**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.

**Data:**

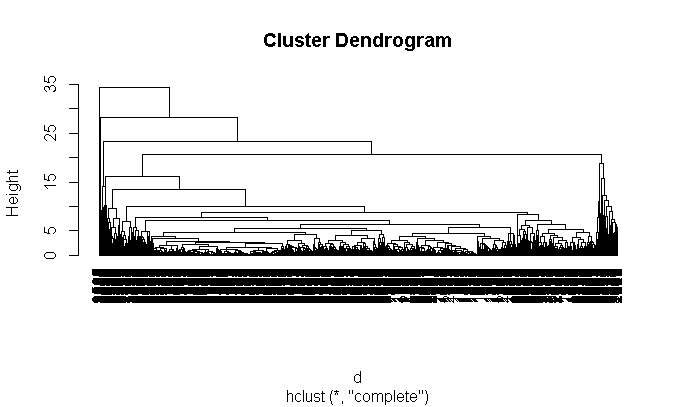
Data in the form of mixed data. It contains the numerical data and dummied data (CC1\_Miles, CC2\_Miles, CC3\_Miles and Award).

**Pre-processing Data:**

All the features are not in same scale. So first of all convert them all into single scale. There was no outlier and NA in the data. Delete the unused feature like ID# from the processing.

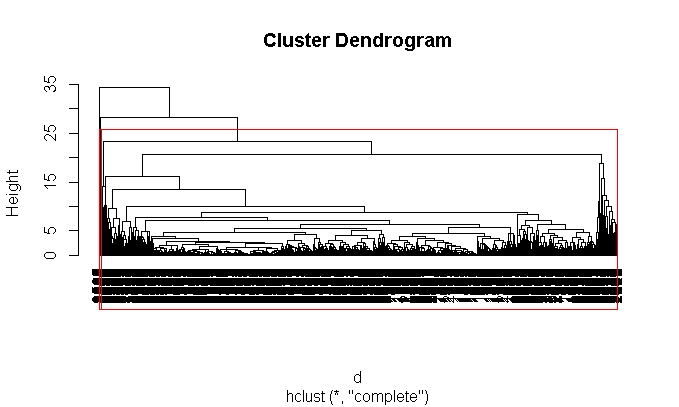
**Building the Model:**

Build the model using Euclidean distance and complete linkage functions. Please find the Dendrogram.



I’m going with Hierarchical Clustering by taking k as 3 and proceeding further. Data are classified in to following Clusters.

|  |
| --- |
| Cluster 1 Cluster 2 Cluster 3 |
| 3980 15 4 |



Look for the medians of each group of the feature.

Cluster Freq Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans

1 3980 42980 0 1.0 1 1 7086 12.0

2 15 108081 0 4.0 1 4 95598 30.0

3 4 106673 250 2.5 1 1 76325 66.5

Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll

1 0 0.0 4099.5

2 0 0.0 4103.0

3 18075 49.5 1822.5

By looking the above data we can categorize the passengers into 3 types.

**Non Frequent travellers:**

Cluster 1 metrics depicts that the flight transactions in last 12 months are zero and enrolling days are very high.

**High frequent travellers:**

Cluster 3 metrics depicts that the flight miles and flight transactions in last 12 months travelled and their bonus miles, bonus transactions, Balance also high and flight enrolling time is less. So These people are travelling very frequently.

**Middle class travellers:**

Cluster 2 metrics depicts that the level of spending is average to cluster 1 and 3.

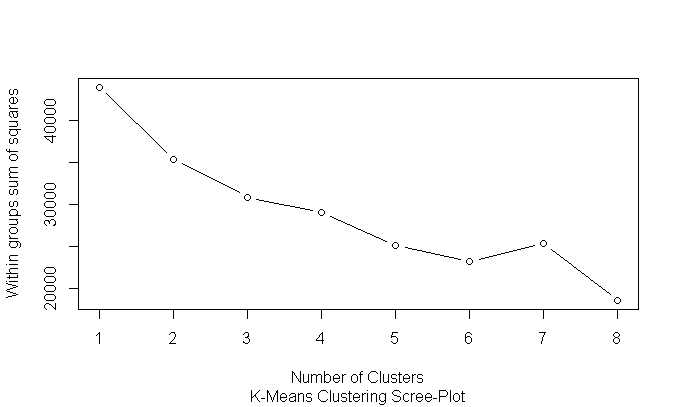
Now treat them as Non Hierarchical cluster. Then we will find out the K by using different techniques.

**By Using K-Selection:**

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

It is giving the value as 2.

**Elbow Curve:**



Elbow Curve value is subjective. As per my analysis 3 is the value for the K.

**By using KMeans:**

Calculate the tot.withinss and betweenss. For a good model tot.withinss should be more and betweenss should be less. If there should be less variation in the difference then we can treat it as final K.

Trails :

K=2 : km <- kmeans(normalized\_data,2)

$ tot.withinss: num 35401

$ betweenss : num 8577

K=3 : km <- kmeans(normalized\_data,3)

$ tot.withinss: num 31001

$ betweenss : num 12977

K=4 : km <- kmeans(normalized\_data,4)

$ tot.withinss: num 29189

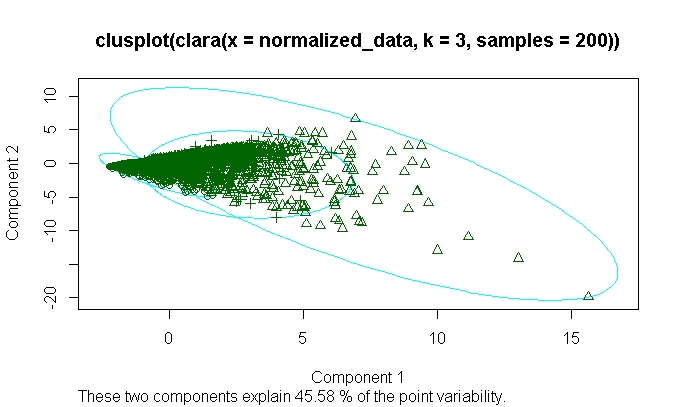
$ betweenss : num 14789

K=5 : km <- kmeans(normalized\_data,5)

$ tot.withinss: num 25173

$ betweenss : num 18805

By looking the data we can consider k=3 for further process. And also check the clustplot.



**Business Problem:**

Perform clustering for the crime data and identify the number of clusters formed and

**Data:**

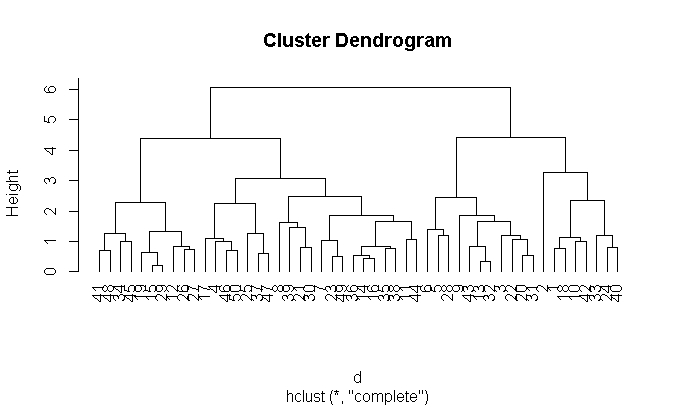
Data in the form of numeric data. It contains the numerical data (Murder, Assault, UrbanPop and Rape)

**Pre-processing Data:**

All the features are not in same scale. So first of all convert them all into single scale. There was no outlier and NA in the data. Delete the unused feature like ‘X’ from the processing.

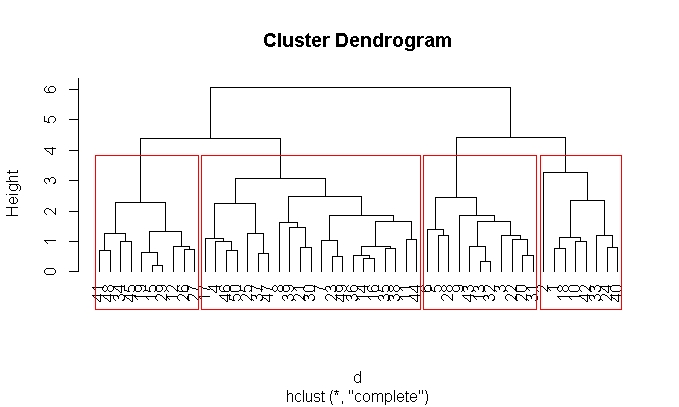
**Building the Model:**

Build the model using Euclidean distance and complete linkage functions. Please find the Dendrogram.



I’m going with Hierarchical Clustering by taking k as 4 and proceeding further. Data are classified in to following Clusters.

|  |
| --- |
| Cluster 1 Cluster 2 Cluster 3 Cluster 4 |
| 8 11 21 10 |



Look for the medians of each group of the feature.

Cluster Freq Murder Assault UrbanPop Rape

1 8 13.8 254 53 22.35

2 11 11.3 255 80 31.90

3 21 6.0 145 70 18.80

4 10 2.4 82 52 11.25

By looking the above data we can categorize the passengers into 4 types.

**Safe Areas**

Cluster 4 metrics depicts that the Areas is little safe, there crime rate is little less compare to all other areas.

**Not Safe Areas:**

Cluster 3 metrics depicts that that Areas is not safe, there Crime rate is not less compare to all other areas.

**Danger Areas:**

Cluster 2 metrics depicts that that Areas is Danger area, there Crime rate is high compare to all other areas. There is no women safety.

**Very Danger Areas:**

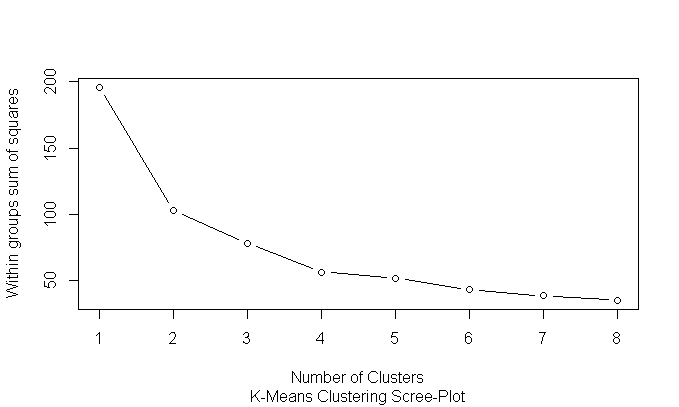
Cluster 1 metrics depicts that that Areas is Danger area, there Crime rate is too high compare to all other areas. There population is less and there is no women safety.Now treat them as Non Hierarchical cluster. Then we will find out the K by using different techniques.

**By Using K-Selection:**

k <- kselection(mydata[,-1], parallel = TRUE, k\_threshold = 0.9, max\_centers=5)

It is giving the value as 2.

**Elbow Curve:**



Elbow Curve value is subjective. As per my analysis 4 is the value for the K.

**By using KMeans:**

Calculate the tot.withinss and betweenss. For a good model tot.withinss should be more and betweenss should be less. If there should be less variation in the difference then we can treat it as final K.

Trails :

K=2 : km <- kmeans(normalized\_data,2)

$ tot.withinss: num 103

$ betweenss : num 93.1

K=3 : km <- kmeans(normalized\_data,3)

$ tot.withinss: num 85.1

$ betweenss : num 111.1

K=4 : km <- kmeans(normalized\_data,4)

$ tot.withinss: num 56.4

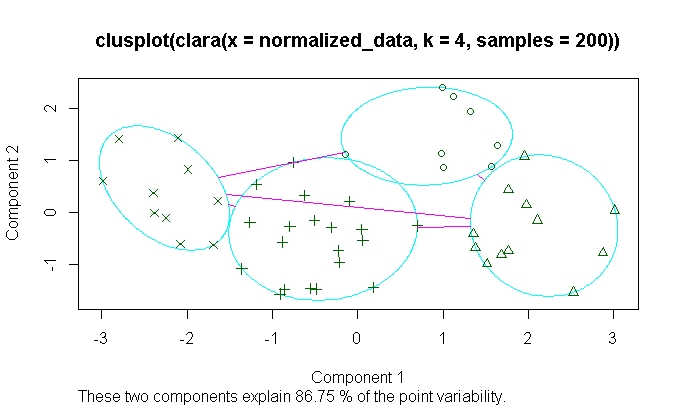
$ betweenss : num 140

K=5 : km <- kmeans(normalized\_data,5)

$ tot.withinss: num 52.9

$ betweenss : num 143

By looking the data We can consider k=4 for further process. And also check the clustplot.

****