Statistical Data Mining
Spring

Assignment -2

Name : Kartik Bapna
UB ID :50291058
Class No 05

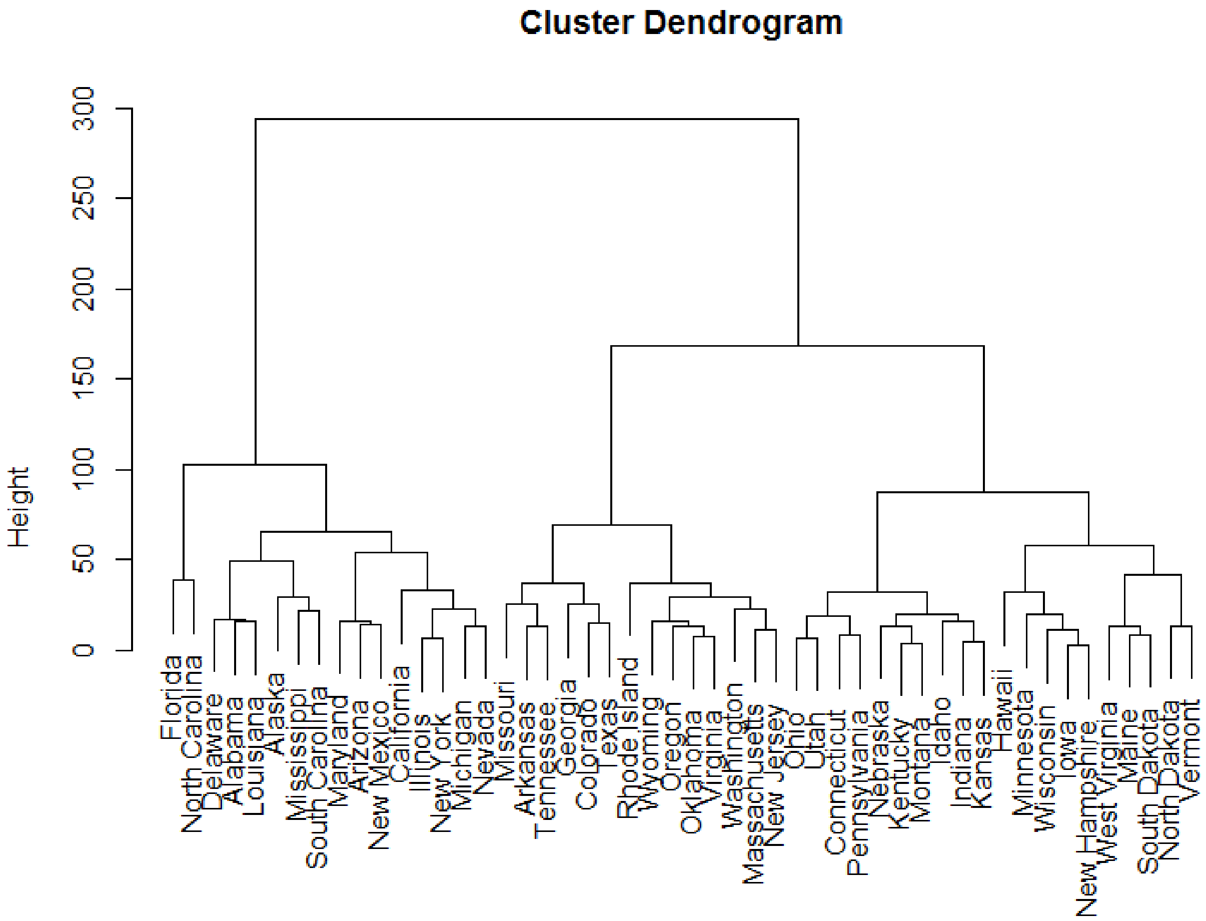
The State University of New York at Buffalo

Engineering Sciences - Data Science

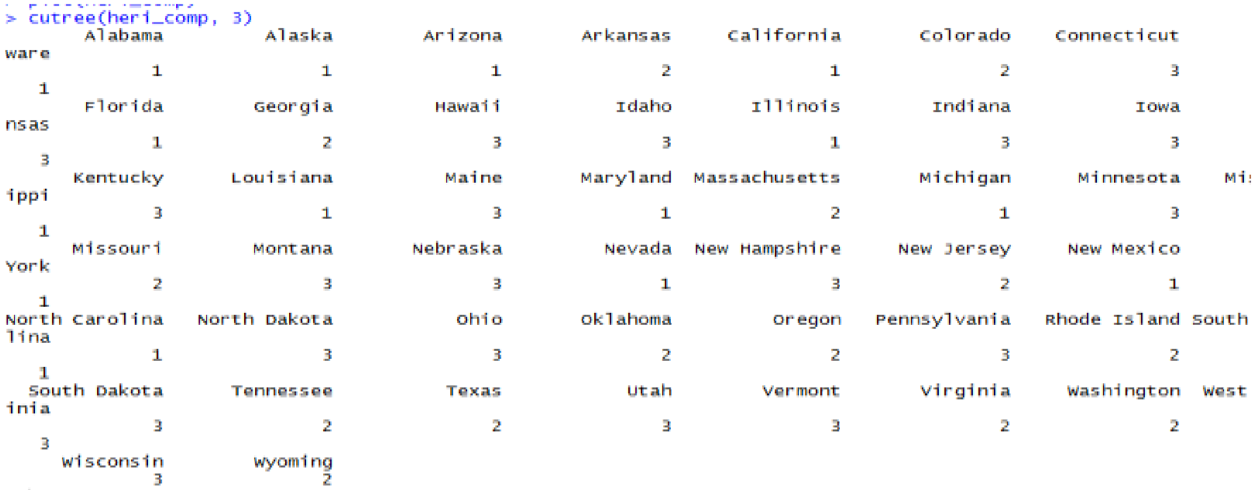
Question 1 **Consider the USArrests data. We will now perform hierarchical clustering on the states.**

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

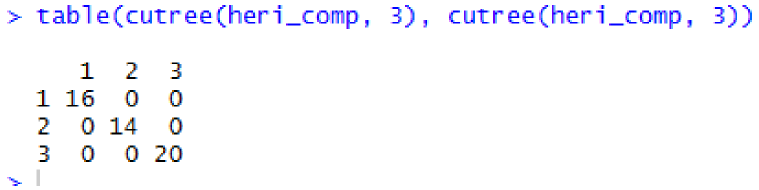
Step1 Hierarchical Clustering



(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?



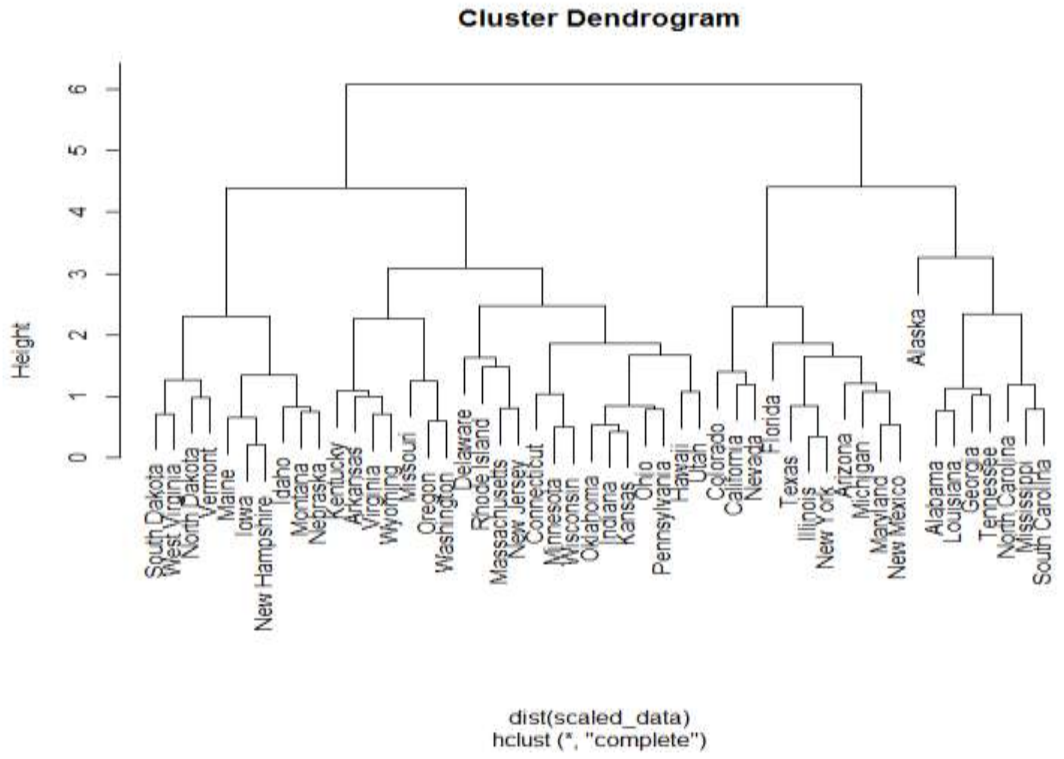
```
> cutree(heri_comp, 3)
       Alabama         Alaska        Arizona       Arkansas     California       Colorado    Connecticut
ware
             1              1              1              2              1              2              3
   1
       Florida        Georgia         Hawaii          Idaho       Illinois        Indiana           Iowa
nsas
             1              2              3              3              1              3              3
   3
      Kentucky      Louisiana          Maine       Maryland  Massachusetts       Michigan      Minnesota    Mi:
ippi
             3              1              3              1              2              1              3
   1
      Missouri        Montana       Nebraska         Nevada  New Hampshire     New Jersey     New Mexico
York
             2              3              3              1              3              2              1
   1
North Carolina   North Dakota           Ohio       Oklahoma         Oregon   Pennsylvania   Rhode Island South
lina
             1              3              3              2              2              3              2
   1
  South Dakota      Tennessee          Texas           Utah        Vermont       Virginia     Washington   West
inia
             3              2              2              3              3              2              2
   3
     Wisconsin        Wyoming
             3              2
```



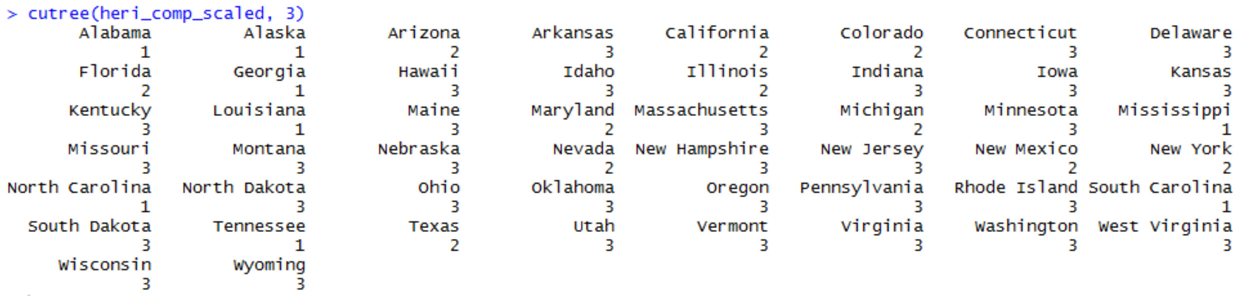
```
> table(cutree(heri_comp, 3), cutree(heri_comp, 3))
    
     1  2  3
  1 16  0  0
  2  0 14  0
  3  0  0 20
> 
```

c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

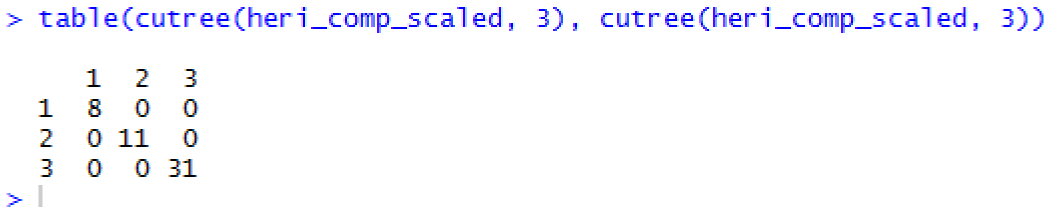
Scaling the data and then doing Hierarchical clustering



d. What effect does scaling the variables have on the hierarchical clustering obtained ? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed ? Provide a justification for your answer.



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



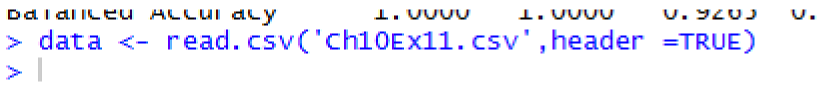
```
> table(cutree(heri_comp_scaled, 3), cutree(heri_comp_scaled, 3))
   
     1  2  3
  1  8  0  0
  2  0 11  0
  3  0  0 31
> 
```

After scaling the cluster are having different value but the tree is similar so it is advisable to scale the data first and then do clustering. As all the features would be on same scale.

Question 2 On the book website, www.StatLearning.com, there is a gene expression data set (Ch10Ex11.csv) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

Solution 2

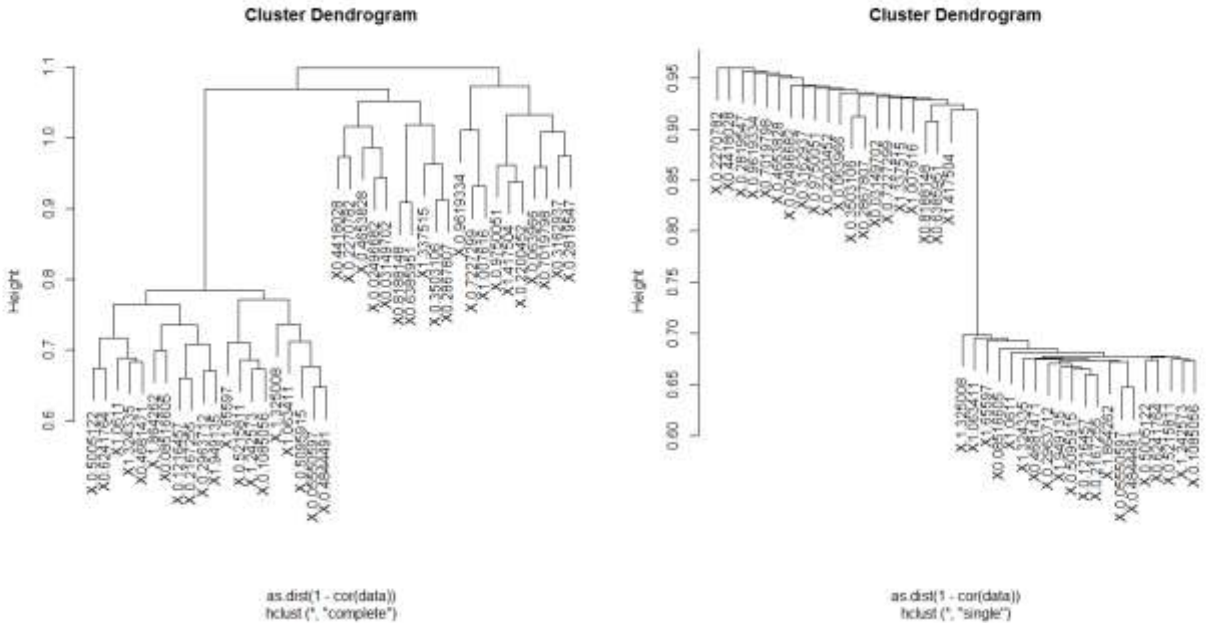
Step 1) Load the data set



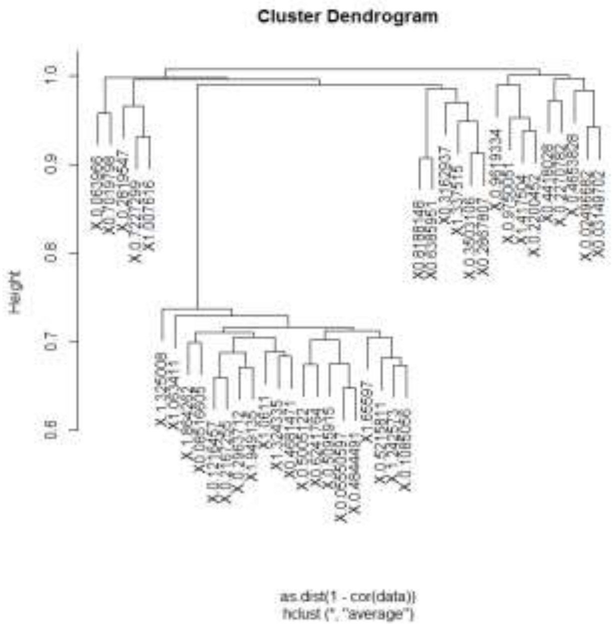
```
Balanced Accuracy       1.0000    1.0000    0.9265   0.
> data <- read.csv('Ch10Ex11.csv',header =TRUE)
> |
```

Step 2) Appling hierarchal clustering using three methods single, complete and average

Complete hierarchal clustering and Single Clustering plots



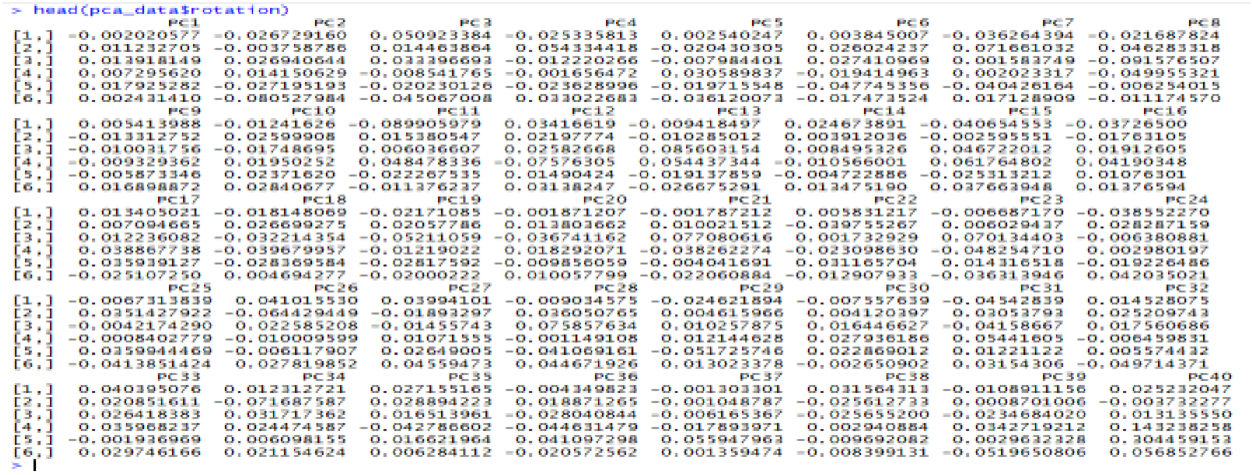
Plot for average hierarchal clustering



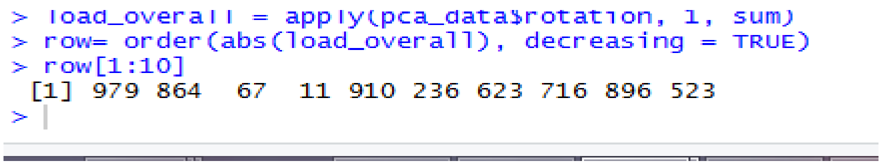
From the above plot we can see that results are different for all three methods

Clustering	single	complete	average
No of cluster	2	2	3

Step 3) Doing PCA can help understand which genes differ the most and will capture the max variation in each components from pc1 onwards as shown



```
> head(pca_data$rotation)
              PC1          PC2          PC3          PC4          PC5          PC6          PC7          PC8
[1,] -0.002020577 -0.026729160  0.050923384 -0.025335813  0.002540247  0.003845007 -0.036264394 -0.021687824
[2,]  0.011232705 -0.003758786  0.014463864  0.054334418 -0.020430305  0.026024237  0.071661032  0.046283318
[3,]  0.013918149  0.026940644  0.033396693 -0.012220266 -0.007984401  0.027410969  0.001583749 -0.091576507
[4,]  0.007295620  0.014150629 -0.008541765 -0.001656472  0.030589837 -0.019414963  0.002023317 -0.049955321
[5,]  0.017925282 -0.027195193 -0.020230126 -0.023628996 -0.019715548 -0.047745356 -0.040426164 -0.006254015
[6,]  0.002431410 -0.080527984 -0.045067008  0.033022683 -0.036120073 -0.017473524  0.017128909 -0.011174570
              PC9        PC10         PC11        PC12         PC13         PC14         PC15        PC16
[1,]  0.005413988 -0.01241626 -0.089905979  0.03416619 -0.009418497  0.024673891 -0.040654553 -0.03726500
[2,] -0.013312752  0.02599908  0.015380547  0.02197774 -0.010285012  0.003912036 -0.002595551 -0.01763105
[3,] -0.010031756 -0.01748695  0.006036607  0.02582668  0.085603154  0.008495326  0.046722012  0.01912605
[4,] -0.009329362  0.01950252  0.048478336 -0.07576305  0.054437344 -0.010566001  0.061764802  0.04190348
[5,] -0.005873346  0.02371620 -0.022267535  0.01490424 -0.019137859 -0.004722886 -0.025313212  0.01076301
[6,]  0.016898872  0.02840677 -0.011376237  0.03138247 -0.026675291  0.013475190  0.037663948  0.01376594
             PC17         PC18        PC19         PC20         PC21         PC22         PC23         PC24
[1,]  0.013405021 -0.018148069 -0.02171085 -0.001871207 -0.001787212  0.005831217 -0.006687170 -0.038552270
[2,]  0.007094665  0.026699275  0.02057786  0.013803662  0.010021512 -0.039755267  0.006029437  0.028287159
[3,]  0.012236082 -0.032214354 -0.05211059 -0.036741162  0.077080616  0.001732929  0.070134403 -0.006380881
[4,]  0.038867738 -0.039679957 -0.01219022  0.018292071 -0.038262274 -0.023098630 -0.048254710  0.002980197
[5,]  0.035939127 -0.028369584 -0.02817592 -0.009856059 -0.004041691  0.031165704  0.014316518 -0.019226486
[6,] -0.025107250  0.004694277 -0.02000222  0.010057799 -0.022060884 -0.012907933 -0.036313946  0.042035021
              PC25         PC26        PC27         PC28         PC29         PC30        PC31         PC32
[1,] -0.0067313839  0.041015530  0.03994101 -0.009034575 -0.024621894 -0.007557639 -0.04542839  0.014528075
[2,]  0.0351427922 -0.064429449 -0.01893297  0.036050765  0.004615966  0.004120397  0.03053793  0.025209743
[3,] -0.0042174290  0.022585208 -0.01455743  0.075857634  0.010257875  0.016446627 -0.04158667  0.017560686
[4,] -0.0008402779 -0.010009599  0.01071555 -0.001149108  0.012144628  0.027936186  0.05441605 -0.006459831
[5,]  0.0359944469 -0.006117907  0.02649005 -0.041069161 -0.051725746  0.022869012  0.01221122  0.005574432
[6,] -0.0413851424  0.027819852  0.04559473  0.044671926  0.013023378 -0.002650902  0.03154306 -0.049714371
             PC33         PC34         PC35         PC36         PC37         PC38          PC39        PC40
[1,]  0.040395076  0.012312721  0.027155165 -0.004349823 -0.001303301  0.031564313 -0.0108911156  0.025232047
[2,]  0.020851611 -0.071687587  0.028894223  0.018871265 -0.001048787 -0.025612733  0.0008701006 -0.003732277
[3,]  0.026418383  0.031717362  0.016513961 -0.028040844 -0.006165367 -0.025655200 -0.0234684020  0.013135550
[4,]  0.035968237  0.024474587 -0.042786602 -0.044631479 -0.017893971  0.002940884  0.0342719212  0.143238258
[5,] -0.001936969  0.006098155  0.016621964  0.041097298  0.055947963 -0.009692082  0.0029632328  0.304459153
[6,]  0.029746166  0.021154624  0.006284112 -0.020572562  0.001359474 -0.008399131 -0.0519650806  0.056852766
> 
```



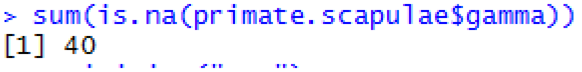
```
> load_overall = apply(pca_data$rotation, 1, sum)
> row= order(abs(load_overall), decreasing = TRUE)
> row[1:10]
 [1] 979 864  67  11 910 236 623 716 896 523
> 
```

Question 3) Access the data “primate.scapulae” a) Cluster the data based on single-linkage, average linkage, and complete-linkage agglomerative hierarchical clustering. Decide on the groupings, and justify it, for all three methods. Calculate the misclassification rate. Which method performed the best and which method performed the worst? Was the result in line with your expectations?

Solution 3)

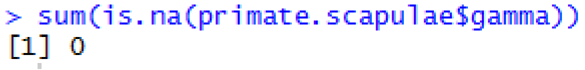
Step1) Load the data

Step 2) Removing the na there are 40 missing values



```
> sum(is.na(primate.scapulae$gamma))
[1] 40
```

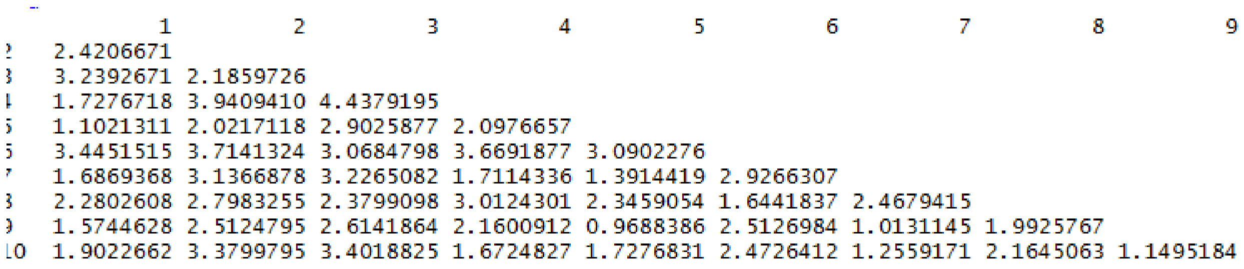
After removing missing value check



```
> sum(is.na(primate.scapulae$gamma))
[1] 0
```

Step 3) Scaling the data

Step 4) Calculating the distance matrix



```
           1         2         3         4         5         6         7         8         9
2  2.4206671
3  3.2392671 2.1859726
4  1.7276718 3.9409410 4.4379195
5  1.1021311 2.0217118 2.9025877 2.0976657
6  3.4451515 3.7141324 3.0684798 3.6691877 3.0902276
7  1.6869368 3.1366878 3.2265082 1.7114336 1.3914419 2.9266307
8  2.2802608 2.7983255 2.3799098 3.0124301 2.3459054 1.6441837 2.4679415
9  1.5744628 2.5124795 2.6141864 2.1600912 0.9688386 2.5126984 1.0131145 1.9925767
10 1.9022662 3.3799795 3.4018825 1.6724827 1.7276831 2.4726412 1.2559171 2.1645063 1.1495184
```

step 5) calculating for single linkage hierarchal clustering, displaying table and printing confusion matrix

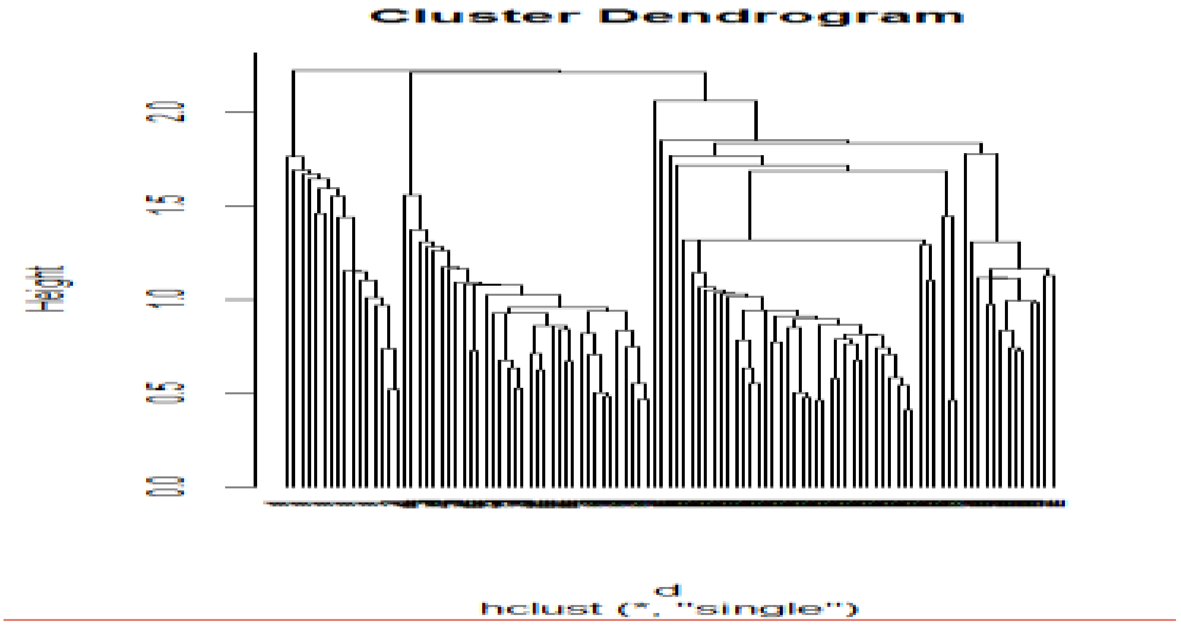
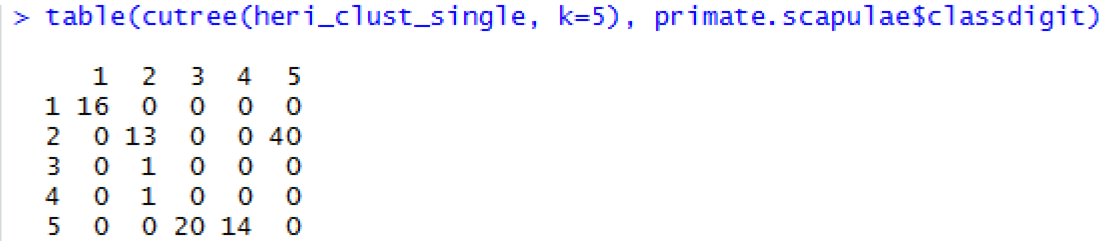
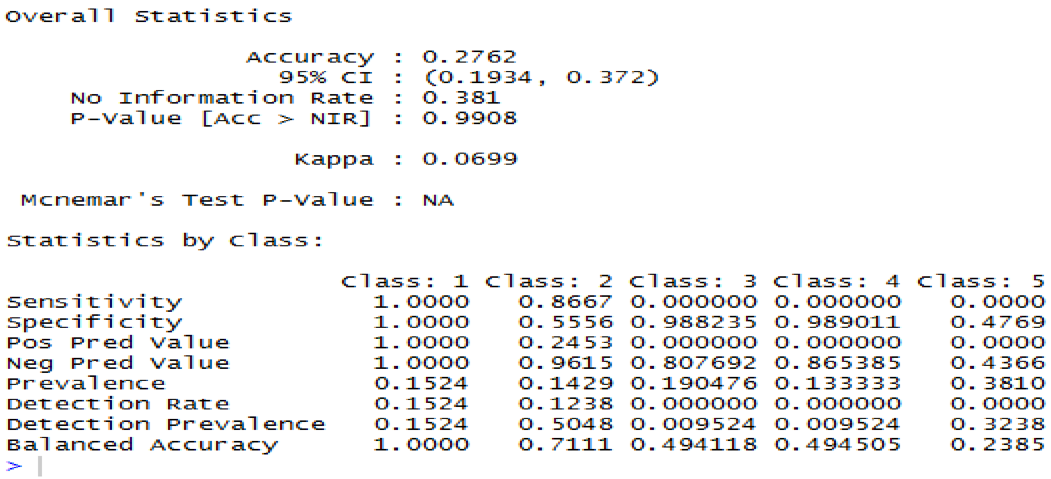


Table for single linkage as we can see 2 samples were misclassified



```
> table(cutree(heri_clust_single, k=5), primate.scapulae$classdigit)

     1  2  3  4  5
  1 16  0  0  0  0
  2  0 13  0  0 40
  3  0  1  0  0  0
  4  0  1  0  0  0
  5  0  0 20 14  0
```



```
Overall Statistics

               Accuracy : 0.2762
                 95% CI : (0.1934, 0.372)
    No Information Rate : 0.381
    P-Value [Acc > NIR] : 0.9908

                  Kappa : 0.0699

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            1.0000   0.8667 0.000000 0.000000   0.0000
Specificity            1.0000   0.5556 0.988235 0.989011   0.4769
Pos Pred Value         1.0000   0.2453 0.000000 0.000000   0.0000
Neg Pred Value         1.0000   0.9615 0.807692 0.865385   0.4366
Prevalence             0.1524   0.1429 0.190476 0.133333   0.3810
Detection Rate         0.1524   0.1238 0.000000 0.000000   0.0000
Detection Prevalence   0.1524   0.5048 0.009524 0.009524   0.3238
Balanced Accuracy      1.0000   0.7111 0.494118 0.494505   0.2385
>
```

Step 6) Calculating **Average clustering** linkage , plotting the cluster and display confusion matrix

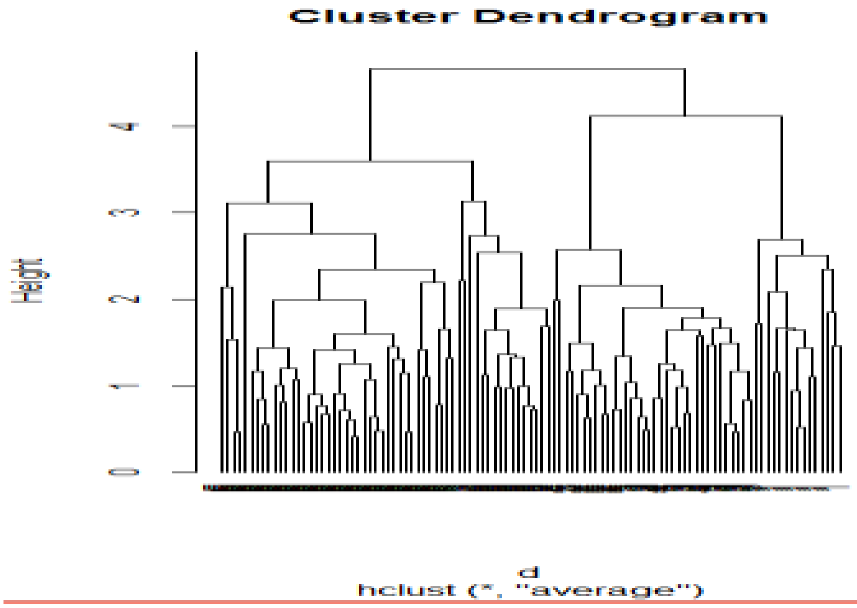
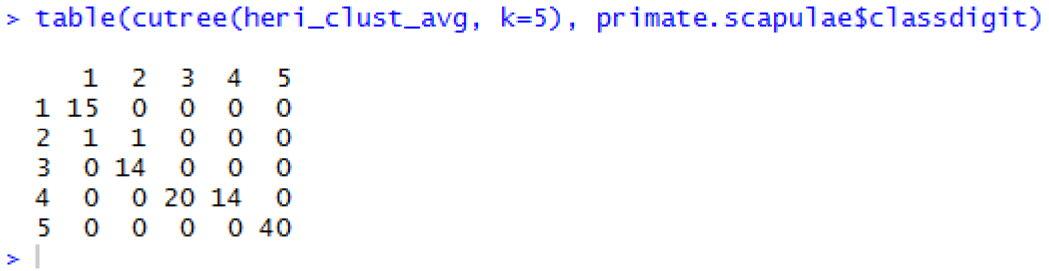
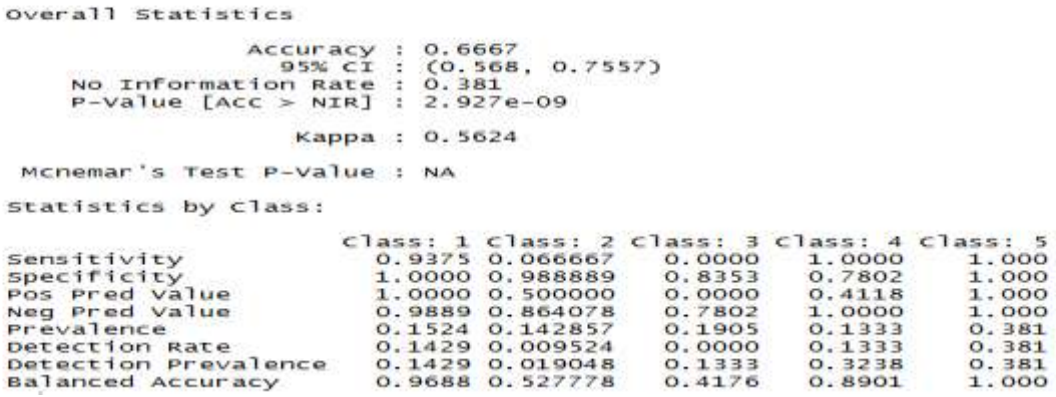


Table for the Average hierarchical clustering as we can see 2 were miss classified



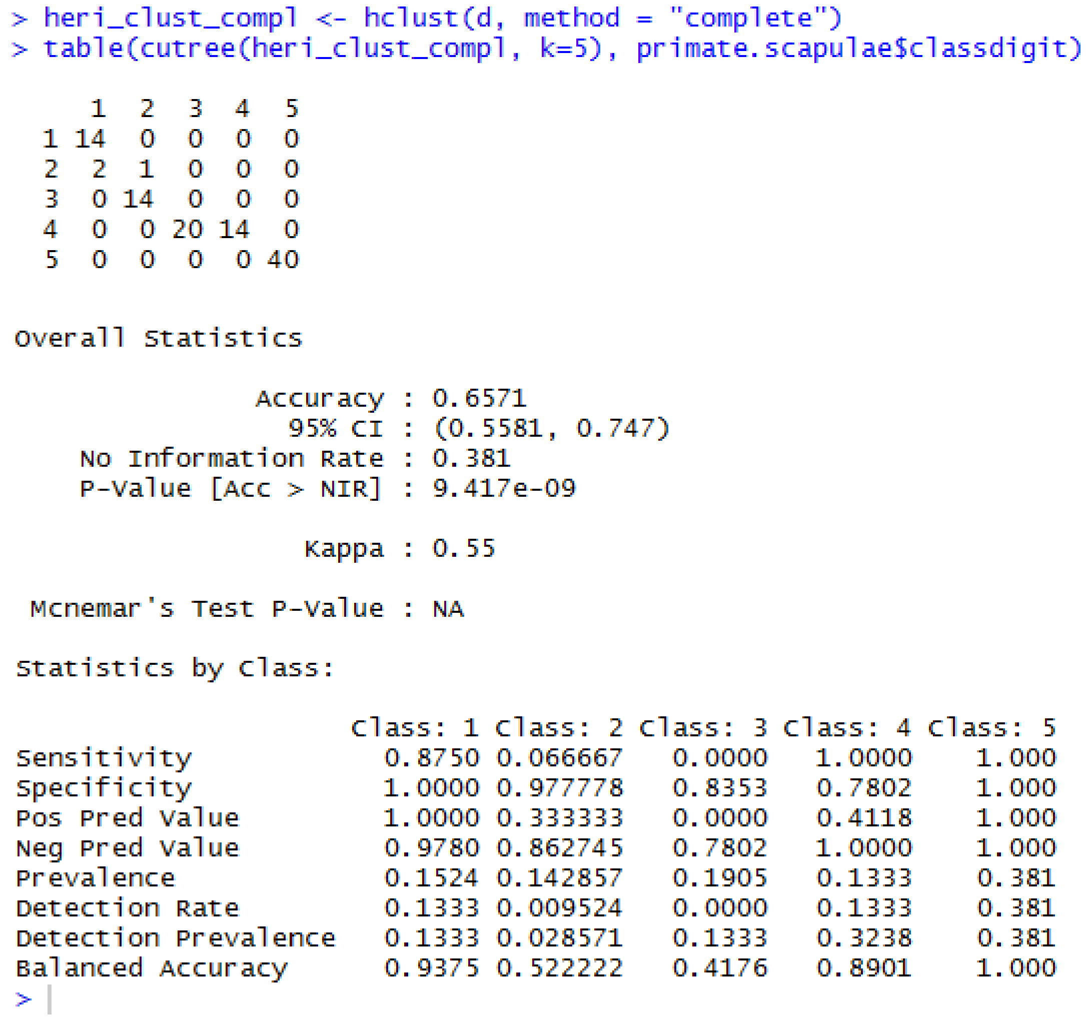
```
> table(cutree(heri_clust_avg, k=5), primate.scapulae$classdigit)

     1  2  3  4  5
  1 15  0  0  0  0
  2  1  1  0  0  0
  3  0 14  0  0  0
  4  0  0 20 14  0
  5  0  0  0  0 40
> 
```

Confusion Matrix for Average hierarchical clustering



```
Overall Statistics

               Accuracy : 0.6667
                 95% CI : (0.568, 0.7557)
    No Information Rate : 0.381
    P-Value [Acc > NIR] : 2.927e-09

                  Kappa : 0.5624

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.9375 0.066667   0.0000   1.0000    1.000
Specificity            1.0000 0.988889   0.8353   0.7802    1.000
Pos Pred Value         1.0000 0.500000   0.0000   0.4118    1.000
Neg Pred Value         0.9889 0.864078   0.7802   1.0000    1.000
Prevalence             0.1524 0.142857   0.1905   0.1333    0.381
Detection Rate         0.1429 0.009524   0.0000   0.1333    0.381
Detection Prevalence   0.1429 0.019048   0.1333   0.3238    0.381
Balanced Accuracy      0.9688 0.527778   0.4176   0.8901    1.000
```

Step 7) Calculating for **complete** hierarchical clustering
as we can see 3 were misclassified



```
> heri_clust_compl <- hclust(d, method = "complete")
> table(cutree(heri_clust_compl, k=5), primate.scapulae$classdigit)

     1  2  3  4  5
  1 14  0  0  0  0
  2  2  1  0  0  0
  3  0 14  0  0  0
  4  0  0 20 14  0
  5  0  0  0  0 40

Overall Statistics

               Accuracy : 0.6571
                 95% CI : (0.5581, 0.747)
    No Information Rate : 0.381
    P-Value [Acc > NIR] : 9.417e-09

                  Kappa : 0.55

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.8750 0.066667   0.0000   1.0000    1.000
Specificity            1.0000 0.977778   0.8353   0.7802    1.000
Pos Pred Value         1.0000 0.333333   0.0000   0.4118    1.000
Neg Pred Value         0.9780 0.862745   0.7802   1.0000    1.000
Prevalence             0.1524 0.142857   0.1905   0.1333    0.381
Detection Rate         0.1333 0.009524   0.0000   0.1333    0.381
Detection Prevalence   0.1333 0.028571   0.1333   0.3238    0.381
Balanced Accuracy      0.9375 0.522222   0.4176   0.8901    1.000
> 
```

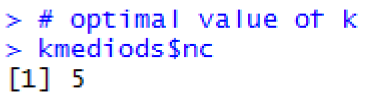
Average hierarchical clustering has the best accuracy with 65% among all the three clustering method, Yes average was the best way to calculate compared to single and complete clustering method.

Question 3 b

Cluster the data based on K-means or K-medoids. Use an analytical technique to justify your choice in “k”. How did the performance compare to the hierarchical clustering of part a? Which did you feel was a better method for this data?

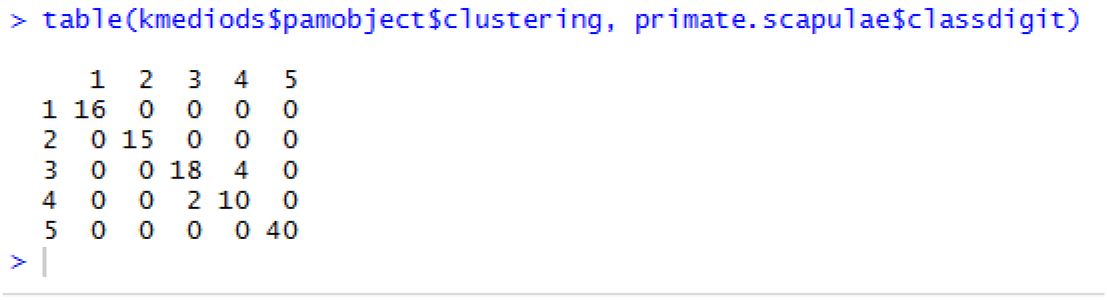
Step1) Clustering the data based on k mediods

Step 2) Optimal value of K



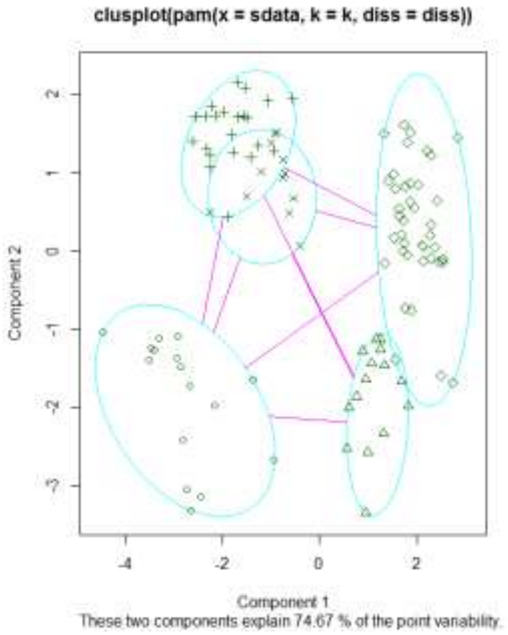
```
> # optimal value of k
> kmediods$nc
[1] 5
```

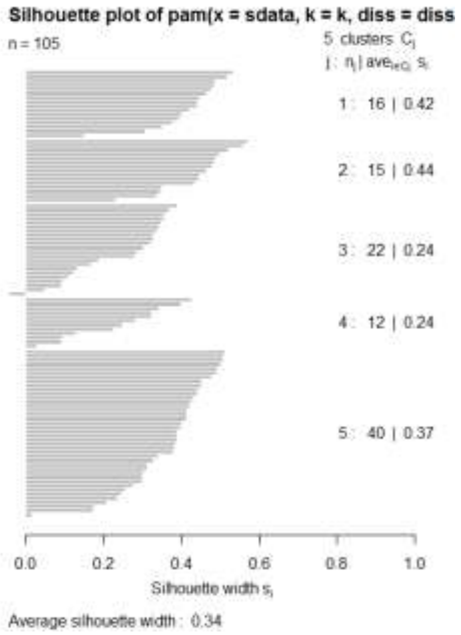
Step 3) tabulating the results



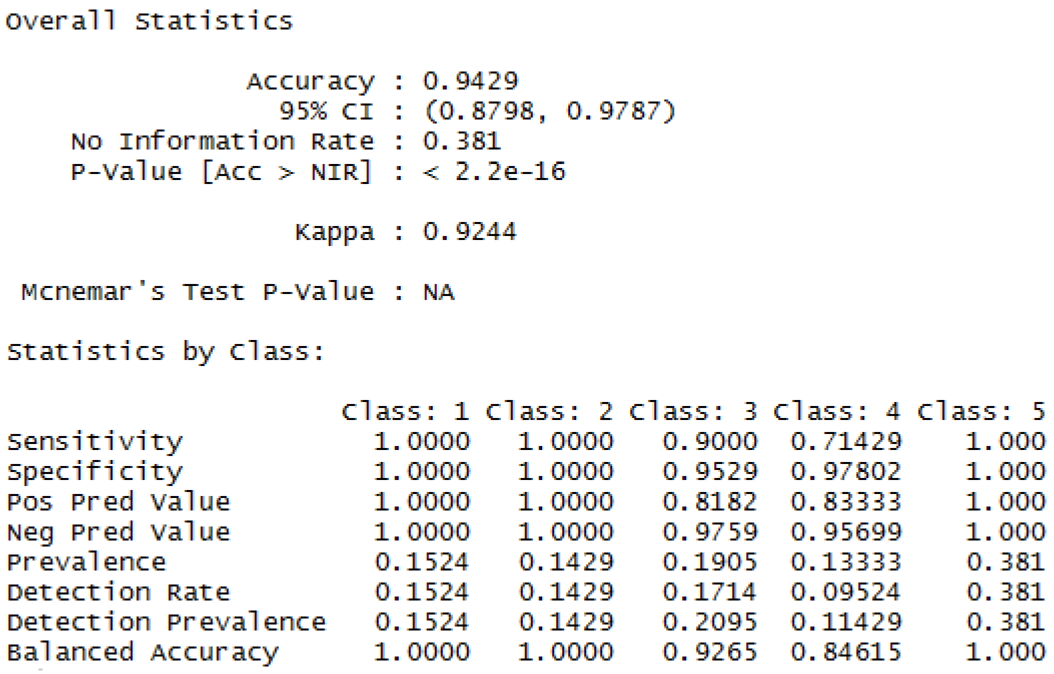
```
> table(kmediods$pamobject$clustering, primate.scapulae$classdigit)

     1  2  3  4  5
  1 16  0  0  0  0
  2  0 15  0  0  0
  3  0  0 18  4  0
  4  0  0  2 10  0
  5  0  0  0  0 40
> 
```

Step 4) Plotting the results





Step 5) Confusion Matrix observations



```
Overall Statistics

               Accuracy : 0.9429
                 95% CI : (0.8798, 0.9787)
    No Information Rate : 0.381
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9244

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            1.0000   1.0000   0.9000  0.71429    1.000
Specificity            1.0000   1.0000   0.9529  0.97802    1.000
Pos Pred Value         1.0000   1.0000   0.8182  0.83333    1.000
Neg Pred Value         1.0000   1.0000   0.9759  0.95699    1.000
Prevalence             0.1524   0.1429   0.1905  0.13333    0.381
Detection Rate         0.1524   0.1429   0.1714  0.09524    0.381
Detection Prevalence   0.1524   0.1429   0.2095  0.11429    0.381
Balanced Accuracy      1.0000   1.0000   0.9265  0.84615    1.000
```

Accuracy is 94% as per the accuracy K-medoids way is better.

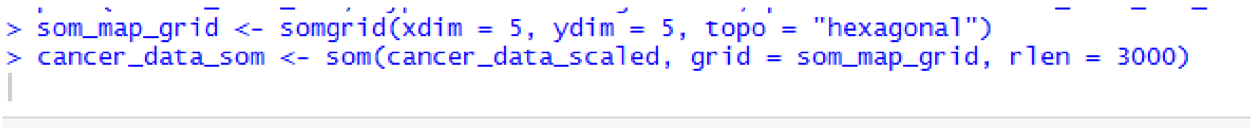
Question 4 Run a batch-SOM analysis on the Wisconsin Breast-Cancer data (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic)). Describe how well the SOM methods cluster the tumor cases into benign and malignant. Compute the U-matrix and discuss its representation for these data.

Step1) Load the data

Step 2) Converting the data types to numeric

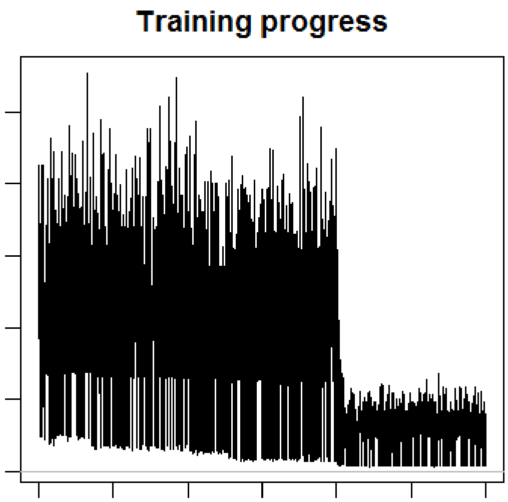
Step 3) Scaling the data

Step4) making the som grid applying the som model

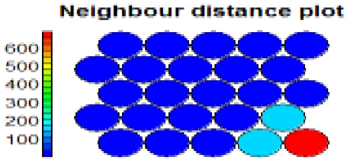


```
> som_map_grid <- somgrid(xdim = 5, ydim = 5, topo = "hexagonal")
> cancer_data_som <- som(cancer_data_scaled, grid = som_map_grid, rlen = 3000)
```

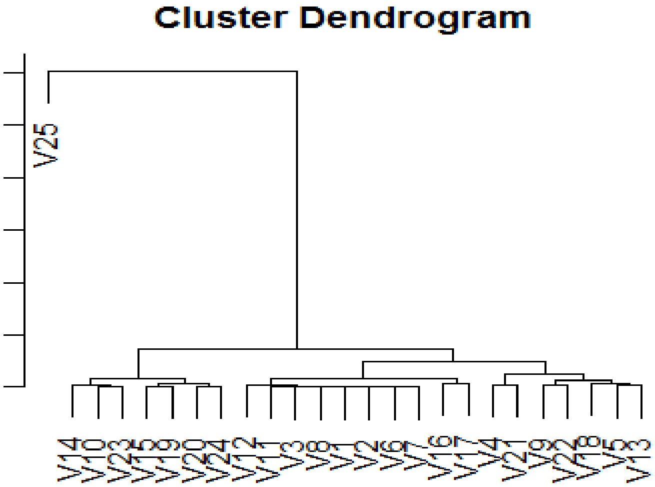
Step 5) Plotting the som codes with type as changes and we can see it converges after 2000 iterations



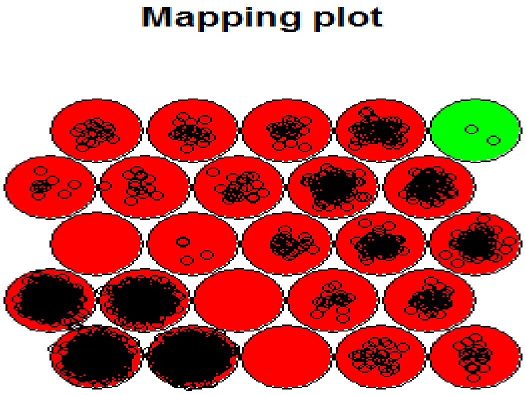
Step 7) SOM U matrix plot as we can see the dark blue nodes are near by based on Euclidean distance , light blue are bit further and red are at extremes.



Step 8) Clustering the codes



Step 9) Plotting the results and we see two clusters are formed clearly


Mapping plot