IRIS

Pramod Verma

16 March 2018

R Markdown

This is the R Markdown file created for introducing the basic building blocks of R to PSA Data Analytics Technical Participants.

```
head(iris)
```

```
##   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
```

```
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 ...
```

```
dim(iris)
```

```
## [1] 150   5
```

```
names(iris)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
## [5] "Species"
```

```
class(iris)
```

```
## [1] "data.frame"
```

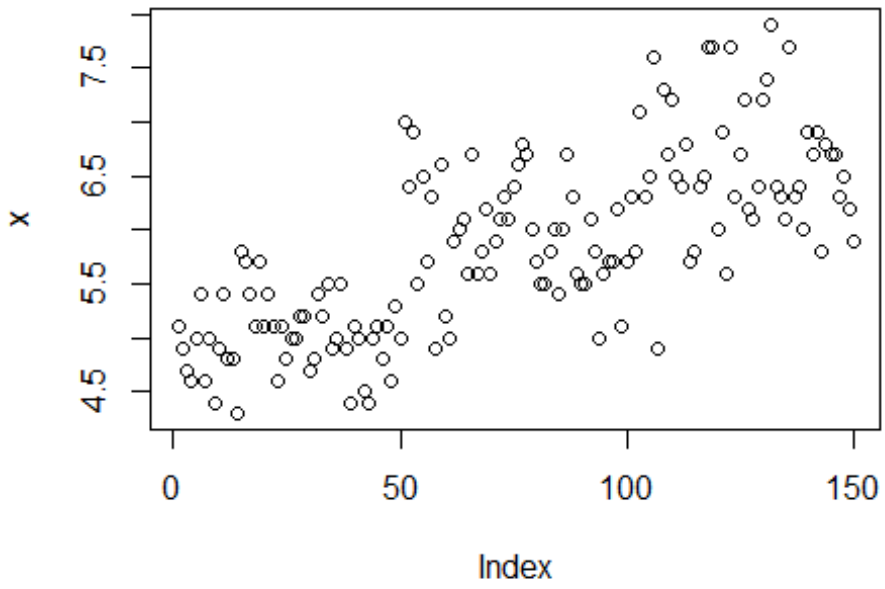
```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
```

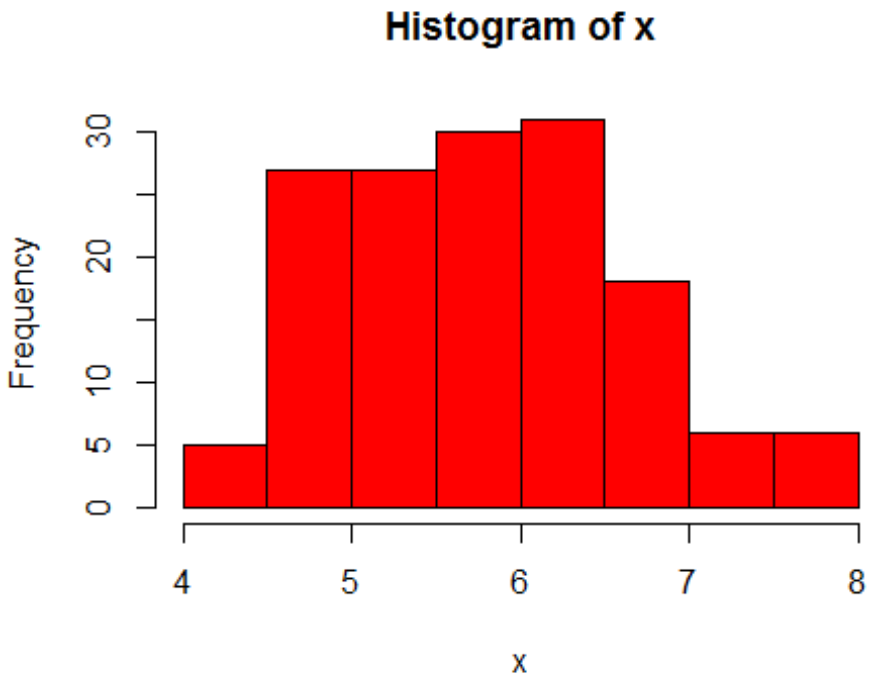
```
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Assigning Variables and Plotting

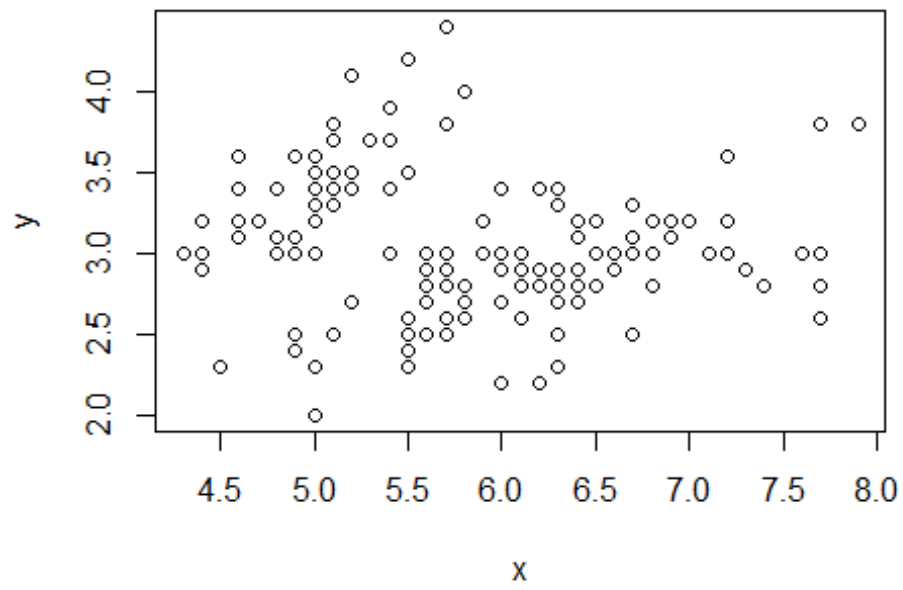
```
x <- iris$Sepal.Length
y <- iris$Sepal.Width
z <- iris$Species

plot(x)
```



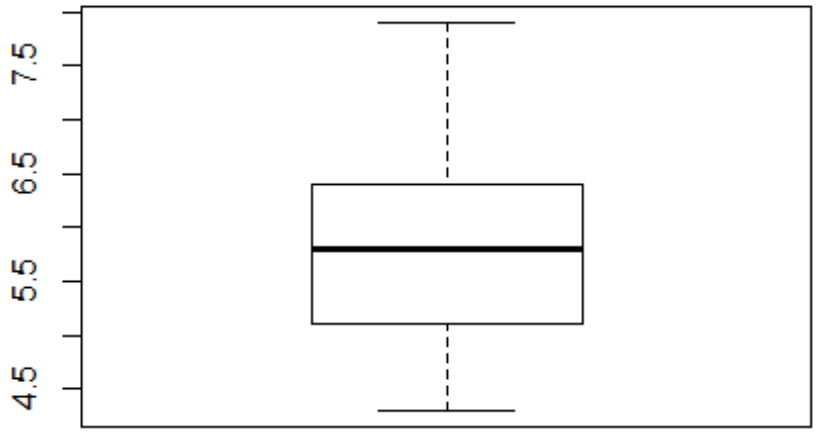

```
hist(x, col = "red")
```




```
plot(x,y)
```

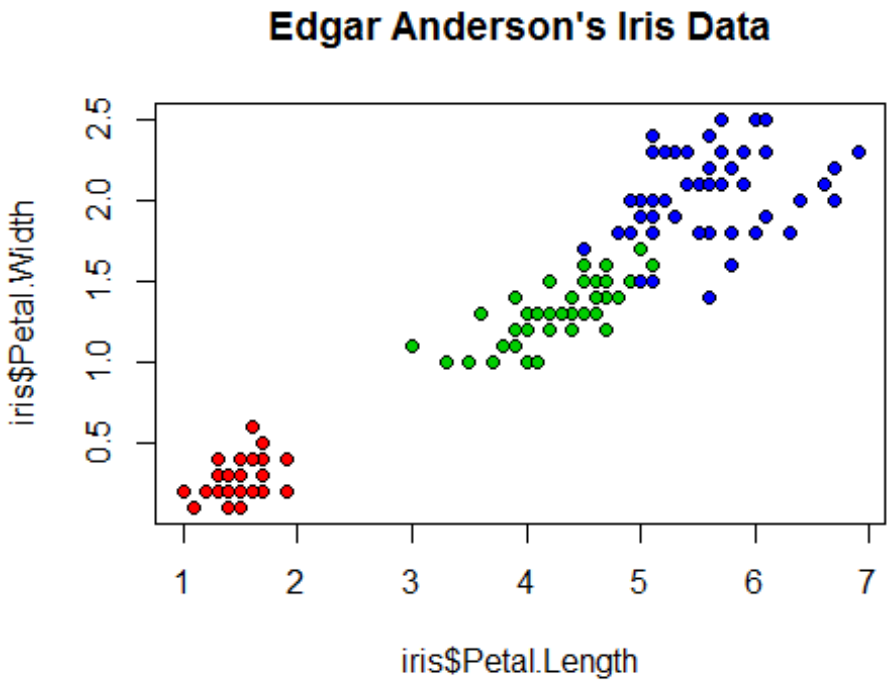



```
boxplot(x)
```



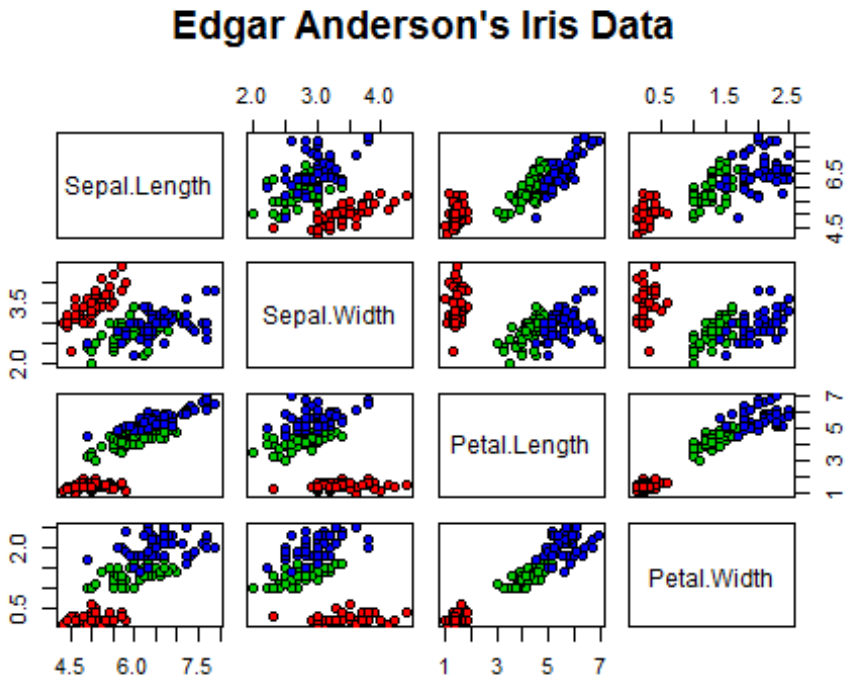
Looking at all the species in the same scatterplot for Petal Length & Width

```
plot(iris$Petal.Length, iris$Petal.Width, pch=21,
bg=c("red","green3","blue")[unclass(iris$Species)],
main="Edgar Anderson's Iris Data")
```



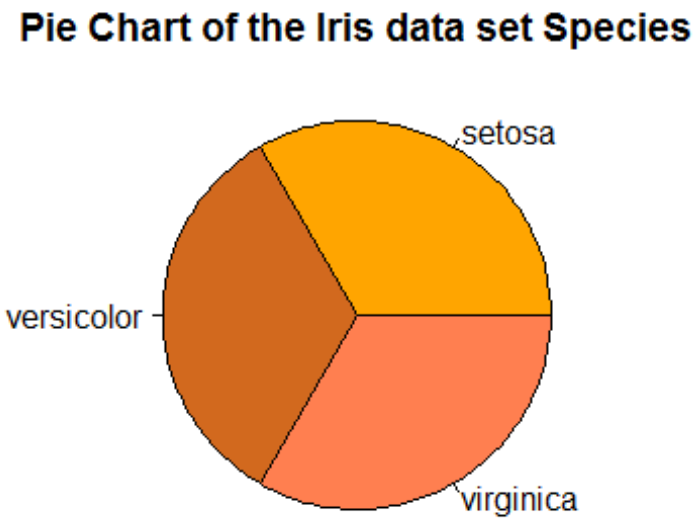
Looking at the scatterplot of all the variables for each species

```
pairs(iris[1:4], main = "Edgar Anderson's Iris Data", pch = 21,
bg = c("red", "green3", "blue")[unclass(iris$Species)])
```



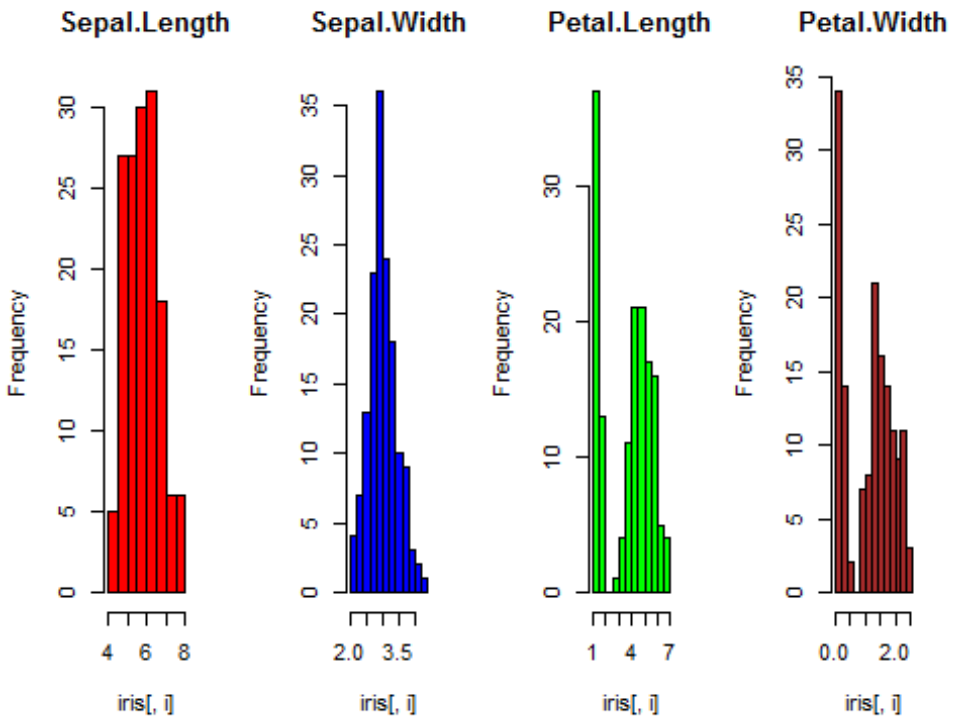
Creating pi chart to show the share of each species

```
pie(table(iris$Species),
main = "Pie Chart of the Iris data set Species",
col = c("orange1", "chocolate", "coral"),
radius = 1)
```



Using for loop to create multiple plots

```
par(mfrow=c(1,4))
color <- c("red", "blue","green","brown")
for(i in 1:4) {
hist(iris[,i], main=names(iris)[i], col = color[i]) }
```

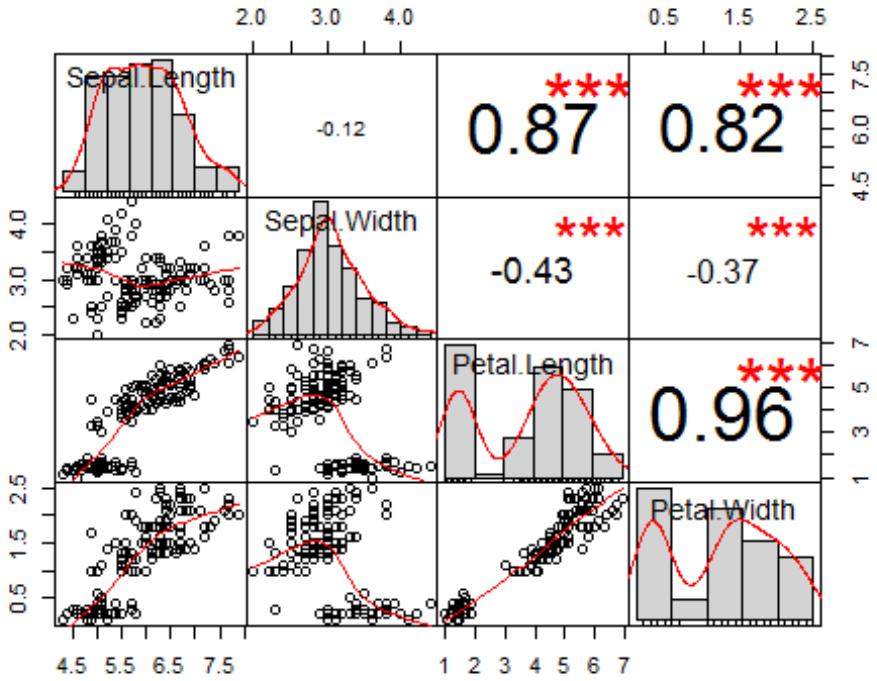


Finding out the correlation among all the numerical variables

```
cor(iris[,c(1:4)])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

```
library(PerformanceAnalytics)
chart.Correlation(iris[,c(1,2,3,4)], histogram=TRUE, pch=19)
```

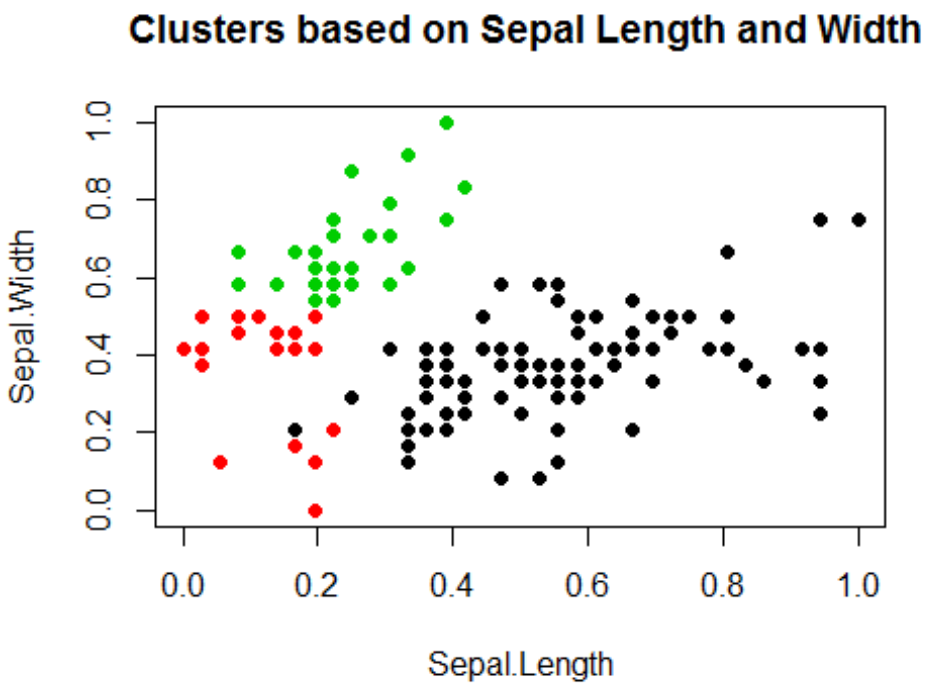


Demostration of K Means Clustering

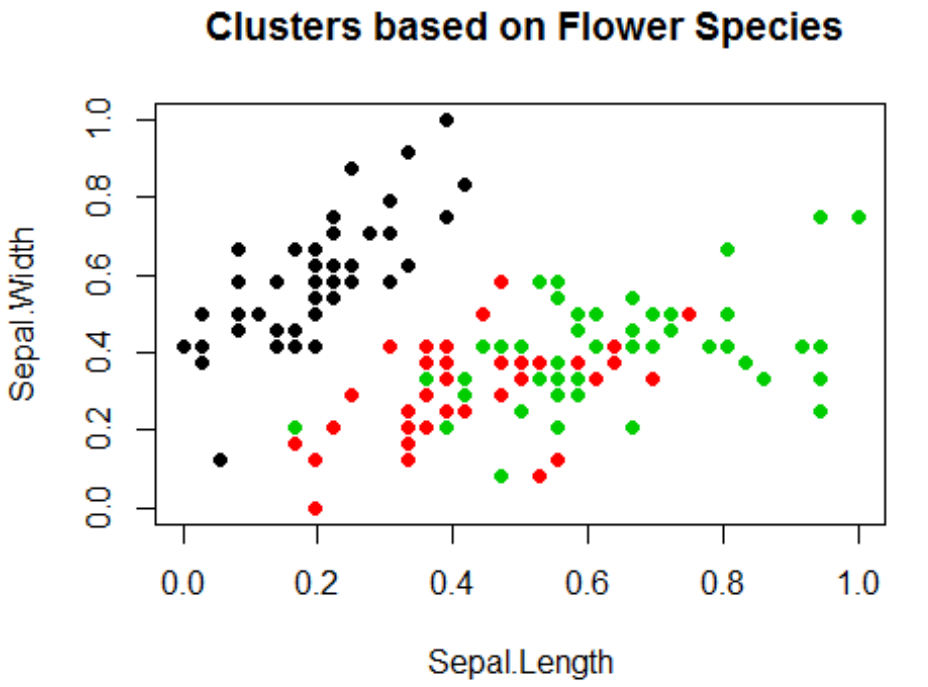
```
normalize <- function(x){
  return ((x-min(x))/(max(x)-min(x)))
}

iris$Sepal.Length<- normalize(iris$Sepal.Length)
iris$Sepal.Width<- normalize(iris$Sepal.Width)
iris$Petal.Length<- normalize(iris$Petal.Length)
iris$Petal.Width<- normalize(iris$Petal.Width)
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1   0.22222222   0.6250000   0.06779661  0.04166667  setosa
## 2   0.16666667   0.4166667   0.06779661  0.04166667  setosa
## 3   0.11111111   0.5000000   0.05084746  0.04166667  setosa
## 4   0.08333333   0.4583333   0.08474576  0.04166667  setosa
## 5   0.19444444   0.6666667   0.06779661  0.04166667  setosa
## 6   0.30555556   0.7916667   0.11864407  0.12500000  setosa
```

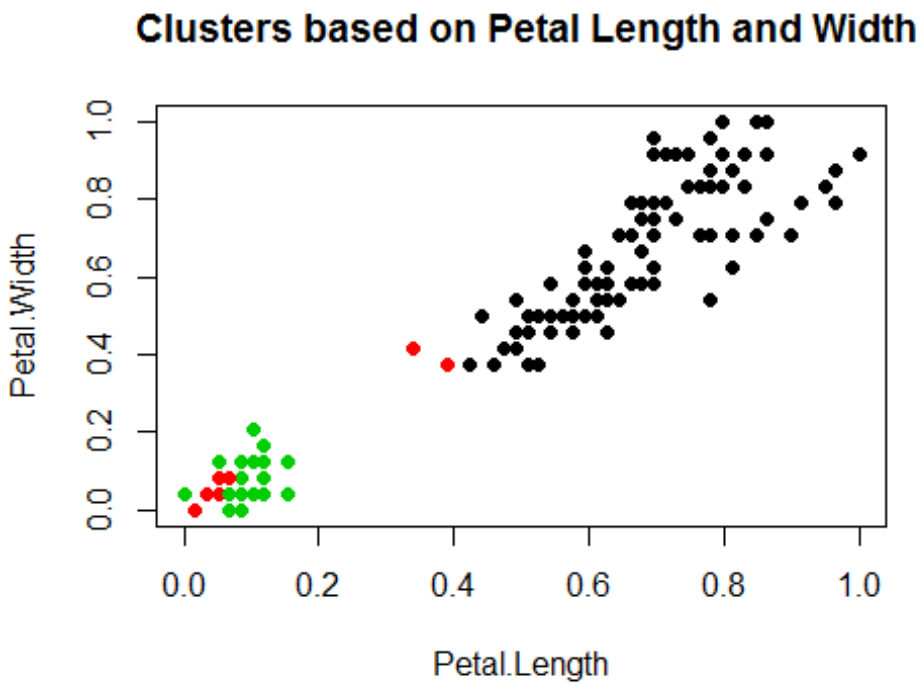
```
result<- kmeans(iris[,c(1,2,3,4)],3)


plot(iris[c(1,2)], col=result$cluster, pch = 19, main =" Clusters based on
Sepal Length and Width")
```



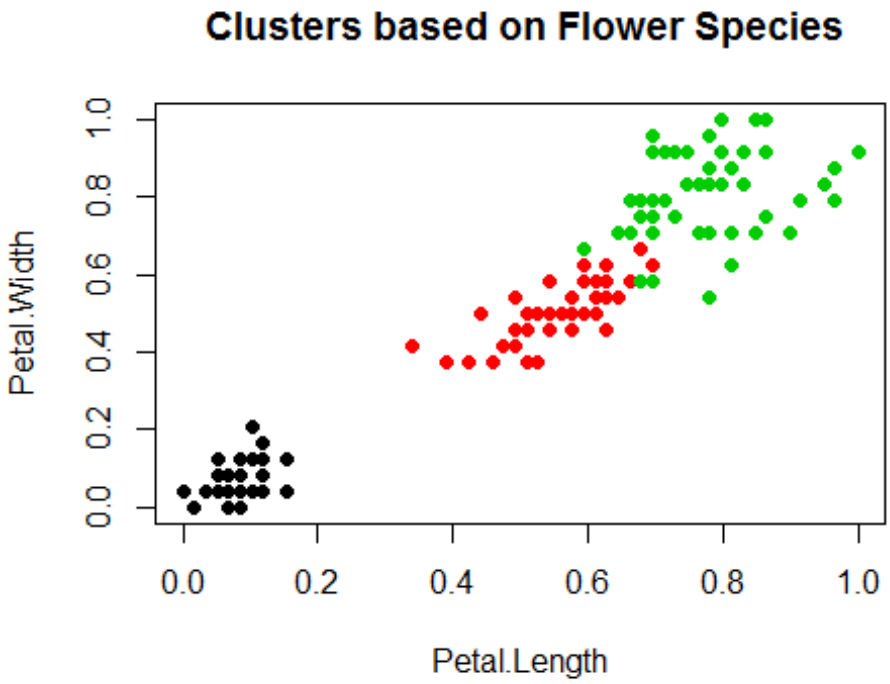

```
plot(iris[c(1,2)], col=iris$Species, pch = 19, main =" Clusters based on
Flower Species")
```



```
plot(iris[c(3,4)], col=result$cluster, pch = 19, main =" Clusters based on
Petal Length and Width")
```

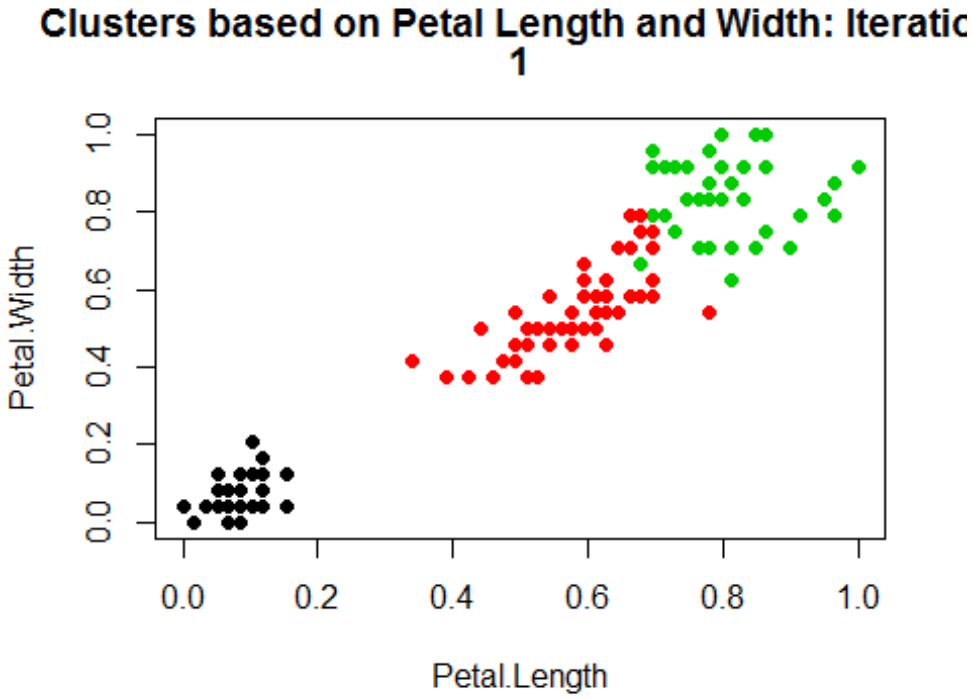


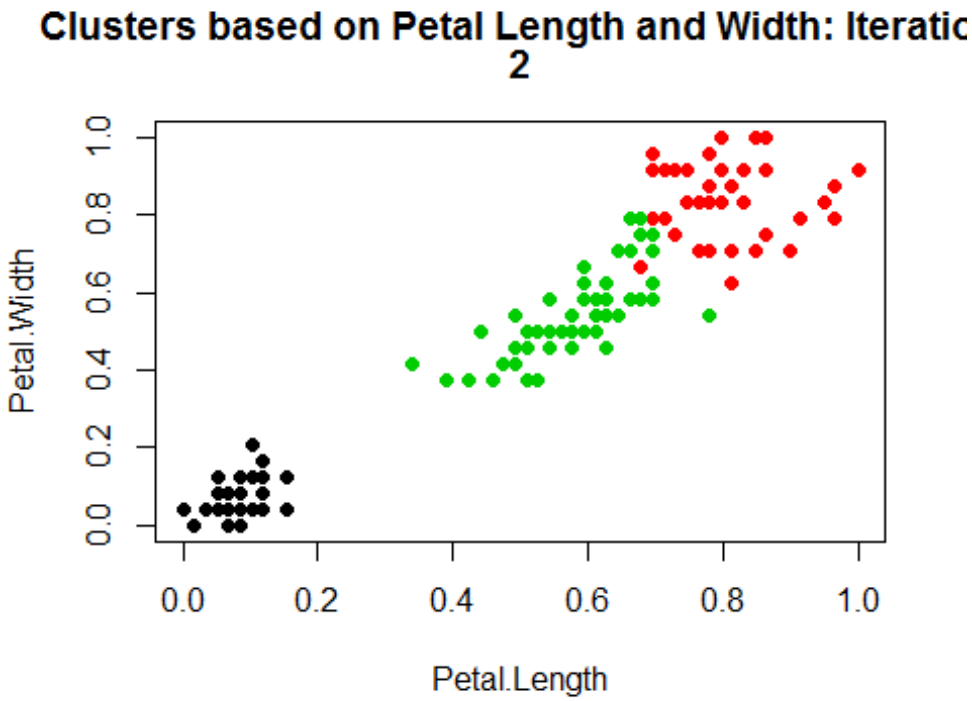

```
plot(iris[c(3,4)], col=iris$Species, pch = 19, main =" Clusters based on
Flower Species")
```

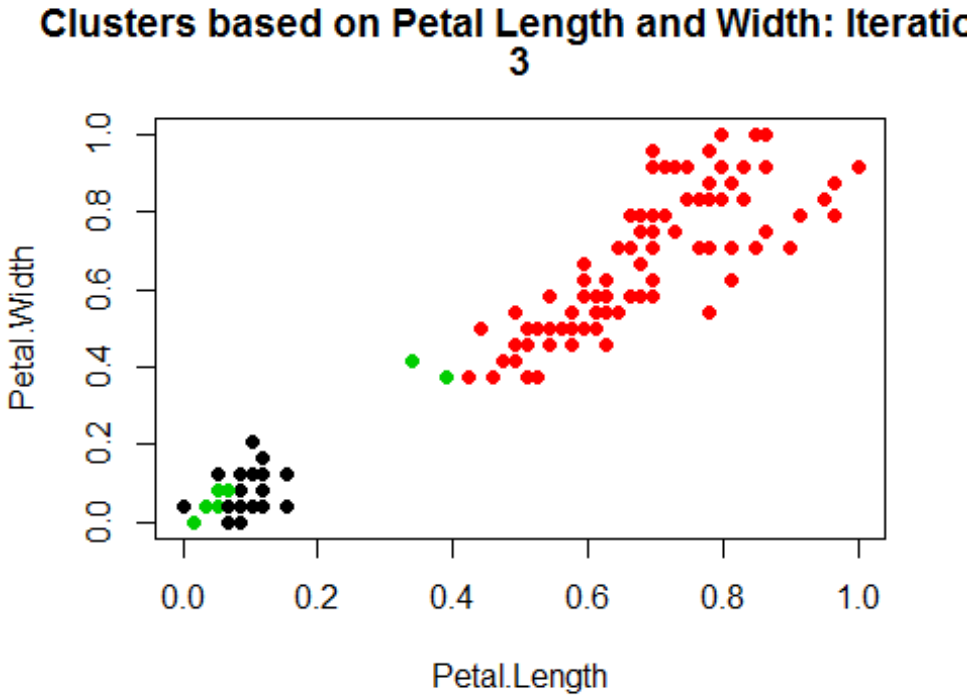


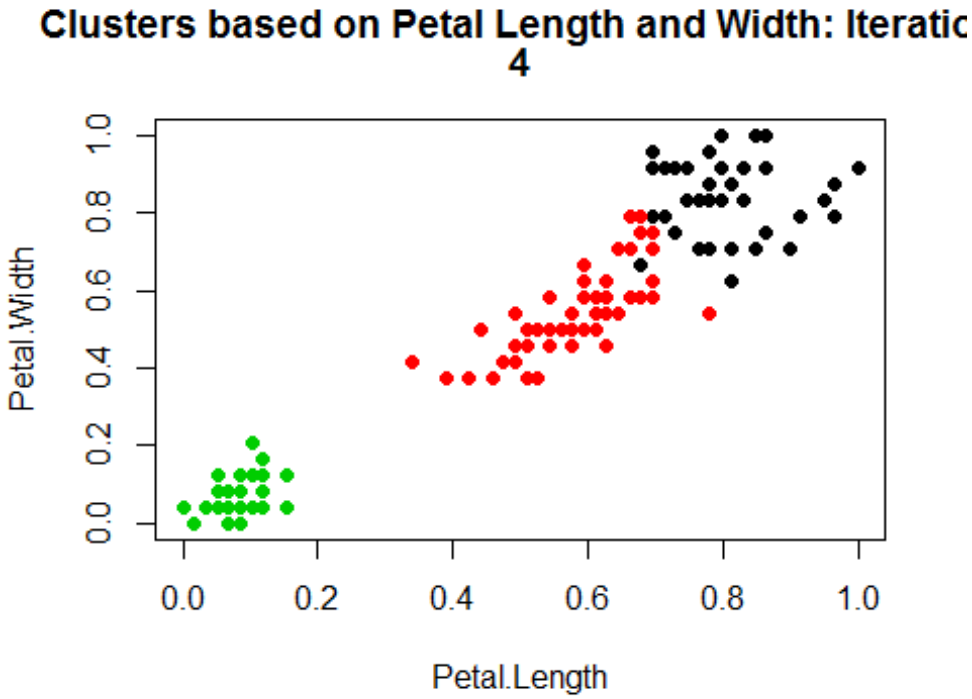
Running the k means algorithms four times on the same dataset

```
for (i in 1:4) {
  set.seed(100 +50*i)
  result<- kmeans(iris[,c(1,2,3,4)],3)
plot(iris[c(3,4)], col=result$cluster, pch = 19, main =c(" Clusters based on
Petal Length and Width: Iteration ", i)) }
```


Clusters based on Petal Length and Width: Iteratic
1
Petal.Width
Petal.Length
0.0
0.2
0.4
0.6
0.8
1.0


Clusters based on Petal Length and Width: Iteratic
2
Petal.Width
Petal.Length
0.0
0.2
0.4
0.6
0.8
1.0





It shows the importance of multiple iteration and averaging them

End of the Script