Final Project Report

Nguyen Viet Hung - 20125058

July 27, 2022

1 Description Statistics

Open data

```
setwd("/mnt/d/Learning/Math/STAT452/Final_Project/")
data <- read.table("star.csv", header=TRUE)
attach(data)
```

Some basic inforamtion about the data:

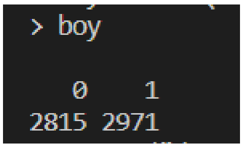
1.1 Qualitative description

1.1.1 Boy

This variable show that the teacher is a boy or not
Run the code below:

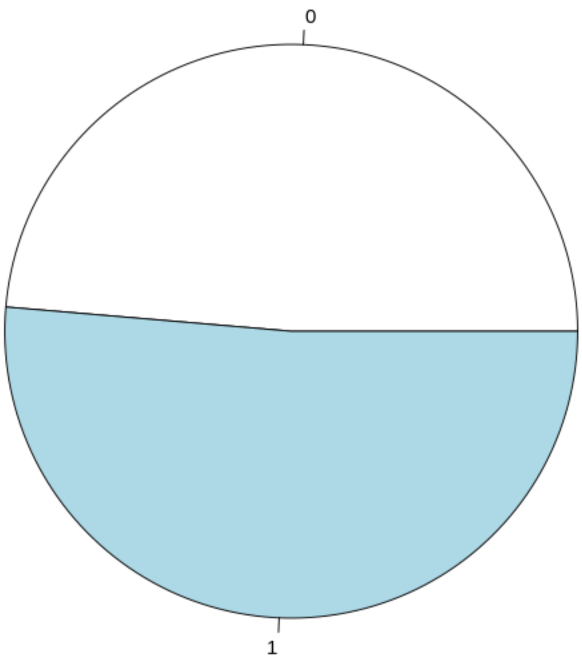
```
boy<-table(data$boy)
boy
pie(boy)
```

We also have the result of the number of boy in the data:



```
> boy

   0    1
2815 2971
```

And the pie plot of the boys number in the data



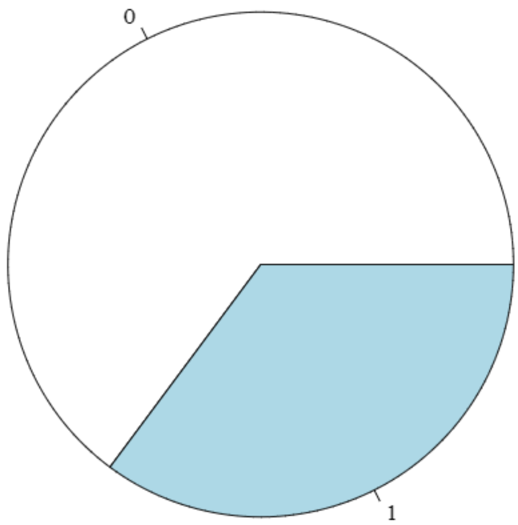
We can see that 51.35% of the teacher in the data are boys, and the other 48.65% are girls

1.1.2 Teacher who has master degree (tchmasters)

This variable show that the teacher has master degree or not
Run the code below

```
tchmasters<-table(data$tchmasters)
tchmasters
pie(tchmasters)
```

We have the circle diagram below z‘’



We can conclude the number of teacher have master degree is less than the numeber of teachers that don’t have master degree

1.1.3 Free lunch provided (freelunch)

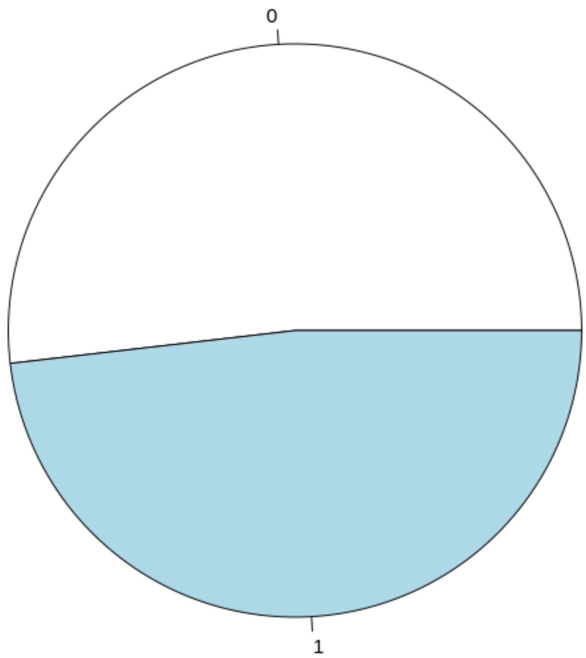
Run the code

```
freelunch<-table(data$freelunch)
freelunch
pie(freelunch)
```

We have the result

```
   0    1
2999 2787
```

And the graph:



In this graph, we can see that the number of teacher have freelunch is less than the number of teachers that don't have it.

1.2 Quantitative statistics

1.2.1 Absence

Run the code:

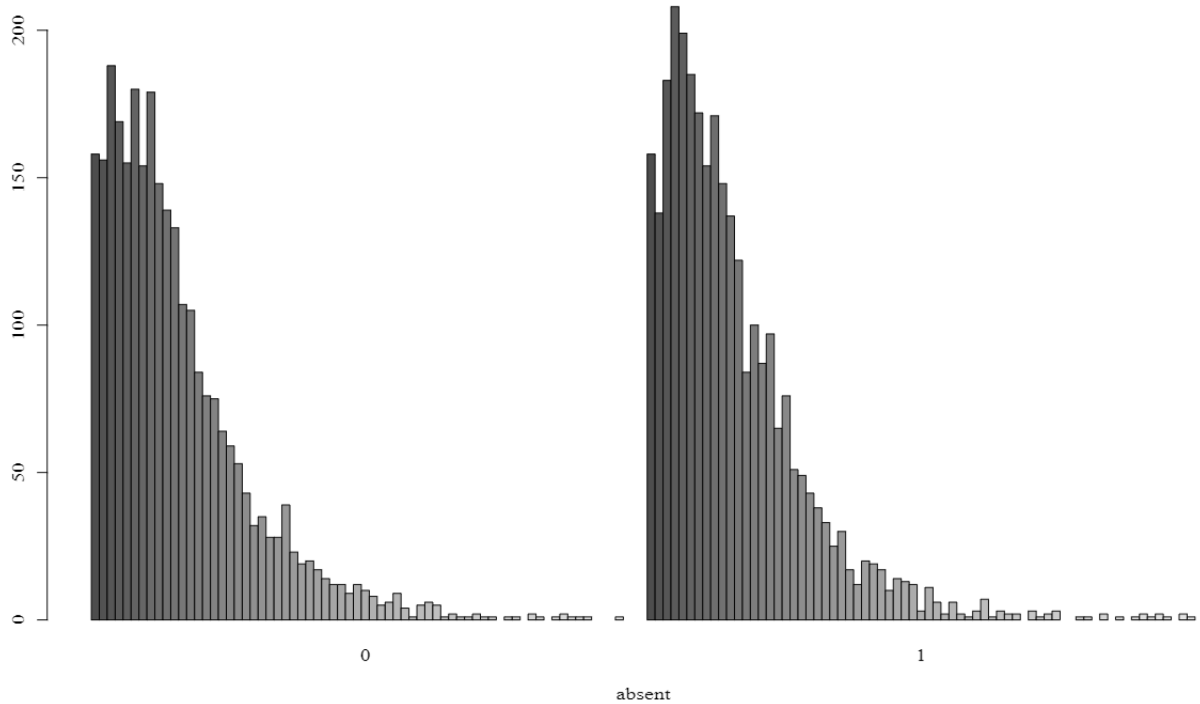
```
tssex.absence<-table(data$absent, data$boy)
tssex.absence
```

Than we have the result:

```
> pie(freelunch)
> tssex.absence<-table(data$absent, data$boy)
> tssex.absence

      0   1
  0  158 158
  1  156 138
  2  188 183
  3  169 208
  4  155 199
  5  180 185
  6  154 172
  7  179 154
  8  148 171
  9  139 148
  10 133 137
  11 107 122
  12 105  84
  13  84 100
  14  76  87
  15  75  97
  16  64  65
  17  59  76
  18  53  51
  19  43  49
  20  32  43
  21  35  38
  22  28  33
  23  28  25
  24  39  30
  25  23  17
  26  19  12
  27  20  20
  28  17  19
  29  14  17
  30  12  10
  31  12  14
  32   9  13
  33  12  12
  34  10   3
  35   8  11
  36   5   6
  37   6   2
  38   9   6
  39   4   2
  40   1   1
  41   5   3
  42   6   7
  43   5   1
  44   1   3
  45   2   2
  46   1   2
  47   1   0
  48   2   3
  49   1   1
  50   1   2
  51   0   3
  52   1   0
  53   1   0
  54   0   1
  55   2   1
  56   1   0
  57   0   2
  58   1   0
  60   2   1
  61   1   0
  63   1   1
  64   1   2
  67   0   1
  69   0   2
  70   0   1
  71   1   0
  74   0   2
  79   0   1
```

Plot the data by some code:

```
barplot(tssex.absence, beside=TRUE, xlab="absent")
```



We make the summary of this property by running in R:

```
summary(tssex.absence)
```

And here is the result:

```
> summary(tssex.absence)
Number of cases in table: 5765
Number of factors: 2
Test for independence of all factors:
        Chisq = 63.48, df = 68, p-value = 0.6328
        Chi-squared approximation may be incorrect
```

From the plot we can conclude that the average absence of female is near the average absence of male

1.2.2 Teaching experience (tchexper)

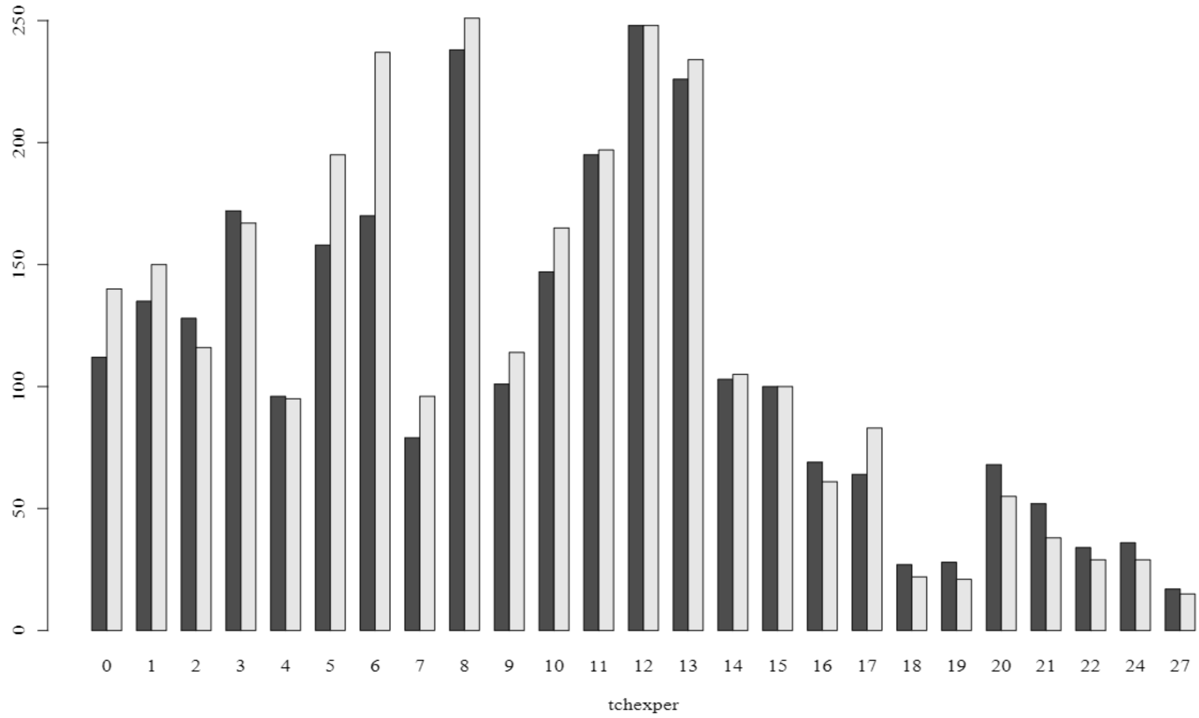
Run the code

```
tssex.teachEx<-table(data$boy, data$tchexper)
tssex.teachEx
barplot(tssex.teachEx, beside=TRUE, xlab="tchexper")
```

We have data from the combine of 2 properties:

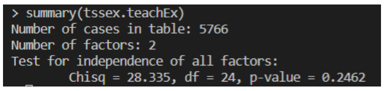
```
> tssex.teachEx

    0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
0 112 135 128 172  96 158 170  79 238 101 147 195 248 226 103 100  69  64  27
1 140 150 116 167  95 195 237  96 251 114 165 197 248 234 105 100  61  83  22

   19  20  21  22  24  27
0  28  68  52  34  36  17
1  21  55  38  29  29  15
```

We have the plot:



From this plot and the summary function in R:

```
summary(tchexper)
```

We have the result: 

```
> summary(tssex.teachEx)
Number of cases in table: 5766
Number of factors: 2
Test for independence of all factors:
        Chisq = 28.335, df = 24, p-value = 0.2462
```

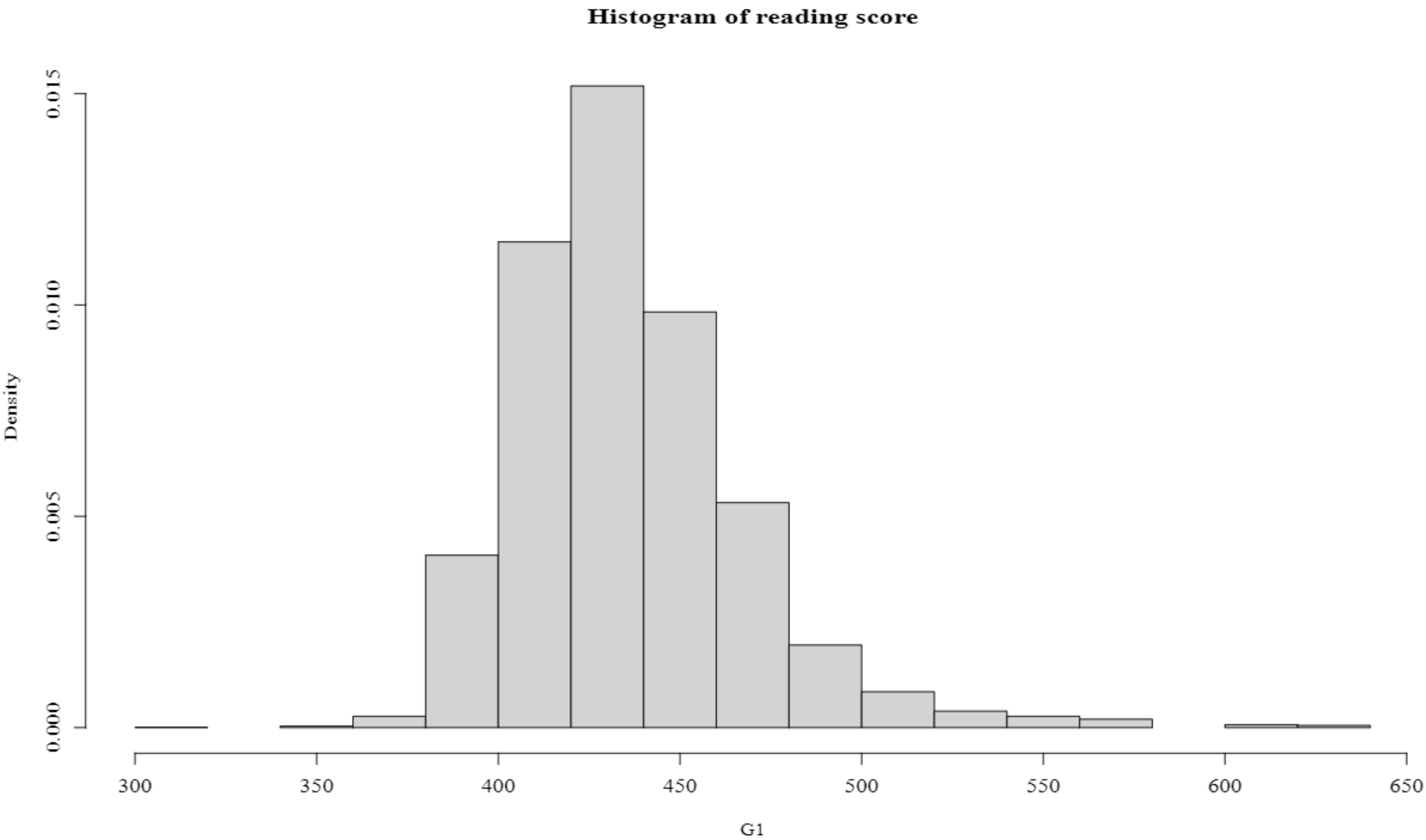
This plot show that the survey is true because the male and female following the teaching experience year is not different so much

1.2.3 Reading score

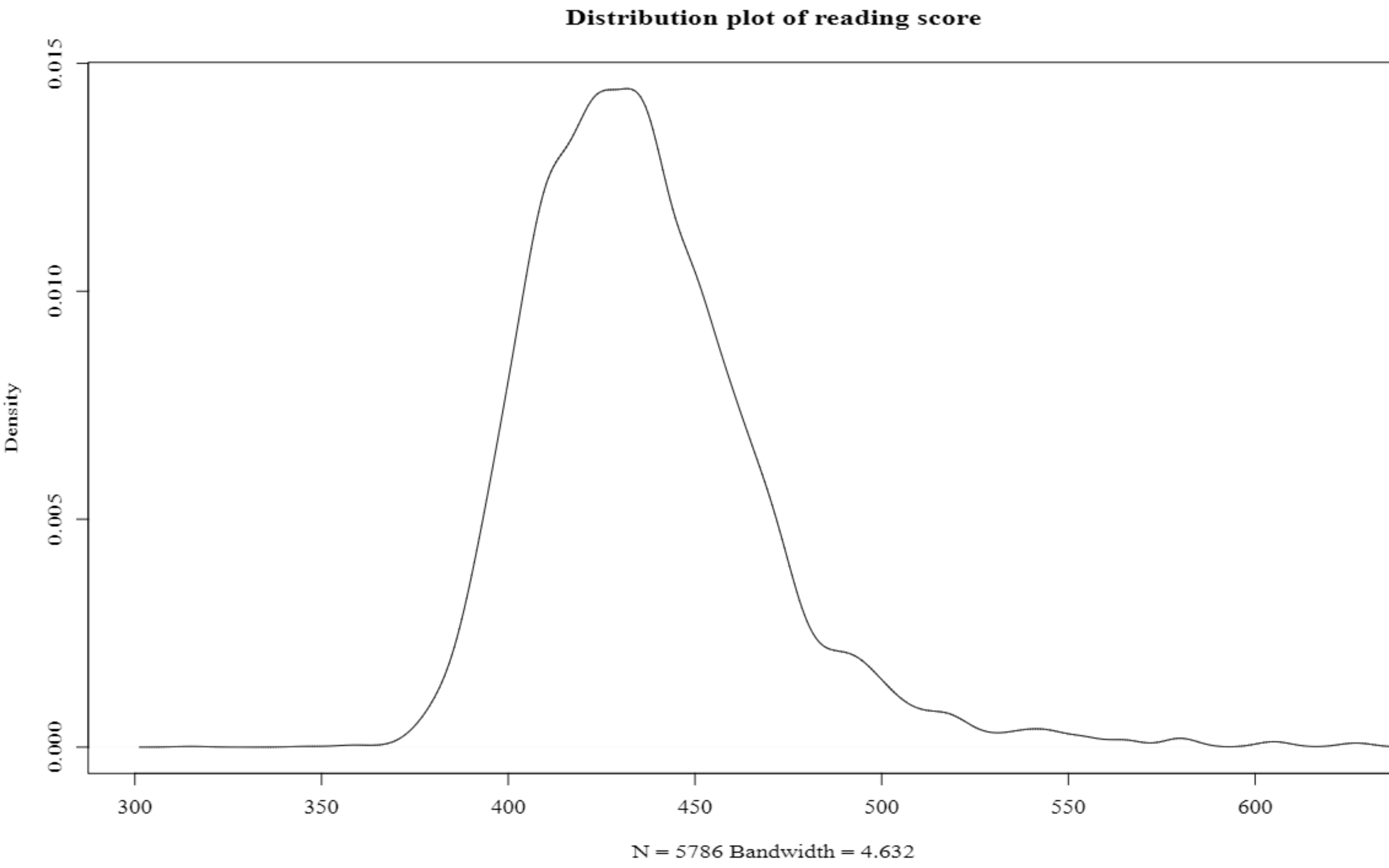
Run the code:

```
G1<-data$readscore
G1
hist(G1, freq=FALSE, main="Histogram of reading score")
plot(density(G1), add=TRUE, main="Distribution plot of reading score")
summary(G1)
```

And now we have 2 plots of the data
First, the histogram:



And the distribution



And the summary of the reading score data:

```
> summary(G1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  315.0   414.0   433.0   436.7   453.0   627.0 
```

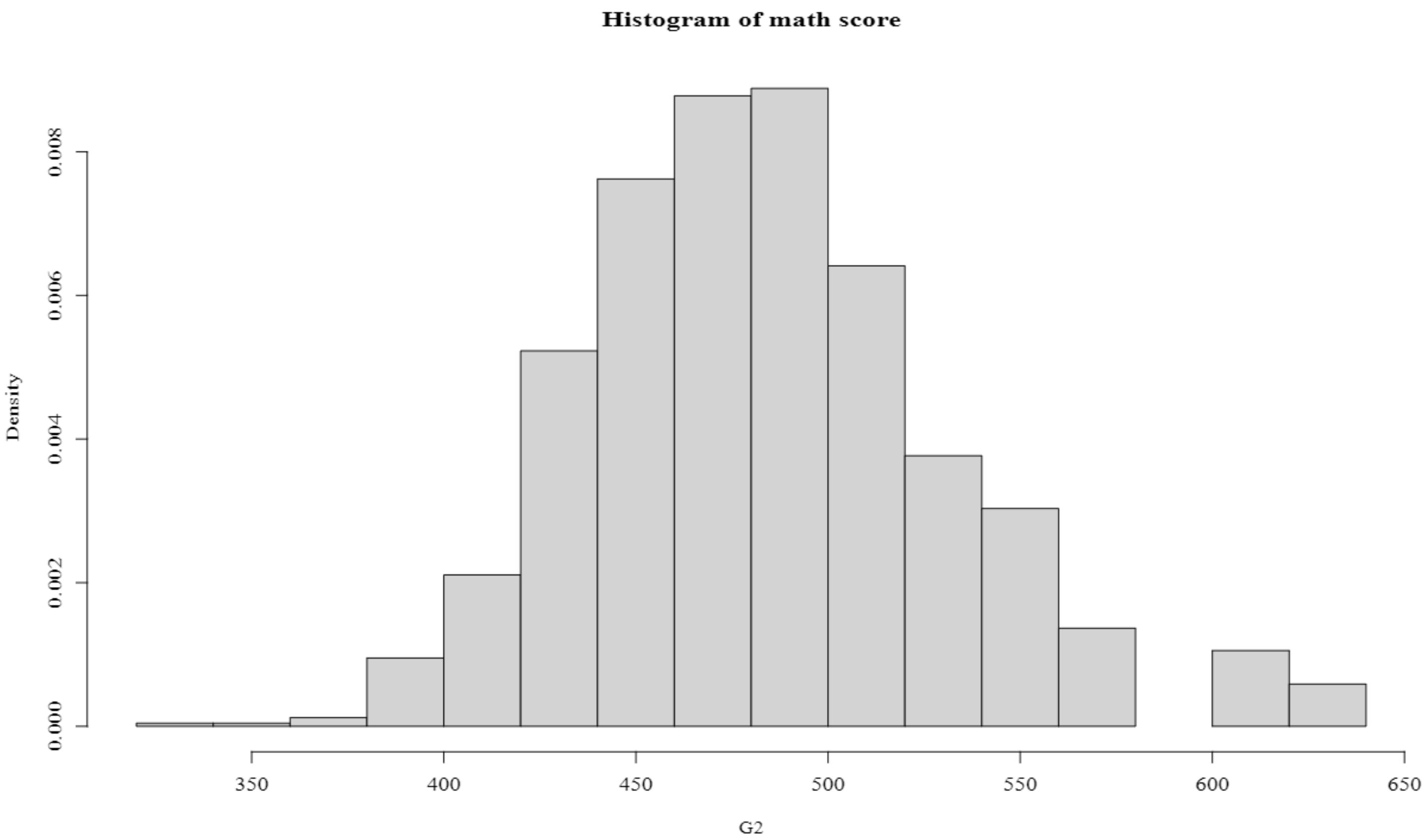
Two above plots and the summary data show that the most point fails into 433 point

1.2.4 Math score

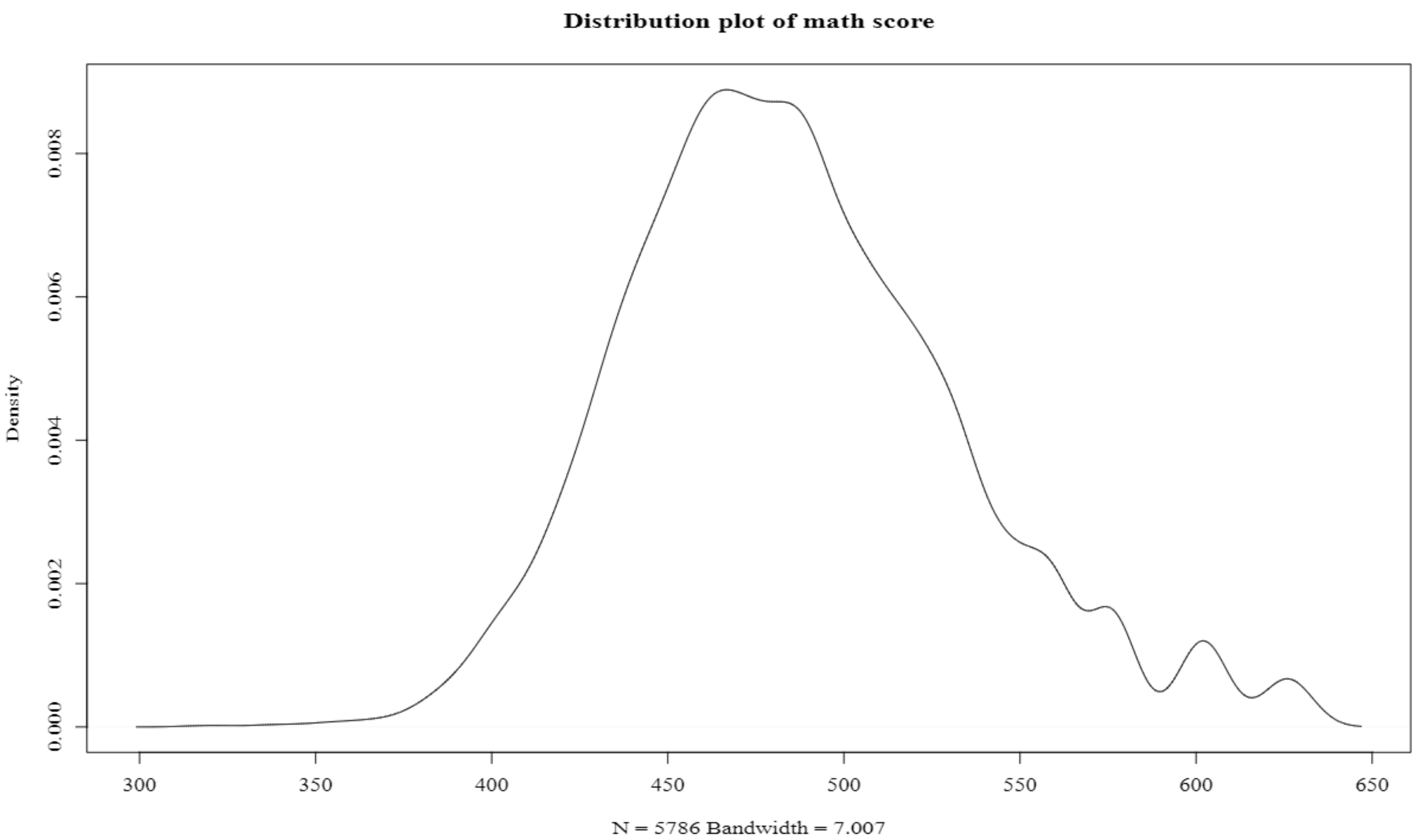
Run the code:

```
G2<-data$mathscore
G2
hist(G2, freq=FALSE, main="Histogram of math score")
plot(density(G2), add=TRUE, main="Distribution plot of math score")
summary(G2)
```

And now we have 2 plots of the data
First, the histogram:



Second, the distribution plot:



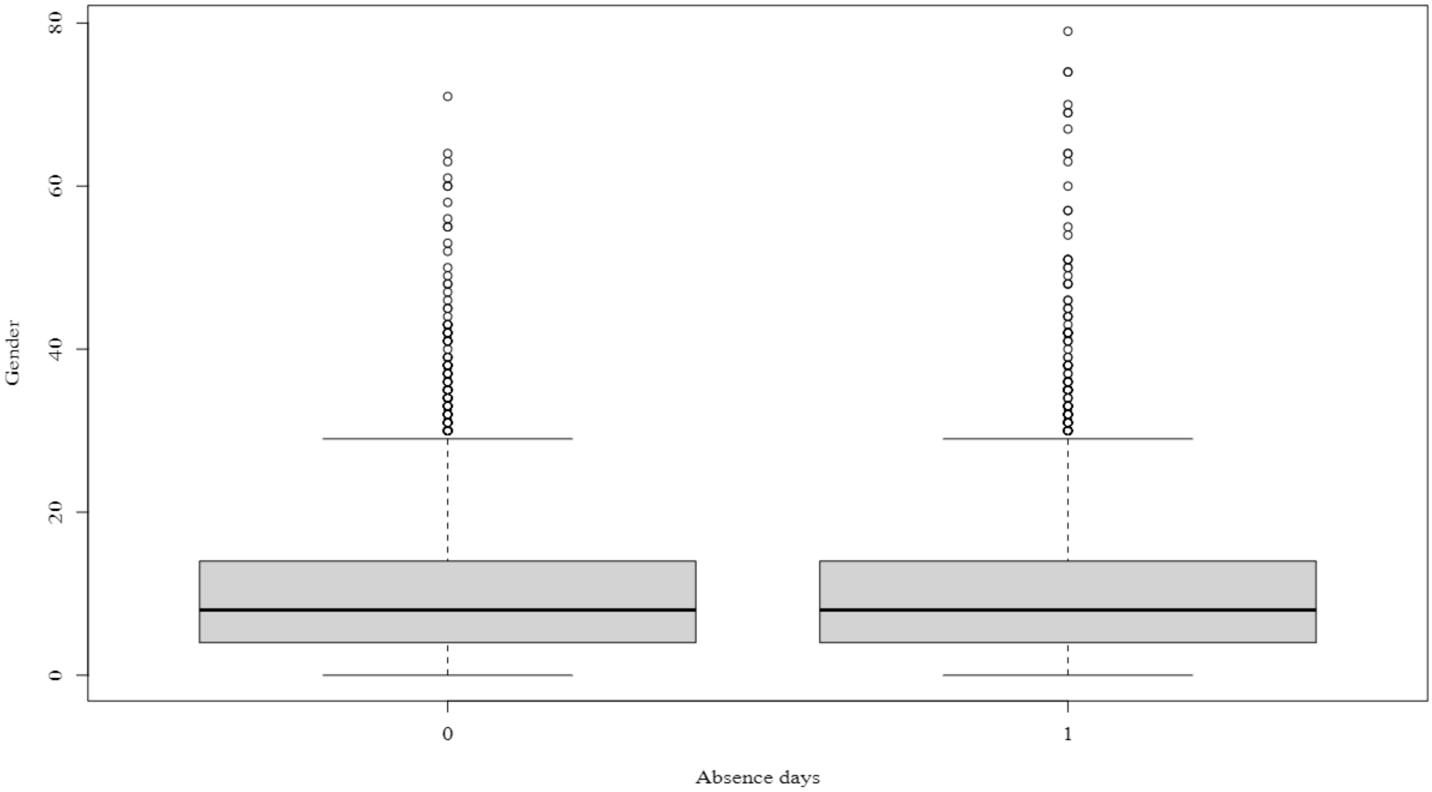
And the summary of the math score data:

```
> summary(G2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  320.0   454.0   484.0   485.6   513.0   626.0
There were 12 warnings (use warnings() to see them)
```

Two above plots and the summary data show that the most point fails into 485.6 point

2 Inferential statistics

2.1 The absence



We will test the absence means between male and female. Let's test the hypothesis that the females are absent more than males

We run the code below:

```
t.test(data$absent~data$boy, alternative="less")
```

The result is:

```
        Welch Two Sample t-test

data:  data$absent by data$boy
t = 0.10041, df = 5752.1, p-value = 0.54
alternative hypothesis: true difference in means between group 0 and group 1 is less than 0
95 percent confidence interval:
      -Inf 0.4263271
sample estimates:
mean in group 0 mean in group 1
       10.28770        10.26318
```

Because the p-value is greater than the $\alpha = 0.05$. So that we accept H_0

2.2 The black

We test the hypothesis that propotion of the black at male is more than female

We use these code to qualitative the black and the boy variable:

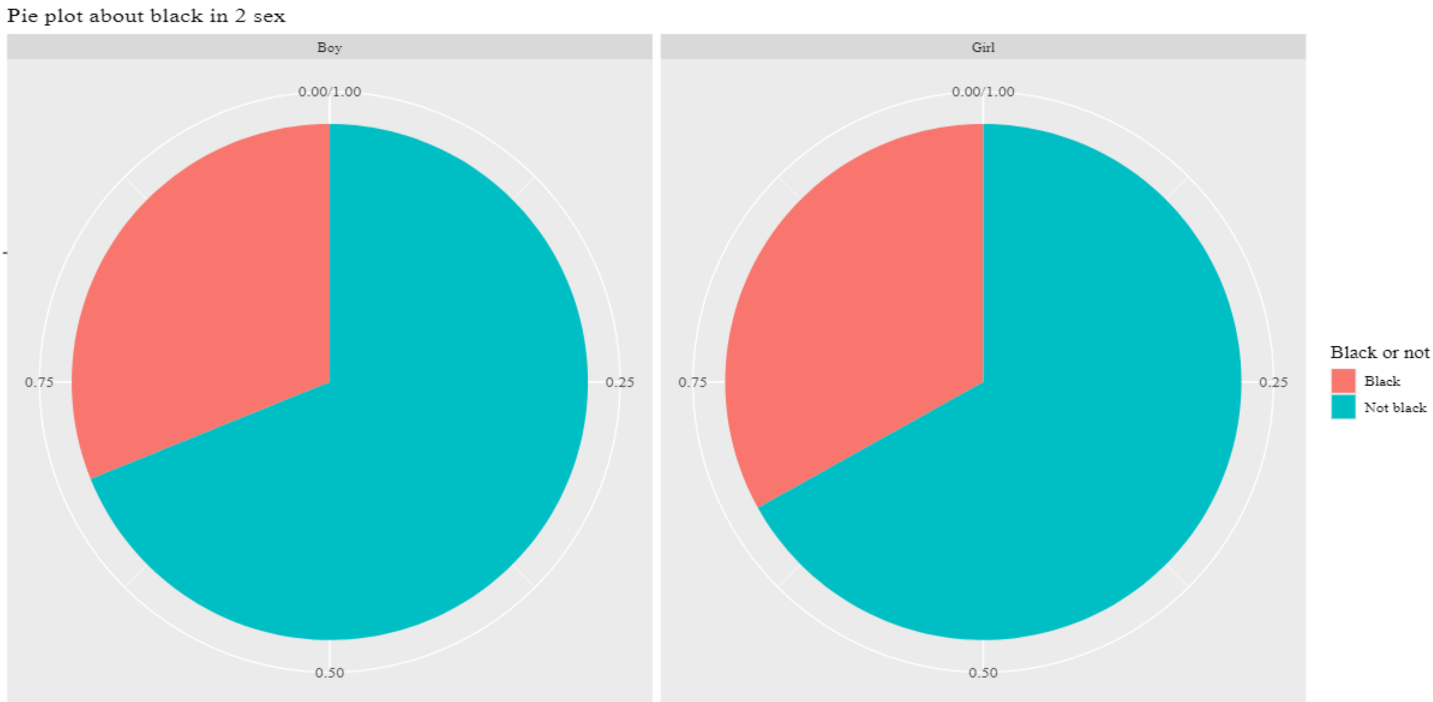
```
library(magrittr)
library(dplyr)
type.data <- data.frame(c = 1:length(boy))
type.black <- 1:length(black)
for (i in 1:length(black)) {
    if (black[i] == 0) {
        type.black[i] <- "Not black"
    } else {
        type.black[i] <- "Black"
    }
```

```
}
type.boy <- 1:length(boy)
for (i in 1:length(boy)) {
    if (boy[i] == 0) {
        type.boy[i] <- "Girl"
    } else {
        type.boy[i] <- "Boy"
    }
}
type.data$type.black <- type.black
blackvsblack <- type.data %>%
    group_by(type.black) %>%
    summarise(count = n()) %>%
    mutate(prec = count / sum(count))
type.data$type.boy <- type.boy
boyvsblack <- type.data %>%
group_by(type.boy,type.black) %>%
summarise(count=n()) %>%
mutate(prec=count/sum(count))
boyvsblack
scale_fill_discrete(name = "Black", labels = c("Black", "Not black"))
```

Then we plot 2 pie chart

```
ggplot(boyvsblack, aes(x="", y= prec, fill=type.black)) +
geom_bar(width = 2, stat = "identity") +
coord_polar("y", start=0) + facet_wrap(~type.boy,ncol = 2,scale =
"fixed")+
ggtitle("Pie plot about black in 2 sex")+
xlab("")+
ylab("")+
scale_fill_discrete(name = "Black or not", labels = c("Black","Not black"))
```

We have the chart below:



We use table function in R to create the statistics.

```
absentSexFreq=table(data$boy, data$black)
```

We have the result:

```
> absentSexFreq

     0    1
0 1883  932
1 2046  925
```

We use prop.test to check the propotion

```
prop.test(absentSexFreq, correct = FALSE, alternative = "less")
```

We have the result:

```
> prop.test(absentSexFreq,correct = FALSE, alternative = "less")

        2-sample test for equality of proportions without continuity
        correction

data:  absentSexFreq
X-squared = 2.5845, df = 1, p-value = 0.05396
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000000  0.0004611622
sample estimates:
   prop 1    prop 2
0.6689165 0.6886570
```

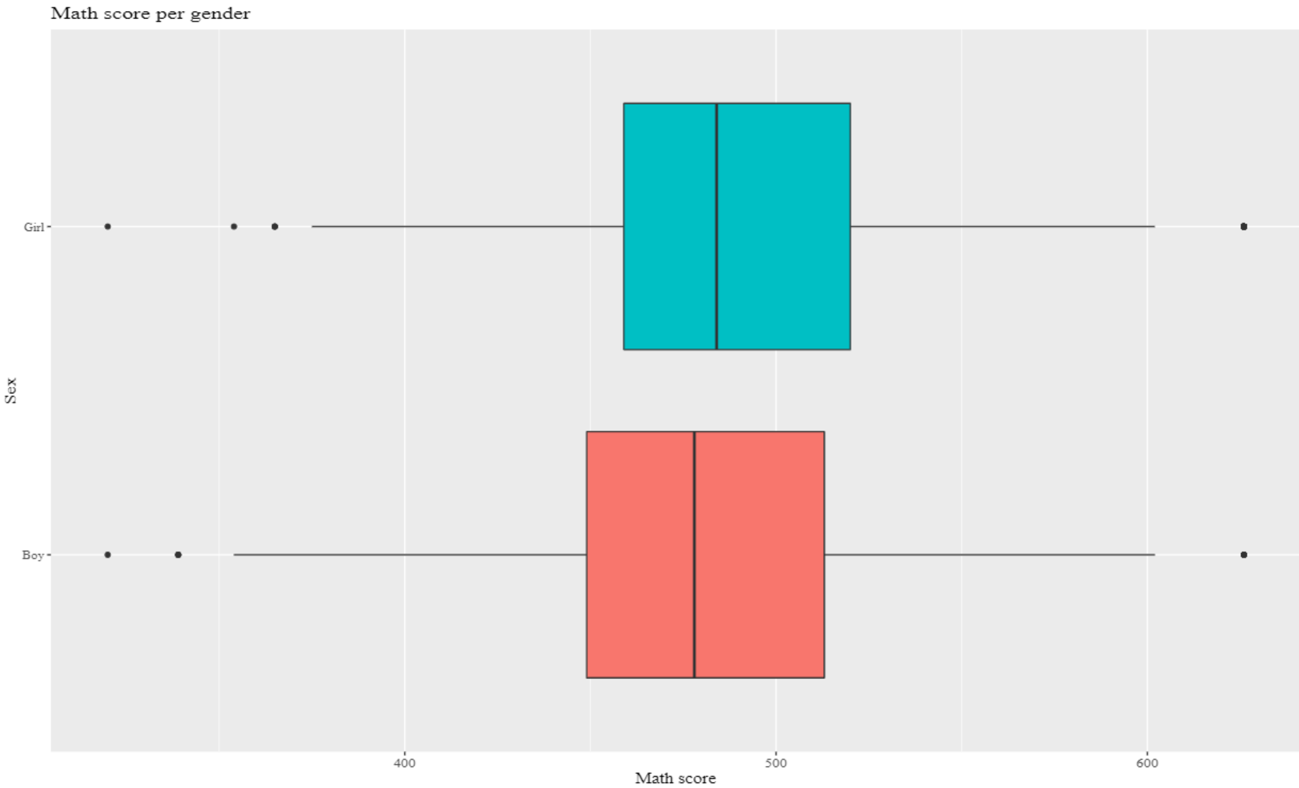
Because the p-value is more than α. So that the assumption above is incorrect

2.3 Math score

We have the box plot ablout the math score per gender using these code below:

```
ggplot(type.data, aes(x=mathscore, y = type.boy, fill = type.boy)) +
geom_boxplot() +
xlab("Math score") +
ylab("Sex") +
ggtitle("Math score per gender") +
theme(legend.position = "none")
```

After running these code, we have the boxplot below:



We will test the hypothesis that the average score of girls is more than the average score of boys
We run the code below:

```
t.test(data$mathscore~data$boy, alternative="greater")
```

We have the result that:

```
all arguments must have the same length
> t.test(data$mathscore~data$boy, alternative="greater")

        Welch Two Sample t-test

data:  data$mathscore by data$boy
t = 6.1224, df = 5767.2, p-value = 4.914e-10
alternative hypothesis: true difference in means between group 0 and group 1 is greater than 0
95 percent confidence interval:
 5.598971      Inf
sample estimates:
mean in group 0 mean in group 1
       489.5304        481.8741
```

With this result, p-value is grater than α, so that we can't reject the hypothesis

2.4 The master degree of teacher

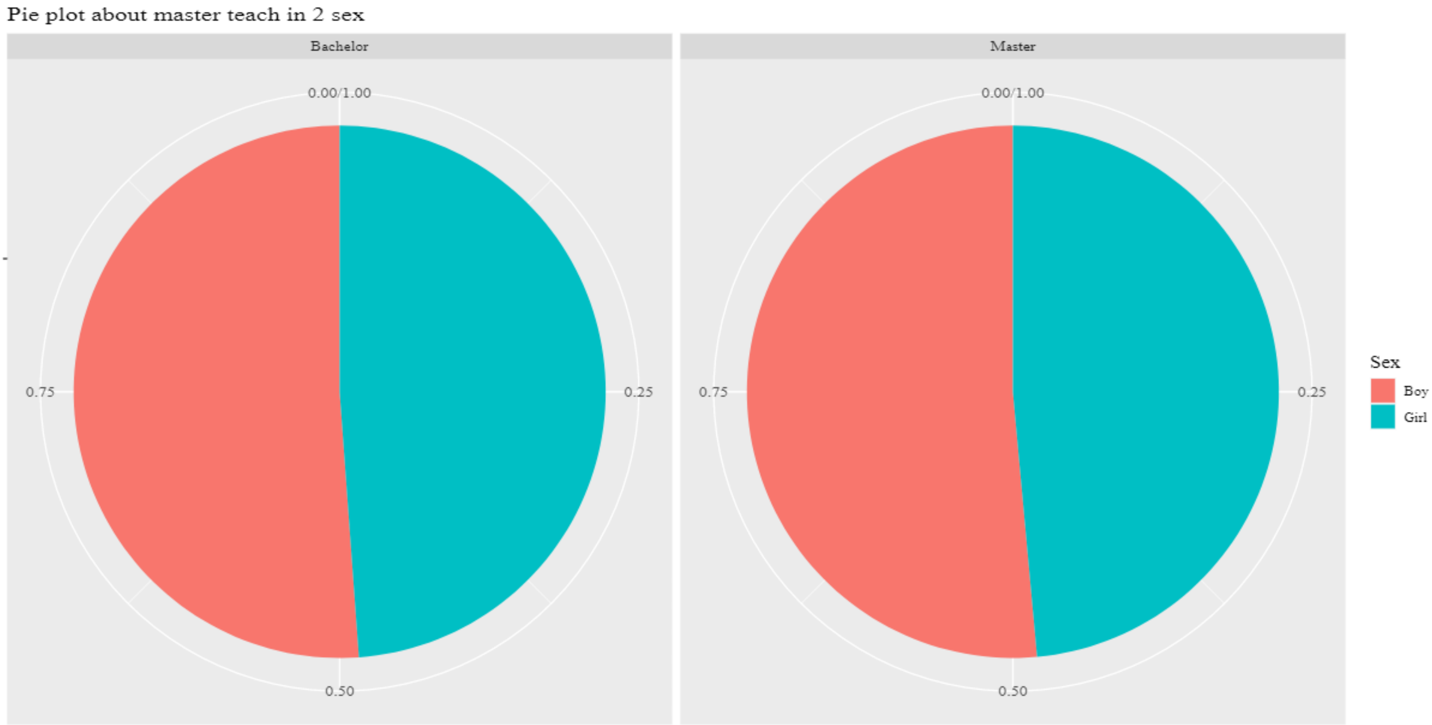
We assumpt that propotion that girl have been taught by master degree is less than male
First, we run these code to quantitive the variable

```
type.tchmasters<- 1:length(tchmasters)
for (i in 1:length(tchmasters)) {
    if (tchmasters[i] == 0) {
        type.tchmasters[i] <- "Master"
    } else {
        type.tchmasters[i] <- "Bachelor"
    }
}
type.boy <- 1:length(boy)
for (i in 1:length(boy)) {
    if (boy[i] == 0) {
        type.boy[i] <- "Girl"
    } else {
        type.boy[i] <- "Boy"
    }
}
type.data$type.tchmasters <- type.tchmasters
```

After that, we run the code to plot the pie graph:

```
boyvsboy <- type.data %>%
  group_by(type.boy) %>%
  summarise(count = n()) %>%
  mutate(prec = count / sum(count))
type.data$type.tchmasters <- type.tchmasters
tchmastersvsboy <- type.data %>%
group_by(type.tchmasters,type.boy) %>%
summarise(count=n()) %>%
mutate(prec=count/sum(count))
tchmastersvsboy
scale_fill_discrete(name = "Sex", labels = c("Boy", "Girl"))
ggplot(tchmastersvsboy, aes(x="", y= prec, fill=type.boy)) +
geom_bar(width = 2, stat = "identity") +
coord_polar("y", start=0) + facet_wrap(~type.tchmasters,ncol = 2,scale =
"fixed")+
ggtitle("Pie plot about master teach in 2 sex")+
xlab("")+
ylab("")+
scale_fill_discrete(name = "Sex", labels = c("Boy","Girl"))
```

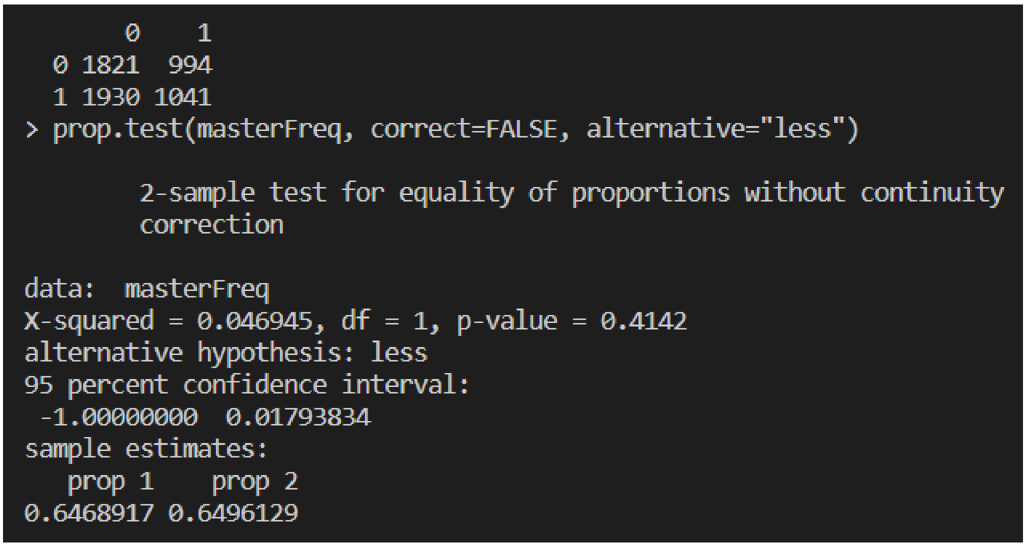
The result that we have a pie graph about the propotion of the master that teach the children in 2 sexes:



We run these code below to check the hypothesis:

```
masterFreq<-table(data$boy, data$tchmasters)
masterFreq
prop.test(masterFreq, correct=FALSE, alternative="less")
```

And then we have the result:



```
       0    1
  0 1821  994
  1 1930 1041
> prop.test(masterFreq, correct=FALSE, alternative="less")

	2-sample test for equality of proportions without continuity
	correction

data:  masterFreq
X-squared = 0.046945, df = 1, p-value = 0.4142
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.01793834
sample estimates:
   prop 1    prop 2 
0.6468917 0.6496129 
```

With that result, P_{value} is more than α, so that we reject this assumption

3 Regression

3.1 Simple Regression

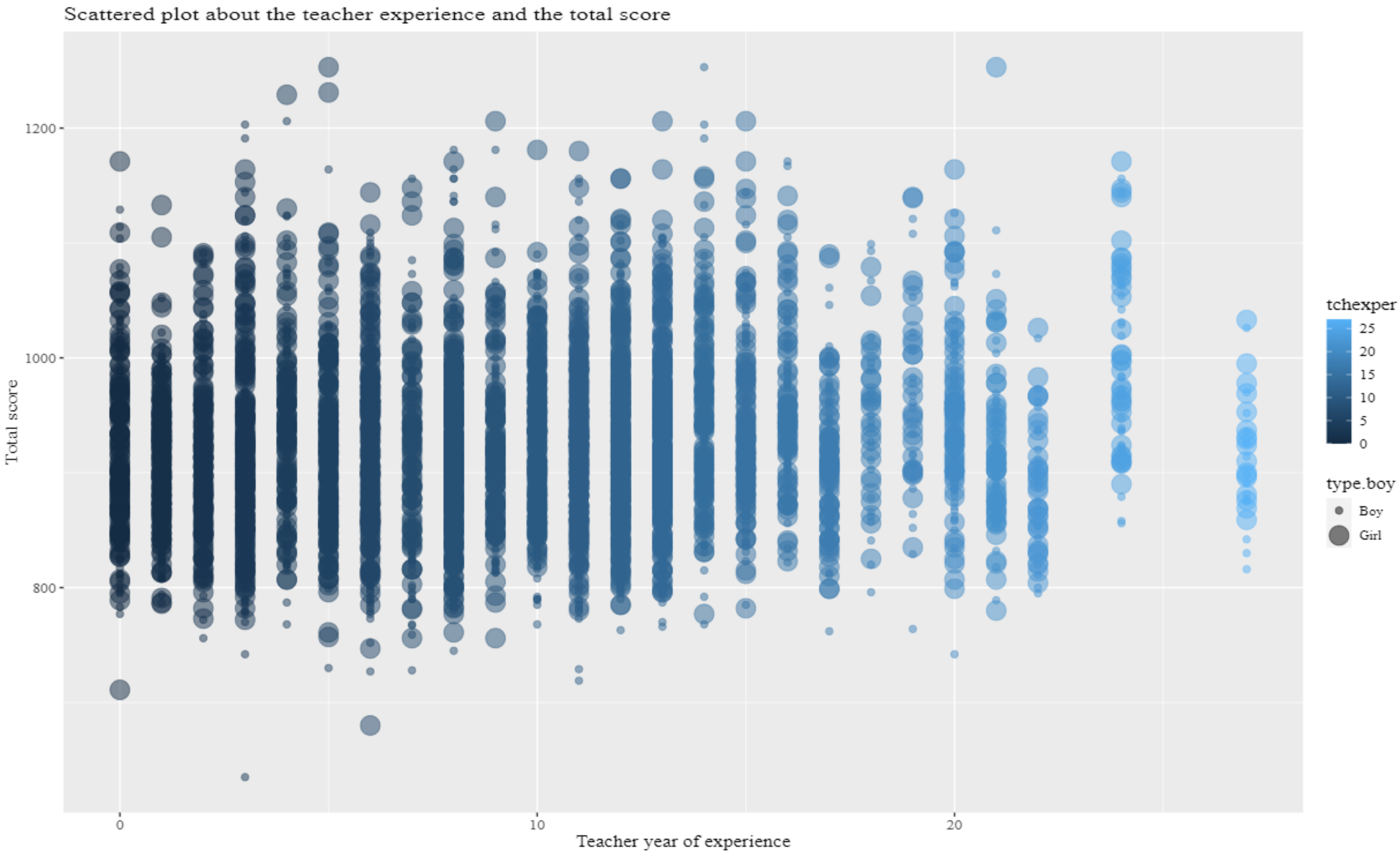
3.1.1 Build simple regression model by build total score over teacher experience

First, we can see the scattered plot to see that how 2 variables depend on together:
We run the code below:

```
ggplot(type.data,aes(tchexper,totalscore,color = tchexper,size = type.boy))+
geom_point(alpha = 0.5)+
```

```
xlab("Total score") +
ylab("Teacher year of experience") +
ggtitle("Scattered plot about the teacher experience and the total score")
```

Run the code, we have the graph:



We can see that the most point in this plot are betwwen 800-1050 points and the teacher experience is about 0-20 years
Between experienced teacher and inexperienced teacher, we can see that their student points are almost distribute on the same range, but more experienced teacher have the peak point of the student's score higher
We will try to see that what total score relates to the teacher experience
We have the equation: totalscore=β_1+$\beta_2$age+ε
We run the code below:

```
model1 <- lm(formula(G1~G2))
summary(model1)
```

Then we have the result:

```
> summary(model1)

Call:
lm(formula = formula(G1 ~ G2))

Residuals:
    Min      1Q  Median      3Q     Max 
-278.41  -51.52   -7.47   41.90  336.74 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 909.1294     1.8342 495.647   <2e-16 ***
G2            1.4264     0.1675   8.514   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73.37 on 5764 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.01242,	Adjusted R-squared:  0.01225 
F-statistic: 72.49 on 1 and 5764 DF,  p-value: < 2.2e-16
```

We have the estimated equation for model1 is:

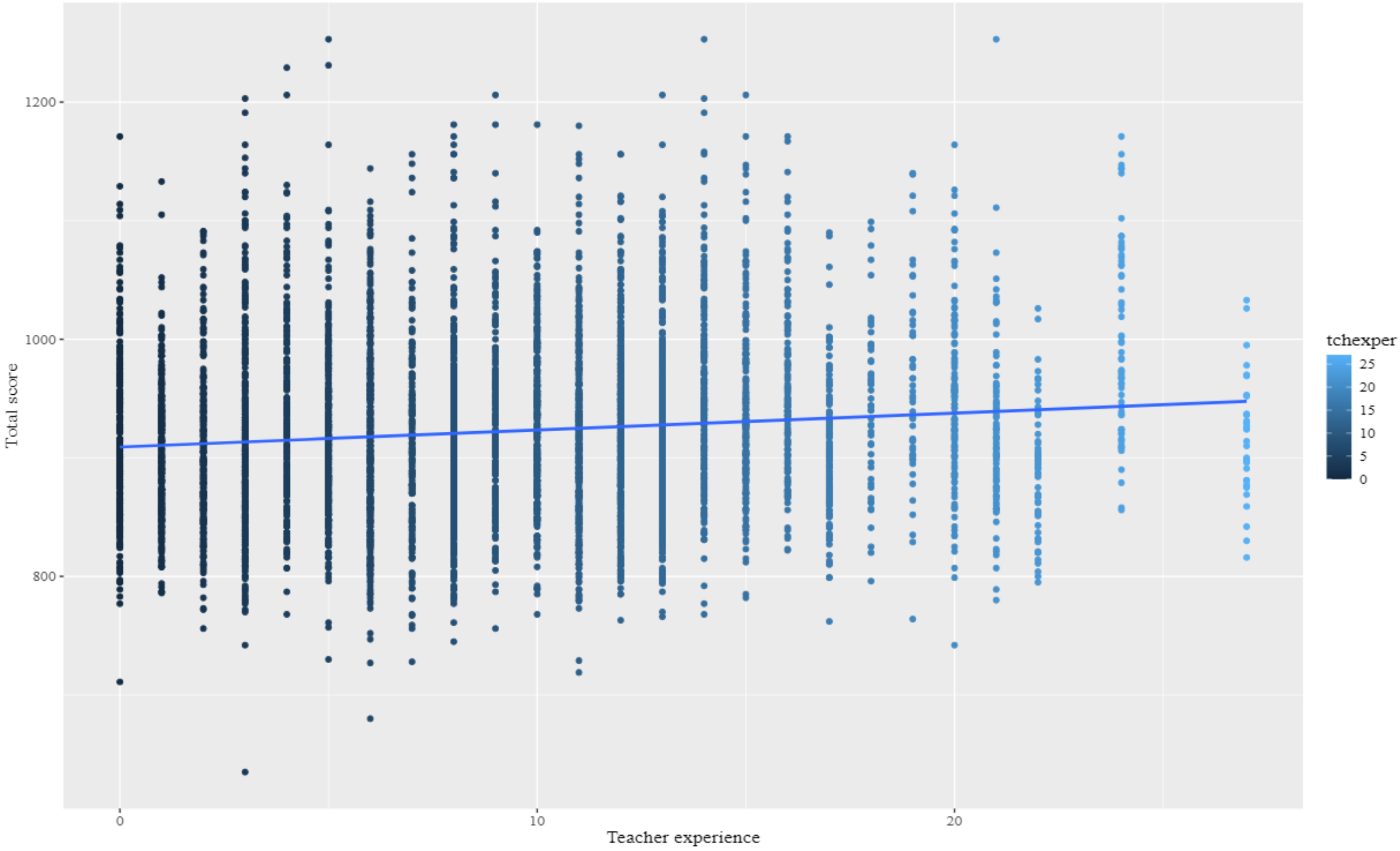
$$\widehat{totalscore} = 909.1294 + 1.4264tchexper$$

With (se) is 1.8342 and 0.1675

We use confint function to estimate 95% confident interval for the coeffiecient
With $\beta_1 = 909.1294, \beta_2 = 1.4264$
We plot the regression graph on the scattered plot using these code:

```
ggplot(type.data, aes(tchexper, totalscore, color = tchexper))+
geom_point(alpha = 1)+
geom_smooth(method = "lm", se = FALSE) +
ylab("Total score") +
xlab("Teacher experience")
```

Here is the result after running these code:



We run the code below:

```
confint(model1)
```

The result:

```
> confint(model1)
                2.5 %     97.5 %
(Intercept) 905.533620 912.725174
G2            1.097934   1.754785
```

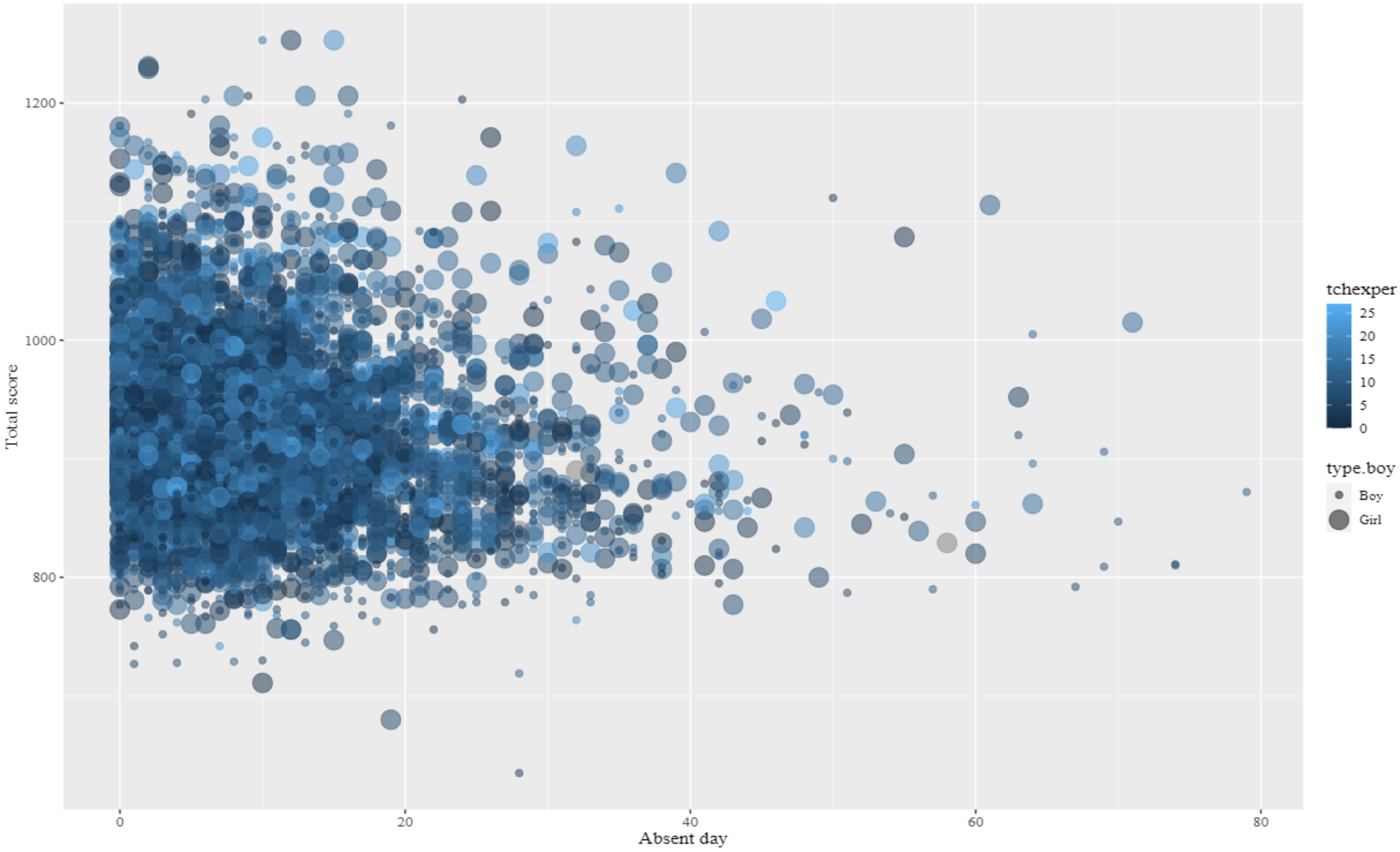
We have the conclusion for the 95% confident interval: $\begin{cases} \beta_1 \in (905.5, 912.7) \\ \beta_2 \in (1.1, 1.8) \end{cases}$

3.1.2 Build regression model from total score and the absent day

First, we have these code to plot the scattered plot about the absent day of the student and the total score

```
ggplot(type.data, aes(absent, totalscore, color = tchexper, size = type.boy))+
geom_point(alpha = 0.5)+
ylab("Total score") +
xlab("Absent day")
```

And than we have the plot after run these code:



We can see that the points distribute almost from 0-40 days
The points of the student that have more absent days are almost lower than the student have less absent day - we can see it significantly from the graphic
We have the base function: totalscore=β_1+$\beta_2$absent+ε
The code below helps we find the full function:

```
model2<-lm(formula(G1~G3))
summary(model2)
```

The result is:

```
lm(formula = formula(G1 ~ G3))

Residuals:
    Min      1Q  Median      3Q     Max 
-272.10  -51.40   -8.02   41.54  334.72 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 931.1801     1.4428  645.38   <2e-16 ***
G3           -0.8601     0.1043   -8.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73.38 on 5763 degrees of freedom
  (21 observations deleted due to missingness)
Multiple R-squared:  0.01167,   Adjusted R-squared:  0.0115 
F-statistic: 68.06 on 1 and 5763 DF,  p-value: < 2.2e-16
```

The result shows that:

$$\begin{cases} \beta_1 = 931.1801\text{: show that the maximum total score is 931.1801 when the student go to school all the days} \\ \beta_2 - 0.8601\text{: show that when the student absent 1 day, the total score reduce by -0.8601} \end{cases}$$

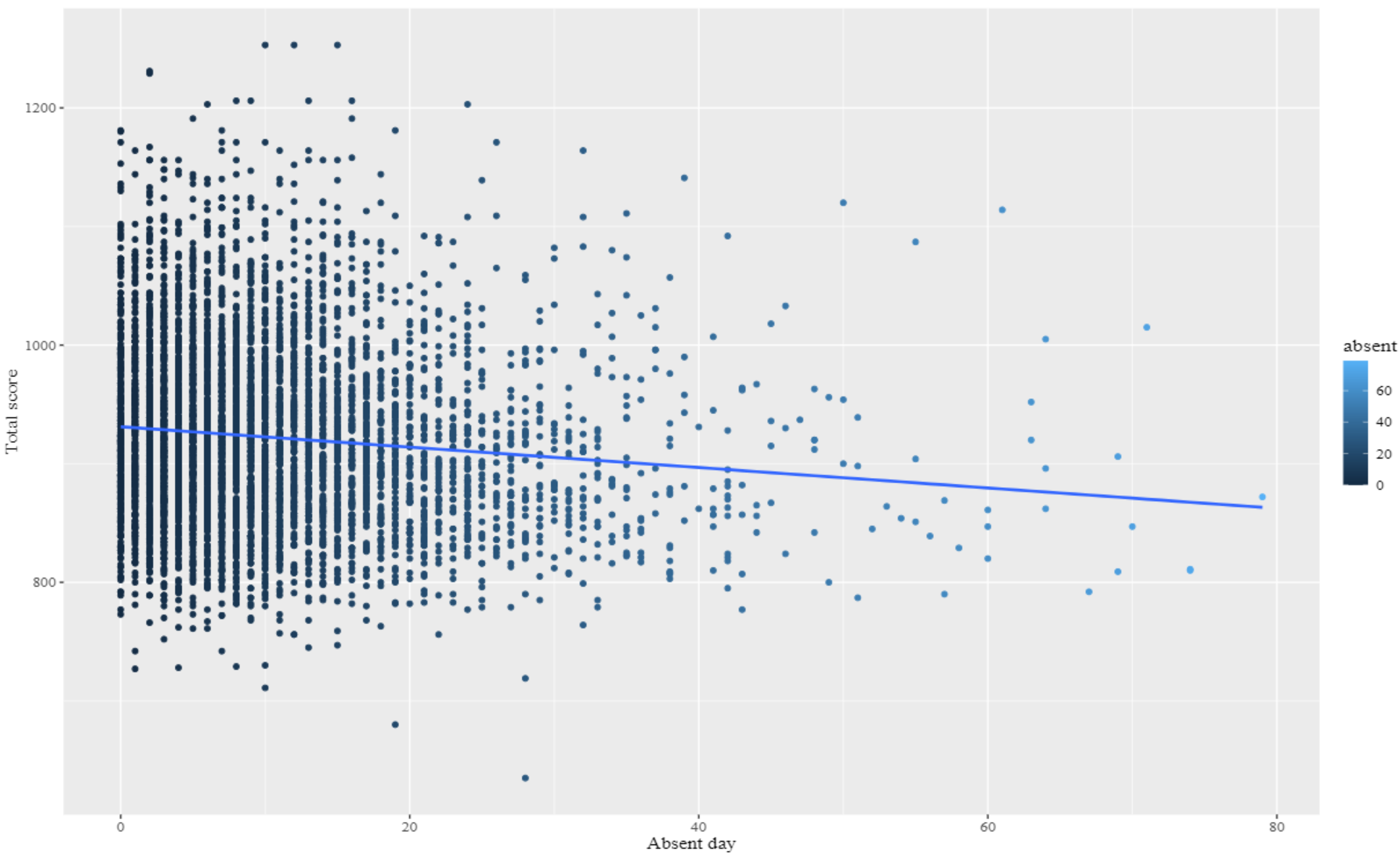
$$\begin{cases} \mathrm{se}(\beta_1) = 1.4428 \\ \mathrm{se}(\beta_2) = 0.1043 \end{cases}$$

$$\begin{cases} t_{\text{value}}\_1 = 645.38 \\ t_{\text{value}}\_2 = -8.25 \end{cases}$$

The estimated function is: totalscore=931.1801-0.8601*absent

We have the regression plot using the code below:

```
ggplot(type.data, aes(absent, totalscore, color = absent))+
geom_point(alpha = 1)+
geom_smooth(method = "lm", se = FALSE) +
ylab("Total score") +
xlab("Absent day")
```



Next, we will do the 95% confident interval of this model

The confint helps us:

```
confint(model2)
```

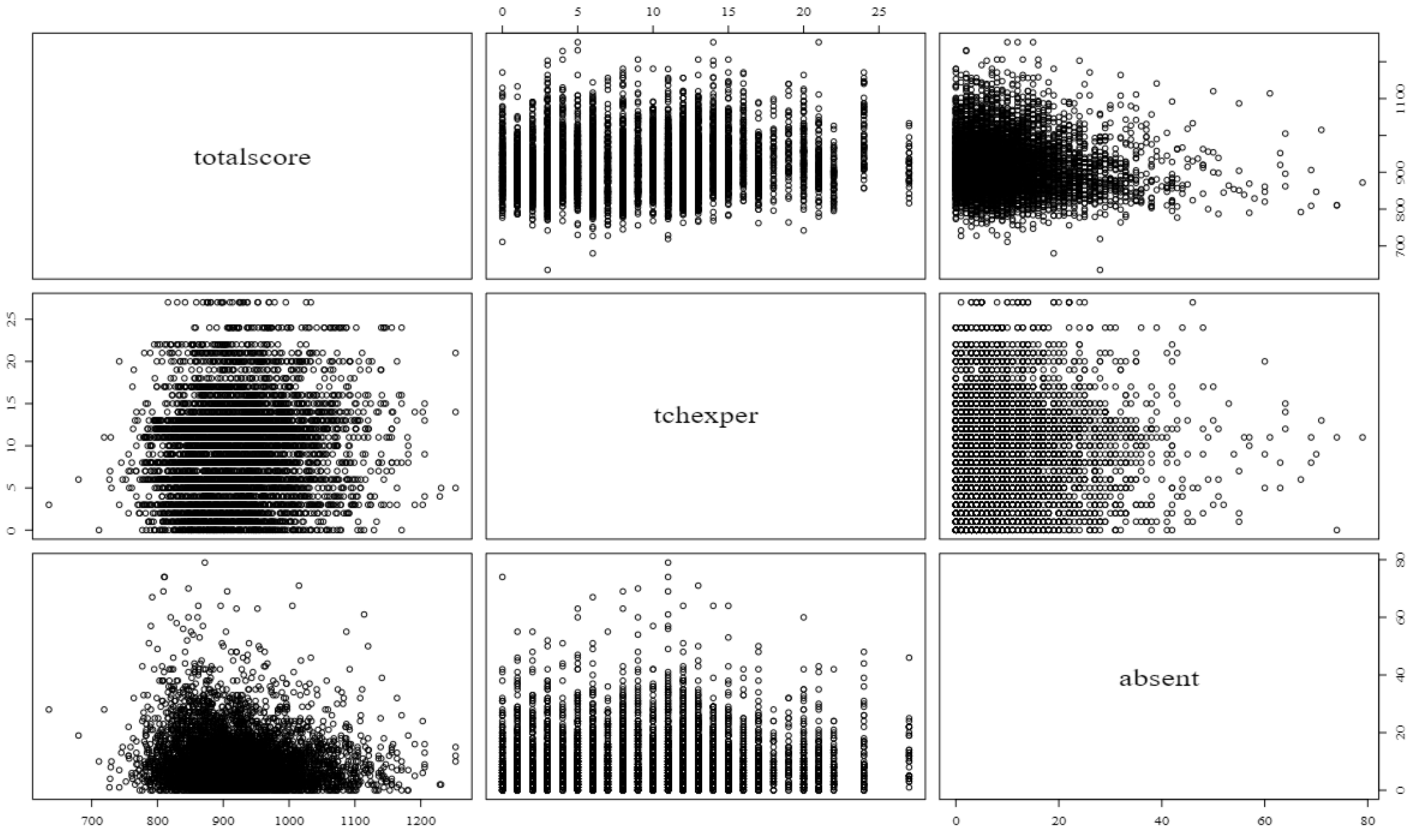
And the confident interval are: $\begin{cases} \beta_1 \in (928.35, 934) \\ \beta_2 \in (-1.06, 0.66) \end{cases}$

3.2 Multiple regression

We have $R_1^2 = 5383.1569$ and $R_2^2 = 5384.6244$ So that the 2 varible have the sam affect to the total score.
We will build the regression model between the totalscore, absent and tchexper
We have this plot by this code:

```
pairs(totalscore~tchexper + absent)
```

And this is the result:



The regression function bases on: totalscore=$\beta_1 + \beta_2$*tchexper+β_3*absent
We run the code below:

```
G1<-data$totalscore
G2<-data$tchexper
G3<-data$absent
model4 <- lm(formula(G1~G2+G3))
summary(model4)
```

After we run the code, we have the result:

```
> G1<-data$totalscore
G2<-data$tchexper
G3<-data$absent
model4 <- lm(formula(G1~G2+G3))
summary(model4)
> G2<-data$tchexper
> G3<-data$absent
> model4 <- lm(formula(G1~G2+G3))
> summary(model4)

Call:
lm(formula = formula(G1 ~ G2 + G3))

Residuals:
    Min      1Q  Median      3Q     Max 
-262.87  -50.84   -8.44   41.30  338.32 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 917.9147     2.1030 436.476   <2e-16 ***
G2            1.4427     0.1671   8.635   <2e-16 ***
G3           -0.8706     0.1041  -8.362   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73 on 5742 degrees of freedom
  (41 observations deleted due to missingness)
Multiple R-squared:  0.02408,	Adjusted R-squared:  0.02374 
F-statistic: 70.85 on 2 and 5742 DF,  p-value: < 2.2e-16
```

Based on the figure, we can see that:

All the coeffiecient of the function: $\begin{cases} \beta_1 = 917.9147 \\ \beta_2 = 1.4427 \\ \beta_3 = -0.8706 \end{cases}$ and $\begin{cases} \text{se}(\beta_1) = 2.1030 \\ \text{se}(\beta_2) = 0.1671 \\ \text{se}(\beta_3) = 0.1041 \end{cases}$ and $\begin{cases} t_{\text{value}}\_1 = 436.476 \\ t_{\text{value}}\_2 = 8.635 \\ t_{\text{value}}\_2 = -8.362 \end{cases}$

I explain some about the coefficient of the function model:

$\begin{cases} \beta_1\text{: when tchexper and absent equal to 0 (new teacher and no absent student), the score is 917.2} \\ \beta_2\text{: when the teacher has one more year of experience, the total score raises by 0.1671} \\ \beta_3\text{: when the student has more 1 day off, the scoer will decrease by 0.1041} \end{cases}$

So the function is: totalscore=$\beta_1 + \beta_2$*tchexper+β_3*absent

We find the 95% confident interval of the coefficient: $\begin{cases} \beta_1 (913.792035, 922.0374330) \\ \beta_2 : (1.115199, 1.7702482) \\ \beta_3 : (-1.074676, -0.6664686) \end{cases}$

4 Test the corelation of 2 pairs of qualitative variable

4.1 About the sex and the freelunch

We assume that the 2 variable is independent, using $\alpha = 0.05$
We run these code below:

```
tb4<-table(data$boy, data$freelunch)
tb4
chisq.test(tb4, correct = FALSE)
```

First, we have the table of 2 variable:

```
> tb4
   
       0    1
  0 1442 1373
  1 1557 1414
```

Secondly, we have the result of the chi-square test:

```
> chisq.test(tb4, correct = FALSE)

        Pearson's Chi-squared test

data:  tb4
X-squared = 0.80753, df = 1, p-value = 0.3689
```

Because the $p_{\text{value}} > \alpha$, so we accept H_0. The 2 variable is independent.

4.2 About the tchmasters and tchwhite

We assume that the 2 variable is independent, using $\alpha = 0.05$
We run these code below:

```
tb5 <- table(data$tchmasters, data$tchwhite)
tb5
chisq.test(tb5, correct = FALSE)
```

We have the table data and the result of the chi-square test:

```
> tb5

       0    1
  0  799 2952
  1  153 1882
> chisq.test(tb5, correct = FALSE)

        Pearson's Chi-squared test

data:  tb5
X-squared = 182.31, df = 1, p-value < 2.2e-16
```

Following the result: the $p_{\text{value}} > \alpha$
So we accept the hypothesis that the 2 variables are independent.