Clustering

K means

Assignment 2

BP: Perform clustering for the crime data and identify the number of clusters formed and draw inferences.

PROCEDURE:

STEP 1: First we have to Exploratory Data Analysis which can be done by plotting scattered plot, box plots and summary.

summary(crime\_data\_1\_)

X1 Murder Assault UrbanPop

Length: 50 Min. : 0.800 Min. : 45.0 Min. :32.00

Class: character 1st Qu.: 4.075 1st Qu.:109.0 1st Qu.:54.50

Mode : character Median : 7.250 Median :159.0 Median :66.00

Mean : 7.788 Mean :170.8 Mean :65.54

3rd Qu.:11.250 3rd Qu.:249.0 3rd Qu.:77.75

Max. :17.400 Max. :337.0 Max. :91.00

Rape

Min. : 7.30

1st Qu.:15.07

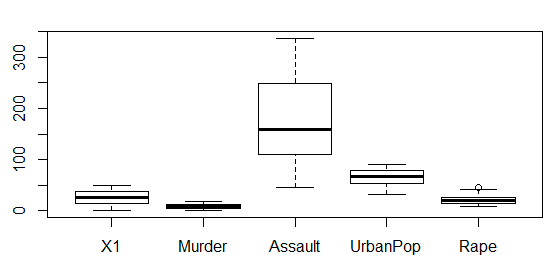
Median :20.10

Mean :21.23

3rd Qu.:26.18

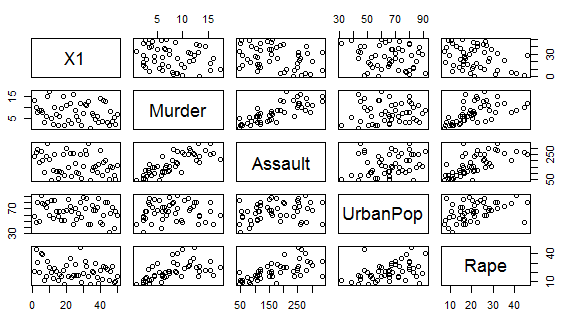
Max. :46.00

From the summary we can see that in assault the difference between mean and max is large so it may be right skewed it can my confirmed by using Box Plot .



It can be confirmed that Assault is right Skewed.

Scatter Plot for the Following data frame is:



STEP 2: Now we can observe from the summary that the data is not in similar scale so now we have to normalize the data into same scale which can be done as follows:

normzalizeddata <- scale(crime\_data\_1\_[,2:5])

View(normzalizeddata)

summary(normzalizeddata)

Murder Assault UrbanPop Rape

Min. :-1.6044 Min. :-1.5090 Min. :-2.31714 Min. :-1.4874

1st Qu.:-0.8525 1st Qu.:-0.7411 1st Qu.:-0.76271 1st Qu.:-0.6574

Median :-0.1235 Median :-0.1411 Median : 0.03178 Median :-0.1209

Mean : 0.0000 Mean : 0.0000 Mean : 0.00000 Mean : 0.0000

3rd Qu.: 0.7949 3rd Qu.: 0.9388 3rd Qu.: 0.84354 3rd Qu.: 0.5277

Max. : 2.2069 Max. : 1.9948 Max. : 1.75892 Max. : 2.6444

From the summary we can see that whole data is converted into Z Scale.

STEP 3: To determine the value of k, we have to do code for elbow curve :

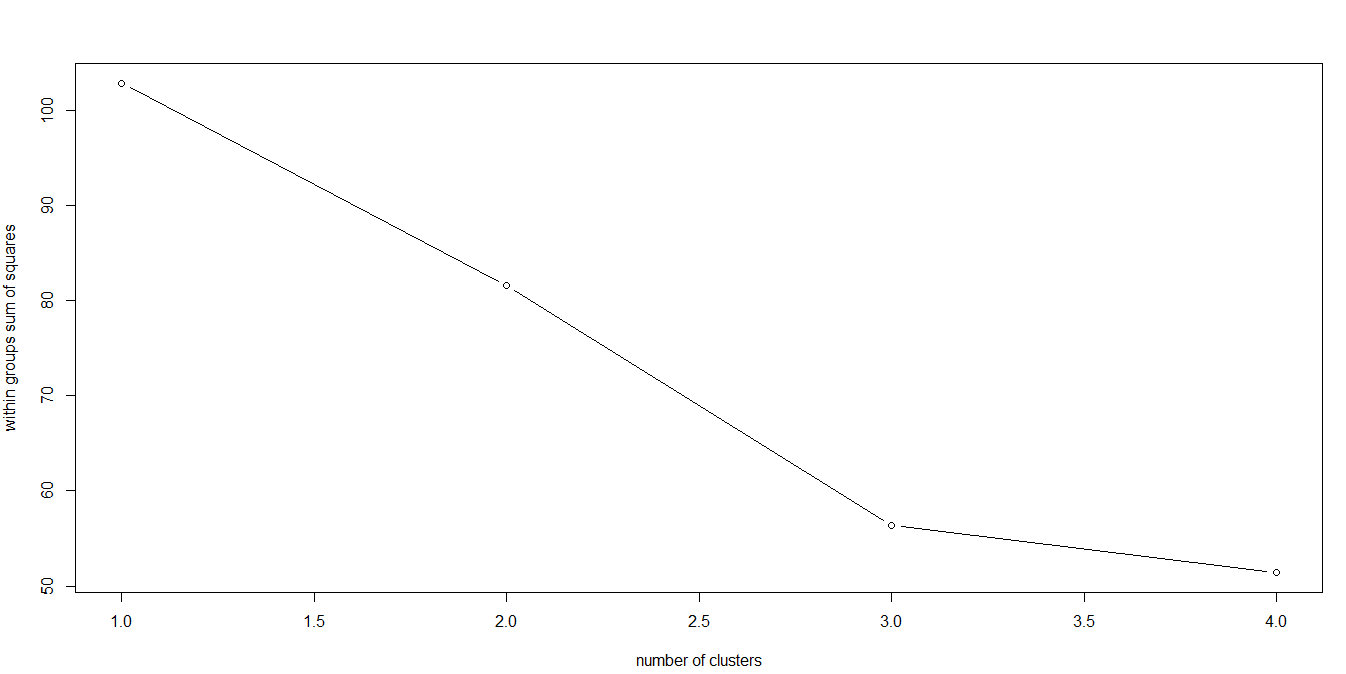
for (i in 1:5) { wss <- c(wss,kmeans(nd,centers=i)$tot.withins)

}

> wss

[1] 102.86240 81.59329 56.40317 51.42912

> plot(wss, type = "b" , xlab ="number of clusters", ylab = "within groups sum of squares")



From the elbow curve we can assume that we can form 2 clusters, to confirm our assumption we can use the auto selection of k by code :

|  |  |  |
| --- | --- | --- |
| |  | | --- | | k <- kselection(crime\_data\_1\_[,2:5] , parallel = TRUE , max\_centers=12)  > k  f(k) finds 2 clusters | |  | |
|  |
| |  | | --- | | Thus we can conclude that we can form 2 clusters. | |

STEP 3: Now we have to cluster it now:

clust <- kmeans(nd,2)

> str(clust)

List of 9

$ cluster : int [1:50] 1 1 1 2 1 1 2 2 1 1 ...

$ centers : num [1:2, 1:4] 1.005 -0.67 1.014 -0.676 0.198 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : chr [1:2] "1" "2"

.. ..$ : chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"

$ totss : num 196

$ withinss : num [1:2] 46.7 56.1

$ tot.withinss: num 103

$ betweenss : num 93.1

$ size : int [1:2] 20 30

$ iter : int 1

$ ifault : int 0

- attr(\*, "class")= chr "kmeans"

> clust$centers

Murder Assault UrbanPop Rape

1 1.004934 1.0138274 0.1975853 0.8469650

2 -0.669956 -0.6758849 -0.1317235 -0.5646433

> clust$cluster

[1] 1 1 1 2 1 1 2 2 1 1 2 2 1 2 2 2 2 1 2 1 2 1 2 1 1 2 2 1 2 2 1 1 1 2 2 2 2 2 2 1

[41] 2 1 1 2 2 2 2 2 2 2

final <- data.frame(clust$cluster , crime\_data\_1\_)

> View(final)

> final

clust.cluster X1 Murder Assault UrbanPop Rape

1 1 Alabama 13.2 236 58 21.2

2 1 Alaska 10.0 263 48 44.5

3 1 Arizona 8.1 294 80 31.0

4 2 Arkansas 8.8 190 50 19.5

5 1 California 9.0 276 91 40.6

6 1 Colorado 7.9 204 78 38.7

7 2 Connecticut 3.3 110 77 11.1

8 2 Delaware 5.9 238 72 15.8

9 1 Florida 15.4 335 80 31.9

10 1 Georgia 17.4 211 60 25.8

11 2 Hawaii 5.3 46 83 20.2

12 2 Idaho 2.6 120 54 14.2

13 1 Illinois 10.4 249 83 24.0

14 2 Indiana 7.2 113 65 21.0

15 2 Iowa 2.2 56 57 11.3

16 2 Kansas 6.0 115 66 18.0

17 2 Kentucky 9.7 109 52 16.3

18 1 Louisiana 15.4 249 66 22.2

19 2 Maine 2.1 83 51 7.8

20 1 Maryland 11.3 300 67 27.8

21 2 Massachusetts 4.4 149 85 16.3

22 1 Michigan 12.1 255 74 35.1

23 2 Minnesota 2.7 72 66 14.9

24 1 Mississippi 16.1 259 44 17.1

25 1 Missouri 9.0 178 70 28.2

26 2 Montana 6.0 109 53 16.4

27 2 Nebraska 4.3 102 62 16.5

28 1 Nevada 12.2 252 81 46.0

29 2 New Hampshire 2.1 57 56 9.5

30 2 New Jersey 7.4 159 89 18.8

31 1 New Mexico 11.4 285 70 32.1

32 1 New York 11.1 254 86 26.1

33 1 North Carolina 13.0 337 45 16.1

34 2 North Dakota 0.8 45 44 7.3

35 2 Ohio 7.3 120 75 21.4

36 2 Oklahoma 6.6 151 68 20.0

37 2 Oregon 4.9 159 67 29.3

38 2 Pennsylvania 6.3 106 72 14.9

39 2 Rhode Island 3.4 174 87 8.3

40 1 South Carolina 14.4 279 48 22.5

41 2 South Dakota 3.8 86 45 12.8

42 1 Tennessee 13.2 188 59 26.9

43 1 Texas 12.7 201 80 25.5

44 2 Utah 3.2 120 80 22.9

45 2 Vermont 2.2 48 32 11.2

46 2 Virginia 8.5 156 63 20.7

47 2 Washington 4.0 145 73 26.2

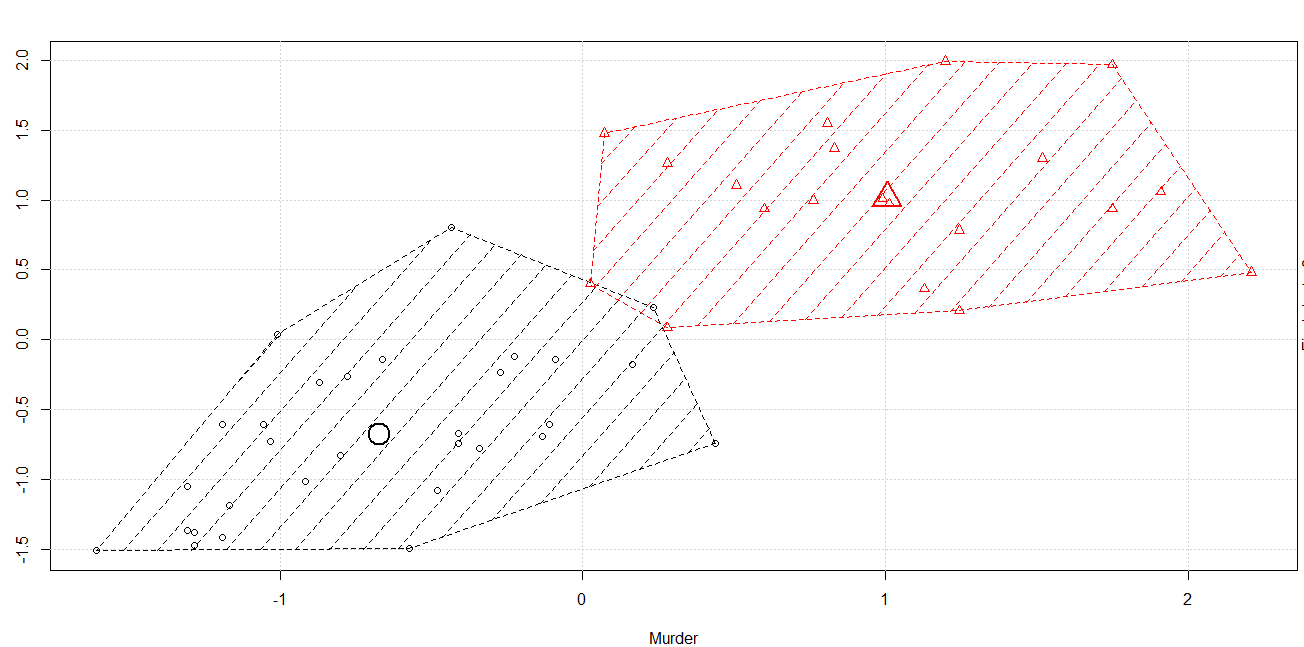
48 2 West Virginia 5.7 81 39 9.3

49 2 Wisconsin 2.6 53 66 10.8

50 2 Wyoming 6.8 161 60 15.6

STEP 4: Now we can visualize the clusters in graphical form by:

km <- kmeans.ani(nd,2)



Therefore, out of 50 observations from the data frame with k clustering method we have formed them into two clusters.

The inferences which we can make from this is that cluster 1 consists of high murder rate and assault compared to that of cluster 2 whereas rape and urban pop is evenly distributed.