**Assignment #8**
Andrew Knight

**Introduction:**

The goal of this assignment is to explore Cluster Analysis using European Employment data in R.

Data File: EuropeanEmployment.csv

Calculations File: EuopeanSkeletonCode_Knight.R

**Assignment Tasks**

**1: The Data**

In an initial view of the data, we see that we have nine distinct industry segments and thirty European countries listed with the employment percentage for each industry. We also have an additional factor variable representing the group, or area of the Europe that to which each county is assigned.

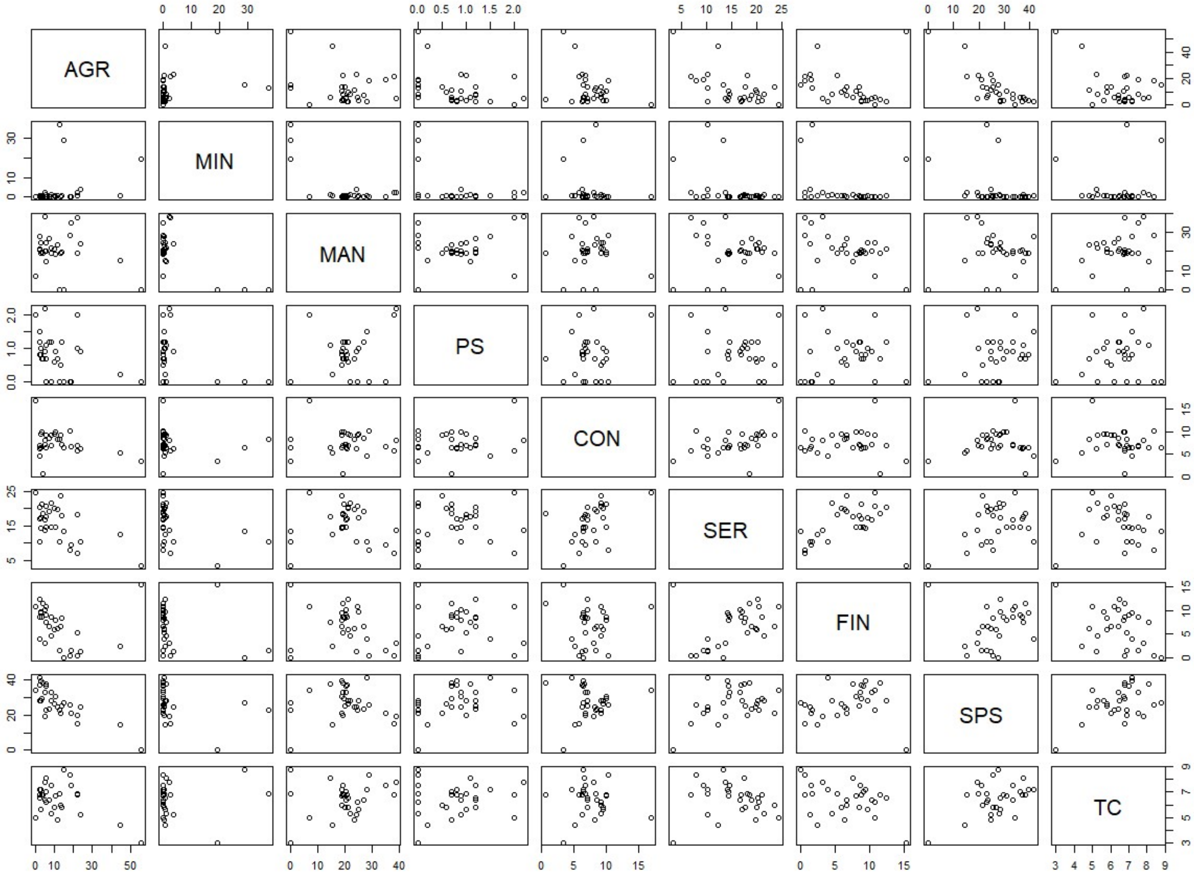
**2: Initial Exploratory Data Analysis**

Because this data set is small, we can start by performing an initial analysis just by looking directly at the raw data. Any trends or useful details we uncover may help inform our segmentation (or cluster) analysis later.

The first thing I notice in the raw data is that the European group that the country is assigned is not necessarily a good indicator of clusters related by industry. For example, the eight countries in the Eastern group do not share many common industry strengths whereas the six countries in the EFTA group do share some similarities.

The next general trend I see in the raw data is that many countries tend to have two or three industries which account for a majority of the employment.  No countries have close to an even distribution of employment across all industries. This seems to make sense given the relative strengths of the countries and the variety in the natural resources for each.

Moving on to a visual review of the data, we start with a pairwise scatter plot.



Unlike linear regression, we are not looking for linear relationships between pairs of industry sectors; we are looking for ‘interesting’ clusters or interesting patterns.

First off, agriculture and mining seems to have a largely binary grouping structure when compared to all other industries. This means that only a few countries have a significantly higher proportion of employment in agriculture or mining than other industries. For agriculture this includes Albania and Turkey. For mining this includes Hungary and Czech/Slovakia. Other pairwise comparisons do not give an immediate revelation. However, we do see some other grouping patterns. For example, when comparing Finance to other industries we see a general dispersion among the various countries, with the exception of mining and to an extent, agriculture. However, it’s difficult to gain much more detail from the big picture view without looking at individual scatterplots.

## 3: Visualizing the Data with Labelled Scatterplots

Using only two dimensions, we compare the financial versus services data and we find that there are basically two clusters of countries, with one outlier exception of Albania in the top left. The first cluster at the bottom is made up of mostly countries in Eastern and Other groups which seem to have higher percentages of services employment and lower percentages of financial employment. There are two exceptions from the Other group that actually fall within the second cluster, Cyprus and Gibraltar. The second cluster consists of countries with higher percentages of both finance and services than the other cluster. This second cluster is made up almost entirely of EFTA and EU groups. Because Albania is such an extreme outlier of the other two clusters, we would need three clusters.

Next, we compare manufacturing to services industry and compare scatterplot for the countries. Here we have less well-defined groupings. We still see a cluster made up almost entirely of EU and EFTA countries right in the middle. This shows relatively high percentage (centered around 20) for services and approximately the same for manufacturing. However, we see a couple Other group countries (Turkey and Cyprus) in close proximity to this cluster, perhaps even should be included in it. Unfortunately, we do not see a nice secondary cluster for the Eastern and Other groups. The Eastern group countries all have comparatively low services percentage but they span the entire spectrum of manufacturing employment. The Other group countries do not seem to show a grouping pattern for either services or manufacturing.

**4: Creating a 2D Projection Using Principal Components Analysis**

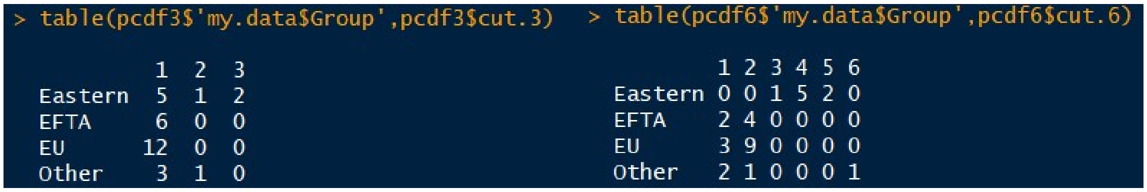
When reducing the dimensionality using PCA, we can plot the first two principal components and again look for agglomerative data. The clustering that we see in the 2D projection is different than in the original data. One way it differs Is that it seems like the separation between the clusters has been amplified a bit. This makes it slightly easier to differentiate the cluster boundaries.

However, for some principal component comparisons it can still be difficult to determine the best boundary definition or the even the number of clusters. For clarity on this, we can take a look at the hierarchical analysis and k-means analysis in the following sections.

**5: Hierarchical Clustering Analysis**

The hierarchical clustering analysis can be performed by producing a dendrogram. This gives the tree-like structure that can be used to determine cutoff lines. The number of clusters can be determined by the height of the cutoff line for relatively small numbers of variables.

When using k=3, we can cut our data into three clusters and each country is placed in one of the three cluster groups. Again, we find that our three clusters do not really correspond to the groups of countries labeled EU, EFTA, etc. For example, our cut cluster 1 has countries from every group and our cut cluster 2 has only two countries from two different groups. It isn’t until cluster 3 that we see two countries from the same country group.



```
> table(pcdf3$'my.data$Group',pcdf3$cut.3)

           1  2  3
  Eastern  5  1  2
  EFTA     6  0  0
  EU      12  0  0
  Other    3  1  0
```

```
> table(pcdf6$'my.data$Group',pcdf6$cut.6)

          1 2 3 4 5 6
  Eastern 0 0 1 5 2 0
  EFTA    2 4 0 0 0 0
  EU      3 9 0 0 0 0
  Other   2 1 0 0 0 1
```

When using k=6, we again cut our data, this time into six distinct clusters. Here again we get the first two clusters with a mix of countries but we see something a little different for the Eastern countries. There are none in the first two, but they exist alone in the next three clusters. Now, we have to be careful about making assumptions based on the above information. For example, without viewing the PC clusters on a scatter plot, we should not assume that these clusters are close to each other.

However, we can get some measure of accuracy of the clustering method used. Calculating the Between Sums of Squares percent for both k=3 and for k=6, we get the following.

| **Hierarchical** | **SS %** |
|---|---|
| k = 3 | 0.5893 |
| k = 6 | 0.8421 |
| k = 7 | 0.8720 |

While the relative measure of accuracy should increase as we add more clusters, we must be mindful of the increasing complexity of more clusters as well.

**6: K-Means Clustering Analysis**

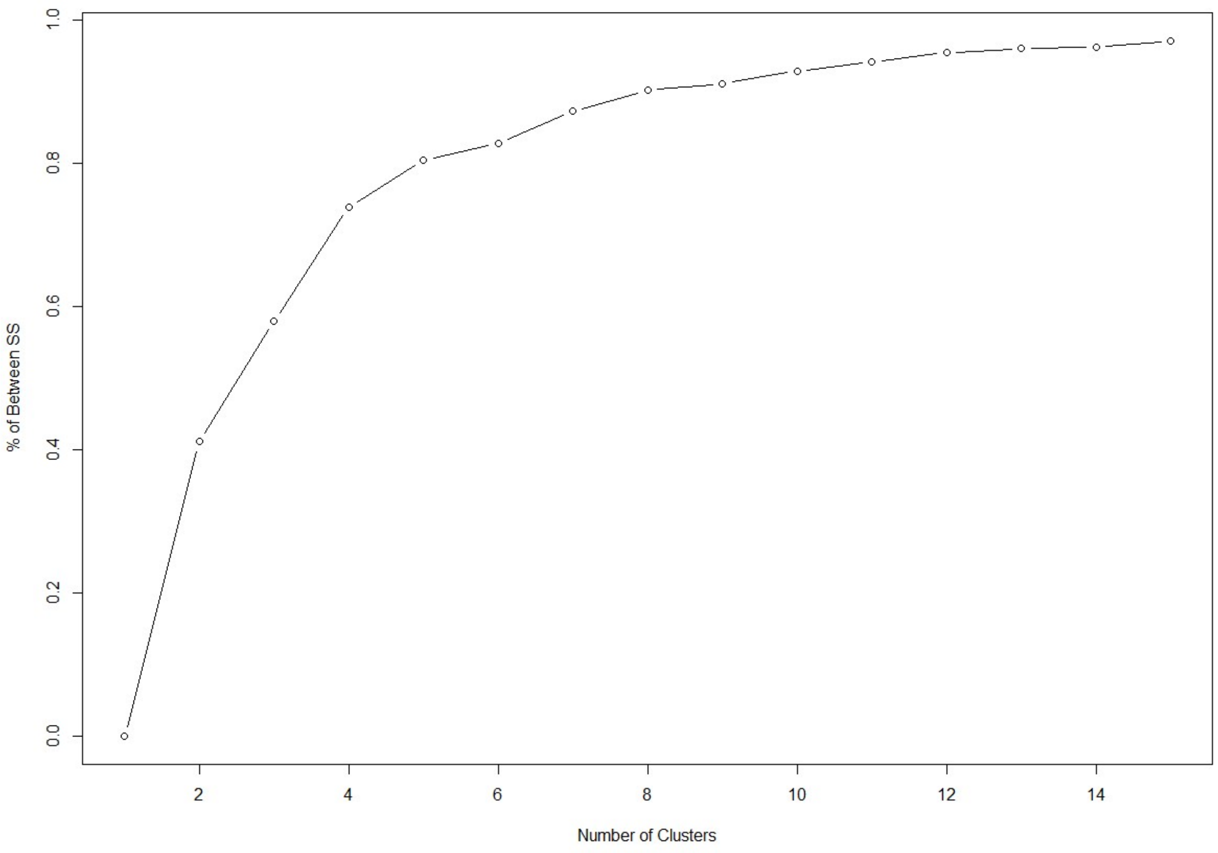
When we use K-Means clustering, we again are interested in determining the ‘right’ number of clusters. K-Means gives another quick way to identify the clusters based on the first two principal components. In this case, it automates the k-value selection but requires many more computing resources to achieve it so we need to be mindful of the size of the dataset we are working on. This is because k-means seeks to minimize the within-group sums of squares over all variables.

| **K-Means** | **SS %** |
|---|---|
| k = 3 | 0.5793 |
| k = 6 | 0.8308 |
| k = 7 | 0.8498 |

Comparing the k-means sums of square percentages to the hierarchical percentages we see they are similar even if k-means are a bit lower. One other thing I notice here is that it does not grow quite as quickly as the hierarchical method as you add additional clusters. This may be data dependent but the important thing to remember is the cluster method is intended to simplify the view of the groupings so that the analyst can make sense of them in an easier way. As you add more clusters, I think it is harder to do that regardless of the clustering method chosen. With this in consideration, the hierarchical method using seven clusters gives the most accurate results based on the sums of squares percentage value.

## 7: Computing the Optimal Number of Clusters

The plots of the number of clusters for both hierarchical and k-means are very similar. I have shown the k-means cluster plot here for reference. A couple of things to note, first of all we will see that if there is a business objective to account for a certain level of variance as we did in principal component analysis, then we could make a decision based on the threshold. If for example we wanted 80% variation explained, then we would choose five clusters and try to identify those clusters in a way that make sense. On the other hand, if we had a number of clusters in mind, for example four groups, then we could see approximately how much variance is explained using those four clusters. In either case, it would be wise to compare these two (and probably other) methods to find the most accurate way to cluster the countries.



As in any clustering exercise, we must be mindful that better accuracy does not usually mean a better choice in the k number. Choosing a higher number of clusters, say above eight, does not help us with a better definition of the trends associated with each cluster. If we try to make some assumptions about two or three different cluster groups, I would think that in most cases, the value of differentiation of groups diminishes significantly. To relate back to the original data, I doubt that splitting countries into 12 or more groups would help us draw any conclusions about employment breakdown between the nine industries. In this case, more than nine groups seems like an unnecessary exercise.

If I were trying to form some insightful conclusions about employment I would take a close look at the nature of work in the various industry segments and see if my assumptions matched the data. For example, could a trend be found that assigns each country a relative strength in each industry based on separating out knowledge workers (traditionally referred to as white collar), vs task workers (colloquially defined as blue collar)? Would we find that countries with higher percentages of employment in agriculture, mining, and manufacturing are clustered together? Perhaps countries more focused on infrastructure would see a higher percentage of employment in construction, transport and power and water supply. Finally, countries traditionally strong in finance and services should be grouped together, right? We don't necessarily need a large number of clusters for the results to useful. Based on my short analysis of these data, this seems to be the case. Countries in the EU and EFTA group tend to have higher employment in finance, services, and social & personal services whereas countries in Eastern tend to have slightly higher concentrations in agriculture and manufacturing.

**Reflections & Conclusion**

We must be careful in using cluster analysis as a method for finding additional insight in the data. It can be a useful tool for breaking down complex data into chunks of discrete buckets with which we can perform further analysis. Caution is needed so as to not misconstrue the clusters or attempt to create to many of them. I think it is similar to principal component analysis in that the idea is to try to find a smallest number of groupings that makes the data more informative and easier to interpret.