Classification

Justin Hardy, Fernando Colman, Linus Fackler, Isabelle Villegas

The Data Set

Starting by reading in the data set. The data set we'll use for the assignment consists of data collected by an airline organization, over their customers' submitted satisfaction surveys, as well as relevant information about their flight and demographic.

If you want to see the data set for yourself, you access it here.

```r
# Read data set
CustomerData_raw <- read.csv("Invistico_Airline.csv")
```

Cleaning Up The Data Set

Cleaning up data set for logistic regression, by converting qualitative columns into factors.

```r
# Create new cleaned CustomerData data frame for scaling (kNN)
CustomerData_scaled <- CustomerData_raw

# Factor columns
CustomerData_scaled$satisfaction <- factor(CustomerData_scaled$satisfaction) # satisfaction
CustomerData_scaled$Gender <- factor(CustomerData_scaled$Gender) # gender
CustomerData_scaled$Customer.Type <- factor(CustomerData_scaled$Customer.Type) # customer type
CustomerData_scaled$Type.of.Travel <- factor(CustomerData_scaled$Type.of.Travel) # travel type
CustomerData_scaled$Class <- factor(CustomerData_scaled$Class) # class

# Normalize factor names
levels(CustomerData_scaled$satisfaction) <- c("Dissatisfied", "Satisfied")
levels(CustomerData_scaled$Customer.Type) <- c("Disloyal", "Loyal")
levels(CustomerData_scaled$Type.of.Travel) <- c("Business", "Personal")

# Create new cleaned CustomerData data frame for full factoring (linear regression)
CustomerData_factored <- CustomerData_scaled

# Continue factoring numeric finite columns
for(i in 8:21) {
  CustomerData_factored[,i] <-
    factor(CustomerData_factored[,i], levels=c(0,1,2,3,4,5)) # out-of-5 ratings
}

# Remove na rows
CustomerData_scaled <- CustomerData_scaled[complete.cases(CustomerData_scaled),]
CustomerData_factored <- CustomerData_factored[complete.cases(CustomerData_factored),]
```

Dividing Into Train/Test

Dividing the data set into train/test…

```
# train/test division
i <- sample(1:nrow(CustomerData_factored), nrow(CustomerData_factored)*0.8, replace=FALSE)
train <- CustomerData_factored[i,]
test <- CustomerData_factored[-i,]

# scaling on separate data frame
train_scaled <- CustomerData_scaled[i,
    names(CustomerData_scaled[i,])[sapply(CustomerData_scaled[i,], is.numeric)]]
test_scaled <- CustomerData_scaled[-i,
    names(CustomerData_scaled[i,])[sapply(CustomerData_scaled[-i,], is.numeric)]]
train_labels <- CustomerData_scaled[i, 1]
test_labels <- CustomerData_scaled[-i, 1]

means <- sapply(train_scaled, mean)
stdvs <- sapply(train_scaled, sd)
train_scaled <- scale(train_scaled, center=means, scale=stdvs)
test_scaled <- scale(test_scaled, center=means, scale=stdvs)
```

Keep in mind that, we have also created a second version of this data set, split into a separate train/test, where the 0-5 Ratings (factored into 6 levels in the main version) are kept numerical/continuous for use in kNN classification.

Data Exploration

Structure

Exploring the train data, we can see that each of our 0-5 Ratings were factored into levels of 6. The reason I opted to factor the data this way is because, although the values are numerical, they're a small finite set of integers. I also noticed higher accuracy in my results after factoring the data this way, which seems to confirm that this was a good decision.

```
##        satisfaction      Gender       Customer.Type        Age
##  Dissatisfied:46724   Female:52584   Disloyal:18889   Min.   : 7.00
##  Satisfied   :56865   Male  :51005   Loyal   :84700   1st Qu.:27.00
##                                                       Median :40.00
##                                                       Mean   :39.44
##                                                       3rd Qu.:51.00
##                                                       Max.   :85.00
##   Type.of.Travel       Class       Flight.Distance Seat.comfort
##  Business:71594   Business:49729   Min.   :  50    0: 3825
##  Personal:31995   Eco     :46345   1st Qu.:1359    1:16737
##                   Eco Plus: 7515   Median :1923    2:23022
##                                    Mean   :1980    3:23131
##                                    3rd Qu.:2543    4:22647
##                                    Max.   :6951    5:14227
##  Departure.Arrival.time.convenient Food.and.drink Gate.location
##  0: 5341                           0: 4716        0:    2
```

```
##  1:16656                          1:16856       1:18062
##  2:18245                          2:21765       2:19673
##  3:18394                          3:22302       3:26591
##  4:23577                          4:21756       4:24020
##  5:21376                          5:16194       5:15241
##  Inflight.wifi.service Inflight.entertainment Online.support
##  0:   93               0: 2383                0:    1
##  1:11725               1: 9431                1:11043
##  2:21605               2:15254                2:13761
##  3:22016               3:19197                3:17273
##  4:25203               4:33486                4:33078
##  5:22947               5:23838                5:28433
##  Ease.of.Online.booking On.board.service Leg.room.service Baggage.handling
##  0:   11                0:    2          0:  354          0:    0
##  1:10640                1:10530          1: 8814          1: 6343
##  2:15890                2:13643          2:17362          2:10723
##  3:17960                3:21559          3:17878          3:19471
##  4:31885                4:32490          4:31693          4:38569
##  5:27203                5:25365          5:27488          5:28483
##  Checkin.service Cleanliness Online.boarding Departure.Delay.in.Minutes
##  0:    1         0:    2     0:   10         Min.   :   0.00
##  1:12215         1: 6166     1:12167         1st Qu.:   0.00
##  2:12359         2:10693     2:14799         Median :   0.00
##  3:28309         3:19087     3:24552         Mean   :  14.53
##  4:29140         4:39001     4:28184         3rd Qu.:  12.00
##  5:21565         5:28640     5:23877         Max.   :1592.00
##  Arrival.Delay.in.Minutes
##  Min.   :   0.00
##  1st Qu.:   0.00
##  Median :   0.00
##  Mean   :  14.95
##  3rd Qu.:  13.00
##  Max.   :1584.00

## 'data.frame':    103589 obs. of  23 variables:
##  $ satisfaction                     : Factor w/ 2 levels "Dissatisfied",..: 1 2 1 2 1 2 2 2 1 1 ...
##  $ Gender                           : Factor w/ 2 levels "Female","Male": 1 2 2 1 1 2 1 2 2 1 ...
##  $ Customer.Type                    : Factor w/ 2 levels "Disloyal","Loyal": 2 2 1 2 2 2 2 2 2 2 ...
##  $ Age                              : int  39 76 22 70 41 33 42 46 43 45 ...
##  $ Type.of.Travel                   : Factor w/ 2 levels "Business","Personal": 1 1 1 2 2 1 1 1 1 1
##  $ Class                            : Factor w/ 3 levels "Business","Eco",..: 1 1 2 2 2 2 3 1 1 1 ..
##  $ Flight.Distance                  : int  2453 3401 1999 396 2587 2106 496 1585 3419 4067 ...
##  $ Seat.comfort                     : Factor w/ 6 levels "0","1","2","3",..: 5 5 2 3 5 5 6 6 4 4 ...
##  $ Departure.Arrival.time.convenient: Factor w/ 6 levels "0","1","2","3",..: 2 3 2 3 4 3 6 6 2 4 ...
##  $ Food.and.drink                   : Factor w/ 6 levels "0","1","2","3",..: 2 5 2 3 5 5 5 6 2 4 ...
##  $ Gate.location                    : Factor w/ 6 levels "0","1","2","3",..: 2 5 4 3 5 5 6 6 2 4 ...
##  $ Inflight.wifi.service            : Factor w/ 6 levels "0","1","2","3",..: 6 5 5 4 6 5 6 4 4 4 ...
##  $ Inflight.entertainment           : Factor w/ 6 levels "0","1","2","3",..: 5 5 2 5 5 5 5 6 4 4 ...
##  $ Online.support                   : Factor w/ 6 levels "0","1","2","3",..: 5 2 5 5 6 5 3 5 5 4 ...
##  $ Ease.of.Online.booking           : Factor w/ 6 levels "0","1","2","3",..: 5 5 5 5 6 5 6 6 4 3 ...
##  $ On.board.service                 : Factor w/ 6 levels "0","1","2","3",..: 5 5 4 5 3 3 6 6 4 2 ...
##  $ Leg.room.service                 : Factor w/ 6 levels "0","1","2","3",..: 5 5 3 5 2 5 6 6 4 4 ...
##  $ Baggage.handling                 : Factor w/ 6 levels "0","1","2","3",..: 5 5 5 5 5 3 6 6 4 4 ...
##  $ Checkin.service                  : Factor w/ 6 levels "0","1","2","3",..: 5 5 6 6 6 4 5 4 5 4 ...
```

```
##  $ Cleanliness                      : Factor w/ 6 levels "0","1","2","3",..: 5 5 6 5 4 4 6 6 4 3 ...
##  $ Online.boarding                  : Factor w/ 6 levels "0","1","2","3",..: 4 4 5 5 6 5 5 4 5 4 ...
##  $ Departure.Delay.in.Minutes       : int  0 0 18 19 0 0 7 5 34 141 ...
##  $ Arrival.Delay.in.Minutes         : int  0 0 14 21 20 15 1 0 37 136 ...

## [1] "Number of NAs: 0"
```

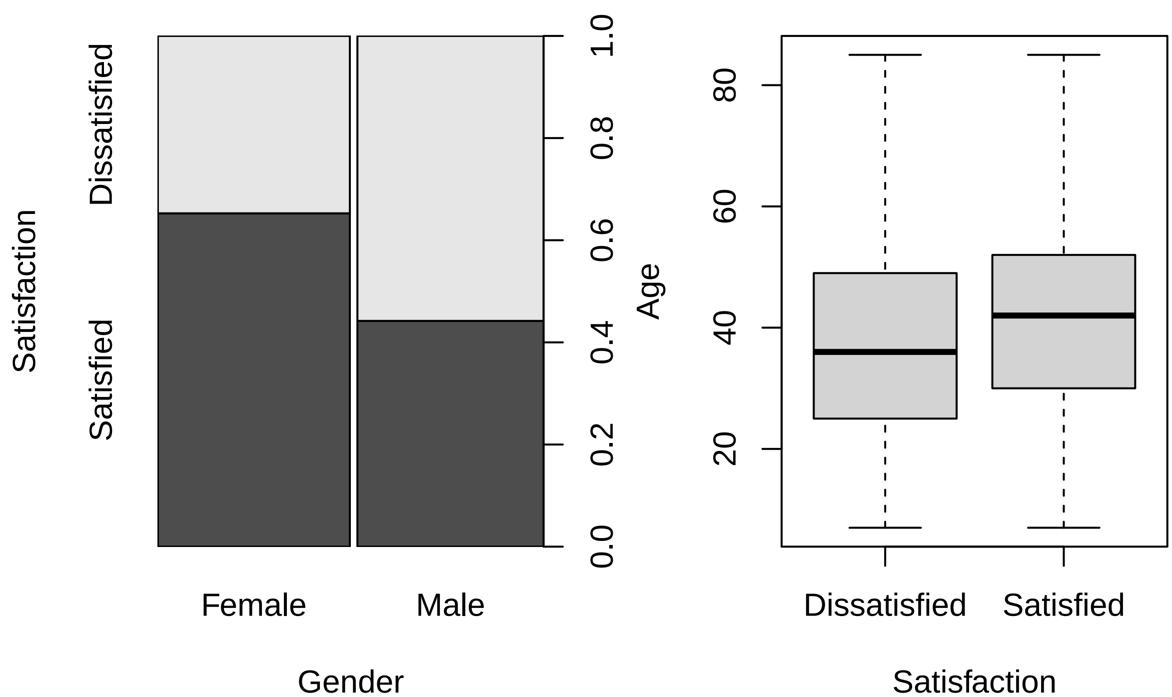
Graphs & Plots

Plotting the data, we can see the relationships between various attributes (or lack thereof):

In the two graphs below, we are seeking to observe for a relationship between the customer's demographics and their satisfaction.

In the left-hand graph, we can observe that females were generally more satisfied with their flights than dissatisfied, as opposed to males who were generally more dissatisfied than satisfied. This may make for a good point of prediction.

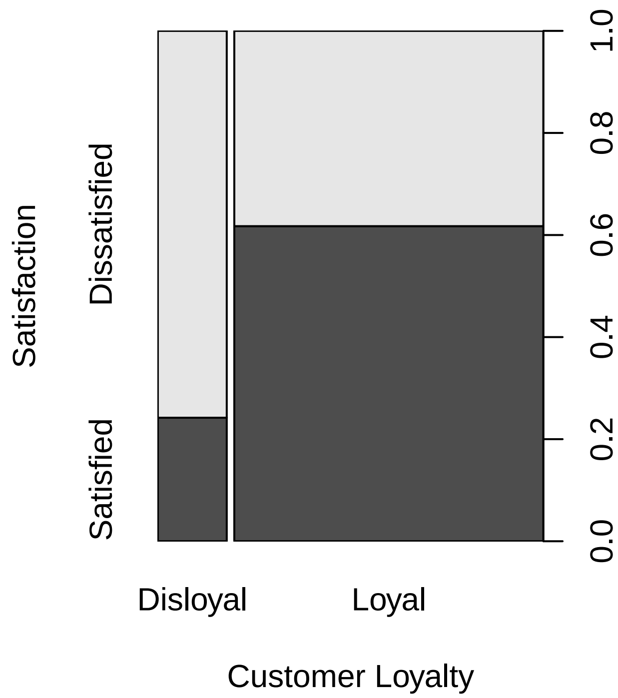
In the right-hand graph, we can observe that those satisfied with their flight were, on average, older than those who were dissatisfied. However, the difference is very small, and the values fall within similar ranges, so it may not make for a good point of prediction.

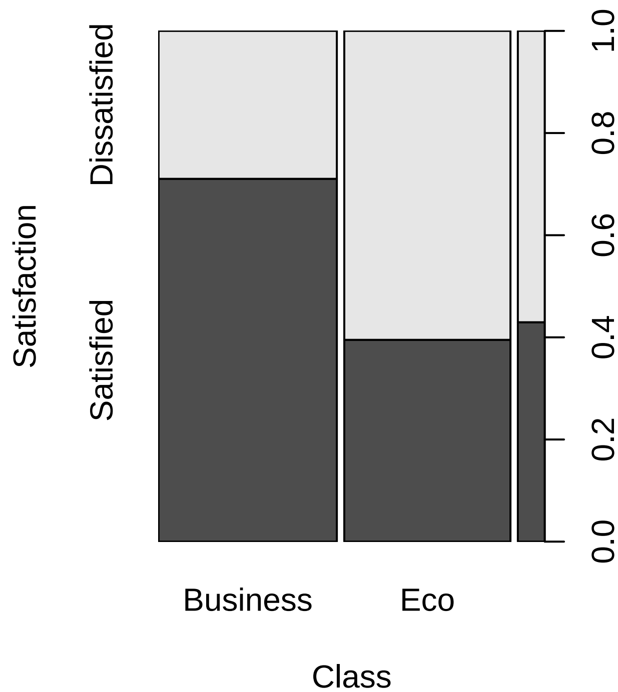


Furthermore, in the next two graphs below, we are seeking to determine if there is a observe for a relationship between the customer's classifications and their satisfaction.

In the left-hand graph, we can observe that loyal customers are significantly likely to be satisfied with their flight, while disloyal customers are significantly likely to be dissatisfied with their flight. The large difference may make a customer's loyalty a good predictor of satisfaction.

In the right-hand graph, we can observe that customers in the Business class are very likely to be satisfied with their flight, while customers in the Eco (Plus) classes are comparatively less likely to be satisfied with their flight. While Eco and Eco Plus lie more near the 50/50 mark, the comparative difference between their satisfaction and the Business class's satisfaction may make for a good point of prediction.

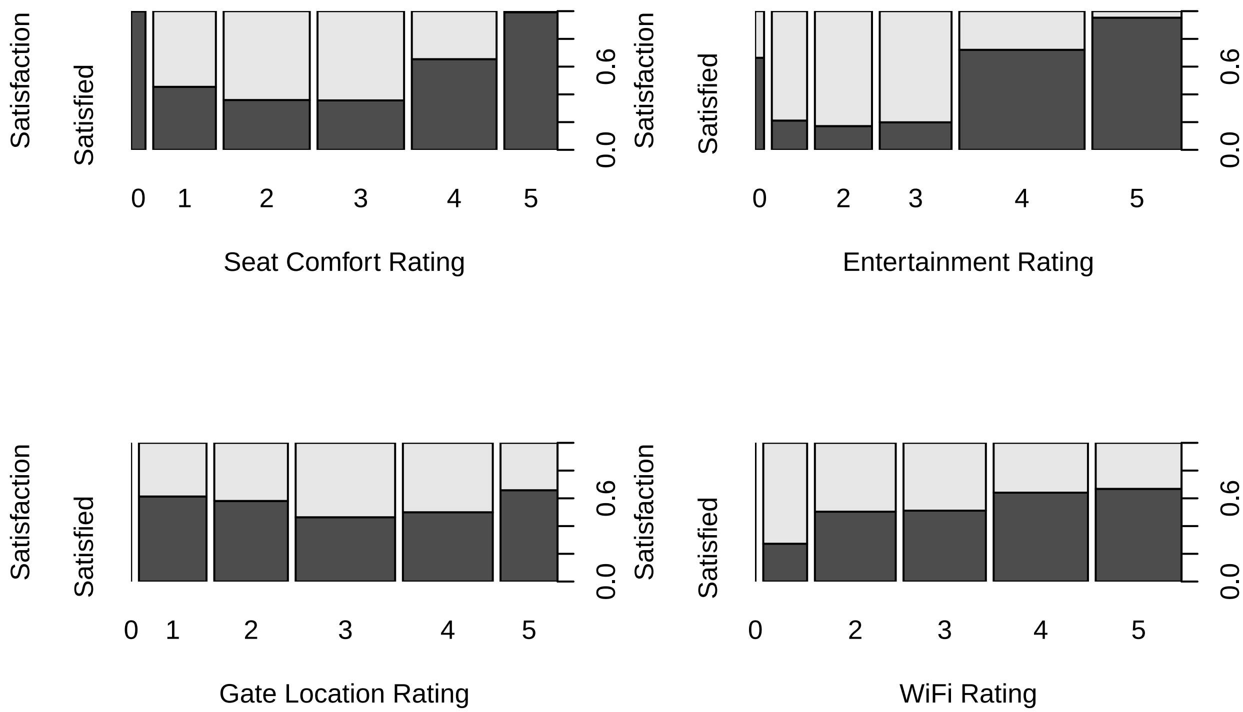




Finally, in the last four graphs below, we are seeking to determine if there is any correlation between the customer's review ratings and their satisfaction.

For obvious reasons, we can assume these will go hand-in-hand, but these graphs help show that generally, the lower the rating, the less likely people are to be satisfied, and the higher the rating, the more likely they are to be satisfied.

This is not true for *all* ratings, however. Such as the bottom-left graph, which implies that Gate Location has little effect on the customer's satisfaction with their flight.



Models

Logistic Regression

Model Training

```
# logistic regression model
glm <- glm(satisfaction~Gender+Customer.Type+Type.of.Travel+Class+Seat.comfort+Leg.room.service
    +Food.and.drink+Inflight.wifi.service+Inflight.entertainment+Departure.Arrival.time.convenient
    +Flight.Distance+Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes, data=train, family=binomial)

# summary
summary(glm)
```

```
##
## Call:
## glm(formula = satisfaction ~ Gender + Customer.Type + Type.of.Travel +
##     Class + Seat.comfort + Leg.room.service + Food.and.drink +
##     Inflight.wifi.service + Inflight.entertainment + Departure.Arrival.time.convenient +
##     Flight.Distance + Departure.Delay.in.Minutes + Arrival.Delay.in.Minutes,
##     family = binomial, data = train)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -4.1950  -0.3582   0.0287   0.3413   3.4845
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          3.334e+00  9.389e-01   3.551 0.000384 ***
## GenderMale                          -8.414e-01  2.201e-02 -38.237  < 2e-16 ***
## Customer.TypeLoyal                   2.112e+00  3.502e-02  60.305  < 2e-16 ***
## Type.of.TravelPersonal              -6.544e-01  3.130e-02 -20.910  < 2e-16 ***
## ClassEco                            -9.512e-01  2.882e-02 -33.007  < 2e-16 ***
## ClassEco Plus                       -1.108e+00  4.485e-02 -24.709  < 2e-16 ***
## Seat.comfort1                       -1.061e+01  5.757e-01 -18.426  < 2e-16 ***
## Seat.comfort2                       -1.107e+01  5.764e-01 -19.199  < 2e-16 ***
## Seat.comfort3                       -1.111e+01  5.764e-01 -19.270  < 2e-16 ***
## Seat.comfort4                       -9.780e+00  5.761e-01 -16.976  < 2e-16 ***
## Seat.comfort5                       -5.102e+00  5.824e-01  -8.761  < 2e-16 ***
## Leg.room.service1                   -4.501e-01  4.688e-01  -0.960 0.337005
## Leg.room.service2                    2.471e-01  4.685e-01   0.528 0.597826
## Leg.room.service3                    2.168e-01  4.684e-01   0.463 0.643517
## Leg.room.service4                    1.434e+00  4.680e-01   3.064 0.002186 **
## Leg.room.service5                    1.797e+00  4.680e-01   3.839 0.000123 ***
## Food.and.drink1                      1.810e+00  3.946e-01   4.587 4.50e-06 ***
## Food.and.drink2                      1.868e+00  3.950e-01   4.729 2.26e-06 ***
## Food.and.drink3                      1.996e+00  3.949e-01   5.055 4.30e-07 ***
## Food.and.drink4                      1.590e+00  3.945e-01   4.029 5.60e-05 ***
## Food.and.drink5                      1.744e+00  3.949e-01   4.416 1.00e-05 ***
## Inflight.wifi.service1               3.284e+00  7.385e-01   4.447 8.70e-06 ***
## Inflight.wifi.service2               3.957e+00  7.384e-01   5.359 8.37e-08 ***
## Inflight.wifi.service3               4.040e+00  7.384e-01   5.471 4.47e-08 ***
## Inflight.wifi.service4               4.345e+00  7.383e-01   5.885 3.98e-09 ***
## Inflight.wifi.service5               4.176e+00  7.384e-01   5.656 1.55e-08 ***
## Inflight.entertainment1             -1.115e+00  3.983e-01  -2.801 0.005100 **
## Inflight.entertainment2             -1.099e+00  3.982e-01  -2.760 0.005783 **
## Inflight.entertainment3             -1.222e+00  3.979e-01  -3.072 0.002128 **
## Inflight.entertainment4              8.158e-01  3.974e-01   2.053 0.040077 *
## Inflight.entertainment5              2.385e+00  3.984e-01   5.986 2.16e-09 ***
## Departure.Arrival.time.convenient1 -2.537e-02  7.378e-02  -0.344 0.730973
## Departure.Arrival.time.convenient2  8.186e-02  7.205e-02   1.136 0.255849
## Departure.Arrival.time.convenient3  4.257e-02  7.138e-02   0.596 0.550945
## Departure.Arrival.time.convenient4 -5.108e-01  6.722e-02  -7.600 2.97e-14 ***
## Departure.Arrival.time.convenient5 -1.550e+00  7.329e-02 -21.150  < 2e-16 ***
## Flight.Distance                     -5.911e-05  1.050e-05  -5.630 1.81e-08 ***
## Departure.Delay.in.Minutes           3.496e-03  1.063e-03   3.289 0.001007 **
## Arrival.Delay.in.Minutes            -8.597e-03  1.047e-03  -8.212  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142610  on 103588  degrees of freedom
## Residual deviance:  57425  on 103550  degrees of freedom
## AIC: 57503
##
## Number of Fisher Scoring iterations: 9
```

Model Predictions

```
##
## pred_glm       Dissatisfied Satisfied
##   Dissatisfied        10299      1546
##   Satisfied            1582     12471

## Accuracy:  0.879218472468916
```

kNN

Model Training

```
# kNN model
pred_kNN <- knn(train=train_scaled, test=test_scaled, cl=train_labels, k=7)
```

Model Predictions

```
##            pred_kNN
## results_kNN Dissatisfied Satisfied
##       FALSE         1363       942
##       TRUE         10939     12654

## Accuracy:  0.910996988184416
```

Decision Tree

Model Training

```
# decision tree model
tree <- tree(satisfaction~., data=train)

# summary
summary(tree)
```

```
##
## Classification tree:
## tree(formula = satisfaction ~ ., data = train)
## Variables actually used in tree construction:
## [1] "Inflight.entertainment"            "Seat.comfort"
## [3] "Departure.Arrival.time.convenient" "Customer.Type"
## [5] "Cleanliness"                       "Ease.of.Online.booking"
## [7] "Class"                             "Food.and.drink"
## Number of terminal nodes:  12
## Residual mean deviance:  0.5679 = 58820 / 103600
## Misclassification error rate: 0.1366 = 14149 / 103589
```

Note that, pruning the tree saw a consistent decrease in the model's accuracy.

Model Predictions

```
##
## pred_tree      Dissatisfied Satisfied
##   Dissatisfied        10121      1655
##   Satisfied            1760     12362

## Accuracy:  0.868136535639818
```

Analysis

Looking at the results of each algorithm, it's clear that kNN performed the best out of all of them.

Knowing how each of the models work, it makes sense that kNN performed the best on this data set, as the columns that use the 0-5 Rating scale are all similar to each other, and are likely classified similarly, honing in on its accuracy. Whereas, the Decision Tree model likely overfitted the data (explaining its comparative inaccuracy), while the Logistic Regression model likely underfitted the data. Despite that, I was able to get both models to give very good prediction accuracies. But will this scale well with other variations of the data? If we are to believe that the models did in fact overfit/underfit the data as previously described, then probably not. However, this may not be the case with kNN, as its classification of the data may transfer over well into other variations of the data.