HW2_2082014_Kim Eunsim_datamining

2024-11-21

```
a2=read.csv("/Users/eunsimkim/Ewha/24_2/datamining/a2_data.csv")
head(a2)
```

```
##   Education JoiningYear      City PaymentTier Age Gender EverBenched
## 1 Bachelors        2017 Bangalore           3  34   Male          No
## 2 Bachelors        2013      Pune           1  28 Female          No
## 3 Bachelors        2014 New Delhi           3  38 Female          No
## 4   Masters        2016 Bangalore           3  27   Male          No
## 5   Masters        2017      Pune           3  24   Male         Yes
## 6 Bachelors        2016 Bangalore           3  22   Male          No
##   ExperienceInCurrentDomain LeaveOrNot
## 1                         0          0
## 2                         3          1
## 3                         2          0
## 4                         5          1
## 5                         2          1
## 6                         0          0
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

#(1)

```
#year
a2$year=2019-a2$JoiningYear
head(a2$year)
```

```
## [1] 2 6 5 3 2 3
```

```
#omit missing data 80% for training data
set.seed(123)
train_index=sample(1:nrow(a2),size=0.8*nrow(a2))
a2_train=a2[train_index,]
a2_test=a2[-train_index,]
```

#(2)

```
library(rsample) #data sampling
library(randomForest)
```

```
## randomForest 4.7-1.2
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```r
library(ranger)
```

```
##
## Attaching package: 'ranger'
```

```
## The following object is masked from 'package:randomForest':
##
##     importance
```

```r
library(caret)
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

```
## Loading required package: lattice
```

```r
library('AmesHousing')
library(xgboost) #
library(pdp) #visualization
library(vtreat)
```

```
## Loading required package: wrapr
```

```r
library(gbm)
```

```
## Loaded gbm 2.2.2
```

```
## This version of gbm is no longer under development. Consider transitioning to gbm3, https://github.c
```

```r
library(ggplot2)
```

```r
#split data



dim(a2_test)
```

```
## [1] 931  10
```

```r
dim(a2_train)
```

```
## [1] 3722   10
```

```r
head(a2_train)
```

```
##      Education JoiningYear      City PaymentTier Age Gender EverBenched
## 2463 Bachelors        2015      Pune           3  28   Male          No
## 2511 Bachelors        2014 Bangalore           3  29   Male          No
## 2227   Masters        2017 New Delhi           3  26   Male          No
## 526  Bachelors        2015 Bangalore           3  25   Male          No
## 4291   Masters        2017 New Delhi           2  31   Male          No
## 2986 Bachelors        2013 Bangalore           3  26   Male          No
##      ExperienceInCurrentDomain LeaveOrNot year
## 2463                         1          0    4
## 2511                         1          0    5
```

```
## 2227                          4          1    2
## 526                           3          0    4
## 4291                          4          0    2
## 2986                          4          0    6
```

#(3)

```
#Fit with binorm
model=glm(a2_train$LeaveOrNot~.,data=a2_train,family=binomial)
```

#(4) Train data.

```
#missclassification rate

a2_pred=predict(model,newdata=a2_train,type="response")
pred_class=ifelse(a2_pred>0.5,1,0)
actual_class=a2_train$LeaveOrNot

#missclassification rate

mean(pred_class!=actual_class)
```

```
## [1] 0.2646427
```

```
#so, the missclassification rate is 26%
```

#(5)

```
a2_pred1=predict(model,newdata=a2_test,type="response")
pred_class=ifelse(a2_pred1>0.5,1,0)
actual_test=a2_test$LeaveOrNot
```

```
#confusion matrix

table(Trueobserved=actual_test,Predicted=pred_class)
```

```
##             Predicted
## Trueobserved   0   1
##            0 537  58
##            1 208 128
```

#(6)

```
#sensitivity(TP/(TP+FN))
sensitivity=537/(537+208)
sensitivity
```

```
## [1] 0.7208054
```

```
#precision=TP/TP+FP
precision=537/(537+58)
precision
```

```
## [1] 0.902521
```

#(7)

```
#F1 score
F1=2*((precision*sensitivity)/(precision+sensitivity))
F1
```

```
## [1] 0.8014925
```

#(8)

#(9) Random Forest model

```
library(randomForest)
```

```
#random forest model
rf=randomForest(LeaveOrNot~.,data=a2_train,ntree=1000)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values.  Are you sure you want to do regression?
```

```
summary(rf)
```

```
##                 Length Class  Mode
## call               4   -none- call
## type               1   -none- character
## predicted       3722   -none- numeric
## mse             1000   -none- numeric
## rsq             1000   -none- numeric
## oob.times       3722   -none- numeric
## importance         9   -none- numeric
## importanceSD       0   -none- NULL
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            11   -none- list
## coefs              0   -none- NULL
## y               3722   -none- numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

```
rf_pred=predict(rf,newdata=a2_test)
pred_class=ifelse(rf_pred>0.5,1,0)
actual_test=a2_test$LeaveOrNot
table(True=a2_test$LeaveOrNot,Predicted=pred_class)
```

```
##     Predicted
## True   0   1
##    0 563  32
##    1  98 238
```

#(11)

```
#sensitivity
sen_rf=562/(562+99)
prec_rf=562/(562+33)
print(sen_rf)
```

```
## [1] 0.8502269
```

```
print(prec_rf)
```

```
## [1] 0.9445378
```

```r
f1_rf=2*((sen_rf*prec_rf)/(sen_rf+prec_rf))
print(f1_rf)
```

```
## [1] 0.8949045
```

#(12) F1-score #f1 score randomforest is better

#2.

#1, dimension

```r
student=read.csv("/Users/eunsimkim/Downloads/student+performance/student/student-mat.csv",sep=";")
dim(student)
```

```
## [1] 395  33
```

```r
print(colSums(is.na(student))) #no na data
```

```
##     school        sex        age    address    famsize    Pstatus       Medu
##          0          0          0          0          0          0          0
##       Fedu       Mjob       Fjob     reason   guardian traveltime  studytime
##          0          0          0          0          0          0          0
##   failures  schoolsup     famsup       paid activities    nursery     higher
##          0          0          0          0          0          0          0
##   internet   romantic     famrel   freetime      goout       Dalc       Walc
##          0          0          0          0          0          0          0
##     health   absences         G1         G2         G3
##          0          0          0          0          0
```

```r
#Yes, or No
student$famsup=ifelse(student$famsup=="Yes",1,0)
head(student)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes      0   no         no
## 2   father          1         2        0        no      0   no         no
## 3   mother          1         2        3       yes      0  yes         no
## 4   mother          1         3        0        no      0  yes        yes
## 5   father          1         2        0        no      0  yes         no
## 6   mother          1         2        0        no      0  yes        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3
## 1        6  5  6  6
## 2        4  5  5  6
## 3       10  7  8 10
```

```
## 4        2 15 14 15
## 5        4  6 10 10
## 6       10 15 15 15
```

```
head(student)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes      0   no         no
## 2   father          1         2        0        no      0   no         no
## 3   mother          1         2        3       yes      0  yes         no
## 4   mother          1         3        0        no      0  yes        yes
## 5   father          1         2        0        no      0  yes         no
## 6   mother          1         2        0        no      0  yes        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3
## 1        6  5  6  6
## 2        4  5  5  6
## 3       10  7  8 10
## 4        2 15 14 15
## 5        4  6 10 10
## 6       10 15 15 15
```

```
student$reason
```

```
##   [1] "course"     "course"     "other"      "home"       "home"
##   [6] "reputation" "home"       "home"       "home"       "home"
##  [11] "reputation" "reputation" "course"     "course"     "home"
##  [16] "home"       "reputation" "reputation" "course"     "home"
##  [21] "reputation" "other"      "course"     "reputation" "course"
##  [26] "home"       "home"       "other"      "home"       "home"
##  [31] "home"       "reputation" "course"     "course"     "home"
##  [36] "other"      "home"       "reputation" "course"     "reputation"
##  [41] "home"       "home"       "course"     "course"     "course"
##  [46] "course"     "home"       "reputation" "home"       "other"
##  [51] "course"     "other"      "other"      "course"     "other"
##  [56] "other"      "reputation" "reputation" "home"       "course"
##  [61] "other"      "course"     "reputation" "home"       "reputation"
##  [66] "course"     "reputation" "course"     "reputation" "reputation"
##  [71] "reputation" "course"     "reputation" "reputation" "home"
##  [76] "home"       "course"     "reputation" "home"       "course"
##  [81] "course"     "home"       "reputation" "home"       "home"
##  [86] "reputation" "course"     "reputation" "reputation" "reputation"
##  [91] "home"       "reputation" "home"       "home"       "reputation"
```

```
##  [96] "home"       "reputation" "course"     "reputation" "course"
## [101] "other"      "other"      "course"     "home"       "course"
## [106] "reputation" "course"     "home"       "home"       "other"
## [111] "course"     "reputation" "home"       "course"     "reputation"
## [116] "course"     "reputation" "home"       "course"     "reputation"
## [121] "course"     "home"       "course"     "course"     "home"
## [126] "home"       "home"       "course"     "reputation" "course"
## [131] "course"     "course"     "course"     "course"     "course"
## [136] "course"     "course"     "course"     "course"     "course"
## [141] "course"     "reputation" "course"     "course"     "home"
## [146] "course"     "home"       "course"     "course"     "course"
## [151] "course"     "course"     "reputation" "home"       "course"
## [156] "course"     "reputation" "course"     "course"     "course"
## [161] "course"     "course"     "course"     "course"     "course"
## [166] "course"     "course"     "home"       "home"       "reputation"
## [171] "course"     "reputation" "reputation" "home"       "reputation"
## [176] "course"     "reputation" "reputation" "other"      "course"
## [181] "home"       "home"       "reputation" "reputation" "reputation"
## [186] "other"      "other"      "course"     "reputation" "home"
## [191] "course"     "course"     "other"      "reputation" "home"
## [196] "course"     "home"       "home"       "home"       "reputation"
## [201] "home"       "reputation" "course"     "reputation" "reputation"
## [206] "home"       "course"     "other"      "home"       "reputation"
## [211] "reputation" "home"       "reputation" "home"       "other"
## [216] "reputation" "reputation" "home"       "home"       "course"
## [221] "reputation" "reputation" "other"      "home"       "home"
## [226] "reputation" "course"     "reputation" "course"     "course"
## [231] "reputation" "course"     "reputation" "reputation" "home"
## [236] "reputation" "home"       "home"       "course"     "reputation"
## [241] "course"     "course"     "course"     "course"     "course"
## [246] "course"     "course"     "other"      "course"     "other"
## [251] "course"     "reputation" "other"      "course"     "course"
## [256] "course"     "reputation" "reputation" "home"       "course"
## [261] "home"       "course"     "course"     "home"       "home"
## [266] "reputation" "other"      "reputation" "reputation" "reputation"
## [271] "home"       "reputation" "home"       "home"       "reputation"
## [276] "course"     "home"       "home"       "reputation" "course"
## [281] "home"       "home"       "reputation" "home"       "course"
## [286] "reputation" "other"      "reputation" "reputation" "reputation"
## [291] "home"       "reputation" "reputation" "reputation" "reputation"
## [296] "home"       "reputation" "home"       "reputation" "home"
## [301] "home"       "home"       "reputation" "reputation" "home"
## [306] "reputation" "course"     "reputation" "reputation" "reputation"
## [311] "home"       "other"      "course"     "reputation" "home"
## [316] "reputation" "course"     "course"     "course"     "course"
## [321] "course"     "course"     "course"     "course"     "home"
## [326] "course"     "reputation" "course"     "course"     "course"
## [331] "course"     "course"     "home"       "home"       "course"
## [336] "course"     "home"       "home"       "home"       "home"
## [341] "home"       "home"       "home"       "home"       "course"
## [346] "other"      "course"     "course"     "reputation" "course"
## [351] "home"       "course"     "course"     "home"       "home"
## [356] "course"     "other"      "reputation" "home"       "course"
## [361] "course"     "other"      "other"      "course"     "course"
```

```
## [366] "course"     "other"      "reputation" "course"     "other"
## [371] "home"       "other"      "home"       "course"     "reputation"
## [376] "home"       "course"     "course"     "home"       "reputation"
## [381] "home"       "other"      "home"       "other"      "home"
## [386] "other"      "reputation" "course"     "course"     "course"
## [391] "course"     "course"     "course"     "course"     "course"
```

```
#I chose the "failures" as y. Also I will do classification model test.
```

2 EDA :

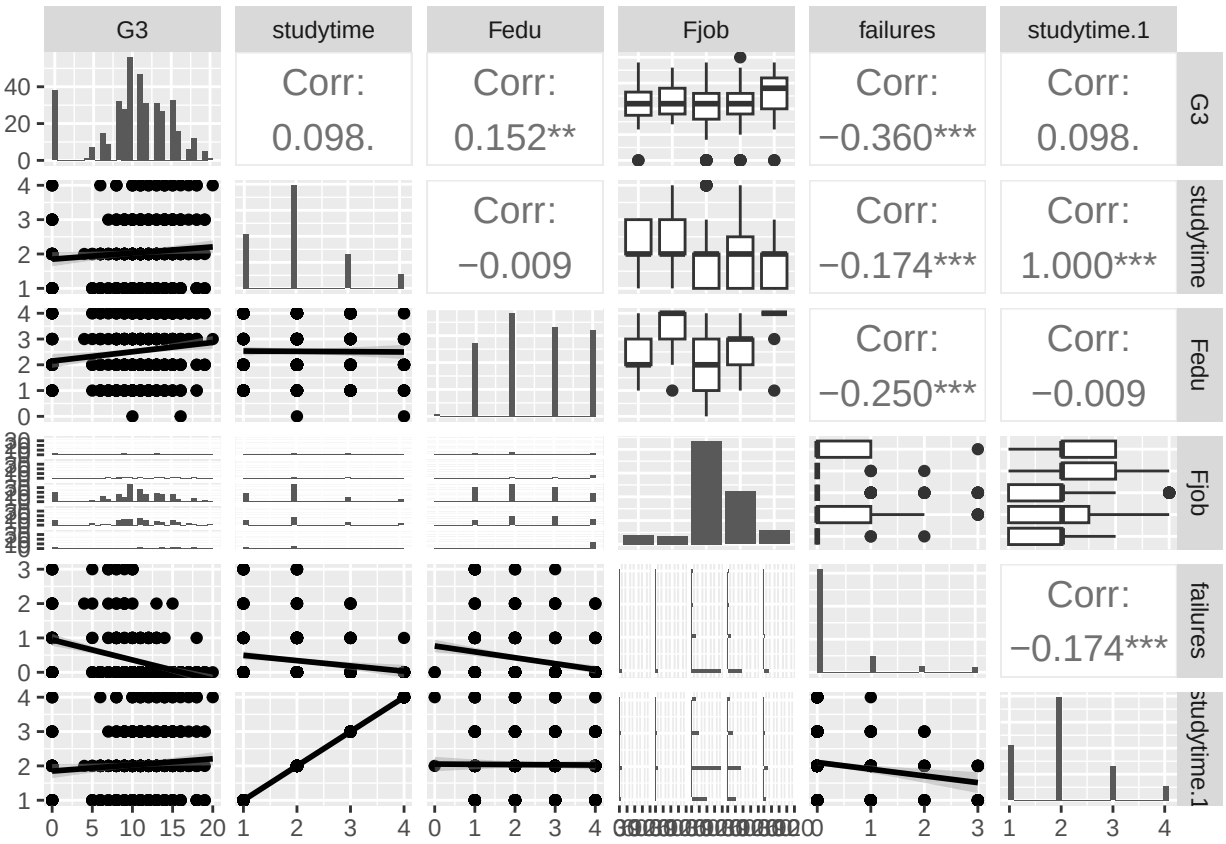
```
#install.packages("GGally")
```

```
library(GGally)
```

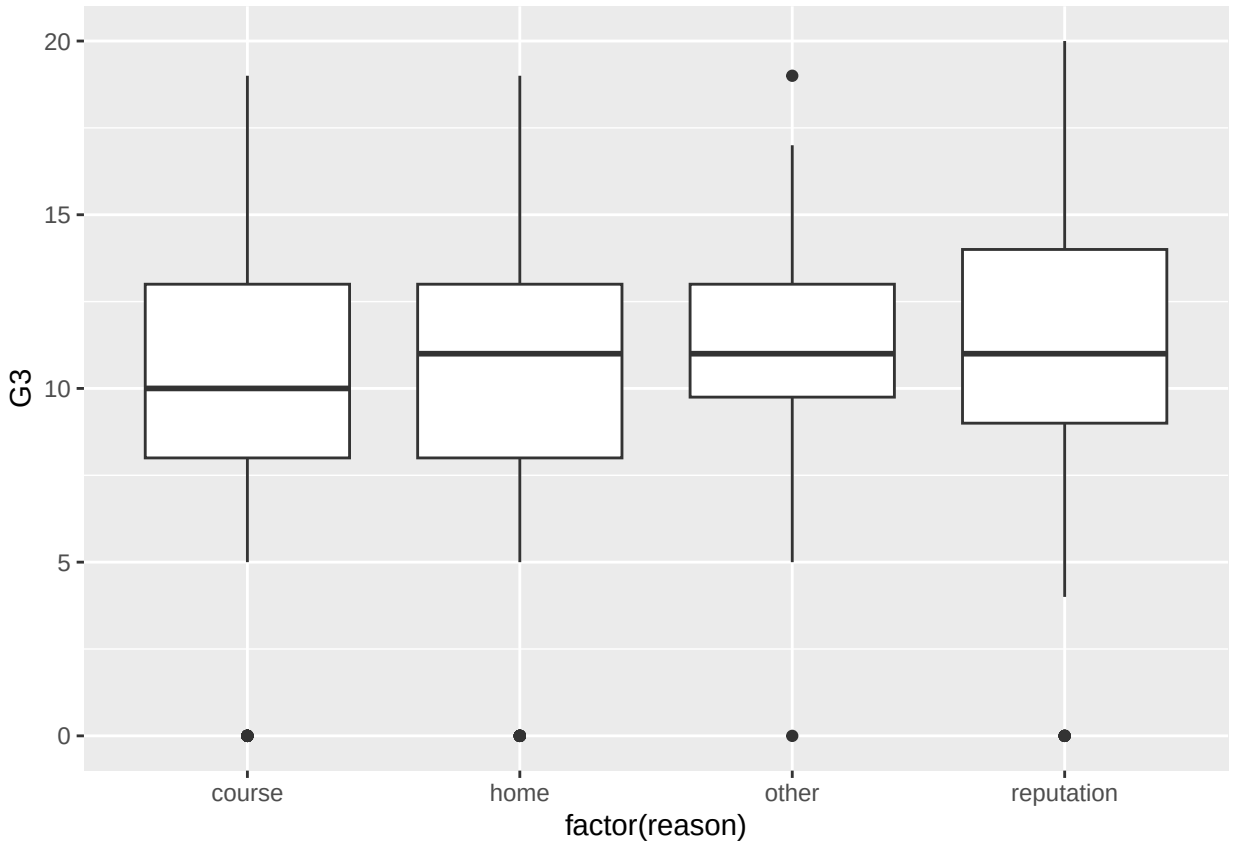
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(student[,c("G3","studytime","Fedu","Fjob","failures","studytime")],upper=list(continuous=wrap("
```

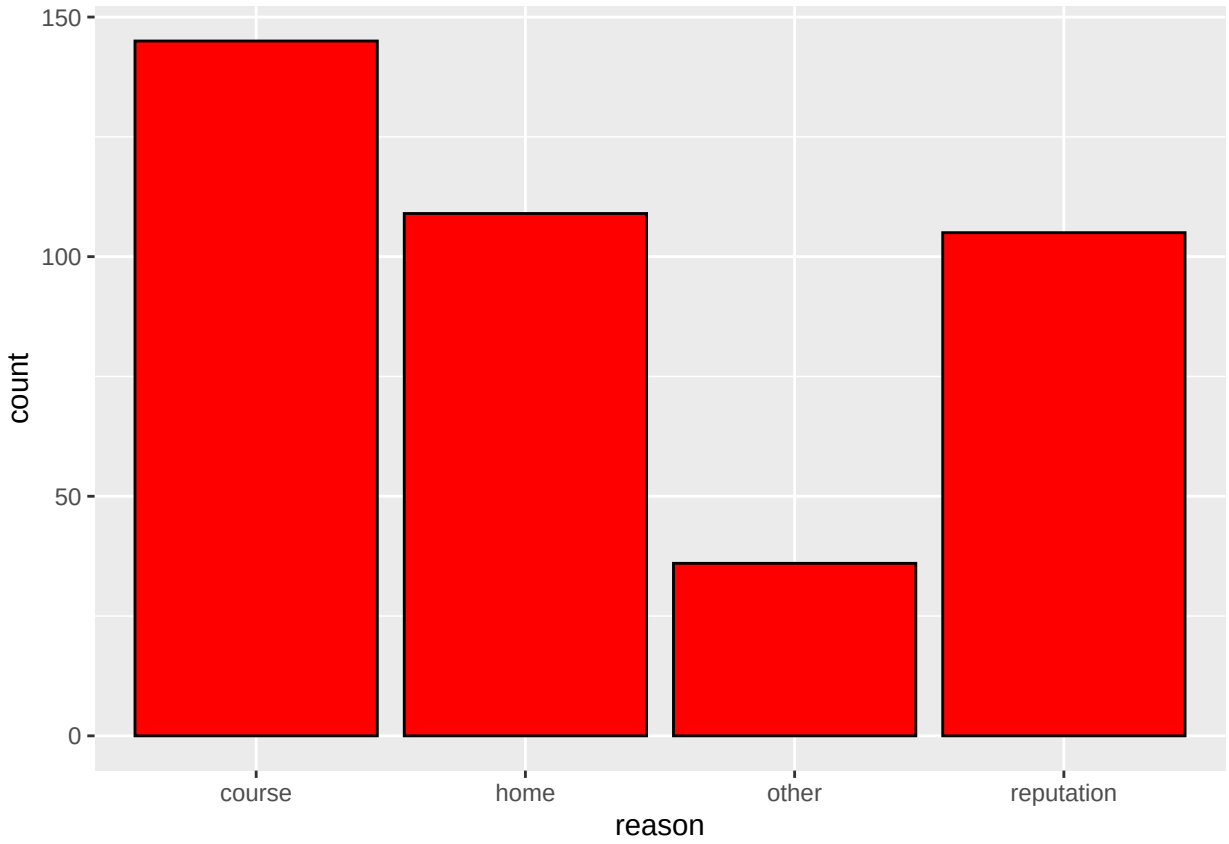
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#ggplot(student,aes(x=G3))+geom_histogram(fill="red",x="G3",y="reason")
ggplot(student,aes(x=factor(reason),y=G3))+geom_boxplot()
```




```
ggplot(student,aes(x=reason))+geom_bar(fill="red",color="black")
```




```
library(caret)
library(lattice)
```

Train/test data

```
set.seed(123)

train_index=sample(1:nrow(student),size=0.8*nrow(student))
train_data=student[train_index,]
test_data=student[-train_index,]


#train_index=createDataPartition(student$famsup,p=0.8)
#train_data=student[train_index,]
#test_data=student[-train_index,]

#logistic regression model


dim(train_data)
```

```
## [1] 316  33
```

```
dim(test_data)
```

```
## [1] 79 33
```

```r
library(nnet)
```

```r
#data

logit_model <- glm(train_data$famsup~ G3+studytime+Fedu+Fjob+studytime+failures,
                   data = train_data,family=binomial)
```

```
## Warning: glm.fit: algorithm did not converge
```

```r
#a2_train$LeaveOrNot~.,data=a2_train,family=binomial)
logit_pred=predict(logit_model,newdata=test_data)
#confusion matrix

#confu_logit=table(True=test_data$failures,Predicted=logit_pred)
pred_class=ifelse(logit_pred>0.5,1,0)
actual_test=test_data$famsup
table(True=test_data$famsup,Predicted=pred_class)
```

```
##     Predicted
## True  0
##    0 79
```

```r
#evaluate model for each class
#prec_logis=diag(confu_logit)/rowSums(confu_logit)
#prec_logis

#sens_logis=diag(confu_logit)/colSums(confu_logit)
#sens_logis
#f1_score=2*(prec_logis*sens_logis)/(prec_logis+sens_logis)
#f1_score
```

```r
rf_stu=randomForest(failures ~ G3+studytime+Fedu+Fjob+studytime,
                   data = train_data,ntree=1000)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values.  Are you sure you want to do regression?
```