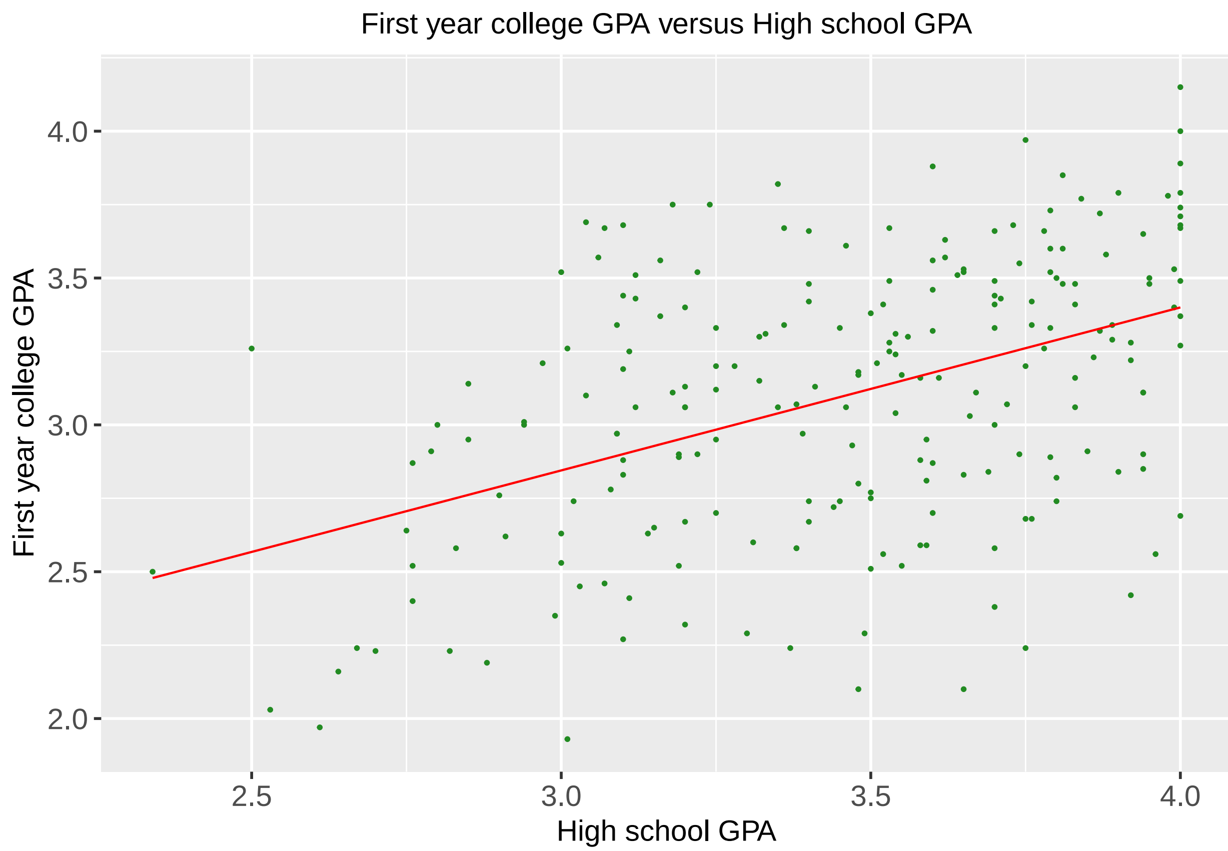
Will students who have high GPA in high school get high GPA in the first year college?

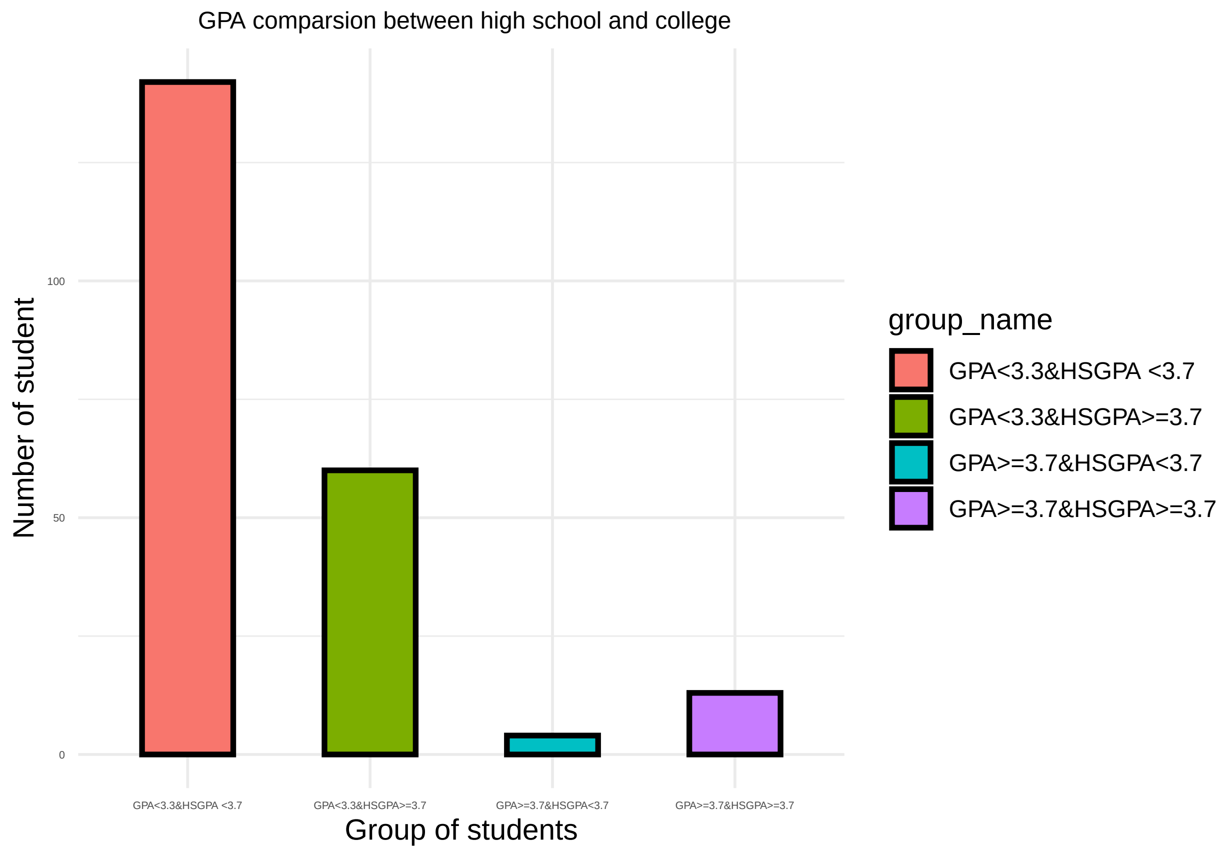
Interpretation and Implementation

```
#primary objective
#set up
#install.packages("ggplot2")
library(ggplot2)

#load the data
student_info<-read.csv("FirstYearGPA.csv",header=T)

#set up variables
#set of high school GPA of students
x<-student_info$HSGPA
#set of first year college GPA of students
y<-student_info$GPA

#student whose GPA scores above or equal 3.7 in high school
HSGPA_greater_than_critical_value<-student_info[student_info$HSGPA>=3.7,]


#student whose GPA scores below 3.7 in high school
HSGPA_less_than_critical_value<-student_info[student_info$HSGPA<3.7,]


#college GPA>=3.7 and high school GPA >=3.7
group_A<-student_info[(student_info$HSGPA>=3.7&student_info$GPA>=3.7),]

#college GPA>3.7 and high school GPA <3.7
group_B<-student_info[student_info$HSGPA<3.7&student_info$GPA>=3.7,]

#college GPA<3.7 and high school GPA >=3.7
group_C<-student_info[student_info$HSGPA>=3.7&student_info$GPA<3.7,]

#college GPA<3.7 and high school GPA <3.7
group_D<-student_info[student_info$HSGPA<3.7&student_info$GPA<3.7,]


#Graph plotting
#scatter plot
graph_GPA<-ggplot(student_info,aes(x=HSGPA,y=GPA))+geom_point(size=0.35,color="forestgreen")+geom_smoot

graph_GPA
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#barchart
#set up
group_name<-c("GPA>=3.7&HSGPA>=3.7","GPA>=3.7&HSGPA<3.7","GPA<3.3&HSGPA>=3.7","GPA<3.3&HSGPA <3.7")
group_size<-c(length(group_A$GPA),length(group_B$GPA),length(group_C$GPA),length(group_D$GPA))
group_data<-data.frame(group_name,group_size)



ggplot(data=group_data,aes(x=group_name,y=group_size,fill=group_name))+geom_bar(stat="identity",width=0
```

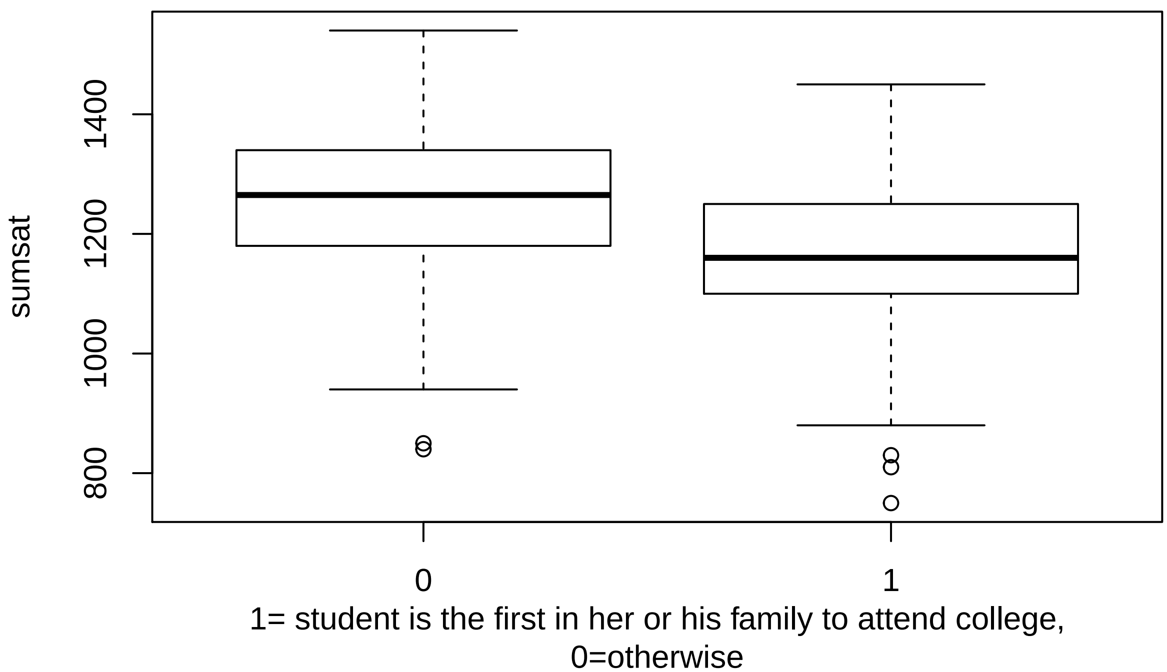



```
#Pearson's Correlation Coefficient

size_of_sample<-length(x)

#mean of sample X
mean_x<-mean(x)

#mean of sample Y
mean_y<-mean(y)

#calculation for Pearson's Correlation Coefficient according to the formula shown in the appendix of th
numerator<-sum((x-mean_x)*(y-mean_y))
denominator_1<-(sqrt(sum(((x-mean_x)^2))))
denominator_2<-(sqrt(sum(((y-mean_y)^2))))


r<-numerator/(denominator_1*denominator_2)

r
```

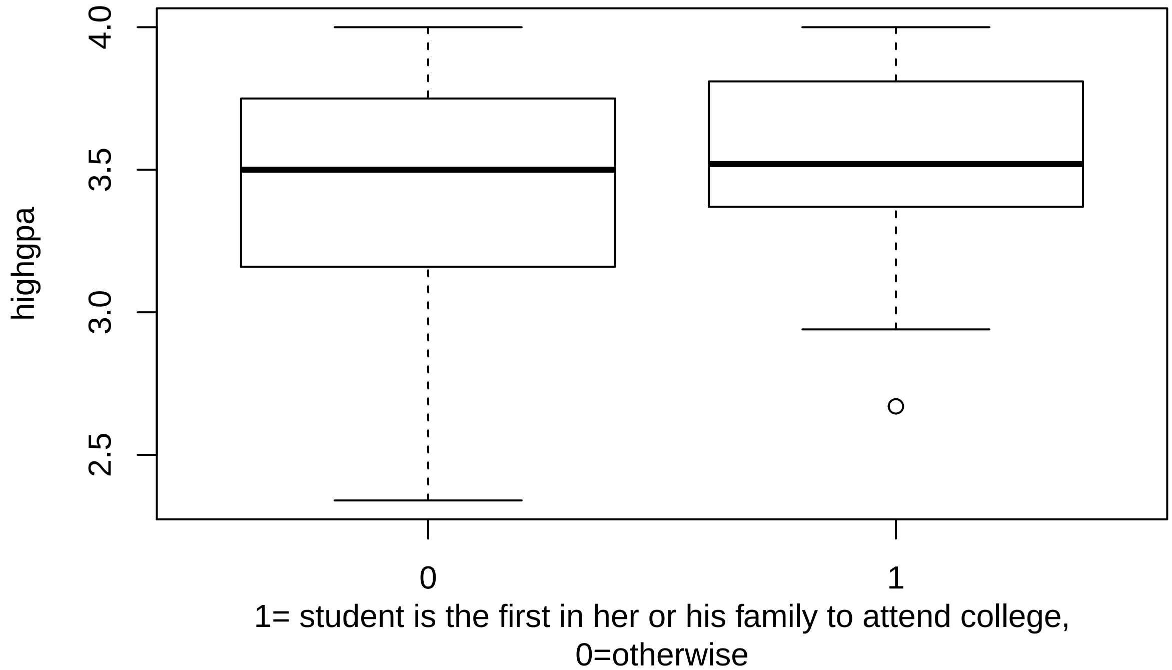
```
## [1] 0.4468873
```

```
#secondary objective
#set up
#load the data
data<- read.csv("FirstYearGPA.csv",header=T)
```

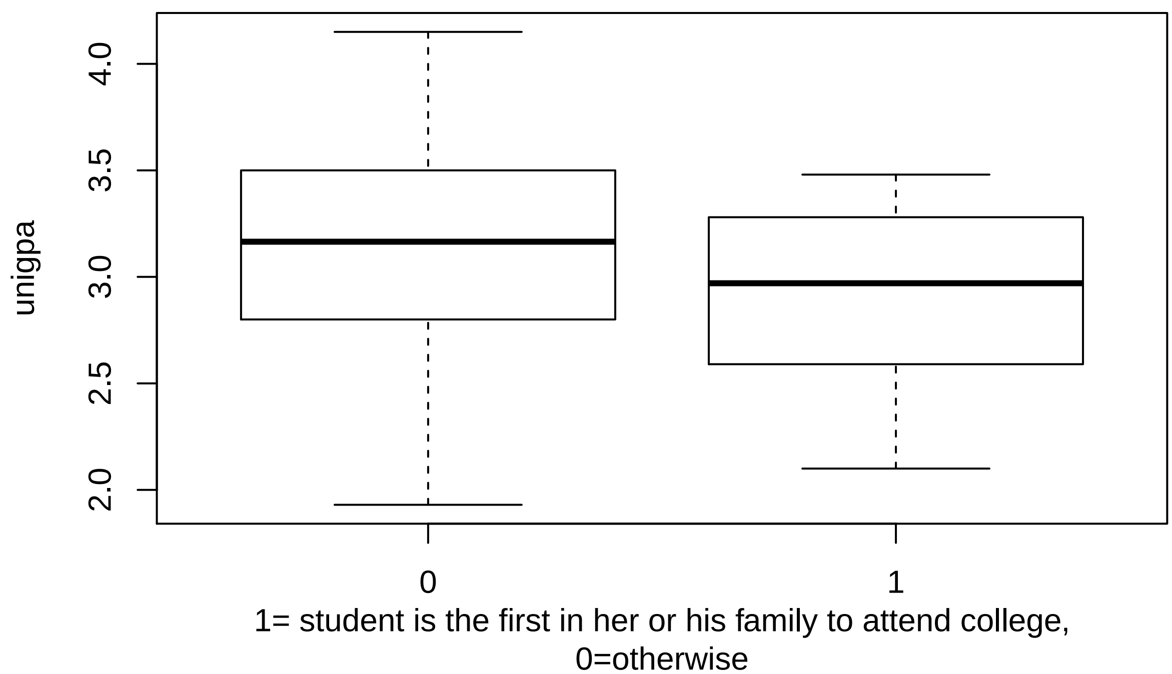
```r
#set up variables
satvnc_1<- data$SATV
satvnc_2<- data$SATM
#sumsat=sum of SAT score
sumsat<- satvnc_1 + satvnc_2
#highgpa=high school GPA
highgpa<- data$HSGPA
#unigpa=first year college GPA
unigpa<- data$GPA
#familyedu=family education level
familyedu<- data$FirstGen

#boxplot:family education level and sumsat
boxplot(sumsat~familyedu, xlab="1= student is the first in her or his family to attend college,
0=otherwise")
```




```r
#boxplot:family education level and high school GPA
boxplot(highgpa~familyedu,xlab="1= student is the first in her or his family to attend college,
0=otherwise")
```




```r
#boxplot:family education level and first year college GPA
boxplot(unigpa~familyedu,xlab="1= student is the first in her or his family to attend college,
0=otherwise")
```




```r
#correlation coefficient:family education level and sumsat
cor(familyedu,sumsat)
```

```
## [1] -0.250557
```

```r
#correlation coefficient:family education level and high school GPA
cor(familyedu,highgpa)
```

```
## [1] 0.06418575
```

```r
#correlation coefficient:family education level and first year college GPA
cor(familyedu,unigpa)
```

```
## [1] -0.1565773
```

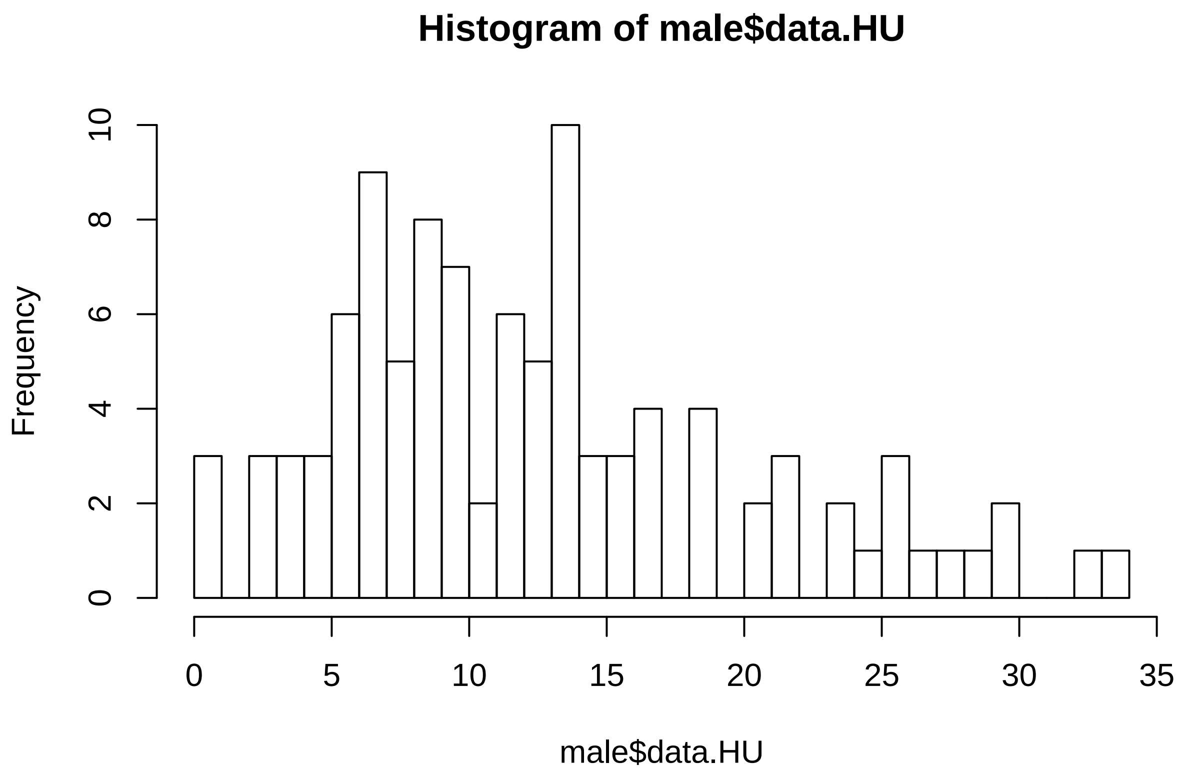
```r
#Tertiary objective
#read data set
data <- read.csv("FirstYearGPA.csv")

#extract the Male, HU, SS from the data set
d <- data.frame(data$Male, data$HU, data$SS)

#seperate set
m <- d$data.Male == 1
male <- d[m,]
female <- d[!m,]

#add gender
nrow(male)
```

```
## [1] 102
```

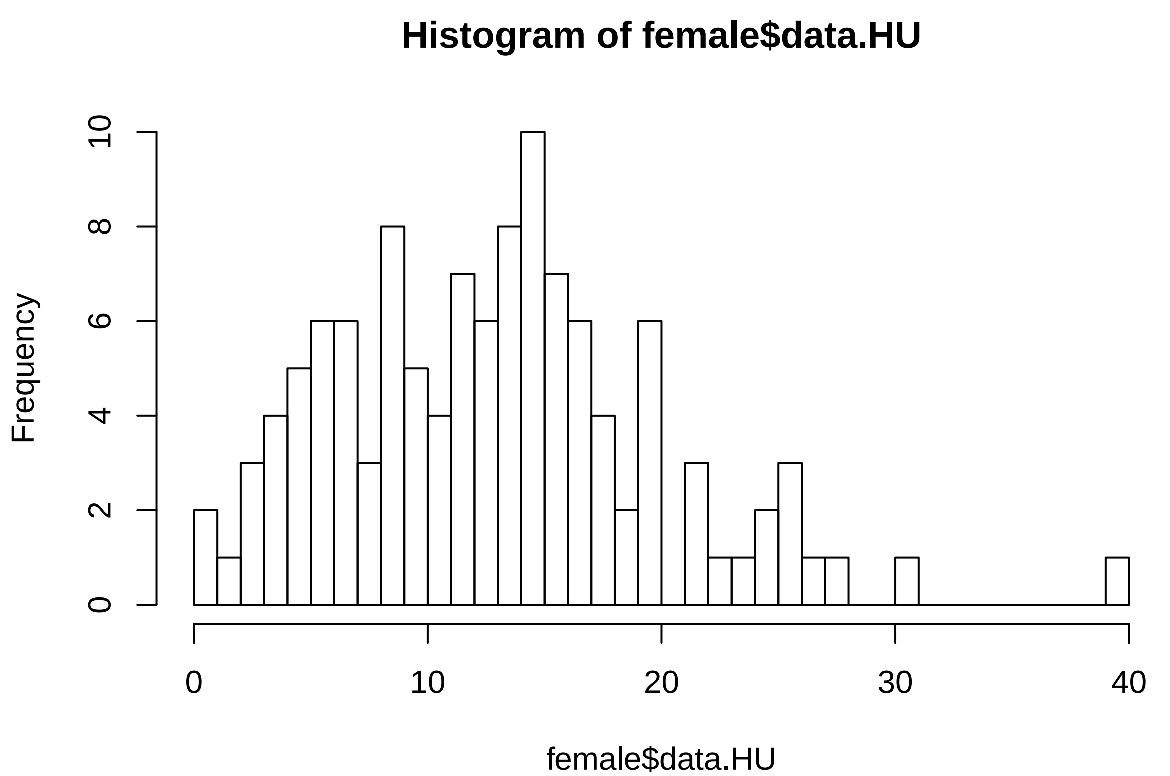
```r
nrow(female)
```

```
## [1] 117
```

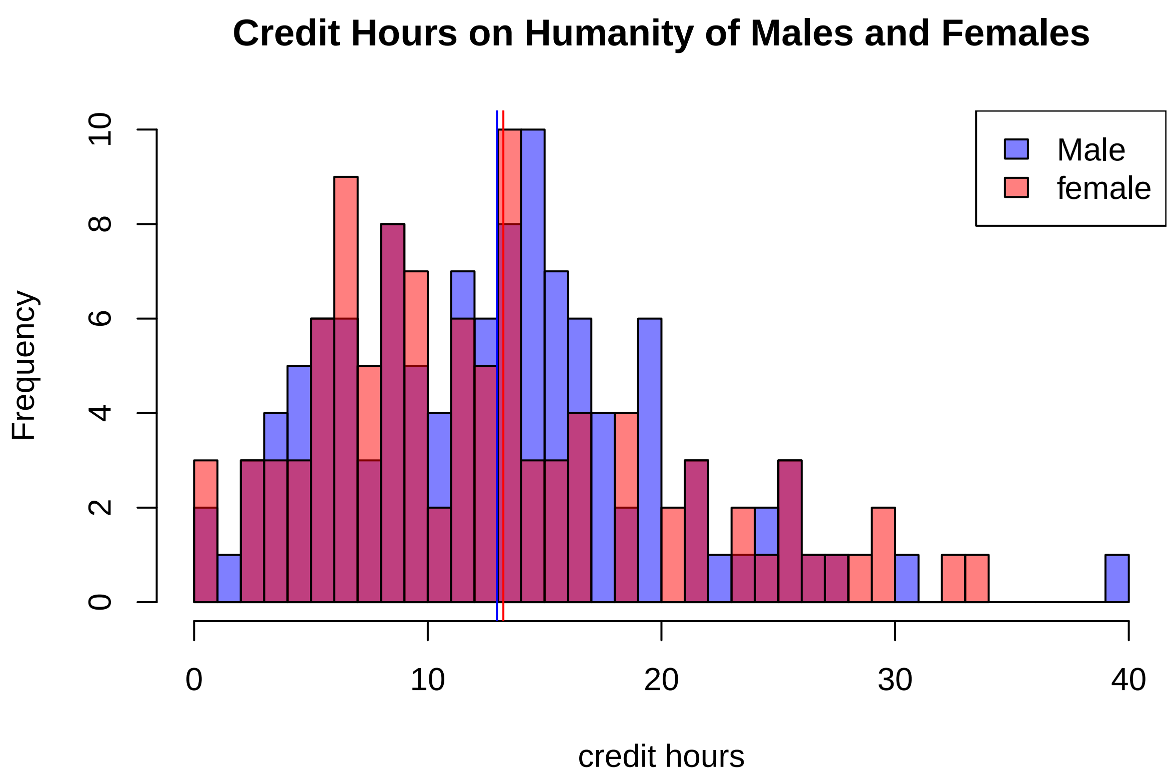
```r
gender <- c(rep("Male",102), rep("Female",117))
HU <- c(male$data.HU, female$data.HU)
SS <- c(male$data.SS, female$data.SS)
df <- data.frame(gender, HU, SS)
genderm <- c(rep("Male",102))
HUm <- c(male$data.HU)
SSm <- c(male$data.SS)
dfm <- data.frame(genderm, HUm, SSm)
genderf <- c(rep("Female",117))
HUf <- c(female$data.HU)
SSf <- c(female$data.SS)
dff <- data.frame(genderf, HUf, SSf)

#graph of gender and cradit on humanity course
p1 <- hist(male$data.HU, breaks = 40)
```



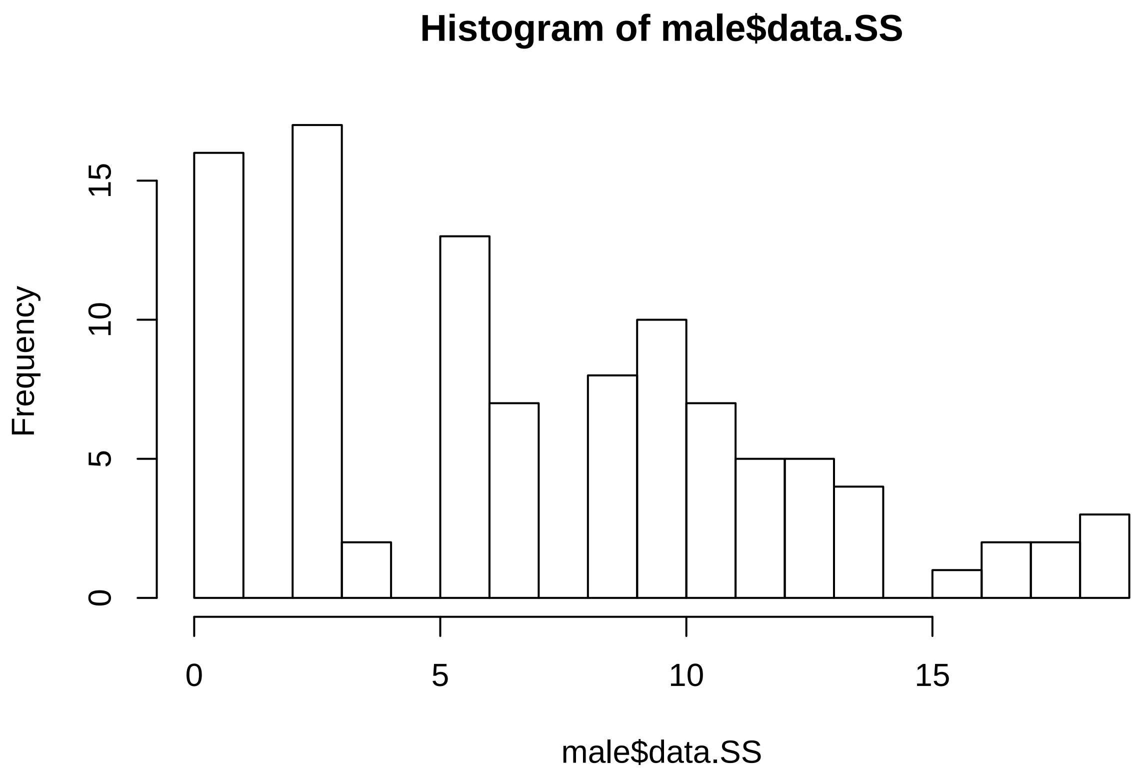

```r
p2 <- hist(female$data.HU, breaks = 40)
```



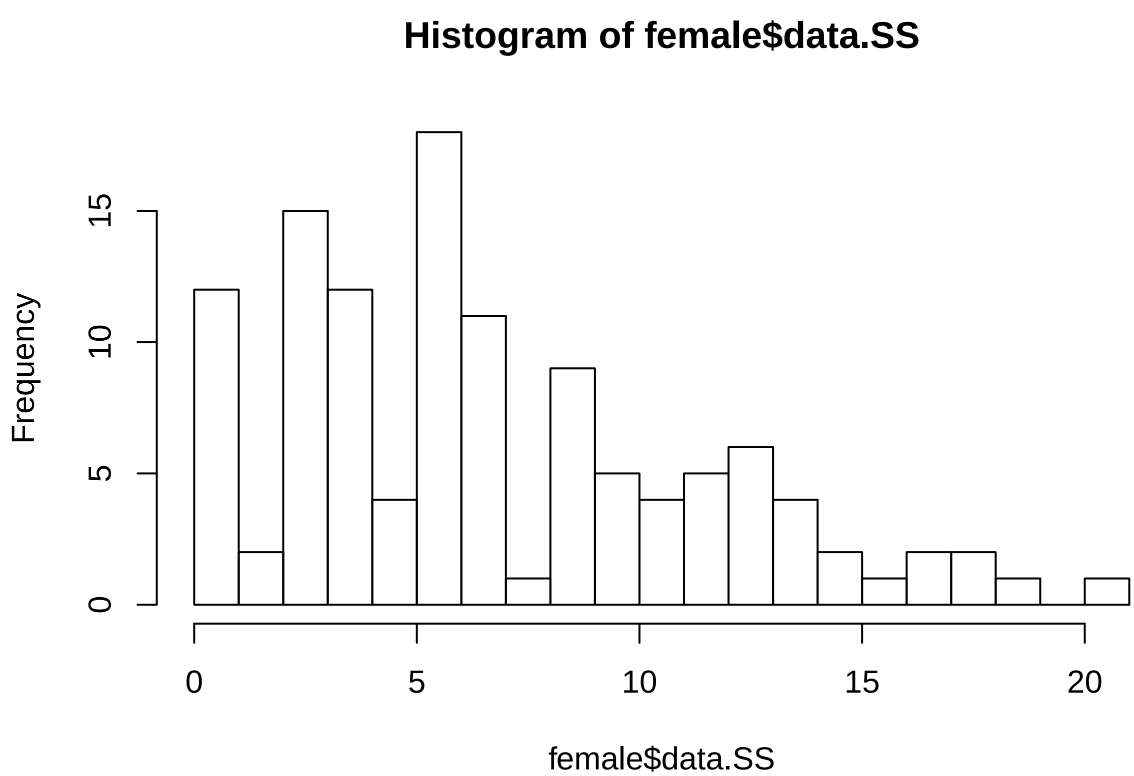
```r
plot(p2, col = rgb(0, 0, 1, 1/2), main="Credit Hours on Humanity of Males and Females", xlab="credit hou
plot(p1, col = rgb(1, 0, 0, 1/2), add = T)
abline(v = mean(male$data.HU), col="blue")
abline(v = mean(female$data.HU), col="red")
legend("topright", c("Male", "female"), fill = c(rgb(0, 0, 1, 1/2), rgb(1, 0, 0, 1/2)))
```



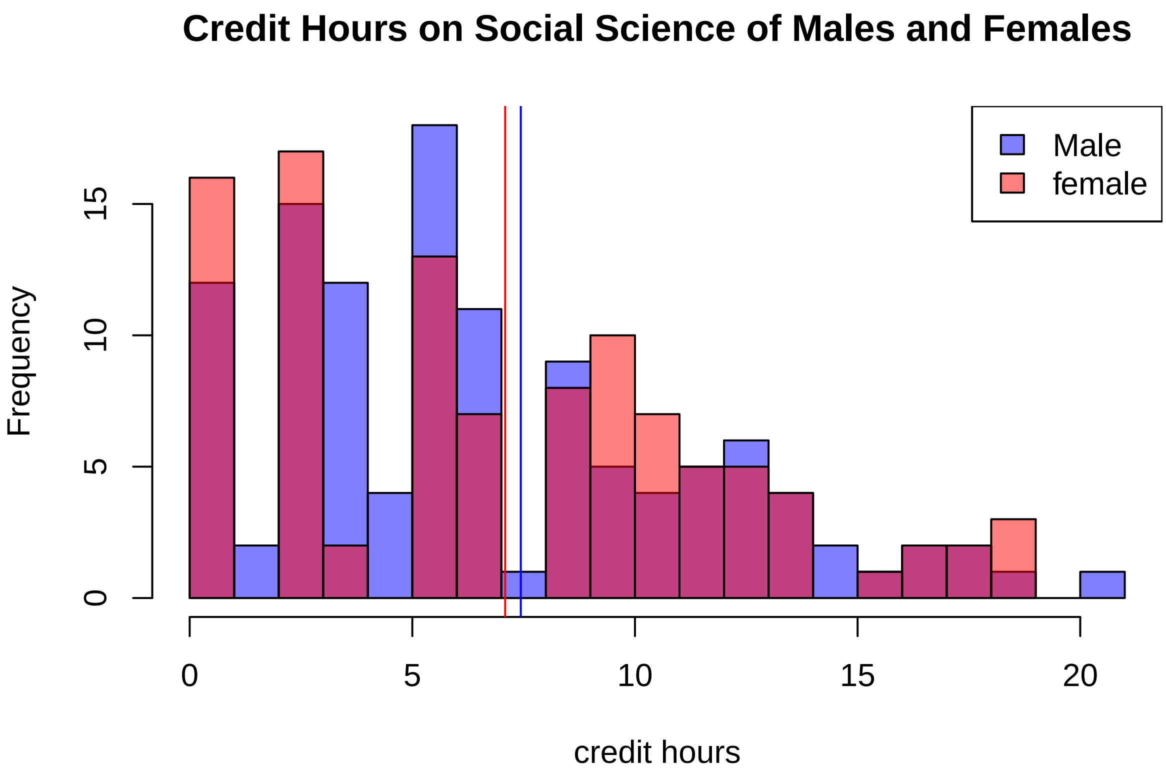

```r
#graph of Credit Hours on Social Science course
p1 <- hist(male$data.SS, breaks = 25)
```



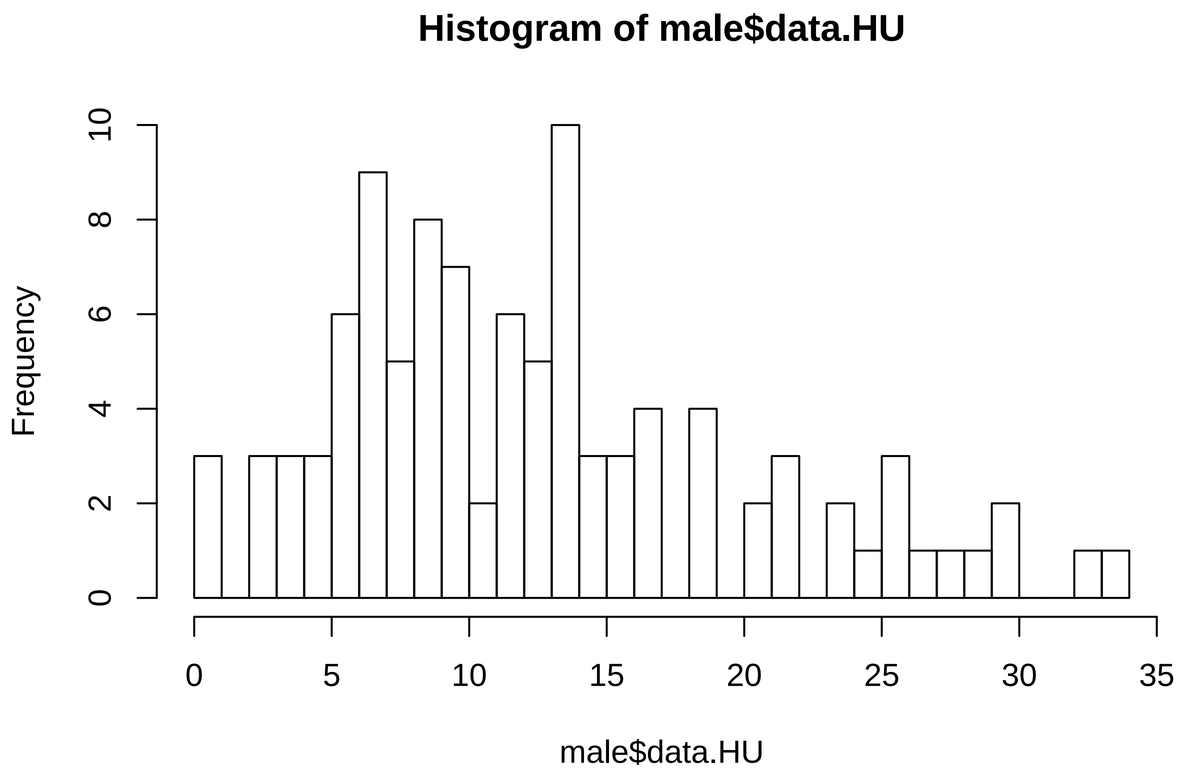

```
p2 <- hist(female$data.SS, breaks = 25)
```



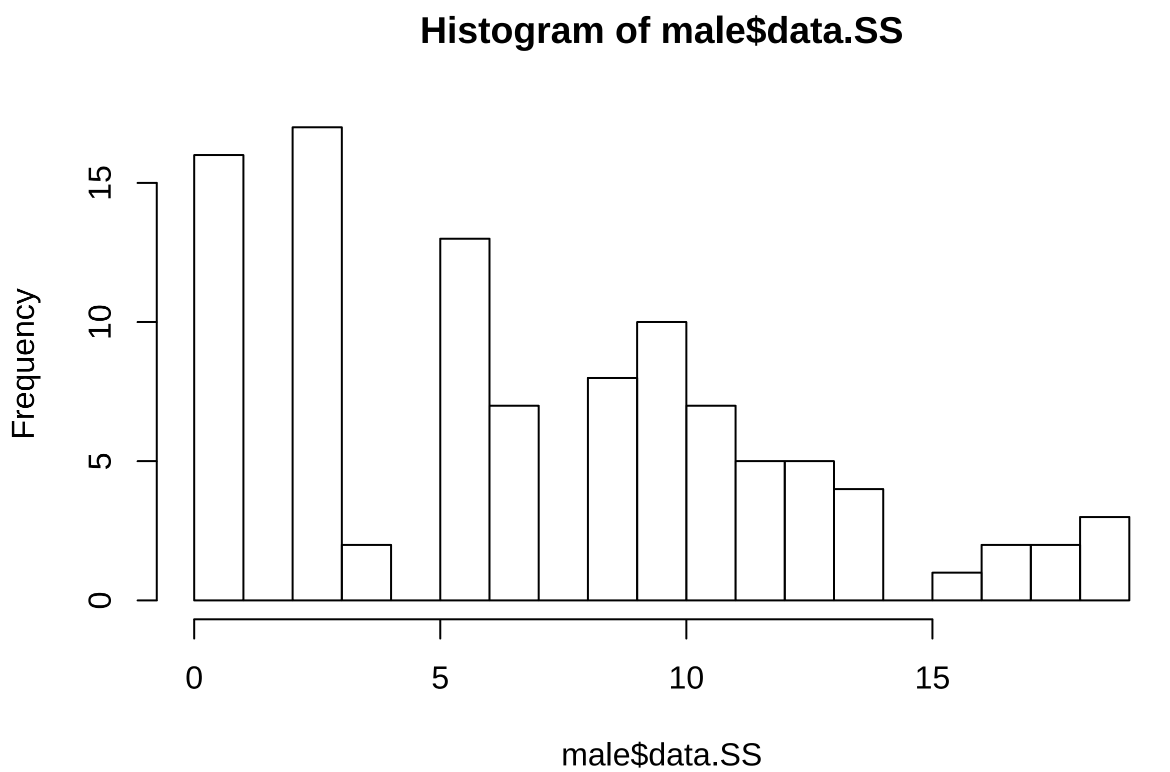
```r
plot(p2, col = rgb(0, 0, 1, 1/2), main="Credit Hours on Social Science of Males and Females", xlab="cre
plot(p1, col = rgb(1, 0, 0, 1/2), add = T)
abline(v = mean(male$data.SS), col="blue")
abline(v = mean(female$data.SS), col="red")
legend("topright", c("Male", "female"), fill = c(rgb(0, 0, 1, 1/2), rgb(1, 0, 0, 1/2)))
```



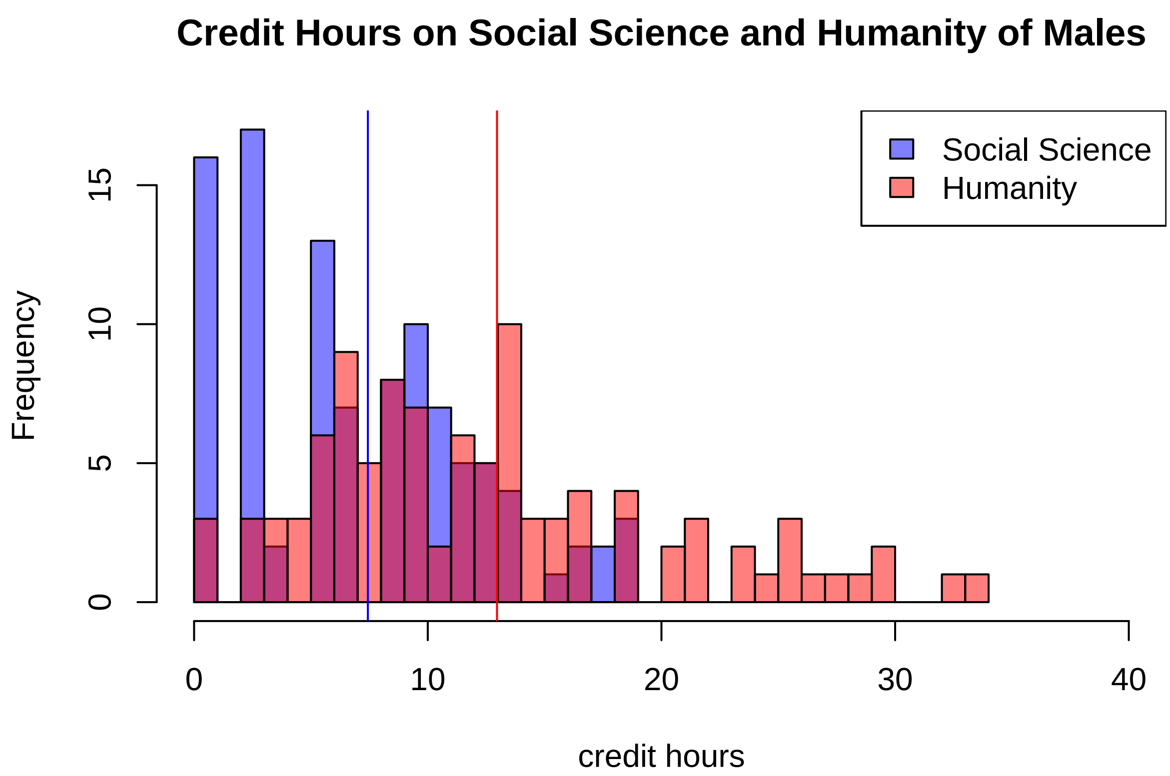

```r
#graph of Credit Hours on Social Science and Humanity of males
p1 <- hist(male$data.HU, breaks = 40)
```



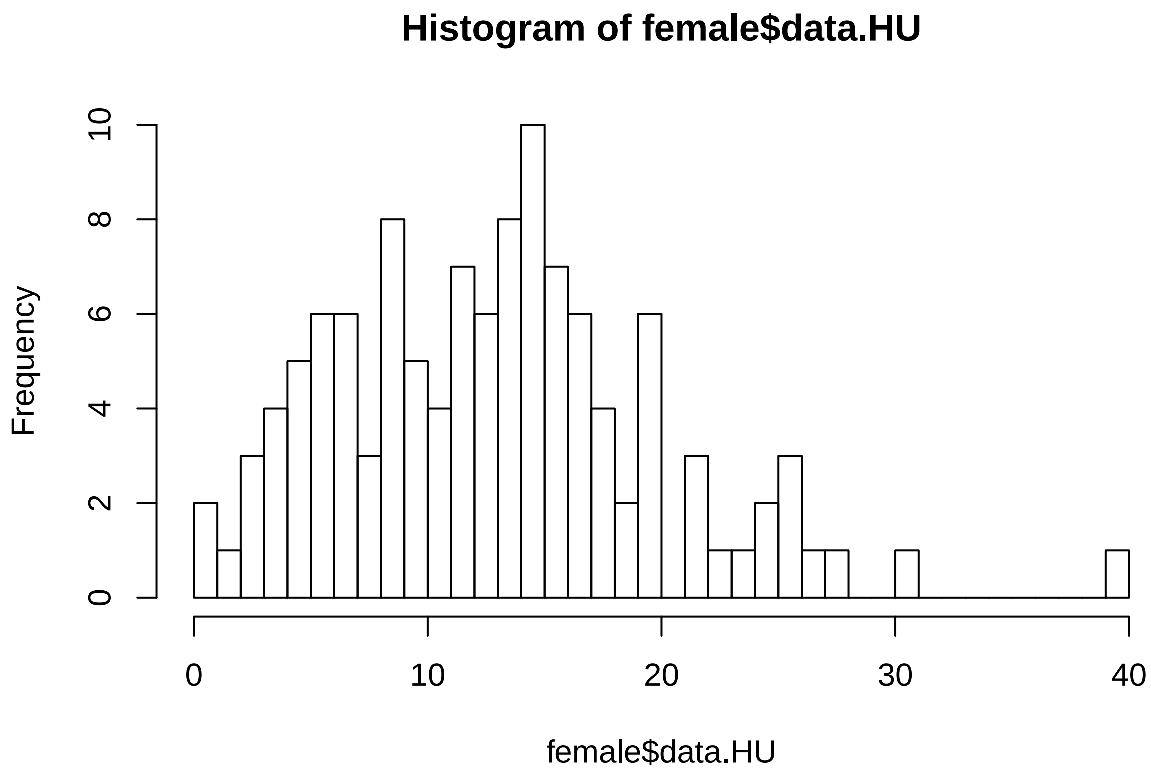

```r
p2 <- hist(male$data.SS, breaks = 20)
```



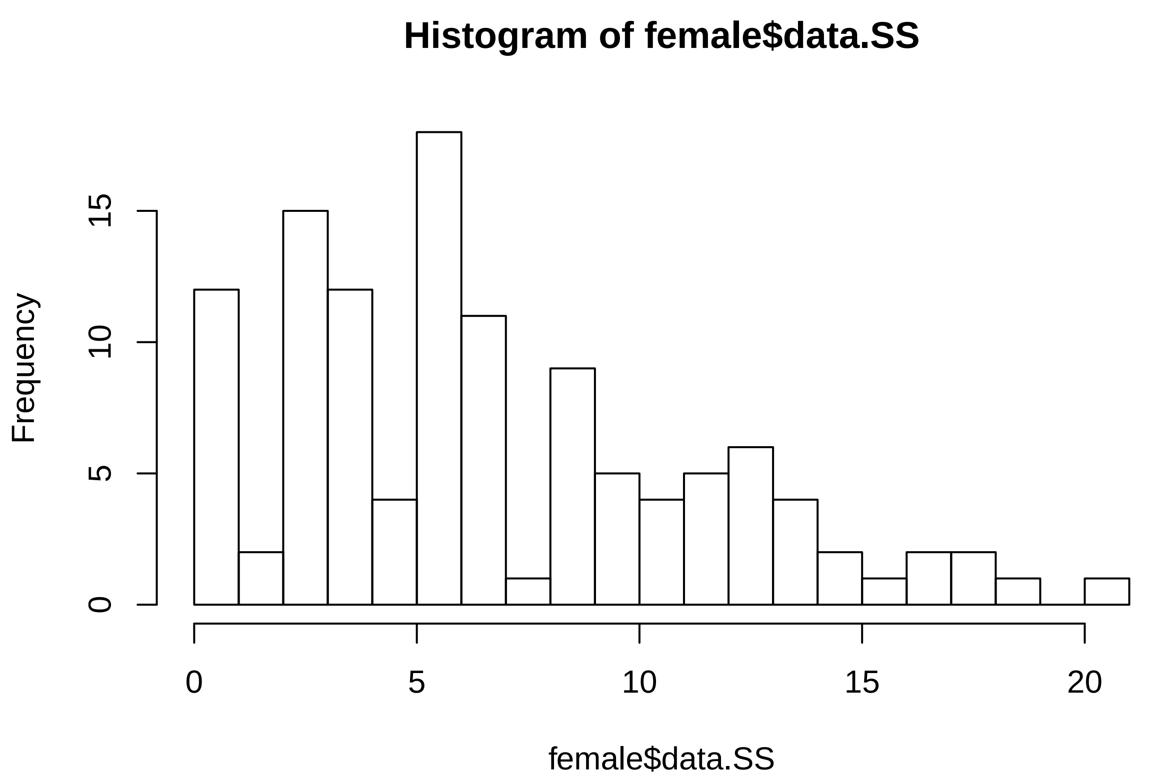
```r
plot(p2, col = rgb(0, 0, 1, 1/2), main="Credit Hours on Social Science and Humanity of Males", xlab="cre
plot(p1, col = rgb(1, 0, 0, 1/2), add = T)
abline(v = mean(male$data.SS), col="blue")
abline(v = mean(male$data.HU), col="red")
legend("topright", c("Social Science", "Humanity"), fill = c(rgb(0, 0, 1, 1/2), rgb(1, 0, 0, 1/2)))
```



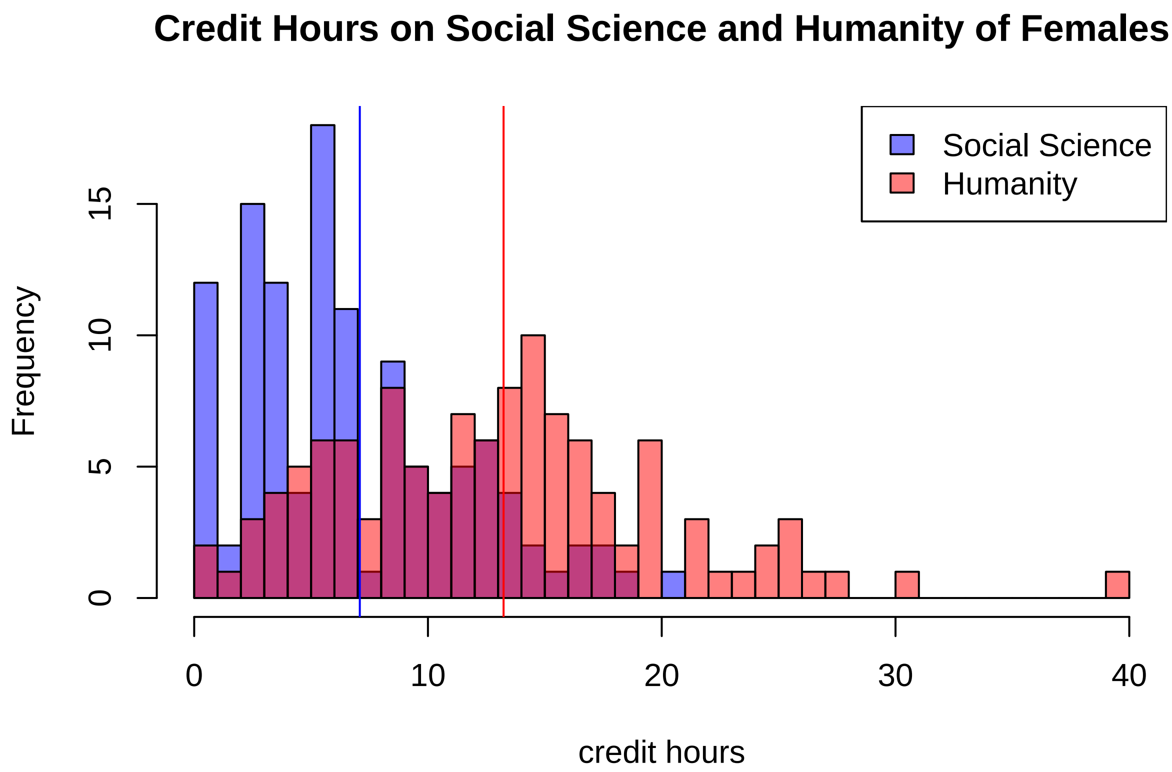

```r
#graph of Credit Hours on Social Science and Humanity of males
p1 <- hist(female$data.HU, breaks = 40)
```




```r
p2 <- hist(female$data.SS, breaks = 25)
```



```r
plot(p2, col = rgb(0, 0, 1, 1/2), main="Credit Hours on Social Science and Humanity of Females", xlab="c
plot(p1, col = rgb(1, 0, 0, 1/2), add = T)
abline(v = mean(female$data.SS), col="blue")
abline(v = mean(female$data.HU), col="red")
legend("topright", c("Social Science", "Humanity"), fill = c(rgb(0, 0, 1, 1/2), rgb(1, 0, 0, 1/2)))
```




```r
#mean
mean(male$data.SS)
```

```
## [1] 7.436275
```

```r
mean(male$data.HU)
```

```
## [1] 12.96275
```

```r
mean(female$data.SS)
```

```
## [1] 7.08547
```

```r
mean(female$data.HU)
```

```
## [1] 13.23504
```

```r
#normality test
shapiro.test(male$data.SS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male$data.SS
## W = 0.94794, p-value = 0.0005277
```

```r
shapiro.test(male$data.HU)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male$data.HU
## W = 0.94264, p-value = 0.000241
```

```r
shapiro.test(female$data.SS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female$data.SS
## W = 0.95042, p-value = 0.0002822
```

```r
shapiro.test(female$data.HU)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female$data.HU
## W = 0.96966, p-value = 0.009397
```

```r
#variance test
bartlett.test(HU ~ gender, df)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  HU by gender
## Bartlett's K-squared = 0.90554, df = 1, p-value = 0.3413
```

```r
bartlett.test(SS ~ gender, df)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  SS by gender
## Bartlett's K-squared = 0.71881, df = 1, p-value = 0.3965
```

document: male: subset that contains the sample data of male students. female: subset that contains the sample data of female students. male$data.HU : Credithoursonhumanityofmalesubsetfemale$data.HU: Credit hours on humanity of female subset male$data.SS : Credithoursonsocialscienceofmalesubsetfemale$data.SS: Credit hours on social science of female subset