Descriptive Statistics

Introduction to Descriptive Statistics

Descriptive Statistics is used in univariate Analysis to summarize and describe a sample.

There are three types of Descriptive statistics

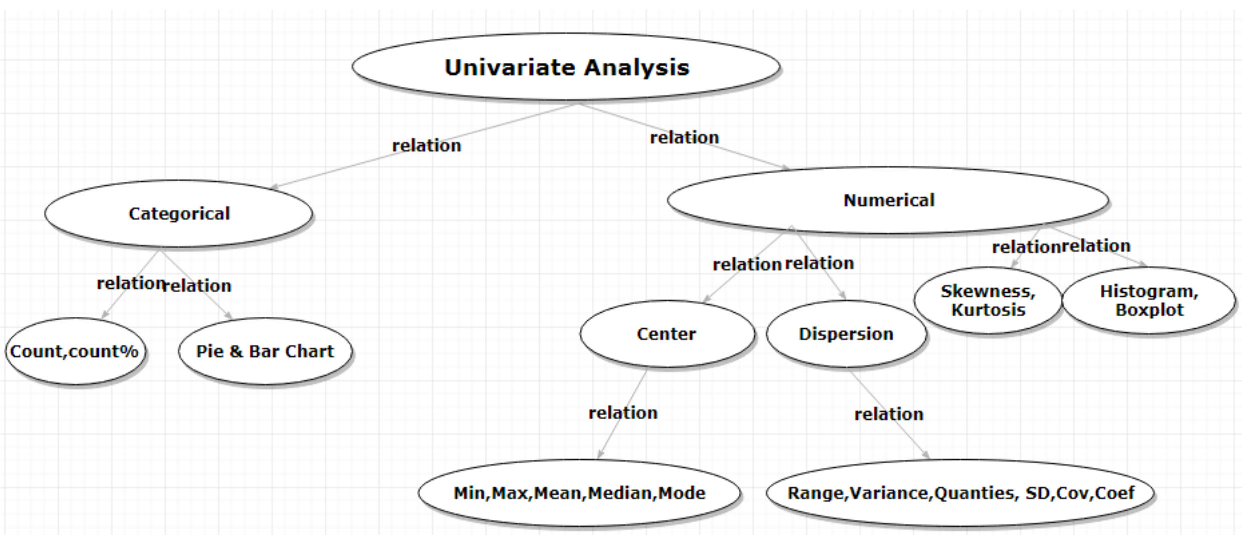
1) Frequence Distribution - A number of times a characteristics of a variable is observed in a sample.

2) Central Tendencey - Central Tendency is an estimate of the "Centre" of a distribution of a value.

- ```
  There are two types of variable in Central Tendency.\newline
  1) continuous Variable - A variable that can take on an infinite number of positive values... i.e
  2) Discrete variable - A variable that's characteristics are seperate from each other.\newline
  ```

3) Dispersion - Dispersion is the spread of the value around a central tendency.

Univariate Analysis Decision Tree


  ```

```
## [1] 2
```

Chapter 1 : Descriptive Statistics

Step 1 : Read Data

```r
LungCapData <- read.csv("LungCapData.txt",header = T,sep = "\t")
attach(LungCapData)
names(LungCapData)
```

```
## [1] "LungCap"   "Age"       "Height"    "Smoke"     "Gender"    "Caesarean"
```

ask for summaries for the variable LungCap

```r
summary(LungCapData)
```

```
##     LungCap            Age            Height         Smoke        Gender
##  Min.   : 0.507   Min.   : 3.00   Min.   :45.30   no :648   female:358
##  1st Qu.: 6.150   1st Qu.: 9.00   1st Qu.:59.90   yes: 77   male  :367
##  Median : 8.000   Median :13.00   Median :65.40
##  Mean   : 7.863   Mean   :12.33   Mean   :64.84
##  3rd Qu.: 9.800   3rd Qu.:15.00   3rd Qu.:70.30
##  Max.   :14.675   Max.   :19.00   Max.   :81.80
##  Caesarean
##  no :561
##  yes:164
##
##
##
##
```

table for praportion

```r
table(Smoke)
```

```
## Smoke
##  no yes
## 648  77
```

```r
table(Smoke)/725
```

```
## Smoke
##        no       yes
## 0.8937931 0.1062069
```

```r
table(Smoke)/length(Smoke)
```

```
## Smoke
##        no       yes
## 0.8937931 0.1062069
```

two way table or contigency table

```r
table(Smoke, Gender)
```

```
##      Gender
## Smoke female male
##   no     314  334
##   yes     44   33
```

```r
mean(LungCap)
```

```
## [1] 7.863148
```

```r
mean(LungCap, trim = 0.10)
```

```
## [1] 7.938081
```

To calculate the median

```r
median(LungCap)
```

```
## [1] 8
```

To calculate the Variance

```r
var(LungCap)
```

```
## [1] 7.086288
```

To calculate the standard deviation

```r
sd(LungCap)
```

```
## [1] 2.662008
```

To calculate the square root

```r
sqrt(var(LungCap))
```

```
## [1] 2.662008
```

```r
sd(LungCap)^2
```

```
## [1] 7.086288
```

```r
min(LungCap)
```

```
## [1] 0.507
```

```r
max(LungCap)
```

```
## [1] 14.675
```

```r
range(LungCap)
```

```
## [1]  0.507 14.675
```

```r
quantile(LungCap, probs = 0.90)
```

```
##    90%
## 11.205
```

```r
quantile(LungCap,probs = c(0.20,0.5,0.9,1))
```

```
##    20%    50%    90%   100%
##  5.645  8.000 11.205 14.675
```

```r
library(e1071)
skewness(LungCap)
```

```
## [1] -0.2269314
```

```r
library(e1071)
kurtosis(LungCap)
```

```
## [1] -0.3259122
```

```r
sum(LungCap)
```

```
## [1] 5700.782
```

count the variables

```r
length(LungCap)
```

```
## [1] 725
```

Calculating Correlation

```r
cor(LungCap,Age)
```

```
## [1] 0.8196749
```

```r
cor(LungCap,Age,method = "spearman")
```

```
## [1] 0.8172464
```

Calculating covariance

```r
cov(LungCap,Age)
```

```
## [1] 8.738289
```

calculating variance

```r
var(LungCap,Age)
```

```
## [1] 8.738289
```

Calculating summary

```r
summary(LungCap)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.507   6.150   8.000   7.863   9.800  14.675
```

```
## [1] 2
```

Chapter 2 : Bar Charts & Pie Charts

Bar charts and pie charts are appropriate for summarizing the distribution of a categorical variable.

A Barchart is a visual display of the frequency for each category of a categorical variable or the relative frequency (%) for each category

```r
table(Gender)
```

```
## Gender
## female   male
##    358    367
```

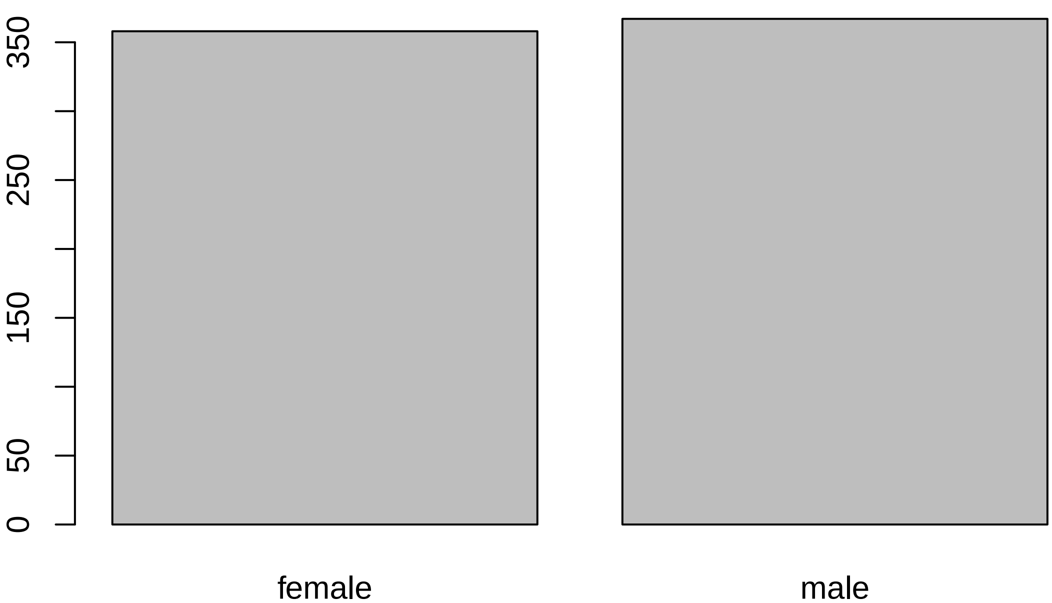
```r
count <- table(Gender)
```

```r
table(Gender)/725
```

```
## Gender
##    female      male
## 0.4937931 0.5062069
```

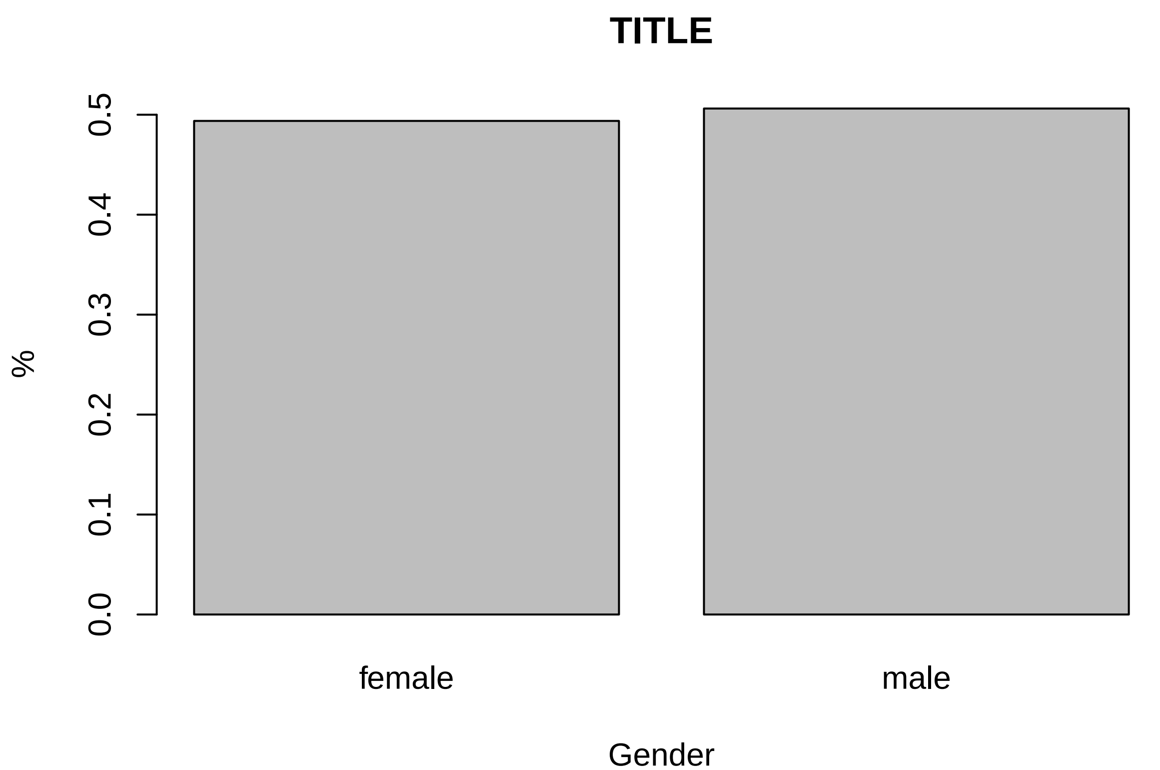
```r
percent <- table(Gender)/725
```

```r
barplot(count)
```

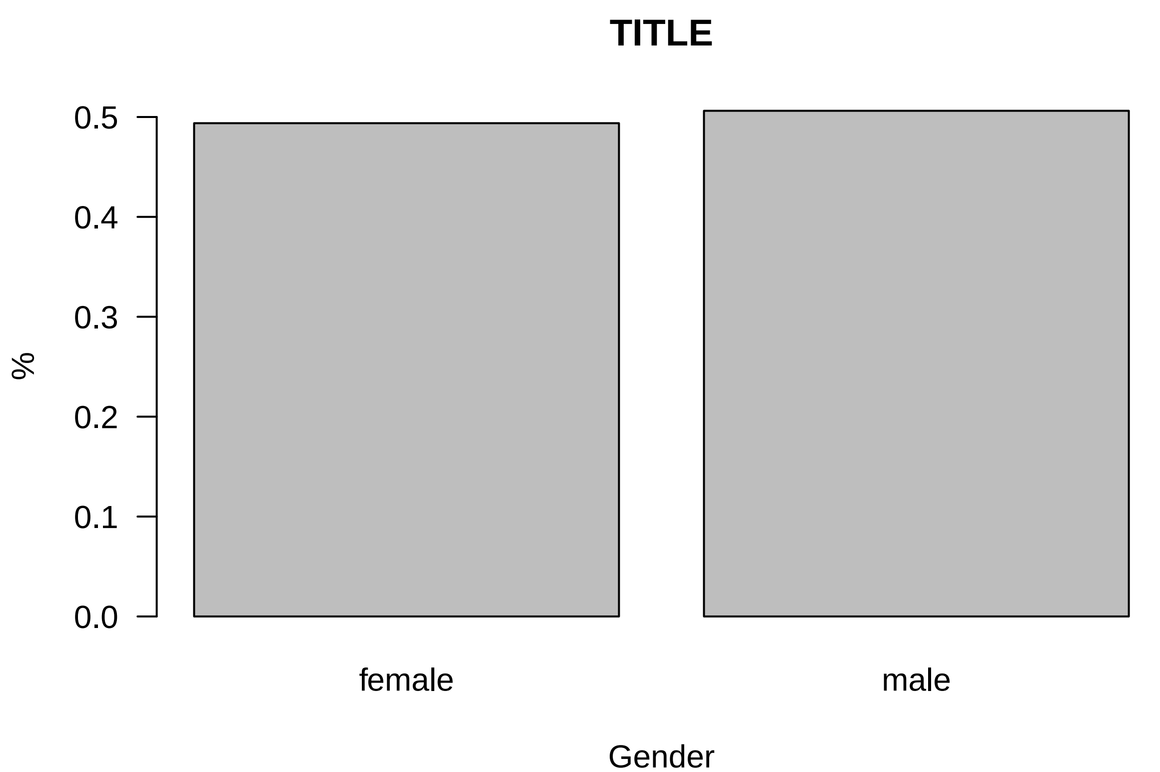


adding title to plot using main command

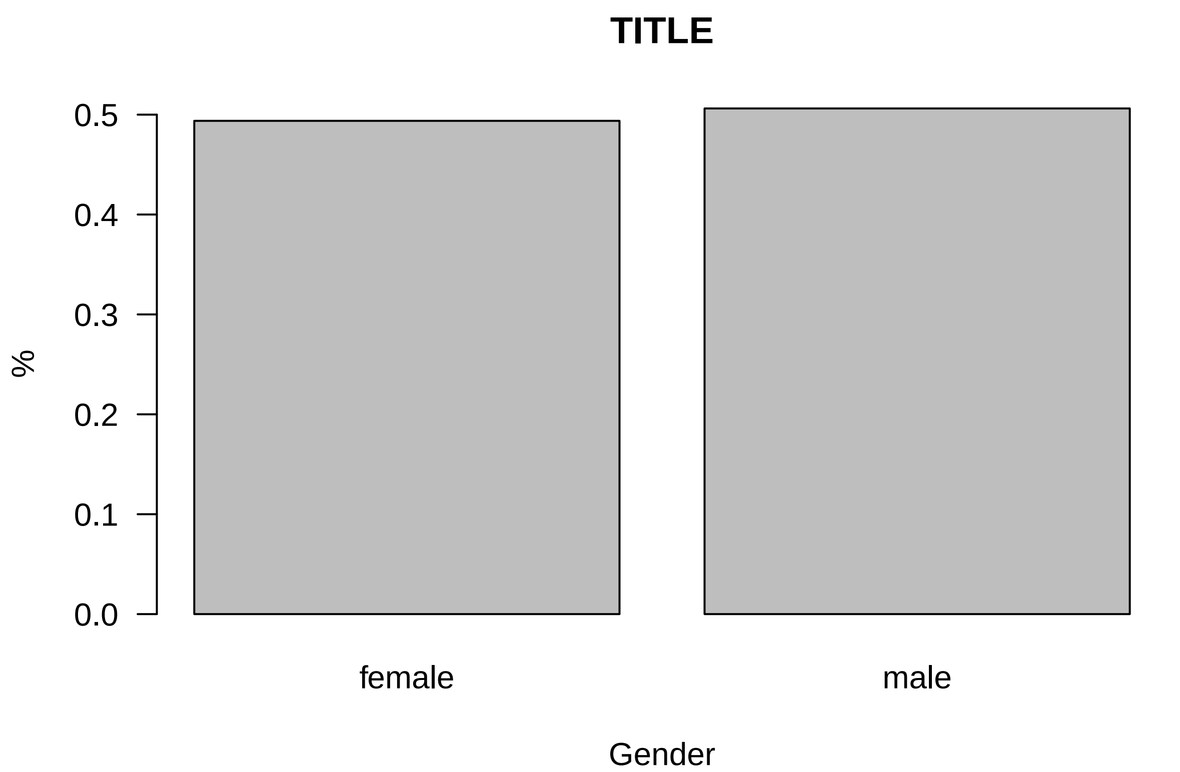
```r
barplot(percent,main = "TITLE", xlab = "Gender",ylab = "%")
```



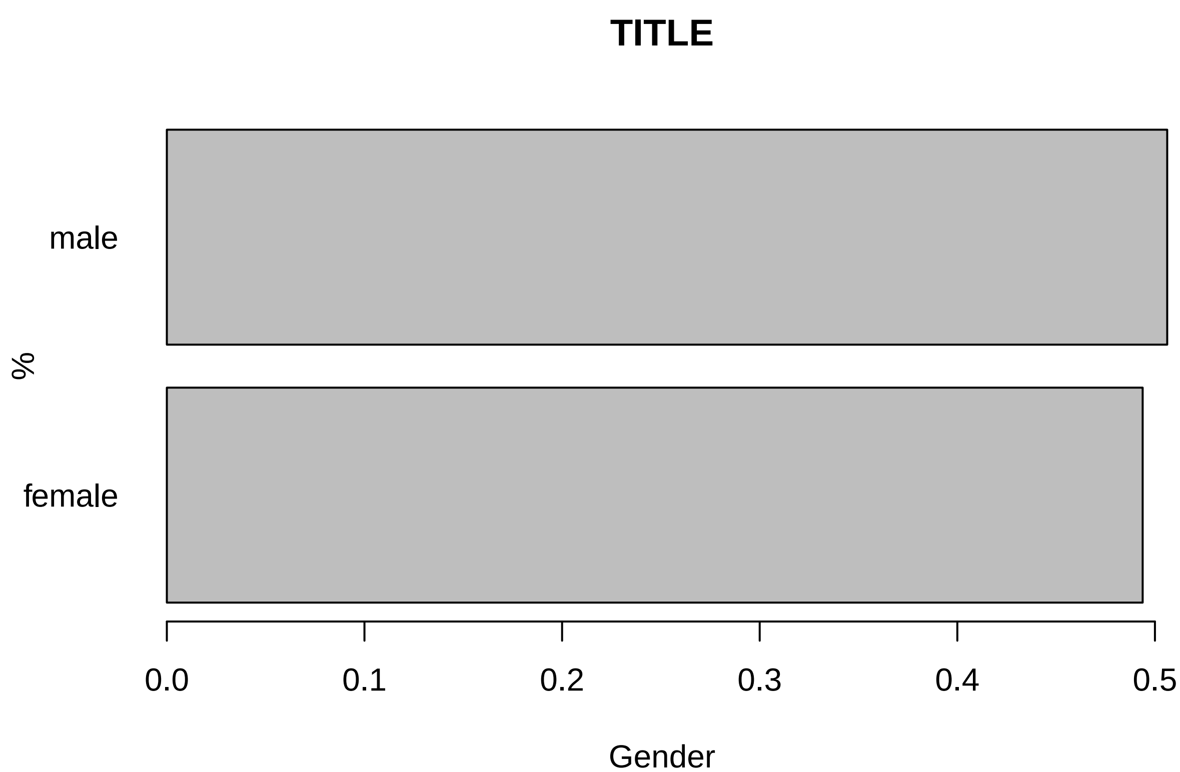

```
barplot(percent,main = "TITLE", xlab = "Gender",ylab = "%",las = 1)
```




```
barplot(percent,main = "TITLE", xlab = "Gender",ylab = "%",las = 1,names.arg = c("female","male"))
```

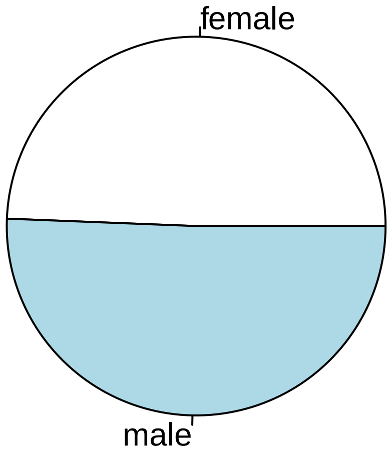



```
barplot(percent,main = "TITLE", xlab = "Gender",ylab = "%",las = 1,names.arg = c("female","male"),horiz
```

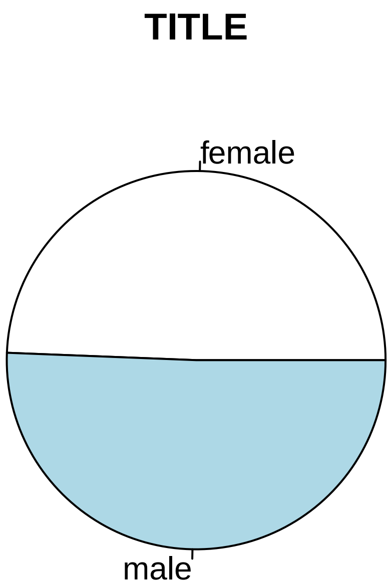


Making Pie Chart

```
pie(count)
```




```
pie(count,main="TITLE")
```

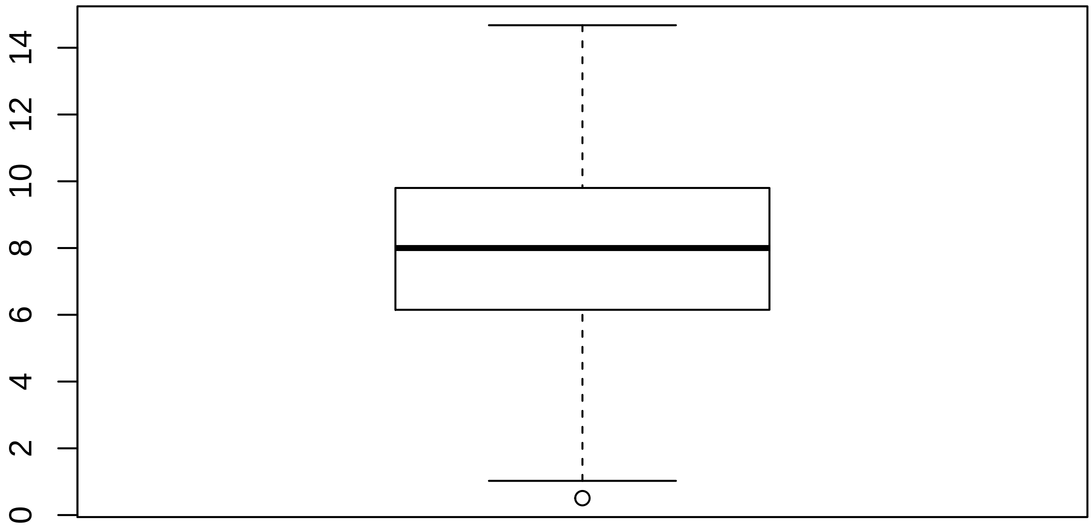


```
## [1] 2
```

Chapter 3 : Making Boxplot and Grouped Boxplots

A Boxplot is appropriate for summarizing the distribution of a numeric variables.

```r
boxplot(LungCap)
```

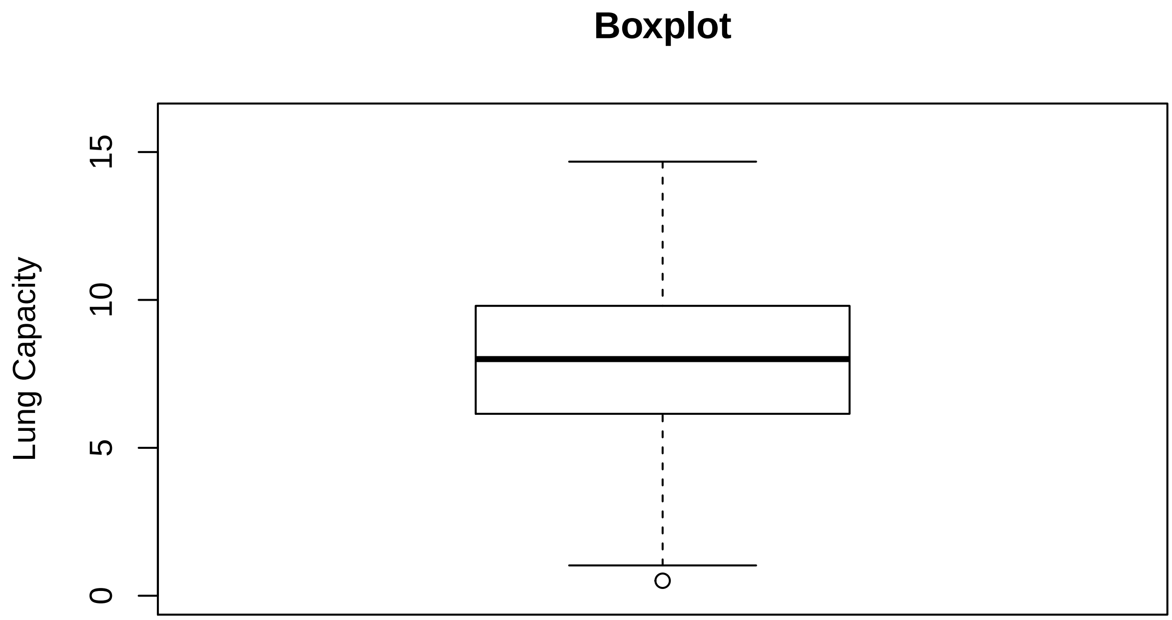


let ask R for mean,median, first quantile,3rd quantile and maximum

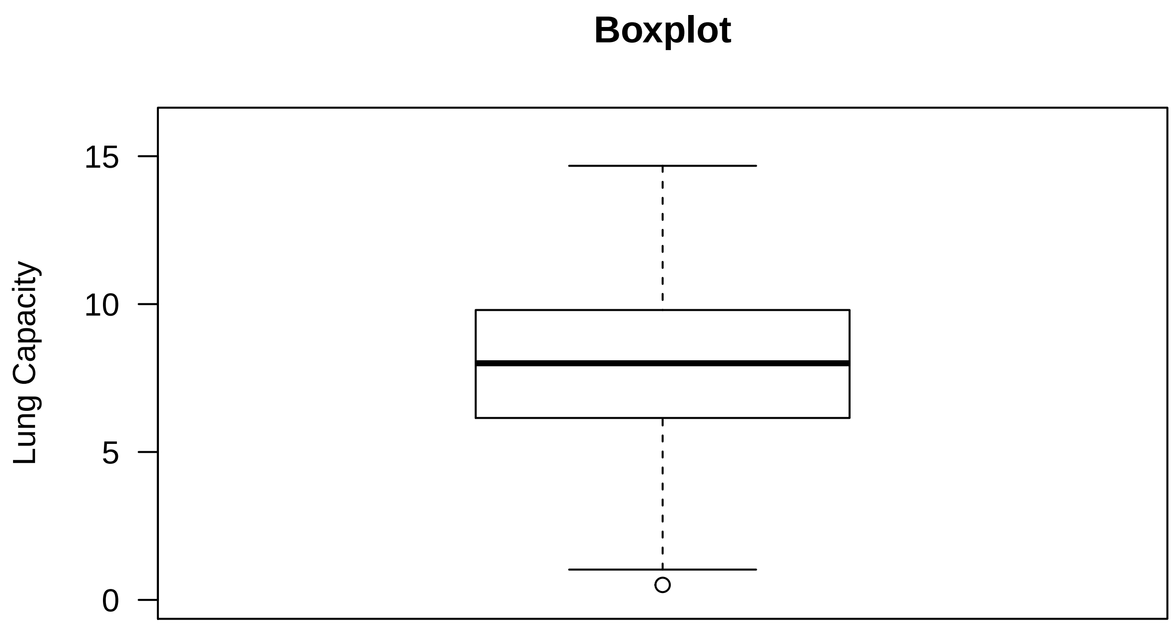
```r
quantile(LungCap,probs = c(0,0.25,0.5,0.75,1))
```

```
##     0%    25%    50%    75%   100%
##  0.507  6.150  8.000  9.800 14.675
```

```r
boxplot(LungCap,main="Boxplot",ylab="Lung Capacity",ylim=c(0,16))
```

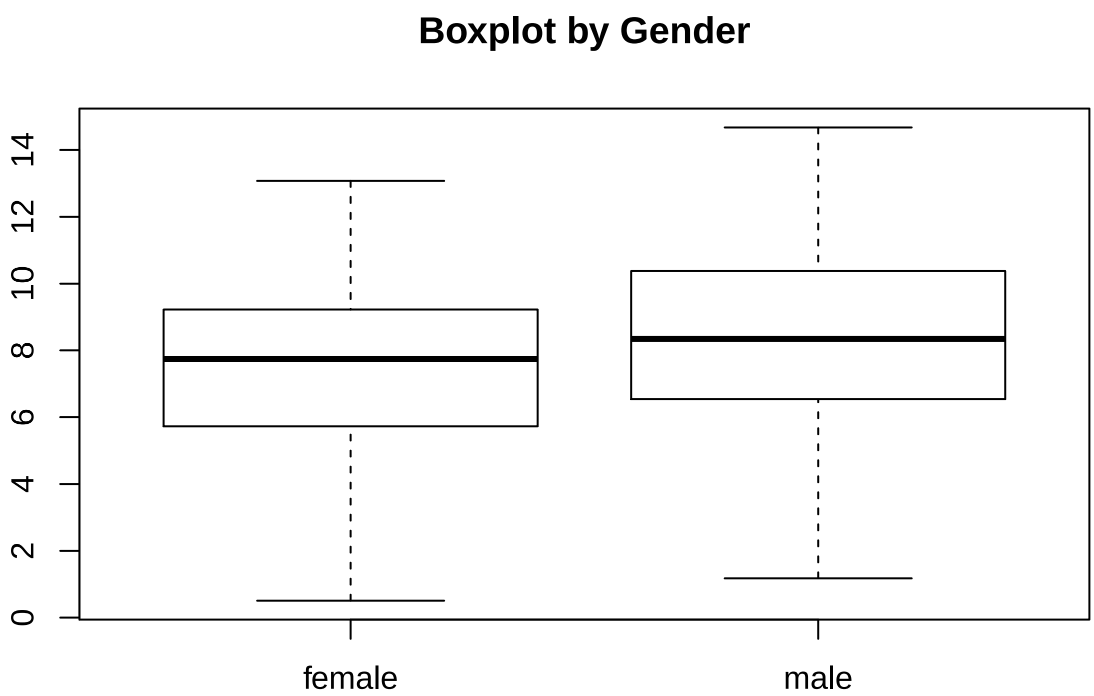



```r
boxplot(LungCap,main="Boxplot",ylab="Lung Capacity",ylim=c(0,16),las=1)
```



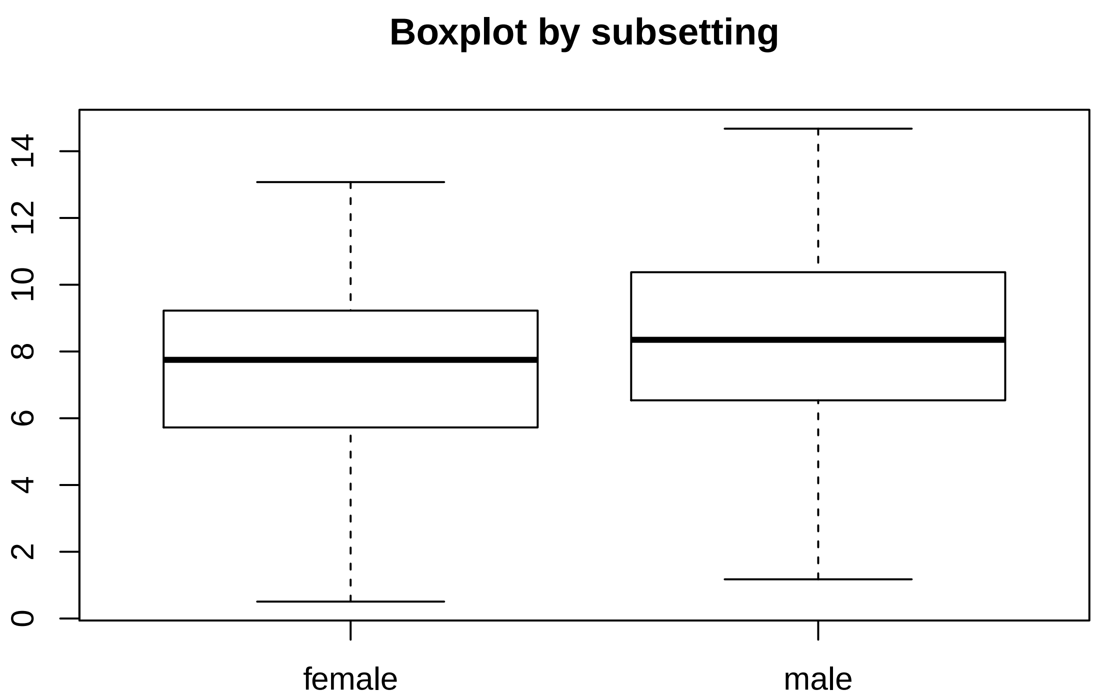
two or more boxplots

```r
boxplot(LungCap ~ Gender,main="Boxplot by Gender")
```



Boxplot with subseting

```
boxplot(LungCap[Gender == "female"],LungCap[Gender == "male"],main="Boxplot by subsetting",names = c("fe
```

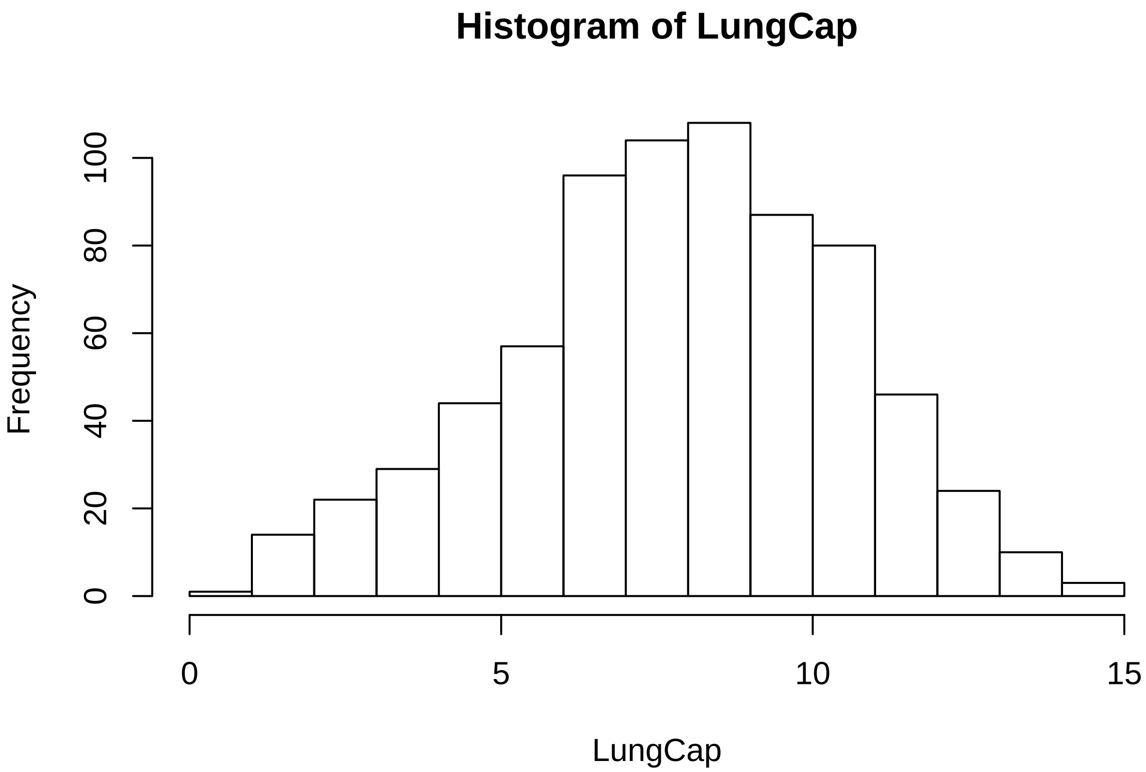


```
## [1] 2
```

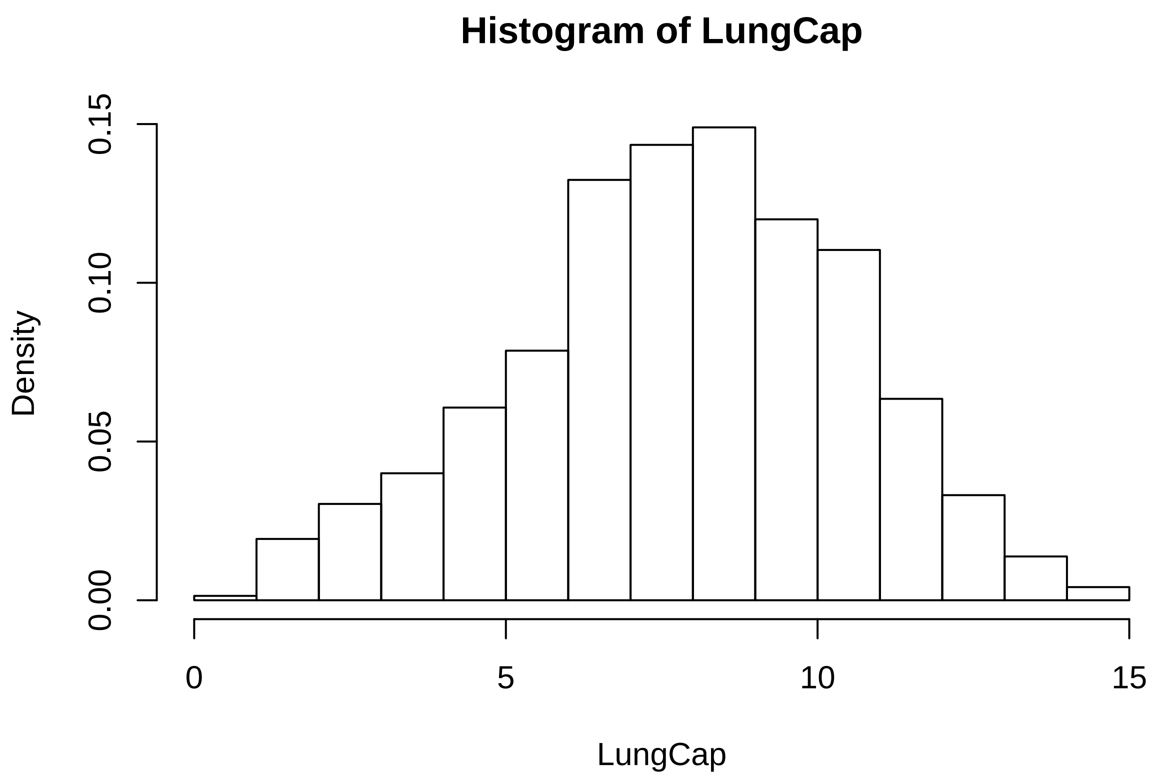
Chapter 4 : Histogram

A histogram is appropriate for summarizing the distribution of a numeric variable. . . .

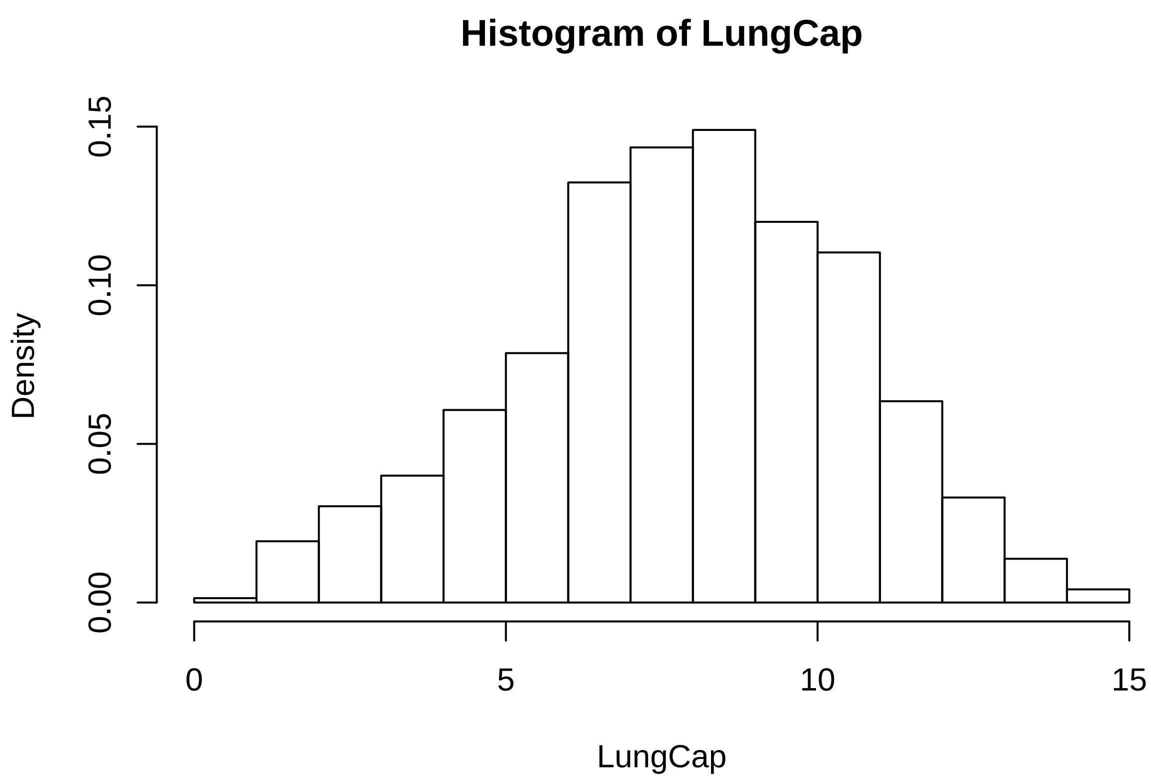
```r
hist(LungCap)
```



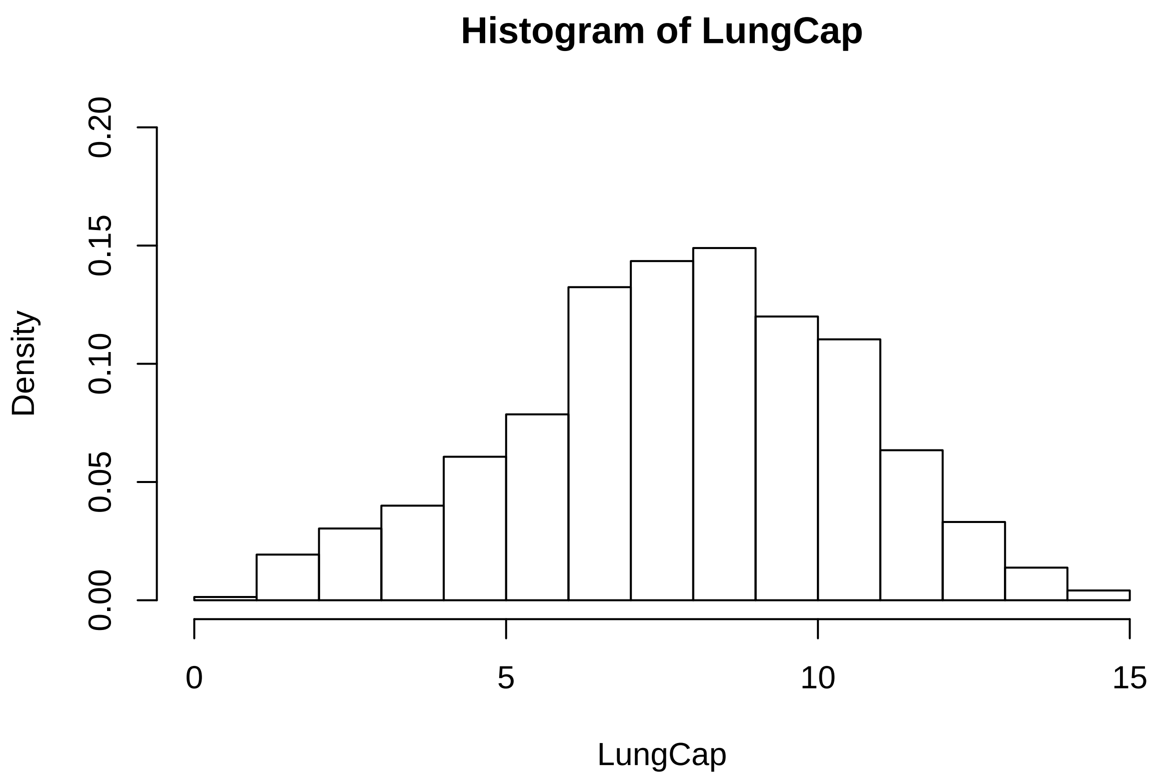

```r
hist(LungCap, freq=FALSE)
```




```r
hist(LungCap,prob =T)
```

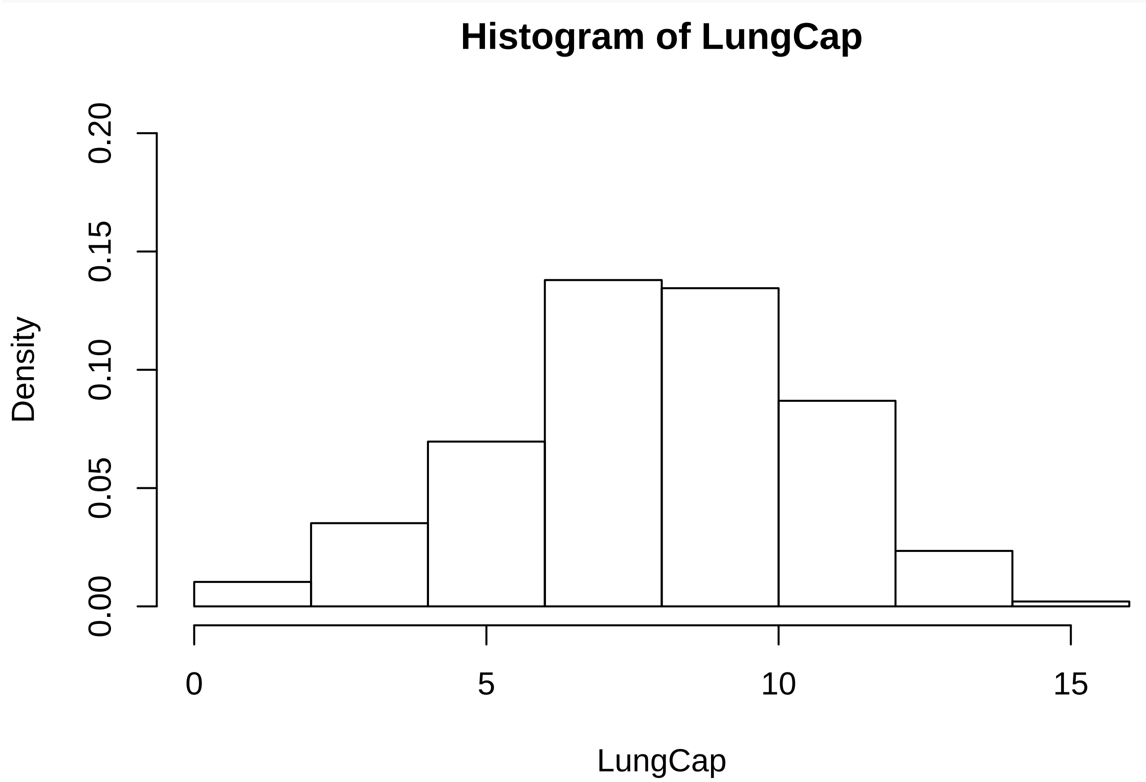



```r
hist(LungCap,prob=T,ylim = c(0,0.2))
```

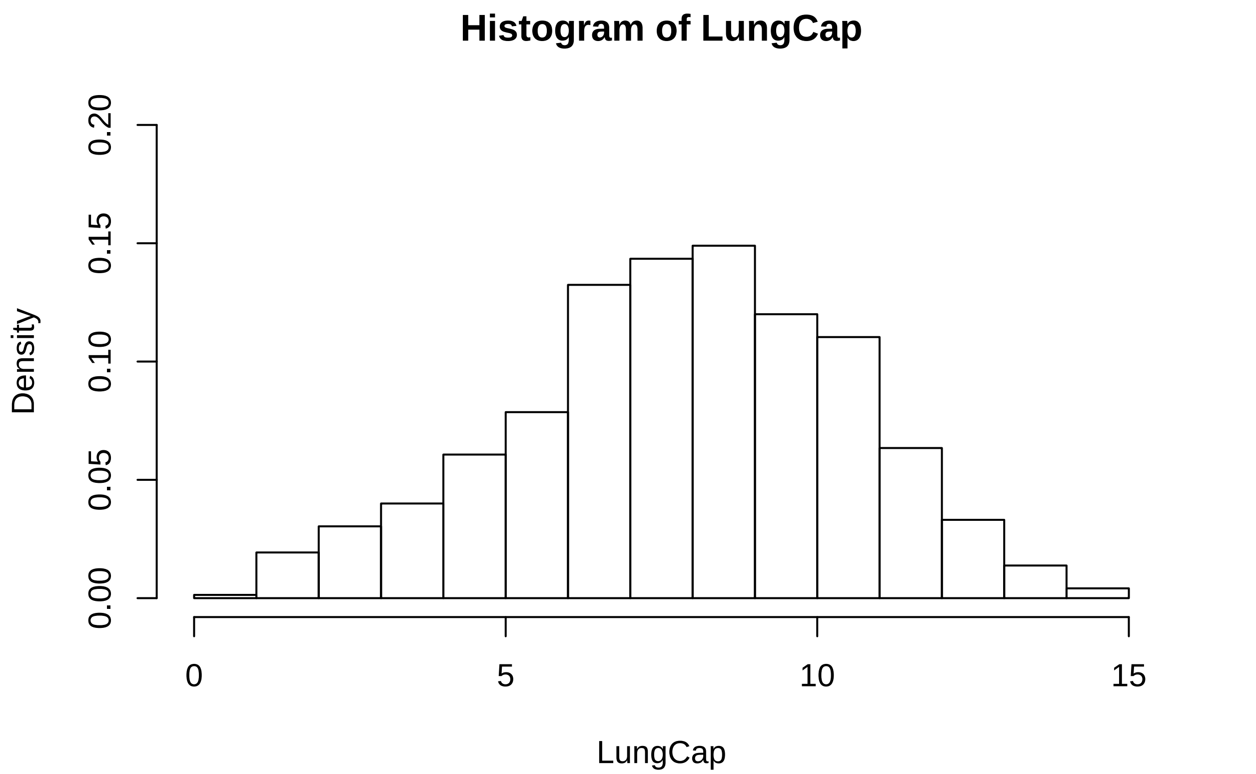


bin width of frequency

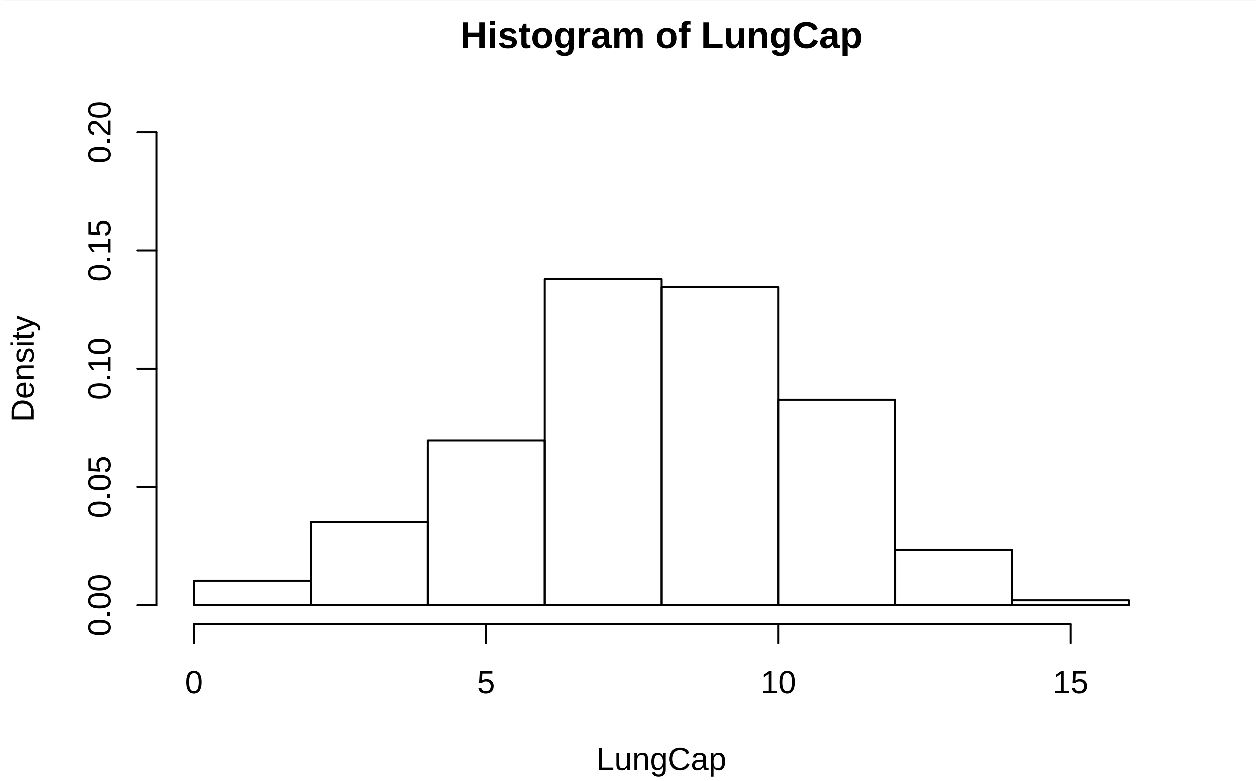
```
hist(LungCap,prob=T,ylim = c(0,0.2),breaks = 7)
```



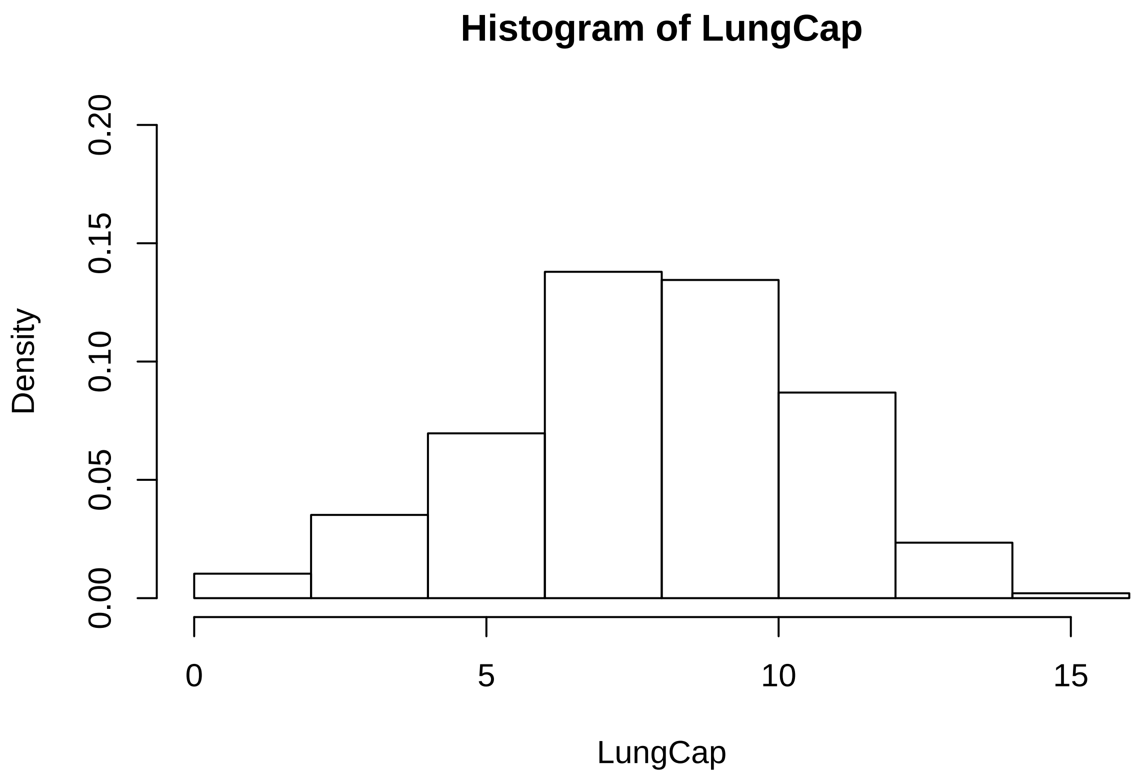
```r
hist(LungCap,prob=T,ylim = c(0,0.2),breaks = 14)
```



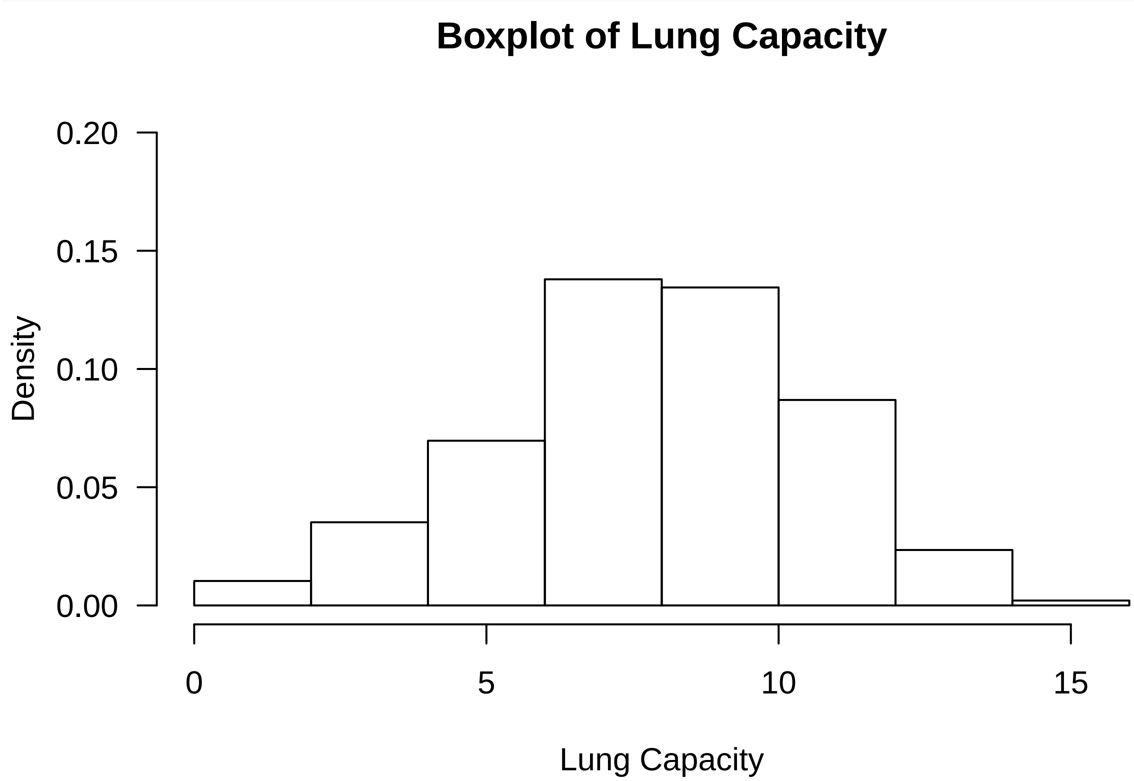

```r
hist(LungCap,prob=T,ylim = c(0,0.2),breaks = c(0,2,4,6,8,10,12,14,16))
```



```
hist(LungCap,prob=T,ylim = c(0,0.2),breaks = seq(from=0,to=16,by=2))
```




```
hist(LungCap,prob=T,ylim = c(0,0.2),breaks = seq(from=0,to=16,by=2),main = "Boxplot of Lung Capacity", :
```



Adding density curve

lines(density(LungCap),col = 2, lwd = 3)

```
## [1] 2
```

Chapter 5 : Stratified Boxplot

Stratified Boxplots are usefull for examining the relationship between a categorical variable and a numeric variable, within strata or groups defined by a third categorical variables. . . .

Create an AgeGroups variable

```r
AgeGroups  <- cut(Age,breaks = c(0,13,15,17,25),labels =c("<13","14/15","16/17","18+"))
```

```r
Age[1:5]
```

```
## [1]  6 18 16 14  5
```

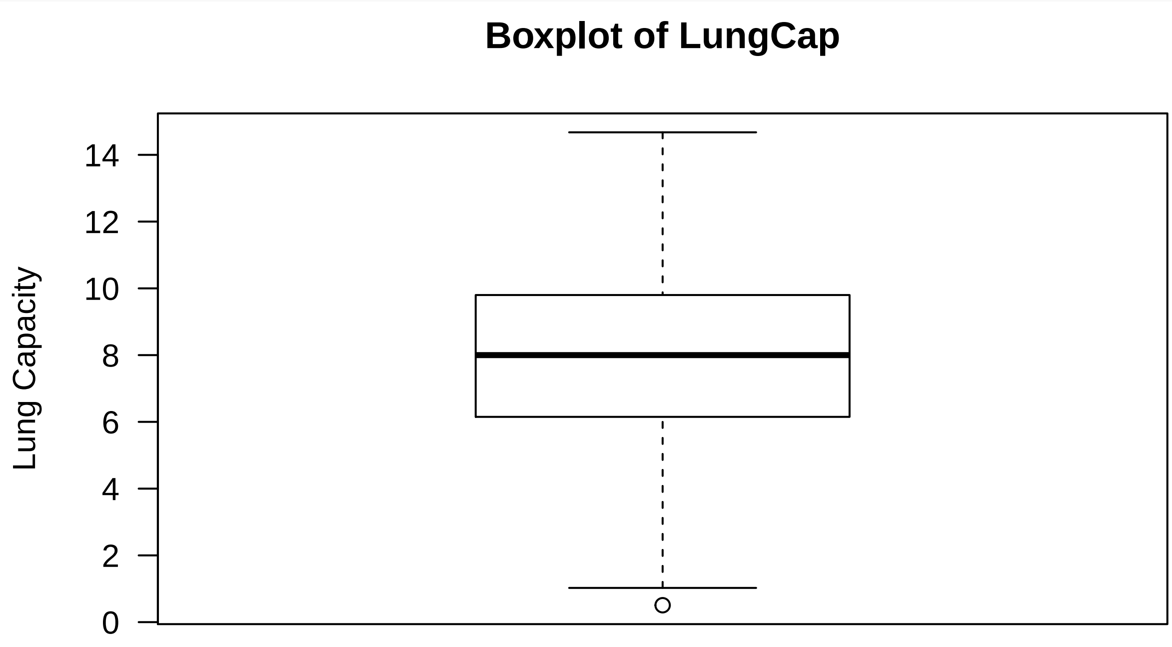
```r
AgeGroups[1:5]
```

```
## [1] <13   18+   16/17 14/15 <13
## Levels: <13 14/15 16/17 18+
```

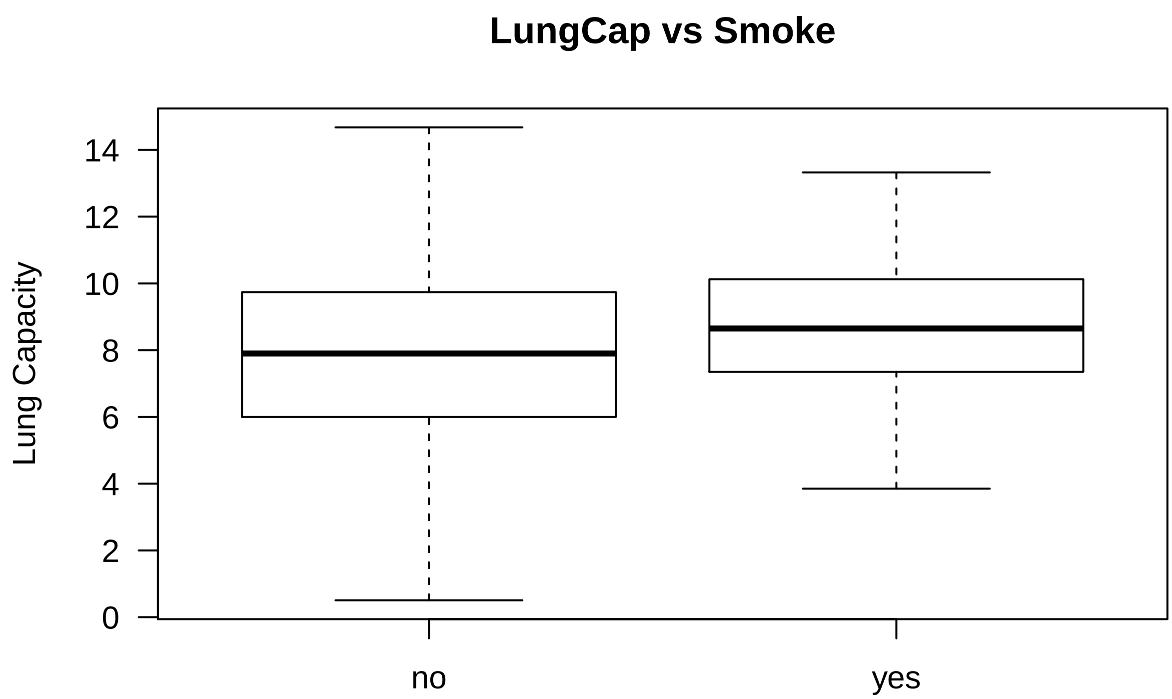
```r
levels(AgeGroups)
```

```
## [1] "<13"   "14/15" "16/17" "18+"
```

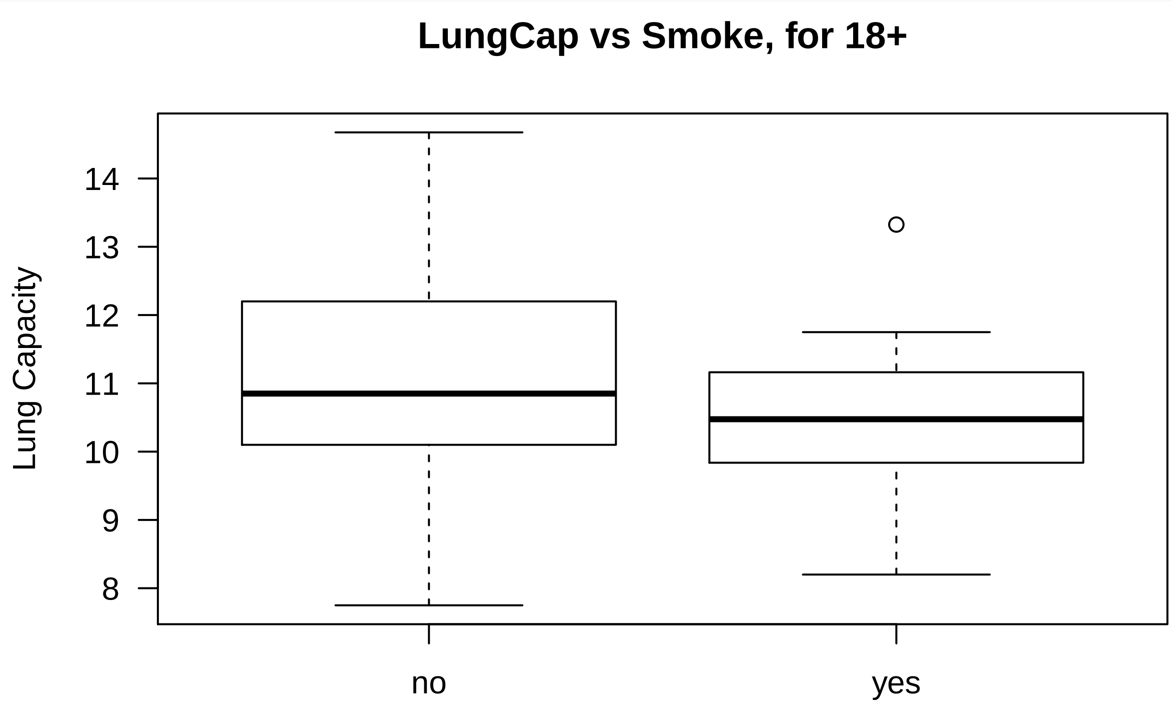
```r
boxplot(LungCap,ylab="Lung Capacity",main="Boxplot of LungCap",las = 1)
```



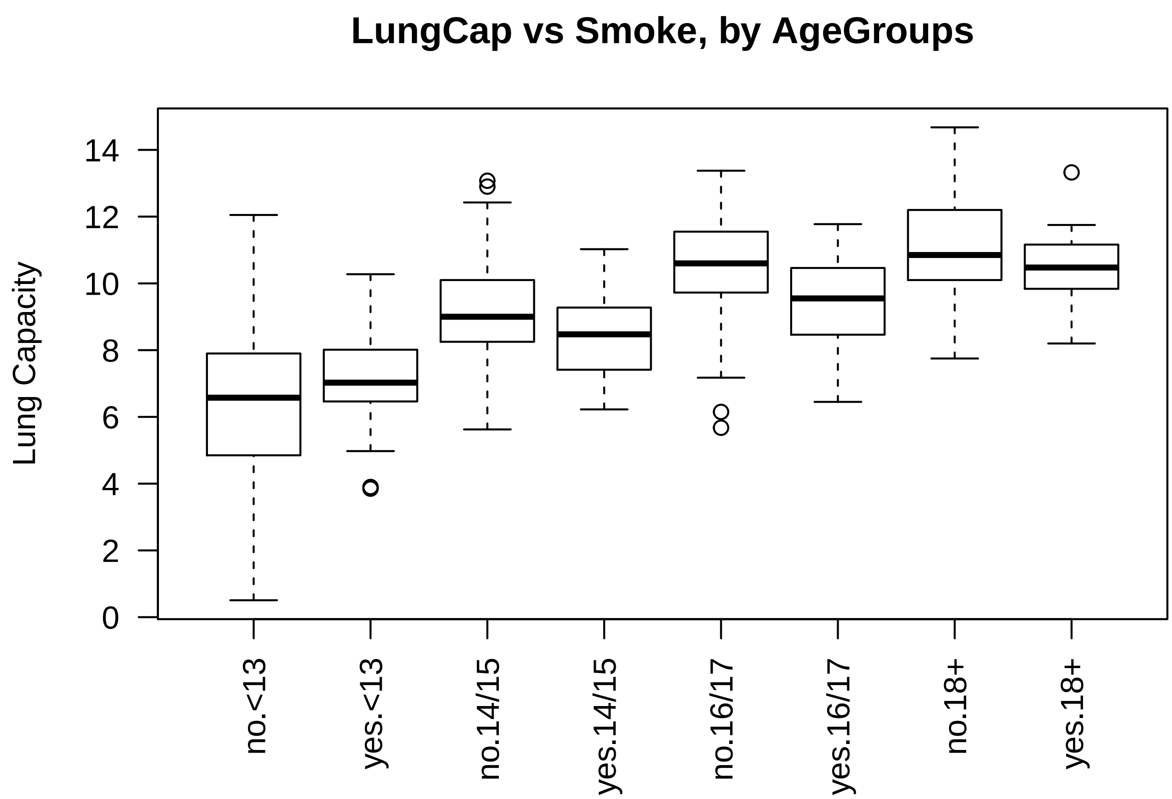

```r
boxplot(LungCap ~ Smoke,ylab="Lung Capacity",main="LungCap vs Smoke",las = 1)
```




```
boxplot(LungCap[Age >= 18] ~ Smoke[Age >= 18],ylab="Lung Capacity",main="LungCap vs Smoke, for 18+",las
```

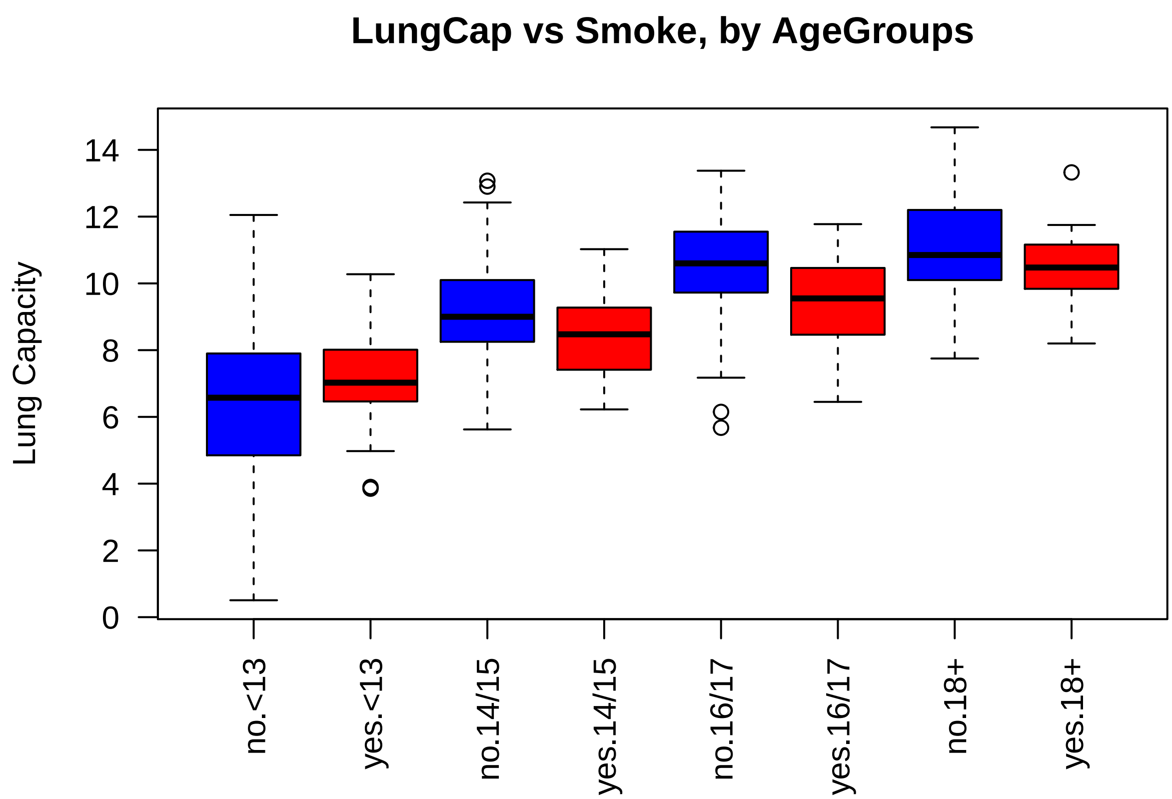



```
boxplot(LungCap ~ Smoke * AgeGroups,ylab="Lung Capacity",main="LungCap vs Smoke, by AgeGroups",las = 2)
```



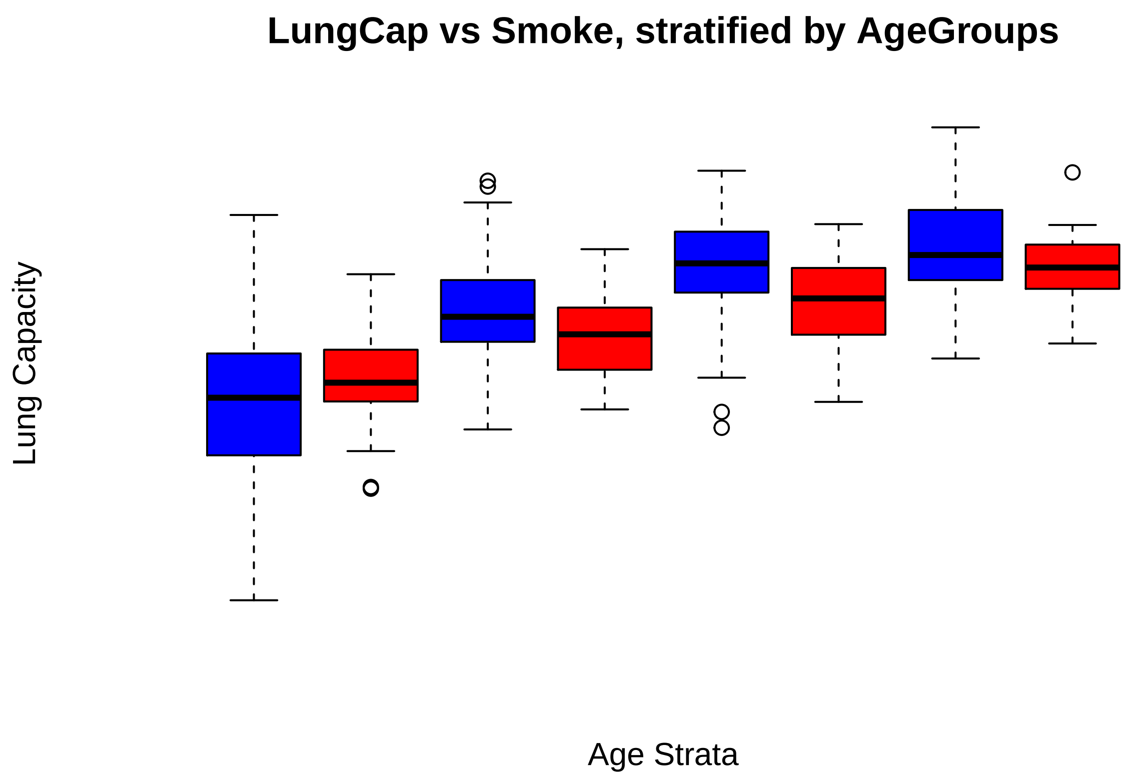
Coloring box plot blue then red

```
boxplot(LungCap ~ Smoke * AgeGroups,ylab="Lung Capacity",main="LungCap vs Smoke, by AgeGroups",las = 2,
```



Make the nice plot, with changed x-axis names, legends . . . .

```
boxplot(LungCap ~ Smoke * AgeGroups,ylab="Lung Capacity",main="LungCap vs Smoke, stratified by AgeGroups
```


LungCap vs Smoke, stratified by AgeGroups
Lung Capacity
Age Strata

```
## [1] 2
```

Chapter 6 : Steam and Leaf Plots

Note : Stem and leaf plots are appropriate for summarizing the distribution of a numeric variable and are most appropriate for smaller datasets...

Extract the lung capacity, for only females and save in female Lungcap

```
femaleLungCap <- LungCap[Gender == "female"]
```

```
stem(femaleLungCap)
```

```
##
##   The decimal point is at the |
##
##    0 | 5
##    1 | 0135689
##    2 | 0033456777789999
##    3 | 0122457788999999
##    4 | 012333344555556666677777899
##    5 | 0000122222334466666677777 8999
##    6 | 000111111122222222223334555555666666777777788888999999
##    7 | 00012333444444444555556666667778888888999999
##    8 | 0000000011111222223333334444445555556666666666777777888888888899
##    9 | 000000001112222333333444555556666777788888999999
##   10 | 00001111122233444555556667777778899
##   11 | 00111223556678888
##   12 | 1222479
##   13 | 1
```

Adjust the scale using scale argument

```
stem(femaleLungCap,scale = 2)
```

```
##
##   The decimal point is at the |
##
##    0 | 5
##    1 | 013
##    1 | 5689
##    2 | 00334
##    2 | 56777789999
##    3 | 01224
##    3 | 57788999999
##    4 | 012333344
##    4 | 555556666677777899
##    5 | 00001222223344
##    5 | 66666777778999
##    6 | 00011111112222222223334
##    6 | 5555556666667777777788888999999
```

```
##      7 | 00012333444444444
##      7 | 5555666667778888888899999
##      8 | 000000001111122222333333444444
##      8 | 555556666666666777777888888888899
##      9 | 00000000111222233333444
##      9 | 55556666777788888999999
##     10 | 00001111122233444
##     10 | 5555666777778899
##     11 | 00111223
##     11 | 556678888
##     12 | 12224
##     12 | 79
##     13 | 1
```

```
## [1] 2
```

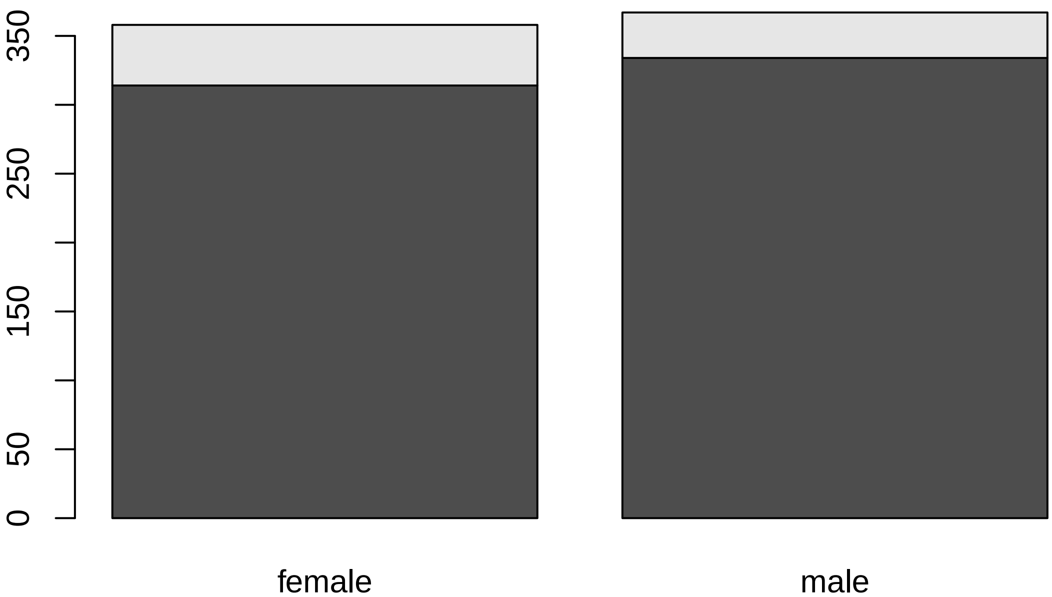
Chapter 7 : Making Stacked Barcharts, Clustered Barcharts, and Mosaic Plots

NOTE : These plots are appropriate for examining the replation between 2 categorical variables . . . .

```r
Table1 <- table(Smoke, Gender)
```

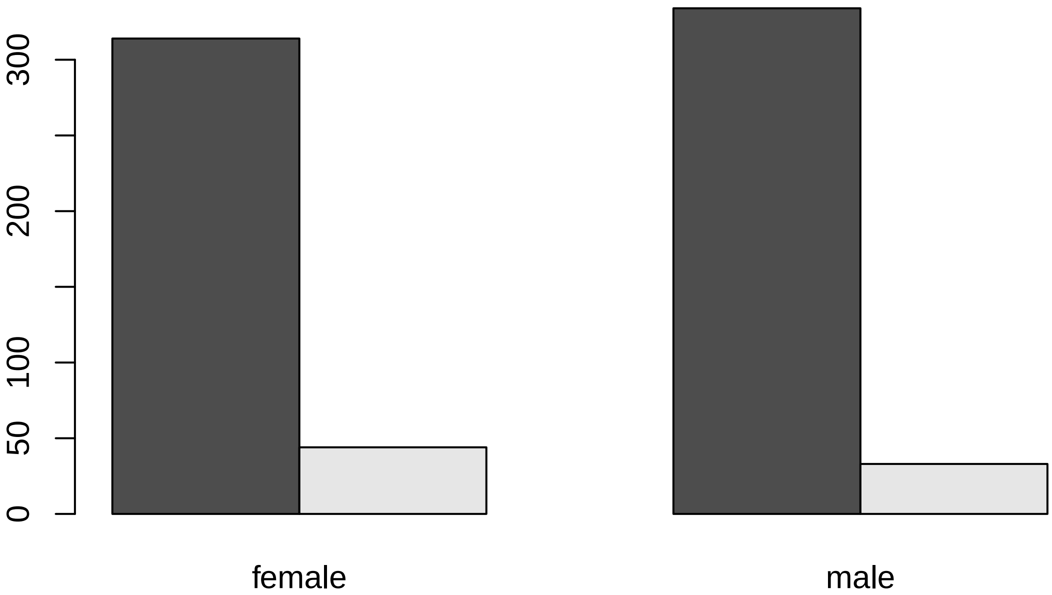
Stack is default

```r
barplot(Table1)
```



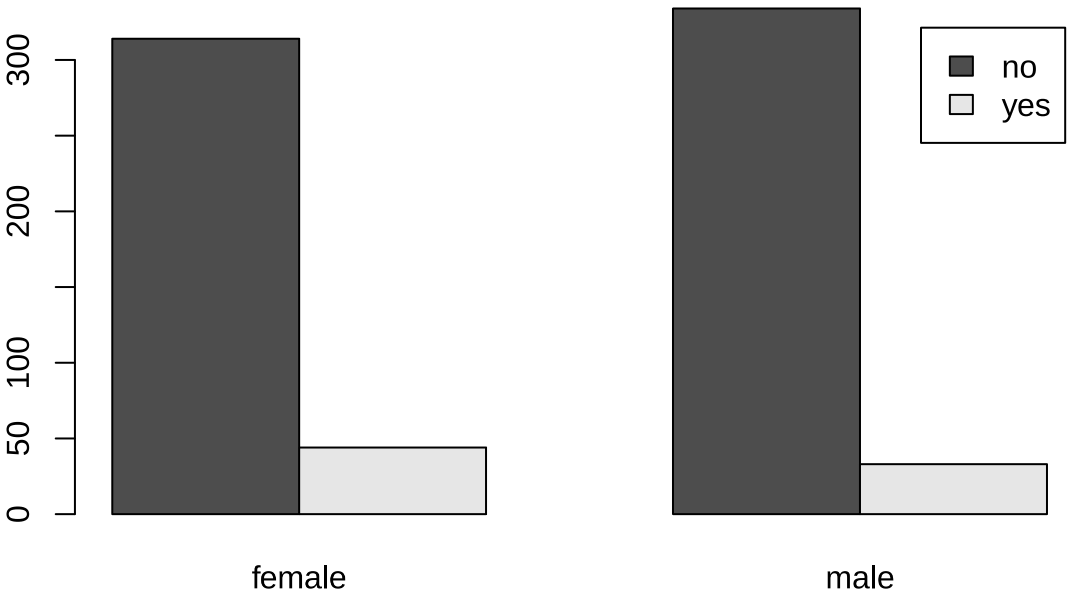
side by side plot using beside argument

```r
barplot(Table1,beside = T)
```



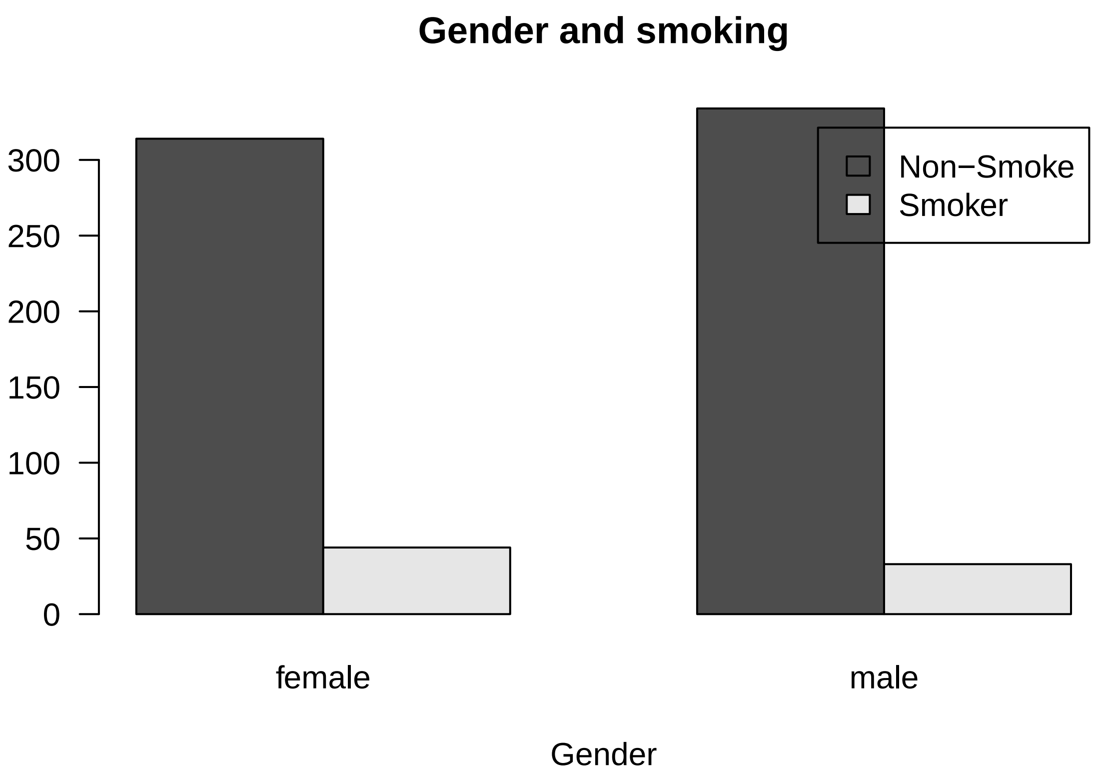
Adding Legends

```
barplot(Table1,beside = T,legend.text = T)
```



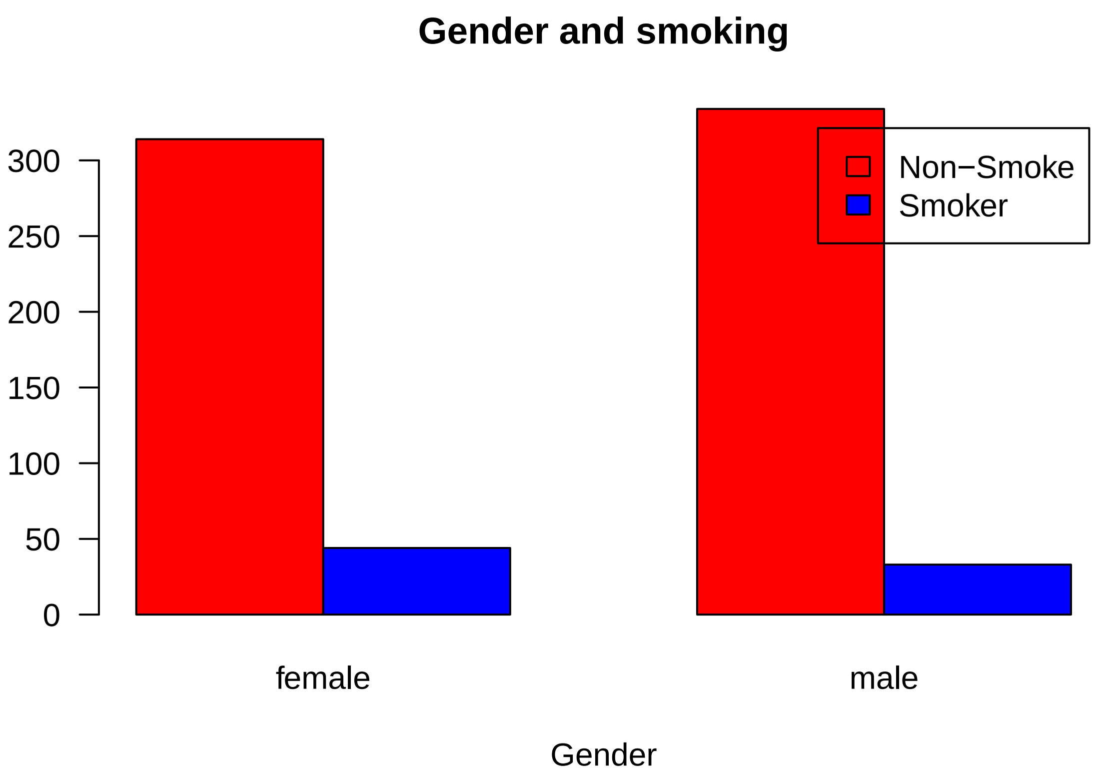
Adding title, xlab, las

```
barplot(Table1,beside = T,legend.text = c("Non-Smoke","Smoker"),main = "Gender and smoking", xlab="Gende
```



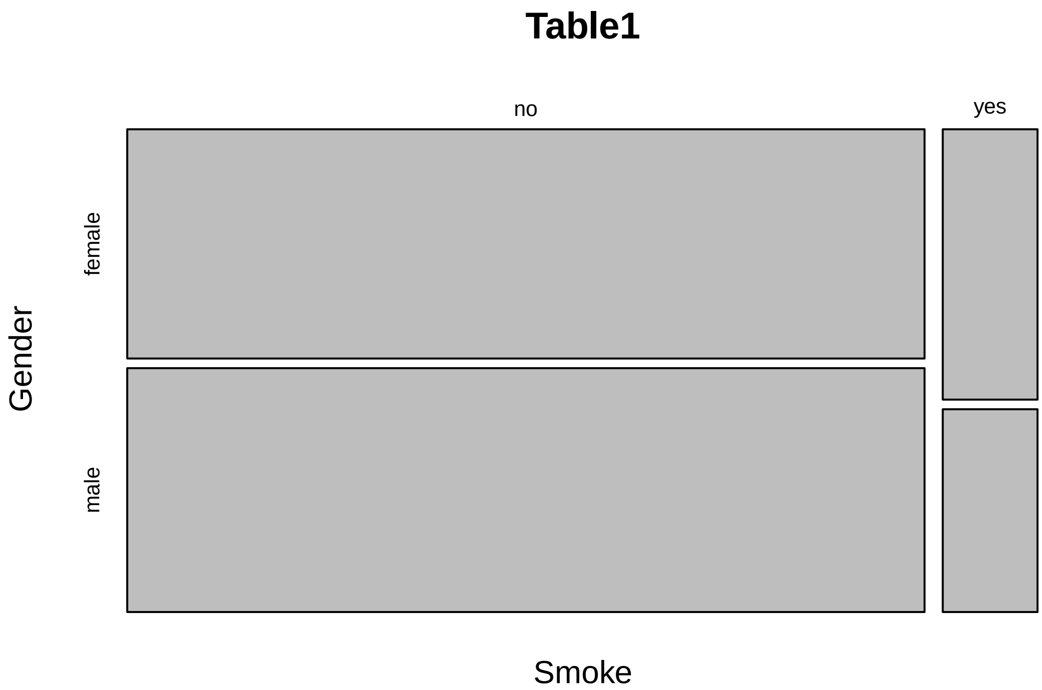
Adding color

```
barplot(Table1,beside = T,legend.text = c("Non-Smoke","Smoker"),main = "Gender and smoking", xlab="Gend
```



Mosaic Plot is one more option to find out difference between two categorical variables

```r
mosaicplot(Table1)
```



```
## [1] 2
```

Chapter 8 : Making Scatterplots

Scatter plots are appropriate for examining the replation between 2 numeric variables . . . .

Exploring the relationship between Height and Age

before producing scatterplot examine the strength of the linear relationship between the 2 numeric variables using pearson's correlation.

- Correlation is Positive when the values increase together, and
- Correlation is Negative when one value decreases as the other increases

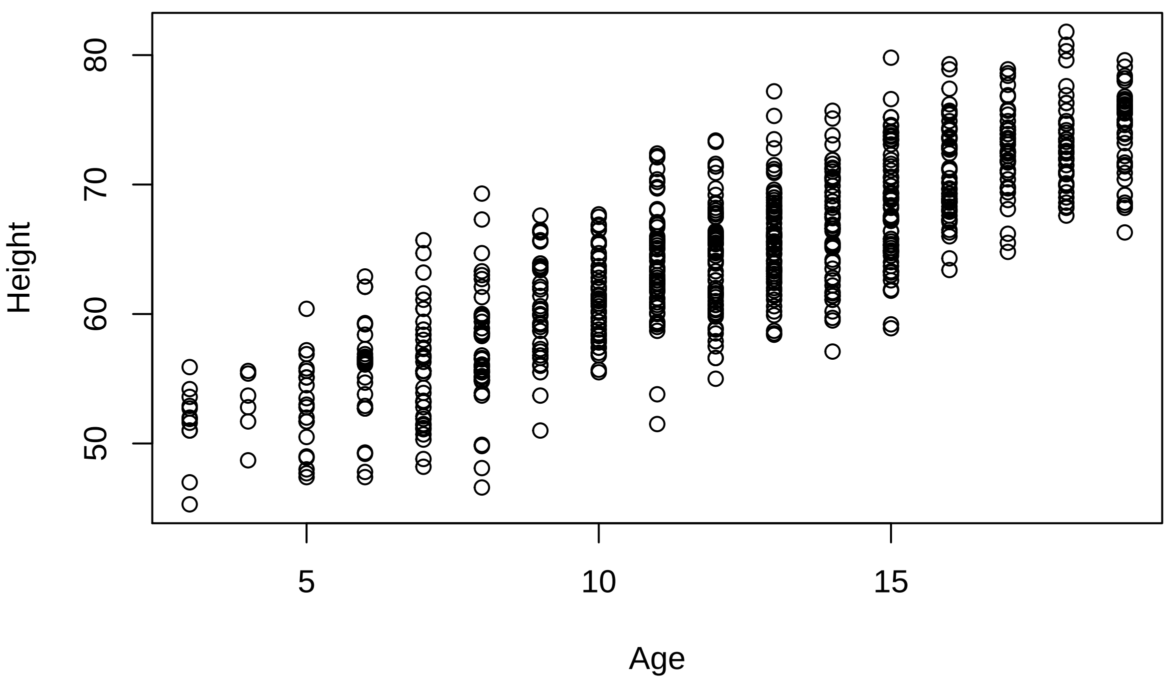
- Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

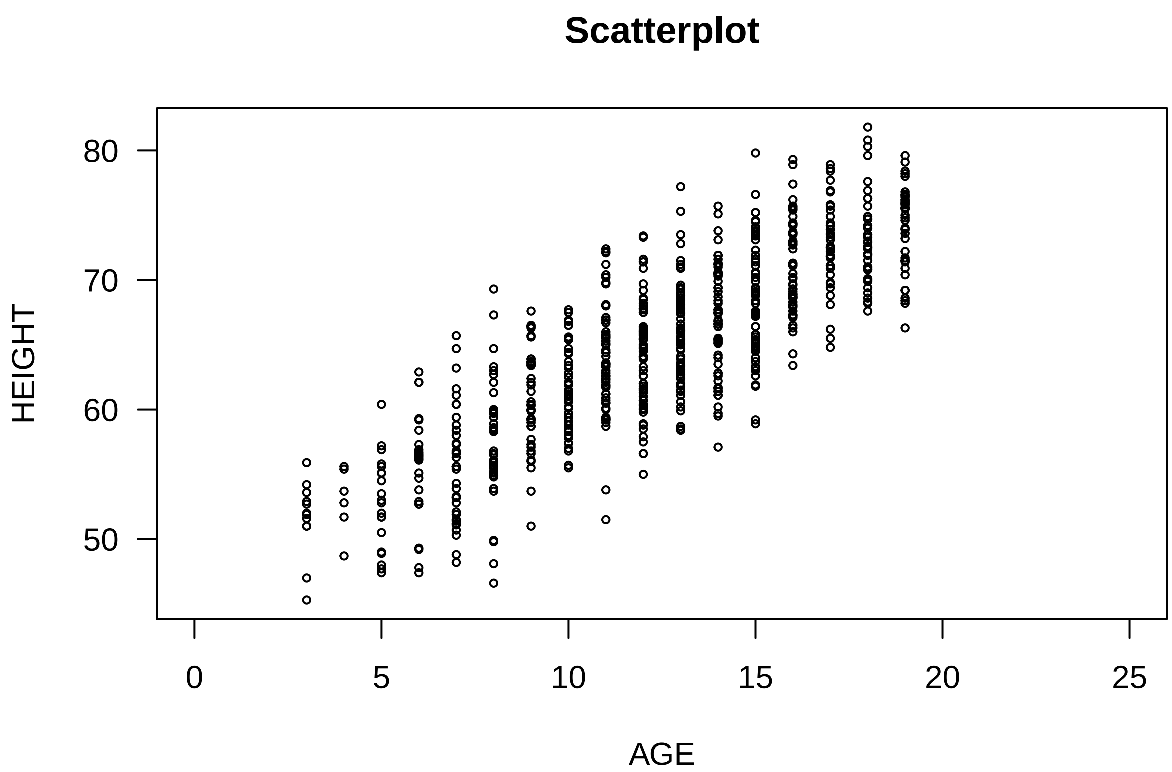
```r
cor(Age,Height)
```

```
## [1] 0.8357368
```

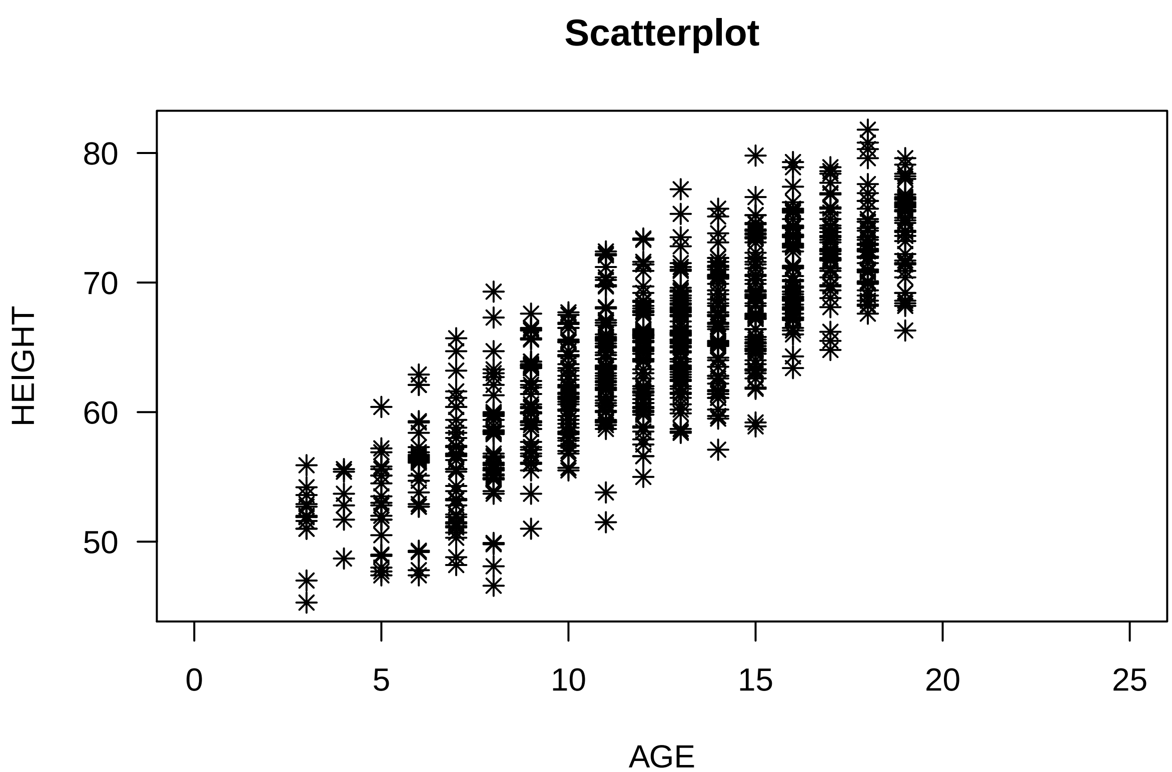
```r
plot(Age,Height)
```



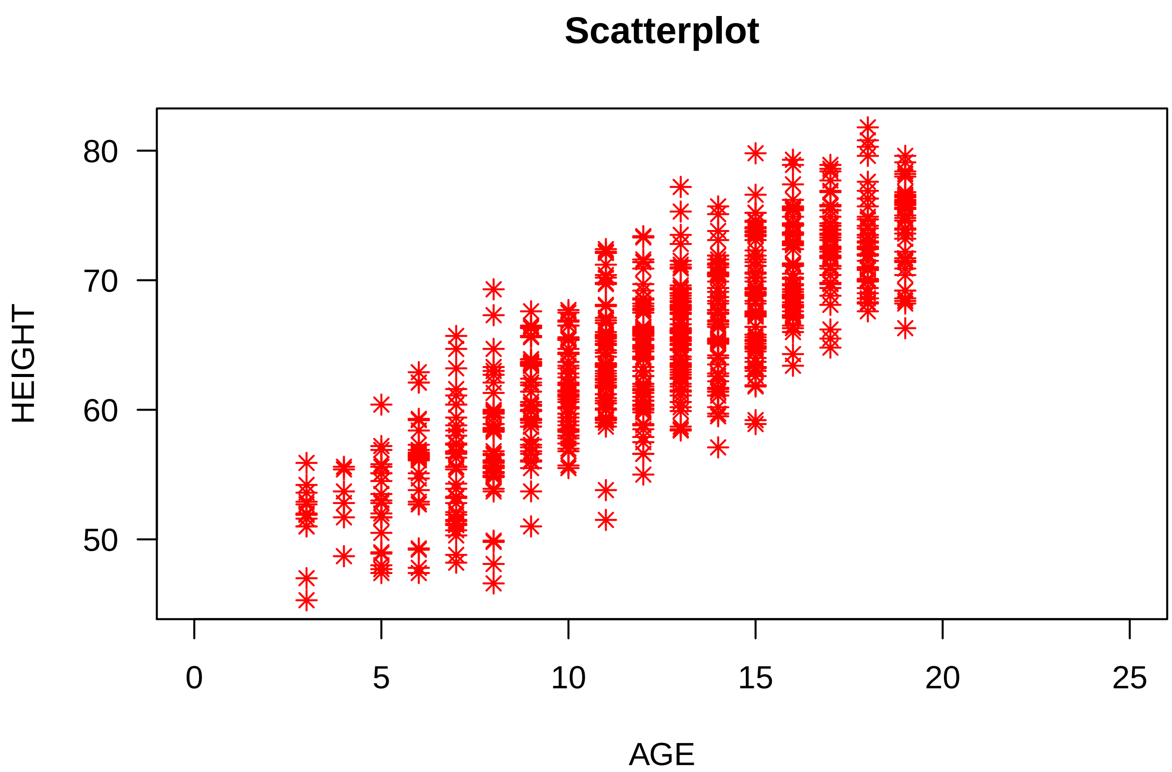

```r
plot(Age,Height,main="Scatterplot",xlab = "AGE", ylab = "HEIGHT",las=1,xlim = c(0,25),cex=0.5)
```




```
plot(Age,Height,main="Scatterplot",xlab = "AGE", ylab = "HEIGHT",las=1,xlim = c(0,25),pch=8)
```




```
plot(Age,Height,main="Scatterplot",xlab = "AGE", ylab = "HEIGHT",las=1,xlim = c(0,25),pch=8,col=2)
```



Add a line

abline(lm(Height ~Age),col=4)
lines(smoot.spline(Age,Height),lty=2,lwd=5)

```
## [1] 2
```

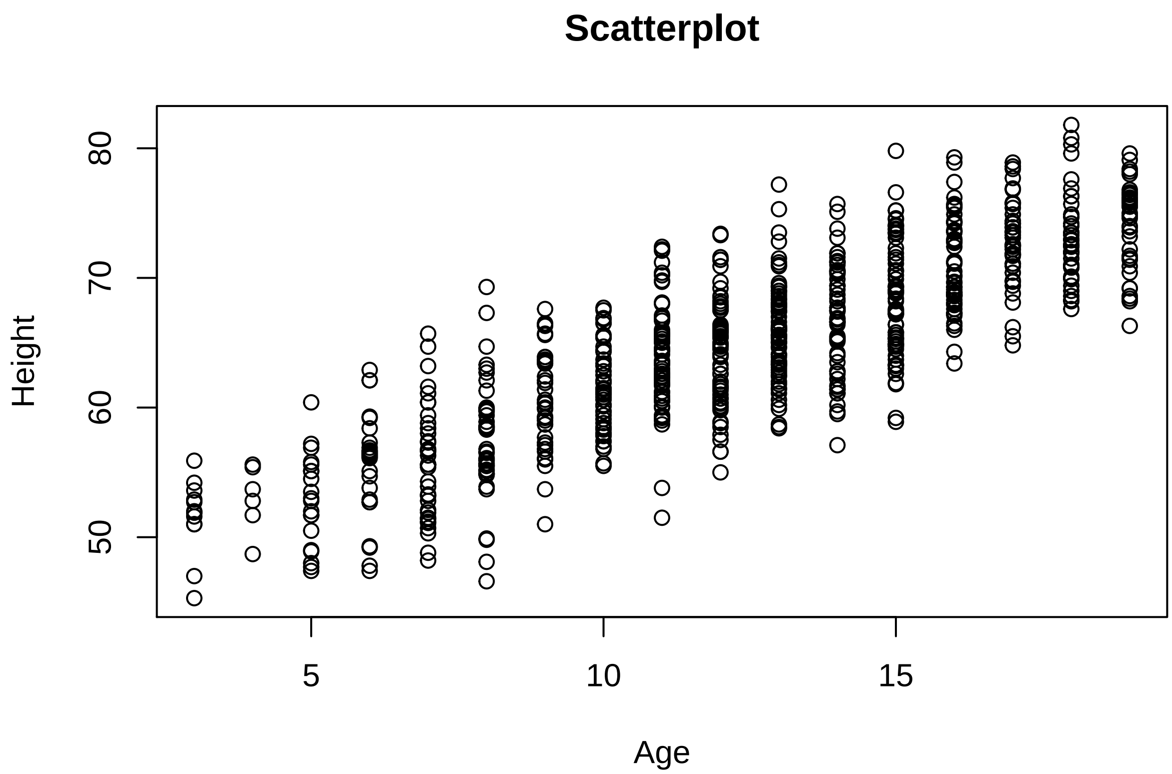
Chapter 9 : Modifying Plots

Working with scatterplot for simplicity

help par or ?par

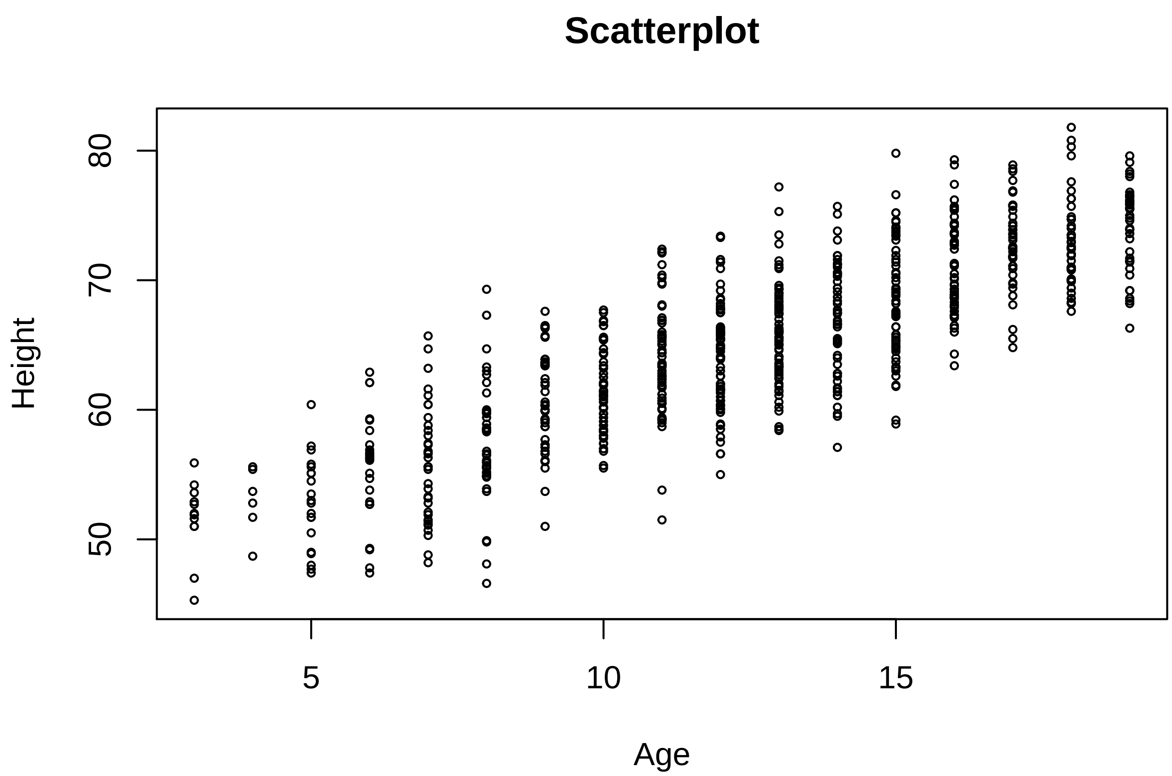
Step 1 :

```r
plot(Age, Height, main = "Scatterplot")
```



Step 2 :

```r
plot(Age, Height, main = "Scatterplot",cex = 0.5)
```



Step 3 :

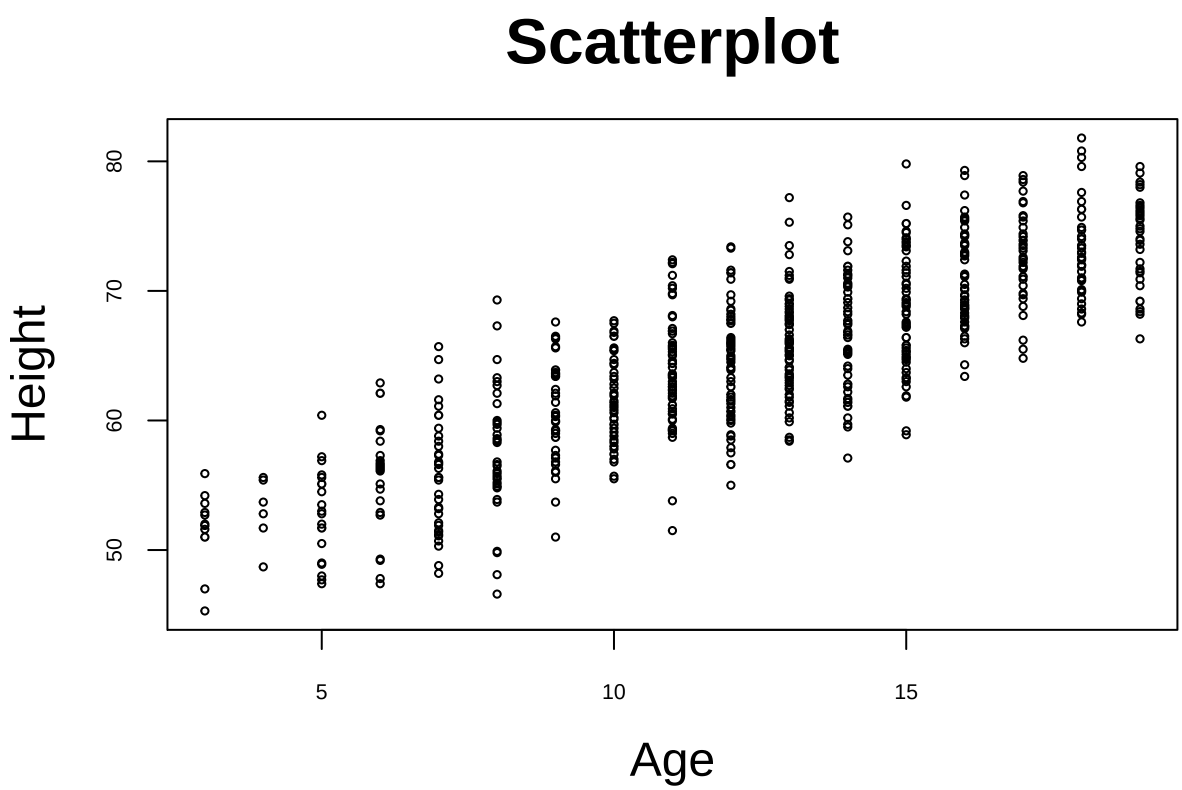
```r
plot(Age, Height, main = "Scatterplot",cex = 0.5,cex.main = 2)
```

Step 4 :

```
plot(Age, Height, main = "Scatterplot",cex = 0.5,cex.main = 2,cex.lab=1.5)
```

Step 5 :

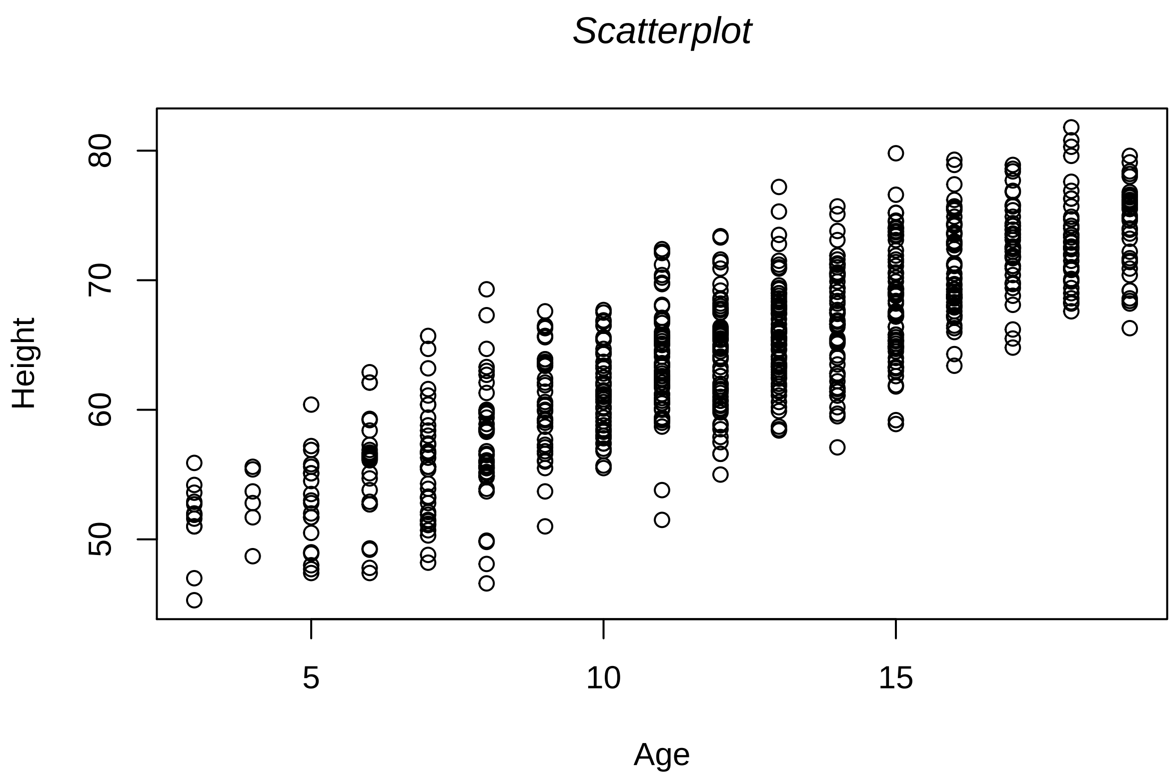
```
plot(Age, Height, main = "Scatterplot",cex = 0.5,cex.main = 2,cex.lab=1.5,cex.axis = 0.7)
```



Changing font

Step 1 : italic

```r
plot(Age, Height, main = "Scatterplot",font.main = 3)
```



Step 2 : Bold

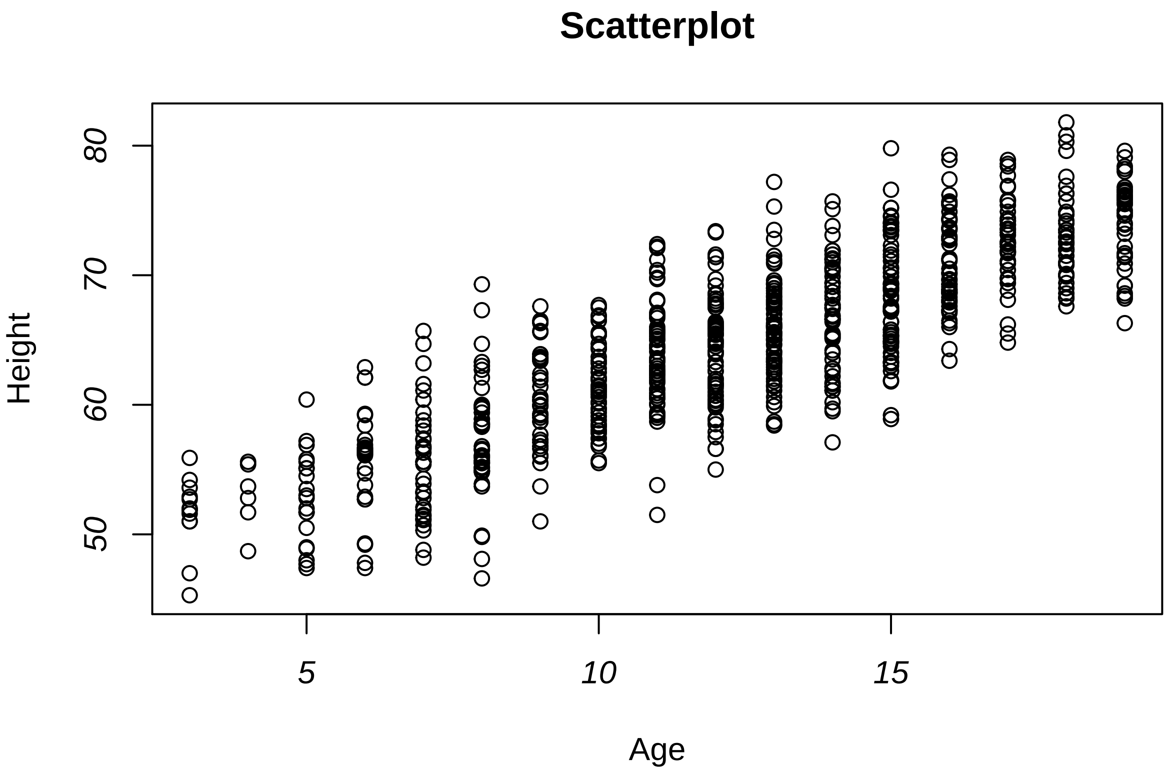
```r
plot(Age, Height, main = "Scatterplot",font.main = 4)
```

Step 3 : x & y label font

```
plot(Age, Height, main = "Scatterplot",font.main = 4,font.lab = 2)
```

Step 4 : axis

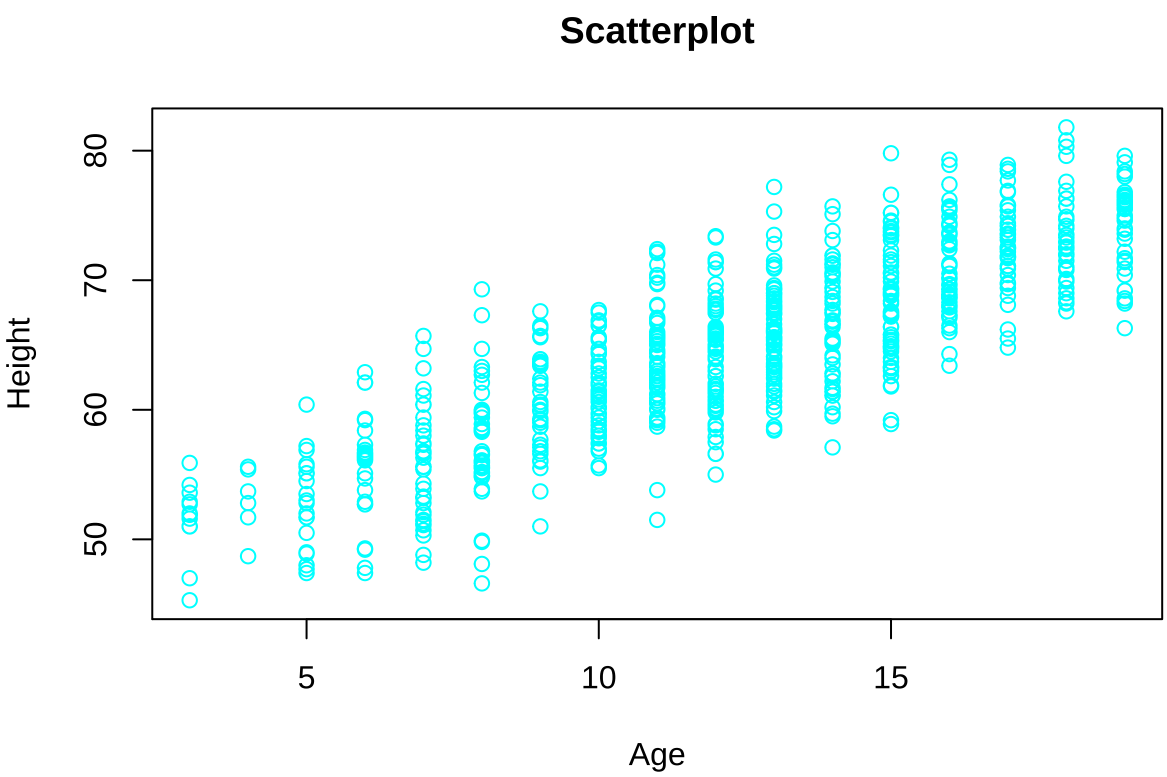
```
plot(Age, Height, main = "Scatterplot",font.axis = 3)
```



Chaning Colors on plots using “col” argument

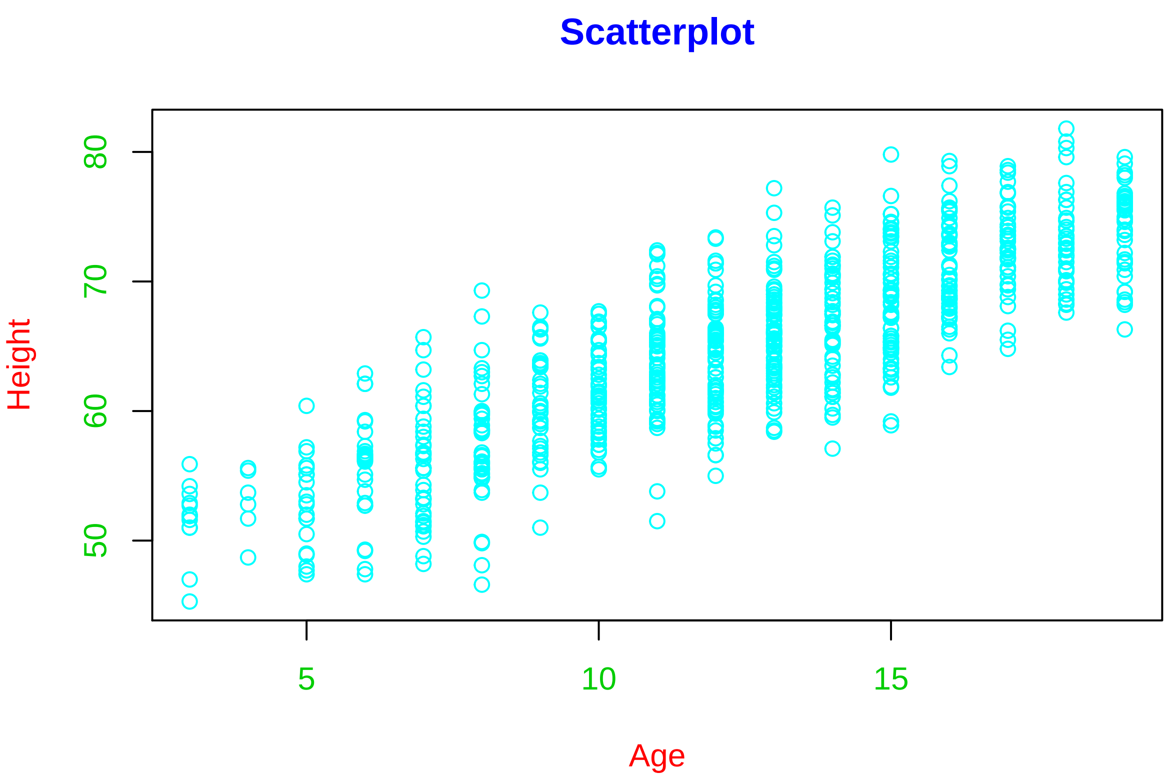
Step 1 :

```
plot(Age, Height, main = "Scatterplot",col = 5)
```



Step 2 : changing color of title,labels,and axis

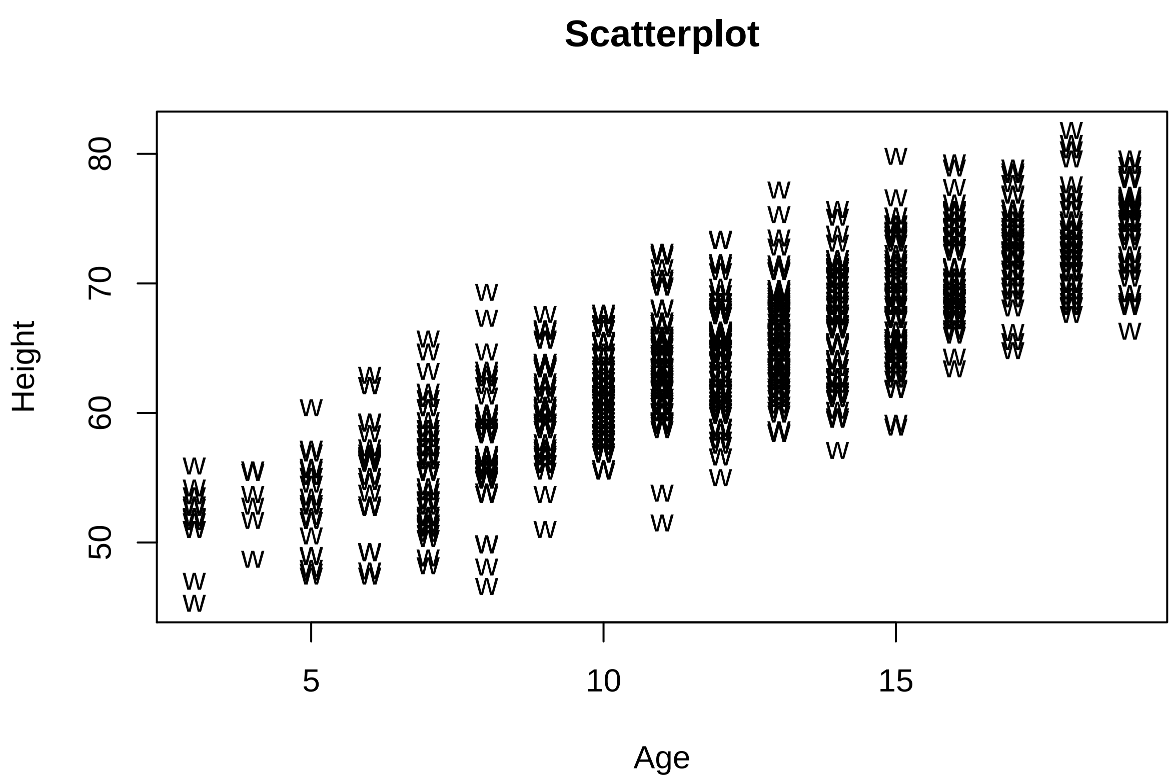
```
plot(Age, Height, main = "Scatterplot",col = 5, col.main=4,col.lab =2,col.axis =3)
```



Chaning plotting characters using "pch" argument

```
plot(Age, Height, main = "Scatterplot",pch=2)
```


```
plot(Age, Height, main = "Scatterplot",pch="w")
```

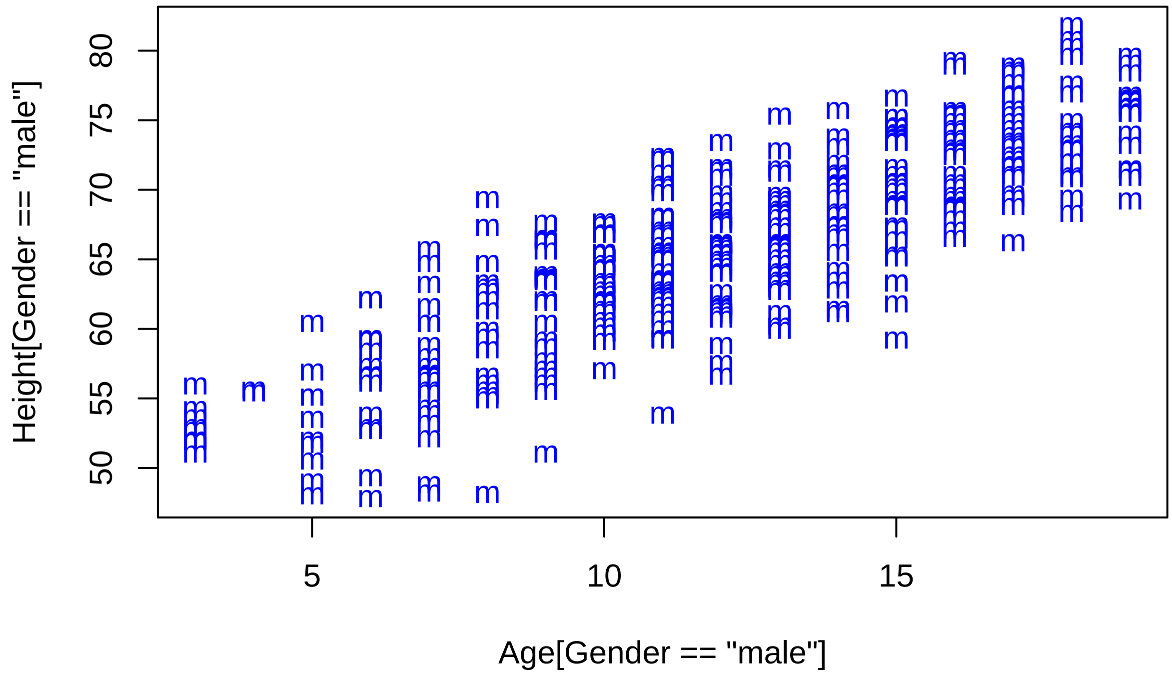


adding linear line

abline(lm(height ~ Age),col = 4,lty =2, lwd 6)

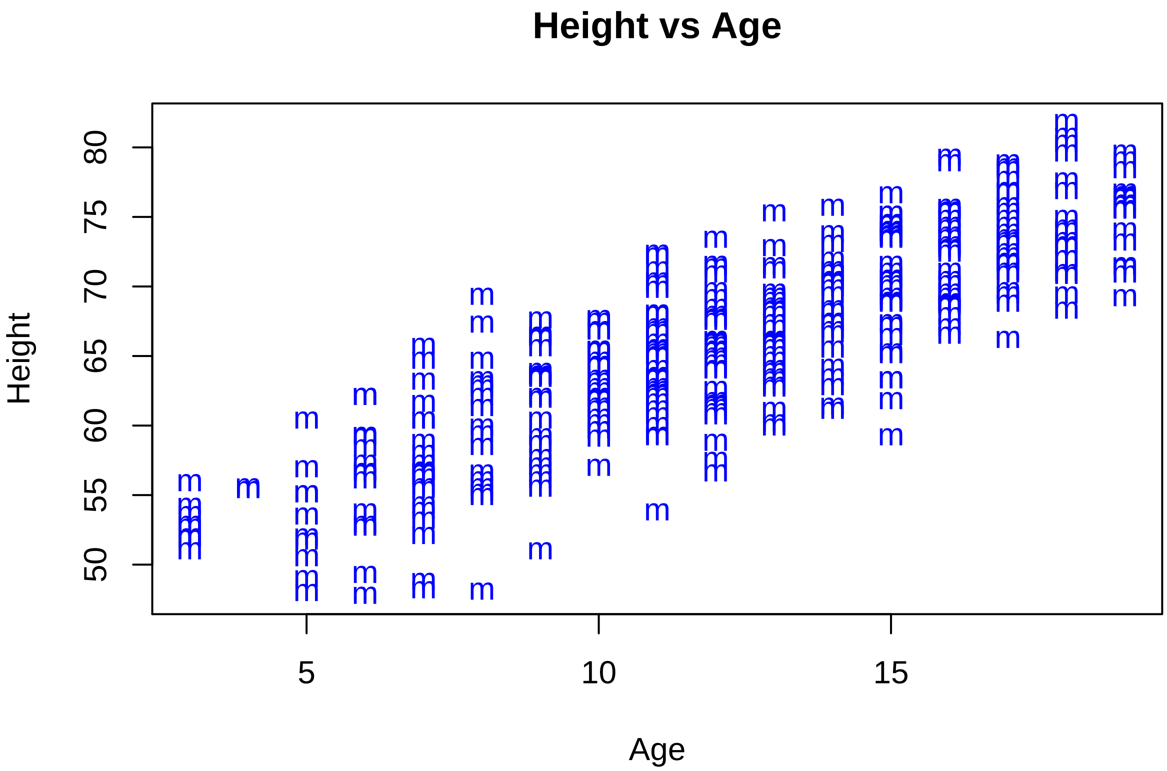
Identifying gender on the same plot using plotting characters and colours ...

```
plot(Age[Gender == "male"],Height[Gender == "male"],col = 4, pch="m")
```



relabel x & y axis

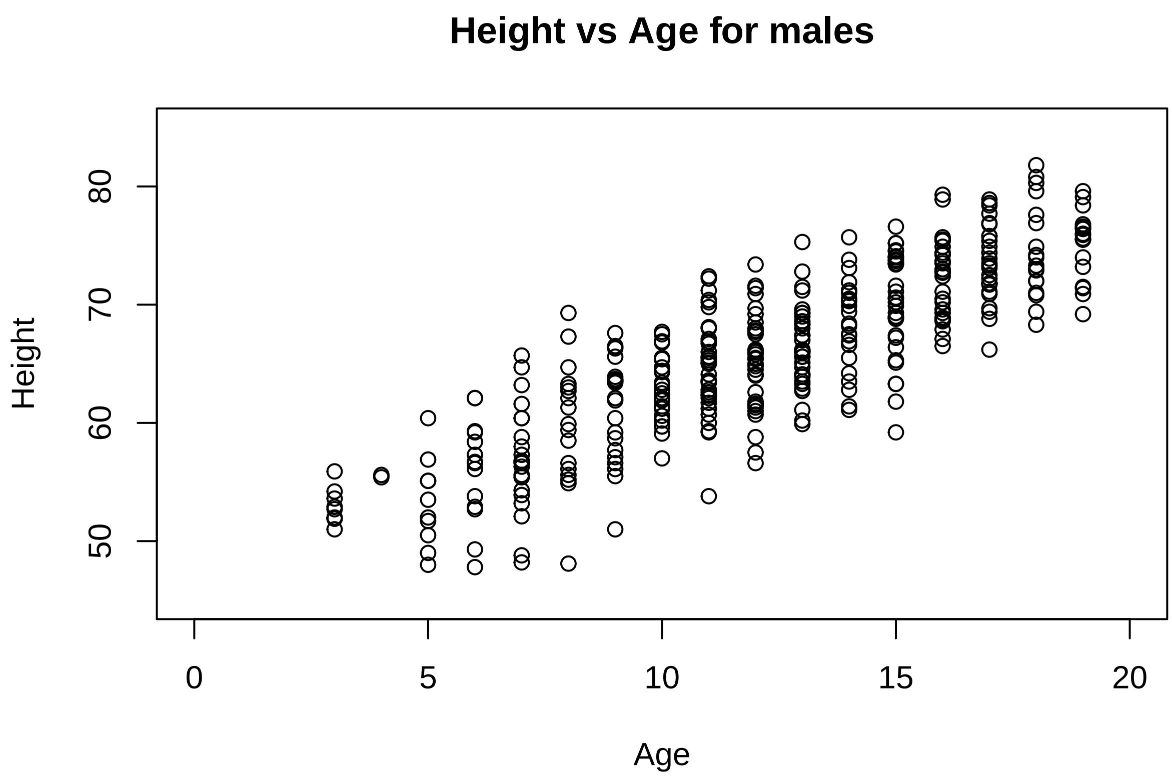
```
plot(Age[Gender == "male"],Height[Gender == "male"],col = 4, pch="m",xlab = "Age", ylab = "Height",main
```



adding female in existing plot

points(Age[Gender == "female"],Height[Gender == "female"],col=6,pch ="f") par(mfrow = c(1,2))

```
plot(Age[Gender == "male"],Height[Gender == "male"],xlab = "Age", ylab = "Height",main = "Height vs Age
```

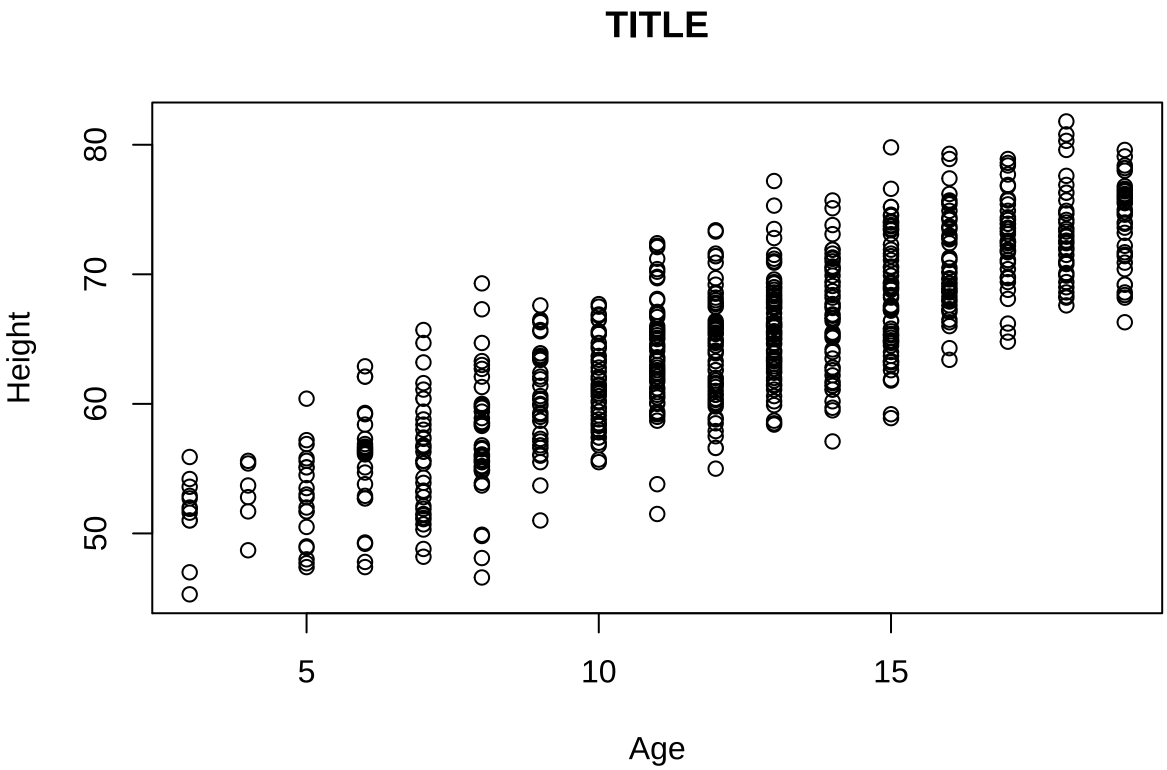



```
plot(Age[Gender == "female"],Height[Gender == "female"],xlab = "Age", ylab = "Height",main = "Height vs
```

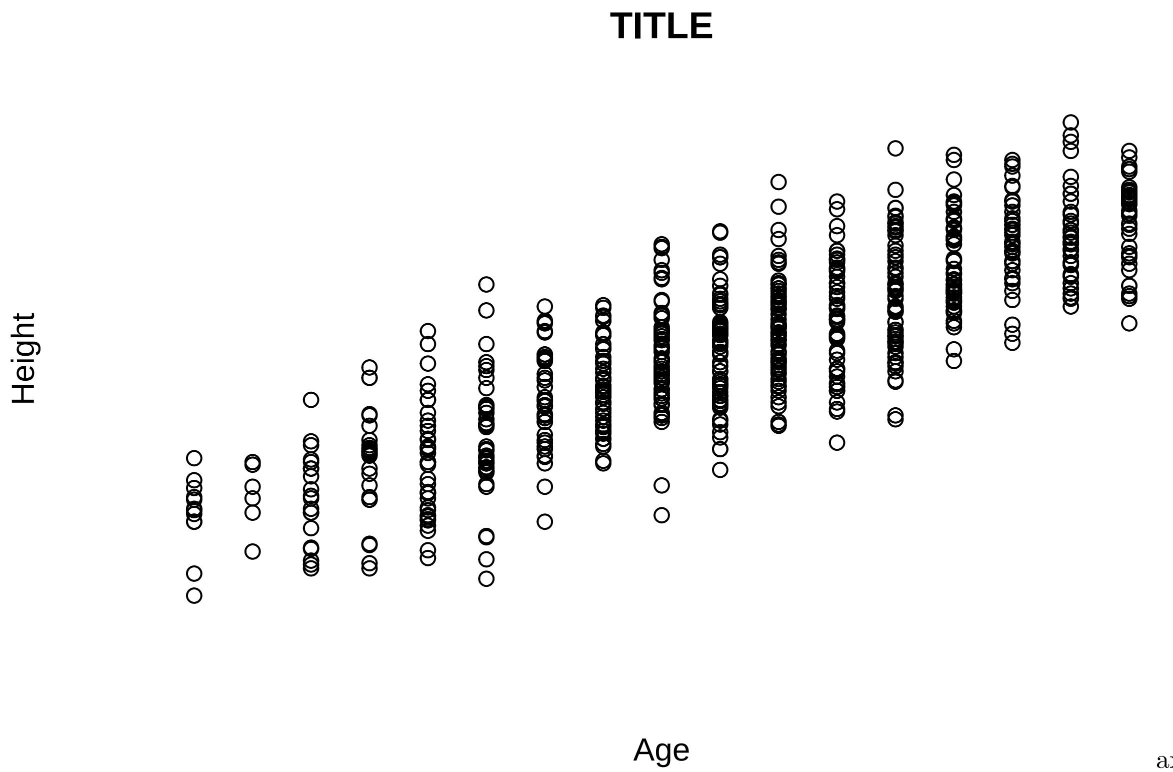
Relabelling the axis

```
par(mfrow =c(1,1))
```

```
plot(Age, Height, main= "TITLE")
```




```
plot(Age, Height, main= "TITLE",axes =F)
```



axis(side = 1, at=c(7,12.3,15), lables =c(“Sev”,“mean”,“15”))
axis(side = 2,at=c(55,65,75),labels=c(55,65,75))
box()
axis(side=4,at=c(50,60,70),labels=c(50,60,70))

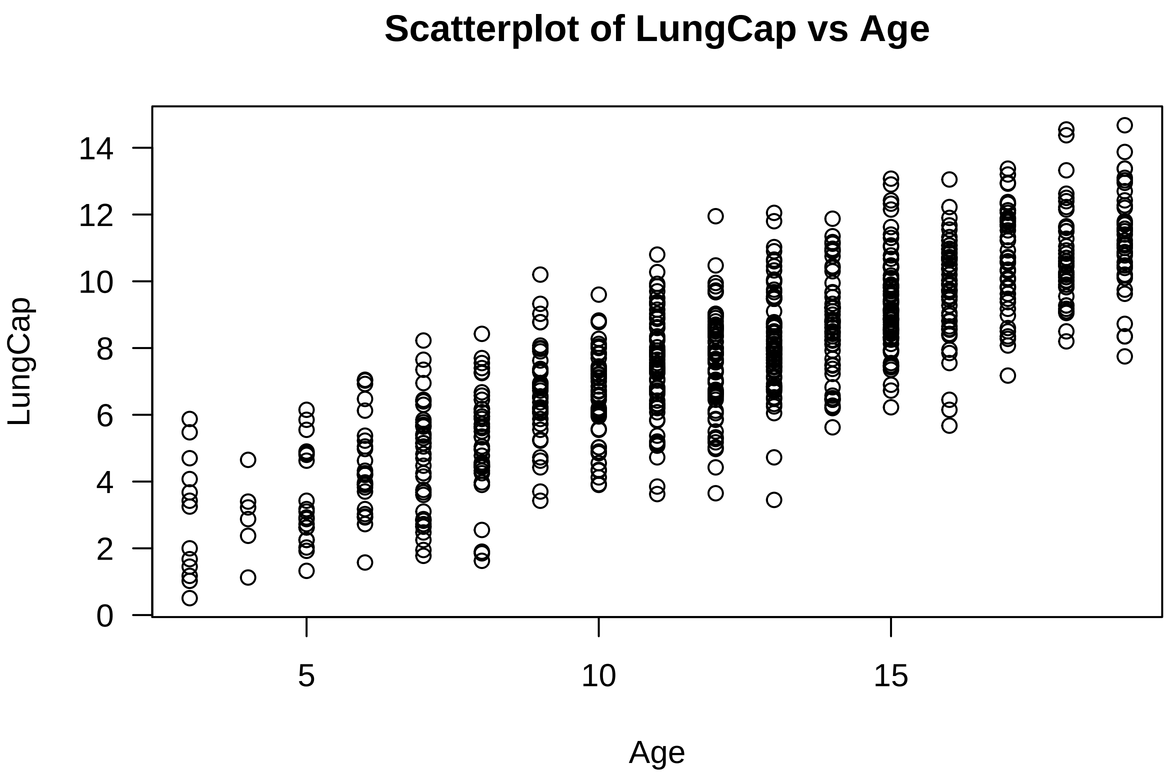
```
## [1] 2
```

Chapter 10 : Adding Text to a Plot

Often one would like to enhance an existing plot by adding some descriptive text to the plot

help(txt) or ?text

```r
plot(Age,LungCap, main= "Scatterplot of LungCap vs Age", las = 1)
```

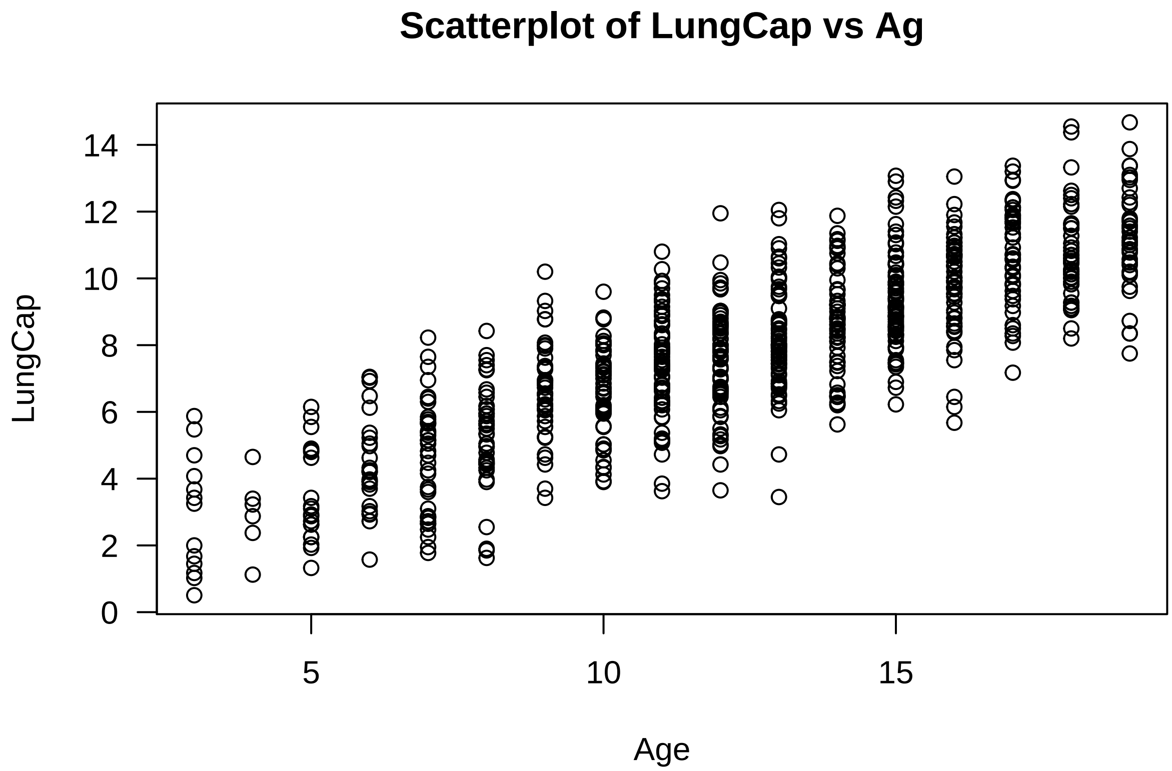



```r
cor(Age, LungCap)
```

```
## [1] 0.8196749
```

text(x=5,y=11,label="r = 0.82") text(x=5,y=11,label="r = 0.82",adj=1)

```r
plot(Age,LungCap,main="Scatterplot of LungCap vs Ag", las =1)
```



text(x=3.5,y=13,adj=0,labels= “r = 0.82”,cex =0.5,col=4)
text(x=3.5,y=13,adj=0,labels= “r = 0.82”,cex =1,col=4,font=4)

Adding horizontal line

abline(h=mean(LungCap),col=2,lwd=2)
text(x=2.5,y=8.5,adj=0,label=“Mean Lung cap”,cex = 0.65,col=2)

plot(Age,LungCap,main=“Scatterplot of LungCap vs Ag”, las =1)
mtext(text=“r = 0.82”, side = 2) –use 1,2,3,4
mtext(text=“r = 0.82”, side = 1,adj=0) –use 1,2,3,4
mtext(text=“r = 0.82”, side = 2,adj=0.75) –use 1,2,3,4
plot(Age,LungCap,main=“Scatterplot of LungCap vs Ag”, las =1)
mtext(text=“r = 0.82”,side = 3, adj=1,col=4,cex=1.25,font=4)

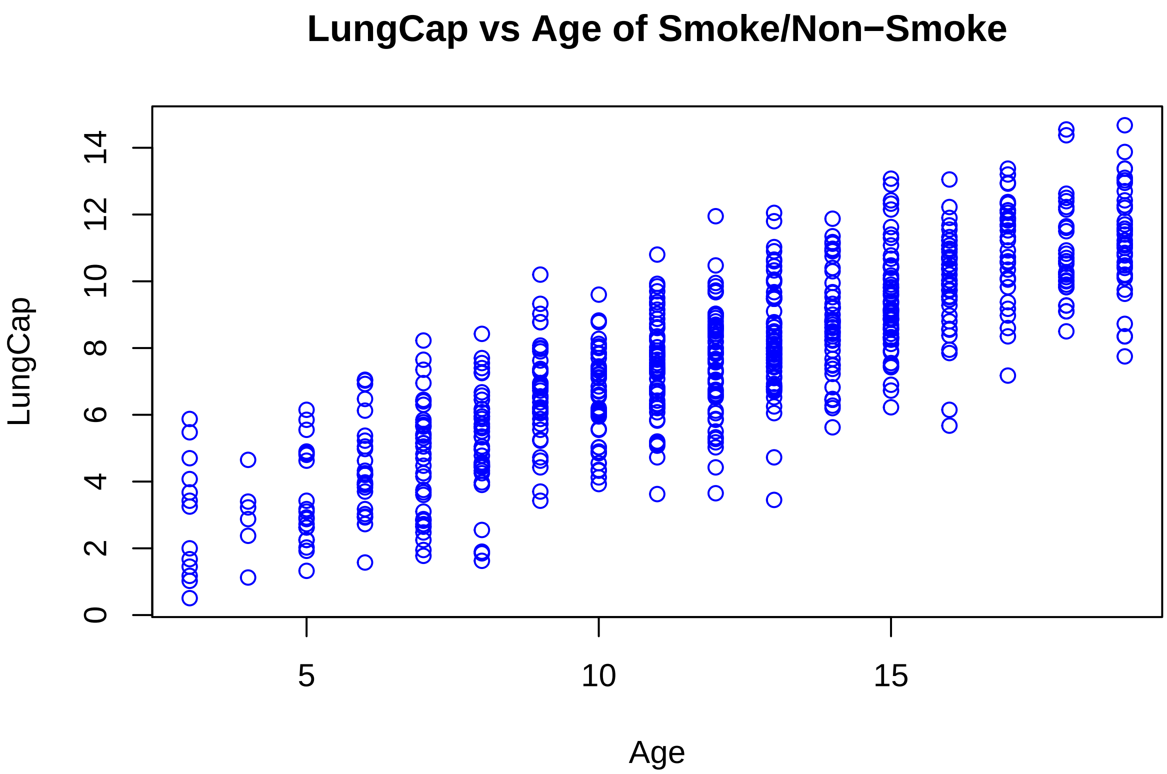
```
## [1] 2
```

Chapter 11 : Adding Legends to Plots

Quite often two or more groups of observations are displayed on a single plot; We will discuss how to add a legend to identify each set . . .

help(legend) or ?legend

```r
plot(Age[Smoke == "no"],LungCap[Smoke == "no"],main="LungCap vs Age of Smoke/Non-Smoke",col=4,xlab = "Ag
```



points(Age[Smoke==“Yes”],LungCap[Smoke == “yes”],col=2)
Adding Legends legend(x = 3.5 , y = 14, legend c(“Non-Smoke”,“Smoke”),fill=c(4,2))
points(Age[Smoke==“Yes”],LungCap[Smoke == “yes”],col=2,pch = 17)
legend(x = 3.5 , y = 14, legend c(“Non-Smoke”,“Smoke”),col=c(4,2),pch=c(16,17))
legend(x = 3.5 , y = 14, legend c(“Non-Smoke”,“Smoke”),col=c(4,2),pch=c(16,17),bty=“n”)
lines(smooth.spline(Age[Smoke == “no”],Lungcap[Smoke == “no”]),col=4,lwd=3)
lines(smooth.spline(Age[Smoke == “yes”],Lungcap[Smoke == “yes”]),col=4,lwd=2)
legend(x = 3.5 , y = 14, legend c(“Non-Smoke”,“Smoke”),col=c(4,2),lty=1,bty=“n”,lwd=3)