Assignment #8
Allison Roeser

Introduction:

The report details a Principal Component Analysis and Custer Analysis study of European employment data. The dataset of thirty countries includes the percentage of European workers that are employed in nine industries. The countries are grouped into four groups: European Union, European Free Trade Association, Eastern European nations, and Other.

First step in the analysis was to perform a correlation study to determine how closely the employment proportions by industry were correlated with one another. Moderate positive and negative correlations were both found in the dataset. The next step was to perform a Principal Component Analysis to reduce the dimensions of the dataset and eliminate multicollinearity. Because the data was only moderately correlated, four Principal Components were required to explain greater than 80% of the variance. The first two Principal Components explained 55% of the variance. Using the first two Principal Components, a Cluster Analysis study was performed using three, four, and five clusters. The results were analyzed to determine the optimal number of clusters. Natural clustering patterns in the data were also visually identified using scatterplots.

Results:

Part 1:

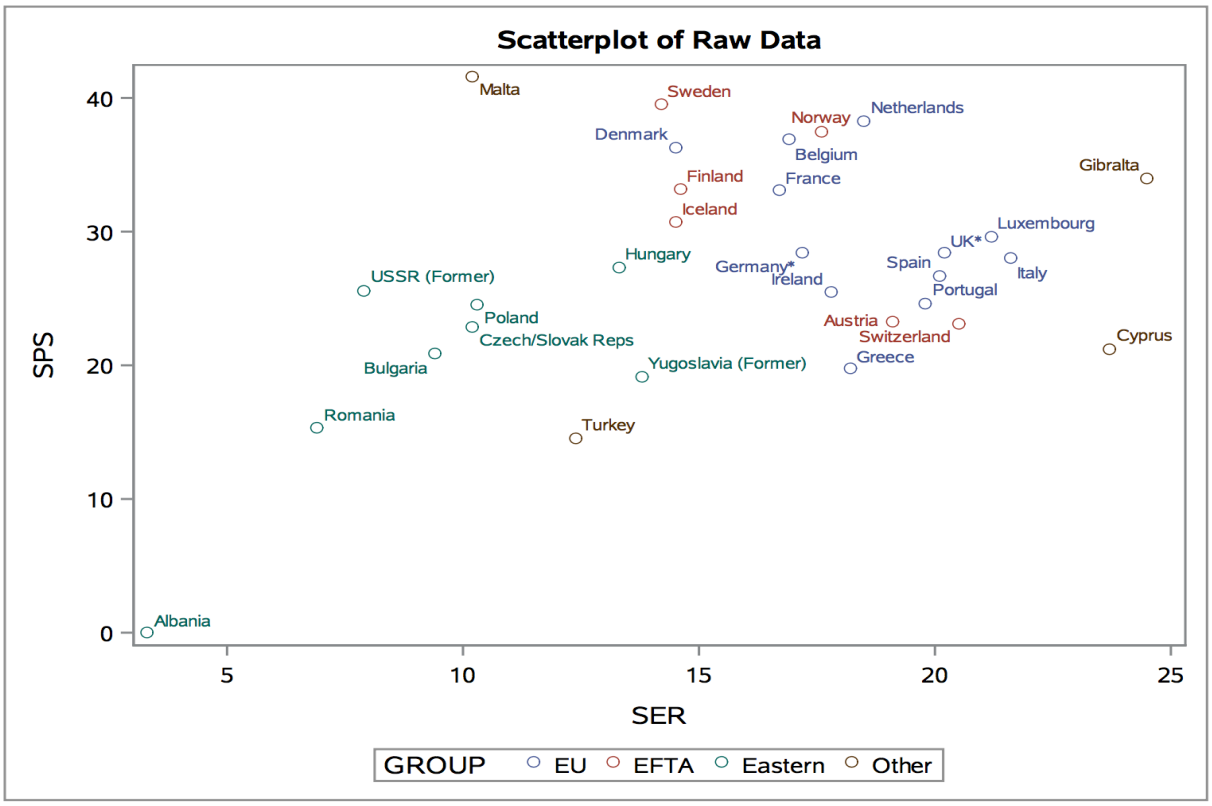
Below is the Pearson Correlation Coefficients table for the nine industries. SPS (Social and Personal Services) and TC (Transport and Communication) are the most positively correlated with a correlation coefficient of 0.475. SPS an AGR (Agriculture) are most negatively correlated with a correlation coefficient of -0.811. The majority of the industries are only moderately correlated. Principal Component Analysis will be used later on to reduce the dimensions of the dataset. PCA is most effective at reducing dimensions when there are high correlations between the variables.

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

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

The Scatterplot matrix below visually displays the relationships between the nine variables. The scatterplot of SPS (Social and Personal Services) vs. SER (services) displays some interesting clustering of points that may be beneficial to investigate further. The correlation coefficient between these two variables is 0.388.

This scatterplot of SPS vs SER is color coded by European grouping. The Eastern group in green forms a good cluster. The other three groups (EU, EFTA, and other) show significant overlap. There may be other underlying factors besides European country group that is causing clustering to occur.

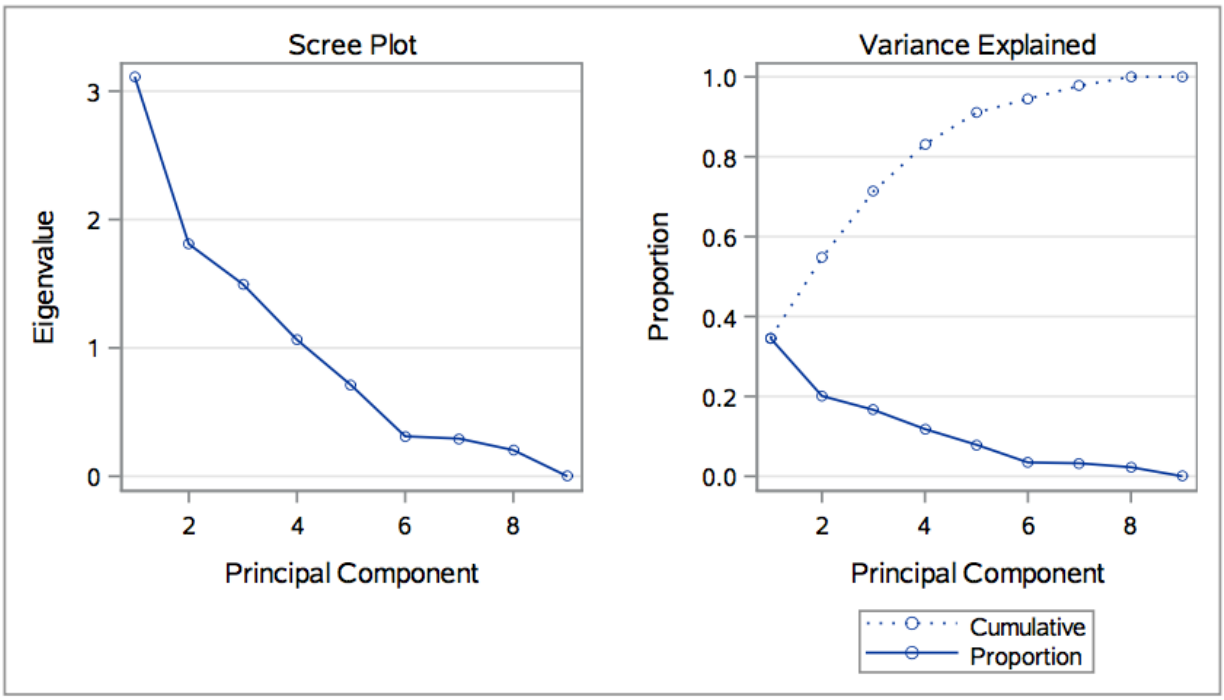


Part 2:

The next step was to perform Principal Component Analysis to reduce the dimensions of the dataset. As mentioned above, the variables are only moderately correlated, which can make PCA less effective. The Eigenvalue results are listed below. The first Principal Component explains 34.6% of the variance. The second Principal Component explains another 20.1% for a cumulative percentage of 54.7%. Four principal components are required to reach the 80% threshold. Using four Principal Components explains 83.1% of the common variance.

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

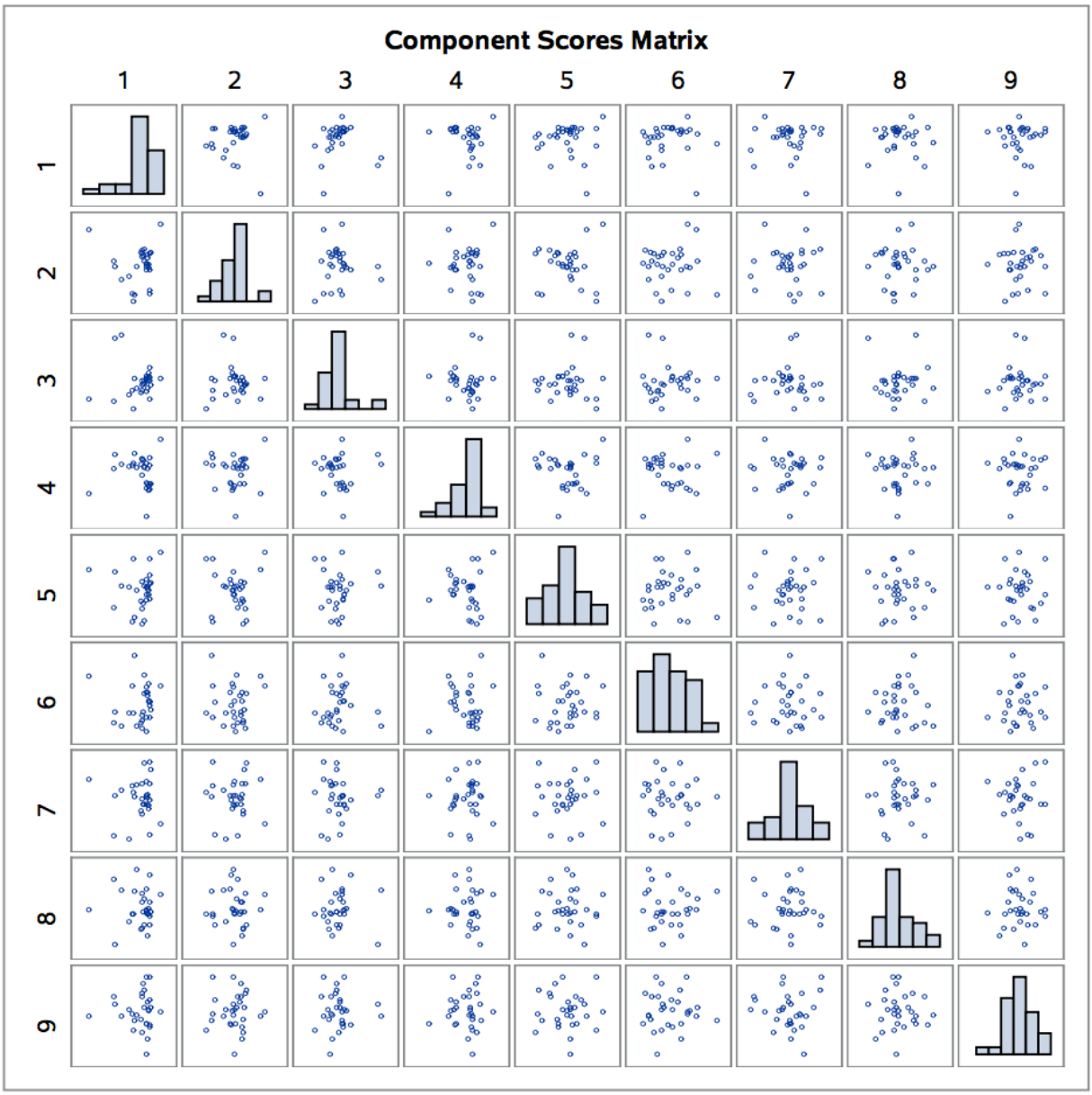
The Scree Plot shows that there are elbows at the points of two and six Principal Components. The Variance Explained plot shows that the Cumulative Variance climbs steadily until the point of about six Principal Components, at which point the graph begins to level off.



The scatterplot matrix below shows the scatterplots of the nine Principal Components plotted against one another. The purpose of the scatterplot matrix of Principal Components is to look for clustering. There is more clustering in the low valued Principal Components (#1, #2, and #3) than in the scatterplots of the higher valued components.

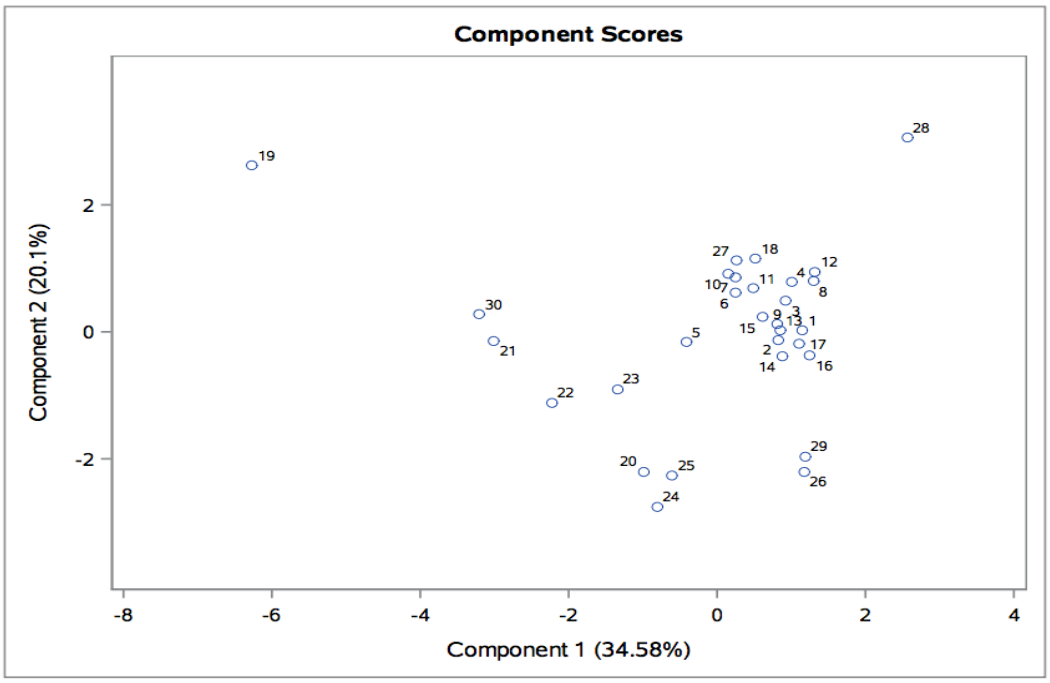
Principal Components Analysis using PROC PRINCOMP

The PRINCOMP Procedure



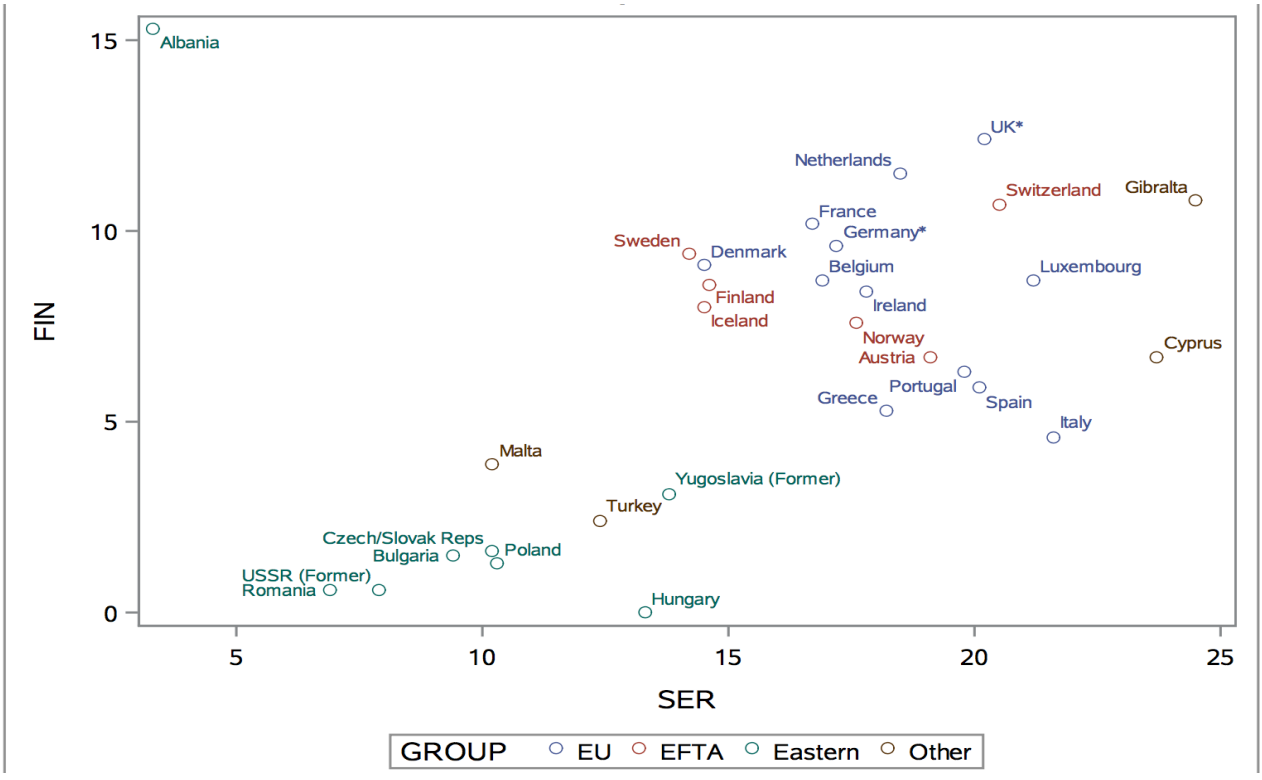
Below is a scatterplot of the first two Principal Components. Principal Component 1 explains 34.6% of the total variance for the dataset and Principal Component 2 explains another 20.1%. By design, these two Principal Components are uncorrelated. The Principal Components are projected into two-dimensional space. If clustering is present in the two-dimensional space it will be present in the nine-dimensional space. Clustering can in fact be seen in the scatterplot. There is a large cluster within the circle bounded by Principal Component Scores 27, 12, 14, and 16.

Anywhere from two to six Principal Components could reasonably be retained. For the purposes of this study, two Principal Components were retained. The clustering present in the scatterplot was be further investigated using Cluster Analysis.

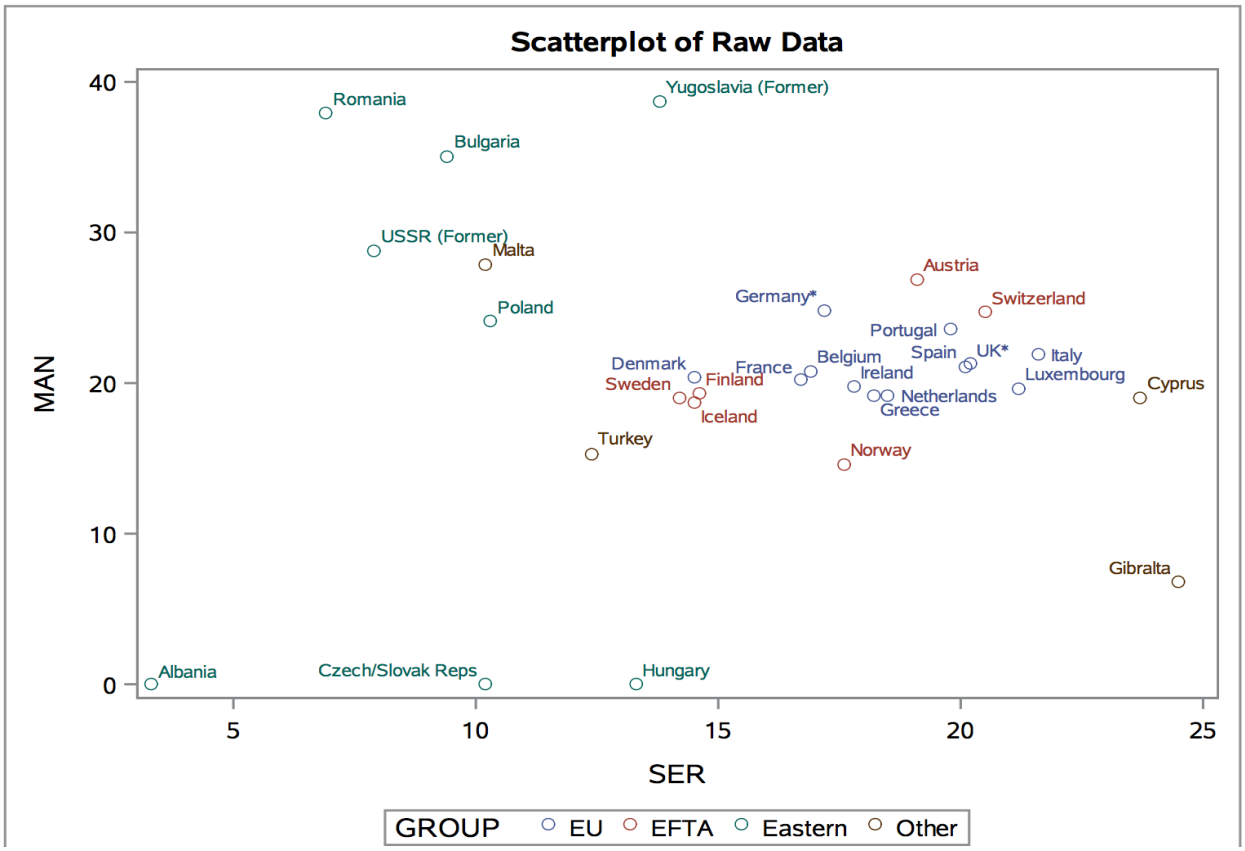


Part 3:

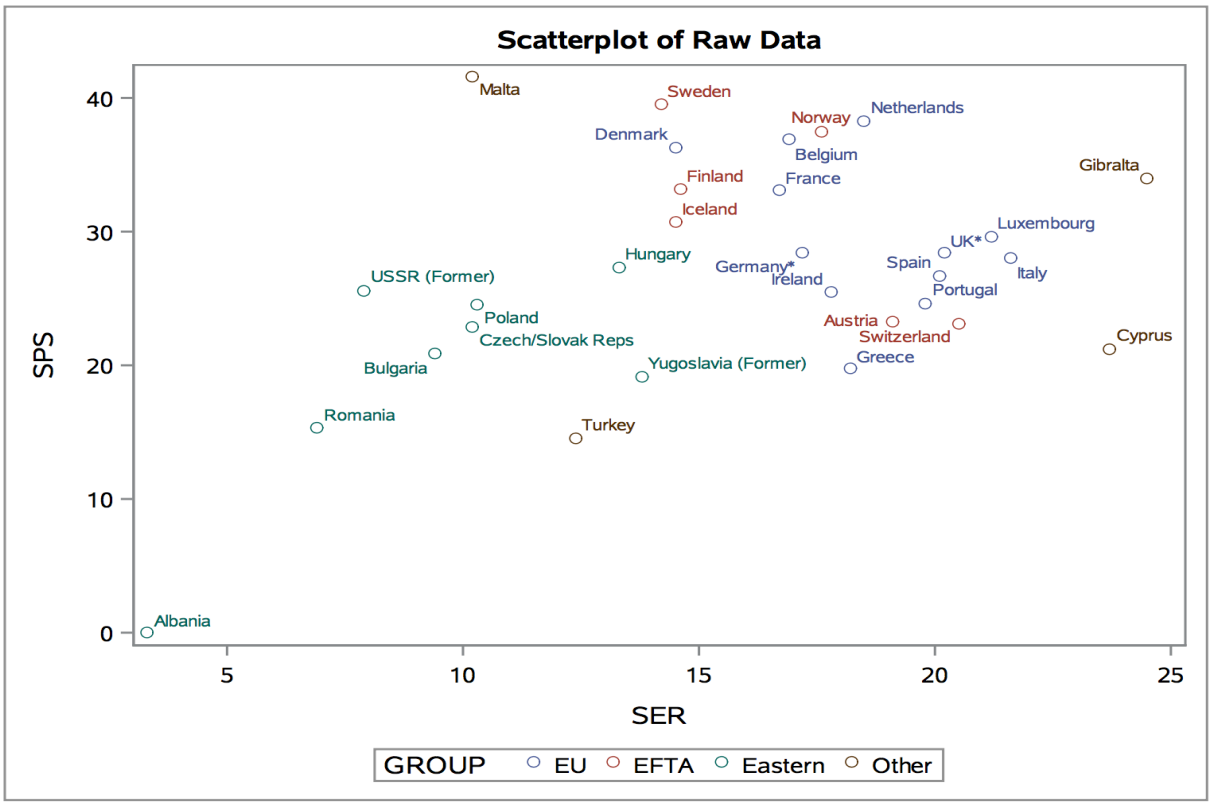
Below is a scatterplot of Financial versus Services % Employment by Country. There is a strong cluster around the Eastern countries such as Bulgaria, Romania, and Poland. There is another more loosely defined cluster around the European Union and EFTA countries. Outlier clusters are also present. For instance, Albania is so far removed that it would form its own cluster. Overall, this dataset seems to have three to four natural clusters.



The scatterplot of Manufacturing versus Services also forms a large cluster around the European Union and EFTA countries. A second cluster formed around the three Eastern countries in the lower left corner (Albania, Czech, and Hungry). A third cluster is formed in the top left corner comprising the remaining Eastern countries. Then Gibraltar goes on to form a fourth cluster of its own.



The scatterplot of Social and Personal Services shows three natural clusters. The first loose, large cluster is around the EU and EFTA countries. Cyprus, Gibraltar, and Malta are also included in this cluster. The second cluster consists of the Eastern countries plus Turkey. And the final cluster is made up of just Albania. The results from these scatterplots illustrates the fact that different projections of the data will create different clustering patterns.



Below is the output from the Cluster Analysis using the first two Principal Components. SAS provides three criteria for determining the optimal number of clusters: Pseudo F Statistic, Cubic Clustering Criterion, and Pseudo T Statistic. The Pseudo F Statistic is similar to R squared. R squared is a measure of the between-cluster variation divided by total variation. The Pseudo F Statistic is normalized by the degrees of freedom. If the cluster is good, a large Pseudo F Statistic is expected. A large Pseudo F Statistic means that the between-cluster variation is larger than the within cluster variation, meaning that the cluster created is significant. The value of the Pseudo F Statistic gradually decreases as more clusters (weaker clusters) are formed. When looking at the graph of two to six clusters, the Pseudo F Statistic is highest at four clusters.

The Cubic Clustering Criterion is the R squared value for the cluster divided by R squared without any cluster. A larger CCC is desired. The CCC is largest when four clusters are present.

For the Pseudo T Statistic, the optimal number of clusters is determined by reading the graph from the right. Find the point in which the line graph rises sharply and then go back one cluster. The graph rises from 4 to 3 (when reading right to left). Therefore, the optimal number of clusters according to the Pseudo T Statistic is four. The line graph for each of these measures is shown below. The order in which the countries were added to the clusters is also shown below. The stronger the cluster relationship, the early the cluster is formed. Albania was the last country added to the list because it does not share any significant relationship with the other countries.

Principal Components Analysis using PROC PRINCOMP

The CLUSTER Procedure
Average Linkage Cluster Analysis

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	81.7893110	60.7024725	0.7950	0.7950
2	21.0868385		0.2050	1.0000

Root-Mean-Square Total-Sample Standard Deviation	7.172034

Root-Mean-Square Distance Between Observations	14.34407

Cluster History

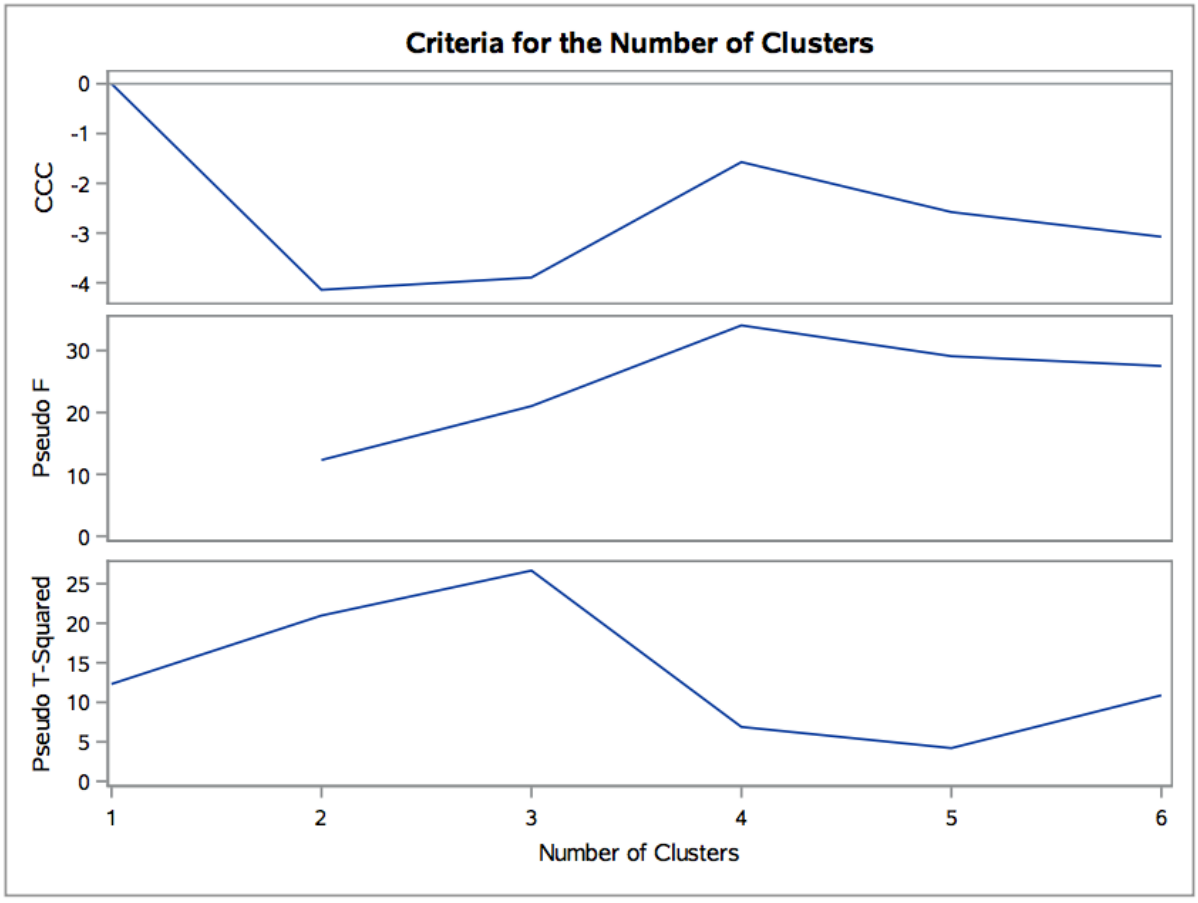
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Norm RMS Distance	Tie
29	Belgium	Norway	2	0.0001	1.00	.	.	251	.	0.0643	
28	Austria	Switzerland	2	0.0003	1.00	.	.	155	.	0.0986	
27	Italy	UK*	2	0.0004	.999	.	.	138	.	0.1015	
26	Portugal	CL28	3	0.0004	.999	.	.	126	1.3	0.1093	
25	Czech/Slovak Reps	Poland	2	0.0004	.998	.	.	122	.	0.1118	
24	CL27	Luxembourg	3	0.0005	.998	.	.	121	1.3	0.112	
23	CL29	Netherlands	3	0.0006	.997	.	.	114	4.3	0.1204	
22	France	Finland	2	0.0007	.996	.	.	108	.	0.1466	
21	CL24	Spain	4	0.0012	.995	.	.	95.4	2.9	0.1633	
20	Ireland	CL26	4	0.0019	.993	.	.	79.9	4.8	0.1988	
19	CL22	Iceland	3	0.0016	.992	.	.	74.1	2.2	0.2024	
18	Bulgaria	CL25	3	0.0019	.990	.	.	69.2	4.4	0.2115	
17	Denmark	Sweden	2	0.0017	.988	.	.	67.9	.	0.2241	
16	Germany*	CL21	5	0.0034	.985	.	.	60.2	5.2	0.2638	
15	CL23	CL17	5	0.0045	.980	.	.	53.2	5.4	0.265	
14	CL18	USSR (Former)	4	0.0031	.977	.	.	52.7	2.6	0.2667	
13	Greece	Yugoslavia (Former)	2	0.0033	.974	.	.	52.7	.	0.3106	
12	CL16	CL20	9	0.0129	.961	.	.	40.2	11.2	0.331	
11	Romania	Turkey	2	0.0052	.956	.	.	41.0	.	0.3875	
10	CL19	Hungary	4	0.0073	.948	.	.	40.8	6.2	0.3917	
9	CL12	Cyprus	10	0.0129	.935	.	.	38.1	4.9	0.4921	
8	CL15	Malta	6	0.0148	.921	.	.	36.5	8.5	0.5266	
7	CL13	CL14	6	0.0264	.894	.	.	32.4	12.1	0.5754	
6	CL8	CL10	10	0.0428	.851	.910	-3.1	27.5	10.9	0.5886	

Principal Components Analysis using PROC PRINCOMP

The CLUSTER Procedure
Average Linkage Cluster Analysis

Cluster History

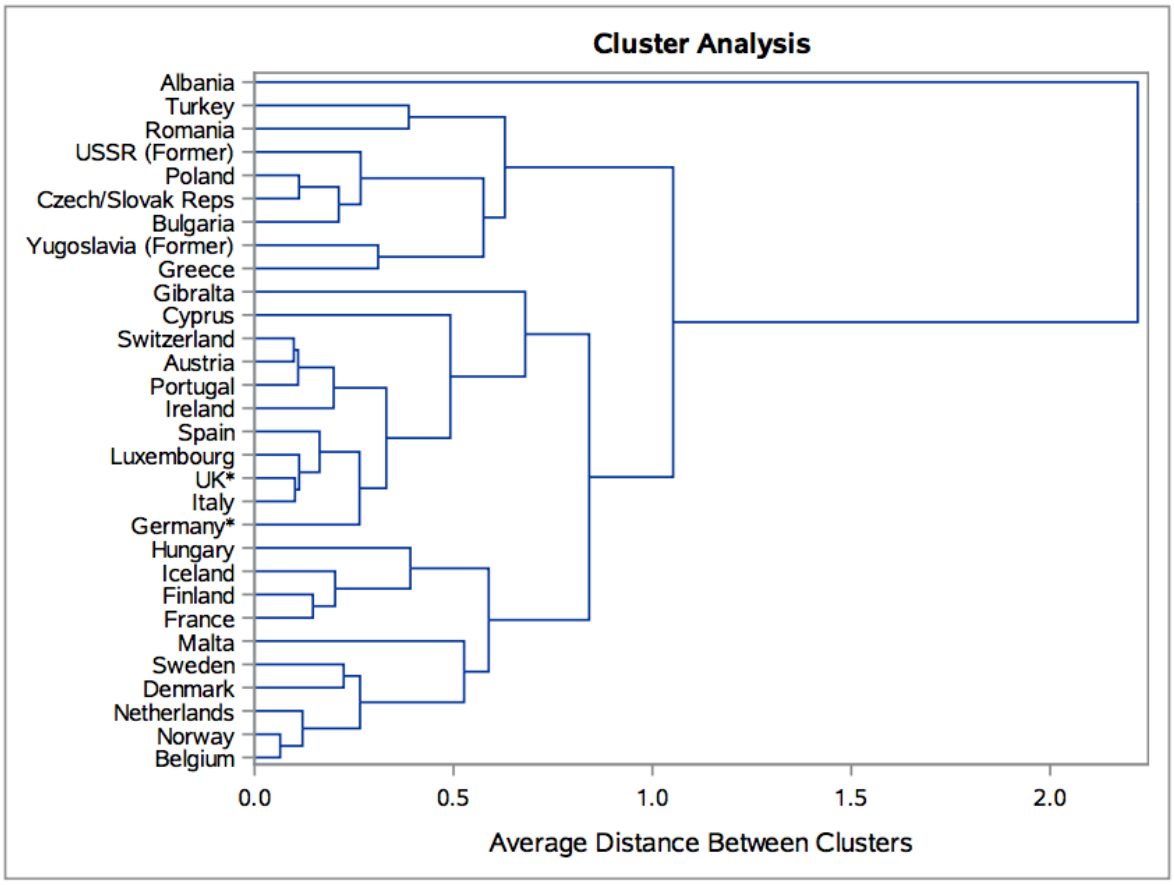
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Norm RMS Distance	Tie
5	CL7	CL11	8	0.0283	.823	.882	-2.6	29.1	4.2	0.6292	
4	CL9	Gibralta	11	0.0259	.797	.839	-1.6	34.1	6.9	0.6803	
3	CL6	CL4	21	0.1882	.609	.767	-3.9	21.0	26.7	0.8411	
2	CL3	CL5	29	0.3036	.305	.620	-4.1	12.3	21.0	1.0521	
1	CL2	Albania	30	0.3054	.000	.000	0.00	.	12.3	2.2201	



Below is the hierarchical Cluster Analysis tree diagram. EFTA / EU cluster and the Eastern cluster get increasingly segmented as the number of clusters increases. Albania forms its own cluster regardless of the number of clusters used.

Principal Components Analysis using PROC PRINCOMP

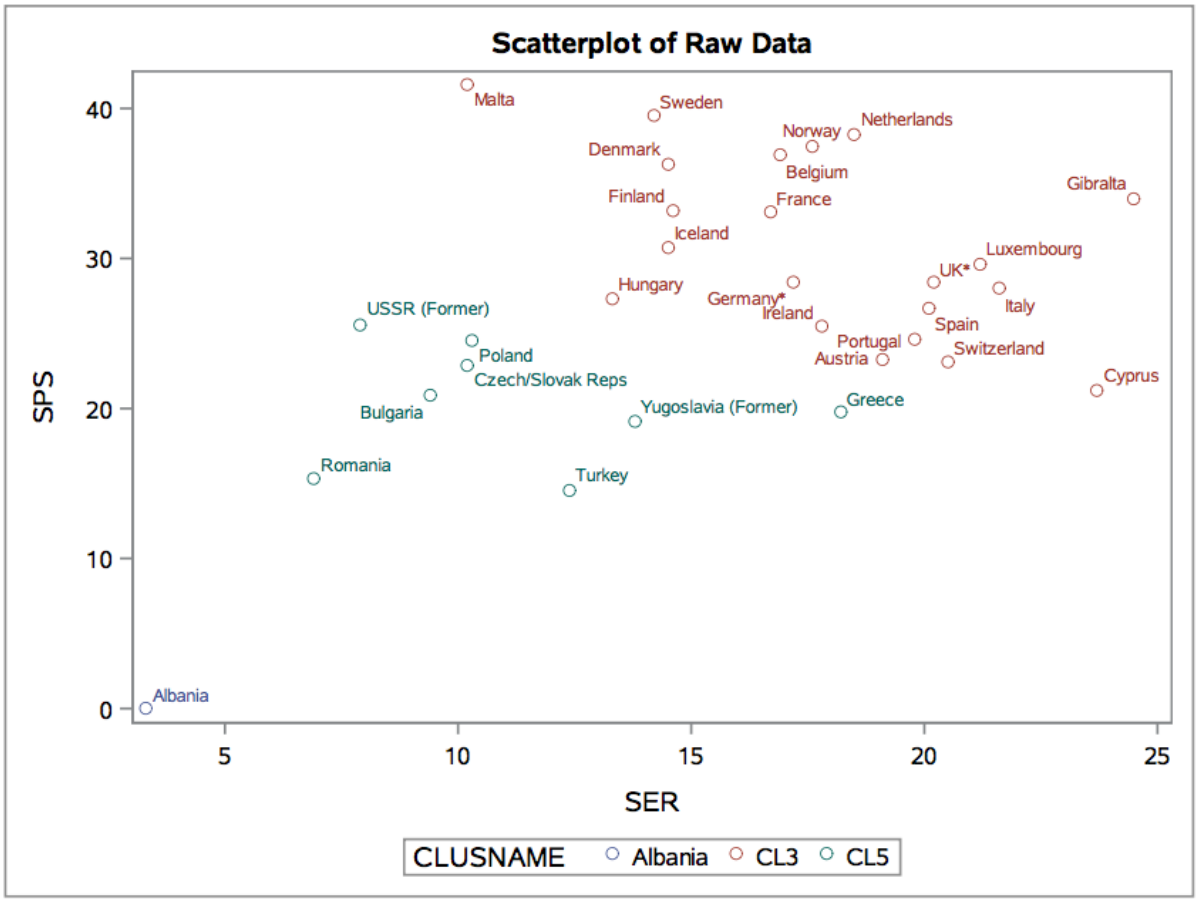
The CLUSTER Procedure
Average Linkage Cluster Analysis



The frequency table for the country groupings when three clusters are present is shown below. Albania forms its own cluster. Cluster CL3 is the largest cluster with all six EFTA countries, eleven of the twelve EU countries, one Eastern country, and three of the other countries. Cluster CL5 has the remaining six Eastern countries, one EU country, and one Other country. The clusters are heavily influenced by country grouping. A scatterplot color coded for the clusters is also shown below.

The FREQ Procedure

Frequency	Table of GROUP by CLUSNAME			
	CLUSNAME			
GROUP	Albania	CL3	CL5	Total
EFTA	0	6	0	6
EU	0	11	1	12
Eastern	1	1	6	8
Other	0	3	1	4
Total	1	21	8	30



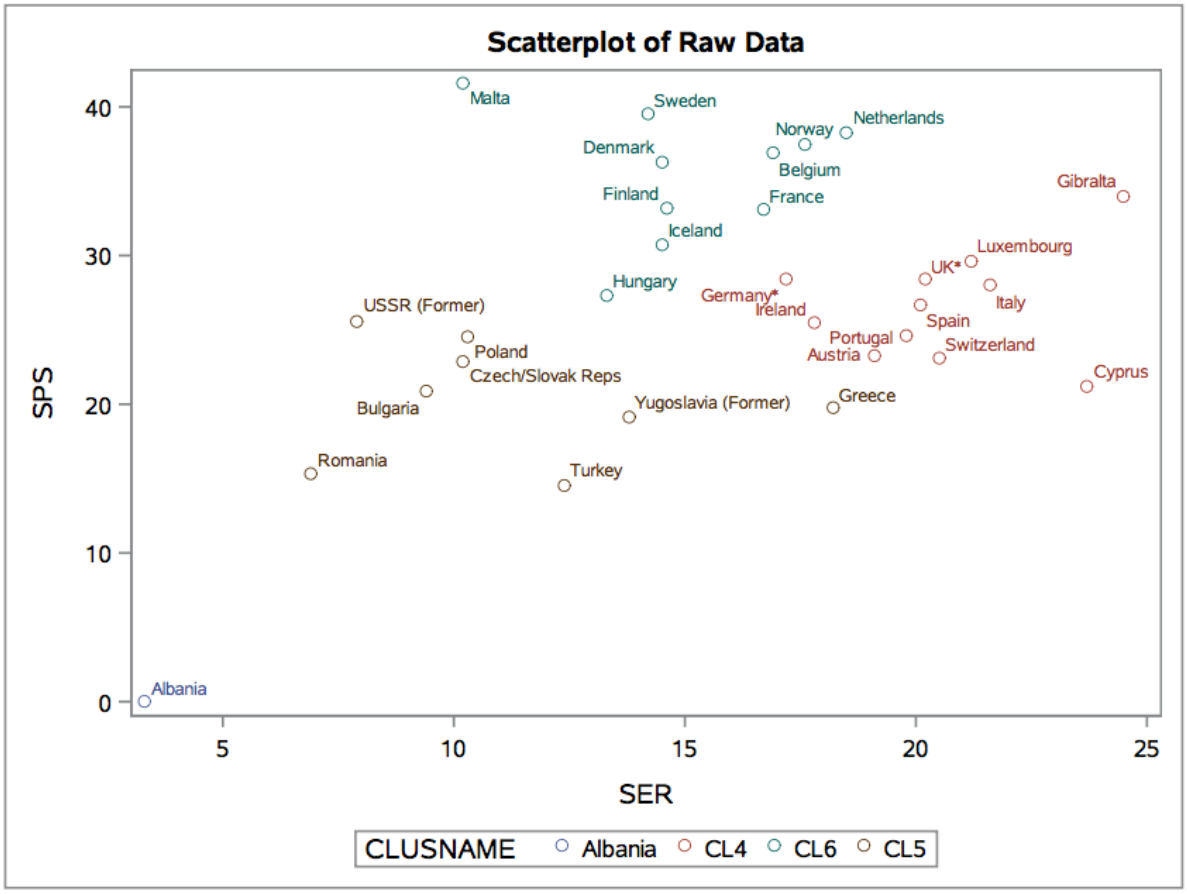
With revised frequency table with four clusters is shown below. Albania remains its own cluster. Cluster CL5 remains unchanged. Cluster CL4 has been divided to create the new cluster of CL6. CL6 consists of four EFTA countries, four EU countries, one Eastern country, and one Other country. Influences other than country grouping are driving the patterns that created this new cluster.

The FREQ Procedure

Frequency

Table of GROUP by CLUSNAME

GROUP	CLUSNAME				
	Albania	CL4	CL5	CL6	Total
EFTA	0	2	0	4	6
EU	0	7	1	4	12
Eastern	1	0	6	1	8
Other	0	2	1	1	4
Total	1	11	8	10	30



The addition of a fifth cluster leaves the Albania, CL5, and CL6 clusters unchanged. Gibraltar is removed from CL4 to create two new clusters. The new clusters are a cluster containing only Gibraltar and a new cluster (CL9) that contains the remaining countries from the CL4 cluster. The addition of the fifth cluster did not provide any new information of value. Four clusters appear to be all that is required. The color-coded scatterplot is shown below.

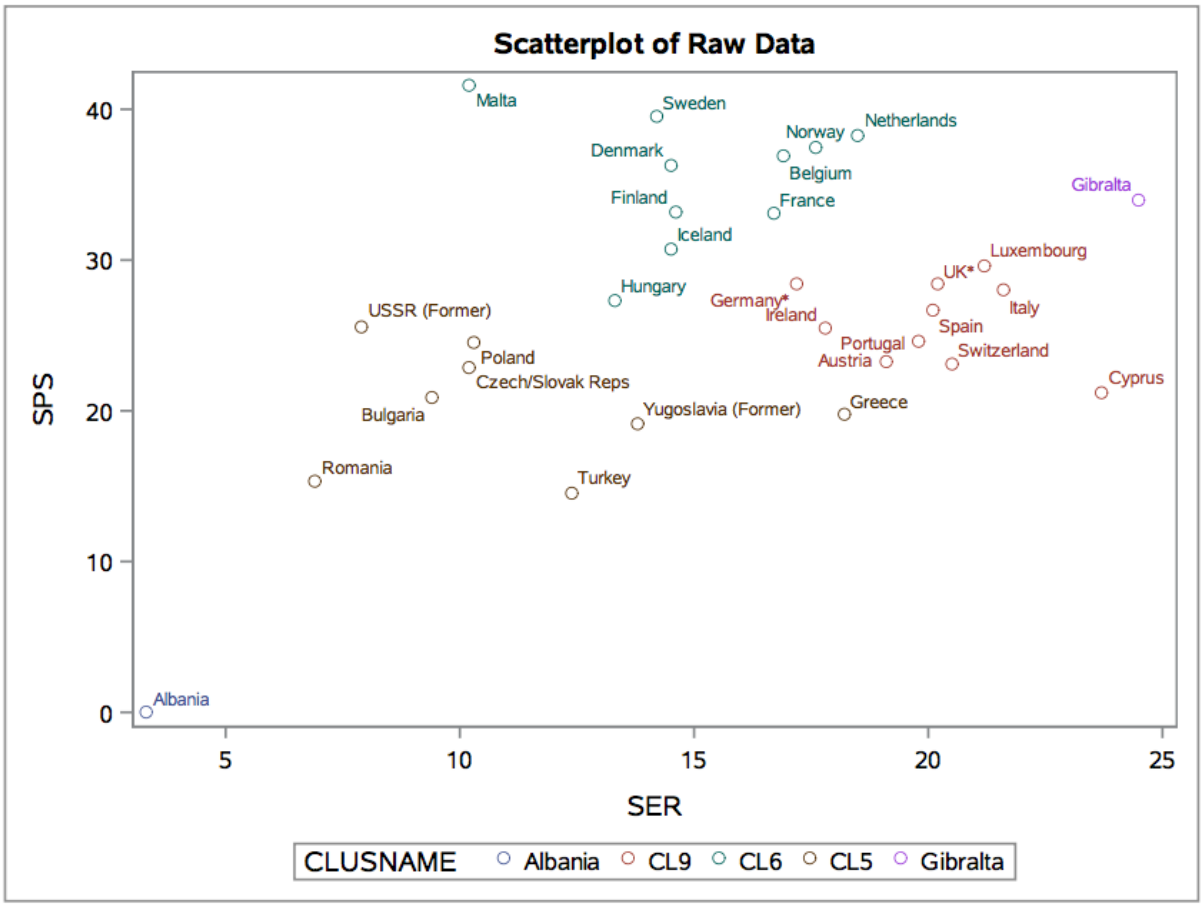
Scatterplot of Raw Data Fr

The FREQ Procedure

Frequency

Table of GROUP by CLUSNAME

GROUP	CLUSNAME					
	Albania	CL5	CL6	CL9	Gibralta	Total
EFTA	0	0	4	2	0	6
EU	0	1	4	7	0	12
Eastern	1	6	1	0	0	8
Other	0	1	1	1	1	4
Total	1	8	10	10	1	30



Conclusions:

Cluster Analysis is a powerful tool for identifying variables that are similar to each other. Clustering can be used to produce homogenous groupings within the datasets that can be used for classification. It is a powerful unsupervised learning method.

The analysis of this dataset of European employment percentage by industry for thirty different countries resulted in four unique clusters. The clusters were influenced by country groupings (EU, EFTA, Eastern, and Other). However, there was considerable overlap between clusters. There were other factors influencing the data besides country groupings. It was found that adding additional clusters was not beneficial. Adding a fifth cluster only segmented away Gibraltar.

The analysis of correlation between the variables showed only moderate correlation. The Principal Components and Clusters identified would have been stronger if the correlation between the variables was stronger.