Assignment #8: Multivariate Analysis (0 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.

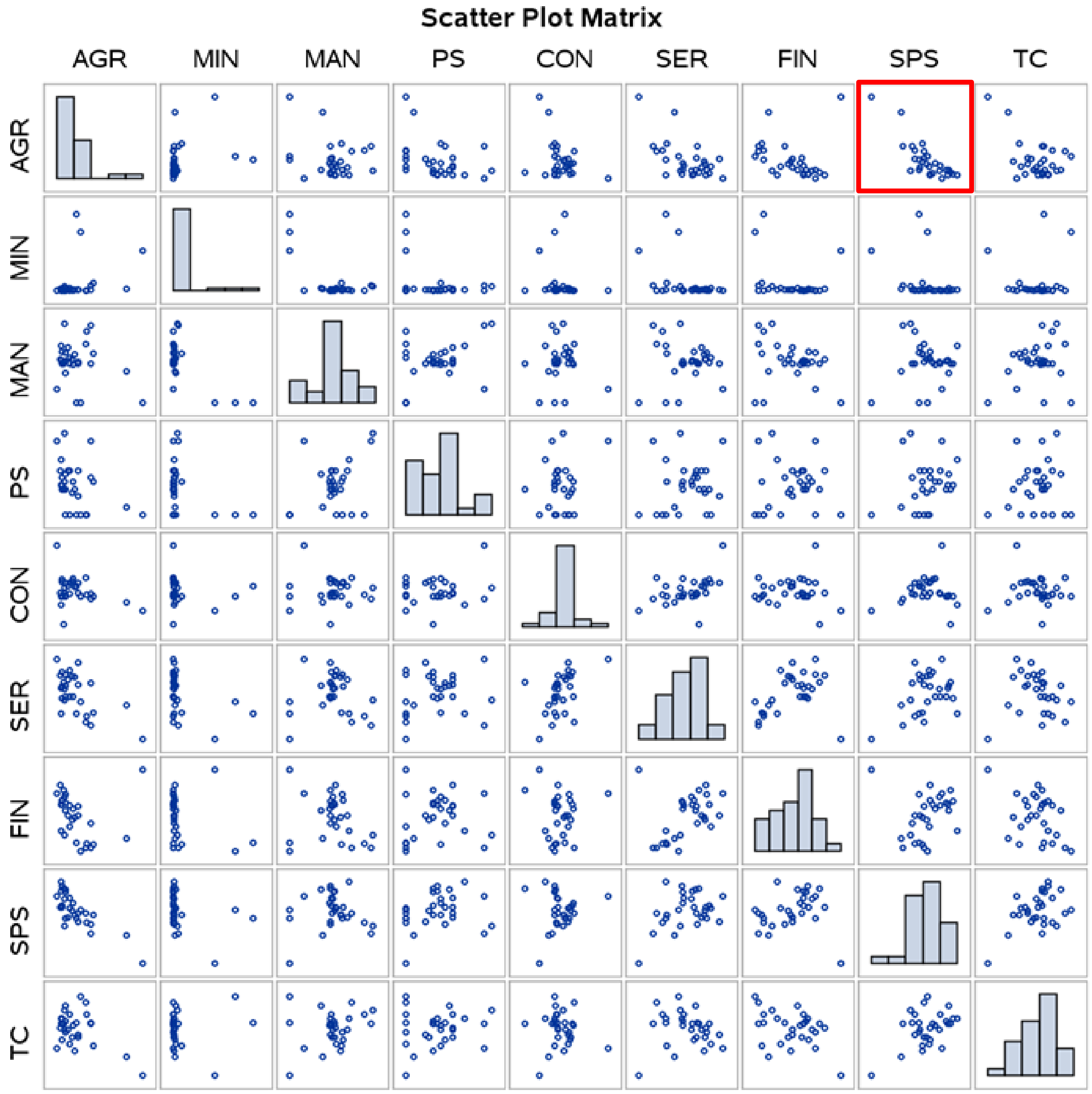
AGR:	agriculture
MIN:	mining
MAN:	manufacturing
PS:	power and water supply
CON:	construction
SER:	services
FIN:	finance
SPS:	social and personal services
TC:	transport and communications

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).

Variable AGR and SPS have a strong negative correlation (-0.8115) and the scatterplot shows the majority of the data in one cluster to the lower bottom and 2 points in the upper left. These will be used in the scatterplot.



```
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.

The four unassigned countries by viewing the scatterplot are Turkey, Cyprus, Gibralta, Malta. Based on the location of these points in relation to the 3 groups, I would assign these countries as follows:

Turkey = Eastern, Cyprus = EU, Gibralta = EU, Malta = EFTA

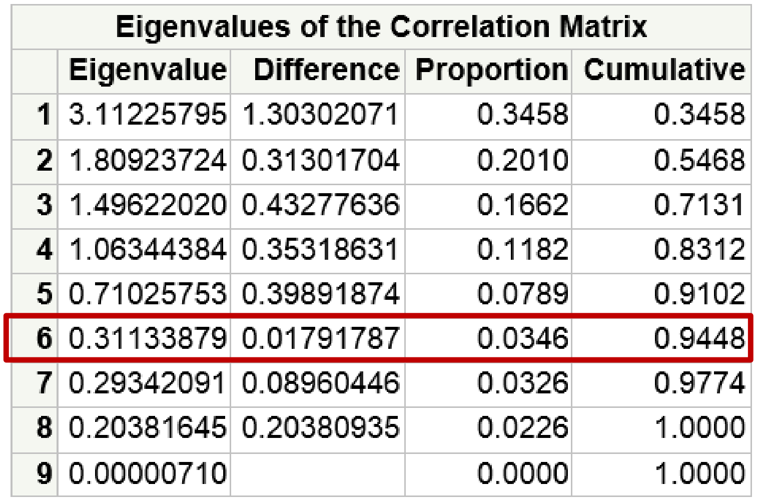
Note: In this assignment our observations are assigned to *classes* or are said to have *labels* (EU, EFTA, Eastern, or Other). Typically we use cluster analysis as an *unsupervised learner* (a situation with no response variable or label) and not as a *supervised learner* (a situation with a response 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

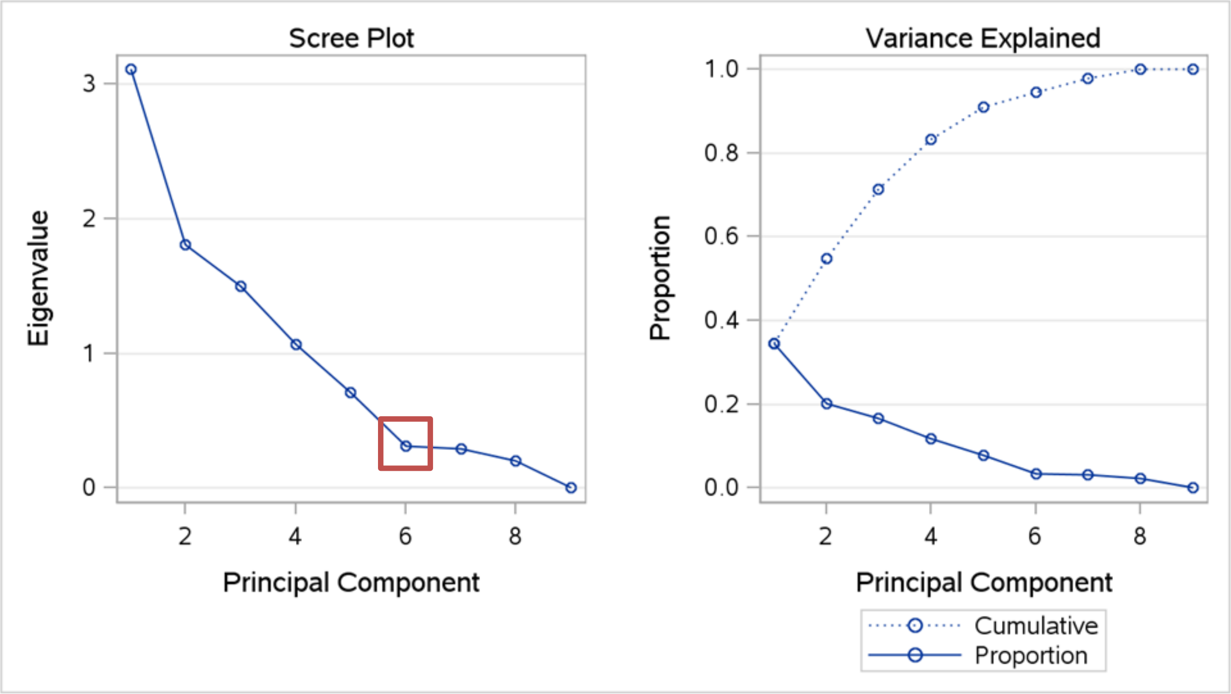
Our data set has nine variables. One method of reducing the dimensionality of our data set is to use principal components analysis. If we perform a principal components analysis, what would the resulting dimensionality be, i.e. how many components should we keep? What decision rule are you using to determine 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;
```

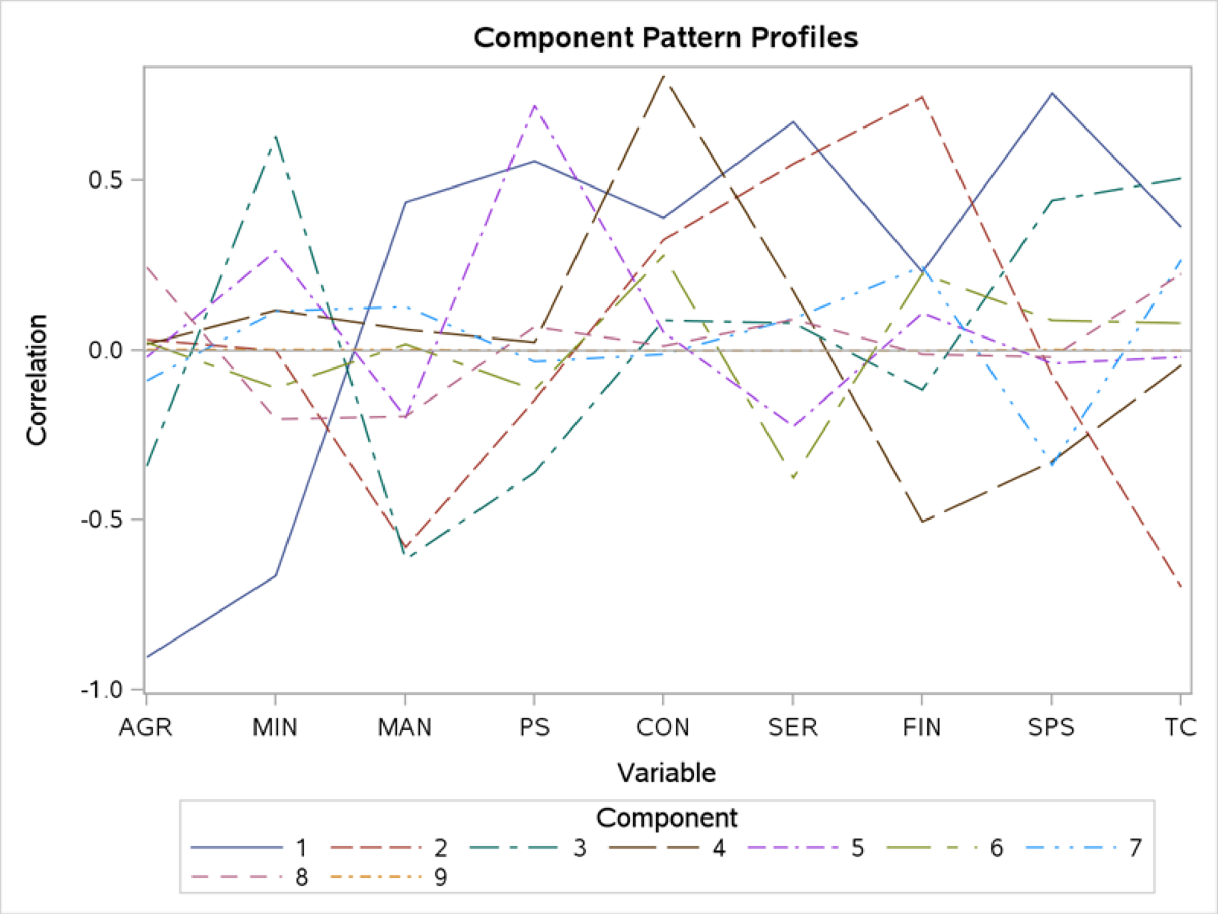
The SAS output is produced below. The decision rule to determine how many variables to keep is to determine where the curve on the scree plot flattens. Based on the scree plot I would use 6 principal components. The principal components are the eigenvalues below that describe the proportion of variation they explain. When 6 components are used and this would explain 94.5% of the variation.



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 component pattern profile shows how each of the principal components are correlated with each of the variables. For example PC1 is highly negatively correlated with AGR while highly correlated with SPS.



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?

PCA is based on numerical linear algebra while factor analysis is based on statistical assumptions with components estimated by maximum likelihood. PCA always produces orthogonal components (no correlation between components) while factor analysis does not. Independent people using PCA will reach the same repeatable conclusion for the number of components to retain while that is not the case for factor analysis.

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;
```

The SAS output using PROC FACTOR is identical to PROC PRINCOMP.

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.

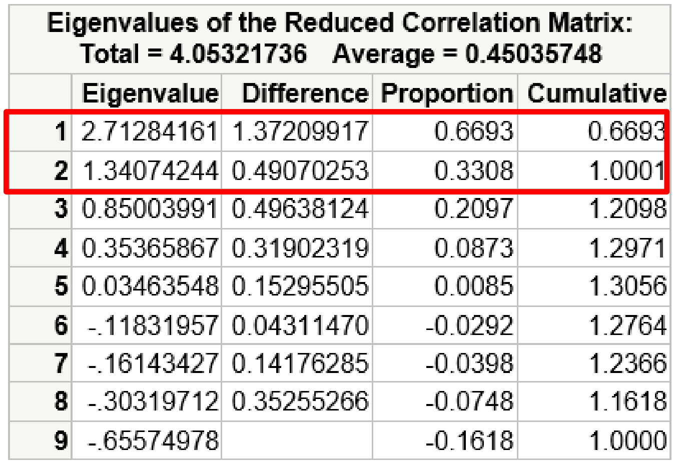
```
ods graphics on;
title Iterated Principal Factor Analysis using PROC FACTOR;
proc factor data=temp method=prinit out=pfa_factors
      nfactors=9 score plots=scree;
run;
ods graphics off;
```

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;
```

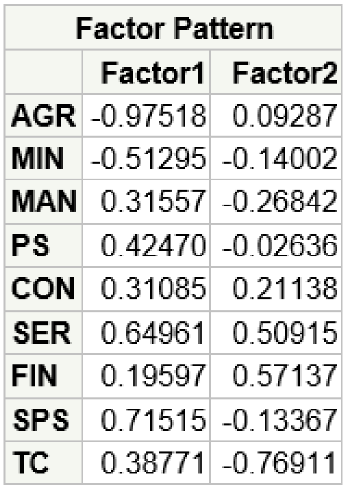
This was not a valid factor analysis at 9 factors as it a produced a warning that there were too many factors for a unique solution and an error that communality was greater than 1.0. Communality is defined as the variance shared with the other variables via common factors. Therefore a value greater than 1 is not a feasible solution. This is an example where the factor analysis model contains as many parameters as there are free parameters. If a communality exceeds 1, it is considered an "ultra-Heywood case (SAS 9.3 User Guide, page 2177). An ultra-Heywood case implies that some unique factor has negative variance, a clear indication that something is wrong (SAS User Guide).

The number of factors was reduced to two and the following SAS output was produced.



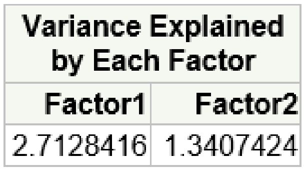
Eigenvalues of the Reduced Correlation Matrix: Total = 4.05321736 Average = 0.45035748				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.71284161	1.37209917	0.6693	0.6693
2	1.34074244	0.49070253	0.3308	1.0001
3	0.85003991	0.49638124	0.2097	1.2098
4	0.35365867	0.31902319	0.0873	1.2971
5	0.03463548	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 is often referred to as the factor loading matrix. Each loading (row) represents the correlation between the observed variable and each factor. For each AGR is almost perfectly negatively correlated with Factor1 and almost no correlation with Factor2.



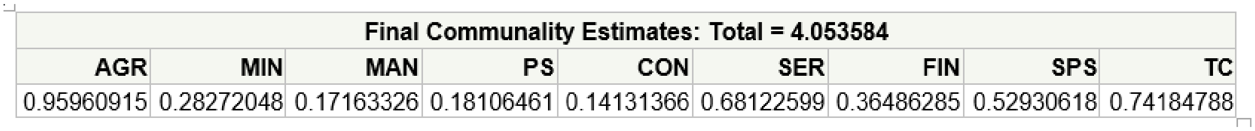
Factor Pattern		
	Factor1	Factor2
AGR	-0.97518	0.09287
MIN	-0.51295	-0.14002
MAN	0.31557	-0.26842
PS	0.42470	-0.02636
CON	0.31085	0.21138
SER	0.64961	0.50915
FIN	0.19597	0.57137
SPS	0.71515	-0.13367
TC	0.38771	-0.76911

The Variance Explained by Each Factor shows that the total variance is explained by the two components. The first and second component account for 2.7128 and 1.3407, respectively, of the total variance of 4.053584.



Variance Explained by Each Factor	
Factor1	Factor2
2.7128416	1.3407424

The Final Communality Estimates show how the observed variables are accounted for by the two components. The total communality of 4.053584 is the sum of the variance explained by the two components.



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

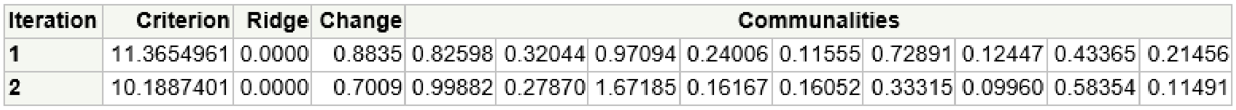
<u>Maximum Likelihood Factor Analysis:</u>

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

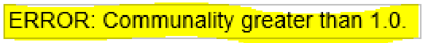
```
ods graphics on;
title Maximum Likelihood Factor Analysis using PROC FACTOR;
proc factor data=temp method=ml out=fa_ml
      outstat=fa_ml_stats mineigen=0 priors=smc nfactors=2 score ;
run;
ods graphics off;
```

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 MLFA produced an error that the Communality is greater than 1.0. The maximum likelihood method is especially susceptible to ultra-Heywood cases. During the iteration process, a variable with high communality is given high weight; this tends to increase its communality, which increases its weight and so on (SAS 9.3 User Guide, page 2166).



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



ERROR: Communality greater than 1.0.

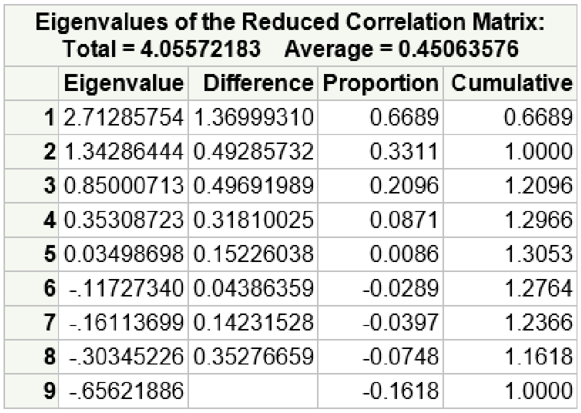
Unweighted Least Squares Factor Analysis:

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?

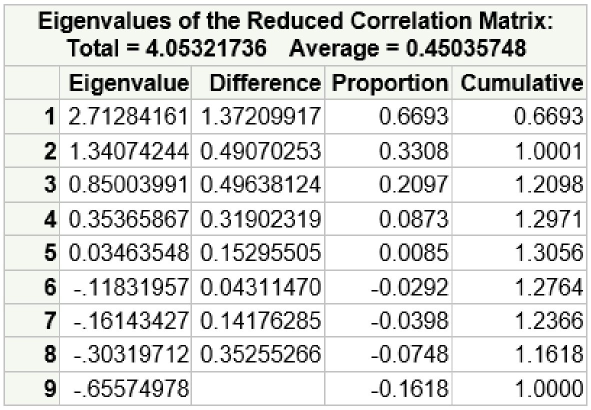
The following SAS output was generated. This analysis produced commonalities less than one and the convergence criterion was satisfied therefore it is a valid factor analysis. The factor weightings are very similar to the iterated principal factor analysis as shown below.



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

Unweighted Least Squares factor loadings

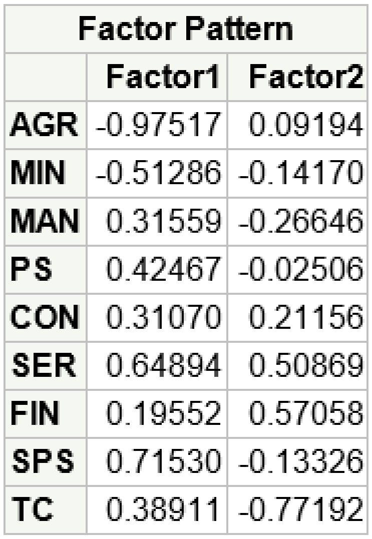


Eigenvalues of the Reduced Correlation Matrix: Total = 4.05321736 Average = 0.45035748

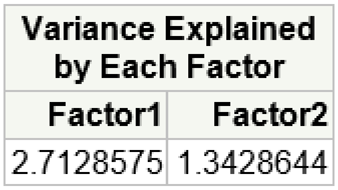
	Eigenvalue	Difference	Proportion	Cumulative
1	2.71284161	1.37209917	0.6693	0.6693
2	1.34074244	0.49070253	0.3308	1.0001
3	0.85003991	0.49638124	0.2097	1.2098
4	0.35365867	0.31902319	0.0873	1.2971
5	0.03463548	0.15295505	0.0085	1.3056
6	-.11831957	0.04311470	-0.0292	1.2764
7	-.16143427	0.14176285	-0.0398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

Iterated PFA loadings

The Factor Pattern table below describes how each observed variable is correlated with Factor1 and Factor2.

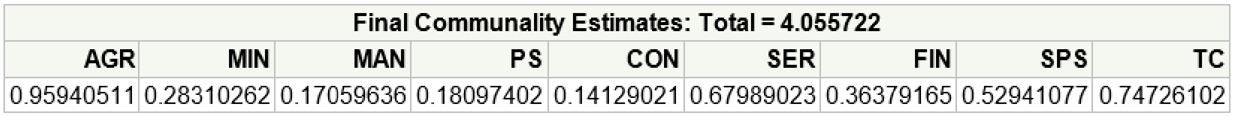


Factor Pattern		
	Factor1	Factor2
AGR	-0.97517	0.09194
MIN	-0.51286	-0.14170
MAN	0.31559	-0.26646
PS	0.42467	-0.02506
CON	0.31070	0.21156
SER	0.64894	0.50869
FIN	0.19552	0.57058
SPS	0.71530	-0.13326
TC	0.38911	-0.77192



Variance Explained by Each Factor	
Factor1	Factor2
2.7128575	1.3428644

The final communality estimates table below outlines the total variance (4.055722) and how each variable contributes to the variance.



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?

Unlike orthogonal factor rotation were the factors are uncorrelated, oblique rotations introduces correlations between the factors that may become simpler to interpret the factor loadings. Oblique rotations often produce more useful patterns than do orthogonal rotations (SAS 9.3 User Guide, page 2125). All rotations, orthogonal or oblique, are equally good statistically (SAS 9.3 User Guide). The preferred rotation is that which is most interpretable (SAS 9.3 User Guide). Obliquely rotated factor solution may be more desirable for those who believe common factor are seldom orthogonal (SAS 9.3 User Guide, page 2189).

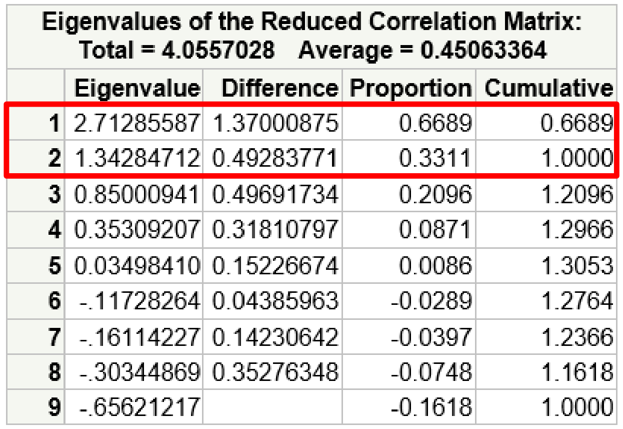
VARIMAX Factor Rotation

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?

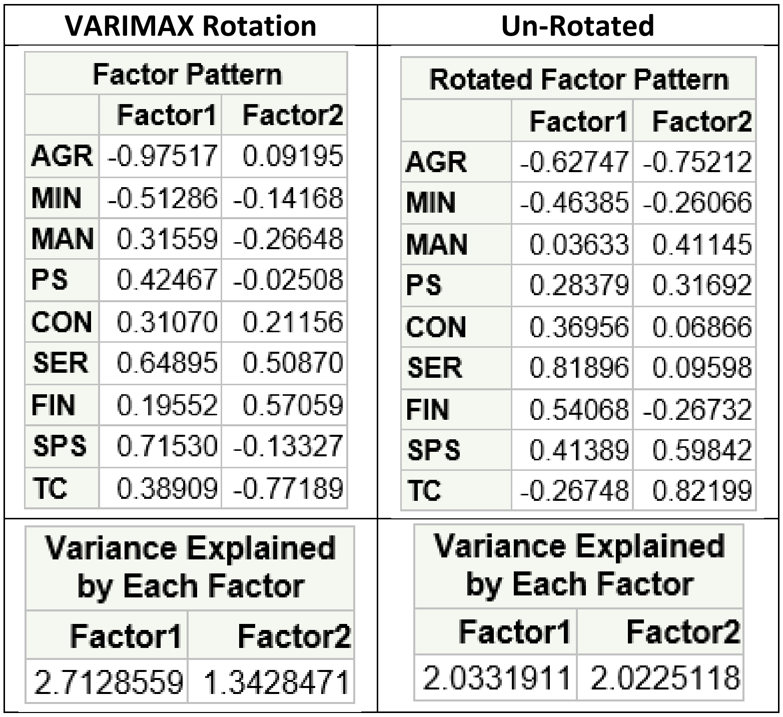
The VARIMAX rotation produced the following SAS output.



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 Varimax rotation did change the variance explained by each factor as shown below. There was approximately an even distribution of variance explained by the factors prior to rotation and after rotation Factor1 became the more dominant explaining the variation (67% vs. 50%). The rotation did alter the interpretation of the factor loadings. The objective of the Varimax rotation is to redistribute the loadings towards either +1 or -1, with fewer loadings in between. A good example is the AGR variable where the weighting for Factor1 changed from -0.62747 to -0.97517. Therefore, it can be deduced the Factor1 is almost perfectly negatively correlated with AGR. Another example is the SPS variable that changed from 0.41389 to 0.71530.



VARIMAX Rotation

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

Variance Explained by Each Factor

Factor1	Factor2
2.7128559	1.3428471

Un-Rotated

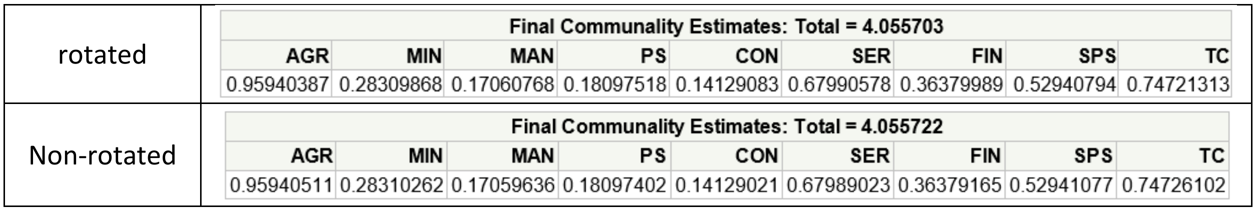
Rotated Factor Pattern

	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

Variance Explained by Each Factor

Factor1	Factor2
2.0331911	2.0225118

There was minimal changes in the final communality estimates when comparing the rotated and original analysis. Therefore the overall variance did not change, just the distribution of the variance that was explained by each factor.



rotated

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

Non-rotated

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

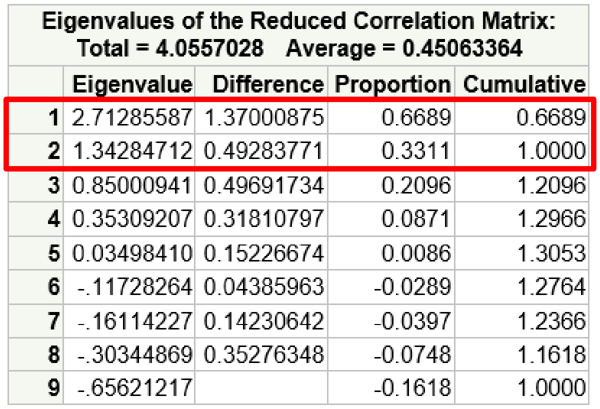
PROMAX Factor Rotation

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 first two factors again explain 100% of variation. In fact the proportion explained by each factor is identical to the Varimax rotation.

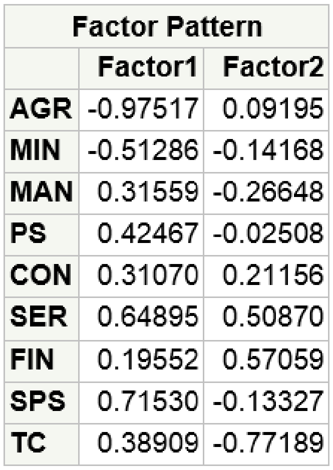


Eigenvalues of the Reduced Correlation Matrix: Total = 4.0557028 Average = 0.45063364

	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

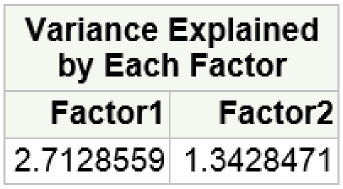
The PROMAX rotation did change the weightings for Factor1 and Factor2 as shown below. In factor the new distribution is identical to the one produced by the VARIMAX rotation.

PROMAX Rotation



Factor Pattern

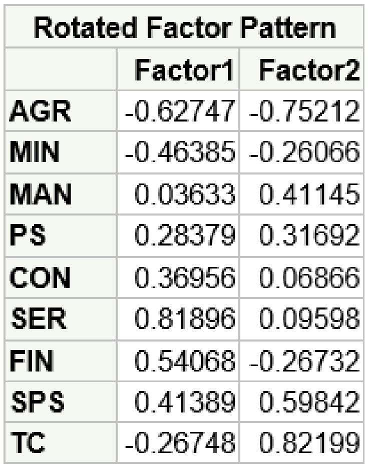
	Factor1	Factor2
AGR	-0.97517	0.09195
MIN	-0.51286	-0.14168
MAN	0.31559	-0.26648
PS	0.42467	-0.02508
CON	0.31070	0.21156
SER	0.64895	0.50870
FIN	0.19552	0.57059
SPS	0.71530	-0.13327
TC	0.38909	-0.77189



Variance Explained by Each Factor

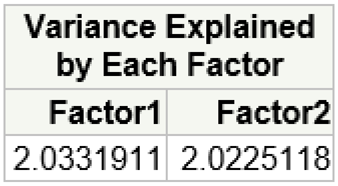
Factor1	Factor2
2.7128559	1.3428471

Un-Rotated



Rotated Factor Pattern

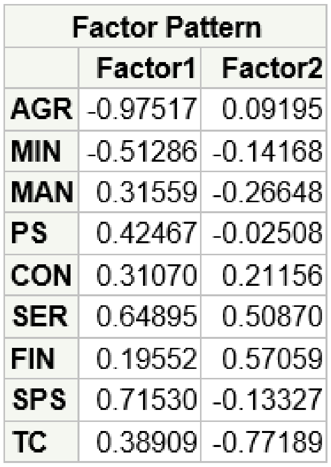
	Factor1	Factor2
AGR	-0.62747	-0.75212
MIN	-0.46385	-0.26066
MAN	0.03633	0.41145
PS	0.28379	0.31692
CON	0.36956	0.06866
SER	0.81896	0.09598
FIN	0.54068	-0.26732
SPS	0.41389	0.59842
TC	-0.26748	0.82199



Variance Explained by Each Factor

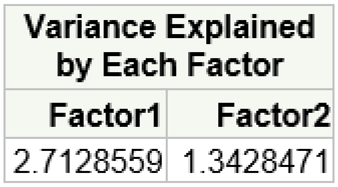
Factor1	Factor2
2.0331911	2.0225118

VARIMAX Rotation



Factor Pattern

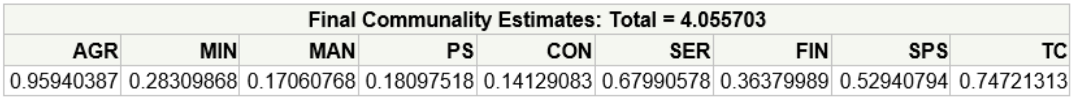
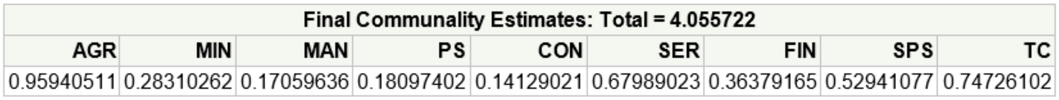
	Factor1	Factor2
AGR	-0.97517	0.09195
MIN	-0.51286	-0.14168
MAN	0.31559	-0.26648
PS	0.42467	-0.02508
CON	0.31070	0.21156
SER	0.64895	0.50870
FIN	0.19552	0.57059
SPS	0.71530	-0.13327
TC	0.38909	-0.77189



Variance Explained by Each Factor

Factor1	Factor2
2.7128559	1.3428471

The total communality was almost identical to non-rotated values with very minimal adjustments to some of the observed values.

<table>
<tr>
<td>rotation</td>
<td>
<table>
<tr><th colspan="9">Final Communality Estimates: Total = 4.055703</th></tr>
<tr><th>AGR</th><th>MIN</th><th>MAN</th><th>PS</th><th>CON</th><th>SER</th><th>FIN</th><th>SPS</th><th>TC</th></tr>
<tr><td>0.95940387</td><td>0.28309868</td><td>0.17060768</td><td>0.18097518</td><td>0.14129083</td><td>0.67990578</td><td>0.36379989</td><td>0.52940794</td><td>0.74721313</td></tr>
</table>
</td>
</tr>
<tr>
<td>Non-rotated</td>
<td>
<table>
<tr><th colspan="9">Final Communality Estimates: Total = 4.055722</th></tr>
<tr><th>AGR</th><th>MIN</th><th>MAN</th><th>PS</th><th>CON</th><th>SER</th><th>FIN</th><th>SPS</th><th>TC</th></tr>
<tr><td>0.95940511</td><td>0.28310262</td><td>0.17059636</td><td>0.18097402</td><td>0.14129021</td><td>0.67989023</td><td>0.36379165</td><td>0.52941077</td><td>0.74726102</td></tr>
</table>
</td>
</tr>
</table>

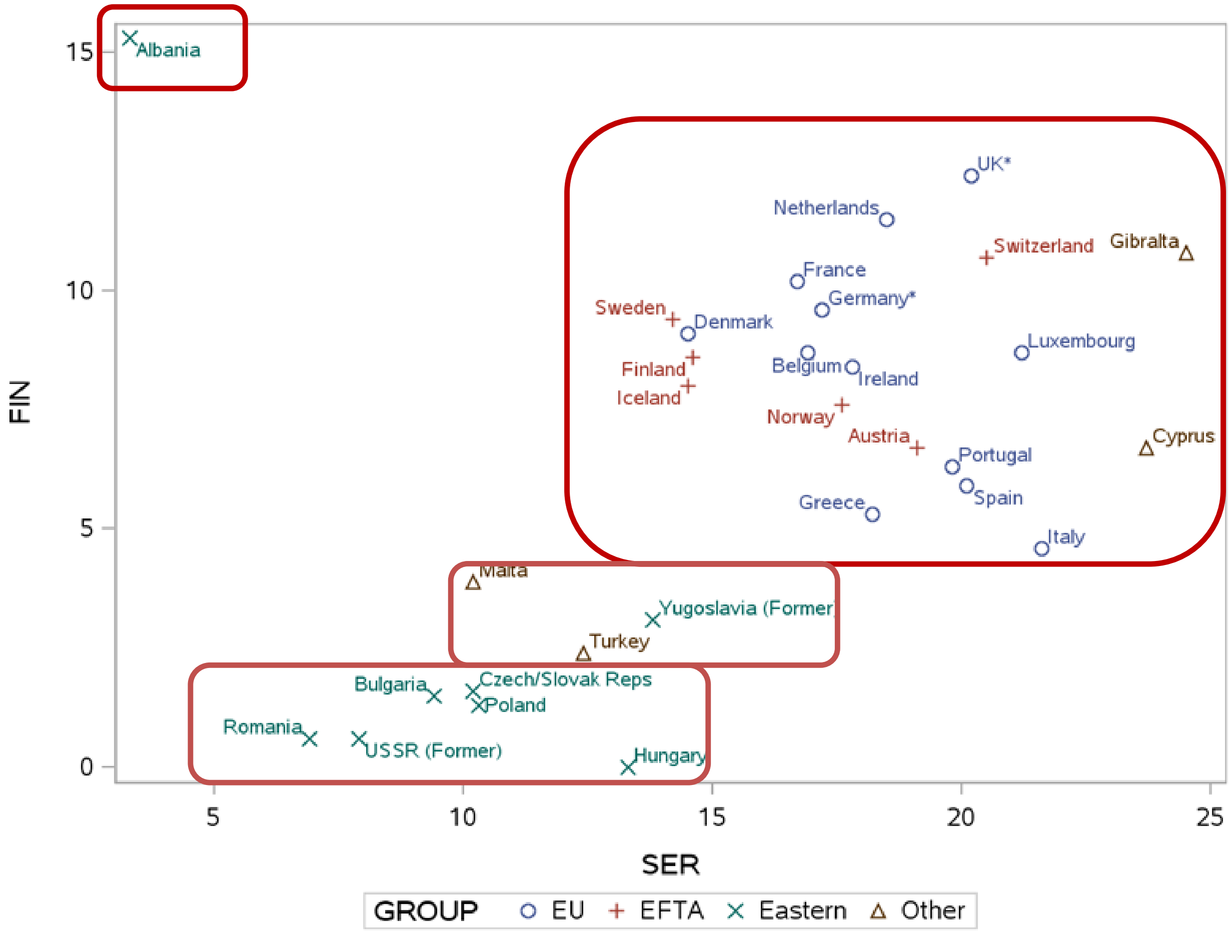
Part 5: Cluster Analysis

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

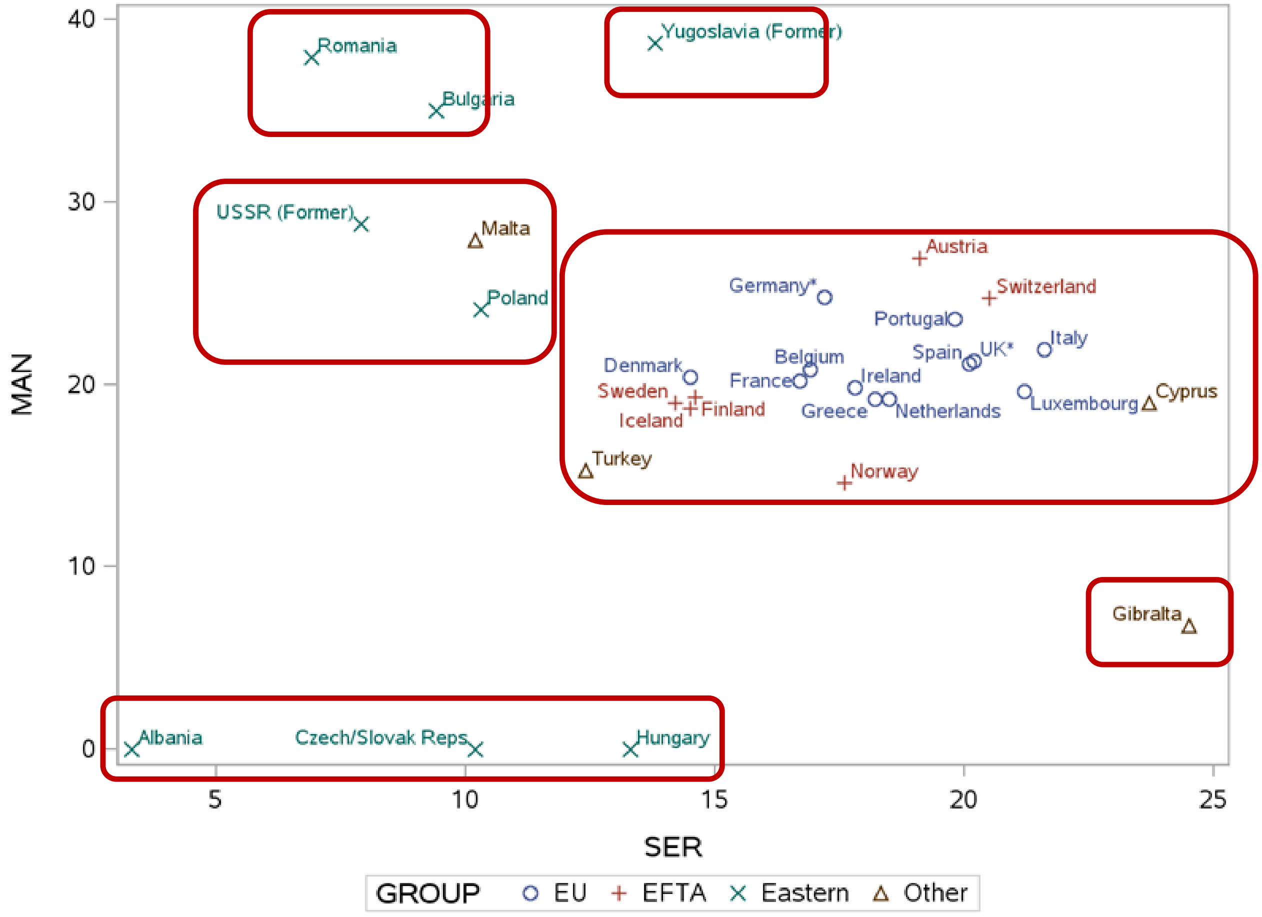
```
ods graphics on;
proc sgplot data=temp;
title 'Scatterplot of Raw Data: FIN*SER';
scatter y=fin x=ser / datalabel=country group=group;
run; quit;
ods graphics off;

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?

There are at least three to five clusters formed for Services by Finance industry groups.


FIN
SER
15
10
5
0
5
10
15
20
25
Albania
UK*
Netherlands
Switzerland
Gibralta
France
Germany*
Sweden
Denmark
Luxembourg
Finland
Belgium
Ireland
Iceland
Norway
Austria
Portugal
Cyprus
Spain
Greece
Italy
Malta
Yugoslavia (Former)
Turkey
Bulgaria
Czech/Slovak Reps
Poland
Romania
USSR (Former)
Hungary
GROUP ○ EU + EFTA × Eastern △ Other

While not exact, I could argue there are five clusters formed for Services by Manufacturing industry groups.



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.

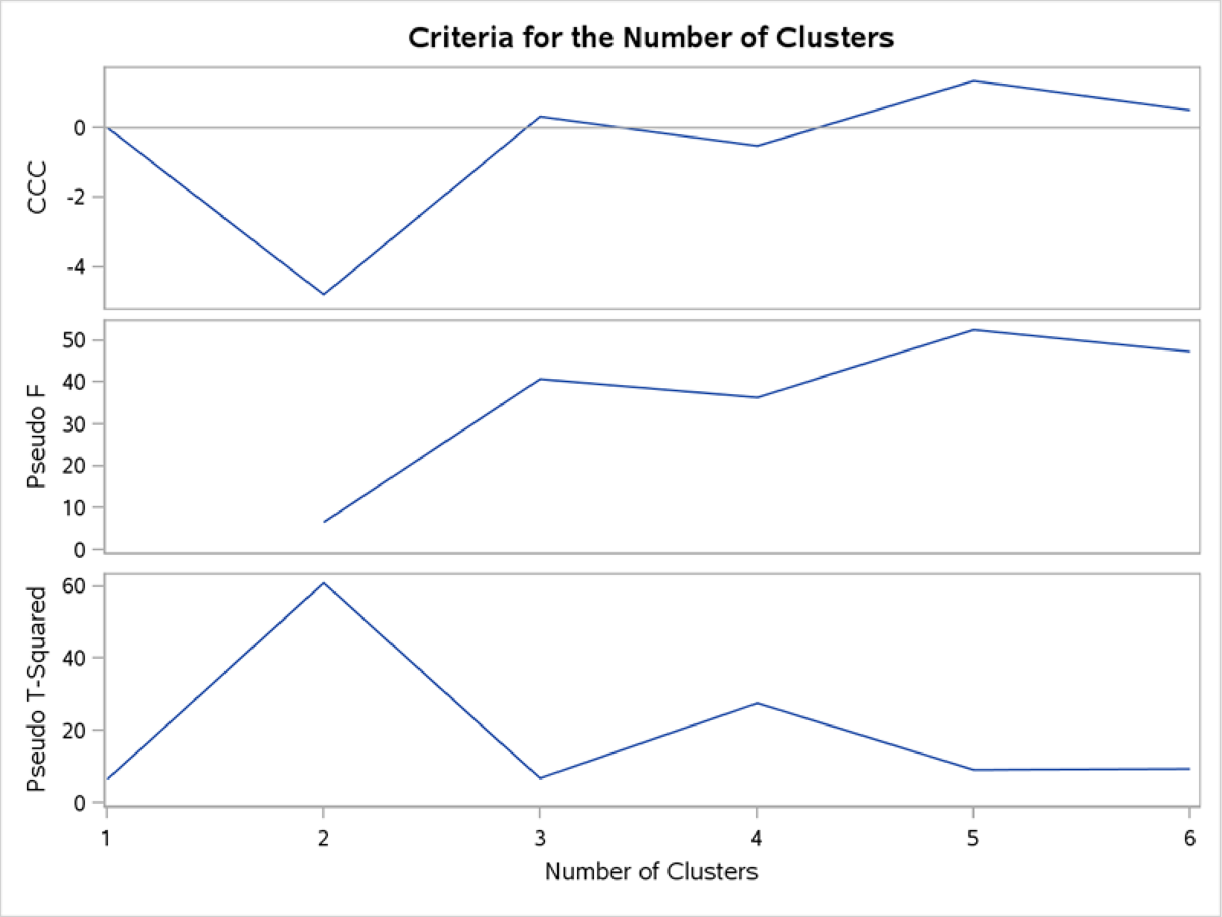
```
ods graphics on;
proc cluster data=temp method=average outtree=tree1 pseudo ccc plots=all;
var fin ser;
id country;
run; quit;
ods graphics off;
```

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

The CCC, PSF and PST2 statistics are useful for estimating the number of clusters in the data (SAS 9.3 User Guide, page 1824).

- CCC – peaks in the plot with values > 2 or 3 indicate good clusters. Based on the plot below its not 100% clear how many good clusters there are but I would state 5.
- PSF – large values indicate good numbers of clusters. This would suggest 5 clusters.
- PST2 – to interpret the plot start from the right and look left until you find the first value that is larger than the previous value, the move back right in the plot to identify that number. The plot would suggest that good clustering exists for 5 and 3.

Based on these plots I would deduce there are 5 clusters.



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;
```

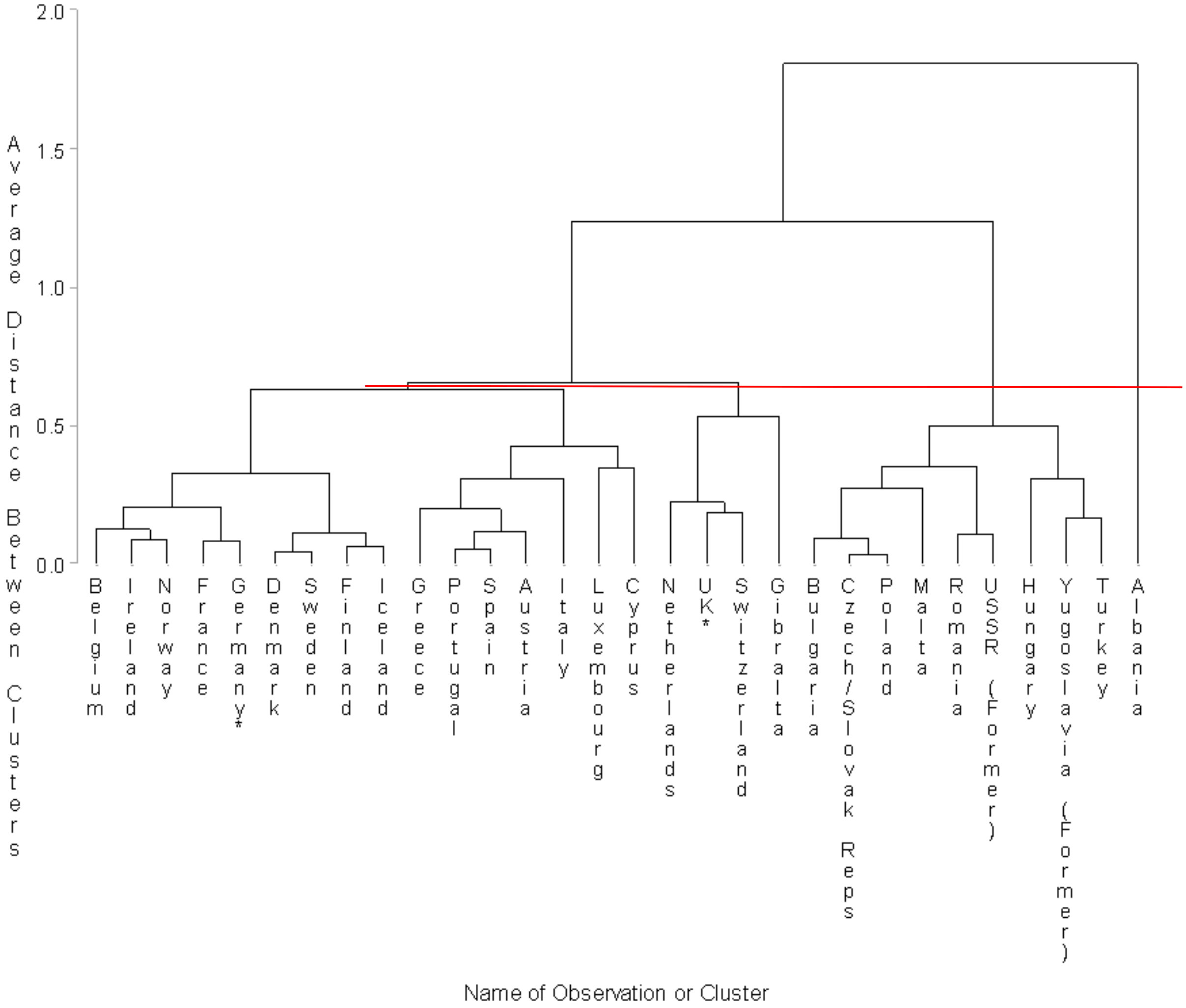


Figure 1: 4 clusters

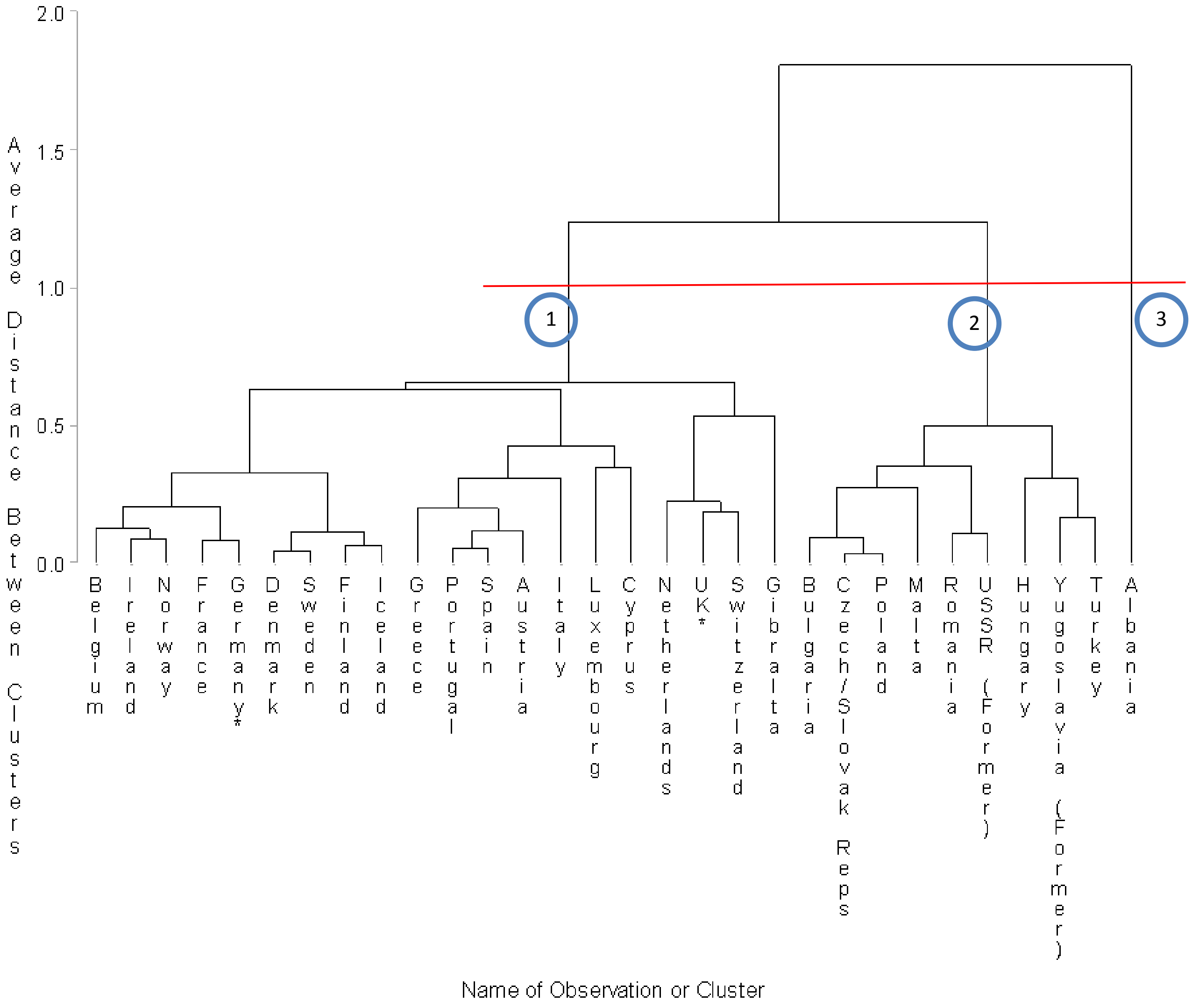


Figure 2: 3 clusters

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?

For three clusters, Albania was one member that did not get clustered with its membership group (Eastern). For four clusters, the EFTA members were grouped into cluster CL4 and CL5 and the EU members were grouped into cluster CL4 and CL5 as well. Albania remained as the one member in its own cluster and not in is membership group of Eastern. More clusters generally provide more homogeneous groups.   Using the dendrograms above there appear minimal differences between 3 and 4 cluster.

3 Clusters

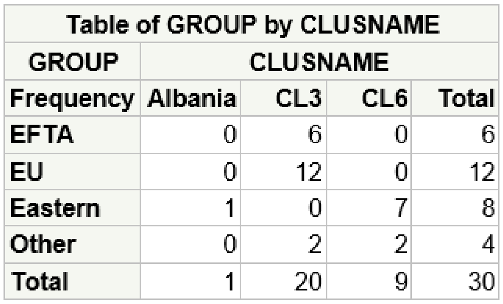


Table of GROUP by CLUSNAME				
GROUP	CLUSNAME			
Frequency	Albania	CL3	CL6	Total
EFTA	0	6	0	6
EU	0	12	0	12
Eastern	1	0	7	8
Other	0	2	2	4
Total	1	20	9	30

4 Clusters

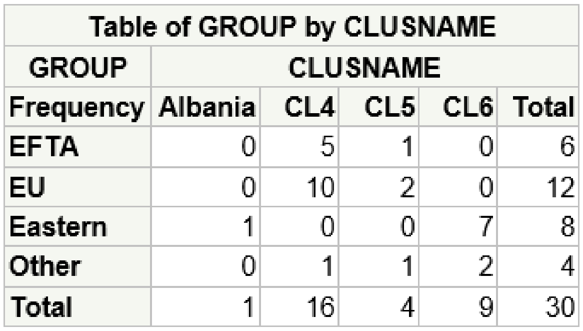
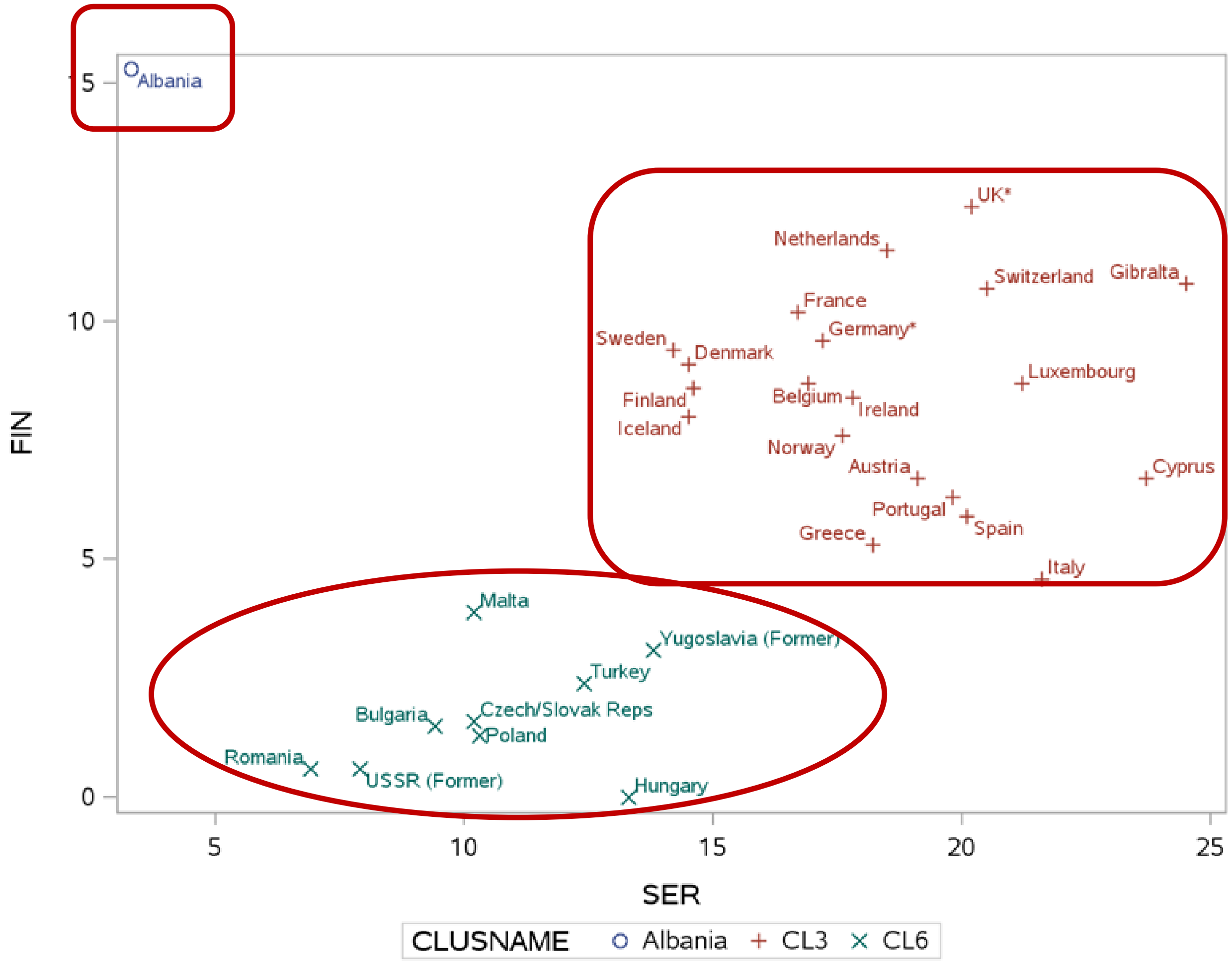
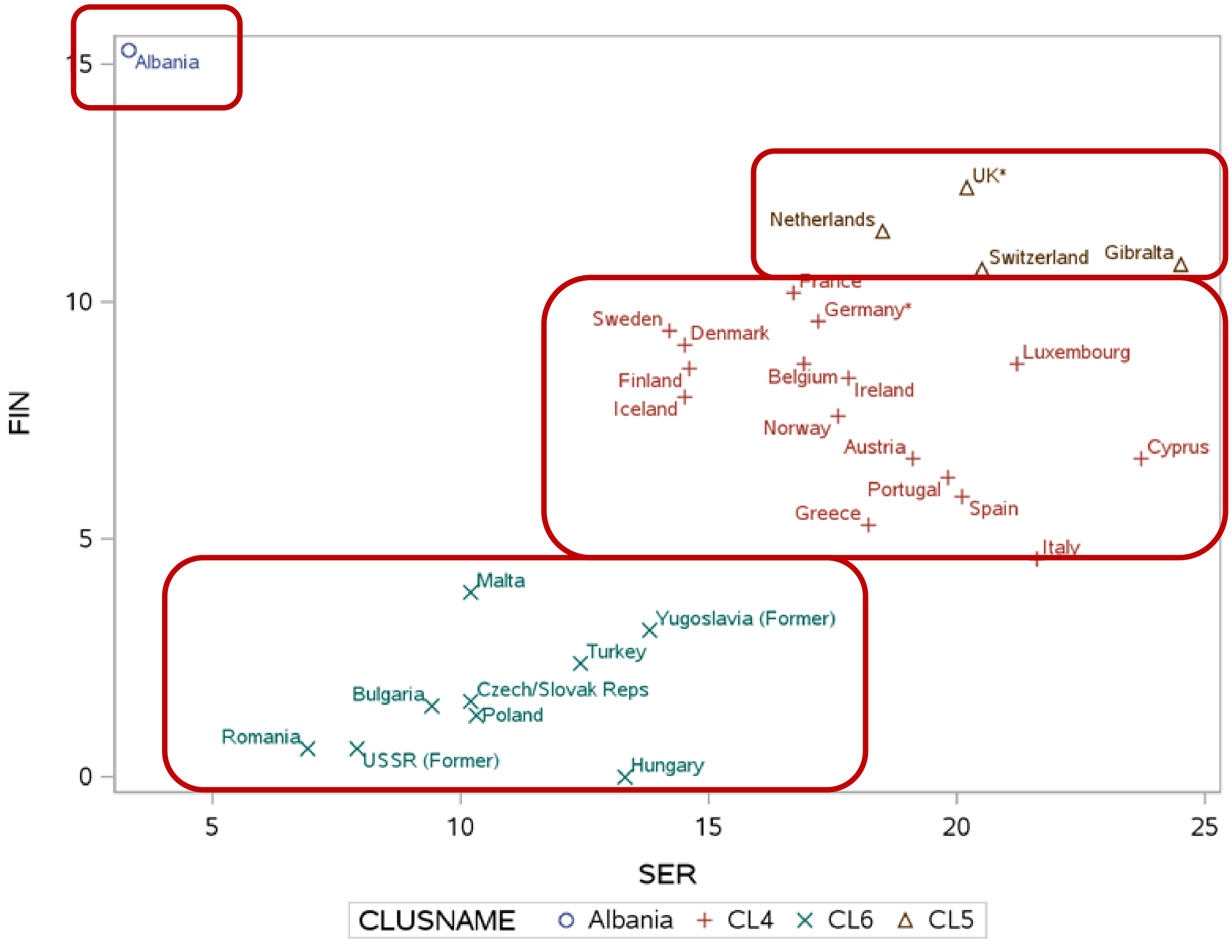


Table of GROUP by CLUSNAME					
GROUP	CLUSNAME				
Frequency	Albania	CL4	CL5	CL6	Total
EFTA	0	5	1	0	6
EU	0	10	2	0	12
Eastern	1	0	0	7	8
Other	0	1	1	2	4
Total	1	16	4	9	30





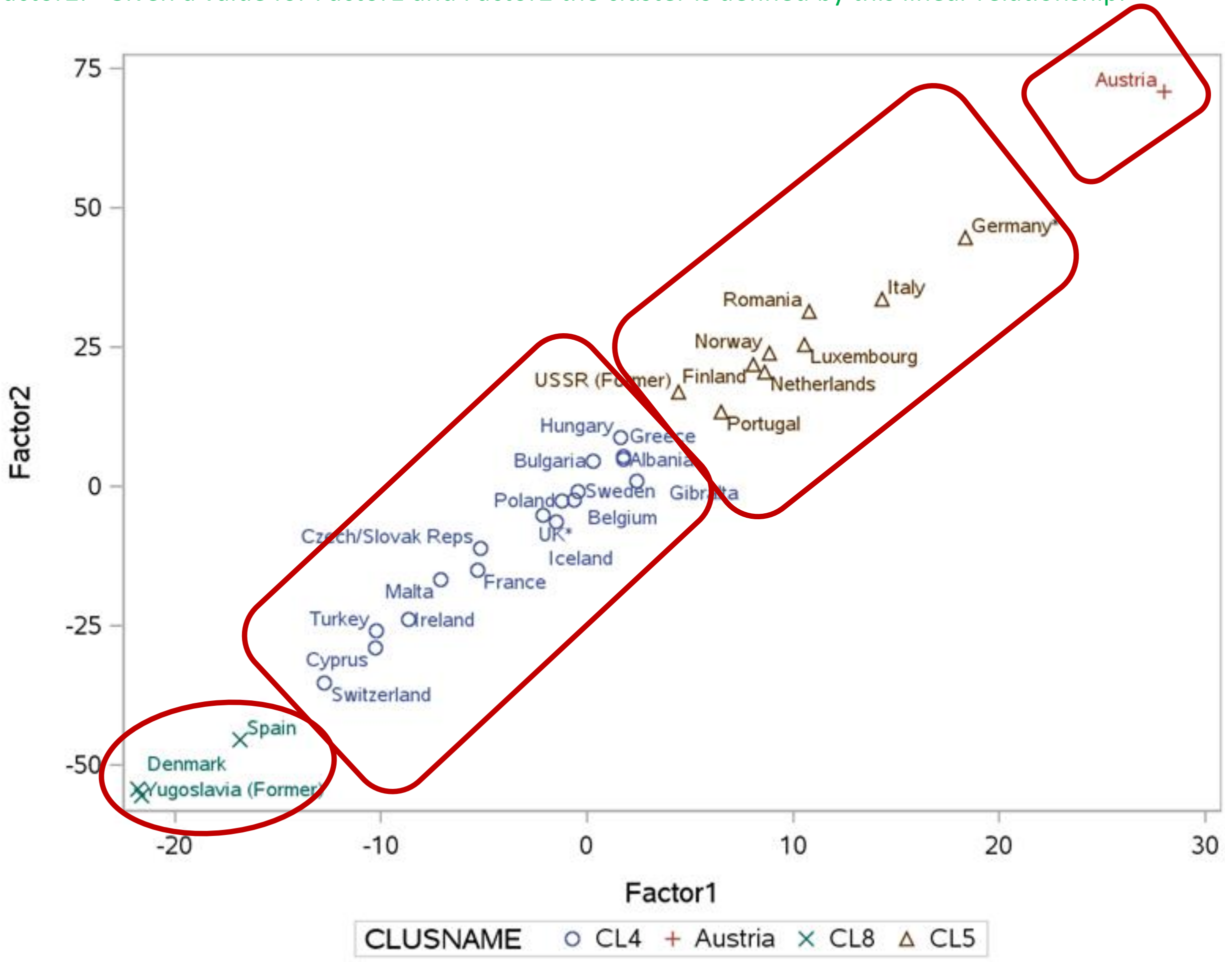
Now perform a similar cluster analysis using the following cluster commands. Which of these four cluster analyses do you prefer?

```
*****************************************************************;
* 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;
```

The above code was run in SAS producing the following output. It would appear that the cluster analysis with VARIMAX rotation is the preferred analysis given the linear relationship between Factor1 and Factor2. Given a value for Factor1 and Factor2 the cluster is defined by this linear relationship.



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.