Assignment #8: Multivariate Analysis (30 points)

Data Directory: Data can be accessed on the SAS OnDemand server using this libname statement.

```
libname mydata        '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/' access=readonly;
```

Data Set: mydata.european_employment

Data Description: Employment in various industry segments reported as a percent for thirty European nations. See the data dictionary for full details. Note that EU stands for European Union, EFTA stands for European Free Trade Association, and Eastern stand for Eastern European nations or the former Eastern Block.

Assignment Instructions: Note that this assignment will not use our assignment template, nor will it follow the guidelines for report writing that we have used all quarter. Instead, you will be able to paste your output and type your answers directly into the Word version of this assignment, convert your solution document to a pdf, and submit your pdf document into Blackboard. **Please color code your answers in green.**

In this assignment we will take a guided tour of the multivariate analysis capabilities in SAS. These capabilities will include PROC PRINCOMP, PROC FACTOR, and PROC CLUSTER. Since none of these methods are covered in our SAS books, our only reference will be the SAS User's Guide.

PROC FACTOR	Chapter 34	SAS 9.3 User's Guide
PROC PRINCOMP	Chapter 72	SAS 9.3 User's Guide
PROC CLUSTER	Chapter 30	SAS 9.3 User's Guide

http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#titlepage.htm

The assignment is broken down into parts for your convenience. Each part will instruct you to generate a particular set of SAS output and interpret this output. In addition a section may have some particular questions that you should address. These questions will be written in **bold black type**.

Note that the SAS code provided in this assignment will produce an extensive amount of output. You will probably want to run the code piece by piece and answer each Part of the assignment completely before moving to the next Part.

For convenience here are the definitions of the abbreviated industries.

AGR: agriculture
MIN: mining
MAN: manufacturing
PS: power and water supply
CON: construction
SER: services
FIN: finance
SPS: social and personal services
TC: transport and communications

Part 1: An Initial Correlation Analysis

We will conclude this tutorial by applying cluster analysis to this data. When we perform a cluster analysis, we will always want to perform the cluster analysis in a low dimensional setting. Only in low dimensions can points be "close together". As we move towards this cluster analysis we want to perform some basic examinations of the data and consider using factor analysis and principal components as means to reduce the dimensionality of our data.

Of course, before we conclude this tutorial we must begin this tutorial. We will begin this tutorial by examining the two dimensional scatterplots of the variables. Use PROC CORR to produce the Pearson correlation coefficients and the scatterplot matrix. Looking at the scatterplots, is there any scatterplot that looks like it would yield interesting cluster results? For the two variables of your choice make this scatterplot (replace Yvar and Xvar with your two variables).

```
data temp;
set mydata.european_employment;
run;

ods graphics on;
proc sgplot data=temp;
title 'Scatterplot of Raw Data';
scatter y=Yvar x=Xvar / datalabel=country group=group;
run; quit;
ods graphics off;
```

In this data set there are four counties that do not belong to any of the three primary groups. If you had to assign each of these countries to a group to which group would you assign each country.

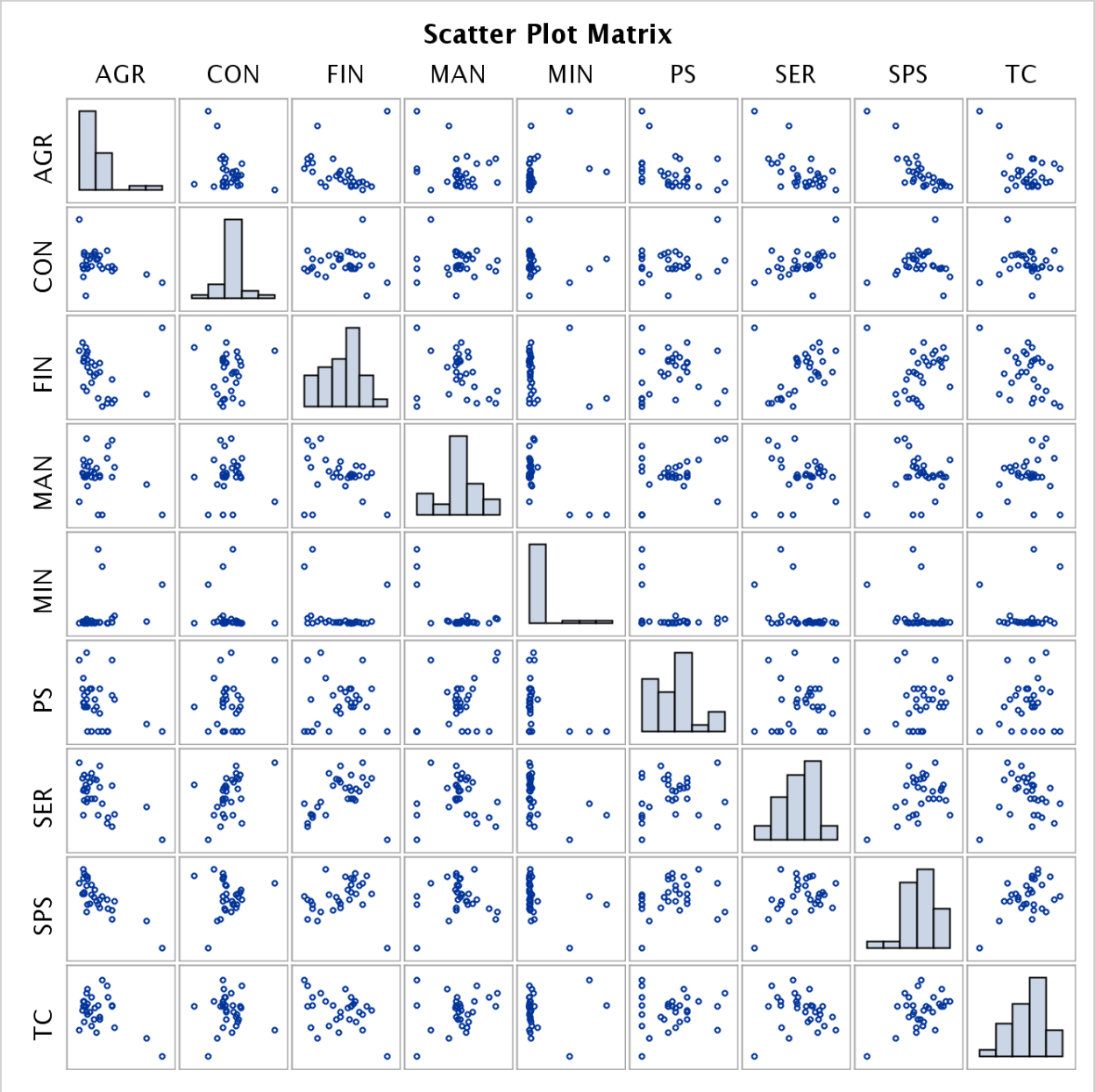
Note: In this assignment our observations are assigned to *classes* or are said to have *labels* (EU, EFTA, Eastern, or Other). Typically we use cluster analysis as an *unsupervised learner* (a situation with no response variable or label) and not as a *supervised learner* (a situation with a response variable or label). If we wanted to be able to correctly assign each country to its group affiliation, then we would define a *classification problem* (see Chapter 11 in *Applied Multivariate Data Analysis*). Throughout this assignment we will be interested in grouping countries together (creating a *segmentation*), but we can also observe their group affiliation to see if these groups have similarities.

As the first step in the initial correlation analysis, the correlation coefficients and accompanying scatterplot matrix were produced using PROC CORR in SAS for all nine predictor variables in the European employment data set. Looking at the correlation coefficients, we can see that there is evidence of collinearity between some of the predictor variables.

Pearson Correlation Coefficients, N = 30 Prob > |r| under H0: Rho=0									
	AGR	**CON**	**FIN**	**MAN**	**MIN**	**PS**	**SER**	**SPS**	**TC**
AGR	1.00000	-0.34861 0.0590	-0.17575 0.3529	-0.25439 0.1749	0.31607 0.0888	-0.38236 0.0370	-0.60471 0.0004	-0.81148 <.0001	-0.48733 0.0063
CON	-0.34861 0.0590	1.00000	-0.01802 0.9247	-0.03446 0.8565	-0.12902 0.4968	0.16480 0.3842	0.47308 0.0083	0.07201 0.7053	-0.05461 0.7744
FIN	-0.17575 0.3529	-0.01802 0.9247	1.00000	-0.27374 0.1433	-0.24806 0.1863	0.09431 0.6201	0.37928 0.0387	0.16602 0.3806	-0.39132 0.0325
MAN	-0.25439 0.1749	-0.03446 0.8565	-0.27374 0.1433	1.00000	-0.67193 <.0001	0.38789 0.0342	-0.03294 0.8628	0.05028 0.7919	0.24290 0.1959
MIN	0.31607 0.0888	-0.12902 0.4968	-0.24806 0.1863	-0.67193 <.0001	1.00000	-0.38738 0.0344	-0.40655 0.0258	-0.31642 0.0885	0.04470 0.8146
PS	-0.38236 0.0370	0.16480 0.3842	0.09431 0.6201	0.38789 0.0342	-0.38738 0.0344	1.00000	0.15498 0.4135	0.23774 0.2059	0.10537 0.5795
SER	-0.60471 0.0004	0.47308 0.0083	0.37928 0.0387	-0.03294 0.8628	-0.40655 0.0258	0.15498 0.4135	1.00000	0.38798 0.0341	-0.08489 0.6556
SPS	-0.81148 <.0001	0.07201 0.7053	0.16602 0.3806	0.05028 0.7919	-0.31642 0.0885	0.23774 0.2059	0.38798 0.0341	1.00000	0.47492 0.0080
TC	-0.48733 0.0063	-0.05461 0.7744	-0.39132 0.0325	0.24290 0.1959	0.04470 0.8146	0.10537 0.5795	-0.08489 0.6556	0.47492 0.0080	1.00000

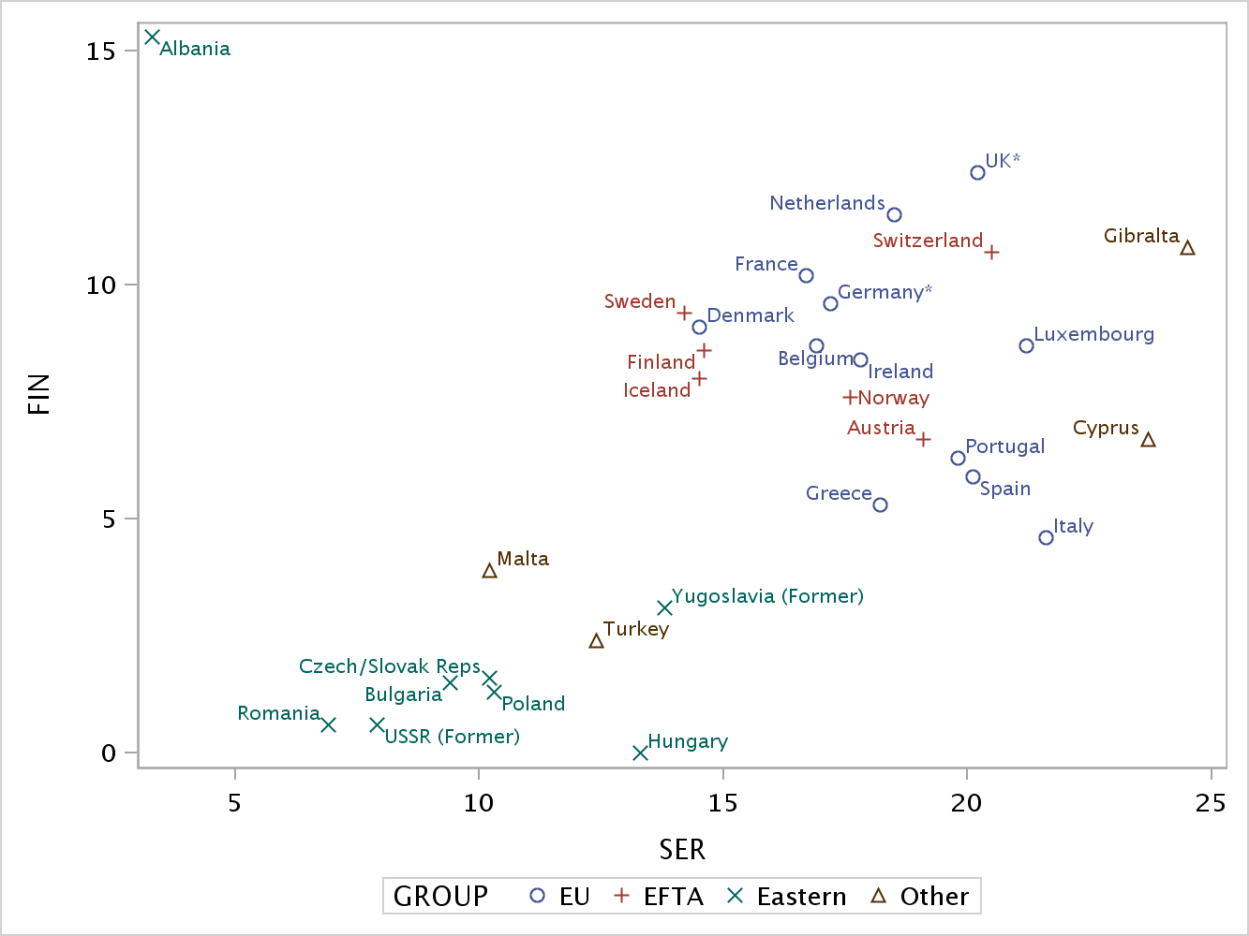
For example, the variable SPS (social and personal services) has a strong negative correlation (-0.8115) with AGR (agriculture). Another example is the variable MAN (manufacturing) has a fairly strong negative correlation with MIN (mining).

The scatterplot matrix was produced to look for potentially interesting areas to conduct cluster analysis. Looking through the matrix, we try to identify a plot that separates the data points into two or more groups (or clusters). This might indicate an underlying difference between the countries in each of the clusters than can be further exploited through cluster analysis and segmentation. The scatterplot matrix is reproduced below.



Looking through the different plots on the matrix, the plot that looked most interesting was the plot of FIN (finance) and SER (services). This plot shows what appears to be two distinct clusters of data points. One is located in the upper right of the graph, and the other is located in the bottom left. There appears to be a divide between both clusters.

Using PROC SGPLOT, we can take a closer look at the scatterplot of FIN on the Y-axis and SER on the X-axis. The plot shows an interesting relationship.



Upon closer inspection, the plot does in fact appear to separate the data into two distinct groups or clusters. On the lower left of the plot, we find most of the Eastern European (Eastern) countries in the dataset, along with the unclassified countries Malta and Turkey. In the upper right of the plot, we find all of the European Union (EU) nations, along with the European Free Trade Association (EFTA) countries and the two unclassified countries Gibraltar and Cyprus. Albania (in the top left corner of the plot) appears to be an outlier.

The fact that the plot of FIN and SER separates the Eastern European countries so nicely from the Western European (EU and EFTA) countries, indicates that this relationship would be an excellent candidate to exploit using cluster analysis.

Based on the plot depicted above, we can also get an idea of where each of the unclassified countries could be assigned, if we were interested in classifying them within one of the existing groups. Malta and Turkey, based on their employment percentages in the areas of finance and services, exhibit the traits of Eastern European countries. Turkey, based on its geographic location and proximity to other Eastern European countries such as Bulgaria and Romania, could be classified as Eastern European.

Malta, an island south of Italy, is a little harder to justify as Eastern European based on geography. Culturally it is also more Western European, having gained its independence from the United Kingdom in

the 1960s. It was also admitted to the European Union in 2004. Despite that last fact, I would classify Malta as an Eastern European country for predictive purposes. It seems to exhibit traits more closely aligned with Eastern European countries, at least upon this initial screening. Given its EU affiliation, I would look at it more closely before assigning it to one of the groups.

Cyprus was admitted to the EU in 2004 and seems to exhibit traits congruent with other EU countries. I would classify Cyprus as a EU nation. Gibraltar is more difficult to classify. It as a self-governing overseas territory of the UK, located in Western Europe just south of Spain. According to our plot, it exhibits traits similar to EU countries, and since it is a territory of the UK, which is a EU country, I would also classify Gibraltar as an EU country.

Part 2: Principal Components Analysis

Our data set has nine variables. One method of reducing the dimensionality of our data set is to use principal components analysis. If we perform a principal components analysis, what would the resulting dimensionality be, i.e. how many components should we keep? What decision rule are you using to determine how many of the principal components to keep? Are there any other competing decision rules that you could use? Include the table of the eigenvalues of the correlation matrix, the scree plot, and the "Component Pattern Profiles" plot. Interpret these plots and make the appropriate comments. See Chapter 3 of *Applied Multivariate Data Analysis* for a statistical reference to principal components analysis.

```
ods graphics on;
title Principal Components Analysis using PROC PRINCOMP;
proc princomp data=temp out=pca_9components outstat=eigenvectors plots=all;
run;
ods graphics off;
```

Principal components analysis is an application of numerical linear algebra that reduces the dimensionality of the data set by combining the predictor variables into principal components that explain the variance in the data and are orthogonal (i.e. uncorrelated with each other).

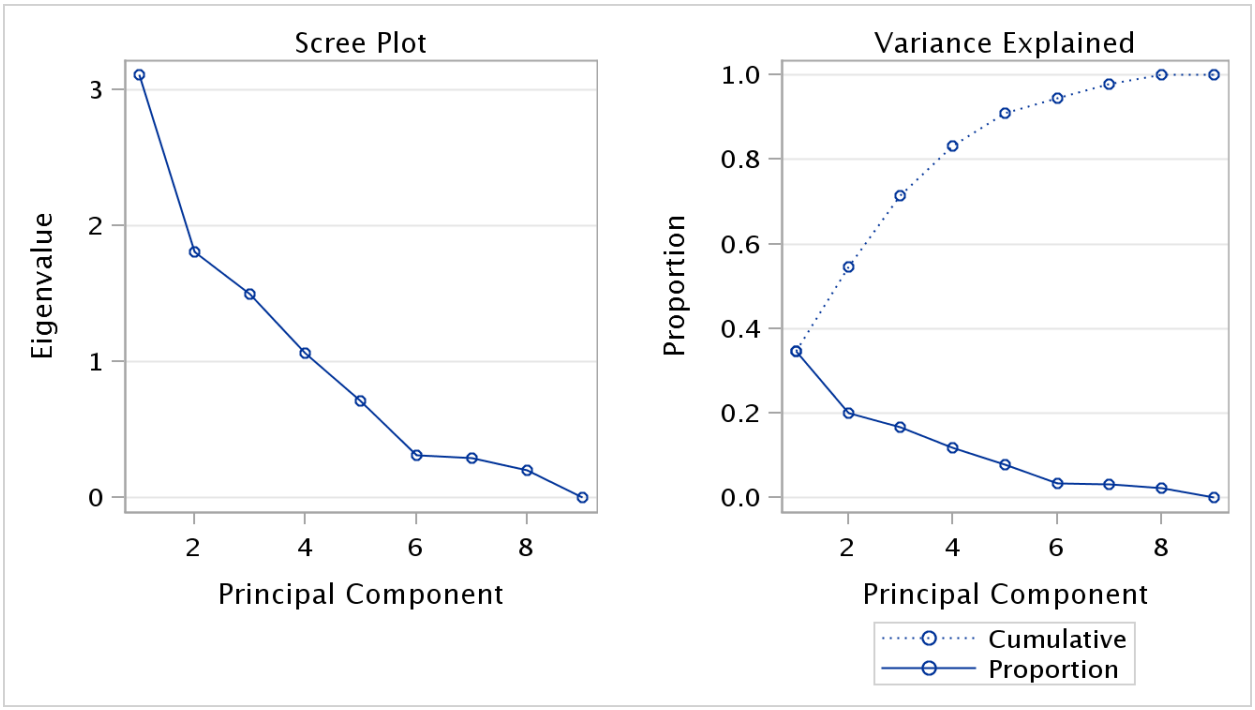
While all of the principal components taken together explain 100 percent of the variance in the data, we can reduce the dimensionality of the data by using only a few of the components. The components are ordered so that they account for a descending proportion of the variance (with the first component accounting for the largest amount of the variance, etc.). By using only the first few components, we can reduce the dimensionality of the data while also explaining a large percentage of the variance.

There are several ways to decide on how many principal components to keep. But first, we'll take a look at the results of running the analysis on the European employment data set.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.11225795	1.30302071	0.3458	0.3458
2	1.80923724	0.31301704	0.2010	0.5468
3	1.49622020	0.43277636	0.1662	0.7131
4	1.06344384	0.35318631	0.1182	0.8312
5	0.71025753	0.39891874	0.0789	0.9102
6	0.31133879	0.01791787	0.0346	0.9448
7	0.29342091	0.08960446	0.0326	0.9774
8	0.20381645	0.20380935	0.0226	1.0000
9	0.00000710		0.0000	1.0000

The eigenvalues of the correlation matrix shows the egeinvalue for each principal component (of which there are nine—the same number of components as original variables in the data set). They also show the proportion of the variance that each component accounts for (along with the cumulative percent of the variation accounted for). We can see, for example, that the first five components account for 91.02 percent of the variance in the data (row 5 under the Cumulative column).

One of the ways to determine how many principal components to keep is to look at the scree plot. The scree plot shows the eigenvalue of each component on the Y-axis plotted against the principal component number on the X-axis. It is often accompanied by the variance explained plot which shows the proportion of variance explained plotted against the principal component number.



The key to evaluating the scree plot is to look for the kink in the plot where the curve flattens out. The rule of thumb is to keep all of the principal components up through the last component before the curve flattens. In this example, we would keep the first six components. The scree plot shows a clear flattening after the sixth component is listed.

The variance explained plot shows the proportion of the variance explained by each of the components, along with the cumulative proportion explained (dotted line). We can see that by using the first six components, we account for a large proportion of the variance (referencing the eigenvalues of the correlation matrix table, we can see that that proportion is equal to 94.48 percent).

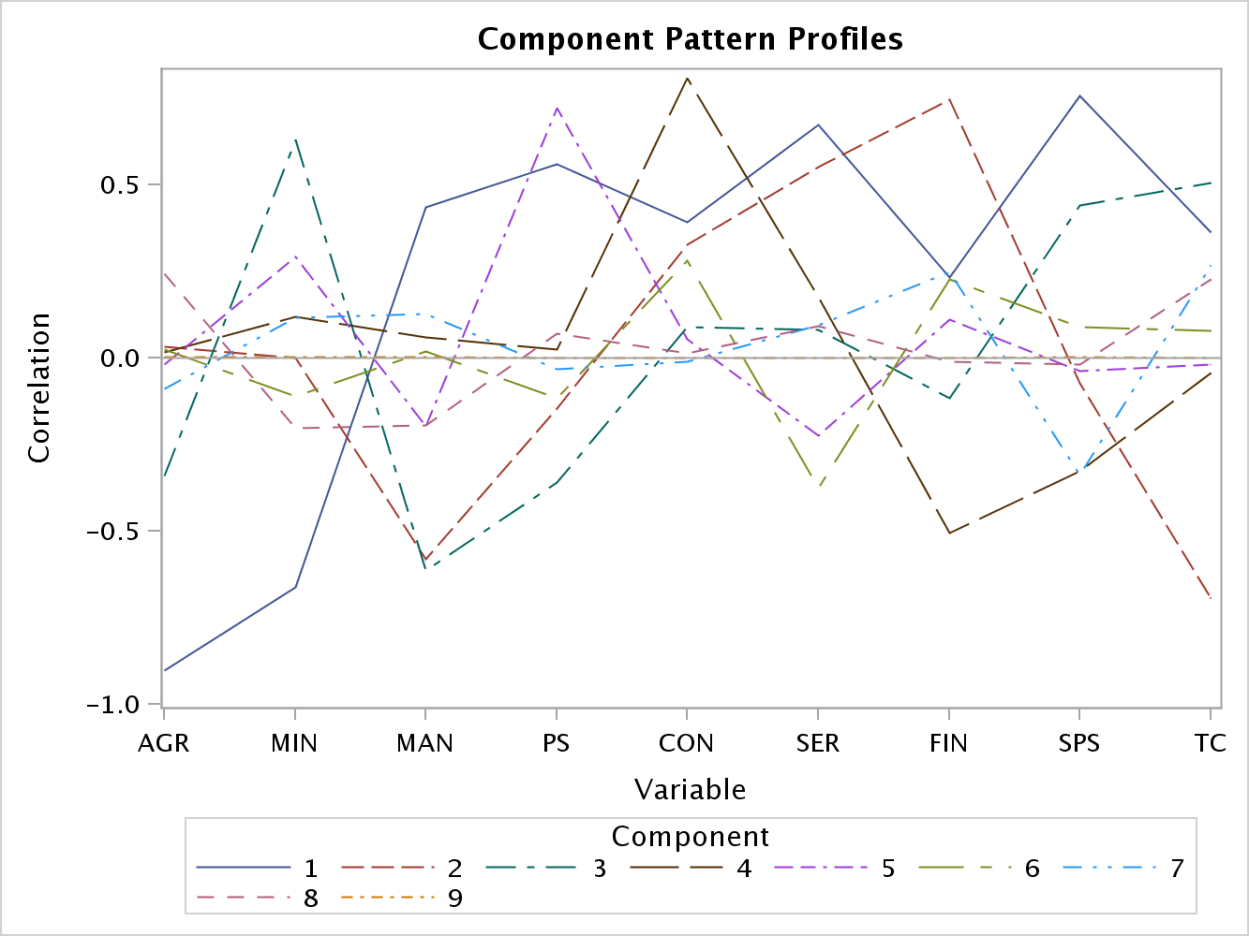
If we use the scree plot as a decision rule, we end up with six principal components that explain almost 95 percent of the variance, reducing the dimensionality of the data set down from its original nine predictor variables.

There are competing decision rules for determining the number of components to keep. In all cases, there is a tradeoff between further reduction in the dimensionality of the data (i.e. using fewer principal components) and explaining less of the variance.

One ad-hoc decision rule is to keep the number of principal components necessary to explain some predetermined proportion of the variance—usually between 70 and 90 percent. In this case, that would lead us to use three (accounting for 71.31 percent of the variance) or four (accounting for 83.12 percent of the variance) principal components, as opposed to six.

Another rule is to keep all principal components whose eigenvalue is larger than the average eigenvalue. When the eigenvalues are derived from the correlation matrix, as they are in this case, the average eigenvalue will also be 1. Using this rule, we would keep the first four principal components. A modification of this rule proposes using 0.7 as the cutoff eigenvalue number, which would lead us to use the first five components.

Principal components analysis is primarily used to reduce the dimensionality of the data and the deal with the presence of multicollinearity. Besides reducing the dimensionality formally through the development of the principal components, the components themselves give the analysts information that can be used heuristically to develop new predictor variables. The Component Patterns Profile plot helps the analyst detect which predictor variables are dominant through analyzing the “makeup” of each component.



The plot above displays a line of each of the nine principal components. Each of the lines shows the correlation coefficient value between each of the original nine predictor variables and the principal component represented by the line.

For example, the AGR is strongly negatively correlated with the first principle component (seen by the plot point for Line 1 on the Y-axis above AGR). Following Line 1 across the plot, we get a picture of what influences the makeup of the first principal component. The first component has negative correlations for percent employment in agriculture and mining, with strong positive correlations in percent employment in services and social and personal services. Informally, we can consider the first principal component (which captures the largest amount of the variance) to be the degree to which a country has a service economy vs. an agrarian or mining economy.

The same type of analysis could be done for each of the principal components. Component number two for example might represent a white collar vs. blue collar distinction (with a strong positive correlation for finance sector employment and negative correlations for manufacturing, transport and communications employment.

This type of analysis is less formal than using the components themselves, but it might also help in reducing the dimensionality of the data by focusing on key predictor variables (such as agriculture and

finance) or developing new predictor variables or indices based on what is learned from the component profiles.

Part 3: Factor Analysis

A second approach to reducing the dimensionality of our data set is to use factor analysis. Before we begin applying a factor analysis, you will need to answer a question? Provide your answer in green.

Are principal components analysis and factor analysis the same statistical method? How are they different?

Principal components and factor analysis are not the same statistical method but they do have the same goal—to explain multivariate data using a reduce number of dimensions that originally found in the data set.

Some of the key differences are as follows. Principal components analysis (PCA) is an application of numerical linear algebra, while factor analysis (FA) is based on statistical assumptions. PCA always produces orthogonal components, while FA does not.

Perhaps most importantly, two analysts using PCA on the same data will almost always arrive at the same conclusion (as to how many components to keep, etc.). The components themselves will always be identical. This is not true of FA. Analysts must make decisions about how much to rotate the factors, which can lead to different results or conclusions being drawn from the same data.

Other differences include: FA postulates a model for the data, whereas PCA does not, and FA tries to explain the correlations or covariance between the variables while PCA focuses on explaining the variance.

PCA is considered the more statistically sound of the two techniques, based on its derivation and repeatable conclusions. FA, when it is favored, is usually chosen for interpretability reasons rather than model performance.

The SAS procedure for performing a variety of implementations of factor analysis is PROC FACTOR. Let's perform a factor analysis on our data using different methods of factor analysis. See Chapter 12 of *Applied Multivariate Data Analysis* for a statistical reference to factor analysis (Exploratory Factor Analysis).

Principal Components Using PROC FACTOR:

In addition to using PROC PRINCOMP to perform a principal components analysis SAS will allow you to perform a principal components analysis using PROC FACTOR. Run this code and compare the output from PROC FACTOR to the output from PROC PRINCOMP.

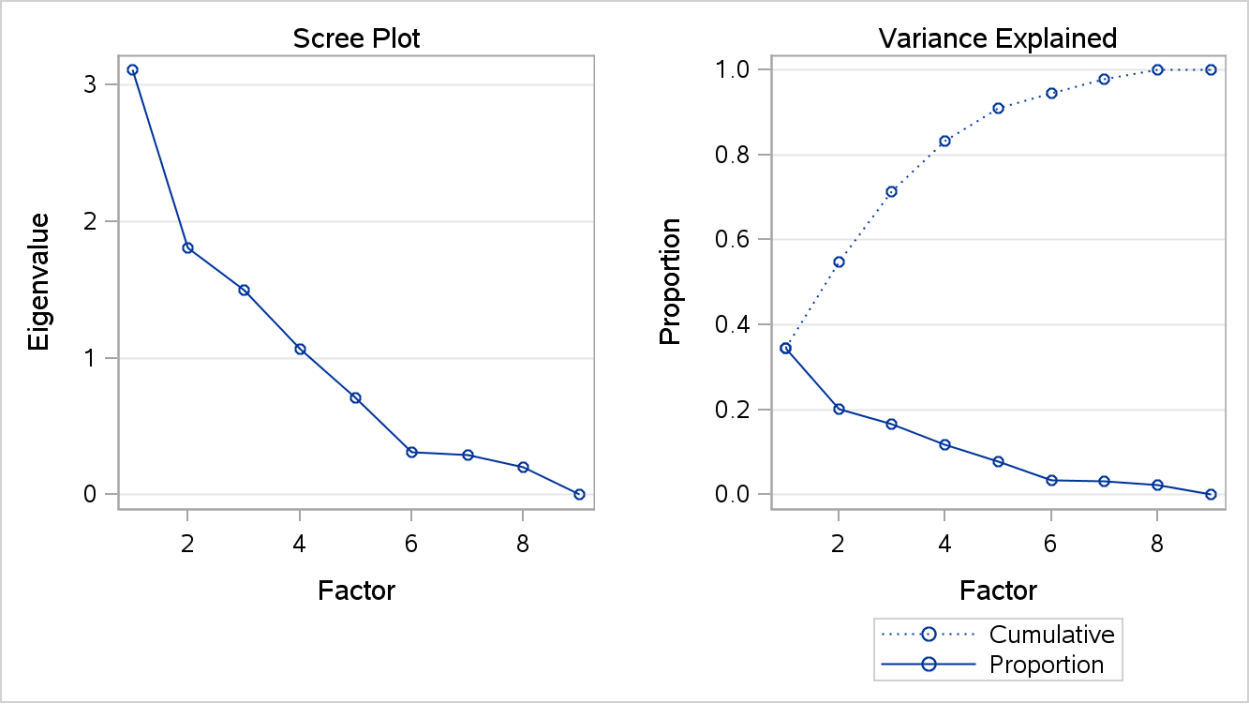
```
ods graphics on;
title Principal Components Analysis using PROC FACTOR;
```

```
proc factor data=temp method=principal out=pca_factors
      nfactors=9 score plots=scree;
run;
ods graphics off;
```

Conducting principal components analysis using PROC FACTOR yields the same results as PROC PRINCOMP. The eigenvalues of the correlation matrix (below) for PROC FACTOR is

Eigenvalues of the Correlation Matrix: Total = 9 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.11225795	1.30302071	0.3458	0.3458
2	1.80923724	0.31301704	0.2010	0.5468
3	1.49622020	0.43277636	0.1662	0.7131
4	1.06344384	0.35318631	0.1182	0.8312
5	0.71025753	0.39891874	0.0789	0.9102
6	0.31133879	0.01791787	0.0346	0.9448
7	0.29342091	0.08960446	0.0326	0.9774
8	0.20381645	0.20380935	0.0226	1.0000
9	0.00000710		0.0000	1.0000

exactly the same as the matrix produced by PROC PRINCOMP in Part 2 of this assignment. The scree plot and other output is also the same.



Iterated Principal Factor Analysis:

Now let's perform a legitimate factor analysis using PROC FACTOR. We will run an Iterated Principal Factor Analysis using the following SAS code.

```
ods graphics on;
title Iterated Principal Factor Analysis using PROC FACTOR;
proc factor data=temp method=prinit out=pfa_factors
      nfactors=9 score plots=scree;
run;
ods graphics off;
```

Is this a valid factor analysis? (Hint: the answer is no.) Why is this not a valid factor analysis? Keep reducing the number for *nfactors* until you get a valid factor analysis. Report the "Eigenvalues of the Reduced Correlation Matrix", the "Factor Pattern", the "Variance Explained by Each Factor", and the "Final Communality Estimates" tables and make the appropriate comments on the results in these tables. As part of your comments do you have an interpretation of the factor loadings.

```
ods graphics on;
title Iterated Principal Factor Analysis using PROC FACTOR;
proc factor data=temp method=prinit out=pfa_factors
      nfactors=2 score plots=scree;
run;
ods graphics off;
```

Factor analysis is a technique that seeks to identify hypothetical variables—known as the common factors—that contribute to the variance of two or more of the predictor variables in a data set. Like principal components analysis, factor analysis is a technique used to reduce the dimensionality of the data set. Unlike the components in PCA, which are observable linear combinations of the predictor variables in the data set, the common factors are unobservable and represent some shared underlying aspects of the predictor variables that can then be used to reduce the dimensionality of the data.

Each original variable is made up of some proportion of a common factor and some proportion of a unique factor (that which is only explained by that variable). For factor analysis to be valid, the unique factors must be uncorrelated with one another and with the common factors. This necessitates that there must be fewer common factors than there are original predictor variables.

The communality of each variable in the factor analysis model is the proportion of the variance shared by common factors. After the communality estimates are found, principal factor analysis requires principal components analysis be performed on the reduced covariance matrix. A unique, valid solution is said to have been found when all of the estimated specific variances in the principal factor analysis are non-negative. This will be the case unless at some point in the iterative process the communality estimate is greater than the corresponding manifest variable.

Running the first PROC FACTOR statement above yields results where the estimated communality is above 1.0, which is not valid. SAS returns an error that says there are too many factors for a unique solution and there is a communality greater than 1.0. SAS continues to yield this error until the number of factors has been reduced to two.

The Eigenvalues of the Reduced Correlation Matrix table for the factor analysis keeping two factors is below.

Eigenvalues of the Reduced Correlation Matrix: Total = 4.05321736 Average = 0.45035748				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.71284161	1.37209917	0.6693	0.6693
2	1.34074244	0.49070253	0.3308	1.0001
3	0.85003991	0.49638124	0.2097	1.2098
4	0.35365867	0.31902319	0.0873	1.2971
5	0.03463548	0.15295505	0.0085	1.3056
6	-.11831957	0.04311470	-0.0292	1.2764
7	-.16143427	0.14176285	-0.0398	1.2366
8	-.30319712	0.35255266	-0.0748	1.1618
9	-.65574978		-0.1618	1.0000

The table shows the eigenvalues for each of the nine principal components estimated from the reduced correlation matrix. Only the first two components are used in our two-factor model. The eigenvalue for each of the first two components represents the variance explained by each of the factors, as seen in the Variance Explained by Each Factor Table, which yields identical values.

Variance Explained by Each Factor	
Factor1	Factor2
2.7128416	1.3407424

The proportion of the variance explained by each of the factors can be found in the principal components table. The first factor accounts for 66.93 percent of the shared variance in the predictor variables, and the second factor accounts for 33.08 percent of the shared variance.

For interpretation purposes, the Factor Pattern table can help the analyst identify what each factor represents.

Factor Pattern		
	Factor1	Factor2
AGR	-0.97518	0.09287
MIN	-0.51295	-0.14002
MAN	0.31557	-0.26842
PS	0.42470	-0.02636
CON	0.31085	0.21138
SER	0.64961	0.50915
FIN	0.19597	0.57137
SPS	0.71515	-0.13367
TC	0.38771	-0.76911

To interpret each factor, we look at the correlation of each of the original predictor variables to the

factor—or the "factor loading". For Factor 1, we can see that employment in the agriculture sector has a very strong negative correlation with the factor, and employment in the service and social and personal services sectors have strong positive correlations. Mining also has a medium strength negative correlation with Factor 1. We might interpret these results by saying that Factor 1 represents the proportion of employment in service sectors as opposed to those involving land resources (agriculture and mining).

Factor 2 has medium-strength correlations with service and finance sector employment, and a strong negative correlation with transport and communications sector employment. This might be interpreted as the health of the finance sector, or perhaps it is related to the urbanization of the nation under study, considering the de-emphasis on transportation and emphasis on traditional urban employment sectors.

These results are fairly similar to those found using principal components analysis.

Final Communality Estimates: Total = 4.053584								
AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
0.95960915	0.28272048	0.17163326	0.18106461	0.14131366	0.68122599	0.36486285	0.52930618	0.74184788

The final communality estimates (above) for the iterated principal factor analysis show the proportion of each of the original variables that is explained by a common factor (as described previously). Here we can see that none of the estimates is above 1.0 (explaining the lack of errors in computing the two-factor model). We can also see the extent to which each variable is accounted for by a common factor.

Agriculture sector employment, in particular, seems to be largely accounted for by common factors in the other predictor variables. The communality estimate indicates that 95.96 percent of the variance in agriculture sector employment can be accounted for by factors common to at least one of the other predictor variables. A possible explanation for this is that agriculture employment tends to be higher in lesser-developed economies. These economies will also tend to have similar rates of finance, service, manufacturing and other sector levels of employment. Transport and communication sector employment and service employment also had high proportions of their variance explained by common factors.

Maximum Likelihood Factor Analysis:

An alternative to iterated principal factor analysis is maximum likelihood factor analysis.

```
ods graphics on;
title Maximum Likelihood Factor Analysis using PROC FACTOR;
proc factor data=temp method=ml out=fa_ml
      outstat=fa_ml_stats mineigen=0 priors=smc nfactors=2 score ;
run;
ods graphics off;
```

Is this a valid factor analysis? If this is not a valid factor analysis, then why is this not a valid factor analysis? If this is a valid factor analysis then report the "Eigenvalues of the Reduced Correlation

Matrix", the "Factor Pattern", the "Variance Explained by Each Factor", and the "Final Communality Estimates" tables and make the appropriate comments on the results in these tables.

This particular maximum likelihood factor analysis is not a valid factor analysis. As the iteration output from the procedure shows, one of the communalities was estimated as 1.67185, which exceeds 1.0 and thus indicates that the estimate for the specific variance would be negative, which is a non-valid result.

Iteration	Criterion	Ridge	Change	Communalities								
1	11.3654961	0.0000	0.8835	0.82598	0.32044	0.97094	0.24006	0.11555	0.72891	0.12447	0.43365	0.21456
2	10.1887401	0.0000	0.7009	0.99882	0.27870	1.67185	0.16167	0.16052	0.33315	0.09960	0.58354	0.11491

The highlighted communality estimate is the estimate that violates the restriction and yields the non-valid result.

Unweighted Least Squares Factor Analysis:

Another type of factor analysis, which is an alternative to both iterated principal factor analysis and maximum likelihood factor analysis, is unweighted least squares factor analysis.

```
ods graphics on;
title Unweighted Least Squares Factor Analysis using PROC FACTOR;
proc factor data=temp method=uls out=fa_uls
      outstat=uls_stats mineigen=0 priors=smc nfactors=2 score ;
run;
ods graphics off;
```

Is this a valid factor analysis? If this is not a valid factor analysis, then why is this not a valid factor analysis? If this is a valid factor analysis then report the "Eigenvalues of the Reduced Correlation Matrix", the "Factor Pattern", the "Variance Explained by Each Factor", and the "Final Communality Estimates" tables and make the appropriate comments on the results in these tables. Are the estimated factor loadings from the unweighted least squares factor analysis significantly different from the factor loadings from iterated principal factor analysis?

The unweighted least squares factor analysis above is a valid factor analysis. All of the communality estimates are below 1.0 and the convergence criteria is satisfied, meaning there is a unique solution to the proposed two-factor unweighted least squares factor model.

Eigenvalues of the Reduced Correlation Matrix: Total = 4.05572183 Average = 0.45063576				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.71285754	1.36999310	0.6689	0.6689
2	1.34286444	0.49285732	0.3311	1.0000
3	0.85000713	0.49691989	0.2096	1.2096
4	0.35308723	0.31810025	0.0871	1.2966
5	0.03498698	0.15226038	0.0086	1.3053
6	-.11727340	0.04386359	-0.0289	1.2764

Eigenvalues of the Reduced Correlation Matrix: Total = 4.05572183 Average = 0.45063576				
	Eigenvalue	Difference	Proportion	Cumulative
7	-.16113699	0.14231528	-0.0397	1.2366
8	-.30345226	0.35276659	-0.0748	1.1618
9	-.65621886		-0.1618	1.0000

The Eigenvalues of the Reduced Correlation Matrix carry the same interpretation as in the iterative principal factor analysis case. The eigenvalues for the first two components—the factors kept by the model—are the same to the one-hundredth decimal place as they were in the previously valid model.

Variance Explained by Each Factor	
Factor1	Factor2
2.7128575	1.3428644

The proportion of the variance explained by each factor varies only slightly in the weighted least squares factor model (Factor 1: ULS factor model proportion of variance = 0.6689 vs. 0.6693 in the IPF model; Factor 2: ULS factor model proportion of variance = 0.3311 vs. 0.3308 in the IPF model).

Factor Pattern		
	Factor1	Factor2
AGR	-0.97517	0.09194
MIN	-0.51286	-0.14170
MAN	0.31559	-0.26646
PS	0.42467	-0.02506
CON	0.31070	0.21156
SER	0.64894	0.50869
FIN	0.19552	0.57058
SPS	0.71530	-0.13326
TC	0.38911	-0.77192

The factor loadings for the unweighted least squares factor model—shown in the Factor Pattern table—are all exactly the same as the iterated principal factor model when rounded to two decimal places. This indicates that they would have the same interpretation as the previous model.

Final Communality Estimates: Total = 4.055722								
AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
0.95940511	0.28310262	0.17059636	0.18097402	0.14129021	0.67989023	0.36379165	0.52941077	0.74726102

The final communality estimates (above) are also nearly identical to the iterated principal factor model.

Part 4: Factor Rotations

We will now consider rotating a set of factors. Before we begin you will need to answer a question? Provide your answer in green.

What is the difference between an oblique and an orthogonal factor rotation? Is there any reason to choose an oblique rotation over an orthogonal rotation, or vice-versa?

Orthogonal rotation of the factor leads to factors that are wholly uncorrelated to one another and ranked in the order of descending proportion of the variance explained (similar to the components in principal components analysis). Oblique rotation leads to factors that are correlated to some degree, but may be easier to interpret.

The reasons to consider the orthogonal rotation are that each of the factors uniquely accounts for the proportion of explained variance. That is, since each factor is uncorrelated with the others, each factor's explanation of the shared variance in the predictor variables is unique to that factor. There is no overlap between the factors.

The problem with the orthogonal rotation is that sometime some of the original predictor variables will be strongly correlated with multiple factors. This means it is difficult to interpret the meaning of each factor, because multiple factors might be influenced by the same variables. An oblique rotation lets the analyst rotate the factors about the origin of their axes, changing the factor loadings to increase interpretability.

There are several problems with oblique rotation. First, it is somewhat arbitrary and leads to non-uniform results for factor analysis. Two different analysts working with the same data could theoretically rotate the factors in such a way as to get contrasting results and conclusions. Additionally, one of the main benefits of the orthogonal rotation—the fact that the factors will account for maximum variance in descending order of importance—will be lost after rotation.

VARIMAX Factor Rotation

First we will perform an orthogonal factor rotation using a VARIMAX rotation.

```
ods graphics on;
title A VARIMAX Rotation of a Unweighted Least Squares Factor Analysis using
PROC FACTOR;
proc factor data=temp method=uls rotate=varimax out=uls_varimax
      outstat=varimax_stats mineigen=0 priors=max nfactors=2 score
      plots=(initloadings preloadings loadings scree) ;
run;
ods graphics off;
```

Report the "Eigenvalues of the Reduced Correlation Matrix", the "Factor Pattern", the "Variance Explained by Each Factor", and the "Final Communality Estimates" tables and the same output for the rotated factors. Make the appropriate comments on the results in these tables. Did the factor rotation

change the variance explained by each factor? Did the factor rotation change the interpretation of the factor loadings? Did the factor rotation change the final communality estimates?

Eigenvalues of the Reduced Correlation Matrix: Total = 4.0557028 Average = 0.45063364				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.71285587	1.37000875	0.6689	0.6689
2	1.34284712	0.49283771	0.3311	1.0000
3	0.85000941	0.49691734	0.2096	1.2096
4	0.35309207	0.31810797	0.0871	1.2966
5	0.03498410	0.15226674	0.0086	1.3053
6	-.11728264	0.04385963	-0.0289	1.2764
7	-.16114227	0.14230642	-0.0397	1.2366
8	-.30344869	0.35276348	-0.0748	1.1618
9	-.65621217		-0.1618	1.0000

The eigenvalues for the varimax rotation factor model are nearly identical to the unweighted least squares factor analysis model. This is likely because the default rotation for factor analysis is the orthogonal rotation. The proportion of the shared variance explained by each of the factors is almost exactly the same, as are the factor loadings and final communality estimates (shown below).

Variance Explained by Each Factor	
Factor1	Factor2
2.7128559	1.3428471

Factor Pattern		
	Factor1	Factor2
AGR	-0.97517	0.09195
MIN	-0.51286	-0.14168
MAN	0.31559	-0.26648
PS	0.42467	-0.02508
CON	0.31070	0.21156
SER	0.64895	0.50870
FIN	0.19552	0.57059
SPS	0.71530	-0.13327
TC	0.38909	-0.77189

Final Communality Estimates: Total = 4.055703								
AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
0.95940387	0.28309868	0.17060768	0.18097518	0.14129083	0.67990578	0.36379989	0.52940794	0.74721313

PROMAX Factor Rotation

Now we will perform an oblique factor rotation using a PROMAX rotation.

```
ods graphics on;
title A PROMAX Rotation of a Unweighted Least Squares Factor Analysis using
PROC FACTOR;
proc factor data=temp method=uls rotate=promax out=uls_promax
      outstat=promax_stats mineigen=0 priors=max nfactors=2 score
      plots=(initloadings preloadings loadings scree) ;
run;
ods graphics off;
```

Report the "Eigenvalues of the Reduced Correlation Matrix", the "Factor Pattern", the "Variance Explained by Each Factor", and the "Final Communality Estimates" tables and the same output for the rotated factors. Make the appropriate comments on the results in these tables. Did the factor rotation change the variance explained by each factor? Did the factor rotation change the interpretation of the factor loadings? Did the factor rotation change the final communality estimates?

Somewhat surprisingly, the oblique promax rotation of the factor model does not produce changes in the explained variance, factor loadings and communality estimates as expected. A look at the Eigenvalues of the Reduced Correlation Matrix and the Variance Explained by Each Factor tables shows some nearly identical results to the original factor model and the varimax rotation of the model.

Eigenvalues of the Reduced Correlation Matrix: Total = 4.0557028 Average = 0.45063364				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.71285587	1.37000875	0.6689	0.6689
2	1.34284712	0.49283771	0.3311	1.0000
3	0.85000941	0.49691734	0.2096	1.2096
4	0.35309207	0.31810797	0.0871	1.2966
5	0.03498410	0.15226674	0.0086	1.3053
6	-.11728264	0.04385963	-0.0289	1.2764
7	-.16114227	0.14230642	-0.0397	1.2366
8	-.30344869	0.35276348	-0.0748	1.1618
9	-.65621217		-0.1618	1.0000

Variance Explained by Each Factor	
Factor1	Factor2
2.7128559	1.3428471

The factor loadings (in the Factor Pattern table) and the communality estimates also remain unchanged with the exception of slight rounding differences.

Factor Pattern		
	Factor1	Factor2
AGR	-0.97517	0.09195
MIN	-0.51286	-0.14168
MAN	0.31559	-0.26648
PS	0.42467	-0.02508
CON	0.31070	0.21156
SER	0.64895	0.50870
FIN	0.19552	0.57059
SPS	0.71530	-0.13327
TC	0.38909	-0.77189

Final Communality Estimates: Total = 4.055703								
AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
0.95940387	0.28309868	0.17060768	0.18097518	0.14129083	0.67990578	0.36379989	0.52940794	0.74721313

Part 5: Cluster Analysis

We will begin our discussion of cluster analysis by making a pair of scatterplots.

```
ods graphics on;
proc sgplot data=temp;
title 'Scatterplot of Raw Data: FIN*SER';
scatter y=fin x=ser / datalabel=country group=group;
run; quit;
ods graphics off;

ods graphics on;
proc sgplot data=temp;
title 'Scatterplot of Raw Data: MAN*SER';
scatter y=man x=ser / datalabel=country group=group;
run; quit;
ods graphics off;
```

How many clusters do you see in the scatterplot of FIN*SER? How many clusters do you see in the scatterplot of MAN*SER?

As previously mentioned, the scatterplot of FIN*SER yields two distinct clusters—one containing the EU and EFTA countries and the other with the Eastern European countries. The clusters are found in the top right and bottom left of the plot, respectively. The caveat here is Albania, which is in the extreme top left of the plot all by itself, and could be considered an outlier or a separate cluster on its own.

The MAN*SER plot shows a different picture. The EU and EFTA countries again appear to be clustered together in the middle right of the graph. The Eastern European countries appear to be split into potentially two different clusters, however, with one in the top left of the plot and the other in the bottom left. That would indicate there are three clusters in the MAN*SER plot.

Clearly different projections of the data will produce different clustering results. We need to be cognizant of this fact.

Now we will use PROC CLUSTER to create a set of clusters algorithmically. Note that PROC CLUSTER performs *hierarchical clustering* (see Chapter 6 in *Applied Multivariate Data Analysis*) so we do not need to specify the number of clusters in advance. We will use the SAS procedure PROC TREE to assign observations to a specified number of clusters after we have performed the hierarchical clustering.

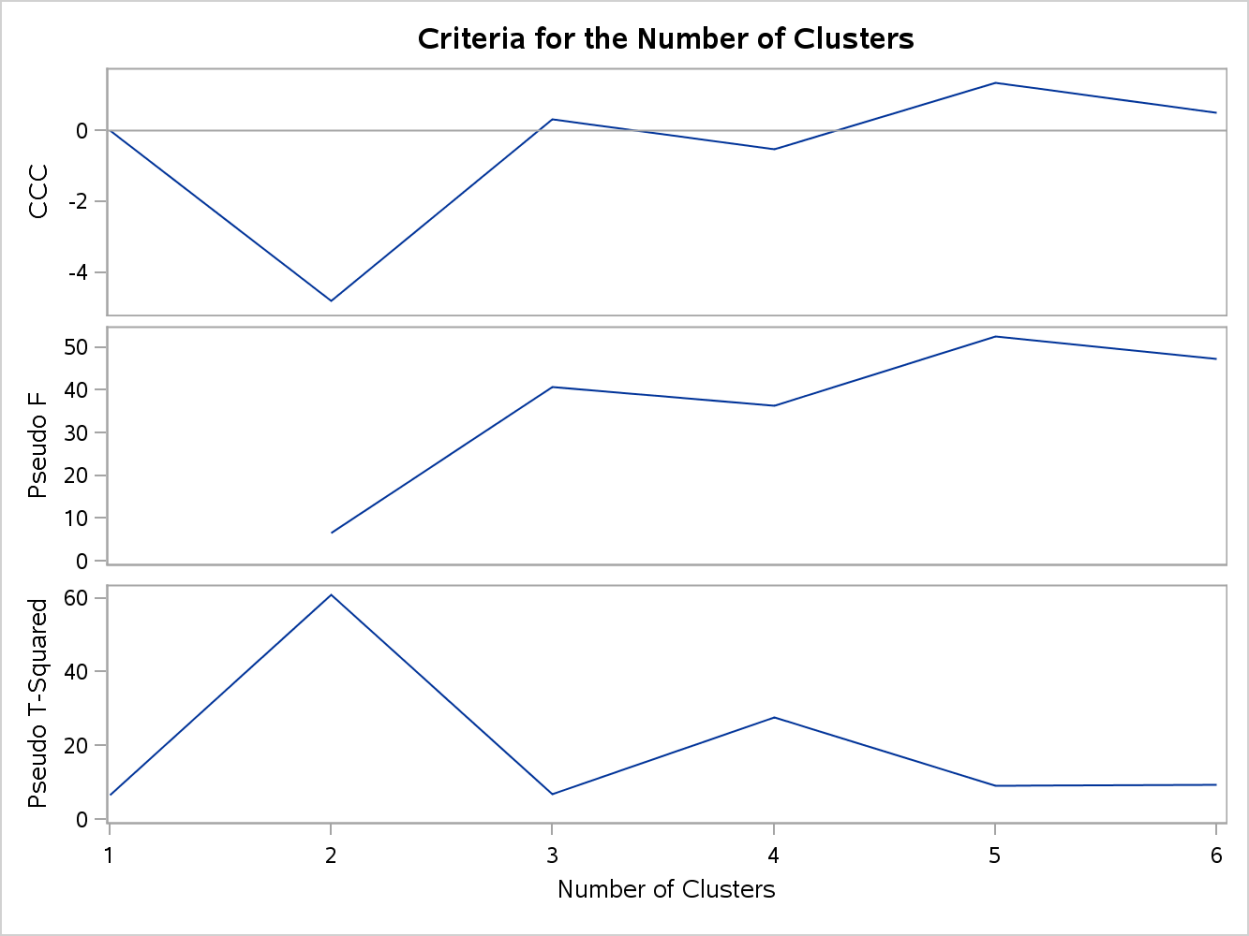
```
ods graphics on;
proc cluster data=temp method=average outtree=tree1 pseudo ccc plots=all;
var fin ser;
id country;
run; quit;
ods graphics off;
```

How do we interpret the measures of CCC, Pseudo F, and Pseudo T-Squared? How do we interpret the plots for these three measures?

CCC is the cubic clustering criterion. It was developed by Warren Sarle in 1983 and can be used for to estimate the number of clusters in a population. The CCC makes assumptions about the shape of the clusters and computers how many clusters are likely given the population size and shape. The CCC is usually accurate in large samples.

The Pseudo F and Pseudo T-Squared statistics are both related to CCC (Pseudo F can be transformed into Pseudo T-Squared and CCC). All of these methods are used to estimate the number of clusters in a population when using hierarchical clustering methods.

It is recommended that the three statistics be used in conjunction with each other to determine how many clusters are present in the population data.



Using the plots (above), the number of clusters is typically estimated to be the number where the CCC and Pseudo F statistics have local peaks while the Pseudo T-Squared statistic has a small value (followed by a larger value).

In our data, this relationship appears to be present when the number of clusters is three. At n=3 clusters, CCC and Pseudo F have local peaks (both values on each side of n=3 are smaller for both statistics) and the Pseudo T-Squared statistic has a small value followed by a larger value when the number of clusters equals four. This suggest that the number of clusters in the population is three, which is consistent with the MAN*SER scatterplot and/or the FIN*SER scatterplot, taking into account Albania.

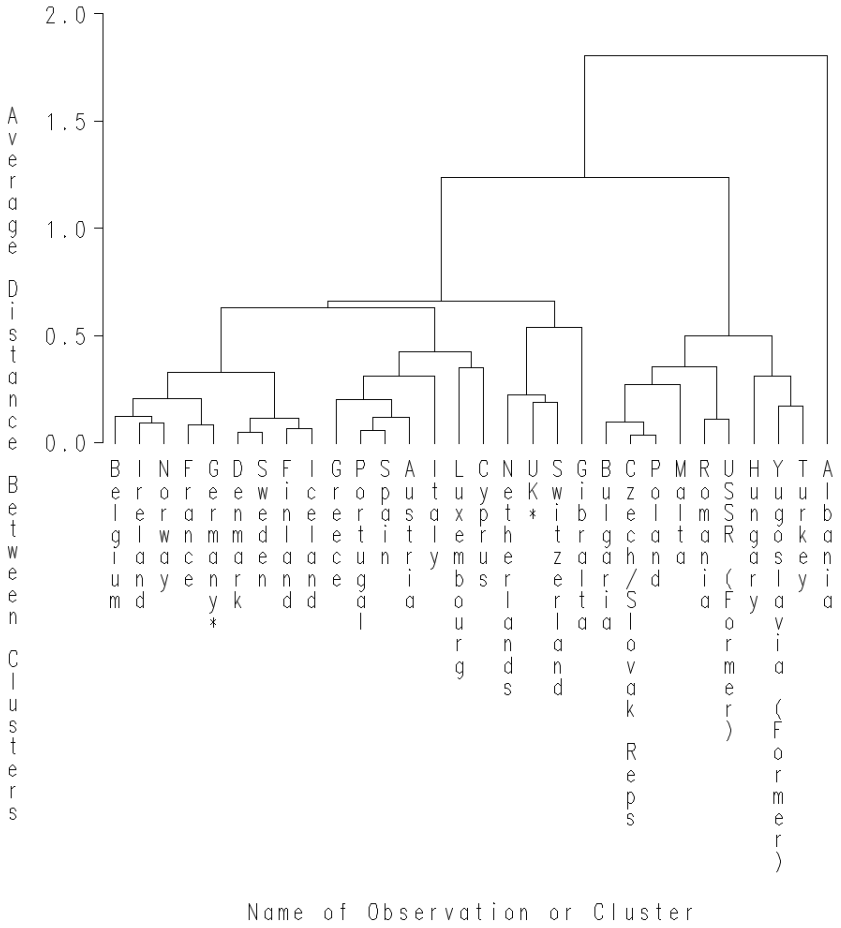
We can use PROC TREE to assign our data to a set number of clusters. Let's compare the output when we assign the observations to four clusters and then to three clusters.

```
ods graphics on;
proc tree data=tree1 ncl=4 out=_4_clusters;
copy fin ser;
run; quit;
ods graphics off;

ods graphics on;
proc tree data=tree1 ncl=3 out=_3_clusters;
```

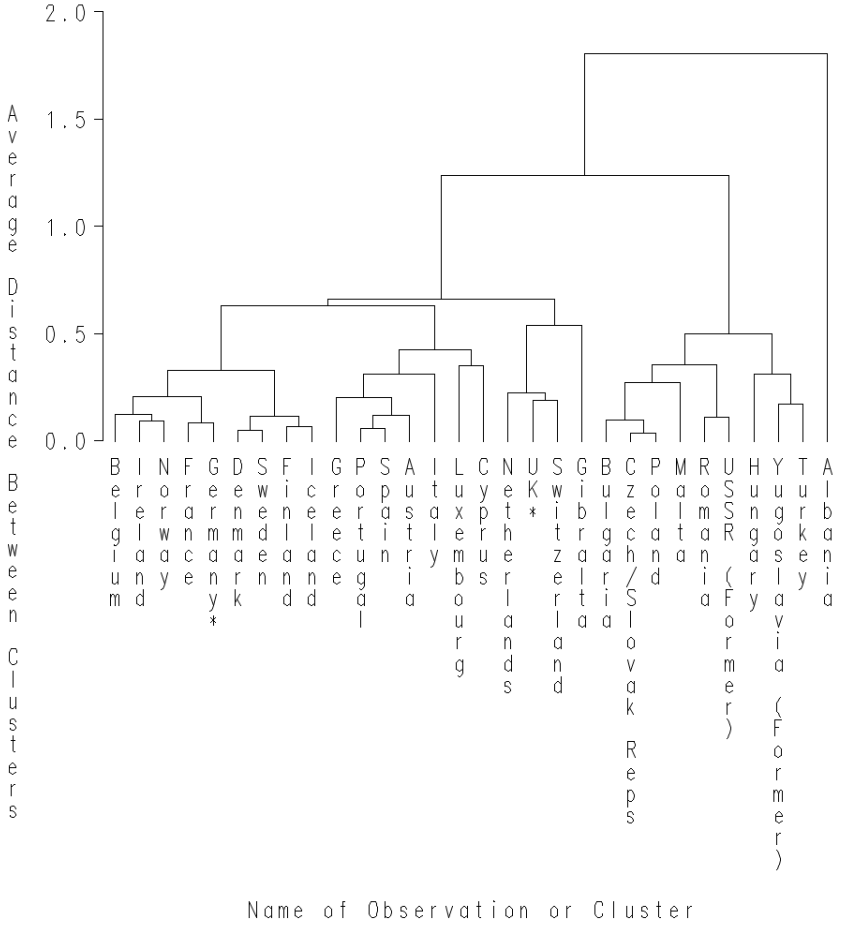
```
copy fin ser;
run; quit;
ods graphics off;
```

Comparing the two different cluster results, we can see that the main difference is whether or not to classify a small group of countries (Netherlands, UK, Switzerland, Gibraltar) as part of a larger EU/EFTA cluster, or whether these countries remain distinct.

Four clusters:



Three clusters:



In the three cluster model, the countries listed above are included with the largely Western European countries. In both cases, one of the clusters represents many of the Eastern European clusters, and one of the clusters represents Albania alone, which was identified as a potential outlier earlier in the analysis.

We will use this macro to make tables displaying the assignment of the observations to the determined clusters.

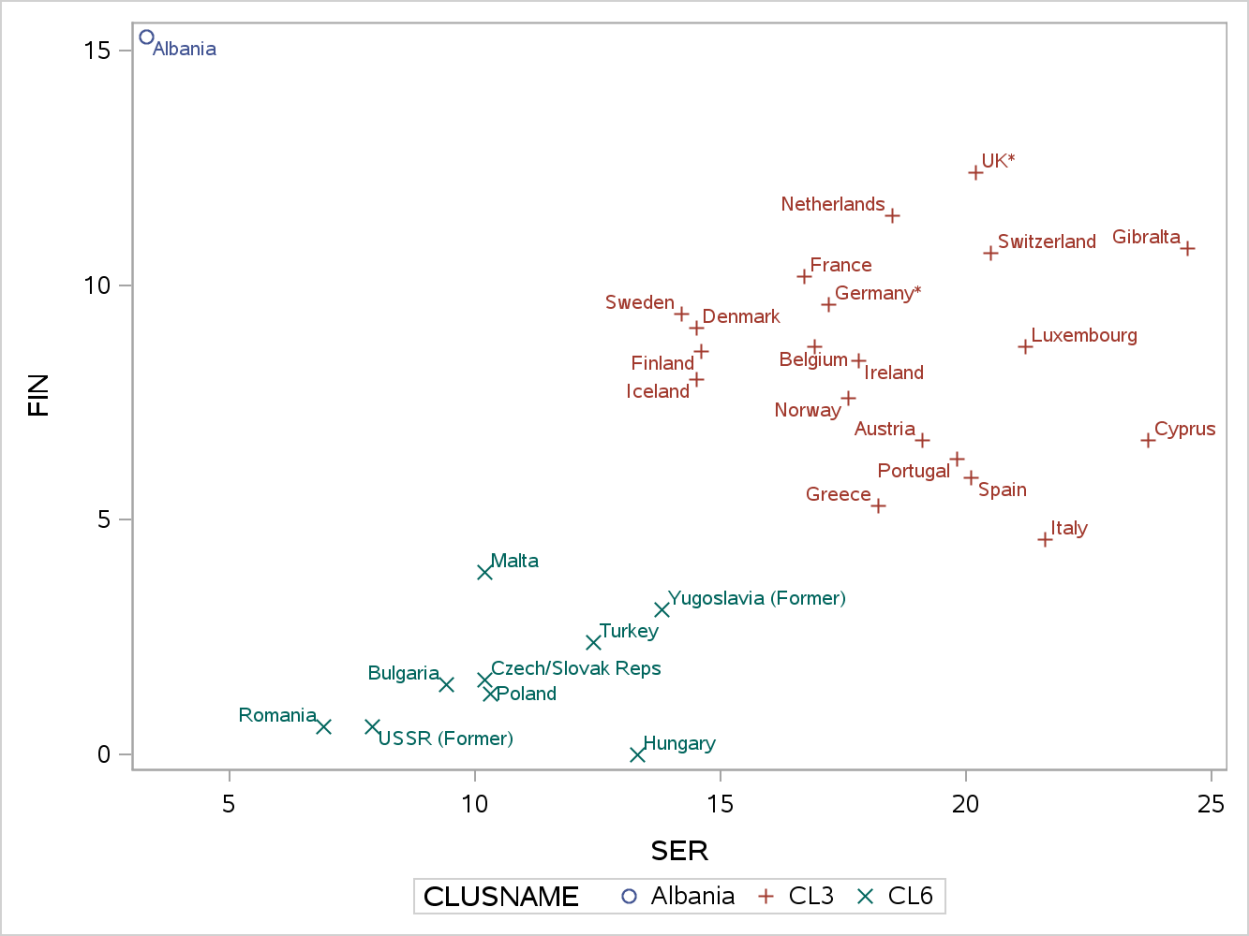
```
%macro makeTable(treeout,group,outdata);
data tree_data;
      set &treeout.(rename=(_name_=country));
run;

proc sort data=tree_data; by country; run; quit;

data group_affiliation;
      set &group.(keep=group country);
run;

proc sort data=group_affiliation; by country; run; quit;
```

```
data &outdata.;
      merge tree_data group_affiliation;
      by country;
run;

proc freq data=&outdata.;
table group*clusname / nopercent norow nocol;
run;
%mend makeTable;


* Call macro function;
%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);

* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_3_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=fin x=ser / datalabel=country group=clusname;
run; quit;
ods graphics off;


%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);

* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_4_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=fin x=ser / datalabel=country group=clusname;
run; quit;
ods graphics off;
```

Display the tables and comment on these results. Did the members of each membership group get clustered into the same cluster? Which number of clusters do you prefer?

The three cluster option is displayed in the following table (and visually in the FIN*SER scatterplot).

Table of GROUP by CLUSNAME				
GROUP	CLUSNAME			
Frequency	Albania	CL3	CL6	Total
EFTA	0	6	0	6
EU	0	12	0	12
Eastern	1	0	7	8
Other	0	2	2	4
Total	1	20	9	30

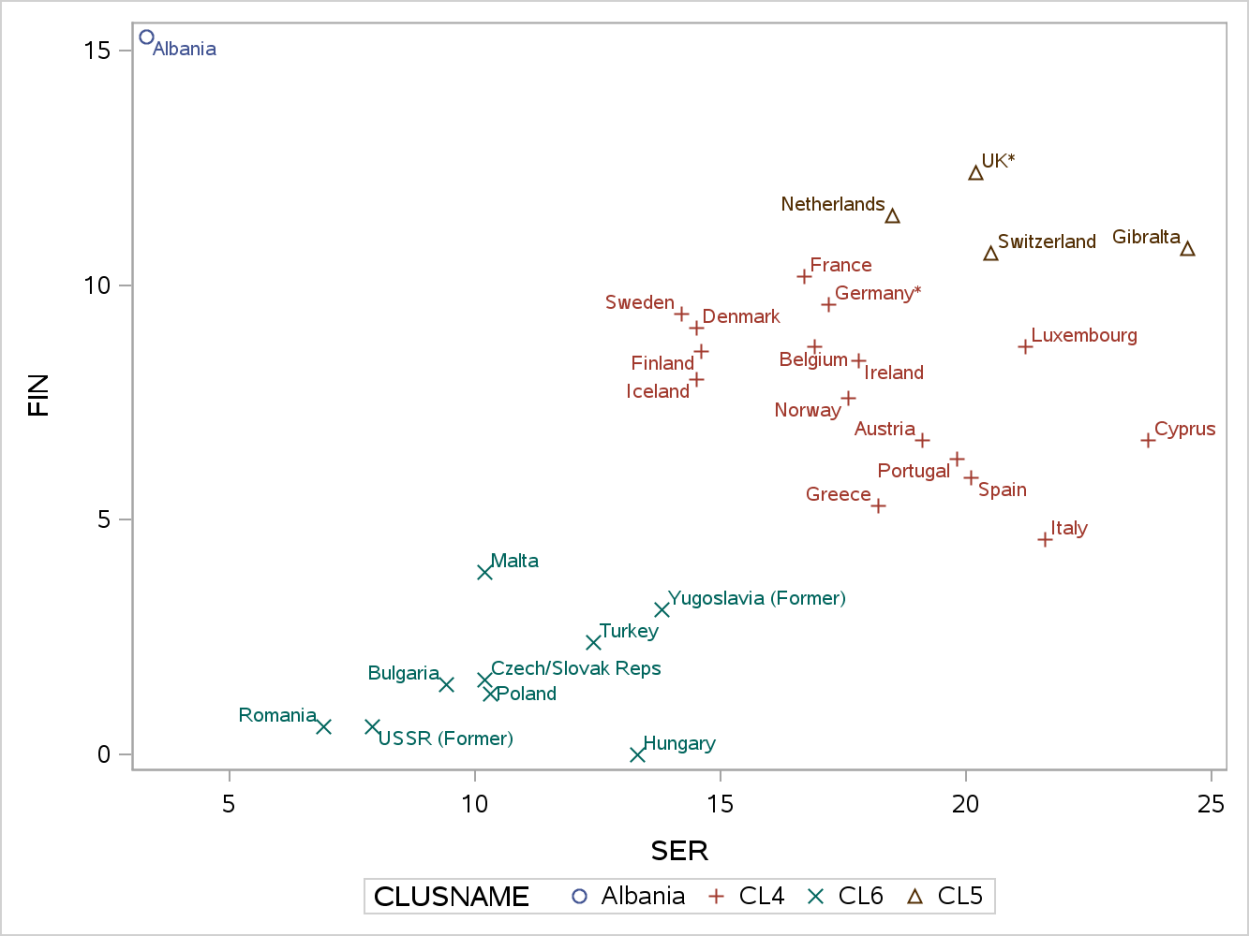


This cluster pattern shows a clear split (identified earlier through the scatterplot matrix) between the Eastern European countries in the middle/left bottom of the plot and the EU/EFTA countries in the middle/right top. Albania, on its own in the extreme top left of the plot, makes up the third and final cluster.

The table is particularly instructive, because it shows that members of the same country groups where clustered together (a good sign). With the exception of Albania, all members of the Eastern European group were classified as members of CL6 (along with 2 Other group member countries). Additionally, all of the EU and EFTA member countries were classified together in CL3.

The four cluster group option is below.

GROUP	CLUSNAME				
Frequency	**Albania**	**CL4**	**CL5**	**CL6**	**Total**
EFTA	0	5	1	0	6
EU	0	10	2	0	12
Eastern	1	0	0	7	8
Other	0	1	1	2	4
Total	1	16	4	9	30



Here we see that two EU, one EFTA and one other country were classified in a new cluster—CL5. With those exceptions, the other clusters remain the same as they were in the three-cluster option.

Because the four-cluster option pulls countries from a variety of member groups for the fourth cluster, I prefer the three-cluster grouping. It cleanly and distinctly divides the Eastern and Western European countries, with Albania on its own. This seems to represent the data better, and also stays true to the original member groupings.

Now perform a similar cluster analysis using the following cluster commands. Which of these four cluster analyses do you prefer?

```
*************************************************************;
* Using the first 2 principal components;
*************************************************************;
ods graphics on;
proc cluster data=pca_9components method=average outtree=tree3 pseudo ccc
plots=all;
var prin1 prin2;
id country;
run; quit;
ods graphics off;
```

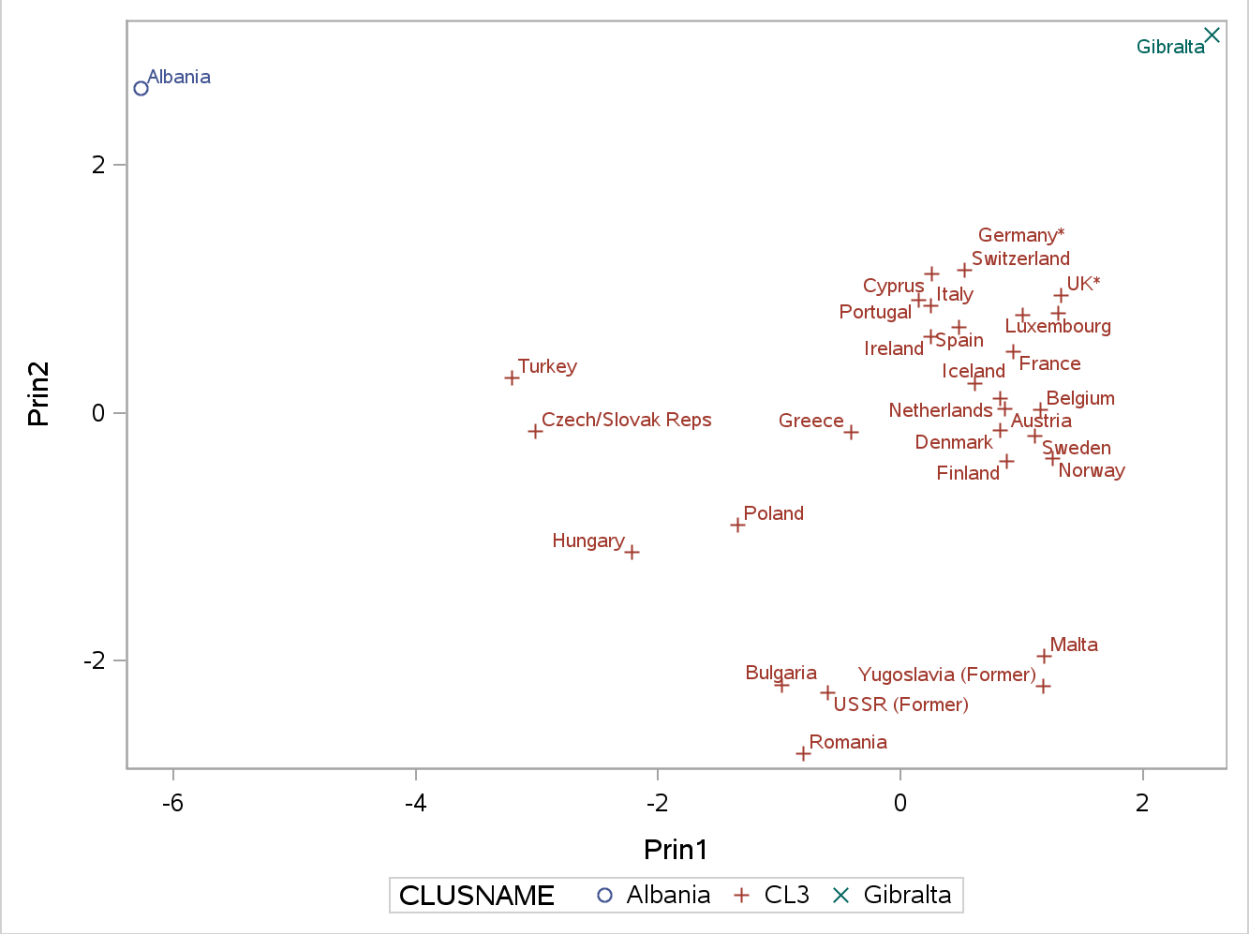
```
ods graphics on;
proc tree data=tree3 ncl=4 out=_4_clusters;
copy prin1 prin2;
run; quit;

proc tree data=tree3 ncl=3 out=_3_clusters;
copy prin1 prin2;
run; quit;
ods graphics off;

%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);
%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);

* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_3_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=prin2 x=prin1 / datalabel=country group=clusname;
run; quit;
ods graphics off;


* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_4_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=prin2 x=prin1 / datalabel=country group=clusname;
run; quit;
ods graphics off;


*********************************************************************;
* Using the first 2 factor components from ULS with VARIMAX rotation;
*********************************************************************;
ods graphics on;
proc cluster data=uls_varimax method=average outtree=tree4 pseudo ccc
plots=all;
var factor1 factor2;
id country;
run; quit;
ods graphics off;

ods graphics on;
proc tree data=tree4 ncl=4 out=_4_clusters;
copy factor1 factor2;
run; quit;

proc tree data=tree4 ncl=3 out=_3_clusters;
copy factor1 factor2;
run; quit;
ods graphics off;

%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);
%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);
```

```
* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_3_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=factor2 x=factor1 / datalabel=country group=clusname;
run; quit;
ods graphics off;

* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_4_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=factor2 x=factor1 / datalabel=country group=clusname;
run; quit;
ods graphics off;
```

The tables and scatterplots for the principle components analysis-based clusters are reproduced below.

Three clusters:

Table of GROUP by CLUSNAME				
GROUP	CLUSNAME			
Frequency	Albania	CL3	Gibralta	Total
EFTA	0	6	0	6
EU	0	12	0	12
Eastern	1	7	0	8
Other	0	3	1	4
Total	1	28	1	30

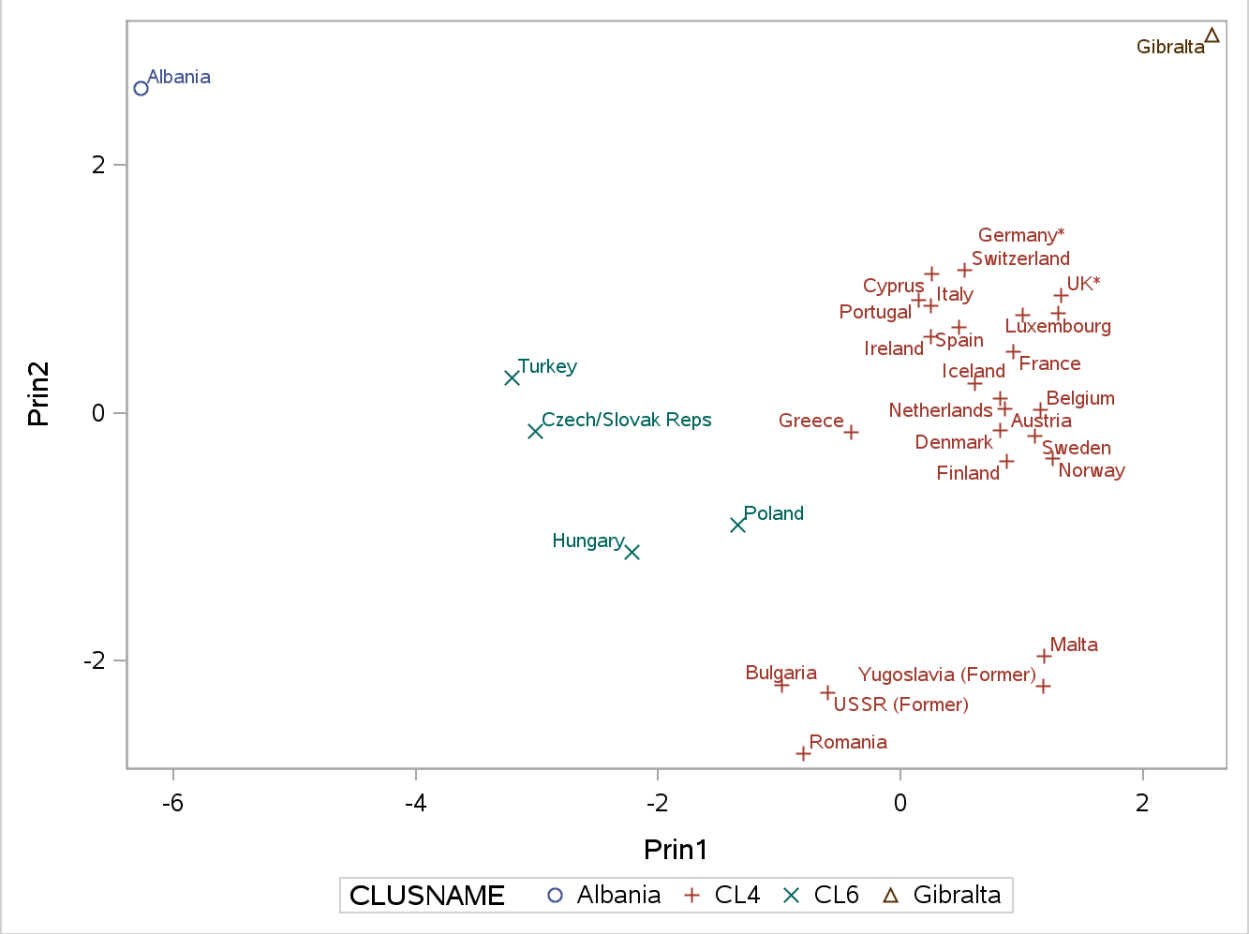


The three cluster analysis based on PCA takes a different approach than the hierarchical algorithm clusters. In addition to Albania—once again alone as its own cluster—the PCA cluster adds Gibraltar as its own cluster (top right of the plot). All of the other countries are grouped together.

When a fourth cluster is introduced, the following split occurs.

Four clusters:

Table of GROUP by CLUSNAME					
GROUP	CLUSNAME				
Frequency	Albania	CL4	CL6	Gibralta	Total
EFTA	0	6	0	0	6
EU	0	12	0	0	12
Eastern	1	4	3	0	8
Other	0	2	1	1	4
Total	1	24	4	1	30



The additional cluster, CL6, pulls three Eastern European countries along with Turkey together. This clustering starts to look like it is taking into account some of the differences we know about the countries into account (such as group membership). The scatterplot looks as if the clustering is incomplete, however. The remaining four Eastern European countries in the data set that are not in CL6 are grouped closely together with Malta, an Other country, in the bottom-right of the plot. It looks as if five clusters might be most appropriate (or three clusters, throwing out Albania and Gibraltar as outliers and removing them from the data set).

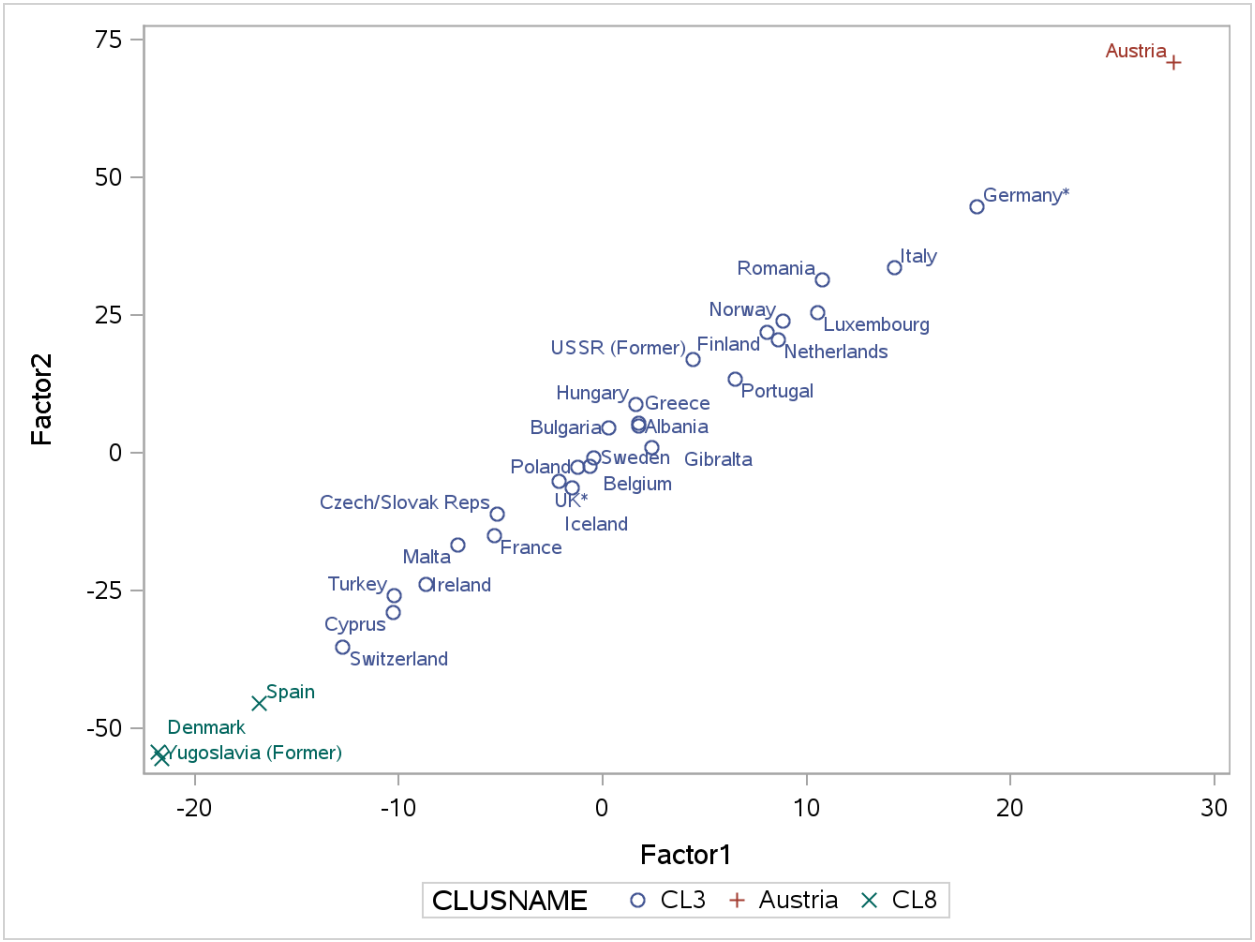
Despite these limitations, I would prefer this cluster analysis over the three-cluster PCA-based analysis because of the further distinction between the mass of countries in the middle-right of the plot.

I would still prefer the three-cluster hierarchical analysis over either of the PCA-based analysis, however, with the possible exception of the five-cluster version of the PCA analysis (described, but not created).

The tables and scatterplots for the factor analysis-based clusters are reproduced below.

Three clusters:

Table of GROUP by CLUSNAME				
GROUP	**CLUSNAME**			
Frequency	**Austria**	**CL3**	**CL8**	**Total**
EFTA	1	5	0	6
EU	0	10	2	12
Eastern	0	7	1	8
Other	0	4	0	4
Total	1	26	3	30



The FA-based three-cluster analysis lacks the obvious geographic/cultural divide present in the two hierarchical cluster analyses or the four-cluster PCA-based analysis. The CL8 cluster isolates two EU

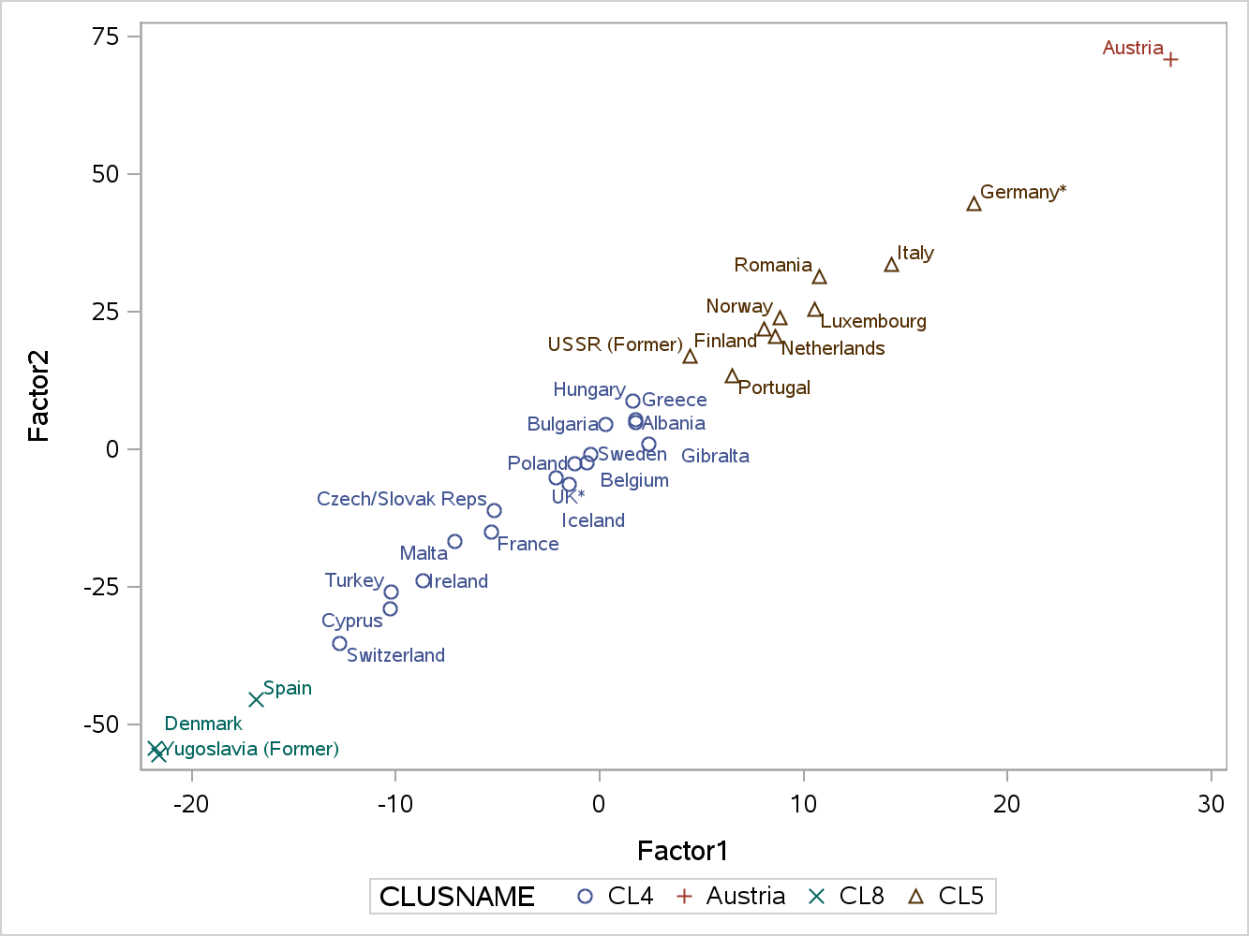
countries and one Eastern European country. Additionally, a very large CL3 cluster has a mix of EFTA, EU, Eastern and Other countries. Austria makes up the final cluster by itself.

It's quite possible that this clustering arrangement hints at an underlying difference between the countries in each cluster. However, the relationship is not obvious. With the exception of putting all four "Other" countries together, these clusters do split up the countries in the traditional member groups. Each cluster is a mix of countries from different areas, cultures and economies. For this reason, I would still prefer the three-cluster analysis produced through the hierarchical algorithm.

The FA-based four-cluster analysis is below.

Four clusters:

Table of GROUP by CLUSNAME					
GROUP	**CLUSNAME**				
Frequency	**Austria**	**CL4**	**CL5**	**CL8**	**Total**
EFTA	1	3	2	0	6
EU	0	5	5	2	12
Eastern	0	5	2	1	8
Other	0	4	0	0	4
Total	1	17	9	3	30



The FA-based four-cluster analysis also splits the countries up regardless of their member group orientation again. Spain, Denmark and Yugoslavia are grouped once again (in CL8). The main difference is the addition of a fourth cluster that groups a variety of countries from Eastern and Western Europe together as an additional cluster. Because of the lack of obvious connections between the countries clustered together, I would still prefer the hierarchical three-cluster model.

Forced to choose from the four additional cluster analyses, I would choose the four-cluster PCS-based analysis because of the way it separates the countries while keeping most aspects of group membership in tact as a defining trait.

Assignment Document:

As mentioned in the beginning we will not be using our typical assignment format. You will be given a Word document of the assignment, and you will write your answers directly into the document near the questions in green. As always the document should be submitted in pdf format.