Hồ Minh Trí: 1751108

Goi Chí Trung: 1751113

Nguyễn Minh Trí: 1751109

Nguyễn Thanh Tùng: 1751119

Statistics Report

This report is about a dataset concerning a sample of students' performance in exams at a public school. We find the dataset on Kaggle, at https://www.kaggle.com/spscientist/students-performance-in-exams.

The dataset consists of 1000 observations on the variables:

- Gender: the gender of the student, qualitative.
- Race/ethnicity: the race or ethnicity group the student falls into, qualitative.
- Parental level of education: the level of education of the parents of the student, qualitative.
- Lunch: whether the student has standard or free/price reduced lunch, qualitative.
- Test preparation course: whether the student has taken a test preparation course before taking the tests, qualitative.
- Math score: the math score of the student, quantitative.
- Reading score: the reading score of the student, quantitative.
- Writing score: the writing score of the student, quantitative.

In this report, we will investigate the many different ways we can make use of the R language to help us better understand the dataset.

The report is divided into two main sections. In the first section, we apply methods of descriptive statistics on the dataset. In the second section, we apply methods of inferential statistics on the dataset.

The table of contents of the report is as follows:

I/ Descriptive statistics:

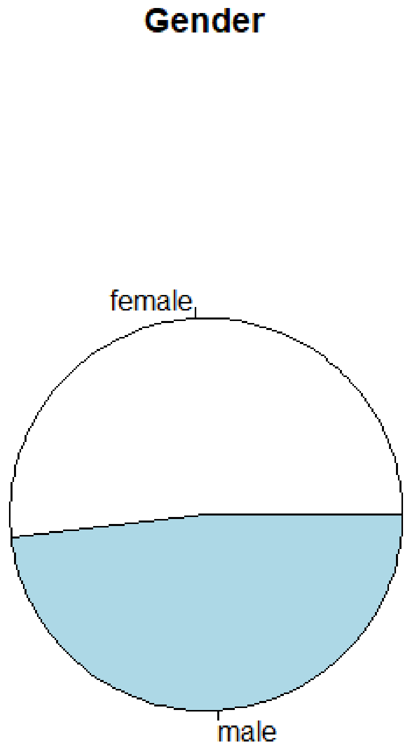
This first section is about descriptive statistics, where we look at methods to summarize and describe important features of the data that we have. The variables in our dataset are either qualitative or quantitative. We will therefore look first into the qualitative variables, then the quantitative variables, and finally all of them together.

1. Qualitative variables:

We will first look into the qualitative variables. To make things simple, we begin by looking at the variables individually, then we move to looking at multiple qualitative variables together.

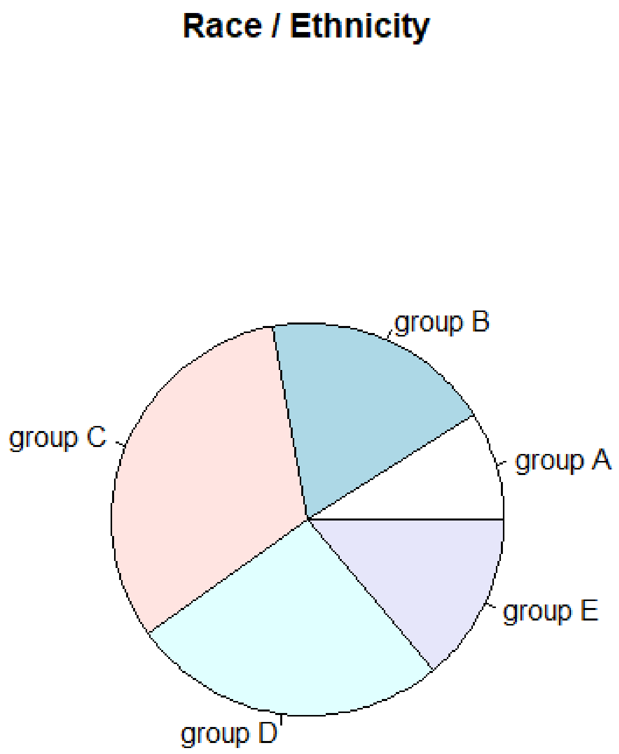
a. Single variable:

We can visualize individual qualitative variables, with the help of pie charts. For example, here is the pie chart for the gender variable.



From the pie chart, we can see that the number of males and females in the dataset is roughly the same, with there being slightly more females than males.

As another example, here is the pie chart for the race / ethnicity variable.



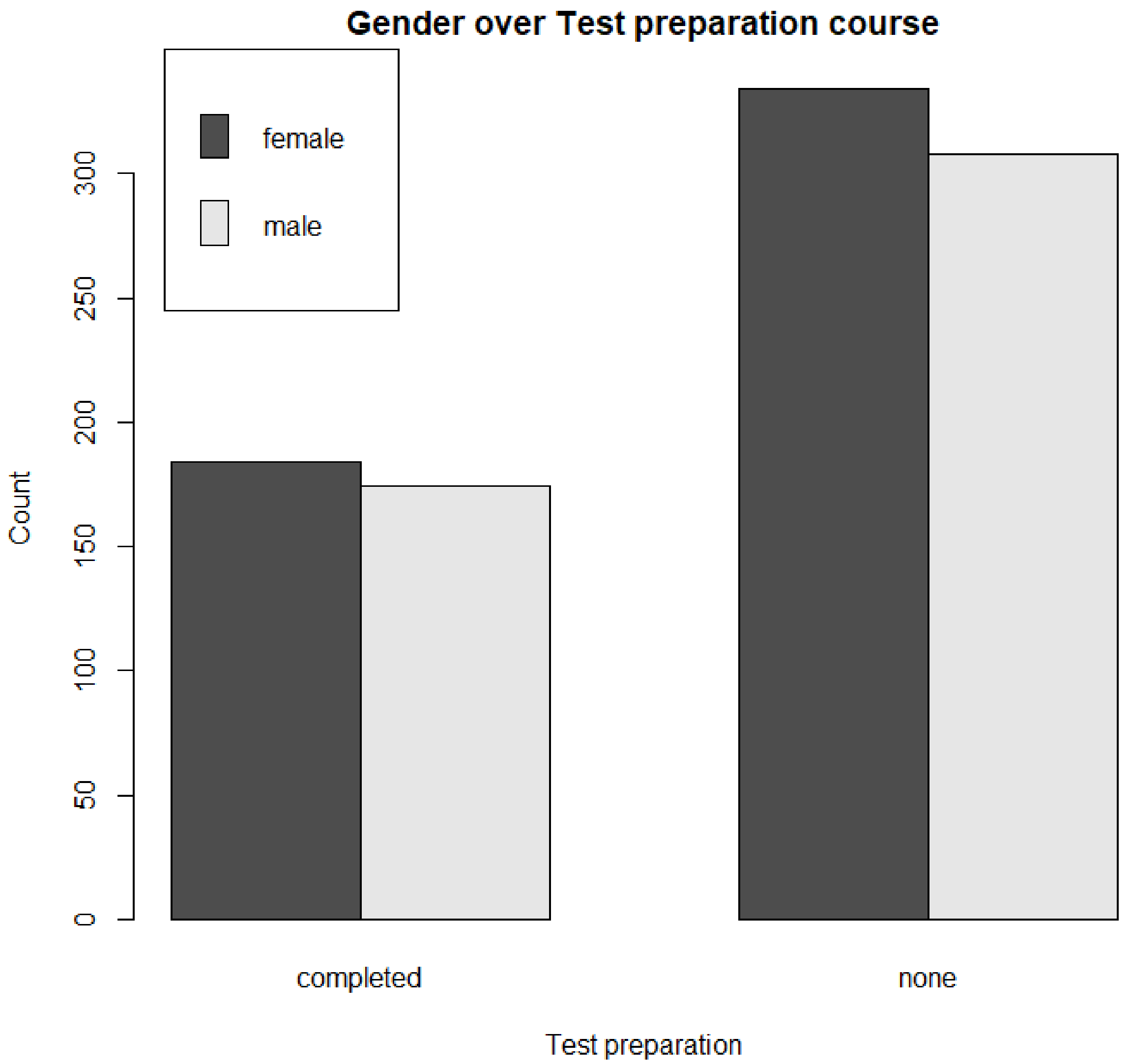
The pie chart helps us to see which group has many students and which group has few. Also, from the pie chart, we can see that group E has about twice as many students as group A.

So as we can see, the pie chart helps us to visualize a qualitative variable, helping us to see in an intuitive way the relative percentage of the dataset of each value of a qualitative variable.

b. Multiple variables:

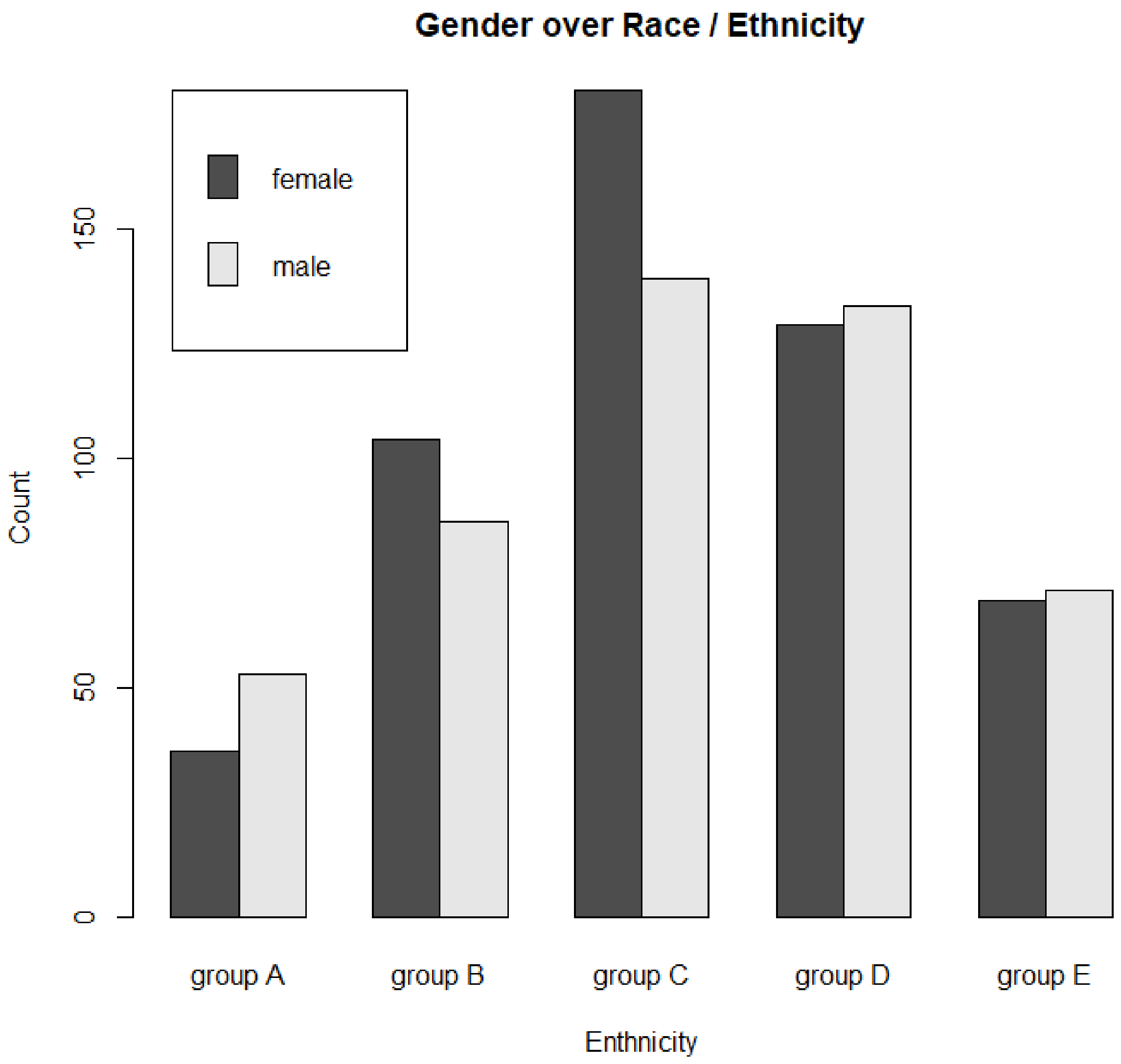
After looking at each qualitative variable individually, we now look at what interesting observations can be made when we look at the variables together. We can do this with the help of the bar plot.

For example, here is a bar plot that shows how gender and test preparation relates to each other.



From the bar plot, we can that in both groups that either took the test preparation course or not, the gender distribution is roughly equal between males and females.

Here is another bar plot that relates gender with ethnicity.



We can see that for group C, the most dominant ethnicity in the dataset, there is a notably higher number of females compared to males.

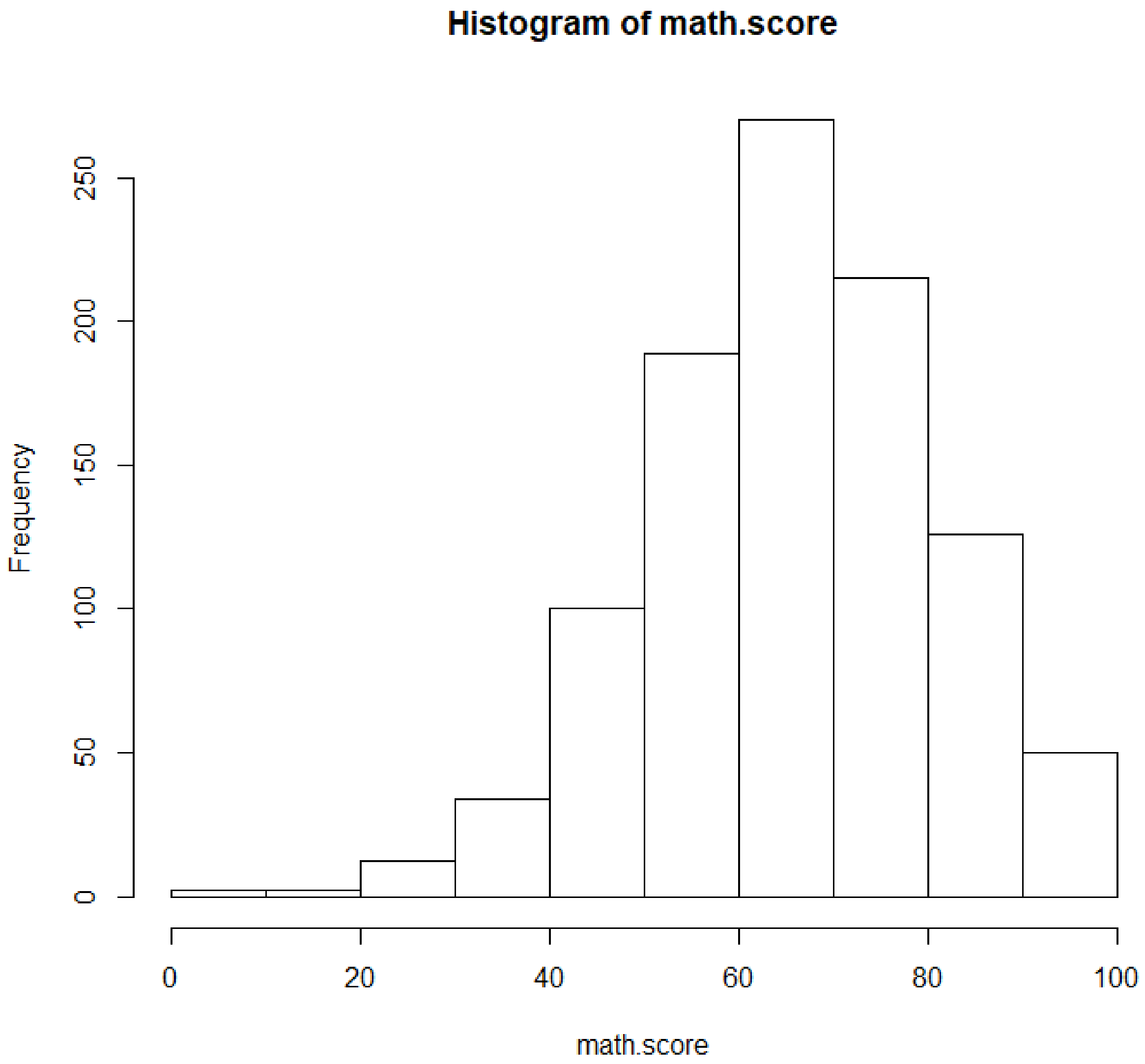
So as we can see, the bar plot can help us to quickly see how pairs of qualitative variables relate to each other.

2. Quantitative variables:

After the qualitative variables, we now look into the quantitative variables. As with the qualitative variables, we begin by looking at the variables individually, then we move to looking at multiple quantitative variables together.

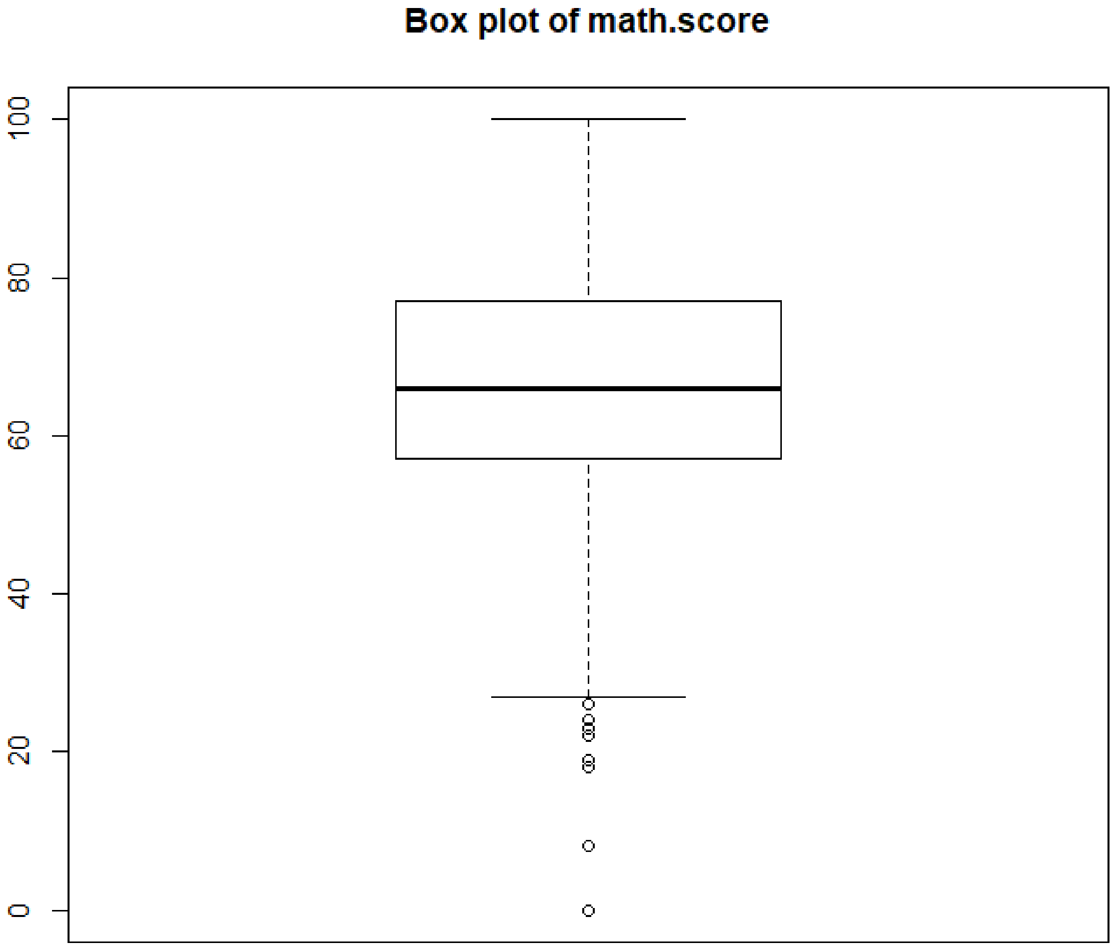
a. Single variable

One way to illustrate a quantitative variable is with the use of a histogram. For example, here is the histogram for the math score variable.



From the histogram, we can estimate the mean of the students' math scores to be somewhere around 60 and 80, possibly higher than 65. We can see that the range of score from 60 to 70 contains the highest number of students, and the farther a range of score from this range, the fewer students that fall into it. We can also see that the histogram is not symmetric and has a left skew, where most students have a math score above 50.

Another way to visualize a quantitative variable is with the help of a box plot. For example, here is the math score variable from earlier in box plot form.



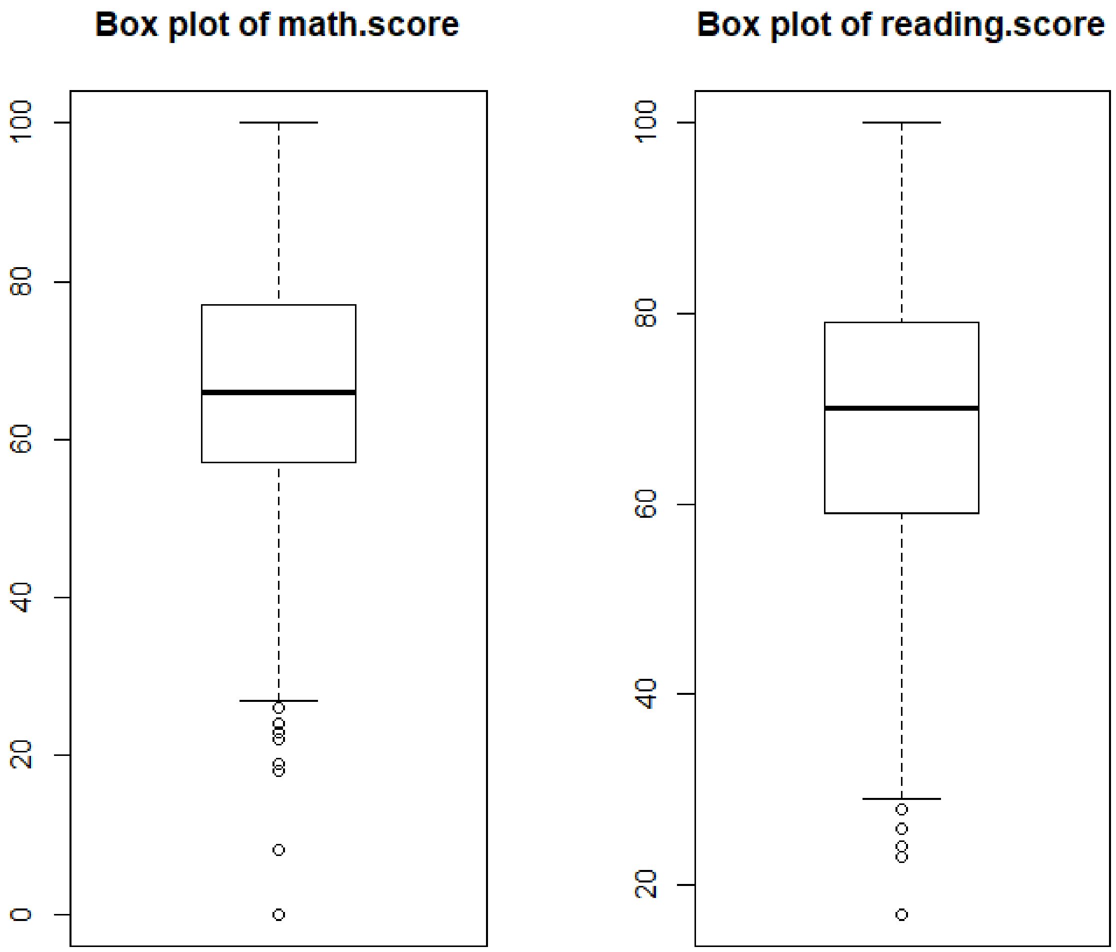
From this, we can see that the median student has a math score of below 70. We can also see that 50% of the students' math scores are concentrated in the somewhat narrow range from 60 to 80. Our dataset also has a few outliers, with scores below 30.

So, the histogram and the box plot help to visualize a quantitative variable in their own ways, with each giving us unique insights into the variable.

b. Multiple variables:

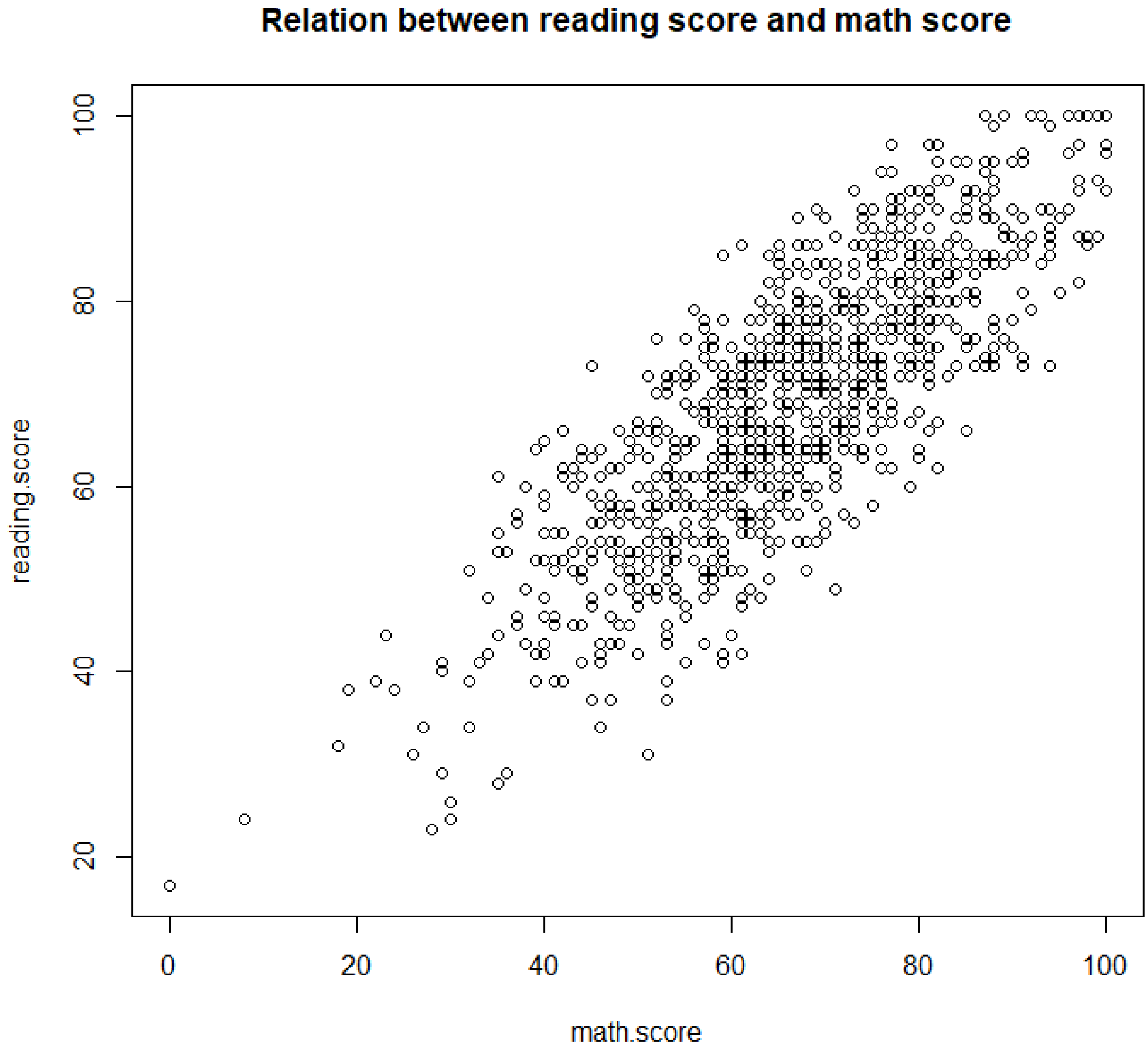
We can also look at the quantitative variables together.

One way is to compare one quantitative variable to another, with the help of box plots. For example, here are the box plots for the math score and reading score variables.



We can derive from the box plots that although the two scores have roughly the same median, reading score has both a wider interquartile range and a wider range.

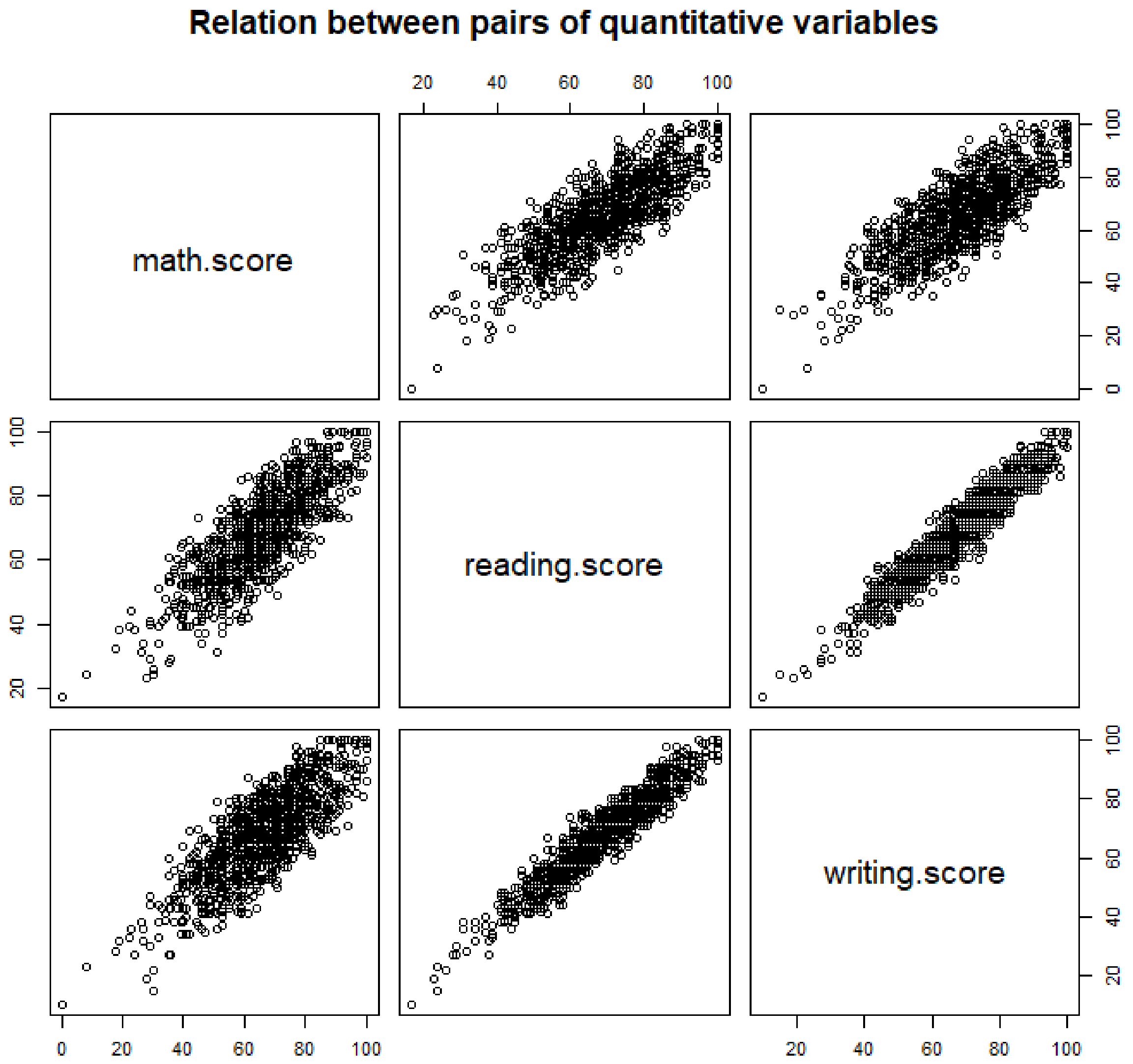
Another way to investigate quantitative variables is to see how one variable is correlated with another. We can visualize this with the help of a scatterplot.



From the plot, we can see that reading score has a roughly linear relationship with math score. This makes sense because a student that is both smart and hardworking enough to get high math score tends to also get high reading score, and vice versa.

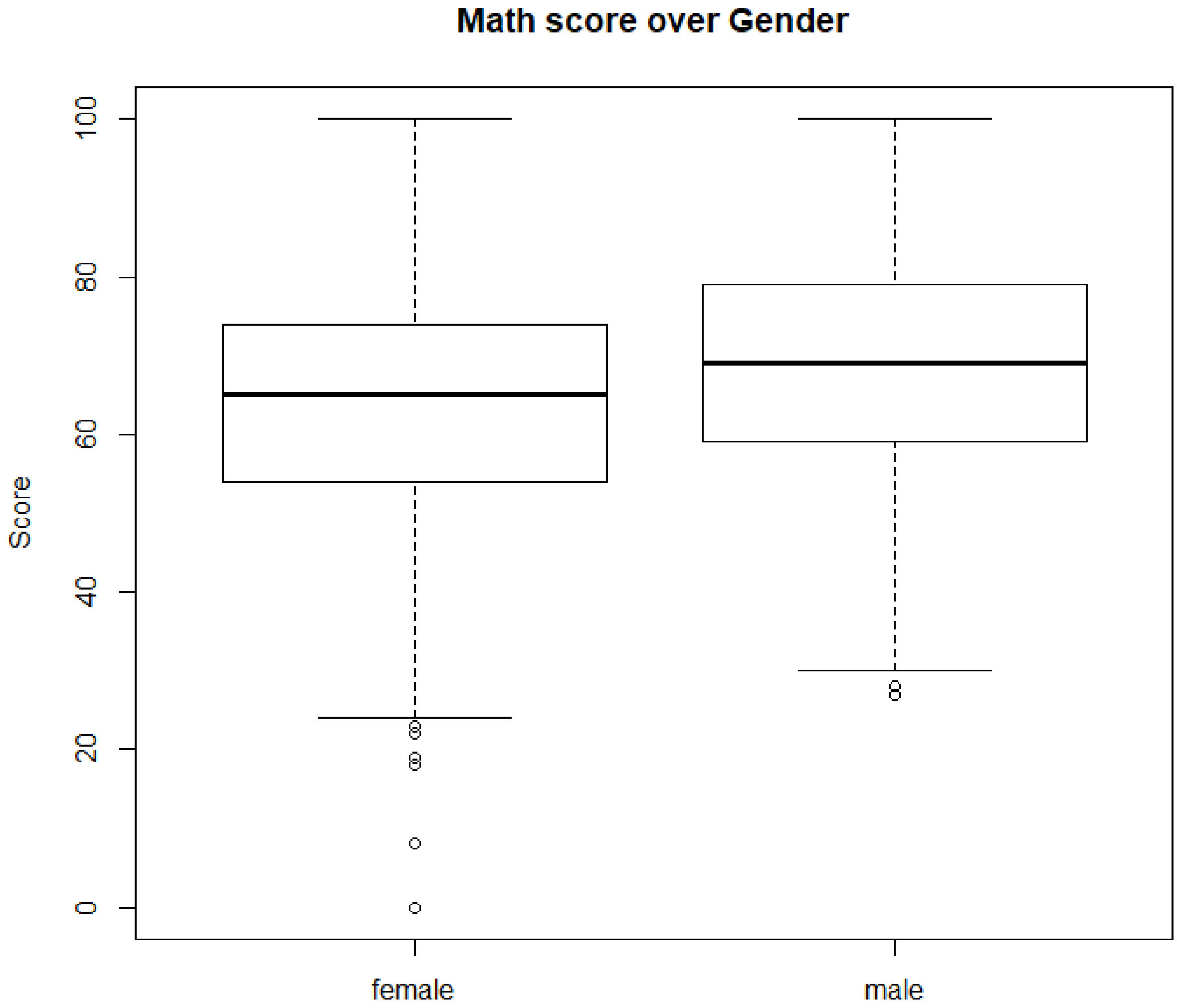
We can verify that the two scores have a roughly linear relationship by calculating their correlation. The cor() function of R returns 0.818, which is quite high, confirming our observation.

Finally, here is a matrix of scatterplots, showing the correlation between all pairs of quantitative variables in our dataset.



3. Qualitative and quantitative variables:

After investigating the qualitative and the quantitative variables separately, we now consider them together. One useful tool for this is the box plot. For example, here is a box plot showing the relationship between gender and math score.

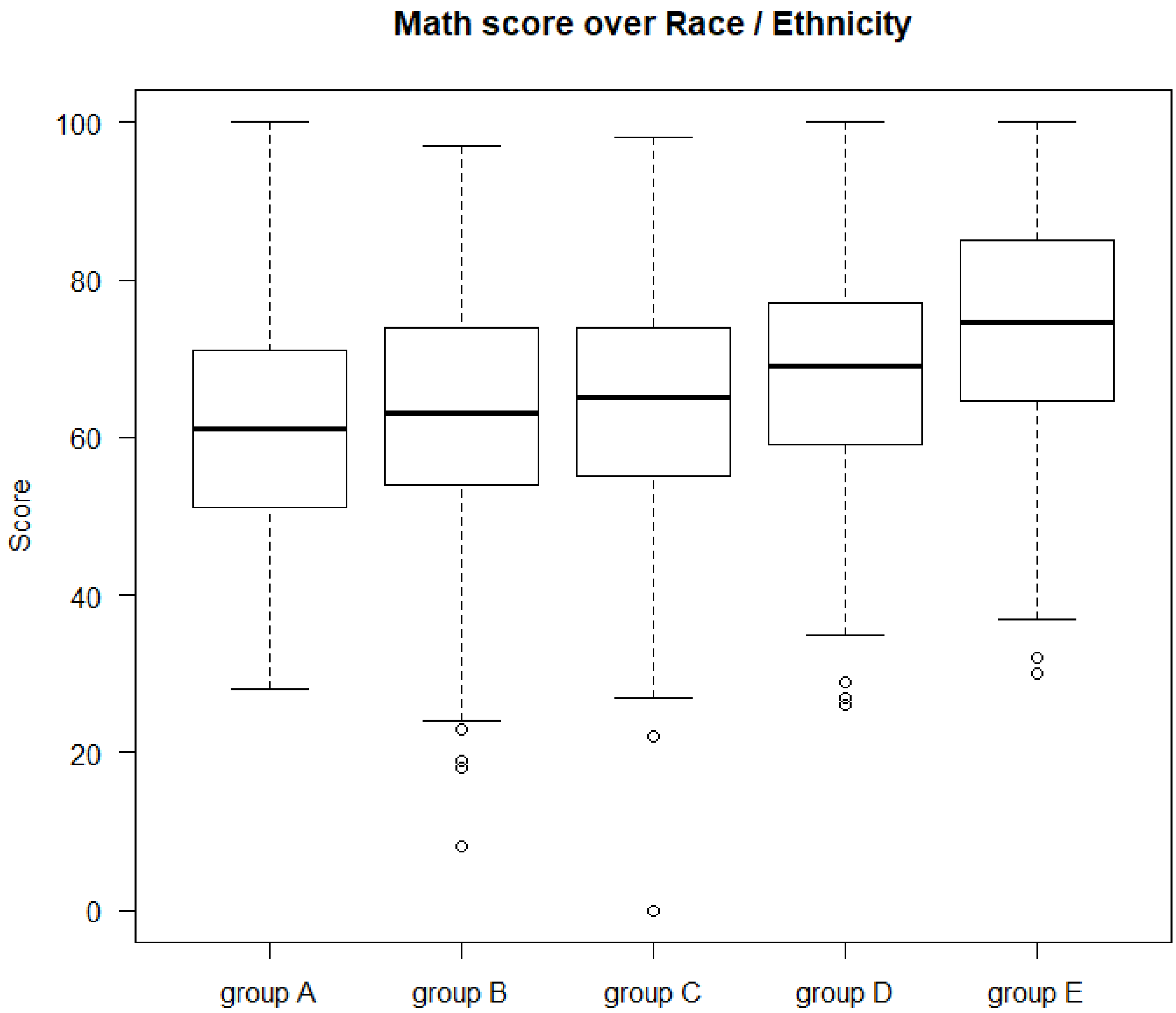


We can make many interesting observations about this. The median, upper quartile, and lower quartile of male math scores are all higher than those of females. This indicates that the median male has a somewhat higher math score than the median female, and that the middle 50% of male scores are also somewhat higher than the middle 50% of female scores.

Also, both the range and the interquartile range of male scores are narrower than those of females, which is possibly because male scores tend to be higher than female scores.

For some reason, there is a notably higher number of outliers for female scores compared to male scores, with some outliers' scores going very low.

As another example, here is a boxplot showing the relationship between math score and ethnicity.



The box plot helps us to see the students group E has considerable higher math scores, measured by median, first quartile, or third quartile.

As we can see, with the help of the box plot, we can make meaningful observations about the relationship between qualitative and quantitative variables.

II/ Inferential statistics:

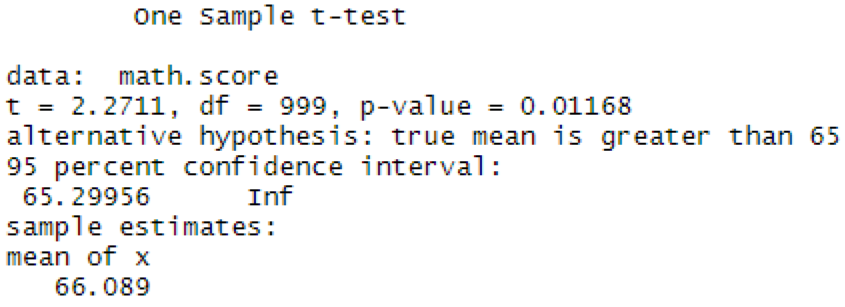
Now that we have made some basic observations on the sample data that we have, it could be useful if we are able to not just describe existing data but also derive some type of conclusion about the population. This lies in the field of inferential statistics.

1. Hypothesis testing:

Along the way, we have made many hypotheses, such that "The mean math score is greater than 65", "There is an equal proportion of males and females". It would be more concrete if we can decide if our sample data supports this statement, with a certain level of confidence.

a. Tests of population mean:

We first focus on those hypotheses concerning population mean. Since we are interested in mean, we make use of the function t.test() of R. We apply t-test on the hypothesis that the mean math score is greater than 65.

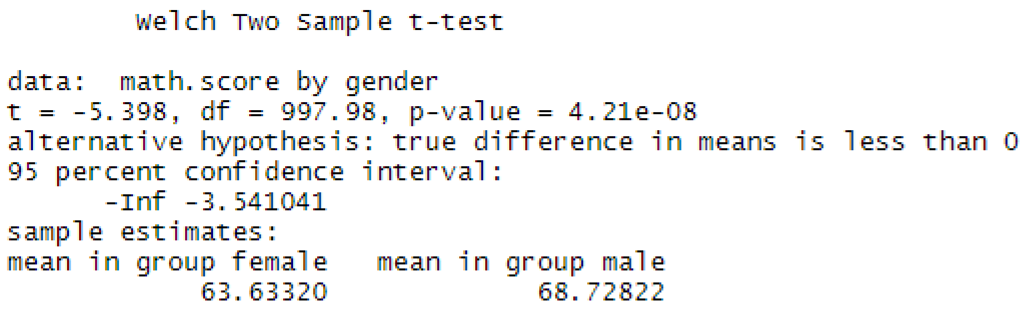


```
        One Sample t-test

data:  math.score
t = 2.2711, df = 999, p-value = 0.01168
alternative hypothesis: true mean is greater than 65
95 percent confidence interval:
 65.29956      Inf
sample estimates:
mean of x
   66.089
```

Because the p-value is 0.012, smaller than 0.05, we can see that at 95% confidence, we reject the null hypothesis that the mean is equal to 65 and accept the alternative hypothesis that the mean is above 65.

The t-test provides us with an estimate of the mean math score of the population, which is 66.089. We also get information about a 95% one-sided confidence interval for the mean, which is larger than 65.300.

We can also use the t-test to compare the means between two populations, such as the math score of males and females. We apply t-test on the hypothesis we made earlier that the mean math score of females is smaller than that of males.



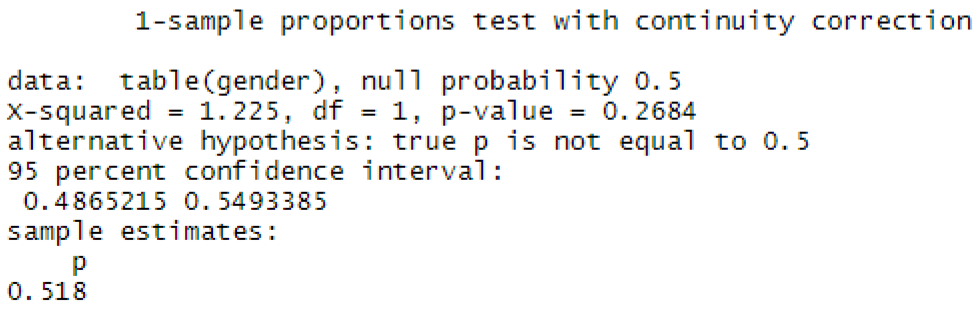
```
        Welch Two Sample t-test

data:  math.score by gender
t = -5.398, df = 997.98, p-value = 4.21e-08
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.541041
sample estimates:
mean in group female   mean in group male 
            63.63320             68.72822 
```

The p-value is very small, strongly suggesting that we reject the null hypothesis that the two means are equal and accept the alternative hypothesis the mean math scores of females is smaller than that of males.

We get an estimate of the mean math score of females and males, which are 63.633 and 68.728, respectively. We also get a 95% one-sided confidence interval for mean_female_math_score – mean_male_math_score, which is smaller than -3.541.

b. Tests of population proportion:

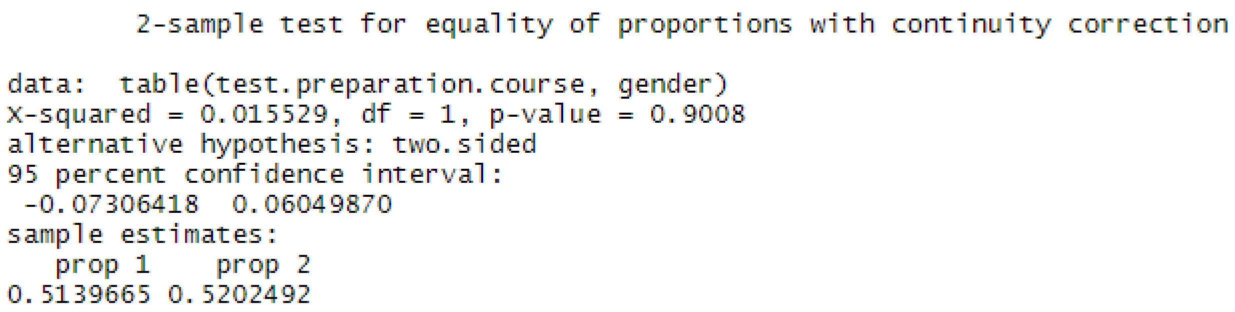
Another type of hypothesis consists of those concerning population proportion. For this, we make use of the prop.test() function. We have previously claimed that the proportion of males and females is the same. We now apply prop.test() to this hypothesis.



```
        1-sample proportions test with continuity correction

data:  table(gender), null probability 0.5
X-squared = 1.225, df = 1, p-value = 0.2684
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4865215 0.5493385
sample estimates:
    p 
0.518 
```

Because the p-value is 0.268, larger than 0.025, we can see that we do not reject the null hypothesis that the proportion of female is 0.5, with confidence 95%. prop.test() also provides us with an estimate of the proportion of females, which is 0.518 and a 95% two-sided confidence interval for the proportion of females, which is from 0.487 to 0.549.

We can also use prop.test() to compare the proportions between two populations. In this case, we choose to compare the proportion of female in the group that does not take the test preparation course and in the group that does.

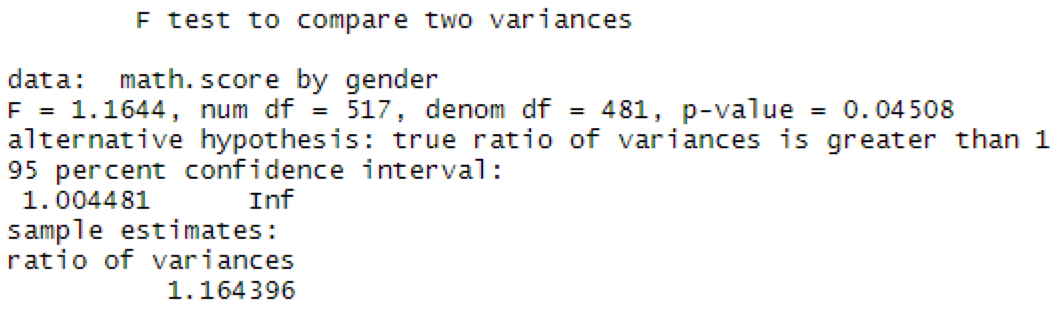


```
        2-sample test for equality of proportions with continuity correction

data:  table(test.preparation.course, gender)
X-squared = 0.015529, df = 1, p-value = 0.9008
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.07306418  0.06049870
sample estimates:
   prop 1    prop 2 
0.5139665 0.5202492 
```

Because the p-value is very close to 1, we accept the null hypothesis that the proportions of females in the two group is equal. We also get estimates for the proportions, 0.514 and 0.520, both approximately 0.5. This agrees with our previous observation that the gender distribution is equal in both groups. And lastly, the 95% two-sided confidence interval for the two proportions goes from -0.073 to 0.060.

c. Tests of population variance:

Yet another type of hypothesis involves the variances of populations. By using var.test(), we can verify the statement we made above, that the variance in math score for females is greater than that for males.



```
        F test to compare two variances

data:  math.score by gender
F = 1.1644, num df = 517, denom df = 481, p-value = 0.04508
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 1.004481      Inf
sample estimates:
ratio of variances 
          1.164396 
```

Because the p-value is 0.045, just under 0.05, we reject the null hypothesis that the two variances are equal and accept the alternative hypothesis that the variance in math score for females is greater than that for males, at 95% confidence. An estimate for the ratio of the variance of females' math scores and males' math scores if 1.164 and a 95% one-sided confidence interval for this value is greater than 1.004.

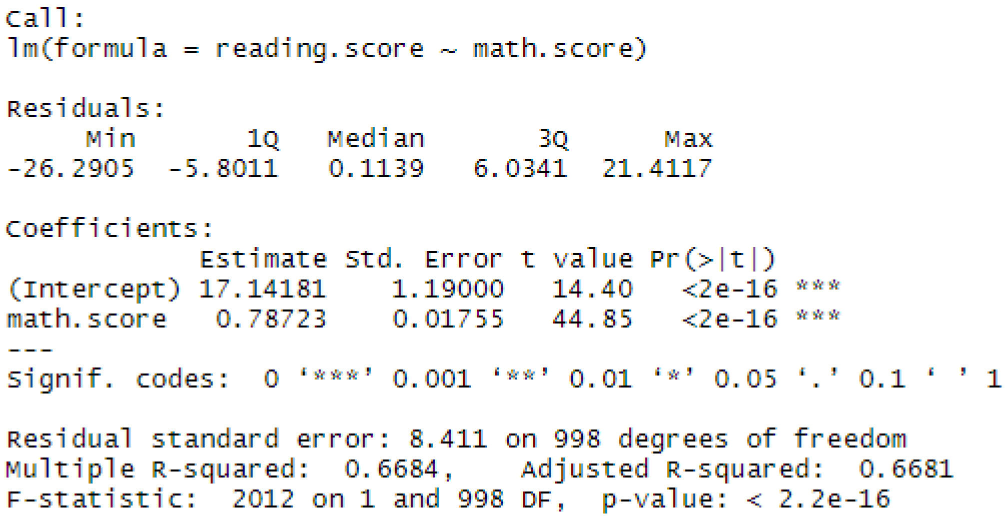
2. Regression:

One thing that would be very useful is to determine how some variables relate to others. This would allow us to answer questions such as "Does a student with high math score usually have high reading score", "For a student with math score

70 and writing score 60, in what interval would their reading score likely fall into”, “Does the math score depends on the reading score more for male or female students?”. Regression analysis allows us to answer these questions.

a. Simple regression:

We first investigate the simplest form of regression, where a variable is linearly related to another single variable. We choose to see how the reading score is linearly determined by the math score. By using the lm() function of R, we can derive a lot of useful information.



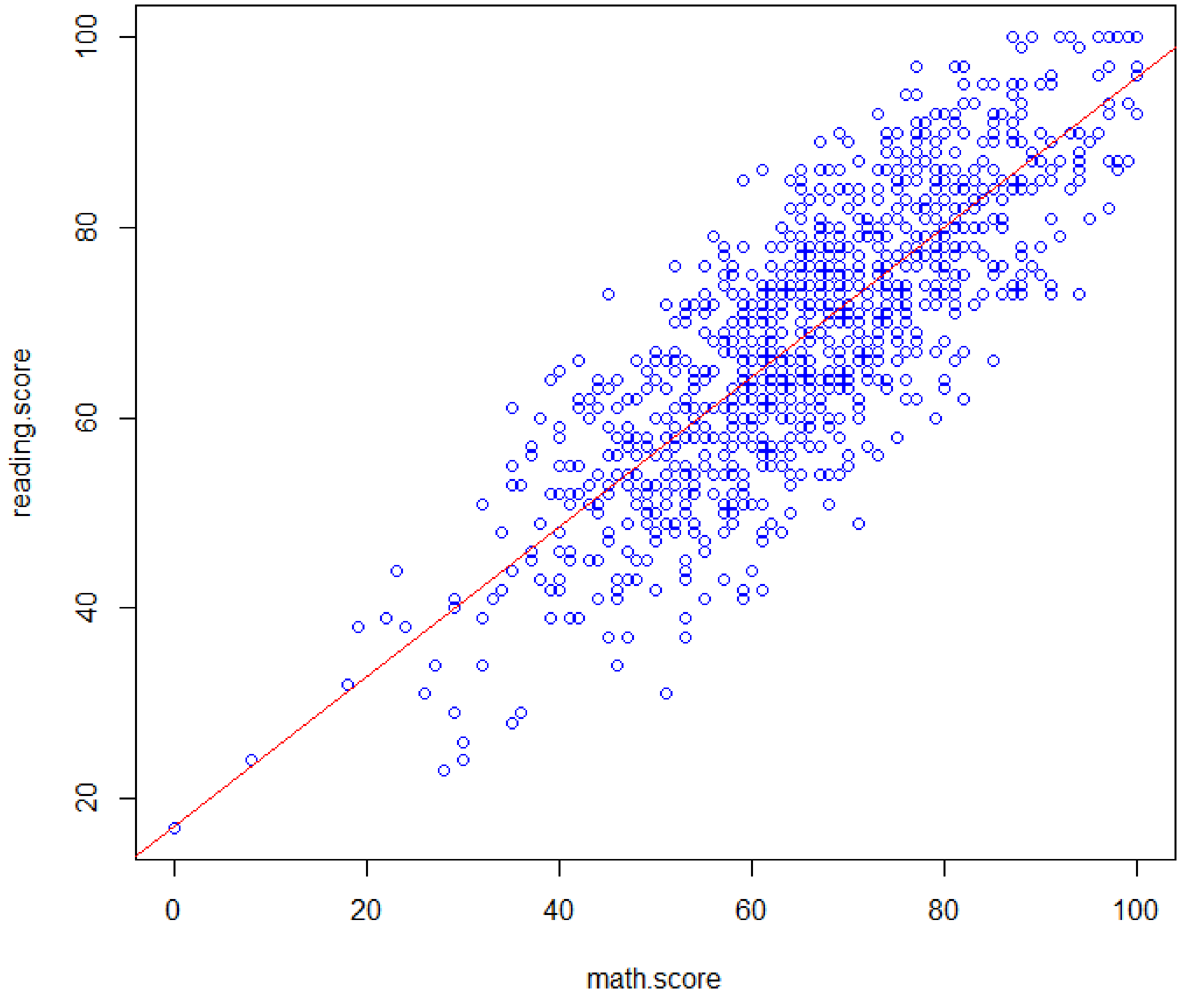
```
Call:
lm(formula = reading.score ~ math.score)

Residuals:
     Min       1Q   Median       3Q      Max 
-26.2905  -5.8011   0.1139   6.0341  21.4117 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 17.14181    1.19000   14.40   <2e-16 ***
math.score   0.78723    0.01755   44.85   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.411 on 998 degrees of freedom
Multiple R-squared:  0.6684,	Adjusted R-squared:  0.6681 
F-statistic:  2012 on 1 and 998 DF,  p-value: < 2.2e-16
```

We can see that the model equation that R calculate is:

reading.score = 17.142 + 0.787 * math.score

with se: 1.190   0.018

We can plot the line of this equation to see how well it fits the data.



As we can see, the line is a pretty good approximation of the data.

With the given t-values for the coefficients, we can conduct some hypothesis testing. For example, we want to test if the reading score is truly related to the math score or if they are unrelated. That is, for reading.score = B1 + B2 * math.score, we want to test if B2 != 0.

Because Pr(>|t|) is very small, we reject the null hypothesis that B2 = 0 and accept the null hypothesis that B2 != 0. In other words, the reading score is related to the math score.

We can also calculate confidence intervals for the coefficients, using the confint() function. We can see that a 95% interval for B2 is from 0.753 to 0.822.

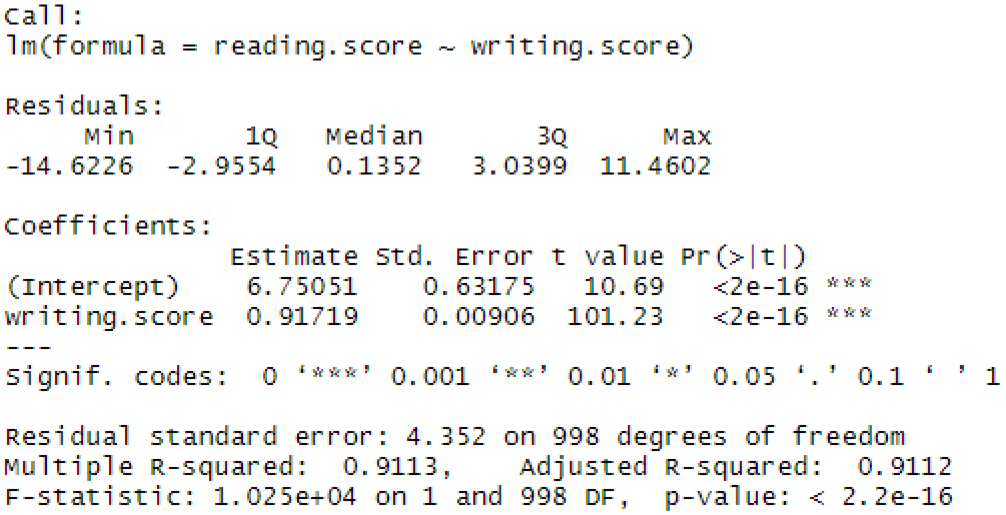
The model equation of the linear relationship between reading score and math score is:

reading.score ~= 17.142 + 0.787 * math.score

This tells us that when the math score of a student increases by 1, their reading score increases by about 0.787. The line also predicts that a student with math score of 76 would have reading score of about 76.971. If we use the predict() function, we can also get a 95% two-sided confidence interval for the reading score, given a math score of 76, which is from 76.348 to 77.595.

But as we can see, because the multiple R-squared value is and the adjusted R-squared value is both quite smaller than 1, the reading score cannot be adequately explained by a linear relationship with the math score.

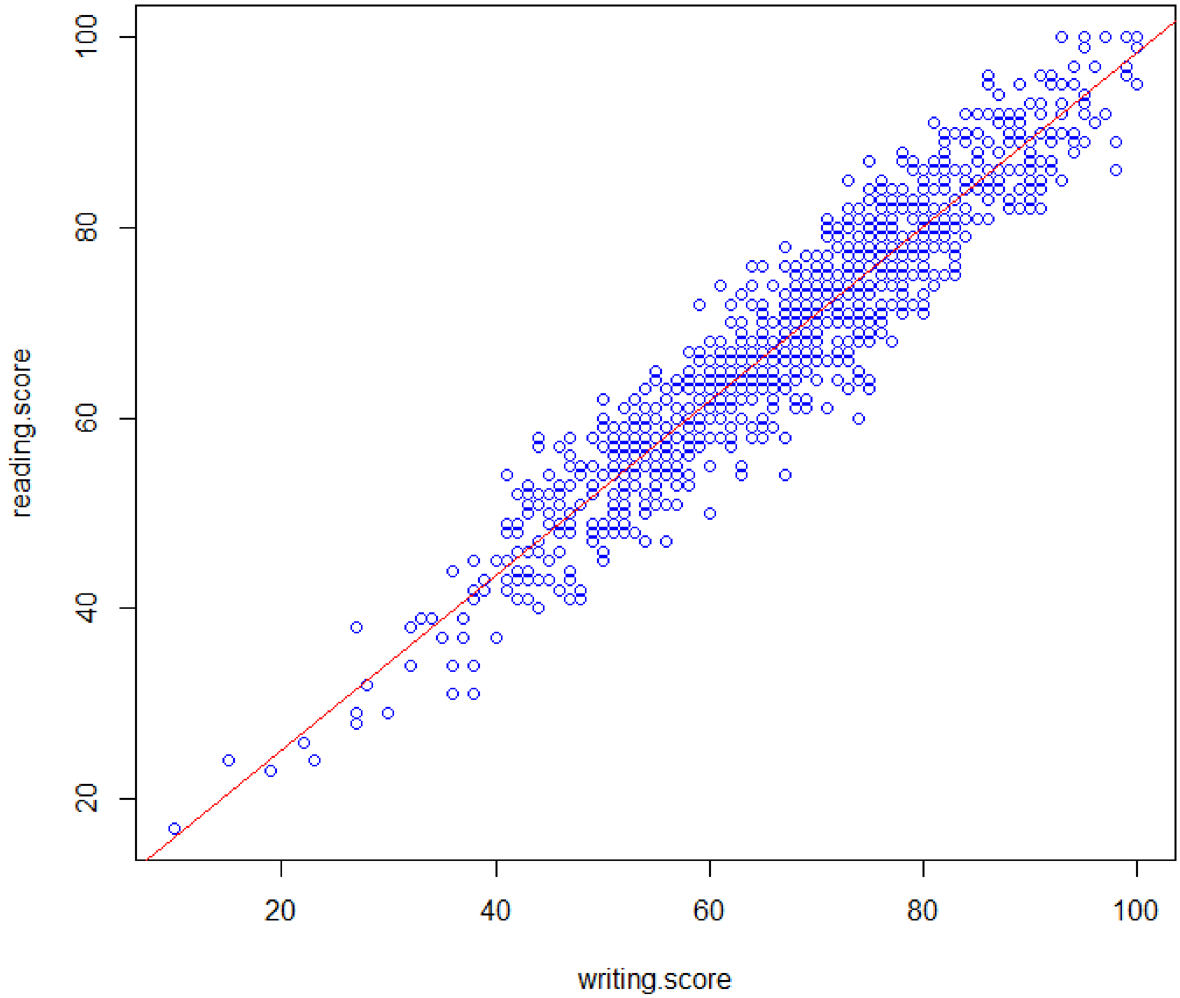
We can therefore try to explain the reading score using a different relationship. In this case, we choose to now find the linear relationship between reading score and writing score. By again applying the lm() function, we get the following results.



```
Call:
lm(formula = reading.score ~ writing.score)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.6226  -2.9554   0.1352   3.0399  11.4602 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    6.75051    0.63175   10.69   <2e-16 ***
writing.score  0.91719    0.00906  101.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.352 on 998 degrees of freedom
Multiple R-squared:  0.9113,	Adjusted R-squared:  0.9112 
F-statistic: 1.025e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

We can make many interesting comparisons between this and the earlier summary of the model between reading score and math score. But one thing that stands out is that both the multiple R-squared value and the adjusted R-squared value of this model is higher than the previous model. This is supported by the fact that B2 is closer to 1, the standard error of B2 is smaller, or that the t-value for B2 is much higher.

But perhaps a plot is the best way to illustrate this.

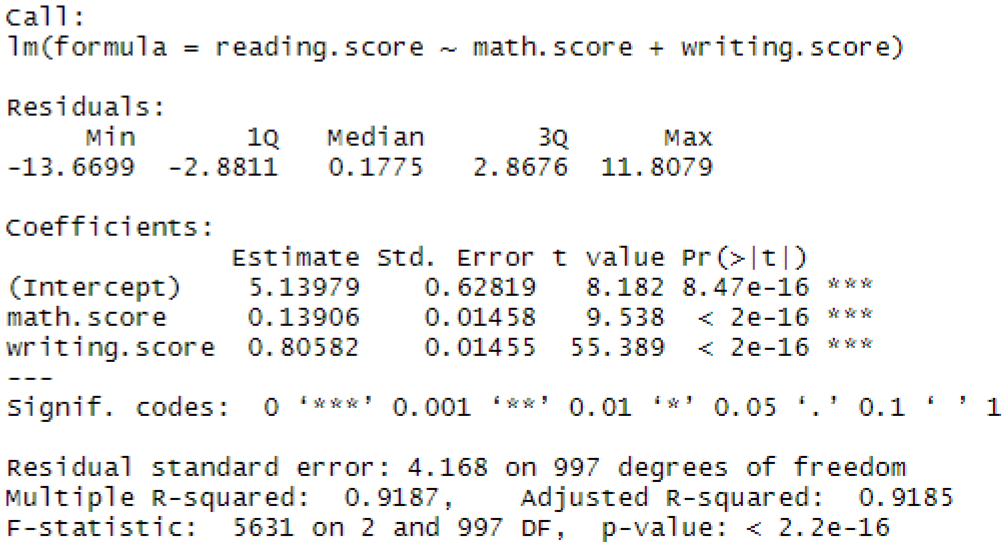


We can see that compared to the previous plot, the reading score is concentrated much closer to the line of best fit in this plot.

These evidences suggest that reading score is better explained by writing score than math score. And we can see that this makes sense, because both reading and writing are language skills, while math is more of a number skill.

b. Multiple regression:

We have just investigated how the reading score can be explained by the math score or the writing score individually. But what if the reading score can be better explained by a linear combination of the math score and the writing score. We therefore move from simple regression to multiple regression. We again apply the lm() function.



```
Call:
lm(formula = reading.score ~ math.score + writing.score)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.6699  -2.8811   0.1775   2.8676  11.8079 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    5.13979    0.62819   8.182 8.47e-16 ***
math.score     0.13906    0.01458   9.538  < 2e-16 ***
writing.score  0.80582    0.01455  55.389  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.168 on 997 degrees of freedom
Multiple R-squared:  0.9187,	Adjusted R-squared:  0.9185 
F-statistic:  5631 on 2 and 997 DF,  p-value: < 2.2e-16
```

We can see that the coefficient for writing score is much higher than that for math score, confirming our previous observation that reading score is better explained by writing score.

However, we can see that both the multiple R-squared and the adjusted R-squared is higher. Considering that the adjusted R-squared has adjusted for the fact that there are now more variables than before, 2 compared to 1, the fact that it is still higher suggests that the model is indeed better thanks to the addition of the math score and not just because we add another variable.

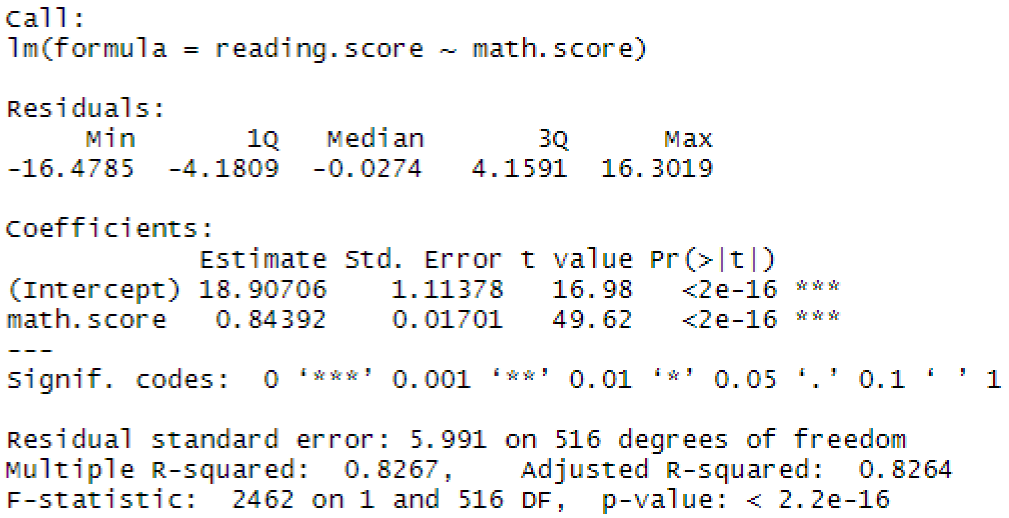
In other words, even though the reading score is better explained by the writing score than the math score, it is best explained by a combination of the writing score and the math score.

c. Conditional regression

What we have investigated so far is unconditional regression. Specifically, we investigate the relationship between the scores of general students. But it might be more meaningful if we investigate the relationship between the scores of a specific subset of the students. We might even make interesting discoveries.

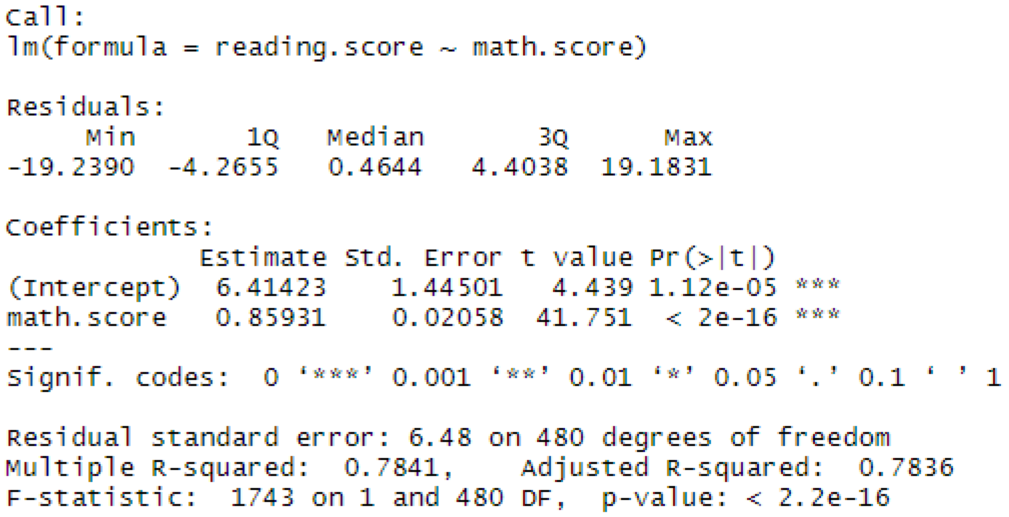
We choose to investigate the relationship between the reading score and the other scores of female students.

We construct the same 3 models as above. One interesting observation that can be made is that apparently, the reading score is better explained by the math score for female students than for students in general. We can see this from the fact the R-squared values, both multiple and adjusted, are higher than the first model.



```
Call:
lm(formula = reading.score ~ math.score)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.4785  -4.1809  -0.0274   4.1591  16.3019 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 18.90706    1.11378   16.98   <2e-16 ***
math.score   0.84392    0.01701   49.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.991 on 516 degrees of freedom
Multiple R-squared:  0.8267,	Adjusted R-squared:  0.8264 
F-statistic:  2462 on 1 and 516 DF,  p-value: < 2.2e-16
```

We repeat this procedure for the relationship between the scores of male students. Once again, we find that the reading score is better explained by the math score for male students than for students in general.



```
Call:
lm(formula = reading.score ~ math.score)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.2390  -4.2655   0.4644   4.4038  19.1831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.41423    1.44501   4.439 1.12e-05 ***
math.score   0.85931    0.02058  41.751  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.48 on 480 degrees of freedom
Multiple R-squared:  0.7841,	Adjusted R-squared:  0.7836 
F-statistic:  1743 on 1 and 480 DF,  p-value: < 2.2e-16
```

This is somewhat confusing because if the reading score of a female student is adequately explained by the math score and the same is true for a male student, why is the math score of a general student poorly explained by their math score?

On closer inspection, we find that the model equation for the relationship between reading score and math score for female students is:

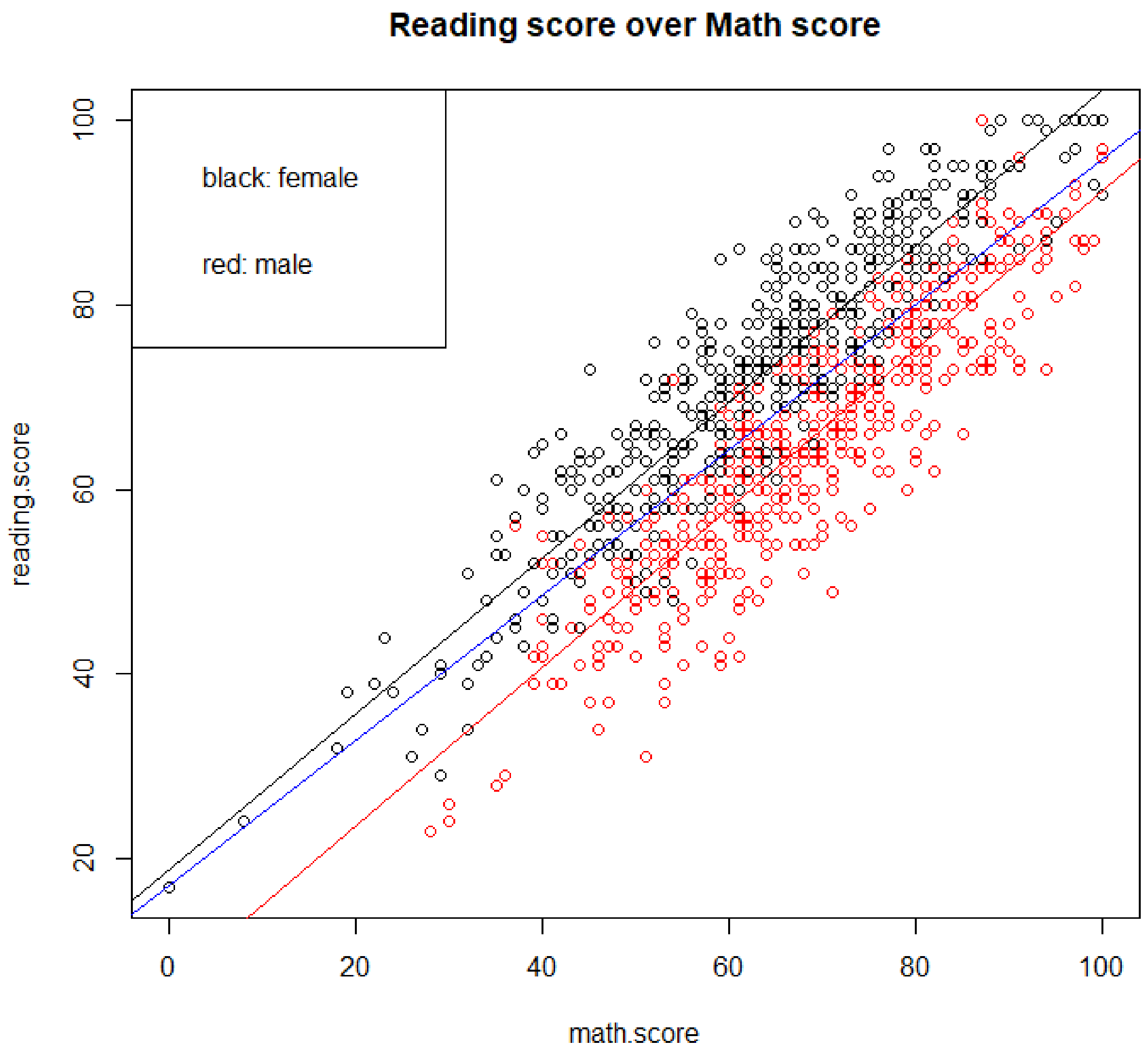
reading.score = 18.907 + 0.844 * math.score

The same equation for male students is:

reading.score = 6.414 + 0.859 * math.score

We can see that even though the B2 coefficients for the two models are somewhat equal, the B1 coefficients are vastly different. Therefore, when we combine the scores of male and female students together, there isn’t really a way to combine the two above equations together to get an accurate model equation for students in general.

We can see it very clearly from this plot.



The plot is a scatterplot of reading scores and math scores, where the dots have been color-coded by gender. On the plot are three lines, the black line is the line of best fit for the black dots, the red line is the line of best fit for the red dots, and the blue line is the line of best fit for all of the dots.

As we can see, the reading score for female students can be adequately explained by the math score and the same is true for male students. Therefore, the dots for the female students are concentrated around a line and so are the dots for male students. But because that for a given math score, female students tend to have higher reading score, we can see that the dots for female students tend to be gathered above those of male students.

We therefore have effectively two separate set of dots, one for female students and one for male students. Each set is concentrated around a line but because the sets are separate from each other, when we combine the two sets together, there isn't a good line of best fit.

We can see that conditional regression has allowed us to discover the important difference between the reading score and the math score of female and male students.

III/ Conclusion:

In this report, we have seen how we can make use of different aspects of statistics to better understand a dataset. More specifically, we have used many functions in R to help us understand the factors that affect the test scores of students at a public school.