Assignment #8

Crystal M. Mosley

Introduction:

The purpose of this data analysis is to perform a cluster analysis by completing a comparison of cluster results from raw predictor data, and cluster results from transformed predictor variables using PCA.

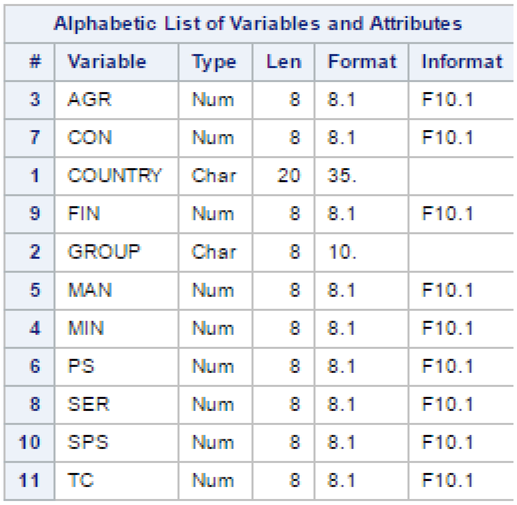
Results:

The dataset consists of employment reporting for various industry segments for 30 European nations as a percentage measurement.

Variable and Industrial Sector:

Variable	Industrial Sector
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

Variables within the dataset:

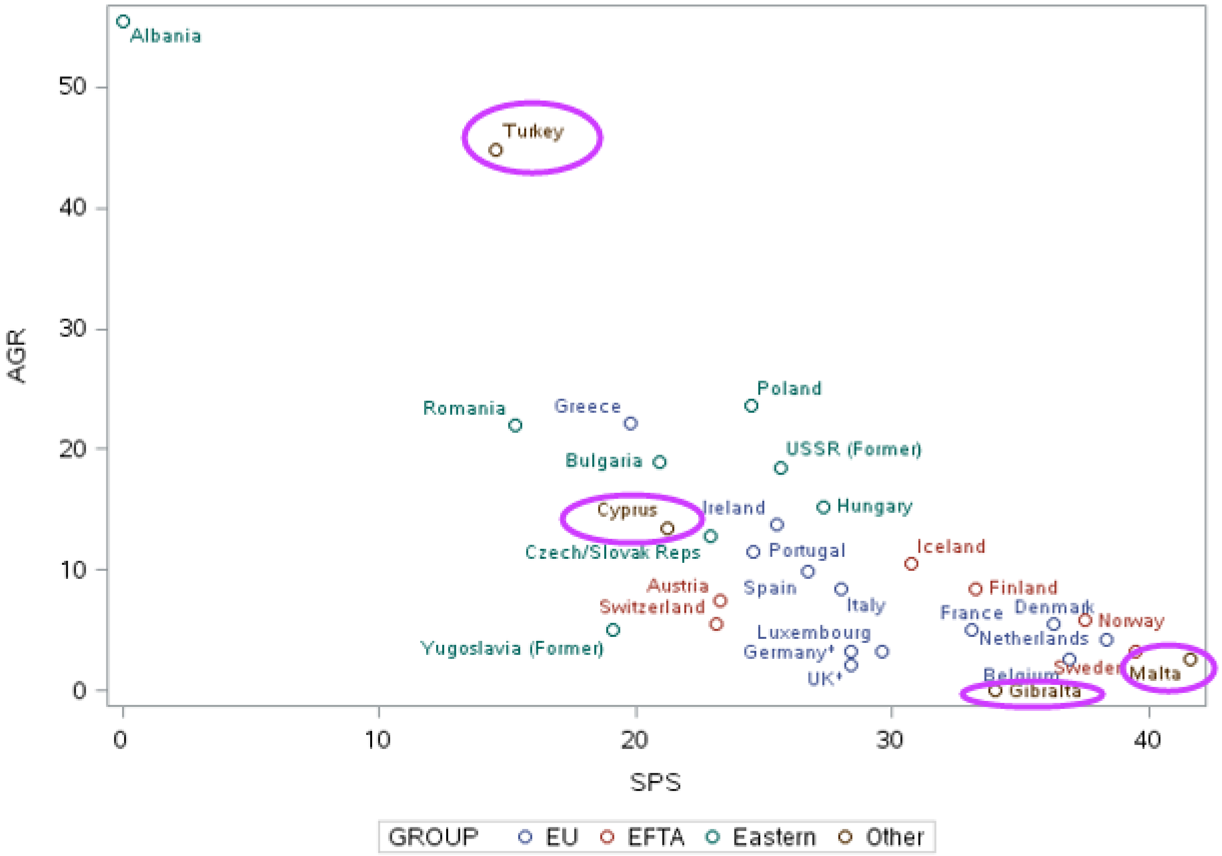


Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
3	AGR	Num	8	8.1	F10.1
7	CON	Num	8	8.1	F10.1
1	COUNTRY	Char	20	35.	
9	FIN	Num	8	8.1	F10.1
2	GROUP	Char	8	10.	
5	MAN	Num	8	8.1	F10.1
4	MIN	Num	8	8.1	F10.1
6	PS	Num	8	8.1	F10.1
8	SER	Num	8	8.1	F10.1
10	SPS	Num	8	8.1	F10.1
11	TC	Num	8	8.1	F10.1

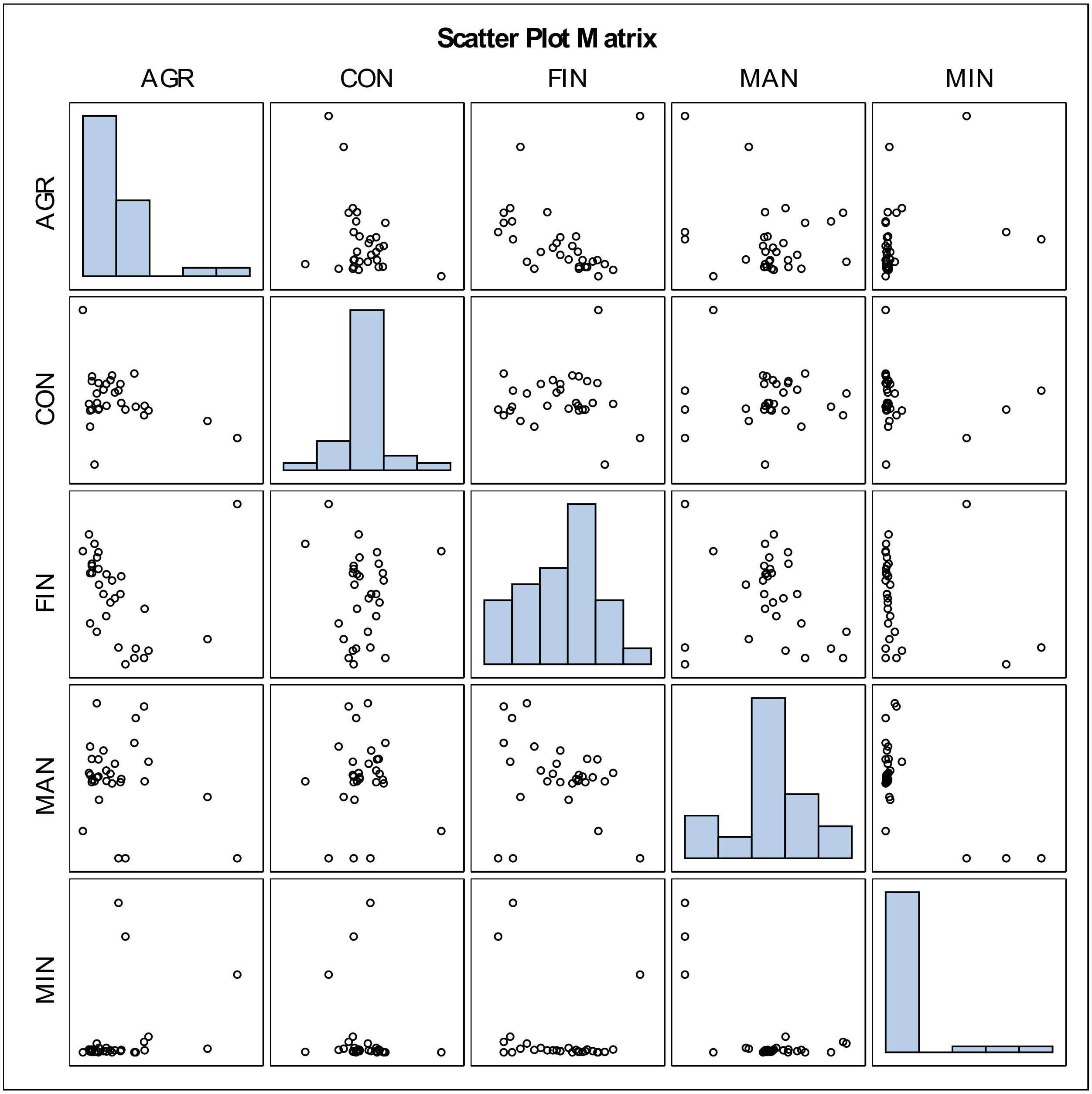
This dataset has a variable, GROUP, which provides subdivision into classes. Examining the contents of GROUP, we see that the subdivision appears to be by trade bloc. Trade bloc is a type of intergovernmental agreement where regional barriers of trading are reduced or eliminated amongst the participating nation-states.

Four countries (Cyprus, Gibralta, Malta, and Turkey) are all within the “Other” group. I would either reassign these countries to the EFTA, or if this assignment was based on location, I’d create another group called the “Mediterranean”.

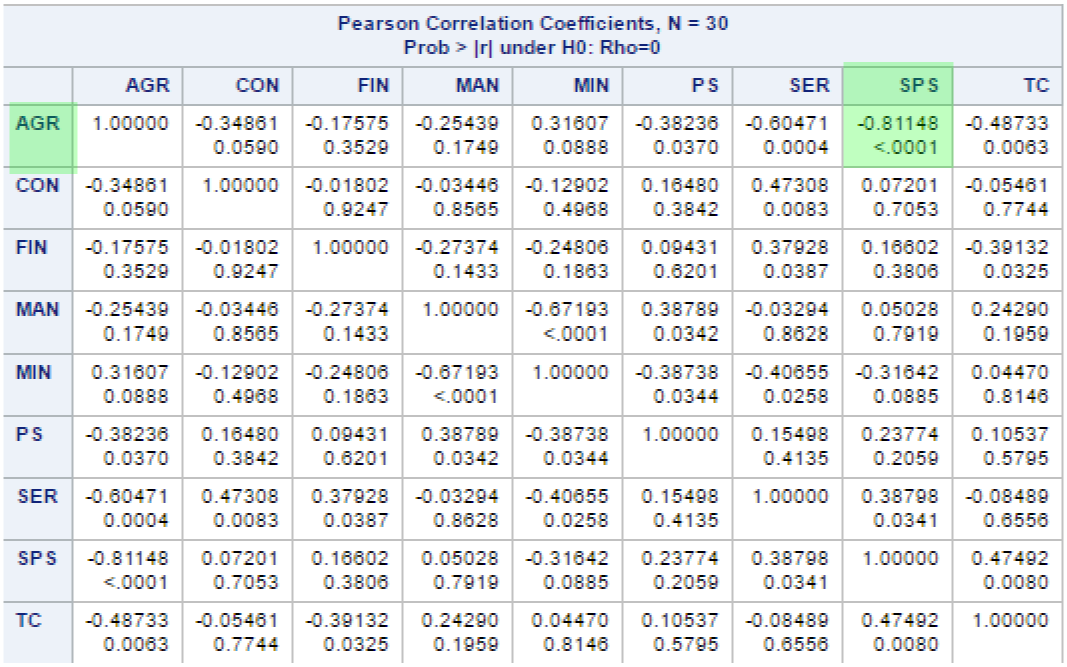


Correlation—

We examine the simple Pearson Correlation for the variables within the dataset, utilizing a PROC CORR, which produces a scatter-plot matrix:



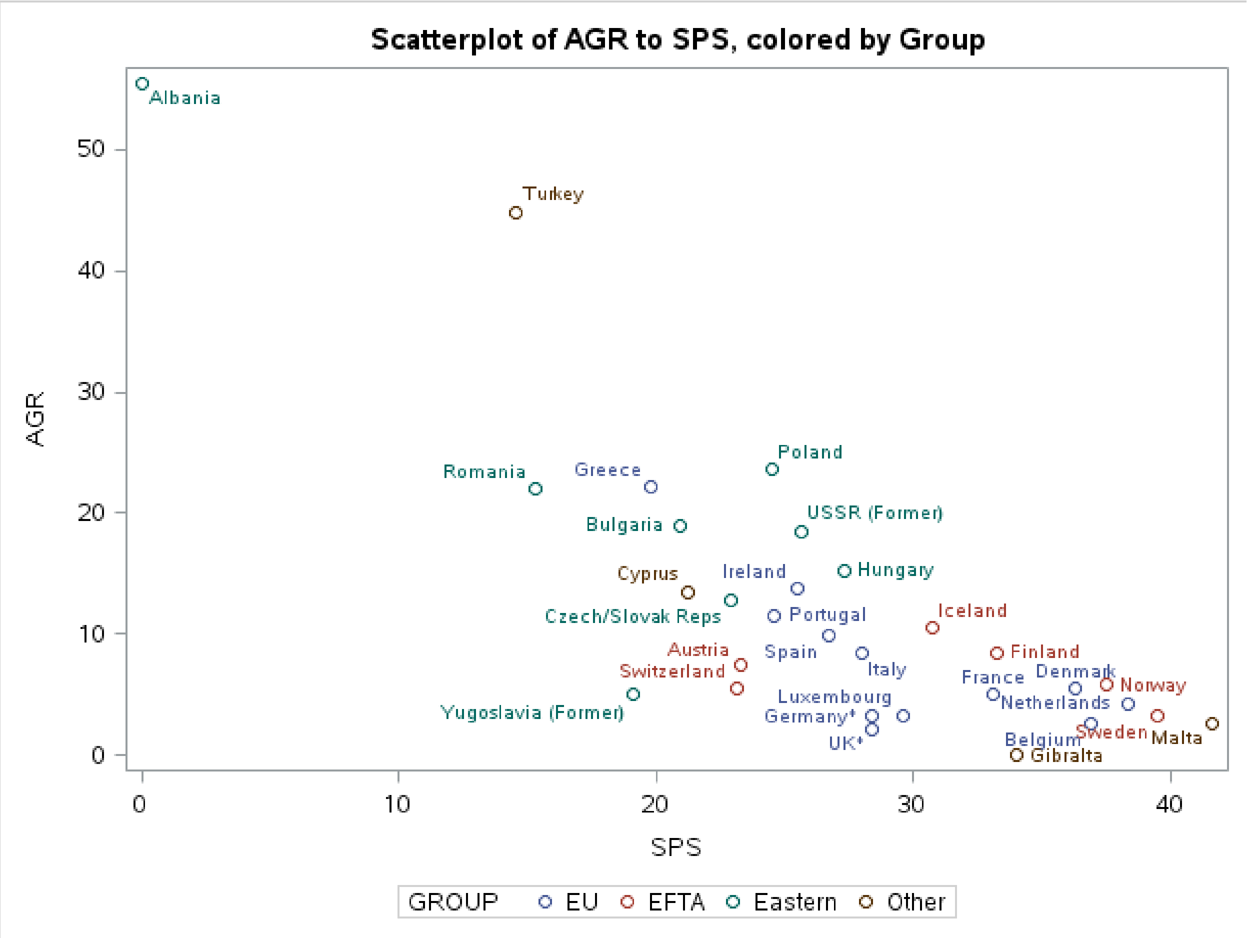
This scatter-plot matrix doesn't fully encompass the correlation matrix. As you can see below, the strongest correlation that's statistically significant, is between AGR and SPS. The correlation being 0.81148 with a probability > |r| under $H_0 : \rho = 0$ test statistic of < 0.0001.



Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

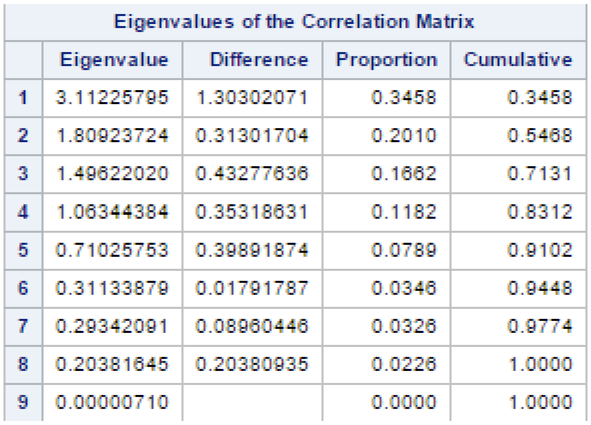
	AGR	CON	FIN	MAN	MIN	PS	SER	SPS	TC
AGR	1.00000	-0.34861 0.0590	-0.17575 0.3529	-0.25439 0.1749	0.31607 0.0888	-0.38236 0.0370	-0.60471 0.0004	-0.81148 <.0001	-0.48733 0.0063
CON	-0.34861 0.0590	1.00000	-0.01802 0.9247	-0.03446 0.8565	-0.12902 0.4968	0.16480 0.3842	0.47308 0.0083	0.07201 0.7053	-0.05461 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

We continue to examine this specific correlation by producing a scatter-plot:



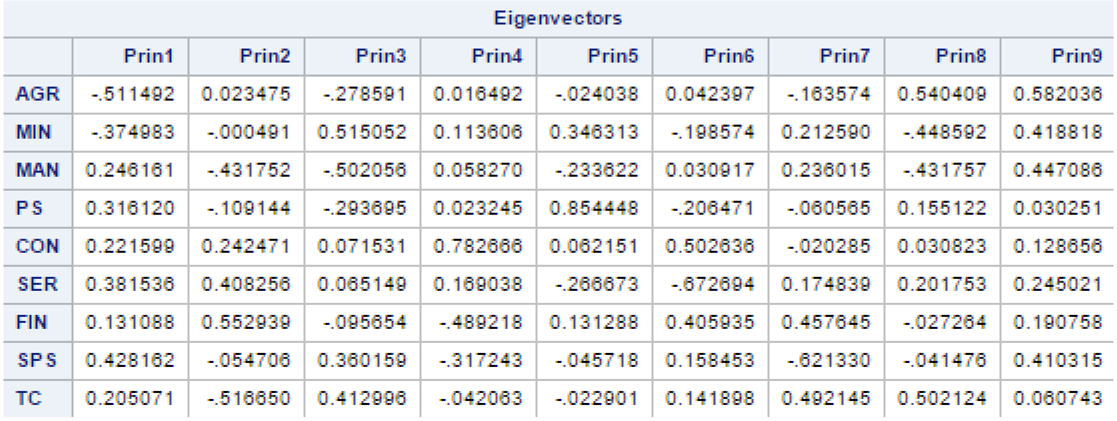
Principal Components, Dimensionality Reduction—

Since nine variables exist within the dataset, we use PCA as a dimensionality reduction method. We use the variability table, or scree plot, to examine how many components we require to account for 90% of the data variability.



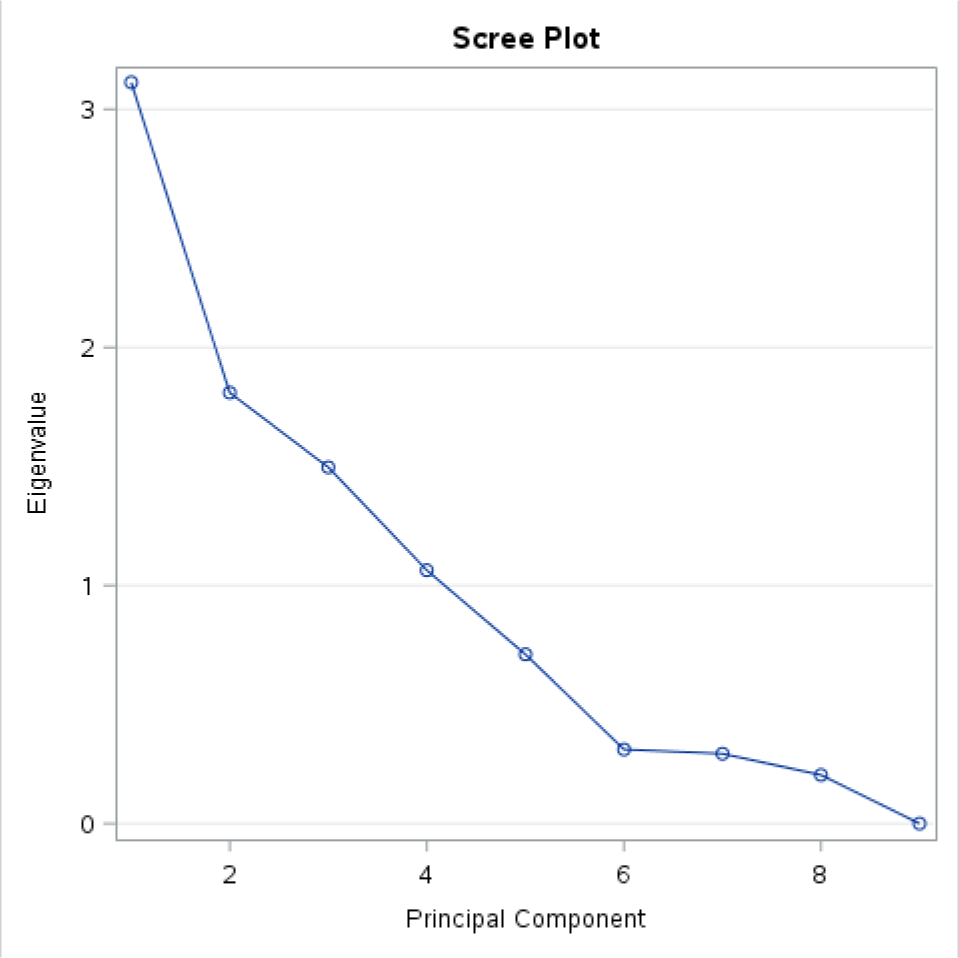
Eigenvalues of the Correlation Matrix

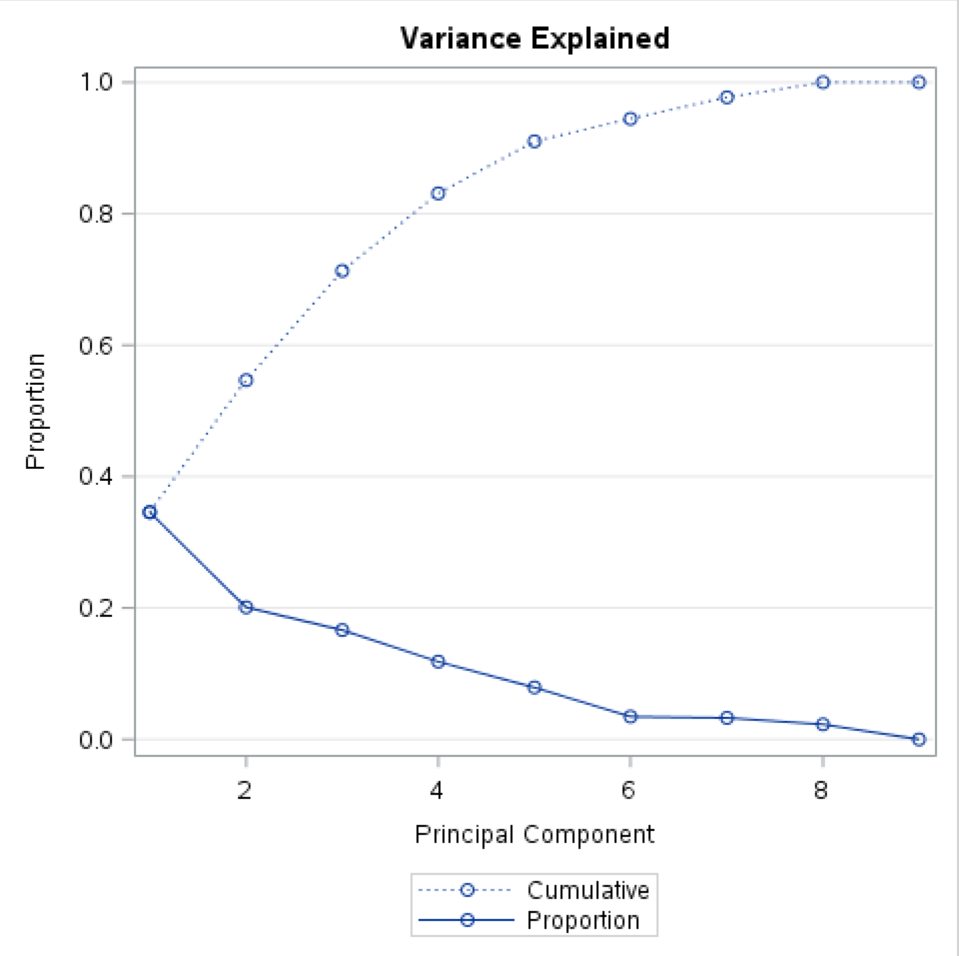
	Eigenvalue	Difference	Proportion	Cumulative
1	3.11225795	1.30302071	0.3458	0.3458
2	1.80923724	0.31301704	0.2010	0.5468
3	1.49622020	0.43277636	0.1662	0.7131
4	1.06344384	0.35318631	0.1182	0.8312
5	0.71025753	0.39891874	0.0789	0.9102
6	0.31133879	0.01791787	0.0346	0.9448
7	0.29342091	0.08960446	0.0326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



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

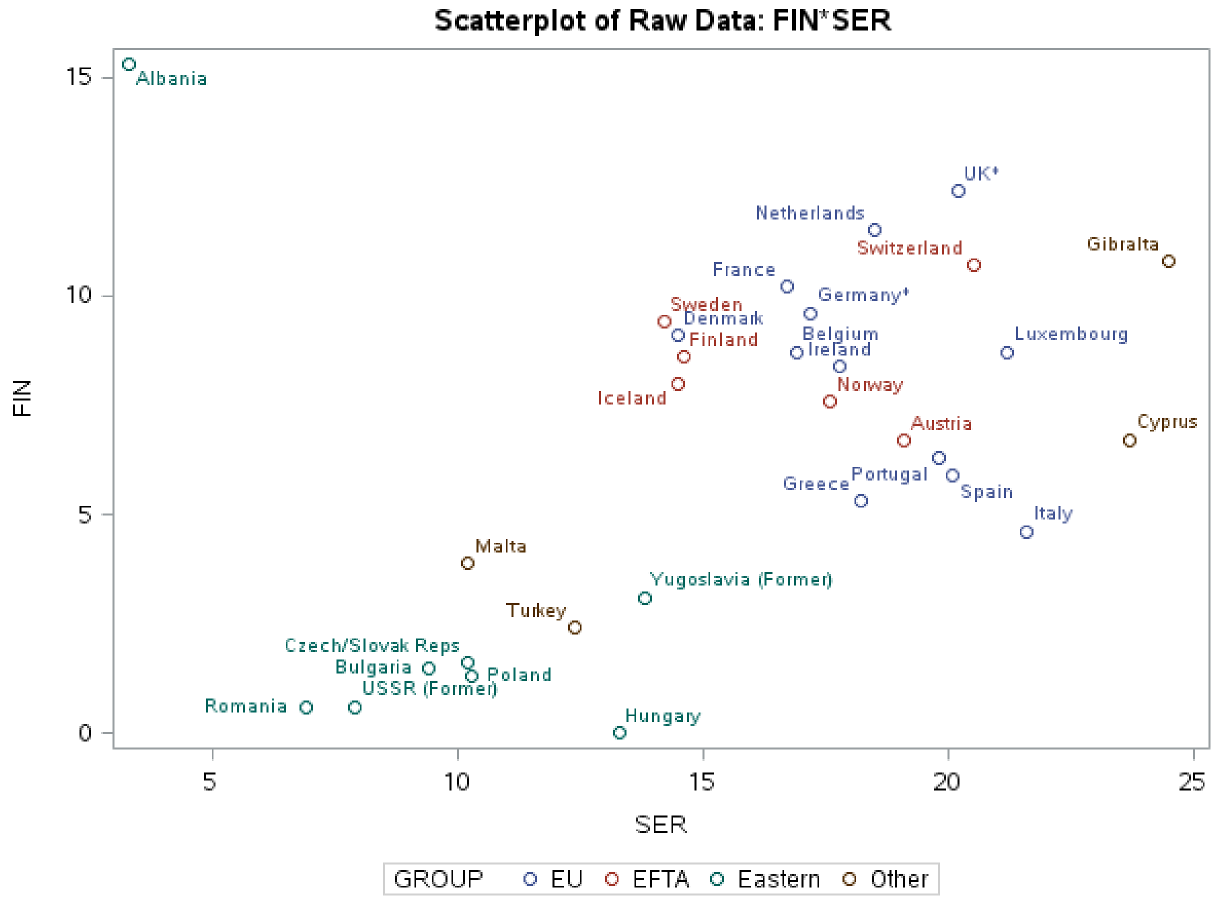


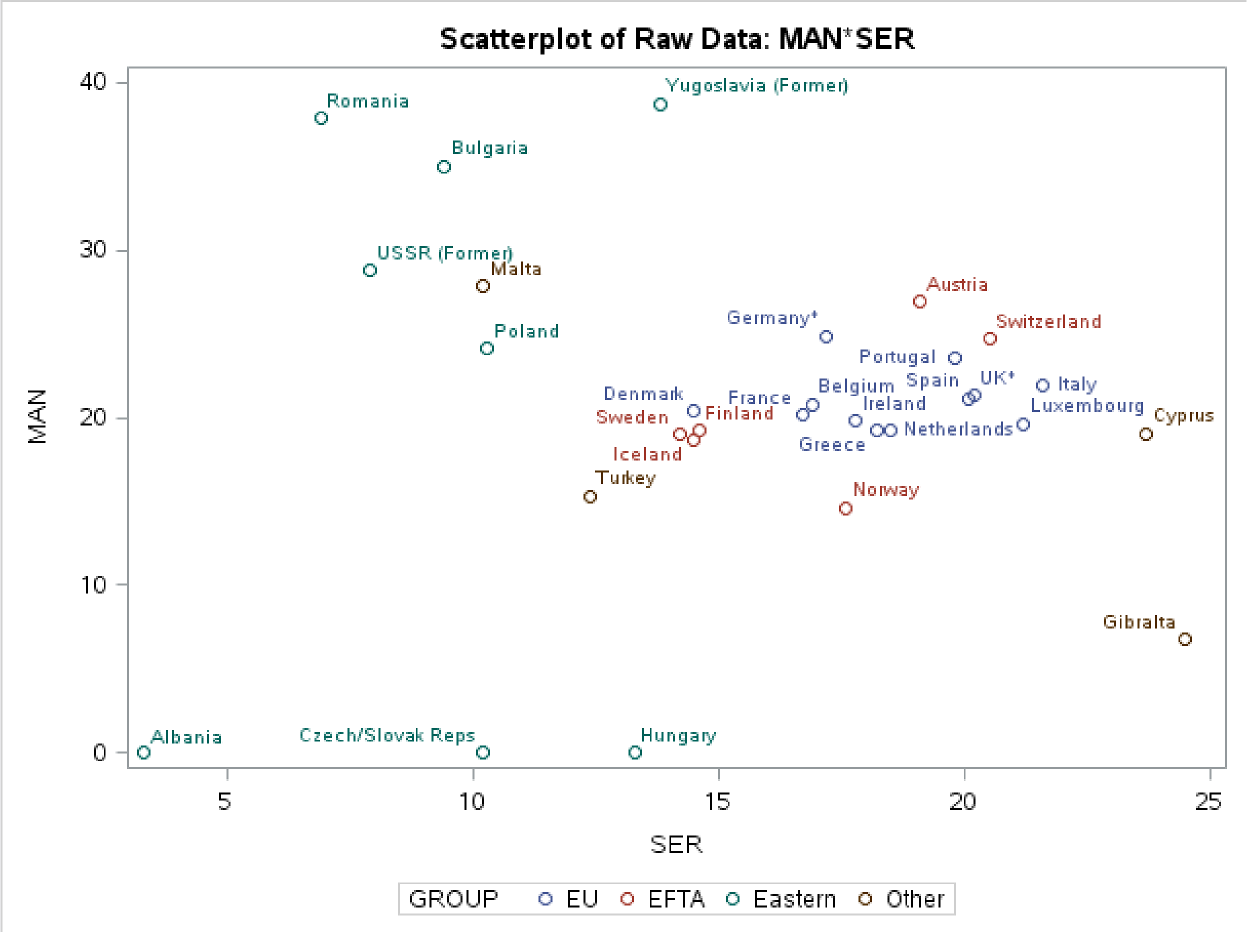


After reviewing the diagnostic output of the PCA procedure, we see that we have to use the first five principal components to explain greater than 90% of the variability in our data. We must accept this since we decided to explain 90% of the *explained variability*. In the future, modeling would be easier if you utilize fewer principal components, which would be less of the *explained variability*.

Cluster Analysis—

We begin by making some of our own scatter-plots, selecting FIN and SER, as well as MAN and SER variables.

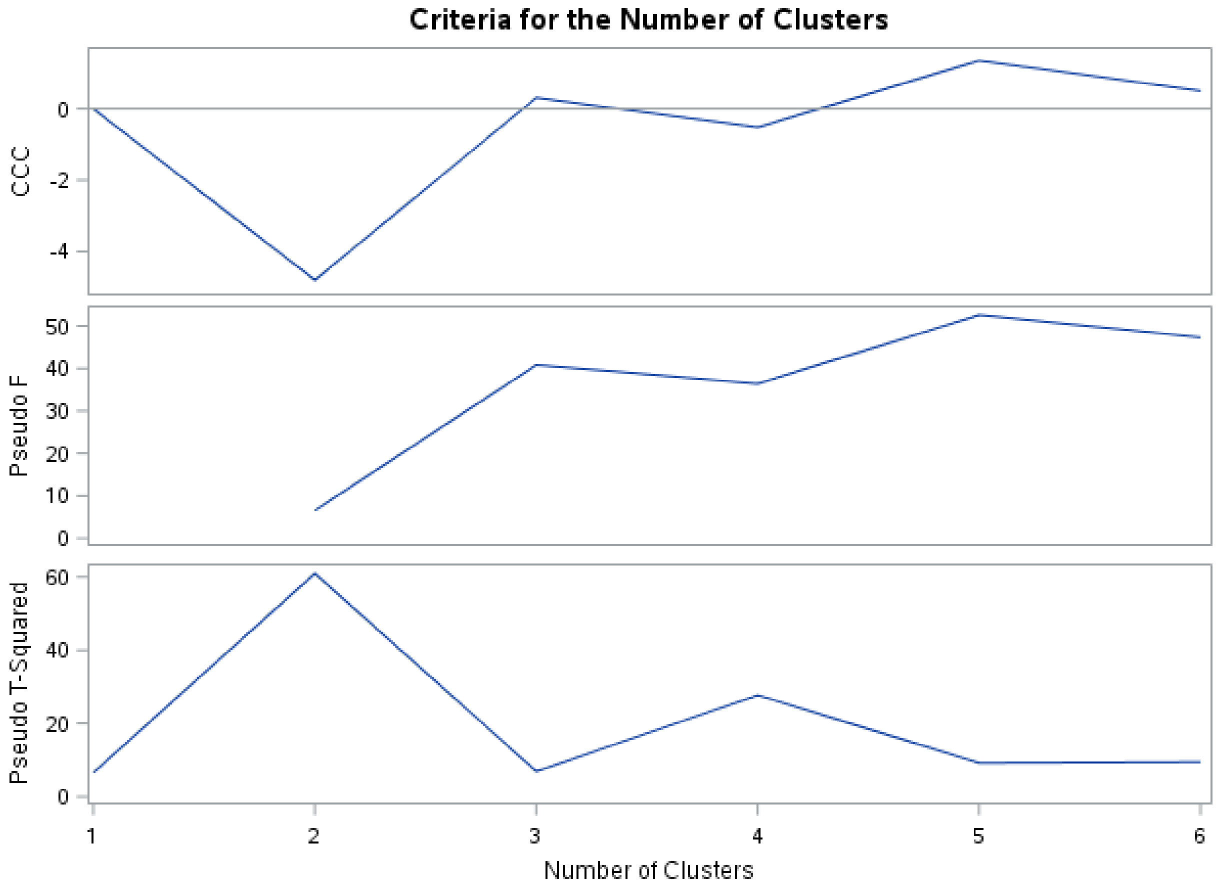




Within each of the above graphs, we can see two clusters. Outlier countries do exist, but in both graphs the Eastern group is clustered, and the EU and EFTA groups are clustered. Gibralta and Albania both look to be outliers.

We will use the cluster procedure within SAS to automatically create clusters with a hierarchical approach. Since this is a hierarchical approach, we do not have to specify the amount of desired clusters. Instead, we can examine diagnostic output for different criteria and then make a decision. There are no complete, adequate methods that can be used for determining the number of population clusters for any type of cluster analysis.

We will examine the diagnostic output of the cluster procedure, and look for the Cubic Clustering Criteria (CCC), Pseudo F, and the Pseudo T-Squared:

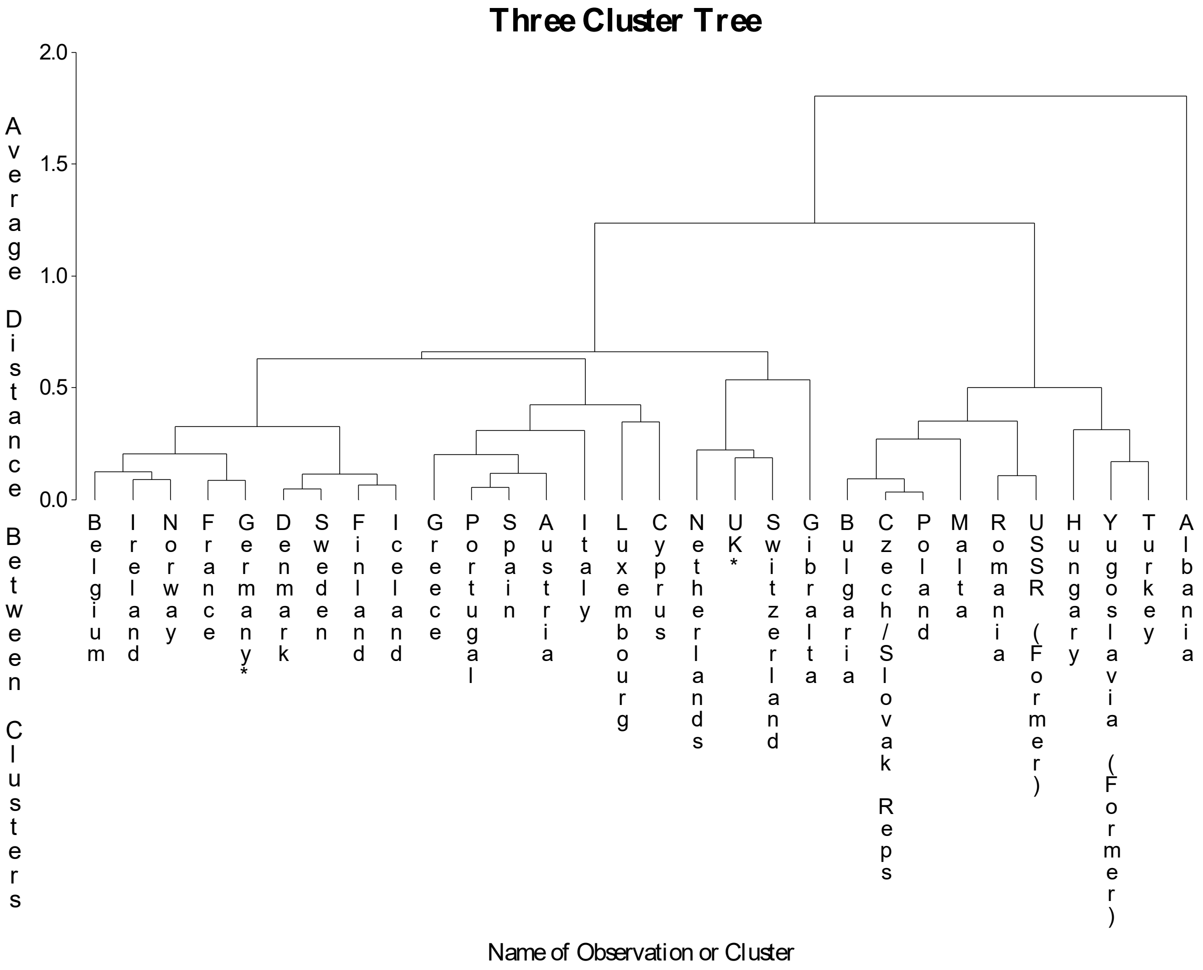


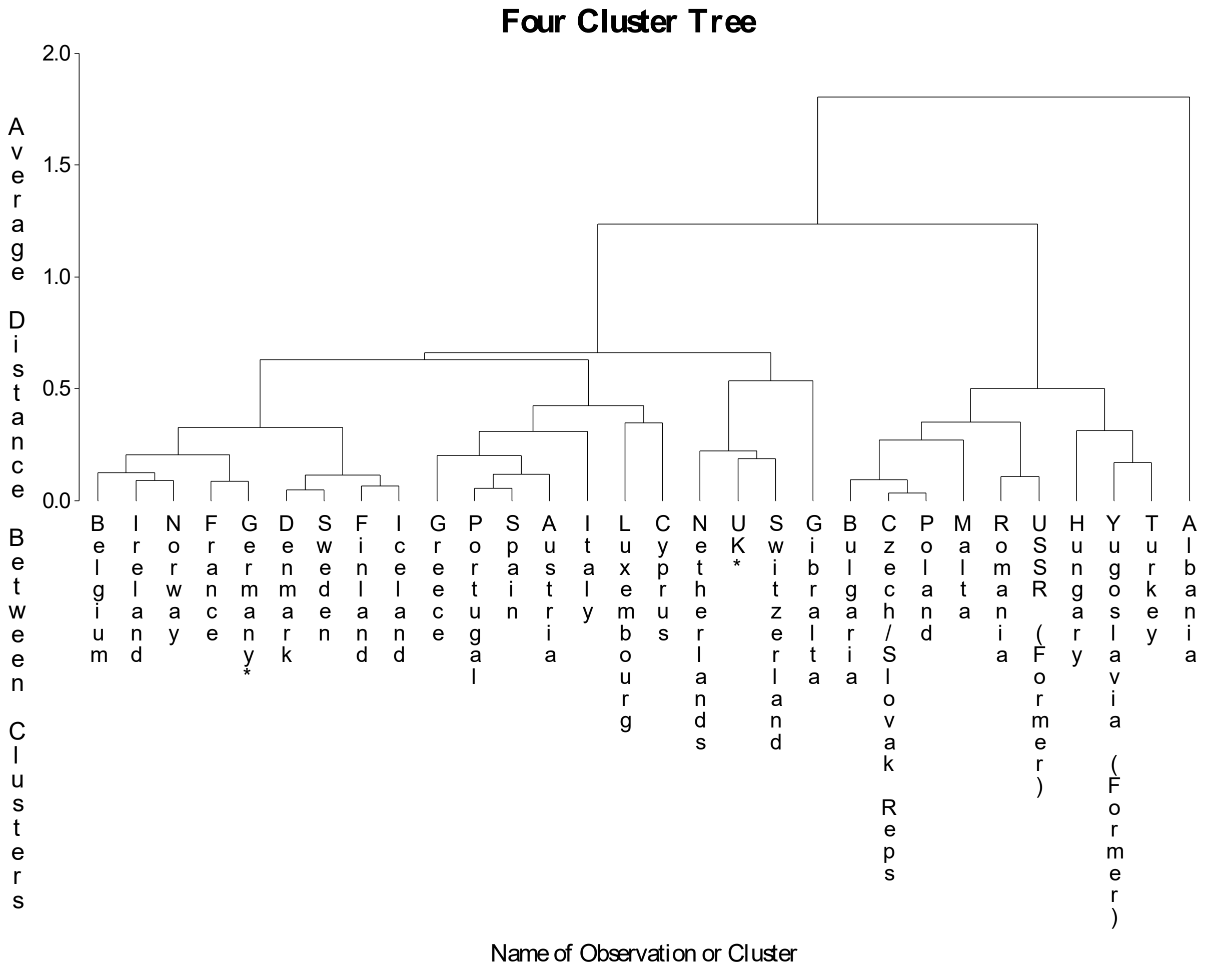
Assuming that the criterion are all graphed in relation to the number of cluster, we will interpret each measurement:

- CCC (Cubic Clustering Criterion)
  - Peaks on the plot with the CCC > 2 or 3 indicate good clusterings
  - Peaks on the plot with CCC between 0 and 2 indicate possible clusters, but should be interpreted cautiously
  - There may be several peaks if the data has a hierarchical structure
  - Distinct non-hierarchical spherical clusters usually show a sharp rise before the peak, followed by a gradual decline
  - Distinct non-hierarchical elliptical clusters usually show a sharp rise to the correct number of clusters, followed by a furthered gradual increase and eventually a gradual decline
  - If all values of the CCC are negative and decreasing for two or more cluster, the distribution is probably unimodal or long-tailed
  - Negative values of the CCC may be due to outliers. Outliers should be removed before clustering
- Pseudo F
  - Look for a comparatively large value

Between CCC and Pseudo F, we conclude that we would be okay with at least three clusters, or four.

We use the tree procedure to assign observations to a specified number of clusters after the hierarchal clustering. We will examine the tabular output between the three cluster tree and four cluster tree:





Frequency of group to cluster with three clusters

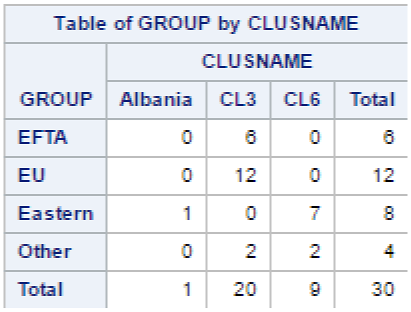


Table of GROUP by CLUSNAME				
	CLUSNAME			
GROUP	Albania	CL3	CL6	Total
EFTA	0	6	0	6
EU	0	12	0	12
Eastern	1	0	7	8
Other	0	2	2	4
Total	1	20	9	30

Frequency of group to cluster with four clusters

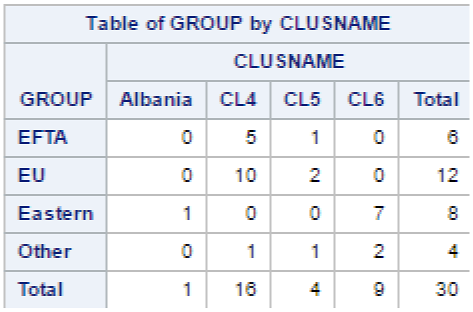
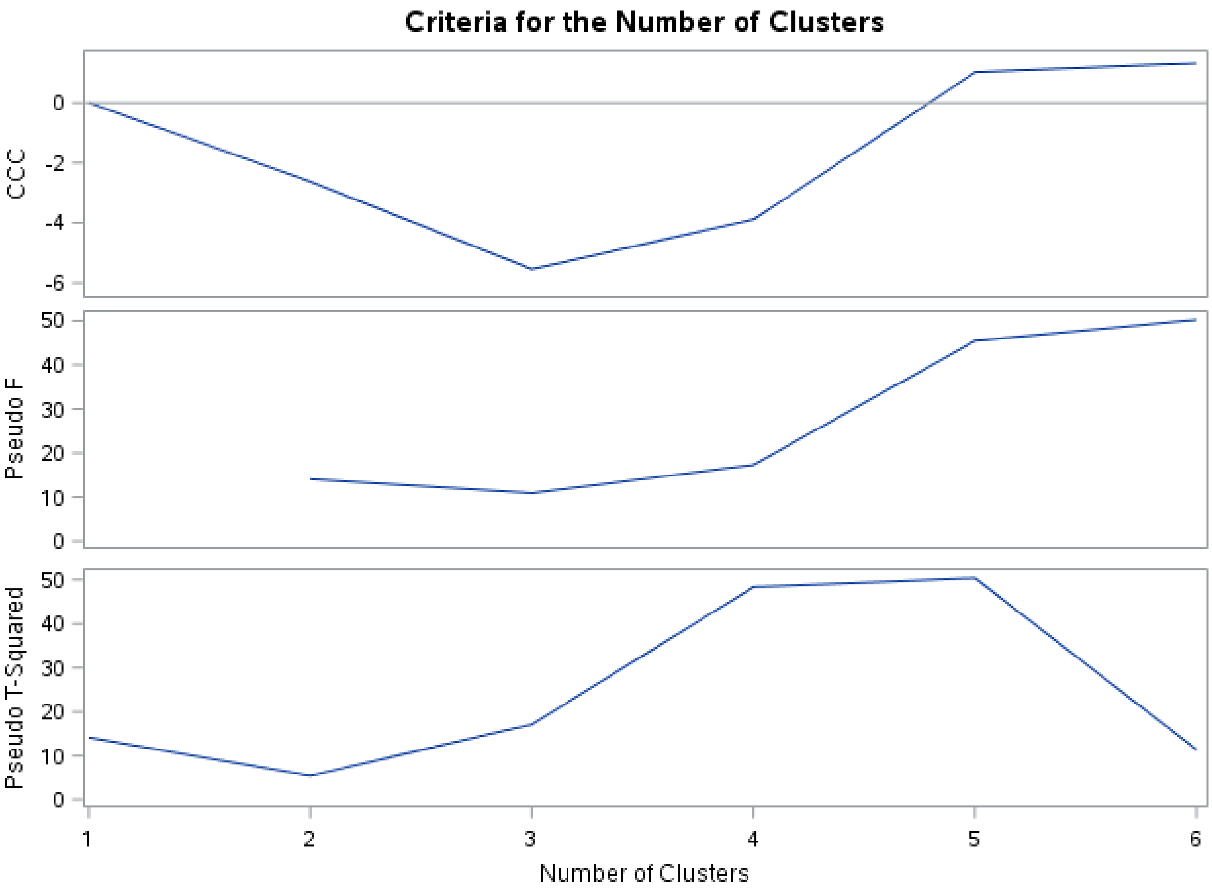
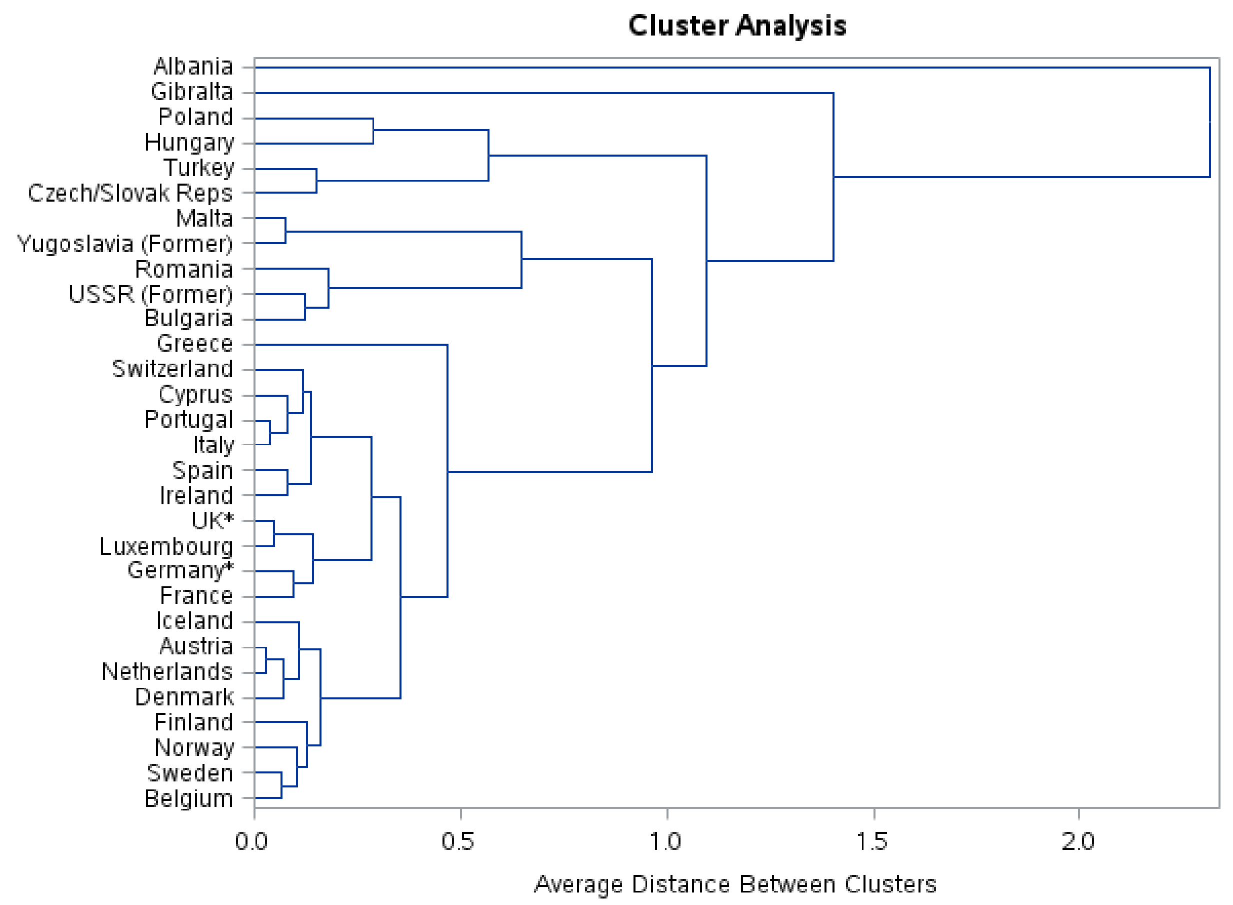


Table of GROUP by CLUSNAME					
	CLUSNAME				
GROUP	Albania	CL4	CL5	CL6	Total
EFTA	0	5	1	0	6
EU	0	10	2	0	12
Eastern	1	0	0	7	8
Other	0	1	1	2	4
Total	1	16	4	9	30

We observe that membership group for this dataset is a guide for where classification into clusters will occur. The three cluster table shows that the existing groups distribute almost solely into a single cluster. Within the four cluster table, the EFTA and EU groups give up some of their members to be distributed amongst other clusters. To reinforce the relative information within the dataset, three clusters is preferred.

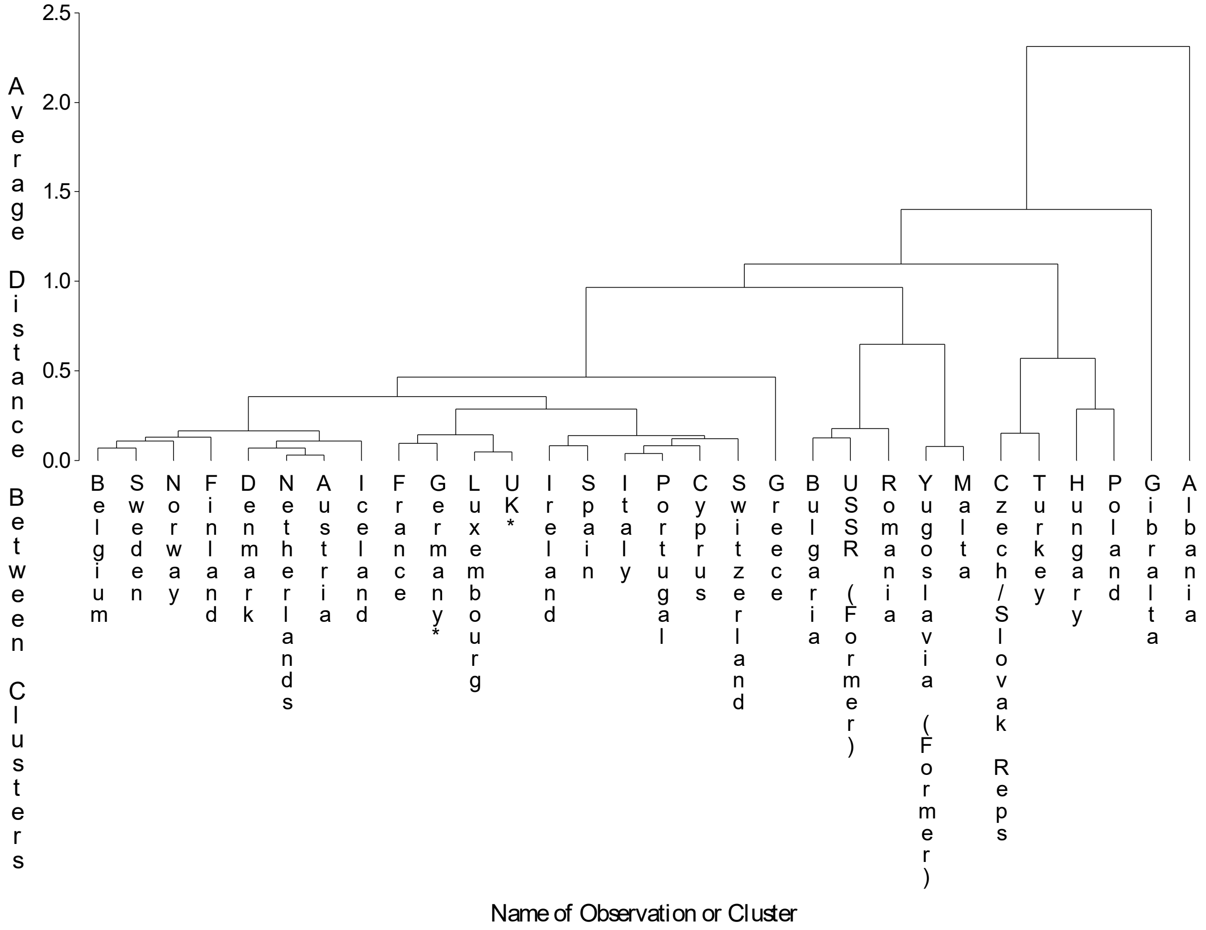
Now, we perform a hierarchical clustering with the PCA dataset:



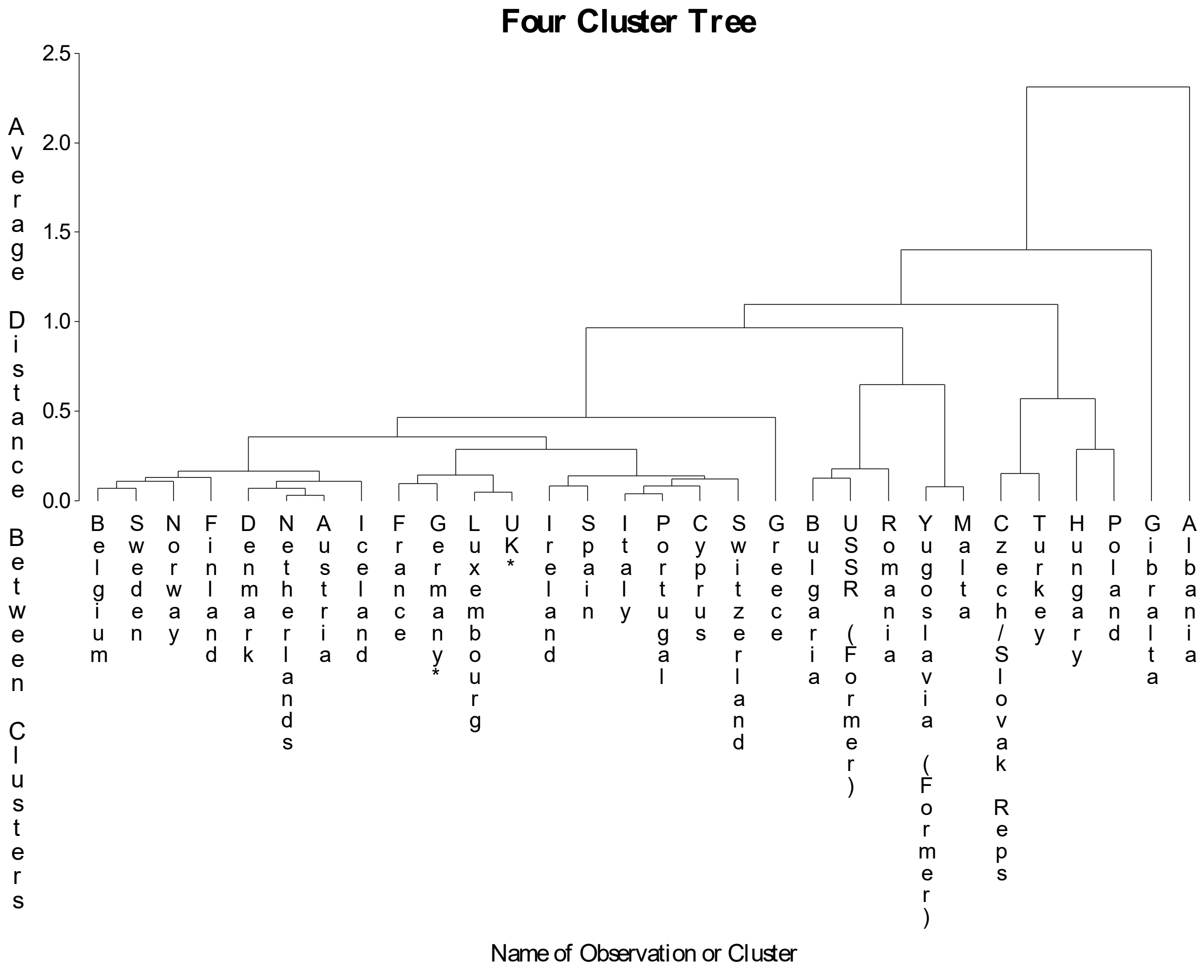


Using previous assumptions for interpreting the criteria, we conclude between the CCC and Pseudo F that we would be okay with at least five or six clusters.

We use the tree procedure to assign observations to a specified number of clusters after the hierarchal clustering. We examine the tabular output between the three and four cluster tree:



Three Cluster Tree



Frequency of group to cluster with three clusters

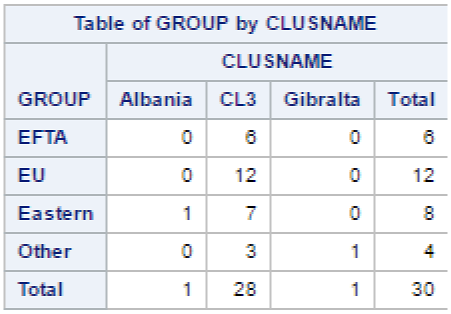


Table of GROUP by CLUSNAME

GROUP	CLUSNAME			
	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

Frequency of group to cluster with four clusters

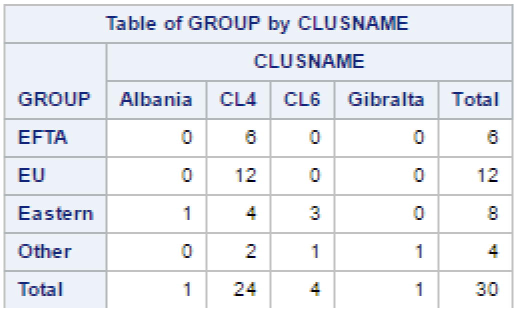


Table of GROUP by CLUSNAME

GROUP	CLUSNAME				
	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

We now see that the membership groups break down a little bit. Using the principal components dataset, we see that this has pushed the clustering towards highlighting the outlier members within the data. Using the raw data is preferred for clustering. If the goal of cluster analysis is to group objects that are more similar (based on some distance method), it would be more useful to have a better distribution of entities into the respective clusters. With the principal components dataset, we see that the clusters have a more skewed amount of membership, as opposed to the raw dataset.

There is likely an assumed bias with this dataset since there is an obvious indication of group membership. In the case of this analysis, we were able to see *some* reinforcement of group membership in the initial clustering. This may have been enough to support analyst bias to a certain level.

Code:

```
libname mydata "/scs/wtm926/" access=readonly;

data temp;
  set mydata.european_employment;

proc contents data=temp;

ods graphics on;
proc corr data=temp nomiss plots=matrix(histogram);
   var AGR CON FIN MAN MIN PS SER SPS TC;

title 'Scatterplot of AGR to SPS, colored by Group';
proc sgplot data=temp;
  scatter y=AGR x=SPS / datalabel=country group=group;

title 'Modeling the Data, Dimensionality Reduction';
proc princomp data=temp out=employ_prin outstat=eigenvectors plots=scree(unpackpanel);

*proc print data=eigenvectors(where=(_TYPE_='SCORE'));


title 'Cluster Analysis: Scatter Plots';

proc sgplot data=temp;
  title 'Scatterplot of Raw Data: FIN*SER';
  scatter y=fin x=ser / datalabel=country group=group;

proc sgplot data=temp;
  title 'Scatterplot of Raw Data: MAN*SER';
  scatter y=man x=ser / datalabel=country group=group;

title 'Cluster Analysis: Automated Cluster Selection';
proc cluster data=temp method=average outtree=tree1 pseudo ccc plots=all;
  var fin ser;
  id country;


proc tree data=tree1 ncl=3 out=_3_clusters;
  title 'Three Cluster Tree';
  copy fin ser;
```

```
proc tree data=tree1 ncl=4 out=_4_clusters;
  title 'Four Cluster Tree';
  copy fin ser;


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

  proc sort data=tree_data; by country;

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

  proc sort data=group_affiliation;
    by country;

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

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

%mend makeTable;

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

proc sgplot data=_3_clusters_with_labels;
  title 'Three Clusters with Labels';
  scatter y=fin x=ser / datalabel=country group=clusname;

proc sgplot data=_4_clusters_with_labels;
  title 'Four Clusters with Labels';
  scatter y=fin x=ser / datalabel=country group=clusname;

proc cluster data=employ_prin method=average outtree=tree3 pseudo ccc plots=all;
  title 'Cluster with Prin1 and Prin2';
  var prin1 prin2;
  id country;

proc tree data=tree3 ncl=3 out=_3_clusters;
  title 'Three Cluster Tree';
  copy prin1 prin2;

proc tree data=tree3 ncl=4 out=_4_clusters;
```

```
  title 'Four Cluster Tree';
  copy prin1 prin2;

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

proc sgplot data=_3_clusters_with_labels;
  title 'Three Clusters with Labels';
  scatter y=prin2 x=prin1 / datalabel=country group=clusname;

proc sgplot data=_4_clusters_with_labels;
  title 'Four Clusters with Labels';
  scatter y=prin2 x=prin1 / datalabel=country group=clusname;

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

run;
ods graphics off;
```