**K Means clustering**

head(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

dim(iris)

## [1] 150 5

names(iris)

## [1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"

sapply(iris,class)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species   
## "numeric" "numeric" "numeric" "numeric" "factor"

str(iris)

## 'data.frame': 150 obs. of 5 variables:  
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  
## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  
## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...  
## $ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

summary(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

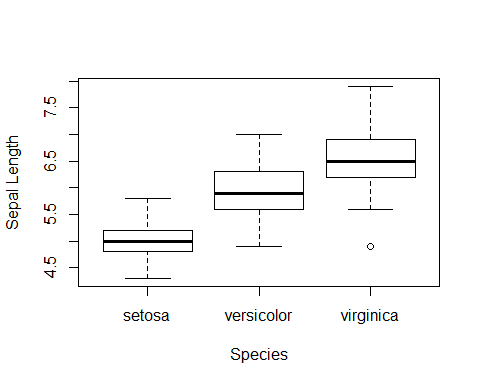
table(iris$Species)

##   
## setosa versicolor virginica   
## 50 50 50

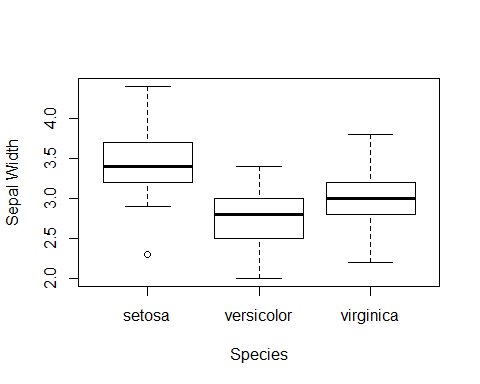
iris[ iris == "?"] <- NA  
colSums(is.na(iris))

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species   
## 0 0 0 0 0

boxplot(iris$Sepal.Length ~ iris$Species,xlab="Species", ylab="Sepal Length")

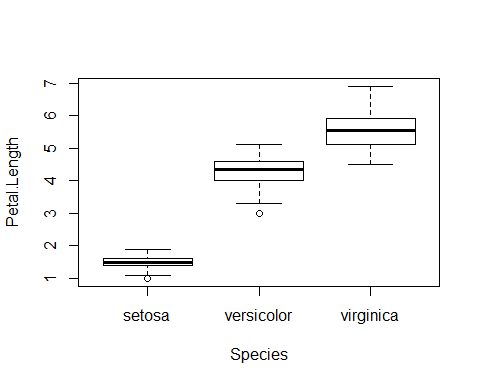
 INFERENCE: There is an overlapping of values of the sepal length of versiclor and virginica.

boxplot(iris$Sepal.Width ~ iris$Species,xlab="Species", ylab="Sepal Width")



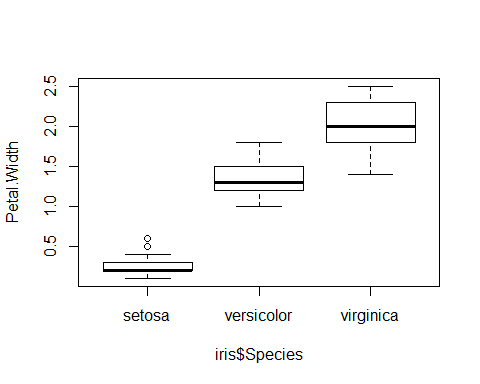
INFERENCE: There is an overlapping of values of the sepal width of versiclor and virginica.

boxplot(iris$Petal.Length ~ iris$Species,xlab="Species",ylab="Petal.Length")



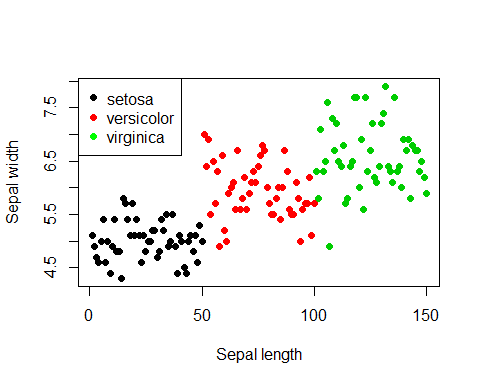
INFERENCE: The overlapping of values of the petal length of versiclor and virginica is less.

boxplot(iris$Petal.Width ~ iris$Species,ylab="Petal.Width")

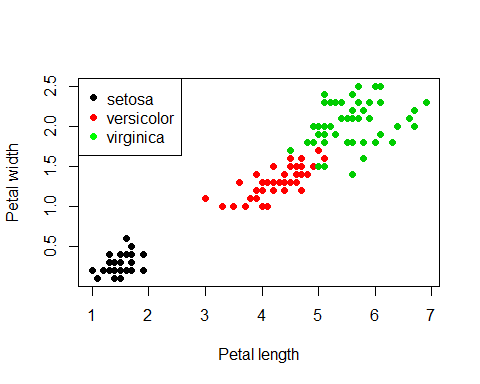
 INFERENCE: The overlapping of values of the petal width of versiclor and virginica is less.

Petal descriptors show a clearer difference compared to sepal descriptors.

plot(iris$Sepal.Length,iris$Sepaal.Width,col=iris$Species,pch=16, xlab="Sepal length", ylab="Sepal width")  
 legend( x="topleft",   
 legend=levels(as.factor(iris$Species)),  
 col=c("black","red","green"),   
 pch=c(16) )



plot(iris$Petal.Length,iris$Petal.Width,col=iris$Species,pch=16, xlab="Petal length", ylab="Petal width")  
 legend( x="topleft",   
 legend=levels(as.factor(iris$Species)),  
 col=c("black","red","green"),   
 pch=c(16) )



#head(iris[,-5])  
iris\_cor=cor(iris[,-5], method = c("pearson", "kendall", "spearman"))  
iris\_cor

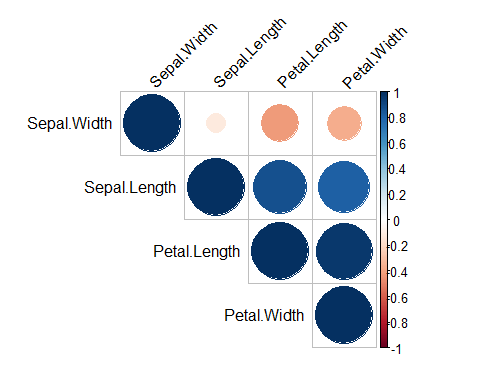
## Sepal.Length Sepal.Width Petal.Length Petal.Width  
## Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411  
## Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259  
## Petal.Length 0.8717538 -0.4284401 1.0000000 0.9628654  
## Petal.Width 0.8179411 -0.3661259 0.9628654 1.0000000

library(corrplot)

## Warning: package 'corrplot' was built under R version 3.6.2

## corrplot 0.84 loaded

corrplot(iris\_cor, type = "upper", order = "hclust",   
 tl.col = "black", tl.srt = 45)



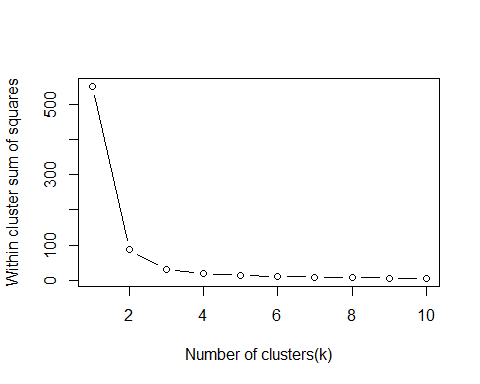
#The correlation matrix is reordered according to the correlation #coefficient using “hclust” method.  
#tl.col (for text label color) and tl.srt (for text label string rotation) #are used to change text colors and rotations.  
#Possible values for the argument type are : “upper”, “lower”, “full”

INFERENCE: Petal length and petal width show a very high positive correlation. Correlation between sepal length and petal length, sepal length and petal width show a high positive correlation.

set.seed(200)  
k.max <- 10  
wss<- sapply(1:k.max,function(k){kmeans(iris[,3:4],k,nstart = 20,iter.max = 20)$tot.withinss})  
wss

## [1] 550.895333 86.390220 31.371359 19.465989 13.916909 11.025145  
## [7] 9.236596 7.674414 6.456495 5.550520

plot(1:k.max,wss, type= "b", xlab = "Number of clusters(k)", ylab = "Within cluster sum of squares")

 INFERENCE: From the graph we conclude that optimal number of clusters is 3 (elbow point). We build a model for k=3, i.e. three clusters.

icluster <- kmeans(iris[,3:4],3,nstart = 20)  
icluster

## K-means clustering with 3 clusters of sizes 50, 52, 48  
##   
## Cluster means:  
## Petal.Length Petal.Width  
## 1 1.462000 0.246000  
## 2 4.269231 1.342308  
## 3 5.595833 2.037500  
##   
## Clustering vector:  
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [75] 2 2 2 3 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 2 3 3 3 3  
## [112] 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3  
## [149] 3 3  
##   
## Within cluster sum of squares by cluster:  
## [1] 2.02200 13.05769 16.29167  
## (between\_SS / total\_SS = 94.3 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

table(icluster$cluster,iris$Species)

##   
## setosa versicolor virginica  
## 1 50 0 0  
## 2 0 48 4  
## 3 0 2 46

INFERNCE: 2 of the versicolor have been put in the cluster with all the virginica and 4 of the verginica have been put in cluster 3 which mostly has versicolor.

library(ggplot2)  
icluster$cluster <- as.factor(icluster$cluster)  
ggplot(iris, aes(Petal.Length, Petal.Width, color = icluster$cluster)) + geom\_point()

