STA6704 Data Mining II

Airline Customer Satisfaction

- Project Report -

by

Aditi Phopale

Sejal Wadekar

Data Set

The dataset consists of reviews from almost 130k passengers who took a flight journey through a particular airline. They were asked to rate various factors about their flight such as seat comfort, food & drinks, ease of online booking etc. They also gave an overall satisfaction score as Satisfied or Dissatisfied. This survey was conducted in 2015 and the dataset does not mention the name of the airline. There are 130k observations with 24 variables which are described below.

ID: Identifies each passenger uniquely

Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

Age: The actual age of the passengers

Gender: The gender of passengers (Female, Male)

Type of Travel : Purpose of the flight of the passengers (Personal Travel, Business Travel)

Class :Travel class of the passengers (Business, Eco, Eco Plus)

Customer Type: The customer type (Loyal customer, Disloyal customer)

Flight distance: The flight distance of that particular journey

Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)"

Ease of Online booking: Satisfaction level of online booking service

Inflight service: Satisfaction level of inflight service

Online boarding: Satisfaction level of online boarding

Inflight entertainment: Satisfaction level of inflight entertainment

Food and drink: Satisfaction level of Food and drink

Seat comfort: Satisfaction level of Seat comfort

On-board service: Satisfaction level of On-board service

Leg room service: Satisfaction level of Leg room service

Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient

Baggage handling: Satisfaction level of baggage handling

Gate location: Satisfaction level of Gate location

Cleanliness: Satisfaction level of Cleanliness

Check-in service: Satisfaction level of Check-in service

Departure Delay in Minutes: delayed in departure

Arrival Delay in Minutes: delayed in Arrival

Problem Statement

The dataset consists of customer ratings for various factors about the airline service. Our aim is to build a model that helps predict whether a customer satisfied or dissatisfied. The airline company can increase or decrease the ratings of various factors and see how it affects the satisfaction of customers. This can help identify which specific services need improvement to have more satisfied customers.

Variable Exploration

Missing Values

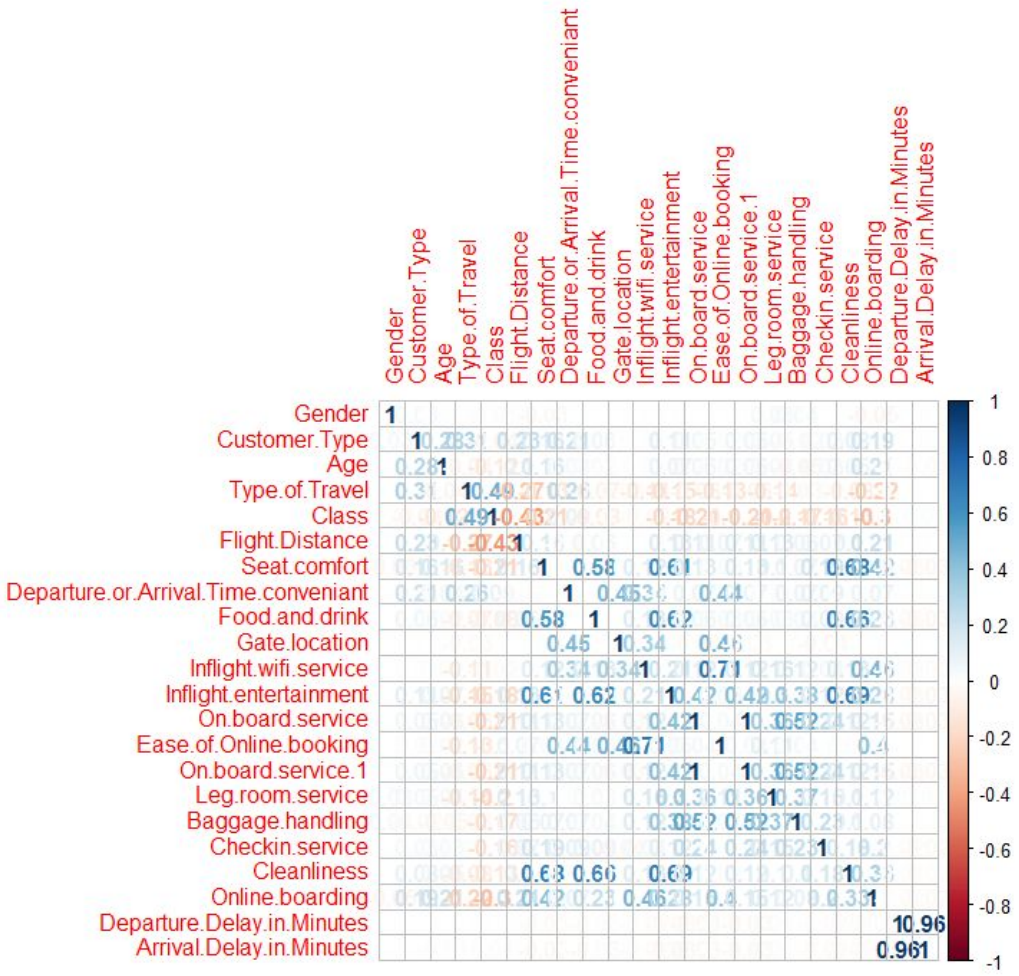
The first step was to check if the dataset consists of any missing or NA values which would not allow us to do any further data processing. There were 393 NA observations in one of the column - ‘Arrival Delay in Minutes’. These NA values were replaced by the column mean.

Renaming Columns and Column Levels

We renamed some columns like ‘satisfaction_v2’, ‘id_’, ‘Departure.Arrival.Time.Conveniant’ to Satisfaction, ID and Departure.or.Arrival.Time.Conveniant. The column Satisfaction had two levels: Satisfied and Neutral or Dissatisfied. The level Neutral or Dissatisfied was renamed to Dissatisfied.

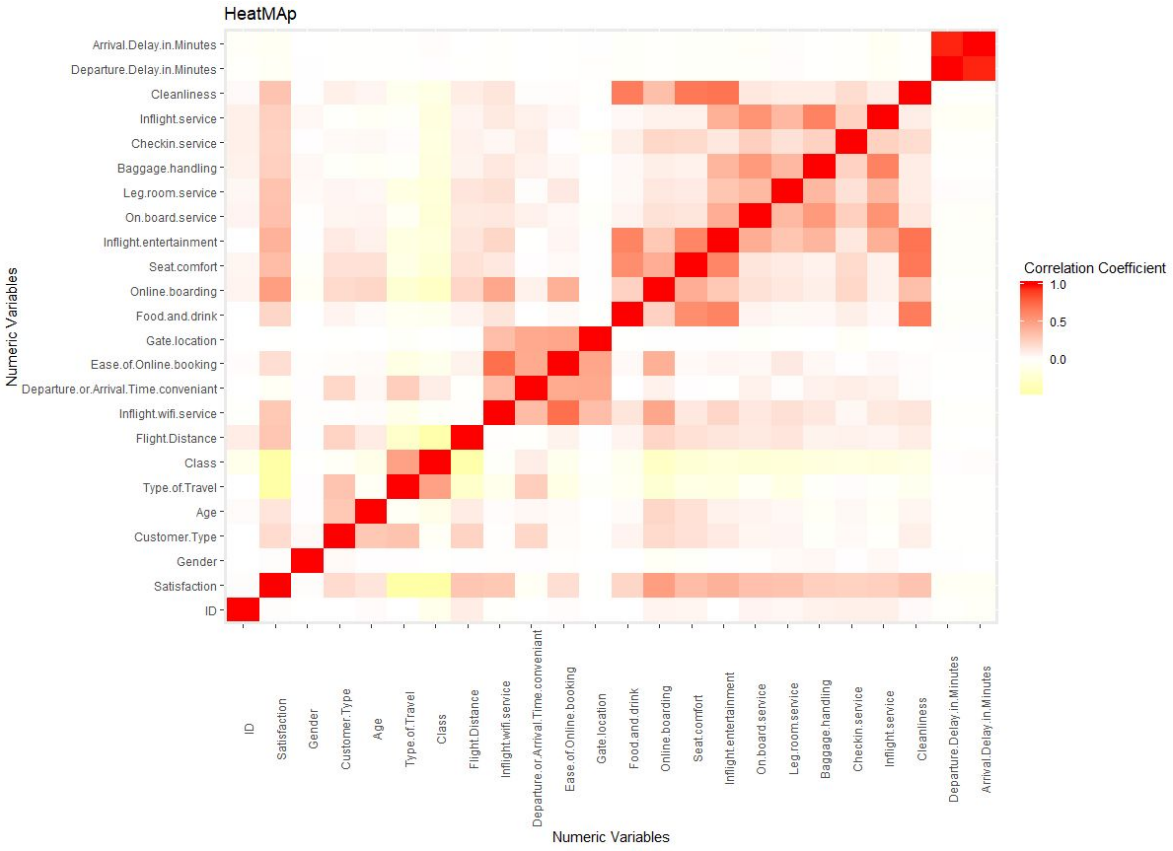
Correlation Matrix and Heatmap

The correlation matrix depicts the collinearity between the variables in the form of a matrix. Variables with correlation values more than 0.80 are said to be highly correlated. Higher the value of the correlation, darker the color in the matrix is.
The quantitative variables are then selected to determine the correlation between them. The correlation matrix is found to be as follows.



From the above matrix, we can see that “Arrival Delay in Minutes” and “Departure Delay in Minutes” are highly correlated followed by “Inflight wifi service” and “Ease of OnlineBooking”.

HeatMap is another way to visualize Hierarchical Clustering. Heat maps allows simultaneous cluster visualization of samples and features.

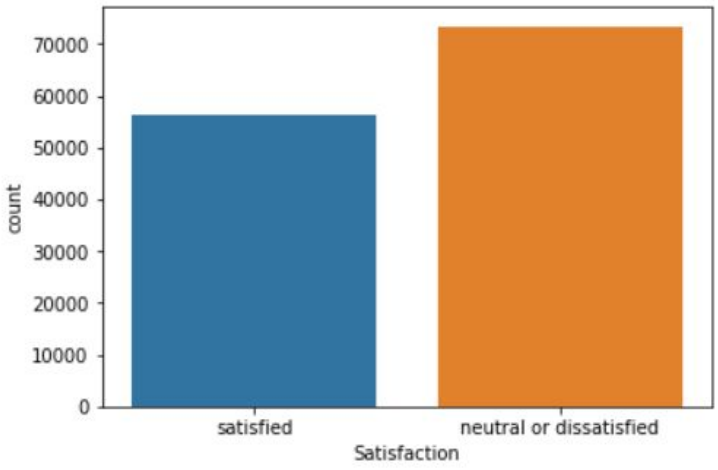


Again, it can be observed that “Arrival Delay in Minutes” and “Departure Delay in Minutes” are highly correlated.

Variable Visualization

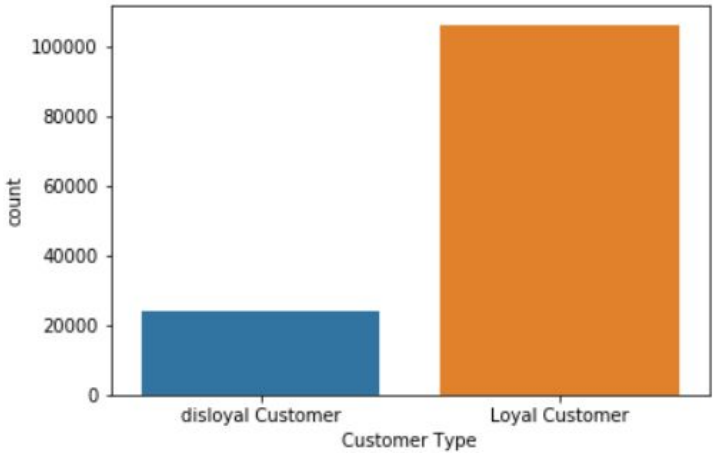
Python 3 was used to observe each variable and the relation between a few of them. Libraries used: Seaborn, Pandas, Matplotlib.

Counting the number of Satisfied vs Dissatisfied passengers:



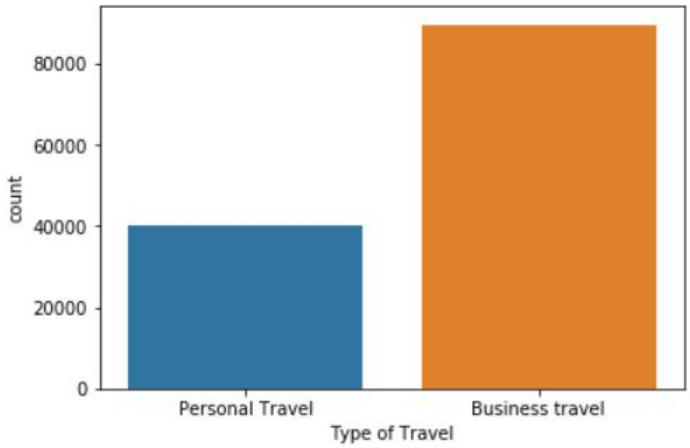
The overall rating shows that more number of passengers are dissatisfied with the airline services. This tells us that the airline company definitely needs to improve their services to provide a better flight journey to their passengers.

Counting the number of Loyal & Disloyal passengers included in the survey:



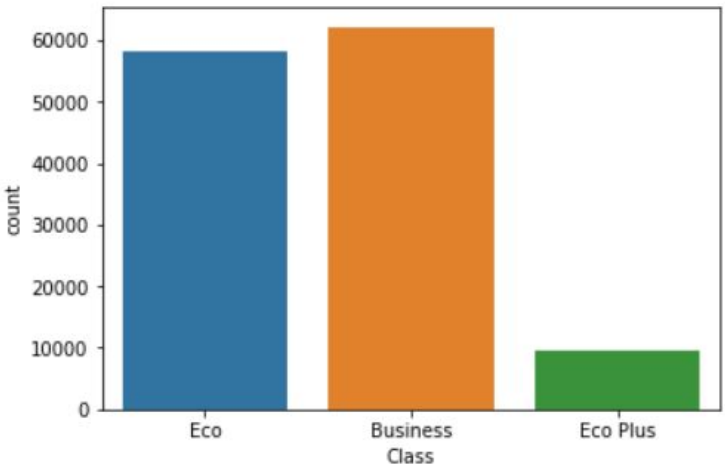
The dataset has a column called ‘Customer Type’ which labels a customer as Loyal or Disloyal. From the above chart we see that most of the passengers surveyed were considered as Loyal Passengers.

Counting the Type of Travel of the passengers:



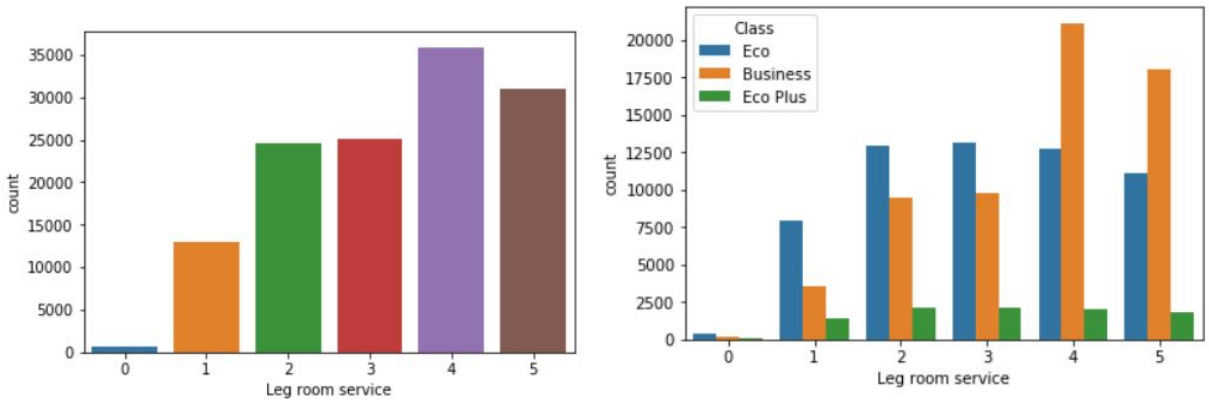
The plot above shows that almost double the number of passengers included in the survey were on a Business Travel rather than Personal.

Counting the distribution of passengers over different Classes:



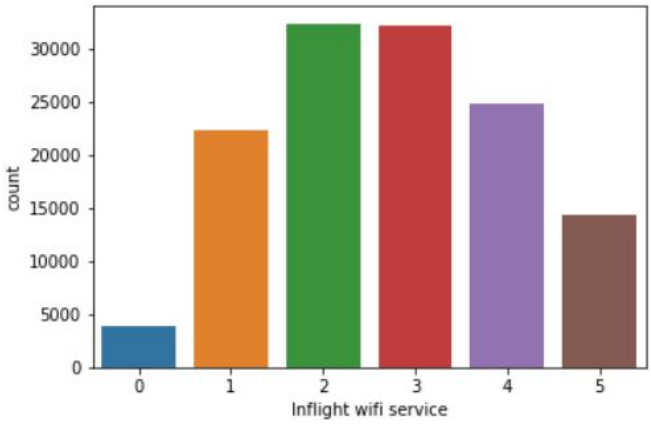
Most of the passengers were from Economy and Business Class and very few from the Economy Plus class.

Ratings for Leg Room Service:



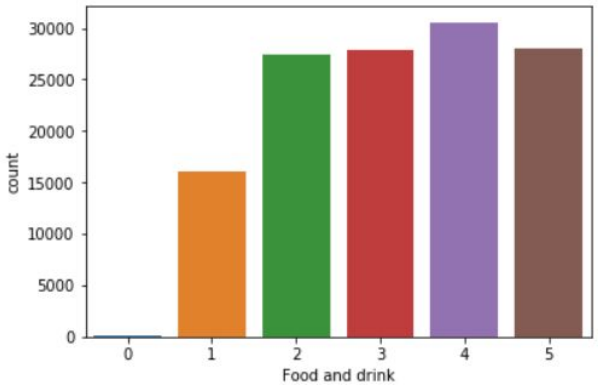
The first plot shows that most customers have rated the leg room space on the higher side. The second plot shows the same chart but divided by Class. We observe that the high leg-room ratings are given mostly by business class passengers. The airline company should try to improve the leg-room for economy class passengers.

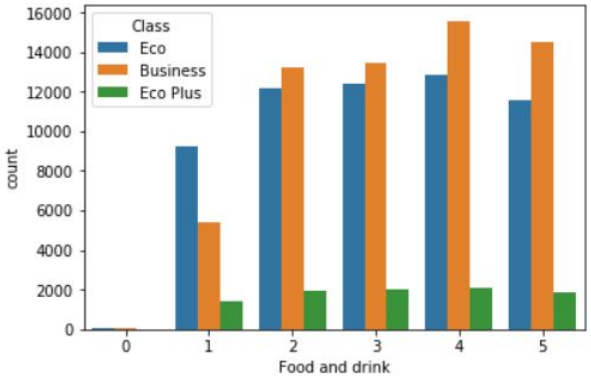
Ratings for In-flight wifi service:



Wifi service has ratings ranging over all the scores, mostly in 2 and 3. There is room for improvement in the on-flight wifi service.

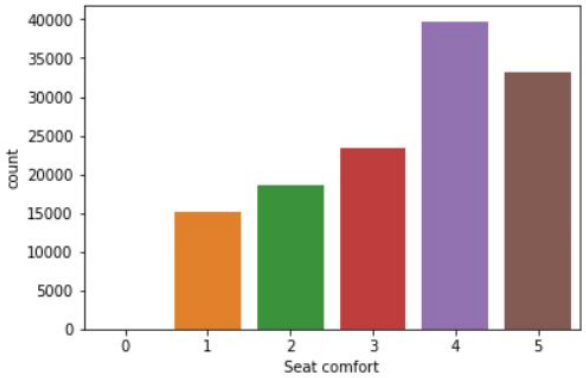
Ratings for Food and Drinks:

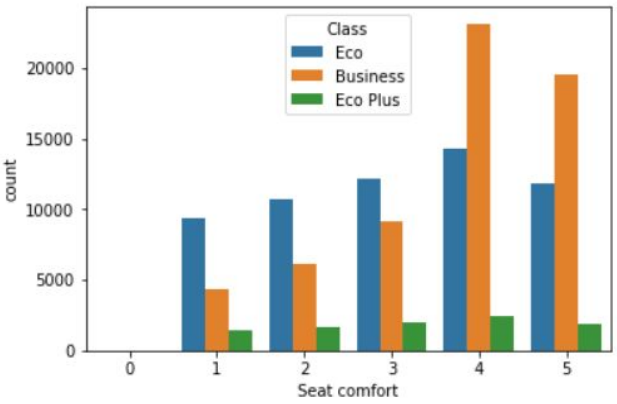




Maximum passengers have given a high rating for food and drink but it can be definitely improved. Second plot tells that the service is rated similarly in both classes.

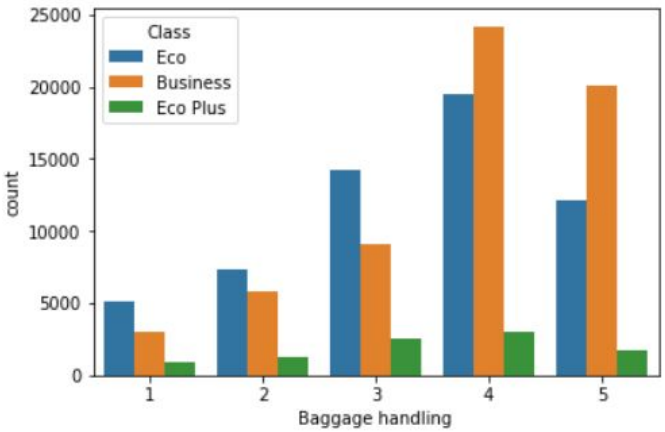
Ratings for Seat Comfort:





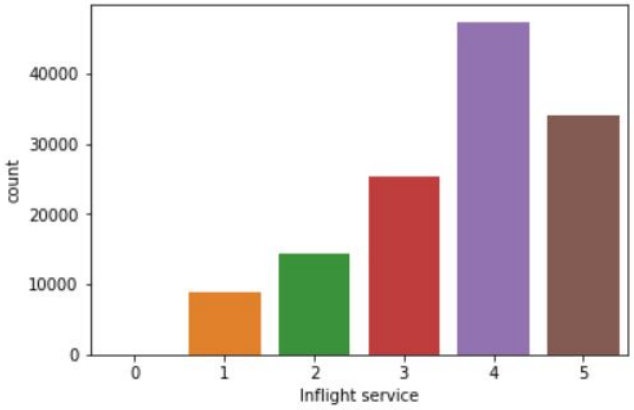
Maximum passengers have given a high rating for seat comfort. From the second plot we see much more business class passengers giving high ratings of 4 and 5. The seat comfort for Economy class can be improved.

Ratings for Baggage Handling:



We see more business class passengers giving high ratings of 4 and 5. Baggage handling can hence be improved.

Ratings for In-flight Services:

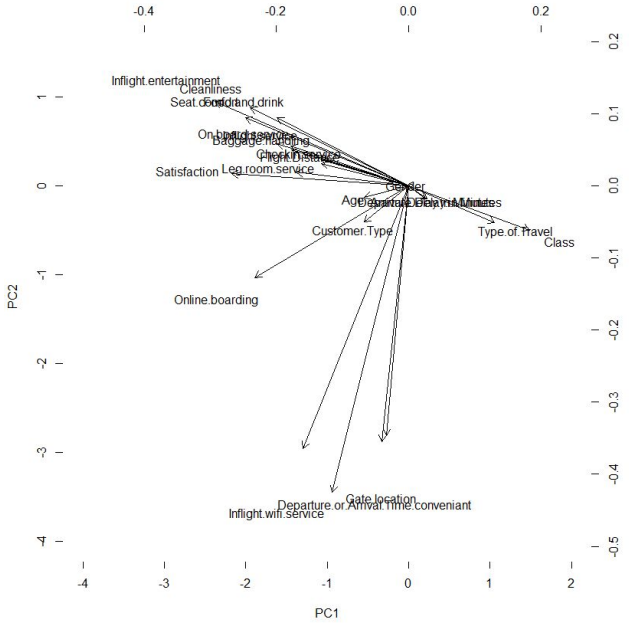


Large number of passengers have rated inflight service 4 out of 5. The ratings on a higher side.

Analysis

Principal Component Analysis

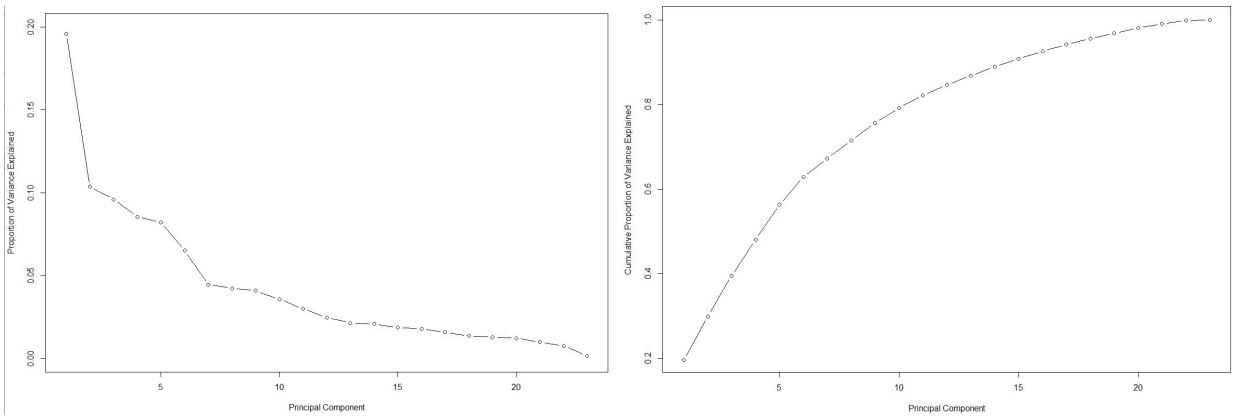
The dataset consists of 24 variables in total. Using PCA we can find less number of components to represent almost all the variation in the dataset. This allows us to work with less number of components and still get accurate results. We scaled the variables because they were measured of different scales: Most of the variables were ratings between 1 to 5, others were categorical variables converted into their numerical representation. prcomp() was used to generate the PC's with Scale = True. The first two PCs and their loading vectors are as shown below, along with the summary of all the PCs.




```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     2.1218 1.5441 1.4860 1.40280 1.3750 1.22486 1.01154 0.98735 0.97016 0.90723 0.83119
Proportion of Variance 0.1957 0.1037 0.0960 0.08556 0.0822 0.06523 0.04449 0.04239 0.04092 0.03579 0.03004
Cumulative Proportion  0.1957 0.2994 0.3954 0.48096 0.5632 0.62839 0.67288 0.71526 0.75618 0.79197 0.82201
                          PC12    PC13   PC14    PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.75540 0.70213 0.6934 0.65716 0.64390 0.60534 0.56041 0.54524 0.53104 0.47701
Proportion of Variance 0.02481 0.02143 0.0209 0.01878 0.01803 0.01593 0.01365 0.01293 0.01226 0.00989
Cumulative Proportion  0.84682 0.86825 0.8891 0.90793 0.92596 0.94189 0.95554 0.96847 0.98073 0.99062
                          PC22    PC23
Standard deviation     0.41956 0.19902
Proportion of Variance 0.00765 0.00172
Cumulative Proportion  0.99828 1.00000
```

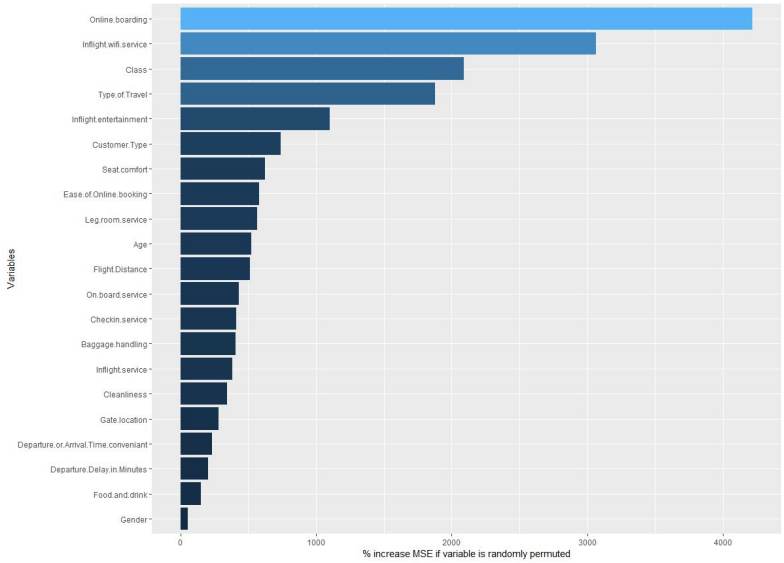
We can see that 88.91% of the variance is explained by only the first 14 PCs. Hence, we can select only these 14 PCs which will act as 14 predictors for the dataset instead of 24 variables in the original dataset.

The scree plot and cumulative scree plot below can also help select the number of PCs to be considered out of the 24 PCs.



Random Forest

Random Forest was applied on 22 variables left after dropping ‘ID’ and highly correlated column ‘Arrival.Delay.in.Minutes’. ‘Satisfaction’ was the output column. Data was divided into 60:40 ratio for train and test respectively. The importance of variables was calculated as shown below. Based on this output, the top 10 variables were selected for modelling.



The train and test results are shown below. The highest training accuracy of 88.73% is obtained. The test accuracy obtained was 94.57% good sensitivity and specificity values of 0.9481 and 0.9439 respectively.

```
Random Forest

77928 samples
   10 predictor
    2 classes: 'dissatisfied', 'satisfied'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 77928, 77928, 77928, 77928, 77928, 77928, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.9447391  0.8873940
   6    0.9422419  0.8823439
  10    0.9364504  0.8705967

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

```
Confusion Matrix and Statistics

             solution_rf
              dissatisfied satisfied
  dissatisfied        28378      1135
  satisfied            1688     20751

               Accuracy : 0.9457
                 95% CI : (0.9437, 0.9476)
    No Information Rate : 0.5787
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8889

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9481
            Specificity : 0.9439
         Pos Pred Value : 0.9248
         Neg Pred Value : 0.9615
             Prevalence : 0.4213
         Detection Rate : 0.3994
   Detection Prevalence : 0.4319
      Balanced Accuracy : 0.9460

       'Positive' Class : satisfied
```

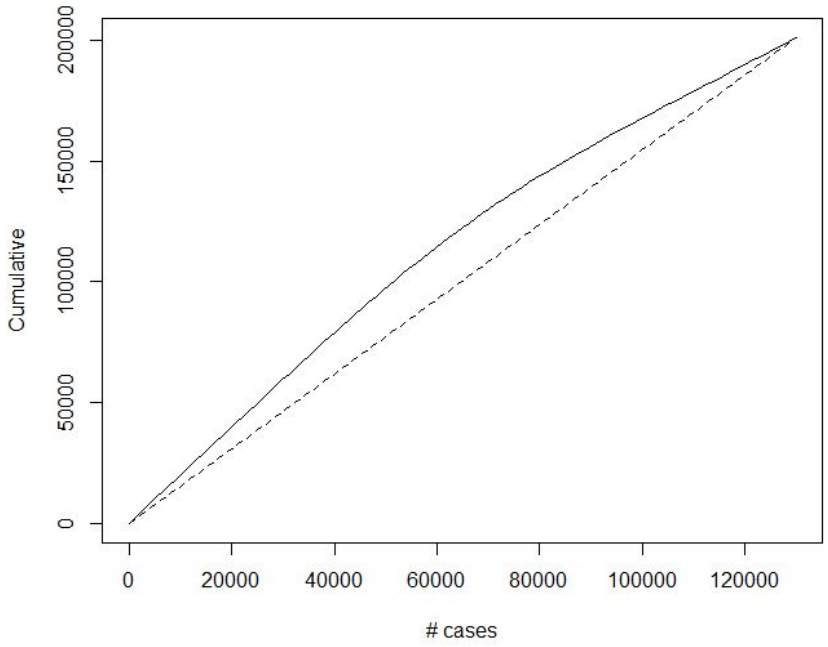
Logistic Regression

Logistic Regression is a predictive analysis method which is used to explain the relationship between one dependent binary variable or the target variable and one or more independent variables or predictors. Logistic Regression model is first implemented in the dataset and the confusion matrices are printed to determine the accuracy.
Following is the summary of the Logistic Regression model.

```
Deviance Residuals:
       Min          1Q      Median          3Q         Max
-3.0387831  -0.5776280   0.1937446   0.5194128   3.6674842

Coefficients:
                                          Estimate       Std. Error   z value               Pr(>|z|)
(Intercept)                             -6.4963849695098  0.0772829488383 -84.05974 < 0.000000000000000222 ***
ID                                      -0.0000032835033  0.0000002181333 -15.05274 < 0.000000000000000222 ***
Gender                                  -0.9730610518696  0.0164622482781 -59.10864 < 0.000000000000000222 ***
Customer.Type                            2.0322508829557  0.0245991188539  82.61478 < 0.000000000000000222 ***
Age                                     -0.0073352522320  0.0005715340921 -12.83432 < 0.000000000000000222 ***
Type.of.Travel                          -0.8942901941693  0.0219272121234 -40.78449 < 0.000000000000000222 ***
Class                                   -0.5148491237325  0.0153394173263 -33.56380 < 0.000000000000000222 ***
Flight.Distance                         -0.0000964534144  0.0000086023781 -11.21241 < 0.000000000000000222 ***
Seat.comfort                             0.2795961258279  0.0092106045338  30.35589 < 0.000000000000000222 ***
Departure.or.Arrival.Time.conveniant    -0.1971159354250  0.0067586403813 -29.16503 < 0.000000000000000222 ***
Food.and.drink                          -0.2163085424491  0.0093625789708 -23.10352 < 0.000000000000000222 ***
Gate.location                            0.1146732783369  0.0076321634614  15.02500 < 0.000000000000000222 ***
Inflight.wifi.service                   -0.0693967162777  0.0088850197314  -7.81053  0.000000000000056948 ***
Inflight.entertainment                   0.6973040862945  0.0083130214569  83.88094 < 0.000000000000000222 ***
Online.support                           0.1018518620459  0.0090331414091  11.27535 < 0.000000000000000222 ***
Ease.of.Online.booking                   0.2015211099059  0.0116510121472  17.29645 < 0.000000000000000222 ***
On.board.service                         0.3152999543962  0.0082422488088  38.25412 < 0.000000000000000222 ***
Leg.room.service                         0.2261286561553  0.0070124822062  32.24659 < 0.000000000000000222 ***
Baggage.handling                         0.1166036505205  0.0092923960823  12.54829 < 0.000000000000000222 ***
Checkin.service                          0.3031007425621  0.0069418779845  43.66264 < 0.000000000000000222 ***
Cleanliness                              0.0922437361790  0.0096720351962   9.53716 < 0.000000000000000222 ***
Online.boarding                          0.1749492380141  0.0099737052590  17.54105 < 0.000000000000000222 ***
Departure.Delay.in.Minutes               0.0033854697774  0.0007507196785   4.50963  0.000006494058153457 ***
Arrival.Delay.in.Minutes                -0.0088178224722  0.0007444543015 -11.84468 < 0.000000000000000222 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Lift charts are used to measure the performance of a predictive classification model. They measure how much better results one can expect with the predictive classification model comparing without a model. From the lift curve plotted for the validation data, we notice that our model is superior to the baseline model.



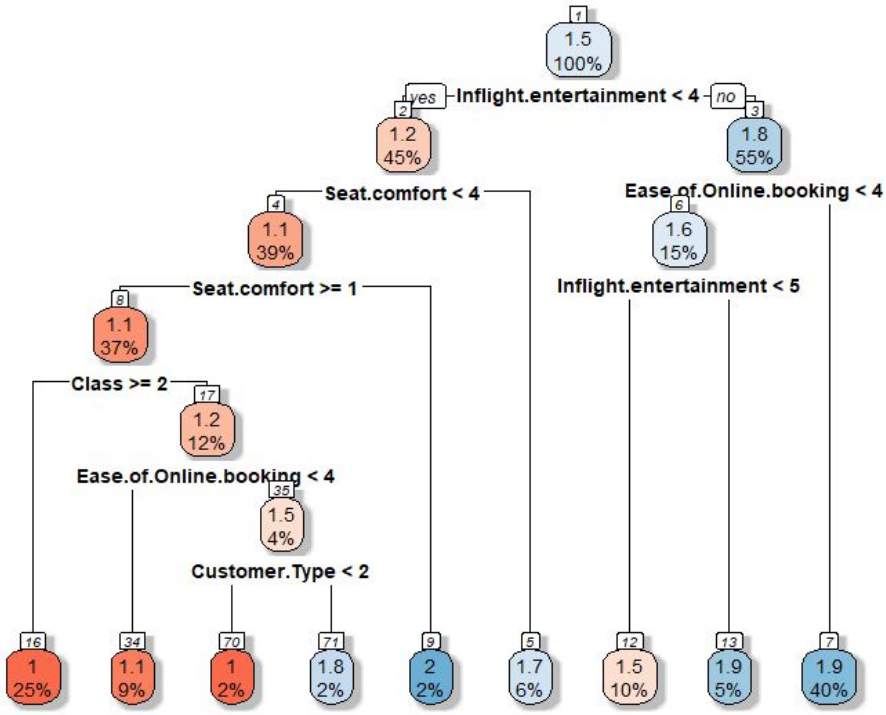
Confusion Matrix was created for the test data and accuracy percentage was noted. It was observed that the accuracy for our validation model was 83.47%

```
> table_test
         Actual
Predicted     1     2
        1 14188  3226
        2  3213 18337
```

```
> acc_test
[1] 83.47449
>
```

Decision Tree

Decision Trees are supervised learning methods used for both Regression and Classification problems. The goal is to create a model which predicts the value of the target variable by learning simple decision tools inferred from the data features.



Decision Tree model was implemented and the summary of the trained model was found as follows.

```
> summary(ndectree)

Regression tree:
tree(formula = Satisfaction ~ ., data = training_data, method = "rpart")
Variables actually used in tree construction:
[1] "Inflight.entertainment" "Seat.comfort"           "Ease.of.Online.booking"
Number of terminal nodes:  6
Residual mean deviance:  0.111 = 10090 / 90910
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.99730 -0.09645  0.09975  0.00000  0.09975  0.90360
```

```
Call:
rpart(formula = Satisfaction ~ ., data = training_data, method = "class")
  n= 90916

          CP nsplit rel error    xerror        xstd
1 0.56547159      0 1.0000000 1.0000000 0.003627684
2 0.05174913      1 0.4345284 0.4345284 0.002901909
3 0.04387321      2 0.3827793 0.3827793 0.002763343
4 0.01349697      3 0.3389061 0.3389061 0.002631423
5 0.01000000      6 0.2984152 0.3086345 0.002531532

Variable importance
Inflight.entertainment           Seat.comfort         Online.support Ease.of.Online.booking
                    33                     20                     15                     12
         Food.and.drink        Online.boarding            Cleanliness        Baggage.handling
                     9                      9                      1                      1
       On.board.service
                     1
```

From the mentioned Complexity Parameter values, we selected the one having the least cross-validated error and chose it to prune the tree. The value chosen for Complexity Parameter should be least so that the cross validated error rate is minimum. Here, the tree was pruned with the CP value as 0.01

Confusion matrix was created for the Decision Tree model and the accuracy percentage was noted. The Accuracy was found to be 86.43% with a good sensitivity and specificity rate of 0.8593 and 0.8683 respectively.

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1 14953  2840
         2  2448 18723

               Accuracy : 0.8643
                 95% CI : (0.8608, 0.8677)
    No Information Rate : 0.5534
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.726

 Mcnemar's Test P-Value : 7.578e-08

            Sensitivity : 0.8593
            Specificity : 0.8683
         Pos Pred Value : 0.8404
         Neg Pred Value : 0.8844
             Prevalence : 0.4466
         Detection Rate : 0.3838
   Detection Prevalence : 0.4567
      Balanced Accuracy : 0.8638

       'Positive' Class : 1
```

Bagging

Bagging, is a machine learning ensemble algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method.
Here, a standard training set consisting of 70% of the data of the Passenger dataset is taken and 10 fold Cross Validation is used to generate samples of the training set by sampling the Passenger dataset uniformly.

Summary of the Bagging classification is as follows:

```
Bagging classification trees with 25 bootstrap replications

Call: bagging.data.frame(formula = Satisfaction ~ ., data = airline_data,
    coob = TRUE)

Out-of-bag estimate of misclassification error:  0.0465
```

Next, the predictions are made. The accuracy of the model is checked with the Confusion matrix and the Kappa statistic.

```
bagPred     1     2
      1 58773    52
      2    20 71035
```

```
> Kappa(keep)
             value       ASE    z Pr(>|z|)
Unweighted 0.9989 0.0001318 7579        0
Weighted   0.9989 0.0001318 7579        0
```

The accuracy seems pretty high. In addition, the Kappa value was found to be almost 0.99 indicating a well-fitted model. To further confirm this, we perform 10 fold Cross Validation using functions from the ‘caret’ package.
The accuracy of our model seems to have increased from 86.43% to 95.64% after Bagging.

```
Bagged CART

129880 samples
    23 predictor
     2 classes: '1', '2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 116892, 116893, 116891, 116892, 116892, 116891, ...
Resampling results:

  Accuracy   Kappa
  0.9564983  0.9123216
```

Boosting

Boosting is similar to bagging but the trees here are grown sequentially. Each tree makes use of the information from the previously built trees. It aims at improving the prediction accuracy. Various parameters such as tree depth, number of trees, shrinkage parameter etc are to be considered. We experimented with a few combinations and then selected the best one. The analysis is as shown below.

Number of Trees	Interaction Depth	Error
1000	10	22.85632
2000	8	22.08299
1000	3	19.39965
2000	5	21.50364
1500	3	19.01067
500	5	21.16397
500	4	19.8437
500	2	16.58375

We selected the model with 500 trees and depth of 2 which has the lowest error of 16.58375.

Result Summary

We ran 5 supervised algorithms on the Passenger dataset, summary of which is listed below.

Models Developed	**Accuracy**
Random Forest	88.73%
Logistic Regression	83.47%
Decision Tree	86.43%
Bagging	95.64%
Boosting	83.41%

The final model selected was Bagging since it provided the maximum test accuracy of 95.64%.