Capstone Project – 1

**Hotel cancellation prediction**

Ananth Gopalakrishna Achar

PGP-DSBA 2020 - 2021

**Table of Contents**

**Contents**

**Defining Problem Statement...................................................................................................................................3**

**Need of the study / project......................................................................................................................................3**

**Understanding business / social opportunity........................................................................................................3**

**Data Report ................................................................................................................................................................3**

**Exploratory Data Analysis..........................................................................................................................................3**

**Variable Transformation..........................................................................................................................................62**

**Correlation using Heatmap.....................................................................................................................................62**

**Train Test Split………………………………………………………………………………………………………………………………..…64**

**Clustering………………………………………………………………………………………………………………………………….……...64**

**Calculating Variance Inflation Factor......................................................................................................................67**

**Outlier Treatment.....................................................................................................................................................69**

**Treating Target Imbalance using ADASYN.............................................................................................................69**

**Model Building...........................................................................................................................................................69**

**Random Forest Classifier..........................................................................................................................................69**

**Decision Tree Classifier..............................................................................................................................................74**

**Logistic Regression model ........................................................................................................................................76**

**Random Forest model using Randomized Search CV..................................................................................................................................................................................78**

**AdaBoost Classifier.....................................................................................................................................................80**

**Gradient Boosting Classifier.......................................................................................................................................83**

**ANN model...................................................................................................................................................................85**

**Comparison of performance metrics of all the models...........................................................................................................................................................................87**

**Insights………………………………………………………………………………………………………………………………………………..88**

**Recommendations………………………………………………………………………………………………………………………………89**

**Defining Problem Statement:**

* A hotel company is trying to predict the hotel booking cancellation based on certain variables.
* The intention is to reduce losses due to booking cancellations.

**Need of the study / project:**

* To predict the booking cancellations so that loss to the hotel can be minimized.
* To optimize revenue – the rooms can be made available for any last-minute bookings.
* To reduce the high commission that is paid to the distribution channels for reselling the room/s.
* In some cases, the rooms that are cancelled are sold at a reduced price to increase the chances of being re-booked.

**Understanding business / social opportunity:**

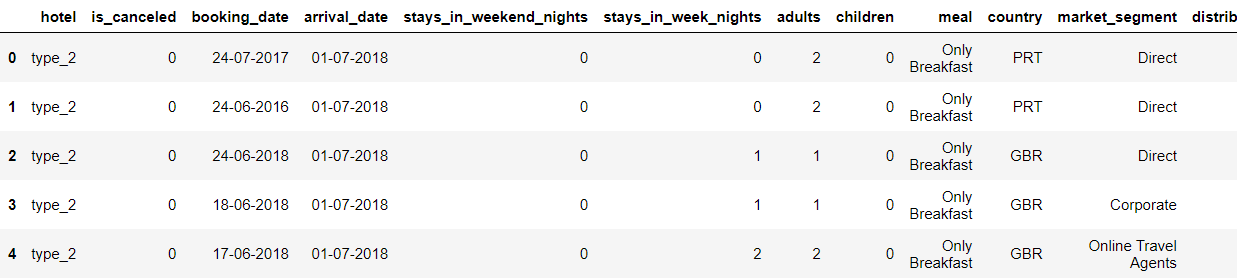
* Increase in profits by having all the rooms booked
* Optimize planning on parking space, laundry wash volumes, etc.
* Reduction in wastage of rooms due to cancellations.
* Study of factors influencing the hotel booking cancellation

**Data Report:**

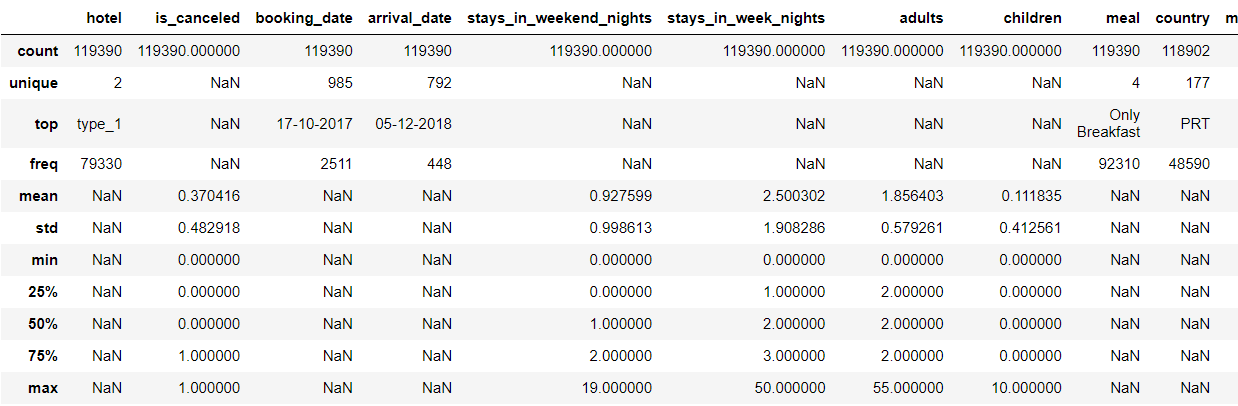
* There are two variables with the names: ‘booking\_date’ and ‘arrival\_date’. The booking\_date is usually before the arrival\_date.
* The dataset contains 119390 observations and 25 features.
* Most of the variables are categorical. days\_in\_waiting\_list, agent are the only two numeric columns. We will also derive a few numeric columns through feature extraction.
* The target variable is: is\_canceled and is a binary variable with 1 indicating that the booking is cancelled and 0 indicating that the booking is not cancelled.

**Exploratory Data Analysis:**

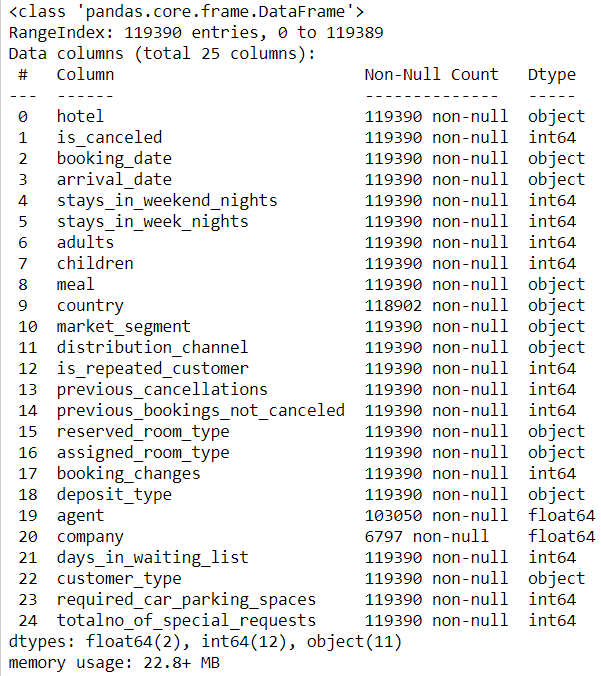
* The head of the data looks like this:



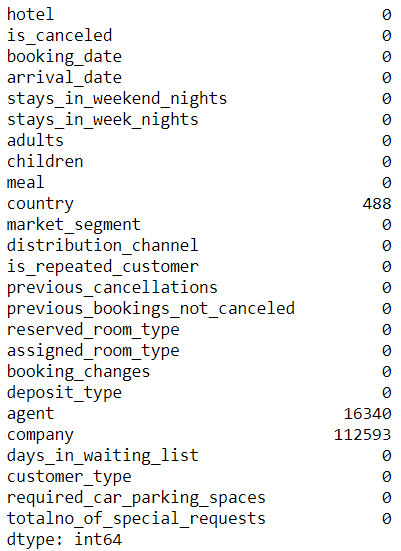
* Many variables are not captured in the screenshot above as the number of variables is large.
* We shall then take a look at the descriptive statistics of our data:



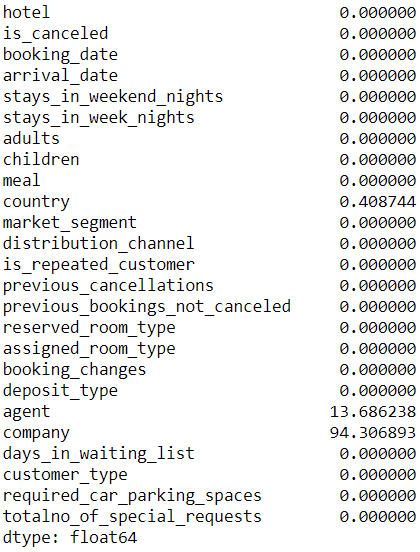
* It is to note that most of the values are NaN as most of our data is categorical in nature.
* We shall now look at the information of each variable:



* The booking\_date and the arrival\_date columns have been recognized as an object type.
* Some columns have missing values.
* We can also see that the size of the data is 22.8 MB of storage space for the client.
* We shall now look at the missing values by each column:



* It can be seen that country has lower missing values and thus can be imputed using the mode.

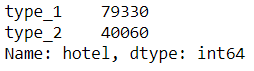


* The above screenshot provides us the percentage of missing information. The agent and company columns have 14% and 94% of missing values and need to be dropped from our analysis. We shall first drop the company column and later shall drop the agent column as well.

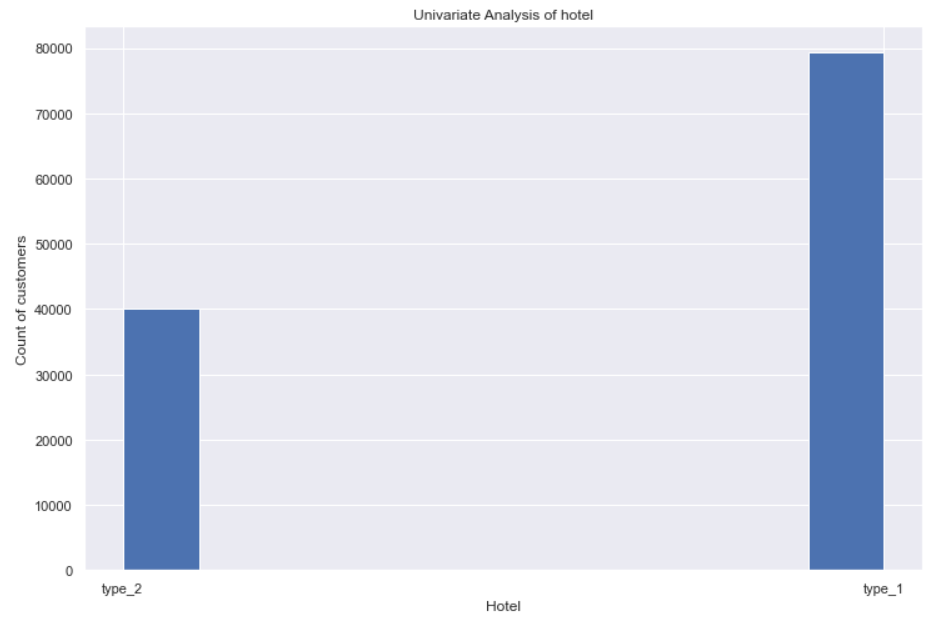
**Let us start by analysing each and every variable and their response to the target variable:**

**Hotel:**

We first check for the value counts of this variable:

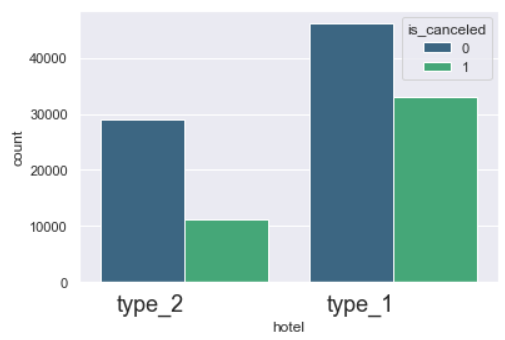


Univariate analysis of Hotel:



The type\_1 of hotels has a count of customers almost twice as that of type\_2.

We then look at hotel with respect to the target variable: is\_canceled

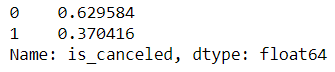


It can be seen from the above plot that the proportion of customers who have cancelled the hotel booking in the type\_2 is almost one third of the count of customers who have not cancelled.

In the type\_1 of hotels, it can be seen that the count of customers who have cancelled their booking is more than 2/3 of the count of customers who have not cancelled.

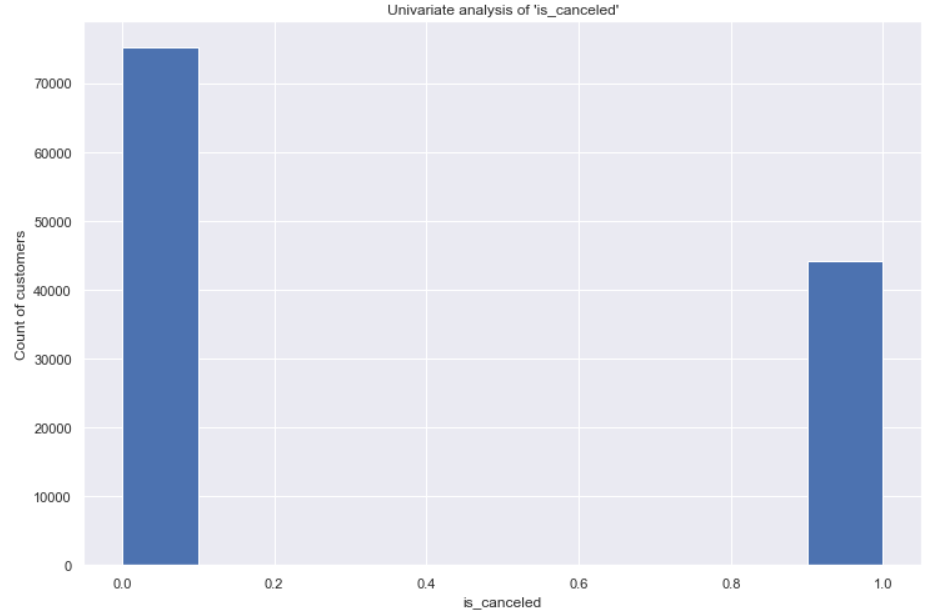
**Target variable: is\_canceled:**

We now look at the value counts of the target variable:



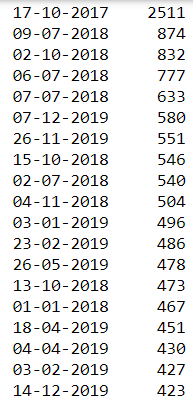
Usually if the proportion in the target variable is 60 - 40 then we conclude that the target variable is balanced. In this case it is close to the above said proportion, however we shall try SMOTE and see if the model performance improves at a later stage.

We shall now look at the univariate visualization of this variable:



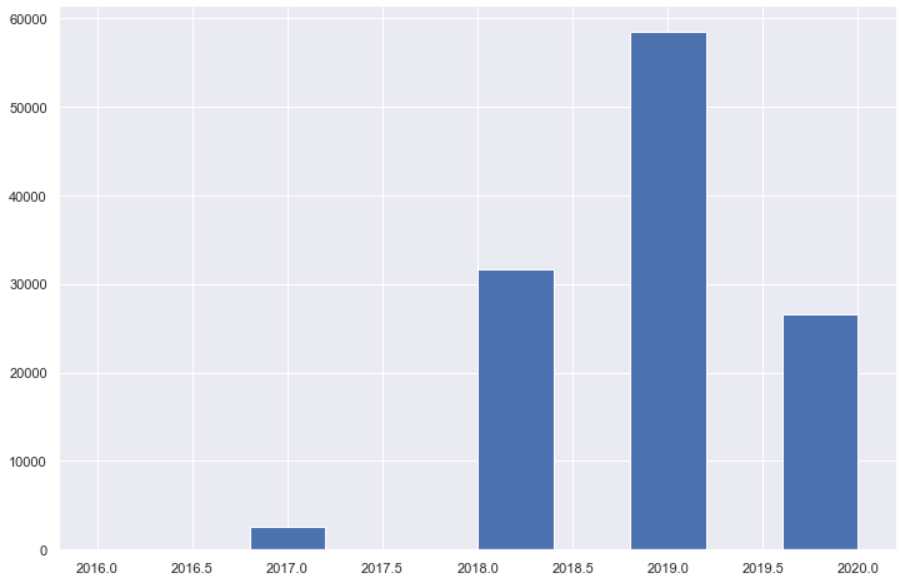
**booking\_date:**

we first look at the value counts of booking\_date since it is shown as an object variable:



From the above, we can make a calculated guess that October and July months are high booking months.

We then extract the booking year from the booking\_date and plot a histogram of the same:



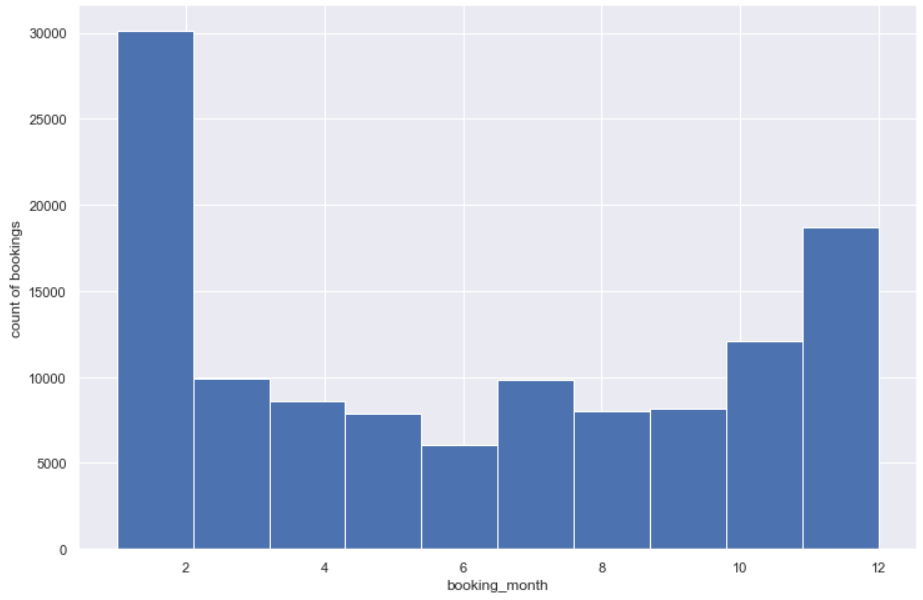
The first half of the year 2019 has seen most number of hotel bookings. Years 2017, 2018 and 2020 have also seen some considerable amount of hotel bookings.

We shall then look at the booking\_year with respect to is\_canceled:



Bookings in the year 2017 have a very high probability of booking cancellation which is close to 0.85

Next is the booking\_year 2018 with a probability of close to 0.4. It can be inferred from the above plot that from the year 2017, as the years progress, the probability of hotel booking cancellation has reduced.

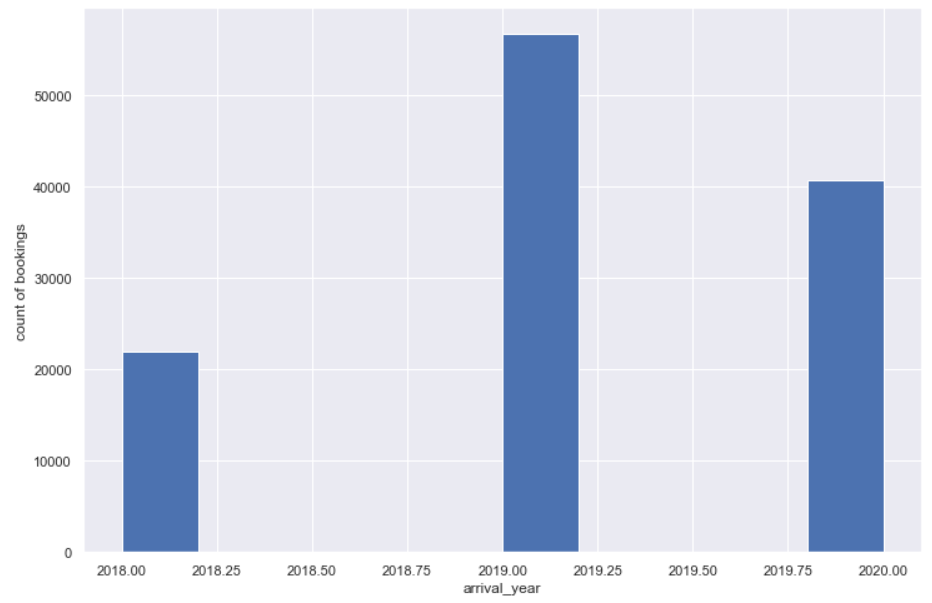


It can be seen from the above plot that the month of January across each year has seen the highest number of hotel bookings followed by December of each year. During the middle of the year, the hotel bookings are the lowest. An exception to this would be the month of July which has count of hotel bookings close to that in the month of February. For the best 3 month period – November, December and January are the months where the hotel has seen high booking volume.



From the above plot, it can be seen that the probability of booking cancellation is highest in the booking\_month of July.

We then extract the arrival\_year and arrival\_month from the arrival\_date column.

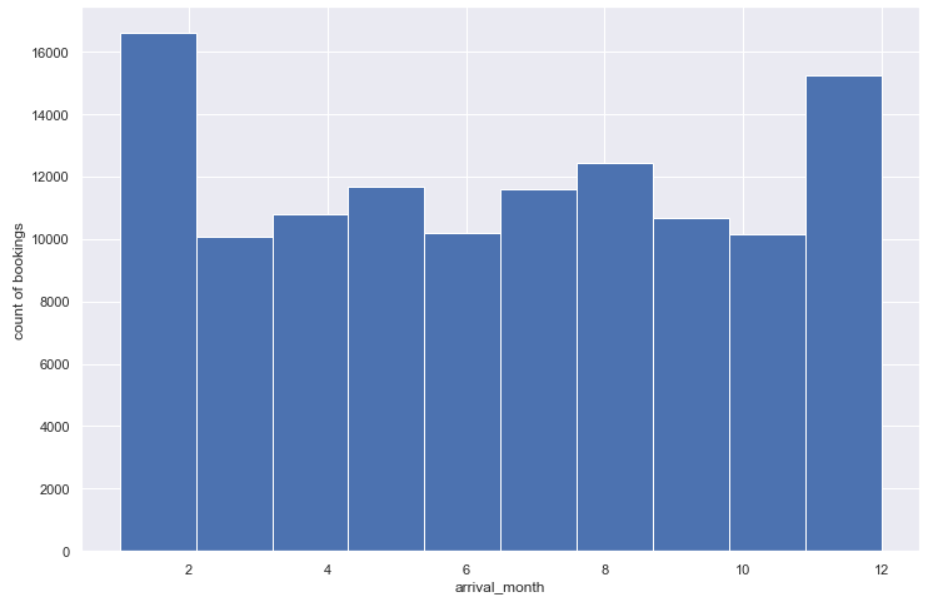


The count of bookings for arrival\_year is the highest in the start of 2019.

The arrival\_year of 2020 has a higher probability of booking cancellation:



We then look at the arrival\_month:



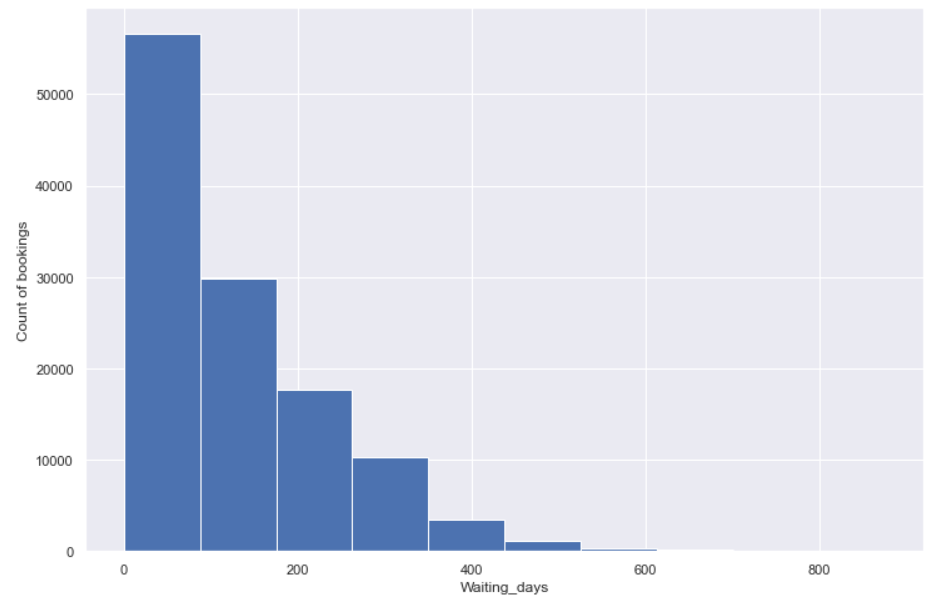
The arrival month is the highest in the month of January and December of each year.

We then look at the arrival\_month with respect to is\_canceled:



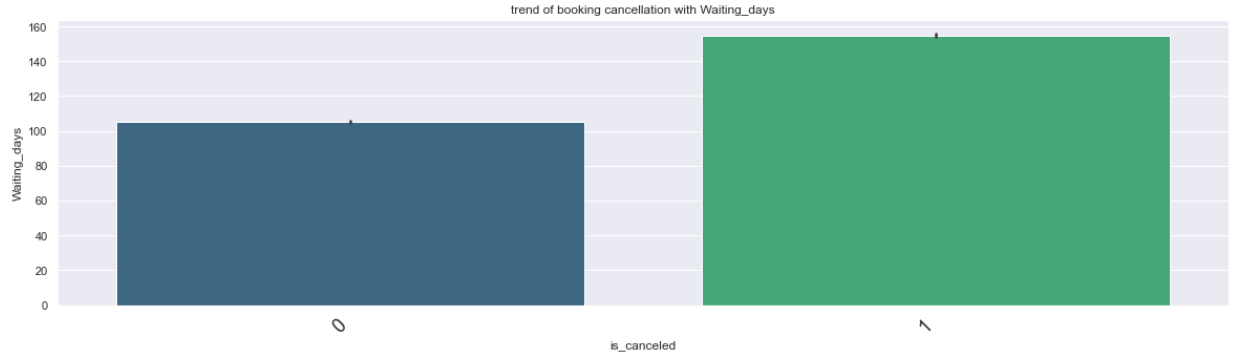
The arrival\_month of April has the highest probability of booking cancellation which is closely followed by the month of May.

We then create a new variable as ‘Waiting\_days’ which is the number of days between the arrival\_date and the booking\_date.



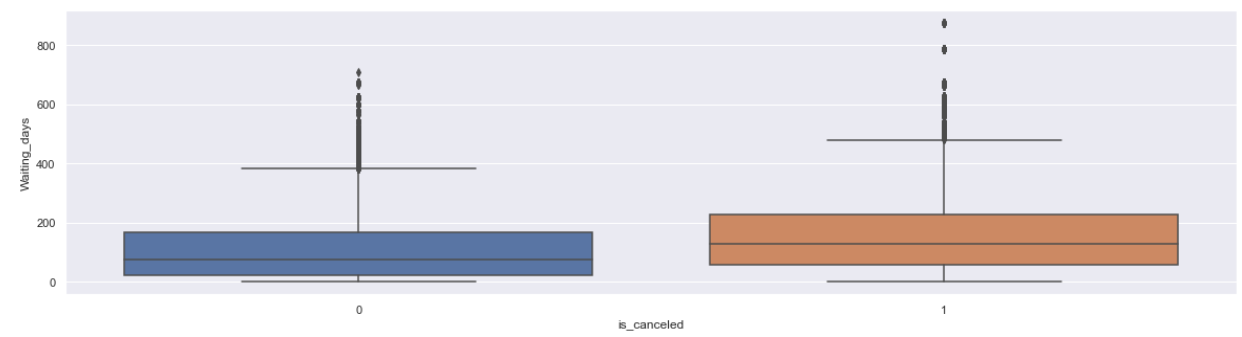
It can be seen from the above plot that the count of bookings for the Waiting\_days between 0 ad 100 is the highest.

We then look at the behaviour of Waiting\_days with the is\_canceled:



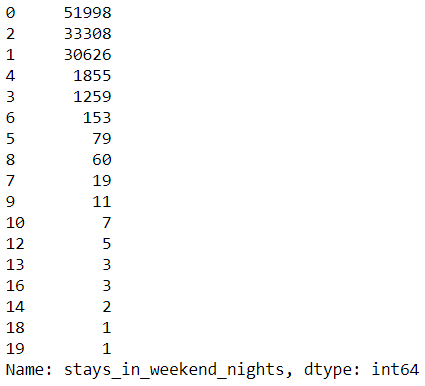
It can be observed from the above plot that when the Waiting\_days is more than 110 days, the chances of the booking being cancelled is high.

We now plot a boxplot to view the same:

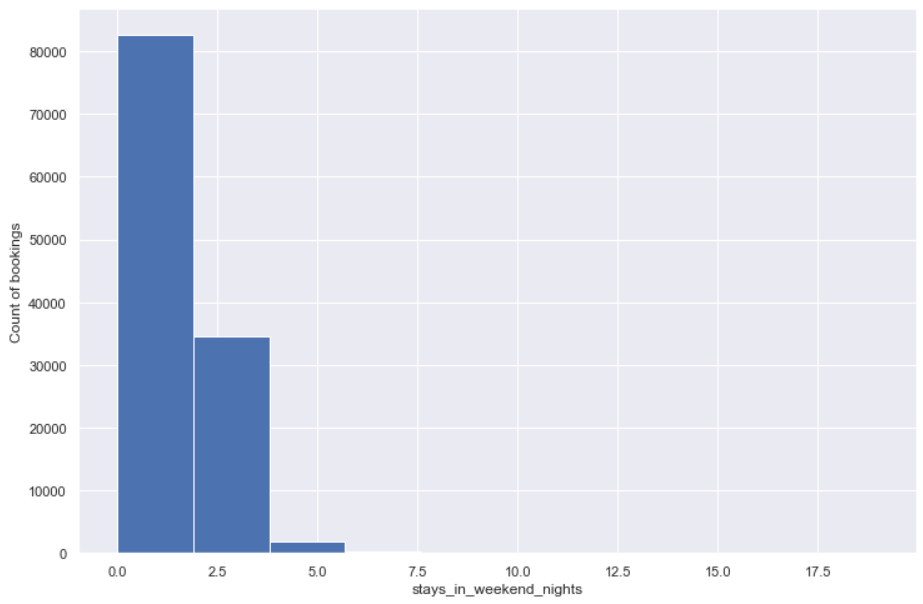


**stays\_in\_weekend\_nights:**

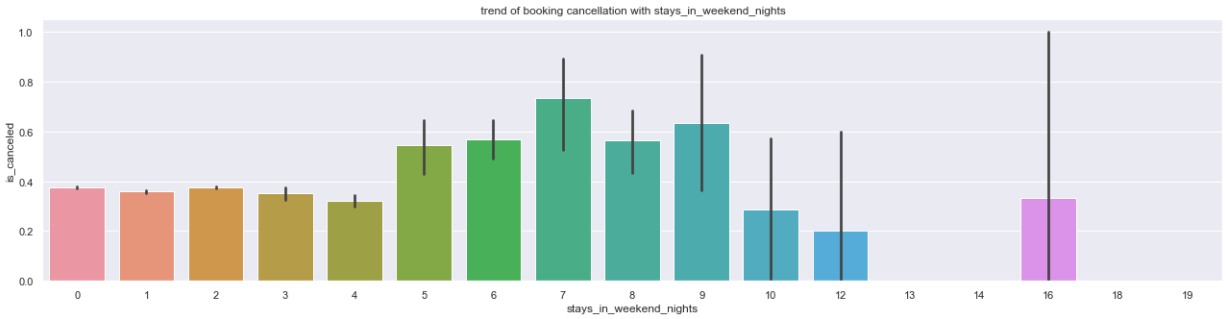
We at first look at the value counts of this variable:



The counts for bookings that have 0 stays\_in\_weekend\_nights is the highest.



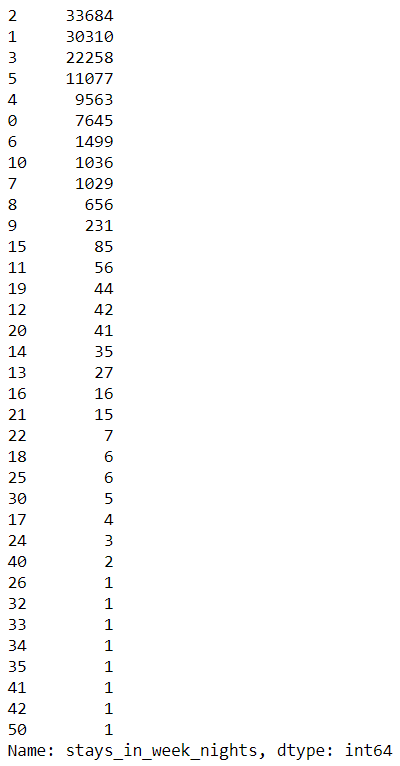
Comparison of stays\_in\_weekend\_nights with is\_canceled:



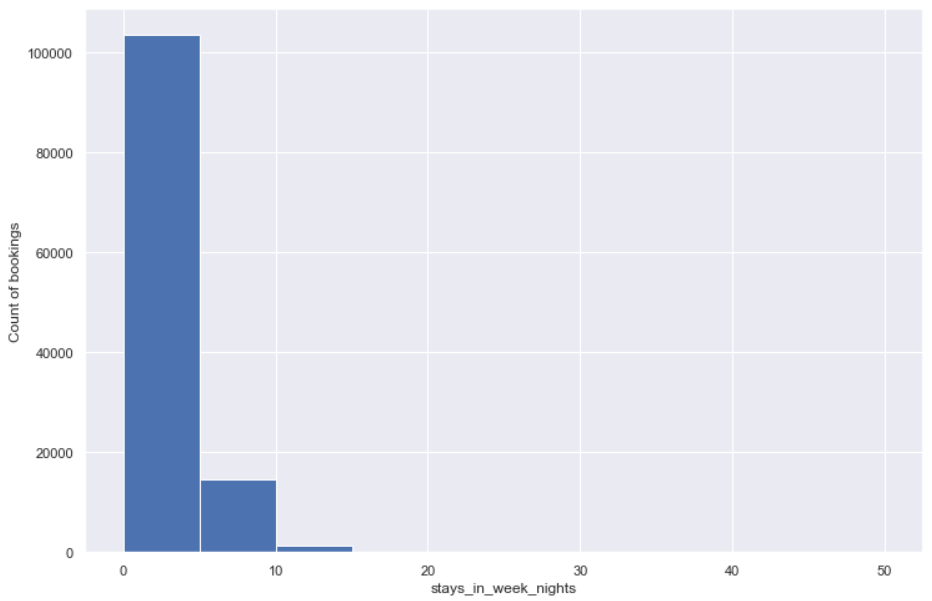
It can be seen that the probability of booking cancellation is highest for the stays\_in\_weekend\_nights of 7. The stays\_in\_weekend\_nights of 16 has the largest variability which is indicated by the long confidence interval. This is probably because of the less count of bookings for this value.

**stays\_in\_week\_nights**

We look at the value counts of this variable:

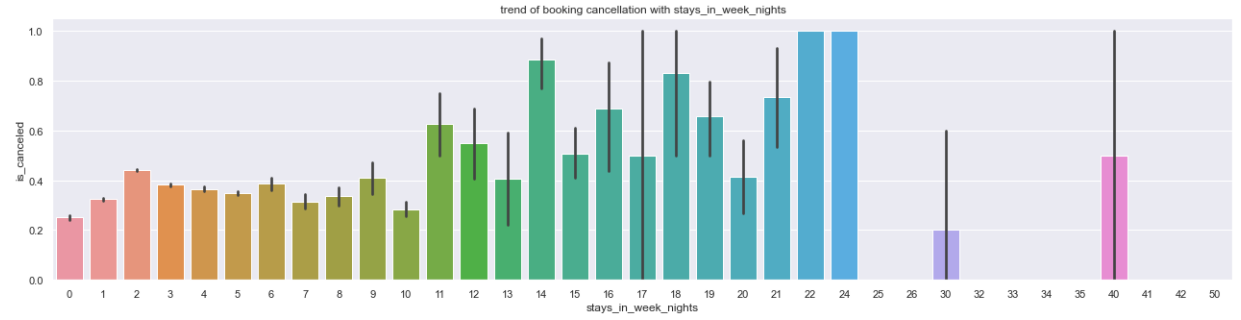


We look at the histogram of this variable:



It is seen that the bulk of the booking distribution lies within the stays\_in\_week\_nights of 0 to 5.

We shall also look at the stays\_in\_week\_nights and is\_canceled:

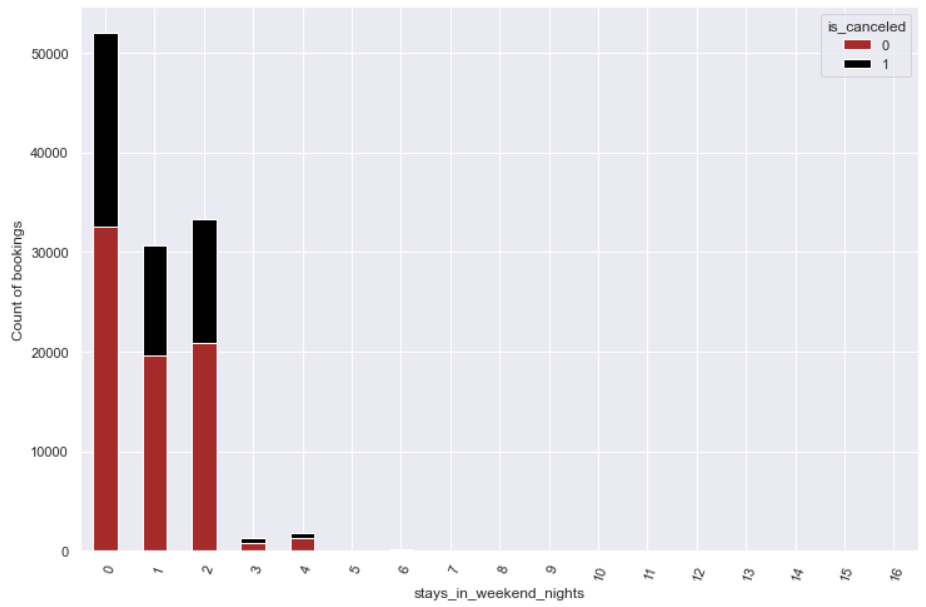


From the above plot it can be inferred that the probability is very high i.e. almost 1 when the stays\_in\_week\_nights is 22 or 24. That is not to be worried about as the count of bookings is low for these values.

As discussed earlier, the bulk of the booking is within the stays\_in\_week\_nights of 5. The stays\_in\_week\_nights of 2 has the highest probability of about 0.5 among the other values in this section.

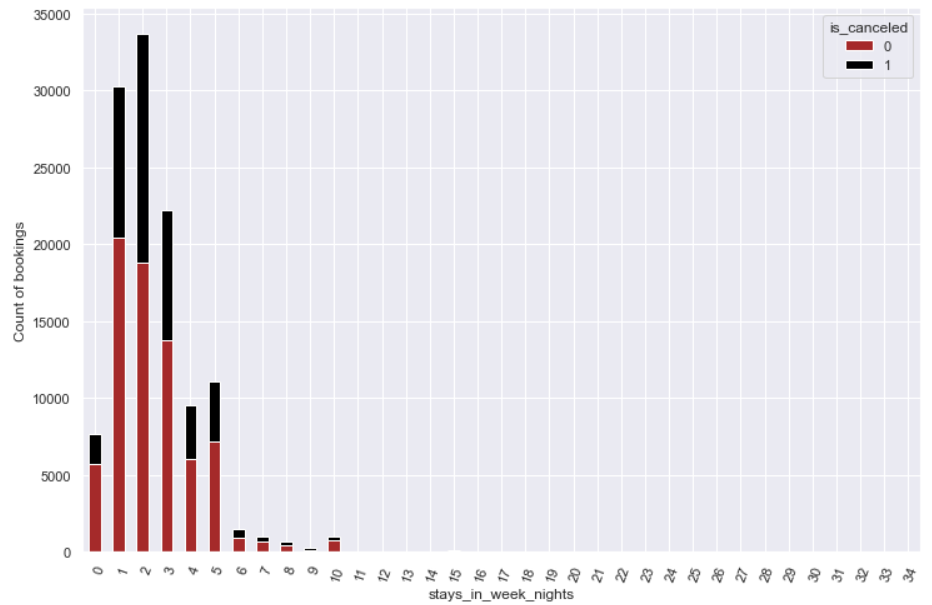
This means that the customer who have booked for stays\_in\_week\_nights of 2 have close to 50% chances of booking cancellation. This is very high and needs to be addressed.

The following plot shows us the stays\_in\_weekend\_nights with respect to is\_canceled:



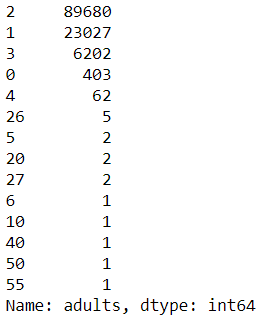
Our previous observation was augmented by the illustration of the above plot. The probability of cancellation is almost the same for bookings with stays\_in\_weekend\_nights of 0 and 2.

We shall now look at a similar plot for stays\_in\_week\_nights:

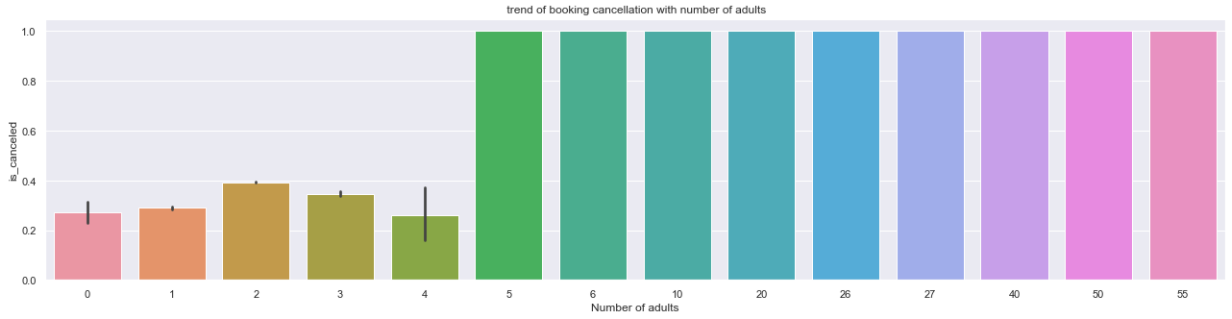


**adults:**

We now look at the value counts of adults:



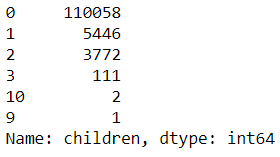
There are 403 bookings with a value of 0 for adults. This does not make much sense as then we would have to wonder who the room is booked for. The sensible guess in this case would be that the name of this column should have been: ‘number of accompanying adults’. We shall look at the distribution of this variable with respect to the target variable:



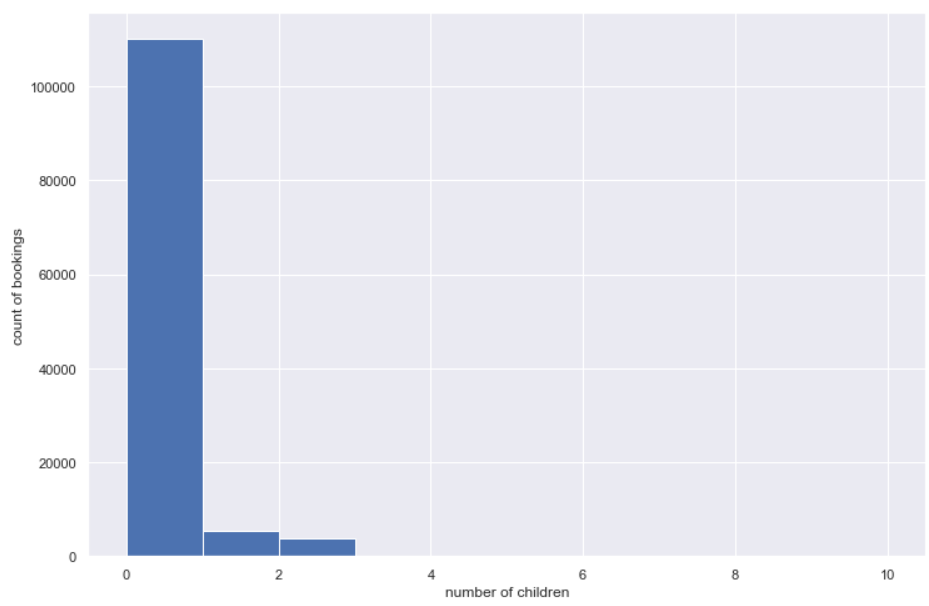
The bulk of the bookings lie within the value of to 4 for the adults. The probability of booking cancellation is highest for adults of 2 as compared to others in this section of the category. The variability of booking cancellation is highest for adults of 4.

**children:**

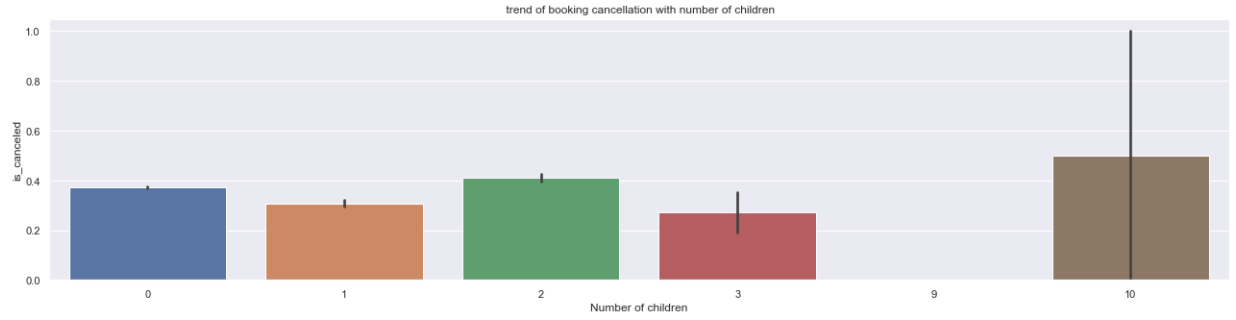
The value counts for this variable is as follows:



Looking at the histogram of this variable:



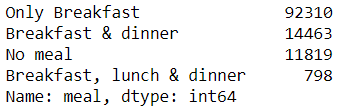
We then look at the distribution of the bookings for children and is\_canceled:

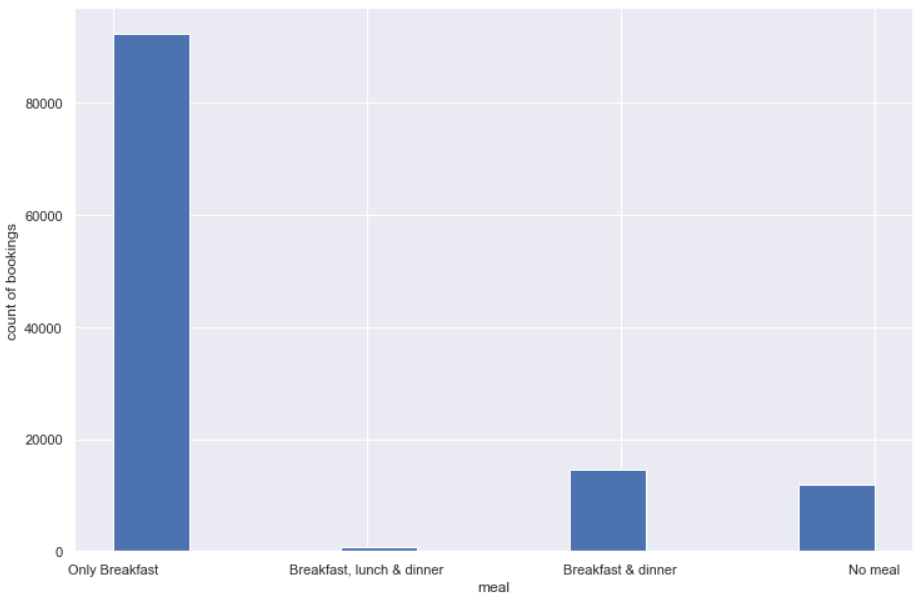


The probability of booking\_cancellation for bookings with 2 children is higher than the bookings with 0 or 1 or 3 children.

**meal:**

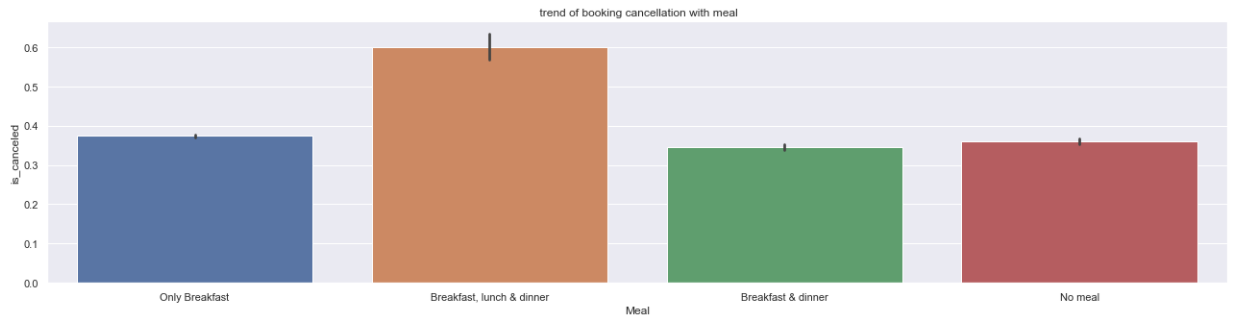
The value counts for meals is as follows:





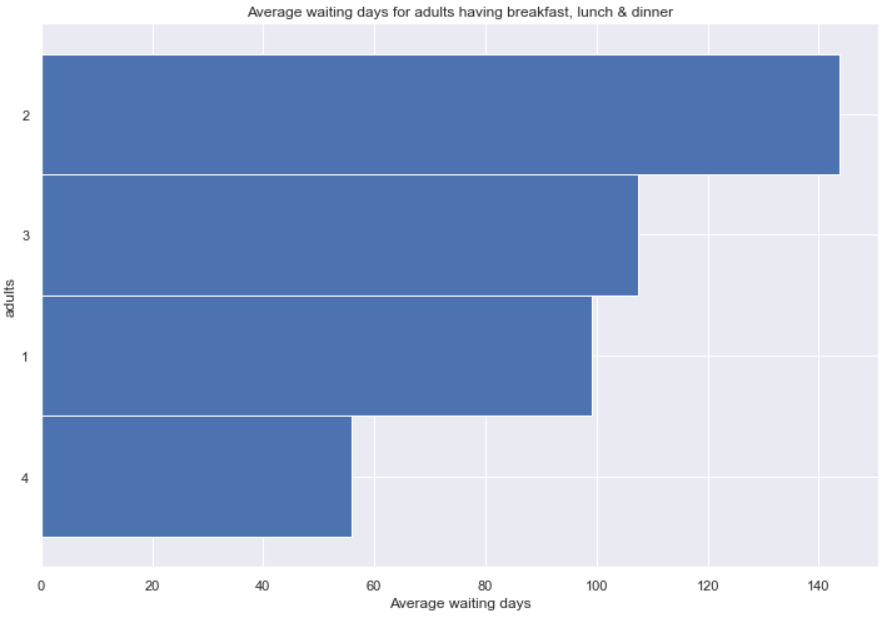
The above plot shows is a histogram of the variable. Only Breakfast is the section that has the highest number of bookings in this category. It is also interesting to note that considerable number of customers have chosen the ‘No meal’ option.

Let now look at this variable with respect to is\_canceled:



It can be seen that the Breakfast, lunch & dinner has the highest probability of the booking being cancelled and is close to 0.6 with also a considerable variability ranging upto 0.7.

We shall now look at the distribution of customers who chose Breakfast, lunch & dinner, average waiting\_days and the number of adults:

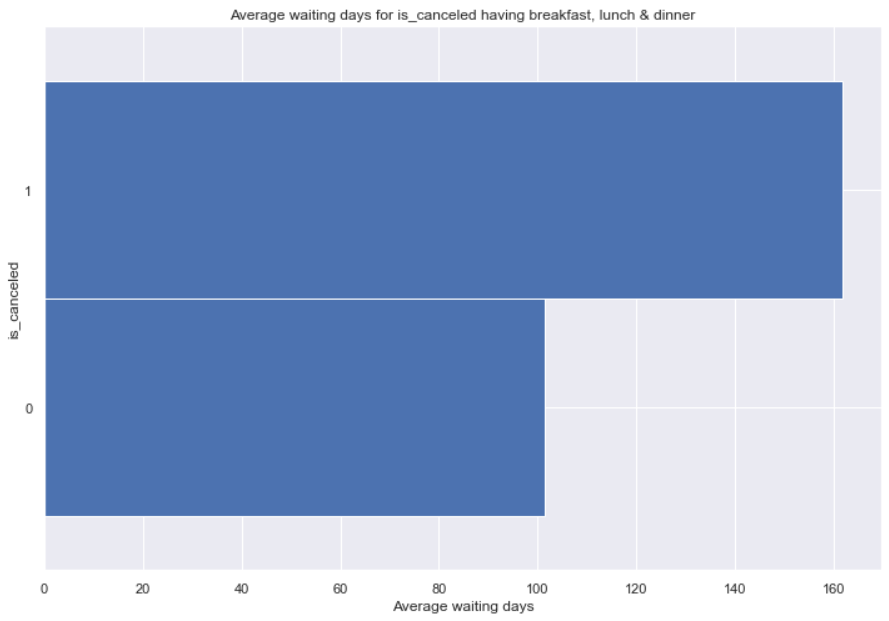


It can be inferred that among the customers who chose Breakfast, lunch & dinner, the average waiting days for 2 adults is the highest and is more than 140 days.

This is followed by the average waiting days for 3 adults and is close to 110.

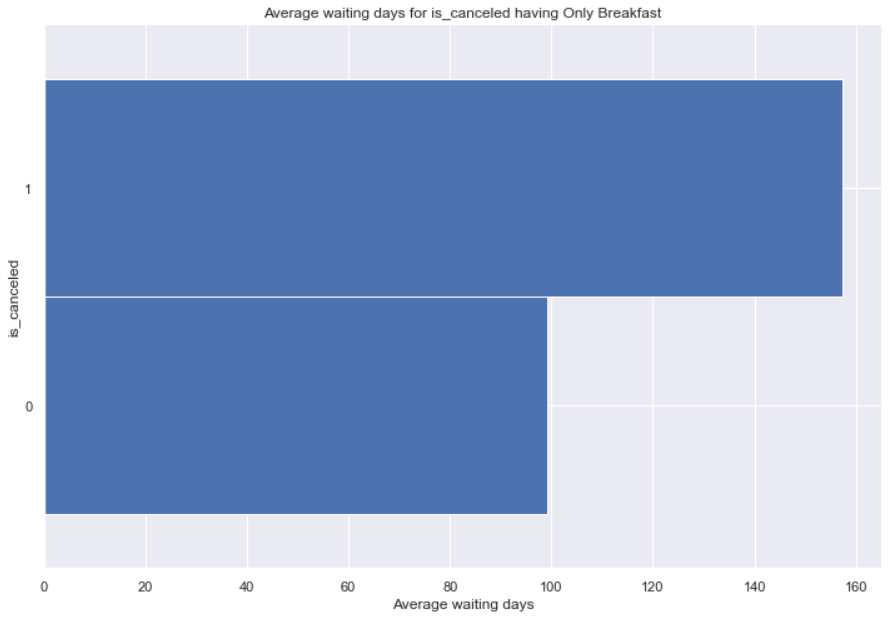
Customers with 4 adults have the least average waiting days in this category.

We now look at the section of customers who chose: Breakfast, lunch & dinner with respect to is\_canceled:



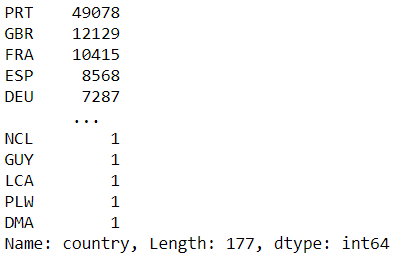
For customers in this category, to can be seen that the average waiting days for customers who cancelled their hotel booking is more than 160 days. And the customers who did not cancel their booking have an average waiting days of just above 100 in this section.

A similar trend is observed for customers who chose: Only Breakfast as their meal preference.

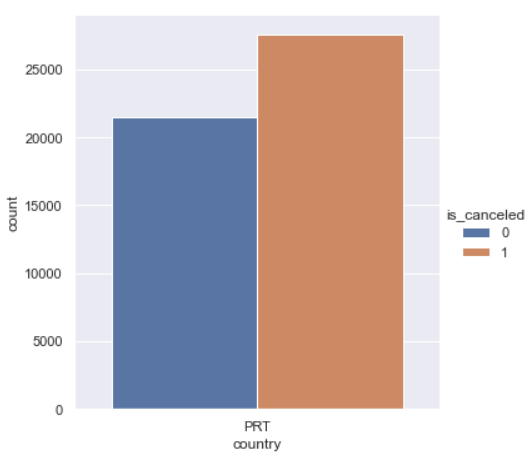


**country:**

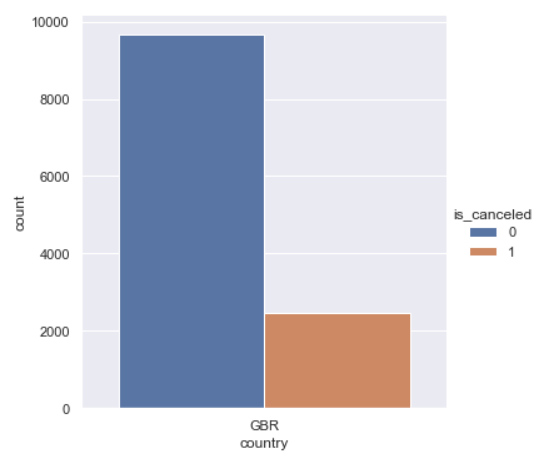
Looking at the value counts of this variable



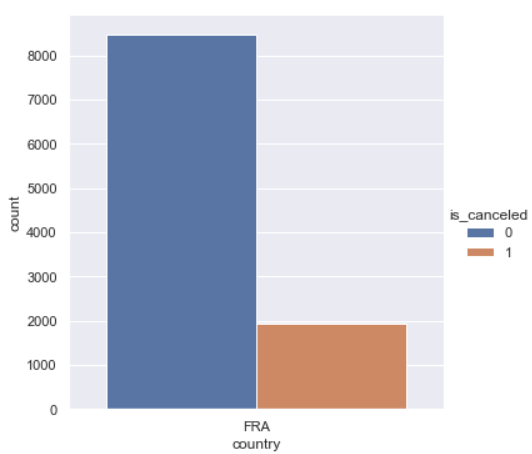
Let us look at the behaviour of the target variable with respect to a some countries that have a high booking count:



The country that has the highest booking count is PRT. In this the proportion of bookings cancelled is more than that of the bookings not being cancelled and is more than 1.25 times.

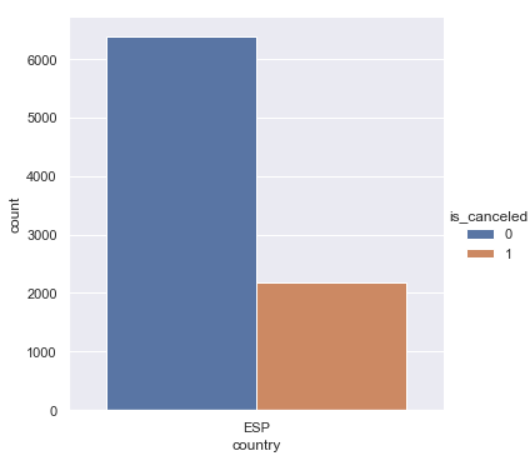


This is for the country GBR and bookings being cancelled by customers from this country is lesser than the bookings not being cancelled.

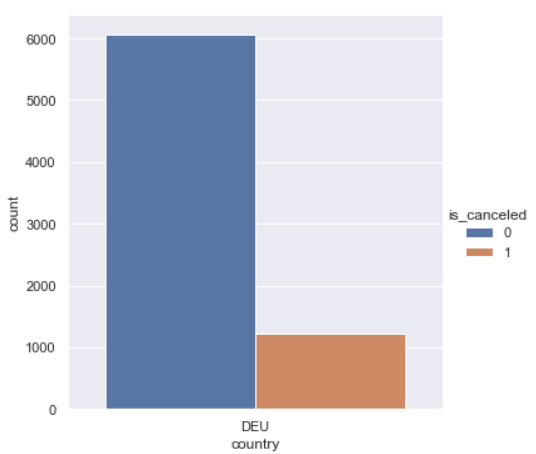


For the country FRA, the inference of the above plot is similar as that of the country GBR, although the count of bookings for FRA is much lower than that of the count of bookings for the country GBR.

This is for the country ESP:

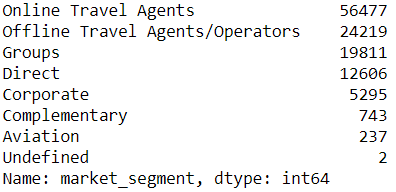


For the country DEU:

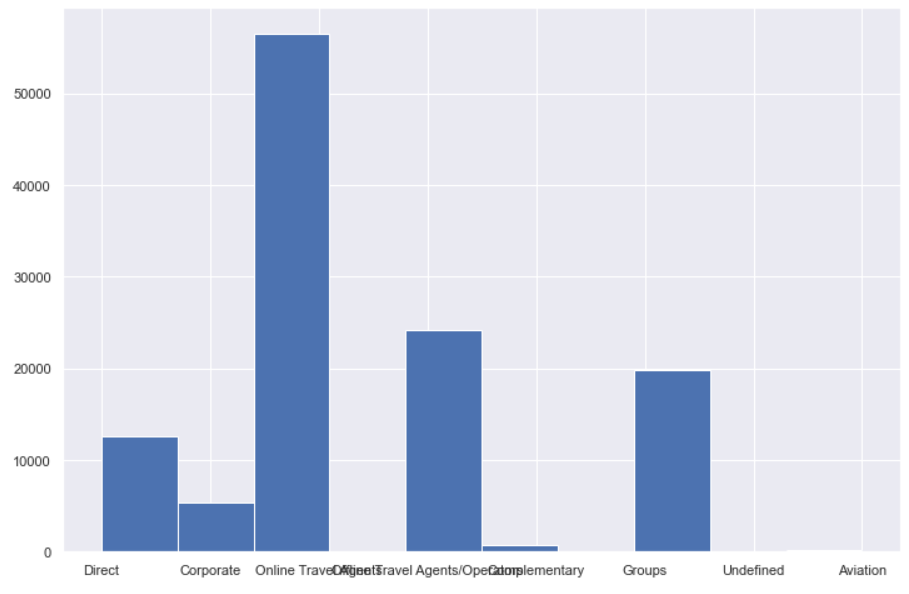


**market\_segment:**

Looking at the value counts of this variable:

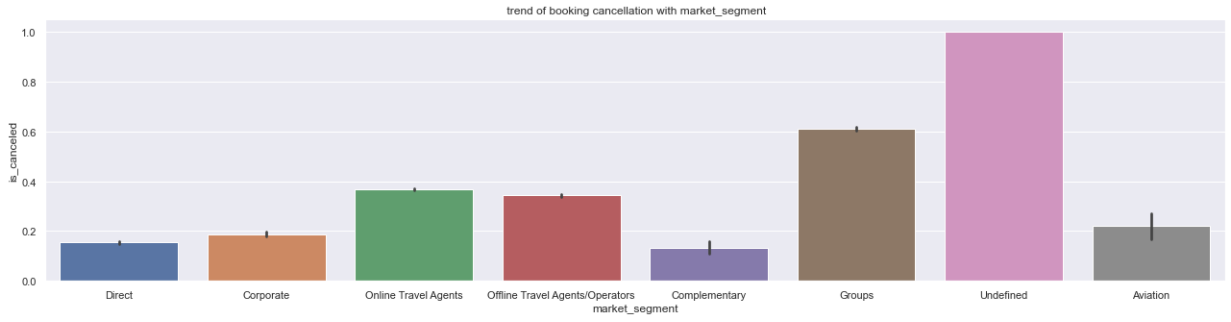


We will now plot the histogram of this variable:



The Online Travel Agents have the highest count of bookings in this category. This is followed by the Offline Travel Agents / Operators.

We will also look at this variable with respect to the target variable:



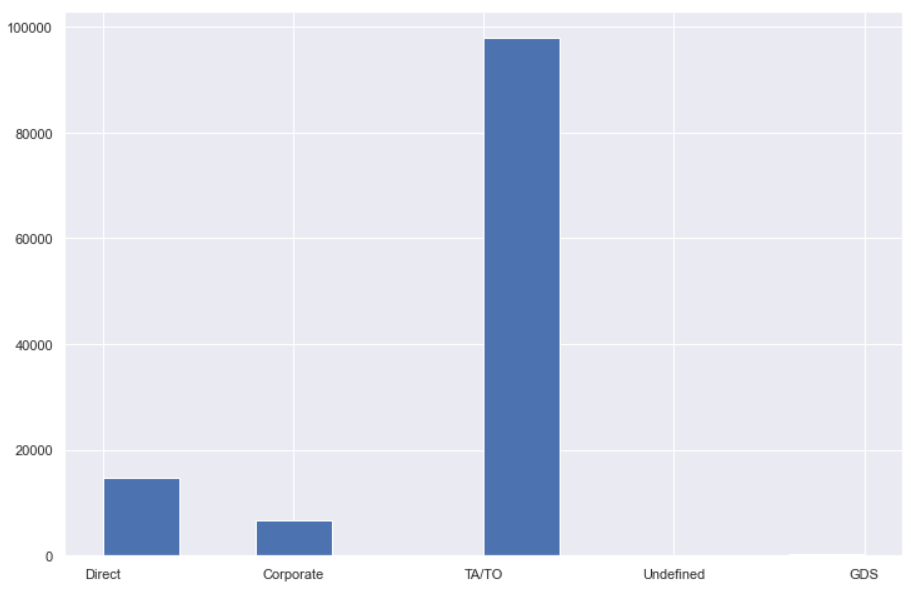
It is to observe that the segment: Undefined has the highest probability of the bookings being cancelled and is close to 1. This is probably because the count of bookings in this category is only 2.

The next segment in this category with a high probability of booking cancellation is the Groups. There are 19811 bookings with groups and they have a booking cancellation probability of 0.6. This is too much when you look at it in terms of numbers. This translates into (19811 x 0.6) = 11886.6 bookings having a high probability of being cancelled. Further investigation into the group size and the reason for cancellation needs to be looked into.

**distribution\_channel:**

Looking into the value counts:





The above plot shows us the histogram of the variable.

We shall now look at the behaviour of this variable with is\_canceled:

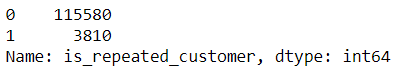


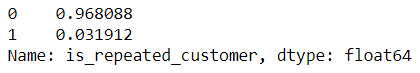
The Undefined section have the highest probability of booking cancellation with 0.8 and also a very high variability.

Next in line is the TA/TO section which have a 0.4 probability of booking cancellation. This means that out of 97870 customers in this section, (0.4 x 97870 = 39148) customers run the risk of having their hotel booking cancelled.

**is\_repeated\_customer:**

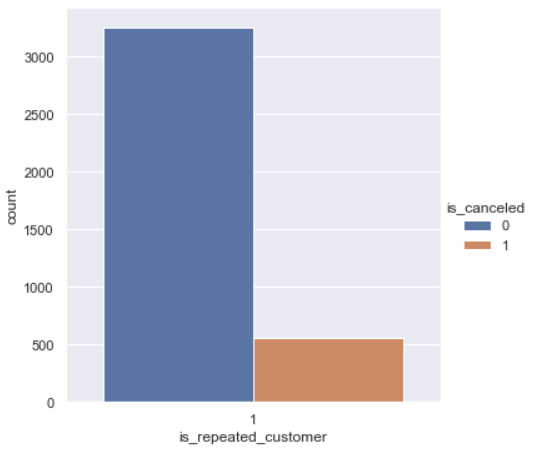
This is a binary variable. The value counts and the value counts with percentage are as follows:



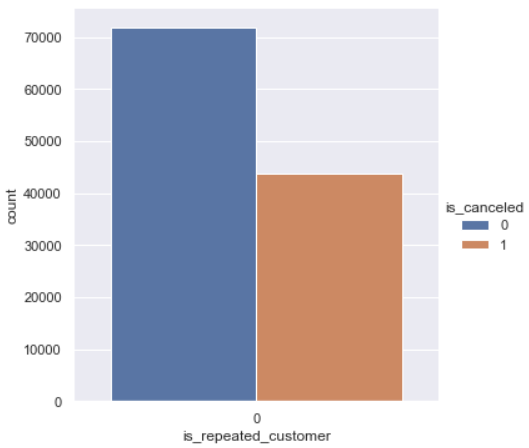


Behaviour of this variable with the target variable which is also binary:

When the customer is a repeated\_customer:



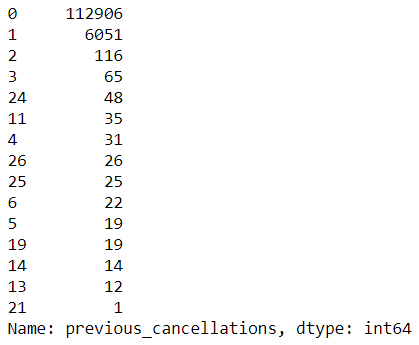
When the customer is not a repeated\_customer:



It can be seen from the above plot that when the customer is not a repeated\_customer for the hotel, the probability of booking cancellation shoots up. This means that the repeated\_customers tend to cancel their hotel booking less often.

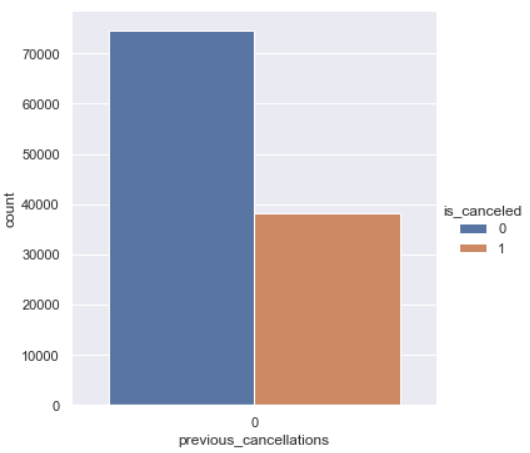
**previous\_cancellations:**

The value counts of this variable is as follows:

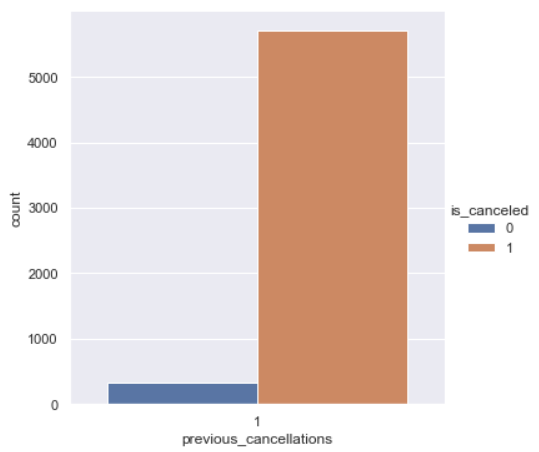


Let us look at the behaviour of this variable in response to the target variable of interest.

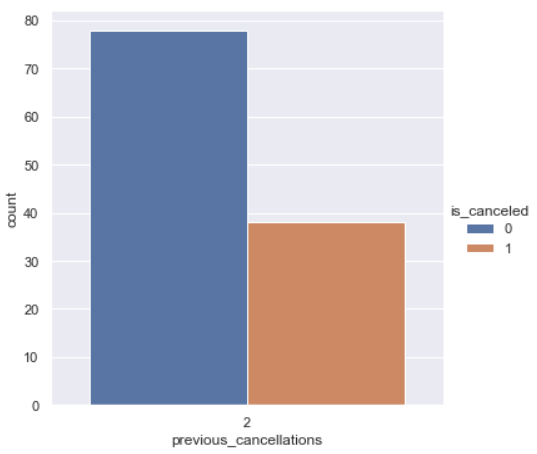
When the previous\_cancellations = 0



When the previous\_cancellations = 1



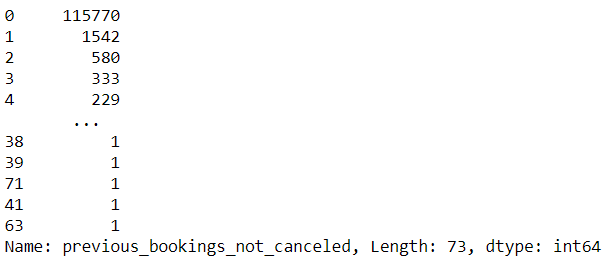
When the previous\_cancellations = 2



It is truly alarming for the hotel to see that most of the customers who have had their previous\_cancellations as 1 have cancelled their hotel booking this time as well.

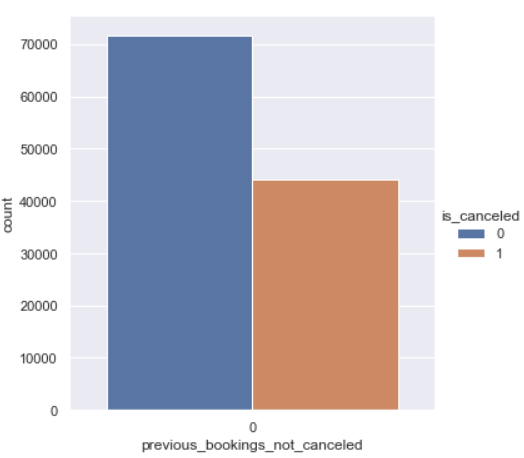
**previous\_bookings\_not\_canceled:**

The value counts of this variable are as follows:

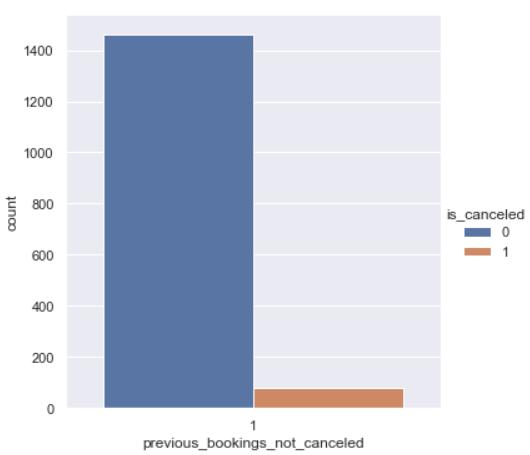


Let us look at this variable in response to the target variable:

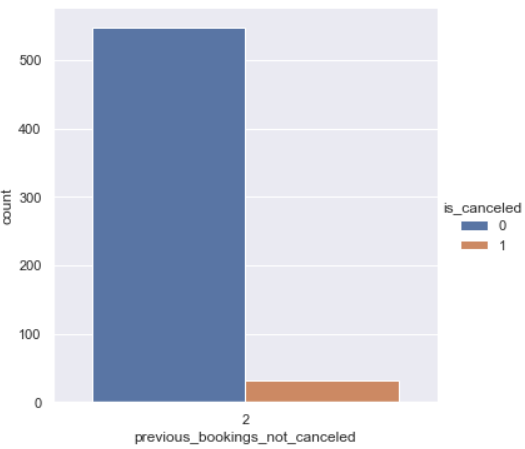
When the previous\_bookings\_not\_canceled is 0:



When the previous\_bookings\_not\_canceled is 1:



When the previous\_bookings\_not\_canceled is 2:

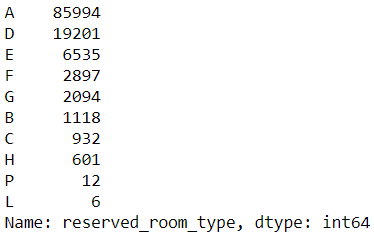


When the previous\_bookings\_not\_canceled is 0, the probability of booking cancellation is high.

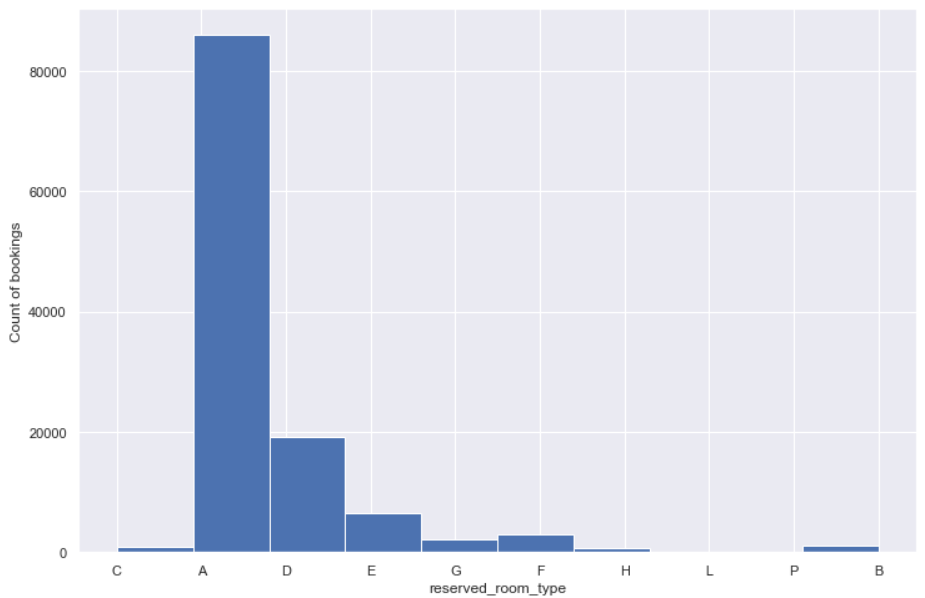
When the previous\_bookings\_not\_canceled is 1 or 2, the booking is very less likely to get cancelled.

**reserved\_room\_type:**

This is a categorical variable and the value counts are as follows:



Let us look at the histogram of this variable:



The reserved\_room\_type A, D and E constitute for the majority of bookings.

Let us now look at the response of this variable with is\_canceled:



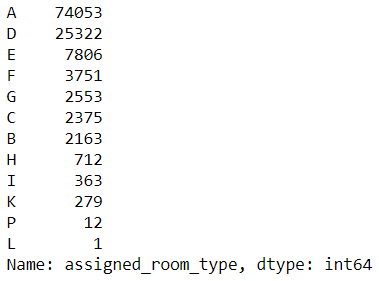
This is to note that the reserved\_room\_type of P has a 100% booking cancellation rate.

The reserved\_room\_type of L has a high variability which is indicated by the long Confidence Interval band.

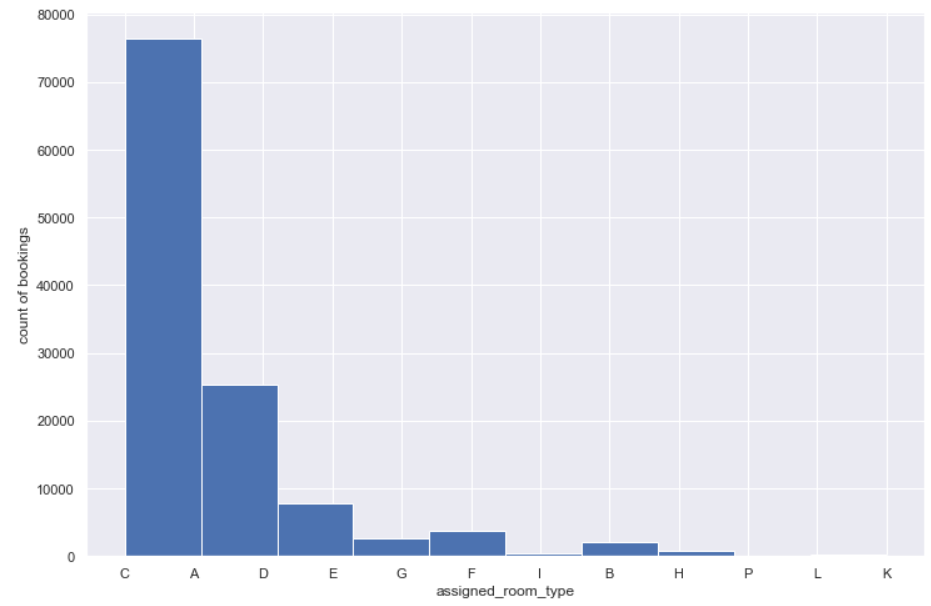
The next reserved\_room\_type with a high booking cancellation probability is H with around 0.4.

**assigned\_room\_type:**

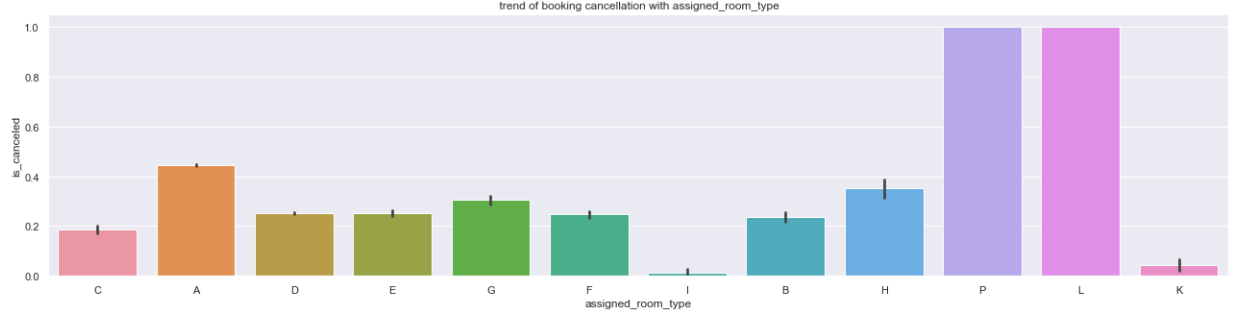
This is again a categorical variable and the value counts of this variable are as follows:



We see a histogram of this variable:



We shall now look at the trend of booking cancellation with the assigned\_room\_type:



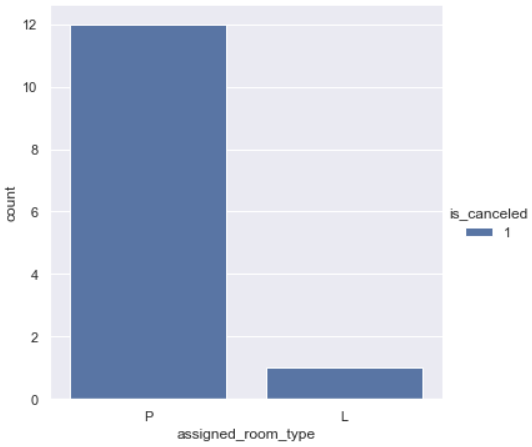
The assigned\_room\_type of P and L have a probability of 1 on booking cancellation.

Though the booking counts are very less for these, the hotel can be rest assured that once either of these rooms are assigned to the customer, the customer is bound to back off.

Another major concern is the assigned\_room\_type of A that sees a probability of booking cancellation of close to 0.45. As the booking volume is high for this assigned\_room\_type, so will be the number of booking cancellations.

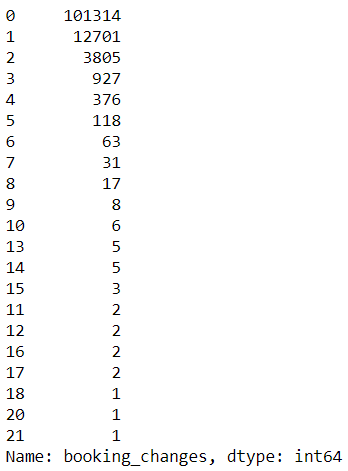
The assigned\_room\_type of I and K are the most dependable in this variable.

Just to showcase the earlier finding of the assigned\_room\_type of P and L:



**booking\_changes:**

Let us look at the value counts of this variable:

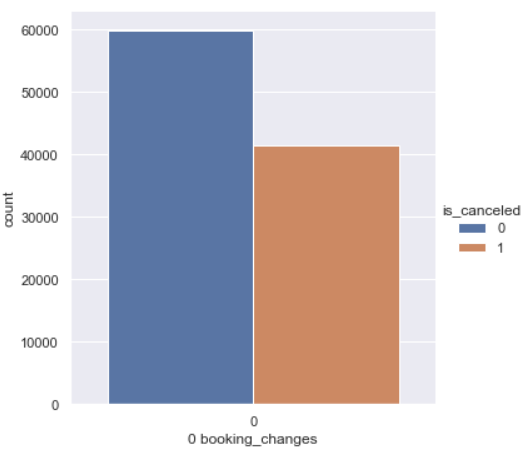


This is a numeric variable and the entries beyond the value of 5 booking\_changes are not much.

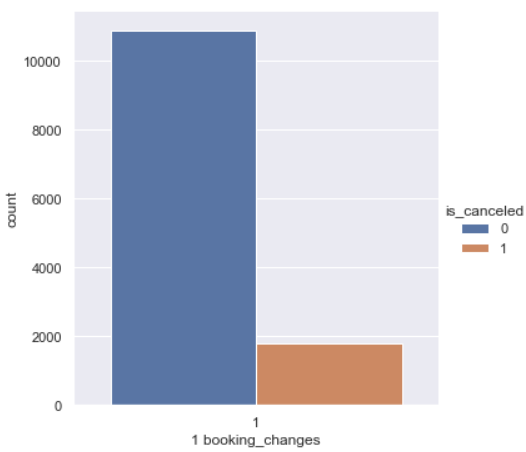
So, ideally speaking, we should consider this as an ordinal variable and only upto 5 booking\_changes.

Let us go ahead and look at the behaviour of this variable with is\_canceled:

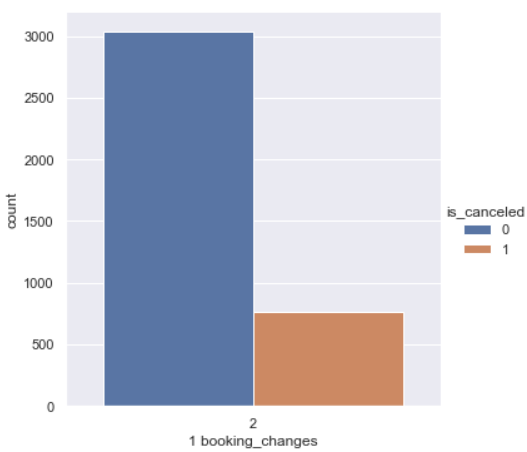
When booking\_changes = 0



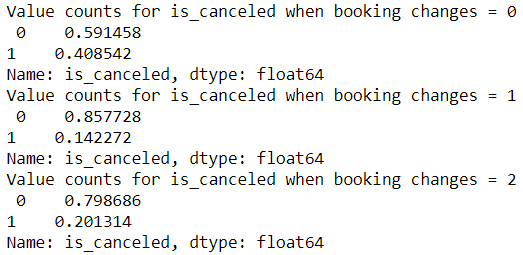
When booking\_changes = 1



When booking\_changes = 2



Let us look at the proportion of booking cancellation in numbers:

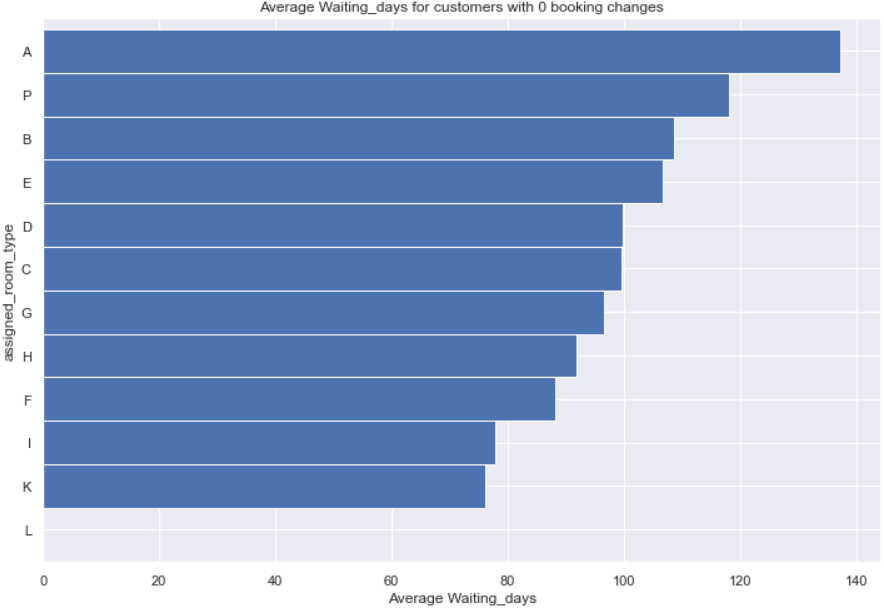


It is to note that when the booking\_changes is 0, the probability of booking cancellation is almost 41%.

Customers who make booking changes once or twice are less prone to cancelling the hotel booking.

On a lighter note: The hotel needs to encourage the customers to make booking\_changes in order to save the booking from being cancelled entirely.

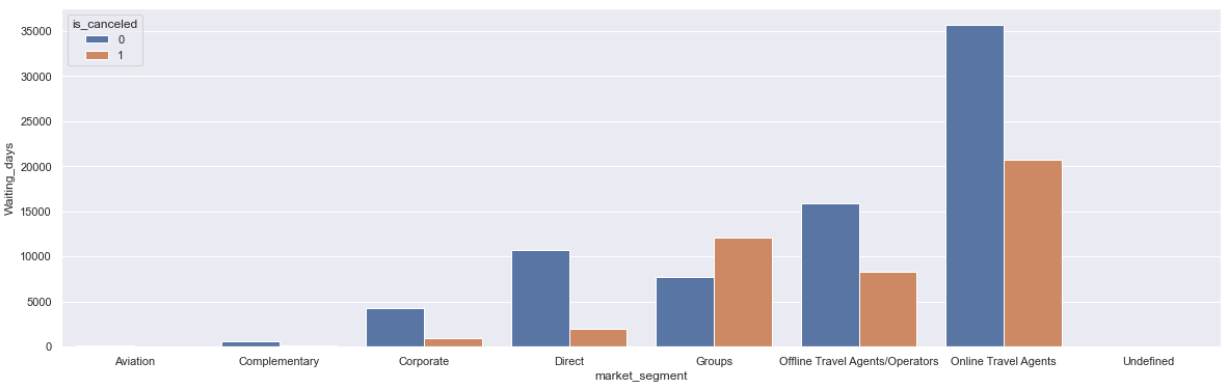
We shall now group the assigned\_room\_type based on average Waiting\_days and sort the same in the descending order:



The average Waiting\_days for assigned\_room\_type: A is the highest followed by the assigned\_room\_type: P and B

It is to note that the assigned\_room\_type of I and K have the lowest Waiting\_days. These were the same assigned\_room\_type that had very low probability of booking cancellation as mentioned earlier.

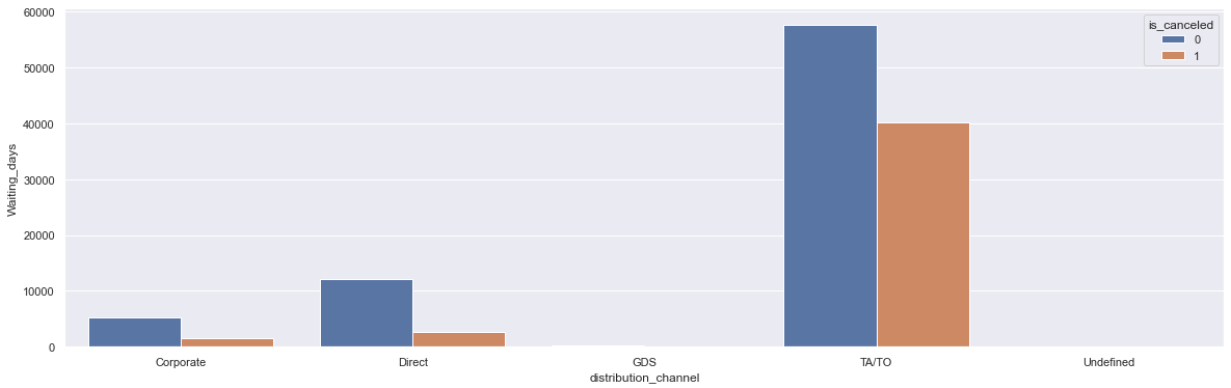
We shall now group the market\_segment based on Waiting\_days with a hue on is\_cancelled:



The Online Travel Agents have the highest Waiting\_days and also have a high proportion of is\_canceled.

The Offline Travel Agents/Operators have a higher proportion of missing values and so does the category of Groups as discussed earlier. It is also to note that the customers in the Group section who cancelled their bookings have had higher Waiting\_days than the customers who did not cancel.

We shall now group the distribution\_channel based on Waiting\_days with a hue on is\_cancelled:

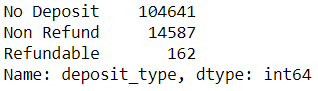


The distribution channel of TA/TO have had the highest Waiting\_days and also a high proportion of booking cancellation.

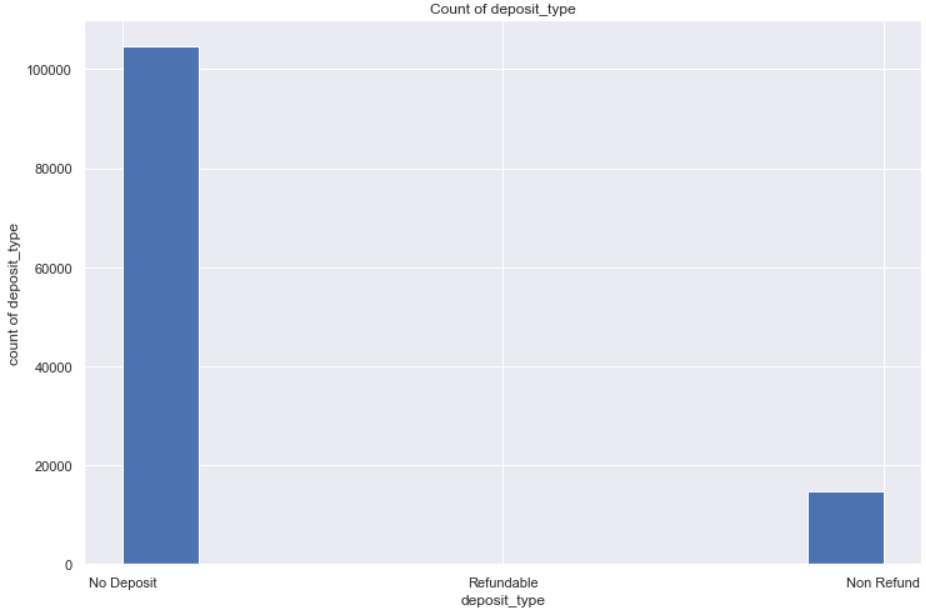
The Waiting\_days of other distribution\_channel segments are low in comparison to the segment of TA/TO.

**deposit\_type:**

We shall now look at the value counts of this variable:



Looking at the histogram of this variable:



Let us look at the behaviour of this variable with respect to the target variable.

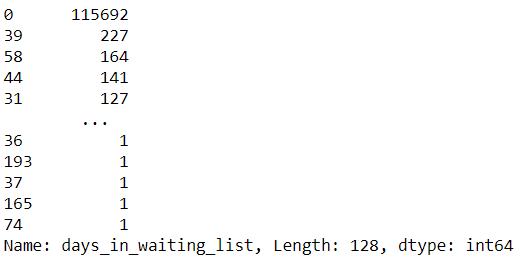


The category of Non Refund has almost a 100% chance of booking cancellation.

This is worrying as it has close to 15000 entries.The Refundable deposit\_type has the lowest probability of booking cancellation.

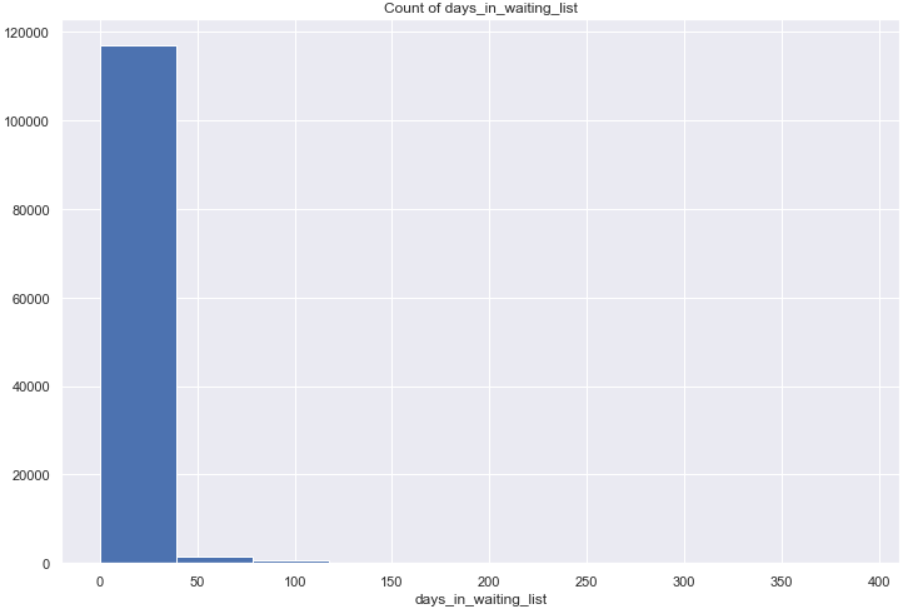
**days\_in\_waiting\_list:**

Value counts is as follows



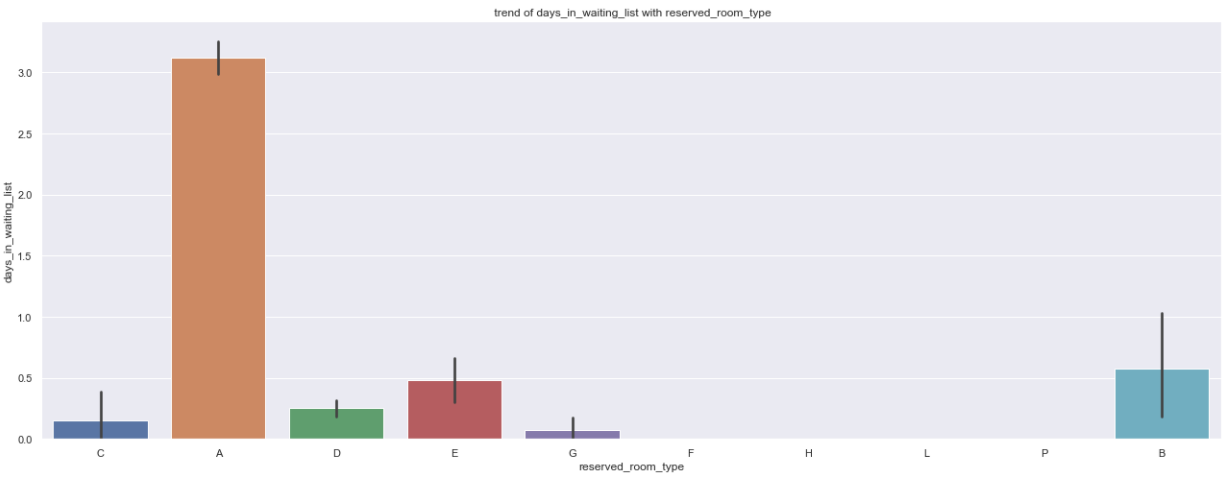
It can be seen that this is more of a skewed distribution as most of the values have 0 as days\_in\_waiting\_list.

We shall now look at the histogram of the same:



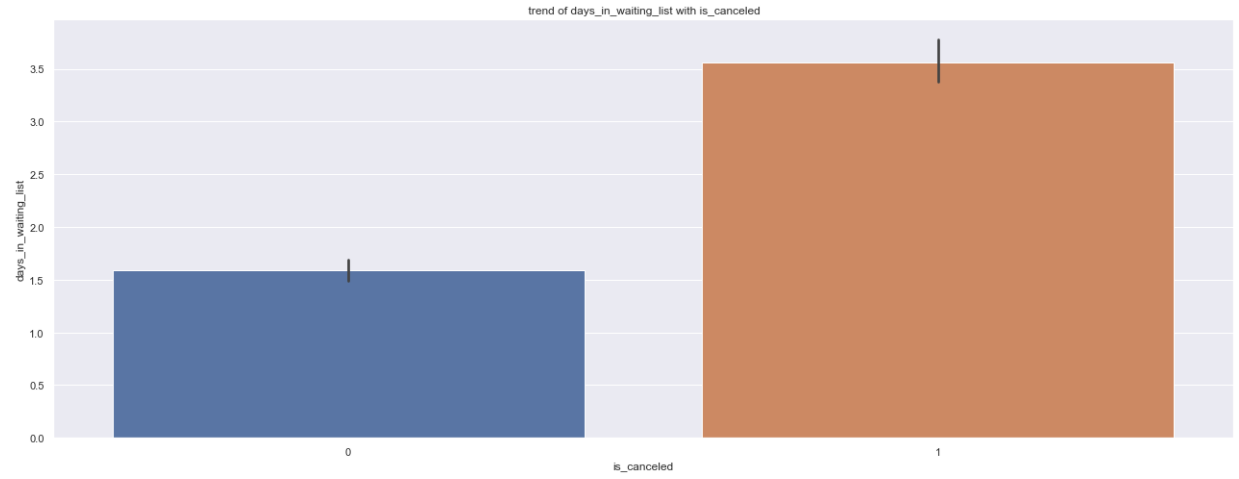
This is a one sided distribution, as stated before.

We shall now look at the distribution of days\_in\_waiting\_list with the reserved\_room\_type:



The bookings on the reserved\_room\_type A have witnessed the maximum days\_in\_waiting\_list. This is followed by the reserved\_room\_type B that has much lesser days\_in\_waiting\_list as compared to that of A.

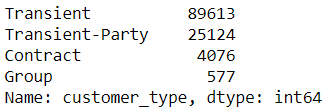
Let us now look at a plot between is\_canceled and days\_in\_waiting\_list:



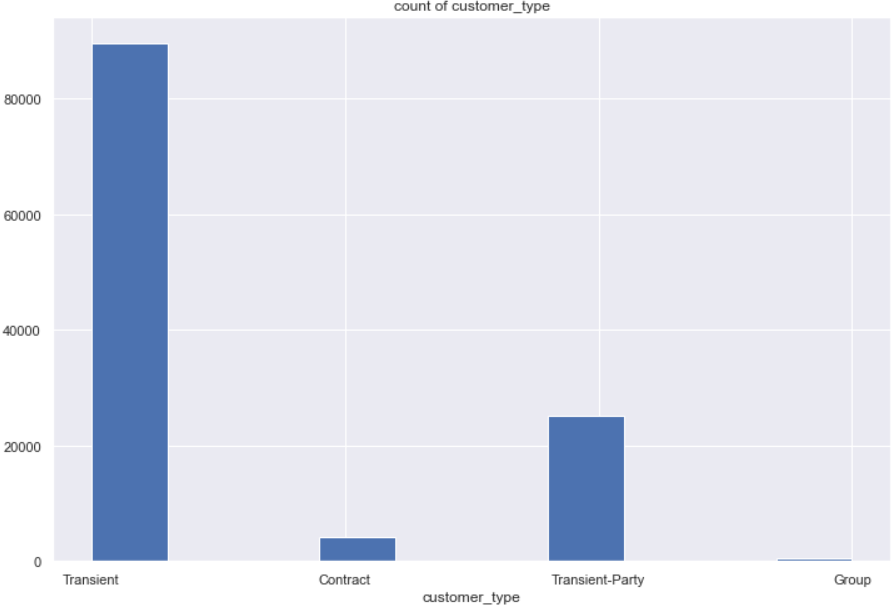
The days\_in\_waiting\_list and is\_canceled are directly proportional beyond around 150 days.

**customer\_type**

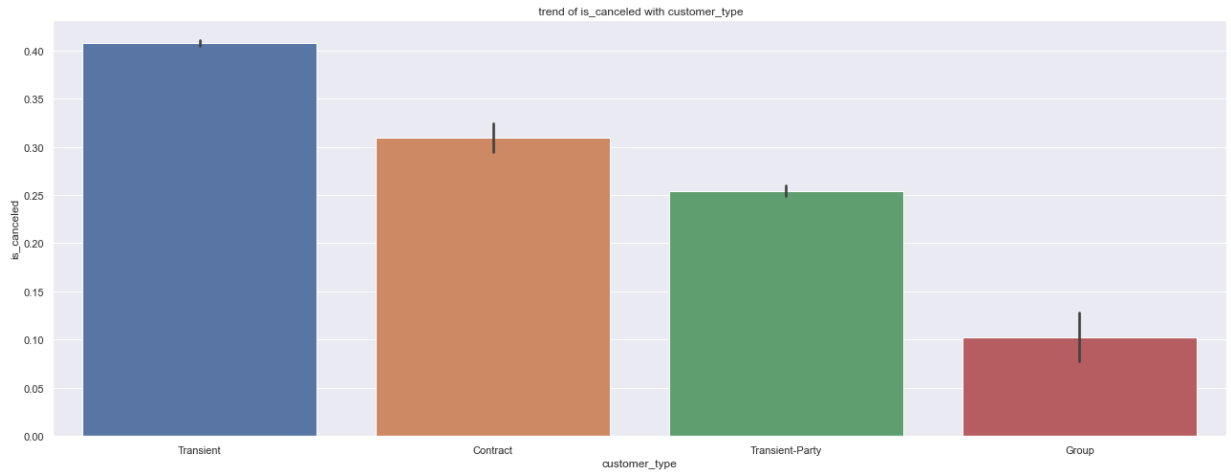
Let us look at the value counts of these variables:



A histogram of the same is shown.



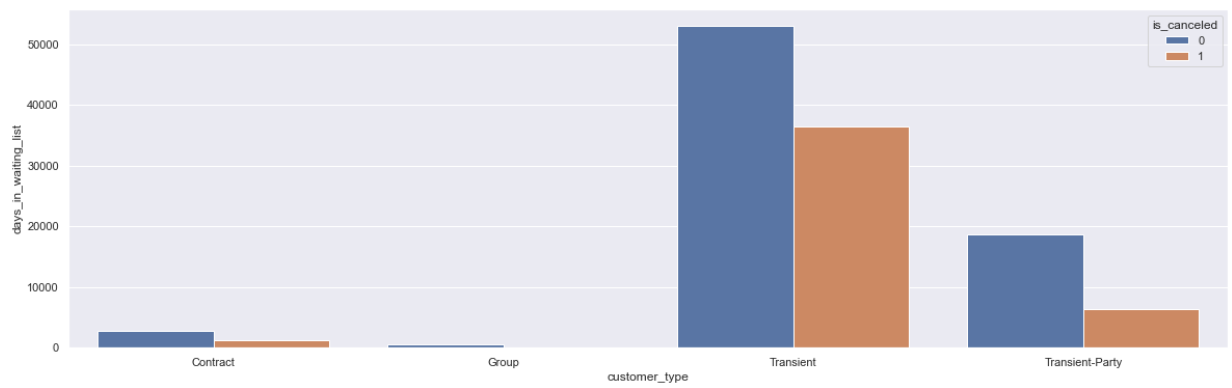
Trend of is\_canceled with customer\_type:



This is an interesting finding that Transient section of the customer\_type has the highest probability of being cancelled and it stays at 0.4

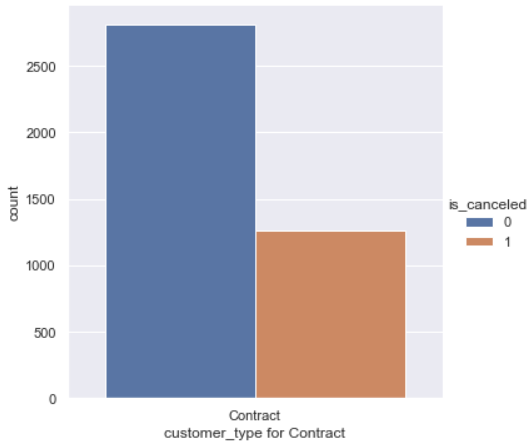
This is then followed by the Contract section and the section that has the least probability of having their hotel booking cancelled is the Group. Group has the largest variability though. It is to note that the scale of hotel bookings between the customer\_type sections: Group and Transient vary by a large margin.

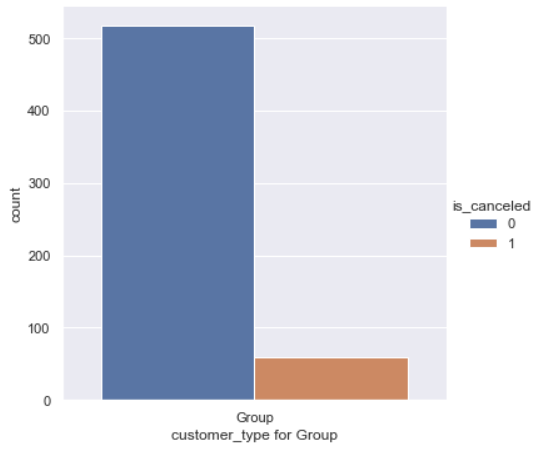
Let us now plot the customer\_type sorted based on days\_in\_waiting\_list with the hue as is\_canceled:

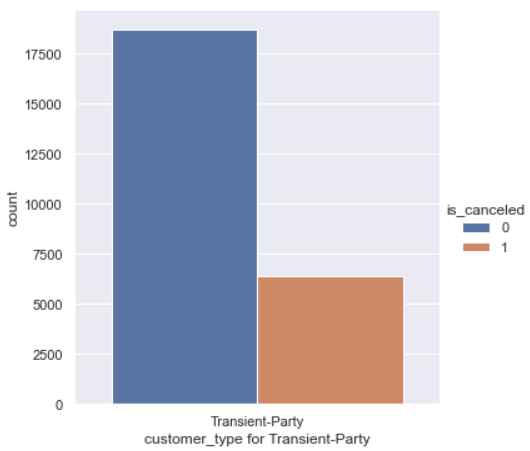


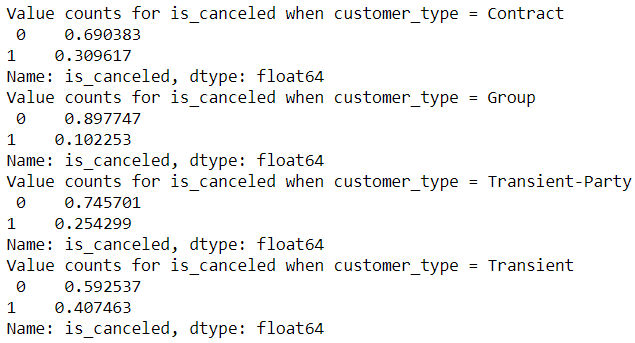
The Transient section have the highest days\_in\_waiting\_list and their proportion of is\_canceled is the highest. This is then followed by the Transient-Party.

We shall now look at the countplot of each section and their proportion of is\_canceled:



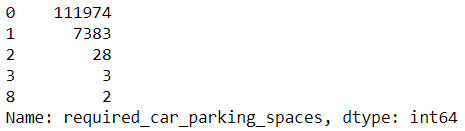






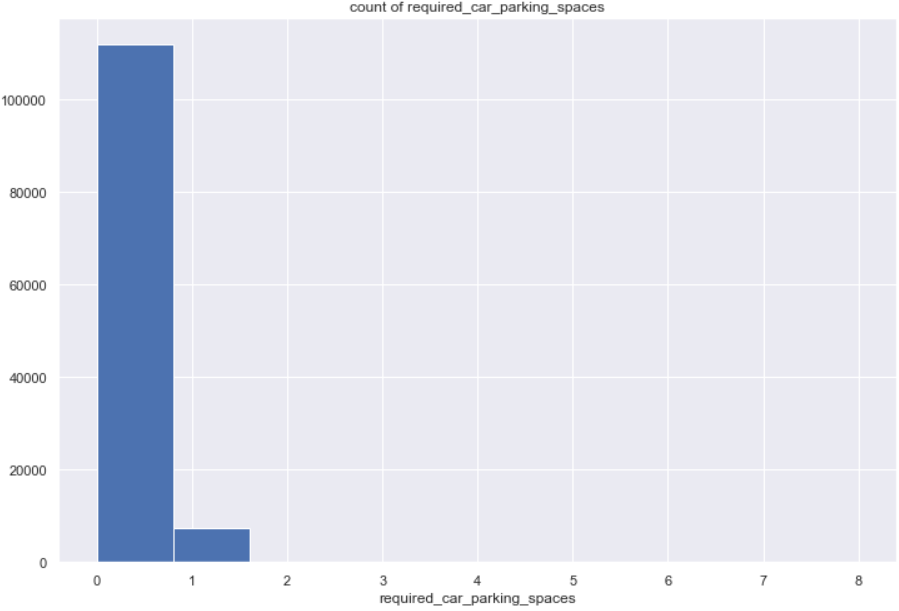
**required\_car\_parking\_spaces**

This is a numeric variable that is almost like a category. Looking at the value counts:

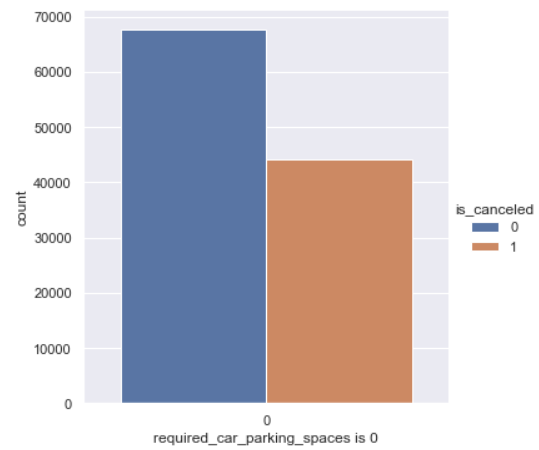


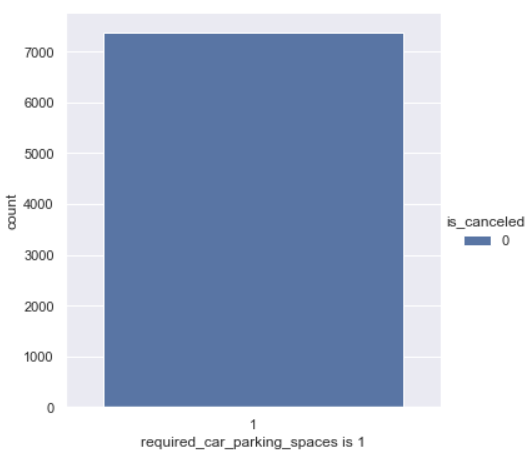
It can be seen that most of the customers do not require car parking space.

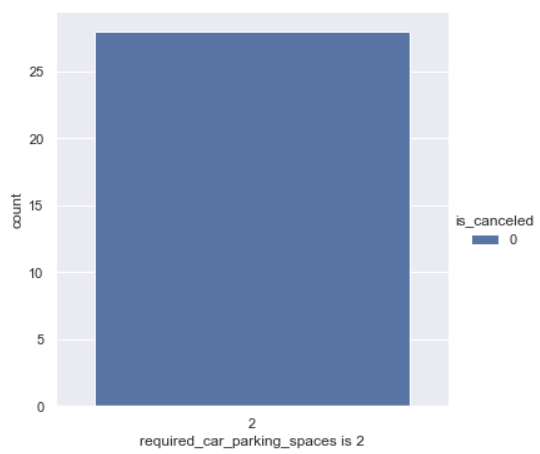
Looking at the histogram of the same:

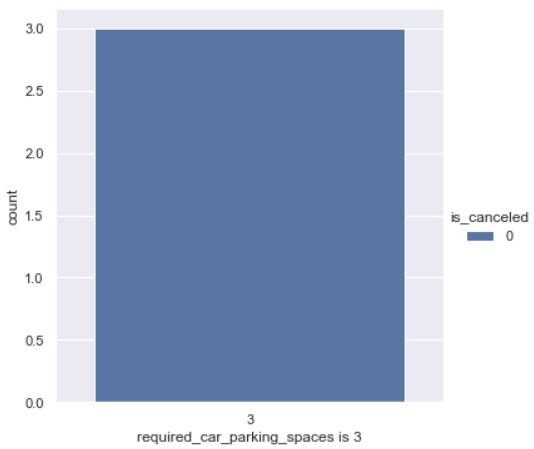


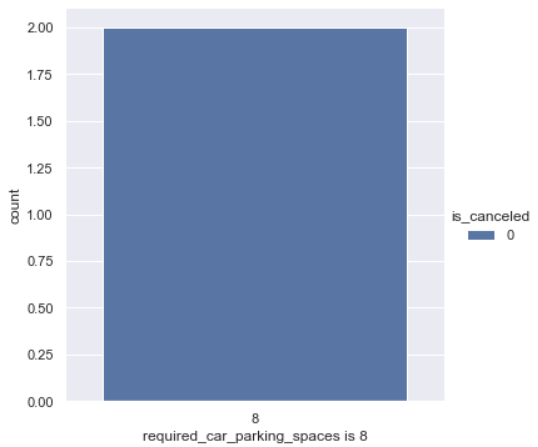
Let us look at the behaviour of required\_car\_parking\_spaces with is\_canceled for each category:







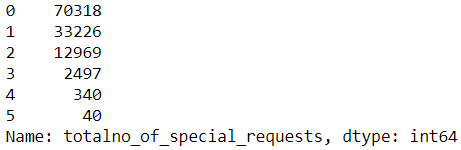




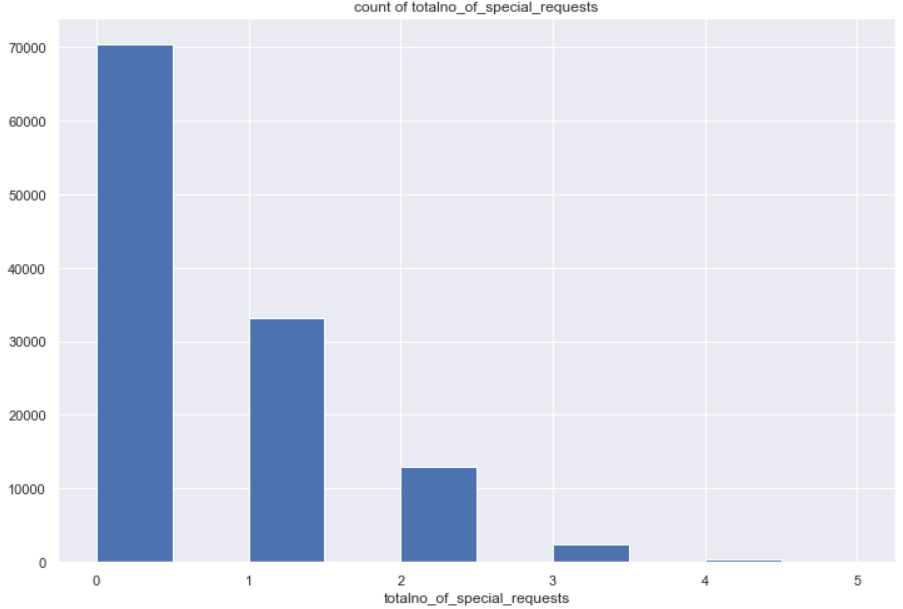
It can be seen that only the customers who have required\_car\_parking\_spaces as 0 are the ones who have a chance of cancelling their hotel booking.

**totalno\_of\_special\_requests**

Firstly, let us look at the value counts of this variable:



Like the previous variable, this is also numeric in expression but categorical in nature. Looking at the histogram of the same:



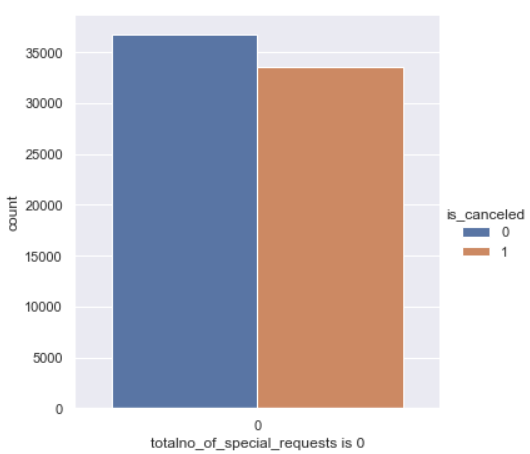
There is a steady decline in the count of bookings across each numerical category of this variable.

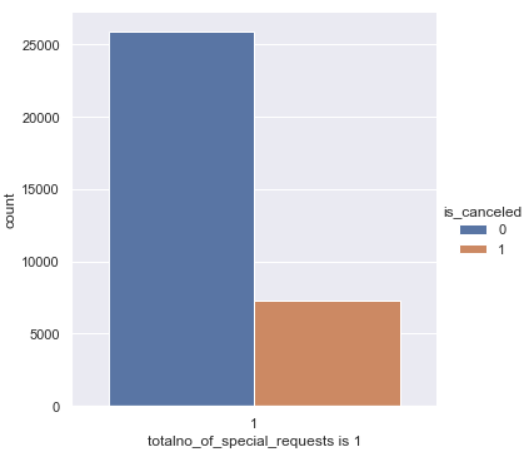
Looking at this variable as a whole with the binary target variable:

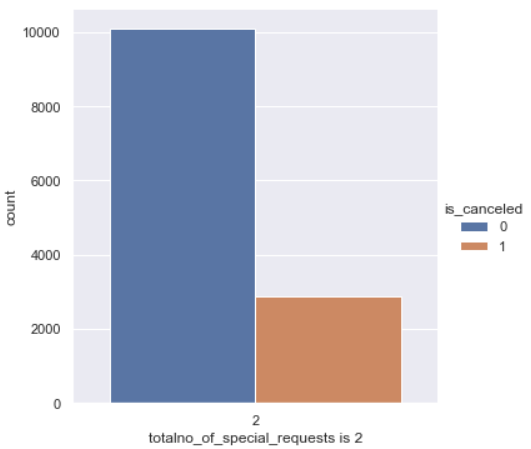


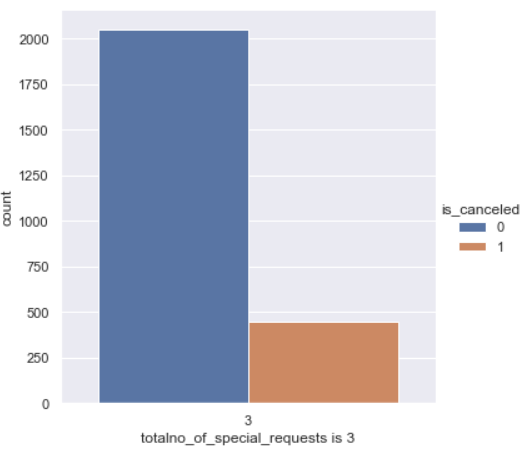
The section that has the highest probability of hotel booking cancellation (close to 0.48) is the totalno\_of\_special\_requests as 0. This is then followed by the people of totalno\_of\_special\_requests as 2 and 1 that have a probability of booking cancellation as slightly more than 0.2.

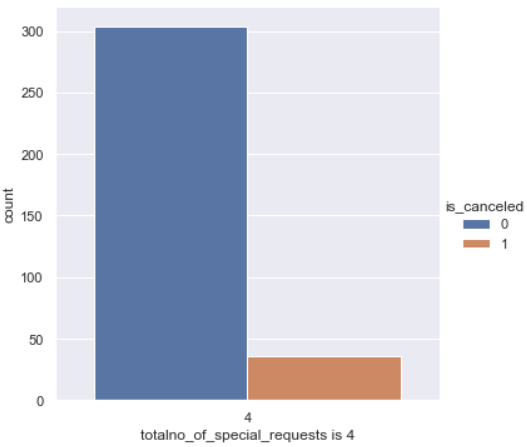
Let us have a look at each category for the is\_cancelled to understand better:

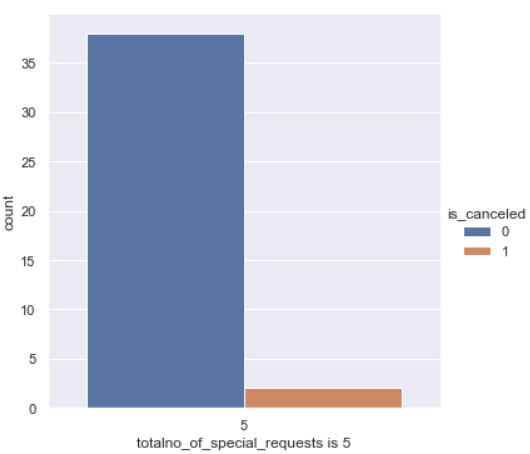




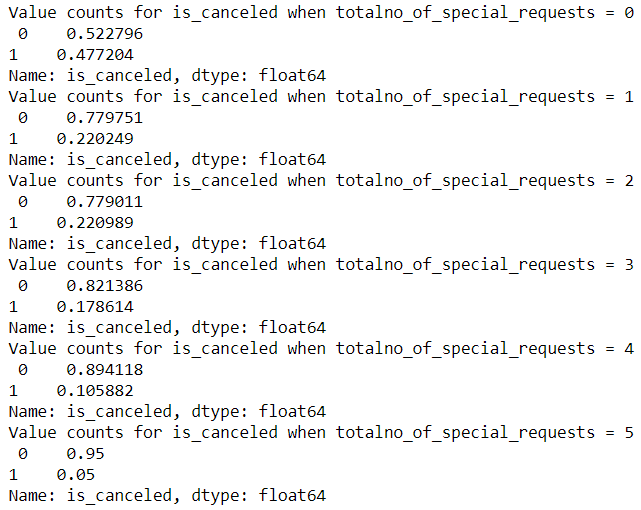








Looking at the proportion of is\_canceled numerically:



**Variable Transformation:**

We shall perform One-hot encoding of the following variables:

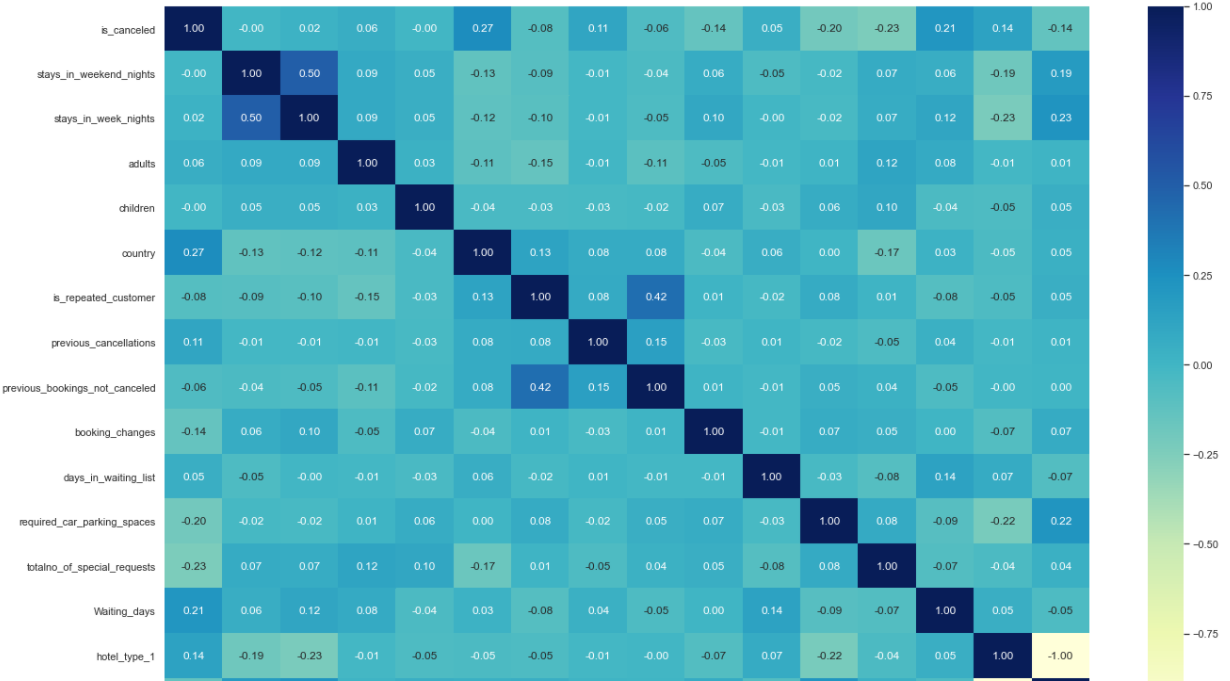
'hotel','meal','booking\_year','booking\_month','arrival\_year','arrival\_month','market\_segment','distribution\_channel','deposit\_type','customer\_type','reserved\_room\_type','assigned\_room\_type'

We shall perform Label Encoding of the variable: ‘country’

**Correlation using Heatmap:**

We use the following variables in the following heatmap to depict the correlation:

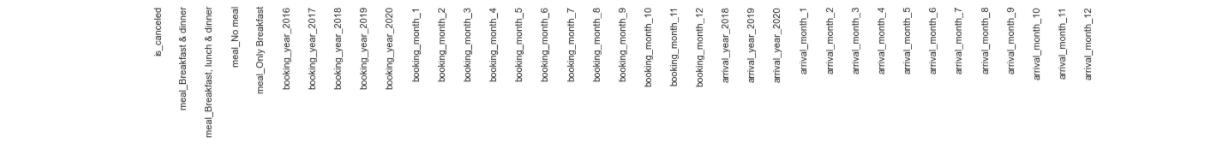
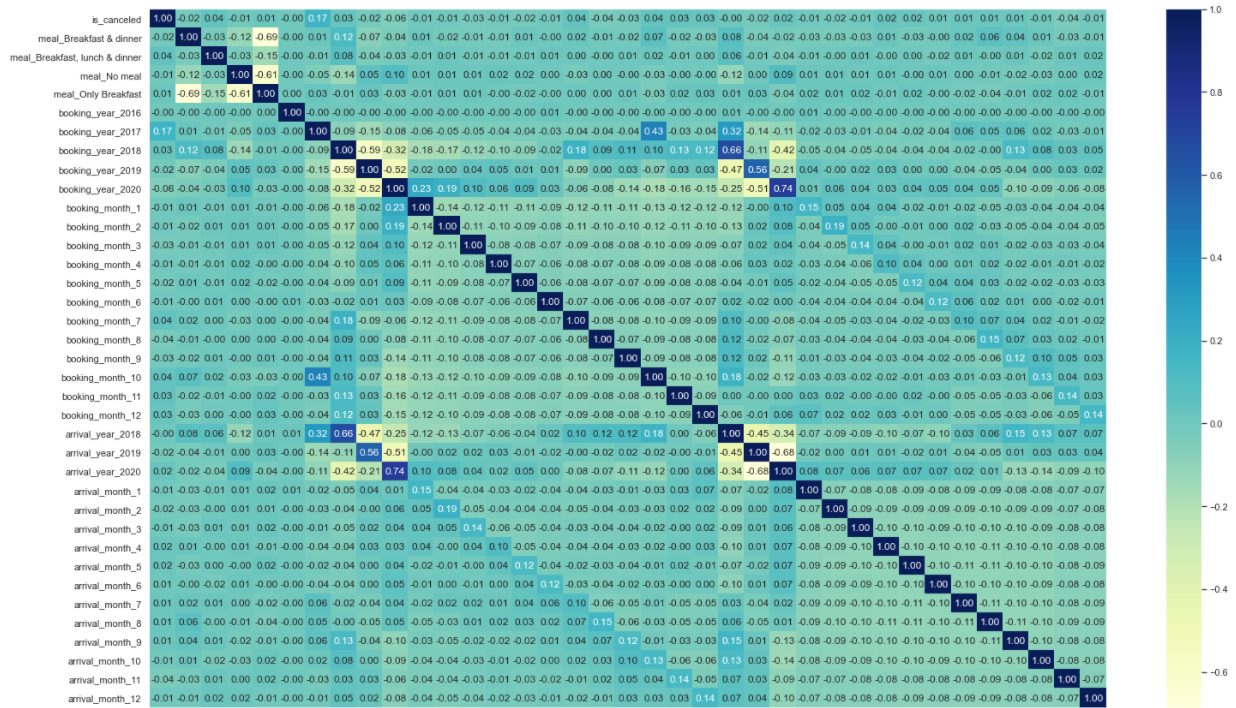
'is\_canceled','stays\_in\_weekend\_nights','stays\_in\_week\_nights','adults','children','country','is\_repeated\_customer','previous\_cancellations','previous\_bookings\_not\_canceled','booking\_changes','days\_in\_waiting\_list','required\_car\_parking\_spaces','totalno\_of\_special\_requests','Waiting\_days','hotel\_type\_1','hotel\_type\_2'



* There is a low correlation between country and is\_canceled.
* A correlation worth mentioning is present between: is\_repeated\_customer and previous\_bookings\_not\_canceled.
* stays\_in\_week\_nights and stays\_in\_weekend\_nights also has a good correlation of 0.5. With this finding, we can probably infer that greater is the stay of the customer during week nights then the customer is highly probable to extend his stay for the weekend nights also.

**Correlation Heatmap of the following variables:**

The variables used for this heatmap are: 'is\_canceled','meal\_Breakfast & dinner','meal\_Breakfast, lunch & dinner','meal\_No meal','meal\_Only Breakfast','booking\_year\_2016','booking\_year\_2017','booking\_year\_2018','booking\_year\_2019','booking\_year\_2020','booking\_month\_1','booking\_month\_2','booking\_month\_3','booking\_month\_4','booking\_month\_5','booking\_month\_6','booking\_month\_7','booking\_month\_8','booking\_month\_9','booking\_month\_10','booking\_month\_11','booking\_month\_12','arrival\_year\_2018','arrival\_year\_2019','arrival\_year\_2020','arrival\_month\_1','arrival\_month\_2','arrival\_month\_3','arrival\_month\_4','arrival\_month\_5','arrival\_month\_6','arrival\_month\_7','arrival\_month\_8','arrival\_month\_9','arrival\_month\_10','arrival\_month\_11','arrival\_month\_12'

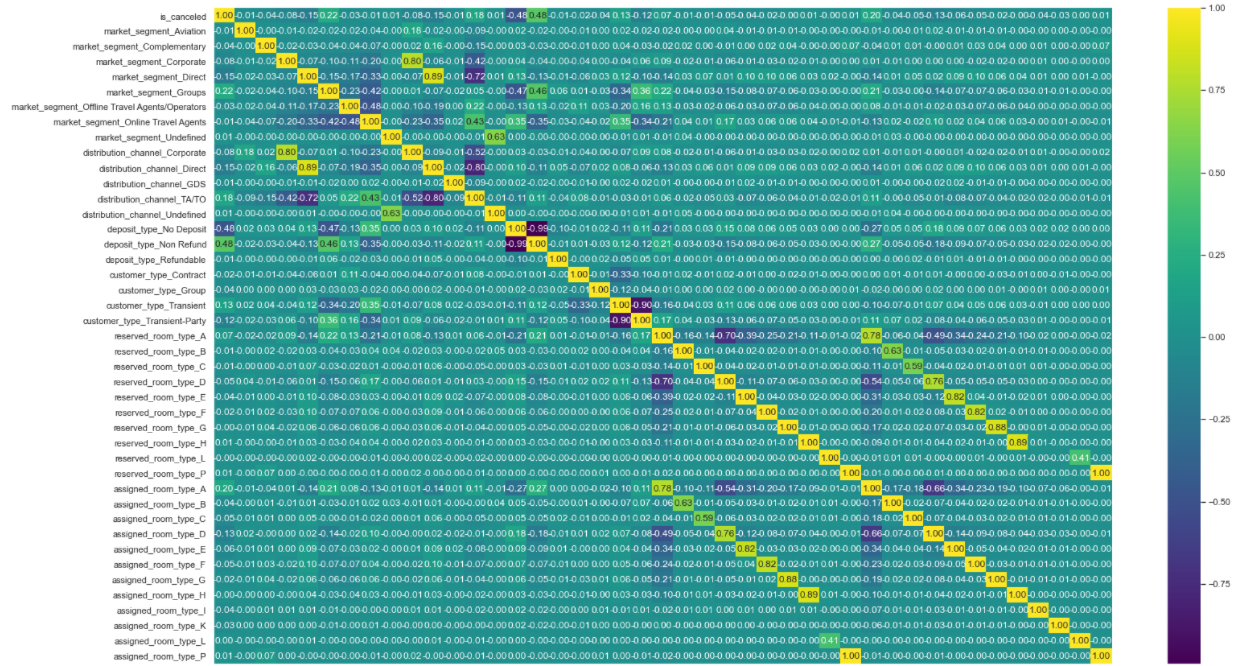


Following variables have a high correlation:

* The booking\_year, months and the arrival\_ year, months are highly correlated as they are derived variables and it is to be expected.
* There is also a high negative correlation between meal\_Only Breakfast, meal\_Breakfast & dinner and meal\_No meal.
* booking\_year\_2017 has a low positive correlation with is\_canceled.

**Correlation Heatmap of the following variables:**

The variables used for this heatmap are:

* Glad to see that the dependent variable is\_canceled has a considerable negative correlation with deposit\_type\_Non\_Refund and a positive correlation with deposit\_type\_Refundable.
* A mild correlation is also observed between is\_canceled and market\_segment\_Groups. Also with is\_canceled and assigned\_room\_type\_A.
* It is also good to see a high correlation between the reserved\_room\_type (A to H) and assigned\_room\_type (A to H).
* A good correlation is observed between market\_segment\_Undefined and distribution\_channel\_Undefined.
* Deposit\_type\_Non Refund and market\_segment\_Groups also have a positive correlation.

**Train Test split:**

Here we split the train and test size in the ratio 70:30 respectively and use the random state as 42. The train data and the test data sizes are computed as follows:



We also separate out the dependent and independent variables for model building and subsequently, model testing.

**Clustering:**

K – Means Clustering method:

We shall now try to cluster the customers based on similarities in attributes, patterns and shall group the same as clusters. In order to find out the best number of clusters that emerge in our data, we shall compute the WSS plot – Within Cluster Sum of Squared Errors plot:

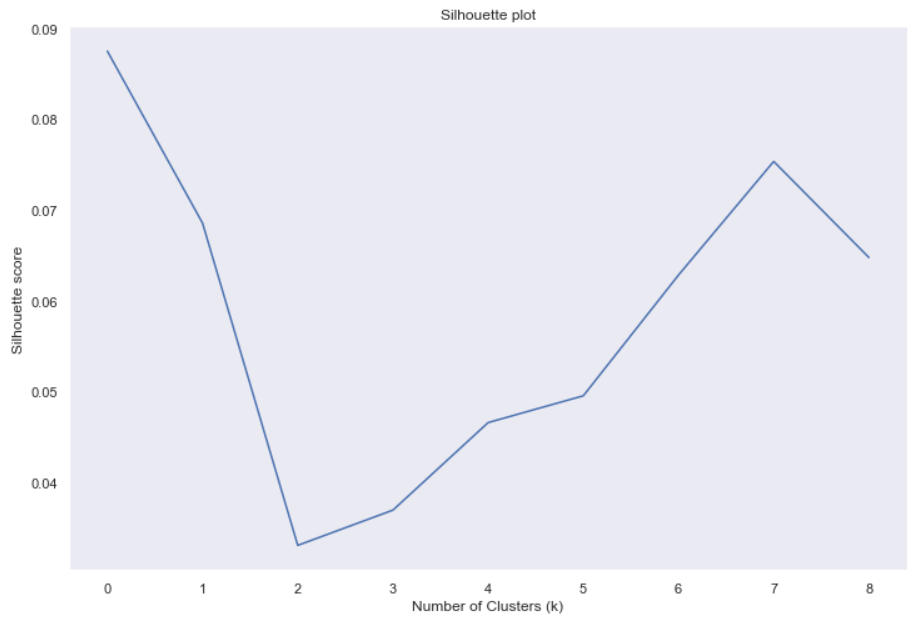


Here we can see that there is a slight bend when the number of clusters is 7. Although, there is no definite indication on the optimum number of clusters that need to be formed for our data. Hence, we shall now try the second approach to derive the optimum number of clusters for our data:

Silhouette method:

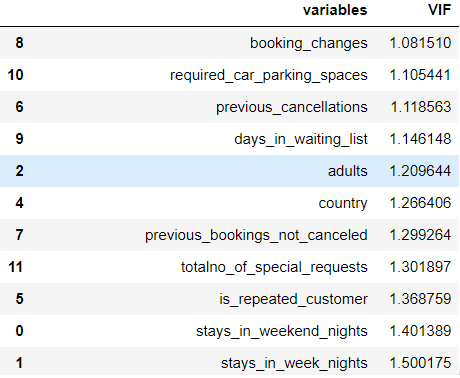
The Silhouette score computes how close an observation is to its own cluster centroid (cohesion) as opposed to the cluster centroid of the neighbouring cluster (separation). The higher the value, the better. We shall take the optimum clusters to be the point where the silhouette score is maximum.

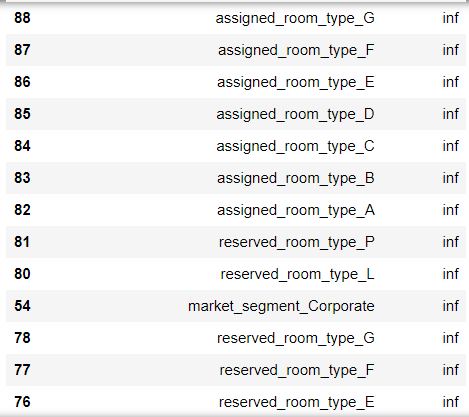
We plot the same using the silhouette score plot. We start with the number of clusters as 2 as the minimum number of clusters and use the distance metric as ‘Euclidean’.

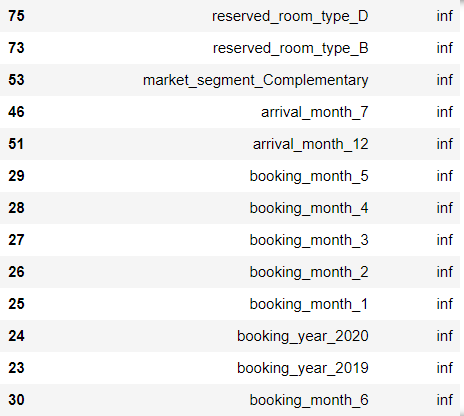


Here, it is evident that the optimum number of clusters is 7 as the silhouette score is maximum for 7 clusters.

**Calculating Variance Inflation Factor:**

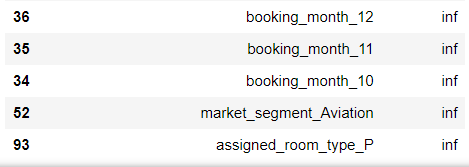
  











* It is observed that a lot of values have a VIF value as inf, meaning infinity.
* This means that there is perfect correlation between the independent variables. It is to be expected as we have one hot encoded the categorical variables. We need to drop one of the variables from this dataset that is resulting in multicollinearity.

**Outlier Treatment:**

Most of the variables in our data are categorical in nature. Even some of the numeric variables are actually categorical in nature and behaviour. It is not advisable to treat outliers in categorical data. One suggestion would be to use booking count that would be less than a certain value to consider that section as an outlier for that variable. For example, in the variable ‘country’, we have a lot of bookings for different countries where the booking count is one. That can be considered as an outlier and it is be dealt on a case-on-case basis. But it is not possible to apply outliers using the strategy that and outlier is 1.5 times the Inter-quartile range on either side of the boxplot. Hence outlier treatment has not been performed in our analysis.

**Treating Target Imbalance using ADASYN:**

Adaptive Synthetic Sampling Method is primarily used to tackle target imbalance in classification problem. In this method, synthetic data is created for the minority data samples based on their distribution. It uses the density distribution as a criterion to decide on the synthetic samples that need to be created.

In our case, the target variable is not completely imbalanced as we have observed that the proportion of distribution is 62.95 for Class 0 and 37.04 for Class 1.

However it does not satisfy the condition for calling the dataset to be balanced: 60% for Class 0 and 40% for Class 1. So, we shall apply SMOTE technique to the data and observe.

Following is the class distribution in the target variable using Counter from collections:



**Model Building:**

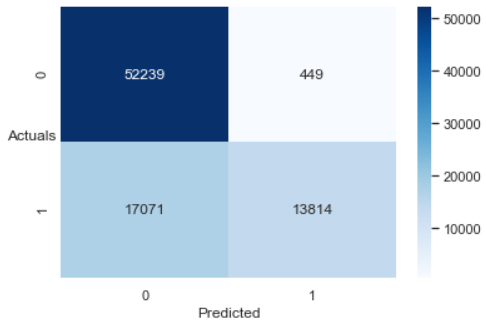
We shall first start with Random forest Classifier model. We will also include the model tuning options such as hyperparameters for this model iteration:

**Random Forest Classifier:**

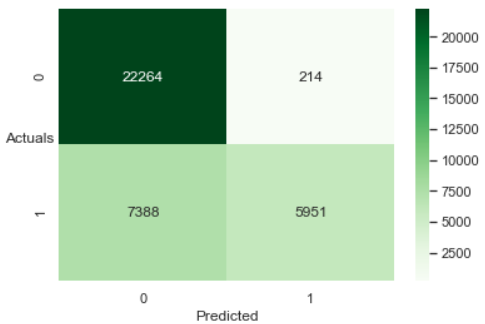
We shall use the Gridsearch cross validation for this. We choose the max\_depth and n\_estimators as the hyperparameters for this. We shall use a range between 3 and 10 for the max\_depth and a list of options as 200, 500 and 1000 for the n\_estimators. We use 3 fold cross validation and the n\_jobs as -1 which means all the processors are to be used. The values specified above have been chosen to give out the best model results after previous iterations with different values.

**Confusion Matrix:**

On Train data:



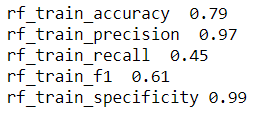
On Test data:



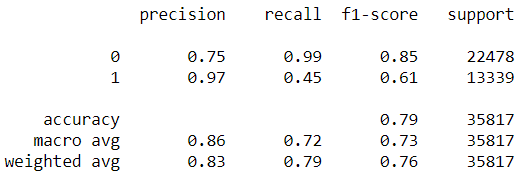
**Classification Report:**

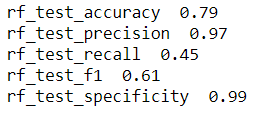
On Train data:



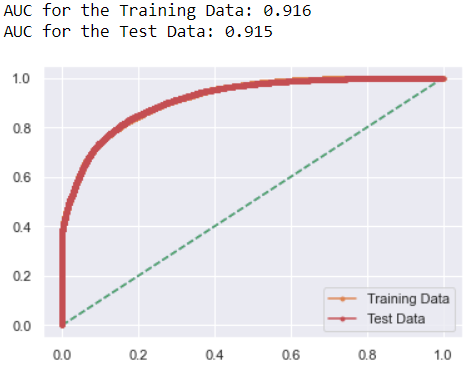


On Test data:

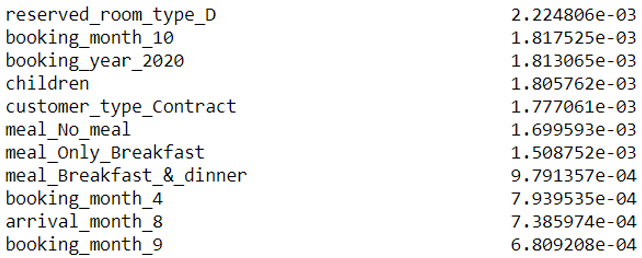
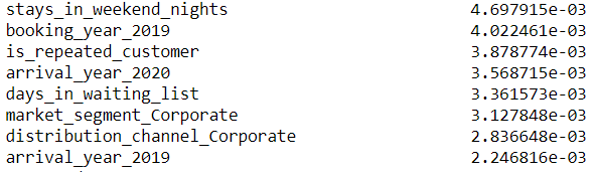
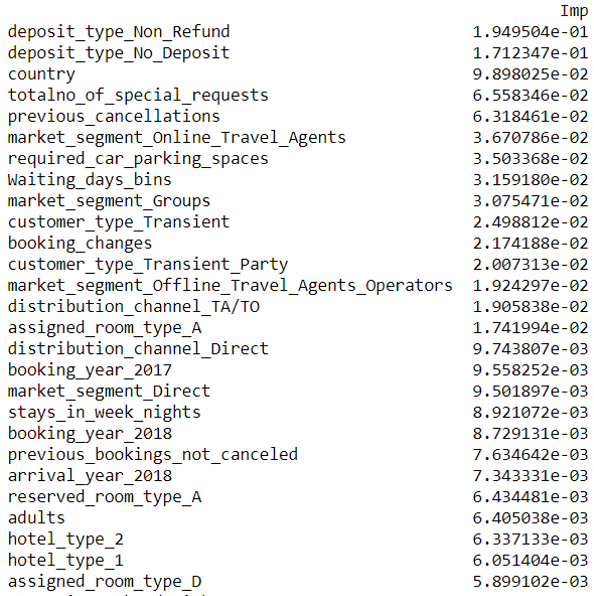


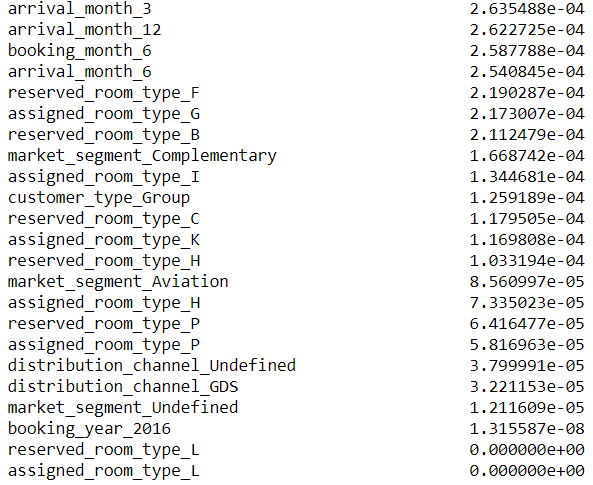
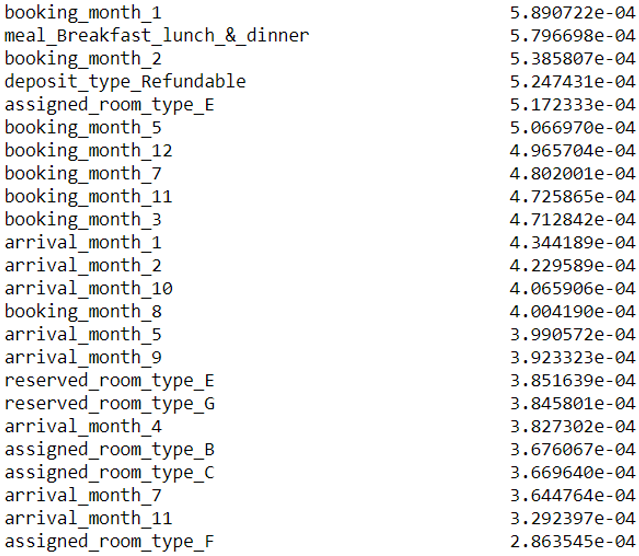


**AUC and ROC for the train and test data:**



**Variable Importance:**





**Take away from the model:**

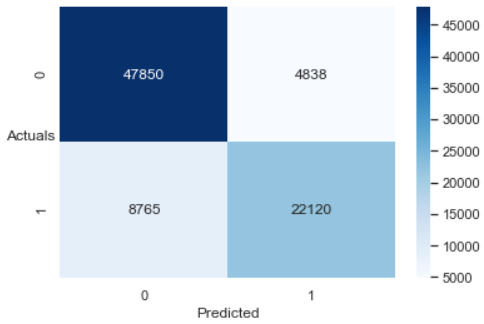
* Accuracy and recall have poor values on the train and test data.
* The model is able to generalize well on unseen data, does not overfit.
* According to this model, the top 5 variables of importance in determining the hotel booking cancellation are: deposit\_type\_Non\_Refund (0.717), deposit\_type\_No\_Deposit (0.63), country (1.33), totalno\_of\_special\_requests (0.88), previous\_cancellations (0.852).
* The AUC for the train and test data is 0.91 which is good.
* Specificity on the train and test data is really good.

**Decision Tree Classifier:**

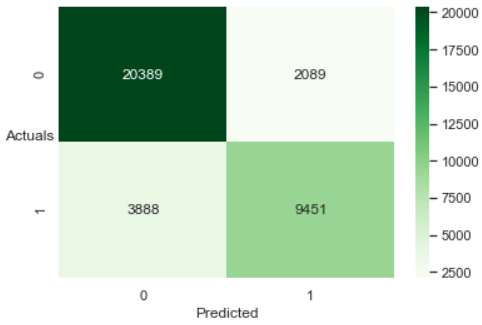
We use the criterion as gini and the random\_state as 1. In this model, we shall use the Grid search cross valiation technique with hyperparamters as max\_depth, min\_samples\_leaf and min\_samples split. We use 3 fold cross validation.

**Confusion Matrix:**

On Train data:

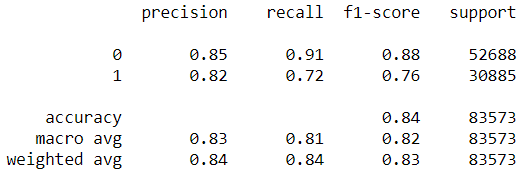


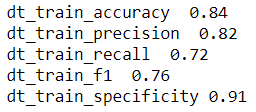
On Test data:



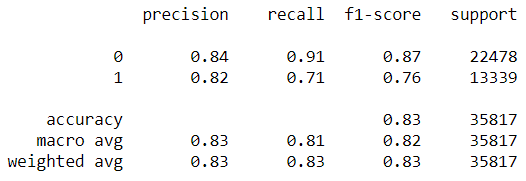
**Classification Report:**

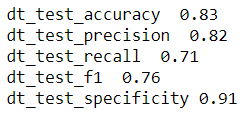
On Train data:





On Test data:





**AUC & ROC for the train and test data:**



**Take away from the model:**

* Good accuracy and precision scores on the train and test data. Recall is however better than the previous model but scope for improvement is present.
* Specificity is also good and constant for both train and test with 0.91.
* The AUC scores for train and test data are similar to the Random Forest Model.

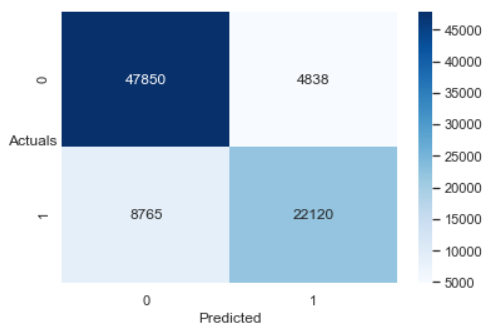
**Logistic Regression model:**

We use the solver as liblinear. The hyperparamters for the grid are penalty as L1 and L2. The ‘C’ is the inverse of regularization strength and the model has been tried out with values as 0.01, 0.1, 1, 10 and 100.

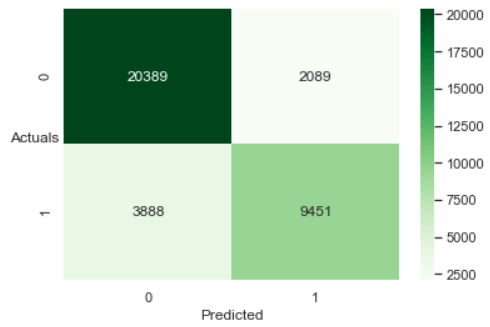
We also use 3 fold cross validation and n\_jobs as -1.

**Confusion Matrix:**

On Train data:

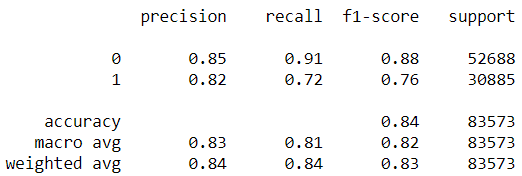


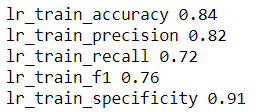
On Test data:



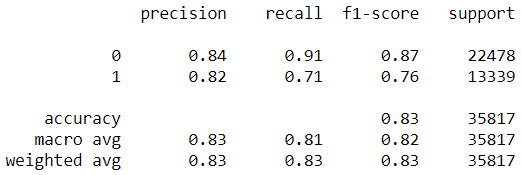
**Classification Report:**

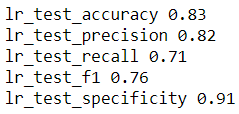
On Train data:





On Test data:





### AUC & ROC for the train and test data:

### 

**Random Forest model using Randomized Search CV:**

We set up a grid with the following parameters: n\_estimators as a list of 10, 100, 200, 500, 1000, 1200.

max\_depth as a list of None, 5, 10, 20, 30.

max\_features as a list of auto and sqrt.

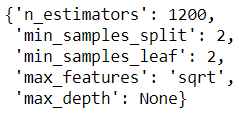
min\_samples\_split as a list of 2, 4 and 6.

min\_samples\_leaf as a list of 1, 2 and 4.

We use the random forest classifier with n\_jobs as 1 indicating the number of processors we wish to run in parallel.

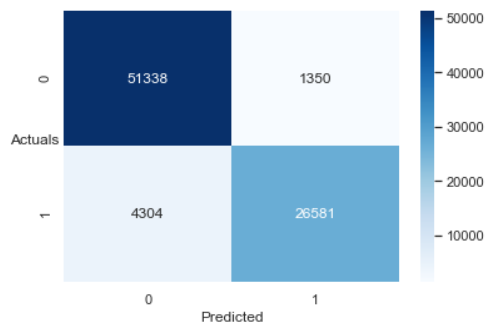
We then import the RandomizedSearchCV with 5 fold cross validation, n\_iter as 10 and verbose as 2 to see the verbosity printed for each estimator.

We get the following output as the best parameters of the model:

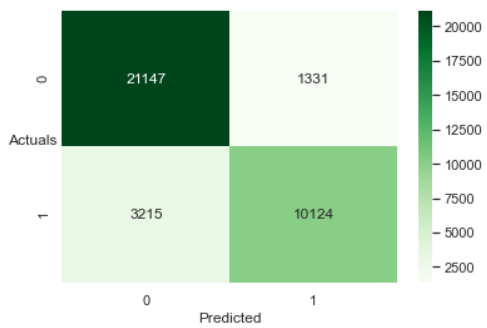


**Confusion Matrix:**

On Train data:

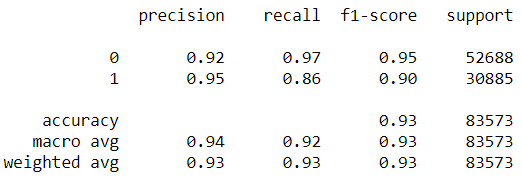


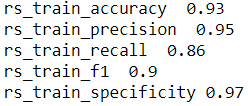
On Test data:



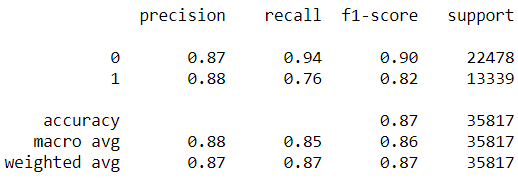
**Classification Report:**

On Train data:



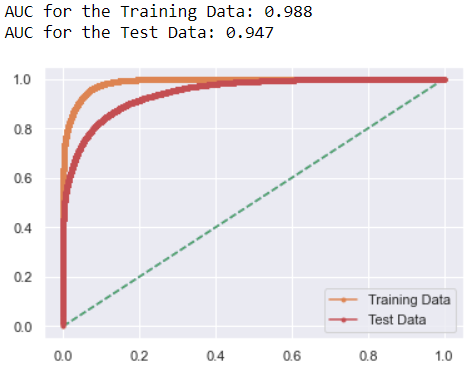


On Test data:





**AUC & ROC for the train and test data:**



**Take away from the model:**

* This model has given us the best scores in terms of accuracy, precision and recall metrics on the test data so far.
* The AUC on the Test data is also the best so far.
* The Logistic regression model did not give us good results as we would expect it to, considering that it is a binary classification problem.
* The specificity values are also very good. The difference between the train data results and the test data results is not too much so it would be a good idea to point out that this model generalizes well on unseen data.

### AdaBoost Classifier:

### We use the n\_estimators as 50. We have tried out different model iterations with increasing the number of estimators but the model results did not improve.

### We are expected to use a weak learner for the base\_model parameter as that is the whole concept of Adaptive boosting classifier. In our case, the decision tree classifier model is the weak learner and hence we will use the same as the base\_model for our AdaBoost Classifier model.

### We use the learning\_rate as 1. We have tried out different model iterations with 0.1 and 0.01 but with no improvement in the scores.

### Confusion Matrix:

### On Train data:

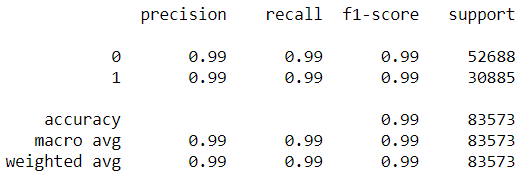
### 

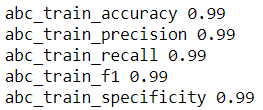
### On Test data:

### 

**Classification Report:**

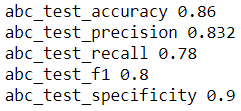
On Train data:



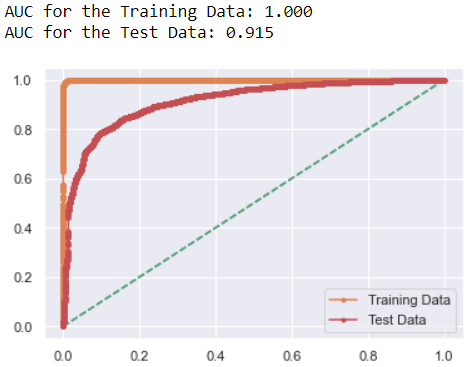


On Test data:





#### **AUC & ROC for the train and test data:**



**Take away points from the model:**

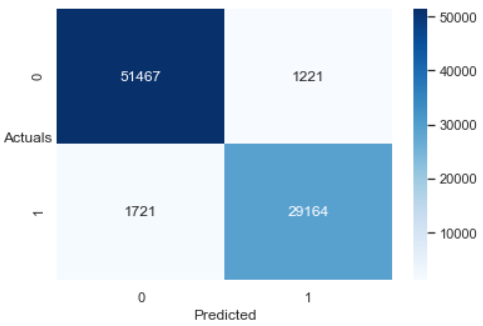
* On the train data this model has given us the best possible results on all the metrics of our interest: accuracy, precision, recall, specificity and AUC score. The AUC is 1 on the train data and is the best ever possible.
* On the test data, however this model has indeed given us precision scores less than the Random forest model with randomized search cv. The recall however is the best so far with 0.78.
* The AUC scores on the test data are also lower than the Random forest model with randomized search cv.
* We are forced to think of this model as an overfit model.

### Gradient Boosting Classifier:

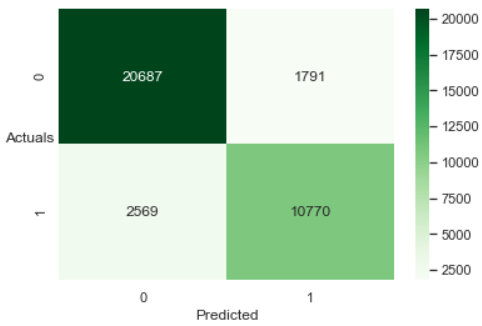
### We again use a grid search cross validation technique with n\_estimators as a list of 50, 250 and 500. Max\_depth is also taken as a list of values comprising of 3, 5, 7, 9. The learning\_rate is taken to be 0.01 and 0.1. Cross validation is taken as 3. It is to note that other values of the before said parameters have been tried out in previous model iterations and the ones specified above have resulted in churning out optimum results.

**Confusion Matrix:**

On Train data:

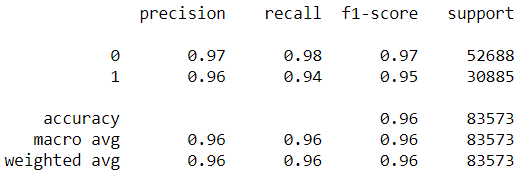


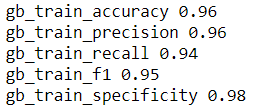
On Test data:



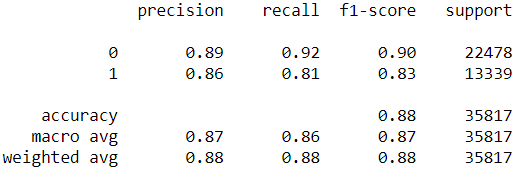
**Classification Report:**

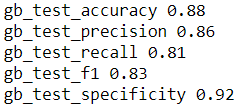
On Train data:



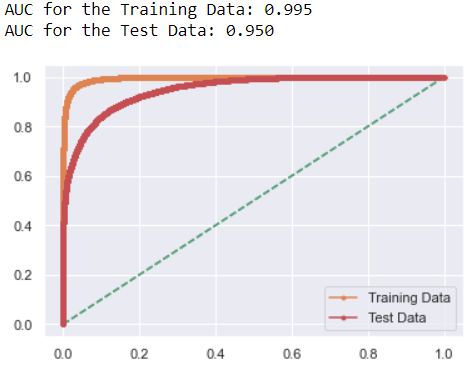


On Test data:





#### **AUC & ROC for the train and test data:**



**Take away points from this model:**

* The model results on the test data show us that this is the best model so far. The accuracy, precision and recall scores are phenomenal.
* Since this is the best model, let us understand the confusion matrix on the test data in greater detail: Out of 35817 data points, our model has correctly predicted that 20687 customers will not cancel their hotel booking. Our model predicts that 1791 out of them will cancel their hotel bookings but they end up not cancelling. This is 5%.
* 2569 customers out of 35817 customers are predicted by the model not to cancel their hotel booking but they end up cancelling the same. This amounts to 7.17%. These are False Negatives in the test data of our confusion matrix and is going to cost the hotel heavily. In our case, the false positives will not affect the hotel as much as the false negatives.
* The recall scores of the test data is 0.81 and is the highest ever obtained. The recall scores on the train data are 0.94. The specificity on the test data is 0.92 which is also very good.

### ANN Model:

### We use the following parameters in the ANN model: hidden\_layer\_sizes as 9. This has been found out in multiple model iterations that the best model accuracy results are obtained when the hidden\_layer\_size is 9. We use the activation\_function as relu, Alpha as 0.1, batch\_size as auto, solver as adam, Learning\_rate as constant. max\_iter as 300 and random\_state as 1.

### Confusion Matrix:

### On Train data:

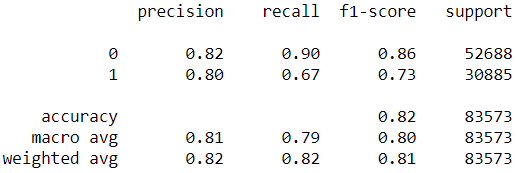
### 

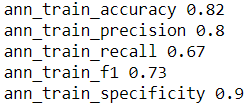
### On Test data:

### 

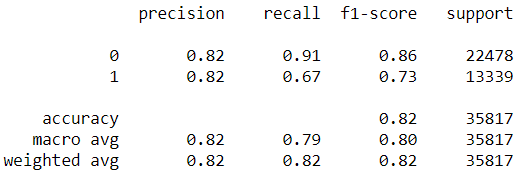
**Classification Report:**

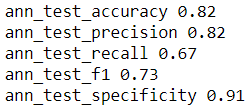
On Train data:



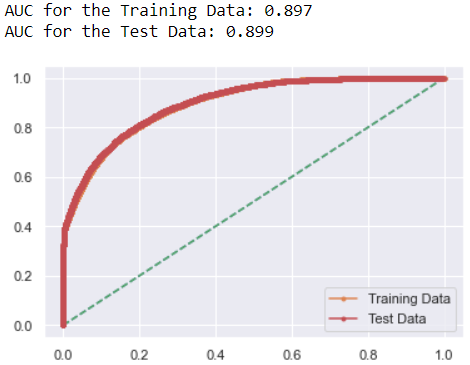


On Test data:





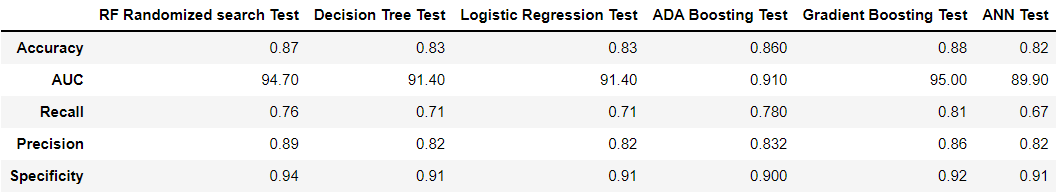
#### **AUC & ROC for the train and test data:**



### Comparison of performance metrics of all the models:

### 

Looking at only the test data of all the models:



The gradient boosting model has given us superior results. We shall recommend the same to the business.

**Insights:**

* Customers with hotel booking in 2017 have mostly cancelled.
* Customers who opt for all the three meals will mostly cancel their booking.
* Customers from Portugal have a high chance of cancellation.
* Market segments Groups and Undefined have a probability of 100% and 60% each to cancel their booking.
* Customers assigned to room types P and L will most certainly cancel their booking.
* Customers who have cancelled their booking once before have a very high chance of cancelling their booking again.
* Customers who make booking changes have a lower chance of cancellation.
* Customers with the refundable deposit type: Non Refund will most certainly cancel their booking.
* Transient type of customers have more than 45% chances of cancelling their booking.
* If waiting\_days > 100 days then greater chances of cancellation. Waiting\_days: difference between the booking date and the arrival date.
* Customers who do not have any special requests have a chance of close to 48% to cancel their booking.
* Least chance of missing the customers who are going to cancel the hotel booking when GB model is applied.
* With the GB model, Cost incurred for False Positives < Cost incurred for False Negatives.
* Some important variables in determining booking cancellation are: booking\_changes, required\_car\_parking\_spaces, previous cancellations, days in waiting list.

**Recommendations:**

* Do not allow the customer to book more than 100 days in advance.
* Impose a restriction on how early a customer can book for the hotel if they have cancelled their booking once previously.
* Encourage customers to place in special requests once they have confirmed the booking.
* Impose a high booking cancellation fee for customers from Portugal.
* Introspect on room types P and L to find the reasons for booking cancellation.
* Increase the hotel prices in the months of December, January and February of each year.
* Entice the hotel inhouse customers with options for add-ons like city tour, visit to important places at a discounted price.
* Use the customer clusters that have been generated for use of targeted marketing depending on the distinct attributes exhibited by the customers of each cluster.
* Use Gradient Boosting model with the before said parameters to predict the booking cancellation to the best.