Project1

Yahui Li

2020/1/13

Step1 Read Data

```r
#install.packages("AER")
library(AER)
data("STAR")
```

Step2 Explore Data

We will only examine the math scores in 1st grade in this project.

```r
data <- data.frame(star1 = STAR$star1, math1 = STAR$math1)

sapply(data,class)
```

```
##     star1     math1
##  "factor" "integer"
```

```r
sapply(data,summary)
```

```
## $star1
##      regular        small regular+aide         NA's
##         2584         1925         2320         4769
##
## $math1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   404.0   500.0   529.0   530.5   557.0   676.0    4998
```

```r
data.star1.na <- data[is.na(data$star1),]
all(is.na(data.star1.na$math1))
```

```
## [1] TRUE
```

Which shows that the math score has not been recorded if class type is not recorded. So we can remove the data where star1 is NA.

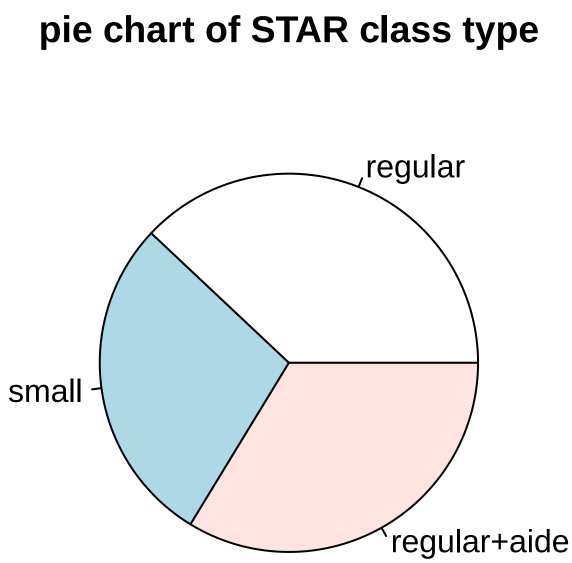
One of the way to deal with NA in math1 is to remove them

```r
data_remove_na <- na.omit(data[-is.na(data$star1),])
```

```r
table(data_remove_na$star1)
```

```
##
##      regular        small regular+aide
##         2507         1868         2225
```

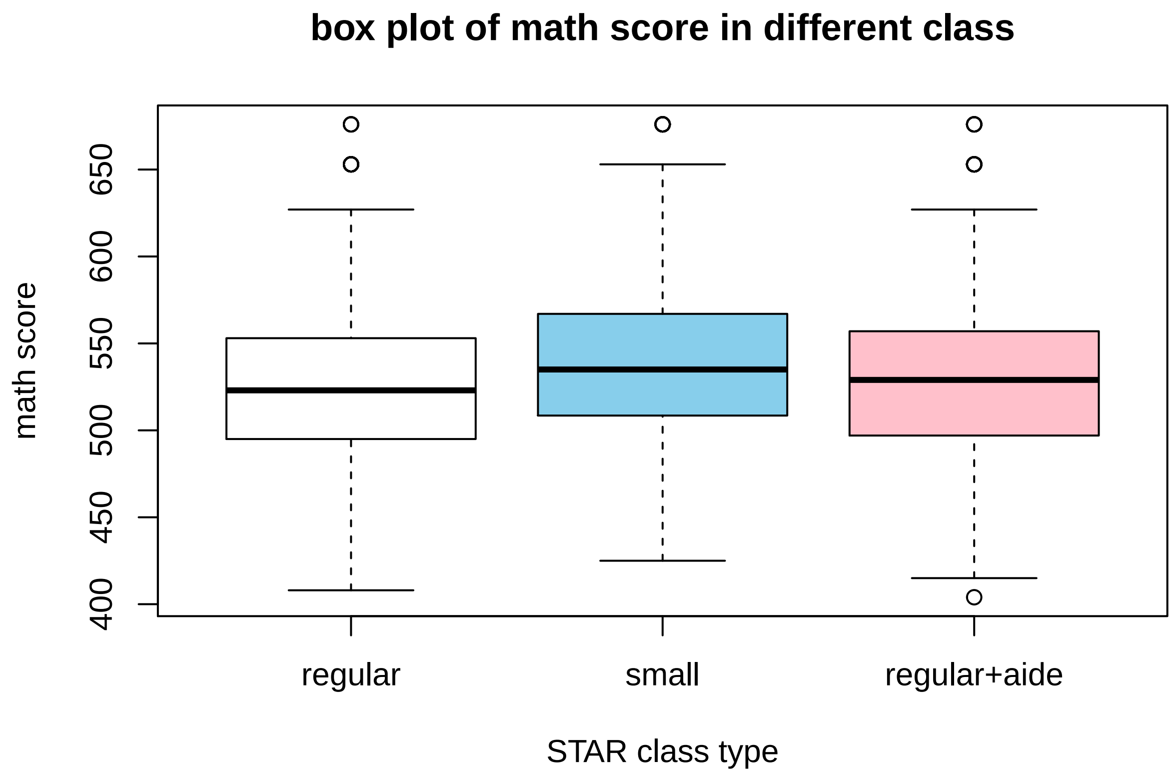
```r
pie(table(data_remove_na$star1),main = "pie chart of STAR class type")
```




```r
tapply(data_remove_na$math1, data_remove_na$star1,summary)
```

```
## $regular
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   408.0   495.0   523.0   525.3   553.0   676.0
##
## $small
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   425.0   509.2   535.0   538.7   567.0   676.0
##
## $`regular+aide`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   404.0   497.0   529.0   529.6   557.0   676.0
```

```r
boxplot(data$math1~data$star1,main = "box plot of math score in different class",
        xlab = "STAR class type", ylab = "math score", col = c("white", "skyblue", "pink"))
```



From the result,

for mean, small > regular+aide > regular;

for all quantile information, small > regular+aide > regular;

for min, small > other two; For max, they are the same.

Something interesting: there are only some certain scores like 601 612 627 653 676.

Step3 One Way ANOVA Model

$$Y_{ij} = \mu_1 + \tau_2 X_{2,ij} + \tau_3 X_{3,ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2), \quad i = 1, 2, 3, j = 1, \cdots, n_i.$$

where $i = 1$ means the class type in 1st grade is regular; $i = 2$ means the class type in 1st grade is small; $i = 3$ means the class type in 1st grade is regular-with-aide.

From the table in step2, $n_1 = 2507, n_2 = 1868, n_3 = 2225, n = 6600$.

$X_{2,ij} = 1$ if $i = 2$, otherwise $X_{2,ij} = 0$. $X_{3,ij} = 1$ if $i = 3$, otherwise $X_{3,ij} = 0$.

Y_{ij} denotes the math grade in 1st grade of the j-th experimental unit in the i-th class type.

μ_i means the population mean of the i-th type class in 1st grade, $i = 1, 2, 3$.

$\tau_i = \mu_i - \mu_1$ means the difference in population mean between i-th type and first type in 1st grade, $i = 2, 3$.

ϵ_{ij} is independent and identically distributed normal random variable with 0 mean and σ^2 variance under normal assumption.

Model Assumption
(a) Response variable residuals are normally distributed.
(b) Variances of populations are equal.
(c) Responses for a given group are independent and identically distributed normal random variables.
All of the assumptions are necessary, because F-test and related procedures are pretty robust to the normality and equal variance assumptions, and pairwise comparisons could be substantially affected by unequal variances. Moreover, non-independence can have serious side effects and is hard to correct. So it is important to apply randomization whenever necessary.

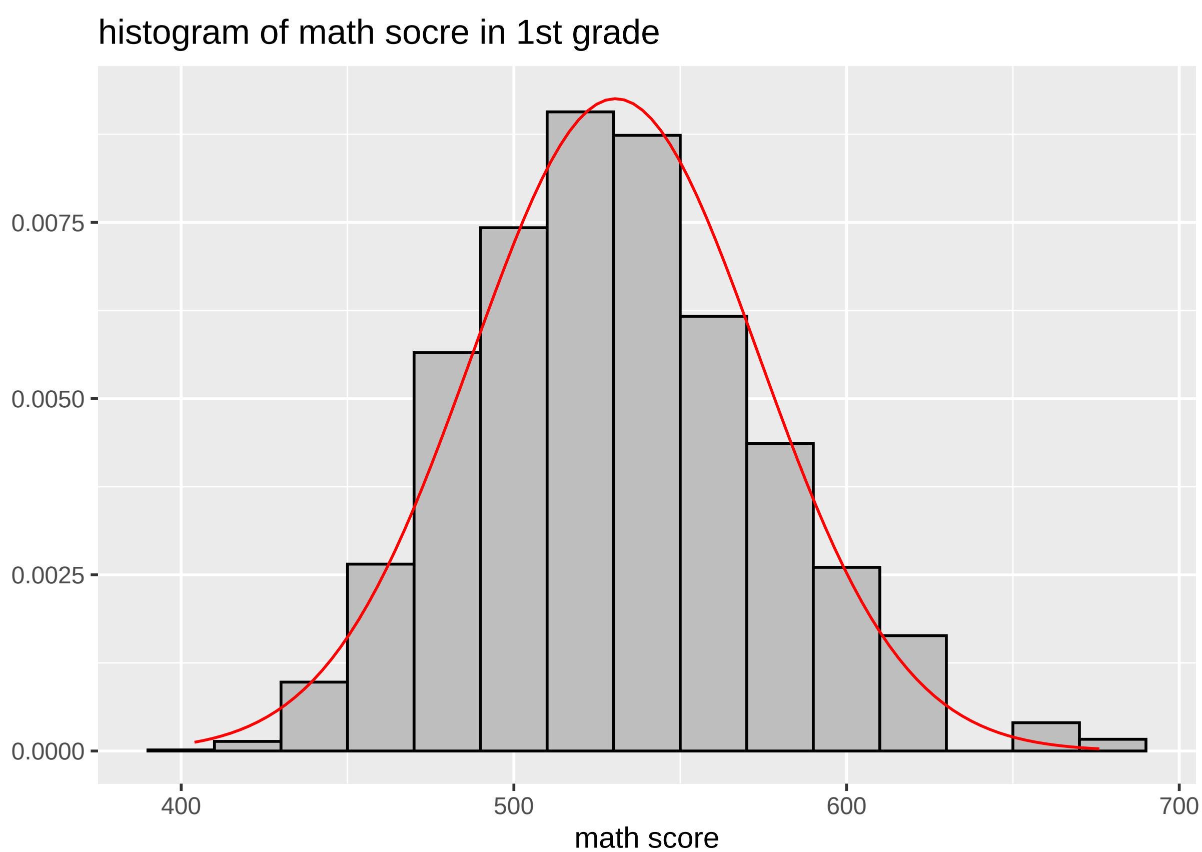
Step4 Appropriate

Before we fit the model, we need to ensure that model is appropriate on this dataset, that is, the response variable satisfies the assumptions of our model. In other words, we will check the normality and equal variance of the response varibales.

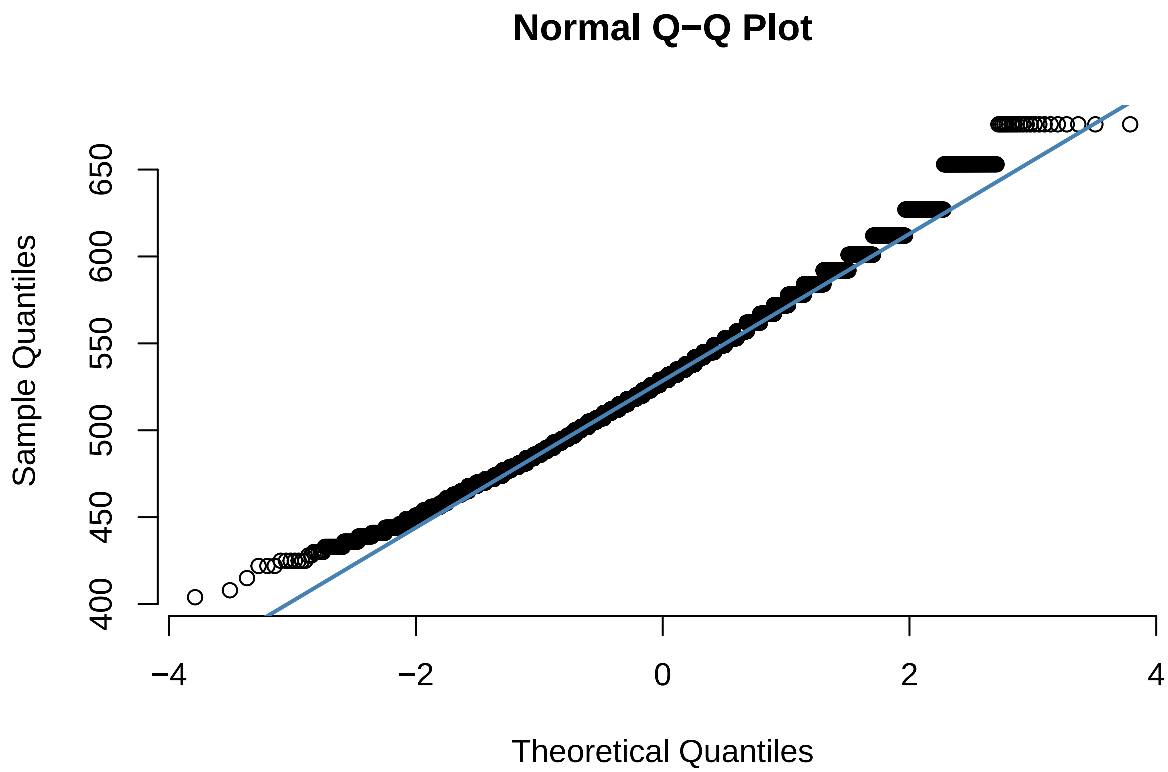
We first make a density plot and a Q-Q plot to check the normality of `math1`.

```
library(ggplot2)
x <- seq(404, 676, length.out=100)
df <- with(data_remove_na, data.frame(x = x, y = dnorm(x, mean(math1), sd(math1))))

ggplot(data_remove_na, aes(x=math1, y = ..density..)) +
  geom_histogram(binwidth = 20, fill = "grey", color = "black") +
  geom_line(data = df, aes(x = x, y = y), color = "red") +
  labs(x="math score",y="",title = "histogram of math socre in 1st grade")
```



```r
qqnorm(data_remove_na$math1, pch = 1, frame = FALSE)
qqline(data_remove_na$math1, col = "steelblue", lwd = 2)
```

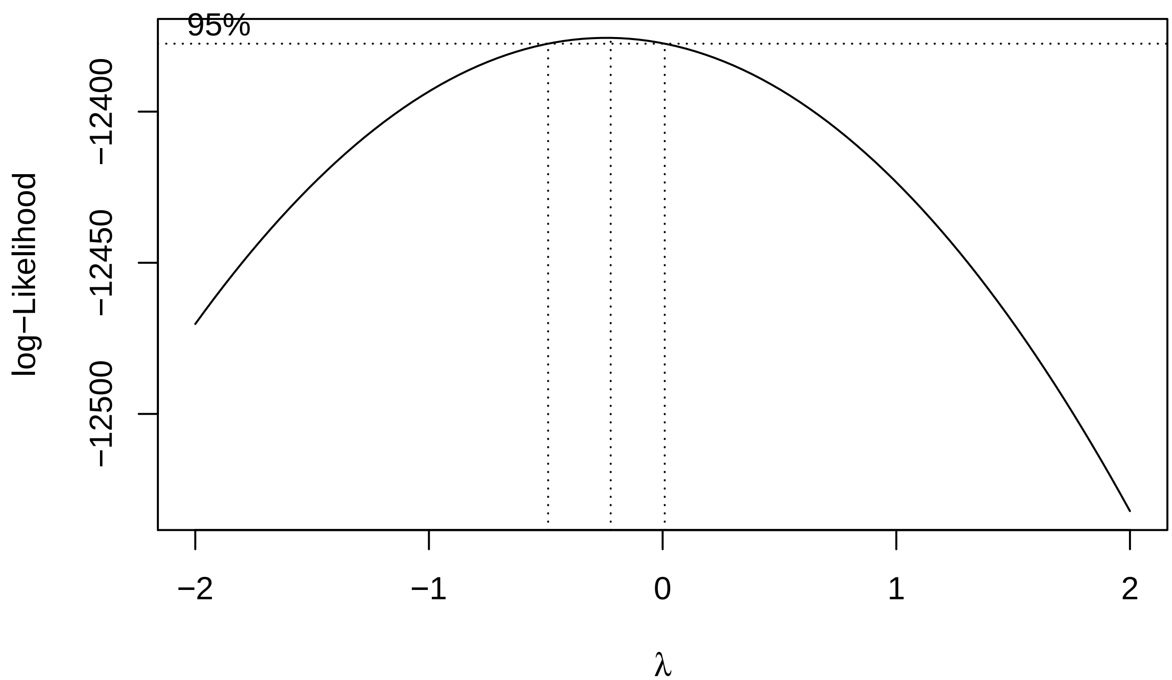


The histogram shows that it seems normal distribution.

The Q-Q plot shows the the distribution of math score is right-skewed.

So we use Box-Cox method to make a transformation on `math1`.

```r
library(MASS)
boxcox(math1 ~ star1 , data = data_remove_na)
```



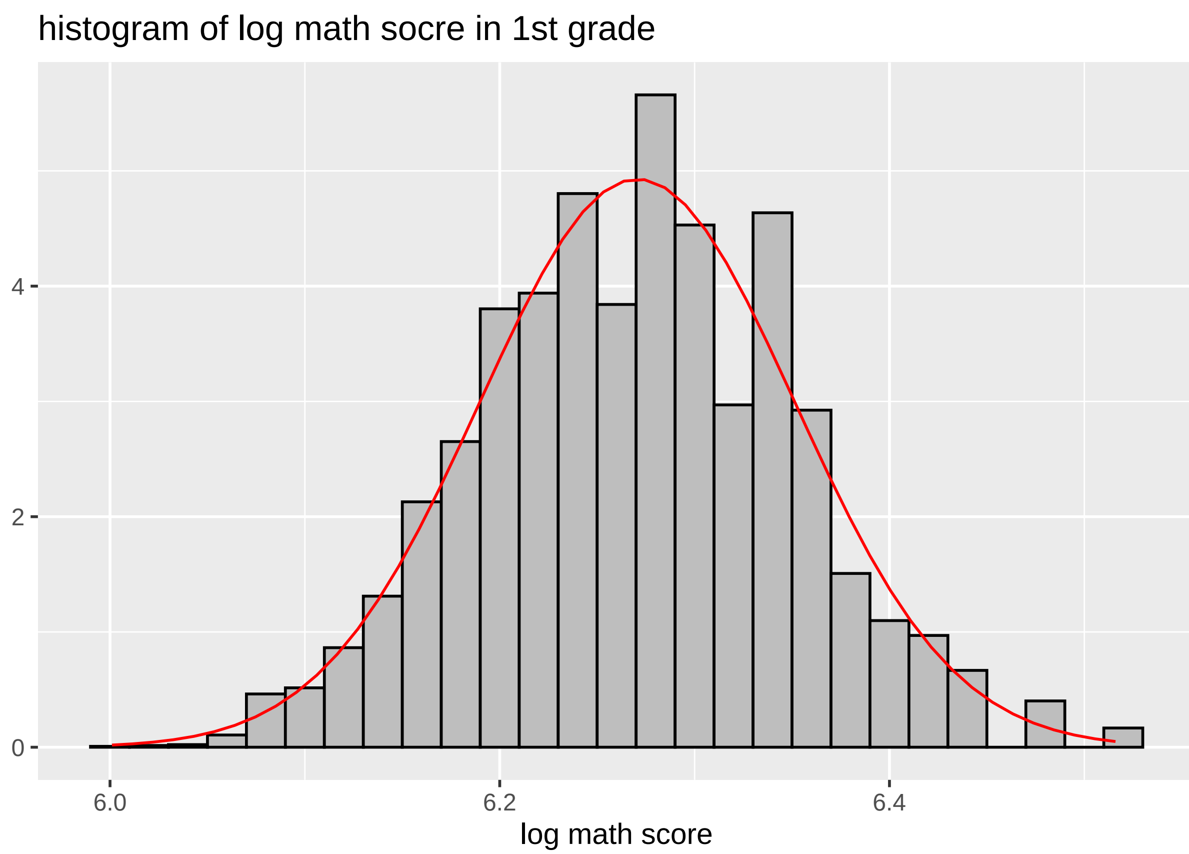
It indicates that we need make a log-transformation for `math1`.

```
summary(log(data_remove_na$math1))
```

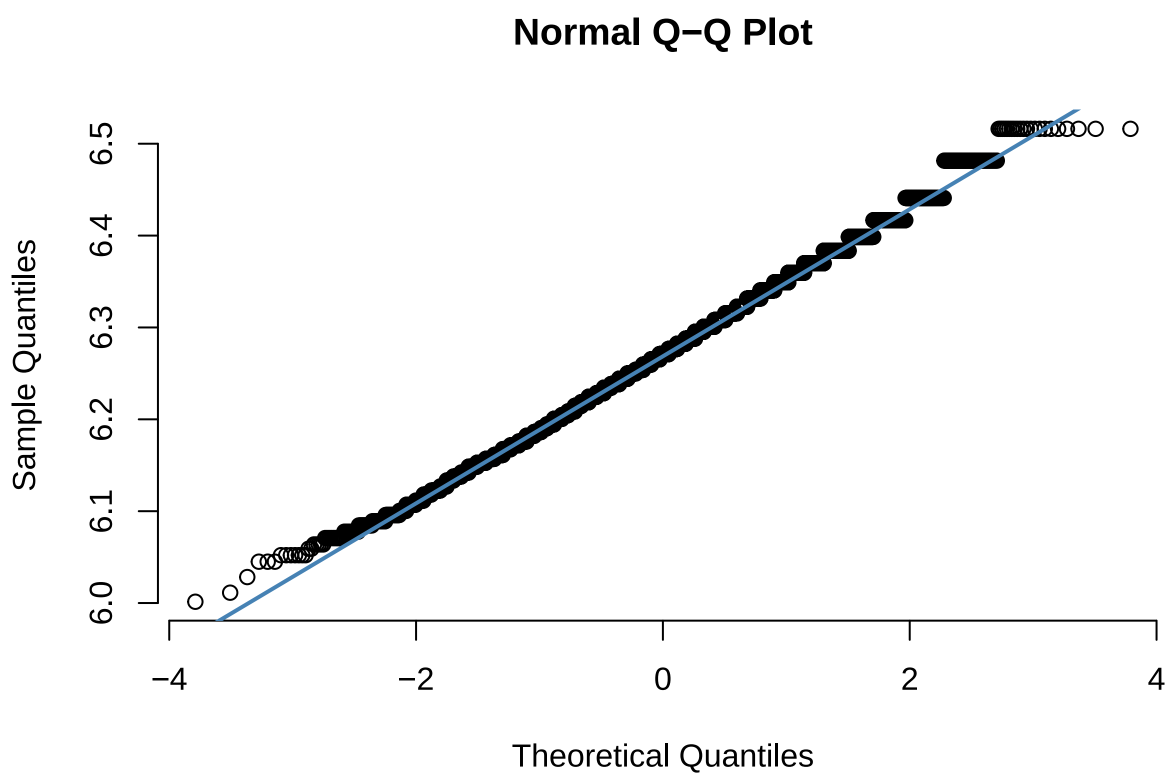
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.001   6.215   6.271   6.271   6.323   6.516
```

```
x <- seq(6.001, 6.516, length.out=50)
df <- with(data_remove_na, data.frame(x = x, y = dnorm(x, mean(log(math1)), sd(log(math1)))))

ggplot(data_remove_na, aes(x=log(math1), y = ..density..)) +
  geom_histogram(binwidth = 0.02, fill = "grey", color = "black") +
  geom_line(data = df, aes(x = x, y = y), color = "red") +
  labs(x="log math score",y="",title = "histogram of log math socre in 1st grade")
```




```r
qqnorm(log(data_remove_na$math1), pch = 1, frame = FALSE)
qqline(log(data_remove_na$math1), col = "steelblue", lwd = 2)
```



The graph shows the distribution of log math score in 1st grade is Normal-like.

Then we calculate the variance of math grade in 1st grade of each class type.

```
library(tidyverse)
```

```
data_remove_na %>%
  group_by(star1) %>%
  summarize(var_math1 = var(log(math1), na.rm = T))
```

```
## # A tibble: 3 x 2
##   star1         var_math1
##   <fct>             <dbl>
## 1 regular         0.00625
## 2 small           0.00666
## 3 regular+aide    0.00650
```

The result shows that they are very small and nearly equal to each other. Therefore, it is appropiate to build our model on this dataset.

Step5 Fit Model

```
anova.fit<- aov(log(math1)~star1,data=data_remove_na)
summary(anova.fit)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## star1           2   0.68  0.3391   52.56 <2e-16 ***
```

```
## Residuals   6597  42.55  0.0065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova.fit$coefficients
```

```
##      (Intercept)        star1small star1regular+aide
##       6.26079457        0.02499081        0.00812069
```

From the result, the fitted model we get is:

$$\log \hat{Y}_{ij} = 6.2608 + 0.0250X_{2,ij} + 0.0081X_{3,ij}$$

with means when the type is regular, the estimate math score is $e^{6.2608} = 523.6377$; when the type is small, the estimate math score is $e^{6.2608+0.0250} = 536.8936$; when the type is regular-with-aide, the estimate math score is $e^{6.2608+0.0081} = 527.8964$.

The following is a ANOVA table for this model.

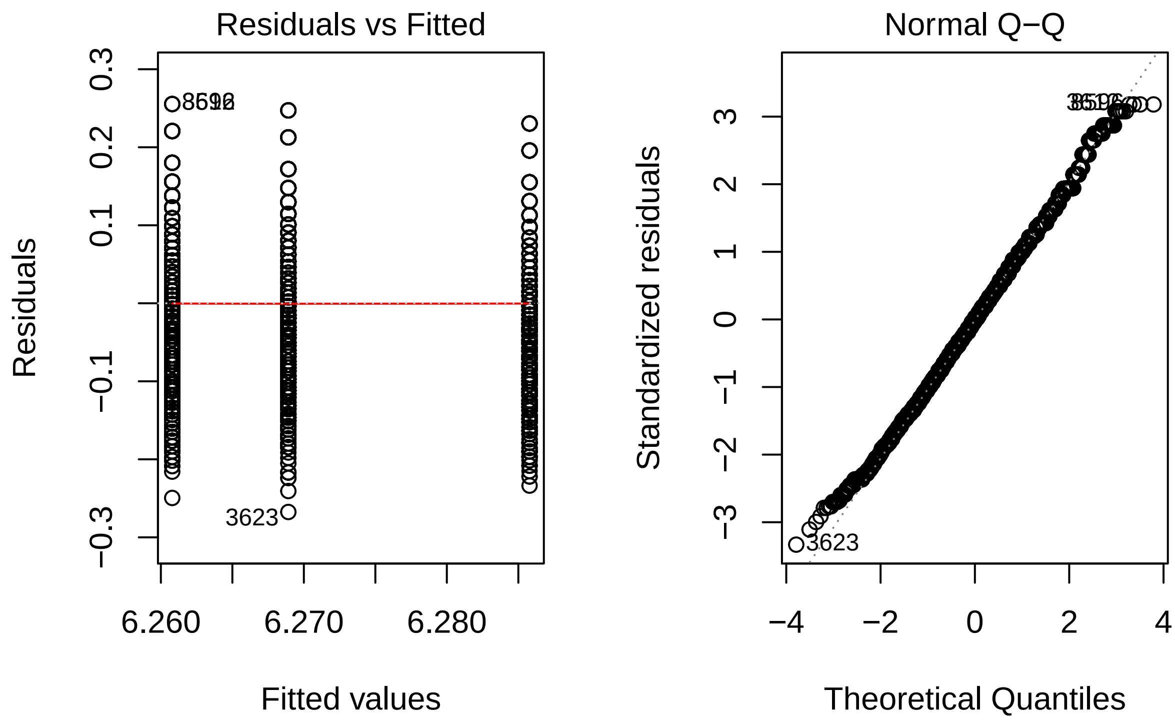
Source of Variation	Sum of Squares	Degrees of Freedom	MS
`Between treatments`	SSTR = 0.68	2	MSTR = 0.3391
`Within treatments`	SSE = 42.55	6597	MSE = 0.0065
`Total`	SSTO = 43.23	6599	

###step6 model diagnostic and sensitivity analysis

Recalling the above assumptions, there are three things we need to check: normality, equal variance and independence.

By Q-Q plot, we can check normality. And By residuals vs fitted value plot, we can check equal variance.

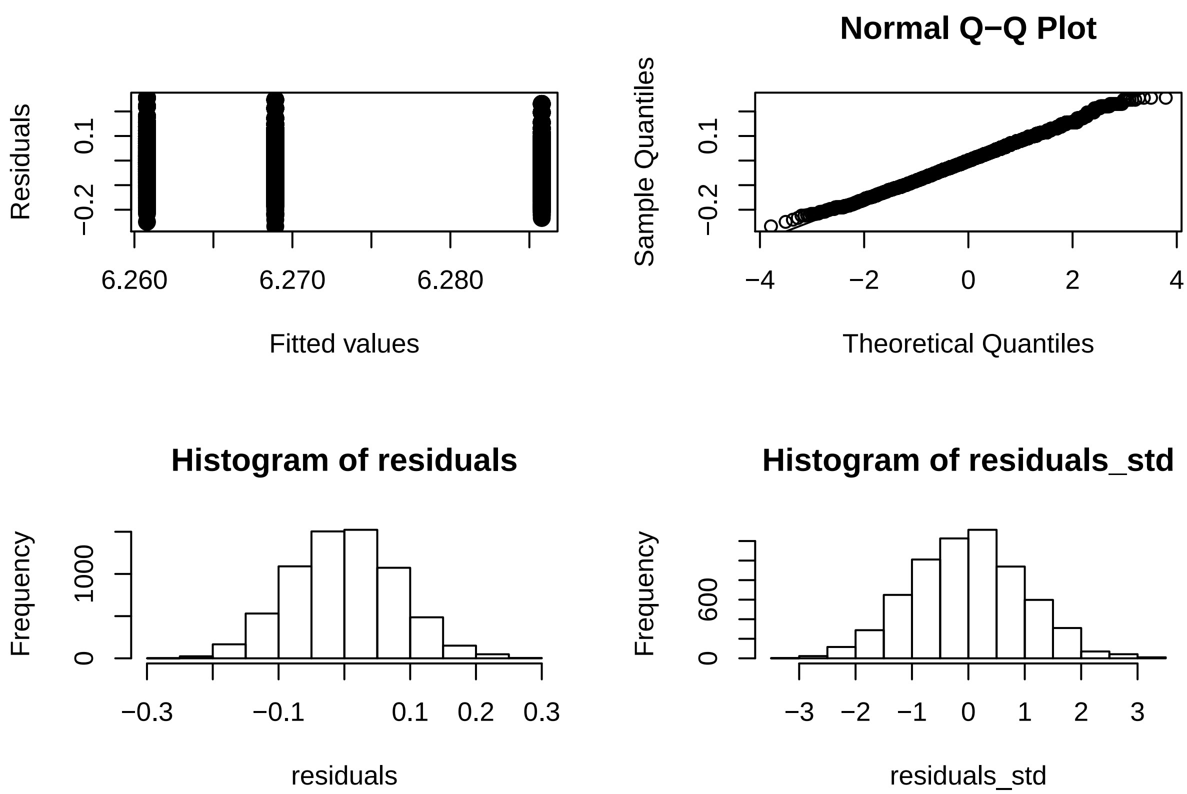
```
par(mfrow=c(1,2))
plot(anova.fit, which = c(1,2))
```



The first figure indicates equal variacne, and the second nearly linear figure supports normality.

shijianversion

```r
par(mfrow=c(2,2))
residuals <- anova.fit$residuals
##Plot the residuals (or the other two versions) against fitted values
plot(anova.fit$fitted.values, anova.fit$residuals,
     type = "p",pch=16,cex=1.5,xlab="Fitted values",ylab="Residuals")
#QQplot
qqnorm(residuals);qqline(residuals)
#residuals
hist(residuals)
#studentized residuals
residuals_std <- rstudent(anova.fit)
hist(residuals_std)
```



From the scatterplot of residuals vs fitted valuues, the residuals are divided into three groups and among each group, these residuals are around the zero, whcih menas that the average residuals equals to zero. According to the histogram of residuals and studentized residuals, we can find that the distribution of the residuals of the fitted model approximates to the normal distribution. Besides, the same conclusion can be obtained by checking the Q-Q Plot of the residuals. Therefore, we can confirm that the residuals of the model are normally distributed.

We now turn to formal tests of the equality of variances. First, we calculate the varianves for each type of class and find that the variances of three types of class are close to each other.

```
# Calculate the variances for each group:
(vars = tapply(data_remove_na$math1,data_remove_na$star1,var))
```

```
##      regular        small regular+aide
##     1734.825     1945.082     1837.493
```

Then, because the sample sizes of the three types of class are not same, we choose two formal tests, which are Bartlett test and Levene test, to check the equality of model variances.

```
data_remove_na$residuals <- residuals
#bartlett test
bartlett.test(residuals ~ star1, data = data_remove_na)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residuals by star1
## Bartlett's K-squared = 2.3004, df = 2, p-value = 0.3166
```

```r
#levene test
leveneTest(residuals ~ star1, data = data_remove_na)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    2  1.1836 0.3062
##       6597
```

From the two tests, both of the P-values are much larger than 0.05, which means that we can not reject the null hyperthesis: the variances of the model are equal.

In conclusion, we confirm that our model satisifies the normality assumption.

Sensitivity Analysis In order to test the sensitivity of our model, we decide to relax the assumption of our model. To be specific, we want to figure out that whether the influence of class size still exists even if the data is not normally distributed. Thus, we conduct the nonparameteric tests as follows, which are The rank test and Krusal-Wallis test.

```r
#rank test
data_remove_na$rank <- rank(data_remove_na$math1)
summary(aov(rank ~ star1, data = data_remove_na))
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## star1          2 3.556e+08 177779655   49.72 <2e-16 ***
## Residuals   6597 2.359e+10   3575611
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#kruskal test
kruskal.test(math1 ~ star1, data = data_remove_na)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  math1 by star1
## Kruskal-Wallis chi-squared = 97.993, df = 2, p-value < 2.2e-16
```

The results both show that the math scores of the different types of class are different at 99% confident level. So, even if the data was not normally distributed, there would still be influence of class size. In a word, our one-way anova model is reasonable in this case.

Step6 hypothese test

```r
TukeyHSD(anova.fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log(math1) ~ star1, data = data_remove_na)
##
## $star1
##                             diff          lwr         upr     p adj
## small-regular         0.02499081  0.019236297  0.03074533 0.0000000
## regular+aide-regular  0.00812069  0.002637091  0.01360429 0.0015104
## regular+aide-small   -0.01687012 -0.022778294 -0.01096196 0.0000000
```

```r
pairwise.t.test(log(data_remove_na$math1),data_remove_na$star1,p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  log(data_remove_na$math1) and data_remove_na$star1
##
##              regular small
## small        < 2e-16 -
## regular+aide 0.0016  7.1e-11
##
## P value adjustment method: bonferroni
```

```r
library(agricolae)
scheffe.test(anova.fit,"star1", group=TRUE,console=TRUE)
```

```
##
## Study: anova.fit ~ "star1"
##
## Scheffe Test for log(math1)
##
## Mean Square Error  : 0.006450156
##
## star1,  means
##
##              log.math1.        std    r      Min      Max
## regular        6.260795 0.07904431 2507 6.011267 6.516193
## regular+aide   6.268915 0.08062925 2225 6.001415 6.516193
## small          6.285785 0.08161399 1868 6.052089 6.516193
##
## Alpha: 0.05 ; DF Error: 6597
## Critical Value of F: 2.997093
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Means with the same letter are not significantly different.
##
##              log(math1) groups
## small          6.285785      a
## regular+aide   6.268915      b
## regular        6.260795      c
```

For task 7, we choose three methods to test the difference in the math scaled score in 1st grade across students in different class types. We use Tukey's Procedure, Bonfeeoni's Procedure and Scheffe's procedure. For Tukey's Procedure, all the p values are less than 0.05, there is statistically significant among three factors. For Bonfeeoni's Procedure, we get the same result as Tukey's Procedure. However, for Scheffe's procedure, we get the different result. It shows that Means with the same letter are not significantly different.

another organization In this part, in order to investigate whether there is a difference among the different factor level means, we use F-test. Moreover, to investigate comparisons between two factor level means simultaneously and control the family-wise type-I error, we use Tukey's Procedure, Bonferroni's Procedure.

F-test
To test null hypotheis $H_0 : \mu_1 = \mu_2 = \mu_3$ against alternative hypothesis: not all μ_i's are equal. We caculate F-statistic: $F^* = \frac{MSTR}{MSE} = 52.16923$ and $F(0.95, 2, 6597) = 2.997093$. Because $F^* > F(0.95, 2, 6597)$, We can thus reject the null hypothesis at the significance level 0.05. We can claim that there exists difference among

factor level means.

Tukey's Procedure

```r
TukeyHSD(anova.fit)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log(math1) ~ star1, data = data_remove_na)
##
## $star1
##                             diff          lwr         upr     p adj
## small-regular         0.02499081  0.019236297  0.03074533 0.0000000
## regular+aide-regular  0.00812069  0.002637091  0.01360429 0.0015104
## regular+aide-small   -0.01687012 -0.022778294 -0.01096196 0.0000000
```

There are three pairwise comparisons of factor levels means. And from the result, we can see that all of them should be declared as being different.

Bonferroni's Procedure

```r
pairwise.t.test(log(data_remove_na$math1),data_remove_na$star1,p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  log(data_remove_na$math1) and data_remove_na$star1
##
##              regular small
## small        < 2e-16 -
## regular+aide 0.0016  7.1e-11
##
## P value adjustment method: bonferroni
```

According to the result, all of p-value of three pairwise comparison is lesser than 0.05. We can make a conclusion that all of factor level means comparisions are different.

Bingdao Version

In order to explore this dataset, we first need to select the variables related to our goal. Because we only examine the math scores in 1st grade in this project, we exclude some variables with information higher than first grade, and we keep some variables to do with kindergarten for the reason that the education in kindergarten inevitably affects performance of children in the first grade. To note that, we don't include all the variables to do with kindergarten. For some factors, for examples: whether the student qualified for free lunch in kindergarten; school type in kindergarten, highest degree of kindergarten teacher, teacher's career ladder level in kindergarten, etc, these things can hardly affect the performance of children in 1st grade. Compared with these factors in 1st grade, the influcen of these factors in kindergarten on first-grade chilidren can be negligible. Aiming to make a good summary to show an overview of this dataset, we narrow the 47 variables to these variables: "gender", "ethnicity", "birth", "stark", "star1", "star2", "star3", "readk", "read1", "read2", "read3", "mathk", "math1", "math2", "math3", "lunch1", "school1", "degree1", "ladder1", ("experience1"), "tethnicity1".

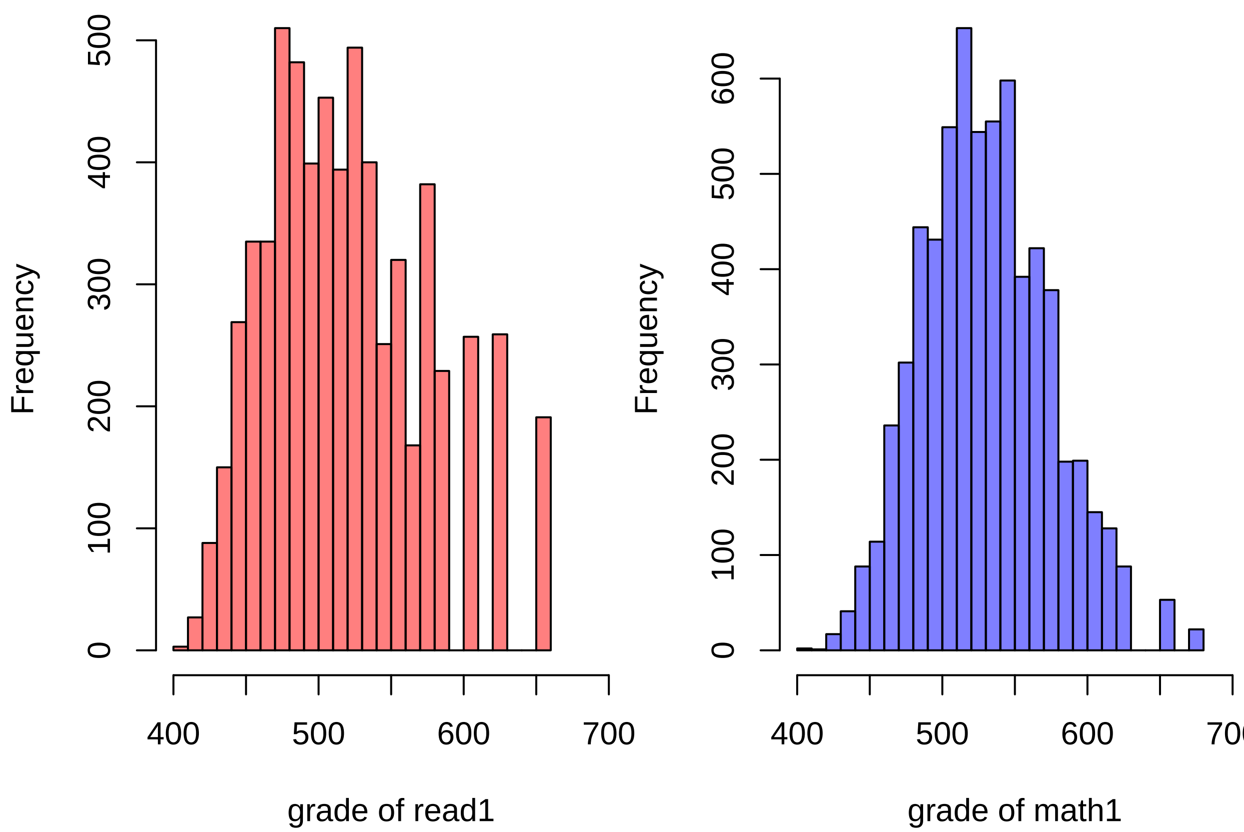
Among them, "readk", "read1", "read2", "read3", "mathk", "math1", "math2", "math3" variables are quantitative variables. The following is a summary table for them.

Variable	Min	Median	Mean	Max	Variance
`readk`	315.0	433.0	436.7	627.0	1005.3

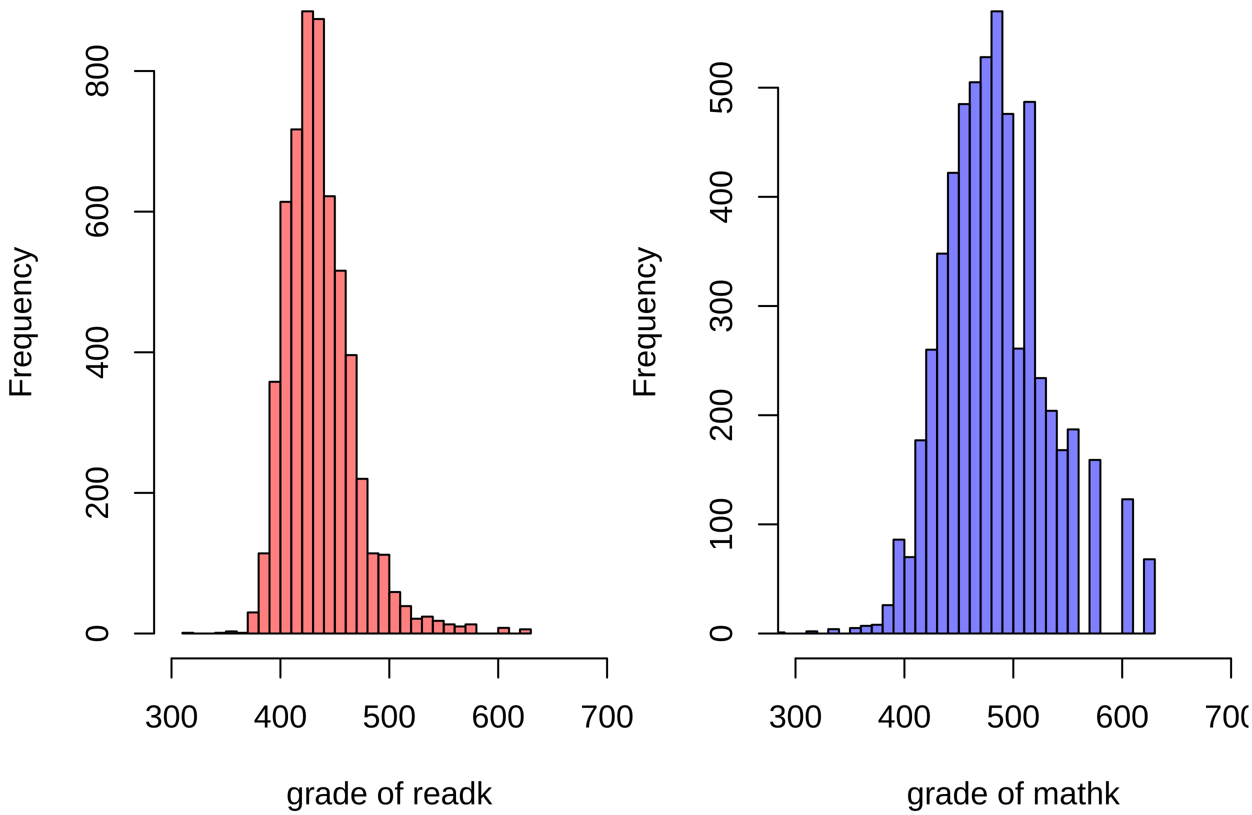
Variable	Min	Median	Mean	Max	Variance
`read1`	404.0	514.0	520.8	651.0	3045.8
`read2`	468.0	582.0	583.9	732.0	2119.962
`read3`	499.0	614.0	615.4	775.0	1487.349
`mathk`	288.0	484.0	485.4	626.0	2275.088
`math1`	404.0	529.0	530.5	676.0	1857.948
`math2`	441.0	579.0	580.6	721.0	1986.84
`math3`	487	616	618	774	1587.08

And we make histograms for "read" and "math" to show its distribution. The following is the figure of "read1" and "math1", and others will be shown in appendix.

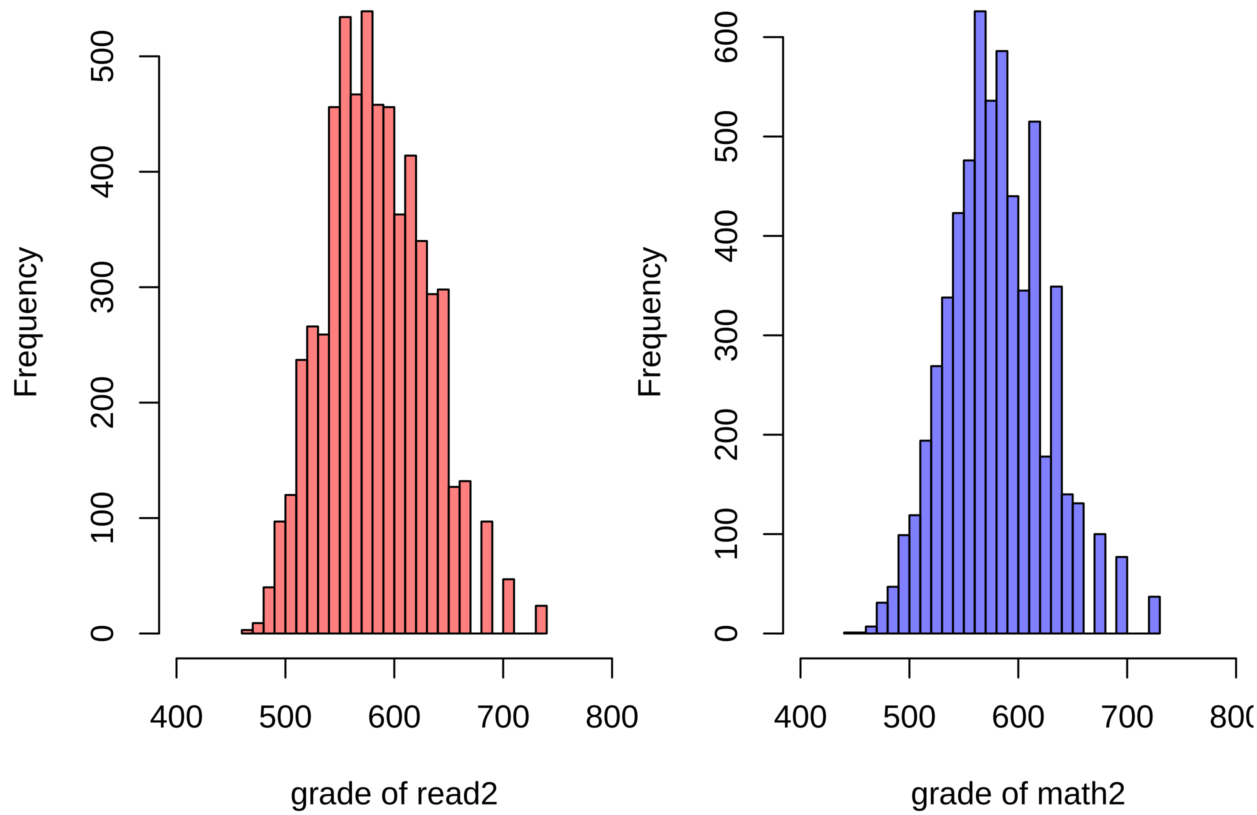
```
par(mfrow=c(1,2),mar=c(4,4,1,0))
hist(STAR$read1, breaks=30 , xlim=c(400,700) , col=rgb(1,0,0,0.5) , xlab="grade of read1", main="")
hist(STAR$math1, breaks=30 , xlim=c(400,700) , col=rgb(0,0,1,0.5) , xlab="grade of math1", main="")
```



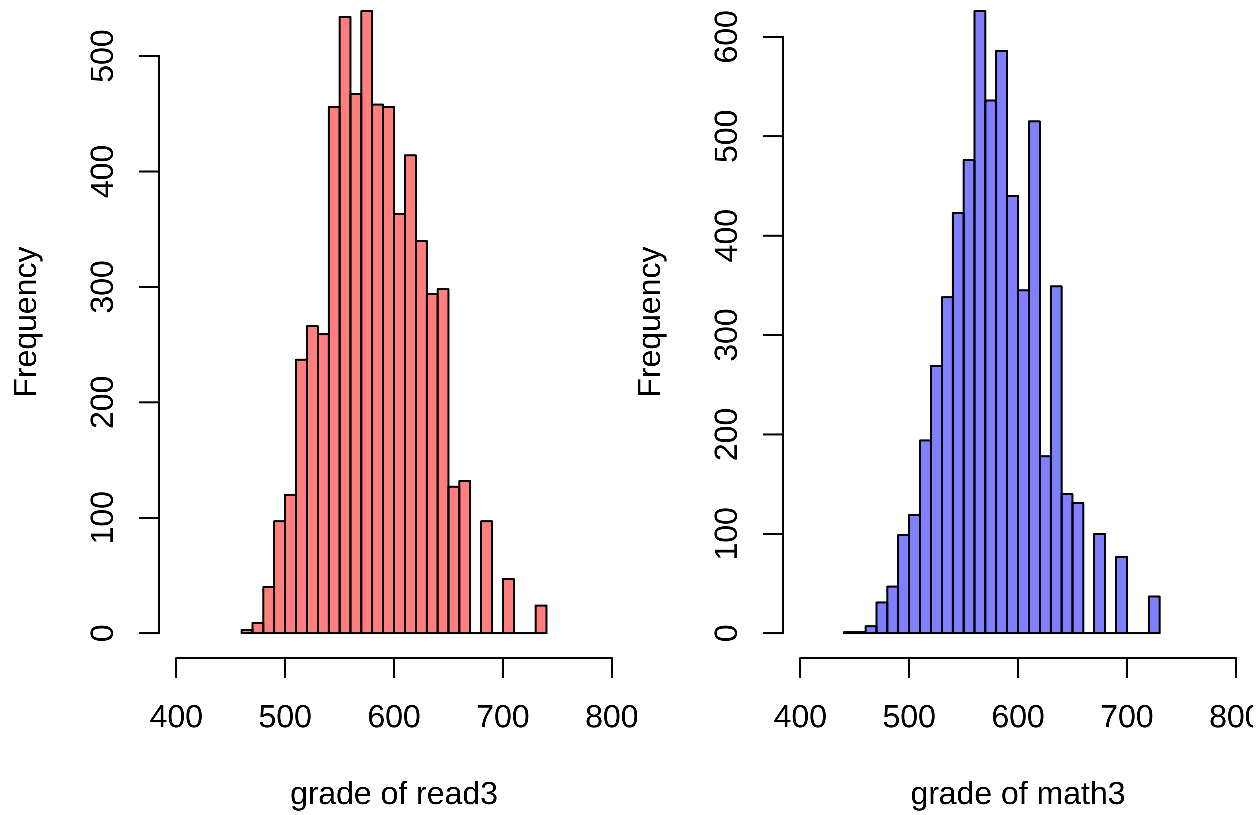

```
#appendix
# for k
par(mfrow=c(1,2),mar=c(4,4,1,0))
hist(STAR$readk, breaks=30 , xlim=c(300,700) , col=rgb(1,0,0,0.5) , xlab="grade of readk", main="" )
hist(STAR$mathk, breaks=30 , xlim=c(300,700) , col=rgb(0,0,1,0.5) , xlab="grade of mathk", main="")
```




```
# for 2
par(mfrow=c(1,2),mar=c(4,4,1,0))
hist(STAR$read2, breaks=30 , xlim=c(400,800) , col=rgb(1,0,0,0.5) , xlab="grade of read2", main="" )
hist(STAR$math2, breaks=30 , xlim=c(400,800) , col=rgb(0,0,1,0.5) , xlab="grade of math2", main="")
```



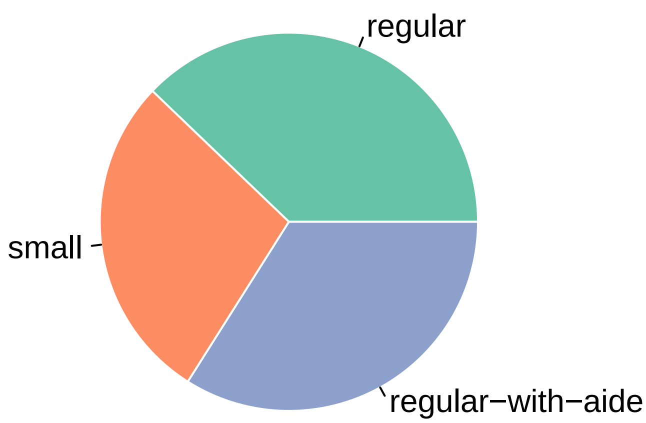

```
#for 3
par(mfrow=c(1,2),mar=c(4,4,1,0))
hist(STAR$read2, breaks=30 , xlim=c(400,800) , col=rgb(1,0,0,0.5) , xlab="grade of read3", main="" )
hist(STAR$math2, breaks=30 , xlim=c(400,800) , col=rgb(0,0,1,0.5) , xlab="grade of math3", main="")
```



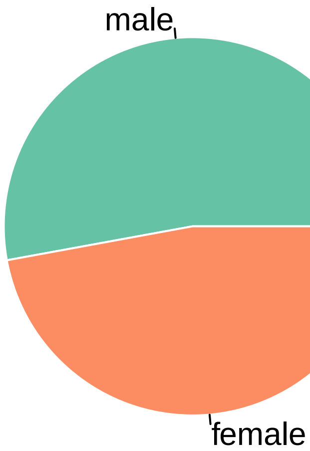
From the plot, it can seen that they are all bell-shaped, and with the higher grade, children's grade of read and math are closer.

Then we will draw pie charts for qualitative variables including "gender", "ethnicity", "birth", "stark", "star1", "star2", "star3", "lunch1", "school1", "degree1", "ladder1", "tethnicity1" variables. Before plotting pie charts, for readability, we collapse levels for "ethnicity" and "birth". We only show the pie chart for "star1" and others can be seen in appendix.

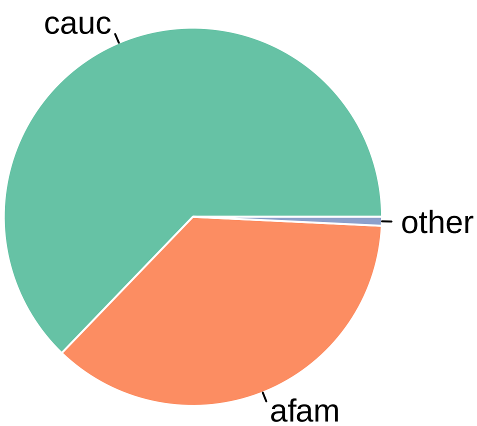
```
library(tidyverse)
library(RColorBrewer)
myPalette <- brewer.pal(6, "Set2")
pie(table(STAR$star1), labels = c("regular","small","regular-with-aide"), border="white", col=myPalette
```



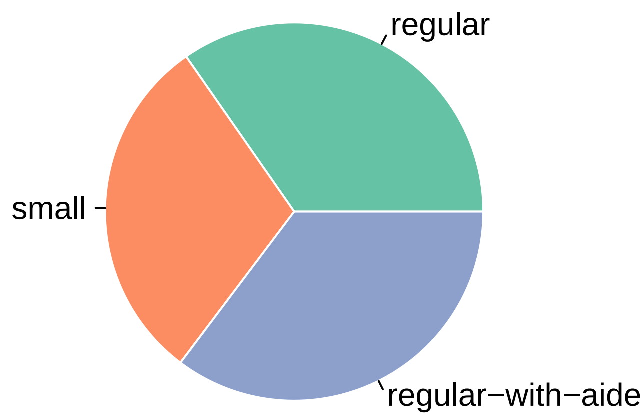

```
#appendix
#for gender
pie(table(STAR$gender), labels = c("male","female"), border="white", col=myPalette )
```



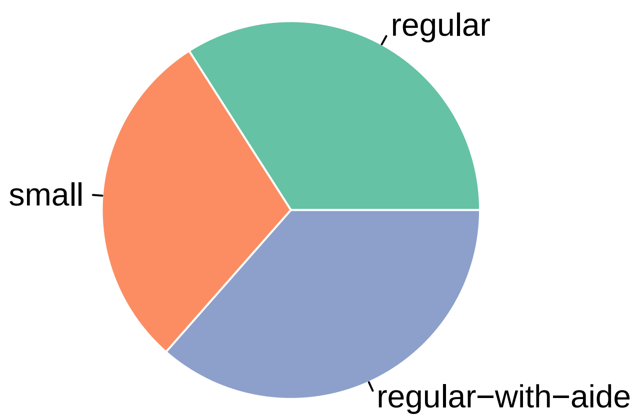

```
#for ethnicity
pie(c(table(STAR$ethnicity)[[1]],table(STAR$ethnicity)[[2]],sum(table(STAR$ethnicity)[[3]]+table(STAR$et
```



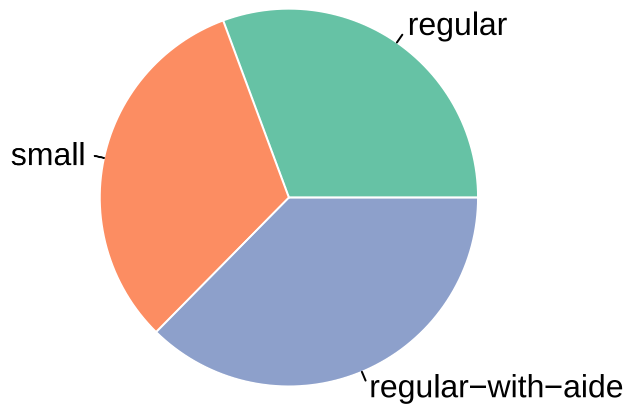

```
#for birth
#for stark
pie(table(STAR$stark), labels = c("regular","small","regular-with-aide"), border="white", col=myPalette
```



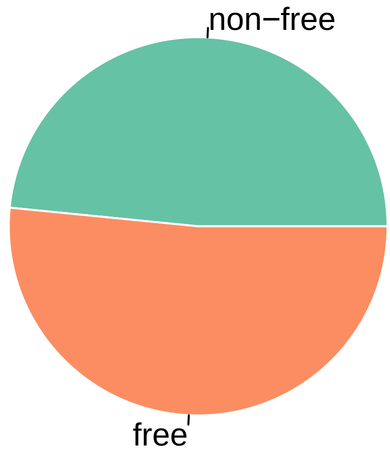

```
#for star2
pie(table(STAR$star2), labels = c("regular","small","regular-with-aide"), border="white", col=myPalette
```



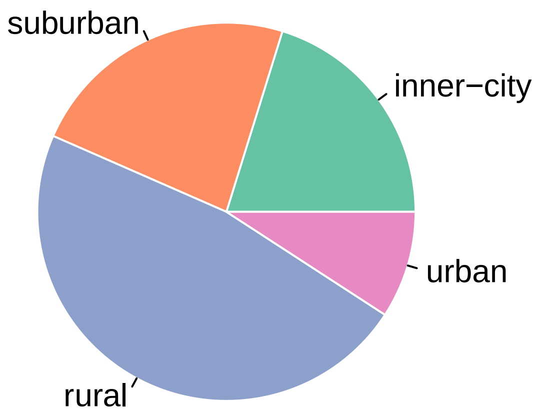

```
#for star3
pie(table(STAR$star3), labels = c("regular","small","regular-with-aide"), border="white", col=myPalette
```



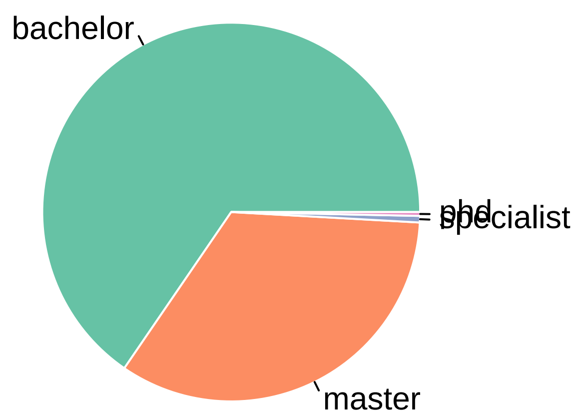

```
#for lunch1
pie(table(STAR$lunch1), labels = c("non-free","free"), border="white", col=myPalette )
```



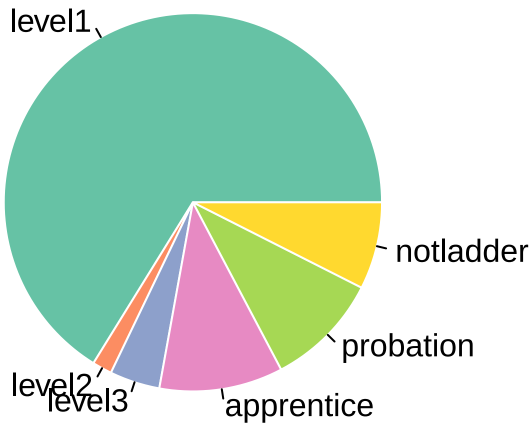

```
#for school1
pie(table(STAR$school1), labels = c("inner-city","suburban","rural", "urban"), border="white", col=myPal
```



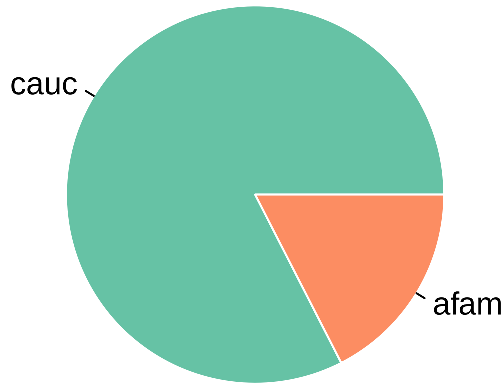

```r
#for degree1
pie(table(STAR$degree1), labels = c("bachelor","master","specialist", "phd"), border="white", col=myPal
```




```
#for ladder1
pie(table(STAR$ladder1), labels = c("level1","level2","level3", "apprentice", "probation", "notladder"))
```

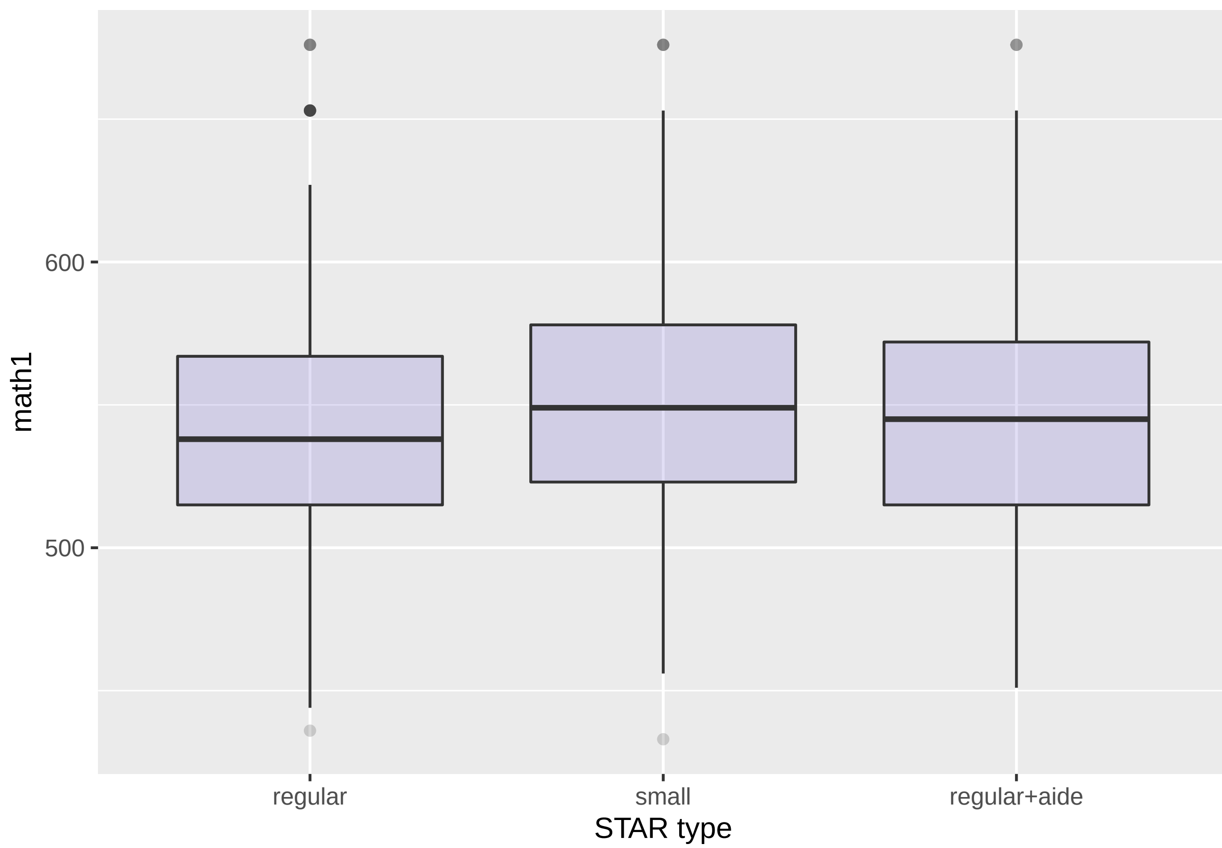



```r
#for tethnicity1
pie(table(STAR$tethnicity1), labels = c("cauc","afam"), border="white", col=myPalette )
```

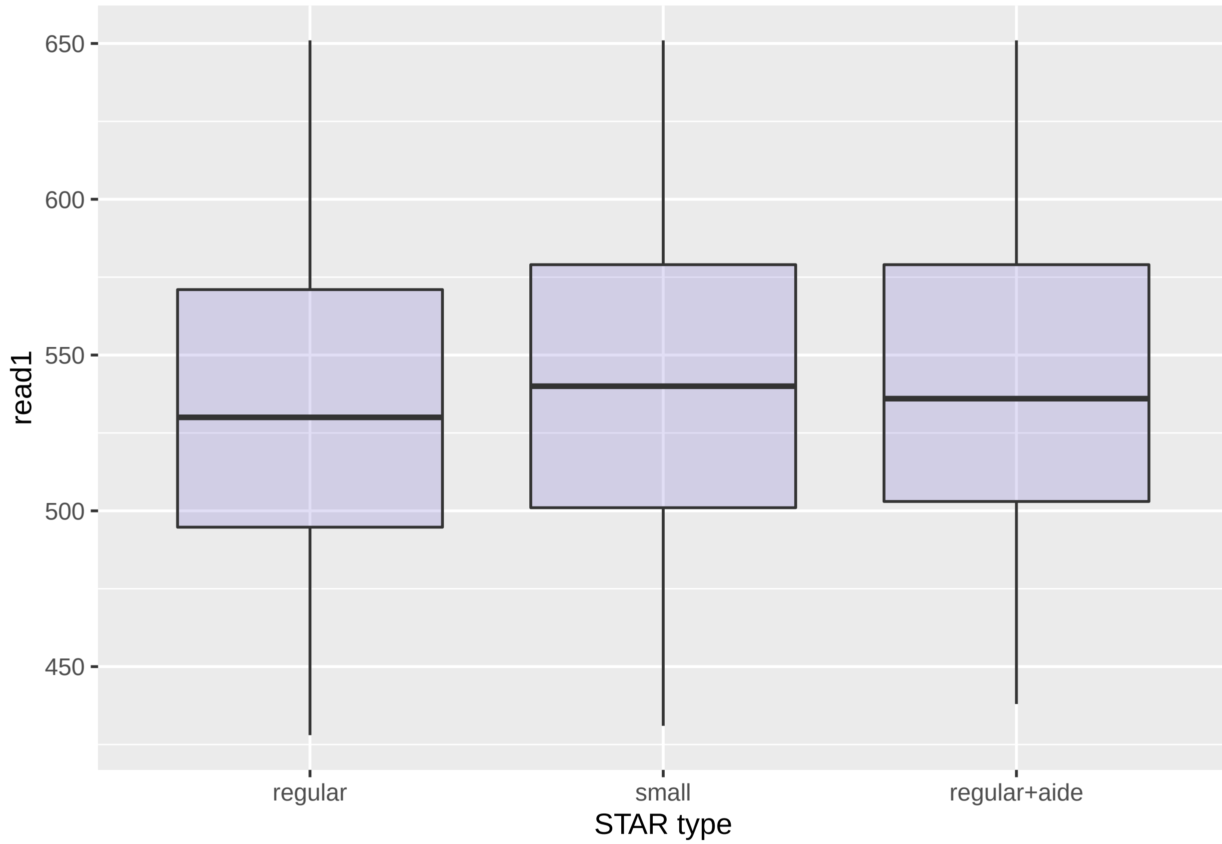


Then we give the boxplots to show the relationships between STAR class type with the grade of read, and math respectively. We only show the boxplot of "star1" and "math1", and others will be shown in appendix.

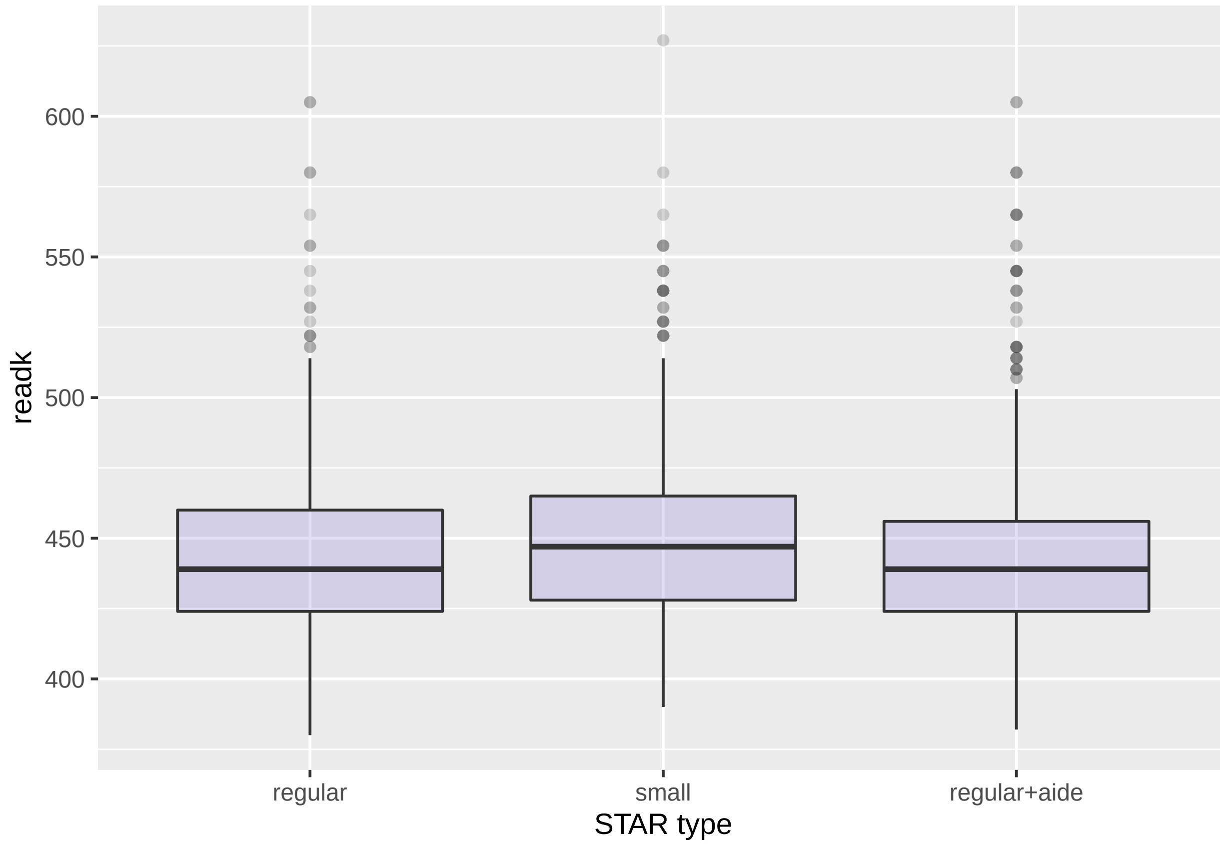
```
library(ggplot2)
ggplot(na.omit(STAR), aes(x=star1, y=math1)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("STAR type")
```



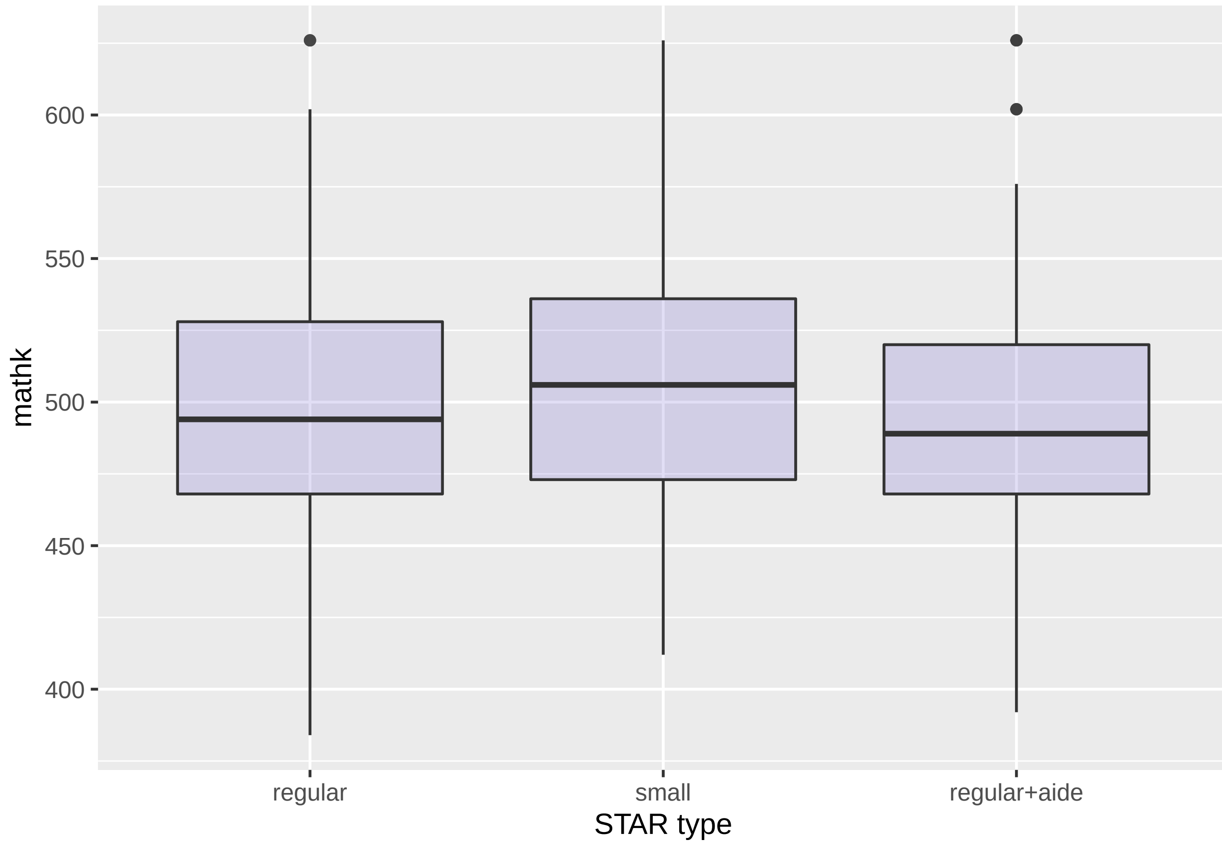

```r
#appendix
#star1 for read1
ggplot(na.omit(STAR), aes(x=star1, y=read1)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("STAR type")
```



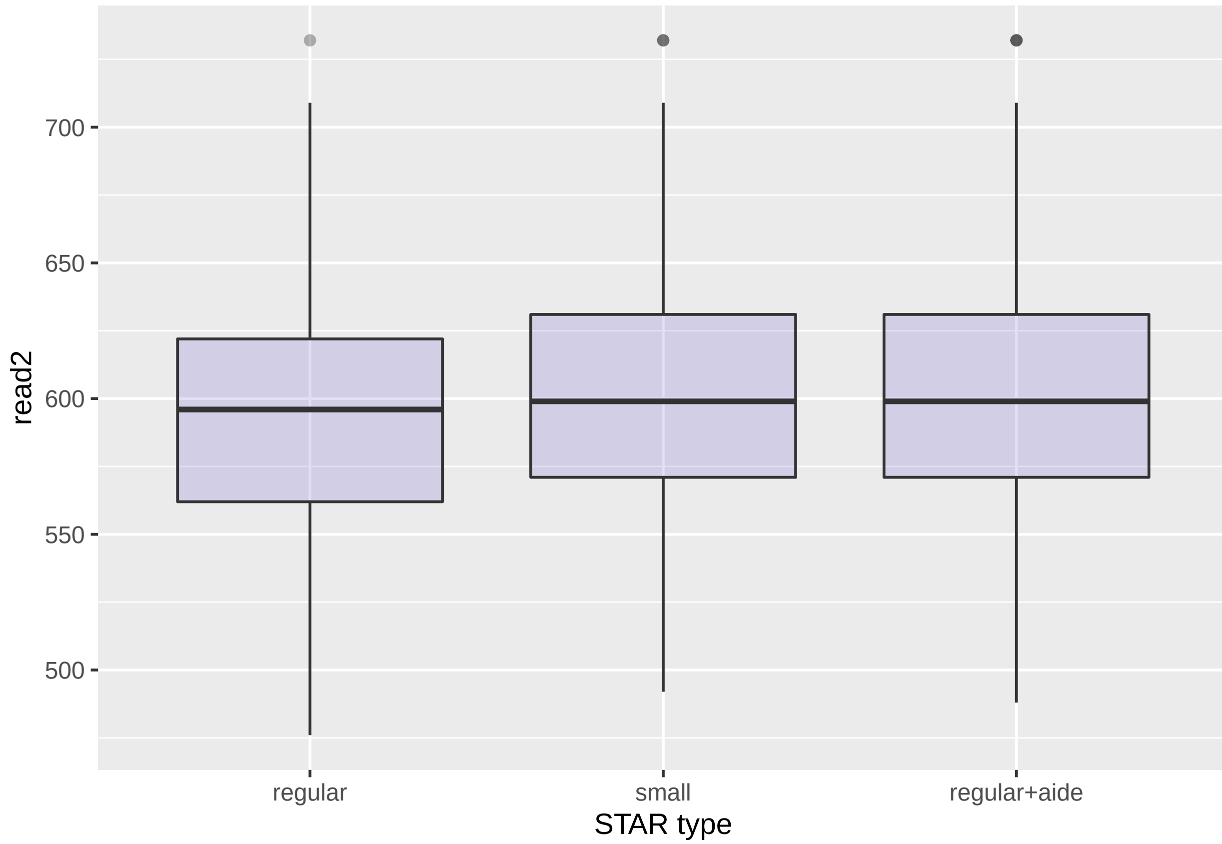

```
#stark for readk
ggplot(na.omit(STAR), aes(x=stark, y=readk)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("STAR type")
```



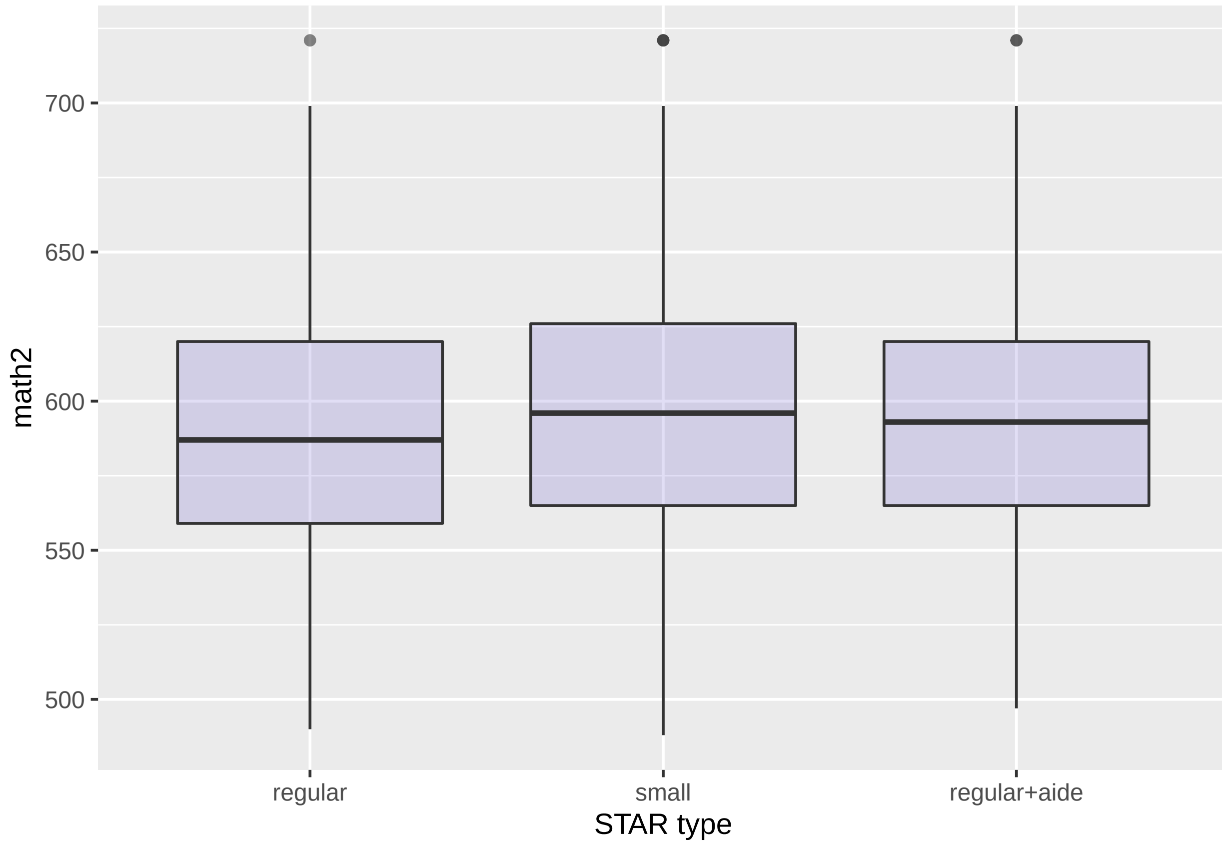

```
#stark for mathk
ggplot(na.omit(STAR), aes(x=stark, y=mathk)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("STAR type")
```



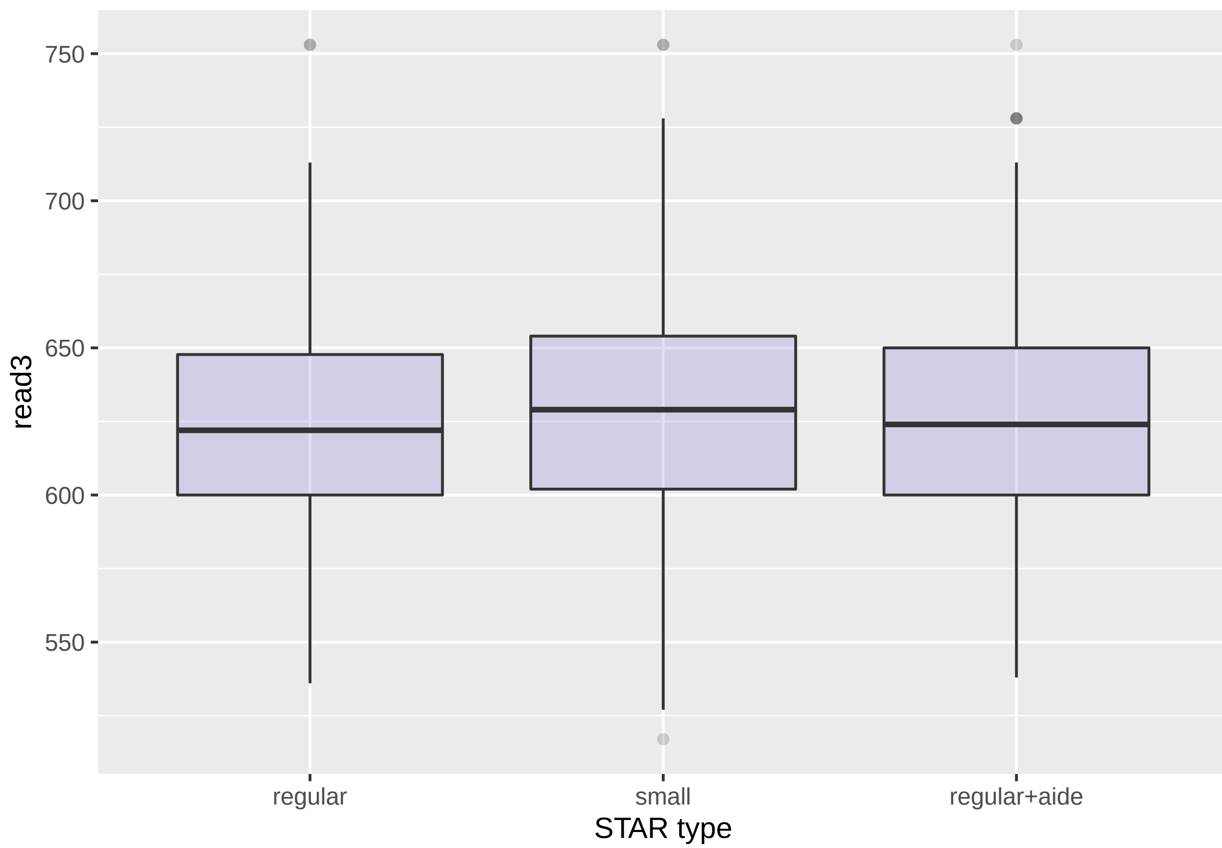

```
#star2 for read2
ggplot(na.omit(STAR), aes(x=star2, y=read2)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("STAR type")
```



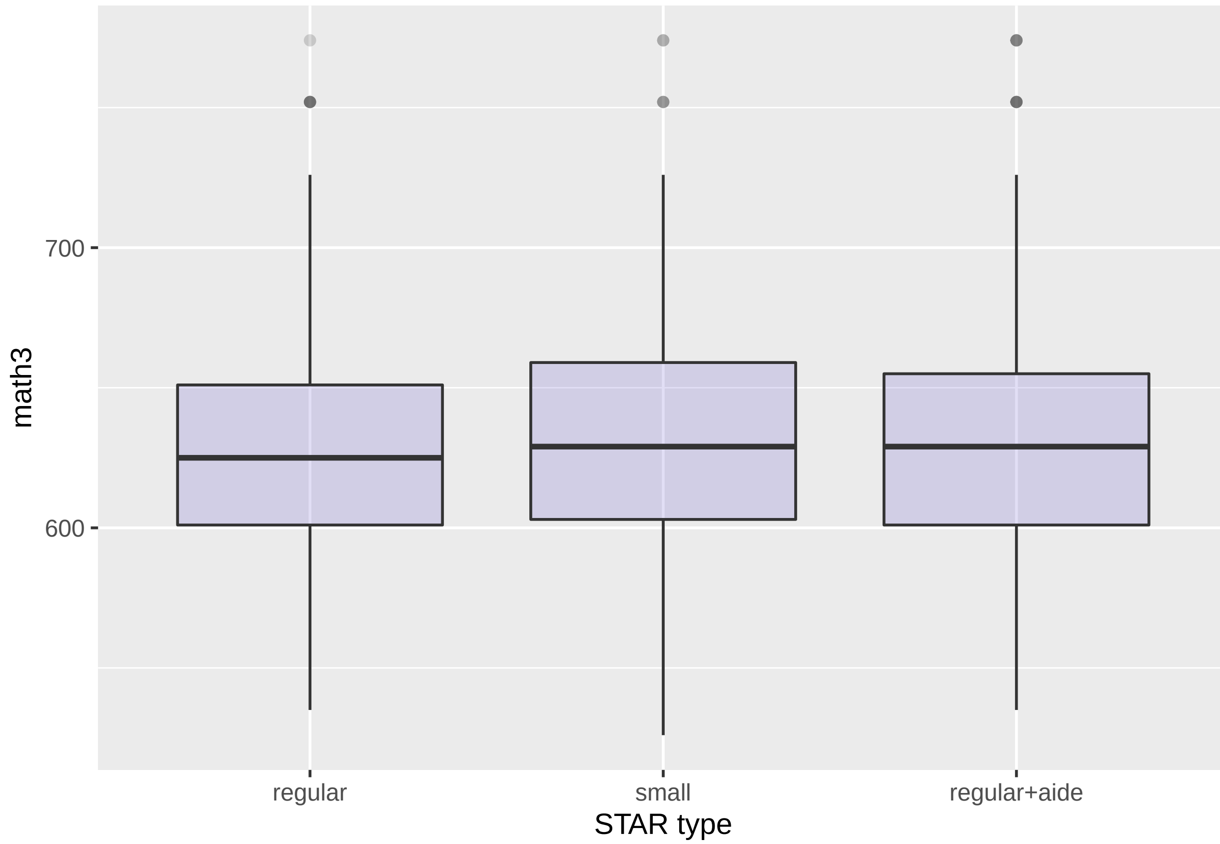

```r
#star2 for math2
ggplot(na.omit(STAR), aes(x=star2, y=math2)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("STAR type")
```




```
#star3 for read3
ggplot(na.omit(STAR), aes(x=star3, y=read3)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("STAR type")
```

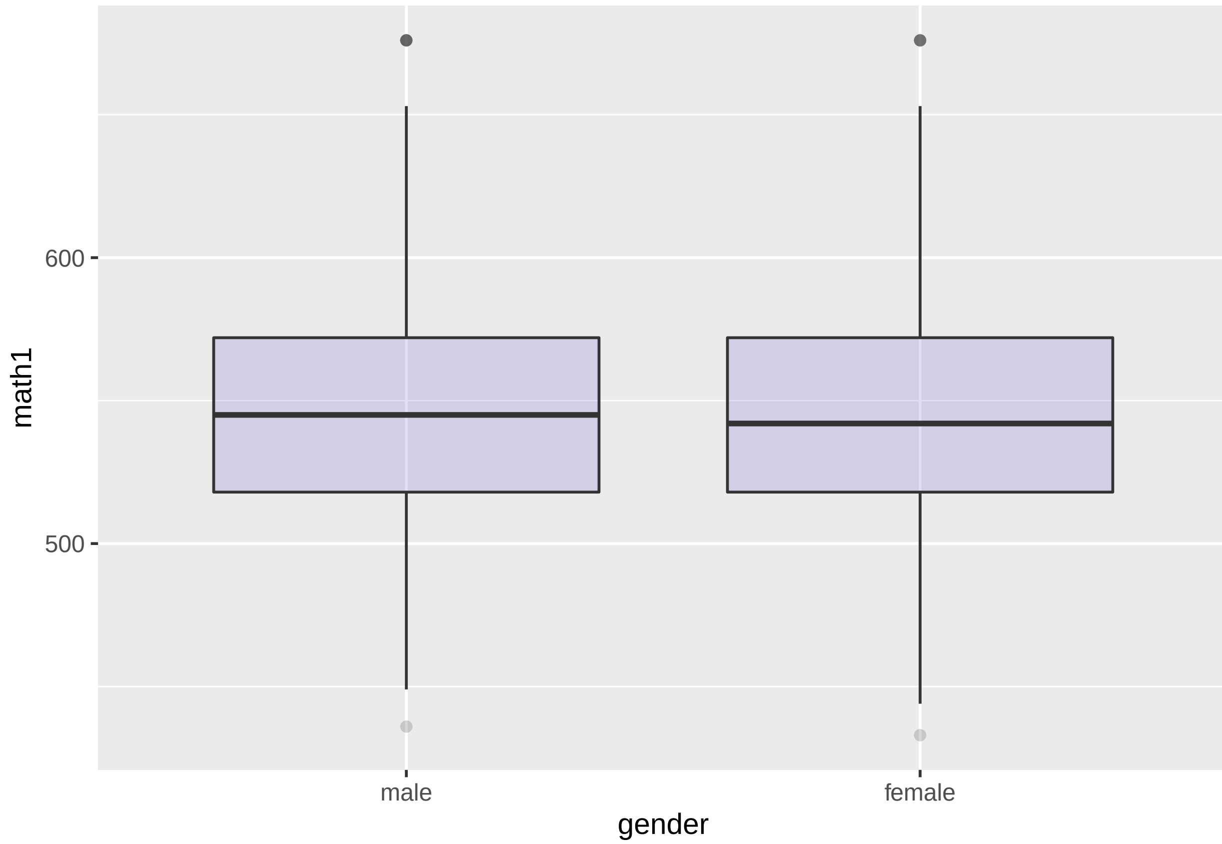



```r
#star3 for read3
ggplot(na.omit(STAR), aes(x=star3, y=math3)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("STAR type")
```

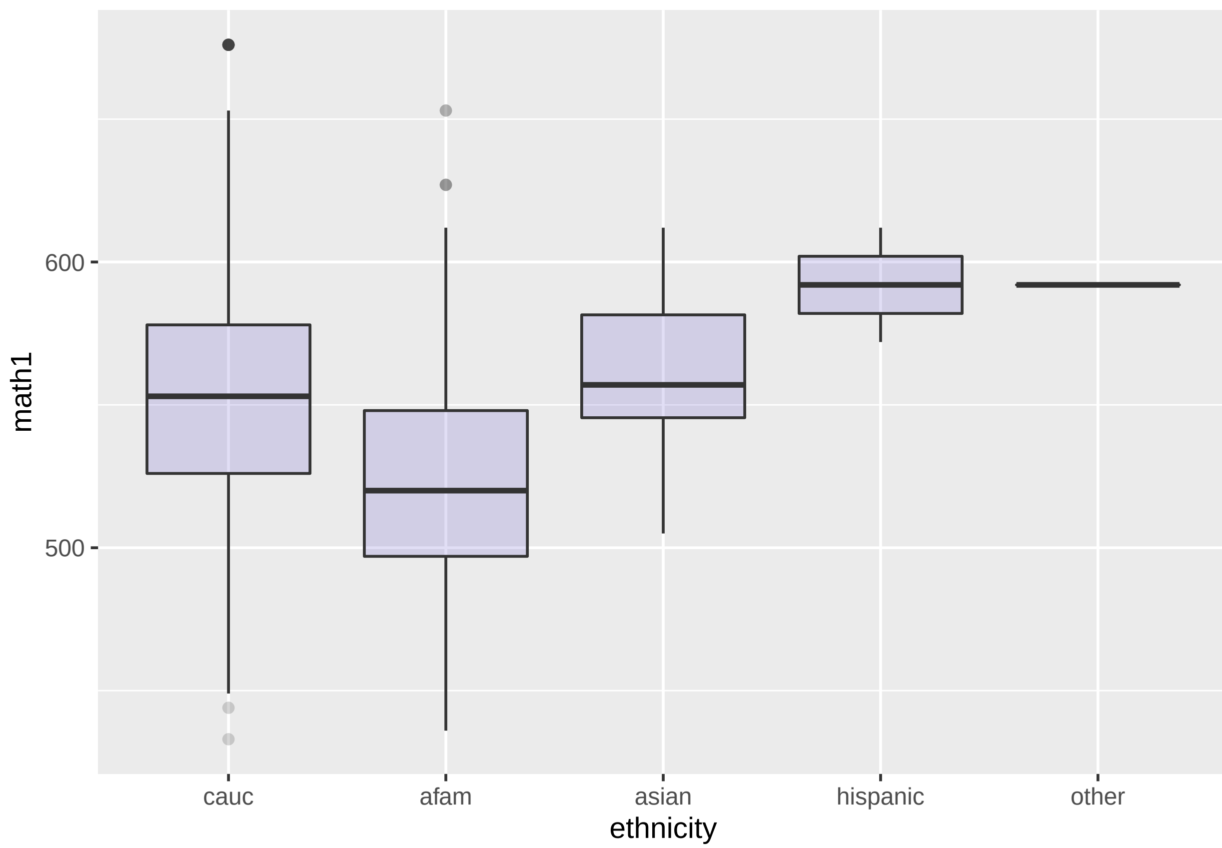


From the boxplot, we can see that children in small type class have a better performance in read and math compared with classes of other types, and with the higher grade, the grades of children in classes of different types are closer.
Next, we will plot the boxplot of "math1" with other qualatative variables, to find out whether these factors affect math grade.

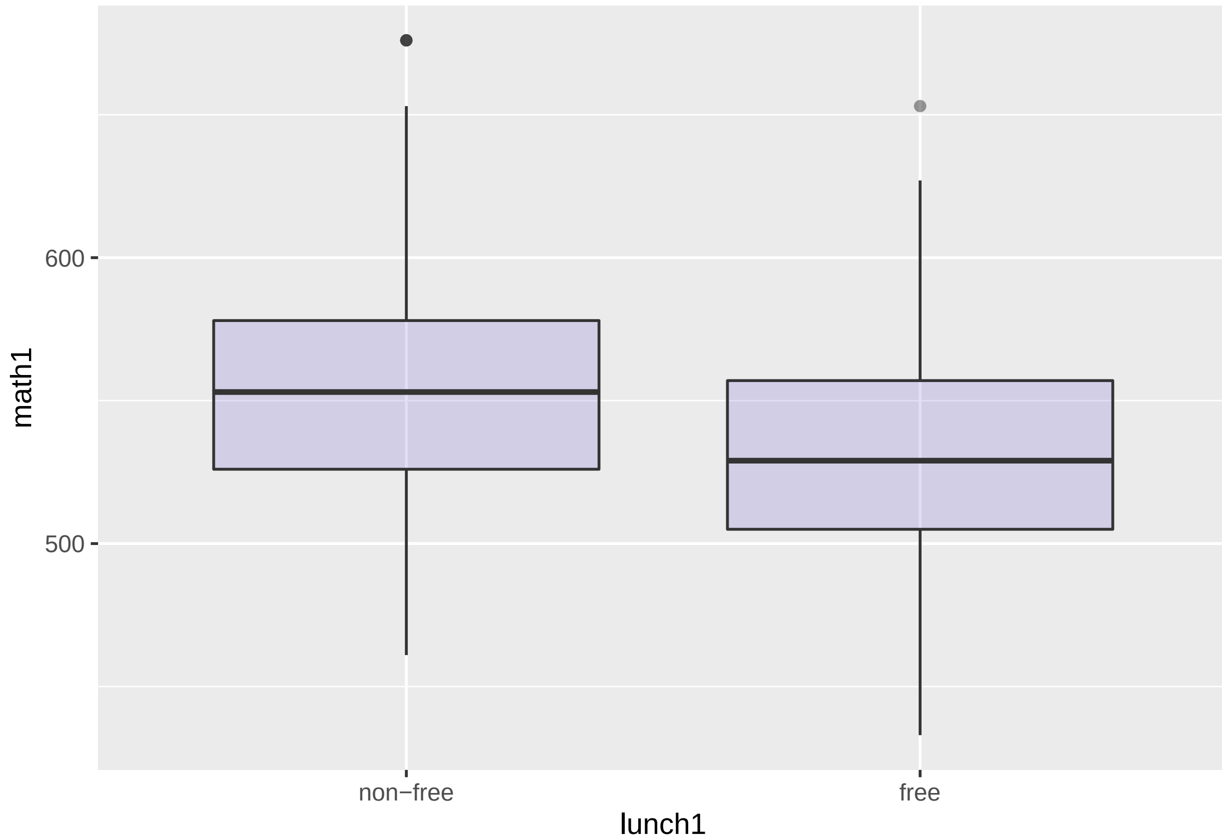
```
library(ggplot2)
#gender
ggplot(na.omit(STAR), aes(x=gender, y=math1)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("gender")
```



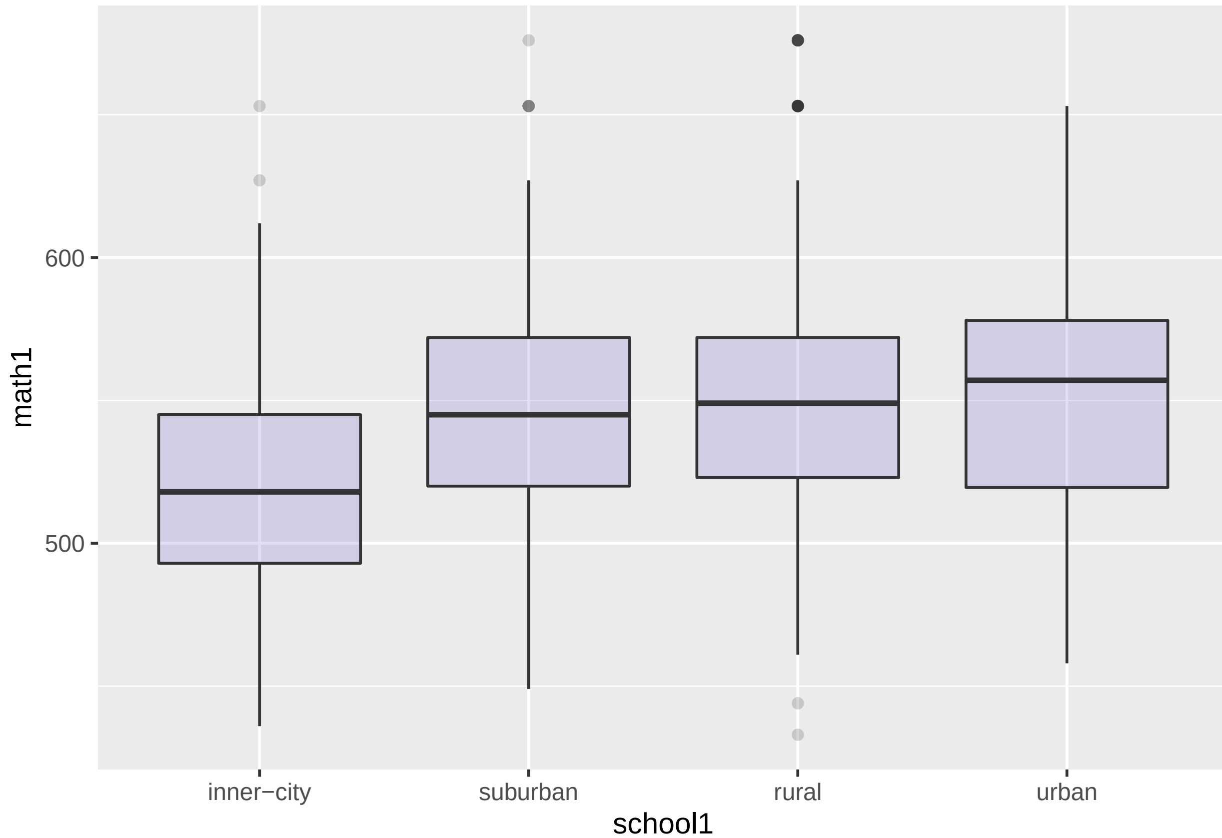

```r
#ethnicity
ggplot(na.omit(STAR), aes(x=ethnicity, y=math1)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("ethnicity")
```



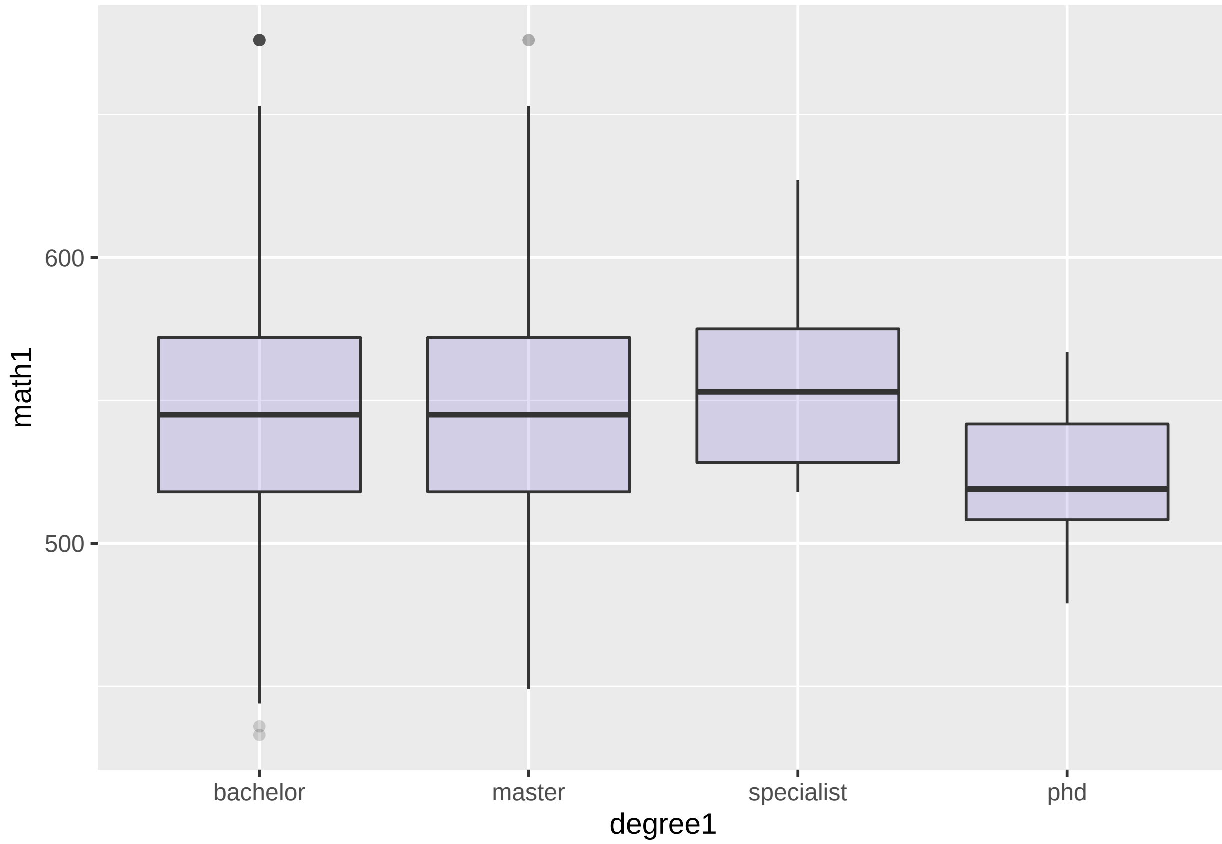

```
#lunch1
ggplot(na.omit(STAR), aes(x=lunch1, y=math1)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("lunch1")
```



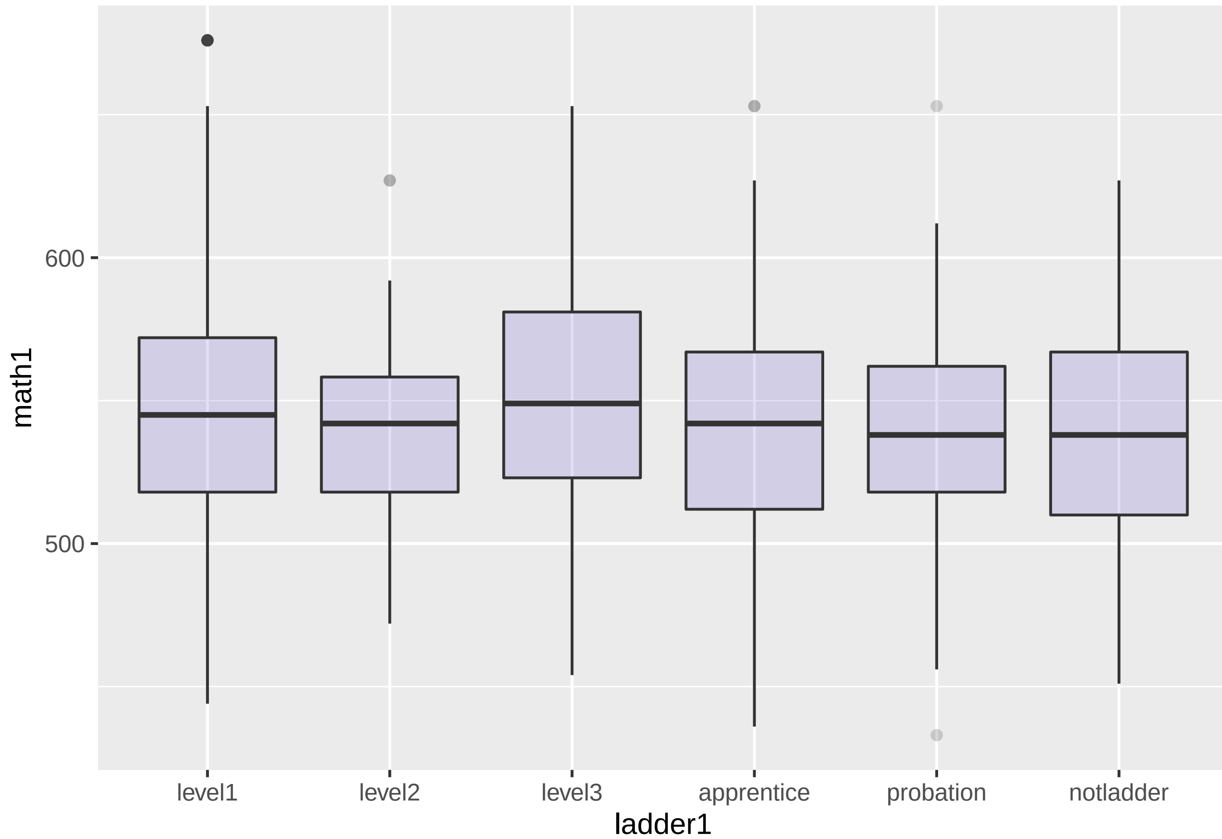

```
#school1
ggplot(na.omit(STAR), aes(x=school1, y=math1)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("school1")
```



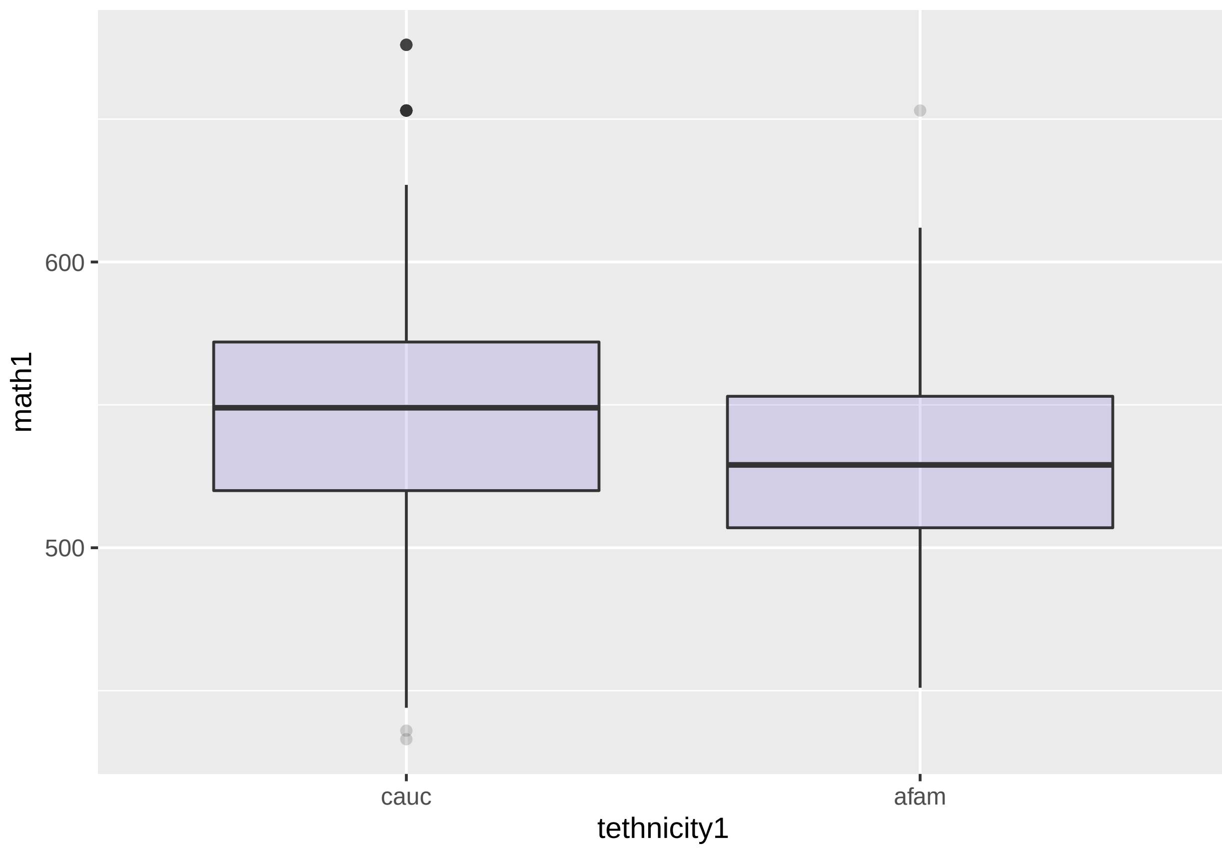

```r
#degree1
ggplot(na.omit(STAR), aes(x=degree1, y=math1)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("degree1")
```




```
#ladder1
ggplot(na.omit(STAR), aes(x=ladder1, y=math1)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("ladder1")
```




```
#experience
#tethnicity1
ggplot(na.omit(STAR), aes(x=tethnicity1, y=math1)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("tethnicity1")
```



As we can see, gender and ladder type don't have much effect on math grade. However, for ethnicity, hispanic's teacher have the best teaching effect, and African-American teacher have the worst teaching effect. For lunch, children having non-free lunch have higher math grade. For school,children from urban scholl have the best grade and children form inner-city school have the lowest grade. For teachers' degree, specialist have the best teaching effect, and phd have the worst teaching effect. For teacher's ethnicity, Caucasian teacher have the better teaching effect than African-American teacher.