Assignment #8: Cluster Analysis (0 points)

Data: The data for this assignment is the European employment data set. This data will be made available by your instructor.

Data Description: Employment in various industry segments reported as a percent for thirty European nations. Note that EU stands for European Union, EFTA stands for European Free Trade Association, and Eastern stand for Eastern European nations or the former Eastern Block.

For convenience here are the definitions of the abbreviated industries.

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

Assignment Instructions:

For this assignment we will perform a cluster analysis starting with an exploratory data analysis and completing with a comparison of cluster results from raw predictor data and cluster results from transformed predictor variables using principal components analysis.

Part 1: An Initial Correlation Analysis

We will conclude this tutorial by applying cluster analysis to this data. When we perform a cluster analysis, we will always want to perform the cluster analysis in a low dimensional setting. Only in low dimensions can points be "close together". As we move towards this cluster analysis we want to perform some basic examinations of the data and consider using principal components as means to reduce the dimensionality of our data.

Of course, before we conclude this tutorial we must begin this tutorial. We will 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. Looking at the scatterplots, is there any scatterplot that looks like it would yield interesting cluster results? For the two variables of your choice make this scatterplot (replace Yvar and Xvar with your two variables).

```
data temp;
set mydata.european_employment;
run;

ods graphics on;
proc sgplot data=temp;
title 'Scatterplot of Raw Data';
scatter y=Yvar x=Xvar / datalabel=country group=group;
run; quit;
ods graphics off;
```

In this data set there are four counties that do not belong to any of the three primary groups. If you had to assign each of these countries to a group to which group would you assign each country.

Note: In this assignment our observations are assigned to *classes* or are said to have *labels* (EU, EFTA, Eastern, or Other). Typically we use cluster analysis as an *unsupervised learner* (a situation with no response variable or label) and not as a *supervised learner* (a situation with a response variable or label). If we wanted to be able to correctly assign each country to its group affiliation, then we would define a *classification problem* (see Chapter 11 in *Applied Multivariate Data Analysis*). Throughout this assignment we will be interested in grouping countries together (creating a *segmentation*), but we can also observe their group affiliation to see if these groups have similarities.

Part 2: Principal Components Analysis

Our data set has nine variables. One method of reducing the dimensionality of our data set is to use principal components analysis. If we perform a principal components analysis, what would the resulting dimensionality be, i.e. how many components should we keep? What decision rule are you using to determine how many of the principal components to keep? Are there any other competing decision rules that you could use? Include the table of the eigenvalues of the correlation matrix, the scree plot, and the "Component Pattern Profiles" plot. Interpret these plots and make the appropriate comments. See Chapter 3 of *Applied Multivariate Data Analysis* for a statistical reference to principal components analysis.

```
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;
```

Part 3: Cluster Analysis

We will begin our discussion of cluster analysis by making a pair of scatterplots.

```
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;
ods graphics off;

ods graphics on;
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;
```

How many clusters do you see in the scatterplot of FIN*SER? How many clusters do you see in the scatterplot of MAN*SER?

Clearly different projections of the data will produce different clustering results. We need to be cognizant of this fact.

Now we will use PROC CLUSTER to create a set of clusters algorithmically. Note that PROC CLUSTER performs *hierarchical clustering* (see Chapter 6 in *Applied Multivariate Data Analysis*) so we do not need to specify the number of clusters in advance. We will use the SAS procedure PROC TREE to assign observations to a specified number of clusters after we have performed the hierarchical clustering.

```
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;
```

How do we interpret the measures of CCC, Pseudo F, and Pseudo T-Squared? How do we interpret the plots for these three measures?

We can use PROC TREE to assign our data to a set number of clusters. Let's compare the output when we assign the observations to four clusters and then to three clusters.

```
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;
```

We will 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;
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;
```

Display the tables and comment on these results. Did the members of each membership group get clustered into the same cluster? Which number of clusters do you prefer?

Now perform a similar cluster analysis using the following cluster commands. Which of these cluster analyses do you prefer?

```
*****************************************************************;
* 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;
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;


%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);

%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);
```

Assignment Document:

All assignment reports should conform to the standards and style of the report template provided to you. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary results or information. The document should be submitted in pdf format. Name your file Assignment8_LastName.pdf.