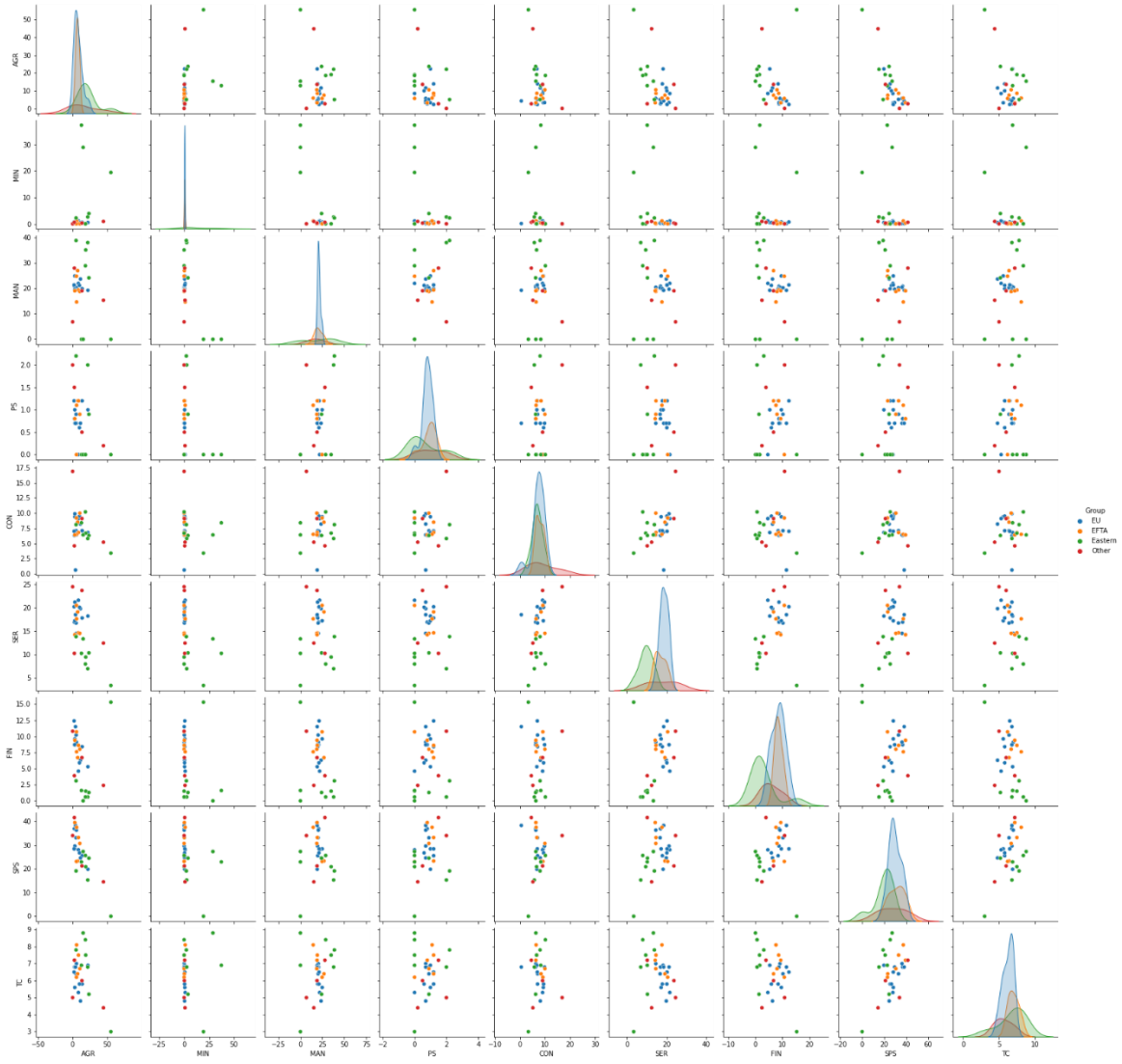
Michael Rocchio

MSDS 411

Assignment 4

Assignment Tasks:

(1) Initial Exploratory Data Analysis: 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. A scatterplot is a more useful statistical graphic when you have 100 data points than when you have 1MM data points.



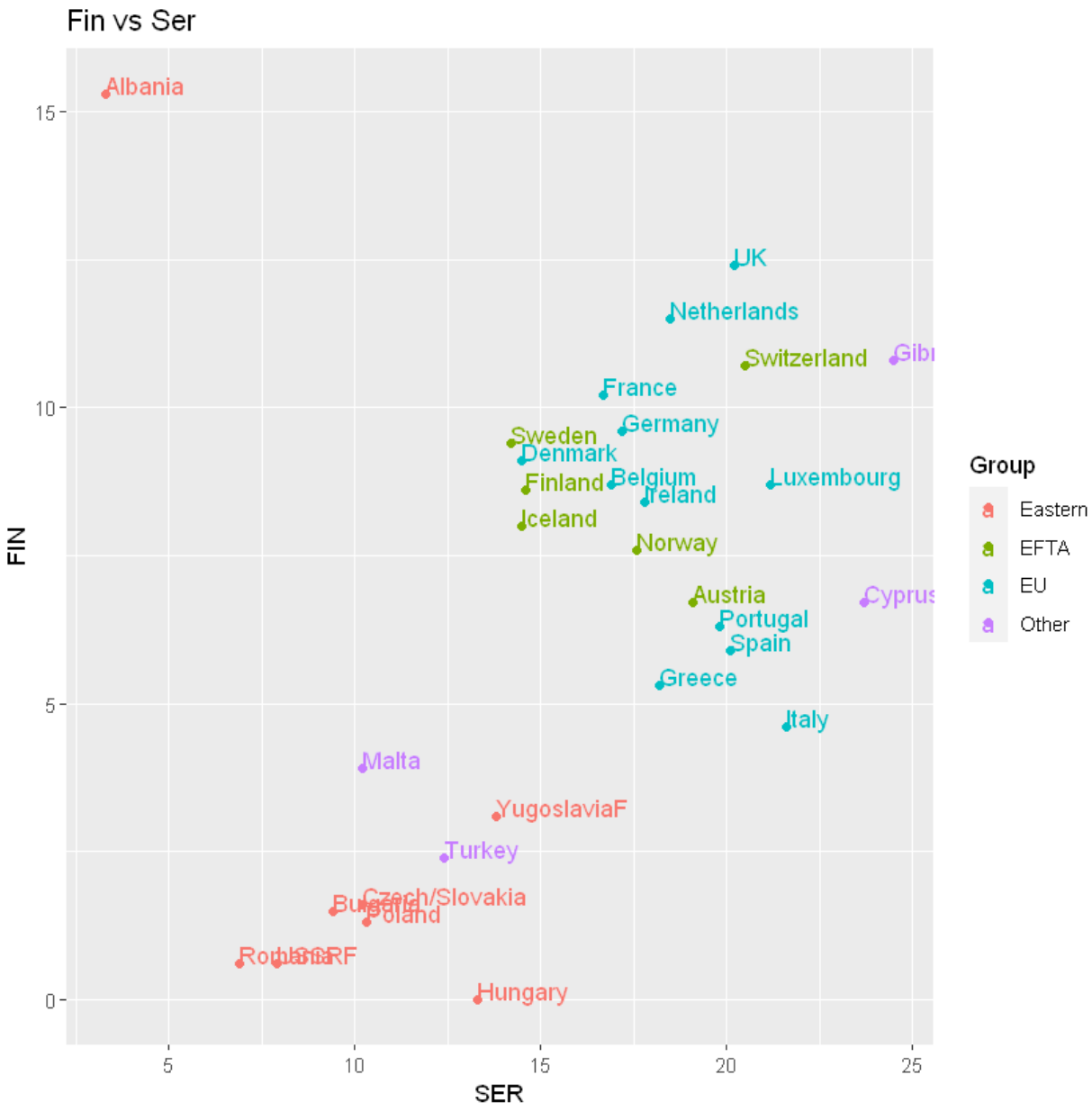
Since we are interested in applying cluster analysis to this data, we can use the pairs plot to scan the individual 2-dimensional views of the data. In cluster analysis we typically focus on 2D and 3D representations of the data in order to avoid the curse of dimensionality. With multivariate data as the dimension grows the distance between the observations grow, and it is difficult for the observations to be 'close' to one another, and hence be grouped into a small number of clusters.

Do you see any interesting 2D views of the data? What would be 'interesting'? Remember, we are interested in applying cluster analysis so 2D plots that show clusters are the plots that would be interesting. Why don't we consider MAN versus SER and SER versus FIN? Do these 2D views look interesting?

What is interesting about these plots is that you can see clusters forming, I see three distinct ones forming in the above plot.

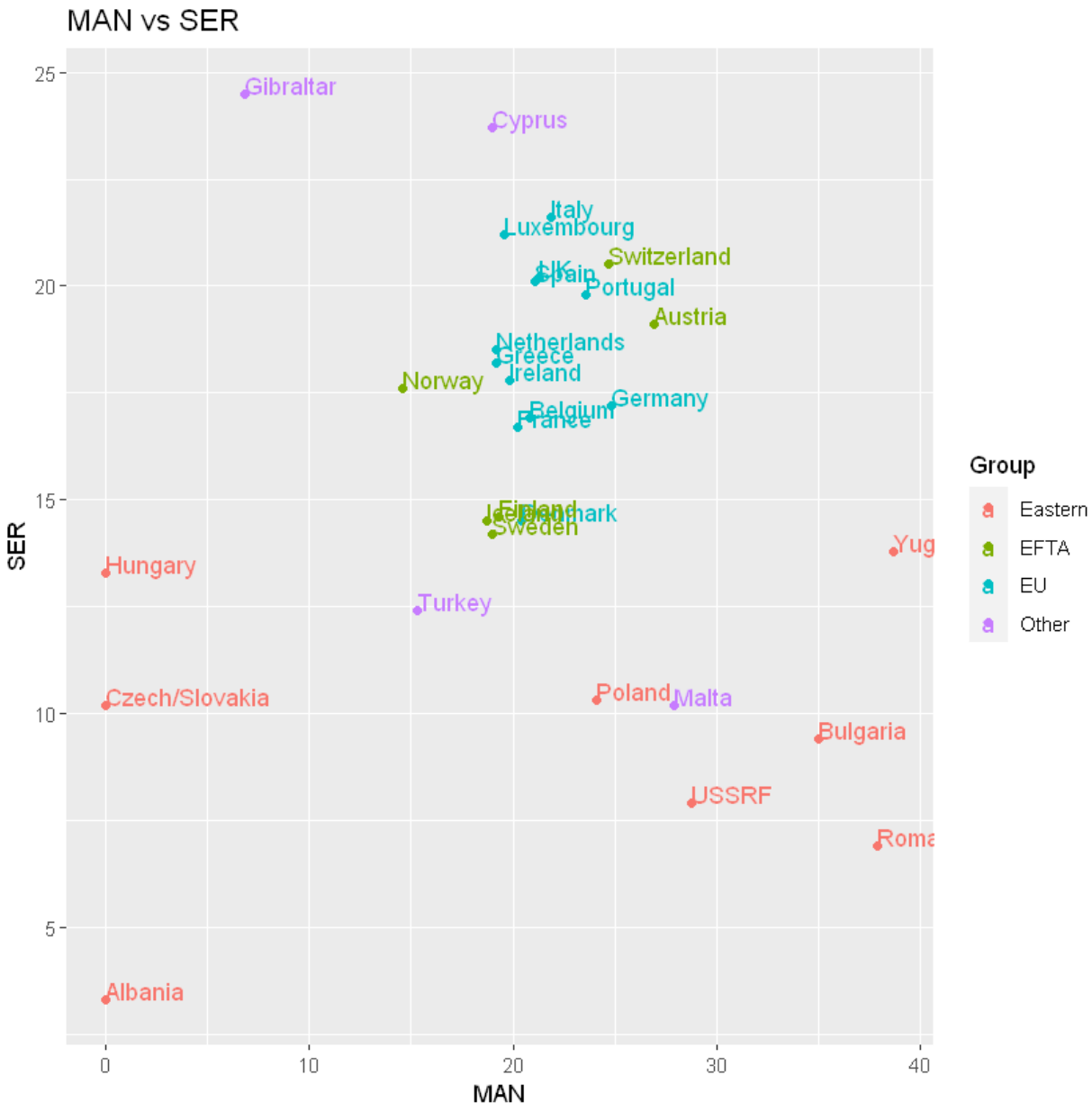
(2) Visualizing the Data with Labelled Scatterplots: While the pairs plot allows us to scan all pairwise scatterplots easily and efficiently, it is not the ideal visualization of the data. After we have homed in on some interesting dimensions we can create more specialized plots for those dimensions. Specialized plots should always include labels and color. The objective is to compress more than two dimensions of information into a two dimensional plot.

a) Plot FIN versus SER. Do we see some clusters in this plot? How many clusters do we have? How many clusters would you have if you were creating a segmentation?



I am seeing 2 distinctive clusters, specifically between the Eastern/Other and the EU/EFTA members.

b) Plot MAN versus SER. Do we see some clusters in this plot? How many clusters do we have? Are they the same clusters as we saw in the previous plot? How many clusters would you have if you were creating a segmentation?



With this analysis I am seeing 3 distinctive clusters, one large one in the upper middle area and two in the lower left and lower right areas of the graph. From this visualization, it appears that the represented data does not have any clustering tendency based on group.

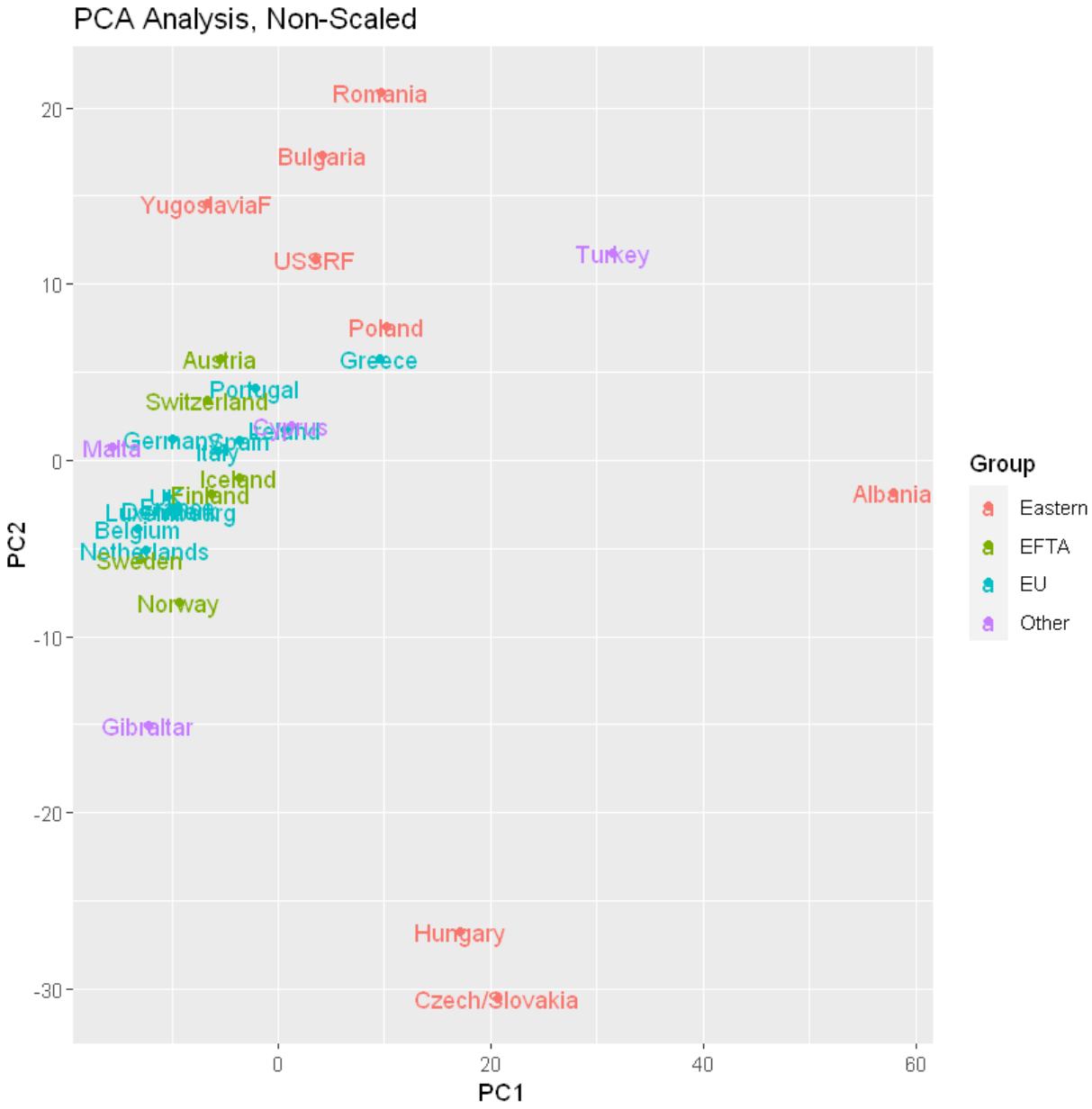
c) Of the two 2D views of the data which one do you think would be the better view for supervised clustering, i.e. using a clustering algorithm to create a classifier that will assign the countries to the correct class/label? Why?

I believe that the first graph would have a better view for supervised clustering as it has two discernible clusters whereas the second graph has one large cluster with two much weaker ones.

(3) Creating a 2D Projection Using Principal Components Analysis: We can use principal components analysis to reduce the dimension of the data. We can project the data down from 9D to

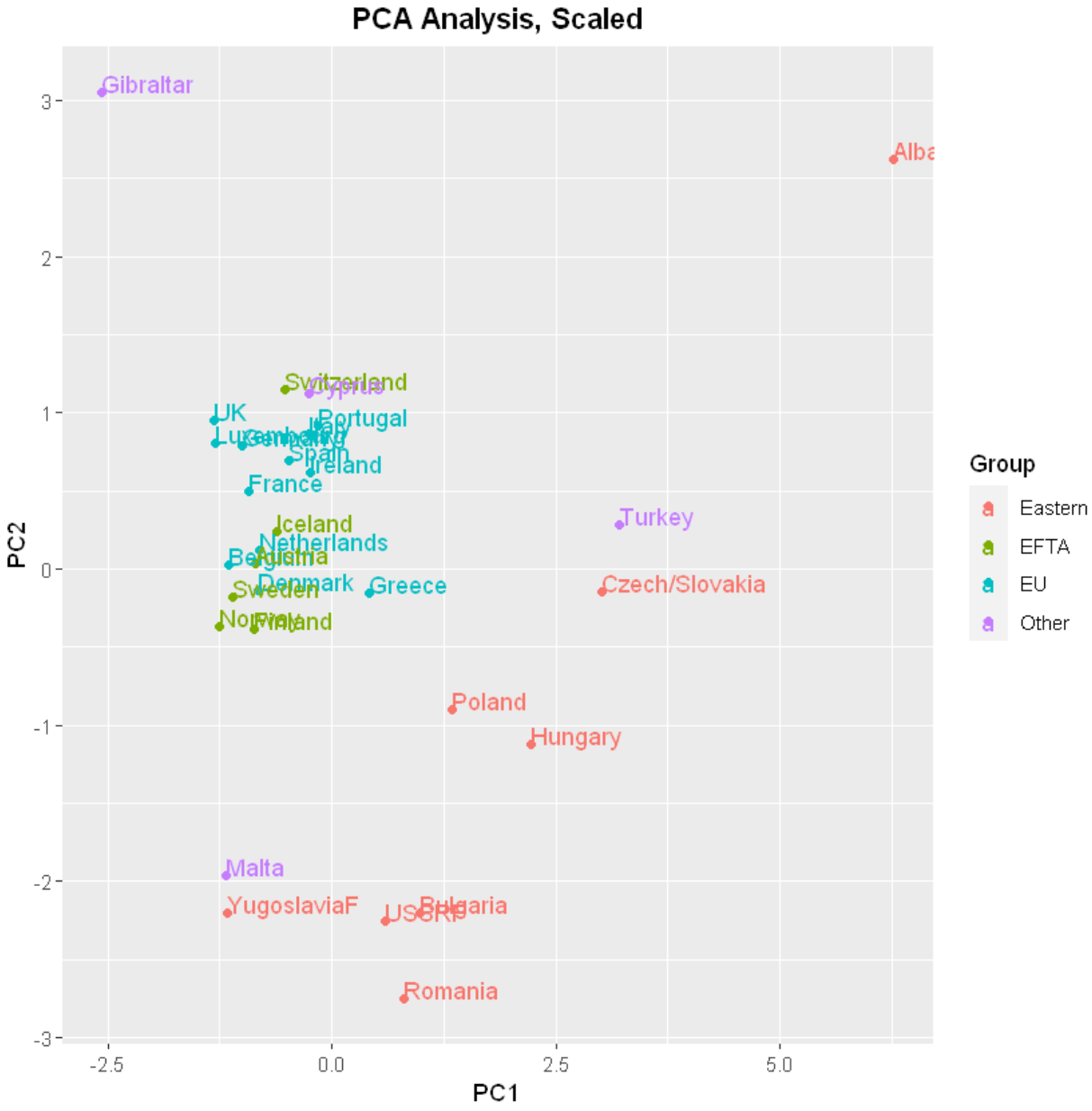
2D by performing PCA and using the first and second principal components. By doing so we are creating a new 2D view of the data, and a view of the data that contains information from more than two dimensions.

a) Use the raw data and conduct a PCA. Plot the first two principal components. How does this 2D projection of the data compare to the two other views of the data that we are considering? How many clusters does this 2D projection have? Clearly, our data can have different degrees of separation in different 2D profiles, and hence some low dimension representations will be better clustered than others.



This shows a number of clusters forming but the outliers are indicative of a lack of scaling due to the distance between the outliers especially with PC1 on the x axis.

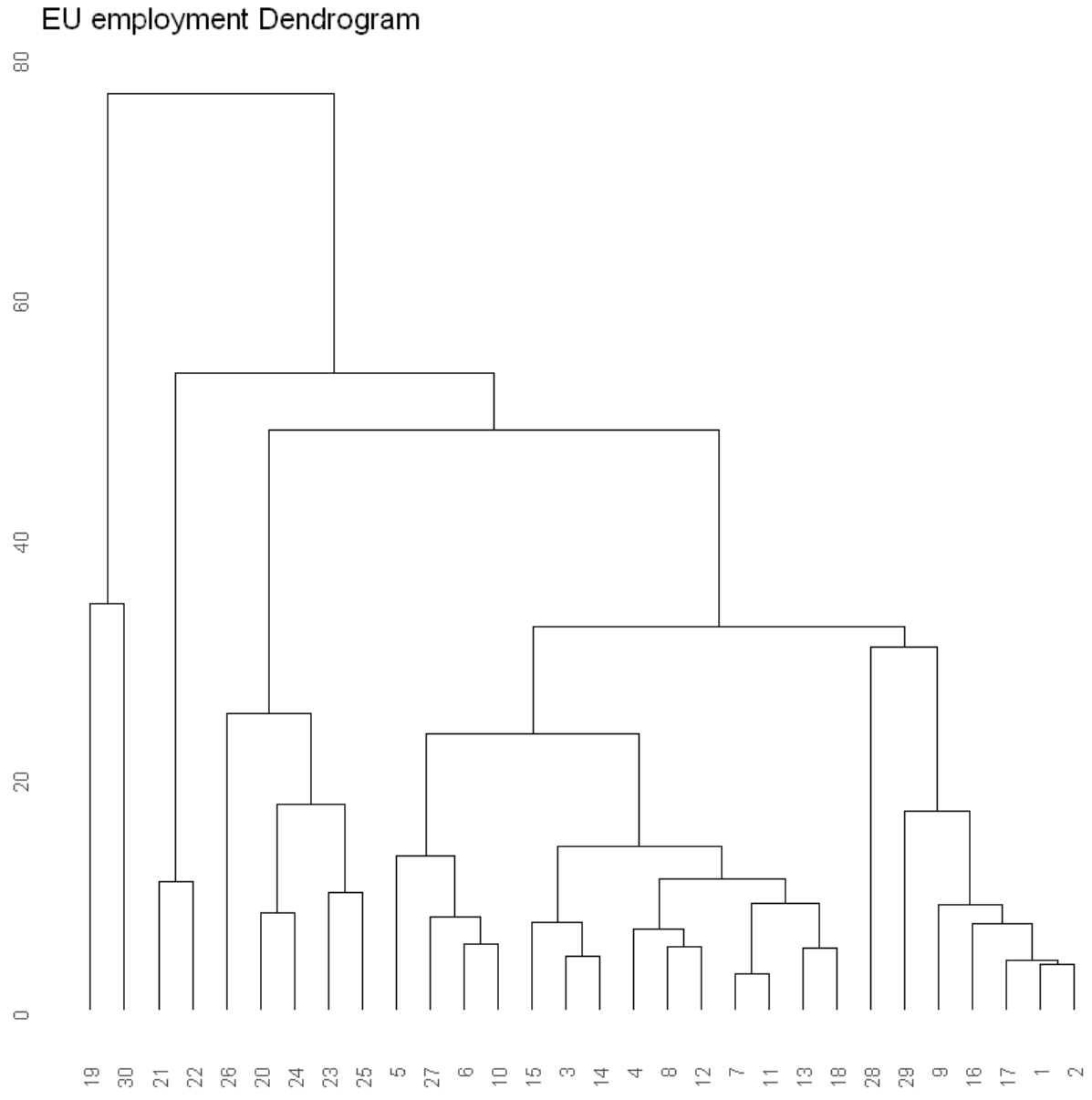
b) Usually, one is supposed to standardize the data to be mean zero with unit variance before performing a PCA. Standardize the data and run a second PCA on the standardized data. Compare the two results. Does standardizing have much of an effect here?

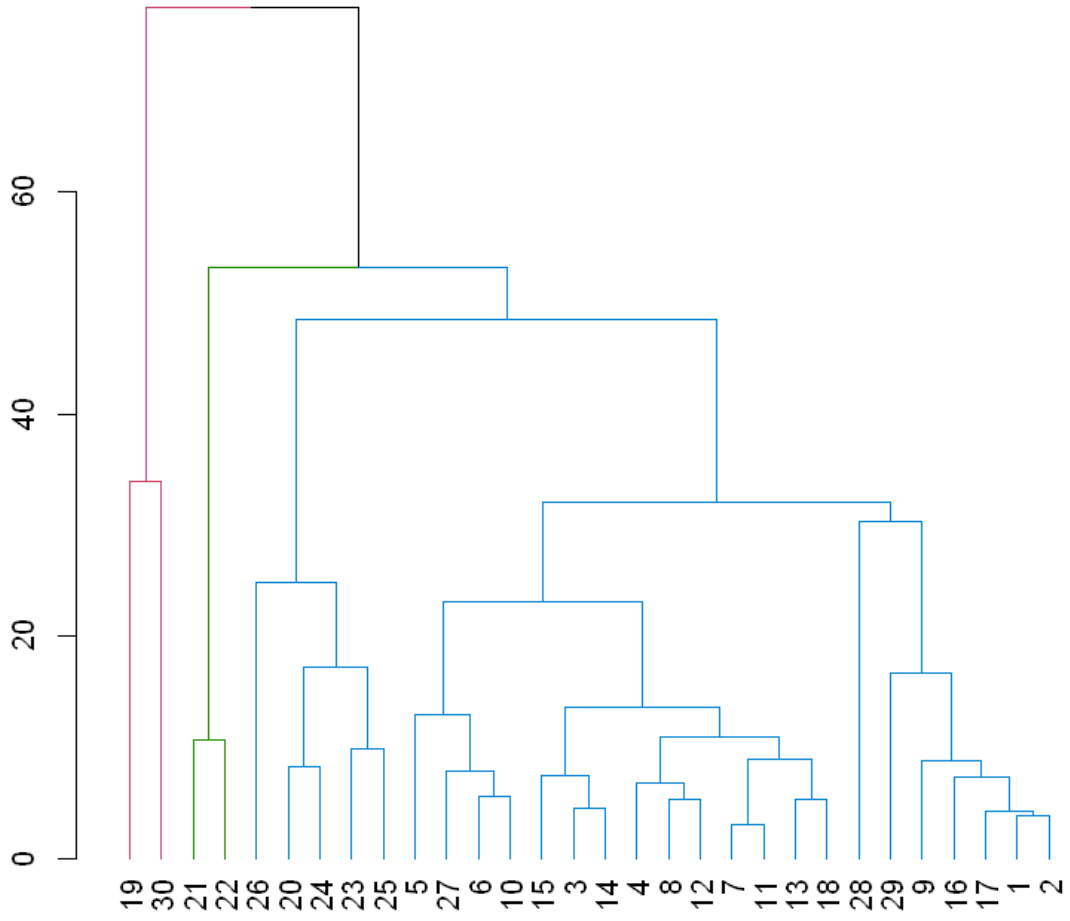


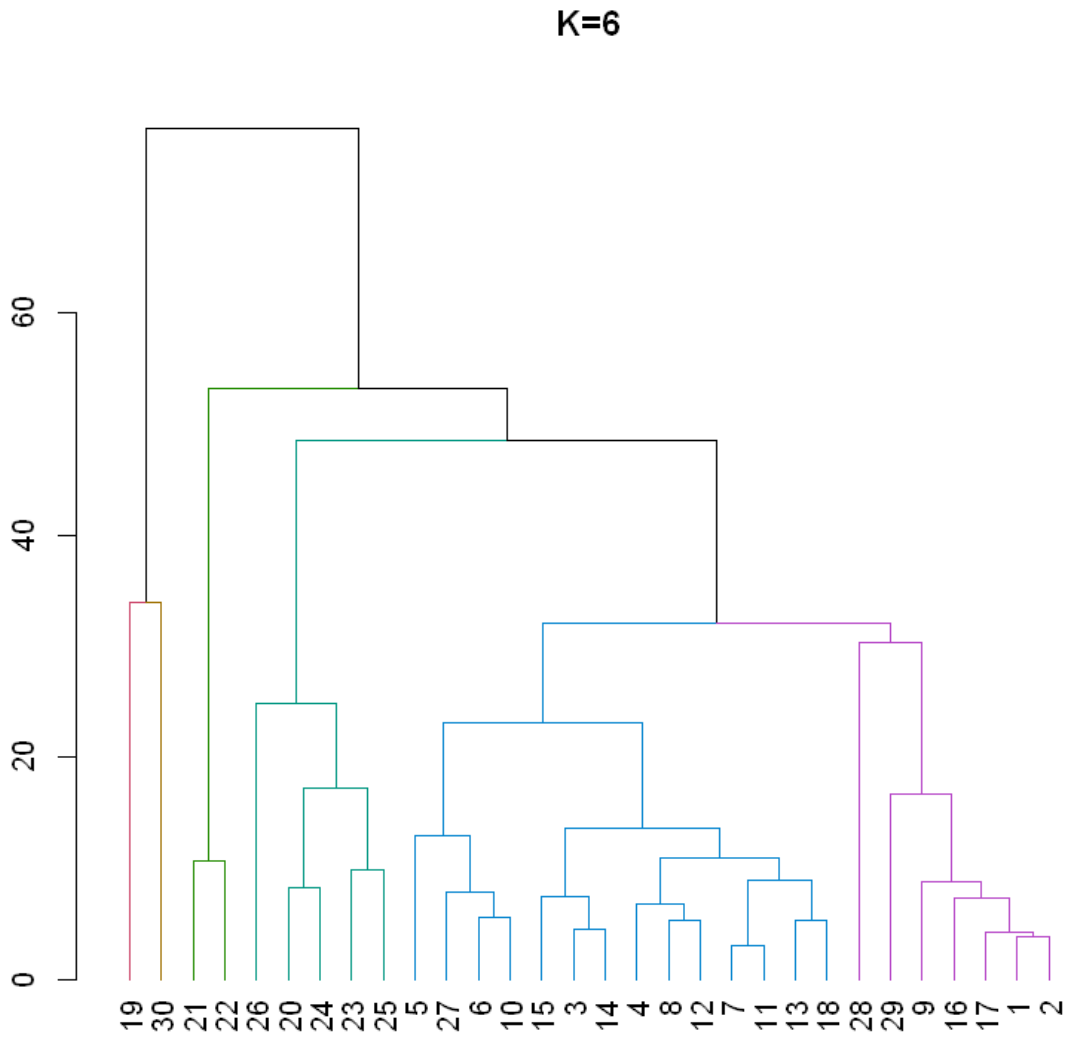
This graph demonstrates the necessity for scaling as the graph is much more balanced on the axes, especially in the case of PC1. There are definite clusters beginning to form with a few outliers.

(4) Hierarchical Clustering Analysis: 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) Perform a hierarchical cluster analysis and obtain a dendrogram. Use the cutree() function to force an assignment of the observations to a particular number of clusters. Use k=3 and k=6 and compare the classification accuracy of two cluster tree cuts. Which set of clusters is more accurate?

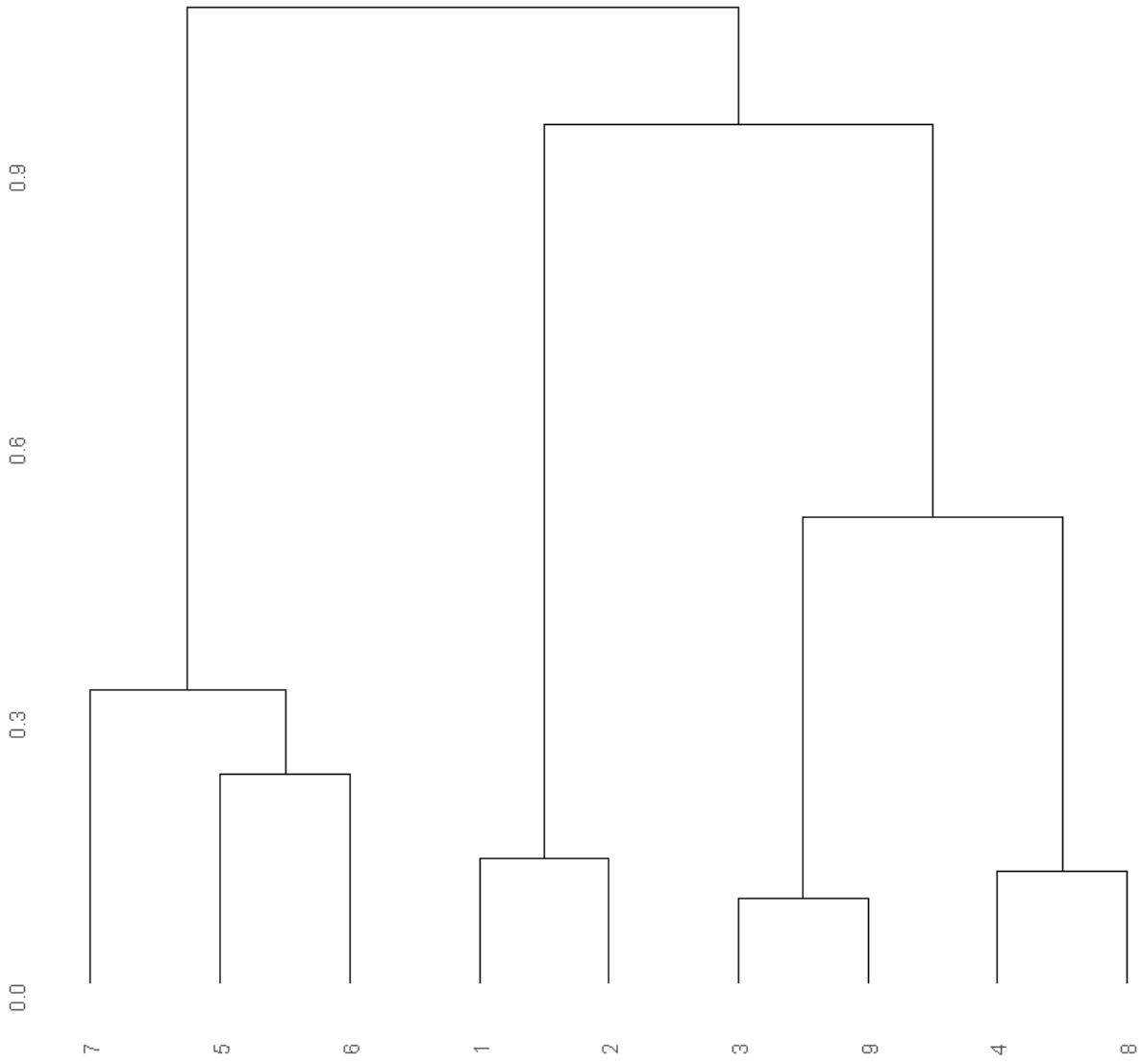


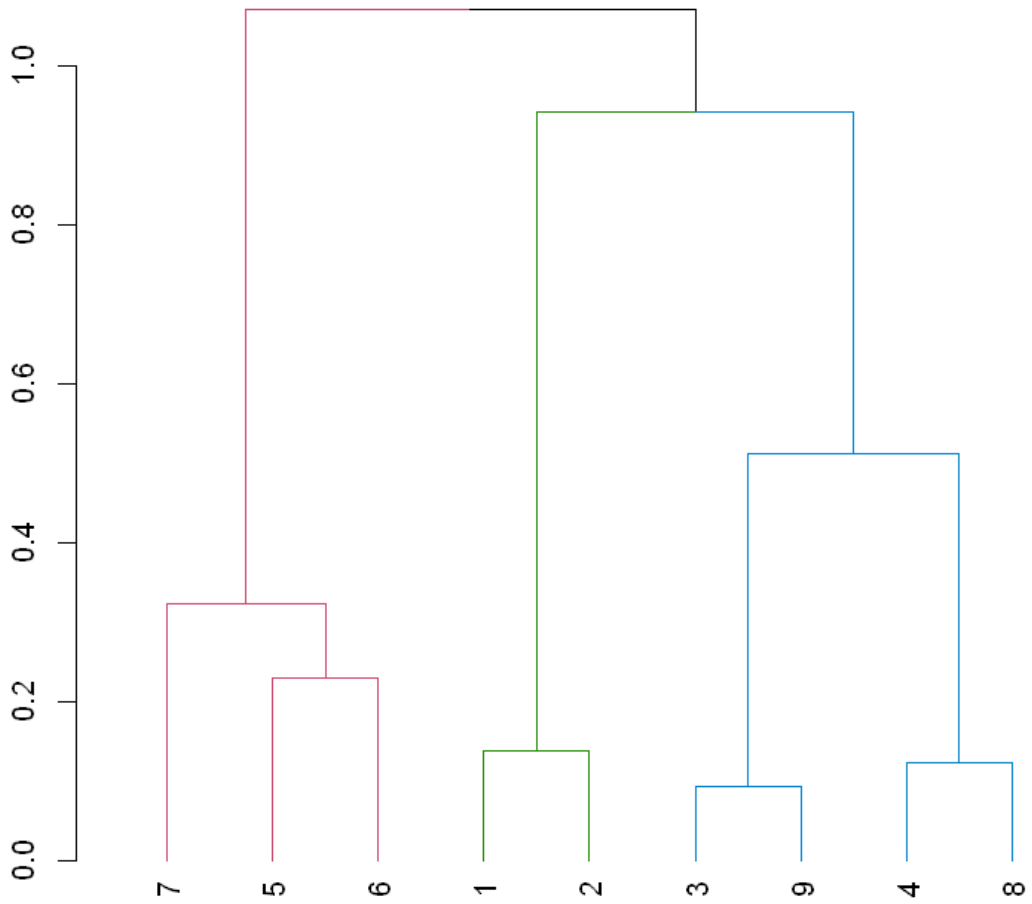

K=3
0
20
40
60
19
30
21
22
26
20
24
23
25
5
27
6
10
15
3
14
4
8
12
7
11
13
18
28
29
9
16
17
1
2

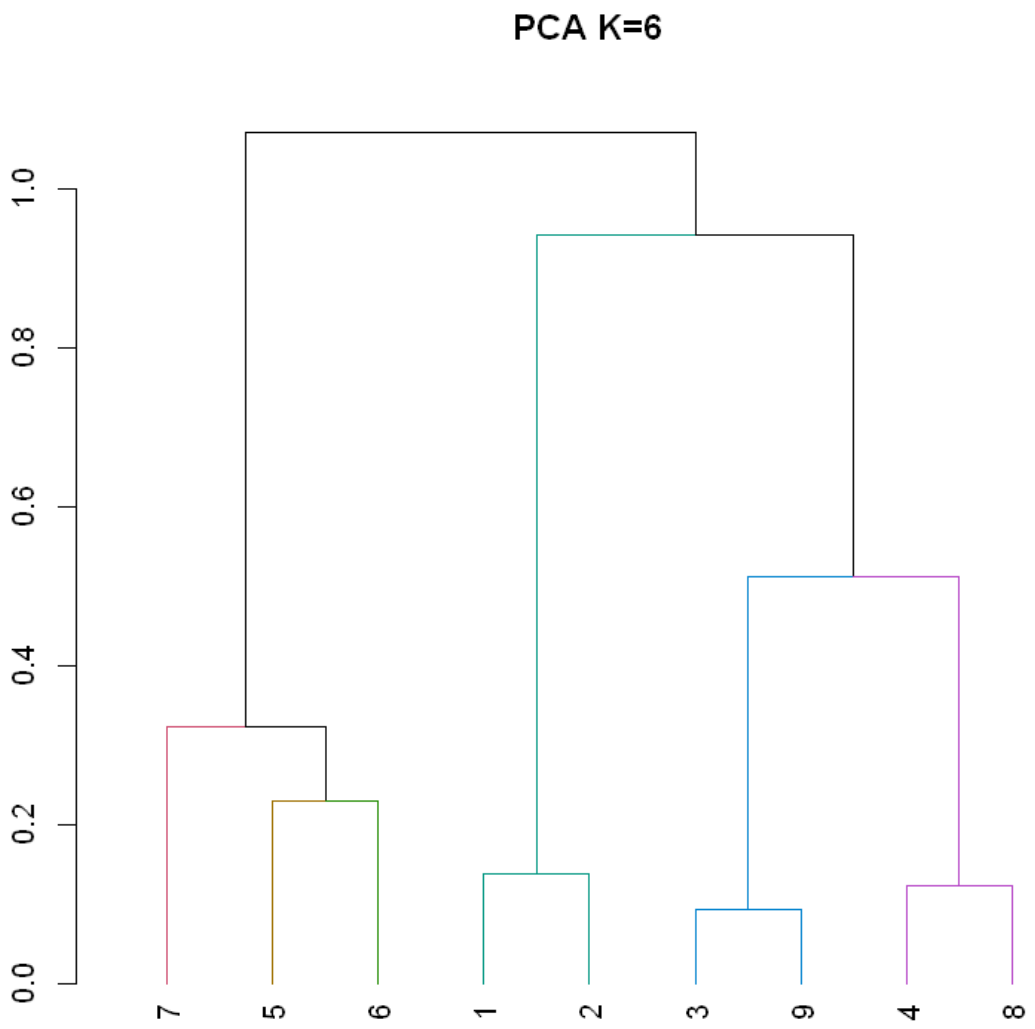


b) Perform the same analysis, but this time use the principal component space using the first and second principal components. Of these four 'cluster models' which one is the most accurate? Make a table to display their accuracy for easy comparison.

EE Employ - PCA




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



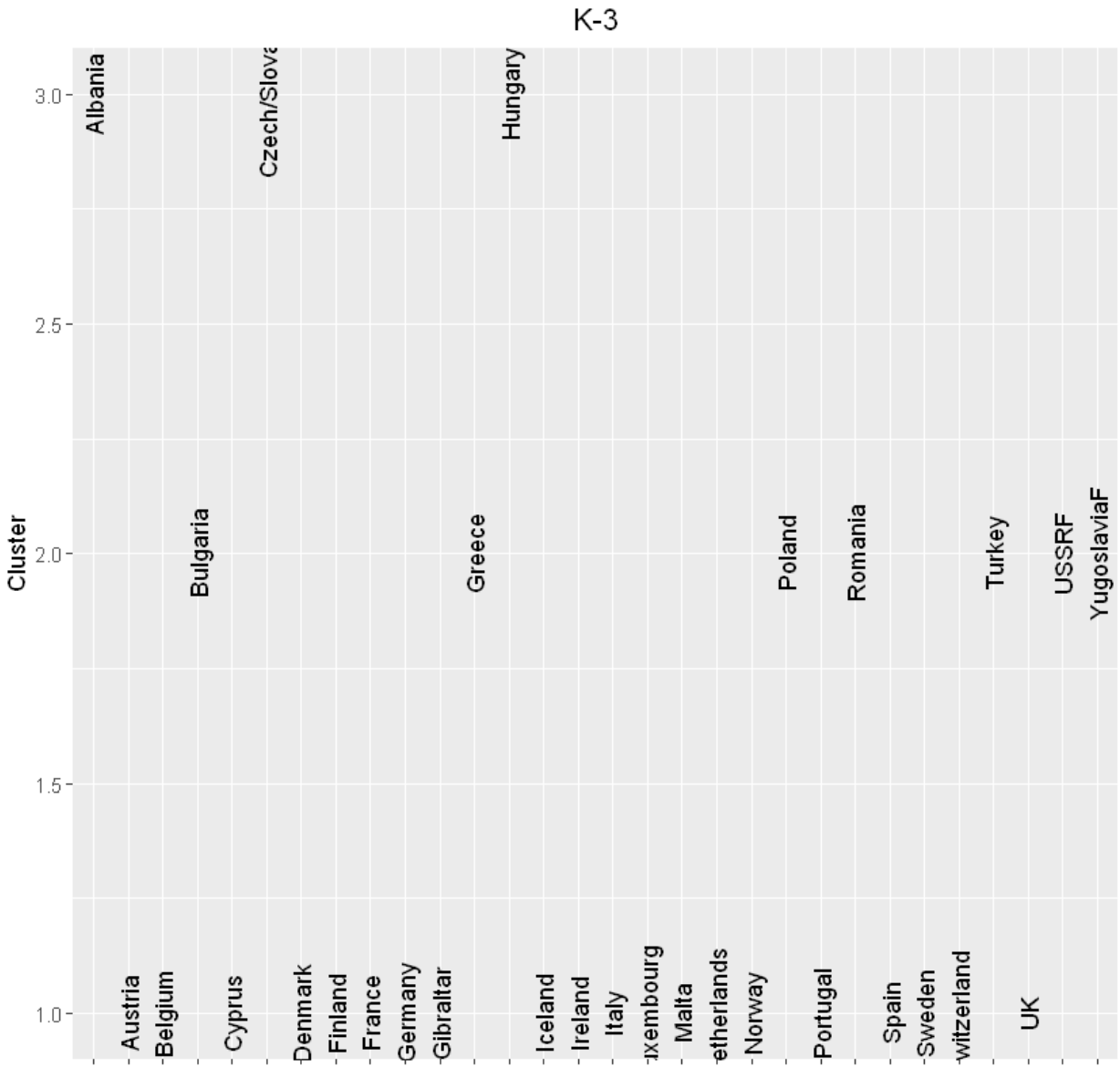
The accuracy values for the K3 analysis were 0.5893 and 0.8421 for normal and PCA, respectively. The accuracy values for the K6 analysis 0.8522 and 0.9884 for normal and PCA, respectively. This means that k6 done with the PCA analysis is the best method.

(5) k-Means Clustering Analysis: Let's perform the analogous cluster analysis and make a comparison.

a) Conduct a K-Means Cluster Analysis on the European Employment data for k=3 and k=6. Compare the classification accuracy of these models with the hierarchical models obtained in task (4).

Method	Pct
<chr>	<dbl>
k=3	0.5893374
k=6	0.8421061
PCA k=3	0.5893374
PCA k=6	0.8421061
K Means k=3	0.5792964
K Means k=6	0.8341866

b) For the k-Means Cluster Models obtain a plot that includes the original labels, their assigned clusters, and the cluster centers. What do you see in these two graphics?



Cluster Centers:

S

```
k3.results$centers
```

A matrix: 3 × 9 of type dbl

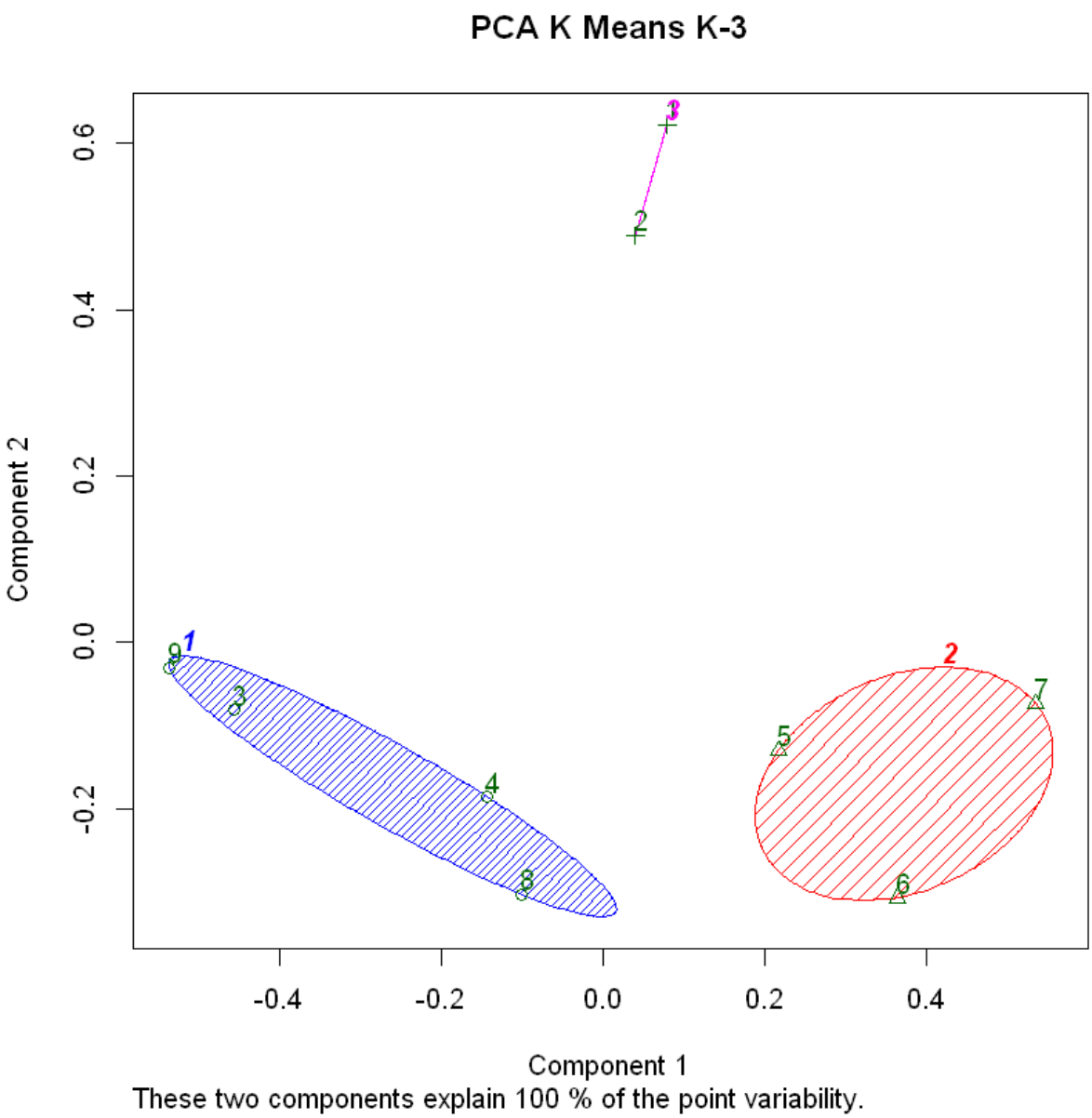
	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
1	6.34500	0.385000	20.48000	0.885	7.930000	18.170000	8.390000	30.99500	6.395000
2	22.15714	1.442857	28.42857	0.900	7.014286	11.271429	2.114286	19.95714	6.714286
3	27.86667	28.533333	0.00000	0.000	6.066667	8.933333	5.633333	16.73333	6.233333

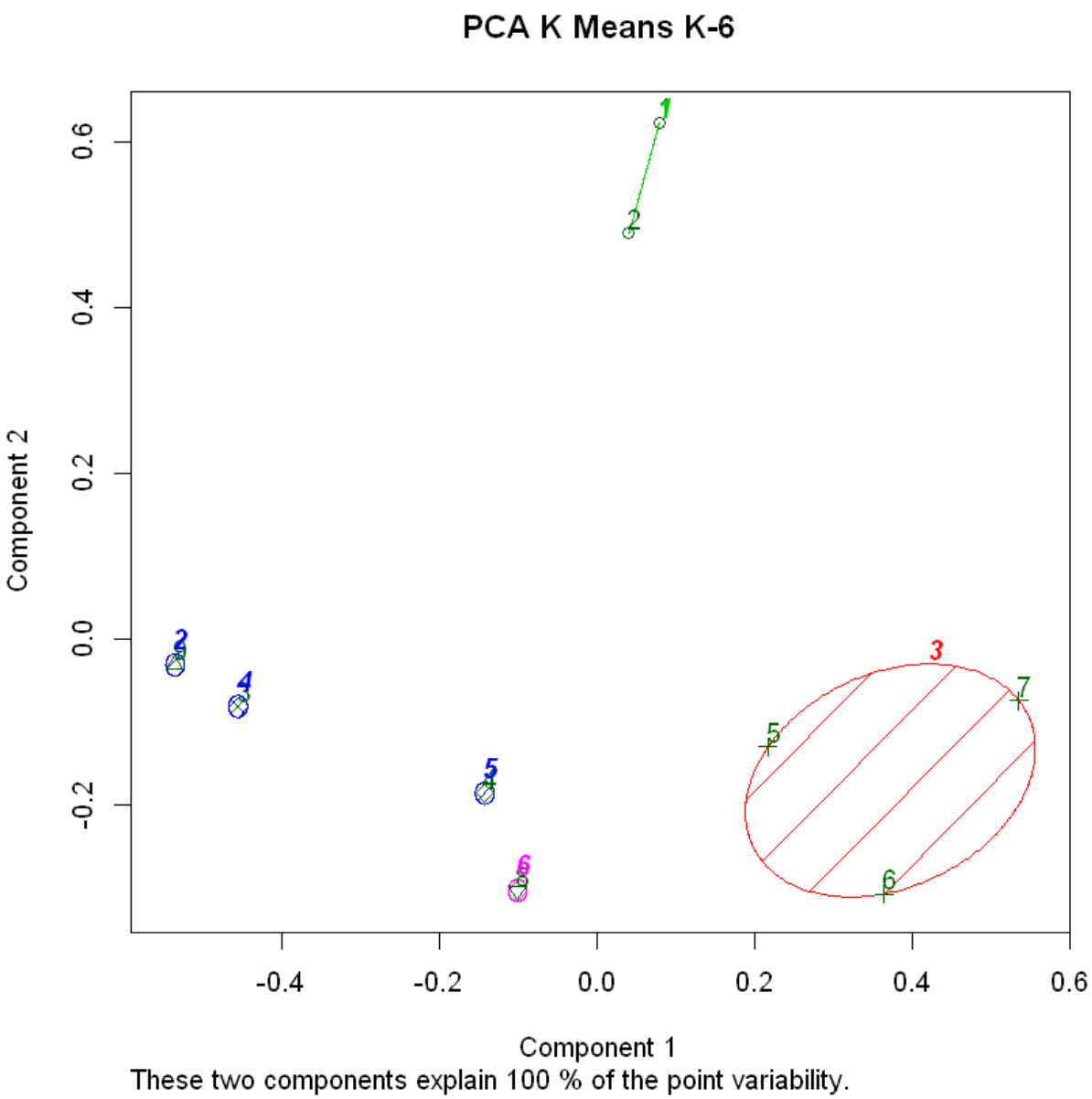
K-6
Cluster
2
4
6
Albania
Austria
Belgium
Bulgaria
Cyprus
Czech/Slova
Denmark
Finland
France
Germany
Gibraltar
Greece
Hungary
Iceland
Ireland
Italy
Luxembourg
Malta
Netherlands
Norway
Poland
Portugal
Romania
Spain
Sweden
Switzerland
Turkey
UK
USSRF
YugoslaviaF

A matrix: 6 × 9 of type dbl

	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
1	50.150000	10.1500000	7.65000	0.1000000	4.300000	7.85000	8.850000	7.25000	3.700000
2	5.000000	2.2000000	38.70000	2.2000000	8.100000	13.80000	3.100000	19.10000	7.800000
3	8.118182	0.4363636	21.94545	0.7272727	8.818182	19.60909	8.000000	26.31818	5.990909
4	4.177778	0.3222222	18.68889	1.0777778	6.844444	16.41111	8.866667	36.71111	6.888889
5	21.060000	1.4000000	29.00000	0.7800000	7.160000	10.54000	1.860000	21.22000	6.960000
6	14.050000	33.1000000	0.00000	0.0000000	7.400000	11.75000	0.800000	25.10000	7.850000

c) Conduct a K-Means Cluster Analysis for k=3 and k=6, but use the Principal Components space.


PCA K Means K-3
Component 2
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.2
0.0
0.2
0.4
Component 1
These two components explain 100 % of the point variability.



These two components explain 100 % of the point variability.

d) What happens as we increase the number of clusters from k=3 to k=6?

It appears to go from having two larger clusters to one large one with several smaller clusters.

e) Of these eight cluster models which is the most accurate? Make a table summarizing the eight models and their accuracy.

A data.frame: 8 × 2

Method	Pct
<fct>	<dbl>
K=3	0.5893374
K=6	0.8421061
PCA k=3	0.5893374
PCA k=6	0.8421106
K Means K=3	0.5792964
K Means K=6	0.8341866
PCA K Means K=3	0.8521550
PCA K Means K=6	0.9520450

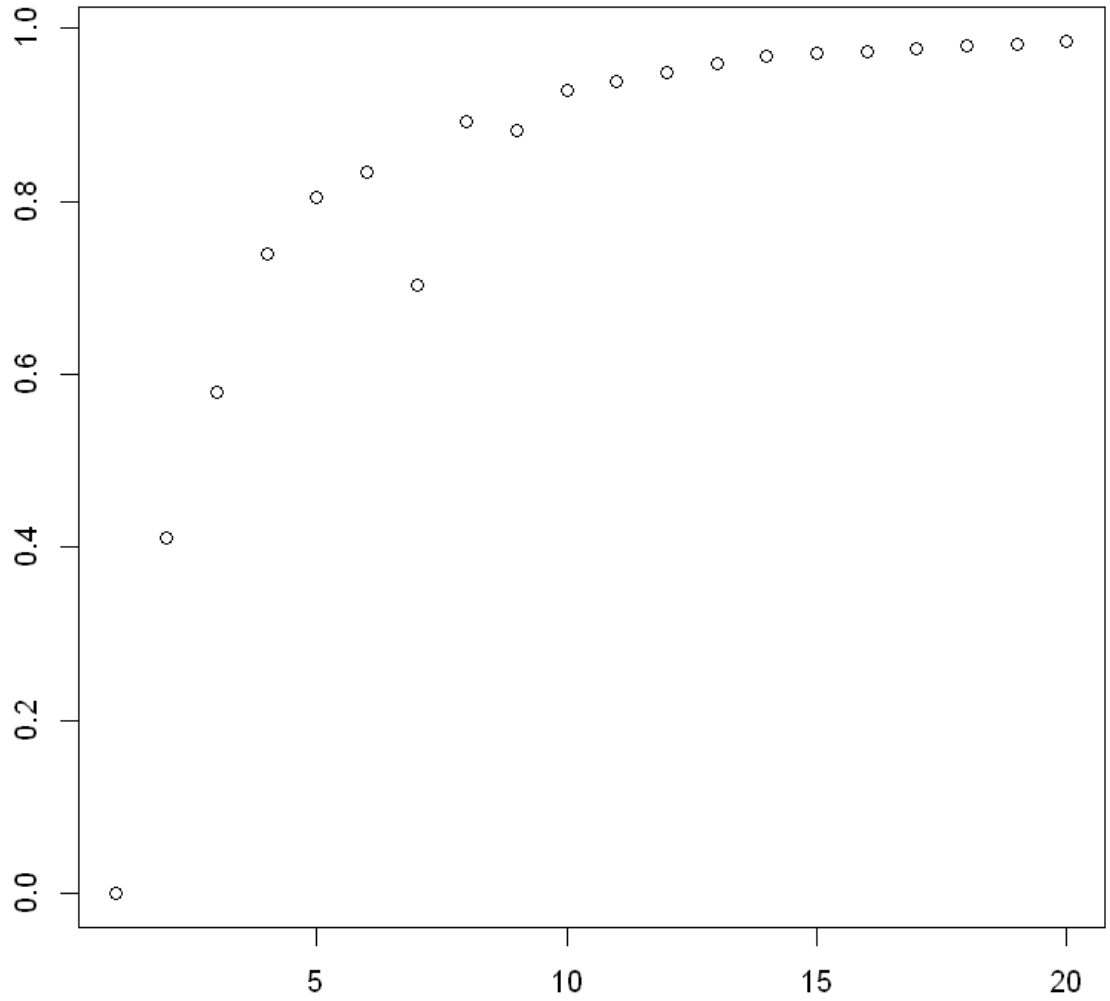
---

f) How do the clusters compare with the original labels (EU, EFTA, Eastern, or Other)?

It's clear that there is a correlation, especially the K=6 method.

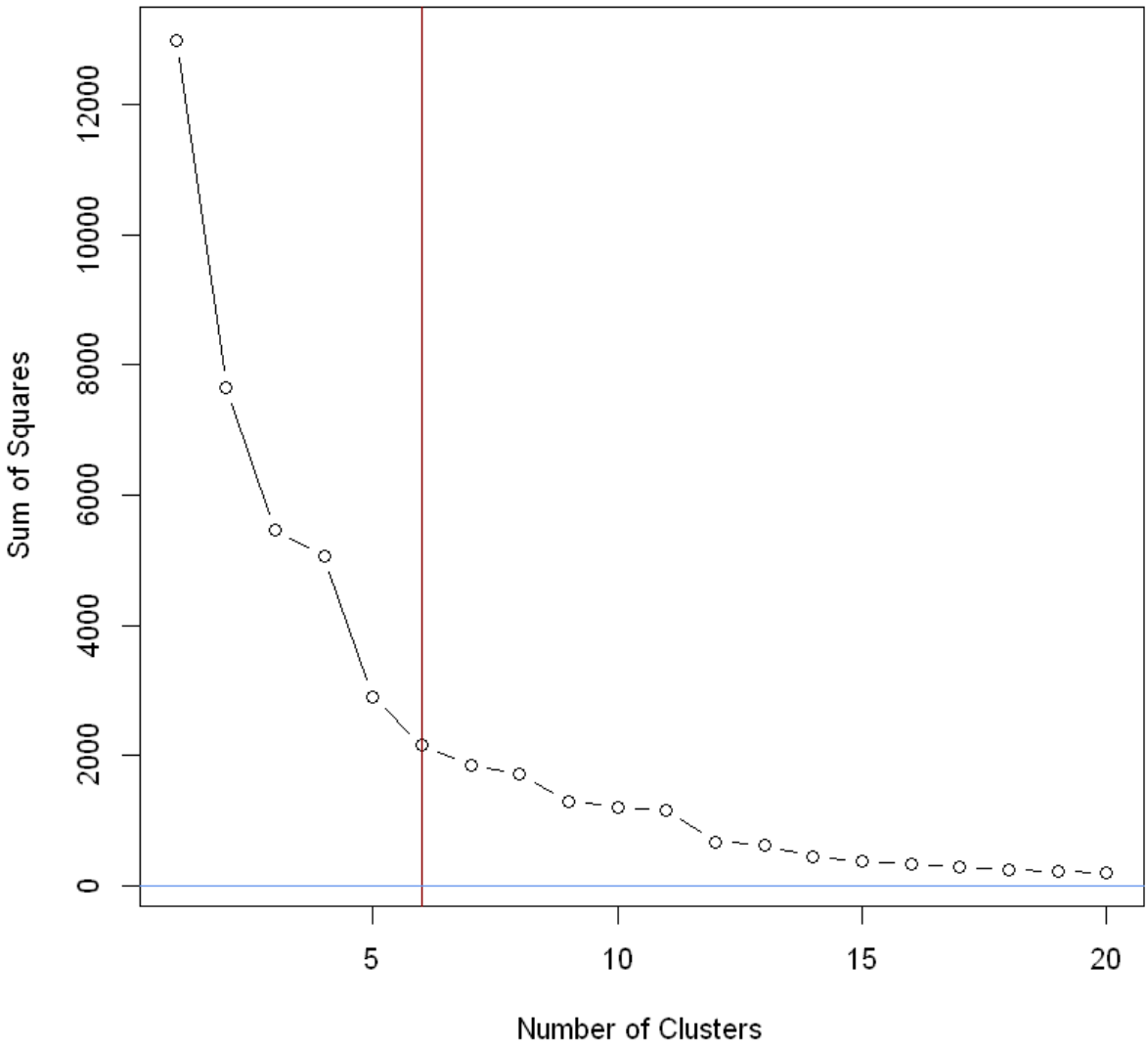
(6) Computing the 'Optimal' Number of Clusters by Brute Force: After completing the above cluster analyses, one should question whether or not k=3 or k=6 is the "best" choice for the number of clusters to retain. Unfortunately, the answer to that question is not as simple as the question. One idea that should be apparent is that we would need to be able to evaluate a large number of clusters bases on some criterion that allows an objective comparison. One option is to use the classification accuracy rate of our clusters.

a) 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?



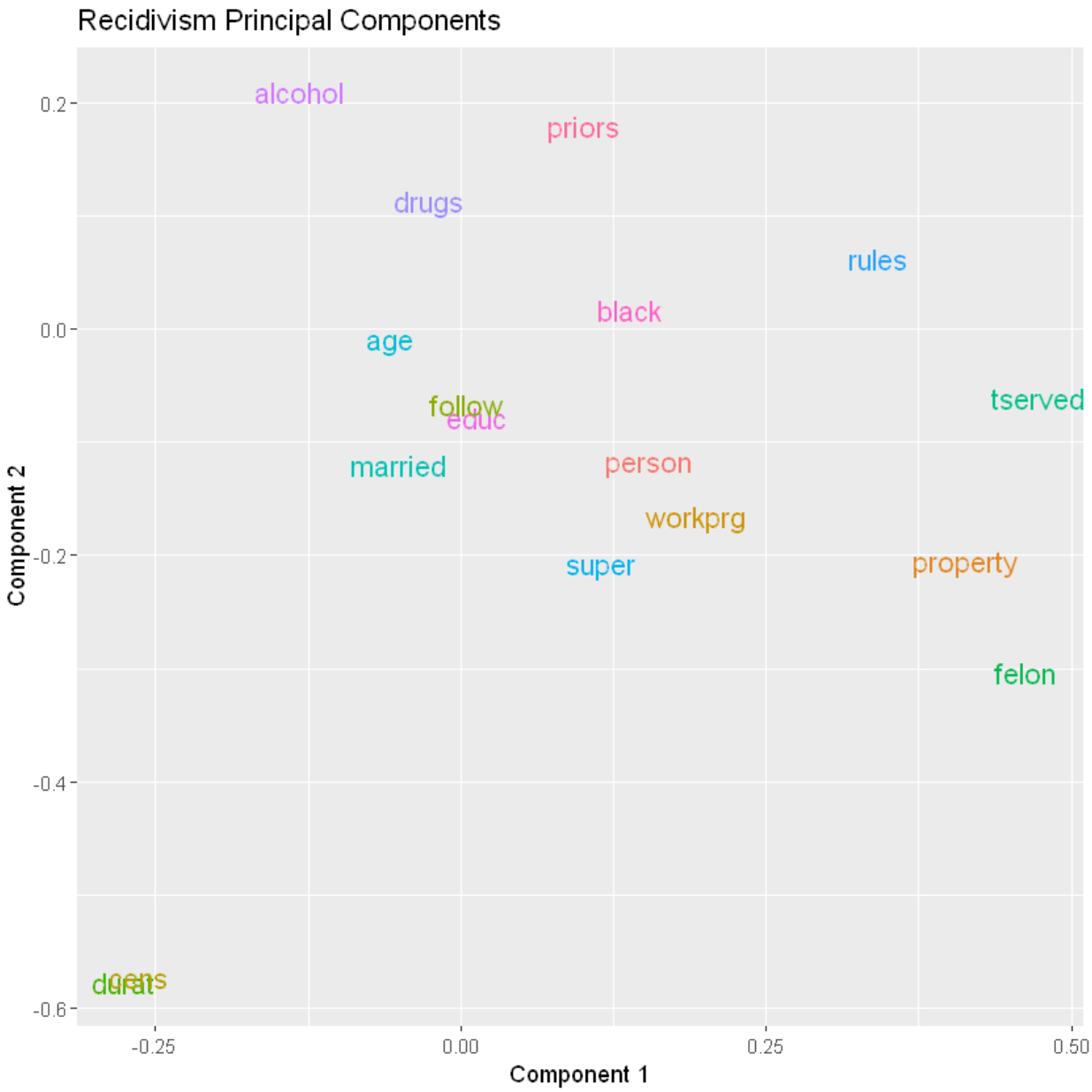
It is clear that at around 15 clusters the classification accuracy begins to level out.

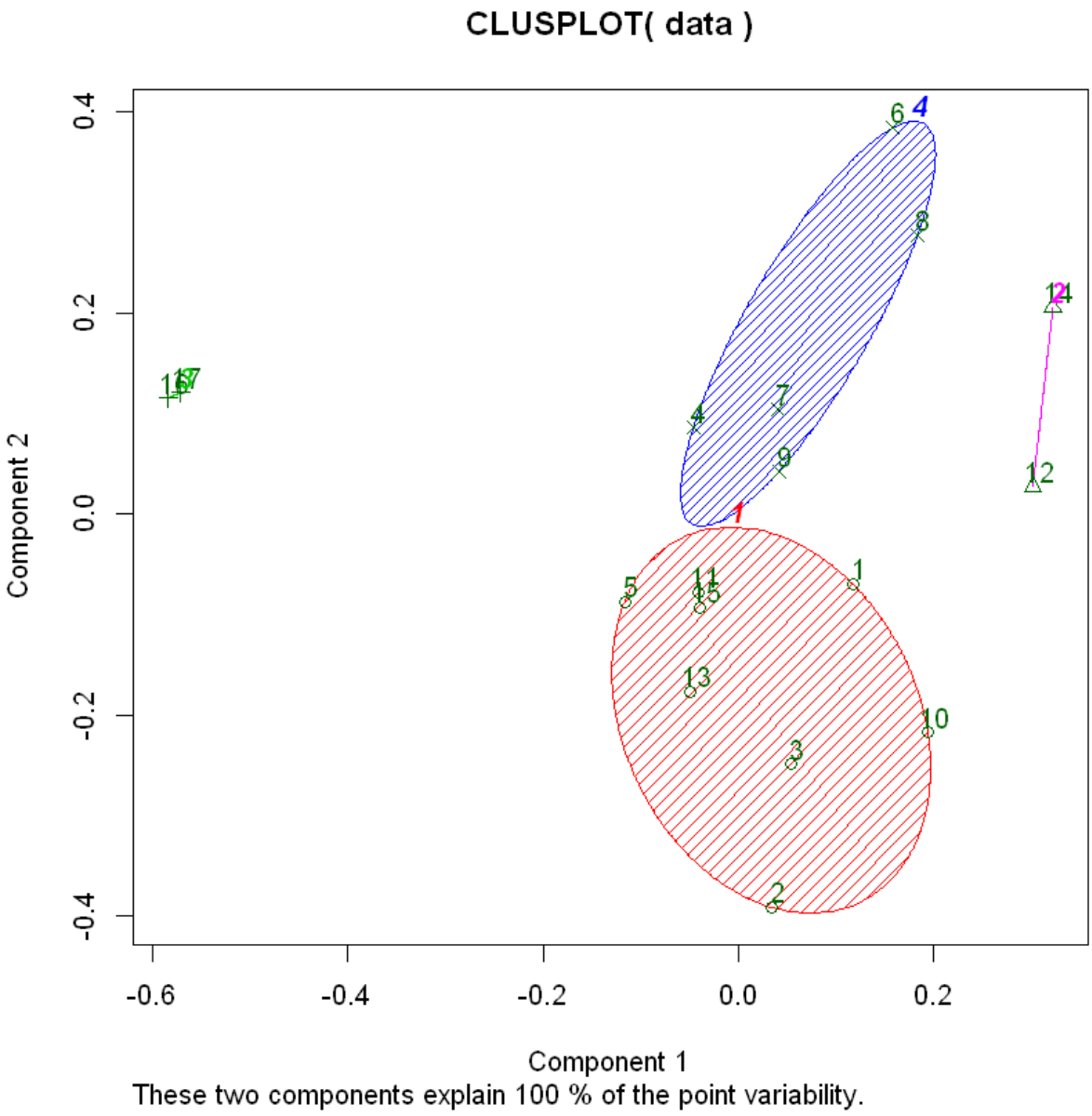
(7) On Your Own Modeling 1: The USSTATES dataset is a 12 variable dataset with n=50 records that you used briefly in MSDS 410. The data, calculated from census data, consists of state-wide average or proportion scores for the non-demographic variables. As such, higher scores for the composite variables translate into having more of that quality. There is no other information available about this data. Use this data set and conduct a hierarchical cluster analysis. Decide on the total number of clusters to retain and describe the differences amongst the clusters.



For this group I decided to use 14 clusters as it was where the sum of squares metric leveled out.

(8) On Your Own Modeling 2: The RECIDIVISM dataset is an 18 variable dataset with n=1445 records. Please see the data description file for the variable definitions and additional information about the dataset. The data consists of a random sample records on convicts released from prison during 1977/1978. Use this data set and conduct a kmeans cluster analysis. Decide on the total number of clusters to retain and describe the differences amongst the clusters.


Recidivism Principal Components
alcohol
priors
drugs
rules
black
age
tserved
follow
educ
married
person
workprg
super
property
felon
durat
ccens
0.2
0.0
-0.2
-0.4
-0.6
-0.25
0.00
0.25
0.50
Component 2
Component 1



I was able generate a clustering analysis with 4 clusters and I got an accuracy of 83.685% which is the best I could achieve after testing multiple different cluster amounts. I input this after a PCA analysis.

(9) Please write a reflection on your cluster modeling experiences.

Forgive my directness but this was an extremely difficult and a ruthlessly long assignment, and to be honest I found it to be a lot of metaphorical hoops jumping. I understand that this is a good program but completing this many pages of work for one assignment, when combined with another class and 60+ hour work weeks has completely torpedoed my life over the past week and a half, every

day I would come home and work on this. Every question got worse and involved more graphical coding. Why did we need to do 3 datasets? Why did I need to graph every single question? I showed this assignment to people in my department who do this stuff for a living and they though it was mega overkill. Once again, I do appreciate the work and it is not a criticism, but I am just reporting how this assignment affected me.

Anyway, I otherwise though that most of the assignment was informative and gave a comprehensive methodology on how to work with smaller data sets which is something that I have very little experience doing. It also drastically increased my skills with R which, prior to this class, had been lacking. The final two modeling exercises we also good to review the concepts that we had learned, the datasets were similarly small and very fit for a clustering analysis.