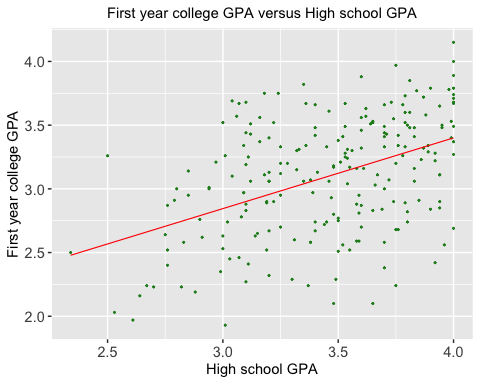
Will students who have high GPA in high school get high GPA in the first year college?

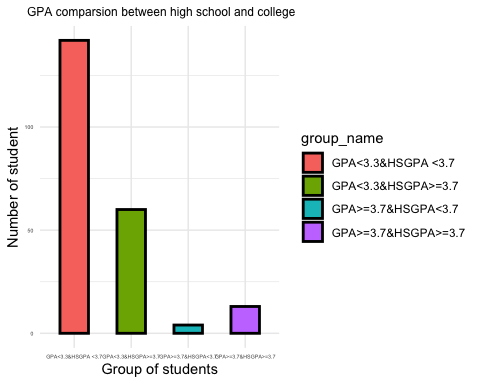
## Interpretation and Implementation

#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\_smooth(size=0.4,method="lm",color="red",se=F)+labs(x="High school GPA",y="First year college GPA",title="First year college GPA versus High school GPA")+theme(axis.text=element\_text(size=11),plot.title=element\_text(size=11,hjust = 0.5))  
  
graph\_GPA

## `geom\_smooth()` using formula 'y ~ x'



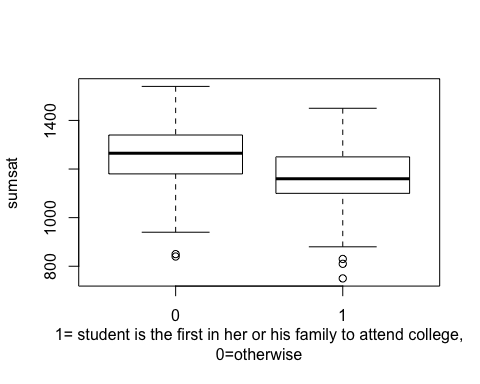
#barchart   
#set up  
group\_name<-c("GPA>=3.7&HSGPA>=3.7","GPA>=3.7&HSGPA<3.7","GPA<3.3&HSGPA>=3.7","GPA<3.3&HSGPA <3.7")  
group\_size<-c(length(group\_A$GPA),length(group\_B$GPA),length(group\_C$GPA),length(group\_D$GPA))  
group\_data<-data.frame(group\_name,group\_size)  
  
   
  
ggplot(data=group\_data,aes(x=group\_name,y=group\_size,fill=group\_name))+geom\_bar(stat="identity",width=0.5,color="black",size=1)+theme\_minimal()+labs(x="Group of students",y="Number of student",title=" GPA comparsion between high school and college")+theme(axis.text=element\_text(size=4),plot.title=element\_text(size=9,hjust = 0.5))



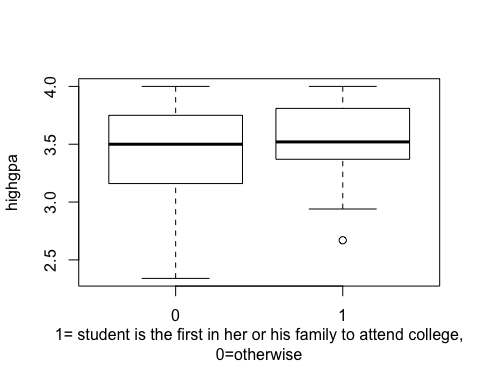
#Pearson’s Correlation Coefficient  
  
size\_of\_sample<-length(x)  
  
#mean of sample X  
mean\_x<-mean(x)  
  
#mean of sample Y  
mean\_y<-mean(y)  
  
#calculation for Pearson’s Correlation Coefficient according to the formula shown in the appendix of the report  
numerator<-sum((x-mean\_x)\*(y-mean\_y))  
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

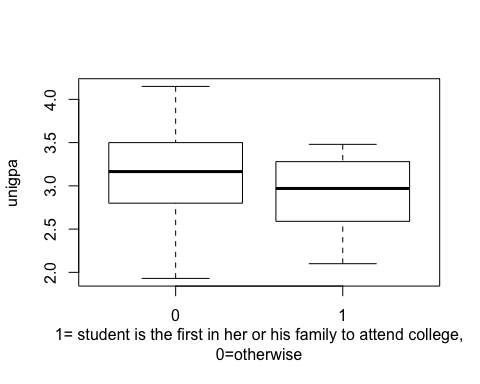
#secondary objective  
#set up  
#load the data  
data<- read.csv("FirstYearGPA.csv",header=T)  
#set up variables  
satvnc\_1<- data$SATV  
satvnc\_2<- data$SATM  
#sumsat=sum of SAT score  
sumsat<- satvnc\_1 + satvnc\_2  
#highgpa=high school GPA  
highgpa<- data$HSGPA  
#unigpa=first year college GPA  
unigpa<- data$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")



#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")



#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")



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

## [1] -0.250557

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

## [1] 0.06418575

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

## [1] -0.1565773

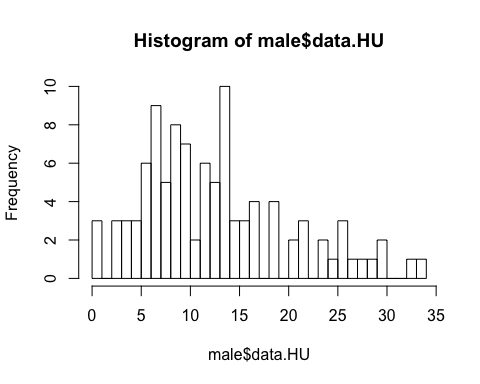
#Tertiary objective  
#read data set  
data <- read.csv("FirstYearGPA.csv")  
  
#extract the Male, HU, SS from the data set  
d <- data.frame(data$Male, data$HU, data$SS)  
  
#seperate set  
m <- d$data.Male == 1  
male <- d[m,]  
female <- d[!m,]  
  
#add gender  
nrow(male)

## [1] 102

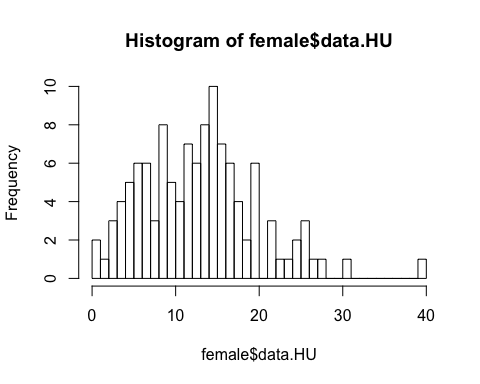
nrow(female)

## [1] 117

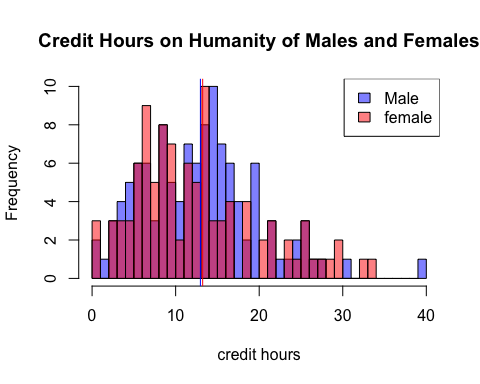
gender <- c(rep("Male",102), rep("Female",117))  
HU <- c(male$data.HU, female$data.HU)  
SS <- c(male$data.SS, female$data.SS)  
df <- data.frame(gender, HU, SS)  
genderm <- c(rep("Male",102))  
HUm <- c(male$data.HU)  
SSm <- c(male$data.SS)  
dfm <- data.frame(genderm, HUm, SSm)  
genderf <- c(rep("Female",117))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



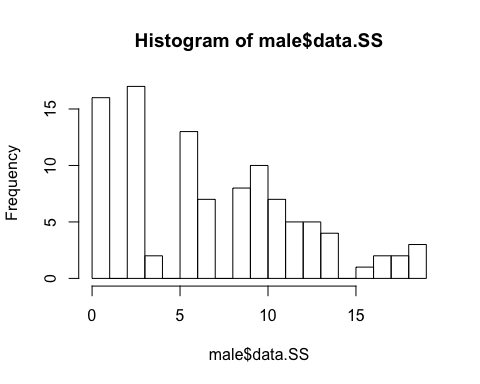
p2 <- hist(female$data.HU, breaks = 40)



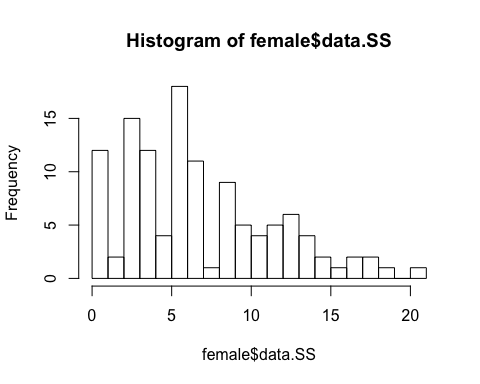
plot(p2, col = rgb(0, 0, 1, 1/2), main="Credit Hours on Humanity of Males and Females", xlab="credit hours", ylim = c(0,10))  
plot(p1, col = rgb(1, 0, 0, 1/2), add = T)  
abline(v = mean(male$data.HU), col="blue")  
abline(v = mean(female$data.HU), col="red")  
legend("topright", c("Male", "female"), fill = c(rgb(0, 0, 1, 1/2), rgb(1, 0, 0, 1/2)))



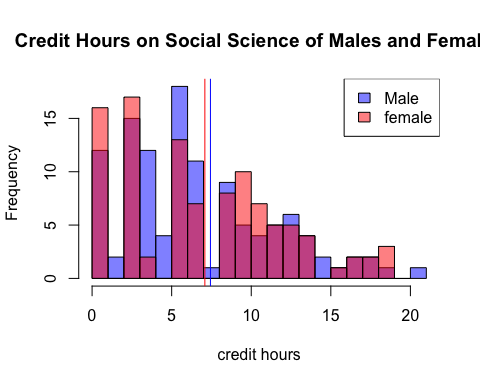
#graph of Credit Hours on Social Science course  
p1 <- hist(male$data.SS, breaks = 25)



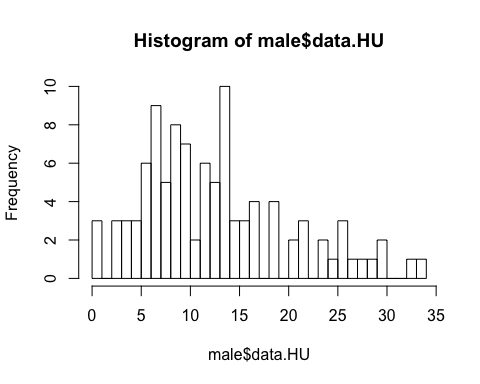
p2 <- hist(female$data.SS, breaks = 25)



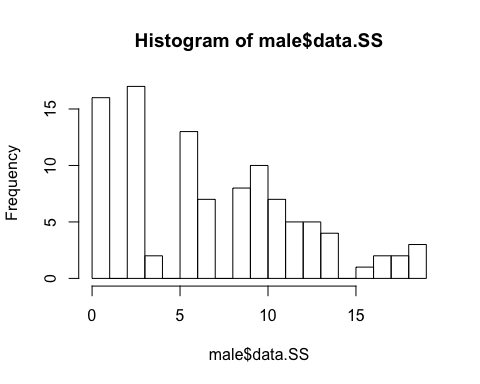
plot(p2, col = rgb(0, 0, 1, 1/2), main="Credit Hours on Social Science of Males and Females", xlab="credit hours", ylim = c(0,18))  
plot(p1, col = rgb(1, 0, 0, 1/2), add = T)  
abline(v = mean(male$data.SS), col="blue")  
abline(v = mean(female$data.SS), col="red")  
legend("topright", c("Male", "female"), fill = c(rgb(0, 0, 1, 1/2), rgb(1, 0, 0, 1/2)))



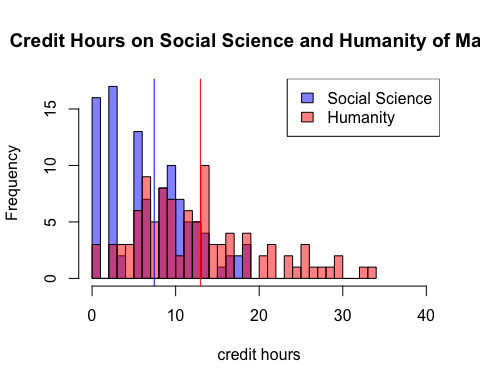
#graph of Credit Hours on Social Science and Humanity of males  
p1 <- hist(male$data.HU, breaks = 40)



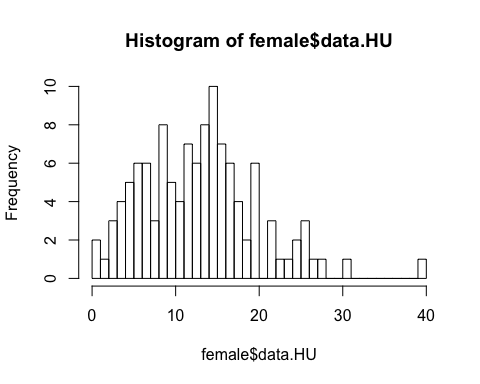
p2 <- hist(male$data.SS, breaks = 20)



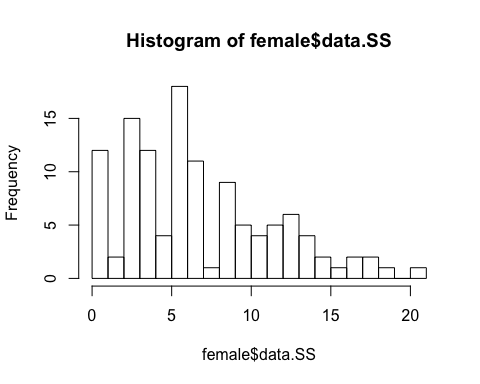
plot(p2, col = rgb(0, 0, 1, 1/2), main="Credit Hours on Social Science and Humanity of Males", xlab="credit hours", xlim = c(0,40), ylim = c(0,17))  
plot(p1, col = rgb(1, 0, 0, 1/2), add = T)  
abline(v = mean(male$data.SS), col="blue")  
abline(v = mean(male$data.HU), col="red")  
legend("topright", c("Social Science", "Humanity"), fill = c(rgb(0, 0, 1, 1/2), rgb(1, 0, 0, 1/2)))



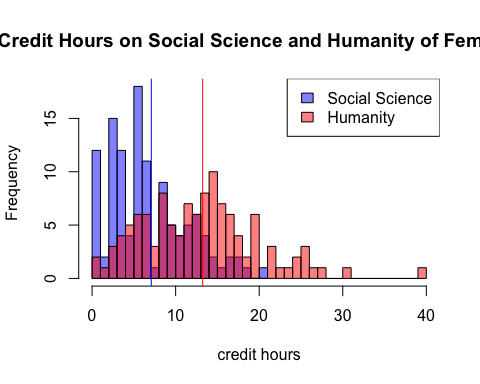
#graph of Credit Hours on Social Science and Humanity of males  
p1 <- hist(female$data.HU, breaks = 40)



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



plot(p2, col = rgb(0, 0, 1, 1/2), main="Credit Hours on Social Science and Humanity of Females", xlab="credit hours", xlim = c(0,40), ylim = c(0,18))  
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)))



#mean  
mean(male$data.SS)

## [1] 7.436275

mean(male$data.HU)

## [1] 12.96275

mean(female$data.SS)

## [1] 7.08547

mean(female$data.HU)

## [1] 13.23504

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

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

shapiro.test(male$data.HU)

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

shapiro.test(female$data.SS)

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

shapiro.test(female$data.HU)

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

#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

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. maledata.HU: Credit hours on humanity of female subset maledata.SS: Credit hours on social science of female subset