STA - 546

STATISTICAL DATA MINING – 2

HW-3

ABINESH SENTHIL KUMAR

#50320934

Question 1

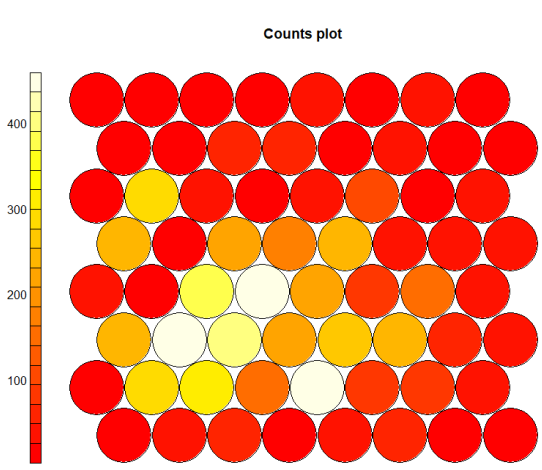
1) (10 points) Consider the tumor microarray data in the package library(ElemStatLearn). >library(ElemStatLearn) >data(nci) >head(nci)

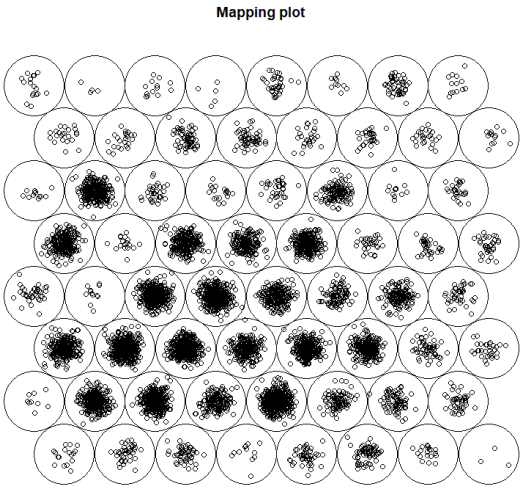
The data consists of several different types of tumor samples. We observed that it many clustering algorithms there are often found to be 2-3 groups/clusters in this well-studied data, although there are 14 subtypes of tumor cells (unique(colnames(nci))). Run a SOM algorithm and present the results (e.g.,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

Running SOM on data,

- Since the variables are of same scale, scaling is not required, and we can perform the analysis on the raw data.
- SOM is run on the data with hexagonal grid size 8*8 and 3000 iterations.
- Since our motive is clustering, we have to define the grid size in a way that the points get spread out evenly and make sure the prototype is not empty or least number of values.

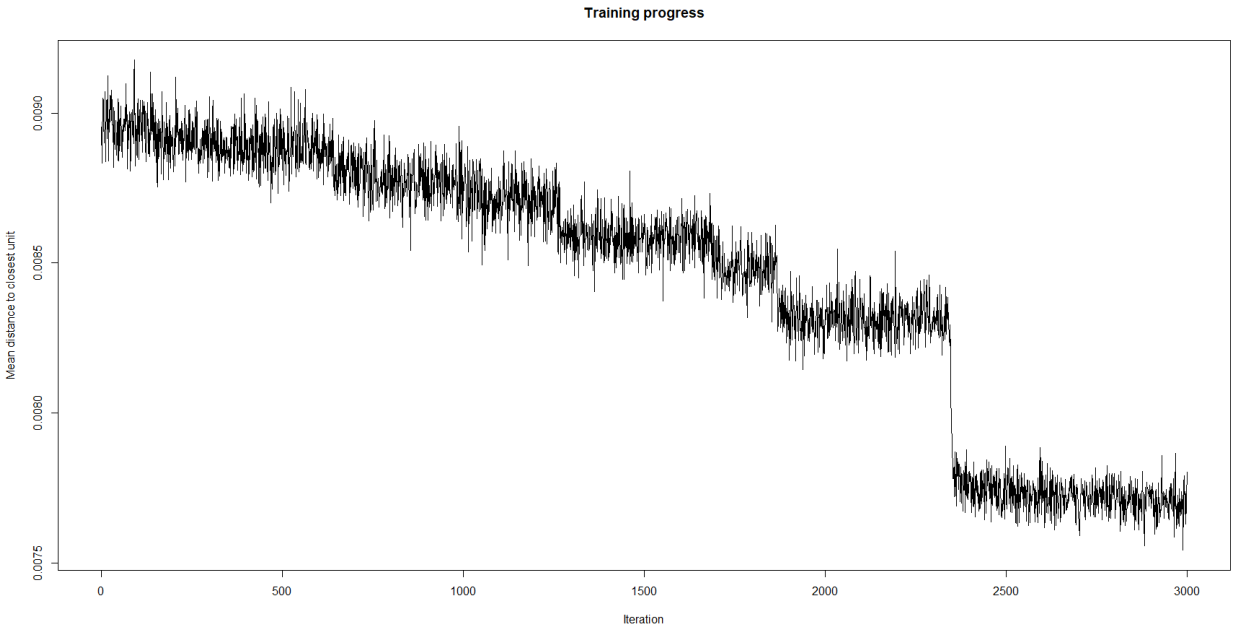
From the counts and mapping plot,





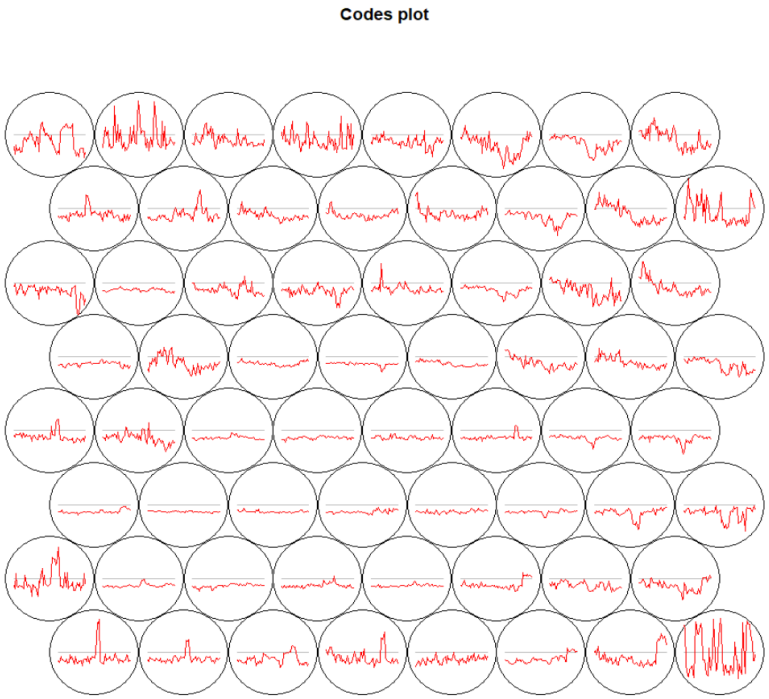
- We can see that the observations are evenly spread and there are no empty prototypes.

From the changes plot,



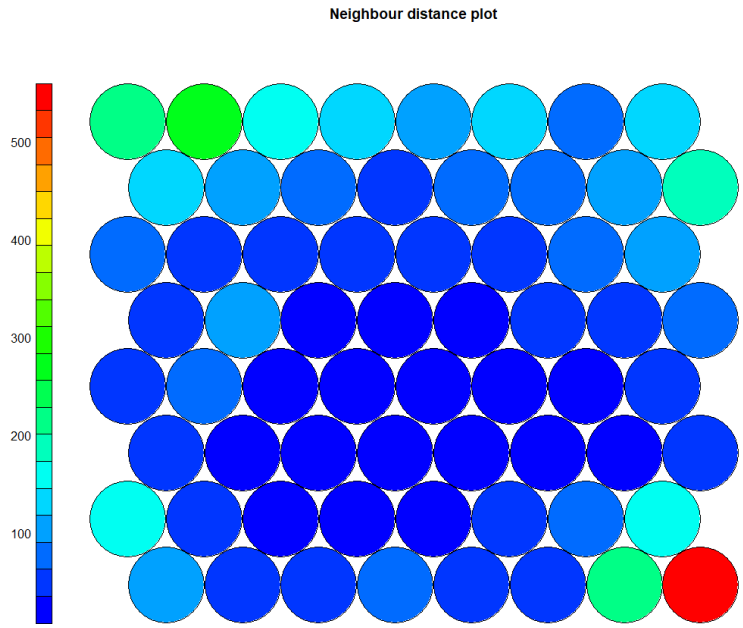
- We can see that the distance has converged around 2500 iterations, so having 3000 iterations will be sufficient.

From the codes plot,



- Since we have 64 variables, we cannot clearly see the magnitude of each variable in this plot.

From the U-matrix/distance. Neighborhood plot,

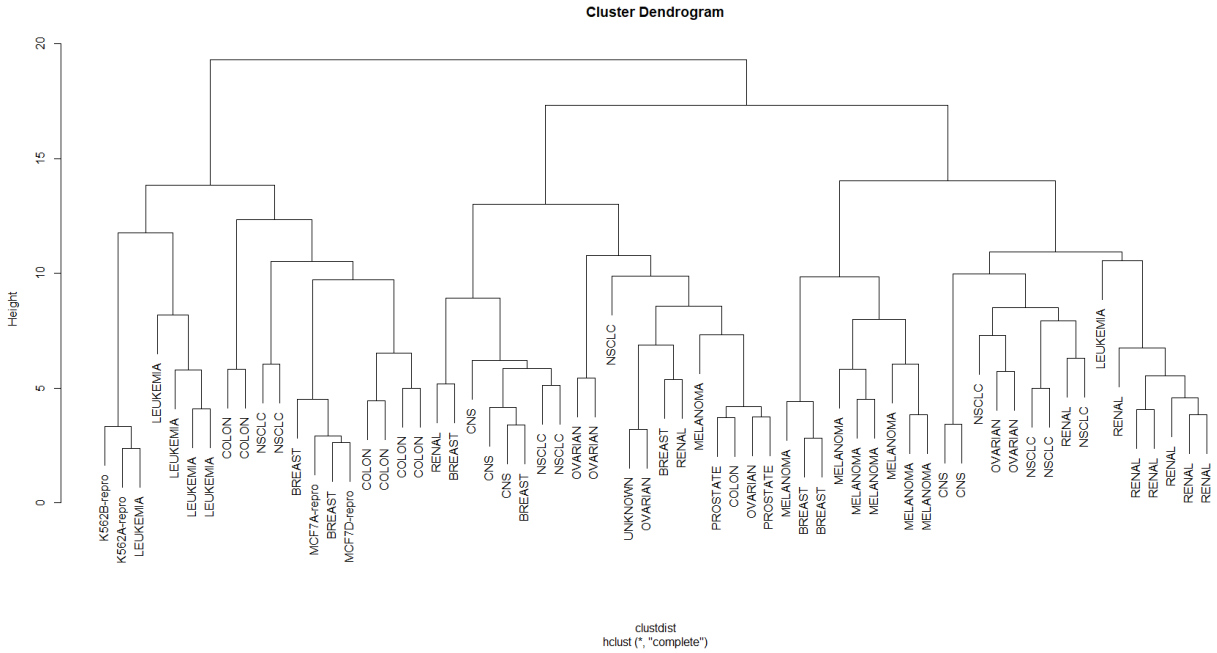


- We can see that the center of the plot has clustered together with similar distance to the neighbors and we can see patterns with three clusters.
- The component plane plots are not appropriate to this dataset since we have many features and we are going to cluster based on the variables instead of the observations.

Applying hierarchical clustering to the prototypes,

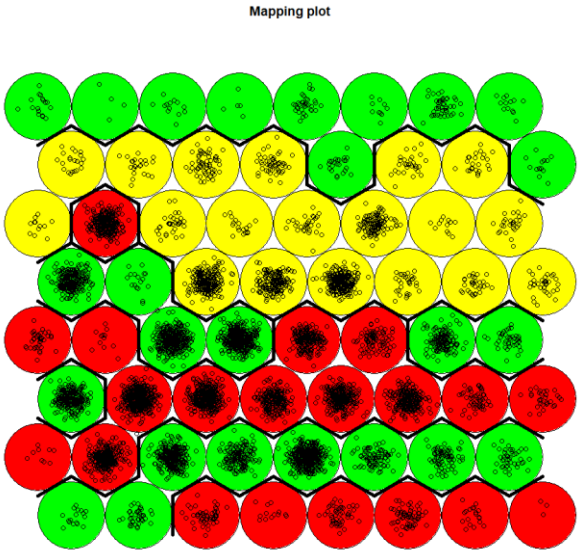
- The observations are clustered using hierarchical clustering using complete linkage.
- For the distance, the codes from the SOM results are obtained and transposed since, the objective is to classify the data based on the tumor cells (columns) and Euclidean distance is used.

The resulting dendrogram,



- From the dendrogram we can see that the prototypes group into three.
- Using the cuttree function we cut the dendrogram at height of 15 to obtain three clusters.

Plotting the som grid mapping with respect to the found cluster,



- As seen in the dendrogram, one of the clusters are well separated while the remaining two are mixed inside each other.
- From the analysis we can say that the SOM method does not well characterize the variables. For this particular dataset

Question 2

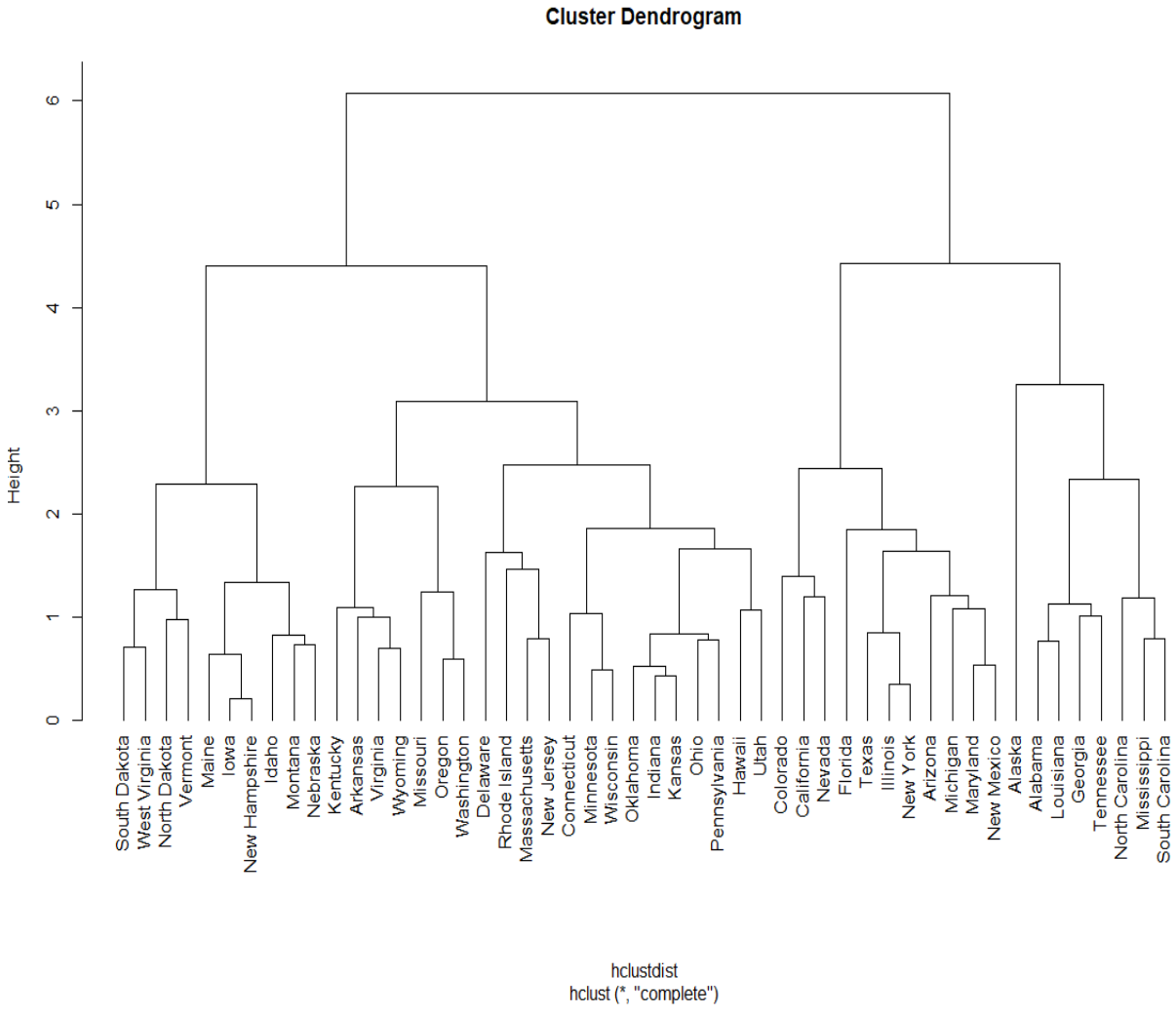
(10 points) Consider the USArrests data. I suggest scaling the data first. > library(ISLR) > data(USArrests) > head(USArrests)

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?

Scaling and performing hierarchical clustering,

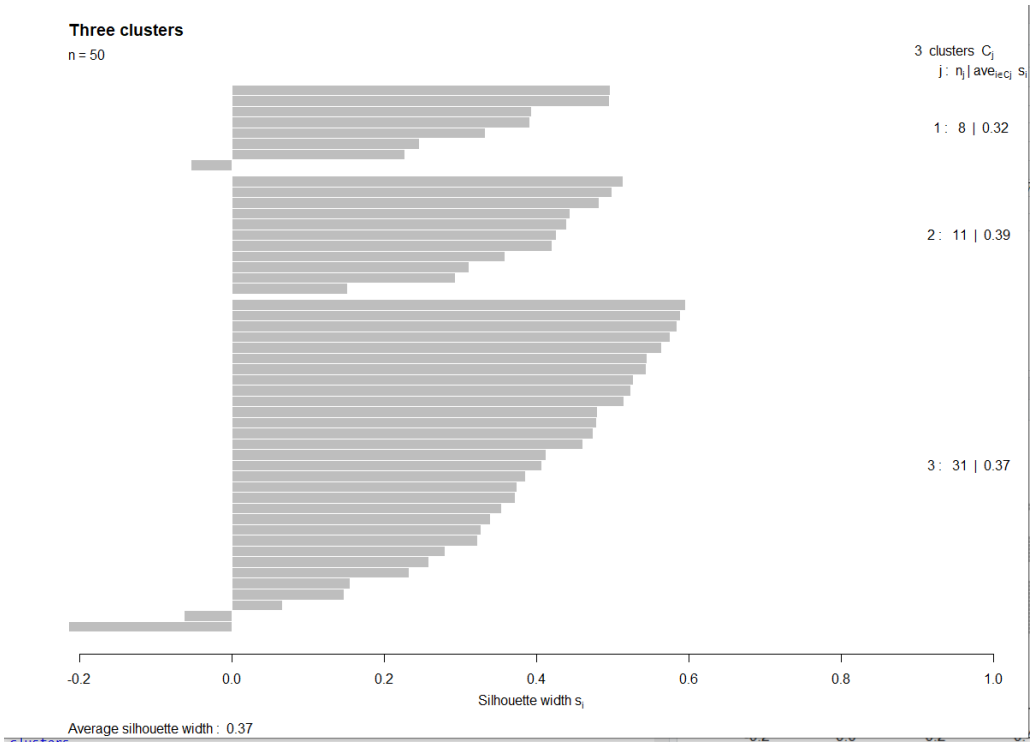
- Since all the variables are not in the same scale we first scale the data using scale() function.
- Dissimilarity matrix is obtained by calculating the **Euclidian** distance and **hierarchal clustering** is performed with **complete linkage**.

The resulting dendrogram,

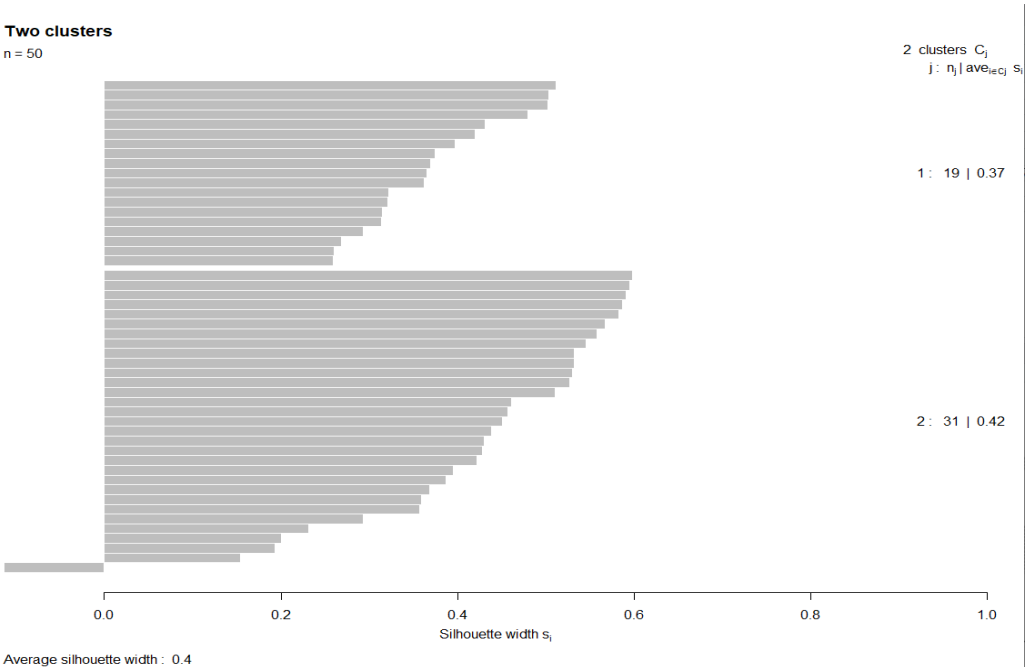


- From the dendrogram it is obvious that we can see **two groups** forming. But we first cut the dendrogram so the result is three clusters and also cut at height that result in two clusters to compare both.
- We then look at the **silhouette plot** to see if we find any issues with clustering.

Silhouette plot for three clusters,



Silhouette plot for two clusters,



- Looking at both the silhouette plots we can see that the average silhouette width for **two clusters is large** (0.4) when compared to the three clusters (0.37). So, we proceed with two clusters.

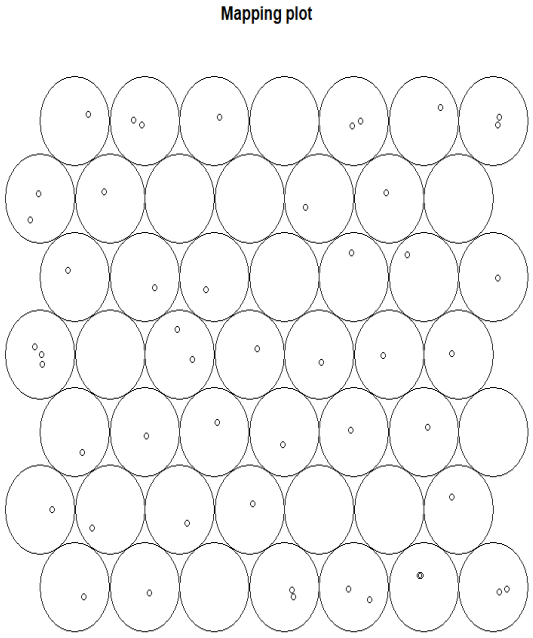
And, yes this is what I would expect.

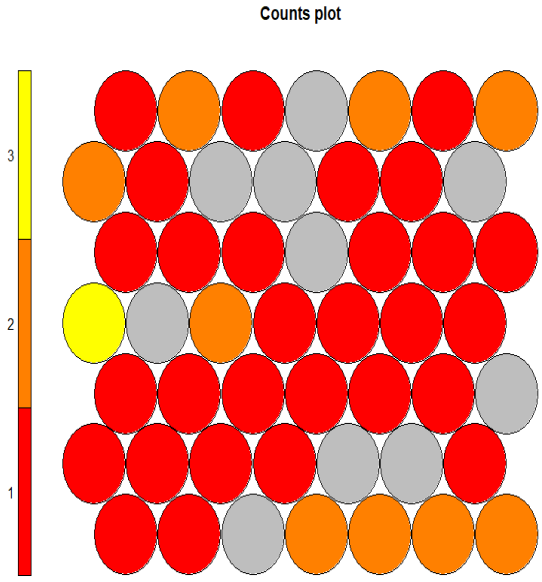
b) Fit a SOM to the data and present the results (e.g., U-matrix, phase plots if appropriate, hclust on prototypes). Is this what you would expect? Does this result generally support your results in Part A.

Fitting SOM to data,

- To fit SOM to the data, we first create a maximum possible grid size of **7*7** to the data. (since there are only 50 observations, we can't go beyond hexagonal grid size of 7*7)
- We then fit self-organizing map to the data using SOM from kohonen package with 3000 iterations (as the mean distance converges around 2500 iterations seen in changes plot).

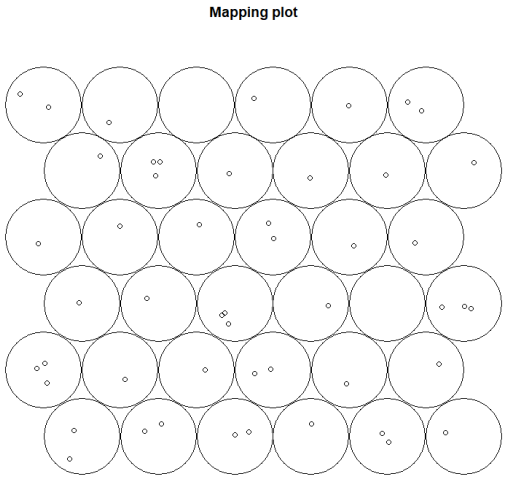
Looking at the mapping and counts plot for 7*7 grid,

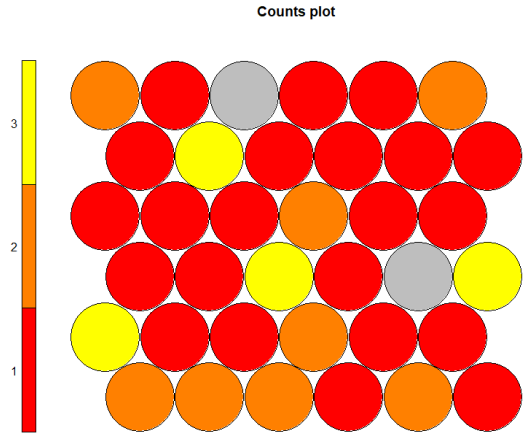




- We can see that there are lot of **empty prototypes** so we should **reduce the grid size.**

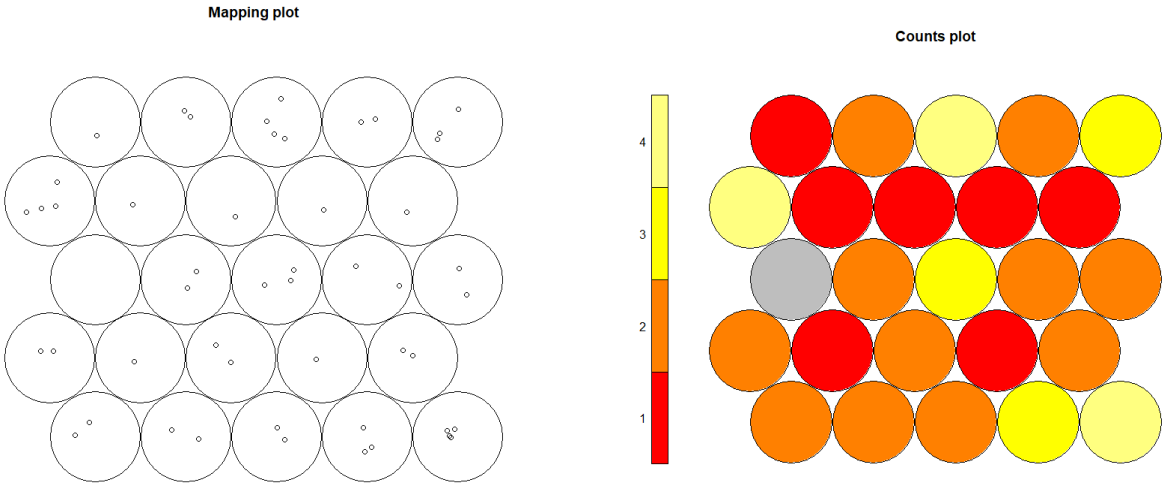
For grid size 6*6,





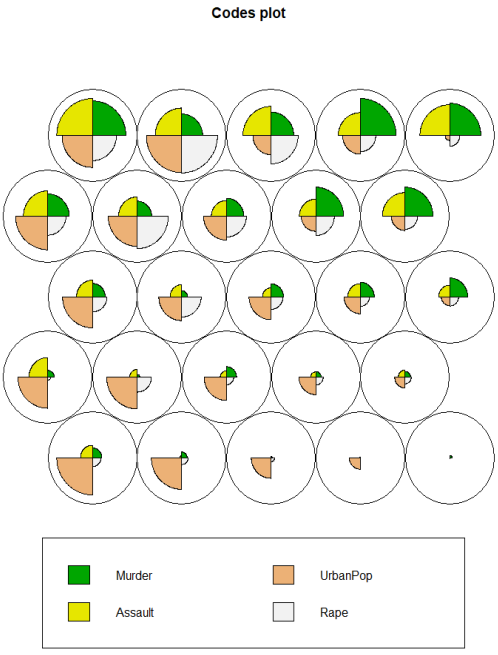
- We can still see **empty prototypes**, so we **reduce** the **grid size** again.

For grid size of 5*5,



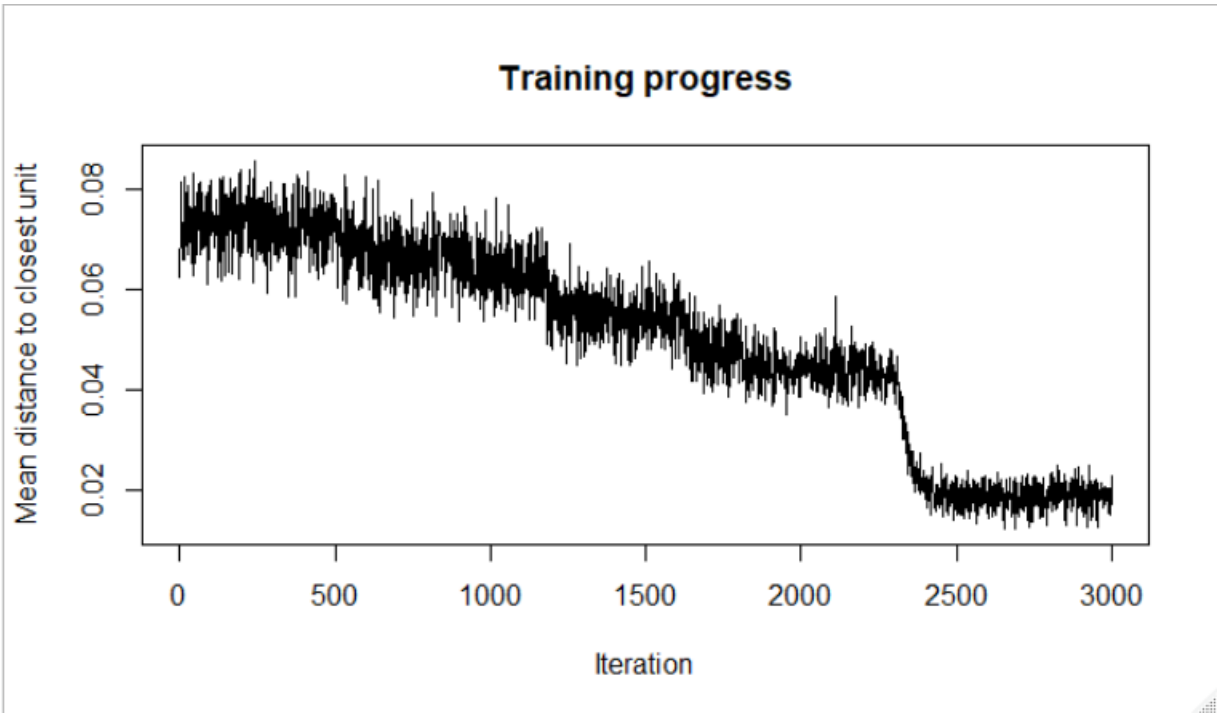
- We can still see one empty prototype, but the counts are more evenly distributed, and the plot looks more stable.
- Since our end goal is clustering, we want the points to be more evenly distributed so, we finalize the grid size to be **hexagonal 5*5**.

Looking at the **codes** plot,



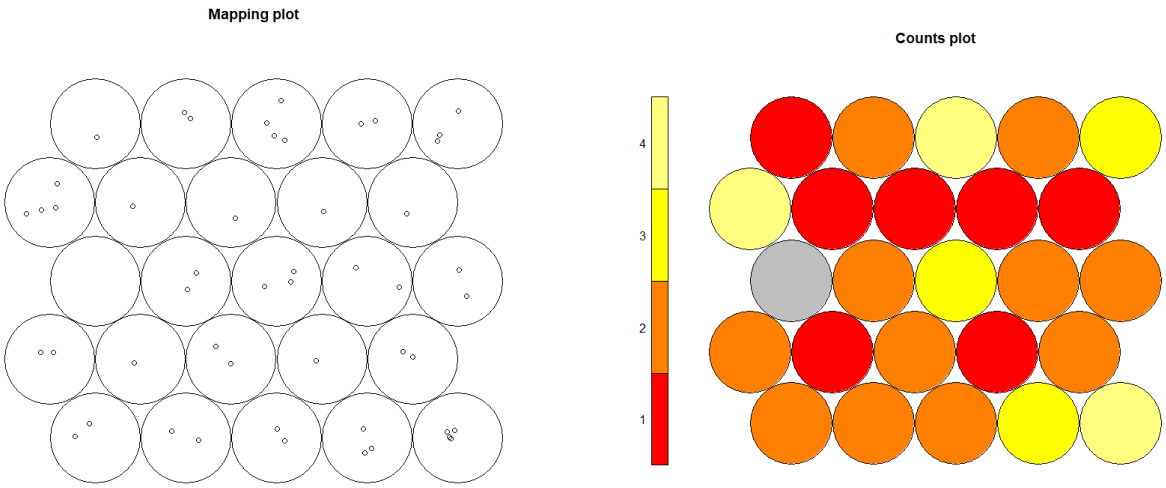
- The magnitude of the variables in each prototype is given by this plot.
- We can see that value of urbanpop is dominated in the lower prototypes and value of murder is dominated in the upper right side prototypes.

From the changes plot,



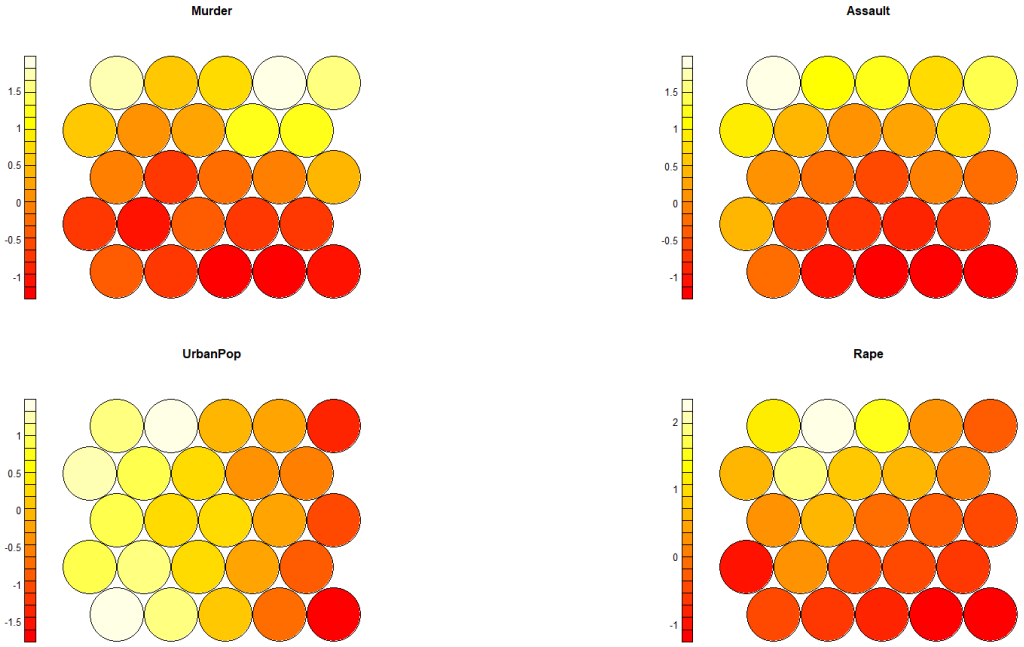
- We can see that the distance converges at around 2500 iterations.

From the **counts and mapping** plot,



- We can see how many **observations** are classified into **each prototype** from these two plots.

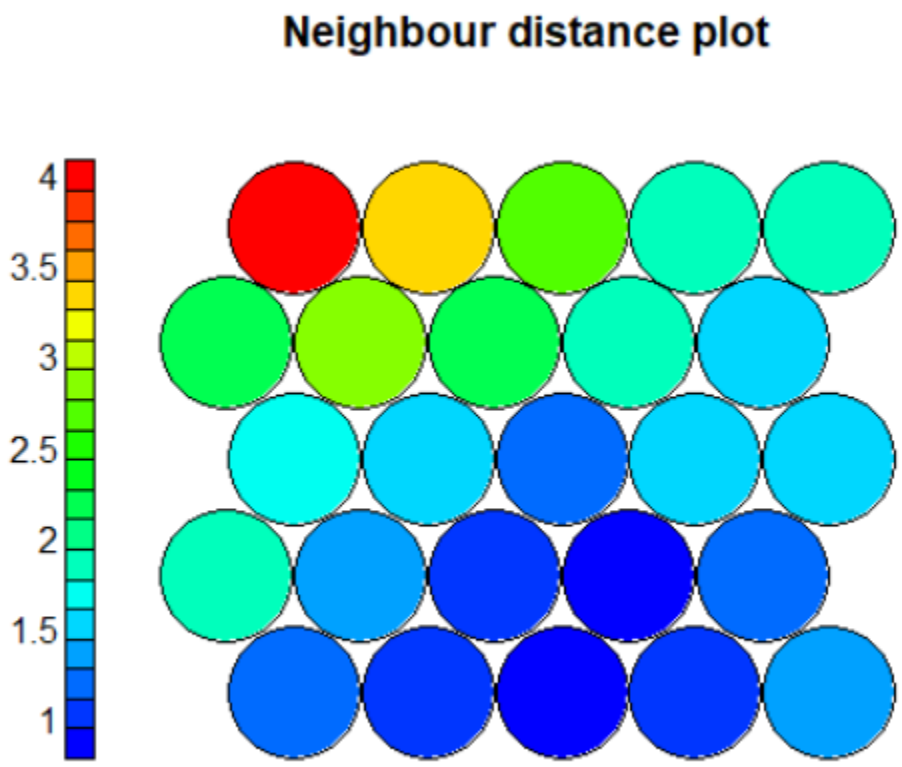
From the component plane plots we can see the **relationship of the variables** with respect to the observations assigned in the prototypes,



We can tell that,

- Murder, assault and rape are positively correlated.
- The bottom right prototypes contain countries with low rape, assault, murder and urbanpop values.
- Top left prototypes contain all high values of murder, rape and assault variables.
- The values of the variables start to decrease for the prototypes from top to bottom for murder, rape and assault.
- For urbanpop the values decrease from left to right.

From the U-matrix/neighborhood distance plot,



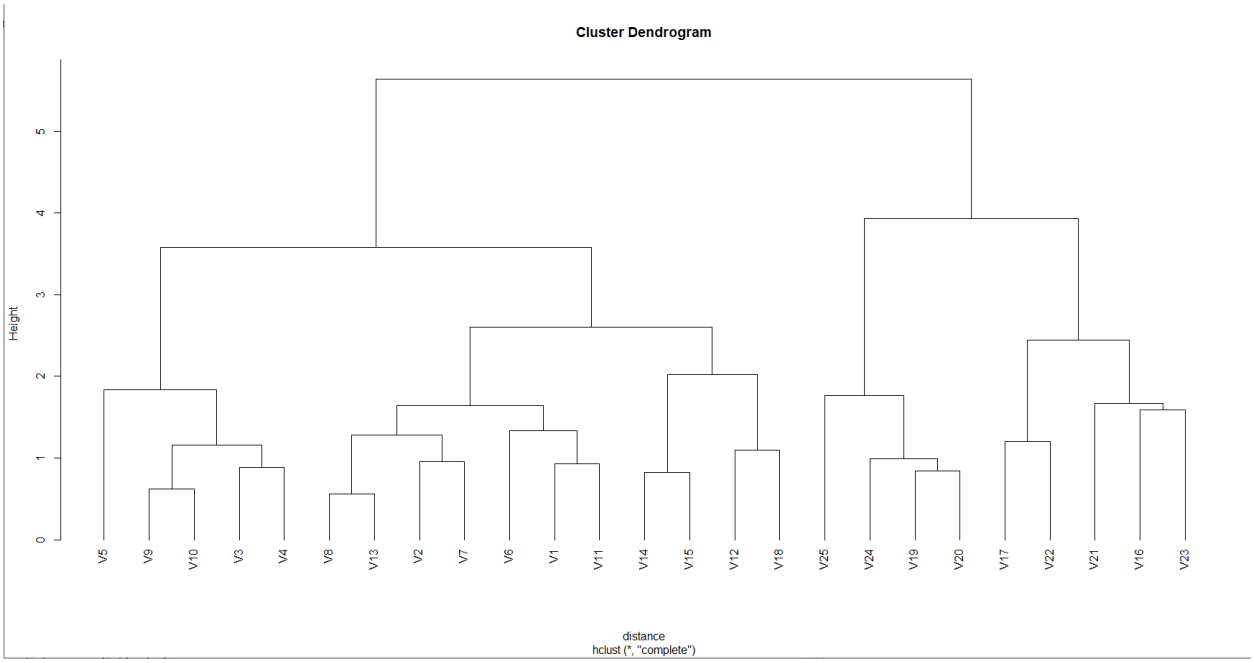
- We can see the blue ones are clustered together and we can see a pattern.
- The top left prototype is very much different among the others.

From the unit.classif function from som we can look at which observations belong to which prototypes.

```
> rownames(usarrestdata)
 [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"       "California"     "Colorado"
 [7] "Connecticut"    "Delaware"       "Florida"        "Georgia"        "Hawaii"         "Idaho"
[13] "Illinois"       "Indiana"        "Iowa"           "Kansas"         "Kentucky"       "Louisiana"
[19] "Maine"          "Maryland"       "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
[25] "Missouri"       "Montana"        "Nebraska"       "Nevada"         "New Hampshire"  "New Jersey"
[31] "New Mexico"     "New York"       "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
[37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina" "South Dakota"   "Tennessee"
[43] "Texas"          "Utah"           "Vermont"        "Virginia"       "Washington"     "West Virginia"
[49] "Wisconsin"      "Wyoming"
> somdatafive$unit.classif
 [1] 20 23 16 15 22 17  2  6 21 24  2 10 16 13  4  8 15 24  4 23  1 23  3 25 18 10  9 22  4  1 23 16 25  5 13 13 12  8  6
[40] 25  5 19 16  7  5 14 12  5  3 14
>
```

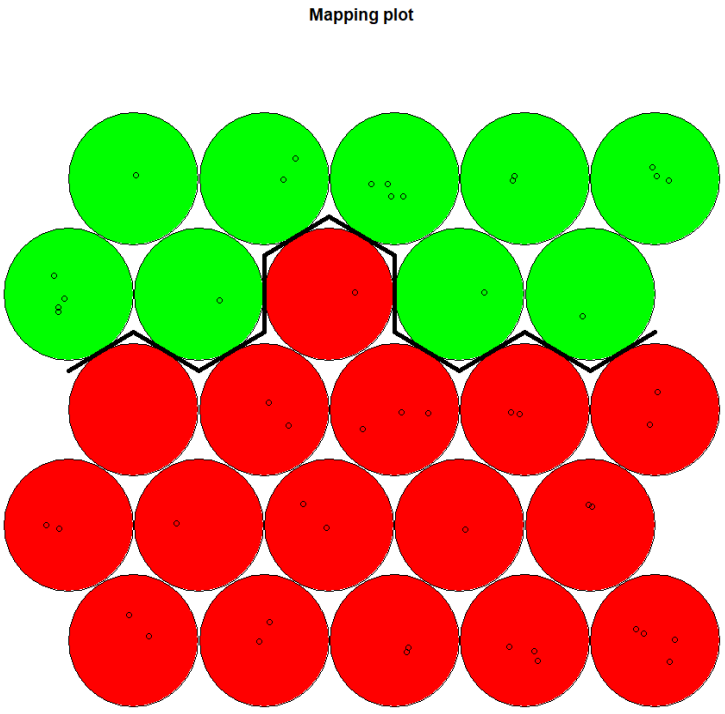
Hierarchial clustering with **complete linkage** is performed on the prototypes with distance calculated from the prototype codes.

The resulting dendrogram,



- We can see that the prototypes split into **two groups**. So the dendrogram is cut at height 5 to get two clusters.

Plotting the SOM grid mapping with respect to the found clustering groups gives,



- From the above analysis we can say that the SOM supports the clustering done previously.
- It is evident from the hierarchical clustering done on the prototypes since, the dendrogram showed two groups.

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

- SOM is more of a visualization algorithm than clustering, we can use other clustering techniques on top of SOM like k-means/k-medoids/hierarchical clustering to cluster the prototypes and analyze how they group together.
- SOM provides good visualization and is preferred when we want to visualize in the context of original variables.
- We can look at the u-matrix to check if any natural groupings fall out and decide on clustering on top of that.
- There is no evaluation metric for grid size, it is done by trial and error which can be perceived differently by different persons. So, this can be considered as a limitation of SOM
- Hierarchical clustering is performed when we don't know the expected clusters in the data and we could look at the dendrogram to decide on number of clusters.

Question 3

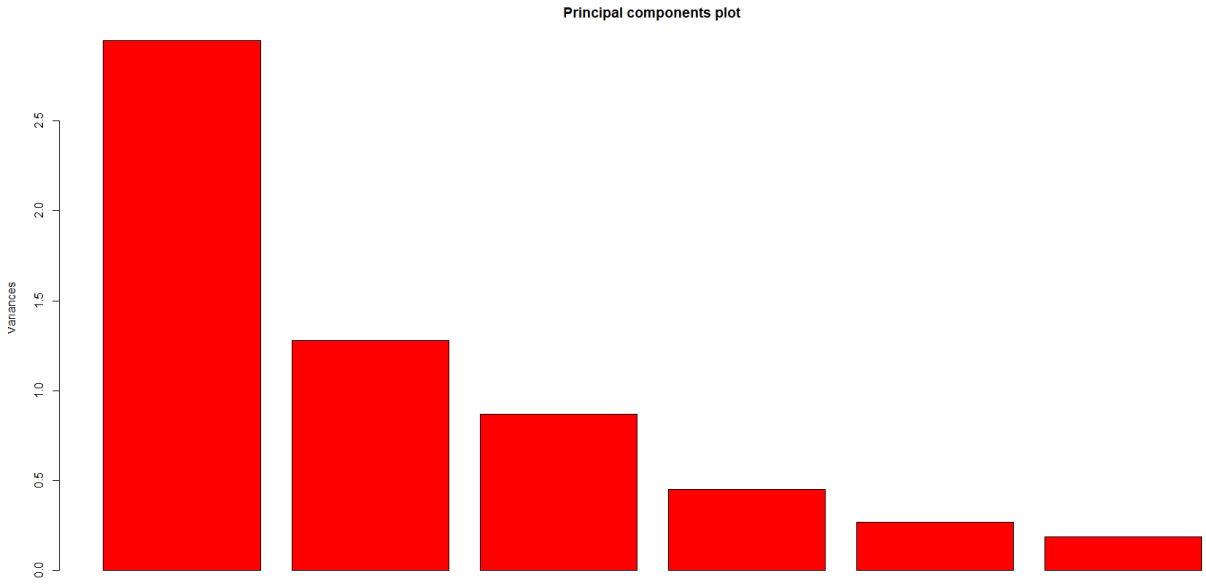
Carry out a PCA of the 100 genuine bank notes, of the 100 counterfeit bank notes, and all of the 200 bank notes combined. Generate some score plots (use colors for the combined). Do you notice any

differences in the results? Show your work, and justify the selection of Principal Components, including diagnostic plots.

PCA for combined data (both genuine and fake)

- We first scale the data using scale() function since the variables are not in common scale. and create references of fake or genuine bank notes.

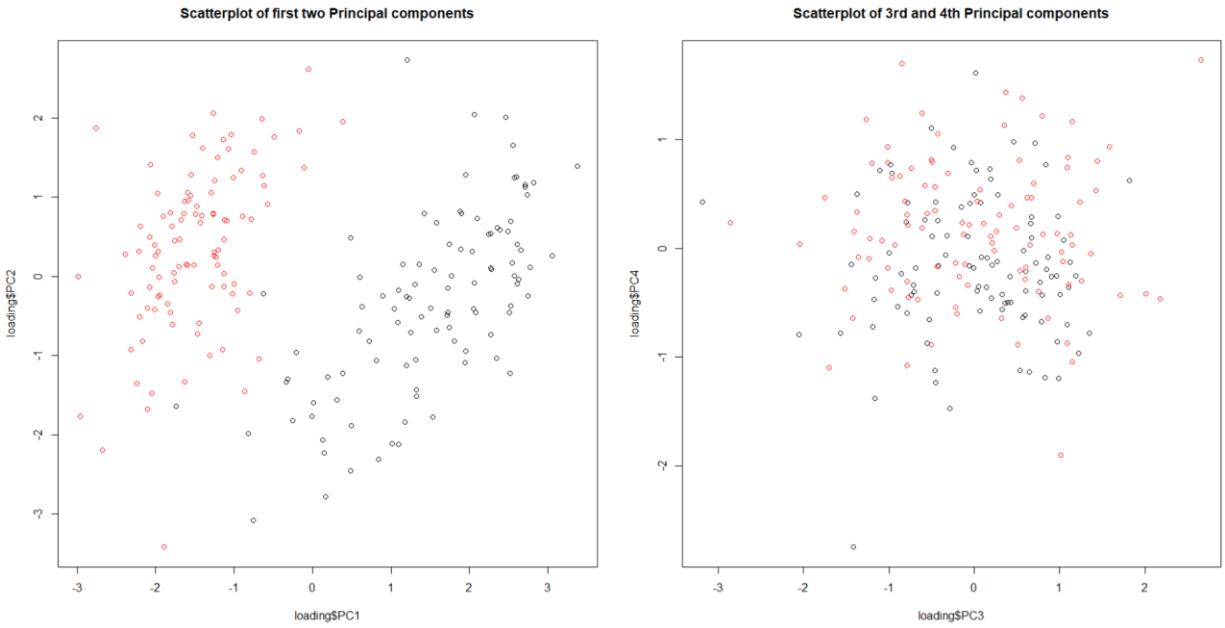
Running the PCA on the combined data gives,



- From this plot we can see the variances for all six principal components, and we can infer that the variance of first principal component is the highest and it gradually decreases.
- From this plot we can say that there is no mistake in the scaling during PCA.
- From the summary diagnostics we can see that the first principal component has explained almost 50% (49.09 %) of variance and the second principal component has described 21.30% variation and they cumulatively explained 70.39% of the total variation in the original data.

```
> summary(princip)
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
>
```

We then obtain the values of each principal component to plot the scatterplot of first four principal components.

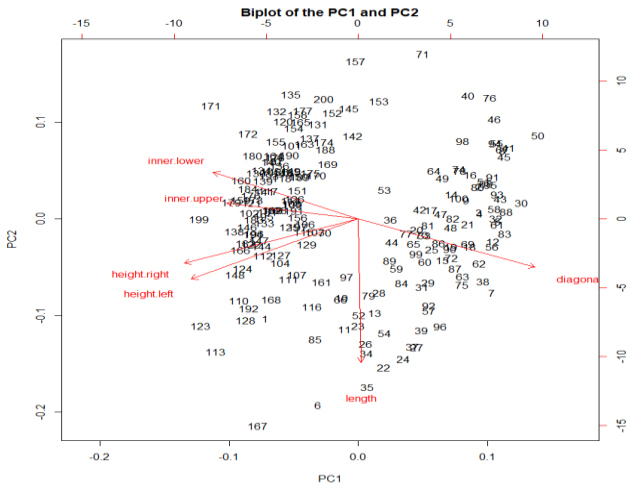


From the scatterplot of first two components,

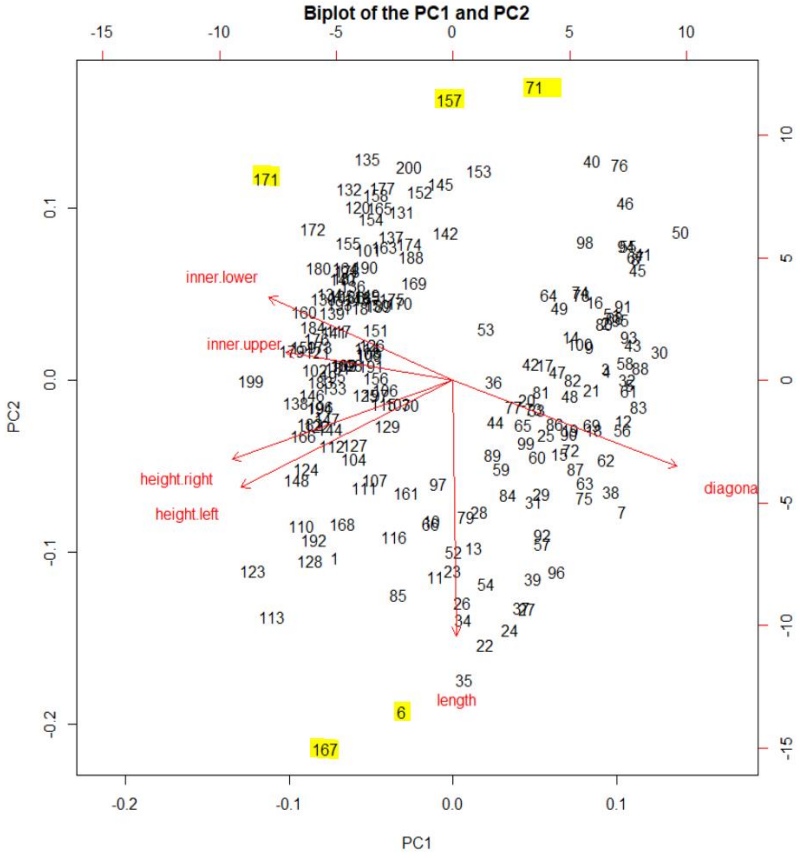
- we can easily observe that the two classes (real and fake notes) separate themselves among the first two principal components. **(Black = original notes, Red = Fake notes)**

Since the first two principal components show clear seperation and they account for the capturing most of the variation in data (70.39%), we **select the first two principal components** for analysis.

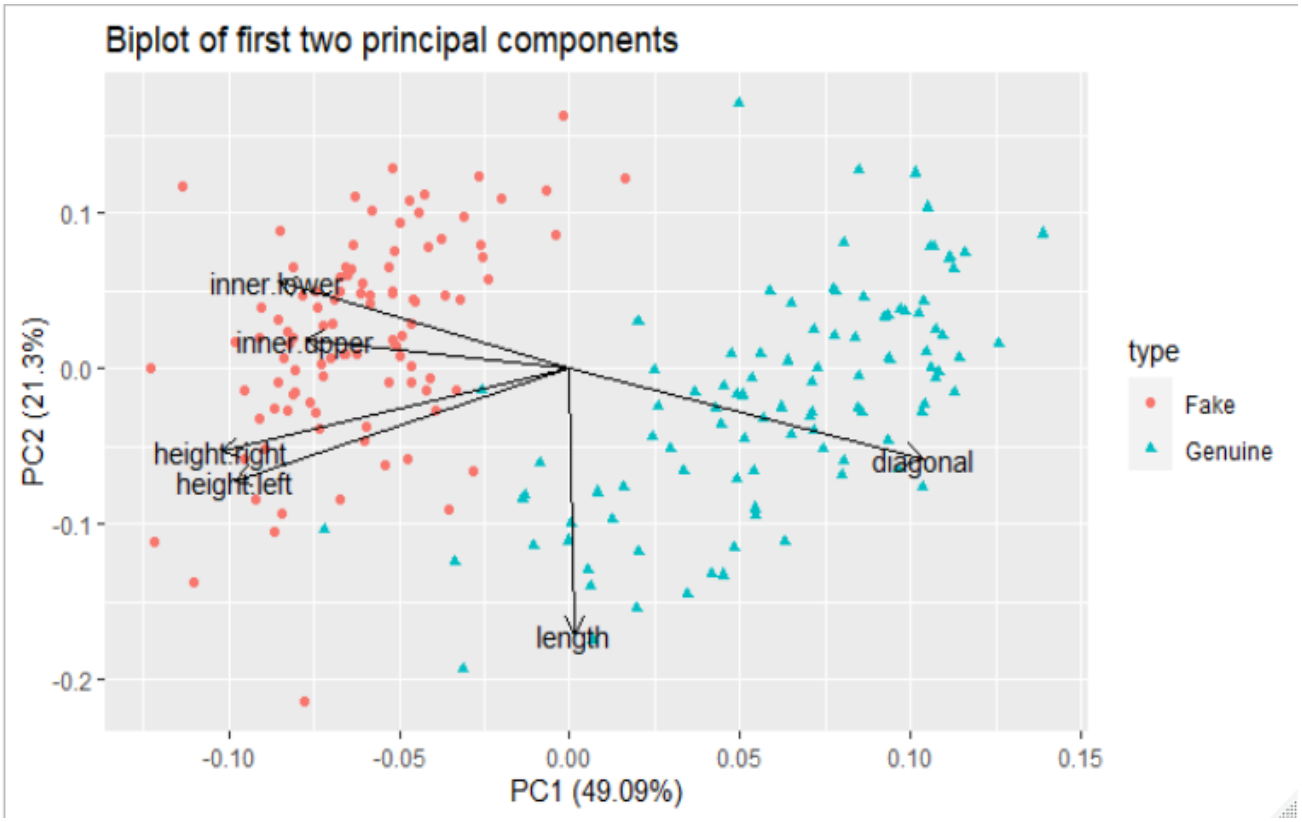
From the biplot of the first two principal components,



- We can see that the most **important contribution** for the **first principal component** is the **diagonal** value and the most important contributor for the **second principal component** is the **length** value.
- Also, we can see that the diagonal and inner lower values have strong negative correlation since they are in opposite directions and inner upper and length are having almost zero correlation.
- We can also see that the observations **167, 6, 171, 157, 71** are potential **outliers** in the data.



Plotting different classes for first two principal components in biplot,

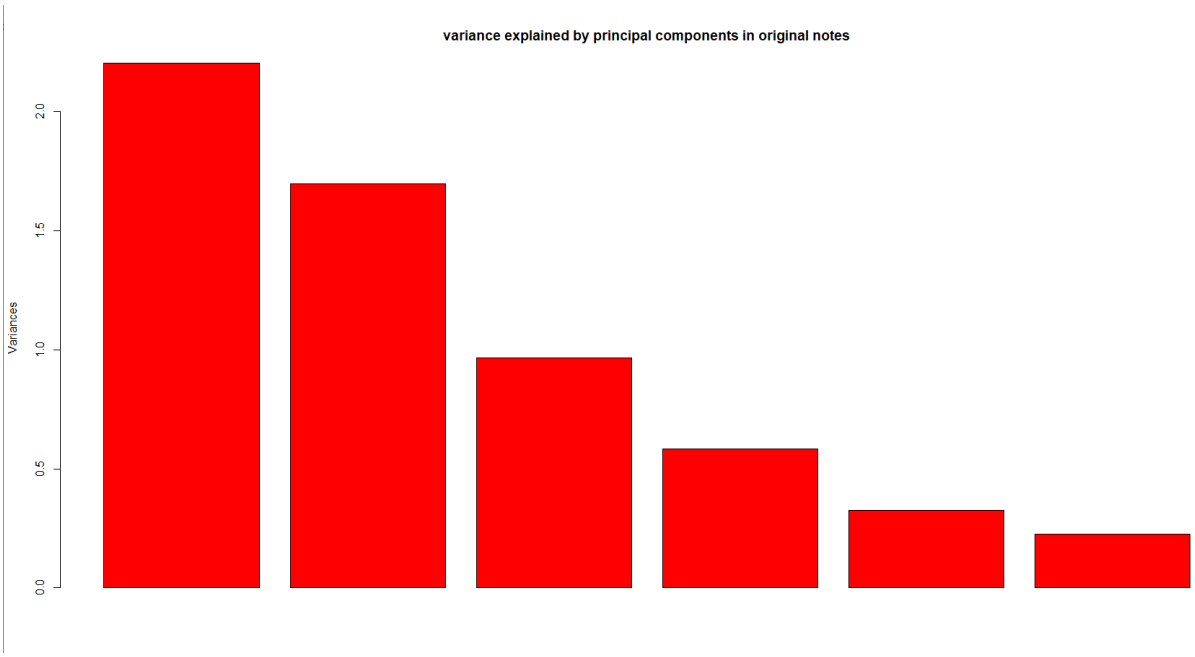


- We can easily see that the classes visibly separate across the length of the diagonal.
- The length of the original notes seems to have high diagonal values and the fake ones have comparatively small diagonal length values.
- Also, the fake ones seem to have high inner upper, inner lower, height right and height left values whereas the original ones have comparatively smaller values.

Running PCA on the genuine notes,

- We first scale the data using scale() function, since the variables do not have the same scale.

After scaling and running the principal component analysis,



- We can see the variance explained by all six principle components.

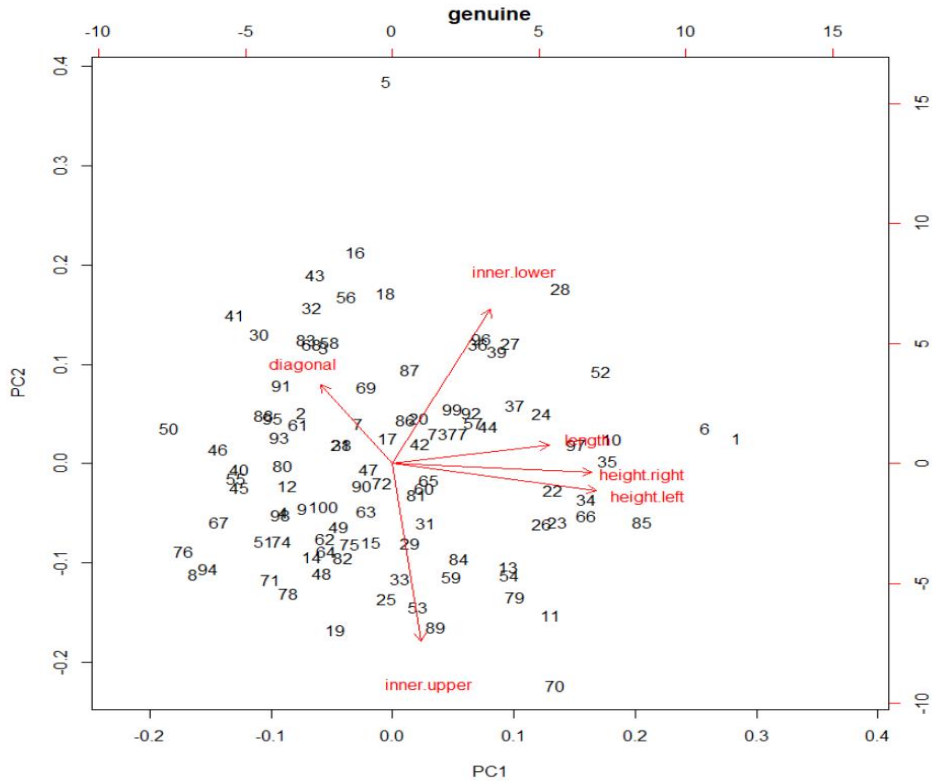
From the summary section we can see that,

```
> summary(principgenuine)
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
>
```

- Principal **component 1** describes **36.73%** variation and the principal **component 2** describes **28.28%** variation and they cumulatively describe 65% variation.

Since the first two principal components show clear seperation and they account for the capturing most of the variation in data (65.01%), we **select the first two principal components** for analysis.

Plotting the biplot of the first two principal components,



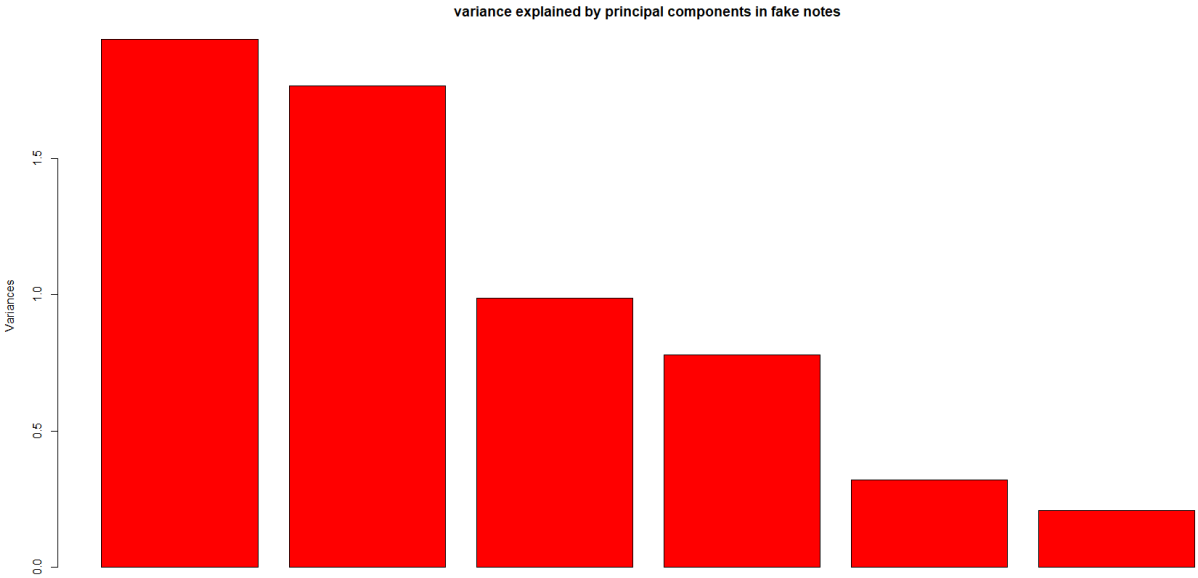
From the biplot we can see that,

- The length values and inner upper values have almost zero correlation.
- For principal component 1, the highest contributor is height left and height right values and for principal component 2, the highest contributor is inner upper and inner lower values.
- Also we can see that the observation 1, 6, 70 are potential outliers in the data.

Running PCA on the fake notes,

- We scale the data using scale() function since all the variable in the data is not in the same scale.

Running PCA on the scaled data,



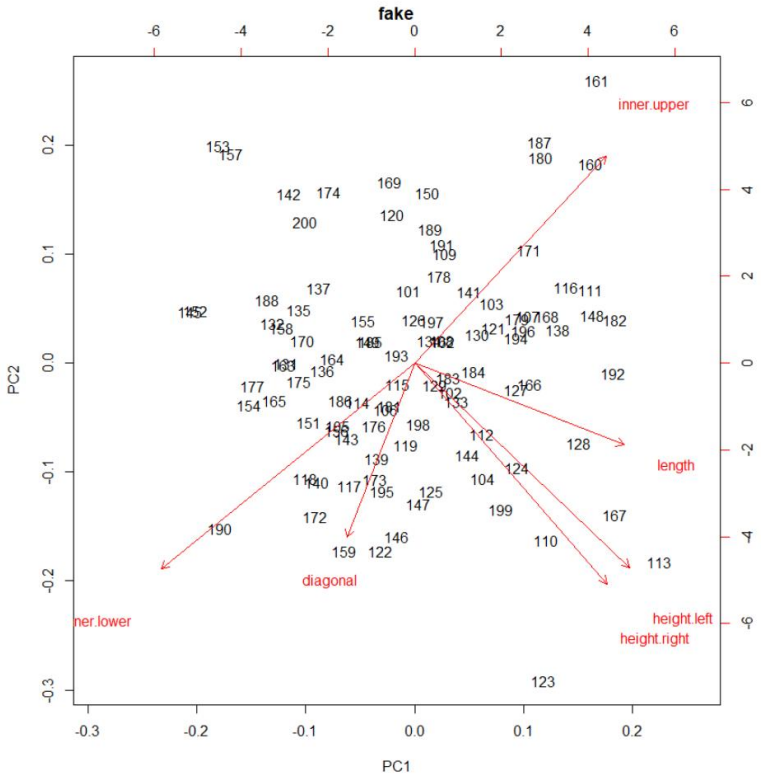
- From this plot we can see the variance explained by each of the six principal components.

```
> summary(principfake)
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
>
```

From the summary we can see that,

- The principal component 1 has explained 32.27 % and principal component 2 has explained 29.41% and cumulatively they have explained 61.69% variance in the data.

  Since the first two components account for capturing most of the variation in the data, we can select those two for analysis.



- From the biplot we can see that inner lower values and inner upper values are negatively correlated and height left and height right variables have almost zero correlation with inner upper and inner lower variables.
- Also, the length and diagonal variables have almost zero correlation.
- The observations 161, 123, 153, 157 are potential outliers.

Conclusion,

- We can see big difference when running PCA on the genuine and fake notes separately.
- When we run together, we can tell that the data itself separates between the genuine and fake notes whereas we are not able to tell them if we run separately.
- Also when running them together, we can see the first two PCs are able to capture almost 75% of variation in data whereas when we run separately they are able to capture little less variation in the data than the other.