Marks_Prediction.R

sudhanshmehta

2020-04-27

```
# loading all the libraries

library(ISLR)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
library(lattice)
library(glmnet)

## Loading required package: Matrix

## Loaded glmnet 3.0-2

library(tree)
library(MASS)

##
## Attaching package: 'MASS'

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

library(ROSE)

## Loaded ROSE 0.0-3

library(rpart)

############################
## Part 2 #################
############################
```

```
#R code to import and prepare the student performance dataset
school1=read.table("student-mat.csv",sep=";",header=TRUE)
school2=read.table("student-por.csv",sep=";",header=TRUE)

############################
## Portugese Performance Analysis
############################

# Quick glance at Data

table(school2$school)

##
##  GP  MS
## 423 226

head(school2)

##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob
reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher
course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other
course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other
other
## 4     GP   F  15       U     GT3       T    4    2   health services
home
## 5     GP   F  16       U     GT3       T    3    3    other    other
home
## 6     GP   M  16       U     LE3       T    4    3 services    other
reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        0       yes     no   no         no
## 4   mother          1         3        0        no    yes   no        yes
## 5   father          1         2        0        no    yes   no         no
## 6   mother          1         2        0        no    yes   no        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3
## 1        4  0 11 11
## 2        2  9 11 11
## 3        6 12 13 12
## 4        0 14 14 14
## 5        0 11 13 13
## 6        6 12 12 13
```

```
colnames(school2)
```

```
##  [1] "school"     "sex"        "age"        "address"    "famsize"   
##  [6] "Pstatus"    "Medu"       "Fedu"       "Mjob"       "Fjob"      
## [11] "reason"     "guardian"   "traveltime" "studytime"  "failures"  
## [16] "schoolsup"  "famsup"     "paid"       "activities" "nursery"   
## [21] "higher"     "internet"   "romantic"   "famrel"     "freetime"  
## [26] "goout"      "Dalc"       "Walc"       "health"     "absences"  
## [31] "G1"         "G2"         "G3"
```

```
summary(school2)
```

```
##  school   sex          age        address famsize   Pstatus      Medu      
##  GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80   Min.   :0.000  
##  MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569   1st Qu.:2.000  
##                   Median :17.00                             Median :2.000  
##                   Mean   :16.74                             Mean   :2.515  
##                   3rd Qu.:18.00                             3rd Qu.:4.000  
##                   Max.   :22.00                             Max.   :4.000  
##       Fedu            Mjob           Fjob            reason      guardian  
##  Min.   :0.000   at_home :135   at_home : 42   course    :285   father:153  
##  1st Qu.:1.000   health  : 48   health  : 23   home      :149   mother:455  
##  Median :2.000   other   :258   other   :367   other     : 72   other : 41  
##  Mean   :2.307   services:136   services:181   reputation:143               
##  3rd Qu.:3.000   teacher : 72   teacher : 36                                
##  Max.   :4.000                                                              
##    traveltime      studytime        failures      schoolsup famsup    
paid    
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :581   no :251   no 
:610  
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 68   yes:398   yes: 
39  
##  Median :1.000   Median :2.000   Median :0.0000                           
##  Mean   :1.569   Mean   :1.931   Mean   :0.2219                           
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000                           
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000                           
##  activities nursery   higher    internet  romantic     famrel     
##  no :334    no :128   no : 69   no :151   no :410   Min.   :1.000  
##  yes:315    yes:521   yes:580   yes:498   yes:239   1st Qu.:4.000  
##                                                     Median :4.000  
##                                                     Mean   :3.931  
##                                                     3rd Qu.:5.000  
##                                                     Max.   :5.000  
##     freetime        goout            Dalc            Walc          health     
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   
:1.000  
```

```
##  1st Qu.:3.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st
Qu.:2.000
##  Median :3.00   Median :3.000   Median :1.000   Median :2.00   Median
:4.000
##  Mean   :3.18   Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean
:3.536
##  3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd
Qu.:5.000
##  Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.
:5.000
##     absences            G1             G2              G3
##  Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00   1st Qu.:10.00
##  Median : 2.000   Median :11.0   Median :11.00   Median :12.00
##  Mean   : 3.659   Mean   :11.4   Mean   :11.57   Mean   :11.91
##  3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :32.000   Max.   :19.0   Max.   :19.00   Max.   :19.00
```

```
############################
## Data Preparation ########
############################

any(is.na(school2))
```

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

```
# There are no missing values in the data set.

# dropping G1 and G2 as they are highly corelated to G3
portuguese_df = subset(school2, select = -c(G1,G2))

summary(portuguese_df)
```

```
##  school   sex          age        address famsize   Pstatus      Medu
##  GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80   Min.   :0.000
##  MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569   1st Qu.:2.000
##                   Median :17.00                             Median :2.000
##                   Mean   :16.74                             Mean   :2.515
##                   3rd Qu.:18.00                             3rd Qu.:4.000
##                   Max.   :22.00                             Max.   :4.000
##       Fedu             Mjob           Fjob            reason      guardian
##  Min.   :0.000   at_home :135   at_home : 42   course    :285   father:153
##  1st Qu.:1.000   health  : 48   health  : 23   home      :149   mother:455
##  Median :2.000   other   :258   other   :367   other     : 72   other : 41
##  Mean   :2.307   services:136   services:181   reputation:143
##  3rd Qu.:3.000   teacher : 72   teacher : 36
##  Max.   :4.000
##    traveltime      studytime        failures      schoolsup famsup
paid
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :581   no :251   no
```

```
:610
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 68   yes:398   yes:
39
##  Median :1.000   Median :2.000   Median :0.0000
##  Mean   :1.569   Mean   :1.931   Mean   :0.2219
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000
##  activities nursery   higher    internet  romantic      famrel
##  no :334    no :128   no : 69   no :151   no :410   Min.   :1.000
##  yes:315    yes:521   yes:580   yes:498   yes:239   1st Qu.:4.000
##                                                     Median :4.000
##                                                     Mean   :3.931
##                                                     3rd Qu.:5.000
##                                                     Max.   :5.000
##     freetime        goout            Dalc            Walc          health
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.
:1.000
##  1st Qu.:3.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st
Qu.:2.000
##  Median :3.00   Median :3.000   Median :1.000   Median :2.00   Median
:4.000
##  Mean   :3.18   Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean
:3.536
##  3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd
Qu.:5.000
##  Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.
:5.000
##     absences            G3
##  Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:10.00
##  Median : 2.000   Median :12.00
##  Mean   : 3.659   Mean   :11.91
##  3rd Qu.: 6.000   3rd Qu.:14.00
##  Max.   :32.000   Max.   :19.00
```

```
# The following variables need to be converted to categorical type:

# Fedu - Father's eductaion
portuguese_df$Fedu = factor(portuguese_df$Fedu,
levels=c("0","1","2","3","4"), ordered=TRUE)
summary(portuguese_df$Fedu)
```

```
##   0   1   2   3   4
##   7 174 209 131 128
```

```
# famrel - quality of family relationships
portuguese_df$famrel = factor(portuguese_df$famrel, levels=1:5, ordered=TRUE)
summary(portuguese_df$famrel)
```

```
##   1   2   3   4   5
##  22  29 101 317 180
```

```r
# traveltime - home to school travel time
portuguese_df$traveltime = factor(portuguese_df$traveltime, levels=0:4,
ordered=TRUE)
summary(portuguese_df$traveltime)
```

```
##   0   1   2   3   4
##   0 366 213  54  16
```

```r
# Medu - Mother's eductaion
portuguese_df$Medu = factor(portuguese_df$Medu,
levels=c("0","1","2","3","4"), ordered=TRUE)
summary(portuguese_df$Medu)
```

```
##   0   1   2   3   4
##   6 143 186 139 175
```

```r
# studytime - weekly study time
portuguese_df$studytime = factor(portuguese_df$studytime, levels=1:4,
ordered=TRUE)
summary(portuguese_df$studytime)
```

```
##   1   2   3   4
## 212 305  97  35
```

```r
# freetime - free time after school
portuguese_df$freetime = factor(portuguese_df$freetime, levels=1:5,
ordered=TRUE)
summary(portuguese_df$freetime)
```

```
##   1   2   3   4   5
##  45 107 251 178  68
```

```r
# goout - going out with friends
portuguese_df$goout = factor(portuguese_df$goout, levels=1:5, ordered=TRUE)
summary(portuguese_df$goout)
```

```
##   1   2   3   4   5
##  48 145 205 141 110
```

```r
# Dalc - workday alcohol consumption
portuguese_df$Dalc = factor(portuguese_df$Dalc, levels=1:5, ordered=TRUE)
summary(portuguese_df$Dalc)
```

```
##   1   2   3   4   5
## 451 121  43  17  17
```

```r
# Walc - weekend alcohol consumption
portuguese_df$Walc = factor(portuguese_df$Walc, levels=1:5, ordered=TRUE)
summary(portuguese_df$Walc)
```

```
##   1   2   3   4   5
## 247 150 120  87  45
```

```
# health - current health status
portuguese_df$health = factor(portuguese_df$health, levels=1:5, ordered=TRUE)
summary(portuguese_df$health)
```

```
##   1   2   3   4   5
##  90  78 124 108 249
```

```
# failures - number of past class failures
portuguese_df$failures = factor(portuguese_df$failures, levels=0:4,
ordered=TRUE)
summary(portuguese_df$failures)
```

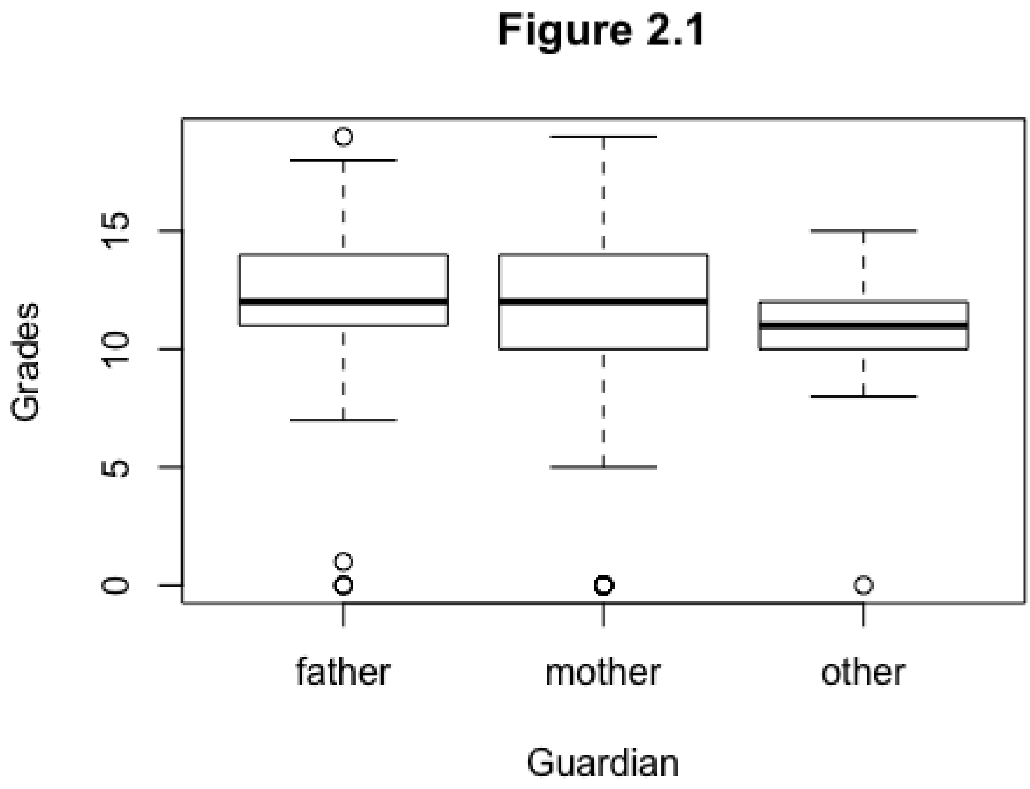
```
##   0   1   2   3   4
## 549  70  16  14   0
```

```
summary(portuguese_df)
```

```
##  school   sex          age        address famsize   Pstatus Medu    Fedu
##  GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80   0:  6   0:  7
##  MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569   1:143   1:174
##                   Median :17.00                             2:186   2:209
##                   Mean   :16.74                             3:139   3:131
##                   3rd Qu.:18.00                             4:175   4:128
##                   Max.   :22.00
##        Mjob          Fjob             reason      guardian   traveltime
##  at_home :135   at_home : 42   course    :285   father:153   0:  0
##  health  : 48   health  : 23   home      :149   mother:455   1:366
##  other   :258   other   :367   other     : 72   other : 41   2:213
##  services:136   services:181   reputation:143                3: 54
##  teacher : 72   teacher : 36                                 4: 16
##
##  studytime failures schoolsup famsup     paid     activities nursery
##  1:212     0:549    no :581   no :251   no :610   no :334    no :128
##  2:305     1: 70    yes: 68   yes:398   yes: 39   yes:315    yes:521
##  3: 97     2: 16
##  4: 35     3: 14
##            4:  0
##
##  higher    internet  romantic  famrel    freetime goout     Dalc      Walc
health
##  no : 69   no :151   no :410   1: 22   1: 45    1: 48   1:451   1:247   1:
90
##  yes:580   yes:498   yes:239   2: 29   2:107    2:145   2:121   2:150   2:
78
##                                3:101   3:251    3:205   3: 43   3:120
3:124
##                                4:317   4:178    4:141   4: 17   4: 87
4:108
##                                5:180   5: 68    5:110   5: 17   5: 45
5:249
##
```

```
##     absences            G3       
##  Min.   : 0.000   Min.   : 0.00  
##  1st Qu.: 0.000   1st Qu.:10.00  
##  Median : 2.000   Median :12.00  
##  Mean   : 3.659   Mean   :11.91  
##  3rd Qu.: 6.000   3rd Qu.:14.00  
##  Max.   :32.000   Max.   :19.00
```

```
############################
## EDA ####################
############################

# Creating box-plots for categorical data
suppressMessages(attach(portuguese_df))

plot(guardian,G3, xlab = "Guardian", ylab = "Grades", main = "Figure 2.1")
```



```
summary(portuguese_df[portuguese_df$guardian=="father",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0    11.0    12.0    12.2    14.0    19.0
```

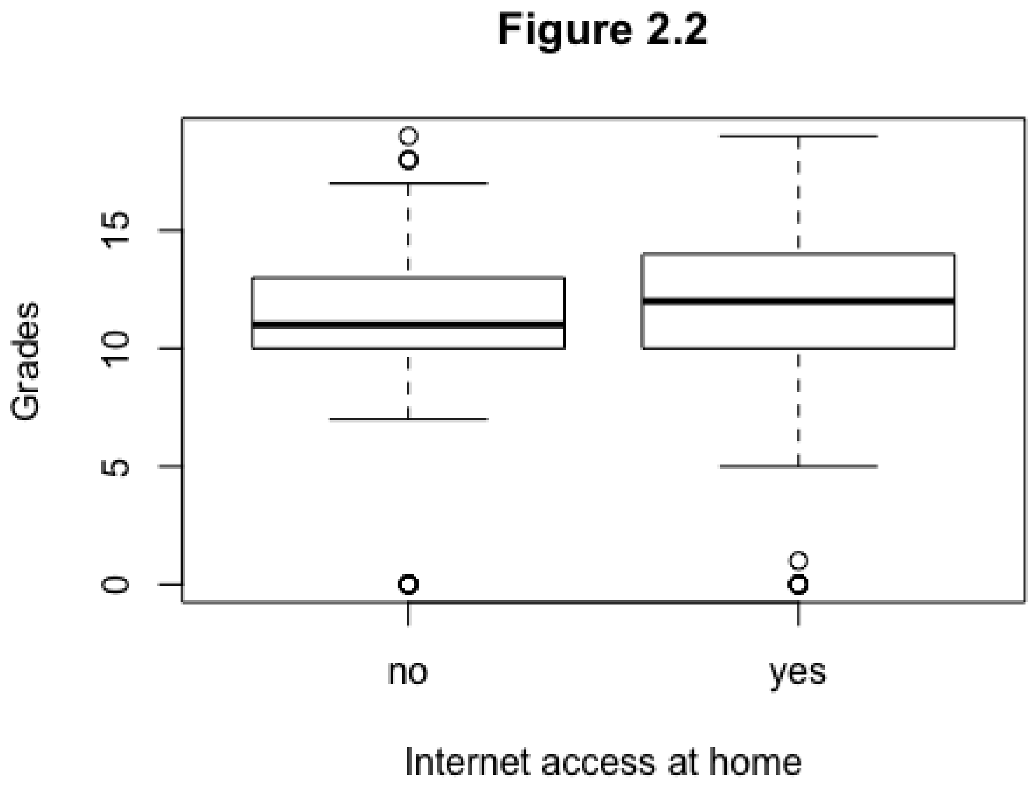
```
summary(portuguese_df[portuguese_df$guardian=="mother",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0    10.0    12.0    11.9    14.0    19.0
```

```
summary(portuguese_df[portuguese_df$guardian=="other",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0    10.0    11.0    10.9    12.0    15.0
```

```
plot(internet,G3, xlab = "Internet access at home", ylab = "Grades", main = 
"Figure 2.2")
```



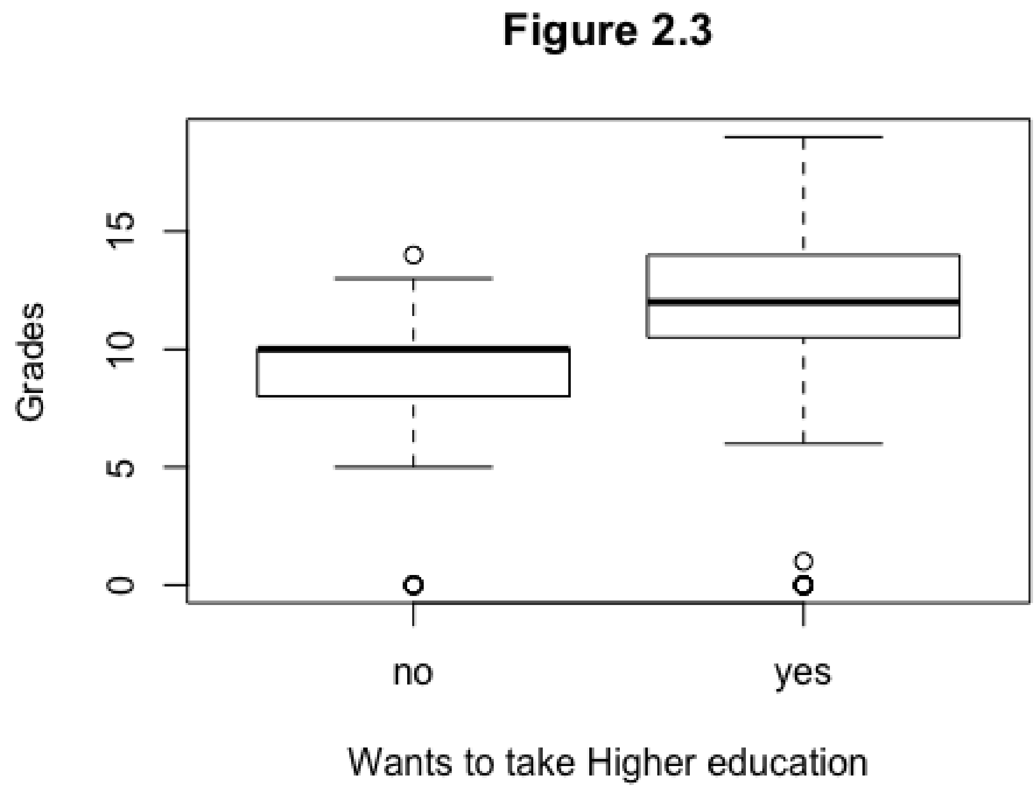
```
summary(portuguese_df[portuguese_df$internet=="yes",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   10.00   12.00   12.17   14.00   19.00
```

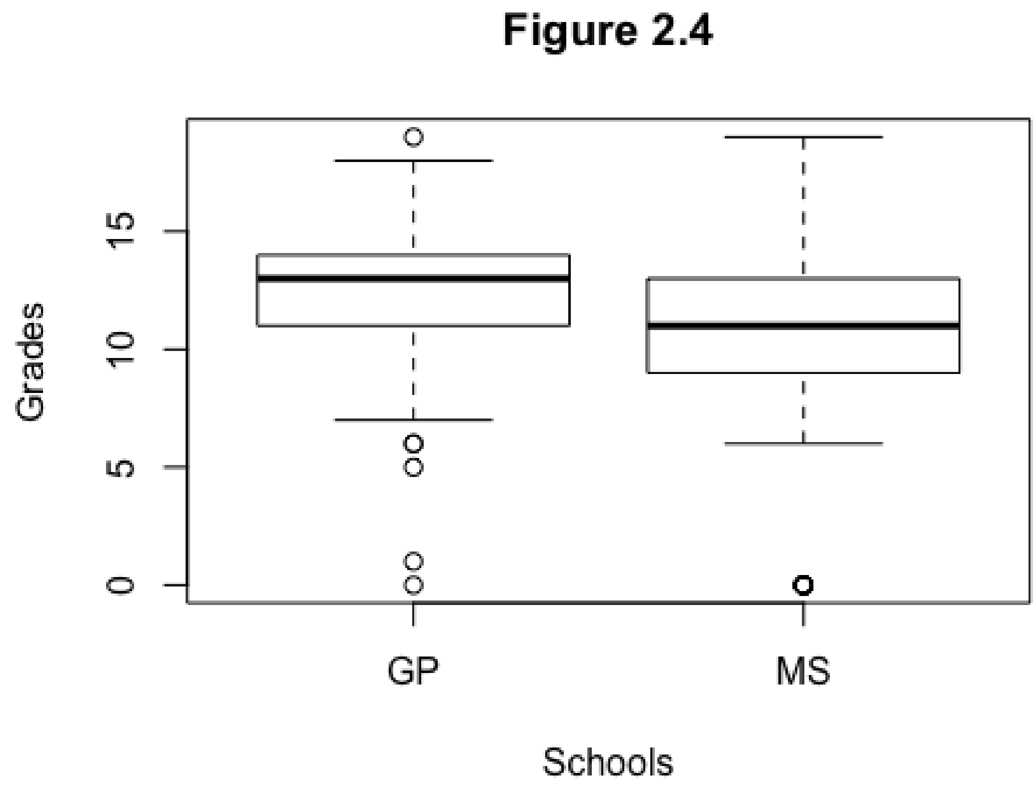
```
summary(portuguese_df[portuguese_df$internet=="no",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   10.00   11.00   11.03   13.00   19.00
```

```
plot(higher,G3, xlab = "Wants to take Higher education", ylab = "Grades", 
main = "Figure 2.3")
```

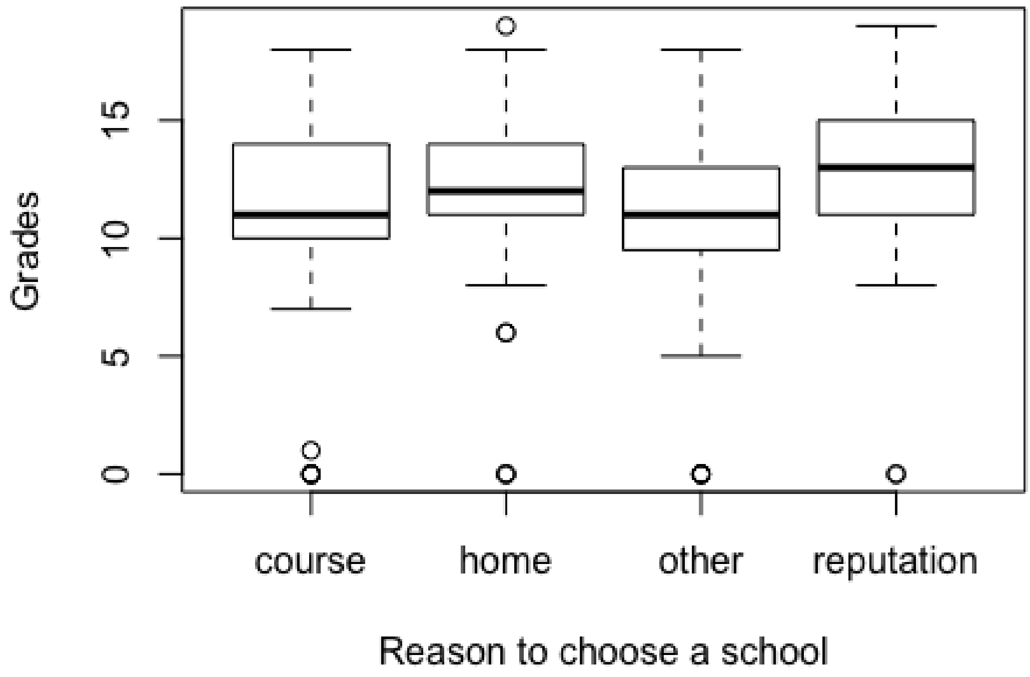


```
summary(portuguese_df[portuguese_df$higher=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.75   12.00   12.28   14.00   19.00

summary(portuguese_df[portuguese_df$higher=="no",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   8.797  10.000  14.000

plot(school, G3, xlab = "Schools", ylab = "Grades", main = "Figure 2.4")
```

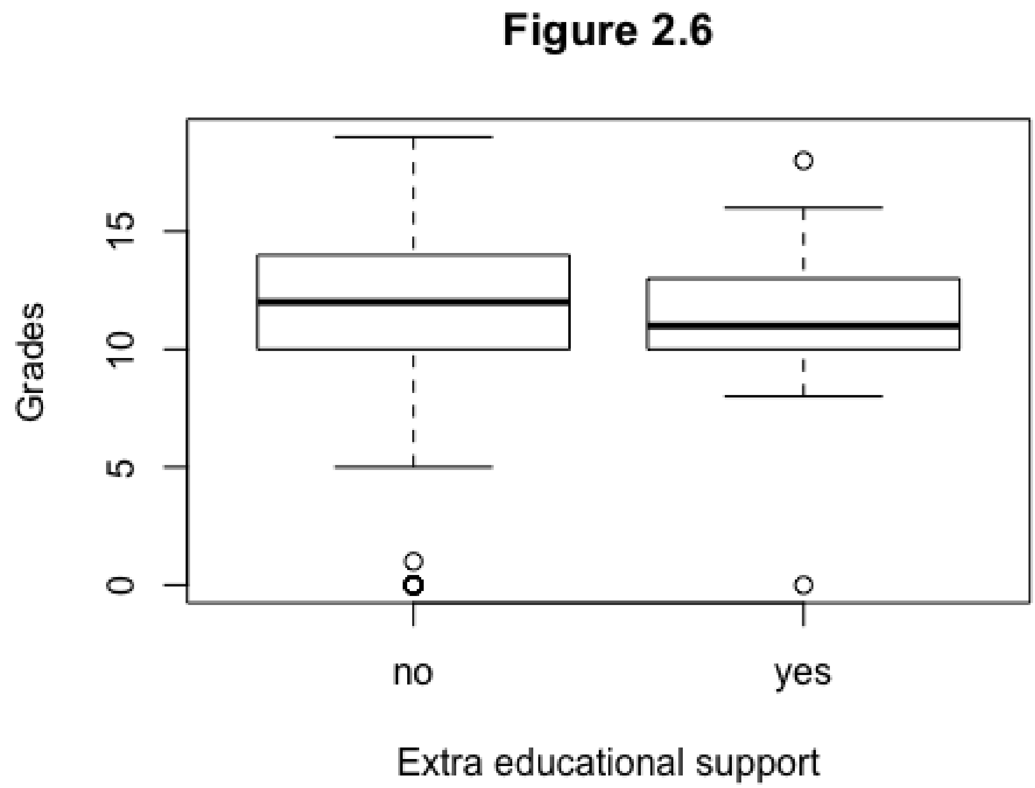


```
summary(portuguese_df[portuguese_df$school=="GP",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   11.00   13.00   12.58   14.00   19.00 

summary(portuguese_df[portuguese_df$school=="MS",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    9.00   11.00   10.65   13.00   19.00 

plot(reason,G3, xlab = "Reason to choose a school", ylab = "Grades", main = 
"Figure 2.5")
```

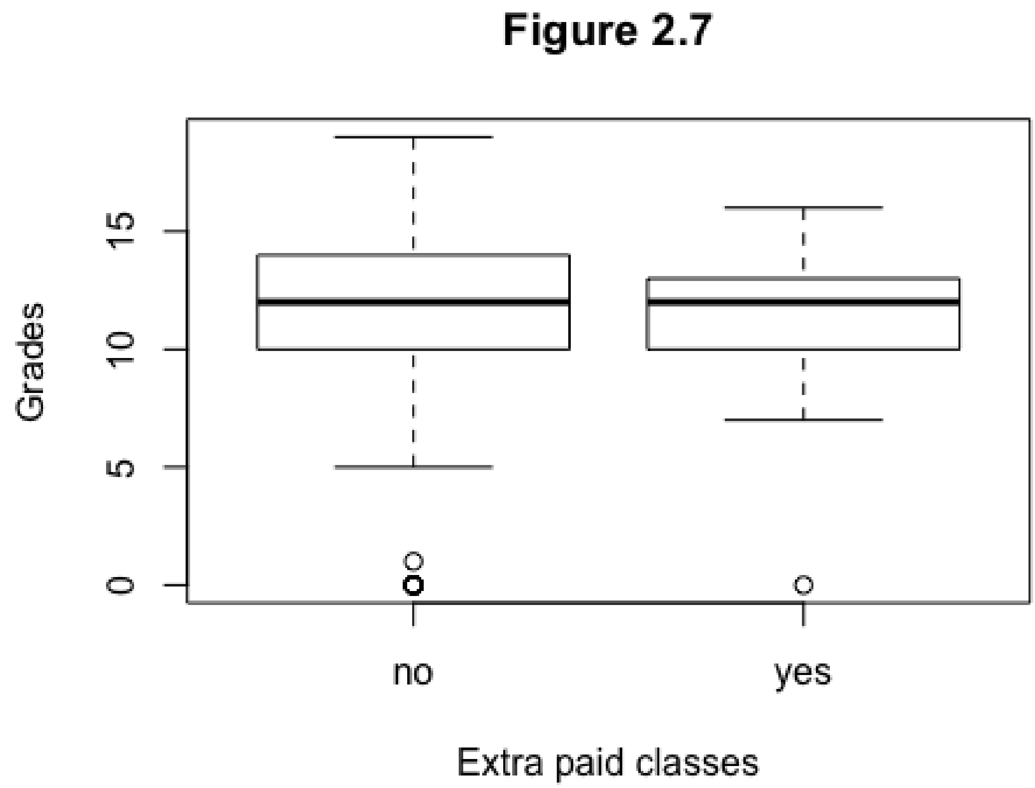
Figure 2.5



```
summary(portuguese_df[portuguese_df$reason=="course",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.55   14.00   18.00

summary(portuguese_df[portuguese_df$reason=="home",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.00   12.00   12.18   14.00   19.00

summary(portuguese_df[portuguese_df$reason=="other",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.75   11.00   10.69   13.00   18.00

summary(portuguese_df[portuguese_df$reason=="reputation",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.00   13.00   12.94   15.00   19.00

plot(schoolsup,G3, xlab = "Extra educational support", ylab = "Grades", main
= "Figure 2.6")
```

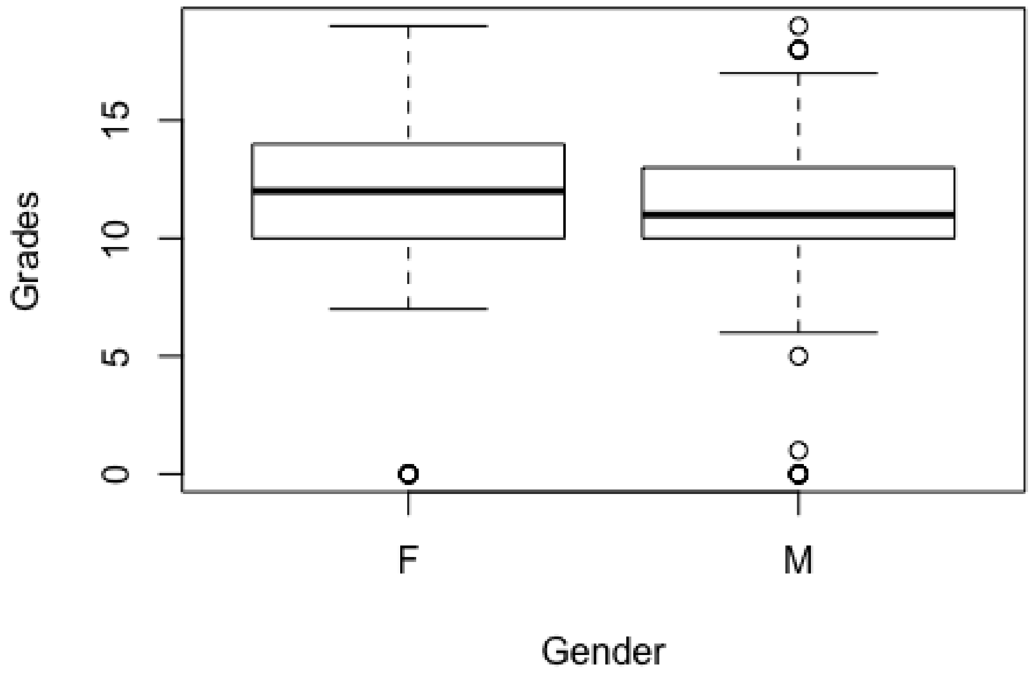


```
summary(portuguese_df[portuguese_df$schoolsup=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.28   13.00   18.00

summary(portuguese_df[portuguese_df$schoolsup=="no",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.98   14.00   19.00

plot(paid, G3, xlab = "Extra paid classes", ylab = "Grades", main = "Figure
2.7")
```

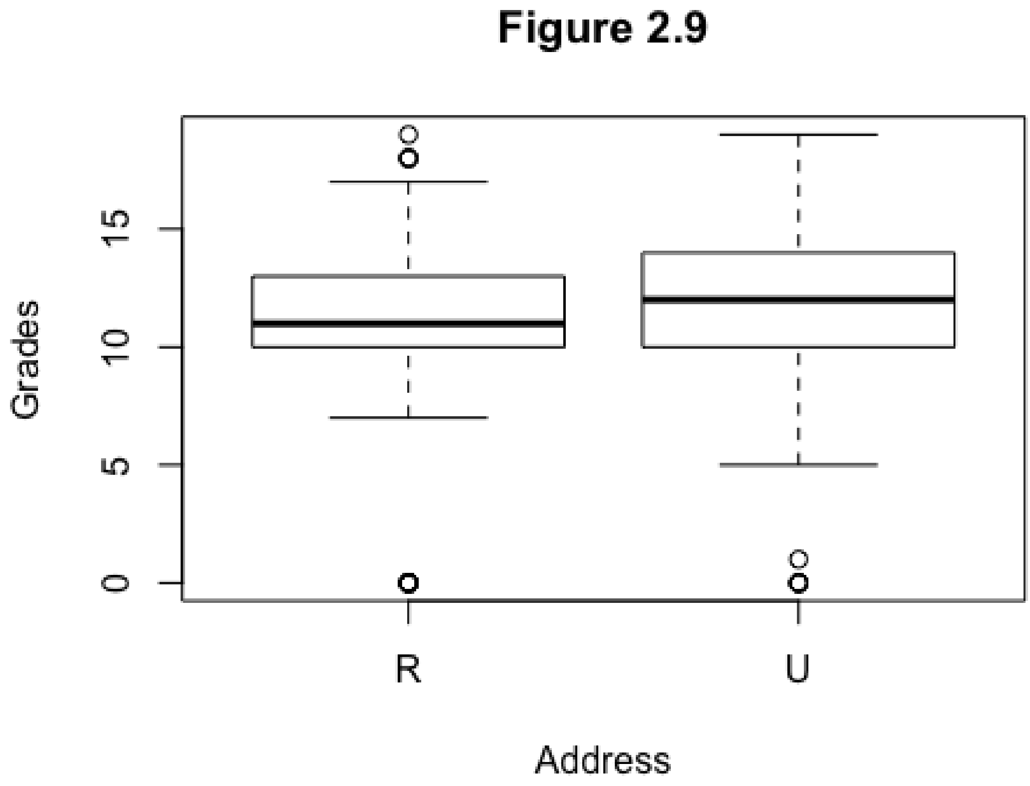


```
summary(portuguese_df[portuguese_df$paid=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.21   13.00   16.00

summary(portuguese_df[portuguese_df$paid=="no",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.95   14.00   19.00

plot(sex,G3, xlab = "Gender", ylab = "Grades", main = "Figure 2.8")
```

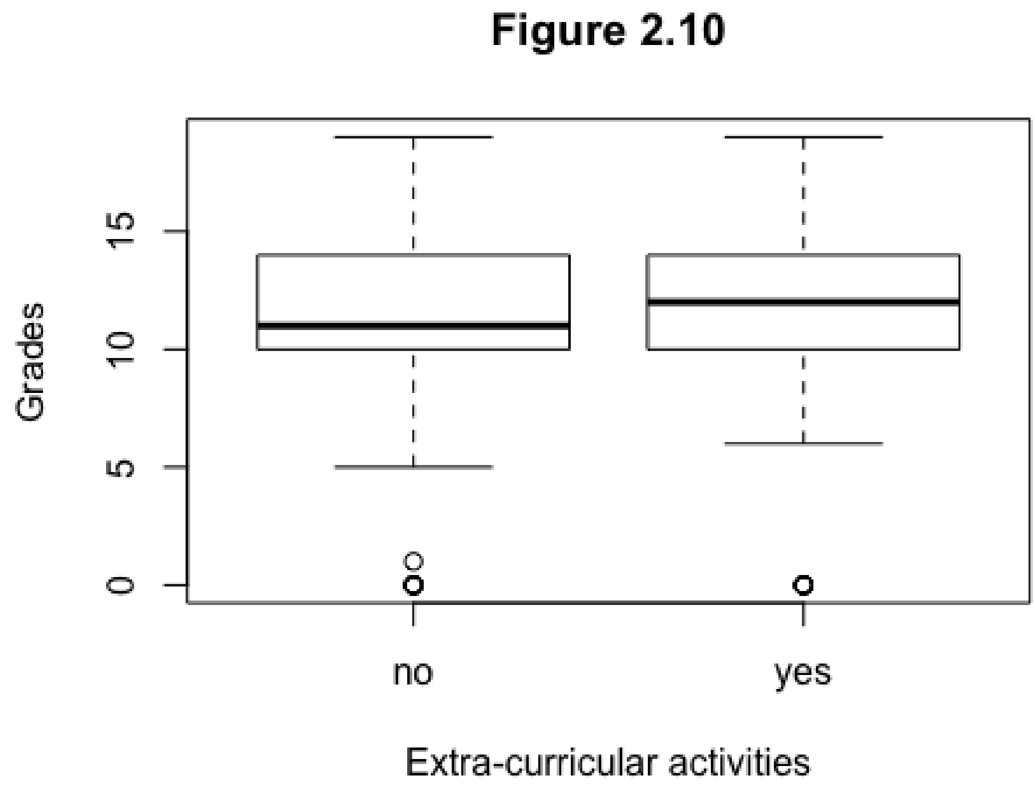
Figure 2.8



```
summary(portuguese_df[portuguese_df$sex=="F",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.25   14.00   19.00

summary(portuguese_df[portuguese_df$sex=="M",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.41   13.00   19.00

plot(address,G3, xlab = "Address", ylab = "Grades", main = "Figure 2.9")
```



```
summary(portuguese_df[portuguese_df$address=="U",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.26   14.00   19.00

summary(portuguese_df[portuguese_df$address=="R",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.09   13.00   19.00

plot(activities,G3, xlab = "Extra-curricular activities", ylab = "Grades",
main = "Figure 2.10")
```



```
summary(portuguese_df[portuguese_df$activities=="yes",]$G3)
```

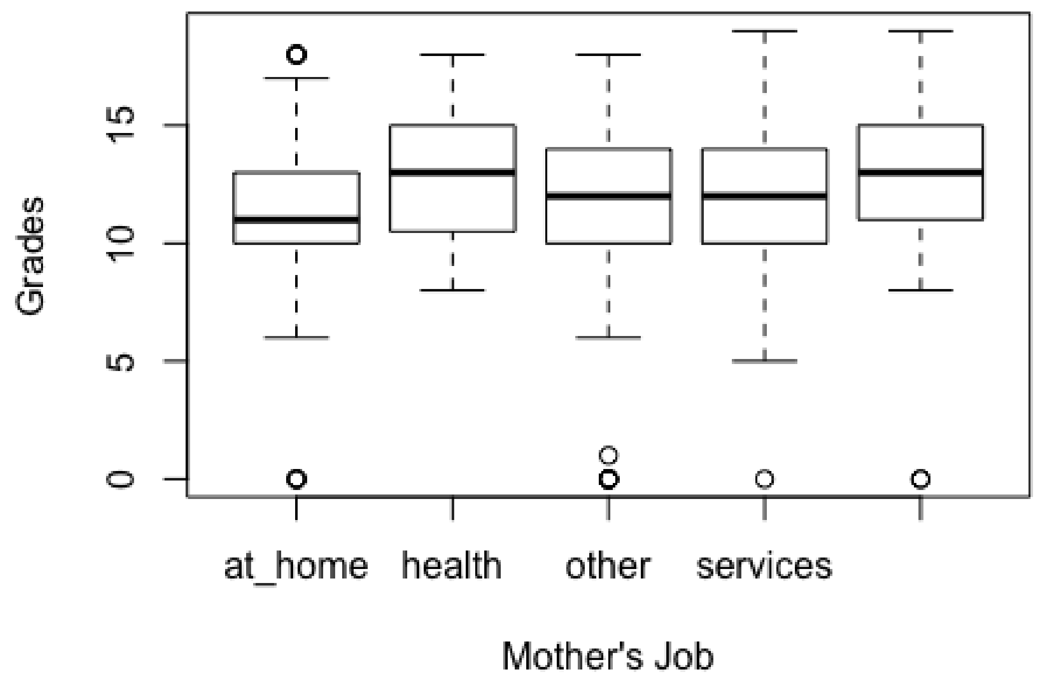
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0    10.0    12.0    12.1    14.0    19.0 
```

```
summary(portuguese_df[portuguese_df$activities=="no",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   10.00   11.00   11.72   14.00   19.00 
```

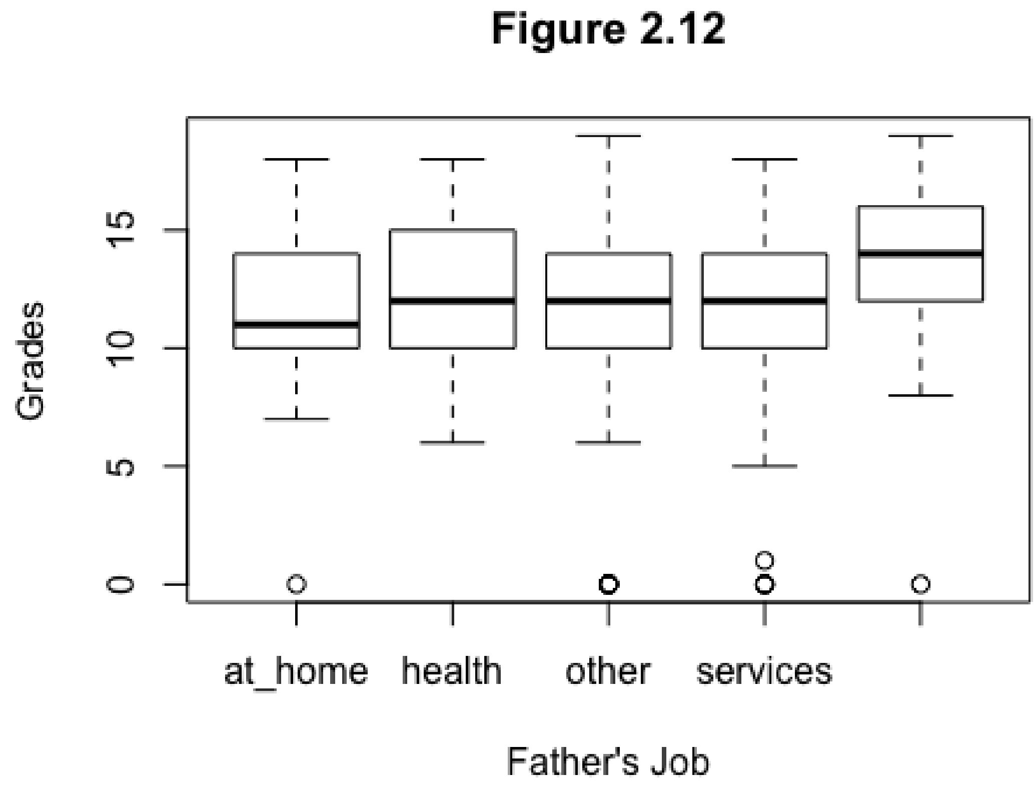
```
plot(Mjob,G3, xlab = "Mother's Job", ylab = "Grades", main = "Figure 2.11")
```

Figure 2.11



```
summary(portuguese_df[portuguese_df$Mjob=="at_home",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.04   13.00   18.00

summary(portuguese_df[portuguese_df$Mjob=="health",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   10.75   13.00   13.06   15.00   18.00

summary(portuguese_df[portuguese_df$Mjob=="other",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.67   14.00   18.00

summary(portuguese_df[portuguese_df$Mjob=="services",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.15   14.00   19.00

summary(portuguese_df[portuguese_df$Mjob=="teacher",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.00   13.00   13.14   15.00   19.00
```

```
plot(Fjob, G3, xlab = "Father's Job", ylab = "Grades", main = "Figure 2.12")
```

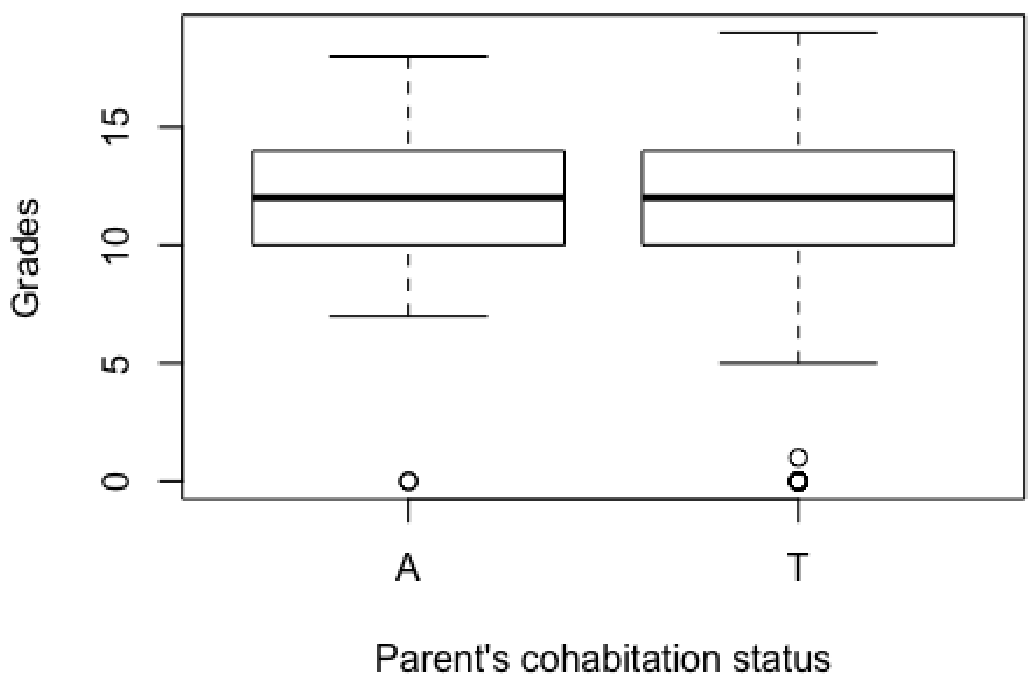


```
summary(portuguese_df[portuguese_df$Fjob=="at_home",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.43   14.00   18.00

summary(portuguese_df[portuguese_df$Fjob=="health",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   10.00   12.00   12.57   15.00   18.00

summary(portuguese_df[portuguese_df$Fjob=="other",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.89   14.00   19.00

summary(portuguese_df[portuguese_df$Fjob=="services",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.63   14.00   18.00

summary(portuguese_df[portuguese_df$Fjob=="teacher",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   14.00   13.58   16.00   19.00
```

```
plot(Pstatus,G3, xlab = "Parent's cohabitation status", ylab = "Grades", main
= "Figure 2.13")
```

Figure 2.13



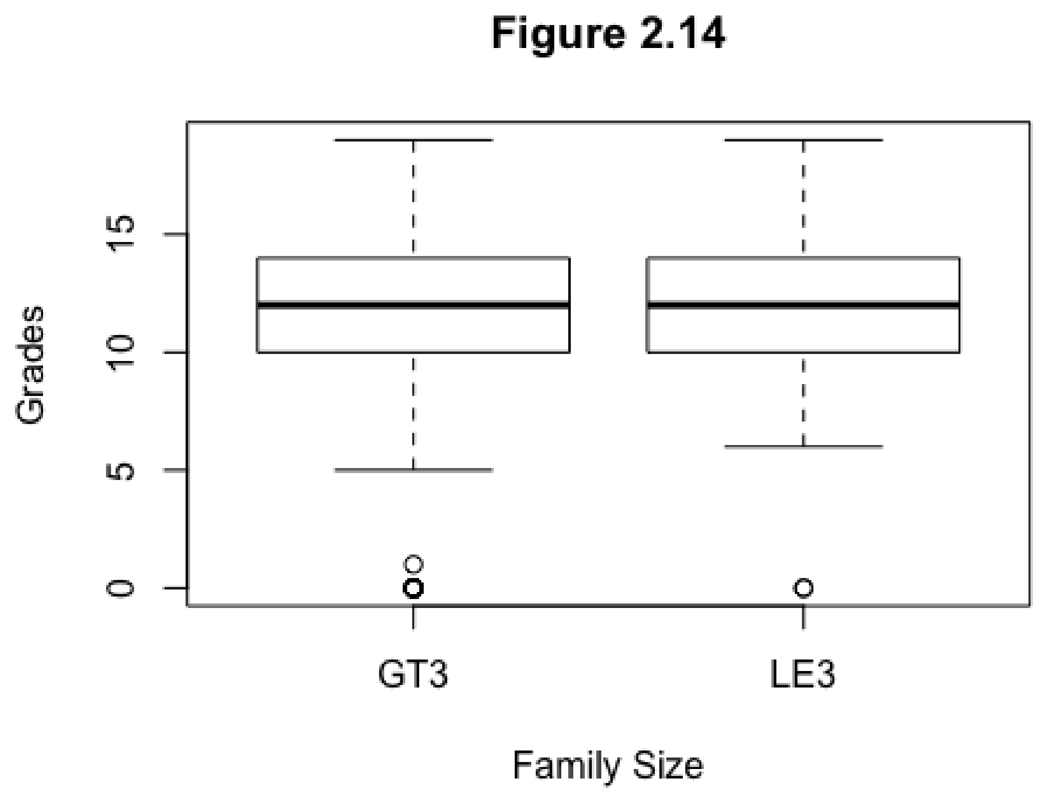
```
summary(portuguese_df[portuguese_df$Pstatus=="A",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.91   14.00   18.00
```

```
summary(portuguese_df[portuguese_df$Pstatus=="T",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.91   14.00   19.00
```

```
plot(famsize, G3, xlab = "Family Size", ylab = "Grades", main = "Figure
2.14")
```



```
summary(portuguese_df[portuguese_df$famsize=="GT3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.81   14.00   19.00

summary(portuguese_df[portuguese_df$famsize=="LE3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.13   14.00   19.00

plot(romantic,G3, xlab = "Romantic relationship", ylab = "Grades", main =
"Figure 2.15" )
```

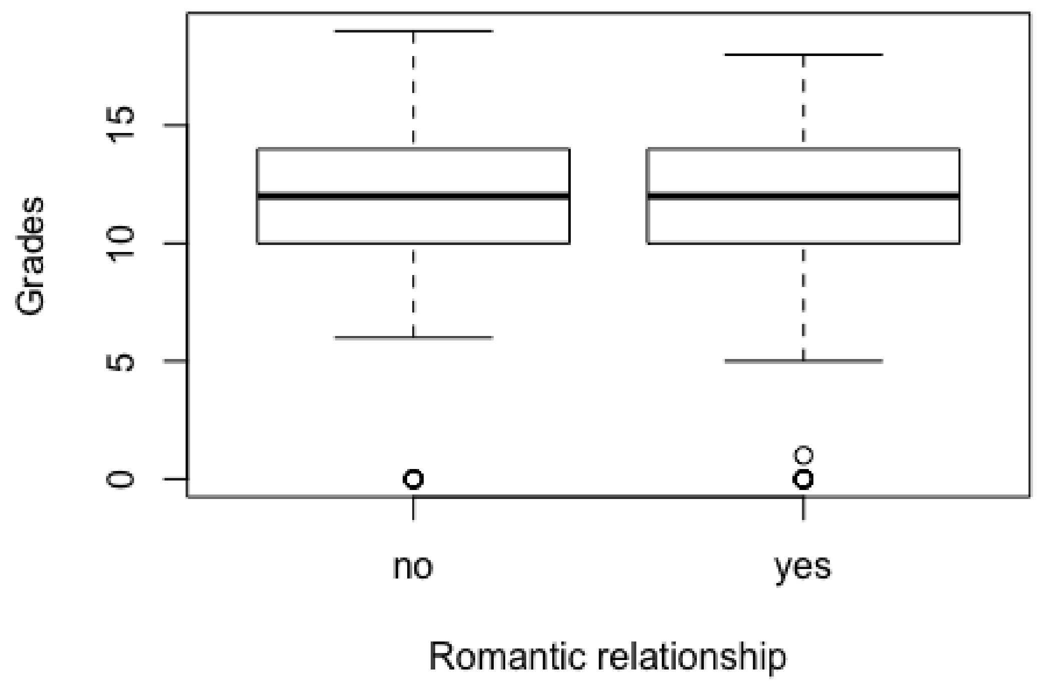
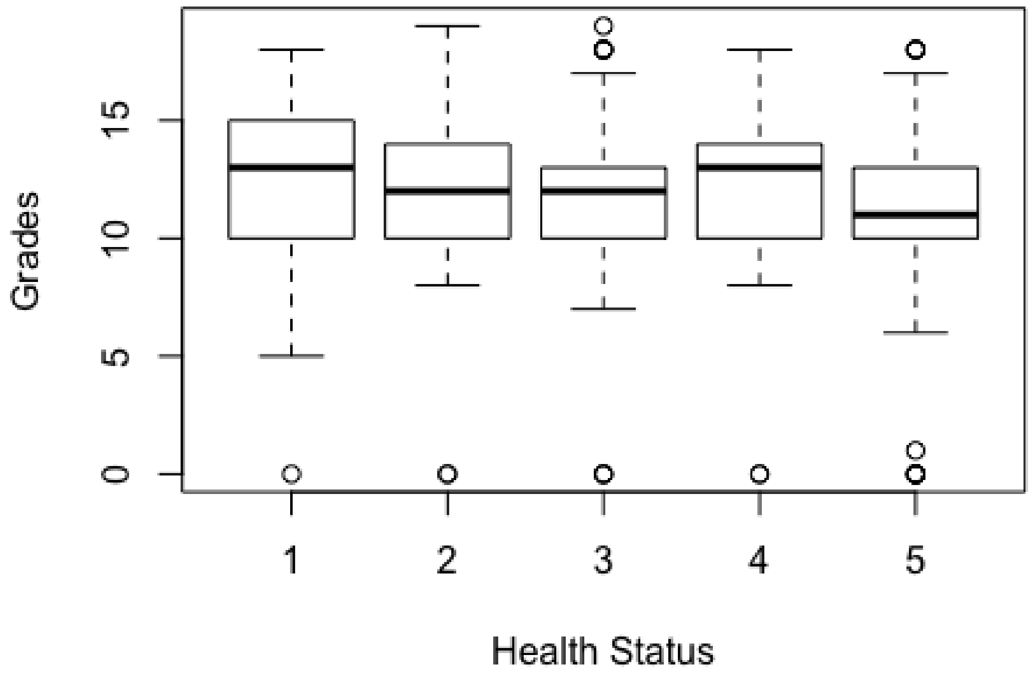


Figure 2.15

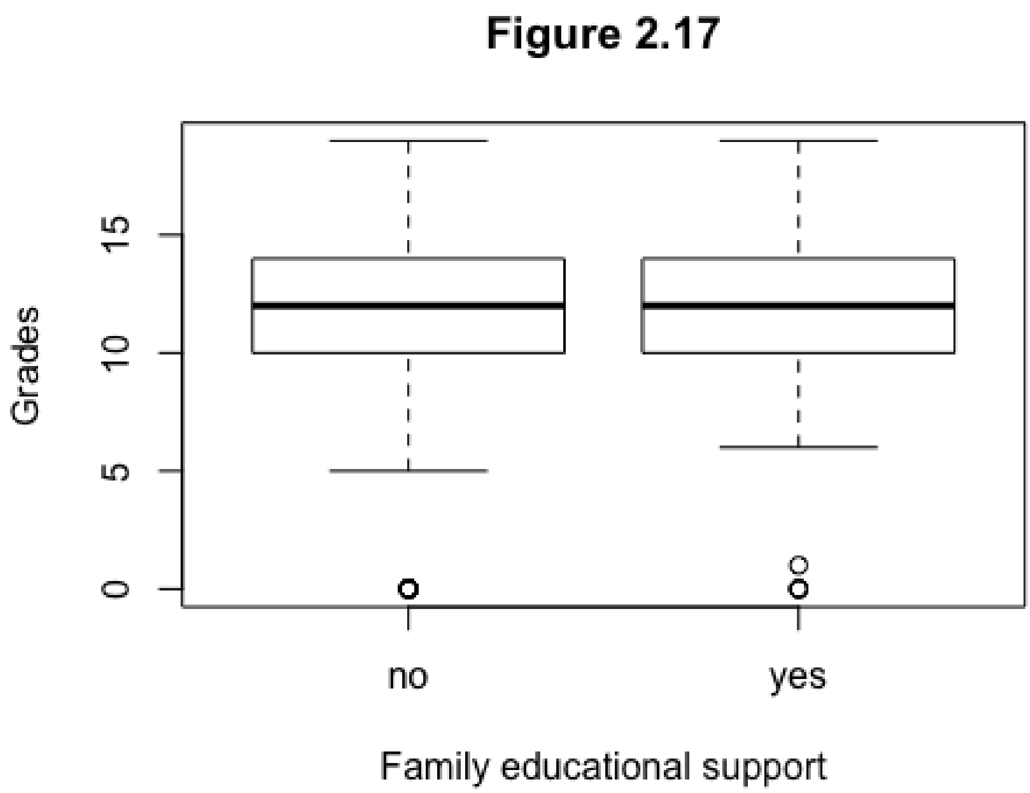
```
summary(portuguese_df[portuguese_df$romantic=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.52   14.00   18.00

summary(portuguese_df[portuguese_df$romantic=="no",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.13   14.00   19.00

plot(health,G3, xlab = "Health Status", ylab = "Grades", main = "Figure
2.16")
```

Figure 2.16

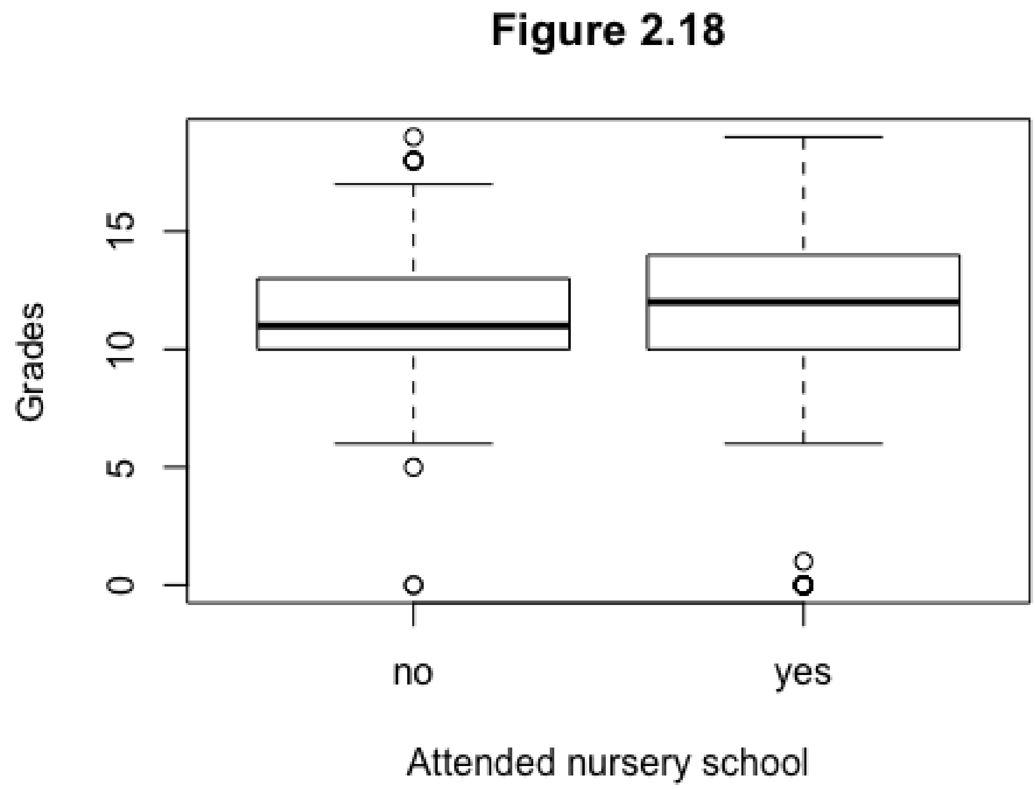


```
summary(portuguese_df[portuguese_df$health=="1",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   13.00   12.48   15.00   18.00

summary(portuguese_df[portuguese_df$health=="2",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.19   14.00   19.00

summary(portuguese_df[portuguese_df$health=="3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.84   13.00   19.00

summary(portuguese_df[portuguese_df$health=="4",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   13.00   12.31   14.00   18.00

summary(portuguese_df[portuguese_df$health=="5",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.47   13.00   18.00
```

```r
plot(famsup,G3, xlab = "Family educational support", ylab = "Grades", main =
"Figure 2.17")
```



```r
summary(portuguese_df[portuguese_df$famsup=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.06   14.00   19.00

summary(portuguese_df[portuguese_df$famsup=="no",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.67   14.00   19.00

plot(nursery,G3, xlab = "Attended nursery school", ylab = "Grades", main =
"Figure 2.18")
```



```
summary(portuguese_df[portuguese_df$activities=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0    10.0    12.0    12.1    14.0    19.0 

summary(portuguese_df[portuguese_df$activities=="no",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   10.00   11.00   11.72   14.00   19.00 

plot(traveltime,G3, xlab = "Traveltime", ylab = "Grades", main = "Figure 2.19")
```

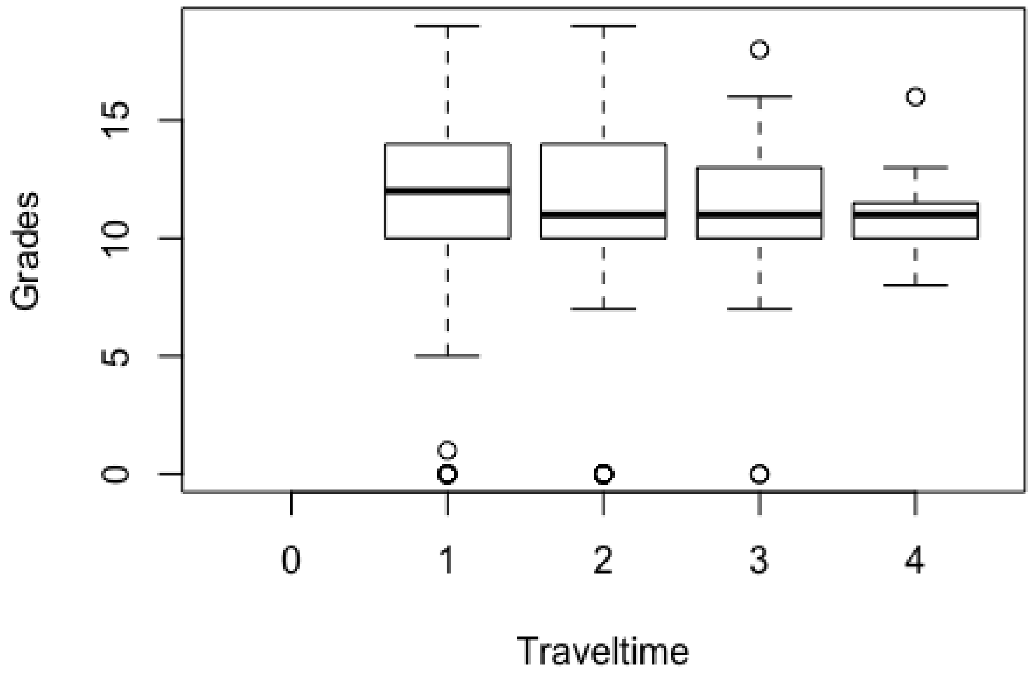
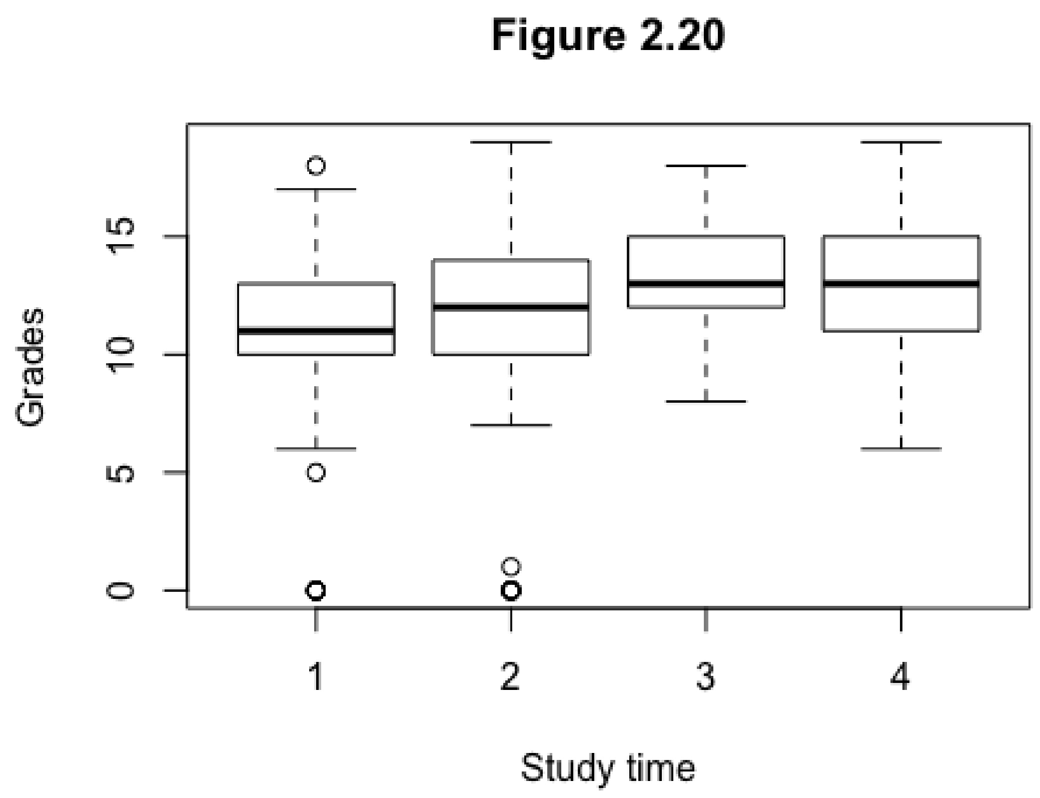


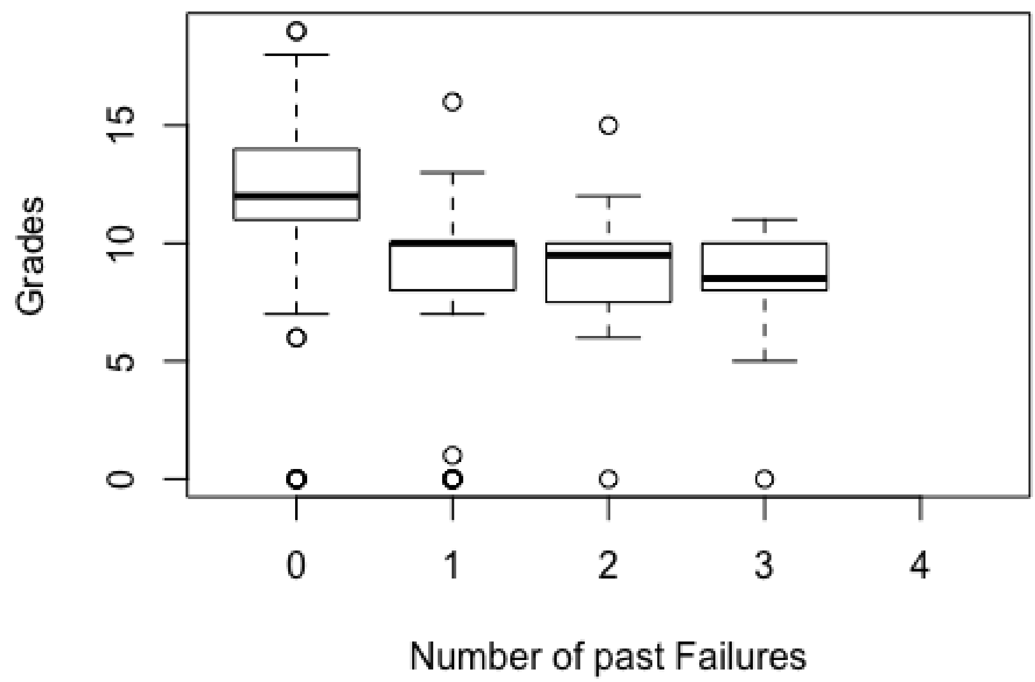
Figure 2.19

```
summary(portuguese_df[portuguese_df$traveltime=="1",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.25   14.00   19.00

summary(portuguese_df[portuguese_df$traveltime=="2",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.58   14.00   19.00

summary(portuguese_df[portuguese_df$traveltime=="3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.17   13.00   18.00

summary(portuguese_df[portuguese_df$traveltime=="4",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   10.00   11.00   10.88   11.25   16.00

plot(studytime,G3, xlab = "Study time", ylab = "Grades", main = "Figure 2.20")
```



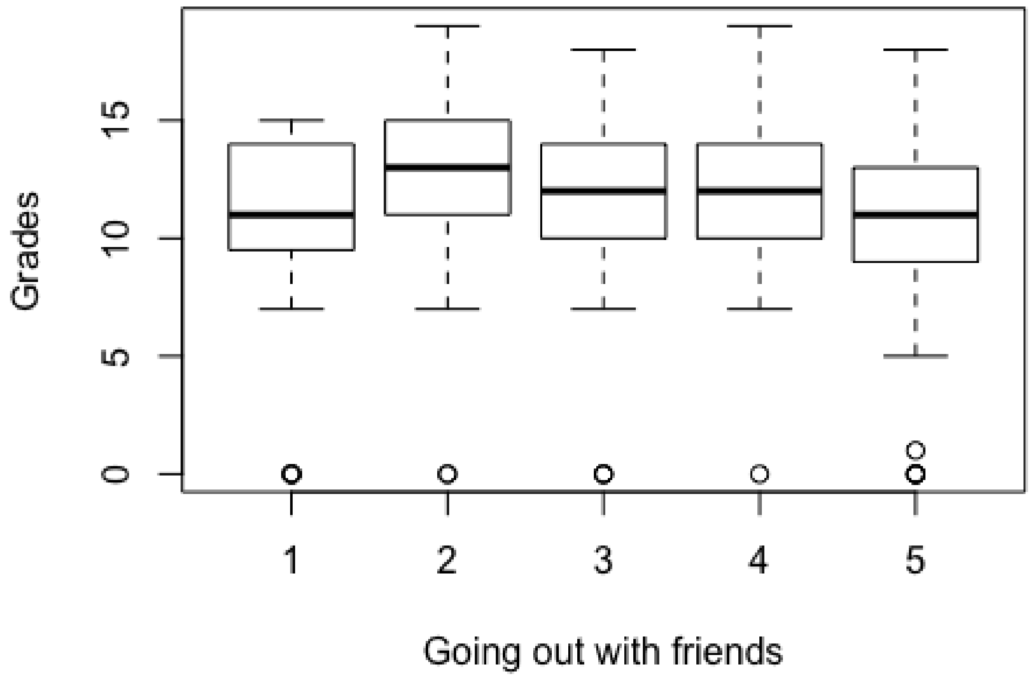
```
summary(portuguese_df[portuguese_df$studytime=="1",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   10.84   13.00   18.00

summary(portuguese_df[portuguese_df$studytime=="2",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.09   14.00   19.00

summary(portuguese_df[portuguese_df$studytime=="3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   12.00   13.00   13.23   15.00   18.00

summary(portuguese_df[portuguese_df$studytime=="4",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   11.00   13.00   13.06   15.00   19.00

plot(failures,G3, xlab = "Number of past Failures", ylab = "Grades", main =
"Figure 2.21")
```

Figure 2.21



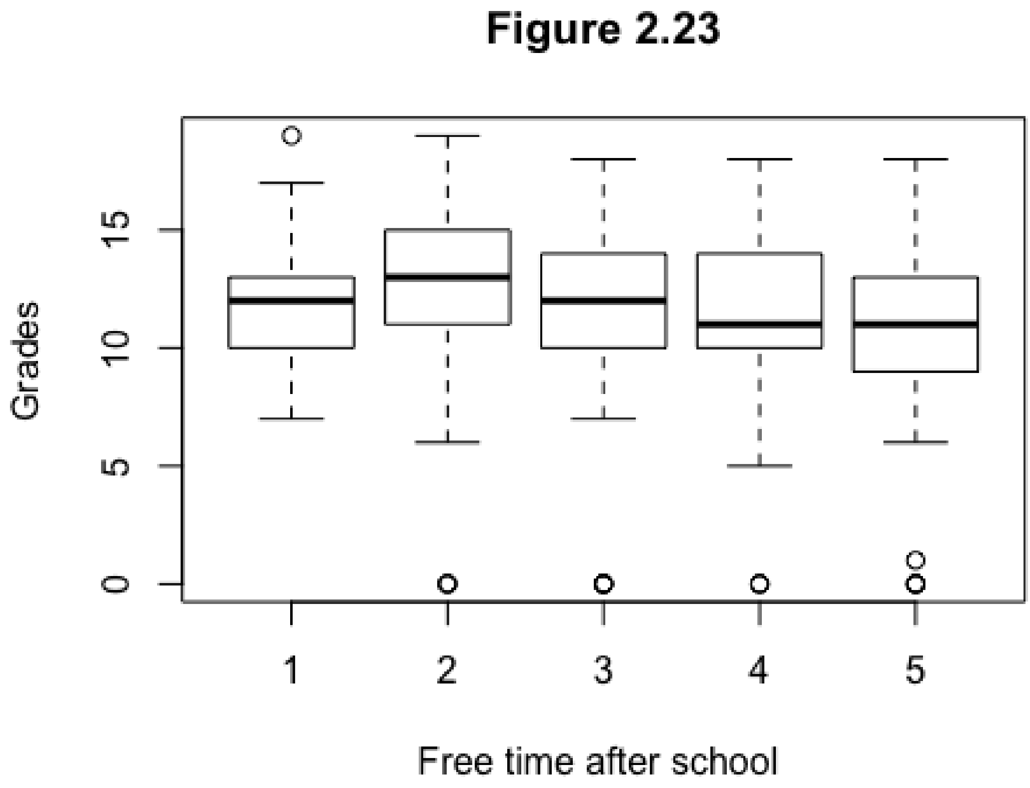
```
summary(portuguese_df[portuguese_df$failures=="0",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.00   12.00   12.51   14.00   19.00

summary(portuguese_df[portuguese_df$failures=="1",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   8.643  10.000  16.000

summary(portuguese_df[portuguese_df$failures=="2",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.750   9.500   8.812  10.000  15.000

summary(portuguese_df[portuguese_df$failures=="3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000   8.500   8.071  10.000  11.000

plot(goout,G3, xlab = "Going out with friends", ylab = "Grades", main =
"Figure 2.22")
```

Figure 2.22



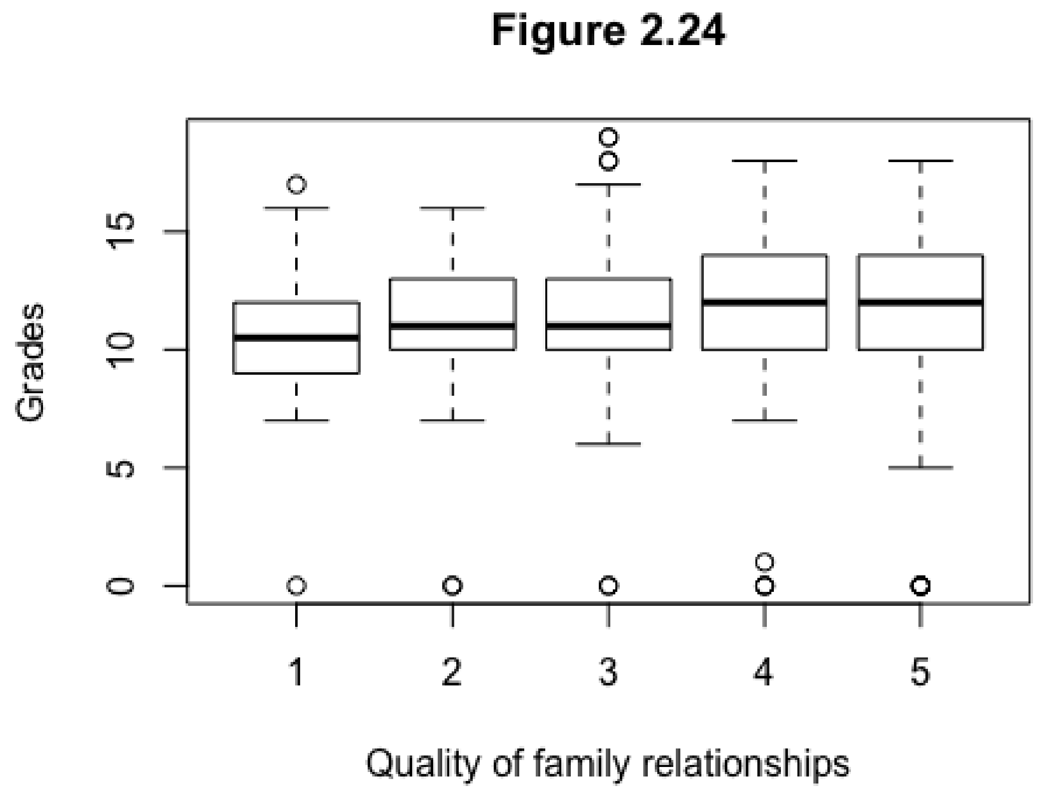
```
summary(portuguese_df[portuguese_df$goout=="1",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.75   11.00   10.73   14.00   15.00

summary(portuguese_df[portuguese_df$goout=="2",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.00   13.00   12.67   15.00   19.00

summary(portuguese_df[portuguese_df$goout=="3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.15   14.00   18.00

summary(portuguese_df[portuguese_df$goout=="4",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.97   14.00   19.00

summary(portuguese_df[portuguese_df$goout=="5",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.25   11.00   10.87   13.00   18.00
```

```
plot(freetime,G3, xlab = "Free time after school ", ylab = "Grades", main =
"Figure 2.23")
```



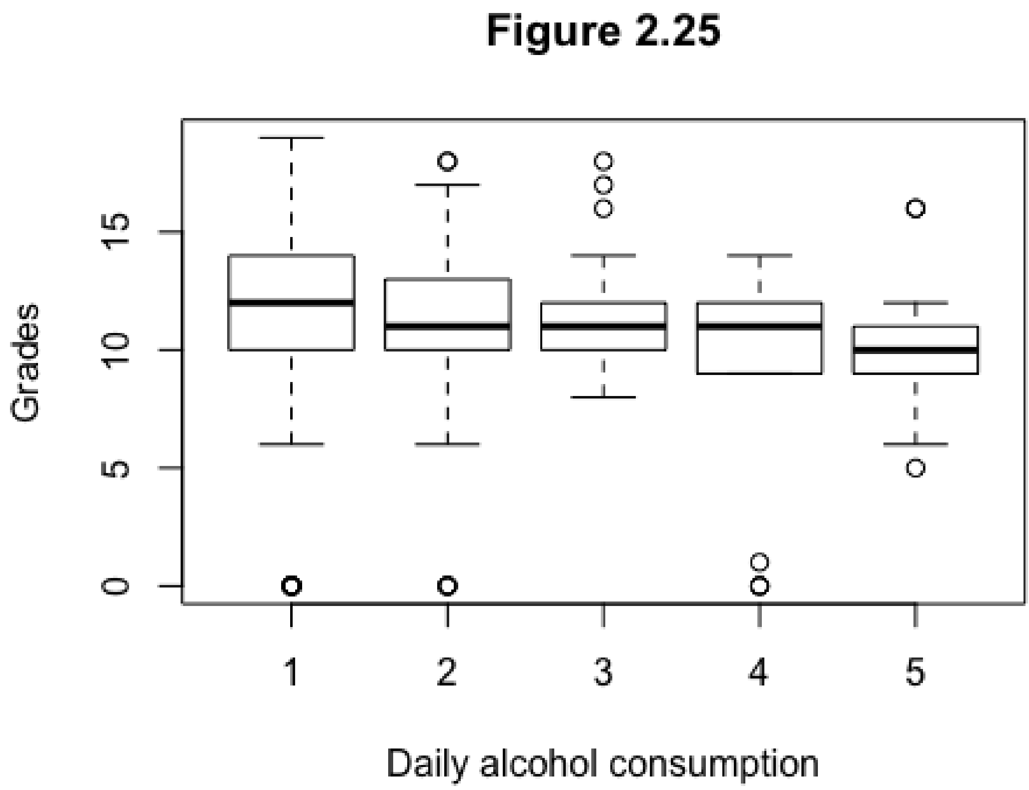
```
summary(portuguese_df[portuguese_df$freetime=="1",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   10.00   12.00   11.73   13.00   19.00

summary(portuguese_df[portuguese_df$freetime=="2",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.00   13.00   12.71   15.00   19.00

summary(portuguese_df[portuguese_df$freetime=="3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.06   14.00   18.00

summary(portuguese_df[portuguese_df$freetime=="4",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.71   14.00   18.00

summary(portuguese_df[portuguese_df$freetime=="5",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    9.00   11.00   10.69   13.00   18.00

plot(famrel,G3, xlab = "Quality of family relationships", ylab = "Grades",
main = "Figure 2.24")
```



```
summary(portuguese_df[portuguese_df$famrel=="1",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    9.00   10.50   10.64   12.00   17.00

summary(portuguese_df[portuguese_df$famrel=="2",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   10.00   11.00   10.86   13.00   16.00

summary(portuguese_df[portuguese_df$famrel=="3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   10.00   11.00   11.59   13.00   19.00

summary(portuguese_df[portuguese_df$famrel=="4",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   10.00   12.00   12.34   14.00   18.00
```

```
summary(portuguese_df[portuguese_df$famrel=="5",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.63   14.00   18.00

plot(Dalc,G3, xlab = "Daily alcohol consumption", ylab = "Grades", main =
"Figure 2.25")
```



```
summary(portuguese_df[portuguese_df$Dalc=="1",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    10.0    12.0    12.3    14.0    19.0

summary(portuguese_df[portuguese_df$Dalc=="2",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.36   13.00   18.00

summary(portuguese_df[portuguese_df$Dalc=="3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   10.00   11.00   11.14   12.00   18.00

summary(portuguese_df[portuguese_df$Dalc=="4",]$G3)
```

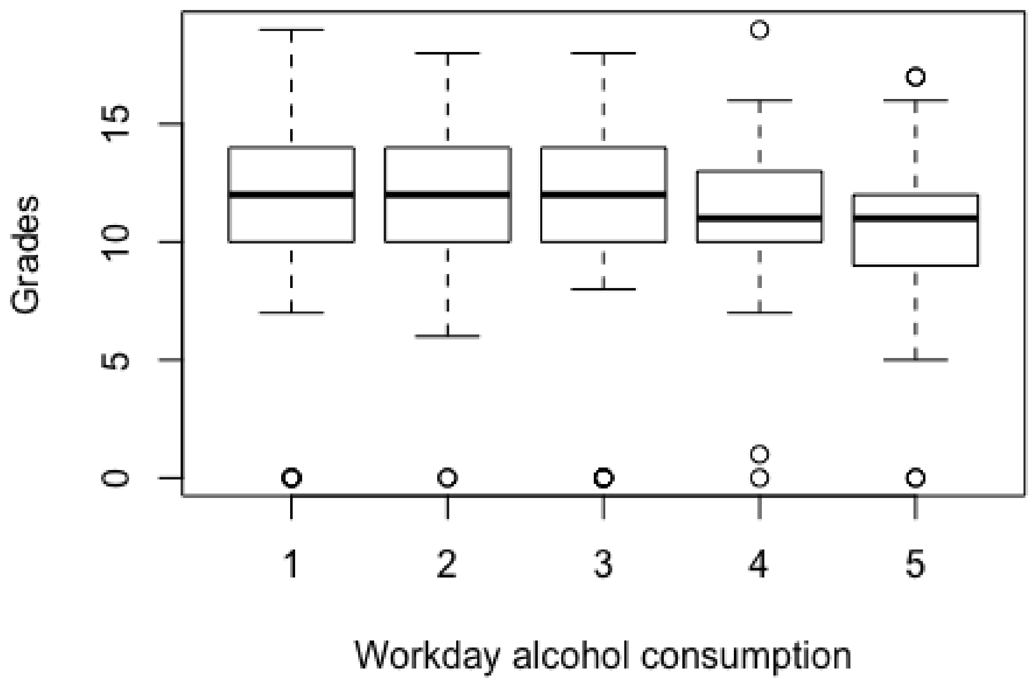
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   9.000  11.000   8.941  12.000  14.000
```

```
summary(portuguese_df[portuguese_df$Dalc=="5",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00    9.00   10.00   10.24   11.00   16.00
```

```
plot(Walc,G3, xlab = "Workday alcohol consumption", ylab = "Grades", main =
"Figure 2.26")
```

Figure 2.26



```
summary(portuguese_df[portuguese_df$Walc=="1",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.36   14.00   19.00
```

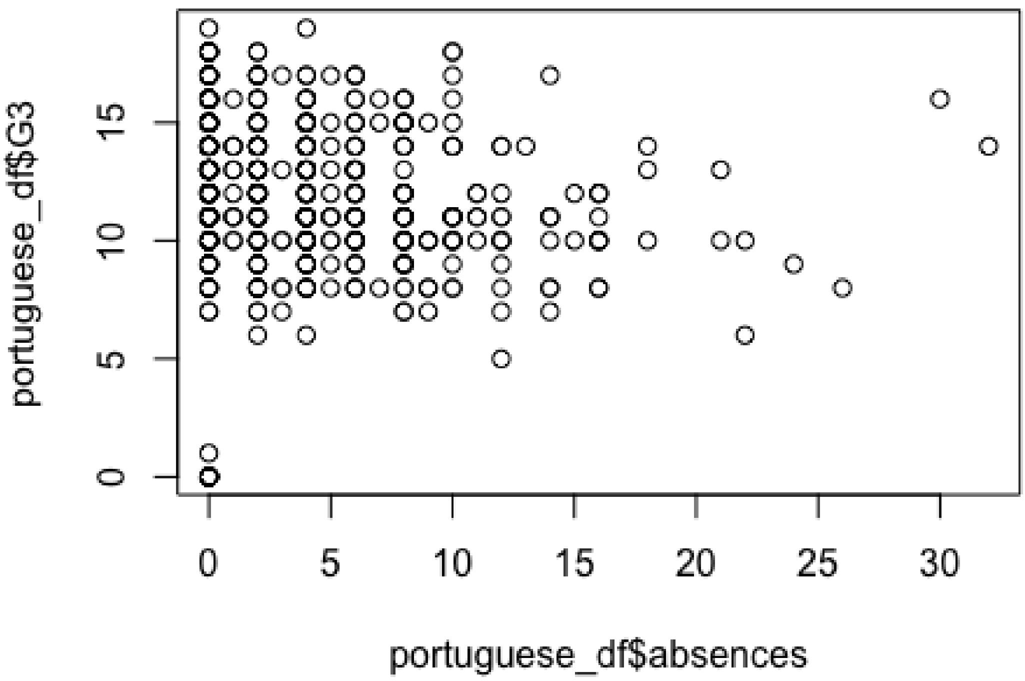
```
summary(portuguese_df[portuguese_df$Walc=="2",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   12.26   14.00   18.00
```

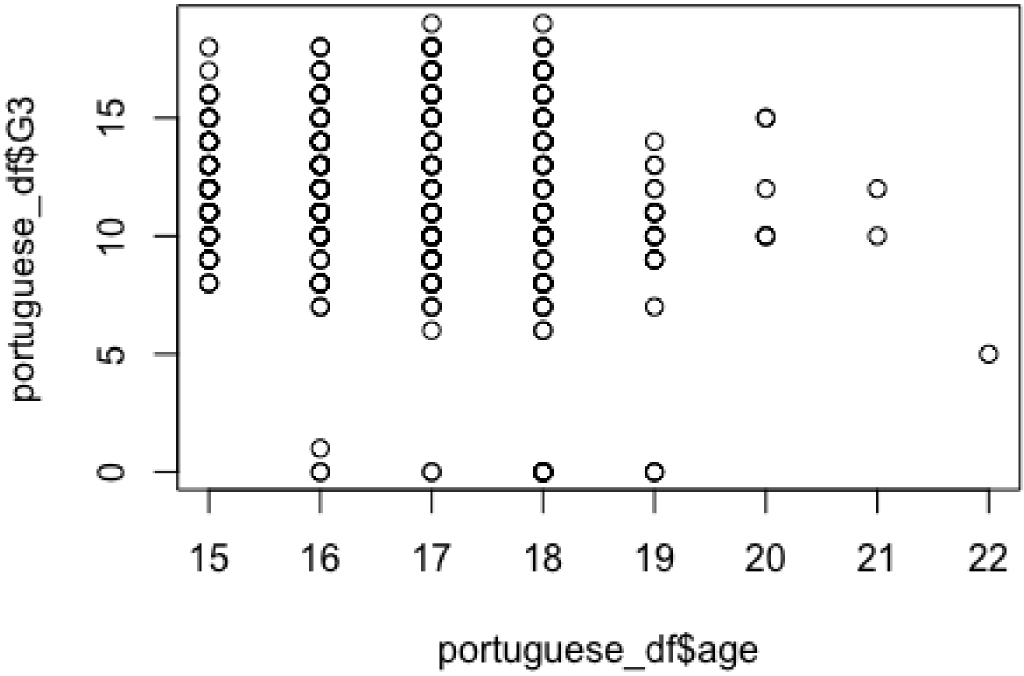
```
summary(portuguese_df[portuguese_df$Walc=="3",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.67   14.00   18.00
```

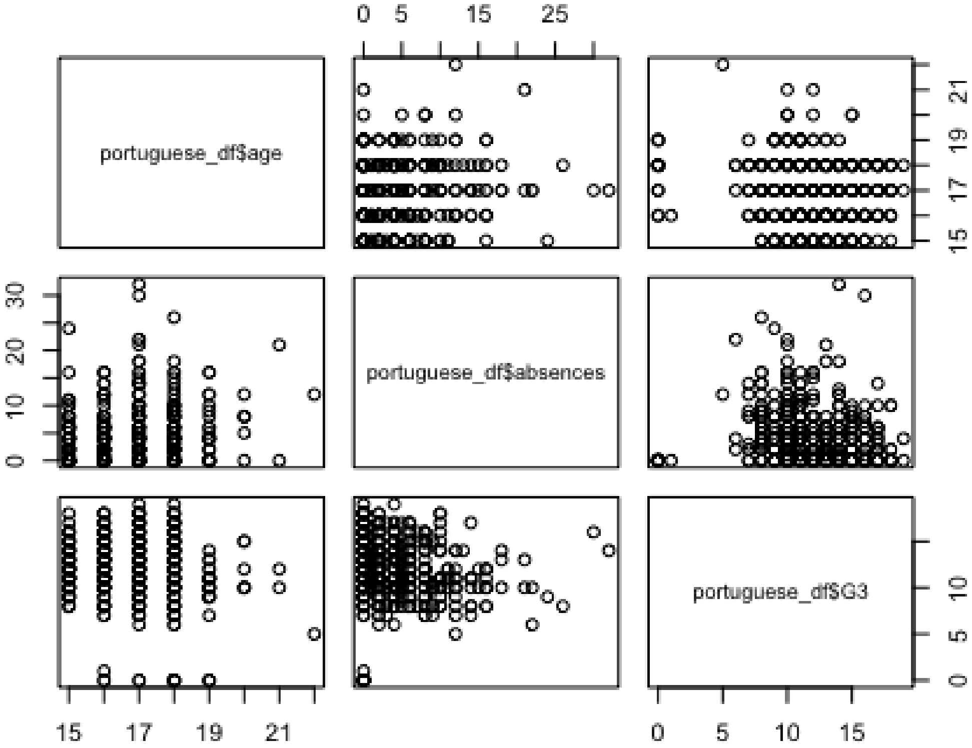
```
summary(portuguese_df[portuguese_df$Walc=="4",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   11.00   11.03   13.00   19.00

summary(portuguese_df[portuguese_df$Walc=="5",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.56   12.00   17.00

# Creating Scatter plots for numerical data
plot(portuguese_df$absences,portuguese_df$G3)
```



```
plot(portuguese_df$age,portuguese_df$G3)
```



```
pairs(~portuguese_df$age+portuguese_df$absences+portuguese_df$G3)
```



```
############################
## Train / Test Split ######
############################

set.seed(1)
train = sample(1:nrow(portuguese_df), 520)
test_g3 = portuguese_df[-train,31]

############################
## Modeling ################
############################

# Linear Model

linear_model_fit <- lm(G3~.,data = portuguese_df[train,])
summary(linear_model_fit)

##
## Call:
## lm(formula = G3 ~ ., data = portuguese_df[train, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8829  -1.3405   0.0279   1.5571   6.7948 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       4.613084   2.365502   1.950 0.051779 .  
## schoolMS         -1.111959   0.315184  -3.528 0.000462 ***
## sexM             -0.520948   0.282418  -1.845 0.065753 .  
## age               0.226252   0.116342   1.945 0.052432 .  
## addressU          0.321364   0.301618   1.065 0.287236    
## famsizeLE3        0.142445   0.274826   0.518 0.604497    
## PstatusT          0.082233   0.384781   0.214 0.830866    
## Medu.L            0.274672   0.885410   0.310 0.756537    
## Medu.Q           -0.054613   0.719701  -0.076 0.939546    
## Medu.C            0.185149   0.474936   0.390 0.696839    
## Medu^4           -0.307310   0.293969  -1.045 0.296408    
## Fedu.L            0.563404   0.821938   0.685 0.493408    
## Fedu.Q            0.164293   0.667540   0.246 0.805704    
## Fedu.C            0.040131   0.452159   0.089 0.929317    
## Fedu^4            0.197860   0.275840   0.717 0.473560    
## Mjobhealth        0.969051   0.598919   1.618 0.106363    
## Mjobother         0.252819   0.340631   0.742 0.458349    
## Mjobservices      0.798819   0.425706   1.876 0.061239 .  
## Mjobteacher       0.571573   0.587806   0.972 0.331382    
## Fjobhealth       -1.030243   0.859377  -1.199 0.231227    
## Fjobother        -0.393582   0.520649  -0.756 0.450078    
## Fjobservices     -1.003251   0.547734  -1.832 0.067666 .  
## Fjobteacher       0.063687   0.773445   0.082 0.934411    
## reasonhome       -0.304664   0.321596  -0.947 0.343969    
## reasonother      -0.452472   0.417329  -1.084 0.278852    
## reasonreputation -0.203243   0.339362  -0.599 0.549543    
## guardianmother   -0.141413   0.303776  -0.466 0.641785    
## guardianother     0.450011   0.607851   0.740 0.459485    
## traveltime.L     -0.321541   0.603797  -0.533 0.594621    
## traveltime.Q     -0.525934   0.502161  -1.047 0.295504    
## traveltime.C     -0.362253   0.371082  -0.976 0.329486    
## studytime.L       0.742378   0.412528   1.800 0.072596 .  
## studytime.Q      -0.004087   0.360846  -0.011 0.990969    
## studytime.C      -0.183806   0.286979  -0.640 0.522182    
## failures.L       -2.235786   0.644455  -3.469 0.000572 ***
## failures.Q        1.166702   0.583399   2.000 0.046119 *  
## failures.C       -0.254607   0.575655  -0.442 0.658492    
## schoolsupyes     -1.008306   0.408043  -2.471 0.013840 *  
## famsupyes        -0.084042   0.264248  -0.318 0.750602    
## paidyes          -0.535484   0.495854  -1.080 0.280755    
## activitiesyes     0.139745   0.254341   0.549 0.582977    
## nurseryyes       -0.265389   0.309511  -0.857 0.391655    
## higheryes         1.964201   0.450148   4.363 1.59e-05 ***
## internetyes       0.498284   0.312918   1.592 0.112002    
## romanticyes      -0.480562   0.260608  -1.844 0.065840 .  
## famrel.L          0.555604   0.502178   1.106 0.269149    
## famrel.Q         -0.340863   0.453391  -0.752 0.452559    
## famrel.C         -0.372056   0.466278  -0.798 0.425333    
## famrel^4         -0.215004   0.384816  -0.559 0.576632    
## freetime.L       -0.640388   0.430882  -1.486 0.137920    
## freetime.Q       -0.262625   0.367495  -0.715 0.475204    
## freetime.C        0.150778   0.311525   0.484 0.628621    
## freetime^4       -0.500048   0.241155  -2.074 0.038689 *  
## goout.L           0.095866   0.399213   0.240 0.810334    
## goout.Q          -1.003146   0.338987  -2.959 0.003247 ** 
## goout.C           0.401323   0.290962   1.379 0.168489    
## goout^4          -0.122062   0.243822  -0.501 0.616883    
## Dalc.L           -0.928698   0.739216  -1.256 0.209649    
## Dalc.Q            0.916809   0.617355   1.485 0.138228    
## Dalc.C            1.617865   0.593348   2.727 0.006648 ** 
## Dalc^4            1.588145   0.531205   2.990 0.002946 ** 
## Walc.L            0.031320   0.522635   0.060 0.952241    
## Walc.Q            0.258706   0.387106   0.668 0.504280    
## Walc.C            0.233010   0.321900   0.724 0.469530    
## Walc^4            0.020608   0.293652   0.070 0.944084    
## health.L         -0.748365   0.281443  -2.659 0.008116 ** 
## health.Q          0.182600   0.287497   0.635 0.525662    
## health.C         -0.352734   0.317469  -1.111 0.267126    
## health^4         -0.123253   0.302338  -0.408 0.683714    
## absences         -0.020404   0.027821  -0.733 0.463695    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.609 on 450 degrees of freedom
## Multiple R-squared:  0.4396, Adjusted R-squared:  0.3537 
## F-statistic: 5.117 on 69 and 450 DF,  p-value: < 2.2e-16
```

```
# Backward Stepwise selection
library(leaps)
back_aic_fit = MASS::stepAIC(linear_model_fit, direction = "backward", trace
= FALSE)
back_aic_fit$anova
```

```
## Stepwise Model Path 
## Analysis of Deviance Table
## 
## Initial Model:
## 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
## 
## Final Model:
## G3 ~ school + sex + age + Fjob + studytime + failures + schoolsup + 
##     higher + internet + romantic + freetime + goout + Dalc + 
##     health
##
##
##              Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1                                      450   3062.844 1062.101
## 2          - Walc  4  5.35494949       454   3068.199 1055.009
## 3          - Medu  4  8.93147245       458   3077.130 1048.520
## 4    - traveltime  3  7.85847236       461   3084.989 1043.847
## 5          - Fedu  4 22.92488520       465   3107.914 1039.697
## 6        - reason  3 11.25018993       468   3119.164 1035.576
## 7        - famsup  1  0.03430424       469   3119.198 1033.581
## 8       - Pstatus  1  0.21604554       470   3119.414 1031.617
## 9      - guardian  2 12.81083344       472   3132.225 1029.748
## 10      - famsize  1  2.28425696       473   3134.509 1028.128
## 11 - activities  1  3.01784163       474   3137.527 1026.628
## 12      - address  1  4.78380647       475   3142.311 1025.420
## 13     - absences  1  5.04216225       476   3147.353 1024.254
## 14      - nursery  1  5.55451929       477   3152.908 1023.171
## 15         - paid  1  6.35212573       478   3159.260 1022.217
## 16       - famrel  4 47.73272757       482   3206.992 1022.015
## 17         - Mjob  4 41.17981615       486   3248.172 1020.650
```

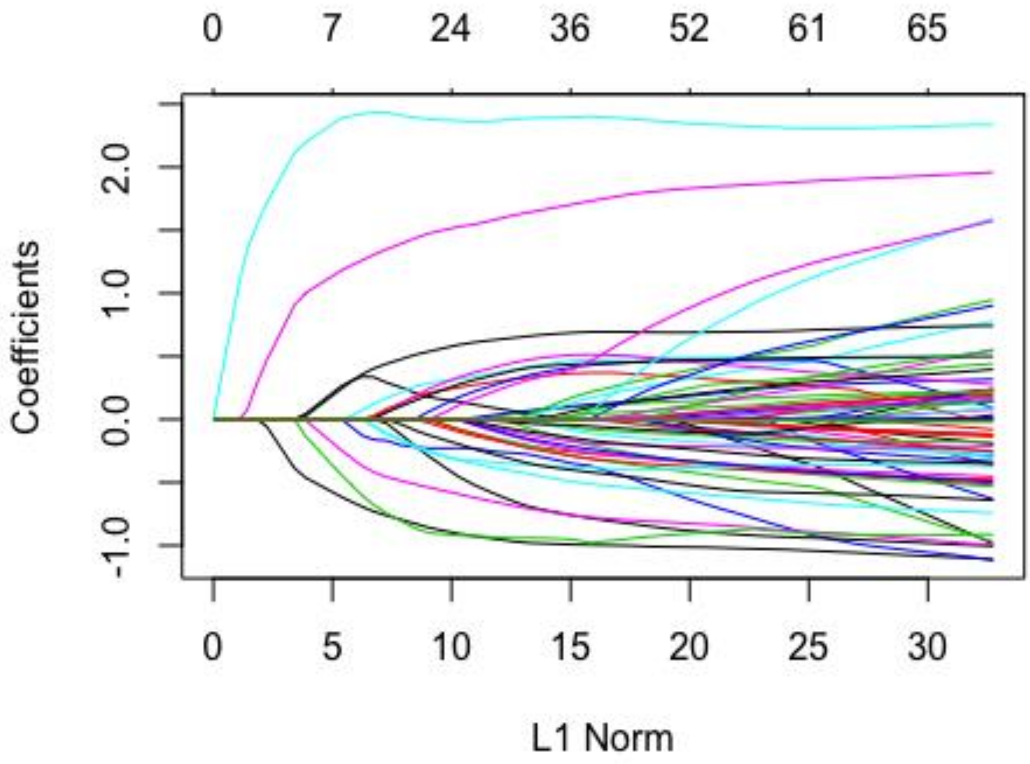
```
summary(back_aic_fit)
```

```
##
## Call:
## lm(formula = G3 ~ school + sex + age + Fjob + studytime + failures +
##     schoolsup + higher + internet + romantic + freetime + goout +
##     Dalc + health, data = portuguese_df[train, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3273  -1.4149   0.0634   1.4910   7.3709
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.5023     2.0962   2.625 0.008940 **
## schoolMS        -1.2845     0.2636  -4.873 1.49e-06 ***
## sexM            -0.4042     0.2616  -1.545 0.122881
## age              0.1981     0.1076   1.841 0.066227 .
## Fjobhealth      -0.5783     0.7799  -0.742 0.458700
## Fjobother       -0.5458     0.4845  -1.126 0.260516
## Fjobservices    -1.0156     0.5087  -1.996 0.046441 *
## Fjobteacher      0.7274     0.6589   1.104 0.270176
## studytime.L      0.7557     0.3841   1.967 0.049703 *
## studytime.Q     -0.1488     0.3439  -0.433 0.665356
## studytime.C     -0.2951     0.2726  -1.083 0.279484
## failures.L      -2.0656     0.6009  -3.437 0.000638 ***
## failures.Q       1.0290     0.5517   1.865 0.062765 .
## failures.C      -0.5571     0.5356  -1.040 0.298853
## schoolsupyes  -1.1486     0.3855  -2.979 0.003036 **
## higheryes      2.1704     0.4274   5.078 5.45e-07 ***
## internetyes    0.7489     0.2861   2.618 0.009119 **
## romanticyes   -0.5671     0.2494  -2.274 0.023423 *
## freetime.L    -0.5677     0.4062  -1.397 0.162913
## freetime.Q    -0.3064     0.3462  -0.885 0.376610
## freetime.C     0.1367     0.2957   0.462 0.644017
## freetime^4    -0.4958     0.2273  -2.182 0.029606 *
## goout.L        0.1059     0.3608   0.294 0.769219
## goout.Q       -0.9920     0.3153  -3.146 0.001756 **
## goout.C        0.3102     0.2771   1.120 0.263433
## goout^4       -0.1816     0.2348  -0.773 0.439720
## Dalc.L        -0.9628     0.6035  -1.595 0.111260
## Dalc.Q         0.9193     0.5496   1.673 0.095014 .
## Dalc.C         1.5708     0.5597   2.807 0.005205 **
## Dalc^4         1.5958     0.5024   3.176 0.001587 **
## health.L      -0.6748     0.2674  -2.524 0.011933 *
## health.Q       0.2298     0.2700   0.851 0.394996
## health.C      -0.2824     0.3008  -0.939 0.348294
## health^4      -0.1516     0.2904  -0.522 0.601873
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.585 on 486 degrees of freedom
## Multiple R-squared:  0.4057, Adjusted R-squared:  0.3654
## F-statistic: 10.05 on 33 and 486 DF,  p-value: < 2.2e-16
```

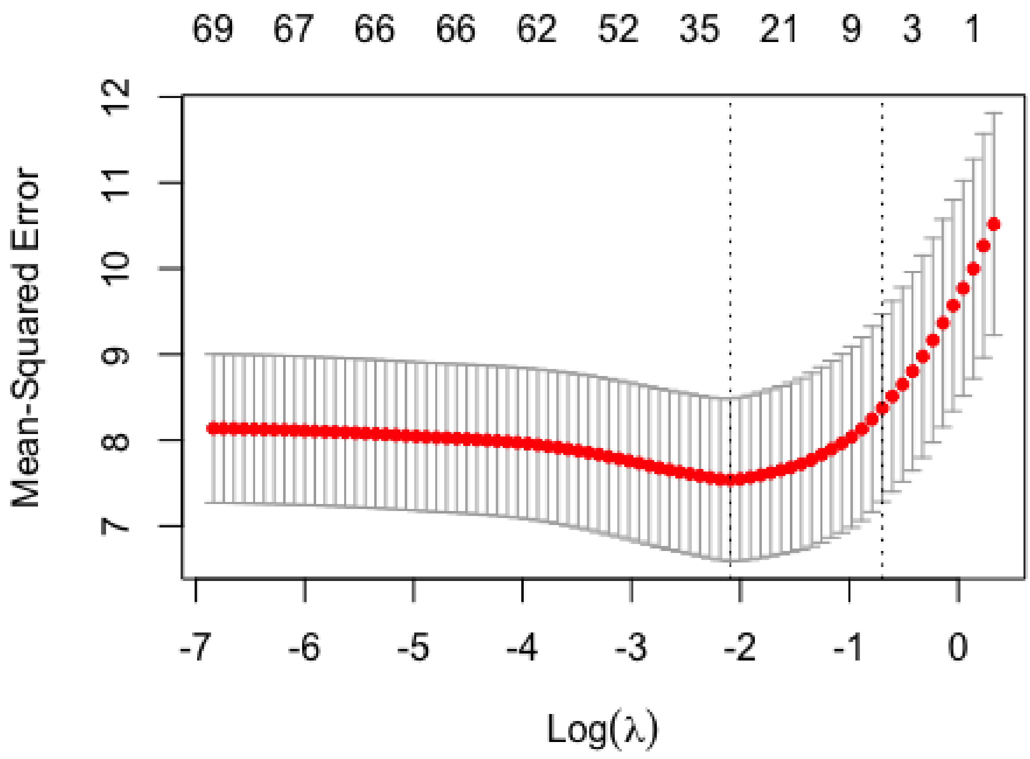
```
back_aic_pred = predict(back_aic_fit, newdata = portuguese_df[-train,1:30])
coef(back_aic_fit)
```

```
##  (Intercept)     schoolMS         sexM          age   Fjobhealth
Fjobother
##    5.5023211   -1.2845022   -0.4042356    0.1981111   -0.5783422   -
0.5457673
## Fjobservices  Fjobteacher  studytime.L  studytime.Q  studytime.C
failures.L
##   -1.0156191    0.7273533    0.7556995   -0.1488175   -0.2950899   -
2.0656223
##   failures.Q   failures.C schoolsupyes    higheryes  internetyes
romanticyes
##    1.0290190   -0.5570627   -1.1485673    2.1704024    0.7489057   -
0.5670927
##   freetime.L   freetime.Q   freetime.C   freetime^4      goout.L
goout.Q
##   -0.5676914   -0.3064003    0.1367181   -0.4958449    0.1059135   -
0.9919739
##      goout.C      goout^4       Dalc.L       Dalc.Q       Dalc.C
Dalc^4
##    0.3102113   -0.1815786   -0.9628167    0.9193336    1.5708497
1.5958420
```

```
##     health.L     health.Q     health.C     health^4
##   -0.6747846    0.2298474   -0.2824263   -0.1515843

mean((back_aic_pred-test_g3)^2)

## [1] 7.589608

# Lasso Regression
library(glmnet)
x_train = model.matrix(G3~., portuguese_df[train,])[,-1]
x_test = model.matrix(G3~., portuguese_df[-train,])[,-1]

y_train = portuguese_df[train,] %>% dplyr::select(G3) %>% unlist() %>%
as.numeric()
y_test = portuguese_df[-train,] %>% dplyr::select(G3) %>% unlist() %>%
as.numeric()

lasso_fit_1 = glmnet(x_train, y_train, alpha = 1)

plot(lasso_fit_1)
```



```
set.seed(1)
cv.out = cv.glmnet(x_train, y_train, alpha = 1)
plot(cv.out)
```

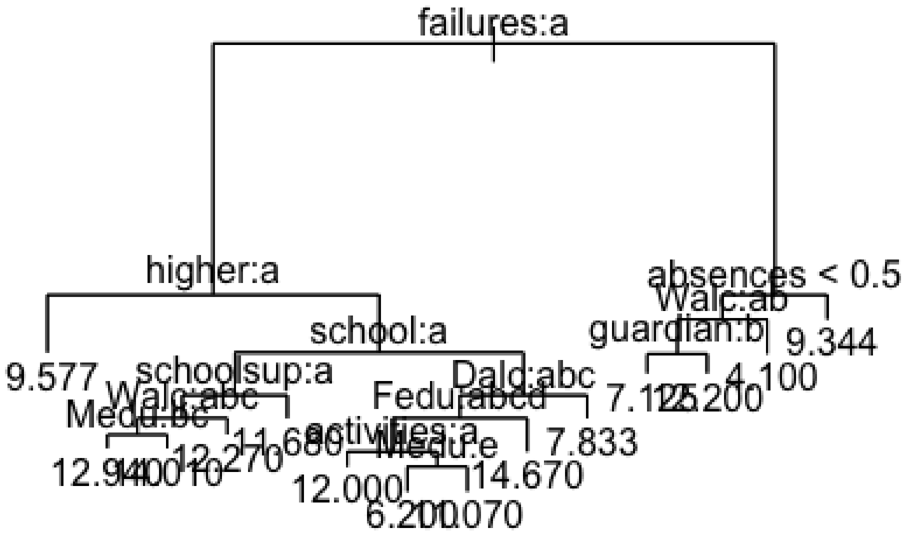


```
bstlambda = cv.out$lambda.min
lasso_pred = predict(lasso_fit_1, s = bstlambda, newx = x_test)
mean((lasso_pred - y_test)^2)

## [1] 6.958149

lasso_bst_fit <- glmnet(x_train, y_train, alpha = 1, lambda = bstlambda)
coef(lasso_bst_fit)

## 72 x 1 sparse Matrix of class "dgCMatrix"
##                            s0
## (Intercept)       8.47327806
## schoolMS         -0.96967226
## sexM             -0.19614891
## age               0.04227054
## addressU          0.05731473
## famsizeLE3        .
## PstatusT          .
## Medu.L            0.06180214
## Medu.Q            0.32120217
## Medu.C            .
## Medu^4            .
## Fedu.L            0.39811260
## Fedu.Q            .
## Fedu.C            .
## Fedu^4            .
## Mjobhealth        .
## Mjobother         .
## Mjobservices      .
## Mjobteacher       .
## Fjobhealth        .
## Fjobother         .
## Fjobservices     -0.18286017
## Fjobteacher       0.31955527
## reasonhome        .
## reasonother      -0.13911050
## reasonreputation  .
## guardianmother    .
## guardianother     .
## traveltime.L      .
## traveltime.Q      .
## traveltime.C      .
## traveltime^4      .
## studytime.L       0.64785306
## studytime.Q       .
## studytime.C       .
## failures.L       -0.24318670
## failures.Q        2.39282563
## failures.C        .
## failures^4        0.44999141
## schoolsupyes     -0.65519965
## famsupyes         .
## paidyes          -0.03381896
## activitiesyes     .
## nurseryyes        .
## higheryes         1.62105082
## internetyes       0.38864511
## romanticyes      -0.20264897
## famrel.L          .
## famrel.Q          .
## famrel.C         -0.35170883
## famrel^4         -0.10258335
## freetime.L       -0.32280454
## freetime.Q        .
## freetime.C        .
## freetime^4       -0.12897210
## goout.L           .
## goout.Q          -0.68723743
## goout.C           .
## goout^4           .
## Dalc.L           -0.94018931
## Dalc.Q            .
## Dalc.C            .
## Dalc^4            0.28074583
## Walc.L           -0.02383276
## Walc.Q            .
## Walc.C            .
## Walc^4            .
## health.L         -0.41325852
## health.Q          .
## health.C         -0.08076236
## health^4          .
## absences          .
```

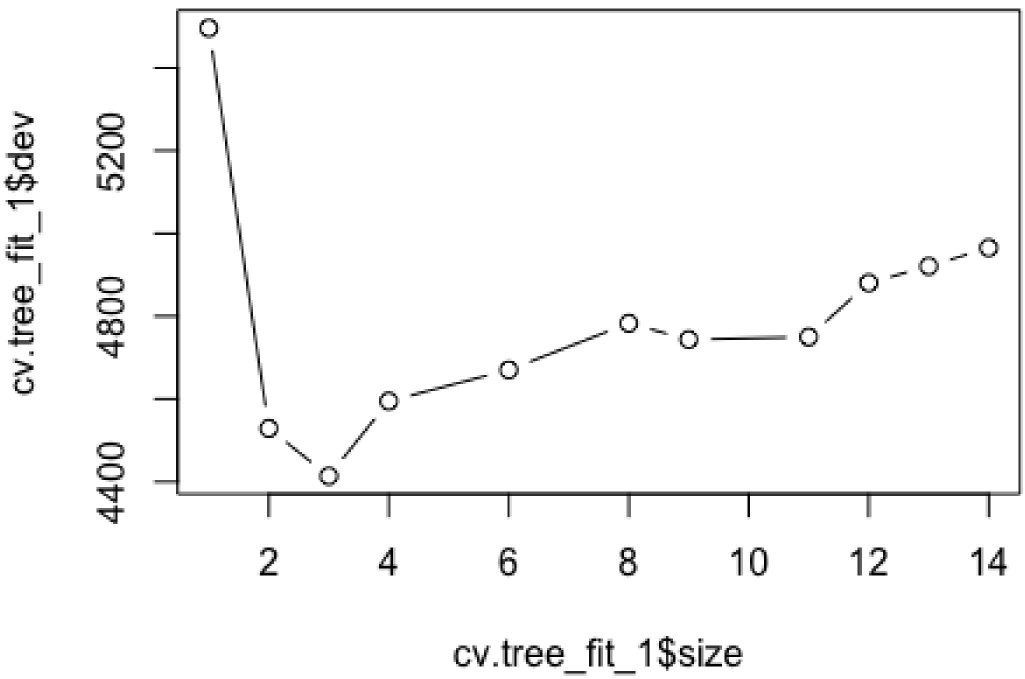
```
############### TREES ##############

library(ISLR)
library(tree)
library(MASS)

tree_fit_1 = tree(G3~., data = portuguese_df , subset = train)
summary(tree_fit_1)
```

```
##
## Regression tree:
## tree(formula = G3 ~ ., data = portuguese_df, subset = train)
## Variables actually used in tree construction:
##  [1] "failures"   "higher"     "school"     "schoolsup"  "Walc"
##  [6] "Medu"       "Dalc"       "Fedu"       "activities" "absences"
## [11] "guardian"
## Number of terminal nodes:  14
## Residual mean deviance:  6.161 = 3118 / 506
## Distribution of residuals:
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -11.070000  -1.577000  -0.008475   0.000000   1.423000   6.800000
```

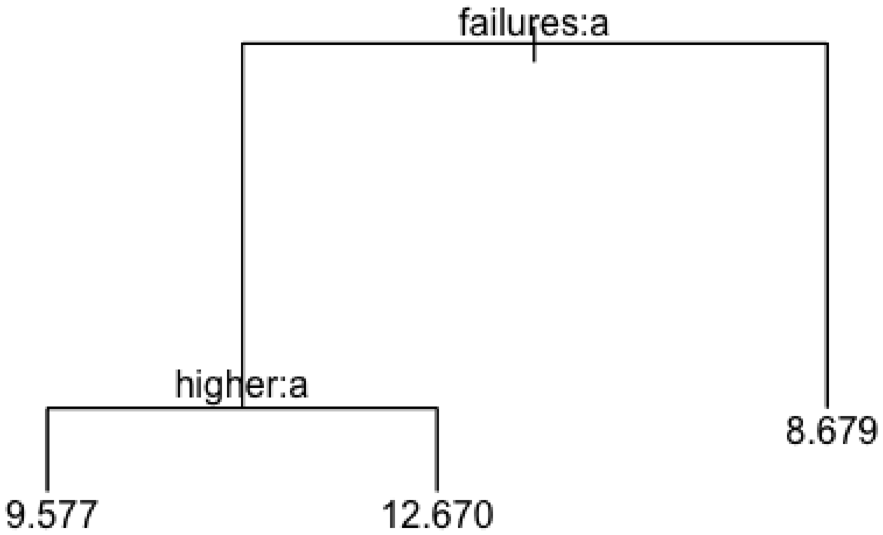
```
plot(tree_fit_1)
text(tree_fit_1)
```



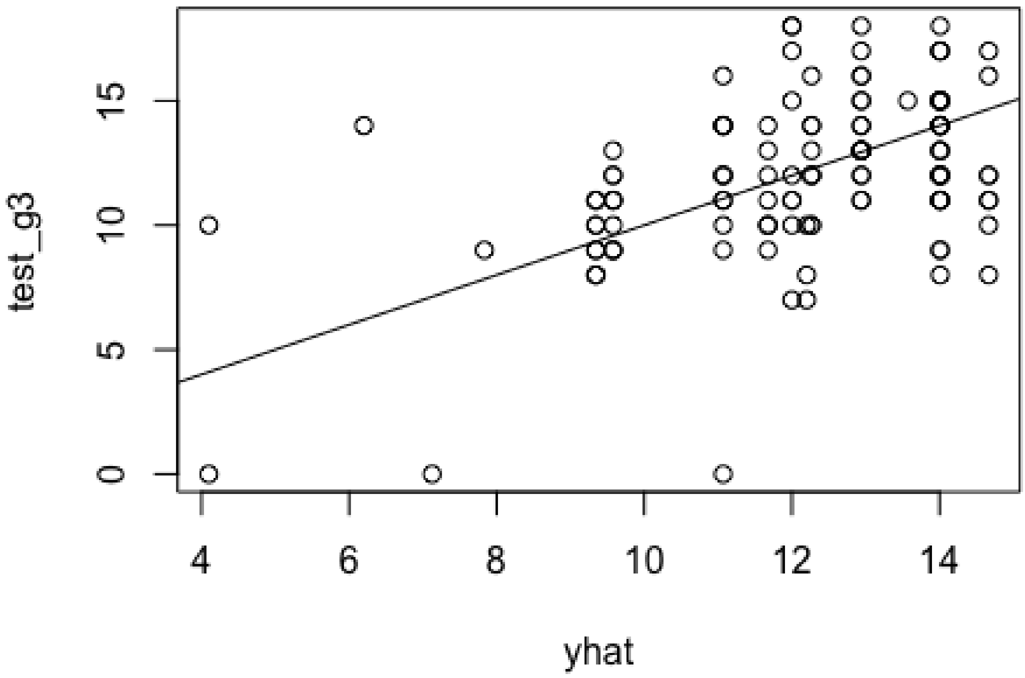
```r
cv.tree_fit_1 = cv.tree(tree_fit_1)
plot(cv.tree_fit_1$size, cv.tree_fit_1$dev, type = 'b')
```



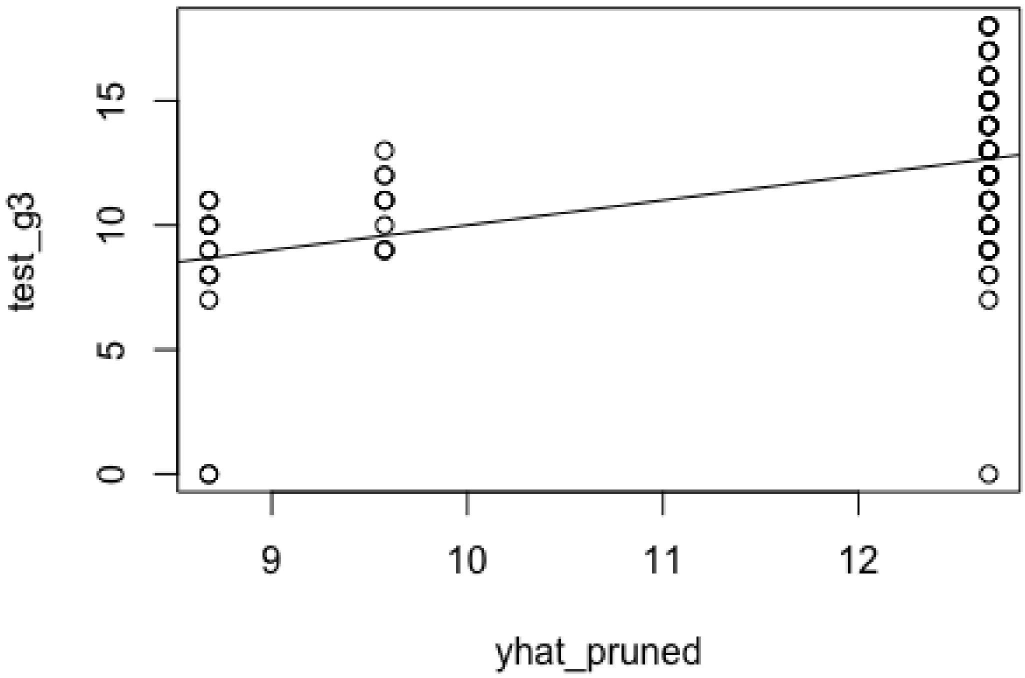
```
prune.tree_fit_1 = prune.tree(tree_fit_1, best = 3)
plot(prune.tree_fit_1)
text(prune.tree_fit_1)
```



```r
yhat = predict(tree_fit_1, newdata = portuguese_df[-train,1:30])
plot(yhat, test_g3)
abline(0,1)
```



```
mean((yhat-test_g3)^2)

## [1] 8.077376

yhat_pruned = predict(prune.tree_fit_1, newdata = portuguese_df[-train,1:30])
plot(yhat_pruned, test_g3)
abline(0,1)
```



```
mean((yhat_pruned-test_g3)^2)

## [1] 7.670547

############## RANDOM FOREST ##############

library(randomForest)

## randomForest 4.6-14

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

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

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

# Bagged DT : m = p predictors i.e. mtry = 30
set.seed(-1)
```

```
bagged_tree_fit = randomForest(G3~., data = portuguese_df[train,], mtry = 30,
ntree= 1000, importance = TRUE)
bagged_tree_fit
```

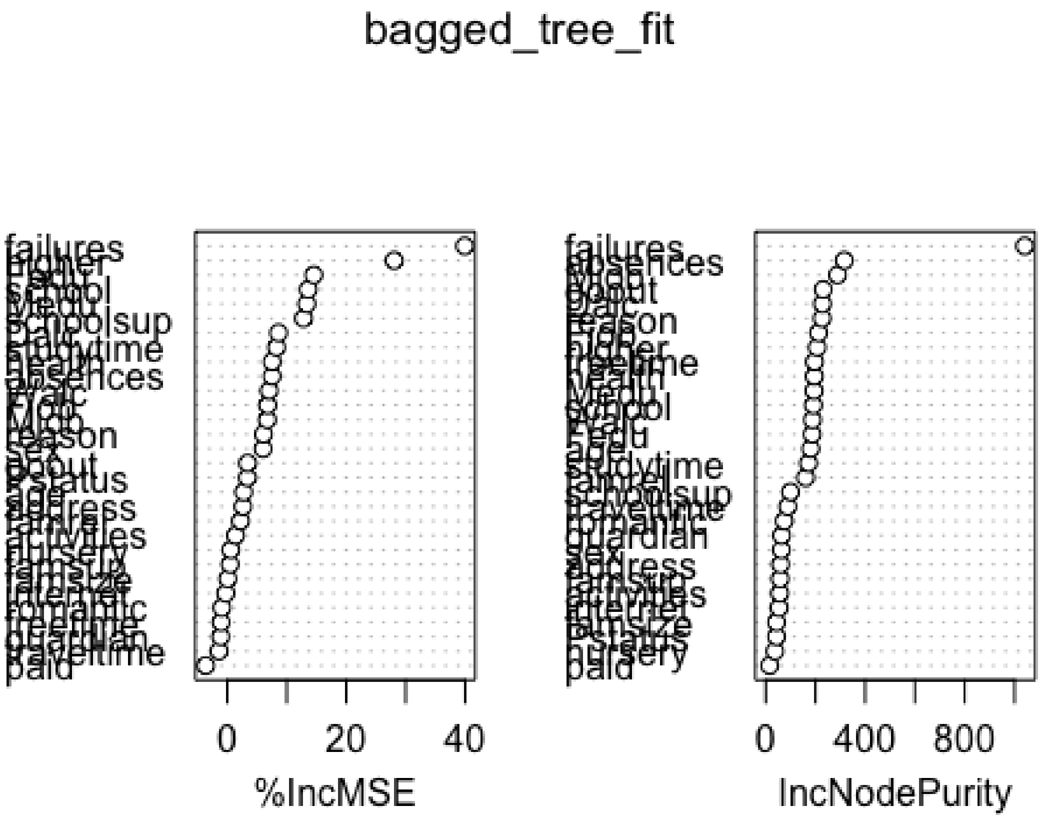
```
##
## Call:
##  randomForest(formula = G3 ~ ., data = portuguese_df[train, ],      mtry =
30, ntree = 1000, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 30
##
##           Mean of squared residuals: 7.682148
##                     % Var explained: 26.91
```

```
yhat_bagged_tree_fit = predict(bagged_tree_fit, newdata = portuguese_df[-
train,1:30])
bagging_test = portuguese_df[-train,"G3"]
mean((yhat_bagged_tree_fit-bagging_test)^2)
```

```
## [1] 6.844891
```

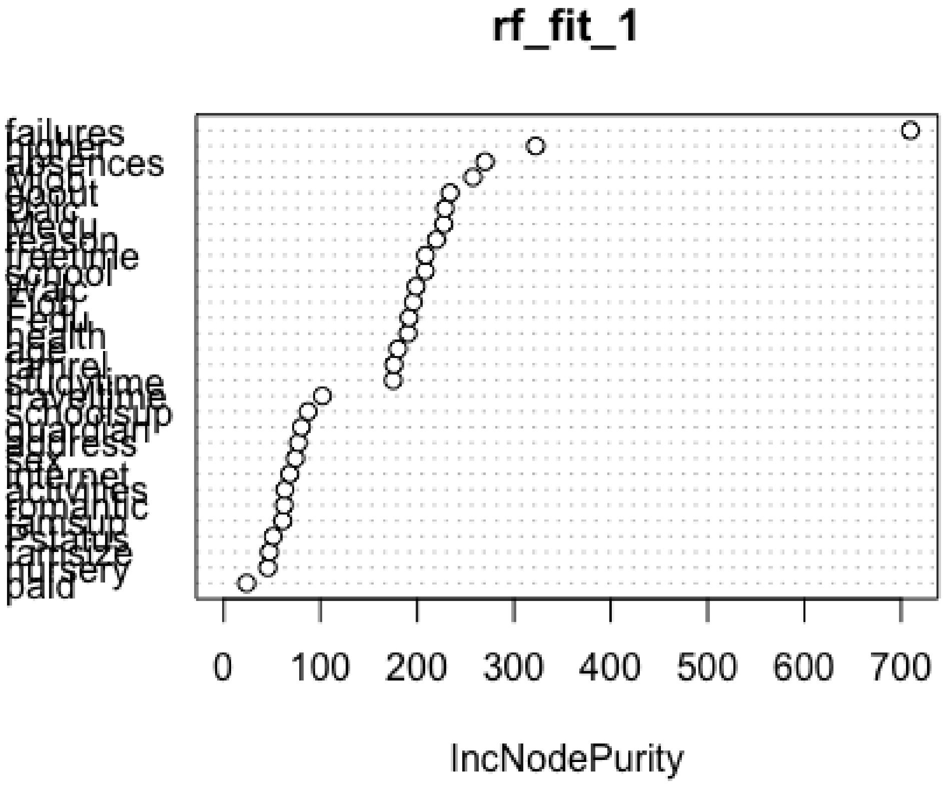
```
importance(bagged_tree_fit)
```

```
##                %IncMSE IncNodePurity
## school     13.61854253     191.57193
## sex         6.03376231      63.25383
## age         2.76914322     181.26489
## address     2.64766862      59.80604
## famsize     0.09636268      43.80555
## Pstatus     3.41834860      43.66349
## Medu       13.39761080     194.45934
## Fedu       14.56729055     186.77956
## Mjob        6.74254116     288.84729
## Fjob        6.79700298     212.65125
## reason      6.09289477     226.95045
## guardian   -1.10087757      64.20830
## traveltime -1.37474894      88.90280
## studytime   8.30495204     171.93630
## failures   39.96175442    1041.57370
## schoolsup  12.84098947      99.87588
## famsup      0.48519301      58.27574
## paid       -3.62527644      16.16196
## activities  1.39746919      56.50043
## nursery     0.60445260      37.08106
## higher     28.09381933     207.36934
## internet   -0.23422326      53.96571
## romantic   -0.89888678      74.30723
## famrel      2.23307245     160.90655
## freetime   -1.03701035     196.17786
## goout       3.43920608     231.35156
## Dalc        8.63629101     228.05784
## Walc        7.05203087     188.29721
## health      7.55977126     195.46607
## absences    7.55176017     315.98248

varImpPlot(bagged_tree_fit)
```



```
# Random Forest - that is with m != p, mtry = p/3 (optimal for regression
trees)
set.seed(-1)
rf_fit_1 = randomForest(G3~., data = portuguese_df[train,], mtry = 10, ntree=
1000, importance = FALSE)
rf_fit_1

##
## Call:
##  randomForest(formula = G3 ~ ., data = portuguese_df[train, ],      mtry =
10, ntree = 1000, importance = FALSE)
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 7.281916
##                     % Var explained: 30.72
```

```r
yhat_rf_fit_1 = predict(rf_fit_1, newdata = portuguese_df[-train,])
bagging_test = portuguese_df[-train,"G3"]
mean((yhat_rf_fit_1-bagging_test)^2)

## [1] 6.849848

varImpPlot(rf_fit_1)
```



```r
###################################  Model Performance
################################

cat("RMSE of Backward Step wise : ", sqrt(mean((back_aic_pred-
test_g3)^2)),"\n")

## RMSE of Backward Step wise :  2.754924

cat("RMSE of Lasso : ", sqrt(mean((lasso_pred - y_test)^2)),"\n")

## RMSE of Lasso :  2.63783

cat("RMSE of Decision Tree : ", sqrt(mean((yhat_pruned-test_g3)^2)),"\n")

## RMSE of Decision Tree :  2.769575
```

```
cat("RMSE of Bagged Decision Trees : ", sqrt(mean((yhat_bagged_tree_fit-
bagging_test)^2)),"\n")
```

```
## RMSE of Bagged Decision Trees :  2.616274
```

```
cat("RMSE of RF : ", sqrt(mean((yhat_rf_fit_1-bagging_test)^2)),"\n")
```

```
## RMSE of RF :  2.617221
```

```
##############################################################################
##########

############################
## Mathematics Performance Analysis
############################

# Quick glance at Data

table(school1$school)
```

```
## 
##  GP  MS 
## 349  46
```

```
head(school1)
```

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

```
colnames(school1)
```

```
##  [1] "school"     "sex"        "age"        "address"    "famsize"
##  [6] "Pstatus"    "Medu"       "Fedu"       "Mjob"       "Fjob"
## [11] "reason"     "guardian"   "traveltime" "studytime"  "failures"
## [16] "schoolsup"  "famsup"     "paid"       "activities" "nursery"
## [21] "higher"     "internet"   "romantic"   "famrel"     "freetime"
## [26] "goout"      "Dalc"       "Walc"       "health"     "absences"
## [31] "G1"         "G2"         "G3"
```

```
summary(school1)
```

```
##  school   sex          age        address famsize   Pstatus      Medu
##  GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   Min.   :0.000
##  MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1st Qu.:2.000
##                   Median :17.0                             Median :3.000
##                   Mean   :16.7                             Mean   :2.749
##                   3rd Qu.:18.0                             3rd Qu.:4.000
##                   Max.   :22.0                             Max.   :4.000
##       Fedu             Mjob           Fjob            reason      guardian
##  Min.   :0.000   at_home : 59   at_home : 20   course    :145   father: 90
##  1st Qu.:2.000   health  : 34   health  : 18   home      :109   mother:273
##  Median :2.000   other   :141   other   :217   other     : 36   other : 32
##  Mean   :2.522   services:103   services:111   reputation:105
##  3rd Qu.:3.000   teacher : 58   teacher : 29
##  Max.   :4.000
##    traveltime      studytime        failures      schoolsup famsup
paid
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344   no :153   no
:214
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51   yes:242
yes:181
##  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
##  activities nursery   higher    internet  romantic      famrel
##  no :194    no : 81   no : 20   no : 66   no :263   Min.   :1.000
##  yes:201    yes:314   yes:375   yes:329   yes:132   1st Qu.:4.000
##                                                     Median :4.000
##                                                      Mean   :3.944
##                                                      3rd Qu.:5.000
##                                                      Max.   :5.000
##     freetime         goout            Dalc            Walc
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median :3.000   Median :1.000   Median :2.000
##  Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##      health          absences           G1              G2
##  Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00
##  1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00
##  Median :4.000   Median : 4.000   Median :11.00   Median :11.00
##  Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71
##  3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00
##  Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00
##        G3
##  Min.   : 0.00
##  1st Qu.: 8.00
##  Median :11.00
##  Mean   :10.42
##  3rd Qu.:14.00
##  Max.   :20.00
```

```
############################
## Data Preparation ########
############################

any(is.na(school1))
```

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

```
# There are no missing values in the data set.

# dropping G1 and G2 as they are highly corelated to G3
mathematics_df = subset(school1, select = -c(G1,G2))
colnames(mathematics_df)
```

```
##  [1] "school"     "sex"        "age"        "address"    "famsize"
##  [6] "Pstatus"    "Medu"       "Fedu"       "Mjob"       "Fjob"
## [11] "reason"     "guardian"   "traveltime" "studytime"  "failures"
## [16] "schoolsup"  "famsup"     "paid"       "activities" "nursery"
## [21] "higher"     "internet"   "romantic"   "famrel"     "freetime"
## [26] "goout"      "Dalc"       "Walc"       "health"     "absences"
## [31] "G3"
```

```
glimpse(mathematics_df)
```

```
## Observations: 395
## Variables: 31
## Registered S3 method overwritten by 'cli':
##   method     from
##   print.tree tree
```

```
## $ school     <fct> GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP,
GP…
## $ sex        <fct> F, F, F, F, F, M, M, F, M, M, F, F, M, M, M, F, F, F,
M, M…
## $ age        <int> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15,
15…
## $ address    <fct> U, U, U, U, U, U, U, U, U, U, U, U, U, U, U, U, U, U,
U, U…
## $ famsize    <fct> GT3, GT3, LE3, GT3, GT3, LE3, LE3, GT3, LE3, GT3, GT3,
GT3…
## $ Pstatus    <fct> A, T, T, T, T, T, T, A, A, T, T, T, T, T, A, T, T, T,
T, T…
## $ Medu       <int> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3,
3, 4…
## $ Fedu       <int> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3,
2, 3…
## $ Mjob       <fct> at_home, at_home, at_home, health, other, services,
other,…
## $ Fjob       <fct> teacher, other, other, services, other, other, other,
teac…
## $ reason     <fct> course, course, other, home, home, reputation, home,
home,…
## $ guardian   <fct> mother, father, mother, mother, father, mother, mother,
mo…
## $ traveltime <int> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3,
1, 1…
## $ studytime  <int> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2,
1, 1…
## $ failures   <int> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
3, 0…
## $ schoolsup  <fct> yes, no, yes, no, no, no, no, yes, no, no, no, no, no,
no,…
## $ famsup     <fct> no, yes, no, yes, yes, yes, no, yes, yes, yes, yes,
yes, y…
## $ paid       <fct> no, no, yes, yes, yes, yes, no, no, yes, yes, yes, no,
yes…
## $ activities <fct> no, no, no, yes, no, yes, no, no, no, yes, no, yes,
yes, n…
## $ nursery    <fct> yes, no, yes, yes, yes, yes, yes, yes, yes, yes, yes,
yes,…
## $ higher     <fct> yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes,
yes…
## $ internet   <fct> no, yes, yes, yes, no, yes, yes, no, yes, yes, yes,
yes, y…
## $ romantic   <fct> no, no, no, yes, no, no, no, no, no, no, no, no, no,
no, y…
```

```
## $ famrel     <int> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5,
5, 3…
## $ freetime   <int> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3,
5, 1…
## $ goout      <int> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2,
5, 3…
## $ Dalc       <int> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
2, 1…
## $ Walc       <int> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1,
4, 3…
## $ health     <int> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4,
5, 5…
## $ absences   <int> 6, 4, 10, 2, 4, 10, 0, 6, 0, 0, 0, 4, 2, 2, 0, 4, 6, 4,
16…
## $ G3         <int> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16,
14…
```

```
# The following variables need to be converted to categorical type:

# Medu - denotes Mother's eductaion
mathematics_df$Medu = factor(mathematics_df$Medu,
levels=c("0","1","2","3","4"), ordered=TRUE)
summary(mathematics_df$Medu)
```

```
##   0   1   2   3   4 
##   3  59 103  99 131
```

```
# Fedu - denotes Father's eductaion
mathematics_df$Fedu = factor(mathematics_df$Fedu,
levels=c("0","1","2","3","4"), ordered=TRUE)
summary(mathematics_df$Fedu)
```

```
##   0   1   2   3   4 
##   2  82 115 100  96
```

```
# famrel - quality of family relationships
mathematics_df$famrel = factor(mathematics_df$famrel, levels=1:5,
ordered=TRUE)
summary(mathematics_df$famrel)
```

```
##   1   2   3   4   5 
##   8  18  68 195 106
```

```
# traveltime - home to school travel time
mathematics_df$traveltime = factor(mathematics_df$traveltime, levels=0:4,
ordered=TRUE)
summary(mathematics_df$traveltime)
```

```
##   0   1   2   3   4 
##   0 257 107  23   8
```

```r
# studytime - weekly study time
mathematics_df$studytime = factor(mathematics_df$studytime, levels=1:4,
ordered=TRUE)
summary(mathematics_df$studytime)
```

```
##   1   2   3   4 
## 105 198  65  27 
```

```r
# freetime - free time after school
mathematics_df$freetime = factor(mathematics_df$freetime, levels=1:5,
ordered=TRUE)
summary(mathematics_df$freetime)
```

```
##   1   2   3   4   5 
##  19  64 157 115  40 
```

```r
# goout - going out with friends
mathematics_df$goout = factor(mathematics_df$goout, levels=1:5, ordered=TRUE)
summary(mathematics_df$goout)
```

```
##   1   2   3   4   5 
##  23 103 130  86  53 
```

```r
# Dalc - workday alcohol consumption
mathematics_df$Dalc = factor(mathematics_df$Dalc, levels=1:5, ordered=TRUE)
summary(mathematics_df$Dalc)
```

```
##   1   2   3   4   5 
## 276  75  26   9   9 
```

```r
# Walc - weekend alcohol consumption
mathematics_df$Walc = factor(mathematics_df$Walc, levels=1:5, ordered=TRUE)
summary(mathematics_df$Walc)
```

```
##   1   2   3   4   5 
## 151  85  80  51  28 
```

```r
# health - current health status
mathematics_df$health = factor(mathematics_df$health, levels=1:5,
ordered=TRUE)
summary(mathematics_df$health)
```

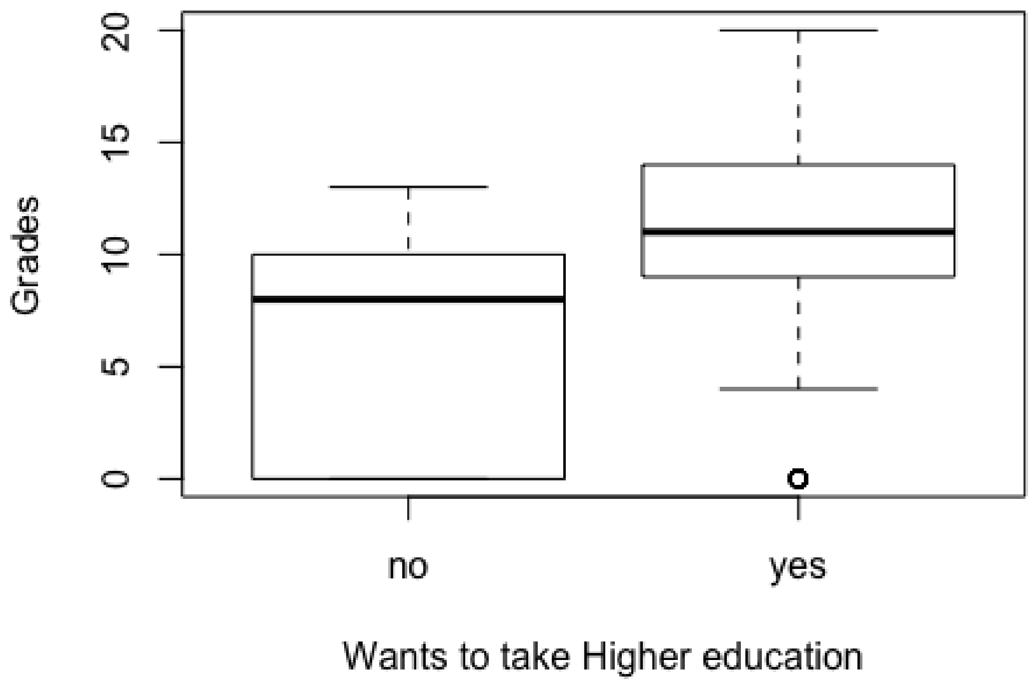
```
##   1   2   3   4   5 
##  47  45  91  66 146 
```

```r
# failures - number of past class failures
mathematics_df$failures = factor(mathematics_df$failures, levels=0:4,
ordered=TRUE)
summary(mathematics_df$failures)
```

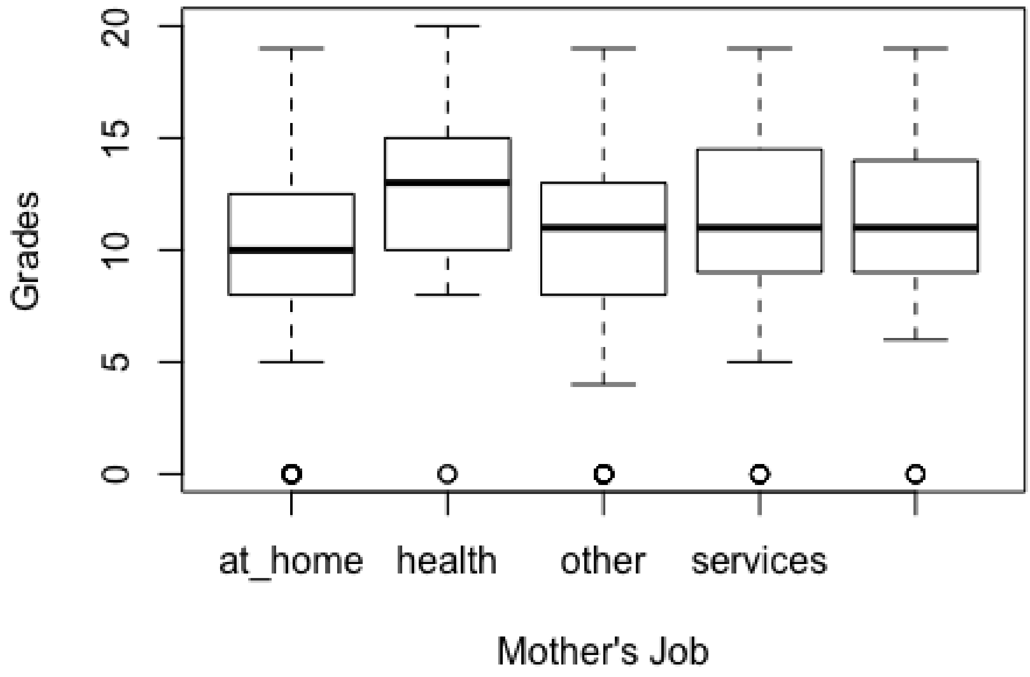
```
##   0   1   2   3   4 
## 312  50  17  16   0 
```

```
summary(mathematics_df)

##  school   sex          age       address famsize   Pstatus Medu    Fedu
##  GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   0:  3   0:  2
##  MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1: 59   1: 82
##                   Median :17.0                             2:103   2:115
##                   Mean   :16.7                             3: 99   3:100
##                   3rd Qu.:18.0                             4:131   4: 96
##                   Max.   :22.0
##        Mjob           Fjob            reason      guardian   traveltime
##  at_home : 59   at_home : 20   course     :145   father: 90   0:  0
##  health  : 34   health  : 18   home       :109   mother:273   1:257
##  other   :141   other   :217   other      : 36   other : 32   2:107
##  services:103   services:111   reputation:105                3: 23
##  teacher : 58   teacher : 29                                 4:  8
##
##  studytime failures schoolsup famsup     paid     activities nursery
##  1:105     0:312    no :344   no :153   no :214   no :194    no : 81
##  2:198     1: 50    yes: 51   yes:242   yes:181   yes:201    yes:314
##  3: 65     2: 17
##  4: 27     3: 16
##            4:  0
##
##  higher    internet  romantic  famrel   freetime goout   Dalc    Walc
health
##  no : 20   no : 66   no :263   1:  8   1: 19    1: 23   1:276   1:151   1:
47
##  yes:375   yes:329   yes:132   2: 18   2: 64    2:103   2: 75   2: 85   2:
45
##                                3: 68   3:157    3:130   3: 26   3: 80   3:
91
##                                4:195   4:115    4: 86   4:  9   4: 51   4:
66
##                                5:106   5: 40    5: 53   5:  9   5: 28
5:146
##
##     absences            G3
##  Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.: 8.00
##  Median : 4.000   Median :11.00
##  Mean   : 5.709   Mean   :10.42
##  3rd Qu.: 8.000   3rd Qu.:14.00
##  Max.   :75.000   Max.   :20.00

########################### Exploratory Data Analysis(EDA)
###########################

# Creating box-plots for categorical data
suppressMessages(attach(mathematics_df))
```

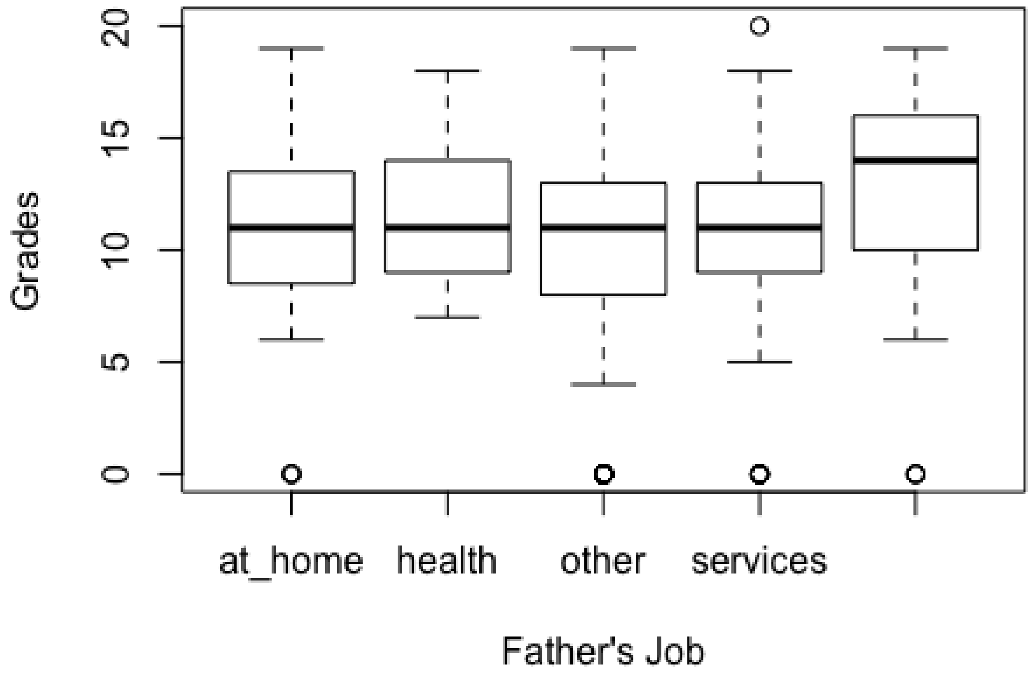
```
plot(higher,G3, xlab = "Wants to take Higher education", ylab = "Grades")
```



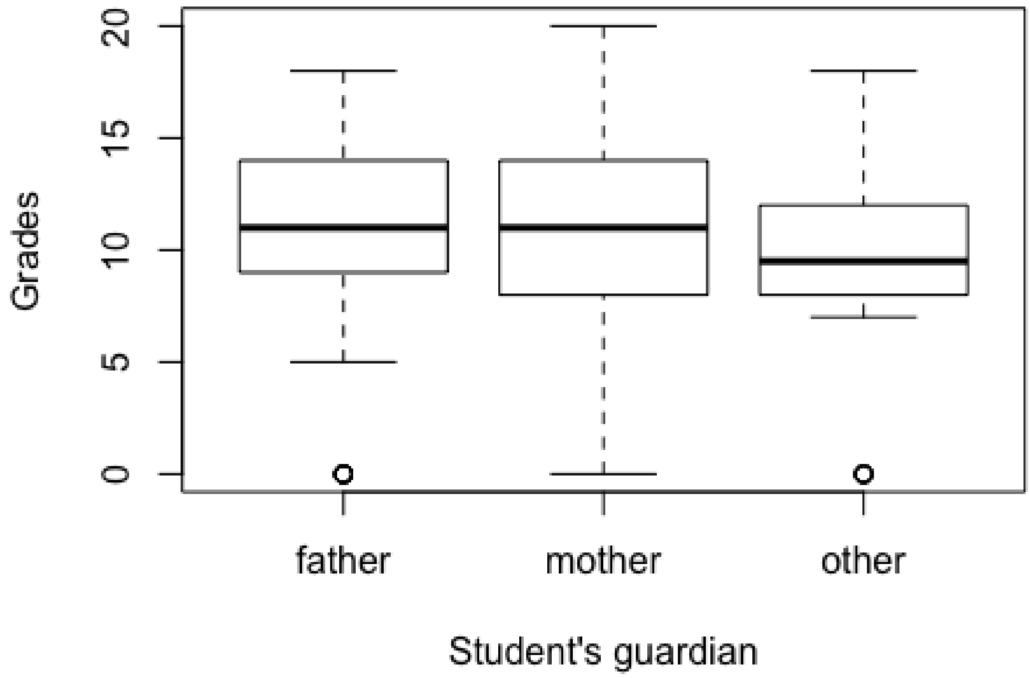
```
plot(Mjob,G3, xlab = "Mother's Job", ylab = "Grades")
```



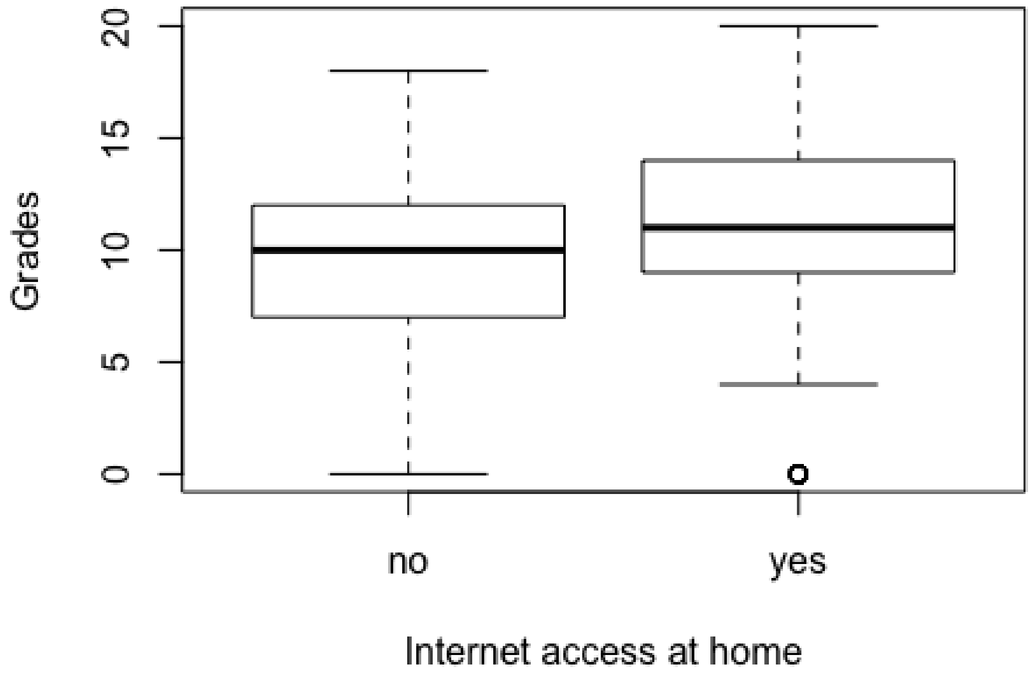
```r
plot(Fjob, G3, xlab = "Father's Job", ylab = "Grades")
```



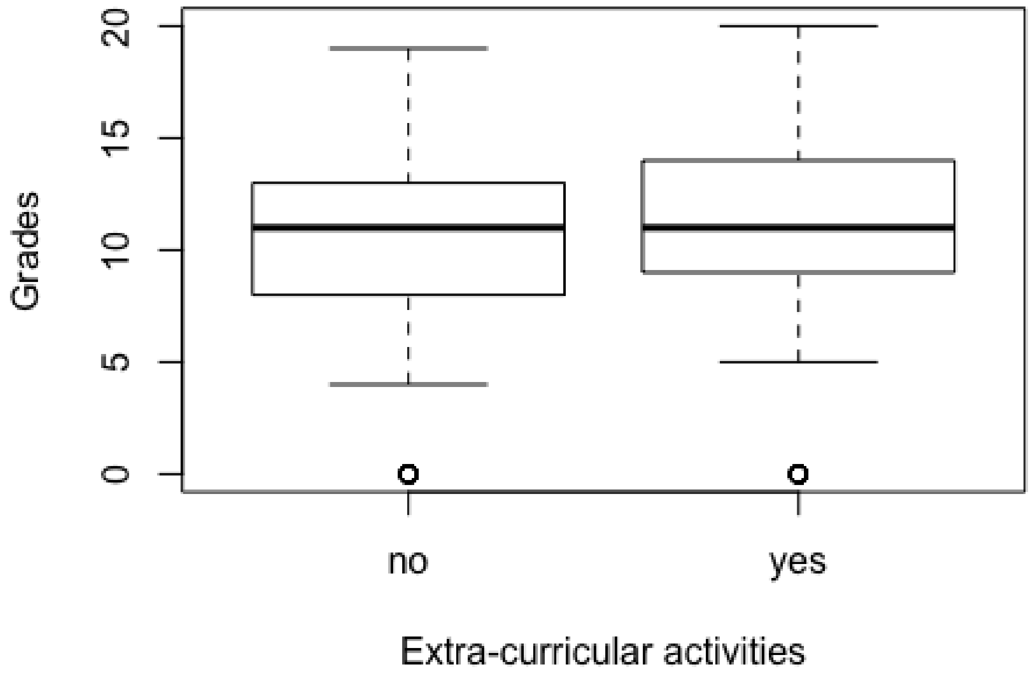
```
plot(guardian,G3, xlab = "Student's guardian", ylab = "Grades")
```



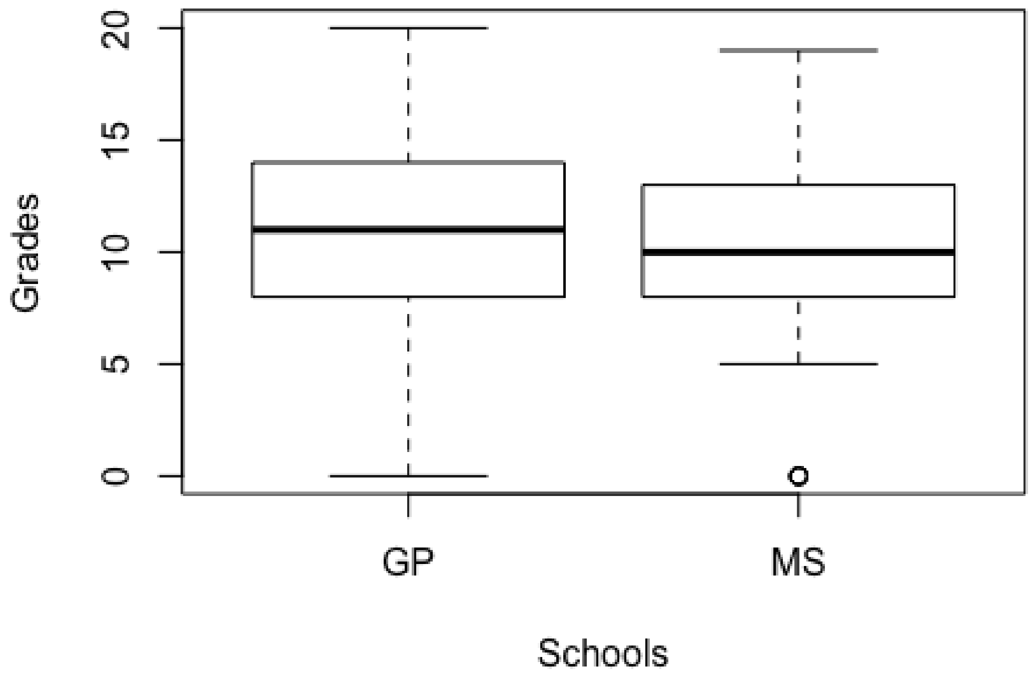
```r
plot(internet,G3, xlab = "Internet access at home", ylab = "Grades")
```



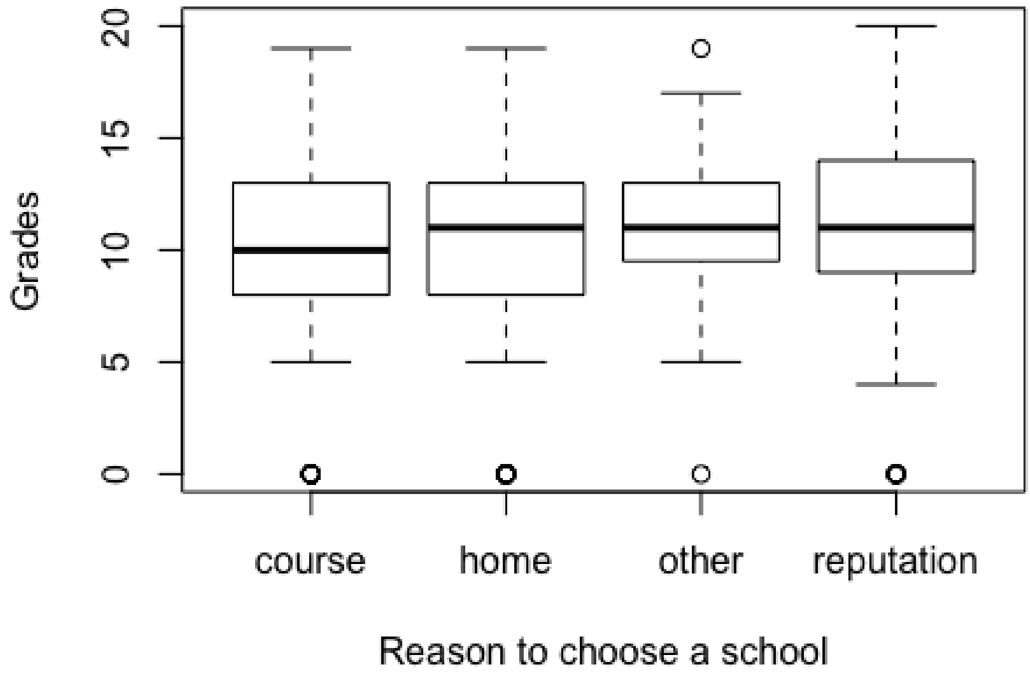
```
plot(activities,G3, xlab = "Extra-curricular activities", ylab = "Grades")
```



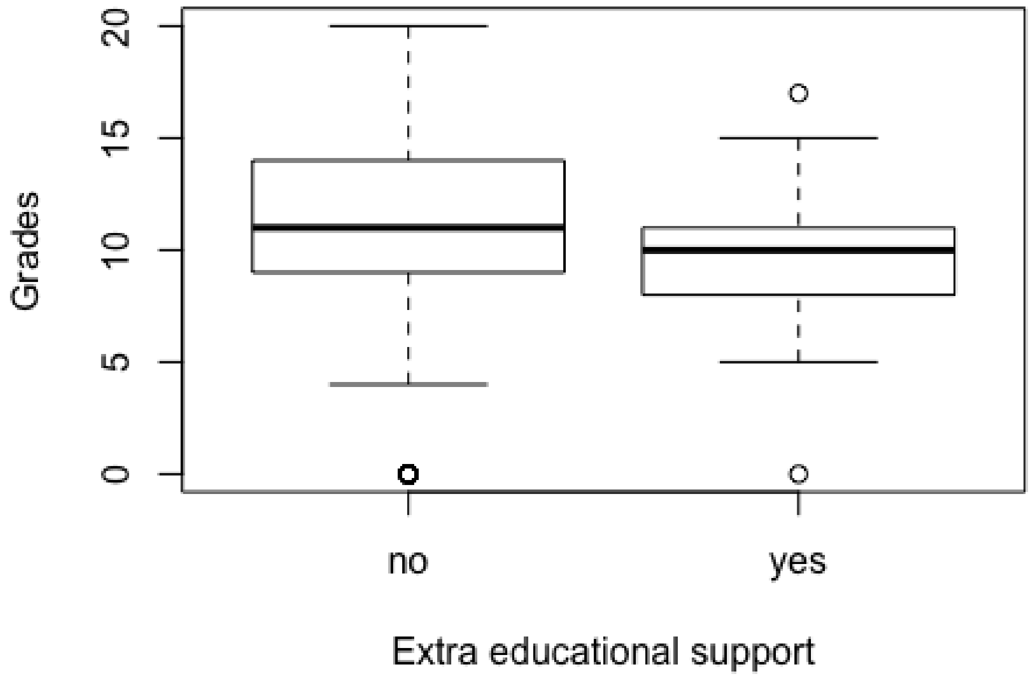
```
plot(school, G3, xlab = "Schools", ylab = "Grades")
```



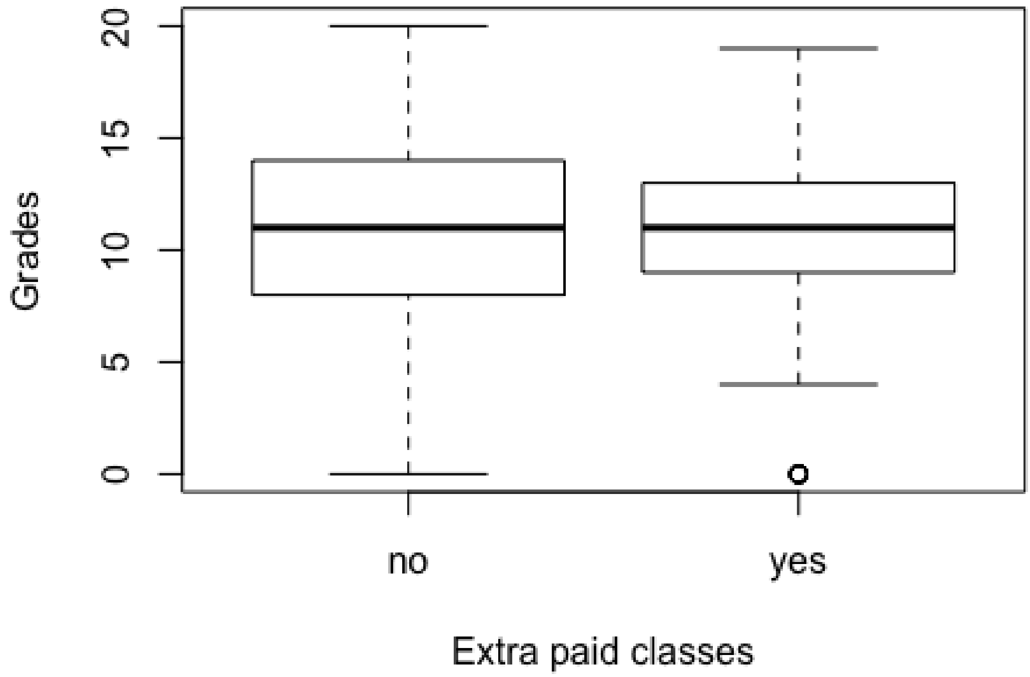
```
plot(reason,G3, xlab = "Reason to choose a school", ylab = "Grades")
```



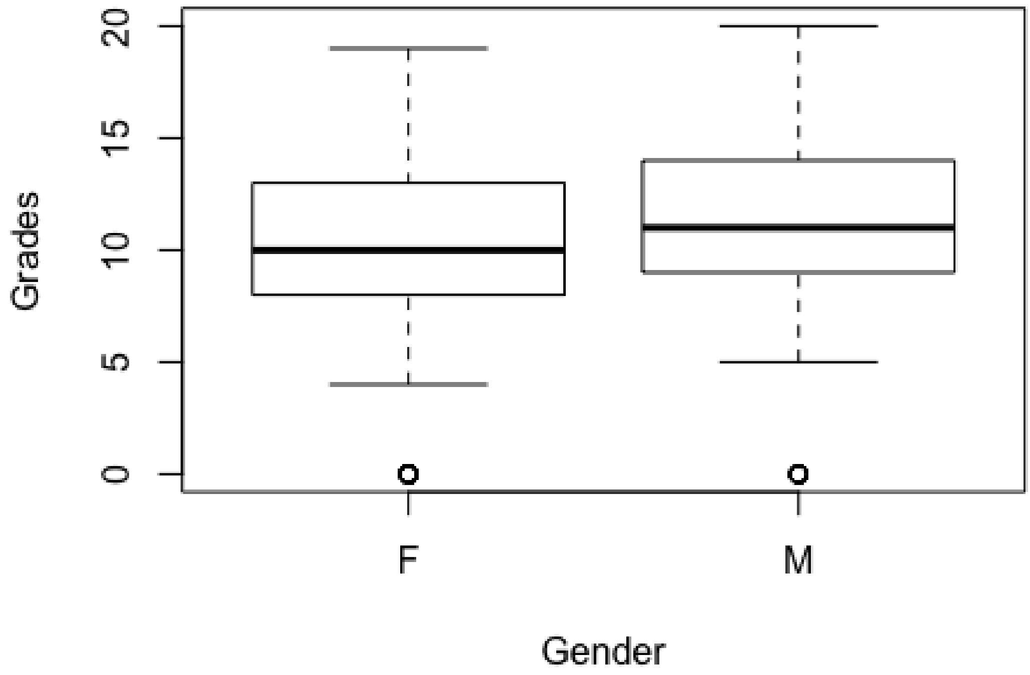
```r
plot(schoolsup,G3, xlab = "Extra educational support", ylab = "Grades")
```



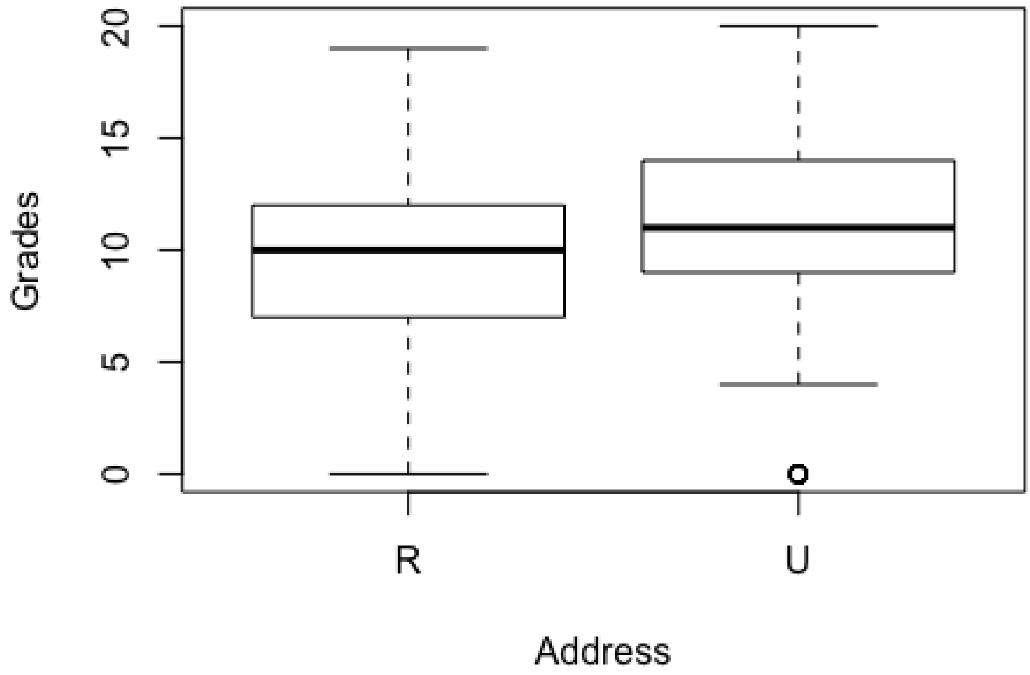
```r
plot(paid, G3, xlab = "Extra paid classes", ylab = "Grades")
```



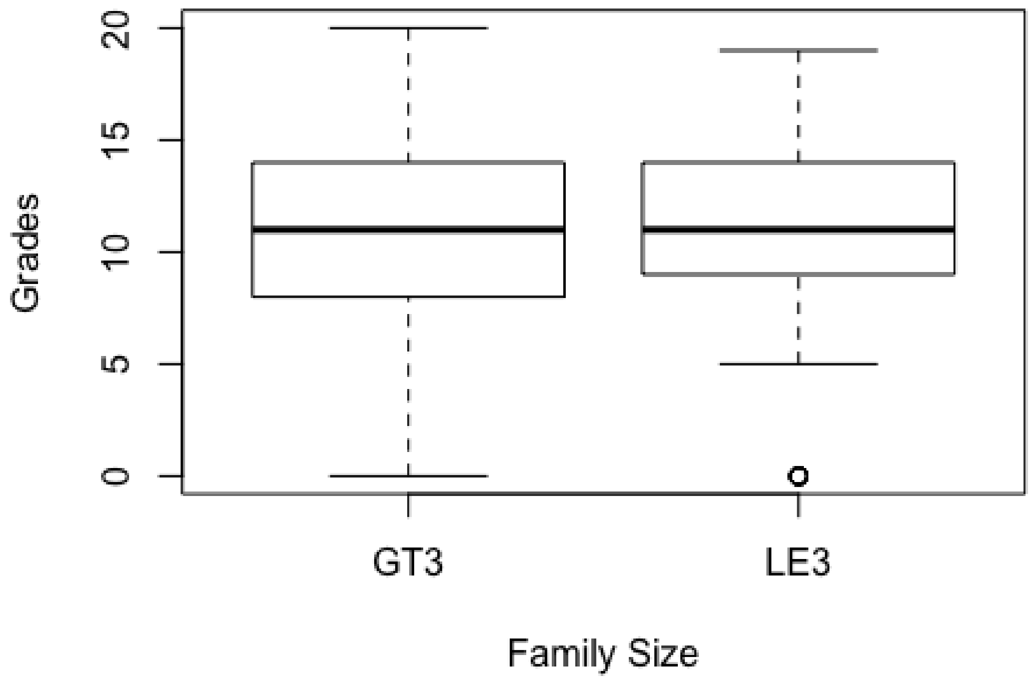
```
plot(sex,G3, xlab = "Gender", ylab = "Grades")
```



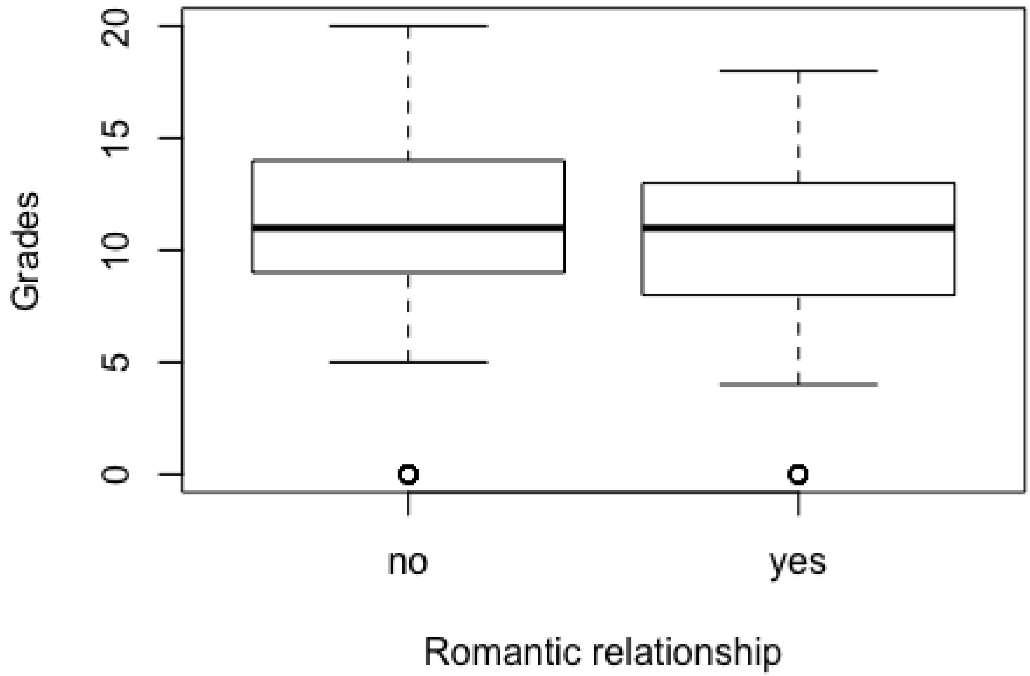
```
plot(address,G3, xlab = "Address", ylab = "Grades")
```



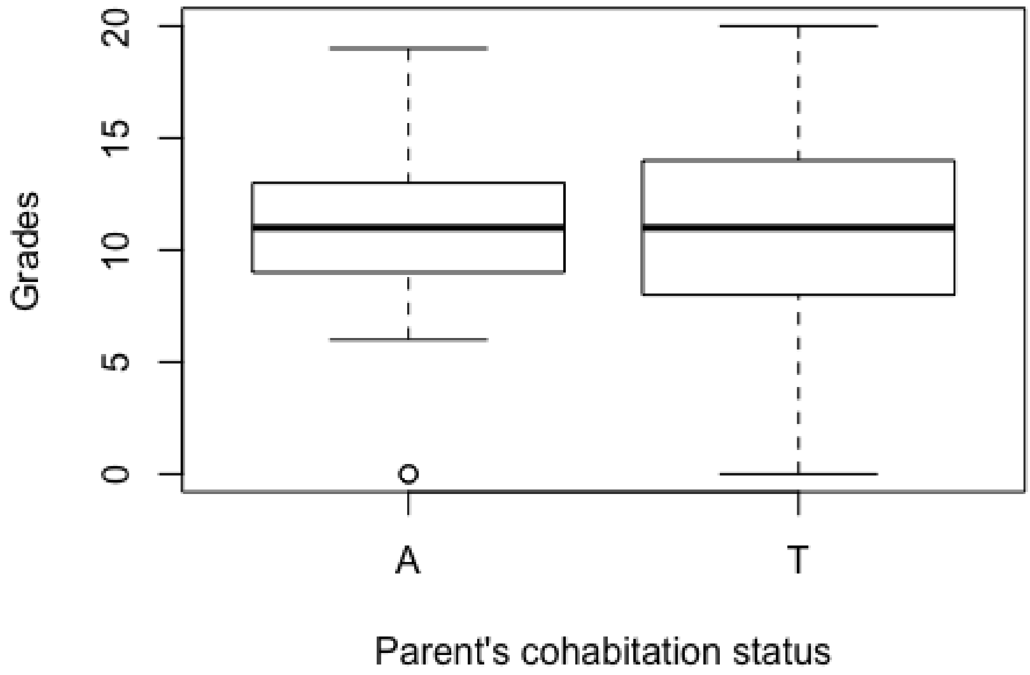
```r
plot(famsize, G3, xlab = "Family Size", ylab = "Grades")
```



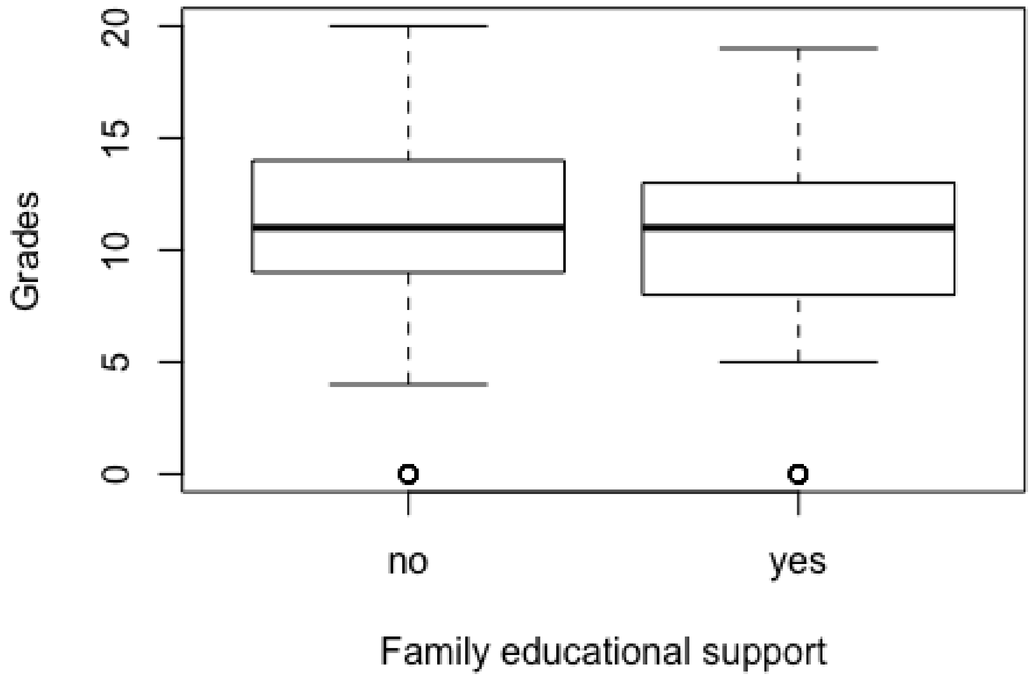
```
plot(romantic,G3, xlab = "Romantic relationship", ylab = "Grades")
```



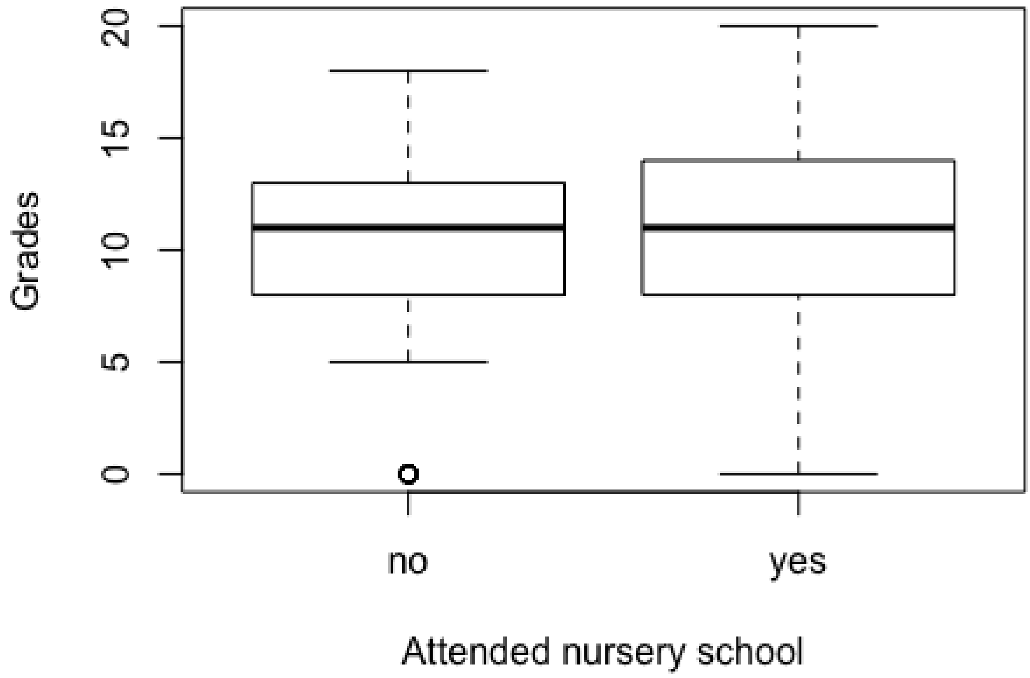
```r
plot(Pstatus,G3, xlab = "Parent's cohabitation status", ylab = "Grades")
```



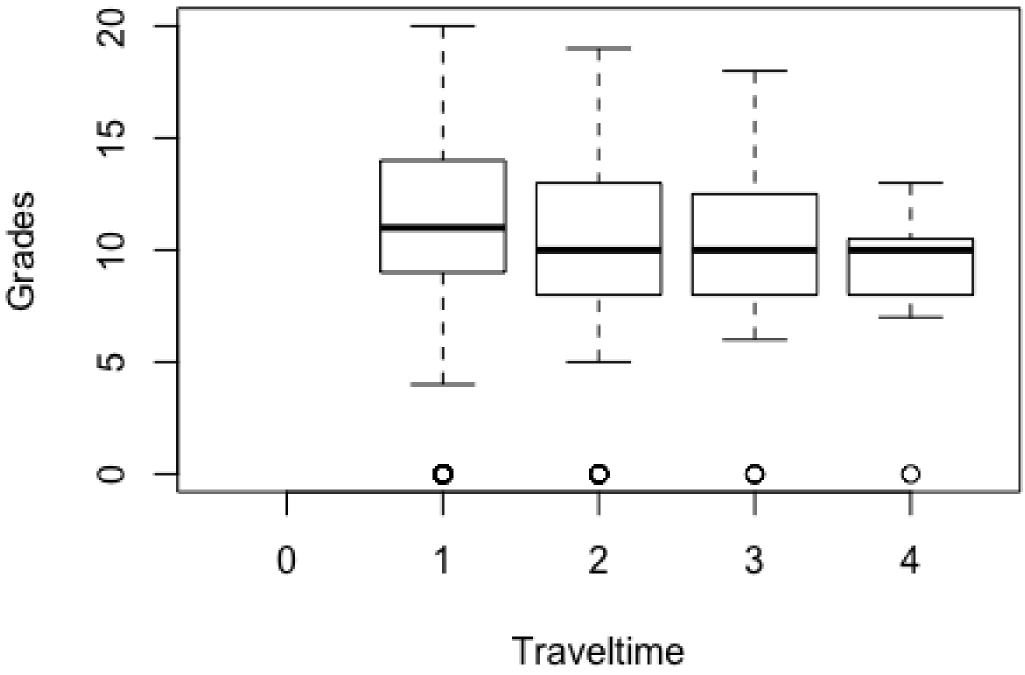
```r
plot(famsup,G3, xlab = "Family educational support", ylab = "Grades")
```



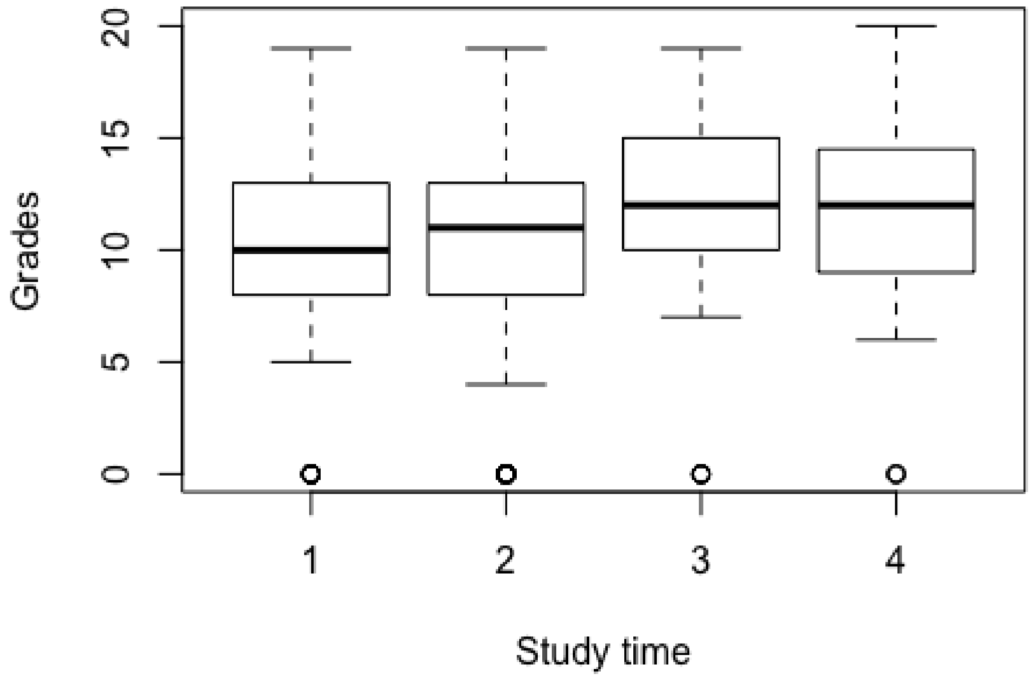
```r
plot(nursery,G3, xlab = "Attended nursery school", ylab = "Grades")
```



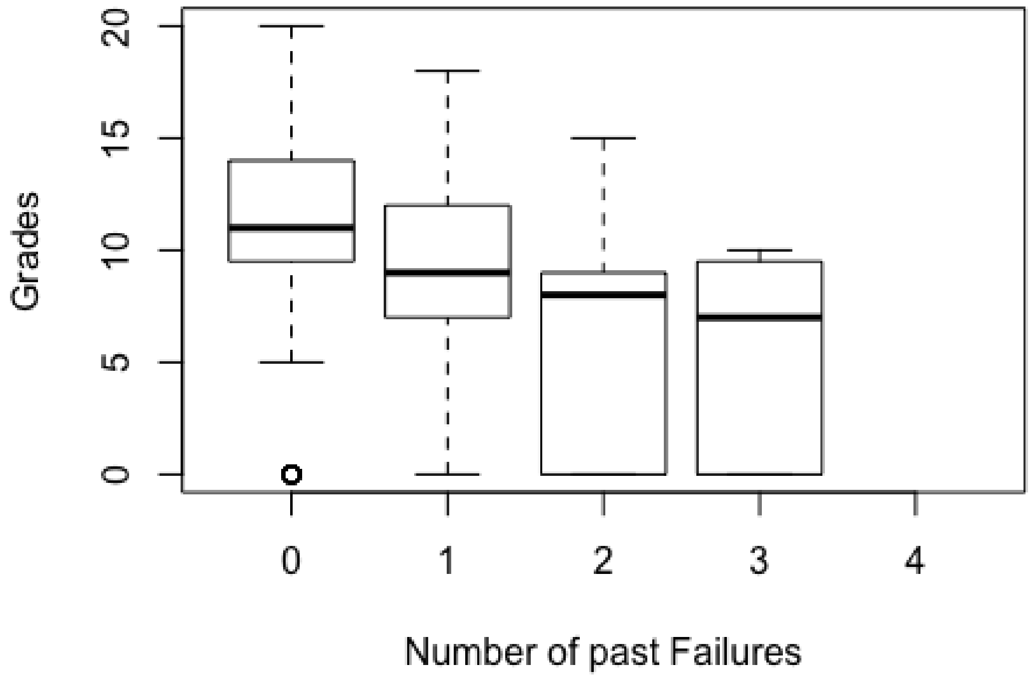
```
plot(traveltime,G3, xlab = "Traveltime", ylab = "Grades")
```



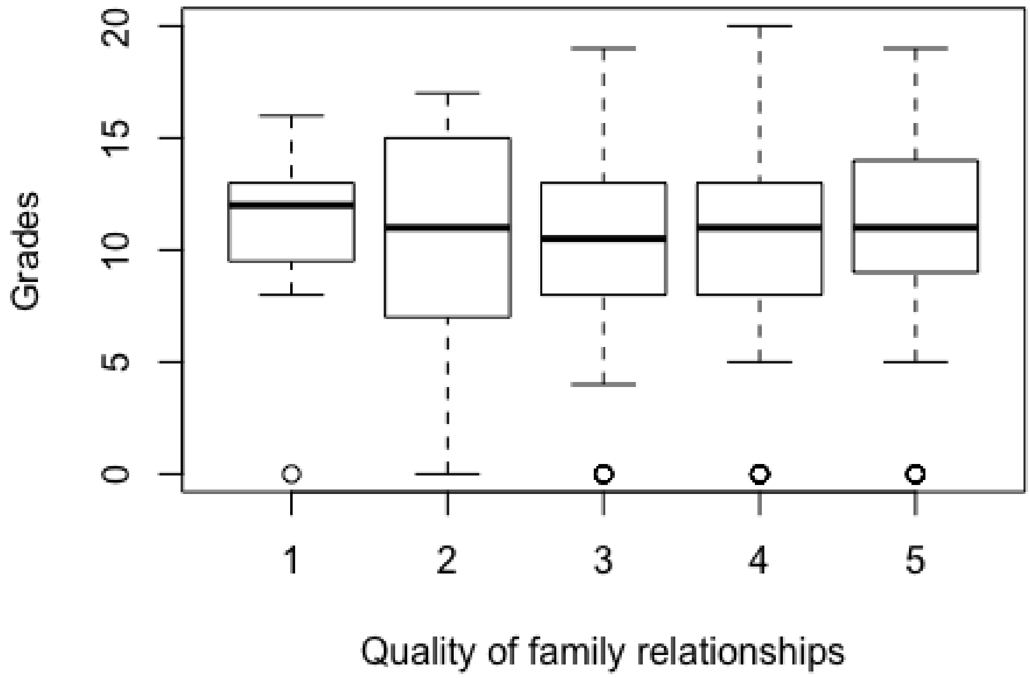
```
plot(studytime,G3, xlab = "Study time", ylab = "Grades")
```



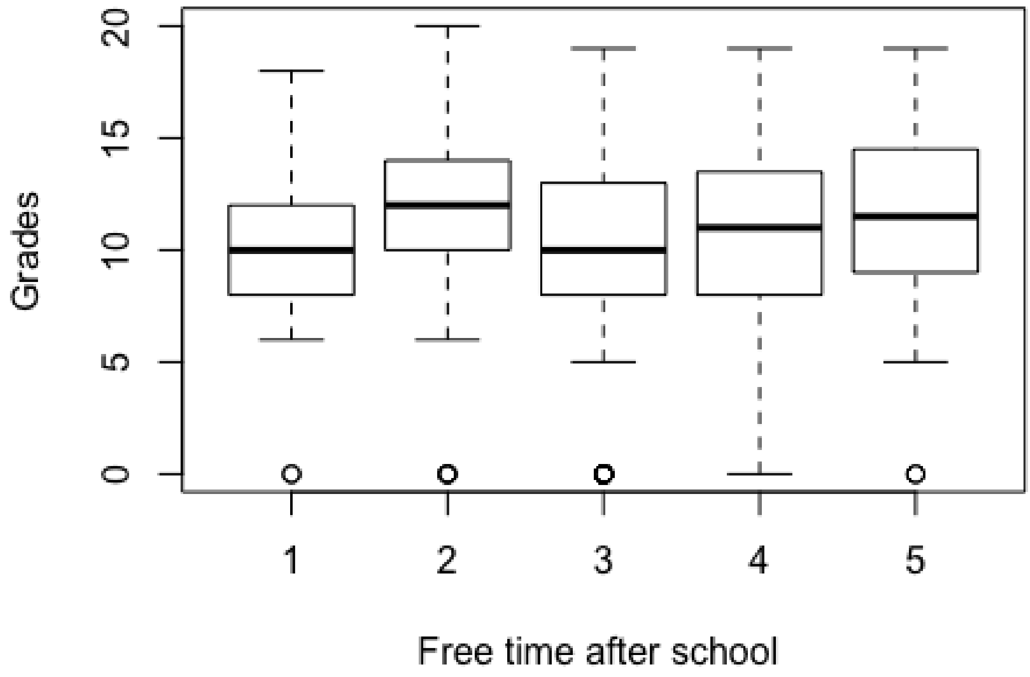
```
plot(failures,G3, xlab = "Number of past Failures", ylab = "Grades")
```



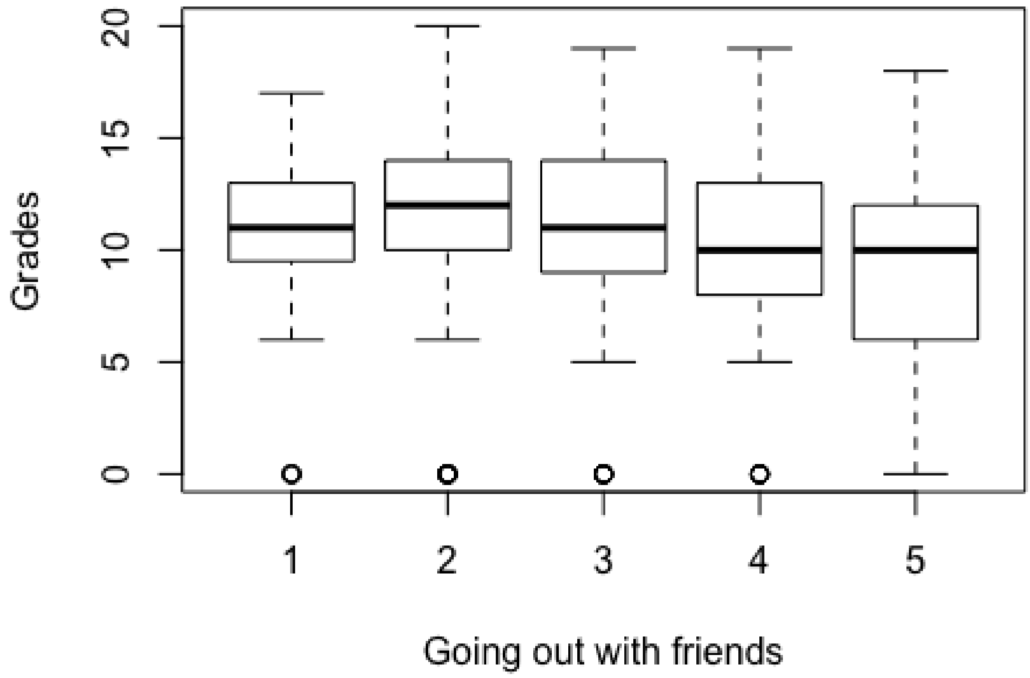
```
plot(famrel,G3, xlab = "Quality of family relationships", ylab = "Grades")
```



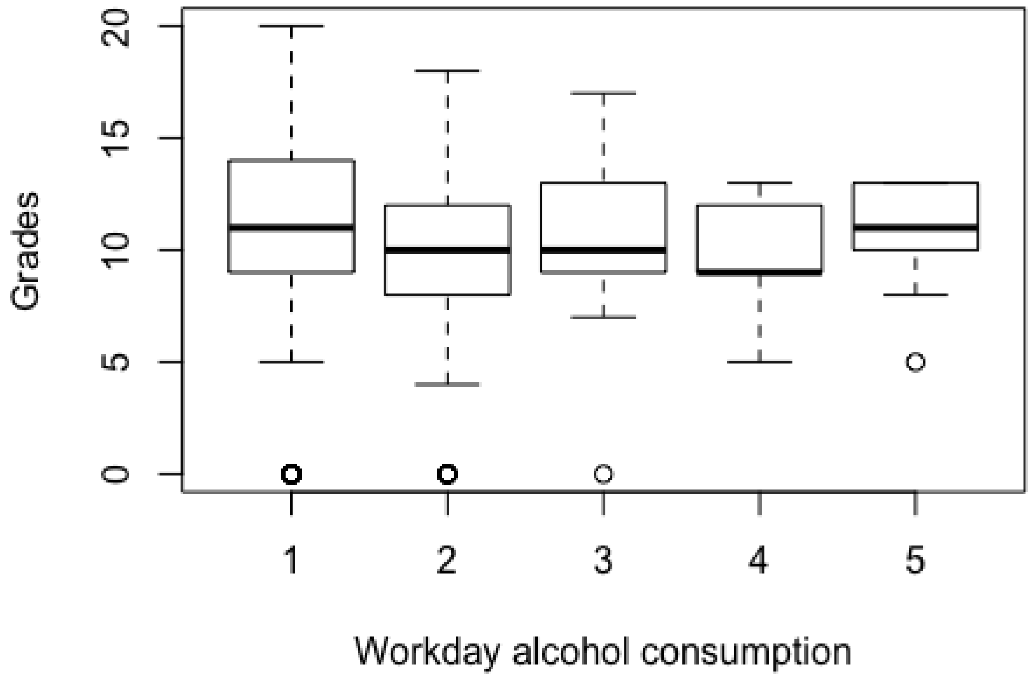
```
plot(freetime,G3, xlab = "Free time after school ", ylab = "Grades")
```



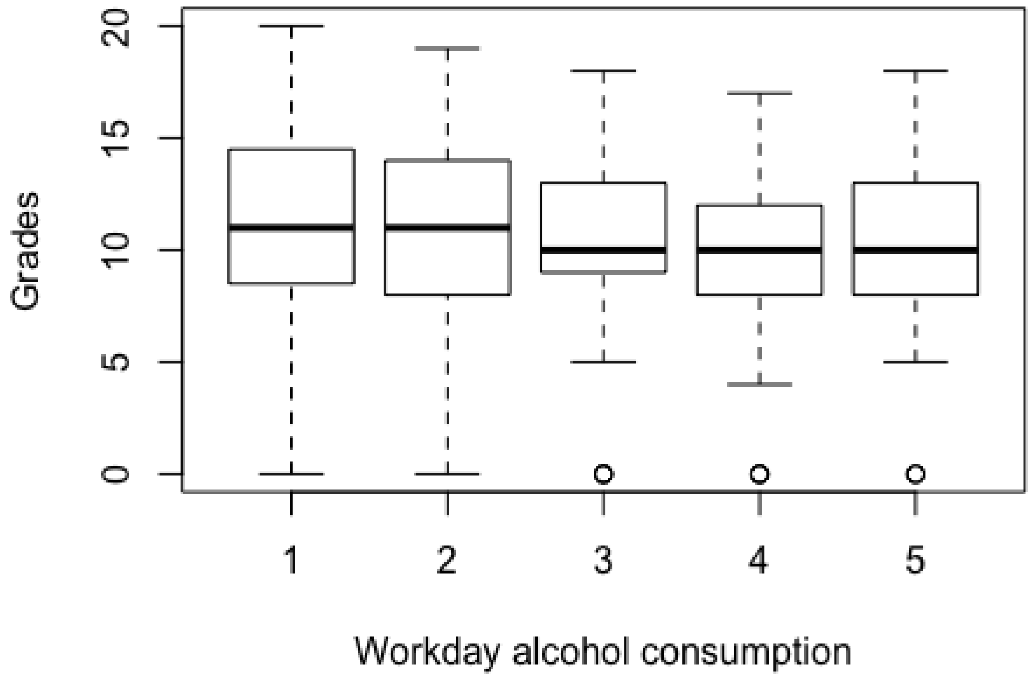
```
plot(goout,G3, xlab = "Going out with friends", ylab = "Grades")
```



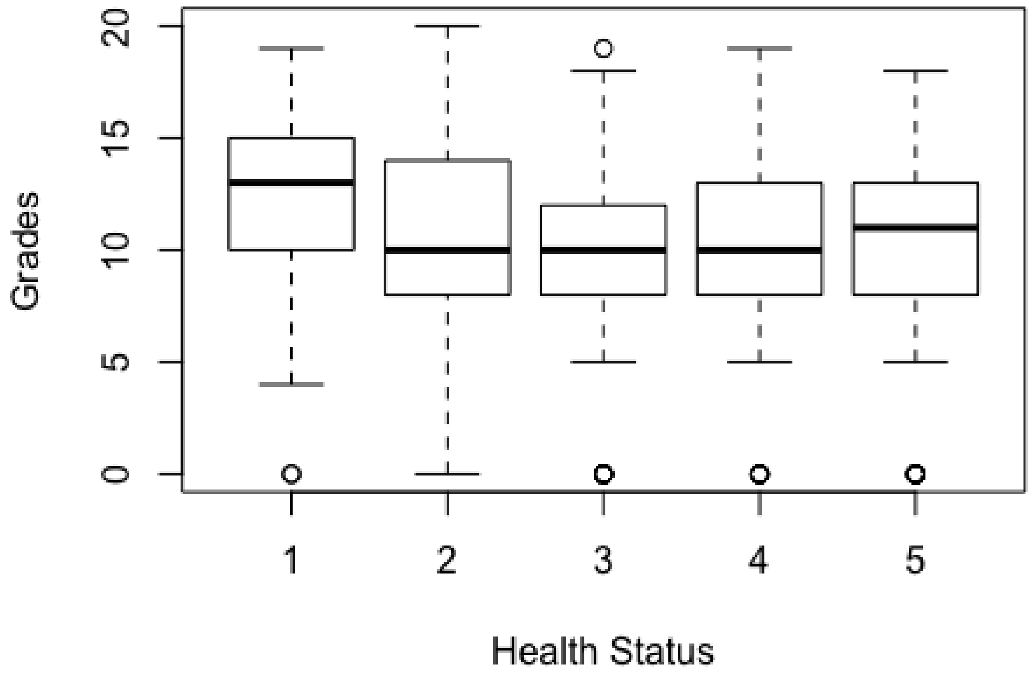
```r
plot(Dalc,G3, xlab = "Workday alcohol consumption", ylab = "Grades")
```



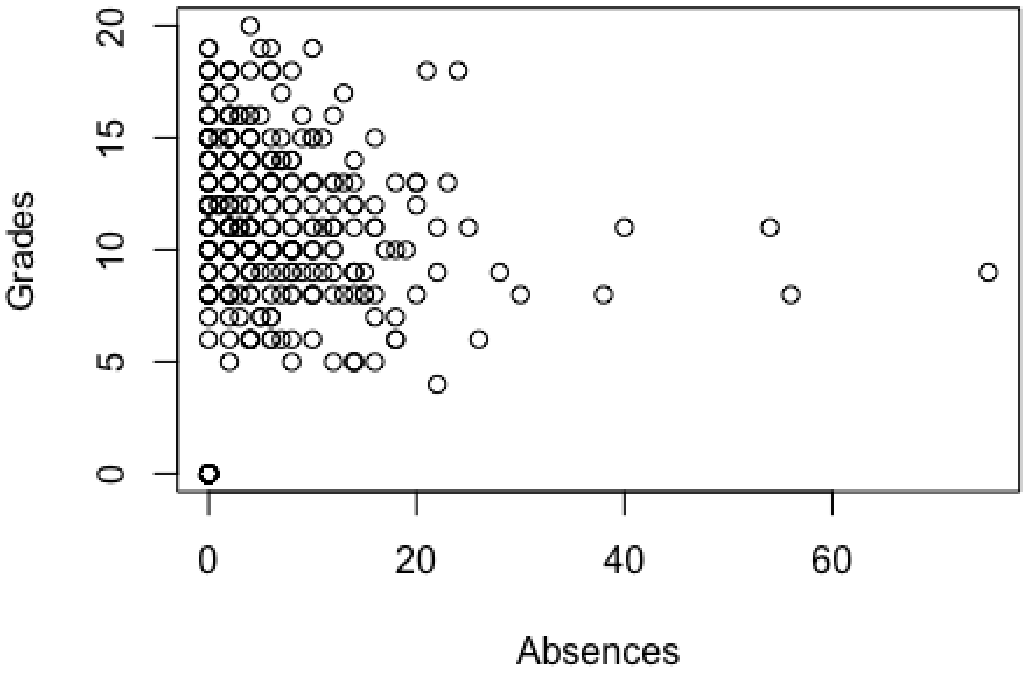
```
plot(Walc,G3, xlab = "Workday alcohol consumption", ylab = "Grades")
```



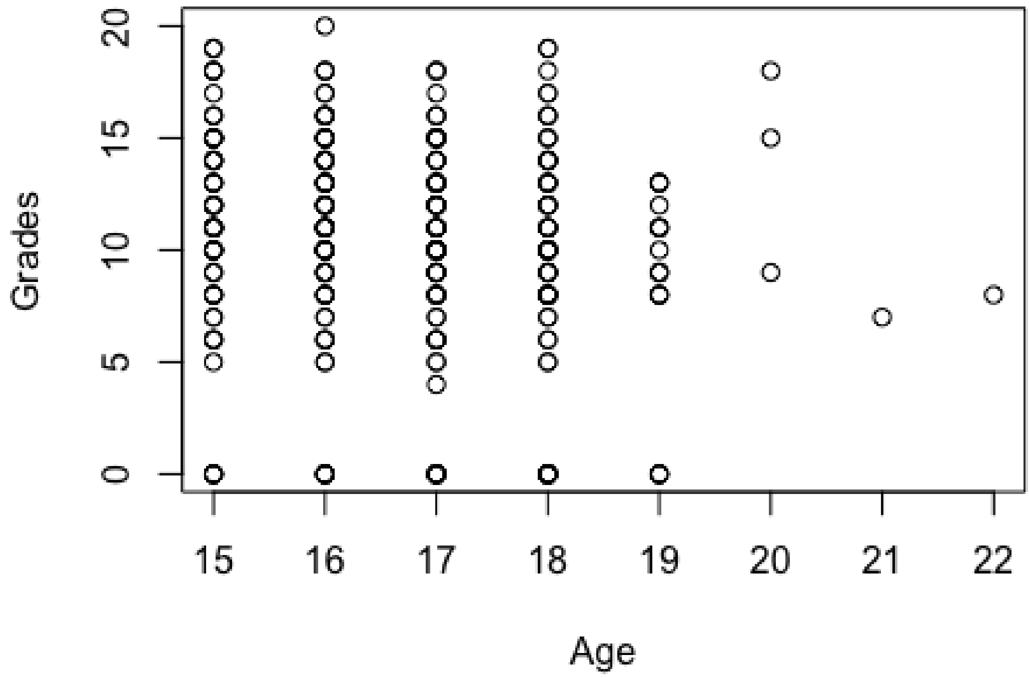
```
plot(health,G3, xlab = "Health Status", ylab = "Grades")
```



```
# Creating Scatter plots for numerical data
plot(mathematics_df$absences,mathematics_df$G3, xlab = "Absences", ylab =
"Grades")
```



```
plot(mathematics_df$age,mathematics_df$G3, xlab = "Age", ylab = "Grades")
```



```
############################
## Train / Test Split ######
############################

set.seed(-2)
train = sample(1:nrow(mathematics_df), 320)
test_g3 = mathematics_df[-train,31]

############################
## Modeling ################
############################

# Linear Model

linear_model_fit <- lm(G3~.,data = mathematics_df[train,])
summary(linear_model_fit)

##
## Call:
## lm(formula = G3 ~ ., data = mathematics_df[train, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0740  -1.8525   0.1554   2.4714   8.3134 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      11.99841    5.22890   2.295 0.022583 *  
## schoolMS          0.78692    1.00839   0.780 0.435910    
## sexM              1.04353    0.60192   1.734 0.084209 .  
## age              -0.30961    0.26388  -1.173 0.241796    
## addressU          0.64222    0.70176   0.915 0.360989    
## famsizeLE3        1.01118    0.59073   1.712 0.088181 .  
## PstatusT          0.08296    0.83306   0.100 0.920756    
## Medu.L           -1.75052    1.87508  -0.934 0.351426    
## Medu.Q            3.07195    1.49758   2.051 0.041281 *  
## Medu.C           -1.36320    1.00536  -1.356 0.176343    
## Medu^4            0.76256    0.62081   1.228 0.220473    
## Fedu.L           -0.97934    2.13375  -0.459 0.646651    
## Fedu.Q            1.18964    1.75318   0.679 0.498046    
## Fedu.C           -0.26437    1.14873  -0.230 0.818171    
## Fedu^4           -0.12624    0.61545  -0.205 0.837647    
## Mjobhealth        1.08935    1.37836   0.790 0.430088    
## Mjobother        -0.26719    0.86769  -0.308 0.758393    
## Mjobservices      0.68975    0.97722   0.706 0.480950    
## Mjobteacher      -1.69539    1.30142  -1.303 0.193870    
## Fjobhealth        0.66727    1.73238   0.385 0.700434    
## Fjobother         0.28872    1.20512   0.240 0.810852    
## Fjobservices      0.29026    1.24906   0.232 0.816428    
## Fjobteacher       1.60954    1.58575   1.015 0.311086    
## reasonhome        0.36022    0.65452   0.550 0.582569    
## reasonother       1.36514    1.01674   1.343 0.180599    
## reasonreputation  0.92527    0.69141   1.338 0.182035    
## guardianmother   -0.06579    0.65099  -0.101 0.919578    
## guardianother     0.42550    1.21348   0.351 0.726148    
## traveltime.L     -0.46672    1.29405  -0.361 0.718655    
## traveltime.Q     -0.66867    1.11129  -0.602 0.547916    
## traveltime.C     -0.99005    0.90950  -1.089 0.277392    
## studytime.L       1.19691    0.82453   1.452 0.147860    
## studytime.Q      -0.79249    0.69185  -1.145 0.253108    
## studytime.C      -0.85269    0.56012  -1.522 0.129189    
## failures.L       -3.36097    0.97519  -3.446 0.000666 ***
## failures.Q        0.93754    1.00675   0.931 0.352622    
## failures.C       -0.05165    0.99110  -0.052 0.958479    
## schoolsupyes     -1.27469    0.77109  -1.653 0.099565 .  
## famsupyes        -1.22693    0.56111  -2.187 0.029699 *  
## paidyes           0.39374    0.58361   0.675 0.500512    
## activitiesyes    -0.08317    0.52344  -0.159 0.873875    
## nurseryyes        0.21285    0.65547   0.325 0.745657    
## higheryes         1.25991    1.19443   1.055 0.292524    
## internetyes       0.03277    0.72959   0.045 0.964213    
## romanticyes      -1.67484    0.57573  -2.909 0.003952 ** 
## famrel.L          0.51286    1.30302   0.394 0.694216    
## famrel.Q          0.12293    1.17400   0.105 0.916689
## famrel.C         -0.23680    1.06497  -0.222 0.824222
## famrel^4         -0.08883    0.81295  -0.109 0.913073
## freetime.L        1.77428    1.02787   1.726 0.085552 .
## freetime.Q        0.92082    0.86007   1.071 0.285365
## freetime.C        1.27695    0.71743   1.780 0.076308 .
## freetime^4       -0.60625    0.51322  -1.181 0.238620
## goout.L          -1.59858    0.92132  -1.735 0.083955 .
## goout.Q          -0.77424    0.79620  -0.972 0.331784
## goout.C           0.52780    0.65011   0.812 0.417641
## goout^4          -0.05078    0.50530  -0.100 0.920035
## Dalc.L            0.03915    1.89035   0.021 0.983495
## Dalc.Q            0.91559    1.53907   0.595 0.552452
## Dalc.C            0.49680    1.38009   0.360 0.719167
## Dalc^4            0.86575    1.15692   0.748 0.454966
## Walc.L            1.26226    1.05867   1.192 0.234274
## Walc.Q            0.83043    0.80156   1.036 0.301195
## Walc.C            0.52345    0.68591   0.763 0.446094
## Walc^4            1.09674    0.58982   1.859 0.064139 .
## health.L         -0.64351    0.65751  -0.979 0.328672
## health.Q          0.76638    0.64503   1.188 0.235910
## health.C         -0.78729    0.68895  -1.143 0.254239
## health^4          0.15445    0.60895   0.254 0.799988
## absences          0.08388    0.03477   2.412 0.016565 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.127 on 250 degrees of freedom
## Multiple R-squared:  0.3886, Adjusted R-squared:  0.2199
## F-statistic: 2.303 on 69 and 250 DF,  p-value: 1.436e-06
```

```
# Backward AIC
library(leaps)
back_aic_fit = MASS::stepAIC(linear_model_fit, direction = "backward", trace
= FALSE)
back_aic_fit$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## 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
##
## Final Model:
## G3 ~ sex + famsize + Medu + Mjob + studytime + failures + schoolsup +
##     famsup + higher + romantic + freetime + goout + absences
```

```
## 
## 
##             Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1                                     250   4257.378 968.1881
## 2       - famrel  4 10.45415819       254   4267.833 960.9729
## 3         - Fedu  4 20.51056049       258   4288.343 954.5071
## 4         - Dalc  4 21.96943047       262   4310.313 948.1423
## 5         - Fjob  4 35.75766574       266   4346.070 942.7860
## 6       - health  4 39.70643076       270   4385.777 937.6963
## 7   - traveltime  3 19.46521297       273   4405.242 933.1134
## 8     - guardian  2  1.06200578       275   4406.304 929.1906
## 9   - activities  1  0.02281112       276   4406.327 927.1922
## 10    - internet  1  0.37423349       277   4406.701 925.2194
## 11     - Pstatus  1  0.80318075       278   4407.504 923.2777
## 12     - nursery  1  1.17093867       279   4408.675 921.3627
## 13      - school  1  5.70074056       280   4414.376 919.7762
## 14        - paid  1  7.40098904       281   4421.777 918.3123
## 15        - Walc  4 98.20091688       285   4519.978 917.3412
## 16     - address  1 14.63238596       286   4534.610 916.3755
## 17         - age  1 20.33106923       287   4554.941 915.8070
## 18      - reason  3 78.74701774       290   4633.688 915.2920
```

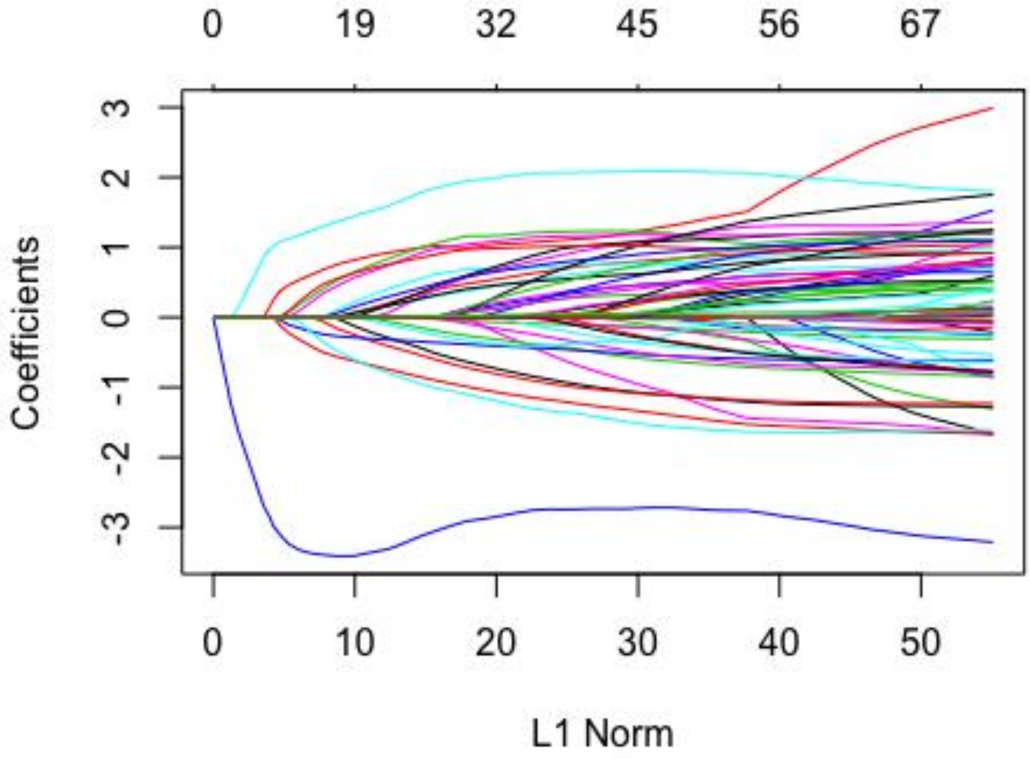
```
summary(back_aic_fit)
```

```
## 
## Call:
## lm(formula = G3 ~ sex + famsize + Medu + Mjob + studytime + failures + 
##     schoolsup + famsup + higher + romantic + freetime + goout + 
##     absences, data = mathematics_df[train, ])
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11.8388  -1.7785   0.3473   2.3933   8.1492 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     7.75582    1.28834   6.020 5.26e-09 ***
## sexM            0.88862    0.52879   1.680   0.0939 .  
## famsizeLE3      0.88643    0.51428   1.724   0.0858 .  
## Medu.L         -1.03514    1.65681  -0.625   0.5326    
## Medu.Q          3.00811    1.33363   2.256   0.0248 *  
## Medu.C         -0.84219    0.90189  -0.934   0.3512    
## Medu^4          0.49662    0.55621   0.893   0.3727    
## Mjobhealth      1.30662    1.14419   1.142   0.2544    
## Mjobother      -0.18330    0.74980  -0.244   0.8070    
## Mjobservices    0.90314    0.83894   1.077   0.2826    
## Mjobteacher    -1.80094    1.12721  -1.598   0.1112    
## studytime.L     1.16637    0.70942   1.644   0.1012    
## studytime.Q    -0.58373    0.61553  -0.948   0.3437    
## studytime.C    -0.64913    0.49971  -1.299   0.1950    
## failures.L    -3.56331    0.86272  -4.130 4.74e-05 ***
## failures.Q     1.41298    0.86155   1.640   0.1021
## failures.C    -0.01880    0.87314  -0.022   0.9828
## schoolsupyes -0.97759    0.69171  -1.413   0.1586
## famsupyes    -1.11724    0.49569  -2.254   0.0249 *
## higheryes     1.58600    1.05950   1.497   0.1355
## romanticyes  -1.56776    0.50967  -3.076   0.0023 **
## freetime.L    1.63414    0.92428   1.768   0.0781 .
## freetime.Q    0.98943    0.78465   1.261   0.2083
## freetime.C    0.97661    0.63081   1.548   0.1227
## freetime^4   -0.66092    0.46581  -1.419   0.1570
## goout.L      -1.69259    0.78911  -2.145   0.0328 *
## goout.Q      -0.66221    0.73031  -0.907   0.3653
## goout.C       0.25449    0.57924   0.439   0.6607
## goout^4       0.12876    0.45404   0.284   0.7769
## absences      0.06907    0.03020   2.287   0.0229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.997 on 290 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.268
## F-statistic: 5.028 on 29 and 290 DF,  p-value: 1.308e-13
```

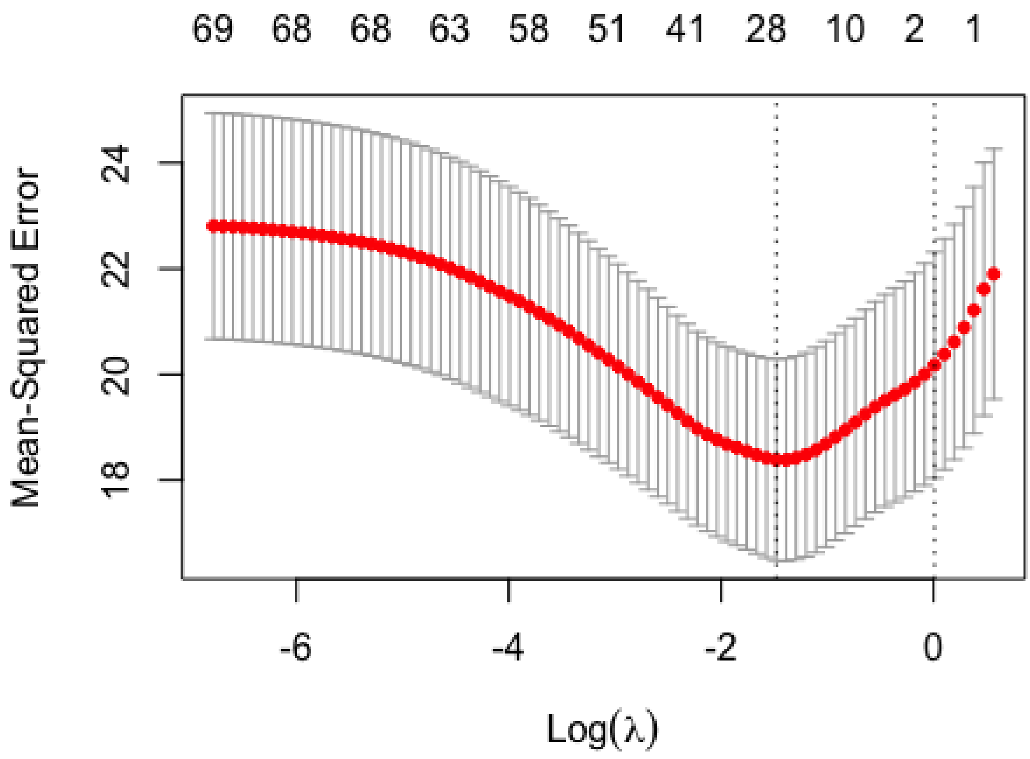
```
back_aic_pred = predict(back_aic_fit, newdata = mathematics_df[-train,1:30])
mean((back_aic_pred-test_g3)^2)
```

```
## [1] 16.89163
```

```
# Lasso Regression
library(glmnet)
x_train = model.matrix(G3~., mathematics_df[train,])[,-1]
x_test = model.matrix(G3~., mathematics_df[-train,])[,-1]

y_train = mathematics_df[train,] %>% dplyr::select(G3) %>% unlist() %>%
as.numeric()
y_test = mathematics_df[-train,] %>% dplyr::select(G3) %>% unlist() %>%
as.numeric()

lasso_fit_1 = glmnet(x_train, y_train, alpha = 1)

plot(lasso_fit_1)
```



```
set.seed(1)
cv.out = cv.glmnet(x_train, y_train, alpha = 1)
plot(cv.out)
```

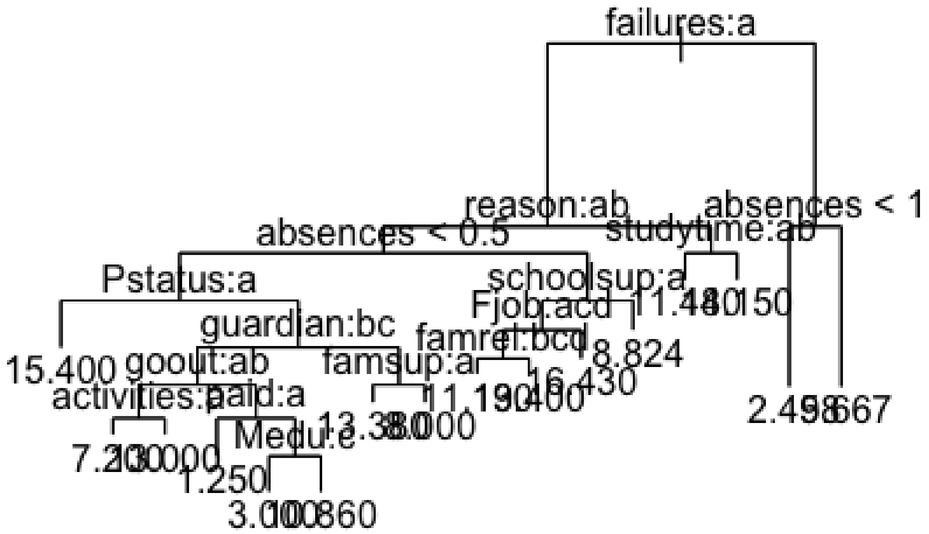


```
bstlambda = cv.out$lambda.min
lasso_pred = predict(lasso_fit_1, s = bstlambda, newx = x_test)
mean((lasso_pred - y_test)^2)

## [1] 16.56345

lasso_bst_fit <- glmnet(x_train, y_train, alpha = 1, lambda = bstlambda)
coef(lasso_bst_fit)

## 72 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)       8.1121081760
## schoolMS          .
## sexM              0.3907867587
## age              -0.0237846575
## addressU          .
## famsizeLE3        0.2979167741
## PstatusT          .
## Medu.L            .
## Medu.Q            1.0200709601
## Medu.C            .
## Medu^4            .
## Fedu.L            .
## Fedu.Q            0.0031428350
## Fedu.C            .
## Fedu^4            .
## Mjobhealth        1.0536655071
## Mjobother         .
## Mjobservices      0.6112917929
## Mjobteacher       .
## Fjobhealth        .
## Fjobother         .
## Fjobservices      .
## Fjobteacher       .
## reasonhome        .
## reasonother       0.3847414836
## reasonreputation  0.3134586003
## guardianmother    .
## guardianother     .
## traveltime.L      .
## traveltime.Q      .
## traveltime.C      .
## traveltime^4      .
## studytime.L       0.4372759550
## studytime.Q       .
## studytime.C      -0.1864573809
## failures.L       -3.0599340303
## failures.Q        1.8409943387
## failures.C        .
## failures^4        .
## schoolsupyes     -0.5091894119
## famsupyes        -0.5986108872
## paidyes           0.0053314532
## activitiesyes     .
## nurseryyes        .
## higheryes         0.9784248643
## internetyes       .
## romanticyes      -0.8957895968
## famrel.L          .
## famrel.Q          .
## famrel.C          .
## famrel^4          .
## freetime.L        .
## freetime.Q        0.8921502321
## freetime.C        0.0040914487
## freetime^4       -0.3587357721
## goout.L          -1.0055458605
## goout.Q          -0.0006621955
## goout.C           .
## goout^4           .
## Dalc.L            .
## Dalc.Q            .
## Dalc.C           -0.2506704166
## Dalc^4            .
## Walc.L              .
## Walc.Q              .
## Walc.C              .
## Walc^4              0.4711815040
## health.L            .
## health.Q            0.0040528849
## health.C            .
## health^4            .
## absences            0.0167907299
```

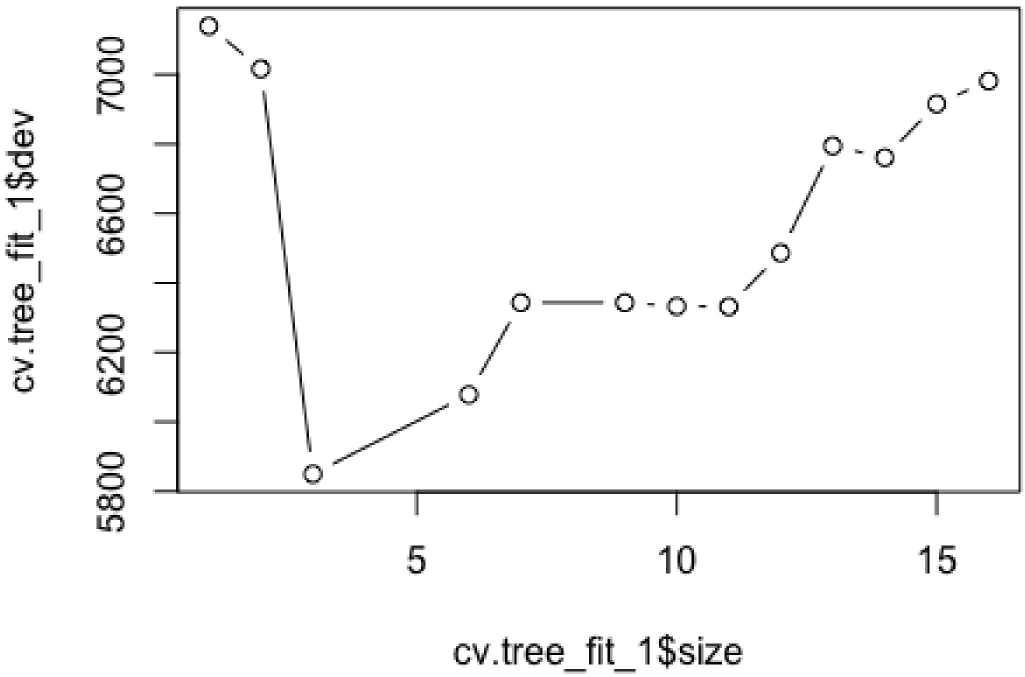
```
# TREES

tree_fit_1 = tree(G3~., data = mathematics_df , subset = train)
summary(tree_fit_1)
```

```
##
## Regression tree:
## tree(formula = G3 ~ ., data = mathematics_df, subset = train)
## Variables actually used in tree construction:
##  [1] "failures"   "reason"     "absences"   "Pstatus"    "guardian"
##  [6] "goout"      "activities" "paid"       "Medu"       "famsup"
## [11] "schoolsup"  "Fjob"       "famrel"     "studytime"
## Number of terminal nodes:  16
## Residual mean deviance:  10.52 = 3197 / 304
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -11.4800  -2.2430  -0.4286   0.0000   2.0000  12.0000
```

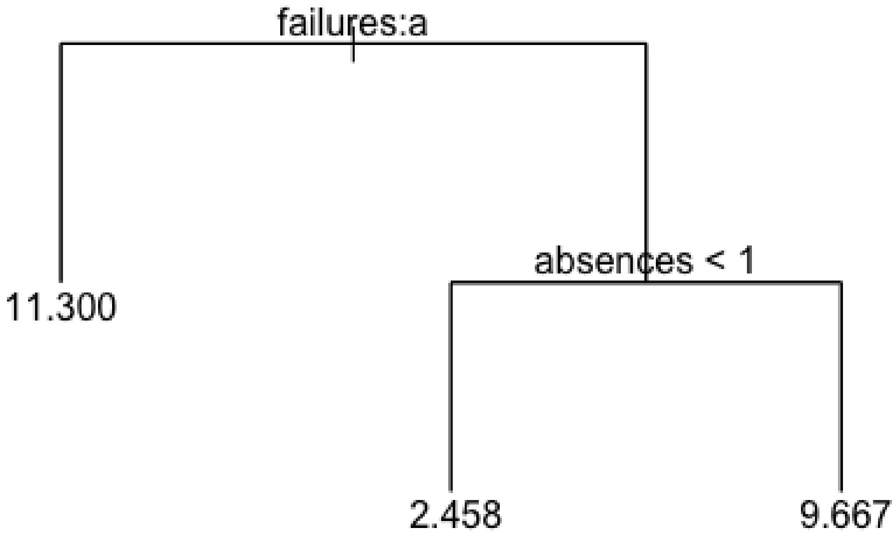
```
plot(tree_fit_1)
text(tree_fit_1)
```



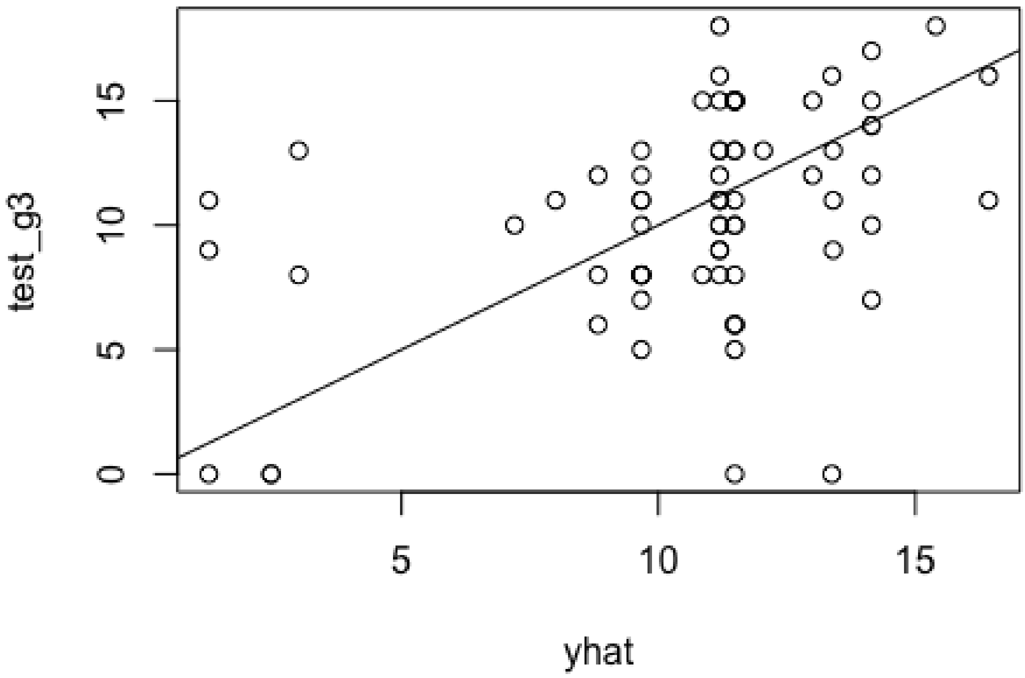
```r
cv.tree_fit_1 = cv.tree(tree_fit_1)
plot(cv.tree_fit_1$size, cv.tree_fit_1$dev, type = 'b')
```



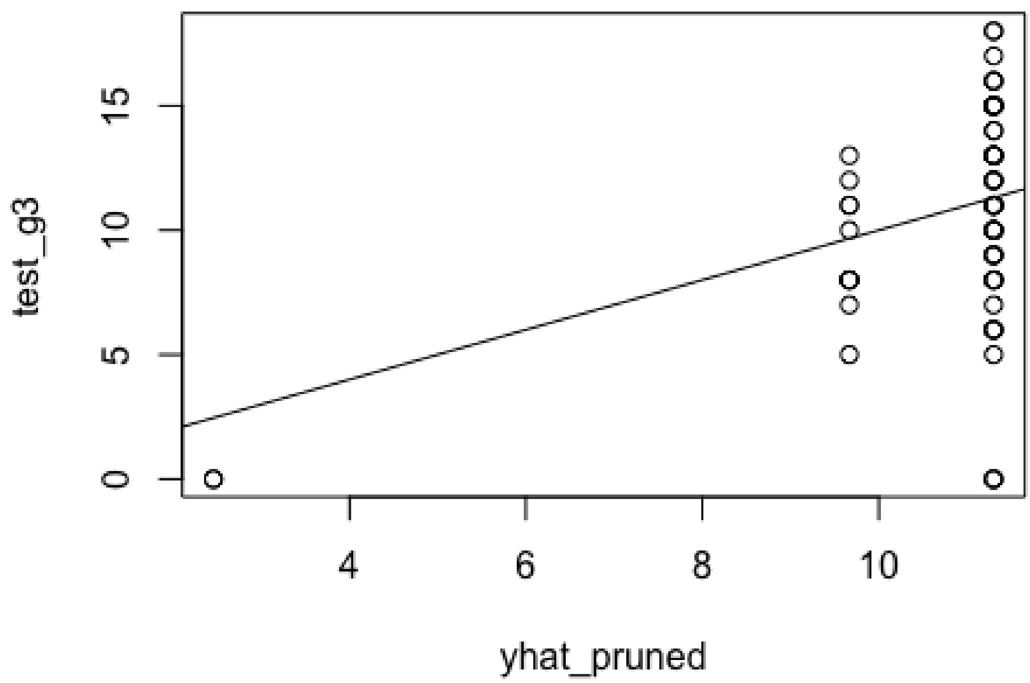
```
prune.tree_fit_1 = prune.tree(tree_fit_1, best = 3)
plot(prune.tree_fit_1)
text(prune.tree_fit_1)
```



```r
yhat = predict(tree_fit_1, newdata = mathematics_df[-train,1:30])
plot(yhat, test_g3)
abline(0,1)
```



```
mean((yhat-test_g3)^2)

## [1] 16.18935

yhat_pruned = predict(prune.tree_fit_1, newdata = mathematics_df[-
train,1:30])
plot(yhat_pruned, test_g3)
abline(0,1)
```



```
mean((yhat_pruned-test_g3)^2)

## [1] 14.18483

############## RANDOM FOREST ##############

library(randomForest)

# Bagged DT : m = p predictors i.e. mtry = 30
set.seed(-1)
bagged_tree_fit = randomForest(G3~., data = mathematics_df[train,], mtry =
30, ntree= 1000, importance = TRUE)
bagged_tree_fit

##
## Call:
##  randomForest(formula = G3 ~ ., data = mathematics_df[train, ],      mtry
= 30, ntree = 1000, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 30
##
##           Mean of squared residuals: 15.73308
##                     % Var explained: 27.7
```

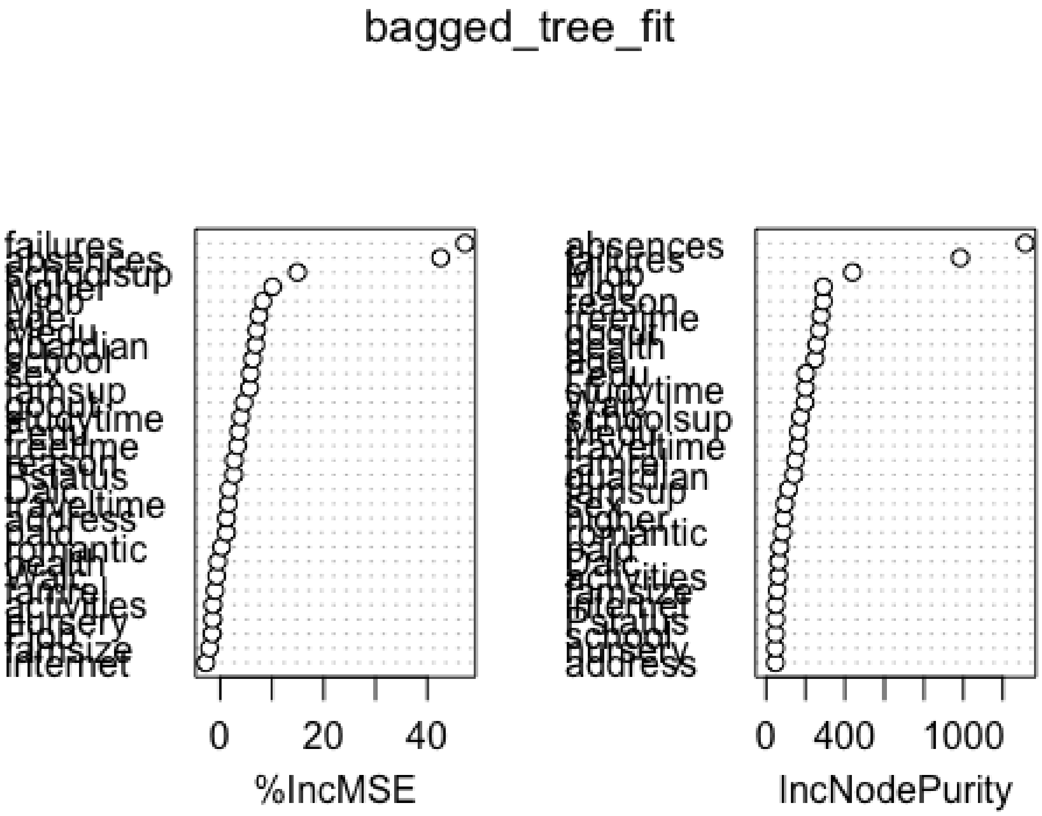
```
yhat_bagged_tree_fit = predict(bagged_tree_fit, newdata = mathematics_df[-
train,1:30])
bagging_test = mathematics_df[-train,"G3"]
mean((yhat_bagged_tree_fit-bagging_test)^2)
```

```
## [1] 13.64493
```

```
importance(bagged_tree_fit)
```

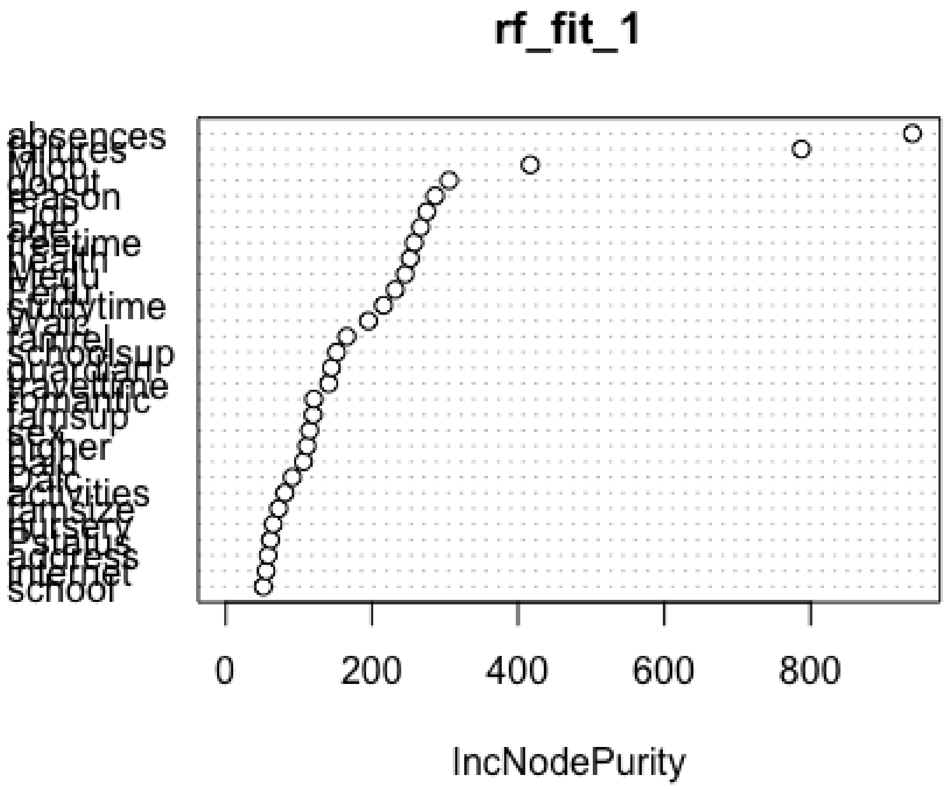
```
##                %IncMSE IncNodePurity
## school       6.2900765      48.08588
## sex          5.8481957      95.71605
## age          7.5323696     248.59311
## address      1.2370782      47.78070
## famsize     -2.0438257      57.15514
## Pstatus      2.6621230      49.83237
## Medu         7.1020693     163.76266
## Fedu         3.6267220     202.62120
## Mjob         8.2467471     438.39734
## Fjob        -1.5411985     291.12677
## reason       2.7200797     290.45428
## guardian     6.8696247     141.77602
## traveltime   1.5927617     163.19483
## studytime    4.0311527     200.69114
## failures    47.2734019     984.65397
## schoolsup   14.9256151     175.38785
## famsup       5.6901665     113.65681
## paid         1.2260391      70.59158
## activities  -1.3800761      63.34758
## nursery     -1.4299752      47.82202
## higher      10.1396247      88.88543
## internet    -2.7663905      50.10340
## romantic     0.2998804      84.55053
## famrel      -0.9745020     148.66097
## freetime     3.3716167     277.18886
## goout        4.6085417     271.19775
## Dalc         1.7327670      64.28213
## Walc        -0.6609624     198.37204
## health      -0.4446879     256.96454
## absences    42.5668597    1314.20257
```

```
varImpPlot(bagged_tree_fit)
```



```
# Random Forest - that is with m != p, mtry = p/3 (optimal for regression
trees)
set.seed(-1)
rf_fit_1 = randomForest(G3~., data = mathematics_df[train,], mtry = 10,
ntree= 1000, importance = FALSE)
rf_fit_1

##
## Call:
##  randomForest(formula = G3 ~ ., data = mathematics_df[train, ],      mtry
= 10, ntree = 1000, importance = FALSE)
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 15.68223
##                     % Var explained: 27.93

yhat_rf_fit_1 = predict(rf_fit_1, newdata = mathematics_df[-train,])
bagging_test = mathematics_df[-train,"G3"]
mean((yhat_rf_fit_1-bagging_test)^2)

## [1] 13.83737
```

```
varImpPlot(rf_fit_1)
```



```
##################################  Model Performance
###############################

cat("RMSE of Backward Step wise : ", sqrt(mean((back_aic_pred-
test_g3)^2)),"\n")

## RMSE of Backward Step wise :  4.109943

cat("RMSE of Lasso : ", sqrt(mean((lasso_pred - y_test)^2)),"\n")

## RMSE of Lasso :  4.069822

cat("RMSE of Decision Tree : ", sqrt(mean((yhat_pruned-test_g3)^2)),"\n")

## RMSE of Decision Tree :  3.766276

cat("RMSE of Bagged Decision Trees : ", sqrt(mean((yhat_bagged_tree_fit-
bagging_test)^2)),"\n")

## RMSE of Bagged Decision Trees :  3.693905

cat("RMSE of RF : ", sqrt(mean((yhat_rf_fit_1-bagging_test)^2)),"\n")
```

```
## RMSE of RF :  3.719861
```