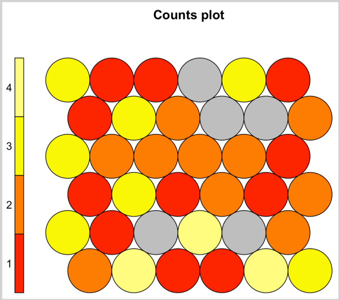
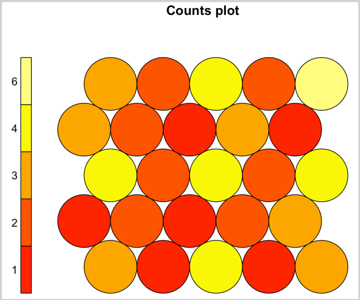
Data Mining II
Homework 3

1) (10 points) Consider the tumor microarray data in the package library(ElemStatLearn).
>library(ElemStatLearn)
>data(nci)
>head(nci)
The data consists of several different types of tumor samples. We observed that it many clustering algorithms there are often found to be 2-3 groups/clusters in this well-studied data, although there are 14 subtypes of tumor cells (unique(colnames(nci))).
Run a SOM algorithm and present the results (e.g., U-matrix, phase plots if appropriate, hclust on prototypes). How well does the SOM method characterize the tumor cells into groups?

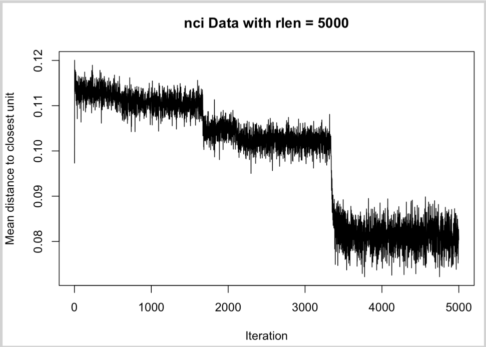
Ans: For the codes plot, because there are large numbers of tumor cells, the default behavior of plotting is to make a line plot and this is not clear as segment plot to identify how SOM method characterize the tumor cells into groups. Also, I found that the dendrograms will be easy to present wired and not fitting well tree, so I tried several different xdim, ydim in somgrid and rlen to find which of the parameter is the best.

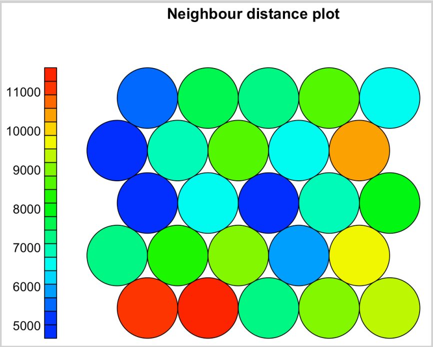
- For xdim = 6, ydim = 6, there are some nodes show null according to the counts plot, so I set xdim and dim to 5.



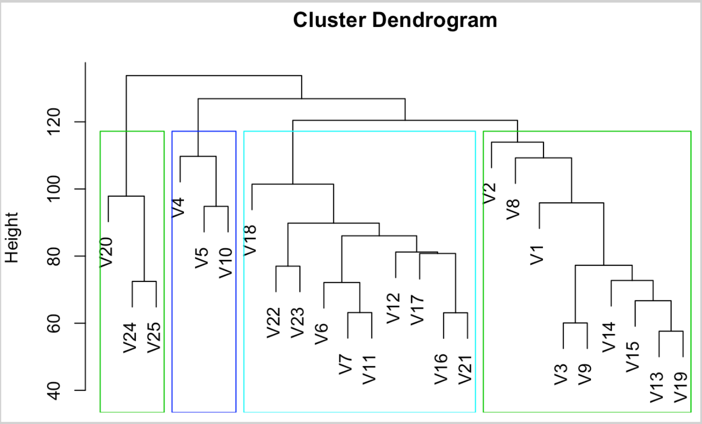


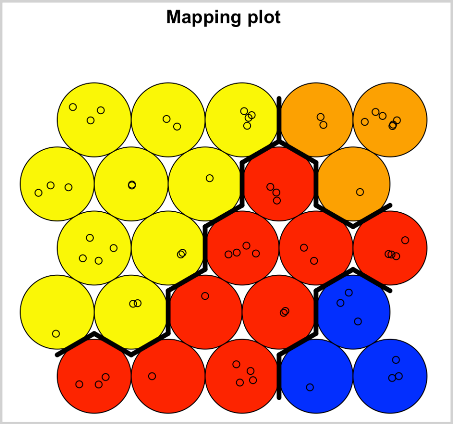
- When rlen = 5000, the phase plot presents that distance will reach to a minimum.
- Neighbour distance plot (often referred to as the U-Matrix) shows that the distance of V1, V2 is quite dissimilar with the surrounding nodes, and V20 is also dissimilar with the adjacent nodes. These two groups of nodes are perhaps to be divided into different clusters with others nodes.



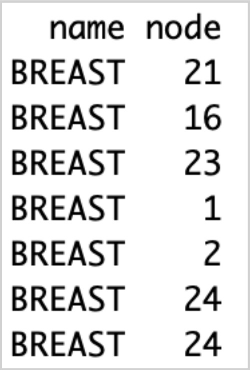


- In the dendrogram of hclust, although the nodes of the tree are mostly on the right, we can still divide them into four clusters.
- In hclust on prototypes, V20 and V25, V26 are classified as a cluster, but V1 and V2 are grouping with adjacent nodes.





- Almost all the same type of tumor cells are classified as the same group, this is because even if the same type of tumor cell, the data difference between each data is still very large, so it will be divided into different nodes



```
  name  node
BREAST    21
BREAST    16
BREAST    23
BREAST     1
BREAST     2
BREAST    24
BREAST    24
```

1. Cluster 1 :
   MCF7A-repro BREAST MCF7D-repro BREAST COLON COLON COLON COLON NSCLC COLON COLON PROSTATE OVARIAN OVARIAN OVARIAN NSCLC NSCLC LEUKEMIA LEUKEMIA NSCLC NSCLC NSCLC MELANOMA PROSTATE
2. Cluster 2 :
   COLON LEUKEMIA LEUKEMIA LEUKEMIA K562B-repro K562A-repro LEUKEMIA
3. Cluster 3 :
   RENAL RENAL RENAL RENAL RENAL RENAL RENAL OVARIAN NSCLC CNS CNS BREAST NSCLC NSCLC OVARIAN RENAL BREAST CNS CNS CNS BREAST RENAL UNKNOWN OVARIAN
4. Cluster 4 :
   MELANOMA BREAST BREAST MELANOMA MELANOMA MELANOMA MELANOMA MELANOMA MELANOMA

2) (10 points) Consider the USArrests data. I suggest scaling the data first.

```
> library(ISLR)
> data(USArrests)
> head(USArrests)
```

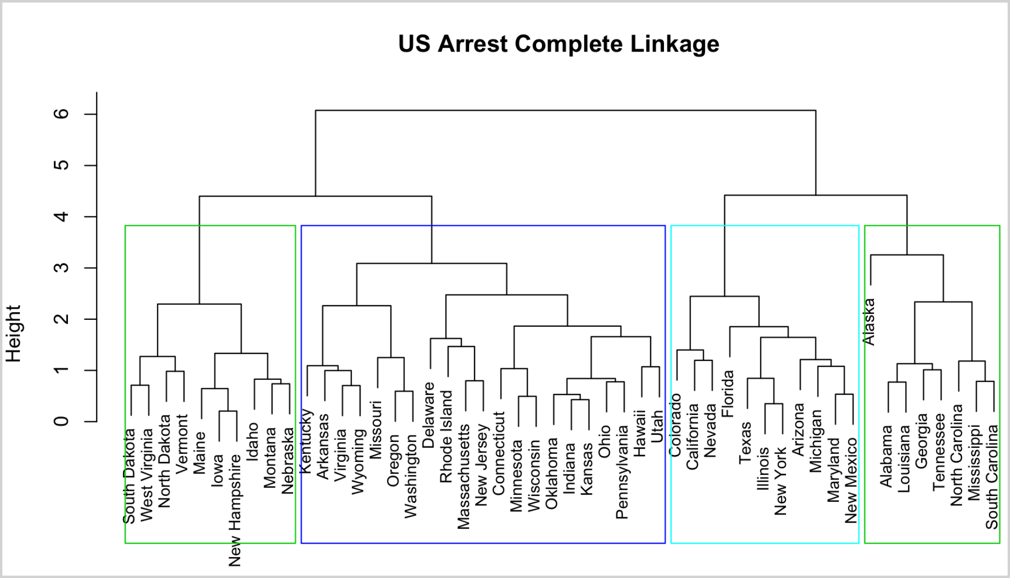
a) Perform hierarchical clustering with complete linkage and Euclidean distance to cluster the states. Cut the dendrogram at a height that results in three clusters. Is this what you would expect?

b) Fit a SOM to the data and present the results (e.g., U-matrix, phase plots if appropriate, hclust on prototypes). Is this what you would expect? Does this result generally support your results in Part A.

c) Comment on the advantages and limitations of hierarchical clustering to SOM, and discuss when one would be preferred over the other.

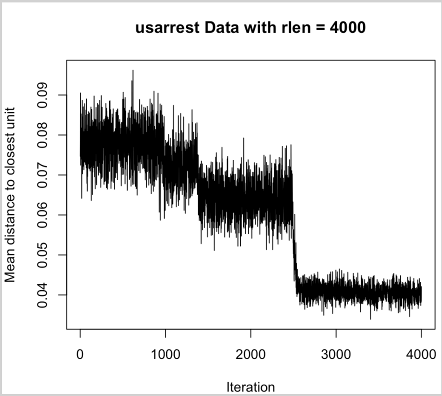
Ans:

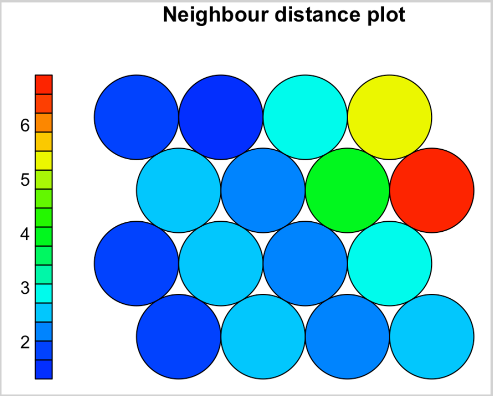
a. For the dendrogram with complete linkage, this tree has good separation with broad shoulders and long branches. But it is better to classify into four clusters rather than three groups expected.



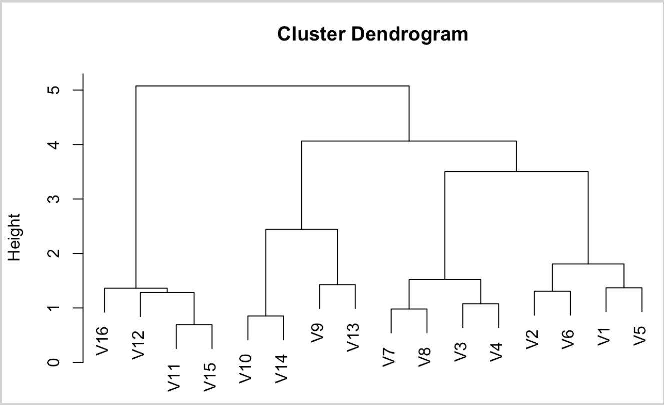
- Cluster 1:
  Alabama Alaska Georgia Louisiana Mississippi North Carolina South Carolina Tennessee
- Cluster 2:
  Arizona California Colorado Florida Illinois Maryland Michigan Nevada New Mexico New York Texas
- Cluster 3:
  Arkansas Connecticut Delaware Hawaii Indiana Kansas Kentucky Massachusetts Minnesota Missouri New Jersey Ohio Oklahoma Oregon Pennsylvania Rhode Island Utah Virginia Washington Wisconsin Wyoming
- Cluster 4:
  Idaho Iowa Maine Montana Nebraska New Hampshire North Dakota South Dakota Vermont West Virginia

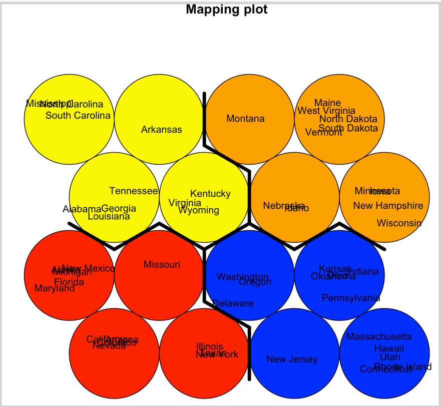
b. The following is the phase plot and neighbor distance plot, I use xdim and ydim= 4 because the higher dim will present null nodes. The phase plot shows that if I set rlen to 4000, this distance will reach to a minimum, and the neighbor distance plot shows that the three nodes on the upper right have quite large distance differences from other nodes, which can be estimated that it will become a cluster.





We can observe that this tree can be cut into four parts in the dendrogram. In hclust on prototypes, we can see that the upper right node is divided into a cluster as estimated, and the other nodes are equally divided into four parts.





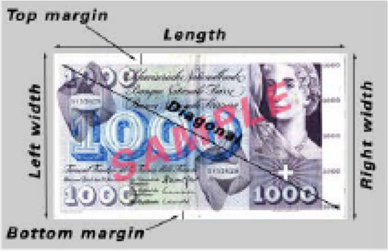
Both hierarchical clustering and SOM methods can be found to be 4 groups as expected, SOM mostly supports hierarchical clustering in this case, yet we can still find that some dots (cities) are classified in different clusters in SOM and hierarchical clustering. For example, Washington, Wisconsin, and Wyoming are classified as one cluster in hierarchical clustering, but Wisconsin and Wyoming are classified in another cluster different from Washington in SOM.

- Cluster 1:
  Arizona California Colorado Nevada Illinois New York Texas Alaska Florida Maryland Michigan New Mexico Missouri
- Cluster 2:
  New JerseyConnecticut Hawaii Massachusetts Rhode Island Utah Delaware Oregon Washington Indiana Kansas Ohio Oklahoma Pennsylvania
- Cluster 3 :
  Alabama Georgia Louisiana Tennessee Kentucky Virginia WyomingMississippi North Carolina South Carolina Arkansas
- Cluster 4 :
  Idaho Nebraska Iowa Minnesota New Hampshire Wisconsin Montana Maine North Dakota South Dakota Vermont West Virginia

c. I think hierarchical clustering is more convenient than SOM in this case, we can review the results of clustering and the nodes included in each cluster more quickly and easily.

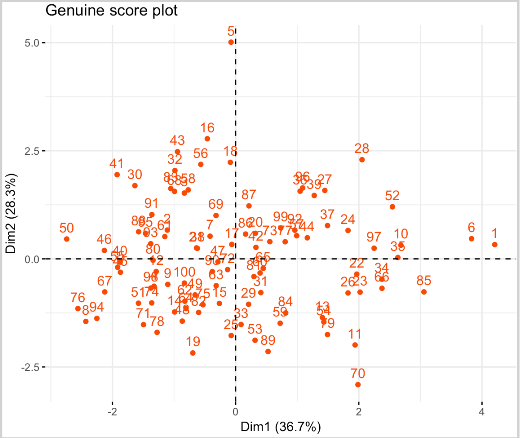
However, if the dataset is large or high-dimensional, it is difficult to use hierarchical clustering for processing and presenting, we should use SOM instead of hierarchical clustering in such cases. On the contrary, if the data is 2D and there are not large number of nodes, hierarchical clustering will be the better choice.

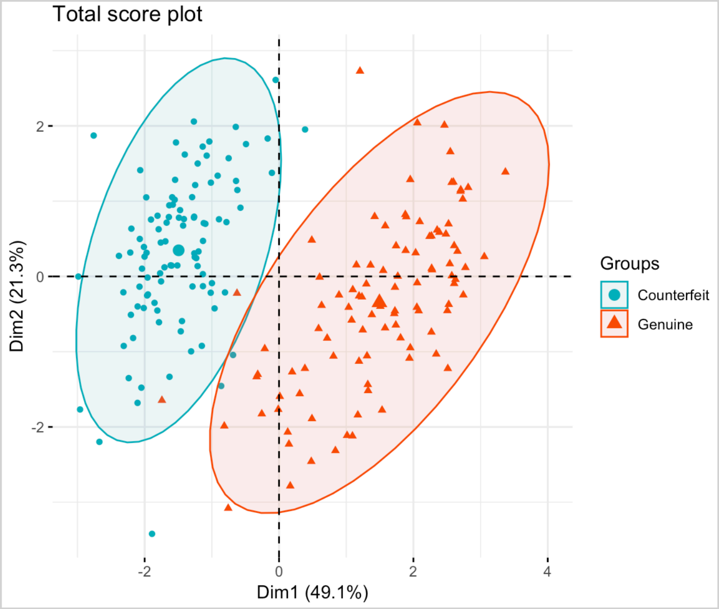
3) (10 points) Access the SwissBankNotes data (posted with assignment). The data consists of six variables measured on 200 old Swiss 1,000-franc bank notes. The first 100 are genuine and the second 100 are counterfeit. The six variables are length of the bank note, height of the bank note, measured on the left, height of the bank note measured on the right, distance of the inner frame to the lower border, distance of inner frame to upper border, and length of the diagonal.

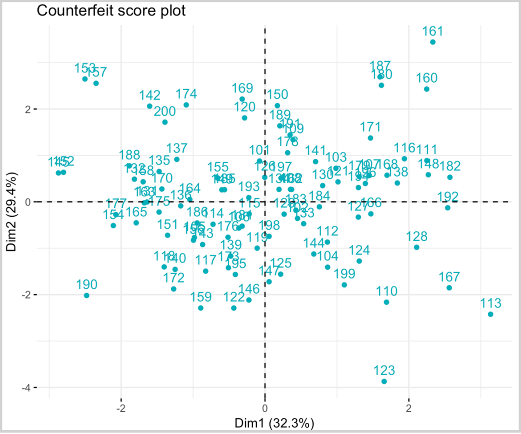


Carry out a PCA of the 100 genuine bank notes, of the 100 counterfeit bank notes, and all of the 200 bank notes combined. Generate some score plots (use colors for the combined). Do you notice any differences in the results? Show your work, and justify the selection of Principal Components, including diagnostic plots.

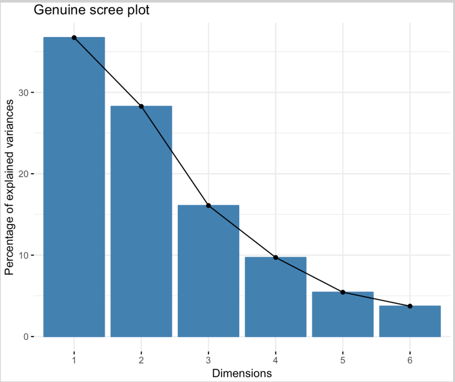
Ans: In the score plot, it is difficult to see the difference from the score plots of genuine and counterfeit bank notes due to scale and center, but in all of the 200 bank notes combined we can find that the score of genuine bank notes in PC1 is higher than counterfeit bank notes.



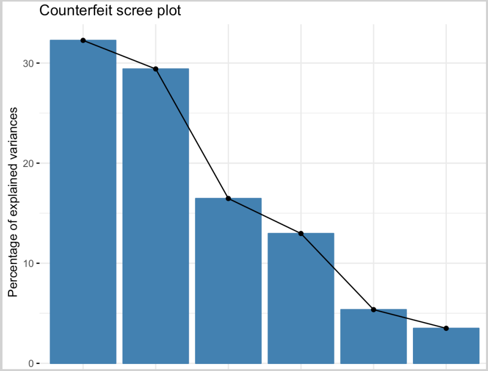




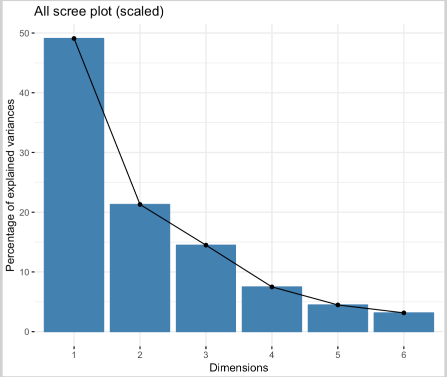
Before starting PCA, I had scaled the data because all elbow will be created more equally. About the selection of Principal Components, in 100 genuine bank notes, 100 counterfeit bank notes, and all of the 200 bank notes combined, we can find that there all have a big jump between PC2 and PC3, so I retained first two component in these three case.



```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.4845 1.3026 0.9827 0.76348 0.57156 0.47340
Proportion of Variance 0.3673 0.2828 0.1610 0.09715 0.05445 0.03735
Cumulative Proportion  0.3673 0.6501 0.8111 0.90820 0.96265 1.00000
```

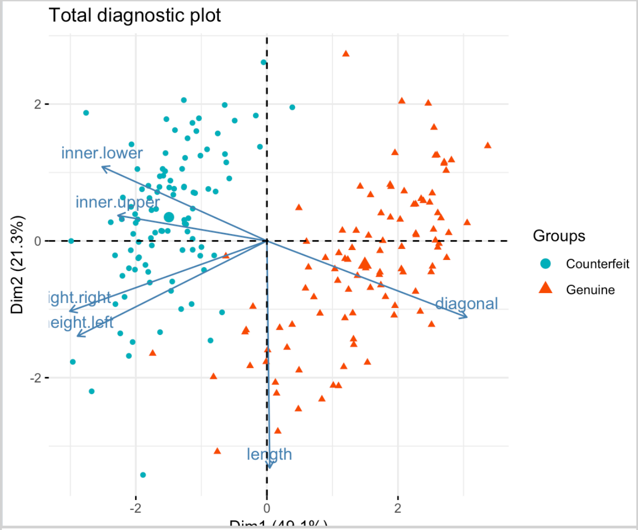


```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.3915 1.3285 0.9941 0.8823 0.56755 0.45840
Proportion of Variance 0.3227 0.2941 0.1647 0.1297 0.05368 0.03502
Cumulative Proportion  0.3227 0.6169 0.7816 0.9113 0.96498 1.00000
```



```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.7163 1.1305 0.9322 0.67065 0.51834 0.43460
Proportion of Variance 0.4909 0.2130 0.1448 0.07496 0.04478 0.03148
Cumulative Proportion  0.4909 0.7039 0.8488 0.92374 0.96852 1.00000
```

The following is the biplots with combined bank notes, we can know that diagonal is the most important variance for the PC1, and the inner.lower is the most important variance for the PC2. We can also recognize that the genuine bank notes’ diagonal and length will be higher than counterfeit bank notes, and counterfeit bank notes' inner.lower and inner.upper will be significantly higher than genuine bank notes.



```
                    PC1         PC2         PC3        PC4        PC5         PC6
length       0.006987029 -0.81549497  0.01768066  0.5746173 -0.0587961  0.03105698
height.left -0.467758161 -0.34196711 -0.10338286 -0.3949225  0.6394961 -0.29774768
height.right -0.486678705 -0.25245860 -0.12347472 -0.4302783 -0.6140972  0.34915294
inner.lower -0.406758327  0.26622878 -0.58353831  0.4036735 -0.2154756 -0.46235361
inner.upper -0.367891118  0.09148667  0.78757147  0.1102267 -0.2198494 -0.41896754
diagonal     0.493458317 -0.27394074 -0.11387536 -0.3919305 -0.3401601 -0.63179849
```