Stats 101A Final Project

Hashim Bhat (705372204)

2023-03-18

Introduction

My research question for this project is: What lifestyle factors make the best linear model for predicting a students grades according to the dataset?

The dataset is from kaggle and it is called "Student Alcohol Consumption" by UCI Machine Learning. Source: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption The data wwas obtained in a survey of students' math and portuguese courses in secondary school. It contains information relating to the social lives, study habits, and family backgrounds of 395 students. Since these variables can all impact a students final grade, I thought it was justified to do multiple regression analysis. The students final grade (The G3 column in the dataset) is the response variable and the lifestyle/social/background factors are the explanatory variables.

Structure

First, I will display the summary statistics of the data such as mean, sd, plots, etc.

Then I will conduct multiple regression analysis and interpret the results as I go along. First I will make a full linear model that includes all of the explanatory variables, and I will reduce it to the ones that show significance. I will then make diagnostic plots and perform any transformations if appropriate. I will check for multicollinearity issues and perform relevant tests for this. I will then use variable selection to select my final model.

Finally, I will state the conclusion of my findings, as well as the limitations of my model and how it could be improved.

Reading the data + Setup

```
student <- read.csv("student-mat.csv")
head(student)
```

```
##   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
```

```
# I didn't include every column because the linear model summary would be too long but I tried to inclu

newstudent <- student[ ,c(7,8,13,14,15,16,17,18,19,25,26,27,28,33)]

newstudent$schoolsup = factor(newstudent$schoolsup,levels = c('no', 'yes'),labels = c(0, 1))
newstudent$schoolsup <- as.numeric(as.character(newstudent$schoolsup))

newstudent$famsup = factor(newstudent$famsup,levels = c('no', 'yes'),labels = c(0, 1))
newstudent$famsup <- as.numeric(as.character(newstudent$famsup))

newstudent$paid = factor(newstudent$paid,levels = c('no', 'yes'),labels = c(0, 1))
newstudent$paid <- as.numeric(as.character(newstudent$paid))

newstudent$activities = factor(newstudent$activities,levels = c('no', 'yes'),labels = c(0, 1))
newstudent$activities <- as.numeric(as.character(newstudent$activities))



# Scale final grade so that it is out of 100
newstudent$G3 <- newstudent$G3 * 5
head(newstudent)
```

```
##   Medu Fedu traveltime studytime failures schoolsup famsup paid activities
## 1    4    4          2         2        0         1      0    0          0
## 2    1    1          1         2        0         0      1    0          0
## 3    1    1          1         2        3         1      0    1          0
## 4    4    2          1         3        0         0      1    1          1
## 5    3    3          1         2        0         0      1    1          0
## 6    4    3          1         2        0         0      1    1          1
##   freetime goout Dalc Walc G3
## 1        3     4    1    1 30
## 2        3     3    1    1 30
## 3        3     2    2    3 50
## 4        2     2    1    1 75
## 5        3     2    1    2 50
## 6        4     2    1    2 75
```

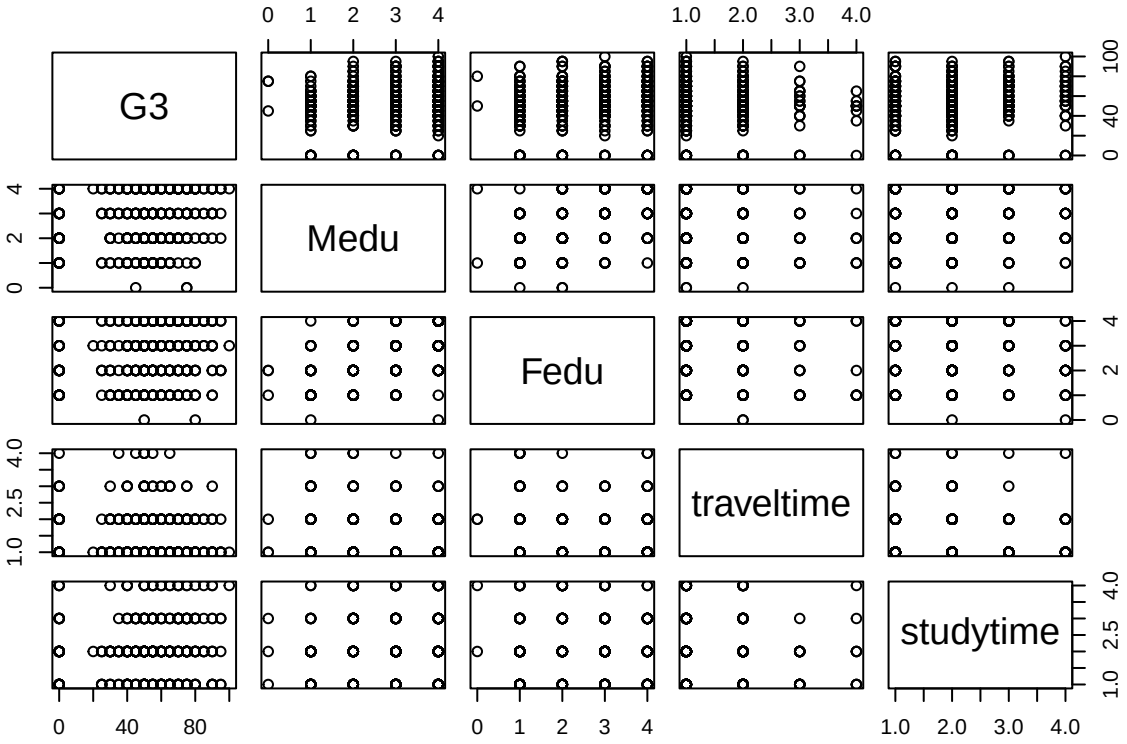
Data Description

Plots

We will be focusing on the first row of these diagrams to examine the effect of the explanatory variables on

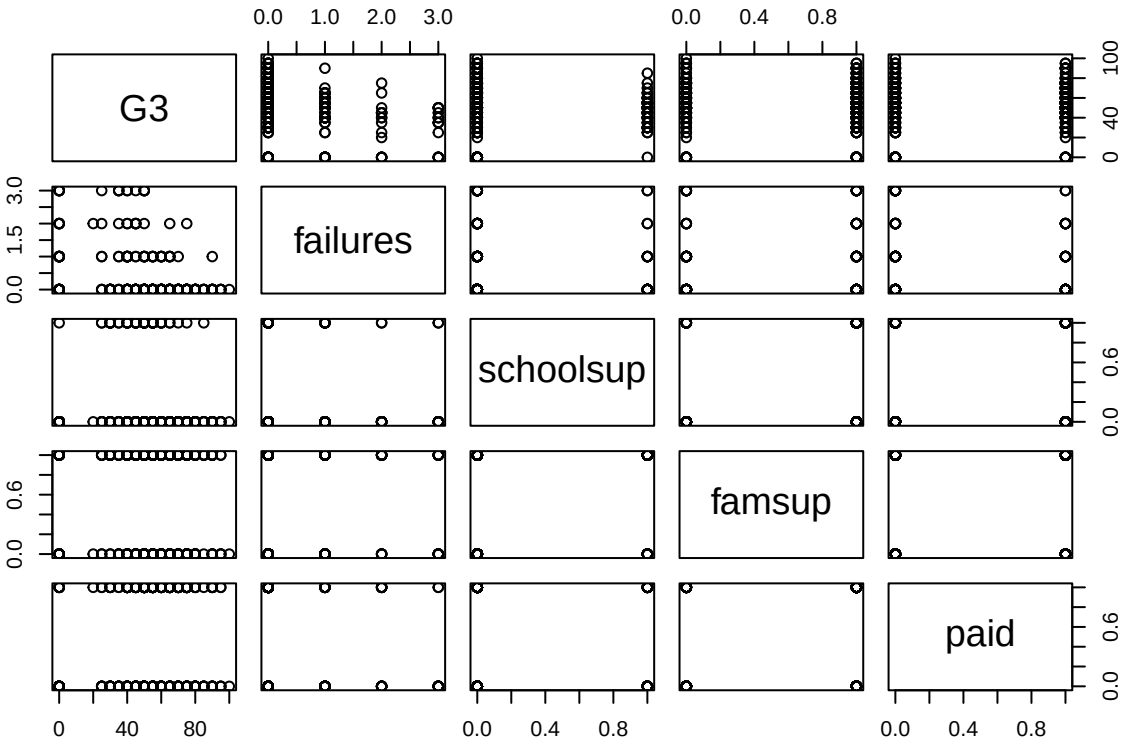
the response variable. For a lot of the explanatory variables, the data was grouped into ranges/brackets and then expressed in terms of those brackets, which makes our plots a lot harder to analyze. Nevertheless, there are still some clear trends.

```
pairs(newstudent[,c(14,1:4)])
```



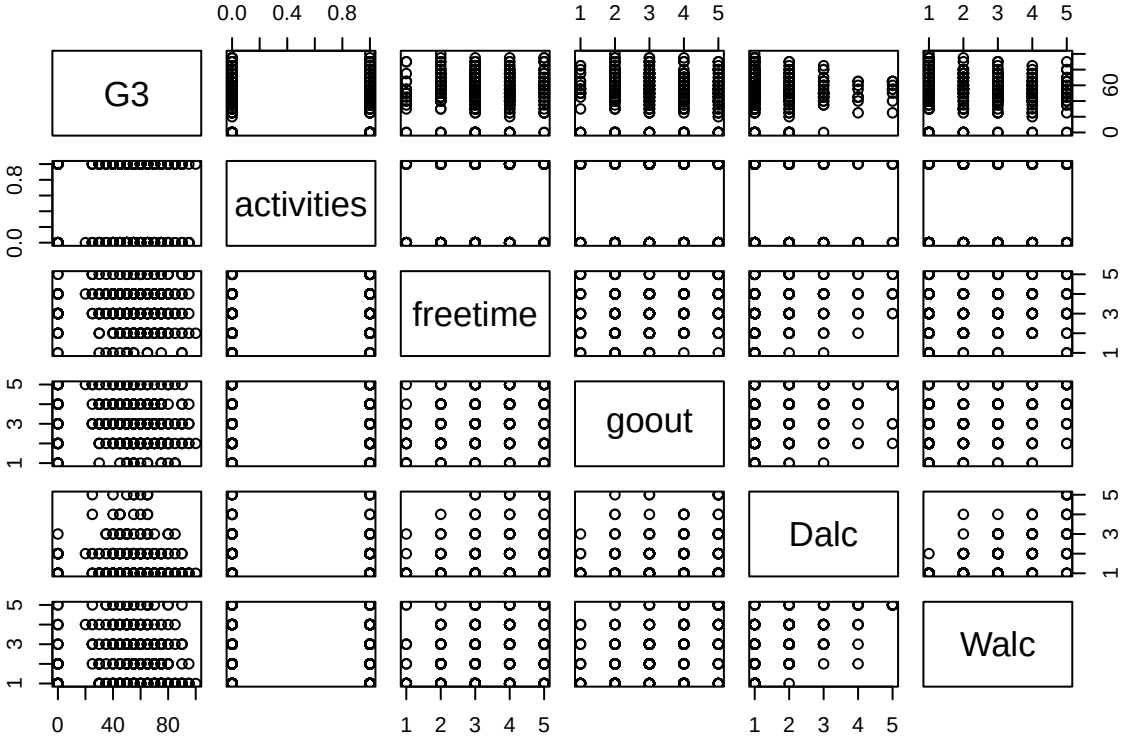
Above, we can see upward trends in Mothers Education, Fathers Education, and downward trend in travel time.

```
pairs(newstudent[,c(14,5:8)])
```



Above, we can see downward trend in failures.

```
pairs(newstudent[,c(14,9:13)])
```



Above, we can see a downward trend in Workday Alcohol Consumption (Dalc)

Correlations

```
cor(newstudent)
```

```
##                   Medu         Fedu   traveltime    studytime      failures
## Medu        1.00000000  0.623455112 -0.171639305  0.064944137 -0.2366799626
## Fedu        0.62345511  1.000000000 -0.158194054 -0.009174639 -0.2504084445
## traveltime -0.17163930 -0.158194054  1.000000000 -0.100909119  0.0922387462
## studytime   0.06494414 -0.009174639 -0.100909119  1.000000000 -0.1735630314
## failures   -0.23667996 -0.250408444  0.092238746 -0.173563031  1.0000000000
## schoolsup  -0.03602948  0.037529649 -0.009246380  0.037762698 -0.0004374907
## famsup      0.18372702  0.185496109 -0.003286261  0.145227617 -0.0550746206
## paid        0.15970038  0.086981416 -0.066420239  0.167219880 -0.1880389659
## activities  0.10827676  0.112642791 -0.007766399  0.089877272 -0.0693405255
## freetime    0.03089087 -0.012845528 -0.017024944 -0.143198407  0.0919874710
## goout       0.06409444  0.043104668  0.028539674 -0.063903675  0.1245609219
## Dalc        0.01983410  0.002386429  0.138325309 -0.196019263  0.1360469312
## Walc       -0.04712346 -0.012631018  0.134115752 -0.253784731  0.1419620300
## G3          0.21714750  0.152456939 -0.117142053  0.097819690 -0.3604149405
##                schoolsup       famsup        paid   activities    freetime
## Medu       -0.0360294775  0.183727022  0.15970038  0.108276762  0.03089087
## Fedu        0.0375296494  0.185496109  0.08698142  0.112642791 -0.01284553
## traveltime -0.0092463805 -0.003286261 -0.06642024 -0.007766399 -0.01702494
## studytime   0.0377626975  0.145227617  0.16721988  0.089877272 -0.14319841
## failures   -0.0004374907 -0.055074621 -0.18803897 -0.069340525  0.09198747
## schoolsup   1.0000000000  0.104681061 -0.02075328  0.046032365 -0.04546543
## famsup      0.1046810614  1.000000000  0.29318434 -0.001500108  0.01053759
## paid       -0.0207532817  0.293184339  1.00000000 -0.021382376 -0.06425287
## activities  0.0460323645 -0.001500108 -0.02138238  1.000000000  0.08972816
## freetime   -0.0454654257  0.010537588 -0.06425287  0.089728164  1.00000000
## goout      -0.0376984912 -0.015631443  0.01049327  0.046087686  0.28501871
## Dalc       -0.0214851000 -0.031575204  0.06246536 -0.066508094  0.20900085
## Walc       -0.0871517384 -0.086687935  0.06045364 -0.037476696  0.14782181
## G3         -0.0827882150 -0.039157145  0.10199624  0.016099701  0.01130724
##                  goout         Dalc        Walc          G3
## Medu        0.06409444  0.019834099 -0.04712346  0.21714750
## Fedu        0.04310467  0.002386429 -0.01263102  0.15245694
## traveltime  0.02853967  0.138325309  0.13411575 -0.11714205
## studytime  -0.06390368 -0.196019263 -0.25378473  0.09781969
## failures    0.12456092  0.136046931  0.14196203 -0.36041494
## schoolsup  -0.03769849 -0.021485100 -0.08715174 -0.08278821
## famsup     -0.01563144 -0.031575204 -0.08668793 -0.03915715
## paid        0.01049327  0.062465362  0.06045364  0.10199624
## activities  0.04608769 -0.066508094 -0.03747670  0.01609970
## freetime    0.28501871  0.209000848  0.14782181  0.01130724
## goout       1.00000000  0.266993848  0.42038575 -0.13279147
## Dalc        0.26699385  1.000000000  0.64754423 -0.05466004
## Walc        0.42038575  0.647544230  1.00000000 -0.05193932
## G3         -0.13279147 -0.054660041 -0.05193932  1.00000000
```

Nothing notable about the correlations of the explanatory variables on G3 (The Response Variable). In regards to the correlations of the explanatory variables with each other, Workday Alcohol Consumption (Dalc) seems to be moderately positively correlated with Weekend Alcohol Consumption (Walc). The correlation is 0.647 so there is a correlation but it might not be high enough to cause a multicolinearity issue but we can check this later on in the report.

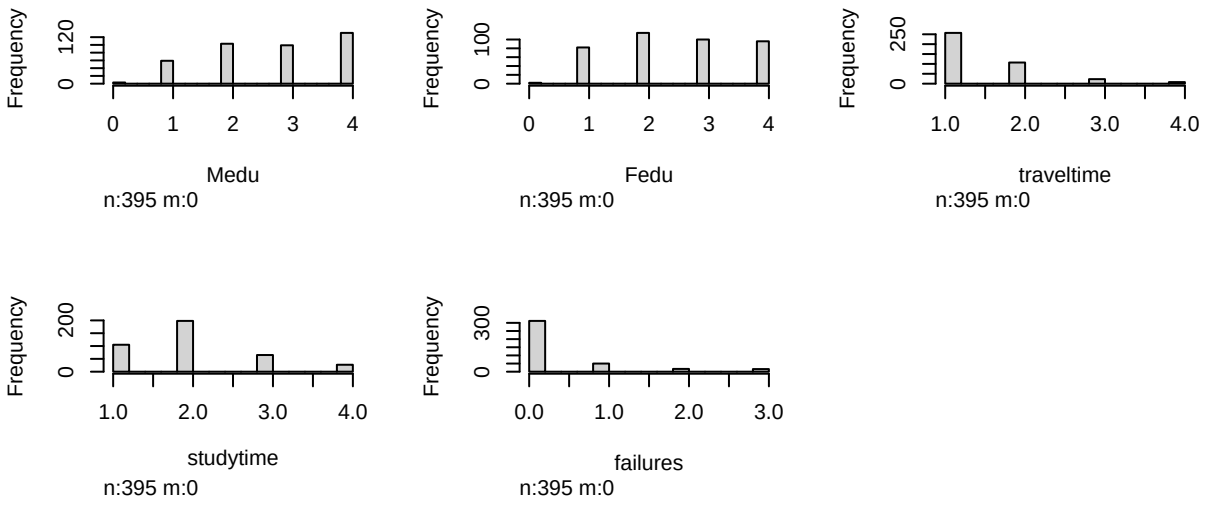
Mean, Standard Deviation, Quartiles

```
summary(newstudent)
```

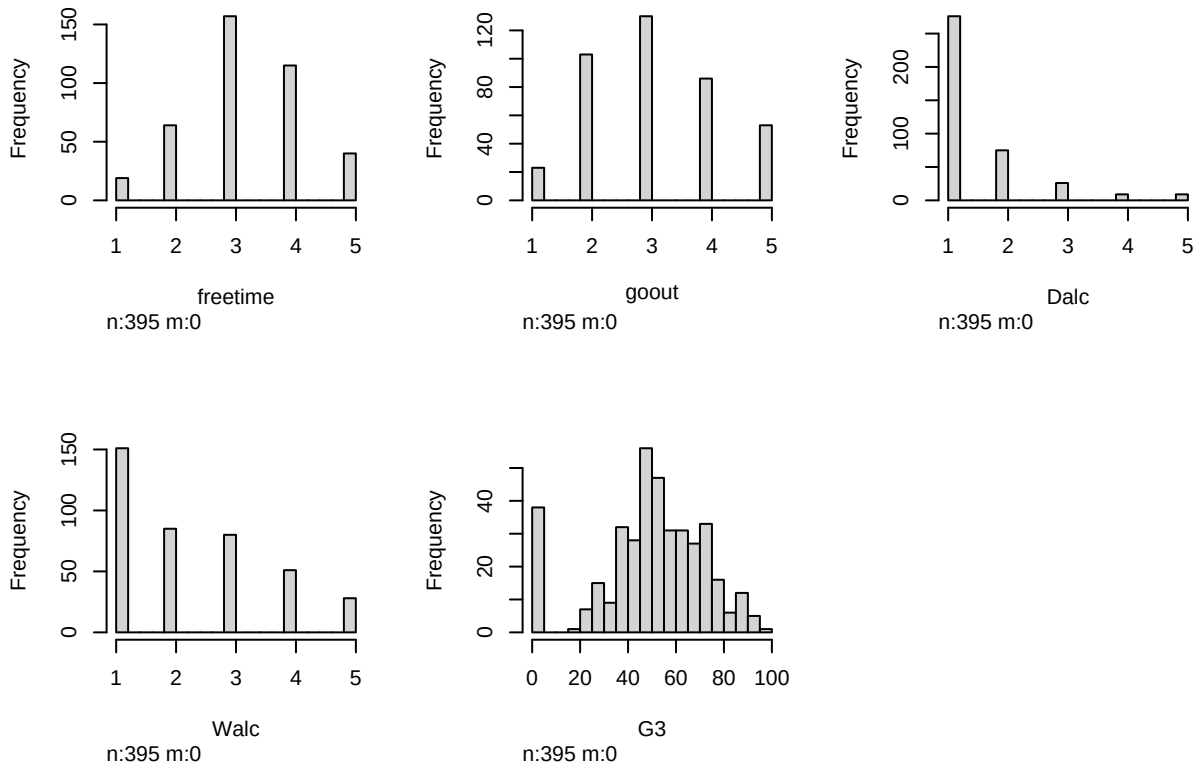
```
##       Medu            Fedu         traveltime      studytime
##  Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median :2.000   Median :1.000   Median :2.000
##  Mean   :2.749   Mean   :2.522   Mean   :1.448   Mean   :2.035
##  3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
##     failures        schoolsup          famsup            paid
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :1.0000   Median :0.0000
##  Mean   :0.3342   Mean   :0.1291   Mean   :0.6127   Mean   :0.4582
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :3.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##    activities        freetime         goout            Dalc
##  Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :1.0000   Median :3.000   Median :3.000   Median :1.000
##  Mean   :0.5089   Mean   :3.235   Mean   :3.109   Mean   :1.481
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000
##  Max.   :1.0000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##       Walc             G3
##  Min.   :1.000   Min.   :  0.00
##  1st Qu.:1.000   1st Qu.: 40.00
##  Median :2.000   Median : 55.00
##  Mean   :2.291   Mean   : 52.08
##  3rd Qu.:3.000   3rd Qu.: 70.00
##  Max.   :5.000   Max.   :100.00
```

Distributions of Variables

```
hist.data.frame(newstudent[,1:8])
```



```
hist.data.frame(newstudent[,9:14])
```



Results and Interpretation

Finding Our Reduced Model

```
regmod <- lm(G3~., data = newstudent)
summary(regmod)
```

```
##
## Call:
## lm(formula = G3 ~ ., data = newstudent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.901 -10.134   1.604  14.233  48.662
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 50.48549    6.81055   7.413 8.07e-13 ***
## Medu         3.25151    1.27719   2.546  0.01129 *
## Fedu        -0.02058    1.28226  -0.016  0.98720
## traveltime  -1.82163    1.57417  -1.157  0.24791
## studytime    1.53713    1.36699   1.124  0.26152
## failures    -9.36790    1.54312  -6.071 3.08e-09 ***
## schoolsup   -4.38472    3.20818  -1.367  0.17252
## famsup      -4.44321    2.35691  -1.885  0.06017 .
## paid         1.70827    2.31938   0.737  0.46187
## activities  -1.07049    2.16697  -0.494  0.62159
## freetime     1.88585    1.14342   1.649  0.09991 .
```

```
## goout       -3.06813    1.09593  -2.800  0.00538 **
## Dalc        -0.80441    1.60139  -0.502  0.61573
## Walc         1.31849    1.18663   1.111  0.26722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.01 on 381 degrees of freedom
## Multiple R-squared:  0.1862, Adjusted R-squared:  0.1585
## F-statistic: 6.707 on 13 and 381 DF,  p-value: 1.295e-11
```

From the Linear Model above, the high Overall F statistic and small p-value of this linear model indicates that at least one of our explanatory variables is statistically significant Medu, failures, famsup, freetime and goout are significant according to a significance level of 0.1

```
anova(regmod)
```

```
## Analysis of Variance Table
##
## Response: G3
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Medu         1   9749  9748.8 22.0769 3.664e-06 ***
## Fedu         1     99    98.6  0.2233  0.636801
## traveltime   1   1317  1316.6  2.9814  0.085035 .
## studytime    1   1263  1263.1  2.8604  0.091604 .
## failures     1  19344 19344.3 43.8068 1.233e-10 ***
## schoolsup    1   1273  1272.8  2.8823  0.090377 .
## famsup       1   1241  1240.6  2.8093  0.094537 .
## paid         1    222   222.5  0.5038  0.478282
## activities   1     88    88.1  0.1995  0.655338
## freetime     1    389   388.9  0.8807  0.348595
## goout        1   2960  2959.8  6.7027  0.009996 **
## Dalc         1     16    15.7  0.0355  0.850630
## Walc         1    545   545.2  1.2346  0.267216
## Residuals  381 168243   441.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the Anova table above, Medu, travel time, study time, failures, schoolsup, famsup, and goout are significant according to a significance level of 0.1

```
# Make Reduced with all variables that were shown to be significant in either of the tables
redmod <- lm(G3~Medu+goout+failures+studytime+freetime+schoolsup+traveltime+famsup, data = newstudent)
summary(redmod)
```

```
##
## Call:
## lm(formula = G3 ~ Medu + goout + failures + studytime + freetime +
##     schoolsup + traveltime + famsup, data = newstudent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.012  -9.332   1.919  13.685  47.906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   52.134      6.450   8.083 8.22e-15 ***
```

```
## Medu           3.158      1.028   3.072  0.00227 **
## goout         -2.567      0.999  -2.569  0.01057 *
## failures      -9.474      1.496  -6.332 6.73e-10 ***
## studytime      1.283      1.306   0.982  0.32672
## freetime       1.695      1.115   1.521  0.12919
## schoolsup     -4.919      3.169  -1.552  0.12147
## traveltime    -1.782      1.546  -1.152  0.24989
## famsup        -4.039      2.238  -1.805  0.07193 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.94 on 386 degrees of freedom
## Multiple R-squared:  0.1813, Adjusted R-squared:  0.1643
## F-statistic: 10.69 on 8 and 386 DF,  p-value: 1.349e-13
```

```
# Anova of reduced model
anova(redmod)
```

```
## Analysis of Variance Table
##
## Response: G3
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Medu         1   9749  9748.8 22.2319 3.381e-06 ***
## goout        1   4468  4468.3 10.1899  0.001528 **
## failures     1  18608 18607.6 42.4343 2.283e-10 ***
## studytime    1    159   159.4  0.3635  0.546903
## freetime     1   1067  1066.6  2.4325  0.119665
## schoolsup    1   1338  1337.5  3.0502  0.081521 .
## traveltime   1    669   669.1  1.5259  0.217478
## famsup       1   1428  1427.9  3.2563  0.071927 .
## Residuals  386 169262   438.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Only difference between the Anova and summary tables of reduced model is the significance of school up. We will include it in the new reduced model, and omit the variables that were shown to be insignificant in both tables.

```
# New reduced model
redmod2 <- lm(G3~Medu+goout+failures+schoolsup+famsup, data = newstudent)
summary(redmod2)
```

```
##
## Call:
## lm(formula = G3 ~ Medu + goout + failures + schoolsup + famsup,
##     data = newstudent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.190  -9.373   2.273  13.810  42.904
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  55.8423     4.1974  13.304  < 2e-16 ***
## Medu          3.3923     1.0162   3.338 0.000924 ***
## goout        -2.2262     0.9624  -2.313 0.021234 *
```

```
## failures     -9.6422     1.4786  -6.521 2.17e-10 ***
## schoolsup    -4.9643     3.1727  -1.565 0.118473
## famsup       -3.7692     2.2199  -1.698 0.090318 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.98 on 389 degrees of freedom
## Multiple R-squared:  0.1716, Adjusted R-squared:  0.1609
## F-statistic: 16.11 on 5 and 389 DF,  p-value: 1.923e-14
```

```
anova(redmod2)
```

```
## Analysis of Variance Table
##
## Response: G3
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Medu        1   9749  9748.8 22.1409 3.527e-06 ***
## goout       1   4468  4468.3 10.1482  0.001561 **
## failures    1  18608 18607.6 42.2606 2.453e-10 ***
## schoolsup   1   1374  1374.3  3.1213  0.078061 .
## famsup      1   1269  1269.4  2.8830  0.090318 .
## Residuals 389 171279   440.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Schoolsup found to be insignificant at a significance level of 0.1 in the summary table but significant in the ANOVA table. I am going to choose to keep it in the model because it is only insignificant by a small margin in the summary table.

```
anova(redmod2, regmod)
```

```
## Analysis of Variance Table
##
## Model 1: G3 ~ Medu + goout + failures + schoolsup + famsup
## Model 2: G3 ~ Medu + Fedu + traveltime + studytime + failures + schoolsup +
##     famsup + paid + activities + freetime + goout + Dalc + Walc
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    389 171279
## 2    381 168243  8    3036.4 0.8595 0.5509
```

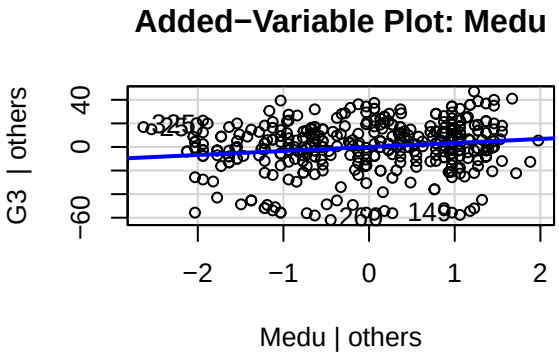
High p value tells us that we should not reject the null hypothesis therefore the reduced model is a better fit

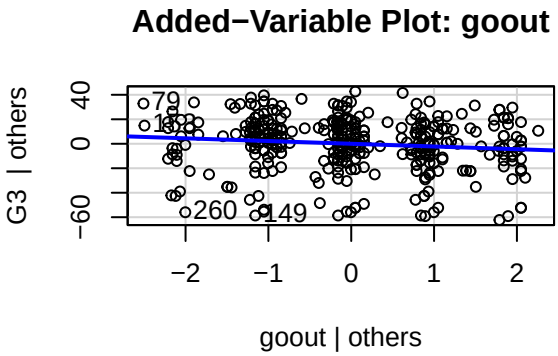
It is odd that Educational Support variables have negative coefficients so we will do added variable plots to observe the true effect

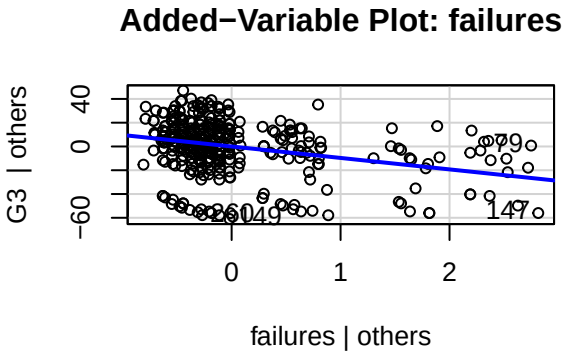
```
library(car)
```

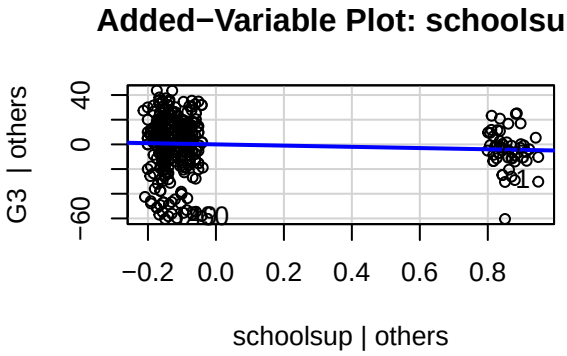
```
## Loading required package: carData
```

```
par(mfrow=c(2,2))
avPlot(redmod2,variable="Medu",ask=FALSE)
avPlot(redmod2,variable="goout",ask=FALSE)
avPlot(redmod2,variable="failures",ask=FALSE)
avPlot(redmod2,variable="schoolsup",ask=FALSE)
```

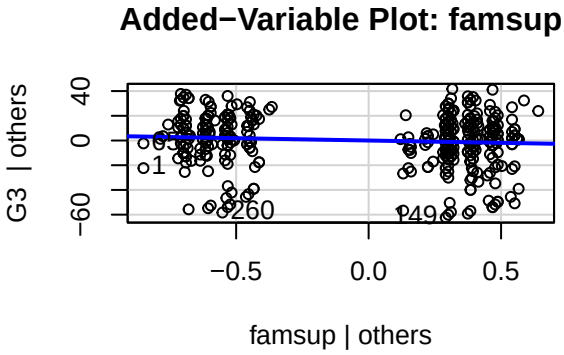









```
avPlot(redmod2,variable="famsup",ask=FALSE)
```



The educational support variables seem to have no effect as the lines on their added variable plots have almost 0 slope. I will remove them from the model.

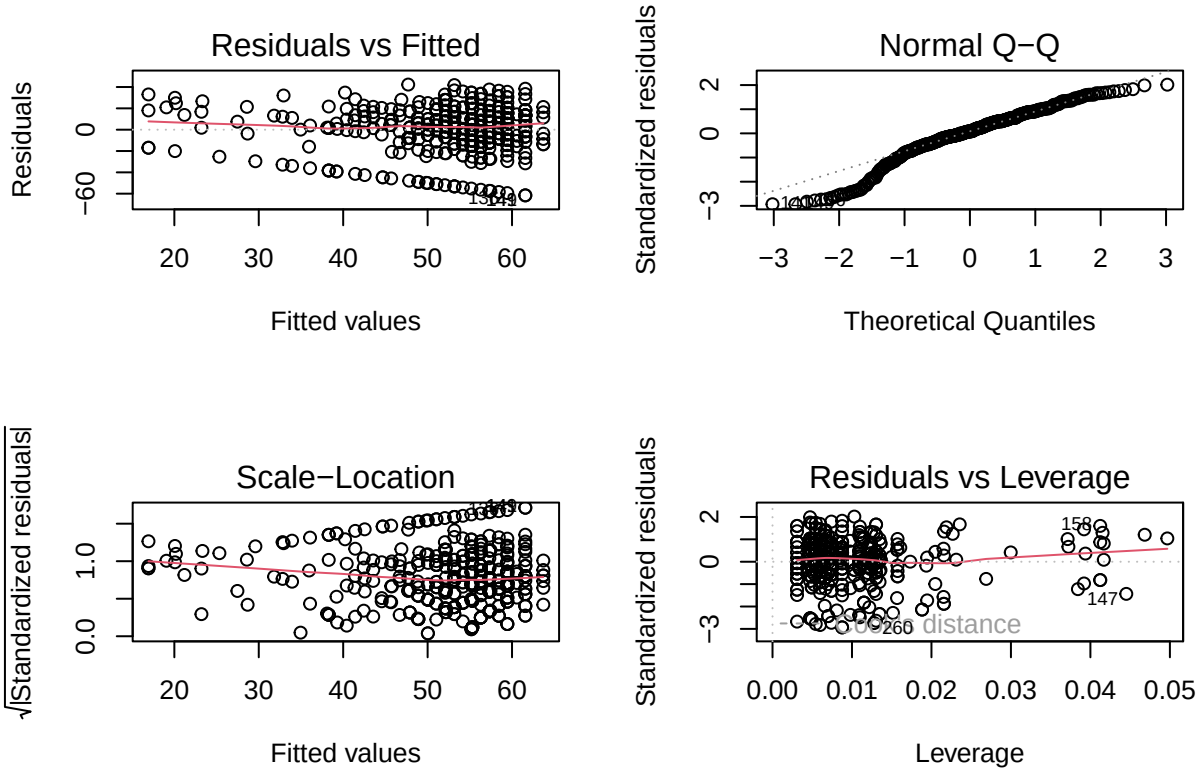
```
# New Reduced Model
redmod3 <- lm(G3~Medu + goout + failures, data = newstudent)
summary(redmod3)
```

```
##
## Call:
## lm(formula = G3 ~ Medu + goout + failures, data = newstudent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.574  -9.813   1.711  13.426  42.299
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.2854     4.0808  13.057  < 2e-16 ***
## Medu          3.1373     1.0037   3.126  0.00191 **
## goout        -2.1304     0.9665  -2.204  0.02809 *
## failures     -9.6118     1.4861  -6.468 2.97e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.09 on 391 degrees of freedom
## Multiple R-squared:  0.1588, Adjusted R-squared:  0.1523
## F-statistic:  24.6 on 3 and 391 DF,  p-value: 1.337e-14
```

Diagnostic Plots and Transformations

```
par(mfrow=c(2,2))
plot(redmod3)
```



Normal QQ Plot is left skewed, Residual vs fitted and standardized residual plots are not too bad but may show a slight downward trend which could indicate a problem, there are a few leverage points in the Residuals vs Leverage plot which could also indicate a problem with our model. Overall, we should try some transformations to see if the diagnostic plots improve.

```
#powertransformation to multicollinearity doesn't work when there are values of 0 so I replaced them wi
G3new <- replace(newstudent$G3, newstudent$G3 == 0, 1e-8)
failuresnew <- replace(newstudent$failures, newstudent$failures == 0, 1e-8)
Medunew <- replace(newstudent$Medu, newstudent$Medu == 0, 1e-8)
```

```
attach(newstudent)
summary(powerTransform(cbind(G3new,Medunew,failuresnew,goout)~1))
```

```
## bcPower Transformations to Multinormality
##             Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
```

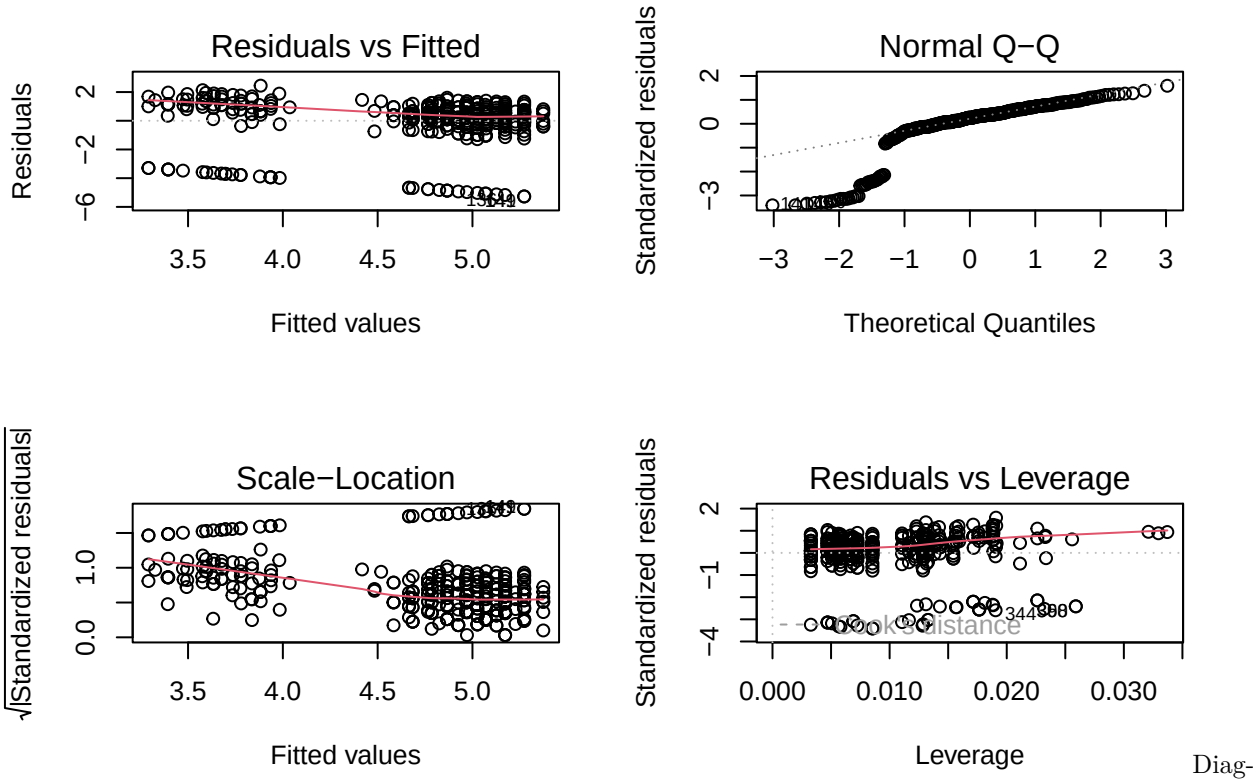
```
## G3new          0.4100        0.41       0.3715       0.4484
## Medunew        0.7502        0.75       0.6118       0.8885
## failuresnew   -0.2408       -0.24      -0.2678      -0.2137
## goout          0.7793        1.00       0.5479       1.0107
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                   LRT df       pval
## LR test, lambda = (0 0 0 0) 2527.442  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                   LRT df       pval
## LR test, lambda = (1 1 1 1) 10548.61  4 < 2.22e-16
```

Indicates we should change powers of G3 to 0.41, failures to -0.24, Medu to 0.75

```
transmod <- lm(I(G3^0.41)~+I(failuresnew^-0.24)+goout+I(Medu^0.75))
summary(transmod)
```

```
##
## Call:
## lm(formula = I(G3^0.41) ~ +I(failuresnew^-0.24) + goout + I(Medu^0.75))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2728 -0.1997  0.3887  0.8571  2.4447
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.516530   0.350270  10.039  < 2e-16 ***
## I(failuresnew^-0.24)  0.014465   0.002417   5.985 4.89e-09 ***
## goout                -0.100640   0.070897  -1.420   0.1565
## I(Medu^0.75)          0.266729   0.121205   2.201   0.0283 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.552 on 391 degrees of freedom
## Multiple R-squared:  0.121,  Adjusted R-squared:  0.1143
## F-statistic: 17.95 on 3 and 391 DF,  p-value: 6.214e-11
```

```
par(mfrow=c(2,2))
plot(transmod)
```



Diagnostic plot look much worse than original. Also the goout variable is no longer significant according to the summary table. So the original model seems better.

```
vif(redmod3)
```

```
##     Medu    goout failures
## 1.069418 1.025422 1.081809
```

No issue with multicolinearity in our reduced model

```
backAIC <- step(redmod3, direction="backward", data=newstudent)
```

```
## Start:  AIC=2412.56
## G3 ~ Medu + goout + failures
##
##             Df Sum of Sq    RSS    AIC
## <none>                   173923 2412.6
## - goout      1    2161.3 176084 2415.4
## - Medu       1    4345.9 178269 2420.3
## - failures   1   18607.6 192531 2450.7
```

This backwards AIC test confirms that there is no multicolinearity because the model stayed the same

Discussion

My Final Model:

```
summary(redmod3)
```

```
##
## Call:
## lm(formula = G3 ~ Medu + goout + failures, data = newstudent)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.574  -9.813   1.711  13.426  42.299
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.2854     4.0808  13.057  < 2e-16 ***
## Medu          3.1373     1.0037   3.126  0.00191 **
## goout        -2.1304     0.9665  -2.204  0.02809 *
## failures     -9.6118     1.4861  -6.468 2.97e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.09 on 391 degrees of freedom
## Multiple R-squared:  0.1588, Adjusted R-squared:  0.1523
## F-statistic:  24.6 on 3 and 391 DF,  p-value: 1.337e-14
```

Overall, my conclusion is that the lifestyle factors from our data that best predict final grades are Frequency of going out with friends which has a negative effect, Mothers Education which has a positive effect, and number of past class failures which has a negative effect.

Holding goout and failures constant, increasing Mothers Education by 1 unit, increases final grades by 3.1373 points (grade is out of 100).

Holding Medu and Failures constant, increasing goout by 1 unit decreases final grade by 2.1304.

Holding Medu and goout constant, increases the number of past classes failed by 1 decreases the final grade by 9.6118.

All of these inferences make sense contextually. Mothers that are more educated could be more involved in their child's school life and could tutor them, therefore the student achieves better grades. Students that have failed a lot of past classes have a record of being unsuccessful when it comes to exams so you would expect them to have lower grades. The most surprising/questionable find is that students who go out more get lower grades. This could make sense however, because students could be spending less time studying if they spend more time going out and therefore get lower grades, and kids who spend a lot of time studying could have less time to go out.

The most surprising find in the analysis was that study time was not found to have a statistically significantly postive effect on final grades. This indicates a clear problem with either the data or the analysis because it is established in the real world that studying more is correlated with getting better grades.

One big limitation of the analysis is the grouping of data for the explanatory variables into categories, rather than just presenting the raw data. If the data was raw, we would have a greater range of numbers that represent more individual observations, therefore we would draw better inferences from our analysis. So in the future, I would pick a dataset that had raw observations to improve my model. Another limitation is that I may not have included every relevant explanatory variable that contributes to a students final grade, therefore the analysis could suffer from an omission bias issue which could make the model inaccurate.