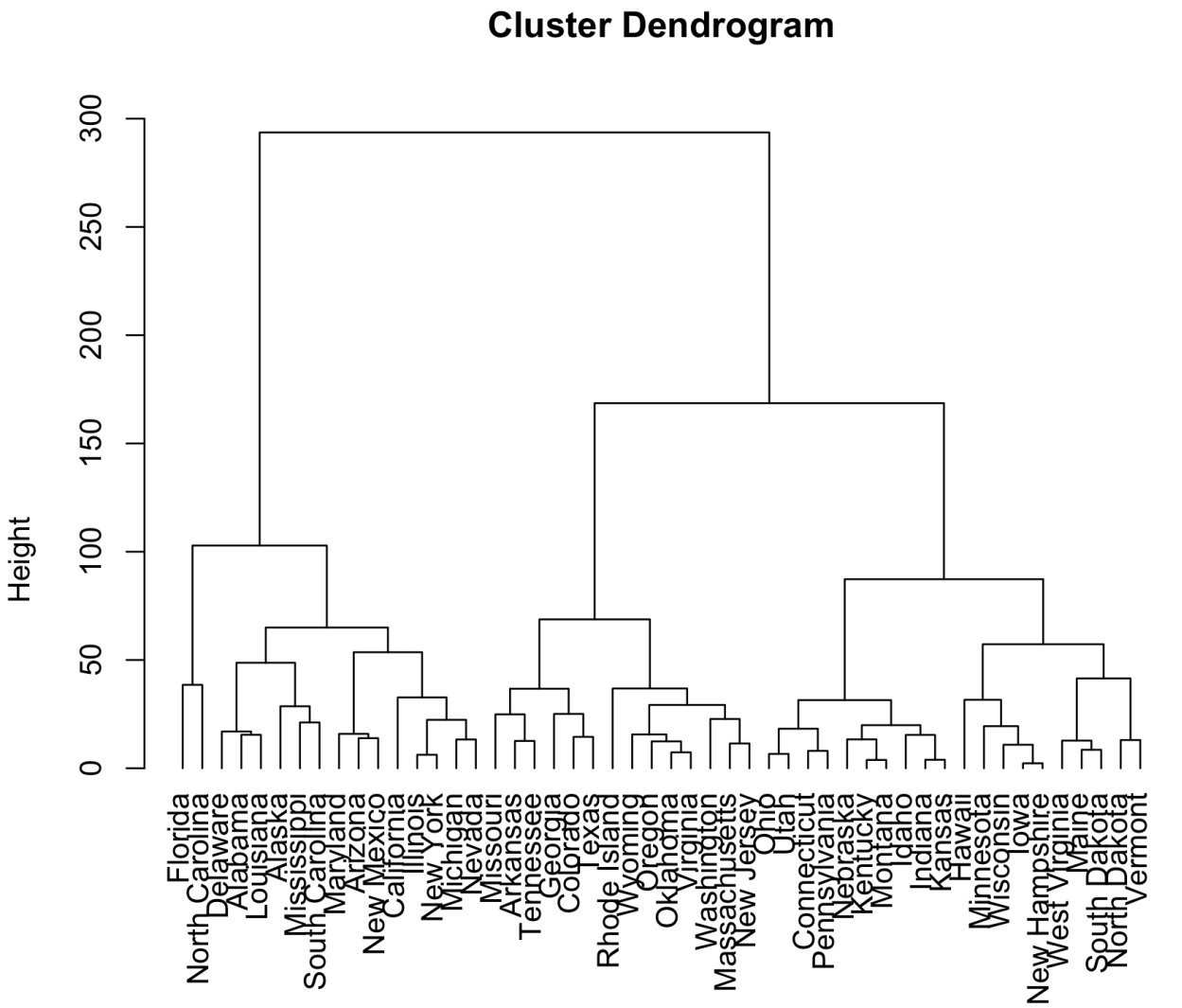
EAS 507

HW -2

By - Mohit Tripathi

Ans – 1

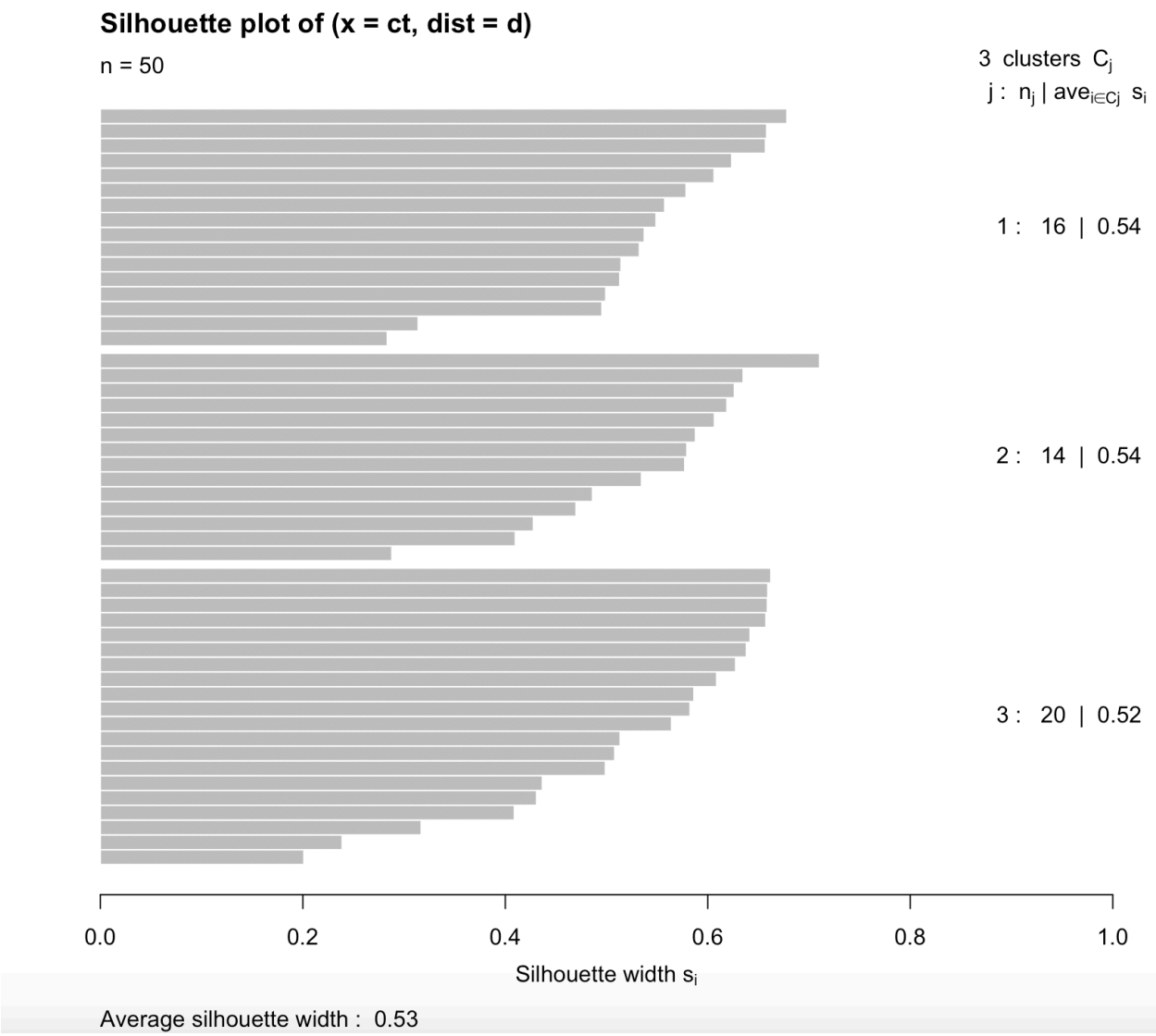
a) Cluster the states using Hierarchal linkage and Euclidian distance



b) Cut tree - 3 clusters

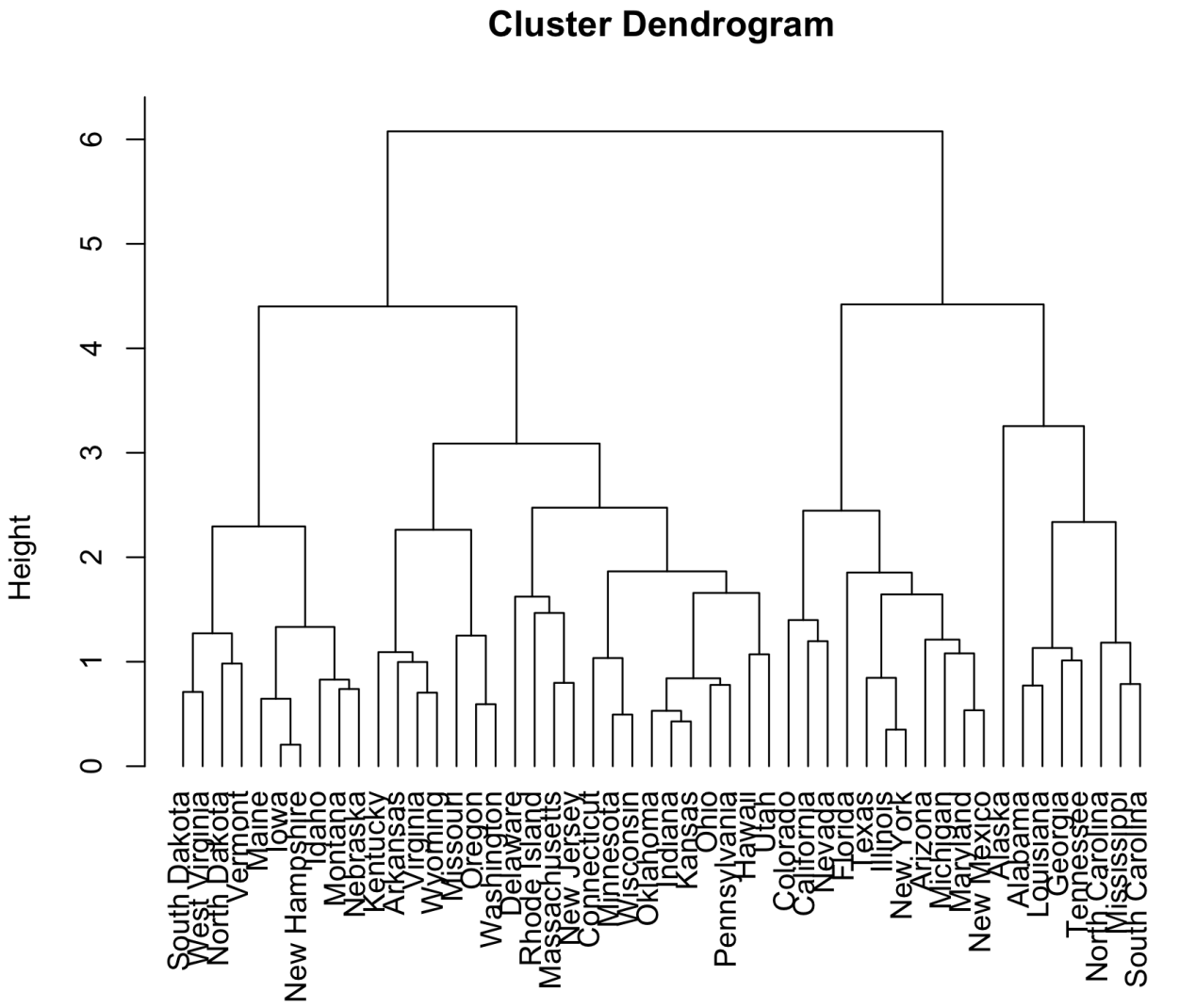
States and Clusters

Alabama	Alaska	Arizona	Arkansas
1	1	1	2
California	Colorado	Connecticut	Delaware
1	2	3	1
Florida	Georgia	Hawaii	Idaho
1	2	3	3
Illinois	Indiana	Iowa	Kansas
1	3	3	3
Kentucky	Louisiana	Maine	Maryland
3	1	3	1
Massachusetts	Michigan	Minnesota	Mississippi
2	1	3	1
Missouri	Montana	Nebraska	Nevada
2	3	3	1
New Hampshire	New Jersey	New Mexico	New York
3	2	1	1
North Carolina	North Dakota	Ohio	Oklahoma
1	3	3	2
Oregon	Pennsylvania	Rhode Island	South Carolina
2	3	2	1
South Dakota	Tennessee	Texas	Utah
3	2	2	3
Vermont	Virginia	Washington	West Virginia
3	2	2	3
Wisconsin	Wyoming		
3	2		



Average silhouette width is 0.53

c.) Scale the data and perform hierarchal clustering



d.)

High frequency crimes will have larger effect on inter-crime dissimilarities and hence on clustering obtained in comparison to rare crimes. This is undesirable

if inter-observation dissimilarities are computed after scaling the variables to have standard deviation as 1 then each variable will have equal importance in hierarchical clustering.

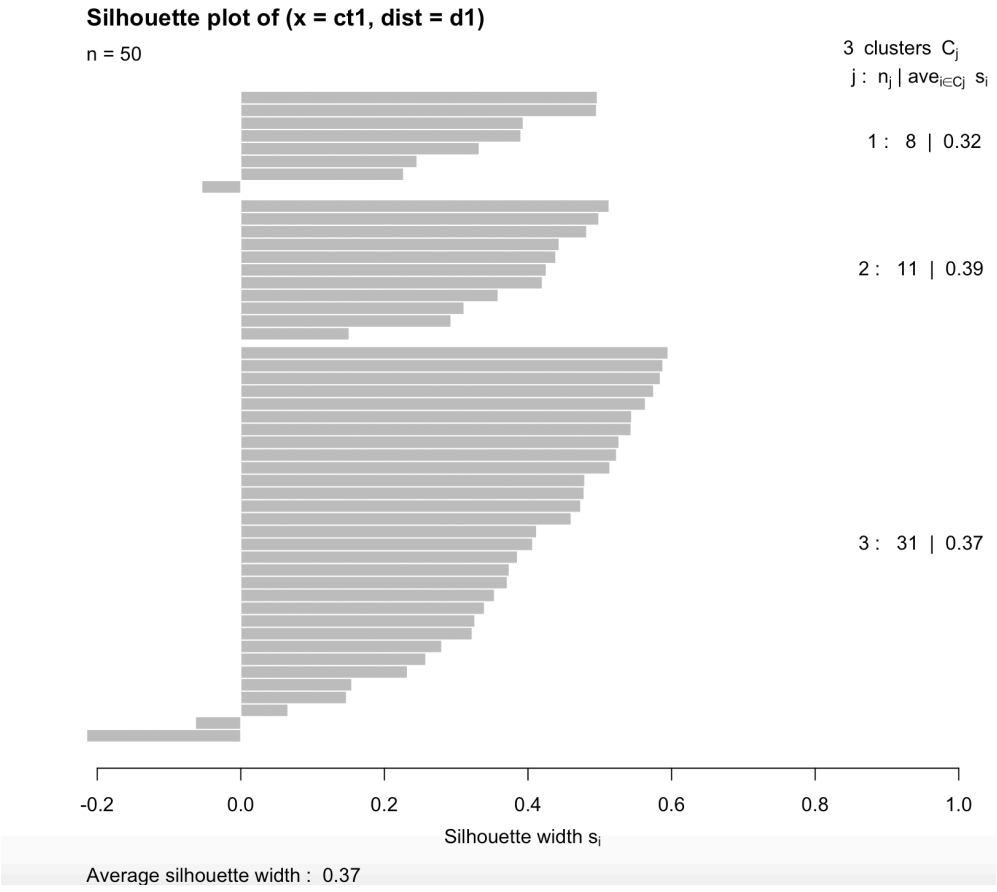
Scaling is helpful when the variables in the data set have different set of measurements like (grams vs kilograms)

Hence the variables should be scaled before the inter- observation dissimilarities are computed so as to give equal weight to all kind of crimes(variables)

If we don't scale the variables, then due to large variance and mean, Assault variable will have a huge impact on the clustering.

Cutree with scaled variables and 3 groups

Alabama	Alaska	Arizona	Arkansas	California
1	1	2	3	2
Colorado	Connecticut	Delaware	Florida	Georgia
2	3	3	2	1
Hawaii	Idaho	Illinois	Indiana	Iowa
3	3	2	3	3
Kansas	Kentucky	Louisiana	Maine	Maryland
3	3	1	3	2
Massachusetts	Michigan	Minnesota	Mississippi	Missouri
3	2	3	1	3
Montana	Nebraska	Nevada	New Hampshire	New Jersey
3	3	2	3	3
New Mexico	New York	North Carolina	North Dakota	Ohio
2	2	1	3	3
Oklahoma	Oregon	Pennsylvania	Rhode Island	South Carolina
3	3	3	3	1
South Dakota	Tennessee	Texas	Utah	Vermont
3	1	2	3	3
Virginia	Washington	West Virginia	Wisconsin	Wyoming
3	3	3	3	3

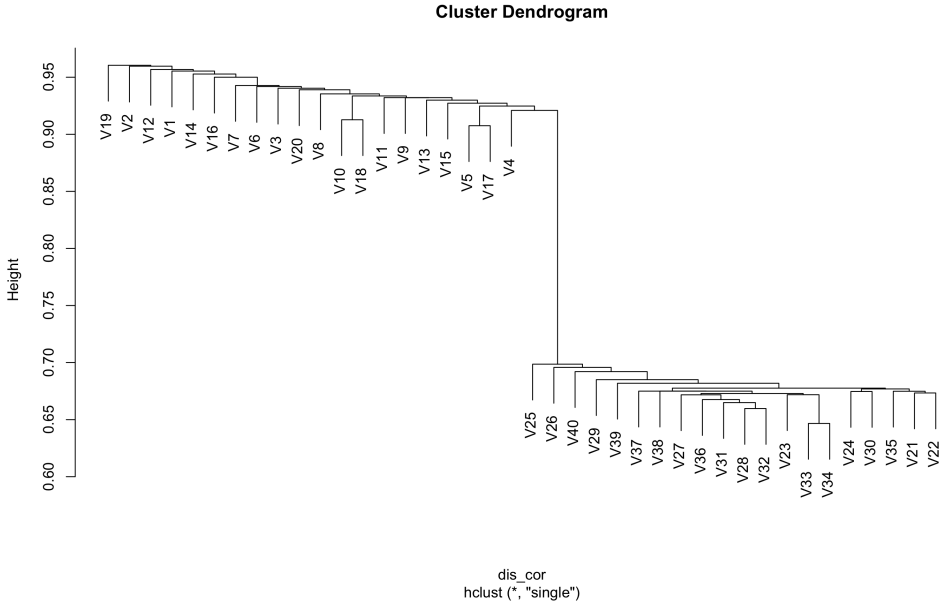


Ans -2

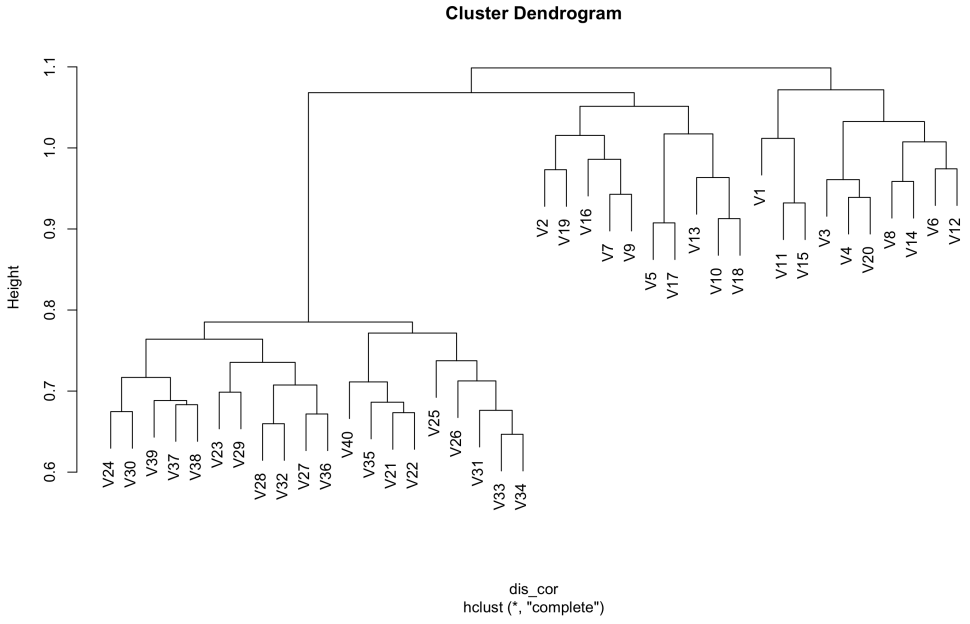
a.) Load the data

b.) Data seems to be already scaled

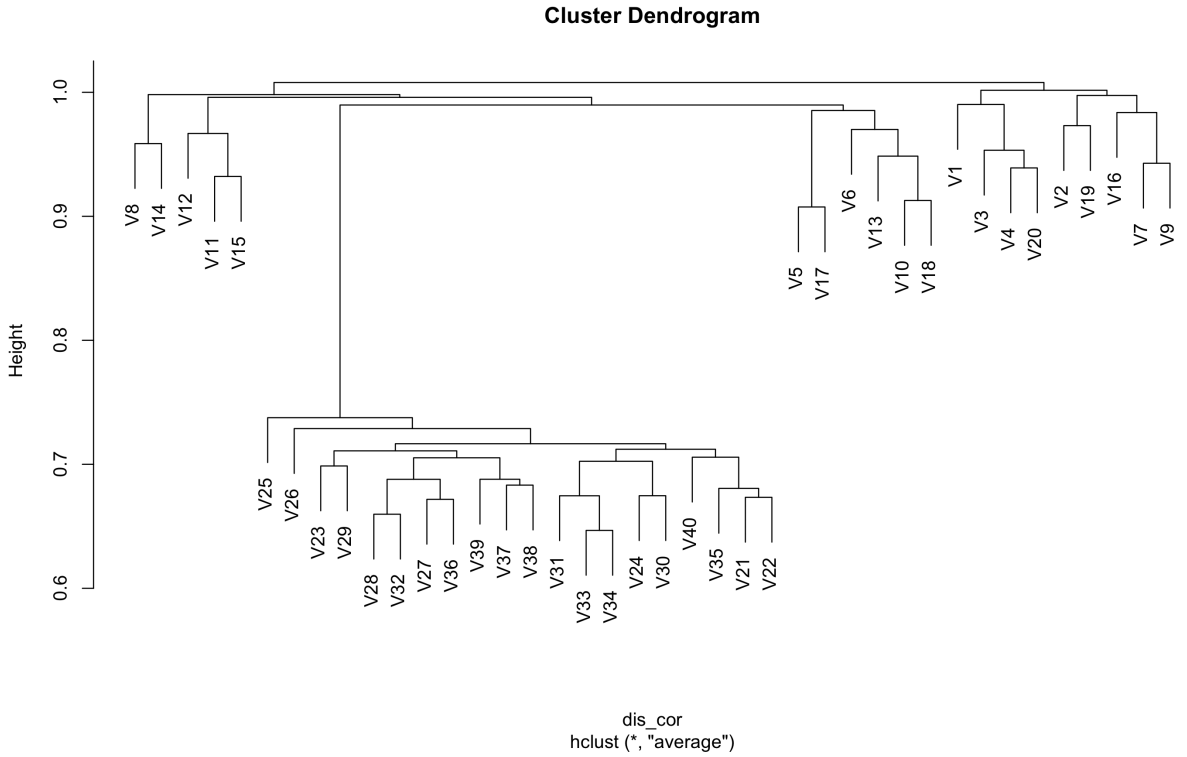
Hierarchical clustering with single linkage (gives 2 clusters)



Hierarchical clustering with complete linkage (gives 2 clusters)



Hierarchical clustering with average linkage (gives 3 clusters)

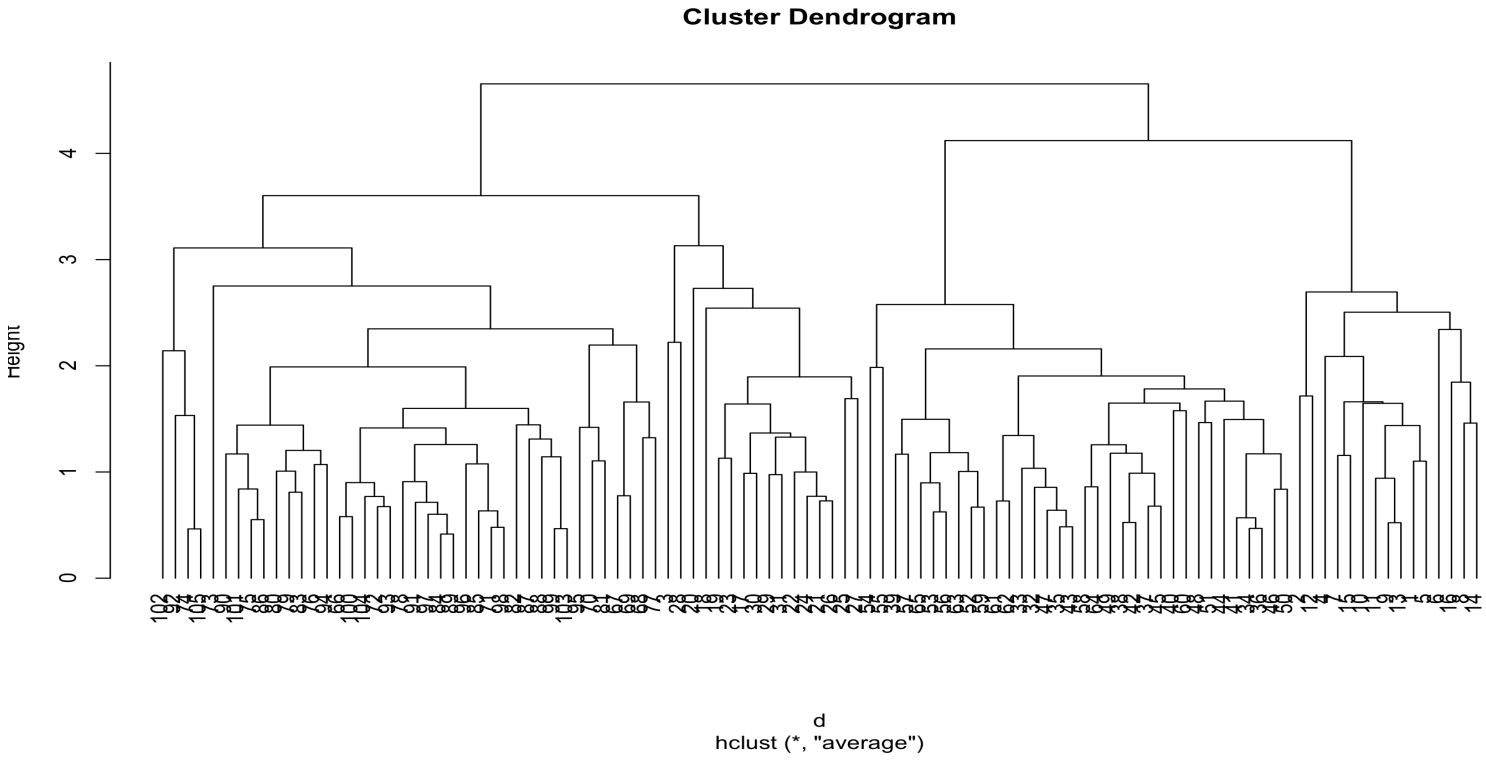


We can see that for complete and single linkages genes separate the samples into two clusters whereas for average linkage, it samples into three clusters.
Hence, we can say that the results obtained depends on the type/method of linkage used.

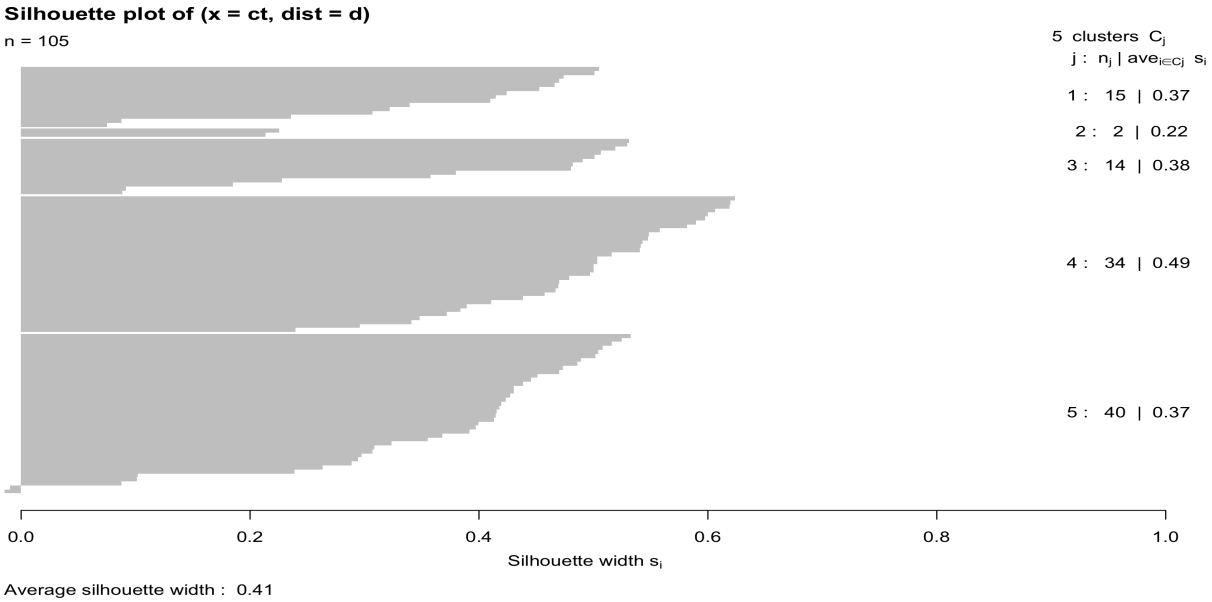
C.) Performed principal component analysis to answer this. Took sum of the absolute value of each principal component of individual observation and have shown top 5 observations with highest principal components value.

12 914 582 549 291 11

Ans-3
a)
Average Linkage

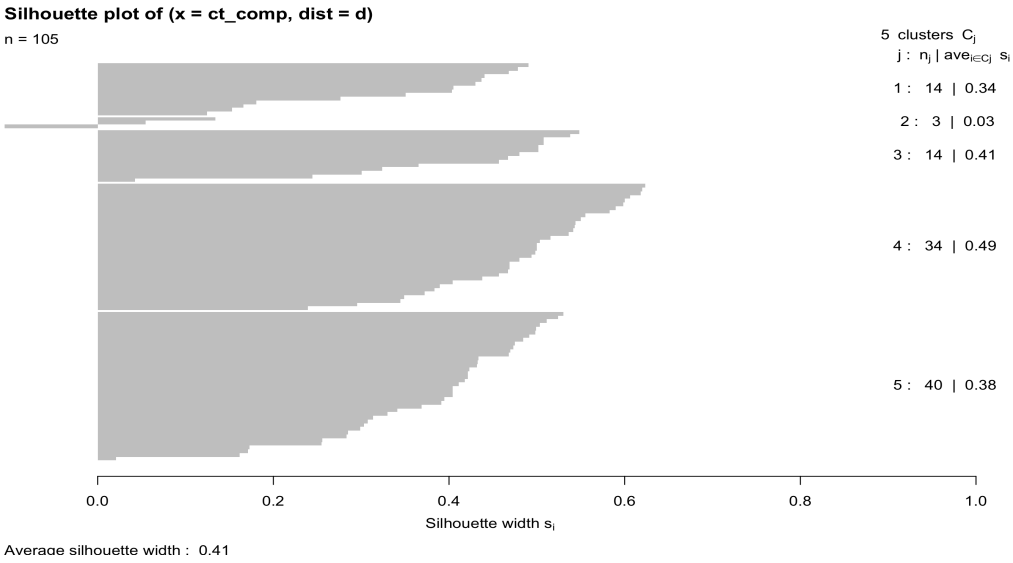


Divided the data into 5 groups/clusters as there are 5 classes in the response variable of the dataset



Average width of silhouette plot is .41 which is maximum at k = 5
Misclassification rate with average linkage is 34%

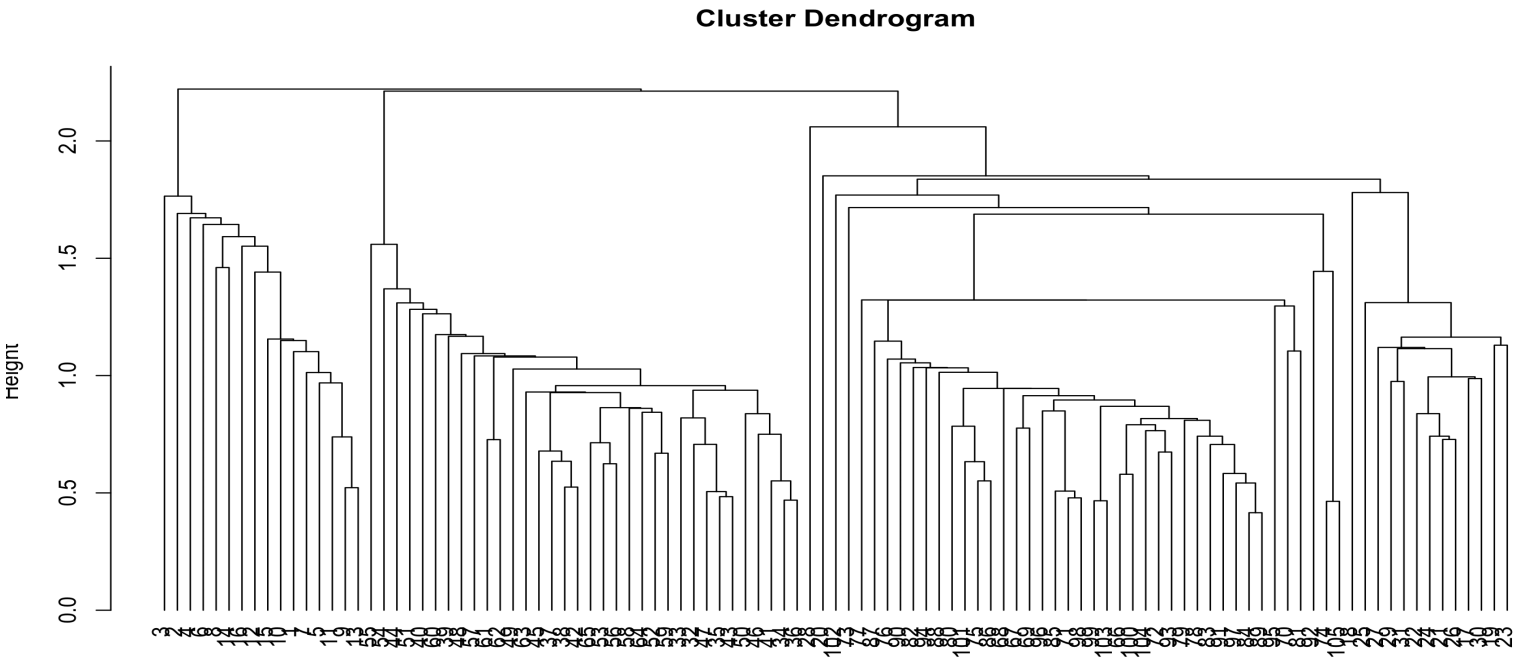
Complete Linkage Divided the data into 5 groups/clusters as there are 5 classes in the response variable of the dataset

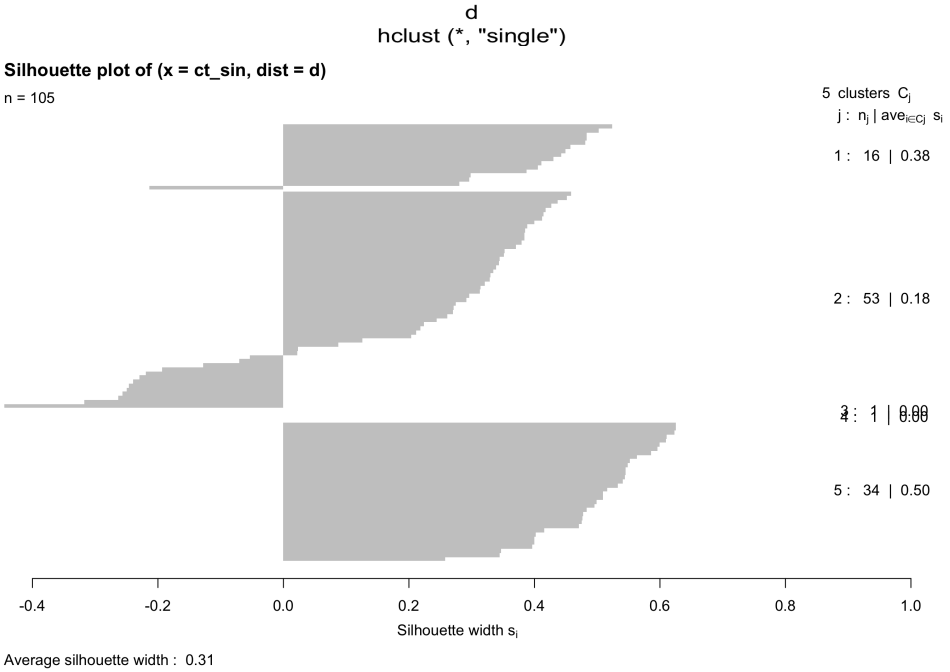


Average width of silhouette plot is .41 which is maximum at k = 5
Misclassification rate with complete linkage is 34.29%

Single Linkage

Divided the data into 5 groups/clusters as there are 5 classes in the response variable of the dataset



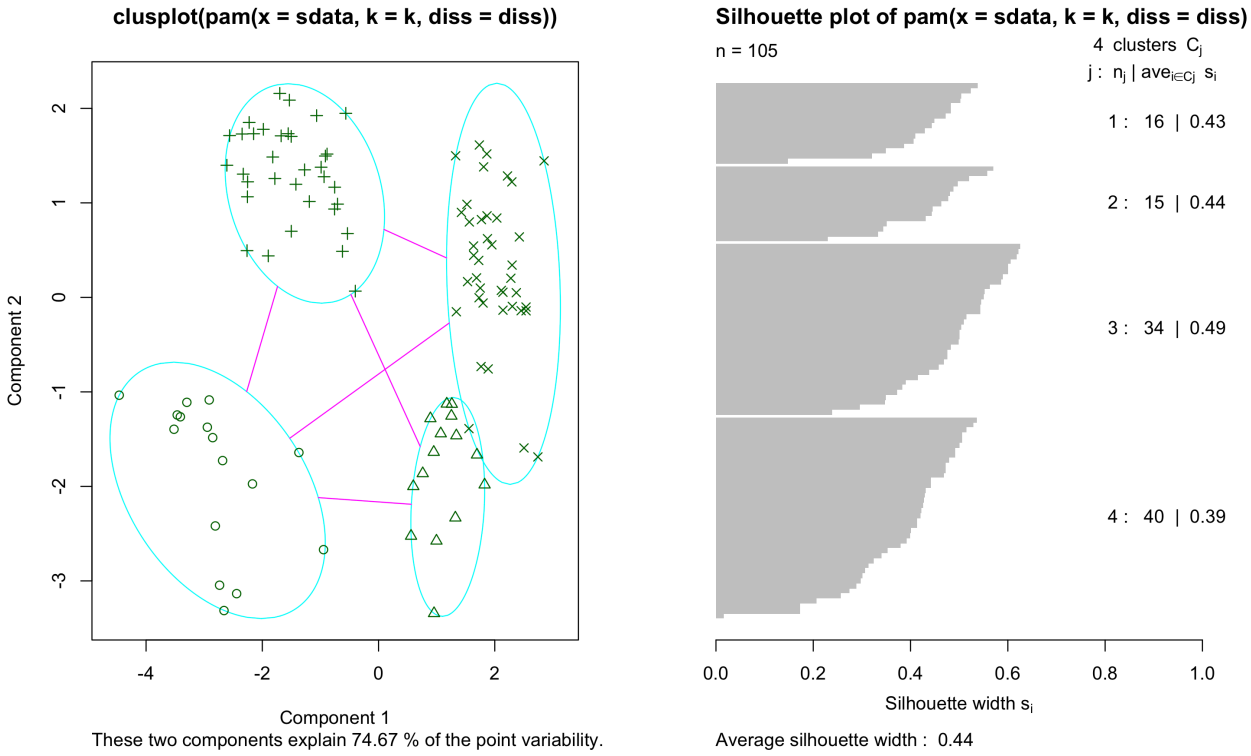


Average width of silhouette plot is .31 which is maximum at k = 5

Misclassification rate with single linkage is 72.38%

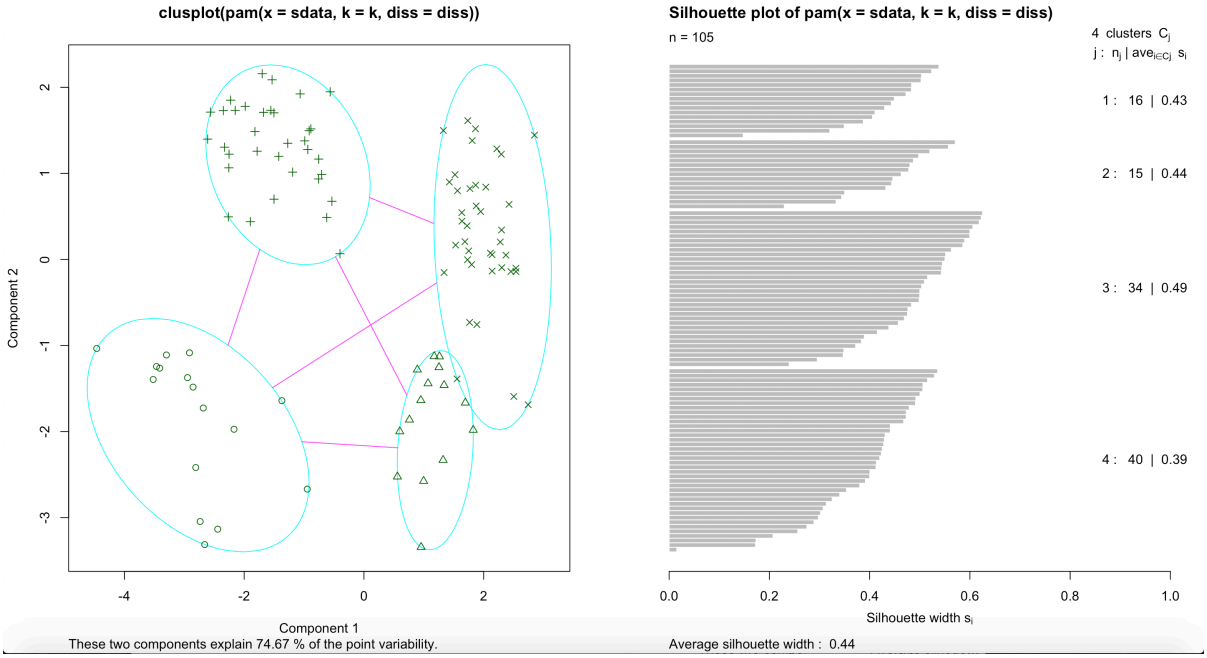
Complete Linkage method performed the best and single linkage method performed the worst.

b) On analytically finding the value of k, it is 4.



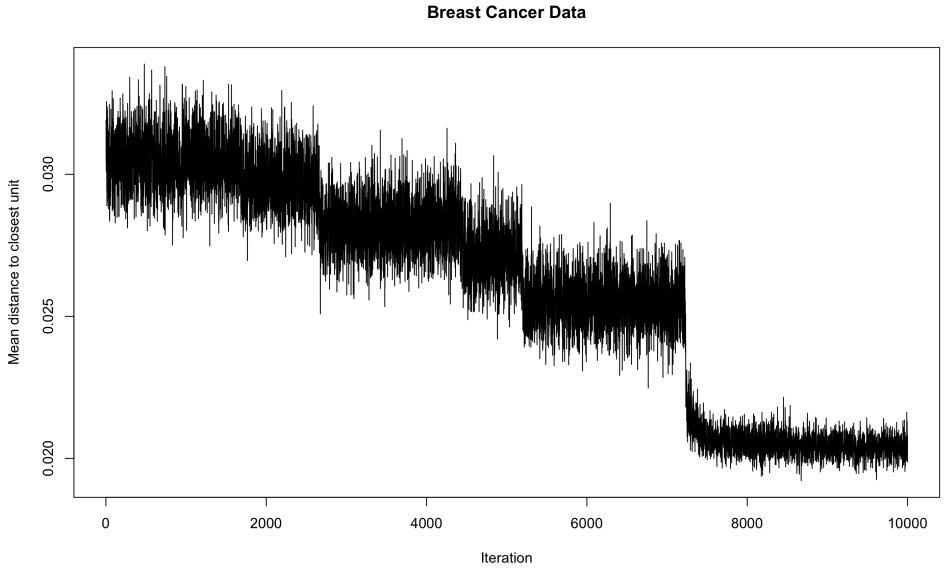
Average width of silhouette plot is .44 which is better than the different methods of hierarchal clustering.

To find the misclassification rate, I have applied k mediods on scaled data and passed k = 5 as argument (since there are 5 classes in the response variable of data set). The misclassification rate is only 6% with k =5 which is the best if we compare the misclassification rate of hierarchal clustering methods and k-mediods



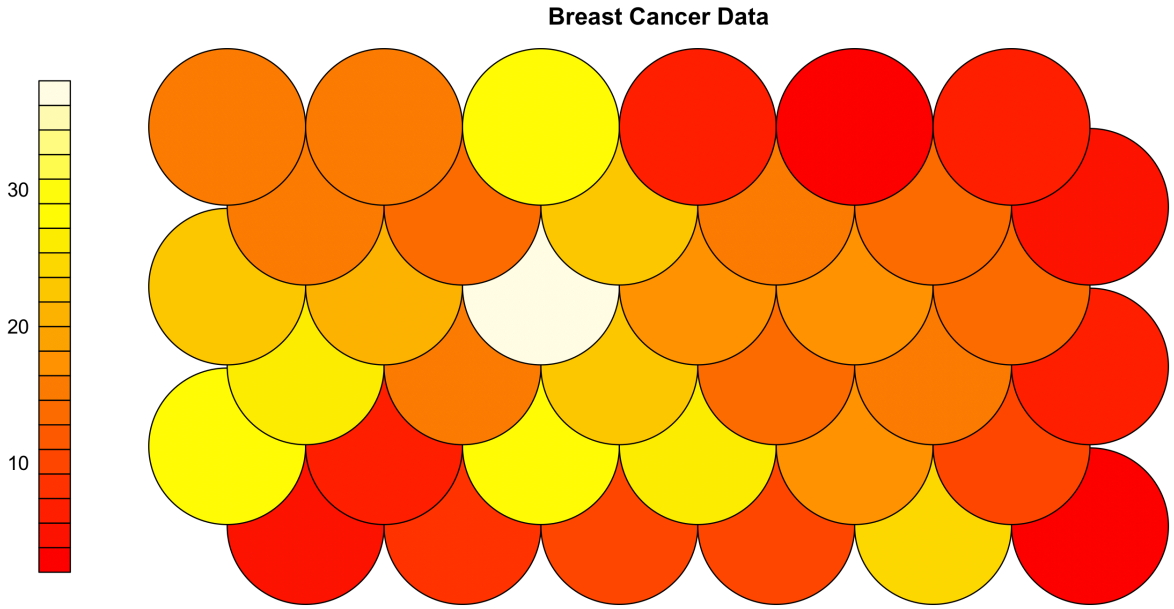
Ans-4

569 rows and 32 variables
cancer is classified as benign and malignant
data has been scaled
xdim and ydim are 6
10000 iterations

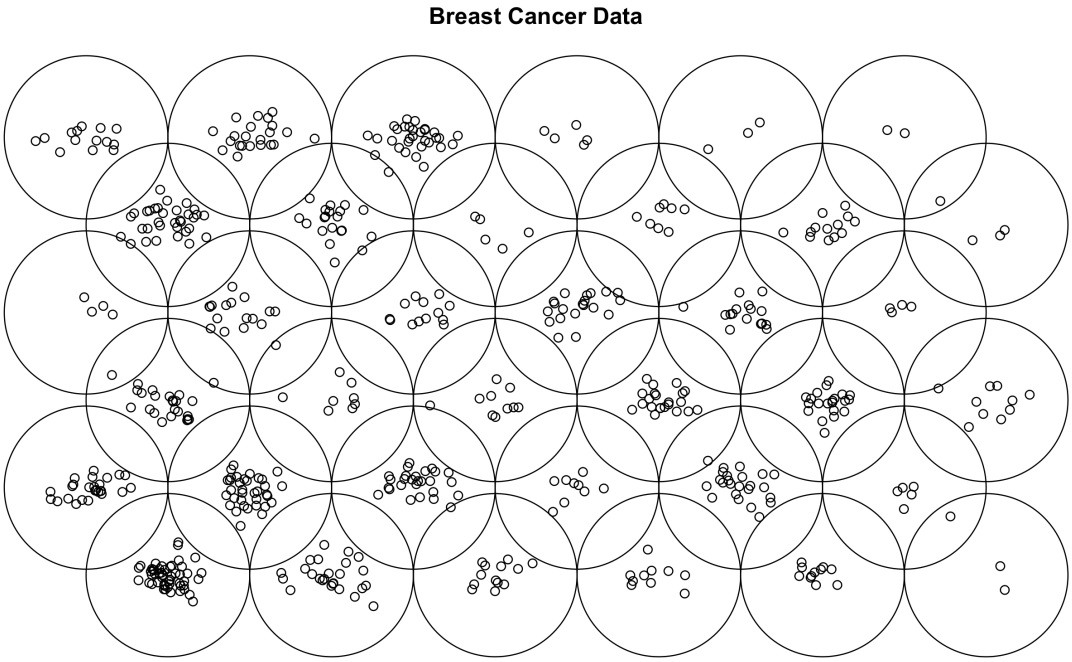


converging at approximately 7500 iterations

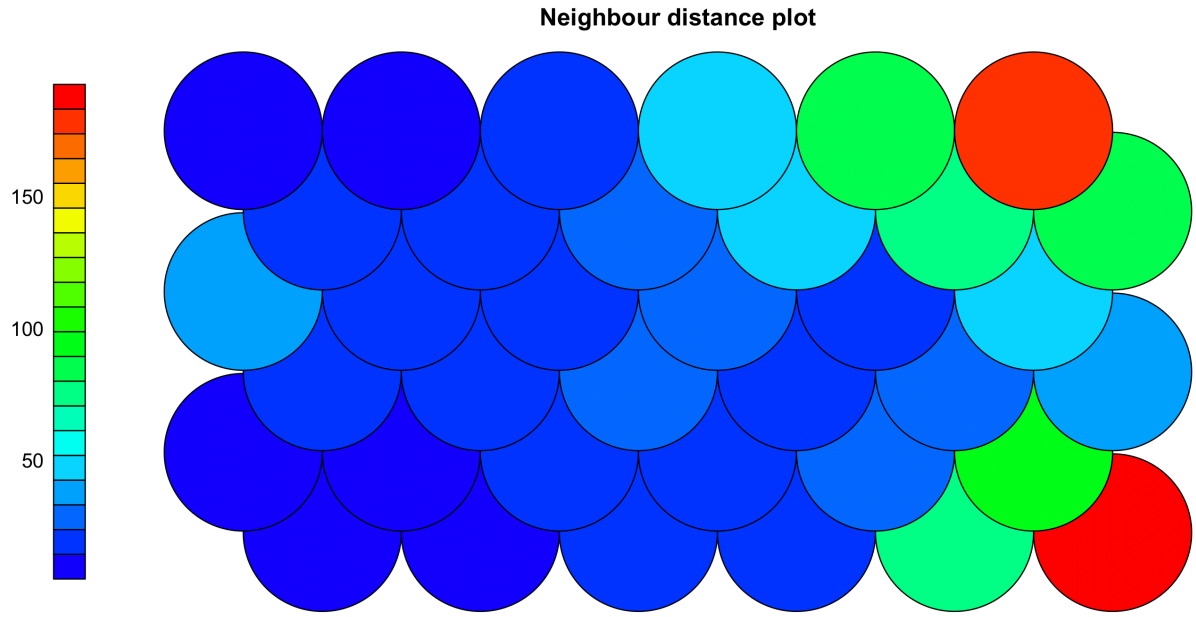
Count type plot



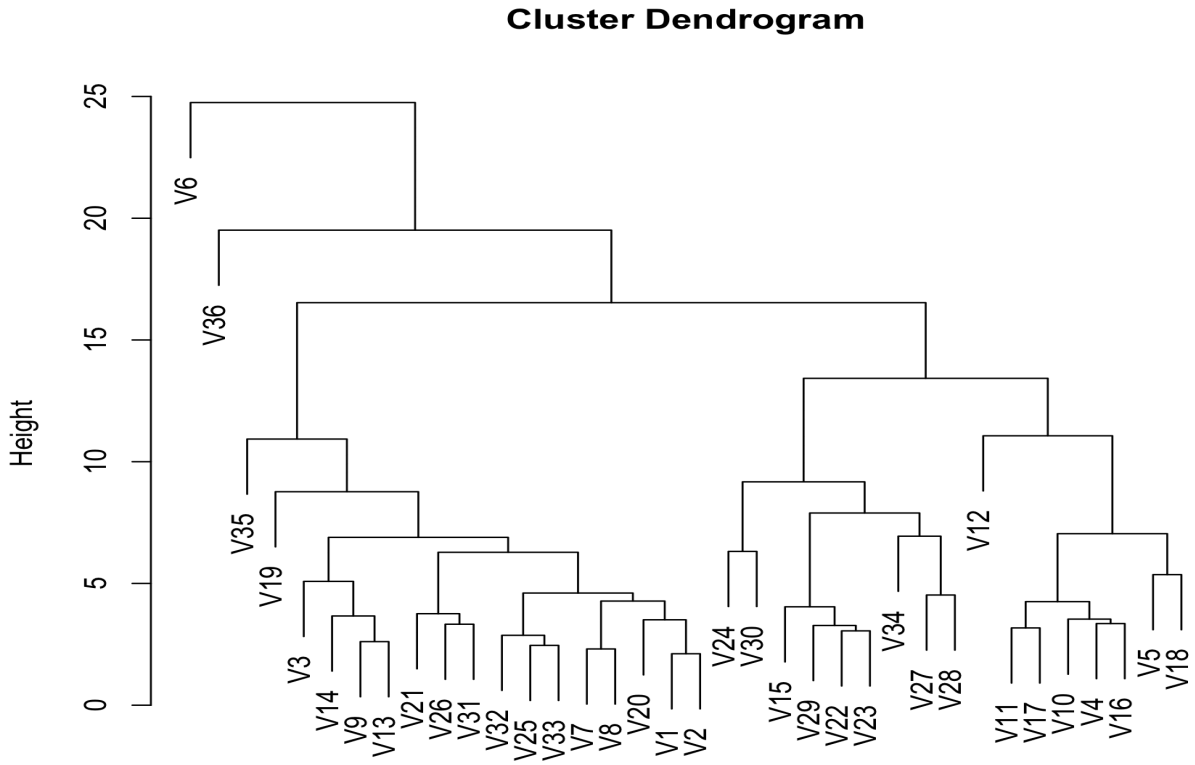
Mapping type plot



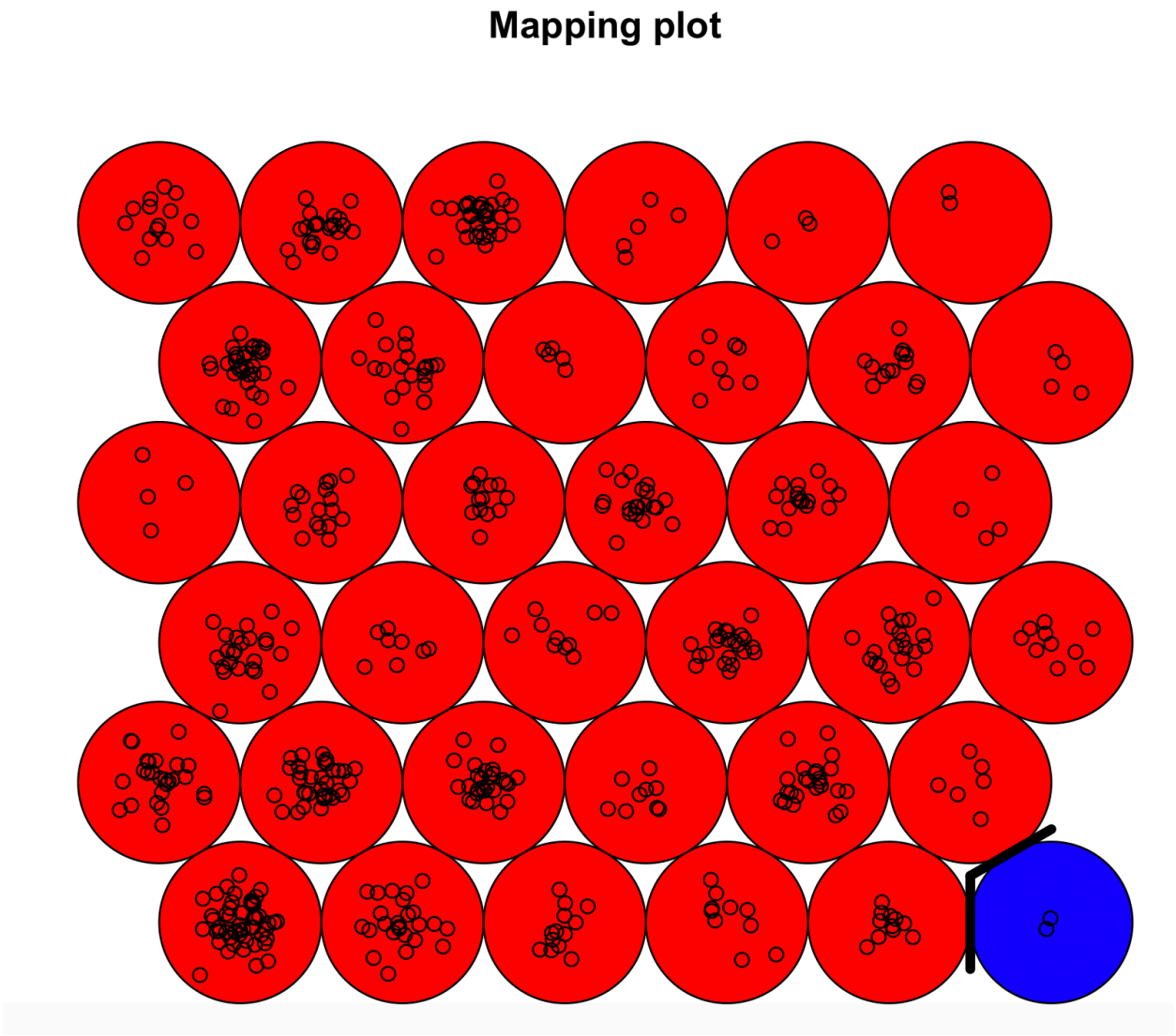
U Matrix Plot (Distant neighbors plot)



Hierarchal clustering plot

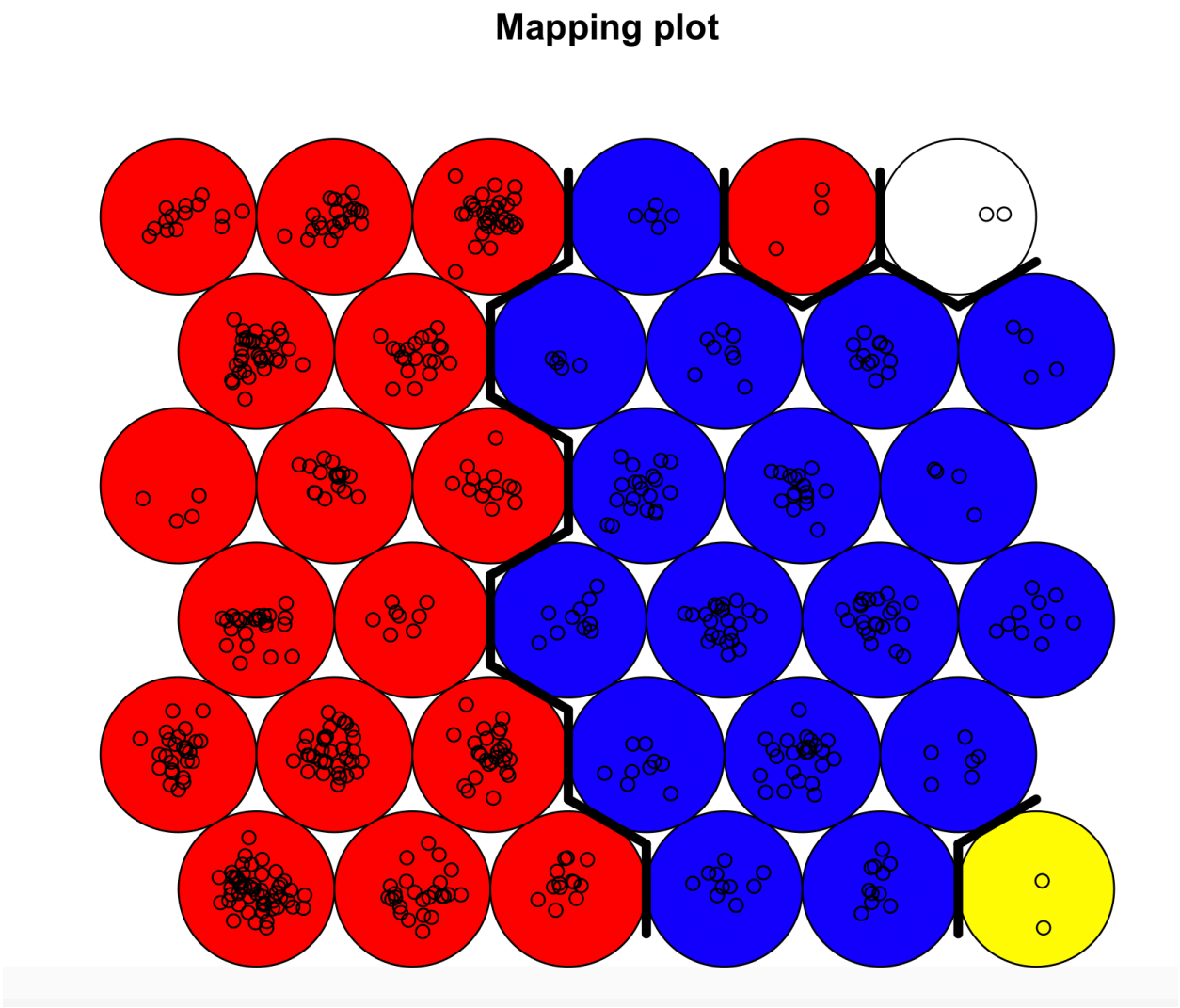


SOM plot with 2 clusters



If we observe the hierarchal clustering plot and SOM plot for k = 2, we can see that most of the observations are clustered in a group and we can say that SOM is not working well.

SOM plot with 4 clusters



If we observe the hierarchal clustering plot and SOM plot for k = 4, we can see that 2 main clusters are there and most probably there are 2 clusters due to outliers (may be because of less observations in these two clusters)