Ji Jung
Assignment #8

Introduction

For this assignment, the principal components analysis will be mainly used to reduce dimensionality of the sample data set as a preprocessor for a cluster analysis. The data set is of an employment report for various industry segments, as a percentage measurement, for thirty European nations.

The variables in the data set are:

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 1: Alphabetic List of Variables and Attributes

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

Table 2: Variable and Industrial Sector

It is observed that this data set has a variable group which provides subdivisions into classes. Examining the contents of group tells that the subdivision appears to be by trade bloc. A trade bloc is a type of intergovernmental agreement where regional barriers to trade are reduced or eliminated amongst the participating nation-states. The tutorial requires to examine the data set in an unsupervised fashion. This group classification would provide a basis to perform an exploratory data analysis in a supervised fashion.

Four countries (Cyprus, Gibralta, Matla, and Turkey) are all within the other group. It could be assumed that the group assignment was purely a basis of local, hence the re-naming this category into something more contextually appropriate, such as Mediterranean. This could be misleading as several nation-states from the EU would be in the Mediterranean region.

Part 1: An Initial Correlation Analysis

First is to examine the simple Pearson correlation for the variables within the data set as this produces the scatter-plot matrix:

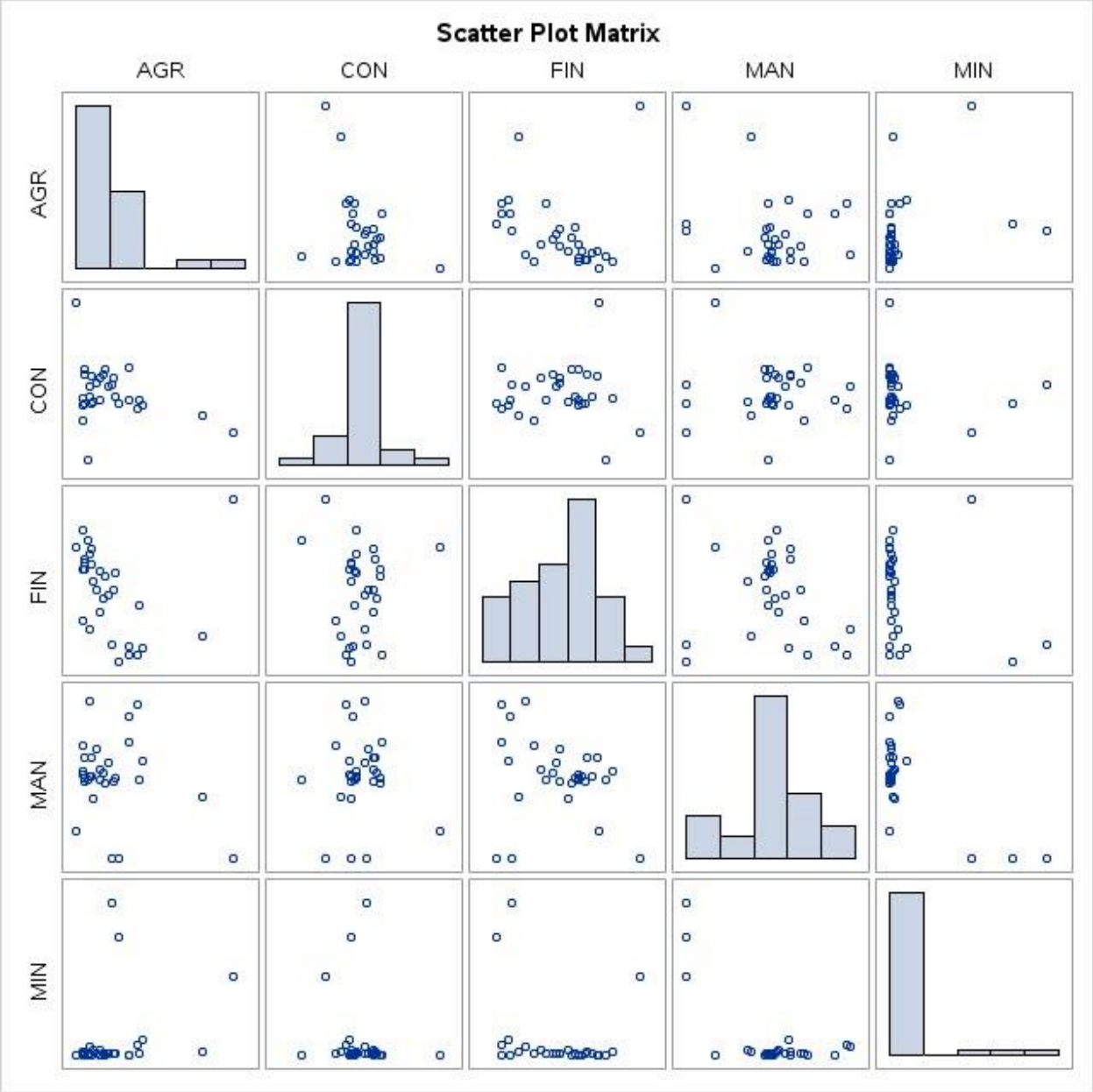


Figure 1: Pearson Correlation Scatter-Plot Matrix

This graphic doesn't fully encompass the correlation matrix. The strongest correlation, with a statistically significant test, is between AGR and SPS. The correlation being 0.81148 with a probability > $|r|$ under $H_0 : \rho = 0$ test statistic of < 0.0001. Next step incurs the examination of this correlation by producing a scatter-plot:

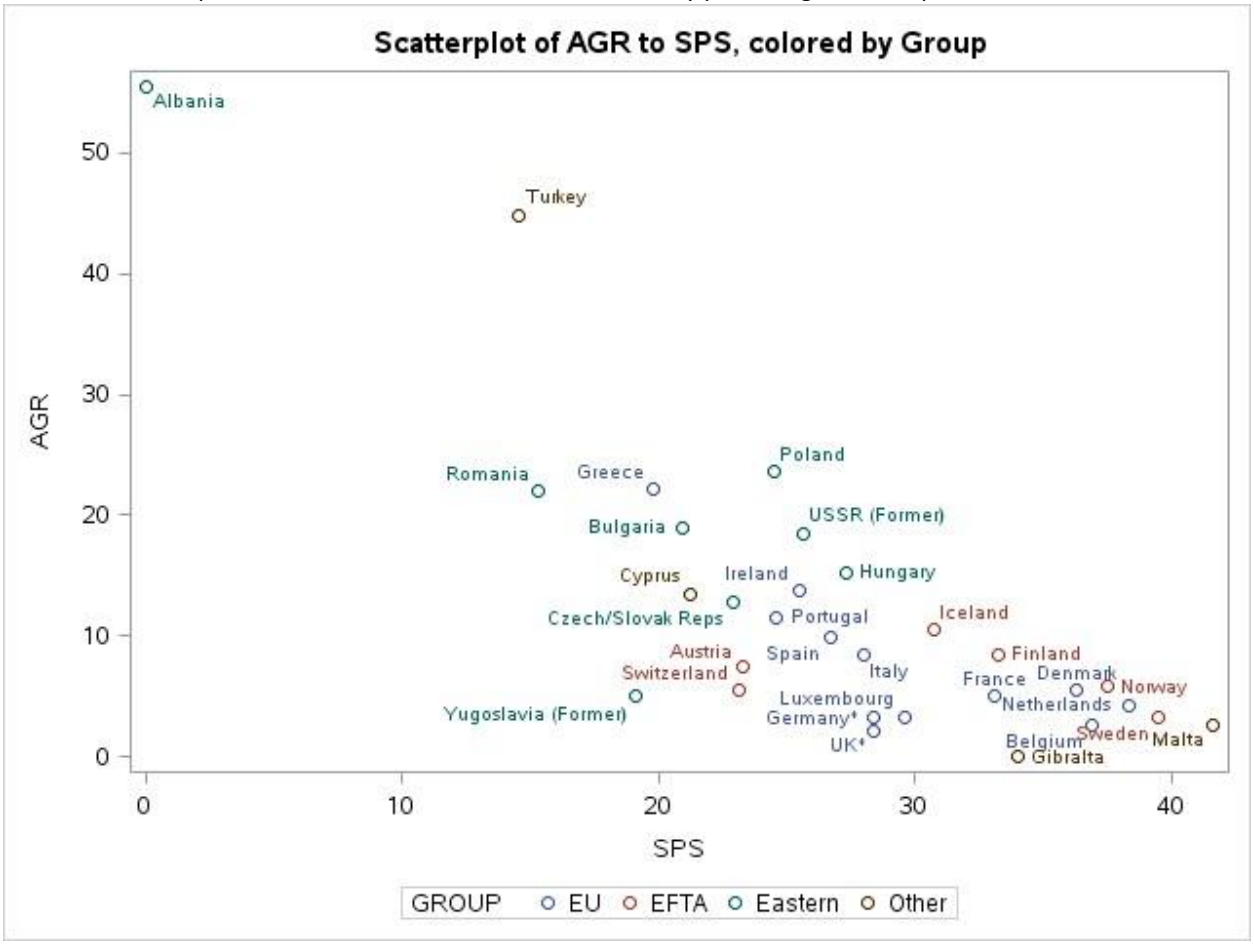


Figure 2: Scatterplot of AGR to SPS, colored by Group

Part 2: Principal Components Analysis

As there are nine variables within the data set, PCA will be used as a dimensionality reduction method. The variability table, or scree plot examines the number of components required to account for 90% of the data variability.

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 3: 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 4: Eigenvectors

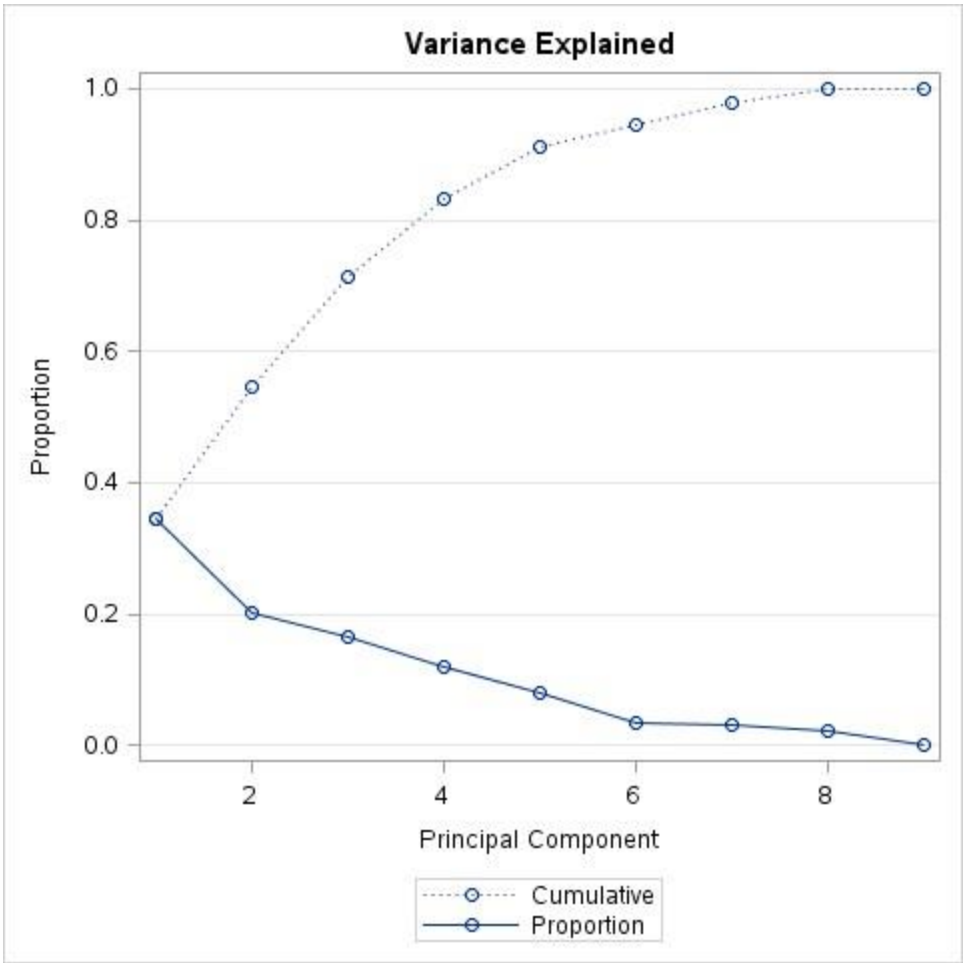
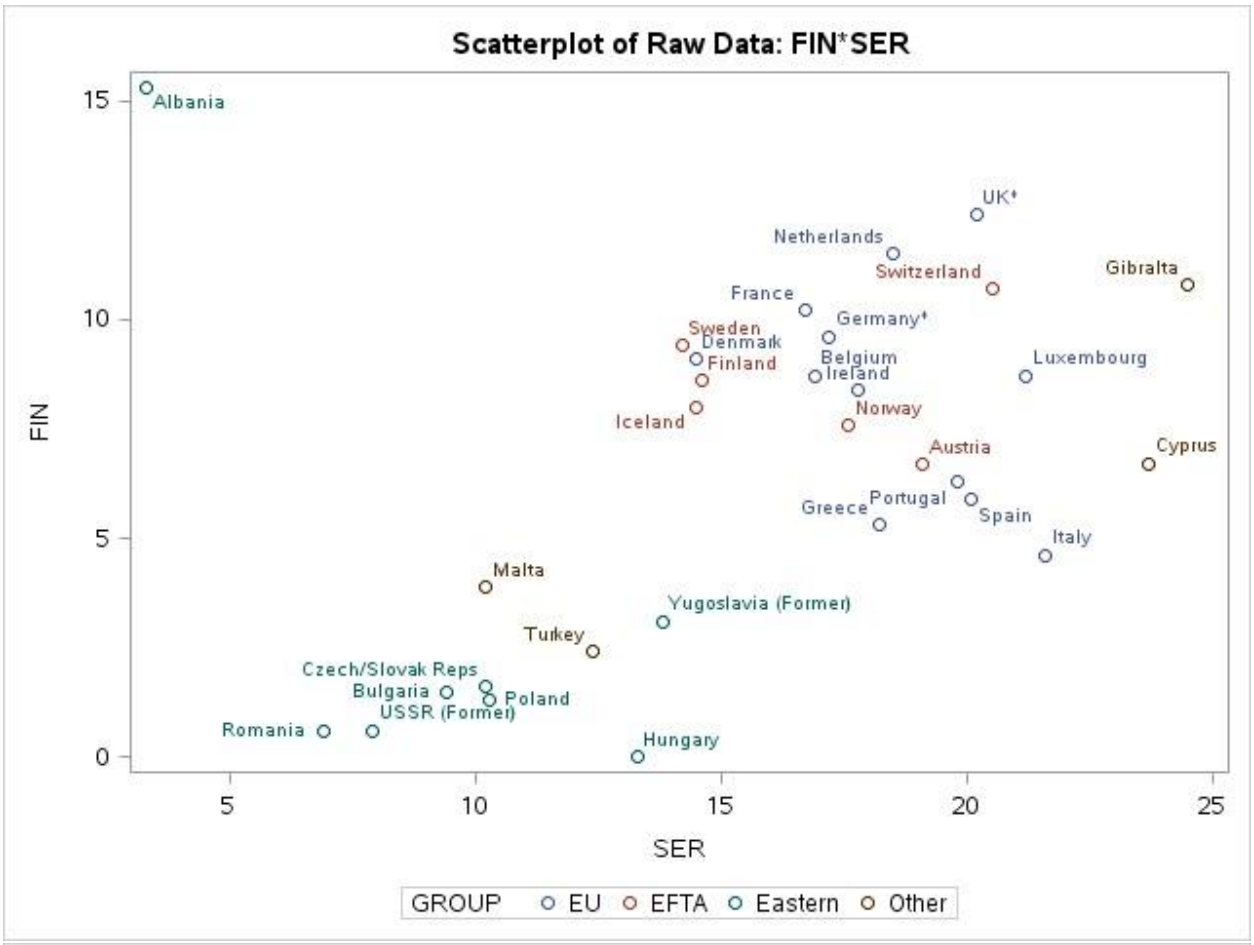


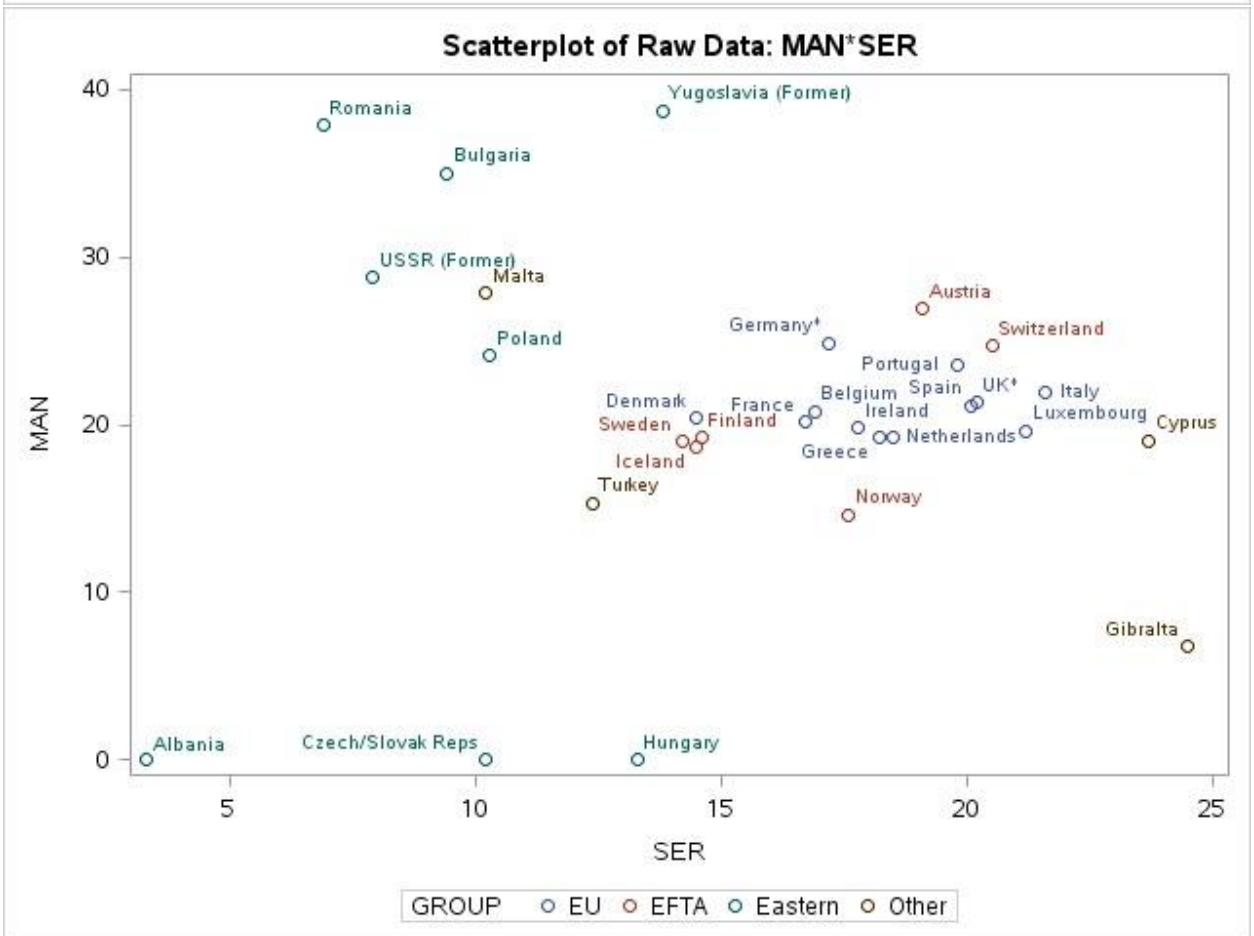
Figure 3: Scree Cumulative Variability

From the diagnostic output of the PCA procedure, it can be reasoned to use the first five principal components to explain greater than 90% of the variability within the data. It can be difficult to accept this as the initial decision, prior to examining output, to take forward 90% of the explained variability. It would be more ideal from an ease-of-modeling perspective to take forward fewer principal components, and in turn less of the explained variability.

Part 3: Cluster Analysis

For next step, scatter plots will be created including selected the FIN and SER, as well as the MAN and SER variables:


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


Scatterplot of Raw Data: MAN*SER
MAN
SER
40
30
20
10
0
5
10
15
20
25
Romania
Yugoslavia (Former)
Bulgaria
USSR (Former)
Malta
Austria
Germany*
Switzerland
Poland
Portugal
Spain
UK*
Italy
Denmark
France
Belgium
Ireland
Luxembourg
Sweden
Finland
Netherlands
Cyprus
Greece
Iceland
Turkey
Norway
Gibralta
Albania
Czech/Slovak Reps
Hungary
GROUP EU EFTA Eastern Other

From the onset, two clusters can be found within each of the graphs. There are some outlier countries, but in both graphs the Eastern group seems to cluster, whereas the EU and EFTA groups seem to cluster. Albania and Gibraltar both seem to be outliers. With the cluster procedure within SAS, clusters with a hierarchical approach can be automatically created. As this is a hierarchal approach, the number of desired clusters does not need to be specified. Instead, a diagnostic output for different criteria can be decided. There are no completely satisfactory methods that can be used for determining the number of population clusters for any type of cluster analysis. The diagnostic output of the cluster procedure will be examined closely, and look for the Cubic Clustering Criterion (CCC), Pseudo F, and the Pseudo T-Squared:

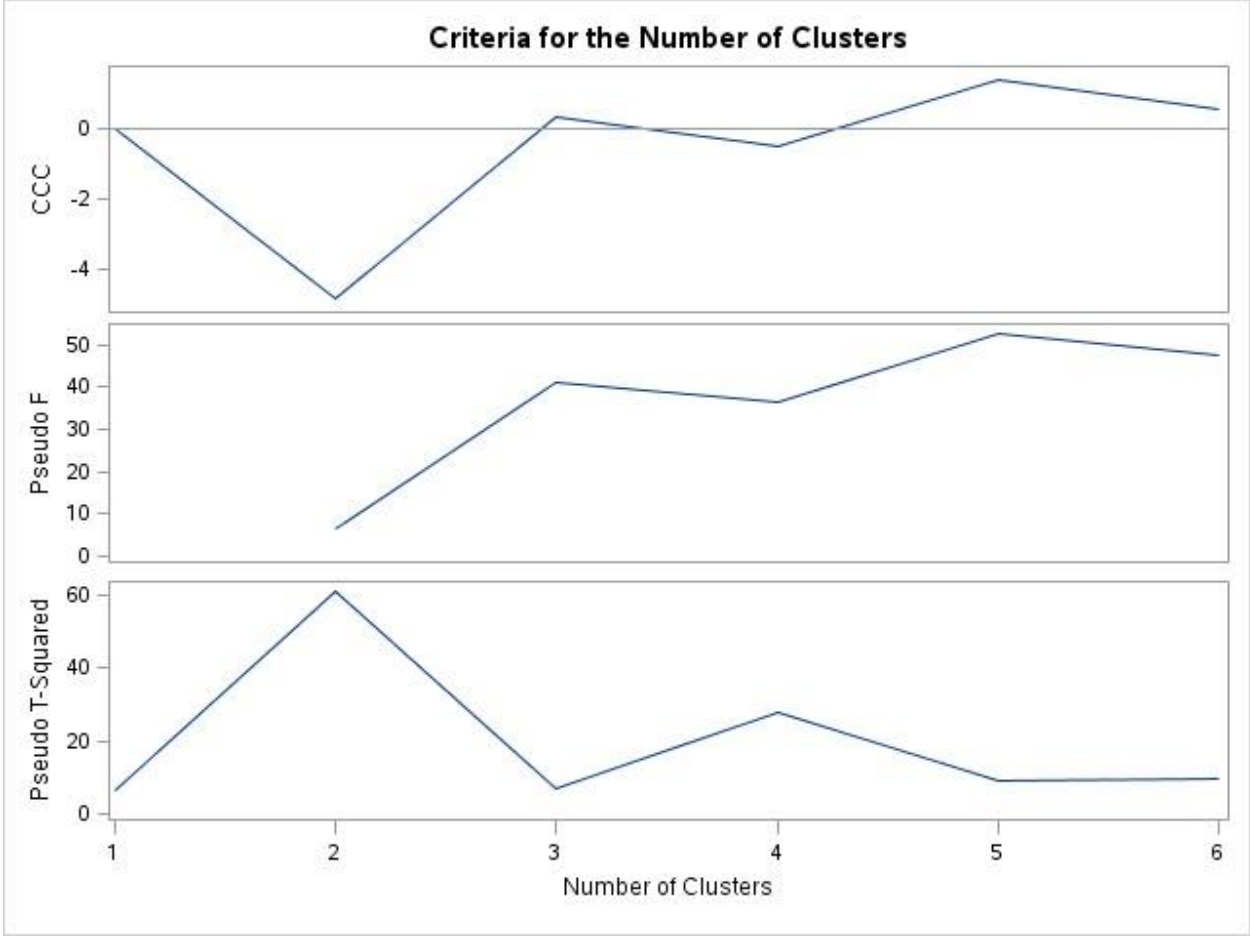


Figure 4: Criteria for Number of Clusters

Interpret the measurements will be done with the following assumptions, if the criterions are all graphed in relation the number of clusters (as it is above):

- Cubic Clustering Criterion
  - Peaks on the plot with the CCC greater than 2 or 3 indicate good clustering.
  - Peaks with the 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.
  - Very distinct non-hierarchical spherical clusters usually show a sharp rise before the peak followed by a gradual decline.
  - Very distinct non-hierarchical elliptical clusters often show a sharp rise to the correct number of clusters followed by a further gradual increase and eventually a gradual decline.

  - If all values of the CCC are negative and decreasing for two or more clusters, the distribution is probably unimodal or long-tailed.
  - Very negative values of the CCC, say, -30, may be due to outliers. Outliers generally should be removed before clustering.

- Pseudo F
  - Look for a relatively large value.

It can be concluded (between the CCC and Pseudo F) that three or four clusters will be optimal.

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

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 5: Frequency of Group to Cluster with Three Clusters

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 6: Frequency of Group to Cluster with Four Clusters

This membership group is observed that, for this data set, is a coherent 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, EFTA and EU give up some of their members to be distributed amongst other clusters. For simplicity, and to reinforce the contextual information within the data set, three-clusters is preferred.
The hierarchical clustering with the principal components data set will be performed next.

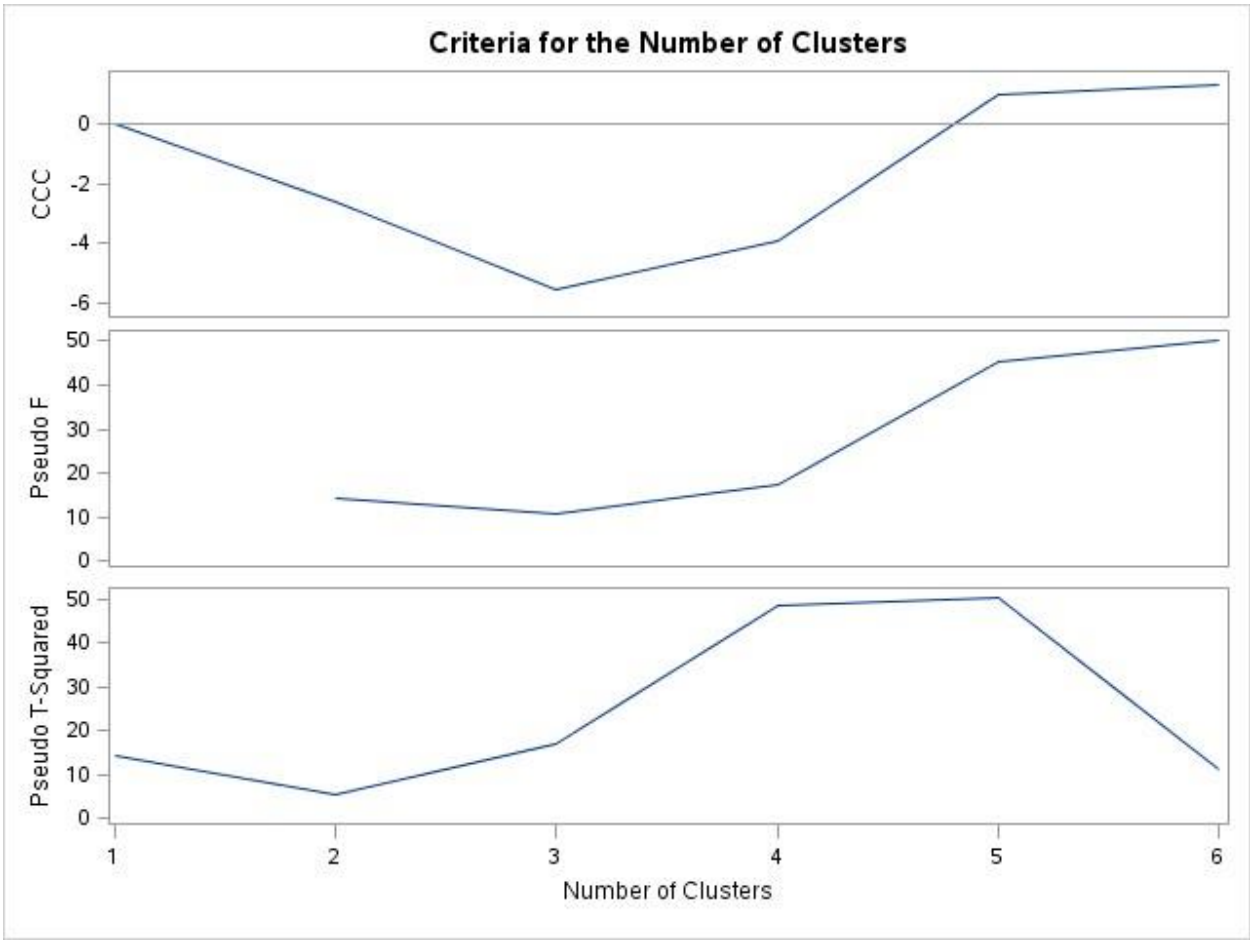


Figure 5: Criteria for Number of Clusters

Using the above assumptions for interpreting the criteria, it is concluded (between the CCC and Pseudo F) that at least five, maybe six clusters will be satisfactory.

From the tree procedure, it is to assign observations to a specified number of clusters after the hierarchal clustering. The tabular output between the three-cluster tree and four cluster tree is examined:

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 7: Frequency of Group to Cluster with Three Clusters

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	30

Table 8: Frequency of Group to Cluster with Four Clusters

The membership group breaks down a bit. Using the principal components data set seems to have pushed the clustering towards accentuating the outlier members within the data. It is preferable to use the raw data for clustering. If the goal of cluster analysis is to group objects that are more similar, which is based on some distance methodology, it would seem to be more useful to have a better distribution of entities into the respective clusters. With the principal components data set, it can be reasoned that the clusters have a more skewed amount of membership, as opposed to the raw data set.

However, it might be possible that an assumed bias exists when working with data that already has an indication of group membership. In the case of this analysis, there is some reinforcement of that group membership in the initial clustering. This may have been enough to reinforce analyst bias to an undoubted level.