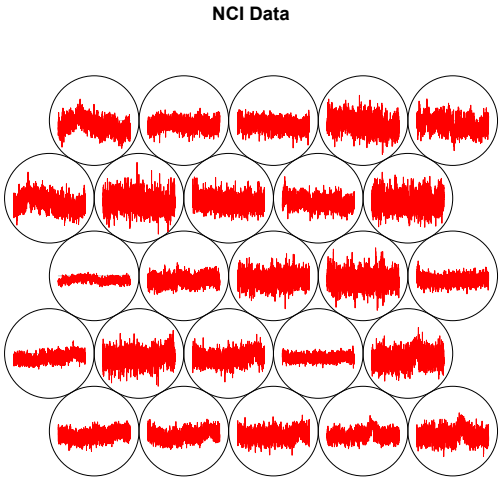
Homework 3 Report

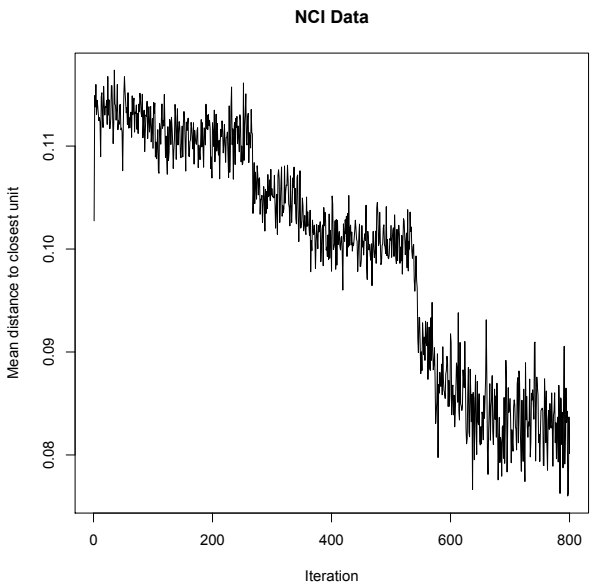
Name: Akshay Adlakha
UBIT: akshayad
Person#: 50317479

Problem 1: In this problem, I was given a tumor microarray data in a package ElemStatLearn. I had to perform the SOM to the data.

After loading the data, did a sanity check by checking NA values. And, explored the data. Then, scaled the data to have standard deviation value equal to 1.

SOM is performed to the data. Plotted different visualization to understand the SOM and the relationships between variables in our dataset. Here is the result:

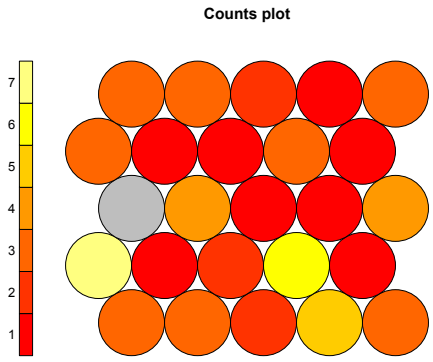




This is training iterations progress plot shows the progress over time, the node which represents the distance from each node's weights to the samples is reduced. Ideally, distance should reach to minimum level. If the curve keeps on decreasing then more iterations are required.

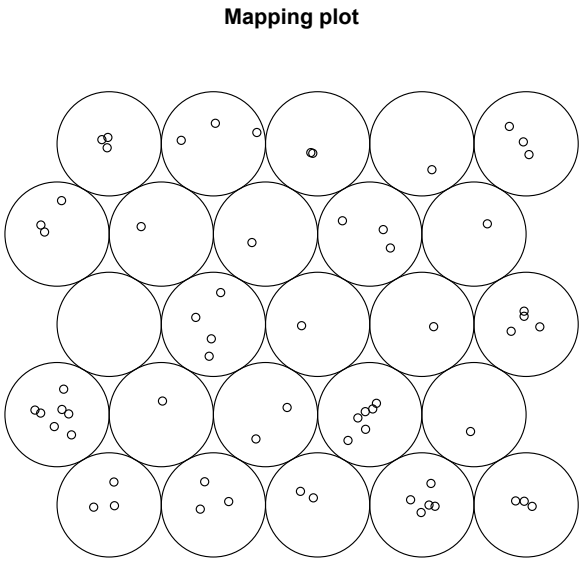
From the above graph, we can see that the curve comes to the minimum level.

Here is the count plot:

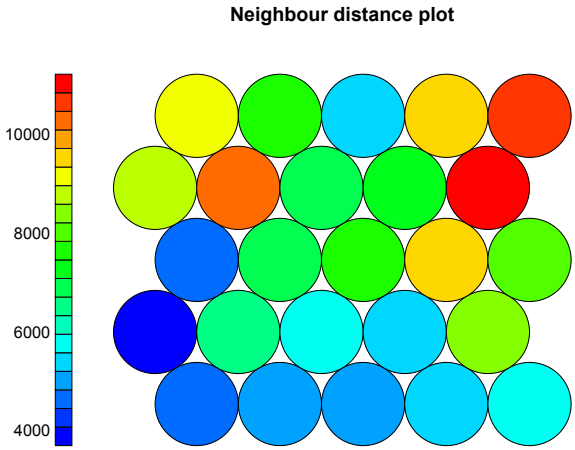


The count plot to visualize the count of samples are mapped to each node on the map. This measures the map quality. Each node in the plot has data.

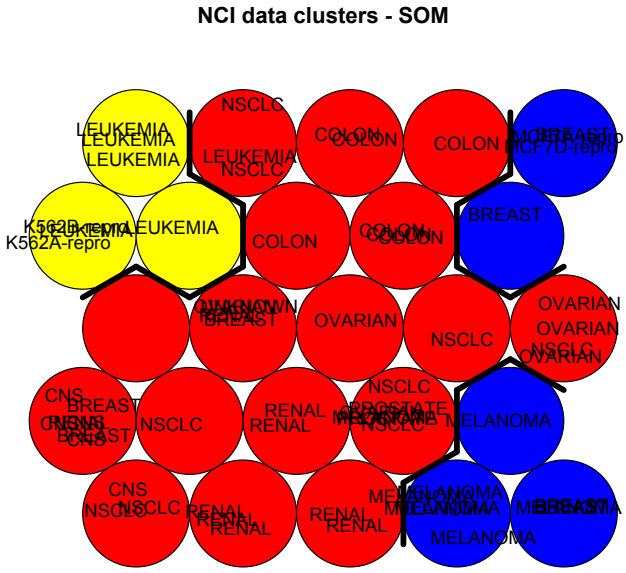
Their corresponding mapping plot:



The U-matrix plot:



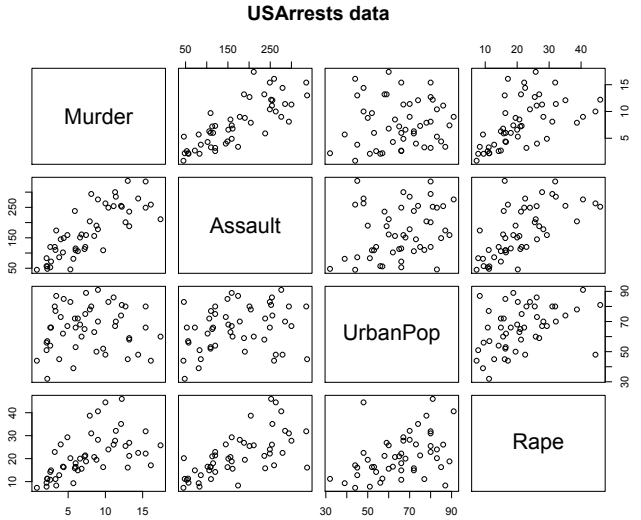
This U-matrix visualization represents the distance between each node and its neighbors. Areas of low neighbor distance indicate groups of nodes that are similar. Areas with large distances indicate the dissimilar nodes – and indicate natural boundaries between node clusters. It can be used to identify clusters within the SOM map.



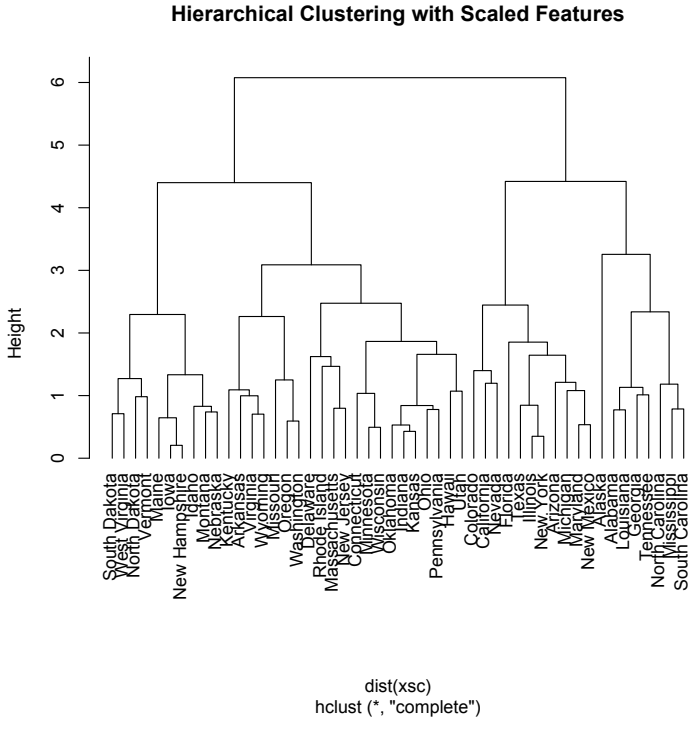
The above plot shows the data separated into groups. Each node has some data. Red cluster 1, blue cluster 2 and yellow cluster 3. And, the result obtained were as mentioned in the question that there are to be 2-3 groups/clusters in the data. It can be seen from the plot that most of the samples belonging to the same cancer type are grouped together.

Problem 2: In this problem, I was given a USArrests dataset from ISLR package. I had to perform the hierarchical clustering with complete linkage and Euclidean distance and SOM to the data.

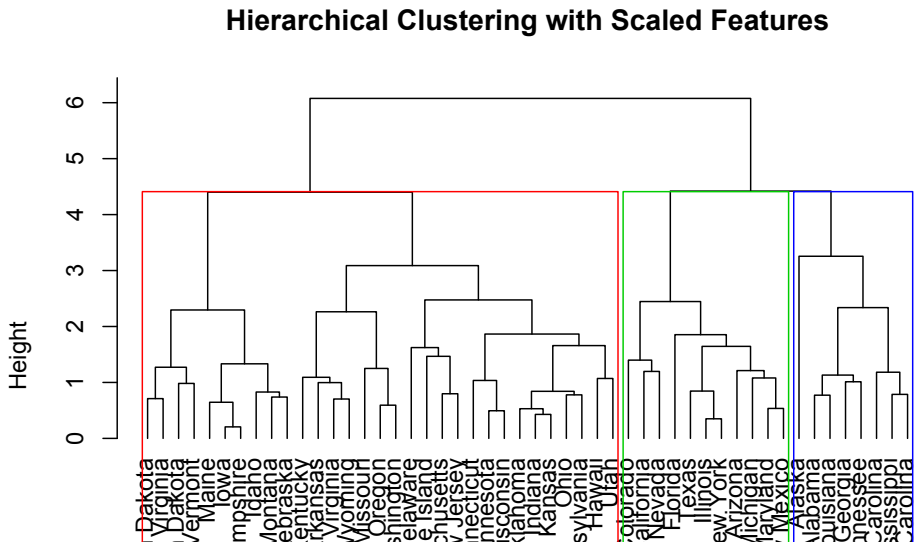
After loading a data, did a sanity check by checking NA values. And, visualize the data.



Then, scaled the data by using the scale method. And, performed hierarchical clustering with complete linkage and euclidean distance to the scaled data. Here is the result:



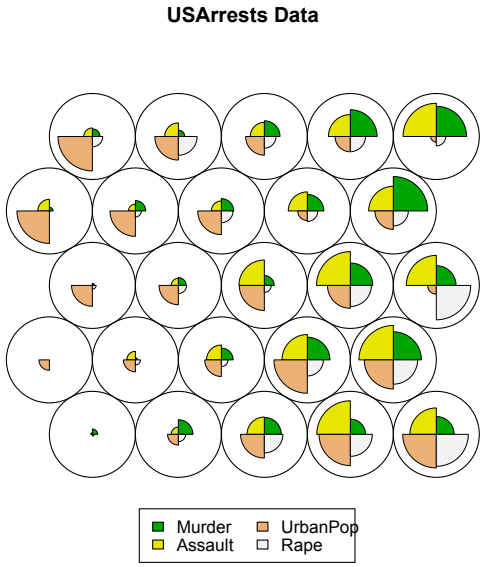
Then, cut this dendrogram to get three clusters. They are differentiated by a color. Here is the plot:



States under red boundary belong to cluster 3. States under green boundary are of cluster 2 and under blue are in cluster 1.

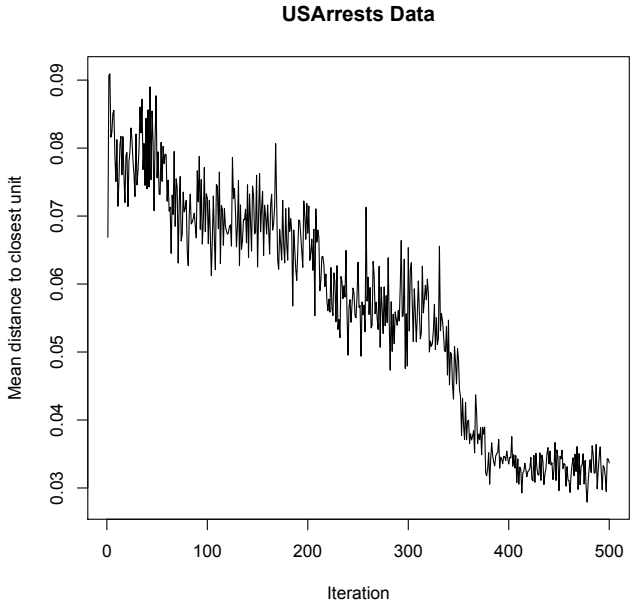
As we can see that it was expected to have high number of states in one of the cluster. The scaling has a significant effect on the clustering. So, it is important to scale the variables first if the feature range is much greater than others, then the distance would be bigger and impact on the overall distance is high. Each variable has given an equal importance in the hierarchical clustering as variables scaling is done before the inter-observation dissimilarities are computed.

Then, performed SOM to the data. Here is the result:



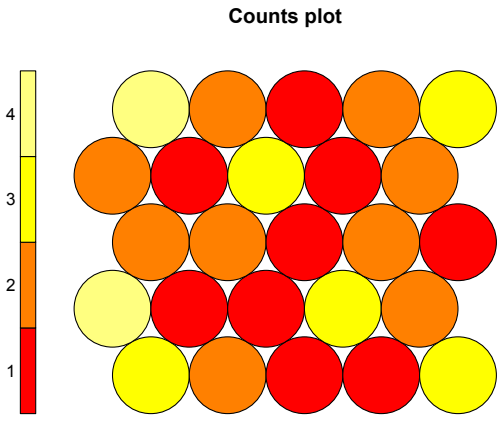
As we can see that the Murder and Assault are more on the right side of the plot and urban pop is more on the left side and lower of the plot.

And, see some other plots:

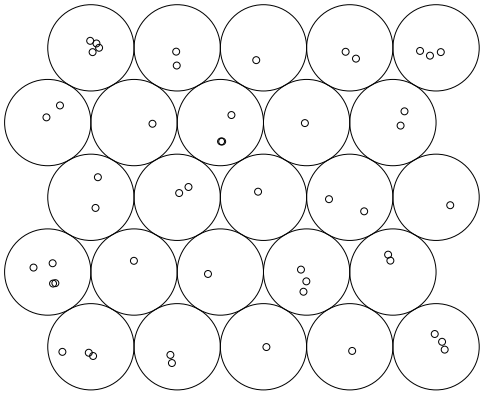


Above training iterations progress plot shows the progress over time. Here, distance reaching to minimum level. If the curve keeps on decreasing then more iterations are required.

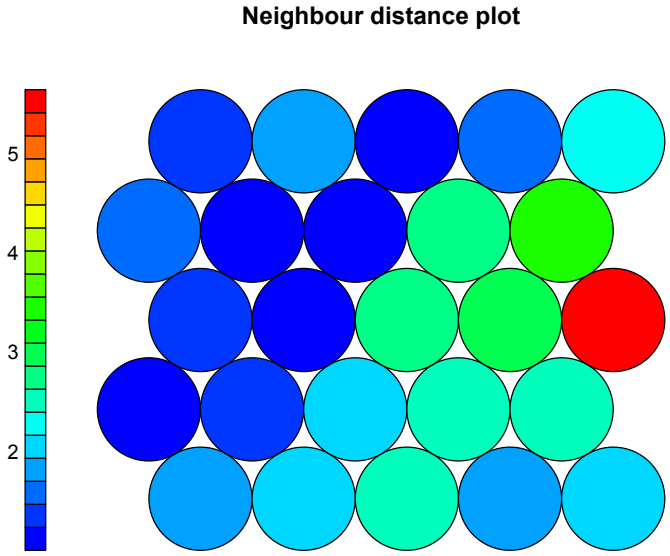
The count plot to see the number of samples mapped to each node.



Here is the mapping plot:



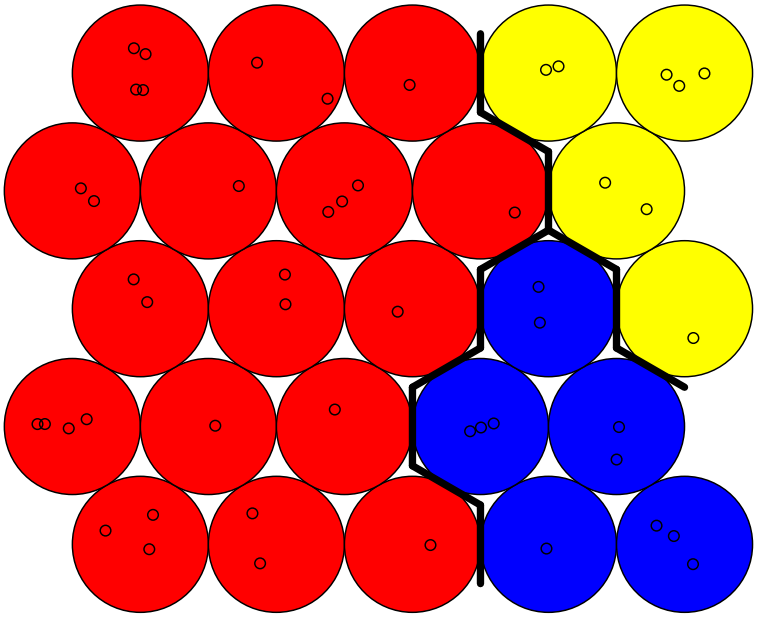
The U-matrix plot:



The lower value is on the left side of the plot and the group of similar nodes and value is high on the right side and has a group of dissimilar nodes.

Then, plotted the SOM with the found cluster. Here is the result:

Mapping plot



As we can see from the above plot, this result was expected as we had small dataset. And, this supports the result of hierarchical clustering in part A.The number of states in each cluster from the both method are same. In general, hierarchical clustering and SOM algorithm give the better results when used with small dataset.

C) Hierarchical clustering and SOM have ambiguity in some noise data when clustered. Hierarchical clustering is more sensitive to noisy data as compared to SOM. Although hierarchical and SOM show good results when used with small dataset.

For comparing the hierarchical clustering with SOM, the hierarchical tree is cut at two different level to obtain corresponding number of clusters(8,16,32). As a result, as a value of k becomes greater the performance of SOM gets lower. And, the accuracy of hierarchical clustering gets better until it reaches the accuracy of SOM algorithm.

They both give the better results when using random dataset.

Problem 3: In this problem, I was given a Swissbanknotes dataset. I had to perform PCA on first 100 data, next 100 data and combined data.

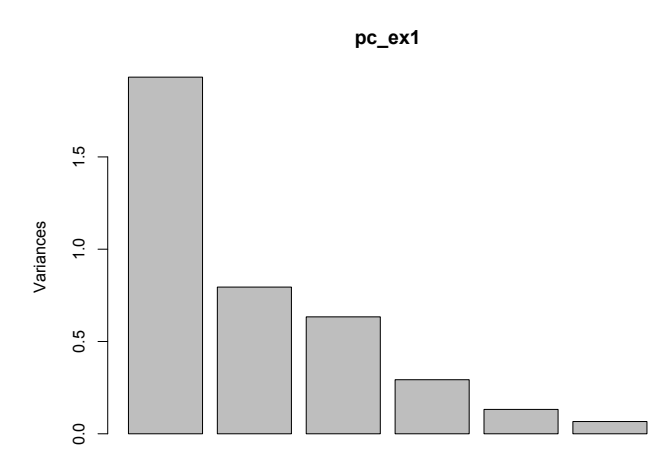
After loading dataset, did a sanity check by checking NA values and explored the data. Then checked mean and variance to see if scaling is required.

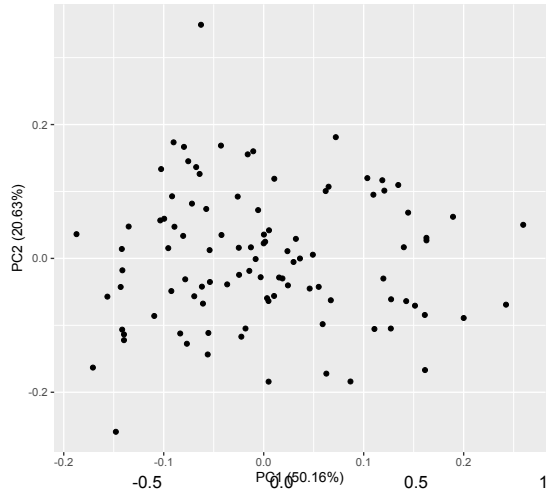
```
apply(dataset , 2, mean)
    length  height.left height.right  inner.lower  inner.upper     diagonal         note
  214.8960     130.1215     129.9565       9.4175      10.6505     140.4835       0.5000
apply(dataset , 2, var)
    length  height.left height.right  inner.lower  inner.upper     diagonal         note
 0.1417930    0.1303394    0.1632741    2.0868781    0.6447234    1.3277163    0.2512563
```

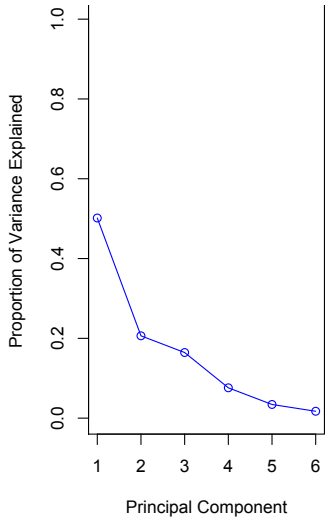
As we can see that, the features have different mean and variance. So, scaling is done to have standard deviation one. Then, performed PCA on 100 bank genuine notes. Here is the result:

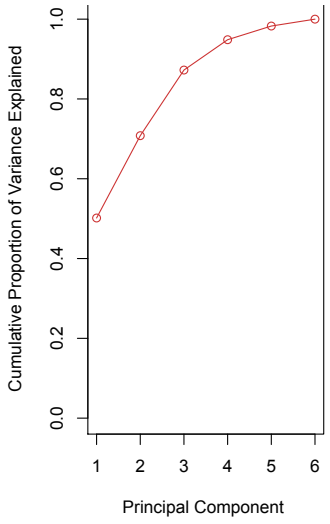
```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.3900 0.8915 0.7959 0.54123 0.36335 0.25779
Proportion of Variance 0.5016 0.2063 0.1645 0.07605 0.03427 0.01725
Cumulative Proportion  0.5016 0.7080 0.8724 0.94847 0.98275 1.00000
```

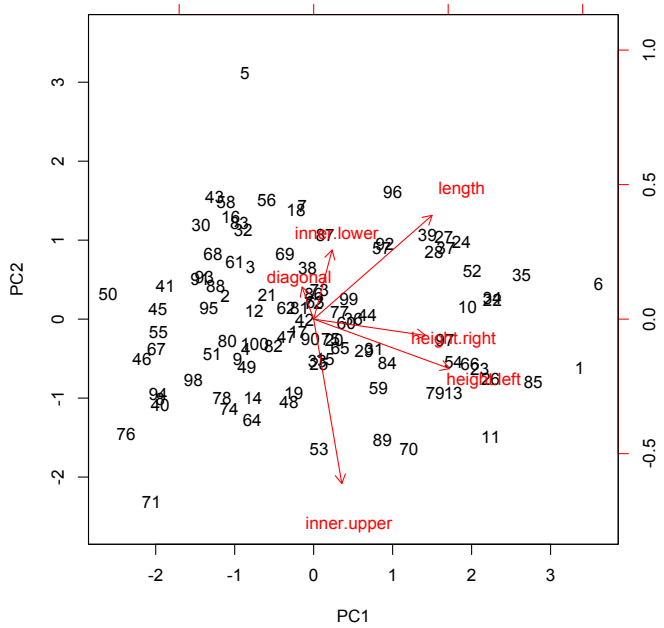
50% variance is explained by PC1 and next 20 by PC2. In total, 87% is explained by first three principal components. Here are some diagnostic plots and score plot to understand the PCA.









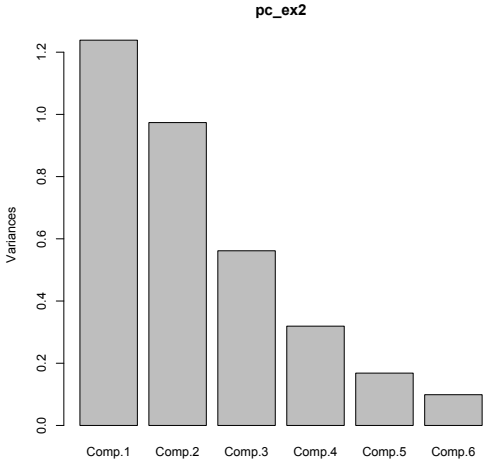


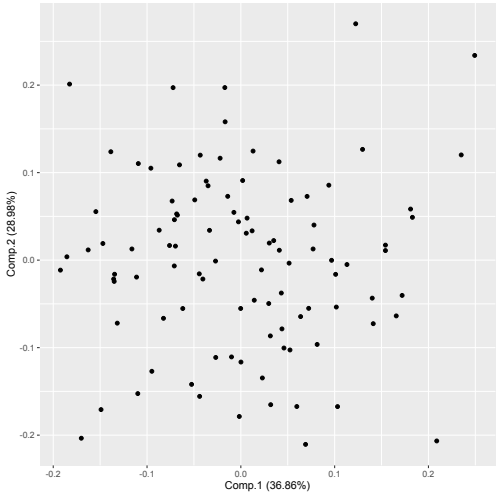
From the above result, we can see that features height.right, height.left are positively correlated. And, the scree plot is third figure explaining the variance by each principal component. Total variance explained is equal to 1.

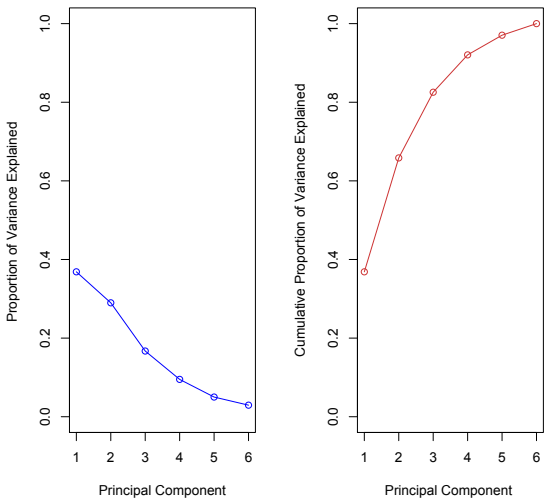
Then, performed on counterfeit set of notes. Here is the result:

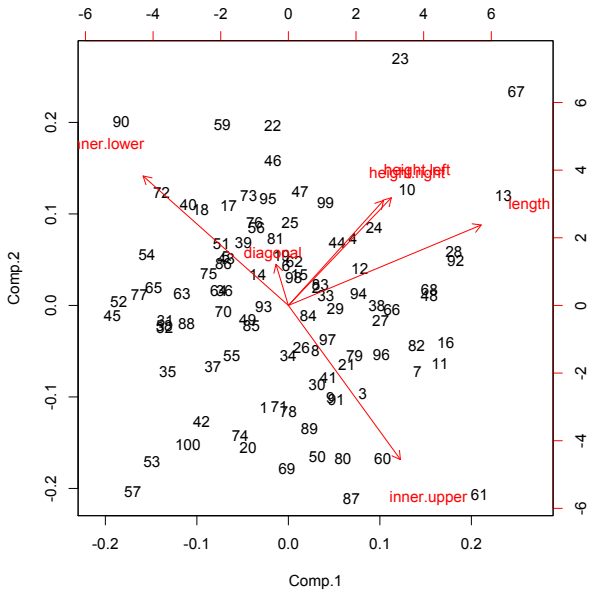
```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6
Standard deviation     1.1129305 0.9867567 0.7493592 0.5649973 0.41019510 0.31441592
Proportion of Variance 0.3686153 0.2897726 0.1671157 0.0950014 0.05007468 0.02942025
Cumulative Proportion  0.3686153 0.6583879 0.8255037 0.9205051 0.97057975 1.00000000
```

Here, first principal component explains the 36% variance and 28% explained by second component. And, first four components are explaining 92% variance of data which is most of the variance.









As we can see that the variance explained by principal components is bit different for this data as compared to the genuine notes data.

And, we can see from the biplot, the orientation of features has changed in this set of data as compared to the set of genuine data.

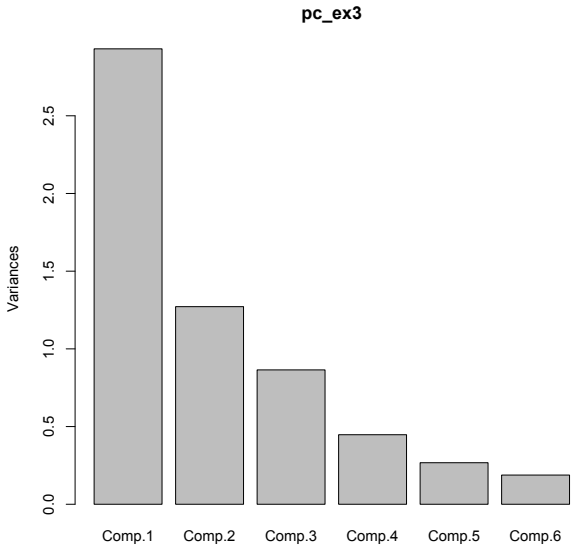
To see the difference in the pattern of the data, performed the PCA on the combined data of genuine and counterfeit bank notes. Here is the result:

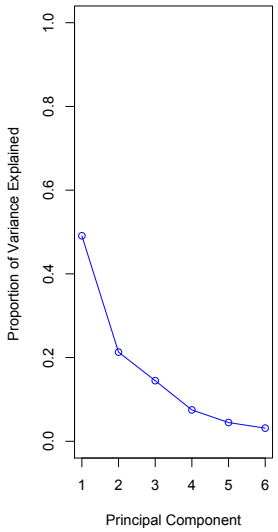
```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     1.7119668 1.1276938 0.9298857 0.66896923 0.51704305 0.43351526
Proportion of Variance 0.4909264 0.2130140 0.1448388 0.07496145 0.04477948 0.03147998
Cumulative Proportion  0.4909264 0.7039403 0.8487791 0.92374054 0.96852002 1.00000000
```

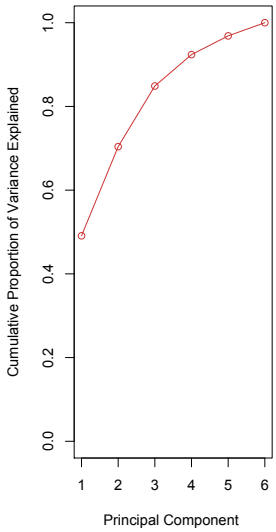
Here, the first three PCs are explaining the 84% of variance. And, they are contributing most of the variance.

As we can see that scaling has significant effect on the data. If we fail to do the scaling the most of the components would have been driven by one or two features.

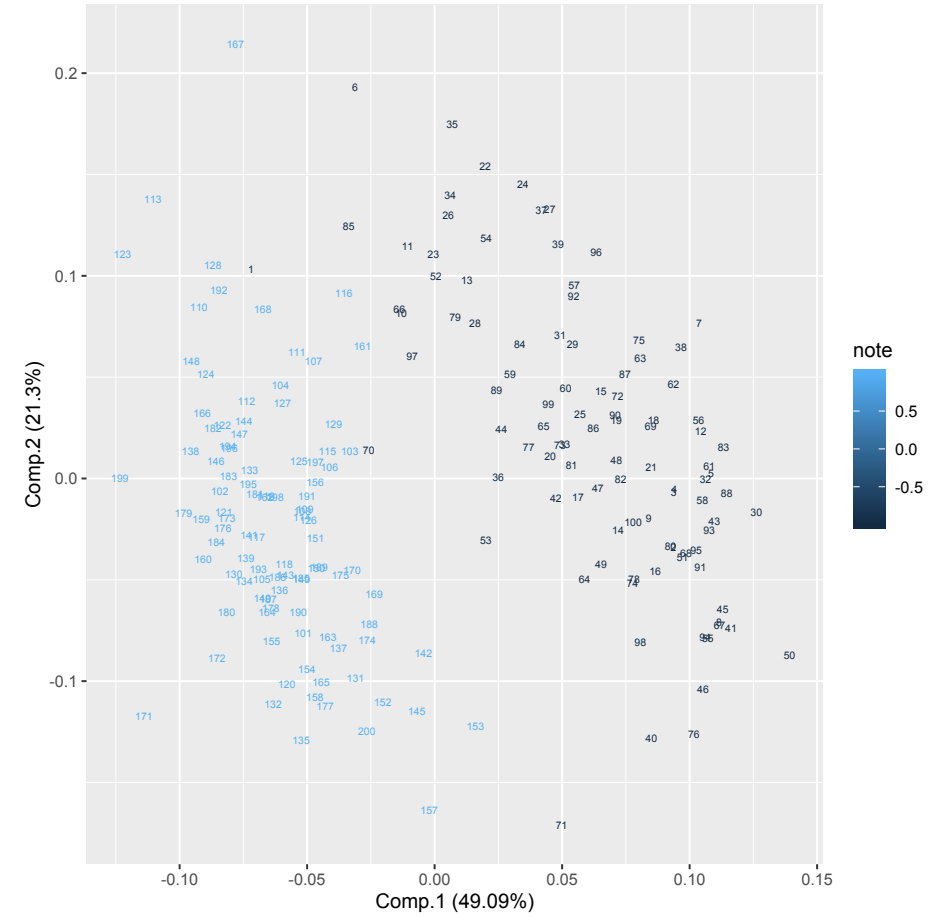
Here are some plots:

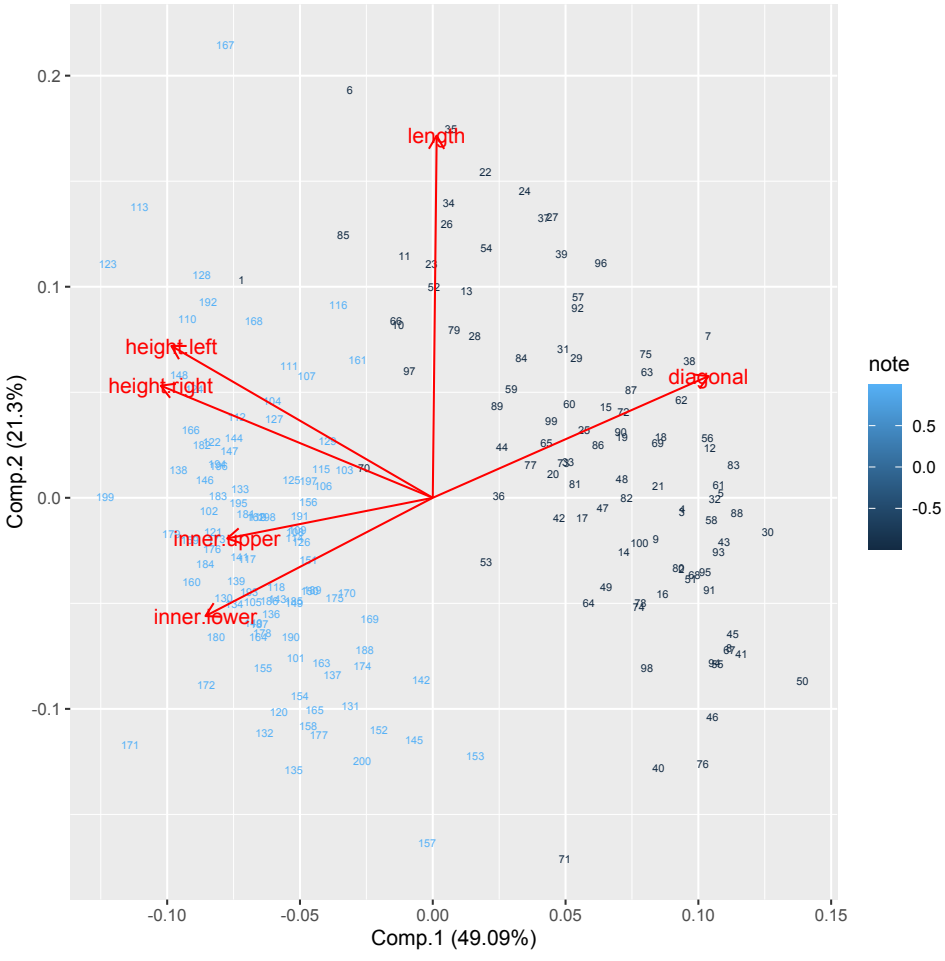






The scree plot and cumsum plot explaining the variance explained by each feature. Most of the variance is explained by first three variance.


Comp.2 (21.3%)
Comp.1 (49.09%)
note
0.5
0.0
-0.5


length
height.left
height.right
diagonal
inner.upper
inner.lower
Comp.2 (21.3%)
Comp.1 (49.09%)
note
0.5
0.0
-0.5

As we can see that there is clear separation between both the set of genuine and counterfeit data.Here, with combined data, features have different orientations. Height.left ,height.right and diagonals are in the opposite direction. They have negative correlation. The principal components pertaining to class-0 which is the genuine notes, has the stronger effect because of feature diagonal. And, the principal components pertaining to class-1 which is the counterfeit notes has the stronger effect because of height.left, height.right.

Feature diagonal and length have different orientations and causing the separation between classes.