Report of Flight Price Predicton

071970225 Zhoukang

2022-11-19

1 Introduction

1.1 The The information about Dataset "Flight Price Prediction"

The information below comes from the author of this dataset,you can check in the website:https://www.kaggle.com/datasets/jillanisofttech/flight-price-prediction-dataset. The objective of the study is to analyze the flight booking dataset obtained from the "Ease My Trip" website and to conduct various statistical hypothesis tests in order to get meaningful information from it. The 'Linear Regression statistical algorithm would be used to train the dataset and predict a continuous target variable. 'Easemytrip' is an internet platform for booking flight tickets, and hence a platform that potential passengers use to buy tickets. A thorough study of the data will aid in the discovery of valuable insights that will be of enormous value to passengers.

1.2 The goal of the project

The goal of this project is to explore the relation between the flight price with other predictors such as airline,date,source,destination and duration.

As a machine learing project,Only the training set of this dataset will be used because the testing set provided by the author did not include the price whereby no test can be implemented on it.

In order to check the model precison,the training set will be de split into two part and one part will be remained as the testing set.

2 Methods and Analysis

2.1 Data cleaning

First the dataset should be cleaned to be fit for analyzing.We can see the first several rows of the dataset:

```
head(Raw)
```

```
##       Airline Date_of_Journey   Source Destination                  Route
## 1      IndiGo      24/03/2019 Banglore   New Delhi            BLR → DEL
## 2   Air India       1/05/2019  Kolkata    Banglore CCU → IXR → BBI → BLR
## 3 Jet Airways       9/06/2019    Delhi      Cochin DEL → LKO → BOM → COK
## 4      IndiGo      12/05/2019  Kolkata    Banglore       CCU → NAG → BLR
## 5      IndiGo      01/03/2019 Banglore   New Delhi       BLR → NAG → DEL
## 6    SpiceJet      24/06/2019  Kolkata    Banglore            CCU → BLR
##   Dep_Time Arrival_Time Duration Total_Stops Additional_Info Price
## 1    22:20 01:10 22 Mar   2h 50m    non-stop         No info  3897
## 2    05:50        13:15   7h 25m     2 stops         No info  7662
## 3    09:25 04:25 10 Jun      19h     2 stops         No info 13882
## 4    18:05        23:30   5h 25m      1 stop         No info  6218
## 5    16:50        21:35   4h 45m      1 stop         No info 13302
```

```
## 6    09:00        11:25   2h 25m    non-stop         No info  3873
```

The general information of the dataset is as below:

```
summary(Raw)
```

```
##    Airline          Date_of_Journey      Source          Destination
##  Length:10683       Length:10683       Length:10683       Length:10683
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Route             Dep_Time         Arrival_Time         Duration
##  Length:10683       Length:10683       Length:10683       Length:10683
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Total_Stops        Additional_Info        Price
##  Length:10683       Length:10683       Min.   : 1759
##  Class :character   Class :character   1st Qu.: 5277
##  Mode  :character   Mode  :character   Median : 8372
##                                        Mean   : 9087
##                                        3rd Qu.:12373
##                                        Max.   :79512
```

We can see that some columns are character and they will be converted into factors and be marked with unique number.

After this work we can see a cleaner dataset:

```
head(Data)
```

```
##   Airline       Date Source Destination Total_Stops Price Duration_hour
## 1       4 2019-03-24      1           6           0  3897      2.833333
## 2       2 2019-05-01      4           1           2  7662      7.416667
## 3       5 2019-06-09      3           2           2 13882     19.000000
## 4       4 2019-05-12      4           1           1  6218      5.416667
## 5       4 2019-03-01      1           6           1 13302      4.750000
## 6       9 2019-06-24      4           1           0  3873      2.416667
```

The we check the basic information of the dataset again:

```
summary(Data)
```

```
##     Airline            Date                Source       Destination
##  Min.   : 1.000   Min.   :2019-03-01   Min.   :1.000   Min.   :1.000
##  1st Qu.: 4.000   1st Qu.:2019-03-27   1st Qu.:3.000   1st Qu.:1.000
##  Median : 5.000   Median :2019-05-15   Median :3.000   Median :2.000
##  Mean   : 4.966   Mean   :2019-05-04   Mean   :2.952   Mean   :2.436
##  3rd Qu.: 5.000   3rd Qu.:2019-06-06   3rd Qu.:4.000   3rd Qu.:3.000
##  Max.   :12.000   Max.   :2019-06-27   Max.   :5.000   Max.   :6.000
##
##   Total_Stops         Price       Duration_hour
##  Min.   :0.0000   Min.   : 1759   Min.   : 1.250
##  1st Qu.:0.0000   1st Qu.: 5277   1st Qu.: 2.833
```

```
##  Median :1.0000   Median : 8372   Median : 8.667
##  Mean   :0.8242   Mean   : 9087   Mean   :10.719
##  3rd Qu.:1.0000   3rd Qu.:12373   3rd Qu.:15.500
##  Max.   :4.0000   Max.   :79512   Max.   :47.667
##  NA's   :1                        NA's   :1
```

Now We see one NA in Duration and Total_Stops which shows there are some unnormal values in the dataset.Have a glimpse at it:

```r
na1<-which(is.na(Data$Duration))
na2<-which(is.na(Data$Total_Stops))
#Check the unnormal value in the Raw dataset
Raw[na1,]
```

```
##        Airline Date_of_Journey Source Destination                 Route
## 6475 Air India       6/03/2019 Mumbai   Hyderabad BOM → GOI → PNQ → HYD
##      Dep_Time Arrival_Time Duration Total_Stops Additional_Info Price
## 6475    16:50        16:55       5m     2 stops         No info 17327
```

```r
Raw[na2,]
```

```
##        Airline Date_of_Journey Source Destination Route Dep_Time Arrival_Time
## 9040 Air India       6/05/2019  Delhi      Cochin  <NA>    09:45 09:25 07 May
##      Duration Total_Stops Additional_Info Price
## 9040  23h 40m        <NA>         No info  7480
```

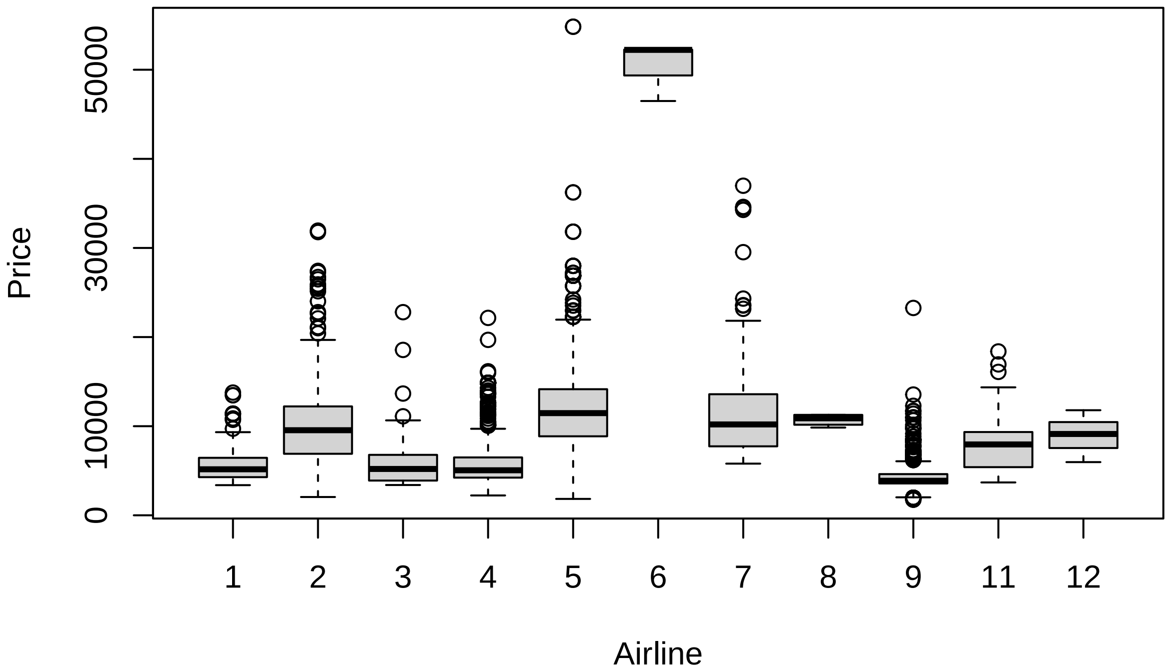
The NA in the Total_Stops is caused by missing value.The time for a flight is impossible to be 5 minutes so we will remove these two rows.

Now we have finished the data cleaning.Then we need to split the dataset into training set and Testing set.The testing set will be 30% of the whole dataset.

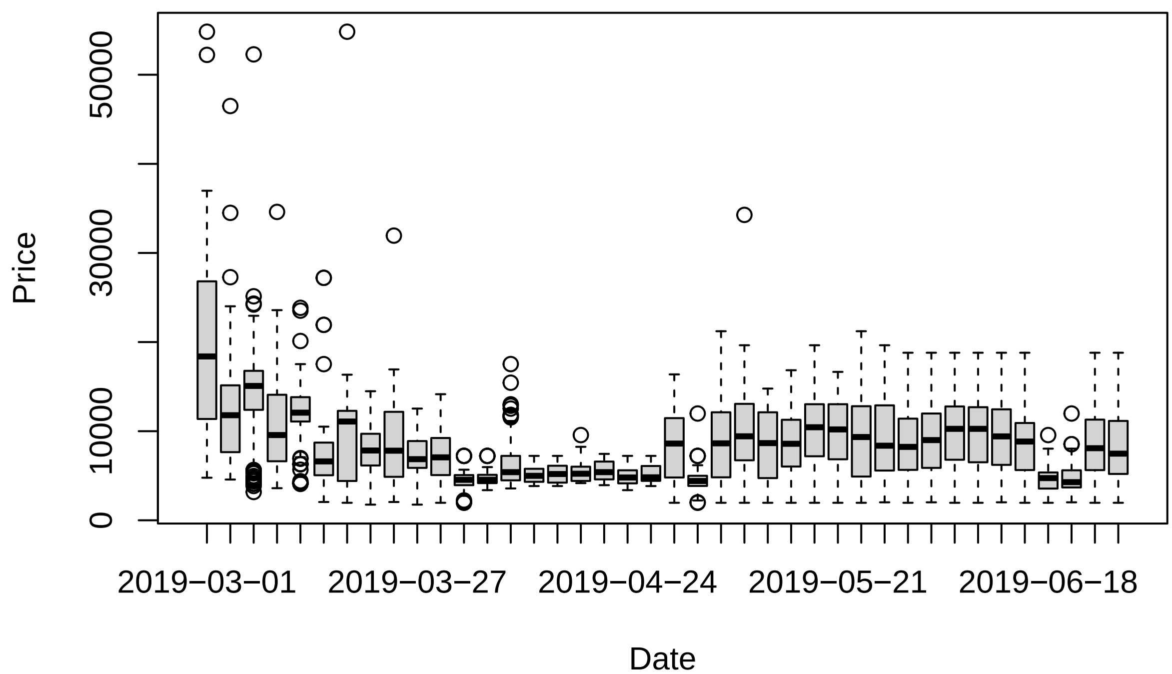
2.2 Visualization

Visualization is a good way to search for some relations between the price and other predictors.

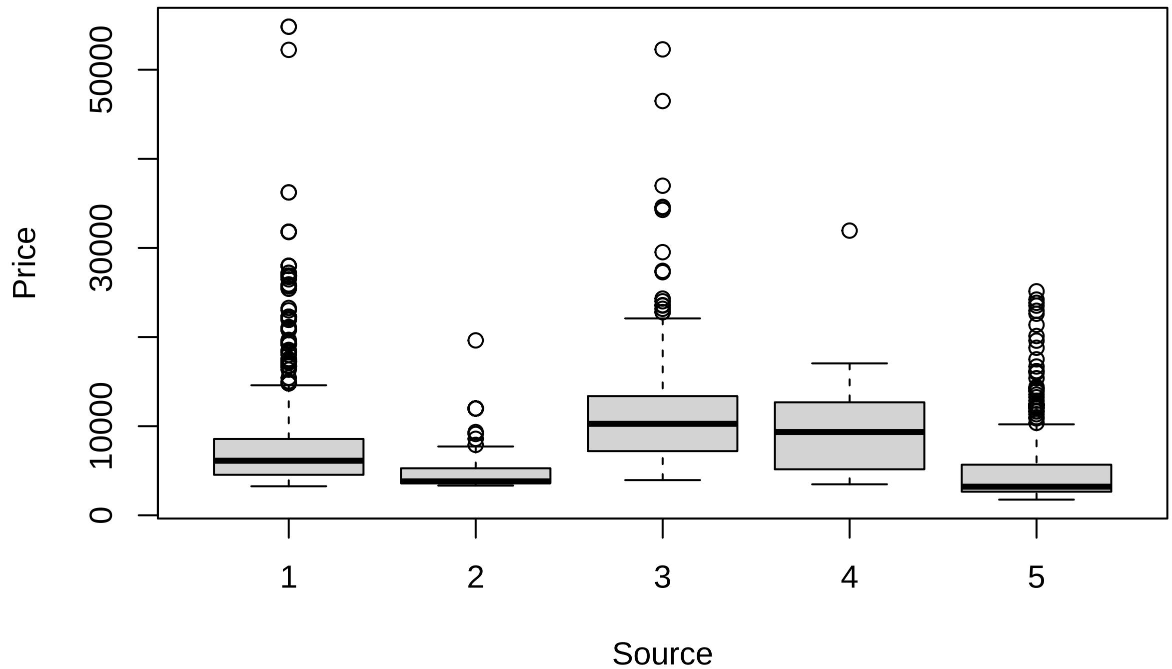
2.2.1 The Airline and Price



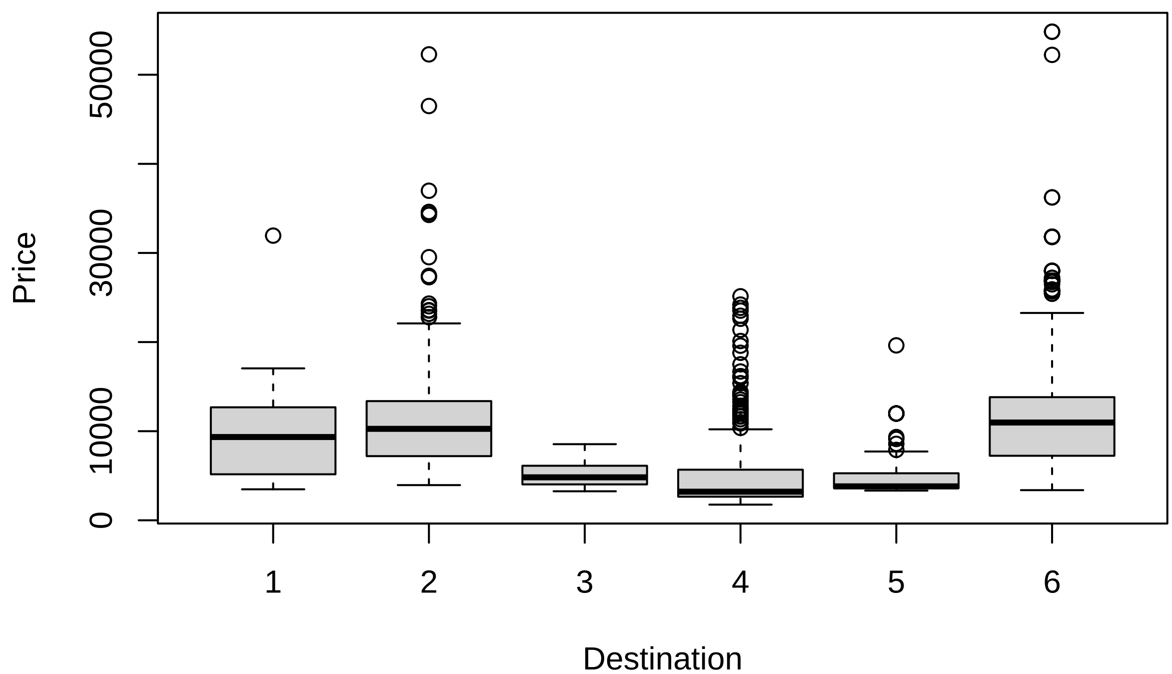
2.2.2 The Date and Price



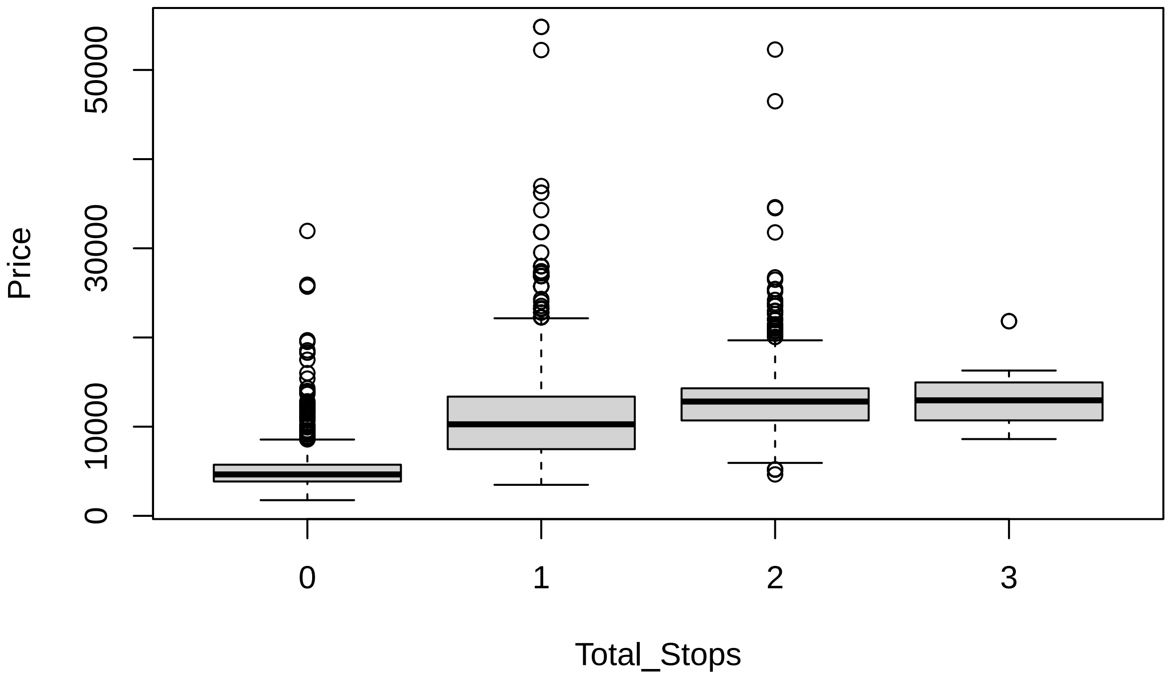
2.2.3 The Source and Price



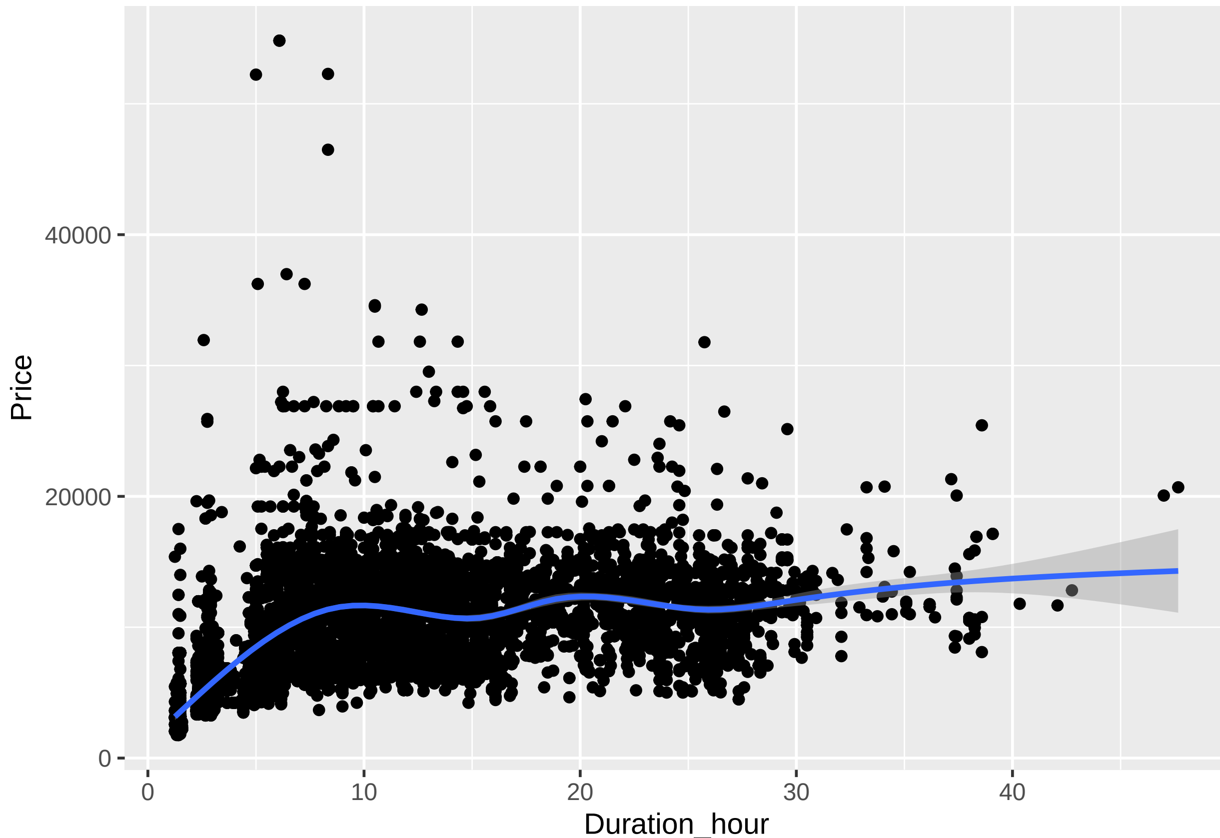
2.2.4 The destination and Price



2.2.5 The Total_Stops and Price



2.2.6 The Duration_hour and Price



2.2.7 Summary of the predictors

To be frank,not any strong relation is showed between the price and other predictors. The analysis be based on all these predictors.

3 The build of the model

3.1 The prediction of price

In order to predict the price,other 6 predictors will be considered and be used in 3 different algorithm which are glm(General liner model),knn(k-Nearest Neighbor) and rf(Random Forest).

For the estimate of the model,we build the RMSE to compute the distance between the estimated price and the true price.

To compute the avarage of all price:

```
mu<-mean(Train_set$Price)
```

Now we can check the RMSE for the first time to see our precision.Define the function RMSE as below:

```
RMSE<-function(pred_value,true_value)
{sqrt(mean((pred_value-true_value)^2))
}
```

And the RMSE for the basic ratings and true ratings is listed:

```
RMSE(mu,Train_set$Price)
```

```
## [1] 4554.445
```

3.2 The knn model

```
model_knn<-train(Price~.,method="knn",data =Data_cor,TuneGrid=data.frame(k=seq(1,20,1)))
pred_knn<-predict(model_knn,Test_set)
```

We can see the RMSE for knn model is:

```
RMSE(pred_knn,Test_set$Price)
```

```
## [1] 2334.089
```

3.3 The glm model

```
model_glm<-train(Price~.,method="glm",data = Data_cor)
pred_glm<-predict(model_glm,Test_set)
```

We can see the RMSE for glm model is:

```
RMSE(pred_glm,Test_set$Price)
```

```
## [1] 3726.003
```

3.4 The rf model

```
set.seed(123,sample.kind = "Rounding")
```

```
## Warning in set.seed(123, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
model_rf<-randomForest(Price ~ ., data = Train_set, importance = TRUE)
pred_rf<-predict(model_rf,Test_set)
```

We can see the RMSE for glm model is:

```
RMSE(pred_rf,Test_set$Price)
```

```
## [1] 2380.458
```

3.5 The ensemble of 3 models

We simply compute the average of every model because we can not use the result generated from the testing set.

```
pred_ensemble<-(pred_knn+pred_glm+pred_rf)/3
```

And we can check the RMSE for the testing dataset:

```
RMSE(pred_ensemble,Test_set$Price)
```

```
## [1] 2575.817
```

We can compare our prediction with just guessing the average price for all price in the testing dataset.

```
mu_test<-mean(Test_set$Price)
RMSE(mu_test,Test_set$Price)
```

```
## [1] 4739.843
```

Now we can see our model do have some improvements in predicting the price of one flight though not very much.

4 The results and conclusion

Though 3 different model are employed and ensemble to predict the price, the result was not satisfying and the RMSE was still quite huge.From the visualization we can not see a obvious relation between the price and other predictors.In normal sense,a longer duration means a longer distance and the price for that flight would be higher while in this dataset it seems not to be in this case.