Cluster Analysis Document

Vivek I(IMT2013017)
Aravind K(IMT2013019)
Nikhilesh P(IMT2013031)
Arun Prasad(IMT2013037)

As part of clustering for the data we have chosen for the project, we have performed the following activites:

(1)K-means clustering

- Decided the number of clusters(k) to be chosen using elbow graph
- Plot of different clusterings.
- Silhoutte plot to evaluate the clusterings we got.

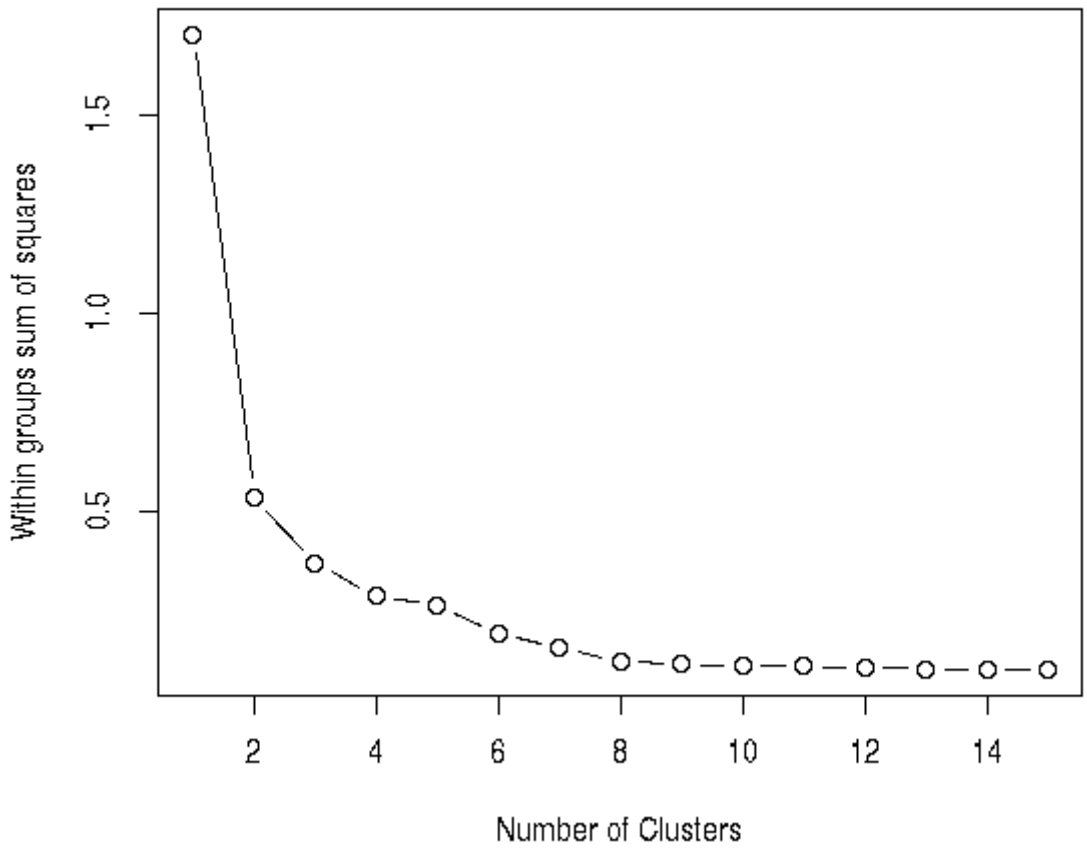
(2)Hirearcheal clustering

- Plot of different clusterings
- Silhoutte plot to evaluate the clusterings we got.

Befor performing he clustering algorithms, we have normalized our data and also converted categorical to numerical attributes to avoid the effect of outliers and to bring all the attribute's values to same scale(or range i.e b/w 0 and 1).

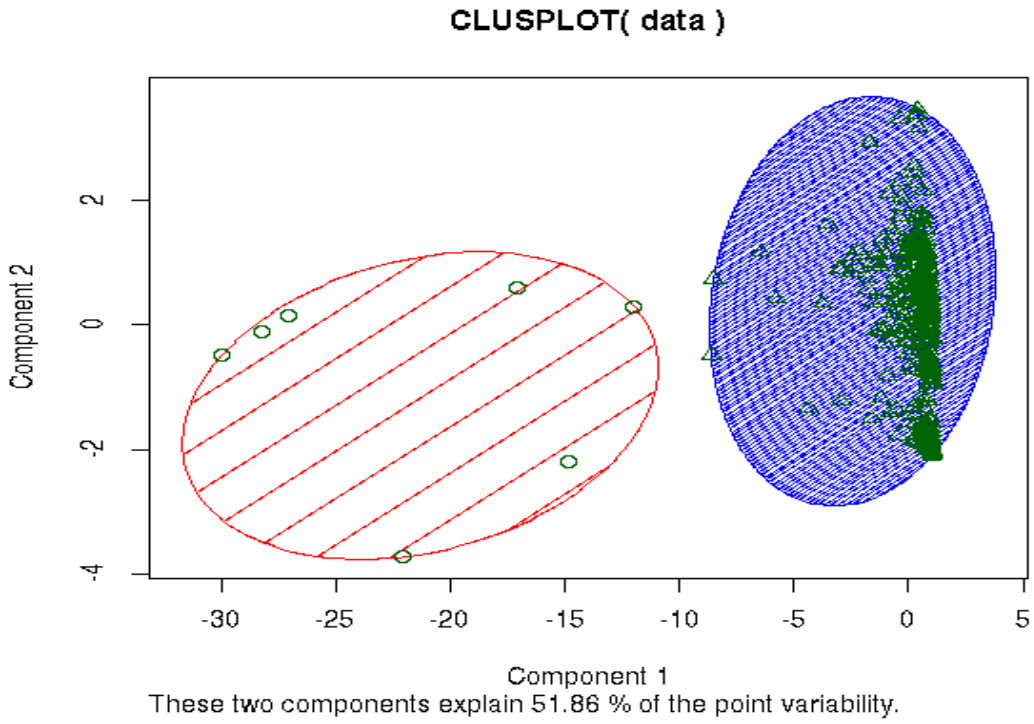
K-means clustering :

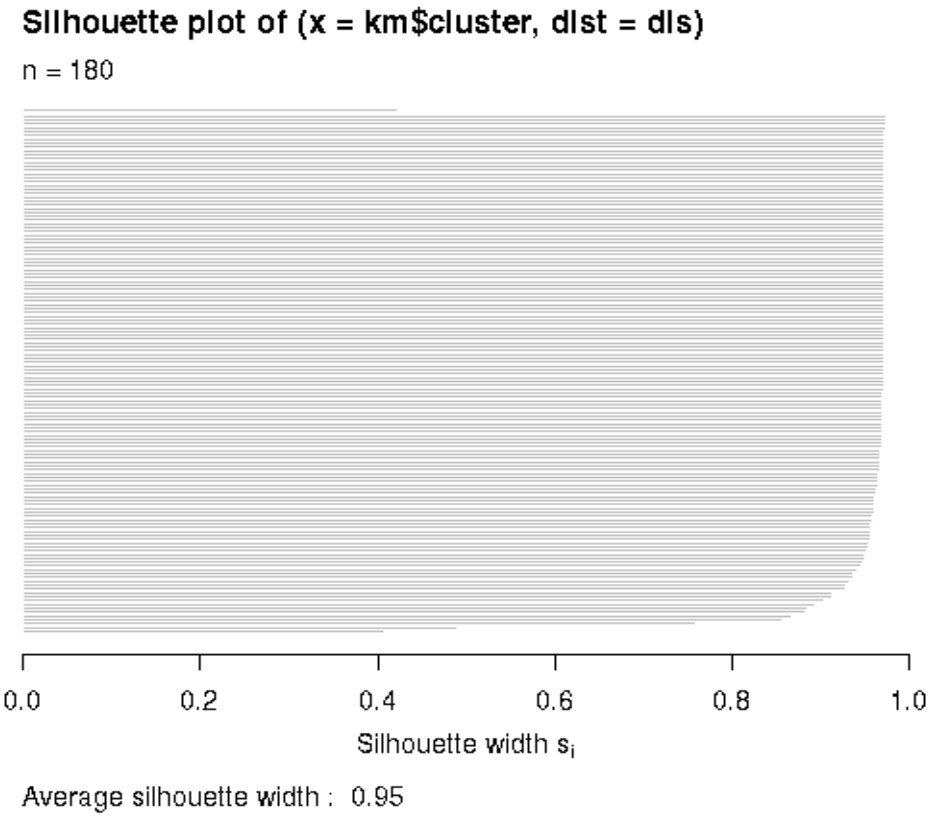
Since for kmeans function, number of clusters(k) needs to be given as input, we have first plotted the elbow graph for deciding this.



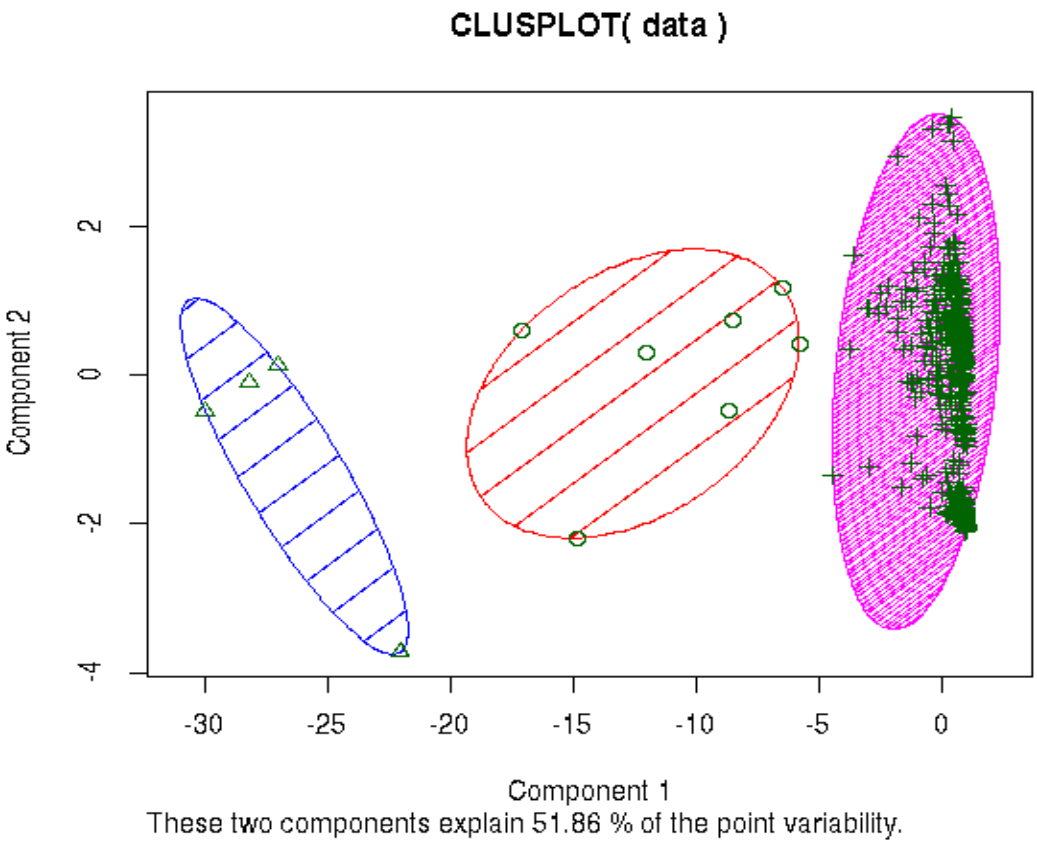
By seeing this graph, we have applied kmeans algorithm for k=2,3,4. The results of these are as follows:

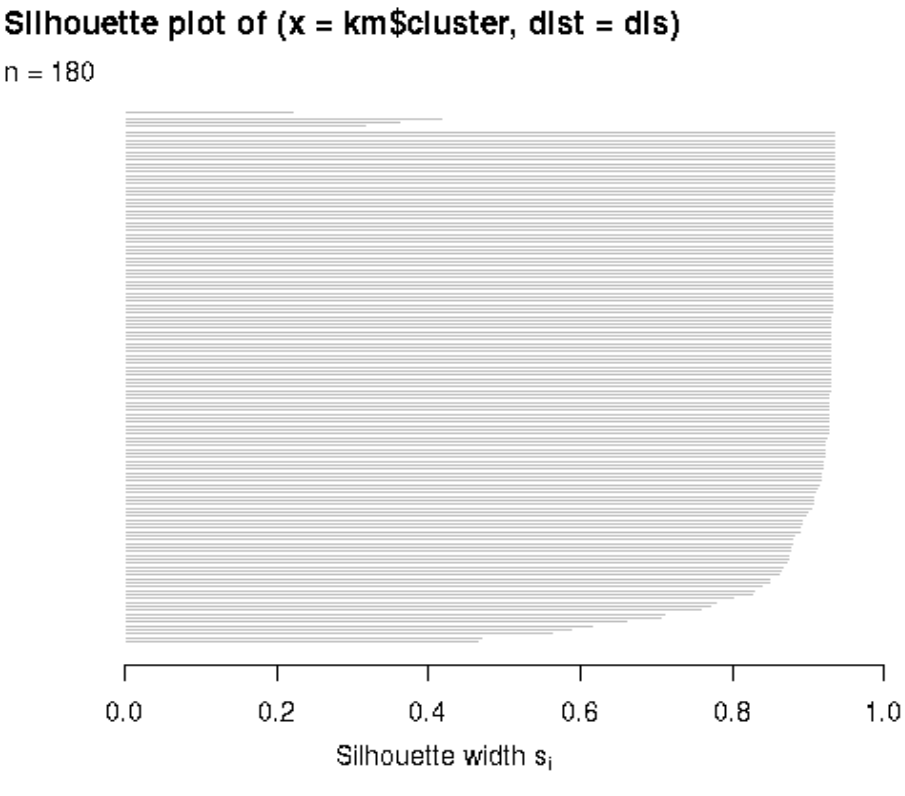
K=2 :



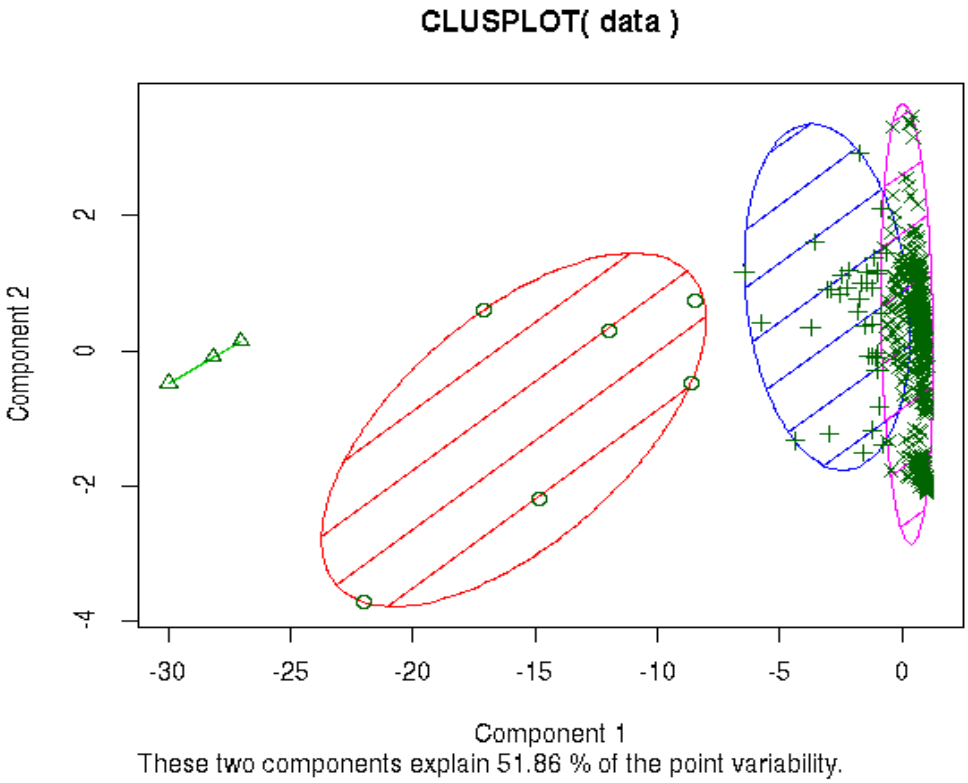


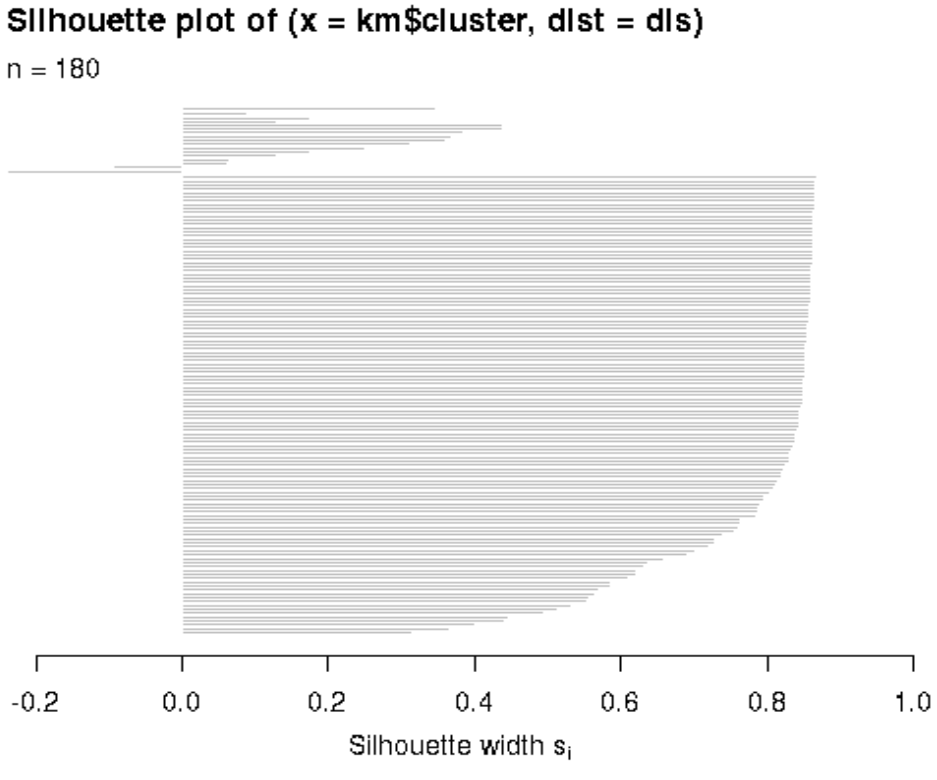
K=3 :


CLUSPLOT(data)
Component 2
2
0
-2
-4
-30
-25
-20
-15
-10
-5
0
Component 1
These two components explain 51.86 % of the point variability.


Silhouette plot of (x = km$cluster, dist = dis)
n = 180
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.87

K=4 :

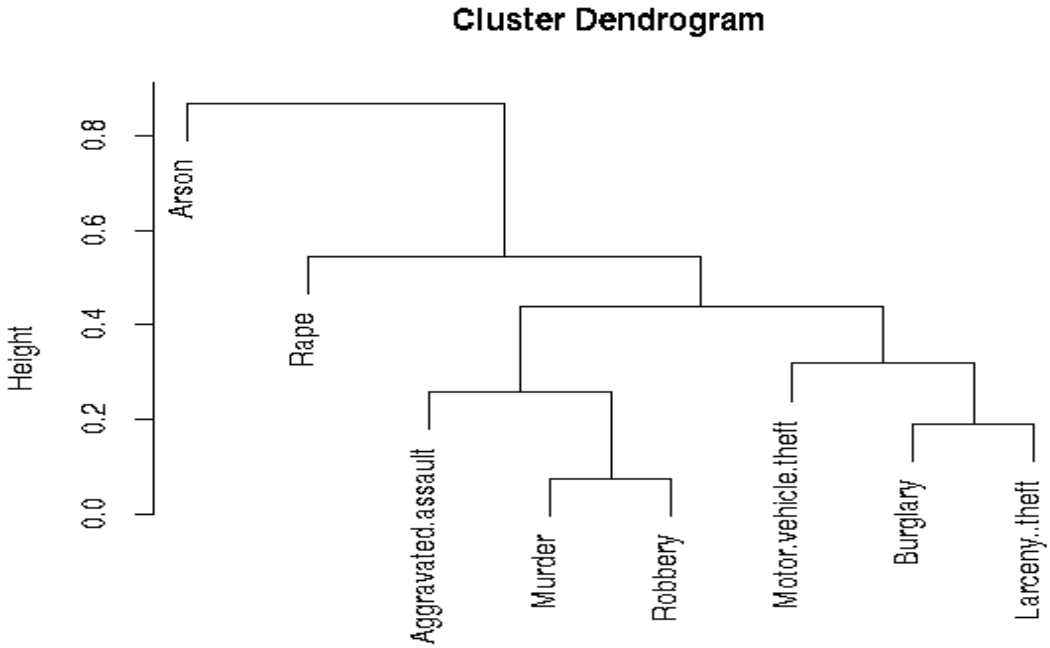

CLUSPLOT(data)
Component 2
2
0
-2
-4
-30
-25
-20
-15
-10
-5
0
Component 1
These two components explain 51.86 % of the point variability.


Silhouette plot of (x = km$cluster, dist = dis)
n = 180
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.71

So, from above silhoutte graphs, it's clear that the average silhotte width is more for k=2.
Also, it's clear from the graph that, more points fall in one cluster and very few fall in te remaining clusters, which implies that all the daa points are close to each other except few data points (or) in oter words, the number of crimes(in all categories) in all the cities almost remain the same except few cities which have more (or) less number of crimes and these fall into othe clusters.

<u>Hirarcheal clustering</u> :

Below shown is the output of hirearcheal clustering which is a cluster dendrogram:



This implies data in the features which are in same subtree of any node are closer(or has similar spread among them) compared to data in the features which are in the other subtree.

Below is the silhoutte plot for the above clustering: