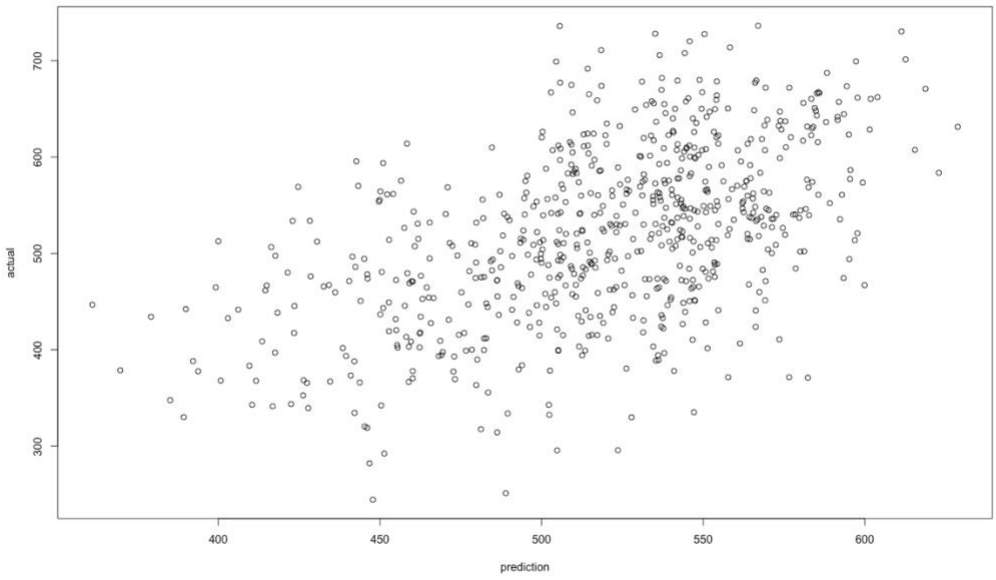
Ximan Liu
1935858
DSC 423
HW7

I have completed this work independently. The solutions given are entirely my own work.

1)
a)
 # R Code:

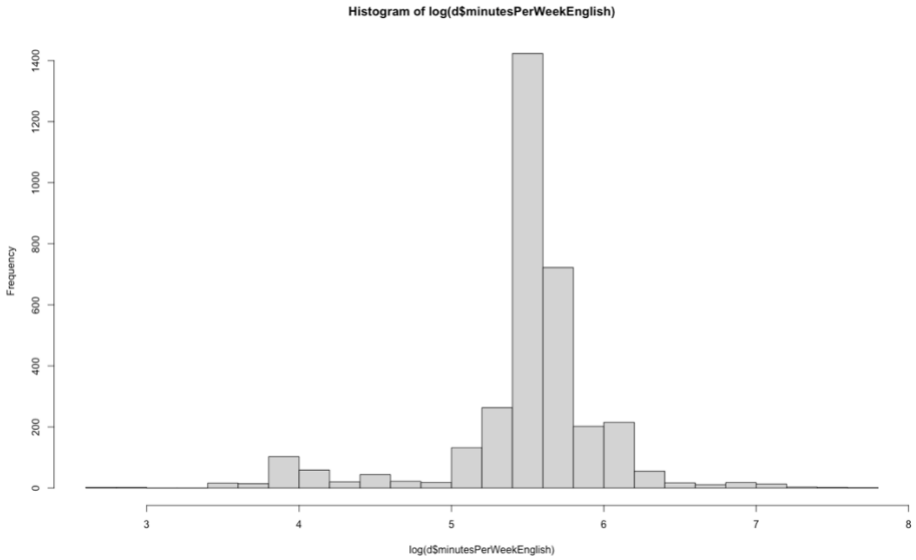
```
d <- Pisa2009[,-c(1,4)]
partition <- sample(2, nrow(d), replace = TRUE, prob = c(0.80, 0.20))
train <- d[partition==1 ,]
test <- d[partition==2 ,]
model <- lm(readingScore ~ ., data = train)
prediction <- predict(model, test)
actual = test$readingScore
# Cross-validation
cor(prediction,actual)
plot(prediction,actual)
```

```
> cor(prediction,actual)
[1] 0.5253573
```

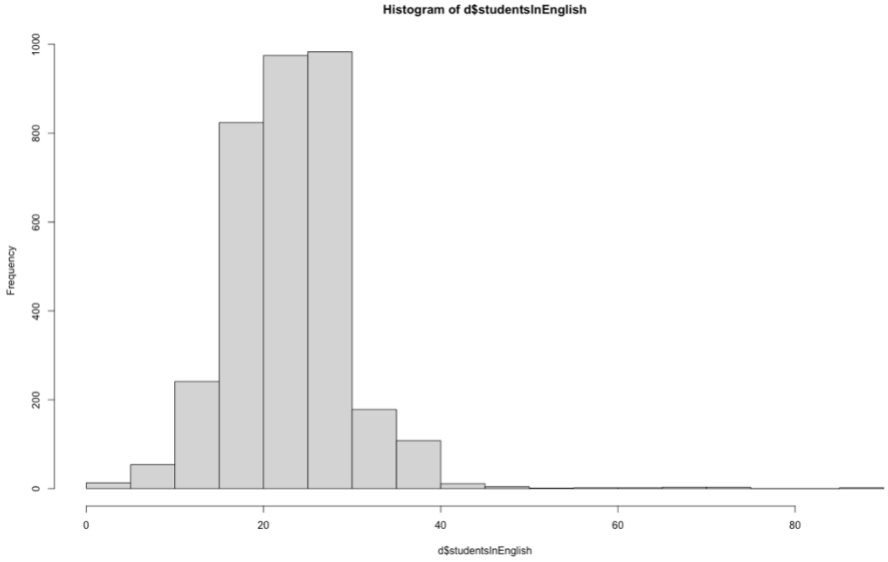


b)
 # R code:
 # Continuous variables:

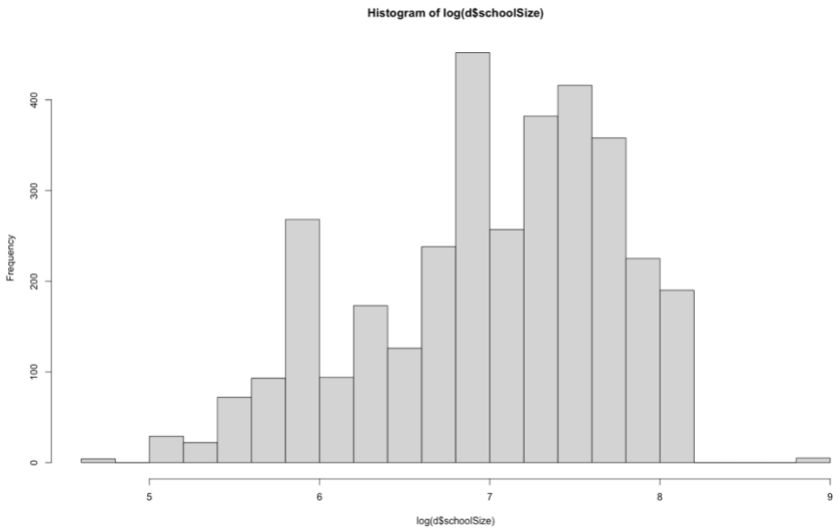
summary(d$minutesPerWeekEnglish)
hist(log(d$minutesPerWeekEnglish), breaks = 20)



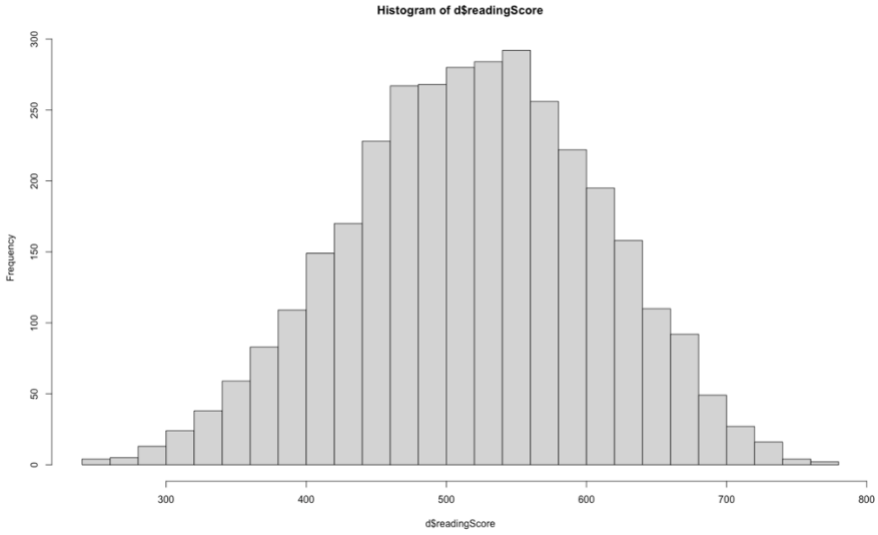
summary(d$studentsInEnglish)
hist(d$studentsInEnglish, breaks = 20)



summary(d$schoolSize)
hist(log(d$schoolSize), breaks = 20)



summary(d$readingScore)
hist(d$readingScore, breaks = 20)



 # Categorical variable:
grade

 # Dummy variables:
male
preschool
expectBachelors
motherHS
motherBachelors
motherWork
fatherHS

fatherBachelors
fatherWork
selfBornUS
motherBornUS
fatherBornUS
englishAtHome
computerForSchoolwork
read30MinsADay
schoolHasLibrary
publicSchool
urban

c)
The VIF values of all variables are less than 10 in the models. In that case, multicollinearity is not obvious here.

 # R code:
install.packages("car")
library(car)

cor(d)

```
> cor(d)
                          grade         male    preschool
grade                 1.000000000 -0.088509655  0.008110970
male                 -0.088509655  1.000000000  0.012025530
preschool             0.008110970  0.012025530  1.000000000
expectBachelors       0.115848353 -0.092327173  0.103052311
motherHS              0.015706074  0.030829424  0.138549635
motherBachelors       0.035358291  0.052540996  0.167373196
motherWork            0.032151185 -0.015030785  0.083064762
fatherHS              0.055521695  0.028284741  0.134133149
fatherBachelors       0.057962570  0.058504910  0.161455867
fatherWork            0.016955315  0.039693866  0.059649111
selfBornUS           -0.028335977  0.026842885  0.089790927
motherBornUS         -0.073731639  0.000600294  0.093708602
fatherBornUS         -0.069321531  0.011960260  0.093035483
englishAtHome        -0.009784131 -0.006461701  0.119919045
computerForSchoolwork 0.083564197 -0.017935135  0.116375374
read30MinsADay        0.041193166 -0.200024132 -0.013157918
minutesPerWeekEnglish 0.038794903 -0.004372457 -0.019019796
studentsInEnglish     0.054907567 -0.036652521 -0.030417181
schoolHasLibrary     -0.026137465  0.032065799  0.006800644
publicSchool         -0.048588328 -0.088921910 -0.100143718
urban                 0.080475246  0.025458816 -0.015045500
schoolSize            0.068044358 -0.002999718 -0.012267662
readingScore          0.222190247 -0.120639795  0.075071533
```

 # Initial full first-order model:
m1 <- lm(readingScore ~ ., data = d)
summary(m1)

```
> m1 <- lm(readingScore ~ ., data = d)
> summary(m1)

Call:
lm(formula = readingScore ~ ., data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-261.888  -50.864    0.936   52.677  269.292

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          140.669025  29.150741   4.826 1.46e-06 ***
grade                 27.617813   2.596531  10.636  < 2e-16 ***
male                 -12.322883   2.745537  -4.488 7.42e-06 ***
preschool             -2.248649   3.064201  -0.734 0.463094
expectBachelors       53.804941   3.713554  14.489  < 2e-16 ***
motherHS               2.991276   5.246017   0.570 0.568580
motherBachelors       11.792804   3.408076   3.460 0.000546 ***
motherWork            -3.509310   3.065809  -1.145 0.252431
fatherHS              11.916119   4.794813   2.485 0.012995 *
fatherBachelors       23.498748   3.492141   6.729 2.00e-11 ***
fatherWork             8.694166   3.816680   2.278 0.022792 *
selfBornUS            -0.529404   6.162315  -0.086 0.931543
motherBornUS          -0.675327   5.623806  -0.120 0.904424
fatherBornUS           7.257378   5.445224   1.333 0.182688
englishAtHome         12.230226   5.949326   2.056 0.039885 *
computerForSchoolwork 26.768734   5.020714   5.332 1.04e-07 ***
read30MinsADay        33.004277   2.973678  11.099  < 2e-16 ***
minutesPerWeekEnglish  0.015511   0.009373   1.655 0.098034 .
studentsInEnglish      0.004265   0.199825   0.021 0.982972
schoolHasLibrary      -3.709011   7.863951  -0.472 0.637209
publicSchool         -26.978189   5.776670  -4.670 3.13e-06 ***
urban                -10.960512   3.384836  -3.238 0.001215 **
schoolSize             0.006934   0.001877   3.693 0.000225 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.06 on 3381 degrees of freedom
Multiple R-squared:  0.2579,    Adjusted R-squared:  0.2531
F-statistic: 53.41 on 22 and 3381 DF,  p-value: < 2.2e-16
```

vif(m1)

```
> vif(m1)
               grade                male            preschool
            1.045874            1.080266             1.070901
     expectBachelors            motherHS      motherBachelors
            1.128145            1.550114             1.523190
          motherWork            fatherHS      fatherBachelors
            1.059733            1.517605             1.571106
          fatherWork          selfBornUS         motherBornUS
            1.034240            1.415586             3.069426
        fatherBornUS        englishAtHome computerForSchoolwork
            2.900185            2.194543             1.097927
      read30MinsADay minutesPerWeekEnglish     studentsInEnglish
            1.064383            1.009872             1.110651
    schoolHasLibrary         publicSchool                urban
            1.040219            1.467492             1.513505
          schoolSize
            1.478538
```

 # Final first-order model:
m2 <- lm(readingScore ~ grade + male + expectBachelors + motherBachelors + fatherHS + fatherBachelors + fatherWork + englishAtHome + computerForSchoolwork + read30MinsADay + publicSchool + urban + schoolSize, data = d)
summary(m2)

```
Call:
lm(formula = readingScore ~ grade + male + expectBachelors +
    motherBachelors + fatherHS + fatherBachelors + fatherWork +
    englishAtHome + computerForSchoolwork + read30MinsADay +
    publicSchool + urban + schoolSize, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-252.914  -51.229    0.834   52.732  266.478

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         141.052061  27.315216   5.164 2.56e-07 ***
grade                27.526139   2.583996  10.653  < 2e-16 ***
male                -12.157285   2.737981  -4.440 9.27e-06 ***
expectBachelors      53.206674   3.695715  14.397  < 2e-16 ***
motherBachelors      11.625893   3.357648   3.463 0.000542 ***
fatherHS             13.806876   4.372171   3.158 0.001603 **
fatherBachelors      23.541143   3.474041   6.776 1.45e-11 ***
fatherWork            8.089382   3.798603   2.130 0.033279 *
englishAtHome        16.559614   4.450749   3.721 0.000202 ***
computerForSchoolwork 26.205402   4.977474   5.265 1.49e-07 ***
read30MinsADay       33.400608   2.966784  11.258  < 2e-16 ***
publicSchool        -25.998915   5.727562  -4.539 5.84e-06 ***
urban               -11.551918   3.299346  -3.501 0.000469 ***
schoolSize            0.006639   0.001823   3.643 0.000274 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.04 on 3390 degrees of freedom
Multiple R-squared:  0.2562,    Adjusted R-squared:  0.2534
F-statistic: 89.84 on 13 and 3390 DF,  p-value: < 2.2e-16
```

vif(m2)

```
> vif(m2)
               grade                male       expectBachelors
            1.036205            1.074748            1.117769
     motherBachelors            fatherHS       fatherBachelors
            1.479025            1.262348            1.555469
          fatherWork        englishAtHome computerForSchoolwork
            1.024866            1.228696            1.079519
       read30MinsADay         publicSchool               urban
            1.059868            1.443212            1.438581
          schoolSize
            1.393886
```

d)

 # Dummy variables:

d$male <- ifelse(d$male == "male", 1, 0)

d$preschool <- ifelse(d$preschool == "preschool", 1, 0)

d$expectBachelors <- ifelse(d$expectBachelors == "expectBachelors", 1, 0)

d$motherHS <- ifelse(d$motherHS == "motherHS", 1, 0)

```
d$motherBachelors <- ifelse(d$motherBachelors == "motherBachelors", 1, 0)
d$motherWork <- ifelse(d$motherWork == "motherWork", 1, 0)
d$fatherHS <- ifelse(d$fatherHS == "fatherHS", 1, 0)
d$fatherBachelors <- ifelse(d$fatherBachelors == "fatherBachelors", 1, 0)
d$fatherWork <- ifelse(d$fatherWork == "fatherWork", 1, 0)
d$selfBornUS <- ifelse(d$selfBornUS == "selfBornUS", 1, 0)
d$motherBornUS <- ifelse(d$motherBornUS == "motherBornUS", 1, 0)
d$fatherBornUS <- ifelse(d$fatherBornUS == "fatherBornUS", 1, 0)
d$englishAtHome <- ifelse(d$englishAtHome == "englishAtHome", 1, 0)
d$computerForSchoolwork <- ifelse(d$computerForSchoolwork ==
"computerForSchoolwork", 1, 0)
d$read30MinsADay <- ifelse(d$read30MinsADay == "read30MinsADay", 1, 0)
d$schoolHasLibrary <- ifelse(d$schoolHasLibrary == "schoolHasLibrary", 1, 0)
d$publicSchool <- ifelse(d$publicSchool == "publicSchool", 1, 0)
d$urban <- ifelse(d$urban == "urban", 1, 0)
```

e)

Model after backward stepwise selection:

readingScore = 129.63847 + 37.86558 * grade + 0.02013 * minutesPerWeekEnglish + e

```
# R code:
install.packages("MASS")
library(MASS)

model_full <- lm(readingScore ~ ., data = d)
step <- stepAIC(model_full, direction = "backward")
step$anova

model_step <- lm(readingScore ~ grade + minutesPerWeekEnglish, data = d)
summary(model_step)
```

```
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
readingScore ~ grade + male + preschool + expectBachelors + motherHS +
    motherBachelors + motherWork + fatherHS + fatherBachelors +
    fatherWork + selfBornUS + motherBornUS + fatherBornUS + englishAtHo
me +
    computerForSchoolwork + read30MinsADay + minutesPerWeekEnglish +
    studentsInEnglish + schoolHasLibrary + publicSchool + urban +
    schoolSize

Final Model:
readingScore ~ grade + minutesPerWeekEnglish
                     Step Df Deviance Resid. Df Resid. Dev
1                                         3399   25684324
2                   - urban  0    0.000      3399   25684324
3            - publicSchool  0    0.000      3399   25684324
4         - schoolHasLibrary  0    0.000      3399   25684324
5           - read30MinsADay  0    0.000      3399   25684324
6    - computerForSchoolwork  0    0.000      3399   25684324
7           - englishAtHome  0    0.000      3399   25684324
8            - fatherBornUS  0    0.000      3399   25684324
9            - motherBornUS  0    0.000      3399   25684324
10             - selfBornUS  0    0.000      3399   25684324
11             - fatherWork  0    0.000      3399   25684324
12        - fatherBachelors  0    0.000      3399   25684324
13                - fatherHS  0    0.000      3399   25684324
14             - motherWork  0    0.000      3399   25684324
15        - motherBachelors  0    0.000      3399   25684324
16                - motherHS  0    0.000      3399   25684324
17        - expectBachelors  0    0.000      3399   25684324
18              - preschool  0    0.000      3399   25684324
19                  - male  0    0.000      3399   25684324
```

```
20     - studentsInEnglish  1 1205.159      3400   25685529
21           - schoolSize  1 6021.438      3401   25691550
        AIC
1  30403.24
2  30403.24
3  30403.24
4  30403.24
5  30403.24
6  30403.24
7  30403.24
8  30403.24
9  30403.24
10 30403.24
11 30403.24
12 30403.24
13 30403.24
14 30403.24
15 30403.24
16 30403.24
17 30403.24
18 30403.24
19 30403.24
20 30401.40
21 30400.20
```

f)

Model after considering second-order terms:

readingScore = (-2.529e+03) + (5.582e+02) * grade + (1.351e-01) * minutesPerWeekEnglish + (-1.121e-04) * minutesPerWeekEnglishSQ + (-2.559e+01) * gradeSQ+ e

```
# R code:
# Second order terms
d$minutesPerWeekEnglishSQ <- (d$minutesPerWeekEnglish)^2
d$studentsInEnglishSQ <- (d$studentsInEnglish)^2
d$schoolSizeSQ <- (d$schoolSize)^2
d$gradeSQ <- (d$grade)^2

d$maleSQ <- (d$male)^2
d$preschoolSQ <- (d$preschool)^2
d$expectBachelorsSQ <- (d$expectBachelors)^2
d$motherHSSQ <- (d$motherHS)^2
d$motherBachelorsSQ <- (d$motherBachelors)^2
d$motherWorkSQ <- (d$motherWork)^2
d$fatherHSSQ <- (d$fatherHS)^2
d$fatherBachelorsSQ <- (d$fatherBachelors)^2
d$fatherWorkSQ <- (d$fatherWork)^2
d$selfBornUSSQ <- (d$selfBornUS)^2
```

```
d$motherBornUSSQ <- (d$motherBornUS)^2
d$fatherBornUSSQ <- (d$fatherBornUS)^2
d$englishAtHomeSQ <- (d$englishAtHome)^2
d$computerForSchoolworkSQ <- (d$computerForSchoolwork)^2
d$read30MinsADaySQ <- (d$read30MinsADay)^2
d$schoolHasLibrarySQ <- (d$schoolHasLibrary)^2
d$publicSchoolSQ <- (d$publicSchool)^2
d$urbanSQ <- (d$urban)^2

model_f <- lm(readingScore ~ ., data = d)
summary(model_f)
step <- stepAIC(model_f, direction = "backward")
step$anova

model_f1 <- lm(readingScore ~ grade + minutesPerWeekEnglish + schoolSize +
minutesPerWeekEnglishSQ + schoolSizeSQ + gradeSQ, data = d)
summary(model_f1)

model_f2 <- lm(readingScore ~ grade + minutesPerWeekEnglish + minutesPerWeekEnglishSQ
+ gradeSQ, data = d)
summary(model_f2)
```

```
Call:
lm(formula = readingScore ~ grade + minutesPerWeekEnglish + minutesPerWeekEng
lishSQ +
    gradeSQ, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-296.90  -58.13    1.10   59.62  340.02

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.529e+03  3.466e+02  -7.298 3.62e-13 ***
grade                    5.582e+02  6.816e+01   8.189 3.69e-16 ***
minutesPerWeekEnglish    1.351e-01  2.185e-02   6.183 7.02e-10 ***
minutesPerWeekEnglishSQ -1.121e-04  1.846e-05  -6.073 1.40e-09 ***
gradeSQ                 -2.559e+01  3.347e+00  -7.645 2.69e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85.7 on 3399 degrees of freedom
Multiple R-squared:  0.07726,   Adjusted R-squared:  0.07617
F-statistic: 71.14 on 4 and 3399 DF,  p-value: < 2.2e-16
```

g)
Model after considering interaction terms:

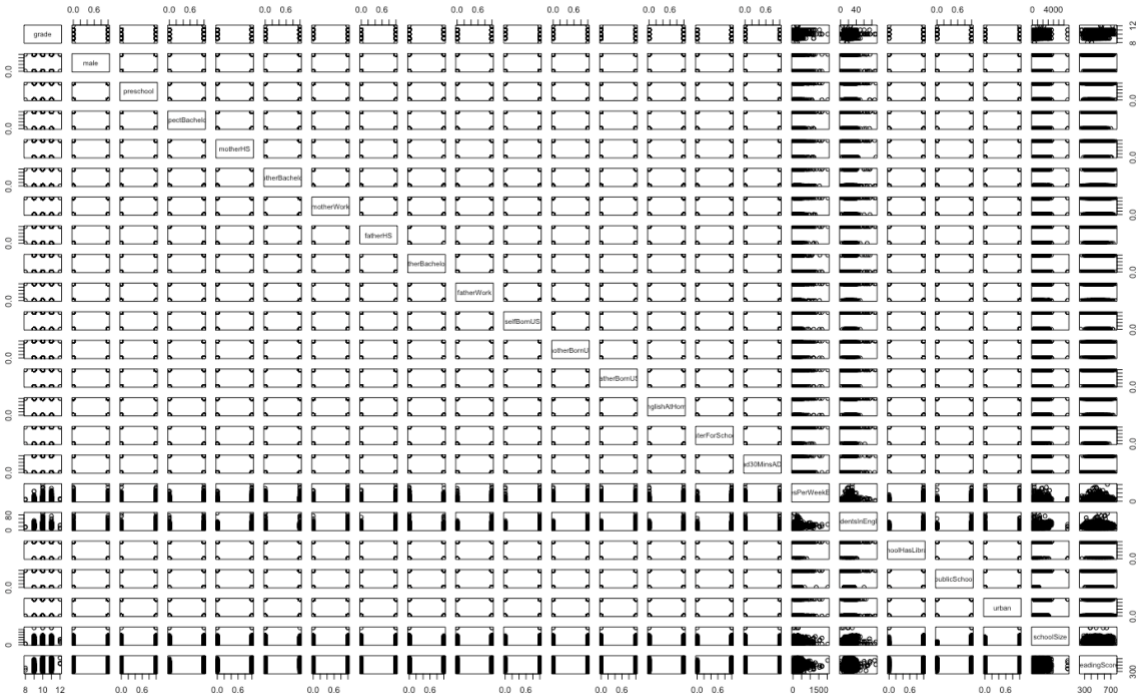
```
readingScore = (-2.529e+03) + (5.582e+02) * grade + (1.351e-01) * minutesPerWeekEnglish +
(-1.121e-04) * minutesPerWeekEnglishSQ + (-2.559e+01) * gradeSQ + e
summary(model_i2)

# R code:
# Interaction terms
d$GM <- d$grade * d$minutesPerWeekEnglish
d$GSE <- d$grade * d$studentsInEnglish
d$GSS <- d$grade * d$schoolSize
d$MSE <- d$minutesPerWeekEnglish * d$studentsInEnglish
d$MSS <- d$minutesPerWeekEnglish * d$schoolSize
d$SESS <- d$studentsInEnglish * d$schoolSize

model_i <- lm(readingScore ~ ., data = d)
summary(model_i)
step <- stepAIC(model_i, direction = "backward")
step$anova

model_i1 <- lm(readingScore ~ grade + minutesPerWeekEnglish + studentsInEnglish +
schoolSize + minutesPerWeekEnglishSQ + schoolSizeSQ + gradeSQ + GSE, data = d)
summary(model_i1)

model_i2 <- lm(readingScore ~ grade + minutesPerWeekEnglish + minutesPerWeekEnglishSQ
+ gradeSQ, data = d)
summary(model_i2)
```




```
Call:
lm(formula = readingScore ~ grade + minutesPerWeekEnglish + minute
sPerWeekEnglishSQ +
    gradeSQ, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-296.90  -58.13    1.10   59.62  340.02

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.529e+03  3.466e+02  -7.298 3.62e-13
grade                    5.582e+02  6.816e+01   8.189 3.69e-16
minutesPerWeekEnglish    1.351e-01  2.185e-02   6.183 7.02e-10
minutesPerWeekEnglishSQ -1.121e-04  1.846e-05  -6.073 1.40e-09
gradeSQ                 -2.559e+01  3.347e+00  -7.645 2.69e-14

(Intercept)             ***
grade                   ***
minutesPerWeekEnglish   ***
minutesPerWeekEnglishSQ ***
gradeSQ                 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85.7 on 3399 degrees of freedom
Multiple R-squared:  0.07726,   Adjusted R-squared:  0.07617
F-statistic: 71.14 on 4 and 3399 DF,  p-value: < 2.2e-16
```

h)
Final model after transform variables:
readingScore = (350.96) + (71.00) * grade9 + (143.57) * grade10 + (161.64) * grade11 + (215.32) * grade12 + (-2104.01) * log(minutesPerWeekEnglish + 1) + (1055.34) * log(minutesPerWeekEnglishSQ + 1) + e
 # R code:
 # set as levels
d$grade <- as.factor(d$grade)

hist(d$readingScore, breaks = 20)
hist(d$minutesPerWeekEnglish, breaks = 20)
hist(log(d$minutesPerWeekEnglish), breaks = 20)
hist(log(d$minutesPerWeekEnglishSQ), breaks = 20)

model_final <- lm(readingScore ~ grade + log(minutesPerWeekEnglish + 1) + log(minutesPerWeekEnglishSQ + 1), data = d)
summary(model_final)

```
Call:
lm(formula = readingScore ~ grade + log(minutesPerWeekEnglish +
    1) + log(minutesPerWeekEnglishSQ + 1), data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-296.809  -58.640    0.745   58.562  254.885

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     350.96      61.97   5.663 1.61e-08
grade9                           71.00      60.45   1.175  0.24025
grade10                         143.57      60.25   2.383  0.01722
grade11                         161.64      60.31   2.680  0.00739
grade12                         215.32      73.75   2.919  0.00353
log(minutesPerWeekEnglish + 1)  -2104.01     395.25  -5.323 1.09e-07
log(minutesPerWeekEnglishSQ + 1) 1055.34     196.88   5.360 8.86e-08

(Intercept)                        ***
grade9
grade10                            *
grade11                            **
grade12                            **
log(minutesPerWeekEnglish + 1)     ***
log(minutesPerWeekEnglishSQ + 1) ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85.16 on 3397 degrees of freedom
Multiple R-squared:  0.08933,   Adjusted R-squared:  0.08772
F-statistic: 55.54 on 6 and 3397 DF,  p-value: < 2.2e-16
```

i)
Final model:
readingScore = (350.96) + (71.00) * grade9 + (143.57) * grade10 + (161.64) * grade11 + (215.32) * grade12 + (-2104.01) * log(minutesPerWeekEnglish + 1) + (1055.34) * log(minutesPerWeekEnglishSQ + 1) + e

P.S.
d$minutesPerWeekEnglishSQ <- (d$minutesPerWeekEnglish)^2

The value of adjusted R-squared (0.08772 or 8.772%) coefficient indicates the quantity of the variation in readingScore explained by the regression line. At this time, Adj-R2 (0.08772 or 8.772%) of the variation in readingScore is explained by grade, minutesPerWeekEnglish and minutesPerWeekEnglishSQ. Adj-R2 (0.08772 or 8.772%) does not increase with the addition of a x-variable that does not improve the regression model. A lower Adj-R2 (0.08772 or 8.772%) typically indicates a less than ideal model.

T-Test:
Our p-value for the final model is 2.2e-16, which is quite close to zero. Usually, a p-value with 0.05 or less is a good sign. Therefore, the small values of p-value for the intercept and slope indicates we can reject the null hypothesis and there is a relationship between readingScore (y) and grade & minutesPerWeekEnglish & minutesPerWeekEnglishSQ (x-var).

F-Test:
Null hypothesis:
Ho: $\beta 1 = \beta 2 = \beta 3 = \ldots = \beta k = 0$
Alternative hypothesis:
Ha: At least one coefficient $\beta j \neq 0$

Test statistic:
F = 55.54
Therefore, F = 55.54 and with p-value less than 0.05 (at alpha=0.05). The null hypothesis of no association between readingScore (y) and grade & minutesPerWeekEnglish & minutesPerWeekEnglishSQ (x-var) is rejected. At least one x-variable has a significant effect on changes in readingScore. F-test gives strong support to the fitted model.

```
Call:
lm(formula = readingScore ~ grade + log(minutesPerWeekEnglish +
    1) + log(minutesPerWeekEnglishSQ + 1), data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-296.809  -58.640    0.745   58.562  254.885

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                      350.96      61.97   5.663 1.61e-08
grade9                            71.00      60.45   1.175  0.24025
grade10                          143.57      60.25   2.383  0.01722
grade11                          161.64      60.31   2.680  0.00739
grade12                          215.32      73.75   2.919  0.00353
log(minutesPerWeekEnglish + 1)  -2104.01     395.25  -5.323 1.09e-07
log(minutesPerWeekEnglishSQ + 1) 1055.34     196.88   5.360 8.86e-08

(Intercept)                        ***
grade9
grade10                            *
grade11                            **
grade12                            **
log(minutesPerWeekEnglish + 1)     ***
log(minutesPerWeekEnglishSQ + 1) ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85.16 on 3397 degrees of freedom
Multiple R-squared:  0.08933,   Adjusted R-squared:  0.08772
F-statistic: 55.54 on 6 and 3397 DF,  p-value: < 2.2e-16
```

j)

The Programme for International Student Assessment (PISA) is a test given every three years to 15-year-old students from around the world to evaluate their performance in mathematics, reading, and science. This test provides a quantitative way to compare the performance of students from different parts of the world. In this homework assignment, we will predict the reading scores of students from the United States of America on the 2009 PISA exam.

According to our preliminary basic analysis of the data, we found that the higher the student's grade (divided into four grades, including grade 9, grade 10, grade 11, grade 12), the higher the corresponding reading score obtained. This variable has a significant impact on the results. At the same time, the longer the number of minutes per week the student spend in English class, the higher the corresponding reading grade they will get in the end.

Grade and minutesPerWeekEnglish, are directly proportional to the readingScore.