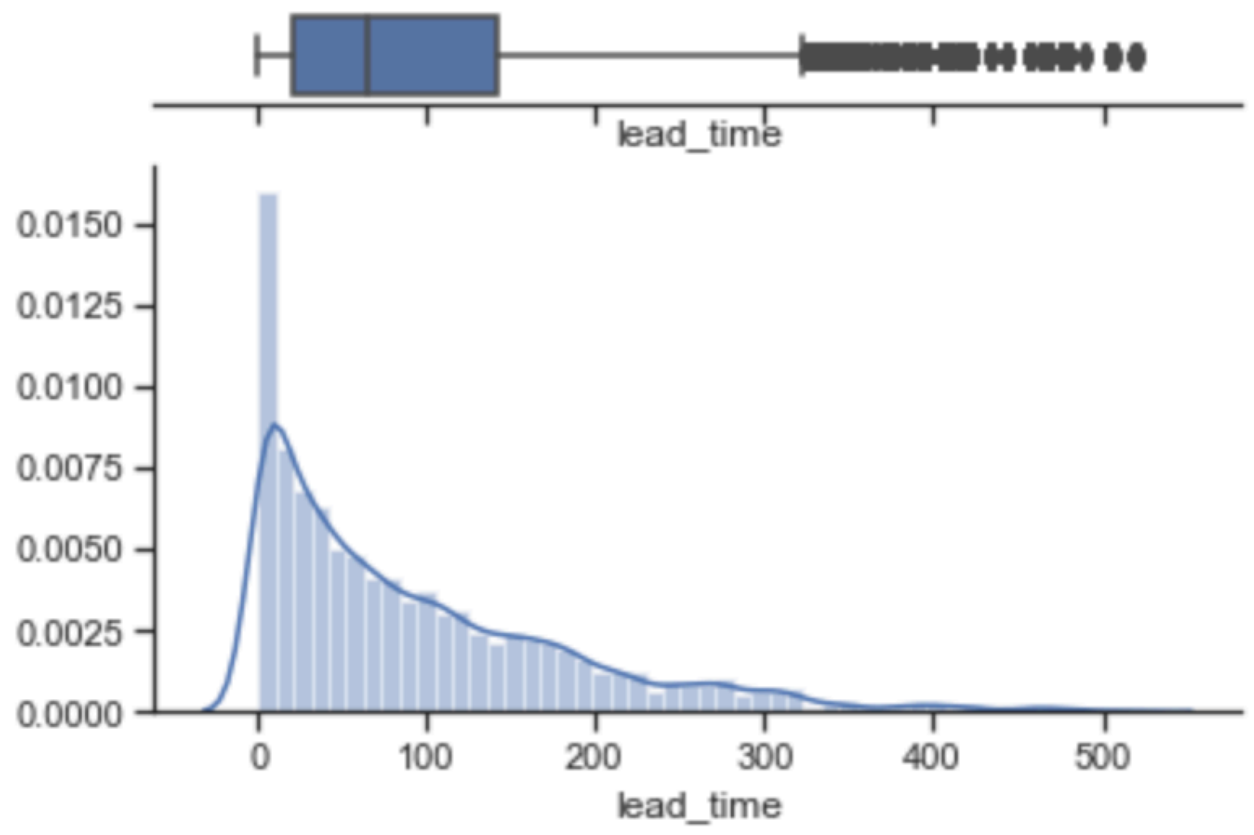
2. hightest no of bookings are in the month of August. But no of spans across other months also, and less bookings are in the mpnth of Jan

3. No of previous cancellations contains only zero value

4.Most of bookings are marketed thru online

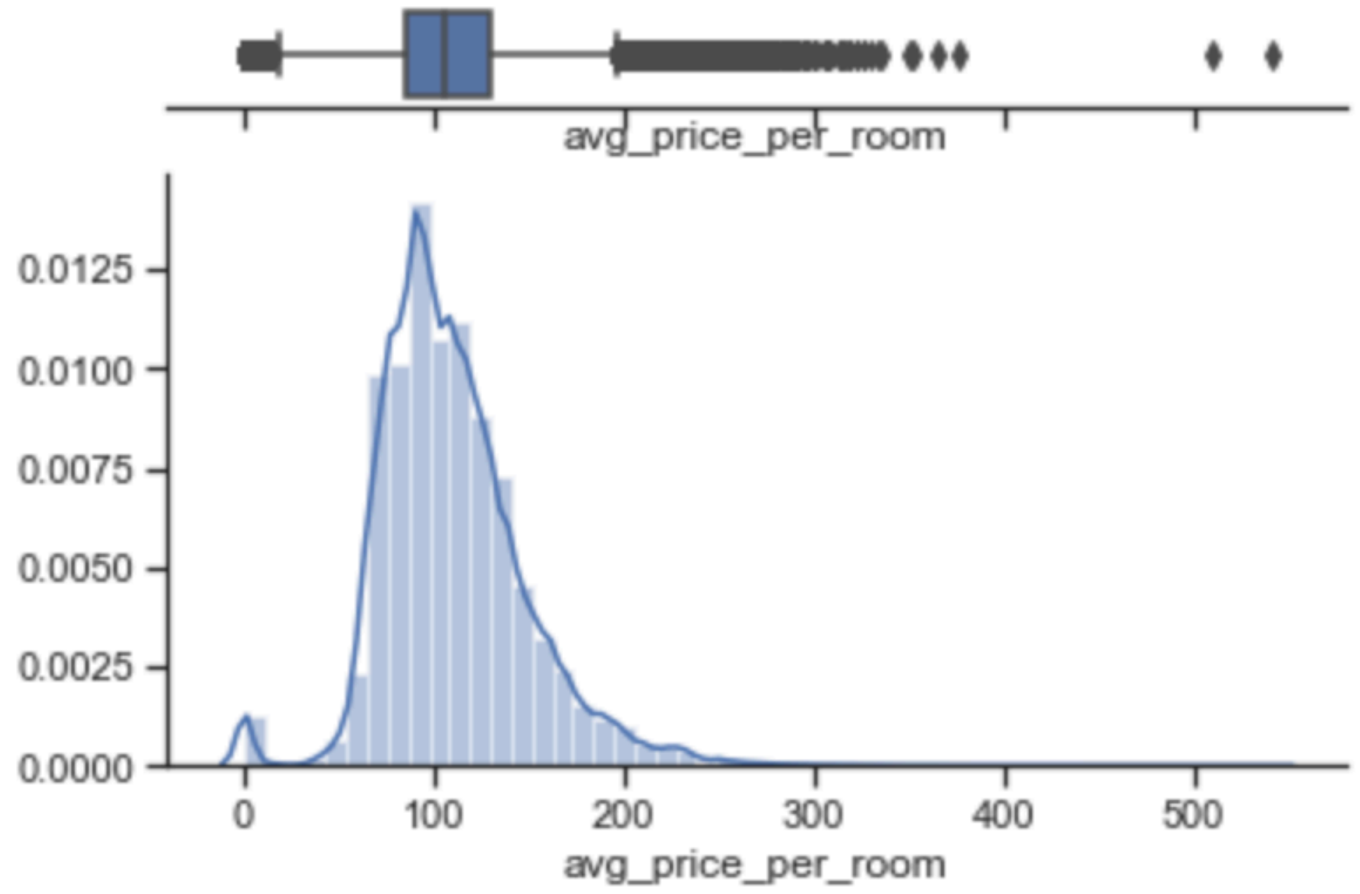
5.Most of bookings are from Room_Type_1

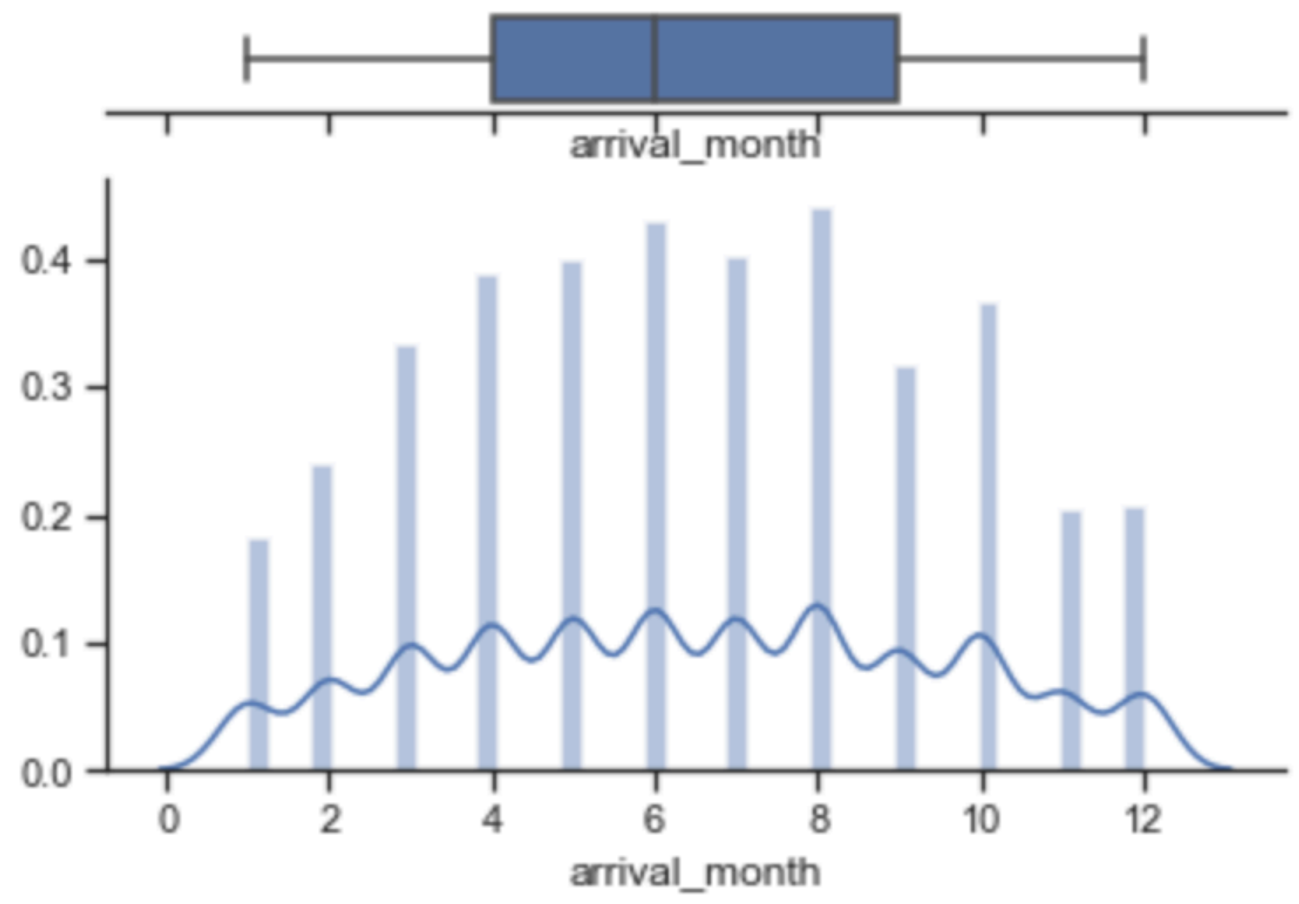
Univariate Analysis



Lead time is right Skewed. So mostly the Booking Lead time is less . Also there are some Outliers

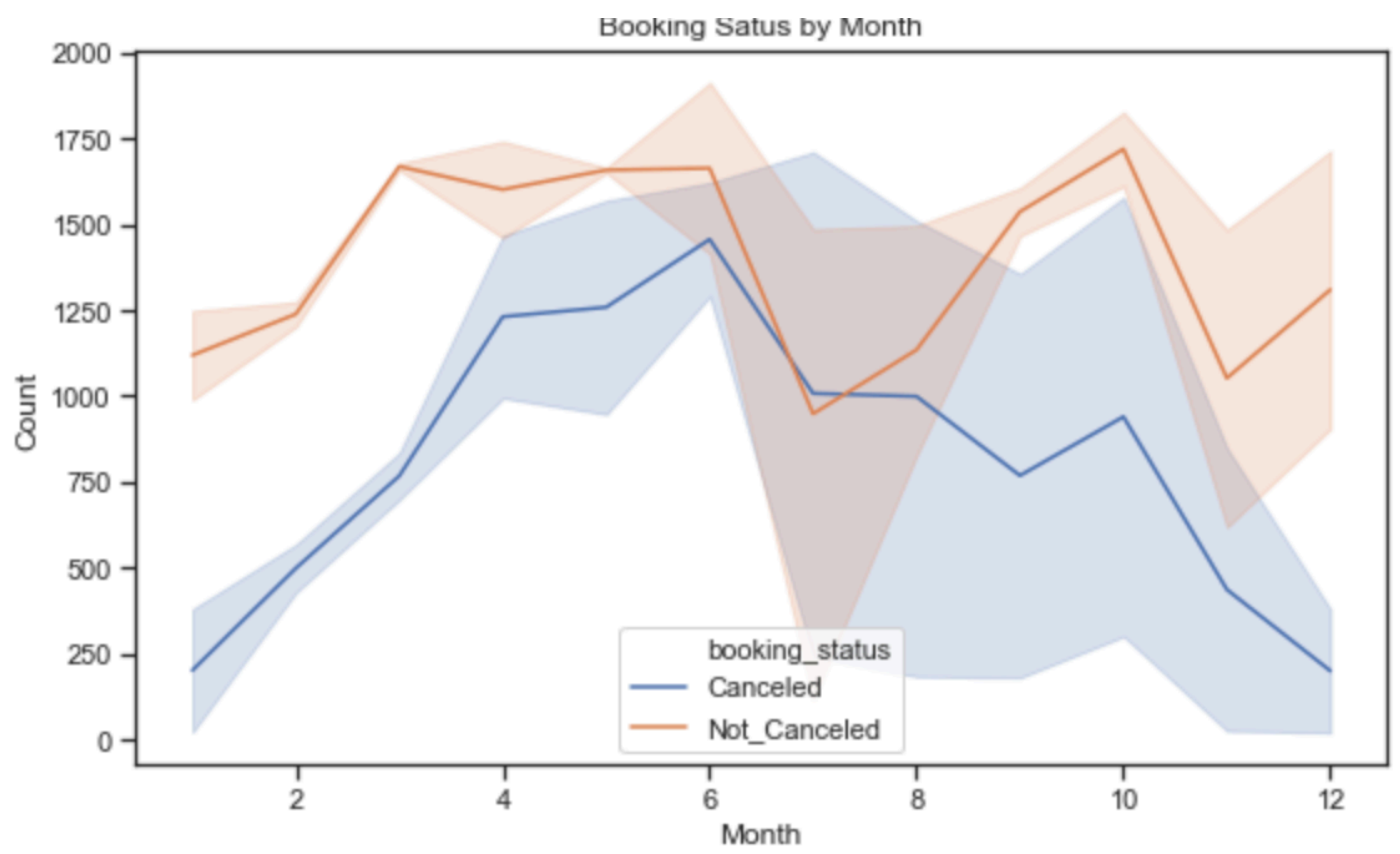
Average Price per room is normal distribution with few outliers



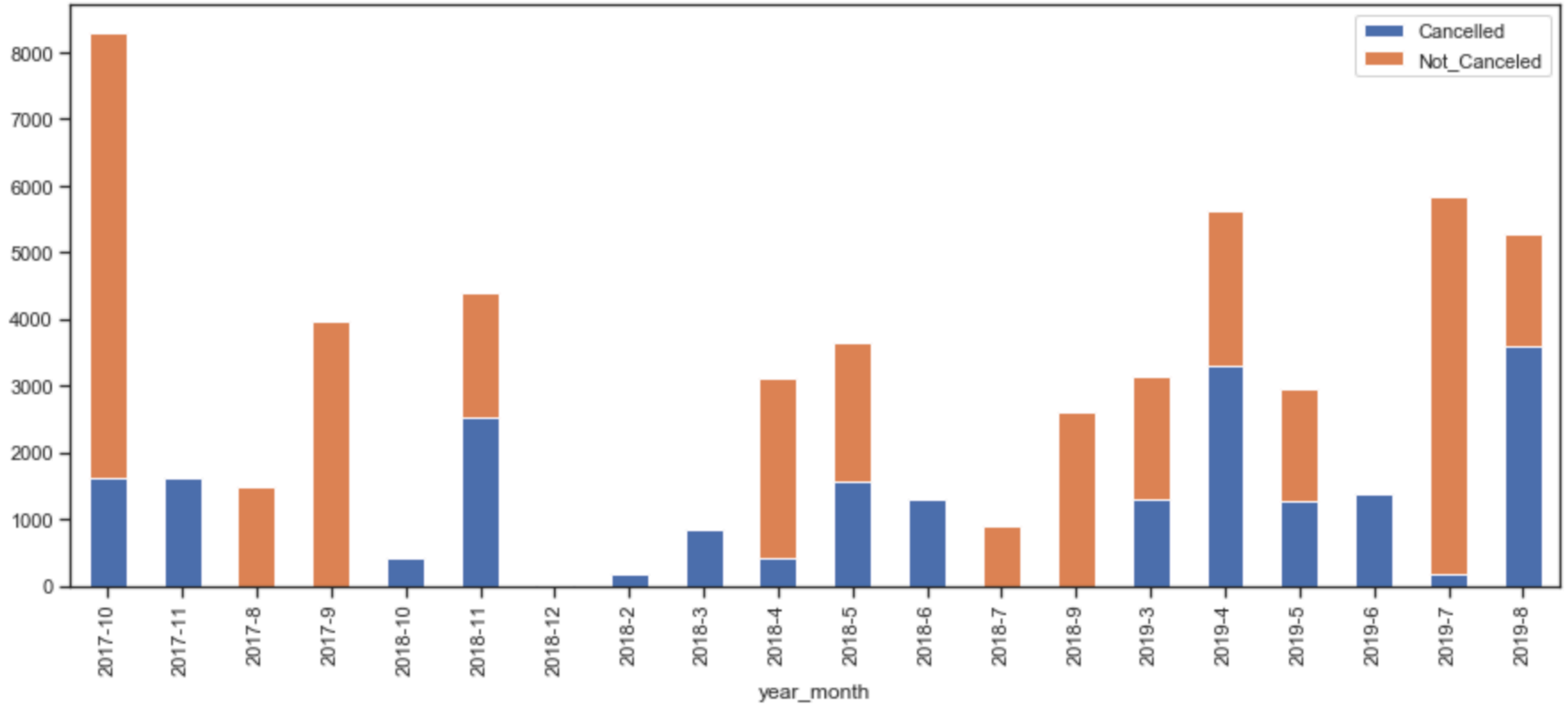


Arrival month spans across the month with leak in June & August

Bivariate Analysis

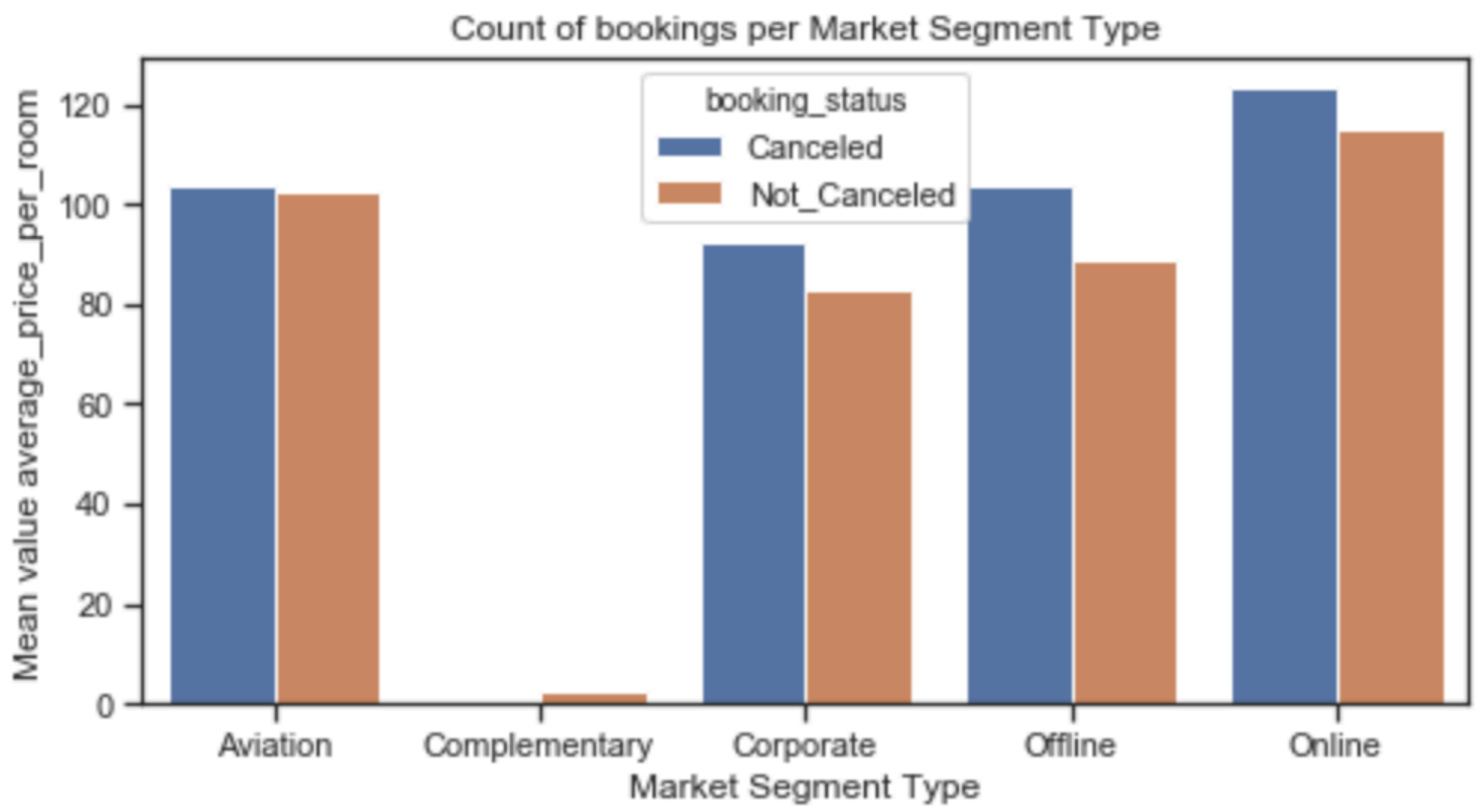


Trend Analysis of Booking Count for the Month



Trend Analysis of Booking Count for the Month , Cancelled bookings are high in 2019-08 , 2019-04,2018-11

Bivariate Analysis



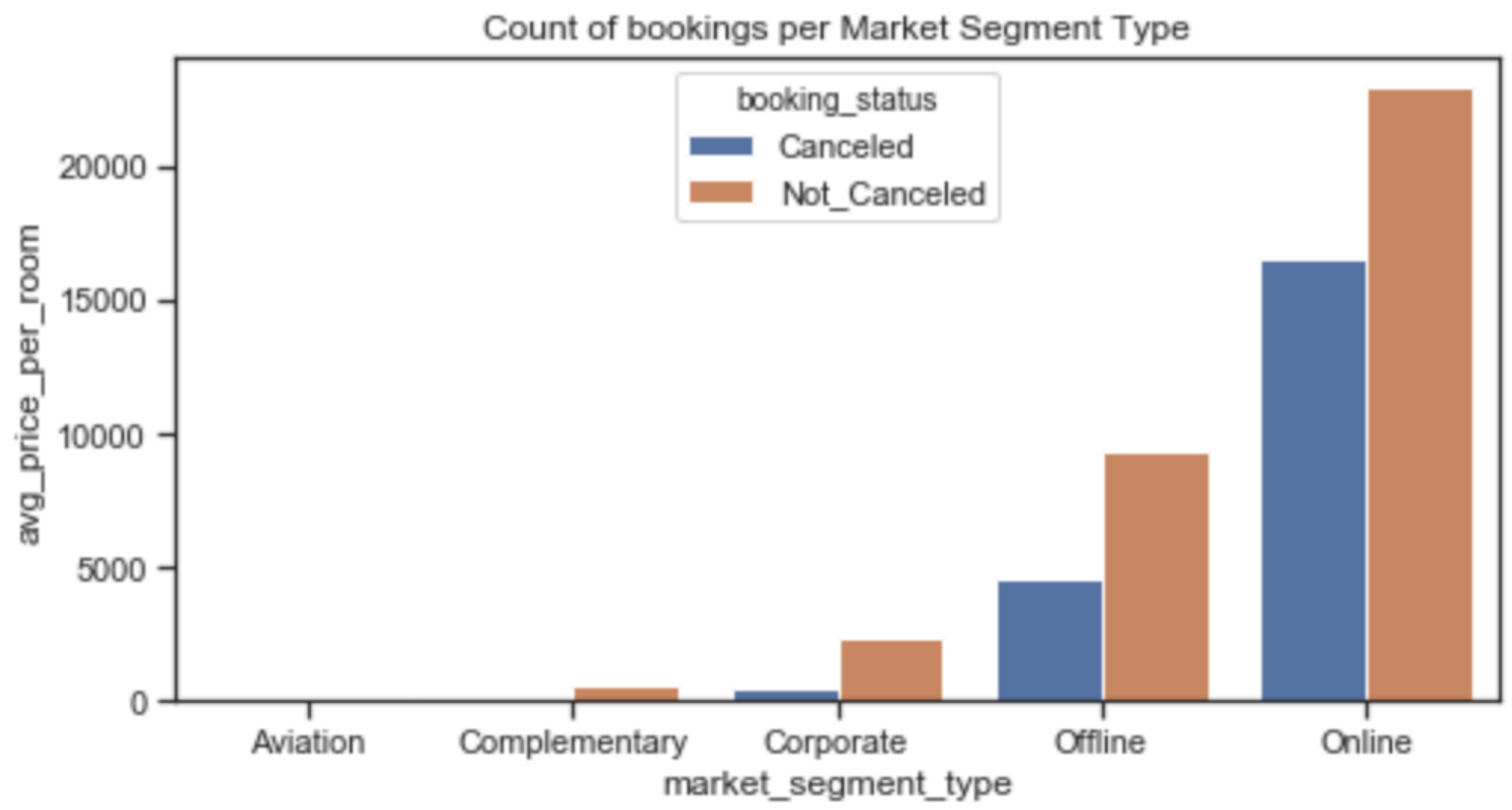
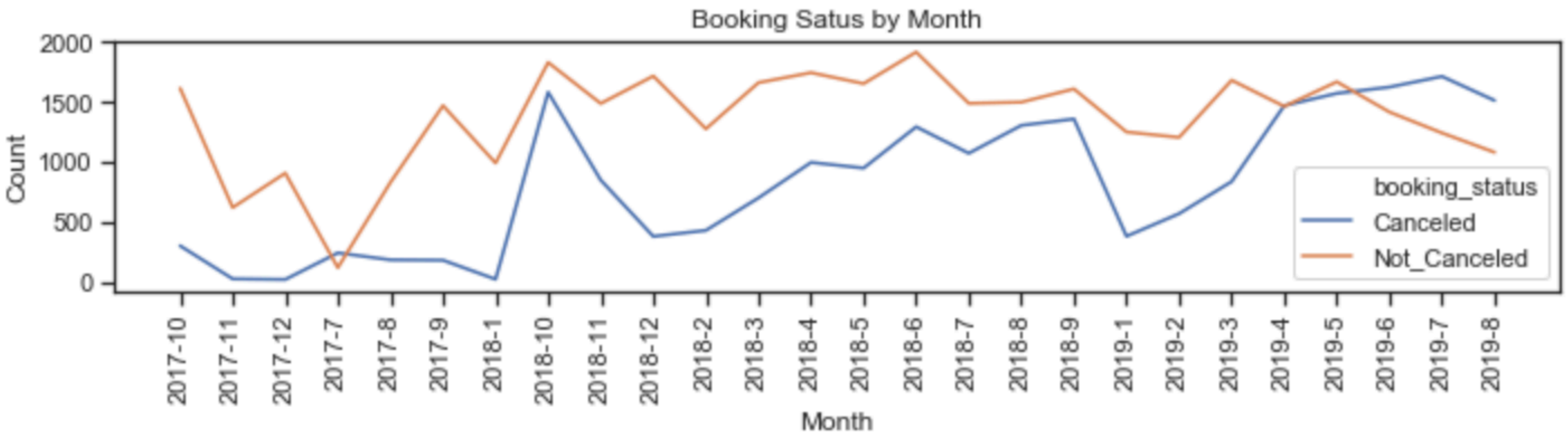


Fig1: Average price of room based on Market Segment Type

Fig2 : Count of Bookings based on Market Segment Type

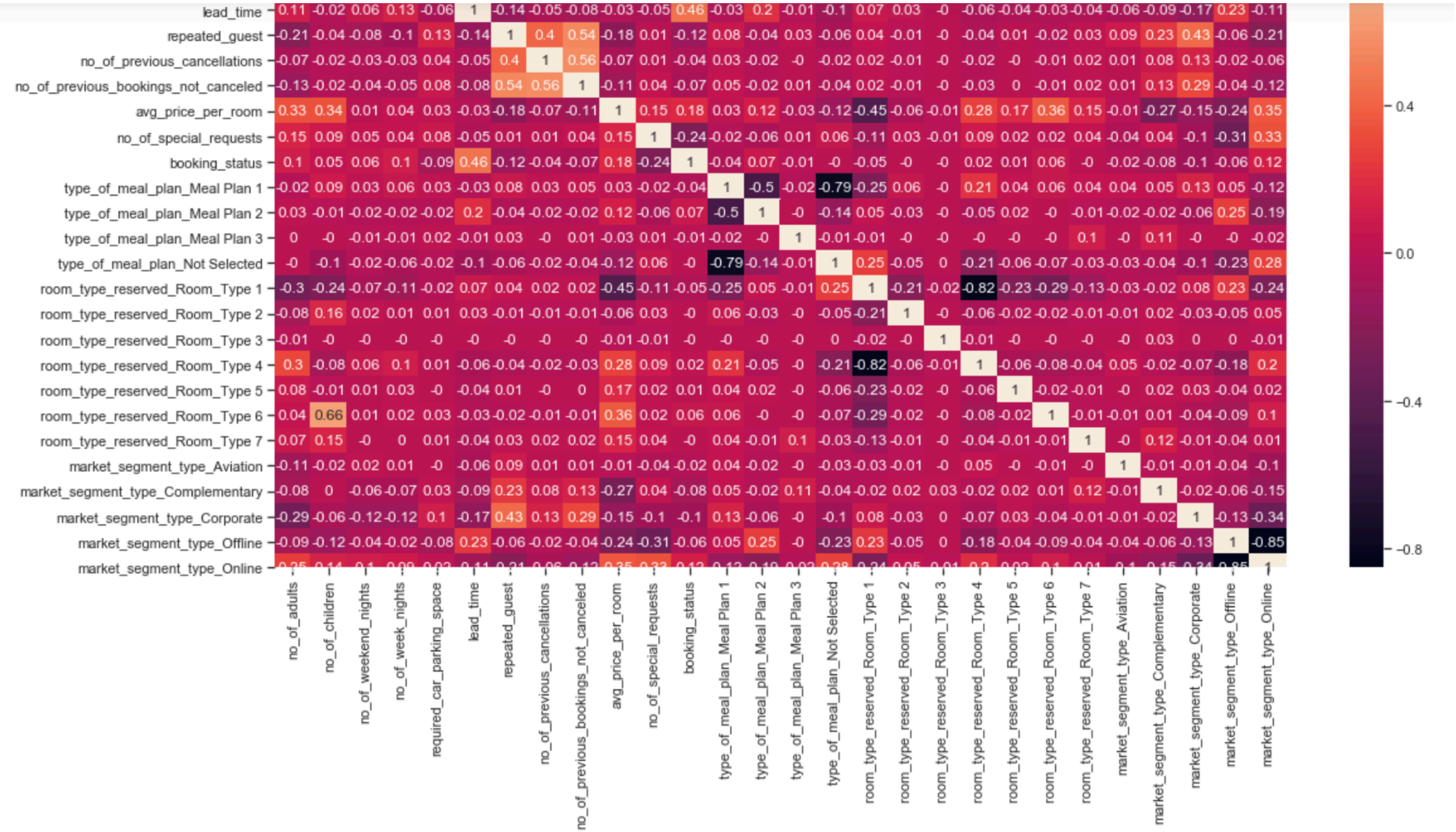
In both the popular one is online.Most Cancelled Bookings are also from Online

Bivariate Analysis

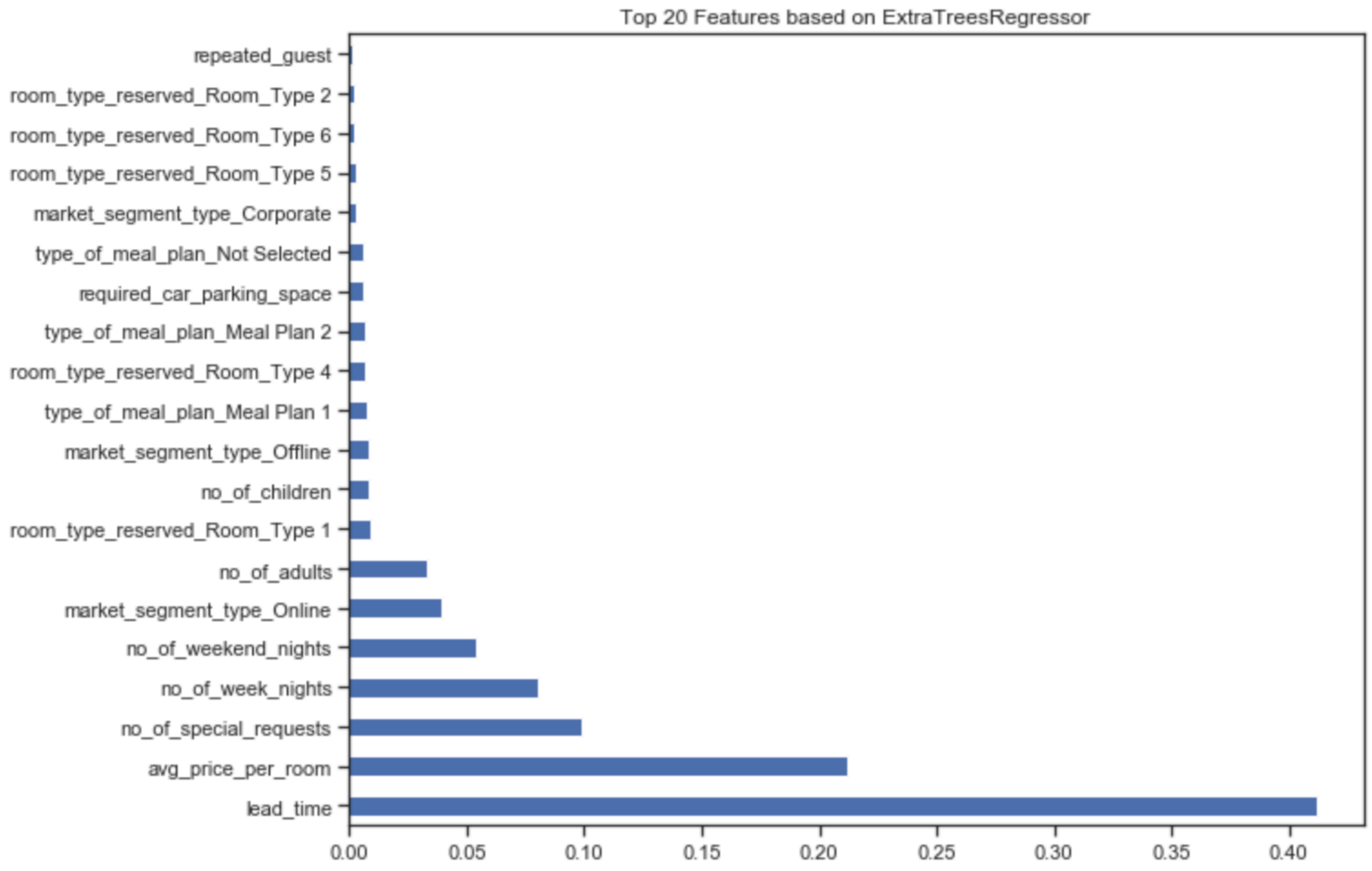


The trend analysis of Number of bookings on month wise

Hest Map

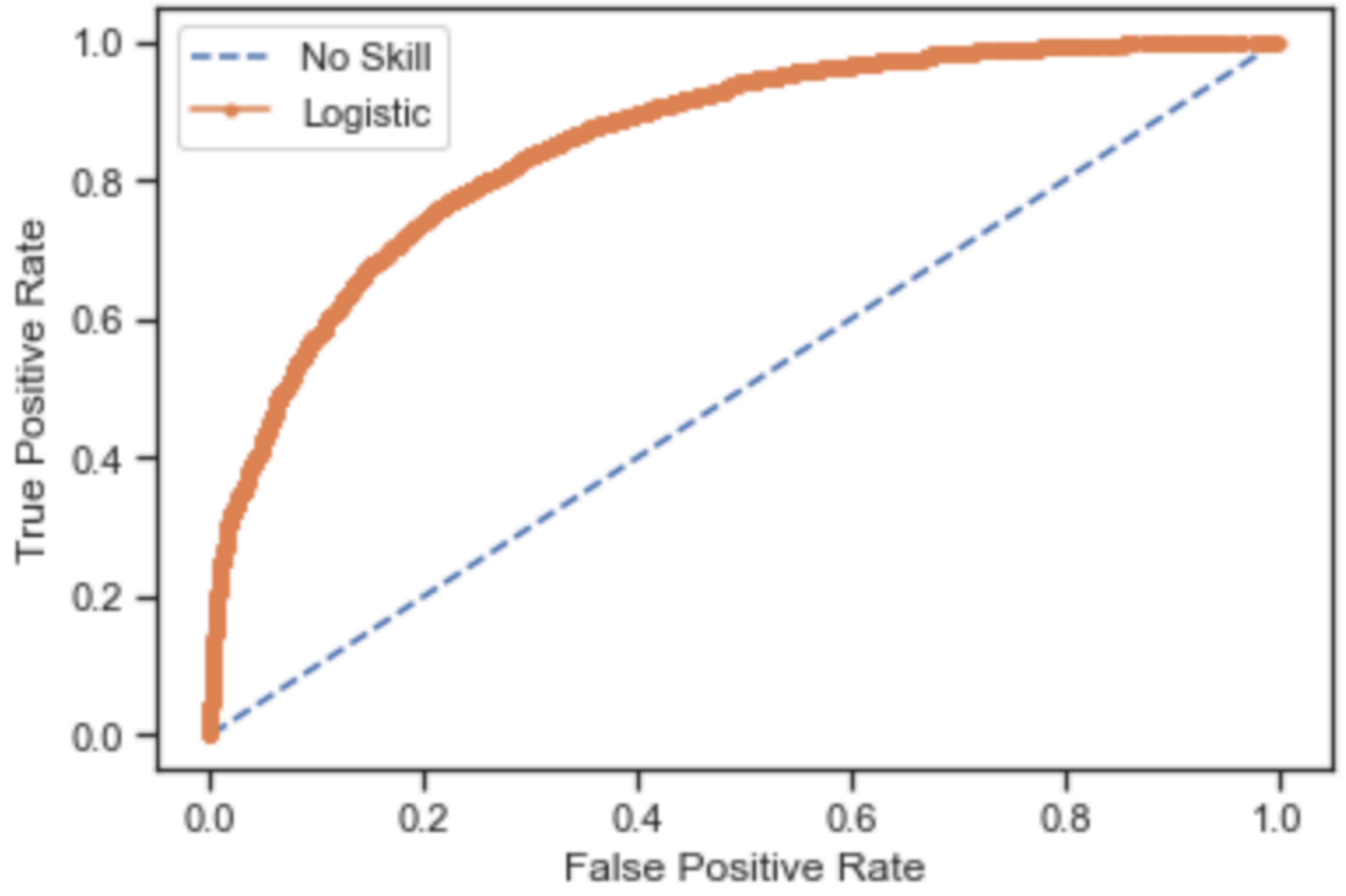

lead_time
repeated_guest
no_of_previous_cancellations
no_of_previous_bookings_not_canceled
avg_price_per_room
no_of_special_requests
booking_status
type_of_meal_plan_Meal Plan 1
type_of_meal_plan_Meal Plan 2
type_of_meal_plan_Meal Plan 3
type_of_meal_plan_Not Selected
room_type_reserved_Room_Type 1
room_type_reserved_Room_Type 2
room_type_reserved_Room_Type 3
room_type_reserved_Room_Type 4
room_type_reserved_Room_Type 5
room_type_reserved_Room_Type 6
room_type_reserved_Room_Type 7
market_segment_type_Aviation
market_segment_type_Complementary
market_segment_type_Corporate
market_segment_type_Offline
market_segment_type_Online
no_of_adults
no_of_children
no_of_weekend_nights
no_of_week_nights
required_car_parking_space
lead_time
repeated_guest
no_of_previous_cancellations
no_of_previous_bookings_not_canceled
avg_price_per_room
no_of_special_requests
booking_status
type_of_meal_plan_Meal Plan 1
type_of_meal_plan_Meal Plan 2
type_of_meal_plan_Meal Plan 3
type_of_meal_plan_Not Selected
room_type_reserved_Room_Type 1
room_type_reserved_Room_Type 2
room_type_reserved_Room_Type 3
room_type_reserved_Room_Type 4
room_type_reserved_Room_Type 5
room_type_reserved_Room_Type 6
room_type_reserved_Room_Type 7
market_segment_type_Aviation
market_segment_type_Complementary
market_segment_type_Corporate
market_segment_type_Offline
market_segment_type_Online
0.4
0.0
−0.4
−0.8

Ensemble Top 20 Features



Model : Logistic Regression

```
No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.857
```



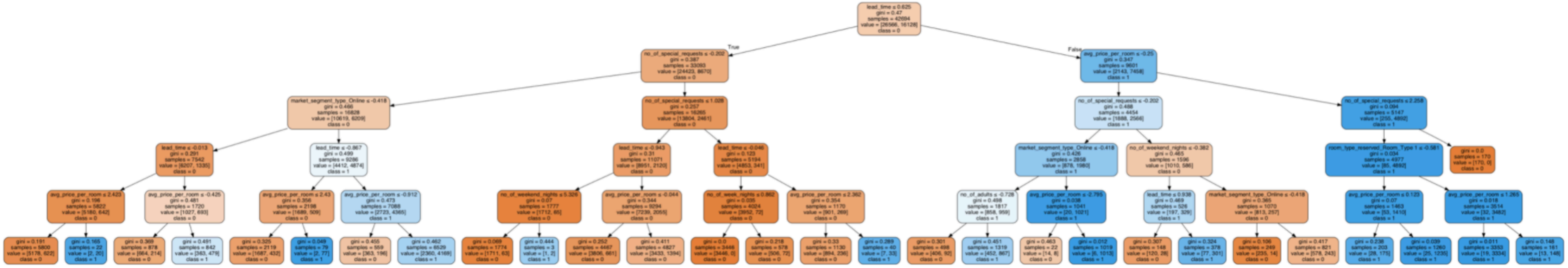

```
Accuracy: 0.8152051714446318
F1 Score of the Logistic Regression Model :  0.6946723336957595
```

Model : Decision Tree

```
GridSearchCV(cv=None, error_score=nan,
             estimator=DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None,
                                              criterion='gini', max_depth=None,
                                              max_features=None,
                                              max_leaf_nodes=None,
                                              min_impurity_decrease=0.0,
                                              min_impurity_split=None,
                                              min_samples_leaf=1,
                                              min_samples_split=2,
                                              min_weight_fraction_leaf=0.0,
                                              presort='deprecated',
                                              random_state=None,
                                              splitter='best'),
             iid='deprecated', n_jobs=None,
             param_grid={'criterion': ['gini', 'entropy'],
                         'max_depth': [2, 3, 4, 5]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
             scoring=None, verbose=0)
```

The GridSearch for Optimized parameers gives criterion as Gini , and no max_depth

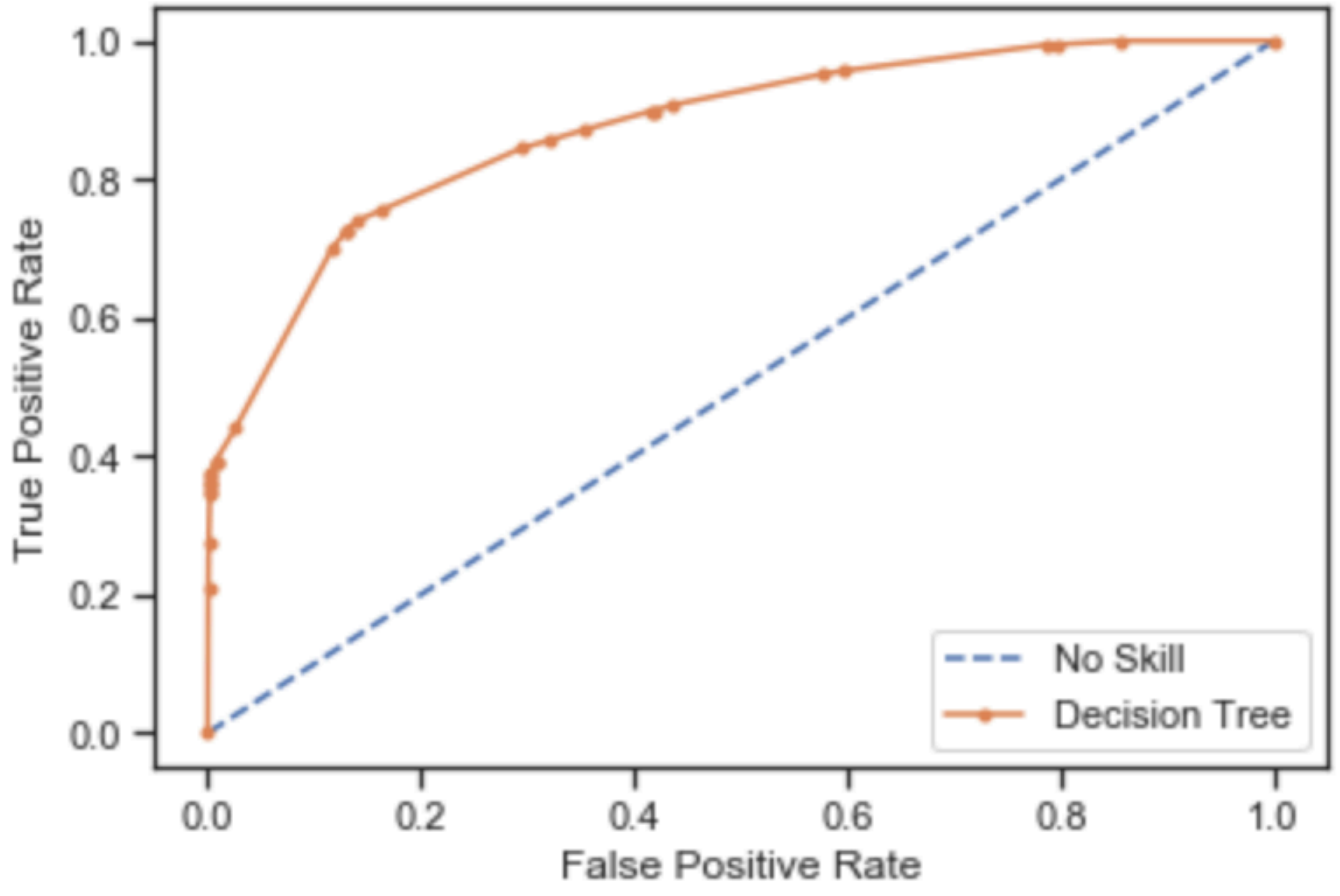
Model : Decision Tree



Model : Decision Tree

The Model with decision tree has the accuracy of 0.815

```
No Skill: ROC AUC=0.500
Decision Tree: ROC AUC=0.874
```




```
Accuracy: 0.8152051714446318
F1 Score of the Decision Regression Model :  0.6946723336957595
```

Conclusion

The Model predicts the cancellation with the acccuracy of .816 , which will be able to predict in cancellation of Bookings based on the createria