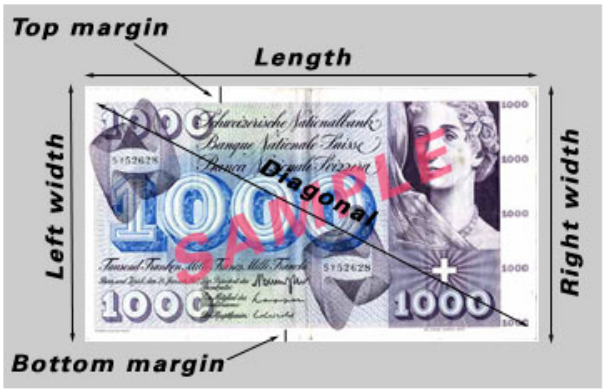
Assignment 3

Question 1

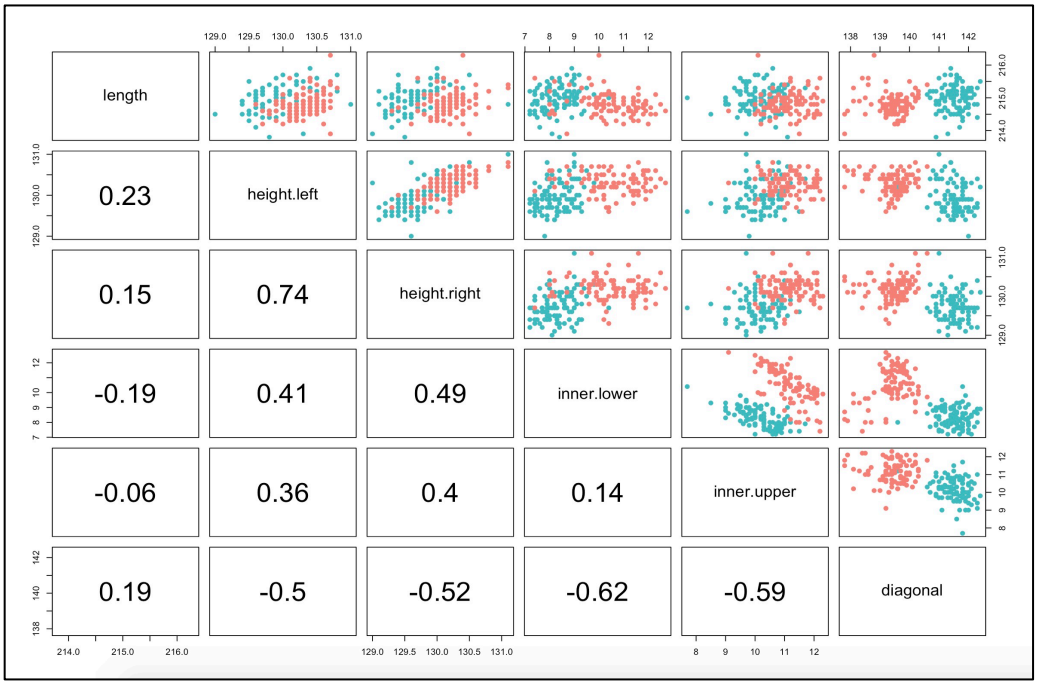
Access the SwissBankNotes data (UB learns). The data consists of six variables measured on 200 old Swiss 1,000-franc bank notes. The first 100 are genuine and the second 100 are counterfeit. The six variables are length of the bank note, height of the bank note, measured on the left, height of the bank 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. Do you notice any differences in the results? Show all work in the selection of Principal Components, including diagnostic plots.



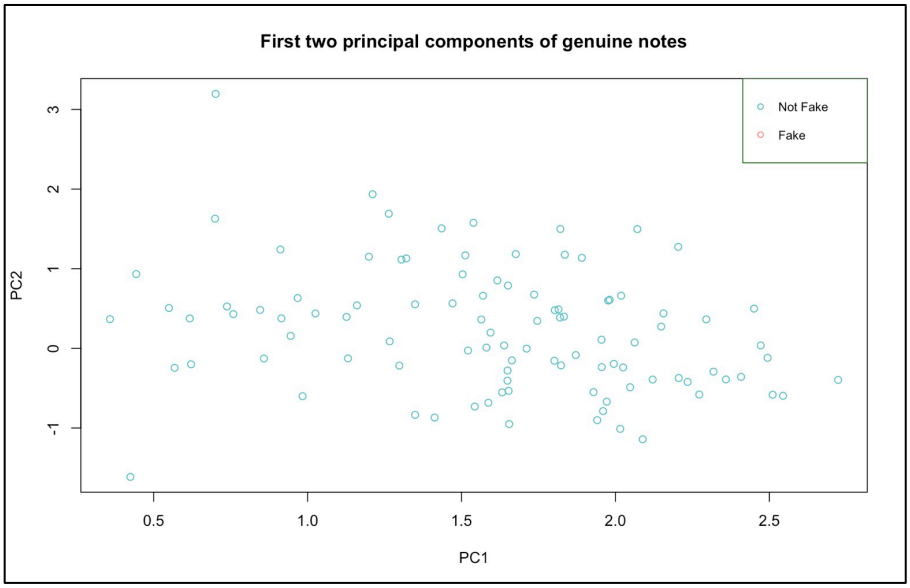
Summary of data:

```
> summary(SwissBankNotes)
     length        height.left     height.right    inner.lower      inner.upper       diagonal      is.fake
 Min.   :213.8   Min.   :129.0   Min.   :129.0   Min.   : 7.200   Min.   : 7.70   Min.   :137.8   0:100
 1st Qu.:214.6   1st Qu.:129.9   1st Qu.:129.7   1st Qu.: 8.200   1st Qu.:10.10   1st Qu.:139.5   1:100
 Median :214.9   Median :130.2   Median :130.0   Median : 9.100   Median :10.60   Median :140.4
 Mean   :214.9   Mean   :130.1   Mean   :130.0   Mean   : 9.418   Mean   :10.65   Mean   :140.5
 3rd Qu.:215.1   3rd Qu.:130.4   3rd Qu.:130.2   3rd Qu.:10.600   3rd Qu.:11.20   3rd Qu.:141.5
 Max.   :216.3   Max.   :131.0   Max.   :131.1   Max.   :12.700   Max.   :12.30   Max.   :142.4
```

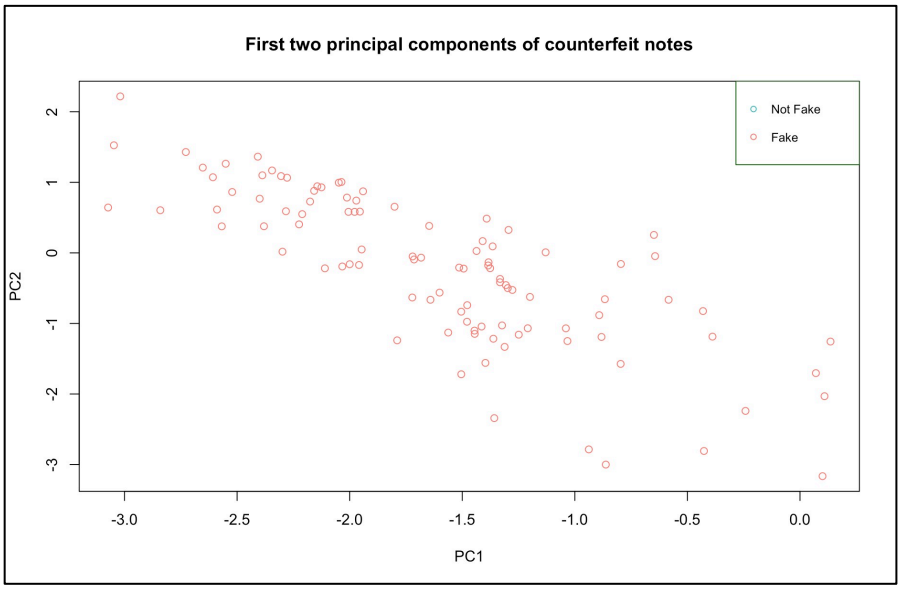
Correlation and Pairs plot for data



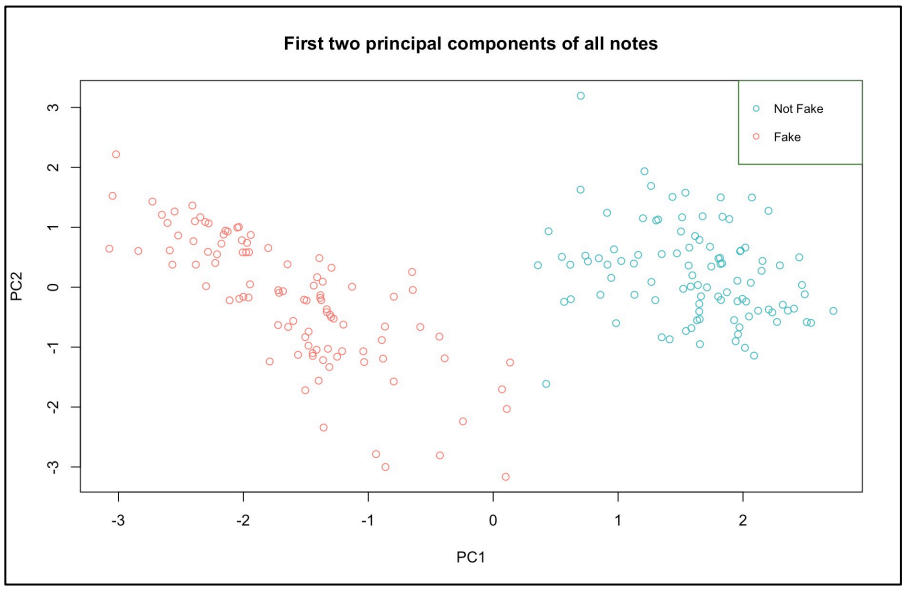
PCA for first 100 notes



PCA for next 100 notes



PCA for all notes



The plot of first two principal components shows clear distinction between genuine and fake notes.

Question 2

Access the data “primate.scapulae” (on UB learns).

a) Cluster the data based on single-linkage, average linkage, and complete- linkage agglomerative hierarchical clustering. Decide on the groupings, and justify it in words, for all three methods. Calculate the misclassification rate for all three methods. Which method performed the best and which method performed the worst? Was the result in line wit your expectations?

b) Cluster the data based on K-means **or** K-medoids. Explain your choice for the number grouping K and calculate the misclassification rate. How did the performance compare to the hierarchical clustering of part a? Which did you feel was a better method for this data?

Clustering result:

cutTreeDt	1	2	3	4	5
1	5	0	0	0	0
2	2	0	4	6	0
3	9	0	0	0	0
4	0	15	0	0	40
5	0	0	16	8	0

Complete

cutTreeDt	1	2	3	4	5
1	13	0	0	0	0
2	2	0	0	0	0
3	1	0	0	0	0
4	0	15	0	0	40
5	0	0	20	14	0

Single

cutTreeDt	1	2	3	4	5
1	13	0	0	0	0
2	2	0	0	0	0
3	1	0	0	0	0
4	0	15	0	0	40
5	0	0	20	14	0

Average

Single & Average clustering perform better.

Kmeans clustering result:

	1	2	3	4	5
1	2	0	9	10	0
2	0	15	0	0	40
3	8	0	0	0	0
4	1	0	11	4	0
5	5	0	0	0	0

K means perform worse than hierarchical clustering for the data.

Question 3:

Write programs to implement K -means clustering and a self-organizing map (SOM), with the prototype lying on a two-dimensional grid. Apply them to the columns of the human tumor microarray data, using K = 2, 5, 10, 20 centroids for both. Demonstrate that as the size of the SOM neighborhood is taken to be smaller and smaller, the SOM solution becomes more similar to the K -means solution.

Similarity between K-Means and SOM for different values

K Means K	SOM Radius	Similarity	K Means K	SOM Radius	Similarity
2	0.1	0.9671702	2	2	0.9584356
5	0.1	0.606243	5	2	0.6066932
10	0.1	0.4918334	10	2	0.4602055
20	0.1	0.3866018	20	2	0.3897954
2	0.4	0.9788491	2	4	0.9797093
5	0.4	0.6030985	5	4	0.6528898
10	0.4	0.444196	10	4	0.4730389
20	0.4	0.3701455	20	4	0.3989395
2	0.8	0.9640614	2	10	0.9711419
5	0.8	0.6426914	5	10	0.6206909
10	0.8	0.4411828	10	10	0.439368
20	0.8	0.3981515	20	10	0.3982031

The values in yellow highlight the similar clustering between two algorithms.