Academic Performance Analysis: Interplay of Socioeconomic and Lifestyle Factors

– Nishant Khanorkar and Prathamesh Khole

Data Overview

- Collected in 2008, from two schools in Portugal.
- Used to understand and study high early school leaving rate in Portugal.
- Highest early school leaving rate in Europe (40% vs. 15%).
- Data covers various aspects of student lives:
  - Academics
  - Social life and status
  - Demographics
  - Background
  - Behavioral nuances
- Data (649 Instances of 30 features), and has 3 decision variables in the form of grades.

Data Recap

- School Identifier (SI):

○ school:	student's school	(binary:	Gabriel Pereira or Mousinho da Silveira)

- Demographic Profile (DP):

○ sex:	student's sex	(binary:	female or male)
○ address:	student's home address type	(binary:	urban or rural)

- Family Background (FB):

○ famsize:	family size	(binary:	≤ 3 or > 3)
○ Pstatus:	parent's cohabitation status	(binary:	living together or apart)
○ Medu:	mother'seducation	(numeric:	from 0 to 4[a])
○ Fedu:	father'seducation	(numeric:	from 0 to 4[a])

*a => (0 – none), 1 – (primary education 4th grade), 2 – (5th to 9th grade), 3 – (secondary education) or 4 – (higher education).

Data Recap

- Educational Factors (EF):

Mjob:	mother’s job	(nominal[b])
Fjob:	father’s job	(nominal[b])
reason:	reason to choose this school	(nominal: close to home, school reputation, course preference or other)
traveltime:	home to school travel time	(numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour)
studytime:	weekly study time	(numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – >10 hours)

- Personal Behaviors (PB):

activities:	extra-curricular activities	(binary: yes or no)
nursery:	attended nursery school	(binary: yes or no)
higher:	wants to take higher education	(binary: yes or no)
internet:	Internet access at home	(binary: yes or no)
romantic:	with a romantic relationship	(binary: yes or no)

*b => teacher, healthcare related, civil services (e.g. administrative or police), at home or other.

Data Recap

- Health and Lifestyle (HL):

Health:	current health status	(numeric: from 1 – very bad to 5 – very good)
Walc:	weekend alcohol consumption	(numeric: from 1 – very low to 5 – very high)
Dalc:	workday alcohol consumption	(numeric: from 1 – very low to 5 – very high)
absences:	number of school absences	(numeric: from 0 to 93)

- Academic Grades (AG):

G1:	first period grade	(numeric: from 0 to 20)
G2:	second period grade	(numeric: from 0 to 20)
G3:	final grade	(numeric: from 0 to 20)

Objective

- Our primary goal is to investigate how socioeconomic factors and lifestyle choices collectively impact academic outcomes in the Portuguese secondary educational context.
- To gain a deeper understanding of the above phenomenon.
- We have grouped our analysis into three broad categories:

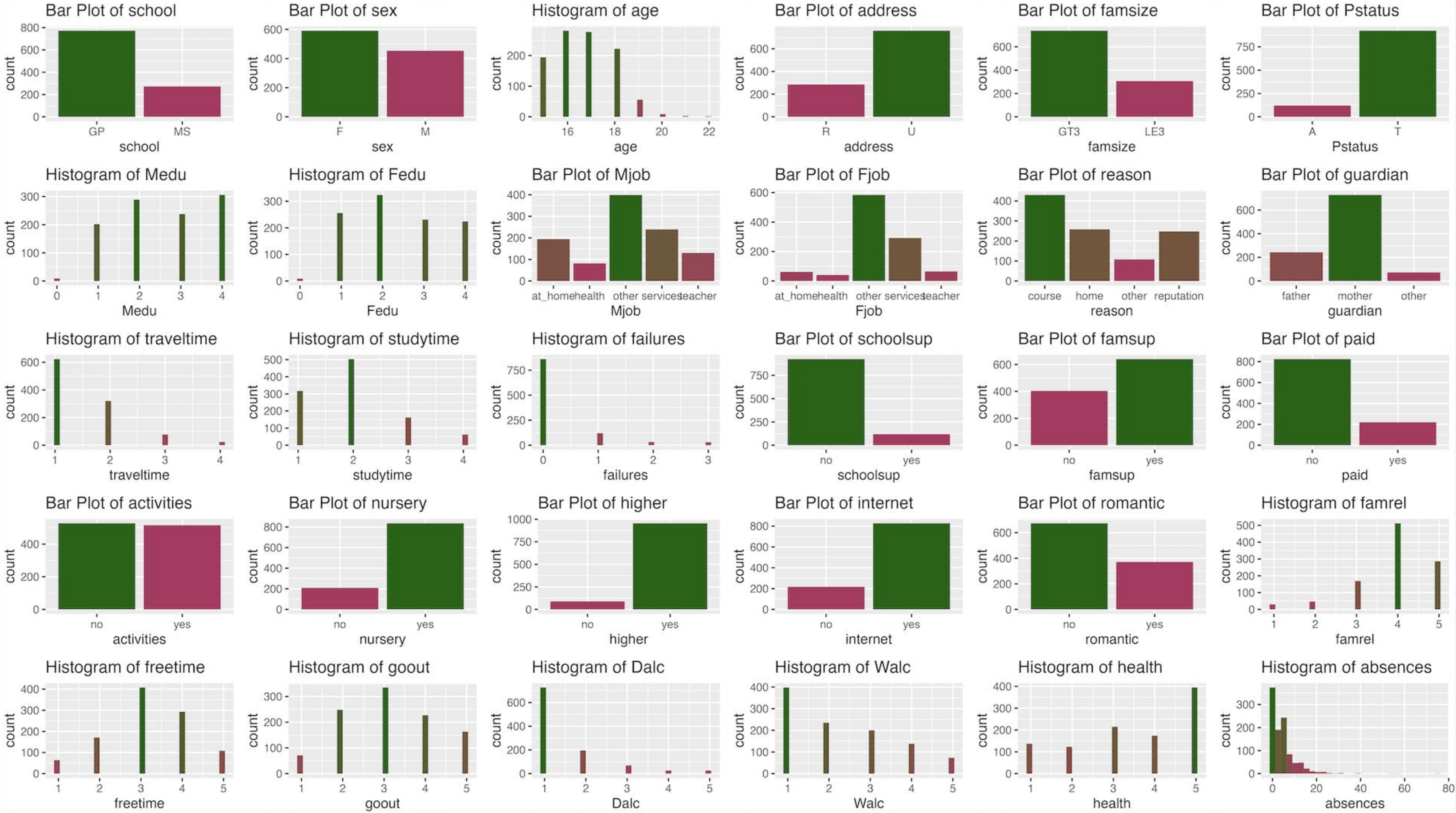
  I. Study the statistical relationships between students' grades and a series of independent variables, including socioeconomic factors, study habits, and personal lifestyle choices.

  II. Evaluate the relative impact of various lifestyle choices and demographic factors on students' academic performances.

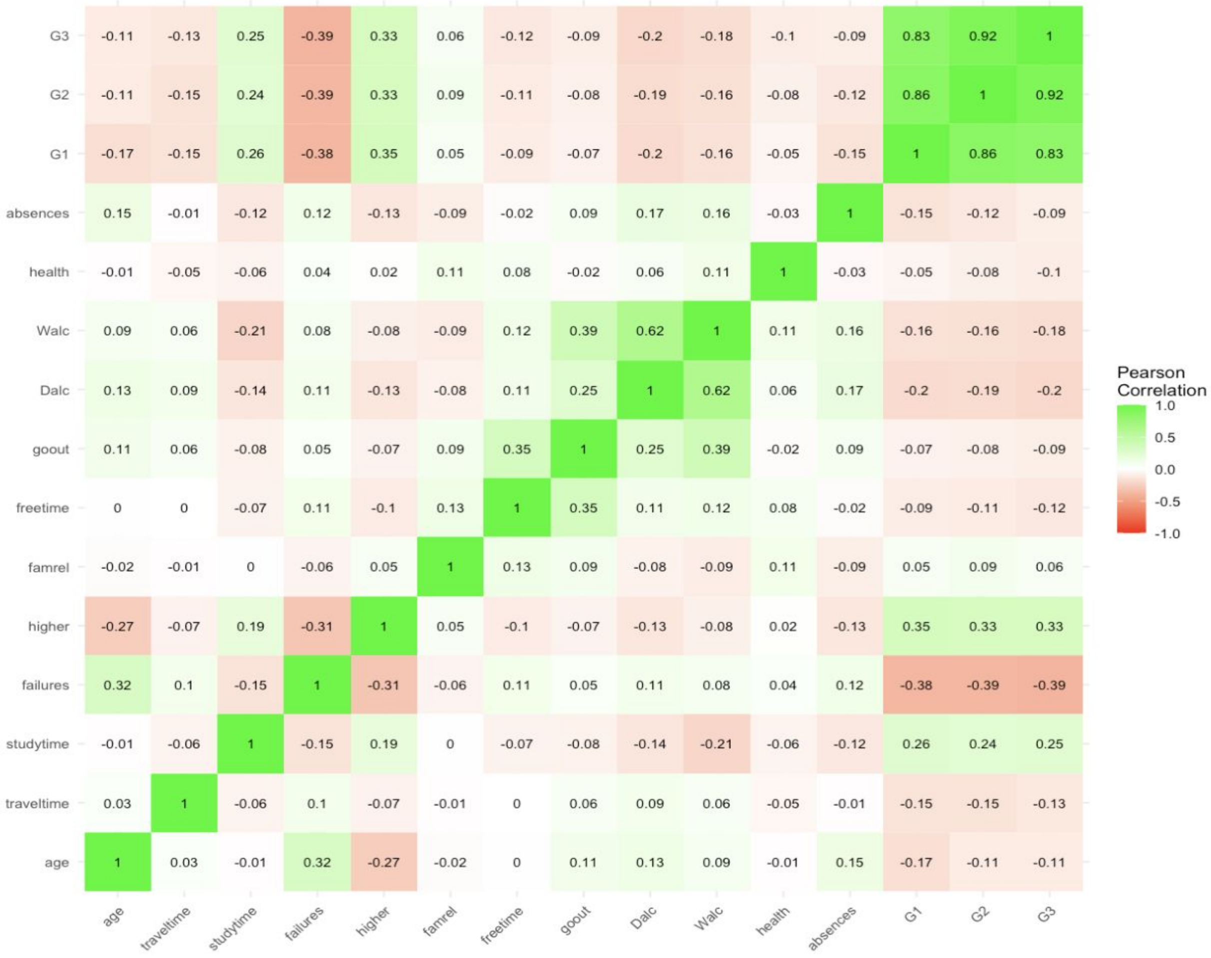
  III. Debunk popular hypotheses for school students.

Exploratory Data Analysis

- Creating or visualizing the variables apart from G1, G2, and G3 to understand the general trends and distribution of data, using histogram and bar plots.
- Creating multiple correlation heat maps to visualize possible data correlations and relationships.
- Selecting highly correlated features to remove from final model.
- Final grade trends based on our observations so far.
- Understanding relationship between final grade and the lifestyle choices made by students.

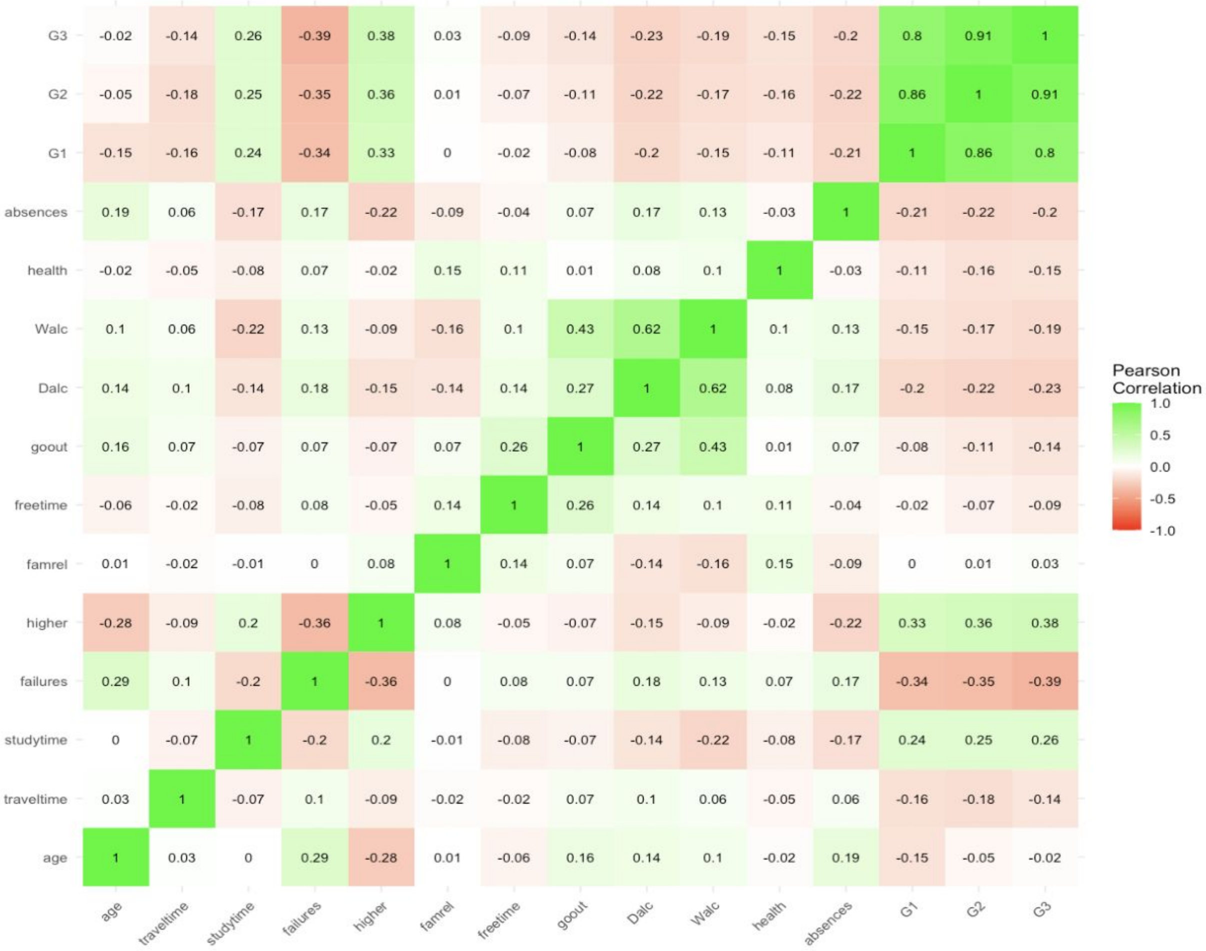

Bar Plot of school
count
0 200 400 600 800
GP MS
school
Bar Plot of sex
count
0 200 400 600
F M
sex
Histogram of age
count
0 100 200
16 18 20 22
age
Bar Plot of address
count
0 200 400 600
R U
address
Bar Plot of famsize
count
0 200 400 600
GT3 LE3
famsize
Bar Plot of Pstatus
count
0 250 500 750
A T
Pstatus
Histogram of Medu
count
0 100 200 300
0 1 2 3 4
Medu
Histogram of Fedu
count
0 100 200 300
0 1 2 3 4
Fedu
Bar Plot of Mjob
count
0 100 200 300 400
at_homehealth other servicesteacher
Mjob
Bar Plot of Fjob
count
0 200 400 600
at_homehealth other servicesteacher
Fjob
Bar Plot of reason
count
0 100 200 300 400
course home other reputation
reason
Bar Plot of guardian
count
0 200 400 600
father mother other
guardian
Histogram of traveltime
count
0 200 400 600
1 2 3 4
traveltime
Histogram of studytime
count
0 100 200 300 400 500
1 2 3 4
studytime
Histogram of failures
count
0 250 500 750
0 1 2 3
failures
Bar Plot of schoolsup
count
0 250 500 750
no yes
schoolsup
Bar Plot of famsup
count
0 200 400 600
no yes
famsup
Bar Plot of paid
count
0 200 400 600 800
no yes
paid
Bar Plot of activities
count
0 200 400
no yes
activities
Bar Plot of nursery
count
0 200 400 600 800
no yes
nursery
Bar Plot of higher
count
0 250 500 750 1000
no yes
higher
Bar Plot of internet
count
0 200 400 600 800
no yes
internet
Bar Plot of romantic
count
0 200 400 600
no yes
romantic
Histogram of famrel
count
0 100 200 300 400 500
1 2 3 4 5
famrel
Histogram of freetime
count
0 100 200 300 400
1 2 3 4 5
freetime
Histogram of goout
count
0 100 200 300
1 2 3 4 5
goout
Histogram of Dalc
count
0 200 400 600
1 2 3 4 5
Dalc
Histogram of Walc
count
0 100 200 300 400
1 2 3 4 5
Walc
Histogram of health
count
0 100 200 300 400
1 2 3 4 5
health
Histogram of absences
count
0 100 200 300
0 20 40 60 80
absences

Correlation Matrix



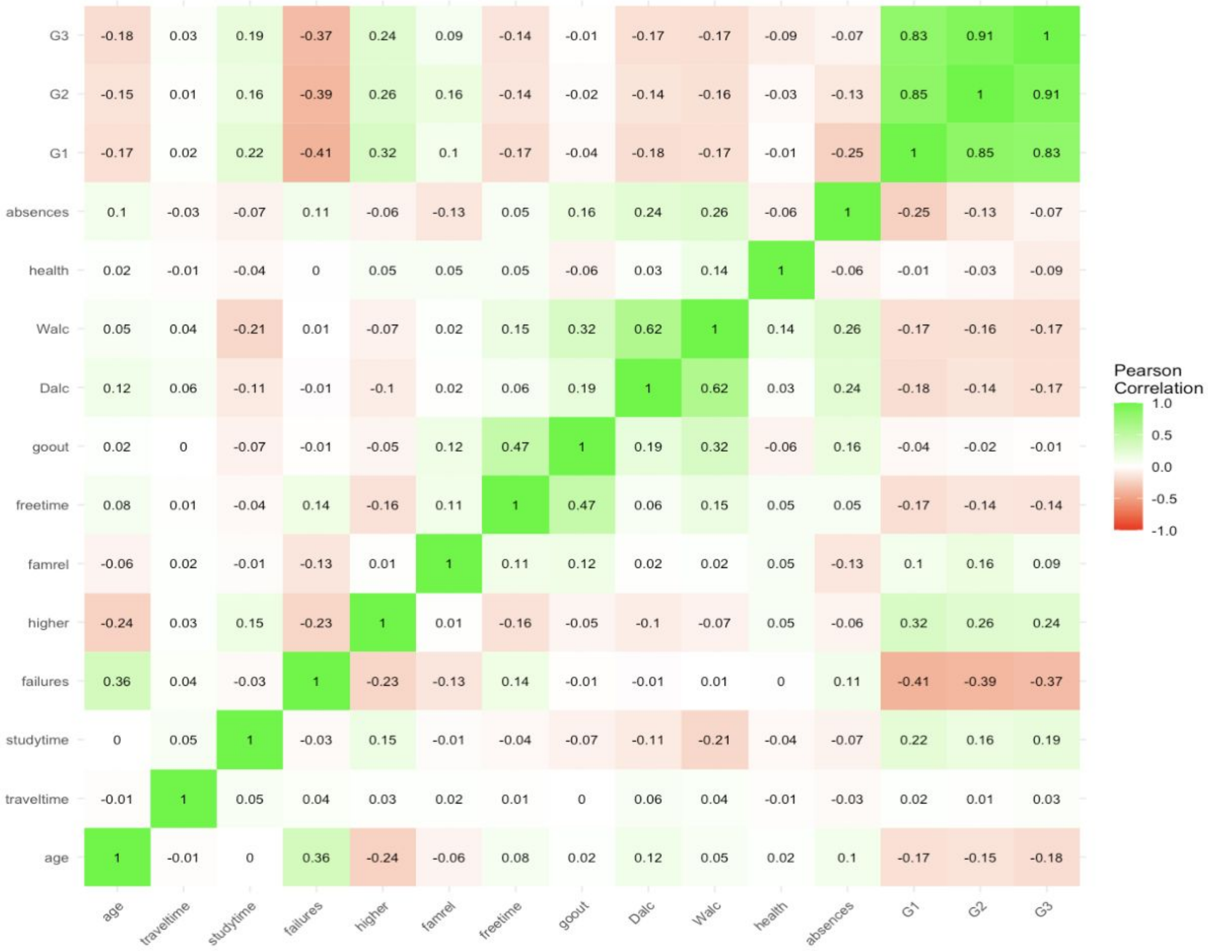
	age	traveltime	studytime	failures	higher	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
G3	-0.11	-0.13	0.25	-0.39	0.33	0.06	-0.12	-0.09	-0.2	-0.18	-0.1	-0.09	0.83	0.92	1
G2	-0.11	-0.15	0.24	-0.39	0.33	0.09	-0.11	-0.08	-0.19	-0.16	-0.08	-0.12	0.86	1	0.92
G1	-0.17	-0.15	0.26	-0.38	0.35	0.05	-0.09	-0.07	-0.2	-0.16	-0.05	-0.15	1	0.86	0.83
absences	0.15	-0.01	-0.12	0.12	-0.13	-0.09	-0.02	0.09	0.17	0.16	-0.03	1	-0.15	-0.12	-0.09
health	-0.01	-0.05	-0.06	0.04	0.02	0.11	0.08	-0.02	0.06	0.11	1	-0.03	-0.05	-0.08	-0.1
Walc	0.09	0.06	-0.21	0.08	-0.08	-0.09	0.12	0.39	0.62	1	0.11	0.16	-0.16	-0.16	-0.18
Dalc	0.13	0.09	-0.14	0.11	-0.13	-0.08	0.11	0.25	1	0.62	0.06	0.17	-0.2	-0.19	-0.2
goout	0.11	0.06	-0.08	0.05	-0.07	0.09	0.35	1	0.25	0.39	-0.02	0.09	-0.07	-0.08	-0.09
freetime	0	0	-0.07	0.11	-0.1	0.13	1	0.35	0.11	0.12	0.08	-0.02	-0.09	-0.11	-0.12
famrel	-0.02	-0.01	0	-0.06	0.05	1	0.13	0.09	-0.08	-0.09	0.11	-0.09	0.05	0.09	0.06
higher	-0.27	-0.07	0.19	-0.31	1	0.05	-0.1	-0.07	-0.13	-0.08	0.02	-0.13	0.35	0.33	0.33
failures	0.32	0.1	-0.15	1	-0.31	-0.06	0.11	0.05	0.11	0.08	0.04	0.12	-0.38	-0.39	-0.39
studytime	-0.01	-0.06	1	-0.15	0.19	0	-0.07	-0.08	-0.14	-0.21	-0.06	-0.12	0.26	0.24	0.25
traveltime	0.03	1	-0.06	0.1	-0.07	-0.01	0	0.06	0.09	0.06	-0.05	-0.01	-0.15	-0.15	-0.13
age	1	0.03	-0.01	0.32	-0.27	-0.02	0	0.11	0.13	0.09	-0.01	0.15	-0.17	-0.11	-0.11

Correlation Matrix for School ‘GP’

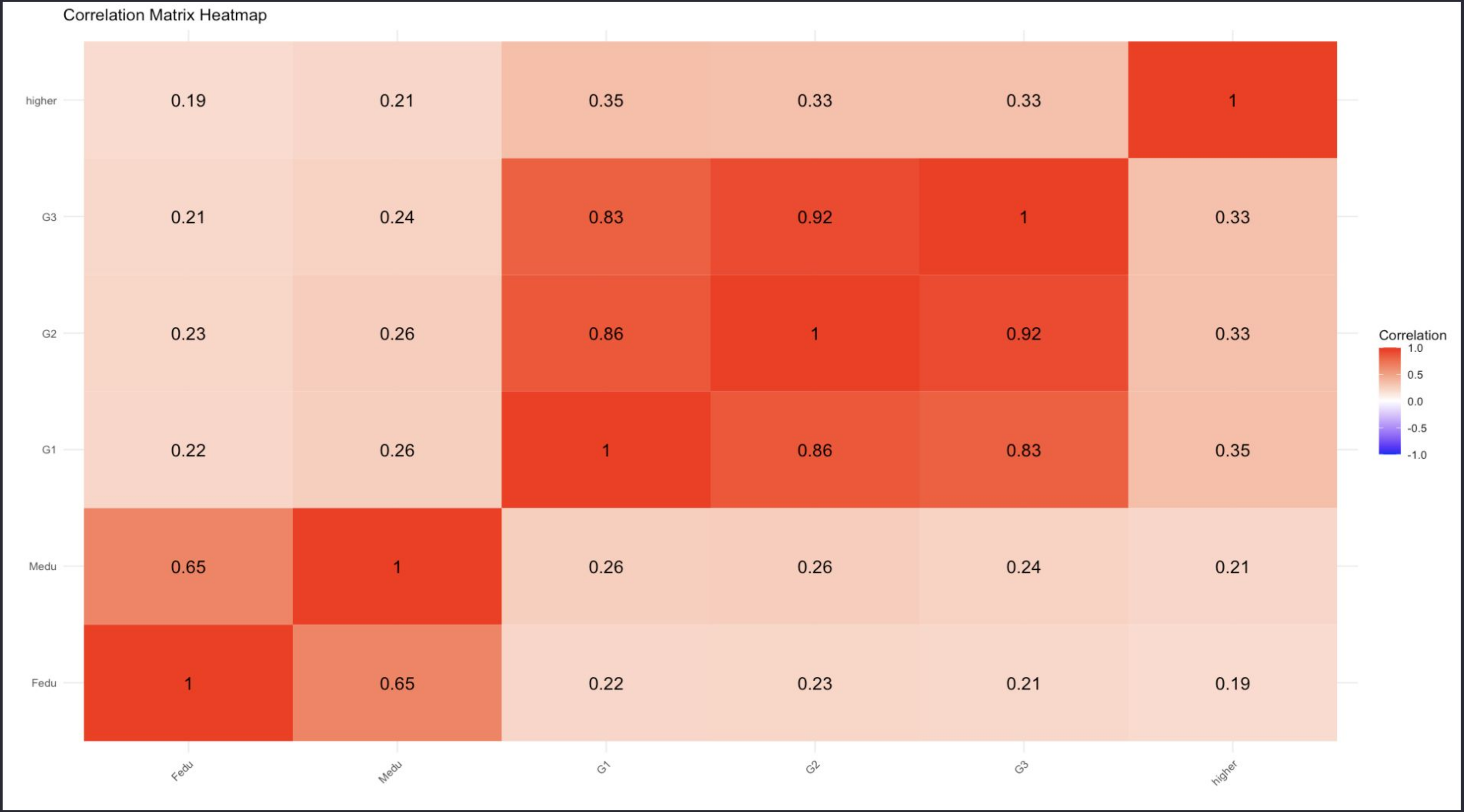


	age	traveltime	studytime	failures	higher	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
G3	-0.02	-0.14	0.26	-0.39	0.38	0.03	-0.09	-0.14	-0.23	-0.19	-0.15	-0.2	0.8	0.91	1
G2	-0.05	-0.18	0.25	-0.35	0.36	0.01	-0.07	-0.11	-0.22	-0.17	-0.16	-0.22	0.86	1	0.91
G1	-0.15	-0.16	0.24	-0.34	0.33	0	-0.02	-0.08	-0.2	-0.15	-0.11	-0.21	1	0.86	0.8
absences	0.19	0.06	-0.17	0.17	-0.22	-0.09	-0.04	0.07	0.17	0.13	-0.03	1	-0.21	-0.22	-0.2
health	-0.02	-0.05	-0.08	0.07	-0.02	0.15	0.11	0.01	0.08	0.1	1	-0.03	-0.11	-0.16	-0.15
Walc	0.1	0.06	-0.22	0.13	-0.09	-0.16	0.1	0.43	0.62	1	0.1	0.13	-0.15	-0.17	-0.19
Dalc	0.14	0.1	-0.14	0.18	-0.15	-0.14	0.14	0.27	1	0.62	0.08	0.17	-0.2	-0.22	-0.23
goout	0.16	0.07	-0.07	0.07	-0.07	0.07	0.26	1	0.27	0.43	0.01	0.07	-0.08	-0.11	-0.14
freetime	-0.06	-0.02	-0.08	0.08	-0.05	0.14	1	0.26	0.14	0.1	0.11	-0.04	-0.02	-0.07	-0.09
famrel	0.01	-0.02	-0.01	0	0.08	1	0.14	0.07	-0.14	-0.16	0.15	-0.09	0	0.01	0.03
higher	-0.28	-0.09	0.2	-0.36	1	0.08	-0.05	-0.07	-0.15	-0.09	-0.02	-0.22	0.33	0.36	0.38
failures	0.29	0.1	-0.2	1	-0.36	0	0.08	0.07	0.18	0.13	0.07	0.17	-0.34	-0.35	-0.39
studytime	0	-0.07	1	-0.2	0.2	-0.01	-0.08	-0.07	-0.14	-0.22	-0.08	-0.17	0.24	0.25	0.26
traveltime	0.03	1	-0.07	0.1	-0.09	-0.02	-0.02	0.07	0.1	0.06	-0.05	0.06	-0.16	-0.18	-0.14
age	1	0.03	0	0.29	-0.28	0.01	-0.06	0.16	0.14	0.1	-0.02	0.19	-0.15	-0.05	-0.02

Correlation Matrix for School ‘MS’

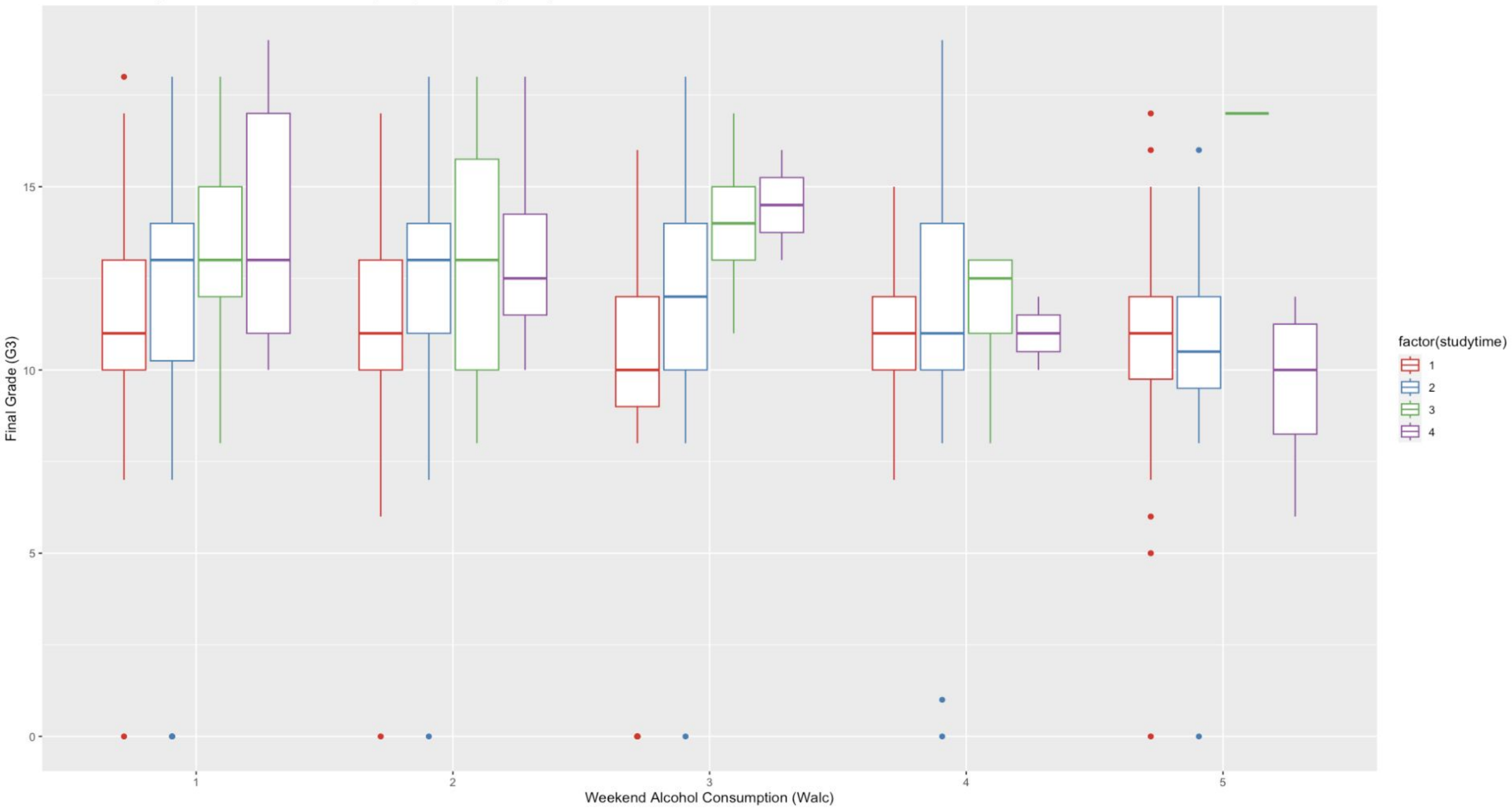


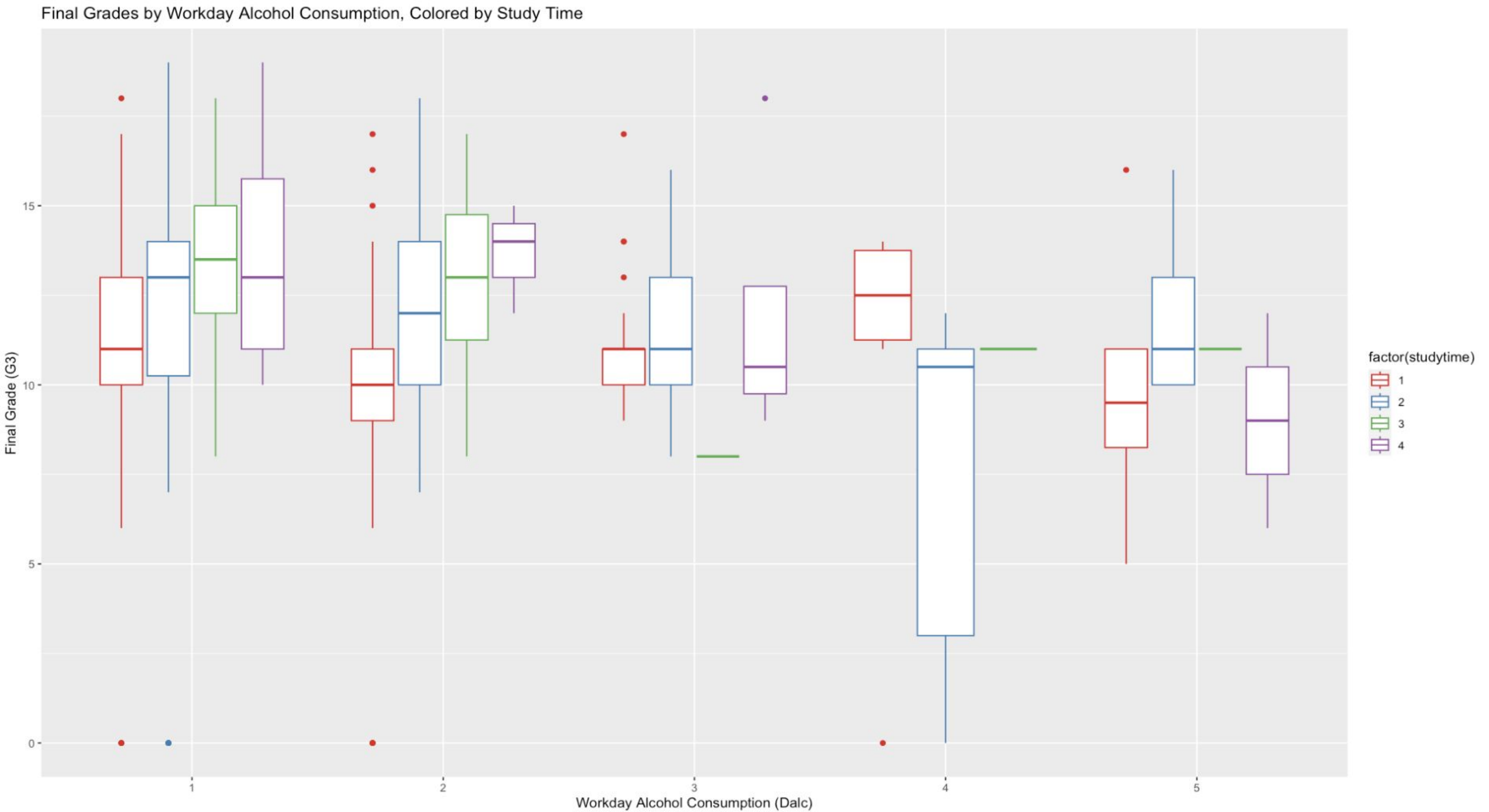
	age	traveltime	studytime	failures	higher	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
G3	-0.18	0.03	0.19	-0.37	0.24	0.09	-0.14	-0.01	-0.17	-0.17	-0.09	-0.07	0.83	0.91	1
G2	-0.15	0.01	0.16	-0.39	0.26	0.16	-0.14	-0.02	-0.14	-0.16	-0.03	-0.13	0.85	1	0.91
G1	-0.17	0.02	0.22	-0.41	0.32	0.1	-0.17	-0.04	-0.18	-0.17	-0.01	-0.25	1	0.85	0.83
absences	0.1	-0.03	-0.07	0.11	-0.06	-0.13	0.05	0.16	0.24	0.26	-0.06	1	-0.25	-0.13	-0.07
health	0.02	-0.01	-0.04	0	0.05	0.05	0.05	-0.06	0.03	0.14	1	-0.06	-0.01	-0.03	-0.09
Walc	0.05	0.04	-0.21	0.01	-0.07	0.02	0.15	0.32	0.62	1	0.14	0.26	-0.17	-0.16	-0.17
Dalc	0.12	0.06	-0.11	-0.01	-0.1	0.02	0.06	0.19	1	0.62	0.03	0.24	-0.18	-0.14	-0.17
goout	0.02	0	-0.07	-0.01	-0.05	0.12	0.47	1	0.19	0.32	-0.06	0.16	-0.04	-0.02	-0.01
freetime	0.08	0.01	-0.04	0.14	-0.16	0.11	1	0.47	0.06	0.15	0.05	0.05	-0.17	-0.14	-0.14
famrel	-0.06	0.02	-0.01	-0.13	0.01	1	0.11	0.12	0.02	0.02	0.05	-0.13	0.1	0.16	0.09
higher	-0.24	0.03	0.15	-0.23	1	0.01	-0.16	-0.05	-0.1	-0.07	0.05	-0.06	0.32	0.26	0.24
failures	0.36	0.04	-0.03	1	-0.23	-0.13	0.14	-0.01	-0.01	0.01	0	0.11	-0.41	-0.39	-0.37
studytime	0	0.05	1	-0.03	0.15	-0.01	-0.04	-0.07	-0.11	-0.21	-0.04	-0.07	0.22	0.16	0.19
traveltime	-0.01	1	0.05	0.04	0.03	0.02	0.01	0	0.06	0.04	-0.01	-0.03	0.02	0.01	0.03
age	1	-0.01	0	0.36	-0.24	-0.06	0.08	0.02	0.12	0.05	0.02	0.1	-0.17	-0.15	-0.18



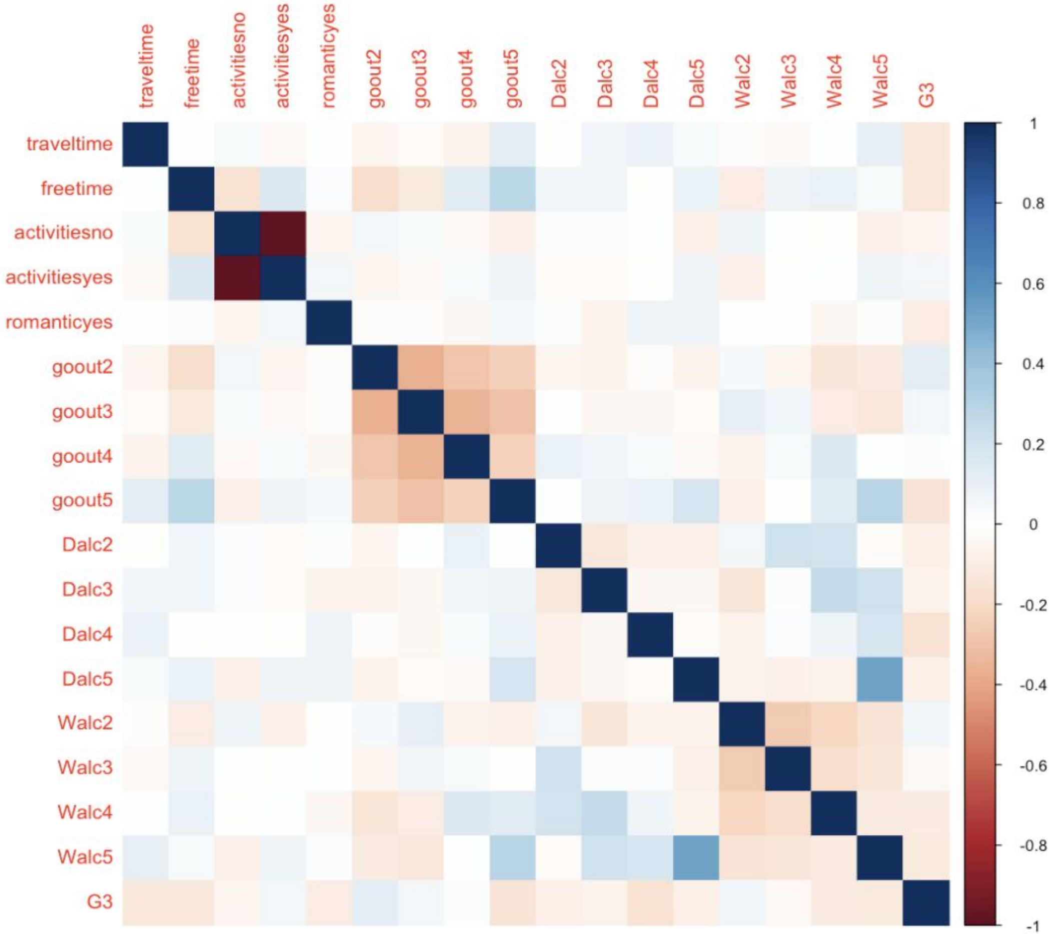
	Fedu	Medu	G1	G2	G3	higher
higher	0.19	0.21	0.35	0.33	0.33	1
G3	0.21	0.24	0.83	0.92	1	0.33
G2	0.23	0.26	0.86	1	0.92	0.33
G1	0.22	0.26	1	0.86	0.83	0.35
Medu	0.65	1	0.26	0.26	0.24	0.21
Fedu	1	0.65	0.22	0.23	0.21	0.19

Final Grades by Weekend Alcohol Consumption, Colored by Study Time




Final Grades by Workday Alcohol Consumption, Colored by Study Time
Final Grade (G3)
15
10
5
0
1
2
3
4
5
Workday Alcohol Consumption (Dalc)
factor(studytime)
1
2
3
4

Correlation Matrix for Lifestyle Choices



Observed Trends and Initial Hypothesis

- Higher workday alcohol consumption leads to lower grades.
- Romantic Interests have a significant influence on Grades.
- Students with higher grades aspire to continue for higher studies.
- Taking paid classes can improve grades.
- Influence of going out more on grades.
- Do absences negatively affect grades.
- Relationship between parents jobs and student grades.
- Impact having activities on grades.
- Relation between grades and free time, and study time.

Fitted Models

- Multiple Linear Regression (MLR)

$$G3 = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n + \epsilon$$

where G3 is the final grade, β_0 is the intercept, $\beta_1,\ldots,\beta_n$ are the coefficients for each predictor variable $X_1,\ldots,X_n$, and ε is the error term. The model assumes that the relationship between the dependent variable and the predictors is linear.

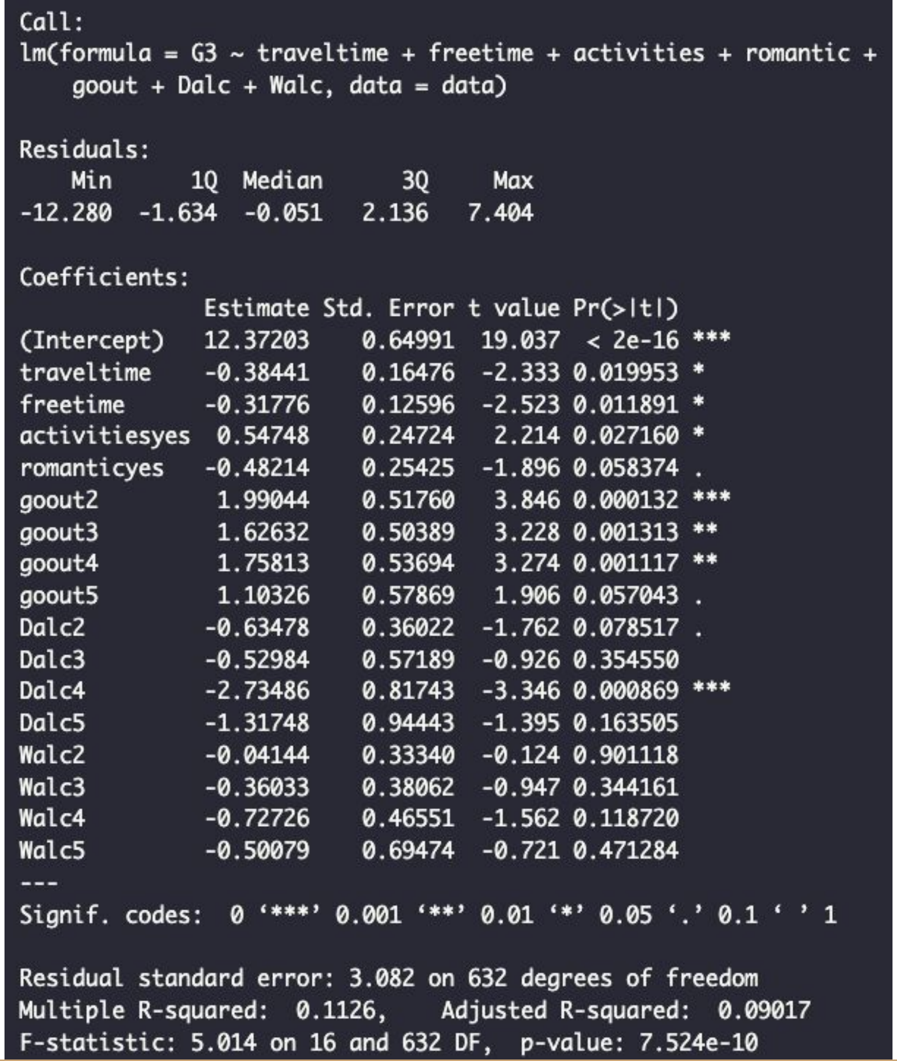
- Logistic Regression (LR)

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n$$

where p is the probability of the outcome (e.g., passing the final grade), p/(1-p) is the odds ratio, and β_0 ,β_1 ,...,β_n are the model coefficients corresponding to each predictor $X_1,\ldots,X_n$.

MLR Model-1

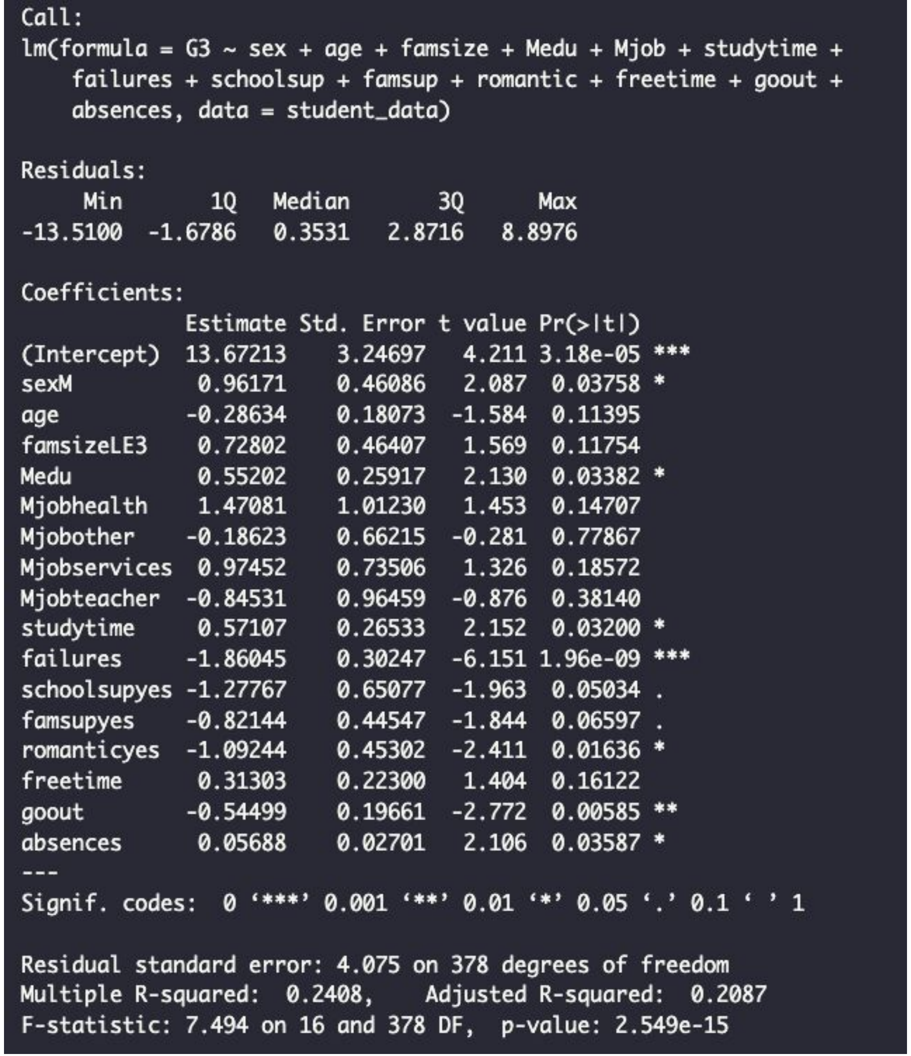
Final grade based on student’s lifestyle factors.



```
Call:
lm(formula = G3 ~ traveltime + freetime + activities + romantic + 
    goout + Dalc + Walc, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.280  -1.634  -0.051   2.136   7.404 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   12.37203    0.64991  19.037  < 2e-16 ***
traveltime    -0.38441    0.16476  -2.333 0.019953 *  
freetime      -0.31776    0.12596  -2.523 0.011891 *  
activitiesyes  0.54748    0.24724   2.214 0.027160 *  
romanticyes   -0.48214    0.25425  -1.896 0.058374 .  
goout2         1.99044    0.51760   3.846 0.000132 ***
goout3         1.62632    0.50389   3.228 0.001313 ** 
goout4         1.75813    0.53694   3.274 0.001117 ** 
goout5         1.10326    0.57869   1.906 0.057043 .  
Dalc2         -0.63478    0.36022  -1.762 0.078517 .  
Dalc3         -0.52984    0.57189  -0.926 0.354550    
Dalc4         -2.73486    0.81743  -3.346 0.000869 ***
Dalc5         -1.31748    0.94443  -1.395 0.163505    
Walc2         -0.04144    0.33340  -0.124 0.901118    
Walc3         -0.36033    0.38062  -0.947 0.344161    
Walc4         -0.72726    0.46551  -1.562 0.118720    
Walc5         -0.50079    0.69474  -0.721 0.471284    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.082 on 632 degrees of freedom
Multiple R-squared:  0.1126,	Adjusted R-squared:  0.09017 
F-statistic: 5.014 on 16 and 632 DF,  p-value: 7.524e-10
```

MLR Model-2

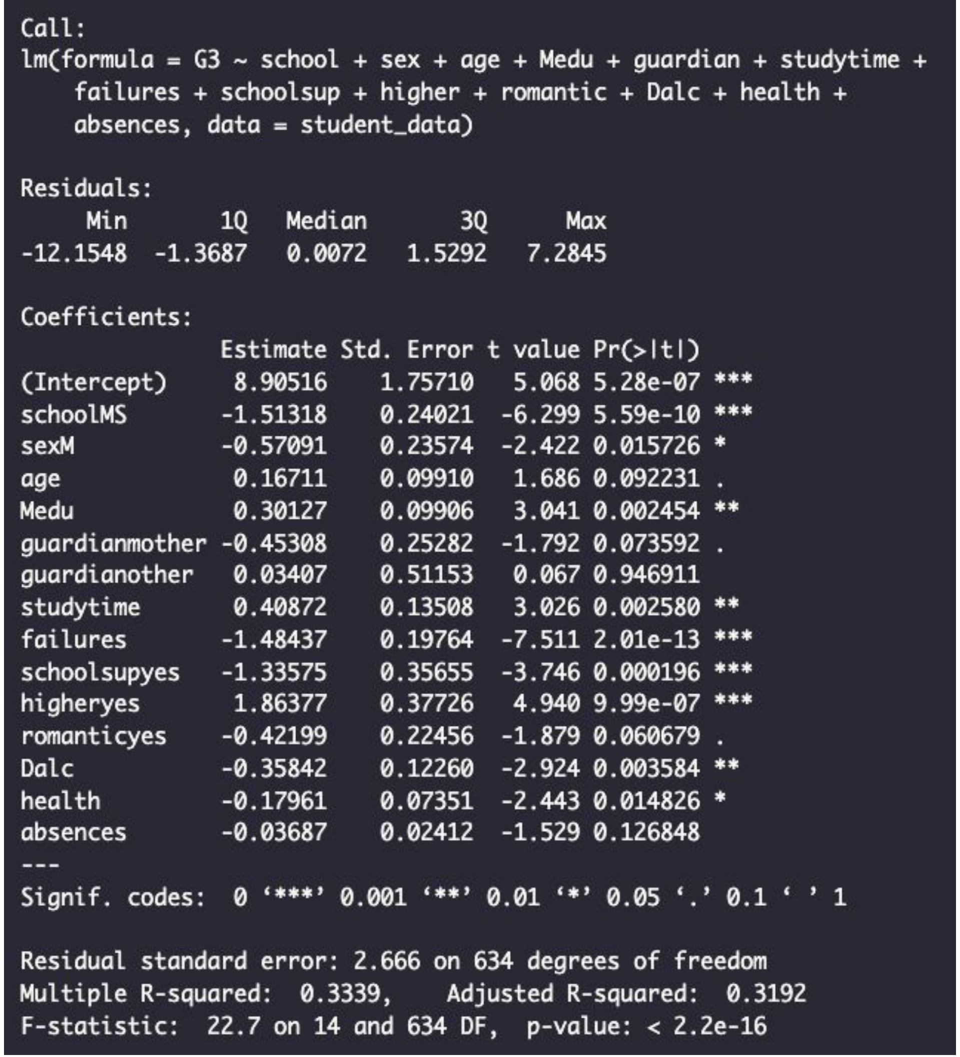
Final Maths course grade based on model created by stepwise feature selection.



```
Call:
lm(formula = G3 ~ sex + age + famsize + Medu + Mjob + studytime +
    failures + schoolsup + famsup + romantic + freetime + goout +
    absences, data = student_data)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5100  -1.6786   0.3531   2.8716   8.8976

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.67213    3.24697   4.211 3.18e-05 ***
sexM         0.96171    0.46086   2.087  0.03758 *
age         -0.28634    0.18073  -1.584  0.11395
famsizeLE3   0.72802    0.46407   1.569  0.11754
Medu         0.55202    0.25917   2.130  0.03382 *
Mjobhealth   1.47081    1.01230   1.453  0.14707
Mjobother   -0.18623    0.66215  -0.281  0.77867
Mjobservices  0.97452    0.73506   1.326  0.18572
Mjobteacher -0.84531    0.96459  -0.876  0.38140
studytime    0.57107    0.26533   2.152  0.03200 *
failures    -1.86045    0.30247  -6.151 1.96e-09 ***
schoolsupyes -1.27767    0.65077  -1.963  0.05034 .
famsupyes   -0.82144    0.44547  -1.844  0.06597 .
romanticyes -1.09244    0.45302  -2.411  0.01636 *
freetime     0.31303    0.22300   1.404  0.16122
goout       -0.54499    0.19661  -2.772  0.00585 **
absences     0.05688    0.02701   2.106  0.03587 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.075 on 378 degrees of freedom
Multiple R-squared:  0.2408,    Adjusted R-squared:  0.2087
F-statistic: 7.494 on 16 and 378 DF,  p-value: 2.549e-15
```

MLR Model-3

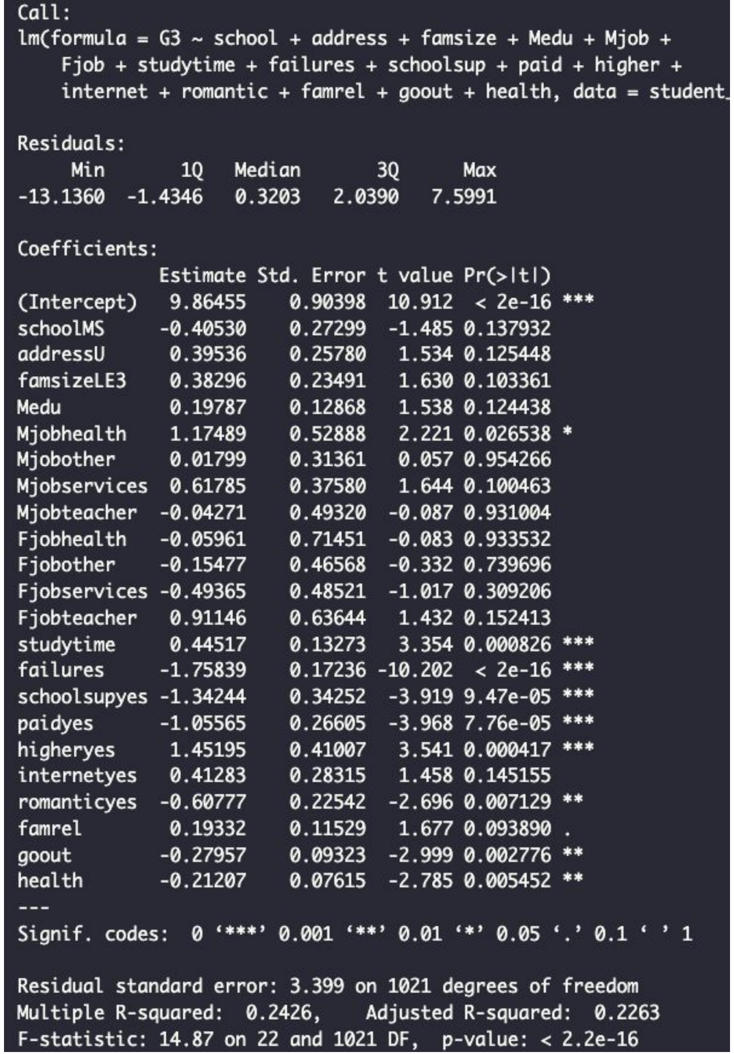
Final Portuguese course grade based on model created by stepwise feature selection.



```
Call:
lm(formula = G3 ~ school + sex + age + Medu + guardian + studytime +
    failures + schoolsup + higher + romantic + Dalc + health +
    absences, data = student_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.1548  -1.3687   0.0072   1.5292   7.2845 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     8.90516    1.75710   5.068 5.28e-07 ***
schoolMS       -1.51318    0.24021  -6.299 5.59e-10 ***
sexM           -0.57091    0.23574  -2.422 0.015726 *  
age             0.16711    0.09910   1.686 0.092231 .  
Medu            0.30127    0.09906   3.041 0.002454 ** 
guardianmother -0.45308    0.25282  -1.792 0.073592 .  
guardianother   0.03407    0.51153   0.067 0.946911    
studytime       0.40872    0.13508   3.026 0.002580 ** 
failures       -1.48437    0.19764  -7.511 2.01e-13 ***
schoolsupyes   -1.33575    0.35655  -3.746 0.000196 ***
higheryes       1.86377    0.37726   4.940 9.99e-07 ***
romanticyes    -0.42199    0.22456  -1.879 0.060679 .  
Dalc           -0.35842    0.12260  -2.924 0.003584 ** 
health         -0.17961    0.07351  -2.443 0.014826 *  
absences       -0.03687    0.02412  -1.529 0.126848    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.666 on 634 degrees of freedom
Multiple R-squared:  0.3339,	Adjusted R-squared:  0.3192 
F-statistic:  22.7 on 14 and 634 DF,  p-value: < 2.2e-16
```

MLR Model-4

Final grade regardless of course based on model created by stepwise feature selection.



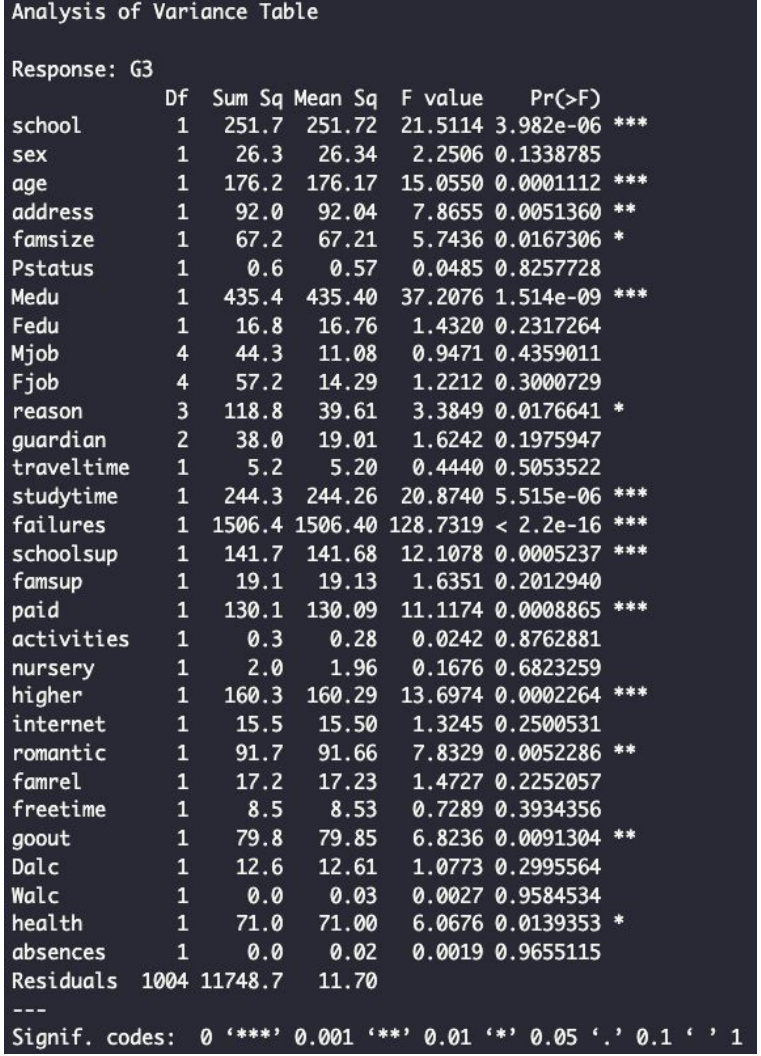
```
Call:
lm(formula = G3 ~ school + address + famsize + Medu + Mjob +
    Fjob + studytime + failures + schoolsup + paid + higher +
    internet + romantic + famrel + goout + health, data = student_

Residuals:
     Min       1Q   Median       3Q      Max 
-13.1360  -1.4346   0.3203   2.0390   7.5991 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.86455    0.90398  10.912  < 2e-16 ***
schoolMS     -0.40530    0.27299  -1.485 0.137932    
addressU      0.39536    0.25780   1.534 0.125448    
famsizeLE3    0.38296    0.23491   1.630 0.103361    
Medu          0.19787    0.12868   1.538 0.124438    
Mjobhealth    1.17489    0.52888   2.221 0.026538 *  
Mjobother     0.01799    0.31361   0.057 0.954266    
Mjobservices  0.61785    0.37580   1.644 0.100463    
Mjobteacher  -0.04271    0.49320  -0.087 0.931004    
Fjobhealth   -0.05961    0.71451  -0.083 0.933532    
Fjobother    -0.15477    0.46568  -0.332 0.739696    
Fjobservices -0.49365    0.48521  -1.017 0.309206    
Fjobteacher   0.91146    0.63644   1.432 0.152413    
studytime     0.44517    0.13273   3.354 0.000826 ***
failures     -1.75839    0.17236 -10.202  < 2e-16 ***
schoolsupyes -1.34244    0.34252  -3.919 9.47e-05 ***
paidyes      -1.05565    0.26605  -3.968 7.76e-05 ***
higheryes     1.45195    0.41007   3.541 0.000417 ***
internetyes   0.41283    0.28315   1.458 0.145155    
romanticyes  -0.60777    0.22542  -2.696 0.007129 ** 
famrel        0.19332    0.11529   1.677 0.093890 .  
goout        -0.27957    0.09323  -2.999 0.002776 ** 
health       -0.21207    0.07615  -2.785 0.005452 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.399 on 1021 degrees of freedom
Multiple R-squared:  0.2426,	Adjusted R-squared:  0.2263 
F-statistic: 14.87 on 22 and 1021 DF,  p-value: < 2.2e-16
```

MLR Model Comparison

Model	Adjusted R-squared	Parameters	Significant Parameters
Model1	0.09017	7	11
Model2	0.2087	13	10
Model3	**0.3192**	**14**	**13**
Model4	0.2263	16	11

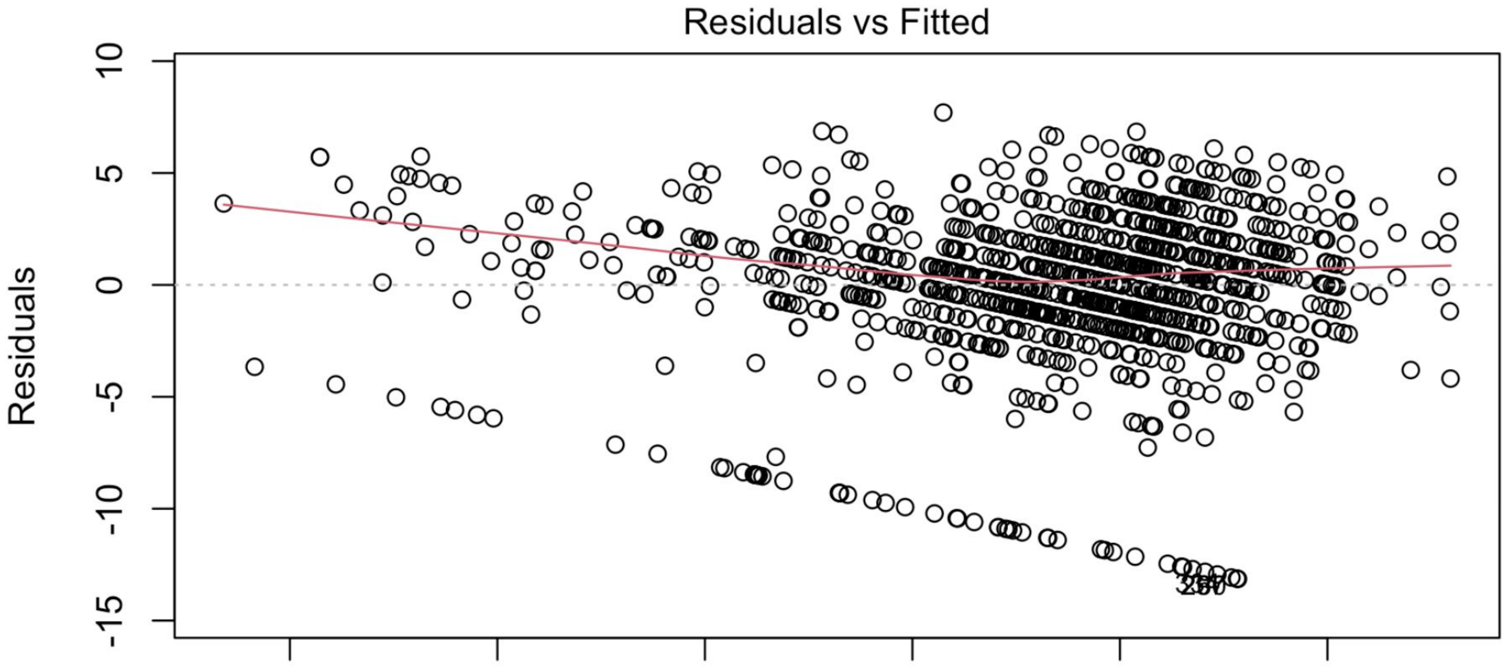
Model Tests

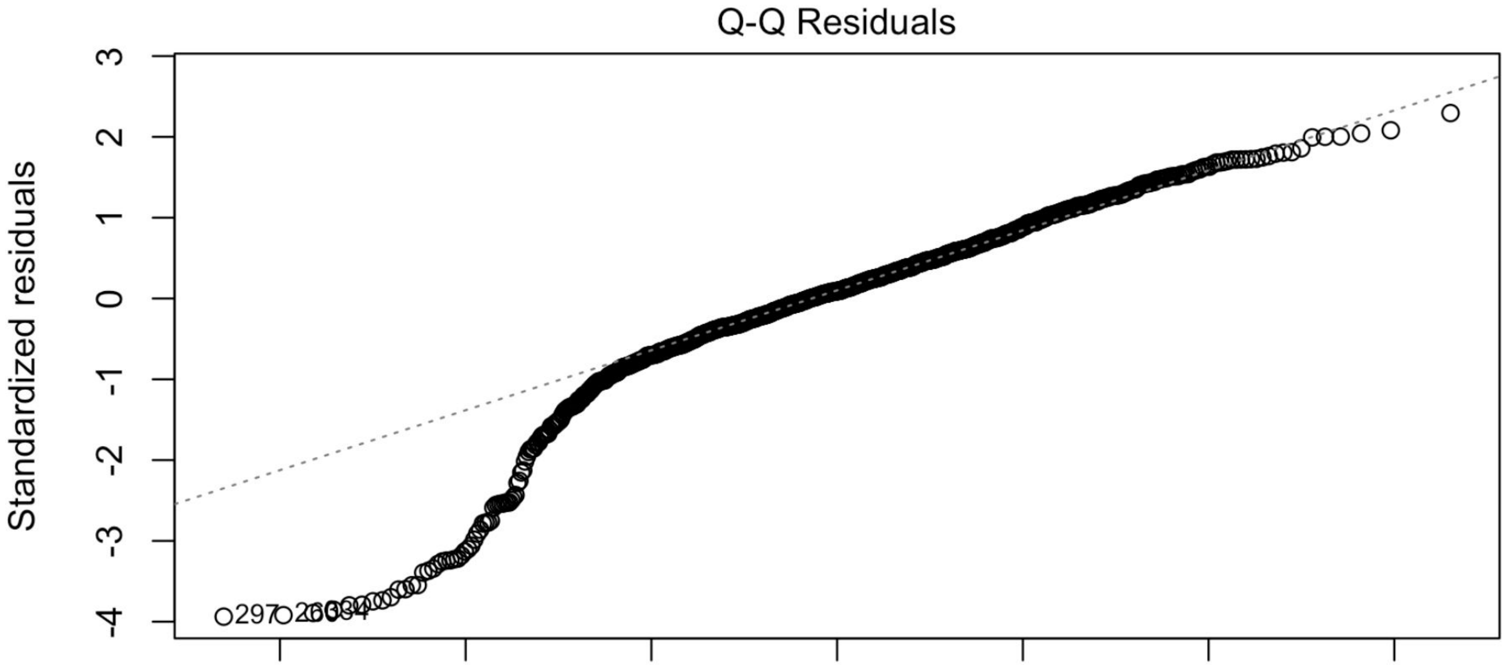
Overall model significance using F-tests for MLR Model

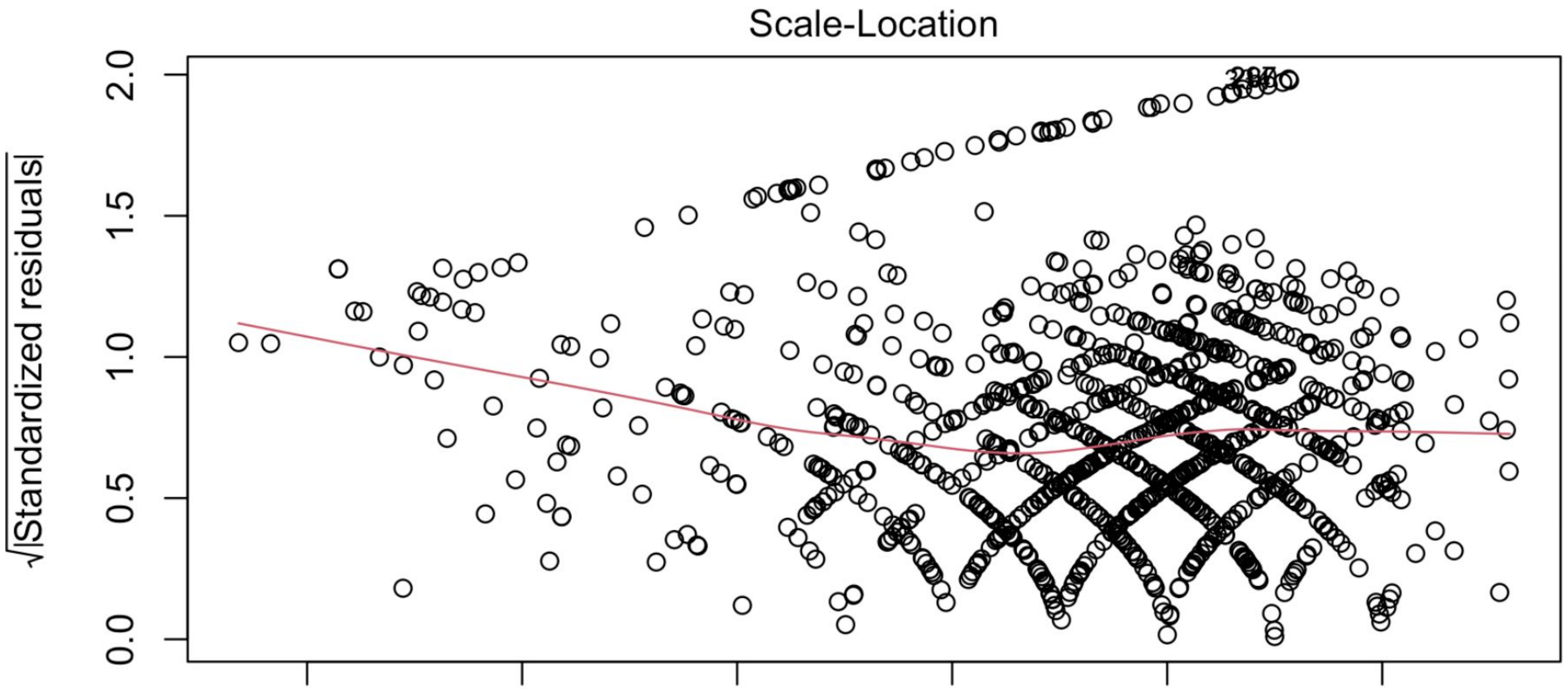


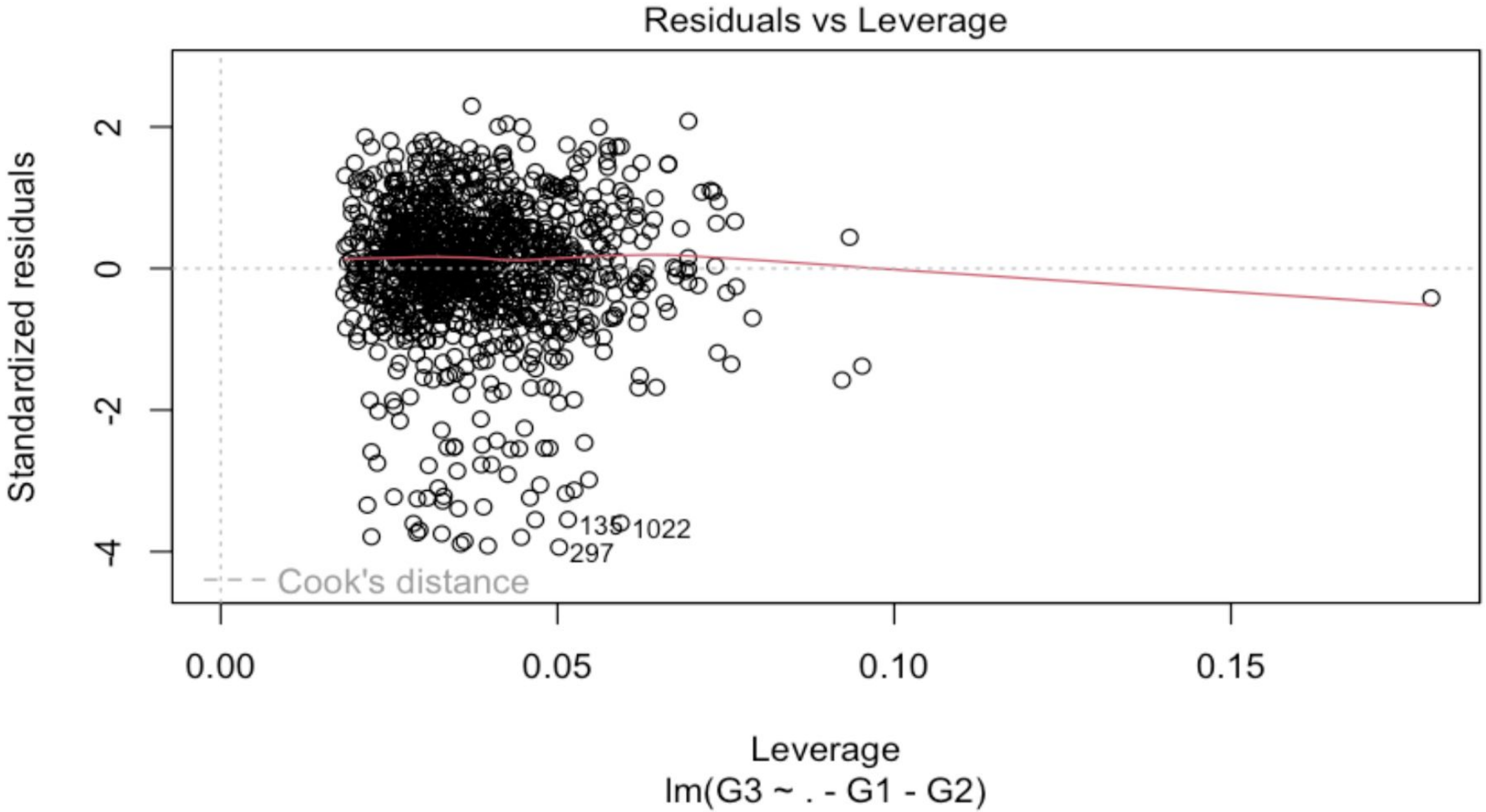
```
Analysis of Variance Table

Response: G3
             Df  Sum Sq Mean Sq  F value    Pr(>F)    
school        1   251.7  251.72  21.5114 3.982e-06 ***
sex           1    26.3   26.34   2.2506 0.1338785    
age           1   176.2  176.17  15.0550 0.0001112 ***
address       1    92.0   92.04   7.8655 0.0051360 ** 
famsize       1    67.2   67.21   5.7436 0.0167306 *  
Pstatus       1     0.6    0.57   0.0485 0.8257728    
Medu          1   435.4  435.40  37.2076 1.514e-09 ***
Fedu          1    16.8   16.76   1.4320 0.2317264    
Mjob          4    44.3   11.08   0.9471 0.4359011    
Fjob          4    57.2   14.29   1.2212 0.3000729    
reason        3   118.8   39.61   3.3849 0.0176641 *  
guardian      2    38.0   19.01   1.6242 0.1975947    
traveltime    1     5.2    5.20   0.4440 0.5053522    
studytime     1   244.3  244.26  20.8740 5.515e-06 ***
failures      1  1506.4 1506.40 128.7319 < 2.2e-16 ***
schoolsup     1   141.7  141.68  12.1078 0.0005237 ***
famsup        1    19.1   19.13   1.6351 0.2012940    
paid          1   130.1  130.09  11.1174 0.0008865 ***
activities    1     0.3    0.28   0.0242 0.8762881    
nursery       1     2.0    1.96   0.1676 0.6823259    
higher        1   160.3  160.29  13.6974 0.0002264 ***
internet      1    15.5   15.50   1.3245 0.2500531    
romantic      1    91.7   91.66   7.8329 0.0052286 ** 
famrel        1    17.2   17.23   1.4727 0.2252057    
freetime      1     8.5    8.53   0.7289 0.3934356    
goout         1    79.8   79.85   6.8236 0.0091304 ** 
Dalc          1    12.6   12.61   1.0773 0.2995564    
Walc          1     0.0    0.03   0.0027 0.9584534    
health        1    71.0   71.00   6.0676 0.0139353 *  
absences      1     0.0    0.02   0.0019 0.9655115    
Residuals  1004 11748.7   11.70                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual Plots MLR Model

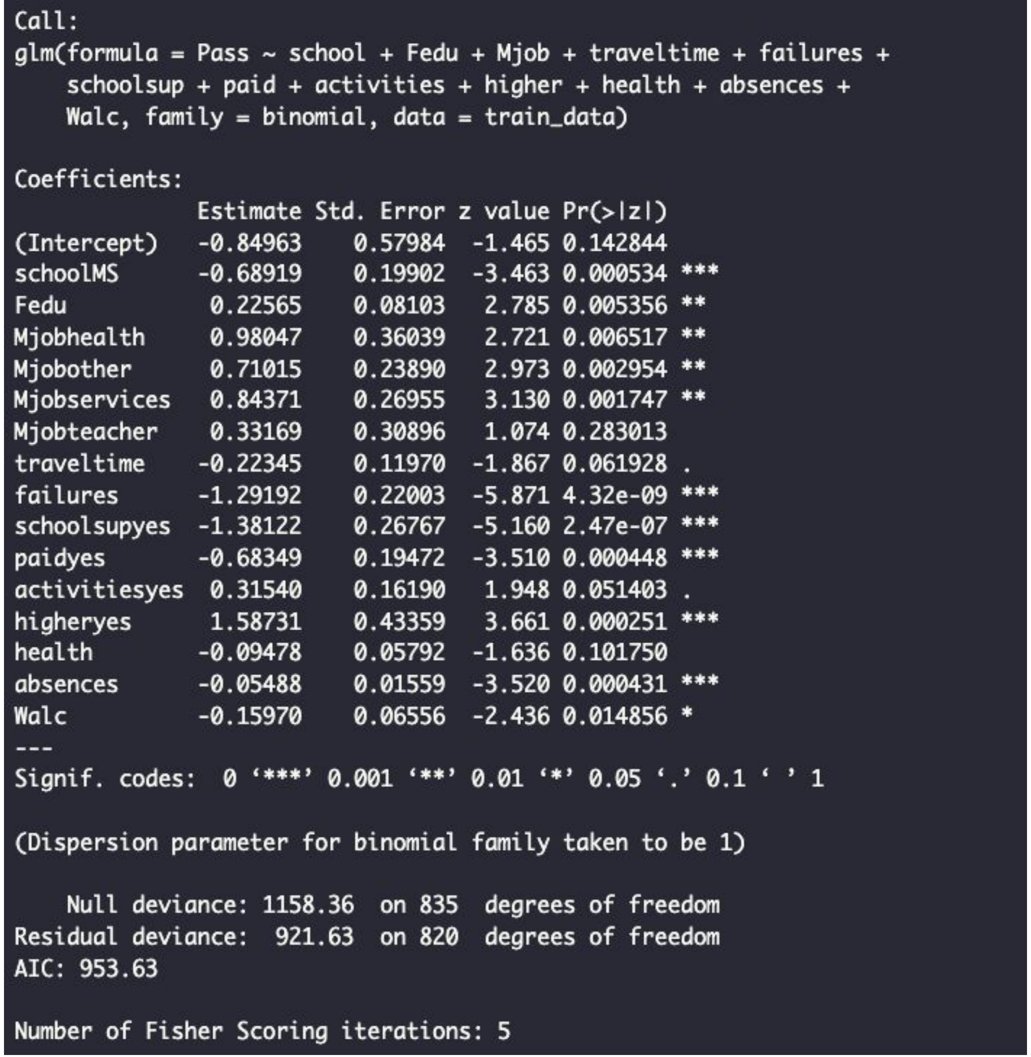



Q-Q Residuals
Standardized residuals
3
2
1
0
-1
-2
-3
-4
297
260
334


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0

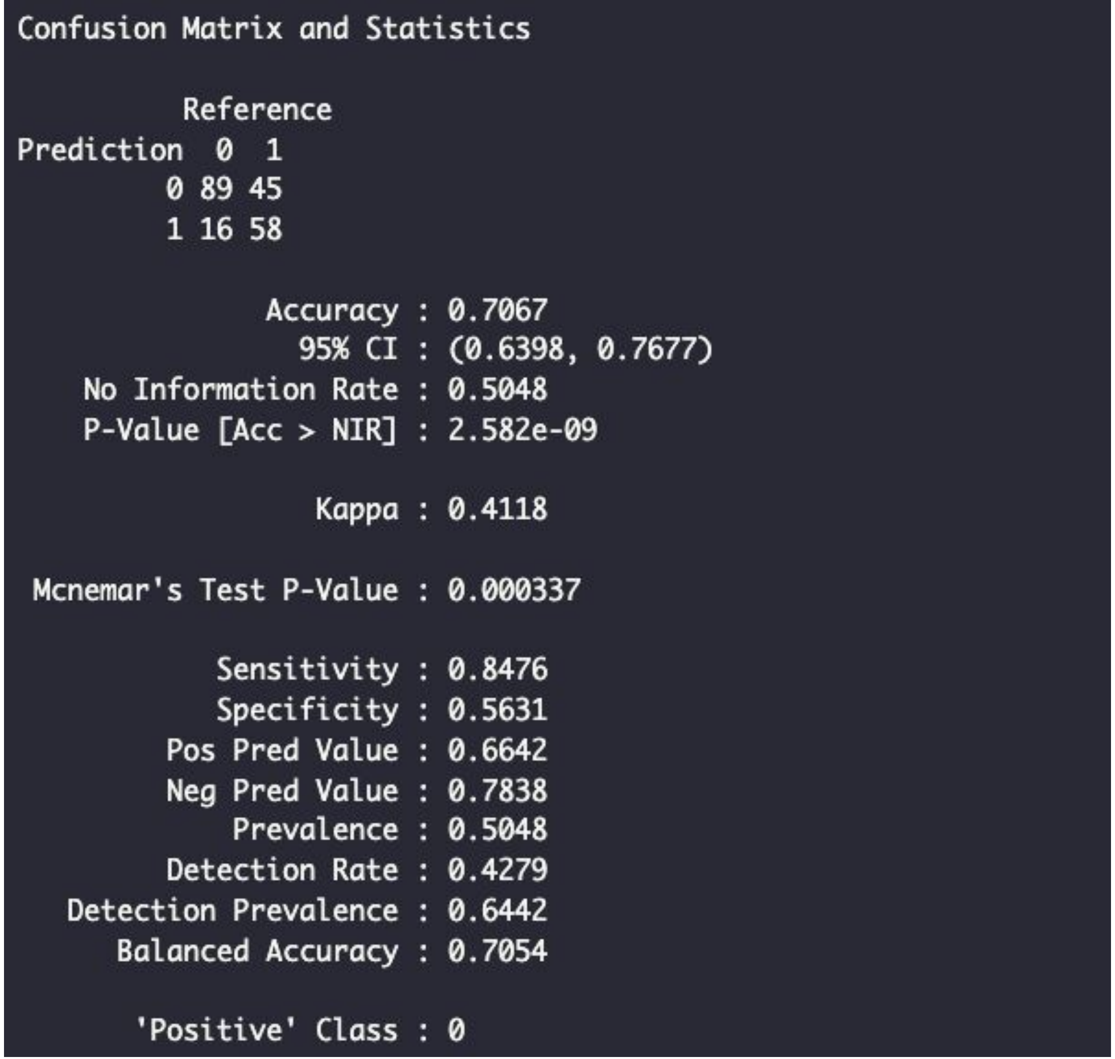

Residuals vs Leverage
Standardized residuals
2
0
-2
-4
135
1022
297
Cook's distance
0.00
0.05
0.10
0.15
Leverage
lm(G3 ~ . - G1 - G2)

Logistic Model



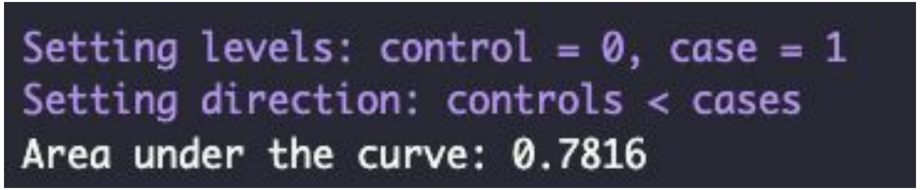
```
Call:
glm(formula = Pass ~ school + Fedu + Mjob + traveltime + failures +
    schoolsup + paid + activities + higher + health + absences +
    Walc, family = binomial, data = train_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.84963    0.57984  -1.465 0.142844
schoolMS      -0.68919    0.19902  -3.463 0.000534 ***
Fedu           0.22565    0.08103   2.785 0.005356 **
Mjobhealth     0.98047    0.36039   2.721 0.006517 **
Mjobother      0.71015    0.23890   2.973 0.002954 **
Mjobservices   0.84371    0.26955   3.130 0.001747 **
Mjobteacher    0.33169    0.30896   1.074 0.283013
traveltime    -0.22345    0.11970  -1.867 0.061928 .
failures      -1.29192    0.22003  -5.871 4.32e-09 ***
schoolsupyes  -1.38122    0.26767  -5.160 2.47e-07 ***
paidyes       -0.68349    0.19472  -3.510 0.000448 ***
activitiesyes  0.31540    0.16190   1.948 0.051403 .
higheryes      1.58731    0.43359   3.661 0.000251 ***
health        -0.09478    0.05792  -1.636 0.101750
absences      -0.05488    0.01559  -3.520 0.000431 ***
Walc          -0.15970    0.06556  -2.436 0.014856 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1158.36  on 835  degrees of freedom
Residual deviance:  921.63  on 820  degrees of freedom
AIC: 953.63

Number of Fisher Scoring iterations: 5
```

Logit Model Evaluation



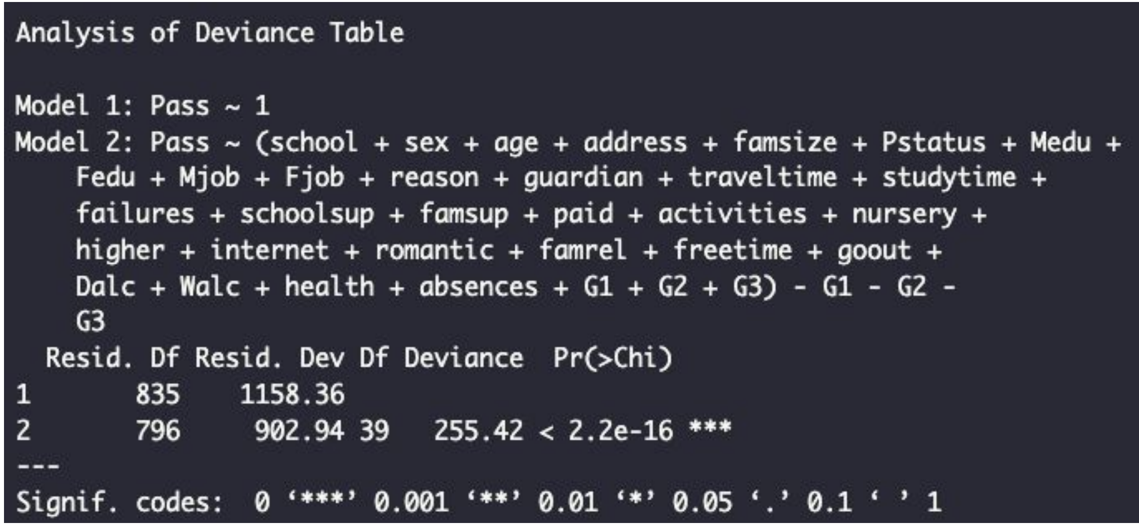
```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 89 45
         1 16 58

               Accuracy : 0.7067
                 95% CI : (0.6398, 0.7677)
    No Information Rate : 0.5048
    P-Value [Acc > NIR] : 2.582e-09

                  Kappa : 0.4118

 Mcnemar's Test P-Value : 0.000337

            Sensitivity : 0.8476
            Specificity : 0.5631
         Pos Pred Value : 0.6642
         Neg Pred Value : 0.7838
             Prevalence : 0.5048
         Detection Rate : 0.4279
   Detection Prevalence : 0.6442
      Balanced Accuracy : 0.7054

       'Positive' Class : 0
```

Area Under the Curve



```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
Area under the curve: 0.7816
```

Logit Model Evaluation using Likelihood Ratio Test



```
Analysis of Deviance Table

Model 1: Pass ~ 1
Model 2: Pass ~ (school + sex + age + address + famsize + Pstatus + Medu + 
    Fedu + Mjob + Fjob + reason + guardian + traveltime + studytime + 
    failures + schoolsup + famsup + paid + activities + nursery + 
    higher + internet + romantic + famrel + freetime + goout + 
    Dalc + Walc + health + absences + G1 + G2 + G3) - G1 - G2 - 
    G3
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
1       835    1158.36                          
2       796     902.94 39   255.42 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Analysis and Conclusions

- High Alcohol consumption negatively impacts grades.
- Romantic Interests do have a significant influence on Grades and impact negatively.
- Aspiration for higher education highly influences the final grades, having a positive influence.
- Enrolling in extra paid classes does contribute to improvement in final grades.
- We may need more data understand the effect of going out on final grades as its behavior is inconsistent across models.
- Absences do have a slight negative impact on the final grades, however we need more data to confirm.
- Mother's education and jobs have a stronger influence on the final grades of students, and mothers having health sector jobs have the most positive influence on the student grades.
- Activities seem to have a slight positive influence on the final grades.
- Higher study time are related to higher grades, and lower free times consequently. However, free time seems to not have a consistent effect on final grades.

Thank you!