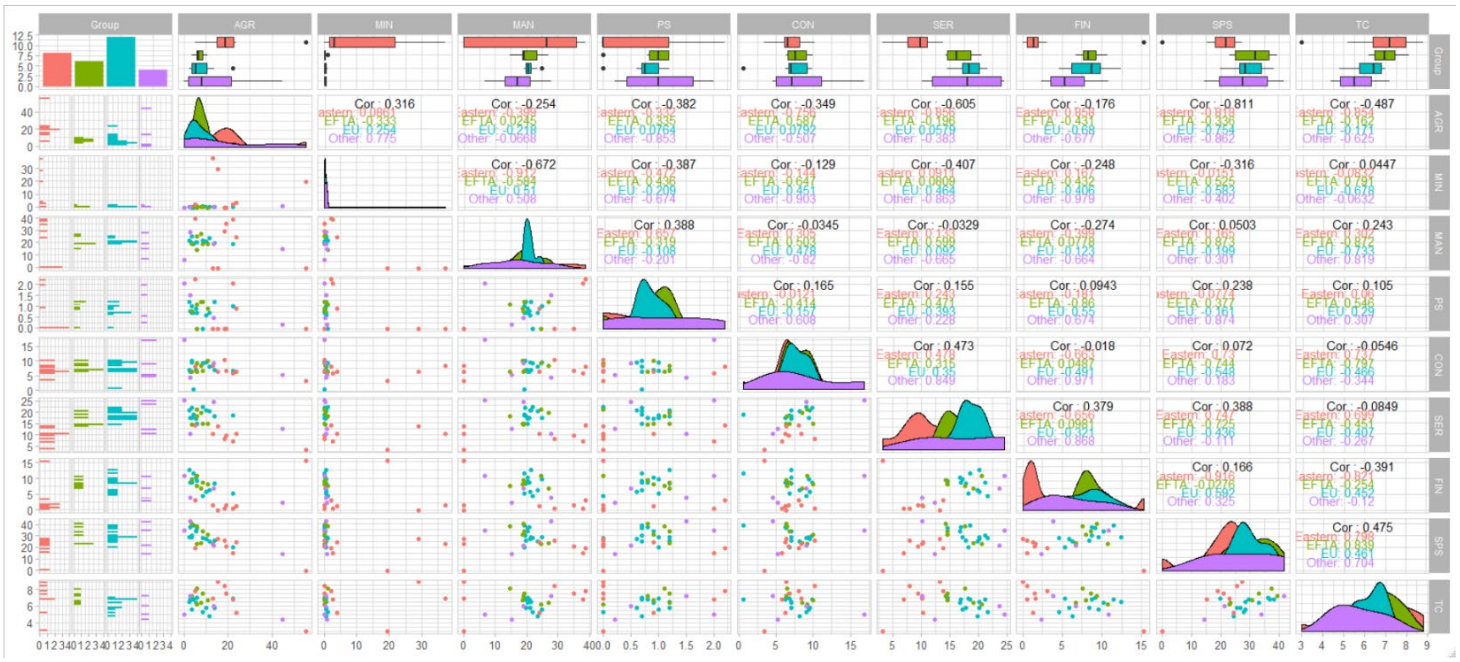
ASSIGMENT #4
BRANDON MORETZ

TASKS

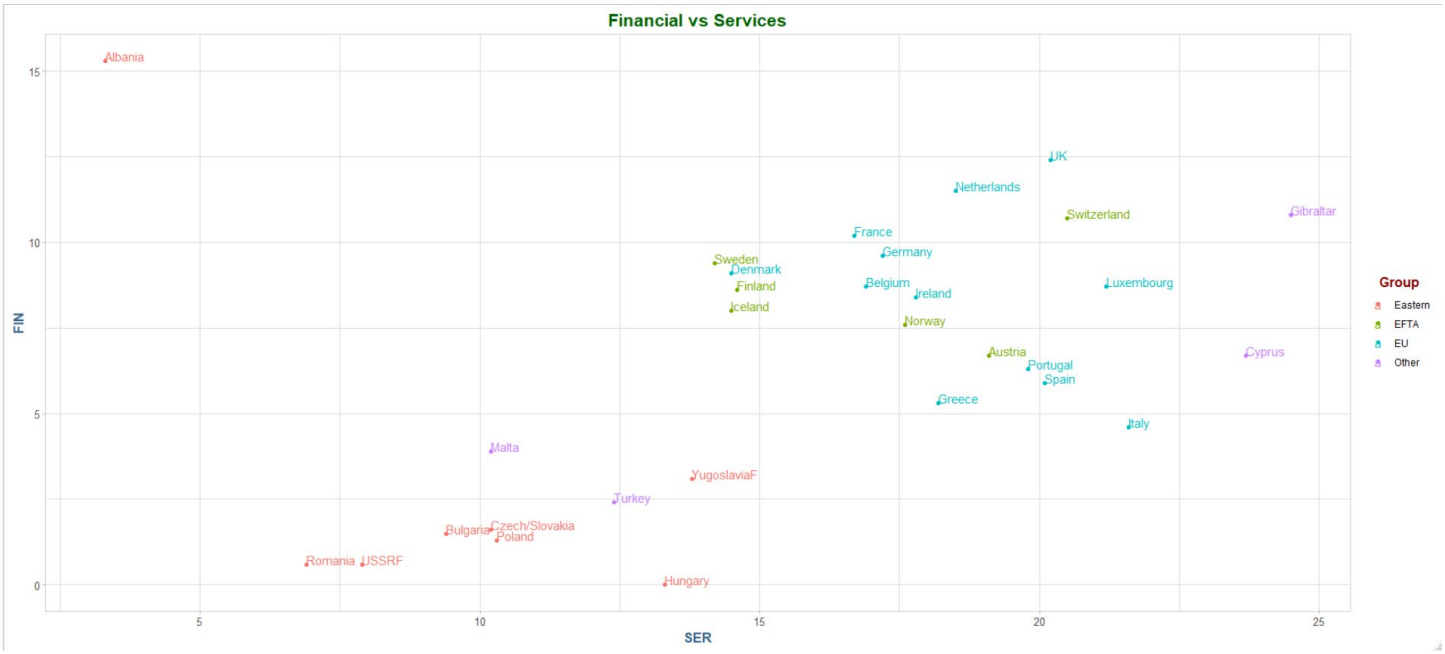
1.) Since we have a relatively small number of variables, we will begin our exploratory data analysis with a pairwise scatterplot. Obtain a pairwise scatterplot of the data. Note that when you have a small number of variables, the pairwise scatterplot is a useful statistical graphic. Another note about scatterplots – they are not very useful when we have too many data points.



We are looking for groupings in the lower diagonal of the above chart. Some notables for further examination include, MAN/SER, SER/FIN, TC/FIN, SPS/SER.

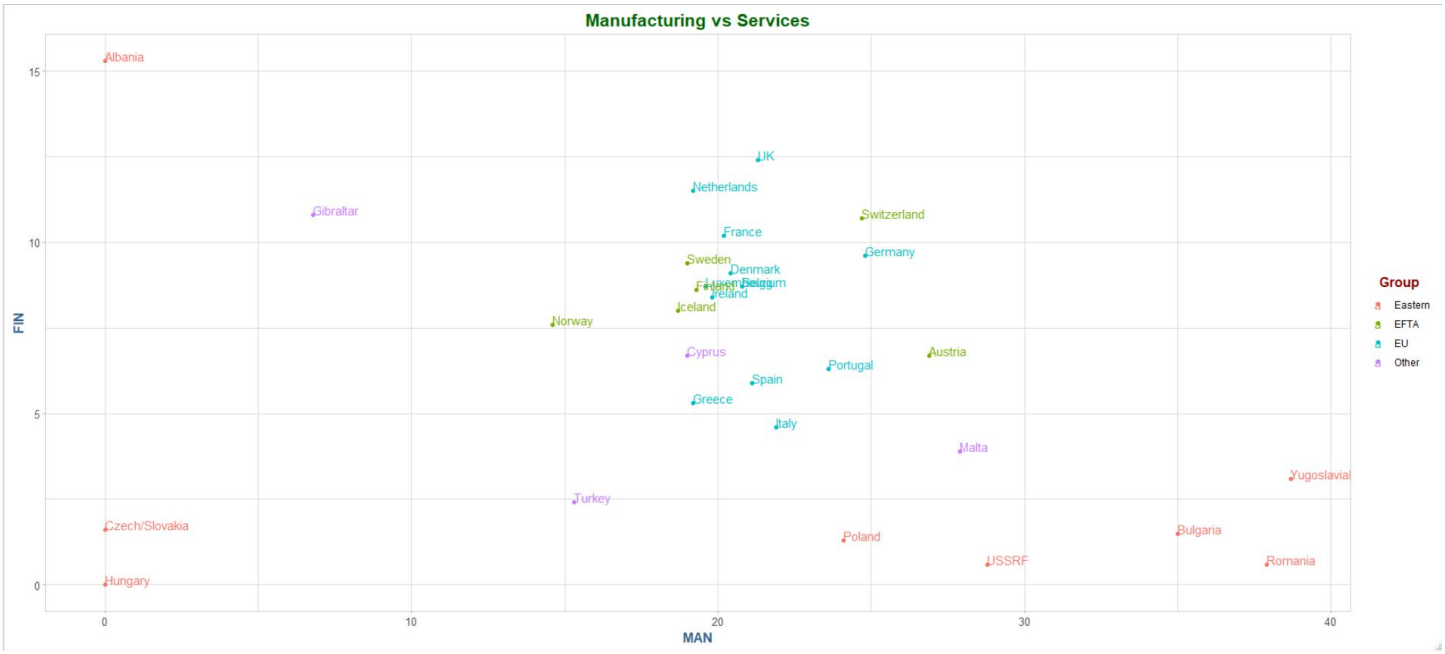
2.) Zooming in on some selections from above:

a. FIN/SER



In the above chart I would think there are two or three distinct clusters/segments.

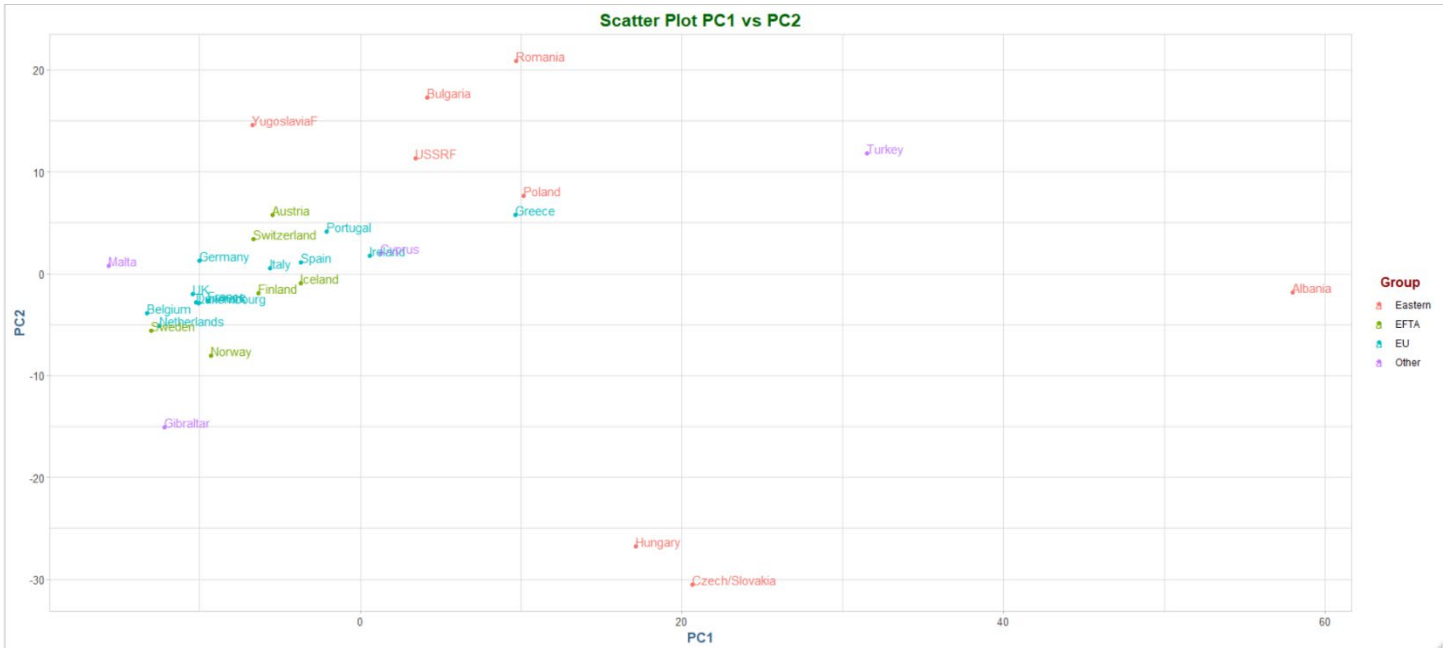
b. MAN/SER



In the above plot I see four distinct clusters, which are much different than the previous plot.

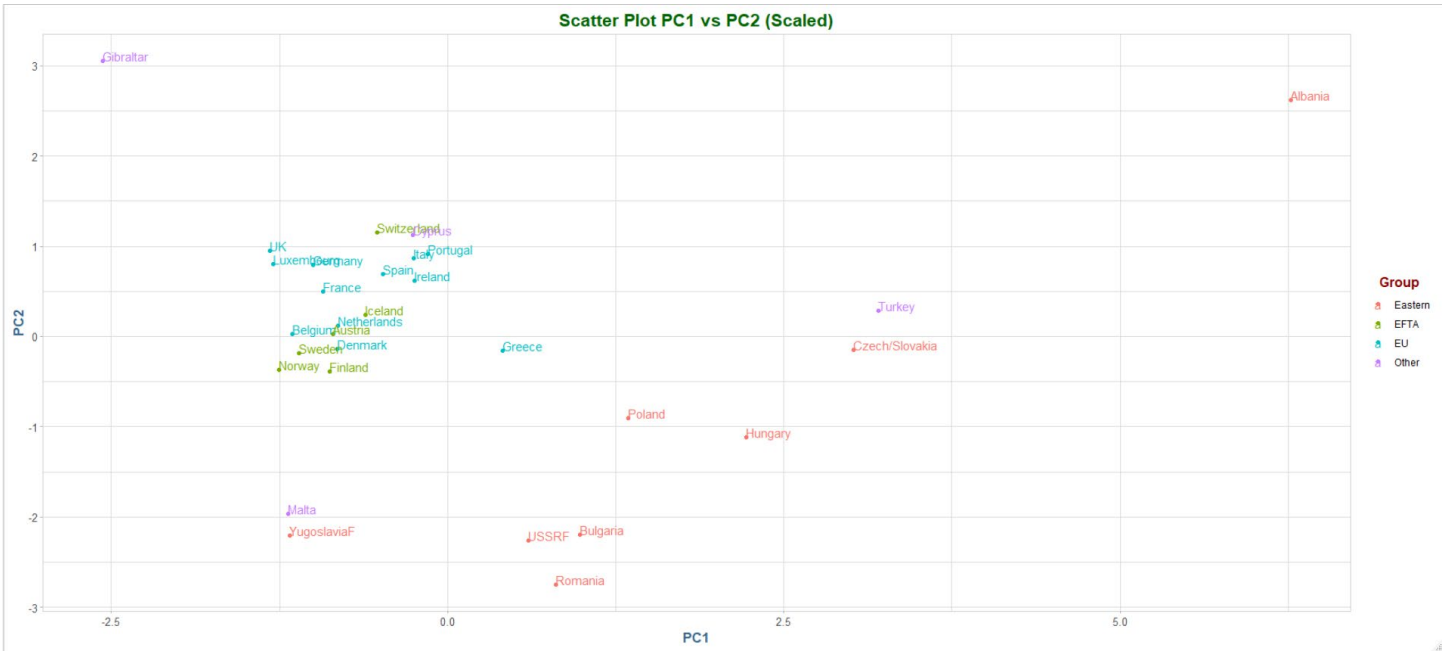
c. I would think that that the first graph would be easier for an algorithm to cluster due to the clearer lines of separation.

3.) We can use principal components analysis to reduce the dimension of the data. We can project the data down from 9D to 2D by performing PCA and using the first and second principal components. By doing so we are creating a new 2D view of the data, and a view of the data that contains information from more than two dimensions.



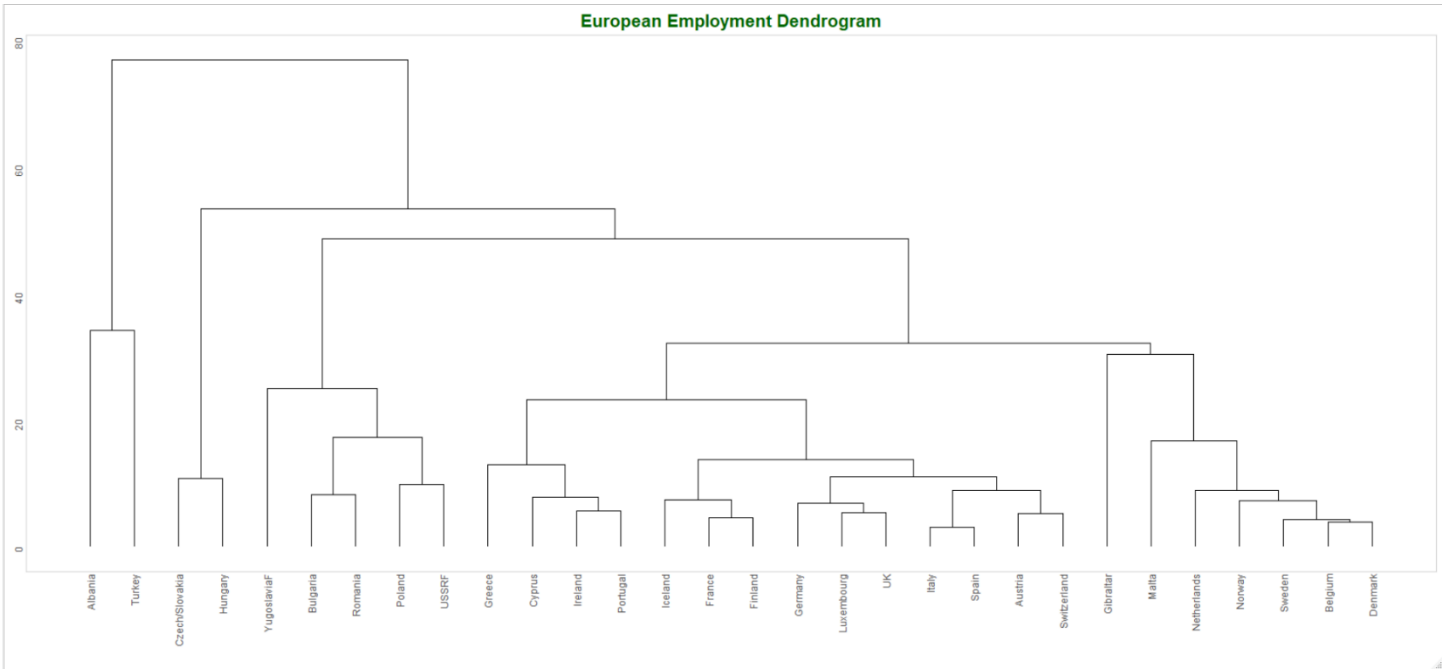
a. The first two principal component loadings:

b. The scaled version of the PCA does show some differences in the clusters.

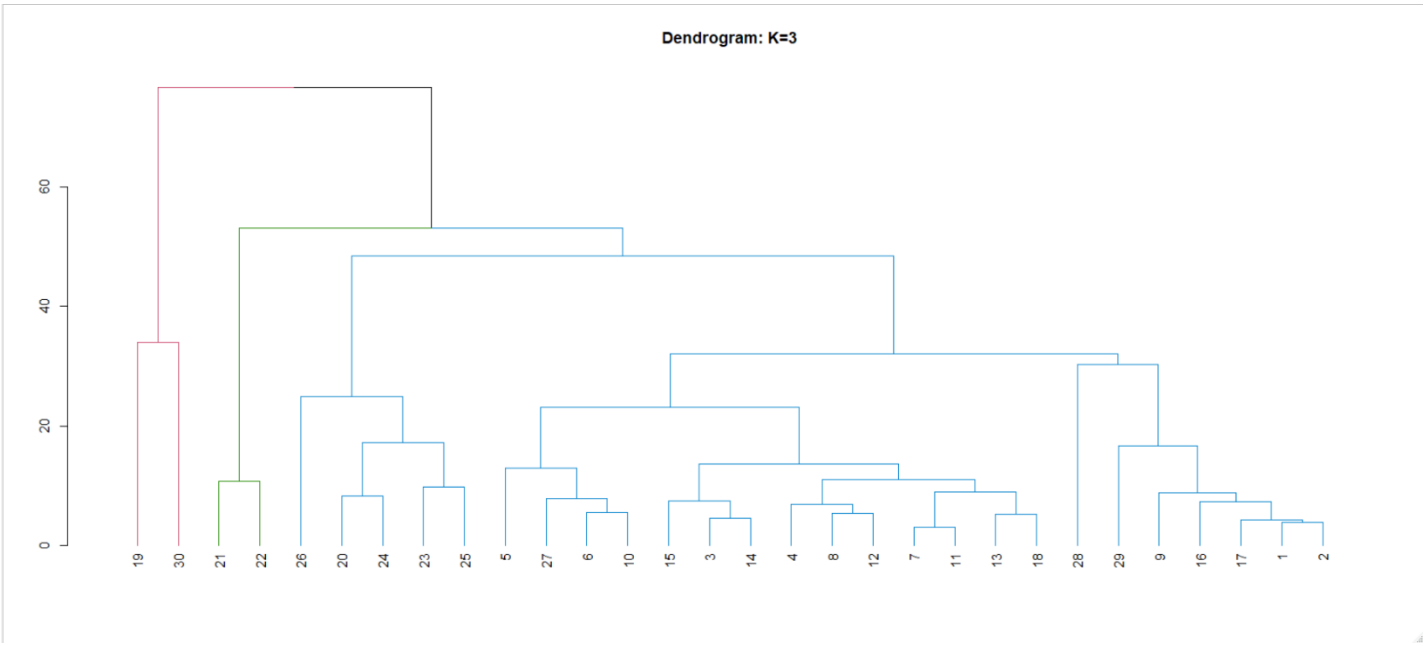


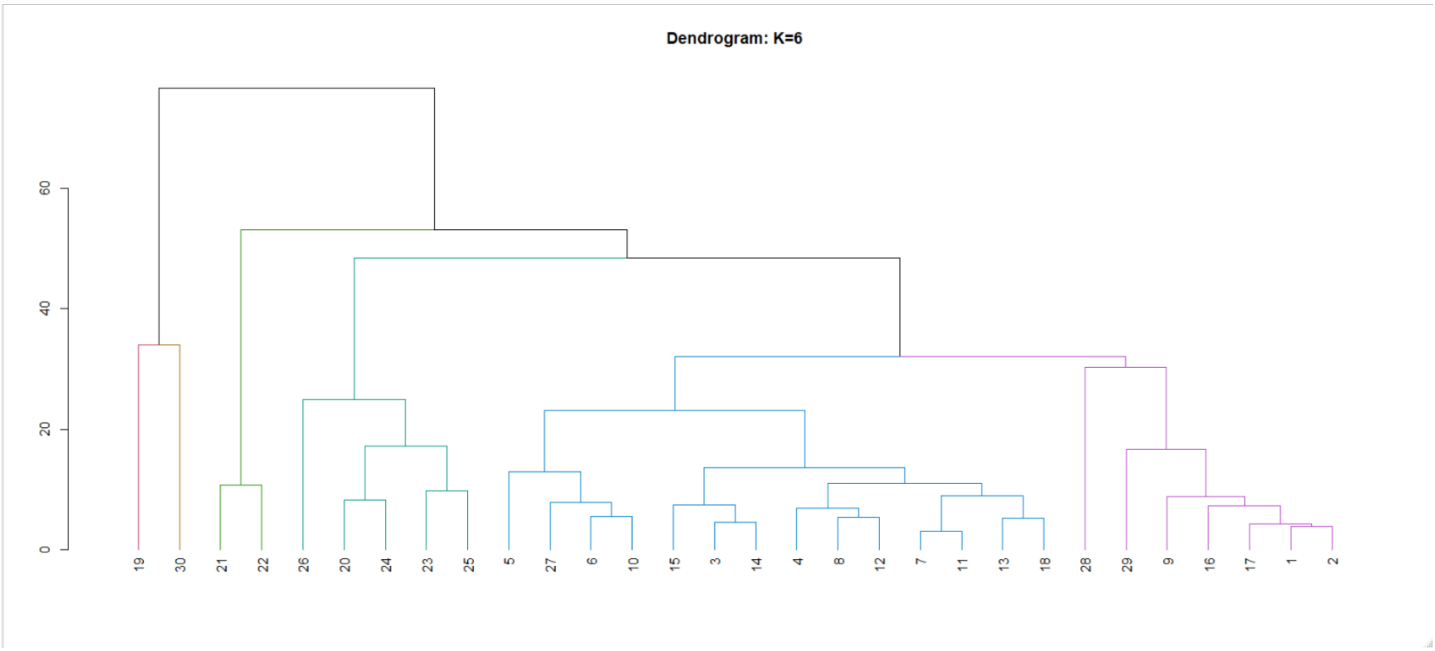
4.) Hierarchical clustering algorithms fit a tree of clusters from k=2 to k=N, where N is the number of data points in the sample. As you know, this tree of clusters can be visualized using a dendrogram. Since the cluster tree stores all possible cluster assignments, we must cut the tree using cutree() to force an assignment of the observations to a particular number of clusters.

a. Obtain a dendrogram:



Dendrogram cuts:



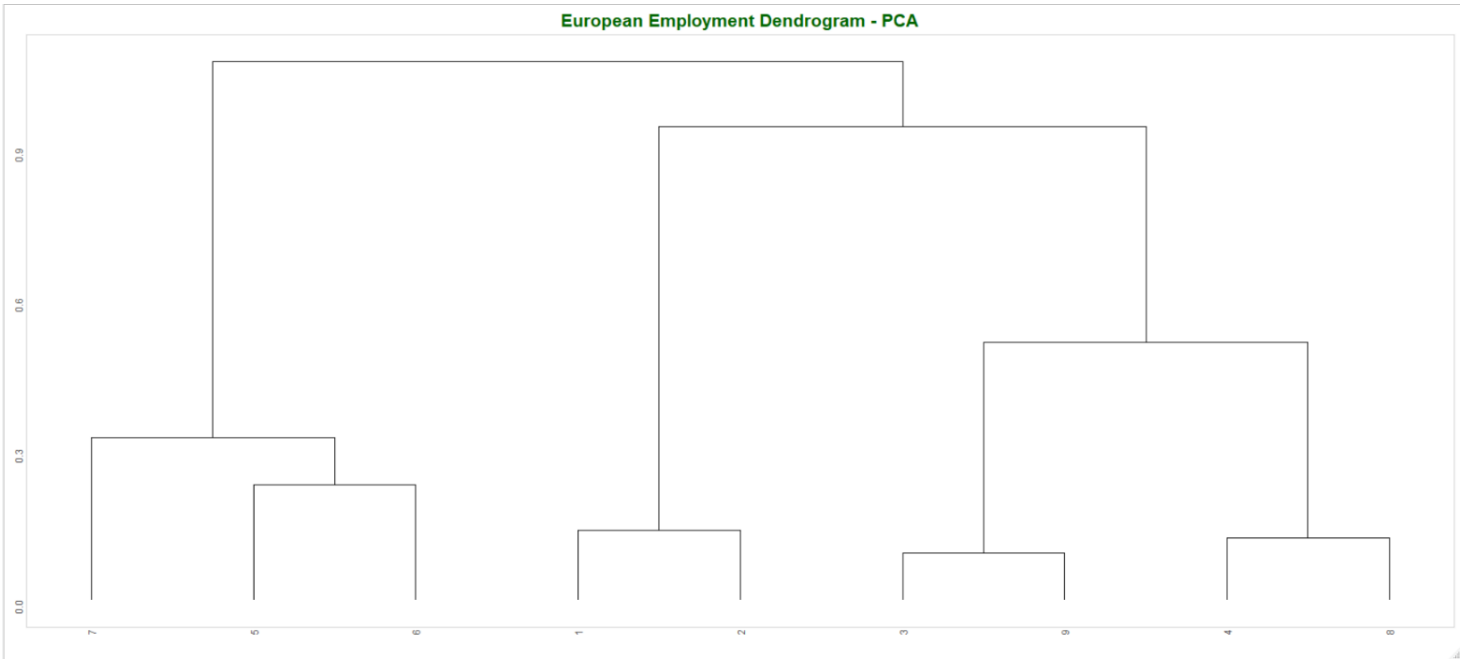


Classification Accuracy

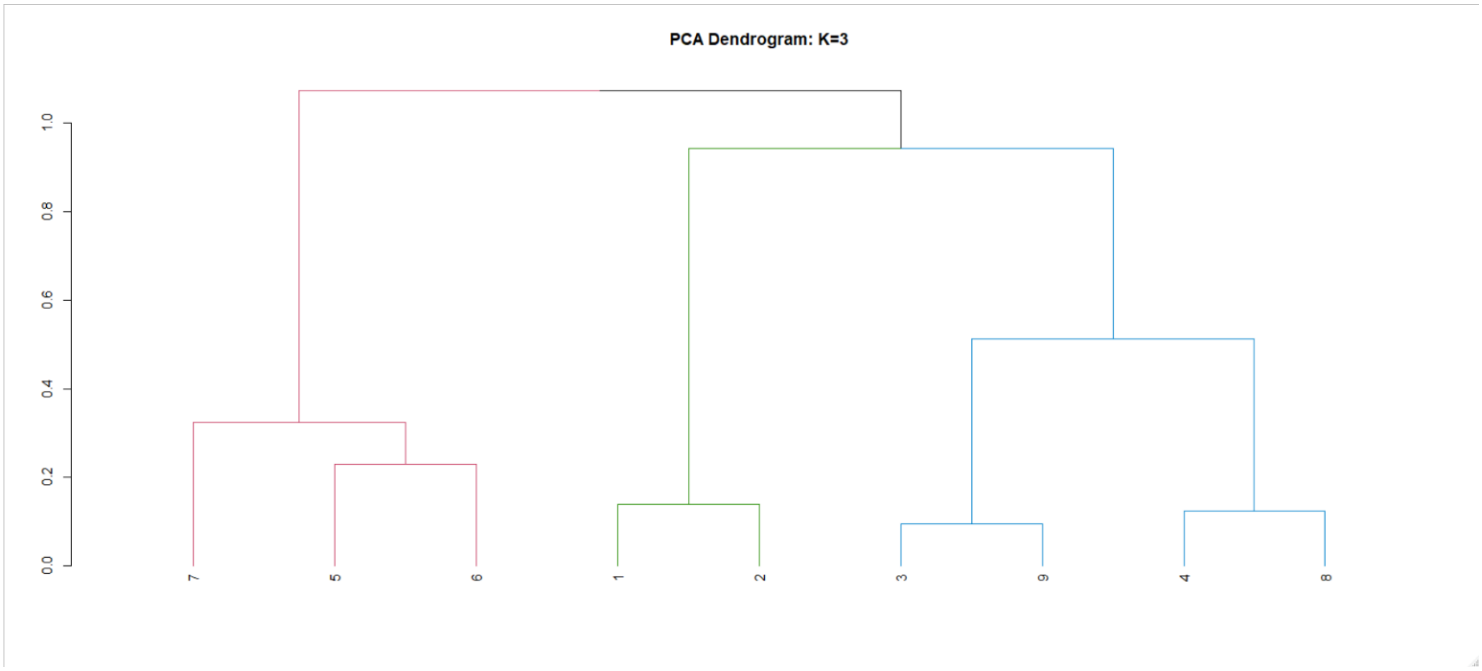
K	Error	Pct
3	5331.018	0.5893374
6	2049.701	0.8421061

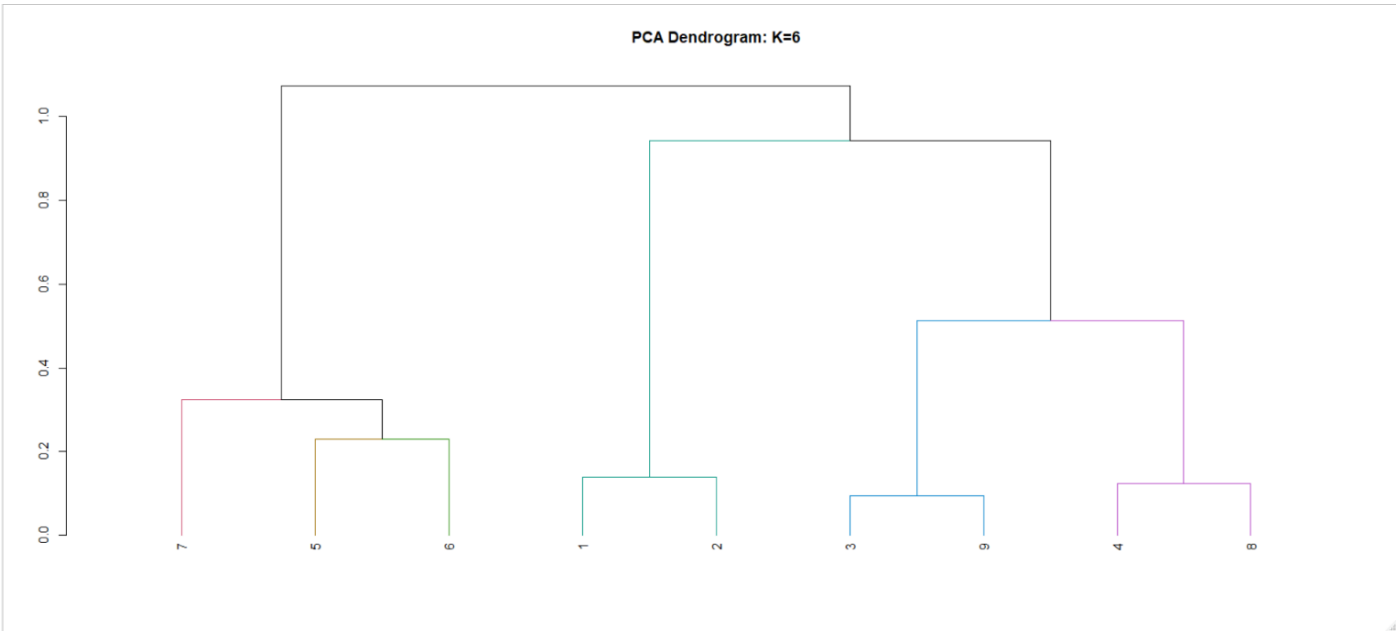
Six clusters are a better fit here.

b. Perform the same analysis, but this time use the principal component space using the first and second principal components. Of these four 'cluster models' which one is the most accurate? Make a table to display their accuracy for easy comparison.



Now, color code the nodes by cut:



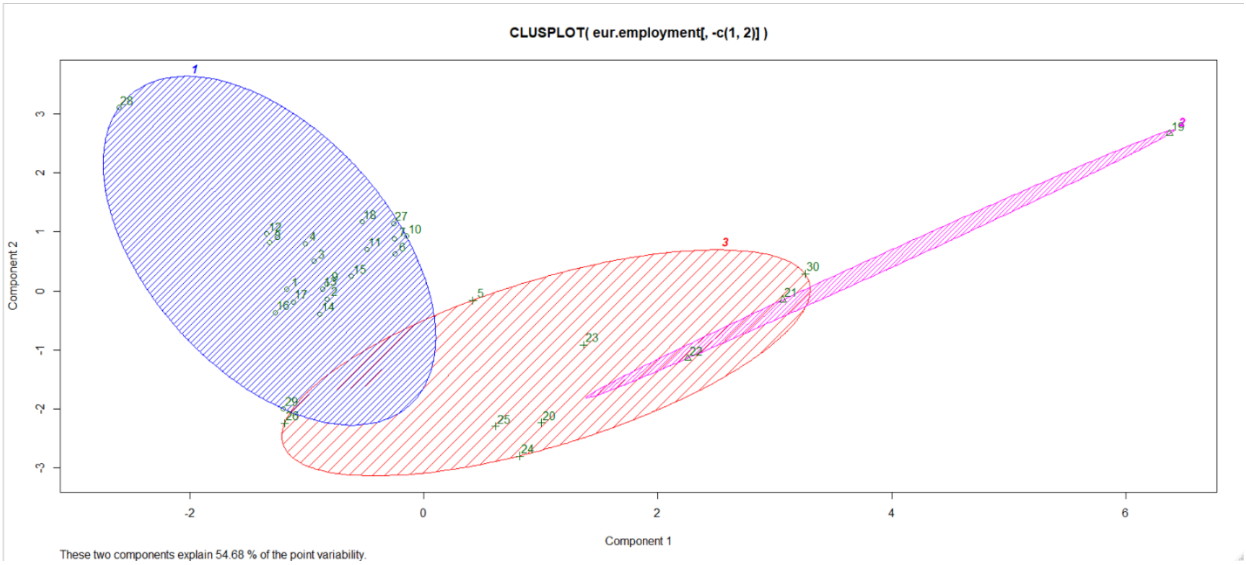


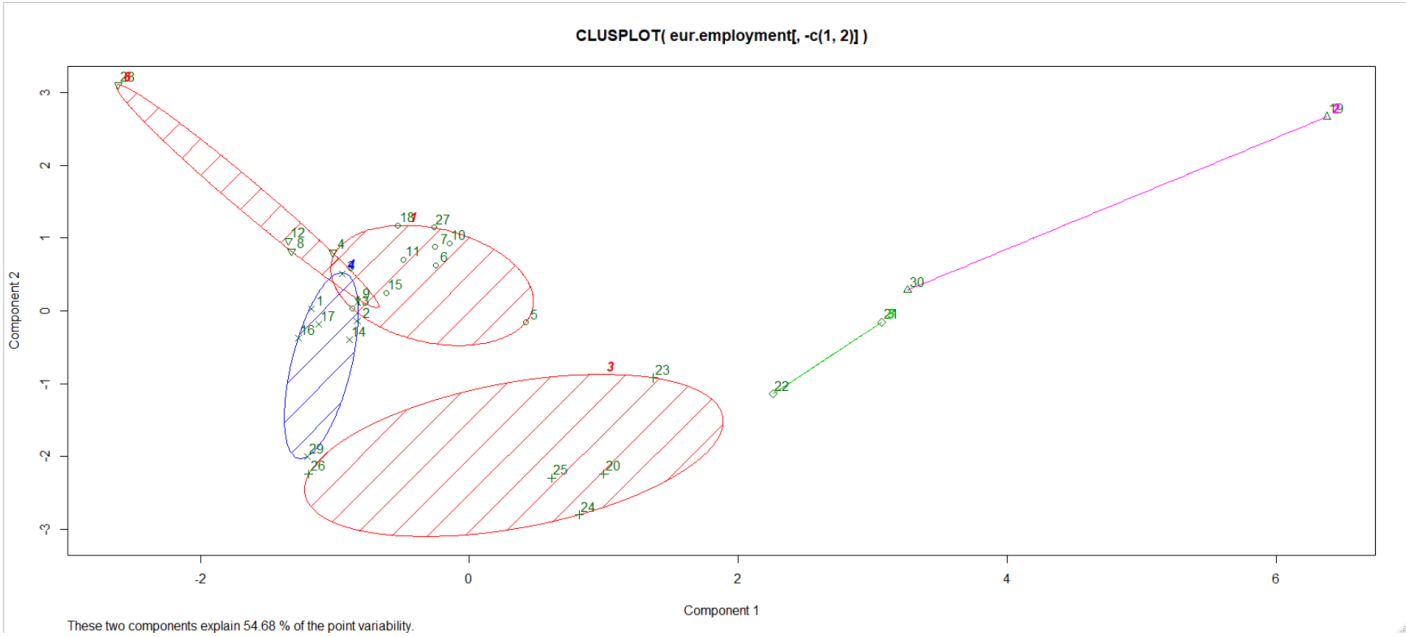
And a summary of the classification accuracy of all four methods:

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

It appears that the PCA clusters with k=6 is by far the best performing.

5.) Let's perform the analogous cluster analysis and make a comparison.





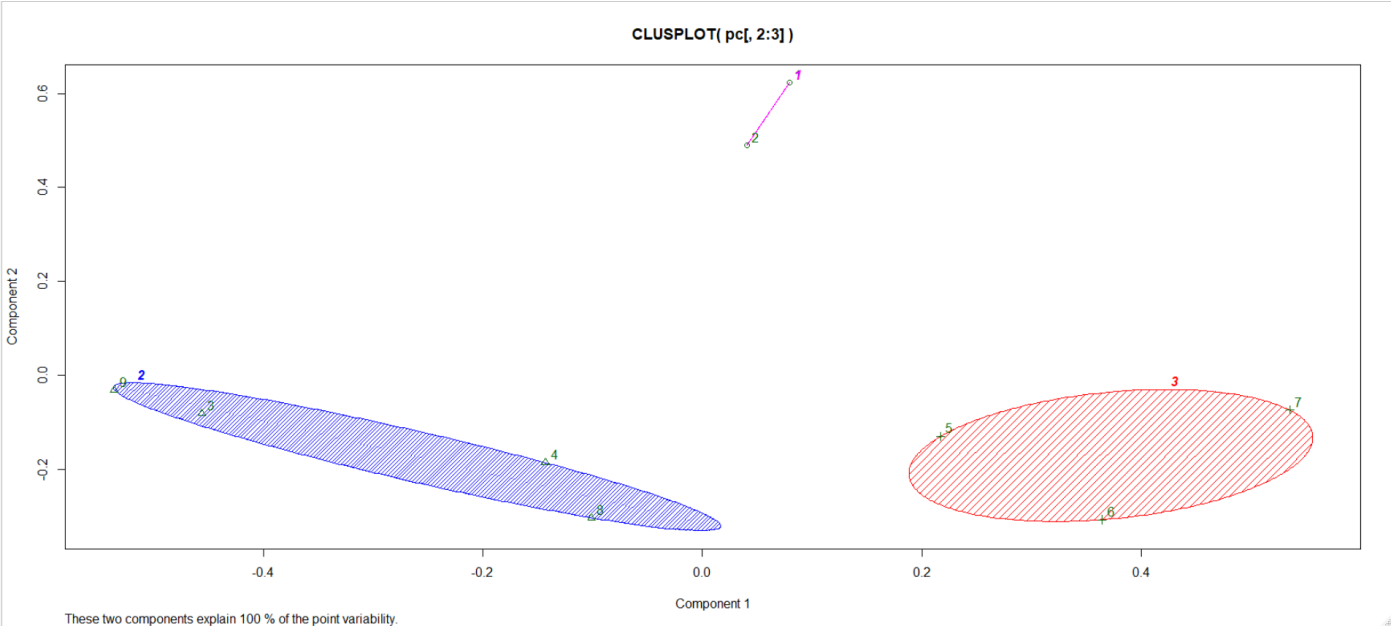
Clustering Method Comparison

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

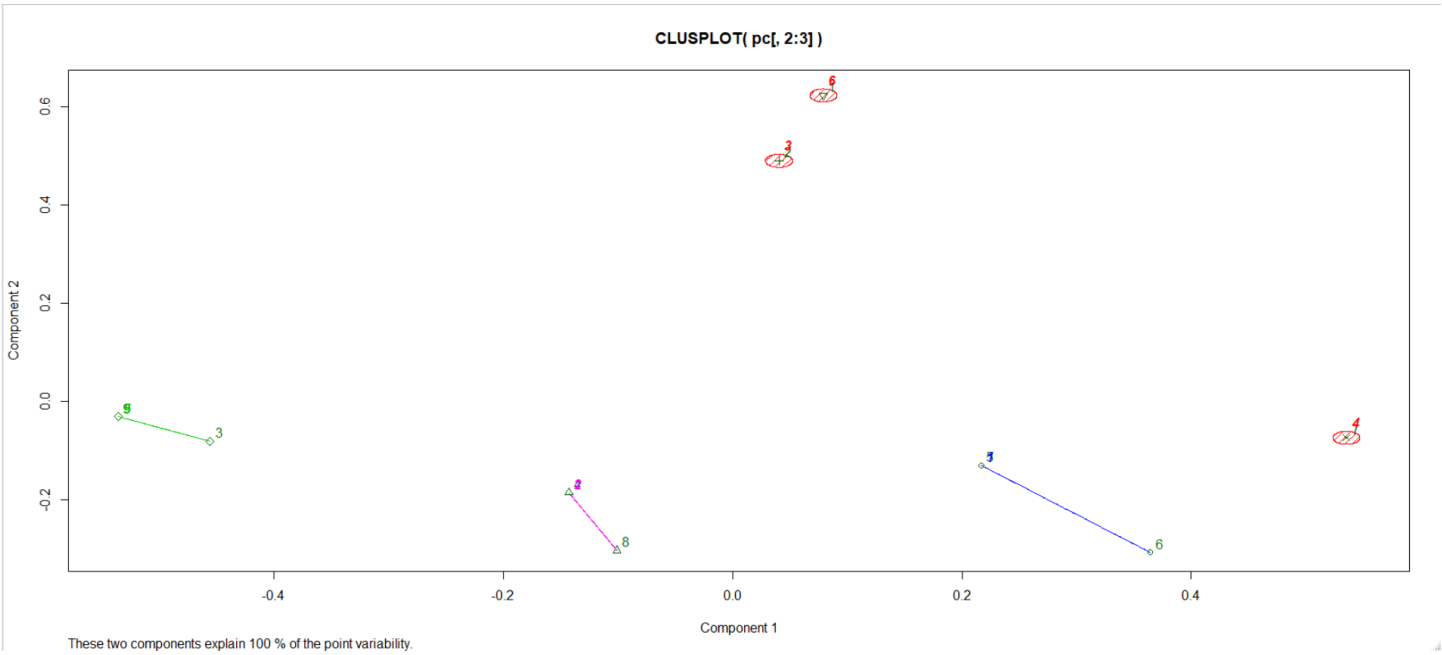
The classification accuracy for KNN methods are slightly worse to their hierarchical peers (non-PCA), although the different is relatively small. The three-cluster plot for KNN has substantial overlapping clusters and looks to be an insufficient fit to the data. The six-cluster plot looks much better and has substantially less overlap in the clusters.

Now, let's do the same KNN analysis on the PCA results.

K=3


CLUSPLOT(pc[, 2:3])
Component 2
Component 1
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
0.4
These two components explain 100 % of the point variability.

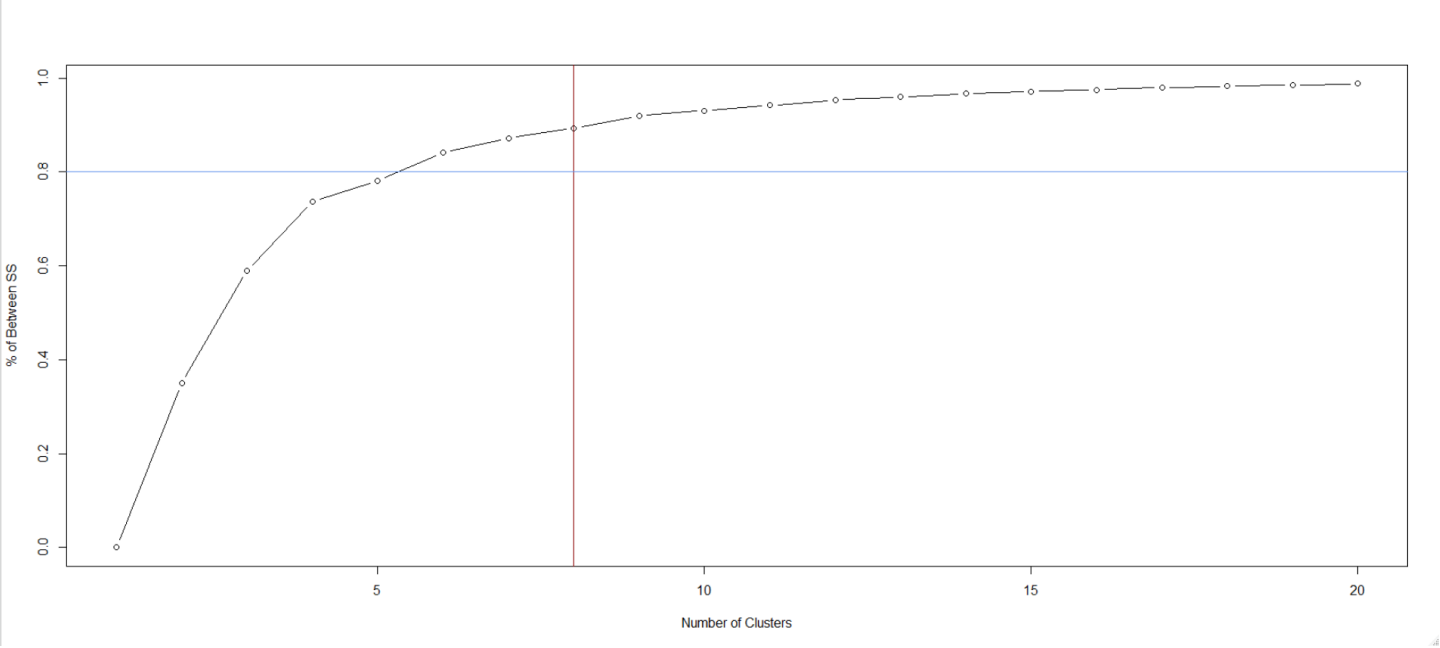
K=6


CLUSPLOT(pc[, 2:3])
Component 2
Component 1
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
0.4
These two components explain 100 % of the point variability.

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 'Eastern' between clusters 1 and 3. The k=6 clustering disperses the countries pretty evenly through the 6 groups, there isn't one cluster that exhibits dominance throughout the groups.

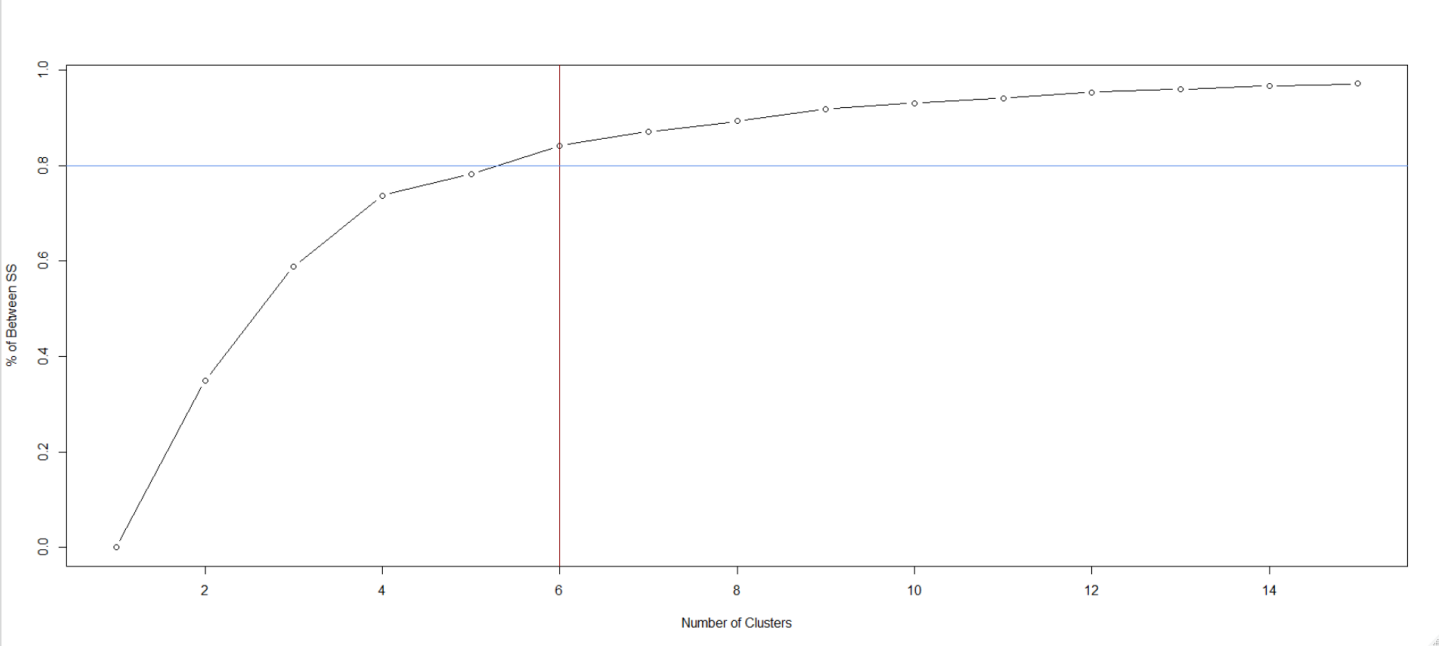
6.) Obtain and plot the classification accuracy for k=1 to k=20 for both hierarchical and k-means clustering algorithms. What can you conclude based on this graph?



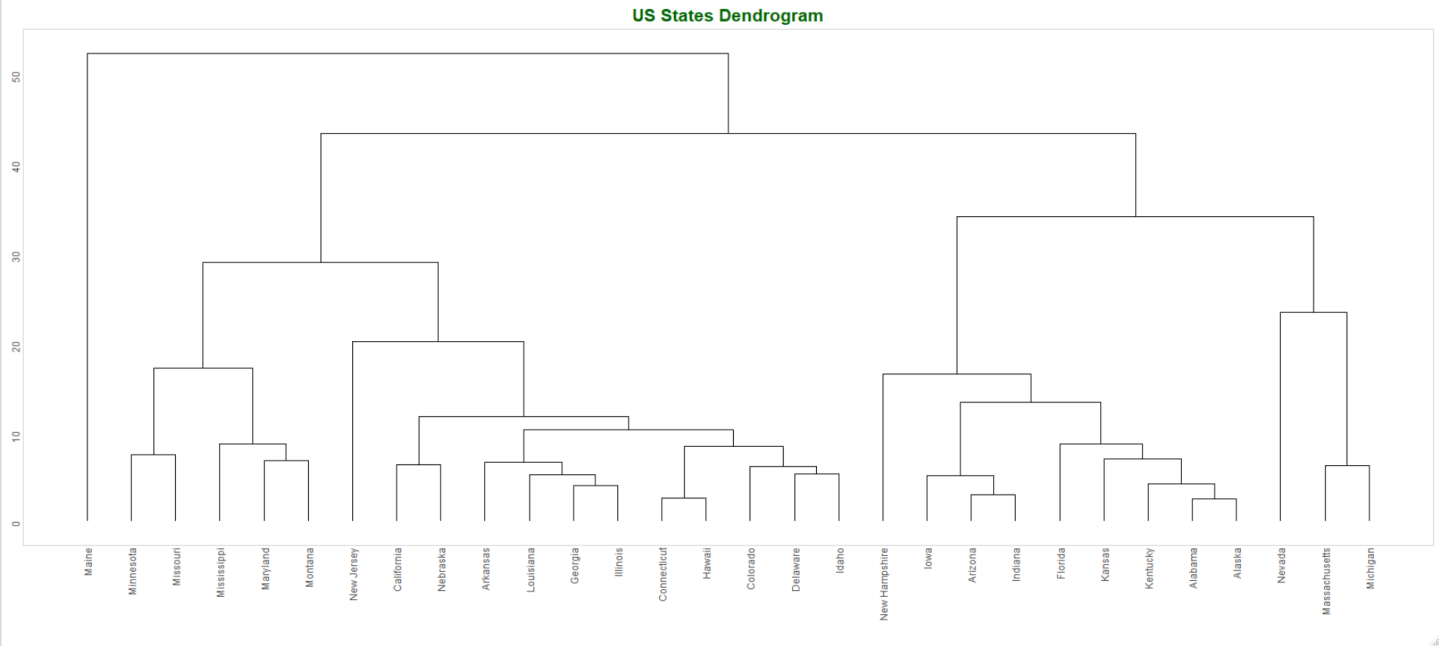
From the graph above, if we are looking to explain the largest proportion of variance, we would choose k=8 (in red), as the diminishing returns after that cutoff are pronounced. However, if we were to set a priori cutoff of say, 80%, we would pick k=5 (blue). I think an optimal number of k in this scenario is 5, as we explain a great deal of the variance in the data without introducing a great deal of bias.

7.) US States Modeling

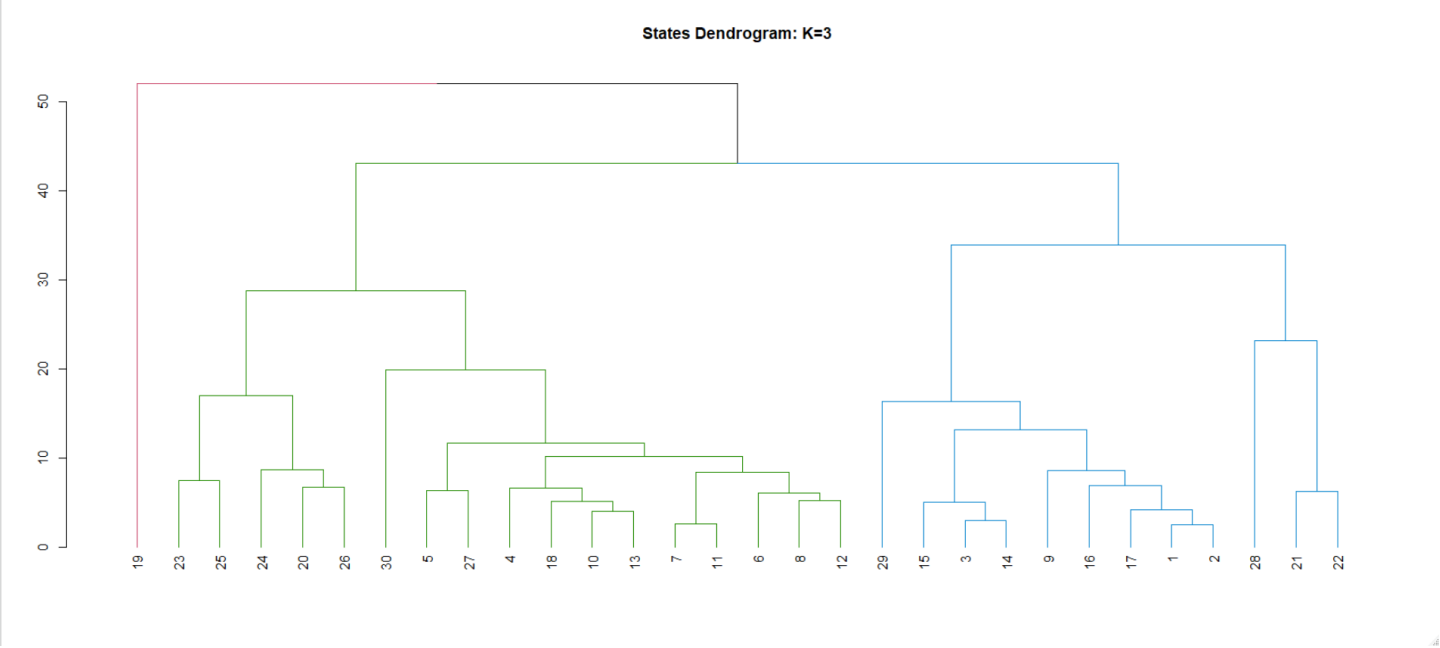
When we load the US States dataset, we can run it through similar analysis that we executed previously to look at the variances explained at different number of clusters.



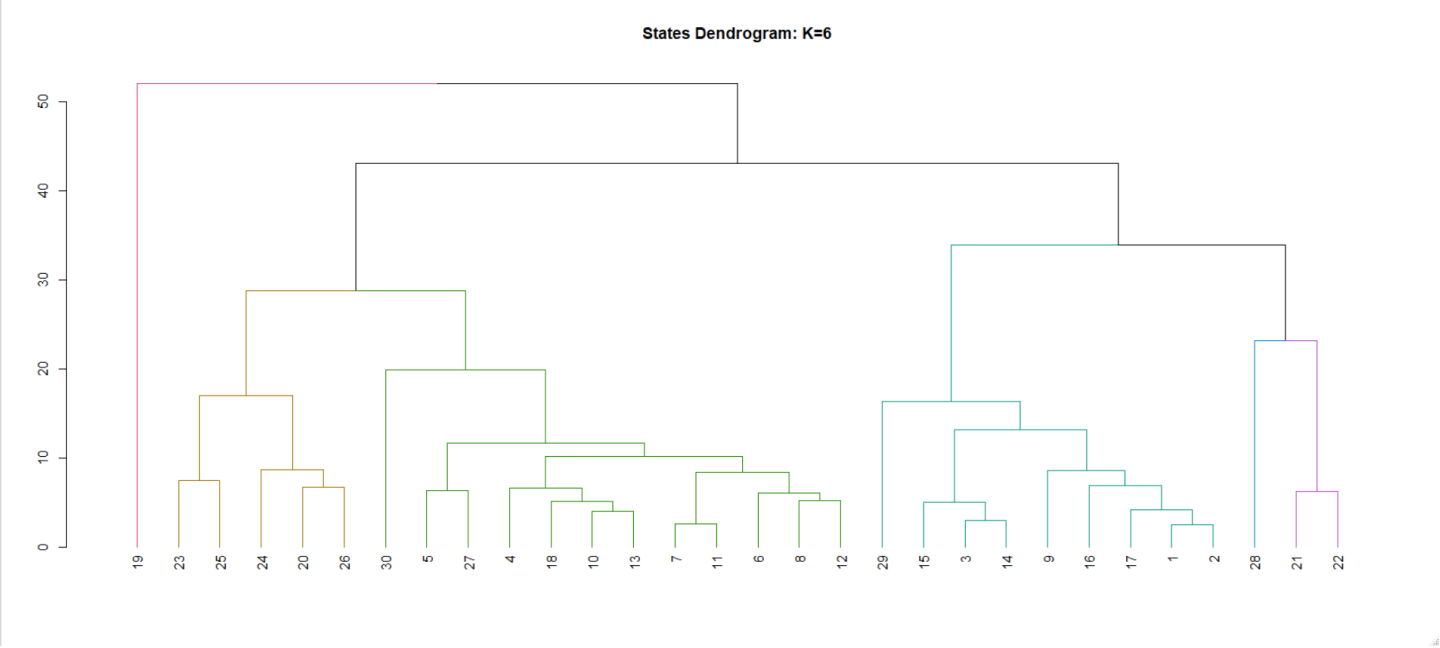
In the plot above, using the priori cutoff of 80% we get right below 6 clusters. However, there is a sharp decline in explained variance after k = 3, so we will explore both. Looking at the dendrogram for the US States data:



While there does seem to be a great deal of clustering in a few regions, there seems to be more dispersion in the leaf's that we can explain with k=3, however, we will see how it looks:



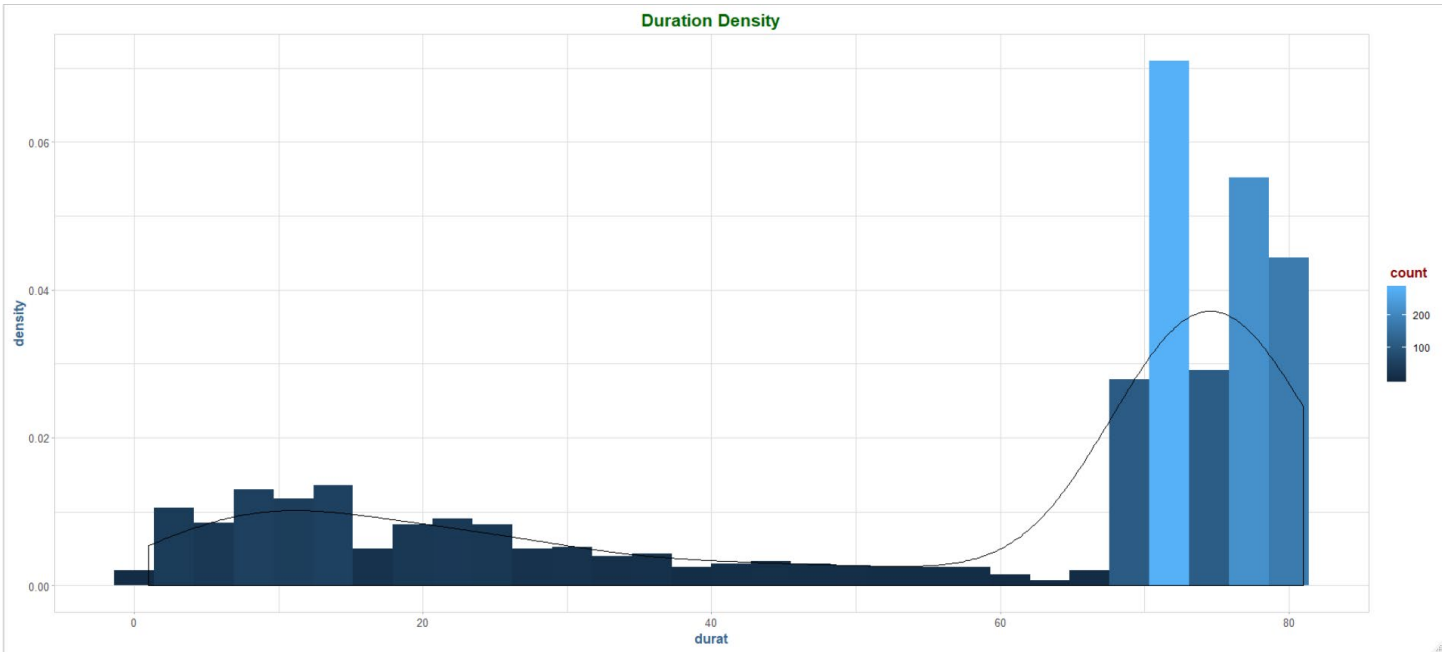
In the k=3 dendrogram, we note that there is a heavily unequal distribution of two of the nodes, leaving Maine on an island. Increasing k=6 we see a much more equal balancing of the nodes:



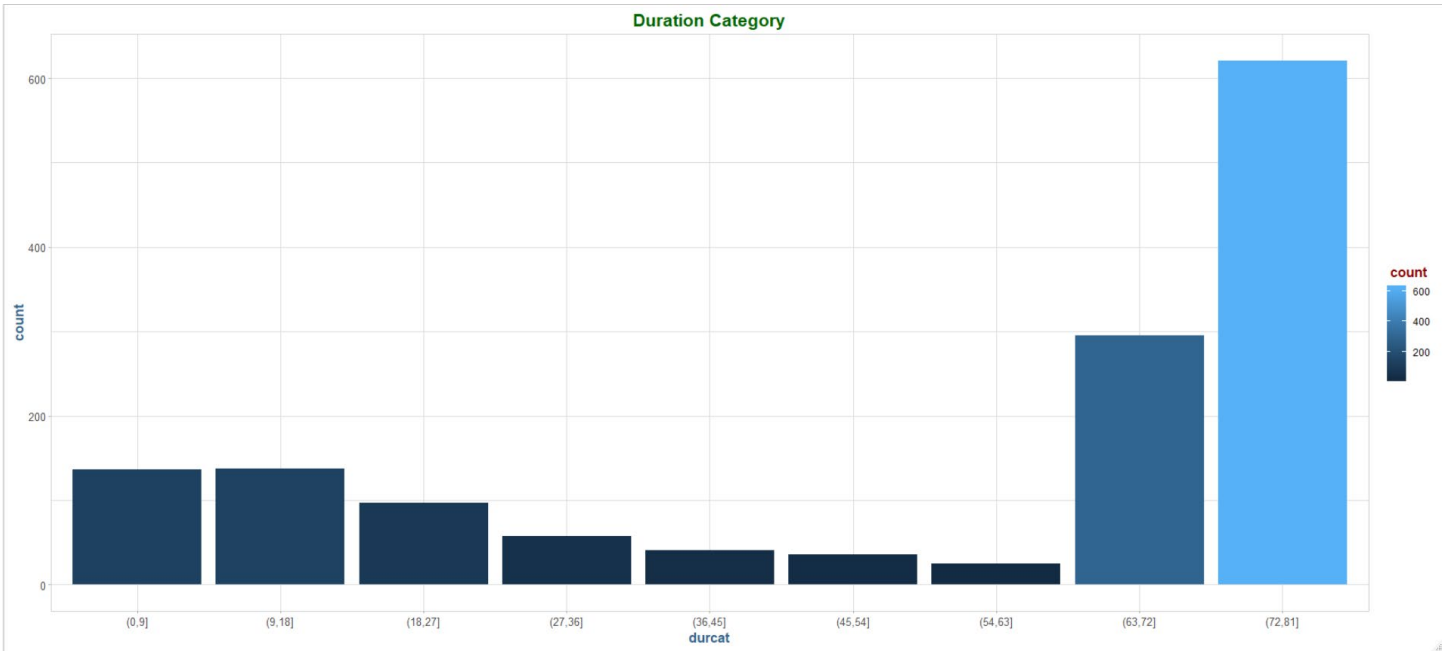
Looking at the errors and the accuracy from the k=3 and k=6 models, we note that k=6 appears to be at/near optimal for this data using a hierarchical clustering model.

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

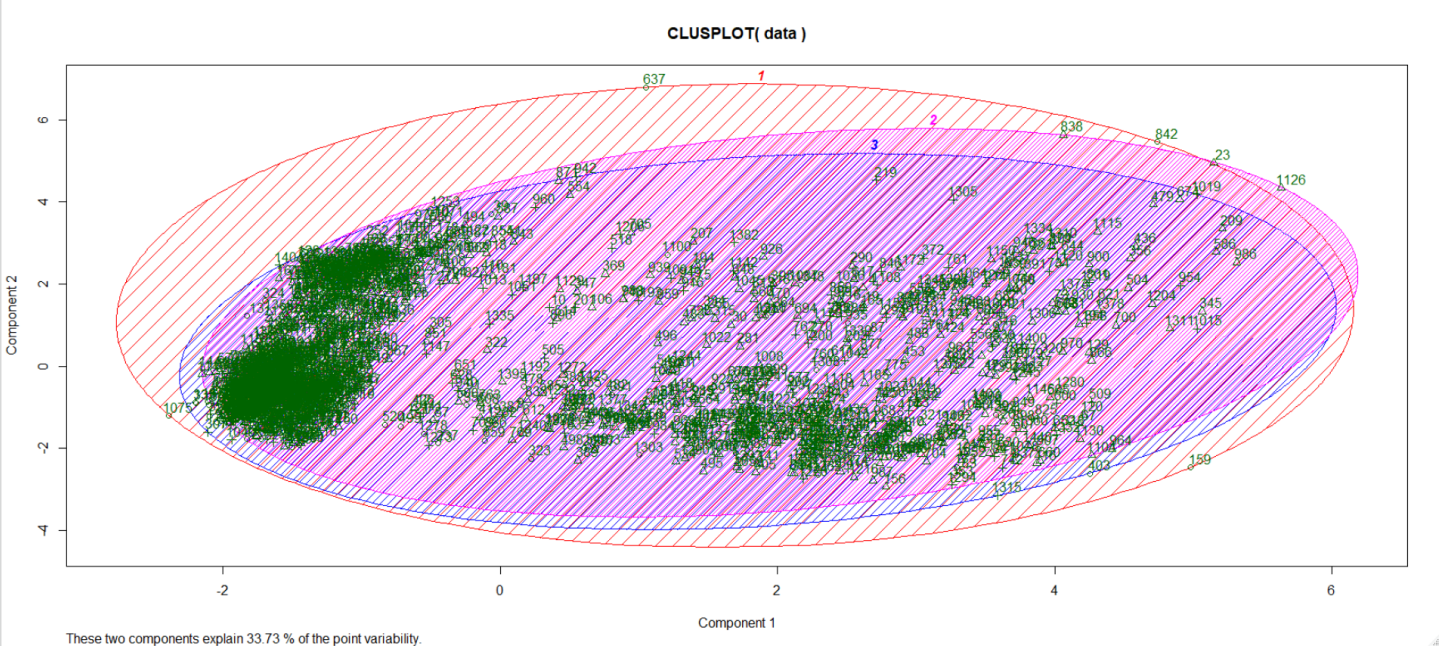
8.) When looking at the recidivism data, we note the lack of a categorical label like we had in the previous datasets. First, we are going to look at the duration variable, as it was the main variable under study for the research it was published with.

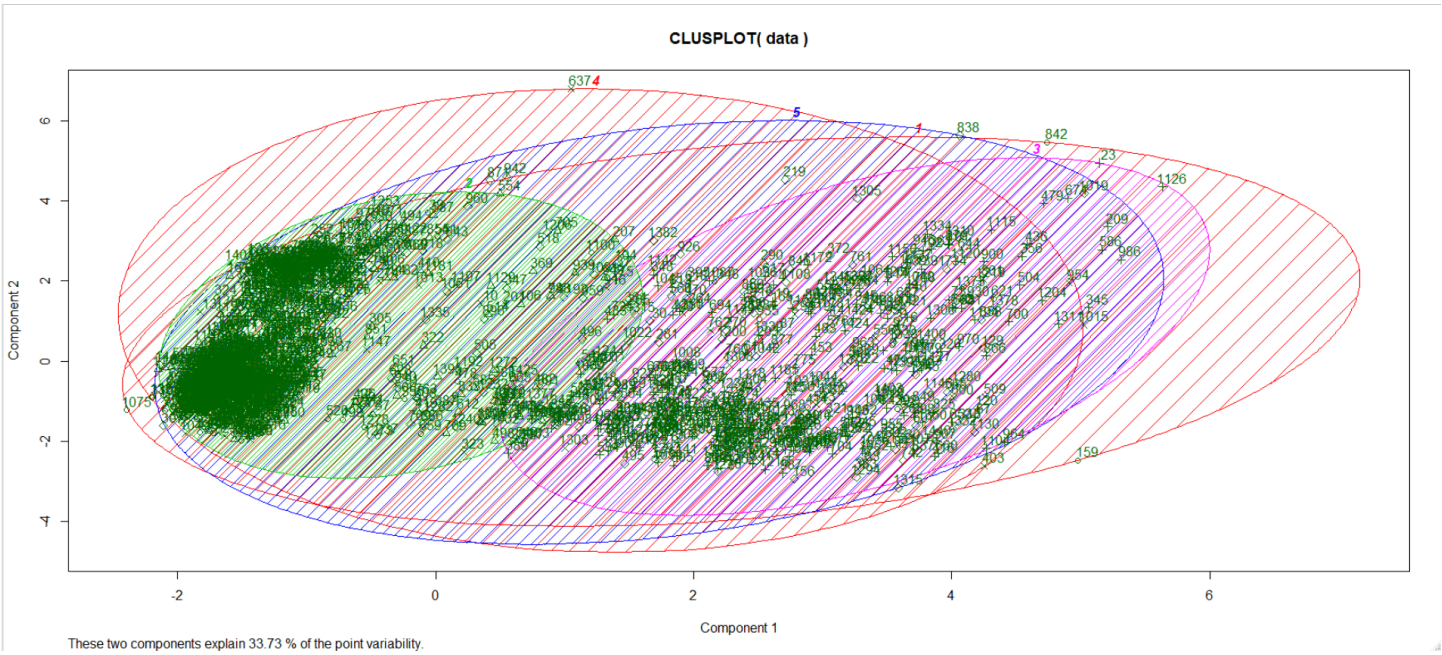


We are going to bin these together in evenly spaced categories, in a new variable called durcat, which we might useful in further analysis:



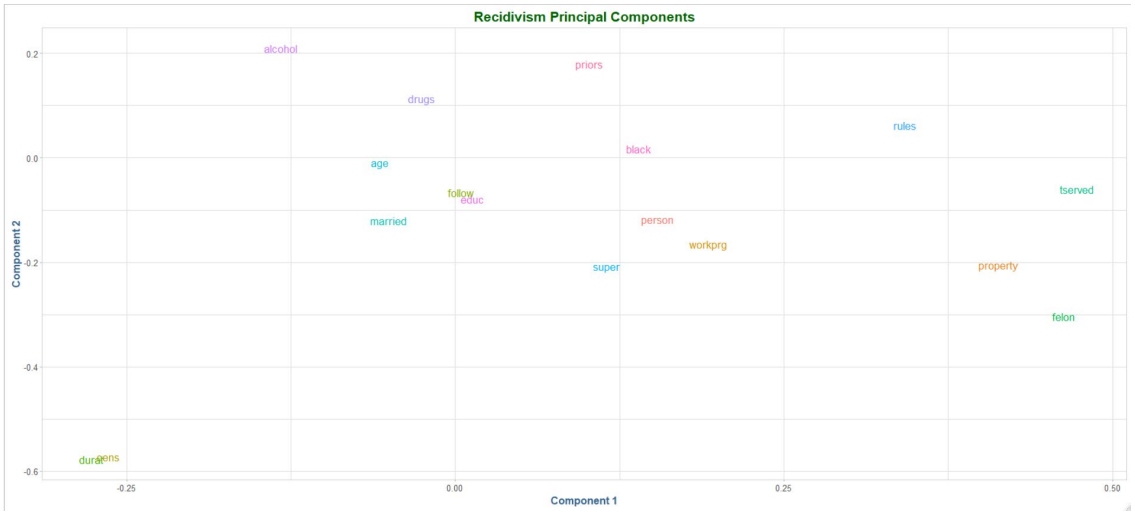
Next, we are going to execute a KNN with a couple of different cluster sizes, k = 3 then 5:



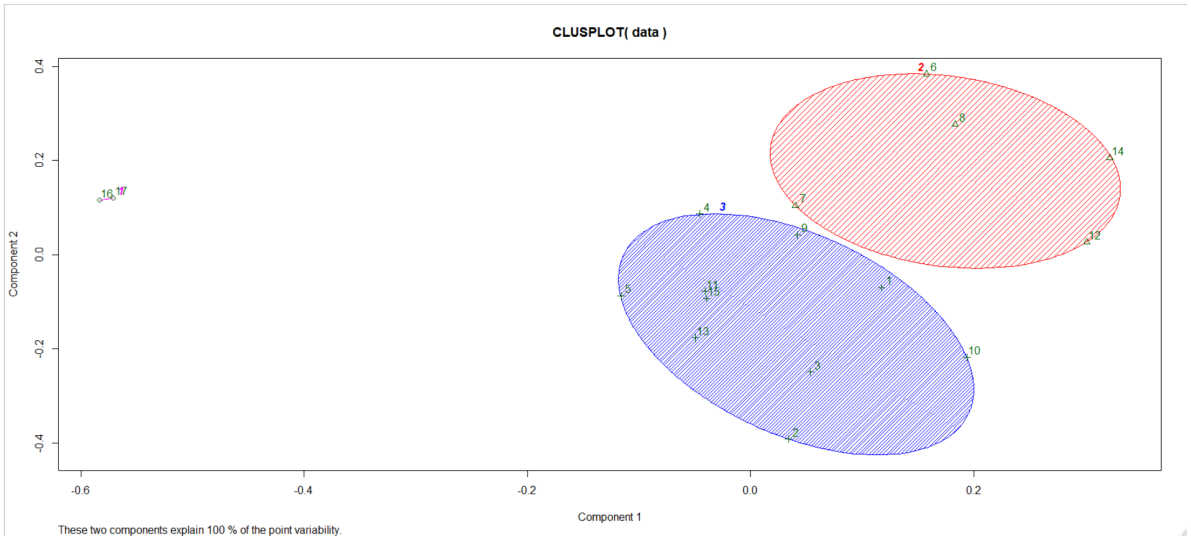


The groups have a significant amount of overlap, although we are starting to see some separation at k=5 within the main clusters in the "middle" of the dataset.

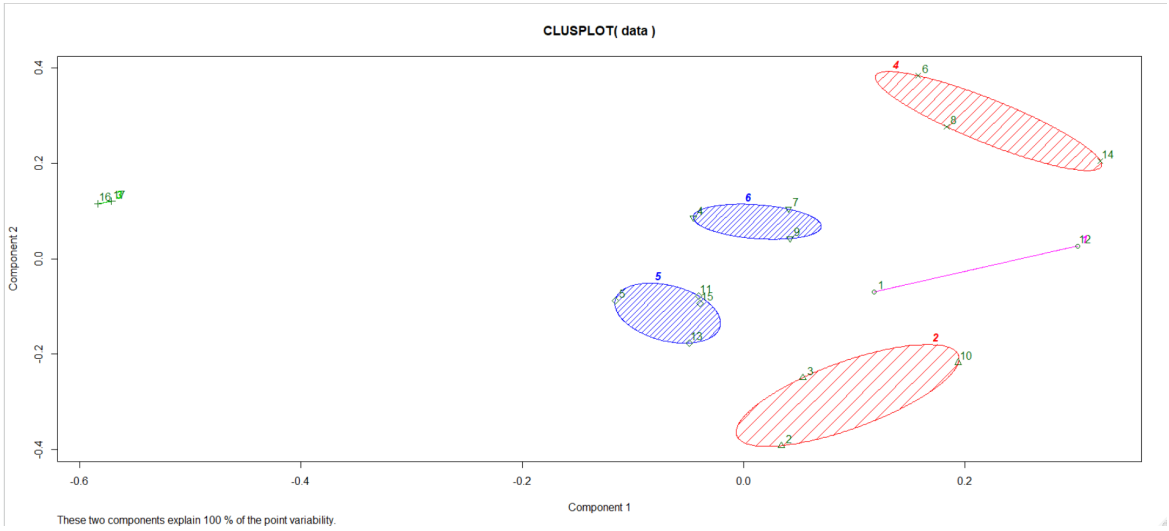
Let's try running PCA first, to see if we can do any better with these groupings.



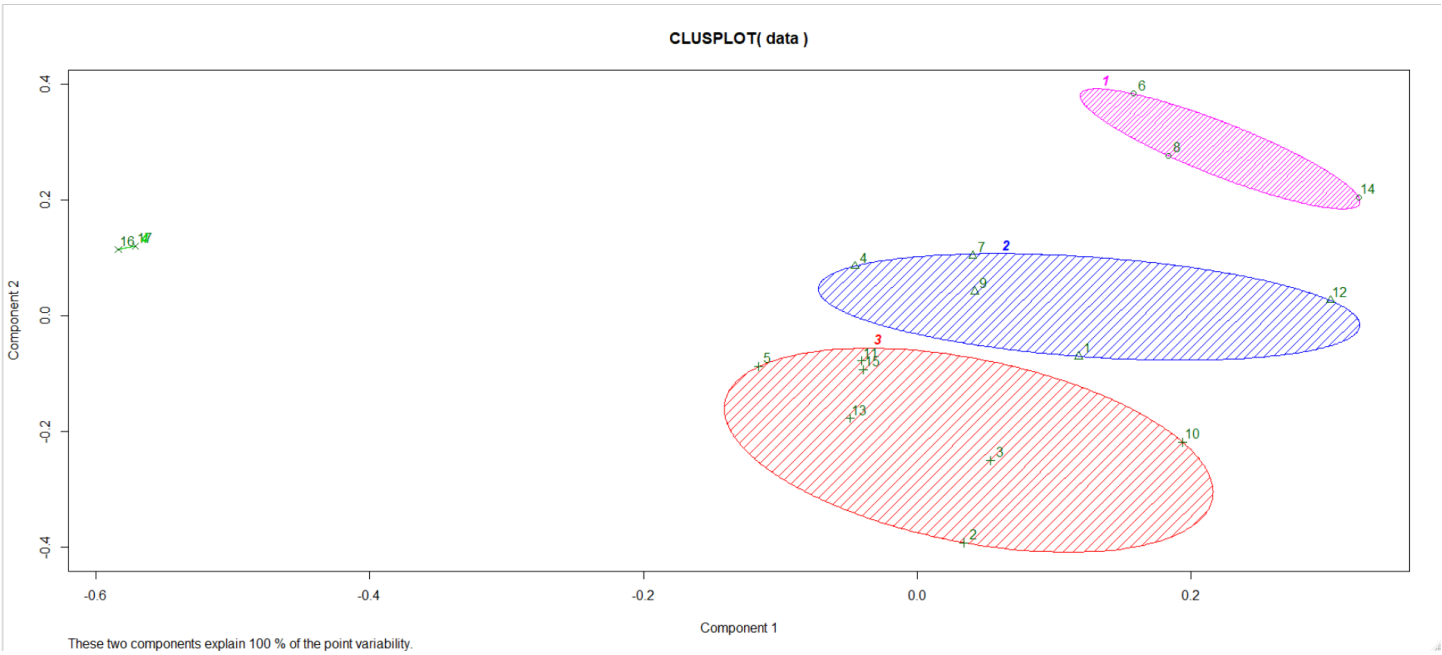
Post PCA, we again run KNN with k=3:



Which is a much better result than we had prior. However, we still see there is a great deal of variance in the 2 clusters, with some outliers on the far end. With k=6 we get more diversity, however almost half the groups appear to contain only a couple of observations:



Finally, we will run it again with k=4:



At k=4 we see the majority of the data being clustered into the 3 primary groups, and only a few observations in the outlier bucket. Here, we will say the optimum number of clusters is 4.

CONCLUSION

This assignment was an informative one, in that it covered two drastically different unsupervised learning techniques to cluster related observations in the data together so that we may uncover hidden relationships. The first example was a bit more interesting to me in that this was really my first exposure to hierarchical clustering methods. The European employment dataset was somewhat different in that most of the datasets we look at are row-major, with many observations, and this dataset is pre-summarized like that of a census. Even with only a handful of observations, there is a great deal of hidden structure in the data, and it would be difficult to find using only traditional "brute-force" analysis (examining n*m statistical plots by hand would be exhausting). Here, the hierarchical analysis with the use of dendrograms made things much easier. The further use of principal components to reduce the dimensionality of the data I did not find particularly useful in this example. However, it did help a great deal with the classification metrics although it came with an interpretability penalty that I am not sure was worth it here.

The second method is the familiar KNN. Coming from a computer science background, I have a great deal of exposure to this algorithm in that I have had to write multiple implementations of it in various classes, as it is one of the simplest, yet surprisingly powerful algorithmic techniques out there. The KNN classification algorithm worked relatively well with the European employment data at k=6 without the use of principal components explaining most of the variance in the data. Again, here I think the use of principal components was not worth the tradeoff in interpretability.

Finally, in the two self-modeling exercises I found the one using the recidivism dataset to be of most interest. Perhaps I could be biased, as this dataset is extremely unwieldly using traditional exploratory data analysis techniques, and the unsupervised methods seem to cut right through it. The use both principal components coupled with KNN broke out well defined groups in this chaos of a dataset. I am glad this was part of the exercise, as I was starting to develop a bias against using PCA in combination with other techniques, and working through this one really drove home the point that you really need to try all avenues for deriving results before you more on. This was an informative exercise and I enjoyed working through it.