**Actionable Insights for Hotel Reservations**

By Yeling Cai, Yunhan Zhang, Zekai Wei

12/11/2021

# **Contents**

[**Introduction**](#_bcfmk07pm5tf) **3**

[**Data Processing**](#_z4sa91yixe8b) **4**

[Initial Analysis](#_bxyppqpygiyc) 4

[Data Cleaning and Transformation](#_sbjpyn13kqir) 5

[**Visualizations**](#_pelbfe1hs9t3) **7**

[Basic Visualizations](#_uwrjp6mn5297) 7

[Mapping](#_xcgn8wkz4naa) 21

[**Models and Prediction (CART & SVM)**](#_afhethz5cyzy) **24**

[Data pre-processing](#_2pm6vn5uwfpb) 24

[Classification and Regression Tree (CART)](#_nnn7gubnnzgk) 25

[Support Vector Machines (SVM)](#_bzbxf9wzerqs) 28

[Model comparison](#_fwxk329vdqwi) 29

[**Associative Rules Mining**](#_lxcv8byl29eo) **31**

[**Recommendations**](#_eaa47nfawtof) **33**

[**Appendix**](#_sjadl1glm0aa) **36**

# **Introduction**

Our team’s goal is to understand some key drivers for why people cancel hotel reservations and better predict who will cancel. This report documents data processing, descriptive statistics, exploratory analysis, data visualization, machine learning, and some substantive recommendations for the hotel based on the data science results.

In this report, we used histogram, bar chart, box plot and mapping methods to visualize the data we have. Based on the results of visualization, we cleaned the data and grabbed essential variables for modeling. We used the Classification and Regression Trees model (CART) and Support Vector Machines model (SVM) to train the training data set and predict the cancelation status of the testing data set. We used the accuracy to judge which model is more suitable for the situation and can be used to predict cancellation status in the future. We also utilized Associative Rules Mining to recognize some vital factors which will affect the cancelation of guests. Finally, we put forward some actionable insights and recommendations for managers.

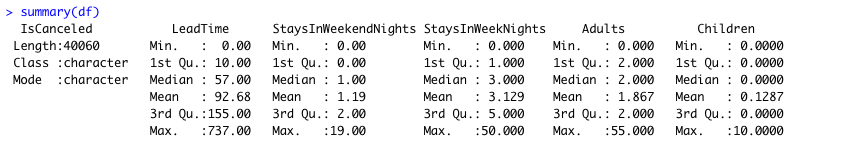
The CSV-file provided was the only source of our analyses. It had 40,060 observations and 20 variables. In this report, the running environment of coding was R version 4.1.1.

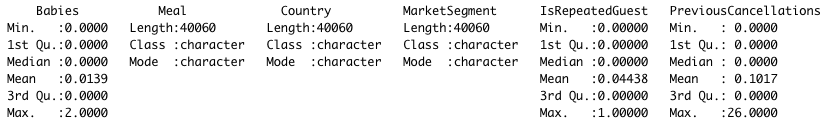
# **Data Processing**

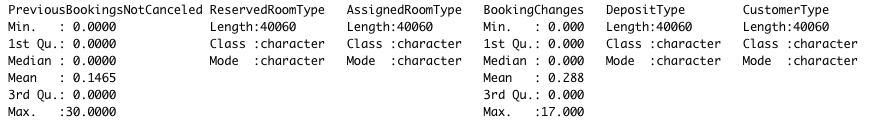
## **Initial Analysis**

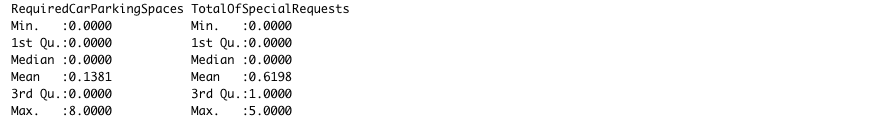
Before we do the analysis, the summary of the raw dataset helps us gain some insights about descriptive statistics and the structure of the entire data set.

Summary of the data set



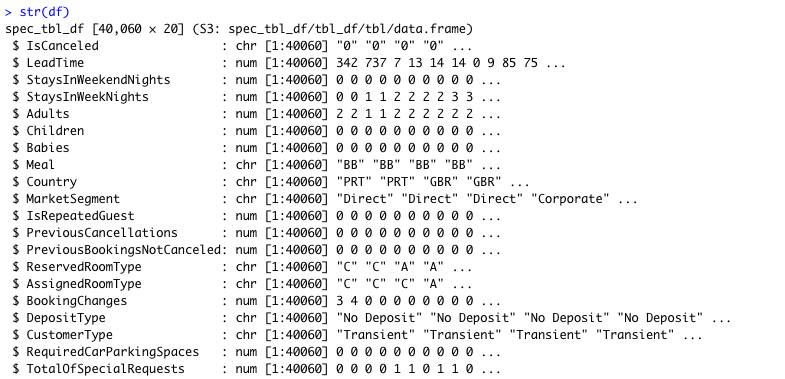






We can see the minimum value, the first quartile, the median, the mean, the third quartile and the maximum value of each numeric variable. Also, we can see the length, the class and the mode of each variable of character type.

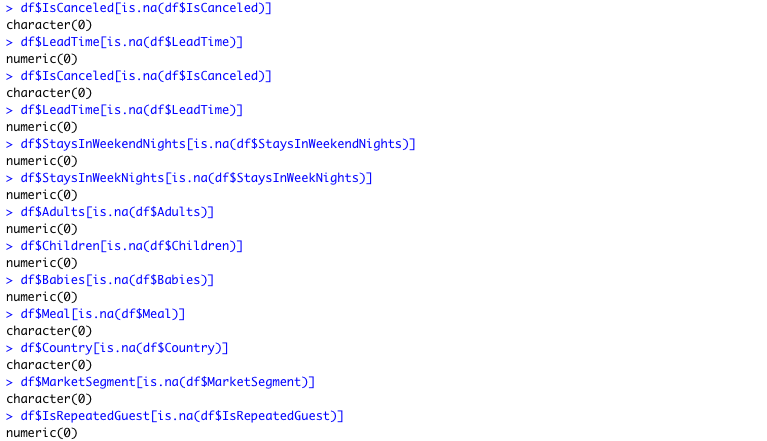
Structure of the data set

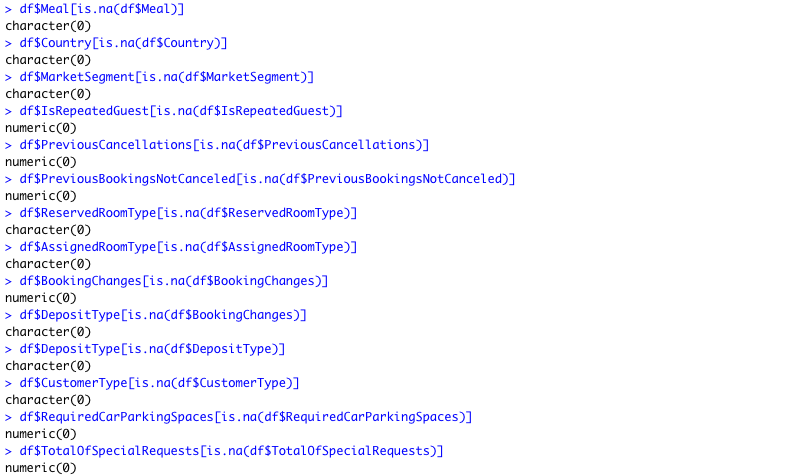


We can see the type and some sample data of each variable.

## **Data Cleaning and Transformation**

Using the “is.null” function in R, we found there is no null value in the data set. Even though we found some null values in the “country” variable in our raw CSV-file, they won’t influence our visualizing result and modeling. Therefore, we didn’t remove these null values.





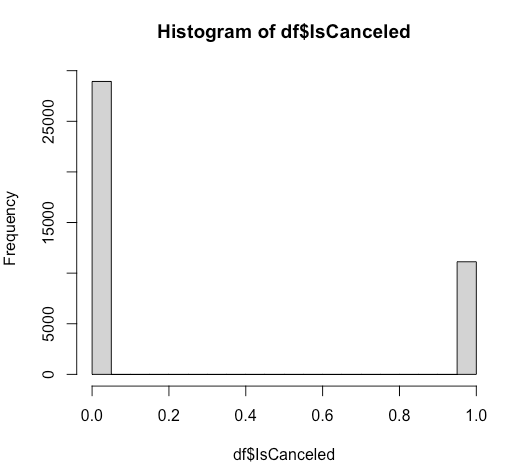
We divided the data set into two subset using the “IsCanceled” variable, to explore the data distribution in cancelled and not-cancelled situations.

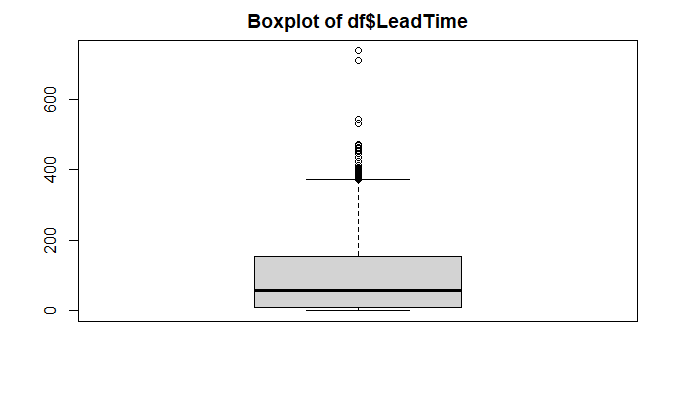


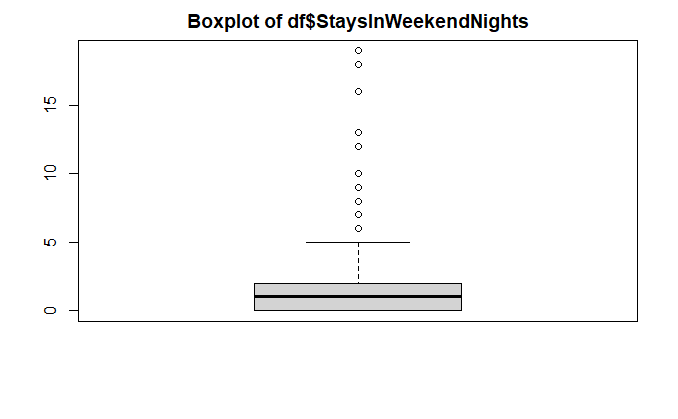
# **Visualizations**

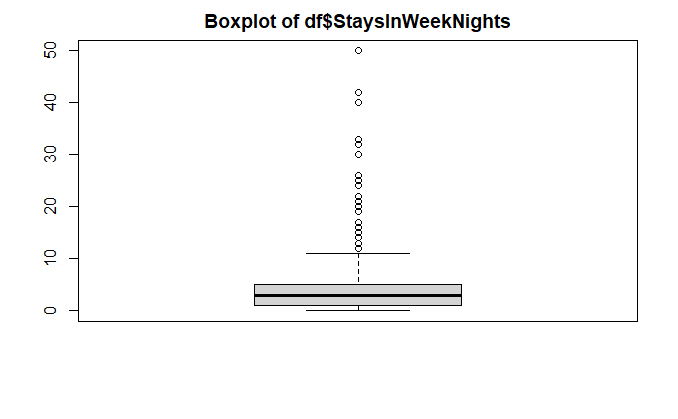
## **Basic Visualizations**

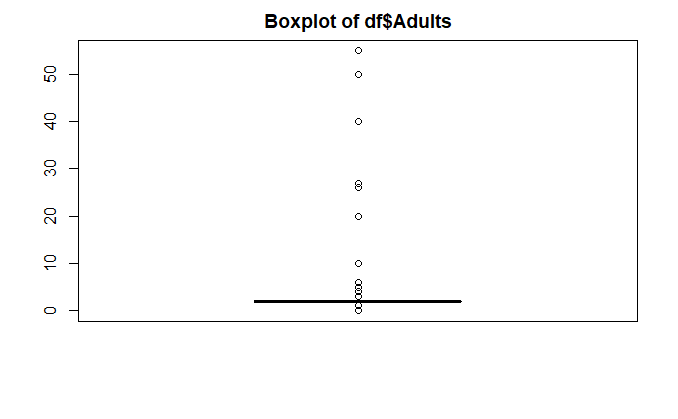
This histogram presented the frequency of cancelled records and not cancelled records.

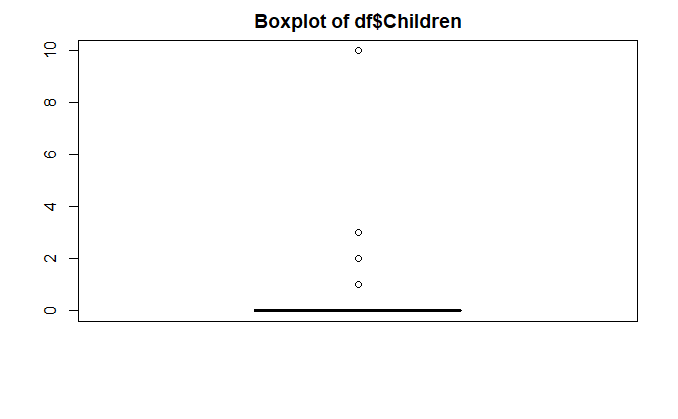


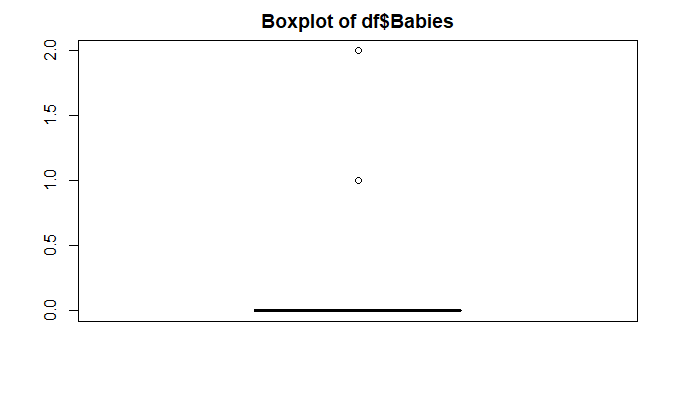
This box plot presented the distribution of lead time.

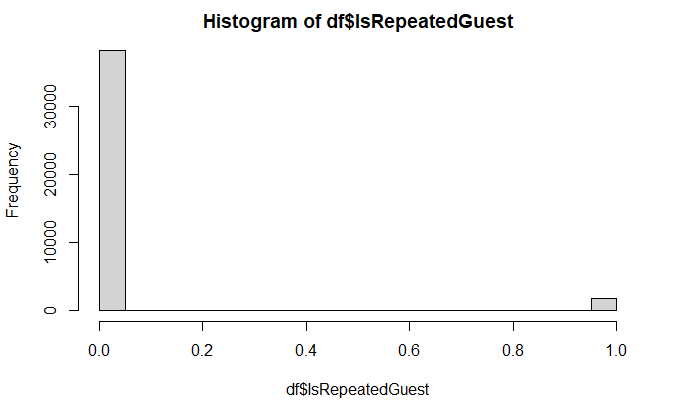
This box plot presented the distribution of “StayInWeekendNights”.

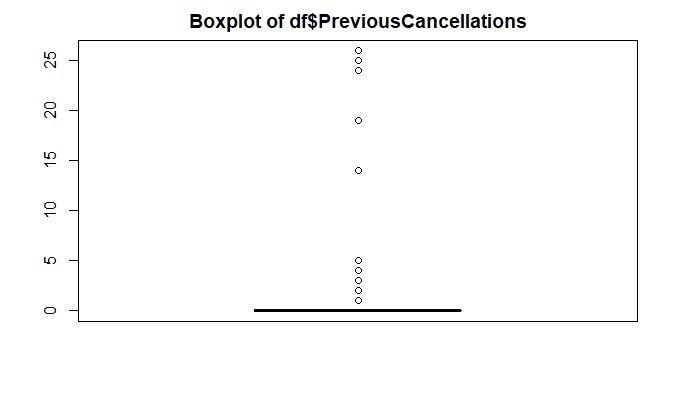
This box plot presented the distribution of “StayInWeekNights”.

This box plot presented the distribution of the number of adults.

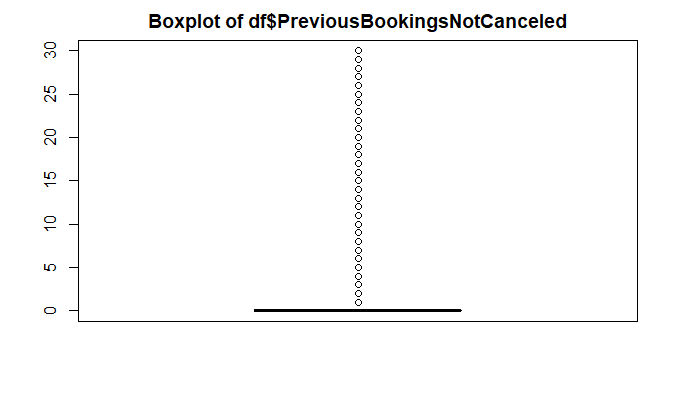
This box plot presented the distribution of the number of children.

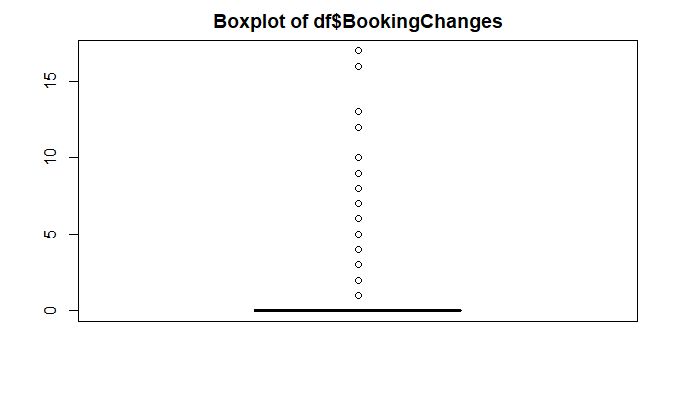
This box plot presented the distribution of the number of babies.

This histogram presented the frequency of repeated guests and not repeated guests.

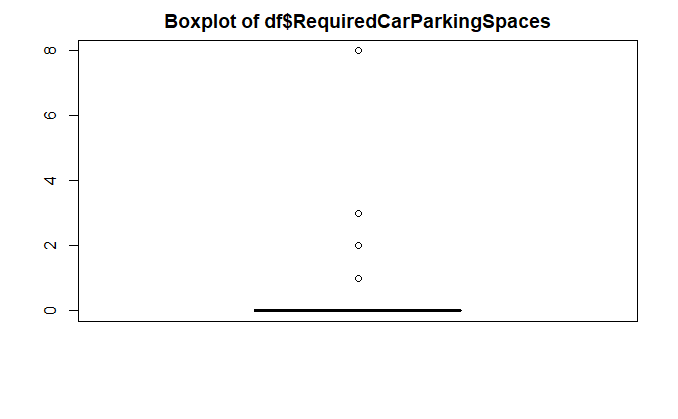
This box plot presented the distribution of the number of previous cancellations.

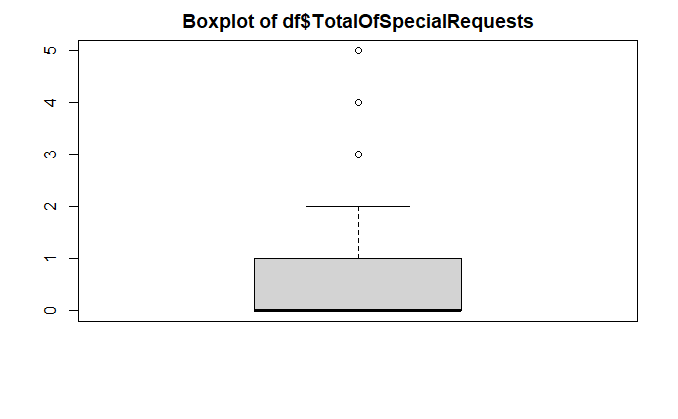
This box plot presented the distribution of the number of previous bookings that are not cancelled.



This box plot presented the distribution of the number of booking changes.

This box plot presented the distribution of the number of required car parking spaces.

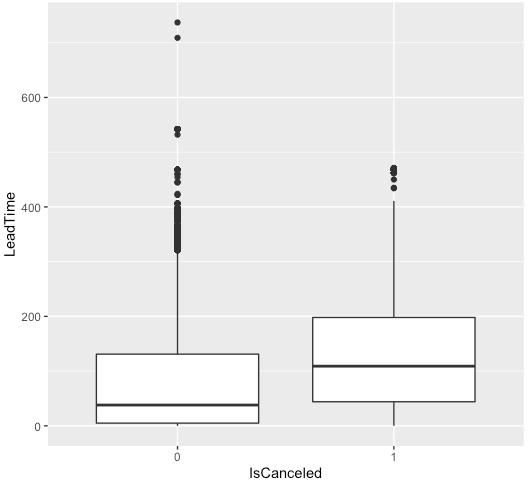


This box plot presented the distribution of the number of total special requests.

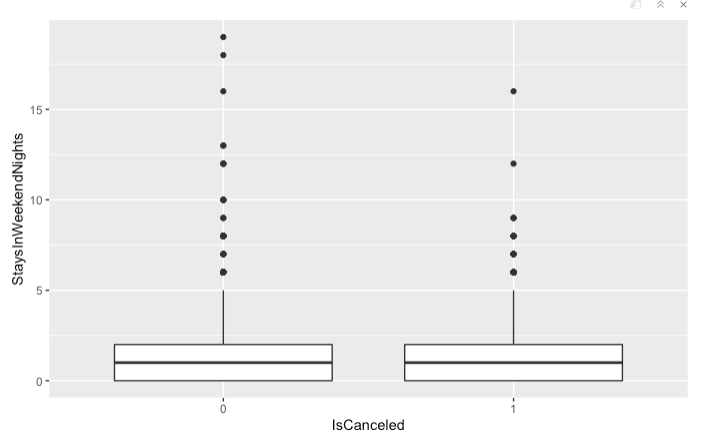
To show the data distribution in two types of situations, we transferred the data type of “IsCanceled” variable from numeric to character.



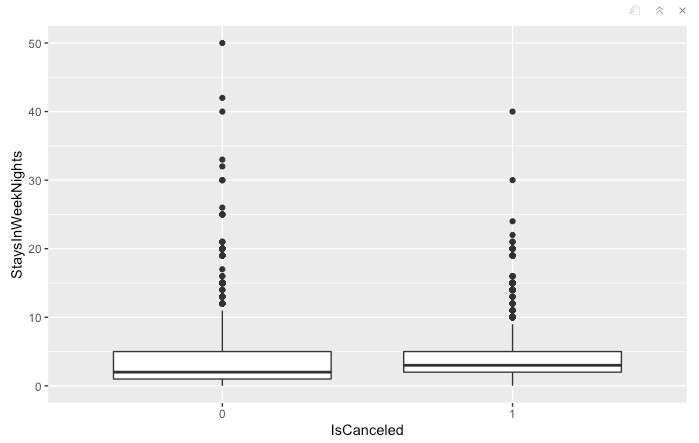
The box plot compared the distributions of lead time for cancelled records and not cancelled records.



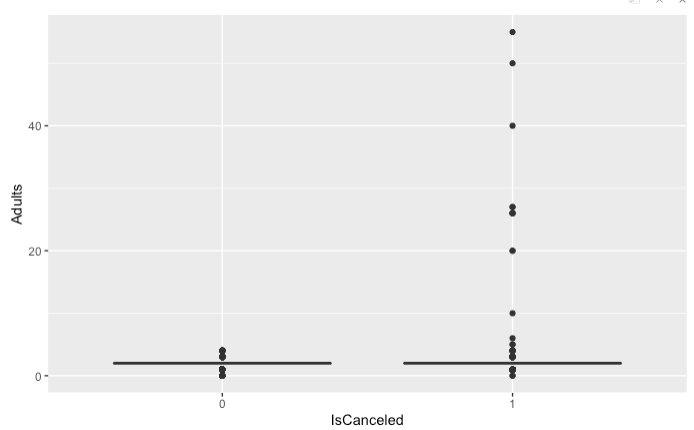
The box plot compared the distributions of ”StayInWeekendNights” for cancelled records and not cancelled records.



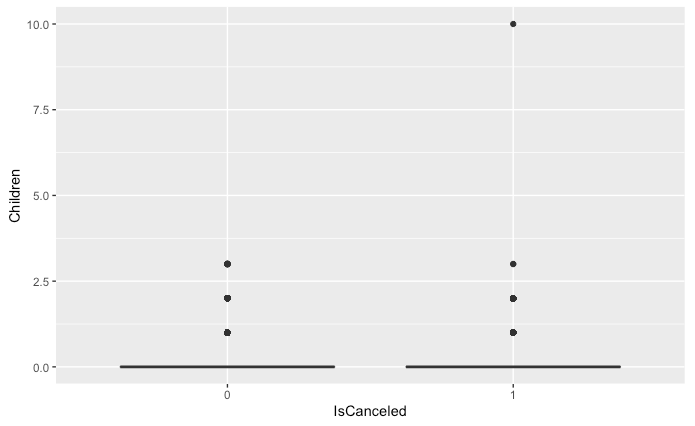
The box plot compared the distributions of “StayInWeekNights” for cancelled records and not cancelled records.



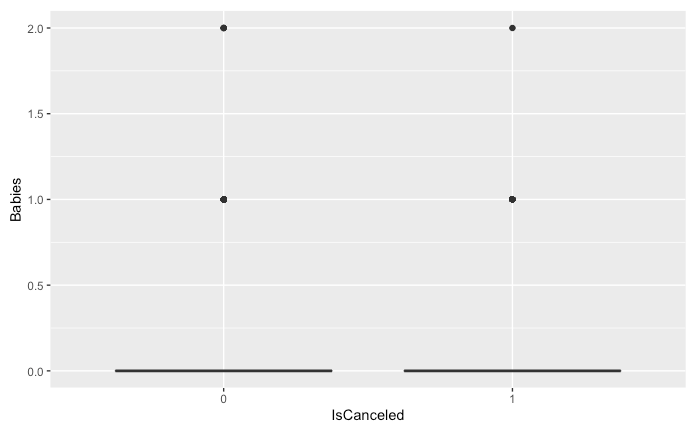
The box plot compared the distributions of the number of adults for cancelled records and not cancelled records.



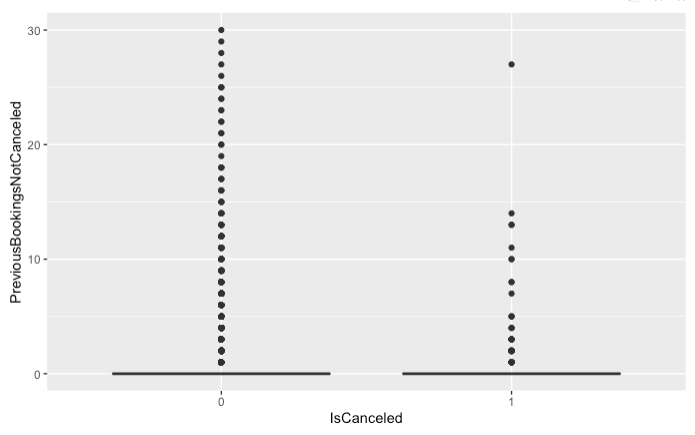
The box plot compared the distributions of the number of children for cancelled records and not cancelled records.



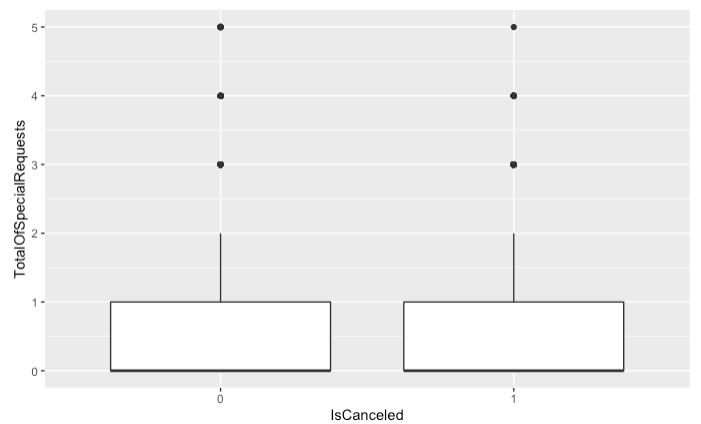
The box plot compared the distributions of the number of babies for cancelled records and not cancelled records.



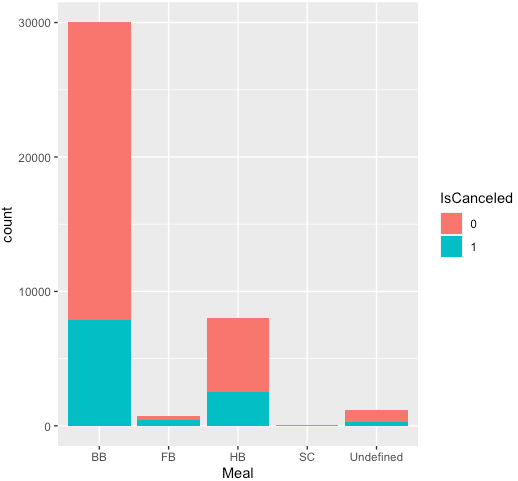
The box plot compared the distributions of the number of previous bookings that are not cancelled for cancelled records and not cancelled records.



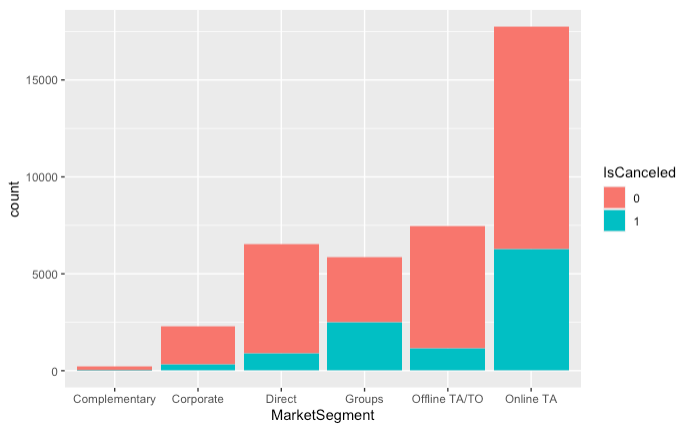
The box plot compared the distributions of the number of total special requests for cancelled records and not cancelled records.



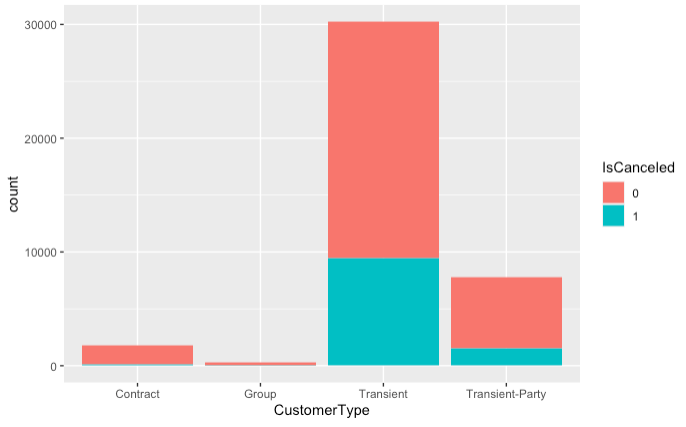
The bar chart compared the distributions of cancelled records and not cancelled records for different meal plans.



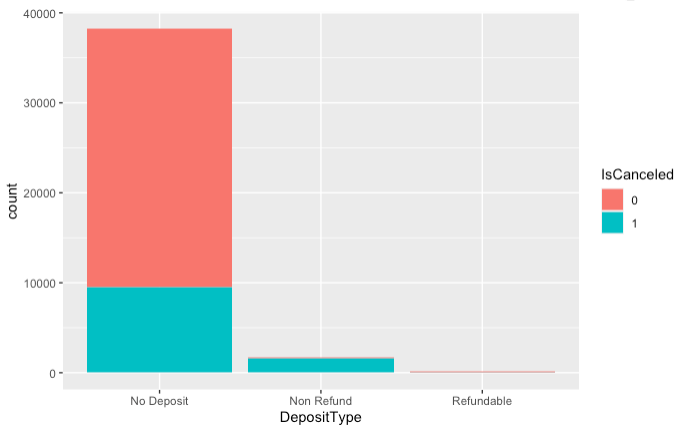
The bar chart compared the distributions of cancelled records and not cancelled records for different market segments.



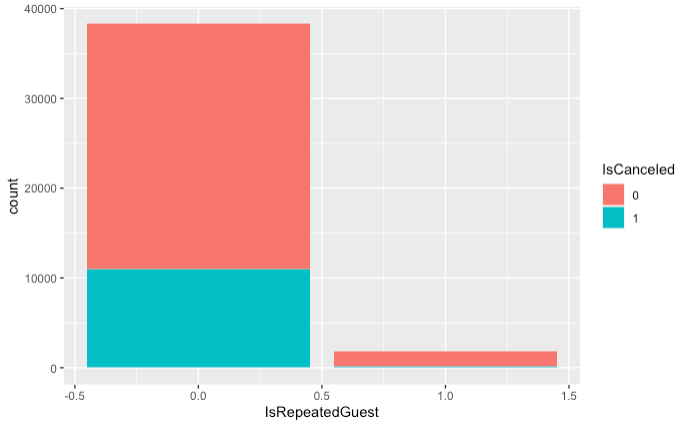
The bar chart compared the distributions of cancelled records and not cancelled records for different customer types.



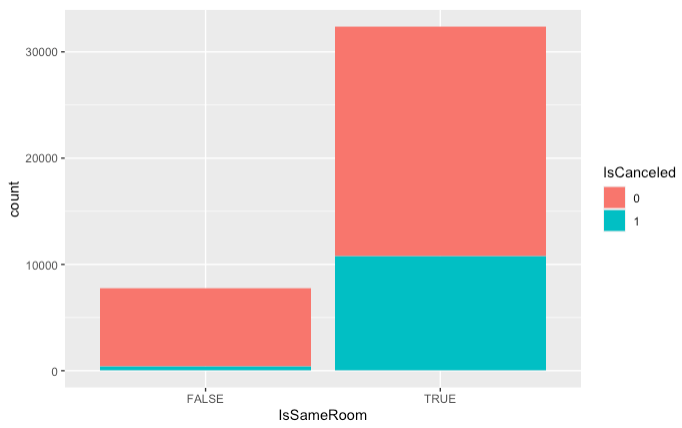
The bar chart compared the distributions of cancelled records and not cancelled records for different deposit types.



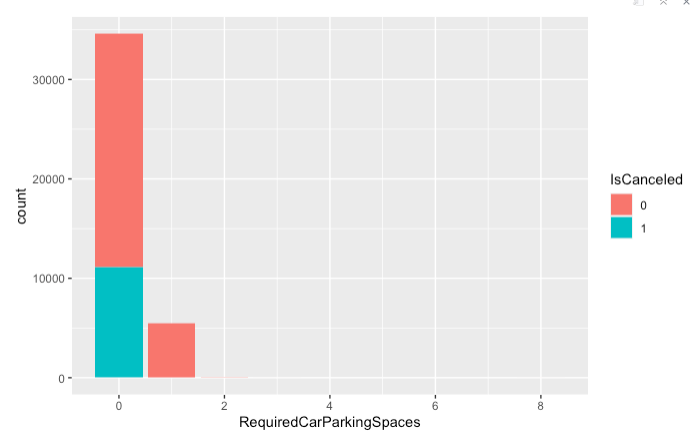
The bar chart compared the distributions of cancelled records and not cancelled records for repeated guests and not repeated guests.



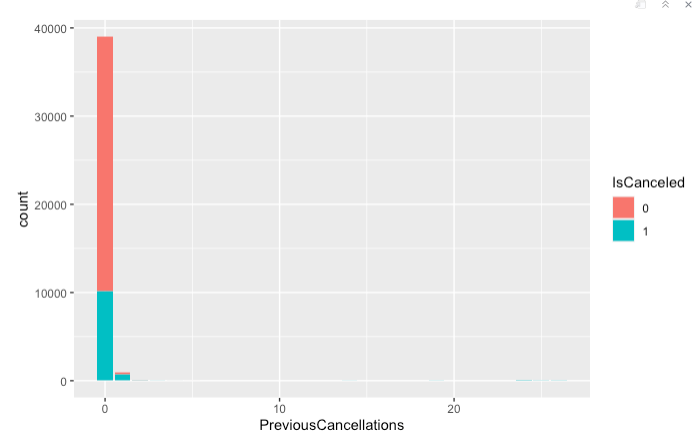
The bar chart compared the distributions of cancelled records and not cancelled records for the coincidence of reserved room and assigned room.



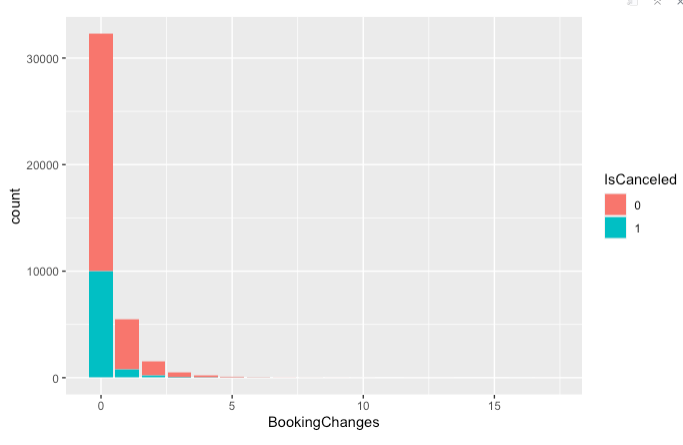
The bar chart compared the distributions of cancelled records and not cancelled records for different numbers of required parking spaces.



The bar chart compared the distributions of cancelled records and not cancelled records for previous number of cancellations.

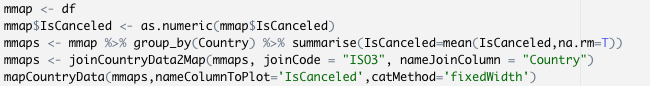


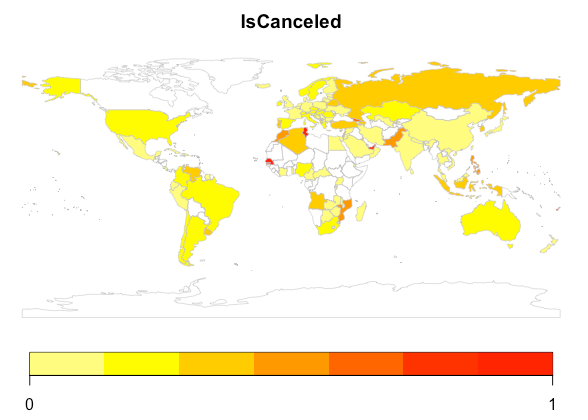
The bar chart compared the distributions of cancelled records and not cancelled records for different numbers of booking changes.



## **Mapping**

To visualize the cancellation data in different countries, we created a color gradient world map.

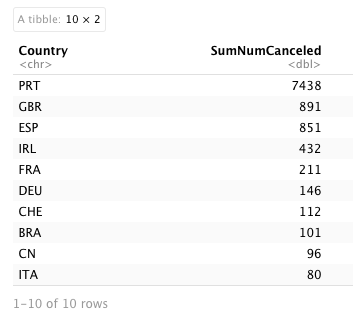




This map is divided by the geographical location of the country. The average cancellation rate for each country is indicated by color shades on the geographic location. In the world map, the darker the color meant the higher the cancellation rate. The cancellation rate in Africa was higher than other continents, even if the data set did not record the data in many countries in Africa.

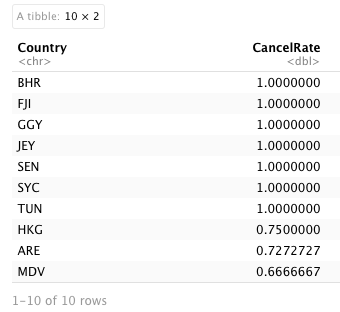
We can also calculate and show the top 10 number of cancellations and the cancelation rate of each country.





The "SumNumCanceled" refers to the total number of canceled reservations from this country. The larger the number, the more cancellations from this country. Portugal had the most number of cancellation records, 7438. The United Kingdom of Great Britain and Northern Ireland had the second highest number of cancellation records. Spain had the third highest number. Therefore, most cancellations happened in Europe. However, the higher number of cancellations did not mean the higher cancellation rate. Because the country which had a higher number of cancellations may just have more and more records of booking. Therefore, the number of cancellation records was big.





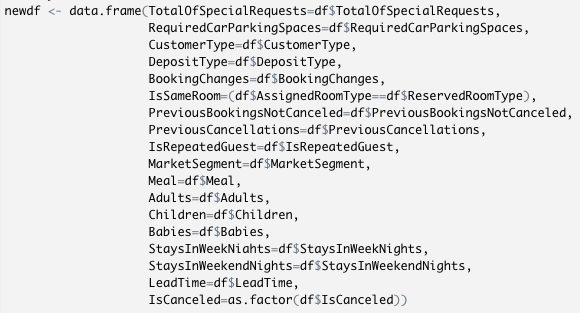
The "CancelRate" is the percentage of reservations from this country that are canceled. When the cancellation rate is 1, it means that all bookings from this country are canceled. The lower the number, the lower the cancellation rate for reservations from that country.

From the above data, we can see that the results of the map visualization can only be used as a reference. In some countries, the average cancellation rate is 0 or 1 due to missing or insufficient data (e.g., only one data for "BHR"), which also shows the extreme nature of the results due to data limitations. Therefore, further data are needed to study and analyze the impact of geographical factors on cancellation rates.

# **Models and Prediction (CART & SVM)**

## **Data pre-processing**

Before creating our models, we had to transform the type of “IsCanceled” from numeric to factor. We also created a new variable named “IsSameRoom” to reflect if the reserved room type is coincident with the assigned room type. Therefore, in this part, we created a new data frame named “newdf”.



We also divided the data set randomly into two parts, a training set for training models and a testing set for evaluating the accuracy of prediction.



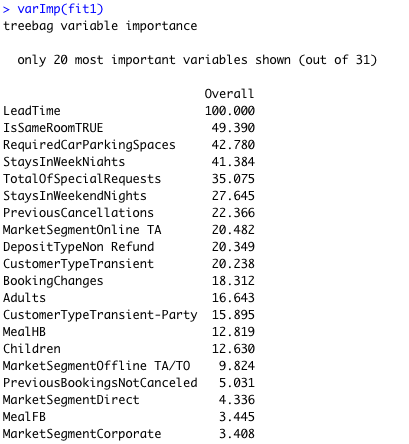
## **Classification and Regression Tree (CART)**

First, we used the CART method to create a model and predict the cancellation of guests.



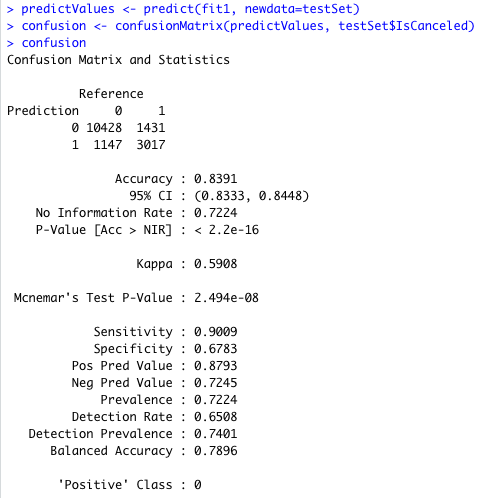
We wanted to figure out the influence of other variables on the “IsCanceled” variable. In this CART medel, we used the “treebag” method. Also, we took the precaution of centering and scaling to put every input variable on the same scale.

We also used the “varImp” function to identify important variables that could influence the cancelation actions of guests.



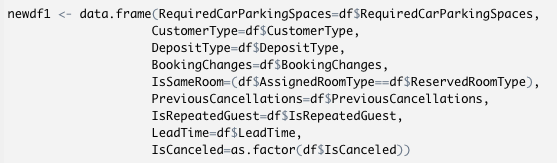
In this table, we can see “LeadTime”, “IsSameRoom” and “RequiredCarParkingSpaces” are the top three variables that affect the cancellation rate.

We used this model to predict cancelation status of the testing data set. We can see the accuracy is around 84%.



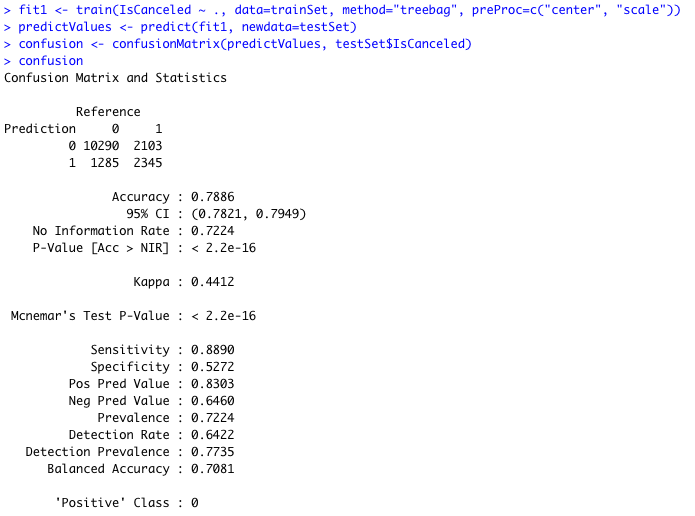
Also, we wanted to see if there was a risk of overfitting. Therefore, we removed several unnecessary variables in “df” and created a new data frame named “newdf1”. Our eliminating action was based on the result of visualization and “varImp” function. We removed following variables in the new data frame, “TotalOfSpecialRequests”, “PreviousBookingsNotCanceled”, “MarketSegment”, “Meal”, “Adults”, “Children”, “Babies”, “StaysInWeekNights” and “​​StaysInWeekendNights”.

If the model we created was overfitting, then the accuracy of the new prediction can be higher as we eliminated several unimportant variables.



We used the same model and found out the accuracy had decreased to around 79%.



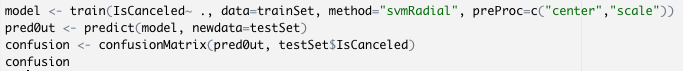


Also, we tried to eliminate less variables. And the accuracies of prediction were all below 84%. Therefore, we could say, the original CART model was an available model to predict the cancellation of guests. It has little risk of overfitting.

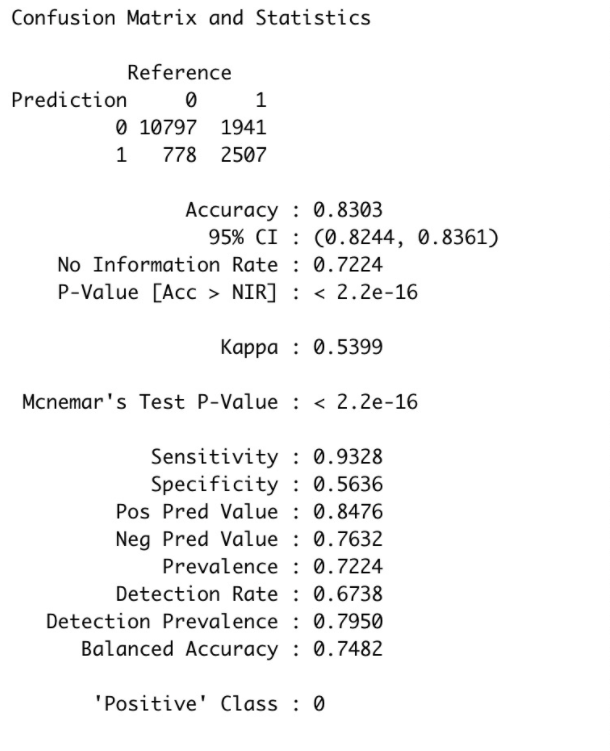
## **Support Vector Machines (SVM)**

Support Vector Machines (SVM) is another supervised machine learning algorithm. It is a kind of generalized linear classifier for binary classification of data. Its decision boundary is the maximum margin hyperplane of learning samples. SVM can be used for nonlinear classification by kernel method.

Therefore, for the second model, we used SVM to predict the cancellation of guests.



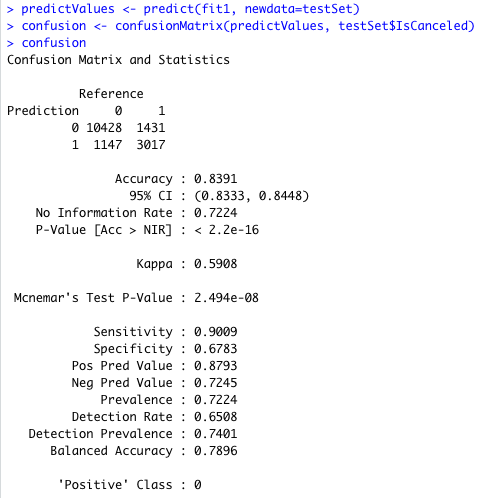
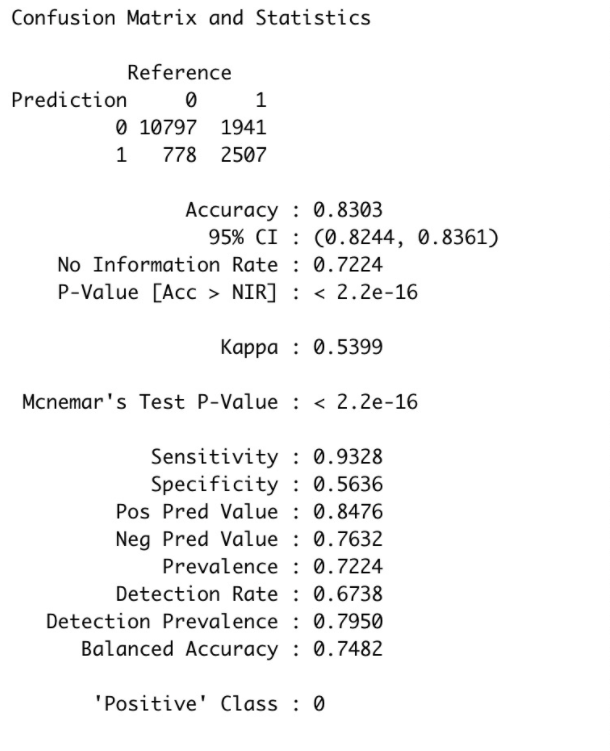
In this model, we used radial basis function as kernel function. We used the training data set to build the model. And we used the model created to predict the cancellation. We also created a confusion matrix, which indicated that the accuracy of the prediction was around 83%. It was a little lower than the CART model.



## **Model comparison**

CART and SVM are both supervised machine learning algorithms. According to the results of training and predicting, we wanted to compare the effect of these two models. We wanted to know which model was more suitable for this case. If managers of the hotel have basic variables in the future, they will have an opportunity to use the model to predict the cancellations.

CART: SVM:

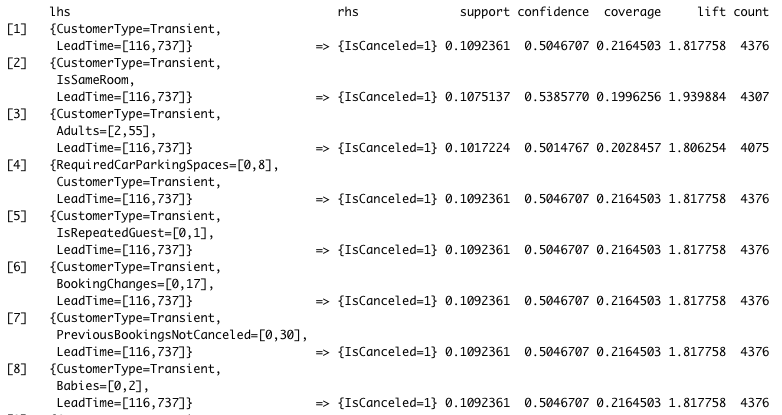
According to these two confusion matrices of CART and SVM models, we could compare the accuracies of them. The accuracy of CART was 83.91%. The accuracy of SVM was 83.03%. The accuracy of CART was better than the accuracy of SVM. The 95% CI of CART was also better than SVM. The no information rates are the same. It was obvious. The P-value of each model was very small, which indicated that the probability that the difference between samples caused by sampling error was really small. Therefore, according to the confusion matrices, the CART model was better to predict the cancellation.

# **Associative Rules Mining**

Associative Rules Mining is one kind of unsupervised machine learning algorithms. It can be used to discover meaningful connections hidden in large data sets. The discovered associations can be represented in the form of association rules or frequent itemsets. In this case, we used Associative Rules Mining to figure out the essential variables that could have a huge impact on the cancellation actions of employees.



In this model, we set the support value at 0.1 and confidence value at 0.5. Because we want to show the influence of other variables to “IsCanceled”, we set “lhs” as default and “rhs” as “IsCanceled=1”. Following are part of the results of rules mining.



Support is the proportion of times that the union of LHS and RHS occurs versus the total number of transactions. It represents the proportion of the total sample in which both events occurred simultaneously.

Confidence is a conditional probability. It is the probability of finding the RHS in

transactions that also contain the LHS.

Lift is the ratio of support for LHS:RHS together versus if they were independent. The lift reflects the correlation between RHS and LHS in the association rules. Lift greater than 1 and higher indicates higher positive correlation. Lift less than 1 and lower indicates higher negative correlation. Lift equals to 1 indicates no correlation between two elements.

The first rule reflects the effect from “Customer Type”and “LeadTime” to the cancellation actions. If the customer is a transient, and the lead time is between 116 and 737, the support can be around 10.9%; the confidence can be around 50.5%; and the lift can be around 1.818. The lift is greater than 1, which reflects the positive correlations between “CustomerType”, “LeadTime”, and “IsCanceled”.

Based on the top 8 rules, we can indicate that the type of customer, lead time, and the consistency between reserved room and assigned room are important variables that can affect the cancelation actions of guests.

However, we cannot use associative rule mining to predict because it is an unsupervised machine learning algorithm. It can only be used to classify clusters of data such that the highest similarity is within the class. It can reflect the influences between variables instead of predicting.

# **Recommendations**

In this dataset, there are a total of 40,060 observations, of which the total number of cancellations is 11,122(27.76%), and the total number of non-cancellations is 28,938(72.24%). The histogram of the number of cancellations shows that the number of cancellations is approximately one-third of the number of non-cancellations and that the number of cancellations represents more than one-quarter of the total number of bookings. This percentage is relatively large among the peers. Therefore, the hotel needs to take some measures to reduce the cancellation rate of customers. In this part, we propose several variables that we think the managers of the hotel should pay more attention to and take some actions.

**- LeadTime**

From the boxplot of the variables "IsCanceled" and "LeadTime", we can see that the median and the distribution interval of the lead time of canceled reservations are larger than those of non-canceled reservations. This means that the greater the number of days between the entry date and the arrival date of a customer's reservation, the more likely he/she is to cancel the reservation. Even though there are several outliers in the visualization plot, which indicates there are several cancellation records that the lead times of them are long, these records are just some exceptions.

Also, in the CART model, through the “varImp” function, we can figure out “LeadTime” is the most important variable that can influence cancellation.

In addition, according to the result of associative rules mining, the top 8 rules all contain the influence from “LeadTime” to cancellation. If the “LeadTime” is larger than 116, the guests will be more likely to cancel their reservations.

In summary, if a guest books the hotel too much time ahead of arrival, he or she will be more likely to cancel the reservations. Therefore, the managers should pay more attention to the records which have a really long lead time. The manager can reduce the price of the hotel room booked close to the entering date, so that more guests will book the hotel closer to the entering date because of the low price.

**- IsSameRoom**

According to the bar chart, the cancellation rate is much lower when the reserved rooms and assigned rooms are different from each other.

According to the result of the “varImp” function in the CART model, the variable of “IsSameRoom” is the second important variable that influences the cancellation. Besides, “IsSameRoom” also appears in the second rule. Consequently, if the reserved room type and the assigned room type are different (or let’s say the hotel approved their requests of room changing) then they will not cancel their appointments.

Therefore, the hotel needs to try to satisfy customers’ needs as much as they can, such as room changing requests. The hotel could also avoid the possibility of overselling by using a more accurate forecasting model. Even if such things can't be avoided, hotels need to put customer experience first and make customers have a better stay experience through proper financial compensation, better service, etc., so as to reduce the cancellation rate.

**- RequiredCarParkingSpaces**

From the CART model, we can figure out that “RequiredCarParkingSpaces” is the third important variable that influences the cancellation rate. The cancellation rate of customers with parking requirements is lower than that of customers without parking requirements.

Therefore, in order to operate better, the hotel should first ensure that the supply of parking spaces is sufficient, and the amount of parking space demand can be predicted by the previous customers' demand for parking spaces.

**- CustomerType**

Even though the variable of “CustomerType” does not show much importance in the CART model, in the Associative Roles Mining model, it plays a significant role in cancellation.

Also, in the bar chart, it is obvious that transient customers are the most likely group to cancel their cancellations.

Therefore, the managers could take actions to transfer transient customers to other types of customers. Decreasing the rate of cancellations in transient customer type is also helpful. The hotel can cut down the price of contract customers and group customers to attract transient customers to become other types of customers. Besides, the hotel should improve the services continuously, so that the hotel will have a very good reputation. As a result, the cancellation can decrease.

# **Appendix**

Attached is a full list of every library used:

library(tidyverse)

library(rworldmap)

library(caret)

library(rpart)

library(rpart.plot)

library(kernlab)

library(arules)

library(arulesViz)

The data file is located at:

<https://intro-datascience.s3.us-east-2.amazonaws.com/Resort01.csv>

*\* All data was considered for exploratory analysis, proper data ethics were maintained for the insights portion of our report.*