Perform Clustering for the crime data and identify the number of clusters formed and draw inferences.

BUSSINESS PROBLEM:

Perform k means to find the value of k and draw the inferences

PREPROCESSING:

Access only numerical values and standardize the values for the further process.

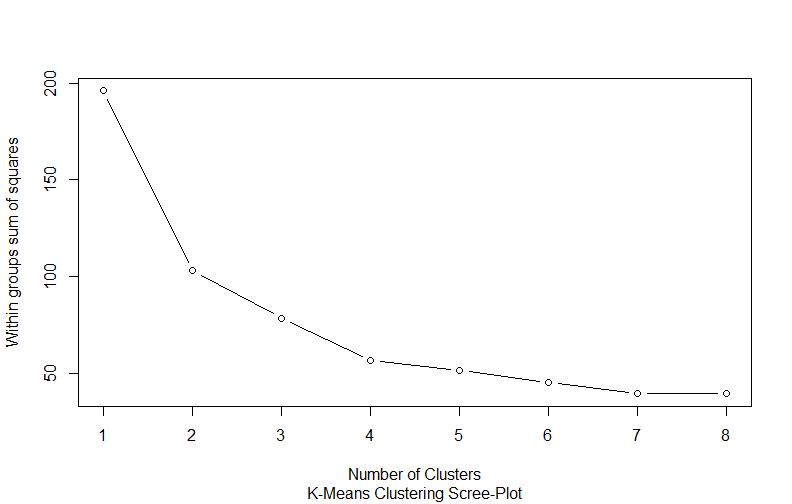
K selection:

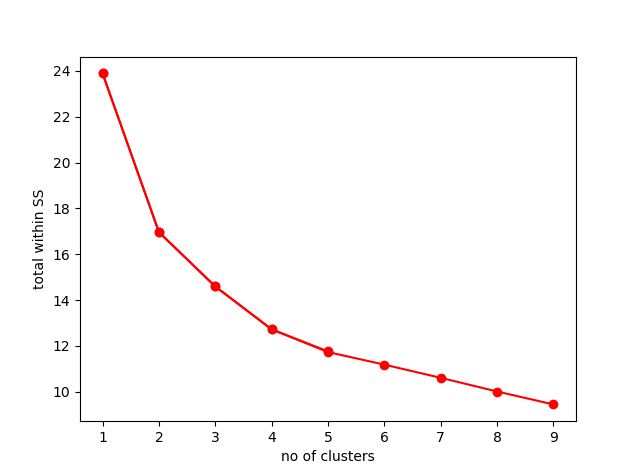
By Using K-Selection we can know the best value to be used .

k <-k selection(mydata[,-1], parallel = TRUE, k\_threshold = 0.9, max\_centers=20)

It is giving the value of k as 2.

SCREWPLOT OR ELBOW CURVE:





By looking at the scree plot we can take the k value based on the elbow in the plot.

I have selected k =4 to form 4 clusters.

Take k value which will be best

Taking k value as 2:

km <- kmeans(norm\_data,2)

km

Within cluster sum of squares by cluster:

56.11445 46.74796

(between\_SS / total\_SS = 47.5 %)

Taking k value as 3:

km3 <- kmeans(norm\_data,3)

km3

Within cluster sum of squares by cluster:

19.922437 8.316061 53.354791

(between\_SS / total\_SS = 58.4 %)

Taking k value as 4:

km4 <- kmeans(norm\_data,4)

km4

Within cluster sum of squares by cluster:

8.316061 11.952463 16.212213 19.922437

(between\_SS / total\_SS = 71.2 %)

taking 4 clusters will be good for the model.

km4$size

8 13 16 13

K-means clustering with 4 clusters of sizes 13, 16, 13, 8

Cluster 1-13

Cluster 2-16

Cluster3-13

Cluster4-8

cluster Murder Assault Urban Pop Rape

1 3.60000 78.53846 52.07692 12.17692

2 5.65625 138.87500 73.87500 18.78125

3 10.81538 257.38462 76.00000 33.19231

4 13.93750 243.62500 53.75000 21.41250

Based on the data, cluster 1 with less no of murders, assaults, urban pops& rapes is good.

Clust plot:

