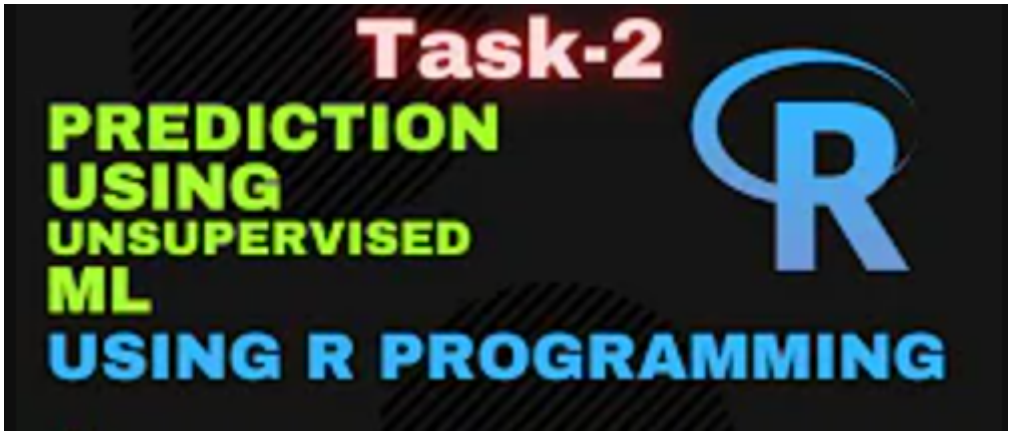
Task-2 SPARK FOUNDATION INTERNSHIP



Prediction Using Unsupervised Machine learning

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

INTRODUCTION : "IRIS" dataset

Iris dataset contains five columns such as:

Petal Length Petal Width Sepal Length Sepal Width Species

Iris is a flowering plant, the researchers have measured various features of the different iris flowers and recorded digitally The iris dataset contains the following data

50 samples of 3 different species of iris (150 samples total)

There are 50 observations of each species for a total of 150 observations with 4 features each (sepal length, sepal width, petal length, petal width).

Measurements: sepal length, sepal width, petal length, petal width

The format for the data: (sepal length, sepal width, petal length, petal width)

The 4 features are

SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm

The target class The flower species type is the target class and it having 3 types

Setosa Versicolor Virginica

In [1]:
```
## loading Data
data(iris)
```

In [2]:
```
#structuring Data set
str(iris)
```

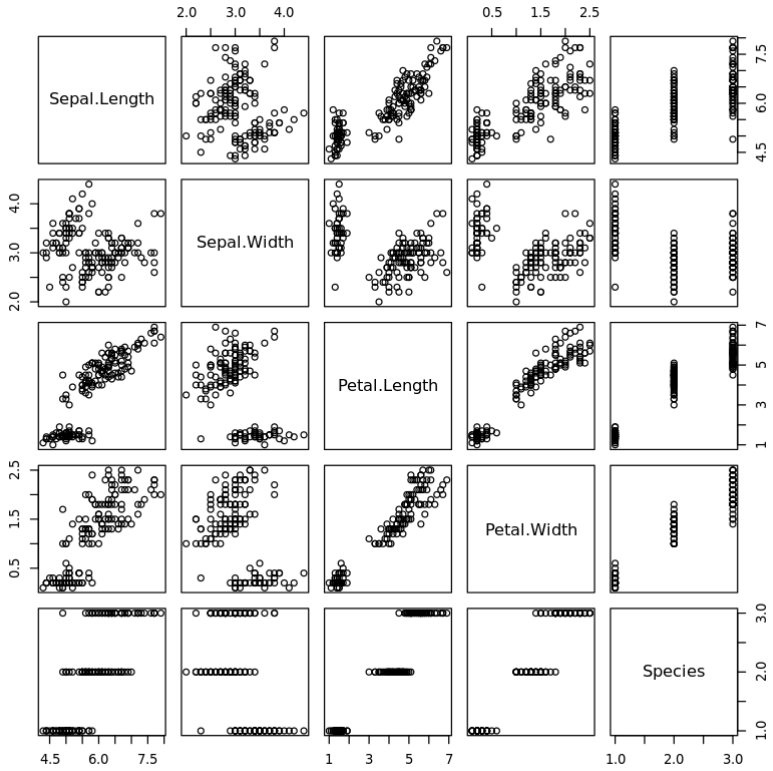
```
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1
...
```

In [9]:
```
#loading packages
library(cluster)
```

In [26]:
```
plot(iris)
    head(iris)
```

A data.frame: 6 × 5

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
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



In [14]:
```
#removing labels from given Data set
iris_1=iris[,-5]
```

In [20]:
```
#fitting K means Clustering
set.seed(200)
k.cluster=kmeans(iris_1,centers=3,nstart=20)
k.cluster
```

```
K-means clustering with 3 clusters of sizes 38, 50, 62

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.850000    3.073684     5.742105    2.071053
2     5.006000    3.428000     1.462000    0.246000
3     5.901613    2.748387     4.393548    1.433871

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
 [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 1 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 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1 1 1 3 1 1 1 1
[112] 1 1 3 3 1 1 1 1 3 1 3 1 3 1 1 3 3 1 1 1 1 3 1 1 1 3 1 1 1 3 1 1 1 3 1
[149] 1 3

Within cluster sum of squares by cluster:
[1] 23.87947 15.15100 39.82097
 (between_SS / total_SS =  88.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

In [21]:
```
# cluster Identification
k.cluster$cluster
```

```
2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·2·
```

2 · 2 · 2 · 2 · 2 · 2 · 2 · 2 · 2 · 2 · 2 · 2 · 2 · 2 · 2 · 2 · 2 · 3 · 3 · 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 · 1 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 3 · 1 · 3 · 1 · 1 · 1 · 1 · 3 · 1 · 1 · 1 · 1 · 1 · 1 · 3 · 3 · 1 · 1 · 1 · 1 · 3 · 1 · 3 · 1 · 3 · 1 · 1 · 3 · 3 · 1 · 1 · 1 · 1 · 1 · 3 · 1 · 1 · 1 · 1 · 3 · 1 · 1 · 1 · 3 · 1 · 1 · 1 · 3 · 1 · 1 · 3

model Achieved the Accuracy of 100 % with P-Value less than 1 which indicates that model is a good fit for prediction

In [22]:

```
# confusion matrix
cmat=table(iris$Species,k.cluster$cluster)
cmat
```

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

Above Matrix States that

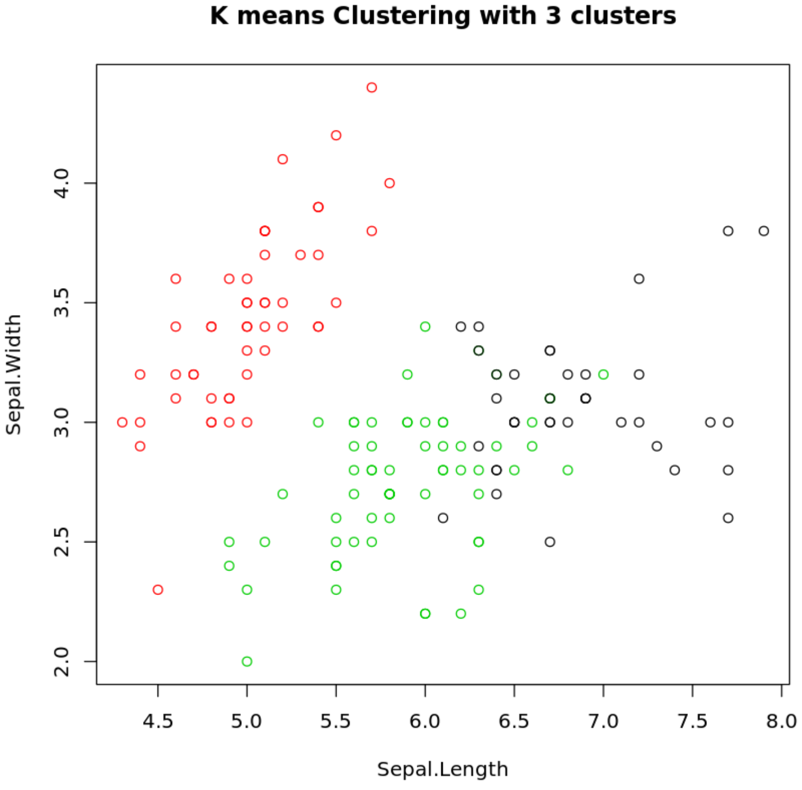
out of 38 Setosa are 2 are classified as vericolor and rest as verginica

out of 50 versicolor all are classified correctly under versicolor

out of 62 verginica 48 are classified under versicolor and rest as viriginca

In [25]:

```
# model Evaluation and visualisation
plot(iris_1[c("Sepal.Length","Sepal.Width")],col=k.cluster$cluster,main = "K means C
```



In [30]:

```
# plotting cluster Centers

k.cluster$centers
k.cluster$centers[,c("Sepal.Length","Sepal.Width")]
```

A matrix: 3 × 4 of type dbl

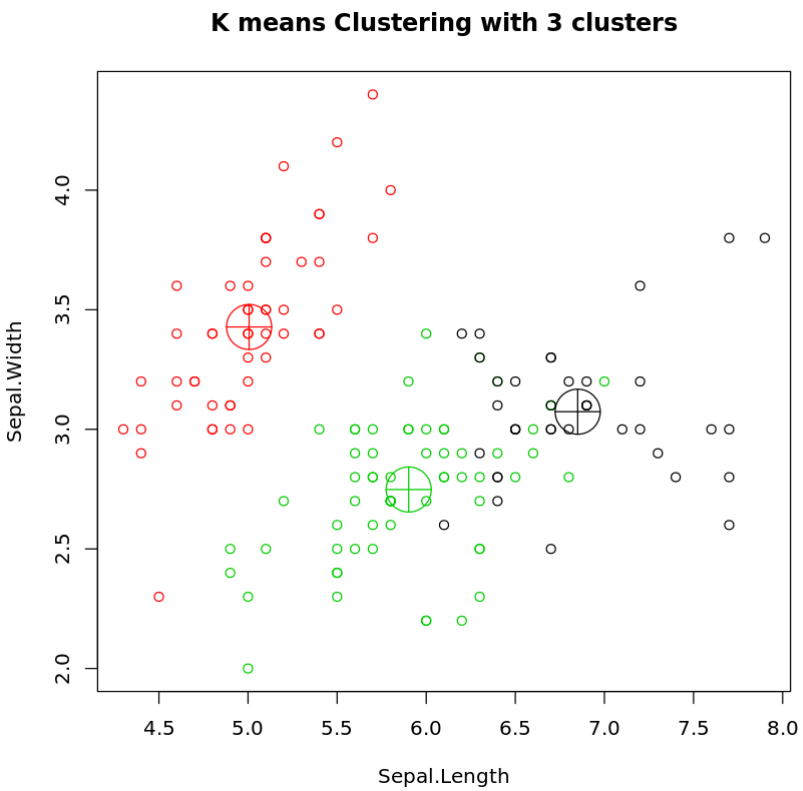
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	6.850000	3.073684	5.742105	2.071053
2	5.006000	3.428000	1.462000	0.246000
3	5.901613	2.748387	4.393548	1.433871

A matrix: 3 × 2 of type dbl

	Sepal.Length	Sepal.Width
1	6.850000	3.073684
2	5.006000	3.428000
3	5.901613	2.748387

In [31]:

```
plot(iris_1[c("Sepal.Length","Sepal.Width")],
     col=k.cluster$cluster,
     main = "K means Clustering with 3 clusters")+
points(k.cluster$centers[,c("Sepal.Length","Sepal.Width")],
      col=1:3 , pch= 10 , cex = 5)
```

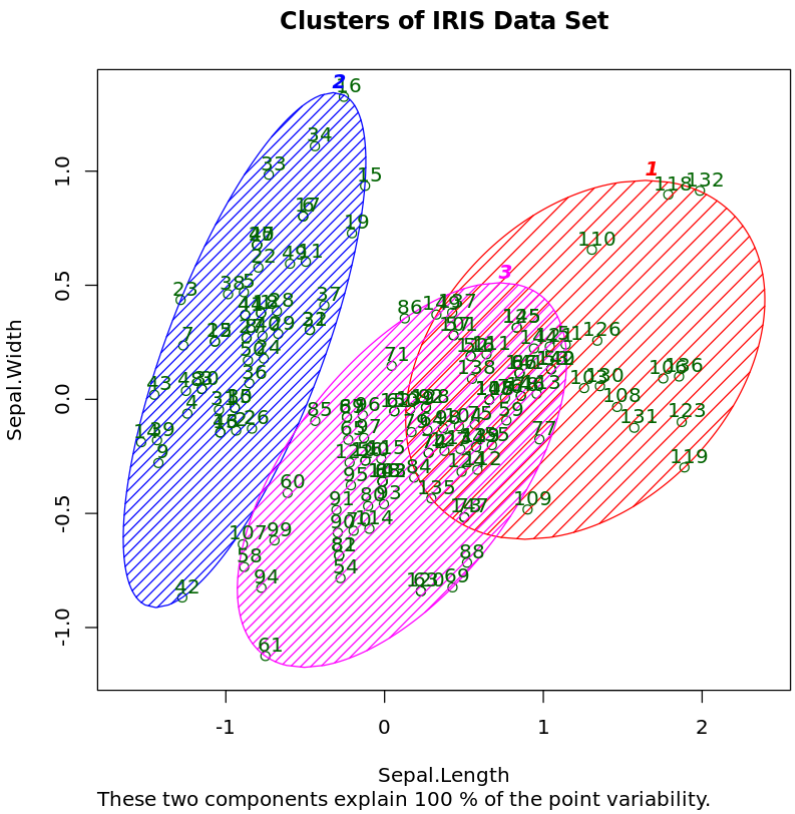


In [34]:

```
# visualising clusters

k=k.cluster$cluster
clusplot(iris_1[,c("Sepal.Length","Sepal.Width")],
         k,
         lines= 0,
         shade=TRUE,
         color = TRUE,
         labels=2,
         plotchar=FALSE,
         span=TRUE,
```

```
        main=paste("Clusters of IRIS Data Set"),
        xlab="Sepal.Length",
        ylab="Sepal.Width")
```

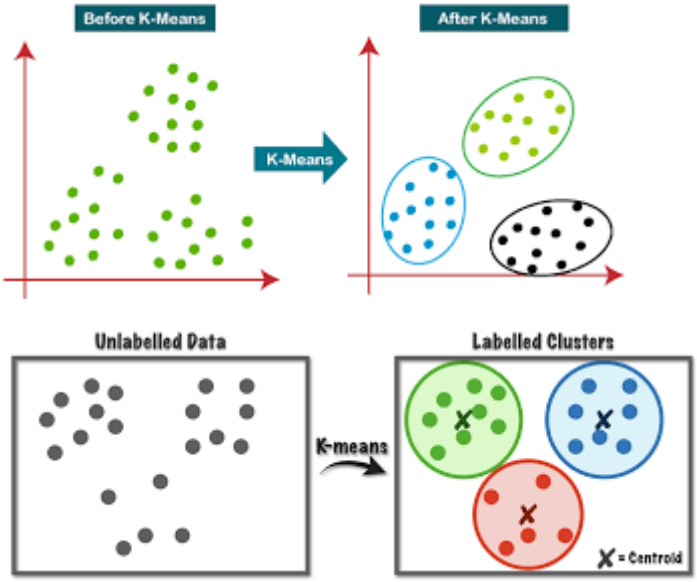


These two components explain 100 % of the point variability.

3 kmean clusters are formed with 100% point variability this means that

I was able to successfully carry-out prediction using Unsupervised Machine Learning task

and was able to evaluate the model's clustering accuracy score.



In [ ]:

In [ ]: