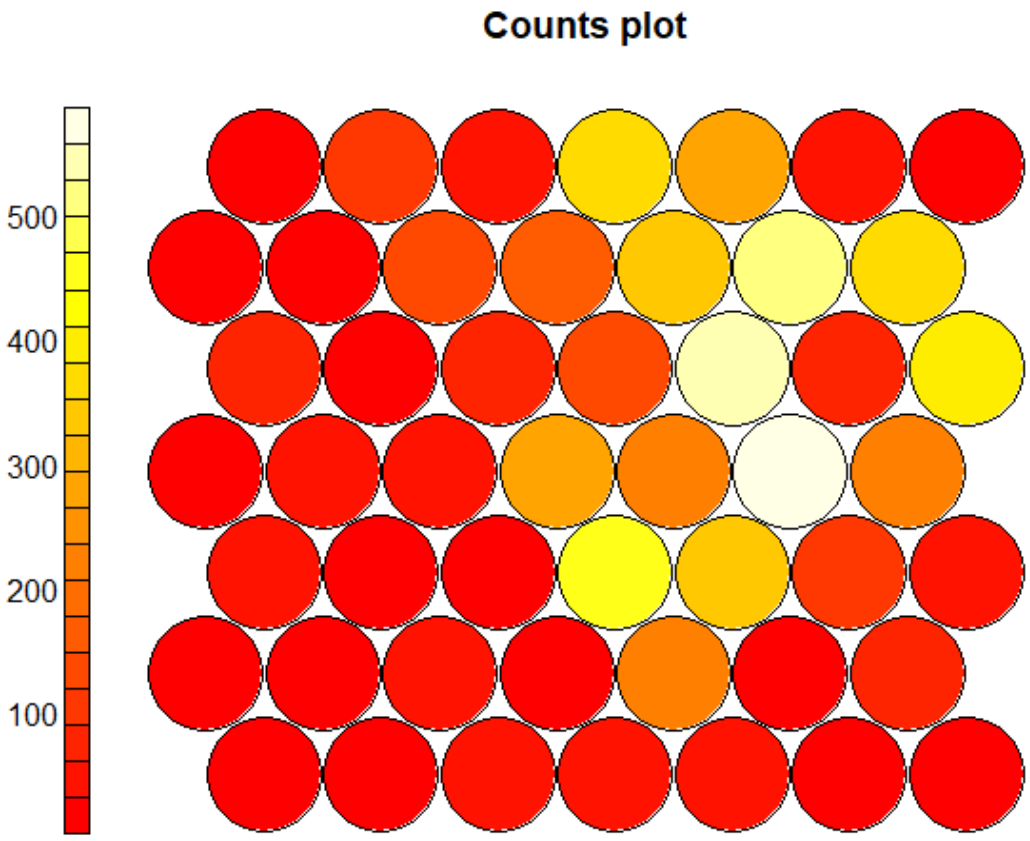
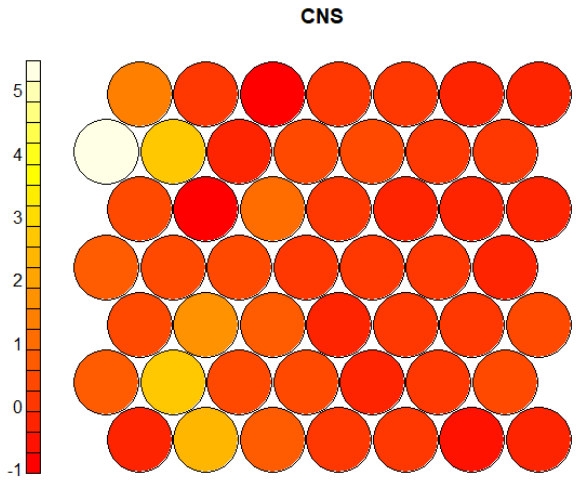
1)

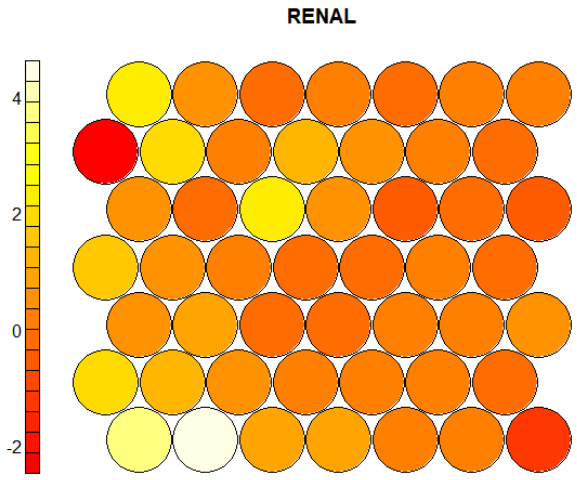
Running a SOM algorithm on the nci data with a 7x7 hexagonal SOM grid.

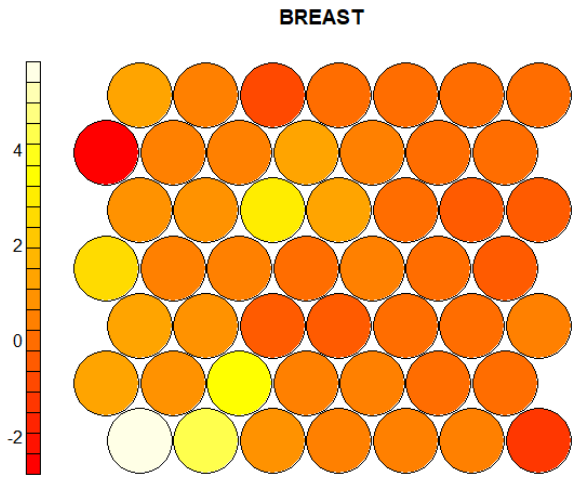


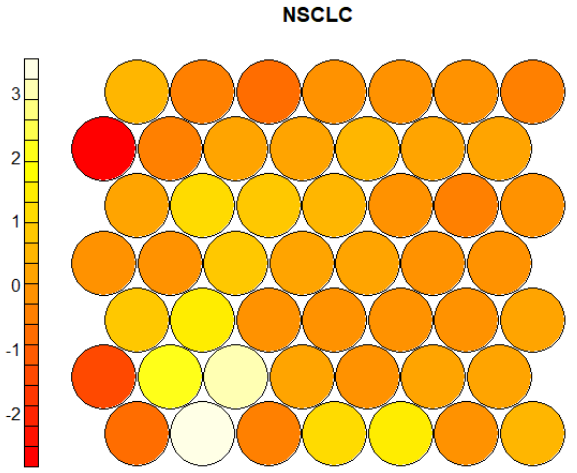

The counts plot indicates that majority of the nodes have a count between 100-300 with the max being around 600.

4 out of the 64 phase plots –



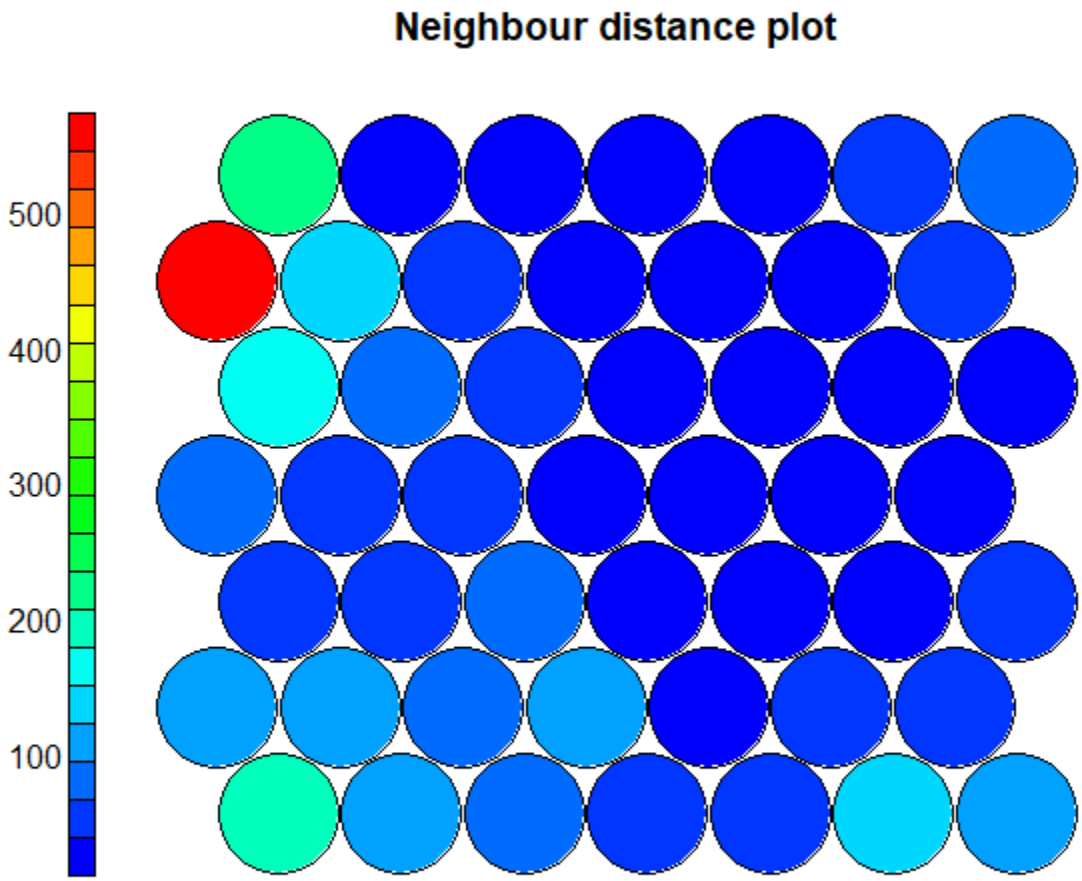



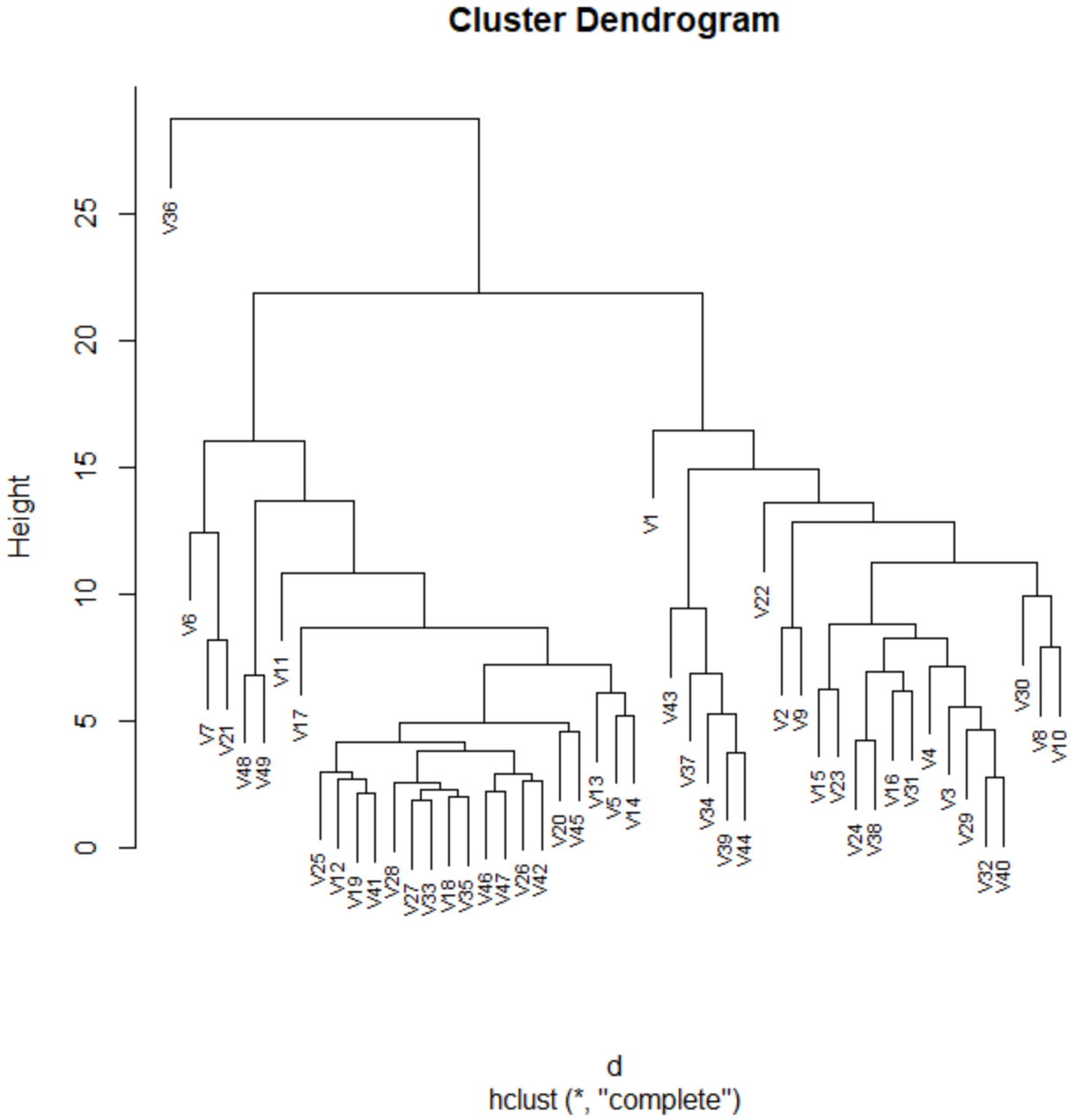




The phase plots for this particular data do not provide significant insight.

U Matrix -




The U matrix shows one particular node which is significantly farther from the rest. This can be explained by looking at V36 in the dendrogram below.

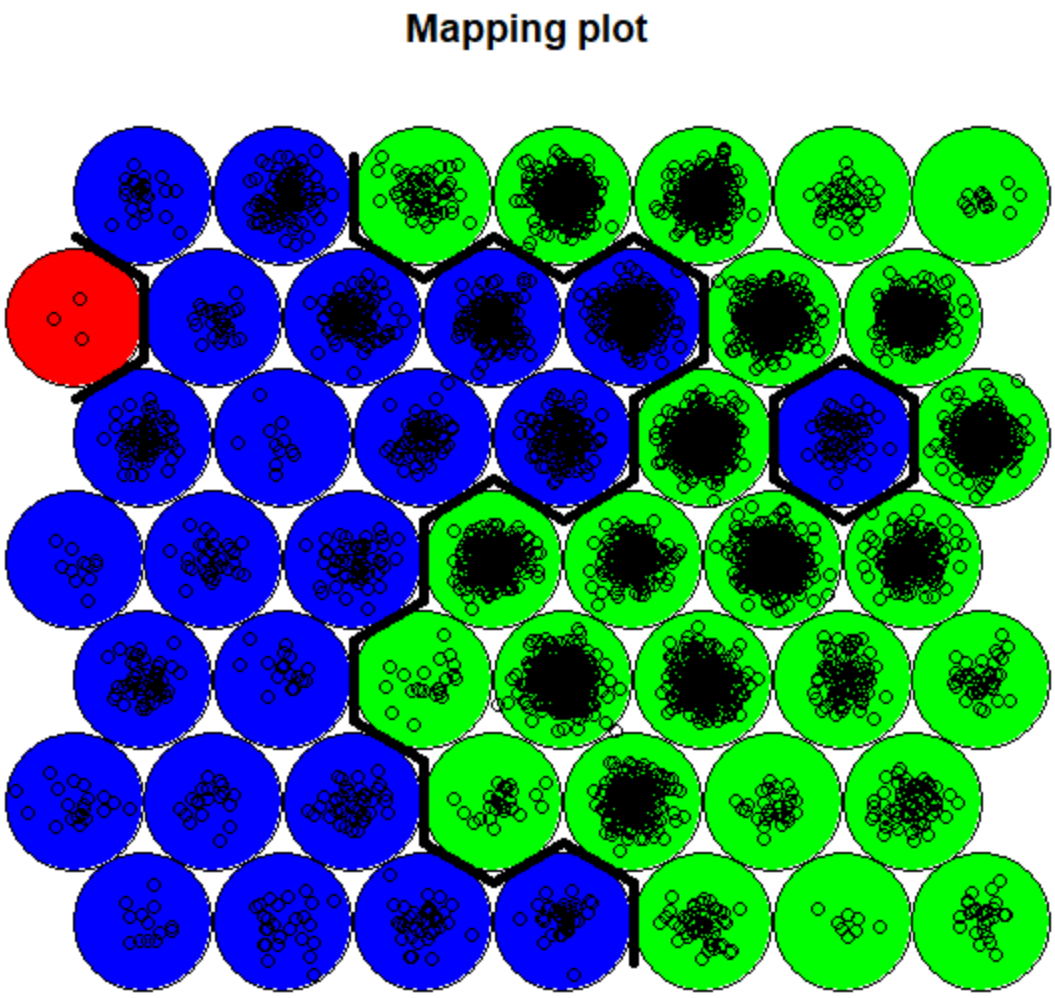



From observing the Dendrogram, cutting it into 3 clusters seems appropriate.

```
> names(cluster_1)
 [1] "V1"  "V2"  "V3"  "V4"  "V8"  "V9"  "V10" "V15" "V16"
[10] "V22" "V23" "V24" "V29" "V30" "V31" "V32" "V34" "V37"
[19] "V38" "V39" "V40" "V43" "V44"
```

```
> names(cluster_2)
 [1] "V5"  "V6"  "V7"  "V11" "V12" "V13" "V14" "V17" "V18"
[10] "V19" "V20" "V21" "V25" "V26" "V27" "V28" "V33" "V35"
[19] "V41" "V42" "V45" "V46" "V47" "V48" "V49"
```

```
> names(cluster_3)
[1] "V36"
```




The mapping plot clearly shows the 3 different clusters obtained.

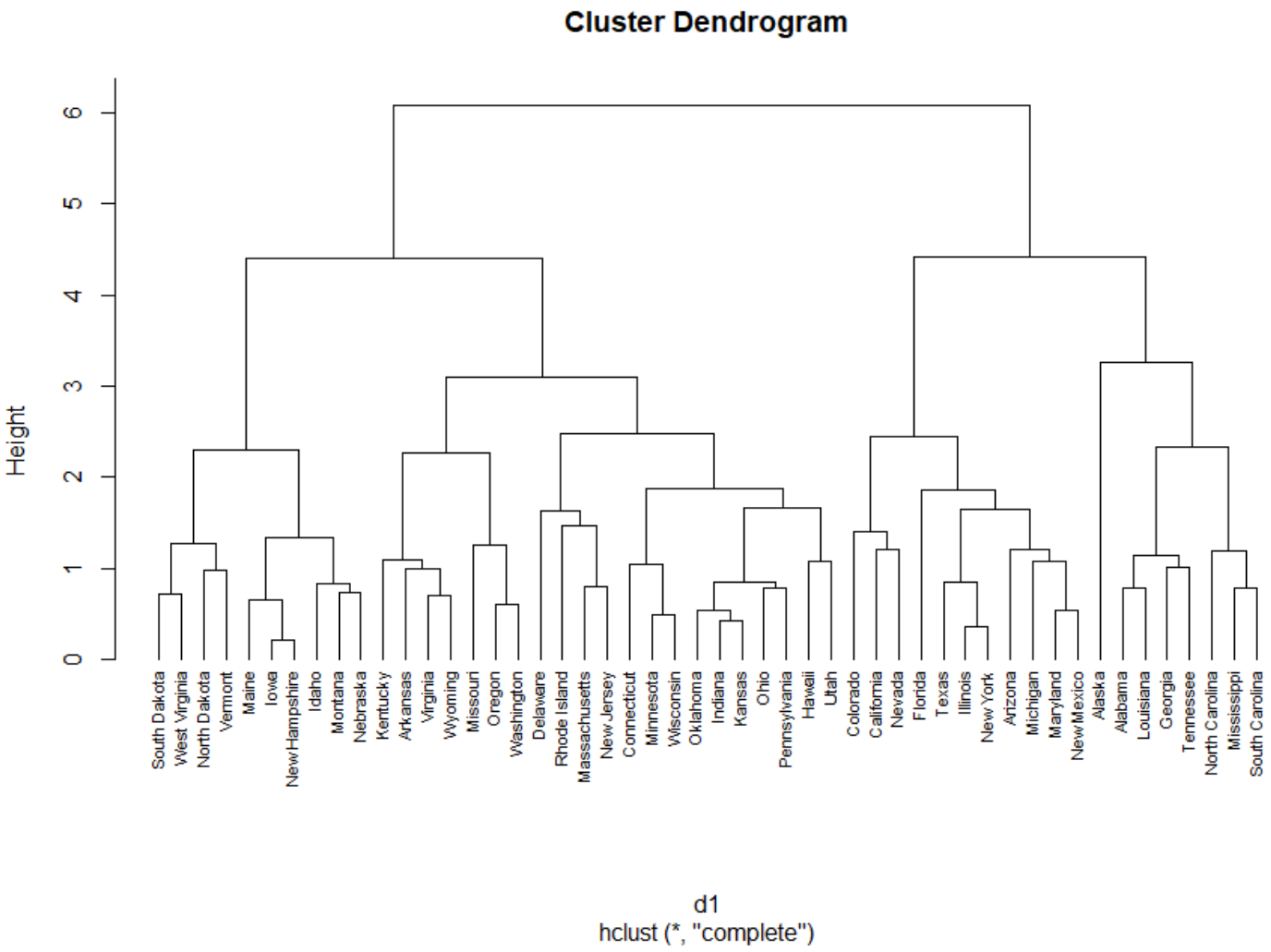
By looking at the results in the different clusters, it is evident that the SOM method characterizes the tumor cells appropriately into 3 groups.

| | CNS | CNS | CNS | RENAL | BREAST | CNS |
|---|---|---|---|---|---|---|
| V35 | -0.19877427 | -0.0674917774 | -0.218272032 | -0.38311598 | -0.540486898 | -0.327138194 |
| V36 | 5.45320459 | 4.9636692638 | -0.821532720 | -2.47479186 | -2.815758923 | -1.496944409 |
| V37 | 2.74428906 | 2.8626805103 | 0.591363724 | 1.93220720 | 0.224603605 | 0.396808678 |

We can slightly perceive why V35, V36 and V37 belong to different clusters.

2)

a) Performing Hierarchical Clustering on the scaled USArrests data with complete linkage and Euclidean linkage.




Cutting the Dendrogram into 3 parts using the cutree function results in clusters with 8, 11 and 31 states.

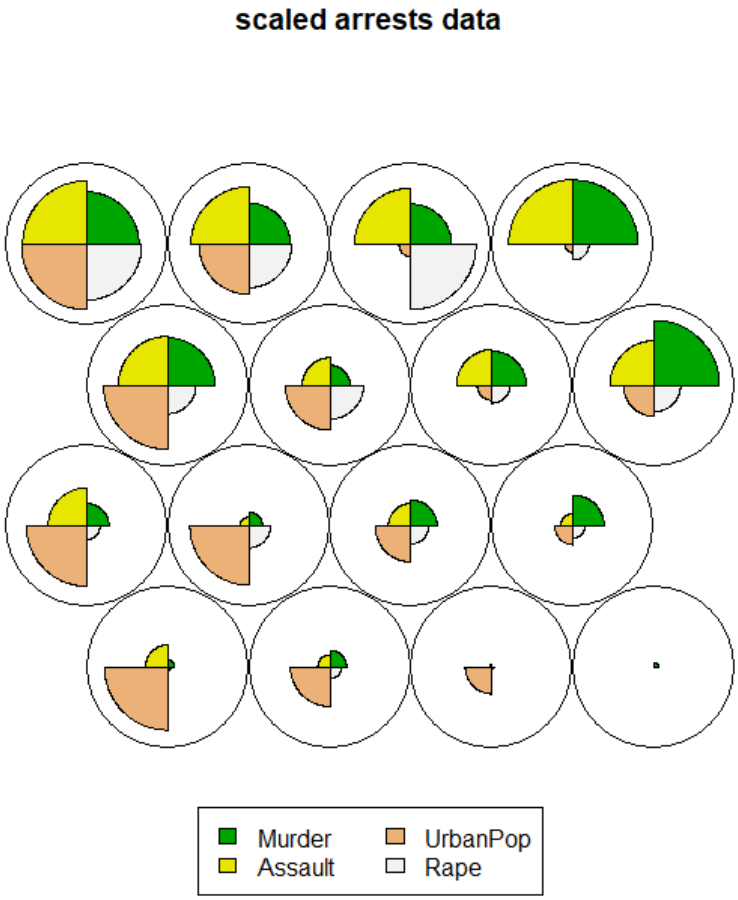
```
> names(cluster_1)
[1] "Alabama"        "Alaska"         "Georgia"        "Louisiana"      "Mississippi"    "North Carolina"
[7] "South Carolina" "Tennessee"
```

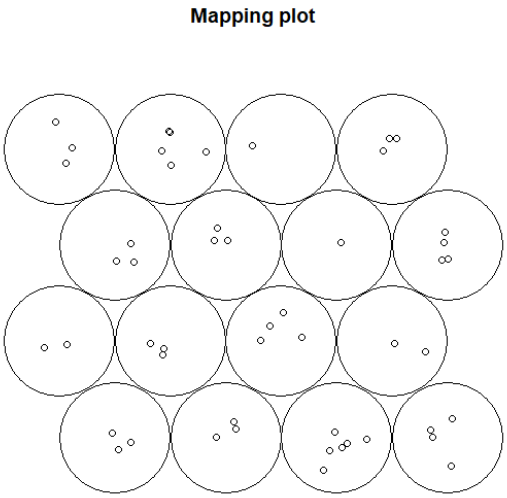
```
> names(cluster_2)
 [1] "Arizona"    "California" "Colorado"   "Florida"    "Illinois"   "Maryland"   "Michigan"   "Nevada"     "New Mexico"
[10] "New York"   "Texas"
```

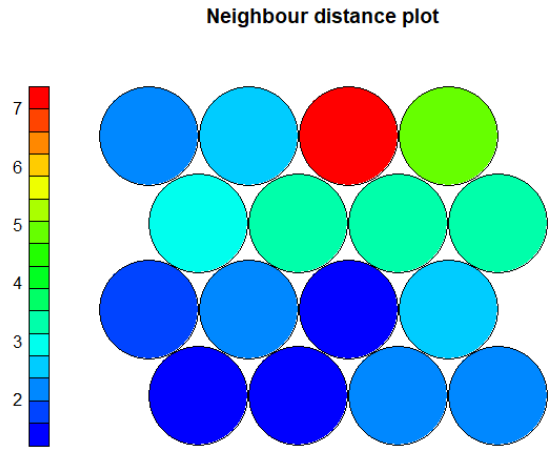
```
> names(cluster_3)
 [1] "Arkansas"      "Connecticut"   "Delaware"      "Hawaii"        "Idaho"         "Indiana"       "Iowa"
 [8] "Kansas"        "Kentucky"      "Maine"         "Massachusetts" "Minnesota"     "Missouri"      "Montana"
[15] "Nebraska"      "New Hampshire" "New Jersey"    "North Dakota"  "Ohio"          "Oklahoma"      "Oregon"
[22] "Pennsylvania"  "Rhode Island"  "South Dakota"  "Utah"          "Vermont"       "Virginia"      "Washington"
[29] "West Virginia" "Wisconsin"     "Wyoming"
```

The scaled data fits as expected within these clusters.

b) Fitting the data on a 4x4 hexagonal grid SOM.

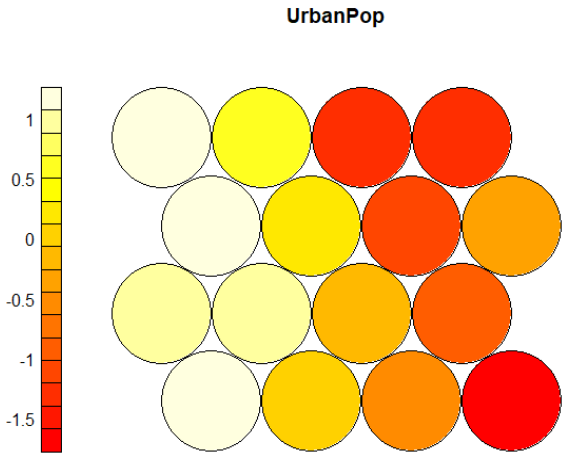


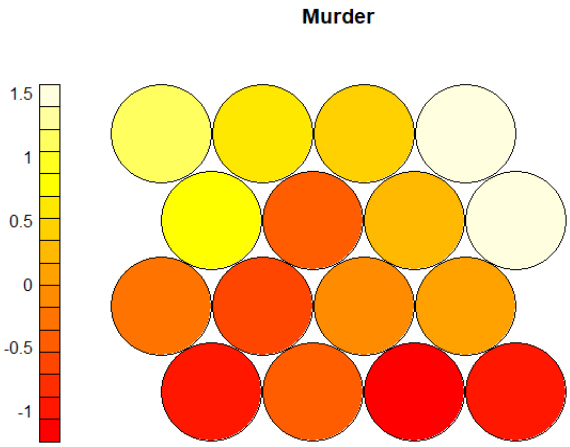



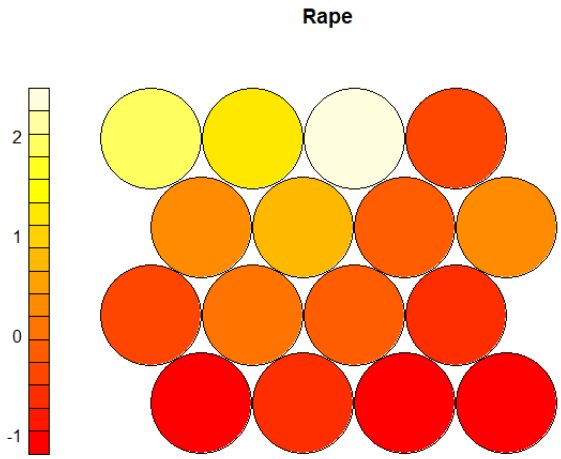



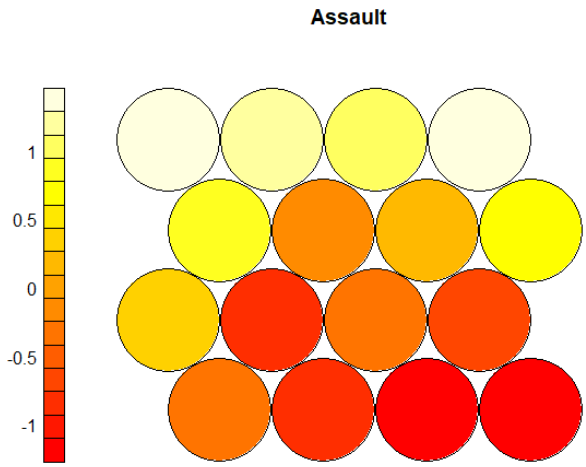

The mapping plot indicates that each node has at least 1 state with a maximum of 7 states. It can also be Noted that the nodes in the lower left are significantly closer to each other.

Phase Plots –

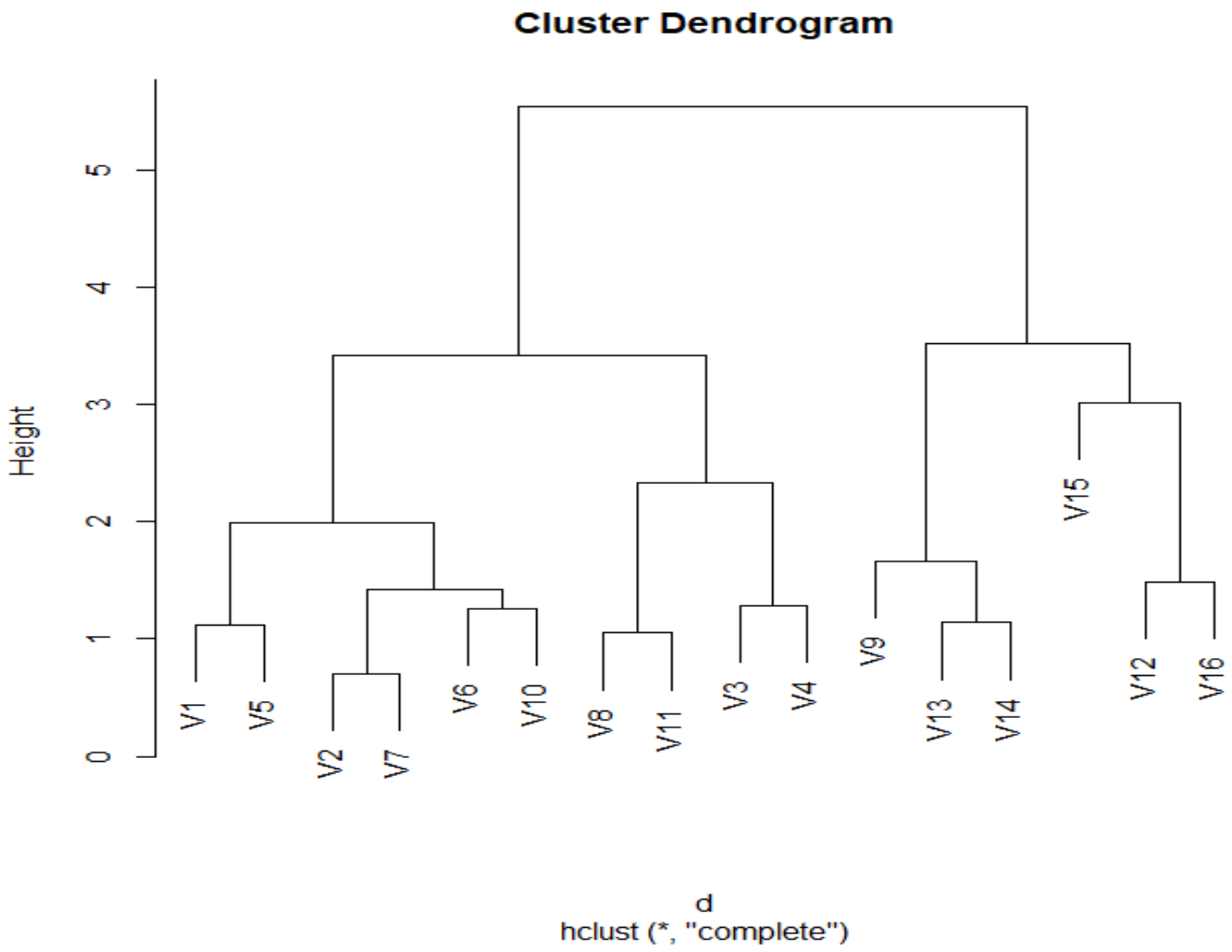


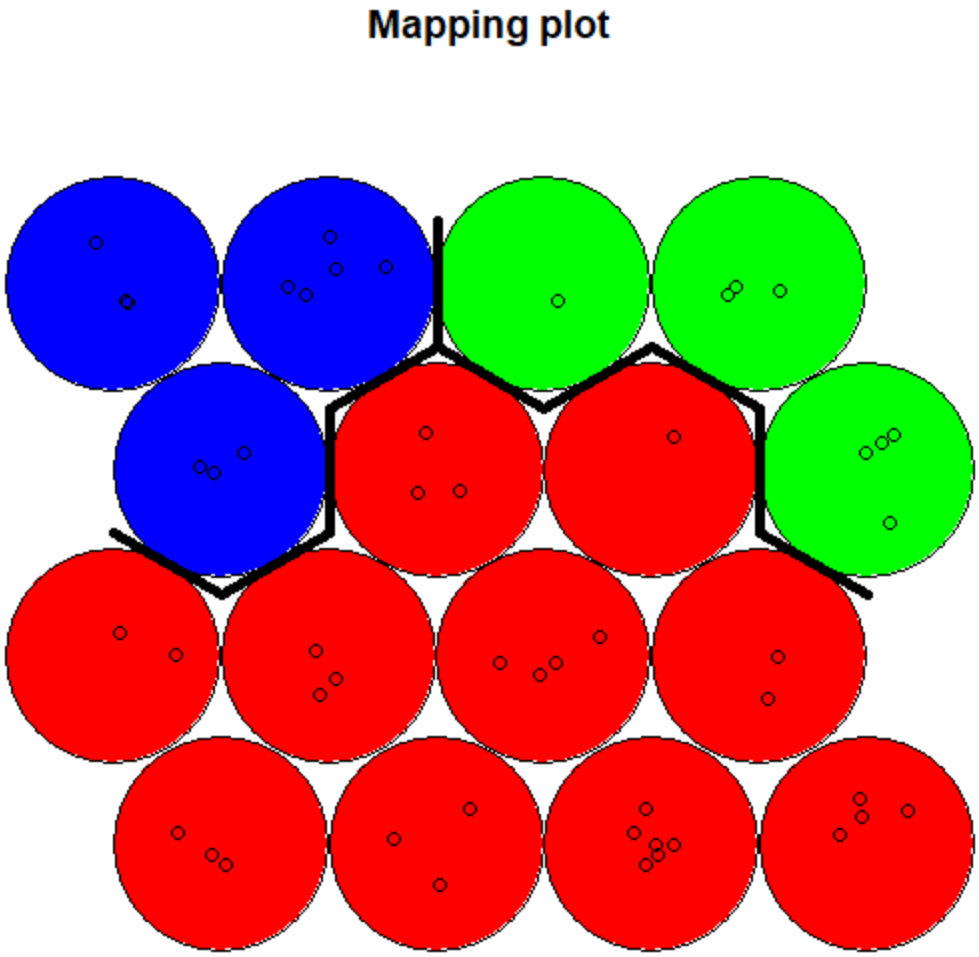




The phase plots in this case gives us an insight into the nature of the data. It can be seen that nodes in the Assault phase plot are somewhat following a similar trend to those in the Murder phase plot.




The cluster dendrogram obtained is similar in structure to dendogram from Part A and it can be seen that the two dendrograms support each other.

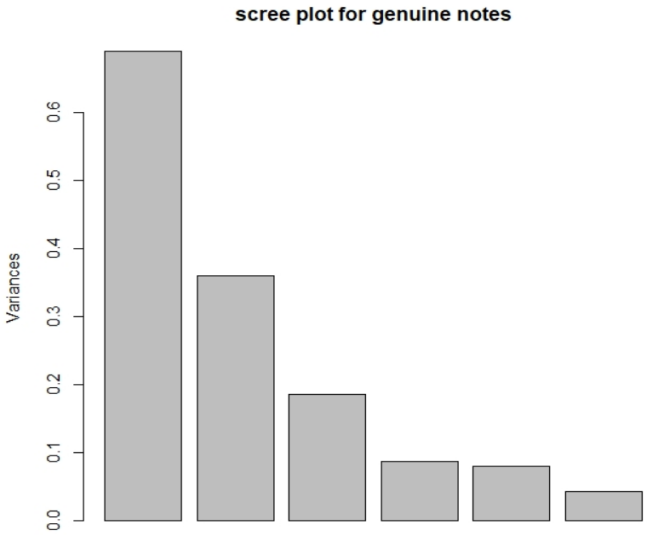



The results obtained using this method support the results obtained in Part A. The mapping plot shows the clear separation between the different group of states with 31, 11 and 8 states in red, blue and green respectively.

c) Hierarchical clustering and SOM are significantly different methods each with their own strengths and limitations. SOM is primarily meant for mapping and visualizing high-dimensional vectors into a 2D space while preserving the topology of the data. Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It is generally not suitable for very large data sets. Furthermore, the main output of hierarchical clustering, the dendrogram is also easy to misinterpret.

3)

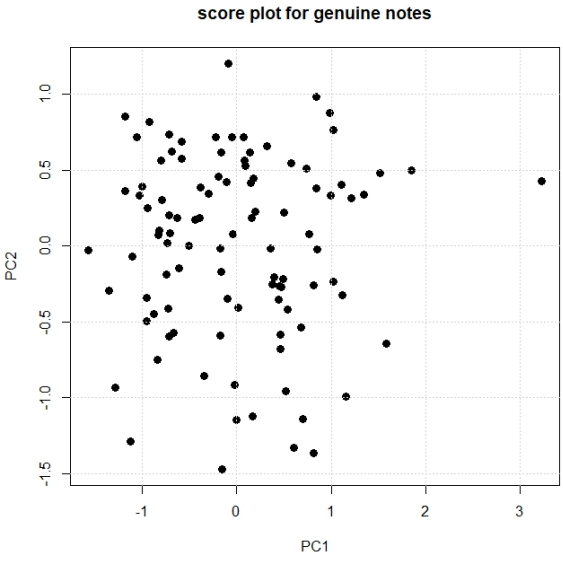
- Applying PCA on the genuine notes data.



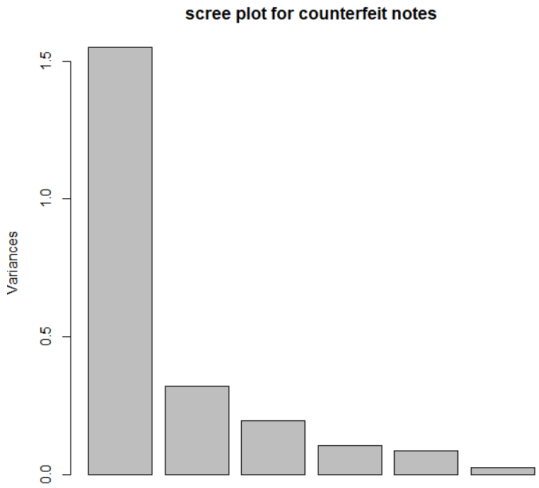

The above plot represents the variance of each principal component in numerical order.

```
> summary(pc1)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     0.8301 0.5994 0.4308 0.29536 0.28317 0.20478
Proportion of Variance 0.4774 0.2489 0.1286 0.06044 0.05556 0.02905
Cumulative Proportion  0.4774 0.7264 0.8549 0.91539 0.97095 1.00000
```

The summary of the principal components of the genuine notes data shows that 47.74% of the variance is explained by the first PC and 24.89% is explained by the second PC.



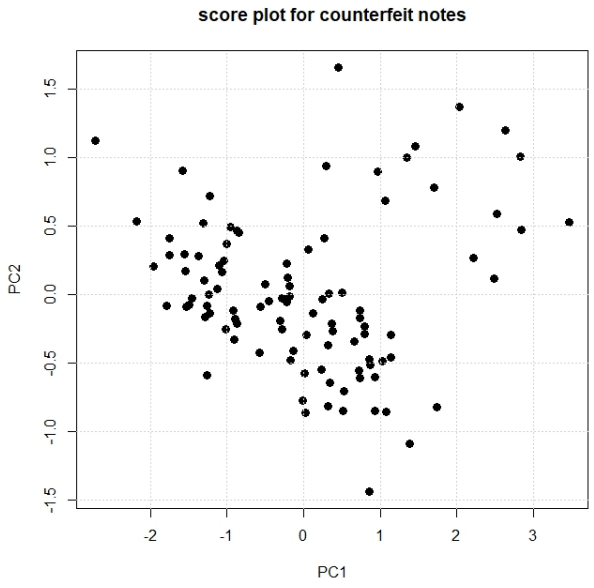
- Applying PCA on the counterfeit notes data.



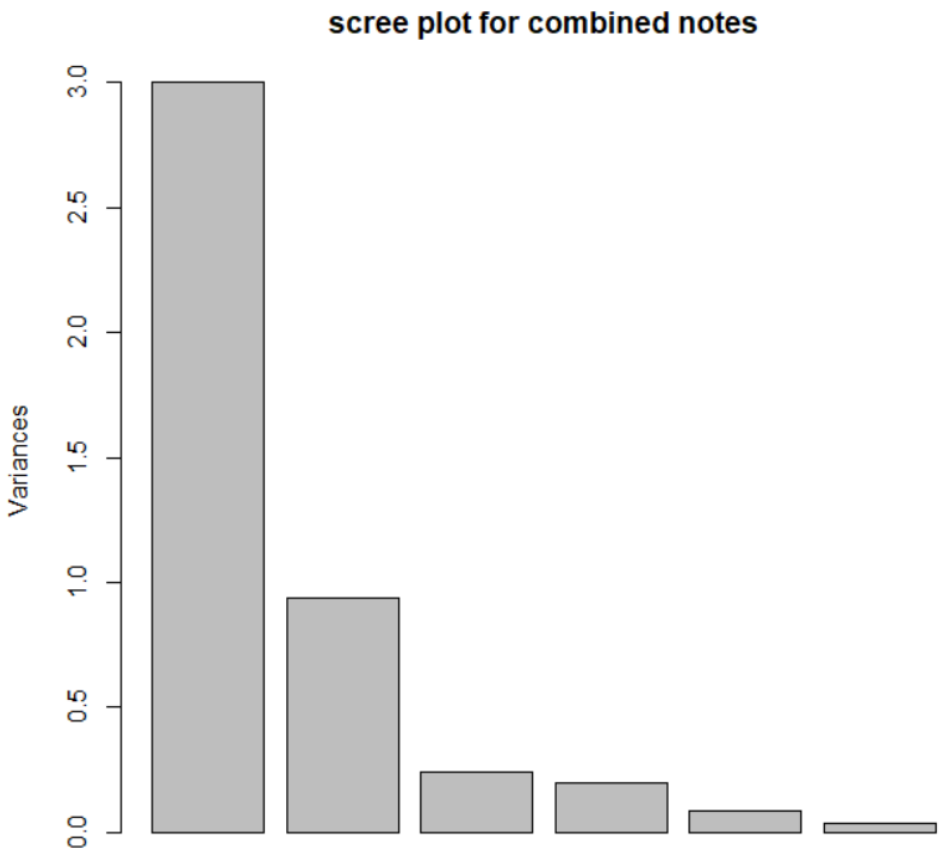

The above plot represents the variance of each principal component in numerical order.

```
> summary(pc2)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     1.245 0.5650 0.44012 0.32227 0.29048 0.15661
Proportion of Variance 0.681 0.1403 0.08515 0.04565 0.03709 0.01078
Cumulative Proportion  0.681 0.8213 0.90648 0.95213 0.98922 1.00000
```

The summary of the principal components of the counterfeit notes data shows that the first PC explains 68.1% of the variance and the second PC explains 14.03% of the variance. This is a significantly different trend from the principal components of the genuine notes data.



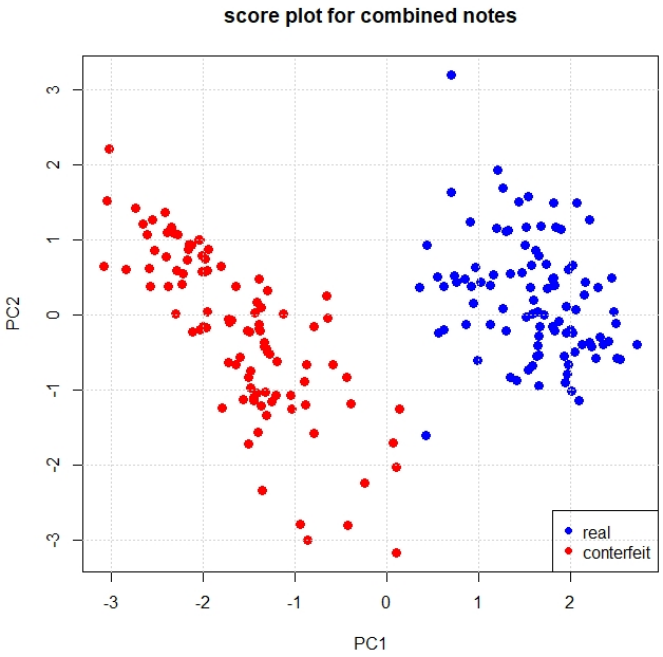
- Applying PCA on the combines notes data.




The above plot represents the variance of each principal component in numerical order.

```
> summary(pc3)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     1.7321 0.9673 0.49337 0.44120 0.29191 0.1885
Proportion of Variance 0.6675 0.2082 0.05416 0.04331 0.01896 0.0079
Cumulative Proportion  0.6675 0.8757 0.92983 0.97314 0.99210 1.0000
```

The summary of the principal components of the combined notes data shows that the first PC explains 66.75% of the variance and the second PC explains 20.85% of the variance.




The above score plot with PC1 as the x axis and PC2 as the y axis shows a clear separation between the real and counterfeit notes. From the biplot below, the principal features are the <u>diagonal, height.left</u> and <u>height.right</u> .