honor statement: "I have completed this work independently. The solutions given are entirely my own work."

1) (50 points) Download the Pisa2009 Dataset from the D2L. 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. The dataset contains information about the demographics and schools for American students taking the exam, derived from 2009 PISA Public-Use Data Files distributed by the United States National Center for Education Statistics (NCES). Each row in the dataset represents one student taking the exam. The datasets have the following variables:

grade: The grade in school of the student (most 15-year-olds in America are in 10th grade)
male: Whether the student is male (1/0)
raceeth: The race/ethnicity composite of the student preschool:
Whether the student attended preschool (1/0)
expectBachelors: Whether the student expects to obtain a bachelor's degree (1/0) motherHS: Whether the student's mother completed high school (1/0)
motherBachelors: Whether the student's mother obtained a bachelor's degree (1/0) motherWork: Whether the student's mother has part-time or full-time work (1/0) fatherHS: Whether the student's father completed high school (1/0)
fatherBachelors: Whether the student's father obtained a bachelor's degree (1/0) fatherWork: Whether the student's father has part-time or full-time work (1/0) selfBornUS: Whether the student was born in the United States of America (1/0) motherBornUS: Whether the student's mother was born in the United States of America (1/0) f
atherBornUS: Whether the student's father was born in the United States of America (1/0) englishAtHome: Whether the student speaks English at home (1/0) computerForSchoolwork: Whether the student has access to a computer for schoolwork (1/0)
read30MinsADay: Whether the student reads for pleasure for 30 minutes/day (1/0) minutesPerWeekEnglish: The number of minutes per week the student spend in English class studentsInEnglish: The number of students in this student's English class at school schoolHasLibrary: Whether this student's school has a library (1/0) publicSchool: Whether this student attends a public school (1/0)
urban: Whether this student's school is in an urban area (1/0)
schoolSize: The number of students in this student's school
readingScore: The student's reading score, on a 1000-point scale

Write a professional report detailing your analysis of the dataset including your efforts to…

a. Create a training and testing set using n-fold cross validation.

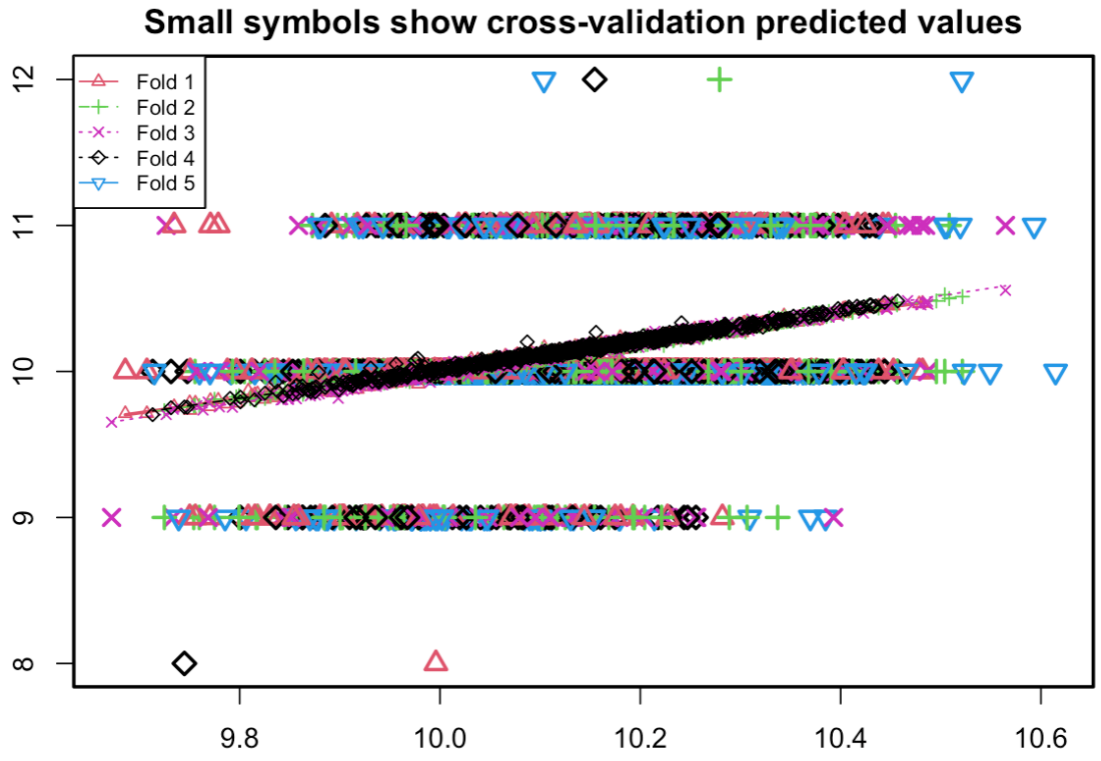
1. Build up the first model with all variables and check the F-test, F-test is good enough to reject the null hypothesis which means at least one H is not equal to zero.
2. Then check ad r-squared : 0.304 .
3.Then Check the p-value of all of the variables,
first I would remove variable of schoolHasLibrary ,and then remove →preschool→ selfBornUS→ urban →fatherBornUS →motherWork→ motherHS→ studentsInEnglish→ fatherWork→ motherBornUS →minutesPerWeekEnglish→ englishAtHome

then I'll have the first - order model with variables :grade,male,raceeth,expectBachelors motherBachelors ,fatherHS,fatherBachelors ,computerForSchoolwork ,read30MinsADay ,publicSchool,schoolSize .

4. Then I'll evaluate the model using 5-fold cross validation ,the overall average of the mean square is 241 which is very large and tells us the regression line is not close to a set of points

```
> Pisa2009_1<-Pisa2009[,-c(1,21,5,13,23,15,9,7,20,12,14,19,16)]
> model <-lm(readingScore~., data = Pisa2009_1)
> summary(model)

Call:
lm(formula = readingScore ~ ., data = Pisa2009_1)

Residuals:
    Min      1Q  Median      3Q     Max
-252.78  -48.55    1.25   49.26  247.80

Coefficients:
                                                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                                     122.56390   28.98059    4.23    2.4e-05 ***
grade                                            26.55105    2.49806   10.63    < 2e-16 ***
male                                            -12.69162    2.64470   -4.80    1.7e-06 ***
raceethAsian                                     55.13496   14.93933    3.69    0.00023 ***
raceethBlack                                     -6.06143   14.11015   -0.43    0.66753
raceethHispanic                                  23.95077   13.77388    1.74    0.08215 .
raceethMore than one race                        40.90995   15.10912    2.71    0.00681 **
raceethNative Hawaiian/Other Pacific Islander  52.38934   19.91875    2.63    0.00857 **
raceethWhite                                     61.59271   13.57375    4.54    5.9e-06 ***
expectBachelors                                  53.98829    3.57089   15.12    < 2e-16 ***
motherBachelors                                  11.33870    3.24582    3.49    0.00048 ***
fatherHS                                         10.72529    4.19269    2.56    0.01057 *
fatherBachelors                                  18.04509    3.36977    5.35    9.1e-08 ***
computerForSchoolwork                            21.64581    4.81178    4.50    7.1e-06 ***
read30MinsADay                                   33.11985    2.86254   11.57    < 2e-16 ***
publicSchool                                    -17.28338    4.98322   -3.47    0.00053 ***
schoolSize                                       0.00667    0.00163    4.10    4.3e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.4 on 3387 degrees of freedom
Multiple R-squared:  0.308,    Adjusted R-squared:  0.304
F-statistic:   94 on 16 and 3387 DF,  p-value: <2e-16
```

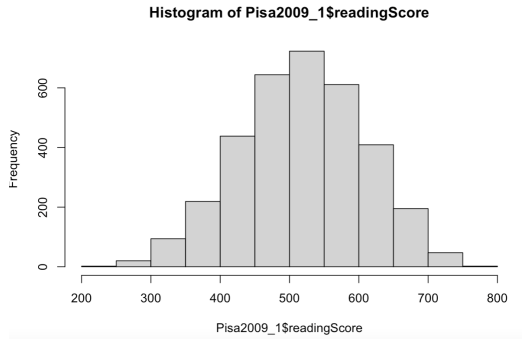
```
> out<- cv.lm(data = Pisa2009_1 , form.lm = (grade~.),plotit = "Observed", m=5)
```



b. Perform appropriate univariate and bivariate analysis on the data.

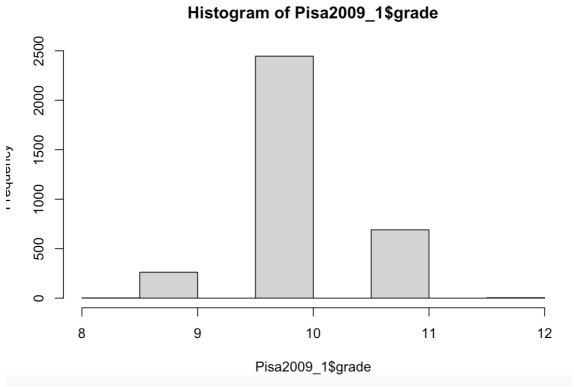
Univariate analysis on the data

The reading Score , this is what we are trying to predict, looks normal.
hist(Pisa2009_1$readingScore)



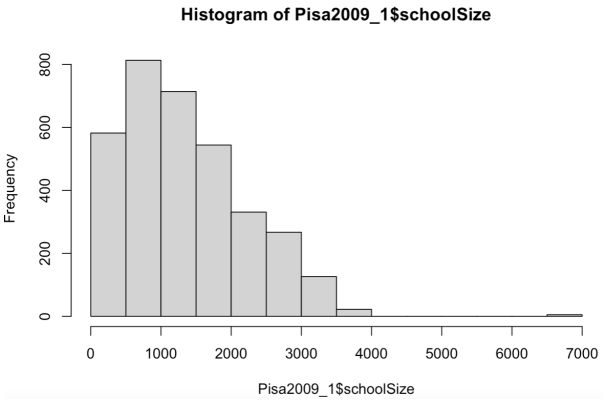
hist(Pisa2009_1$grade)

Grade looks normal



hist(Pisa2009_1$schoolSize)

Scoolsize looks normal with a right - skewed tail.



Bivariate analysis on the data, look at the correlation in the entire dataset

> cor(Pisa2009_1)

From the chart below, we did not see any variables that is strongly correlated with readingScore

	grade	male	expectBachelors	motherBachelors	fatherHS	fatherBachelors
grade	1.0000	-0.0885	0.1158	0.03536	0.0555	0.0580
male	-0.0885	1.0000	-0.0923	0.05254	0.0283	0.0585
expectBachelors	0.1158	-0.0923	1.0000	0.17717	0.1605	0.2202
motherBachelors	0.0354	0.0525	0.1772	1.00000	0.2030	0.5502
fatherHS	0.0555	0.0283	0.1605	0.20297	1.0000	0.2724
fatherBachelors	0.0580	0.0585	0.2202	0.55020	0.2724	1.0000
computerForSchoolwork	0.0836	-0.0179	0.1534	0.13795	0.1651	0.1600
read30MinsADay	0.0412	-0.2000	0.1138	0.02985	0.0389	0.0484
publicSchool	-0.0486	-0.0889	-0.1099	-0.18633	-0.0839	-0.1920
schoolSize	0.0680	-0.0030	0.0385	-0.00374	-0.0807	0.0206
readingScore	0.2222	-0.1206	0.3433	0.22864	0.1950	0.2790

	computerForSchoolwork	read30MinsADay	publicSchool	schoolSize	readingScore
grade	0.0836	0.0412	-0.0486	0.06804	0.2222
male	-0.0179	-0.2000	-0.0889	-0.00300	-0.1206
expectBachelors	0.1534	0.1138	-0.1099	0.03853	0.3433
motherBachelors	0.1379	0.0299	-0.1863	-0.00374	0.2286
fatherHS	0.1651	0.0389	-0.0839	-0.08072	0.1950
fatherBachelors	0.1600	0.0484	-0.1920	0.02060	0.2790
computerForSchoolwork	1.0000	-0.0196	-0.0716	0.06666	0.1786
read30MinsADay	-0.0196	1.0000	0.0104	-0.01574	0.2242
publicSchool	-0.0716	0.0104	1.0000	0.25832	-0.1187
schoolSize	0.0667	-0.0157	0.2583	1.00000	0.0302

readingScore	0.1786	0.2242	-0.1187	0.03023	1.0000

c. Check for multicollinearity.

Since I did not see the strong correlation between the variables, so I will check the multicollinearity by VIF.
We are worried about anything over 10, so we did not have to worried about any variables below, their VIF value are low.

```
> vif(model)
                        GVIF Df GVIF^(1/(2*Df))
grade                   1.04  1          1.02
male                    1.08  1          1.04
raceeth                 1.35  6          1.03
expectBachelors         1.12  1          1.06
motherBachelors         1.48  1          1.22
fatherHS                1.25  1          1.12
fatherBachelors         1.57  1          1.25
computerForSchoolwork 1.08  1          1.04
read30MinsADay          1.06  1          1.03
publicSchool            1.17  1          1.08
schoolSize              1.20  1          1.09
```

d. Create appropriate dummy variables.

```
Pisa2009_1$male<- as.factor(Pisa2009_1$male)
Pisa2009_1$raceeth<- as.factor(Pisa2009_1$raceeth)
Pisa2009_1$expectBachelors<- as.factor(Pisa2009_1$expectBachelors)
Pisa2009_1$fatherHS<- as.factor(Pisa2009_1$fatherHS)

Pisa2009_1$fatherBachelors<- as.factor(Pisa2009_1$fatherBachelors)
Pisa2009_1$computerForSchoolwork<- as.factor(Pisa2009_1$computerForSchoolwork)
Pisa2009_1$read30MinsADay<- as.factor(Pisa2009_1$read30MinsADay)
Pisa2009_1$publicSchool<- as.factor(Pisa2009_1$publicSchool)
Pisa2009_1$motherBachelors<- as.factor(Pisa2009_1$motherBachelors)
```

e. Perform feature selection.

1. Build up the first model with all variables and check the F-test, F-test is good enough to reject the null hypothesis which means at least one H is not equal to zero.
2. Then check ad r-squared : 0.304 .
3.Then Check the p-value of all of the variables,
first I would remove variable of schoolHasLibrary ,and then remove →preschool→ selfBornUS→ urban →fatherBornUS →motherWork→ motherHS→ studentsInEnglish→ fatherWork→ motherBornUS →minutesPerWeekEnglish→ englishAtHome

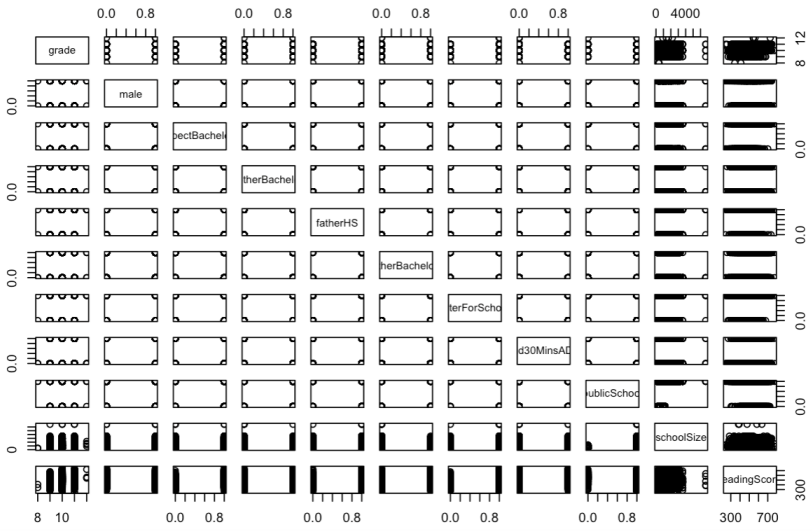
then I'll have the first - order model with variables :grade,male,raceeth,expectBachelors motherBachelors ,fatherHS,fatherBachelors ,computerForSchoolwork ,read30MinsADay ,publicSchool,schoolSize .

4. Then I'll evaluate the model using 5-fold cross validation ,the overall average of the mean square is 241 which is very large and tells us the regression line is not close to a set of points

```
> Pisa2009_1<-Pisa2009[,-c(1,21,5,13,23,15,9,7,20,12,14,19,16)]
> model <-lm(readingScore~., data = Pisa2009_1)
> summary(model)

Call:
lm(formula = readingScore ~ ., data = Pisa2009_1)
```

f. Check for appropriate second order terms.

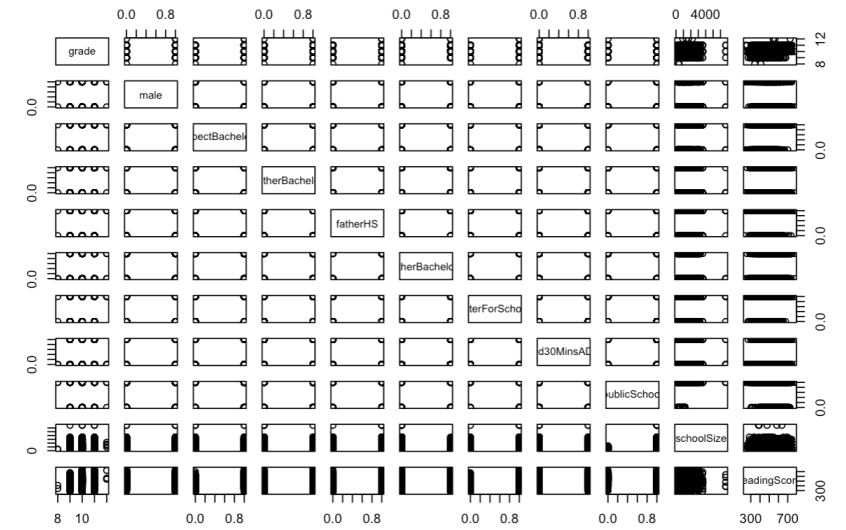
Since there is a slightly linear correlation between reading school and school size/grade, I would add second order term by using variables of school size and grade. There is a slight improvement on the adjusted r-square from 30.5% to 31 %.



```
> Pisa2009_1$schoolSizeSQ<- (Pisa2009_1$schoolSize)^2/1000
> Pisa2009_1$gradeSQ<- (Pisa2009_1$grade)^2
> model <-lm(readingScore~., data = Pisa2009_1)
> summary(model)

Call:
lm(formula = readingScore ~ ., data = Pisa2009_1)

Residuals:
    Min      1Q  Median      3Q     Max
-256.86  -48.89    1.53   49.76  244.84

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    -1.47e+03   3.03e+02   -4.85  1.3e-06 ***
grade                                           3.41e+02   5.96e+01    5.72  1.2e-08 ***
male1                                          -1.25e+01   2.63e+00   -4.74  2.2e-06 ***
raceethAsian                                    5.72e+01   1.49e+01    3.84  0.00013 ***
raceethBlack                                   -4.88e+00   1.41e+01   -0.35  0.72889
raceethHispanic                                 2.60e+01   1.37e+01    1.89  0.05861 .
raceethMore than one race                       4.23e+01   1.51e+01    2.81  0.00502 **
raceethNative Hawaiian/Other Pacific Islander   5.36e+01   1.98e+01    2.70  0.00694 **
raceethWhite                                    6.22e+01   1.35e+01    4.60  4.4e-06 ***
expectBachelors1                                5.28e+01   3.57e+00   14.81  < 2e-16 ***
motherBachelors1                                1.14e+01   3.23e+00    3.53  0.00042 ***
fatherHS1                                       9.97e+00   4.18e+00    2.38  0.01726 *
fatherBachelors1                                1.77e+01   3.36e+00    5.26  1.6e-07 ***
computerForSchoolwork1                          2.00e+01   4.80e+00    4.15  3.4e-05 ***
read30MinsADay1                                 3.29e+01   2.85e+00   11.53  < 2e-16 ***
publicSchool1                                  -1.88e+01   5.01e+00   -3.75  0.00018 ***
schoolSize                                      1.33e-02   4.21e-03    3.16  0.00162 **
schoolSizeSQ                                   -1.84e-03   1.09e-03   -1.69  0.09101 .
gradeSQ                                        -1.54e+01   2.92e+00   -5.28  1.4e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74.1 on 3385 degrees of freedom
Multiple R-squared:  0.314,    Adjusted R-squared:  0.31
F-statistic:   86 on 18 and 3385 DF,  p-value: <2e-16
```

g. Check for appropriate interaction terms.

By checking the plot, it seems to have a correlation between variables of schoolsize and grade, so I would try to add interaction term by using these two variables, however the adjusted r-square did not improve, so I would not include the interaction term into my model.




```
> Pisa2009_1$gdschsize<- (Pisa2009_1$grade)*(Pisa2009_1$schoolSize)
> model <-lm(readingScore~., data = Pisa2009_1)
> summary(model)

Call:
lm(formula = readingScore ~ ., data = Pisa2009_1)

Residuals:
   Min     1Q Median     3Q    Max
-256.9  -48.6    1.7   49.6  244.9

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     -1.48e+03   3.04e+02   -4.88  1.1e-06 ***
grade                                            3.39e+02   5.96e+01    5.69  1.4e-08 ***
male1                                           -1.25e+01   2.64e+00   -4.75  2.1e-06 ***
raceethAsian                                     5.76e+01   1.49e+01    3.86  0.00012 ***
raceethBlack                                    -4.69e+00   1.41e+01   -0.33  0.73899
raceethHispanic                                  2.61e+01   1.37e+01    1.90  0.05733 .
raceethMore than one race                        4.25e+01   1.51e+01    2.82  0.00479 **
raceethNative Hawaiian/Other Pacific Islander  5.32e+01   1.98e+01    2.68  0.00735 **
raceethWhite                                     6.24e+01   1.35e+01    4.61  4.1e-06 ***
expectBachelors1                                 5.28e+01   3.57e+00   14.79  < 2e-16 ***
motherBachelors1                                 1.15e+01   3.23e+00    3.54  0.00040 ***
fatherHS1                                        9.93e+00   4.18e+00    2.37  0.01766 *
```

```
fatherBachelors1                          1.76e+01   3.36e+00    5.24  1.7e-07 ***
computerForSchoolwork1                    2.00e+01   4.80e+00    4.15  3.3e-05 ***
read30MinsADay1                           3.29e+01   2.85e+00   11.54  < 2e-16 ***
publicSchool1                            -1.86e+01   5.02e+00   -3.69  0.00022 ***
schoolSize                                3.75e-02   2.82e-02    1.33  0.18284
schoolSizeSQ                             -1.78e-03   1.09e-03   -1.64  0.10213
gradeSQ                                  -1.52e+01   2.94e+00   -5.18  2.4e-07 ***
gdschsize                                -2.41e-03   2.77e-03   -0.87  0.38435
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74.1 on 3384 degrees of freedom
Multiple R-squared:  0.314,    Adjusted R-squared:  0.31
F-statistic: 81.5 on 19 and 3384 DF,  p-value: <2e-16
```

h. Transform variables as needed.

I would log the readingscore to see its outcome; the Adjusted R-squared has been improved from 0.31 to 0.314.

```
> model <-lm(log(readingScore)~., data = Pisa2009_1)
> summary(model)

Call:
lm(formula = log(readingScore) ~ ., data = Pisa2009_1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6106 -0.0869  0.0138  0.1010  0.4113

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               1.72e+00   6.11e-01    2.82  0.00480 **
grade                                     7.75e-01   1.20e-01    6.46  1.2e-10 ***
male1                                    -2.60e-02   5.31e-03   -4.90  9.9e-07 ***
raceethAsian                              1.15e-01   3.00e-02    3.82  0.00013 ***
raceethBlack                             -1.15e-02   2.83e-02   -0.41  0.68404
raceethHispanic                           5.47e-02   2.76e-02    1.98  0.04785 *
raceethMore than one race                 9.05e-02   3.03e-02    2.98  0.00286 **
raceethNative Hawaiian/Other Pacific Islander  1.12e-01   4.00e-02    2.79  0.00525 **
raceethWhite                              1.28e-01   2.72e-02    4.70  2.7e-06 ***
```

```
expectBachelors1            1.12e-01   7.18e-03   15.64  < 2e-16 ***
motherBachelors1            1.79e-02   6.51e-03    2.76  0.00587 **
fatherHS1                   1.97e-02   8.42e-03    2.34  0.01922 *
fatherBachelors1            3.28e-02   6.77e-03    4.84  1.3e-06 ***
computerForSchoolwork1      4.05e-02   9.67e-03    4.19  2.8e-05 ***
read30MinsADay1             6.45e-02   5.74e-03   11.24  < 2e-16 ***
publicSchool1              -3.66e-02   1.01e-02   -3.61  0.00030 ***
schoolSize                  7.73e-05   5.67e-05    1.36  0.17298
schoolSizeSQ               -3.83e-06   2.19e-06   -1.75  0.08082 .
gradeSQ                    -3.50e-02   5.91e-03   -5.93  3.4e-09 ***
gdschsize                  -4.97e-06   5.58e-06   -0.89  0.37354
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.149 on 3384 degrees of freedom
Multiple R-squared:  0.318,    Adjusted R-squared:  0.314
F-statistic: 82.9 on 19 and 3384 DF,  p-value: <2e-16
```

i. Evaluate your final model as if for a data scientist.
j. Write a summary as if for a layman

First, we could check the F-test, it looks good, something in the model appears to be working. Adjusted R-squared is 31.4% which means 31.4% of variability of reading score is explained by the model.
Then look at the t-test of each variable which is good enough to keep them in the model.