Final Project

Introduction

The Data

We will be using data that has extensive information on secondary school students in their math class.

```
data <- read.csv("data/student-mat.csv")
```

Creation of New Variables

In order to provide more insight, we saw room to create informative variables based upon the given data.

The given variables Medu and Fedu give information about the student's parents education history. Using this, we created a new variable "first_gen_college" that indicates if the student would be a first generation college student if they decided to pursue higher education. This will give more tangible and clear insight to how parental education impacts student's performance.

```
data <- data %>%
  mutate(first_gen_college = case_when(
    Medu < 4 & Fedu < 4 ~ "yes",
    TRUE ~"no"
  ))
data[["first_gen_college"]] <- as.factor(data[["first_gen_college"]])
```

Additionally, many variables are self reported ratings from the students on a scale of 1-5. We decided that instead factoring these variables so that scores of 1-3 would be "low" and scores of 4-5 would be "high" would be beneficial to our analysis as it would be more interpretable in context.

```
data <- data %>%
  mutate(famrel = case_when(
    famrel == 1 ~ "low",
    famrel == 2 ~ "low",
    famrel == 3 ~ "low",
    famrel == 4 ~ "high",
    famrel == 5 ~"high"
  ))

data <- data %>%
  mutate(freetime = case_when(
    freetime == 1 ~ "low",
    freetime == 2 ~ "low",
    freetime == 3 ~ "low",
    freetime == 4 ~ "high",
    freetime == 5 ~"high"
  ))

data <- data %>%
  mutate(goout = case_when(
    goout == 1 ~ "low",
    goout == 2 ~ "low",
```

```r
    goout == 3 ~ "low",
    goout == 4 ~ "high",
    goout == 5 ~"high"
  ))

data <- data %>%
  mutate(Dalc = case_when(
    Dalc == 1 ~ "low",
    Dalc == 2 ~ "low",
    Dalc == 3 ~ "low",
    Dalc == 4 ~ "high",
    Dalc == 5 ~"high"
  ))

data <- data %>%
  mutate(Walc = case_when(
    Walc == 1 ~ "low",
    Walc == 2 ~ "low",
    Walc == 3 ~ "low",
    Walc == 4 ~ "high",
    Walc == 5 ~"high"
  ))

data <- data %>%
  mutate(health = case_when(
    health == 1 ~ "low",
    health == 2 ~ "low",
    health == 3 ~ "low",
    health == 4 ~ "high",
    health == 5 ~"high"
  ))

data[["sex"]] <- as.factor(data[["sex"]])
data[["address"]] <- as.factor(data[["address"]])
data[["famsize"]] <- as.factor(data[["famsize"]])
data[["Pstatus"]] <- as.factor(data[["Pstatus"]])
data[["Mjob"]] <- as.factor(data[["Mjob"]])
data[["Fjob"]] <- as.factor(data[["Fjob"]])
data[["reason"]] <- as.factor(data[["reason"]])
data[["guardian"]] <- as.factor(data[["guardian"]])
data[["schoolsup"]] <- as.factor(data[["schoolsup"]])
data[["famsup"]] <- as.factor(data[["famsup"]])
data[["paid"]] <- as.factor(data[["paid"]])
data[["activities"]] <- as.factor(data[["activities"]])
data[["nursery"]] <- as.factor(data[["nursery"]])
data[["higher"]] <- as.factor(data[["higher"]])
data[["internet"]] <- as.factor(data[["internet"]])
data[["romantic"]] <- as.factor(data[["romantic"]])
data[["famrel"]] <- as.factor(data[["famrel"]])
data[["freetime"]] <- as.factor(data[["freetime"]])
data[["goout"]] <- as.factor(data[["goout"]])
data[["Dalc"]] <- as.factor(data[["Dalc"]])
data[["Walc"]] <- as.factor(data[["Walc"]])
```

```r
data[["health"]] <- as.factor(data[["health"]])
```

Additionally, using information from the famsup and internet variables, we created a variable called "stable_learning_env". If famsup is "yes" and internet is "yes", then stable_learning_env is "yes", otherwise "no".

```r
data <- data %>%
  mutate(stable_learning_env = case_when(
    internet =="yes" & famsup =="yes" ~"yes",
    TRUE ~"no"
  ))
data[["stable_learning_env"]] <- as.factor(data[["stable_learning_env"]])
```

Also, we created a new variable "high_freq_absent", which if absences >= 10 for a student, we considered them a highly frequent student.

```r
data <- data %>%
  mutate(high_freq_absent = case_when(
    absences >= 10 ~"yes",
    TRUE ~"no"
  ))
data[["high_freq_absent"]] <- as.factor(data[["high_freq_absent"]])
```

We also created a "failed" variable, which was "yes" if failures > 0, and "no" otherwise.

```r
data <- data %>%
  mutate(failed = case_when(
    failures > 0 ~"yes",
    TRUE ~"no"
  ))
data[["failed"]] <- as.factor(data[["failed"]])
```

Exploratory Data Analysis

```r
summary(data)
```

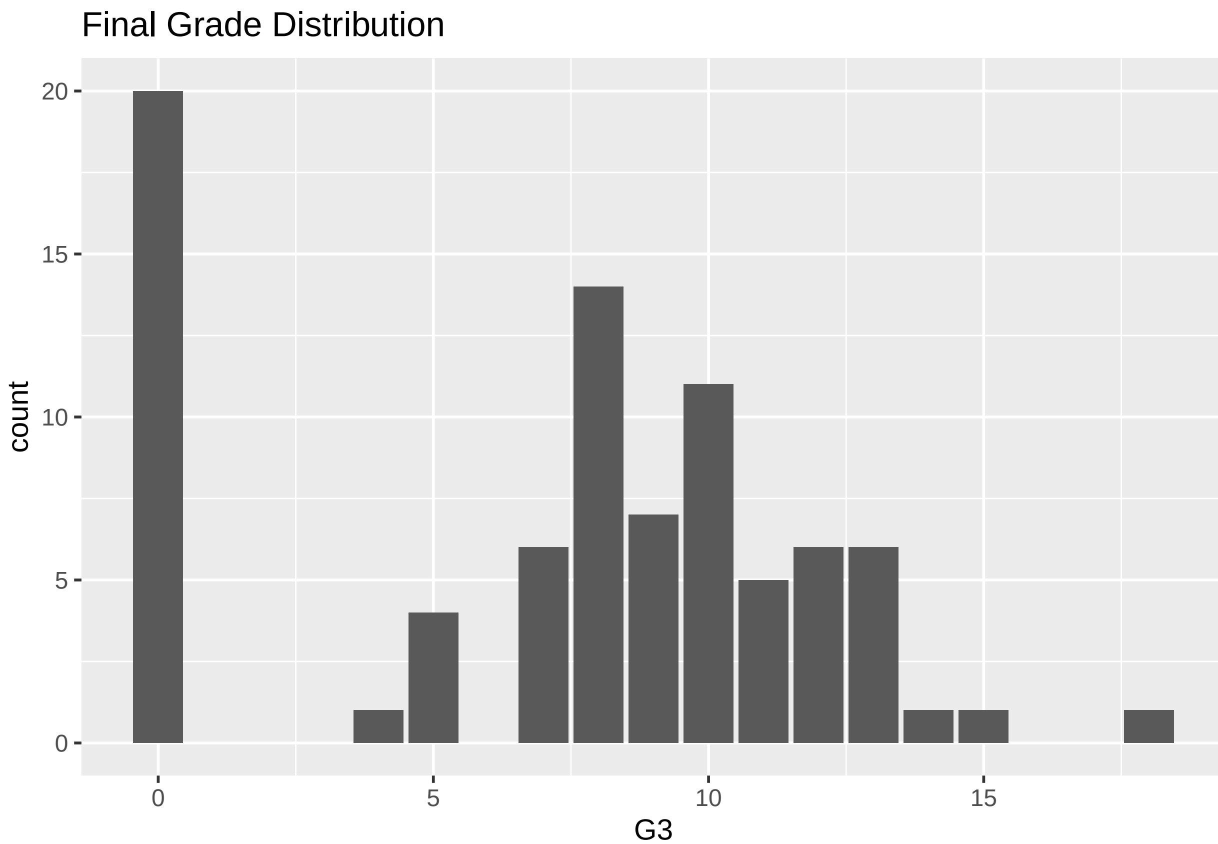
```
##     school              sex          age         address famsize   Pstatus
##  Length:395         F:208   Min.   :15.0   R: 88   GT3:281   A: 41
##  Class :character   M:187   1st Qu.:16.0   U:307   LE3:114   T:354
##  Mode  :character           Median :17.0
##                             Mean   :16.7
##                             3rd Qu.:18.0
##                             Max.   :22.0
##       Medu            Fedu             Mjob           Fjob           reason
##  Min.   :0.000   Min.   :0.000   at_home : 59   at_home : 20   course    :145
##  1st Qu.:2.000   1st Qu.:2.000   health  : 34   health  : 18   home      :109
##  Median :3.000   Median :2.000   other   :141   other   :217   other     : 36
##  Mean   :2.749   Mean   :2.522   services:103   services:111   reputation:105
##  3rd Qu.:4.000   3rd Qu.:3.000   teacher : 58   teacher : 29
##  Max.   :4.000   Max.   :4.000
##    guardian      traveltime      studytime        failures       schoolsup
##  father: 90   Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344
##  mother:273   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51
##  other : 32   Median :1.000   Median :2.000   Median :0.0000
##               Mean   :1.448   Mean   :2.035   Mean   :0.3342
##               3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
```

```
##                Max.   :4.000   Max.   :4.000   Max.   :3.0000
##  famsup     paid     activities nursery    higher    internet  romantic
##  no :153   no :214   no :194    no : 81   no : 20   no : 66   no :263
##  yes:242   yes:181   yes:201    yes:314   yes:375   yes:329   yes:132
##
##
##
##
##   famrel    freetime    goout       Dalc       Walc      health
##  high:301   high:155   high:139   high: 18   high: 79   high:212
##  low : 94   low :240   low :256   low :377   low :316   low :183
##
##
##
##
##     absences            G1              G2              G3
##  Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
##  Median : 4.000   Median :11.00   Median :11.00   Median :11.00
##  Mean   : 5.709   Mean   :10.91   Mean   :10.71   Mean   :10.42
##  3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00
##  first_gen_college stable_learning_env high_freq_absent failed
##  no :157           no :186             no :312          no :312
##  yes:238           yes:209             yes: 83          yes: 83
##
##
##
##
```

First, I will start off with univariate and bivariate plots of the response variable and key predictors I see being important.

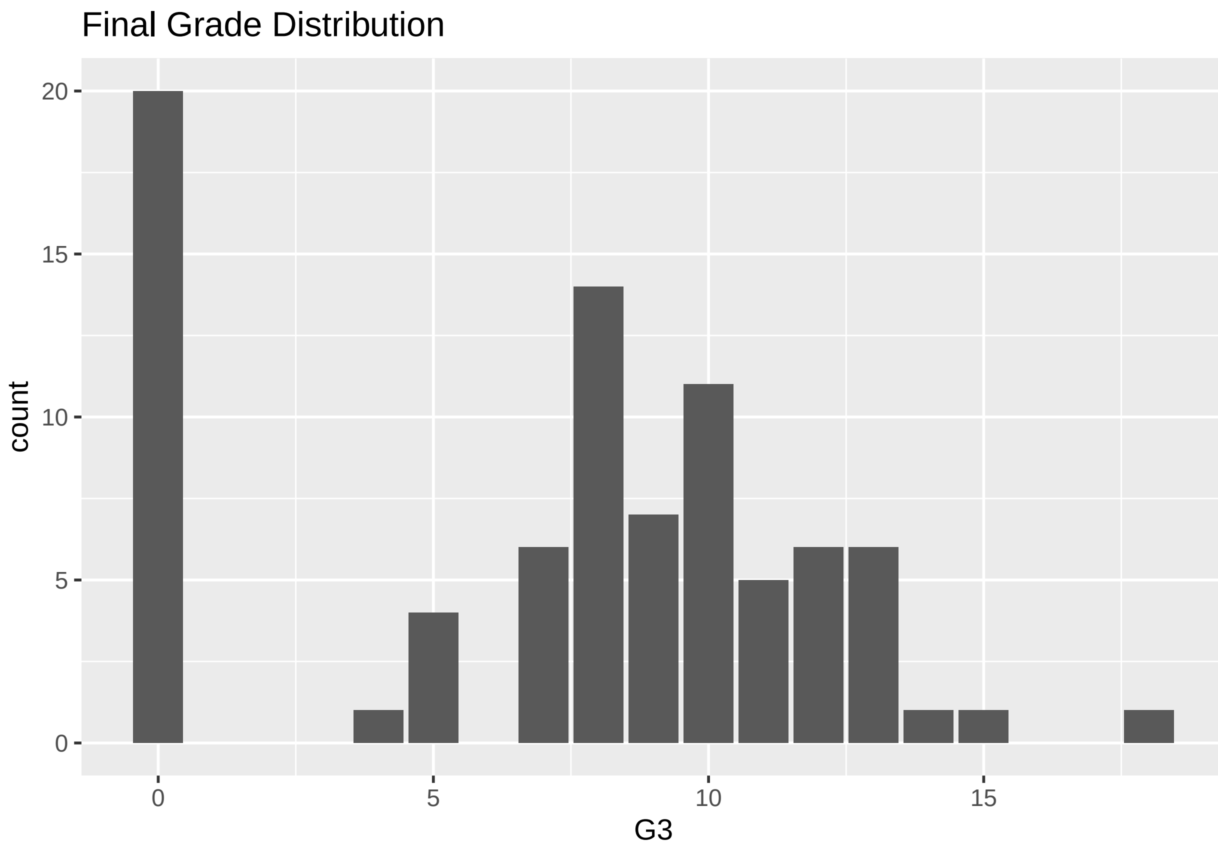
```
data %>%
  filter(failed =="yes") %>%
  ggplot(aes(G3)) +
  geom_histogram(stat = "count") +
  labs(title="Final Grade Distribution")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



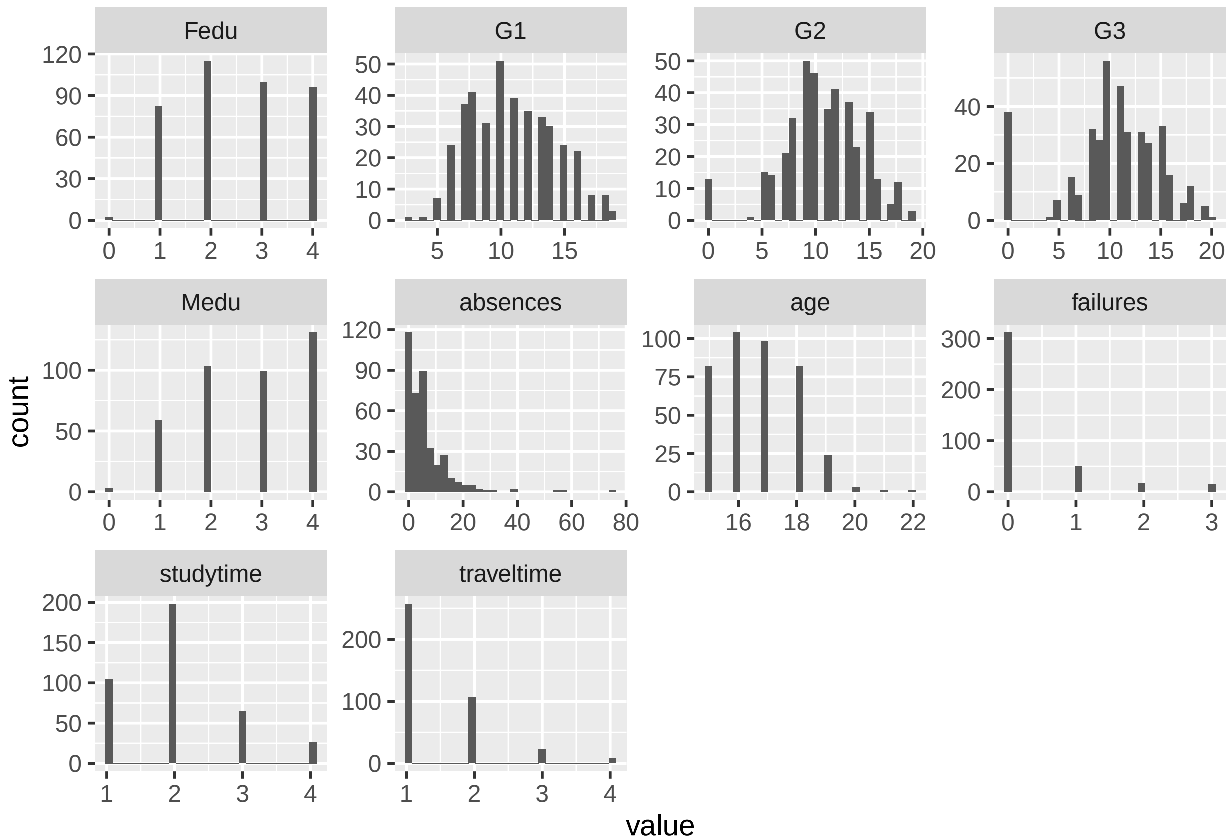

```r
data %>%
  filter(failed =="yes") %>%
  ggplot(aes(G3)) +
  geom_histogram(stat = "count") +
  labs(title="Final Grade Distribution")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



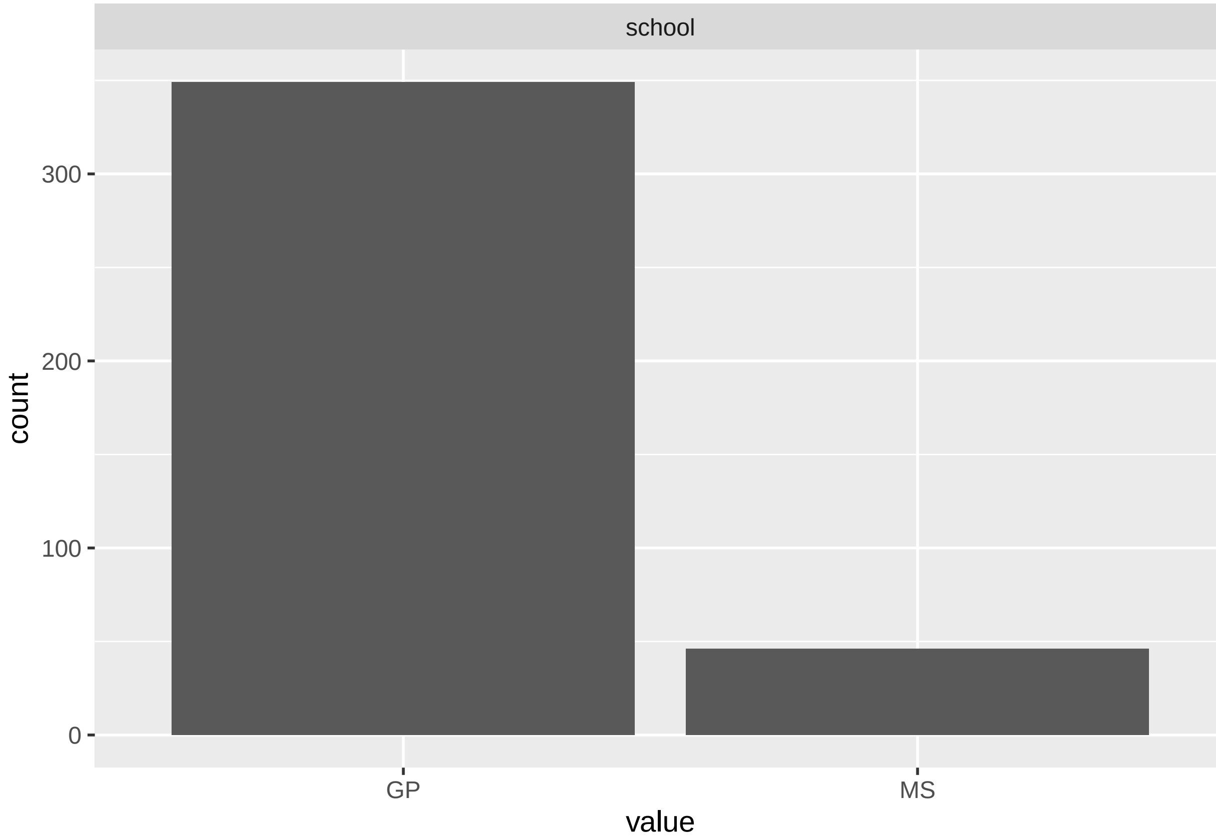

```r
data %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram()
```

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



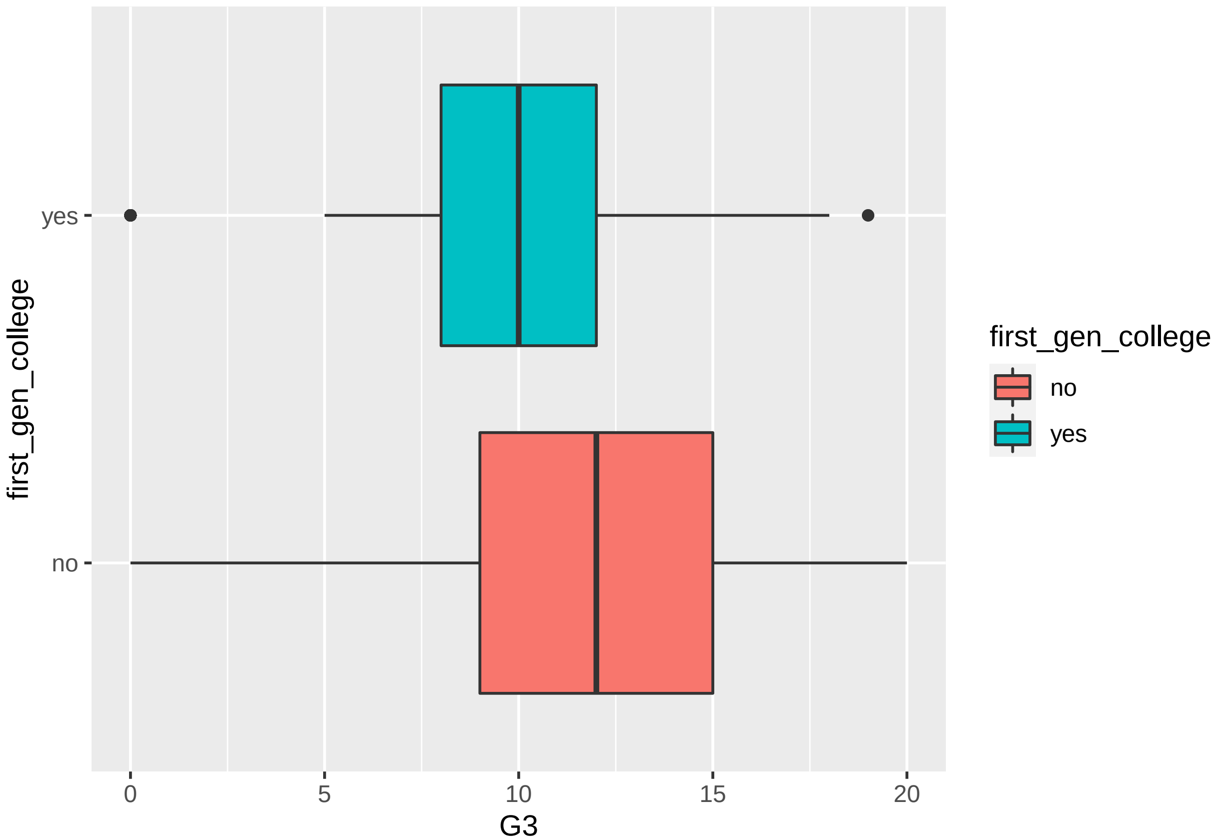

```r
data %>%
  keep(is.character) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(stat="count")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

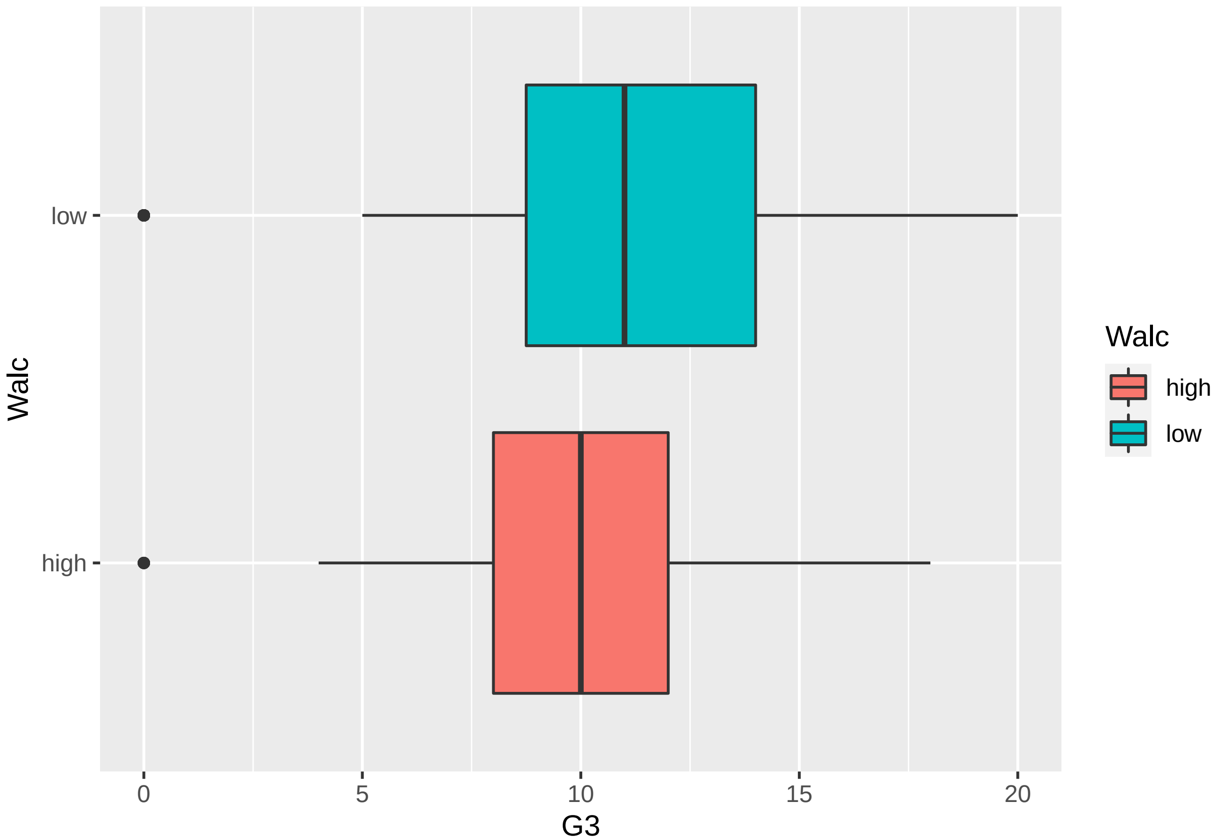


Above we see that the response variable, G3, is pretty normally distributed, thus no transformation is necessary,

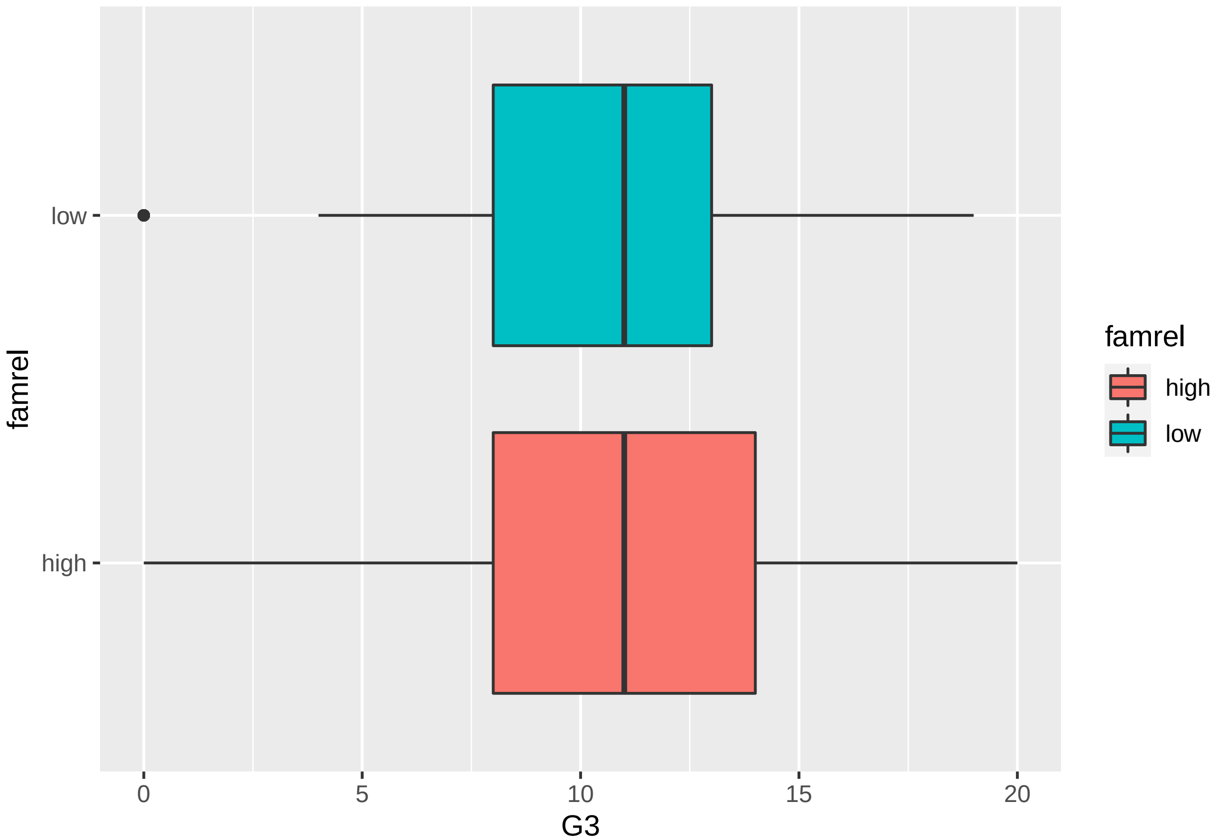
```
ggplot(data = data, aes(x = G3, y = first_gen_college, fill=first_gen_college)) +
  geom_boxplot()
```



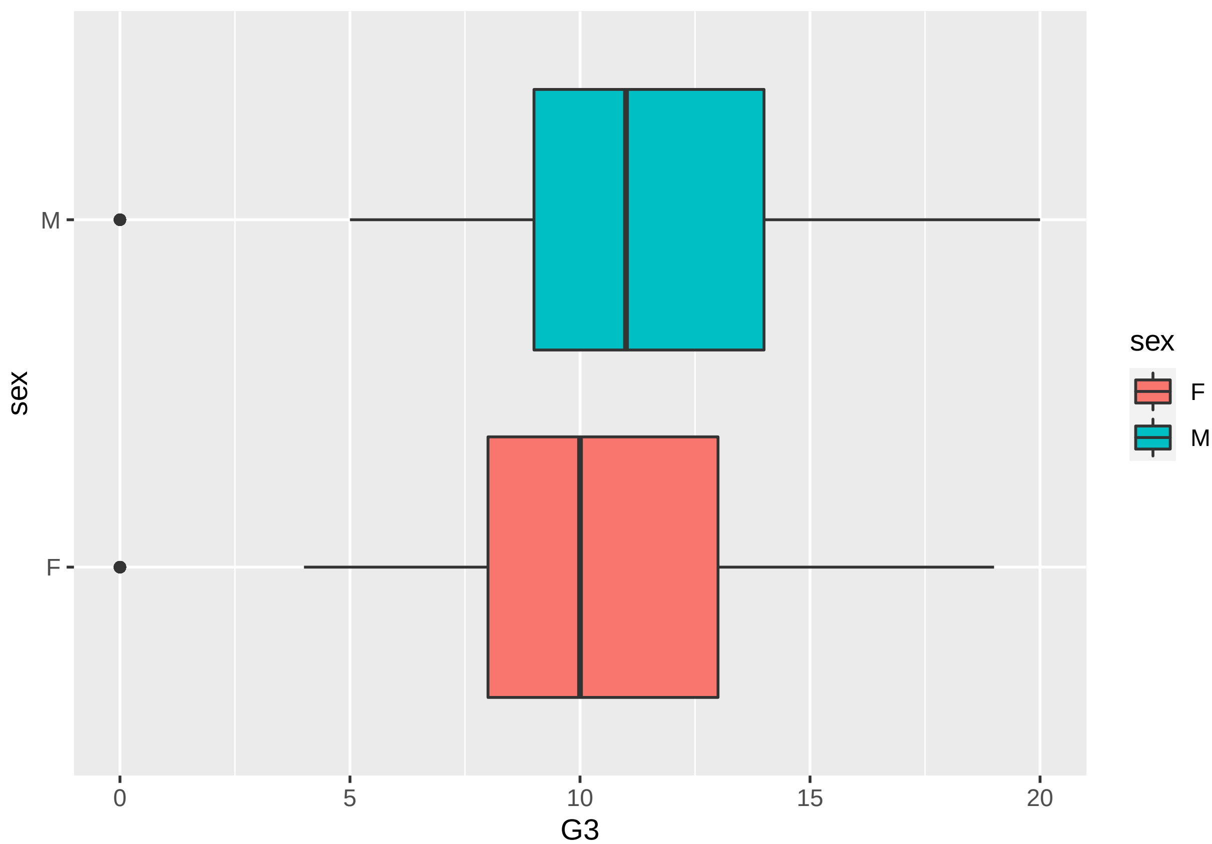

```
ggplot(data = data, aes(x = G3, y = Walc, fill = Walc)) +
  geom_boxplot()
```



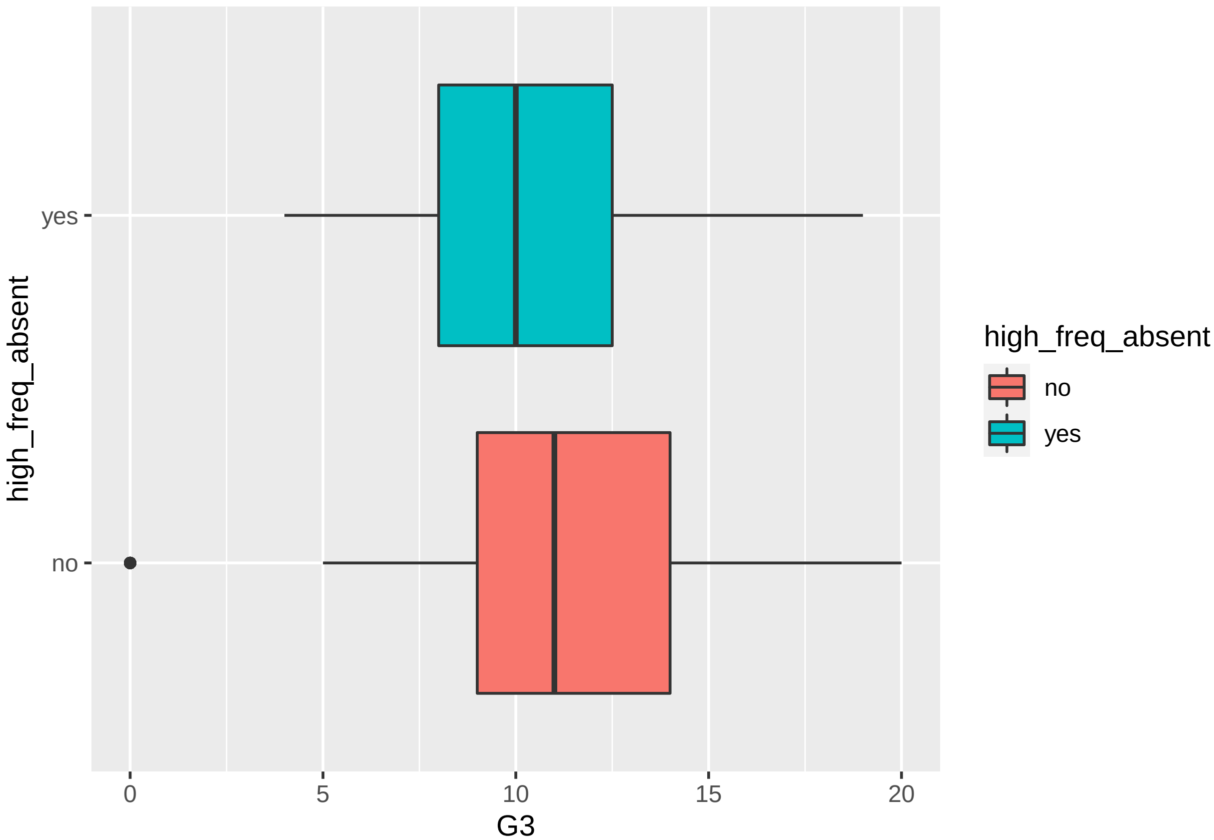

```
ggplot(data = data, aes(x = G3, y = famrel, fill = famrel)) +
  geom_boxplot()
```



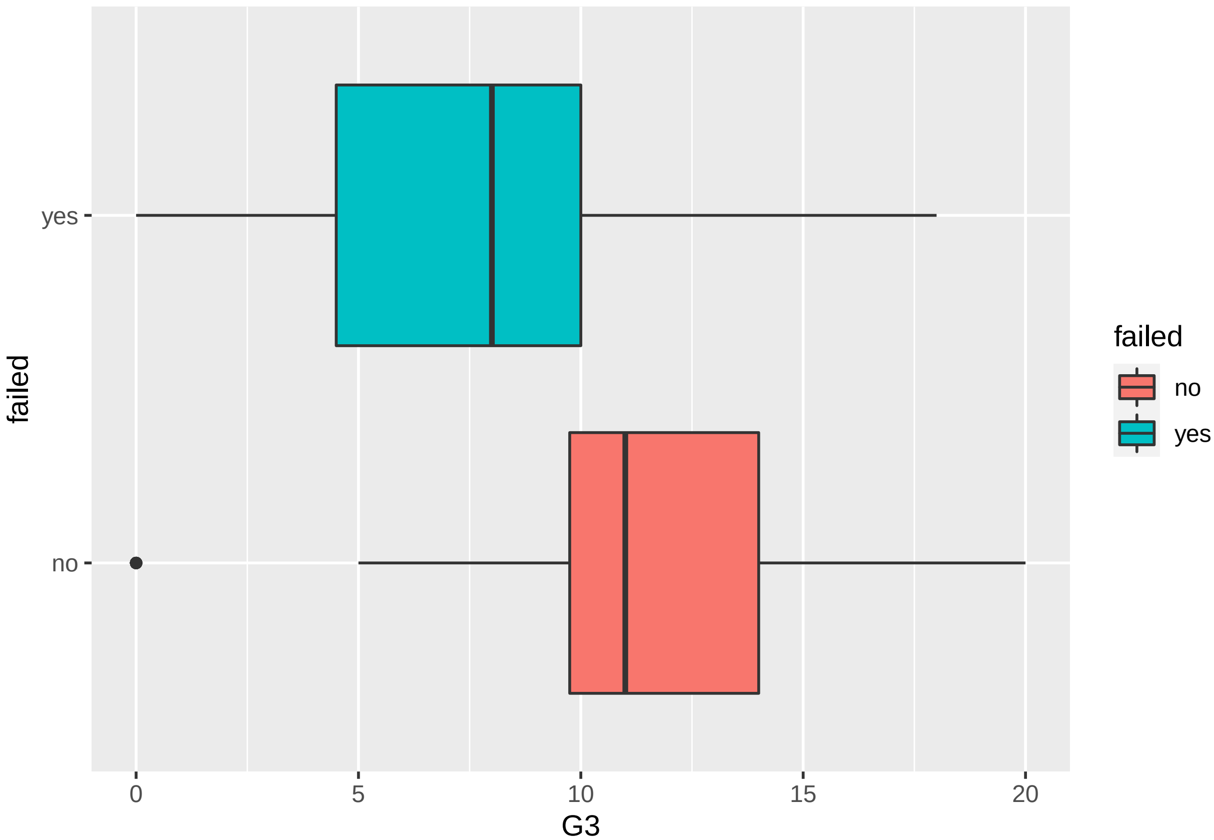

```r
ggplot(data = data, aes(x = G3, y= sex, fill = sex)) +
  geom_boxplot()
```



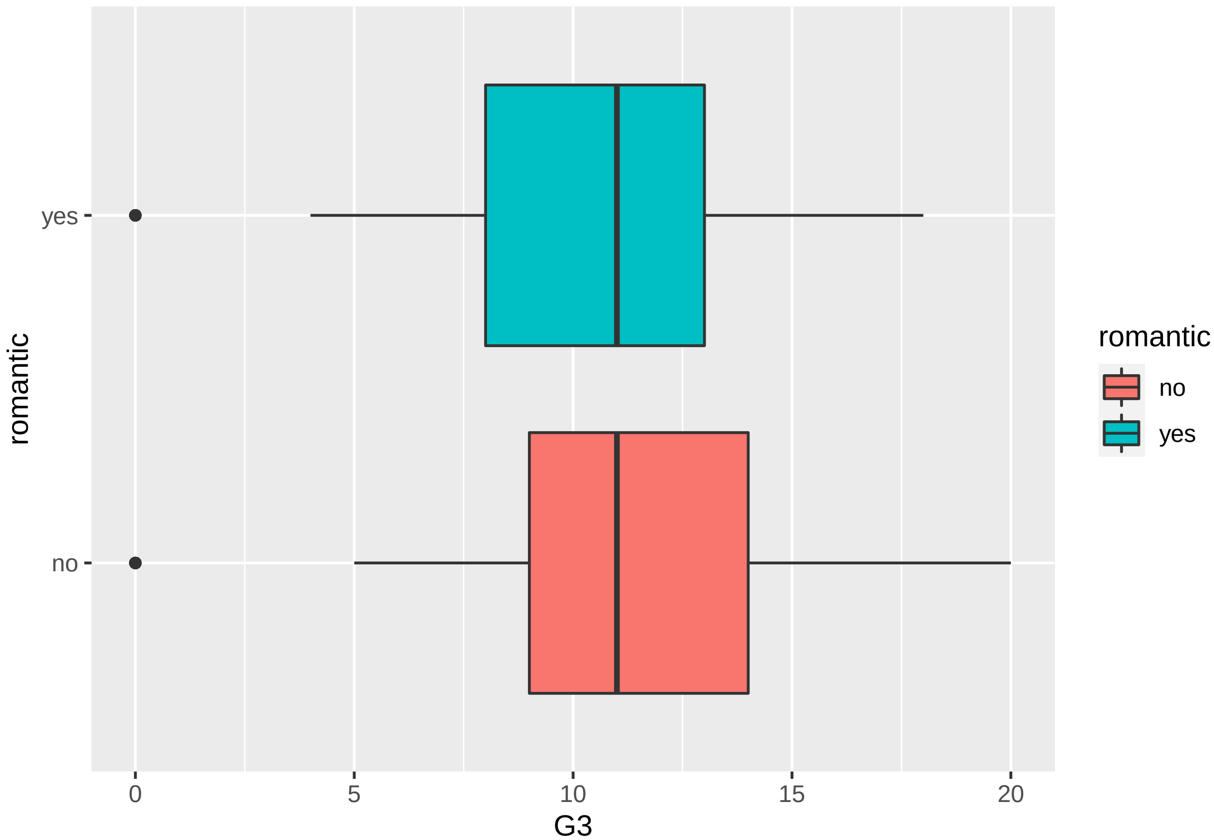

```r
ggplot(data = data, aes(x = G3, y = high_freq_absent, fill = high_freq_absent)) +
  geom_boxplot()
```



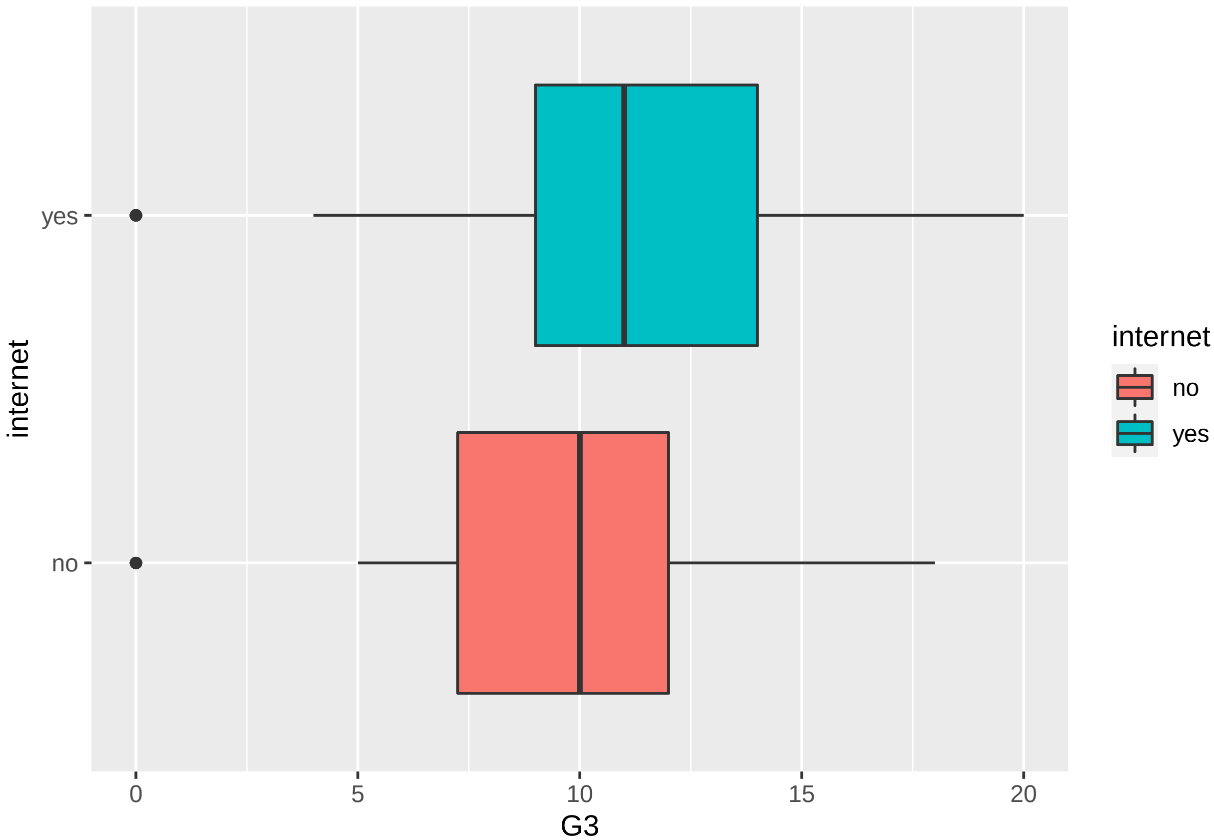

```r
ggplot(data = data, aes(x = G3, y=failed, fill = failed)) +
  geom_boxplot()
```



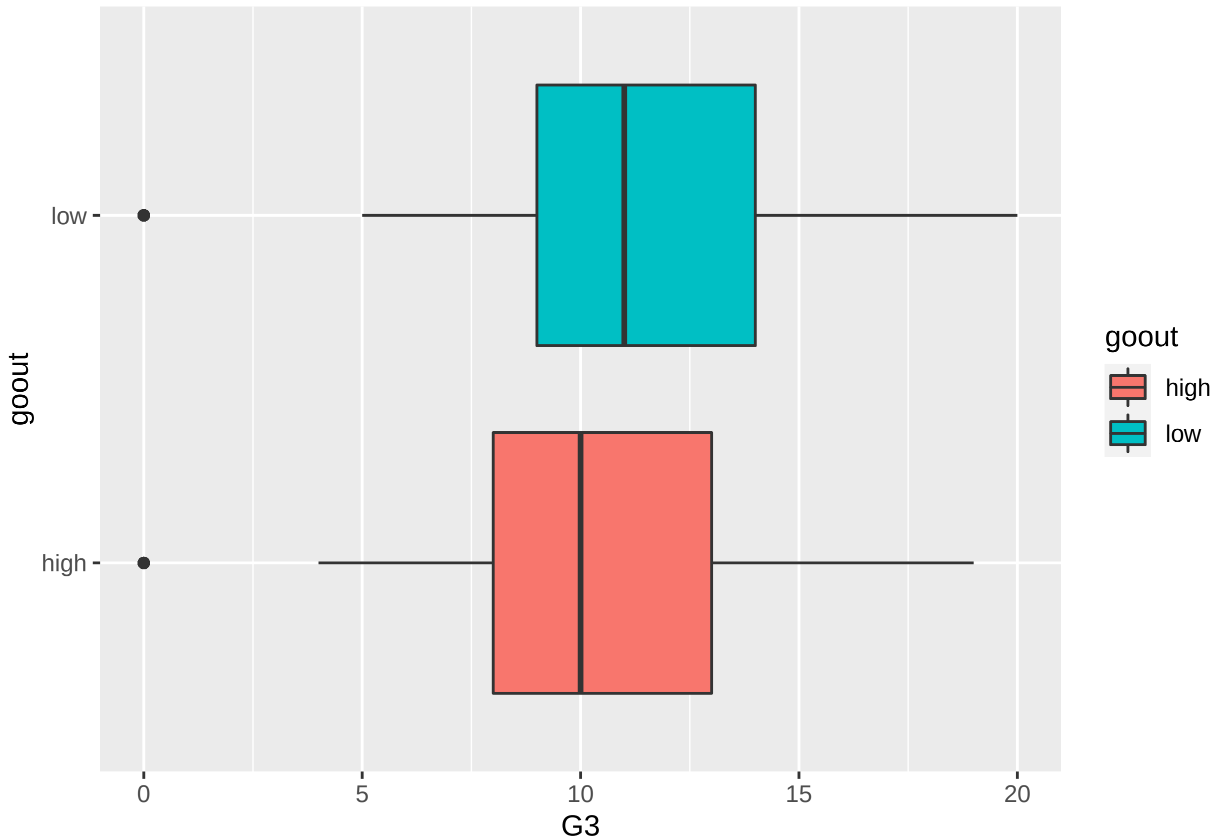

```r
ggplot(data = data, aes(x = G3, y=romantic, fill = romantic)) +
  geom_boxplot()
```




```r
ggplot(data = data, aes(x = G3, y=internet, fill = internet)) +
  geom_boxplot()
```




```r
ggplot(data = data, aes(x = G3, y=goout, fill = goout)) +
  geom_boxplot()
```



From the initial explorations above, we can see a few possible trends. Students who had at least one of the following traits: failed a class previously, were a highly frequent absent student, frequently went out, without internet, were frequent drinkers on the weekend, were in romantic relationships, and were first generation students, on average had lower final grades than their counterparts.

```r
names(data)
```

```
##  [1] "school"            "sex"                "age"
##  [4] "address"           "famsize"            "Pstatus"
##  [7] "Medu"              "Fedu"               "Mjob"
## [10] "Fjob"              "reason"             "guardian"
## [13] "traveltime"        "studytime"          "failures"
## [16] "schoolsup"         "famsup"             "paid"
## [19] "activities"        "nursery"            "higher"
## [22] "internet"          "romantic"           "famrel"
## [25] "freetime"          "goout"              "Dalc"
## [28] "Walc"              "health"             "absences"
## [31] "G1"                "G2"                 "G3"
## [34] "first_gen_college"   "stable_learning_env" "high_freq_absent"
## [37] "failed"
```

```r
num_cols <- unlist(lapply(data, is.numeric))
quant_vars <- data[,num_cols]
cor(quant_vars)
```

```
##                     age        Medu         Fedu  traveltime    studytime
## age         1.000000000 -0.16365842 -0.163438069  0.07064072 -0.004140037
## Medu       -0.163658419  1.00000000  0.623455112 -0.17163930  0.064944137
## Fedu       -0.163438069  0.62345511  1.000000000 -0.15819405 -0.009174639
```

```
## traveltime  0.070640721 -0.17163930 -0.158194054  1.00000000 -0.100909119
## studytime  -0.004140037  0.06494414 -0.009174639 -0.10090912  1.000000000
## failures    0.243665377 -0.23667996 -0.250408444  0.09223875 -0.173563031
## absences    0.175230079  0.10028482  0.024472887 -0.01294378 -0.062700175
## G1         -0.064081497  0.20534100  0.190269936 -0.09303999  0.160611915
## G2         -0.143474049  0.21552717  0.164893393 -0.15319796  0.135879999
## G3         -0.161579438  0.21714750  0.152456939 -0.11714205  0.097819690
##               failures    absences          G1         G2          G3
## age         0.24366538  0.17523008 -0.06408150 -0.1434740 -0.16157944
## Medu       -0.23667996  0.10028482  0.20534100  0.2155272  0.21714750
## Fedu       -0.25040844  0.02447289  0.19026994  0.1648934  0.15245694
## traveltime  0.09223875 -0.01294378 -0.09303999 -0.1531980 -0.11714205
## studytime  -0.17356303 -0.06270018  0.16061192  0.1358800  0.09781969
## failures    1.00000000  0.06372583 -0.35471761 -0.3558956 -0.36041494
## absences    0.06372583  1.00000000 -0.03100290 -0.0317767  0.03424732
## G1         -0.35471761 -0.03100290  1.00000000  0.8521181  0.80146793
## G2         -0.35589563 -0.03177670  0.85211807  1.0000000  0.90486799
## G3         -0.36041494  0.03424732  0.80146793  0.9048680  1.00000000
```

```
#library(corr)
#quant_vars %>% correlate() %>% network_plot(min_cor=0.2)
```

Creating variables for an ordinal final grade, 6-category final grade, and binary final grade

We'd like to examine final grades in multiple ways. The first is as a continuous numerical variable as G3 is.

The second is final grades as an ordered factor variable in order to perform multicategory ordinal logit modeling to see if we could improve fit and predictive power. However, this was unsuccessful.

```
data <- data %>%
  mutate(ord_g3 = factor(G3, ordered=T)
  )
```

The third is final grades as a 6-category ordered factor variable according to the Portuguese education system's classifications. We believe this could address some of the outliers and abnormality in the data (for example, many students received 0's, but no one received a 1, 2, or 3).

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

## The following object is masked from 'package:openintro':
##
##     densityPlot
```

```
data <- data %>%
  mutate(cat_g3 = case_when(
```

```r
    G3 == 0 ~ "Poor",
    G3 <= 9 ~ "Weak",
    G3 <= 13 ~ "Sufficient",
    G3 <= 15 ~ "Good",
    G3 <= 17 ~"Very Good",
    G3 <= 20 ~ "Excellent"
  ))
data <- data %>%
  mutate(cat_g3 = factor(cat_g3, levels=c("Poor", "Weak", "Sufficient", "Good", "Very Good", "Excellent
```

The fourth is final grades as a binary factor variable. This is done based on the previous categories in the Portuguese classification system. If the student receives a "poor" or "weak" grade, or G3 $<$ 10, this is considered a "low" grade. If the student received a "sufficient" "good" "very good" or "excellent" grade, this is a high grade.

```r
data <- data %>%
  mutate(pf = case_when(
    G3 >= 10 ~ "high",
    G3 < 10 ~ "low"
  ))
data <- data %>%
  mutate(pf = factor(pf, levels=c("high", "low"), ordered = FALSE))
```

Splitting data into training and testing sets

```r
attach(data)
set.seed(3)
train_ind <- sample(x = nrow(data), size = 0.8 * nrow(data))
test_ind_neg <- -train_ind
training <- data[train_ind, ]
testing <- data[test_ind_neg, ]
```

Linear model

```r
base_lm <- lm(G3 ~ . -G2 -G1 -ord_g3 -cat_g3 -pf, data=training)
vif(base_lm)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## school              1.578437  1        1.256358
## sex                 1.464307  1        1.210086
## age                 1.910030  1        1.382038
## address             1.453502  1        1.205613
## famsize             1.129040  1        1.062563
## Pstatus             1.210511  1        1.100232
## Medu                3.864208  1        1.965759
## Fedu                2.747919  1        1.657685
## Mjob                3.907751  4        1.185744
## Fjob                2.470252  4        1.119677
## reason              1.689574  3        1.091347
## guardian            1.957382  2        1.182821
## traveltime          1.399147  1        1.182855
## studytime           1.382126  1        1.175639
## failures            4.908405  1        2.215492
```

```
## schoolsup          1.258712  1        1.121923
## famsup             6.719082  1        2.592119
## paid               1.357612  1        1.165166
## activities         1.148293  1        1.071584
## nursery            1.246017  1        1.116251
## higher             1.420646  1        1.191909
## internet           2.697131  1        1.642294
## romantic           1.183239  1        1.087768
## famrel             1.217355  1        1.103338
## freetime           1.309708  1        1.144425
## goout              1.465603  1        1.210621
## Dalc               1.419632  1        1.191483
## Walc               1.767170  1        1.329349
## health             1.187784  1        1.089855
## absences           2.489997  1        1.577972
## first_gen_college  3.648994  1        1.910234
## stable_learning_env 8.598173  1        2.932264
## high_freq_absent   2.555965  1        1.598739
## failed             4.979120  1        2.231394
```

```r
summary(base_lm)
```

```
##
## Call:
## lm(formula = G3 ~ . - G2 - G1 - ord_g3 - cat_g3 - pf, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.521  -2.044   0.376   2.489   9.741
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          18.203829   5.365987   3.392 0.000796 ***
## schoolMS              1.150014   0.867133   1.326 0.185878
## sexM                  1.322519   0.556764   2.375 0.018225 *
## age                  -0.484580   0.241157  -2.009 0.045483 *
## addressU              0.469861   0.656448   0.716 0.474752
## famsizeLE3            1.084061   0.538517   2.013 0.045095 *
## PstatusT             -0.938572   0.804552  -1.167 0.244403
## Medu                  0.109100   0.415204   0.263 0.792932
## Fedu                 -0.198417   0.350209  -0.567 0.571476
## Mjobhealth            0.867372   1.261872   0.687 0.492436
## Mjobother            -0.267653   0.802816  -0.333 0.739094
## Mjobservices          0.851853   0.890079   0.957 0.339390
## Mjobteacher          -1.408800   1.187437  -1.186 0.236491
## Fjobhealth           -0.610959   1.658086  -0.368 0.712808
## Fjobother            -0.391708   1.059598  -0.370 0.711911
## Fjobservices         -0.165284   1.115764  -0.148 0.882346
## Fjobteacher           0.801340   1.452904   0.552 0.581714
## reasonhome            0.449578   0.618690   0.727 0.468058
## reasonother           0.908713   0.889464   1.022 0.307859
## reasonreputation      1.042423   0.663933   1.570 0.117561
## guardianmother       -0.077074   0.625039  -0.123 0.901953
## guardianother         0.375031   1.092420   0.343 0.731635
## traveltime           -0.395034   0.395929  -0.998 0.319292
```

```
## studytime                0.605942   0.320779   1.889 0.059959 .
## failures                -0.718930   0.657655  -1.093 0.275285
## schoolsupyes            -1.484089   0.750939  -1.976 0.049130 *
## famsupyes               -0.465877   1.229810  -0.379 0.705117
## paidyes                  0.765679   0.536344   1.428 0.154557
## activitiesyes           -0.308944   0.491852  -0.628 0.530450
## nurseryyes              -0.363035   0.626732  -0.579 0.562899
## higheryes                0.008231   1.180199   0.007 0.994440
## internetyes              1.106572   1.016492   1.089 0.277286
## romanticyes             -1.349381   0.529977  -2.546 0.011445 *
## famrellow               -0.107383   0.600760  -0.179 0.858270
## freetimelow             -1.525774   0.537121  -2.841 0.004842 **
## gooutlow                 1.589621   0.578545   2.748 0.006404 **
## Dalclow                  0.241061   1.376739   0.175 0.861135
## Walclow                  0.106693   0.763599   0.140 0.888981
## healthlow                0.626273   0.501042   1.250 0.212395
## absences                 0.092750   0.043125   2.151 0.032380 *
## first_gen_collegeyes    -1.342709   0.900456  -1.491 0.137083
## stable_learning_envyes -0.901463   1.351184  -0.667 0.505232
## high_freq_absentyes     -0.563752   0.902594  -0.625 0.532763
## failedyes               -2.151741   1.233152  -1.745 0.082130 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.08 on 272 degrees of freedom
## Multiple R-squared:  0.3484, Adjusted R-squared:  0.2453
## F-statistic: 3.382 on 43 and 272 DF,  p-value: 6.586e-10
```

From the vif, it is easy to see that stable_learning_env and famsup have high VIF values. This is likely because famsup and was used to create stable_learning_env. Failures and failed also have higher VIF values, likely because failures was used to create failed. Because we believe that failures is much more explanatory than failed, we will choose to include failures in the model. In order to combat multicollinearity and increase interpretability, we will exclude Medu and Fedu as well from the model, as these were used to create first_gen_college.

We will then perform stepwise selection.

```r
base_lm1 <- lm(G3 ~ . -G2 -G1 -ord_g3 -cat_g3 -pf -famsup -failed -Medu -Fedu, data=training)
vif(base_lm1)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## school              1.568012  1        1.252203
## sex                 1.429650  1        1.195680
## age                 1.886543  1        1.373515
## address             1.417926  1        1.190767
## famsize             1.113004  1        1.054990
## Pstatus             1.186562  1        1.089294
## Mjob                2.919758  4        1.143321
## Fjob                2.182393  4        1.102470
## reason              1.612403  3        1.082876
## guardian            1.849833  2        1.166227
## traveltime          1.360444  1        1.166381
## studytime           1.318872  1        1.148421
## failures            1.449164  1        1.203812
## schoolsup           1.251563  1        1.118733
## paid                1.320423  1        1.149097
```

```
## activities          1.135452  1        1.065576
## nursery             1.232963  1        1.110388
## higher              1.400887  1        1.183591
## internet            1.561243  1        1.249497
## romantic            1.178830  1        1.085740
## famrel              1.206921  1        1.098599
## freetime            1.295403  1        1.138158
## goout               1.455626  1        1.206493
## Dalc                1.409833  1        1.187364
## Walc                1.763148  1        1.327836
## health              1.163763  1        1.078778
## absences            2.439558  1        1.561908
## first_gen_college   1.958251  1        1.399375
## stable_learning_env 1.642204  1        1.281485
## high_freq_absent    2.526284  1        1.589429
```

```r
summary(base_lm1)
```

```
##
## Call:
## lm(formula = G3 ~ . - G2 - G1 - ord_g3 - cat_g3 - pf - famsup -
##     failed - Medu - Fedu, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5624  -2.1927   0.4004   2.5123  10.5373
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             17.79859    5.04615   3.527 0.000492 ***
## schoolMS                 1.04850    0.86362   1.214 0.225754
## sexM                     1.32764    0.54972   2.415 0.016381 *
## age                     -0.50485    0.23949  -2.108 0.035932 *
## addressU                 0.50691    0.64788   0.782 0.434641
## famsizeLE3               1.03186    0.53428   1.931 0.054467 .
## PstatusT                -1.04561    0.79595  -1.314 0.190054
## Mjobhealth               1.08398    1.15446   0.939 0.348575
## Mjobother               -0.11206    0.76049  -0.147 0.882965
## Mjobservices             0.90872    0.81142   1.120 0.263726
## Mjobteacher             -1.19095    1.07084  -1.112 0.267033
## Fjobhealth              -0.73039    1.63411  -0.447 0.655248
## Fjobother               -0.28368    1.05558  -0.269 0.788329
## Fjobservices            -0.08299    1.10893  -0.075 0.940398
## Fjobteacher              0.68683    1.42000   0.484 0.628993
## reasonhome               0.40798    0.61497   0.663 0.507626
## reasonother              0.93393    0.87811   1.064 0.288452
## reasonreputation         0.97534    0.65634   1.486 0.138412
## guardianmother           0.05160    0.61482   0.084 0.933172
## guardianother            0.23516    1.08695   0.216 0.828877
## traveltime              -0.34200    0.39012  -0.877 0.381438
## studytime                0.65563    0.31312   2.094 0.037182 *
## failures                -1.63949    0.35708  -4.591 6.69e-06 ***
## schoolsupyes            -1.54992    0.74824  -2.071 0.039248 *
## paidyes                  0.84526    0.52855   1.599 0.110917
## activitiesyes           -0.36125    0.48873  -0.739 0.460440
```

```
## nurseryyes             -0.36948    0.62297  -0.593 0.553602
## higheryes              -0.14749    1.17108  -0.126 0.899868
## internetyes             1.43957    0.77279   1.863 0.063550 .
## romanticyes            -1.38423    0.52859  -2.619 0.009314 **
## famrellow              -0.08263    0.59773  -0.138 0.890151
## freetimelow            -1.55447    0.53378  -2.912 0.003882 **
## gooutlow                1.62051    0.57614   2.813 0.005265 **
## Dalclow                 0.12807    1.37095   0.093 0.925638
## Walclow                 0.12246    0.76216   0.161 0.872470
## healthlow               0.71110    0.49558   1.435 0.152453
## absences                0.08563    0.04265   2.007 0.045676 *
## first_gen_collegeyes   -1.26069    0.65915  -1.913 0.056834 .
## stable_learning_envyes -1.43472    0.59006  -2.431 0.015674 *
## high_freq_absentyes    -0.59115    0.89666  -0.659 0.510269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.076 on 276 degrees of freedom
## Multiple R-squared:  0.3398, Adjusted R-squared:  0.2465
## F-statistic: 3.642 on 39 and 276 DF,  p-value: 1.819e-10
```

```
step.model <- stepAIC(base_lm1, direction="both")
```

```
## Start:  AIC=925.34
## G3 ~ (school + sex + age + address + famsize + Pstatus + Medu +
##     Fedu + Mjob + Fjob + reason + guardian + traveltime + studytime +
##     failures + schoolsup + famsup + paid + activities + nursery +
##     higher + internet + romantic + famrel + freetime + goout +
##     Dalc + Walc + health + absences + G1 + G2 + first_gen_college +
##     stable_learning_env + high_freq_absent + failed + ord_g3 +
##     cat_g3 + pf) - G2 - G1 - ord_g3 - cat_g3 - pf - famsup -
##     failed - Medu - Fedu
##
##                         Df Sum of Sq    RSS    AIC
## - Fjob                   4     18.10 4604.6 918.59
## - guardian               2      0.78 4587.3 921.40
## - reason                 3     44.49 4631.0 922.39
## - Dalc                   1      0.15 4586.7 923.35
## - higher                 1      0.26 4586.8 923.36
## - famrel                 1      0.32 4586.8 923.36
## - Walc                   1      0.43 4586.9 923.37
## - nursery                1      5.85 4592.4 923.74
## - high_freq_absent       1      7.22 4593.7 923.84
## - activities             1      9.08 4595.6 923.97
## - address                1     10.17 4596.7 924.04
## - traveltime             1     12.77 4599.3 924.22
## - school                 1     24.49 4611.0 925.03
## - Pstatus                1     28.68 4615.2 925.31
## <none>                               4586.5 925.34
## - health                 1     34.21 4620.7 925.69
## - paid                   1     42.50 4629.0 926.26
## - Mjob                   4    134.94 4721.5 926.51
## - internet               1     57.67 4644.2 927.29
## - first_gen_college      1     60.79 4647.3 927.50
## - famsize                1     61.98 4648.5 927.58
```

```
## - absences            1     66.97 4653.5 927.92
## - schoolsup           1     71.30 4657.8 928.22
## - studytime           1     72.86 4659.4 928.32
## - age                 1     73.84 4660.4 928.39
## - sex                 1     96.93 4683.4 929.95
## - stable_learning_env  1     98.25 4684.8 930.04
## - romantic            1    113.96 4700.5 931.10
## - goout               1    131.47 4718.0 932.27
## - freetime            1    140.93 4727.5 932.91
## - failures            1    350.33 4936.8 946.60
##
## Step:  AIC=918.59
## G3 ~ school + sex + age + address + famsize + Pstatus + Mjob +
##     reason + guardian + traveltime + studytime + failures + schoolsup +
##     paid + activities + nursery + higher + internet + romantic +
##     famrel + freetime + goout + Dalc + Walc + health + absences +
##     first_gen_college + stable_learning_env + high_freq_absent
##
##                      Df Sum of Sq    RSS    AIC
## - guardian            2      0.26 4604.9 914.60
## - reason              3     47.42 4652.0 915.82
## - Dalc                1      0.06 4604.7 916.59
## - higher              1      0.08 4604.7 916.59
## - famrel              1      0.13 4604.7 916.60
## - Walc                1      2.46 4607.1 916.76
## - nursery             1      6.34 4611.0 917.02
## - activities          1      8.90 4613.5 917.20
## - high_freq_absent    1      9.56 4614.2 917.24
## - address             1     10.89 4615.5 917.33
## - traveltime          1     10.92 4615.5 917.34
## - school              1     25.19 4629.8 918.31
## <none>                              4604.6 918.59
## - Pstatus             1     31.30 4635.9 918.73
## - paid                1     34.55 4639.2 918.95
## - health              1     36.56 4641.2 919.09
## - Mjob                4    133.29 4737.9 919.60
## - famsize             1     55.82 4660.4 920.39
## - internet            1     59.44 4664.1 920.64
## - studytime           1     70.92 4675.5 921.42
## - schoolsup           1     71.10 4675.7 921.43
## - age                 1     71.43 4676.1 921.45
## - absences            1     75.03 4679.6 921.69
## - first_gen_college   1     82.83 4687.5 922.22
## - sex                 1    100.82 4705.4 923.43
## - stable_learning_env  1    104.04 4708.7 923.65
## - romantic            1    112.26 4716.9 924.20
## - goout               1    125.94 4730.6 925.11
## + Fjob                4     18.10 4586.5 925.34
## - freetime            1    142.96 4747.6 926.25
## - failures            1    345.98 4950.6 939.48
##
## Step:  AIC=914.6
## G3 ~ school + sex + age + address + famsize + Pstatus + Mjob +
##     reason + traveltime + studytime + failures + schoolsup +
```

```
##     paid + activities + nursery + higher + internet + romantic +
##     famrel + freetime + goout + Dalc + Walc + health + absences +
##     first_gen_college + stable_learning_env + high_freq_absent
##
##                       Df Sum of Sq    RSS    AIC
## - reason               3     47.77 4652.6 911.87
## - higher               1      0.04 4604.9 912.61
## - Dalc                 1      0.05 4604.9 912.61
## - famrel               1      0.16 4605.0 912.62
## - Walc                 1      2.63 4607.5 912.79
## - nursery              1      7.04 4611.9 913.09
## - activities           1      8.87 4613.7 913.21
## - high_freq_absent     1      9.45 4614.3 913.25
## - traveltime           1     10.76 4615.6 913.34
## - address              1     11.60 4616.5 913.40
## - school               1     24.98 4629.9 914.31
## <none>                              4604.9 914.60
## - Pstatus              1     32.70 4637.6 914.84
## - paid                 1     35.19 4640.1 915.01
## - health               1     37.68 4642.6 915.18
## - Mjob                 4    134.49 4739.4 915.70
## - famsize              1     55.76 4660.6 916.41
## - internet             1     59.19 4664.1 916.64
## - schoolsup            1     70.91 4675.8 917.43
## - studytime            1     71.30 4676.2 917.46
## - absences             1     75.41 4680.3 917.74
## - age                  1     78.21 4683.1 917.93
## - first_gen_college    1     83.21 4688.1 918.26
## + guardian             2      0.26 4604.6 918.59
## - sex                  1    100.86 4705.7 919.45
## - stable_learning_env  1    103.97 4708.8 919.66
## - romantic             1    112.47 4717.4 920.23
## + Fjob                 4     17.58 4587.3 921.40
## - goout                1    130.82 4735.7 921.46
## - freetime             1    148.38 4753.3 922.63
## - failures             1    362.01 4966.9 936.52
##
## Step:  AIC=911.87
## G3 ~ school + sex + age + address + famsize + Pstatus + Mjob +
##     traveltime + studytime + failures + schoolsup + paid + activities +
##     nursery + higher + internet + romantic + famrel + freetime +
##     goout + Dalc + Walc + health + absences + first_gen_college +
##     stable_learning_env + high_freq_absent
##
##                       Df Sum of Sq    RSS    AIC
## - higher               1      0.19 4652.8 909.88
## - Dalc                 1      0.37 4653.0 909.89
## - famrel               1      0.67 4653.3 909.91
## - Walc                 1      3.45 4656.1 910.10
## - activities           1      4.73 4657.4 910.19
## - nursery              1      5.43 4658.1 910.23
## - high_freq_absent     1      6.59 4659.2 910.31
## - address              1      7.15 4659.8 910.35
## - traveltime           1     12.89 4665.5 910.74
```

```
## - school              1     22.32 4675.0 911.38
## <none>                            4652.6 911.87
## - Pstatus             1     32.12 4684.8 912.04
## - paid                1     45.97 4698.6 912.97
## - health              1     51.03 4703.7 913.31
## - famsize             1     57.06 4709.7 913.72
## - internet            1     59.18 4711.8 913.86
## - Mjob                4    152.59 4805.2 914.06
## + reason              3     47.77 4604.9 914.60
## - schoolsup           1     70.95 4723.6 914.65
## - studytime           1     73.95 4726.6 914.85
## - age                 1     77.33 4730.0 915.07
## - absences            1     80.72 4733.4 915.30
## - first_gen_college   1     83.01 4735.7 915.45
## + guardian            2      0.61 4652.0 915.82
## - sex                 1     94.04 4746.7 916.19
## - romantic            1    106.92 4759.6 917.05
## - stable_learning_env  1    108.83 4761.5 917.17
## + Fjob                4     20.72 4631.9 918.46
## - goout               1    134.28 4786.9 918.86
## - freetime            1    147.31 4800.0 919.72
## - failures            1    375.51 5028.2 934.39
##
## Step:  AIC=909.88
## G3 ~ school + sex + age + address + famsize + Pstatus + Mjob +
##     traveltime + studytime + failures + schoolsup + paid + activities +
##     nursery + internet + romantic + famrel + freetime + goout +
##     Dalc + Walc + health + absences + first_gen_college + stable_learning_env +
##     high_freq_absent
##
##                       Df Sum of Sq    RSS    AIC
## - Dalc                1      0.37 4653.2 907.90
## - famrel              1      0.67 4653.5 907.92
## - Walc                1      3.39 4656.2 908.11
## - activities          1      4.86 4657.7 908.21
## - nursery             1      5.41 4658.2 908.25
## - high_freq_absent    1      6.91 4659.8 908.35
## - address             1      7.15 4660.0 908.36
## - traveltime          1     12.75 4665.6 908.74
## - school              1     22.14 4675.0 909.38
## <none>                            4652.8 909.88
## - Pstatus             1     32.03 4684.9 910.05
## - paid                1     45.81 4698.6 910.98
## - health              1     51.01 4703.8 911.32
## - famsize             1     57.06 4709.9 911.73
## + higher              1      0.19 4652.6 911.87
## - internet            1     60.09 4712.9 911.93
## - Mjob                4    152.60 4805.4 912.08
## + reason              3     47.91 4604.9 912.61
## - schoolsup           1     70.89 4723.7 912.66
## - studytime           1     73.86 4726.7 912.86
## - age                 1     77.75 4730.6 913.12
## - absences            1     82.38 4735.2 913.43
## - first_gen_college   1     82.86 4735.7 913.46
```

```
## + guardian             2      0.52 4652.3 913.84
## - sex                  1     96.68 4749.5 914.38
## - romantic             1    107.36 4760.2 915.09
## - stable_learning_env  1    109.97 4762.8 915.26
## + Fjob                 4     20.58 4632.3 916.48
## - goout                1    134.38 4787.2 916.88
## - freetime             1    147.26 4800.1 917.72
## - failures             1    390.65 5043.5 933.35
##
## Step:  AIC=907.9
## G3 ~ school + sex + age + address + famsize + Pstatus + Mjob +
##     traveltime + studytime + failures + schoolsup + paid + activities +
##     nursery + internet + romantic + famrel + freetime + goout +
##     Walc + health + absences + first_gen_college + stable_learning_env +
##     high_freq_absent
##
##                       Df Sum of Sq    RSS    AIC
## - famrel               1      0.72 4653.9 905.95
## - Walc                 1      3.03 4656.2 906.11
## - activities           1      4.80 4658.0 906.23
## - nursery              1      6.08 4659.3 906.32
## - high_freq_absent     1      6.57 4659.8 906.35
## - address              1      7.43 4660.6 906.41
## - traveltime           1     12.56 4665.8 906.76
## - school               1     22.66 4675.9 907.44
## <none>                             4653.2 907.90
## - Pstatus              1     32.91 4686.1 908.13
## - paid                 1     47.96 4701.2 909.14
## - health               1     50.72 4703.9 909.33
## - famsize              1     57.16 4710.4 909.76
## + Dalc                 1      0.37 4652.8 909.88
## + higher               1      0.18 4653.0 909.89
## - internet             1     60.49 4713.7 909.98
## - Mjob                 4    152.76 4806.0 910.11
## + reason               3     48.23 4605.0 910.61
## - schoolsup            1     70.53 4723.7 910.66
## - studytime            1     74.08 4727.3 910.90
## - age                  1     77.45 4730.7 911.12
## - absences             1     82.09 4735.3 911.43
## - first_gen_college    1     83.04 4736.2 911.49
## + guardian             2      0.53 4652.7 911.87
## - sex                  1     97.53 4750.7 912.46
## - romantic             1    107.07 4760.3 913.09
## - stable_learning_env  1    110.38 4763.6 913.31
## + Fjob                 4     20.92 4632.3 914.48
## - goout                1    134.45 4787.7 914.91
## - freetime             1    149.23 4802.4 915.88
## - failures             1    391.65 5044.9 931.44
##
## Step:  AIC=905.95
## G3 ~ school + sex + age + address + famsize + Pstatus + Mjob +
##     traveltime + studytime + failures + schoolsup + paid + activities +
##     nursery + internet + romantic + freetime + goout + Walc +
##     health + absences + first_gen_college + stable_learning_env +
```

```
##     high_freq_absent
##
##                       Df Sum of Sq    RSS    AIC
## - Walc                 1      3.64 4657.6 904.20
## - activities           1      4.81 4658.7 904.28
## - nursery              1      5.89 4659.8 904.35
## - high_freq_absent     1      7.01 4660.9 904.43
## - address              1      7.60 4661.5 904.47
## - traveltime           1     12.36 4666.3 904.79
## - school               1     22.05 4676.0 905.45
## <none>                             4653.9 905.95
## - Pstatus              1     32.56 4686.5 906.16
## - paid                 1     48.47 4702.4 907.23
## - health               1     50.03 4703.9 907.33
## - famsize              1     57.21 4711.1 907.81
## + famrel               1      0.72 4653.2 907.90
## + Dalc                 1      0.42 4653.5 907.92
## + higher               1      0.18 4653.7 907.94
## - internet             1     60.48 4714.4 908.03
## - Mjob                 4    153.33 4807.2 908.20
## + reason               3     48.78 4605.1 908.62
## - schoolsup            1     70.28 4724.2 908.69
## - studytime            1     74.34 4728.3 908.96
## - age                  1     76.78 4730.7 909.12
## - absences             1     83.22 4737.1 909.55
## - first_gen_college    1     84.45 4738.4 909.64
## + guardian             2      0.53 4653.4 909.92
## - sex                  1     99.62 4753.5 910.65
## - romantic             1    106.82 4760.7 911.12
## - stable_learning_env  1    109.76 4763.7 911.32
## + Fjob                 4     20.56 4633.4 912.55
## - goout                1    133.89 4787.8 912.92
## - freetime             1    153.98 4807.9 914.24
## - failures             1    400.20 5054.1 930.02
##
## Step:  AIC=904.2
## G3 ~ school + sex + age + address + famsize + Pstatus + Mjob +
##     traveltime + studytime + failures + schoolsup + paid + activities +
##     nursery + internet + romantic + freetime + goout + health +
##     absences + first_gen_college + stable_learning_env + high_freq_absent
##
##                       Df Sum of Sq    RSS    AIC
## - nursery              1      4.73 4662.3 902.52
## - activities           1      4.84 4662.4 902.53
## - address              1      7.81 4665.4 902.73
## - high_freq_absent     1      8.48 4666.0 902.77
## - traveltime           1     13.44 4671.0 903.11
## - school               1     22.66 4680.2 903.73
## <none>                             4657.6 904.20
## - Pstatus              1     31.44 4689.0 904.33
## - paid                 1     47.50 4705.1 905.41
## - health               1     53.07 4710.6 905.78
## + Walc                 1      3.64 4653.9 905.95
## - famsize              1     56.38 4713.9 906.00
```

```
## + famrel              1      1.34 4656.2 906.11
## + higher              1      0.11 4657.5 906.19
## + Dalc                1      0.00 4657.6 906.20
## - internet            1     60.22 4717.8 906.26
## - Mjob                4    155.05 4812.6 906.55
## + reason              3     49.50 4608.1 906.82
## - schoolsup           1     69.70 4727.3 906.89
## - age                 1     74.42 4732.0 907.21
## - studytime           1     75.33 4732.9 907.27
## - absences            1     82.24 4739.8 907.73
## - first_gen_college   1     82.97 4740.5 907.78
## + guardian            2      0.72 4656.8 908.15
## - sex                 1     96.40 4754.0 908.67
## - romantic            1    106.64 4764.2 909.35
## - stable_learning_env  1    106.93 4764.5 909.37
## + Fjob                4     22.60 4635.0 910.66
## - freetime            1    158.03 4815.6 912.74
## - goout               1    185.17 4842.7 914.52
## - failures            1    407.34 5064.9 928.69
##
## Step:  AIC=902.52
## G3 ~ school + sex + age + address + famsize + Pstatus + Mjob +
##     traveltime + studytime + failures + schoolsup + paid + activities +
##     internet + romantic + freetime + goout + health + absences +
##     first_gen_college + stable_learning_env + high_freq_absent
##
##                      Df Sum of Sq    RSS    AIC
## - activities          1      4.90 4667.2 900.85
## - address             1      7.47 4669.8 901.03
## - high_freq_absent    1      9.19 4671.5 901.14
## - traveltime          1     13.58 4675.9 901.44
## - school              1     24.23 4686.5 902.16
## <none>                            4662.3 902.52
## - Pstatus             1     30.52 4692.8 902.58
## - paid                1     45.61 4707.9 903.60
## - health              1     53.10 4715.4 904.10
## - famsize             1     53.74 4716.0 904.14
## + nursery             1      4.73 4657.6 904.20
## + Walc                1      2.49 4659.8 904.35
## + famrel              1      0.99 4661.3 904.45
## + Dalc                1      0.25 4662.0 904.50
## + higher              1      0.11 4662.2 904.51
## - internet            1     61.57 4723.9 904.67
## - Mjob                4    153.35 4815.7 904.75
## + reason              3     48.00 4614.3 905.25
## - schoolsup           1     71.23 4733.5 905.31
## - studytime           1     72.39 4734.7 905.39
## - age                 1     73.32 4735.6 905.45
## - first_gen_college   1     79.91 4742.2 905.89
## - absences            1     83.68 4746.0 906.14
## + guardian            2      1.42 4660.9 906.42
## - sex                 1     95.32 4757.6 906.92
## - stable_learning_env  1    104.76 4767.1 907.54
## - romantic            1    106.56 4768.9 907.66
```

```
## + Fjob                 4     23.03 4639.3 908.96
## - freetime             1    161.98 4824.3 911.31
## - goout                1    186.83 4849.1 912.94
## - failures             1    406.77 5069.1 926.95
##
## Step:  AIC=900.85
## G3 ~ school + sex + age + address + famsize + Pstatus + Mjob +
##     traveltime + studytime + failures + schoolsup + paid + internet +
##     romantic + freetime + goout + health + absences + first_gen_college +
##     stable_learning_env + high_freq_absent
##
##                       Df Sum of Sq    RSS    AIC
## - address              1      8.38 4675.6 899.42
## - high_freq_absent     1      8.69 4675.9 899.44
## - traveltime           1     13.25 4680.4 899.75
## - school               1     25.74 4692.9 900.59
## <none>                             4667.2 900.85
## - Pstatus              1     33.23 4700.4 901.09
## - paid                 1     47.24 4714.4 902.03
## - health               1     53.74 4720.9 902.47
## - famsize              1     53.98 4721.2 902.49
## + activities           1      4.90 4662.3 902.52
## + nursery              1      4.79 4662.4 902.53
## + Walc                 1      2.51 4664.7 902.68
## + famrel               1      1.00 4666.2 902.78
## + higher               1      0.21 4667.0 902.84
## + Dalc                 1      0.21 4667.0 902.84
## - internet             1     61.36 4728.6 902.98
## - Mjob                 4    155.16 4822.4 903.19
## - studytime            1     69.35 4736.5 903.51
## - schoolsup            1     71.26 4738.5 903.64
## - age                  1     71.39 4738.6 903.65
## + reason               3     43.81 4623.4 903.87
## - first_gen_college    1     77.58 4744.8 904.06
## - absences             1     82.72 4749.9 904.40
## + guardian             2      1.33 4665.9 904.76
## - sex                  1     92.08 4759.3 905.03
## - stable_learning_env  1    105.11 4772.3 905.89
## - romantic             1    108.79 4776.0 906.13
## + Fjob                 4     23.01 4644.2 907.29
## - freetime             1    159.47 4826.7 909.47
## - goout                1    189.67 4856.9 911.44
## - failures             1    404.33 5071.5 925.11
##
## Step:  AIC=899.42
## G3 ~ school + sex + age + famsize + Pstatus + Mjob + traveltime +
##     studytime + failures + schoolsup + paid + internet + romantic +
##     freetime + goout + health + absences + first_gen_college +
##     stable_learning_env + high_freq_absent
##
##                       Df Sum of Sq    RSS    AIC
## - high_freq_absent     1      7.82 4683.4 897.95
## - traveltime           1     20.55 4696.1 898.81
## - school               1     20.80 4696.4 898.82
```

```
## <none>                           4675.6 899.42
## - Pstatus             1     34.46 4710.0 899.74
## - paid                1     47.52 4723.1 900.62
## + address             1      8.38 4667.2 900.85
## + activities          1      5.81 4669.8 901.03
## + nursery             1      4.43 4671.1 901.12
## - famsize             1     56.28 4731.9 901.20
## - health              1     56.68 4732.3 901.23
## + Walc                1      2.73 4672.8 901.23
## + famrel              1      1.25 4674.3 901.34
## + Dalc                1      0.38 4675.2 901.39
## + higher              1      0.21 4675.4 901.41
## - studytime           1     65.31 4740.9 901.80
## - Mjob                4    158.79 4834.4 901.97
## - age                 1     72.01 4747.6 902.25
## - schoolsup           1     72.20 4747.8 902.26
## - internet            1     74.03 4749.6 902.38
## - first_gen_college   1     76.27 4751.8 902.53
## - absences            1     78.57 4754.1 902.69
## + reason              3     39.19 4636.4 902.76
## + guardian            2      2.51 4673.1 903.25
## - sex                 1     88.92 4764.5 903.37
## - romantic            1    108.06 4783.6 904.64
## - stable_learning_env 1    113.75 4789.3 905.02
## + Fjob                4     24.15 4651.4 905.78
## - freetime            1    159.80 4835.4 908.04
## - goout               1    187.68 4863.3 909.86
## - failures            1    411.75 5087.3 924.09
##
## Step:  AIC=897.95
## G3 ~ school + sex + age + famsize + Pstatus + Mjob + traveltime +
##     studytime + failures + schoolsup + paid + internet + romantic +
##     freetime + goout + health + absences + first_gen_college +
##     stable_learning_env
##
##                      Df Sum of Sq    RSS    AIC
## - traveltime          1     17.95 4701.3 897.16
## - school              1     21.12 4704.5 897.37
## <none>                           4683.4 897.95
## - Pstatus             1     33.43 4716.8 898.20
## - paid                1     47.35 4730.7 899.13
## + high_freq_absent    1      7.82 4675.6 899.42
## + address             1      7.51 4675.9 899.44
## - famsize             1     53.42 4736.8 899.53
## + activities          1      5.24 4678.2 899.59
## - health              1     54.41 4737.8 899.60
## + nursery             1      5.09 4678.3 899.60
## + Walc                1      3.87 4679.5 899.69
## + famrel              1      1.95 4681.4 899.82
## + higher              1      0.55 4682.8 899.91
## + Dalc                1      0.01 4683.4 899.95
## - Mjob                4    155.91 4839.3 900.30
## - studytime           1     68.56 4752.0 900.54
## - schoolsup           1     69.87 4753.3 900.63
```

```
## - first_gen_college    1     73.06 4756.5 900.84
## - internet             1     73.57 4757.0 900.87
## - age                  1     74.42 4757.8 900.93
## + reason               3     36.67 4646.7 901.46
## + guardian             2      2.23 4681.2 901.80
## - sex                  1     88.92 4772.3 901.89
## - absences             1     93.93 4777.3 902.22
## - romantic             1    108.41 4791.8 903.18
## - stable_learning_env  1    117.57 4801.0 903.78
## + Fjob                 4     25.93 4657.5 904.19
## - freetime             1    156.00 4839.4 906.30
## - goout                1    191.68 4875.1 908.62
## - failures             1    418.05 5101.4 922.97
##
## Step:  AIC=897.16
## G3 ~ school + sex + age + famsize + Pstatus + Mjob + studytime +
##     failures + schoolsup + paid + internet + romantic + freetime +
##     goout + health + absences + first_gen_college + stable_learning_env
##
##                       Df Sum of Sq    RSS    AIC
## - school               1     13.12 4714.5 896.04
## <none>                             4701.3 897.16
## - Pstatus              1     37.31 4738.7 897.65
## + traveltime           1     17.95 4683.4 897.95
## + address              1     14.13 4687.2 898.21
## - famsize              1     50.02 4751.4 898.50
## - paid                 1     51.57 4752.9 898.60
## + activities           1      5.28 4696.1 898.80
## + high_freq_absent     1      5.22 4696.1 898.81
## + Walc                 1      5.04 4696.3 898.82
## + nursery              1      5.00 4696.4 898.82
## - health               1     55.38 4756.7 898.86
## + famrel               1      1.65 4699.7 899.05
## + higher               1      0.18 4701.2 899.14
## + Dalc                 1      0.00 4701.3 899.16
## - schoolsup            1     66.57 4767.9 899.60
## - age                  1     68.18 4769.5 899.71
## - studytime            1     71.57 4772.9 899.93
## - first_gen_college    1     72.21 4773.6 899.97
## - internet             1     79.66 4781.0 900.47
## - Mjob                 4    172.82 4874.2 900.56
## + reason               3     38.17 4663.2 900.58
## - sex                  1     86.00 4787.4 900.88
## + guardian             2      1.03 4700.3 901.09
## - absences             1     93.80 4795.1 901.40
## - romantic             1    114.69 4816.0 902.77
## - stable_learning_env  1    127.21 4828.6 903.59
## + Fjob                 4     23.17 4678.2 903.60
## - freetime             1    163.28 4864.6 905.94
## - goout                1    201.83 4903.2 908.44
## - failures             1    427.96 5129.3 922.69
##
## Step:  AIC=896.04
## G3 ~ sex + age + famsize + Pstatus + Mjob + studytime + failures +
```

```
##     schoolsup + paid + internet + romantic + freetime + goout +
##     health + absences + first_gen_college + stable_learning_env
##
##                       Df Sum of Sq    RSS    AIC
## <none>                             4714.5 896.04
## - Pstatus              1     35.18 4749.6 896.39
## + school               1     13.12 4701.3 897.16
## + traveltime           1      9.95 4704.5 897.37
## - paid                 1     51.98 4766.5 897.50
## - famsize              1     52.22 4766.7 897.52
## + address              1      6.79 4707.7 897.58
## + nursery              1      6.38 4708.1 897.61
## + activities           1      6.13 4708.3 897.63
## + high_freq_absent     1      5.96 4708.5 897.64
## + Walc                 1      5.08 4709.4 897.70
## - age                  1     55.64 4770.1 897.74
## - health               1     58.46 4772.9 897.93
## + famrel               1      0.75 4713.7 897.99
## + Dalc                 1      0.03 4714.4 898.04
## + higher               1      0.02 4714.5 898.04
## - studytime            1     66.05 4780.5 898.43
## - schoolsup            1     67.44 4781.9 898.53
## - first_gen_college    1     71.41 4785.9 898.79
## - Mjob                 4    169.68 4884.2 899.21
## - internet             1     78.13 4792.6 899.23
## - sex                  1     83.06 4797.5 899.56
## + reason               3     36.35 4678.1 899.59
## - absences             1     84.74 4799.2 899.67
## + guardian             2      0.71 4713.8 899.99
## - romantic             1    110.51 4825.0 901.36
## + Fjob                 4     25.81 4688.7 902.30
## - stable_learning_env  1    133.19 4847.7 902.84
## - freetime             1    168.19 4882.7 905.11
## - goout                1    206.60 4921.1 907.59
## - failures             1    436.16 5150.6 922.00
```

```r
summary(step.model)
```

```
##
## Call:
## lm(formula = G3 ~ sex + age + famsize + Pstatus + Mjob + studytime +
##     failures + schoolsup + paid + internet + romantic + freetime +
##     goout + health + absences + first_gen_college + stable_learning_env,
##     data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8038  -2.0340   0.5122   2.7302   9.4009
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            15.33022    3.53863   4.332 2.03e-05 ***
## sexM                    1.14936    0.50417   2.280 0.023338 *
## age                    -0.36877    0.19764  -1.866 0.063055 .
## famsizeLE3              0.92434    0.51135   1.808 0.071682 .
```

```
## PstatusT              -1.11457    0.75127  -1.484 0.138984
## Mjobhealth             1.45918    1.05599   1.382 0.168076
## Mjobother              0.12173    0.70292   0.173 0.862633
## Mjobservices           1.18821    0.75599   1.572 0.117087
## Mjobteacher           -1.02186    1.01413  -1.008 0.314461
## studytime              0.60088    0.29557   2.033 0.042949 *
## failures              -1.69872    0.32517  -5.224 3.31e-07 ***
## schoolsupyes          -1.47696    0.71899  -2.054 0.040835 *
## paidyes                0.88974    0.49333   1.804 0.072323 .
## internetyes            1.60975    0.72805   2.211 0.027800 *
## romanticyes           -1.32462    0.50372  -2.630 0.008995 **
## freetimelow           -1.61053    0.49645  -3.244 0.001314 **
## gooutlow               1.79373    0.49888   3.595 0.000379 ***
## healthlow              0.88593    0.46320   1.913 0.056762 .
## absences               0.06509    0.02826   2.303 0.021987 *
## first_gen_collegeyes  -1.26333    0.59763  -2.114 0.035363 *
## stable_learning_envyes -1.61319    0.55879  -2.887 0.004178 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.998 on 295 degrees of freedom
## Multiple R-squared:  0.3214, Adjusted R-squared:  0.2753
## F-statistic: 6.984 on 20 and 295 DF,  p-value: 7.548e-16
```

The new base model had Multiple R-squared: 0.3105, Adjusted R-squared: 0.2687. It had low VIF values for all predictors.

The model chosen by stepwise selection has Multiple R-squared: 0.3214, Adjusted R-squared: 0.2753.

Based on the stepwise regression model, we can see that the variables sex, studytime, failures, schoolsup, romantic, internet, freetime, goout, absences, first__gen__college, stable__learning__environment seem to be significant active predictors.

Based on these active variables, some interactions that we think could be significant are: schoolsup*failed, famsup*first__gen__college, higher*first__gen__college. Let us fit an active model with all interaction effects.

```r
activelm <- lm(G3 ~ (sex + studytime + failures + schoolsup + internet + romantic + freetime +
    goout + absences + first_gen_college + stable_learning_env)^2, data=training)

summary(activelm)
```

```
##
## Call:
## lm(formula = G3 ~ (sex + studytime + failures + schoolsup + internet +
##     romantic + freetime + goout + absences + first_gen_college +
##     stable_learning_env)^2, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1952  -2.2440   0.2061   2.2805   7.7945
##
## Coefficients: (1 not defined because of singularities)
##                                              Estimate Std. Error t value
## (Intercept)                                 12.771641   3.233918   3.949
## sexM                                        -1.208869   2.143622  -0.564
## studytime                                   -1.725767   1.210747  -1.425
## failures                                    -3.280658   1.814571  -1.808
```

```
## schoolsupyes                              2.173669   2.948642   0.737
## internetyes                               5.483829   2.984898   1.837
## romanticyes                               0.696648   2.803658   0.248
## freetimelow                              -3.173678   2.265160  -1.401
## gooutlow                                  1.418082   2.335342   0.607
## absences                                  0.190491   0.223493   0.852
## first_gen_collegeyes                     -5.716076   2.169500  -2.635
## stable_learning_envyes                   -4.185889   2.087174  -2.006
## sexM:studytime                           -0.351308   0.635536  -0.553
## sexM:failures                            -1.802727   0.861508  -2.093
## sexM:schoolsupyes                        -2.625377   1.606956  -1.634
## sexM:internetyes                         -0.236036   1.534400  -0.154
## sexM:romanticyes                          1.368767   1.075689   1.272
## sexM:freetimelow                          1.143424   1.063825   1.075
## sexM:gooutlow                            -0.782701   1.126458  -0.695
## sexM:absences                            -0.115427   0.098945  -1.167
## sexM:first_gen_collegeyes                 4.590234   1.062775   4.319
## sexM:stable_learning_envyes               2.005373   1.166237   1.720
## studytime:failures                       -0.017140   0.580027  -0.030
## studytime:schoolsupyes                   -2.612719   0.897222  -2.912
## studytime:internetyes                     0.466131   1.036462   0.450
## studytime:romanticyes                     0.309851   0.751706   0.412
## studytime:freetimelow                     0.619871   0.659915   0.939
## studytime:gooutlow                        1.122108   0.698920   1.605
## studytime:absences                       -0.004017   0.055576  -0.072
## studytime:first_gen_collegeyes            1.111382   0.660136   1.684
## studytime:stable_learning_envyes          0.635920   0.732275   0.868
## failures:schoolsupyes                     1.328830   0.967865   1.373
## failures:internetyes                     -1.724431   1.027744  -1.678
## failures:romanticyes                      0.460922   0.723757   0.637
## failures:freetimelow                     -0.679171   0.748854  -0.907
## failures:gooutlow                        -0.128550   0.746079  -0.172
## failures:absences                         0.140412   0.050705   2.769
## failures:first_gen_collegeyes             2.879308   1.086380   2.650
## failures:stable_learning_envyes           0.542664   0.820805   0.661
## schoolsupyes:internetyes                 -0.103316   2.457053  -0.042
## schoolsupyes:romanticyes                  1.031672   1.960012   0.526
## schoolsupyes:freetimelow                  1.469428   1.689940   0.870
## schoolsupyes:gooutlow                    -1.336499   1.848278  -0.723
## schoolsupyes:absences                    -0.219656   0.102927  -2.134
## schoolsupyes:first_gen_collegeyes         4.369126   1.676085   2.607
## schoolsupyes:stable_learning_envyes       0.938376   1.837391   0.511
## internetyes:romanticyes                  -1.445013   1.877426  -0.770
## internetyes:freetimelow                   0.647936   1.516222   0.427
## internetyes:gooutlow                     -1.277221   1.617826  -0.789
## internetyes:absences                     -0.286884   0.162690  -1.763
## internetyes:first_gen_collegeyes         -2.861207   1.682607  -1.700
## internetyes:stable_learning_envyes              NA         NA      NA
## romanticyes:freetimelow                  -0.896032   1.117300  -0.802
## romanticyes:gooutlow                     -1.170004   1.162587  -1.006
## romanticyes:absences                     -0.054333   0.082551  -0.658
## romanticyes:first_gen_collegeyes         -0.414515   1.116332  -0.371
## romanticyes:stable_learning_envyes       -0.390913   1.226070  -0.319
## freetimelow:gooutlow                     -1.135263   1.028015  -1.104
```

```
## freetimelow:absences                          0.105695   0.079253   1.334
## freetimelow:first_gen_collegeyes              1.027093   1.073311   0.957
## freetimelow:stable_learning_envyes           -0.883874   1.204153  -0.734
## gooutlow:absences                            -0.049473   0.088874  -0.557
## gooutlow:first_gen_collegeyes                 0.760329   1.095314   0.694
## gooutlow:stable_learning_envyes               0.416446   1.236313   0.337
## absences:first_gen_collegeyes                 0.103663   0.077193   1.343
## absences:stable_learning_envyes               0.149713   0.064099   2.336
## first_gen_collegeyes:stable_learning_envyes  0.218672   1.177958   0.186
##                                              Pr(>|t|)
## (Intercept)                                  0.000102 ***
## sexM                                         0.573302
## studytime                                    0.155296
## failures                                     0.071816 .
## schoolsupyes                                 0.461706
## internetyes                                  0.067369 .
## romanticyes                                  0.803968
## freetimelow                                  0.162429
## gooutlow                                     0.544252
## absences                                     0.394843
## first_gen_collegeyes                         0.008946 **
## stable_learning_envyes                       0.045984 *
## sexM:studytime                               0.580912
## sexM:failures                                0.037400 *
## sexM:schoolsupyes                            0.103568
## sexM:internetyes                             0.877868
## sexM:romanticyes                             0.204393
## sexM:freetimelow                             0.283490
## sexM:gooutlow                                0.487804
## sexM:absences                                0.244490
## sexM:first_gen_collegeyes                    2.26e-05 ***
## sexM:stable_learning_envyes                  0.086757 .
## studytime:failures                           0.976450
## studytime:schoolsupyes                       0.003916 **
## studytime:internetyes                        0.653293
## studytime:romanticyes                        0.680548
## studytime:freetimelow                        0.348473
## studytime:gooutlow                           0.109649
## studytime:absences                           0.942435
## studytime:first_gen_collegeyes               0.093514 .
## studytime:stable_learning_envyes             0.385999
## failures:schoolsupyes                        0.170998
## failures:internetyes                         0.094620 .
## failures:romanticyes                         0.524808
## failures:freetimelow                         0.365308
## failures:gooutlow                            0.863340
## failures:absences                            0.006041 **
## failures:first_gen_collegeyes                0.008553 **
## failures:stable_learning_envyes              0.509134
## schoolsupyes:internetyes                     0.966493
## schoolsupyes:romanticyes                     0.599105
## schoolsupyes:freetimelow                     0.385399
## schoolsupyes:gooutlow                        0.470291
## schoolsupyes:absences                        0.033809 *
```

```
## schoolsupyes:first_gen_collegeyes          0.009690 **
## schoolsupyes:stable_learning_envyes        0.610004
## internetyes:romanticyes                    0.442218
## internetyes:freetimelow                    0.669503
## internetyes:gooutlow                       0.430587
## internetyes:absences                       0.079059 .
## internetyes:first_gen_collegeyes           0.090288 .
## internetyes:stable_learning_envyes               NA
## romanticyes:freetimelow                    0.423337
## romanticyes:gooutlow                       0.315206
## romanticyes:absences                       0.511027
## romanticyes:first_gen_collegeyes           0.710715
## romanticyes:stable_learning_envyes         0.750119
## freetimelow:gooutlow                       0.270514
## freetimelow:absences                       0.183536
## freetimelow:first_gen_collegeyes           0.339523
## freetimelow:stable_learning_envyes         0.463624
## gooutlow:absences                          0.578253
## gooutlow:first_gen_collegeyes              0.488223
## gooutlow:stable_learning_envyes            0.736516
## absences:first_gen_collegeyes              0.180523
## absences:stable_learning_envyes            0.020301 *
## first_gen_collegeyes:stable_learning_envyes 0.852880
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.812 on 250 degrees of freedom
## Multiple R-squared:  0.477,  Adjusted R-squared:  0.341
## F-statistic: 3.508 on 65 and 250 DF,  p-value: 7.111e-13
```

Significant interactions exist between failures and absences, first_gen_college and failures, absences and stable_learning_env, schoolsup and absences, schoolsup and first_gen_college, sex and first_gen_college, sex and failures, studytime and schoolsup. Interestingly, in this model, the most active predictors that are not interaction terms are first_gen_college and stable_learning_env.

Fitting a pared-down active model with interaction effects:

```r
inter_lm <- lm(G3 ~ first_gen_college + stable_learning_env + failures * absences + first_gen_college*fa
summary(inter_lm)
```

```
##
## Call:
## lm(formula = G3 ~ first_gen_college + stable_learning_env + failures *
##     absences + first_gen_college * failures + absences * stable_learning_env +
##     schoolsup * absences + schoolsup * first_gen_college + sex *
##     first_gen_college + sex * failures + studytime * schoolsup,
##     data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8346  -2.2165   0.5038   2.6839   9.7555
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         11.549075   0.953942  12.107  < 2e-16 ***
## first_gen_collegeyes                -3.876900   0.734764  -5.276 2.53e-07 ***
```

```
## stable_learning_envyes           -1.066369   0.584885  -1.823 0.069266 .
## failures                         -4.082968   0.903745  -4.518 9.00e-06 ***
## absences                          0.001917   0.044024   0.044 0.965294
## schoolsupyes                     -0.039524   2.050748  -0.019 0.984636
## sexM                             -0.386046   0.748703  -0.516 0.606500
## studytime                         0.783966   0.317022   2.473 0.013956 *
## failures:absences                 0.122270   0.043933   2.783 0.005725 **
## first_gen_collegeyes:failures     2.343693   0.903151   2.595 0.009924 **
## stable_learning_envyes:absences   0.055433   0.057981   0.956 0.339818
## absences:schoolsupyes            -0.082094   0.080280  -1.023 0.307324
## first_gen_collegeyes:schoolsupyes  5.249086   1.422565   3.690 0.000266 ***
## first_gen_collegeyes:sexM         3.741658   0.988312   3.786 0.000185 ***
## failures:sexM                    -1.620103   0.621436  -2.607 0.009590 **
## schoolsupyes:studytime           -1.700906   0.764949  -2.224 0.026923 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.012 on 300 degrees of freedom
## Multiple R-squared:  0.3049, Adjusted R-squared:  0.2701
## F-statistic: 8.772 on 15 and 300 DF,  p-value: < 2.2e-16
```

```r
AIC(inter_lm)
```

```
## [1] 1792.389
```

Unfortunately even with interaction effects, the Multiple R-squared: 0.3049, Adjusted R-squared: 0.2701 and AIC is 1792.389. T

Using the model on the testing set:

```r
pred.lm <- predict(inter_lm, testing)
mse_test <- mean((pred.lm - testing$G3)^2)
mse_test
```

```
## [1] 16.71672
```

Test MSE of 16.7167.

Regression random forest

The linear model did not seem a good fit to the data. Let us try a regression random forest. Because we would prefer simpler categories in this case, we will exclude variables that have been recoded as stable_learning_env and first_gen_college. We will also include failed instead of failures.

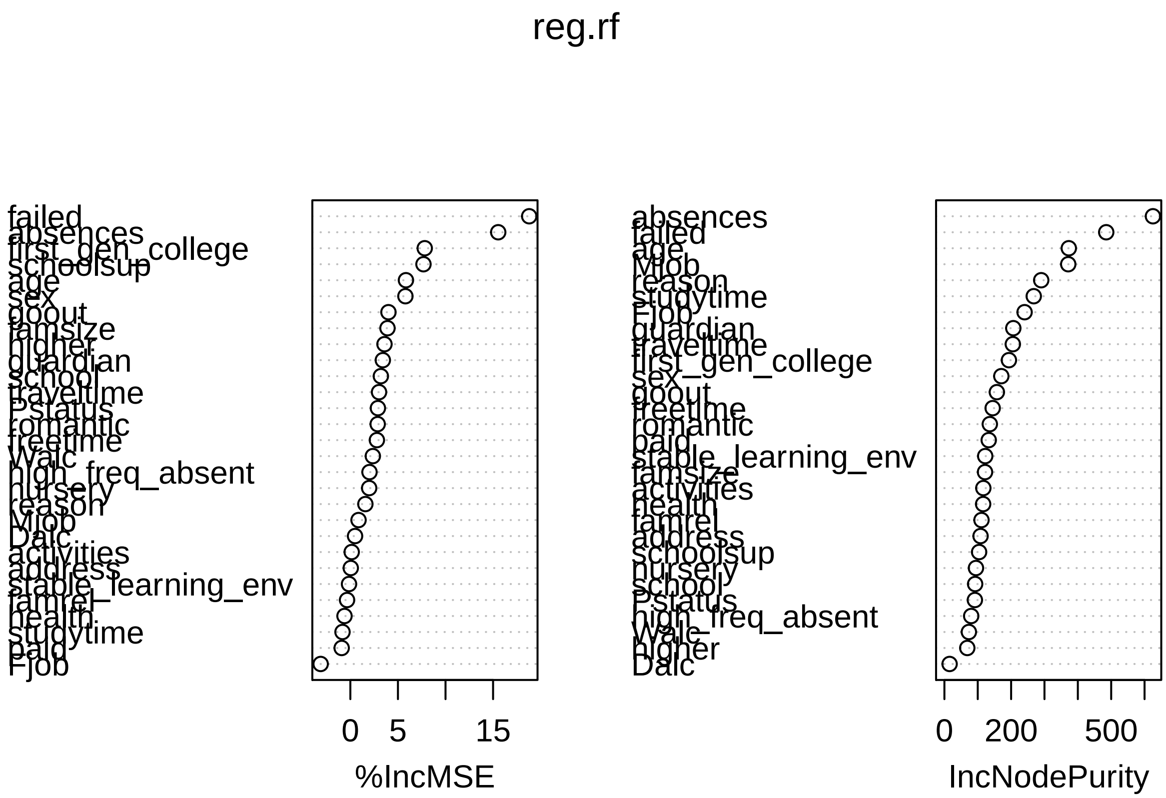
```r
library(randomForest)
reg.rf <- randomForest(G3 ~ . -G1 -G2 -G3 -ord_g3 -pf -cat_g3 -famsup -internet -failures -Medu -Fedu, d
                         importance=TRUE, na.action=na.omit)
print(reg.rf)
```

```
##
## Call:
##  randomForest(formula = G3 ~ . - G1 - G2 - G3 - ord_g3 - pf -      cat_g3 - famsup - internet - failu
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 16.60549
##                     % Var explained: 24.46
```

```r
importance(reg.rf)
```

```
##                         %IncMSE IncNodePurity
## school               3.20498228      92.14556
## sex                  5.78422413     170.60954
## age                  5.84112070     372.90154
## address              0.04182065     108.36491
## famsize              3.90407681     121.79232
## Pstatus              2.89480341      91.37820
## Mjob                 0.85982282     371.24996
## Fjob                -3.12031556     240.37053
## reason               1.57813107     290.21952
## guardian             3.41192477     206.53911
## traveltime           3.01449913     205.14705
## studytime           -0.82579688     267.85746
## schoolsup            7.68530713     104.09379
## paid                -0.91299056     132.93413
## activities           0.14021304     116.65076
## nursery              1.97637739      94.59997
## higher               3.58747512      68.54767
## romantic             2.87220658     136.18840
## famrel              -0.34544771     111.17359
## freetime             2.78903533     144.70798
## goout                4.00439708     157.30312
## Dalc                 0.49519450      15.62897
## Walc                 2.36209607      73.37783
## health              -0.61656502     116.12179
## absences            15.53928061     625.12730
## first_gen_college    7.81750171     193.41632
## stable_learning_env -0.13941147     122.34641
## high_freq_absent     2.01280460      80.33724
## failed              18.79465146     485.13377
```

```r
varImpPlot(reg.rf)
```




```r
yhat_rf <- predict(reg.rf, newdata = testing)
mse_test.rf <- mean((yhat_rf - testing$G3)^2)

mse_test.rf
```

```
## [1] 14.01952
```

Improved test MSE compared to the linear model. test MSE = 13.90083 24.94% variation explained; mean of squared residuals is 16.5.

A pared-down random forest fit with the most important predictors according to Node purity and % increase in MSE.

```r
reg.rf1 <- randomForest(G3 ~ failed + absences + schoolsup + first_gen_college + age + studytime + Psta
                        importance=TRUE, na.action=na.omit)
print(reg.rf1)
```

```
##
## Call:
##  randomForest(formula = G3 ~ failed + absences + schoolsup + first_gen_college +      age + studytim
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 15.48296
##                     % Var explained: 29.57
```

```r
importance(reg.rf1)
```

```
##                     %IncMSE IncNodePurity
## failed            24.274412      723.5997
```

```
## absences          21.464334      975.8273
## schoolsup         11.919815      180.7827
## first_gen_college  7.780720      255.6447
## age                4.578481      535.1025
## studytime          2.204100      386.4779
## Pstatus            8.244009      139.3850
## famsize            6.039078      203.2685
## guardian           6.628225      303.8222
## freetime           5.648642      215.2646
## Mjob               2.484538      574.2635
## romantic           3.745939      198.3632
## paid               2.989026      198.0420
## sex                8.073657      231.3207
## goout              2.176087      225.6308
```

```
varImpPlot(reg.rf1)
```


```
yhat_rf1 <- predict(reg.rf1, newdata = testing)
mse_test.rf1 <- mean((yhat_rf1 - testing$G3)^2)

mse_test.rf1
```

```
## [1] 14.34316
```

Test MSE of 14.39569; 30.49% of var explained by model; mean of squared residuals: 15.28173.

Overall, the random forest on regression has improved Test MSE compared to linear modeling, but still has a relatively poor fit. This indicates that perhaps considering G3 as a continuous response variable is inadequate to examine relationships between final grades and other variables.

The 4 most important factors seem to be: failed absences schoolsup first_gen_college

Multicategory ordinal logit model

Due to the way grades are assigned as values between 0 and 20, we would like to consider G3 as an ordered categorical variable with 21 levels. This would allow us to fit a multicategory ordinal logistic model to the data.

We examine the EDA and active variables in the linear model to choose the predictors in our base model.

Fitting the base model:

```
mod <-polr(ord_g3 ~ . -G1 -G2 -G3 -cat_g3 -pf, data = training)
summary(mod)
```

```
##
## Re-fitting to get Hessian
## Call:
## polr(formula = ord_g3 ~ . - G1 - G2 - G3 - cat_g3 - pf, data = training)
##
## Coefficients:
##                         Value Std. Error t value
## schoolMS              0.32605    0.38236  0.8527
## sexM                  0.56589    0.24600  2.3004
## age                  -0.28037    0.10765 -2.6045
## addressU              0.18757    0.28015  0.6695
## famsizeLE3            0.50097    0.23647  2.1185
## PstatusT             -0.32472    0.34653 -0.9371
## Medu                  0.06932    0.18259  0.3796
## Fedu                 -0.08955    0.15325 -0.5844
## Mjobhealth            0.60023    0.54994  1.0915
## Mjobother            -0.05752    0.35774 -0.1608
## Mjobservices          0.54672    0.39896  1.3704
## Mjobteacher          -0.61096    0.52074 -1.1732
## Fjobhealth           -0.40733    0.69420 -0.5868
## Fjobother            -0.17855    0.48037 -0.3717
## Fjobservices         -0.05255    0.49757 -0.1056
## Fjobteacher           0.46618    0.66467  0.7014
## reasonhome            0.23177    0.26518  0.8740
## reasonother           0.25966    0.38080  0.6819
## reasonreputation      0.39833    0.28845  1.3809
## guardianmother        0.04580    0.26486  0.1729
## guardianother         0.46437    0.48249  0.9624
## traveltime           -0.14252    0.17642 -0.8079
## studytime             0.37461    0.14709  2.5467
## failures             -0.40709    0.28580 -1.4244
## schoolsupyes         -1.07299    0.33629 -3.1907
## famsupyes            -0.49298    0.54604 -0.9028
## paidyes               0.20052    0.23526  0.8523
## activitiesyes        -0.11540    0.21736 -0.5309
## nurseryyes           -0.17096    0.26950 -0.6343
## higheryes            -0.15919    0.49945 -0.3187
## internetyes           0.39537    0.44970  0.8792
## romanticyes          -0.50355    0.23510 -2.1419
## famrellow             0.05791    0.26647  0.2173
## freetimelow          -0.67456    0.23496 -2.8709
## gooutlow              0.77564    0.24852  3.1210
## Dalclow               0.07991    0.59539  0.1342
```

```
## Walclow                0.25638    0.32121  0.7982
## healthlow              0.29549    0.21720  1.3604
## absences               0.02002    0.02009  0.9966
## first_gen_collegeyes  -0.70199    0.38869 -1.8060
## stable_learning_envyes -0.16982    0.59977 -0.2831
## high_freq_absentyes   -0.16074    0.38422 -0.4184
## failedyes             -0.91443    0.54086 -1.6907
##
## Intercepts:
##       Value   Std. Error t value
## 0|4   -7.1685  2.3463    -3.0552
## 4|5   -7.1297  2.3461    -3.0390
## 5|6   -6.9426  2.3448    -2.9609
## 6|7   -6.5614  2.3424    -2.8012
## 7|8   -6.3439  2.3413    -2.7096
## 8|9   -5.7195  2.3376    -2.4467
## 9|10  -5.3221  2.3345    -2.2798
## 10|11 -4.6194  2.3278    -1.9844
## 11|12 -3.9471  2.3222    -1.6997
## 12|13 -3.5499  2.3203    -1.5299
## 13|14 -3.0007  2.3201    -1.2933
## 14|15 -2.4425  2.3202    -1.0527
## 15|16 -1.6383  2.3195    -0.7063
## 16|17 -1.0749  2.3222    -0.4629
## 17|18 -0.7489  2.3251    -0.3221
## 18|19  0.2523  2.3411     0.1078
## 19|20  2.1350  2.5097     0.8507
##
## Residual Deviance: 1513.116
## AIC: 1633.116
```

```r
acc.ord <- predict(mod, training)
ctable <- table(training$G3, acc.ord)
round((sum(diag(ctable))/sum(ctable))*100,2)
```

```
## [1] 20.89
```

```r
ctable
```

```
##     acc.ord
##       0  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##   0  18  0  0  0  0  0  0 10  5  0  0  0  0  0  0  0  0  0
##   4   0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
##   5   3  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0
##   6   1  0  0  0  0  0  0  9  1  0  0  0  1  0  0  0  0  0
##   7   4  0  0  0  0  0  0  1  3  0  0  0  0  0  0  0  0  0
##   8  11  0  0  0  0  0  0  8  7  0  0  0  1  0  0  0  0  0
##   9   6  0  0  0  0  0  0  8  6  0  0  0  0  0  0  0  0  0
##   10  8  0  0  0  0  0  0 14 12  0  0  0  5  0  0  0  0  0
##   11  2  0  0  0  0  0  0 12 20  0  1  0  5  0  0  0  0  0
##   12  3  0  0  0  0  0  0  3 12  0  0  0  5  0  0  0  0  0
##   13  5  0  0  0  0  0  0  2 17  0  0  0  4  0  0  0  0  0
##   14  1  0  0  0  0  0  0  2 12  0  0  0  8  0  0  0  0  0
##   15  0  0  0  0  0  0  0  3  7  0  1  0 14  0  0  0  0  0
##   16  0  0  0  0  0  0  0  1  4  0  0  0  7  0  0  0  0  0
##   17  0  0  0  0  0  0  0  0  3  0  0  0  2  0  0  0  0  0
```

```
##   18  0  0  0  0  0  0  0  1  4  0  0  0  4  0  0  0  0  0
##   19  0  0  0  0  0  0  0  0  1  0  0  0  4  0  0  0  0  0
##   20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0
```

```
mod1 <- polr(ord_g3 ~ failed + high_freq_absent + romantic + internet + goout + first_gen_college + Wal
summary(mod1)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = ord_g3 ~ failed + high_freq_absent + romantic +
##     internet + goout + first_gen_college + Walc + sex + schoolsup +
##     famsup + absences + studytime + higher, data = training)
##
## Coefficients:
##                          Value Std. Error t value
## failedyes             -1.34171    0.27018 -4.9659
## high_freq_absentyes    0.03565    0.34970  0.1019
## romanticyes           -0.45766    0.22322 -2.0502
## internetyes            0.36012    0.27651  1.3024
## gooutlow               0.58461    0.22879  2.5552
## first_gen_collegeyes -0.66042    0.22335 -2.9568
## Walclow                0.33622    0.27525  1.2215
## sexM                   0.55216    0.22206  2.4865
## schoolsupyes          -0.69882    0.28905 -2.4177
## famsupyes             -0.48594    0.21594 -2.2503
## absences               0.01048    0.01726  0.6071
## studytime              0.25616    0.13427  1.9078
## higheryes              0.37412    0.45201  0.8277
##
## Intercepts:
##       Value   Std. Error t value
## 0|4   -1.6658  0.6572    -2.5347
## 4|5   -1.6293  0.6566    -2.4816
## 5|6   -1.4556  0.6534    -2.2276
## 6|7   -1.1027  0.6485    -1.7004
## 7|8   -0.9014  0.6467    -1.3939
## 8|9   -0.3219  0.6437    -0.5000
## 9|10   0.0449  0.6427     0.0698
## 10|11  0.6847  0.6438     1.0635
## 11|12  1.2954  0.6464     2.0040
## 12|13  1.6591  0.6479     2.5610
## 13|14  2.1640  0.6516     3.3212
## 14|15  2.6641  0.6579     4.0496
## 15|16  3.3940  0.6711     5.0577
## 16|17  3.9280  0.6858     5.7276
## 17|18  4.2419  0.6982     6.0754
## 18|19  5.2132  0.7670     6.7971
## 19|20  7.0544  1.1934     5.9112
##
## Residual Deviance: 1562.392
## AIC: 1622.392
```

```
(ctable <- coef(summary(mod1)))
```

```
##
## Re-fitting to get Hessian
```

```
##                            Value Std. Error     t value
## failedyes            -1.34171397 0.27018479 -4.96591237
## high_freq_absentyes   0.03565031 0.34969709  0.10194625
## romanticyes          -0.45766071 0.22322374 -2.05023314
## internetyes           0.36011512 0.27651008  1.30235803
## gooutlow              0.58461188 0.22879147  2.55521711
## first_gen_collegeyes -0.66042320 0.22335452 -2.95683831
## Walclow               0.33622440 0.27525303  1.22151026
## sexM                  0.55215988 0.22206496  2.48647905
## schoolsupyes         -0.69881781 0.28904608 -2.41766920
## famsupyes            -0.48593653 0.21594151 -2.25031553
## absences              0.01047719 0.01725748  0.60710999
## studytime             0.25615867 0.13426713  1.90782854
## higheryes             0.37411910 0.45201352  0.82767238
## 0|4                  -1.66581605 0.65721302 -2.53466683
## 4|5                  -1.62930927 0.65655494 -2.48160386
## 5|6                  -1.45558081 0.65342826 -2.22760613
## 6|7                  -1.10268571 0.64846971 -1.70044290
## 7|8                  -0.90144383 0.64670148 -1.39391027
## 8|9                  -0.32186448 0.64369588 -0.50002569
## 9|10                  0.04485165 0.64272846  0.06978319
## 10|11                 0.68472806 0.64384467  1.06349884
## 11|12                 1.29538659 0.64638677  2.00404255
## 12|13                 1.65914607 0.64785012  2.56100295
## 13|14                 2.16402771 0.65158608  3.32116935
## 14|15                 2.66414639 0.65787988  4.04959395
## 15|16                 3.39401724 0.67106437  5.05766269
## 16|17                 3.92801161 0.68580175  5.72761970
## 17|18                 4.24194365 0.69821069  6.07544927
## 18|19                 5.21317802 0.76696900  6.79711702
## 19|20                 7.05442354 1.19339185  5.91123825
```

Calculate and store p-values:

```r
p1 <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p1))
```

```
##                            Value Std. Error     t value      p value
## failedyes            -1.34171397 0.27018479 -4.96591237 6.837883e-07
## high_freq_absentyes   0.03565031 0.34969709  0.10194625 9.187993e-01
## romanticyes          -0.45766071 0.22322374 -2.05023314 4.034169e-02
## internetyes           0.36011512 0.27651008  1.30235803 1.927940e-01
## gooutlow              0.58461188 0.22879147  2.55521711 1.061216e-02
## first_gen_collegeyes -0.66042320 0.22335452 -2.95683831 3.108111e-03
## Walclow               0.33622440 0.27525303  1.22151026 2.218929e-01
## sexM                  0.55215988 0.22206496  2.48647905 1.290142e-02
## schoolsupyes         -0.69881781 0.28904608 -2.41766920 1.562027e-02
## famsupyes            -0.48593653 0.21594151 -2.25031553 2.442892e-02
## absences              0.01047719 0.01725748  0.60710999 5.437779e-01
## studytime             0.25615867 0.13426713  1.90782854 5.641338e-02
## higheryes             0.37411910 0.45201352  0.82767238 4.078561e-01
## 0|4                  -1.66581605 0.65721302 -2.53466683 1.125543e-02
## 4|5                  -1.62930927 0.65655494 -2.48160386 1.307926e-02
```

```
## 5|6                  -1.45558081 0.65342826 -2.22760613 2.590679e-02
## 6|7                  -1.10268571 0.64846971 -1.70044290 8.904765e-02
## 7|8                  -0.90144383 0.64670148 -1.39391027 1.633447e-01
## 8|9                  -0.32186448 0.64369588 -0.50002569 6.170570e-01
## 9|10                  0.04485165 0.64272846  0.06978319 9.443662e-01
## 10|11                 0.68472806 0.64384467  1.06349884 2.875558e-01
## 11|12                 1.29538659 0.64638677  2.00404255 4.506550e-02
## 12|13                 1.65914607 0.64785012  2.56100295 1.043705e-02
## 13|14                 2.16402771 0.65158608  3.32116935 8.964113e-04
## 14|15                 2.66414639 0.65787988  4.04959395 5.130658e-05
## 15|16                 3.39401724 0.67106437  5.05766269 4.244263e-07
## 16|17                 3.92801161 0.68580175  5.72761970 1.018496e-08
## 17|18                 4.24194365 0.69821069  6.07544927 1.236411e-09
## 18|19                 5.21317802 0.76696900  6.79711702 1.067334e-11
## 19|20                 7.05442354 1.19339185  5.91123825 3.395455e-09
```

Confidence intervals for parameter estimates:

```
(ci1 <- confint(mod1))
```

```
## Waiting for profiling to be done...

##
## Re-fitting to get Hessian

##                              2.5 %      97.5 %
## failedyes             -1.876197497 -0.81556618
## high_freq_absentyes   -0.656411452  0.71764477
## romanticyes           -0.897663611 -0.02188466
## internetyes           -0.180649012  0.90497042
## gooutlow               0.137337761  1.03514595
## first_gen_collegeyes -1.100400976 -0.22416437
## Walclow               -0.203641426  0.87718770
## sexM                   0.118264255  0.98945878
## schoolsupyes          -1.268171885 -0.13267874
## famsupyes             -0.910849856 -0.06369438
## absences              -0.021851056  0.04656485
## studytime             -0.006577242  0.52033547
## higheryes             -0.505211121  1.27490555
```

Analyzing the p-values and confidence intervals allows us to determine whether the coefficient estimates are significant. Based on these, failed, romantic, goout, first_gen_college, sex, schoolsup, famsup, studytime seem to be active. (Studytime is dubious, but we will include it in the next model)

Refitting a model with these predictors:

```
mod2 <- polr(ord_g3 ~ failed + romantic + goout + first_gen_college + studytime + sex + schoolsup + fams
summary(mod2)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = ord_g3 ~ failed + romantic + goout + first_gen_college +
##     studytime + sex + schoolsup + famsup, data = training)
##
## Coefficients:
##                         Value Std. Error t value
```

```
## failedyes           -1.3882     0.2606  -5.327
## romanticyes         -0.4157     0.2160  -1.924
## gooutlow             0.6874     0.2119   3.245
## first_gen_collegeyes -0.6953     0.2189  -3.177
## studytime            0.2749     0.1333   2.062
## sexM                 0.4689     0.2127   2.205
## schoolsupyes        -0.6642     0.2881  -2.305
## famsupyes           -0.4203     0.2133  -1.970
##
## Intercepts:
##       Value   Std. Error t value
## 0|4   -2.5429  0.4453    -5.7101
## 4|5   -2.5068  0.4443    -5.6425
## 5|6   -2.3344  0.4396    -5.3103
## 6|7   -1.9834  0.4324    -4.5866
## 7|8   -1.7842  0.4294    -4.1546
## 8|9   -1.2081  0.4228    -2.8573
## 9|10  -0.8412  0.4200    -2.0027
## 10|11 -0.2072  0.4175    -0.4963
## 11|12  0.3981  0.4182     0.9519
## 12|13  0.7604  0.4192     1.8142
## 13|14  1.2610  0.4225     2.9847
## 14|15  1.7560  0.4295     4.0886
## 15|16  2.4801  0.4464     5.5558
## 16|17  3.0130  0.4678     6.4414
## 17|18  3.3261  0.4857     6.8484
## 18|19  4.2924  0.5793     7.4098
## 19|20  6.1276  1.0817     5.6650
##
## Residual Deviance: 1567.543
## AIC: 1617.543
```

```r
(ctable <- coef(summary(mod2)))
```

```
##
## Re-fitting to get Hessian
```

```
##                           Value Std. Error    t value
## failedyes           -1.3881639  0.2605997 -5.3268047
## romanticyes         -0.4156752  0.2160358 -1.9241039
## gooutlow             0.6873737  0.2118547  3.2445526
## first_gen_collegeyes -0.6953188  0.2188570 -3.1770460
## studytime            0.2748992  0.1332997  2.0622646
## sexM                 0.4689357  0.2126871  2.2048153
## schoolsupyes        -0.6642084  0.2881418 -2.3051443
## famsupyes           -0.4203058  0.2133383 -1.9701374
## 0|4                 -2.5429074  0.4453387 -5.7100521
## 4|5                 -2.5067847  0.4442706 -5.6424726
## 5|6                 -2.3344286  0.4396061 -5.3102731
## 6|7                 -1.9834475  0.4324423 -4.5866175
## 7|8                 -1.7841536  0.4294407 -4.1545983
## 8|9                 -1.2081500  0.4228275 -2.8573116
## 9|10                -0.8412191  0.4200487 -2.0026706
## 10|11               -0.2072168  0.4175067 -0.4963196
## 11|12                0.3980517  0.4181620  0.9519078
```

```
## 12|13                0.7604341  0.4191666  1.8141570
## 13|14                1.2609616  0.4224815  2.9846550
## 14|15                1.7559528  0.4294754  4.0885992
## 15|16                2.4801341  0.4464006  5.5558481
## 16|17                3.0129643  0.4677514  6.4413788
## 17|18                3.3260960  0.4856779  6.8483581
## 18|19                4.2923566  0.5792787  7.4098294
## 19|20                6.1275584  1.0816527  5.6649962
```

```
p2 <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p2))
```

```
##                           Value Std. Error    t value      p value
## failedyes            -1.3881639  0.2605997 -5.3268047 9.995554e-08
## romanticyes          -0.4156752  0.2160358 -1.9241039 5.434156e-02
## gooutlow              0.6873737  0.2118547  3.2445526 1.176353e-03
## first_gen_collegeyes -0.6953188  0.2188570 -3.1770460 1.487835e-03
## studytime             0.2748992  0.1332997  2.0622646 3.918255e-02
## sexM                  0.4689357  0.2126871  2.2048153 2.746706e-02
## schoolsupyes         -0.6642084  0.2881418 -2.3051443 2.115849e-02
## famsupyes            -0.4203058  0.2133383 -1.9701374 4.882263e-02
## 0|4                  -2.5429074  0.4453387 -5.7100521 1.129416e-08
## 4|5                  -2.5067847  0.4442706 -5.6424726 1.676252e-08
## 5|6                  -2.3344286  0.4396061 -5.3102731 1.094611e-07
## 6|7                  -1.9834475  0.4324423 -4.5866175 4.504849e-06
## 7|8                  -1.7841536  0.4294407 -4.1545983 3.258595e-05
## 8|9                  -1.2081500  0.4228275 -2.8573116 4.272462e-03
## 9|10                 -0.8412191  0.4200487 -2.0026706 4.521266e-02
## 10|11                -0.2072168  0.4175067 -0.4963196 6.196689e-01
## 11|12                 0.3980517  0.4181620  0.9519078 3.411438e-01
## 12|13                 0.7604341  0.4191666  1.8141570 6.965355e-02
## 13|14                 1.2609616  0.4224815  2.9846550 2.838983e-03
## 14|15                 1.7559528  0.4294754  4.0885992 4.339860e-05
## 15|16                 2.4801341  0.4464006  5.5558481 2.762670e-08
## 16|17                 3.0129643  0.4677514  6.4413788 1.183930e-10
## 17|18                 3.3260960  0.4856779  6.8483581 7.470236e-12
## 18|19                 4.2923566  0.5792787  7.4098294 1.264620e-13
## 19|20                 6.1275584  1.0816527  5.6649962 1.470278e-08
```

```
(ci2 <- confint(mod2))
```

```
## Waiting for profiling to be done...
##
## Re-fitting to get Hessian
```

```
##                            2.5 %       97.5 %
## failedyes            -1.90462685 -0.881708189
## romanticyes          -0.84097272  0.006566793
## gooutlow              0.27365841  1.104747592
## first_gen_collegeyes -1.12660028 -0.268012697
## studytime             0.01411035  0.537164395
## sexM                  0.05297154  0.887306542
## schoolsupyes         -1.23199282 -0.100060789
## famsupyes            -0.83978194 -0.002875816
```

AIC has decreased.

Based on the p-values and confidence intervals, romantic does not seem to be significant. Let's try excluding it.

Pared-down model again:

```
mod3 <- polr(ord_g3 ~ failed + goout + first_gen_college + sex + schoolsup + studytime, data = training
summary(mod3)
```

```
## Call:
## polr(formula = ord_g3 ~ failed + goout + first_gen_college +
##     sex + schoolsup + studytime, data = training, Hess = TRUE)
##
## Coefficients:
##                        Value Std. Error t value
## failedyes            -1.4470     0.2594  -5.577
## gooutlow              0.6862     0.2115   3.244
## first_gen_collegeyes -0.5623     0.2119  -2.654
## sexM                  0.5365     0.2106   2.547
## schoolsupyes         -0.6138     0.2822  -2.175
## studytime             0.2189     0.1311   1.670
##
## Intercepts:
##       Value   Std. Error t value
## 0|4   -2.1140  0.4116    -5.1354
## 4|5   -2.0782  0.4106    -5.0620
## 5|6   -1.9083  0.4060    -4.7005
## 6|7   -1.5631  0.3995    -3.9125
## 7|8   -1.3668  0.3969    -3.4435
## 8|9   -0.7982  0.3913    -2.0401
## 9|10  -0.4358  0.3892    -1.1199
## 10|11  0.1921  0.3879     0.4952
## 11|12  0.7943  0.3898     2.0379
## 12|13  1.1538  0.3918     2.9448
## 13|14  1.6485  0.3967     4.1551
## 14|15  2.1387  0.4054     5.2759
## 15|16  2.8588  0.4247     6.7305
## 16|17  3.3897  0.4478     7.5696
## 17|18  3.7015  0.4667     7.9310
## 18|19  4.6606  0.5642     8.2598
## 19|20  6.4828  1.0738     6.0371
##
## Residual Deviance: 1574.549
## AIC: 1620.549
```

```
(ctable <- coef(summary(mod3)))
```

```
##                           Value Std. Error    t value
## failedyes            -1.4470044  0.2594410 -5.5773921
## gooutlow              0.6862095  0.2115274  3.2440692
## first_gen_collegeyes -0.5623425  0.2119141 -2.6536341
## sexM                  0.5364682  0.2106266  2.5470106
## schoolsupyes         -0.6138372  0.2821635 -2.1754659
## studytime             0.2188984  0.1310851  1.6698957
## 0|4                  -2.1139669  0.4116495 -5.1353568
## 4|5                  -2.0782289  0.4105540 -5.0620109
## 5|6                  -1.9083008  0.4059759 -4.7005269
```

```
## 6|7                   -1.5630672  0.3995011 -3.9125480
## 7|8                   -1.3667896  0.3969204 -3.4434854
## 8|9                   -0.7982483  0.3912773 -2.0401089
## 9|10                  -0.4358495  0.3891718 -1.1199410
## 10|11                  0.1920664  0.3878675  0.4951856
## 11|12                  0.7942916  0.3897682  2.0378563
## 12|13                  1.1537589  0.3917941  2.9448095
## 13|14                  1.6485090  0.3967448  4.1550867
## 14|15                  2.1387107  0.4053722  5.2759186
## 15|16                  2.8587993  0.4247500  6.7305461
## 16|17                  3.3897018  0.4478022  7.5696401
## 17|18                  3.7014831  0.4667133  7.9309568
## 18|19                  4.6605957  0.5642498  8.2598092
## 19|20                  6.4827907  1.0738176  6.0371431
```

```
p3 <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p3))
```

```
##                            Value Std. Error    t value      p value
## failedyes             -1.4470044  0.2594410 -5.5773921 2.441512e-08
## gooutlow               0.6862095  0.2115274  3.2440692 1.178351e-03
## first_gen_collegeyes -0.5623425  0.2119141 -2.6536341 7.963013e-03
## sexM                   0.5364682  0.2106266  2.5470106 1.086501e-02
## schoolsupyes          -0.6138372  0.2821635 -2.1754659 2.959522e-02
## studytime              0.2188984  0.1310851  1.6698957 9.494000e-02
## 0|4                   -2.1139669  0.4116495 -5.1353568 2.816093e-07
## 4|5                   -2.0782289  0.4105540 -5.0620109 4.148574e-07
## 5|6                   -1.9083008  0.4059759 -4.7005269 2.594911e-06
## 6|7                   -1.5630672  0.3995011 -3.9125480 9.132736e-05
## 7|8                   -1.3667896  0.3969204 -3.4434854 5.742675e-04
## 8|9                   -0.7982483  0.3912773 -2.0401089 4.133948e-02
## 9|10                  -0.4358495  0.3891718 -1.1199410 2.627389e-01
## 10|11                  0.1920664  0.3878675  0.4951856 6.204691e-01
## 11|12                  0.7942916  0.3897682  2.0378563 4.156431e-02
## 12|13                  1.1537589  0.3917941  2.9448095 3.231536e-03
## 13|14                  1.6485090  0.3967448  4.1550867 3.251642e-05
## 14|15                  2.1387107  0.4053722  5.2759186 1.320927e-07
## 15|16                  2.8587993  0.4247500  6.7305461 1.690275e-11
## 16|17                  3.3897018  0.4478022  7.5696401 3.742596e-14
## 17|18                  3.7014831  0.4667133  7.9309568 2.174638e-15
## 18|19                  4.6605957  0.5642498  8.2598092 1.459055e-16
## 19|20                  6.4827907  1.0738176  6.0371431 1.568666e-09
```

```
(ci3 <- confint(mod3))
```

```
## Waiting for profiling to be done...
```

```
##                             2.5 %      97.5 %
## failedyes            -1.96148845 -0.94318819
## gooutlow              0.27307924  1.10288470
## first_gen_collegeyes -0.97950820 -0.14817187
## sexM                  0.12485281  0.95095558
## schoolsupyes         -1.16913473 -0.06070744
## studytime            -0.03766773  0.47670960
```

All predictors are significant, but AIC has increased compared to mod2.

Evaluating accuracy of the model for the training set:

```
acc.ord3 <- predict(mod3, training)
ctable <- table(training$G3, acc.ord3)
round((sum(diag(ctable))/sum(ctable))*100,2)
```

```
## [1] 17.09
```

```
ctable
```

```
##     acc.ord3
##        0  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##   0   16  0  0  0  0  0  0  7  7  0  0  0  3  0  0  0  0  0
##   4    1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   5    2  0  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0
##   6    1  0  0  0  0  0  0  6  5  0  0  0  0  0  0  0  0  0
##   7    5  0  0  0  0  0  0  0  1  0  0  0  2  0  0  0  0  0
##   8   12  0  0  0  0  0  0  4  9  0  0  0  2  0  0  0  0  0
##   9    6  0  0  0  0  0  0  4  9  0  0  0  1  0  0  0  0  0
##   10   8  0  0  0  0  0  0  6 20  0  0  0  5  0  0  0  0  0
##   11   4  0  0  0  0  0  0  8 23  0  0  0  5  0  0  0  0  0
##   12   3  0  0  0  0  0  0  5 11  0  0  0  4  0  0  0  0  0
##   13   5  0  0  0  0  0  0  3 17  0  0  0  3  0  0  0  0  0
##   14   1  0  0  0  0  0  0  3  8  0  0  0 11  0  0  0  0  0
##   15   0  0  0  0  0  0  0  2 14  0  0  0  9  0  0  0  0  0
##   16   0  0  0  0  0  0  0  2  6  0  0  0  4  0  0  0  0  0
##   17   0  0  0  0  0  0  0  1  3  0  0  0  1  0  0  0  0  0
##   18   1  0  0  0  0  0  0  1  5  0  0  0  2  0  0  0  0  0
##   19   0  0  0  0  0  0  0  0  3  0  0  0  2  0  0  0  0  0
##   20   0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0
```

Very terrible accuracy even for the training set.

What if we add interaction terms?

Let's base our interaction terms on the discussion for the linear model.

```
mod4 <- polr(ord_g3 ~ failed + goout + romantic +  first_gen_college + sex + schoolsup + sex*schoolsup +
summary(mod4)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = ord_g3 ~ failed + goout + romantic + first_gen_college +
##     sex + schoolsup + sex * schoolsup + sex * first_gen_college +
##     schoolsup * failed + schoolsup * studytime + schoolsup *
##     first_gen_college + studytime * famsup, data = training)
##
## Coefficients:
##                                     Value Std. Error  t value
## failedyes                        -1.72945     0.2981 -5.80124
## gooutlow                          0.67401     0.2169  3.10806
## romanticyes                      -0.48910     0.2204 -2.21930
## first_gen_collegeyes             -1.43664     0.3180 -4.51738
## sexM                             -0.01446     0.3391 -0.04265
## schoolsupyes                      0.73242     0.9305  0.78712
## studytime                         0.36175     0.2487  1.45475
```

```
## famsupyes                         -0.76421     0.5894 -1.29664
## sexM:schoolsupyes                 -1.06747     0.6217 -1.71701
## first_gen_collegeyes:sexM          1.06194     0.4232  2.50904
## failedyes:schoolsupyes             1.15774     0.6593  1.75607
## schoolsupyes:studytime            -1.11414     0.3374 -3.30178
## first_gen_collegeyes:schoolsupyes  1.59232     0.6093  2.61341
## studytime:famsupyes                0.14783     0.2850  0.51876
##
## Intercepts:
##       Value   Std. Error t value
## 0|4   -3.0101  0.5949    -5.0599
## 4|5   -2.9717  0.5939    -5.0034
## 5|6   -2.7903  0.5898    -4.7309
## 6|7   -2.4230  0.5833    -4.1537
## 7|8   -2.2159  0.5806    -3.8163
## 8|9   -1.6123  0.5757    -2.8006
## 9|10  -1.2173  0.5741    -2.1202
## 10|11 -0.5423  0.5738    -0.9451
## 11|12  0.0965  0.5744     0.1680
## 12|13  0.4842  0.5739     0.8438
## 13|14  1.0192  0.5741     1.7752
## 14|15  1.5345  0.5779     2.6551
## 15|16  2.2702  0.5892     3.8531
## 16|17  2.8082  0.6045     4.6455
## 17|18  3.1241  0.6180     5.0554
## 18|19  4.0937  0.6937     5.9014
## 19|20  5.9292  1.1498     5.1569
##
## Residual Deviance: 1537.762
## AIC: 1599.762
```

```r
(ctable <- coef(summary(mod4)))
```

```
##
## Re-fitting to get Hessian
```

```
##                                         Value Std. Error     t value
## failedyes                         -1.72945368  0.2981178 -5.80124295
## gooutlow                           0.67400726  0.2168580  3.10805786
## romanticyes                       -0.48909643  0.2203829 -2.21930313
## first_gen_collegeyes              -1.43663982  0.3180252 -4.51737778
## sexM                              -0.01446239  0.3390652 -0.04265371
## schoolsupyes                       0.73241720  0.9305008  0.78712151
## studytime                          0.36174519  0.2486653  1.45474731
## famsupyes                         -0.76421191  0.5893770 -1.29664348
## sexM:schoolsupyes                 -1.06746616  0.6217004 -1.71701055
## first_gen_collegeyes:sexM          1.06194139  0.4232461  2.50904014
## failedyes:schoolsupyes             1.15773926  0.6592797  1.75606684
## schoolsupyes:studytime            -1.11413949  0.3374358 -3.30178248
## first_gen_collegeyes:schoolsupyes  1.59232262  0.6092889  2.61341153
## studytime:famsupyes                0.14782789  0.2849627  0.51876227
## 0|4                               -3.01008158  0.5948850 -5.05993887
## 4|5                               -2.97166463  0.5939322 -5.00337322
## 5|6                               -2.79029263  0.5898075 -4.73085276
## 6|7                               -2.42295680  0.5833207 -4.15373034
```

```
## 7|8                            -2.21585268  0.5806252 -3.81632165
## 8|9                            -1.61234628  0.5757246 -2.80055115
## 9|10                           -1.21732081  0.5741468 -2.12022571
## 10|11                          -0.54231022  0.5738069 -0.94510930
## 11|12                           0.09649365  0.5743725  0.16799839
## 12|13                           0.48423331  0.5738879  0.84377688
## 13|14                           1.01921806  0.5741418  1.77520259
## 14|15                           1.53449674  0.5779426  2.65510251
## 15|16                           2.27018077  0.5891862  3.85307839
## 16|17                           2.80819116  0.6045010  4.64546978
## 17|18                           3.12405658  0.6179583  5.05544881
## 18|19                           4.09365582  0.6936810  5.90135194
## 19|20                           5.92920227  1.1497705  5.15685736
```

```r
p4 <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p4))
```

```
##                                      Value Std. Error     t value
## failedyes                      -1.72945368  0.2981178 -5.80124295
## gooutlow                        0.67400726  0.2168580  3.10805786
## romanticyes                    -0.48909643  0.2203829 -2.21930313
## first_gen_collegeyes           -1.43663982  0.3180252 -4.51737778
## sexM                           -0.01446239  0.3390652 -0.04265371
## schoolsupyes                    0.73241720  0.9305008  0.78712151
## studytime                       0.36174519  0.2486653  1.45474731
## famsupyes                      -0.76421191  0.5893770 -1.29664348
## sexM:schoolsupyes              -1.06746616  0.6217004 -1.71701055
## first_gen_collegeyes:sexM       1.06194139  0.4232461  2.50904014
## failedyes:schoolsupyes          1.15773926  0.6592797  1.75606684
## schoolsupyes:studytime         -1.11413949  0.3374358 -3.30178248
## first_gen_collegeyes:schoolsupyes  1.59232262  0.6092889  2.61341153
## studytime:famsupyes             0.14782789  0.2849627  0.51876227
## 0|4                            -3.01008158  0.5948850 -5.05993887
## 4|5                            -2.97166463  0.5939322 -5.00337322
## 5|6                            -2.79029263  0.5898075 -4.73085276
## 6|7                            -2.42295680  0.5833207 -4.15373034
## 7|8                            -2.21585268  0.5806252 -3.81632165
## 8|9                            -1.61234628  0.5757246 -2.80055115
## 9|10                           -1.21732081  0.5741468 -2.12022571
## 10|11                          -0.54231022  0.5738069 -0.94510930
## 11|12                           0.09649365  0.5743725  0.16799839
## 12|13                           0.48423331  0.5738879  0.84377688
## 13|14                           1.01921806  0.5741418  1.77520259
## 14|15                           1.53449674  0.5779426  2.65510251
## 15|16                           2.27018077  0.5891862  3.85307839
## 16|17                           2.80819116  0.6045010  4.64546978
## 17|18                           3.12405658  0.6179583  5.05544881
## 18|19                           4.09365582  0.6936810  5.90135194
## 19|20                           5.92920227  1.1497705  5.15685736
##                                     p value
## failedyes                      6.582515e-09
## gooutlow                       1.883212e-03
## romanticyes                    2.646611e-02
## first_gen_collegeyes           6.261014e-06
## sexM                           9.659776e-01
```

```
## schoolsupyes                      4.312107e-01
## studytime                         1.457392e-01
## famsupyes                         1.947539e-01
## sexM:schoolsupyes                 8.597724e-02
## first_gen_collegeyes:sexM         1.210597e-02
## failedyes:schoolsupyes            7.907700e-02
## schoolsupyes:studytime            9.607254e-04
## first_gen_collegeyes:schoolsupyes 8.964329e-03
## studytime:famsupyes               6.039265e-01
## 0|4                               4.193909e-07
## 4|5                               5.633572e-07
## 5|6                               2.235787e-06
## 6|7                               3.270986e-05
## 7|8                               1.354559e-04
## 8|9                               5.101542e-03
## 9|10                              3.398702e-02
## 10|11                             3.446031e-01
## 11|12                             8.665845e-01
## 12|13                             3.987941e-01
## 13|14                             7.586444e-02
## 14|15                             7.928431e-03
## 15|16                             1.166420e-04
## 16|17                             3.393034e-06
## 17|18                             4.293798e-07
## 18|19                             3.605349e-09
## 19|20                             2.511290e-07
```

```r
(ci4 <- confint(mod4))
```

```
## Waiting for profiling to be done...
##
## Re-fitting to get Hessian
```

```
##                                        2.5 %      97.5 %
## failedyes                         -2.3214364 -1.15123396
## gooutlow                           0.2503920  1.10114465
## romanticyes                       -0.9230151 -0.05837766
## first_gen_collegeyes              -2.0646870 -0.81670972
## sexM                              -0.6799620  0.65071516
## schoolsupyes                      -1.0998497  2.56413384
## studytime                         -0.1227550  0.85203523
## famsupyes                         -1.9233145  0.38932033
## sexM:schoolsupyes                 -2.2962304  0.15377624
## first_gen_collegeyes:sexM          0.2363092  1.89633281
## failedyes:schoolsupyes            -0.1337302  2.46425196
## schoolsupyes:studytime            -1.7820611 -0.45214925
## first_gen_collegeyes:schoolsupyes  0.4009686  2.79629741
## studytime:famsupyes               -0.4093914  0.70866271
```

AIC has decreased significantly compared to the previous models without interaction terms, by nearly 20. However, in this model, sex, its interaction with schoolsup, and its interaction with first_gen_college all seem to be insignificant. The interaction between studytime and famsup and failed and schoolsup do not seem significant either, so let us remove it to pare down the model:

```r
mod5 <- polr(ord_g3 ~ failed + goout + romantic + schoolsup + first_gen_college + schoolsup * studytime
summary(mod5)
```

```
##
## Re-fitting to get Hessian
## Call:
## polr(formula = ord_g3 ~ failed + goout + romantic + schoolsup +
##     first_gen_college + schoolsup * studytime + schoolsup * first_gen_college,
##     data = training)
##
## Coefficients:
##                                     Value Std. Error  t value
## failedyes                        -1.39146     0.2622 -5.30667
## gooutlow                          0.59649     0.2134  2.79476
## romanticyes                      -0.50391     0.2163 -2.32943
## schoolsupyes                      0.03364     0.8300  0.04052
## first_gen_collegeyes             -0.85007     0.2317 -3.66947
## studytime                         0.29305     0.1393  2.10388
## schoolsupyes:studytime           -0.85318     0.3244 -2.62985
## schoolsupyes:first_gen_collegeyes  1.52426     0.5720  2.66499
##
## Intercepts:
##       Value   Std. Error t value
## 0|4   -2.6921  0.4042    -6.6607
## 4|5   -2.6558  0.4029    -6.5914
## 5|6   -2.4839  0.3975    -6.2485
## 6|7   -2.1376  0.3891    -5.4942
## 7|8   -1.9409  0.3852    -5.0386
## 8|9   -1.3590  0.3772    -3.6032
## 9|10  -0.9824  0.3739    -2.6277
## 10|11 -0.3366  0.3707    -0.9081
## 11|12  0.2744  0.3708     0.7400
## 12|13  0.6418  0.3719     1.7259
## 13|14  1.1514  0.3752     3.0688
## 14|15  1.6489  0.3824     4.3124
## 15|16  2.3707  0.4009     5.9140
## 16|17  2.9048  0.4245     6.8436
## 17|18  3.2188  0.4440     7.2496
## 18|19  4.1810  0.5450     7.6720
## 19|20  6.0069  1.0636     5.6479
##
## Residual Deviance: 1563.245
## AIC: 1613.245
```

```
(ctable <- coef(summary(mod5)))
```

```
##
## Re-fitting to get Hessian
##                                       Value Std. Error     t value
## failedyes                        -1.39146287  0.2622101 -5.30667226
## gooutlow                          0.59648735  0.2134306  2.79475999
## romanticyes                      -0.50390805  0.2163229 -2.32942552
## schoolsupyes                      0.03363511  0.8300358  0.04052249
## first_gen_collegeyes             -0.85006782  0.2316595 -3.66947101
## studytime                         0.29304592  0.1392885  2.10387705
## schoolsupyes:studytime           -0.85318089  0.3244225 -2.62984500
## schoolsupyes:first_gen_collegeyes  1.52425650  0.5719558  2.66498990
```

```
## 0|4                           -2.69208825  0.4041720 -6.66074948
## 4|5                           -2.65581637  0.4029201 -6.59142141
## 5|6                           -2.48392676  0.3975209 -6.24854363
## 6|7                           -2.13758424  0.3890606 -5.49421935
## 7|8                           -1.94093740  0.3852119 -5.03862266
## 8|9                           -1.35895482  0.3771508 -3.60321358
## 9|10                          -0.98243182  0.3738798 -2.62766733
## 10|11                         -0.33664864  0.3707238 -0.90808488
## 11|12                          0.27441039  0.3708471  0.73995555
## 12|13                          0.64181542  0.3718808  1.72586325
## 13|14                          1.15140976  0.3751961  3.06882090
## 14|15                          1.64894914  0.3823753  4.31238440
## 15|16                          2.37070007  0.4008641  5.91397458
## 16|17                          2.90477643  0.4244524  6.84358578
## 17|18                          3.21877485  0.4439939  7.24959282
## 18|19                          4.18096474  0.5449623  7.67202626
## 19|20                          6.00685744  1.0635640  5.64785684
```

```r
p5 <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p5))
```

```
##                                      Value Std. Error     t value
## failedyes                      -1.39146287  0.2622101 -5.30667226
## gooutlow                        0.59648735  0.2134306  2.79475999
## romanticyes                    -0.50390805  0.2163229 -2.32942552
## schoolsupyes                    0.03363511  0.8300358  0.04052249
## first_gen_collegeyes           -0.85006782  0.2316595 -3.66947101
## studytime                       0.29304592  0.1392885  2.10387705
## schoolsupyes:studytime         -0.85318089  0.3244225 -2.62984500
## schoolsupyes:first_gen_collegeyes  1.52425650  0.5719558  2.66498990
## 0|4                            -2.69208825  0.4041720 -6.66074948
## 4|5                            -2.65581637  0.4029201 -6.59142141
## 5|6                            -2.48392676  0.3975209 -6.24854363
## 6|7                            -2.13758424  0.3890606 -5.49421935
## 7|8                            -1.94093740  0.3852119 -5.03862266
## 8|9                            -1.35895482  0.3771508 -3.60321358
## 9|10                           -0.98243182  0.3738798 -2.62766733
## 10|11                          -0.33664864  0.3707238 -0.90808488
## 11|12                           0.27441039  0.3708471  0.73995555
## 12|13                           0.64181542  0.3718808  1.72586325
## 13|14                           1.15140976  0.3751961  3.06882090
## 14|15                           1.64894914  0.3823753  4.31238440
## 15|16                           2.37070007  0.4008641  5.91397458
## 16|17                           2.90477643  0.4244524  6.84358578
## 17|18                           3.21877485  0.4439939  7.24959282
## 18|19                           4.18096474  0.5449623  7.67202626
## 19|20                           6.00685744  1.0635640  5.64785684
##                                     p value
## failedyes                      1.116447e-07
## gooutlow                       5.193826e-03
## romanticyes                    1.983653e-02
## schoolsupyes                   9.676766e-01
## first_gen_collegeyes           2.430529e-04
## studytime                      3.538917e-02
## schoolsupyes:studytime         8.542381e-03
```

```
## schoolsupyes:first_gen_collegeyes 7.699064e-03
## 0|4                              2.724347e-11
## 4|5                              4.356354e-11
## 5|6                              4.142975e-10
## 6|7                              3.924425e-08
## 7|8                              4.688939e-07
## 8|9                              3.143071e-04
## 9|10                             8.597255e-03
## 10|11                            3.638334e-01
## 11|12                            4.593270e-01
## 12|13                            8.437202e-02
## 13|14                            2.149054e-03
## 14|15                            1.615033e-05
## 15|16                            3.339494e-09
## 16|17                            7.723504e-12
## 17|18                            4.180264e-13
## 18|19                            1.693003e-14
## 19|20                            1.624604e-08
```

```
(ci5 <- confint(mod5))
```

```
## Waiting for profiling to be done...
##
## Re-fitting to get Hessian
```

```
##                                          2.5 %      97.5 %
## failedyes                          -1.91099915 -0.88183866
## gooutlow                            0.17940546  1.01668340
## romanticyes                        -0.92995496 -0.08125812
## schoolsupyes                       -1.60642853  1.65956106
## first_gen_collegeyes               -1.30698473 -0.39813983
## studytime                           0.02039631  0.56697874
## schoolsupyes:studytime             -1.49158352 -0.21490640
## schoolsupyes:first_gen_collegeyes  0.40814640  2.65646321
```

This has resulted in an increase in the AIC, which is still lower than the first three models.

Let's check the accuracy of this model with interaction terms:

```
acc.ord4 <- predict(mod4, training)
ctable <- table(training$G3, acc.ord4)
round((sum(diag(ctable))/sum(ctable))*100,2)
```

```
## [1] 19.94
```

```
ctable
```

```
##     acc.ord4
##       0  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##   0  18  0  0  0  0  0  0 10  3  0  1  0  1  0  0  0  0  0
##   4   1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   5   4  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
##   6   1  0  0  0  0  0  0 10  1  0  0  0  0  0  0  0  0  0
##   7   4  0  0  0  0  0  0  1  1  0  0  0  2  0  0  0  0  0
##   8  14  0  0  0  0  0  0  7  3  0  0  0  3  0  0  0  0  0
##   9   8  0  0  0  0  0  0  4  6  0  2  0  0  0  0  0  0  0
##   10  5  0  0  0  0  0  0 14 10  0  2  0  8  0  0  0  0  0
##   11  5  0  0  0  0  0  0 13 15  0  1  0  6  0  0  0  0  0
```

```
##   12  3  0  0  0  0  0  0  7  9  0  1  0  3  0  0  0  0  0
##   13  4  0  0  0  0  0  0  5 10  0  3  0  6  0  0  0  0  0
##   14  0  0  0  0  0  0  0  2  8  0  4  0  9  0  0  0  0  0
##   15  0  0  0  0  0  0  0  3  8  0  1  0 13  0  0  0  0  0
##   16  0  0  0  0  0  0  0  0  9  0  0  0  3  0  0  0  0  0
##   17  0  0  0  0  0  0  0  0  3  0  0  0  2  0  0  0  0  0
##   18  1  0  0  0  0  0  0  0  4  0  2  0  2  0  0  0  0  0
##   19  0  0  0  0  0  0  0  0  2  0  0  0  3  0  0  0  0  0
##   20  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0
```

The accuracy is even lower than mod3, at only 19.94% for the training set.

Checking on testing set:

```
pred.ord3 <- predict(mod3, testing)
ctable <- table(testing$G3, pred.ord3)
round((sum(diag(ctable))/sum(ctable))*100,2)
```

```
## [1] 8.86
```

```
pred.ord4 <- predict(mod4, testing)
ctable <- table(testing$G3, pred.ord4)
round((sum(diag(ctable))/sum(ctable))*100,2)
```

```
## [1] 11.39
```

```
pred.ord5 <- predict(mod5, testing)
ctable <- table(testing$G3, pred.ord5)
round((sum(diag(ctable))/sum(ctable))*100,2)
```

```
## [1] 10.13
```

Accuracy rates are even lower, at 8.86%, 11.39%, and 10.13%.

Highly inaccurate model, not a good fit for the data.

6-category grades modeling

```
set.seed(3)
train_ind <- sample(x = nrow(data), size = 0.8 * nrow(data))
test_ind_neg <- -train_ind
ftrain <- data[train_ind, ]
ftest <- data[test_ind_neg, ]
```

Trying out a multicat ordinal logit on this:

```
mod6 <- polr(cat_g3 ~ failed + goout + romantic + schoolsup + first_gen_college + schoolsup * studytime
summary(mod6)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = cat_g3 ~ failed + goout + romantic + schoolsup +
##     first_gen_college + schoolsup * studytime + schoolsup * first_gen_college,
##     data = ftrain)
##
## Coefficients:
##                                        Value Std. Error  t value
```

```
## failedyes                          -1.55916     0.2803 -5.56248
## gooutlow                            0.58202     0.2260  2.57501
## romanticyes                        -0.68049     0.2310 -2.94618
## schoolsupyes                        0.01527     0.8809  0.01734
## first_gen_collegeyes               -0.87441     0.2449 -3.56977
## studytime                           0.31024     0.1444  2.14898
## schoolsupyes:studytime             -0.89744     0.3472 -2.58483
## schoolsupyes:first_gen_collegeyes  1.78767     0.6142  2.91072
##
## Intercepts:
##                     Value   Std. Error t value
## Poor|Weak           -2.8407  0.4234    -6.7090
## Weak|Sufficient     -1.0846  0.3904    -2.7781
## Sufficient|Good      1.1195  0.3899     2.8716
## Good|Very Good       2.3538  0.4154     5.6661
## Very Good|Excellent  3.2057  0.4575     7.0075
##
## Residual Deviance: 873.2587
## AIC: 899.2587
```

```r
(ctable <- coef(summary(mod6)))
```

```
##
## Re-fitting to get Hessian
```

```
##                                         Value Std. Error     t value
## failedyes                          -1.55915962  0.2802994 -5.56247945
## gooutlow                            0.58201779  0.2260254  2.57501100
## romanticyes                        -0.68048602  0.2309721 -2.94618281
## schoolsupyes                        0.01527404  0.8808503  0.01734011
## first_gen_collegeyes               -0.87441246  0.2449495 -3.56976600
## studytime                           0.31023730  0.1443649  2.14898050
## schoolsupyes:studytime             -0.89744352  0.3471959 -2.58483360
## schoolsupyes:first_gen_collegeyes  1.78767301  0.6141683  2.91072192
## Poor|Weak                          -2.84071658  0.4234218 -6.70895259
## Weak|Sufficient                    -1.08462029  0.3904224 -2.77806907
## Sufficient|Good                     1.11949859  0.3898567  2.87156458
## Good|Very Good                      2.35375515  0.4154121  5.66607292
## Very Good|Excellent                 3.20565589  0.4574631  7.00746349
```

```r
p6 <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p6))
```

```
##                                         Value Std. Error     t value
## failedyes                          -1.55915962  0.2802994 -5.56247945
## gooutlow                            0.58201779  0.2260254  2.57501100
## romanticyes                        -0.68048602  0.2309721 -2.94618281
## schoolsupyes                        0.01527404  0.8808503  0.01734011
## first_gen_collegeyes               -0.87441246  0.2449495 -3.56976600
## studytime                           0.31023730  0.1443649  2.14898050
## schoolsupyes:studytime             -0.89744352  0.3471959 -2.58483360
## schoolsupyes:first_gen_collegeyes  1.78767301  0.6141683  2.91072192
## Poor|Weak                          -2.84071658  0.4234218 -6.70895259
## Weak|Sufficient                    -1.08462029  0.3904224 -2.77806907
## Sufficient|Good                     1.11949859  0.3898567  2.87156458
## Good|Very Good                      2.35375515  0.4154121  5.66607292
```

```
## Very Good|Excellent               3.20565589  0.4574631  7.00746349
##                                       p value
## failedyes                        2.659684e-08
## gooutlow                         1.002369e-02
## romanticyes                      3.217222e-03
## schoolsupyes                     9.861653e-01
## first_gen_collegeyes             3.573003e-04
## studytime                        3.163595e-02
## schoolsupyes:studytime           9.742600e-03
## schoolsupyes:first_gen_collegeyes 3.605948e-03
## Poor|Weak                        1.960263e-11
## Weak|Sufficient                  5.468299e-03
## Sufficient|Good                  4.084453e-03
## Good|Very Good                   1.461074e-08
## Very Good|Excellent              2.426773e-12
```

```r
(ci5 <- confint(mod6))
```

```
## Waiting for profiling to be done...
##
## Re-fitting to get Hessian
```

```
##                                        2.5 %     97.5 %
## failedyes                        -2.11603937 -1.0156926
## gooutlow                          0.14100166  1.0278318
## romanticyes                      -1.13625714 -0.2299976
## schoolsupyes                     -1.71834716  1.7492857
## first_gen_collegeyes             -1.35840403 -0.3972182
## studytime                         0.02754372  0.5941023
## schoolsupyes:studytime           -1.58076549 -0.2136889
## schoolsupyes:first_gen_collegeyes  0.58713044  3.0005349
```

```r
acc.ord6 <- predict(mod6, ftrain)
ctable <- table(ftrain$cat_g3, acc.ord6)
ctable
```

```
##              acc.ord6
##               Poor Weak Sufficient Good Very Good Excellent
##   Poor           3   10         20    0         0         0
##   Weak           5   22         46    0         0         0
##   Sufficient     2   13        113    2         0         0
##   Good           0    0         47    1         0         0
##   Very Good      0    0         17    0         0         0
##   Excellent      0    0         14    1         0         0
```

Still not very accurate for the training

Random forest:

```r
rf.cat<-randomForest(cat_g3~. -G1 -G2 -G3 -ord_g3 -pf -famsup -internet -Medu -Fedu,data = ftrain, mtry
print(rf.cat)
```

```
##
## Call:
##  randomForest(formula = cat_g3 ~ . - G1 - G2 - G3 - ord_g3 - pf -      famsup - internet - Medu - Fe
##                Type of random forest: classification
##                      Number of trees: 50
## No. of variables tried at each split: 3
```

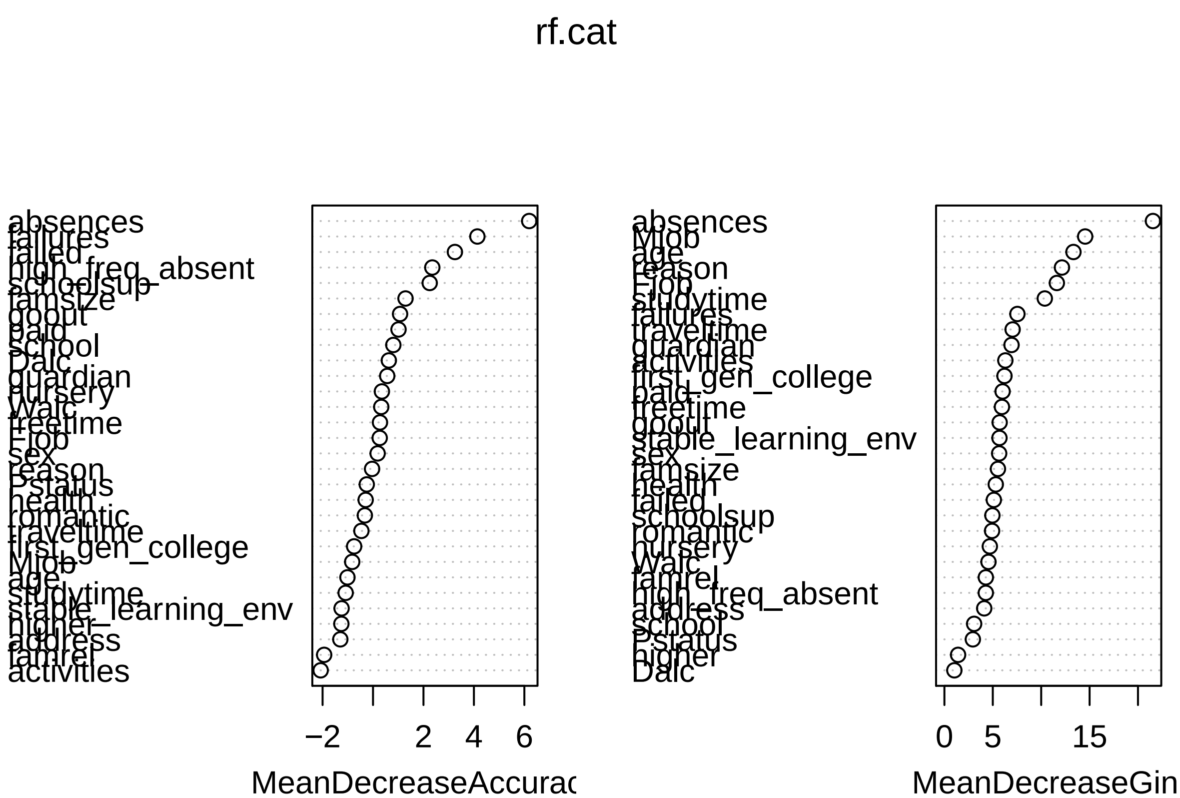
```
##
##         OOB estimate of  error rate: 57.91%
## Confusion matrix:
##            Poor Weak Sufficient Good Very Good Excellent class.error
## Poor         15    7          8    2         1         0   0.5454545
## Weak          5   18         48    2         0         0   0.7534247
## Sufficient    3   29         93    3         2         0   0.2846154
## Good          4    5         32    7         0         0   0.8541667
## Very Good     0    0         11    5         0         1   1.0000000
## Excellent     1    0         12    1         1         0   1.0000000
```

```r
importance(rf.cat)
```

```
##                             Poor        Weak  Sufficient        Good
## school               1.418224026  0.40894251 -0.57960533  2.54388488
## sex                 -0.653501187  0.60370568  0.24977571  0.48066823
## age                  0.448344220 -1.17420634 -0.24533484 -0.59042922
## address             -0.693495819 -0.21330911 -1.50717981  0.96654289
## famsize              1.764145807  0.98725385  0.28687545  0.38019026
## Pstatus              0.437039727 -0.53762447  0.07045264 -0.87903414
## Mjob                -1.109135394 -0.44161141 -1.50681345  0.95931824
## Fjob                 1.297625783  0.40257487  0.05823661  0.09064474
## reason              -0.097894846 -1.71667103  0.18204719  1.70951479
## guardian             0.147158285  0.15798101  1.23071197 -0.92995689
## traveltime          -0.425294519 -0.28006441  0.14534919 -1.43228661
## studytime            0.057173173  0.38568536 -0.50241649 -2.03731931
## failures             2.556769774  3.08729766  0.45793818  2.53571001
## schoolsup            1.753438366  1.25517987  0.78379658  1.70510374
## paid                 1.118627552  0.26189803  0.51657208 -0.63798601
## activities          -1.370551184 -1.20699948 -1.13224555 -0.95825557
## nursery             -1.298410047  1.02781486  0.46410186 -1.32884393
## higher              -2.316940279 -0.04532634 -0.68585552  1.78471688
## romantic            -0.680407593  0.30727455 -0.33609909 -0.40894821
## famrel              -1.078387545 -0.50289296 -1.24361414 -0.90902151
## freetime            -0.001088676  0.02740684  0.56238463 -0.82769673
## goout               -2.022831566  1.07316905  0.87287338  0.79658049
## Dalc                 0.000000000  0.41953213  0.20864014  1.01015254
## Walc                -0.858439370 -0.04346963 -0.01723483  2.44025218
## health               0.793248289  0.58067311 -0.45732022 -0.18750654
## absences             7.448999273  1.58862031  2.78077467  2.04587569
## first_gen_college    0.416878406  0.25922961 -1.89180829 -0.13036804
## stable_learning_env -0.765481584  0.04294296 -0.09970439 -1.10686914
## high_freq_absent     2.915807853  0.34734604  0.95599262  3.03573666
## failed               2.885781564  1.52474214  0.81895718  2.58167937
##                       Very Good  Excellent MeanDecreaseAccuracy
## school              -1.01015254  1.0101525           0.80498543
## sex                 -1.34359993 -0.1754656           0.18194613
## age                 -1.44845467  1.2963037          -1.01353341
## address             -1.01015254  1.0101525          -1.29671970
## famsize              0.76232872  1.0101525           1.28901079
## Pstatus              0.00000000  0.4678229          -0.24896753
## Mjob                 1.14907792  0.5579040          -0.82550818
## Fjob                -1.21657276 -0.4272368           0.26508220
## reason              -0.22067177  0.5833694          -0.03728804
## guardian            -0.66011578 -1.0101525           0.56306717
```

```
## traveltime           1.00191205 -1.0101525          -0.46143791
## studytime            0.51054941  0.3558777          -1.08499648
## failures             0.00000000  1.0101525           4.13700025
## schoolsup           -1.43177092  1.0101525           2.24716923
## paid                 0.97943770  1.0980312           1.01047467
## activities          -0.51987524  0.0000000          -2.07632814
## nursery              0.75798367  1.4229360           0.35161402
## higher               0.00000000  0.0000000          -1.25637567
## romantic            -0.31832142  1.0101525          -0.32461487
## famrel               1.01015254 -1.0101525          -1.93774260
## freetime             1.01015254 -0.6024591           0.27578889
## goout               -0.60245906  1.0101525           1.07199306
## Dalc                 0.00000000  0.0000000           0.62635885
## Walc                 0.49601049  1.0101525           0.32577261
## health              -1.56071254  0.1562119          -0.29254478
## absences             0.57155643  0.3742818           6.19172692
## first_gen_college    0.06283591  0.9200461          -0.74819158
## stable_learning_env -1.01015254 -1.4400461          -1.24639512
## high_freq_absent     1.76776695 -0.8904292           2.35017255
## failed               1.01015254  1.4229360           3.24673911
##                     MeanDecreaseGini
## school                      3.072587
## sex                         5.656577
## age                        13.331134
## address                     4.112408
## famsize                     5.529408
## Pstatus                     2.937808
## Mjob                       14.542602
## Fjob                       11.614861
## reason                     12.143577
## guardian                    6.934204
## traveltime                  7.047091
## studytime                  10.372108
## failures                    7.541551
## schoolsup                   4.955187
## paid                        6.012111
## activities                  6.294656
## nursery                     4.692421
## higher                      1.414178
## romantic                    4.924411
## famrel                      4.280805
## freetime                    5.942191
## goout                       5.707473
## Dalc                        1.028423
## Walc                        4.552716
## health                      5.306674
## absences                   21.543826
## first_gen_college           6.205268
## stable_learning_env         5.686258
## high_freq_absent            4.278214
## failed                      5.106471
```

```
varImpPlot(rf.cat)
```




```r
rf.acc<- predict(rf.cat, ftrain, type = 'class')
t<-table(predictions=rf.acc, actual=ftrain$cat_g3)
t
```

```
##              actual
## predictions  Poor Weak Sufficient Good Very Good Excellent
##   Poor         33    0          0    0         0         0
##   Weak          0   71          0    0         0         0
##   Sufficient    0    2        130    0         1         0
##   Good          0    0          0   48         0         0
##   Very Good     0    0          0    0        16         0
##   Excellent     0    0          0    0         0        15
```

```r
sum(diag(t))/sum(t)
```

```
## [1] 0.9905063
```

Very fitted model with accuracy for training data >99%.

Let's see what the accuracy rate for the testing set is:

```r
rf.pred<- predict(rf.cat, ftest, type = 'class')
t<-table(predictions=rf.pred, actual=ftest$cat_g3)
t
```

```
##              actual
## predictions  Poor Weak Sufficient Good Very Good Excellent
##   Poor          2    1          2    0         0         0
##   Weak          1    0          4    0         1         1
##   Sufficient    2   17         24    9         2         1
##   Good          0    1          5    3         1         1
##   Very Good     0    0          0    0         1         0
```

```
##   Excellent     0    0          0    0         0         0
```

```r
sum(diag(t))/sum(t)
```

```
## [1] 0.3797468
```

43.03% accuracy, which is an improvement.

Let's choose the most important variables, as well as interaction effects we believe to be important based on previous exploration:

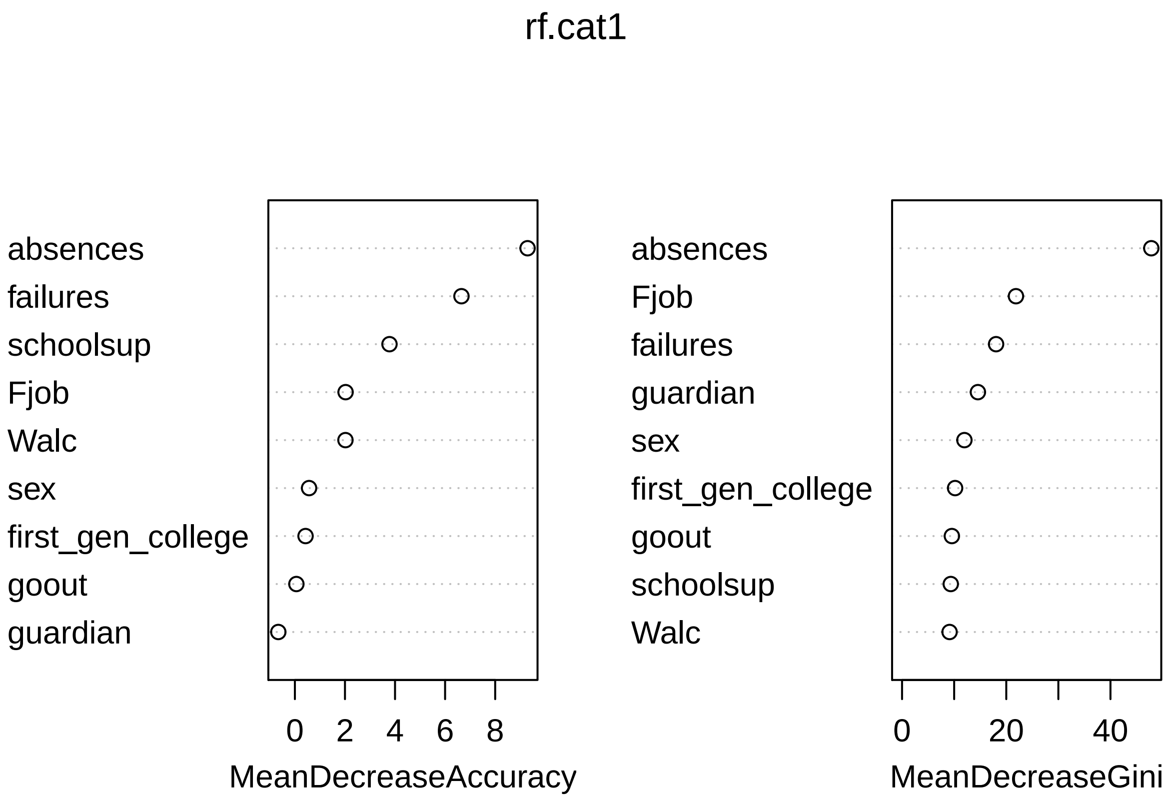
```r
rf.cat1<-randomForest(cat_g3~failures + absences + sex + Walc + Fjob +goout + schoolsup + first_gen_col
print(rf.cat1)
```

```
##
## Call:
##  randomForest(formula = cat_g3 ~ failures + absences + sex + Walc +      Fjob + goout + schoolsup + 
##                Type of random forest: classification
##                      Number of trees: 50
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 52.22%
## Confusion matrix:
##            Poor Weak Sufficient Good Very Good Excellent class.error
## Poor         23    3          4    2         0         1   0.3030303
## Weak          5   27         38    2         0         1   0.6301370
## Sufficient    5   23         87   11         2         2   0.3307692
## Good          2    5         23   13         1         4   0.7291667
## Very Good     1    1         10    3         1         1   0.9411765
## Excellent     0    1         10    3         1         0   1.0000000
```

```r
importance(rf.cat1)
```

```
##                         Poor       Weak Sufficient       Good  Very Good
## failures           4.2424935  4.1450188  1.0631210  3.9687210  2.9769975
## absences          14.7974760  1.7270164  3.9337529  1.4591986  1.8980916
## sex                3.3357840  1.5103454 -2.3359910  1.0936902  1.5714323
## Walc               2.4844928 -0.3991997 -0.6085898  3.7842420  3.8025511
## Fjob               0.3852836  2.2506175  1.2083110 -1.8090465  2.1483694
## goout             -1.5310686  1.6782038 -1.3062458  0.8902641  3.6881474
## schoolsup          4.0204822  4.3824578 -1.2463088  3.2099174  1.0610934
## first_gen_college  0.6975070  0.4266225 -2.2457970  4.6074291 -0.9355315
## guardian           1.2666280 -2.5306220 -0.3286237 -1.1966790  3.0171426
##                    Excellent MeanDecreaseAccuracy MeanDecreaseGini
## failures           1.0101525           6.65313915        18.045934
## absences          -1.4509532           9.29408139        47.834972
## sex               -0.9379438           0.56632438        11.968161
## Walc               1.0101525           2.01966213         9.122728
## Fjob               0.1039052           2.02501814        21.861115
## goout              0.8013456           0.05896317         9.559409
## schoolsup          0.0000000           3.77886103         9.351629
## first_gen_college  0.7912918           0.42524516        10.182273
## guardian          -0.8099566          -0.66267503        14.557878
```

```r
varImpPlot(rf.cat1)
```




```r
rf.acc<- predict(rf.cat1, ftrain, type = 'class')
t<-table(predictions=rf.acc, actual=ftrain$cat_g3)
t
```

```
##              actual
## predictions   Poor Weak Sufficient Good Very Good Excellent
##   Poor          30    4          2    1         0         0
##   Weak           0   55          2    0         0         0
##   Sufficient     2   12        123   13         6         6
##   Good           1    1          1   34         1         2
##   Very Good      0    0          1    0         9         0
##   Excellent      0    1          1    0         1         7
```

```r
sum(diag(t))/sum(t)
```

```
## [1] 0.8164557
```

54.75% OOB estimate of error rate and 83.5% accuracy rate for the training data.

```r
rf.pred1<- predict(rf.cat1, ftest, type = 'class')
t<-table(predictions=rf.pred1, actual=ftest$cat_g3)
t
```

```
##              actual
## predictions   Poor Weak Sufficient Good Very Good Excellent
##   Poor           2    1          2    0         0         0
##   Weak           1    1          6    1         0         0
##   Sufficient     1   15         23    7         3         1
##   Good           0    2          4    3         1         2
##   Very Good      0    0          0    1         0         0
##   Excellent      1    0          0    0         1         0
```

```
sum(diag(t))/sum(t)
```

```
## [1] 0.3670886
```

37.97% Accuracy, which is less than the full RF model.

The RF models indicate that for grade categorization, the most important variables are absences, failed, guardian, studytime, Mjob and Fjob, schoolsup, age, goout, first_gen_college (not in that order).

Modeling for low-high grades

Considering final grades as a continuous variable and ordinal categorical variable gave poor results. Therefore, we'd like to model a binary variable that indicates whether the student has a high grade (grade >= 10) or low grade (<10).

```
set.seed(3)
train_ind1 <- sample(x = nrow(data), size = 0.8 * nrow(data))
test_ind_neg1 <- -train_ind1
ftrain1 <- data[train_ind1, ]
ftest1 <- data[test_ind_neg1, ]
```

Fitting a decision tree on pass-fail

```
data[["pf"]] <- as.factor(data[["pf"]])
training[["pf"]] <- as.factor(training[["pf"]])
testing[["pf"]] <- as.factor(testing[["pf"]])
treepf <- tree(pf ~ . -G1 -G2 -G3 -ord_g3 -failures -reason -health -age -nursery -ord_g3, data=training
```

```
## Warning in tree(pf ~ . - G1 - G2 - G3 - ord_g3 - failures - reason - health - :
## NAs introduced by coercion
```

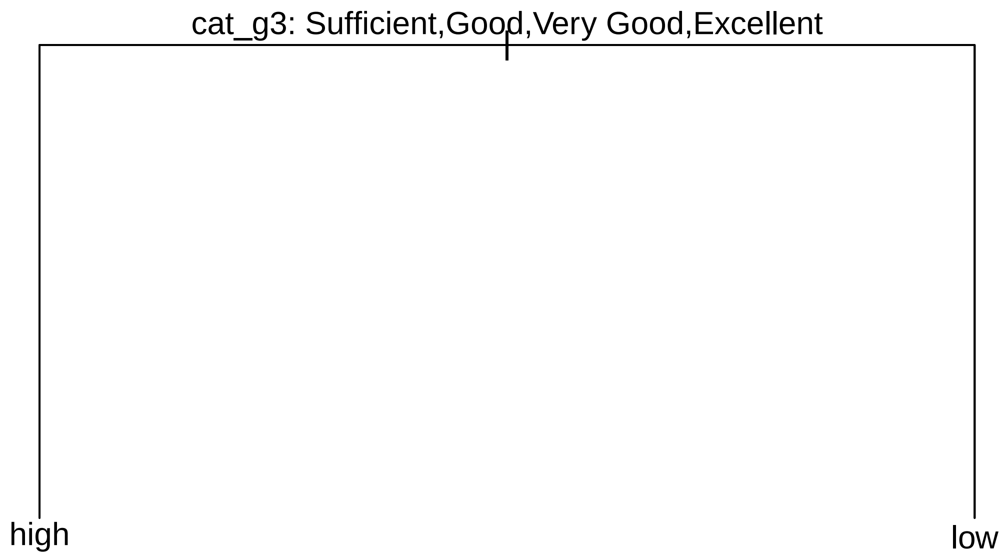
```
treepf
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
## 1) root 316 403.2 high ( 0.6646 0.3354 )
##   2) cat_g3: Sufficient,Good,Very Good,Excellent 210   0.0 high ( 1.0000 0.0000 ) *
##   3) cat_g3: Poor,Weak 106   0.0 low ( 0.0000 1.0000 ) *
```

```
summary(treepf)
```

```
##
## Classification tree:
## tree(formula = pf ~ . - G1 - G2 - G3 - ord_g3 - failures - reason -
##     health - age - nursery - ord_g3, data = training)
## Variables actually used in tree construction:
## [1] "cat_g3"
## Number of terminal nodes:  2
## Residual mean deviance:  0 = 0 / 314
## Misclassification error rate: 0 = 0 / 316
```

```
plot(treepf)
text(treepf, pretty = 0)
```



Initial Tree Diagnostic

```
tree.pred <- predict(treepf, testing, type = "class")
```

```
## Warning in pred1.tree(object, tree.matrix(newdata)): NAs introduced by coercion
```

```
table(tree.pred, testing$pf)
```

```
##
## tree.pred high low
##      high   55   0
##      low     0  24
```

```
sum(diag(table(tree.pred, testing$pf)))/79
```

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

Misclassification rate: 0.38. This can likely be decreased with other methods- using all variables likely overfits.

###Pruning

```
set.seed(3)
cv.pf <- cv.tree(treepf, FUN = prune.misclass)
```

```
## Warning in tree(model = m[rand != i, , drop = FALSE]): NAs introduced by
## coercion

## Warning in pred1.tree(tree, tree.matrix(nd)): NAs introduced by coercion

## Warning in tree(model = m[rand != i, , drop = FALSE]): NAs introduced by
## coercion

## Warning in pred1.tree(tree, tree.matrix(nd)): NAs introduced by coercion

## Warning in tree(model = m[rand != i, , drop = FALSE]): NAs introduced by
## coercion

## Warning in pred1.tree(tree, tree.matrix(nd)): NAs introduced by coercion

## Warning in tree(model = m[rand != i, , drop = FALSE]): NAs introduced by
## coercion
```

```
## Warning in pred1.tree(tree, tree.matrix(nd)): NAs introduced by coercion

## Warning in tree(model = m[rand != i, , drop = FALSE]): NAs introduced by
## coercion

## Warning in pred1.tree(tree, tree.matrix(nd)): NAs introduced by coercion

## Warning in tree(model = m[rand != i, , drop = FALSE]): NAs introduced by
## coercion

## Warning in pred1.tree(tree, tree.matrix(nd)): NAs introduced by coercion

## Warning in tree(model = m[rand != i, , drop = FALSE]): NAs introduced by
## coercion

## Warning in pred1.tree(tree, tree.matrix(nd)): NAs introduced by coercion

## Warning in tree(model = m[rand != i, , drop = FALSE]): NAs introduced by
## coercion

## Warning in pred1.tree(tree, tree.matrix(nd)): NAs introduced by coercion

## Warning in tree(model = m[rand != i, , drop = FALSE]): NAs introduced by
## coercion

## Warning in pred1.tree(tree, tree.matrix(nd)): NAs introduced by coercion

## Warning in tree(model = m[rand != i, , drop = FALSE]): NAs introduced by
## coercion

## Warning in pred1.tree(tree, tree.matrix(nd)): NAs introduced by coercion
```

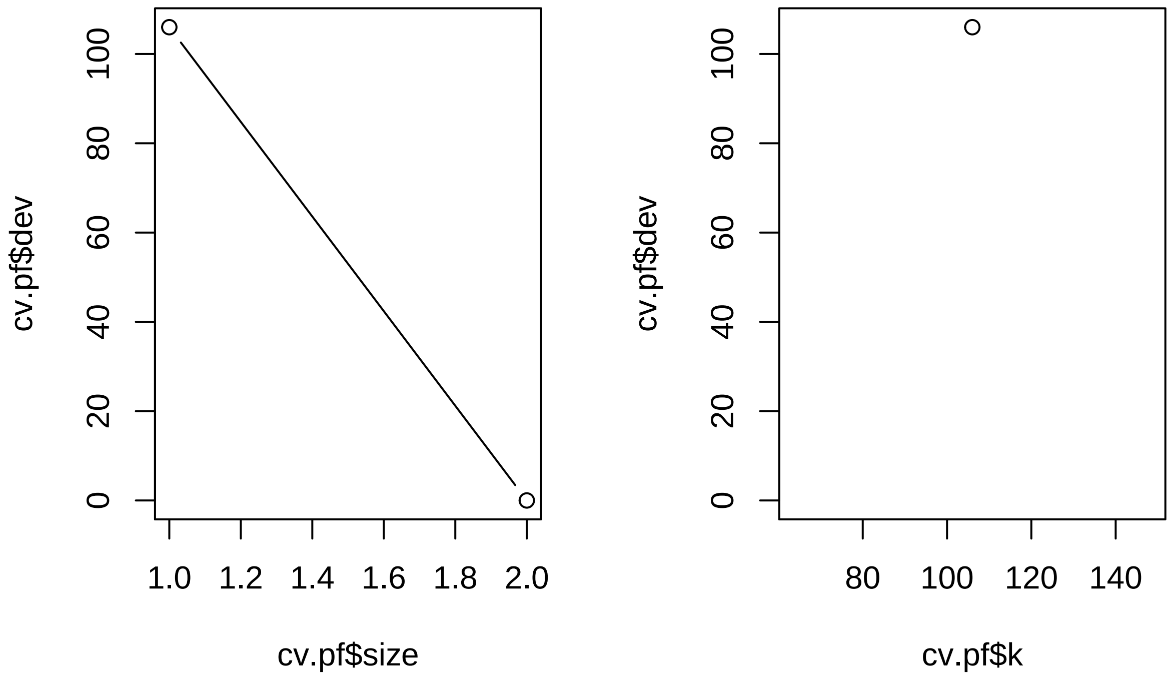
```
names(cv.pf)
```

```
## [1] "size"   "dev"    "k"      "method"
```

```
cv.pf
```

```
## $size
## [1] 2 1
##
## $dev
## [1]   0 106
##
## $k
## [1] -Inf  106
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

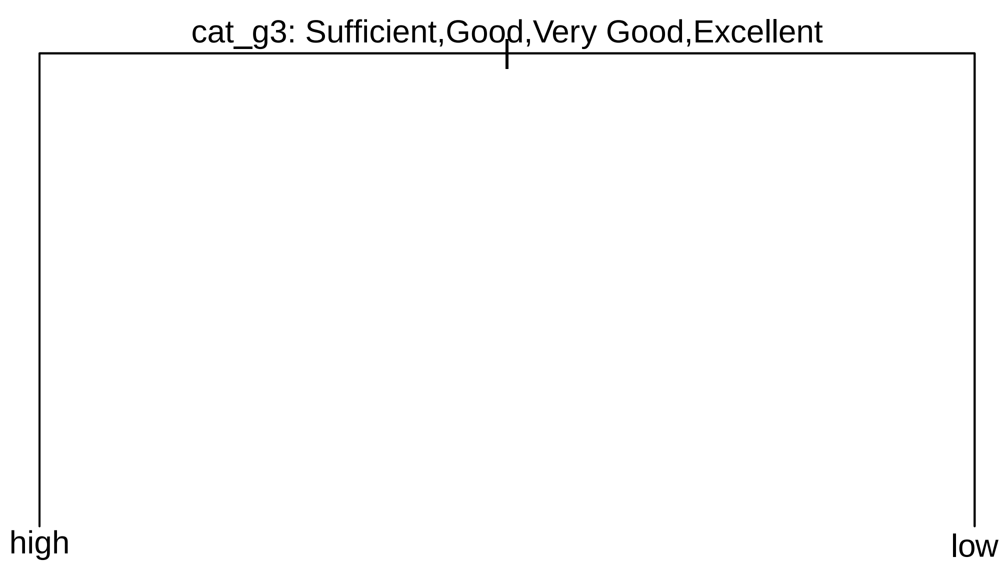
```
par(mfrow = c(1,2))
plot(cv.pf$size, cv.pf$dev, type = "b")
plot(cv.pf$k, cv.pf$dev, type = "b")
```



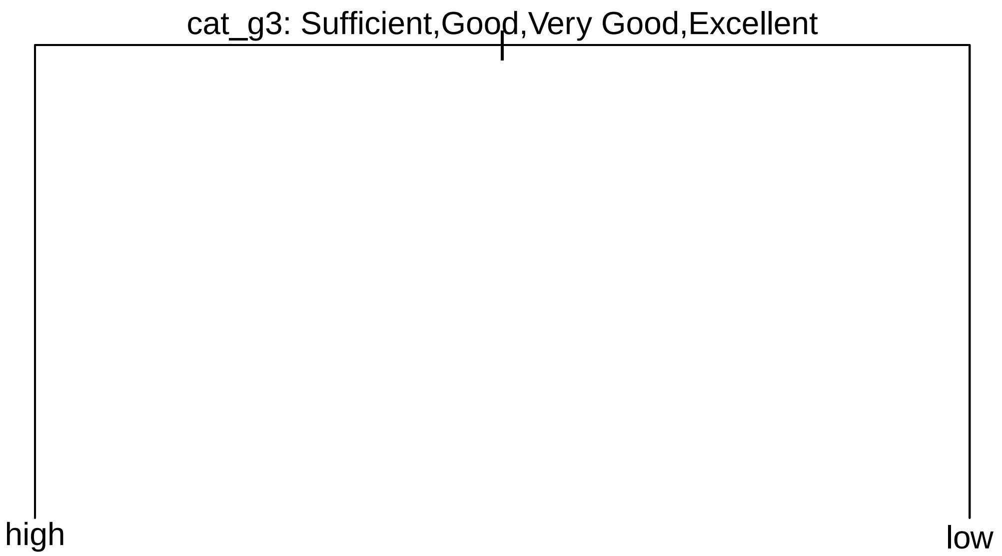

```
prune.pf <- prune.misclass(treepf, best = 3)
```

```
## Warning in prune.tree(tree = treepf, best = 3, method = "misclass"): best is
## bigger than tree size
```

```
plot(prune.pf)
text(prune.pf, pretty = 0)
```




```
prune.short <- prune.misclass(treepf, best = 2)
plot(prune.short)
text(prune.short, pretty = 0)
```




```
treepred2 <- predict(prune.pf, testing, type = "class")
```

```
## Warning in pred1.tree(object, tree.matrix(newdata)): NAs introduced by coercion
```

```
table(treepred2, testing$pf)
```

```
##
## treepred2 high low
##      high   55   0
##      low     0  24
```

```
sum(diag(table(treepred2, testing$pf)))/79
```

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

```
treepred3 <- predict(prune.short, testing, type = "class")
```

```
## Warning in pred1.tree(object, tree.matrix(newdata)): NAs introduced by coercion
```

```
table(treepred3, testing$pf)
```

```
##
## treepred3 high low
##      high   55   0
##      low     0  24
```

```
sum(diag(table(treepred3, testing$pf)))/79
```

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

Misclassification rateL .32.

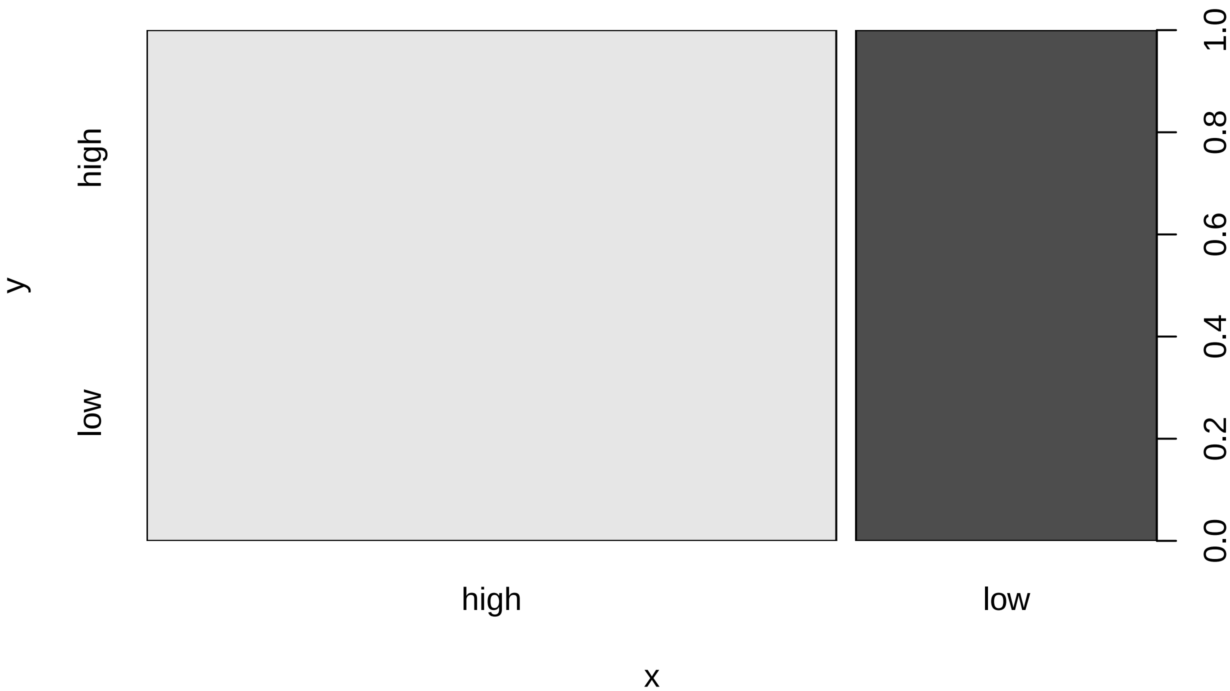
Bagging

```
library(randomForest)
set.seed(1)
bag.pf <- randomForest(pf ~ . -G1 -G2 -G3 -ord_g3 -failures -reason -health -age -nursery -ord_g3, data=
bag.pf
```

```
##
## Call:
```

```
##  randomForest(formula = pf ~ . - G1 - G2 - G3 - ord_g3 - failures -      reason - health - age - nurs
##                Type of random forest: classification
##                      Number of trees: 75
## No. of variables tried at each split: 28
##
##         OOB estimate of  error rate: 0%
## Confusion matrix:
##      high low class.error
## high  210   0           0
## low     0 106           0
```

```r
yhat.bag <- predict(bag.pf, testing)
plot(yhat.bag, testing$pf)
```




```r
table(yhat.bag, testing$pf)
```

```
##
## yhat.bag high low
##     high   55   0
##     low     0  24
```

```r
sum(diag(table(yhat.bag, testing$pf)))/79
```

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

Boosting

```r
library(gbm)
```

```
## Loaded gbm 2.1.8
```

```r
attach(data)
```

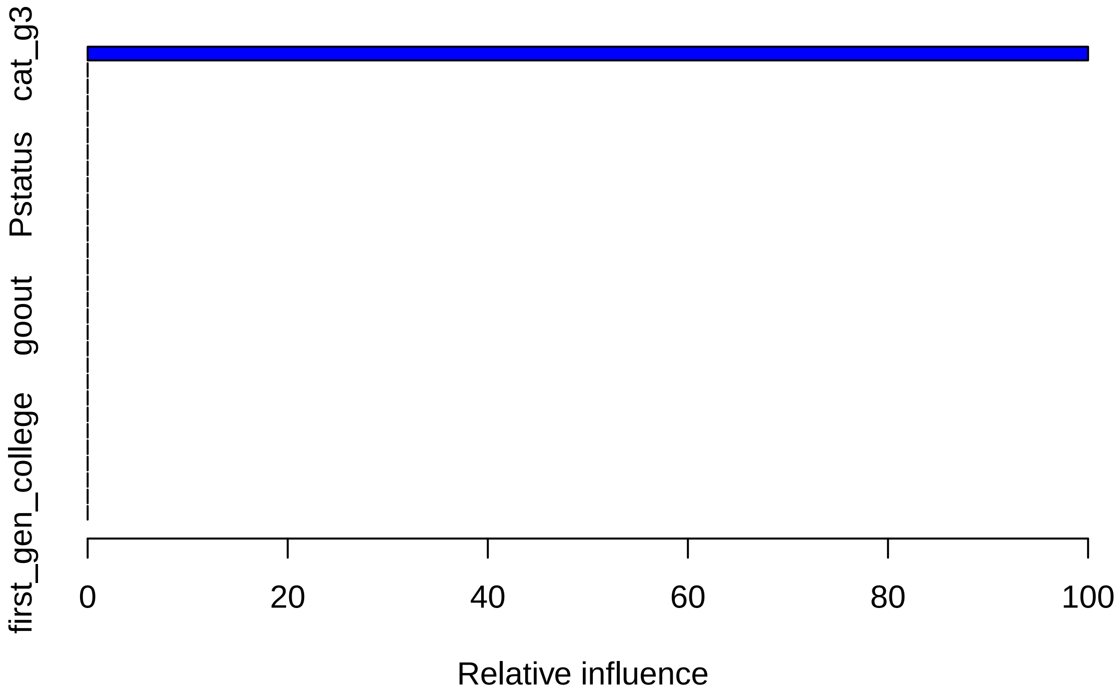
```
## The following objects are masked from data (pos = 4):
##
```

```
##     Dalc, Fedu, Fjob, G1, G2, G3, Medu, Mjob, Pstatus, Walc, absences,
##     activities, address, age, cat_g3, failed, failures, famrel,
##     famsize, famsup, first_gen_college, freetime, goout, guardian,
##     health, high_freq_absent, higher, internet, nursery, ord_g3, paid,
##     pf, reason, romantic, school, schoolsup, sex, stable_learning_env,
##     studytime, traveltime
```

```r
data[["pf_factor"]] <- as.factor(data[["pf"]])
data[["pf_bin"]] <- as.numeric(data[["pf_factor"]])-1
training[["pf_factor"]] <- as.factor(training[["pf"]])
training[["pf_bin"]] <- as.numeric(training[["pf_factor"]])-1
testing[["pf_factor"]] <- as.factor(testing[["pf"]])
testing[["pf_bin"]] <- as.numeric(testing[["pf_factor"]])-1

set.seed(1)
boost.pf <- gbm(pf_bin ~ . -pf_factor -pf -school -G1 -G2 -G3 -ord_g3 -failures -reason -health -age -nu
                    distribution = "bernoulli", n.trees = 500,
                    interaction.depth = 2)

summary(boost.pf)
```




```
##                                 var      rel.inf
## cat_g3                       cat_g3 1.000000e+02
## absences                   absences 1.527612e-27
## traveltime               traveltime 7.700951e-29
## Mjob                           Mjob 5.879732e-29
## studytime                 studytime 8.341861e-30
## Medu                           Medu 4.589843e-30
## schoolsup                 schoolsup 3.356194e-30
## Fjob                           Fjob 3.083552e-30
## Pstatus                     Pstatus 1.407552e-31
## famsize                     famsize 1.559680e-32
```

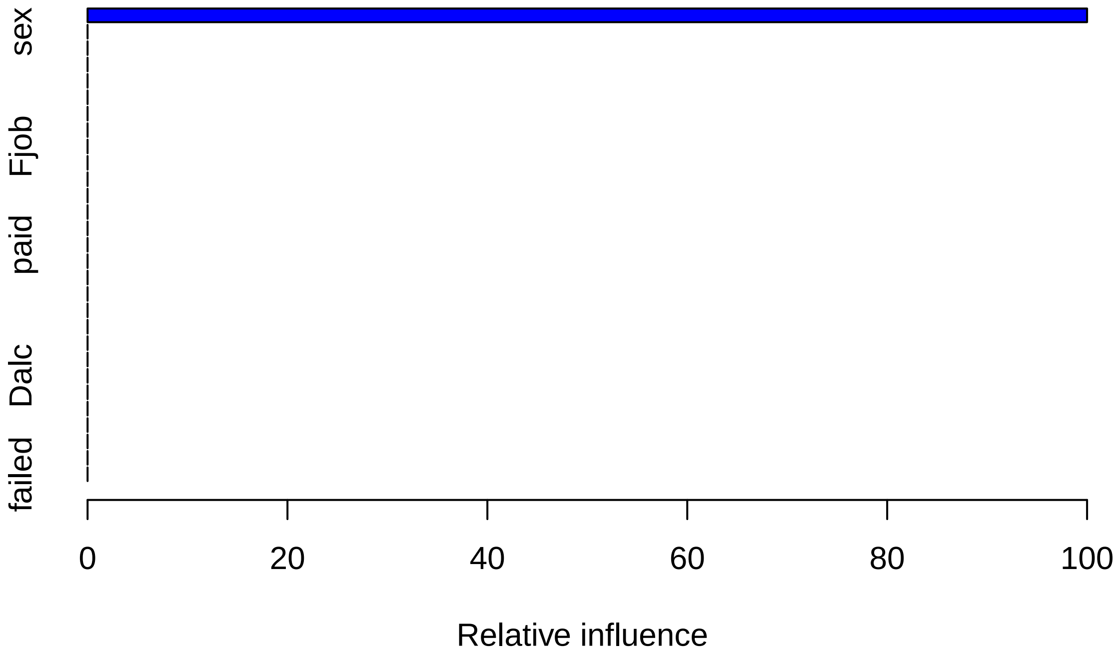
```
## address                             address 8.849969e-34
## internet                           internet 1.157148e-34
## high_freq_absent       high_freq_absent 2.399331e-38
## Walc                                   Walc 2.156494e-42
## failed                               failed 4.545526e-46
## Fedu                                   Fedu 1.215371e-49
## goout                                 goout 4.070774e-61
## guardian                           guardian 2.196326e-62
## paid                                   paid 4.185927e-66
## famrel                               famrel 4.076438e-66
## romantic                           romantic 5.379314e-68
## stable_learning_env stable_learning_env 8.809047e-72
## activities                       activities 2.816272e-73
## sex                                     sex 0.000000e+00
## famsup                               famsup 0.000000e+00
## higher                               higher 0.000000e+00
## freetime                           freetime 0.000000e+00
## Dalc                                   Dalc 0.000000e+00
## first_gen_college     first_gen_college 0.000000e+00
```

```
predboost1 <- predict(boost.pf, testing,
                      n.trees = 500)
table(predboost1, testing$pf_bin)
```

```
##
## predboost1           0  1
##   -51.1132477456973 55  0
##   36.7658750128531   0 24
```

Lower interaction depth

```
boost.pf1 <- gbm(pf_bin ~ . -pf_factor -pf -school -G1 -G2 -G3 -ord_g3 -failures -reason -health -age -r
                    distribution = "bernoulli", n.trees = 500,
                    interaction.depth = 1)
summary(boost.pf1)
```




```
##                                     var rel.inf
## cat_g3                           cat_g3     100
## sex                                 sex       0
## address                         address       0
## famsize                         famsize       0
## Pstatus                         Pstatus       0
## Medu                               Medu       0
## Fedu                               Fedu       0
## Mjob                               Mjob       0
## Fjob                               Fjob       0
## guardian                       guardian       0
## traveltime                   traveltime       0
## studytime                     studytime       0
## schoolsup                     schoolsup       0
## famsup                           famsup       0
## paid                               paid       0
## activities                   activities       0
## higher                           higher       0
## internet                       internet       0
## romantic                       romantic       0
## famrel                           famrel       0
## freetime                       freetime       0
## goout                             goout       0
## Dalc                               Dalc       0
## Walc                               Walc       0
## absences                       absences       0
## first_gen_college     first_gen_college       0
## stable_learning_env stable_learning_env       0
## high_freq_absent       high_freq_absent       0
## failed                           failed       0
```

```
predboost1 <- predict(boost.pf, testing,
                      n.trees = 500)
```

```r
table(predboost1, testing$pf_bin)
```

```
##
## predboost1          0  1
##   -51.1132477456973 55  0
##   36.7658750128531   0 24
```

Not much difference.

Fitting random forest on low-high binary

Fitting with ALL predictors:

```r
rf.bin<-randomForest(pf~. -G1 -G2 -G3 -ord_g3 - cat_g3 -Medu -Fedu,data = ftrain1,mtry=3, ntree=50, impo
print(rf.bin)
```

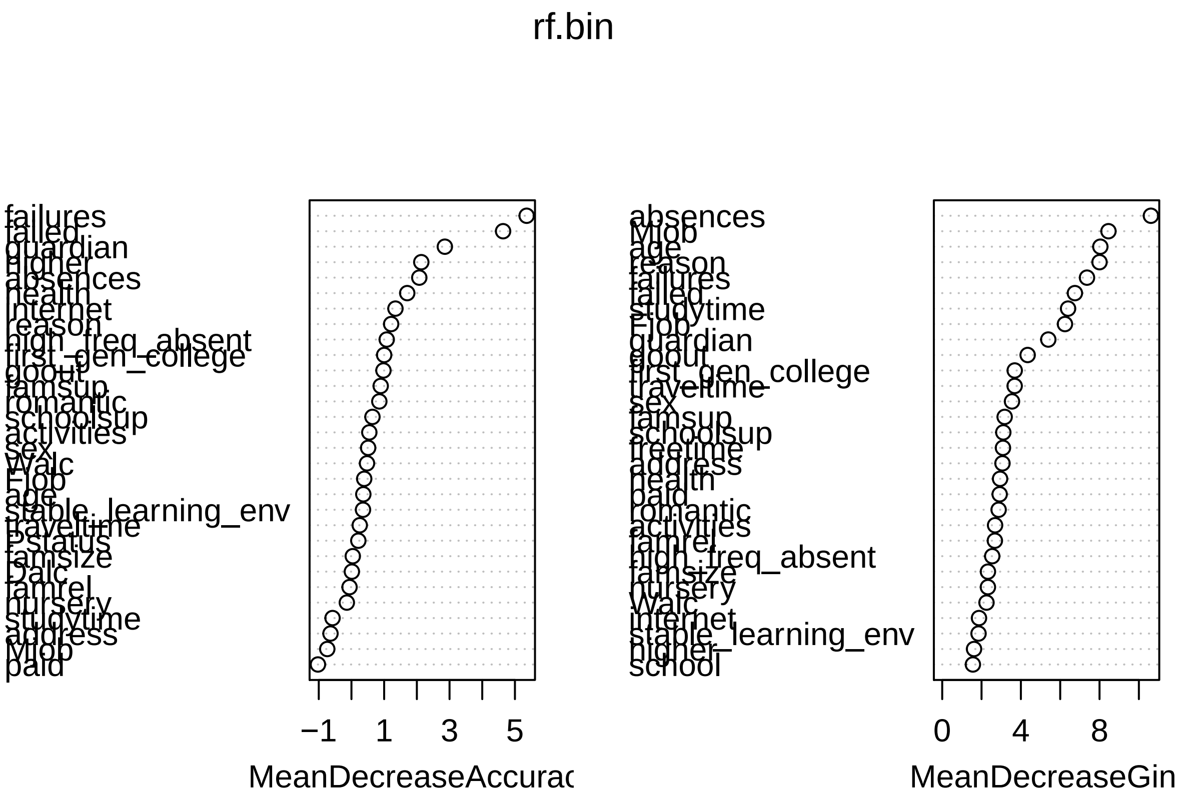
```
##
## Call:
##  randomForest(formula = pf ~ . - G1 - G2 - G3 - ord_g3 - cat_g3 -      Medu - Fedu, data = ftrain1, m
##                Type of random forest: classification
##                      Number of trees: 50
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 27.85%
## Confusion matrix:
##      high low class.error
## high  190  20   0.0952381
## low    68  38   0.6415094
```

```r
importance(rf.bin)
```

```
##                          high         low MeanDecreaseAccuracy
## school             -2.485581508 -1.52673306          -2.63710303
## sex                -0.965165302  2.93414745           0.51049234
## age                 0.139033230  0.40753652           0.36305088
## address            -1.265315512  0.56919163          -0.64142255
## famsize             0.313937542 -0.15509877           0.04017503
## Pstatus             0.300092074 -0.04079301           0.21024630
## Mjob               -0.665033163 -0.73081250          -0.73948831
## Fjob               -0.379109665  1.06891868           0.39015400
## reason              0.714816698  1.24057039           1.21413248
## guardian            2.173626595  1.84588051           2.85619441
## traveltime          0.646069637 -0.43241156           0.25330303
## studytime          -0.342003707 -0.08909346          -0.57883698
## failures            4.346277380  3.55520510           5.35786996
## schoolsup          -0.411547231  1.69806708           0.64211123
## famsup              1.635973757 -0.52964534           0.89440050
## paid               -1.202828598  0.12282011          -1.02521760
## activities         -0.233423760  1.23855081           0.54681514
## nursery             0.380920261 -0.51327766          -0.14170752
## higher              1.297165181  1.66139689           2.13552428
## internet            1.517377344  0.09930989           1.34488620
## romantic            0.326795259  1.39052208           0.85260390
## famrel              0.258084483 -0.34734701          -0.05859069
## freetime           -1.697979099 -1.19984074          -1.89631703
## goout              -0.366115745  2.29795989           0.97810480
```

```
## Dalc                -0.227459308  0.20745016           0.01294779
## Walc                -1.455065051  1.97279834           0.47546418
## health               1.035644277  1.46258333           1.70557566
## absences             2.505092087  0.74901892           2.07446675
## first_gen_college    0.002474154  1.88146528           1.00016786
## stable_learning_env -0.137473745  1.10737207           0.35268623
## high_freq_absent     0.028780867  1.83391235           1.07867783
## failed               4.759518594  3.25610088           4.63730082
##                     MeanDecreaseGini
## school                     1.5600159
## sex                        3.5491078
## age                        8.0402404
## address                    3.0645379
## famsize                    2.3253121
## Pstatus                    1.4532774
## Mjob                       8.4544034
## Fjob                       6.2418644
## reason                     7.9999579
## guardian                   5.3937605
## traveltime                 3.6843505
## studytime                  6.3999760
## failures                   7.3620293
## schoolsup                  3.1061548
## famsup                     3.1764733
## paid                       2.9223962
## activities                 2.6885872
## nursery                    2.3242116
## higher                     1.6192776
## internet                   1.8714043
## romantic                   2.8689542
## famrel                     2.6783574
## freetime                   3.0896971
## goout                      4.3452508
## Dalc                       0.8336794
## Walc                       2.2521656
## health                     2.9419715
## absences                  10.6122048
## first_gen_college          3.6871370
## stable_learning_env        1.8476062
## high_freq_absent           2.5439946
## failed                     6.7471392
```

```r
varImpPlot(rf.bin)
```




```r
rf.acc<- predict(rf.bin, ftrain1, type = 'class')
t<-table(predictions=rf.acc, actual=ftrain1$pf)
t
```

```
##            actual
## predictions high low
##        high  210   2
##        low     0 104
```

```r
sum(diag(t))/sum(t)
```

```
## [1] 0.9936709
```

Predictions on testing set:

```r
rf.pred2<- predict(rf.bin, ftest1, type = 'class')
t<-table(predictions=rf.pred2, actual=ftest1$pf)
t
```

```
##            actual
## predictions high low
##        high   48  20
##        low     7   4
```

```r
sum(diag(t))/sum(t)
```

```
## [1] 0.6582278
```

72.15% accuracy rate.

Finding the best random forest model by including important predictors:

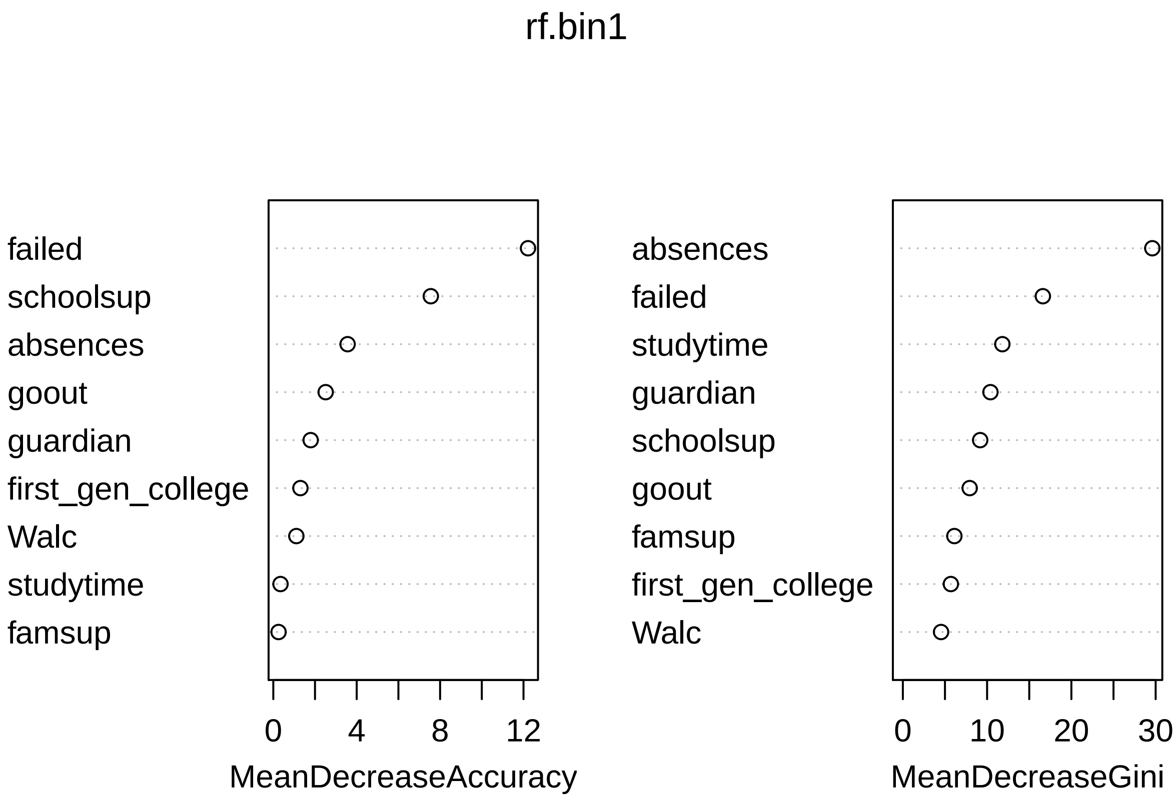
```r
rf.bin1<-randomForest(pf~failed + absences+ guardian + studytime + goout + schoolsup + first_gen_college
print(rf.bin1)
```

```
##
## Call:
##  randomForest(formula = pf ~ failed + absences + guardian + studytime +      goout + schoolsup + firs
##                Type of random forest: classification
##                      Number of trees: 50
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 27.85%
## Confusion matrix:
##      high low class.error
## high  182  28   0.1333333
## low    60  46   0.5660377
```

```r
importance(rf.bin1)
```

```
##                         high        low MeanDecreaseAccuracy MeanDecreaseGini
## failed            8.76476089 10.7432931           12.2176093        16.615519
## absences          3.10547085  2.0738397            3.5634858        29.604750
## guardian          2.58574861 -0.5726153            1.7926224        10.396056
## studytime        -0.04201686  0.5169681            0.3448446        11.809784
## goout             1.57374996  2.8259365            2.5109677         7.934012
## schoolsup         5.17404427  5.9413297            7.5551707         9.177453
## first_gen_college 0.20283606  1.8521200            1.2991851         5.697752
## Walc              0.86179592  0.5995871            1.1049734         4.536099
## famsup            1.32940595 -0.9431626            0.2515182         6.114414
```

```r
varImpPlot(rf.bin1)
```




```
rf.acc1<- predict(rf.bin1, ftrain1, type = 'class')
t<-table(predictions=rf.acc1, actual=ftrain1$pf)
t
```

```
##             actual
## predictions high low
##        high  208  25
##        low     2  81
```

```
sum(diag(t))/sum(t)
```

```
## [1] 0.914557
```

The pared-down model has an 00B estimate of error rate of 25.95% and a training set prediction accuracy rate of 90.19%.

Predictions on testing set:

```
rf.pred3<- predict(rf.bin1, ftest1, type = 'class')
t<-table(predictions=rf.pred3, actual=ftest1$pf)
t
```

```
##             actual
## predictions high low
##        high   48  18
##        low     7   6
```

```
sum(diag(t))/sum(t)
```

```
## [1] 0.6835443
```

72.15% prediction accuracy rate.

Overall the random-forests for pass-fail indicate that the most important factors affecting whether the student passes/fails are failed, absences, guardian, studytime, goout, schoolsup, first_gen_college.