Noé Flores
Cluster Analysis

Introduction:

This report will utilize and apply cluster analysis techniques on a data set consisting of various employment industry segments across several European nations. The data is broken down by the individual percentage of each employment segment for each of the thirty European Nations represented in the data. Our analysis will begin with an exploratory data analysis and conclude with different cluster analysis techniques we can compare. By initiating our basic examination of the data, we can consider and enlist the use of principal component analysis as a method for reducing the dimensionality of the dataset. The comparison of the results from our cluster analysis methods will stem from both the raw data and include the transformed predictor variables from using principal component analysis.

Results and Analysis:

Our dataset contains 30 observations of 11 different variables. The table below provides a simple breakdown and description of those variables and their attributed variable types.

Figure 1: Variables and Descriptions.

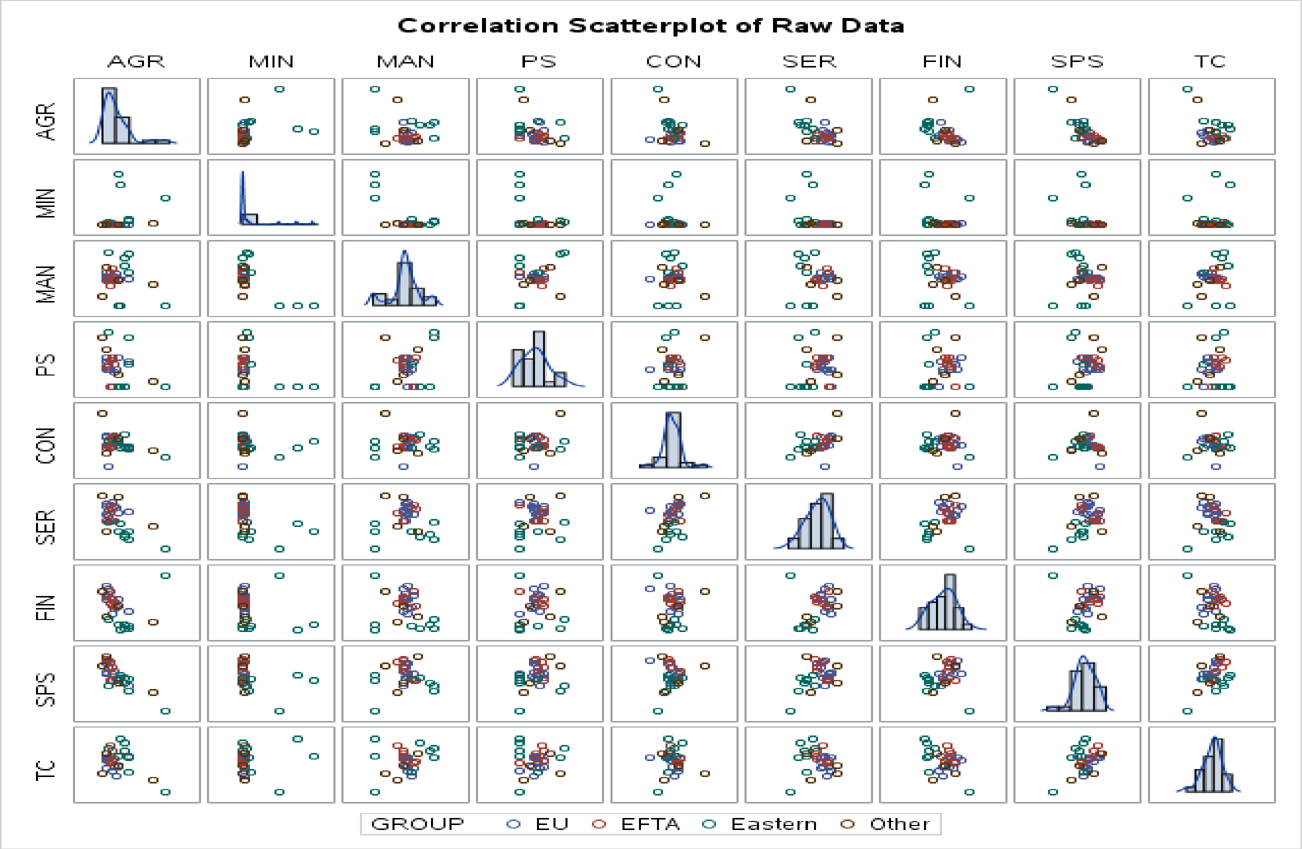
List of Variables and Attributes			
	Variable	Type	Descriptions
1	COUNTRY	Char	Country
2	GROUP	Char	Primary-Group
3	AGR	Num	Agriculture
4	CON	Num	Construction
5	FIN	Num	Finance
6	MAN	Num	Manufacturing
7	MIN	Num	Mining
8	PS	Num	Power and Water Supply
9	SER	Num	Services
10	SPS	Num	Social and Personal Services
11	TC	Num	Transport and Communications

Given that our data is composed of reported employment percentages within various industry segments in each country, we can start our analysis with a simple correlation matrix and graphical display of the plots to examine any possible relationships between the employment variables.

Figure 2: Correlation Matrix.

	Pearson Correlation Coefficients, N = 30								
	AGR	**MIN**	**MAN**	**PS**	**CON**	**SER**	**FIN**	**SPS**	**TC**
AGR	 1.0000	0.3161 0.0888	-0.2544 0.1749	-0.3824 0.0370	-0.3486 0.0590	-0.6047 0.0004	-0.1758 0.3529	-0.8115 <.0001	-0.4873 0.0063
MIN	0.3161 0.0888	 1.0000	-0.6719 <.0001	-0.3874 0.0344	-0.1290 0.4968	-0.4066 0.0258	-0.2481 0.1863	-0.3164 0.0885	0.0447 0.8146
MAN	-0.2544 0.1749	-0.6719 <.0001	 1.0000	0.3879 0.0342	-0.0345 0.8565	-0.0329 0.8628	-0.2737 0.1433	0.0503 0.7919	0.2429 0.1959
PS	-0.3824 0.0370	-0.3874 0.0344	0.3879 0.0342	 1.0000	0.1648 0.3842	0.1550 0.4135	0.0943 0.6201	0.2377 0.2059	0.1054 0.5795
CON	-0.3486 0.0590	-0.1290 0.4968	-0.0345 0.8565	0.1648 0.3842	 1.0000	0.4731 0.0083	-0.0180 0.9247	0.0720 0.7053	-0.0546 0.7744
SER	-0.6047 0.0004	-0.4066 0.0258	-0.0329 0.8628	0.1550 0.4135	0.4731 0.0083	 1.0000	0.3793 0.0387	0.3880 0.0341	-0.0849 0.6556
FIN	-0.1758 0.3529	-0.2481 0.1863	-0.2737 0.1433	0.0943 0.6201	-0.0180 0.9247	0.3793 0.0387	 1.0000	0.1660 0.3806	-0.3913 0.0325
SPS	-0.8115 <.0001	-0.3164 0.0885	0.0503 0.7919	0.2377 0.2059	0.0720 0.7053	0.3880 0.0341	0.1660 0.3806	 1.0000	0.4749 0.0080
TC	-0.4873 0.0063	0.0447 0.8146	0.2429 0.1959	0.1054 0.5795	-0.0546 0.7744	-0.0849 0.6556	-0.3913 0.0325	0.4749 0.0080	 1.0000

Figure 3: Correlation Scatterplots.



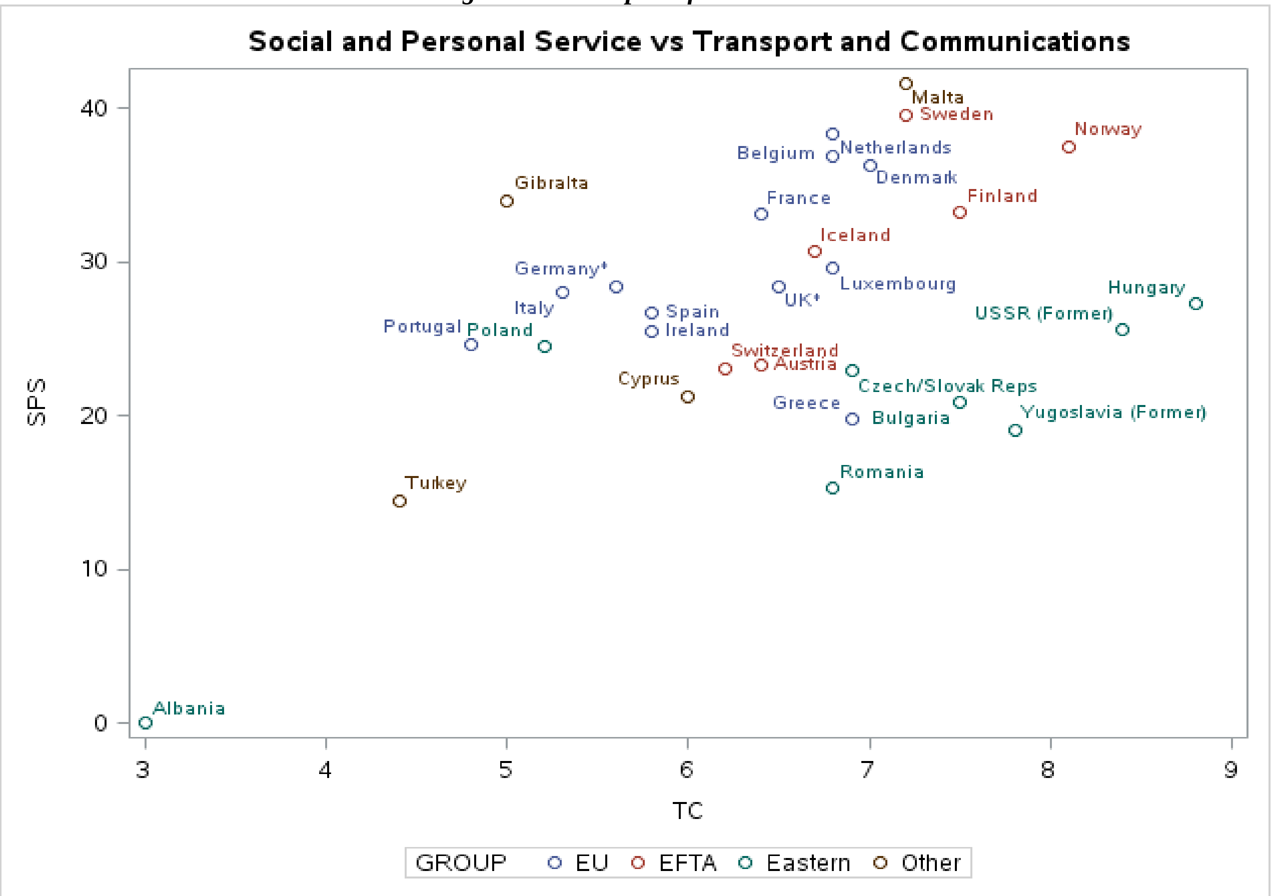
The strongest correlation in the matrix was that between AGR and SPS, but that correlation was negative. Having a negative correlation isn't problematic, it's an indicator of one variable increasing while the other variable decreases. For simplicity, I have chosen to examine Social and Personal Services (SPS) and Transport and Communications (TC) since these two variables exhibited one of the highest positive correlations while maintaining a significant P-value below .05.

Figure 4: SPS vs TC correlation and P-Value.

SPS VS TC	
Correlation	**0.4749**
P-Value	**0.0080**

Within this dataset, there are three primary groups, European Union (EU), European Free Trade Association (EFTA), Eastern Europe (Eastern), and Other. Below we have a scatter plot with the breakdown of Social and Personal Services (SPS) plotted against Transport and Communications (TC) displaying color codes of the grouping for each country. These groupings possibly appear to be broken down by the various trade agreements in place across Europe. There is also some visual clustering to that effect as well. There are four countries that find themselves outside of the three main groups. It doesn't appear indicative that location would have anything to do with the groupings, and given that information, it may be useful to bucket these countries according to their major financial trading partners, most likely to be their neighboring European countries.

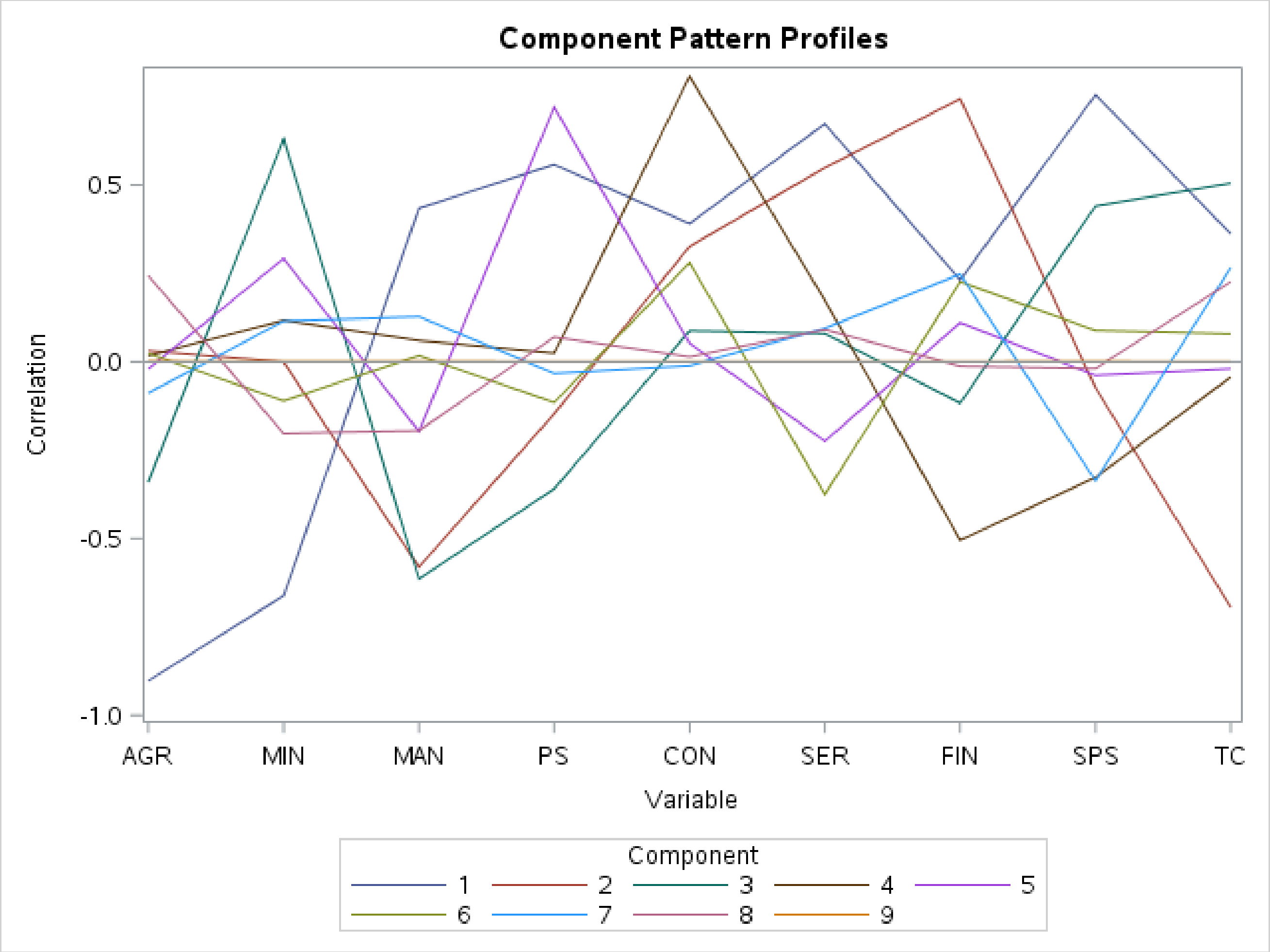
Figure 5: Scatterplot of SPS vs TC.



To move forward with the analysis, we will now perform a principal component analysis (PCA) on our data set. The incorporation of (PCA) is used to find underlying variables that can best differentiate our data and to combat multicollinearity by reducing the dimensionality projection of our data.

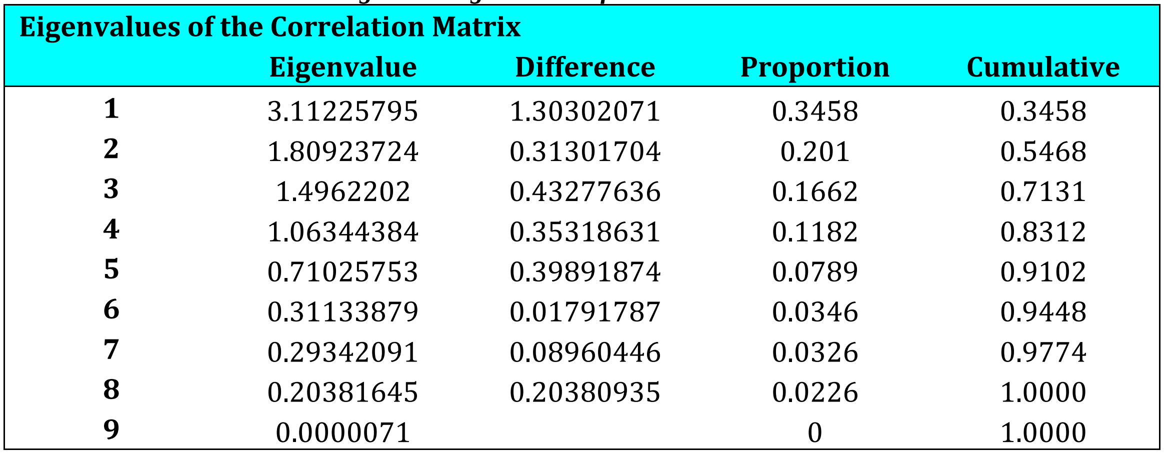
Our first visual from our PCA is the Component pattern Profiles plot. This plot provides us with a representation which compares our component's correlation across all the variables. We can observe that the correlations for each component differ in strength across each variable.

Figure 8: Component Pattern Profiles



The table below provides our components, their eigenvalues, and it also provides us with the proportion of variation each variable accounts for.

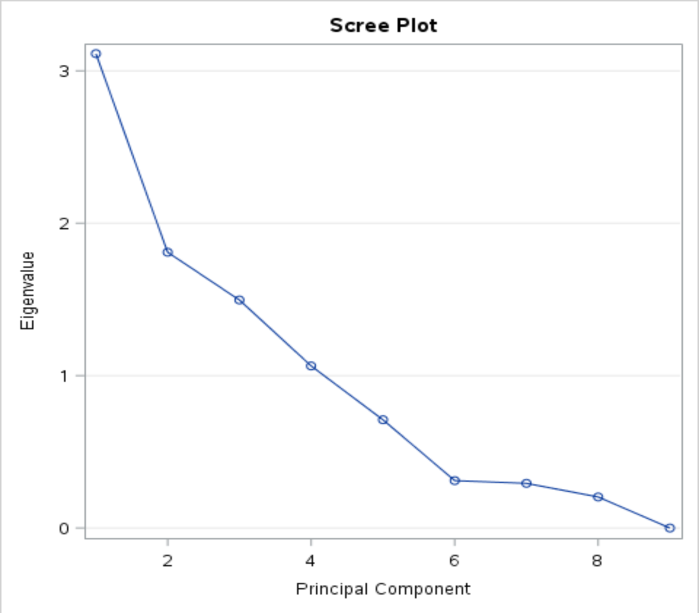
Figure 6: Eigenvalues of Correlation matrix.



Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.11225795	1.30302071	0.3458	0.3458
2	1.80923724	0.31301704	0.201	0.5468
3	1.4962202	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	0.0000071		0	1.0000

One of the most common decision rules associated with PCA states that we should keep just enough components to explain a large enough cumulative percentage of the total variation of the original variable. The suggested percentage is in the realm of 70% to 90%. Based on this guidance we would likely be safe selecting 4 or 5 principal components.

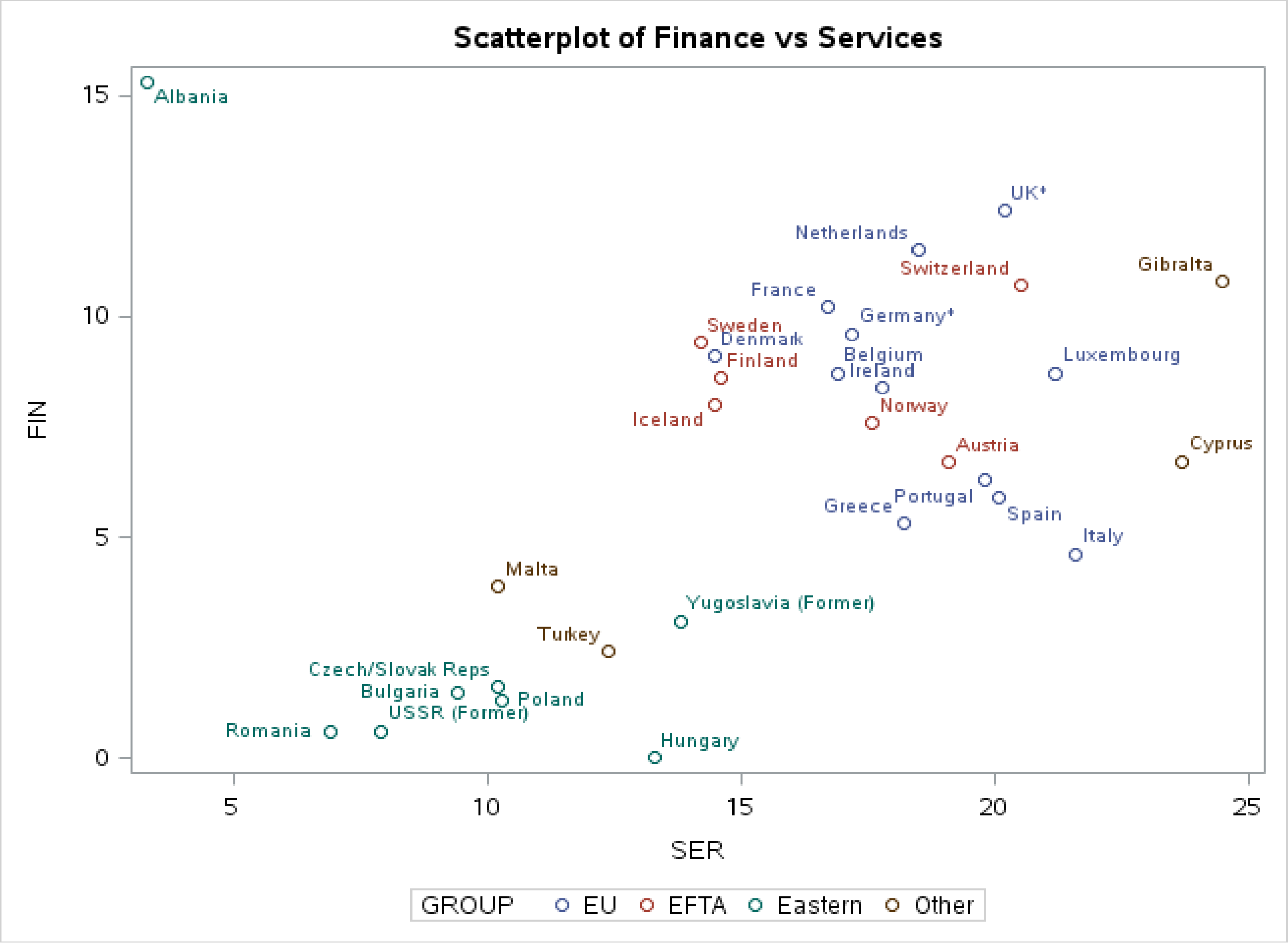
Figure 7: Scree Diagram



The Scree Plot is another decision tool in PCA. The number of components selected is the value corresponding to an "elbow" in the curve, or a change in slope from steep to shallow. Looking at the Scree plot above, you can see that after the sixth component the slope of the line begins to flatten out. A different rule brings a different perspective to our decision. Given the controls available for each selection method, it appears that selecting 5 components will achieve our desired 90% of the variability, but the fact that the Scree plot is leaning towards 6 components is a bit troublesome. In the effort to produce a simple analysis, the decision would be to select 5 variable.

Our cluster analysis will begin with a pair of scatterplots utilizing some of the variables in our dataset. Finance will be compared with Services and Manufacturing will be compared against Services as well. Figure.8 has our first plot of Finance against Services and gives us the scatter representation by for each group.

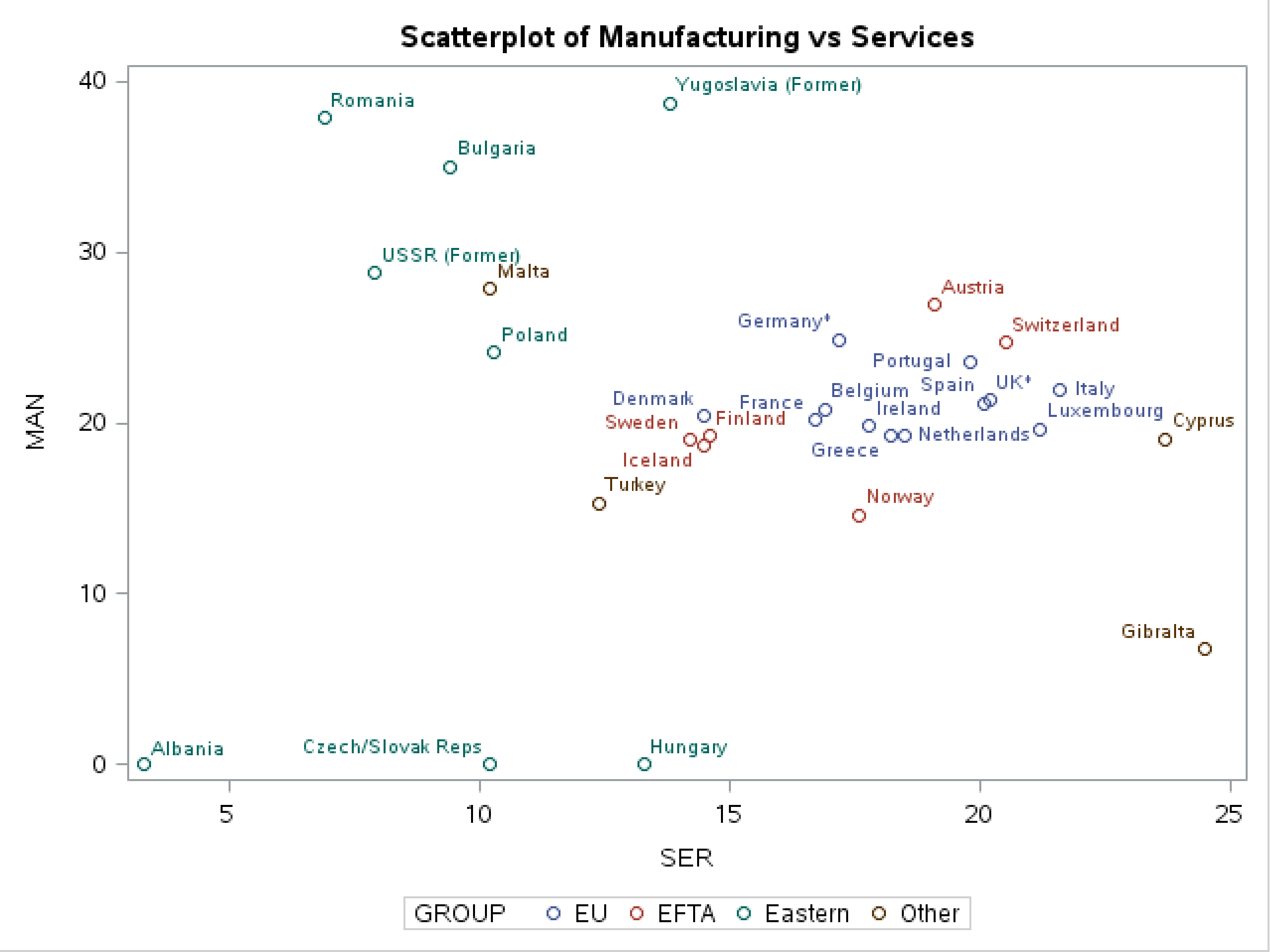
Figure 8: Scatterplot FIN vs. SER



The basic idea of Cluster analysis revolves around the sorting of similar objects into specific categories. The most straightforward approach to identifying groups or clusters in multivariate data is through the examination of scatter plots. In our first plot above, without the use of any specific statistical clustering methods, we can make out a few clusters in the plot. The Eastern European (Eastern) countries appear to be clustering together near the bottom of the plot. European Union (EU) and the European Free Trade countries (EFTA) seem to be forming a cluster of their own. There are some outlying data points as well that are peculiar. Albania is off on its own near to top left and Cyrus and Gibralta are off to the right side of the plot but in this case, we can possibly make an argument to consider this another small cluster.

The scatter plot below compares Manufacturing against Services. The plot once again gives us the scatter representation by group.

Figure 9: Scatterplot MAN vs SER



In this plot, the clusters appear to be gathering a bit counter-intuitively from our purely visual observation. We still aren't utilizing any formal clustering methods to achieve any concrete conclusions on the clusters within the plot. We can observe that different clusters of Eastern European countries are gathering near the bottom of the plot and the top portion. The EU and EFTA countries appear to be in a tighter cluster in this plot, while we continue to see Cyrus and Girbralta off to the far right forming their own group or cluster. Somewhat not surprising is that in each of these two plots the clusters were different, but at the same time they maintained some familiarity with respect to which groups were contained within each cluster.

We can now proceed to utilize proper cluster analysis techniques and create our clusters algorithmically and not through pure visual observation. In this first run of our cluster analysis, the statistical software will incorporate hierarchical clustering, so it won't be necessary to specify the exact number of clusters in advance. Once the hierarchical clustering has been performed the software will assign the observations to a specified number of clusters.

Figure 10: Clustering Criteria

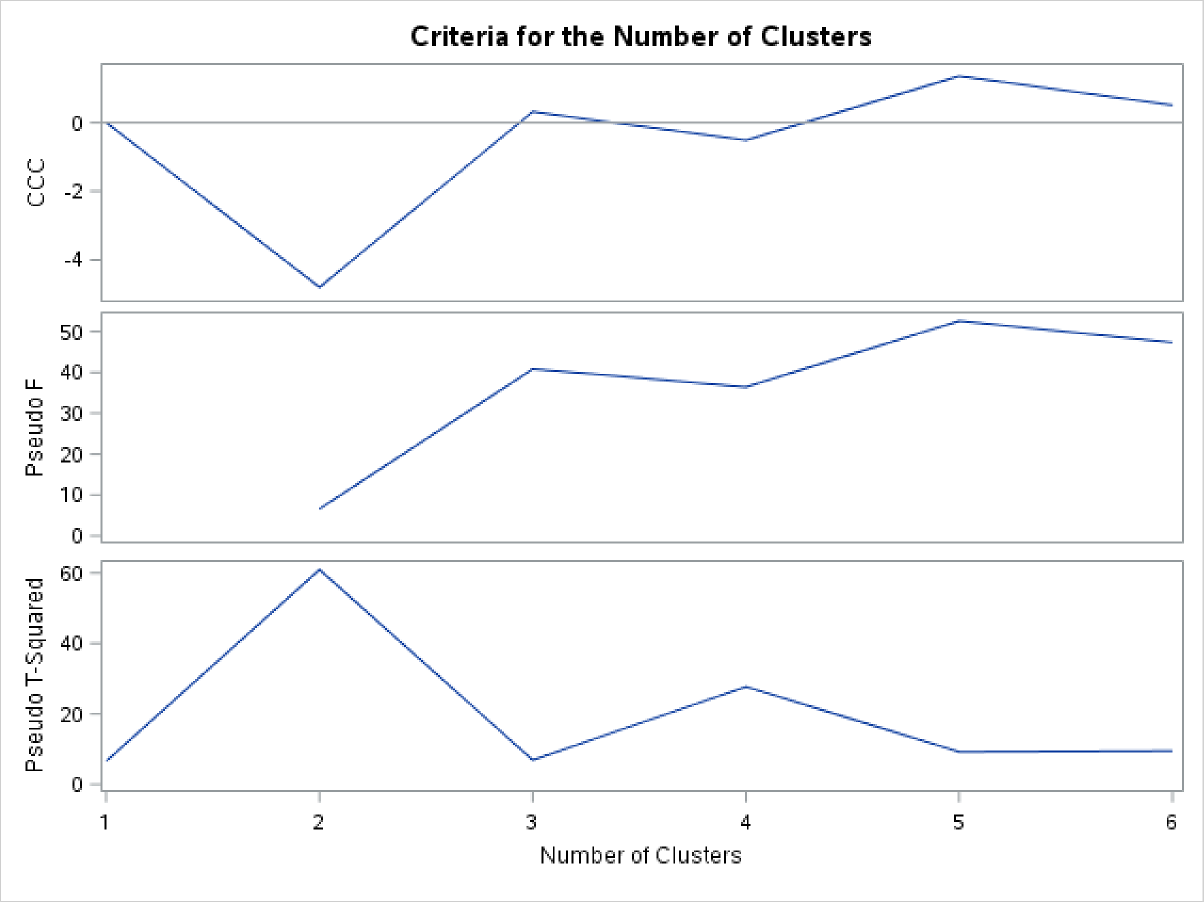


Figure.10 provides a visual output from the cluster analysis with three different measures for interpretation. The Cubic Clustering Criterion (CCC), Pseudo F, and Pseudo T-Squared. The CCC can be used to estimate the number of clusters by minimizing the sum of squares. What we are looking for is peaks on the plot with values of CCC which are great than 2. Our CCC plot appears to peak near 5. The pseudo F statistic measures the separation among the clusters at their current hierarchical level. We want to find peaks in our plot which correspond to the optimal number of clusters. Once again, our Pseudo F plot is peaking at5. The Pseudo T-Squared attempts to quantify the difference between two clusters that are merged at a certain point. In terms of the plot, if the Pseudo T-Square value is significant, then those clusters should not be considered. If the value of PseudoT-Square is small, then those clusters can safely be combined and used. Pseudo T-Square bottoms out at 5 clusters corroborating the indications of the previous two measures.

As of now, our analysis was built through the use of a statistical algorithm that performs hierarchical clustering, basically skipping over the necessity of specifying the desired amount of clusters. We will now move to assign our data to a fixed number of clusters. Specifically, we will assign our observations to 3 and 4 clusters and compare the results of the groupings.

Figure 11: Frequency table for 3 Clusters.

Frequency Table of Group by Cluster name				
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**

Figure 12: Frequency table for 4 Clusters.

Frequency Table of Group by Cluster name					
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**

The results of our cluster analysis show some interesting similarities between the utilization of three or four clusters. The three cluster Frequency table in Figure.11 shows all of the European Free Trade Agreement (EFTA) and European Union (EU) countries bucketed into one cluster. Eastern European (Eastern) and Other group are mostly contained within a different cluster. This information appears to be reasonably consistent with the visual observations we noticed earlier. The ever persistent outlier, Albania, has found itself in a single cluster, which is also compatible with what we observed on the initial plot where we attempted to asses clusters through purely visual mechanisms.

The Frequency table with four clusters in Figure.12 shows a bit more fracturing of the groups. Specifically, we see that the EFTA and EU give up some of their members and form a new cluster with one of the countries from the Other group. Everything else remains the same, and the initial visual assumptions we made appears to be confirmed in this table as well. Given the fact that there is very little difference in the distribution of members in three or four clusters, it would appear preferable to use three clusters and simplify our analysis.

We will now perform similar cluster analysis, but will incorporate the principal component analysis we ran earlier. The PCA is included as means for dimensionality reduction. In this first iteration of our cluster analysis, the statistical software will incorporate hierarchical clustering meaning it won't be necessary to specify the exact number of clusters in advance

Figure 13: Clustering Criteria for PCA clustering

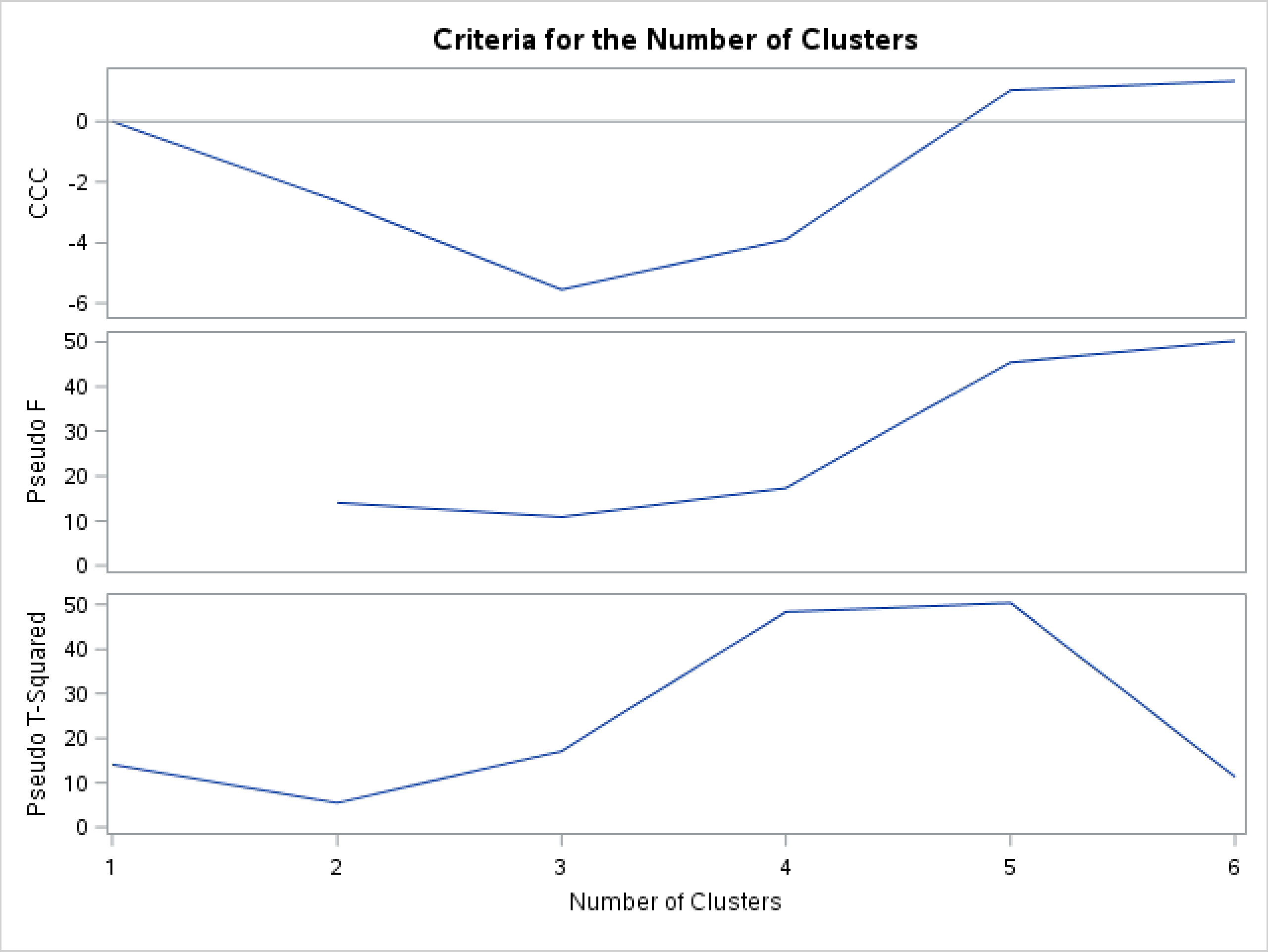
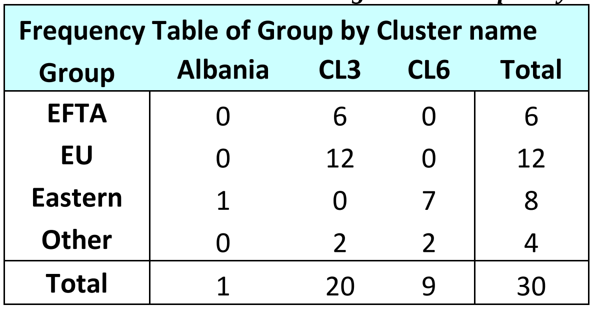


Figure.13 provides a visual output from the cluster utilizing the Principal Component Analysis. Once again, the Cubic Clustering Criterion (CCC), Pseudo F, and Pseudo T-Squared are used to estimate the number of clusters. We want to find the peaks in the CCC plot and hopefully, have values greater than two. Our CCC plot appears to be peaking near five. The pseudo F statistic measure wants to find peaks in our plot which correspond to the optimal number of clusters. Our Pseudo F plot is also peaking near five. The Pseudo T-Squared attempts to quantify the difference between two clusters, if the Pseudo T-Square value is large, then those clusters should not be considered. Our Pseudo T-Square bottoms out at six clusters. Given the information present in these metrics, utilizing five or six clusters would be optimal in this particular cluster analysis.

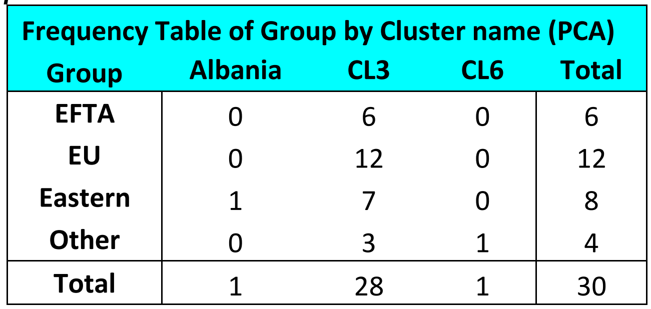
We will now move to assign our data to a fixed number of clusters based on the data in our Principal Component Analysis (PCA). Our observations will once again be segregated to 3 and 4 clusters and compared for the results of the groupings. In order to get a better visual representation of this comparison we will display the tables side by side

Figure 14: Frequency table for 3 Clusters Initial vs PCA.



Frequency Table of Group by Cluster name

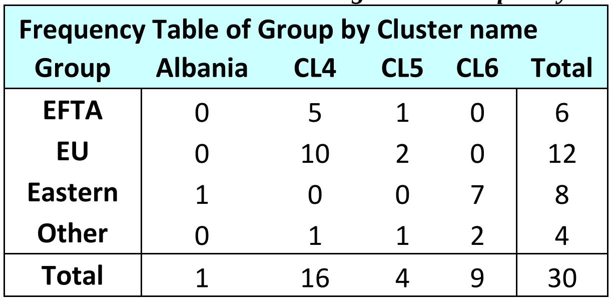
Group	Albania	CL3	CL6	Total
EFTA	0	6	0	6
EU	0	12	0	12
Eastern	1	0	7	8
Other	0	2	2	4
Total	1	20	9	30



Frequency Table of Group by Cluster name (PCA)

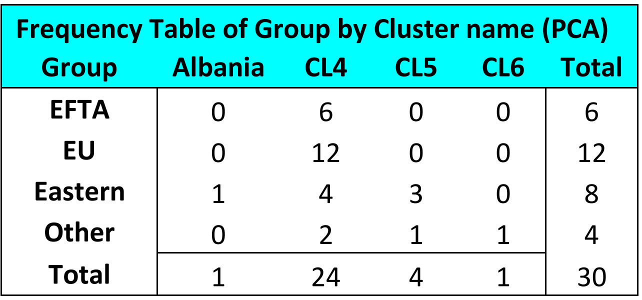
Group	Albania	CL3	CL6	Total
EFTA	0	6	0	6
EU	0	12	0	12
Eastern	1	7	0	8
Other	0	3	1	4
Total	1	28	1	30

Figure 15: Frequency table for 4 Clusters Initial vs PCA.



Frequency Table of Group by Cluster name

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



Frequency Table of Group by Cluster name (PCA)

Group	Albania	CL4	CL5	CL6	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

In each case, there are some similarities between the two cluster analysis techniques when we compare our initial results on the left against those run with the principal component analysis on the right. The three cluster PCA Frequency table in Figure.14 shows us that once again all of the European Free Trade Agreement (EFTA) and European Union (EU) countries were bucketed into one cluster but we now have the majority of the Eastern European (Eastern) countries as well as three out of the four other labeled nations within that very same cluster. Albania continues to be a solitary outlying cluster, but the groupings here have been significantly consolidated.

The PCA Frequency table with four clusters in Figure.15 shows more consolidation of the groups. The EFTA and EU kept the same members as the three cluster group and consolidated when compared to the original cluster analysis. In this analysis, we do have more fracturing in the Eastern countries with a small majority located in the main group.

Given the object of cluster analysis to group variables based on similar observations, it would still appear logical to use the raw data with three clusters as opposed the PCA which produced more tight groupings and coincided more with our visual non-algorithmic observations of where we thought each country would be grouped.

Conclusion:

This analysis examined two different cluster analysis modeling techniques. We began each cluster analysis technique by plotting two of our variables against each other differentiated by their individual groups in order to visually quantify and observe the presence of similarities and group clusters. The formal algorithmic hierarchical cluster technique was performed on the raw data produced results where we were able to visualize the optimal cluster selection through the examination of Cubic Clustering Criteria (CCC), Pseudo F, and Pseudo T-Square. Our analysis continued by specifying the number of clusters and segregated the results into three or four clusters. The use of Principal component analysis was used as a means to reduce dimensionality, and the first nine principal components were used in our second cluster analysis. A comparison of results of the Raw data cluster analysis for three and four clusters, versus the one involving our principal components with three or four clusters, lead to our final opinion stating that the use of three clusters compromised of the raw data would be sufficient for our segmentation needs.

Apendix:SAS Code

```
/*Read file into SAS and identify it as mydata*/
libname mydata "/scs/wtm926/" access=readonly;
data temp;
set mydata.european_employment;
run;
/***Check contents of data***/
proc contents data = temp;
run; quit;
/***Part 1: An Initial Correlation Analysis***/
/***perform some basic examinations of the data and consider using principal components as
means to reduce
the dimensionality of our data.begin this tutorial by examining the two dimensional scatterplots of
the
variables. Use PROC CORR to produce the Pearson correlation coefficients and the scatterplot
matrix***/
proc corr data=temp;
var AGR MIN MAN PS CON SER FIN SPS TC;
/***Correlation matrix with histogram grouped by group***/
ods graphics on;
proc sgscatter data=temp;
matrix AGR MIN MAN PS CON SER FIN SPS TC /group=group diagonal=(histogram kernel);
title 'Correlation Scatterplot of Raw Data';
run; quit;
ods graphics off;
/***Scatterplot of Social and Personal Services vs Transport and Communications***/
ods graphics on;
proc sgplot data=temp;
```

```
title 'Social and Personal Services vs Transport and Communications';
scatter y= SPS x= TC / datalabel=country group=group;
run; quit;
ods graphics off;
/***Part 2: Principal Component Analysis***/
/***Include the table of the eigenvalues of the correlation matrix, the scree plot,
and the “Component Pattern Profiles” plot.***/
/***PCA***/
ods graphics on;
title 'Principal Components Analysis using PROC PRINCOMP';
proc princomp data=temp out=pca_9components outstat=eigenvectors plots(unpack)=all;
run; quit;
ods graphics off;
/***Part 3:Cluster Analysis - Begin discussion of cluster analysis by making a pair of scatterplots***/
/***Scatterplot of Finance vs Services***/
ods graphics on;
proc sgplot data=temp;
title 'Scatterplot of Finance vs Services';
scatter y=fin x=ser / datalabel=country group=group; run; quit;
ods graphics off;
/***Scatterplot of Manufacturing vs Services***/
ods graphics on;
proc sgplot data=temp;
title 'Scatterplot of Manufacturing vs Services';
scatter y=man x=ser / datalabel=country group=group; run; quit;
ods graphics off;
```

```
/***Now we will use PROC CLUSTER to create a set of clusters algorithmically***/

/***PROC CLUSTER performs hierarchical clustering we do not need to specify the # of clusters
in advance***/

/***Cluster Analysis hierarchical***/

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;

/***Use PROC TREE to assign our data to a set number of clusters***/

/***4 clusters***/

ods graphics on;

proc tree data = tree1 ncl = 4 out = _4_clusters;

        copy fin ser;

run; quit;

ods graphics off;

/***3 Clusters***/

ods graphics on;

proc tree data = tree1 ncl = 3 out = _3_clusters;

        copy fin ser;

run; quit;

ods graphics off;

/***Use this macro to make tables displaying the assignment of the observations to the determined
clusters***/

%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 (1)***/
ods graphics on;
proc sgplot data=_3_clusters_with_labels;
title 'Scatterplot of Finance vs Services';
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 (2)***/
ods graphics on;
proc sgplot data=_4_clusters_with_labels;
title 'Scatterplot of Finance vs Services';
scatter y=fin x=ser / datalabel=country group=clusname; run; quit;
ods graphics off;
/***Perform a similar cluster analysis using the following cluster commands***/
/***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=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);
/***Plot the clusters for a visual display (3)***/
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 (4)***/
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;
%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);
%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);
quit;
```

References

(1) Everitt, B. S., Dunn, G., (2001). Applied Multivariate Data Analysis. Chichester, West Sussex United Kingdom: Wiley and Sons.

(2) Montgomery, D. C., Peck, E. A., Vinning, G. G., (2012). Introduction to Linear Regression Analysis Hoboken, NJ: Wiley.

(3) Everitt, B. S. (2010). Basic Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences. Sound Parkway, NW: CRC Press.

(4) Cody, R. (2011). SAS: Statistics by Example. Carey, NC: SAS Institute Inc.

(5) http://support.sas.com/kb/22/addl/fusion_22540_1_a108_5903.pdf (accessed February 28, 2017)