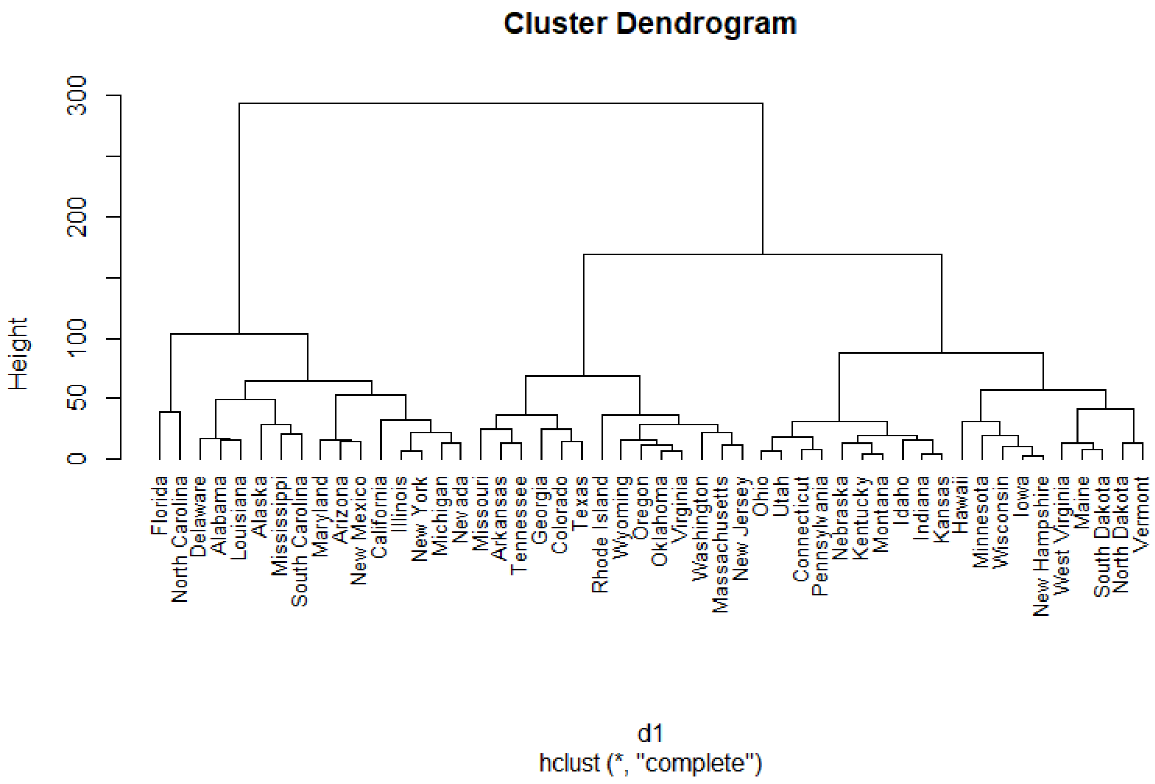
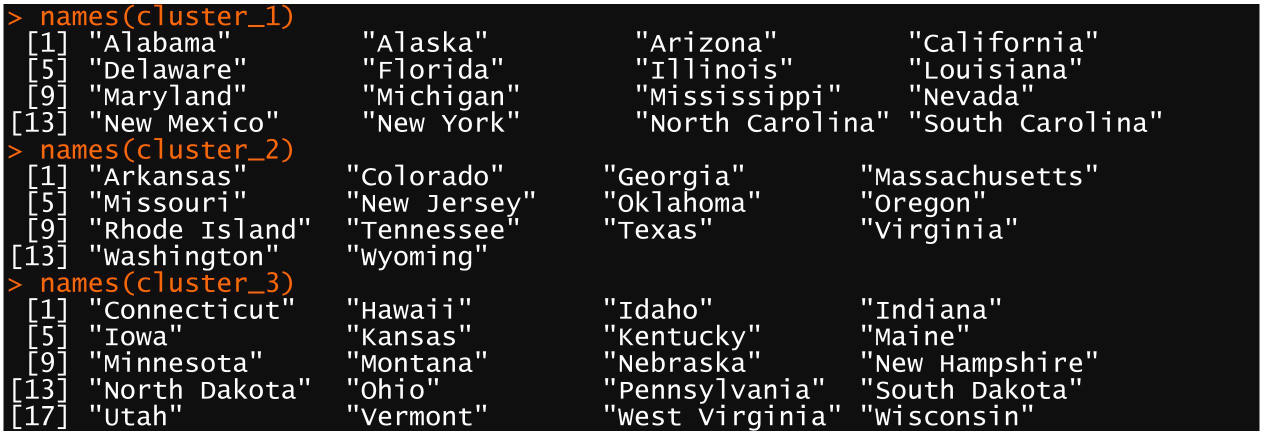
Data Mining II HW-2 Report

ISLR Chapter 10, P9 –

a) Clustering the states using Hierarchal Clustering with complete linkage and Euclidean distance

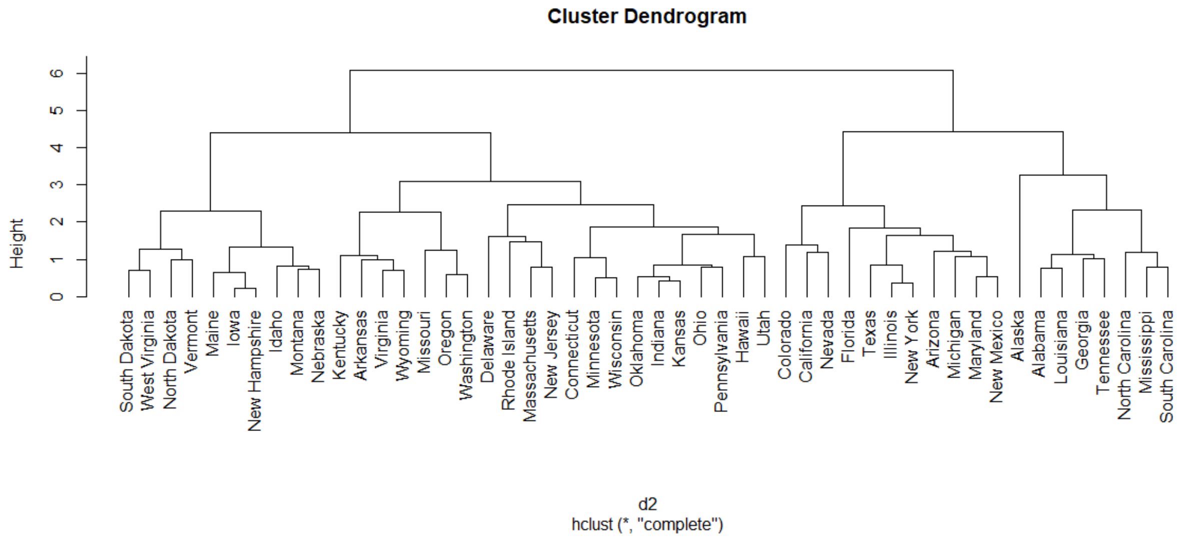


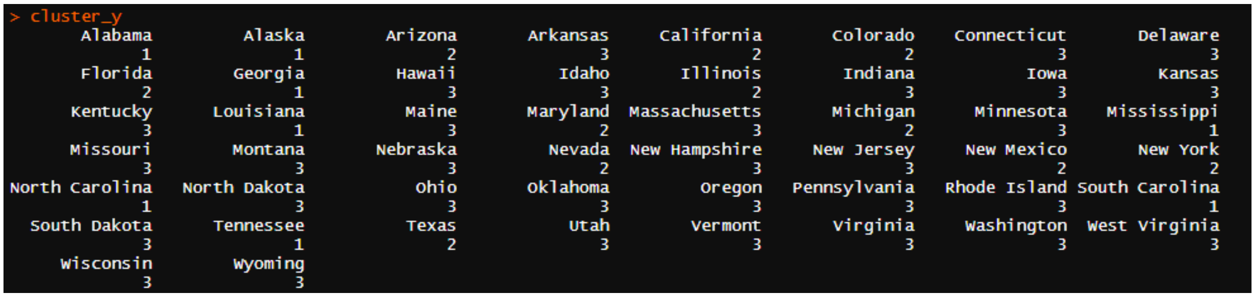
b) Cutting the dendrogram at a height that results in 3 different clusters.



```
> names(cluster_1)
 [1] "Alabama"        "Alaska"         "Arizona"        "California"
 [5] "Delaware"       "Florida"        "Illinois"       "Louisiana"
 [9] "Maryland"       "Michigan"       "Mississippi"    "Nevada"
[13] "New Mexico"     "New York"       "North Carolina" "South Carolina"
> names(cluster_2)
 [1] "Arkansas"      "Colorado"      "Georgia"       "Massachusetts"
 [5] "Missouri"      "New Jersey"    "Oklahoma"      "Oregon"
 [9] "Rhode Island"  "Tennessee"     "Texas"         "Virginia"
[13] "Washington"    "Wyoming"
> names(cluster_3)
 [1] "Connecticut"   "Hawaii"        "Idaho"         "Indiana"
 [5] "Iowa"          "Kansas"        "Kentucky"      "Maine"
 [9] "Minnesota"     "Montana"       "Nebraska"      "New Hampshire"
[13] "North Dakota"  "Ohio"          "Pennsylvania"  "South Dakota"
[17] "Utah"          "Vermont"       "West Virginia" "Wisconsin"
```

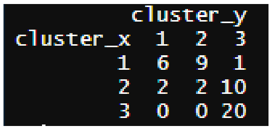
c) Scaling the variables to have a standard deviation of 1 and hierarchically clustering the states using complete linkage and Euclidean distance.





```
> cluster_y
       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 
```

d) Table of scaled and unscaled clusters



```
         cluster_y
cluster_x  1  2  3
        1  6  9  1
        2  2  2 10
        3  0  0 20
```

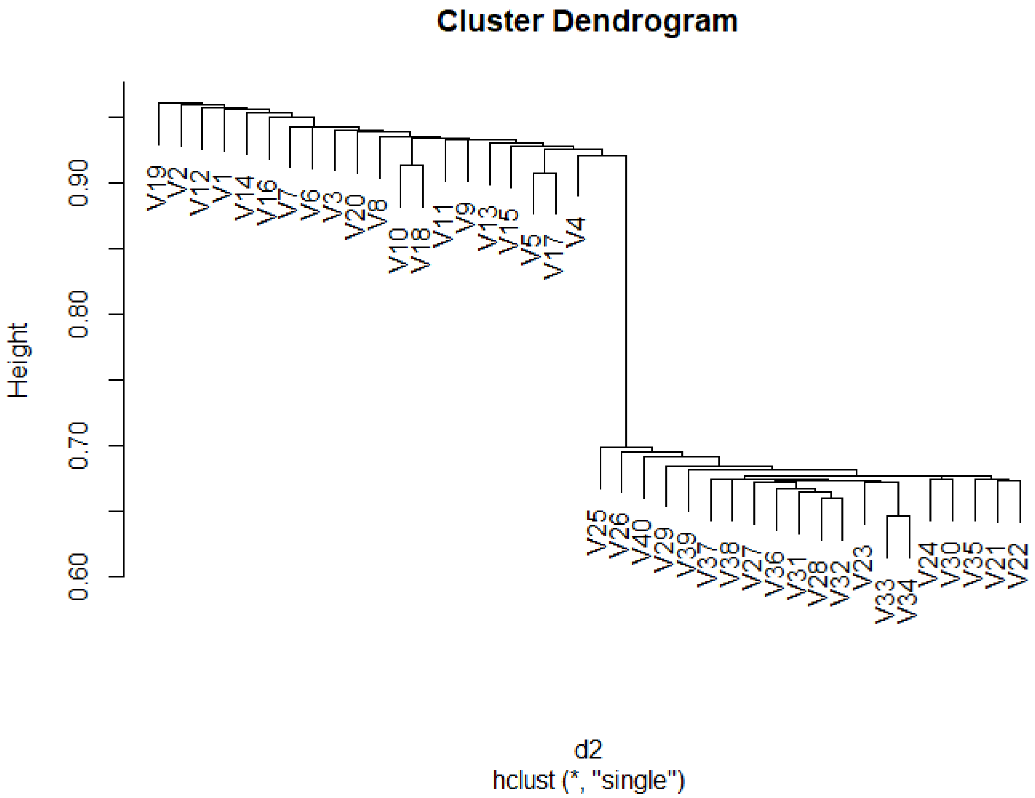
It is evident that scaling the data has contributed to clusters of different states as compared to clustering from unscaled data. Furthermore, the number of states within each cluster has changed significantly. Scaling the variables should depend upon the particular dataset and the unit of measurement of its variables. Since Murder, Assault , Rape and Urban Population are quantified using incomparable units of measurements, scaling the variables will provide better results for this case. As the choice of the measuring units gives rise to relative weights of the variable, scaling attempts to give all variables a similar weight which might lead to better results although depending on the particular application and the data, some variables might be intrinsically more or less important and require a higher or lower weightage.

ISLR Chapter 10, P10 –

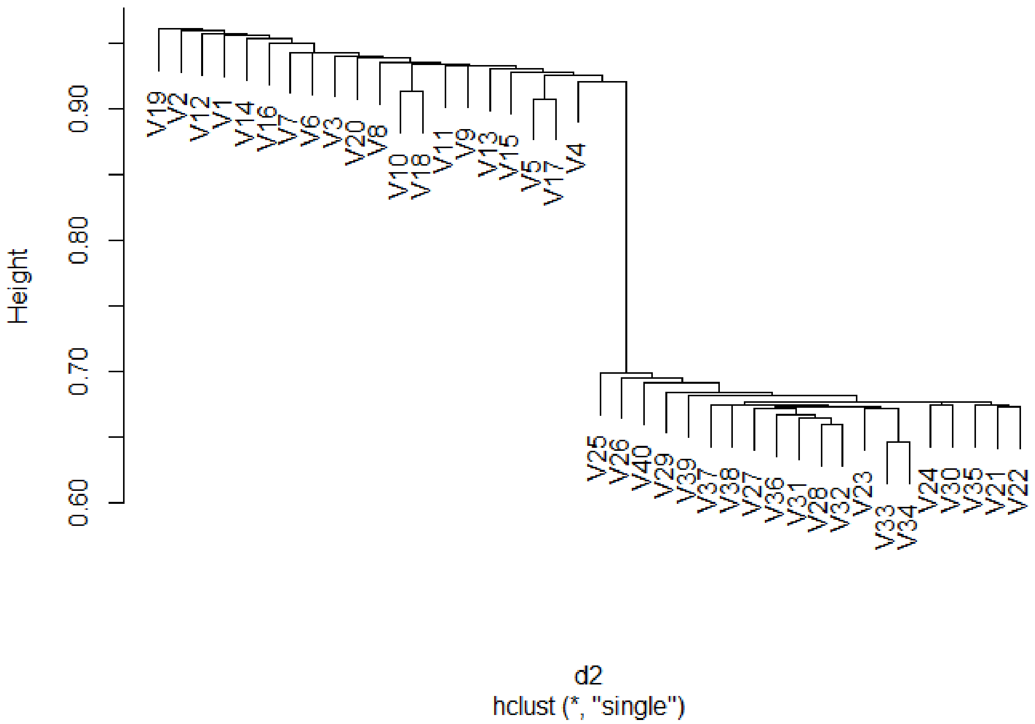
a) Using read.csv to load the data from Ch10Ex11.csv.

```
data<-read.csv("C:/Users/X/Desktop/Ch10Ex11.csv", header = F)
```

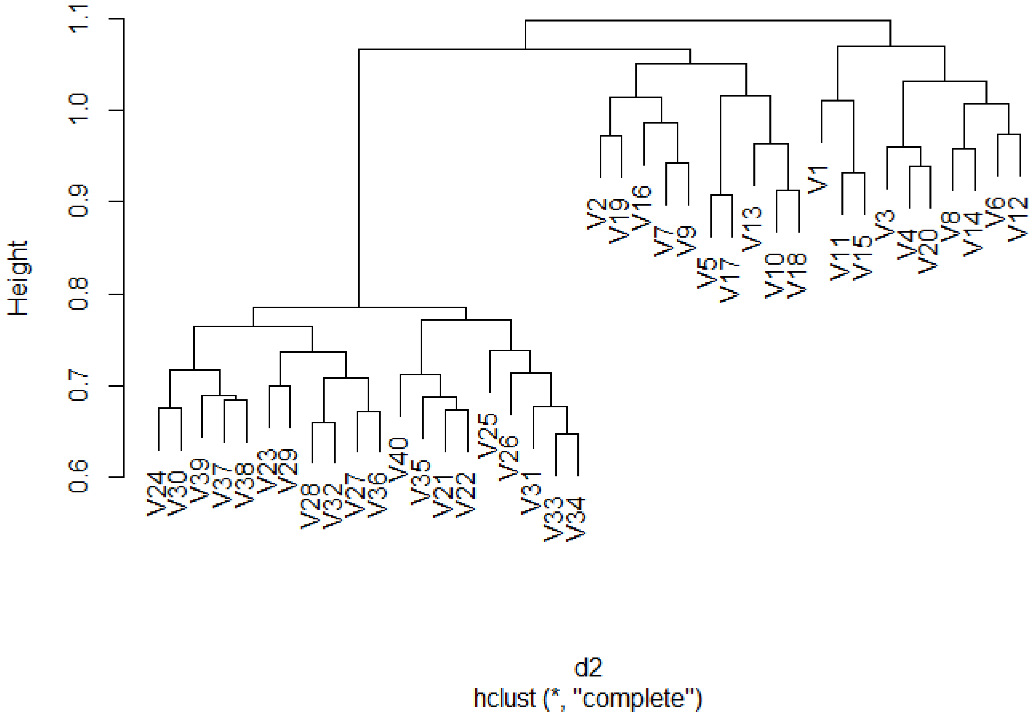
b) Hierarchical Clustering with correlation based distance



Cluster Dendrogram

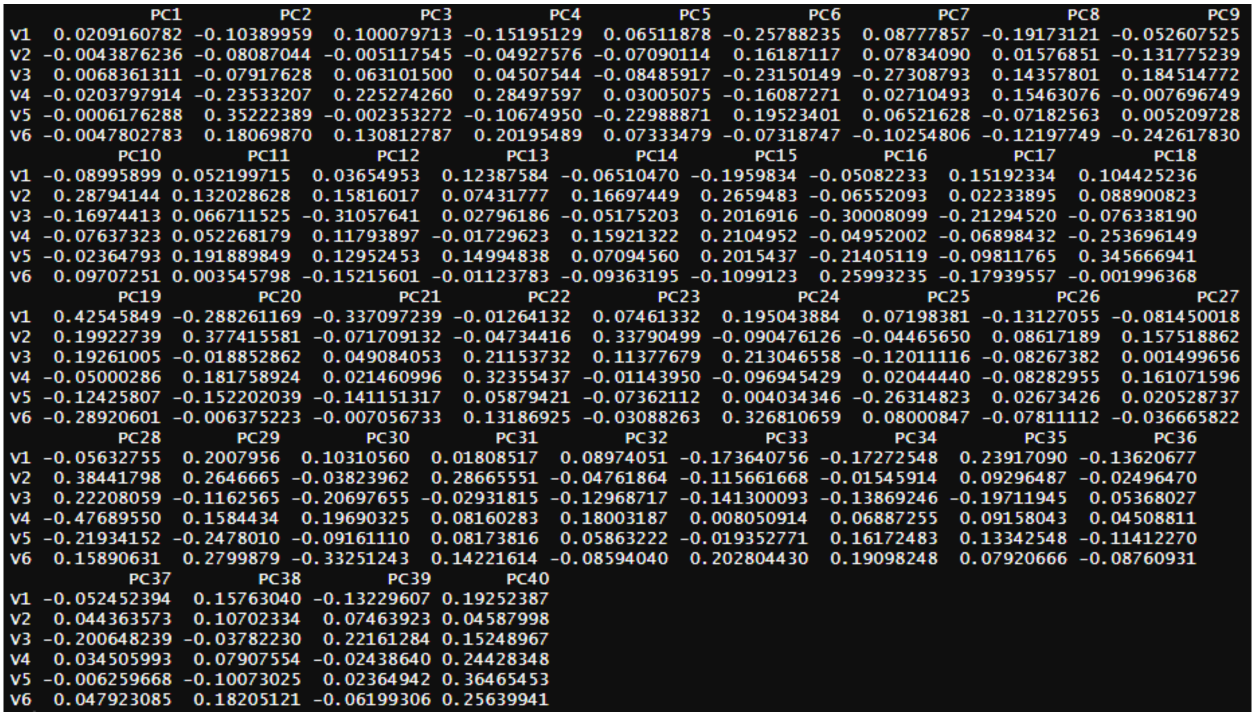


Cluster Dendrogram



From the dendrograms it is evident that different linkages provide different outcomes. The number of clusters for single and complete linkages are indeed 2 whereas for average it is 3.

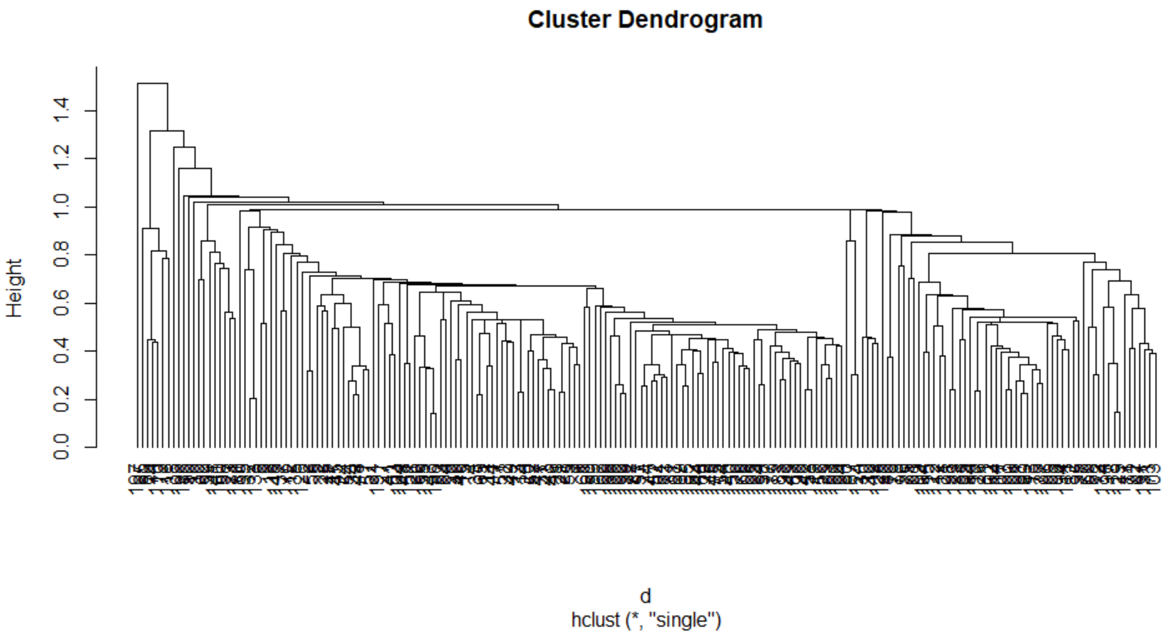
c) To find out the genes that differ the most amongst the two groups, we use PCA with the prcomp() function and scale set as TRUE.
The rotation matrix gives the principal component loadings.

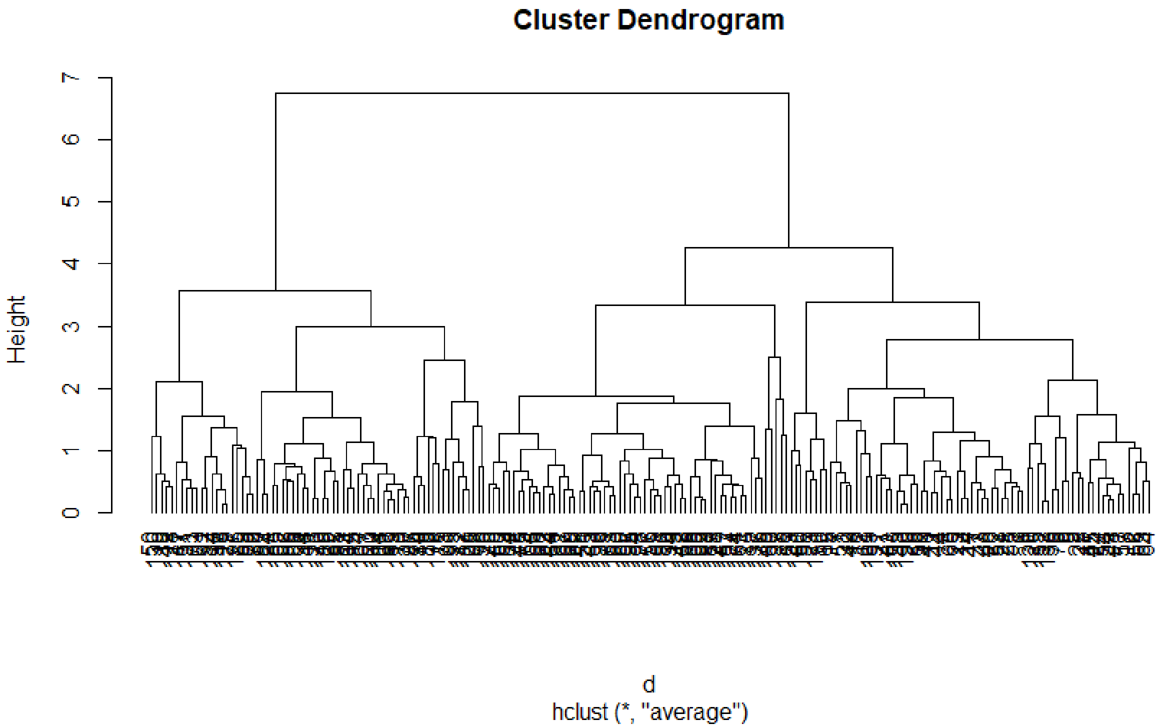


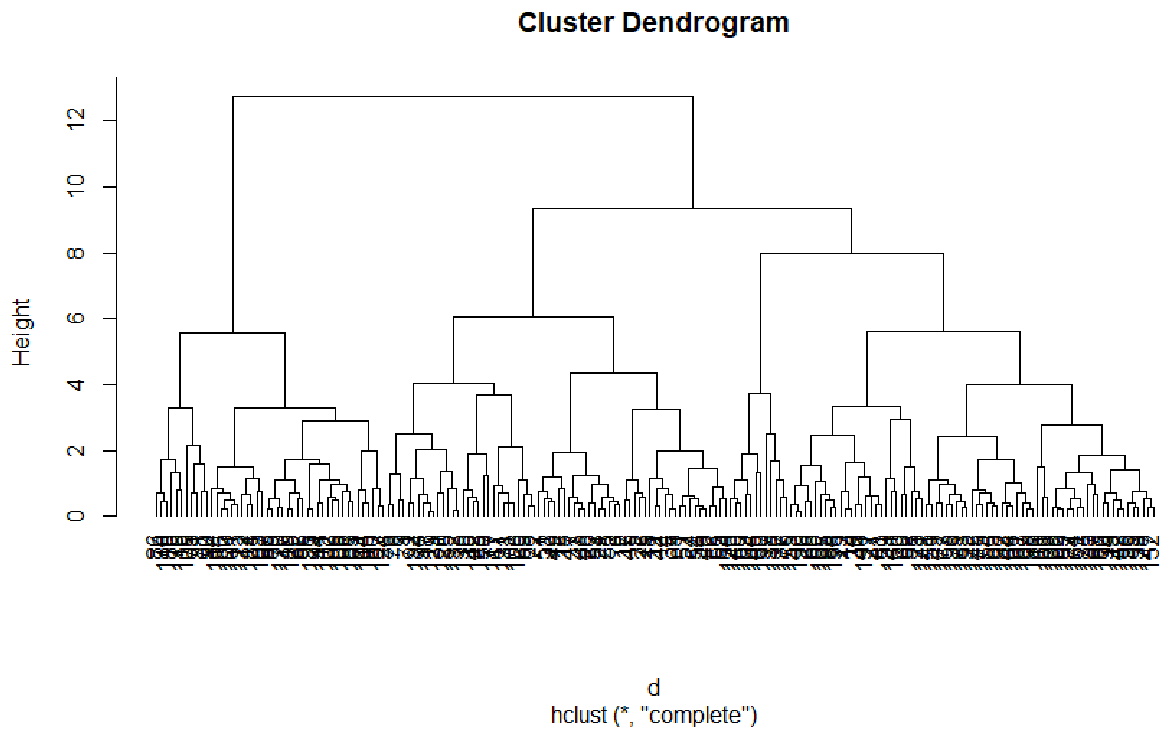
```
             PC1         PC2          PC3         PC4         PC5         PC6         PC7         PC8          PC9
V1  0.0209160782 -0.10389959  0.100079713 -0.15195129  0.06511878 -0.25788235  0.08777857 -0.19173121 -0.052607525
V2 -0.0043876236 -0.08087044 -0.005117545 -0.04927576 -0.07090114  0.16187117  0.07834090  0.01576851 -0.131775239
V3  0.0068361311 -0.07917628  0.063101500  0.04507544 -0.08485917 -0.23150149 -0.27308793  0.14357801  0.184514772
V4 -0.0203797914 -0.23533207  0.225274260  0.28497597  0.03005075 -0.16087271  0.02710493  0.15463076 -0.007696749
V5 -0.0006176288  0.35222389 -0.002353272 -0.10674950 -0.22988871  0.19523401  0.06521628 -0.07182563  0.005209728
V6 -0.0047802783  0.18069870  0.130812787  0.20195489  0.07333479 -0.07318747 -0.10254806 -0.12197749 -0.242617830
          PC10        PC11        PC12        PC13        PC14       PC15        PC16        PC17         PC18
V1 -0.08995899 0.052199715  0.03654953  0.12387584 -0.06510470 -0.1959834 -0.05082233  0.15192334  0.104425236
V2  0.28794144 0.132028628  0.15816017  0.07431777  0.16697449  0.2659483 -0.06552093  0.02233895  0.088900823
V3 -0.16974413 0.066711525 -0.31057641  0.02796186 -0.05175203  0.2016916 -0.30008099 -0.21294520 -0.076338190
V4 -0.07637323 0.052268179  0.11793897 -0.01729623  0.15921322  0.2104952 -0.04952002 -0.06898432 -0.253696149
V5 -0.02364793 0.191889849  0.12952453  0.14994838  0.07094560  0.2015437 -0.21405119 -0.09811765  0.345666941
V6  0.09707251 0.003545798 -0.15215601 -0.01123783 -0.09363195 -0.1099123  0.25993235 -0.17939557 -0.001996368
          PC19         PC20         PC21        PC22        PC23         PC24        PC25        PC26         PC27
V1  0.42545849 -0.288261169 -0.337097239 -0.01264132  0.07461332  0.195043884  0.07198381 -0.13127055 -0.081450018
V2  0.19922739  0.377415581 -0.071709132 -0.04734416  0.33790499 -0.090476126 -0.04465650  0.08617189  0.157518862
V3  0.19261005 -0.018852862  0.049084053  0.21153732  0.11377679  0.213046558 -0.12011116 -0.08267382  0.001499656
V4 -0.05000286  0.181758924  0.021460996  0.32355437 -0.01143950 -0.096945429  0.02044440 -0.08282955  0.161071596
V5 -0.12425807 -0.152202039 -0.141151317  0.05879421 -0.07362112  0.004034346 -0.26314823  0.02673426  0.020528737
V6 -0.28920601 -0.006375223 -0.007056733  0.13186925 -0.03088263  0.326810659  0.08000847 -0.07811112 -0.036665822
          PC28       PC29        PC30        PC31        PC32         PC33        PC34        PC35        PC36
V1 -0.05632755  0.2007956  0.10310560  0.01808517  0.08974051 -0.173640756 -0.17272548  0.23917090 -0.13620677
V2  0.38441798  0.2646665 -0.03823962  0.28665551 -0.04761864 -0.115661668 -0.01545914  0.09296487 -0.02496470
V3  0.22208059 -0.1162565 -0.20697655 -0.02931815 -0.12968717 -0.141300093 -0.13869246 -0.19711945  0.05368027
V4 -0.47689550  0.1584434  0.19690325  0.08160283  0.18003187  0.008050914  0.06887255  0.09158043  0.04508811
V5 -0.21934152 -0.2478010 -0.09161110  0.08173816  0.05863222 -0.019352771  0.16172483  0.13342548 -0.11412270
V6  0.15890631  0.2799879 -0.33251243  0.14221614 -0.08594040  0.202804430  0.19098248  0.07920666 -0.08760931
           PC37        PC38        PC39       PC40
V1 -0.052452394  0.15763040 -0.13229607 0.19252387
V2  0.044363573  0.10702334  0.07463923 0.04587998
V3 -0.200648239 -0.03782230  0.22161284 0.15248967
V4  0.034505993  0.07907554 -0.02438640 0.24428348
V5 -0.006259668 -0.10073025  0.02364942 0.36465453
V6  0.047923085  0.18205121 -0.06199306 0.25639941
```

3)

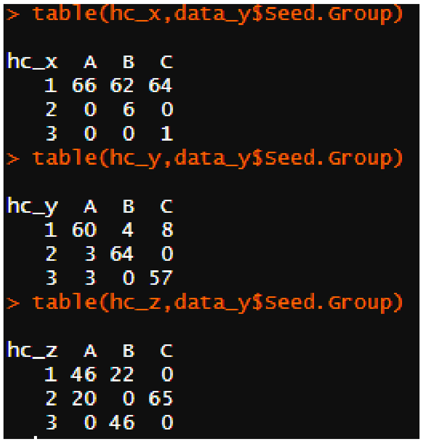
a) Reading data using read.delim(). Removing the seed group from consideration. Applying single linkes, average linked and complete linked hierarchical clustering.


Cluster Dendrogram
Height
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
d
hclust (*, "single")


Cluster Dendrogram
Height
0
1
2
3
4
5
6
7
d
hclust (*, "average")



Cutting the Dendrograms with k =3



```
> table(hc_x,data_y$Seed.Group)

hc_x  A  B  C
   1 66 62 64
   2  0  6  0
   3  0  0  1
> table(hc_y,data_y$Seed.Group)

hc_y  A  B  C
   1 60  4  8
   2  3 64  0
   3  3  0 57
> table(hc_z,data_y$Seed.Group)

hc_z  A  B  C
   1 46 22  0
   2 20  0 65
   3  0 46  0
```