학업성취도 분석
17010668 황창현

목차

- 문제정의
- EDA
- 결과 분석
- modeling
- re-modeling
- 한계점

1. 문제정의

데이터 변수 정의

```
1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinh
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" - greate
6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A
7 Medu - mother's education (numeric: 0 - none,  1 - primary education (4th g
8 Fedu - father's education (numeric: 0 - none,  1 - primary education (4th g
9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "serv
10 Fjob - father's job (nominal: "teacher", "health" care related, civil "ser
11 reason - reason to choose this school (nominal: close to "home", school "r
12 guardian - student's guardian (nominal: "mother", "father" or "other")
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3
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 (Math or Portuguese) (
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
25 freetime - free time after school (numeric: from 1 - very low to 5 - very
26 goout - going out with friends (numeric: from 1 - very low to 5 - very hig
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very
29 health - current health status (numeric: from 1 - very bad to 5 - very goo
30 absences - number of school absences (numeric: from 0 to 93)

# grades
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)
```

school,sex,address,famsize,Mjob,Fjob,famsup,Pstatus,resson,guardian,nursery,paid,higher,internet,romantic,activities,schoolsup은 각각의 unique 값들이 범주형이자 독립적이다.

age,traveltime,studytime,failures,famrel,freetime,gout,Dalc,Walc,health 은 숫자형이지만, 범주형이고 순서형이라고도 볼 수 있을 것 같다.

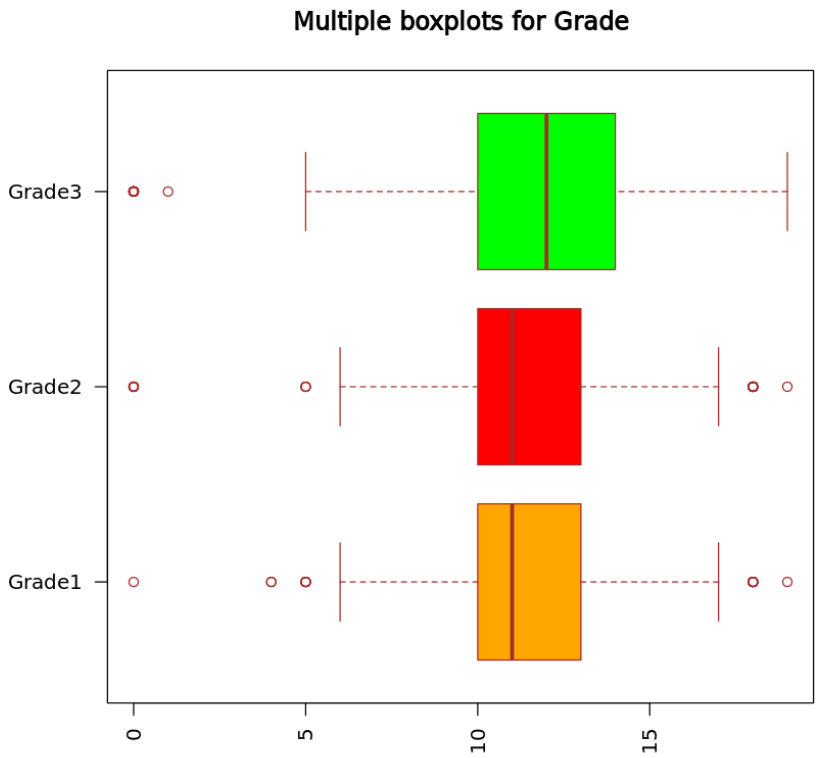
absences와 G1,G2,G3는 연속형 변수다.

찾아보니까 위 학교들의 grade는 10을 넘어야 pass라고 했다. Pass grade is 10. [16]

이번엔 학업성취를 위해 노력을 했는가를 분석하려고 한다. Grade를 통해 열심히 하고자 하는 학생들을 걸러내고자 했다.

2. EDA

Grade



G1	G2	G3
<int>	<int>	<int>
0	11	11
9	11	11
12	13	12
14	14	14
11	13	13
12	12	13

Grade의 순서를 나타낸다. 각각의 행은 독립적으로 학생 개개인의 성적을 나타낸다. 처음에 못했다가 잘하는 학생도 있고, 잘했다가 못해진 학생도 있다.

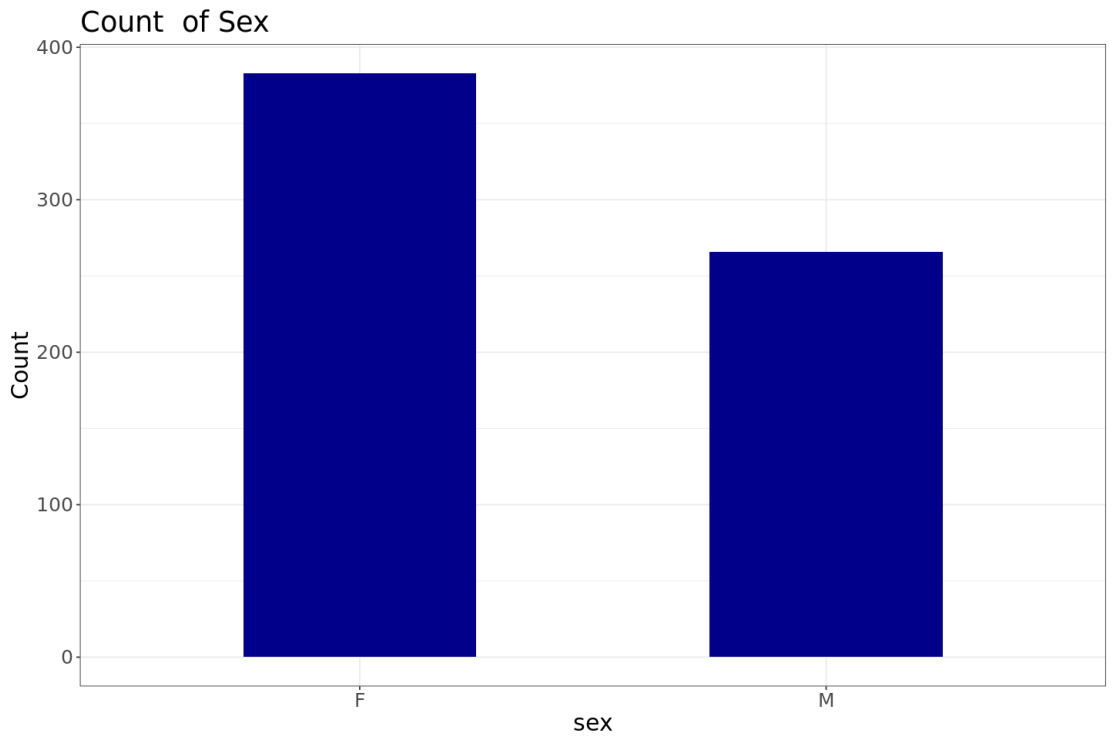
열심히 하고자 하는 학생들을 최종 성적 G3를 기준으로 G1와G2 보다 같거나 큰 범주변수 comp를 하나 만들었다.

그리고 comp를 기준으로 학생들의 특징을 살펴봤다.

```
## 좋은 성적을 받았으면 yes 아니면 no
## 기준은 첫번째 grade또는 두번째 grade가 마지막 grade보다 같거나 작다면 yes.

student$comp=ifelse((student$G1<=student$G3)|(student$G2<=student$G3),'yes','no')
```

2. EDA

Sex




```
        Pearson's Chi-squared test with Yates' continuity correction

data:  sex_comp_table
X-squared = 3.177, df = 1, p-value = 0.07468
```

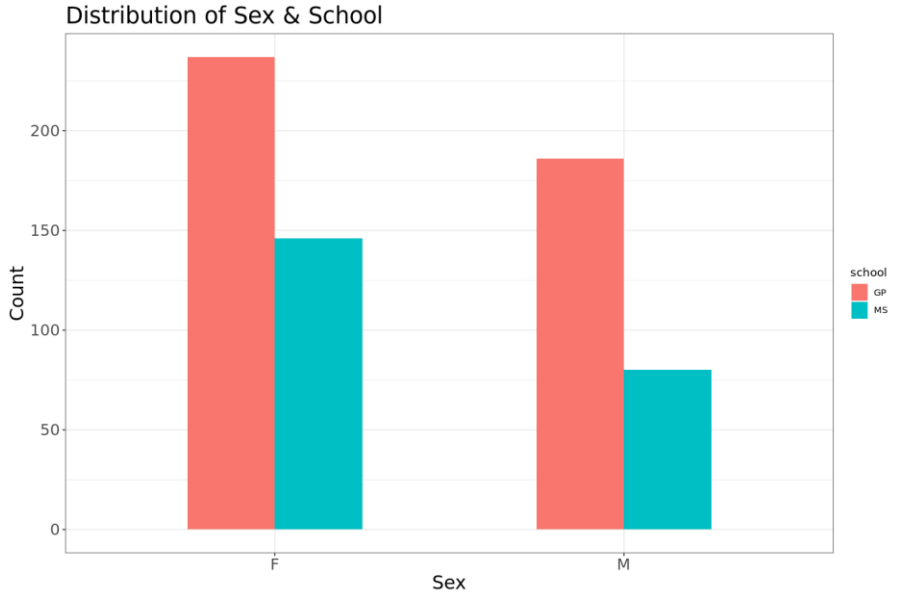
먼저 학생의 성비를 봤고, 여자가 약 1.5배 많은 것을 확인했다.

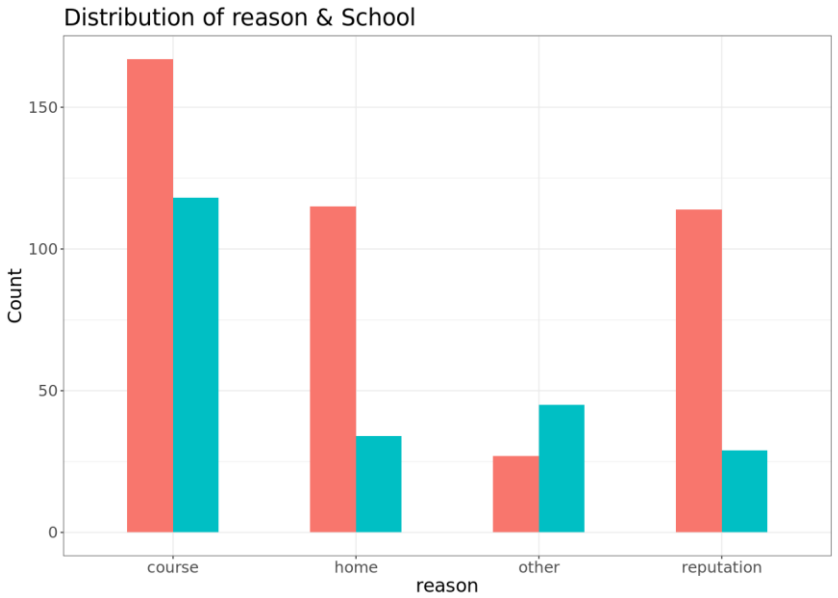
후에 comp와 sex의 빈도차이를 카이제곱검정을 통해 확인했다.

보통 유의수준을 0.05 혹은 0.01로 낮게 잡아야한다고 하는데 0.075정도면 어느정도 빈도의 차이가 있다는 것이다.

예를 들어, 남자에서의 열심히 하는 학생, 열심히 안하는 학생의 빈도는 여자에서의 열심히 하는 학생, 열심히 안하는 학생의 차이가 있다는 뜻이다.

Sex & School






```
prop.table(sex_school_table,2)
```

```
     school
sex         GP        MS
  F 0.5602837 0.6460177
  M 0.4397163 0.3539823
```

성별과 학교의 차이를 봤다. GP학교가 더 많았다.

GP학교의 특징으로는 course를 많이 보고 온다는 것을 알았다.
명성도 MS가 오는 이유보다 좋았다.

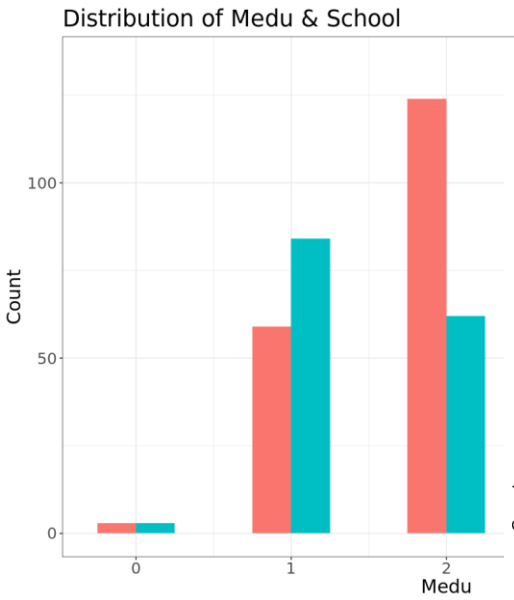
reason에 따라 열심히 하는 학생의 비율차이도 볼 수 있었다.

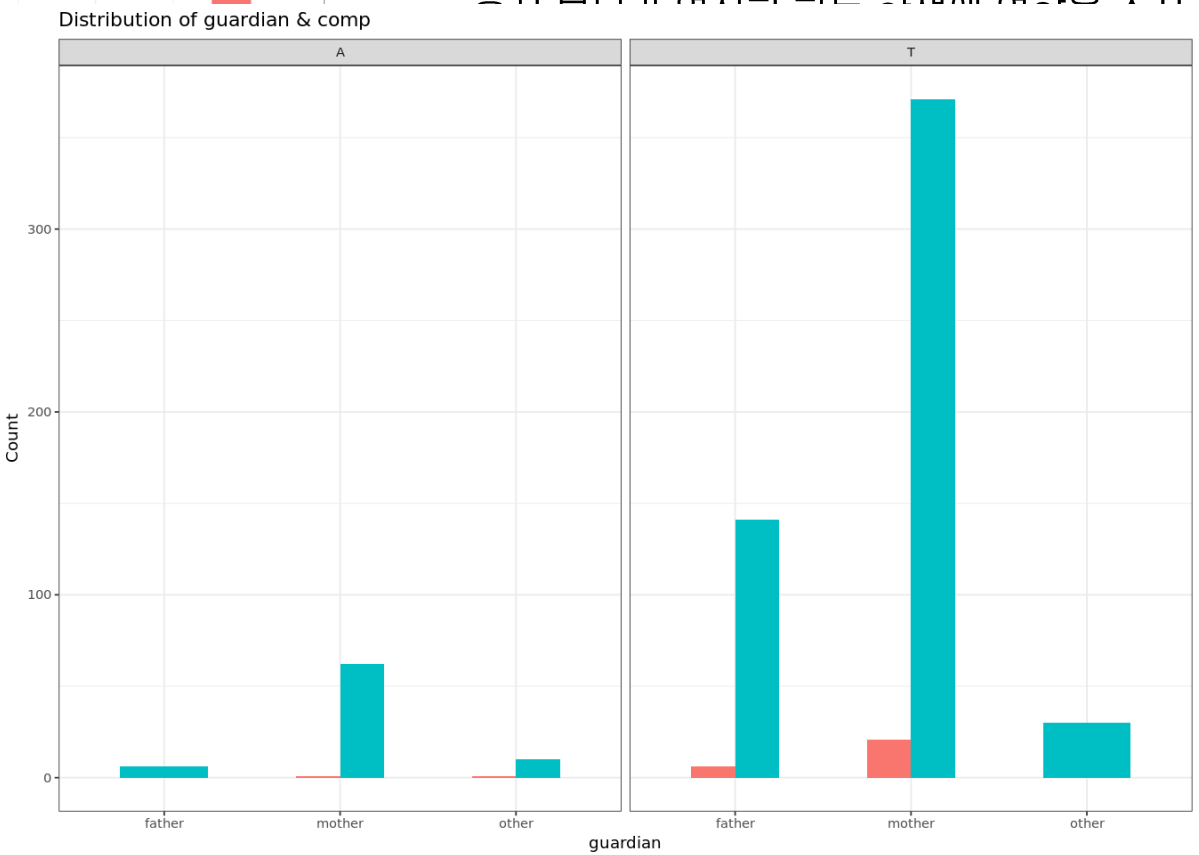
일단 열심히 하는 학생들이 대부분이다. 열심히 안하는 학생들은 소수였다. unbalanced 한 데이터셋이다.

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  sex_school_table
X-squared = 4.1289, df = 1, p-value = 0.04216
```

성별과 학교의 빈도차이는 있었다.
즉, 테이블에서도 볼 수 있듯이 성비의 차이가 분명히 있었다.

Medu & school





까 가정했다.
지만, 크게 눈에 띄

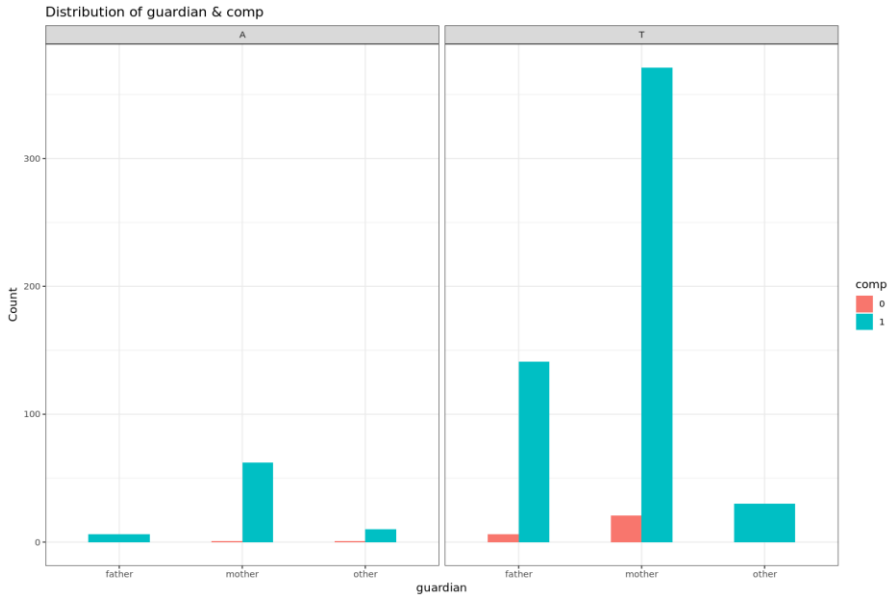
```
isq.test(Medu_comp_table):
mation may be incorrect"
```

-squared test

e
f = 4, p-value = 0.6224

했다.

Guardian & Pstatus & comp

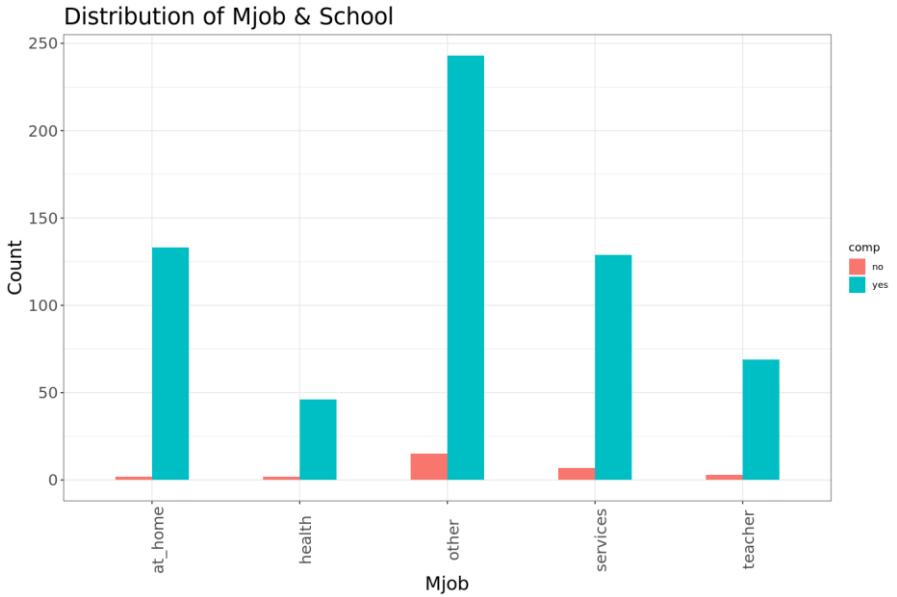


대부분 같이 살고 따로 떨어져 산다면 other이 많이 보호자가 되는것도 봤다. 대부분 어머니가 많이 보호자가 된다.

따로 떨어져 산다면 아버지가 보호자가 되면 열심히 하는 학생이다.

카이제곱검정을 했지만, guardian과 comp의 빈도차이는 없었다.

Mjob & School



어머니의 직업에 따라 차이가 있는지 봤는데 대부분 other이고 큰 차이가 없었다.

```
Warning message in chisq.test(Mjob_comp_table):
 "Chi-squared approximation may be incorrect"
```

```
        Pearson's Chi-squared test

data:  Mjob_comp_table
X-squared = 4.0881, df = 4, p-value = 0.3942
```

```
fisher.test(Mjob_comp_table)
```

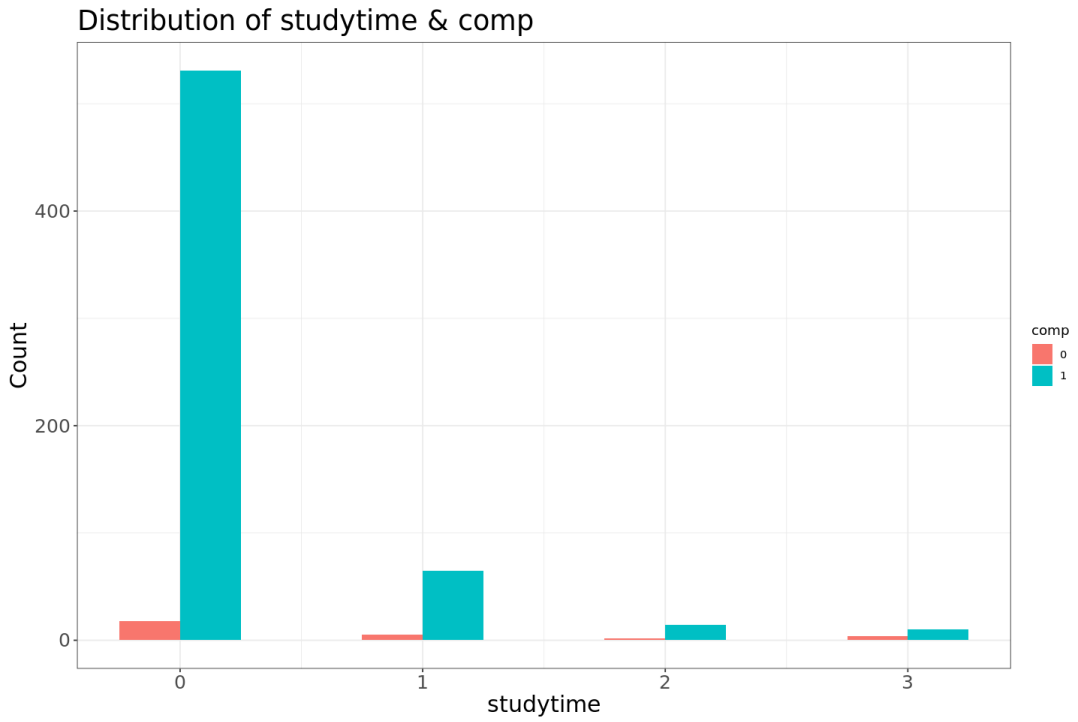
```
        Fisher's Exact Test for Count Data

data:  Mjob_comp_table
p-value = 0.3396
alternative hypothesis: two.sided
```

```
Mjob_comp_table
```

Mjob	comp 0	1
at_home	2	133
health	2	46
other	15	243
services	7	129
teacher	3	69

테이블에서 0이 나타나서 그렇다. 가정에는 틀리지 않는데 fisher test로도 해봤지만, 의미가 없었다.

Studytime & comp




```
Warning message in chisq.test(traveltime_comp_table):
 “Chi-squared approximation may be incorrect”


        Pearson's Chi-squared test

data:  traveltime_comp_table
X-squared = 0.22656, df = 3, p-value = 0.9732
```

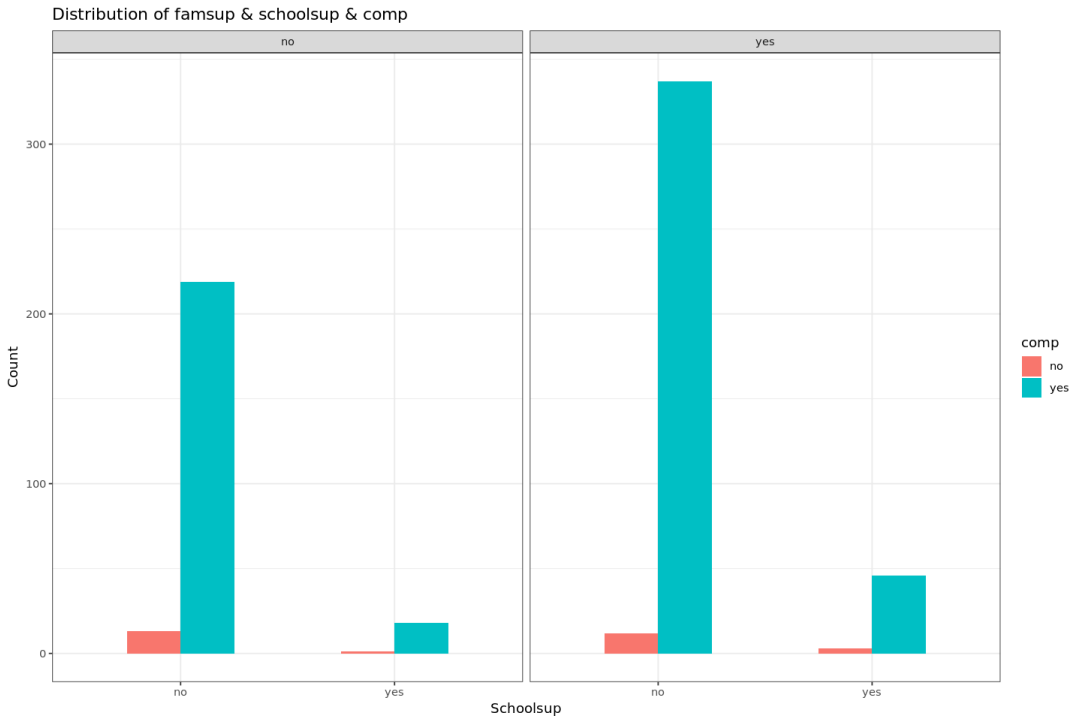
```
Warning message in chisq.test(studytime_comp_table):
 “Chi-squared approximation may be incorrect”


        Pearson's Chi-squared test

data:  studytime_comp_table
X-squared = 12.535, df = 3, p-value = 0.005758
```

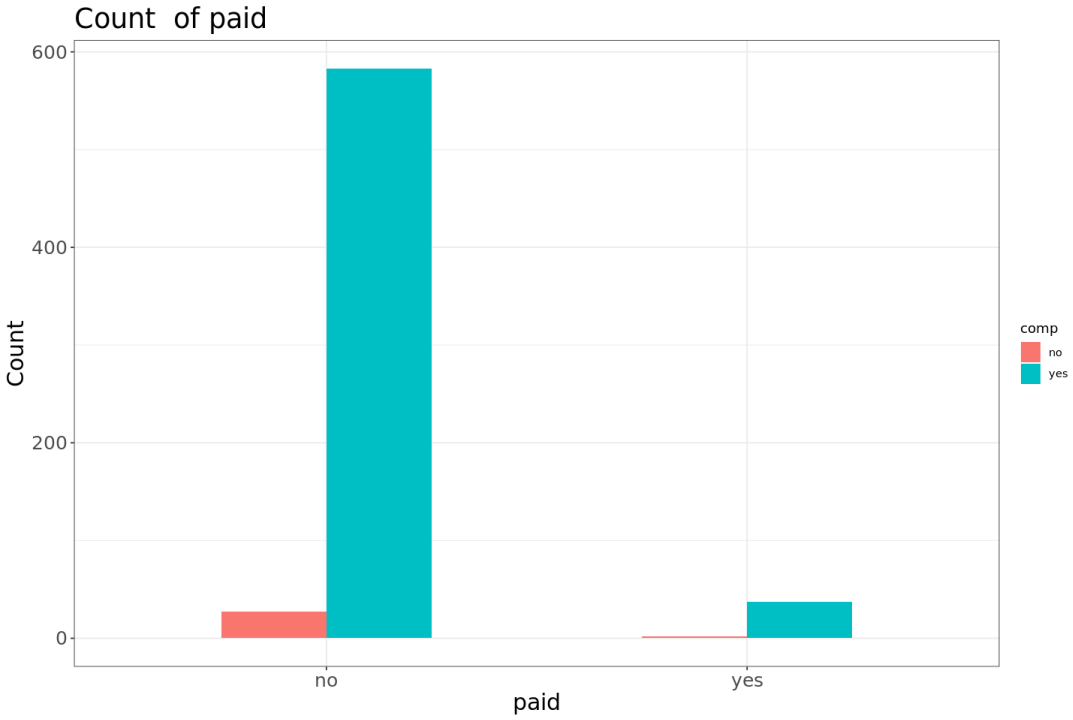
대부분 0(<2hours) 이다.

traveltime 보다는 studytime과 열심히 하는 학생과 아닌 학생의 차이가 두드러졌다. 당연했다.

2. EDA

famsup & schoolsup & paid

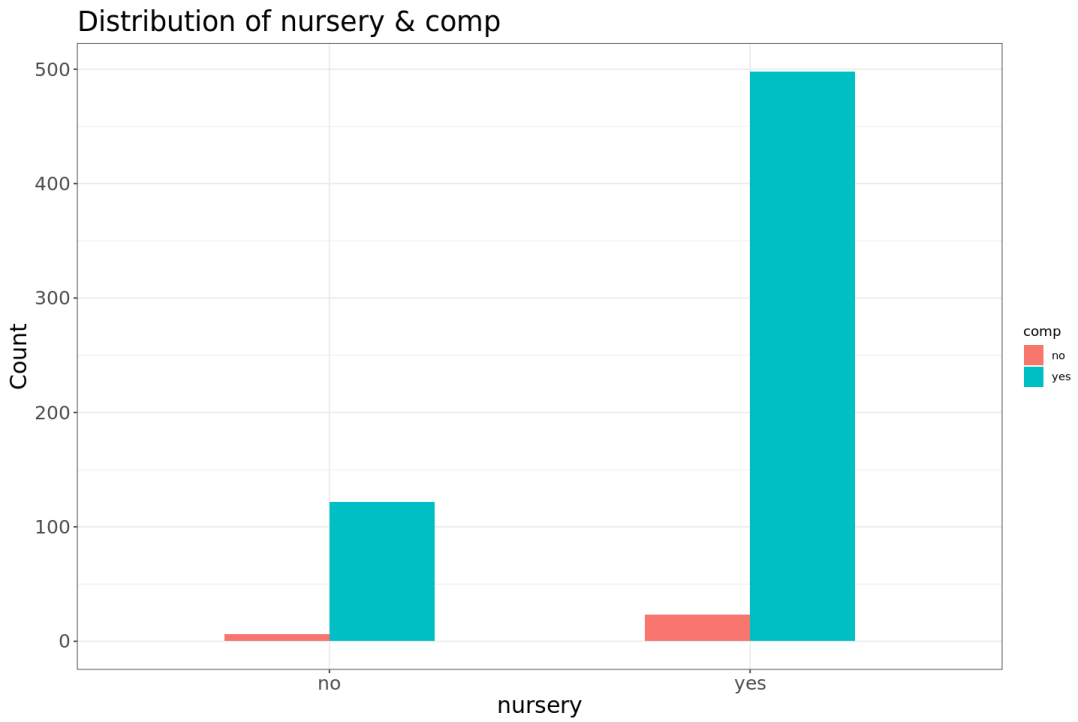


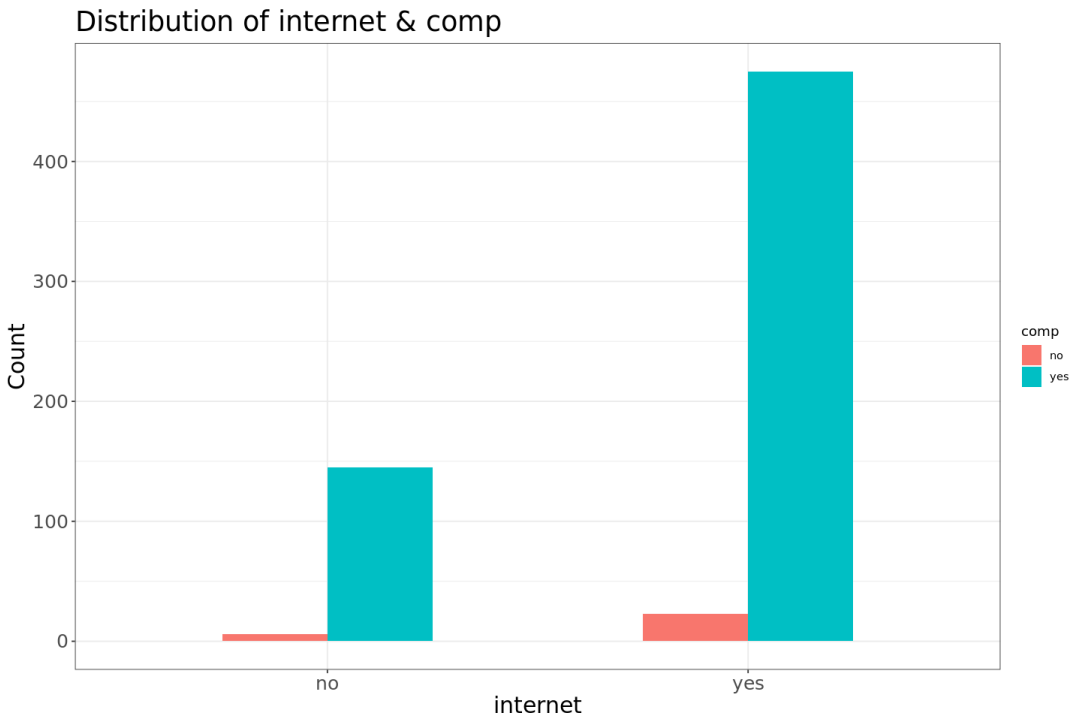


대부분 famsup은 받지만, schoolsup은 받지 않았다. 또한, paid class는 거의 듣지 않았다.

paid를 들은 학생 중에 열심히 하지 않은 학생들은 거의 없었다.

nursery & comp & paid



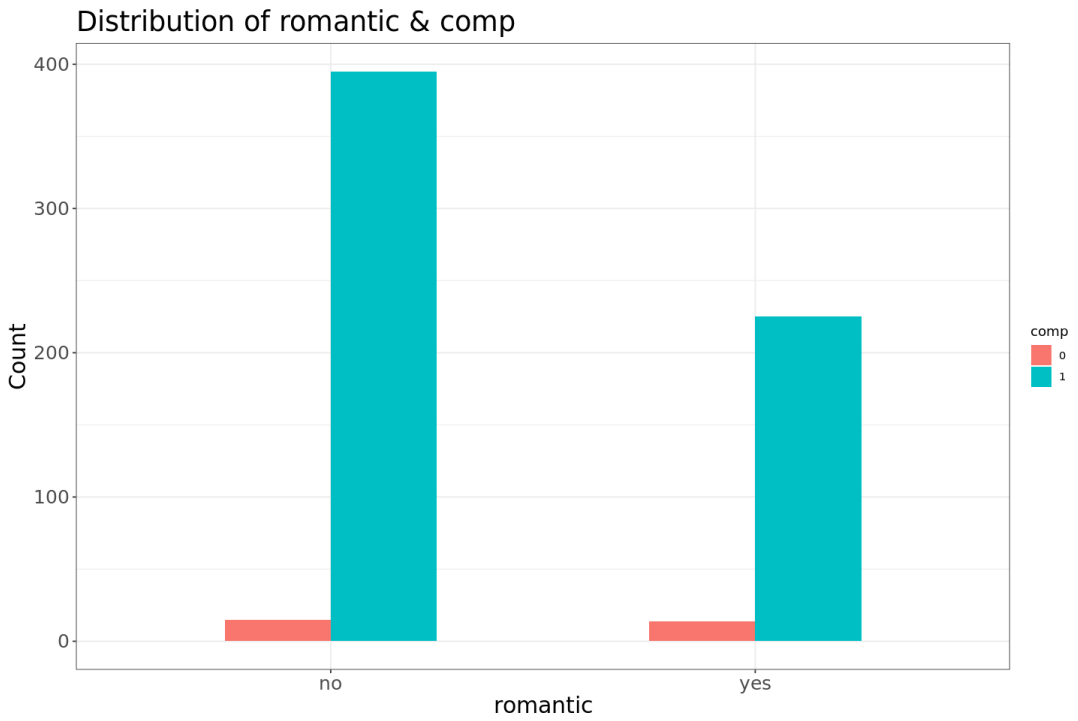


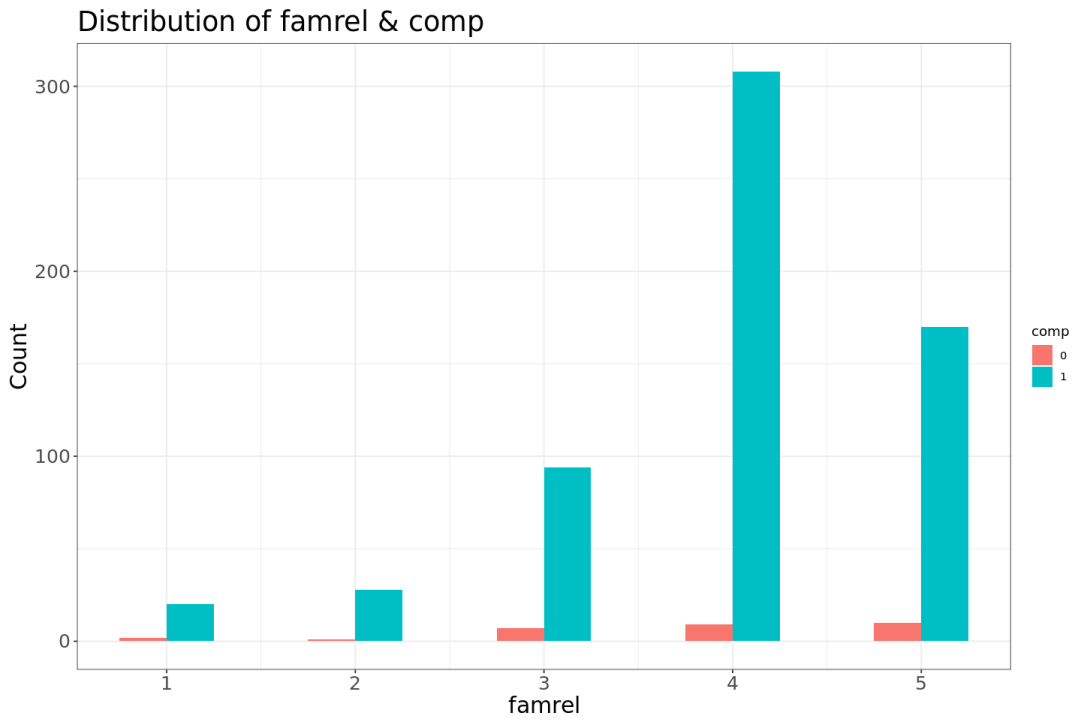
대부분 nursery(보육원)을 다닌다. 조금 이상하지만, 그렇다.

또한, internet 보급률이 잘 되어있다.

2. EDA

romantic & comp & famrel



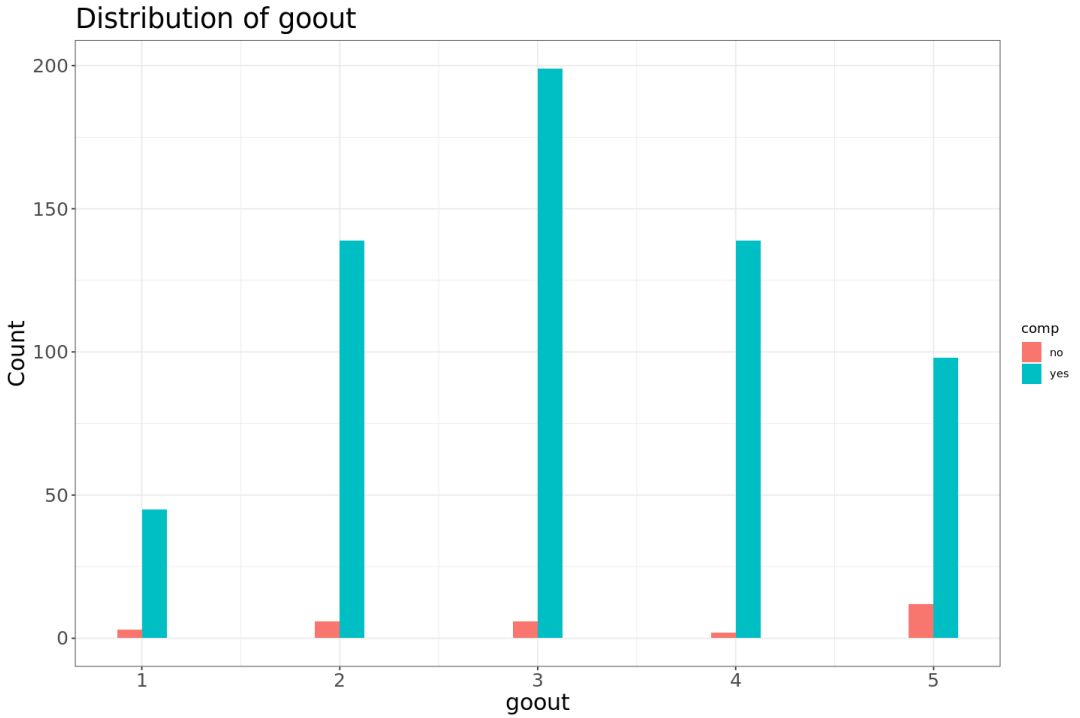


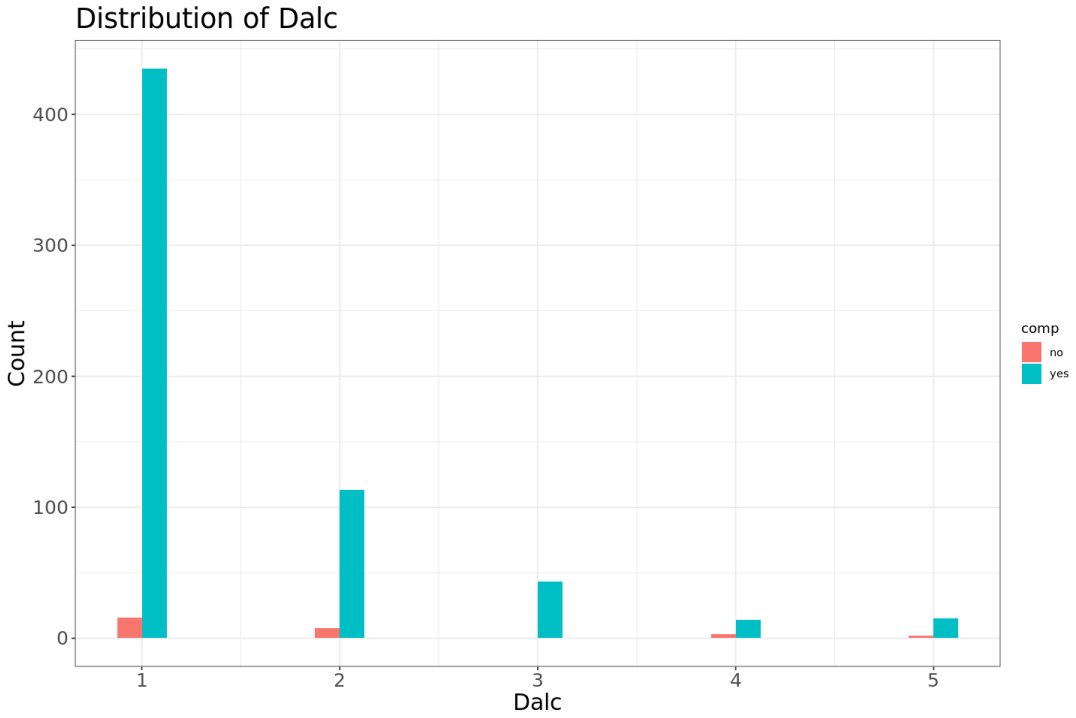
romantic(여자친구)이 대부분 없다. romantic이 없는 아이들이 열심히 하는 학생들이 많다.

famrel은 대부분 좋다. 다만, 열심히 안하는 학생들 중에 관계가 2보다 1이 더 많다. 즉, 가족관계가 이유일 수도 있다.

2. EDA

romantic & comp & famrel





열심히 하는 학생들은 goout이 보통이 많다. 다만, 그렇지 않은 학생들은 goout이 5다. 대부분.

Dalc도 열심히 하는 학생들은 1이 많지만, 그렇지 않은 학생들은 5,4의 비율이 크다.

```
prop.table(xtabs(~Dalc+comp,student),2)
```

```
    comp
Dalc          0          1
   1 0.55172414 0.70161290
   2 0.27586207 0.18225806
   3 0.00000000 0.06935484
   4 0.10344828 0.02258065
   5 0.06896552 0.02419355
```

2. EDA

numeric ordinal variables

	freetime	goout	Dalc	Walc	health
freetime	1.00000000	0.3543453	0.12717156	0.1201480	0.09510536
goout	0.35434528	1.0000000	0.23397658	0.3724547	-0.01212210
Dalc	0.12717156	0.2339766	1.00000000	0.6130561	0.08494647
Walc	0.12014798	0.3724547	0.61305611	1.0000000	0.11428202
health	0.09510536	-0.0121221	0.08494647	0.1142820	1.00000000

	G1	G2	G3
G1	1.0000000	0.8930649	0.8832876
G2	0.8930649	1.0000000	0.9444512
G3	0.8832876	0.9444512	1.0000000

연속형 변수들의 spearman 방법을 사용하여 순서 상관계수를 구했다.

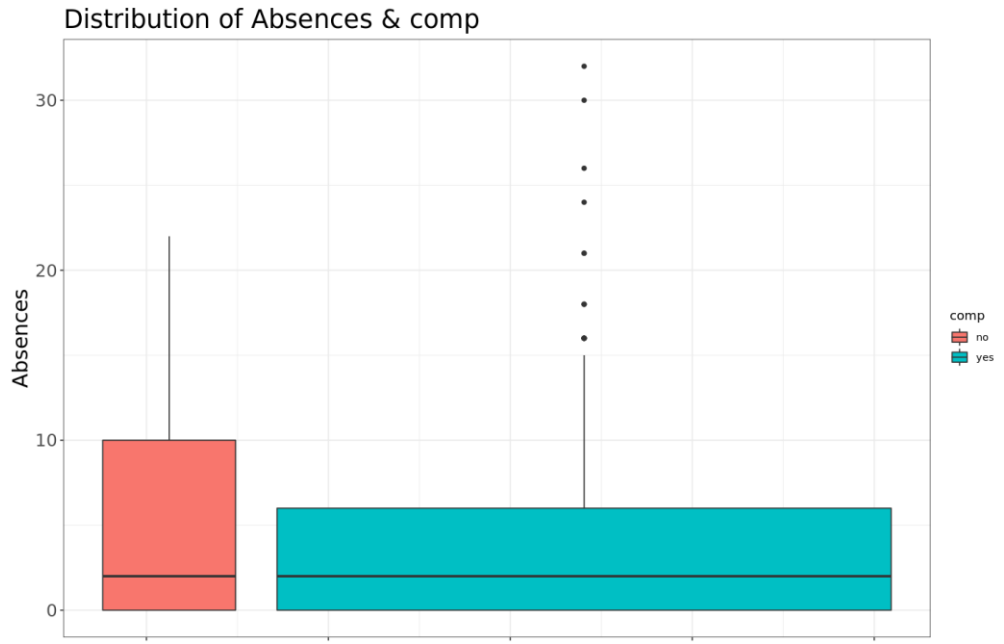
freetime과 goout이 약한 양의 상관관계가 있다.
goout과 Dalc가 약한 양의 상관관계가 있다.

후에 grade간에 상관관계가 있기 때문에 다중공선성문제를 해결하기 위해 제거한다.

2. EDA

absences

comp	x
<chr>	<dbl>
no	5.172414
yes	3.588710



absences만 봤을 때 열심히 하는 학생들과 아닌 학생들의 평균 결석 수가 다르다.

3. 결과분석

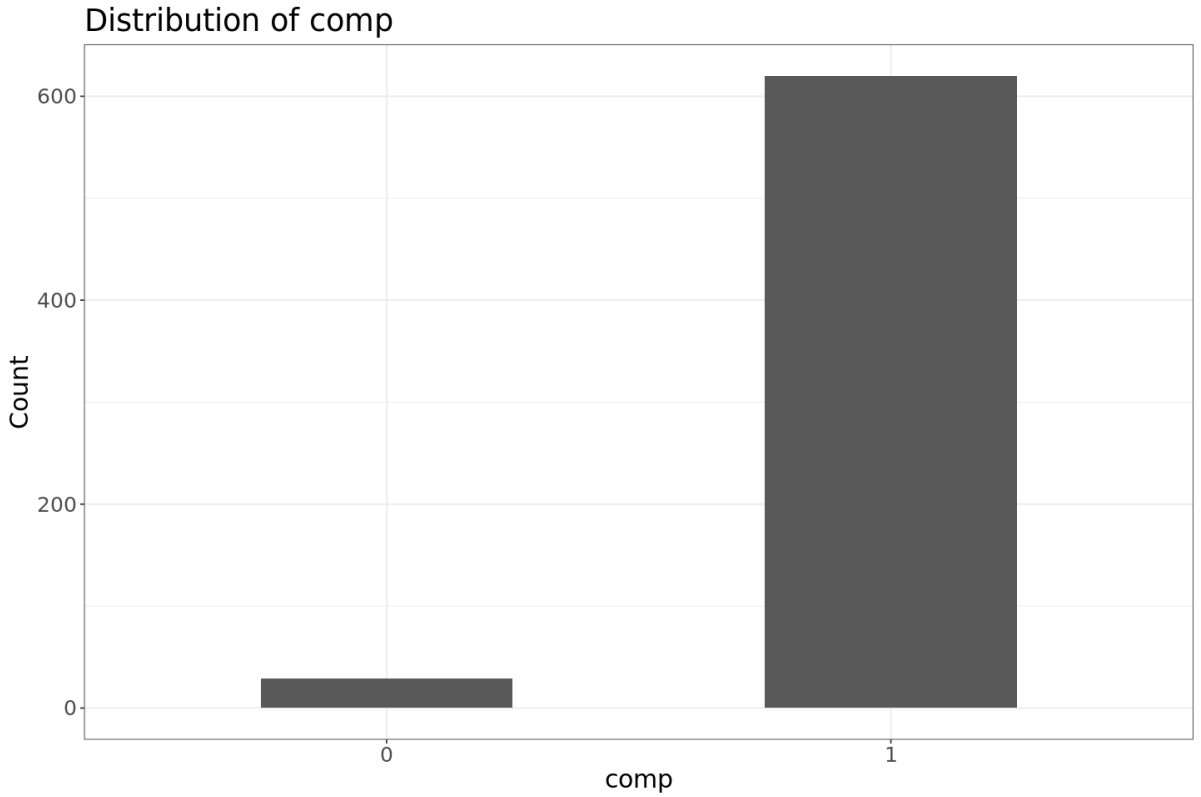
대부분 GP학교에 다녔고, 학생들의 대부분은 여자이다.
나이는 15~18세라고 되어있지만, 넘는 학생들도 있었다.
대부분 도시에 살며, 거의 부모님과 같이 살았다. 만약 따로 살게 된다면 그래도 어머니가 보호자인 경우가 많았지만, 아버지가 보호자인 경우 열심히 했다.
그래서 그런지 아버지의 교육수준에 영향이 있었고, 아버지의 높은 교육 수준은 대부분의 아이들은 GP에 다녔다. 하지만, 직업에 따른 차이는 없었다.

이유또한 집과 가까운 것부터 course등 학교를 많이 보는. 보호자 및 부모의 의견이 중요하다고 가정했다. 집과의 거리가 열심히하는 것과는 안하는것에 대한 빈도는 통계적으로 차이가 없었지만 대부분 가까울수록 열심히 하는 학생들이었다.
grade에 통과 못한 적이 있는 경우에 열심히 안한 학생들일 가능성이 높았다.

가족과의 관계도 대부분 좋았지만, 열심히 안하는 학생들은 가족관계가 안좋은 경우도 있었다.
인터넷도 대부분 보급되어있었고, 장학금은 거의 없었지만 부모로 부터 도움을 받았다.

결석이 많을 수록 열심히 안하는 학생들이었고, 술 많이 마시고, 친구들이 많을 수록 그랬다.

4. modeling



종속변수로 해놓은 comp가 0,1 의 unbalance 문제를 바로 잡아야한다.

그래서 모델링은 stratified hold out 방법을 사용하여 사이즈는 0.85:0.15로 하여 걸러내는 훈련을 시킨다.

모델은 logistic regression을 통해 확률로 구한다.

여러 가정에 의한 모델들을 실험으로 알아낸다.

가정 1. 외부요인1(주소,학교와의 거리, 이유, 시간)에 있다.
가정 2. 외부요인3(학교)에 있다.
가정 3. 외부요인2(가족)에 있다.
가정 4. 내부요인(자신)에 있다.

4-1. modeling 결과 – 외부요인1

```
Call:
glm(formula = comp ~ address + reason + traveltime, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6439   0.2525   0.2812   0.3127   0.5518

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.44306    0.71253   3.429 0.000606 ***
addressU         0.47039    0.45606   1.031 0.302341
reasonhome       0.04887    0.56660   0.086 0.931270
reasonother     -0.88892    0.53920  -1.649 0.099230 .
reasonreputation 0.26588    0.60565   0.439 0.660667
traveltime       0.25095    0.30815   0.814 0.415438
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 203.58  on 551  degrees of freedom
Residual deviance: 198.75  on 546  degrees of freedom
AIC: 210.75

Number of Fisher Scoring iterations: 6
```

address,reason,traveltime이 외부요인(거리)가 됐다.

reason이 조금이나마 영향이 끼침을 알 수 있다.

외부요인은 아니라고 할 수 있을 만큼 추정치가 낮았다.

4-1. modeling 결과 – 외부요인2

```
Call:
glm(formula = comp ~ school + reason + schoolsup + paid + activities +
    absences, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7595   0.2289   0.2660   0.3198   0.7848

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.78497    0.51648   7.328 2.33e-13 ***
schoolMS        -0.05555    0.46765  -0.119   0.9055
reasonhome       0.17416    0.57438   0.303   0.7617
reasonother     -1.05339    0.55593  -1.895   0.0581 .
reasonreputation  0.29771    0.61465   0.484   0.6281
schoolsupyes    -0.26168    0.65263  -0.401   0.6884
paidyes         -0.35388    0.78818  -0.449   0.6534
activitiesyes   -0.46139    0.42897  -1.076   0.2821
absences        -0.07814    0.03526  -2.216   0.0267 *
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 203.58  on 551  degrees of freedom
Residual deviance: 194.35  on 543  degrees of freedom
AIC: 212.35

Number of Fisher Scoring iterations: 6
```

학교와 관련된 변수들로 해봤는데 결석수가 negative한 영향을 끼쳤다.

결석이 중요한 변수임을 알 수 있다.

4-1. modeling 결과 – 외부요인3

```
Call:
glm(formula = comp ~ famsize + Pstatus + Medu + Fedu + Mjob +
    Fjob + guardian + famsup + nursery + famrel, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2207   0.1502   0.2351   0.3348   0.8567

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.971e+01  1.028e+03   0.019   0.9847
famsizeLE3     -4.631e-01  4.647e-01  -0.997   0.3190
PstatusT       -7.207e-01  7.914e-01  -0.911   0.3625
Medu           -9.120e-02  2.948e-01  -0.309   0.7570
Fedu            7.174e-01  2.981e-01   2.406   0.0161 *
Mjobhealth     -3.820e-01  1.369e+00  -0.279   0.7802
Mjobother      -1.555e+00  7.852e-01  -1.980   0.0477 *
Mjobservices   -1.268e+00  8.807e-01  -1.440   0.1500
Mjobteacher    -9.749e-01  1.210e+00  -0.806   0.4203
Fjobhealth     -1.735e+01  1.028e+03  -0.017   0.9865
Fjobother      -1.519e+01  1.028e+03  -0.015   0.9882
Fjobservices   -1.597e+01  1.028e+03  -0.016   0.9876
Fjobteacher    -1.647e+01  1.028e+03  -0.016   0.9872
guardianmother -6.039e-01  5.946e-01  -1.016   0.3098
guardianother  -2.710e-01  1.194e+00  -0.227   0.8205
famsupyes       3.667e-01  4.293e-01   0.854   0.3929
nurseryyes     -2.233e-03  5.340e-01  -0.004   0.9967
famrel         -3.796e-02  2.230e-01  -0.170   0.8649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 203.58  on 551  degrees of freedom
Residual deviance: 180.37  on 534  degrees of freedom
AIC: 216.37

Number of Fisher Scoring iterations: 17
```

가족과 관련된 변수들로 모델을 만들어봤고, eda에서 얘기했었던

아버지의 교육수준이 중요한 영향을 끼쳤다. 또한, 어머니의 직업도 영향이 있는것 같다.

4-1. modeling 결과 – 내부요인

```
Call:
glm(formula = comp ~ sex + age + studytime + failures + higher +
    romantic + freetime + goout + Dalc + health + absences, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1239   0.1752   0.2212   0.3132   1.1432

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.75787    3.49989   1.645  0.09994 .
sexM        -0.25598    0.47490  -0.539  0.58988
age         -0.03692    0.18940  -0.195  0.84544
studytime    0.51376    0.34831   1.475  0.14021
failures    -0.74274    0.27738  -2.678  0.00741 **
higheryes   -0.10728    0.61931  -0.173  0.86248
romanticyes -0.20652    0.44708  -0.462  0.64413
freetime    -0.35080    0.22932  -1.530  0.12608
goout       -0.10742    0.19271  -0.557  0.57726
Dalc         0.06300    0.23518   0.268  0.78880
health      -0.18376    0.15930  -1.154  0.24868
absences    -0.05427    0.03663  -1.482  0.13845
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 203.58  on 551  degrees of freedom
Residual deviance: 177.59  on 540  degrees of freedom
AIC: 201.59

Number of Fisher Scoring iterations: 6
```

자기 자신에 관련된 변수들로 모델을 했더니 failures가 중요한 영향을 미쳤다. 확실히 실패한 경험이 많을 수록 열심히 하는 학생이 아니라는 것을 직관적으로 알 수 있다.

4-1. modeling 결과 – 성적

```
Call:
glm(formula = comp ~ G1 + G2 + G3, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.71673   0.00610   0.02382   0.08143   3.02773

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.4162     1.7541   3.658 0.000254 ***
G1           -0.7852     0.2531  -3.102 0.001921 **
G2           -3.9767     0.9287  -4.282 1.85e-05 ***
G3            4.6990     1.0460   4.492 7.05e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 203.584  on 551  degrees of freedom
Residual deviance:  59.284  on 548  degrees of freedom
AIC: 67.284

Number of Fisher Scoring iterations: 10
```

종속변수가 G1,G2,G3로 만들었기 때문에 많은 관련이 있다.

하지만, eda에서도 말했다시피 셋의 상관계수가 매우 높기 때문에 이상한 결과를 가져올 수도 있다.

4-2. 최종 modeling 결과

```
Call:
glm(formula = comp ~ absences + failures + Fedu + Mjob + reason +
    G3 - 1, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3042   0.0876   0.1362   0.2398   1.4773

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
absences        -0.10141    0.03917  -2.589  0.00962 **
failures        -0.23535    0.33623  -0.700  0.48394
Fedu             0.27406    0.28547   0.960  0.33704
Mjobat_home      1.02720    1.10583   0.929  0.35294
Mjobhealth      -1.03682    1.39703  -0.742  0.45799
Mjobother       -0.86063    0.90212  -0.954  0.34008
Mjobservices    -0.99478    1.11267  -0.894  0.37129
Mjobteacher     -0.87576    1.31932  -0.664  0.50682
reasonhome       0.06382    0.68198   0.094  0.92544
reasonother     -0.68102    0.71572  -0.952  0.34134
reasonreputation -0.04238    0.71721  -0.059  0.95288
G3               0.39440    0.06627   5.952 2.66e-09 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 765.23  on 552  degrees of freedom
Residual deviance: 130.42  on 540  degrees of freedom
AIC: 154.42

Number of Fisher Scoring iterations: 7
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  1  0
         1  3 93

               Accuracy : 0.9691
                 95% CI : (0.9123, 0.9936)
    No Information Rate : 0.9588
    P-Value [Acc > NIR] : 0.4293

                  Kappa : 0.3899

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 0.25000
            Specificity : 1.00000
         Pos Pred Value : 1.00000
         Neg Pred Value : 0.96875
             Prevalence : 0.04124
         Detection Rate : 0.01031
   Detection Prevalence : 0.01031
      Balanced Accuracy : 0.62500

       'Positive' Class : 0
```

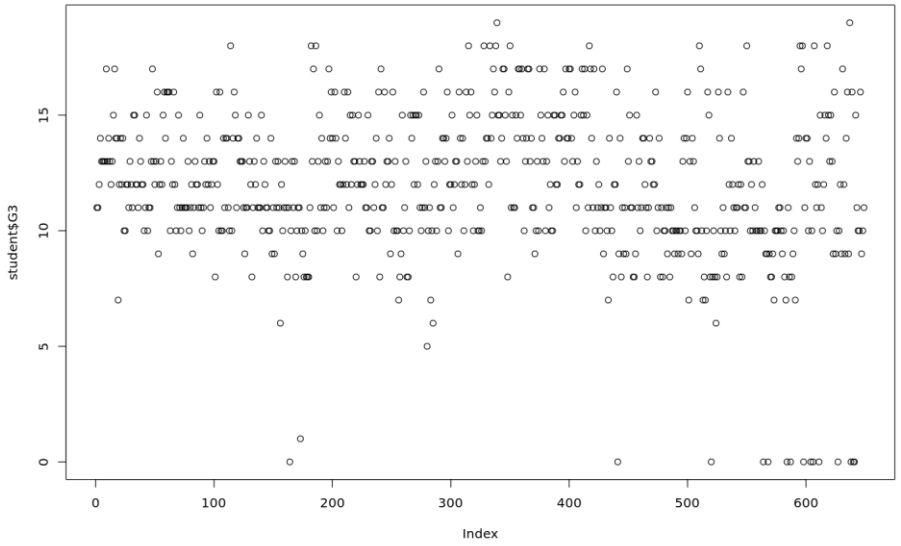
처음 나눴던 valid dataset에 적용을 했을 때 그냥 1로만 예측을 해도 95.8%의 정확도를 보일 수 있다.

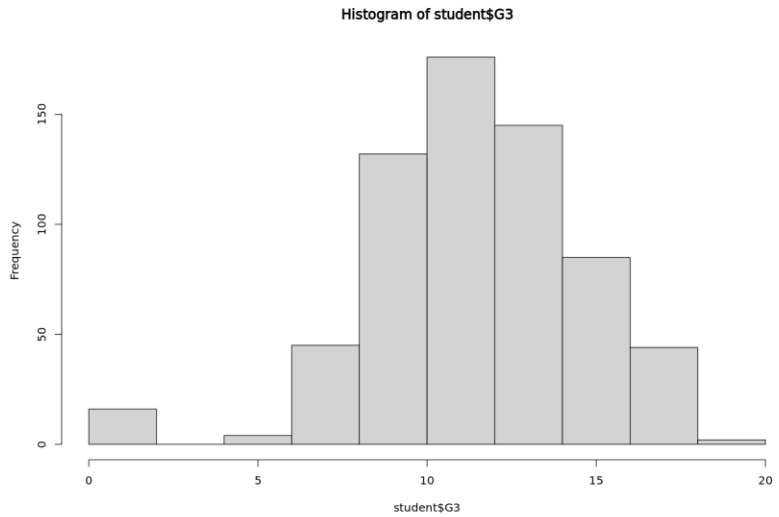
model은 변수들로만 이루어진 모델을 만들기 위해서 intercept를 제거했고,

confusion matrix 결과는 3개는 예측 못했지만, 1개는 맞췄다.

기존 1로만 했을 때보다 하나 더 맞춘것이다.

5. modeling





종속변수는 G3로 은근한 정규분포를 띄고 있다.

모델링은 stratified hold out 방법을 사용하여 사이즈는 0.85:0.15로 하여 걸러내는 훈련을 시킨다.

모델은 multi-regression을 통해 구한다.

아까 했던 가정과 같이 진행했다.

가정 1. 외부요인1(주소,학교와의 거리, 이유, 시간)에 있다.
가정 2. 외부요인3(학교)에 있다.
가정 3. 외부요인2(가족)에 있다.
가정 4. 내부요인(자신)에 있다.

5-1. modeling 결과 – 외부요인1

```
Call:
lm(formula = G3 ~ address + reason + traveltime, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-12.444  -1.464  -0.028   1.862   7.398

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      11.6368     0.4870  23.897  < 2e-16 ***
addressU          0.8205     0.3153   2.602  0.00951 **
reasonhome        0.1717     0.3519   0.488  0.62569
reasonother      -0.6227     0.4533  -1.374  0.17006
reasonreputation  1.1522     0.3582   3.217  0.00137 **
traveltime       -0.3451     0.1941  -1.777  0.07607 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.182 on 546 degrees of freedom
Multiple R-squared:  0.05689,   Adjusted R-squared:  0.04825
F-statistic: 6.587 on 5 and 546 DF,  p-value: 5.76e-06
```

address,reason,traveltime이 외부요인(거리)가 됐다.

reason이 조금이나마 영향이 끼침을 알 수 있다.

p value를 통해 어느정도 유의수준에서 Urban(도시)인지 중요하고, 명성을 보고 찾아온것도 그렇고, 등교시간이 짧을수록 G3가 큼을 알 수 있다.

5-1. modeling 결과 – 외부요인2

```
Call:
lm(formula = G3 ~ famsize + Pstatus + Medu + Fedu + Mjob + Fjob +
    guardian + famsup + nursery + famrel, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-12.681  -1.618  -0.043   1.822   7.447

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20867    0.99433   9.261   <2e-16 ***
famsizeLE3    0.39829    0.30693   1.298   0.1950
PstatusT      0.27510    0.43109   0.638   0.5236
Medu          0.30176    0.19212   1.571   0.1169
Fedu          0.34881    0.17385   2.006   0.0453 *
Mjobhealth    1.17358    0.66640   1.761   0.0788 .
Mjobother     0.17897    0.37207   0.481   0.6307
Mjobservices  0.59353    0.45712   1.298   0.1947
Mjobteacher   0.58056    0.61911   0.938   0.3488
Fjobhealth   -0.61021    0.97503  -0.626   0.5317
Fjobother     0.05783    0.57184   0.101   0.9195
Fjobservices -0.58131    0.60745  -0.957   0.3390
Fjobteacher   0.68604    0.84428   0.813   0.4168
guardianmother -0.32796    0.33200  -0.988   0.3237
guardianother  -0.88868    0.62517  -1.422   0.1558
famsupyes     0.20977    0.28054   0.748   0.4549
nurseryyes   -0.19213    0.34500  -0.557   0.5778
famrel        0.21012    0.14317   1.468   0.1428
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 3.153 on 534 degrees of freedom
Multiple R-squared:  0.09471,   Adjusted R-squared:  0.06589
F-statistic: 3.286 on 17 and 534 DF,  p-value: 1.04e-05
```

가족 관련 변수들로 구성했다.

관련이 있는건 아버지의 교육수준정도 그 외에는 관련이 없어서

G3를 설명할 수 있는 요인 중에 가족요인은 아닌라고 분석했다.

p value도 높기 때문에 변수 개개인의 뒷받침할 근거도 거의 없다.

5-1. modeling 결과 – 외부요인3

```
Call:
lm(formula = G3 ~ school + reason + schoolsup + paid + activities +
    absences, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-12.762  -1.533   0.034   1.838   8.281

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     12.76165    0.30090  42.411  < 2e-16 ***
schoolMS        -1.90141    0.29188  -6.514 1.67e-10 ***
reasonhome       0.30216    0.34269   0.882 0.378309
reasonother     -0.32517    0.44614  -0.729 0.466403
reasonreputation  0.90265    0.35246   2.561 0.010705 *
schoolsupyes    -1.14129    0.42968  -2.656 0.008137 **
paidyes         -0.32336    0.59035  -0.548 0.584098
activitiesyes    0.13397    0.27026   0.496 0.620287
absences        -0.10011    0.02876  -3.480 0.000541 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 3.09 on 543 degrees of freedom
Multiple R-squared:  0.1159,    Adjusted R-squared:  0.1029
F-statistic:   8.9 on 8 and 543 DF,  p-value: 1.713e-11
```

학교가 MS보다 GP에서의 성적이 좋았어서 이렇게 나온것같다.

또, 명성 나왔다. 장학금을 덜 받을 수록 G3가 컸다.? 이렇게 해석하면 안되고 그냥 장학금을 받은 아이들이 거의 없었다. eda에서 장학금보다 부모지원금이 더 많았어서 이렇게 나온것 같다.

그외로 결석도 큰 영향을 끼쳤다. eda 한 것 처럼 결석이 없을수록 성적이 높았다.

5-1. modeling 결과 – 내부요인

```
Call:
lm(formula = G3 ~ sex + age + studytime + failures + higher +
    romantic + freetime + goout + Dalc + health + absences, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9605  -1.4417   0.0883   1.5841   6.4033

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.326063   1.977401   3.199  0.00146 **
sexM        -0.264861   0.264994  -0.999  0.31800
age          0.258174   0.110081   2.345  0.01937 *
studytime    0.699797   0.157321   4.448 1.05e-05 ***
failures    -1.813560   0.245851  -7.377 6.15e-13 ***
higheryes    2.184916   0.427852   5.107 4.55e-07 ***
romanticyes -0.397557   0.253237  -1.570  0.11702
freetime    -0.095536   0.122950  -0.777  0.43748
goout       -0.102252   0.111454  -0.917  0.35932
Dalc        -0.323295   0.148502  -2.177  0.02991 *
health      -0.099712   0.084096  -1.186  0.23627
absences     0.008282   0.026302   0.315  0.75297
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 2.808 on 540 degrees of freedom
Multiple R-squared:  0.2738,    Adjusted R-squared:  0.259
F-statistic:  18.5 on 11 and 540 DF,  p-value: < 2.2e-16
```

저번과 같이 자기 자신에 관련된 변수들로 모델을 했더니 failures가 중요한 영향을 미쳤다. 확실히 실패한 경험이 많을 수록 열심히 하는 학생이 아니라는 것을 직관적으로 알 수 있다.

또한 studytime도 많을 수록, 실패 덜하고 수준높은 교육을 원할 수록 술도 적게 먹을 수록 성적이 높았다.

직관적으로 알 수 있지만 p value도 뒷받침해준다.

신기했던건 나이가 높을 수록 성적이 높다.

5-1. modeling 결과 – 성적

```
Call:
lm(formula = G3 ~ G1 + G2 + G3, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5064 -0.4306 -0.0658  0.6700  2.6783

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23170    0.23603  -0.982 0.326711
G1           0.14263    0.03873   3.683 0.000253 ***
G2           0.90768    0.03667  24.755  < 2e-16 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 1.286 on 549 degrees of freedom
Multiple R-squared:  0.8452,    Adjusted R-squared:  0.8446
F-statistic:  1498 on 2 and 549 DF,  p-value: < 2.2e-16
```

종속변수가 G1,G2,G3로 만들었기 때문에 많은 관련이 있다.

신기한건 보통 G1과 G2가 높을 수록 G3도 높은 성적을 받는다.

꾸준히 잘하는 학생들은 꾸준히 높은 성적을 받는 다는 뜻이다.

5-2. 최종 modeling 결과

```
Call:
lm(formula = G3 ~ G1 + schoolsup + absences + failures + studytime +
    higher + Dalc + G2 + school + reason - 1, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8639 -0.4644  0.0307  0.6109  3.0240

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
G1              0.11466    0.03949   2.903  0.00384 **
schoolsupno     0.21986    0.36155   0.608  0.54338
schoolsupyes   -0.01759    0.38661  -0.046  0.96372
absences        0.02512    0.01225   2.050  0.04083 *
failures       -0.29565    0.11152  -2.651  0.00826 **
studytime       0.13401    0.07214   1.858  0.06376 .
higheryes       0.17994    0.19609   0.918  0.35922
Dalc           -0.12959    0.06336  -2.045  0.04132 *
G2              0.88635    0.03654  24.255  < 2e-16 ***
schoolMS       -0.10313    0.12486  -0.826  0.40918
reasonhome     -0.04322    0.14184  -0.305  0.76067
reasonother    -0.31411    0.18329  -1.714  0.08715 .
reasonreputation -0.15785    0.14611  -1.080  0.28044
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 1.264 on 539 degrees of freedom
Multiple R-squared:  0.9897,    Adjusted R-squared:  0.9895
F-statistic:  3986 on 13 and 539 DF,  p-value: < 2.2e-16
```

```
rmse(predict(reg,valid),valid$G3)
```

1.16189516973206

rmse는 코드를 통해서 구현했고, 검증 데이터 셋과 rmse한 결과 차이는 1.16으로 근소한 차이를 보여줬다.

6. 한계점

- 위 데이터 셋의 이상치? age, absences 등 수치에서의 이상치라고 할 것들을 제거 안하고 했다.
  데이터의 수도 적었을 뿐더러 age같은 경우도 18세 넘어도 입학 가능하다고 홈페이지에 써있고, absence도 많이 한것이 말이 되기 때문이다. 잡음에 많이
- 틀린 경우를 봤다. 결석과 goout하고 Dalc와 G3를 봤을 때는 이해를 했지만 전체적인 체크를 봤을 때 그냥 외향적인 아이다.
  higher한 교육도 받고 싶어하고 가족관계도 좋았다.
  다른 학생은 결석도 없고 성적도 엄청 잘받았지만 comp기준에 틀린 이유는 G3이 G1,G2보다 작다.
  점수는 잘받았는데 노력을 안했다고 단정지어버려서 틀린 결과가 나온것 같다.