6315305_assignment_prediction_model

Andreas Sakapetis

January 18, 2019

Loading neccessary libraries

```
#loading the library "readxl" to be able to use the function read_xlsx().
library(readxl)
#loading the library "formattable" to introduce a table with all the variables.
library(formattable)
#loading the library "ggplot2" to create usefull plots.
library(ggplot2)
#loading the library "GGally" to create usefull plots.
library(GGally)
#loading the library "tidyverse" to perform varius manipulations.
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.1
```

```
## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.2     v dplyr   0.7.8
## v readr   1.1.1     v stringr 1.3.1
## v tibble  1.4.2     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------- tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#loading the library "car" to perform VIF.
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.2
```

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

```
## Warning: package 'carData' was built under R version 3.5.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
#loading the library "leaps" to perform subset selection.
```

Import the data from local folder to Rstudio in p.student variable. ___The dataset can be found from kaggle(https://www.kaggle.com/spscientist/students-performance-in-exams/version/1)___

```
p.student <-
  read.csv("C:\\Users\\Andreas\\Desktop\\Assignment B\\StudentsPerformance.csv",
           header = TRUE)
```

View the names of the columns in the p.student dataset.

```
names(p.student)
```

```
## [1] "gender"                       "race.ethnicity"
## [3] "parental.level.of.education" "lunch"
## [5] "test.preparation.course"      "math.score"
## [7] "reading.score"                "writing.score"
```

Use the head() function to examine at the first few rows of the p.student dataset.

```
head(p.student)
```

```
##   gender race.ethnicity parental.level.of.education        lunch
## 1 female        group B           bachelor's degree     standard
## 2 female        group C                some college     standard
## 3 female        group B             master's degree     standard
## 4   male        group A          associate's degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B          associate's degree     standard
##   test.preparation.course math.score reading.score writing.score
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

Use the function tail() to examine the last few rows of the p.student dataset.

```
tail(p.student)
```

```
##      gender race.ethnicity parental.level.of.education        lunch
## 995    male        group A                high school     standard
## 996  female        group E             master's degree     standard
## 997    male        group C                high school free/reduced
## 998  female        group C                high school free/reduced
## 999  female        group D               some college     standard
## 1000 female        group D               some college free/reduced
##      test.preparation.course math.score reading.score writing.score
## 995                     none         63            63            62
## 996                completed         88            99            95
## 997                     none         62            55            55
## 998                completed         59            71            65
## 999                completed         68            78            77
## 1000                    none         77            86            86
```

This data set includes scores from three exams and a variety of personal, social, and economic factors that have interaction effects upon them.The dataset consists of 8 variables, with a total of 1000 observations. There are no missing values. Within the dataset exist variables that with the type being either an interger or a factor.

In the study sample, 482/1000 (48.2%) of the students are male 518/1000 (51.8%) are female. The mean math.score is 66.089 points with a standard deviation of 15.16 points (range 0-100 points). The mean of reading.score is 69.169 points with a standard deviation of 14 points (range 0-100 points).

Approximately 31.9% of the sudents are enthnicity group C, 26.2% are group D, 19% are group B, 14% are group E and 8.9% group A. 64.2% of the students did none preparation and 35.8% did a complete preparation.

64.5 % of the students had a standard lunch, while 35.5% of the students had free or reduced lunch. 22.6% of the stundets has parents that acquired some college education, 22.2% associate's degree education, 19.6% of the students have parents with high scool education and 17.9% some high scool education. 11.8% of the students have parents with a bachelor's degree and the most scarce of all 5.9% of the students have parents with a master's degree.

This function counts missing values

```
sapply(p.student, function(x) sum(length(which(is.na(x)))))
```

```
##                       gender             race.ethnicity
##                            0                          0
## parental.level.of.education                      lunch
##                            0                          0
##     test.preparation.course                 math.score
##                            0                          0
##               reading.score               writing.score
##                            0                          0
```

With sapply function it is possible to examine the type of the variables.

```
sapply(p.student,class)
```

```
##                       gender             race.ethnicity
##                     "factor"                   "factor"
## parental.level.of.education                      lunch
##                     "factor"                   "factor"
##     test.preparation.course                 math.score
##                     "factor"                  "integer"
##               reading.score               writing.score
##                    "integer"                  "integer"
```

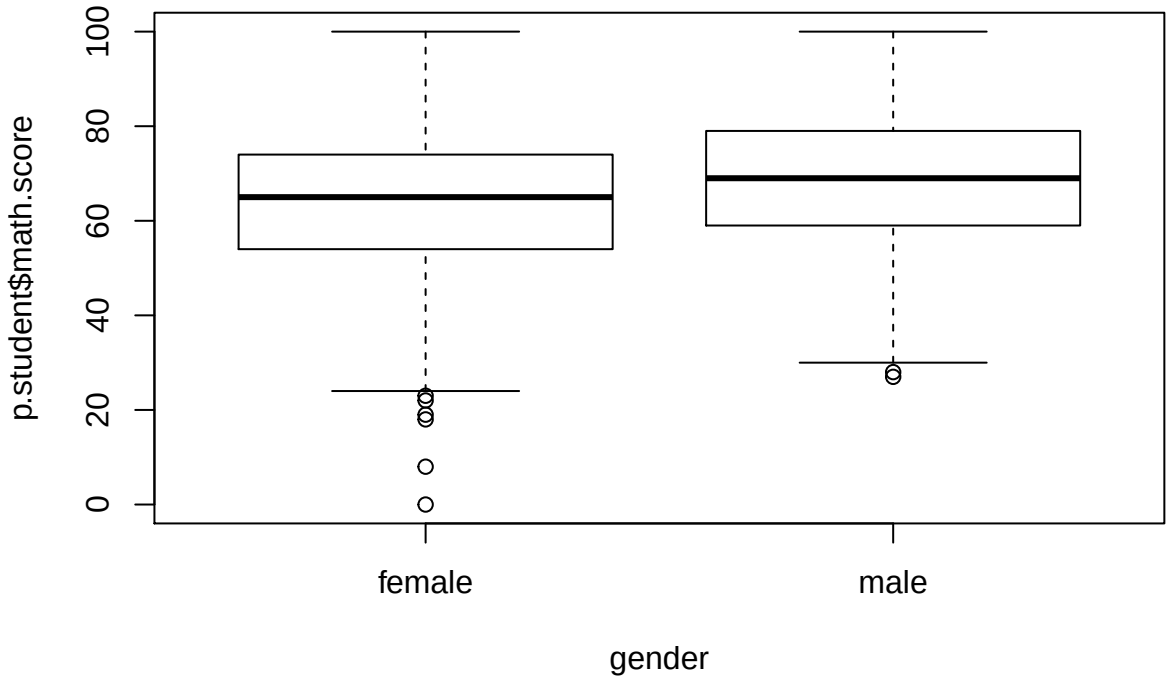
With the summary fucntion, the summary statistics for the p.student dataset are obtained.

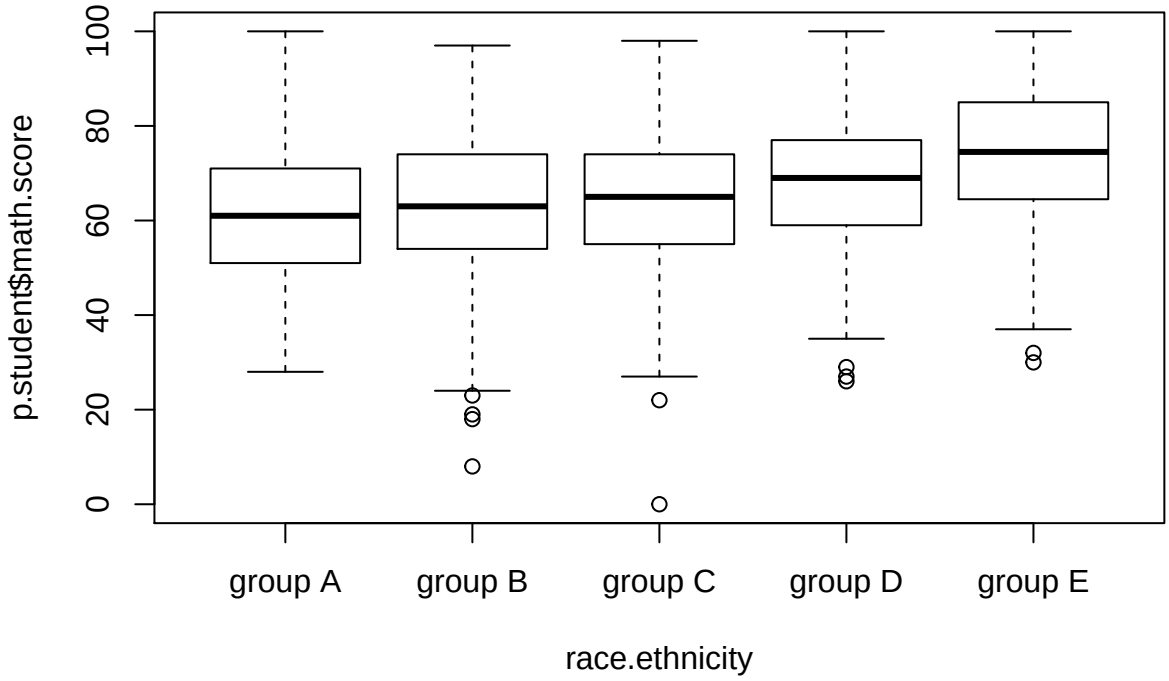
```
summary(p.student)
```

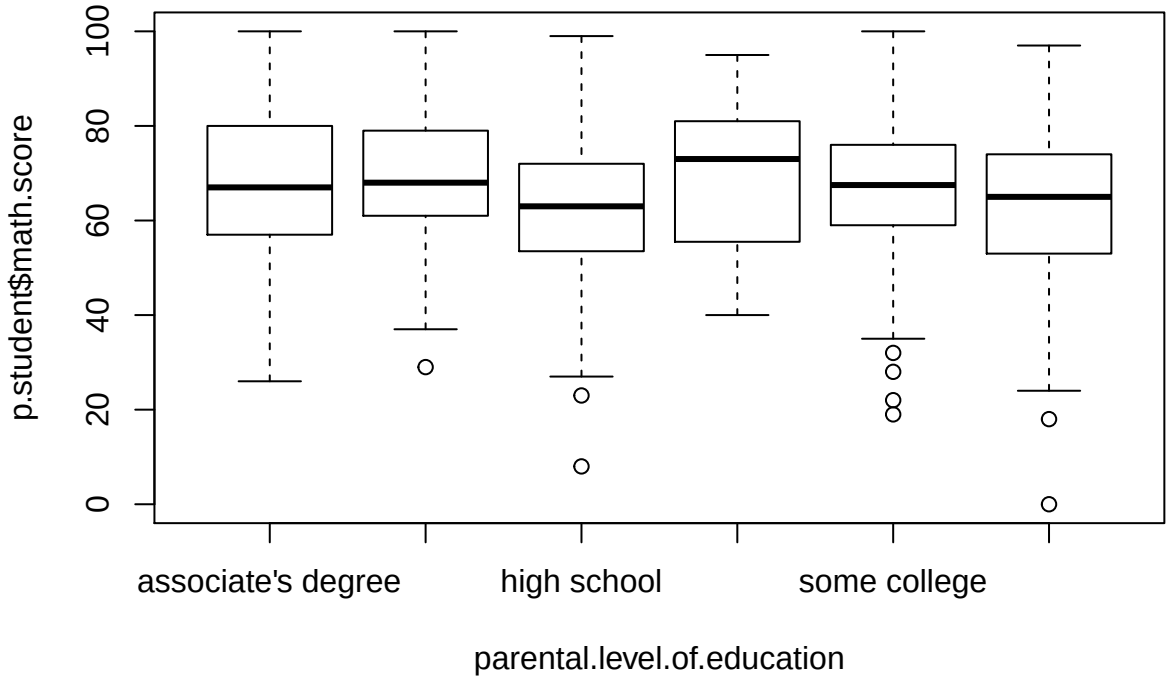
```
##     gender    race.ethnicity     parental.level.of.education
##  female:518   group A: 89    associate's degree:222
##  male  :482   group B:190    bachelor's degree :118
##               group C:319    high school       :196
##               group D:262    master's degree   : 59
##               group E:140    some college      :226
##                              some high school  :179
##           lunch     test.preparation.course   math.score
##  free/reduced:355   completed:358           Min.   :  0.00
##  standard    :645   none     :642           1st Qu.: 57.00
##                                             Median : 66.00
##                                             Mean   : 66.09
##                                             3rd Qu.: 77.00
##                                             Max.   :100.00
##  reading.score    writing.score
##  Min.   : 17.00   Min.   : 10.00
##  1st Qu.: 59.00   1st Qu.: 57.75
##  Median : 70.00   Median : 69.00
##  Mean   : 69.17   Mean   : 68.05
##  3rd Qu.: 79.00   3rd Qu.: 79.00
##  Max.   :100.00   Max.   :100.00
```

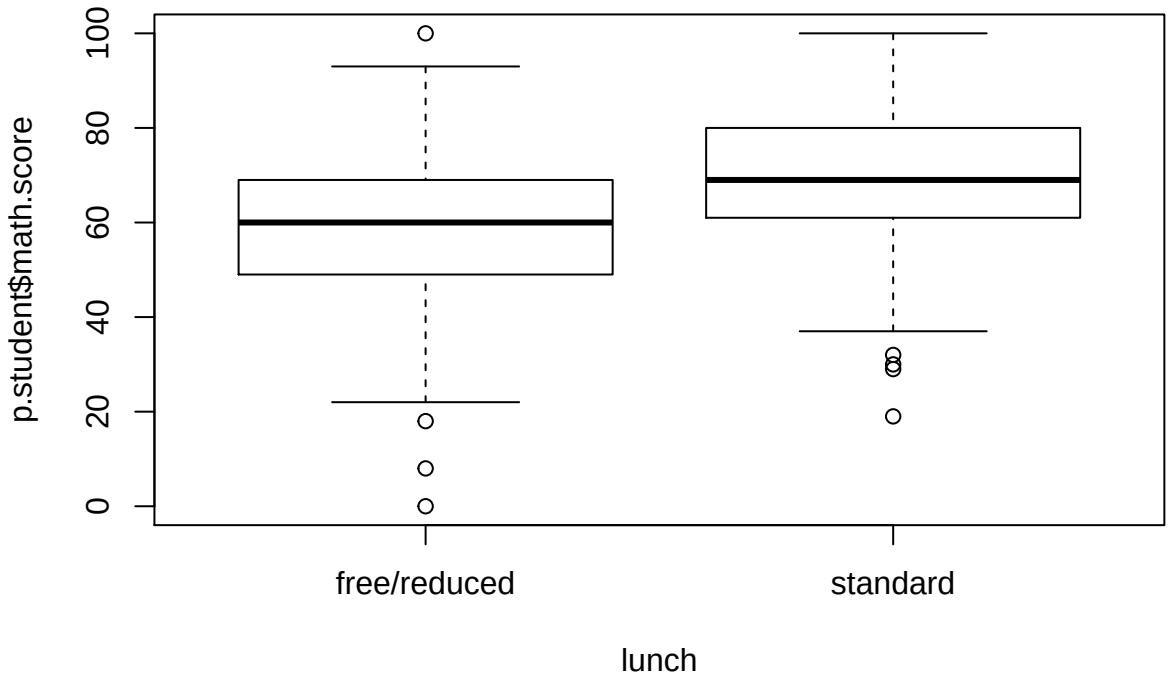
These plots are assisting in identifying trends in the data, using as response variable "Math score".

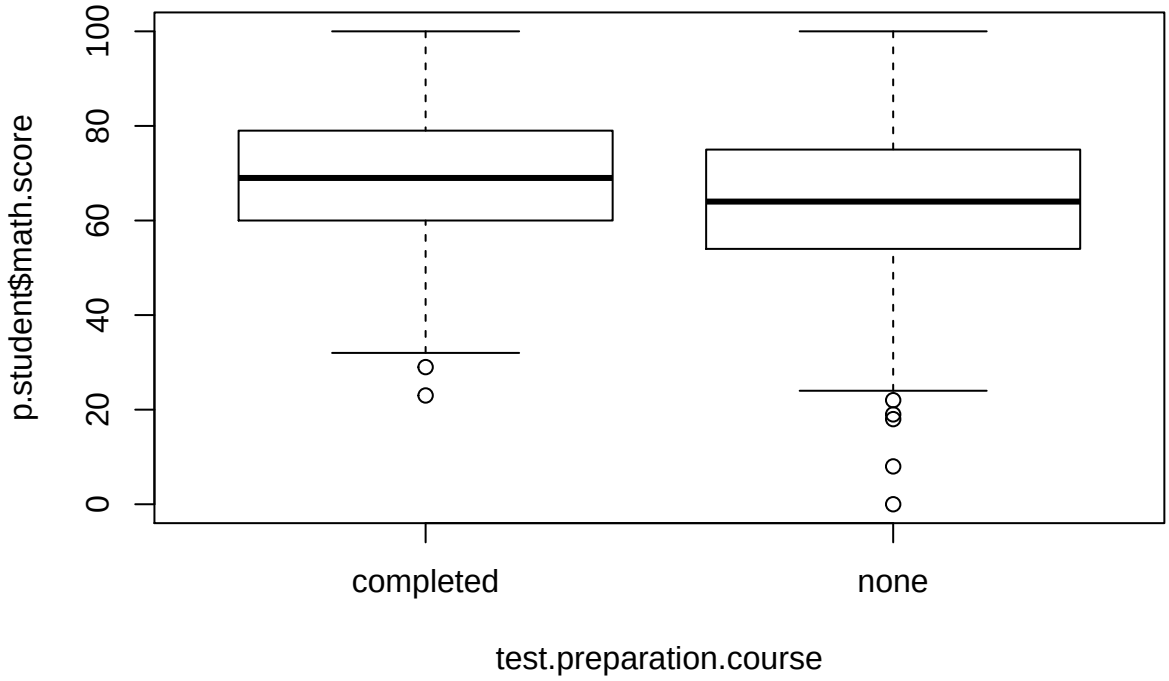
```
plot(p.student$math.score~.,data=p.student)
```

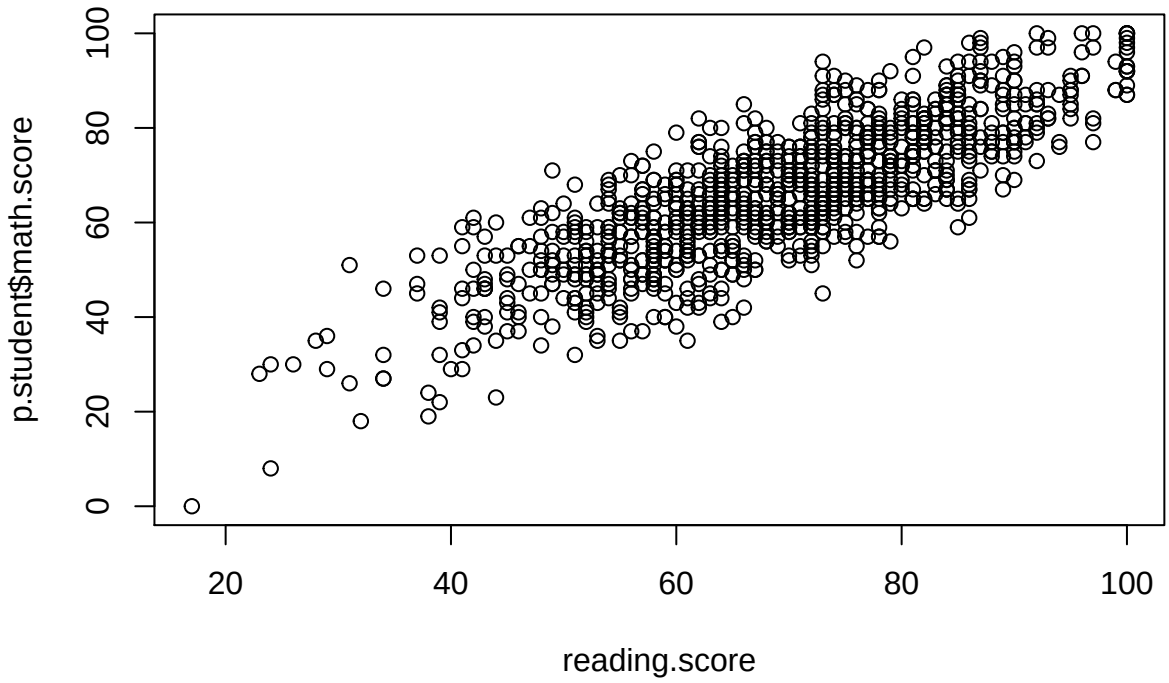


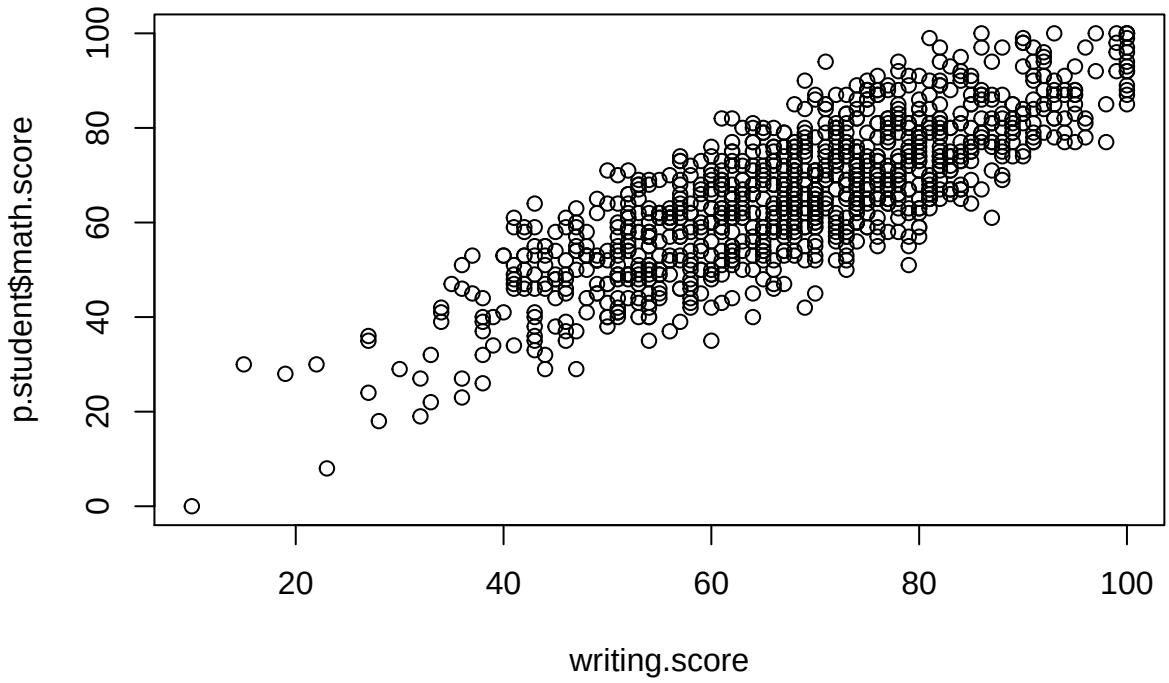

p.student$math.score
0
20
40
60
80
100
group A
group B
group C
group D
group E
race.ethnicity


p.student$math.score
0
20
40
60
80
100
associate's degree
high school
some college
parental.level.of.education


p.student$math.score
0
20
40
60
80
100
free/reduced
standard
lunch

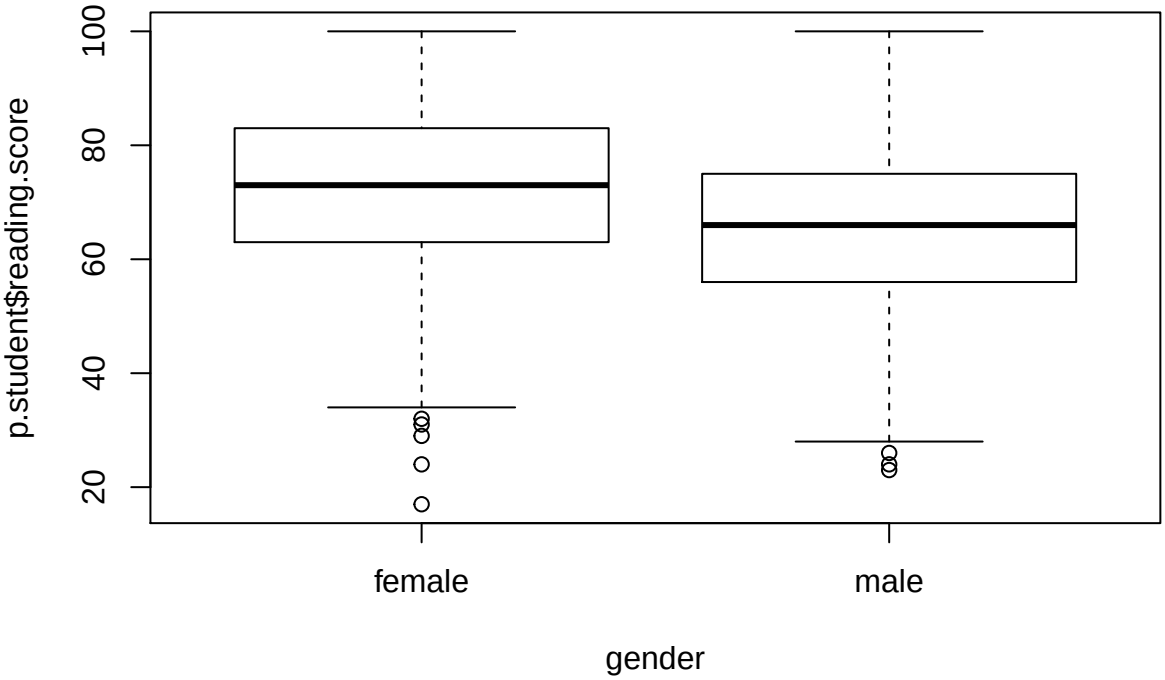

p.student$math.score
0
20
40
60
80
100
completed
none
test.preparation.course

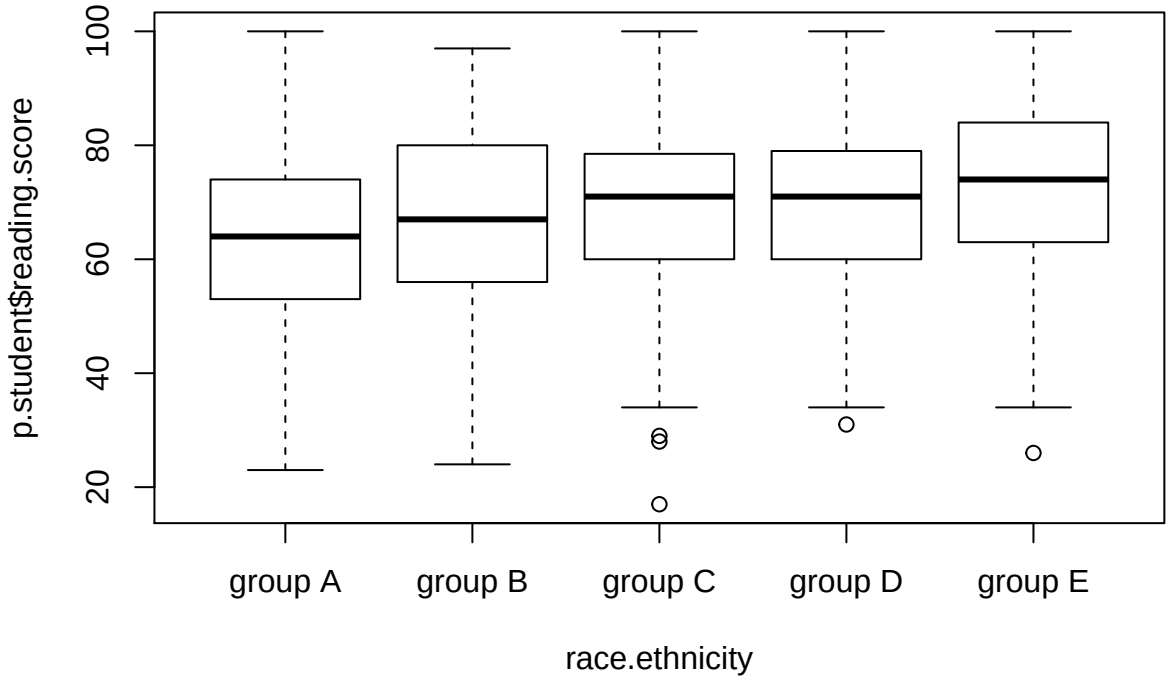

p.student$math.score
0
20
40
60
80
100
20
40
60
80
100
reading.score

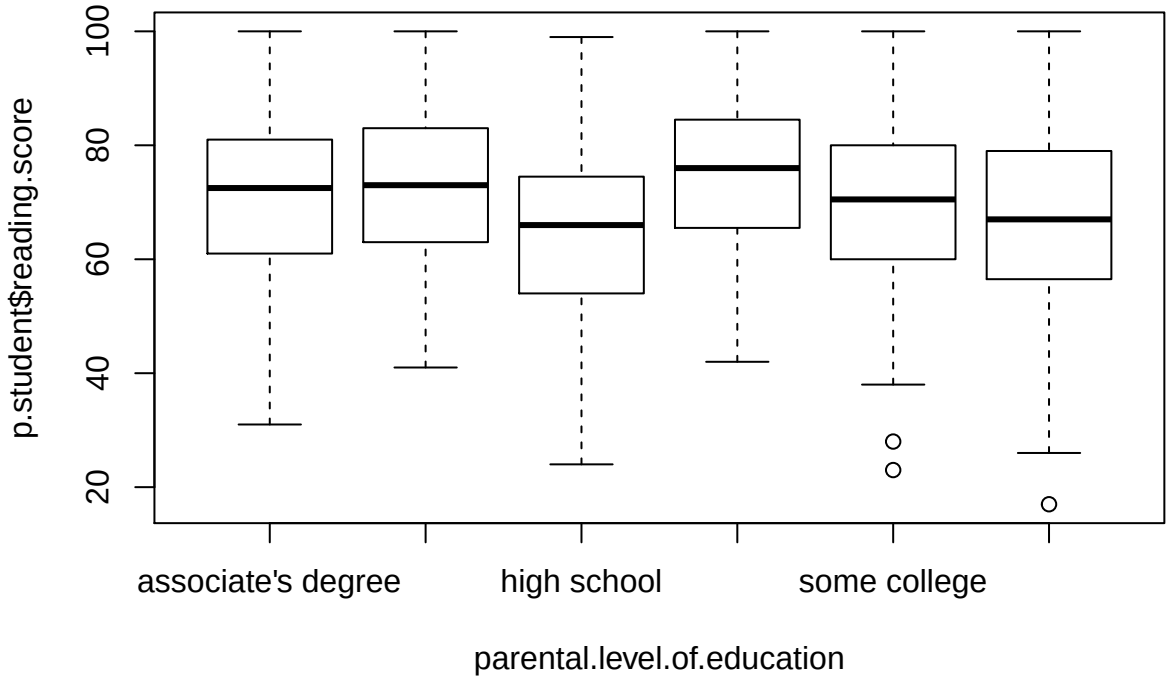


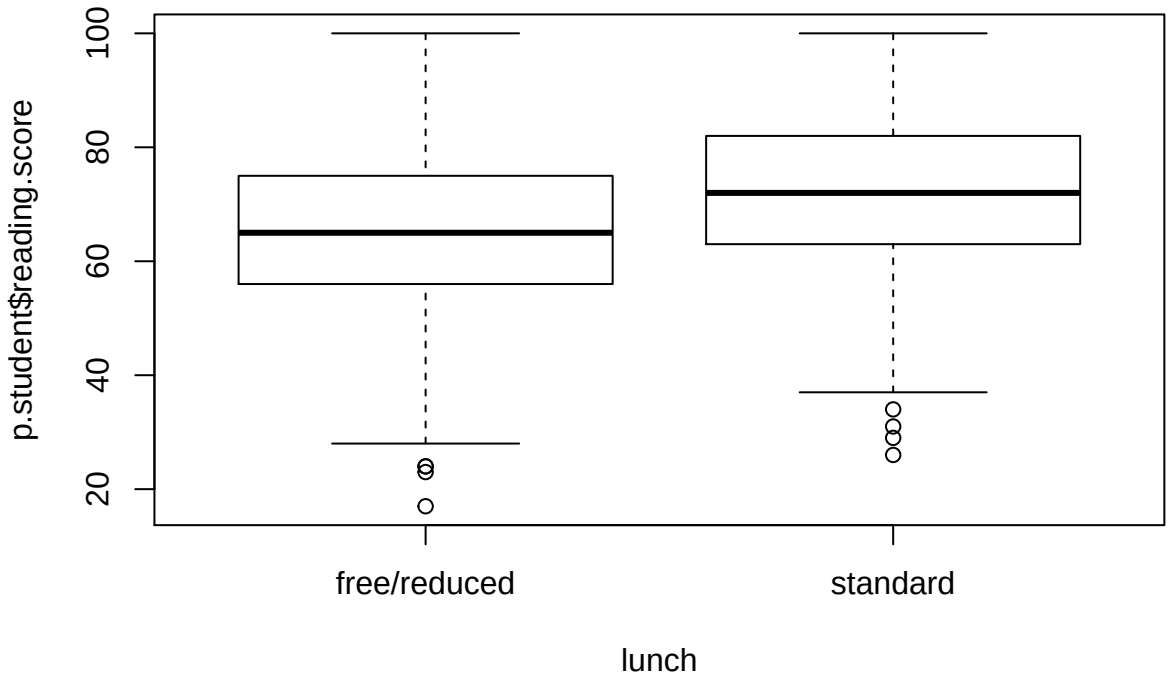
These plots are in identifying trends in the data, using as response variable "Reading score".

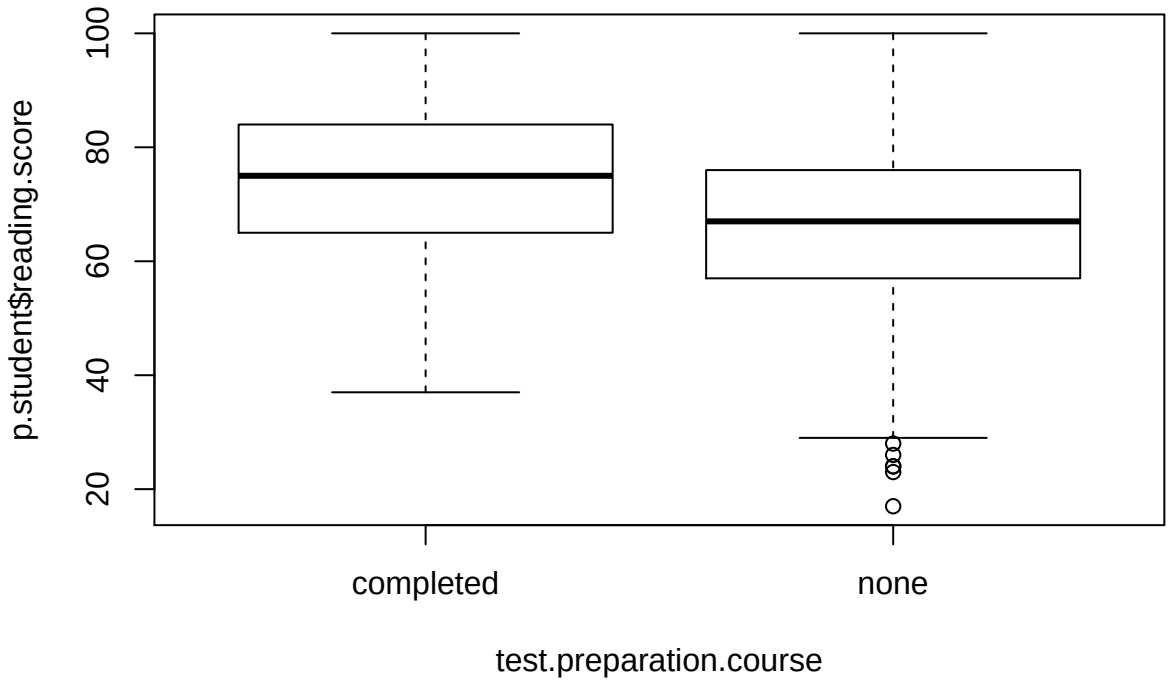
```
plot(p.student$reading.score~.,data=p.student)
```

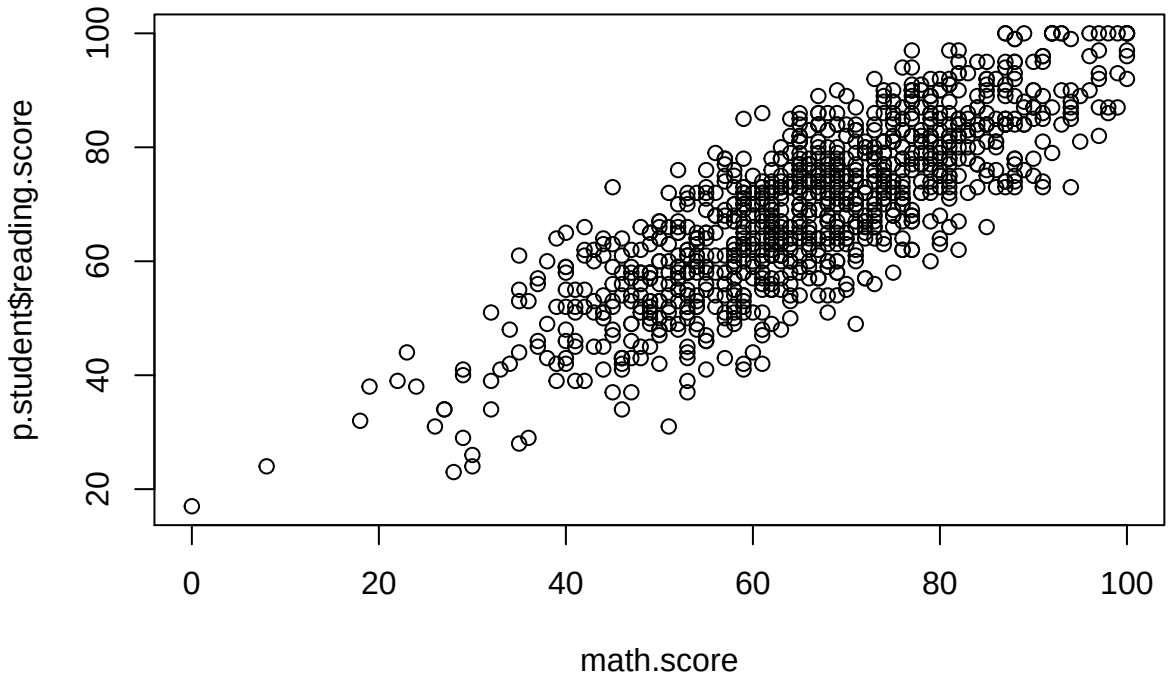

p.student$reading.score
20
40
60
80
100
female
male
gender

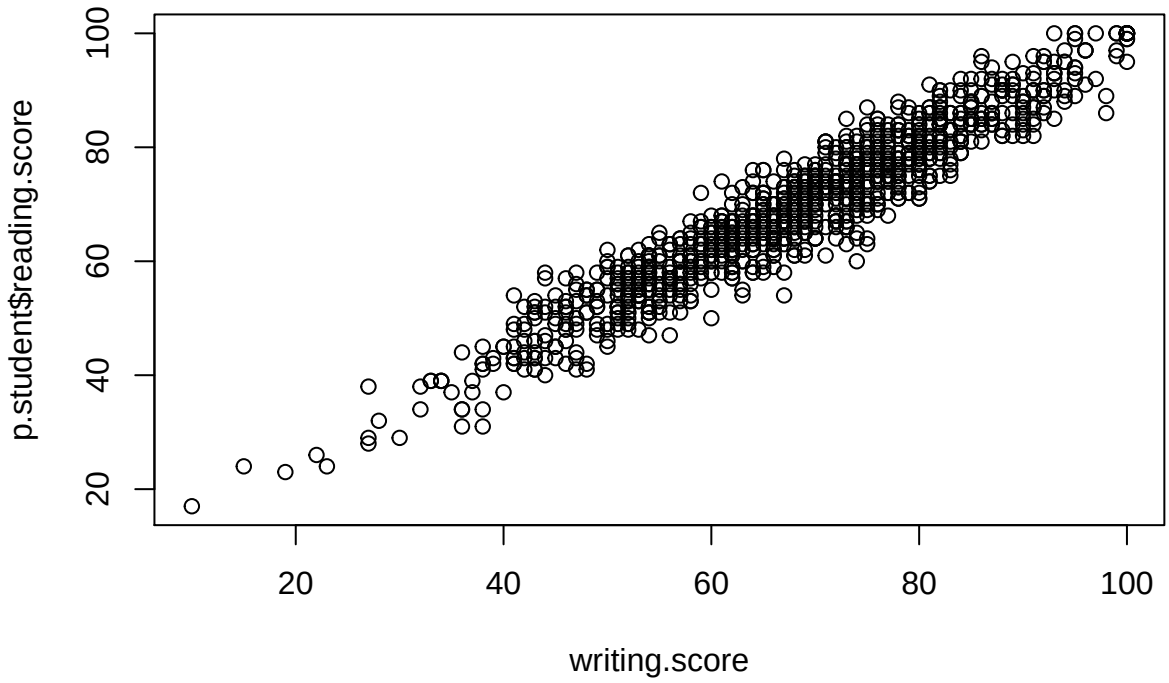

p.student$reading.score
20
40
60
80
100
group A
group B
group C
group D
group E
race.ethnicity


p.student$reading.score
20
40
60
80
100
associate's degree
high school
some college
parental.level.of.education


p.student$reading.score
20
40
60
80
100
free/reduced
standard
lunch

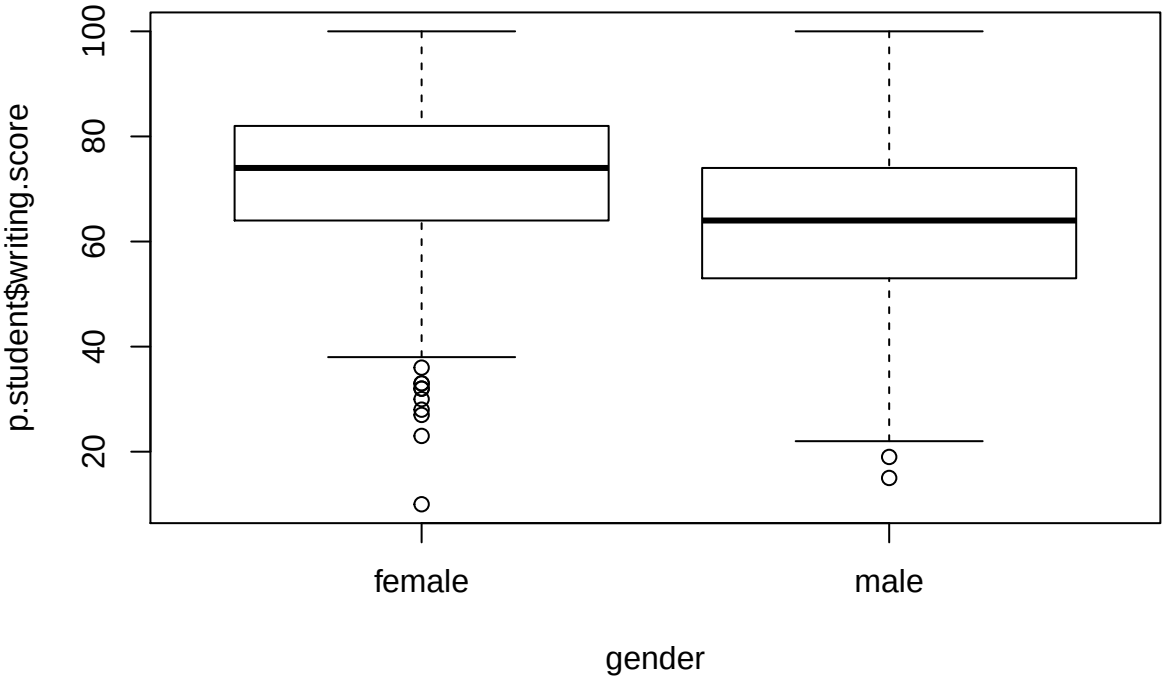

p.student$reading.score
20
40
60
80
100
completed
none
test.preparation.course

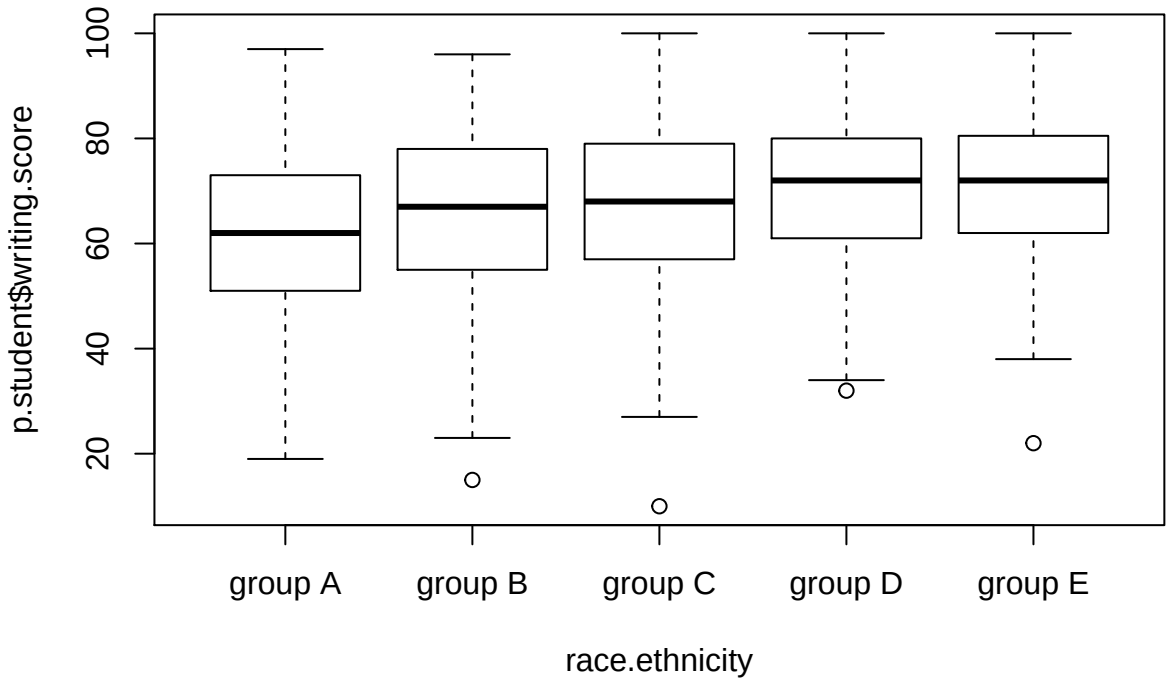

p.student$reading.score
20
40
60
80
100
0
20
40
60
80
100
math.score

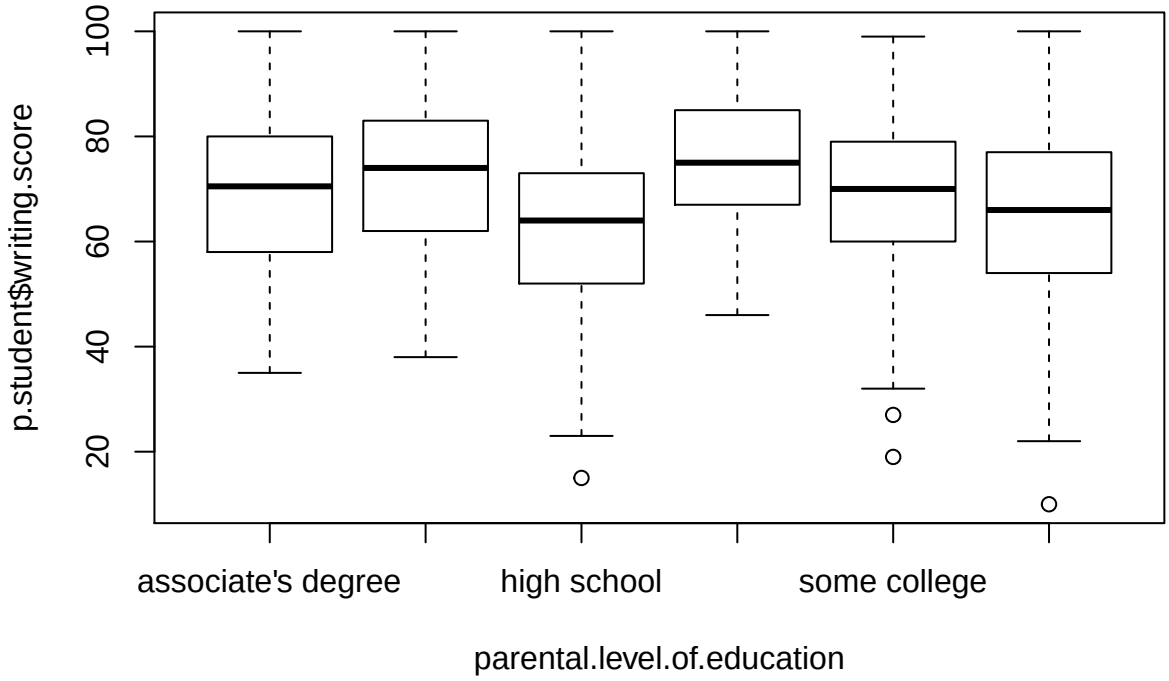


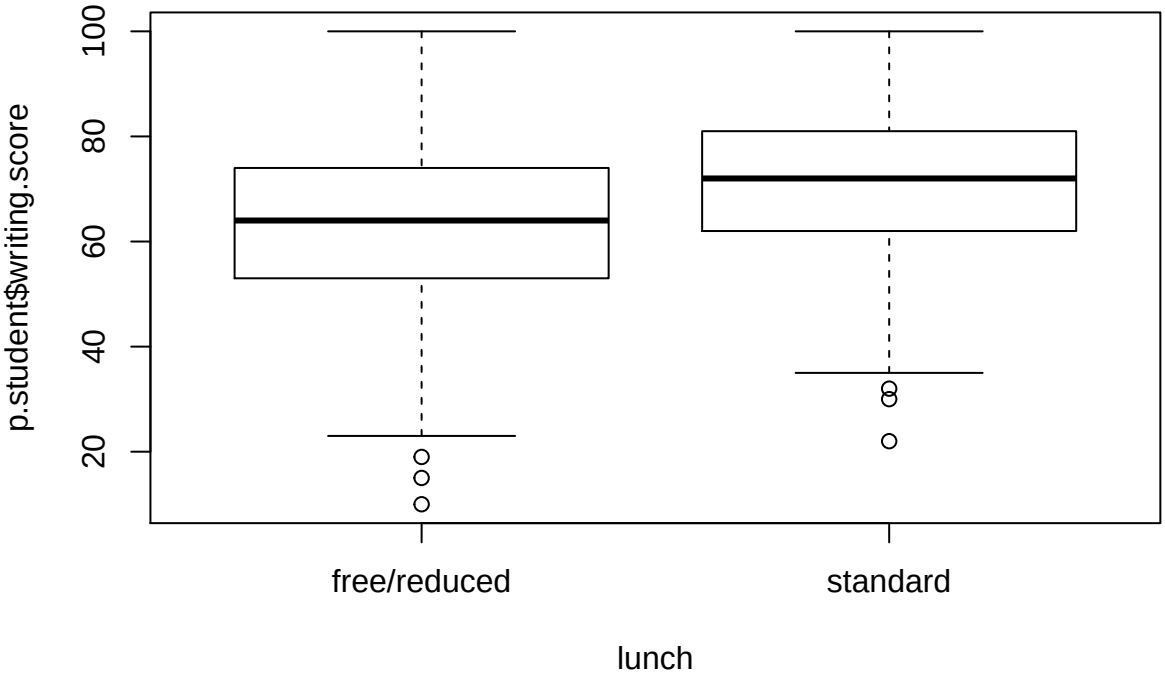
These plots arein identifying trends in the data, using as response variable "Writing score".

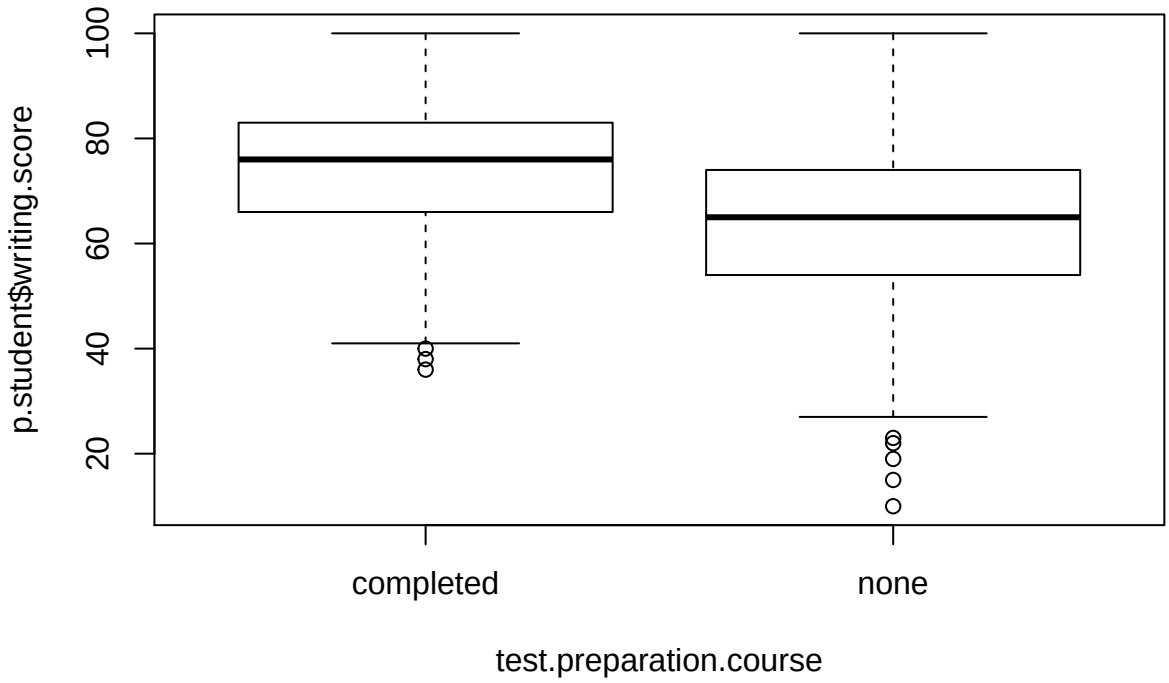
```
plot(p.student$writing.score~.,data=p.student)
```

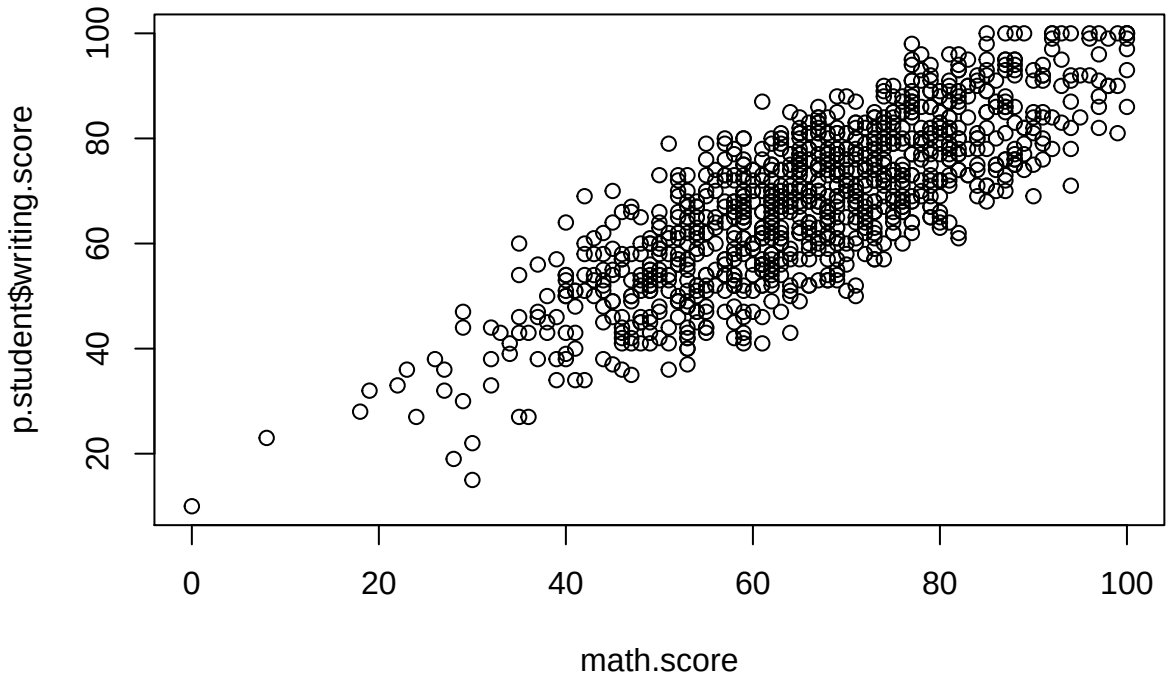

p.students$writing.score
20
40
60
80
100
female
male
gender

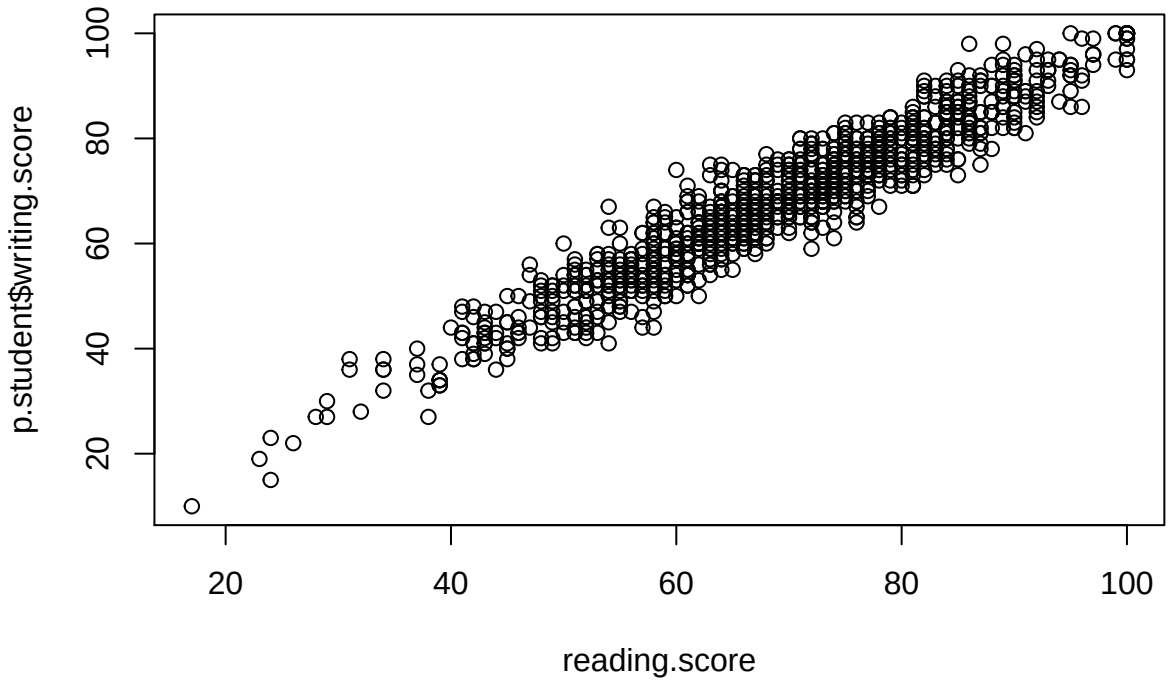

p.students$writing.score
20
40
60
80
100
group A
group B
group C
group D
group E
race.ethnicity


p.students$writing.score
20
40
60
80
100
associate's degree
high school
some college
parental.level.of.education


p.students$writing.score
20
40
60
80
100
free/reduced
standard
lunch


p.students$writing.score
20
40
60
80
100
completed
none
test.preparation.course

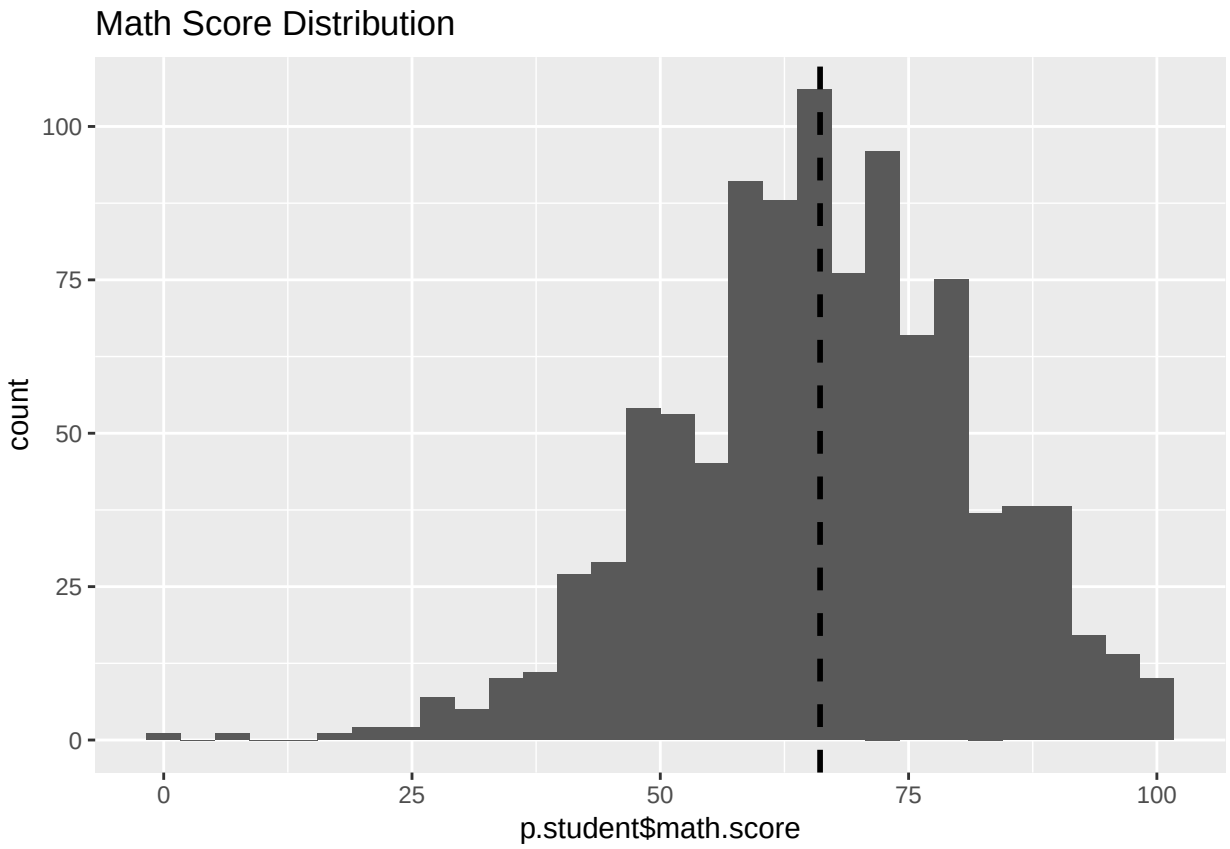

p.student$writing.score
20
40
60
80
100
0
20
40
60
80
100
math.score



From the scatterplots we can see that the variables reading.score, writing.score and math.score appear to have a linear relationship.

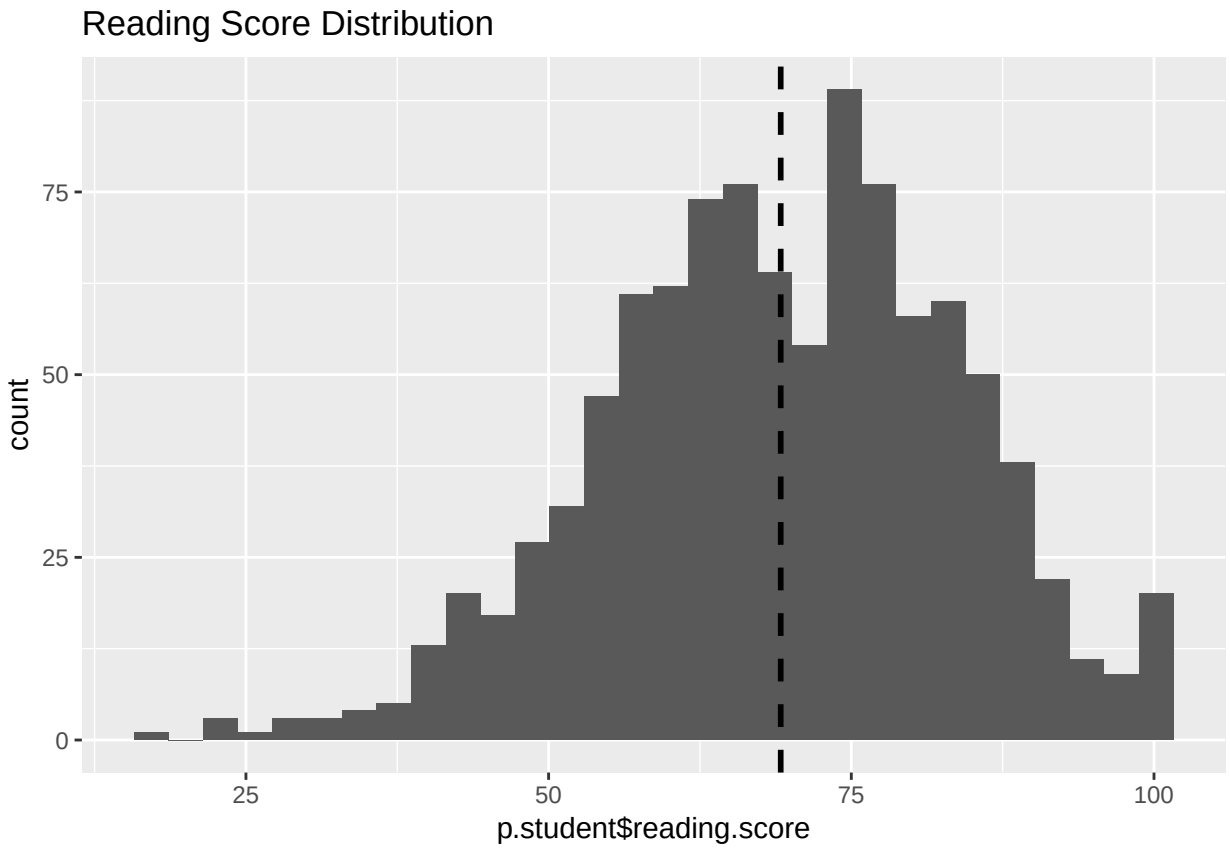
```
ggplot(p.student)+
geom_histogram(mapping = aes(x = p.student$math.score))+
geom_vline(xintercept=mean(p.student$math.score), lwd=1, linetype=2, color="black")+
  ggtitle("Math Score Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



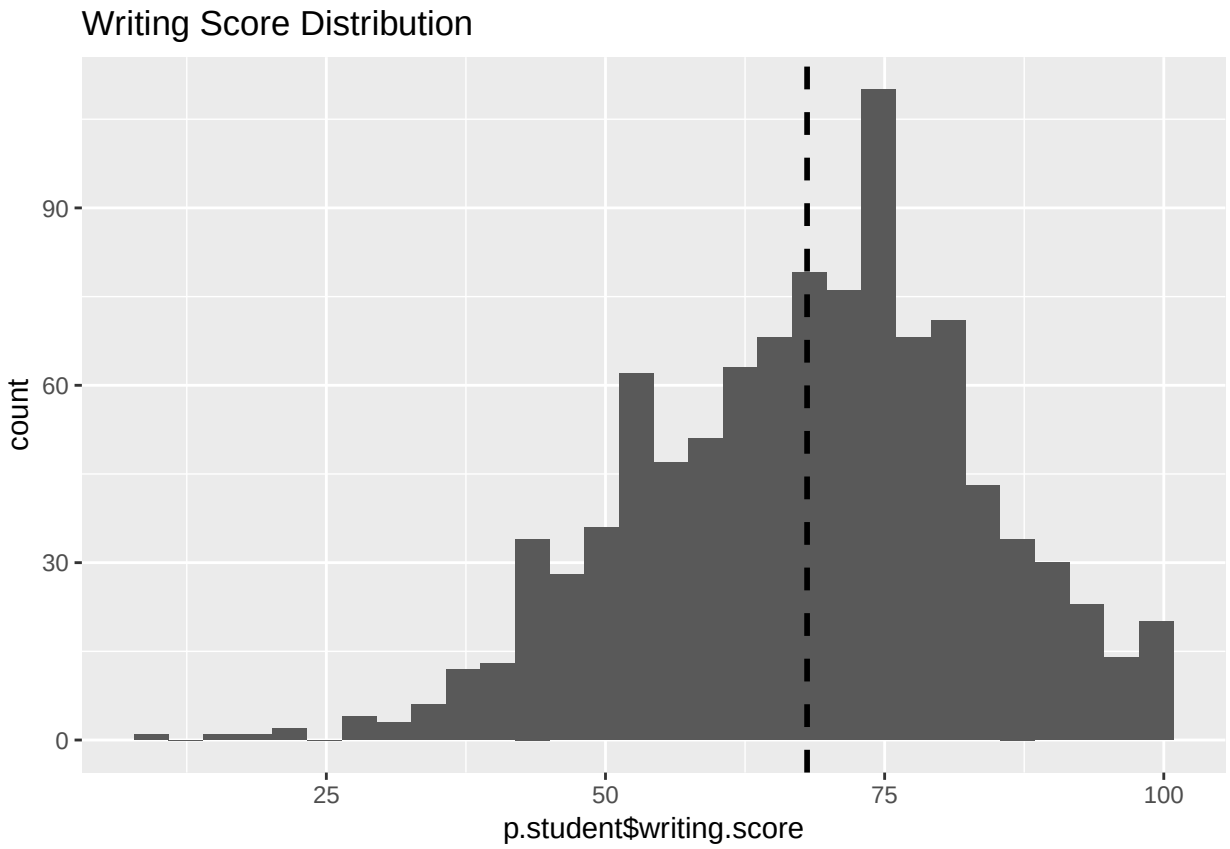

```
ggplot(p.student)+
geom_histogram(mapping = aes(x = p.student$reading.score))+
geom_vline(xintercept=mean(p.student$reading.score), lwd=1, linetype=2, color="black")+
  ggtitle("Reading Score Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

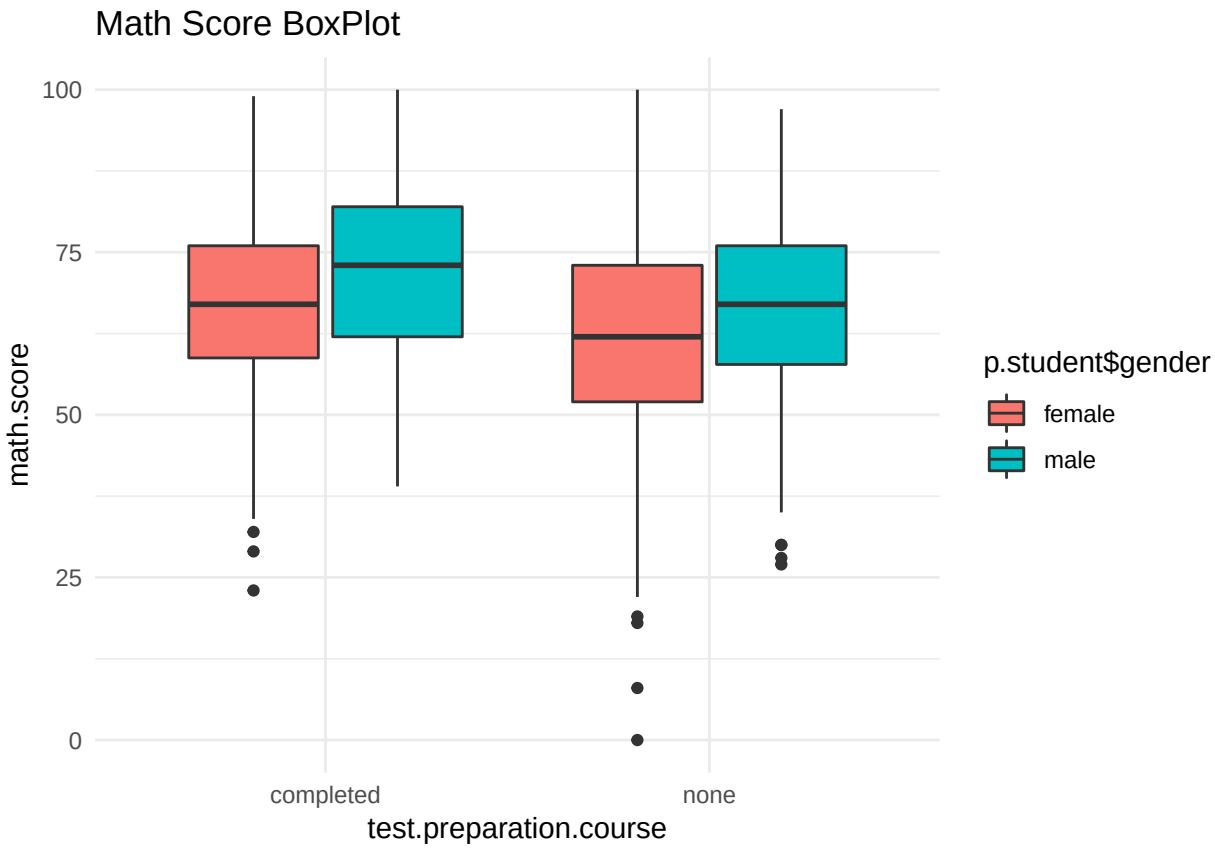



```
ggplot(p.student)+
geom_histogram(mapping = aes(x = p.student$writing.score))+
geom_vline(xintercept=mean(p.student$writing.score), lwd=1, linetype=2, color="black")+
  ggtitle("Writing Score Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

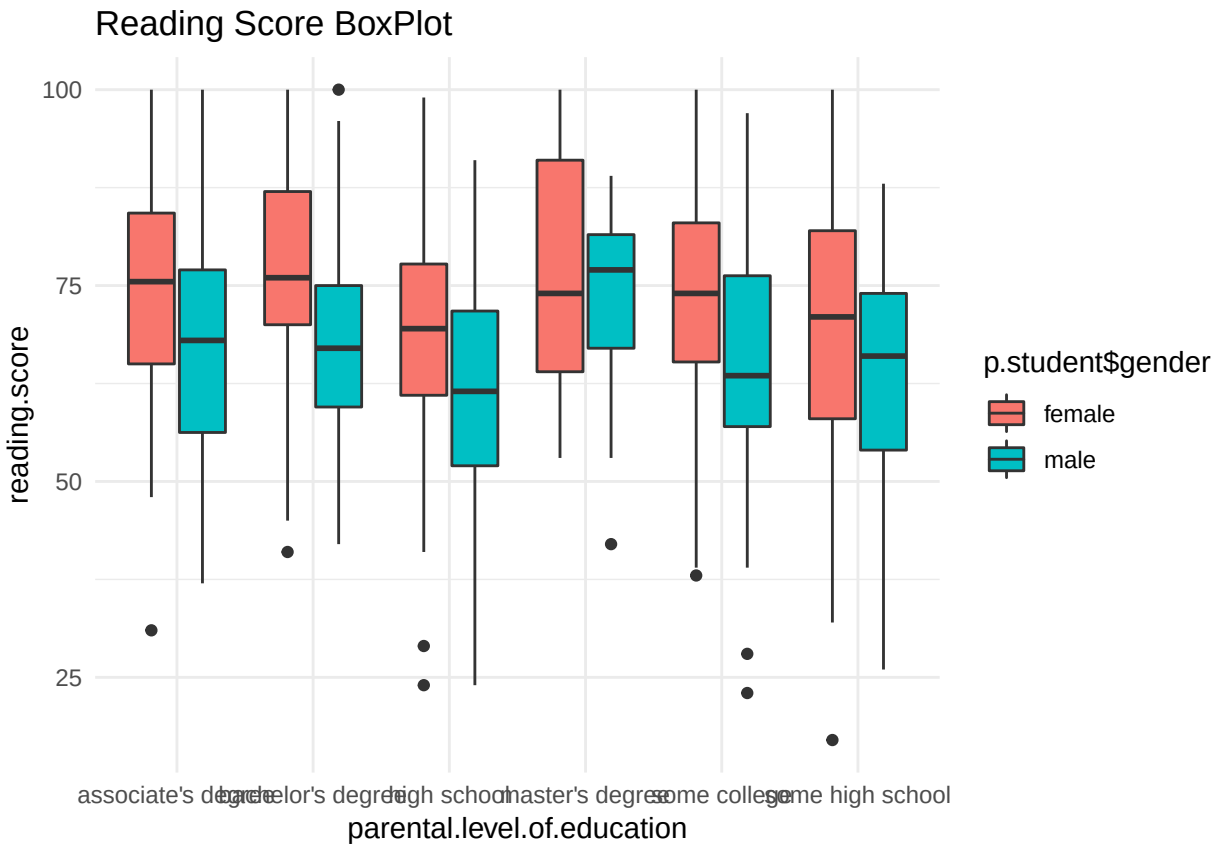



```
p.student %>%
  ggplot(aes(x = test.preparation.course , y = math.score, fill = p.student$gender)) +
  geom_boxplot() +
  theme_minimal()+
  ggtitle("Math Score BoxPlot")
```

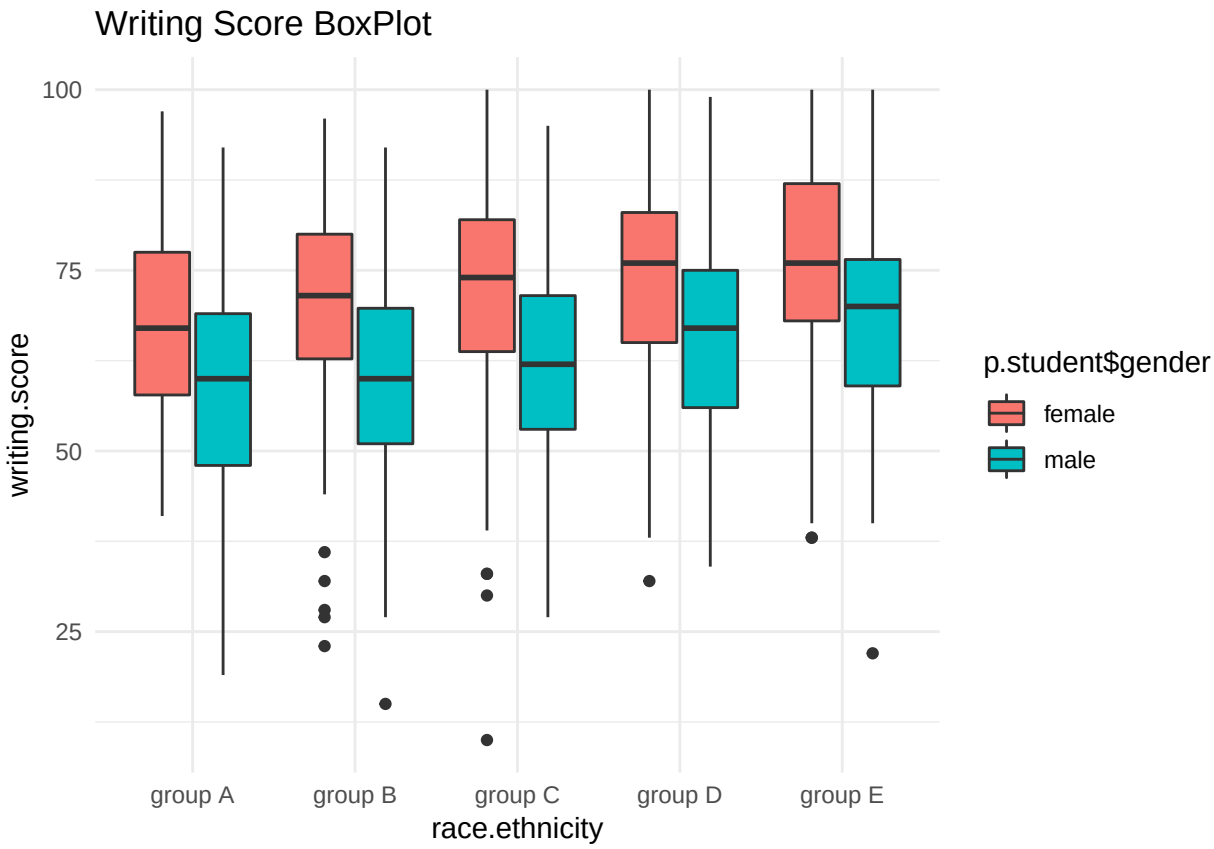


From the boxplot examination females seem to do less better than males on math. From the plot completed preparation is concentrating better scores in math than none preperation. Outliers are spotted even though from visual examination of the dataset only a single value was "0", which can be interpreated as poor performace. The outliers are more spread for females and more concentrated for males.

```
p.student %>%
  ggplot(aes(x = parental.level.of.education , y = reading.score, fill = p.student$gender)) +
  geom_boxplot() +
  theme_minimal()+
  ggtitle("Reading Score BoxPlot")
```




```
p.student %>%
  ggplot(aes(x = race.ethnicity , y = writing.score, fill = p.student$gender)) +
  geom_boxplot() +
  theme_minimal()+
  ggtitle("Writing Score BoxPlot")
```



Preparing the dataset to execute the multiple linear regression.

```
#Splitting the dataset into a training(75%) and testing set(25%).

#Create a variable "split" that contains 750 times the word train and 250 the
#word test.
split <- c(rep("train", 750), rep("test",  250))

#Divide the dataset based on the words train and test.
p.student <- p.student %>% mutate(split = sample(split))

#Delete the column split
p.student_train <- p.student %>% filter(split == "train")
p.student_train$split <-  NULL

#Delete the column split
p.student_test <- p.student %>% filter(split == "test")
p.student_test$split <-  NULL
```

Executing the multiple linear regression.

```
#The multiple regression model is selected since the input variable is quantitative.
model1 <- lm(formula=p.student_train$math.score ~ . , data=p.student_train)
#Aquire the summary statistics of the multiple linear regression.
summary(model1)
```

```
##
## Call:
```

```
## lm(formula = p.student_train$math.score ~ ., data = p.student_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.106  -3.608   0.137   3.579  14.310
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                                -11.87135    1.45361  -8.167
## gendermale                                  13.34126    0.43093  30.959
## race.ethnicitygroup B                        0.86157    0.77571   1.111
## race.ethnicitygroup C                       -0.05201    0.73209  -0.071
## race.ethnicitygroup D                        0.12267    0.75908   0.162
## race.ethnicitygroup E                        4.87210    0.84857   5.742
## parental.level.of.educationbachelor's degree -0.42444    0.71061  -0.597
## parental.level.of.educationhigh school       0.37325    0.63288   0.590
## parental.level.of.educationmaster's degree  -1.96025    0.95000  -2.063
## parental.level.of.educationsome college      0.97805    0.59079   1.656
## parental.level.of.educationsome high school  0.88949    0.63992   1.390
## lunchstandard                                2.94613    0.43579   6.760
## test.preparation.coursenone                  3.64031    0.46005   7.913
## reading.score                                0.24053    0.04930   4.879
## writing.score                                0.72639    0.05111  14.213
##                                             Pr(>|t|)
## (Intercept)                                 1.38e-15 ***
## gendermale                                   < 2e-16 ***
## race.ethnicitygroup B                          0.2671
## race.ethnicitygroup C                          0.9434
## race.ethnicitygroup D                          0.8717
## race.ethnicitygroup E                       1.37e-08 ***
## parental.level.of.educationbachelor's degree   0.5505
## parental.level.of.educationhigh school         0.5555
## parental.level.of.educationmaster's degree     0.0394 *
## parental.level.of.educationsome college        0.0982 .
## parental.level.of.educationsome high school    0.1649
## lunchstandard                               2.80e-11 ***
## test.preparation.coursenone                 9.25e-15 ***
## reading.score                               1.31e-06 ***
## writing.score                                < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.44 on 735 degrees of freedom
## Multiple R-squared:  0.8748, Adjusted R-squared:  0.8724
## F-statistic: 366.8 on 14 and 735 DF,  p-value: < 2.2e-16
```

As it is possible to examine from the summary(), regression coefficient gender for male is associated with an increase of 13 points on math.score against females. The regression coefficient Race ethnicity E has also significant association with an increase of 5 points on math.score against Race ethnicity A.

However,the regression coefficient race.ethnicities B,C and D do not reach statistical significance. Same applies for the regression coefficient parental.level.of.education for the education levels of: some high school,some college,master's degree high school. On the contrast regression coefficient parental.level.of.education bachelor's degree is associated with a decrease of

1.7 points on math.score against asscociate's degree.

The regression coefficient Lunch is associated with an increase of 3.3 points on math.score against free/reduced. The regression coefficient Reading.score is associated with an increase of 0.22 points on math.score. The regression coefficient writing.score is associated with an increas of 0.71 on math.score.

The regression coefficient test.preparation none is associated with an increase of 3.4 against completed preparation.

The multiple regression estimates β_0 (the intercept) and β_1,β_2,β_3,β_4,β_5,β_6 and β_7 in the following equation:

$$\boldsymbol{y} = \beta_0 + \beta_1 \cdot \boldsymbol{x}_1 + \beta_2 \cdot \boldsymbol{x}_2 + \beta_3 \cdot \boldsymbol{x}_3 + \beta_4 \cdot \boldsymbol{x}_4 + \beta_5 \cdot \boldsymbol{x}_5 + \beta_6 \cdot \boldsymbol{x}_6 + \beta_7 \cdot \boldsymbol{x}_7 + \boldsymbol{\epsilon}$$

Through the examination of the multiple regression analysis. The Fstatistic p-value is < 2.2e-16. This leads to the conclusion that, there is at least, one predictor variable which is significantly related to the outcome variable.

The adjusted R-squared indicates that 87% of the variation in math score can be explained by the model containing gender,race.ethnicity,parental.level.of.education, lunch, test.preparation, reading.score and writing score.

To examine which predictor variables are significant, we require estimate of regression beta coefficients and the associated t-statitic p-values:

```
summary(model1)$coefficient
```

```
##                                              Estimate Std. Error
## (Intercept)                                -11.87135162 1.45360940
## gendermale                                  13.34125523 0.43092954
## race.ethnicitygroup B                        0.86156949 0.77571497
## race.ethnicitygroup C                       -0.05201041 0.73209331
## race.ethnicitygroup D                        0.12266509 0.75907816
## race.ethnicitygroup E                        4.87210175 0.84857415
## parental.level.of.educationbachelor's degree -0.42444395 0.71060967
## parental.level.of.educationhigh school       0.37324846 0.63288353
## parental.level.of.educationmaster's degree  -1.96025175 0.94999616
## parental.level.of.educationsome college      0.97805458 0.59078656
## parental.level.of.educationsome high school  0.88949056 0.63991852
## lunchstandard                                2.94613290 0.43578765
## test.preparation.coursenone                  3.64030889 0.46004809
## reading.score                                0.24052975 0.04929586
## writing.score                                0.72639051 0.05110779
##                                                t value      Pr(>|t|)
## (Intercept)                                -8.16680988  1.377855e-15
## gendermale                                 30.95924940 2.377810e-135
## race.ethnicitygroup B                       1.11067791  2.670702e-01
## race.ethnicitygroup C                      -0.07104342  9.433825e-01
## race.ethnicitygroup D                       0.16159744  8.716673e-01
## race.ethnicitygroup E                       5.74151566  1.372918e-08
## parental.level.of.educationbachelor's degree -0.59729549  5.504941e-01
## parental.level.of.educationhigh school       0.58975853  5.555338e-01
## parental.level.of.educationmaster's degree  -2.06343123  3.942226e-02
## parental.level.of.educationsome college      1.65551258  9.824743e-02
## parental.level.of.educationsome high school  1.39000596  1.649479e-01
## lunchstandard                                6.76047818  2.803725e-11
```

```
## test.preparation.coursenone                   7.91288769  9.254318e-15
## reading.score                                 4.87930905  1.306031e-06
## writing.score                                14.21291304  1.108770e-40
```

The variables that were selected for the model based on p-value are: gender, lunch, test.preparation.course,reading.score and writing.score

Before proceding with further investigation, VIF will be examined to find whether or not there is colinearity between the predictors.

```
vif(model1)
```

```
##                                  GVIF Df GVIF^(1/(2*Df))
## gender                       1.175355  1        1.084138
## race.ethnicity               1.133902  4        1.015832
## parental.level.of.education  1.174067  5        1.016177
## lunch                        1.112281  1        1.054647
## test.preparation.course      1.252851  1        1.119308
## reading.score               12.758294  1        3.571875
## writing.score               14.756446  1        3.841412
```

The mean VIF is well under 10, thus there is not multicollinearity .

```
reduced <-
  lm(formula=p.student_train$math.score ~gender+lunch+test.preparation.course+reading.score+writing.sco
     data=p.student_train)

summary(reduced)
```

```
##
## Call:
## lm(formula = p.student_train$math.score ~ gender + lunch + test.preparation.course +
##     reading.score + writing.score, data = p.student_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7596  -3.5920   0.0561   3.9204  15.6561
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -10.63975    1.26646  -8.401 2.23e-16 ***
## gendermale                  13.39477    0.44331  30.215  < 2e-16 ***
## lunchstandard                3.29343    0.45094   7.304 7.23e-13 ***
## test.preparation.coursenone  3.20687    0.47049   6.816 1.93e-11 ***
## reading.score                0.28113    0.05001   5.622 2.67e-08 ***
## writing.score                0.68403    0.05069  13.493  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.69 on 744 degrees of freedom
## Multiple R-squared:  0.8614, Adjusted R-squared:  0.8605
## F-statistic: 924.8 on 5 and 744 DF,  p-value: < 2.2e-16
```

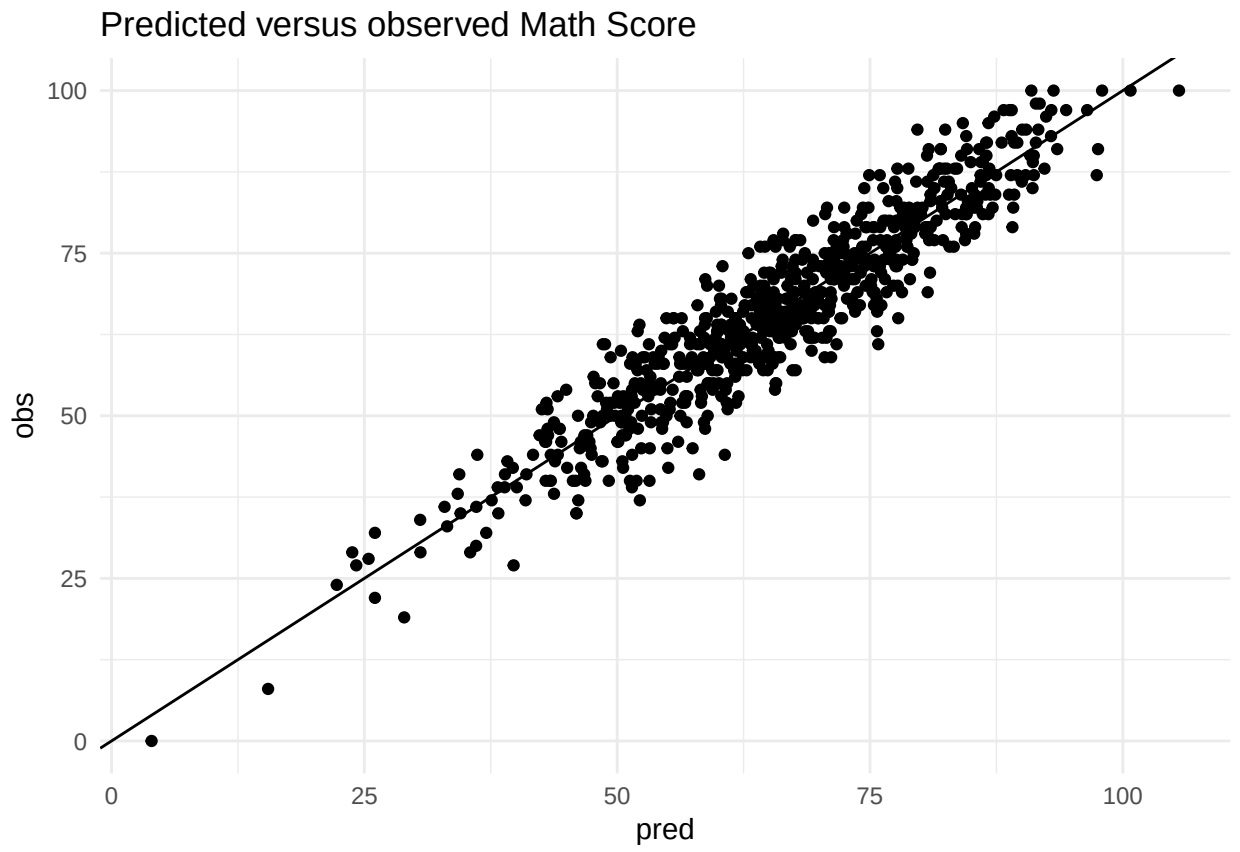
There is a reduction in R-squared from 87% to 86%, thus model1 is preferred.

With the function prediction(), values for the math.score will be obtained, against the observed math.score values.

```
y_pred <- predict(model1,newdata = p.student_train)
```

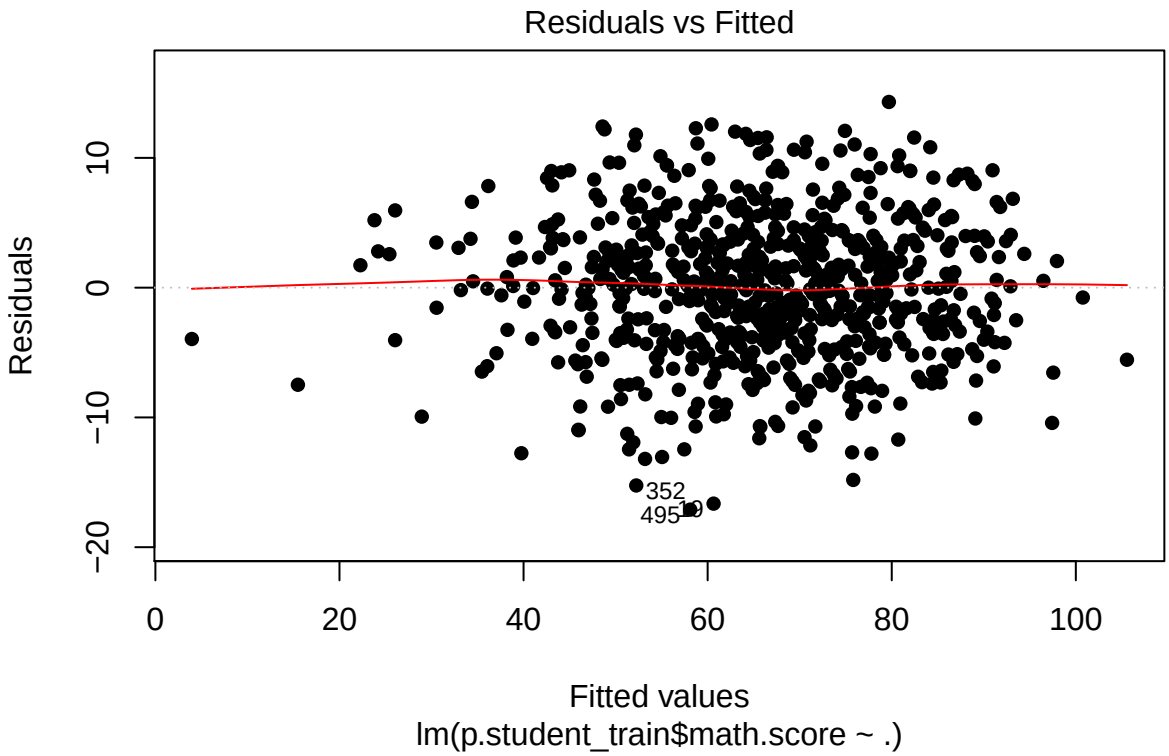
Create a tibble with the predicted and the observed values. Construct a plot y_pred mapped to the x position and the true y value (p.student_train$math.score) mapped to the y value to examine the fit. The line indicates that the fit is perfect.

```
tibble(pred = y_pred,
       obs  = p.student_train$math.score) %>%
  ggplot(aes(x = pred, y = obs)) +
  geom_point() +
  theme_minimal() +
  geom_abline(slope = 1)+
  labs(title = "Predicted versus observed Math Score")
```



Residuals plot

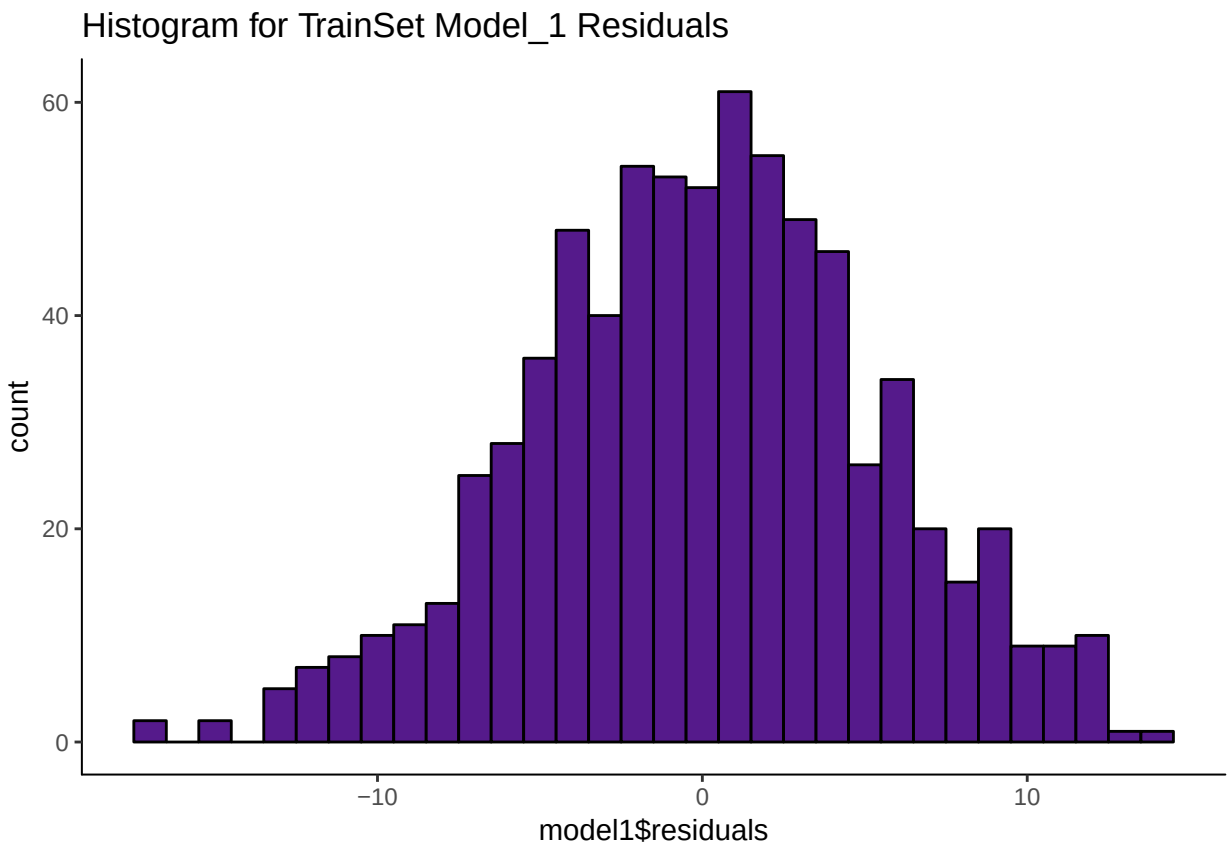
```
plot(model1, pch=16, which=1)
```



This residual plot is constructed from the multiple linear regression "model1". The red line is a smooth fit to the residuals.

Also, from the histogram on the residuals we can see that they form a normal distribution.

```
ggplot(data=p.student_train, aes(model1$residuals)) +
    geom_histogram(binwidth = 1, color = "black", fill = "purple4")+
    theme(panel.background = element_rect(fill = "white"),
          axis.line.x=element_line(),
          axis.line.y=element_line()) +
    ggtitle("Histogram for TrainSet Model_1 Residuals")
```



Performing multiple regression to the test set for cross validation.

```
model2 <- lm(formula=math.score ~ . , data=p.student_test)
summary(model2)
```

```
##
## Call:
## lm(formula = math.score ~ ., data = p.student_test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3111  -3.6941  -0.0203   3.1033  14.6202
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                   -10.37186    2.46860  -4.202
## gendermale                                     13.03477    0.75613  17.239
## race.ethnicitygroup B                           0.55315    1.57437   0.351
## race.ethnicitygroup C                           0.73638    1.46474   0.503
## race.ethnicitygroup D                          -0.24205    1.49746  -0.162
## race.ethnicitygroup E                           5.58863    1.59568   3.502
## parental.level.of.educationbachelor's degree   -3.50901    1.25008  -2.807
## parental.level.of.educationhigh school          0.78604    1.00869   0.779
## parental.level.of.educationmaster's degree     -2.19615    1.46245  -1.502
## parental.level.of.educationsome college        -1.44474    1.00927  -1.431
## parental.level.of.educationsome high school    -0.61306    1.09113  -0.562
## lunchstandard                                   3.88819    0.74276   5.235
```

```
## test.preparation.coursenone                     2.89273    0.79367   3.645
## reading.score                                   0.28033    0.08256   3.395
## writing.score                                   0.68062    0.08484   8.023
##                                                Pr(>|t|)
## (Intercept)                                    3.77e-05 ***
## gendermale                                      < 2e-16 ***
## race.ethnicitygroup B                          0.725644
## race.ethnicitygroup C                          0.615622
## race.ethnicitygroup D                          0.871728
## race.ethnicitygroup E                          0.000552 ***
## parental.level.of.educationbachelor's degree 0.005420 **
## parental.level.of.educationhigh school       0.436604
## parental.level.of.educationmaster's degree   0.134519
## parental.level.of.educationsome college      0.153623
## parental.level.of.educationsome high school  0.574747
## lunchstandard                                  3.66e-07 ***
## test.preparation.coursenone                    0.000329 ***
## reading.score                                  0.000804 ***
## writing.score                                  4.90e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.107 on 235 degrees of freedom
## Multiple R-squared:  0.8901, Adjusted R-squared:  0.8835
## F-statistic: 135.9 on 14 and 235 DF,  p-value: < 2.2e-16
```

As it is possible to examine from the summary(), regression coefficient gender and for male is associated with an increase of 13 points on math.score against females. The regression coefficient Race ethnicity E has also significant association with an increase of 4 points on math.score against Race ethnicity A.

However,regression coefficient race.ethnicities B,C and D do not reach statistical significance. Same applies for the regression coefficient parental.level.of.education for the education levels of: bachelor's degree,some college,master's degree high school.

On the contrast regression coefficient parental.level.of.education some high school is associated with a decrease of 2.4 points on math.score against asscociate's degree.

The regression coefficient Lunch is associated with an increase of 2.8 points on math.score against free/reduced.The regression coefficient Reading.score is associated with an increase of 3.7 points on math.score.

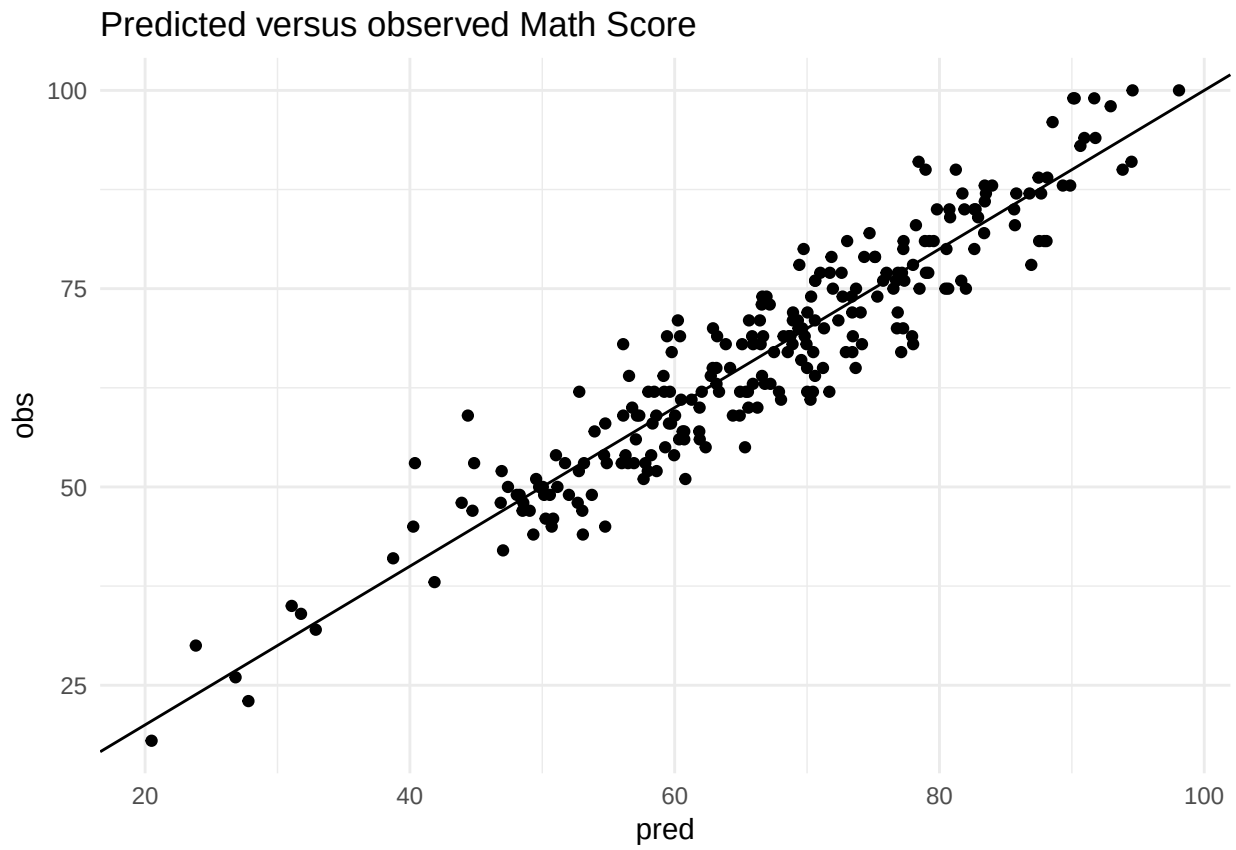
The regression coefficient writing.score is associated with an increas of 0.35 on math.score. The regression coefficient test.preparation none is associated with an increase of 0.67 against completed preparation.

```
y_pred2 <- predict(model2,newdata = p.student_test)
```

Create a tibble with the predicted and the observed values. Construct a plot y_pred mapped to the x position and the true y value (p.student_train$math.score) mapped to the y value to examine the fit. The line indicates that the fit is perfect.

```
tibble(pred = y_pred2,
       obs  = p.student_test$math.score) %>%
  ggplot(aes(x = pred, y = obs)) +
  geom_point() +
  theme_minimal() +
```

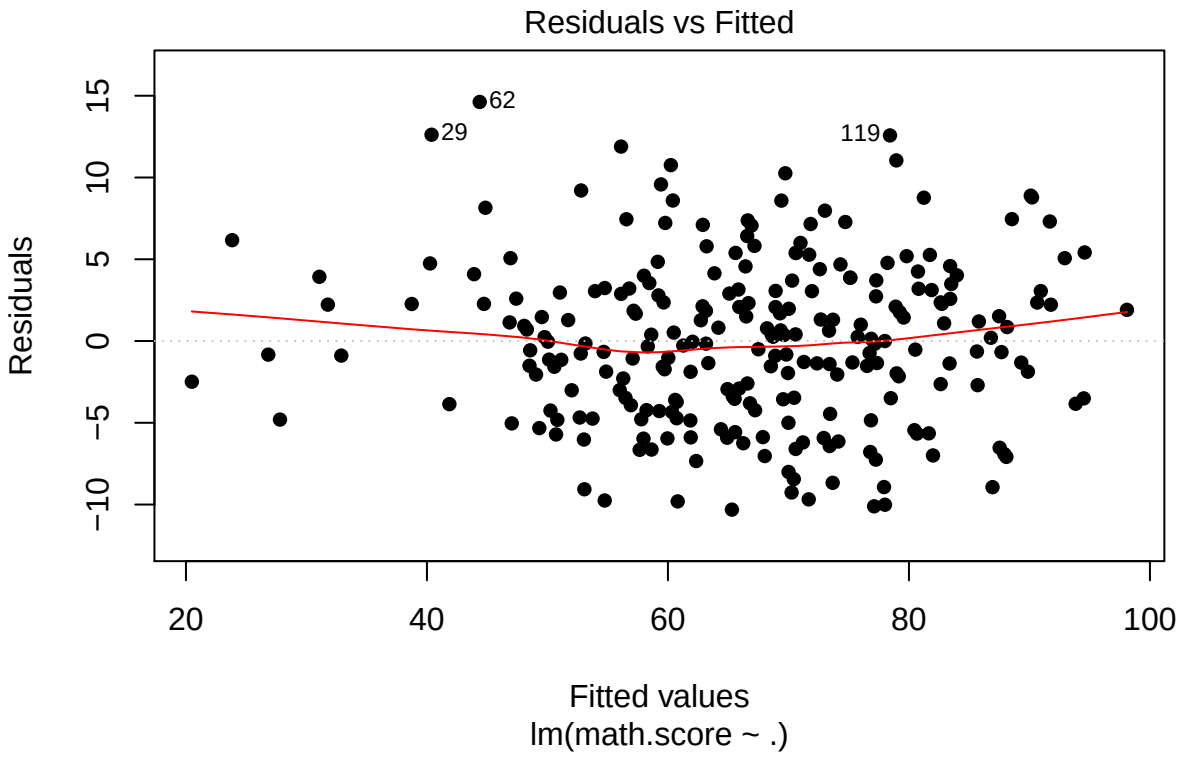
```
  geom_abline(slope = 1)+
  labs(title = "Predicted versus observed Math Score")
```




```
#95% confidence interval
prediction1 <- predict(model2, p.student_test, interval="confidence", level = 0.95)
```

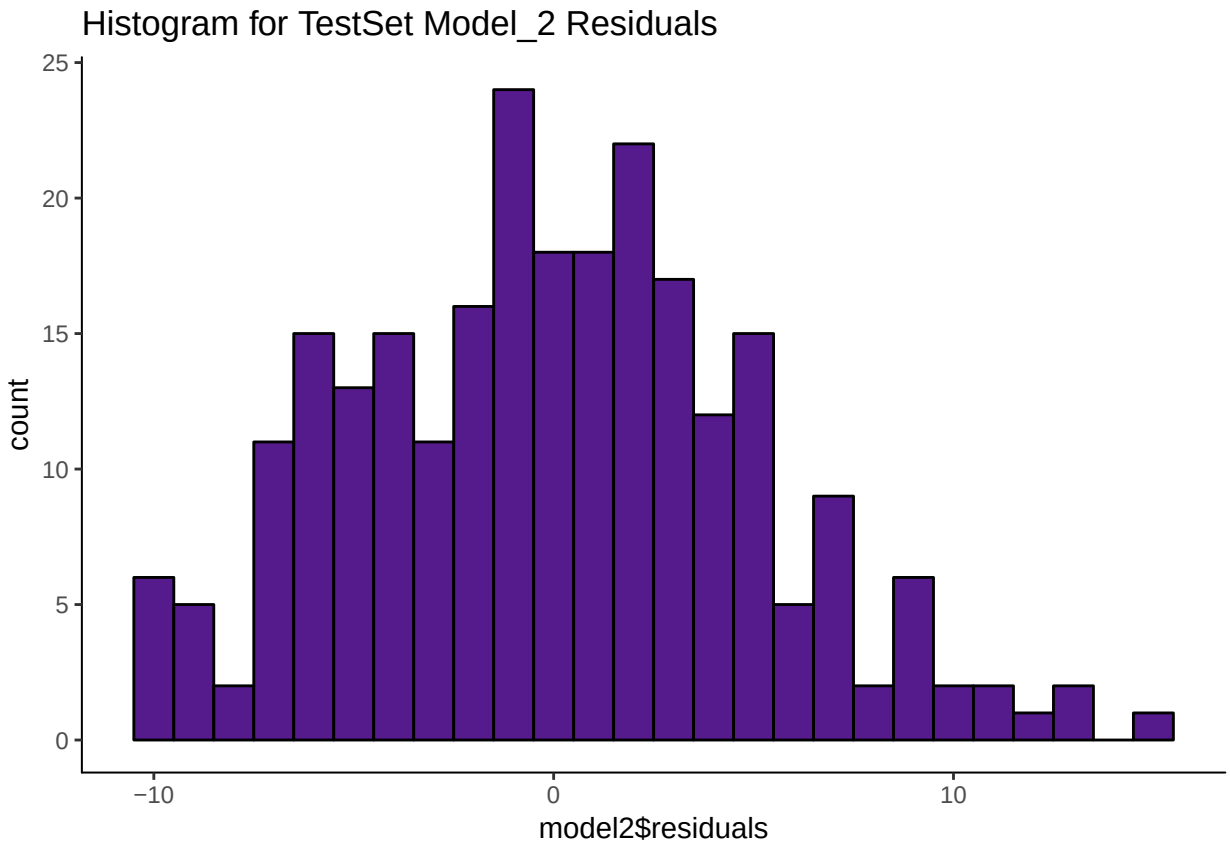
Residuals plot

```
plot(model2, pch=16, which=1)
```



This residual plot is constructed from the multiple linear regression "model2". The red line is a smooth fit to the residuals. From the examination of the plot, the line is a smooth fit to the residuals. Also, from the histogram on the residuals we can see that they form normal distribution.

```
ggplot(data=p.student_test, aes(model2$residuals)) +
    geom_histogram(binwidth = 1, color = "black", fill = "purple4")+
    theme(panel.background = element_rect(fill = "white"),
          axis.line.x=element_line(),
          axis.line.y=element_line()) +
    ggtitle("Histogram for TestSet Model_2 Residuals")
```



The training model and the test model have same R-squared value.