Student Performance Data Analysis

AGH, 2021, Group 1
Marcin Baranek, Kamil Bartocha

Dataset: Student Performance Data Set
https://archive.ics.uci.edu/ml/datasets/Student+Performance
Data description: Data approach student achievement(grades) in secondary education of Portuguese schools.

- Number of Attributes: 33
- Number of Instances: 395
- Target Variables: G1, G2, G3 (final grade)

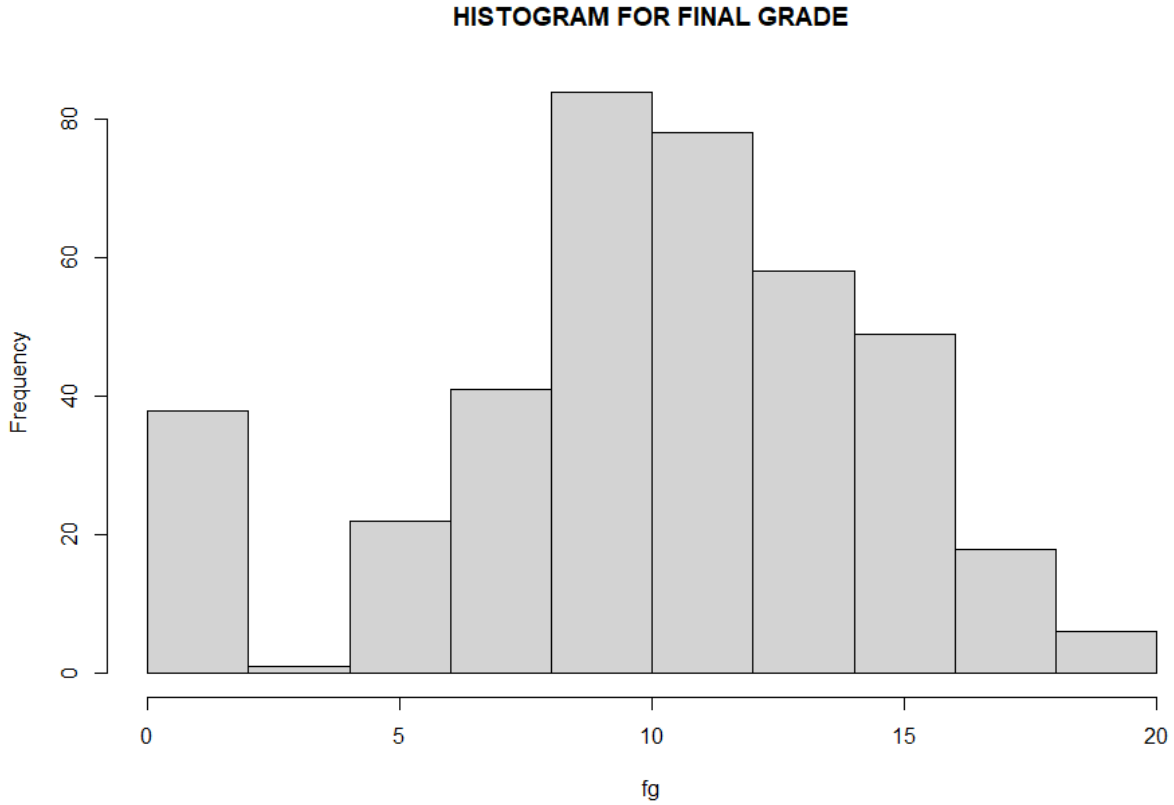
For better understanding further data analysis it will be crucial to know that the variables stands for:

1) school - student's school (binary: 'GP' or 'MS')
2) sex - student's sex (binary: 'F' - female or 'M' - male)
3) age - student's age (numeric: from 15 to 22)
4) address - student's home address type (binary: 'U' - urban or 'R' - rural)
5) famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6) Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7) Medu - mother's education (numeric: 0 - 4)
8) Fedu - father's education (numeric: 0 - 4)
9) Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services', 'at_home' or 'other')
10) Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services', 'at_home' or 'other')
11) reason - reason to choose this school
12) guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13) traveltime - home to school (num: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 h, or 4 - >1 h)
14) studytime - weekly study (num: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 h, or 4 - >1 h)
15) failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16) schoolsup - extra educational support (binary: yes or no)
17) famsup - family educational support (binary: yes or no)
18) paid - extra paid classes within the course subject (binary: yes or no)
19) activities - extra-curricular activities (binary: yes or no)
20) nursery - attended nursery school (binary: yes or no)
21) higher - wants to take higher education (binary: yes or no)
22) internet - Internet access at home (binary: yes or no)
23) romantic - with a romantic relationship (binary: yes or no)
24) famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25) freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26) goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27) Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28) Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29) health - current health status (numeric: from 1 - very bad to 5 - very good)
30) absences - number of school absences (numeric: from 0 to 93)
31) G1 - first period grade (numeric: from 0 to 20)
32) G2 - second period grade (numeric: from 0 to 20)
33) G3 - final grade (numeric: from 0 to 20, output target)

1) Data overview with G3 as target variable

summary of G3 values - final grades

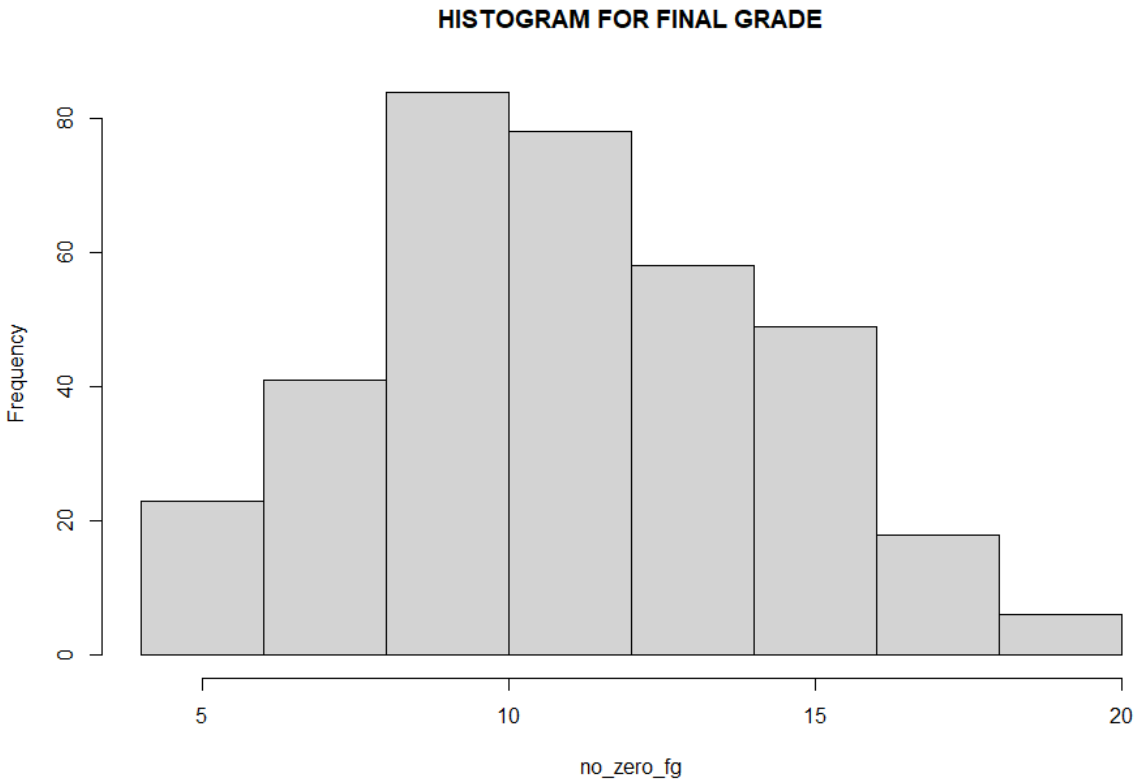
```
> summary(data)
     Min.    1st Qu.  Median   Mean     3rd Qu.    Max.
     0.00    8.00     11.00    10.42    14.00      20.00
```



As we can see there are grade values equals 0 (probably the person did not take the exam). For better data analysis we decide to remove “0” scores from dataset

Removing “0” values:

```
> no_zero_data <- data[data$G3 != 0,]

> summary(no_zero_fg)
   Min.    1st Qu.   Median    Mean    3rd Qu.    Max.
   4.00    9.00      11.00     11.52   14.00      20.00
```



2) Data analysis for each variable

At this stage we would like to clean our dataset from variables that has no dependencies for final grade score

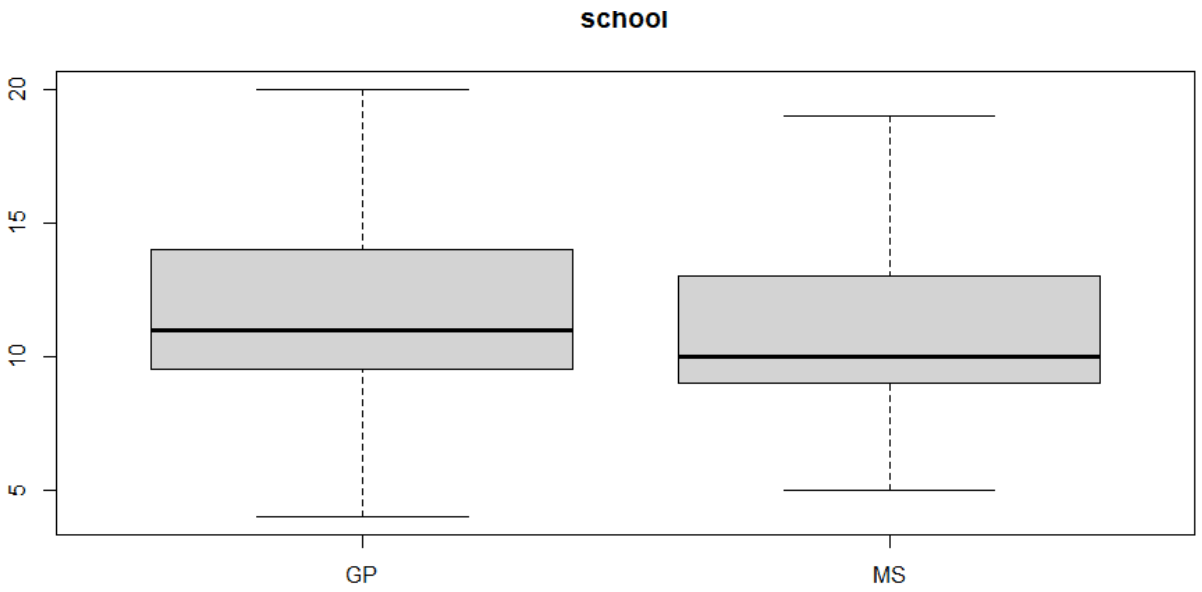
For each variable we are using Anova with respect to G3 to decide which predictors have no visible dependencies with final grade

17 Variables have no visible dependencies:

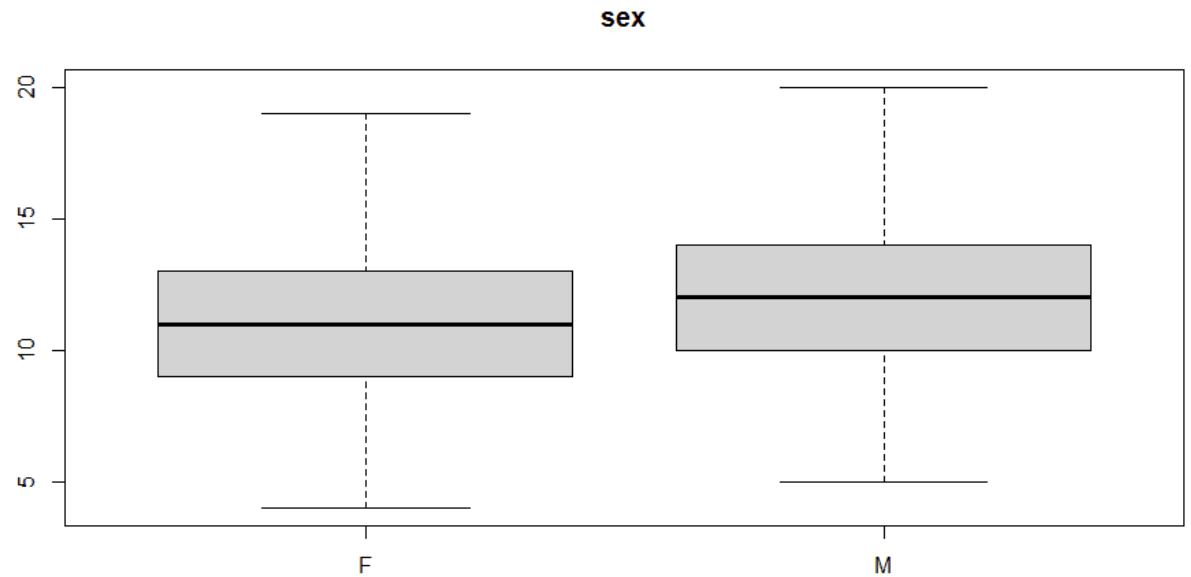
- school
- sex
- age
- famsize
- pstatus
- reason
- guardian
- traveltime
- famsup
- paid
- activities
- noursery
- romantic
- famrel
- freetime
- dalc
- health

Example of analysis first two covariates:

```
"SUMMARY OF ANOVA FOR:" "school"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1     26   25.93   2.499  0.115
Residuals   355   3683   10.38
```




```
"SUMMARY OF ANOVA FOR:" "sex"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1     39   38.93   3.765 0.0531
Residuals   355   3670   10.34
```

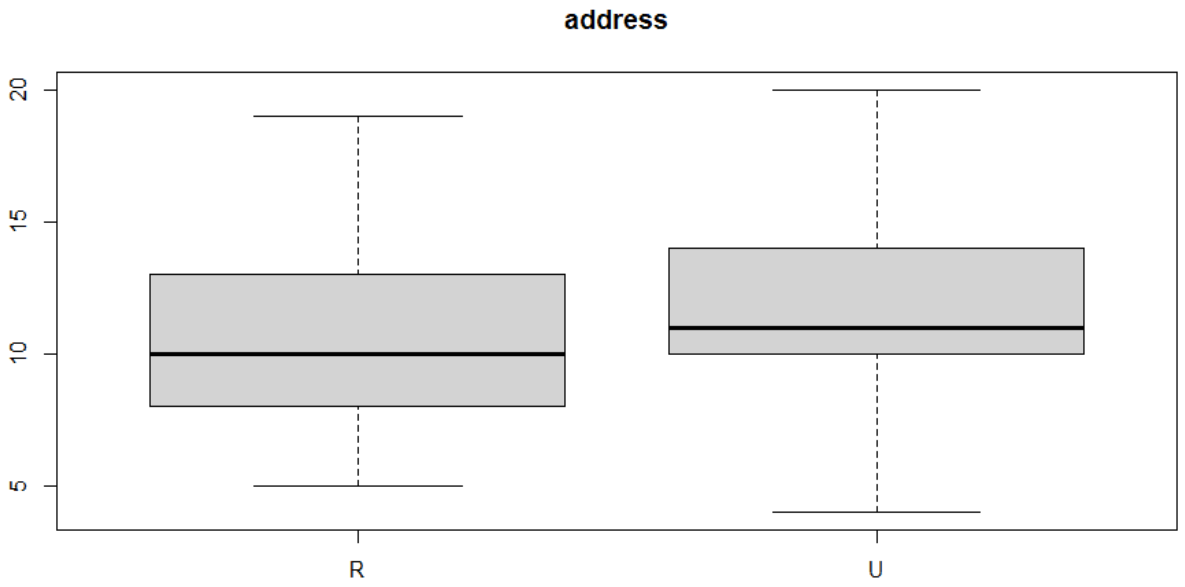


See analysis of all 17 variables in Appendix.

Variables with Sufficient Dependencies:

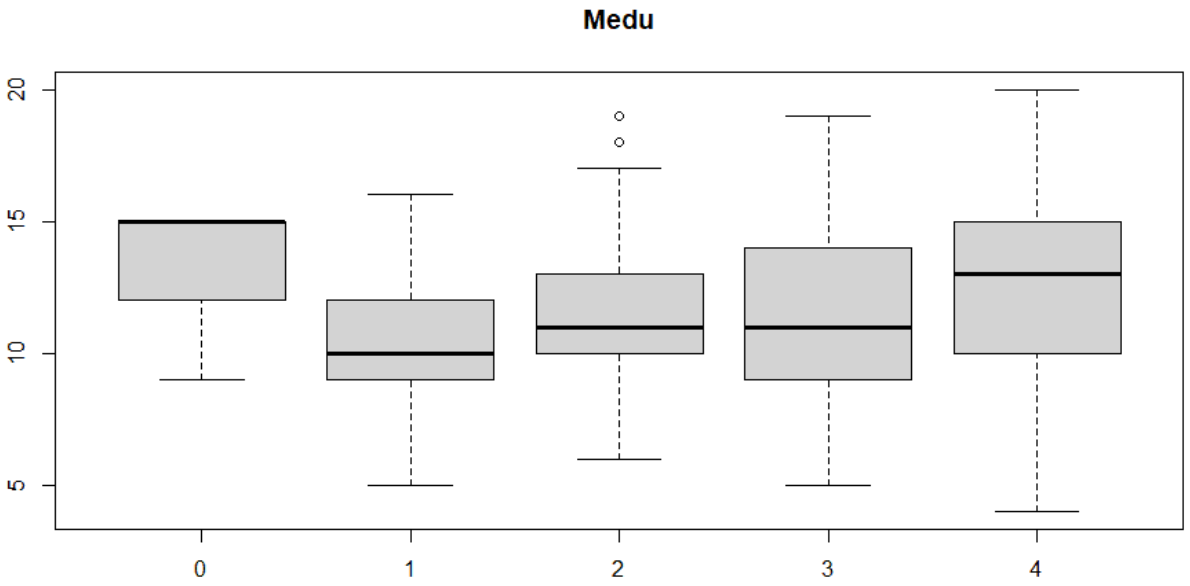
"SUMMARY OF ANOVA FOR:" **"address" (U - Urban, R - rural)**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	1	63	62.77	6.111	0.0139	*
Residuals	355	3646	10.27			



"SUMMARY OF ANOVA FOR:" **"Medu" - Mothers education**

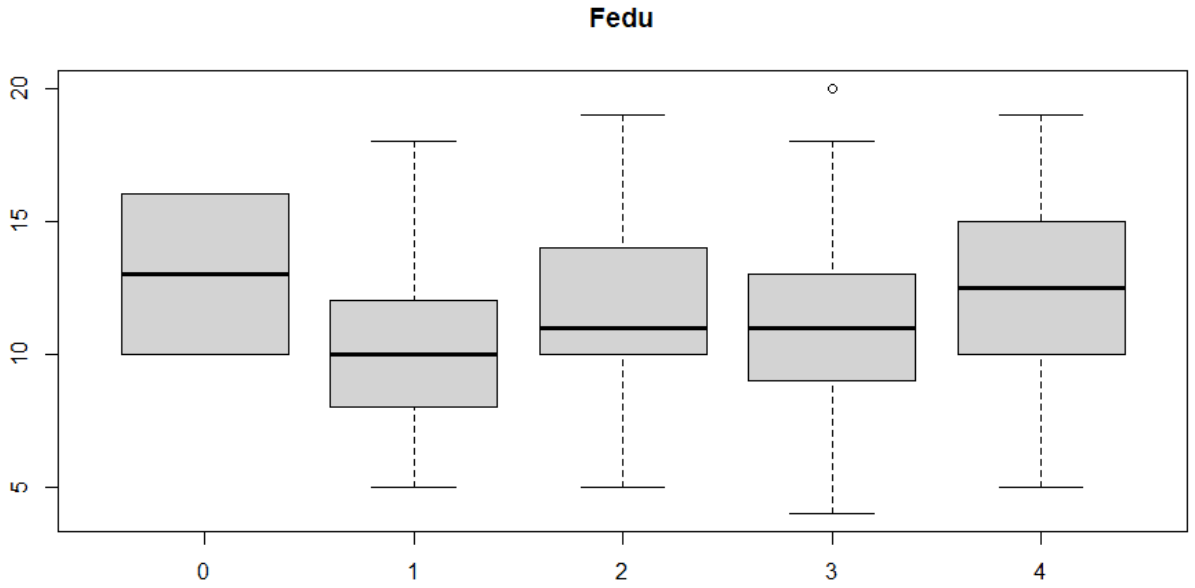
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	4	179	44.83	4.471	0.00155	**
Residuals	352	3530	10.03			



0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade,
3 - secondary education or 4 - higher education)

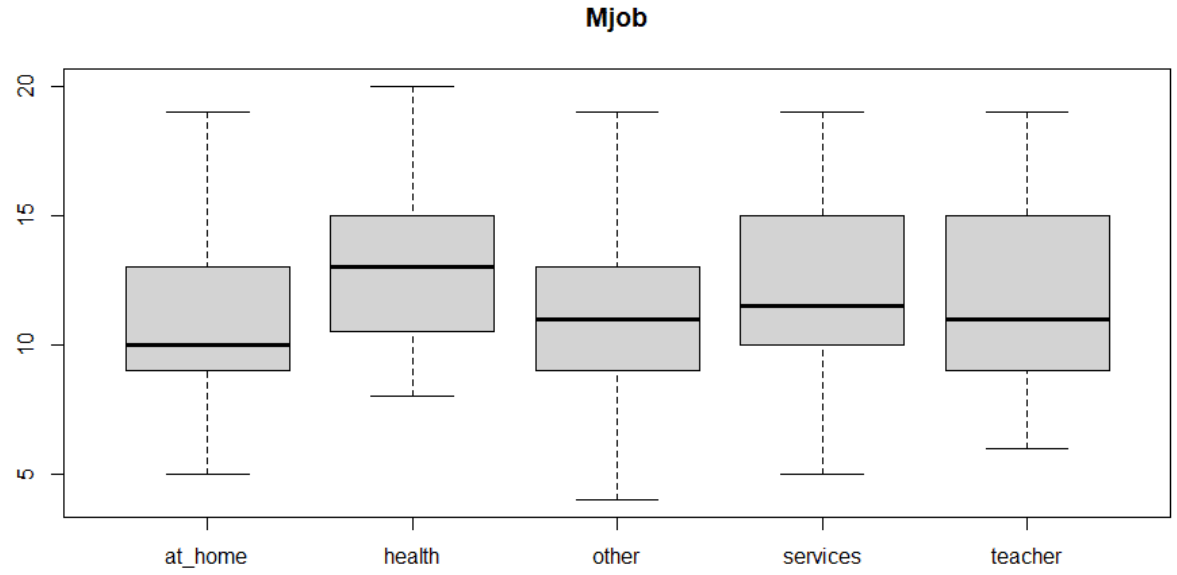
"SUMMARY OF ANOVA FOR:" **"Fedu" - Fathers education**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	4	139	34.63	3.414	0.00933	**
Residuals	352	3571	10.14			



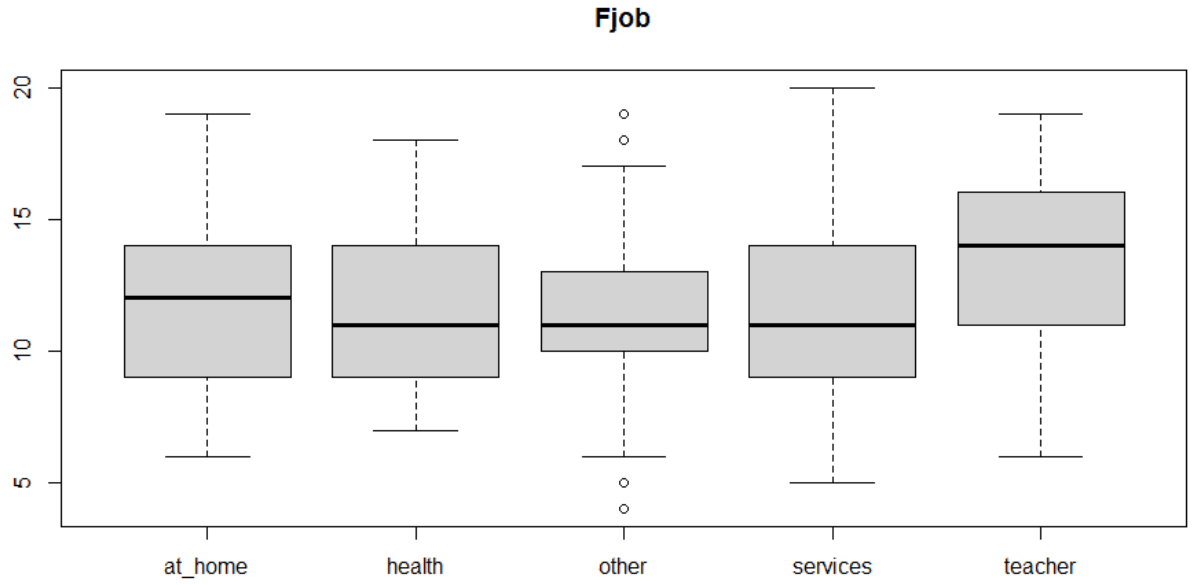
"SUMMARY OF ANOVA FOR:" **"Mjob" - Mothers job**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	4	171	42.72	4.25	0.00226	**
Residuals	352	3538	10.05			



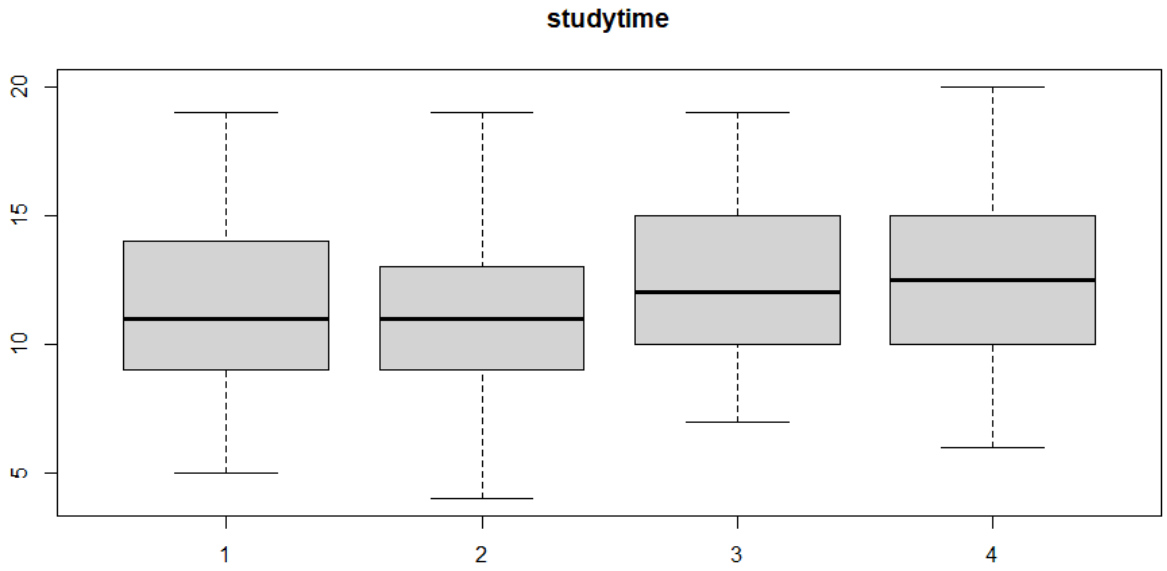
"SUMMARY OF ANOVA FOR:" **"Fjob" - Fathers job**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	4	101	25.36	2.474	0.0442	*
Residuals	352	3608	10.25			



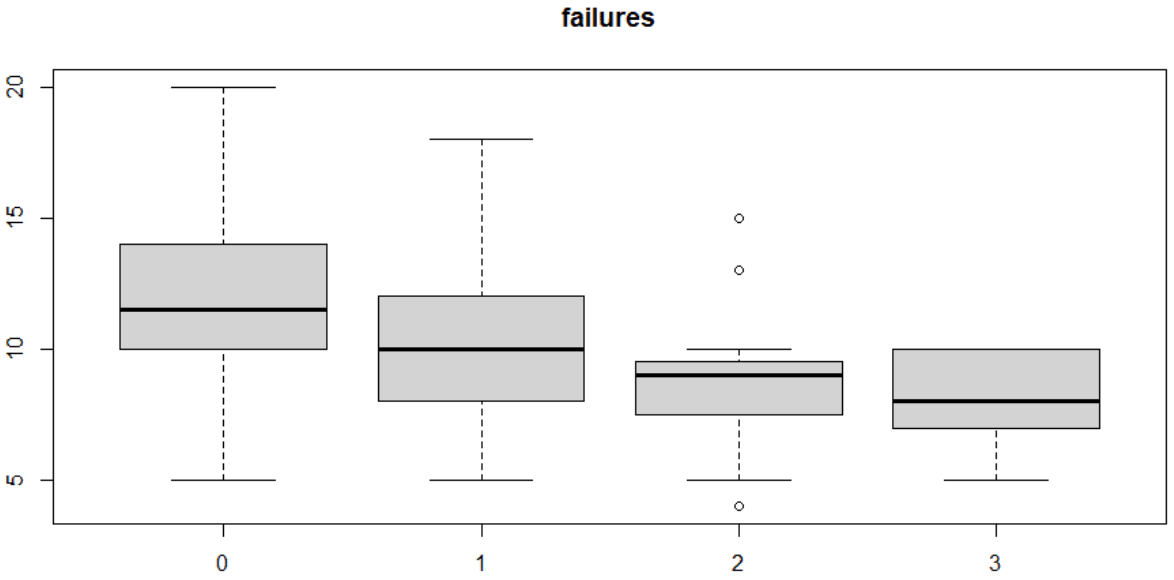
"SUMMARY OF ANOVA FOR:" **"studytime"**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	3	133	44.35	4.378	0.00482	**
Residuals	353	3576	10.13			



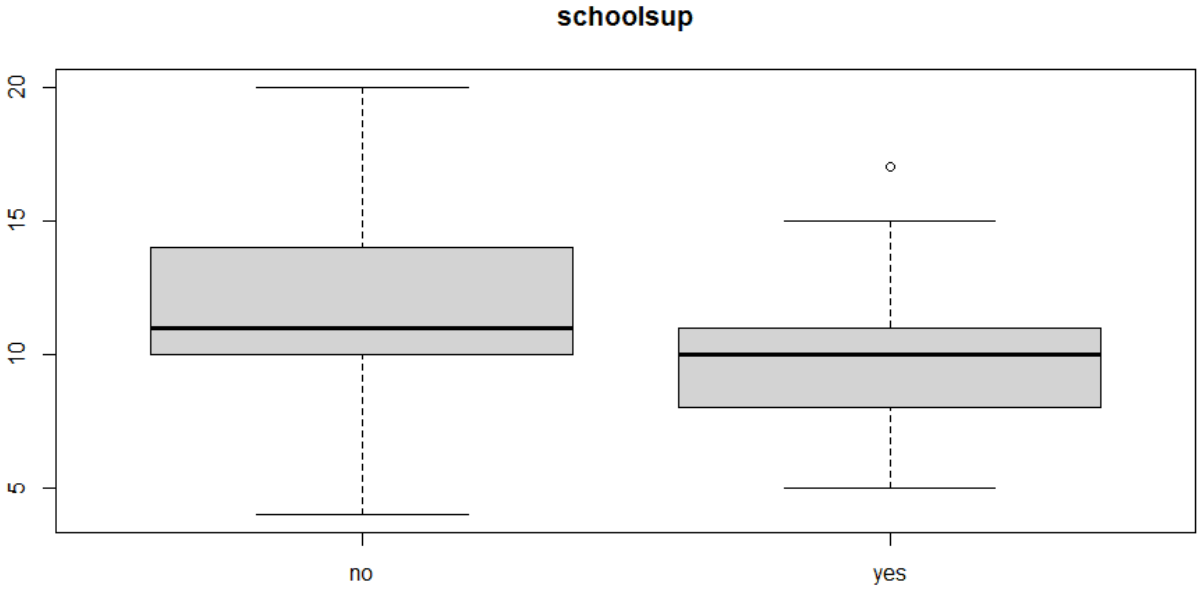
"SUMMARY OF ANOVA FOR:" "**failures**" - number of past class failures

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	3	330	110.03	11.49	3.3e-07	***
Residuals	353	3379	9.57			



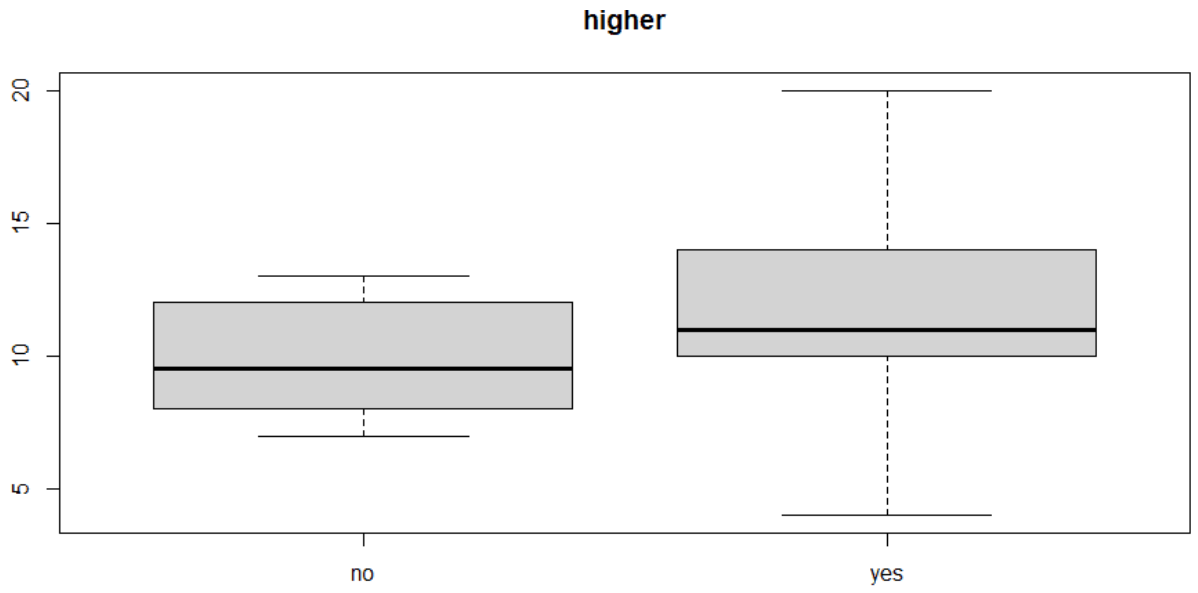
"SUMMARY OF ANOVA FOR:" "**schoolsup**" - extra edu support

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	1	211	210.74	21.39	5.27e-06	***
Residuals	355	3498	9.85			



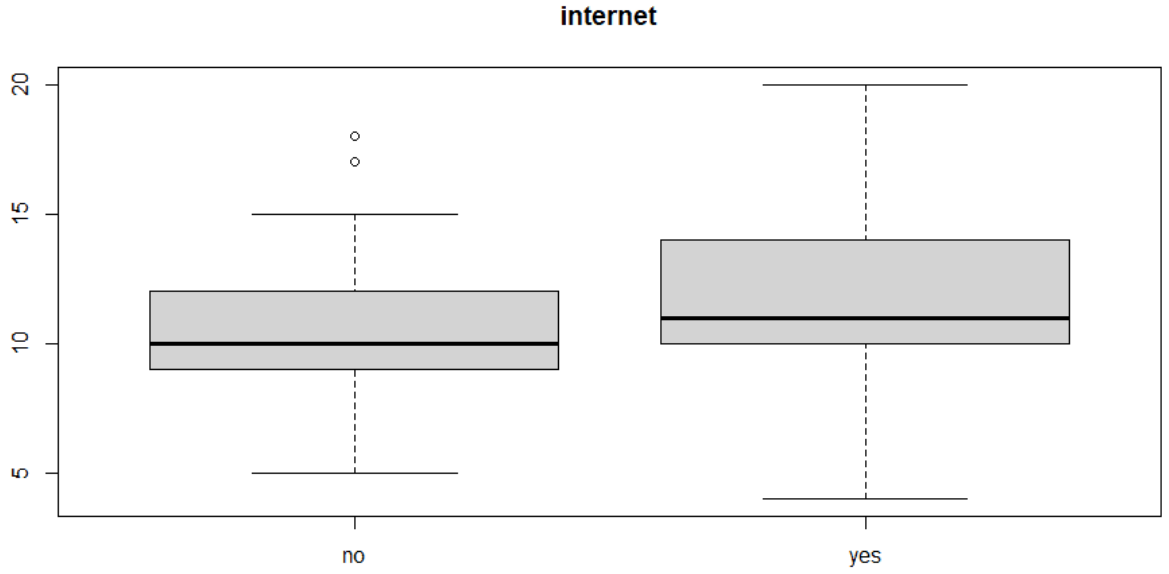
"SUMMARY OF ANOVA FOR:" **"higher"** - wants to take higher edu

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	1	48	47.71	4.626	0.0322	*
Residuals	355	3661	10.31			



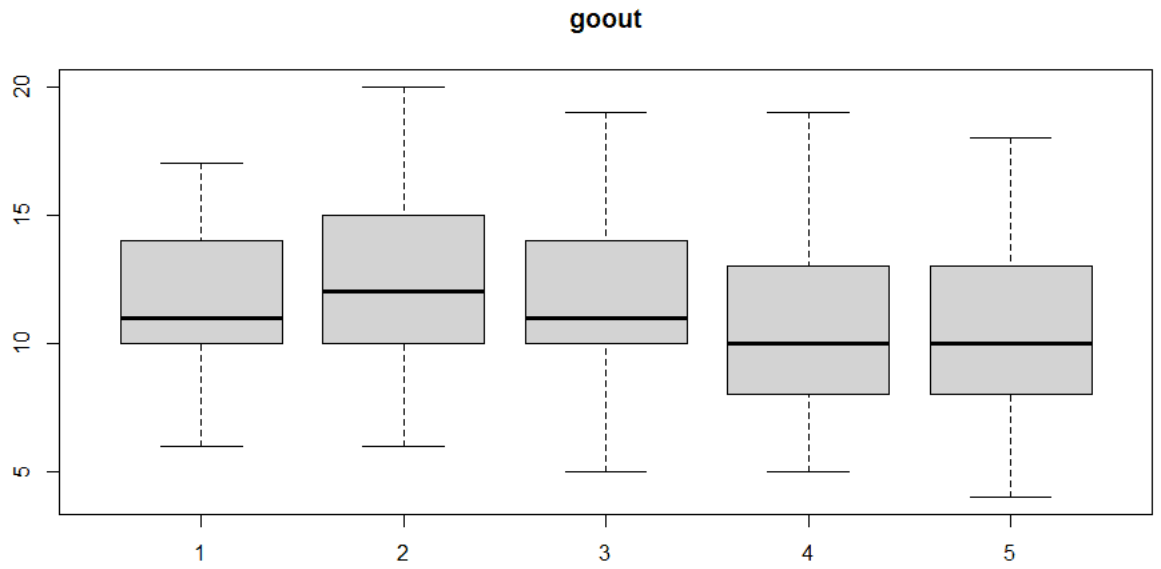
"SUMMARY OF ANOVA FOR:" **"internet"** - internet access

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	1	46	46.21	4.479	0.035	*
Residuals	355	3663	10.32			



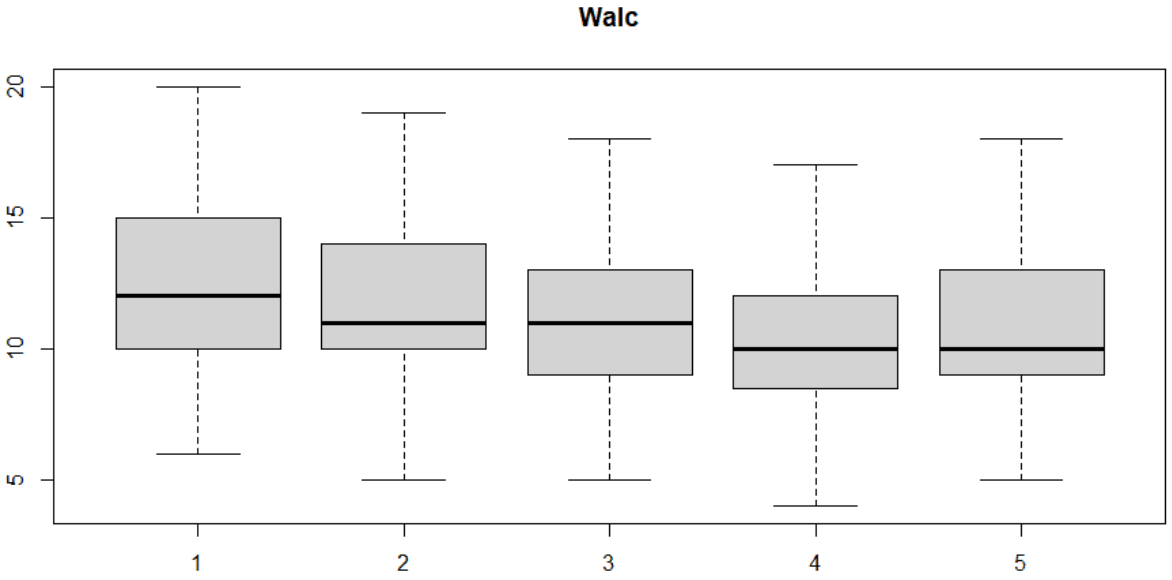
"SUMMARY OF ANOVA FOR:" **"goout"**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	4	136	33.91	3.341	0.0106	*
Residuals	352	3573	10.15			



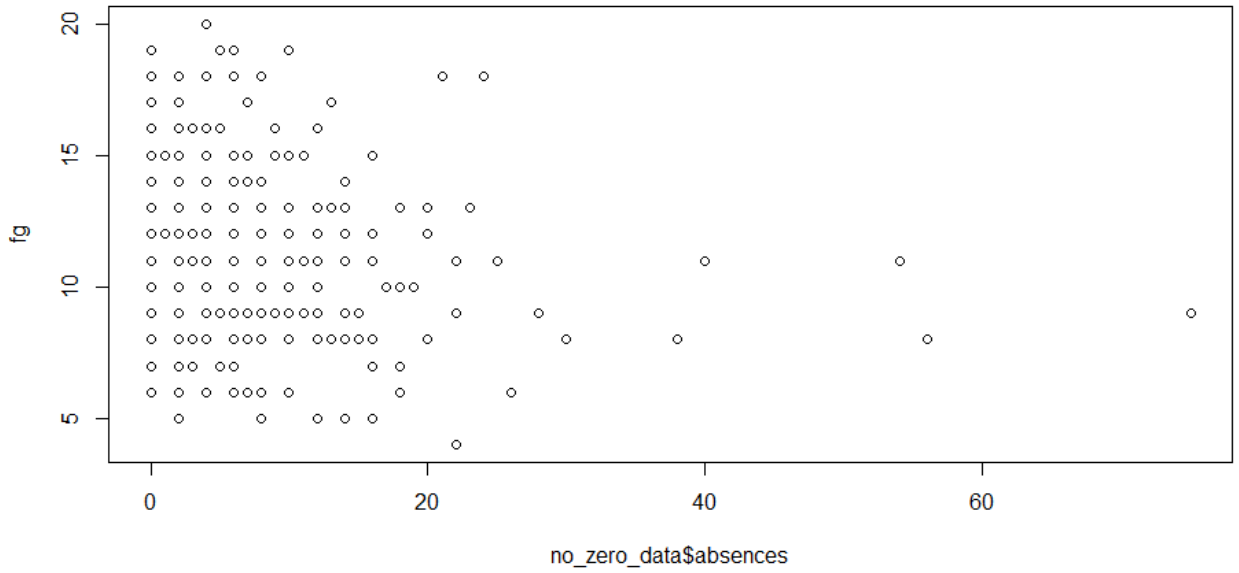
"SUMMARY OF ANOVA FOR:" **"Walc"** - weekend alcohol consumption

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
variable	4	156	38.88	3.851	0.00446	**
Residuals	352	3554	10.10			



1 - very low to 5 - very high

```
"SUMMARY OF ANOVA FOR:" "absences" - school absences,
                                     from 0 to 93
             Df Sum Sq Mean Sq F value Pr(>F)
variable     33    520  15.759   1.596 0.0231 *
Residuals   323   3189   9.873
```




```
"SUMMARY OF ANOVA FOR:" "G1"
             Df Sum Sq Mean Sq F value Pr(>F)
variable     15 2975.0  198.33   92.13 <2e-16 ***
Residuals   341  734.1    2.15

"SUMMARY OF ANOVA FOR:" "G2"
             Df Sum Sq Mean Sq F value Pr(>F)
variable     14   3468  247.68   350.7 <2e-16 ***
Residuals   342    242    0.71
```

Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful.

It is obvious that G1 and G2 are correlated with G3. We decided to remove them from our predictors

Summary:
We are left with 11 covariates.

```
"SUMMARY OF ANOVA FOR:" "address"
variable      1     63   62.77   6.111 0.0139 *

"SUMMARY OF ANOVA FOR:" "Medu"
variable      4    179   44.83   4.471 0.00155 **

"SUMMARY OF ANOVA FOR:" "Fedu"
variable      4    139   34.63   3.414 0.00933 **

"SUMMARY OF ANOVA FOR:" "Mjob"
variable      4    171   42.72    4.25 0.00226 **

"SUMMARY OF ANOVA FOR:" "Fjob"
variable      4    101   25.36   2.474 0.0442 *

"SUMMARY OF ANOVA FOR:" "studytime"
variable      3    133   44.35   4.378 0.00482 **

"SUMMARY OF ANOVA FOR:" "failures"
variable      3    330  110.03   11.49 3.3e-07 ***

"SUMMARY OF ANOVA FOR:" "schoolsup"
variable      1    211  210.74   21.39 5.27e-06 ***

"SUMMARY OF ANOVA FOR:" "higher"
variable      1     48   47.71   4.626 0.0322 *

"SUMMARY OF ANOVA FOR:" "internet"
variable      1     46   46.21   4.479  0.035 *

"SUMMARY OF ANOVA FOR:" "goout"
variable      4    136   33.91   3.341 0.0106 *

"SUMMARY OF ANOVA FOR:" "Walc”
variable      4    156   38.88   3.851 0.00446 **

"SUMMARY OF ANOVA FOR:" "absences"
variable     33    520  15.759   1.596 0.0231 *
```

Questions: Which of them affect final grade? Do we need all of them in our model?

To find out answers we are using LASSO

3) All possible regression and best subset. LASSO.

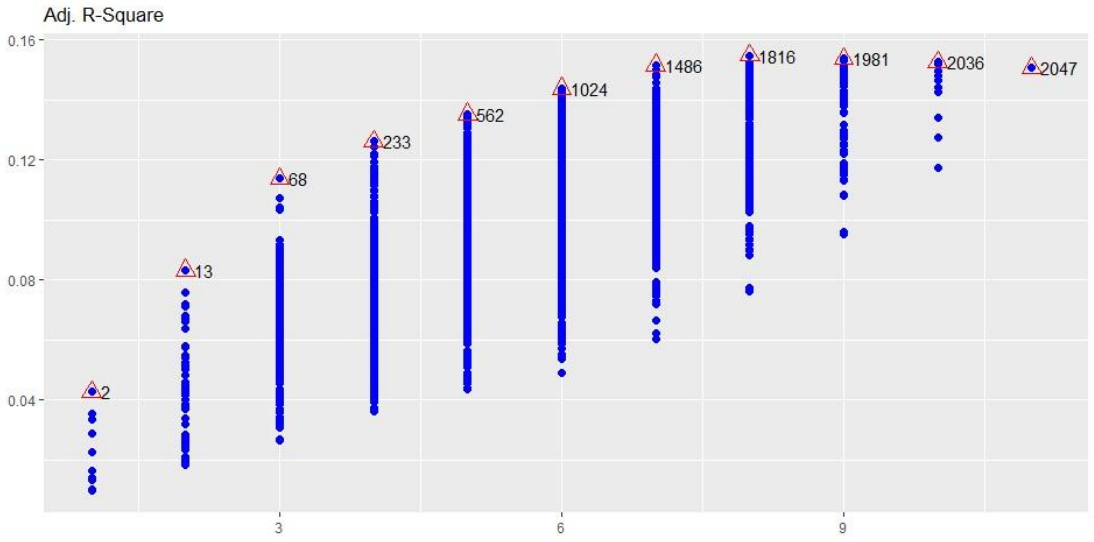
Purpose, choose less that 11 parameters

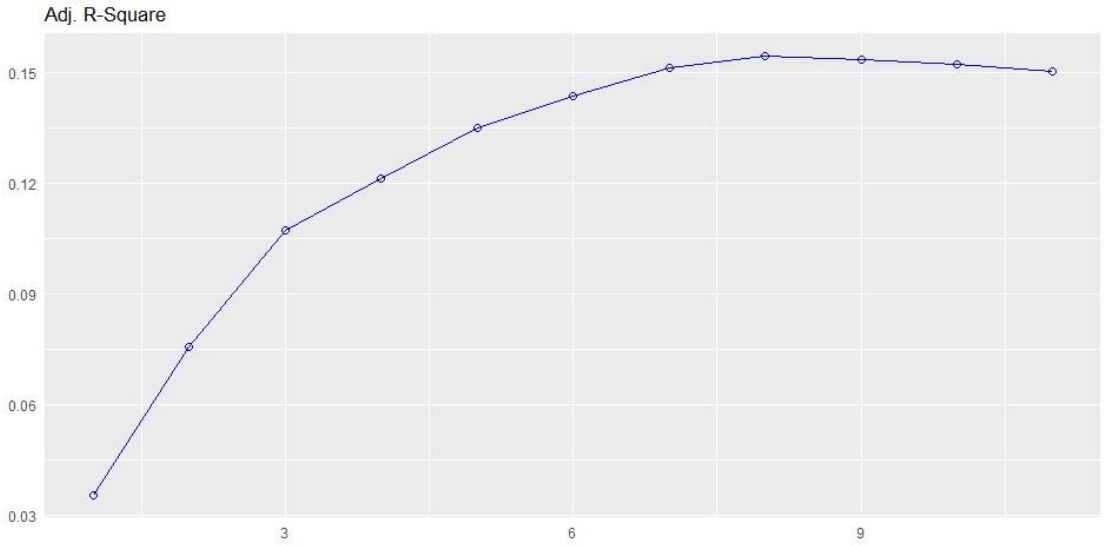
```
> library(olsrr)
> model <- lm(no_zero_data$G3 ~ no_zero_data$address + no_zero_data$Medu
            + no_zero_data$Fedu + no_zero_data$Mjob + no_zero_data$Fjob
            + no_zero_data$studytime + no_zero_data$higher
            + no_zero_data$internet + no_zero_data$goout
            + no_zero_data$Walc + no_zero_data$absences,no_zero_data)
> models <- ols_step_all_possible(model)
> plot(models)
```

Subsets Regression Summary

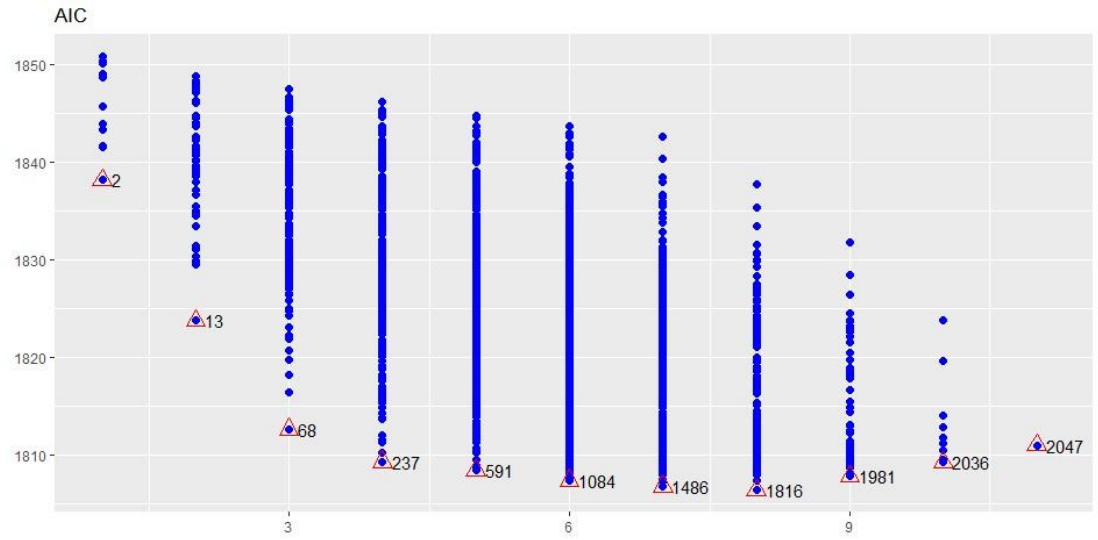
Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC
1	0.0461	0.0352	0.0192	46.8042	1843.9463
2	0.0888	0.0758	0.0569	30.9152	1829.6041
3	0.1223	0.1073	0.088	18.8371	1818.1958
4	0.1460	0.1213	0.0893	10.9192	1816.4380
5	0.1620	0.1353	0.1013	6.2233	1811.6959
6	0.1727	0.1438	0.1071	3.7366	1809.1061
7	**0.1826**	**0.1516**	**0.113**	**1.6000**	**1806.8214**
8	0.1880	0.1548	0.1128	1.3210	1806.4387
9	0.1893	0.1537	0.1088	2.7564	1807.8460
10	0.1906	0.1525	0.1051	4.2373	1809.3001
11	0.1912	0.1506	0.1025	6.0000	1811.0503

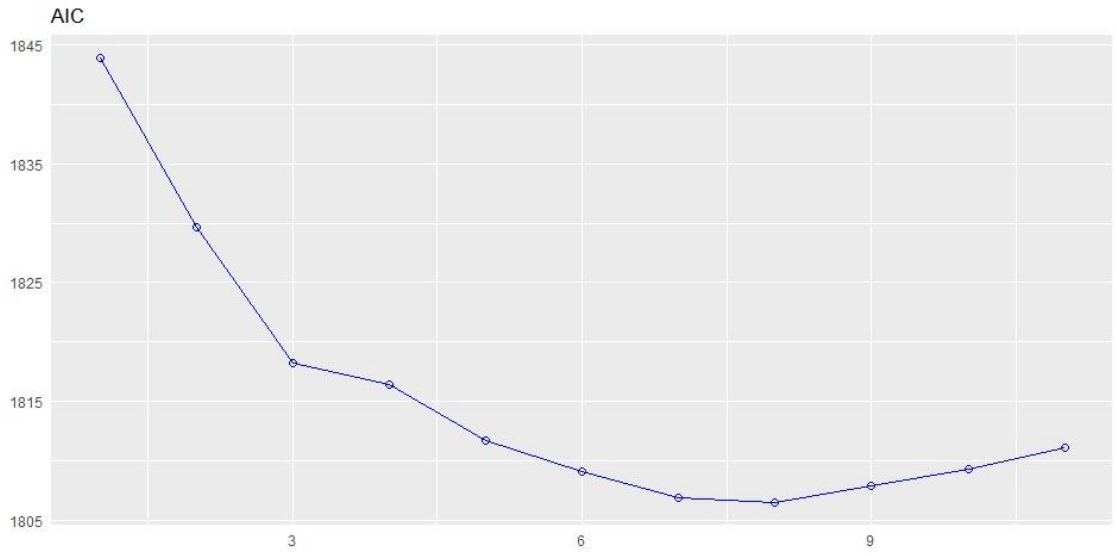
Adj. R-Square:




Adj. R-Square
0.15
0.12
0.09
0.06
0.03
3
6
9

AIC


AIC
1850
1840
1830
1820
1810
2
13
68
237
591
1084
1486
1816
1981
2036
2047
3
6
9


AIC
1845
1835
1825
1815
1805
3
6
9

In the above best subset summary statement we can see best predictors for each number of variables.

```
> best_subset <- ols_step_best_subset(model)
> plot(best_subset)
                    Best Subsets Regression
--------------------------------------------------------------------
M Index    Predictors
     1 Mjob
     2 Mjob   absences
     3 Mjob   goout   absences
     4 Mjob   Fjob   goout   absences
     5 Mjob   Fjob   studytime   goout   absences
     6 Medu   Mjob   Fjob   studytime   goout   absences
     7 Medu   Mjob   Fjob   studytime   internet   goout   absences
     8 address   Medu   Mjob   Fjob   studytime   internet  goout
       absences
     9 address   Medu   Fedu   Mjob   Fjob   studytime   internet
       goout absences
    10 address   Medu   Fedu   Mjob   Fjob   studytime   internet
       goout Walc   absences
    11 address   Medu   Fedu   Mjob   Fjob   studytime   higher
       internet   goout   Walc   absences
```

After analysis we choose to keep 7 predictors:

```
Medu   Mjob   Fjob   studytime   internet   goout   absences

     Coefficients:
                    Estimate   Std. Error t_value Pr(>|t|)
     (Intercept)    11.41945    1.03418  11.042  < 2e-16 ***
      Medu           0.39901    0.19427   2.054   0.0407 *
      Mjob-health    1.39780    0.79993   1.747   0.0815 .
      Mjob-other    -0.20565    0.52727  -0.390   0.6968
      Mjob-services  0.80140    0.58419   1.372   0.1710
      Mjob-teacher  -0.19291    0.74104  -0.260   0.7948
      Fjob-health   -1.19947    1.03470  -1.159   0.2472
      Fjob-other    -0.62297    0.76201  -0.818   0.4142
      Fjob-services -0.77694    0.79049  -0.983   0.3264
      Fjob-teacher   1.02319    0.96341   1.062   0.2890
      studytime      0.40503    0.19511   2.076   0.0386 *
      internet-yes   0.91731    0.45075   2.035   0.0426 *
      goout         -0.57564    0.14623  -3.937   0.0001 ***
      absences      -0.08097    0.01966  -4.118 4.78e-05 ***
```

Results: Going out and absence have a negative impact. Time for study, and internet access has a positive effect. Parents job is also important

4) Data analysis for the sake of grade improvement

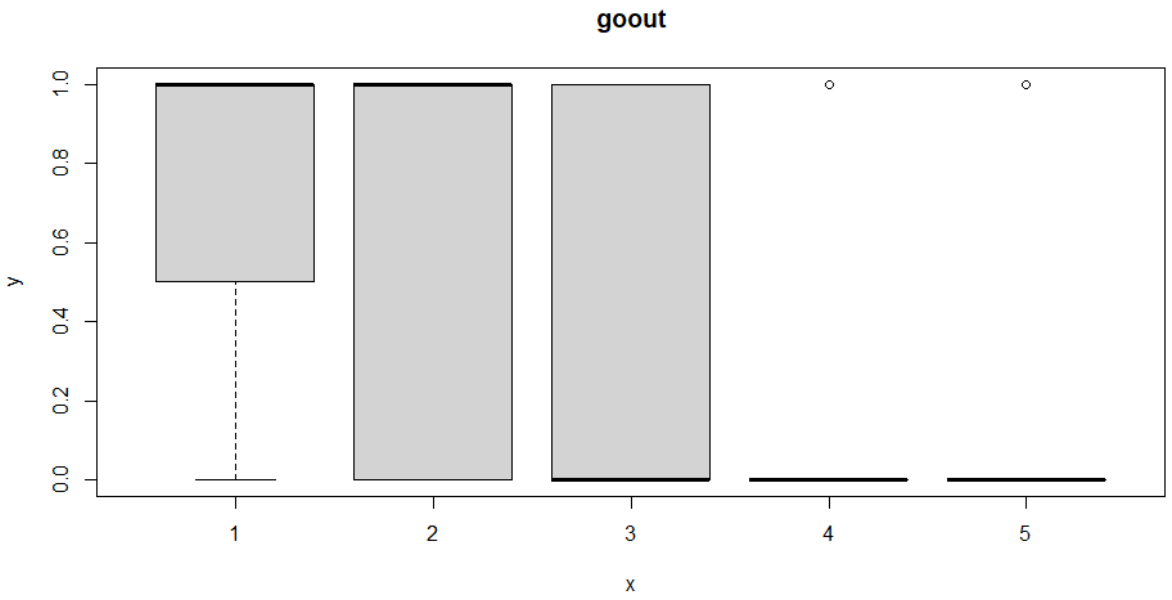
In this section we analyze which predictors influenced grade improvement.

- Grades in Portugal are distributed on scale of 0 - 20, with 10 being the lowest passing grade.
- We choose students who achieved less than 10 points in G1
- Out of these students, we select only those who have passed in G3 (10 or more points)
- Our goal is to find covariates with sufficient dependency on improvement

```
          Summary for new variable - improvement
 Min.    1st Qu.  Median    Mean    3rd Qu.    Max.
0.0000  0.0000   0.0000    0.3423  1.0000     1.0000
```

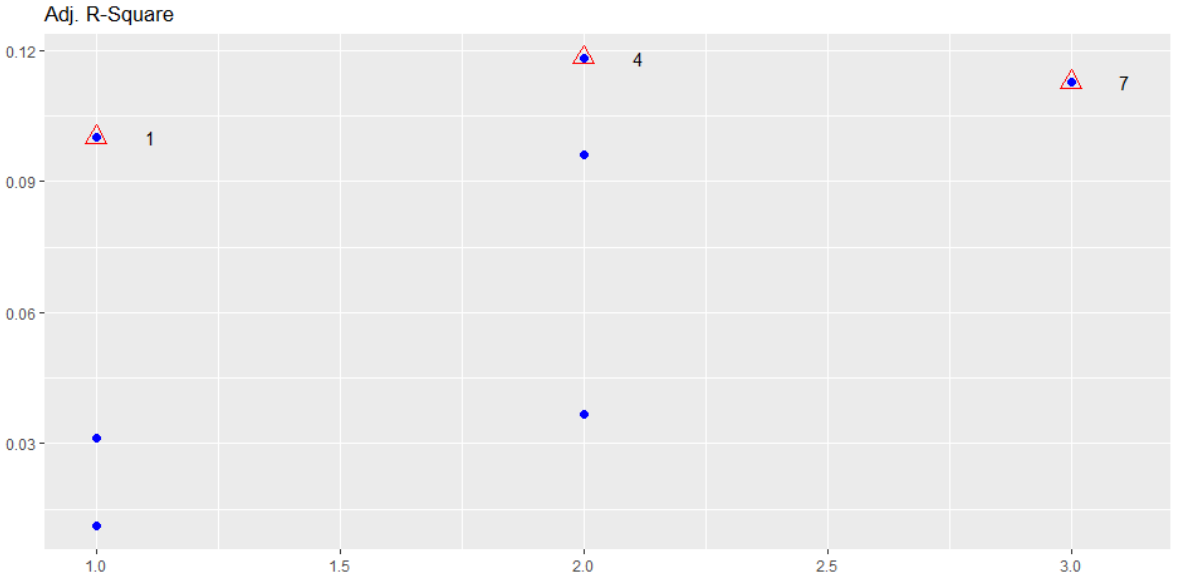
From 357 students 111 achieved score below 50% in G1
Only 3 variables have visible effect on improvement

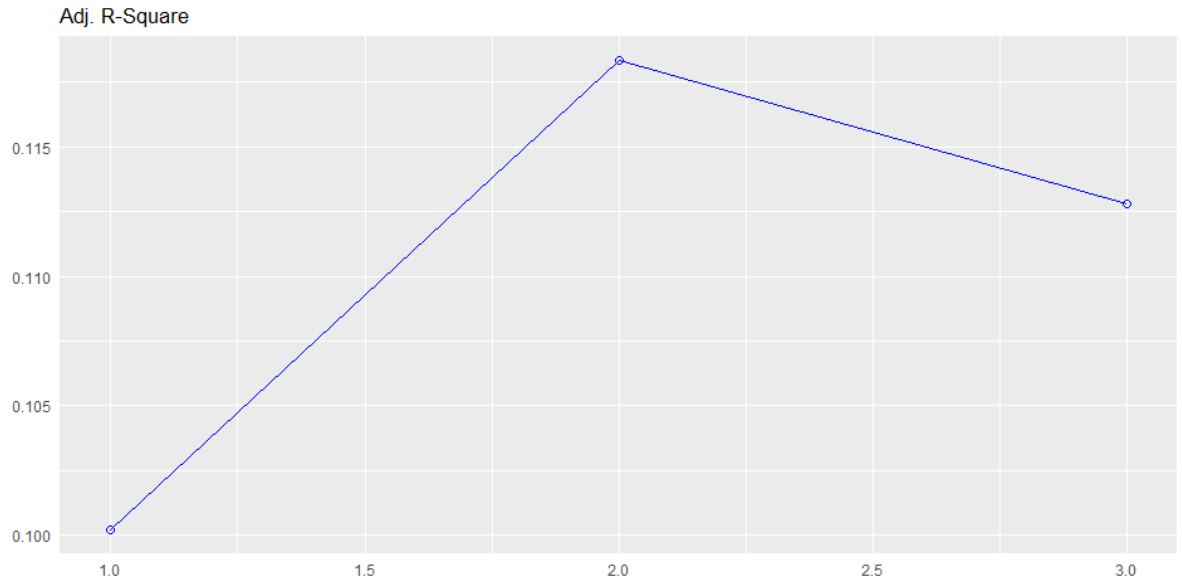
```
"SUMMARY OF ANOVA FOR:" "failures"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      3  1.934  0.6446   2.991 0.0342 *
Residuals   107 23.057  0.2155

"SUMMARY OF ANOVA FOR:" "romantic"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1  1.004  1.0043   4.564 0.0349 *
Residuals   109 23.987  0.2201

"SUMMARY OF ANOVA FOR:" "goout"
             Df Sum Sq Mean Sq F value  Pr(>F)
variable      4  3.177  0.7942   3.859 0.00575 **
Residuals   106 21.814  0.2058
```



All possible regression and best subset. LASSO.

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC
1	0.1084	0.1002	0.0756	3.5475	142.7697
2	0.1344	**0.1184**	0.0873	2.3203	141.4811
3	0.1370	0.1128	0.0694	4.0000	143.1493





Based on above data and plots we select 2 predictors for our model:
`romantic` and `gout`

```
           Best Subsets Regression
Model Index    Predictors
      1          goout
      2          romantic  goout
      3          failures  romantic  goout
```

Final model:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.82252    0.13332   6.169 1.22e-08 ***
romantic-yes      -0.17760    0.09855  -1.802  0.07432 .
goout             -0.13461    0.03926  -3.428  0.00086 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4475 on 108 degrees of freedom
Multiple R-squared:  0.1344,    Adjusted R-squared:  0.1184
F-statistic: 8.384 on 2 and 108 DF,  p-value: 0.0004123
```

Results: Going out and being in a romantic relationship have a negative impact on the improvement.

Go out	**Improved**	**Not improved**
1. (very low)	3	1
2.	16	12
3.	10	23
4.	6	24
5. (very high)	3	13

In relationship	**Improved**	**Not improved**
yes	5	24
no	34	48

APPENDIX

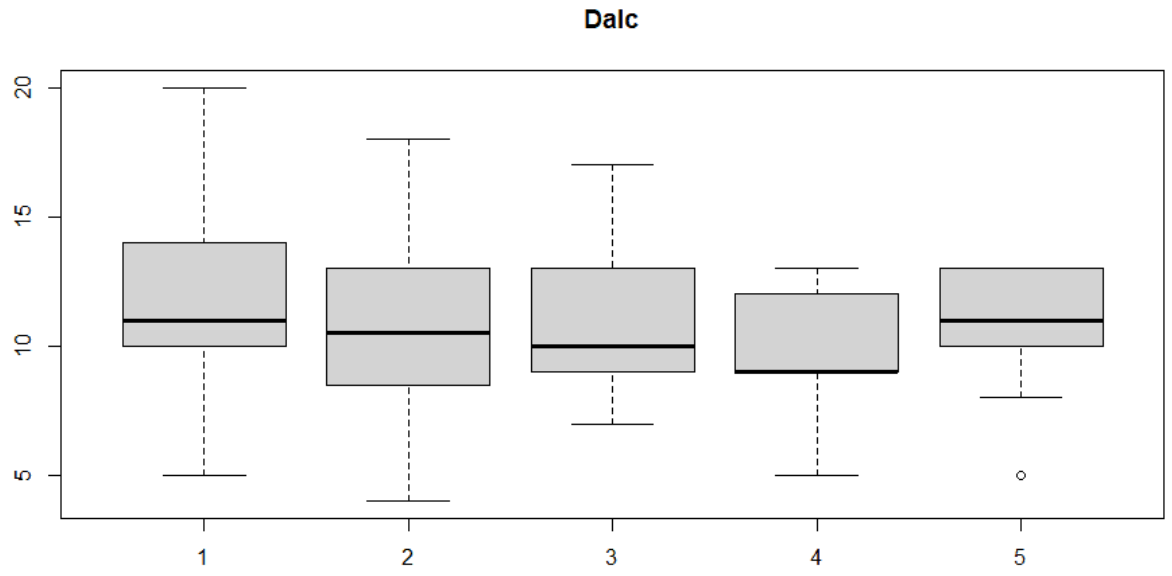
Variables with Sufficient Dependencies on G3

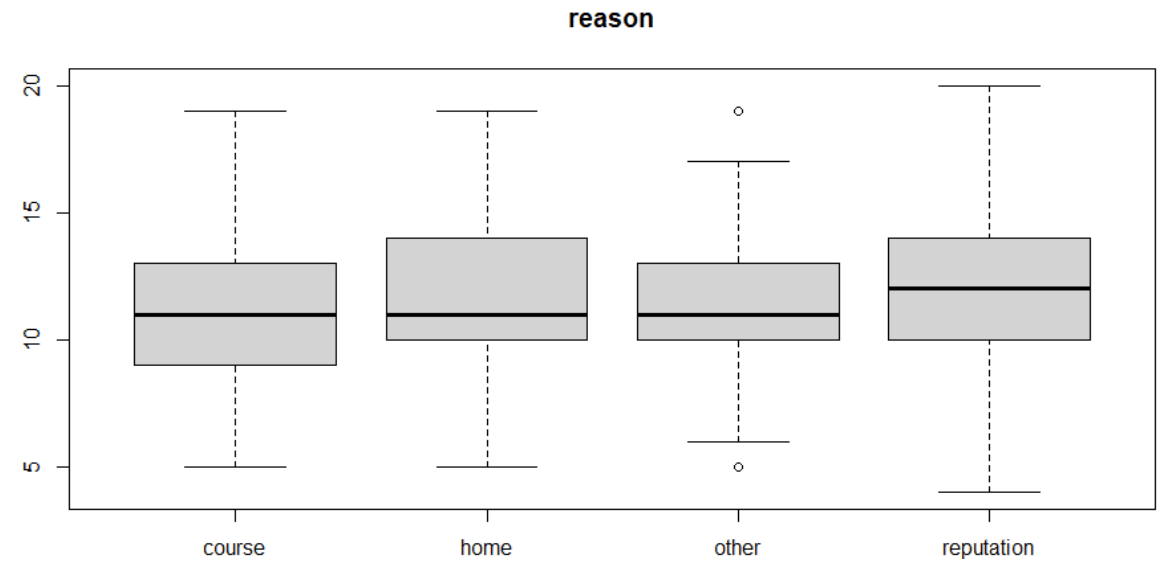
ANOVA:

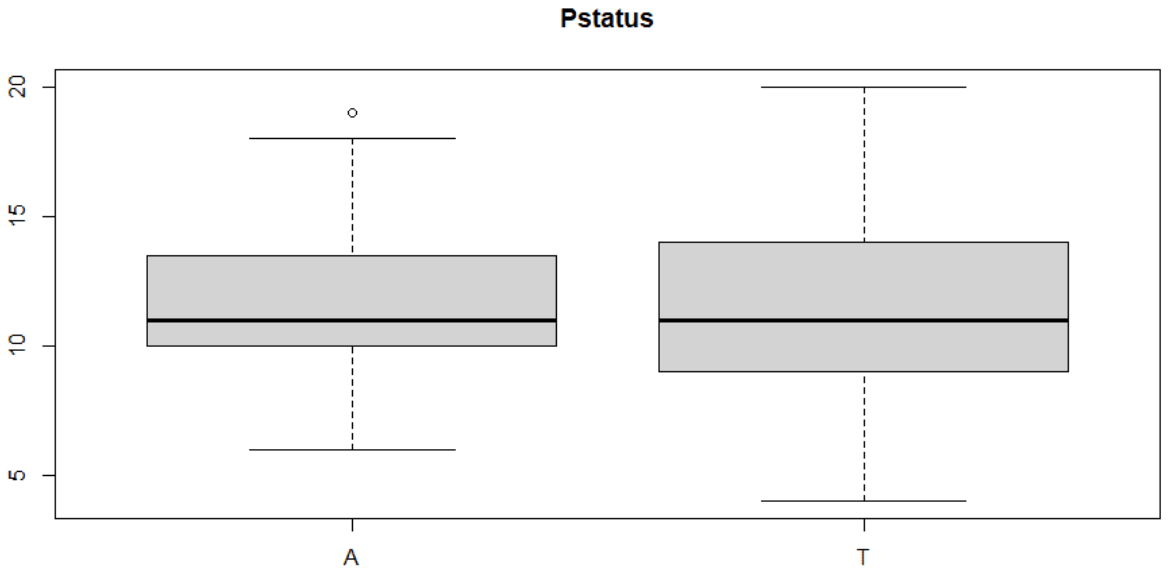
```
"SUMMARY OF ANOVA FOR:" "school"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1     26   25.93   2.499  0.115
Residuals   355   3683   10.38

"SUMMARY OF ANOVA FOR:" "sex"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1     39   38.93   3.765 0.0531
Residuals   355   3670   10.34

"SUMMARY OF ANOVA FOR:" "age"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      7    133   18.98   1.852 0.0767
Residuals   349   3576   10.25

"SUMMARY OF ANOVA FOR:" "famsize"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1      6   5.859   0.562  0.454
Residuals   355   3703  10.432

"SUMMARY OF ANOVA FOR:" "Pstatus"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1      3   2.637   0.253  0.616
Residuals   355   3706  10.441

"SUMMARY OF ANOVA FOR:" "reason"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      3     15   4.951   0.473  0.701
Residuals   353   3694  10.465

"SUMMARY OF ANOVA FOR:" "guardian"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      2     20   10.08   0.968  0.381
Residuals   354   3689   10.42

"SUMMARY OF ANOVA FOR:" "traveltime"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      3     42   14.01   1.348  0.258
Residuals   353   3667   10.39

"SUMMARY OF ANOVA FOR:" "famsup"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1     17    16.8   1.615  0.205
Residuals   355   3692    10.4
```

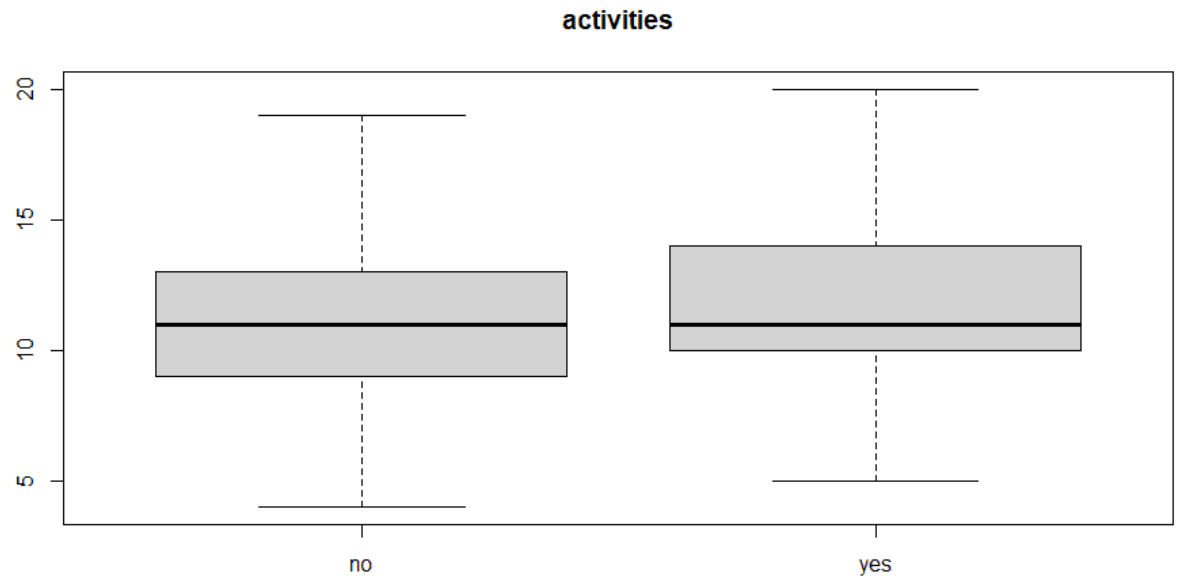
```
"SUMMARY OF ANOVA FOR:" "paid"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1      3   3.098   0.297  0.586
Residuals   355   3706  10.439

"SUMMARY OF ANOVA FOR:" “activities"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1     13   12.74   1.223  0.269
Residuals   355   3696   10.41

"SUMMARY OF ANOVA FOR:" "nursery"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1      3   2.613    0.25  0.617
Residuals   355   3706  10.441


"SUMMARY OF ANOVA FOR:" "romantic"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      1      9   9.252   0.888  0.347
Residuals   355   3700  10.422

"SUMMARY OF ANOVA FOR:" "famrel"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      4     35   8.645   0.828  0.508
Residuals   352   3674  10.439

"SUMMARY OF ANOVA FOR:" "freetime"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      4     61   15.29   1.476  0.209
Residuals   352   3648   10.36

"SUMMARY OF ANOVA FOR:" "Dalc"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      4     96   23.91    2.33 0.0558
Residuals   352   3613   10.27

"SUMMARY OF ANOVA FOR:" "health"
             Df Sum Sq Mean Sq F value Pr(>F)
variable      4     74   18.49    1.79   0.13
Residuals   352   3635   10.33
```

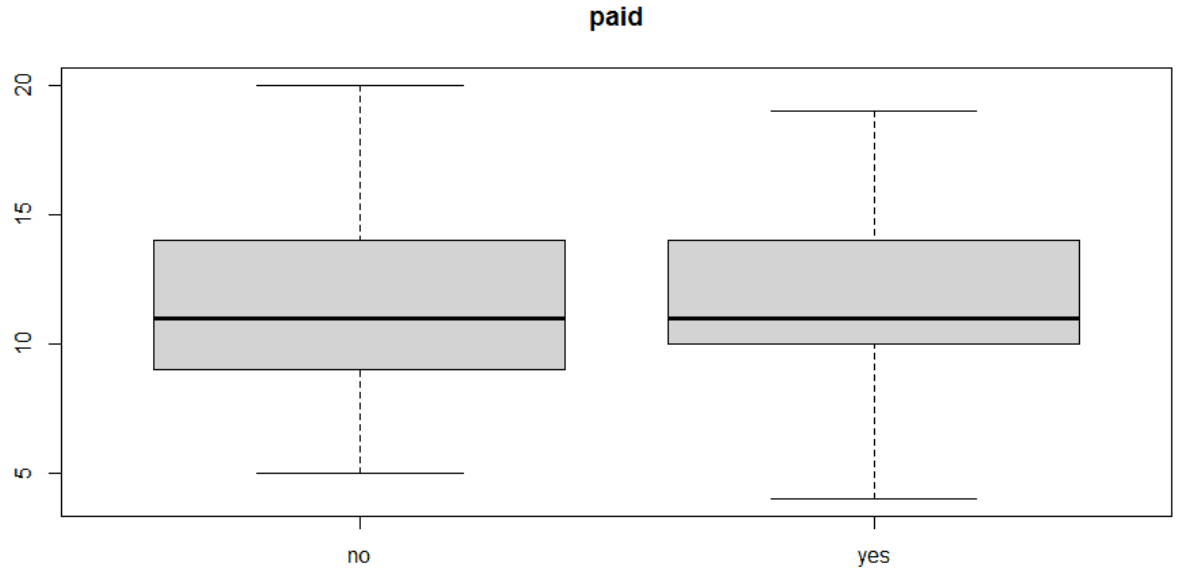
Plots:

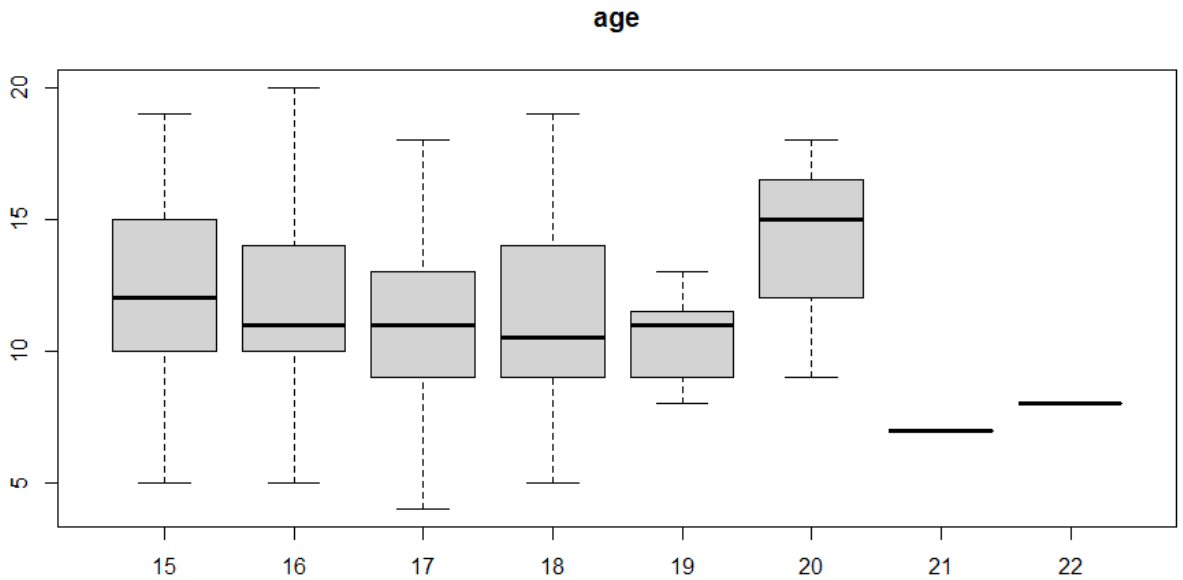

Dalc
20
15
10
5
1
2
3
4
5

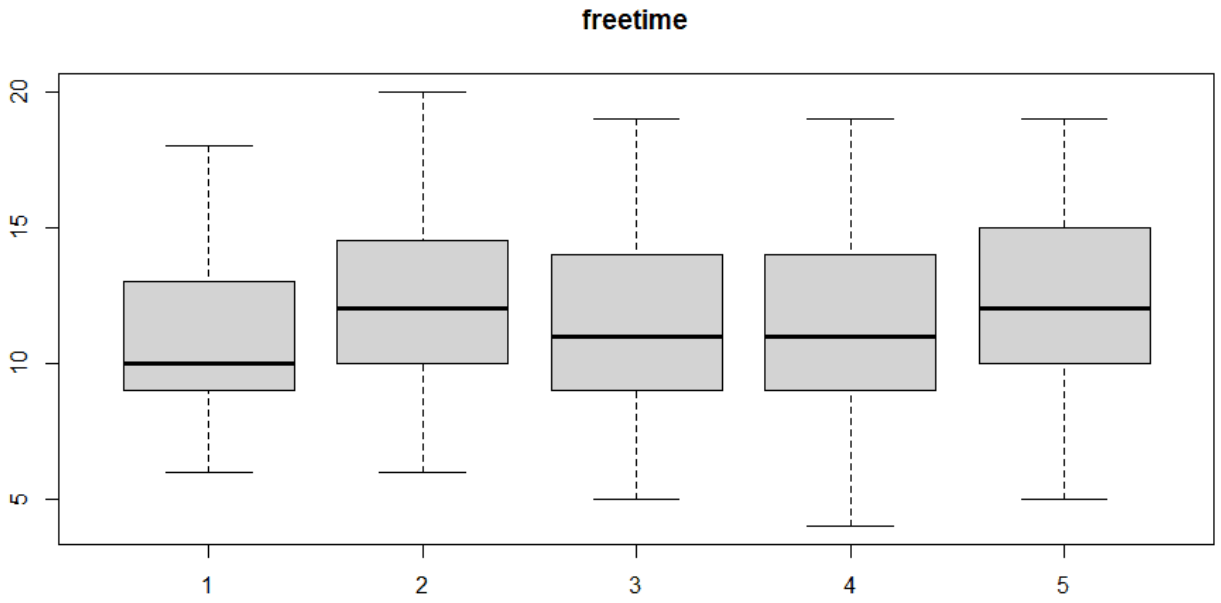

reason
20
15
10
5
course
home
other
reputation

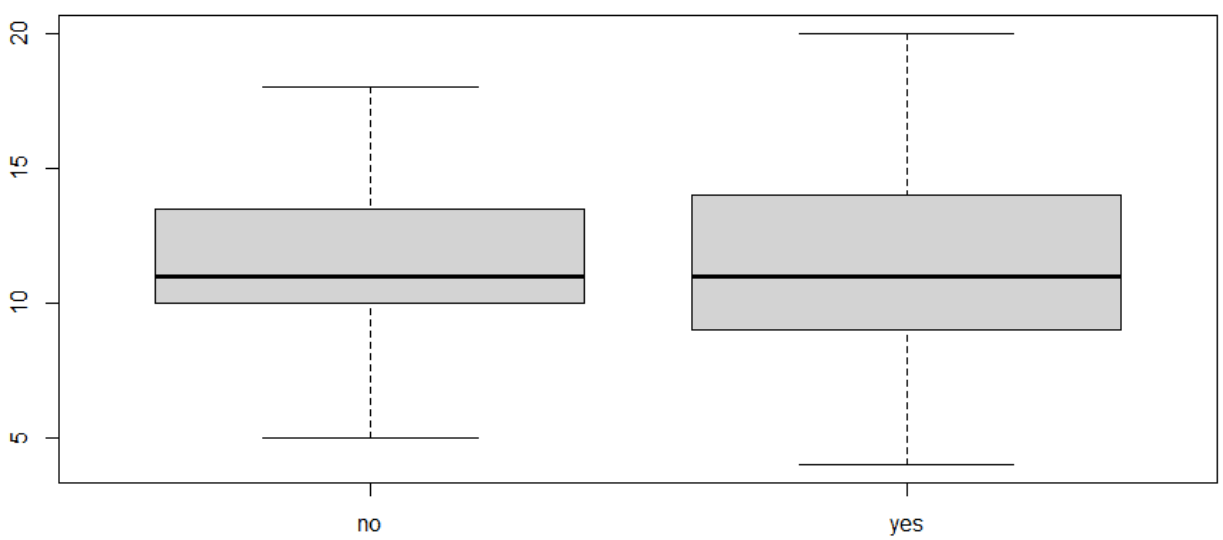

Pstatus
20
15
10
5
A
T

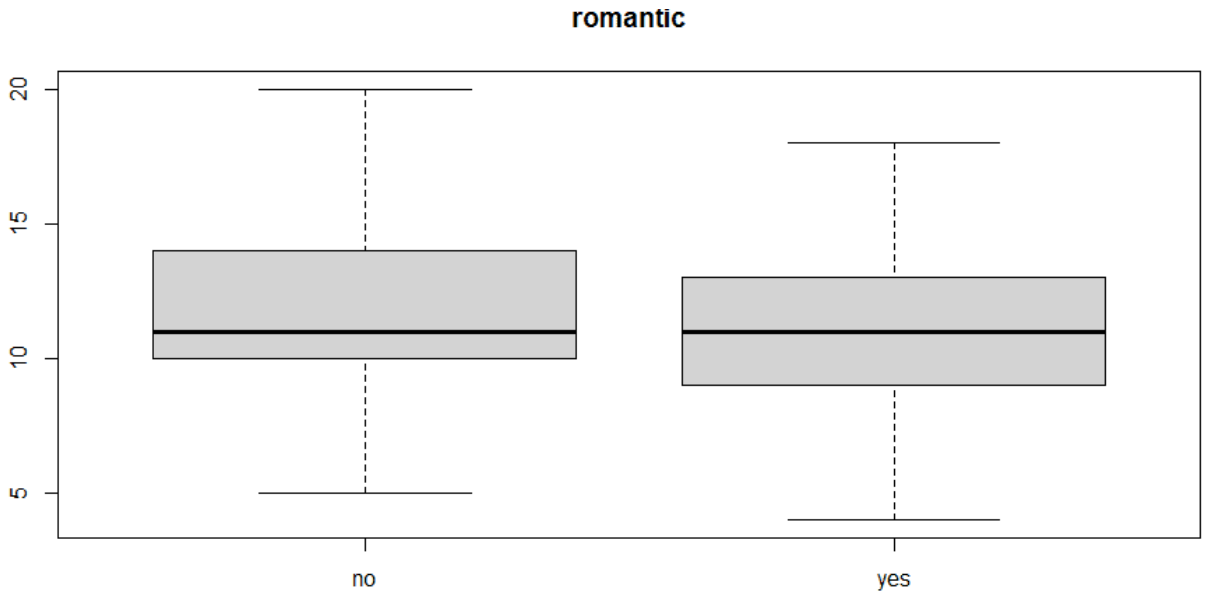

activities
20
15
10
5
no
yes

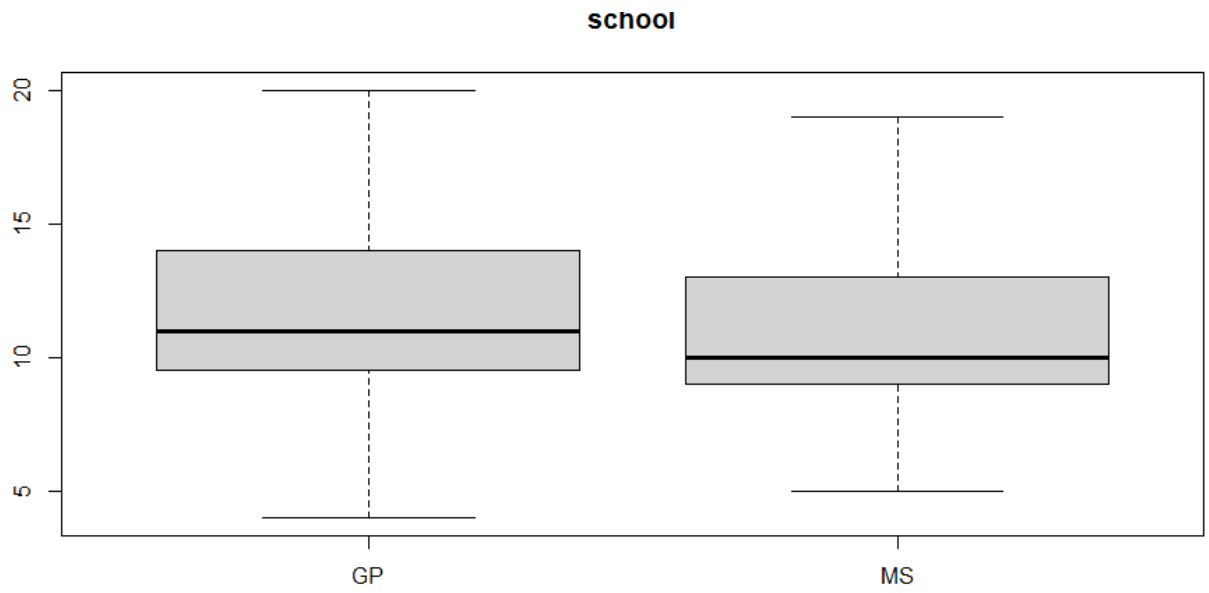

paid
20
15
10
5
no
yes

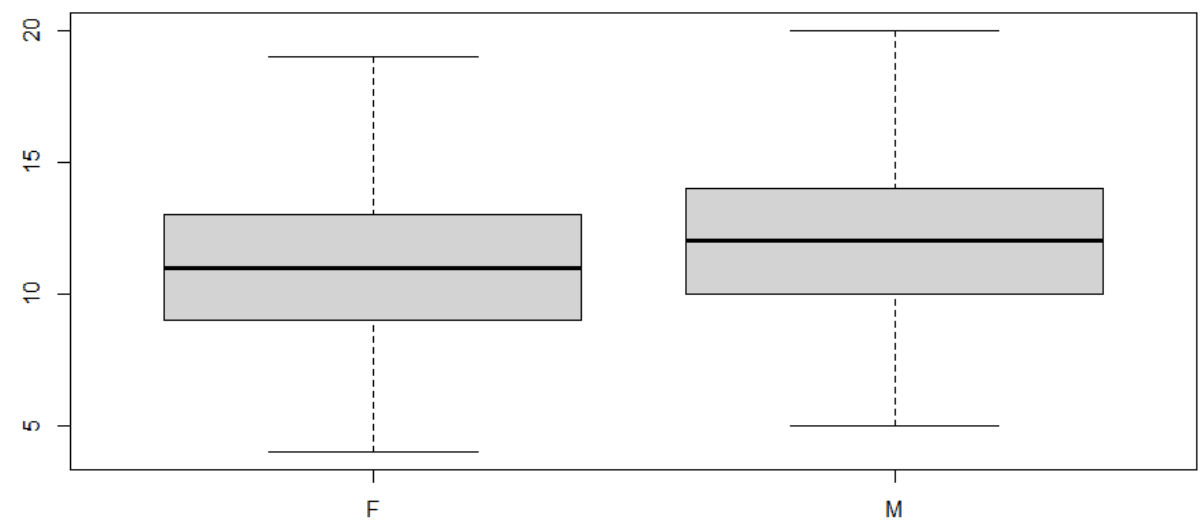

age
20
15
10
5
15
16
17
18
19
20
21
22

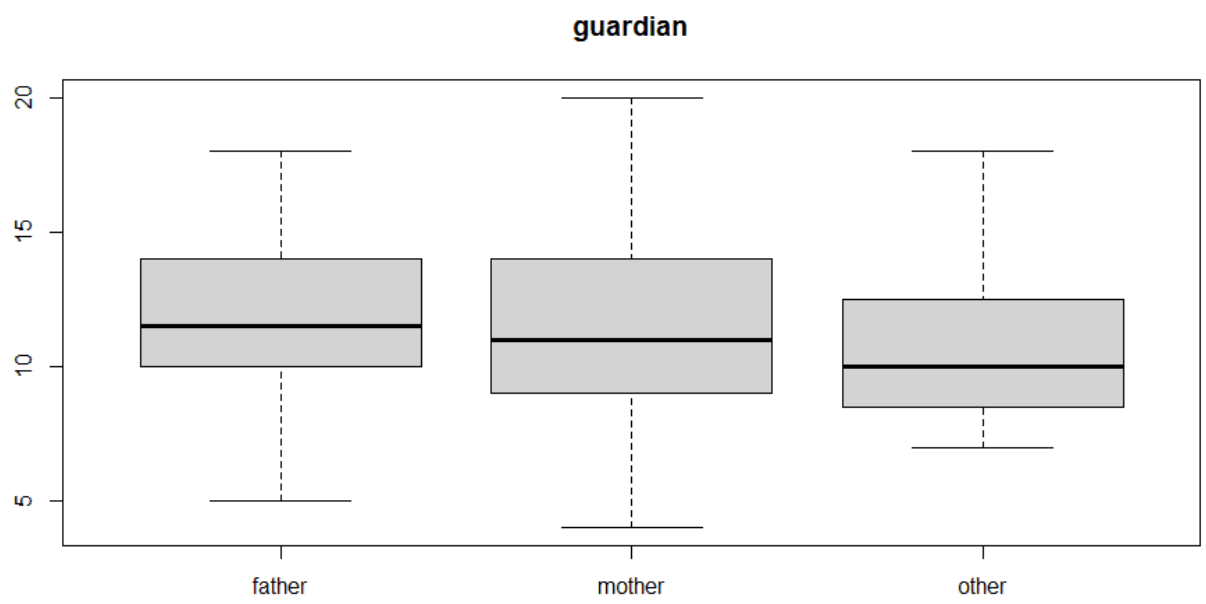

freetime
20
15
10
5
1
2
3
4
5

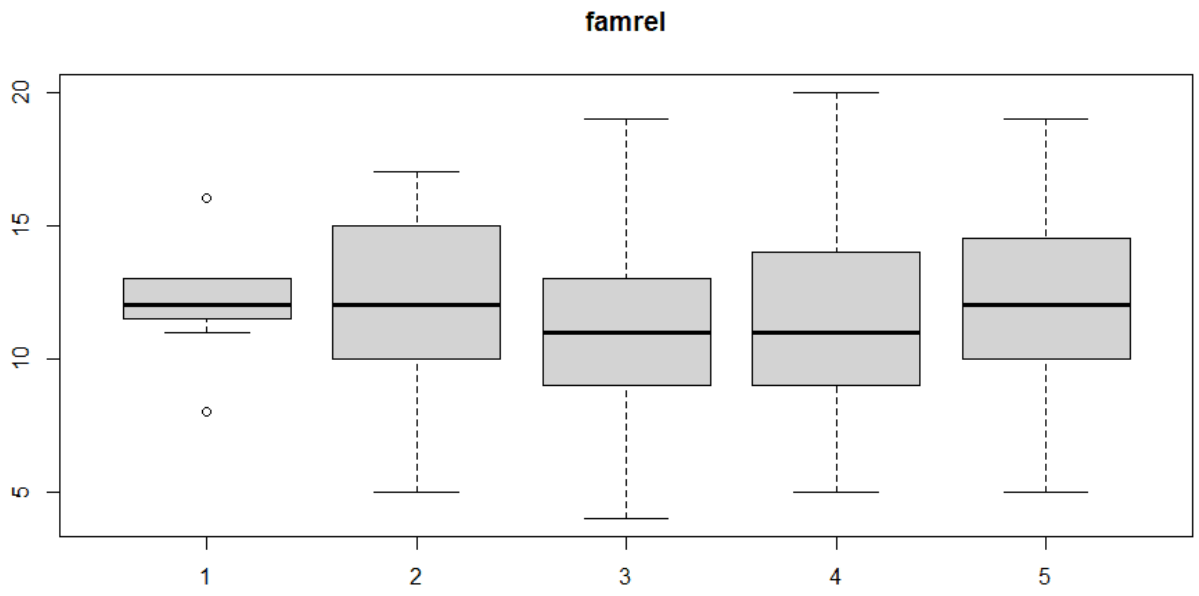

nursery
20
15
10
5
no
yes

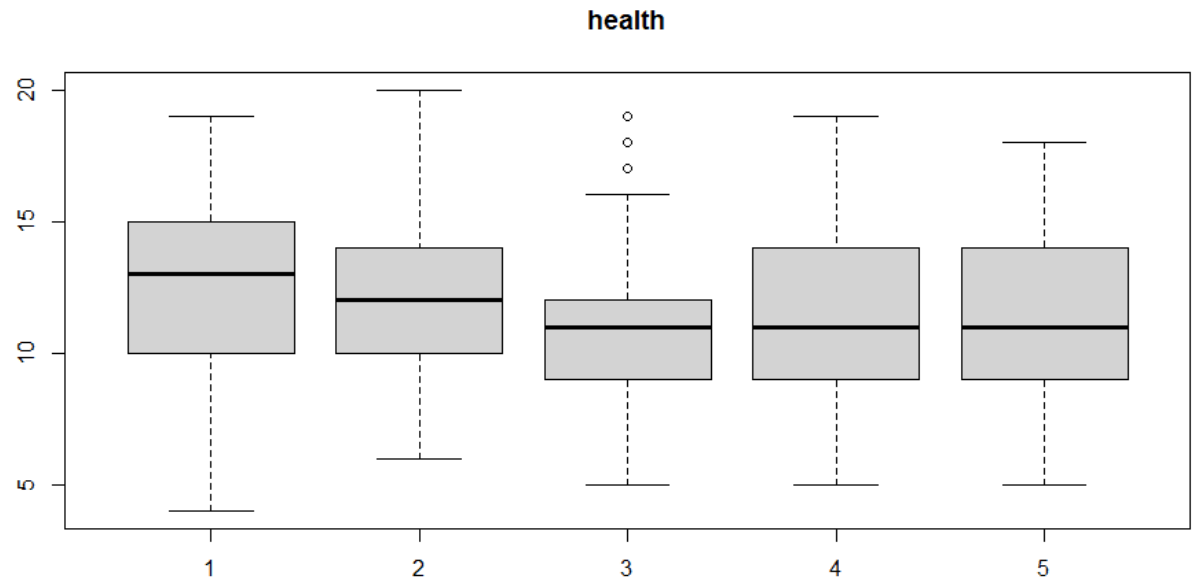

romantic
20
15
10
5
no
yes

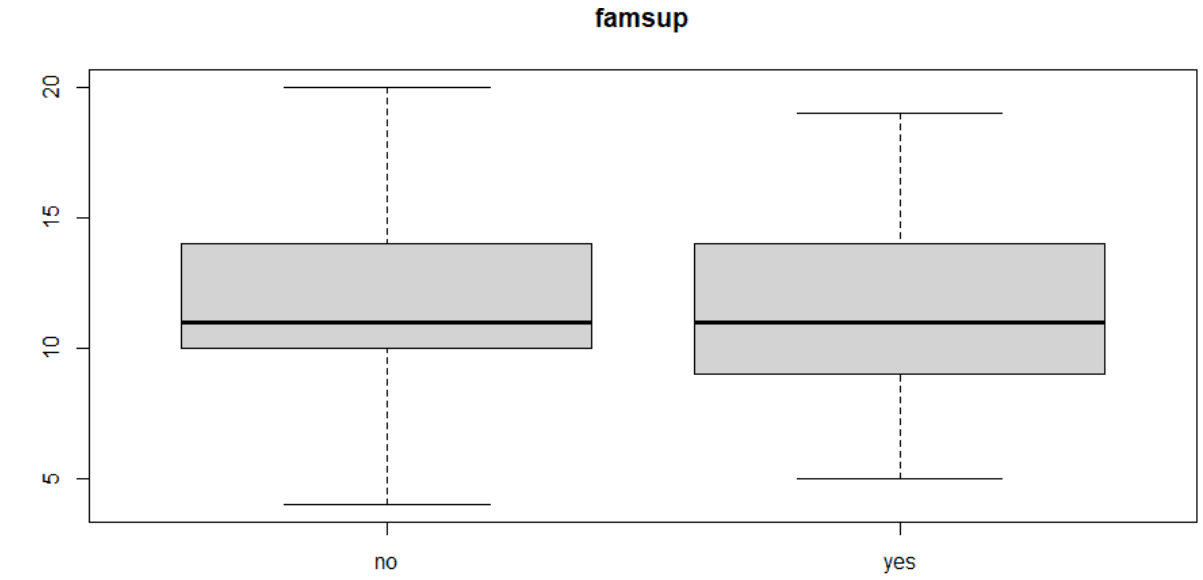

school
20
15
10
5
GP
MS

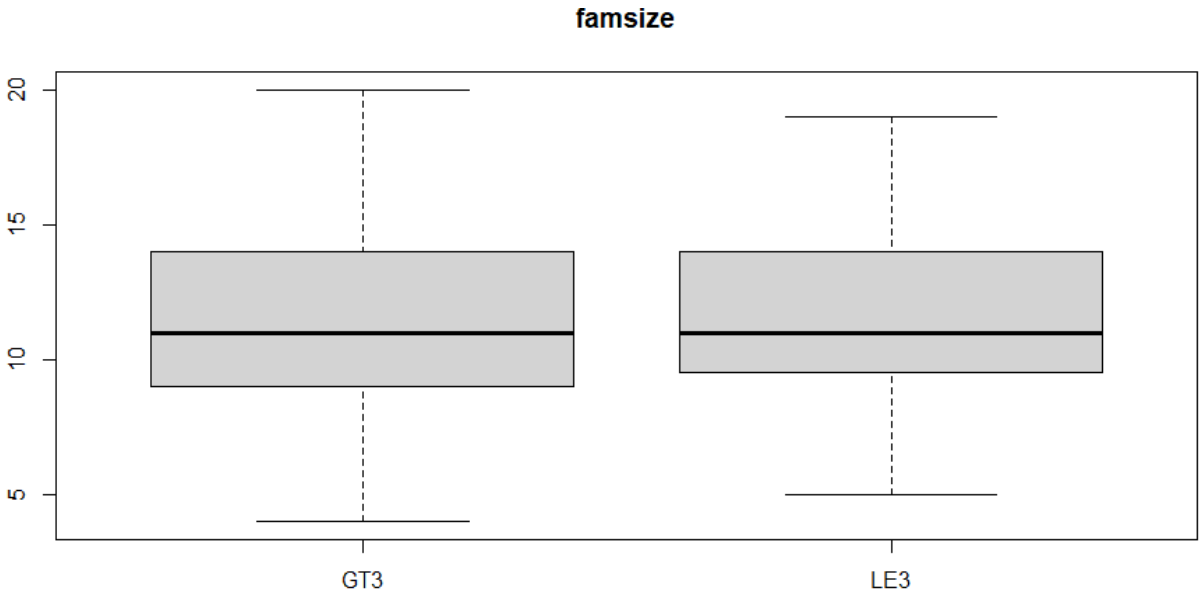

sex
20
15
10
5
F
M


guardian
20
15
10
5
father
mother
other


famrel
20
15
10
5
1
2
3
4
5


health
20
15
10
5
1
2
3
4
5


famsup
20
15
10
5
no
yes


famsize
20
15
10
5
GT3
LE3