Clustering Notes

library(datasets)  
library(tidyverse)

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --  
## v dplyr 1.1.2 v readr 2.1.4  
## v forcats 1.0.0 v stringr 1.5.0  
## v ggplot2 3.4.2 v tibble 3.2.1  
## v lubridate 1.9.2 v tidyr 1.3.0  
## v purrr 1.0.1   
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(caret)

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(grid)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(rattle)

## Loading required package: bitops  
## Rattle: A free graphical interface for data science with R.  
## Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.

library(MASS)

##   
## Attaching package: 'MASS'  
##   
## The following object is masked from 'package:dplyr':  
##   
## select

library(cluster)  
library(e1071)  
library(kernlab)

##   
## Attaching package: 'kernlab'  
##   
## The following object is masked from 'package:purrr':  
##   
## cross  
##   
## The following object is masked from 'package:ggplot2':  
##   
## alpha

library(caTools)  
library(factoextra)

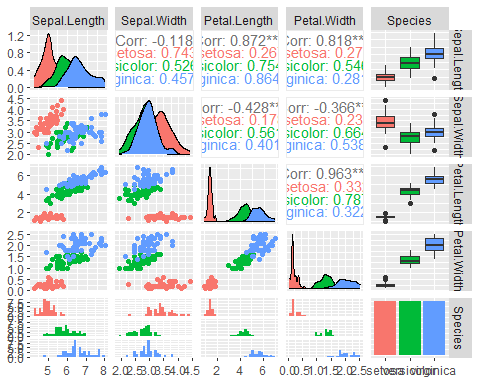
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

myiris = data.frame(iris)  
head(myiris)

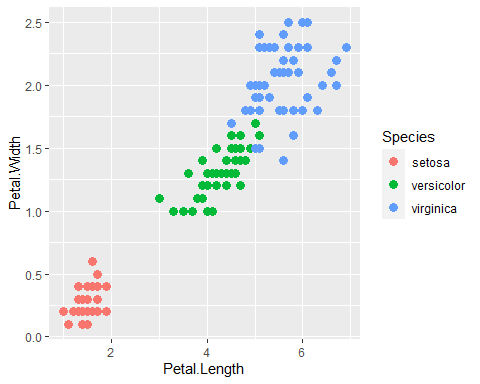
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

ggpairs(myiris, aes(color = Species))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggplot(data=myiris,aes(x=Petal.Length,y=Petal.Width,color=Species)) +geom\_point(size=3)

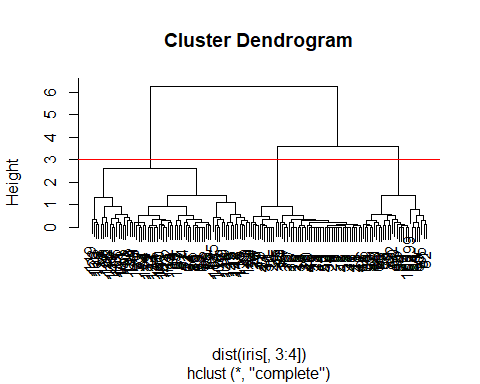


## Hierarchal clustering

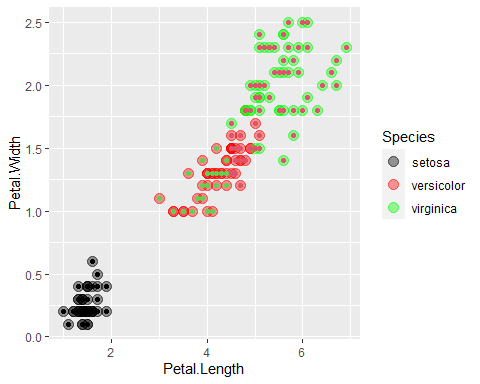
clusters <- hclust(dist(iris[, 3:4]), method = "complete")  
plot(clusters)  
clusterCut <- cutree(clusters, 3)  
table(clusterCut, iris$Species)

##   
## clusterCut setosa versicolor virginica  
## 1 50 0 0  
## 2 0 21 50  
## 3 0 29 0

abline(h=3,col="red")



ggplot(iris, aes(Petal.Length, Petal.Width, color = Species)) +   
 geom\_point(alpha = 0.4, size = 3.5) + geom\_point(col = clusterCut) +   
 scale\_color\_manual(values = c('black', 'red', 'green'))



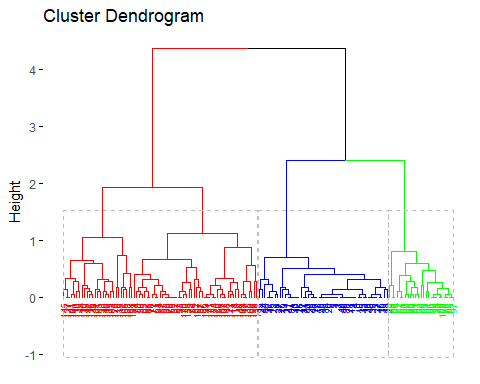
# extra hierarchal clustering

h2 = hcut(iris[,3:4], k=3,hc\_func = "hclust",stand=TRUE, hc\_method = "complete")  
h2

##   
## Call:  
## stats::hclust(d = x, method = hc\_method)  
##   
## Cluster method : complete   
## Distance : euclidean   
## Number of objects: 150

fviz\_dend(h2,rect=TRUE,cex=0.5,k\_colors=c("red", "blue", "green"))

## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as  
## of ggplot2 3.3.4.  
## i The deprecated feature was likely used in the factoextra package.  
## Please report the issue at <https://github.com/kassambara/factoextra/issues>.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



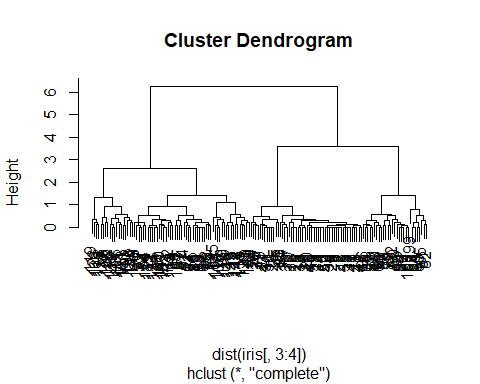
h2$cluster

## [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 3 2 2 2 3 2 3 3 2 3 2 3 2 2 3 2 3 2 3 2 2  
## [75] 3 2 2 2 2 3 3 3 3 2 2 2 2 2 3 3 3 2 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2  
## [112] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [149] 2 2

h2$size

## [1] 50 75 25

h2c=hclust(dist(iris[,3:4]),method="complete")  
plot(h2c)



## K-mean Clustering

set.seed(3920)  
kmi = kmeans(iris[,3:4],centers = 3, nstart = 15)  
kmi

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

kmi$tot.withinss

## [1] 31.37136

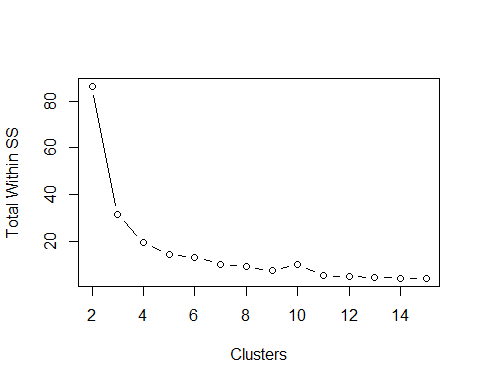
sum(kmi$withinss)

## [1] 31.37136

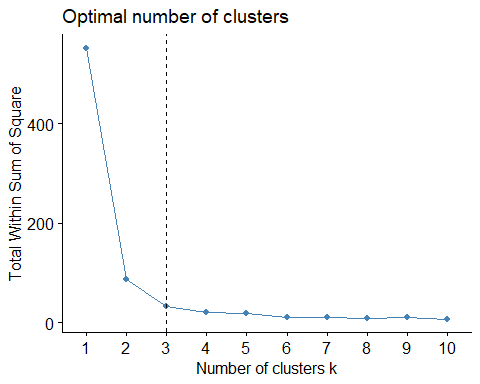
table(kmi$cluster,iris$Species)

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

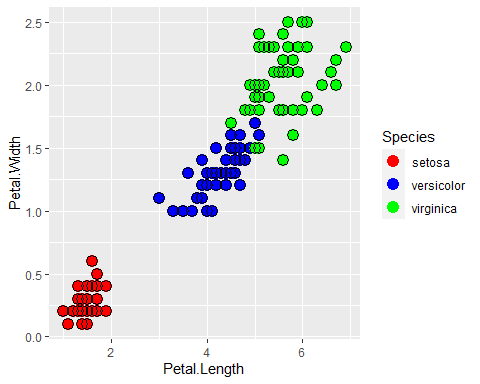
totwss = NULL  
for(i in 2:15){  
 totwss = append(totwss,kmeans(iris[,3:4], centers=i)$tot.withinss)  
}  
plot(x=2:15,y=totwss, type="b",xlab="Clusters",ylab="Total Within SS")



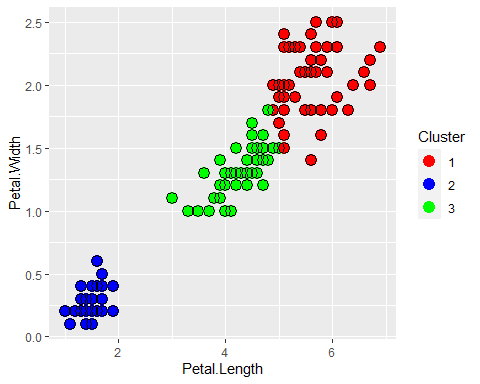
fviz\_nbclust(iris[,3:4], kmeans,method="wss") + geom\_vline(xintercept = 3, linetype=2)



ggplot(iris,aes(x=Petal.Length, y=Petal.Width,color=Species))+geom\_point(size=4)+scale\_color\_manual(values=c("red","blue","green")) +  
 geom\_point(shape=1,size=4,color="black")



ggplot(iris,aes(x=Petal.Length, y=Petal.Width,color=as.factor(kmi$cluster)))+geom\_point(size=4)+scale\_color\_manual(name="Cluster",values=c("red","blue","green")) +  
 geom\_point(shape=1,size=4,color="black")



nkm = cbind(myiris,cluster=kmi$cluster)  
class(nkm[,-5])

## [1] "data.frame"

fviz\_cluster(kmi, nkm[,-5], geom="point",show.clust.cent = TRUE,pointsize = 2)

