Student Performance Dataset Analysis

Benedikt Willecke
IES19372

Introduction

This is a dataset describing the exam score in math, reading and writing. There are many categorical variables: Gender, race/ethnicity, parental level of education, lunch and test preparation course. The ethnicity I divided into groups A-E but it is not specified which ethnicity it is. In total this dataset includes 8 variables with 1000 data points.

Exploratory Data Analysis

First, we should load the data and take a first look at it. Here we can see what values the variables have. The scores seem to have a out-of-100 score. For the lunch plan there seems to be at least one free/reduced plan. The level of parental education is described be the degree.

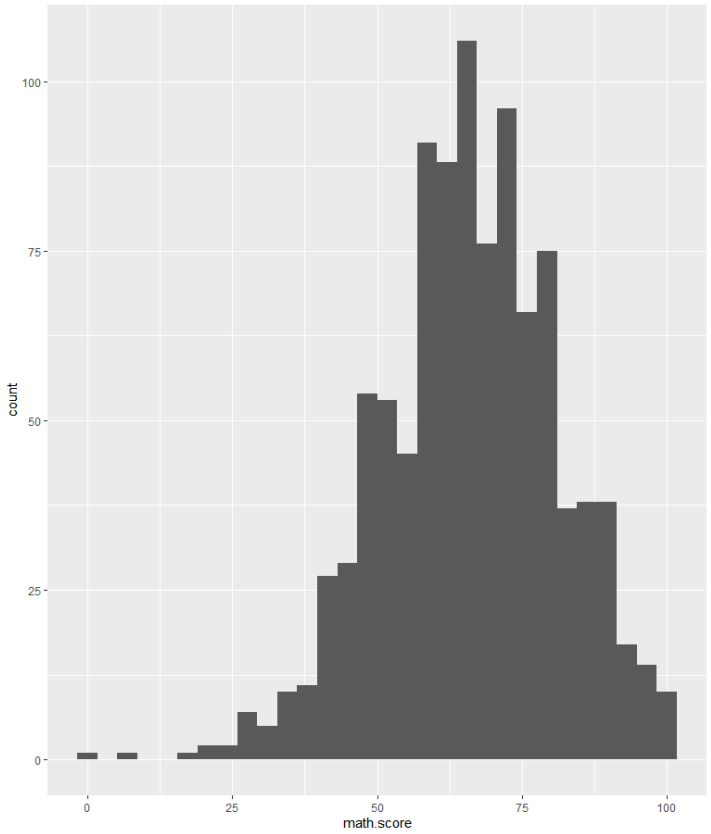
```
> data<-read.csv("C:\\Users\\HaHaK\\Google Drive\\Studium\\01 Bachelor\\6. Semester SS20\\Multivariate Statistical Analysis\\Project - Stu
dent Performance Data\\StudentsPerformance.csv")
> head(data)
  gender race.ethnicity parental.level.of.education       lunch test.preparation.course math.score reading.score writing.score
1 female        group B           bachelor's degree    standard                  none         72            72            74
2 female        group C                some college    standard             completed         69            90            88
3 female        group B            master's degree    standard                  none         90            95            93
4   male        group A         associate's degree free/reduced                  none         47            57            44
5   male        group C                some college    standard                  none         76            78            75
6 female        group B         associate's degree    standard                  none         71            83            78
>
```

Next, let's take a look at the summary of the data. Here we can observe several things. Firstly, there are more datapoints for females than for males. Secondly, the ethnicity groups are not equally distributed, however they seem to be pretty diverse. For the parental level of education surprisingly there seems to be no parent with a doctoral degree. Other than that, except for bachelors and Masters degree all other degrees seem to be pretty equally distributed. For the lunch there is only the standard and free/reduced option where 2/3 have the former and 1/3 the latter lunch plan. For the preparation only 1/3 of students took the course. The test scores are indeed on a scale from 0 to 100. The mean is at about 70% with math slightly lower. In all exams at least one student got 100 points and only in the math exam at least one student got 0 points.

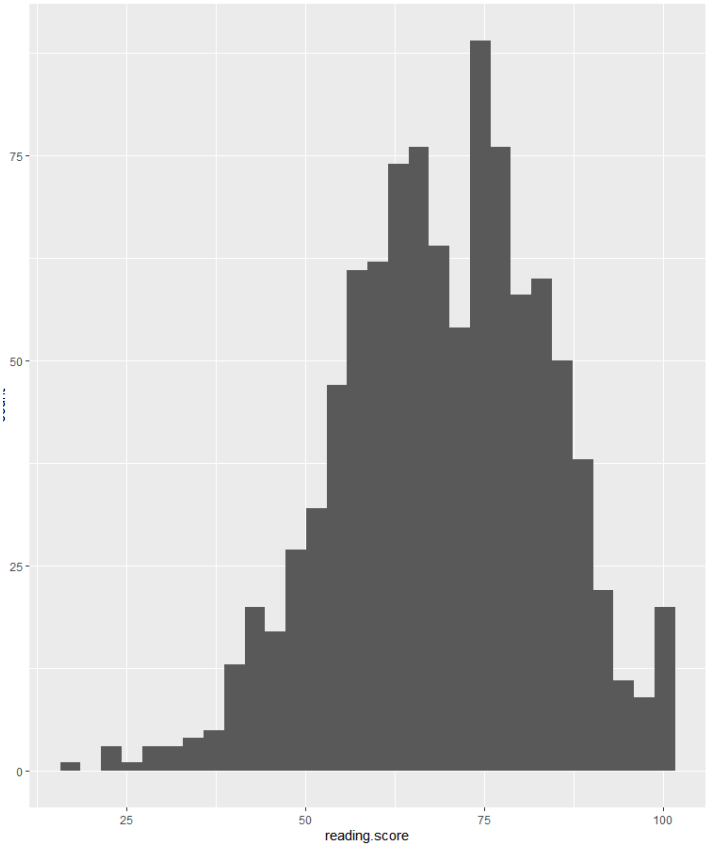
```
> summary(data)
   gender     race.ethnicity      parental.level.of.education         lunch     test.preparation.course   math.score
female:518   group A: 89    associate's degree:222          free/reduced:355   completed:358           Min.   :  0.00
male  :482   group B:190    bachelor's degree :118          standard    :645   none     :642           1st Qu.: 57.00
             group C:319    high school       :196                                                    Median : 66.00
             group D:262    master's degree   : 59                                                    Mean   : 66.09
             group E:140    some college      :226                                                    3rd Qu.: 77.00
                            some high school  :179                                                    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
```

Let's now look at how the test scores are distributed. Here we can see that all test scores are negatively skewed so the peak on the positive said. In fact, the peak is quite sharp. We can also see that in the math and reading score there are two or more peaks where the writing score as one very sharp peak. Also, the graph of the reading and writing score actually goes up when reaching 100 points, meaning more people got perfect score than let's say 90.

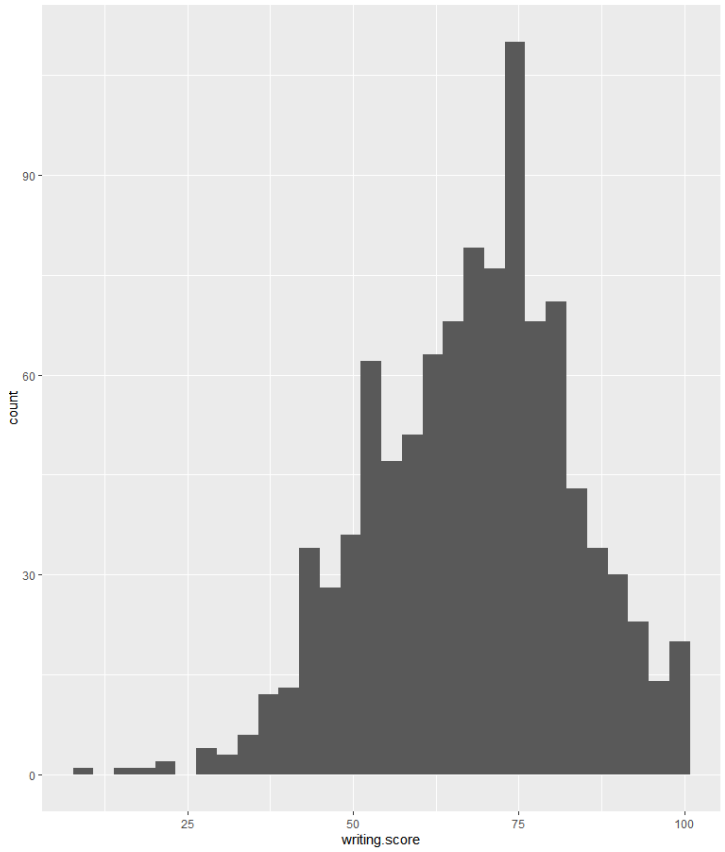
```
> ggplot(data=data, aes(x=math.score))+geom_histogram()
```




```
ggplot(data=data, aes(x=reading.score))+geom_histogram()
```

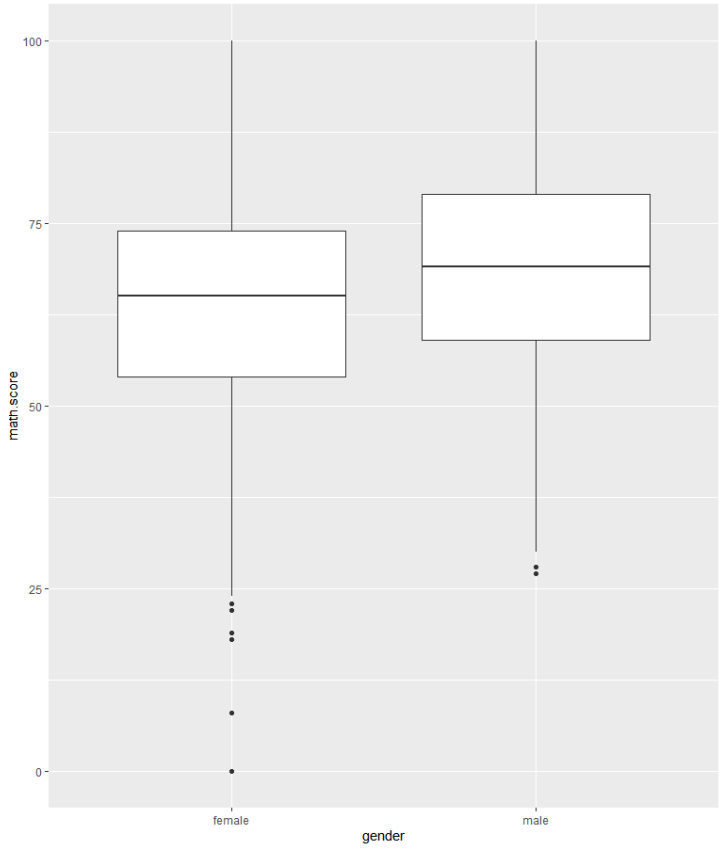


```
ggplot(data=data, aes(x=writing.score))+geom_histogram()
```

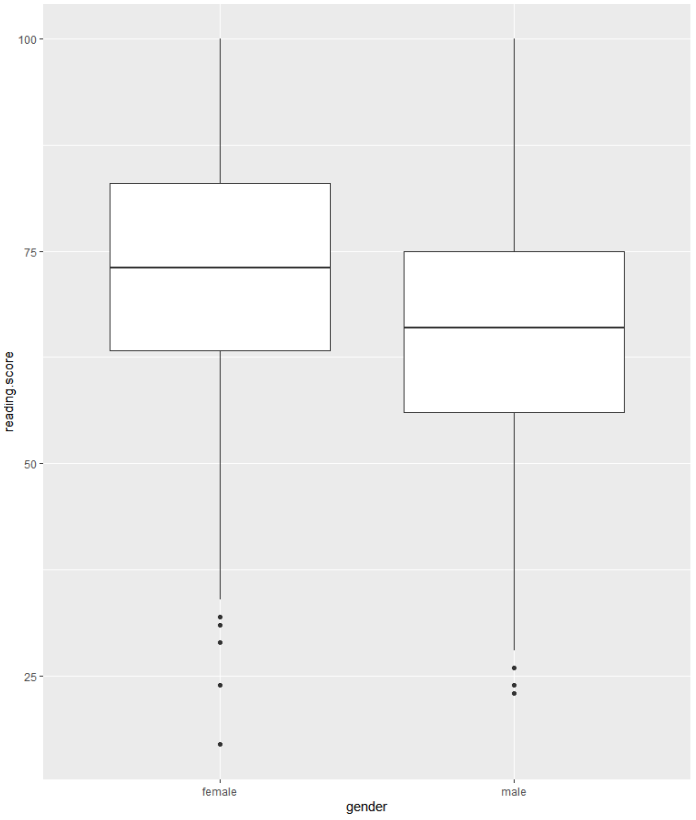


Next let's look at the boxplots of the scores grouped by gender. Firstly, we can confirm our previous observation about the general distribution of the data. Additionally, we can observe a lot of outliers on the lower score side and never on the upper side. This seems to be the case throughout subject and gender. Secondly, we can observe a clear difference between male and female. In general males seem to get better scores in math whereas females score better in reading and writing. But the difference of means in math is less than the difference of means in reading and writing. However, there seem to be more female low score outliers than male ones.

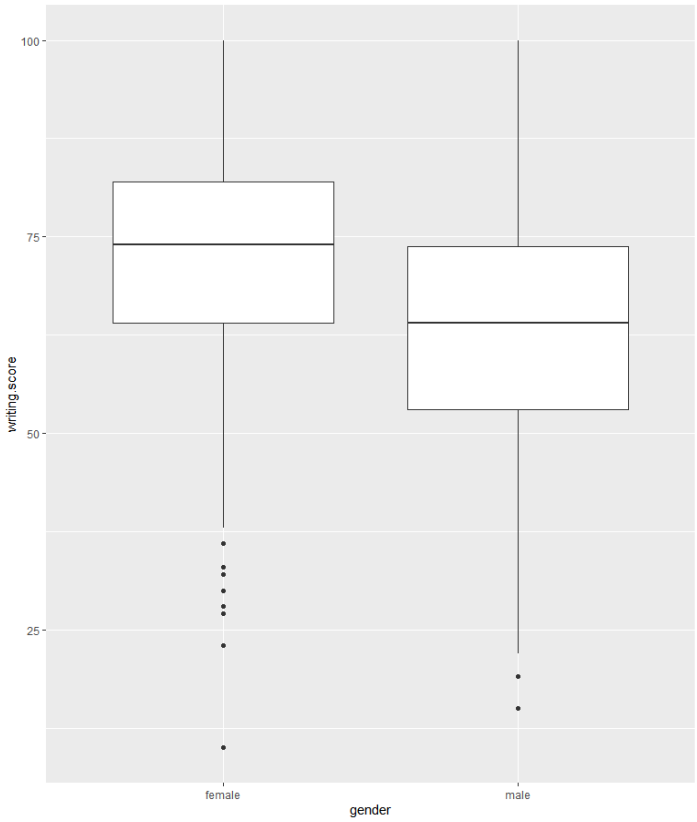
```
ggplot(data=data, aes(x=gender, y=math.score))+geom_boxplot()
```



```
ggplot(data=data, aes(x=gender, y=reading.score))+geom_boxplot()
```




```
ggplot(data=data, aes(x=gender, y=writing.score))+geom_boxplot()
```



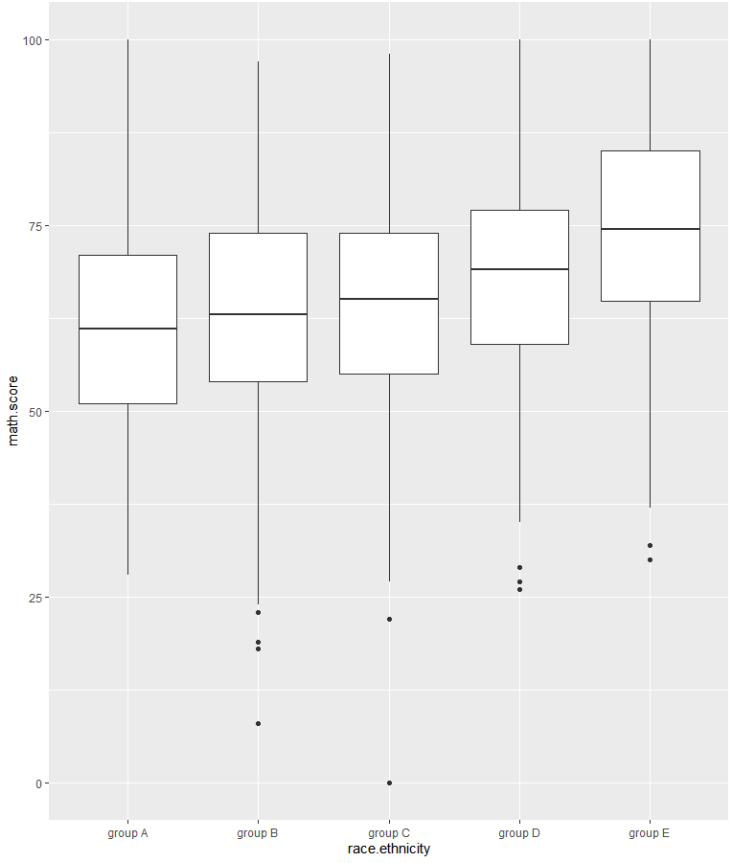
Future Analysis

In my future analysis I would like to do more comparisons like in the group gender. This could be done for example to test the effectiveness of the preparation course and what the major factors are that contribute to a good test score. The factors could be analyzed via PCA. Also, I could see how to distinguish some group like ethnicity, gender or parental educational background by using LDA. In general, I would like to explore the difference between certain groups and see what important factors for a good score are. Ideally, I would also like to find out what the best way is to improve the test scores.

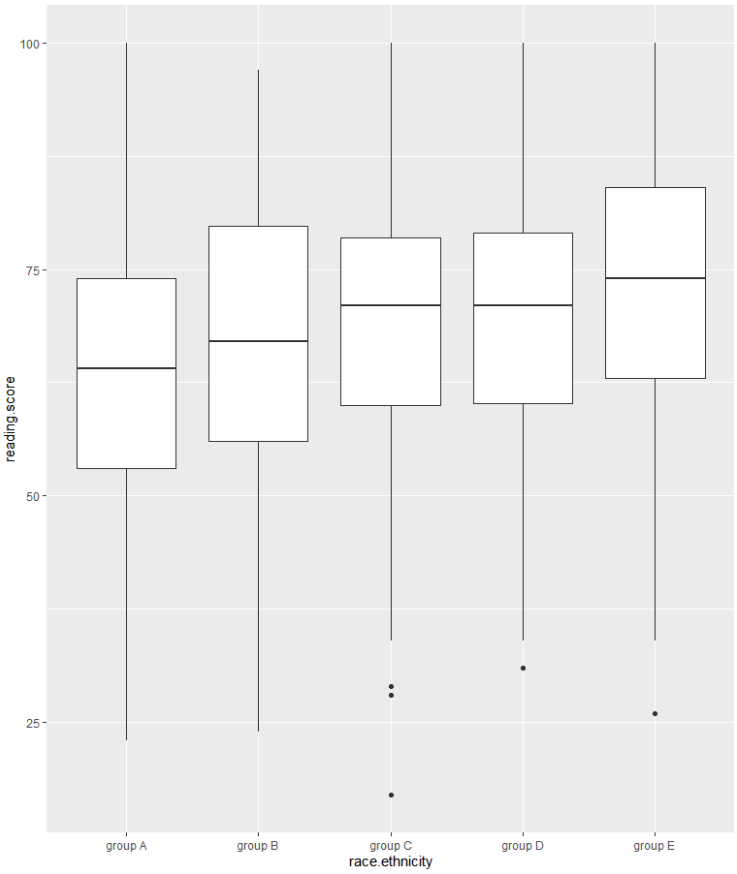
Test Scores and Ethnicity

From this analysis we can conclude that every ethnicity group has higher score than the previous one. Although the difference is not as large in reading and writing as in math. However, throughout all tests between group A and group E there seems to be a difference of at least 10 points.

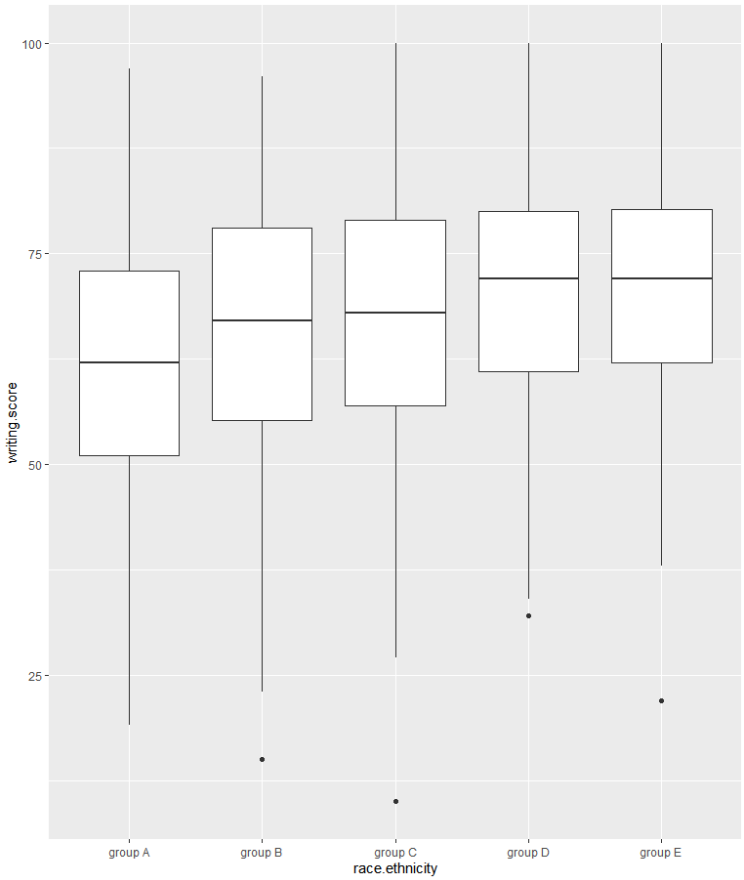
```
ggplot(data=data, aes(x=race.ethnicity, y=math.score))+geom_boxplot()
```



```
ggplot(data=data, aes(x=race.ethnicity, y=reading.score))+geom_boxplot()
```



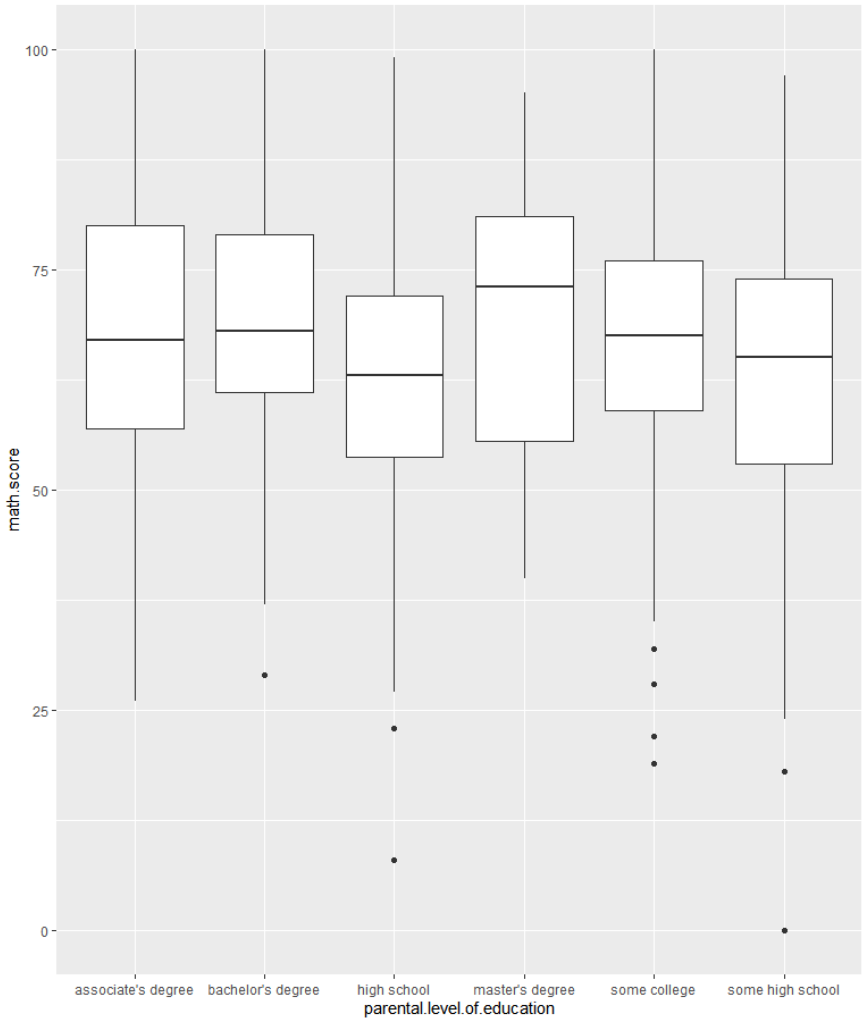

```
ggplot(data=data, aes(x=race.ethnicity, y=writing.score))+geom_boxplot()
```



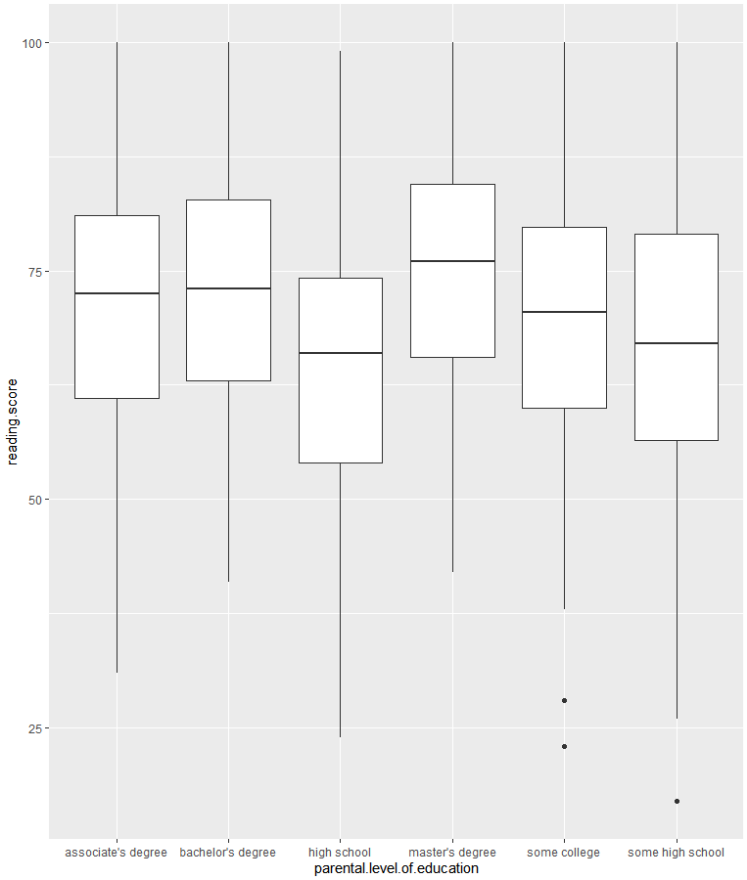
Test Scores and Parental Level of Education

Generally, there seems to be a correlation between the parental level of education and their children's score. Kids with parents that have a Masters degree seem to score higher throughout test compared to other students. The difference seems to be same in reading, writing and math as well. The only difference is that in reading and writing students whose parents have a bachelors or associates degree have a higher average score in reading and writing in comparison to their math score.

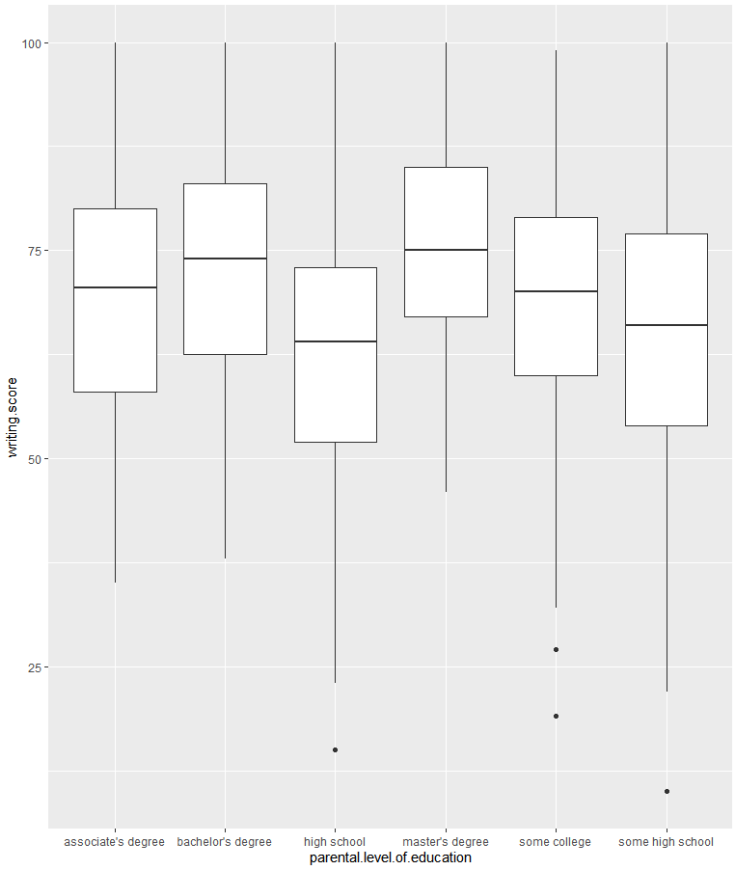
```
ggplot(data=data, aes(x=parental.level.of.education, y=math.score))+geom_boxplot()
```



```
ggplot(data=data, aes(x=parental.level.of.education, y=reading.score))+geom_boxplot()
```



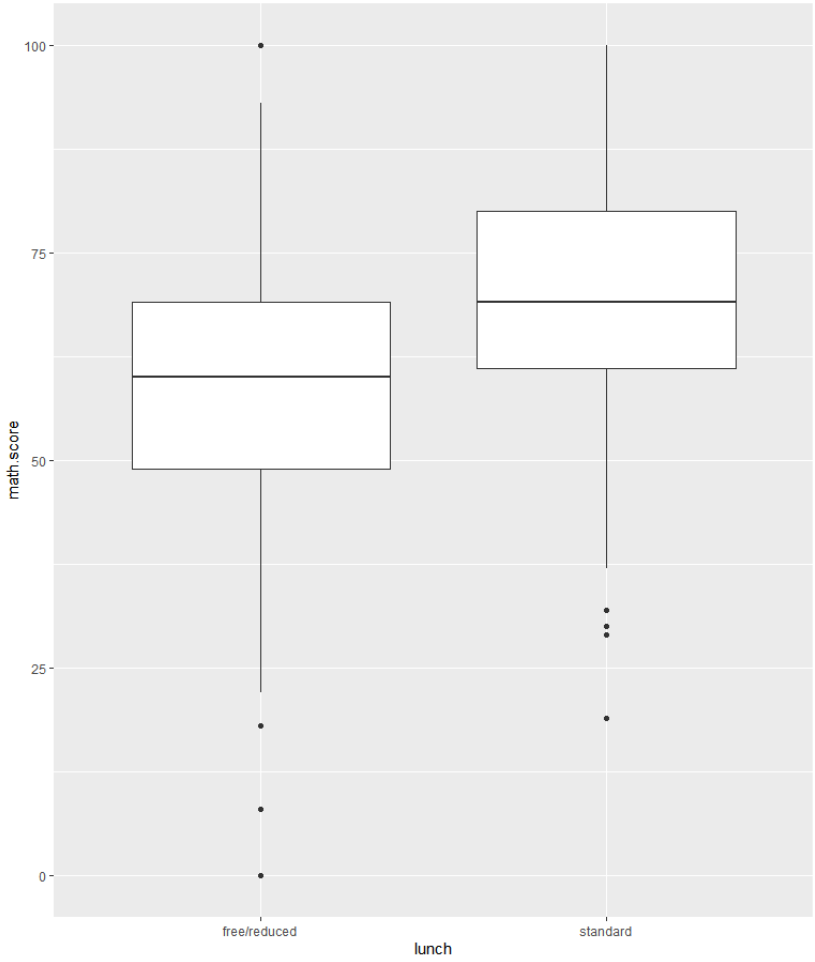

```
ggplot(data=data, aes(x=parental.level.of.education, y=writing.score))+geom_boxplot()
```



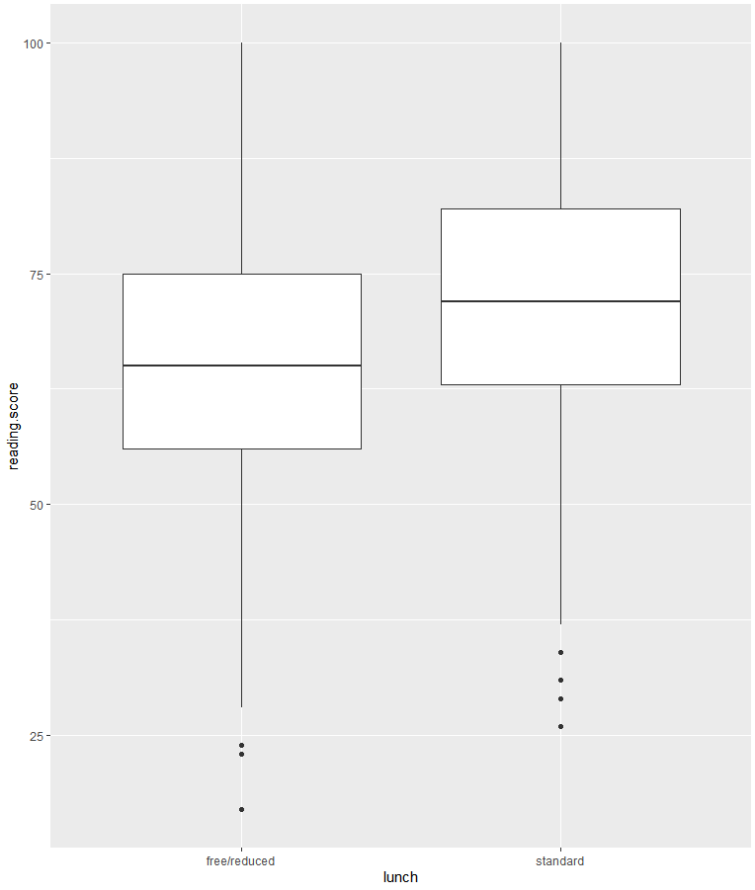
Test Scores and Lunch Plan

From this analysis we can conclude that students with free/reduced lunch plan score lower than students with the standard plan. The difference is more severe in math than in reading and writing. However, it is important to know that this only shows correlation and not causation.

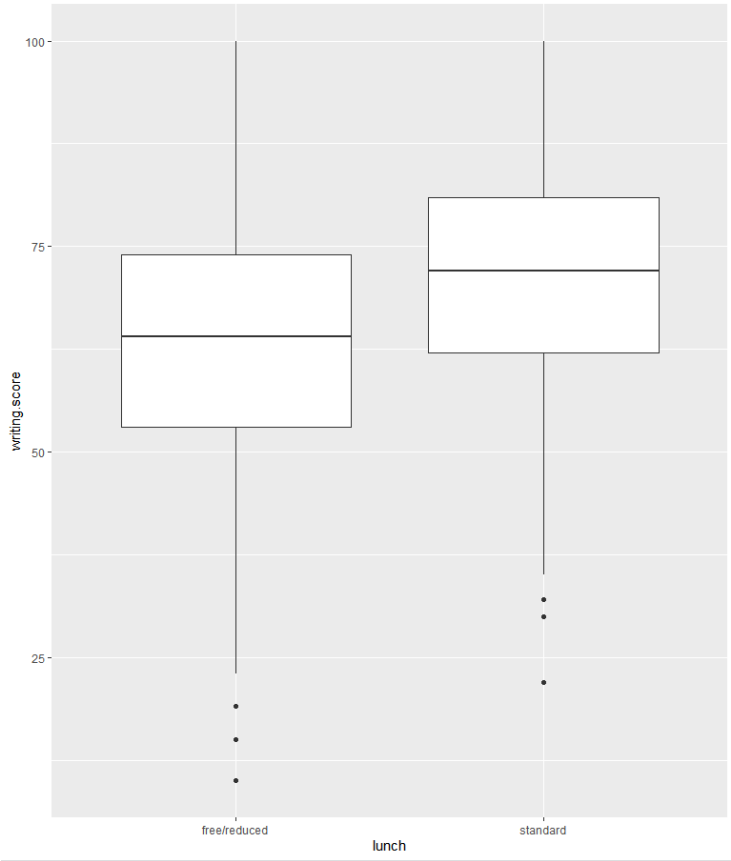
```
ggplot(data=data, aes(x=lunch, y=math.score))+geom_boxplot()
```



```
ggplot(data=data, aes(x=lunch, y=reading.score))+geom_boxplot()
```



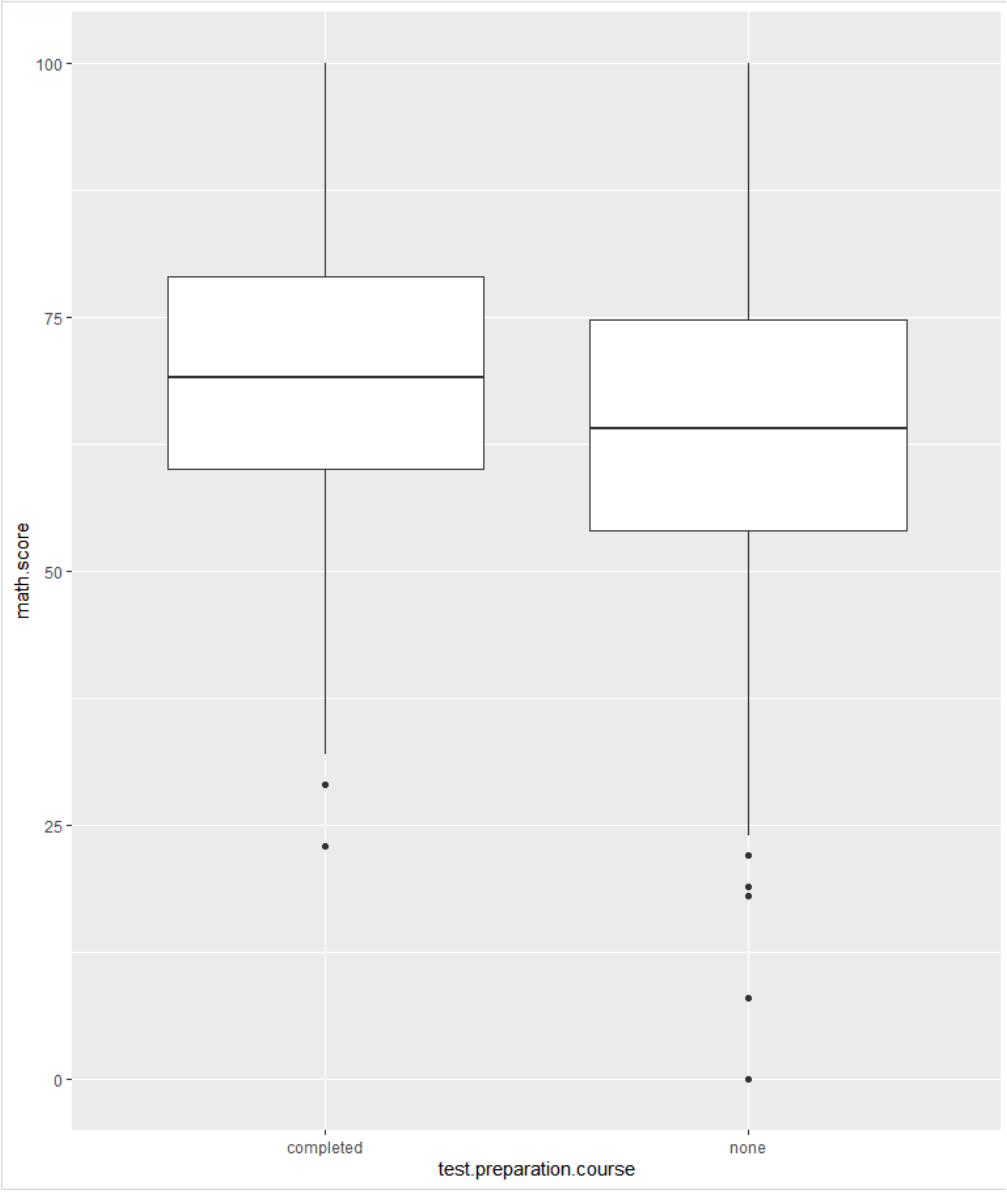

```
ggplot(data=data, aes(x=lunch, y=writing.score))+geom_boxplot()
```



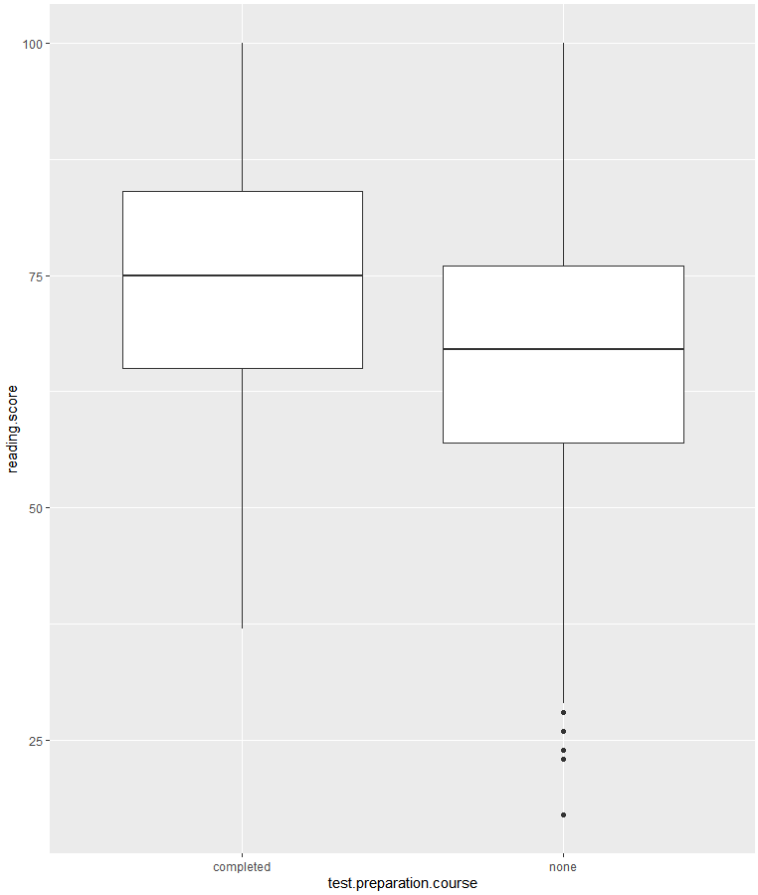
Test Scores and Test Preparation Course

From this analysis we can see how effective the preparation course is at improving the test course. We can see that the course generally improves the student's performance. However, the reading and writing scores saw more improvement than the math score. We also get more insights when looking at the gender differences here. Here we can confirm our previous observation that on average male students performed better in math whereas female students got better scores in reading and writing. We can see that the worse performing gender in the respective test has the same performance with the preparation course as the opposite gender did without. From this we can conclude that indeed the course is effective at improving students' scores. However, I would suggest a more extensive course for the gender that historically has done worse in the respective category in order to give everybody the support they need.

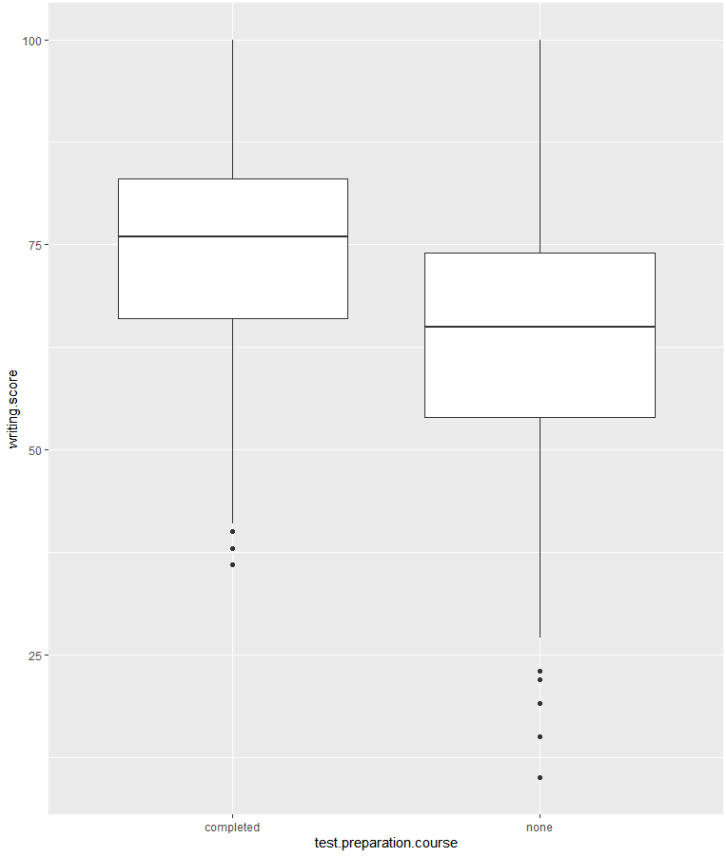
```
ggplot(data=data, aes(x=test.preparation.course, y=math.score))+geom_boxplot()
```



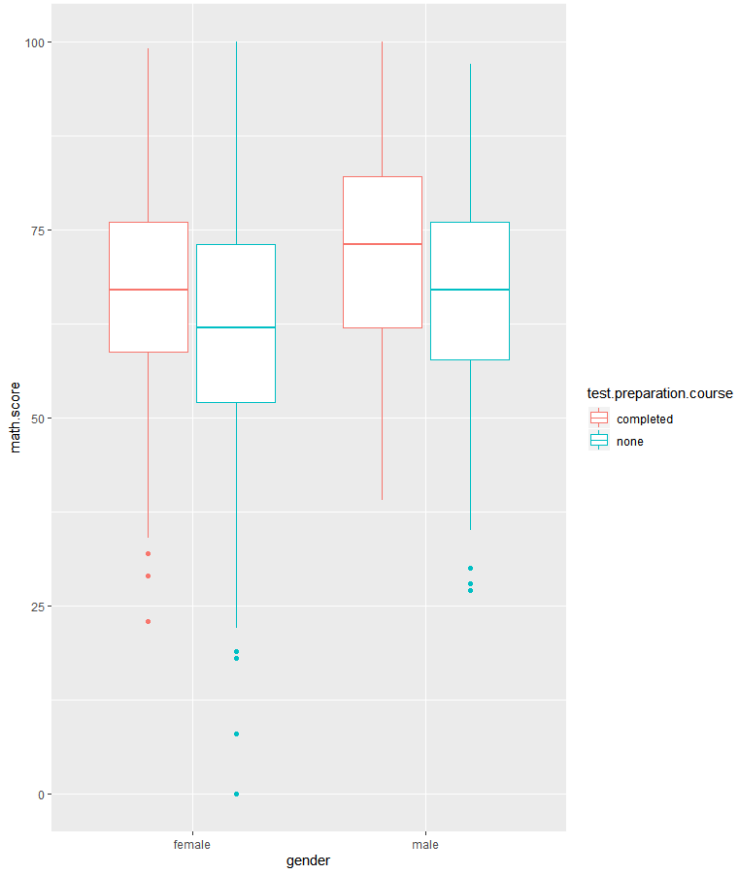
```
ggplot(data=data, aes(x=test.preparation.course, y=reading.score))+geom_boxplot()
```



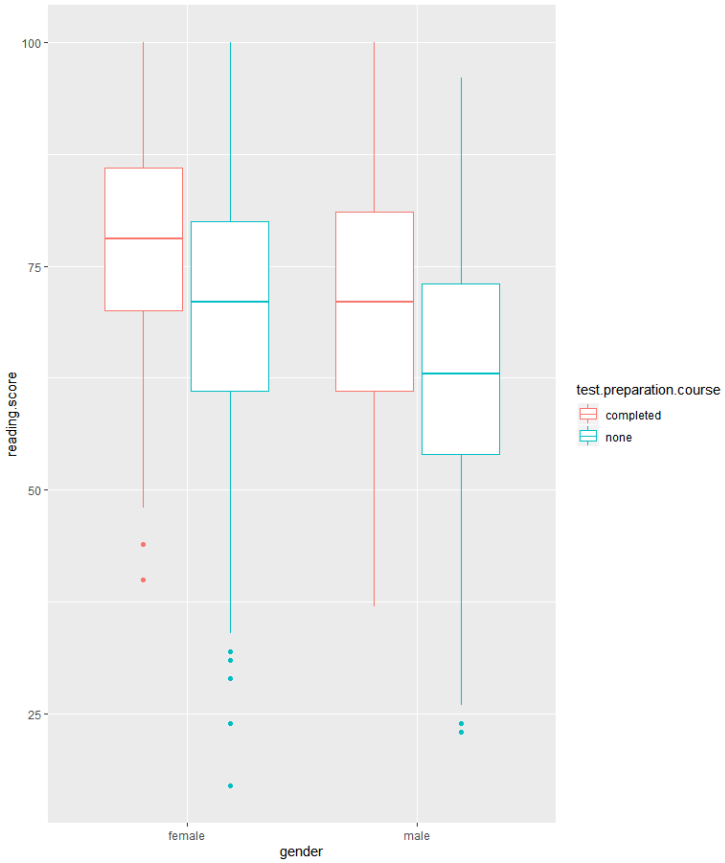

```
ggplot(data=data, aes(x=test.preparation.course, y=writing.score))+geom_boxplot()
```



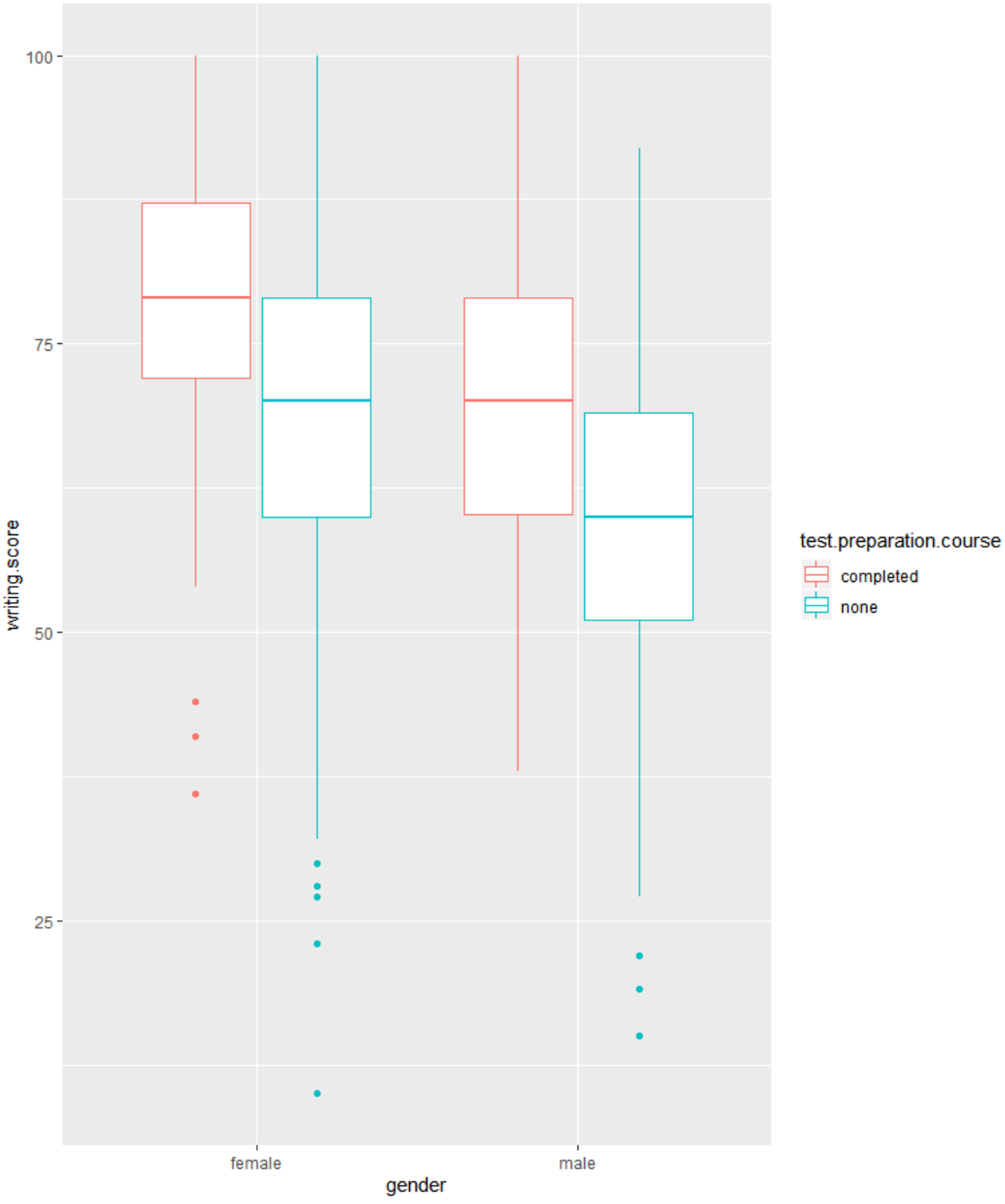
```
ggplot(data, aes(gender, math.score, color = test.preparation.course))+geom_boxplot()
```




```
ggplot(data, aes(gender, reading.score, color = test.preparation.course))+geom_boxplot()
```



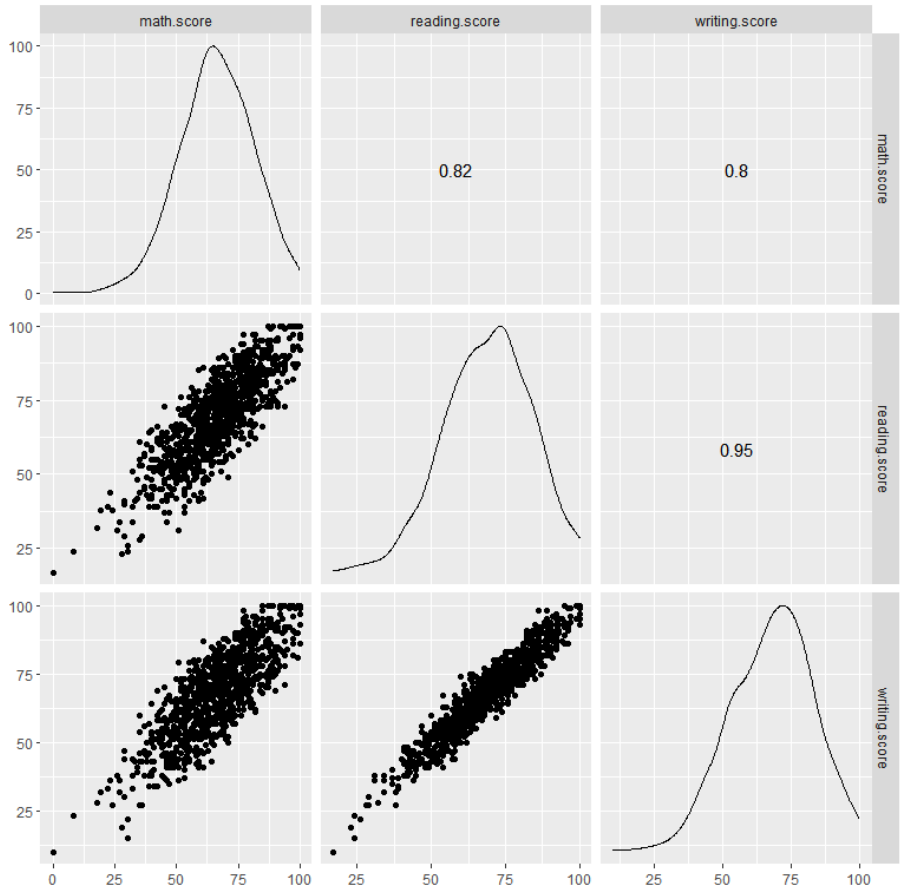
```
ggplot(data, aes(gender, writing.score, color = test.preparation.course))+geom_boxplot()
```



Test Score Correlation

From this we can simply see that reading and writing are more correlated with each other than with mathematics. This makes sense logically as well. However, we can also see that generally good students get good scores in all exams whereas generally bad students get bad grades throughout all tests.

```
ggscatmat(data)
```



PCA

We can see that the first principal component is enough to describe the data with 90% accuracy. The first PC is the overall performance. The second one is the difference between the math and reading/writing score.

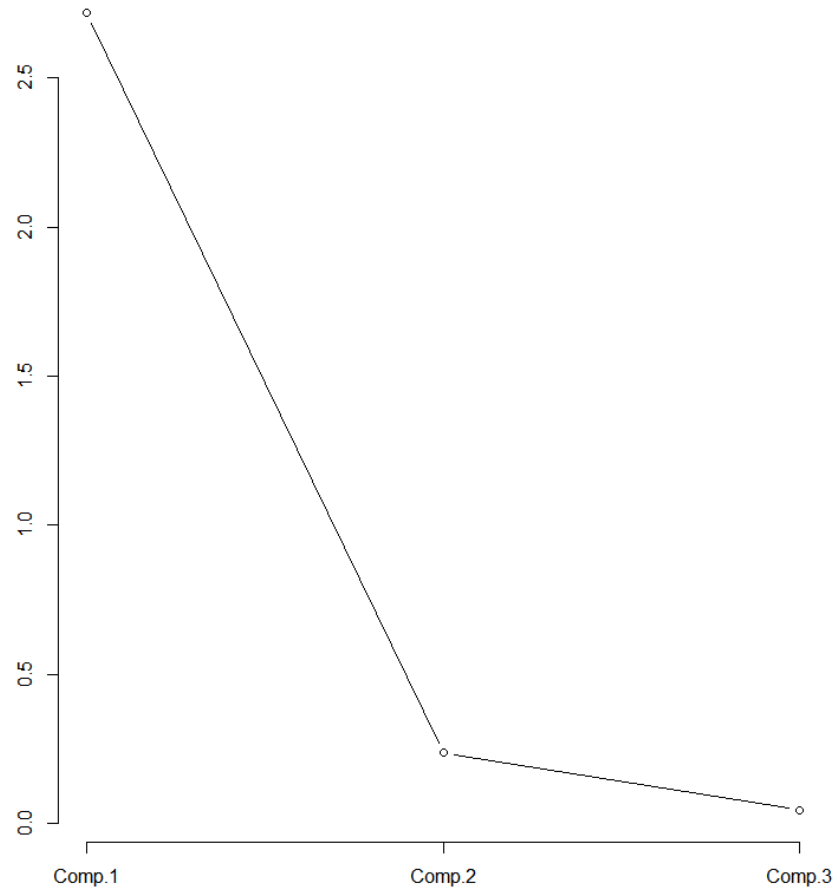
```
> pca.result<-princomp(data[,-1:-5],cor=TRUE)
> pca.result
Call:
princomp(x = data[, -1:-5], cor = TRUE)

Standard deviations:
   Comp.1    Comp.2    Comp.3
1.6487661 0.4864002 0.2120970

 3  variables and  1000 observations.
> summary(pca.result)
Importance of components:
                          Comp.1     Comp.2     Comp.3
Standard deviation     1.6487661 0.48640017 0.21209696
Proportion of Variance 0.9061433 0.07886171 0.01499504
Cumulative Proportion  0.9061433 0.98500496 1.00000000
> pca.result$loadings

Loadings:
              Comp.1 Comp.2 Comp.3
math.score     0.555  0.831
reading.score  0.590 -0.359 -0.723
writing.score  0.587 -0.424  0.690

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000
> screeplot(pca.result,type="lines")
```



LDA

Here we can see that we can predict the gender with 90% accuracy given all the other values. This indicates that all these variables are very dependent on each other/affect each other.

```
> test.data.id<-c(sample(1:500,100),sample(501:1000,100))
> train.data<-data[-test.data.id,]
> test.data<-data[test.data.id,]
```

```
> lda.result<-lda(gender~.,data=train.data)
> lda.result
Call:
lda(gender ~ ., data = train.data)

Prior probabilities of groups:
female   male
  0.52   0.48

Group means:
       race.ethnicitygroup B race.ethnicitygroup C race.ethnicitygroup D race.ethnicitygroup E
female             0.1971154             0.3653846             0.2403846             0.1298077
male               0.1718750             0.2968750             0.2708333             0.1510417
       parental.level.of.educationbachelor's degree parental.level.of.educationhigh school
female                                    0.1225962                                0.1899038
male                                      0.1145833                                0.2083333
       parental.level.of.educationmaster's degree parental.level.of.educationsome college
female                                  0.06971154                                0.2139423
male                                    0.04687500                                0.2213542
       parental.level.of.educationsome high school lunchstandard test.preparation.coursenone math.score reading.score writing.score
female                                   0.1778846     0.6346154                    0.6370192   63.51442      72.45913      72.41587
male                                     0.1927083     0.6822917                    0.6458333   68.88281      65.74479      63.59896

Coefficients of linear discriminants:
                                                LD1
race.ethnicitygroup B                     -0.344879441
race.ethnicitygroup C                     -0.254502473
race.ethnicitygroup D                      0.062613027
race.ethnicitygroup E                     -0.984254979
parental.level.of.educationbachelor's degree  0.298777419
parental.level.of.educationhigh school    -0.150116017
parental.level.of.educationmaster's degree    0.305207825
parental.level.of.educationsome college   -0.013180187
parental.level.of.educationsome high school  -0.259638269
lunchstandard                             -0.242453653
test.preparation.coursenone               -0.924102741
math.score                                 0.176357487
reading.score                             -0.005014391
writing.score                             -0.187776221
```

```
> train.pc<-predict(lda.result,train.data)$class
> test.pc<-predict(lda.result,test.data)$class
```

```
> table(train.data$gender,train.pc)
        train.pc
         female male
  female    381   35
  male       41  343
```

```
> train.miss.rate<-mean(train.pc!=train.data$gender)
> train.miss.rate
[1] 0.095
```

```
> table(test.data$gender,test.pc)
        test.pc
         female male
  female     94    8
  male       10   88
> test.miss.rate<-mean(test.pc!=test.data$gender)
> test.miss.rate
[1] 0.09
```

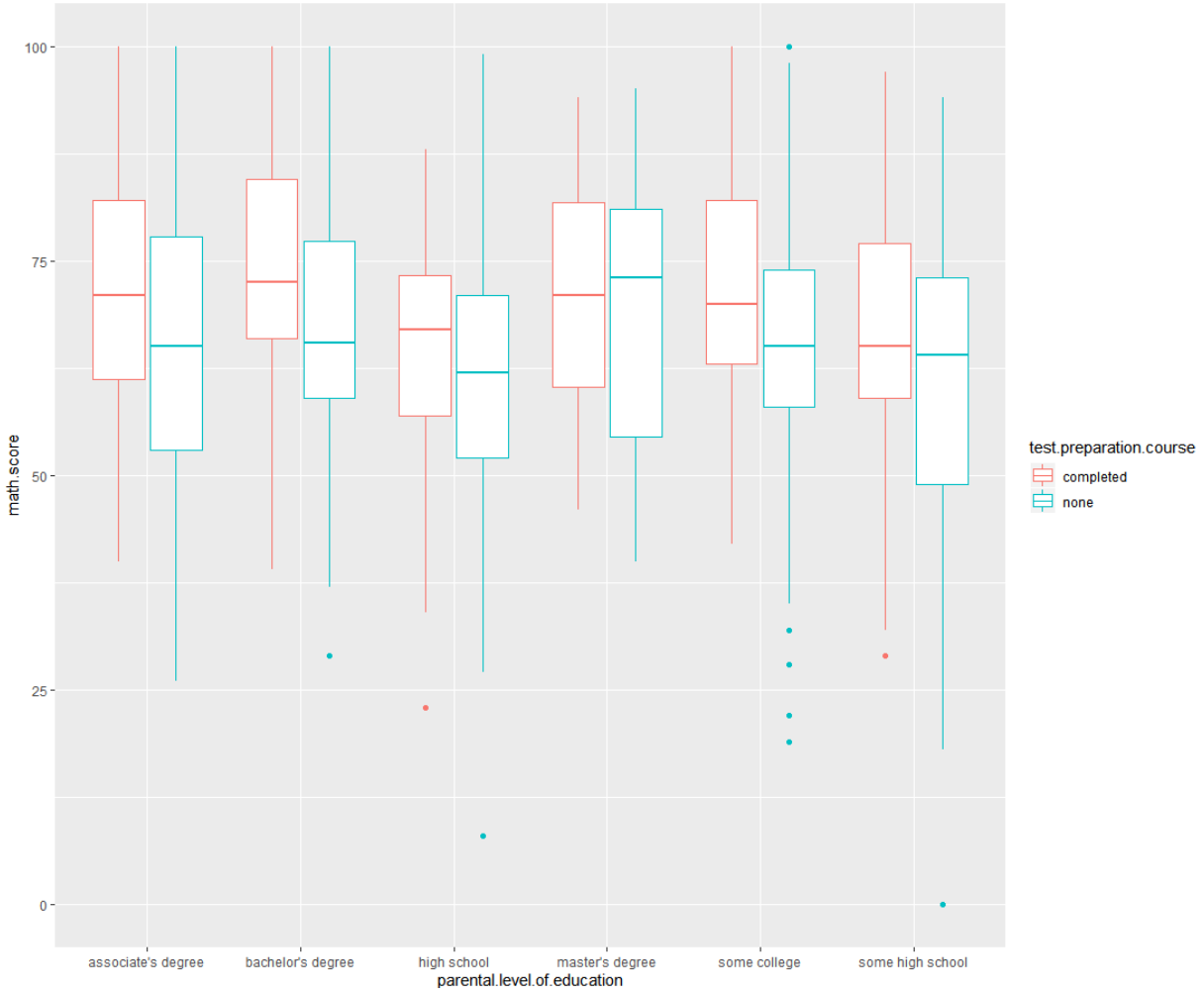
Factor for Test Scores

We saw that all categories(gender, lunch plan etc.) affect the student's performance. In order to provide all students, the support they need the school wants to improve on the preparation course, as this is the only variable that the school can change. In the following I analyses which groups are not targeted well by the course. We already saw that gender plays a role. Boys seem to need more support in reading and writing whereas girls need more in math.

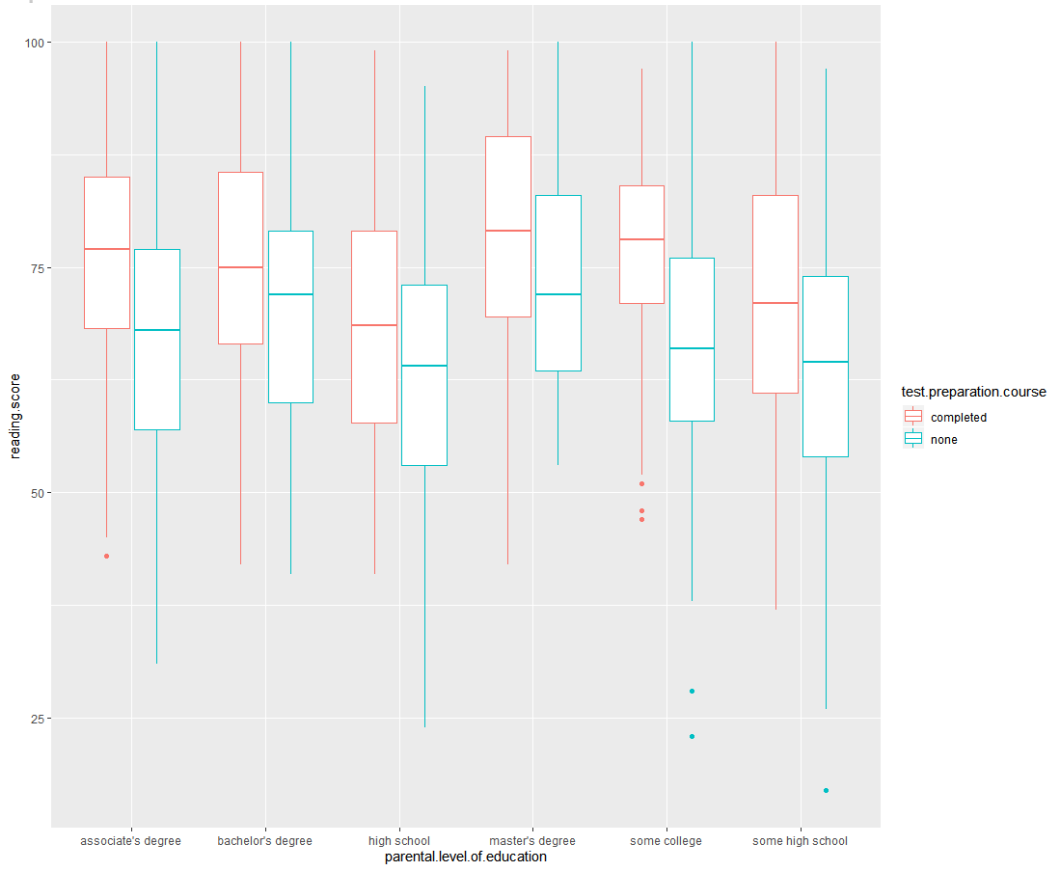
Parental Level of Education

While the reading and writing scores improve throughout all parental educational backgrounds, we see a lot of difference in the math score improvements. Kids with parents who have a Masters degree already perform well and see almost no improvement. Kids with parents who only have a high school degree already perform bad compared to their peers and see no improvement either. "some college" and "Bachelors degree" backgrounds on the other hand benefit a lot from this course and "associate's degree" as well as "some high school" backgrounds see an intermediate improvement. From this we can conclude that the course is designed in a way that does not benefit everybody in the same way.

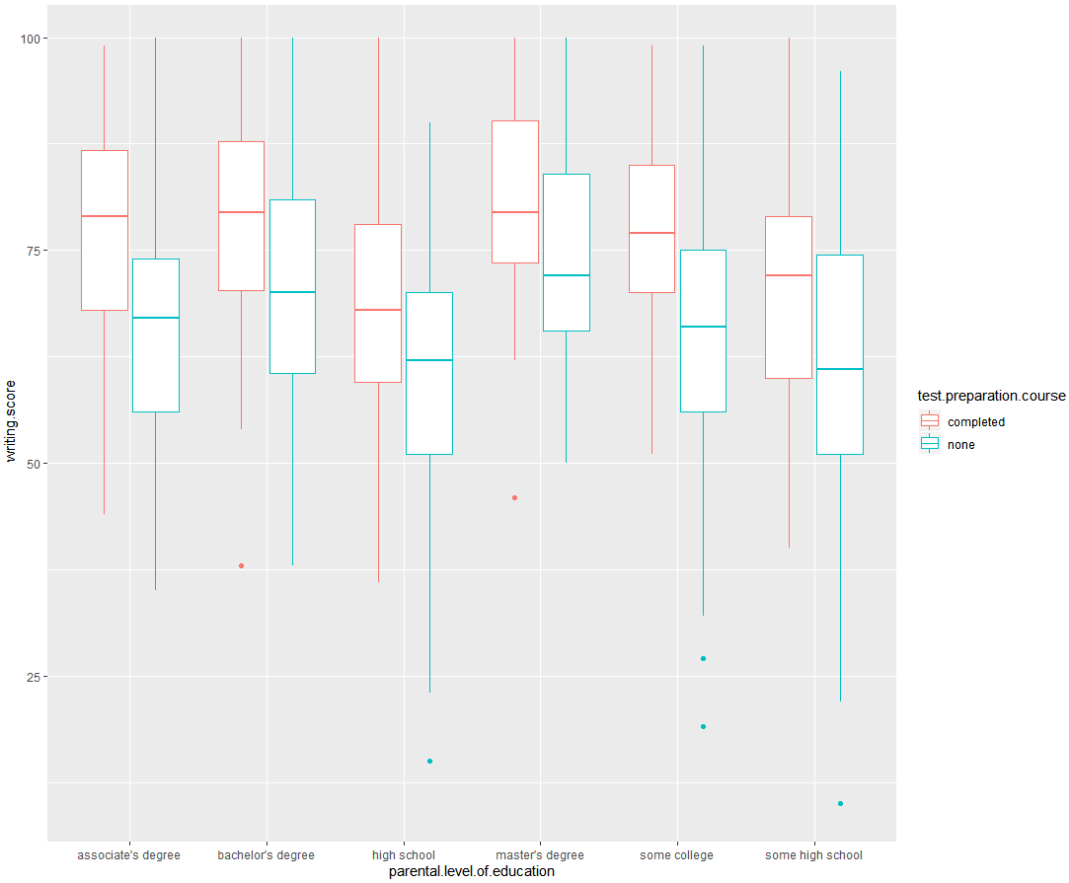
```
> ggplot(data, aes(parental.level.of.education, math.score, color=test.preparation.course))+geo
m_boxplot()
```



```
> ggplot(data, aes(parental.level.of.education, reading.score, color=test.preparation.course))+
geom_boxplot()
```



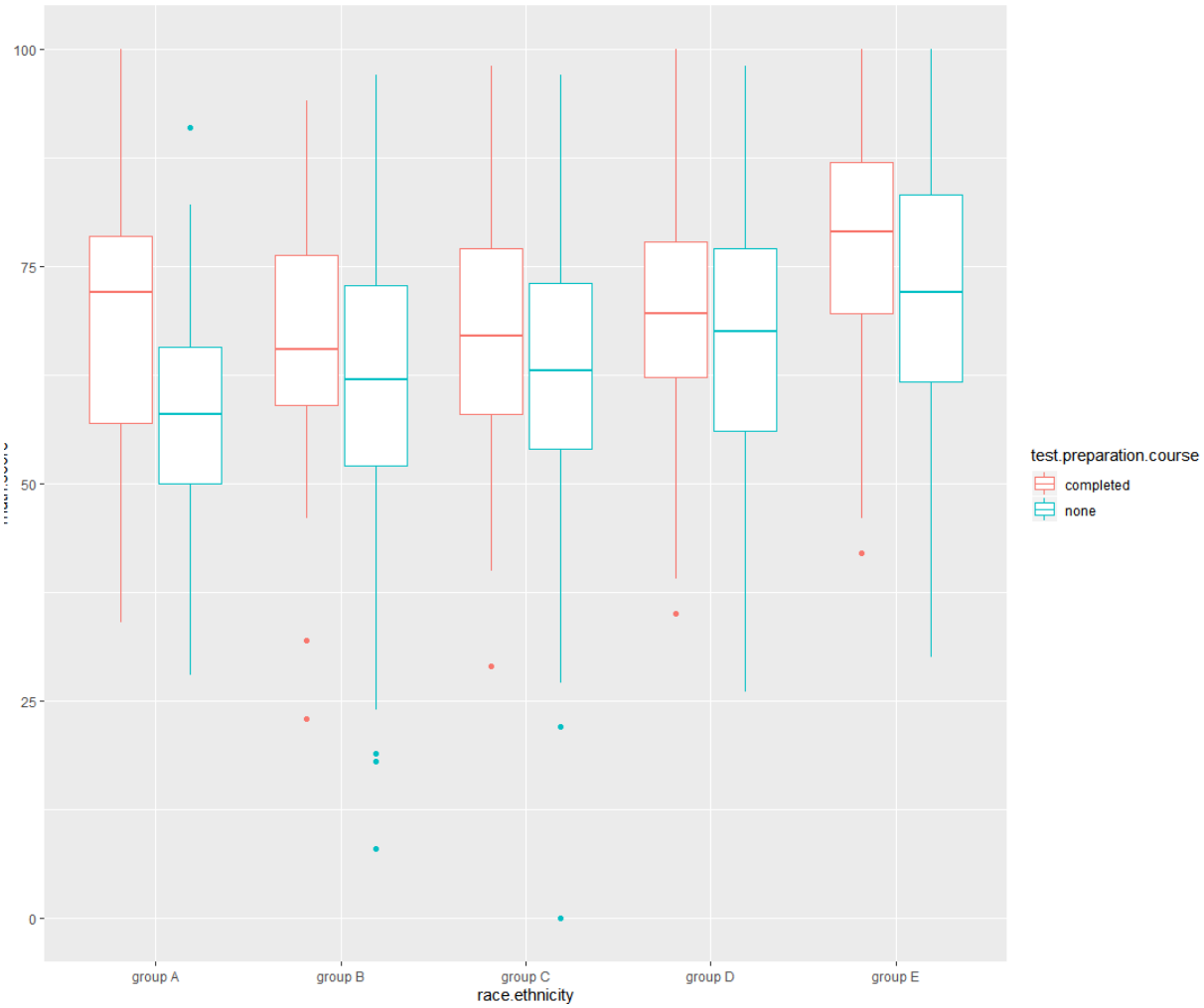

```
> ggplot(data, aes(parental.level.of.education, writing.score, color=test.preparation.course))+
geom_boxplot()
```



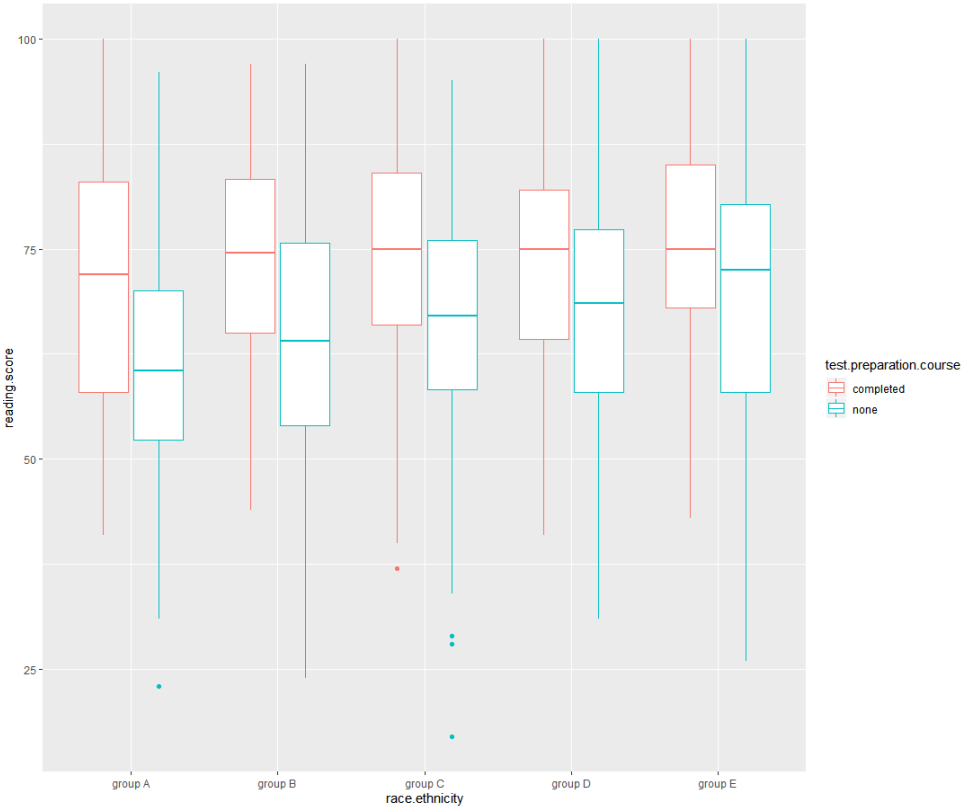
Ethnicity

As per ethnicity this course seems to be doing a good job. Especially the worst performing ethnicity group gets the support they need and are on par with the other groups. The advantage that group E seems to have cannot be made up however.

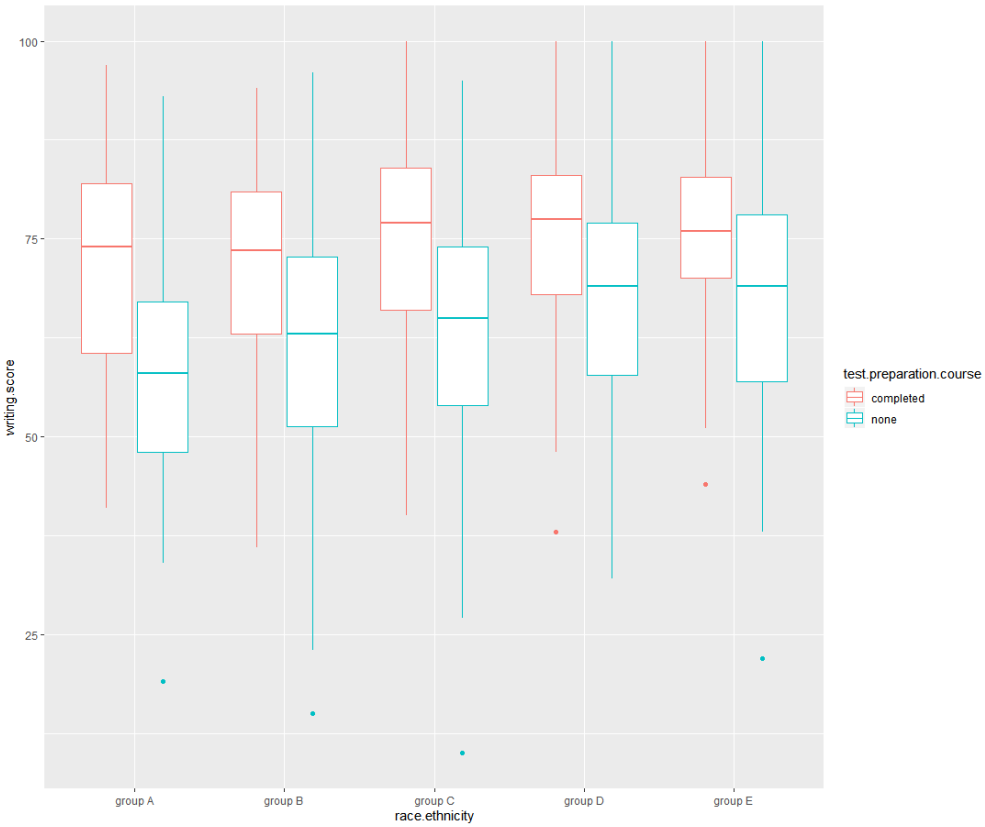
```
ggplot(data, aes(race.ethnicity, math.score, color=test.preparation.course))+geom_boxplot()
```



```
ggplot(data, aes(race.ethnicity, reading.score, color=test.preparation.course))+geom_boxplot()
```



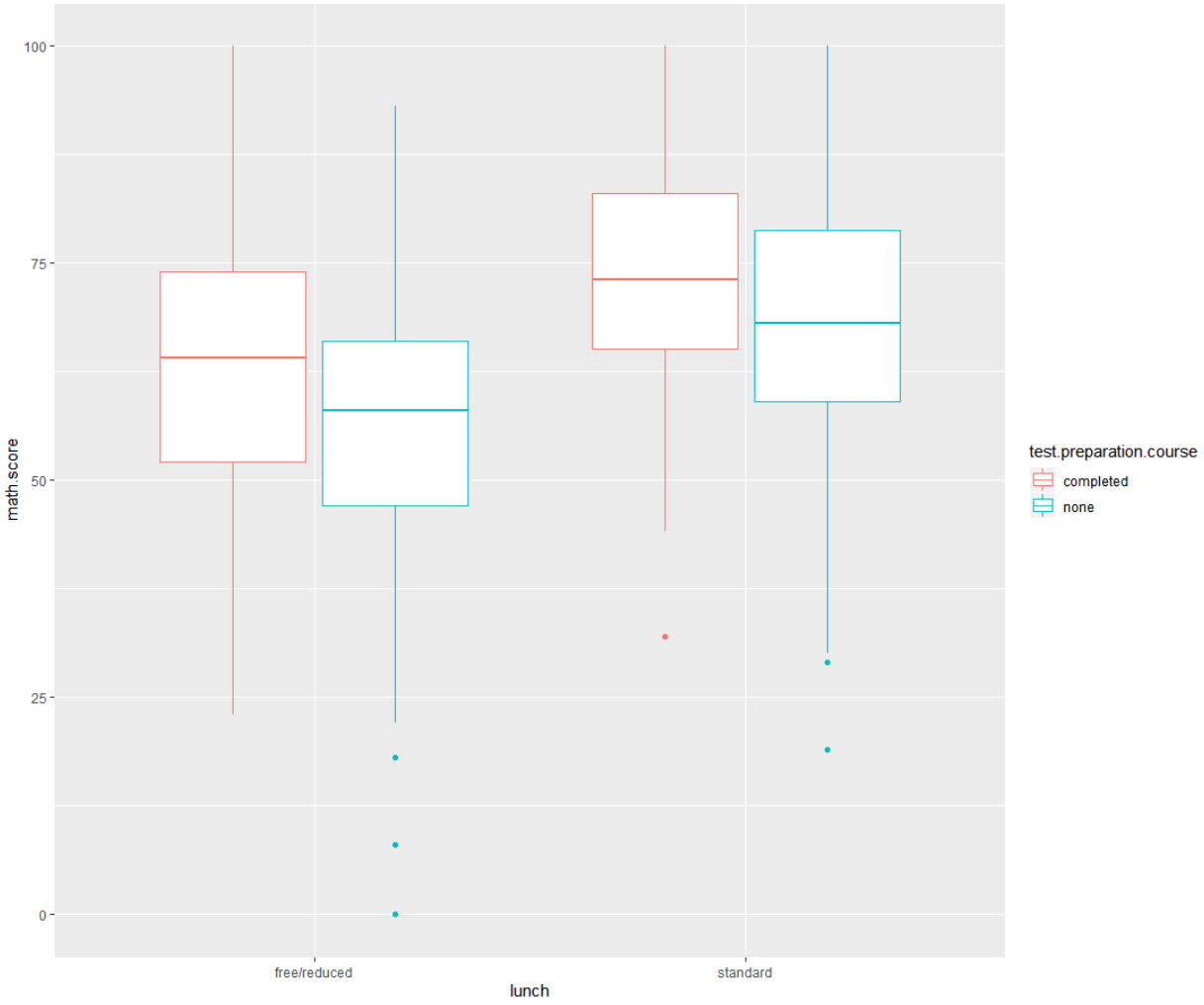

```
ggplot(data, aes(race.ethnicity, writing.score, color=test.preparation.course))+geom_boxplot()
```



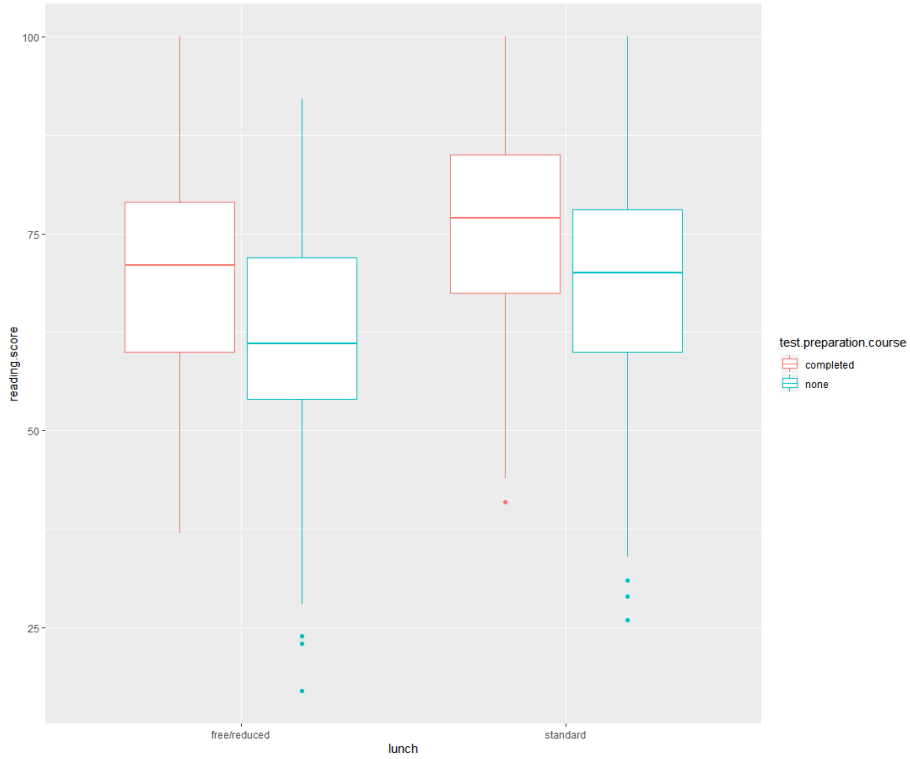
Lunch Plan

In terms of the lunch plan all students seem to get more or less the same boost. However, the students with the free/reduced plan still seem to be performing worse. This might be an indicator that these students come from low income/low parental educational background. This would also explain why they respond very well to the support they get from the school with the preparation course. But They still have a disadvantage compared to students from a more privileged background who took the course as well.

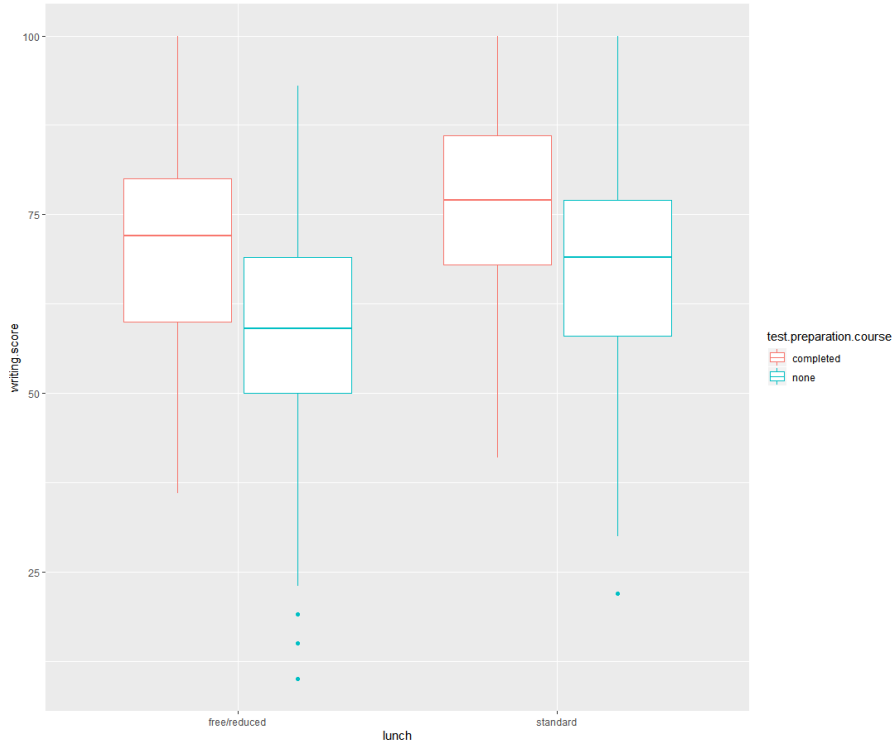
```
ggplot(data, aes(lunch, math.score, color=test.preparation.course))+geom_boxplot()
```



```
ggplot(data, aes(lunch, reading.score, color=test.preparation.course))+geom_boxplot()
```




```
ggplot(data, aes(lunch, writing.score, color=test.preparation.course))+geom_boxplot()
```



Conclusion

We can see that all test scores are strongly correlated, especially reading and writing. Thus, we can explain the data very well by only using the overall performance as a measure, like a GPA. The LDA shows that the data is fairly predictable. As an example, we could even predict the gender with a 90% accuracy rate by using the other values as input. We can also see that many factors have a huge impact on the student's performance. I wanted to find out if the course is successful at helping students achieve better grades and what aspects could be improved. From my analysis I conclude that the school should focus more on women in their math section and men in their reading/writing section. Also, students from a background of low level education of the parents or low-income families (indicated by lunch plan and parental educational background) need more support especially in math as well as ethnic groups other than group A.