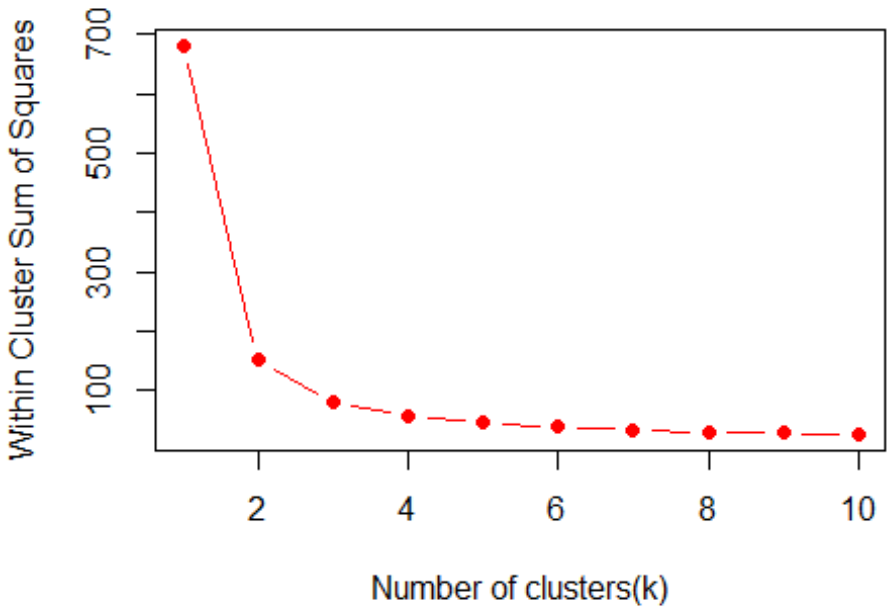
Optimum_Clusters_Prediction.R

Valli Subha R

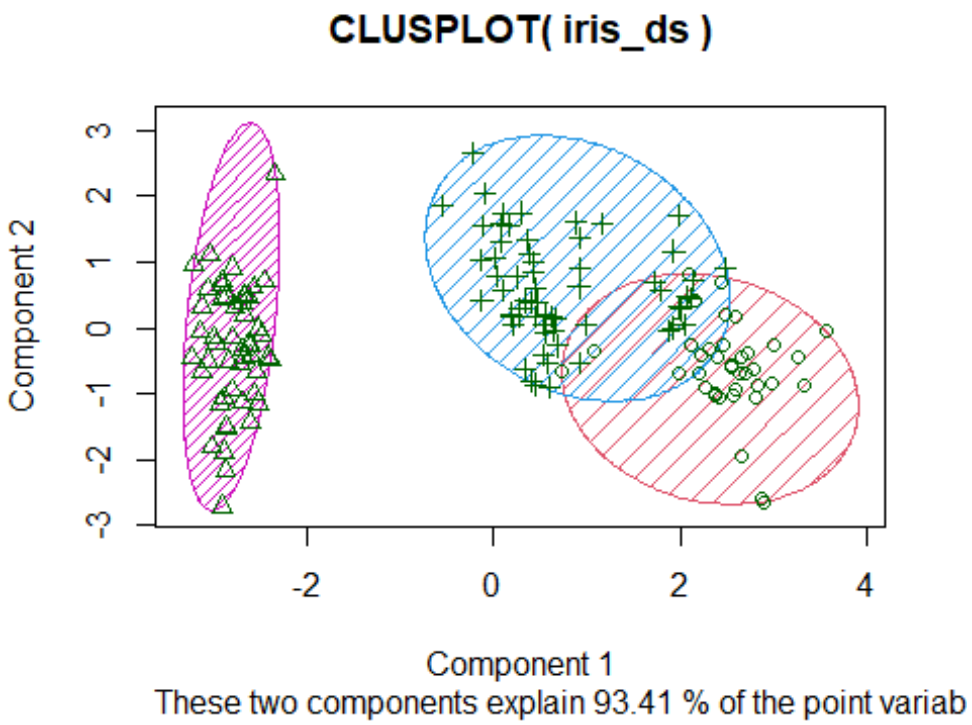
2021-01-11

```
iris_ds= read.csv("Iris.csv")

str(iris_ds)

## 'data.frame':    150 obs. of  6 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ SepalLengthCm: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ SepalWidthCm : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ PetalLengthCm: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ PetalWidthCm : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species      : chr  "Iris-setosa" "Iris-setosa" "Iris-setosa" "Iris-
setosa" ...

summary(iris_ds)

##        Id         SepalLengthCm    SepalWidthCm   PetalLengthCm
##  Min.   :  1.00   Min.   :4.300   Min.   :2.000   Min.   :1.000
##  1st Qu.: 38.25   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600
##  Median : 75.50   Median :5.800   Median :3.000   Median :4.350
##  Mean   : 75.50   Mean   :5.843   Mean   :3.054   Mean   :3.759
##  3rd Qu.:112.75   3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100
##  Max.   :150.00   Max.   :7.900   Max.   :4.400   Max.   :6.900
##   PetalWidthCm     Species
##  Min.   :0.100   Length:150
##  1st Qu.:0.300   Class :character
##  Median :1.300   Mode  :character
##  Mean   :1.199
##  3rd Qu.:1.800
##  Max.   :2.500

#(without normalization)

library(ggplot2)

tot.withinss = NULL
for (i in 1:10){
  iris_cluster = kmeans(iris_ds[,2:5], center=i, nstart=20)
  tot.withinss[i] = iris_cluster$tot.withinss
}

plot(x=1:10, y=tot.withinss, type="b", pch=19, col= "red",
```

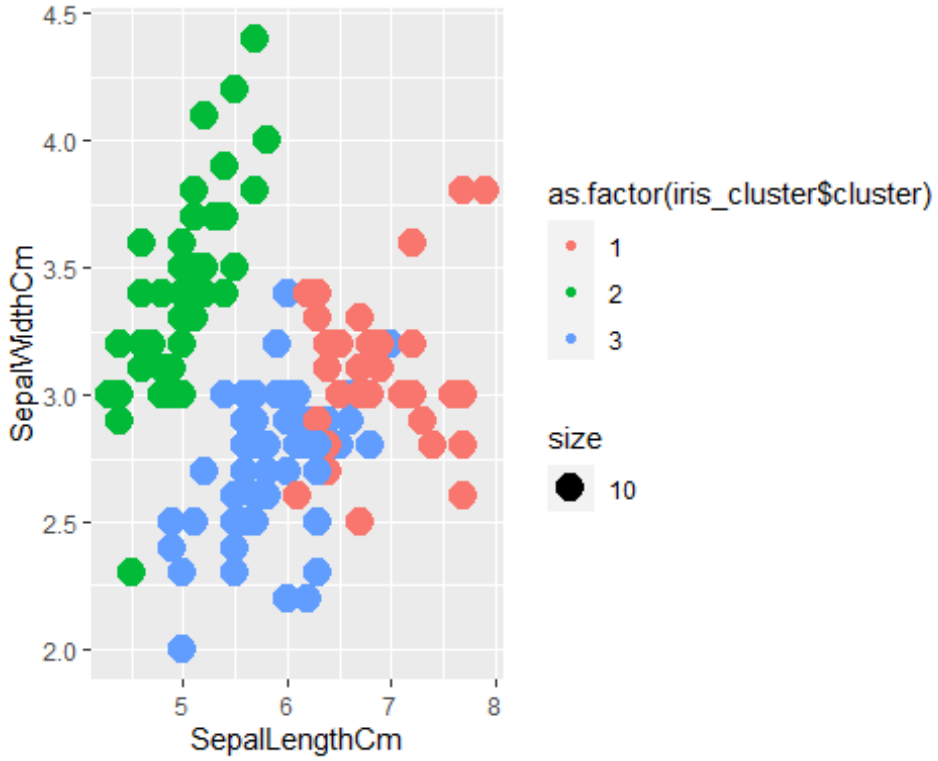
```
     xlab = "Number of clusters(k)",
     ylab = "Within Cluster Sum of Squares")
```




```
library(cluster)
iris_cluster = kmeans(iris_ds[,2:5], center=3, nstart=20)
clusplot(x=iris_ds, clus= iris_cluster$cluster,
         color=T, shade=T, labels=0, lines=0)
```

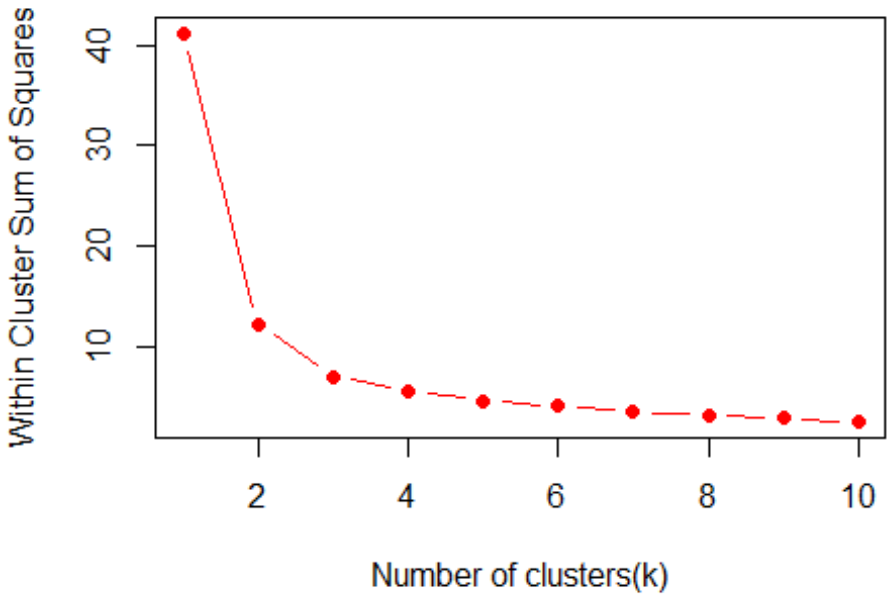



```
table(iris_cluster$cluster, iris_ds$Species)

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

ggplot(iris_ds,aes(x = SepalLengthCm, y = SepalWidthCm,
                   col= as.factor(iris_cluster$cluster),size=10))+
  geom_point() + scale_color_discrete()
```

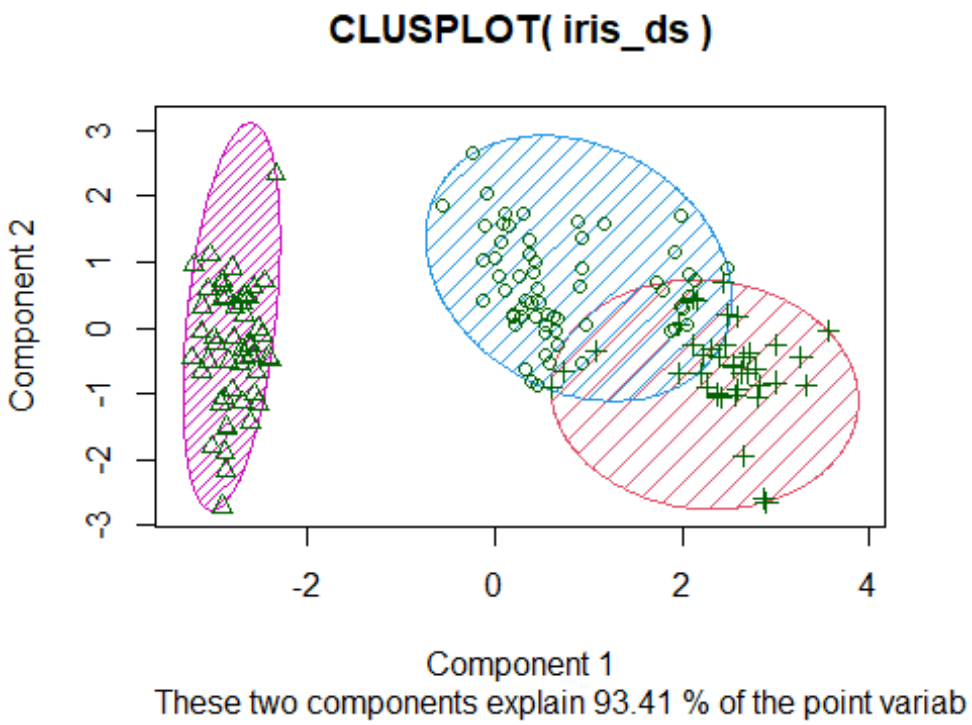



```
##########################################



iris_ds= read.csv("Iris.csv")

str(iris_ds)

## 'data.frame':    150 obs. of  6 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ SepalLengthCm: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ SepalWidthCm : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ PetalLengthCm: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ PetalWidthCm : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species      : chr  "Iris-setosa" "Iris-setosa" "Iris-setosa" "Iris-
setosa" ...

summary(iris_ds)

##        Id         SepalLengthCm    SepalWidthCm   PetalLengthCm
##  Min.   :  1.00   Min.   :4.300   Min.   :2.000   Min.   :1.000
##  1st Qu.: 38.25   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600
##  Median : 75.50   Median :5.800   Median :3.000   Median :4.350
##  Mean   : 75.50   Mean   :5.843   Mean   :3.054   Mean   :3.759
##  3rd Qu.:112.75   3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100
##  Max.   :150.00   Max.   :7.900   Max.   :4.400   Max.   :6.900
##   PetalWidthCm     Species
```

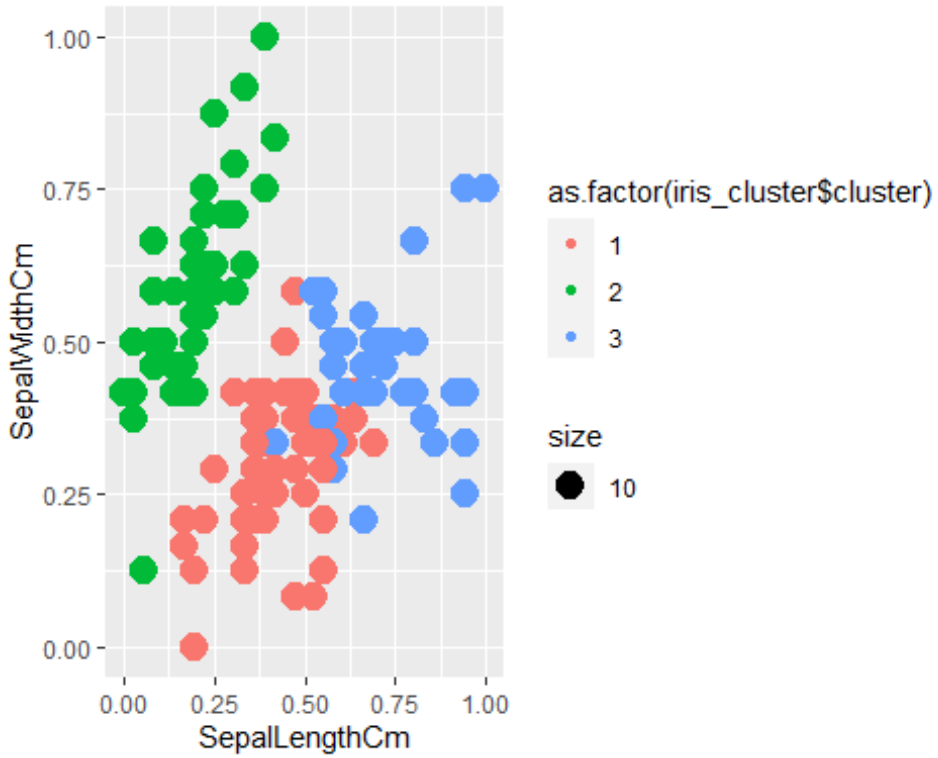
```
##  Min.   :0.100   Length:150
##  1st Qu.:0.300   Class :character
##  Median :1.300   Mode  :character
##  Mean   :1.199
##  3rd Qu.:1.800
##  Max.   :2.500

normalize <- function(x){
  return ((x-min(x))/(max(x)-min(x)))
}

iris_ds$SepalLengthCm = normalize(iris_ds$SepalLengthCm)
iris_ds$SepalWidthCm = normalize(iris_ds$SepalWidthCm)
iris_ds$PetalLengthCm = normalize(iris_ds$PetalLengthCm)
iris_ds$PetalWidthCm = normalize(iris_ds$PetalWidthCm)

library(ggplot2)

tot.withinss = NULL
for (i in 1:10){
  iris_cluster = kmeans(iris_ds[,2:5], center=i, nstart=20)
  tot.withinss[i] = iris_cluster$tot.withinss
}

plot(x=1:10, y=tot.withinss, type="b", pch=19, col= "red",
     xlab = "Number of clusters(k)",
     ylab = "Within Cluster Sum of Squares")
```




```r
library(cluster)
iris_cluster = kmeans(iris_ds[,2:5], center=3, nstart=20)
clusplot(x=iris_ds, clus= iris_cluster$cluster,
         color=T, shade=T, labels=0, lines=0)
```




```
table(iris_cluster$cluster, iris_ds$Species)

##
##     Iris-setosa Iris-versicolor Iris-virginica
##   1           0              47             14
##   2          50               0              0
##   3           0               3             36

ggplot(iris_ds,aes(x = SepalLengthCm, y = SepalWidthCm,
                   col= as.factor(iris_cluster$cluster),size=10))+
  geom_point() + scale_color_discrete()
```


1.00
0.75
0.50
0.25
0.00
SepalWidthCm
0.00
0.25
0.50
0.75
1.00
SepalLengthCm
as.factor(iris_cluster$cluster)
1
2
3
size
10