K-Means Clustering in R Programming

K-means in R-programming is an Unsupervised Non-linear algorithm that cluster data based on similarity or similar groups. It seeks to partition the observations into a pre-specified number of clusters. Segmentation of data takes place to assign each training example to a segment called a cluster. In the unsupervised algorithm, high reliance on raw data is given with large expenditure on manual review for review of relevance is given. It is used in a variety of fields like Banking, healthcare, retail, Media, etc.

Theory
K-Means clustering groups the data on similar groups. The algorithm is as follows:

1. Choose the number **K** clusters.
2. Select at random K points, the centroids(Not necessarily from the given data).
3. Assign each data point to closest centroid that forms K clusters.
4. Compute and place the new centroid of each centroid.
5. Reassign each data point to new cluster.

After final reassignment, name the cluster as Final cluster.

The Dataset
Iris dataset consists of 50 samples from each of 3 species of Iris(Iris setosa, Iris virginica, Iris versicolor) and a multivariate dataset introduced by British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems. Four features were measured from each sample i.e length and width of the sepals and petals and based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

CODE:

Performing K-Means Clustering on Dataset
Using K-Means Clustering algorithm on the dataset which includes 11 persons and 6 variables or attributes

```
#Install
install.packages(“stats”)
install.packages(“dplyr”)
install.packages(“ggplot2”)
install.packages(“ggfortify”)

#Load
library(stats)
library(dplyr)
library(ggplot2)
library(ggfortify)

#Unsupervised
view(iris)
mydata = select(iris,c(1,2,3,4))

#WSS Plot
Wssplot <- function(data, nc = 15, seed = 1234)
{
 wss <- (nrow(data)-1)*sum(apply(data,2,var))
 for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
plot(1:nc, wss, type=”b”, xlab= “Number of clusters”, ylab = “within groups sum of square”)
}

#Wss plot to maximum number of clusters
wssplot(mydata)

#K-means cluster
KM = kmeans(mydata,2)

#Cluster plot
autoplot(KM,mydata,frame= TRUE)

#Cluster centre
KM$centers
```