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

Data Description:

Murder -- Muder rates in different places of United States

Assualt- Assualt rate in different places of United States

UrbanPop - urban population in different places of United States

Rape - Rape rate in different places of United States

**Ans**:

> library(readr)

> crime\_data <- read\_csv("crime\_data.csv")

> head(crime\_data)

# A tibble: 6 x 5

X1 Murder Assault UrbanPop Rape

<chr> <dbl> <dbl> <dbl> <dbl>

1 Alabama 13.2 236 58 21.2

2 Alaska 10 263 48 44.5

3 Arizona 8.1 294 80 31

4 Arkansas 8.8 190 50 19.5

5 California 9 276 91 40.6

6 Colorado 7.9 204 78 38.7

**Convert data Into Normalized form.**

> normalized\_data<-scale(crime\_data[,2:5])

**Distance matrix Formation**

> d <- dist(normalized\_data, method = "euclidean")

**Hierarchical Clustering.**

> fit <- hclust(d, method="complete")

> fit$order

[1] 41 48 34 45 19 15 29 12 26 27 17 4 46 50 25 37 47 8 39 21 30 7 23 49 36 14 16 35 38 11 44 6

[33] 5 28 9 43 13 32 3 22 20 31 2 1 18 10 42 33 24 40

> fit$height

[1] 0.2058539 0.3502188 0.4287712 0.4940832 0.5303259 0.5353893 0.5935343 0.6457158 0.7038309

[10] 0.7108812 0.7389936 0.7722224 0.7781298 0.7865674 0.7977642 0.8286936 0.8412900 0.8457697

[19] 0.9824857 0.9971035 1.0122252 1.0354597 1.0709720 1.0800988 1.0918624 1.1314351 1.1826891

[28] 1.1968261 1.2117356 1.2502752 1.2716808 1.3329504 1.3988595 1.4668378 1.6230495 1.6448574

[37] 1.6585736 1.8537984 1.8649801 2.2630242 2.2952287 2.3374653 2.4458600 2.4748807 3.0883430

[46] 3.2554326 4.4005416 4.4200736 6.0766416

> plot(fit, hang=-1)



> rect.hclust(fit, k=6, border="red")



**Cutting and Grouping into 6 clusters**

> groups <- cutree(fit, k=6)

> membership<-as.matrix(groups)

> final <- data.frame(crime\_data, membership)

Final Dataset:

> final

X1 Murder Assault UrbanPop Rape membership

1 Alabama 13.2 236 58 21.2 1

2 Alaska 10.0 263 48 44.5 2

3 Arizona 8.1 294 80 31.0 3

4 Arkansas 8.8 190 50 19.5 4

5 California 9.0 276 91 40.6 3

6 Colorado 7.9 204 78 38.7 3

7 Connecticut 3.3 110 77 11.1 5

8 Delaware 5.9 238 72 15.8 5

9 Florida 15.4 335 80 31.9 3

10 Georgia 17.4 211 60 25.8 1

11 Hawaii 5.3 46 83 20.2 5

12 Idaho 2.6 120 54 14.2 6

13 Illinois 10.4 249 83 24.0 3

14 Indiana 7.2 113 65 21.0 5

15 Iowa 2.2 56 57 11.3 6

16 Kansas 6.0 115 66 18.0 5

17 Kentucky 9.7 109 52 16.3 4

18 Louisiana 15.4 249 66 22.2 1

19 Maine 2.1 83 51 7.8 6

20 Maryland 11.3 300 67 27.8 3

21 Massachusetts 4.4 149 85 16.3 5

22 Michigan 12.1 255 74 35.1 3

23 Minnesota 2.7 72 66 14.9 5

24 Mississippi 16.1 259 44 17.1 1

25 Missouri 9.0 178 70 28.2 4

26 Montana 6.0 109 53 16.4 6

27 Nebraska 4.3 102 62 16.5 6

28 Nevada 12.2 252 81 46.0 3

29 New Hampshire 2.1 57 56 9.5 6

30 New Jersey 7.4 159 89 18.8 5

31 New Mexico 11.4 285 70 32.1 3

32 New York 11.1 254 86 26.1 3

33 North Carolina 13.0 337 45 16.1 1

34 North Dakota 0.8 45 44 7.3 6

35 Ohio 7.3 120 75 21.4 5

36 Oklahoma 6.6 151 68 20.0 5

37 Oregon 4.9 159 67 29.3 4

38 Pennsylvania 6.3 106 72 14.9 5

39 Rhode Island 3.4 174 87 8.3 5

40 South Carolina 14.4 279 48 22.5 1

41 South Dakota 3.8 86 45 12.8 6

42 Tennessee 13.2 188 59 26.9 1

43 Texas 12.7 201 80 25.5 3

44 Utah 3.2 120 80 22.9 5

45 Vermont 2.2 48 32 11.2 6

46 Virginia 8.5 156 63 20.7 4

47 Washington 4.0 145 73 26.2 4

48 West Virginia 5.7 81 39 9.3 6

49 Wisconsin 2.6 53 66 10.8 5

50 Wyoming 6.8 161 60 15.6 4

> for(i in 2:8) wss[i]=sum(kmeans(norm\_crime,centers=i)$withinss)

> plot(1:8,wss,type = "b",xlab = "clusters",ylab = "with in sum of squares",main = "K-means clustering")

>



> fviz\_nbclust(norm\_crime,method = 'wss',FUNcluster = kmeans)



> fviz\_nbclust(norm\_crime,method = 'silhouette',FUNcluster = kmeans)



> fviz\_nbclust(norm\_crime,method = 'gap\_stat',FUNcluster = kmeans)

Clustering k = 1,2,..., K.max (= 10): .. done

Bootstrapping, b = 1,2,..., B (= 100) [one "." per sample]:

.................................................. 50

.................................................. 100

>



> final <- kmeans(norm\_crime,2)



> finalanim <- kmeans.ani(norm\_crime,2)

> fviz\_cluster(final,data = crime[-1])



> View(finaldata)

## Group.1 Murder Assault UrbanPop Rape groups

## 1 1 14.087500 252.7500 53.50000 24.53750 1

## 2 2 11.054545 264.0909 79.09091 32.61818 2

## 3 3 5.871429 134.4762 70.76190 18.58095 3

## 4 4 3.180000 78.7000 49.30000 11.63000 4

As per summary we can say group 2 have the higher rate of crime.