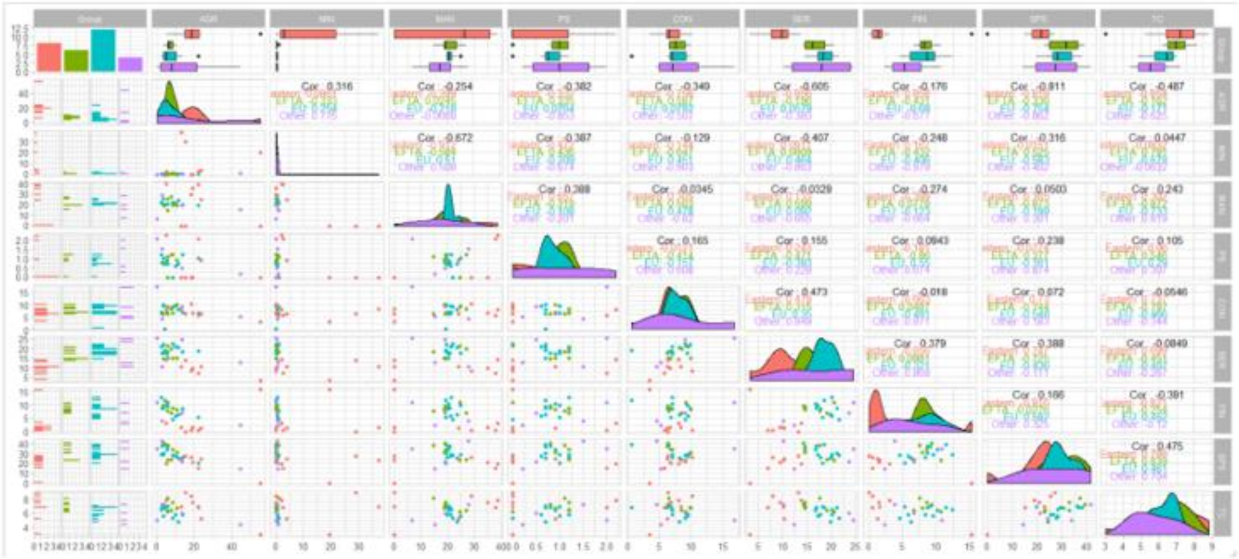
Assignment #4

May 2021

1)

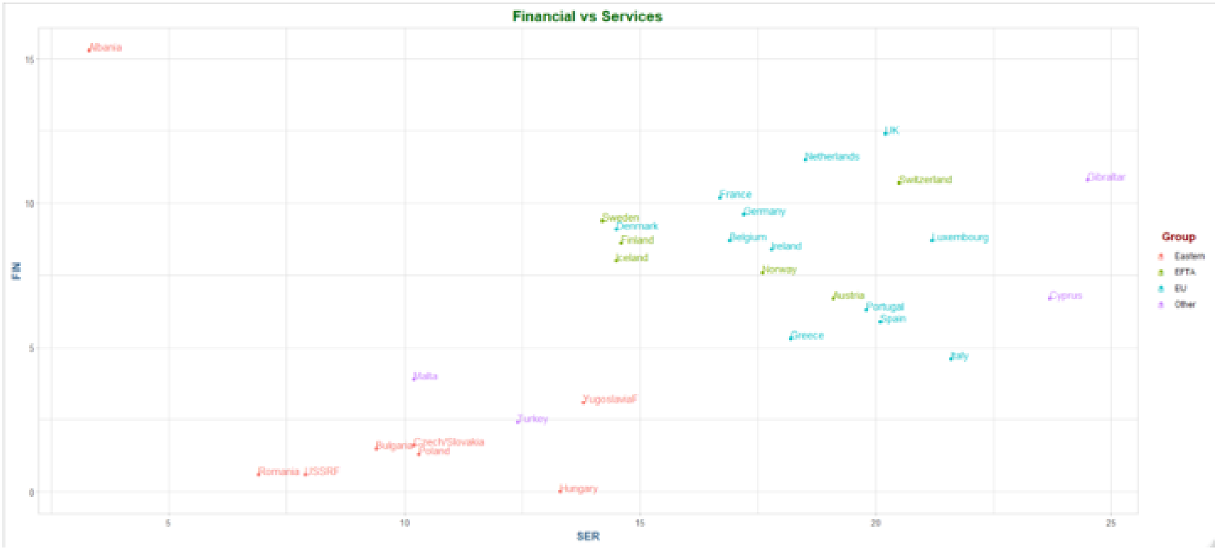
There are only a small number of variables so I will start with a pairwise scatterplot, and it should be noted that when there are a smaller number of variables like in our dataset the pairwise is a useful and viewable graphic. This is not the case when there are too many datapoints.



The groupings I focus on are in the lower part of the chart, specifically man/ser, ser/fin, tc/fin, sps/ser.

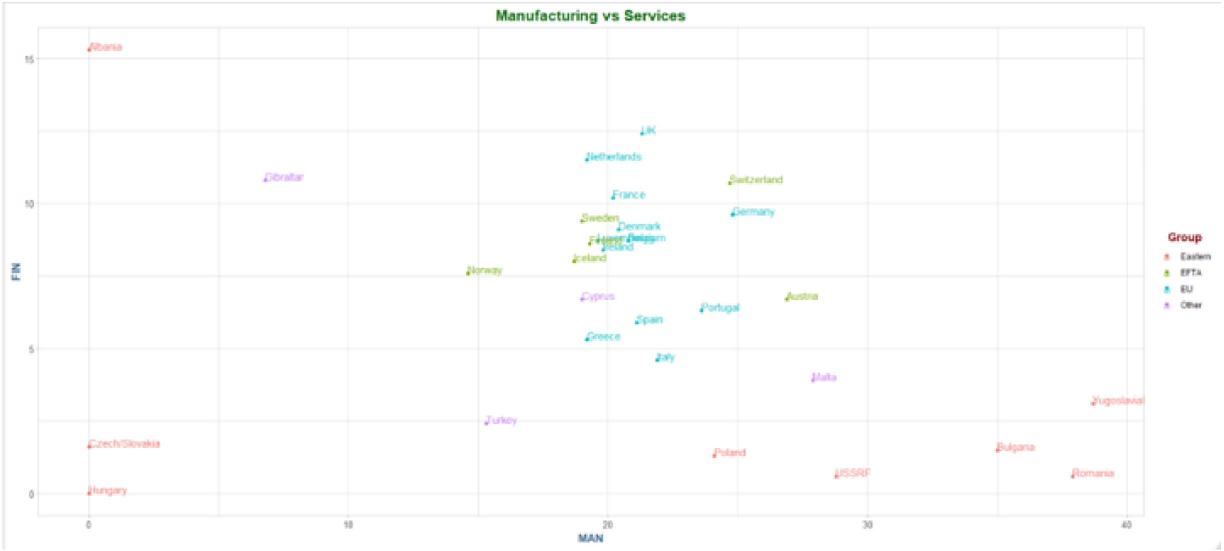
2)

a) Fin/ser:



There are only two distinct clusters in this graph.

b) Man/Ser:

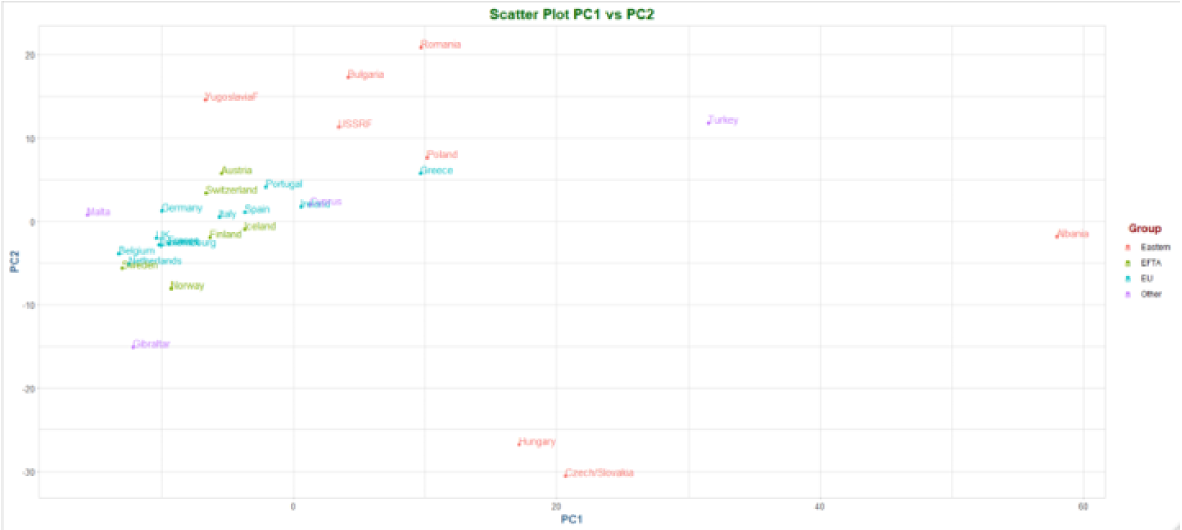


The seem to be four distinct clusters in this plot.

The first graph might be an easier algorithm to cluster due to the clearer lines of separation.

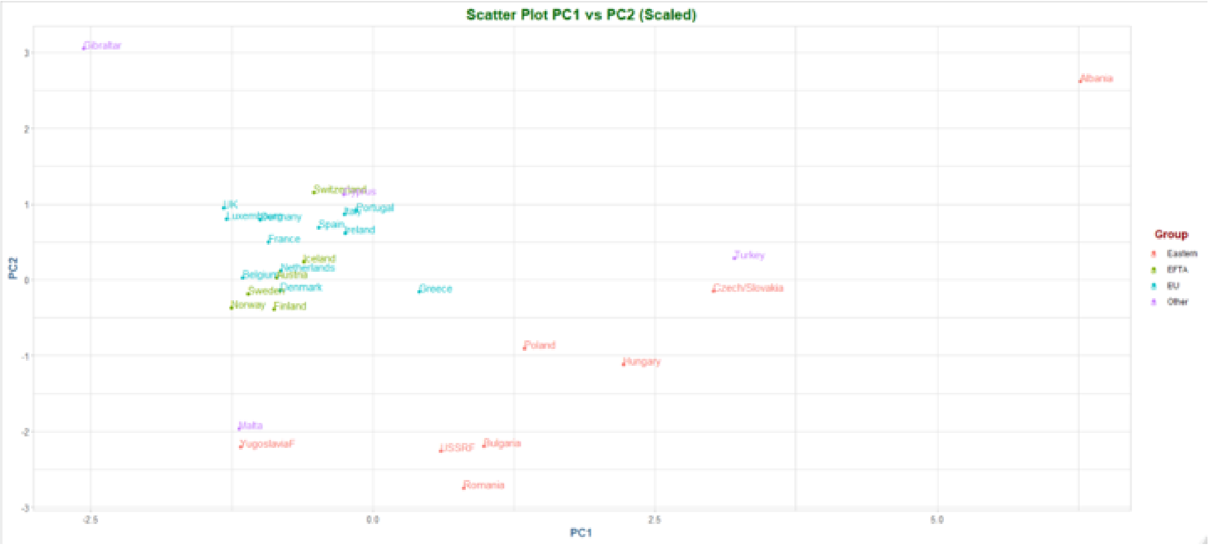
3)

We can use the PCA to reduce the dimensions of the data and project the data down from 9D to 2D by using the PCA and using first and second principal components. By doing this, we are creating a new 2D view of the data ana a view of the data that contains info from more than these two dimensions.



a) The first two principal component loadings

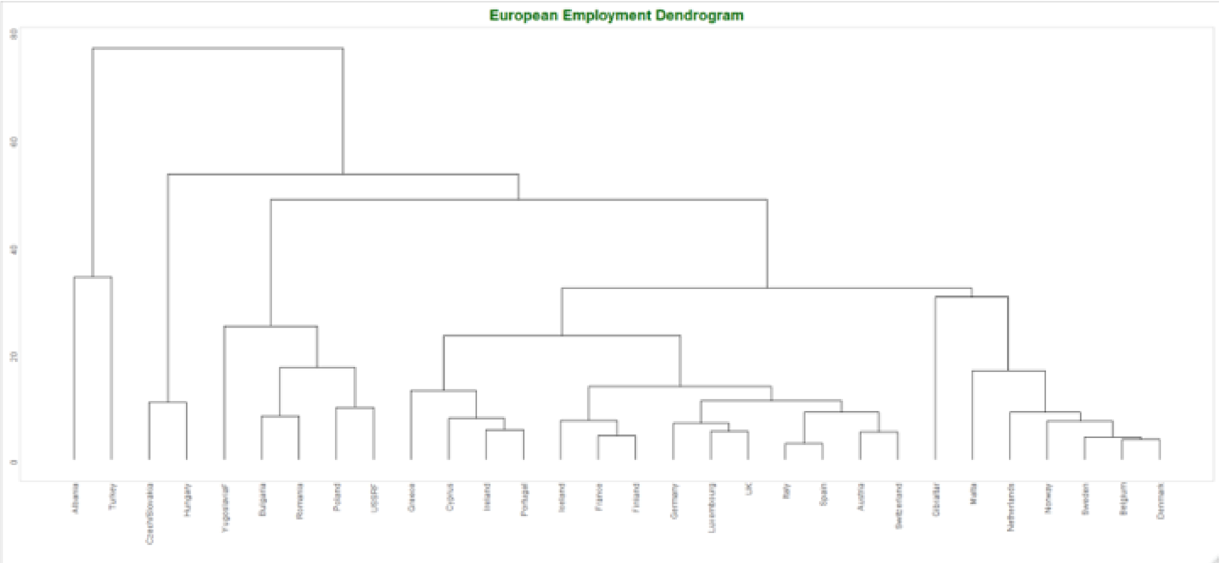
b) The scaled version of the PCA shows some difference in the clusters.



4)

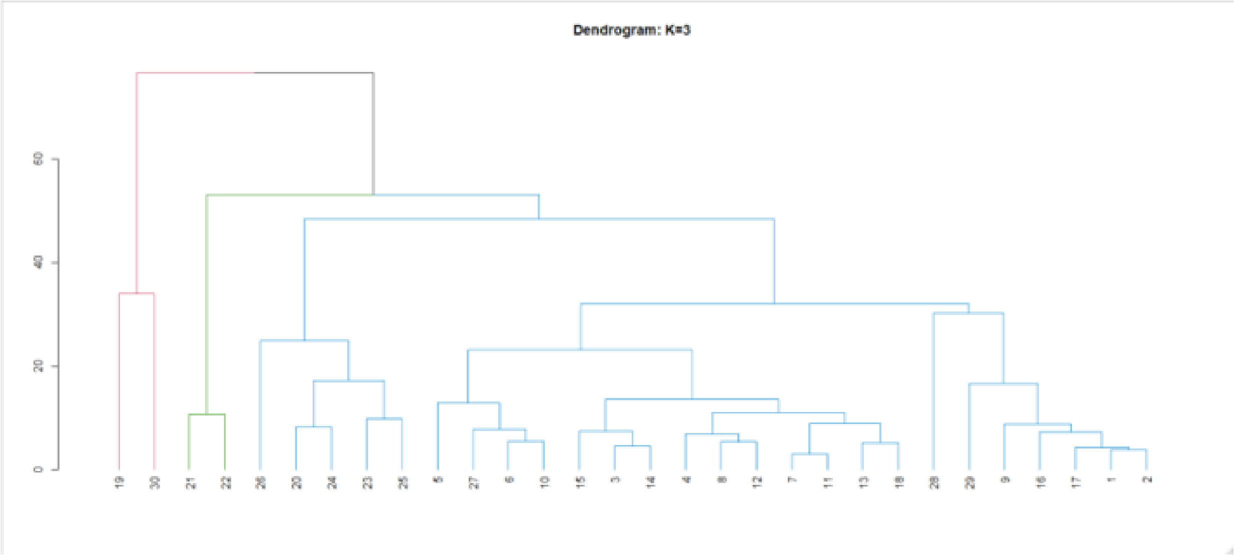
Hierarchical clustering algorithms fit a tree of clusters from k=2 to k=n, where n is the number of data points in the sample. This tree of clusters can be visualized using a dendogram and since the cluster tree stores all cluster assignment, we need time the tree using cutree() to force an assignment of the observations to a particular number of clusters.

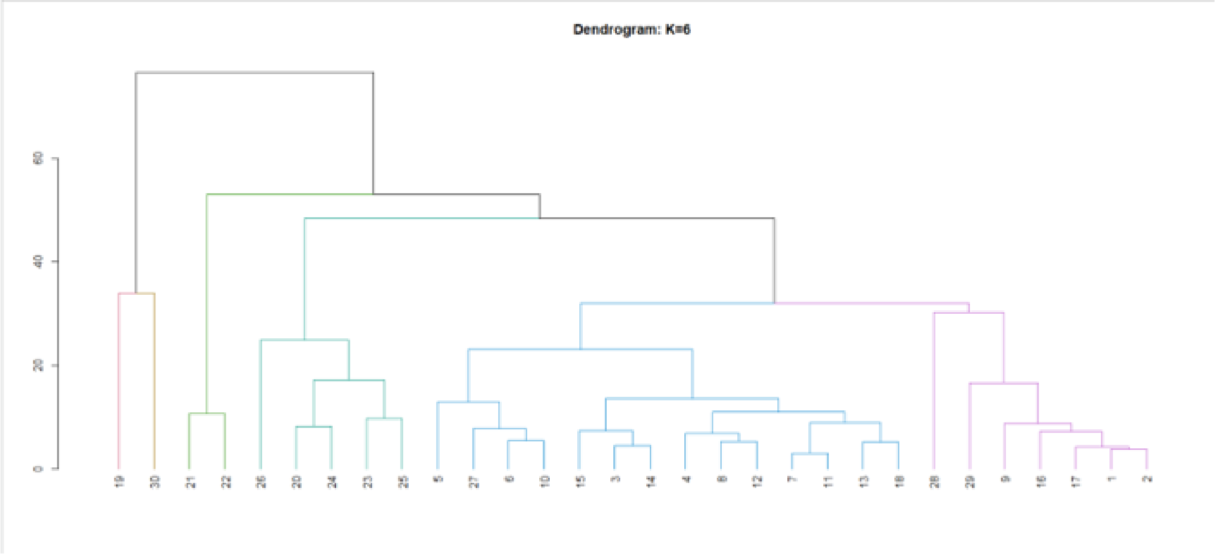
a)

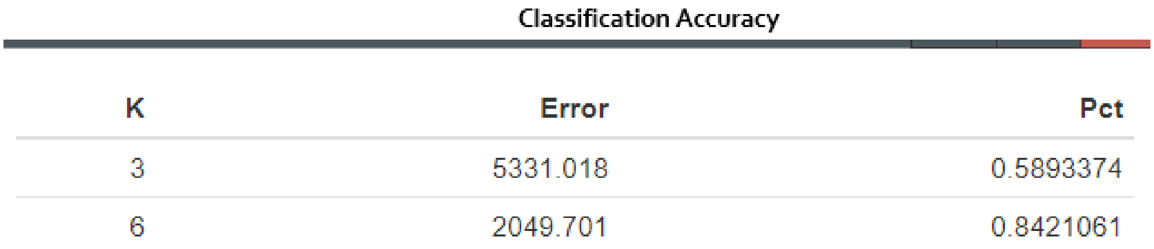


Dendogram cuts:

b)





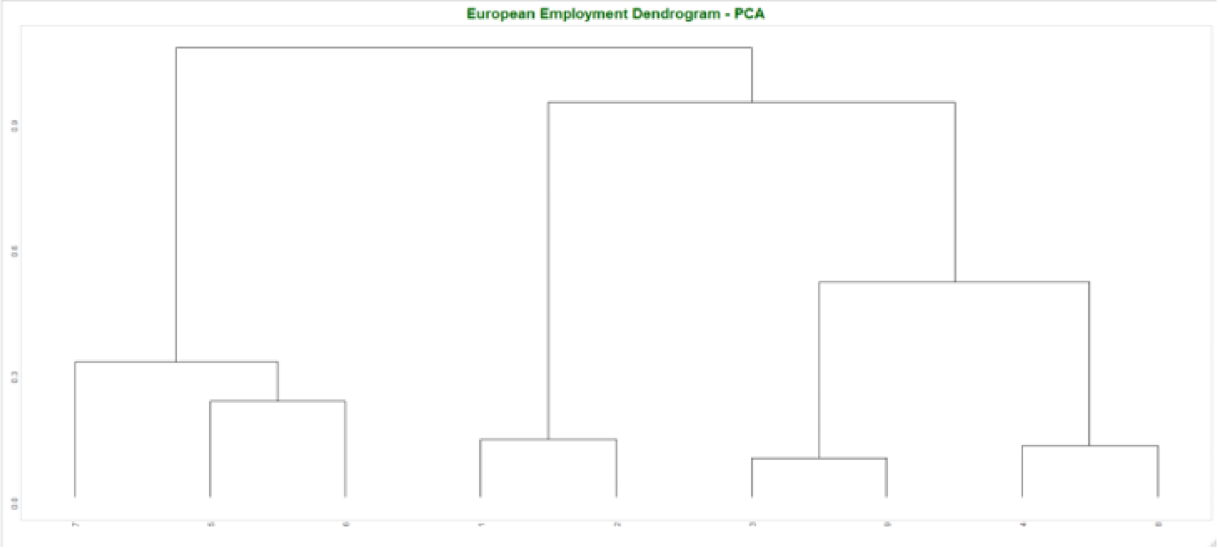


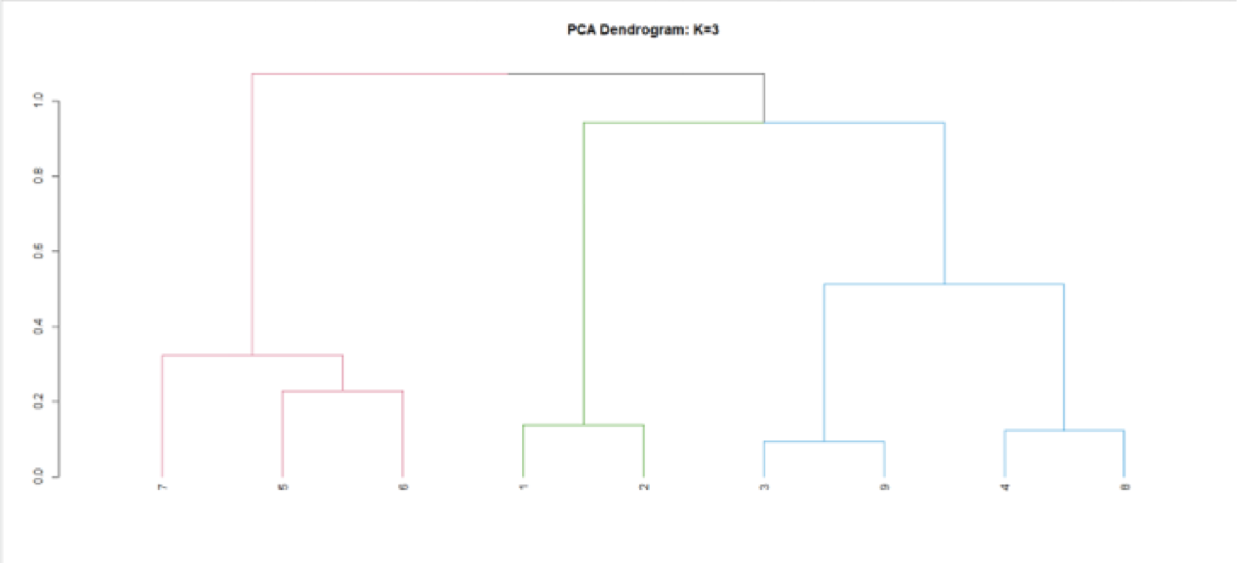
Classification Accuracy

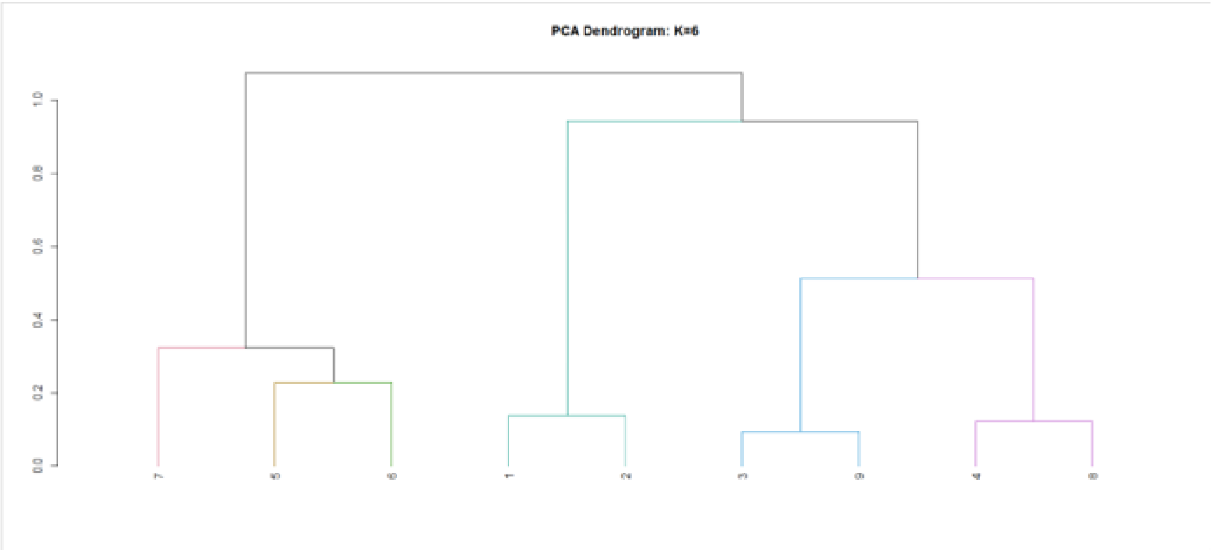
K	Error	Pct
3	5331.018	0.5893374
6	2049.701	0.8421061

Six clusters clearly has better accuracy than just 3.

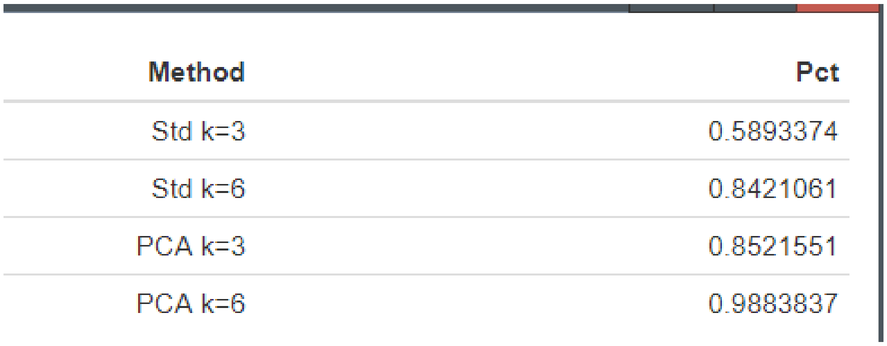
b)


European Employment Dendrogram - PCA
0.9
0.6
0.3
0.0
7
5
6
1
2
3
9
4
8


PCA Dendrogram: K=3
1.0
0.8
0.6
0.4
0.2
0.0
7
5
6
1
2
3
9
4
8



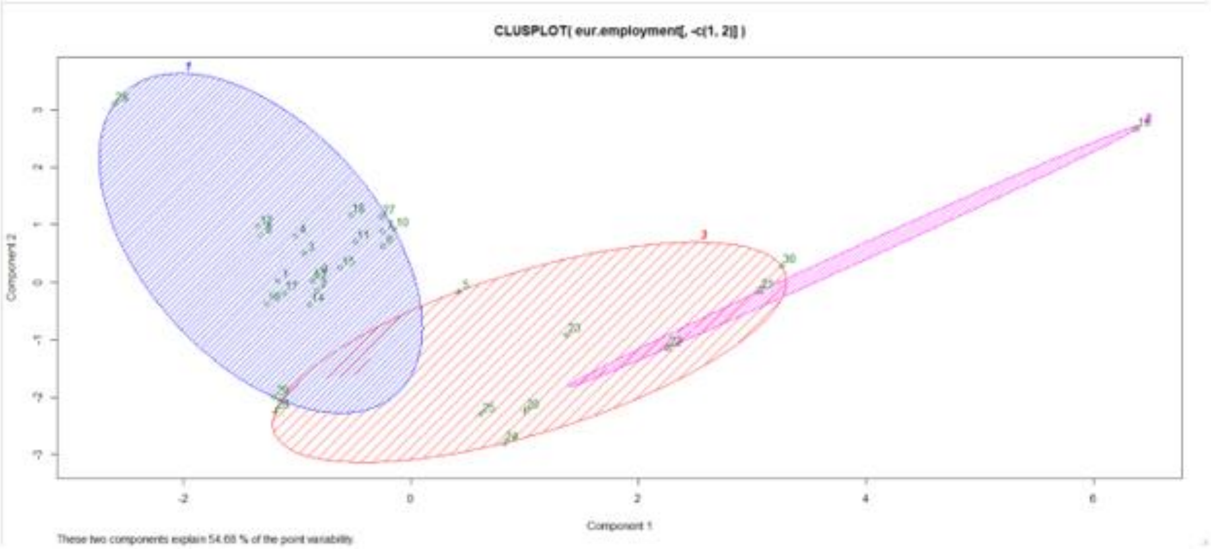
Summary of the classification accuracy of the four models:

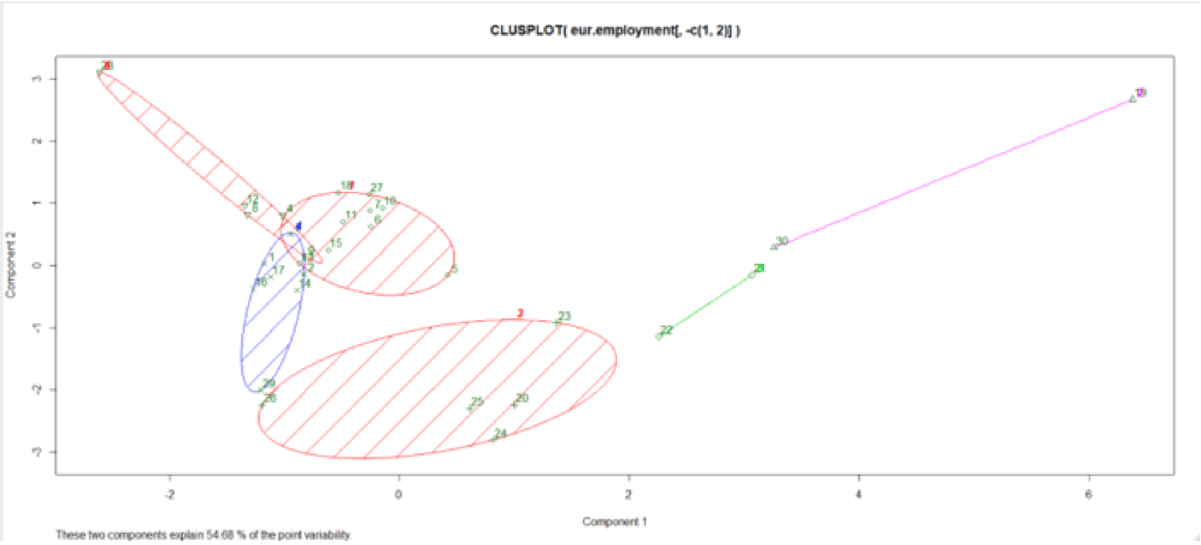


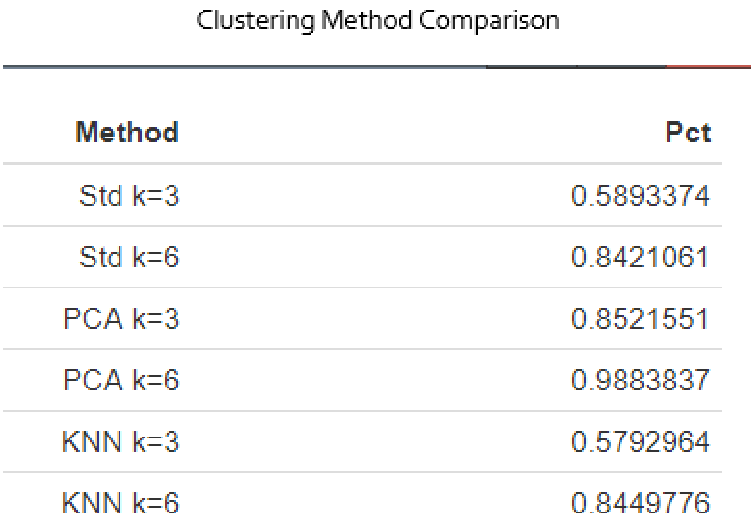
Method	Pct
Std k=3	0.5893374
Std k=6	0.8421061
PCA k=3	0.8521551
PCA k=6	0.9883837

Six clusters still seems to be the best performing model with near perfect accuracy.

5)


CLUSPLOT(eur.employment[, -c(1, 2)])
Component 1
Component 2
-2
0
2
4
6
-3
-2
-1
0
1
2
3
These two components explain 54.68 % of the point variability.


CLUSPLOT(eur.employment[, -c(1, 2)])
Component 1
Component 2
-2
0
2
4
6
-3
-2
-1
0
1
2
3
These two components explain 54.68 % of the point variability.



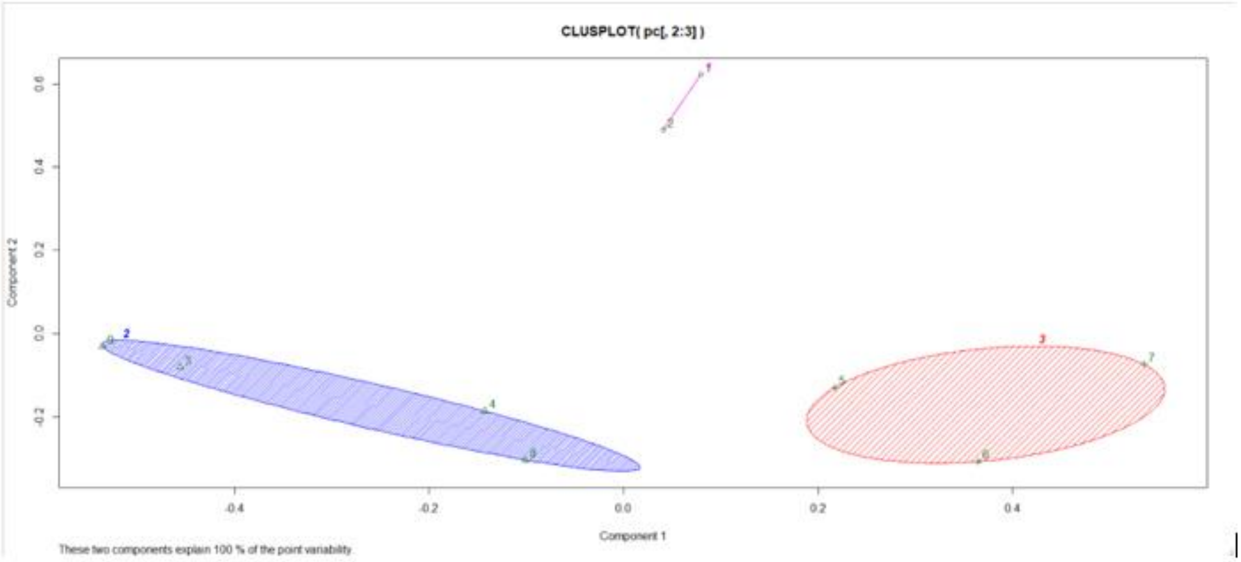
Clustering Method Comparison

Method	Pct
Std k=3	0.5893374
Std k=6	0.8421061
PCA k=3	0.8521551
PCA k=6	0.9883837
KNN k=3	0.5792964
KNN k=6	0.8449776

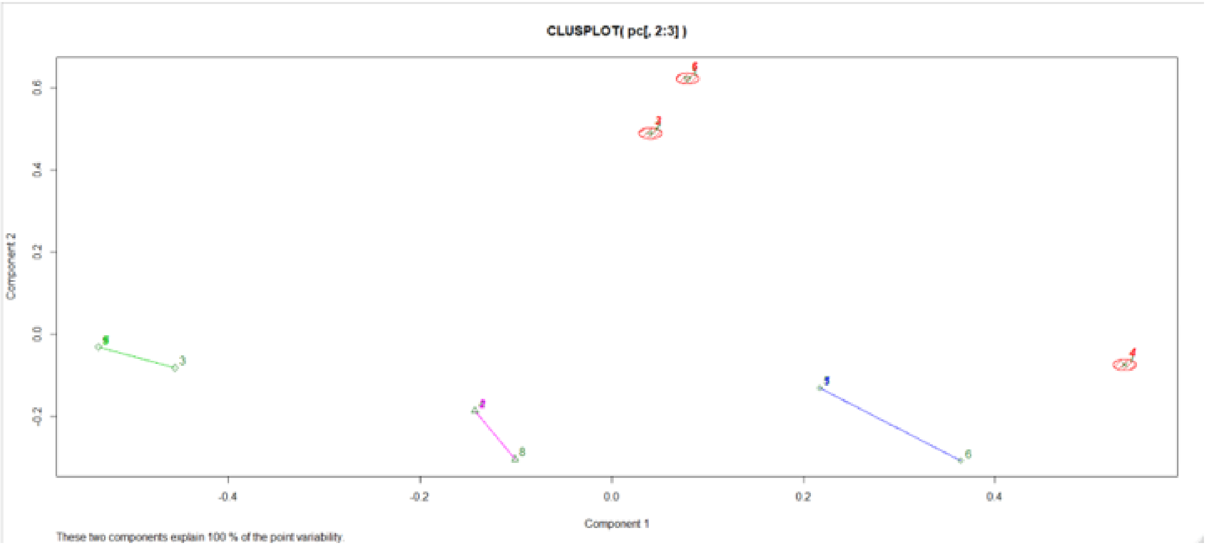
The classification accuracy for the KNN models are a bit worse than the hierarchical (non-pca) models although they are pretty similar. The three cluster plot has substantial overlapping clusters and doesn't look to be a good fit for the data. The six cluster plots like the previous version, performs much better and has substantially less overlap in the clusters.

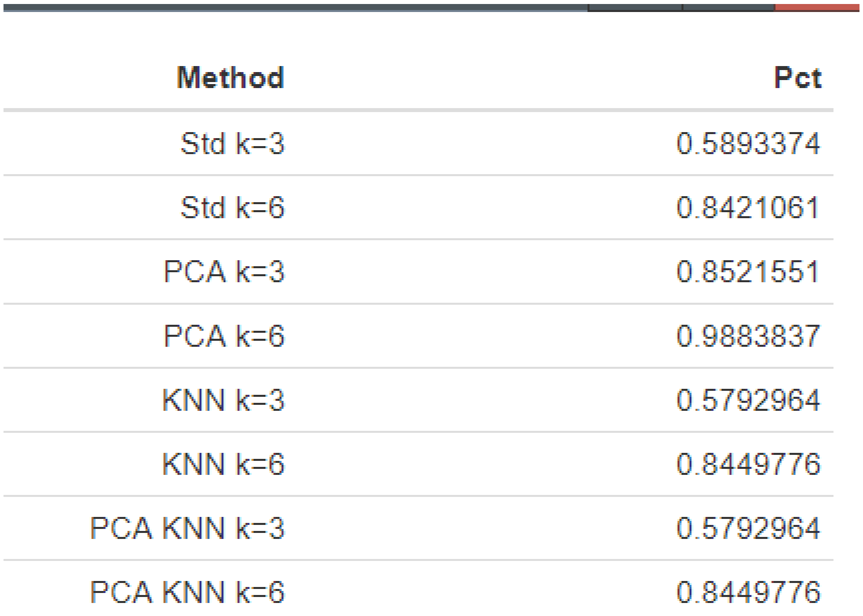
PCA:

K = 3:



K = 6:

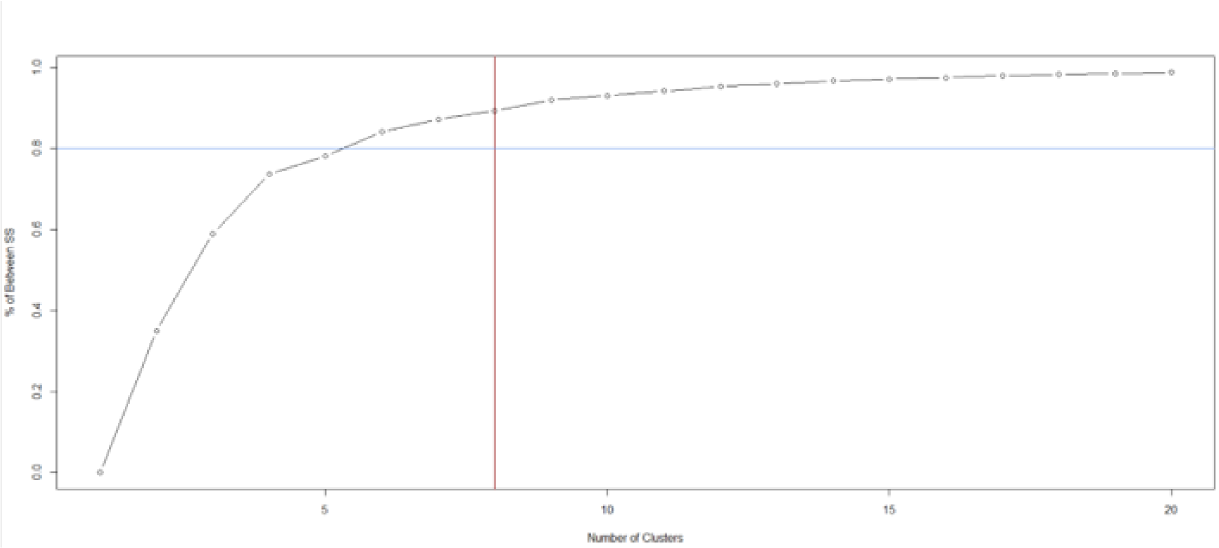




Method	Pct
Std k=3	0.5893374
Std k=6	0.8421061
PCA k=3	0.8521551
PCA k=6	0.9883837
KNN k=3	0.5792964
KNN k=6	0.8449776
PCA KNN k=3	0.5792964
PCA KNN k=6	0.8449776

The KNN clustering with k-3 combines the eu/efta/other together and splits easter between clusters 1 and 3. The k=6 clustering disperse the countries evenly throughout the 6 groups but there isn’t one cluster that exhibits dominance through the groups.

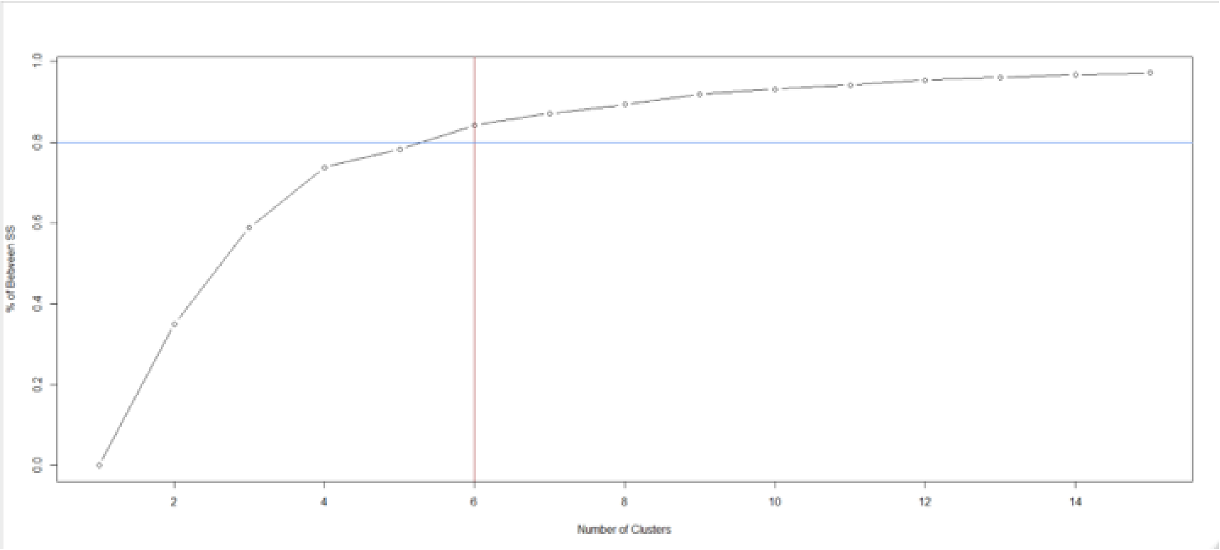
6)



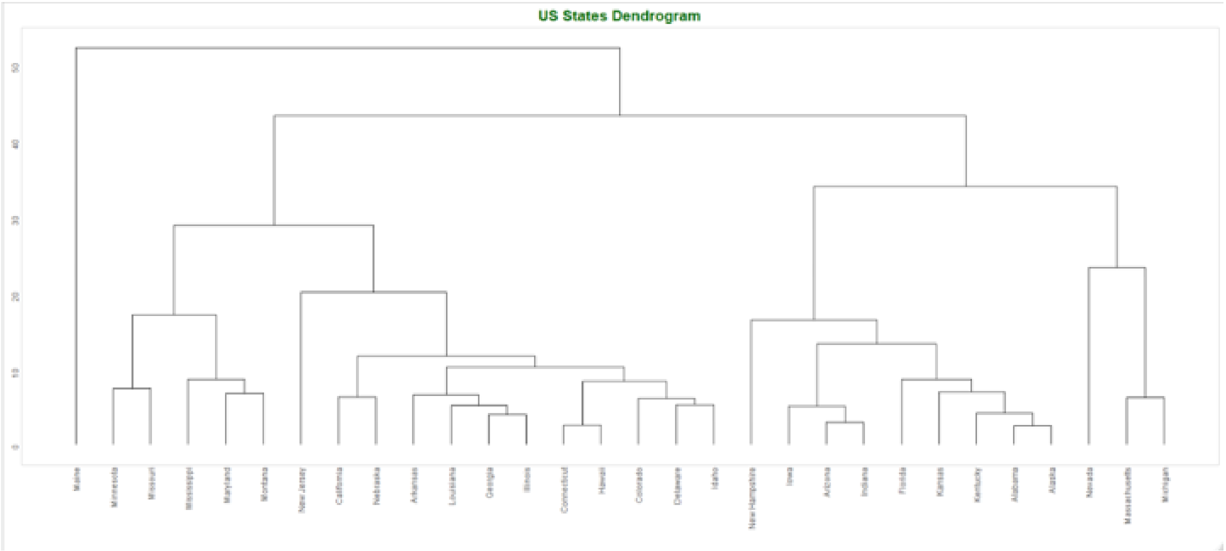
I would choose k-8 to explain the largest proportion in variance as the diminishing returns after the cutoff are pronounced. A prior cutoff at 80% suggests we would pick k-5. The optimal number seems to be as we can explain a large amount of the variance in the data without introducing a lot of bias.

7)

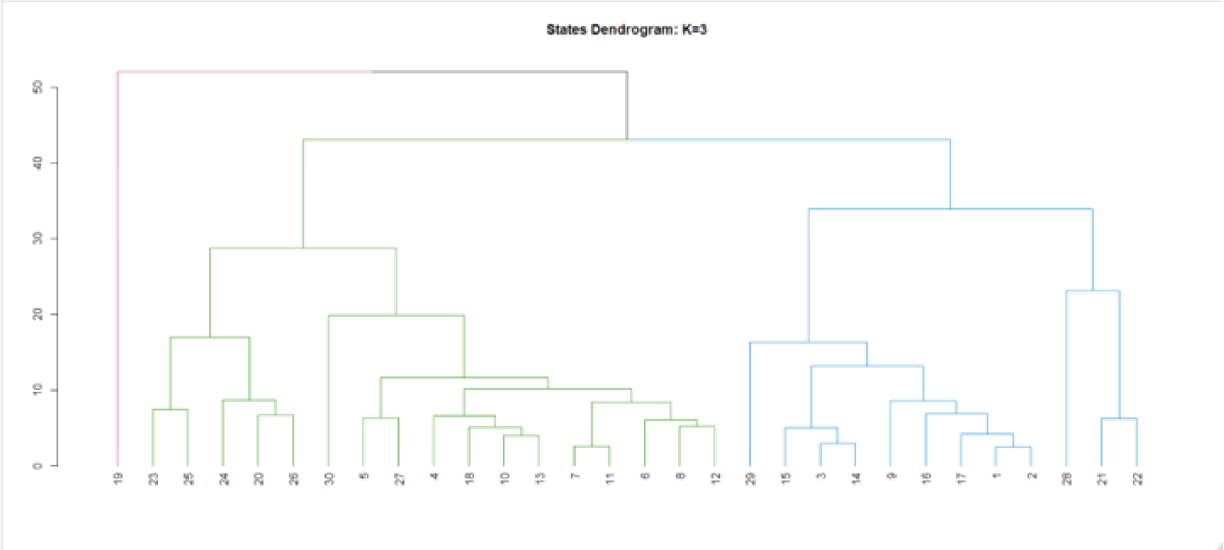
After viewing the US States dataset, we can run ti through similar analysis that we conducted to look at the variances explained in the different number of clusters.



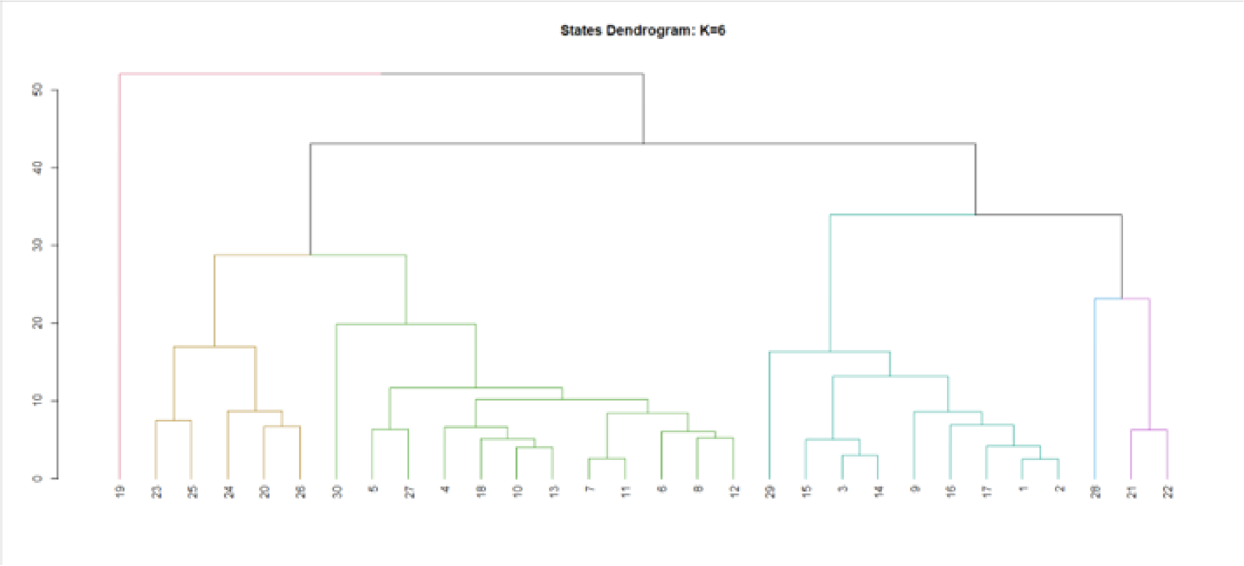
Using the priori cutoff of 80%, we get close to 6 clusters, but there is a sharp decline in explained variance after k = 3, so we will explore both.



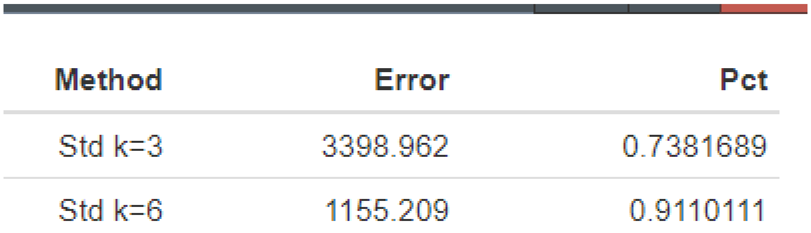
There is some clustering in a few regions but there seems to be more dispersion in the leafs that we can explain with k-3.



In this dendrogram we can see a heavily unequal distribution of two of the nodes, with Maine as an outlier, but increasing the nodes to 6 increases the balance of the nodes.



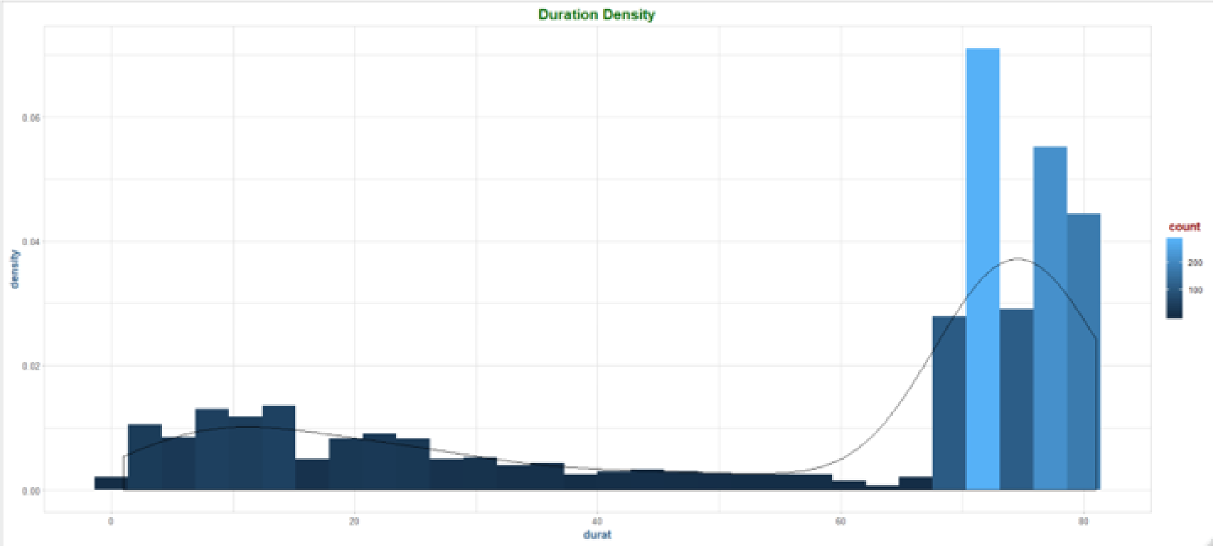
The errors from k-3 and k-6 show that k=6 is a good model for this data using a hierarchical clustering model.



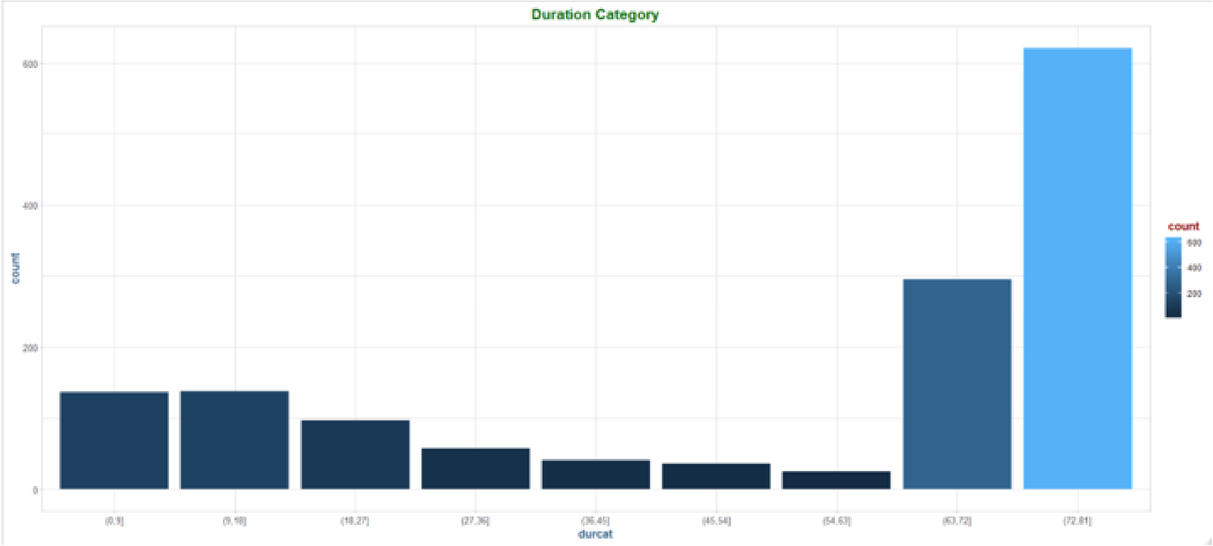
Method	Error	Pct
Std k=3	3398.962	0.7381689
Std k=6	1155.209	0.9110111

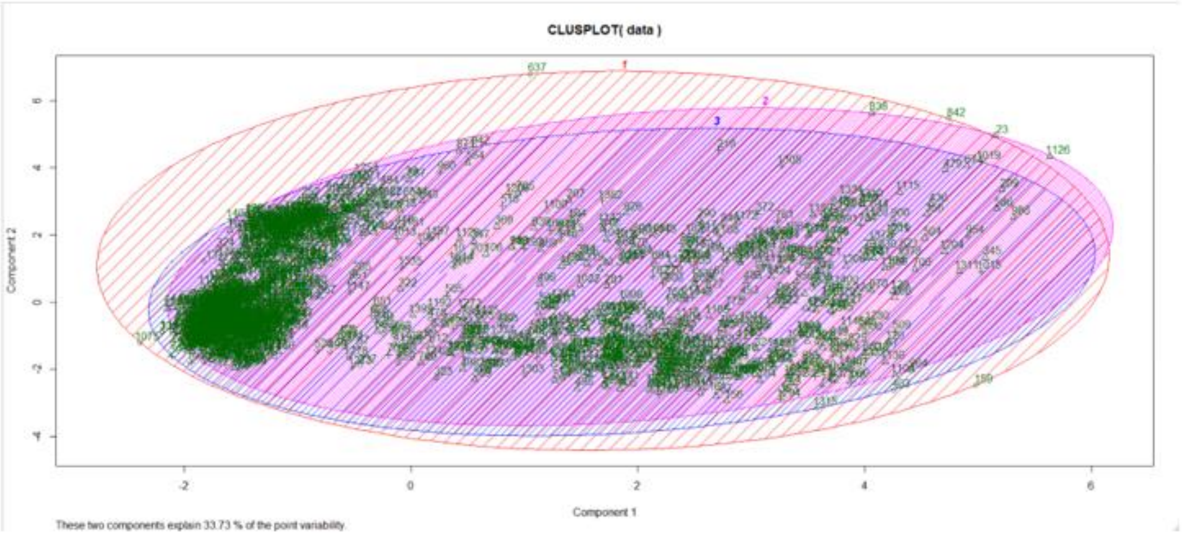
8)

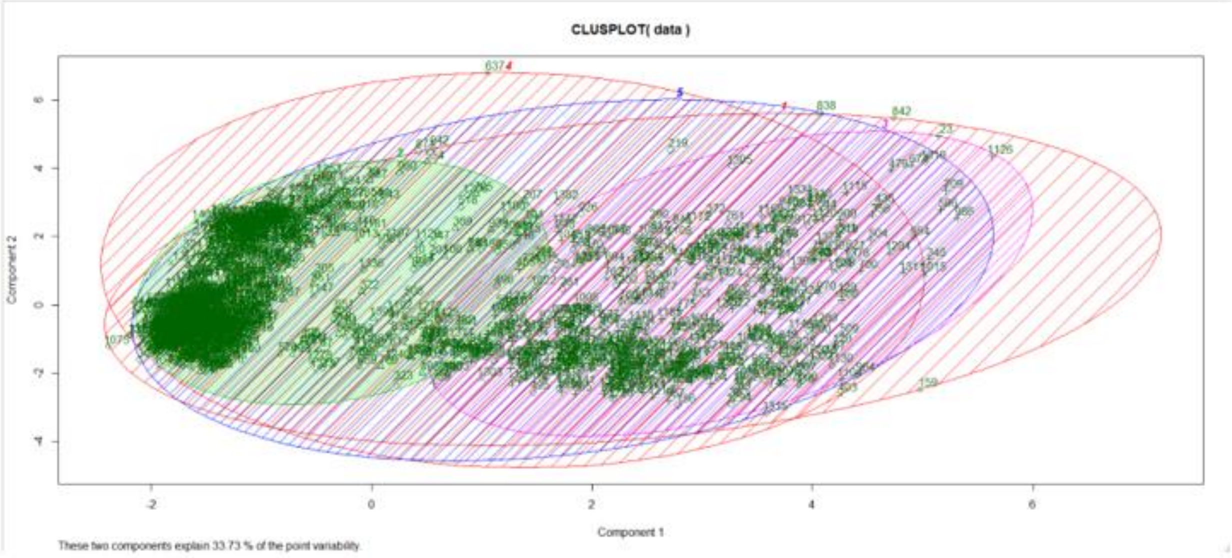
The recidivism data from the past assignment had a lack of categorical labels that were present in previous datasets. First, we look at the duration variable the main response variable used in the study.



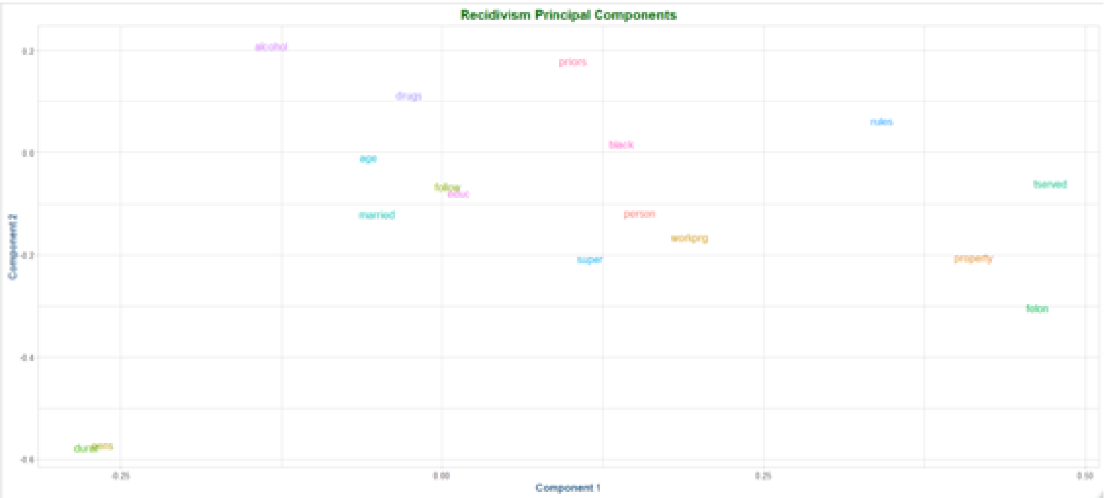
I binned them together in categories with a new variable durcat.

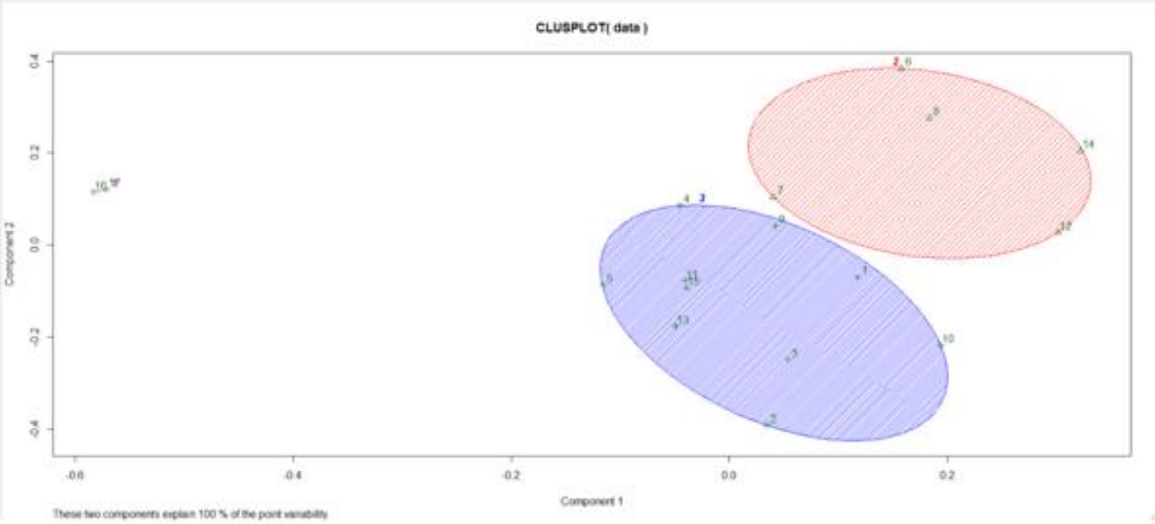




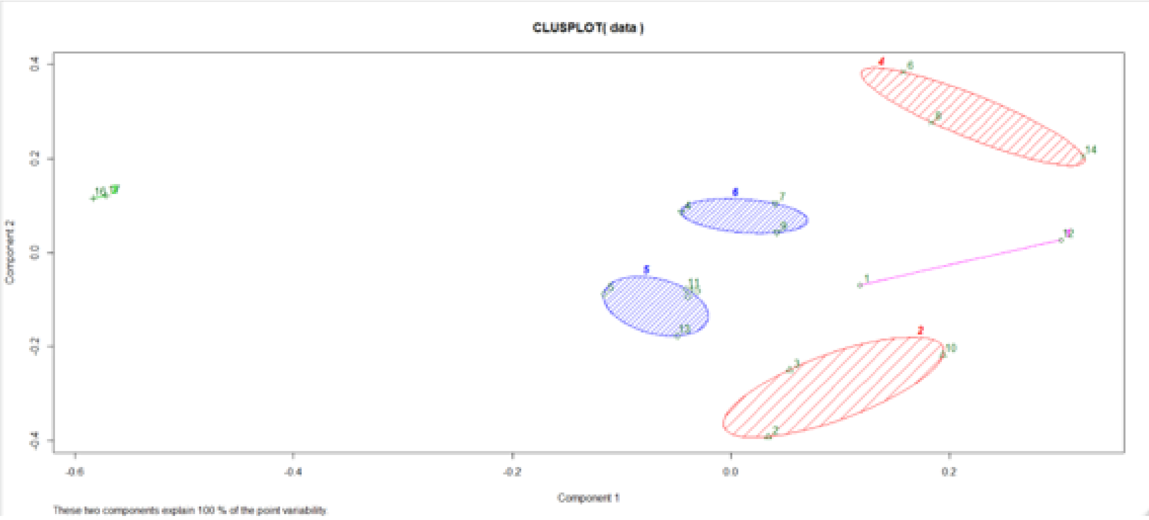


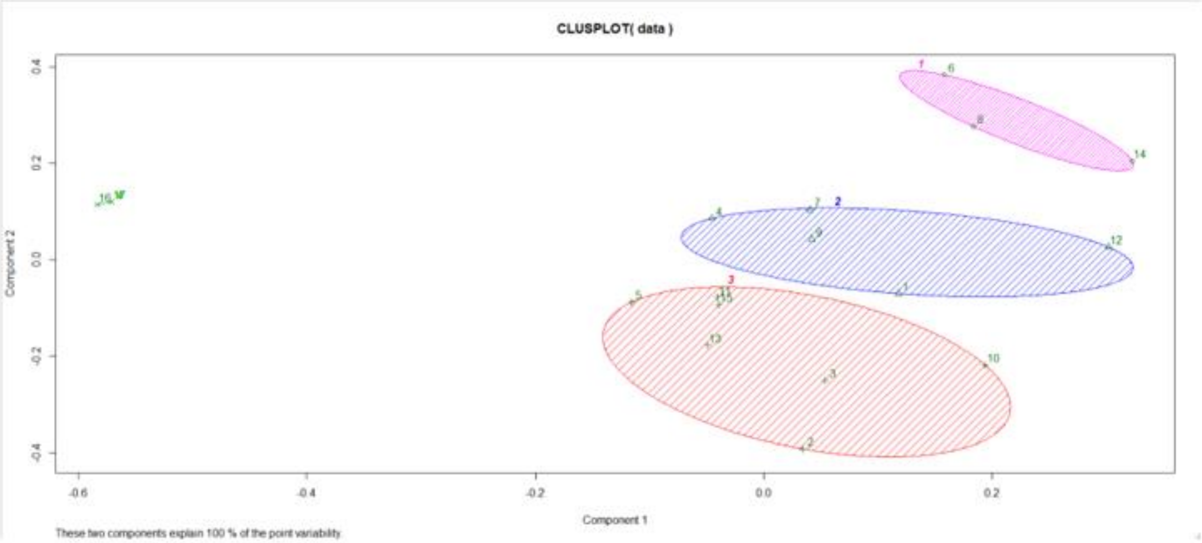
The groups have a large amount of overlap although there begins to be some separation around k=5 within the main clusters in the dataset.





These are better results than we had in previous models but there is still a large amount of variance in the 2 clusters with some outliers. With k=6 there is more diversity but half the groups appear to have a couple of observations.





The majority of the clustering is at k=4 with the data being clustered into 3 primary groups and only a few outliers. The optimum number of clusters is 4.

9)

This assignment was another challenging and fulfilling one. I have had some experience with KNN and clustering and it was good to expound on it with a different dataset. The first model was new because I hadn't had experience with hierarchical clustering. The European employment dataset seemed different than the other datasets with many observations. There also seems to be a large amount of hidden structure in the data. The hierarchical analysis was interesting when using dendrograms. Using PCA was also good to reduce dimensionality of the data and it helped with classification metrics although some of it was difficult to interpret.

The second model I had more experience with and it worked pretty well with the European data at k=6 without the use of PCA explaining most of the variance in the data. Lastly, the last two models were also interesting. Applying both the KNN and PCA had good results on this data.