Hierarchal Clustering documentation

1.Ans)

Business problem: Perform clustering (Both hierarchical and K means clustering) for the airlines data to obtain optimum number of clusters. Draw the inferences from the clusters obtained.

Hierarchal clustering:

No. of rows=3999

No. of columns=12

Null values=0

Normalizing continuous columns to bring them under same scale

distance matrix

Summary:

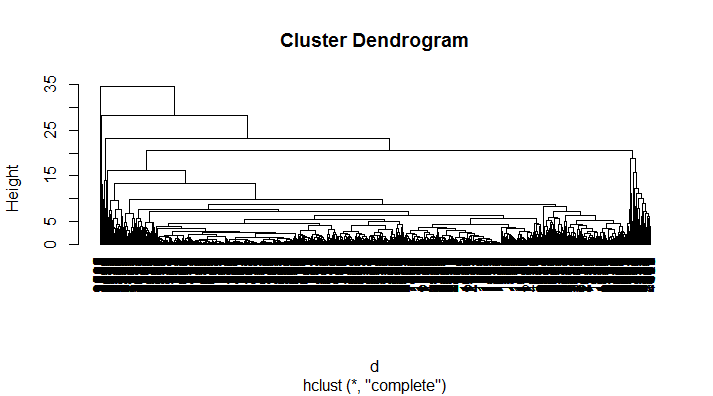
|  |
| --- |
| Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles  Min. : 0 Min. : 0.0 Min. :1.00 Min. :1.000 Min. :1.000  1st Qu.: 18528 1st Qu.: 0.0 1st Qu.:1.00 1st Qu.:1.000 1st Qu.:1.000  Median : 43097 Median : 0.0 Median :1.00 Median :1.000 Median :1.000  Mean : 73601 Mean : 144.1 Mean :2.06 Mean :1.015 Mean :1.012  3rd Qu.: 92404 3rd Qu.: 0.0 3rd Qu.:3.00 3rd Qu.:1.000 3rd Qu.:1.000  Max. :1704838 Max. :11148.0 Max. :5.00 Max. :3.000 Max. :5.000    Bonus\_miles Bonus\_trans Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll  Min. : 0 Min. : 0.0 Min. : 0.0 Min. : 0.000 Min. : 2  1st Qu.: 1250 1st Qu.: 3.0 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.:2330  Median : 7171 Median :12.0 Median : 0.0 Median : 0.000 Median :4096  Mean : 17145 Mean :11.6 Mean : 460.1 Mean : 1.374 Mean :4119  3rd Qu.: 23801 3rd Qu.:17.0 3rd Qu.: 311.0 3rd Qu.: 1.000 3rd Qu.:5790  Max. :263685 Max. :86.0 Max. :30817.0 Max. :53.000 Max. :8296    Award.  Min. :0.0000  1st Qu.:0.0000  Median :0.0000  Mean :0.3703  3rd Qu.:1.0000  Max. :1.0000 |
|  |
| |  | | --- | |  | |

By the above summary we will get to know that, inputs have different ranges, so its better to normalize, so that we will get the values of every input in the range 0-1

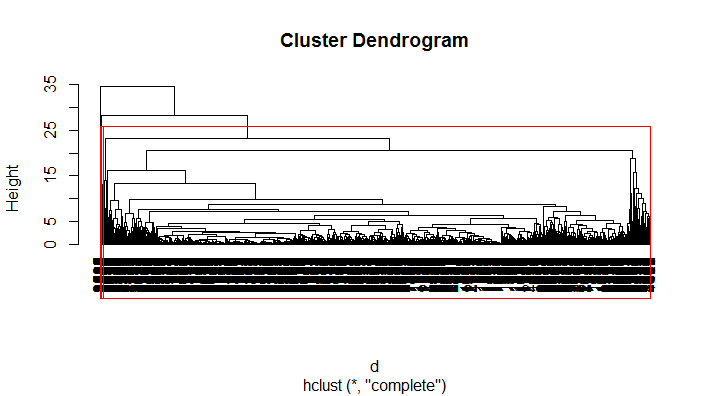
Clustering:

In hierarchal clustering technique, we will group the similar records, by calculating the Euclidian distance between each record to record.

Dendogram :



For k=3,



K-means-clustering:

No. of rows=3999

No. of columns=12

Null values=0

Summary:

|  |
| --- |
| Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles  Min. : 0 Min. : 0.0 Min. :1.00 Min. :1.000 Min. :1.000  1st Qu.: 18528 1st Qu.: 0.0 1st Qu.:1.00 1st Qu.:1.000 1st Qu.:1.000  Median : 43097 Median : 0.0 Median :1.00 Median :1.000 Median :1.000  Mean : 73601 Mean : 144.1 Mean :2.06 Mean :1.015 Mean :1.012  3rd Qu.: 92404 3rd Qu.: 0.0 3rd Qu.:3.00 3rd Qu.:1.000 3rd Qu.:1.000  Max. :1704838 Max. :11148.0 Max. :5.00 Max. :3.000 Max. :5.000    Bonus\_miles Bonus\_trans Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll  Min. : 0 Min. : 0.0 Min. : 0.0 Min. : 0.000 Min. : 2  1st Qu.: 1250 1st Qu.: 3.0 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.:2330  Median : 7171 Median :12.0 Median : 0.0 Median : 0.000 Median :4096  Mean : 17145 Mean :11.6 Mean : 460.1 Mean : 1.374 Mean :4119  3rd Qu.: 23801 3rd Qu.:17.0 3rd Qu.: 311.0 3rd Qu.: 1.000 3rd Qu.:5790  Max. :263685 Max. :86.0 Max. :30817.0 Max. :53.000 Max. :8296    Award.  Min. :0.0000  1st Qu.:0.0000  Median :0.0000  Mean :0.3703  3rd Qu.:1.0000  Max. :1.0000  By the above summary we will get to know that, inputs have different ranges, so need to do normalization, so that we will get the values of every input in the range 0-1  K-means clus  tering:  This technique we will group the similar records, by calculating the Euclidian distance from centre to record.  K value will be selected by the totalwithinness and betweenness  Withinness is within cluster difference, nothing but distance between centre to record in a cluster.  Betweenness is nothing but cluster to cluster difference  Total withinness should be less and betweenness should be high  At k=4,  tot.withinss: num 28902  betweenss : num 15076    We will select the no. of clusters by using Elbow curve  For centres=4, clusters are formed here.  The elbow curve is given below.  C:\Users\sravyarachana\Desktop\ASSIGNMENTS\clustering\Rplot03.png |

2.Ans)

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

No. of rows=50

No. of columns=5

NA values=0

Summary:

Murder Assault UrbanPop Rape

Min. : 0.800 Min. : 45.0 Min. :32.00 Min. : 7.30

1st Qu.: 4.075 1st Qu.:109.0 1st Qu.:54.50 1st Qu.:15.07

Median : 7.250 Median :159.0 Median :66.00 Median :20.10

Mean : 7.788 Mean :170.8 Mean :65.54 Mean :21.23

3rd Qu.:11.250 3rd Qu.:249.0 3rd Qu.:77.75 3rd Qu.:26.18

Max. :17.400 Max. :337.0 Max. :91.00 Max. :46.00

By the above summary we will get to know that, inputs have different ranges, so its better to normalize, so that we will get the values of every input in the range 0-1

Clustering:

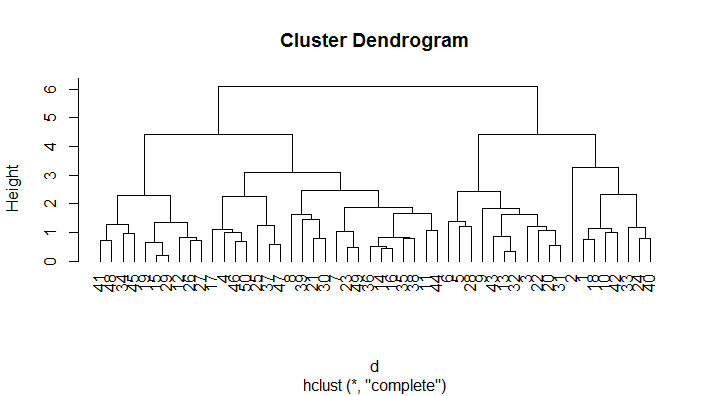
In hierarchal clustering technique, we will group the similar records, by calculating the Euclidian distance between each record to record.

Dendogram :

Clustering:

In hierarchal clustering technique, we will group the similar records, by calculating the Euclidian distance between each record to record.

Dendogram :



For k=3,the 3 clusters are formed

