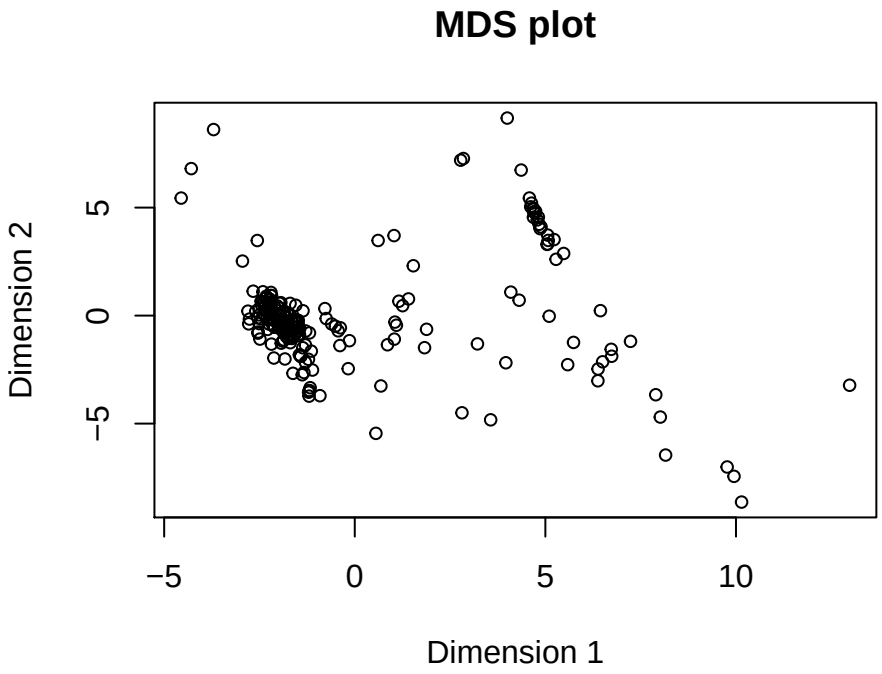
ex5_Palombarini

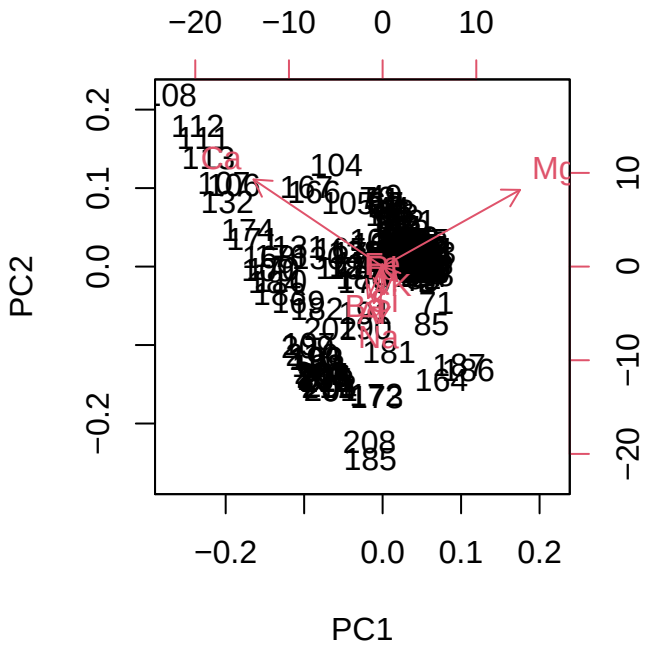
jacopo

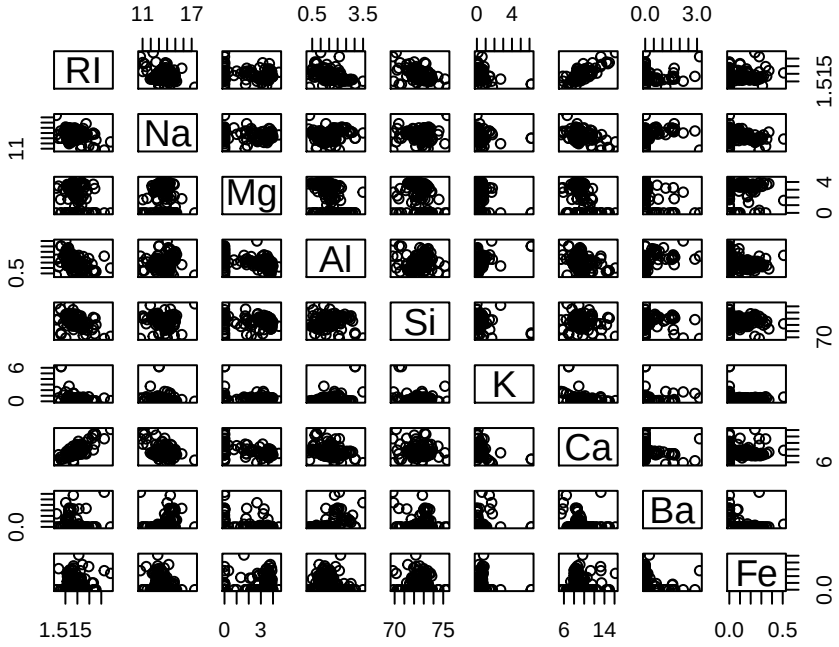
2023-11-13

Ex 1

Exploratory analysis

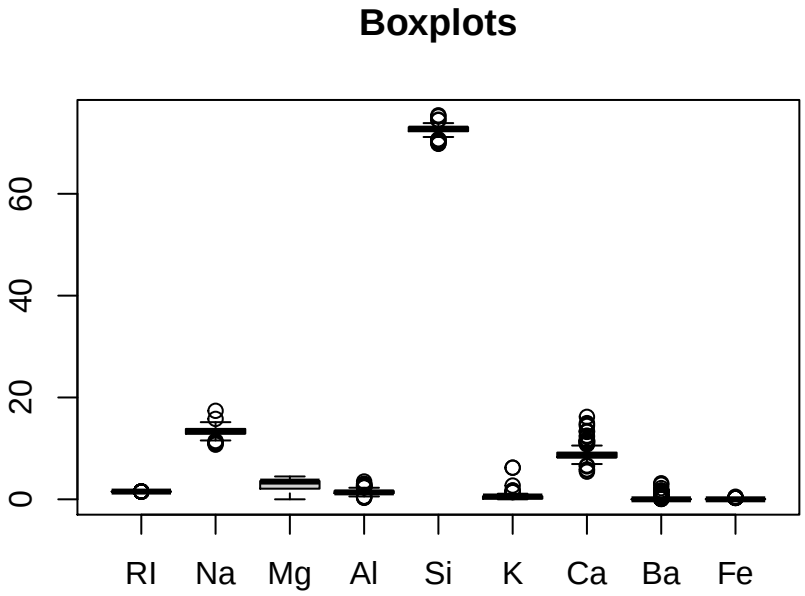


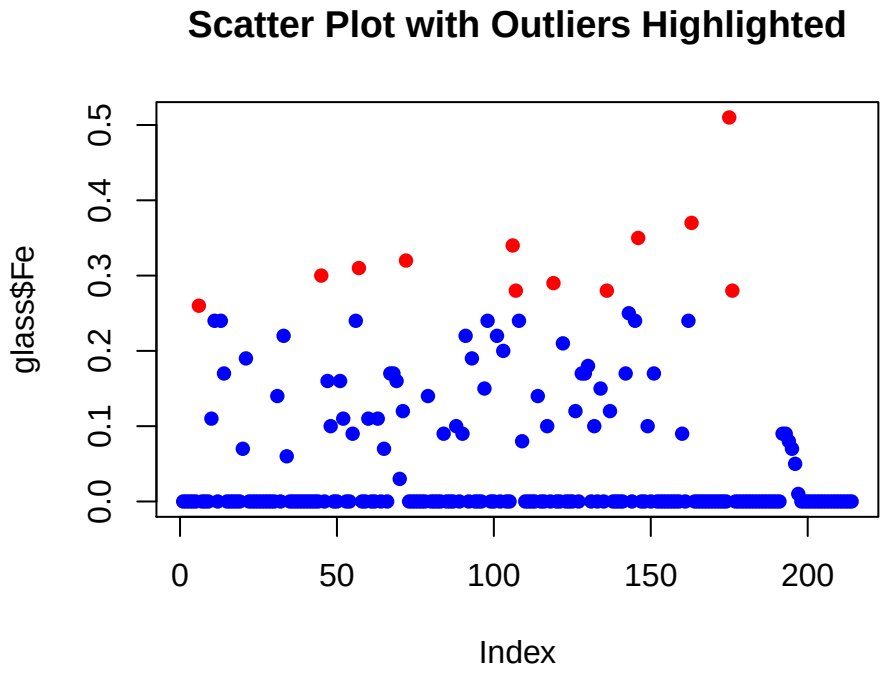




Data seem only partially spherical, one reason not to choose k-means maybe.

Assess variability and outliers





Data seem anough variable

```
##    Variance                Mean
## RI "9.22254137159407e-06" "Mean   :1.518  "
## Na "0.666841367206353"    "Mean   :13.41  "
## Mg "2.08054039094379"     "Mean   :2.685  "
## Al "0.249270179018033"    "Mean   :1.445  "
## Si "0.59992118818832"     "Mean   :72.65  "
## K  "0.425354203413628"    "Mean   :0.4971  "
## Ca "2.0253658483612"      "Mean   : 8.957  "
## Ba "0.247226993111316"    "Mean   :0.175  "
## Fe "0.00949430038172963"  "Mean   :0.05701  "
```

with a relevant number of outliers, another reason for which k-means could not be a good method to choose, while maybe PAM could.

Skeweness of data

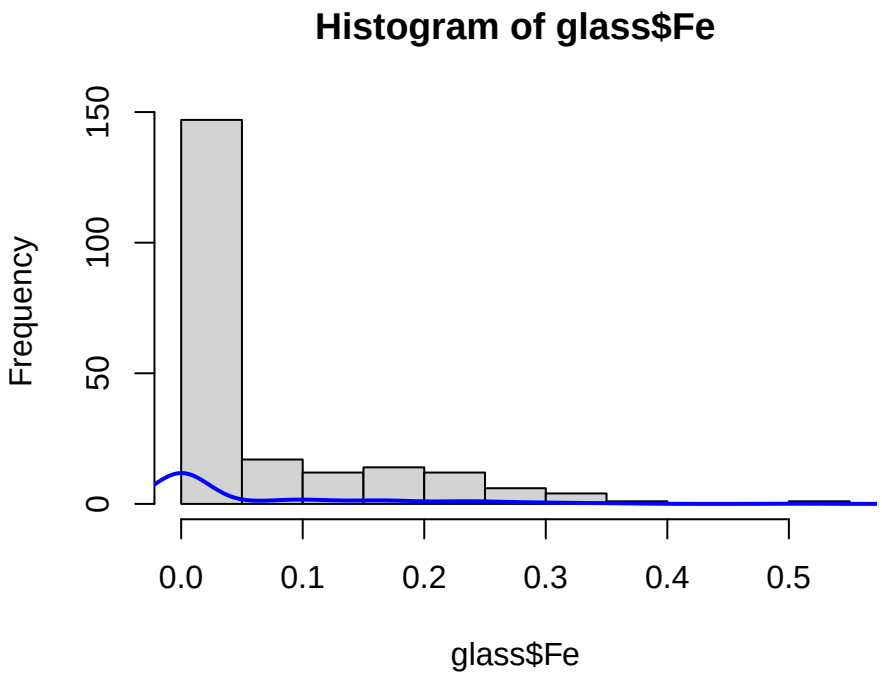
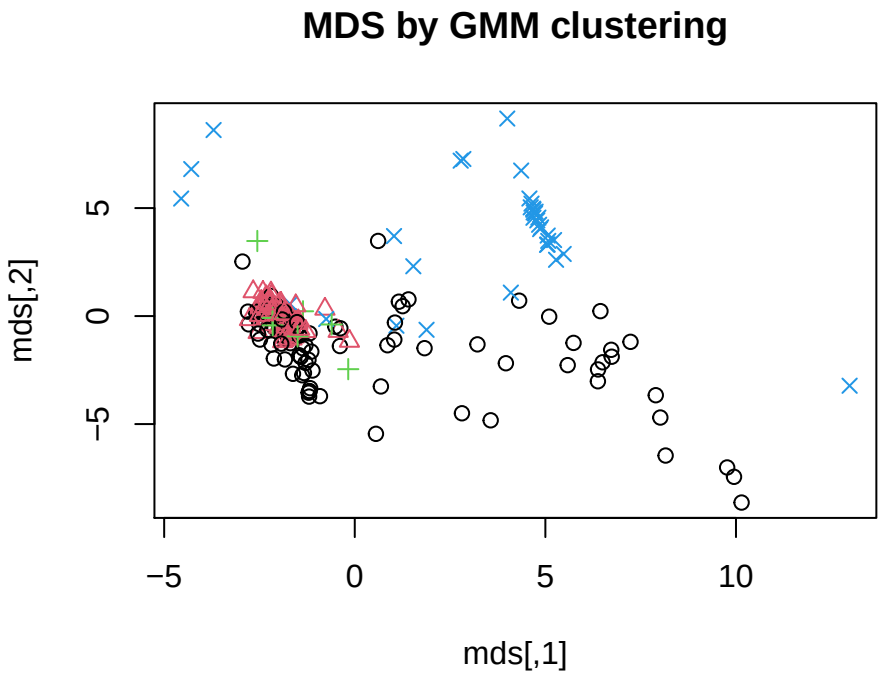


Figure 1: Barplot of one of the variables, Fe in particular

If the histogram and the density plot exhibit similar patterns, the notion of the density being relatively equal. So, we have not eqaul densities, and k-means coul be not performing. Variables also don't seem all gaussian distributed, this worn us that GMM could not perform well. We don't know the original clusterization and data of this type could assume a nested configuration, for these reasons we could use hierarchical clustering also, PAM could give us good results due to its ability to handle skewed, nonlinear and non spherical data.

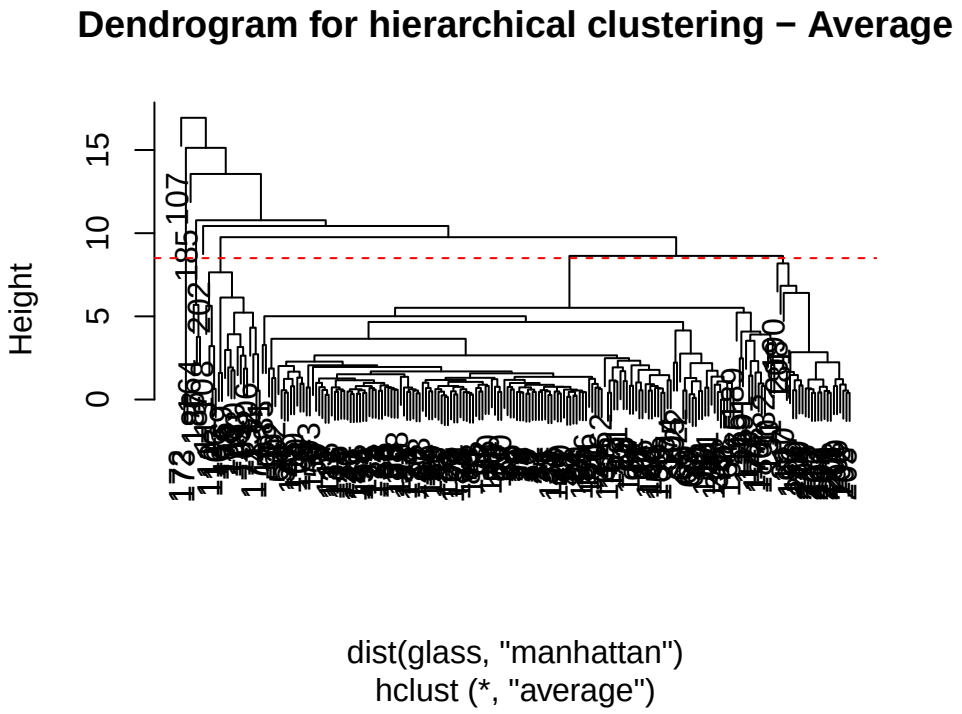
Gaussian mixture

GMM assumes that the underlying distribution of each cluster is Gaussian. If data distribution are significantly different, GMM might not perform well. That's why we could have some problem with it.



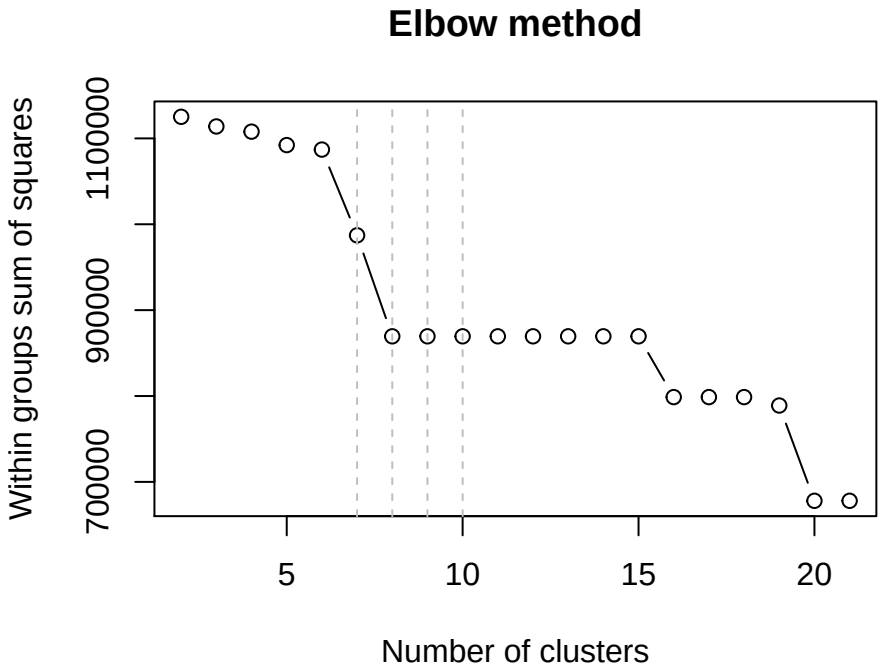
Hierarchical clustering

Average linkage might be more suitable as it is more robust to noise and outliers.

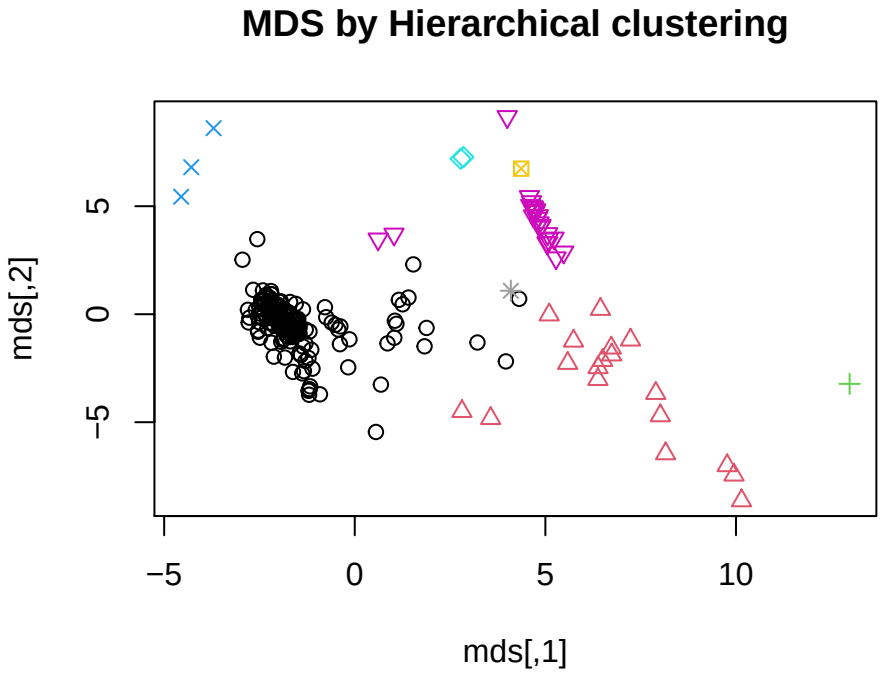


8 clusters seem to be the best option

Elbow method

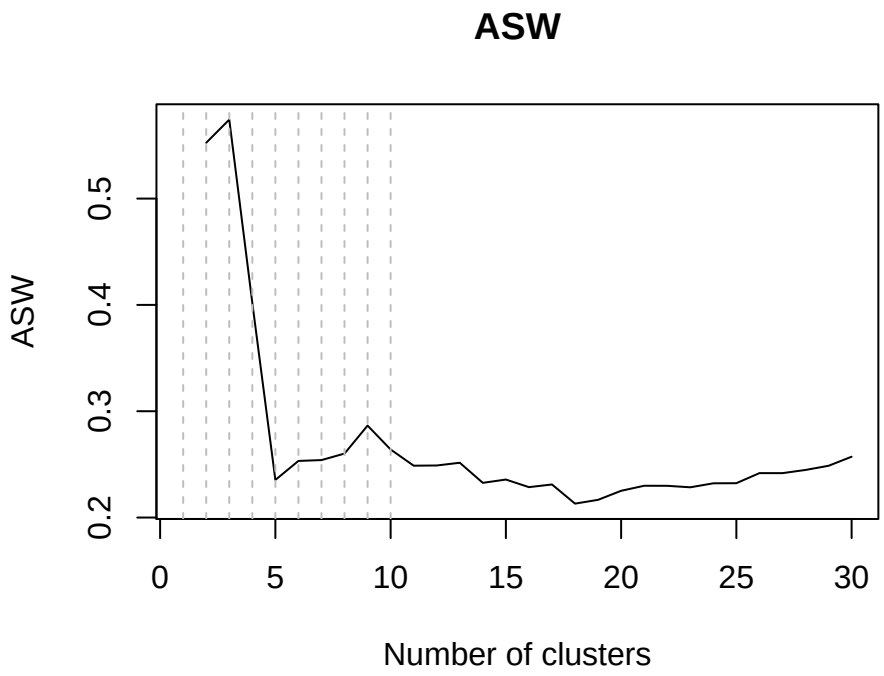


It confirms what we saw in the dendrogram

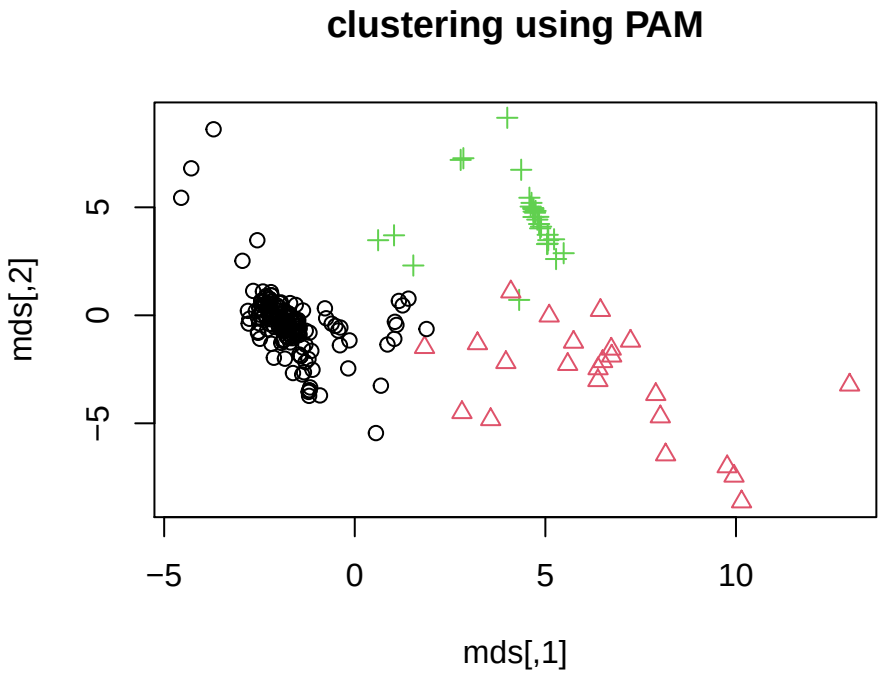


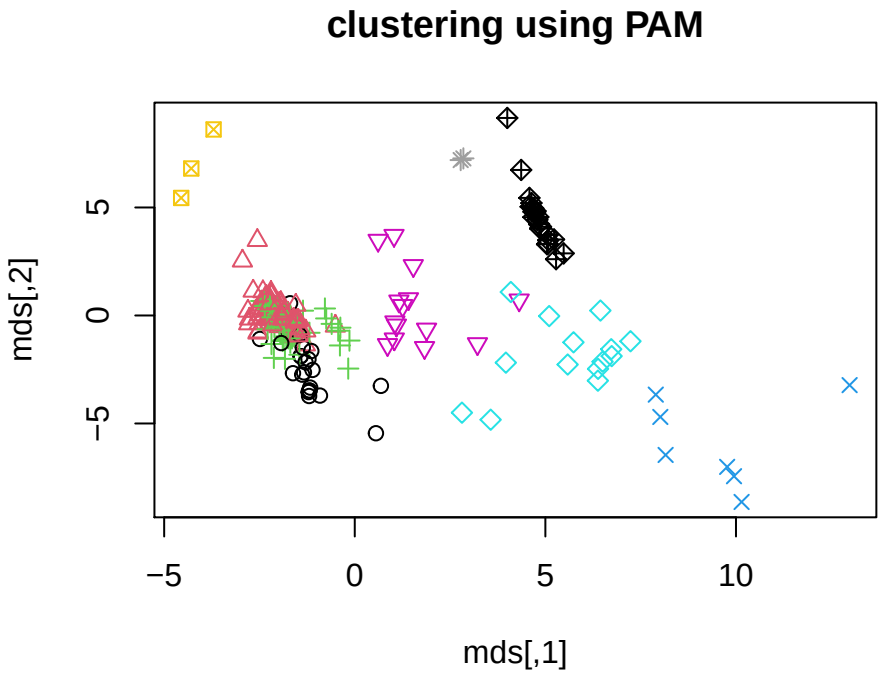
PAM

ASW to find the right K



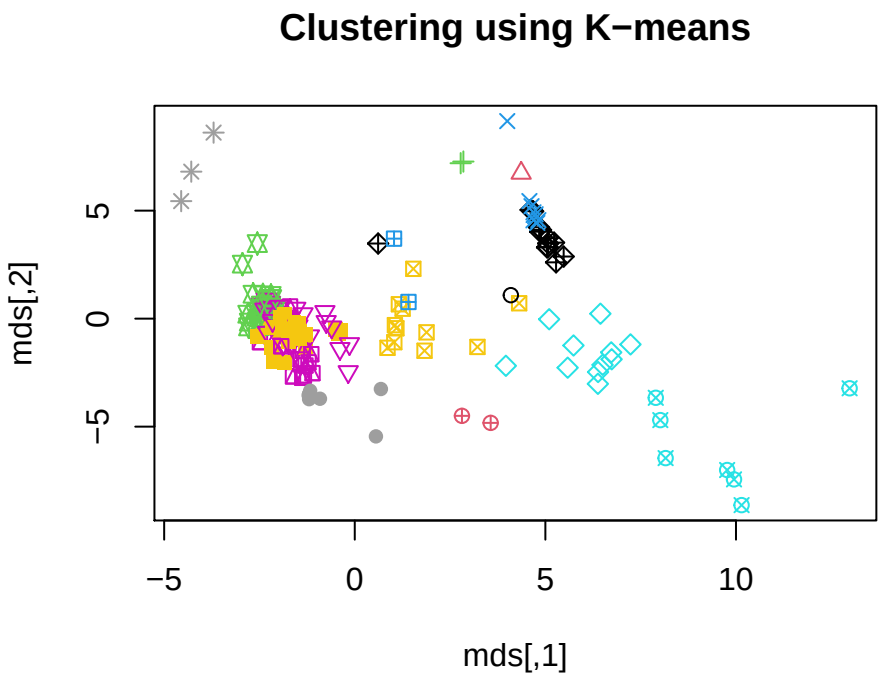
3 seems to be the best cliustering, anyway it could not make makes sense in this context. I decided to take into account also the second best value, that is for k = 9





in 2 dimentions, 3 clusters appear clearer, not forcely the best option.

K-means



Compare the clusterings (how meaningful and useful are them)

```
        PAM_HC    PAM_GMM Kmeans_PAM Kmeans_HC    GMM_HC GMM_Kmeans
[1,] 0.2763526 0.1911103   0.5523431 0.1936323 0.2100446  0.1268865
```

These indexes are just to compare the different methods' results and it seems that PAM and K-means are the only two similar, respect to the others.

- Selection of the best one

I would choose PAM clustering, basing my choice on the mds visualizations by clusters, it seems to be the most clear.

- Interpretation of the clusters (Comment on other aspects of the data set that could be relevant)

Tanking PAM clustering with 3 clusters we could imagine that they represent types of glass divided, for example, as: resistent glass, decoration glass and glass used for lamps and similar objects.

2a

We minimize $\bar{\tau}_\eta = \sum_i \sum_k p_{ik}^{(t)} (\log \pi_k + \log p(a_k; \theta^2(x_i)))$

Given $\pi_k^{NOTE} = \frac{1}{n} \sum_i p_{ik}$ $\Rightarrow$ conditioned maximization

$$\sum_k = 1 \Rightarrow \sum_k \pi_k - 1 = 0$$

we take deriv. with respect to π_k AND d

$$\sum_i \sum_k p_{ik} (\log \pi_k + \log p(a_k; \theta_k^2(x_i))) - d \left(\sum_k \pi_k - 1\right)$$

$$\sum_i \sum_k p_{ik} = \sum_i \sum_k \mathbb{1}(z_i = k)(\eta, x_i) = \sum_i \mathbb{1} = n \text{ dts}$$

$$\rightarrow \frac{\partial}{\partial} = 0 - 1 \left(\sum_k \pi_k - 1\right) = \sum_k \pi_k - 1 \Rightarrow \sum_k \pi_k = 1$$

$$- \frac{\partial}{\partial \pi_k} = \frac{1}{\pi_k} \sum_i p_{ik} - d = 0 \Rightarrow \sum_i p_{ik} - d\pi_k = 0$$

$$\rightarrow \sum_k \frac{1}{n} \sum_i p_{ik} =$$

$$\rightarrow d\pi_k = \sum_i p_{ik} \rightarrow \pi_k = \frac{\sum_i p_{ik}}{d}$$

NOTE

$$\sum_k \pi_k = \frac{\sum_k \sum_i p_{ik}}{d} \rightarrow 1 - \frac{n}{d} \rightarrow b = n$$

$$\Rightarrow \pi_k^{NOTE} = \frac{\sum_i p_{ik}}{n}$$

Ex 3

Results for ex 3 are in the script, here are some considerations:
Sorry, i don't know why but my rmd file for this exercice didn't work.
To sum up:
I preferred manual method for iris data and iterative one for glass and oliveoil datasets.
Manual parameters have been tuned one by one looking for the best look in the plots.
It seems that for glass dataset we obtain worst results.