**PROJECT REPORT**

**Varun Kavalipurapu**

**rot893 /(@01519517)**

**Description:** In this assignment, we are focusing on the clustering algorithms and for this we are choosing one data set and applying kmeans clustering algorithms on it. To implement the algorithm, I am using a statistical tool called R programming.

**Partition-based clustering:**

Data clustering, by deﬁnition, is an exploratory and descriptive data analysis technique, which has gained a lot of attention. Partitioning-based clustering methods are ﬂexible methods based on iterative relocation of data points between clusters.

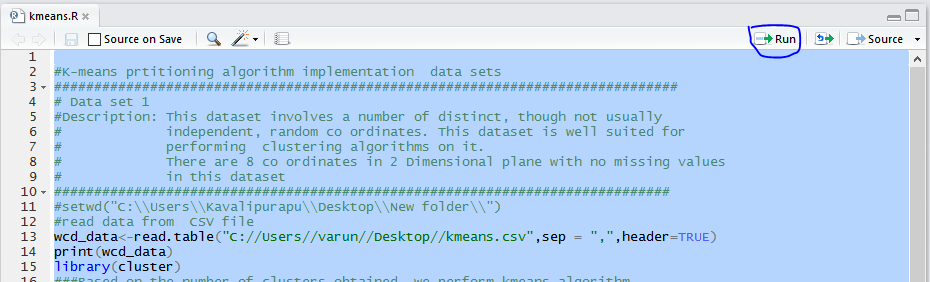
**Packages used:**

**K-means:**

Nbclust: This package is used for determining the best amount of clusters. It provides 30 indices for determining the optimal number of clusters in a dataset. We used this package to find the number of clusters for the dataset.

**Steps to run the program:**

* Download kmeans.R and input.csv files into a folder
* Install Rstudio or rtools to run the program
* Edit the first line of the file “kmeans.R” to set the directory path to “input.csv”
* Select all lines in kmeans.R file and hit run button on the top as shown below

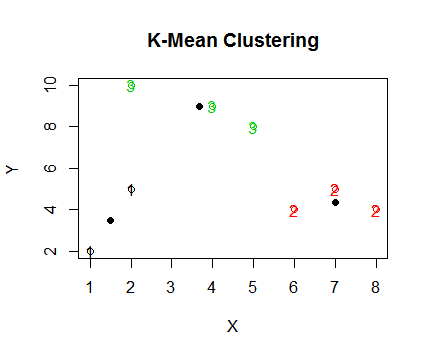


**Output :**

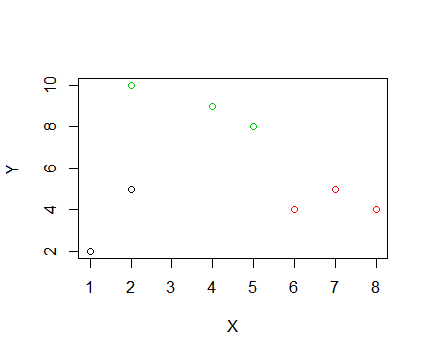
* Input given to this algorithm was a set of 8 co ordinates
* We have given the number of clusters as 3.
* Output is as follows

**Screenshots of output :**

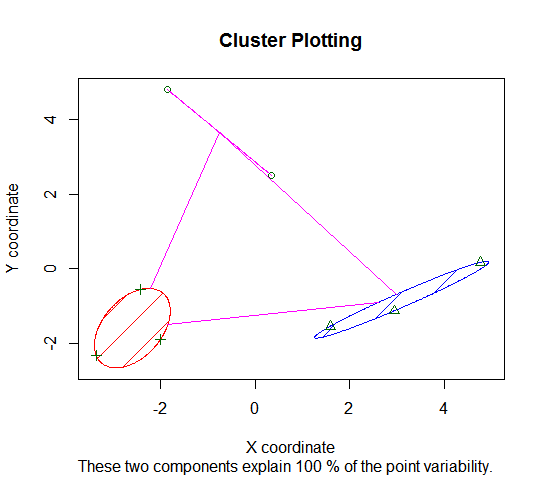
All the points were divided into three clusters as shown below

****

* Three different colors are showed to differentiate the clusters with 3 points in cluster1, 2 points in cluster 2, and 3 points in cluster 3.



* If all the points are plotted in a cluster, they look like as following



**Source code:**

#K-means prtitioning algorithm implementation

##############################################################################

# Data set 1

#Description: This dataset involves a number of distinct, though not usually

# independent, random co ordinates. This dataset is well suited for

# performing clustering algorithms on it.

# There are 8 co ordinates in 2 Dimensional plane with no missing values

# in this dataset

#############################################################################

#setwd("C:\\Users\\Kavalipurapu\\Desktop\\New folder\\")

#read data from CSV file

wcd\_data<-read.table("C://Users//varun//Desktop//kmeans.csv",sep = ",",header=TRUE)

print(wcd\_data)

library(cluster)

###Based on the number of clusters obtained, we perform kmeans algorithm

# Start the clock!

ptm <- proc.time()

kmeans <- kmeans(wcd\_data, 3, method=manhattan)

# Stop the clock

proc.time() - ptm

print(kmeans)

kmeans$size

kmeans$centers

plot(wcd\_data,col=kmeans$cluster)

points(kmeans$centers,pch=32)

plot(wcd\_data[c("X","Y")],col=kmeans$cluster,main="K-Mean Clustering")

text(wcd\_data, labels=kmeans$cluster, col=kmeans$cluster)

points(kmeans$centers,pch=16)

clusplot(wcd\_data,kmeans$cluster,main='Cluster Plotting',color=TRUE,shade=TRUE,xlab = "X coordinate", ylab = "Y coordinate")

##Calculating Time complexity for the dataset K-Means

# Start the clock!

ptm <- proc.time()

kmeans <- kmeans(wcd\_data, 3,nstart=25)

# Stop the clock