Task 2 - Prediction using Unsupervised ML

SANJEETA CHAKRABORTY

10/04/2021

From the given ‘Iris’ dataset, predict the optimum number of clusters and represent it visually.

Loading Dataset

```
data <- read.csv("Iris.csv", header=TRUE)
head(data)
```

```
##   Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm     Species
## 1  1           5.1          3.5           1.4          0.2 Iris-setosa
## 2  2           4.9          3.0           1.4          0.2 Iris-setosa
## 3  3           4.7          3.2           1.3          0.2 Iris-setosa
## 4  4           4.6          3.1           1.5          0.2 Iris-setosa
## 5  5           5.0          3.6           1.4          0.2 Iris-setosa
## 6  6           5.4          3.9           1.7          0.4 Iris-setosa
```

Loading Packages

```
library(ClusterR)
```

```
## Warning: package 'ClusterR' was built under R version 4.0.5
```

```
## Loading required package: gtools
```

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 4.0.5
```

```
library(ggplot2)
library(funModeling)
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## funModeling v.1.9.4 :)
## Examples and tutorials at livebook.datascienceheroes.com
##  / Now in Spanish: librovivodecienciadedatos.ai
```

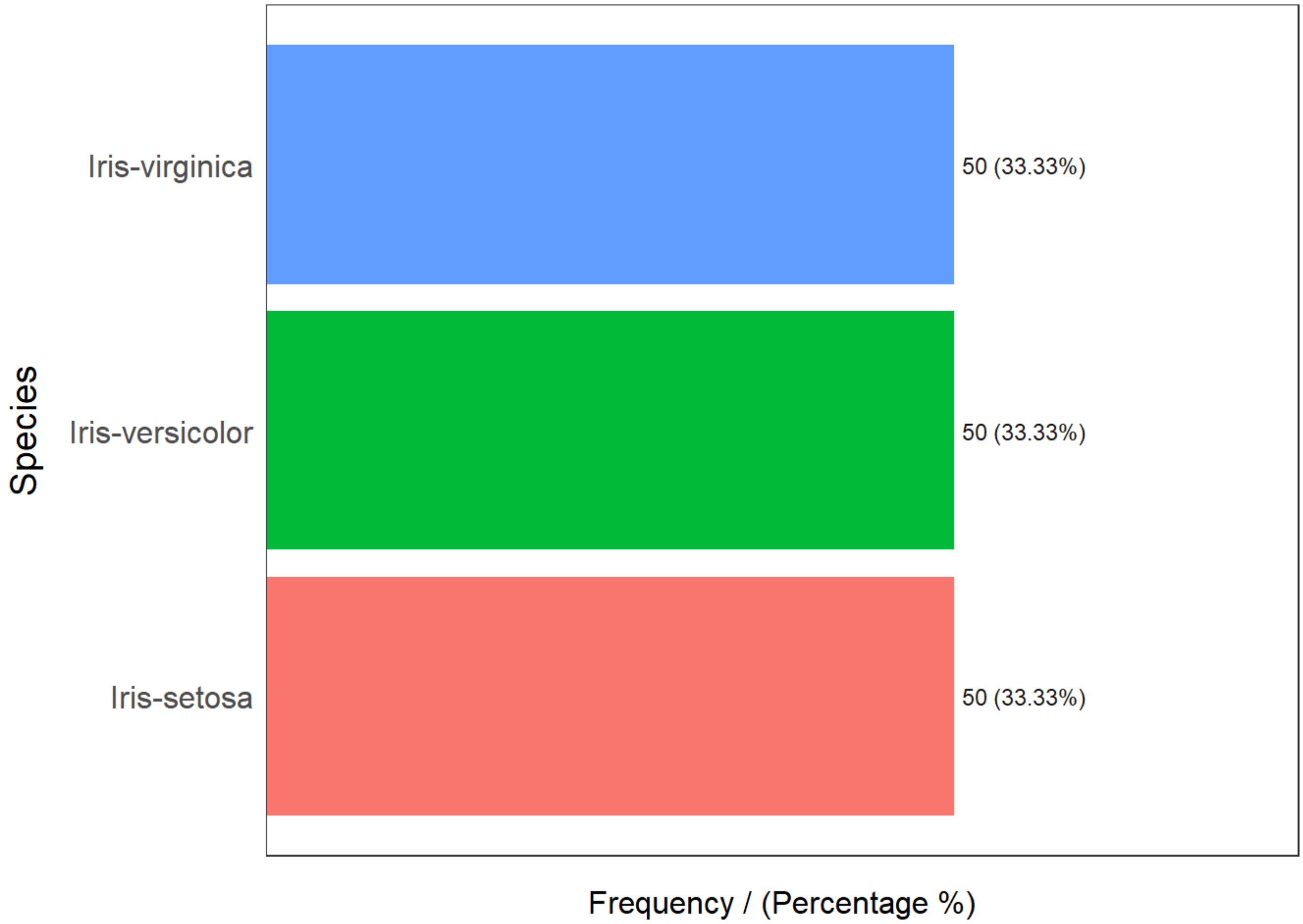
Getting Insights from the data

```
describe(data)
```

```
## data
##
##  6  Variables      150  Observations
## --------------------------------------------------------------------------------
## Id
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      150        1     75.5    50.33     8.45    15.90
##      .25      .50      .75      .90      .95
##    38.25    75.50   112.75   135.10   142.55
##
## lowest :   1   2   3   4   5, highest: 146 147 148 149 150
## --------------------------------------------------------------------------------
## SepalLengthCm
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0       35    0.998    5.843   0.9462    4.600    4.800
##      .25      .50      .75      .90      .95
##    5.100    5.800    6.400    6.900    7.255
##
## lowest : 4.3 4.4 4.5 4.6 4.7, highest: 7.3 7.4 7.6 7.7 7.9
## --------------------------------------------------------------------------------
## SepalWidthCm
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0       23    0.991    3.054   0.4837    2.345    2.500
##      .25      .50      .75      .90      .95
##    2.800    3.000    3.300    3.610    3.800
##
## lowest : 2.0 2.2 2.3 2.4 2.5, highest: 3.9 4.0 4.1 4.2 4.4
## --------------------------------------------------------------------------------
## PetalLengthCm
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0       43    0.998    3.759    1.978     1.30     1.40
##      .25      .50      .75      .90      .95
##     1.60     4.35     5.10     5.80     6.10
##
## lowest : 1.0 1.1 1.2 1.3 1.4, highest: 6.3 6.4 6.6 6.7 6.9
## --------------------------------------------------------------------------------
## PetalWidthCm
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0       22    0.991    1.199   0.8688      0.2      0.2
##      .25      .50      .75      .90      .95
##      0.3      1.3      1.8      2.2      2.3
##
## lowest : 0.1 0.2 0.3 0.4 0.5, highest: 2.1 2.2 2.3 2.4 2.5
## --------------------------------------------------------------------------------
## Species
##        n  missing distinct
##      150        0        3
##
## Value           Iris-setosa Iris-versicolor  Iris-virginica
## Frequency                50              50              50
## Proportion            0.333           0.333           0.333
## --------------------------------------------------------------------------------
```

Analysis of categorical variables

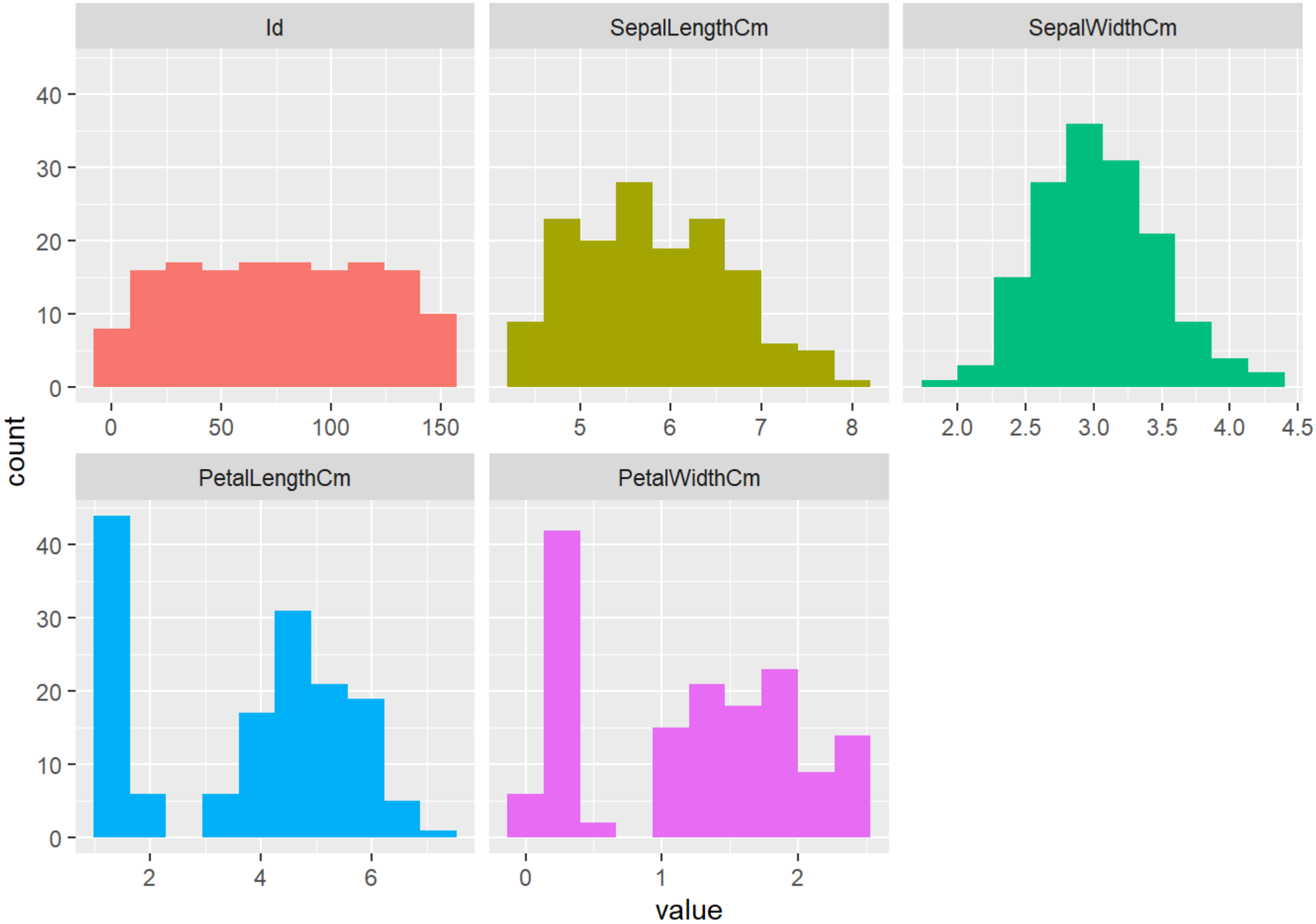
```
freq(data)
```




```
##           Species frequency percentage cumulative_perc
## 1     Iris-setosa        50      33.33           33.33
## 2 Iris-versicolor        50      33.33           66.66
## 3  Iris-virginica        50      33.33          100.00
```

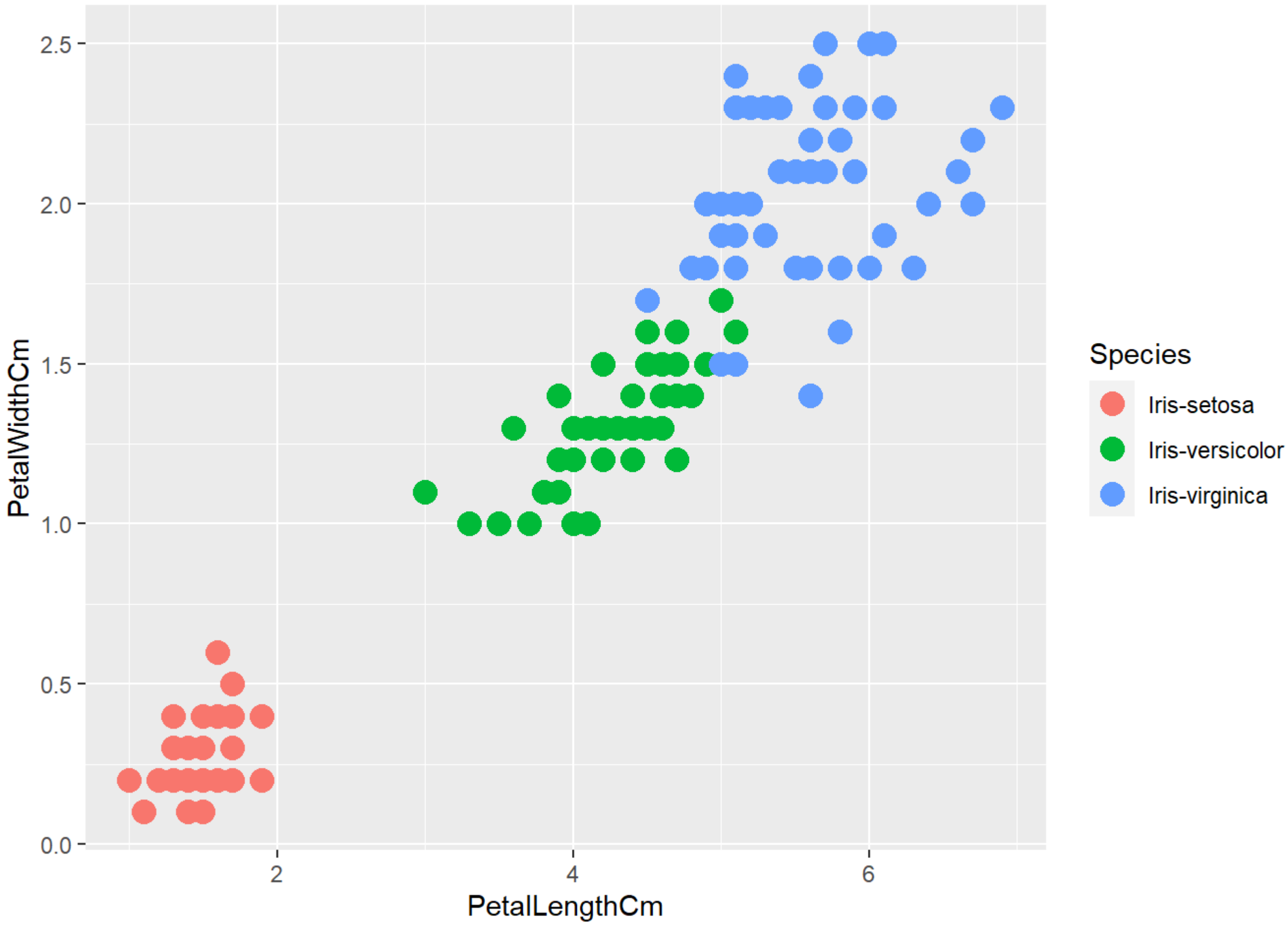
Analysis of Numerical Variables

```
plot_num(data)
```

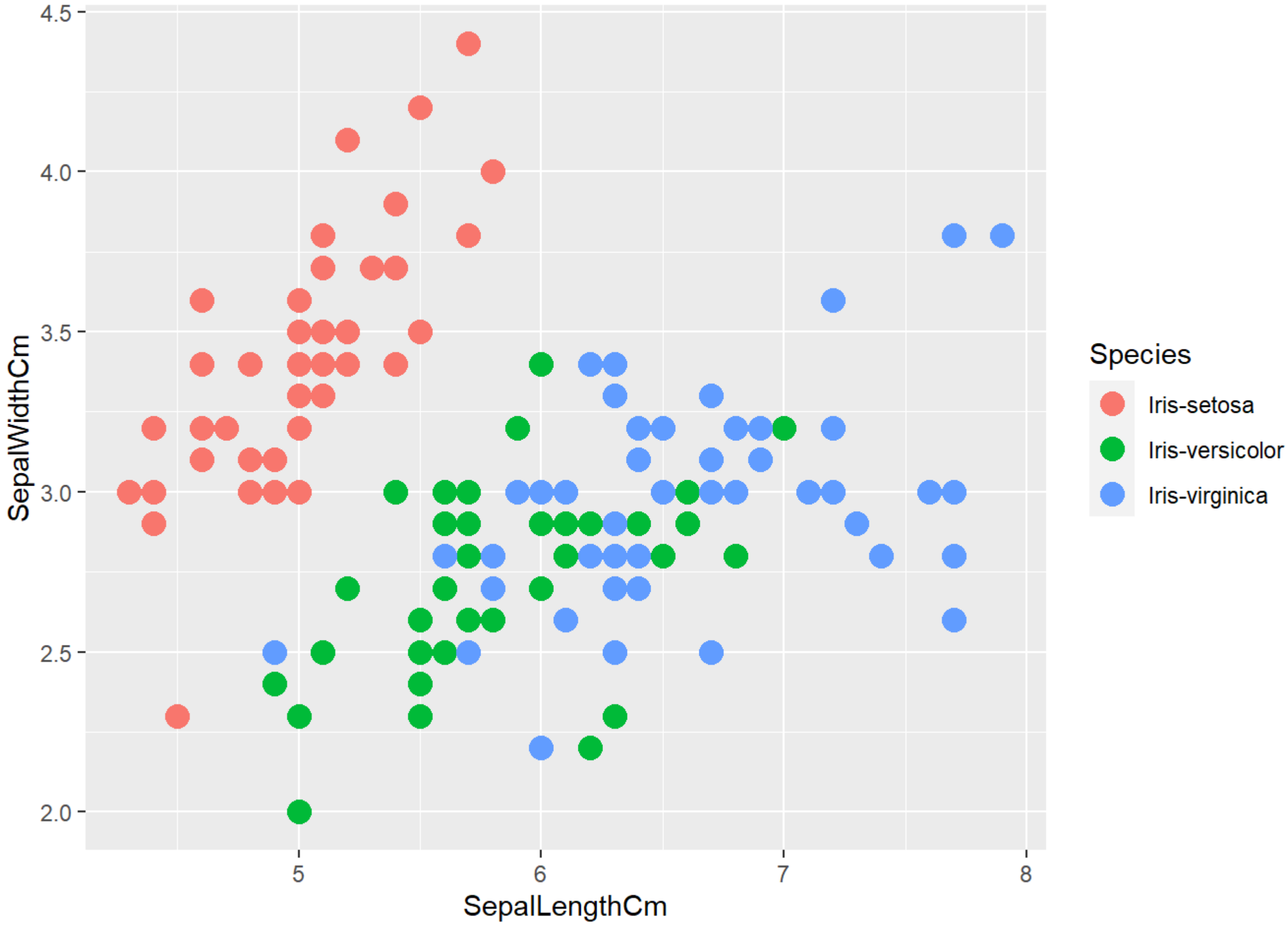


Scatterplot

```
ggplot(data, aes(PetalLengthCm, PetalWidthCm)) + geom_point(aes(col=Species), size=4)
```




```
ggplot(data, aes(SepalLengthCm, SepalWidthCm)) + geom_point(aes(col=Species), size=4)
```



From the above plots we can observe that the species setosa can be easily clustered while versicolor and virginica are overlapping.

Removing the label

```
df = data[,2:5]
head(df)
```

```
##   SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm
## 1           5.1          3.5           1.4          0.2
## 2           4.9          3.0           1.4          0.2
## 3           4.7          3.2           1.3          0.2
## 4           4.6          3.1           1.5          0.2
## 5           5.0          3.6           1.4          0.2
## 6           5.4          3.9           1.7          0.4
```

Fitting K-Means Model

```
set.seed(45)
model <- kmeans(df, centers = 3, nstart = 20)
model
```

```
## K-means clustering with 3 clusters of sizes 62, 38, 50
##
## Cluster means:
##   SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm
## 1      5.901613     2.748387      4.393548     1.433871
## 2      6.850000     3.073684      5.742105     2.071053
## 3      5.006000     3.418000      1.464000     0.244000
##
## Clustering vector:
##   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [38] 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2
## [112] 2 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2
## [149] 2 1
##
## Within cluster sum of squares by cluster:
## [1] 39.82097 23.87947 15.24040
##  (between_SS / total_SS =  88.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Comparing the clusters with the species

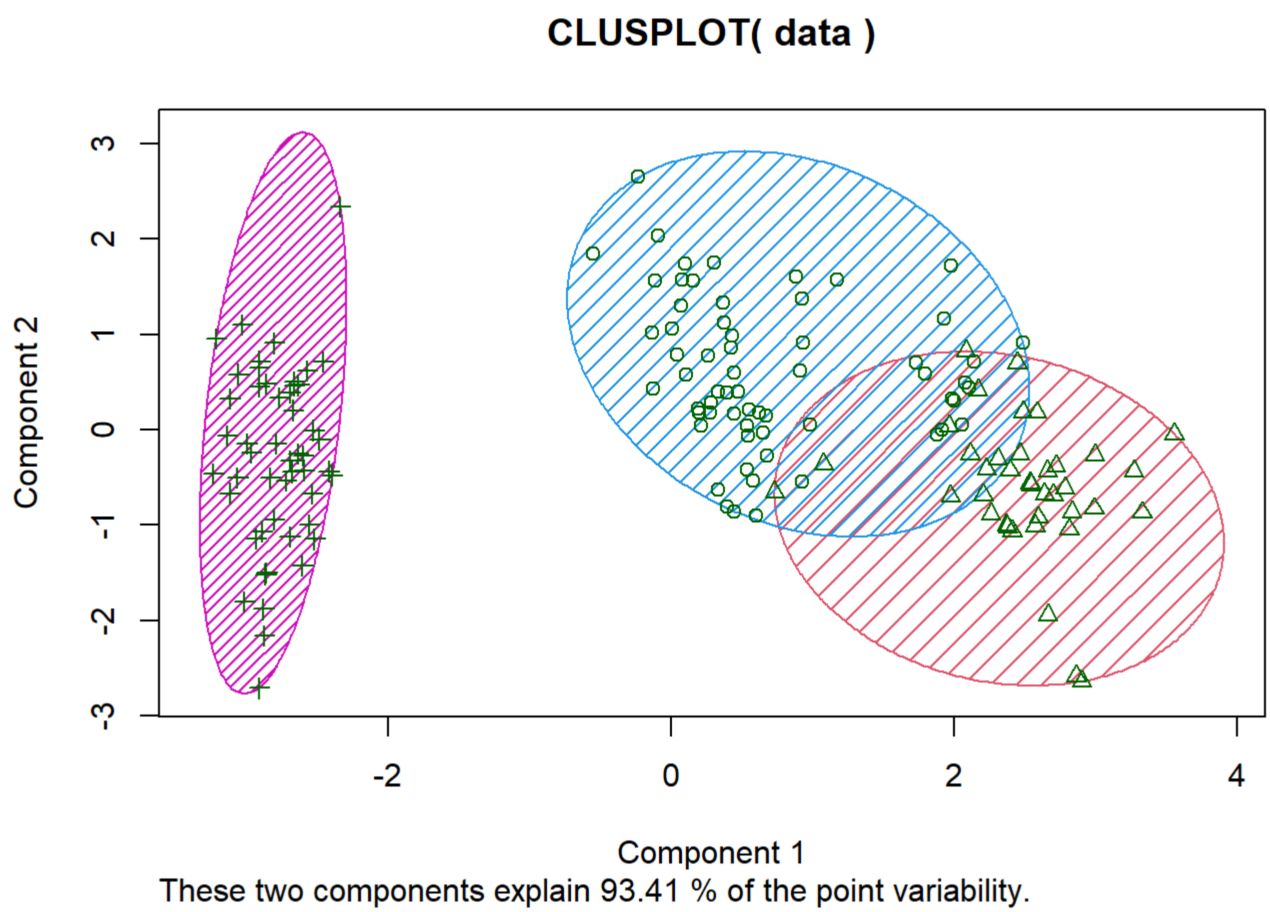
```
cm <- table(model$cluster, data$Species)
cm
```

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

We can observe, the data belonging to the setosa species got grouped into cluster 3, versicolor into cluster 1, and virginica into cluster 2. The algorithm wrongly classified two data points belonging to versicolor and fourteen data points belonging to virginica.

Clusterplot

```
clusplot(data, model$cluster, color=T, shade=T, labels=0, lines=0)
```

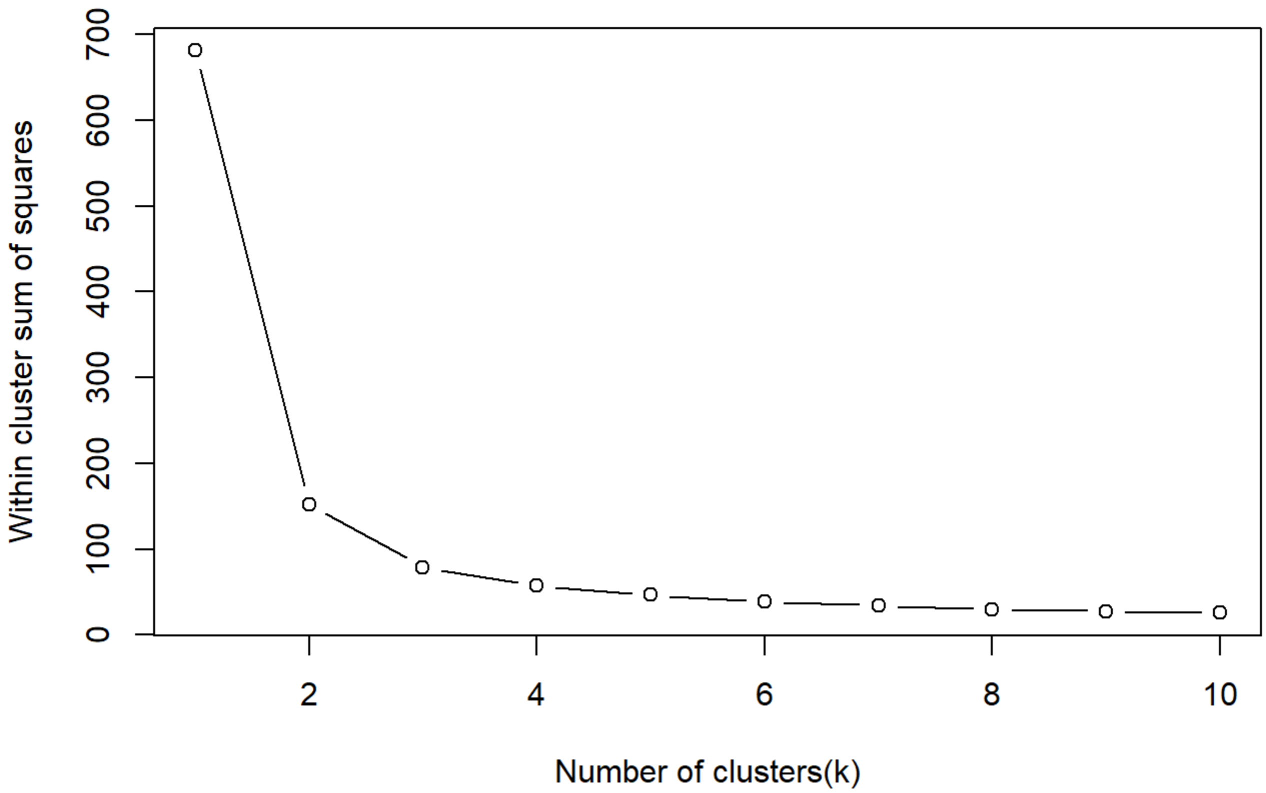


Optimum number of clusters

```
k.max <- 10
wss<- sapply(1:k.max,function(k){kmeans(data[,2:5],k,nstart = 20,iter.max = 20)$tot.withinss})
wss
```

```
##  [1] 680.82440 152.36871  78.94084  57.31787  46.53558  38.93096  34.18921
##  [8]  29.95409  27.76542  25.82880
```

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



From the above plot we can say that the optimum value for k is 3.