Grades_Prediction.R

harshitmehta

2020-04-02

```
# loading all the libraries

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(ROSE)

## Loaded ROSE 0.0-3

#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)

################### Understanding the 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 Cleaning & Preparation
#########################################
```

```r
# to check if there are any missing values
any(is.na(school1))
```

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

```r
# Thus we have no missing values in the data set.

# dropping G1 and G2 from school1 (math)
df_math = subset(school1, select = -c(G1,G2))
colnames(df_math)
```

```
##  [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"
```

```r
glimpse(df_math)
```

```
## Observations: 395
## Variables: 31
## $ 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…
```

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

# Medu - denotes Mother's eductaion - 5 levels
df_math$Medu = factor(df_math$Medu, levels=c("0","1","2","3","4"),
ordered=TRUE)
summary(df_math$Medu)
```

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

```r
# Fedu - denotes Father's eductaion - 5 levels
df_math$Fedu = factor(df_math$Fedu, levels=c("0","1","2","3","4"),
ordered=TRUE)
summary(df_math$Fedu)
```

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

```
# famrel - denotes - quality of family relationships
# 1 - very bad to 5 - excellent
df_math$famrel = factor(df_math$famrel, levels=1:5, ordered=TRUE)
summary(df_math$famrel)
```

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

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

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

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

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

```
# freetime - free time after school (1 - very low to 5 - very high)
df_math$freetime = factor(df_math$freetime, levels=1:5, ordered=TRUE)
summary(df_math$freetime)
```

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

```
# goout - going out with friends ( 1 - very low to 5 - very high)
df_math$goout = factor(df_math$goout, levels=1:5, ordered=TRUE)
summary(df_math$goout)
```

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

```
# Dalc - workday alcohol consumption (from 1 - very low to 5 - very high)
df_math$Dalc = factor(df_math$Dalc, levels=1:5, ordered=TRUE)
summary(df_math$Dalc)
```

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

```
# Walc - weekend alcohol consumption ( 1 - very low to 5 - very high)
df_math$Walc = factor(df_math$Walc, levels=1:5, ordered=TRUE)
summary(df_math$Walc)
```

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

```
# health - current health status ( 1 - very bad to 5 - very good)
df_math$health = factor(df_math$health, levels=1:5, ordered=TRUE)
summary(df_math$health)
```

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

```
# failures - number of past class failures (n if 1<=n<3, else 4)
df_math$failures = factor(df_math$failures, levels=0:4, ordered=TRUE)
summary(df_math$failures)
```

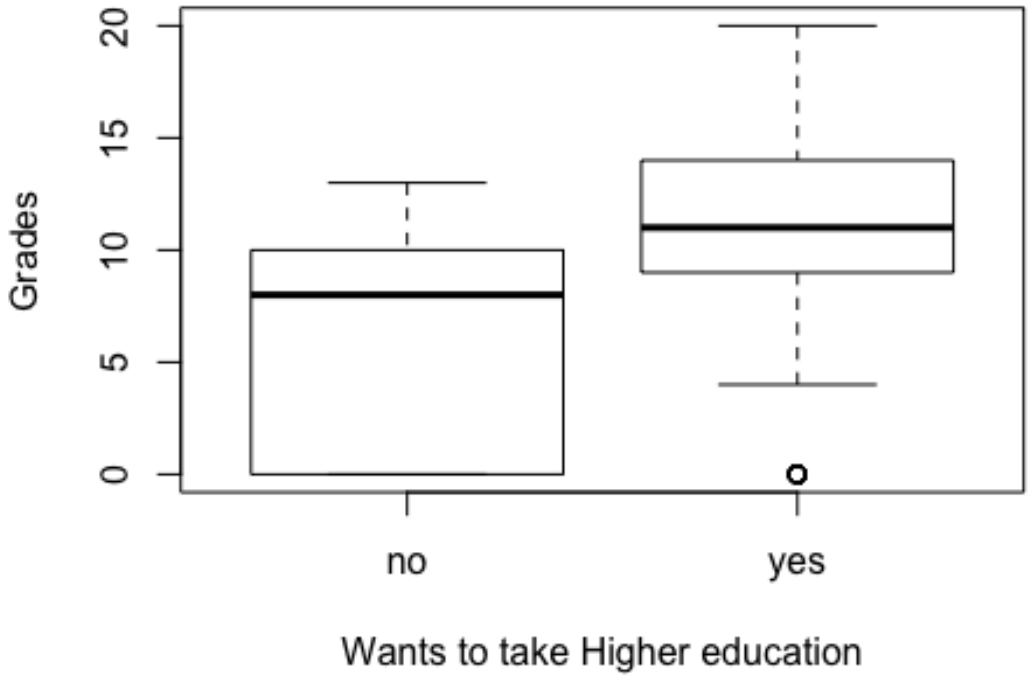
```
##   0   1   2   3   4
## 312  50  17  16   0
```

```
summary(df_math)
```

```
##  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(df_math))

plot(higher,G3, xlab = "Wants to take Higher education", ylab = "Grades",
main = "Figure 2.1")
```

Figure 2.1

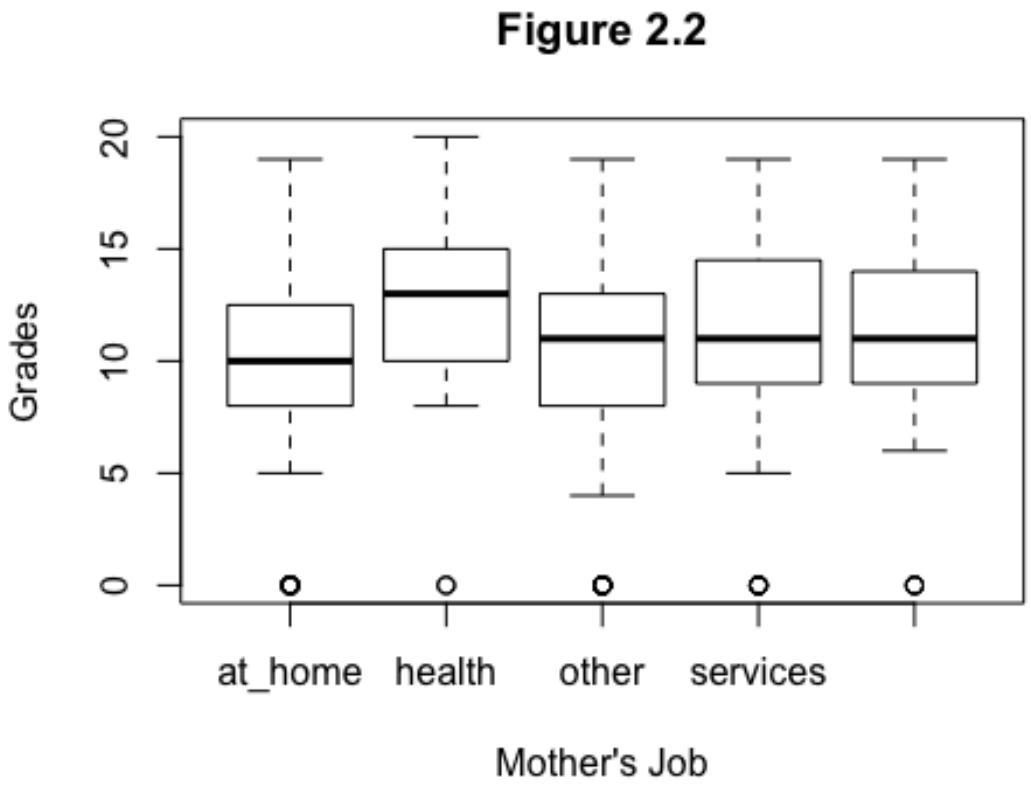



```
summary(df_math[df_math$higher=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.61   14.00   20.00
```

```r
summary(df_math[df_math$higher=="no",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     8.0     6.8    10.0    13.0
```

```r
plot(Mjob,G3, xlab = "Mother's Job", ylab = "Grades", main = "Figure 2.2")
```




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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.153  12.500  19.000
```

```r
summary(df_math[df_math$Mjob=="health",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   13.00   12.15   15.00   20.00
```

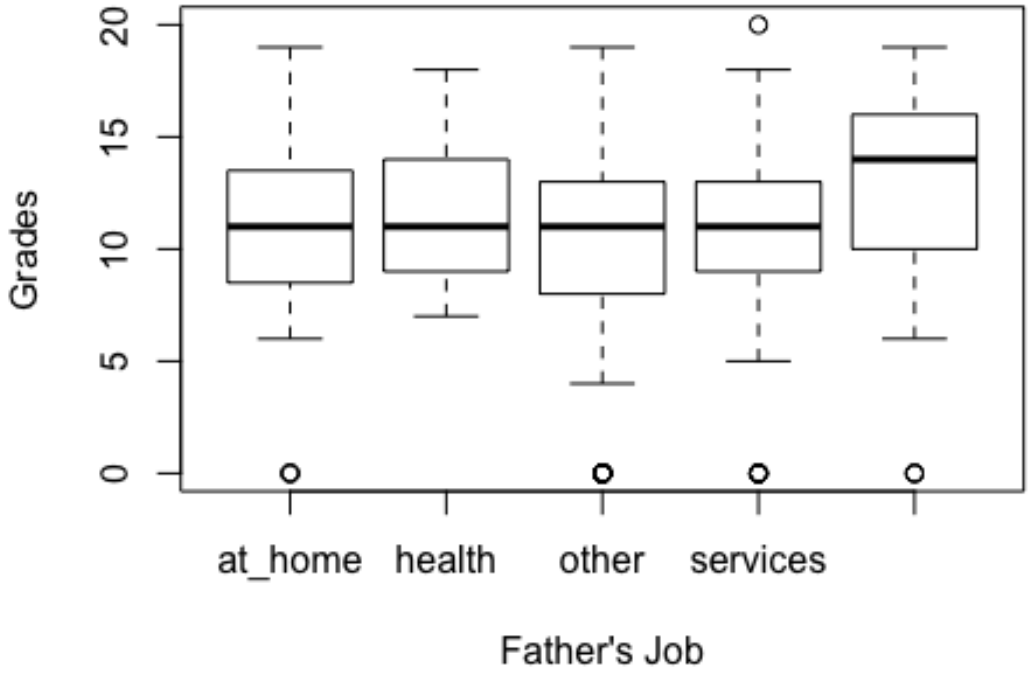
```r
summary(df_math[df_math$Mjob=="other",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  11.000   9.823  13.000  19.000
```

```r
summary(df_math[df_math$Mjob=="services",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   11.02   14.50   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   11.05   14.00   19.00

plot(Fjob, G3, xlab = "Father's Job", ylab = "Grades", main = "Figure 2.3")
```

Figure 2.3




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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.75   11.00   10.15   13.25   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00    9.00   11.00   11.61   14.00   18.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.19   13.00   19.00
```

```r
summary(df_math[df_math$Fjob=="services",]$G3)
```

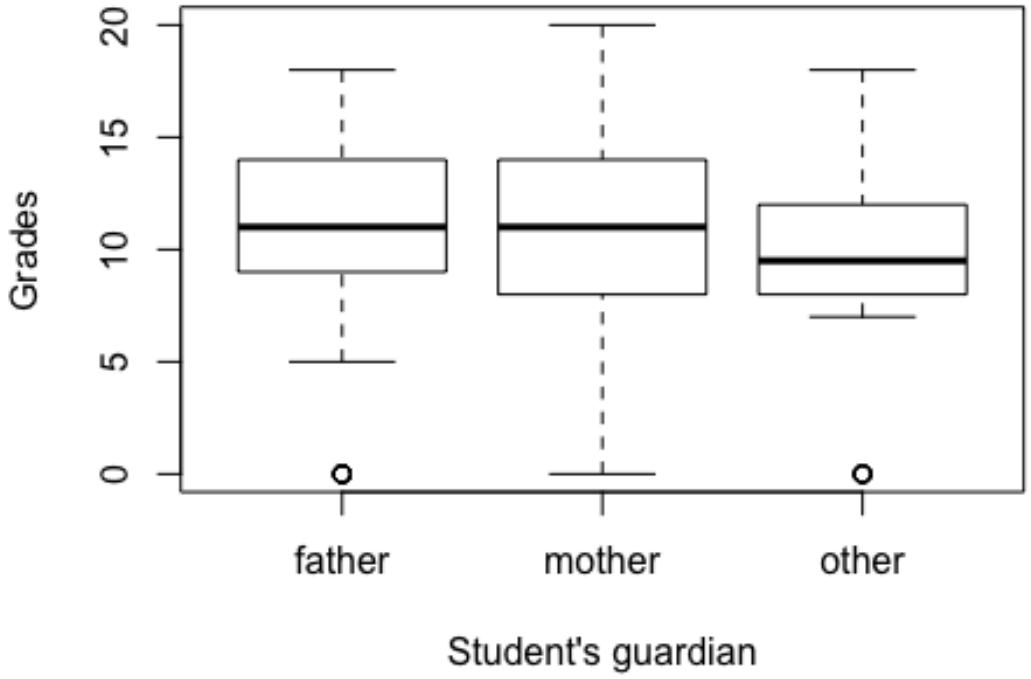
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     9.0    11.0    10.3    13.0    20.0
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   14.00   11.97   16.00   19.00
```

```r
plot(guardian,G3, xlab = "Student's guardian", ylab = "Grades", main =
"Figure 2.4")
```

Figure 2.4




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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.69   14.00   18.00
```

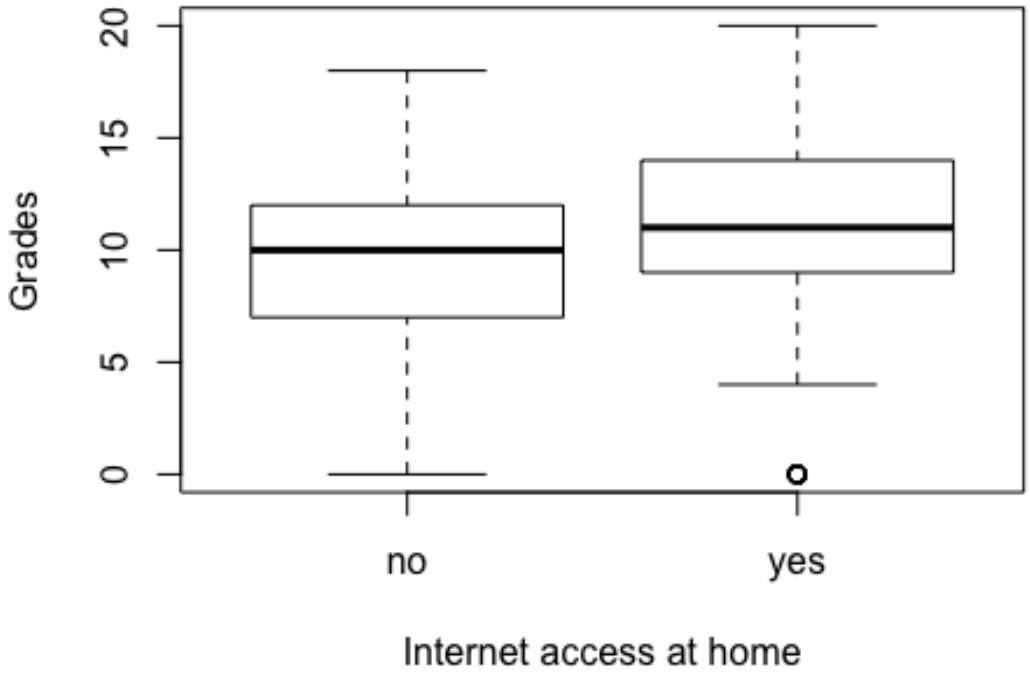
```r
summary(df_math[df_math$guardian=="mother",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.48   14.00   20.00
```

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

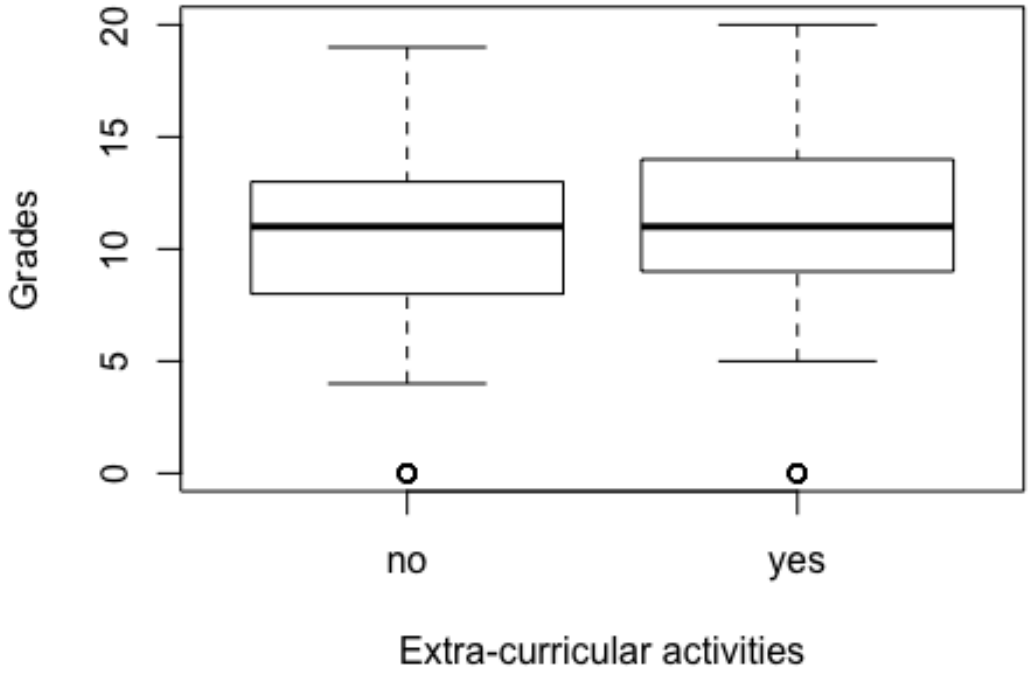
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000   9.500   9.062  12.000  18.000

plot(internet,G3, xlab = "Internet access at home", ylab = "Grades", main =
"Figure 2.5")
```

Figure 2.5



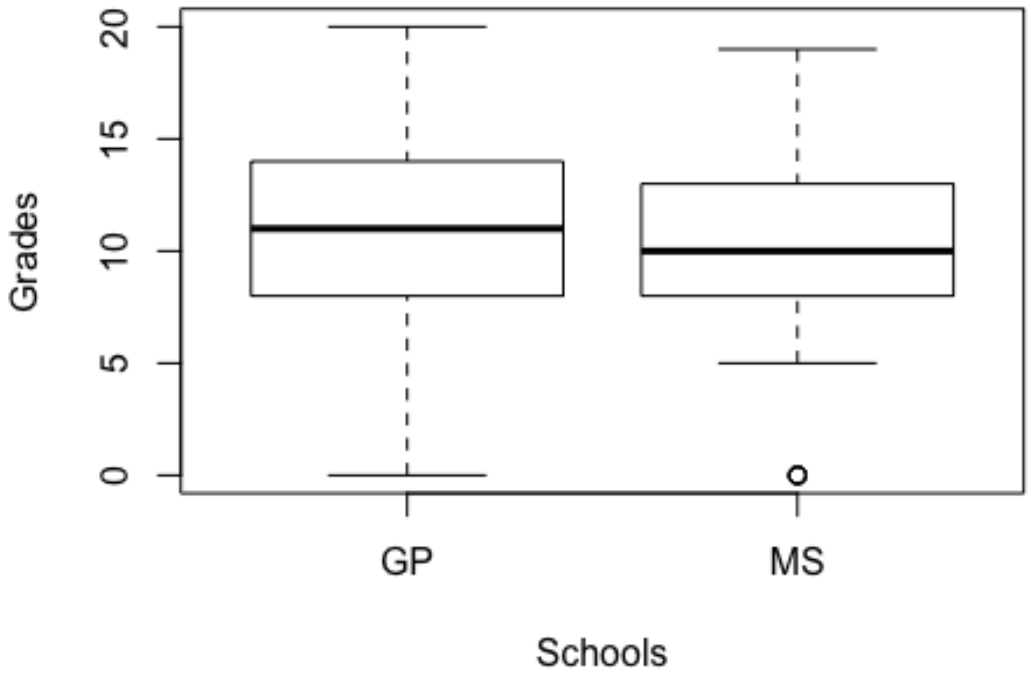

```
summary(df_math[df_math$internet=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.62   14.00   20.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.250  10.000   9.409  12.000  18.000

plot(activities,G3, xlab = "Extra-curricular activities", ylab = "Grades",
main = "Figure 2.6")
```

Figure 2.6



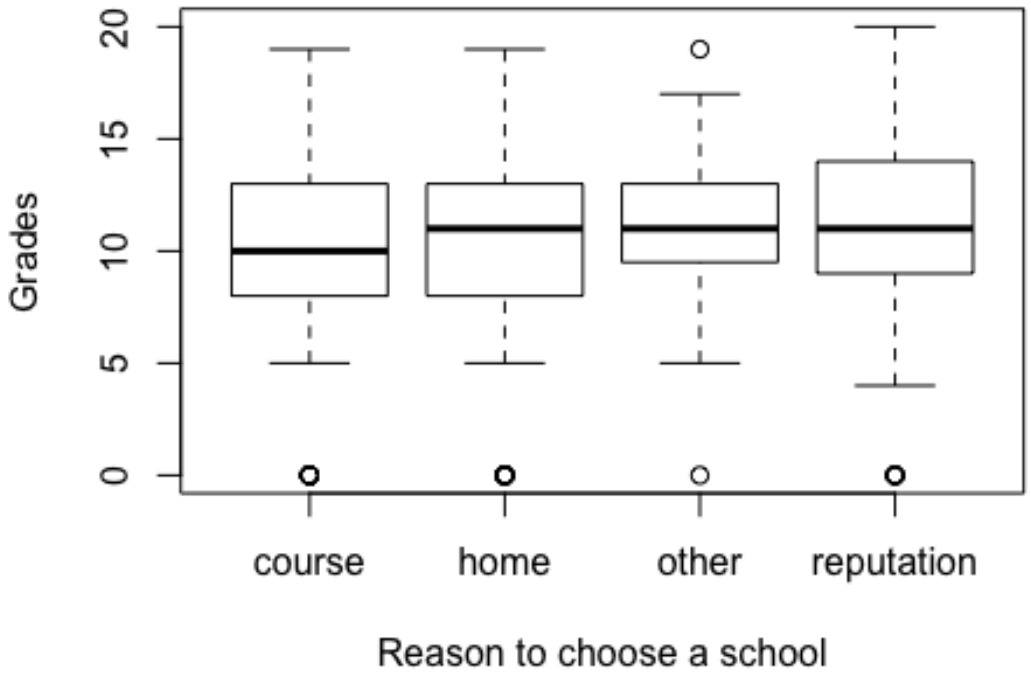

```
summary(df_math[df_math$activities=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.49   14.00   20.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.34   13.00   19.00

plot(school, G3, xlab = "Schools", ylab = "Grades", main = "Figure 2.7")
```

Figure 2.7



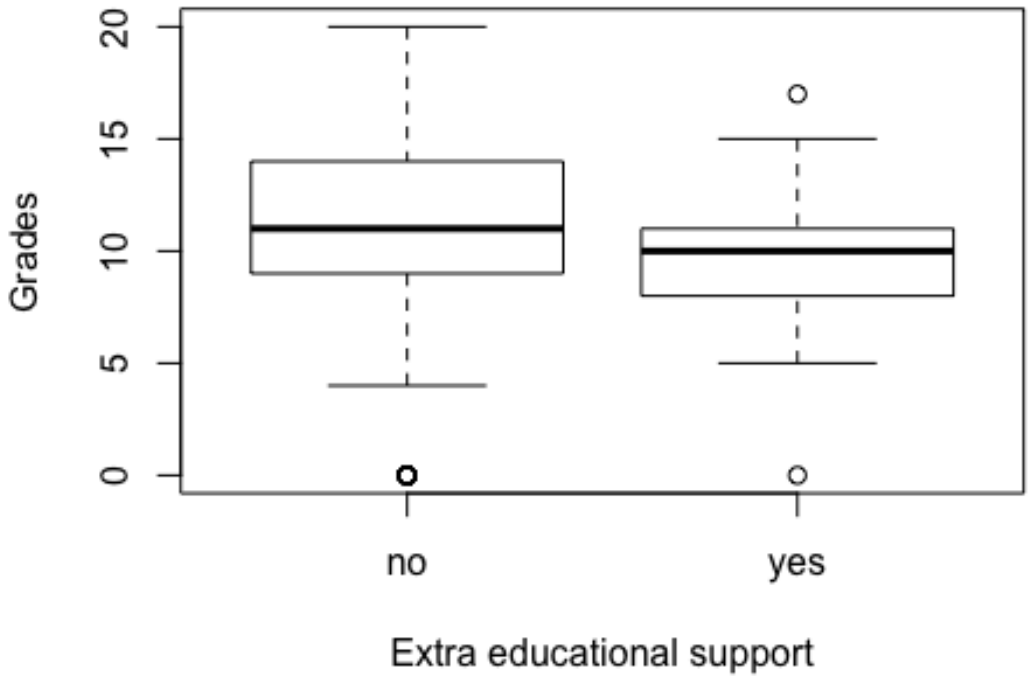

```
summary(df_math[df_math$school=="GP",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.49   14.00   20.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.848  12.750  19.000

plot(reason,G3, xlab = "Reason to choose a school", ylab = "Grades", main =
"Figure 2.8")
```

Figure 2.8




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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.821  13.000  19.000

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.26   13.00   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.75   11.00   11.17   13.00   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   11.14   14.00   20.00

plot(schoolsup,G3, xlab = "Extra educational support", ylab = "Grades", main
= "Figure 2.9")
```

Figure 2.9

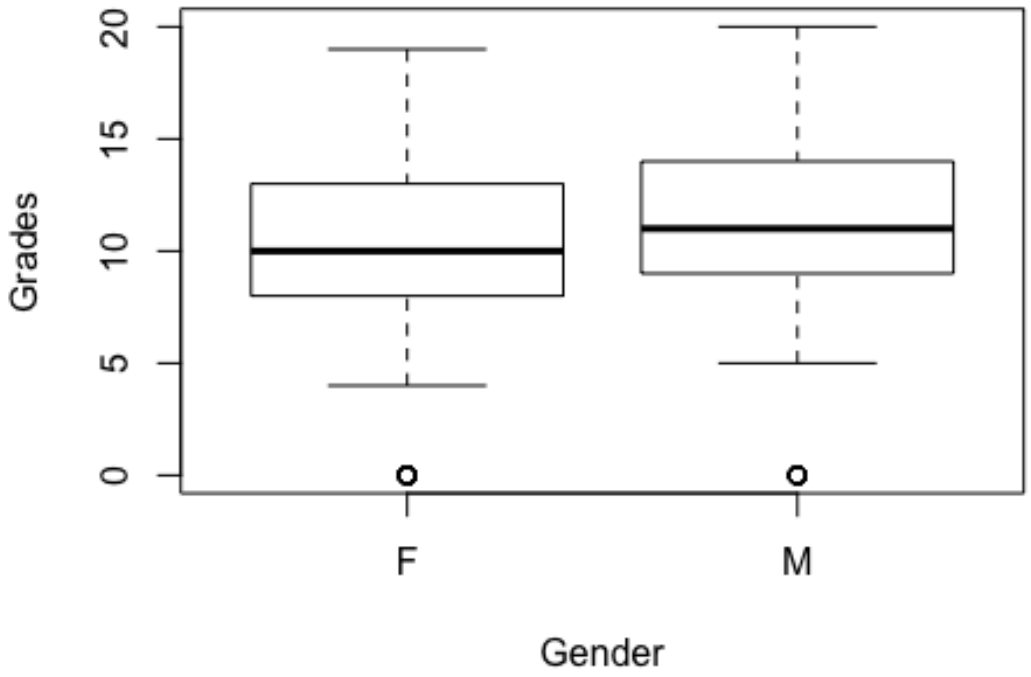



```
summary(df_math[df_math$schoolsup=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.431  11.000  17.000

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.56   14.00   20.00

plot(paid, G3, xlab = "Extra paid classes", ylab = "Grades", main = "Figure
2.10")
```

Figure 2.10

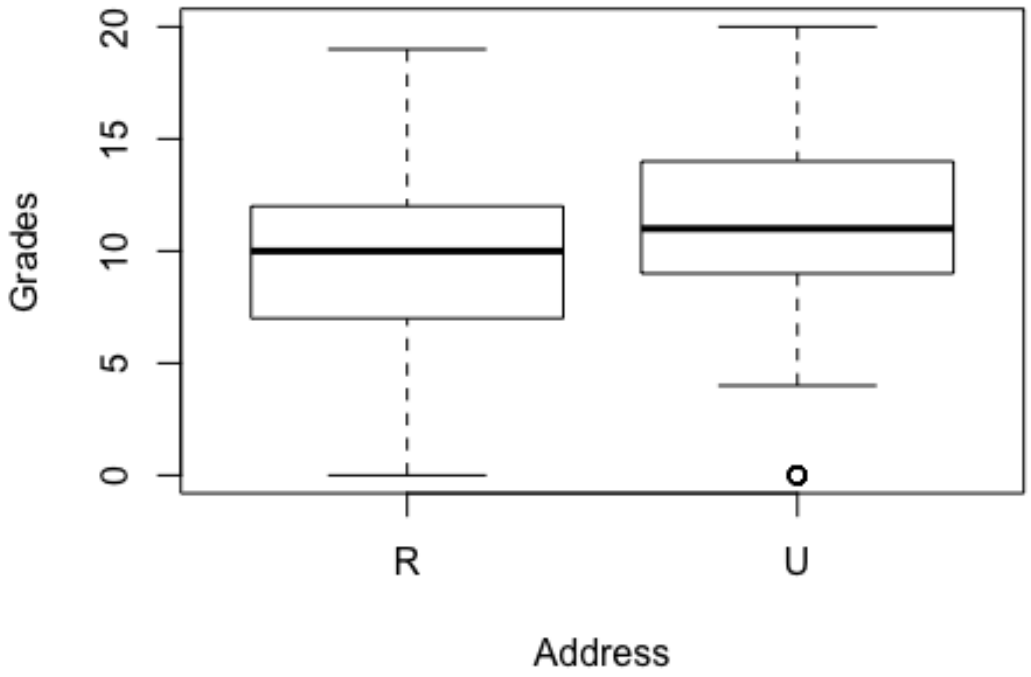

```
summary(df_math[df_math$paid=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.92   13.00   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  11.000   9.986  14.000  20.000

plot(sex,G3, xlab = "Gender", ylab = "Grades", main = "Figure 2.11")
```

Figure 2.11



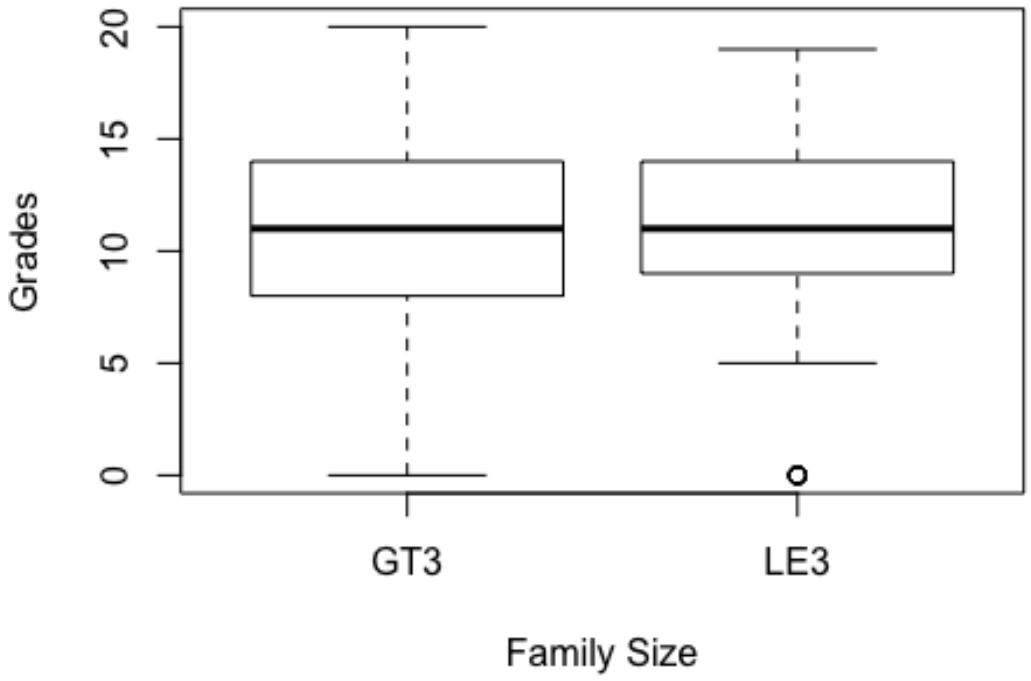

```
summary(df_math[df_math$sex=="F",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.966  13.000  19.000

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.91   14.00   20.00

plot(address,G3, xlab = "Address", ylab = "Grades", main = "Figure 2.12")
```

Figure 2.12



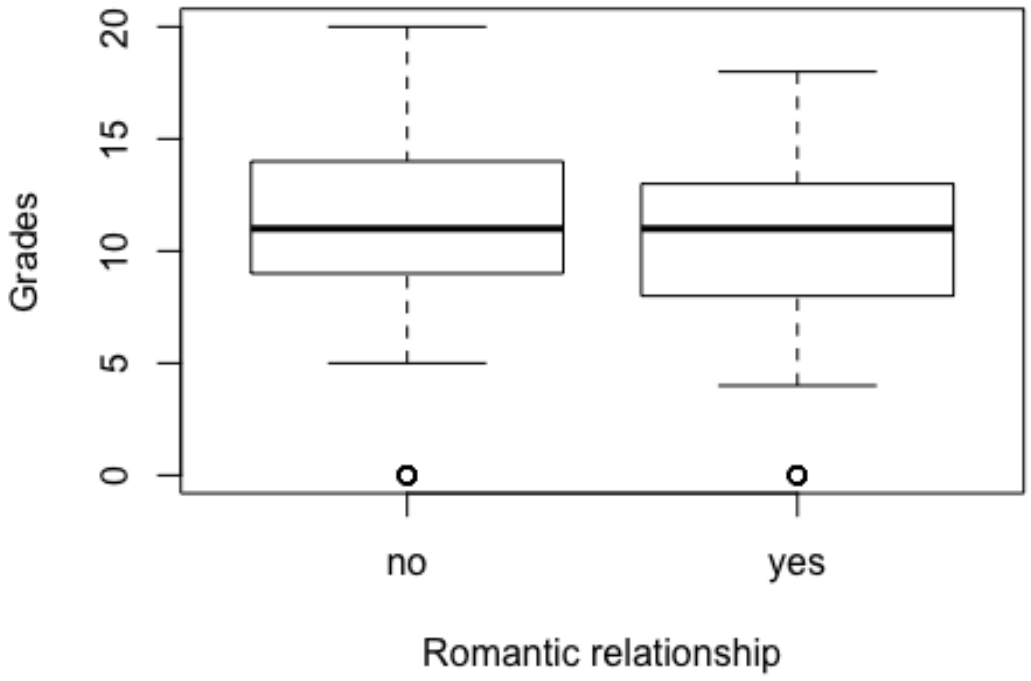

```
summary(df_math[df_math$address=="U",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.67   14.00   20.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.000  10.000   9.511  12.000  19.000

plot(famsize, G3, xlab = "Family Size", ylab = "Grades", main = "Figure
2.13")
```

Figure 2.13



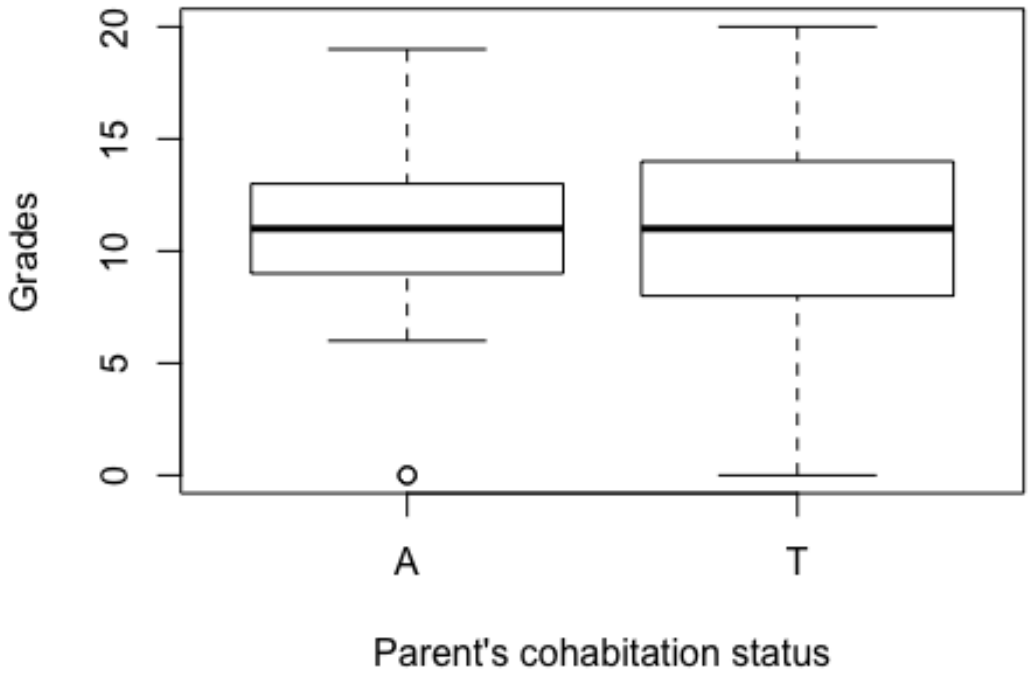

```
summary(df_math[df_math$famsize=="GT3",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.18   14.00   20.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   11.00   13.75   19.00

plot(romantic,G3, xlab = "Romantic relationship", ylab = "Grades", main =
"Figure 2.14" )
```

Figure 2.14



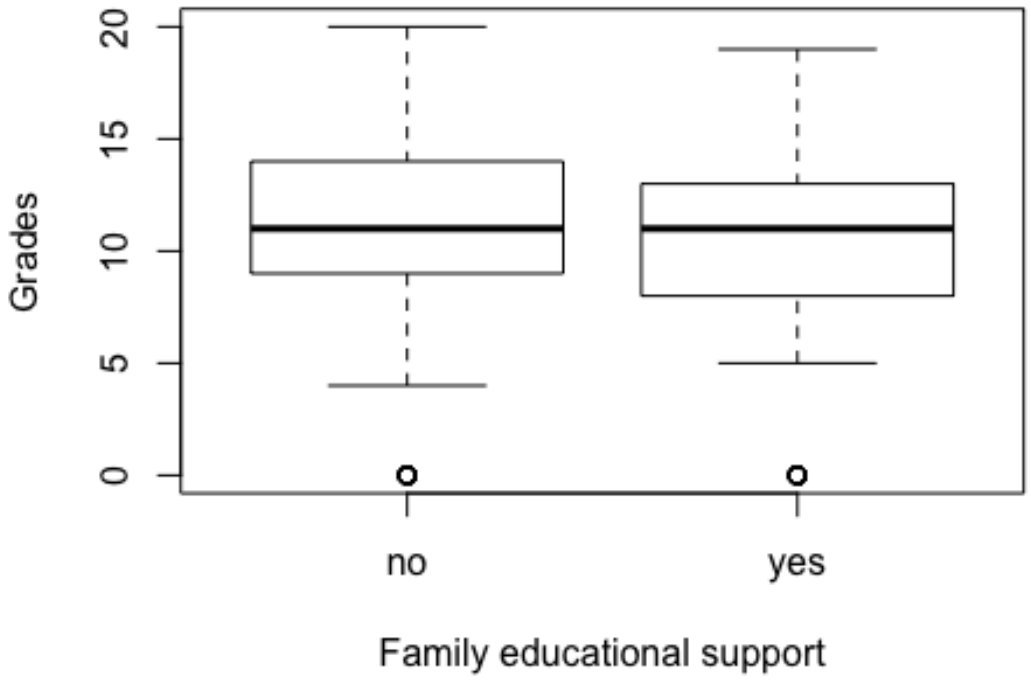

```
summary(df_math[df_math$romantic=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  11.000   9.576  13.000  18.000

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.84   14.00   20.00

plot(Pstatus,G3, xlab = "Parent's cohabitation status", ylab = "Grades", main
= "Figure 2.15")
```

Figure 2.15



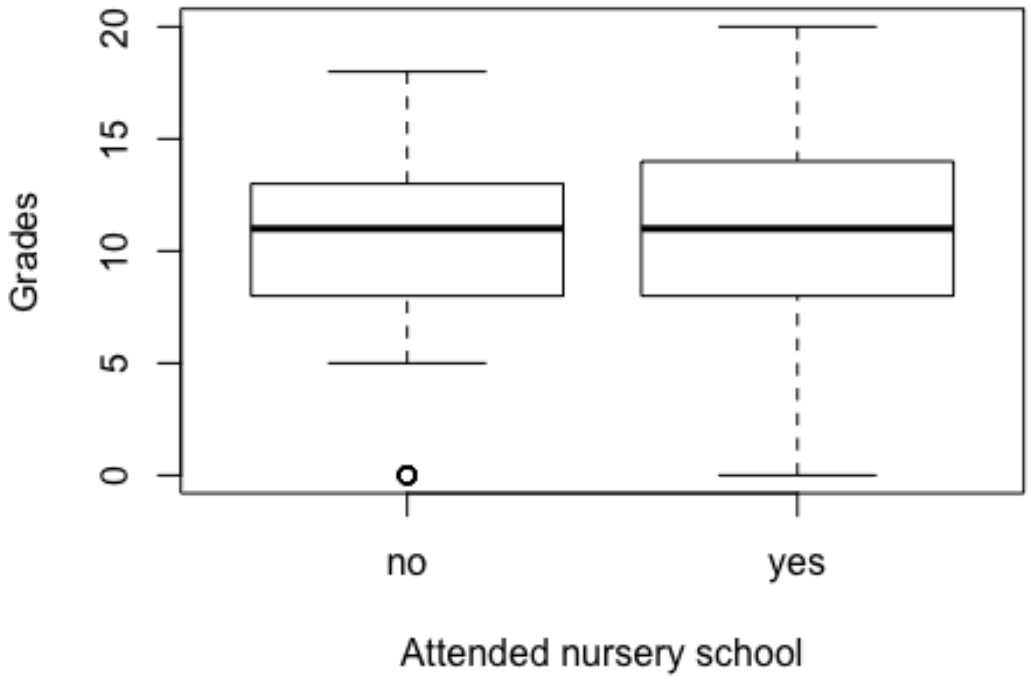

```
summary(df_math[df_math$Pstatus=="A",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     9.0    11.0    11.2    13.0    19.0

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.32   14.00   20.00

plot(famsup,G3, xlab = "Family educational support", ylab = "Grades", main =
"Figure 2.16")
```

Figure 2.16



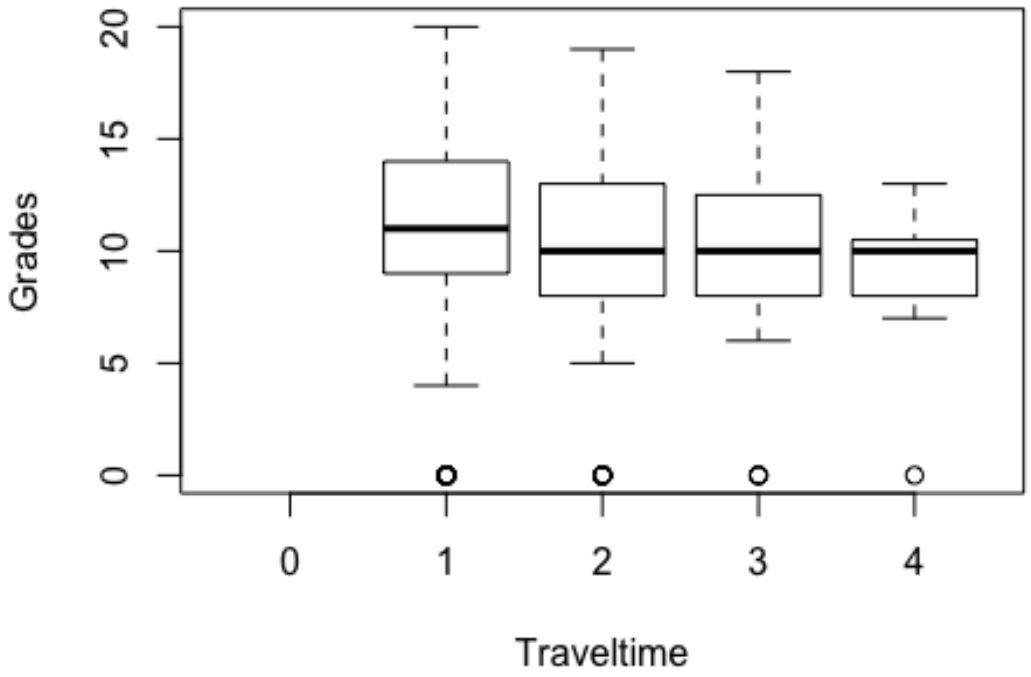

```
summary(df_math[df_math$famsup=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.27   13.00   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.64   14.00   20.00

plot(nursery,G3, xlab = "Attended nursery school", ylab = "Grades", main =
"Figure 2.17")
```

Figure 2.17



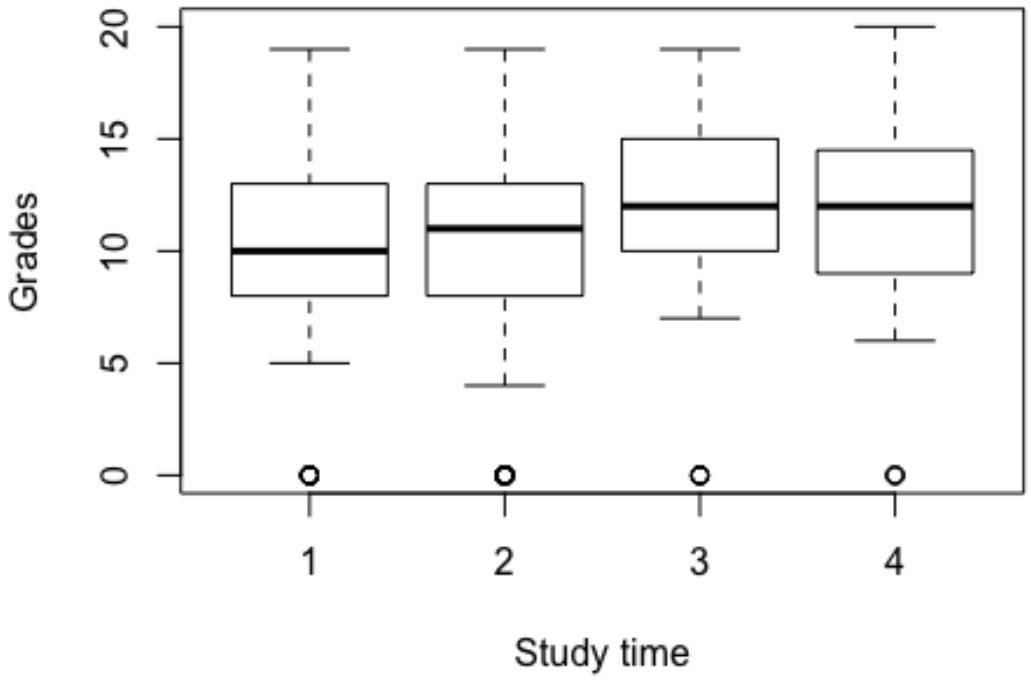

```
summary(df_math[df_math$activities=="yes",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.49   14.00   20.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.34   13.00   19.00

plot(traveltime,G3, xlab = "Traveltime", ylab = "Grades", main = "Figure
2.18")
```

Figure 2.18



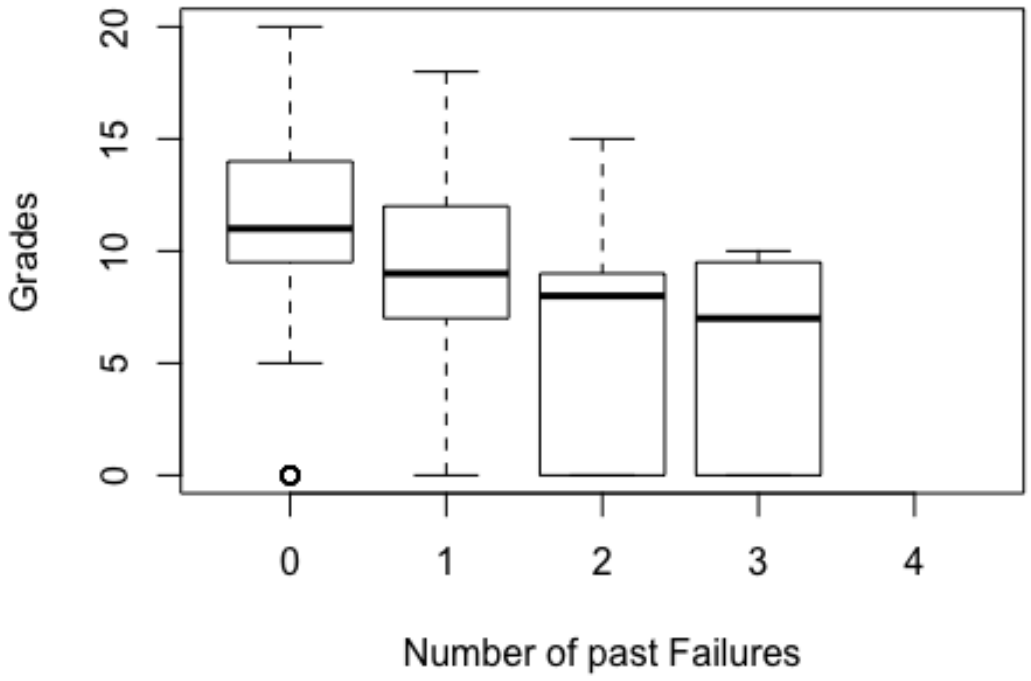

```
summary(df_math[df_math$traveltime=="1",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.78   14.00   20.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.907  13.000  19.000

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.261  12.500  18.000

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.50   10.00    8.75   10.25   13.00

plot(studytime,G3, xlab = "Study time", ylab = "Grades", main = "Figure
2.19")
```

Figure 2.19



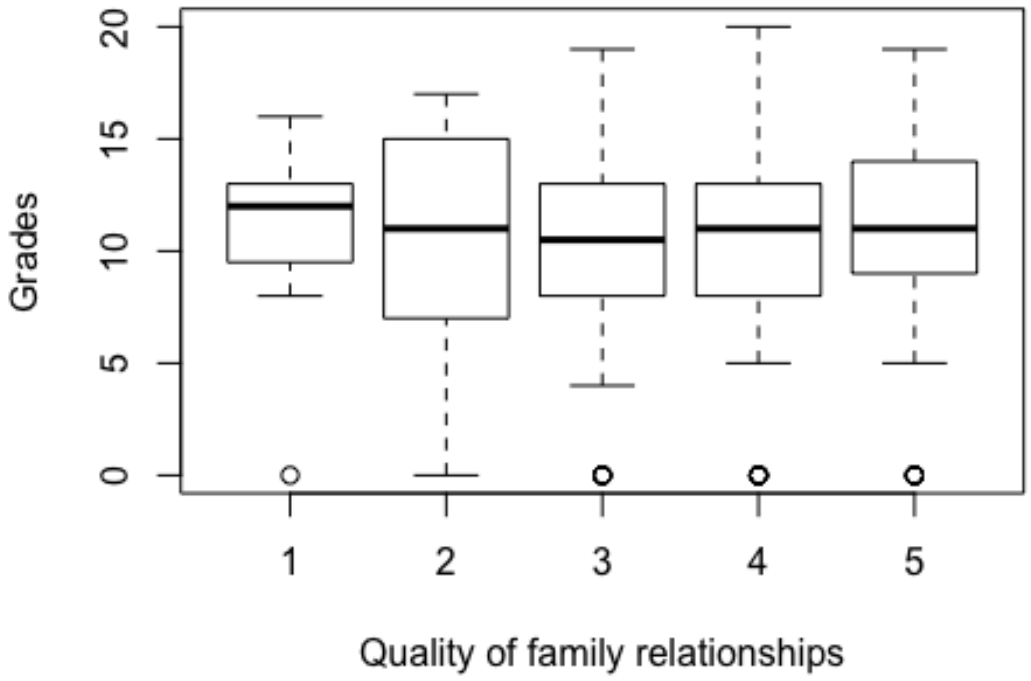

```
summary(df_math[df_math$studytime=="1",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   10.00   10.05   13.00   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.17   13.00   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    10.0    12.0    11.4    15.0    19.0

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   12.00   11.26   14.50   20.00

plot(failures,G3, xlab = "Number of past Failures", ylab = "Grades", main =
"Figure 2.20")
```

Figure 2.20



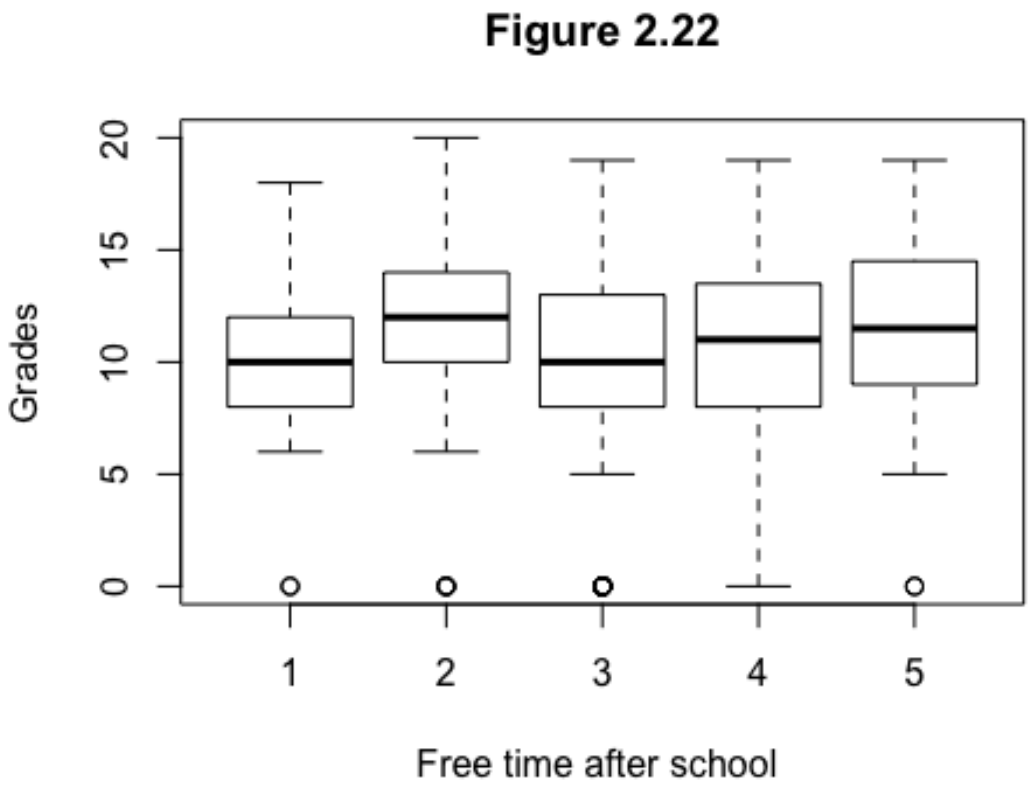

```
summary(df_math[df_math$failures=="0",]$G3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.75   11.00   11.25   14.00   20.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.00    9.00    8.12   11.75   18.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   8.000   6.235   9.000  15.000

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   7.000   5.688   9.250  10.000

plot(famrel,G3, xlab = "Quality of family relationships", ylab = "Grades",
main = "Figure 2.21")
```

Figure 2.21




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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.25   12.00   10.62   13.00   16.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.250  11.000   9.889  14.500  17.000

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   10.50   10.04   13.00   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.36   13.00   20.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.83   14.00   19.00
```

```r
plot(freetime,G3, xlab = "Free time after school ", ylab = "Grades", main =
"Figure 2.22")
```




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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.842  12.000  18.000
```

```r
summary(df_math[df_math$freetime=="2",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.56   14.00   20.00
```

```r
summary(df_math[df_math$freetime=="3",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.783  13.000  19.000
```

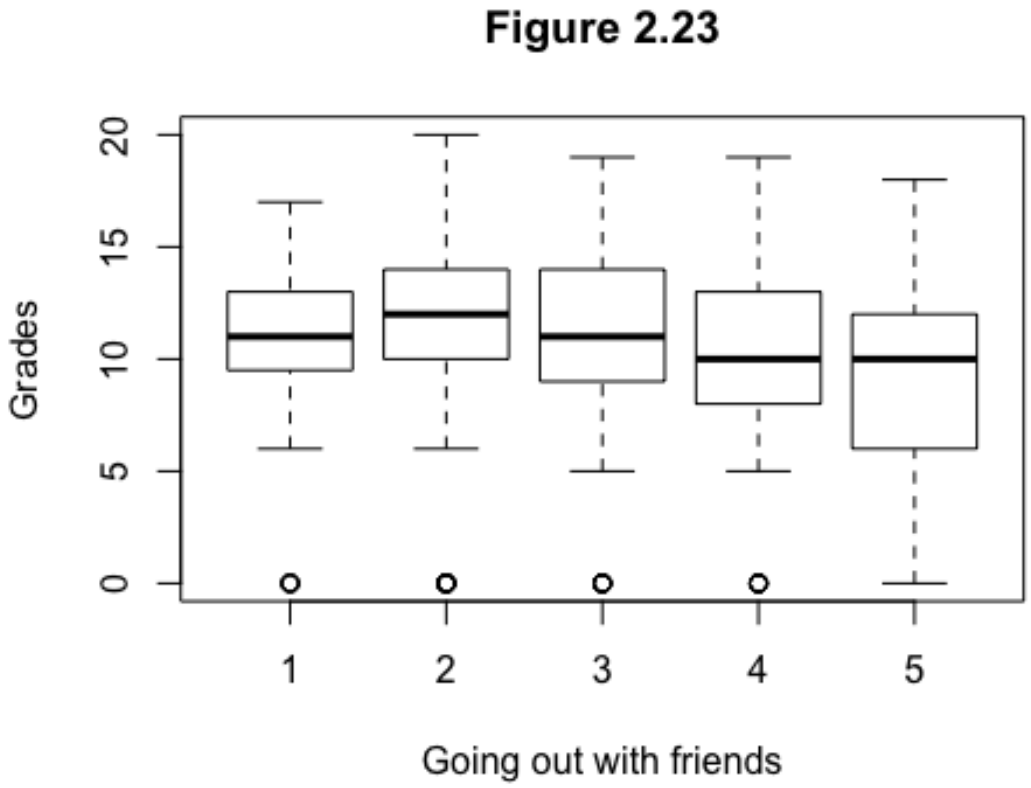
```r
summary(df_math[df_math$freetime=="4",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.43   13.50   19.00
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.50   11.30   14.25   19.00
```

```
plot(goout,G3, xlab = "Going out with friends", ylab = "Grades", main =
"Figure 2.23")
```




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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.50   11.00    9.87   13.00   17.00
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   12.00   11.19   14.00   20.00
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.96   14.00   19.00
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.651  13.000  19.000
```

```r
summary(df_math[df_math$goout=="5",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.000  10.000   9.038  12.000  18.000
```

```r
plot(Dalc,G3, xlab = "Workday alcohol consumption", ylab = "Grades", main =
"Figure 2.24")
```

Figure 2.24


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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   11.00   10.73   14.00   20.00
```

```r
summary(df_math[df_math$Dalc=="2",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.253  12.000  18.000
```

```r
summary(df_math[df_math$Dalc=="3",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   10.00   10.50   12.75   17.00
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   9.000   9.000   9.889  12.000  13.000
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   10.00   11.00   10.67   13.00   13.00
```

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

Figure 2.25


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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.50   11.00   10.74   14.50   20.00
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.08   14.00   19.00
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   10.00   10.72   13.00   18.00
```

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

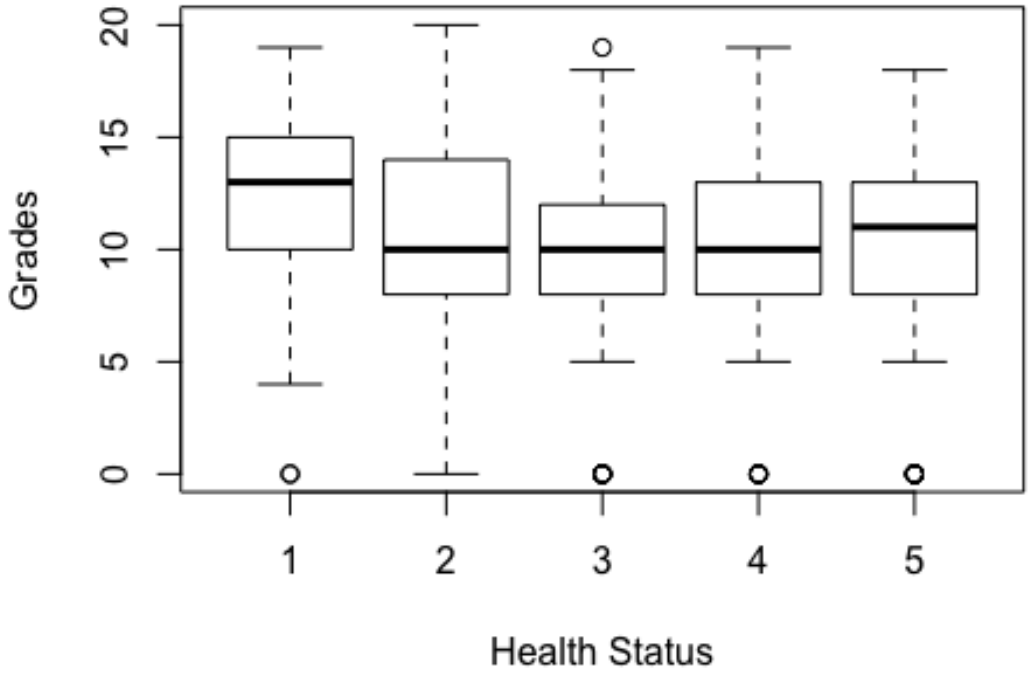
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.686  12.000  17.000
```

```r
summary(df_math[df_math$Walc=="5",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   10.00   10.14   13.00   18.00
```

```r
plot(health,G3, xlab = "Health Status", ylab = "Grades", main = "Figure
2.26")
```

Figure 2.26




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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   13.00   11.87   15.00   19.00
```

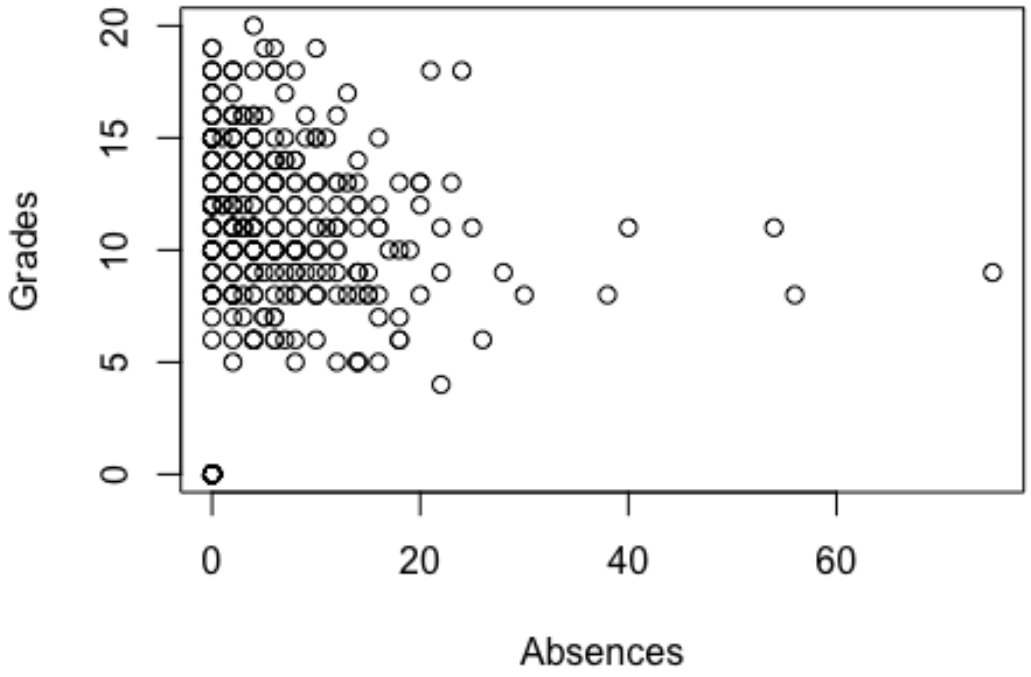
```r
summary(df_math[df_math$health=="2",]$G3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   10.00   10.22   14.00   20.00
```

```r
summary(df_math[df_math$health=="3",]$G3)
```

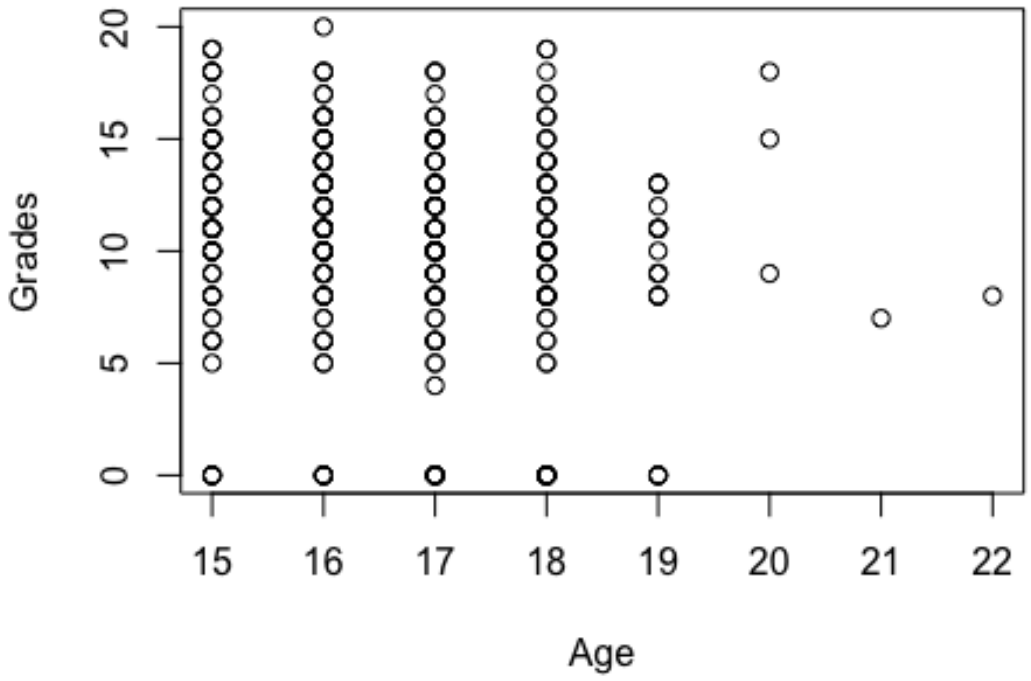
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   10.00   10.01   12.00   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   10.00   10.11   13.00   19.00

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

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     8.0    11.0    10.4    13.0    18.0

# Creating Scatter plots for numerical data
plot(df_math$absences,df_math$G3, xlab = "Absences", ylab = "Grades", main =
"Figure 2.27")
```

Figure 2.27



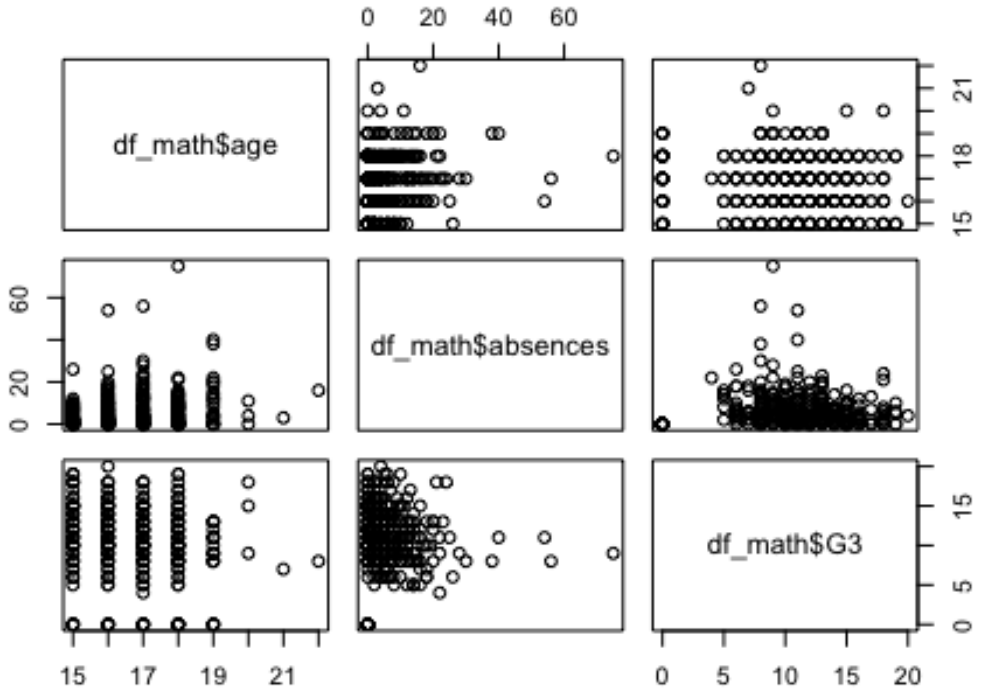

```
plot(df_math$age,df_math$G3, xlab = "Age", ylab = "Grades", main = "Figure
2.28")
```

Figure 2.28




```
pairs(~df_math$age+df_math$absences+df_math$G3, main = "Figure 2.29")
```

Figure 2.29




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

set.seed(1)
train = sample(1:nrow(df_math), 320)
actual_g3 = df_math[-train,31]

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


### Subset Selection
# Stepwise Selection
# Linear Model

full_model_fit <- lm(G3~.,data = df_math[train,])
summary(full_model_fit)

##
## Call:
## lm(formula = G3 ~ ., data = df_math[train, ])
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -10.5860  -1.8884   0.2391   2.6444   8.0591
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.37356    5.11660   2.223  0.02712 *
## schoolMS          0.65789    0.94091   0.699  0.48507
## sexM              1.15858    0.59665   1.942  0.05328 .
## age              -0.20451    0.25397  -0.805  0.42144
## addressU          0.12356    0.70627   0.175  0.86126
## famsizeLE3        0.97671    0.57385   1.702  0.08999 .
## PstatusT         -0.81746    0.85598  -0.955  0.34050
## Medu.L           -1.52919    1.82857  -0.836  0.40380
## Medu.Q            2.23057    1.46727   1.520  0.12972
## Medu.C           -1.16472    0.97779  -1.191  0.23471
## Medu^4            0.94051    0.61918   1.519  0.13003
## Fedu.L           -1.46270    2.05986  -0.710  0.47831
## Fedu.Q            1.10225    1.72715   0.638  0.52393
## Fedu.C           -0.54347    1.12109  -0.485  0.62826
## Fedu^4           -0.17967    0.60142  -0.299  0.76539
## Mjobhealth        2.12008    1.40387   1.510  0.13226
## Mjobother        -0.54948    0.80732  -0.681  0.49674
## Mjobservices      0.69879    0.92062   0.759  0.44854
## Mjobteacher      -1.59562    1.27357  -1.253  0.21142
## Fjobhealth       -0.36757    1.62357  -0.226  0.82108
## Fjobother        -0.08255    1.20039  -0.069  0.94523
## Fjobservices     -0.11539    1.21530  -0.095  0.92443
## Fjobteacher       1.53708    1.51282   1.016  0.31060
## reasonhome        0.26257    0.63376   0.414  0.67900
## reasonother       0.81177    0.91259   0.890  0.37458
## reasonreputation  0.91837    0.69779   1.316  0.18934
## guardianmother   -0.42092    0.64031  -0.657  0.51154
## guardianother     0.46071    1.14483   0.402  0.68771
## traveltime.L     -0.94323    1.24677  -0.757  0.45004
## traveltime.Q     -0.99117    1.05577  -0.939  0.34874
## traveltime.C     -1.20528    0.86244  -1.398  0.16349
## studytime.L       0.92732    0.78637   1.179  0.23942
## studytime.Q      -1.05931    0.65868  -1.608  0.10904
## studytime.C      -0.74552    0.54513  -1.368  0.17266
## failures.L       -2.99928    0.94420  -3.177  0.00168 **
## failures.Q        1.73921    0.98321   1.769  0.07813 .
## failures.C        0.18844    0.99219   0.190  0.84953
## schoolsupyes     -0.77841    0.78885  -0.987  0.32471
## famsupyes        -1.07931    0.55002  -1.962  0.05084 .
## paidyes           0.46739    0.57421   0.814  0.41644
## activitiesyes    -0.42149    0.50490  -0.835  0.40463
## nurseryyes       -0.26205    0.62421  -0.420  0.67499
## higheryes         1.26431    1.18850   1.064  0.28845
## internetyes       0.29083    0.73336   0.397  0.69202
## romanticyes      -1.35548    0.54899  -2.469  0.01422 *
```

```
## famrel.L          0.08489    1.27556   0.067  0.94699
## famrel.Q          0.25468    1.13228   0.225  0.82222
## famrel.C          0.08684    1.04378   0.083  0.93376
## famrel^4         -0.03926    0.79678  -0.049  0.96075
## freetime.L        1.10101    1.01953   1.080  0.28122
## freetime.Q        1.09158    0.83651   1.305  0.19312
## freetime.C        0.58974    0.68170   0.865  0.38781
## freetime^4       -0.48997    0.51717  -0.947  0.34435
## goout.L          -0.43631    0.93275  -0.468  0.64036
## goout.Q          -1.53998    0.79207  -1.944  0.05299 .
## goout.C           1.73408    0.65516   2.647  0.00864 **
## goout^4           0.02533    0.49002   0.052  0.95882
## Dalc.L           -1.31993    1.56210  -0.845  0.39893
## Dalc.Q            0.60406    1.19176   0.507  0.61270
## Dalc.C            0.37700    1.16197   0.324  0.74587
## Dalc^4            0.95203    1.07762   0.883  0.37784
## Walc.L            1.72659    1.07680   1.603  0.11010
## Walc.Q            1.17585    0.80871   1.454  0.14720
## Walc.C            0.38949    0.68004   0.573  0.56733
## Walc^4            0.55933    0.58026   0.964  0.33601
## health.L         -0.78715    0.67047  -1.174  0.24150
## health.Q          1.24852    0.63774   1.958  0.05137 .
## health.C         -0.96898    0.65268  -1.485  0.13890
## health^4          0.09574    0.58221   0.164  0.86951
## absences          0.05328    0.03228   1.651  0.10008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.038 on 250 degrees of freedom
## Multiple R-squared:  0.3861, Adjusted R-squared:  0.2167
## F-statistic: 2.279 on 69 and 250 DF,  p-value: 1.94e-06

# Backward AIC
library(leaps)
backward_aic_fit = MASS::stepAIC(full_model_fit, direction = "backward",
trace = FALSE)
backward_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 + Pstatus + Mjob + studytime + failures + famsup + romantic +
```

```
##     freetime + goout + absences
##
##
##            Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                                  250   4076.247 954.2756
## 2      - famrel  4  5.233160       254   4081.481 946.6861
## 3        - Fedu  4 18.984935       258   4100.466 940.1711
## 4        - Fjob  4 42.390463       262   4142.856 935.4623
## 5        - Dalc  4 50.448908       266   4193.305 931.3355
## 6        - Medu  4 46.262989       270   4239.568 926.8466
## 7  - traveltime  3 25.641283       273   4265.209 922.7762
## 8        - Walc  4 52.516915       277   4317.726 918.6922
## 9    - guardian  2 19.240726       279   4336.967 916.1150
## 10     - health  4 82.566743       283   4419.534 914.1499
## 11     - school  1  2.942378       284   4422.476 912.3629
## 12   - internet  1  3.775709       285   4426.252 910.6360
## 13       - paid  1  5.509443       286   4431.761 909.0340
## 14    - nursery  1  6.563109       287   4438.324 907.5076
## 15        - age  1  6.364914       288   4444.689 905.9661
## 16 - activities  1  5.518295       289   4450.207 904.3632
## 17    - address  1 10.692323       290   4460.900 903.1311
## 18  - schoolsup  1 19.100651       291   4480.000 902.4984
## 19    - famsize  1 21.536479       292   4501.537 902.0330
## 20     - higher  1 26.299165       293   4527.836 901.8971
## 21     - reason  3 80.071689       296   4607.908 901.5066
```

```
summary(backward_aic_fit)
```

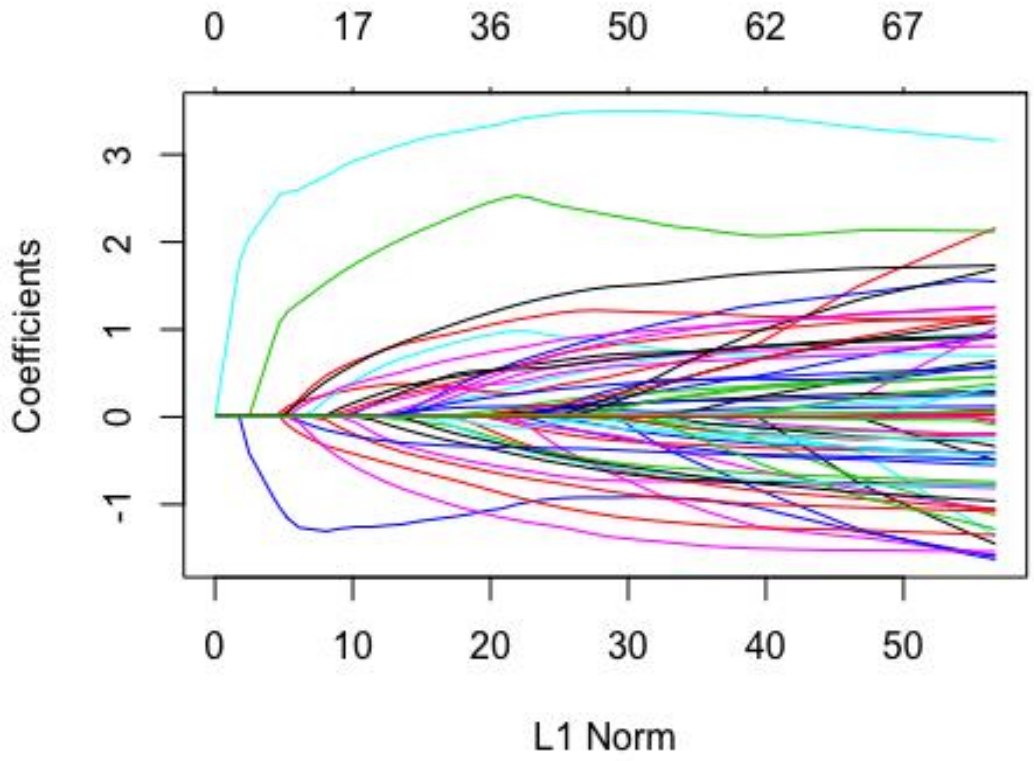
```
##
## Call:
## lm(formula = G3 ~ sex + Pstatus + Mjob + studytime + failures +
##     famsup + romantic + freetime + goout + absences, data = df_math[train,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2211  -1.9812   0.1642   2.7946   7.9781
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.92834    1.09295   8.169 9.11e-15 ***
## sexM          0.86646    0.51130   1.695  0.09120 .
## PstatusT     -1.12437    0.77544  -1.450  0.14812
## Mjobhealth    3.20104    0.99078   3.231  0.00137 **
## Mjobother    -0.23765    0.68764  -0.346  0.72989
## Mjobservices  1.24358    0.72849   1.707  0.08886 .
## Mjobteacher  -0.30811    0.85323  -0.361  0.71828
## studytime.L   0.87973    0.68863   1.278  0.20243
## studytime.Q  -0.81192    0.59721  -1.360  0.17502
## studytime.C  -0.56955    0.47995  -1.187  0.23631
```

```
## failures.L   -3.37409    0.81577  -4.136 4.61e-05 ***
## failures.Q    2.05074    0.83749   2.449  0.01492 *
## failures.C   -0.01937    0.87278  -0.022  0.98231
## famsupyes    -1.05749    0.48283  -2.190  0.02929 *
## romanticyes  -1.19695    0.49006  -2.442  0.01517 *
## freetime.L    0.78757    0.91421   0.861  0.38967
## freetime.Q    1.08514    0.76493   1.419  0.15706
## freetime.C    0.35184    0.61168   0.575  0.56560
## freetime^4   -0.83026    0.46503  -1.785  0.07522 .
## goout.L      -0.06599    0.80150  -0.082  0.93443
## goout.Q      -1.61272    0.70899  -2.275  0.02364 *
## goout.C       1.70817    0.58412   2.924  0.00372 **
## goout^4       0.09673    0.44974   0.215  0.82986
## absences      0.04833    0.02811   1.719  0.08664 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.946 on 296 degrees of freedom
## Multiple R-squared:  0.306,  Adjusted R-squared:  0.2521
## F-statistic: 5.675 on 23 and 296 DF,  p-value: 1.196e-13
```

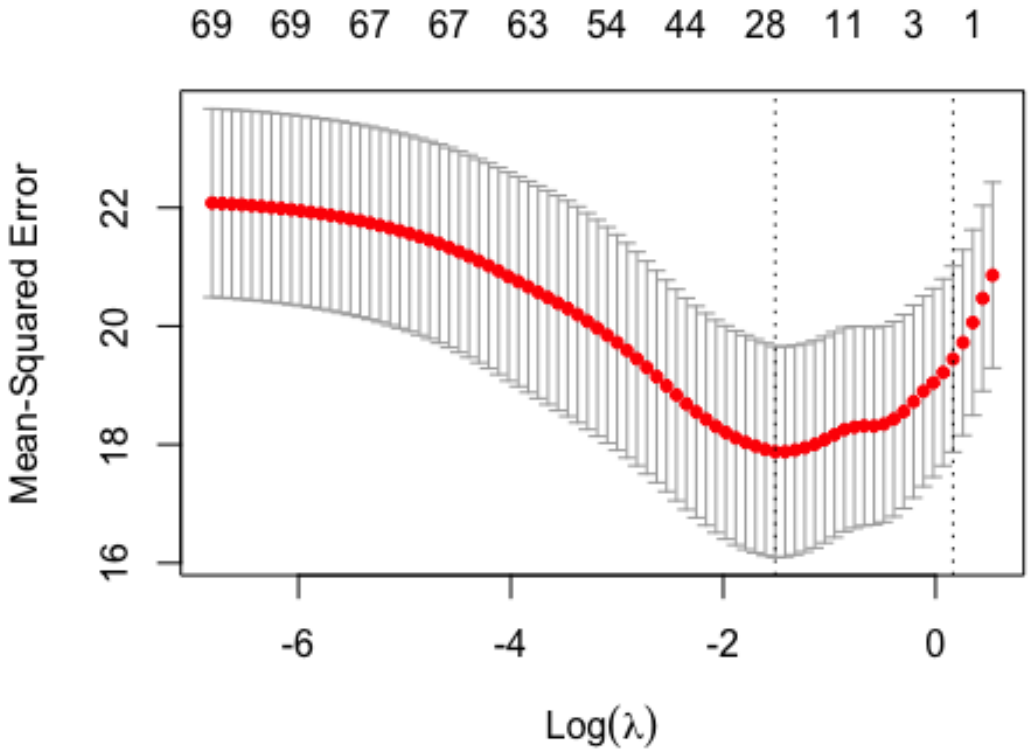
```
backward_aic_pred = predict(backward_aic_fit, newdata = df_math[-train,1:30])
mean((backward_aic_pred-actual_g3)^2)
```

```
## [1] 21.85308
```

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

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

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

lasso_mod = glmnet(x_train,
                   y_train,
                   alpha = 1) # Fit lasso model on training data

plot(lasso_mod)    # Draw plot of coefficients
```




```r
set.seed(1)
cv.out = cv.glmnet(x_train, y_train, alpha = 1) # Fit lasso model on training data
plot(cv.out) # Draw plot of training MSE as a function of lambda
```




```
best_lambda = cv.out$lambda.min # Select lamda that minimizes training MSE
lasso_pred = predict(lasso_mod, s = best_lambda, newx = x_test) # Use best
lambda to predict test data
mean((lasso_pred - y_test)^2) # Calculate test MSE
```

```
## [1] 20.28559
```

```
lasso_best <- glmnet(x_train, y_train, alpha = 1, lambda = best_lambda)
coef(lasso_best)
```

```
## 72 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## (Intercept)       8.455645174
## schoolMS          .
## sexM              0.299358675
## age               .
## addressU          .
## famsizeLE3        0.112234374
## PstatusT         -0.376687614
## Medu.L            .
## Medu.Q            0.369604152
## Medu.C            .
## Medu^4            .
## Fedu.L            .
```

```
## Fedu.Q            .
## Fedu.C            .
## Fedu^4            .
## Mjobhealth        2.127988558
## Mjobother         .
## Mjobservices      0.735031677
## Mjobteacher       .
## Fjobhealth        .
## Fjobother         .
## Fjobservices      .
## Fjobteacher       0.140756790
## reasonhome        .
## reasonother       0.069050470
## reasonreputation  0.388551582
## guardianmother    .
## guardianother     .
## traveltime.L      .
## traveltime.Q      .
## traveltime.C      .
## traveltime^4      .
## studytime.L       0.356476058
## studytime.Q       .
## studytime.C      -0.057850338
## failures.L       -1.205404763
## failures.Q        3.166106933
## failures.C        .
## failures^4        .
## schoolsupyes     -0.191050033
## famsupyes        -0.452459426
## paidyes           .
## activitiesyes     .
## nurseryyes        .
## higheryes         0.563982876
## internetyes       .
## romanticyes      -0.626745797
## famrel.L          .
## famrel.Q          0.071738036
## famrel.C          .
## famrel^4          .
## freetime.L        .
## freetime.Q        0.875188261
## freetime.C        .
## freetime^4       -0.309710624
## goout.L           .
## goout.Q          -0.886401091
## goout.C           0.923590573
## goout^4           .
## Dalc.L            .
## Dalc.Q            .
## Dalc.C            .
```

```
## Dalc^4            0.258062645
## Walc.L            .
## Walc.Q            .
## Walc.C            .
## Walc^4            0.087220553
## health.L          .
## health.Q          0.149656366
## health.C         -0.068064336
## health^4          .
## absences          0.004981549
```

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

```
library(ISLR)
library(tree)
```

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

```
library(MASS)
```

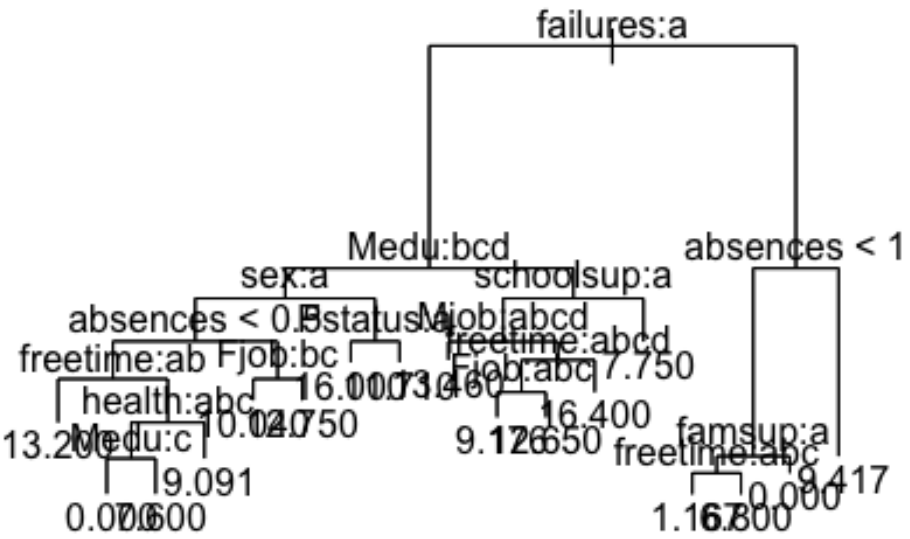
```
##
## Attaching package: 'MASS'
```

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

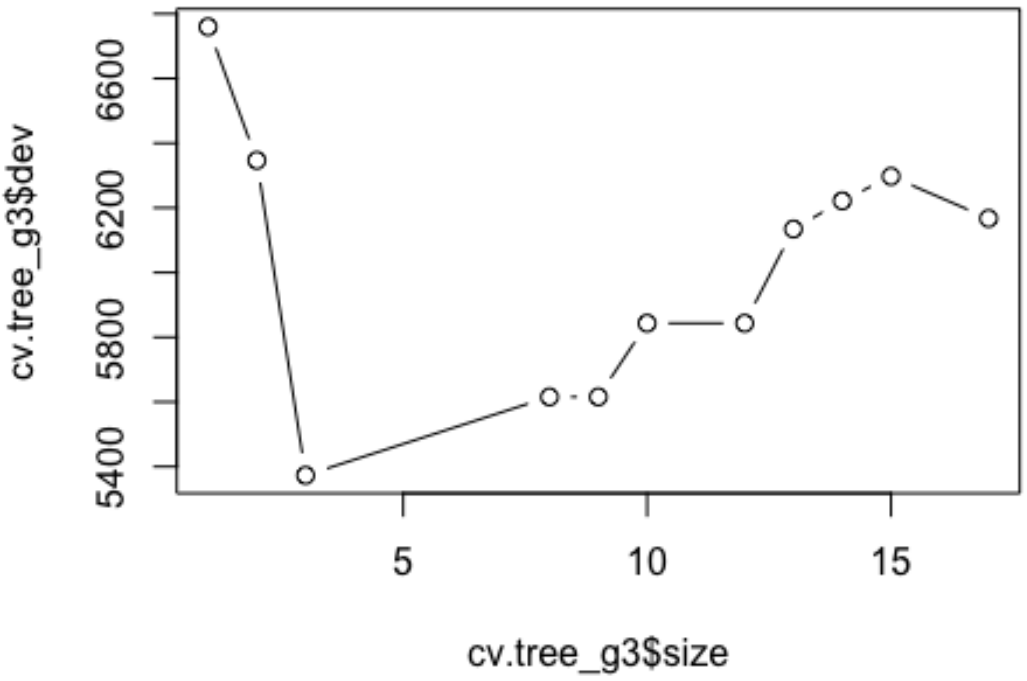
```
tree_g3 = tree(G3~., data = df_math , subset = train)
summary(tree_g3)
```

```
##
## Regression tree:
## tree(formula = G3 ~ ., data = df_math, subset = train)
## Variables actually used in tree construction:
##  [1] "failures"  "Medu"      "sex"       "absences"  "freetime"  "health"
##  [7] "Fjob"      "Pstatus"   "schoolsup" "Mjob"      "famsup"
## Number of terminal nodes:  17
## Residual mean deviance:  10.68 = 3235 / 303
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -13.460  -1.710   0.000   0.000   1.957   7.957
```

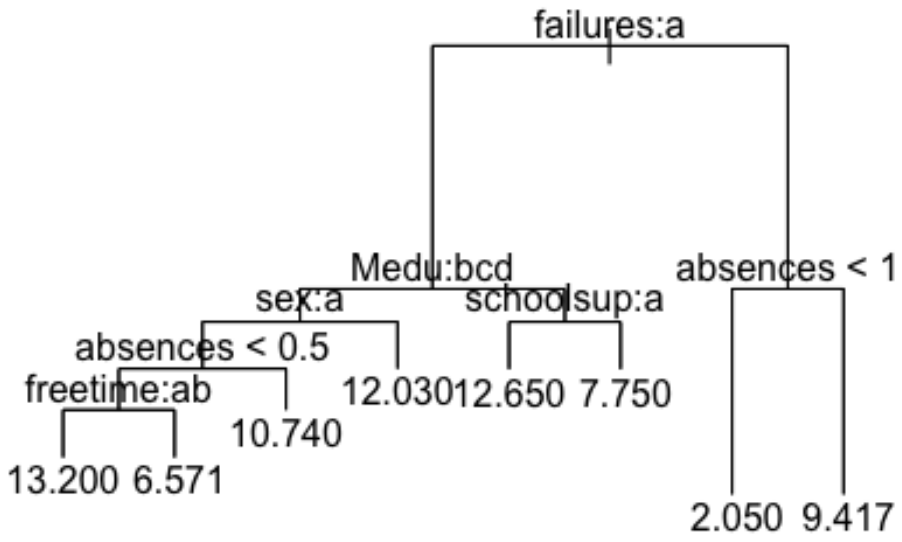
```
plot(tree_g3)
text(tree_g3)
```



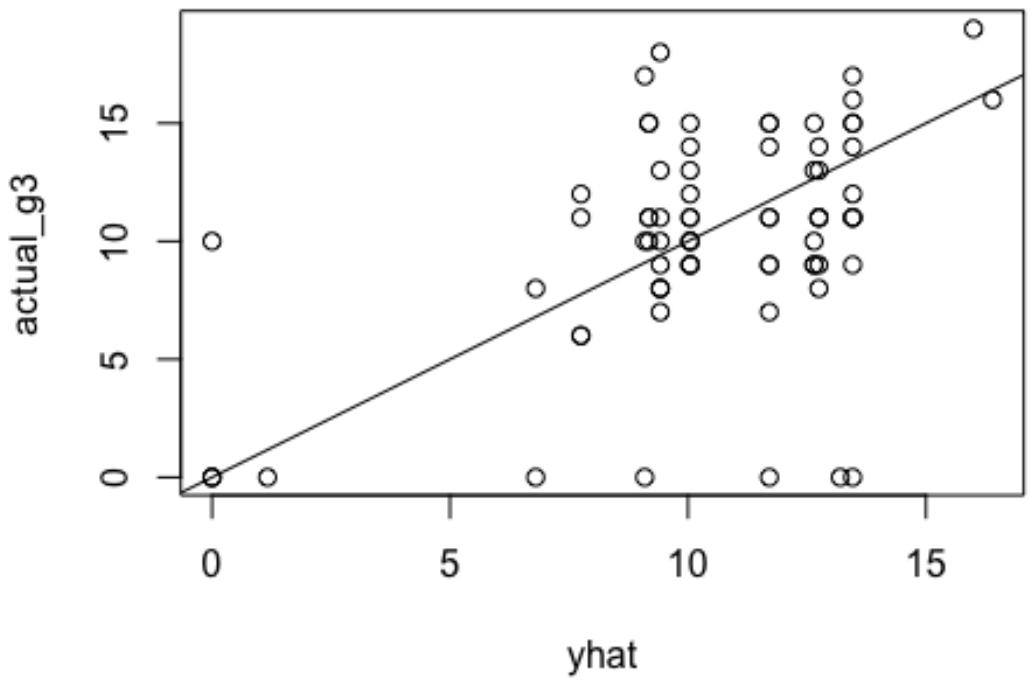

```
cv.tree_g3 = cv.tree(tree_g3)
plot(cv.tree_g3$size, cv.tree_g3$dev, type = 'b')
```



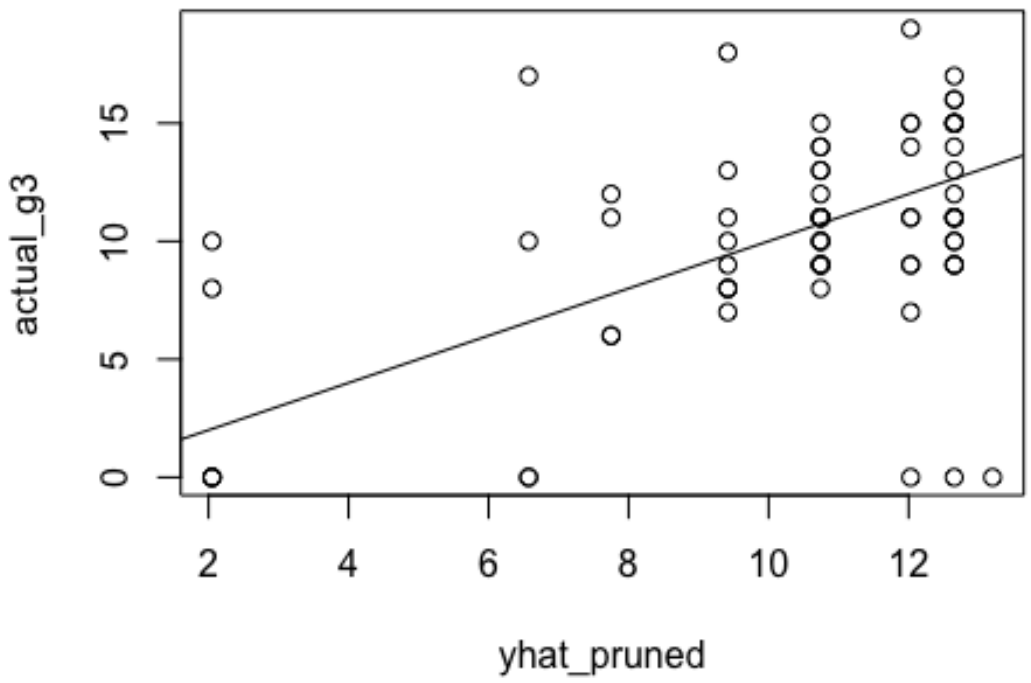

```
prune.tree_g3 = prune.tree(tree_g3, best = 4)
plot(prune.tree_g3)
text(prune.tree_g3)
```




```
yhat = predict(tree_g3, newdata = df_math[-train,1:30])
plot(yhat, actual_g3)
abline(0,1)
```




```
mean((yhat-actual_g3)^2)

## [1] 17.0883

yhat_pruned = predict(prune.tree_g3, newdata = df_math[-train,1:30])
plot(yhat_pruned, actual_g3)
abline(0,1)
```




```
mean((yhat_pruned-actual_g3)^2)
```

```
## [1] 16.86808
```

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

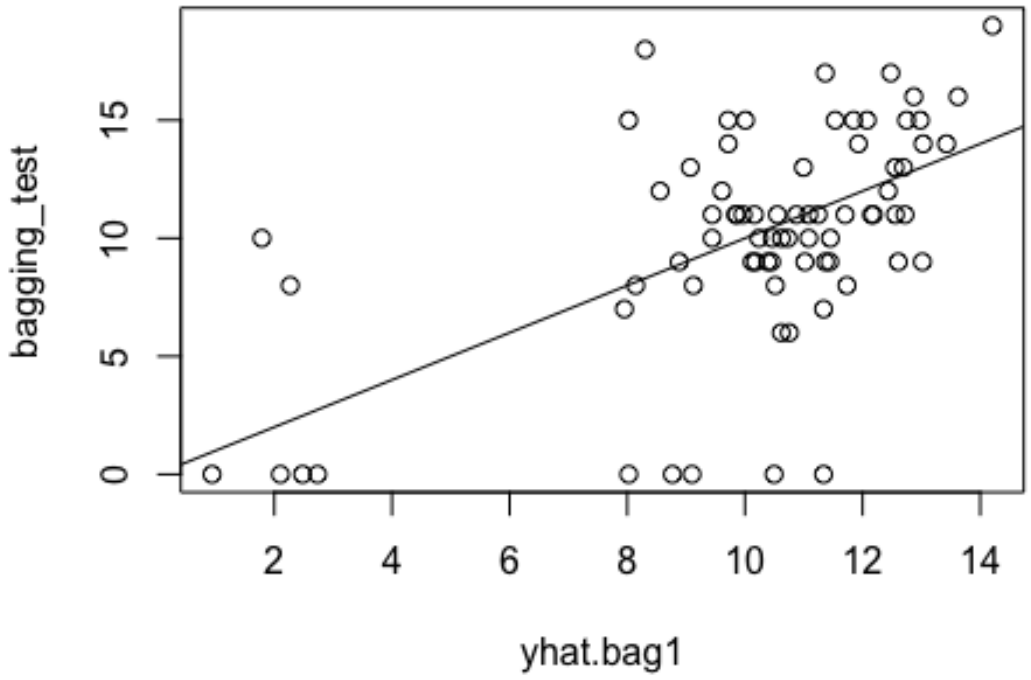
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
# We are performing bagging - by considering all the predictors i.e. mtry =
30
```

```r
set.seed(-1)
bagging_g3 = randomForest(G3~., data = df_math[train,], mtry = 30, ntree=
1000, importance = TRUE)
bagging_g3
```

```
##
## Call:
##  randomForest(formula = G3 ~ ., data = df_math[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: 15.51685
##                     % Var explained: 25.22
```

```r
yhat.bag1 = predict(bagging_g3, newdata = df_math[-train,1:30])
bagging_test = df_math[-train,"G3"]
plot(yhat.bag1, bagging_test)
abline(0,1)
```



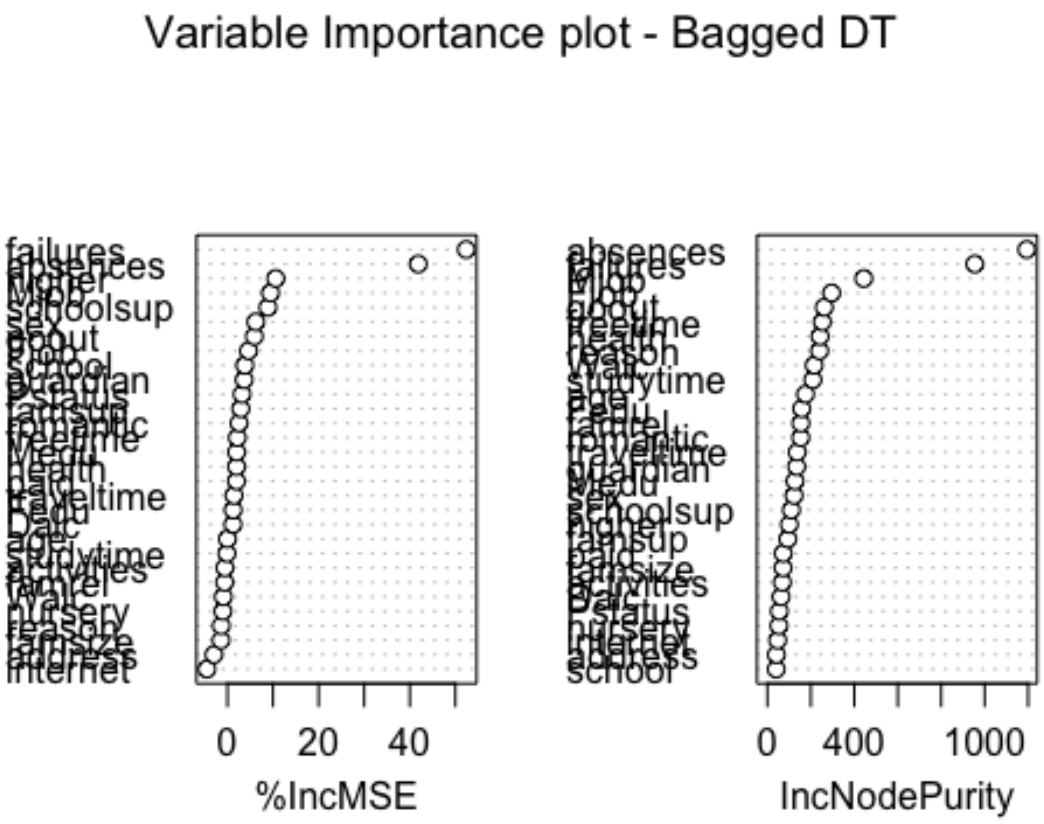

```r
mean((yhat.bag1-bagging_test)^2)
```

```
## [1] 15.01827
```

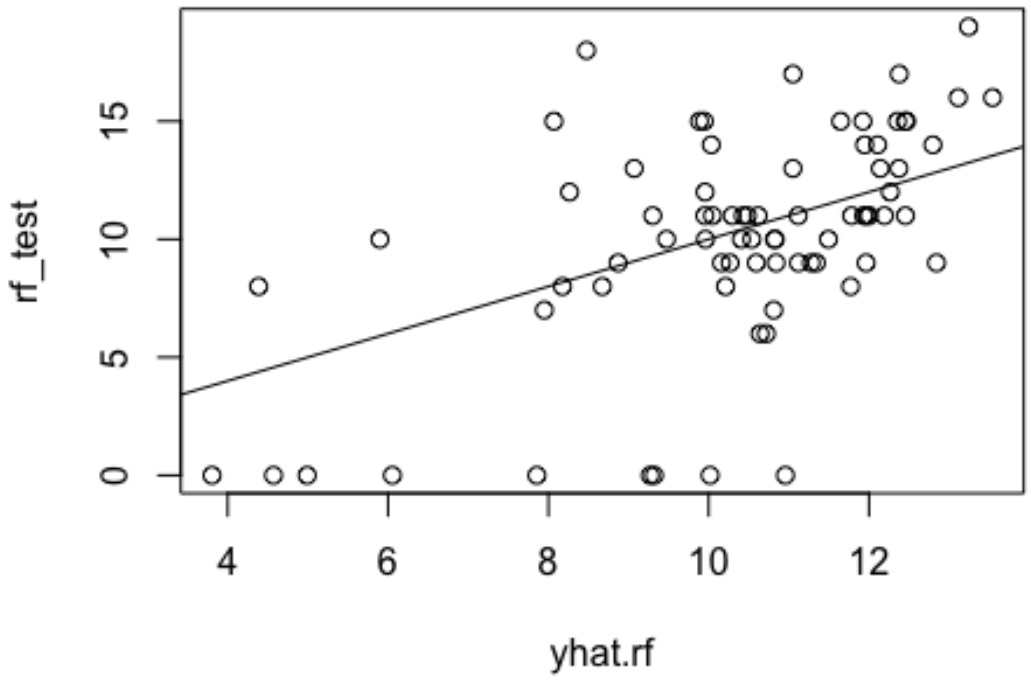
```r
importance(bagging_g3)
```

```
##                %IncMSE IncNodePurity
## school      3.82397290      40.49613
## sex         6.23140629     122.98733
## age         0.01575303     175.54999
## address    -2.96528408      41.32419
## famsize    -1.39511289      71.92401
## Pstatus     3.20968961      54.92950
## Medu        2.03416169     127.30552
## Fedu        1.27188014     158.26101
## Mjob        9.46257318     443.68111
## Fjob        4.63089555     295.82755
## reason     -1.34263759     240.42183
## guardian    3.68187814     134.83284
## traveltime  1.44992378     137.21260
## studytime  -0.10877801     209.41492
## failures   52.39339174     954.87230
## schoolsup   8.89207111     110.37334
## famsup      2.93470021      94.11851
## paid        1.99683443      72.86714
## activities -0.54920171      68.93737
## nursery    -0.99070078      53.22999
## higher     10.60895891     100.86497
## internet   -4.49754632      47.28975
## romantic    2.55600586     153.66563
## famrel     -0.59428748     154.20489
## freetime    2.17528816     252.81159
## goout       5.91166719     261.65538
## Dalc        1.23922397      62.95308
## Walc       -0.89603917     215.08513
## health      2.01666148     242.26011
## absences   41.86074096    1193.79785
```

```r
varImpPlot(bagging_g3, main = "Variable Importance plot - Bagged DT")
```



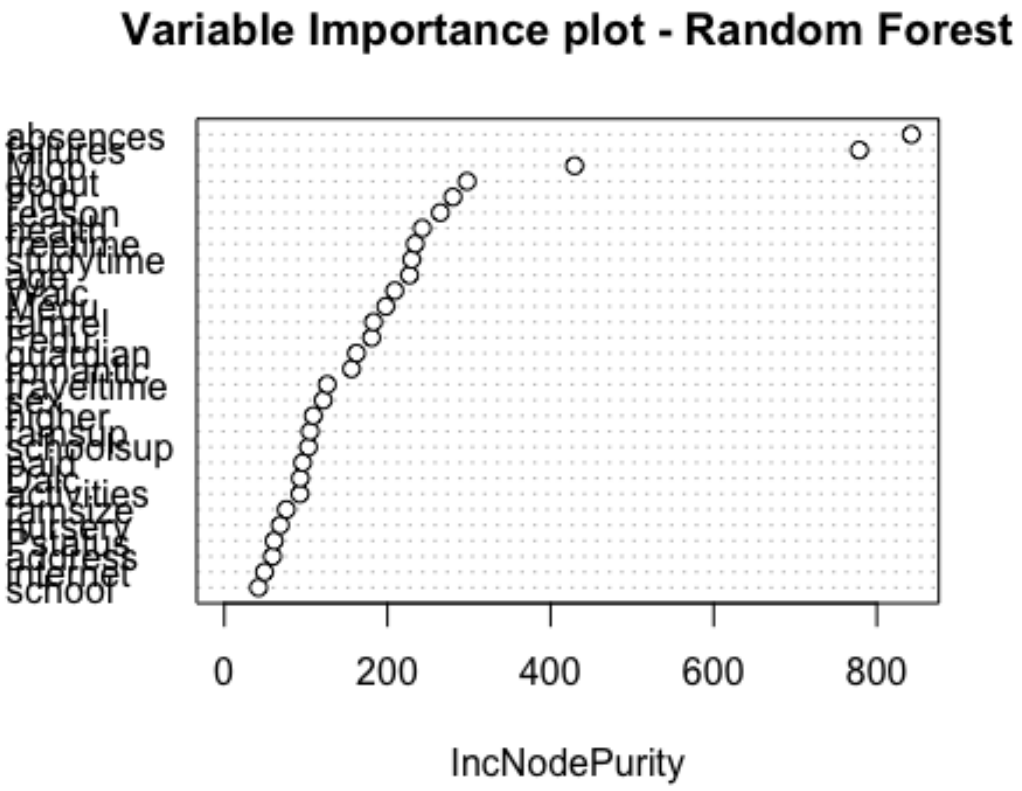

```
# RF - that is with m != p. mtry = p/3
set.seed(-1)
rf_g3 = randomForest(G3~., data = df_math[train,], mtry = 10, ntree= 1000,
importance = FALSE)
rf_g3

##
## Call:
##  randomForest(formula = G3 ~ ., data = df_math[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: 15.44804
##                     % Var explained: 25.55

yhat.rf = predict(rf_g3, newdata = df_math[-train,])
rf_test = df_math[-train,"G3"]
plot(yhat.rf, rf_test)
abline(0,1)
```




```
mean((yhat.rf-rf_test)^2)
```

```
## [1] 14.9953
```

```
varImpPlot(rf_g3, main = "Variable Importance plot - Random Forest")
```




```
#################################  RESULTS
##########################################

cat("\n\n Model Performance : \n\n")

##
##
##  Model Performance :

cat("RMSE of Backward Step wise : ", sqrt(mean((backward_aic_pred-
actual_g3)^2)),"\n")

## RMSE of Backward Step wise :  4.674728

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

## RMSE of Lasso :  4.503953

cat("RMSE of Decision Tree : ", sqrt(mean((yhat_pruned-actual_g3)^2)),"\n")

## RMSE of Decision Tree :  4.107076

cat("RMSE of Bagged Decision Trees : ", sqrt(mean((yhat.bag1-
bagging_test)^2)),"\n")
```

```
## RMSE of Bagged Decision Trees :  3.875341
```

```
cat("RMSE of RF : ", sqrt(mean((yhat.rf-rf_test)^2)),"\n")
```

```
## RMSE of RF :  3.872377
```

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

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

```
############################### Understanding the 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 Cleaning & Preparation
############################
```

```
# to check if there are any missing values
any(is.na(school2))
```

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

```
# Thus we have no missing values in the data set.
```

```
# dropping G1 and G2 from school2 (portuguese)
df_port = subset(school2, select = -c(G1,G2))
colnames(df_port)
```

```
##  [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(df_port)
```

```
## Observations: 649
## Variables: 31
## $ 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, 0, 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, no, no, no, no, no, no, no, no, no, no, no, no,
no…
## $ 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> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 10, 2,
2, …
## $ G3         <int> 11, 11, 12, 14, 13, 13, 13, 13, 17, 13, 14, 13, 12, 13,
15…
```

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

# Medu - denotes Mother's eductaion - 5 levels
df_port$Medu = factor(df_port$Medu, levels=c("0","1","2","3","4"),
ordered=TRUE)
summary(df_port$Medu)
```

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

```
# Fedu - denotes Father's eductaion - 5 levels
df_port$Fedu = factor(df_port$Fedu, levels=c("0","1","2","3","4"),
ordered=TRUE)
summary(df_port$Fedu)
```

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

```
# famrel - denotes - quality of family relationships
# 1 - very bad to 5 - excellent
df_port$famrel = factor(df_port$famrel, levels=1:5, ordered=TRUE)
summary(df_port$famrel)
```

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

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

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

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

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

```r
# freetime - free time after school (1 - very low to 5 - very high)
df_port$freetime = factor(df_port$freetime, levels=1:5, ordered=TRUE)
summary(df_port$freetime)
```

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

```r
# goout - going out with friends ( 1 - very low to 5 - very high)
df_port$goout = factor(df_port$goout, levels=1:5, ordered=TRUE)
summary(df_port$goout)
```

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

```r
# Dalc - workday alcohol consumption (from 1 - very low to 5 - very high)
df_port$Dalc = factor(df_port$Dalc, levels=1:5, ordered=TRUE)
summary(df_port$Dalc)
```

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

```r
# Walc - weekend alcohol consumption ( 1 - very low to 5 - very high)
df_port$Walc = factor(df_port$Walc, levels=1:5, ordered=TRUE)
summary(df_port$Walc)
```

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

```r
# health - current health status ( 1 - very bad to 5 - very good)
df_port$health = factor(df_port$health, levels=1:5, ordered=TRUE)
summary(df_port$health)
```

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

```r
# failures - number of past class failures (n if 1<=n<3, else 4)
df_port$failures = factor(df_port$failures, levels=0:4, ordered=TRUE)
summary(df_port$failures)
```

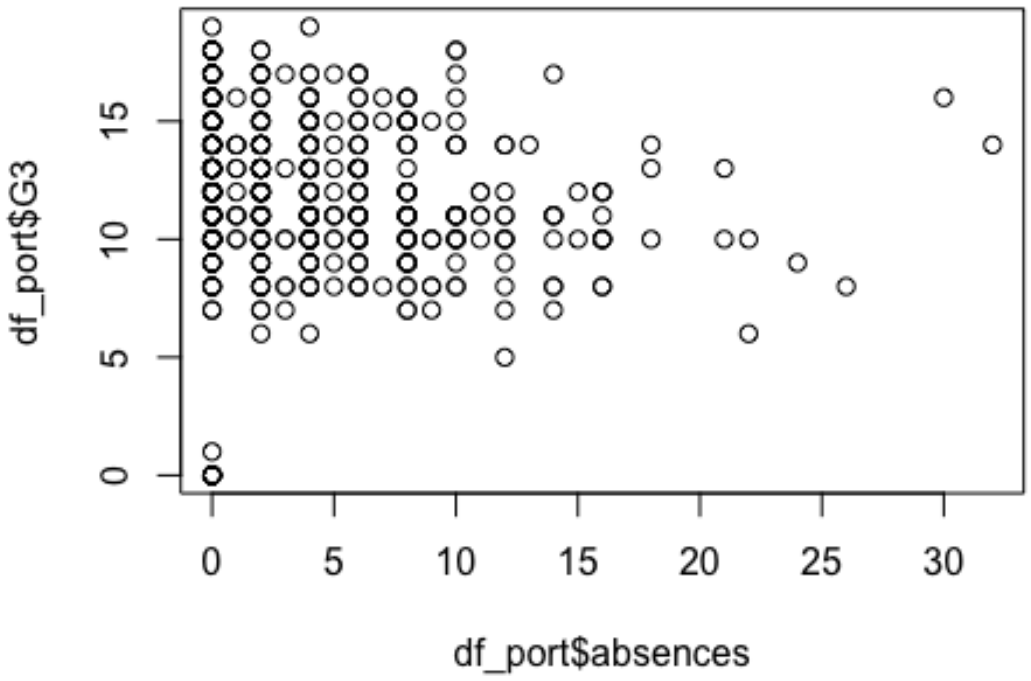
```
##   0   1   2   3   4
## 549  70  16  14   0
```

```r
summary(df_port)
```

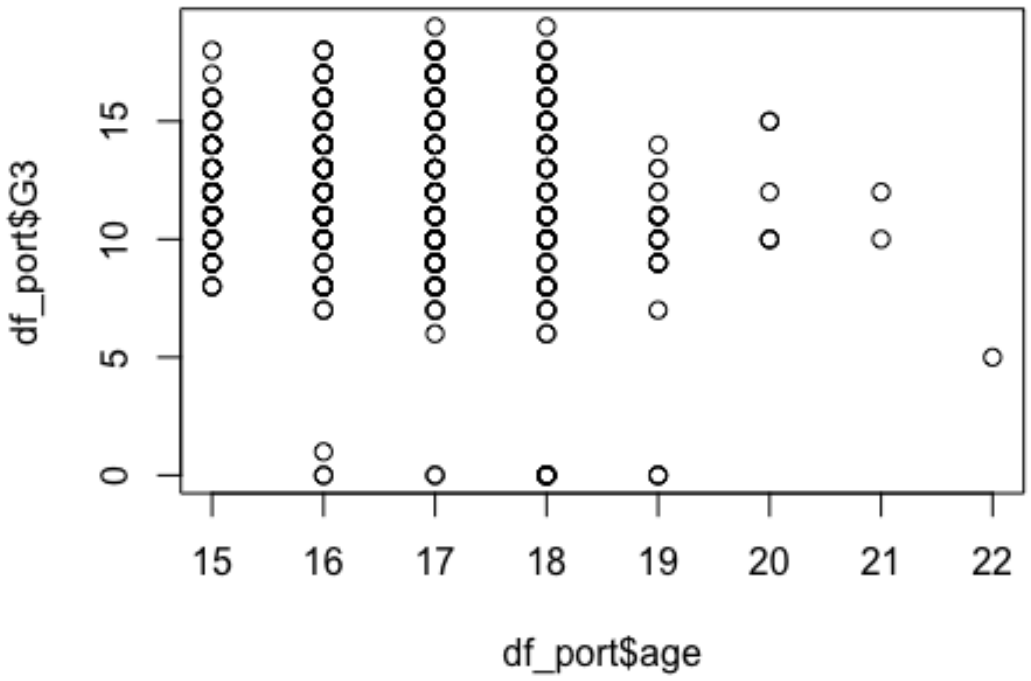
```
##  school   sex          age        address famsize   Pstatus Medu    Fedu
##  GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80   0:  6   0:  7
##  MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569   1:143   1:174
##                   Median :17.00                             2:186   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
```

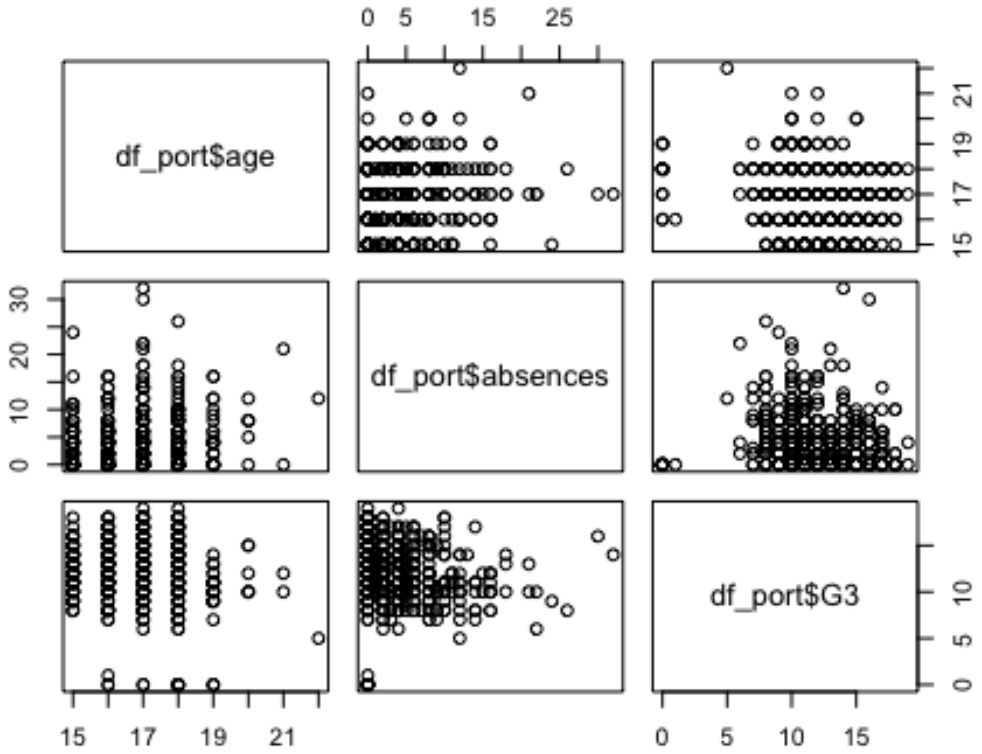
```r
############################## Exploratory Data Analysis(EDA)
#########################

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



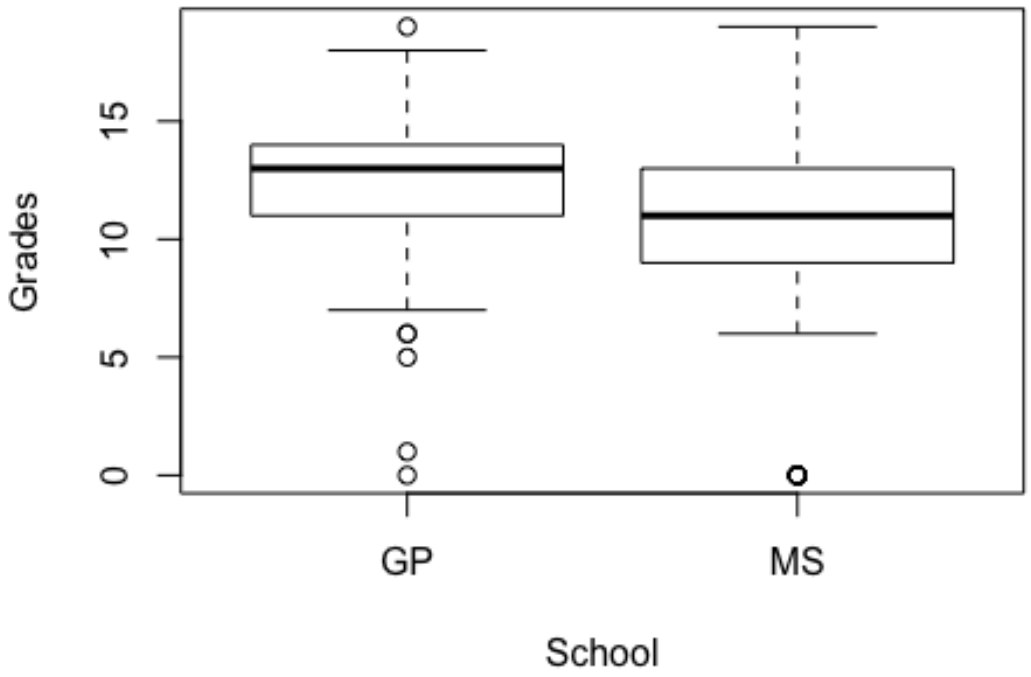

```
plot(df_port$age,df_port$G3)
```



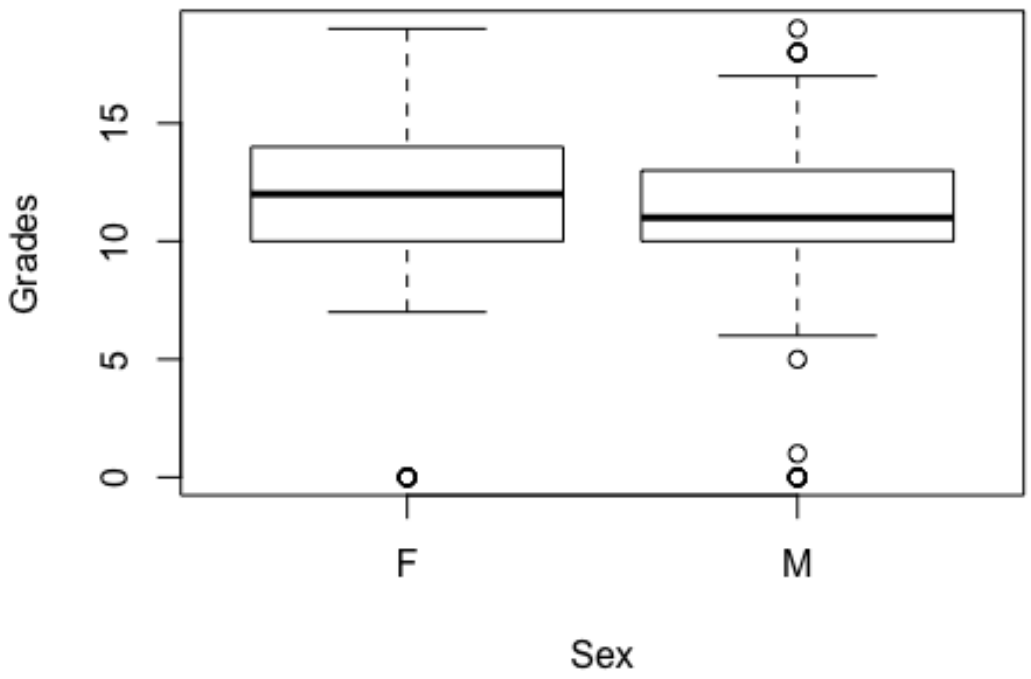

```
pairs(~df_port$age+df_port$absences+df_port$G3)
```



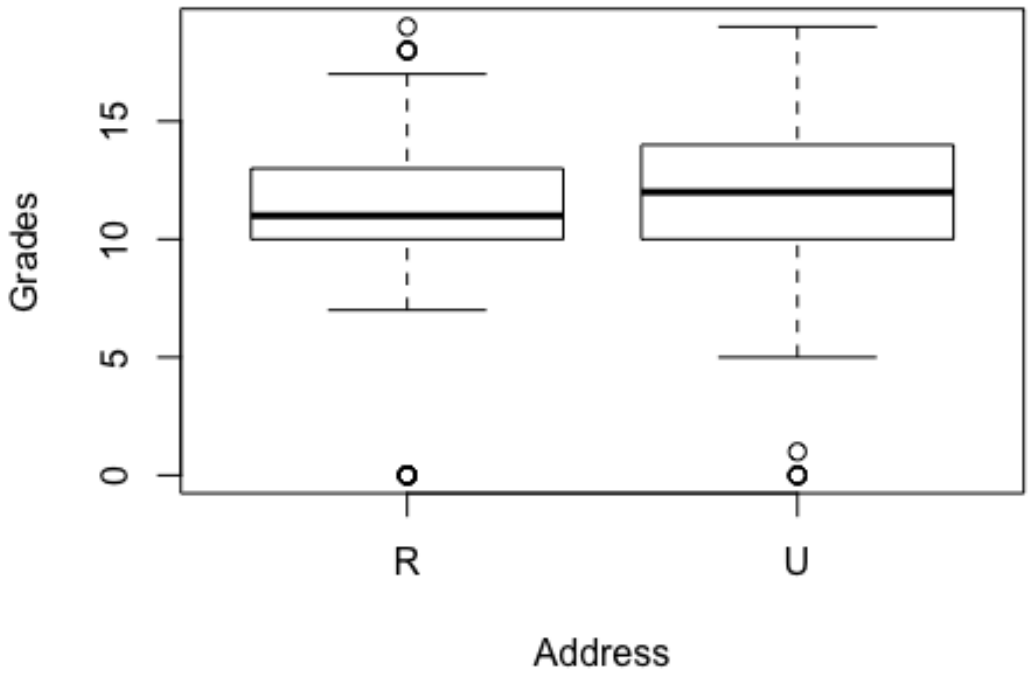

```r
# Creating box-plots for categorical data
suppressMessages(attach(df_port))
plot(school, G3, xlab = "School", ylab = "Grades")
```



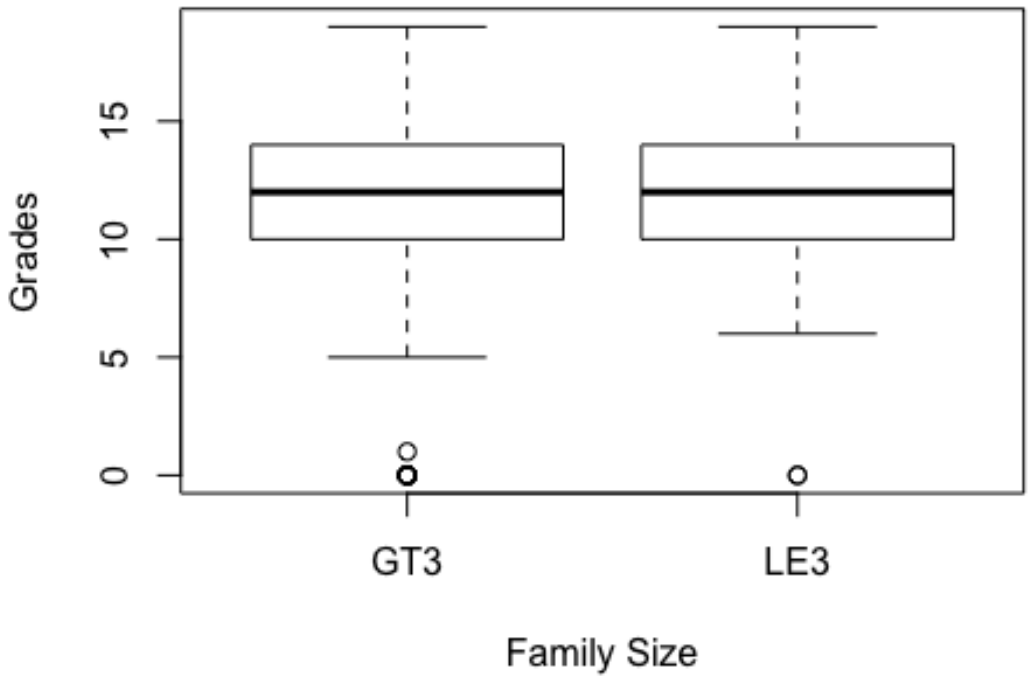

```
plot(sex,G3, xlab = "Sex", ylab = "Grades")
```



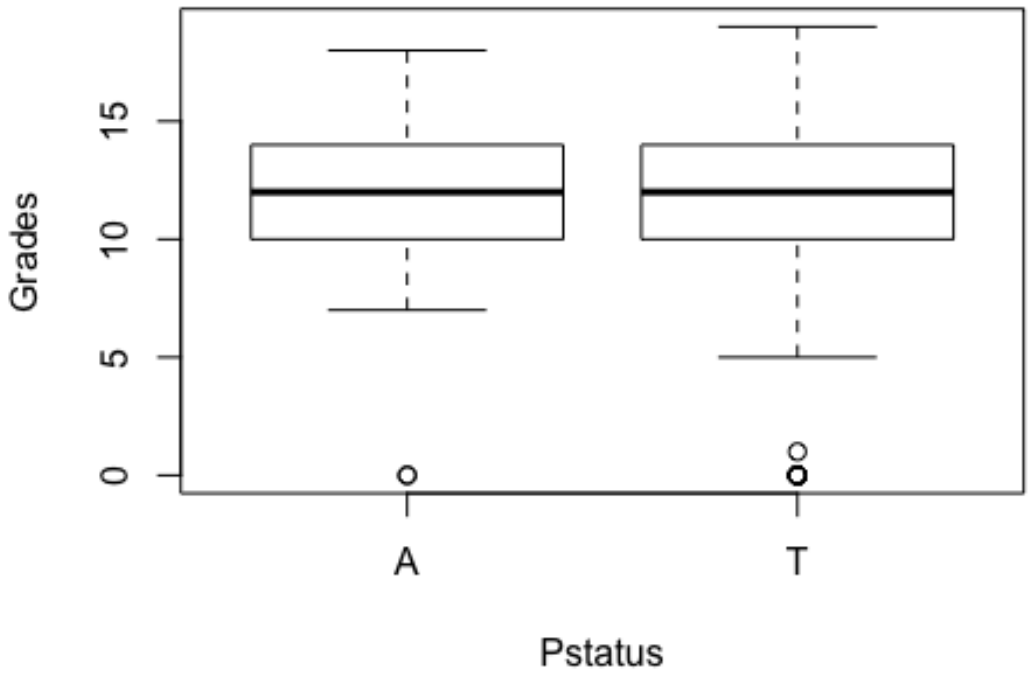

```
plot(address,G3, xlab = "Address", ylab = "Grades")
```



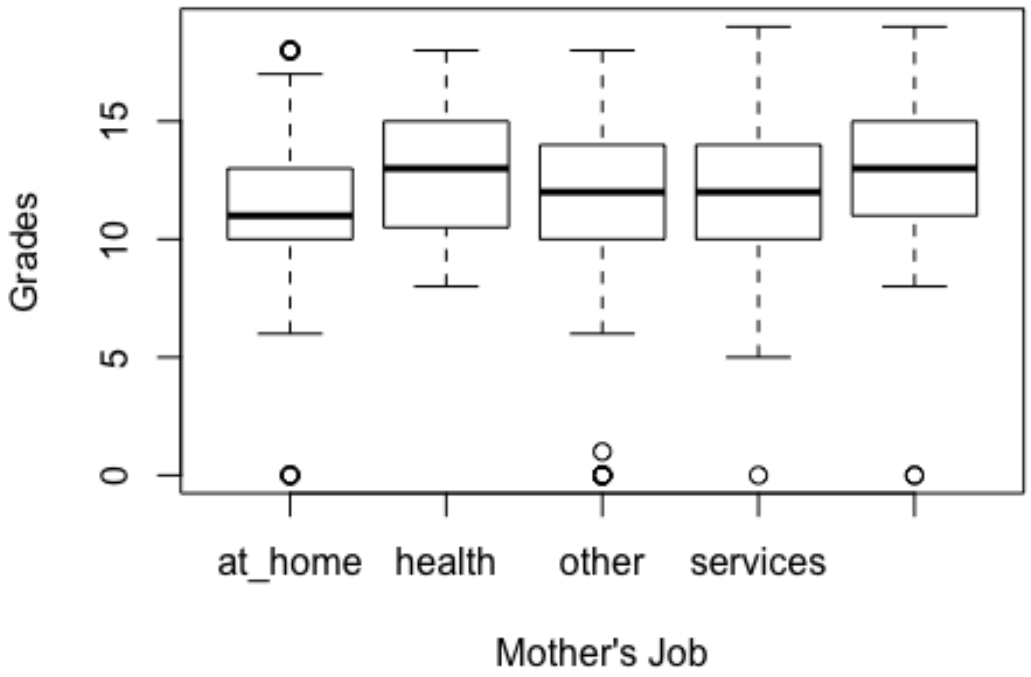

```r
plot(famsize, G3, xlab = "Family Size", ylab = "Grades")
```



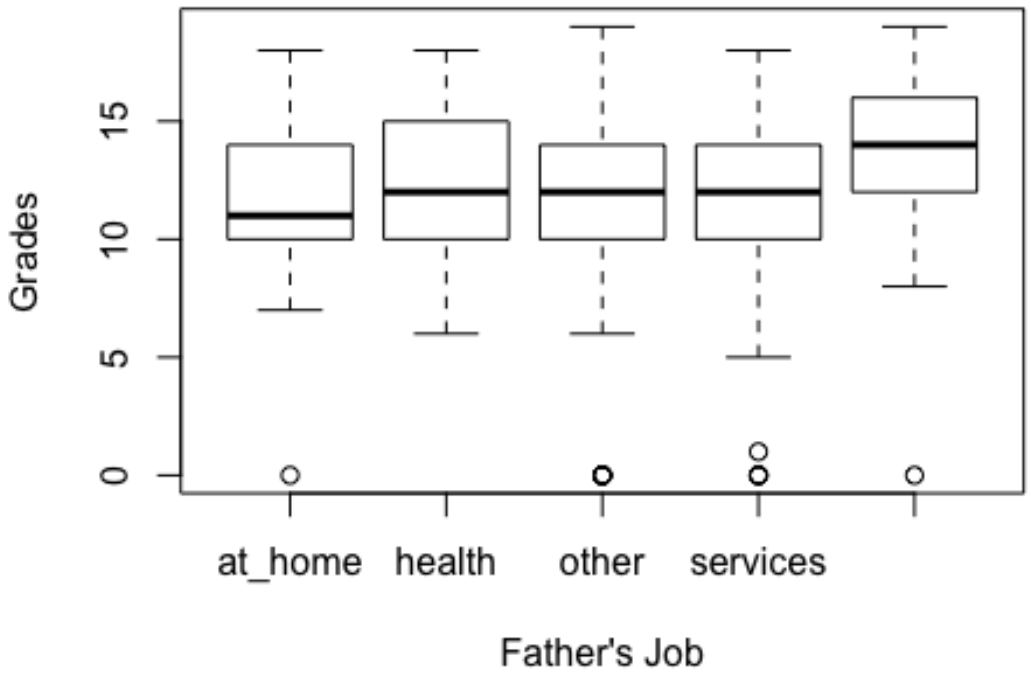

```
plot(Pstatus,G3, xlab = "Pstatus", ylab = "Grades")
```



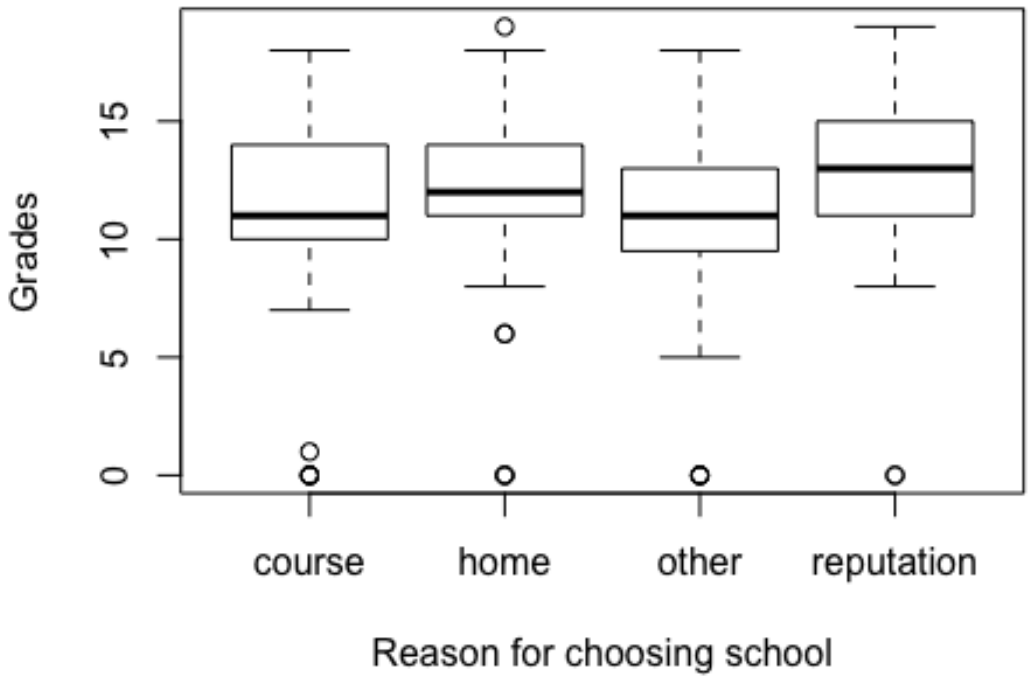

```
plot(Mjob,G3, xlab = "Mother's Job", ylab = "Grades")
```



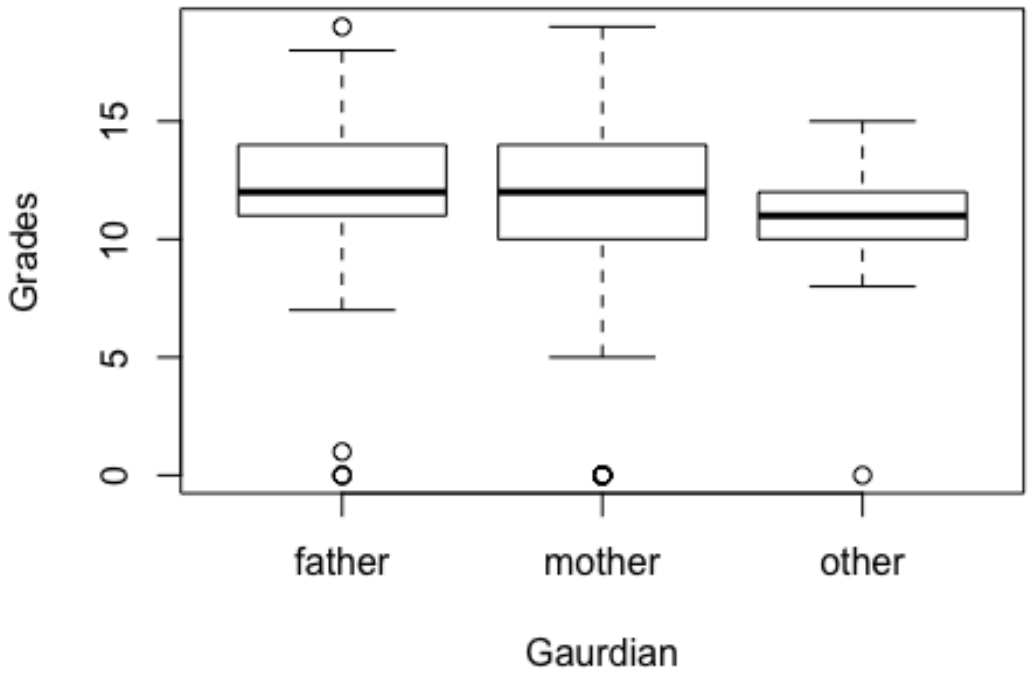

```
plot(Fjob, G3, xlab = "Father's Job", ylab = "Grades")
```



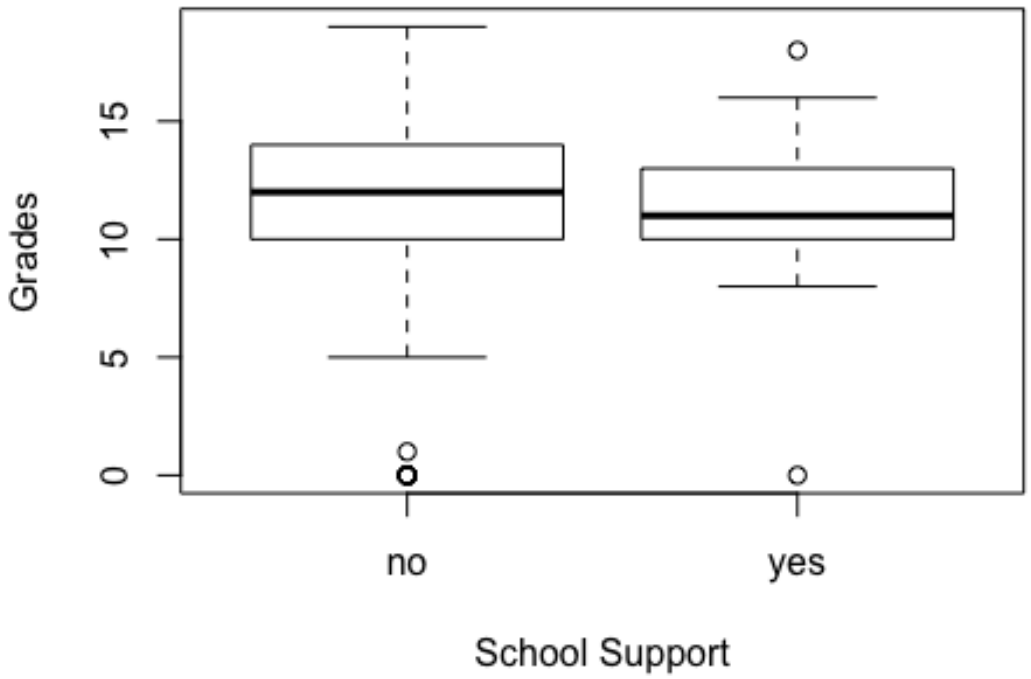

```r
plot(reason,G3, xlab = "Reason for choosing school", ylab = "Grades")
```



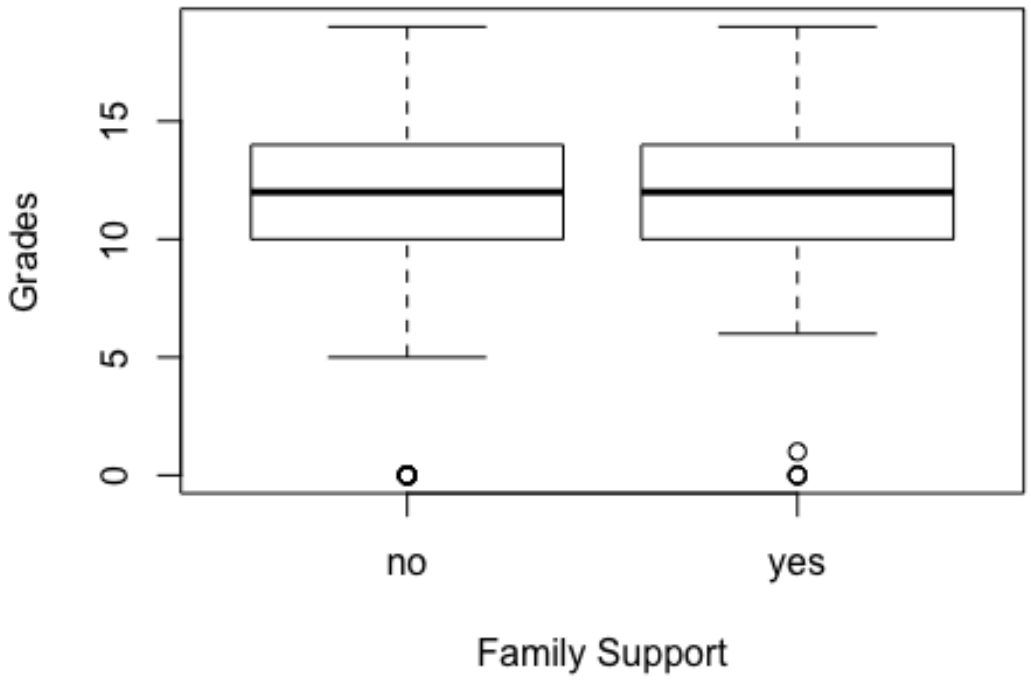

```
plot(guardian,G3, xlab = "Gaurdian", ylab = "Grades")
```




```
plot(schoolsup,G3, xlab = "School Support", ylab = "Grades")
```

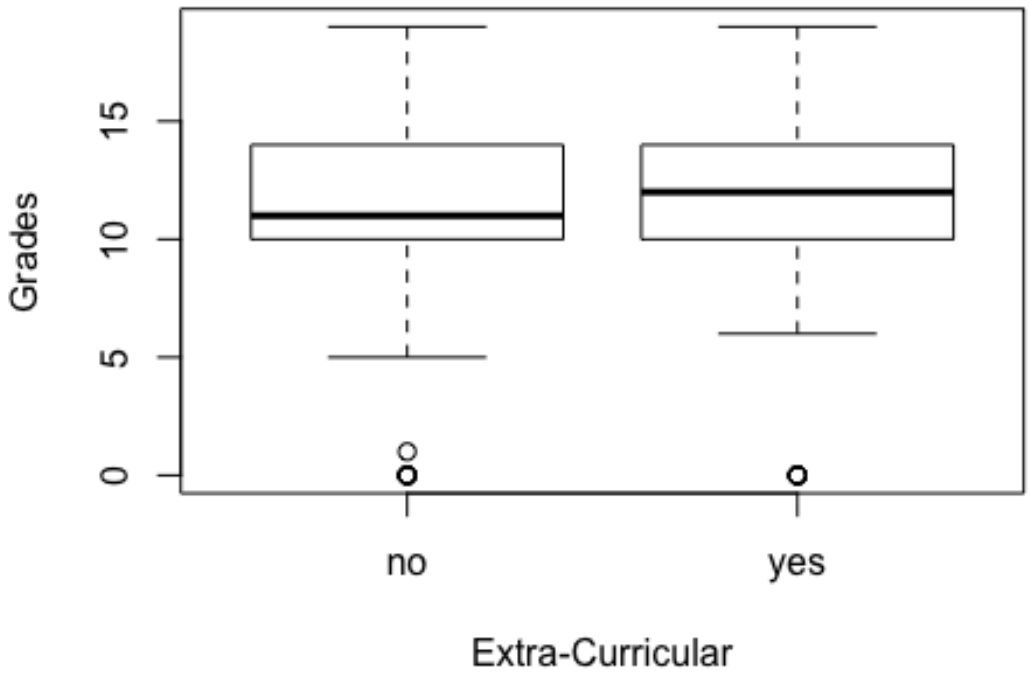



```r
plot(famsup,G3, xlab = "Family Support", ylab = "Grades")
```

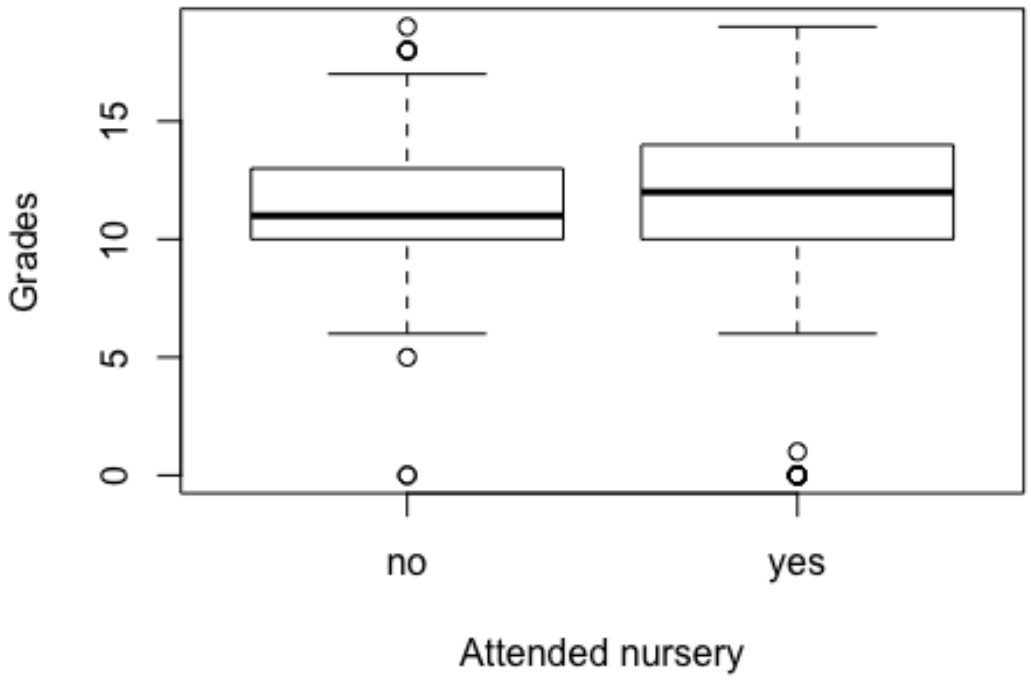



```r
plot(paid, G3, xlab = "Extra Paid classes", ylab = "Grades")
```

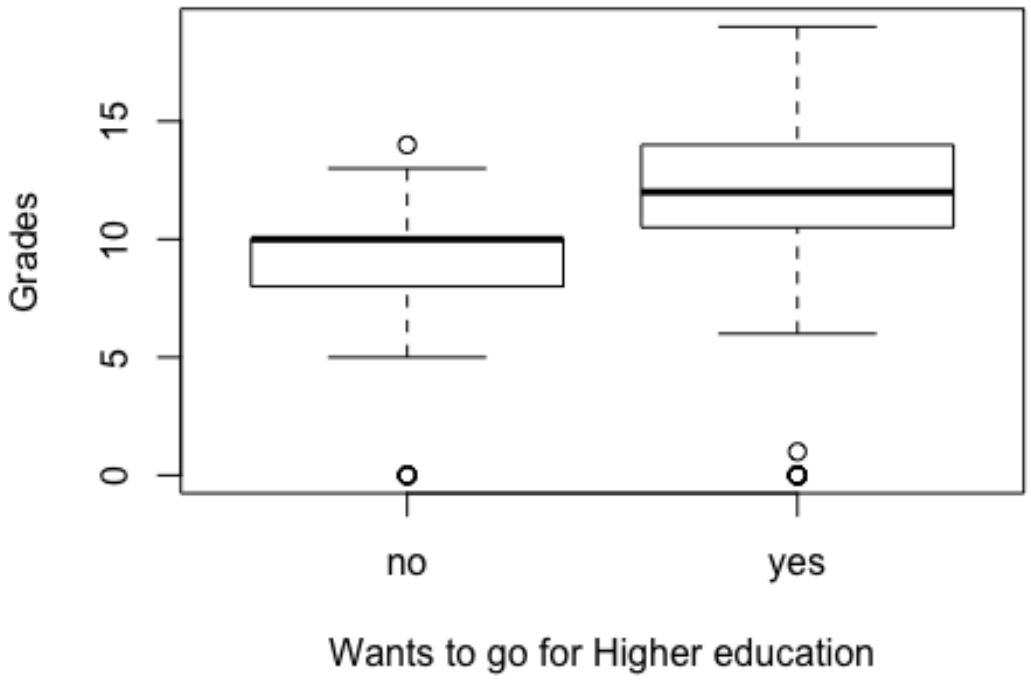

```
plot(activities,G3, xlab = "Extra-Curricular", ylab = "Grades")
```



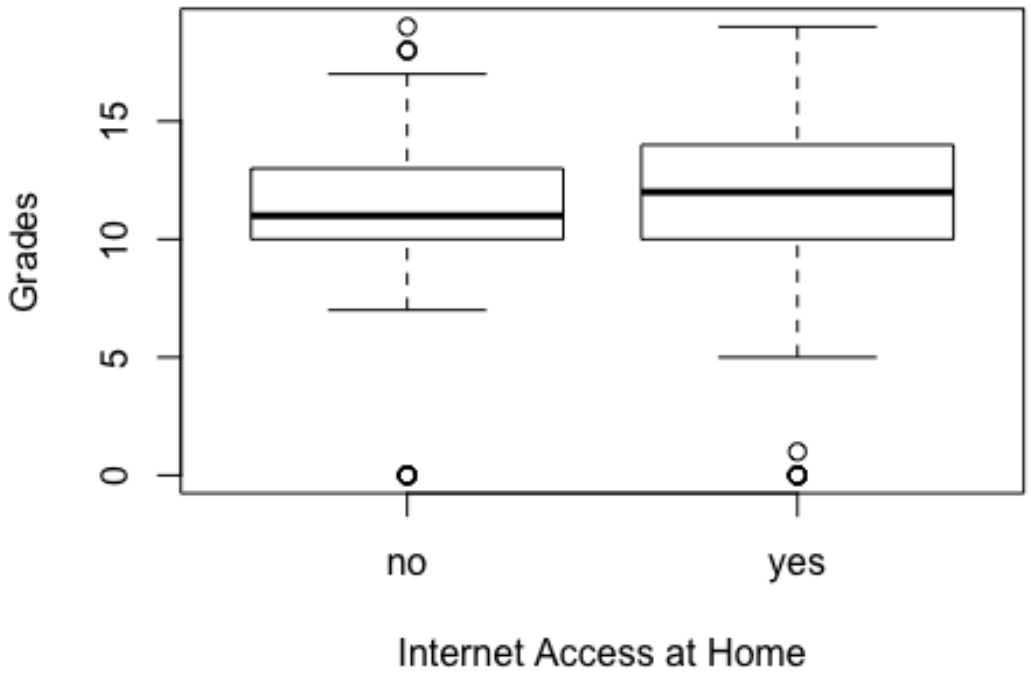

```
plot(nursery,G3, xlab = "Attended nursery", ylab = "Grades")
```



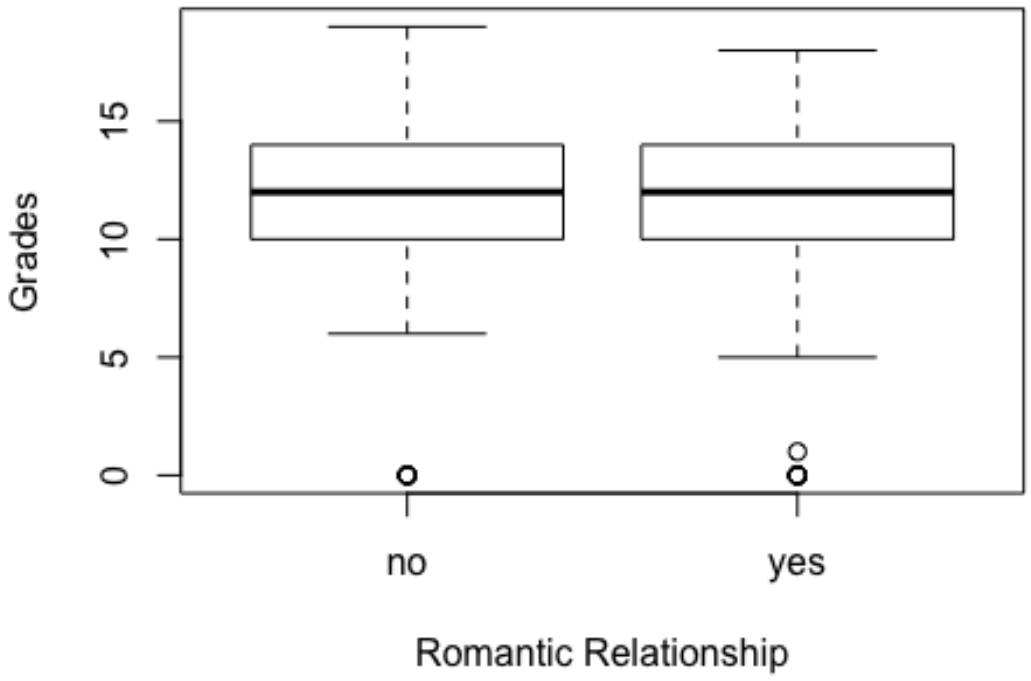

```
plot(higher,G3, xlab = "Wants to go for Higher education", ylab = "Grades")
```




```r
plot(internet,G3, xlab = "Internet Access at Home", ylab = "Grades")
```




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




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

set.seed(0)
train = sample(1:nrow(df_port), 520)
actual_g3 = df_port[-train,31]

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


### Subset Selection
# Stepwise Selection
# Linear Model

full_model_fit <- lm(G3~.,data = df_port[train,])
summary(full_model_fit)

##
## Call:
## lm(formula = G3 ~ ., data = df_port[train, ])
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -10.8970  -1.3576   0.0306   1.5394   7.0173
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.286837   2.427050   2.178 0.029902 *
## schoolMS         -1.236081   0.310938  -3.975 8.19e-05 ***
## sexM             -0.929916   0.285790  -3.254 0.001224 **
## age               0.222049   0.117555   1.889 0.059548 .
## addressU          0.602488   0.298302   2.020 0.044004 *
## famsizeLE3        0.161134   0.279973   0.576 0.565218
## PstatusT          0.338964   0.384518   0.882 0.378502
## Medu.L            0.240117   0.890602   0.270 0.787582
## Medu.Q           -0.165668   0.722559  -0.229 0.818756
## Medu.C            0.273096   0.476120   0.574 0.566533
## Medu^4           -0.240390   0.295462  -0.814 0.416300
## Fedu.L            0.447196   0.819861   0.545 0.585712
## Fedu.Q           -0.041832   0.668963  -0.063 0.950167
## Fedu.C           -0.181309   0.449047  -0.404 0.686577
## Fedu^4            0.259150   0.277065   0.935 0.350116
## Mjobhealth        1.163257   0.618820   1.880 0.060781 .
## Mjobother        -0.050762   0.354829  -0.143 0.886305
## Mjobservices      0.649493   0.432471   1.502 0.133845
## Mjobteacher       0.319050   0.588172   0.542 0.587782
## Fjobhealth       -1.191533   0.834765  -1.427 0.154162
## Fjobother        -0.119392   0.523210  -0.228 0.819602
## Fjobservices     -0.719502   0.549442  -1.310 0.191028
## Fjobteacher       0.237905   0.769851   0.309 0.757444
## reasonhome       -0.626721   0.324164  -1.933 0.053820 .
## reasonother      -0.912354   0.415149  -2.198 0.028482 *
## reasonreputation -0.282935   0.332751  -0.850 0.395615
## guardianmother   -0.247235   0.301435  -0.820 0.412540
## guardianother     0.652389   0.615950   1.059 0.290096
## traveltime.L     -0.259024   0.607194  -0.427 0.669880
## traveltime.Q     -0.468250   0.503319  -0.930 0.352701
## traveltime.C     -0.447902   0.376681  -1.189 0.235037
## studytime.L       0.985274   0.411305   2.395 0.017006 *
## studytime.Q       0.065065   0.358638   0.181 0.856117
## studytime.C      -0.072305   0.287161  -0.252 0.801317
## failures.L       -2.502752   0.652013  -3.839 0.000142 ***
## failures.Q        1.221516   0.591141   2.066 0.039365 *
## failures.C       -0.142311   0.587298  -0.242 0.808647
## schoolsupyes     -0.932009   0.413212  -2.256 0.024580 *
## famsupyes        -0.088939   0.265171  -0.335 0.737479
## paidyes          -0.451010   0.514370  -0.877 0.381053
## activitiesyes     0.061259   0.254191   0.241 0.809667
## nurseryyes       -0.007752   0.308185  -0.025 0.979943
## higheryes         1.471070   0.431743   3.407 0.000715 ***
## internetyes       0.322294   0.318565   1.012 0.312223
## romanticyes      -0.597855   0.262583  -2.277 0.023265 *
```

```
## famrel.L          0.792788   0.489213   1.621 0.105817
## famrel.Q         -0.288644   0.445691  -0.648 0.517553
## famrel.C         -0.369077   0.448682  -0.823 0.411182
## famrel^4         -0.035236   0.376841  -0.094 0.925545
## freetime.L       -0.316000   0.417671  -0.757 0.449699
## freetime.Q        0.123765   0.362532   0.341 0.732969
## freetime.C        0.242942   0.308669   0.787 0.431660
## freetime^4       -0.319454   0.241143  -1.325 0.185928
## goout.L          -0.052904   0.404295  -0.131 0.895949
## goout.Q          -0.955408   0.346955  -2.754 0.006131 **
## goout.C           0.576262   0.288035   2.001 0.046029 *
## goout^4          -0.157121   0.243331  -0.646 0.518798
## Dalc.L           -0.721008   0.705684  -1.022 0.307465
## Dalc.Q            0.296003   0.598409   0.495 0.621089
## Dalc.C            1.320514   0.565183   2.336 0.019907 *
## Dalc^4            1.566476   0.511779   3.061 0.002339 **
## Walc.L           -0.102940   0.511026  -0.201 0.840447
## Walc.Q            0.371599   0.383915   0.968 0.333605
## Walc.C            0.149434   0.324384   0.461 0.645259
## Walc^4           -0.205316   0.291177  -0.705 0.481098
## health.L         -0.569721   0.286146  -1.991 0.047083 *
## health.Q          0.143876   0.285706   0.504 0.614801
## health.C         -0.417965   0.317341  -1.317 0.188481
## health^4         -0.270124   0.297538  -0.908 0.364436
## absences         -0.034142   0.028816  -1.185 0.236709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.613 on 450 degrees of freedom
## Multiple R-squared:  0.4434, Adjusted R-squared:  0.3581
## F-statistic: 5.195 on 69 and 450 DF,  p-value: < 2.2e-16

# Backward AIC
library(leaps)
backward_aic_fit = MASS::stepAIC(full_model_fit, direction = "backward",
trace = FALSE)
backward_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 + address + Mjob + Fjob + reason + studytime +
```

```
##     failures + schoolsup + higher + romantic + goout + Dalc +
##     health + absences
##
##
##            Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1                                    450   3072.943 1063.812
## 2        - Medu  4  6.45568686       454   3079.399 1056.904
## 3        - Walc  4 14.42954705       458   3093.828 1051.335
## 4        - Fedu  4 19.56169176       462   3113.390 1046.612
## 5  - traveltime  3 10.93598099       465   3124.326 1042.435
## 6    - freetime  4 27.05105643       469   3151.377 1038.918
## 7     - nursery  1  0.00842128       470   3151.386 1036.920
## 8  - activities  1  0.02566774       471   3151.411 1034.924
## 9      - famsup  1  1.11245230       472   3152.524 1033.107
## 10    - famsize  1  2.37660293       473   3154.900 1031.499
## 11    - Pstatus  1  2.94256694       474   3157.843 1029.984
## 12       - paid  1  3.51678822       475   3161.360 1028.563
## 13     - famrel  4 41.64123202       479   3203.001 1027.368
## 14   - guardian  2 19.91486675       481   3222.916 1026.591
## 15   - internet  1 11.94897061       482   3234.865 1026.515
```

```
summary(backward_aic_fit)
```

```
##
## Call:
## lm(formula = G3 ~ school + sex + age + address + Mjob + Fjob +
##     reason + studytime + failures + schoolsup + higher + romantic +
##     goout + Dalc + health + absences, data = df_port[train, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1899  -1.4314   0.0418   1.5145   7.1138
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.72269    2.15587   2.654  0.00821 **
## schoolMS         -1.29200    0.28658  -4.508 8.21e-06 ***
## sexM             -0.81771    0.26550  -3.080  0.00219 **
## age               0.24530    0.10903   2.250  0.02491 *
## addressU          0.55751    0.27568   2.022  0.04370 *
## Mjobhealth        1.22920    0.52976   2.320  0.02074 *
## Mjobother        -0.01998    0.33347  -0.060  0.95225
## Mjobservices      0.62204    0.38802   1.603  0.10956
## Mjobteacher       0.55585    0.46489   1.196  0.23241
## Fjobhealth       -0.96093    0.78135  -1.230  0.21936
## Fjobother        -0.17907    0.49109  -0.365  0.71554
## Fjobservices     -0.78313    0.51753  -1.513  0.13088
## Fjobteacher       0.37668    0.69536   0.542  0.58827
## reasonhome       -0.55555    0.31031  -1.790  0.07403 .
## reasonother      -0.94626    0.40185  -2.355  0.01894 *
```

```
## reasonreputation -0.23136    0.32020  -0.723  0.47030
## studytime.L       1.01009    0.38625   2.615  0.00920 **
## studytime.Q       0.00815    0.33927   0.024  0.98084
## studytime.C      -0.20499    0.27333  -0.750  0.45365
## failures.L       -2.66610    0.60830  -4.383 1.44e-05 ***
## failures.Q        1.13322    0.56265   2.014  0.04456 *
## failures.C       -0.19619    0.55358  -0.354  0.72320
## schoolsupyes     -1.01646    0.39425  -2.578  0.01023 *
## higheryes         1.48743    0.41477   3.586  0.00037 ***
## romanticyes      -0.62136    0.25179  -2.468  0.01394 *
## goout.L          -0.15308    0.34942  -0.438  0.66150
## goout.Q          -0.96766    0.31173  -3.104  0.00202 **
## goout.C           0.57378    0.27748   2.068  0.03919 *
## goout^4          -0.21029    0.23595  -0.891  0.37326
## Dalc.L           -0.55279    0.59075  -0.936  0.34988
## Dalc.Q            0.48000    0.52842   0.908  0.36413
## Dalc.C            1.25724    0.53140   2.366  0.01838 *
## Dalc^4            1.58175    0.48732   3.246  0.00125 **
## health.L         -0.60405    0.27302  -2.212  0.02740 *
## health.Q          0.08943    0.27213   0.329  0.74259
## health.C         -0.32091    0.30060  -1.068  0.28625
## health^4         -0.28560    0.28847  -0.990  0.32265
## absences         -0.04068    0.02763  -1.472  0.14168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.591 on 482 degrees of freedom
## Multiple R-squared:  0.4141, Adjusted R-squared:  0.3691
## F-statistic: 9.206 on 37 and 482 DF,  p-value: < 2.2e-16
```

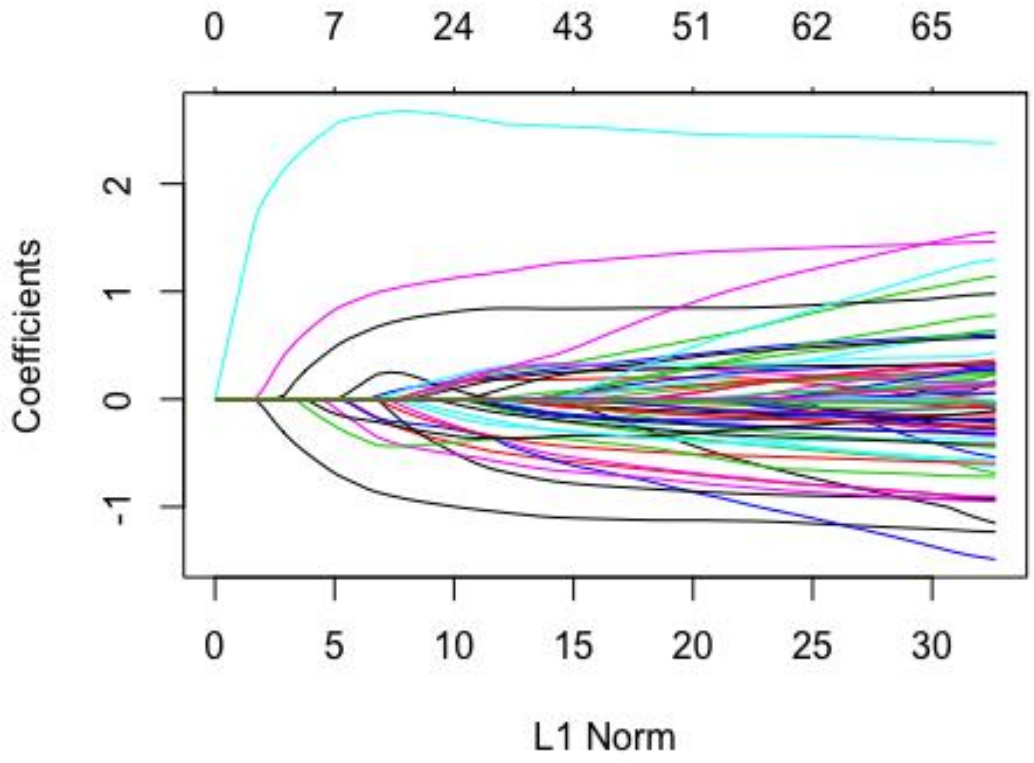
```
backward_aic_pred = predict(backward_aic_fit, newdata = df_port[-train,1:30])
mean((backward_aic_pred-actual_g3)^2)
```

```
## [1] 7.505688
```

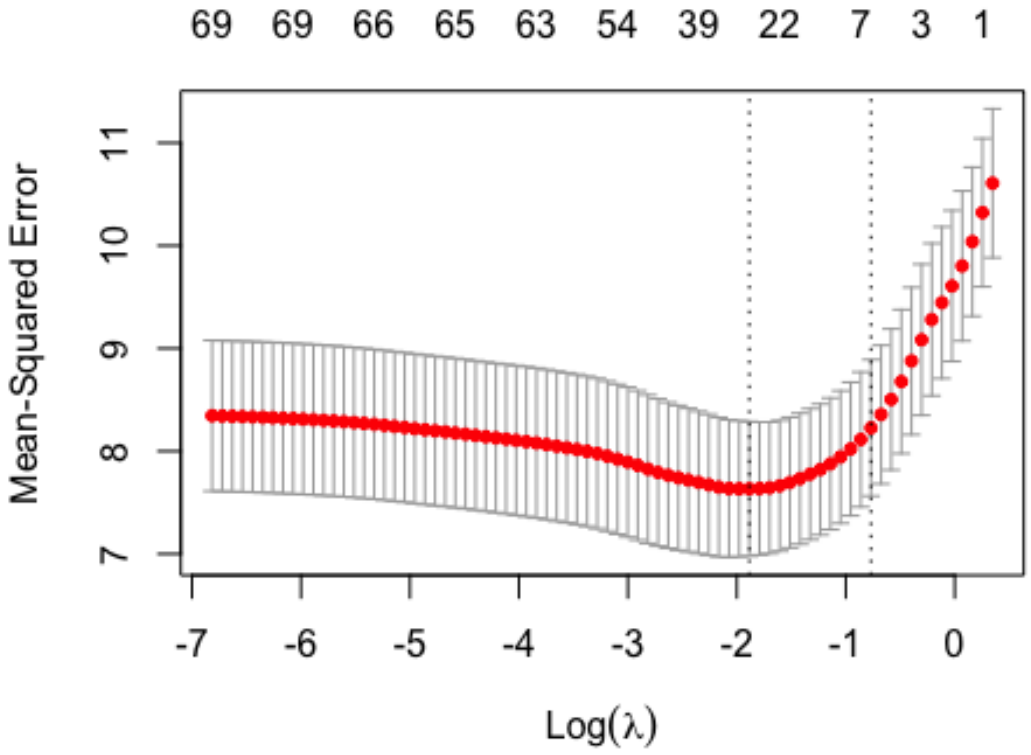
```
coef(backward_aic_fit)
```

```
##      (Intercept)         schoolMS             sexM              age
##      5.722690553     -1.291997872     -0.817709646      0.245298665
##         addressU       Mjobhealth        Mjobother     Mjobservices
##      0.557510601      1.229200513     -0.019980701      0.622039441
##      Mjobteacher       Fjobhealth        Fjobother     Fjobservices
##      0.555854483     -0.960932120     -0.179072625     -0.783126926
##      Fjobteacher        reasonhome       reasonother reasonreputation
##      0.376678764     -0.555548428     -0.946263931     -0.231358764
##      studytime.L      studytime.Q      studytime.C       failures.L
##      1.010088222      0.008150383     -0.204989605     -2.666103814
##       failures.Q       failures.C     schoolsupyes        higheryes
##      1.133216683     -0.196186827     -1.016459008      1.487427881
##      romanticyes           goout.L          goout.Q           goout.C
##     -0.621363866     -0.153084935     -0.967661700      0.573783162
```

```
##          goout^4           Dalc.L           Dalc.Q           Dalc.C
##     -0.210285968     -0.552790698      0.480000755      1.257235940
##           Dalc^4         health.L         health.Q         health.C
##      1.581752091     -0.604045745      0.089425272     -0.320912708
##         health^4         absences
##     -0.285595807     -0.040677823

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

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

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

lasso_mod = glmnet(x_train,
                   y_train,
                   alpha = 1) # Fit lasso model on training data

plot(lasso_mod)    # Draw plot of coefficients
```




```r
set.seed(1)
cv.out = cv.glmnet(x_train, y_train, alpha = 1) # Fit lasso model on training data
plot(cv.out) # Draw plot of training MSE as a function of lambda
```




```
best_lambda = cv.out$lambda.min # Select lamda that minimizes training MSE
lasso_pred = predict(lasso_mod, s = best_lambda, newx = x_test) # Use best
lambda to predict test data
mean((lasso_pred - y_test)^2) # Calculate test MSE
```

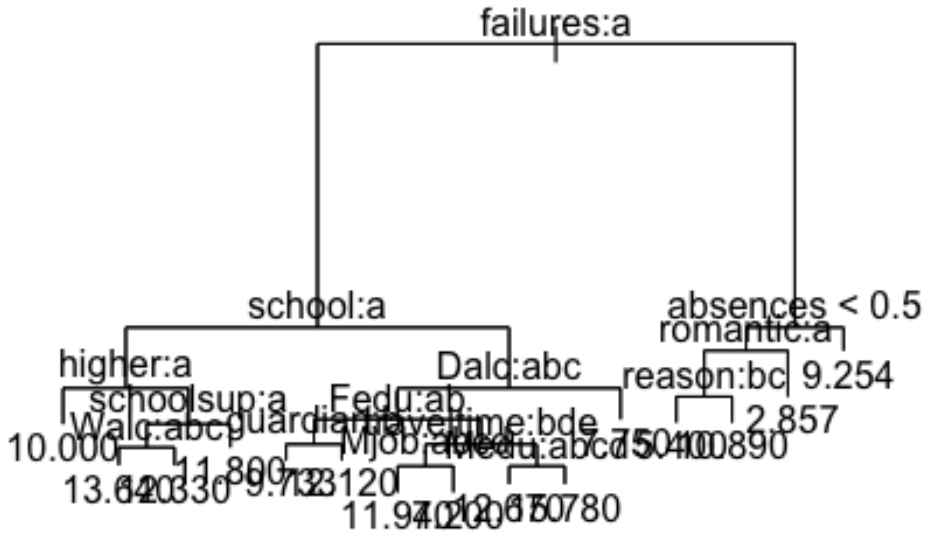
```
## [1] 6.702807
```

```
lasso_best <- glmnet(x_train, y_train, alpha = 1, lambda = best_lambda)
coef(lasso_best)
```

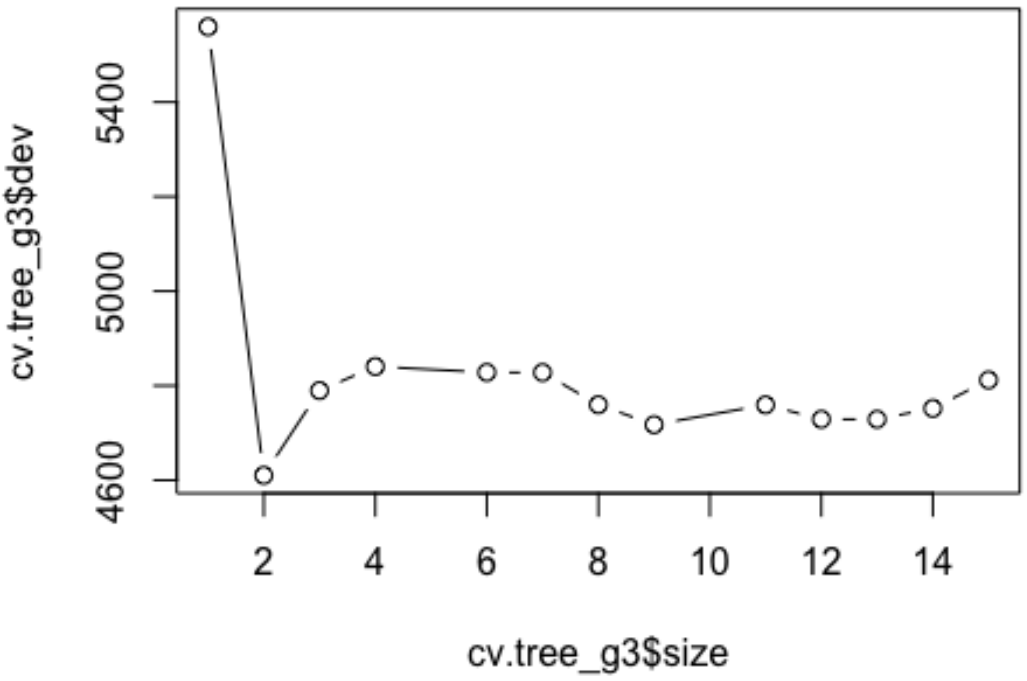
```
## 72 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)      10.0385233925
## schoolMS         -1.0230860417
## sexM             -0.4529522051
## age               .
## addressU          0.1883057671
## famsizeLE3        .
## PstatusT          .
## Medu.L            0.0341045822
## Medu.Q            0.1747250570
## Medu.C            .
## Medu^4            .
## Fedu.L            0.2626432819
```

```
## Fedu.Q            .
## Fedu.C            .
## Fedu^4            .
## Mjobhealth         0.2075531268
## Mjobother         -0.0319172053
## Mjobservices       .
## Mjobteacher        .
## Fjobhealth         .
## Fjobother          .
## Fjobservices      -0.0326211085
## Fjobteacher        .
## reasonhome         .
## reasonother       -0.3150062983
## reasonreputation   .
## guardianmother     .
## guardianother      .
## traveltime.L       .
## traveltime.Q       .
## traveltime.C       .
## traveltime^4       .
## studytime.L        0.8327425305
## studytime.Q        .
## studytime.C        .
## failures.L        -0.3102033493
## failures.Q         2.6058281753
## failures.C         .
## failures^4         0.2374217271
## schoolsupyes      -0.5954414690
## famsupyes          .
## paidyes            .
## activitiesyes      .
## nurseryyes         .
## higheryes          1.1525230248
## internetyes        0.1441149315
## romanticyes       -0.3116377098
## famrel.L           .
## famrel.Q           .
## famrel.C          -0.2117012191
## famrel^4           .
## freetime.L        -0.0224087693
## freetime.Q         .
## freetime.C         .
## freetime^4        -0.0003148453
## goout.L            .
## goout.Q           -0.5425210922
## goout.C            0.0243958220
## goout^4            .
## Dalc.L            -0.3831428900
## Dalc.Q             .
## Dalc.C             .
```

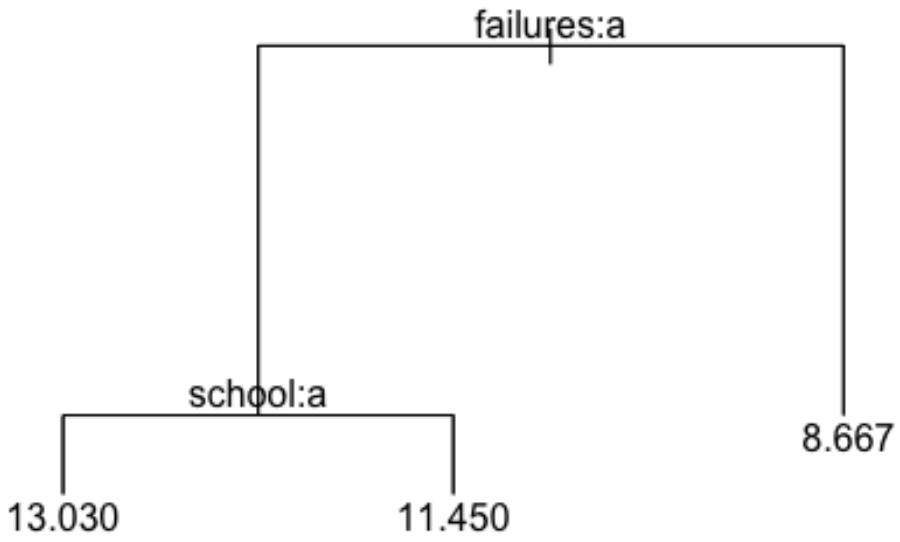
```
## Dalc^4            0.2074990754
## Walc.L           -0.3603219087
## Walc.Q            .
## Walc.C            .
## Walc^4            .
## health.L         -0.1623315935
## health.Q          .
## health.C         -0.0881602442
## health^4          .
## absences         -0.0016493008

############### TREES ##############

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

tree_g3 = tree(G3~., data = df_port , subset = train)
summary(tree_g3)

##
## Regression tree:
## tree(formula = G3 ~ ., data = df_port, subset = train)
## Variables actually used in tree construction:
##  [1] "failures"   "school"     "higher"     "schoolsup"  "Walc"
##  [6] "Dalc"       "Fedu"       "guardian"   "traveltime" "Mjob"
## [11] "Medu"       "absences"   "romantic"   "reason"
## Number of terminal nodes:  15
## Residual mean deviance:  6.08 = 3070 / 505
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -9.73300 -1.64400  0.08497  0.00000  1.35600  7.14300

plot(tree_g3)
text(tree_g3)
```



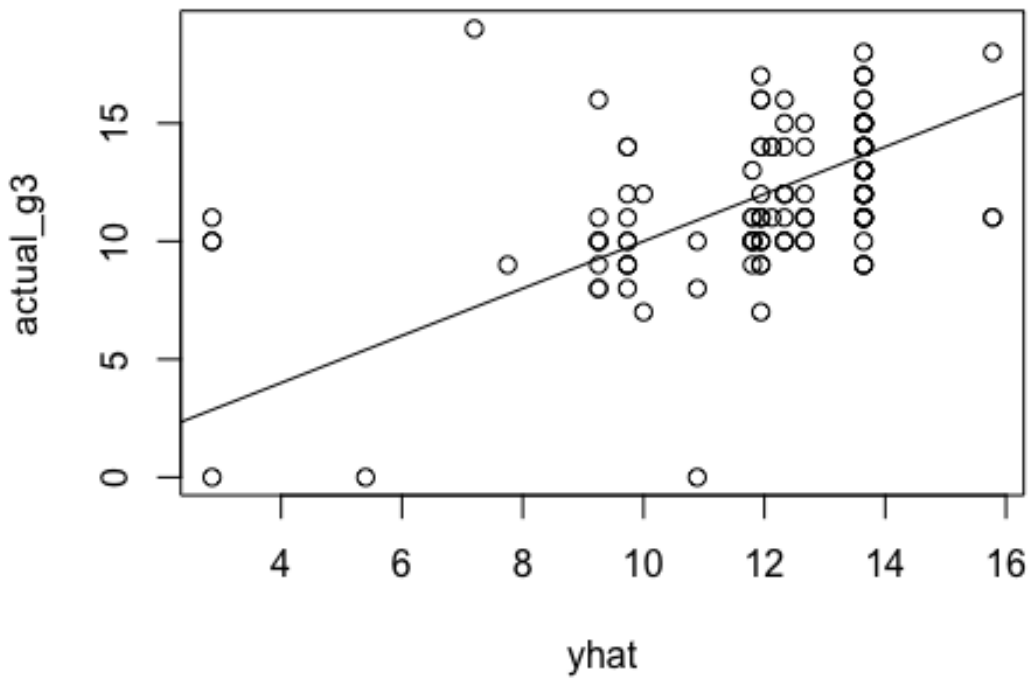

```
cv.tree_g3 = cv.tree(tree_g3)
plot(cv.tree_g3$size, cv.tree_g3$dev, type = 'b')
```



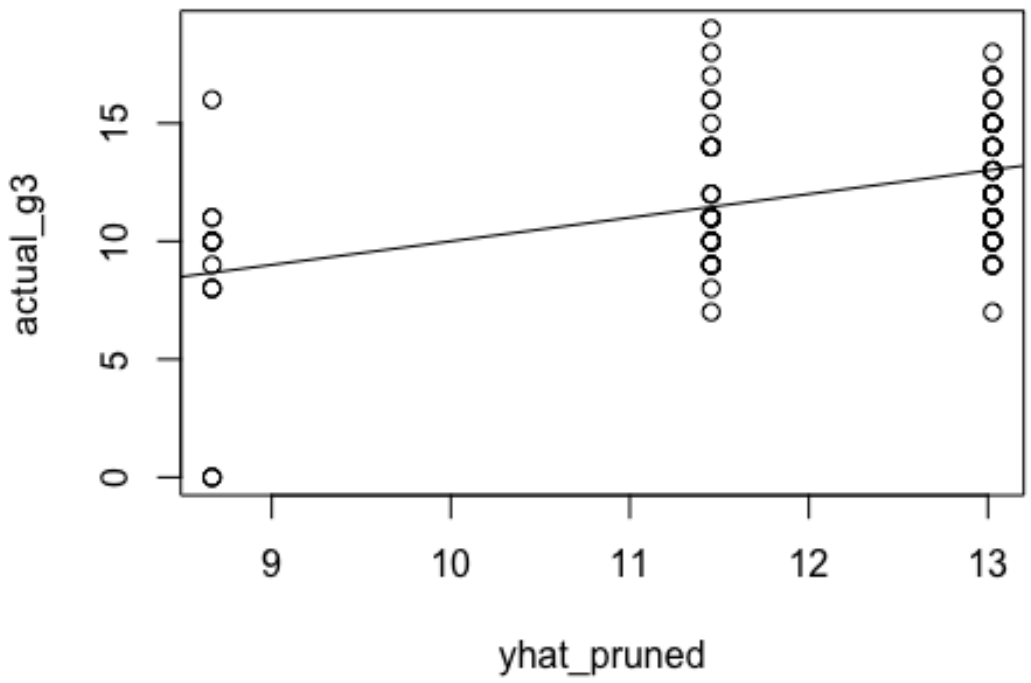

```
prune.tree_g3 = prune.tree(tree_g3, best = 3)
plot(prune.tree_g3)
text(prune.tree_g3)
```




```
yhat = predict(tree_g3, newdata = df_port[-train,1:30])
plot(yhat, actual_g3)
abline(0,1)
```




```
mean((yhat-actual_g3)^2)

## [1] 8.573731

yhat_pruned = predict(prune.tree_g3, newdata = df_port[-train,1:30])
plot(yhat_pruned, actual_g3)
abline(0,1)
```




```
mean((yhat_pruned-actual_g3)^2)
```

```
## [1] 7.72787
```

```
############### RANDOM FOREST ################
```

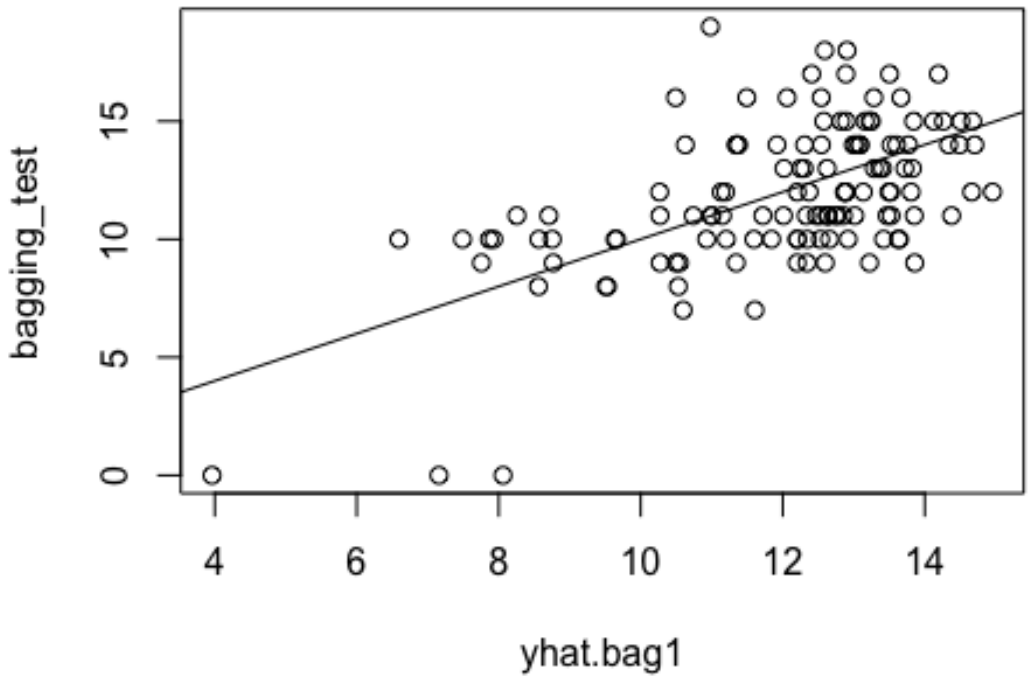
```
library(randomForest)
```

```
# We are performing bagging - by considering all the predictors i.e. mtry =
30
set.seed(-1)
bagging_g3 = randomForest(G3~., data = df_port[train,], mtry = 30, ntree=
1000, importance = TRUE)
bagging_g3
```

```
##
## Call:
##  randomForest(formula = G3 ~ ., data = df_port[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: 7.635967
##                     % Var explained: 28.08
```

```
yhat.bag1 = predict(bagging_g3, newdata = df_port[-train,1:30])
bagging_test = df_port[-train,"G3"]
plot(yhat.bag1, bagging_test)
abline(0,1)
```




```
mean((yhat.bag1-bagging_test)^2)
```

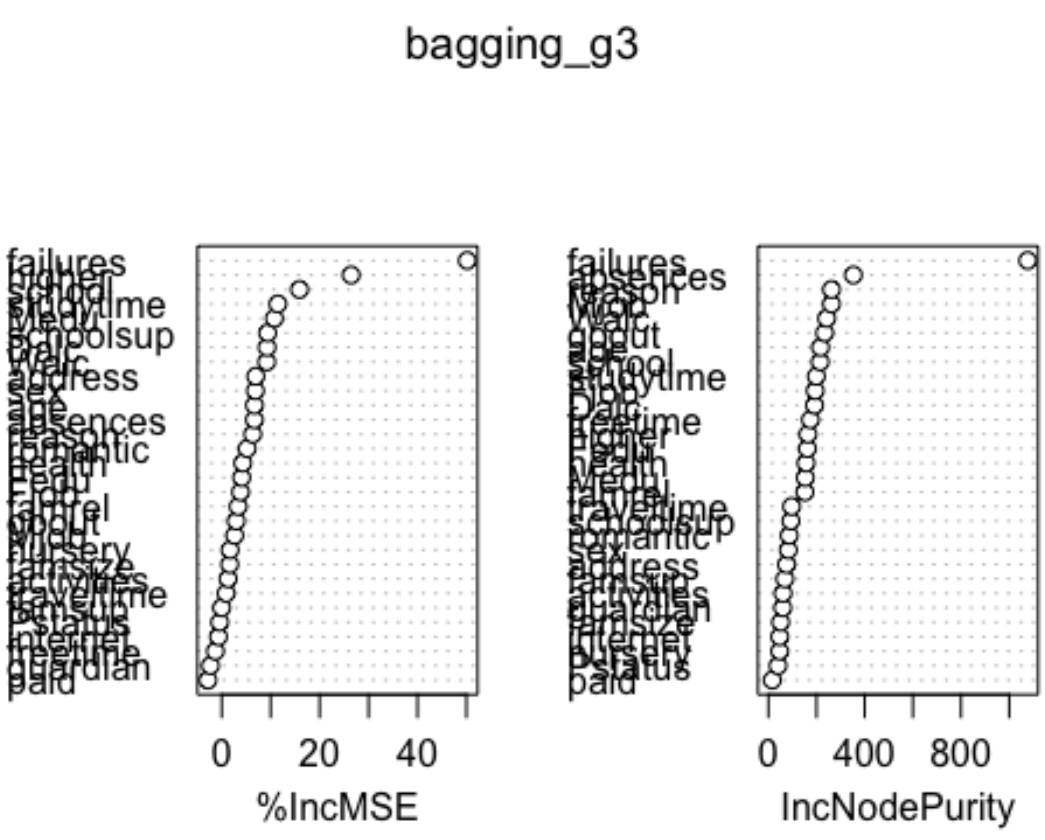
```
## [1] 6.471414
```

```
importance(bagging_g3)
```

```
##                %IncMSE IncNodePurity
## school     15.90394019     215.31111
## sex         6.88546475      82.98756
## age         6.69778044     216.38610
## address     6.96126153      75.78787
## famsize     1.62963563      50.94322
## Pstatus    -0.40849522      39.89862
## Medu       10.76322256     153.28604
## Fedu        4.07577466     162.02026
## Mjob        2.65975416     261.96249
```

```
## Fjob         3.85716091     194.40672
## reason       6.33798251     262.29793
## guardian    -2.31838280      59.72885
## traveltime   0.89752085      96.57556
## studytime   11.46661111     199.21714
## failures    50.17322080    1077.23288
## schoolsup    9.38586317      93.52542
## famsup      -0.01954312      68.54344
## paid        -2.82740264      16.16808
## activities   1.29470206      60.99892
## nursery      1.70250886      47.23048
## higher      26.43926110     164.16047
## internet    -0.66696898      47.34818
## romantic     5.18674843      86.44411
## famrel       3.36767222     151.43816
## freetime    -1.29545418     174.81081
## goout        2.96839684     233.94433
## Dalc         9.26281749     191.61963
## Walc         9.19939116     242.05873
## health       4.30018341     156.36127
## absences     6.68787890     352.92494
```

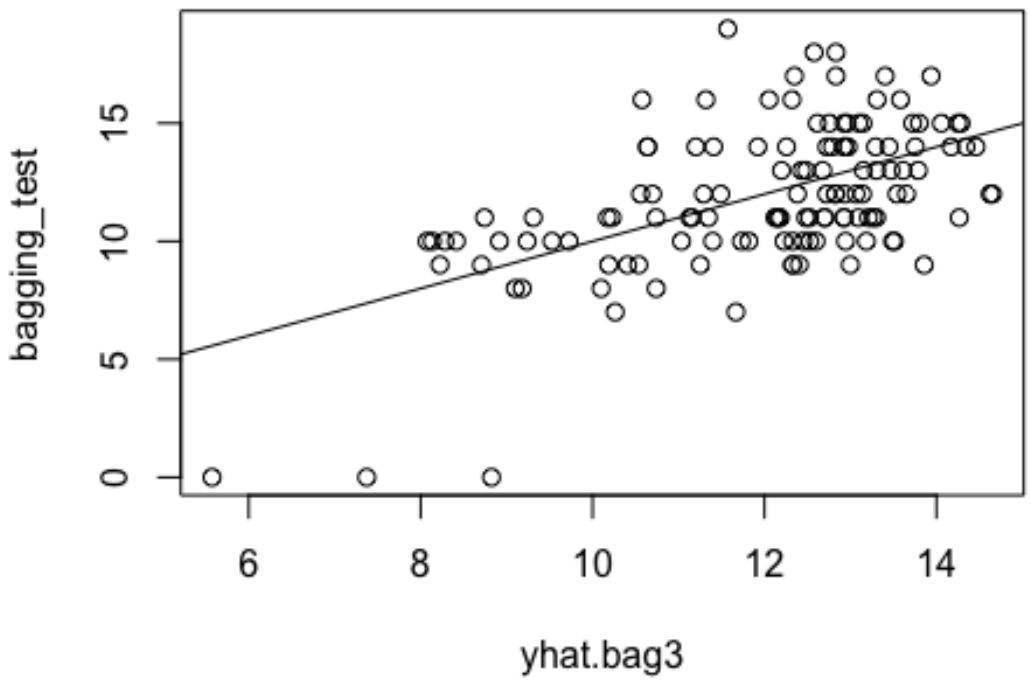
```
varImpPlot(bagging_g3)
```



```r
# trying RF - that is with m != p, and setting importance = False
set.seed(-1)
bagging_g3 = randomForest(G3~., data = df_port[train,], mtry = 10, ntree=
1000, importance = FALSE)
bagging_g3
```

```
##
## Call:
##  randomForest(formula = G3 ~ ., data = df_port[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: 7.303379
##                     % Var explained: 31.21
```

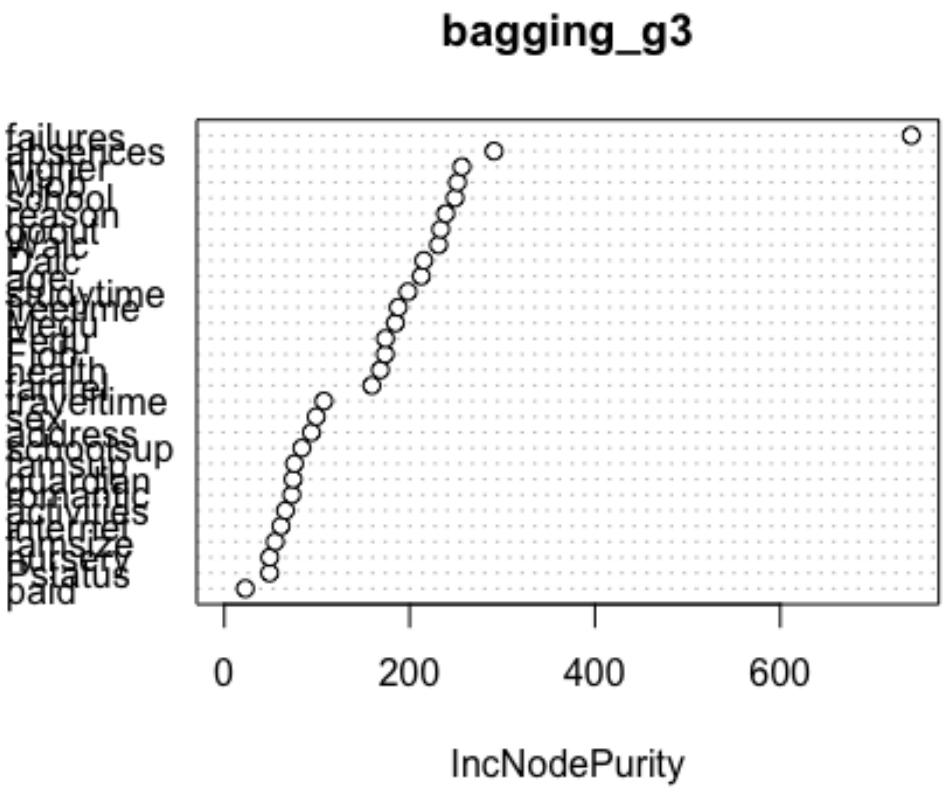
```r
yhat.bag3 = predict(bagging_g3, newdata = df_port[-train,])
bagging_test = df_port[-train,"G3"]
plot(yhat.bag3, bagging_test)
abline(0,1)
```




```r
mean((yhat.bag3-bagging_test)^2)
```

```
## [1] 6.480404
```

```
varImpPlot(bagging_g3)
```




```
###################################  RESULTS
##########################################
```

```
cat("\n\n Model Performance : \n\n")
```

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

```
cat("RMSE of Backward Step wise : ", sqrt(mean((backward_aic_pred-
actual_g3)^2)),"\n")
```

```
## RMSE of Backward Step wise :  2.739651
```

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

```
## RMSE of Lasso :  2.588978
```

```
cat("RMSE of Decision Tree : ", sqrt(mean((yhat_pruned-actual_g3)^2)),"\n")
```

```
## RMSE of Decision Tree :  2.779905
```

```r
cat("RMSE of Bagged Decision Trees : ", sqrt(mean((yhat.bag1-
bagging_test)^2)),"\n")
```

```
## RMSE of Bagged Decision Trees :  2.543897
```

```r
cat("RMSE of RF : ", sqrt(mean((yhat.bag3-bagging_test)^2)),"\n")
```

```
## RMSE of RF :  2.545664
```