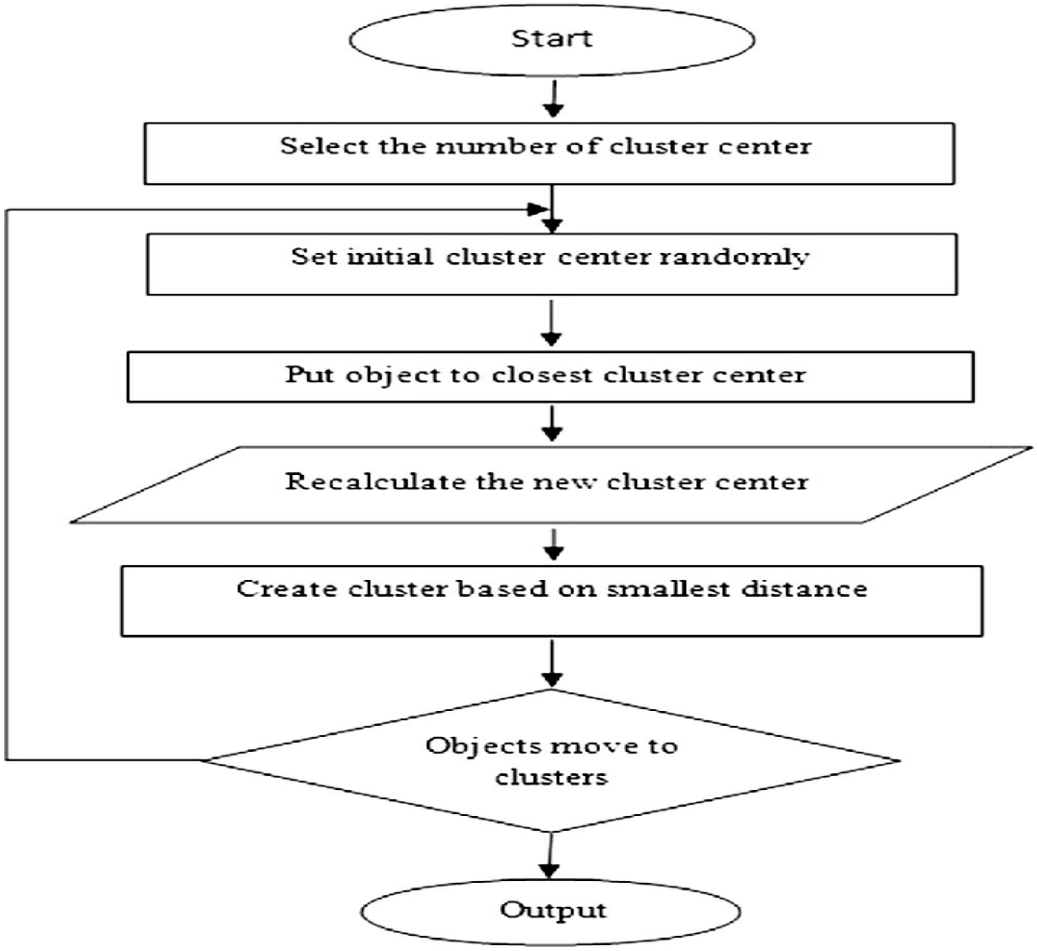
Practical No.8

Perform the Data Clustering using Clustering Algorithm(Clustering : k-means Algorithm)

Clustering

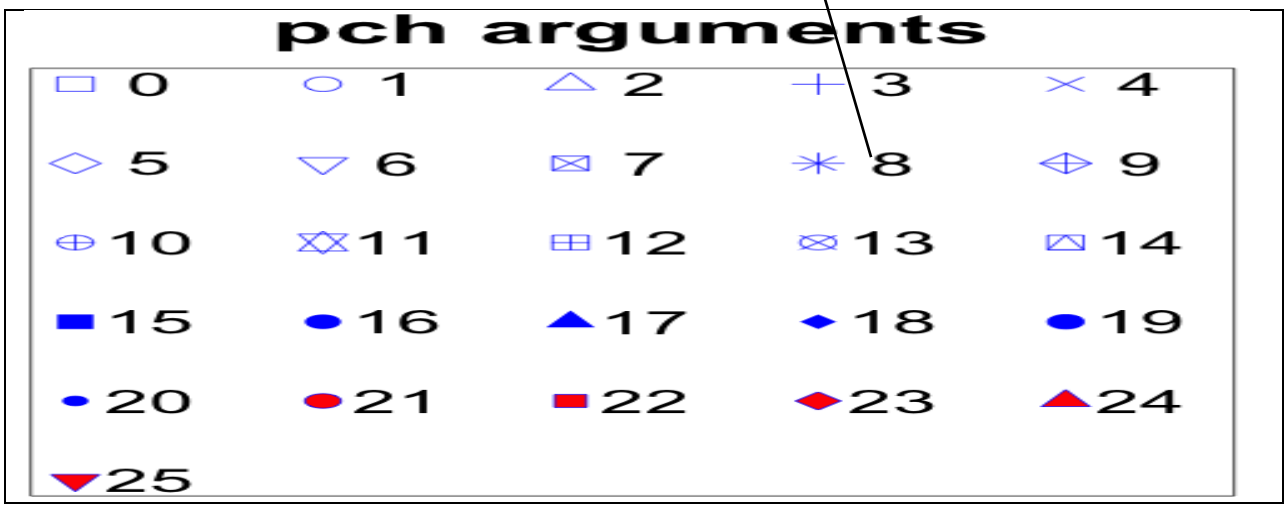
- **Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups.**
- **Let's understand this with an example.**
- **Suppose, you are the head of a rental store and wish to understand preferences of your Customers to scale up your business.**
- **Is it possible for you to look at details of each costumer and devise a unique business strategy for each one of them?**
- **Definitely not. But, what you can do is to cluster all of your Customers into say 10 groups based on their purchasing habits and use a separate strategy for costumers in each of these 10 groups. And this is what we call clustering.**

K-Means Clustering



Iris is a genus of 260–300 species of flowering plants with showy flowers.


```
newiris <- iris
newiris$Species <- NULL
(kc <- kmeans(newiris,3))
print(kc)
# Compare the Species label with the clustering result.
table (iris$Species,kc$cluster)
plot(newiris[c("Sepal.Length","Sepal.Width")],col=kc$cluster)

points(kc$centers[,c("Sepal.Length","Sepal.Width")],col=1:3,pch=8,cex=2)
```

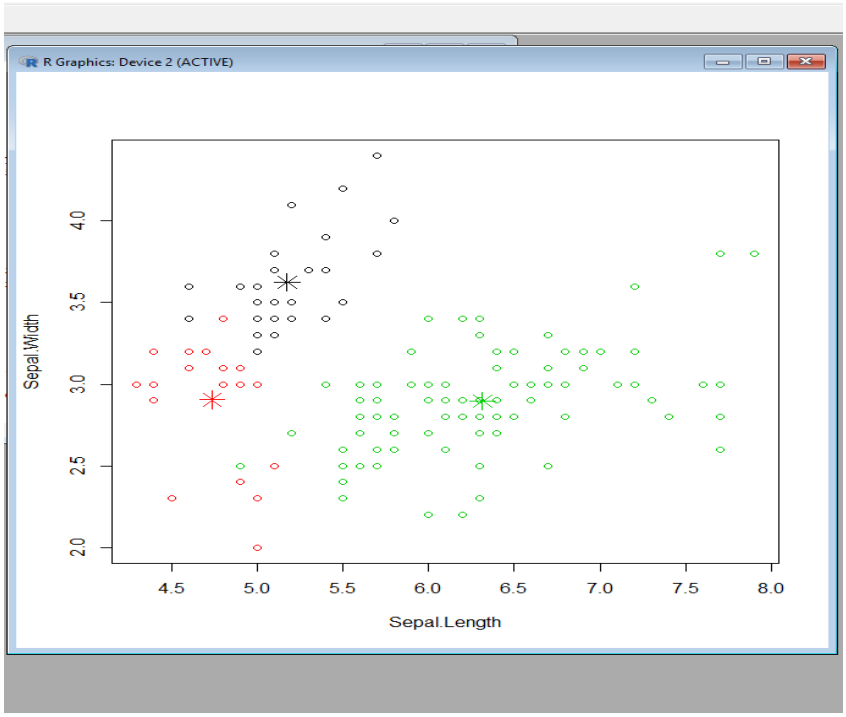


R Console

```
> # Apply K mean to iris and store result
> newiris <- iris
> newiris$Species <- NULL
> (kc <- kmeans(newiris,3))
K-means clustering with 3 clusters of sizes 62, 38, 50

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

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 3 3 3 3 3 3
 [47] 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 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [93] 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2
[139] 1 2 2 2 1 2 2 2 1 2 2 1

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

Available components:

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

R Console

```
> #Compare the Species label with the clustering result
> table (iris$Species,kc$cluster)

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


R Graphics: Device 2 (ACTIVE)
Sepal.Width
Sepal.Length
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0