Math156-Project

Huy Nguyen

5/22/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
```

Load data

```
train <- suppressMessages(suppressWarnings(read_csv("/Users/huynguyen/Downloads/Math 156/train.csv")))
train_no_id <- train[,-c(1,2)] #rid of index and id
test <- suppressMessages(suppressWarnings(read_csv("/Users/huynguyen/Downloads/Math 156/test.csv"))) #t
test_no_id <- train[,-c(1,2)]

#deal with categorical variables
train_no_id$Gender <- factor(train_no_id$Gender)
train_no_id$`Customer Type` <- factor(train_no_id$`Customer Type`)
train_no_id$`Type of Travel` <- factor(train_no_id$`Type of Travel`)
train_no_id$Class <- factor(train_no_id$Class)
train_no_id$`Inflight wifi service` <- factor(train_no_id$`Inflight wifi service`)
train_no_id$`Departure/Arrival time convenient` <- factor(train_no_id$`Departure/Arrival time convenien
train_no_id$`Ease of Online booking` <- factor(train_no_id$`Ease of Online booking`)
train_no_id$`Gate location` <- factor(train_no_id$`Gate location`)
train_no_id$`Food and drink` <- factor(train_no_id$`Food and drink`)
train_no_id$`Online boarding` <- factor(train_no_id$`Online boarding`)
train_no_id$`Seat comfort` <- factor(train_no_id$`Seat comfort`)
train_no_id$`Inflight entertainment` <- factor(train_no_id$`Inflight entertainment`)
train_no_id$`On-board service` <- factor(train_no_id$`On-board service`)
train_no_id$`Leg room service` <- factor(train_no_id$`Leg room service`)
train_no_id$`Baggage handling` <- factor(train_no_id$`Baggage handling`)
train_no_id$`Checkin service` <- factor(train_no_id$`Checkin service`)
train_no_id$`Inflight service` <- factor(train_no_id$`Inflight service`)
train_no_id$Cleanliness <- factor(train_no_id$Cleanliness)
```

```
test_no_id$Gender <- factor(test_no_id$Gender)
test_no_id$`Customer Type` <- factor(test_no_id$`Customer Type`)
test_no_id$`Type of Travel` <- factor(test_no_id$`Type of Travel`)
test_no_id$Class <- factor(test_no_id$Class)
test_no_id$`Inflight wifi service` <- factor(test_no_id$`Inflight wifi service`)
test_no_id$`Departure/Arrival time convenient` <- factor(test_no_id$`Departure/Arrival time convenient`
test_no_id$`Ease of Online booking` <- factor(test_no_id$`Ease of Online booking`)
test_no_id$`Gate location` <- factor(test_no_id$`Gate location`)
test_no_id$`Food and drink` <- factor(test_no_id$`Food and drink`)
test_no_id$`Online boarding` <- factor(test_no_id$`Online boarding`)
test_no_id$`Seat comfort` <- factor(test_no_id$`Seat comfort`)
test_no_id$`Inflight entertainment` <- factor(test_no_id$`Inflight entertainment`)
test_no_id$`On-board service` <- factor(test_no_id$`On-board service`)
test_no_id$`Leg room service` <- factor(test_no_id$`Leg room service`)
test_no_id$`Baggage handling` <- factor(test_no_id$`Baggage handling`)
test_no_id$`Checkin service` <- factor(test_no_id$`Checkin service`)
test_no_id$`Inflight service` <- factor(test_no_id$`Inflight service`)
test_no_id$Cleanliness <- factor(test_no_id$Cleanliness)
```

Preliminary Plots

```
length(is.na(train_no_id)) # Number of NAs in train
```

```
## [1] 2389792
```

```
length(is.na(test_no_id)) #Number f NAs in test
```

```
## [1] 2389792
```

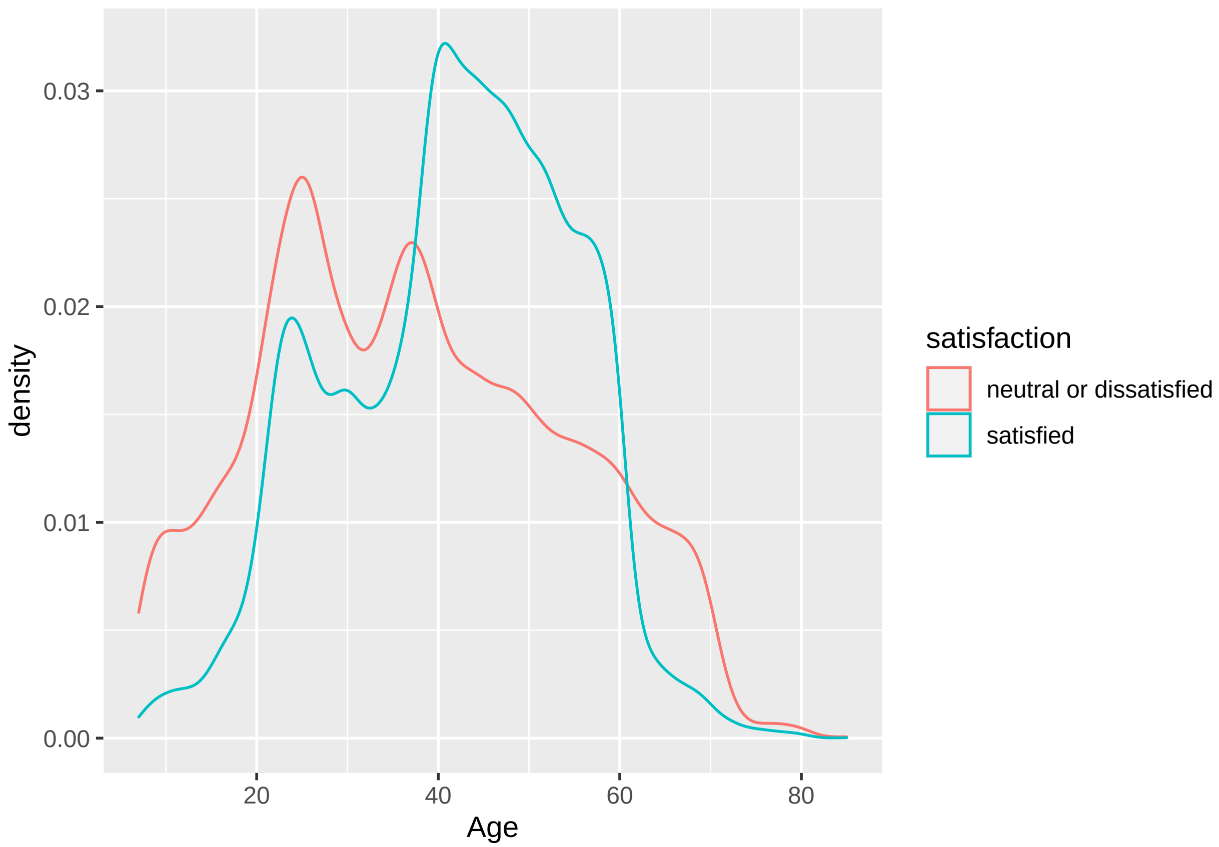
```
set.seed(123) #Set seed to 123
#plot()
```

```
# Numerical variables
library(gridExtra)
```

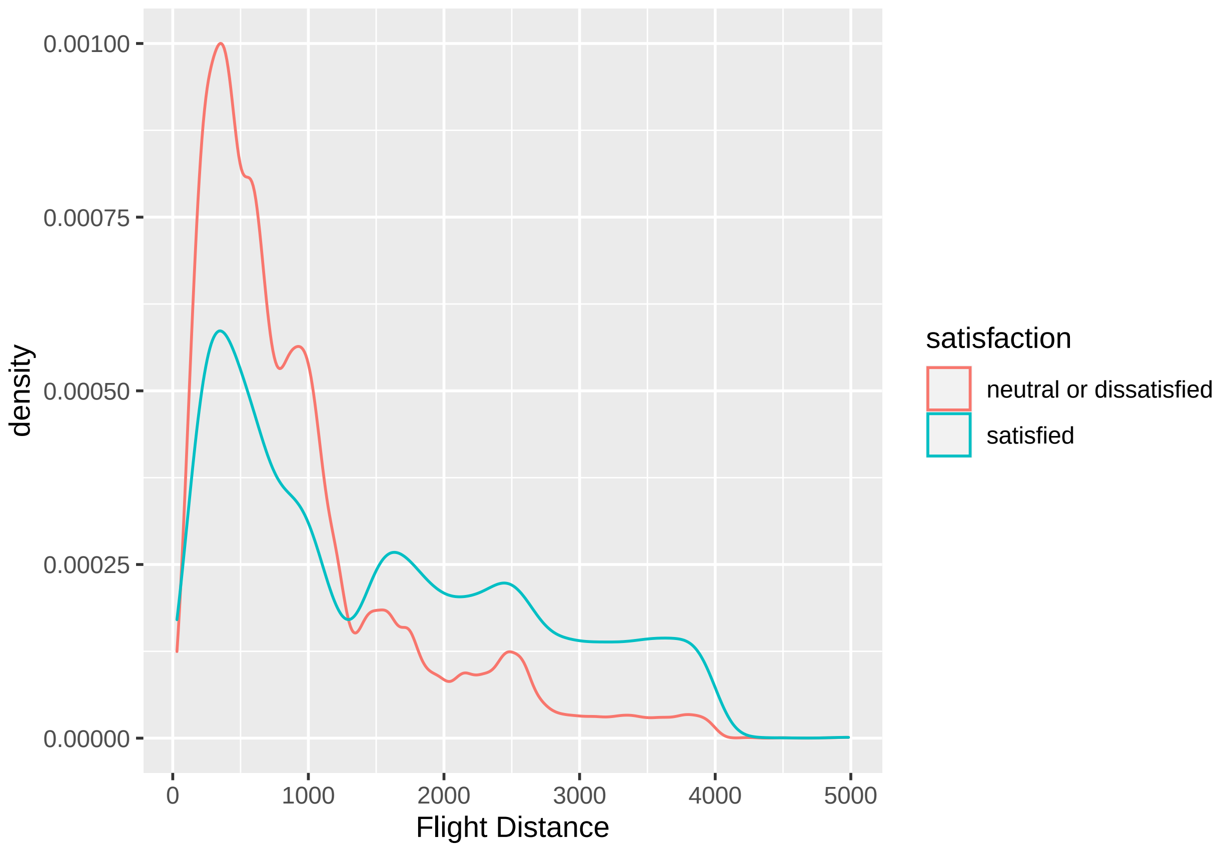
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

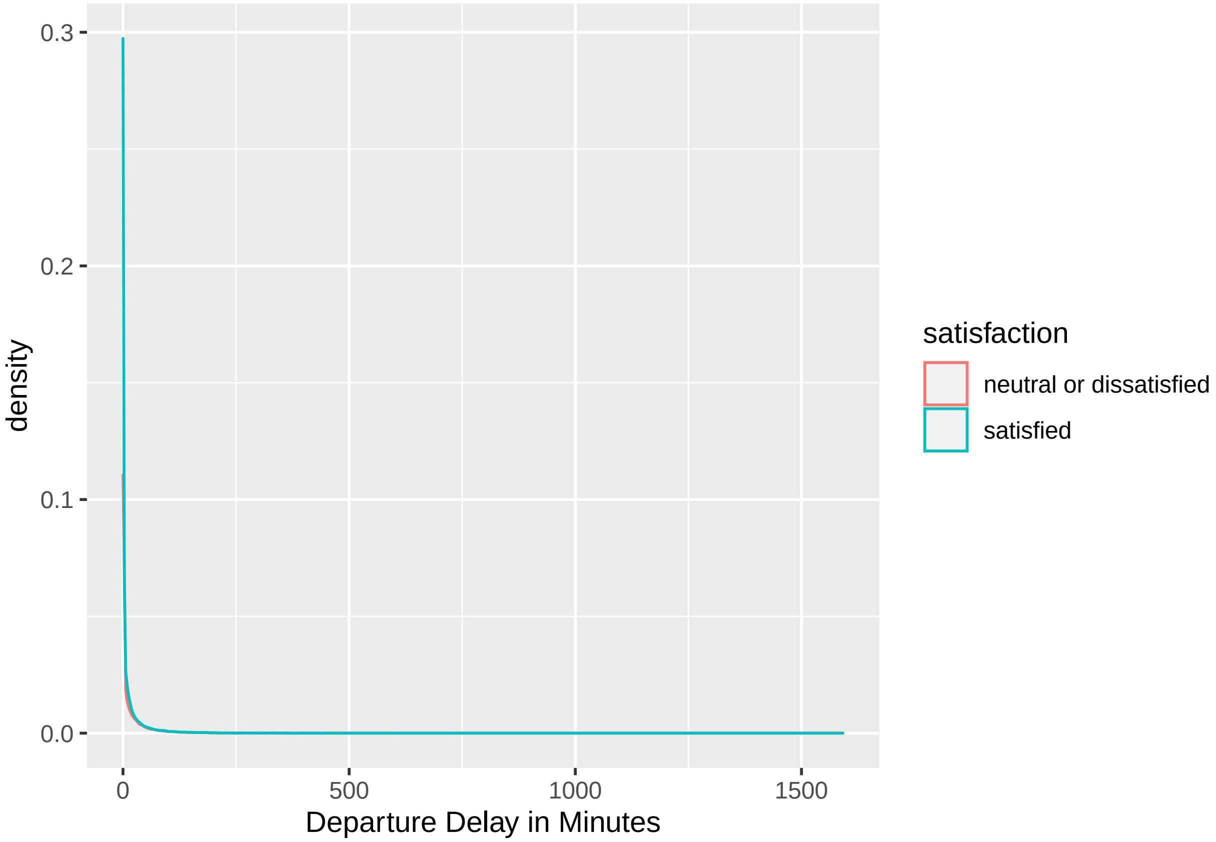
```
g1<-ggplot(train_no_id, aes(x = Age))+geom_density(aes(color = satisfaction)) #definitely in model
g2<-ggplot(train_no_id, aes(x = `Flight Distance`))+geom_density(aes(color = satisfaction)) # may not i
g3<-ggplot(train_no_id, aes(x = `Departure Delay in Minutes`))+geom_density(aes(color = satisfaction))
g4<-ggplot(train_no_id, aes(x = `Arrival Delay in Minutes`))+geom_density(aes(color = satisfaction)) #
#grid.arrange(g1,g2,ncol=2)
g1
```


0.03
0.02
0.01
0.00
density
20
40
60
80
Age
satisfaction
neutral or dissatisfied
satisfied

```
g2
```

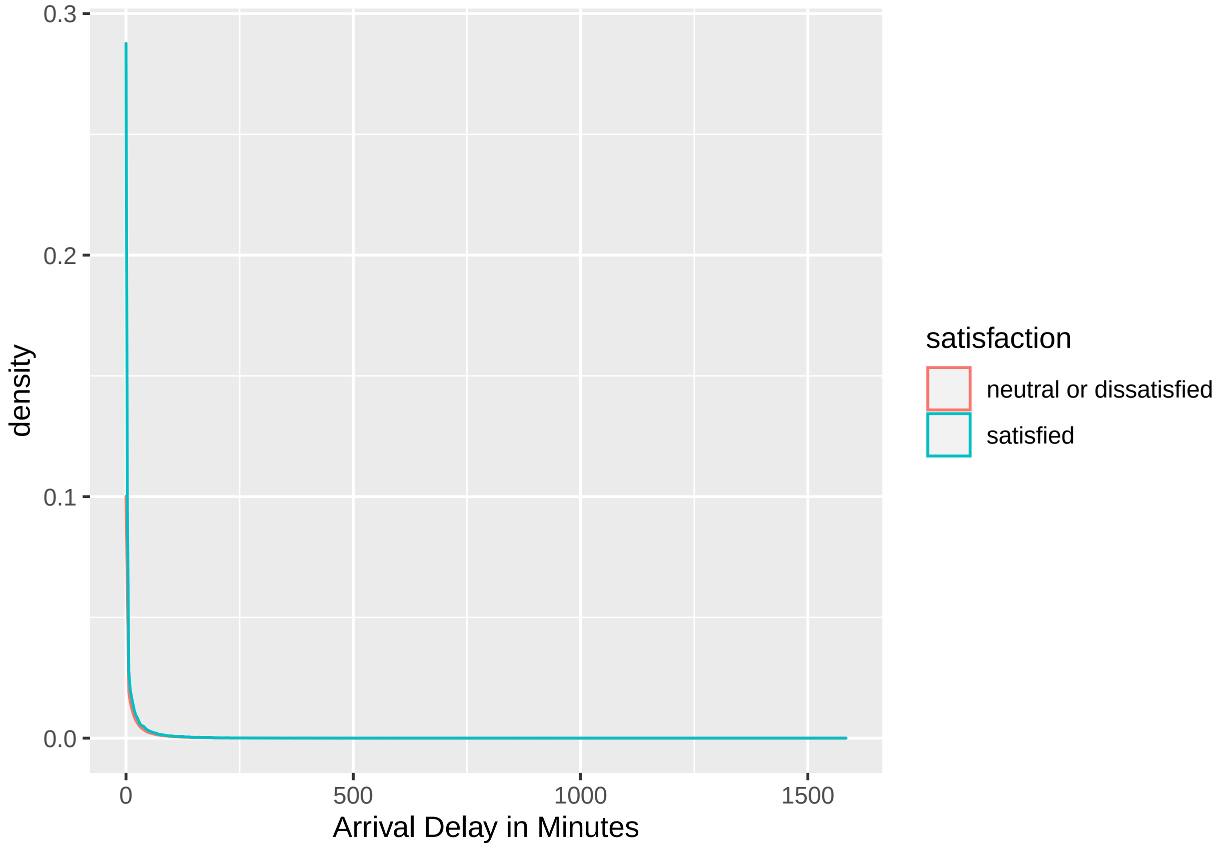

0.00100
0.00075
0.00050
0.00025
0.00000
density
0
1000
2000
3000
4000
5000
Flight Distance
satisfaction
neutral or dissatisfied
satisfied

```
g3
```

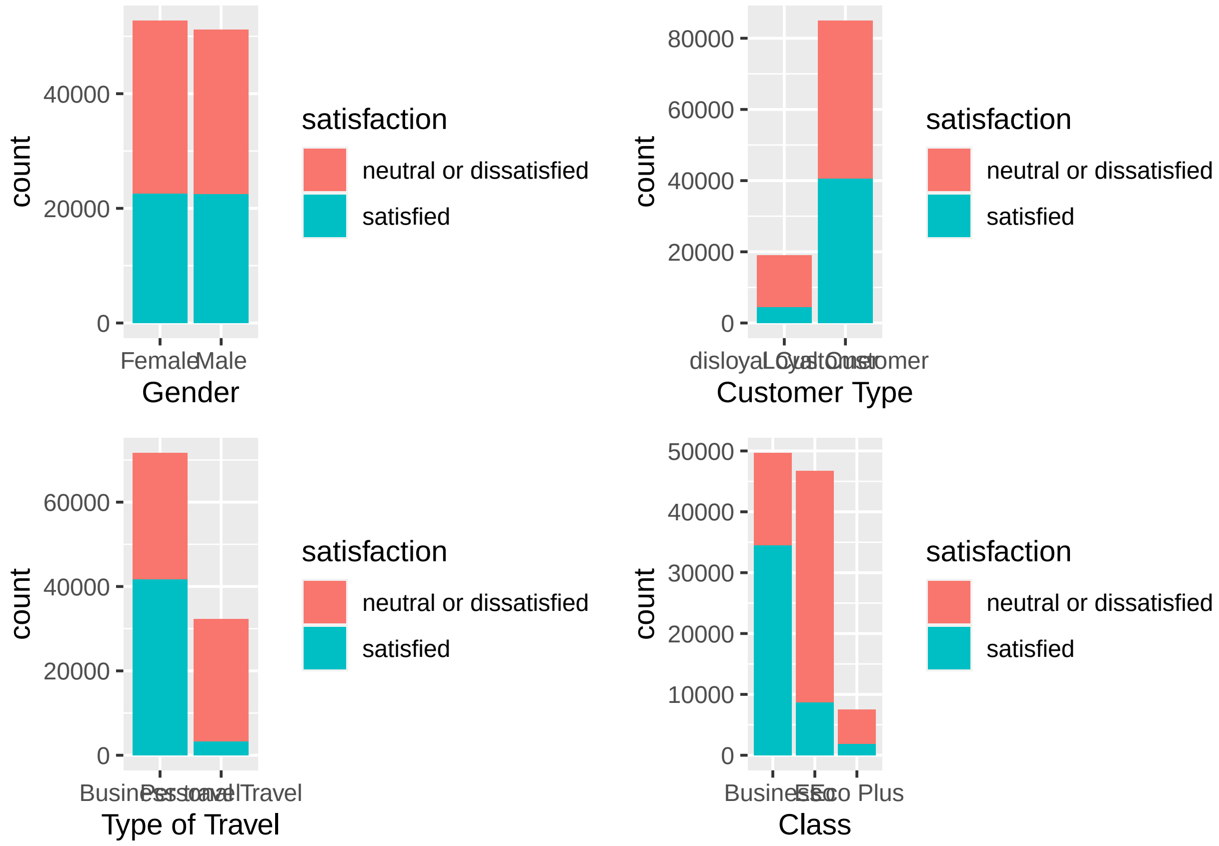


```
g4
```

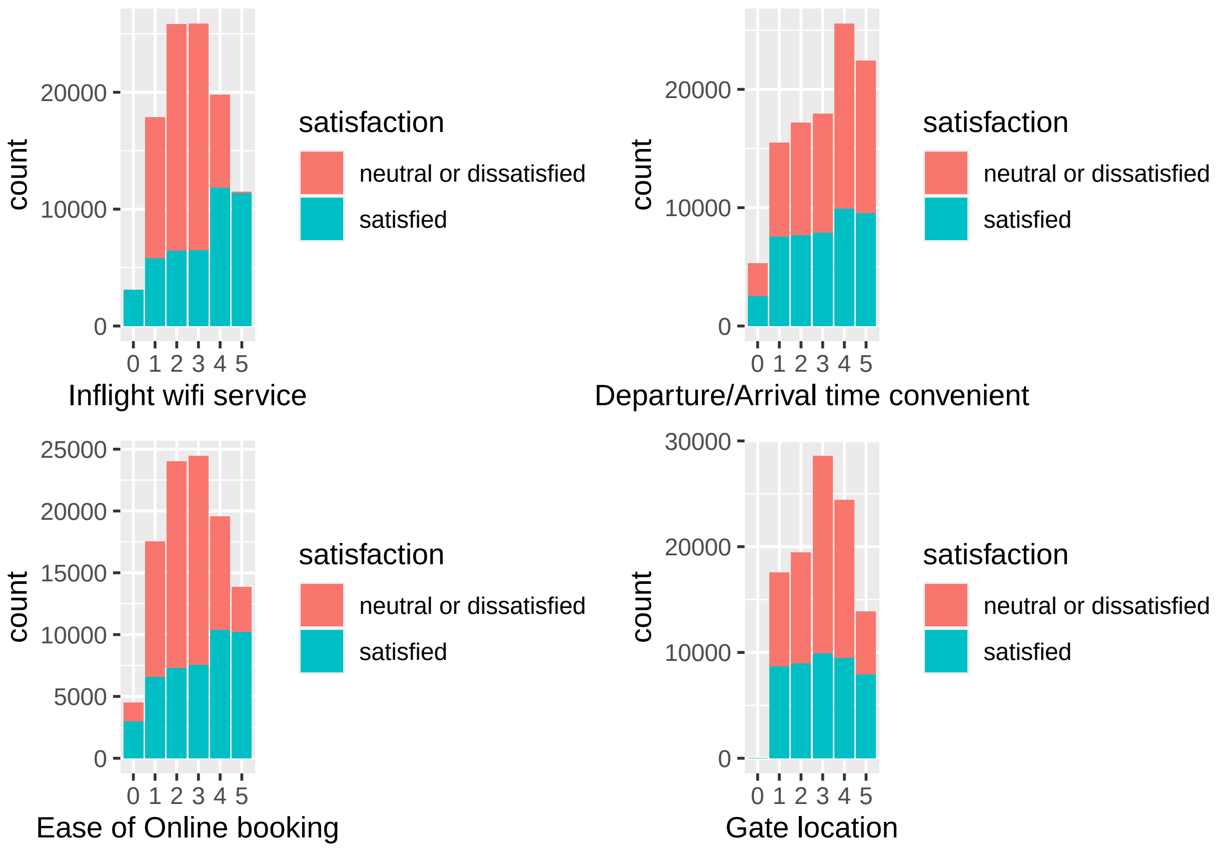
```
## Warning: Removed 310 rows containing non-finite values (stat_density).
```



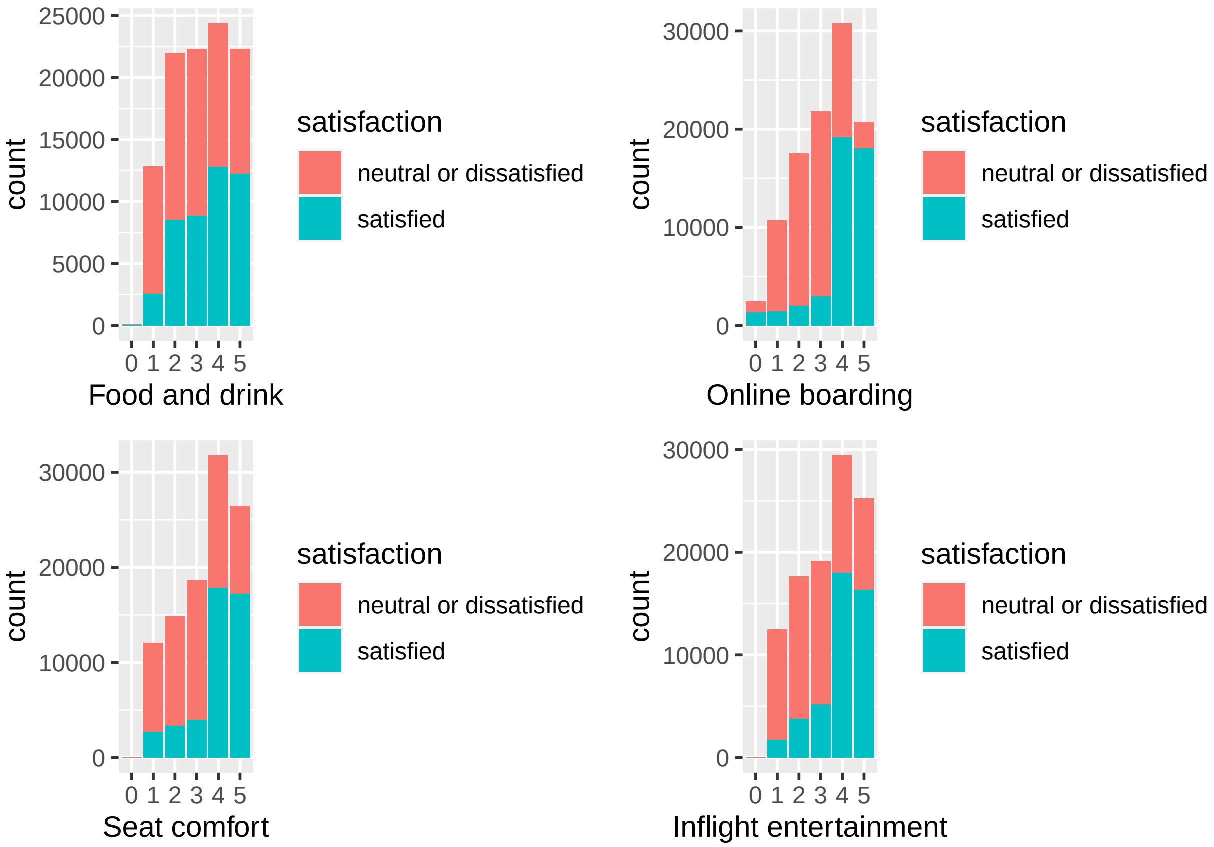
```
# categorical variable
g5<-ggplot(train_no_id, aes(x=Gender,fill=satisfaction)) + geom_bar() # not in model
g6<-ggplot(train_no_id, aes(x=`Customer Type`,fill=satisfaction)) + geom_bar() # in model
g7<-ggplot(train_no_id, aes(x=`Type of Travel`,fill=satisfaction)) + geom_bar() # in model
g8<-ggplot(train_no_id, aes(x=Class,fill=satisfaction)) + geom_bar() # in model
g9<-ggplot(train_no_id, aes(x=`Inflight wifi service`,fill=satisfaction)) + geom_bar() # in model
g10<-ggplot(train_no_id, aes(x=`Departure/Arrival time convenient`,fill=satisfaction)) + geom_bar() # i
g11<-ggplot(train_no_id, aes(x=`Ease of Online booking`,fill=satisfaction)) + geom_bar() # in model
g12<-ggplot(train_no_id, aes(x=`Gate location`,fill=satisfaction)) + geom_bar() # in model
g13<-ggplot(train_no_id, aes(x=`Food and drink`,fill=satisfaction)) + geom_bar() # in model
g14<-ggplot(train_no_id, aes(x=`Online boarding`,fill=satisfaction)) + geom_bar() # in model
g15<-ggplot(train_no_id, aes(x=`Seat comfort`,fill=satisfaction)) + geom_bar() # in model
g16<-ggplot(train_no_id, aes(x=`Inflight entertainment`,fill=satisfaction)) + geom_bar() # in model
g17<-ggplot(train_no_id, aes(x=`On-board service`,fill=satisfaction)) + geom_bar() # in model
g18<-ggplot(train_no_id, aes(x=`Leg room service`,fill=satisfaction)) + geom_bar() # in model
g19<-ggplot(train_no_id, aes(x=`Baggage handling`,fill=satisfaction)) + geom_bar() # in model
g20<-ggplot(train_no_id, aes(x=`Checkin service`,fill=satisfaction)) + geom_bar() # in model
g21<-ggplot(train_no_id, aes(x=`Inflight service`,fill=satisfaction)) + geom_bar() # in model
g22<-ggplot(train_no_id, aes(x=Cleanliness,fill=satisfaction)) + geom_bar() # in model
grid.arrange(g5,g6,g7,g8,ncol=2)
```



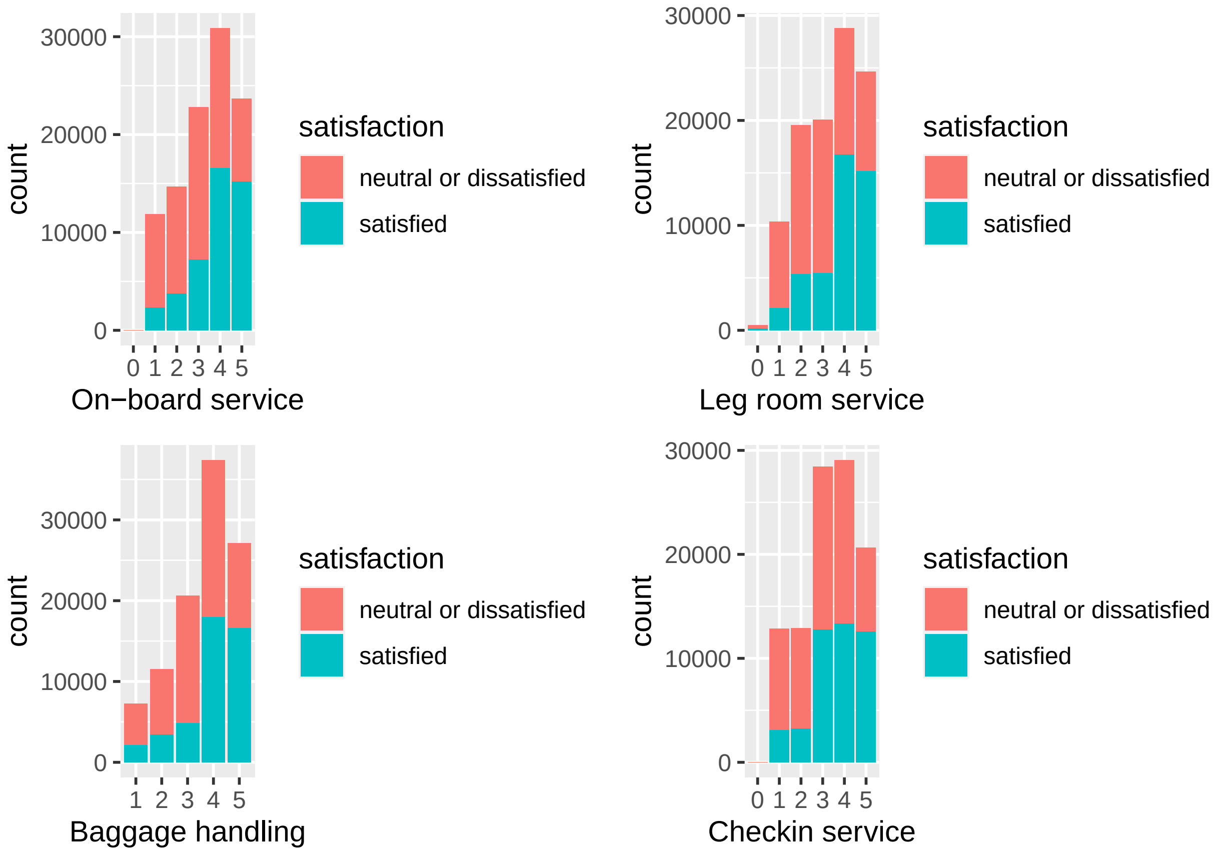
```
grid.arrange(g9,g10,g11,g12,ncol=2)
```



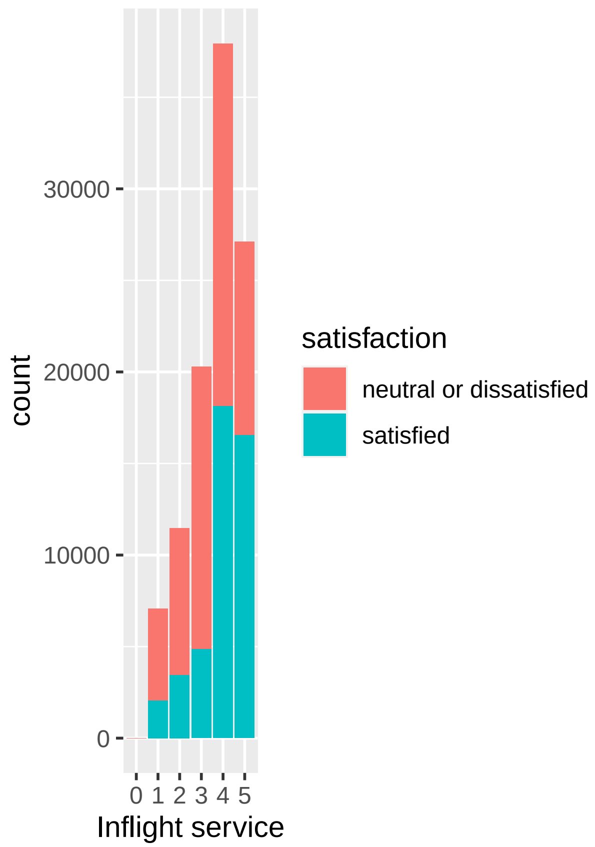
```
grid.arrange(g13,g14,g15,g16,ncol=2)
```

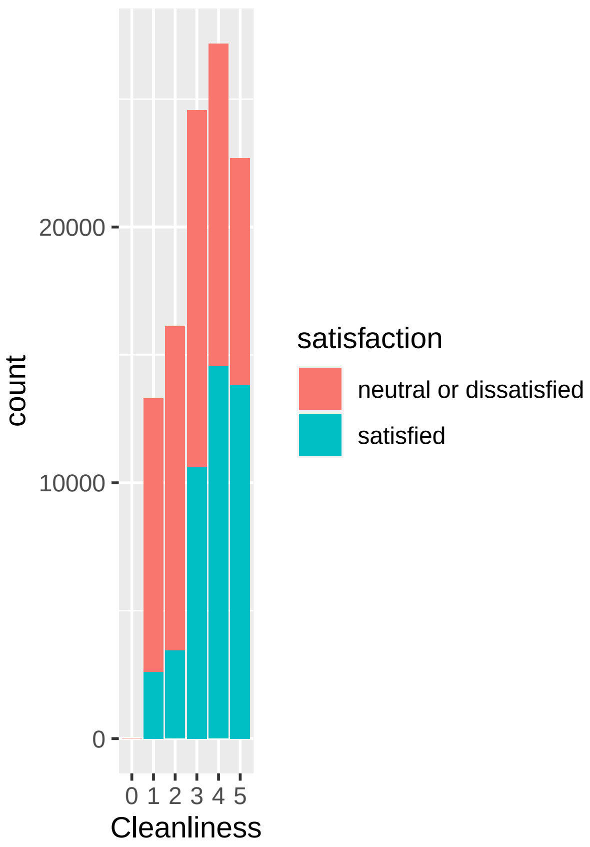


```
grid.arrange(g17,g18,g19,g20,ncol=2)
```



```
grid.arrange(g21,g22,ncol=2)
```





Logistic Model

```
library(leaps)
train_removeDelay <- train_no_id[-c(21,22)]
train_removeDelay <- train_removeDelay[rowSums(is.na(train_removeDelay)) == 0, ]
#Remove Delays since not important + No more NAs
test_removeDelay <- test_no_id[-c(21,22)]
#train_NoNA <- train_no_id[rowSums(is.na(train_no_id)) == 0, ]
# mfull<-glm(as.factor(satisfaction)~.,data=train_removeDelay,family=binomial())
# bBIC=step(mfull,direction="backward",log=nrow(train_removeDelay)) #Backward BIC
# bBIC
glm.mod <- glm(as.factor(satisfaction) ~ ., data = train_removeDelay,
               family=binomial())
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(glm.mod)
```

```
##
## Call:
## glm(formula = as.factor(satisfaction) ~ ., family = binomial(),
##     data = train_removeDelay)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6734  -0.2140  -0.0474   0.1356   4.4193
```

```
##
## Coefficients: (3 not defined because of singularities)
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          4.937e+00  9.961e+03   0.000 0.999605
## GenderMale                           4.445e-02  2.723e-02   1.632 0.102639
## ‘Customer Type‘Loyal Customer        3.341e+00  4.945e-02  67.558  < 2e-16 ***
## Age                                 -1.961e-03  1.013e-03  -1.936 0.052851 .
## ‘Type of Travel‘Personal Travel     -4.253e+00  5.493e-02 -77.420  < 2e-16 ***
## ClassEco                            -6.342e-01  3.714e-02 -17.073  < 2e-16 ***
## ClassEco Plus                       -8.484e-01  6.035e-02 -14.058  < 2e-16 ***
## ‘Flight Distance‘                    7.269e-06  1.530e-05   0.475 0.634743
## ‘Inflight wifi service‘1            -2.413e+01  8.832e+01  -0.273 0.784728
## ‘Inflight wifi service‘2            -2.437e+01  8.832e+01  -0.276 0.782568
## ‘Inflight wifi service‘3            -2.442e+01  8.832e+01  -0.276 0.782186
## ‘Inflight wifi service‘4            -2.287e+01  8.832e+01  -0.259 0.795713
## ‘Inflight wifi service‘5            -1.731e+01  8.832e+01  -0.196 0.844609
## ‘Departure/Arrival time convenient‘1  3.126e-01  9.296e-02   3.362 0.000773 ***
## ‘Departure/Arrival time convenient‘2  4.220e-01  8.956e-02   4.711 2.46e-06 ***
## ‘Departure/Arrival time convenient‘3  2.413e-01  8.635e-02   2.794 0.005205 **
## ‘Departure/Arrival time convenient‘4 -6.831e-01  7.736e-02  -8.830  < 2e-16 ***
## ‘Departure/Arrival time convenient‘5 -9.216e-01  8.495e-02 -10.849  < 2e-16 ***
## ‘Ease of Online booking‘1            3.073e+00  9.164e-01   3.354 0.000797 ***
## ‘Ease of Online booking‘2            3.002e+00  9.164e-01   3.275 0.001055 **
## ‘Ease of Online booking‘3            3.501e+00  9.162e-01   3.821 0.000133 ***
## ‘Ease of Online booking‘4            4.360e+00  9.160e-01   4.760 1.94e-06 ***
## ‘Ease of Online booking‘5            3.730e+00  9.163e-01   4.071 4.68e-05 ***
## ‘Gate location‘1                    -1.883e+01  6.523e+03  -0.003 0.997697
## ‘Gate location‘2                    -1.874e+01  6.523e+03  -0.003 0.997707
## ‘Gate location‘3                    -1.892e+01  6.523e+03  -0.003 0.997686
## ‘Gate location‘4                    -1.919e+01  6.523e+03  -0.003 0.997653
## ‘Gate location‘5                    -1.938e+01  6.523e+03  -0.003 0.997629
## ‘Food and drink‘1                    1.427e-01  1.715e+00   0.083 0.933701
## ‘Food and drink‘2                    4.266e-01  1.715e+00   0.249 0.803556
## ‘Food and drink‘3                    3.005e-01  1.715e+00   0.175 0.860887
## ‘Food and drink‘4                    3.259e-01  1.715e+00   0.190 0.849288
## ‘Food and drink‘5                    2.157e-01  1.715e+00   0.126 0.899923
## ‘Online boarding‘1                  -3.666e+00  9.196e-01  -3.986 6.71e-05 ***
## ‘Online boarding‘2                  -3.580e+00  9.195e-01  -3.894 9.88e-05 ***
## ‘Online boarding‘3                  -3.804e+00  9.192e-01  -4.139 3.49e-05 ***
## ‘Online boarding‘4                  -2.154e+00  9.188e-01  -2.345 0.019051 *
## ‘Online boarding‘5                  -9.335e-01  9.191e-01  -1.016 0.309744
## ‘Seat comfort‘1                      2.143e+01  6.523e+03   0.003 0.997379
## ‘Seat comfort‘2                      2.089e+01  6.523e+03   0.003 0.997444
## ‘Seat comfort‘3                      1.984e+01  6.523e+03   0.003 0.997573
## ‘Seat comfort‘4                      2.054e+01  6.523e+03   0.003 0.997487
## ‘Seat comfort‘5                      2.138e+01  6.523e+03   0.003 0.997385
## ‘Inflight entertainment‘1            3.920e+01  1.519e+03   0.026 0.979418
## ‘Inflight entertainment‘2            3.997e+01  1.519e+03   0.026 0.979012
## ‘Inflight entertainment‘3            4.078e+01  1.519e+03   0.027 0.978588
## ‘Inflight entertainment‘4            4.049e+01  1.519e+03   0.027 0.978741
## ‘Inflight entertainment‘5            3.970e+01  1.519e+03   0.026 0.979155
## ‘On-board service‘1                 -2.286e+01  4.053e+03  -0.006 0.995499
## ‘On-board service‘2                 -2.273e+01  4.053e+03  -0.006 0.995524
## ‘On-board service‘3                 -2.218e+01  4.053e+03  -0.005 0.995633
## ‘On-board service‘4                  -2.210e+01  4.053e+03  -0.005 0.995649
## ‘On-board service‘5                  -2.158e+01  4.053e+03  -0.005 0.995752
## ‘Leg room service‘1                  -2.420e+00  9.604e-01  -2.520 0.011750 *
## ‘Leg room service‘2                  -2.143e+00  9.599e-01  -2.233 0.025573 *
## ‘Leg room service‘3                  -2.280e+00  9.598e-01  -2.375 0.017529 *
## ‘Leg room service‘4                  -1.593e+00  9.599e-01  -1.660 0.097012 .
## ‘Leg room service‘5                  -1.418e+00  9.596e-01  -1.477 0.139620
## ‘Baggage handling‘2                  -2.309e-01  7.576e-02  -3.048 0.002305 **
## ‘Baggage handling‘3                  -8.651e-01  7.070e-02 -12.237  < 2e-16 ***
## ‘Baggage handling‘4                  -2.860e-01  6.870e-02  -4.164 3.13e-05 ***
## ‘Baggage handling‘5                   4.733e-01  7.312e-02   6.473 9.58e-11 ***
## ‘Checkin service‘1                   -1.415e+00  5.412e-02 -26.145  < 2e-16 ***
## ‘Checkin service‘2                   -1.232e+00  5.385e-02 -22.872  < 2e-16 ***
## ‘Checkin service‘3                   -7.155e-01  4.333e-02 -16.515  < 2e-16 ***
## ‘Checkin service‘4                   -7.491e-01  4.317e-02 -17.353  < 2e-16 ***
## ‘Checkin service‘5                           NA         NA      NA       NA
## ‘Inflight service‘1                  -5.495e-01  7.583e-02  -7.247 4.27e-13 ***
## ‘Inflight service‘2                  -7.498e-01  6.886e-02 -10.889  < 2e-16 ***
## ‘Inflight service‘3                  -1.424e+00  5.718e-02 -24.896  < 2e-16 ***
## ‘Inflight service‘4                  -7.035e-01  4.496e-02 -15.647  < 2e-16 ***
## ‘Inflight service‘5                          NA         NA      NA       NA
## Cleanliness1                         -9.731e-01  7.466e-02 -13.033  < 2e-16 ***
## Cleanliness2                         -9.368e-01  7.258e-02 -12.907  < 2e-16 ***
## Cleanliness3                         -4.387e-01  6.096e-02  -7.196 6.18e-13 ***
## Cleanliness4                         -5.902e-01  5.979e-02  -9.871  < 2e-16 ***
## Cleanliness5                                 NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142189  on 103903  degrees of freedom
## Residual deviance:  37168  on 103830  degrees of freedom
## AIC: 37316
##
## Number of Fisher Scoring iterations: 17
```

Predict on test model:

```
test_feat <- test_removeDelay[,-c(21)]
dim(test_feat) # 103904 rows
```

```
## [1] 103904     20
```

```
test_y <- test_removeDelay[,c(21)]
predtest <- predict(glm.mod,test_feat,type="response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
glm.predtest=rep("satisfied",103904)
glm.predtest[predtest<0.5]="neutral or dissatisfied"
mean(glm.predtest == test_y)
```

```
## [1] 0.9339775
```