Cluster Analysis

FY 2016

Author Eric Lewis

TABLE OF CONTENTS

Contents

Introduction

The objective of this cluster analysis is to identify the clusters within the European employment data set. The end result of cluster analysis is for the purpose of market segmentation to identify clusters to match either products of services.

Data Survey

Data Preparation

The European employment data set is utilized for this Cluster Analysis. This data set contains 11 variables having 30 observations. The Cluster Analysis will begin with an EDA and completed with a comparison of cluster results from transformed predictor variables using principal components analysis.

Initial Correlation Analysis

This analysis begins with an initial examination of the European employment data set. The data set variables are as follows, table 1 and table 2 including the industry sector.

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

Table 1: Alphabetic List of Variables and Attributes

Variable	Industry 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
AGR:	Agriculture
MIN:	Mining

Table 2: Variables and Industry Sectors

Upon examination of the observations of the data set, we observe a variable named group. This variable is composed of trade blocs: EU, EFTA, Eastern, and Other. The trade bloc is an intergovernmental agreement in which the barriers to trade are reduced among the participating countries. Observations 27 thru 30 show that Cyprus, Gibralta, Malta, and Turkey are in the "Other" group. We may find these countries may potentially not group accordingly in our cluster analysis.

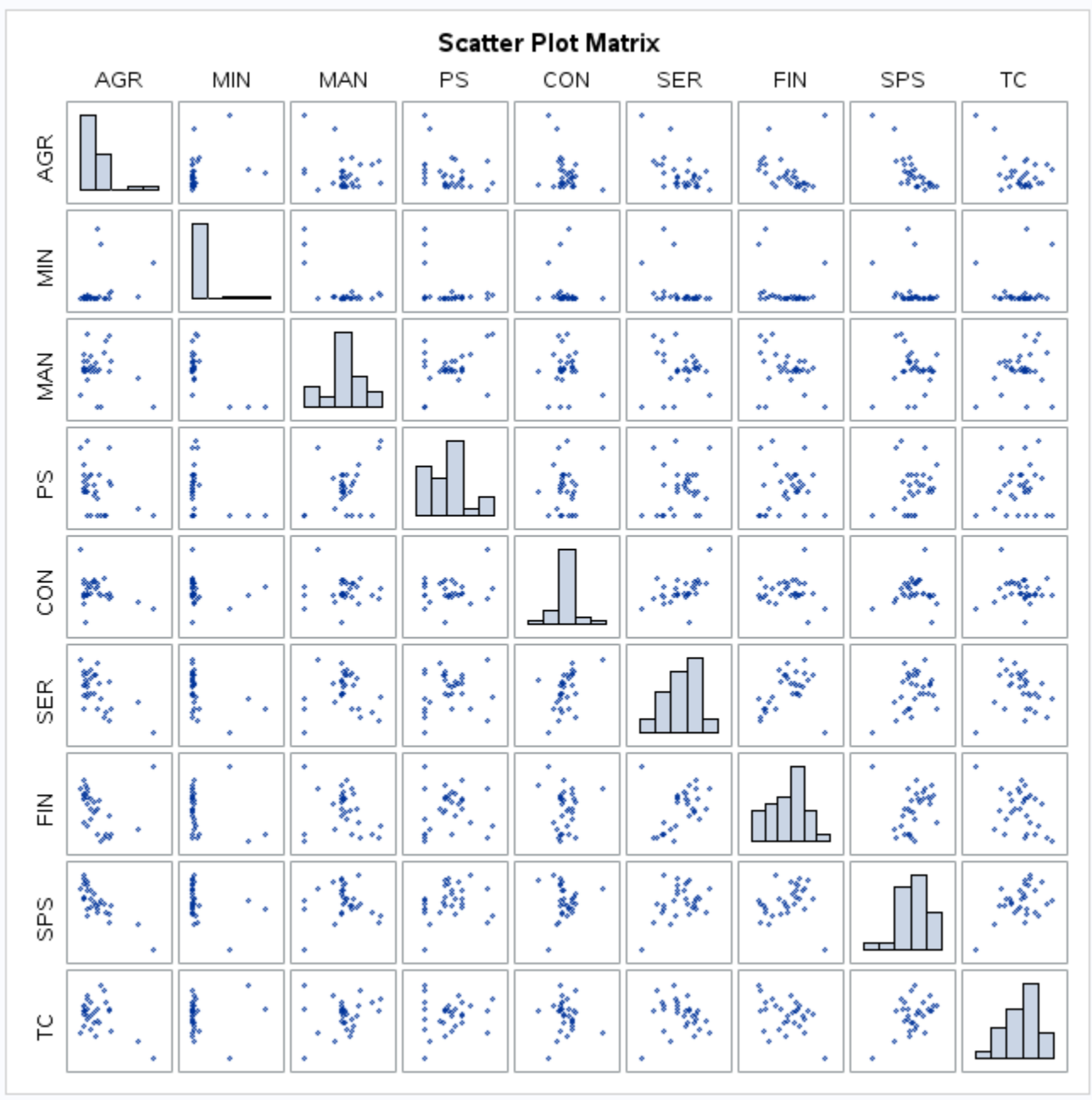

Figure 1: Pearson Correlation Coefficients Scatter Plot Matrix

The Scatter Plot Matrix is demonstrating 9 dimensions showing that some clustering is evident. For example in Figure 1 there is evidence of clustering within the variables SER and FIN. Along the same lines the strong negative linear relationship of AGR and SPS is -.81148, with a p-value of .001. The remaining correlation coefficients are weak to moderate downhill negative linear relationships and weak to moderate uphill positive linear relationships, ~ 30/70 mix.

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

Table 3: Pearson Correlation Coefficients

We observe similar results as above, some clustering with Fin and SER; having a moderate uphill positive linear relationships of 0.37928, and not a strong p-value of 0.0387. The initial results of the initial correlation analysis indicates there are not significant findings within the variable group as indicated by the lack of significant clusters shown in the Pearson Correlation Coefficients, Scatter Plot Matrix, and the Scatterplot of Raw Data: FIN & SER.

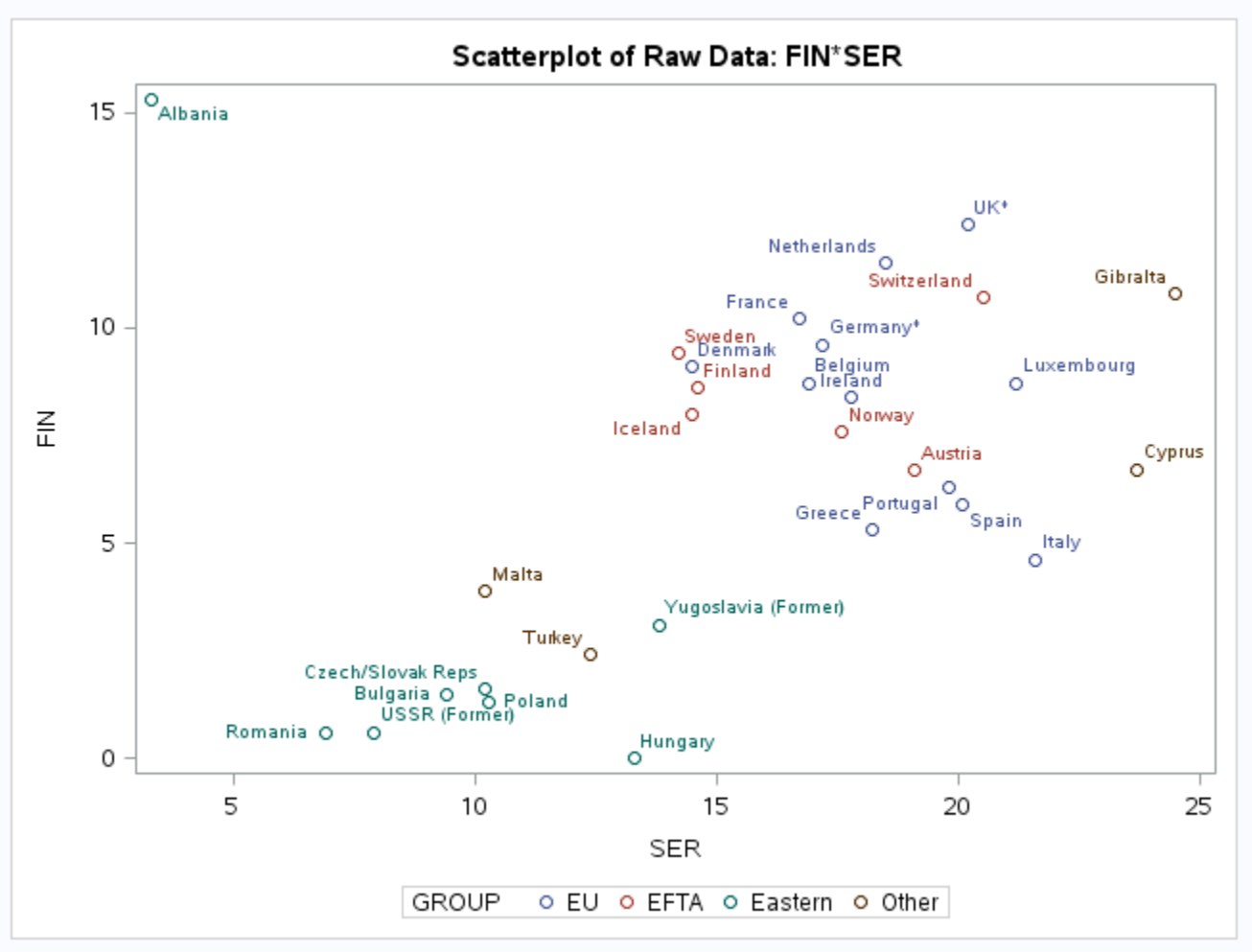

Figure 2: Scatterplot of Variables FIN & SER

If it is necessary to assign the four countries aligned to the "Other" group, based upon the scatterplots and tree analysis, the following alignment is recommended:

Country	Group
Cyprus	EU
Gibralta	EU
Malta	Eastern

Turkey	Eastern

Principal Components Analysis

The Principle Components Analysis with be executed using nine variables. The method for reducing dimensionality is Principal Components Analysis. Based upon the Eigenvalues of the Correlation Matrix, Scree Plot, and the Variance Explained Plot, the 5 principal components are being selected primarily based upon diminishing returns using the "Proportion" column in the Eigenvalues of the Correlation Matrix. After 5 components the proportion is significantly reduced from 0.0789 to 0.0346, as shown in table 4.

	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

Table 4: Eigenvalues of the Correlation Matrix

	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

Table 5: Eigenvectors

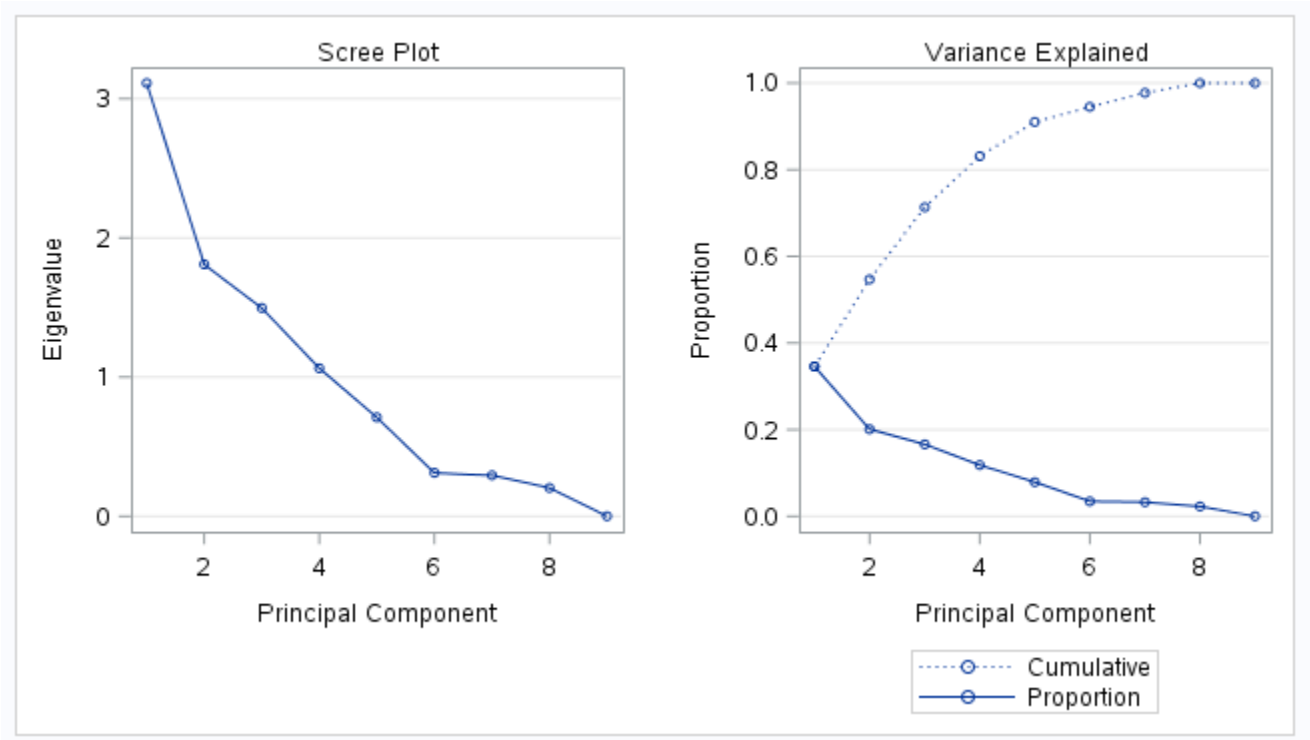

Figure 3: Scree Plot & Variance Explained

Other competing decision rules that could be utilized for determining the quantity of principal components could be a set percentage of the variance explained, for example 80% would be 4 components, and 90% would be 5 components. The Scree Plot and Variance Explained are not showing strong indication as to how many principal components to select, such as an elbow in the plot, shown in Figure 3. Similarly to analyzing Rotated Factor Patterns to define groupings, table 5: Eigenvectors demonstrates the greatest quantity of highest correlated Principal Components are in components 1 thru 5.

Cluster Analysis

The Cluster Analysis is initiated with the Scatterplot Analysis of FIN & SER, and MAN & SER.

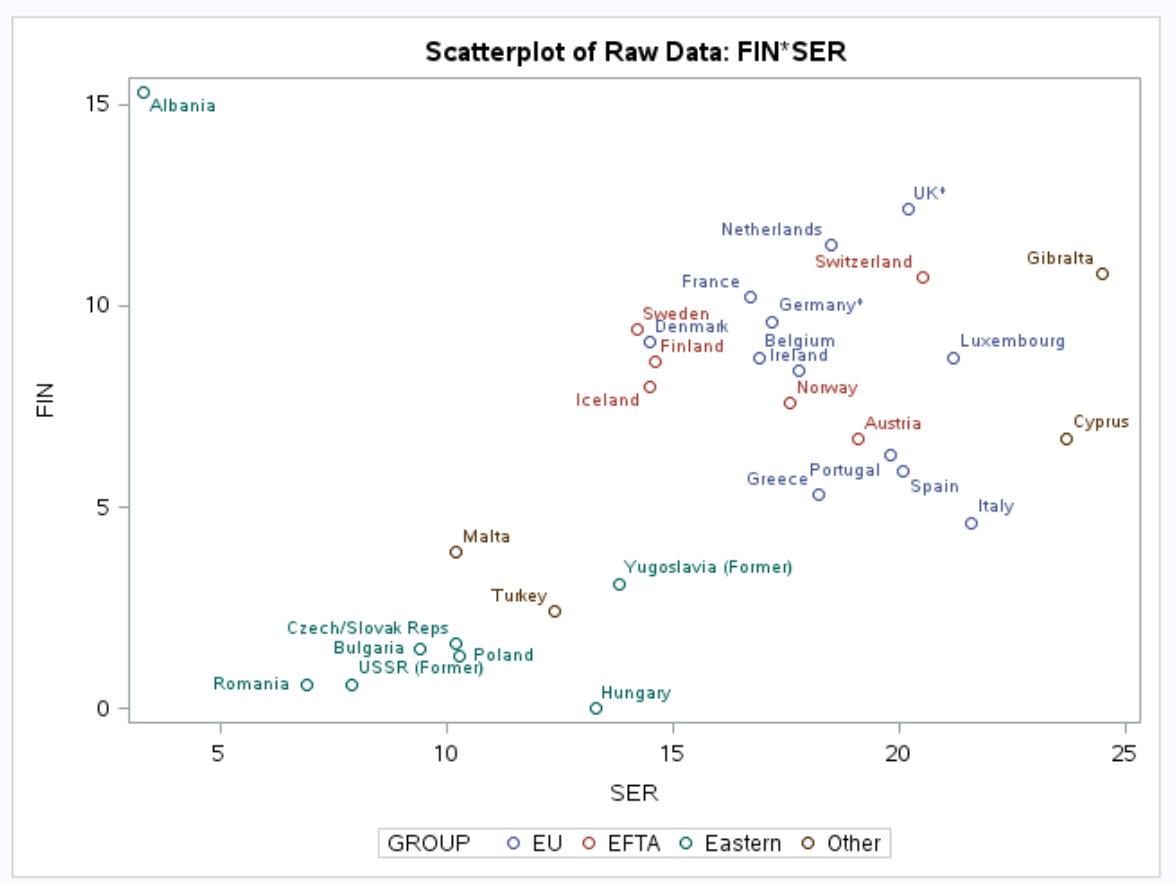

Figure 4: Scatterplot of FIN & SER

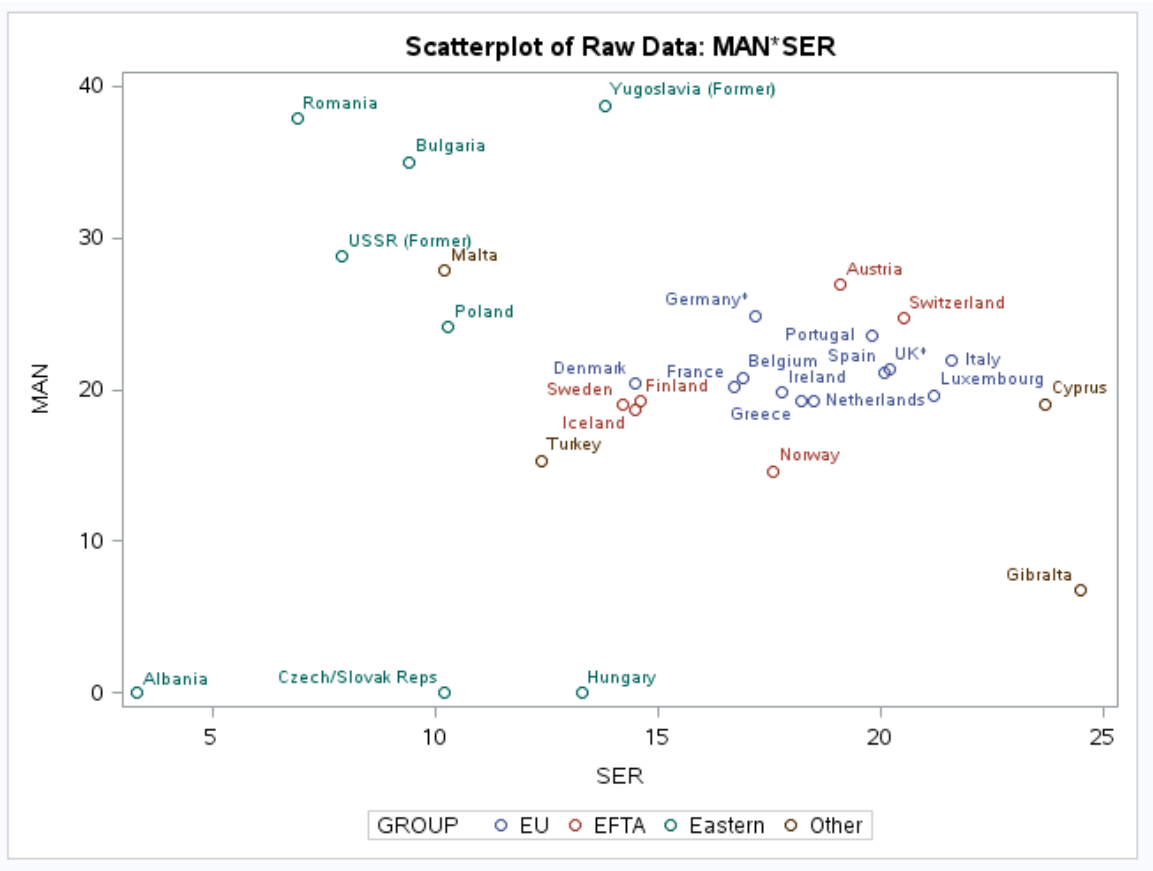

Figure 5: Scatterplot of MAN & SER

The two Scatterplots demonstrate at least two cluster patterns within each plot. The EA and EFTA are both demonstrating clustering; while, Albania and Gibralta are distinct outliers. The next step will be use the SAS PROC CLUSTER to automatically generate clusters using average linkage group average, unweighted pair-group method using arithmetic averages. An important point is that there is no completely ideal method which may be used to determine the number of population clusters within a cluster analysis. We delve deeper into our cluster analysis through interpreting the measure of CCC, Pseudo F, and Pseudo T-Squared.

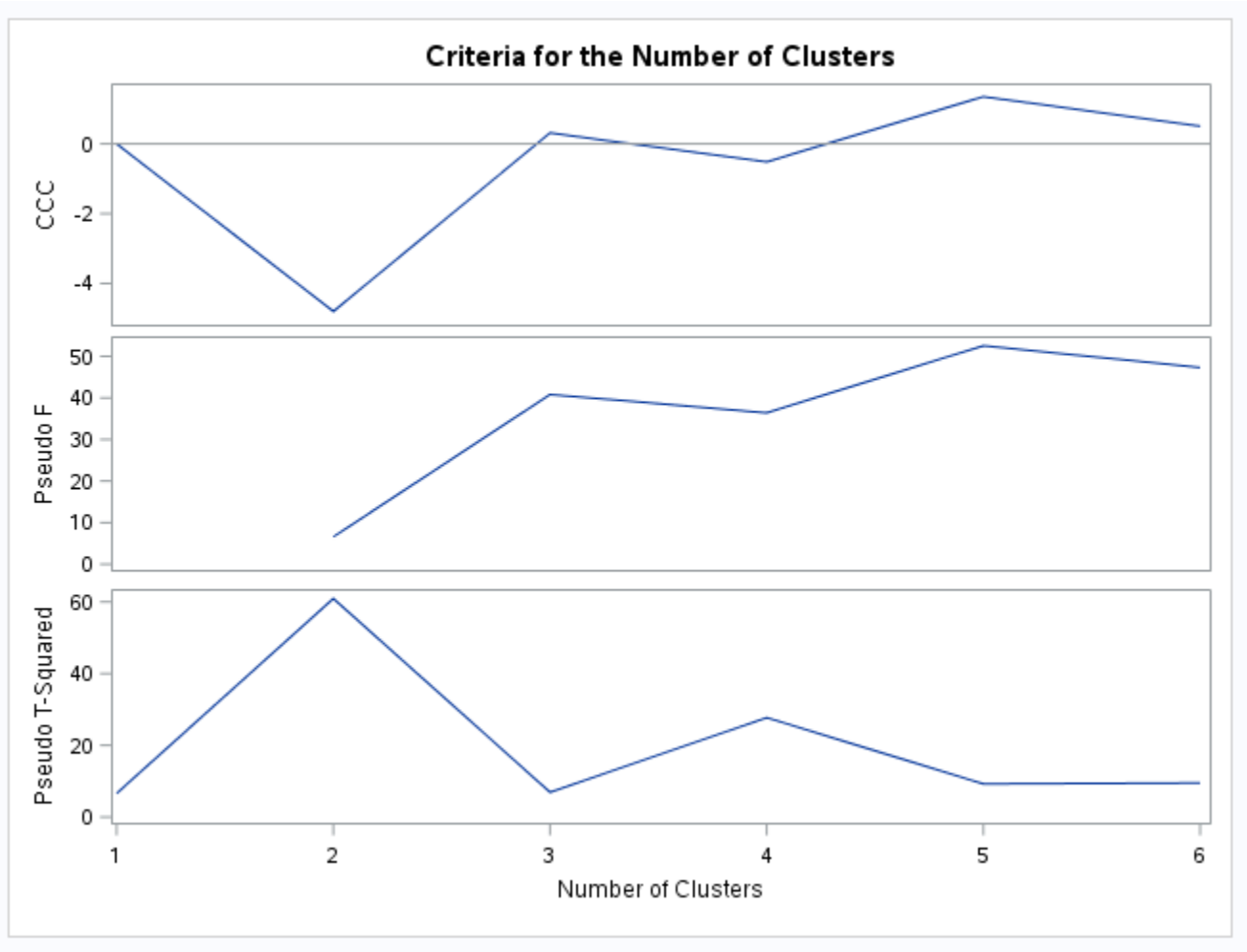

Figure 6: Criteria for the Number of Clusters

Cubic Clustering Criterion:

- Plot peaks that are greater than 2 or 3, generally indicate good clusters.
- The plot may demonstrate several peaks if the data is hierarchical in nature.
- If the plot demonstrates negative values it may be as a result of outliers. Take into consideration outlier removal prior to clustering. [1]

Pseudo F Metrics:

- Seek large values in the plot statistics in the Pseudo F Metrics. [1]

Pseudo T-Statistic:

- Start at the top of the printed output and look for the first relatively large value, then move back up one cluster. [1]

The interpretations from the Cubic Clustering Criterion, Pseudo F Metric, and Pseudo T-Statistic indicate that somewhere between 2 to 3 clusters are evident. The tree procedure provides observations of the of the three and four cluster trees.

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

Table 6: Frequency Group Cluster Three

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

Table 7: Frequency Group Cluster Four

Group	Albania	CL10	CL5	CL6	CL7	Total
EFTA	0	4	1	0	1	6
EU	0	5	2	0	5	12
Eastern	1	0	0	7	0	8
Other	0	0	1	2	1	4
Total	1	9	4	9	7	30

Table 8: Frequency Group Cluster Five

Based upon the observations of three, four, and five frequencies of groups to clusters. It appears that the three cluster frequency table shows that EFTA, EU, and Eastern are forming exclusively into one cluster, CL3. The four and five cluster tables are showing the EFTA and EU are losing some of their members to other clusters. Based upon these findings, three clusters appears to be the best option.

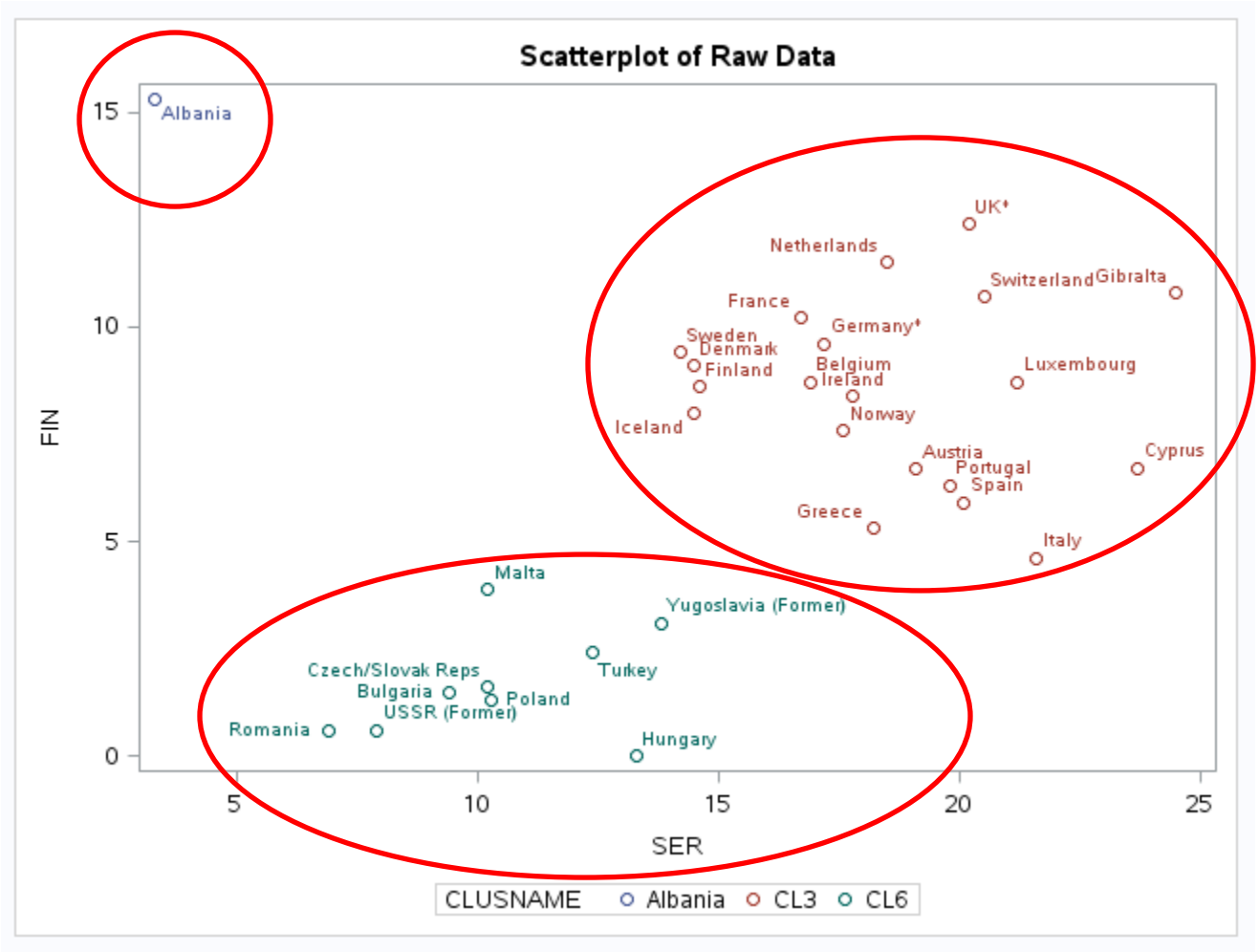

Figure 7: Scatterplot of Three Clusters

Clustering using the first two principal components:

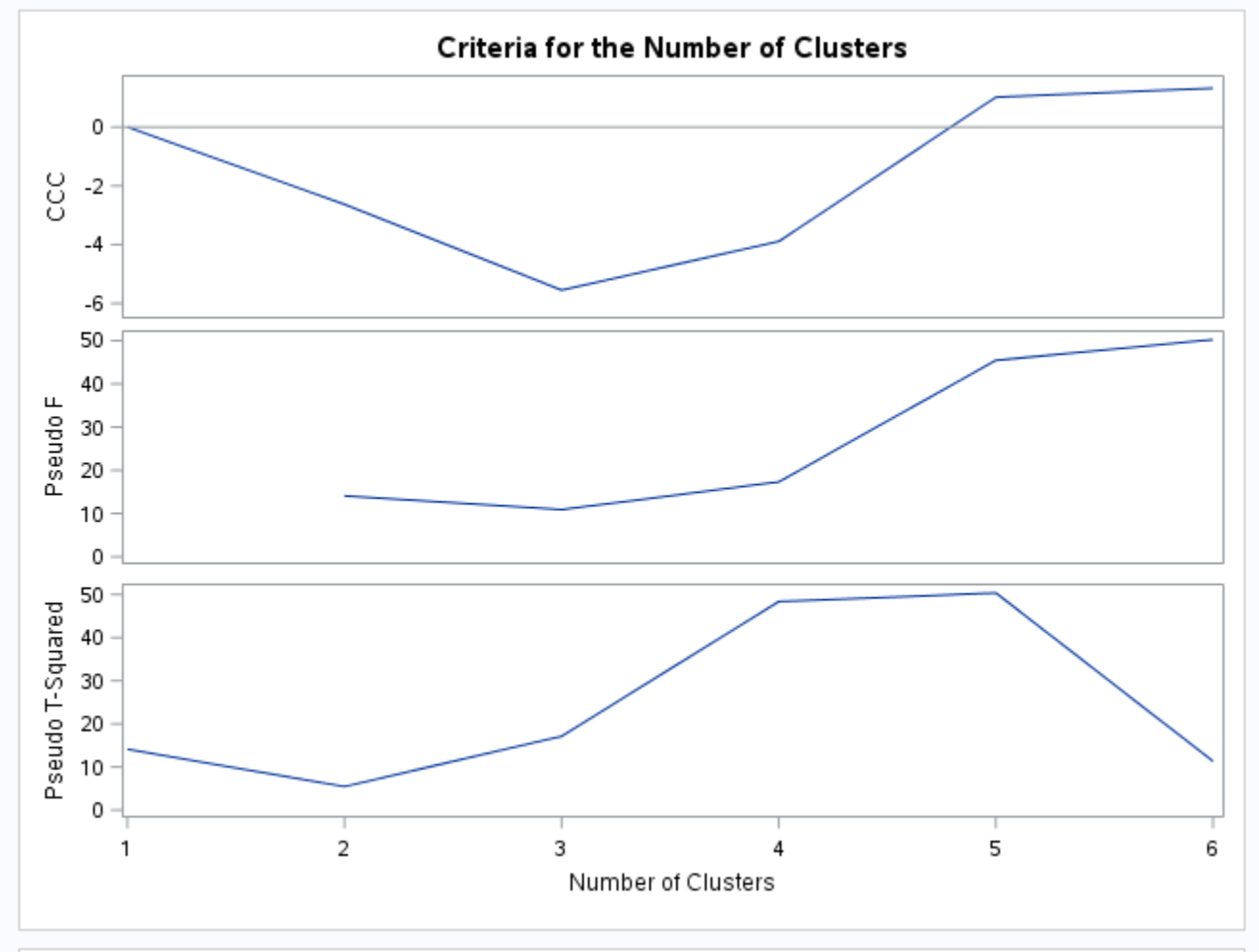

Figure 8: Criteria for the Number of Clusters

Based upon the observations of the above Criteria for the Number of Clusters plot, focusing on the results from CCC and Pseudo T, around 5 clusters would be optimal. Next we will review the observations from the output of the from the cluster trees.

Group	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

Table 9: Frequency Group Cluster Three

Group	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	3

Table 10: Frequency Group Cluster Four

Group	Albania	CL5	CL6	CL7	Gibralta	Total
EFTA	0	0	0	6	0	6
EU	0	0	0	12	0	12
Eastern	1	4	3	0	0	8
Other	0	1	1	1	1	4
Total	1	5	4	19	1	30

Table 11: Frequency Group Cluster Five

Group	Albania	CL11	CL24	CL6	CL7	Gibralta	Total
EFTA	0	0	0	0	6	0	6
EU	0	0	0	0	12	0	12
Eastern	1	3	1	3	0	0	8
Other	0	0	1	1	1	1	4
Total	1	3	2	4	19	1	30

Table 12: Frequency Group Cluster Six

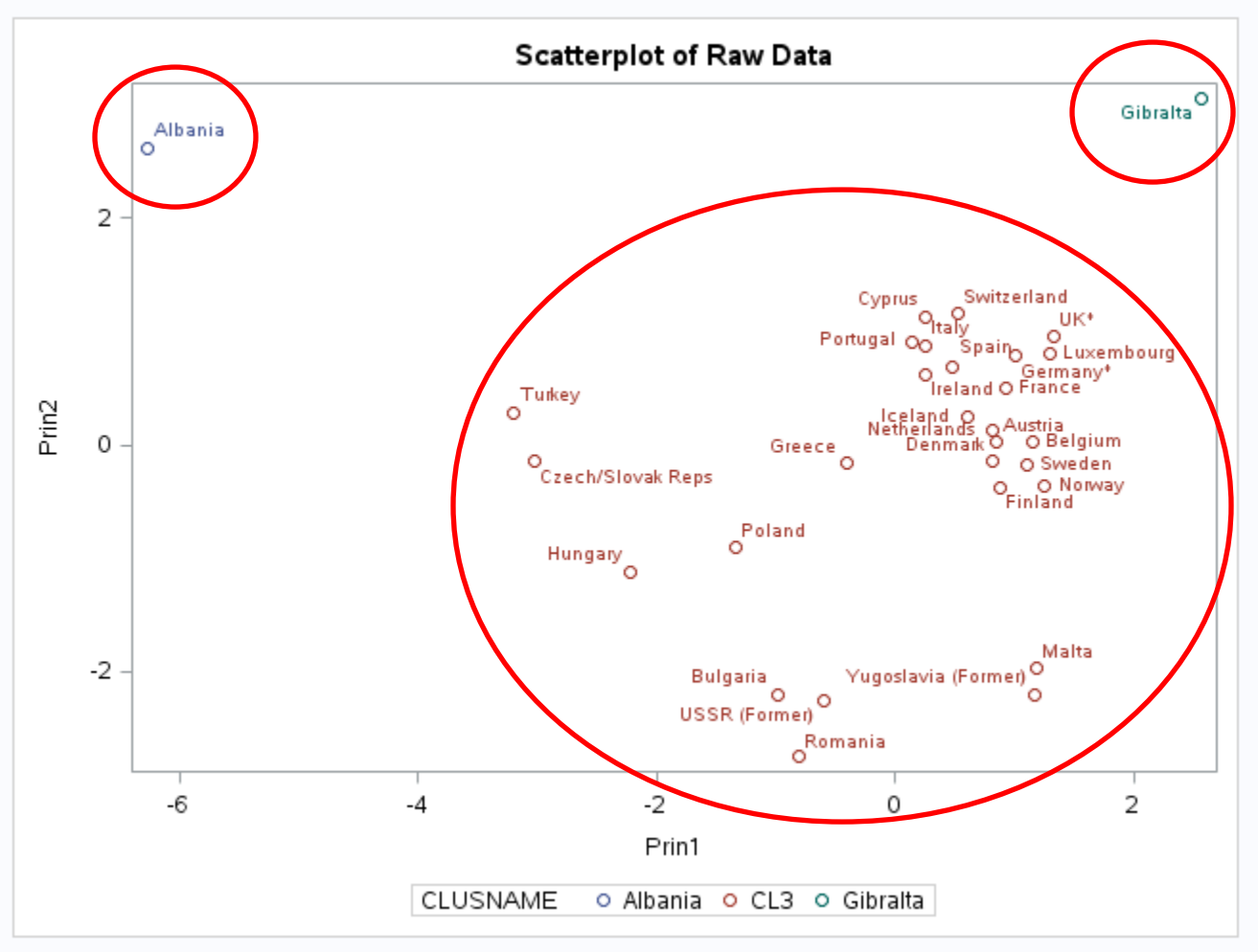

Figure 9: Scatterplot of Three Clusters

Based upon the observations of three, four, five, and six frequencies of groups to clusters. It appears that the three cluster frequency table shows that EFTA, EU, and Eastern are forming exclusively into one cluster, CL3. The four, five, and six cluster tables are showing the EFTA and EU are losing some of their members to other clusters. The principal components data set does not indicate any improvement for cluster analysis.

Conclusion

The recommendation is to use the original data set for cluster analysis. The intended outcome of cluster analysis is to group a set of objects into groups, (clusters) in order to find patterns in data most beneficial in segmentation for marketing purposes. Popular notions of cluster analysis are using patterns associated to some type of distance. Perhaps the European employment data set has inherent clusters built in by design. Meaning the EFTA, EU, and Eastern have existing trade agreements, so that industry sector variables have natural tendency to cluster. The same conclusion holds true for the four countries aligned to the "other" group, they show on the Scatterplots as outliers.

Appendix

1. SAS/STAT 9.2 User's Guide, The CLUSTER Procedure, Page 1235 - 1245

SAS Code

```
* 02.18.2015;
* european_employment;

* Assignment # 8;

%let path=xxx;
libname orion "&path";

data temp; set orion.european_employment;
proc contents; run;
proc print; run;

* Produce the scatterplot matrix;
ods graphics on;
title Correlation Structure of the Raw Data;
proc corr data=temp plot=matrix(histogram nvar=all);
run; quit;
title ;
ods graphics off;

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;


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;

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;

title '';
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;


ods graphics on;
proc tree data=tree1 ncl=5 out=_5_clusters; copy fin ser;
run; quit;
proc print data=_5_clusters; run;
```

```
ods graphics off;
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;

%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;
%makeTable(treeout=_5_clusters,group=temp,outdata=_5_clusters_with_labels);
* Plot the clusters for a visual display; ods graphics on;
proc sgplot data=_5_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=fin x=ser / datalabel=country group=clusname; run;



*****************************************************************;
* 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=6 out=_6_clusters; copy prin1 prin2;
run; quit;
proc tree data=tree3 ncl=5 out=_5_clusters; copy prin1 prin2;
run; quit;
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);
%makeTable(treeout=_5_clusters,group=temp,outdata=_5_clusters_with_labels);
%makeTable(treeout=_6_clusters,group=temp,outdata=_6_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;
* Plot the clusters for a visual display; ods graphics on;
proc sgplot data=_5_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=prin2 x=prin1 / datalabel=country group=clusname; run; quit;
proc sgplot data=_6_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=prin2 x=prin1 / datalabel=country group=clusname; run; quit;
```