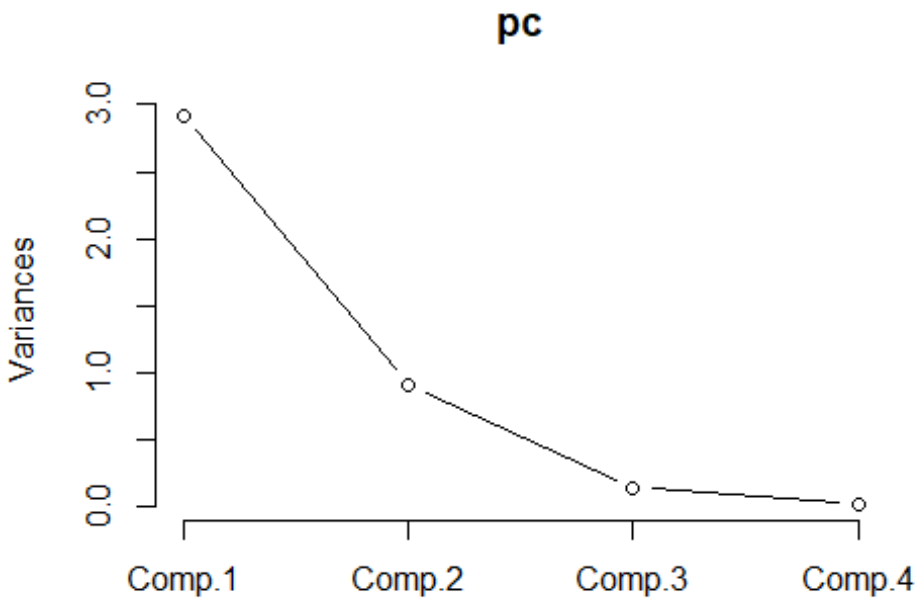
Barrett_quiz2.R

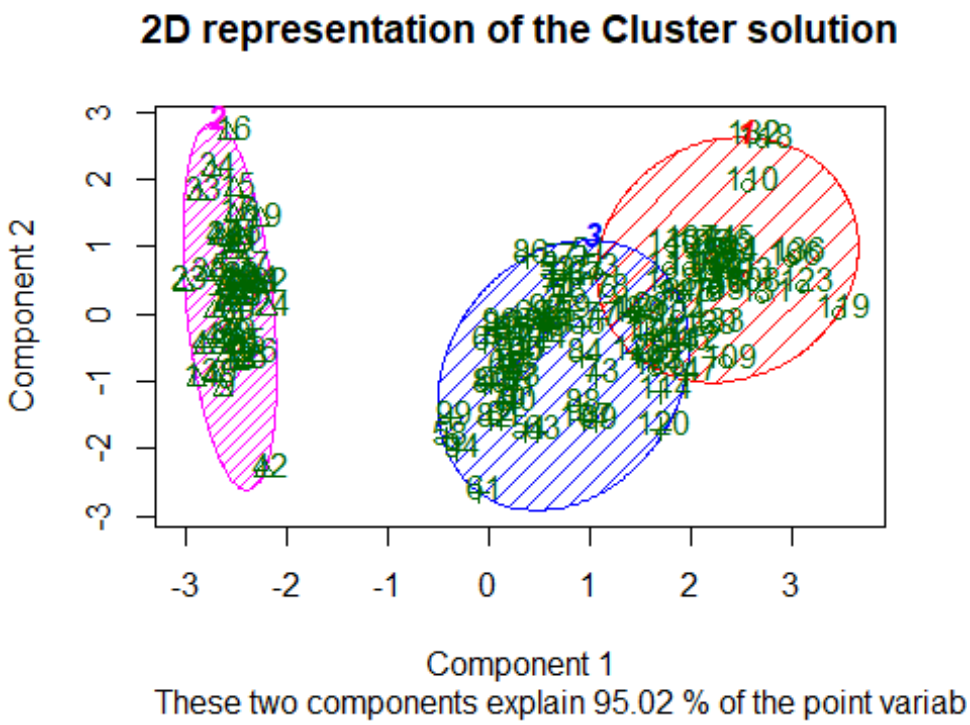
Nick

2021-03-18

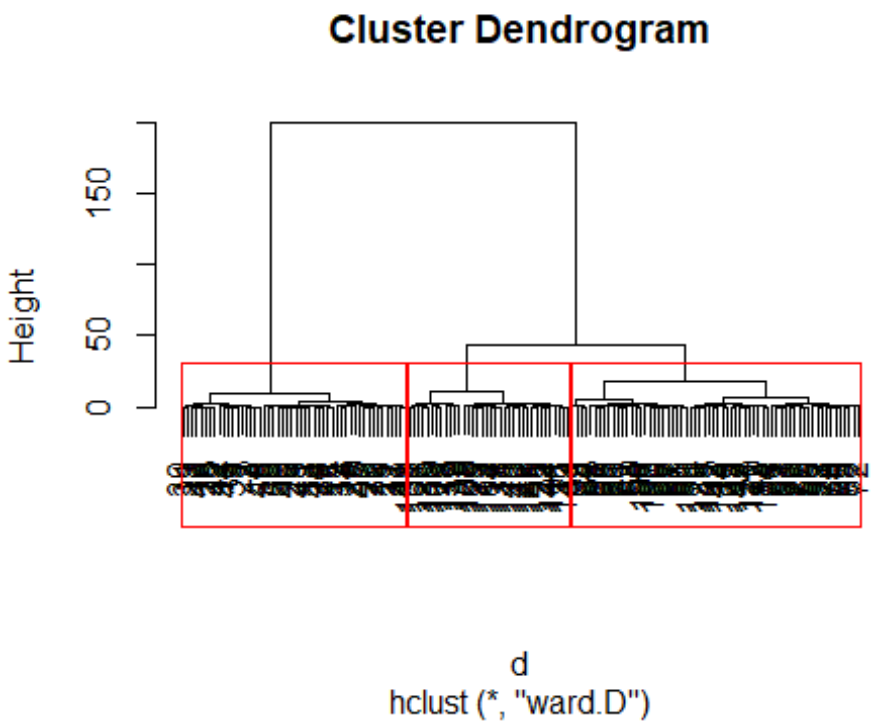
```
#AMS 580 Quiz 2 Nicholas Barrett
library("cluster")
#library("factoextra")
data = iris
pc = princomp(data[,1:4], cor = T)
screeplot(pc, type = "lines")
```



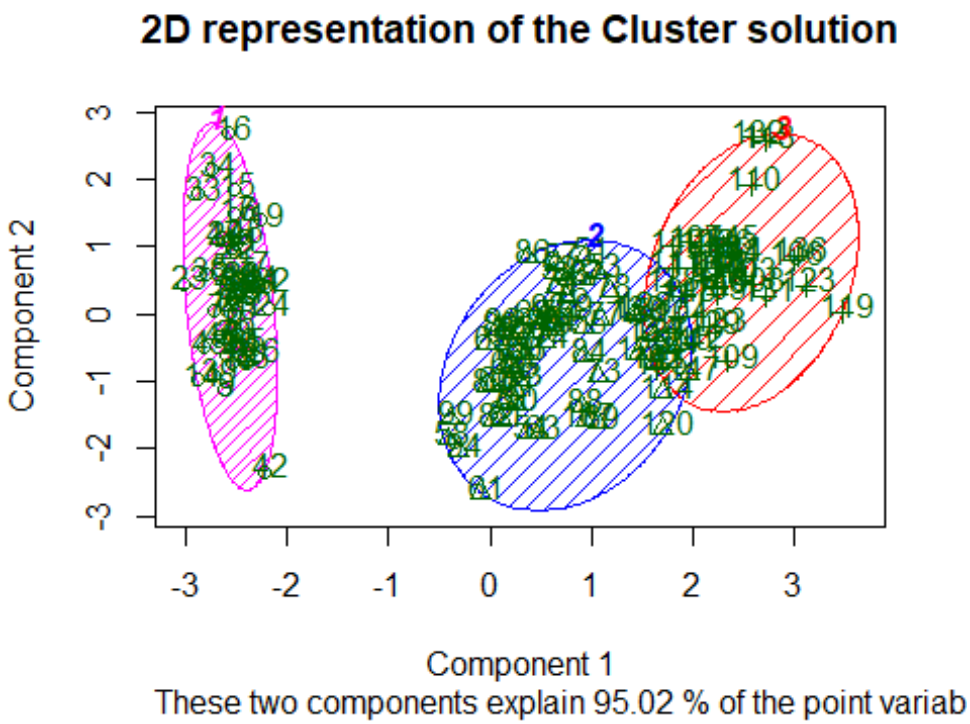

```
k.means.fit <- kmeans(data[,1:4], 3)
library(cluster)
clusplot(data, k.means.fit$cluster, main='2D representation of the Cluster
solution',color=TRUE, shade=TRUE,labels=2, lines=0)
```



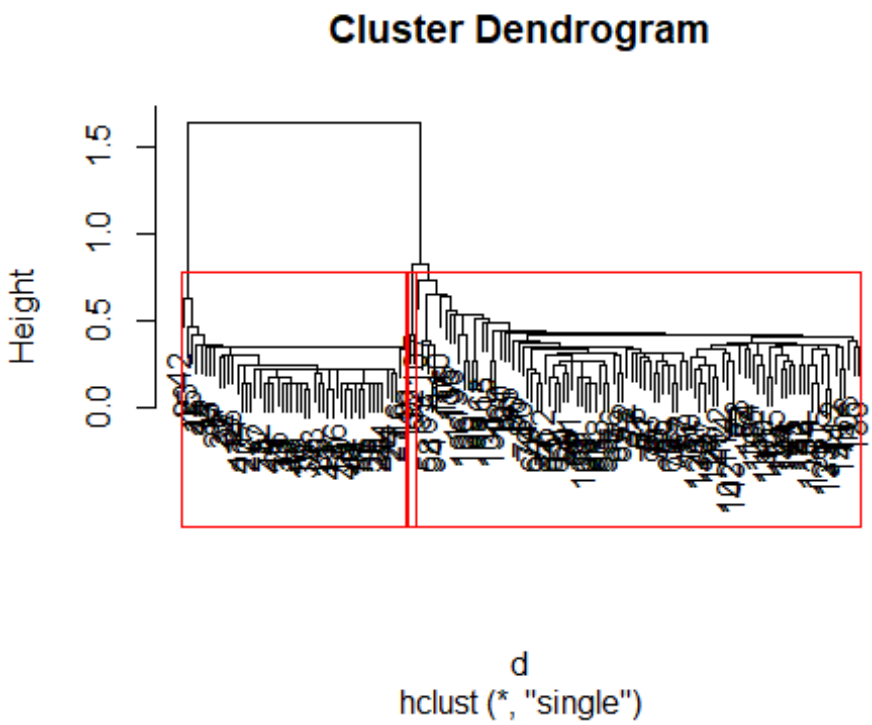

```
table(data[,5],k.means.fit$cluster)

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

#Wald
d <- dist(data[,1:4], method = "euclidean")
H.fit <- hclust(d, method="ward.D")
plot(H.fit)
groups <- cutree(H.fit, k=3)
rect.hclust(H.fit, k=3, border="red")
```



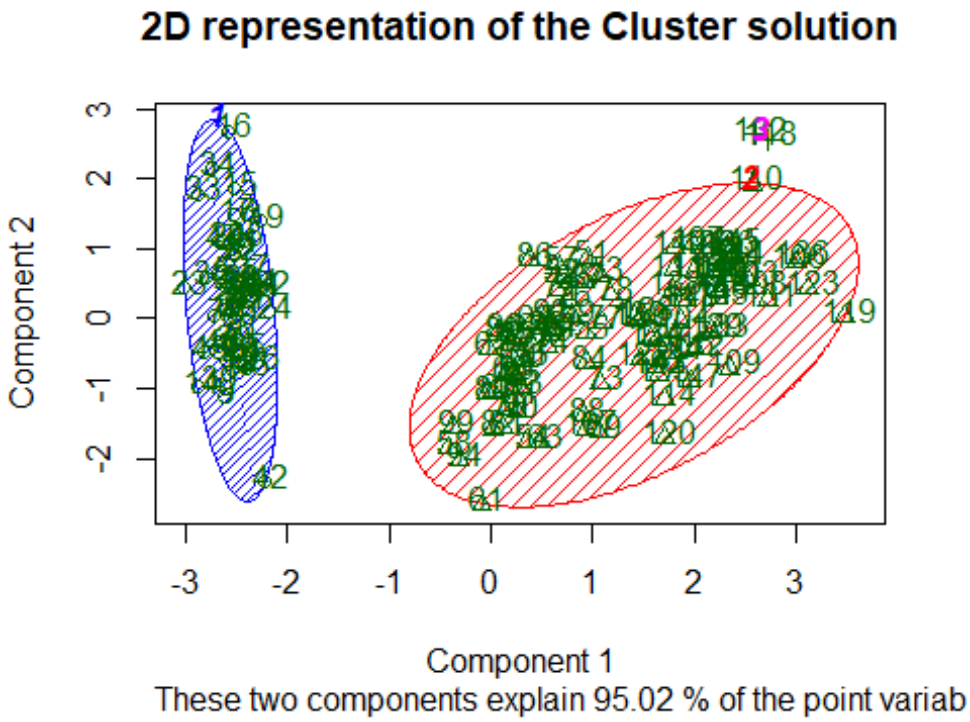

```
table(data[,5],groups)

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

clusplot(data, groups, main='2D representation of the Cluster
solution',color=TRUE, shade=TRUE,labels=2, lines=0)
```



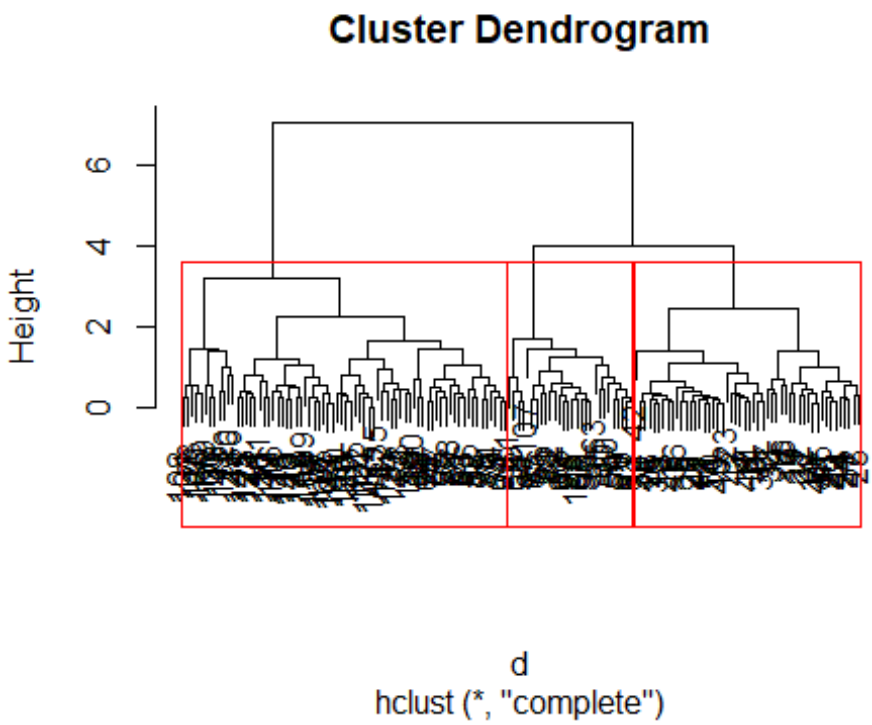

```
#136/150

#H.Single
H.fit <- hclust(d, method="single")
plot(H.fit)
groups <- cutree(H.fit, k=3)
rect.hclust(H.fit, k=3, border="red")
```



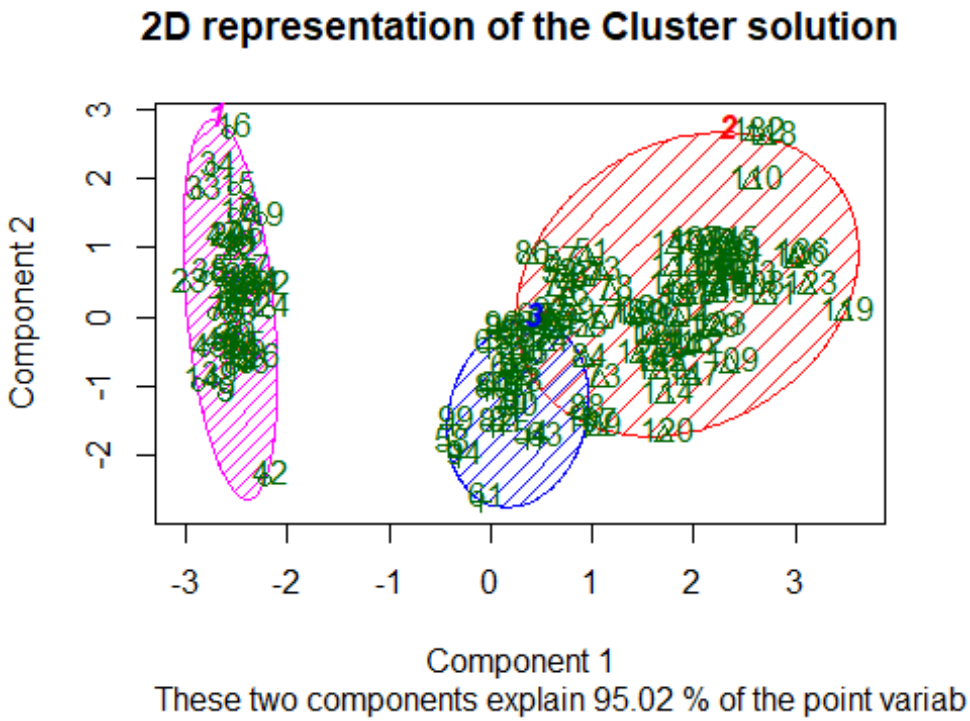

```
table(data[,5],groups)

##             groups
##               1  2  3
##   setosa     50  0  0
##   versicolor  0 50  0
##   virginica   0 48  2

clusplot(data, groups, main='2D representation of the Cluster
solution',color=TRUE, shade=TRUE,labels=2, lines=0)
```



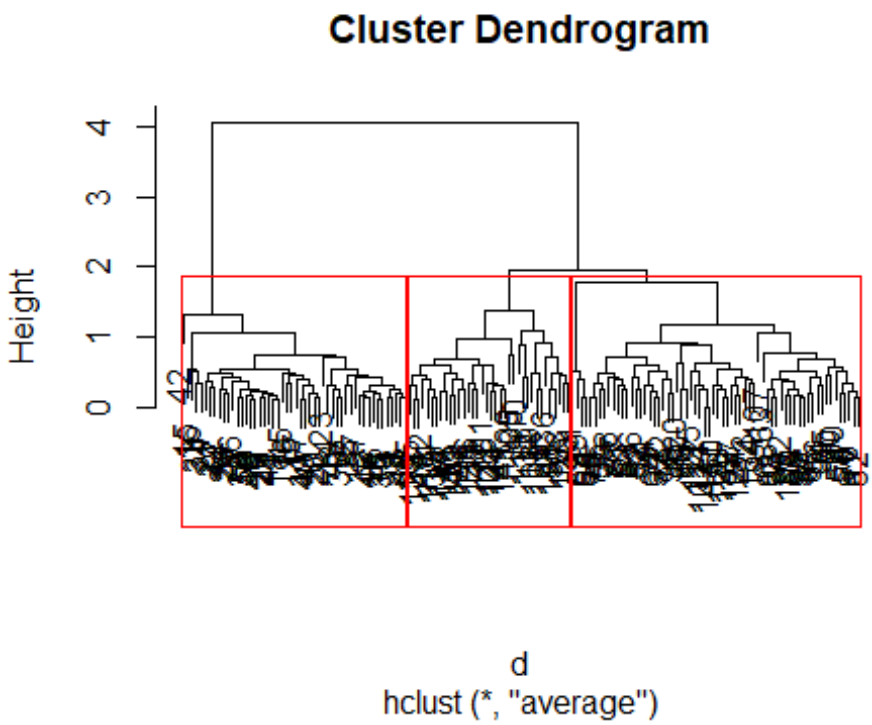

```
#102/150


#H.Complete
H.fit <- hclust(d, method="complete")
plot(H.fit)
groups <- cutree(H.fit, k=3)
rect.hclust(H.fit, k=3, border="red")
```



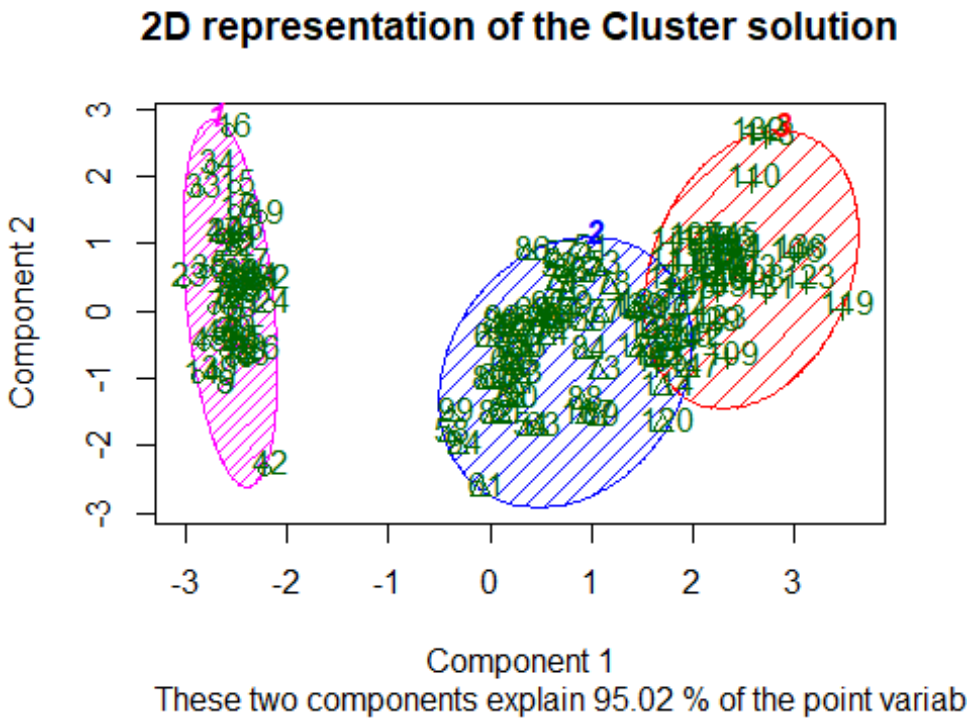

```
table(data[,5],groups)

##             groups
##               1  2  3
##   setosa     50  0  0
##   versicolor  0 23 27
##   virginica   0 49  1

clusplot(data, groups, main='2D representation of the Cluster
solution',color=TRUE, shade=TRUE,labels=2, lines=0)
```



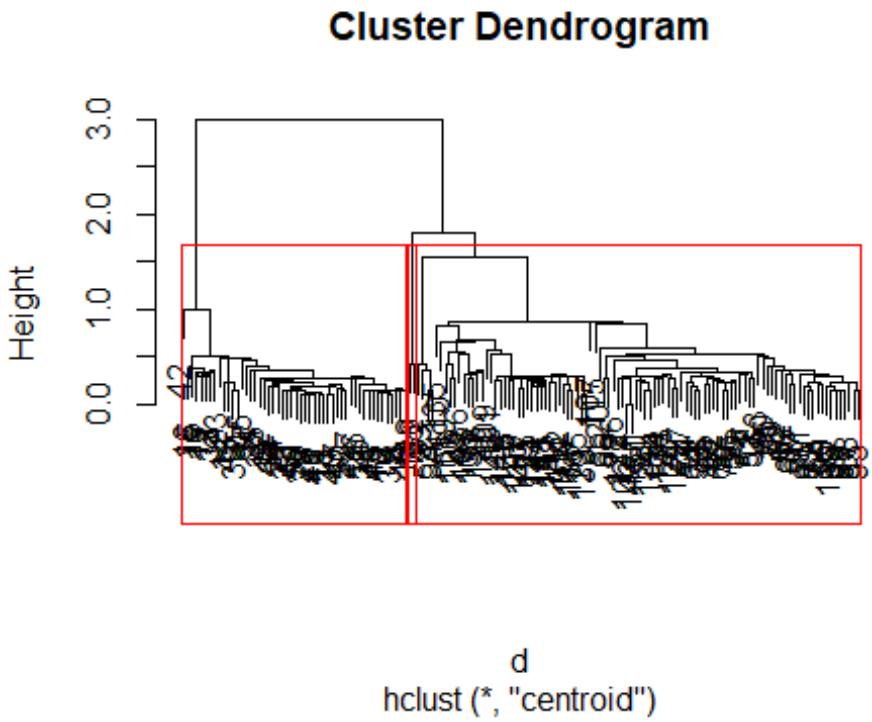

```
#74/150


#H.Average
H.fit <- hclust(d, method="average")
plot(H.fit)
groups <- cutree(H.fit, k=3)
rect.hclust(H.fit, k=3, border="red")
```



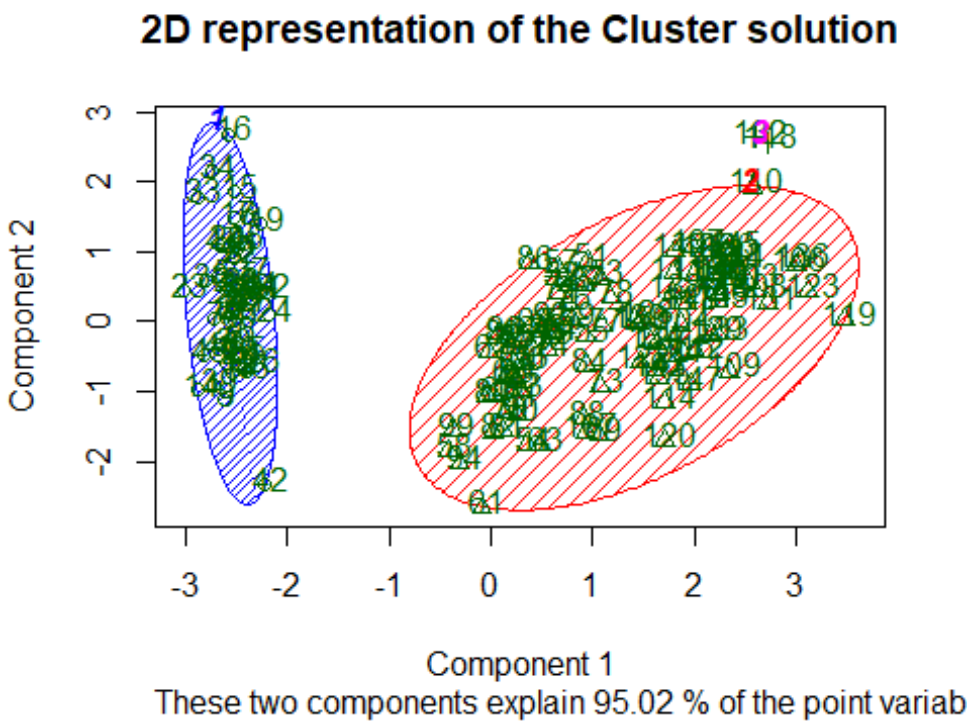

```
table(data[,5],groups)

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

clusplot(data, groups, main='2D representation of the Cluster
solution',color=TRUE, shade=TRUE,labels=2, lines=0)
```




```
#136/150

#H.centroid
H.fit <- hclust(d, method="centroid")
plot(H.fit)
groups <- cutree(H.fit, k=3)
rect.hclust(H.fit, k=3, border="red")
```




```
table(data[,5],groups)

##             groups
##               1  2  3
##   setosa     50  0  0
##   versicolor  0 50  0
##   virginica   0 48  2

clusplot(data, groups, main='2D representation of the Cluster
solution',color=TRUE, shade=TRUE,labels=2, lines=0)
```




```
#102/150

#Wald or H average would work best for this data, both achievinve 50/50 for
the first 2 groups
# and putting to many species in group 2 from group 3
# this method is only viable for determining group 1 and so is not
recommended

#Q2
install.packages("devtools")
library(devtools)
install_github("vqv/ggbiplot")
library(ggbiplot)
ir.pca = prcomp(iris[,1:4], center = TRUE, scale. = TRUE)
summary(ir.pca)

## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7084 0.9560 0.38309 0.14393
## Proportion of Variance 0.7296 0.2285 0.03669 0.00518
## Cumulative Proportion  0.7296 0.9581 0.99482 1.00000

ggbiplot(ir.pca)
ggbiplot(ir.pca,ellipse=TRUE,circle=TRUE, labels=rownames(iris),
groups=iris$Species)
ir.pca$rotation[,1] #PC1
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.5210659   -0.2693474    0.5804131    0.5648565

#Yes, the variance is easily soaked by few principal components
```