Assignment 8

Introduction:

Our data set consists of employment reporting on various industry segments reported as a percent for thirty European nations. In order to perform a cluster analysis, we will first reduce the dimensionality of our data set through utilization of principal components analysis. The variables contained in our data set are listed below:

Table 1: Alphabetic List of Variables and Attributes

Variable	Type	Length	Format	Informat
AGR	Num	8	8.1	F10.1
CON	Num	8	8.1	F10.1
COUNTRY	Char	20	35.	
FIN	Num	8	8.1	F10.1
GROUP	Char	8	10.	
MAN	Num	8	8.1	F10.1
MIN	Num	8	8.1	F10.1
PS	Num	8	8.1	F10.1
SER	Num	8	8.1	F10.1
SPS	Num	8	8.1	F10.1
TC	Num	8	8.1	F10.1

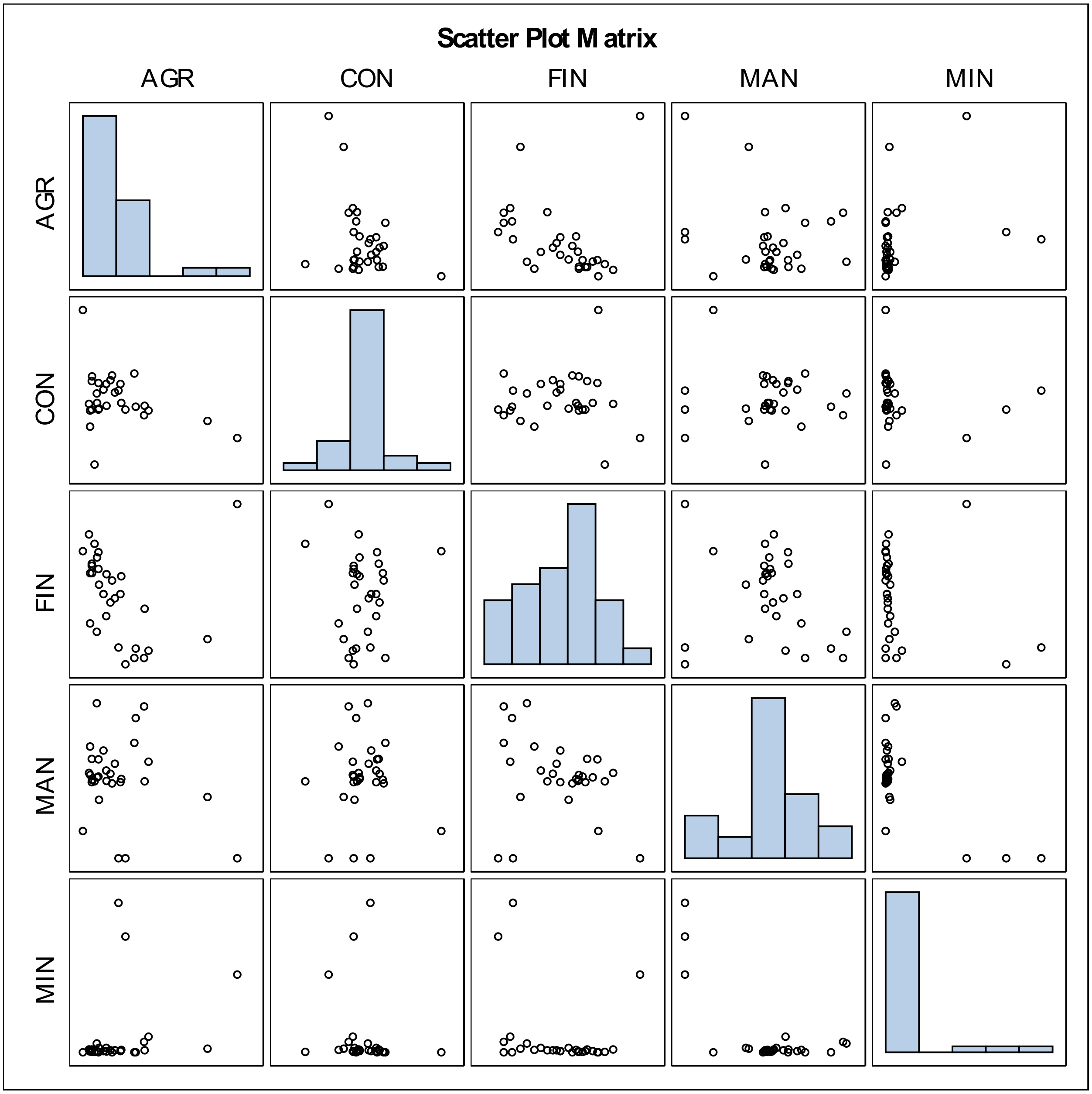
Table 2: Variables listed by Industry Sector

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

Observed Correlation:

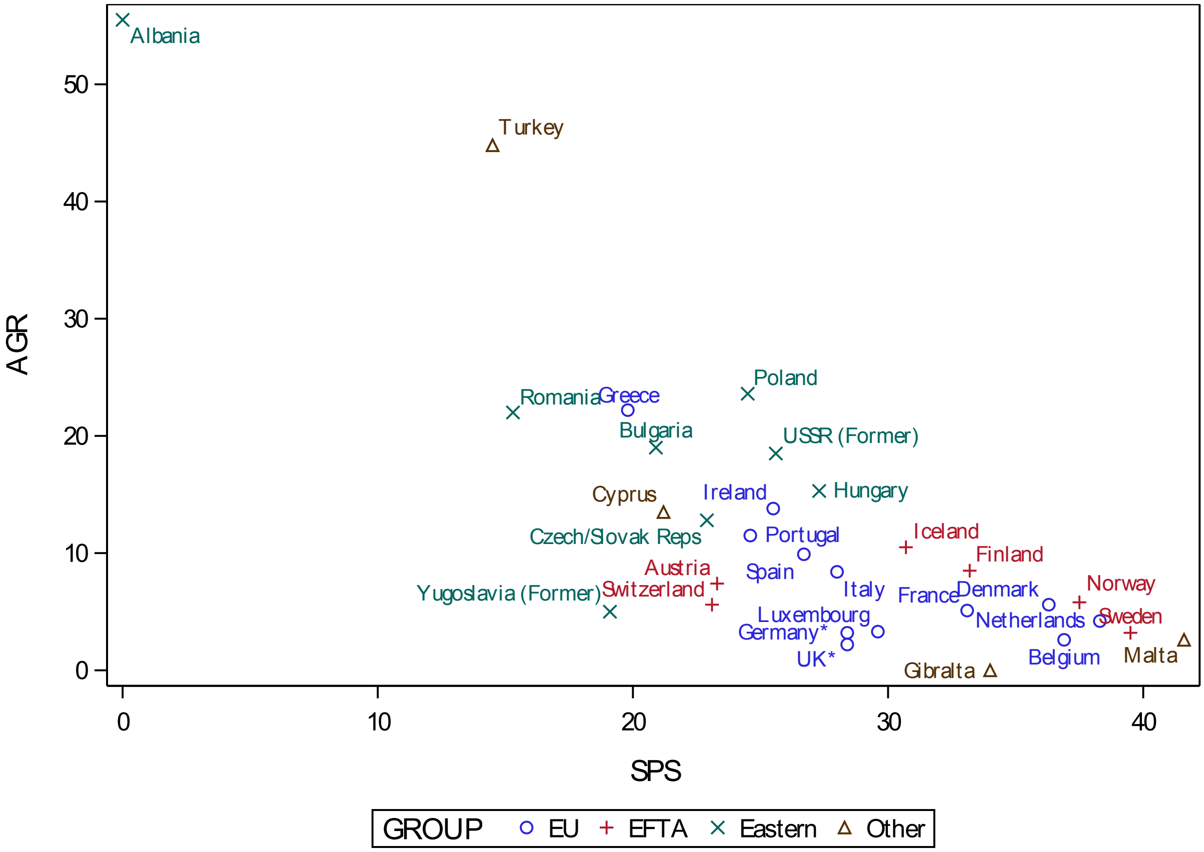
In observing our data set we can see that our variable group provides a subdivision into classes, which segments our list of countries by trade block. This grouping may provide a basis for performing our exploratory data analysis in a supervised fashion. But first, we'll examine the correlation of the variables within our data set which is visualized in the scatter-plot matrix below:

Figure 1: Pearson Correlation Scatter-Plot Matrix



In observing the results in Figure 1 above, the correlation matrix doesn't give us a good indication of which variables are most strongly correlated to each other. The variables that appear to have the strongest correlation are AGR and SPS, so we'll examine this relationship further by producing a scatter-plot. This is presented in Figure 2 below:

Figure 2: Scatterplot of AGR to SPS, colored by Group



Principal Components Analysis:

In order to reduce the dimensionality of our data set, which contains nine variables, we'll use the principal components method. This will allow us to determine the eigenvalues, along with the results from our scree plot, to identify the number of components required that will account of 90% of the variability in our data set.

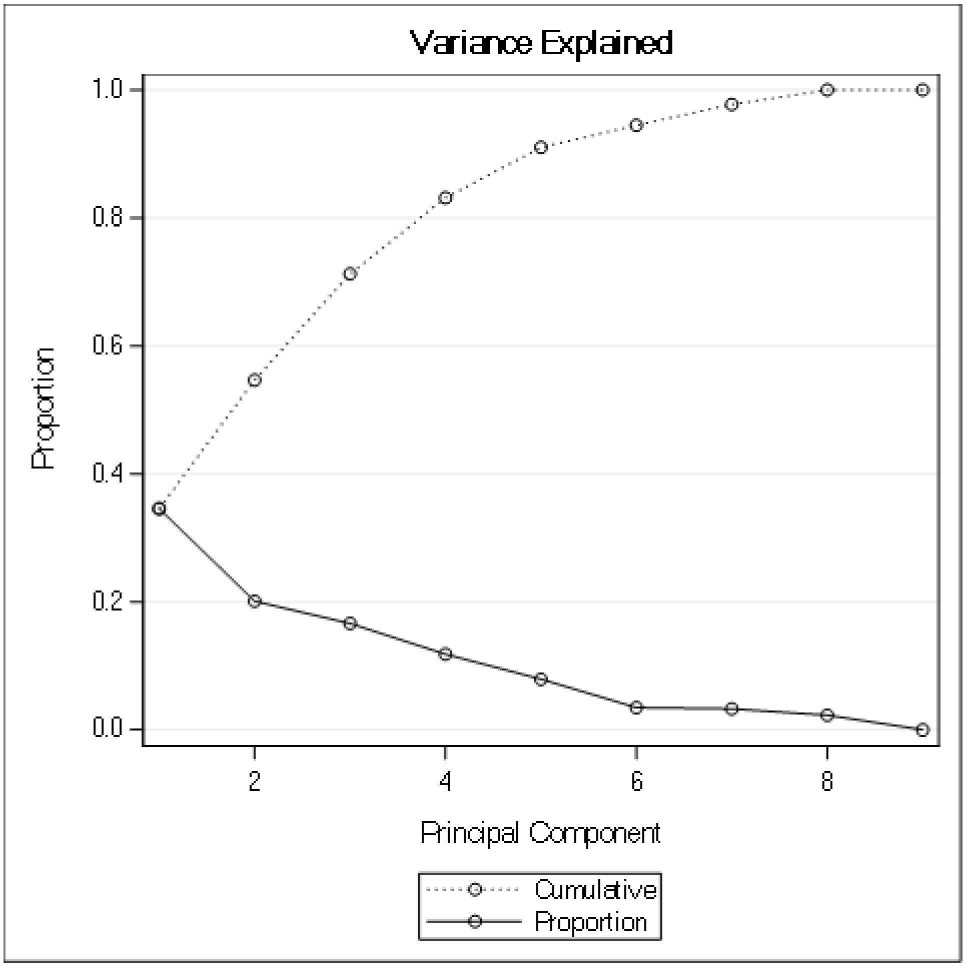
Table 3: Eigenvalues of the Correlation Matrix

Eigenvalues of the Correlation Matrix				
Observation	**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: Eigenvectors

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

Figure 3: Scree Plot of Cumulative Variability



After observing the result from our Principal Components Analysis it appears that the first five variables explain 90% of the variability in our data set. Although it would be ideal to include fewer than five principal components, we had set 90% as our initial threshold. Thus, we'll include the first five variables.

Cluster Analysis:

Now that we have decided on five variables to include in our further exploration of this data set, we will graph scatter plots of the relationship between our FIN and SER variables, as well as our MAN and SER variables.

Figure 4: Cluster Analysis Scatter Plot: FIN by SER

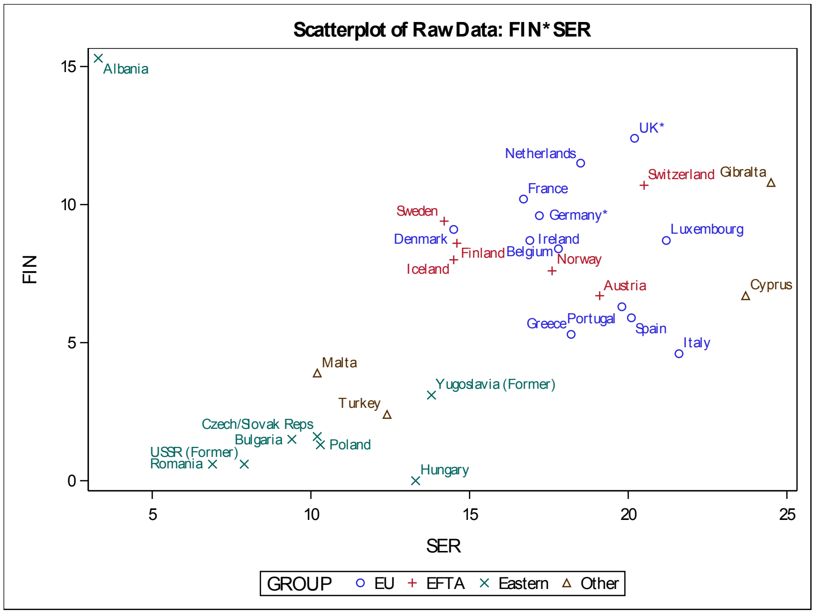


Figure 5: Cluster Analysis: MAN by SER

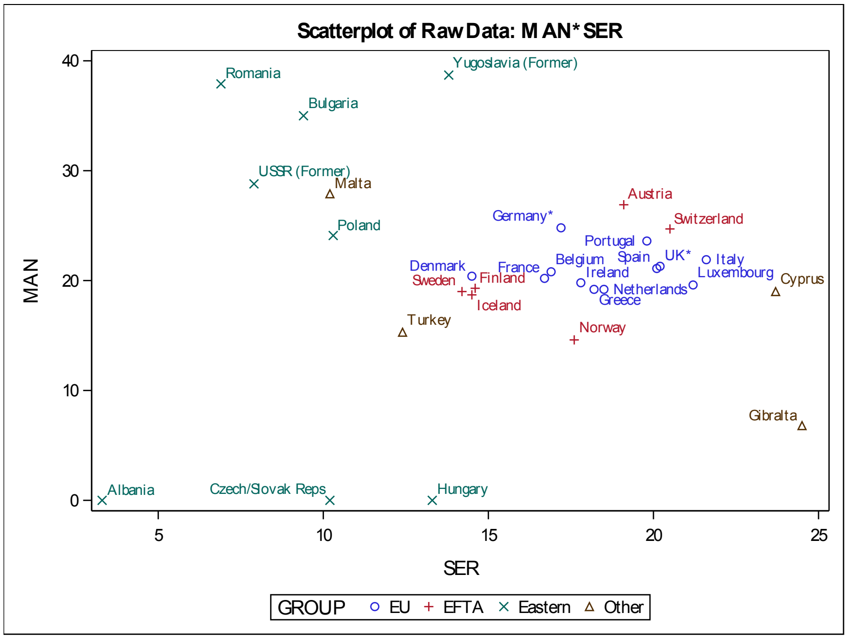
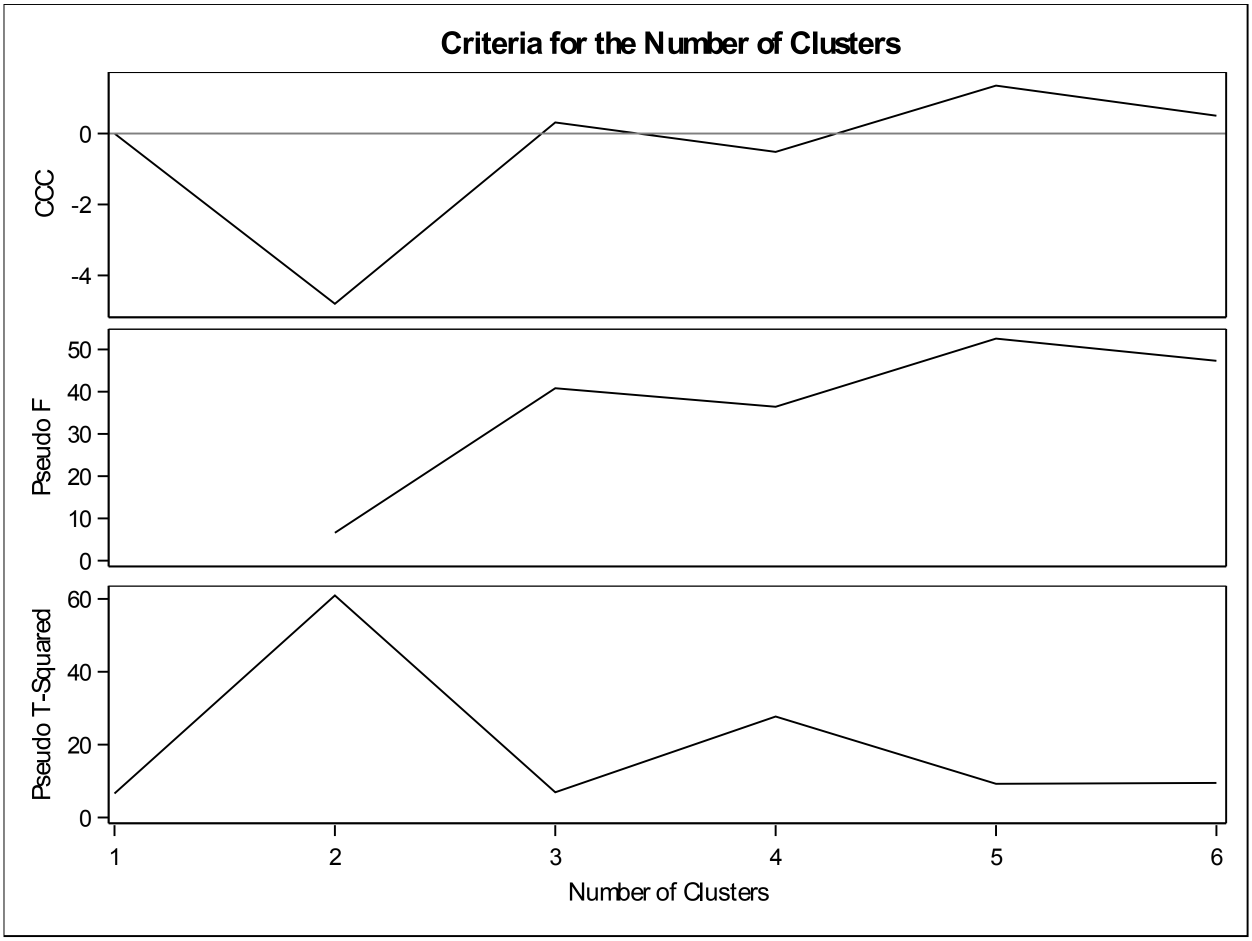


Figure 6: Criterion for Number of Clusters



In interpreting the plots above, we'll utilize the criteria below in further determining the proper number of clusters to include.

The Cubic Clustering Criterion criteria states that the peaks of the plot with the CCC greater than two or three indicate good clustering. This method also tells us that peaks between zero and two are indications of variables that should be removed or interpreted cautiously. Additionally, the CCC method states that very distinct non-hierarchical clusters often show a sharp rise to the proper number of clusters followed by a gradual increase that follows and then eventually declines. Furthermore, the Pseudo F value is relatively large, which tells us that the proper number of clusters for our data set would fall between three or four.

In further determining the proper number of clusters to include in our data set, we'll use the tree procedure to assign each of our observations to a specific number of clusters. We'll examine the output that will allow us to compare the differences between the three cluster tree and four cluster tree below:

Table 5: Frequency of Group to Cluster with Three Clusters

GROUP				
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

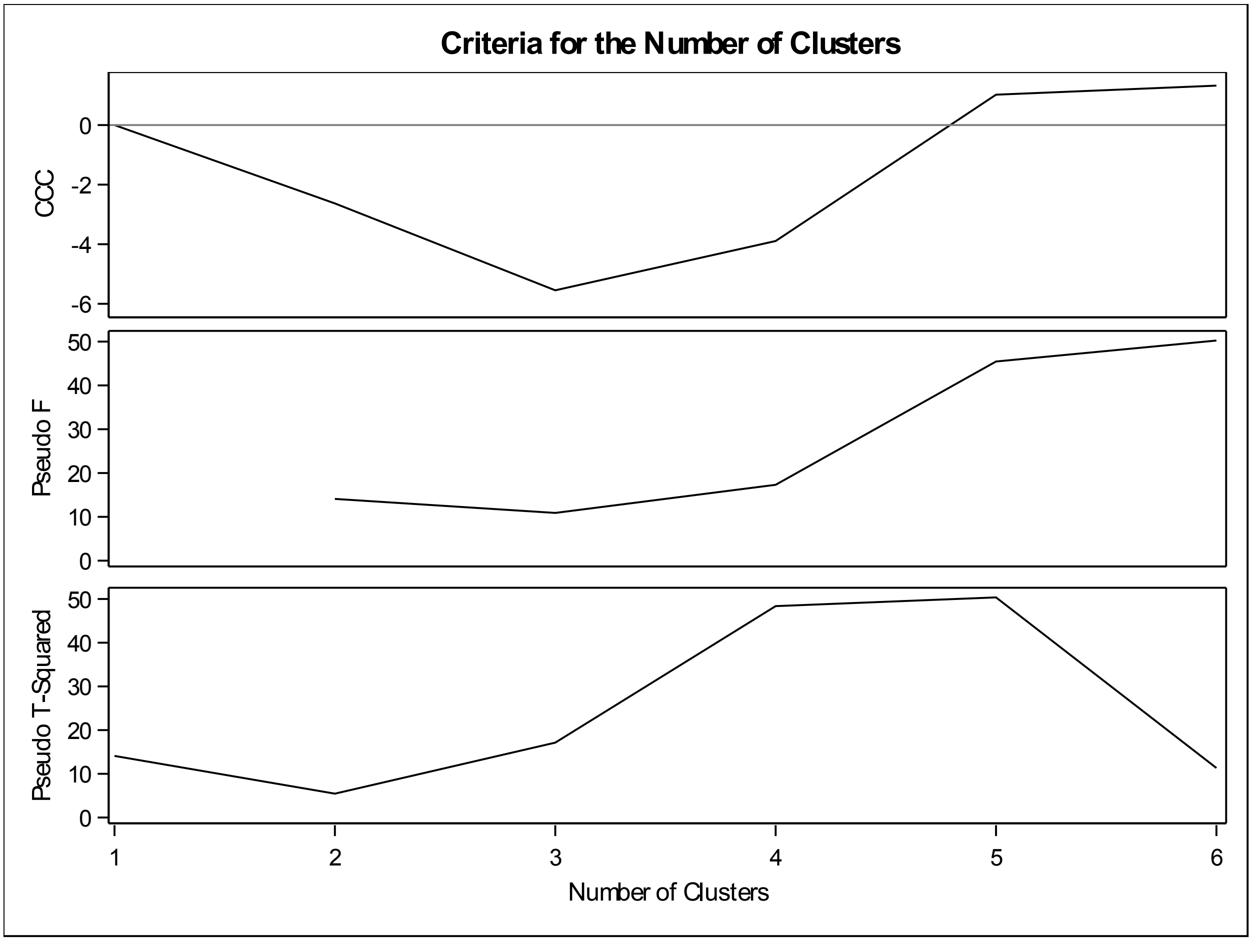
Table 6: Frequency of Group to Cluster with Four Clusters

GROUP					
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

After interpreting the observations in the table above, we can see that the three cluster table tells us that our existing groups are unevenly distributed as most of the observations fall into a single cluster. However, in examining the observations seen in Table 6, we can see that the EFTA and EU observations start to fall within the other cluster groupings.

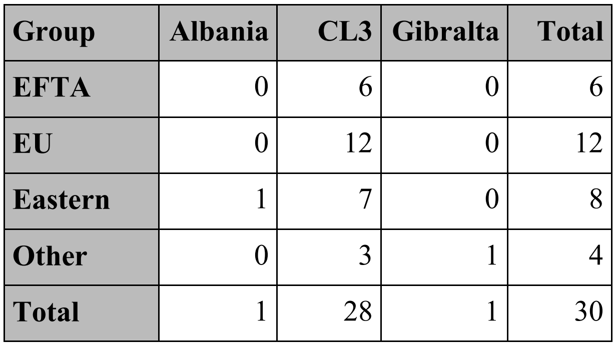
In order to further validate our decision making criterion, we'll perform a hierarchical cluster of our principal components data set. This criterion is presented below in Figure 7.

Figure 7: Criteria for Number of Clusters



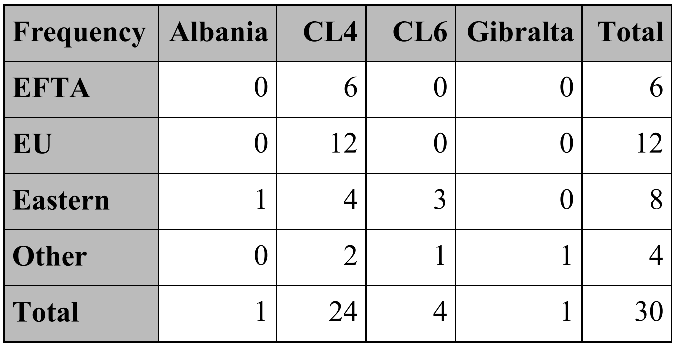
The results of our interpretation from the above visualization would lead us to conclude that the proper number of clusters to include within our data set would be five clusters. In order to explore this finding further, we'll use the tree procedure to assign each of our observations to an assigned number of groupings within our hierarchical clustering. We can examine these differences in Table 7 and Table 8 presented below.

Table 7: Frequency of Group to Cluster with Four Clusters



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 8: Frequency of Group to Cluster with Four Clusters



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

These results illustrate that the groupings are further dispersed. After using the principal components analysis, we can see that this seems to push our groupings toward the outliers within our data set. Thus, we would be better served in using our raw data set for clustering since the PCA data set skews our clusters more so than our raw data set. Although an analyst should take into an account an assumed bias with the predisposed groupings, the results of our analysis strongly reinforce our conclusion that the optimal number of clusters would be found utilizing the raw data set, rather than using the principal components analysis method.

Code:

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


data employ;
	set mydata.european_employment;


proc contents data=employ;


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


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


title 'Modeling the Data, Dimensionality Reduction';


proc princomp data=employ out=employ_prin outstat=eigenvectors plots=scree(unpackpanel);
run;


title 'Cluster Analysis Scatter Plots';


proc sgplot data=employ;
	title 'Scatterplot of Raw Data: FIN*SER';
	scatter y=FIN x=SER / datalab=country group=group;


proc sgplot data=employ;
```

```
        title 'Scatterplot of Raw Data: MAN*SER';
        scatter y=MAN x=SER / datalab=country group=group;

title 'Cluster Analysis: Automated Cluster Selection';
proc cluster data=employ method=average outtree=tree1 psuedo ccc plots=all;
        var FIN SER;
        id country;

proc tree data=tree1 nc1=3 out=_3_clusters;
        title 'Three Cluster Tree';
        copy FIN SER;

proc tree data=tree1 nc1=4 out=_4_clusters;
        title 'Four Cluster Tree';
        copy FIN SER;
run;

%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;

* Call macro function;
%makeTable(treeout=_3_clusters,group=employ,outdata=_3_clusters_with_labels);
%makeTable(treeout=_4_clusters,group=employ,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 'Foud Cluster Tree';
  copy prin1 prin2;

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

run;
ods graphics off
```