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?

Based on the analysis of the scatterplot, 3 combinations looked interesting. I initially chose the combination of SPS and AGR, since its correlation was highest. However, I found that SER and FIN actually provided a better cluster segmentation and provided a better "guess" for placing the 4 "other" countries into groups.

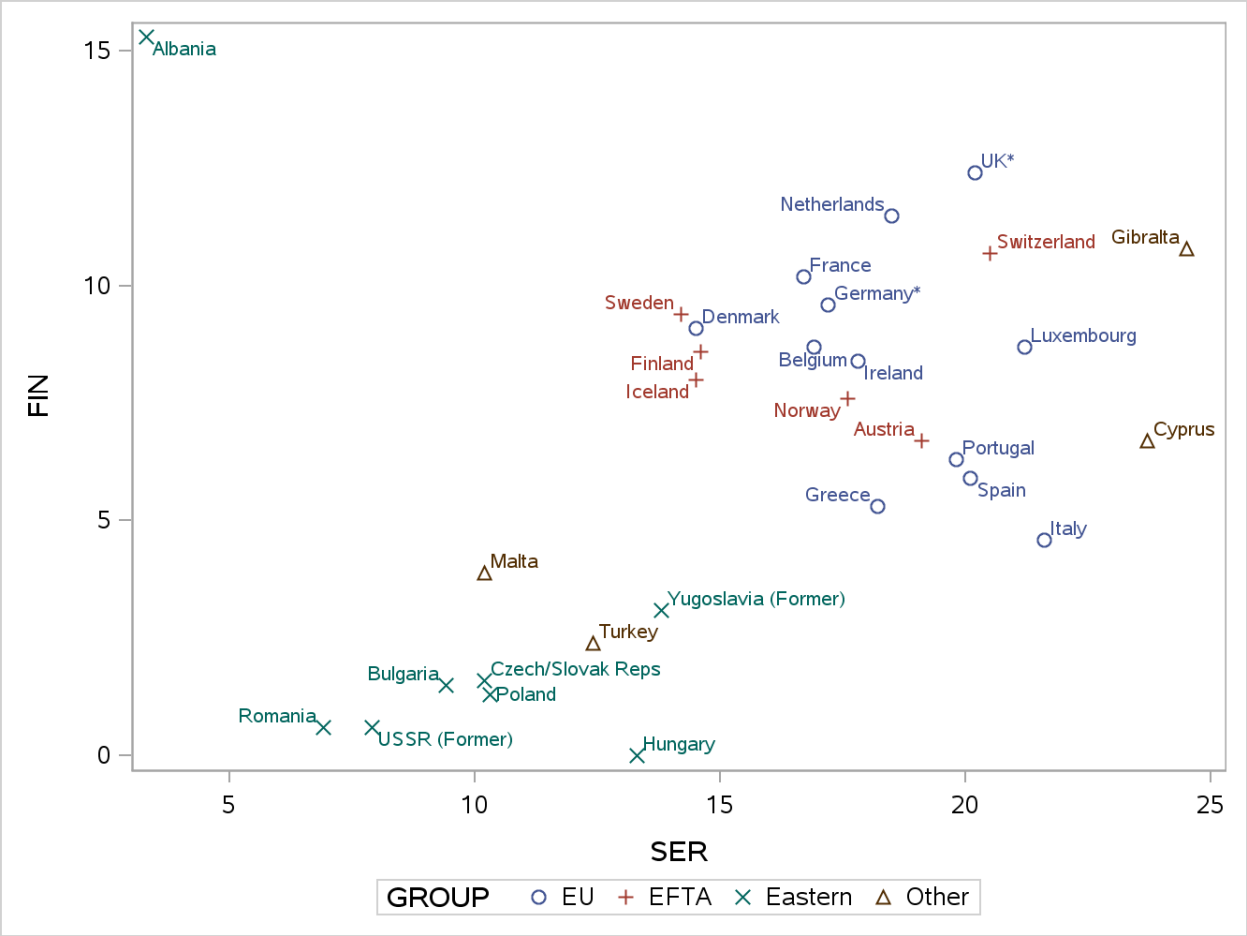
Pearson Correlation Coefficients, N = 30 Prob > |r| under H0: Rho=0									
	AGR	**MIN**	**MAN**	**PS**	**CON**	**SER**	**FIN**	**SPS**	**TC**
AGR	1.00000	0.31607 0.0888	-0.25439 0.1749	-0.38236 0.0370	-0.34861 0.0590	-0.60471 0.0004	-0.17575 0.3529	-0.81148 <.0001	-0.48733 0.0063
MIN	0.31607 0.0888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

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=FIN x=SER / 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.

Based on the scatterplot provided below, I would place Malta in Eastern, Turkey in Eastern, Cyprus in EU, and Gibralta in EU. (The only country that appears truly out of place is Switzerland.)



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.

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;
```

Based on a threshold of ensuring that 80% of the variance is represented, I would accept 4 components, since 83.12% of the variance would then be explained.

I could also choose the number of components based on those with Eigenvalues > 1. Using this approach, I would also select the first 4 components.

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

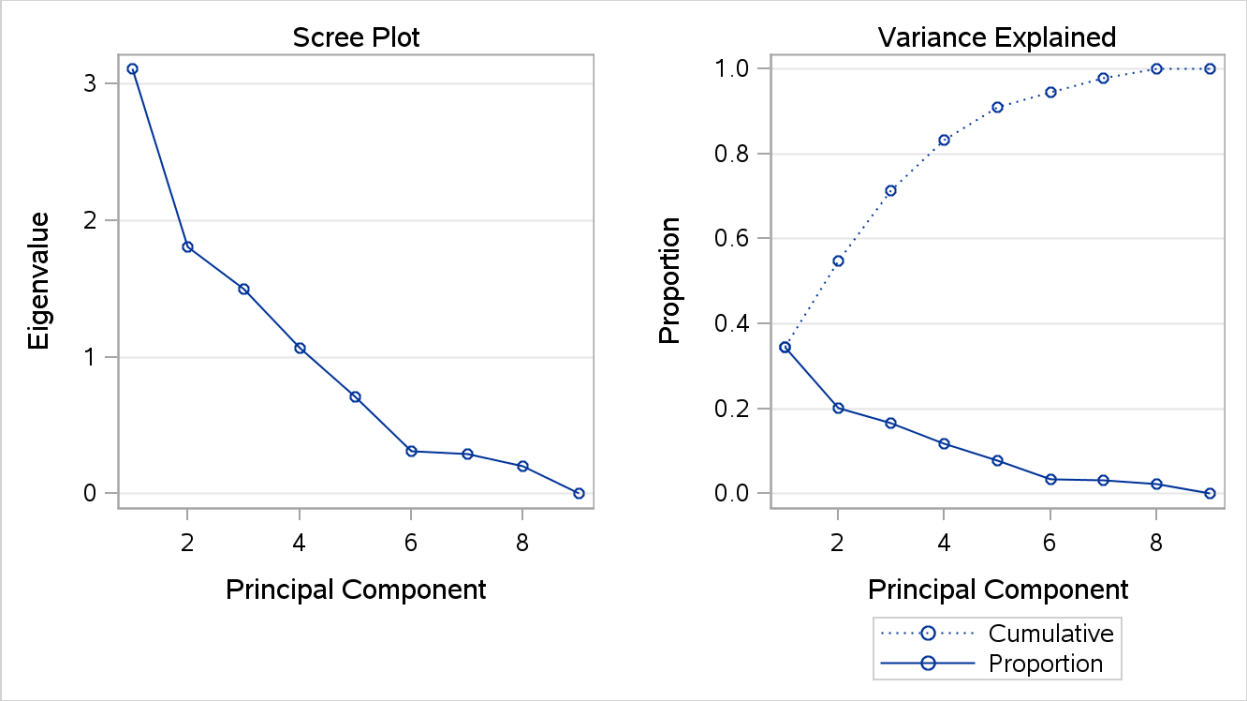
The Eigenvectors allow me to interpret the principal components by examining the "loadings". Based on these loadings, PC1 has relatively high values for SPS and SER and Low values for AGR and MIN. PC2 has High values for SER and FIN and Low Values for TC and MAN. PC4 appears to have the largest separation, in which CON is a very high 0.78 and FIN is a very low -.49, relative to the other loadings.

Eigenvectors

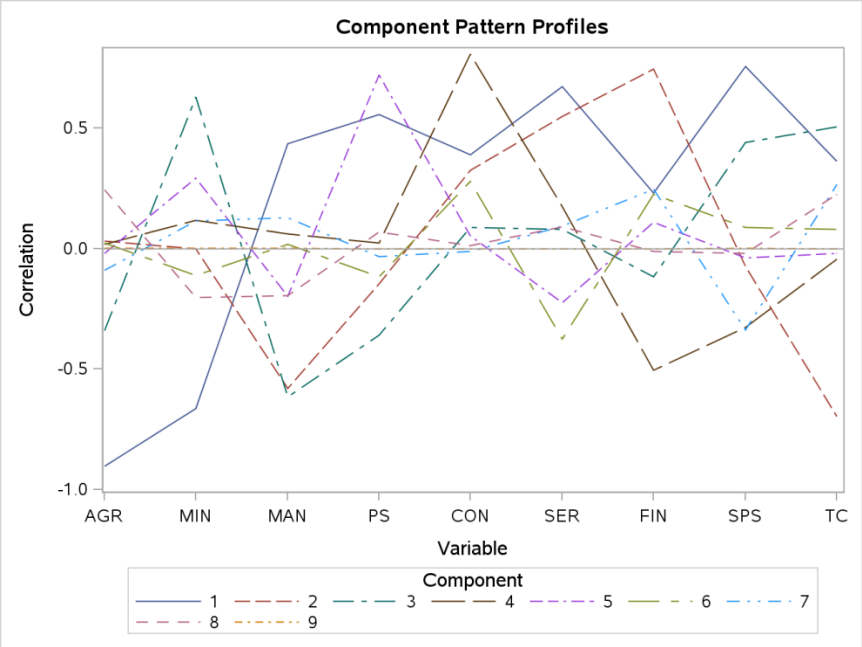
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
AGR	-.511492	0.023475	-.278591	0.016492	-.024038	0.042397	-.163574	0.540409	0.582036
MIN	-.374983	-.000491	0.515052	0.113606	0.346313	-.198574	0.212590	-.448592	0.418818
MAN	0.246161	-.431752	-.502056	0.058270	-.233622	0.030917	0.236015	-.431757	0.447086
PS	0.316120	-.109144	-.293695	0.023245	0.854448	-.206471	-.060565	0.155122	0.030251
CON	0.221599	0.242471	0.071531	0.782666	0.062151	0.502636	-.020285	0.030823	0.128656
SER	0.381536	0.408256	0.065149	0.169038	-.266673	-.672694	0.174839	0.201753	0.245021
FIN	0.131088	0.552939	-.095654	-.489218	0.131288	0.405935	0.457645	-.027264	0.190758
SPS	0.428162	-.054706	0.360159	-.317243	-.045718	0.158453	-.621330	-.041476	0.410315
TC	0.205071	-.516650	0.412996	-.042063	-.022901	0.141898	0.492145	0.502124	0.060743

The Scree Plot would allow me to determine the number of components to accept by viewing where the line bends or flattens. Based on this rule, I would accept the first 6 components.

The second graph shows the scree plot along with the cumulative variance accounted for by the components. As shown, the first 4 components explain over 80% of the variance.



In the Component Pattern Profiles graph, each variable is plotted as an observation whose coordinates are correlations between the variable and the two corresponding components on the plot. Since I ran this for all 9 components, it is a difficult graph to interpret.



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 Component Analysis and Factor Analysis are not the same statistical method. Factor Analysis is based on statistical methods and assumptions, whereas PCA is based on linear algebra. Factor analysis postulates a model for the data – PCA does not. Factor analysis tries to explain the covariances of the observed variables, whereas PCA is primarily concerned with explaining the variance in the observed variables. In PCA, if the number of components is increased, the components remain unchanged. In factor analysis, though, there could be substantial changes.

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:

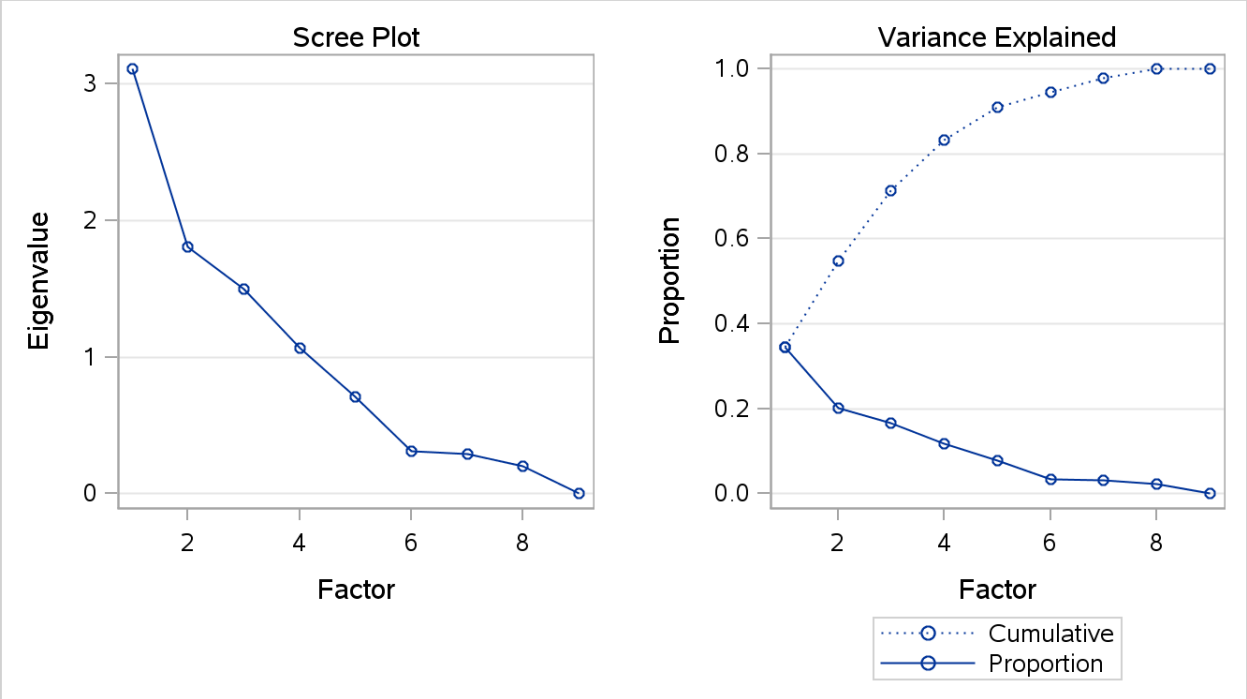
In addition to using PROC PRINCOMP to perform a principal components analysis SAS will allow you to perform a principal components analysis using PROC FACTOR. Run this code and compare the output from PROC FACTOR to the output from PROC PRINCOMP.

```
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;
```

These tables created by PROC FACTOR seem to match the results produced by PRINCOMP. The Eigenvalues match exactly.

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

The scree plot and the variance explained graph also match the output of the PRINCOMP. (This makes sense, since the eigenvalues matched.)



The factor pattern table below matches the component pattern graph displayed earlier with the PRINCOMP procedure.

Factor Pattern									
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9
AGR	-0.90235	0.03158	-0.34077	0.01701	-0.02026	0.02366	-0.08861	0.24397	0.00155
MIN	-0.66153	-0.00066	0.63001	0.11715	0.29186	-0.11080	0.11516	-0.20252	0.00112
MAN	0.43427	-0.58074	-0.61412	0.06009	-0.19689	0.01725	0.12785	-0.19492	0.00119
PS	0.55769	-0.14681	-0.35925	0.02397	0.72010	-0.11521	-0.03281	0.07003	0.00008
CON	0.39094	0.32614	0.08750	0.80711	0.05238	0.28046	-0.01099	0.01392	0.00034
SER	0.67309	0.54914	0.07969	0.17432	-0.22474	-0.37535	0.09471	0.09108	0.00065
FIN	0.23126	0.74375	-0.11700	-0.50450	0.11065	0.22650	0.24790	-0.01231	0.00051
SPS	0.75535	-0.07358	0.44055	-0.32715	-0.03853	0.08841	-0.33656	-0.01872	0.00109
TC	0.36178	-0.69493	0.50518	-0.04338	-0.01930	0.07918	0.26659	0.22669	0.00016

The scoring coefficients are normalized to give Principal Component Scores with unit variance, wherease PRINCOMP produces component scores with variance equal to the corresponding eigenvalue.

Standardized Scoring Coefficients									
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9
AGR	-0.2899352	0.01745251	-0.2277561	0.01599266	-0.0285226	0.0759832	-0.3019739	1.19702447	218.469776
MIN	-0.2125564	-0.0003648	0.42106913	0.11016531	0.41092306	-0.3558825	0.39246238	-0.9936465	157.205162
MAN	0.13953465	-0.3209864	-0.4104446	0.05650521	-0.2772078	0.05540933	0.43570729	-0.9563572	167.815802
PS	0.17919035	-0.0811435	-0.2401037	0.02254142	1.01385982	-0.3700338	-0.1118089	0.34360137	11.3549322
CON	0.12561144	0.1802652	0.05847831	0.75896054	0.07374625	0.9008172	-0.0374475	0.068275	48.2914946
SER	0.21627065	0.30351818	0.0532614	0.16391795	-0.3164255	-1.2055929	0.32276959	0.44688931	91.9695736
FIN	0.07430646	0.41108292	-0.0782001	-0.4744002	0.15578188	0.72751139	0.84485715	-0.0603896	71.6018867
SPS	0.24270018	-0.0406711	0.29444003	-0.3076338	-0.0542477	0.28397703	-1.1470359	-0.0918699	154.013441
TC	0.11624271	-0.3841038	0.33763521	-0.0407893	-0.0271733	0.25430786	0.90854769	1.11222071	22.8002052

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?

The factor analysis is invalid, because the communality exceeds 1, which indicates that some unique factor has negative variance – so something must be wrong. It is most likely that the cause is too many common factors.

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;
```

It is interesting that the cumulative proportion exceeds 1 after 2 components. This is due to the negative eigenvalues for components 6-9. Based on this, it seems appropriate for us to keep the first 2 factors.

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 factor pattern provides the "loadings", which aid interpretation of the factors. Factor 1 has high positive values for SER and SPS and High Negative Values for AGR and MIN. This suggests that Factor1 seems to measure the differences between those countries that rely on the grounds resources vs. those that are primarily service-driven. Factor 2 seems to measure differences between countries heavy in financial services vs. those in transportation and communications.

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

This table simply provides the proportion of variance explained by each of the 2 factors.

Variance Explained by Each Factor	
Factor1	Factor2
2.7128416	1.3407424

The final communality estimates table below lists the proportion of variance of the variables accounted for by the common factors.

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

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.

The factor analysis is invalid, because the communality exceeds 1, which indicates that some unique factor has negative variance – so something must be wrong. It is most likely that the cause is too many common factors.

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?

This is a valid factor analysis.

The eigenvalues resulting from ULS are nearly identical to those of the Iterated Principal Factor Analysis. Again, it appears that factors 1 and 2 would be most appropriate.

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
7	-.16113699	0.14231528	-0.0397	1.2366
8	-.30345226	0.35276659	-0.0748	1.1618
9	-.65621886		-0.1618	1.0000

The "loadings" are nearly identical to the IPFA method, and I would interpret them in the same manner.

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 variance explained by each factor table and the Final Communality Estimates are also very similar to the results of the Iterated Principal Factor Analysis.

Variance Explained by Each Factor	
Factor1	Factor2
2.7128575	1.3428644

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

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?

If factors are rotated by an orthogonal factor rotation, the factors are also uncorrelated. If factors are rotated by an oblique factor rotation, the factors become correlated.

Oblique rotations often produce more useful patterns. However, since the factors are then correlated, there is no single unambiguous measure of the importance of a factor in explaining the variable.

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?

The eigenvalues listed below are the same as we saw with ULS.

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

Orthogonal Transformation Matrix		
	1	2
1	0.70986	0.70435
2	0.70435	-0.70986

When analyzing the factor loadings, the rotation has possibly added to the interpretability of Factor 2, though it somewhat lessened the interpretability of Factor 1.

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

Rotated Factor Pattern		
	Factor1	Factor2
AGR	-0.62747	-0.75212
MIN	-0.46385	-0.26066
MAN	0.03633	0.41145
PS	0.28379	0.31692
CON	0.36956	0.06866
SER	0.81896	0.09598
FIN	0.54068	-0.26732
SPS	0.41389	0.59842
TC	-0.26748	0.82199

The amount of variance explained by each factor was increased for Factor 2, but it was decreased for Factor 1. I believe this somewhat supports the observations of the rotated Factor Pattern as well.

Variance Explained by Each Factor	
Factor1	Factor2
2.7128559	1.3428471

Variance Explained by Each Factor (ROTATED)	
Factor1	Factor2
2.0331911	2.0225118

The final communality estimates remained unchanged by the rotation. This makes sense, because the process used an orthogonal factor rotation.

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

Final Communality Estimates: Total = 4.055703 (ROTATED)								
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?

The Eigenvalues are the same as in ULS. However, the "Normalized Oblique Transformation Matrix" holds completely different values for its transformation matrix than did the Orthogonal Matrix.

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

Normalized Oblique Transformation Matrix

	1	2
1	0.64586	0.63123
2	0.79798	-0.80961

Using a threshold of 0.50 for the rotated factor pattern, Factor 2 is improved vs. the unrotated values. However, I do not necessarily find the output easy to interpret.

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

Rotated Factor Pattern (Standardized Regression Coefficients)		
	Factor1	Factor2
AGR	-0.55645	-0.69000
MIN	-0.44429	-0.20903
MAN	-0.00882	0.41495
PS	0.25427	0.28837
CON	0.36949	0.02484
SER	0.82506	-0.00221
FIN	0.58160	-0.33853
SPS	0.35563	0.55941
TC	-0.36466	0.87054

The factor structure produces nearly the same information as the rotated factor pattern, though the factor structure is helpful in that it essentially provides correlation coefficients for the variables.

Factor Structure (Correlations)		
	Factor1	Factor2
AGR	-0.71251	-0.81585
MIN	-0.49157	-0.30951
MAN	0.08503	0.41296
PS	0.31949	0.34588
CON	0.37511	0.10841
SER	0.82456	0.18440
FIN	0.50503	-0.20699
SPS	0.48216	0.63985
TC	-0.16777	0.78806

As shown by the "variance explained" tables, rotating the factors improved factor 2, though it decreased the explanation from factor 1. It interests me that the rotation essentially made the variance explanations equal to each other.

Variance Explained by Each Factor	
Factor1	Factor2
2.7128559	1.3428471

Variance Explained by Each Factor Eliminating Other Factors	
Factor1	Factor2
1.8850888	1.8608050

Variance Explained by Each Factor Ignoring Other Factors	
Factor1	Factor2
2.1948980	2.1706142

The communality estimates continue to be the same. This is because factor rotations only redistribute the variance explained by the factors. The total variance (communality) explained by the factors for any variable remains unchanged.

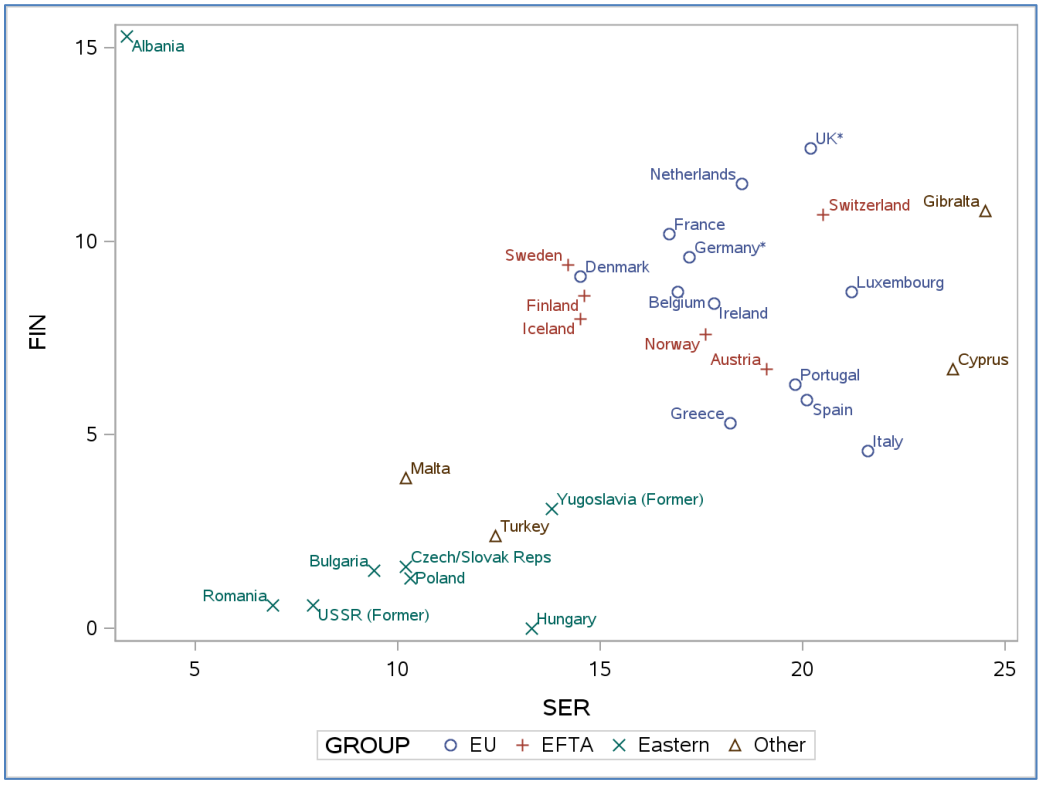
Final Communality Estimates: Total = 4.055703								
AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
0.95940387	0.28309868	0.17060768	0.18097518	0.14129083	0.67990578	0.36379989	0.52940794	0.74721313

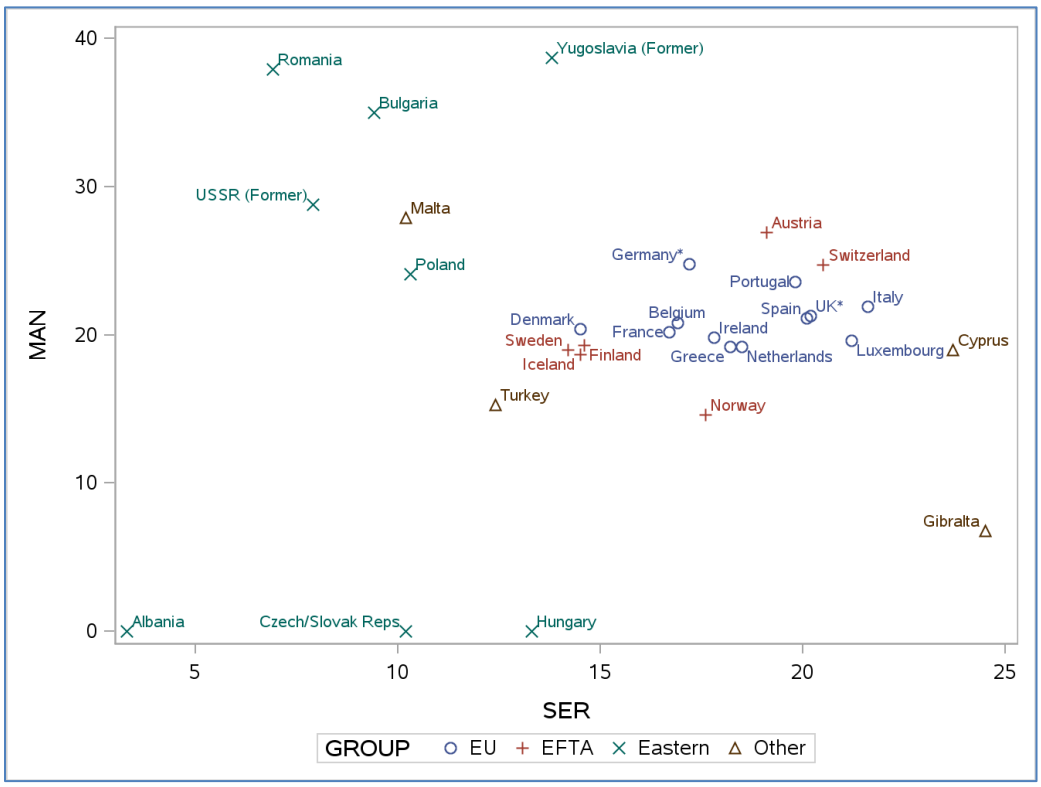
Final Communality Estimates: Total = 4.055703 (ROTATED)								
AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
0.95940387	0.28309868	0.17060768	0.18097518	0.14129083	0.67990578	0.36379989	0.52940794	0.74721313

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?

In the above graphs, I see 3 clusters in the FIN*SER graph, though I believe I see 3 or 4 clusters in the MAN*SER graph.

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, Pseudo F, and Pseudo T-Squared help us estimate the number of clusters we should use. We can interpret the measures of CCC by examining where the peaks appear in the plots.

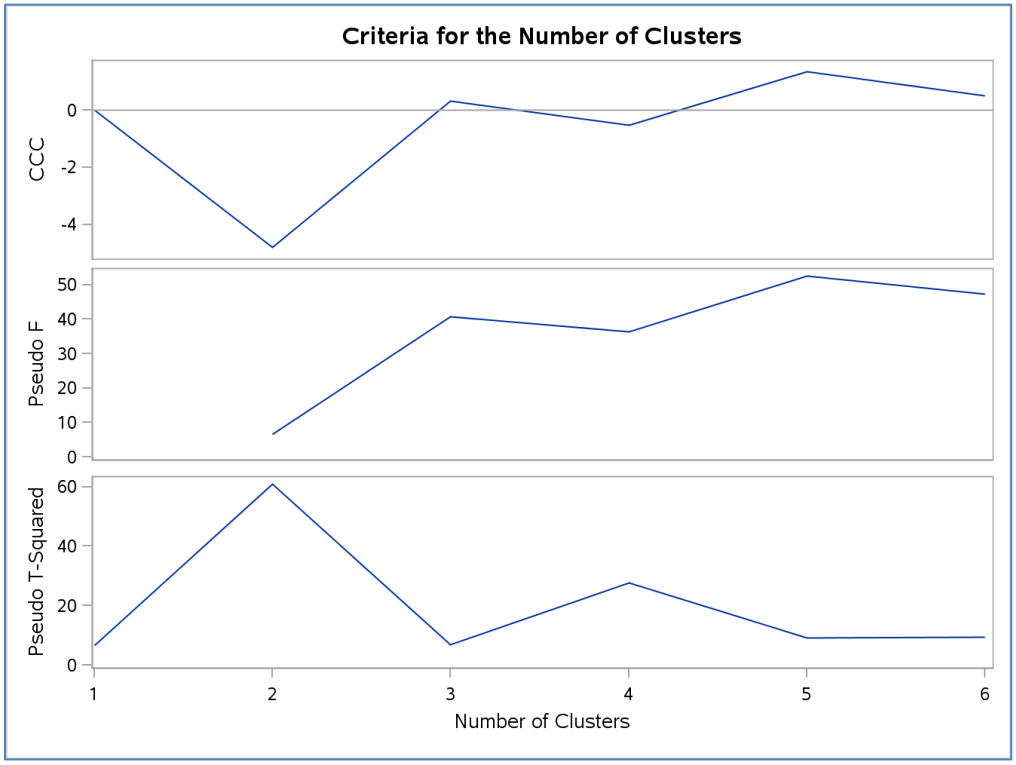
For CCC, peaks with values greater than 2 or 3 indicate good clusters, and peaks with values between 0 and 2 indicate possibly good clusters.

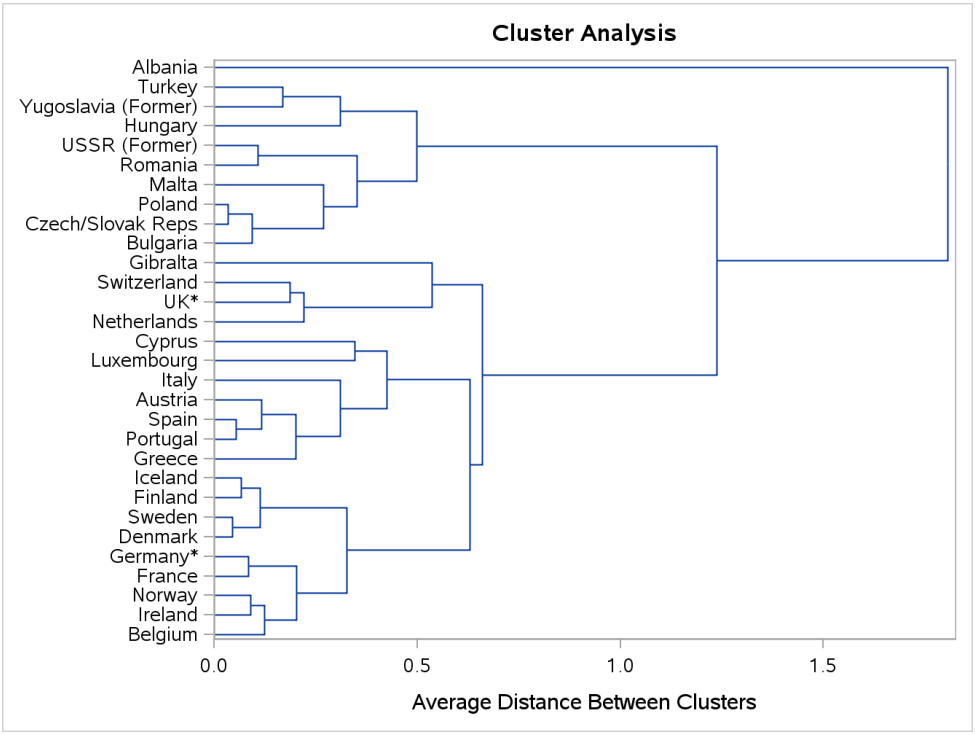
The Pseudo *F* statistic (PSF) is useful in that relatively large values indicate good numbers of clusters. Again, we can find potentially optimal clusters by finding peaks.

The pseudo t^2 statistic can be examined by finding spots in which the previous value is much larger than the next value. The number of clusters represented by the "next" value in the cluster history is a potentially "optimal" solution.

Cluster History											
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Norm RMS Distance	Tie
29	Czech/Slovak Reps	Poland	2	0.0000	1.00	.	.	881	.	0.0343	
28	Denmark	Sweden	2	0.0001	1.00	.	.	652	.	0.046	
27	Portugal	Spain	2	0.0001	1.00	.	.	537	.	0.0542	
26	Finland	Iceland	2	0.0002	1.00	.	.	438	.	0.066	
25	France	Germany*	2	0.0002	.999	.	.	340	.	0.0847	
24	Ireland	Norway	2	0.0003	.999	.	.	294	.	0.0894	
23	Bulgaria	CL29	3	0.0004	.999	.	.	248	9.7	0.0939	
22	Romania	USSR (Former)	2	0.0004	.998	.	.	226	.	0.1084	
21	CL28	CL26	4	0.0008	.998	.	.	183	6.9	0.1127	
20	CL27	Austria	3	0.0006	.997	.	.	173	5.8	0.116	
19	Belgium	CL24	3	0.0006	.996	.	.	167	2.2	0.1236	
18	Yugoslavia (Former)	Turkey	2	0.0010	.995	.	.	151	.	0.1697	
17	UK*	Switzerland	2	0.0012	.994	.	.	138	.	0.1872	
16	Greece	CL20	4	0.0019	.992	.	.	119	5.6	0.2009	
15	CL19	CL25	5	0.0029	.989	.	.	99.1	7.8	0.2037	
14	Netherlands	CL17	3	0.0019	.987	.	.	96.9	1.5	0.2214	
13	CL23	Malta	4	0.0036	.984	.	.	86.1	16.8	0.269	
12	CL16	Italy	5	0.0048	.979	.	.	76.4	5.5	0.3104	
11	Hungary	CL18	3	0.0041	.975	.	.	73.8	4.1	0.3108	
10	CL15	CL21	9	0.0141	.961	.	.	54.5	19.5	0.3272	
9	Luxembourg	Cyprus	2	0.0042	.957	.	.	58.0	.	0.3472	
8	CL13	CL22	6	0.0098	.947	.	.	56.0	8.8	0.3526	
7	CL12	CL9	7	0.0127	.934	.	.	54.4	5.5	0.4254	
6	CL8	CL11	9	0.0263	.908	.900	0.50	47.3	9.5	0.4996	
5	CL14	Gibralta	4	0.0141	.894	.869	1.35	52.6	9.2	0.5369	
4	CL10	CL7	16	0.0859	.808	.822	-.52	36.4	27.7	0.6305	
3	CL4	CL5	20	0.0564	.751	.741	0.31	40.8	6.9	0.6595	
2	CL3	CL6	29	0.5610	.190	.570	-4.8	6.6	60.9	1.2374	
1	CL2	Albania	30	0.1904	.000	.000	0.00	.	6.6	1.806	

Based on CCC plotted below, I may find optimal solutions with 3 or 5 clusters, since those are areas in which the graph peaks. Using the Pseudo F statistic, I would choose 3 or 5 clusters, also identifying peaks in those locations. Using the Pseudo T-Squared graph, I would also choose 3 or 5 clusters, since those counts follow the 2 peaks shown.





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;
```

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;
```

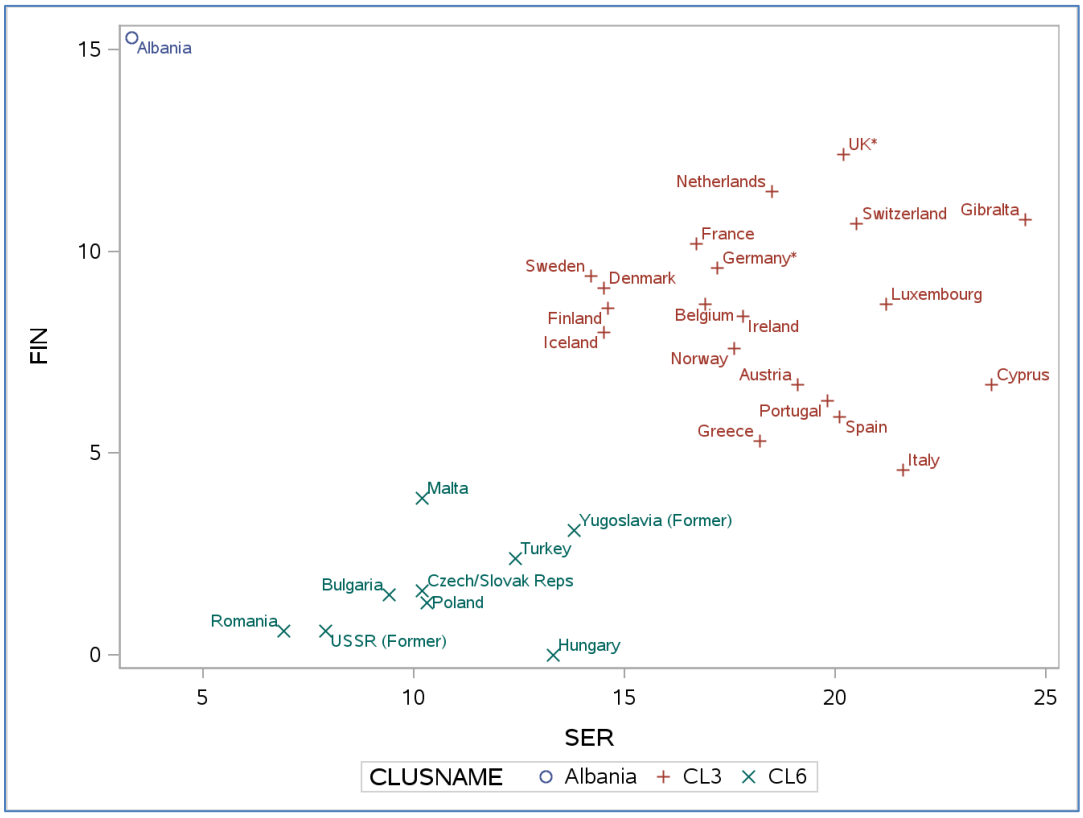
Display the tables and comment on these results. Did the members of each membership group get clustered into the same cluster? Which number of clusters do you prefer?

As shown in the following charts, the countries of the UK, Netherlands, Switzerland, and Gibralta were clustered separately, when we requested 4 clusters rather than 3.

I personally prefer the 3 clusters, and I feel it provides a better analysis, since the 4 countries that were segmented when moving to 4 clusters were not necessarily far removed from cluster 3.

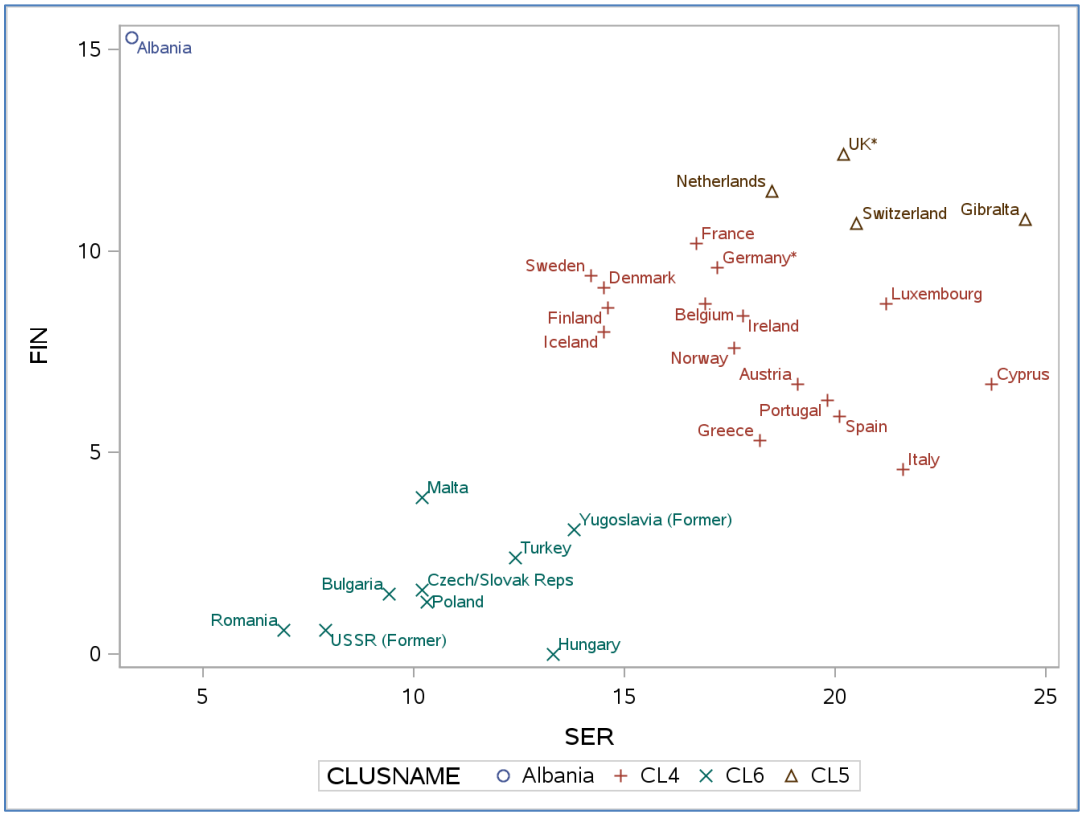
3 Clusters (using Macro)

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



4-Clusters (using Macro)

Table of GROUP by CLUSNAME					
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



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;
```

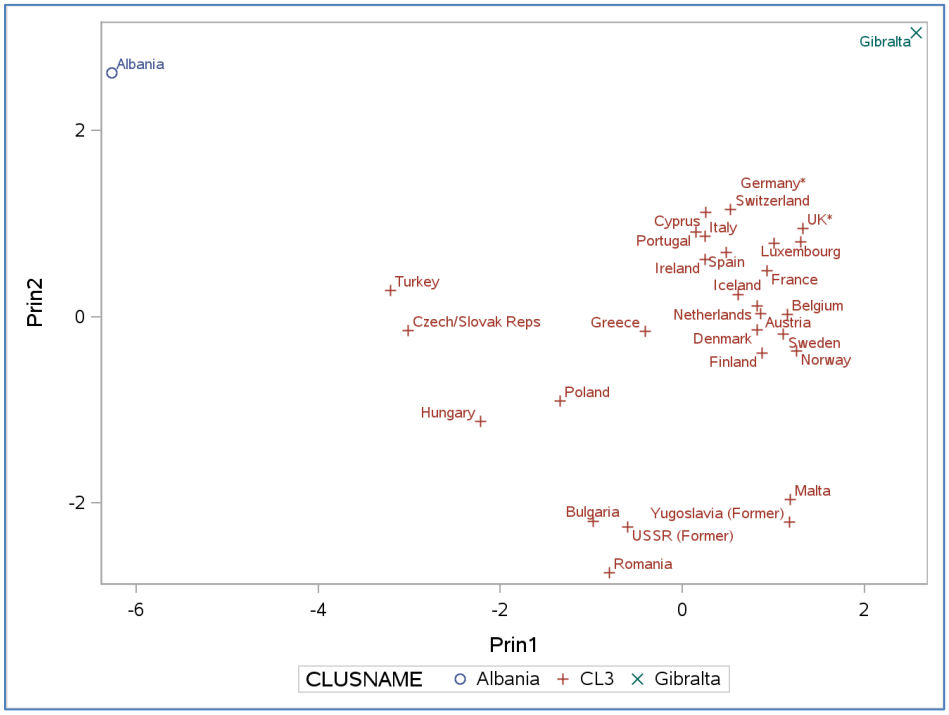
```
************************************************************************;
* Using the first 2 factor components from ULS with VARIMAX rotation;
************************************************************************;
ods graphics on;
proc cluster data=uls_varimax method=average outtree=tree4 pseudo 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;
```

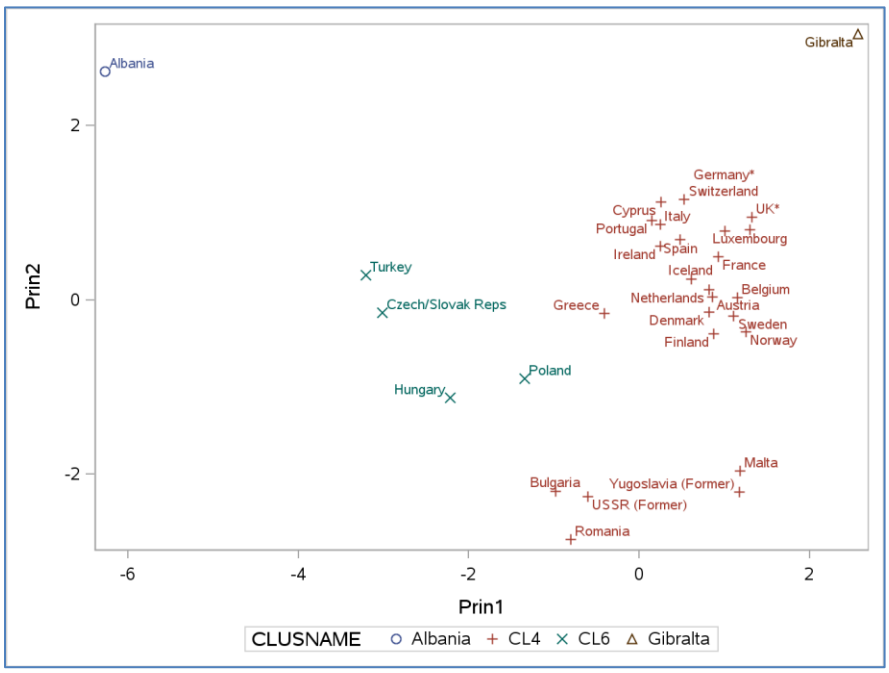
```
%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);
%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);

* 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;
```

When evaluating these 4 graphs, I prefer the final graph, which provides 4 clusters using ULS with varimax rotation. I feel this graph provides a very clear delineation of the countries.

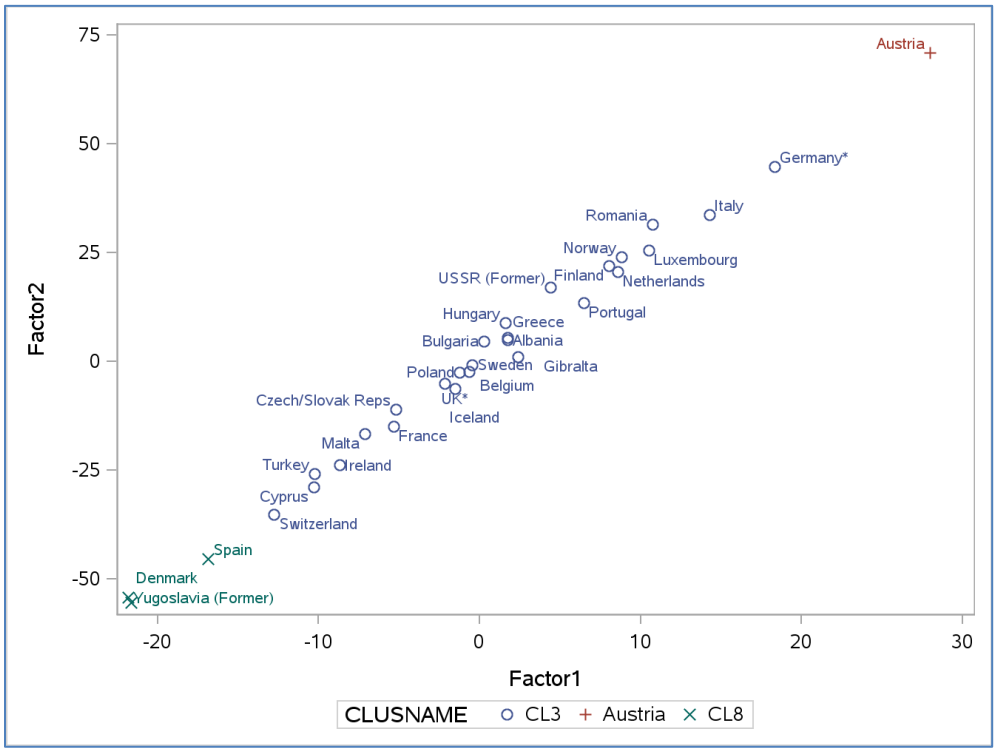
In fact, both "rotation" graphs do a great job of helping me visualize the clusters. Either solution seems like it would work well.

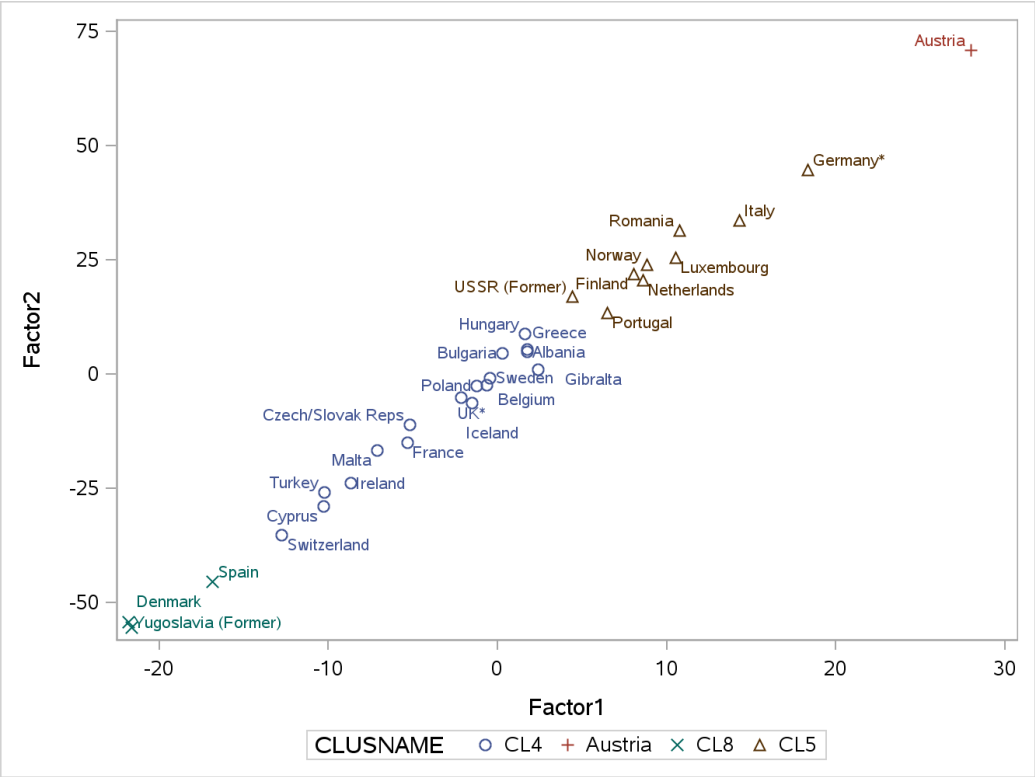
```
**************************************************************************;
* Using the first 2 principal components;
**************************************************************************;
```






```
*************************************************************************;
* Using the first 2 factor components from ULS with VARIMAX rotation;
*************************************************************************;
```





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.