Perfomance Analysis of students using R

Jeevan-1740256

20 December 2018

The performance of students during exams are influenced by many factors. Read the data set and try to get as many insights you can and hence suggest how the various factors influence the marks of different subjects.

library(readxl)

## Warning: package 'readxl' was built under R version 3.5.1

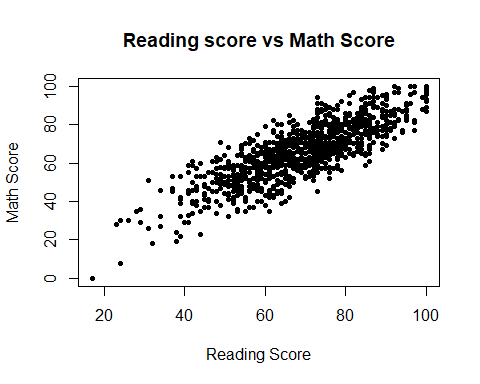
first<- read\_excel("C:/Users/Jeevan/Desktop/Christ University/R Studio/Stud\_Perf.xlsx")  
first

## # A tibble: 1,000 x 8  
## gender `race/ethnicity` `parental level~ lunch `test preparati~  
## <chr> <chr> <chr> <chr> <chr>   
## 1 female group B bachelor's degr~ stan~ none   
## 2 female group C some college stan~ completed   
## 3 female group B master's degree stan~ none   
## 4 male group A associate's deg~ free~ none   
## 5 male group C some college stan~ none   
## 6 female group B associate's deg~ stan~ none   
## 7 female group B some college stan~ completed   
## 8 male group B some college free~ none   
## 9 male group D high school free~ completed   
## 10 female group B high school free~ none   
## # ... with 990 more rows, and 3 more variables: `math score` <dbl>,  
## # `reading score` <dbl>, `writing score` <dbl>

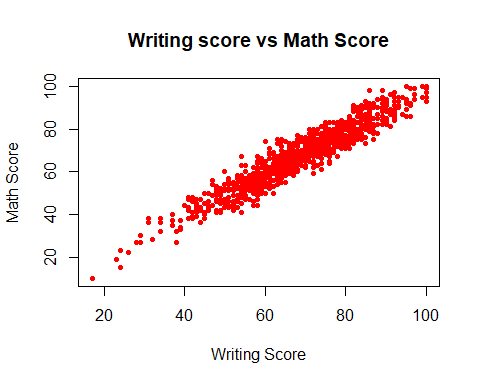
attach(first)  
head(first)

## # A tibble: 6 x 8  
## gender `race/ethnicity` `parental level~ lunch `test preparati~  
## <chr> <chr> <chr> <chr> <chr>   
## 1 female group B bachelor's degr~ stan~ none   
## 2 female group C some college stan~ completed   
## 3 female group B master's degree stan~ none   
## 4 male group A associate's deg~ free~ none   
## 5 male group C some college stan~ none   
## 6 female group B associate's deg~ stan~ none   
## # ... with 3 more variables: `math score` <dbl>, `reading score` <dbl>,  
## # `writing score` <dbl>

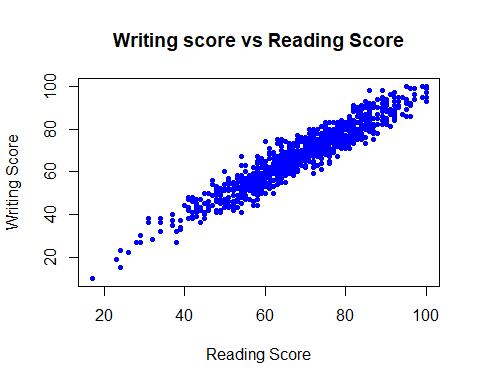
plot(`reading score`,`math score`,main = "Reading score vs Math Score",xlab = "Reading Score",ylab = "Math Score",pch=20)



plot(`reading score`,`writing score`,main = "Writing score vs Math Score",xlab = "Writing Score",ylab = "Math Score",pch=20,col="red")



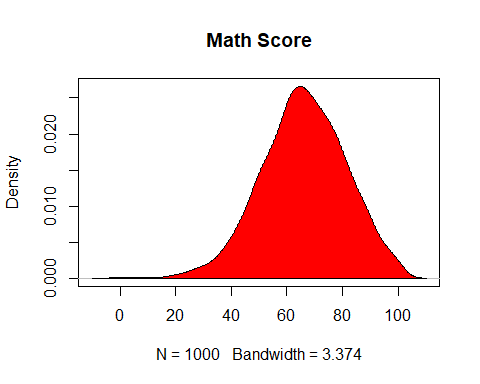
plot(`reading score`,`writing score`,main = "Writing score vs Reading Score",xlab = "Reading Score",ylab = "Writing Score",pch=20,col="blue")



d<-density(first$`math score`)  
d

##   
## Call:  
## density.default(x = first$`math score`)  
##   
## Data: first$`math score` (1000 obs.); Bandwidth 'bw' = 3.374  
##   
## x y   
## Min. :-10.12 Min. :1.331e-06   
## 1st Qu.: 19.94 1st Qu.:2.310e-04   
## Median : 50.00 Median :3.945e-03   
## Mean : 50.00 Mean :8.308e-03   
## 3rd Qu.: 80.06 3rd Qu.:1.595e-02   
## Max. :110.12 Max. :2.660e-02

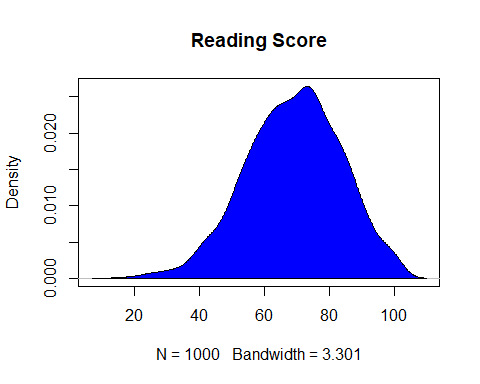
plot(d,main="Math Score")  
polygon(d,col="red",border="black")



m<-density(first$`reading score`)  
m

##   
## Call:  
## density.default(x = first$`reading score`)  
##   
## Data: first$`reading score` (1000 obs.); Bandwidth 'bw' = 3.301  
##   
## x y   
## Min. : 7.098 Min. :1.369e-06   
## 1st Qu.: 32.799 1st Qu.:9.030e-04   
## Median : 58.500 Median :6.085e-03   
## Mean : 58.500 Mean :9.718e-03   
## 3rd Qu.: 84.201 3rd Qu.:1.904e-02   
## Max. :109.902 Max. :2.645e-02

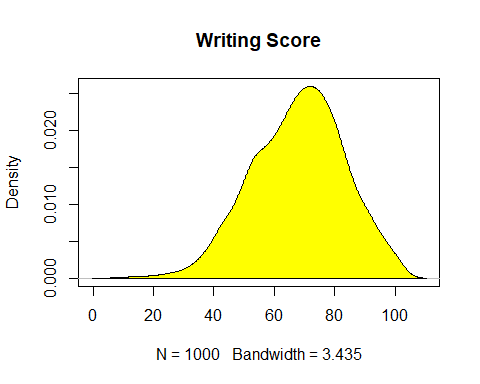
plot(m,main = "Reading Score")  
polygon(m,col="blue",border="black")



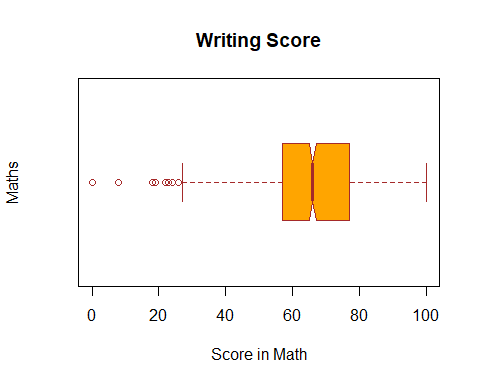
w<-density(first$`writing score`)  
w

##   
## Call:  
## density.default(x = first$`writing score`)  
##   
## Data: first$`writing score` (1000 obs.); Bandwidth 'bw' = 3.435  
##   
## x y   
## Min. : -0.3058 Min. :1.312e-06   
## 1st Qu.: 27.3471 1st Qu.:5.172e-04   
## Median : 55.0000 Median :5.799e-03   
## Mean : 55.0000 Mean :9.032e-03   
## 3rd Qu.: 82.6529 3rd Qu.:1.731e-02   
## Max. :110.3058 Max. :2.602e-02

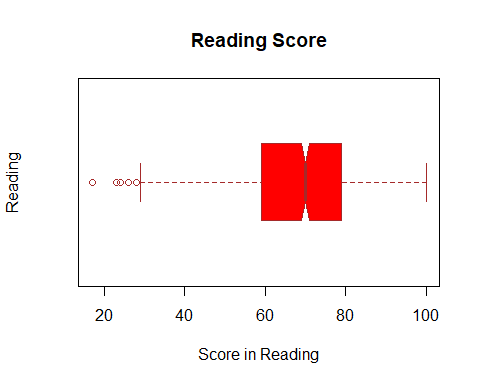
plot(w,main = "Writing Score")  
polygon(w,col="yellow",border="black")



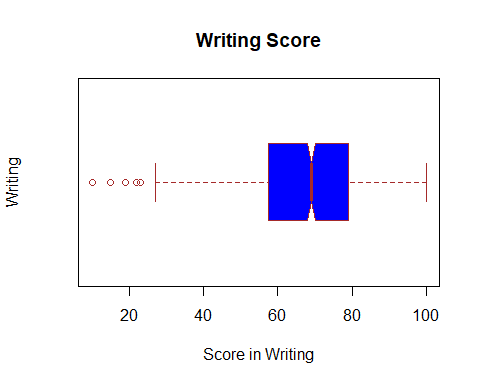
boxplot(first$`math score`,main = "Writing Score",xlab="Score in Math",ylab="Maths",col="orange",border="brown",horizontal = TRUE,notch = TRUE)



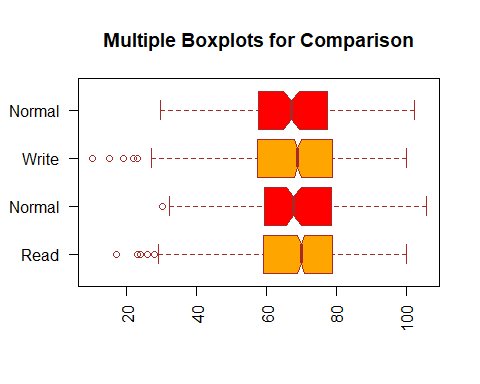
boxplot(first$`reading score`,main = "Reading Score",xlab="Score in Reading",ylab="Reading",col="red",border="brown",horizontal = TRUE,notch = TRUE)



boxplot(first$`writing score`,main = "Writing Score",xlab="Score in Writing",ylab="Writing",col="blue",border="brown",horizontal = TRUE,notch = TRUE)



read<-first$`reading score`  
write<-first$`writing score`  
read\_norm<-rnorm(200,mean=mean(read),sd=sd(read))  
write\_norm<-rnorm(200,mean=mean(write),sd=sd(write))  
boxplot(read,read\_norm,write,write\_norm,main="Multiple Boxplots for Comparison",names = c("Read","Normal","Write","Normal"),las=2,col = c("orange","red"),border = "brown",horizontal = TRUE,notch = TRUE)



summary(first)

## gender race/ethnicity parental level of education  
## Length:1000 Length:1000 Length:1000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## lunch test preparation course math score   
## Length:1000 Length:1000 Min. : 0.00   
## Class :character Class :character 1st Qu.: 57.00   
## Mode :character Mode :character 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

plot(`reading score`,`writing score`,main = "Reading score vs Writing Score",xlab = "Reading Score",ylab = "Writing Score",pch=20)  
cor(`reading score`,`writing score`,method = "pearson")

## [1] 0.9545981

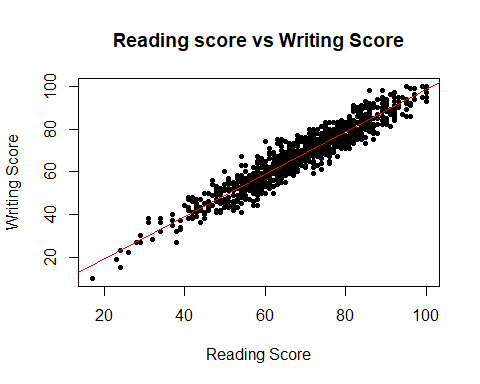
model<-lm(`writing score`~`reading score`)  
model

##   
## Call:  
## lm(formula = `writing score` ~ `reading score`)  
##   
## Coefficients:  
## (Intercept) `reading score`   
## -0.6676 0.9935

summary.lm(model)

##   
## Call:  
## lm(formula = `writing score` ~ `reading score`)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.9573 -2.9573 0.0363 3.1026 15.0557   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.667554 0.693792 -0.962 0.336   
## `reading score` 0.993531 0.009814 101.233 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.529 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

abline(model,col='red')



t.test(`reading score`,`writing score`)

##   
## Welch Two Sample t-test  
##   
## data: reading score and writing score  
## t = 1.6732, df = 1994.8, p-value = 0.09445  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.1918904 2.4218904  
## sample estimates:  
## mean of x mean of y   
## 69.169 68.054

t.test(`reading score`,`math score`)

##   
## Welch Two Sample t-test  
##   
## data: reading score and math score  
## t = 4.6271, df = 1995.1, p-value = 3.947e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.774566 4.385434  
## sample estimates:  
## mean of x mean of y   
## 69.169 66.089

t.test(`writing score`,`math score`)

##   
## Welch Two Sample t-test  
##   
## data: writing score and math score  
## t = 2.8946, df = 1998, p-value = 0.003837  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.6336869 3.2963131  
## sample estimates:  
## mean of x mean of y   
## 68.054 66.089

plot(`reading score`,`math score`,main = "Reading score vs Math Score",xlab = "Reading Score",ylab = "Math Score",pch=20)  
cor(`reading score`,`math score`,method = "pearson")

## [1] 0.8175797

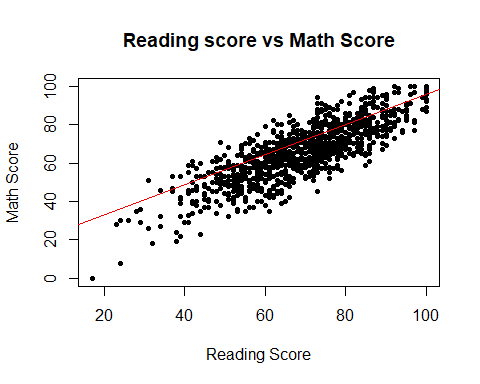
model1<-lm(`reading score`~`math score`)  
model1

##   
## Call:  
## lm(formula = `reading score` ~ `math score`)  
##   
## Coefficients:  
## (Intercept) `math score`   
## 17.1418 0.7872

summary.lm(model1)

##   
## 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

abline(model1,col='red')



var.test(`reading score`,`math score`)

##   
## F test to compare two variances  
##   
## data: reading score and math score  
## F = 0.92713, num df = 999, denom df = 999, p-value = 0.232  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.8189358 1.0496262  
## sample estimates:  
## ratio of variances   
## 0.9271335

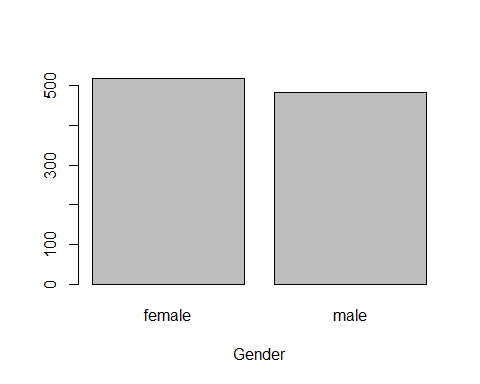
var.test(`reading score`,`writing score`)

##   
## F test to compare two variances  
##   
## data: reading score and writing score  
## F = 0.92316, num df = 999, denom df = 999, p-value = 0.2066  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.8154283 1.0451306  
## sample estimates:  
## ratio of variances   
## 0.9231625

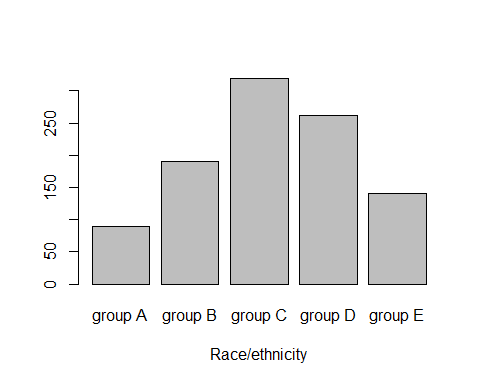
var.test(`writing score`,`math score`)

##   
## F test to compare two variances  
##   
## data: writing score and math score  
## F = 1.0043, num df = 999, denom df = 999, p-value = 0.9459  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.8870982 1.1369896  
## sample estimates:  
## ratio of variances   
## 1.004301

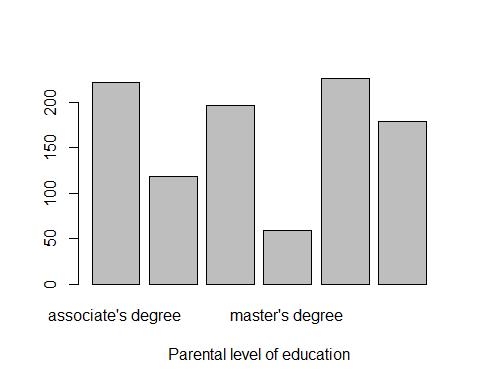
counts<-table(first$gender)  
barplot(counts,xlab="Gender")



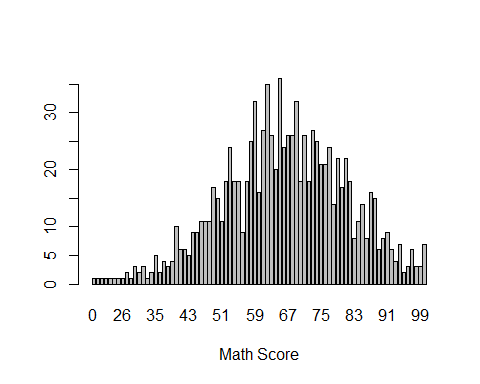
counts1<-table(first$`race/ethnicity`)  
barplot(counts1,xlab="Race/ethnicity")



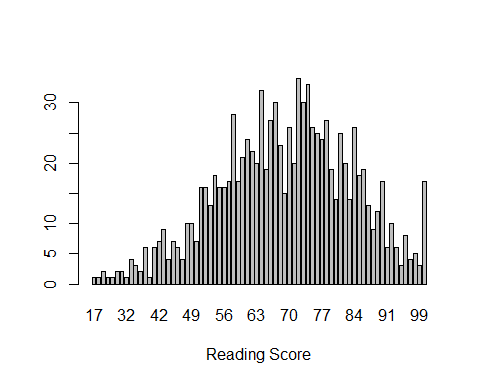
counts2<-table(first$`parental level of education`)  
barplot(counts2,xlab = "Parental level of education")



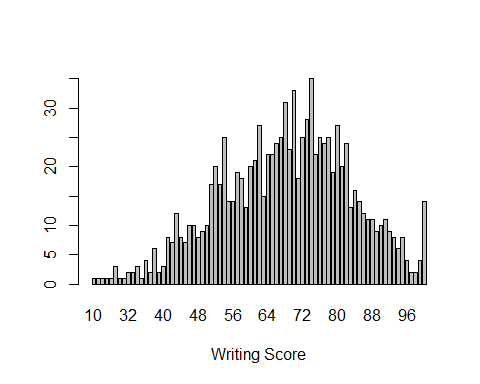
counts3<-table(first$`math score`)  
barplot(counts3,xlab = "Math Score")



counts4<-table(first$`reading score`)  
barplot(counts4,xlab = "Reading Score")



counts5<-table(first$`writing score`)  
barplot(counts5,xlab="Writing Score")

 The excel sheet on the Student perfomances was imported into RStudio. Multiple graphs such as scatter plots were plotted among the reading, writing & Math scores. A kernel density plot was plotted with each of the students scoes such as their reading scores, math scores and their writing scores. The density result shows the minimum value, 1st Quartile deviation, meadian, mean, 3rd quartile deviation and maximum value. Different boxplots have been created with the different scores such as the students perfomance of their reading score, writing score and math score. Another boxplot has been created to compare the scores together. The summary has been shown with the length of each column, the datatype, and regarding the scores - it shows the minimum value, 1st Quartile deviation, meadian, mean, 3rd quartile deviation and maximum value. Using Pearson’s method of correlation, the correlation between the reading score and the writing score is 0.9545981. A linear model has been fitted to the graph and the multiple R Squared value is 0.9113 which means that the linear model is a good fit to the data. t-tests such as Welch Two Sample t-test have been used to test the difference in means of the reading score and writing score. At 95% confidence interval - the Alternative hypothesis states that the true difference in means is not equal to 0. Since the p-value of the t-test between the reading score and math score is greater than 0.05 which is the alpha value - we accept the null hypothesis and reject the alternative hypothesis which means that there is no significant difference between the means of the reading score and the math score. Another Welch Two Sample t-test has been implemented between the writing score and the math score. The p-value which is 0.003837 is less than the alpha value which is 0.05 at 95% confidence interval telling us that there is a significant difference between the mean of the writing score and the mean of the math score which implies that we reject the mull hypothesis and accept the alternative hypothesis. The F-test has been used to compare the 2 variances between reading score and math score. The p - value which is 0.232 is greater than the alpha value which is 0.05 at 95% confidence level which means that we accept the null hypothesis and reject the alternative hypothesis which implies that the true ratio of variances is equal to 1. The F-test has been used to compare the 2 variances between writing score and math score. The p - value which is 0.9459 is greater than the alpha value which is 0.05 at 95% confidence level which means that we accept the null hypothesis and reject the alternative hypothesis which implies that the true ratio of variances is equal to 1. The F-test has been used to compare the 2 variances between reading score and writing score. The p - value which is 0.2066 is greater than the alpha value which is 0.05 at 95% confidence level which means that we accept the null hypothesis and reject the alternative hypothesis which implies that the true ratio of variances is equal to 1. A few barplots have been created on the perfomance of the students.