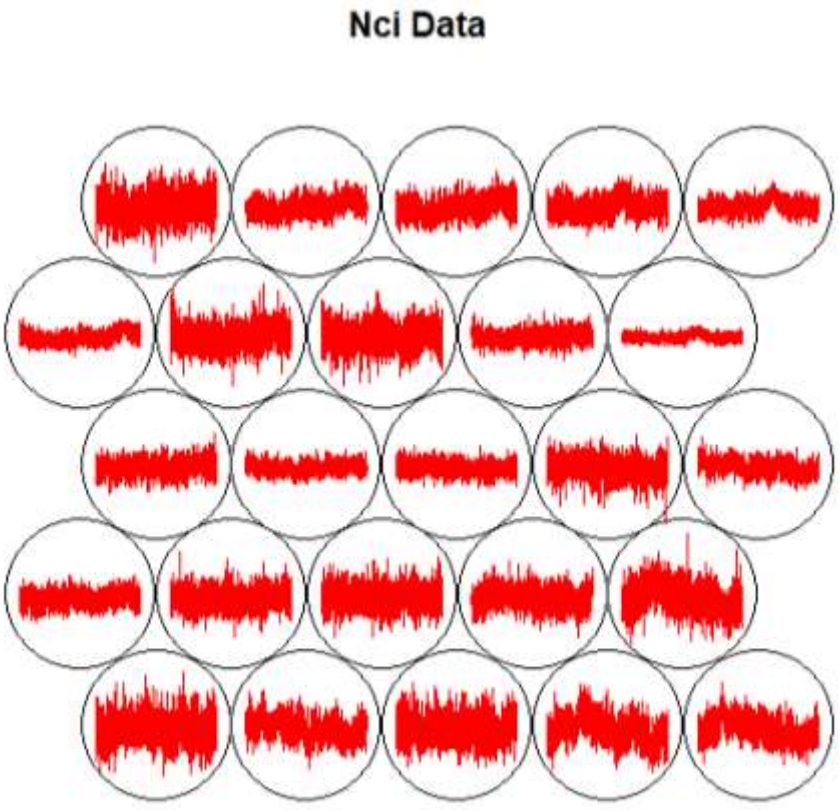
Statistical Data Mining II Project 3

Problem 1

In this problem I had to consider the tumor microarray data in the package library(ElemStatLearn). The data consists of several different types of tumor samples. I observed that in many clustering algorithms there are often found to be 2-3 groups/clusters in this Ill-studied data, although there are 14 subtypes of tumor cells, as shown using (unique(colnames(nci))), as there are some repeated column names.

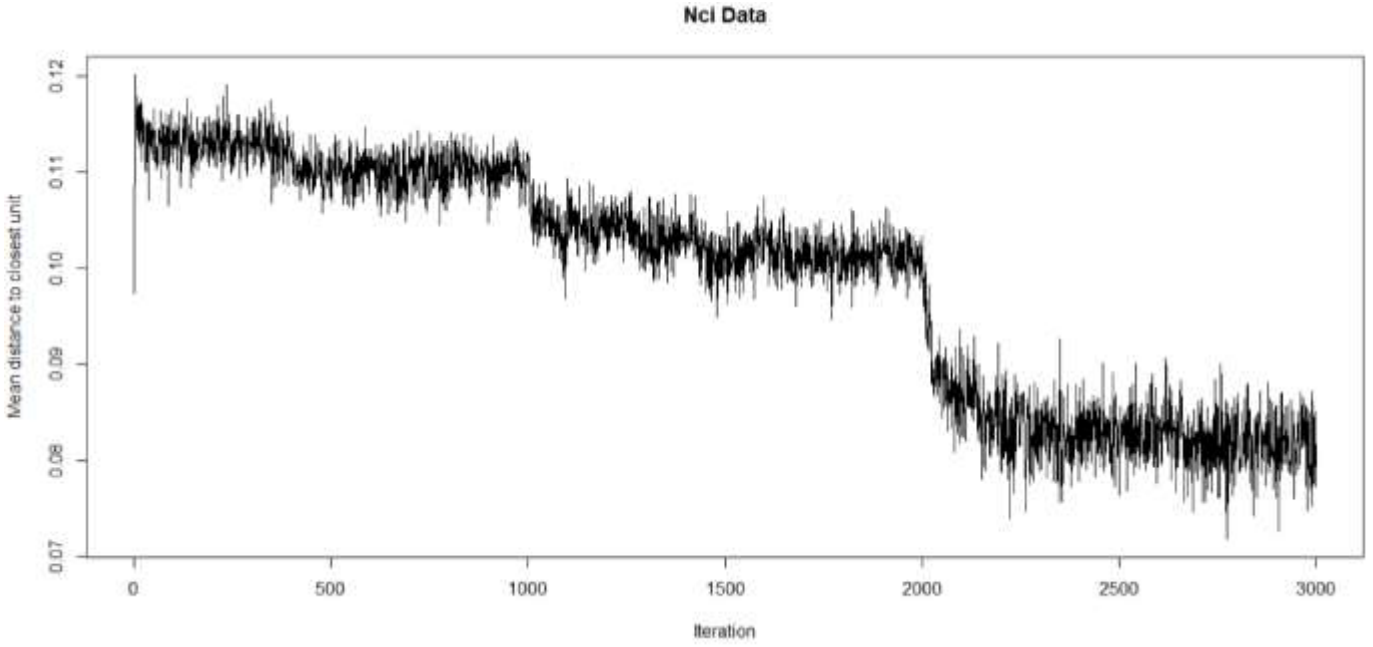
I had to run a SOM algorithm. First I scaled the nci data by centering and dividing by std i.e. finding the z-score of columns. This makes everything on same scale for SOM and then I transposed the data i.e. swapped the rows and columns.

I plotted the SOM, with the default code type plot. I plotted a 5 by 5 hexagonal grid, with all the 64 variables and all code book vectors for each prototype showing only magnitude, as shown below. In the SOM plot, because I have a lot of variables/features (64 variables) I are not able to see pie charts and make it difficult to interpret. Yet, from the plot I can see that the code books look very similar with some porotypes more flat than others.

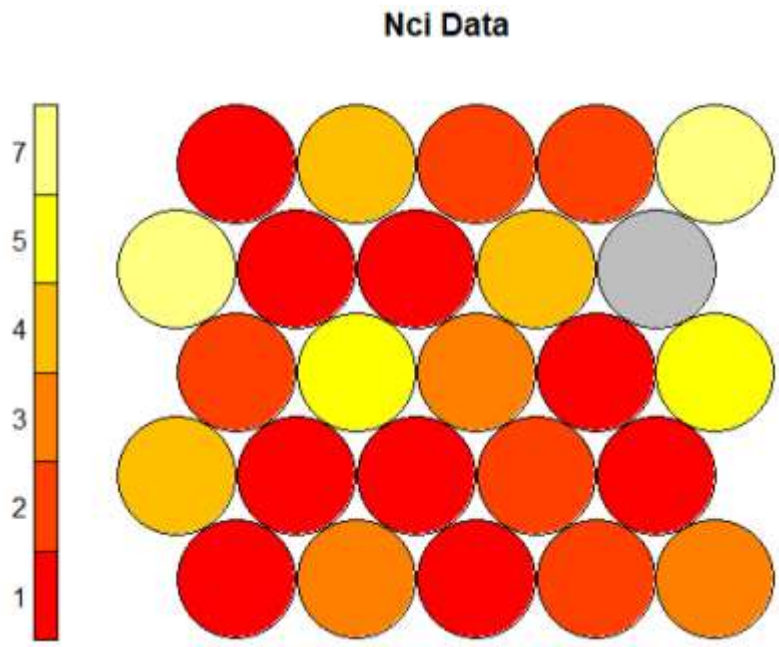


I also plotted the SOM with the changes plot type. This gives the relative changes of the code book vectors as procced through the 3000 iterations, this tells us whether we are converging.

The graph below shows the average distance for the closest unit for all the different distances between the assigned data points and the prototypes. From the graph we can see that the mean distance to the closest unit/prototype decreases and converges, with 3000 iterations.

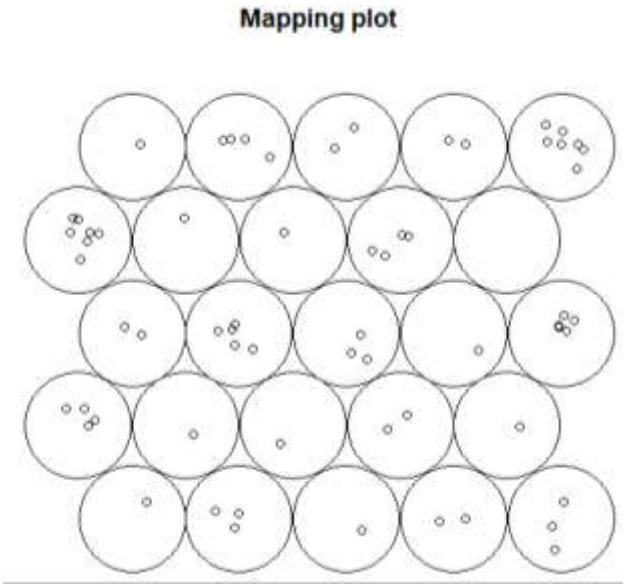


Afterwards, I plotted the SOM with the count type plot to get a better idea of the features driving the SOM. This tells us how many different items map into the different units. This gives us an impression of the distribution of the data points over where they map over the grid. Each node/prototype color gives the average of the observations that are ending up in that prototype over all the 64 variables. We can see that most of data points are in the yellow to peach prototypes, as they have around 5-7 data points.



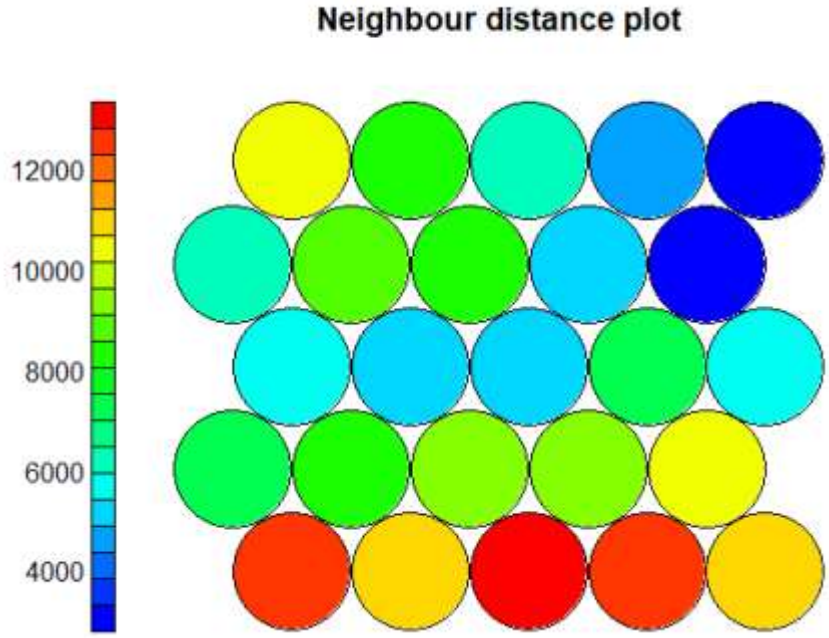
Afterwards, I plotted an SOM plot using the mapping type plot. This is similar to the plot with color mapping above, but it allows us to count the dots which represent the data points in the NCI data. This gives an idea where points sit with respect to the prototypes.

In the nodes where we have 2 or more types of cancers , each which is 6,000 long, the porotype is the is an avenge of the data points being mapped to it.

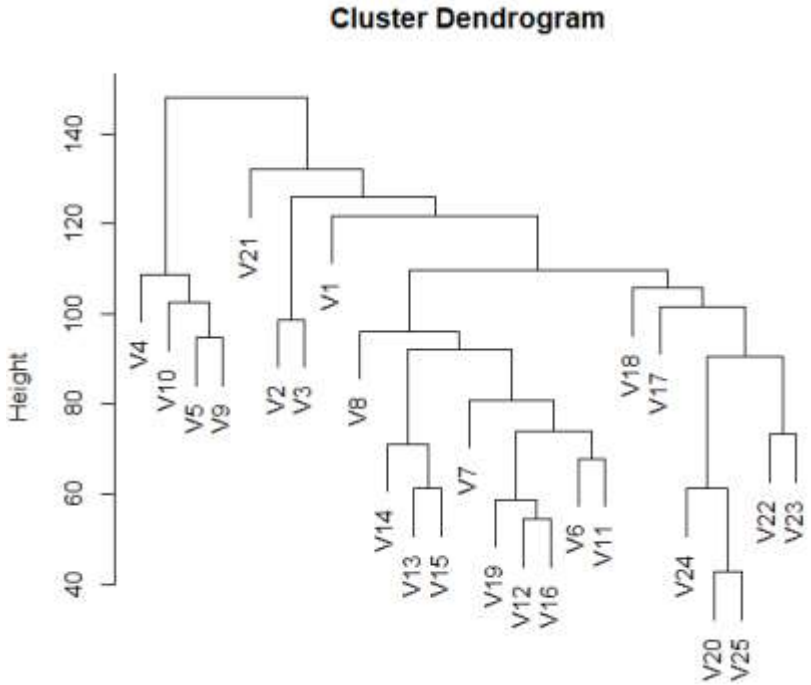


Afterwards, the neighbor distance plot was plotted as it gives impression of distance to neighbors, where blue are closest ones and the red dots are very different from the neighbors. From the plot we can see the green are the closest ones in the lower and upper

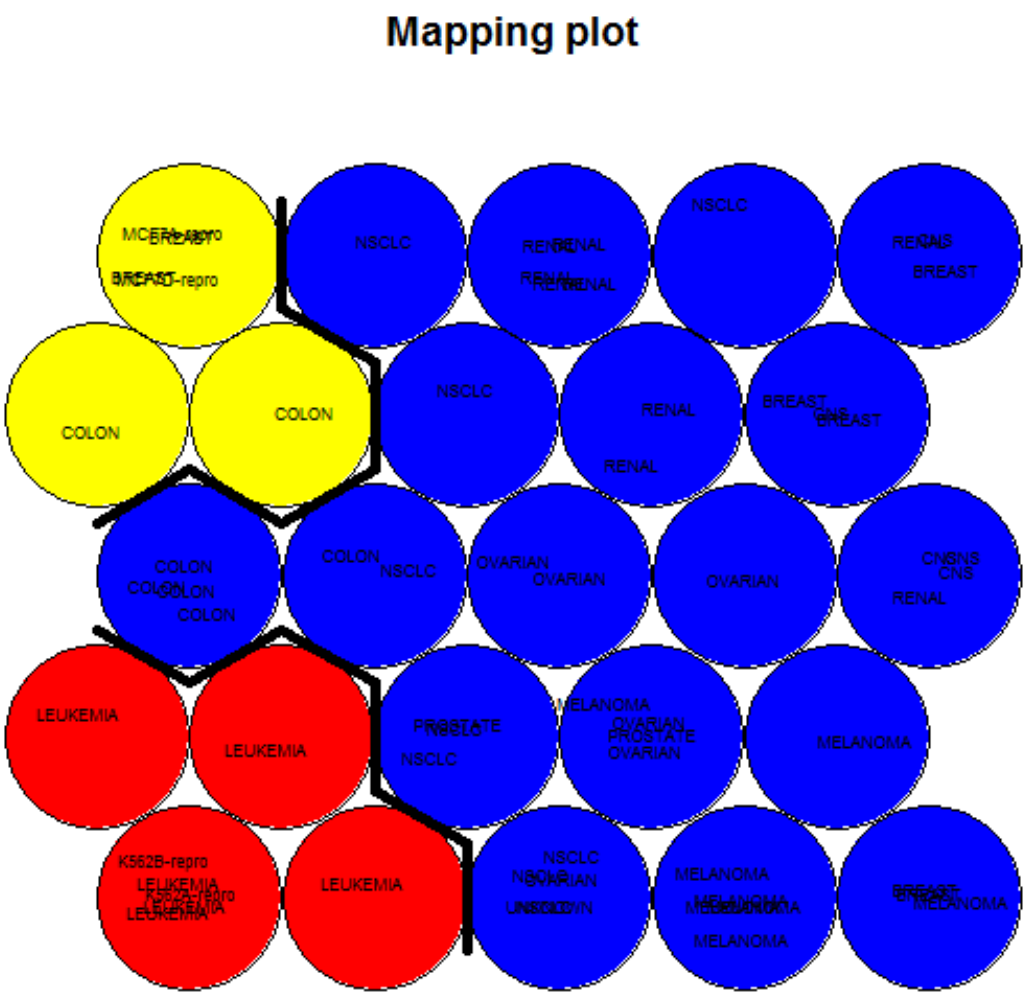
left hand corner. Also, we see that the blue porotypes in the right hand upper corner are similar, while the red are highly dissimilar prototypes in the lower bottom.



The results were presented with hclust on prototypes. In this part I performed hierarchical clustering of the observations using complete linkage. Euclidean distance was used as the dissimilarity measure. I clustered the cacner types as shown below in the cluster dendrogram, derived from the dissimilarity between the code book vectors/prototypes.



I later cut at the height to give 3 clusters/classes at height of 130 as this is where there are we can see the broad shoulders in the dendrogram. Also, this is supported as there are 3 groups/clusters in this well-studied data. We plotted the SOM with the found clusters, as shown below.



The above U matrix looks at how similar the prototypes are in value and looks at Euclidian distance between prototypes. Big Euclidian distance means the bigger the difference between the prototypes. The colored bright spots in the U matrix help to determine the groupings. Each of the node has a prototype associated with it and each porotype has a number of variables associated with it. The convergence is the average of the variables.

The SOM method characterizes the tumor cells well into groups. We can see from the 5 by 5 grid of SOM that we get 3 clusters of tumor cell groups, and this is what we expect as there are 3 groups/clusters in this well-studied data.
We have enough points in the prototypes to estimate good dissimilarity between them, empty prototypes are different in their profiles from the rest and with 5 by 5 we have more spread with the original points compared to smaller grids.

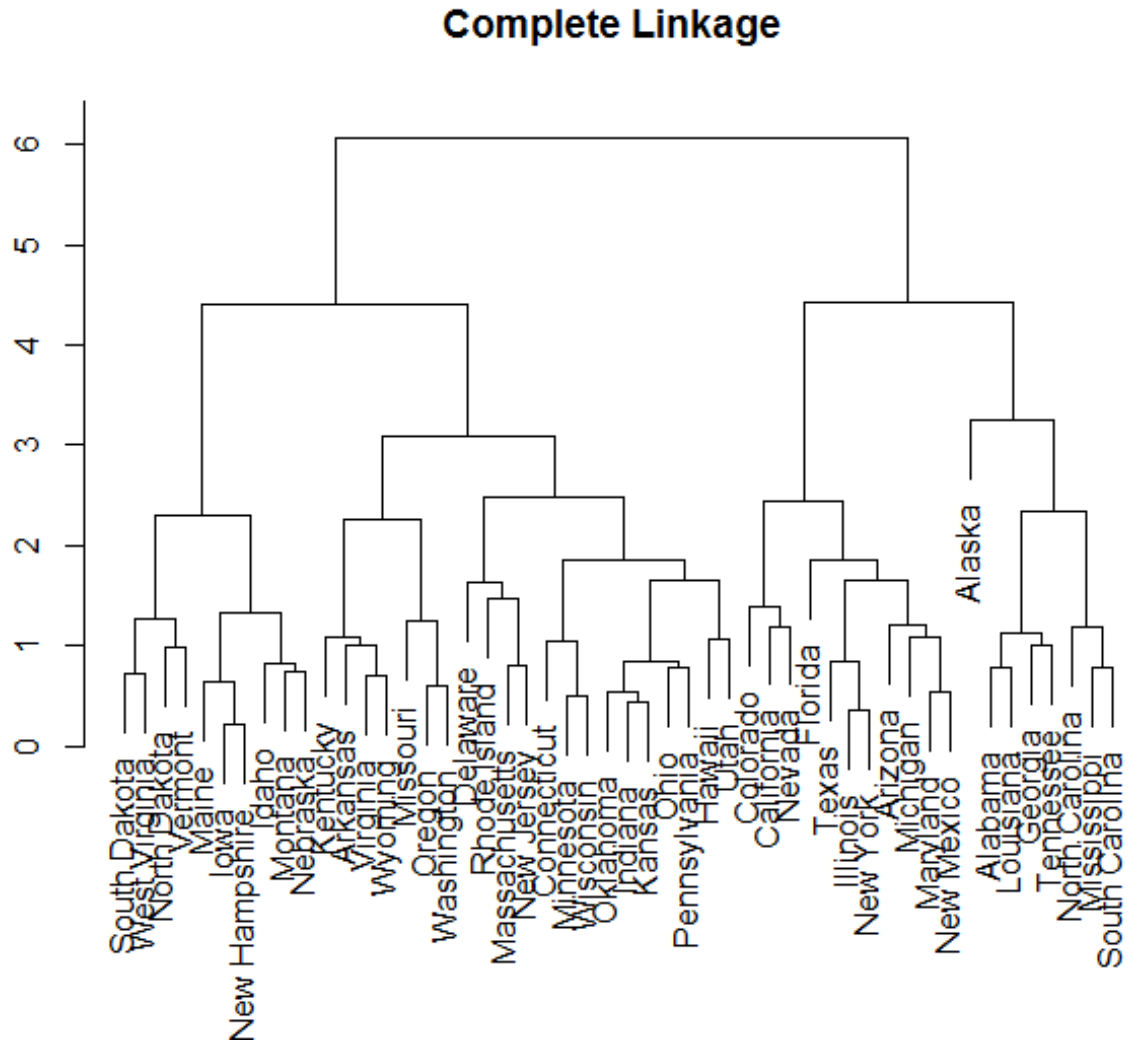
We are trying to understand the groupings in terms of their original variable, and we know the labels and we are trying to recover them. In this case we have leukemia cluster mainly in the red cluster, the melanoma, ovarian, renal, NSCLC, breast, prostate, CNS cancers are clustered mainly in the blue cluster and the colon, K562B.repro cancer is mainly clustered in the yellow clusters.

Problem 2

In this problem I had to consider the USArrests data to perform clustering. First I had to scale the data by centering and divide by standard deviation i.e. determining the z-score of columns. This was done so everything on same scale for clustering using SOM.

Part A

In this part we had to perform hierarchical clustering with complete linkage and Euclidean distance to cluster the states. By looking at the structure of the dendrogram, I cut it at a height that resulted in three clusters, as this is the height where the dendrogram has broad shoulders and big gaps of no connections and this is where we get big groups rather than small groups.



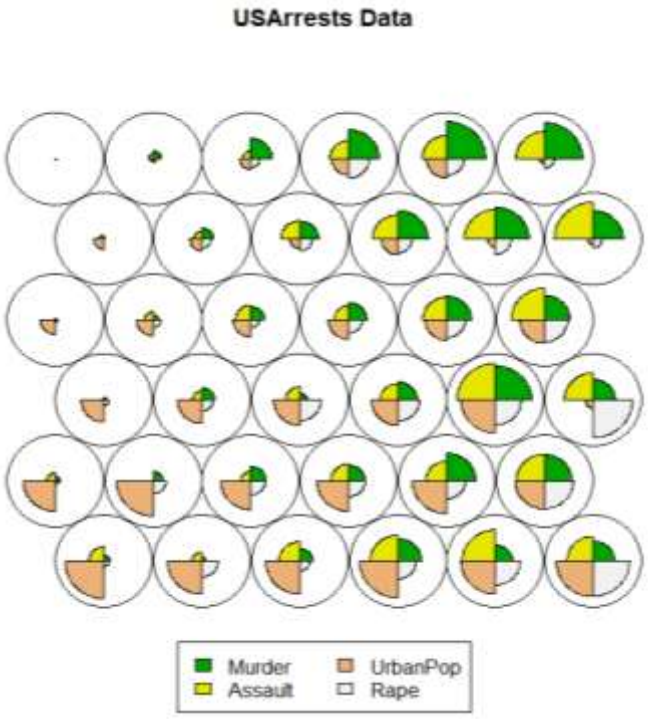
From the hierarchical clustering, we get the following four clusters shown below, each with their corresponding US states.

```
> #Cluster1
> cluster.data[which(cluster.data$four.clusters == 1),]$state.name
[1] Alabama        Alaska         Georgia        Louisiana      Mississippi    North Carolina South Carolina
[8] Tennessee
50 Levels: Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware Florida ... Wyoming
> #Cluster2
> cluster.data[which(cluster.data$four.clusters == 2),]$state.name
 [1] Arizona    California Colorado   Florida    Illinois   Maryland   Michigan   Nevada     New Mexico
[10] New York   Texas
50 Levels: Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware Florida ... Wyoming
> #Cluster 3
> cluster.data[which(cluster.data$four.clusters == 3),]$state.name
 [1] Arkansas      Connecticut   Delaware      Hawaii        Indiana       Kansas        Kentucky
 [8] Massachusetts Minnesota     Missouri      New Jersey    Ohio          Oklahoma      Oregon
[15] Pennsylvania  Rhode Island  Utah          Virginia      Washington    Wisconsin     Wyoming
50 Levels: Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware Florida ... Wyoming
> #Cluster 4
> cluster.data[which(cluster.data$four.clusters == 4),]$state.name
 [1] Idaho         Iowa          Maine         Montana       Nebraska      New Hampshire North Dakota
 [8] South Dakota  Vermont       West Virginia
50 Levels: Alabama Alaska Arizona Arkansas California Colorado Connecticut Delaware Florida ... Wyoming
```

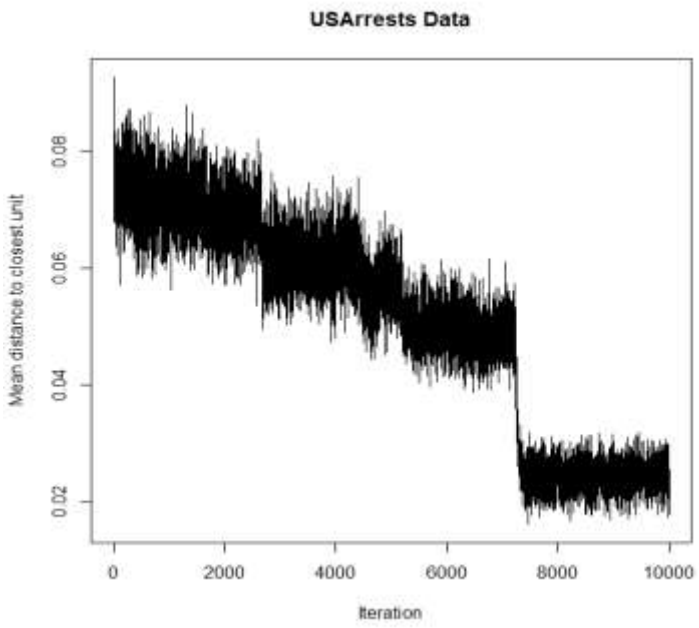
Based on the US Census Bureau; there are four regions of the US: the Northeast, the Midwest, the South, and the West. I would expect the US states to cluster based on their geographic location, as shown in the map below. We expect similar crime trends in states closer to each other. Cluster 1 has AL, GA, LA, MS, NC, SC and TN clustered correctly according to green cluster in the map. Cluster 2 has CA, NM, CO, NV that are correctly clustered as shown in the grey cluster in the map below. Cluster 3 has IN, WI, KS, MO, MN that are correctly clustered as shown in the blue cluster in the map. Cluster 4 has ME, NH, VT that that are correctly clustered as shown in the orange cluster in the map below. Therefore, the hierarchical clustering does cluster some of the states based on the crime rates somewhat based on their location.

PART B

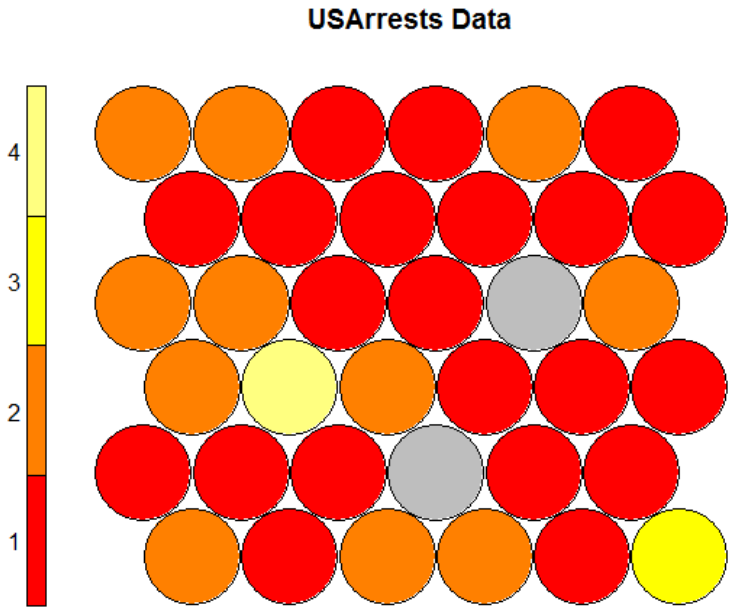
In this part we had to fit a SOM to the US Arrest data and present the results obtained. We plotted the SOM, with the default code type plot. I plotted a 6 by 6 hexagonal grid, with all the 4 variables and all code book vectors for each prototype showing only magnitude, as shown below. In the SOM plot, we see that the prototype values are pie chart, this describes the prototypes. On the left side we see that the peach is mostly dominant. We see that on the top right side, the green and yellow together dominate and as we move to the lower right corner, the yellow, green and peach and white variables dominant the prototypes.



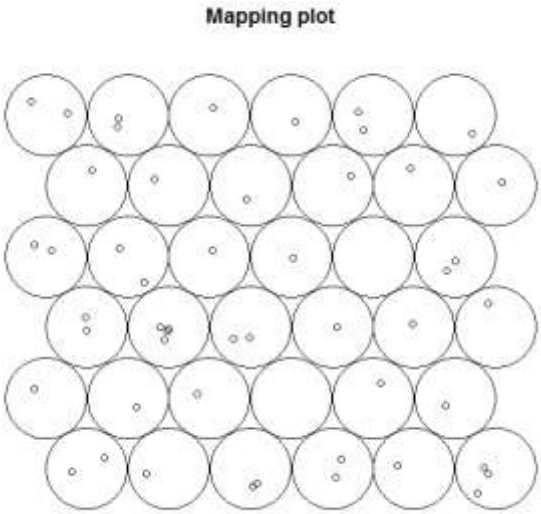
I also plotted the SOM with the changes plot type, as shown below. This gives the relative changes of the code book vectors as procced through the 1000 iterations, and tells us whether we are converging. From the graph we can see that the mean distance to the closest unit/prototype decreases and converges, with 1000 iterations.



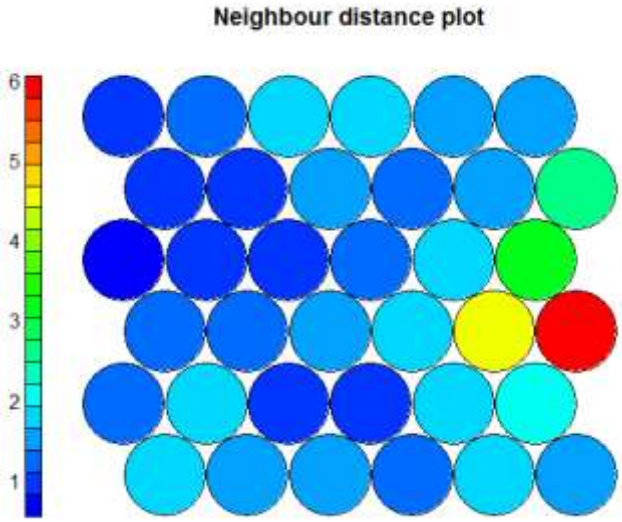
Afterwards, we plotted the SOM with the count type plot to get a better idea of the features driving the SOM. This tells us how many different items map into the different units. This gives us an impression of the distribution of the data points over where they map over the grid. Each prototype color gives the average of the observations that are ending up in that prototype over all the 4 variables. We can see that most of data points are in the yellow prototypes, as they have around 3-4 data points.



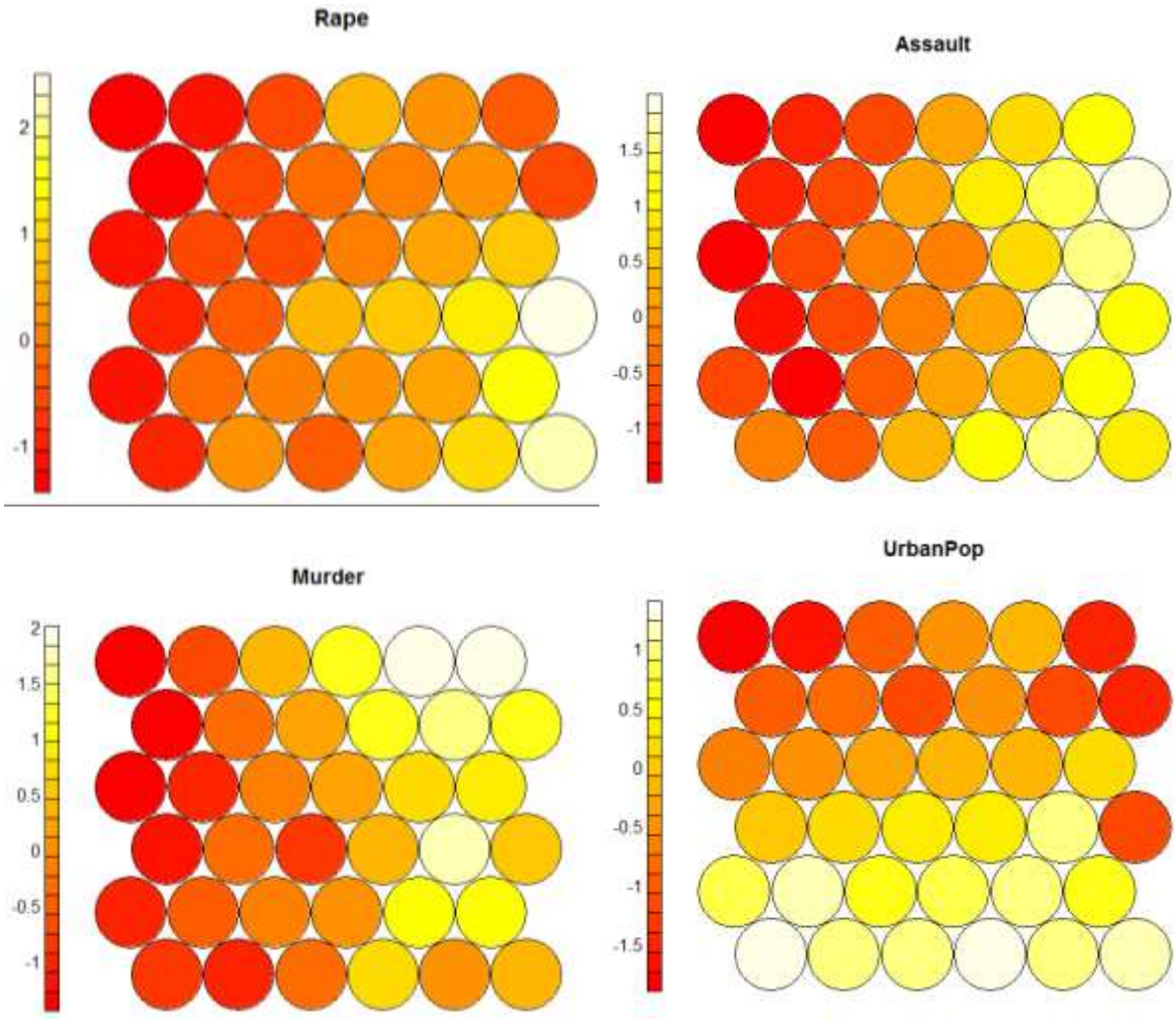
Afterwards, I plotted an SOM plot using the mapping type plot. This is similar to the plot color mapping above, but it allows us to count the dots which represent the data points in the US Arrest data. This gives an idea where points sit with respect to the prototypes.



Afterwards, the neighbor distance plot was plotted as it gives impression of distance to neighbors, where blue are closest ones and the red dots are very different from the neighbors. From the plot we can see the blue prototypes are dominating and they are the most similar. With the 2 green prototypes which are less similar in the right side, while the yellow and red are highly dissimilar prototypes in the center right.

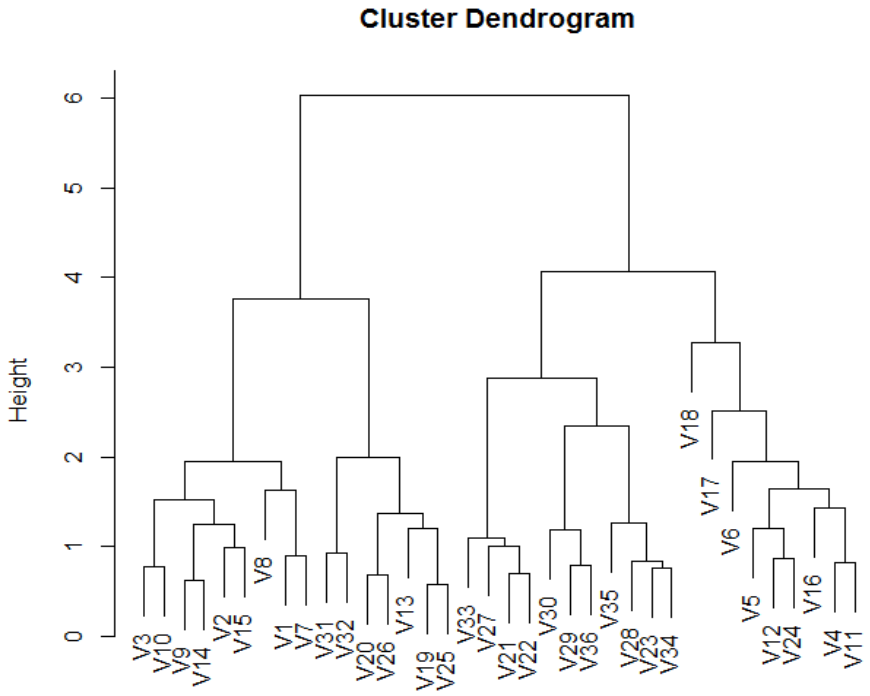


I later plotted a component plane plots to visualize original data over the SOM .Since there are 4 variables, we plotted 4 SOM component plane plots to see how the variables change over the grid.

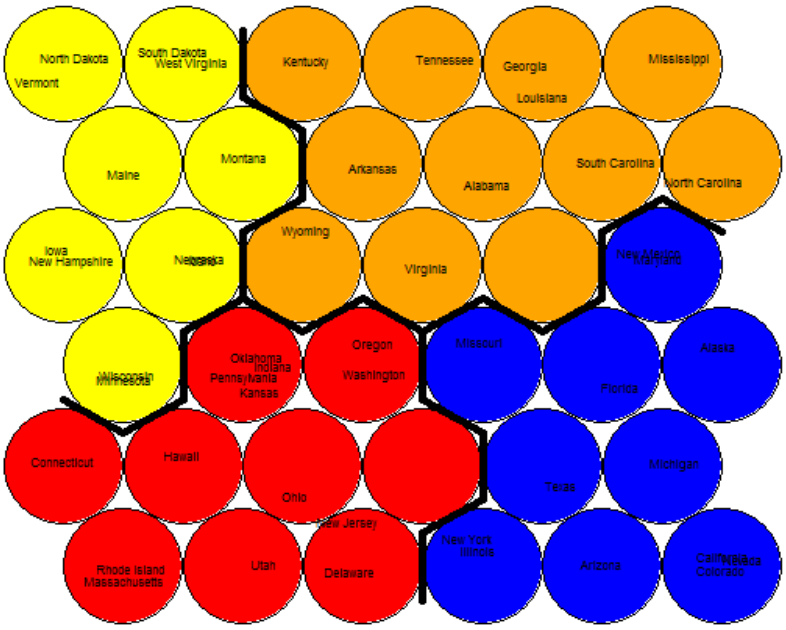


We see that the assault and murder variables both have the data points spread over red and yellow prototype. With the yellow dominating i.e. more data points, which have closer Euclidian distance to the prototype value on the right side of the grid, whereas rape variable has mostly red prototypes spread, with the yellow dominating only the lower right corner. Urban pop has the red spread mostly in the upper right and left side and the yellow prototype spread in lower right and left.

After performing the SOM and assigning the 50 states to each prototype (6x6=36 prototypes) , I performed hierarchical clustering of the observations using complete linkage on the prototype, with the Euclidean distance used as the dissimilarity measure.
I cut at the dendrogram at height of 3.9 to give 4 clusters/groups, as this is where dendrogram has broad shoulders and big gaps of no connections and this is where we get big groups rather than small groups.



Mapping plot



Based on the US Census Bureau; there are four regions of the US: the Northeast, the Midwest, the South, and the West. I would expect the US states to cluster based on their geographic location, as shown in the map below. We expect similar crime and urban population trends in states closer to each other. Orange cluster in the SOM has AL, GA LA, MS NC, SC, TN, WV, KY which correctly to the green region in the map below. Furthermore, the blue cluster in the SOM has MI, MO, IL which map the blue region in the map and the red cluster in the SOM has WA, OR, HI, UT, which maps to the grey region in the map. The yellow cluster in SOM has ME, NH, VT which map into the orange region in the map; therefore these are the results that are expected, however both miss out the states which have to be included in their clustering.

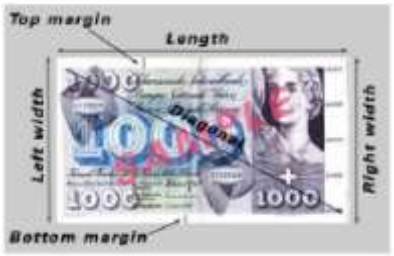
These results generally support the results in part A. The SOM cluster has similar clustering to the hierarchical clustering. Both had AL, GA, LA, MS, NC, SC and TN clustered correctly according to green cluster in the map, but SOM had WV ad KT as well. The SOM cluster had WA, OR, HI, UT, whereas hierarchical clustering of the states had CA, NM, CO, NV clustered as in the grey cluster in the map. The SOM cluster had MI, MO, IL whereas, hierarchical clustering had IN, WI, KS, MO, MN clustered in the blue cluster in the map. Lastly, both the SOM clustering and hierarchical clustering had ME, NH, VT as states which map onto the orange region in the map.

PART C

The advantages of hierarchical clustering to SOM, is that easy to implement and outputs visualization called a dendrogram. Based on the dendrogram structure (i.e. big gaps and broad shoulders in dendrogram) we select the number of clusters. However, this is not preferred when the data is typically in high dimensions and the data cannot be visualized in a plot like the dendrogram and these results in poor results.
On the other hand, an SOM uses high dimensional feature space and maps it down to a 2D grid to define the clusters. SOM also provides a nice visualization like hierarchical clustering where we get dendorgram.
However, SOM also has a special representation, as we can interpret the cluster in the context of the original variables. When using hierarchical clustering we get dendorgram, but it is not obvious why the group is splitting from this group, as we cannot take this back to the original set of measurements for which we have computed the dissimilarity, but we can do this with SOM. Using the U-matrix helps visualize the original set of variable on the SOM which makes it unique.

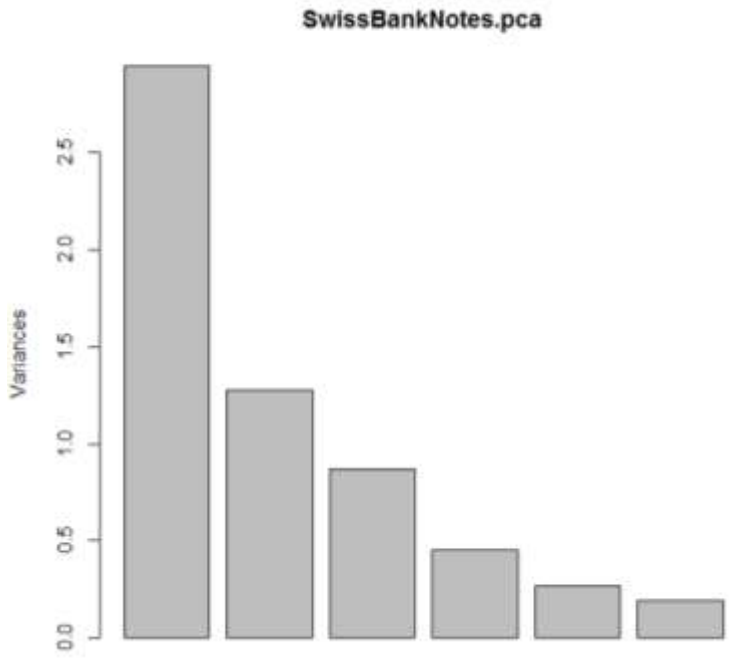
Problem 3

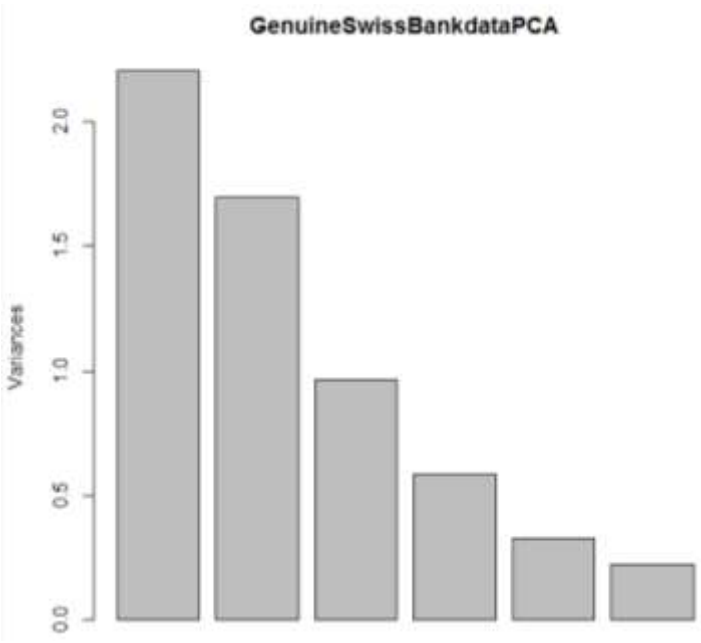
In this problem I had to access the SwissBankNotes data. The data consists of six variables measured on 200 old Swiss 1,000-franc bank notes. The first 100 are genuine and the second 100 are counterfeit. The six variables are length of the bank note, height of the bank note, measured on the left, height 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, as shown below.

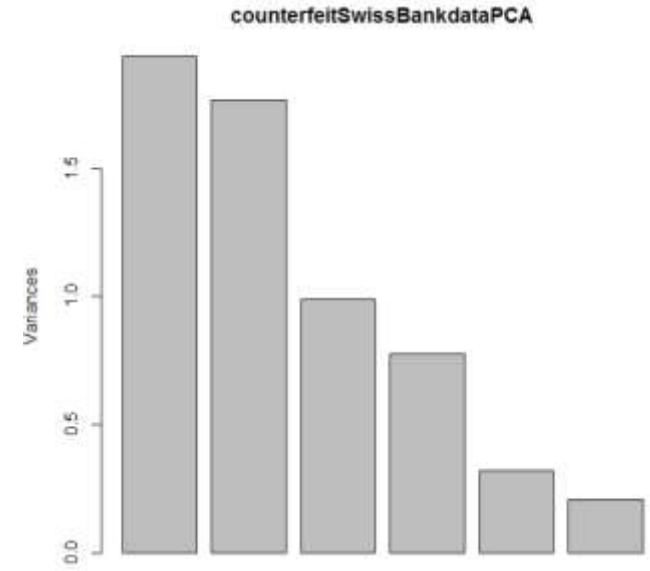


We had to carry out a PCA of the 100 genuine bank notes, of the 100 counterfeit bank notes, and all of the 200 bank notes combined. We ran the PCA alogirthm and generated the following pca plots to see which principal compenents contribute to the largest variation/variance in the data set (the x-axis has the principal compenets PC1 to PC6).

The PCA plots for the combined SwissBanknotes, Genuine Banknotes and Counterfietabnk notes are shown below.





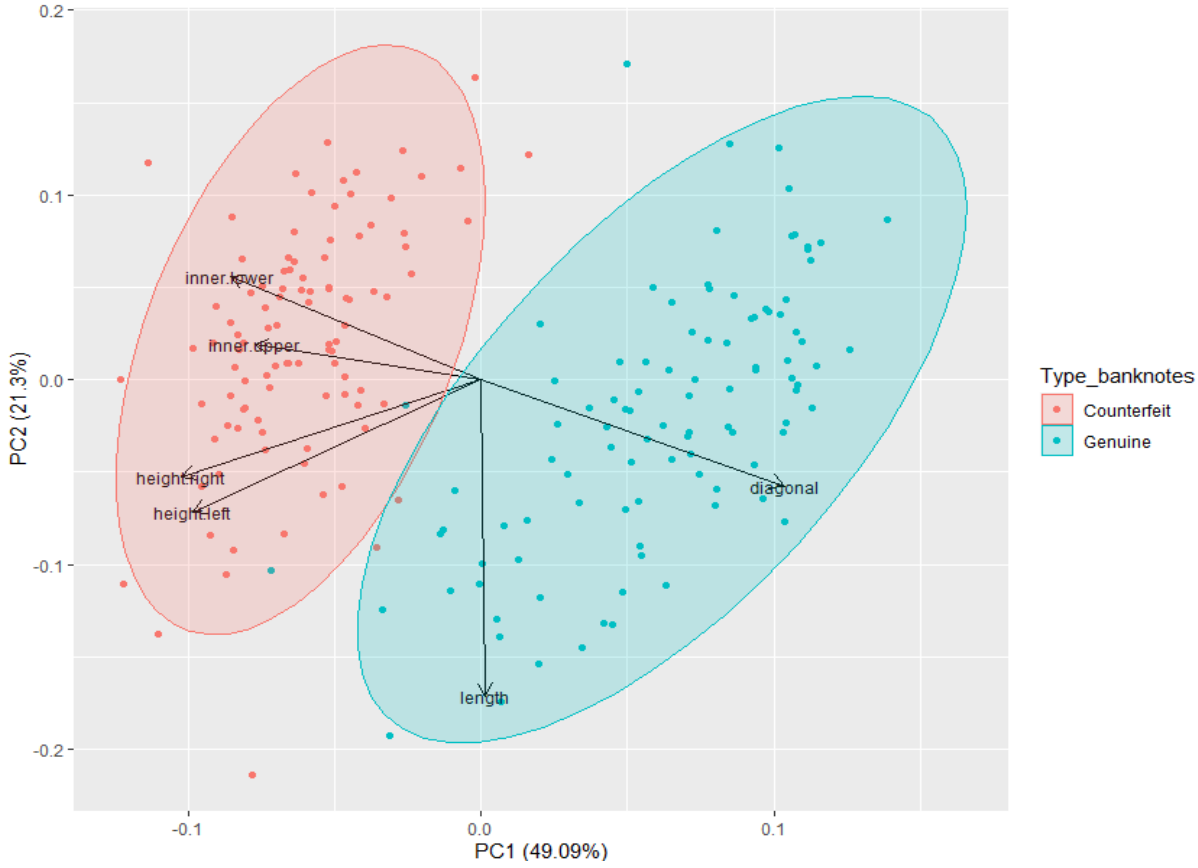


From the above plots, we can see the elbow structure in from PC1 to PC2 occurring in the combined SwissBankNotes. This shows that in the combined data, the PC1 has largest variance in the data. From the counterfeit and genuine Bank Notes PCA plots, we can see the elbow structure occurs from P2 to PC3. This shows that PC1 and PC2 describe the most variation in the data set and thus I selected these principal Components to produce the PCA biplots.

I decided to use biplots as diagnostic tools because Principal Component Biplots can be used to detect multivariate outliers, determine correlation between variables, and used to determine which variables contribute to the largest source(s) of variation in the dataset. They are used to visualize

variation of data and they show observations (principal component score) and variables (principal competent loading) simultaneously. The axis is usually the first componentsPC1 and PC2 as they represent the most variation in a data set. The observations that are extreme are not clustered around 0 and are majorly separated away from their variable/loading dataset i.e. are extremely above or below the average values of the loading would be outliers in biplot.

The biplot for the for all the 200 combined SwissBanknotes, 100 Genuine Banknotes and 100 Counterfietabnk notes we plotted as shown below.

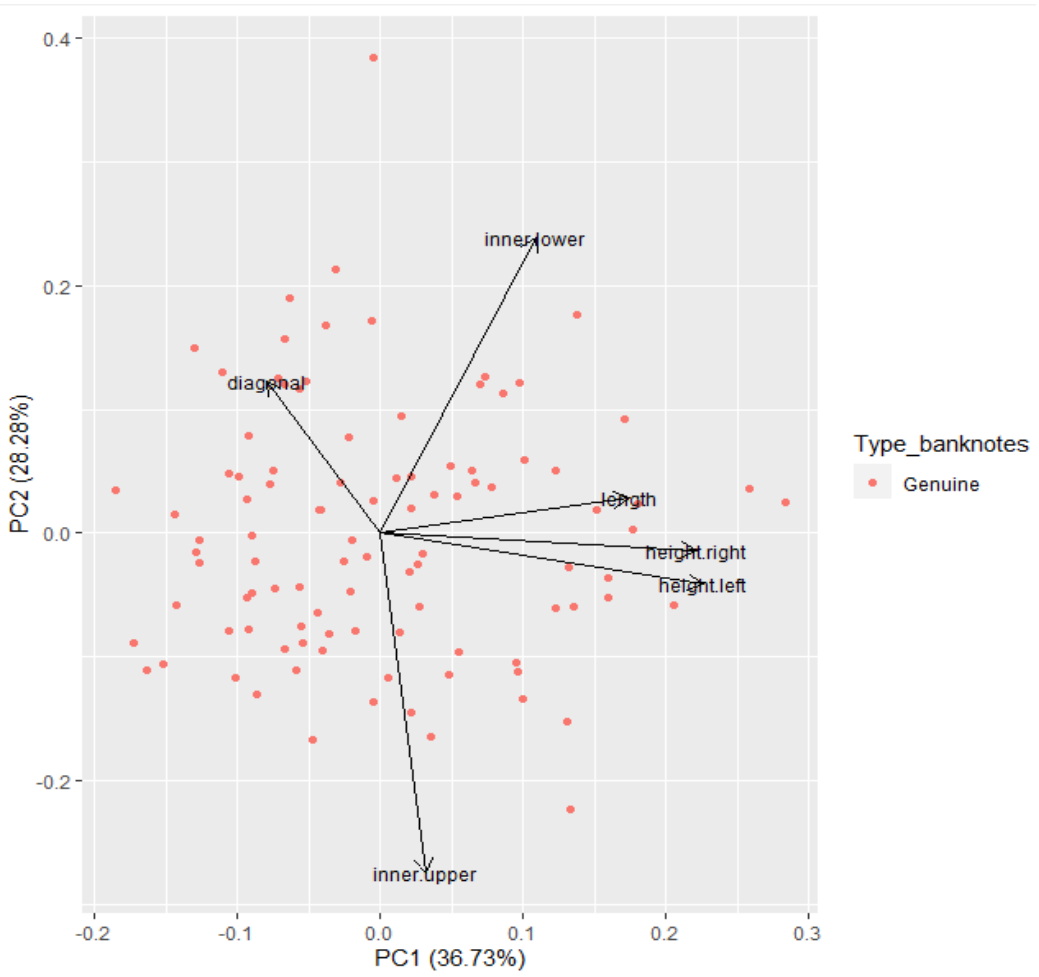


From the 200 combined bank notes biplot above, we can see that the loadings in the biplot are the variables of the dataset, which are length, diagonal, heightleft, heightright, innerupper and innerlower. From the biplot we can see that the height right and height left are positively correlated and innerupper and inner lower are also positively correlated, while length is not correlated to diagonal or height or inner as it is perpendicular to them. Also, diagonal is not correlated with height of the bank note and is negatively correlated with inner bank note. The variables contribute to the largest source of variation in the dataset are diagonal, heightright, and height left, this is because there are the variables that have most extreme values for PC1 (x-axis) of about 0.1 (in terms of magnitude).

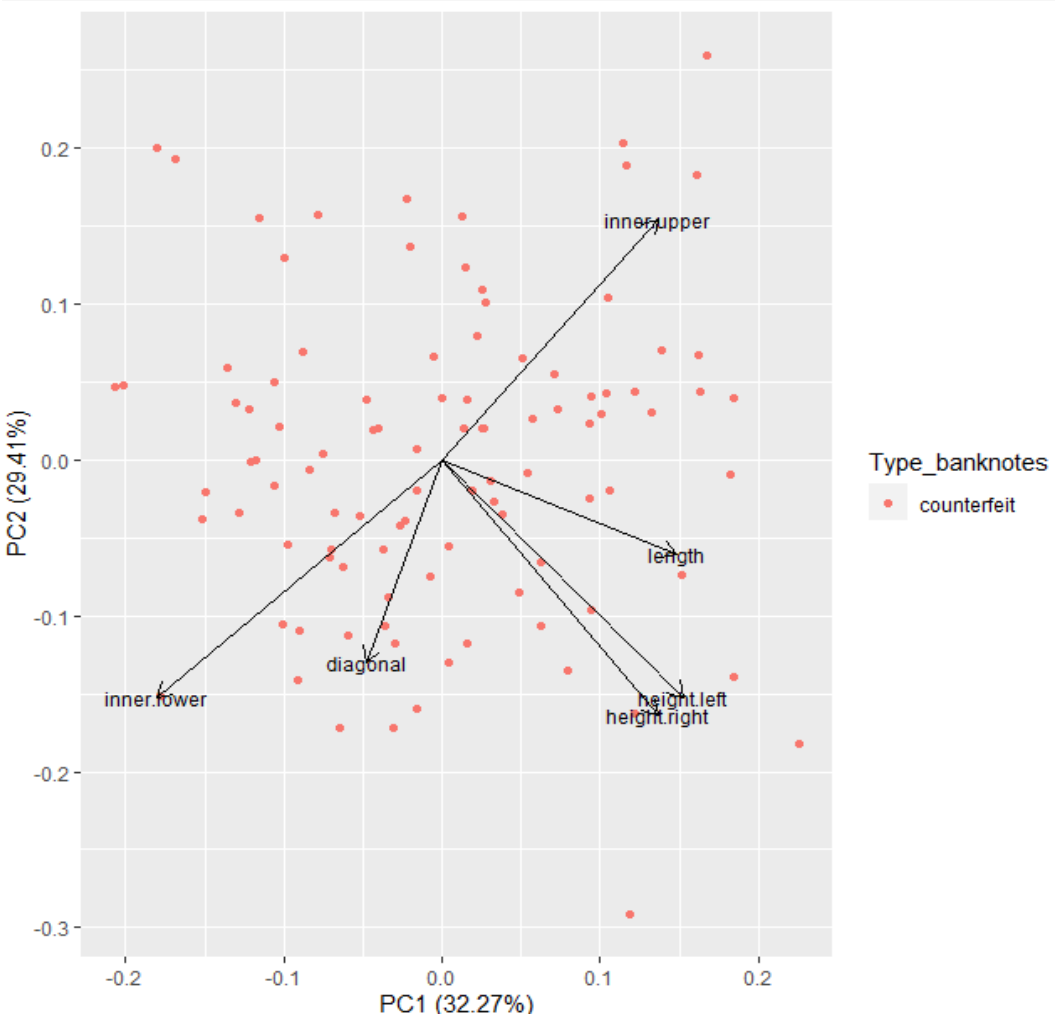
From the biplot we can see that the coiunterfiet and the genuine bank notes observations separate well into two groups. The cointerfiet bank notes has more data points which are not

clustered/overlapping other observations and are not clustered around zero. This is due to the factors/variable loadings of height.right and height left. These observations are opposite of the direction of the loading height.right and height left and are below the average value. The counterfeit observations have average values of length and diagonal variable. On the other hand, the genuine bank notes have observations more clustered and overlapping compared to the counterfeit bank notes. The genuine bank note observations have average values for height.left, height.right, inner.upper and inner.lower, as they are clustered around them, but below average values for the diagonal and length variables.

The biplot for 100 genuine bank notes was plotted as shown below. There are differences compared to the 200 bank notes combined biplot. The inner.lower are not positively correlated but nearly negatively correlated. The length and diagonal still have no correlation; and the height right and height lower still are positively correlated. The observations are more clustered and overlapping, with data points having average values for the variables, expect for some observations, which have the diagonal values as above average value (as the data points are beyond the diagonal arrow, both in positive and negative direction). The variables contribute to the largest source of variation in the dataset are height.right, and height left, this is because there are the variables that have most extreme values for PC1 (x-axis) of about 0.2 (in terms of magnitude).



The biplot for 100 counterfeit bank notes was plotted as shown below. There are differences compared to the 200 bank notes combined biplot. The inner.lower are not positively correlated but nearly negatively correlated. The length and diagonal still have no correlation; and the height right and height lower still are positively correlated. The observations are still less clustered, more spread out and less overlapping. The data points have above and below average values for the nearly all the variables e.g. inner.upper, height left and height.right, diagonal and length as they observations lie above the arrow in the positive and negative direction. The variables contribute to the largest source of variation in the dataset are diagonal, inner.lower, height.right, and height left, this is because there are the variables that have most extreme values for PC1 (x-axis) close to 0.2 (in terms of magnitude).


0.2
0.1
0.0
-0.1
-0.2
-0.3
PC2 (29.41%)
-0.2
-0.1
0.0
0.1
0.2
PC1 (32.27%)
inner.upper
length
diagonal
inner.lower
height.left
height.right
Type_banknotes
counterfeit