# check dimensions of data set

dim(iris)

# [1] 150 5

names(iris)  
#[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"

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 ...

# now we remove the 'Species' column from the data  
# this is to make sure the data is unlabeled

iris\_numerical =iris[,-5]  
str(iris\_numerical)

#'data.frame': 150 obs. of 4 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 ...

# iris dataset contains three classes so we set the cluster number to 3  
# use kmeans clustering for 3 clusters

kmeansClustering = kmeans(iris\_numerical,3)

kmeansClustering

# K-means clustering with 3 clusters of sizes 50, 62, 38

# Cluster means:  
# Sepal.Length Sepal.Width Petal.Length Petal.Width  
# 1 5.006000 3.428000 1.462000 0.246000  
# 2 5.901613 2.748387 4.393548 1.433871  
# 3 6.850000 3.073684 5.742105 2.071053

# 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  
# [30] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 2 2 2 2 2  
# [59] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2  
# [88] 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3 2 3 3 3 3 3 3 2 2 3  
# [117] 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 3 3 3 3 2 3 3 3 2 3 3  
# [146] 3 2 3 3 2

# Within cluster sum of squares by cluster:  
# [1] 15.15100 39.82097 23.87947  
# (between\_SS / total\_SS = 88.4 %)

# Available components:

# [1] "cluster" "centers" "totss" "withinss"  
# [5] "tot.withinss" "betweenss" "size" "iter"  
# [9] "ifault"

# there are four dimensions to the data but we use only two  
# plot Sepal length & width and colour as per cluster  
# membership

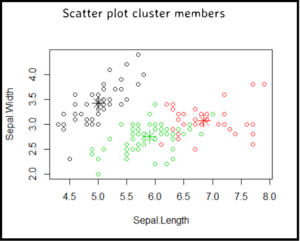
plot(iris\_numerical[c("Sepal.Length","Sepal.Width")],  
col = kmeansClustering$cluster)

#plot cluster centers for the clusters

points(kmeansClustering$centers[,c("Sepal.Length", "Sepal.Width")], col =1:3, pch=8, cex=2)

## Interpret cluster members scatter plot

Some red points are closer to the green center, this is due to remaining two dimensions viz. Petal.Length and Petal.Widthwhich are not included in the scatter plot.



#Lets compare the kmeans cluster with the Species classes

table(iris$Species, kmeansClustering$cluster)

#                             1          2         3  
# setosa                 50       0        0  
# versicolor          0         48      2  
# virginica            0         14       36

# cluster setosa can be separated from other clusters  
# clusters versicolor and virginica have a small degree of overlap

# now lets apply hierarchical clustering  
# we take a random sample of 40 records to avoid overcrowding

index = sample(1:nrow(iris\_numerical), 40)  
index  
iris\_numerical\_40 = iris\_numerical[index,]  
iris\_numerical\_40

# calculate Ecuclidean distance

distance = dist(iris\_numerical\_40)  
distance

print(distance, digits=3)

iris\_numerical\_40\_hc = hclust(distance)

iris\_numerical\_40\_hc

# Call:  
# hclust(d = distance)

# Cluster method : complete  
# Distance : euclidean  
# Number of objects: 40

# plot dendogram for hierarchical clustering

plot(iris\_numerical\_40\_hc)

# plot(iris\_numerical\_40\_hc, hang=-1, labels=iris$Species[index])

plot(iris\_numerical\_40\_hc,   
                              labels=iris$Species[index],  
                                                cex=0.6)

## Interpret cluster member dendogram

Similar to k-means cluster here also we can see that setosa can be separated from remaining two clusters, clusters virginca and versicolor are overlapping

