

|  |
| --- |
| Assignment #4: Cluster Analysis  *MSDS 411* |

# **Data:**

The data for this assignment is the European employment data set. This data is posted in Canvas.

# 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:**

In this assignment we will learn how to perform an exploratory data analysis for a clustering problem, fit a hierarchical cluster analysis, fit a k-means cluster analysis, how to integrate principal components analysis and cluster analysis, how to use cluster analysis as a predictive model, and how to make a variety of R graphics applicable to cluster analysis and multivariate analysis in general.

# Assignment Tasks:

1. Read the European Employment data into R. Since the data set is small enough, you can easily view and comprehend the entire data set by simply printing it out. Print out the data. Be sure you understand the nature of this data and how it is similar/different from other datasets we’ve worked with.
2. Initial Exploratory Data Analysis: Since we have a relatively small number of variables, we will begin our exploratory data analysis with a pairwise scatterplot. Obtain a pairwise scatterplot of the data. Note that when you have a small number of variables, the pairwise scatterplot is a useful statistical graphic. Another note about scatterplots – they are not very useful when we have too many data points. A scatterplot is a more useful statistical graphic when you have 100 data points than when you have 1MM data points.

Since we are interested in applying cluster analysis to this data, we can use the pairs plot to scan the individual 2-dimensional views of the data. In cluster analysis we typically focus on 2D and 3D representations of the data in order to avoid the *curse of dimensionality*. With multivariate data as the dimension grows the distance between the observations grow, and it is difficult for the observations to be ‘close’ to one another, and hence be grouped into a small number of clusters.

Do you see any interesting 2D views of the data? What would be ‘interesting’? Remember, we are interested in applying cluster analysis so 2D plots that show clusters are the plots that would be interesting. Why don’t we consider MAN versus SER and SER versus FIN? Do these 2D views look interesting?

1. Visualizing the Data with Labelled Scatterplots**:** While the pairs plot allows us to scan all pairwise scatterplots easily and efficiently, it is not the ideal visualization of the data. After we have homed in on some interesting dimensions we can create more specialized plots for those dimensions. Specialized plots should always include labels and color. The objective is to compress more than two dimensions of information into a two dimensional plot.
2. Plot FIN versus SER. Do we see some clusters in this plot? How many clusters do we have? How many clusters would you have if you were creating a segmentation?
3. Plot MAN versus SER. Do we see some clusters in this plot? How many clusters do we have? Are they the same clusters as we saw in the previous plot? How many clusters would you have if you were creating a segmentation?
4. Of the two 2D views of the data which one do you think would be the better view for supervised clustering, i.e. using a clustering algorithm to create a classifier that will assign the countries to the correct class/label? Why?
5. Creating a 2D Projection Using Principal Components Analysis: We can use principal components analysis to reduce the dimension of the data. We can project the data down from 9D to 2D by performing PCA and using the first and second principal components. By doing so we are creating a new 2D view of the data, and a view of the data that contains information from more than two dimensions.
6. Use the raw data and conduct a PCA. Plot the first two principal components. How does this 2D projection of the data compare to the two other views of the data that we are considering? How many clusters does this 2D projection have? Clearly, our data can have different degrees of separation in different 2D profiles, and hence some low dimension representations will be better clustered than others.
7. Usually, one is supposed to standardize the data to be mean zero with unit variance before performing a PCA. Standardize the data and run a second PCA on the standardized data. Compare the two results. Does standardizing have much of an effect here?

Note: In general, we almost always want to standardize our data before we perform PCA to keep the variables with the largest scales from getting the largest loadings. Remember – large scale means large variance. However, in this case we have a type of data called *compositional data*. Compositional data represent the components of a whole. In our case the dimensions sum to 100, and each dimension represents a component of the economy. Here, large components in some dimensions will require small components in other dimensions in order to sum to 100. The nature of compositional data can cause a variety of problems. Most statistical methods are designed for continuous data, and the question is how ‘continuous’ is our compositional data. Hence, the need to check.

1. Hierarchical Clustering Analysis: Hierarchical clustering algorithms fit a tree of clusters from k=2 to k=N, where N is the number of data points in the sample.  As you know, this tree of clusters can be visualized using a dendrogram. Since the cluster tree stores all possible cluster assignments, we must cut the tree using cutree() to force an assignment of the observations to a particular number of clusters.
2. Perform a hierarchical cluster analysis and obtain a dendrogram. Use the cutree() function to force an assignment of the observations to a particular number of clusters. Use k=3 and k=6 and compare the classification accuracy of two cluster tree cuts. Which set of clusters is more accurate?
3. Perform the same analysis, but this time use the principal component space using the first and second principal components. Of these four ‘cluster models’ which one is the most accurate? Make a table to display their accuracy for easy comparison.
4. k-Means Clustering Analysis: Let’s perform the analogous cluster analysis and make a comparison.
5. Conduct a K-Means Cluster Analysis on the European Employment data for k=3 and k=6. Compare the classification accuracy of these models with the hierarchical models obtained in task (4).
6. For the k-Means Cluster Models obtain a plot that includes the original labels, their assigned clusters, and the cluster centers. What do you see in these two graphics?
7. Conduct a K-Means Cluster Analysis for k=3 and k=6, but use the Principal Components space.
8. What happens as we increase the number of clusters from k=3 to k=6?
9. Of these eight cluster models which is the most accurate? Make a table summarizing the eight models and their accuracy.
10. How do the clusters compare with the original *labels* (EU, EFTA, Eastern, or Other)?
11. Computing the ‘Optimal’ Number of Clusters by Brute Force: After completing the above cluster analyses, one should question whether or not k=3 or k=6 is the “best” choice for the number of clusters to retain. Unfortunately, the answer to that question is not as simple as the question. One idea that should be apparent is that we would need to be able to evaluate a large number of clusters bases on some criterion that allows an objective comparison. One option is to use the classification accuracy rate of our clusters.
12. Obtain and plot the classification accuracy for k=1 to k=20 for both hierarchical and k-means clustering algorithms. What can you conclude based on this graph?
13. On Your Own Modeling 1**:** The USSTATES dataset is a 12 variable dataset with n=50 records that you used briefly in MSDS 410. The data, calculated from census data, consists of state-wide average or proportion scores for the non-demographic variables. As such, higher scores for the composite variables translate into having more of that quality. There is no other information available about this data. Use this data set and conduct a hierarchical cluster analysis. Decide on the total number of clusters to retain and describe the differences amongst the clusters.
14. On Your Own Modeling 2: The RECIDIVISM dataset is an 18 variable dataset with n=1445 records. Please see the data description file for the variable definitions and additional information about the dataset. The data consists of a random sample records on convicts released from prison during 1977/1978. Use this data set and conduct a kmeans cluster analysis. Decide on the total number of clusters to retain and describe the differences amongst the clusters.
15. Please write a reflection on your cluster modeling experiences.

# Assignment Document:

All assignment reports should answer each of the questions separately. Please be sure to clearly indicate which question is being addressed. 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 R-code, intermediary computations, R-results, or non-essential information. The document should be submitted in pdf format. Name your file Assign4\_LastName.pdf.