Assignment #8: Cluster Analysis
Prabhat Thakur

Introduction:

The objective of this assignment is 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.

Data: The data for this assignment is the European employment data set posted in Canvas.
This data set contains employment in various industry segments reported as a percent for thirty European nations.

Code: Sample code for cluster analysis on European employment data set was provided in EuropeanSkeletonCode.R

Assignment Tasks:

(1) The Data:

After loading the data set in R data frame we can see that, there are 30 observations and 11 variables. 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.

With this type of data we could perform a segmentation, an *unsupervised learning* problem where we assign the countries to different groups based on their similarity as determined by a clustering algorithm, or we could define a classification problem and use a clustering algorithm as a *supervised learning* algorithm that would predict the class/label of each observation based on its cluster assignment.

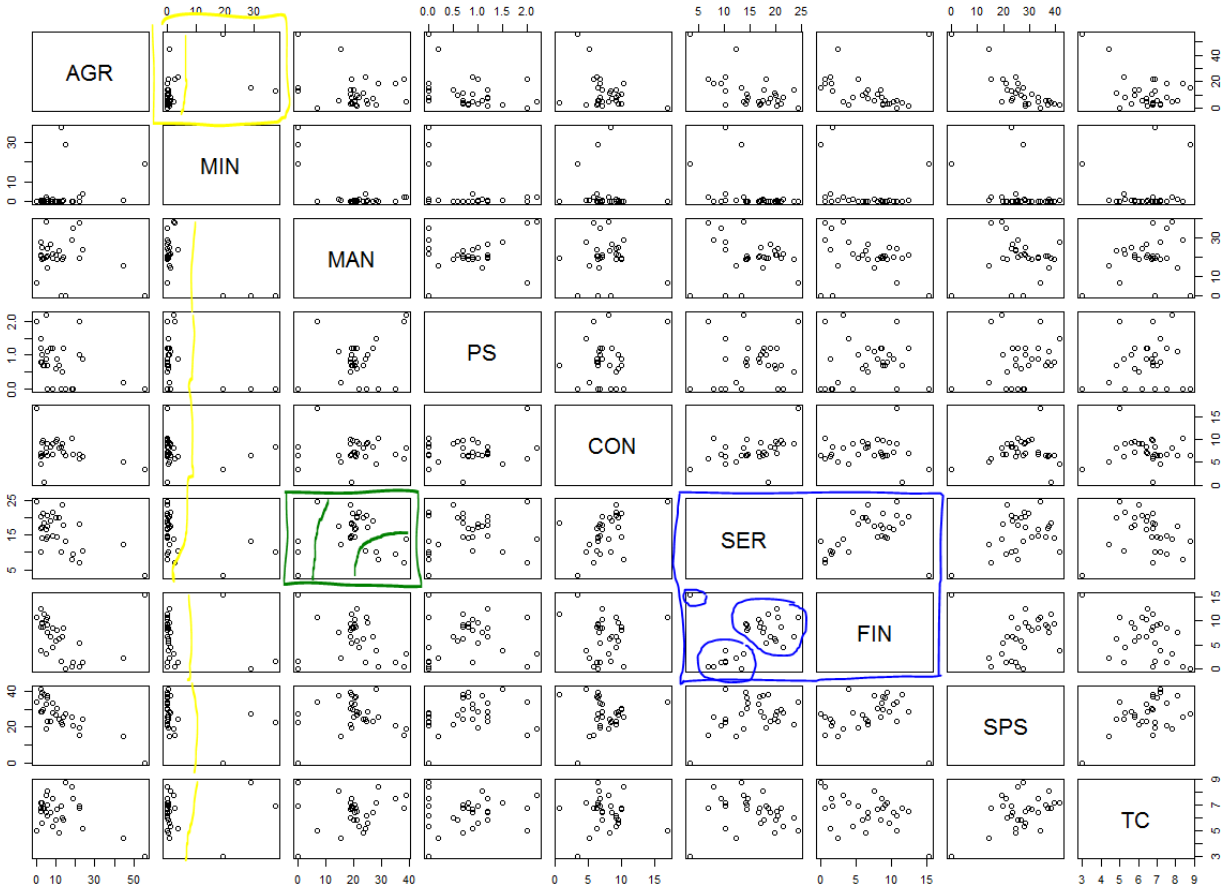
(2) Initial Exploratory Data Analysis:

Since the data set is small enough, we can easily view and comprehend the entire data set by simply printing it out. We can begin our exploratory data 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.

We are trying to cluster European countries based on employment in different industries. From looking at data, Countries classified in European group 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, Czech and Hungary have low employment in AGR and high employment

in MIN compared to other countries in that group, similar differences can be seen in other groups. Another trend we can see that employment % is not distributed evenly across all industries. Many countries tend to have two or three industries which accounts for a majority of the employment which make sense given countries relative strengths in different industry sectors and available natural resources.

For visual review of our data, given small number of variables, we can use pairwise scatterplot for further analysis. 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. Unlike linear regression, we are not looking for linear relationships/correlation between pairs of industry sectors; we are looking for 'interesting' clusters or interesting patterns.



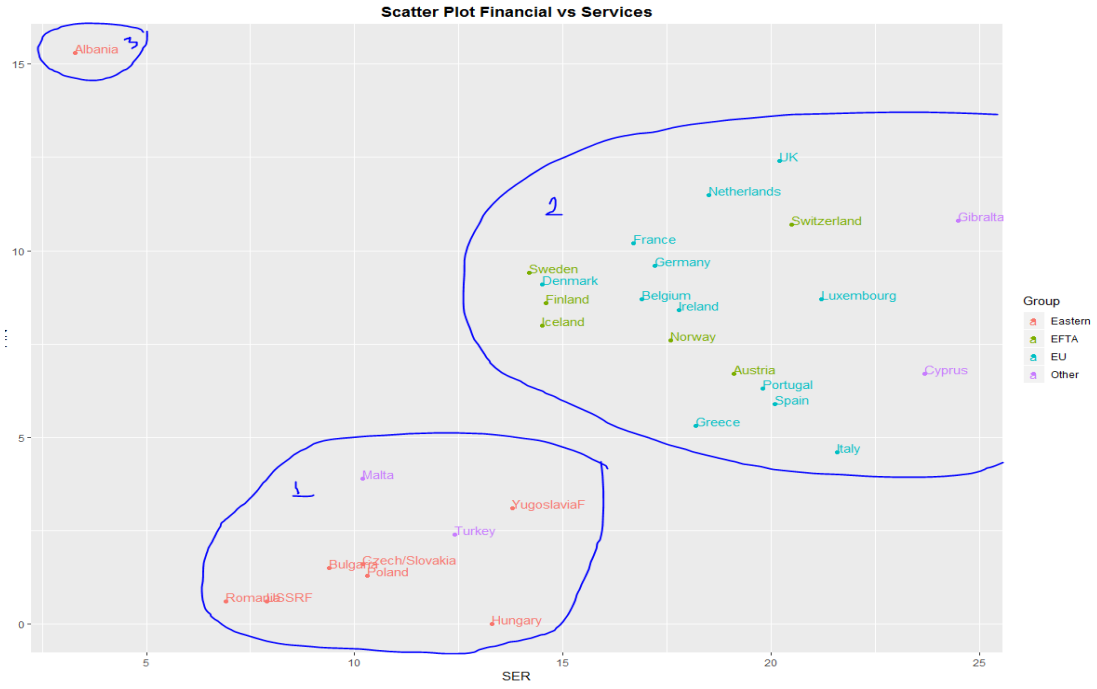
Looking at above pairwise scatter plots, we can see some interesting cluster patterns in some industry pairs. Scatter plot of Mining vs all the industries all seems to have largely two clusters located in two sides of the plot. Which suggests only a few countries have a significantly higher proportion of employment in mining than other industries. Also, countries where employment in Mining sector is high, other industries are not so strong. For mining this includes Albania, Hungary and Czech/Slovakia which are in the right side of the yellow line in above pairwise scatter plots.

Looking at other pairwise scatter plots, we can see some cluster patterns in MAN vs SER, FIN vs SER and AGR vs FIN. From SER vs FIN, it appears that there are largely two groups of countries, one with low employment in SER and FIN and second in high employment in both. Other pairwise comparisons do not revile much information. We will look into these discussed pairwise scatter plots in more details in next section.

(3) Visualizing the Data with Labelled Scatterplots:

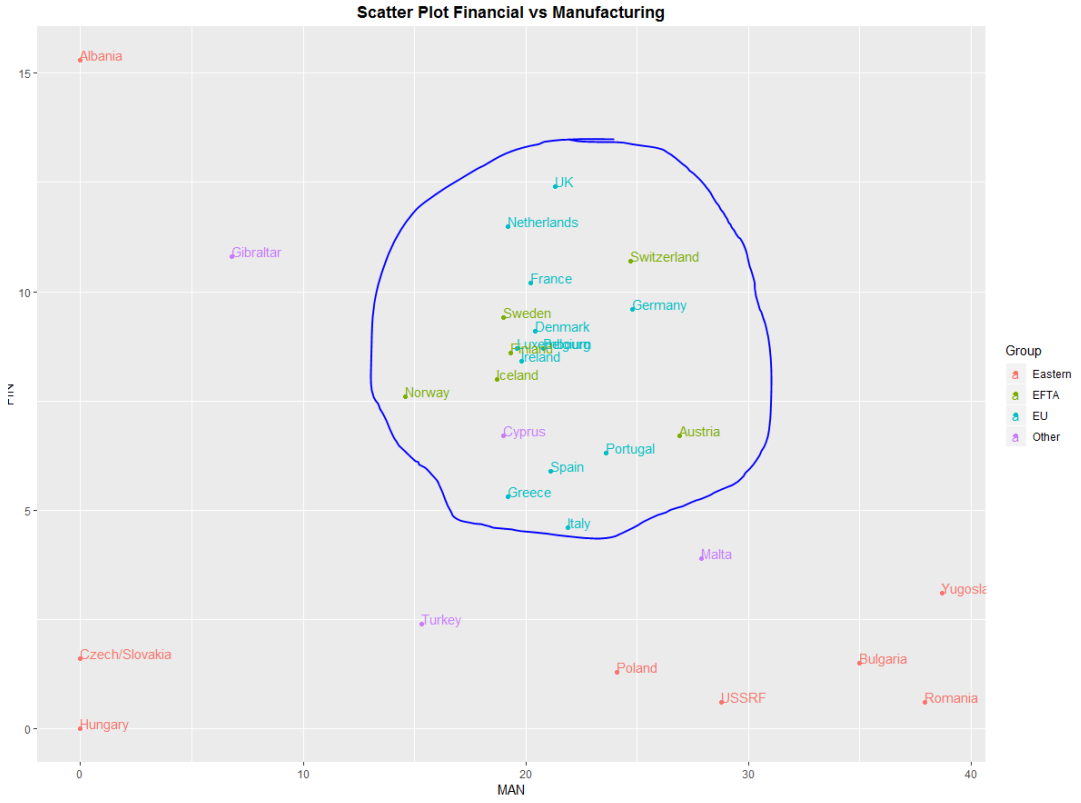
While the pairs plot allows us to scan all of the pairwise scatterplots easily and efficiently, it is not the ideal visualization of the data. After we have honed in on some interesting dimensions we can create more specialized plots for those dimensions.

Let's review FIN versus SER plot. We can see that there are basically two clusters of countries, with one 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. Looking at the second cluster, there are two exceptions from the Other group that actually fall in it, 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.

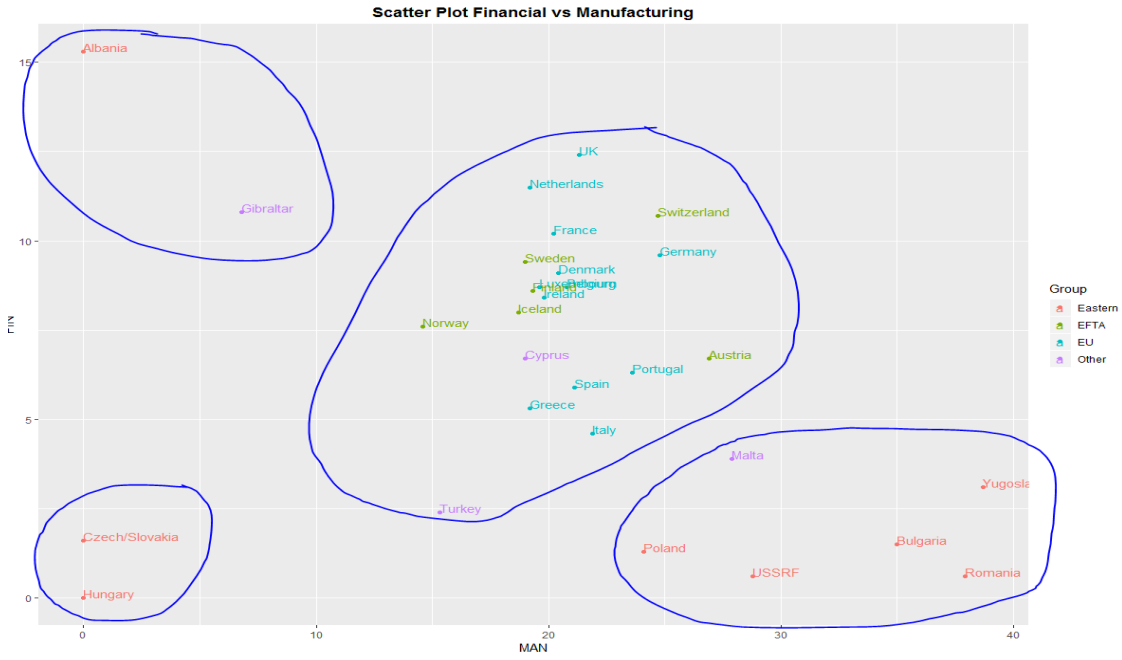


Next we will look at MAN versus SER plot. Here we have less well-defined groupings and countries are scattered all over the plot. However, we can see a cluster made up almost entirely of EU and EFTA countries right in the middle. This shows relatively high percentage (centered around 20) for services and approximately 8% for manufacturing. One country Cyprus from Other group countries is

part of this cluster. 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.



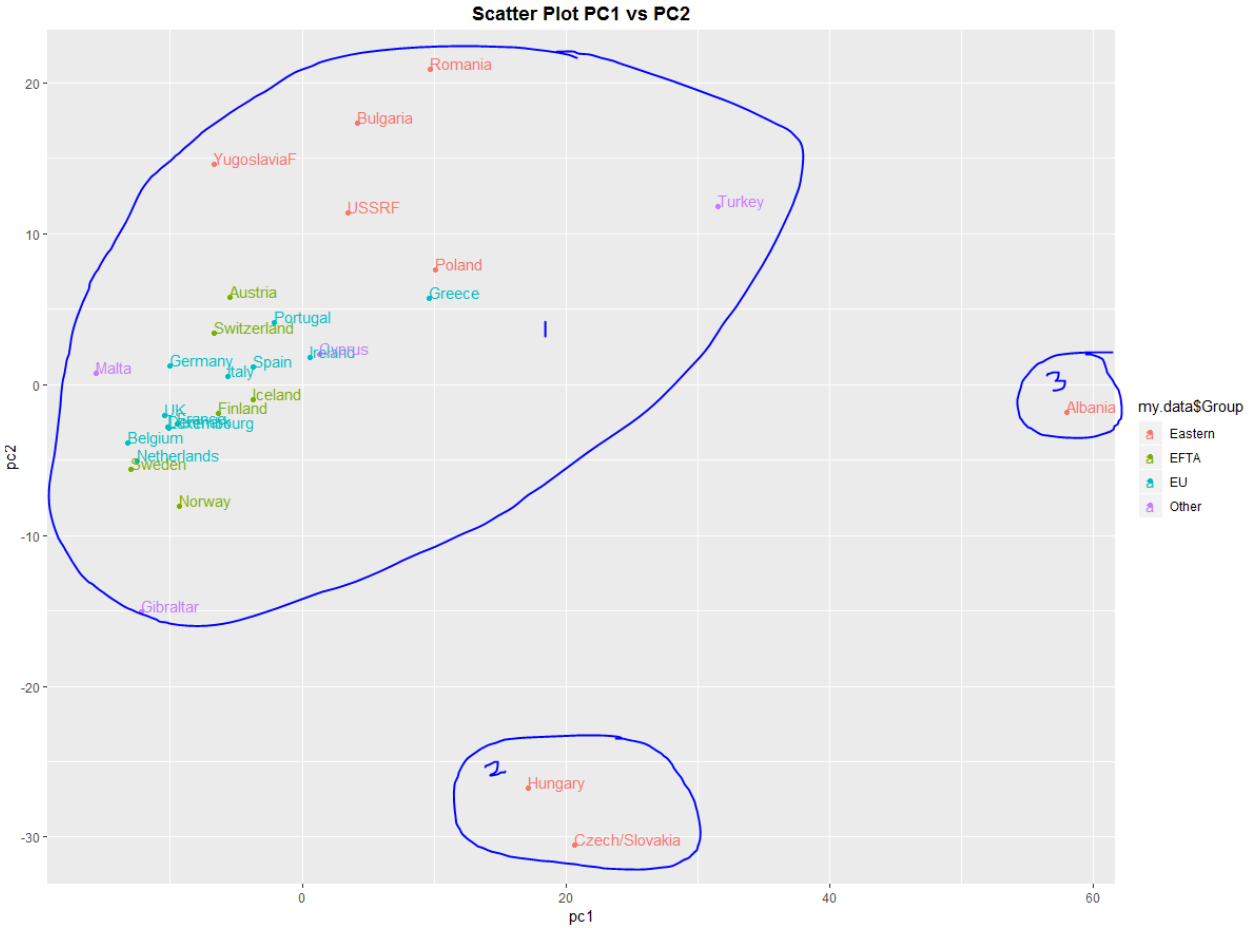
I would create at least 4 clusters to group all countries in different segments.



If we have to use these predictor for classification of countries using supervised clustering algorithm, FIN vs SER (1st scatter plot) will be a good choice. As explained above, there are three well separated clusters of countries. The algorithm can classify countries to one of these clusters with more accuracy.

(4) Creating a 2D Projection Using Principal Components Analysis:

In this step we will use principal components analysis (PCA) to reduce the dimension of the data from 9D to 2D and use 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.



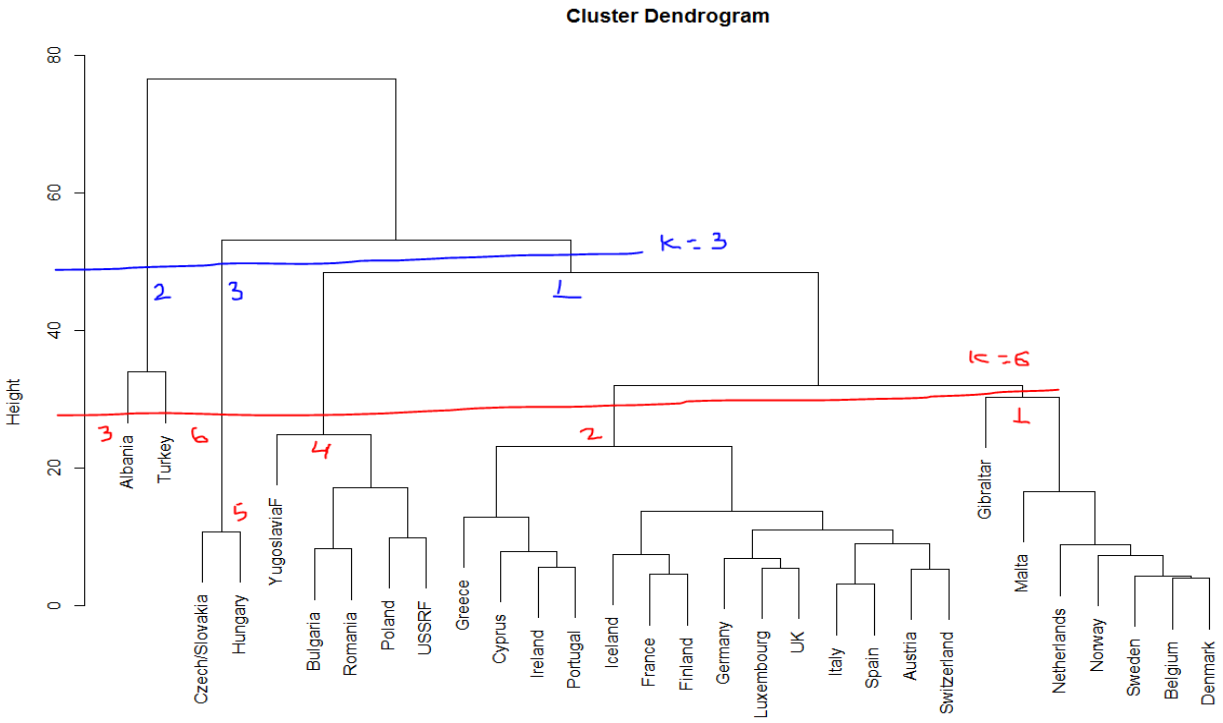
The clustering that we see in the 2D projection using 1st and 2nd principle components is quite different than FIN versus SER but very similar to what we saw in MAN versus SER plot from the original data. It does show 3 cluster and possibly 4. Cluster 1 marked in above could be divided in 2 clusters but it is difficult to create a clean grouping due to their close proximity. It is interesting to see that we can see similar grouping using 2D projection.

(5) Hierarchical Clustering Analysis:

In this step we will use 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. This tree of clusters can

be visualized using a dendrogram, and all software programs that have a hierarchical clustering algorithm should produce a dendrogram. When the data is small enough, then dendrograms are useful for visualizing the tree of clusters. However, like many statistical graphics, when the data gets large (large N) the tree, and hence the dendogram, becomes too large to be an effective display of the clusters.

Let's cut the tree to k=3 and k=6 and compare the classification accuracy of two cluster tree cuts.



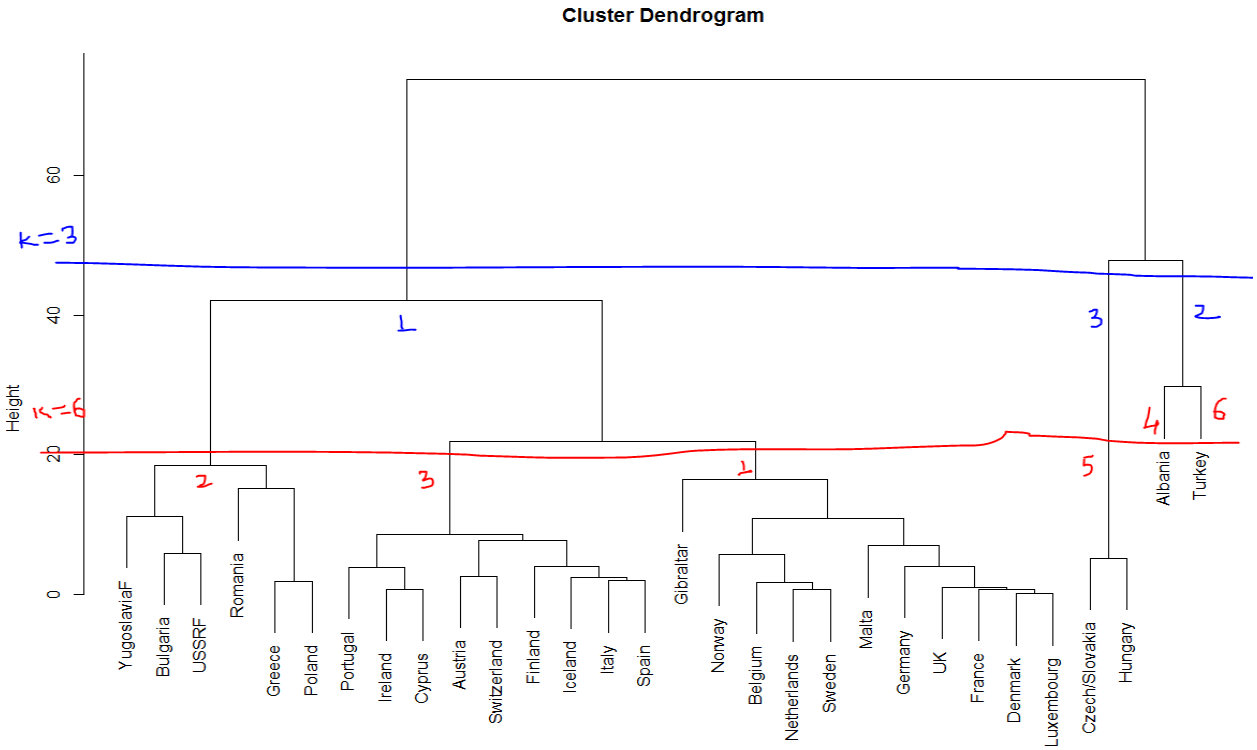
When using k=3, we can cut our data into three clusters and each country is placed in one of the three cluster groups. Again, we find that our three clusters do not really correspond to the groups of countries labeled EU, EFTA, etc. For example, our cut cluster 1 has countries from every group and our 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
```

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 not in the first two, but they exist alone in the next three clusters.

```
> 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
```

Now let's perform the same analysis in the principal component space using the first and second principal components.
Below is the dendogram with K=3 and K=6 cut.



When K=3, PCA clusters are exactly same as original data however there are some difference in when K=6. For example in cluster 2, there is one EU country along with 5 Eastern countries. Where as in original data, cluster 4 only has 5 Eastern countries.

```
> table(PCApcdf3$'my.data$Group',PCApcdf3$cut.3)

          1  2  3
  Eastern  5  1  2
  EFTA     6  0  0
  EU      12  0  0
  Other    3  1  0
> table(PCApcdf6$'my.data$Group',PCApcdf6$cut.6)

          1 2 3 4 5 6
  Eastern 0 5 0 1 2 0
  EFTA    2 0 4 0 0 0
  EU      7 1 4 0 0 0
  Other   2 0 1 0 0 1
```

Let's compare the accuracy of these models by calculating percentage of within-group sums of squares. Closer to 100% means we got good clusters in our model and closer to 0% means model doesn't have good clusters. Below is the matrix of Between Sums of Squares percent for both 9D and 2D data for k=3 and for k=6. From the matrix we can see that Model with 2PCs and 6 clusters is most accurate.

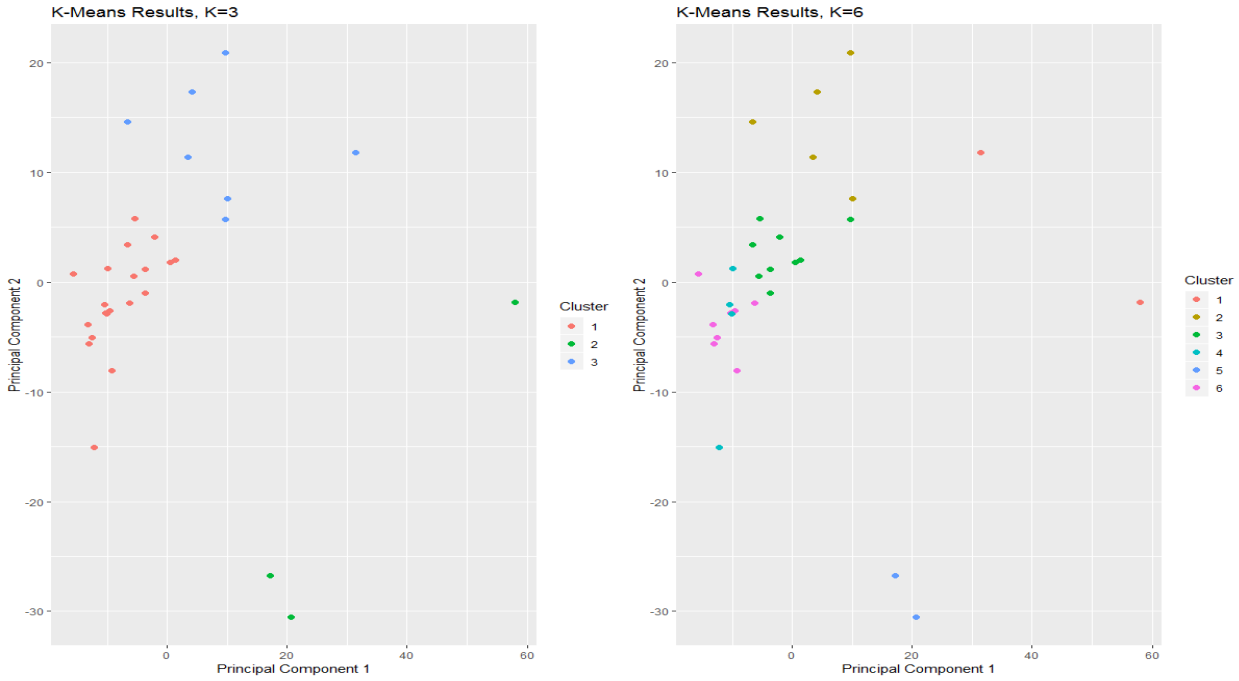
Accuracy Matrix (SS%)	Original Data	PCA with two PCs
K=3	0.5893374	0.6759739
K=6	0.8421061	**0.9291546**

(6) k-Means Clustering Analysis:

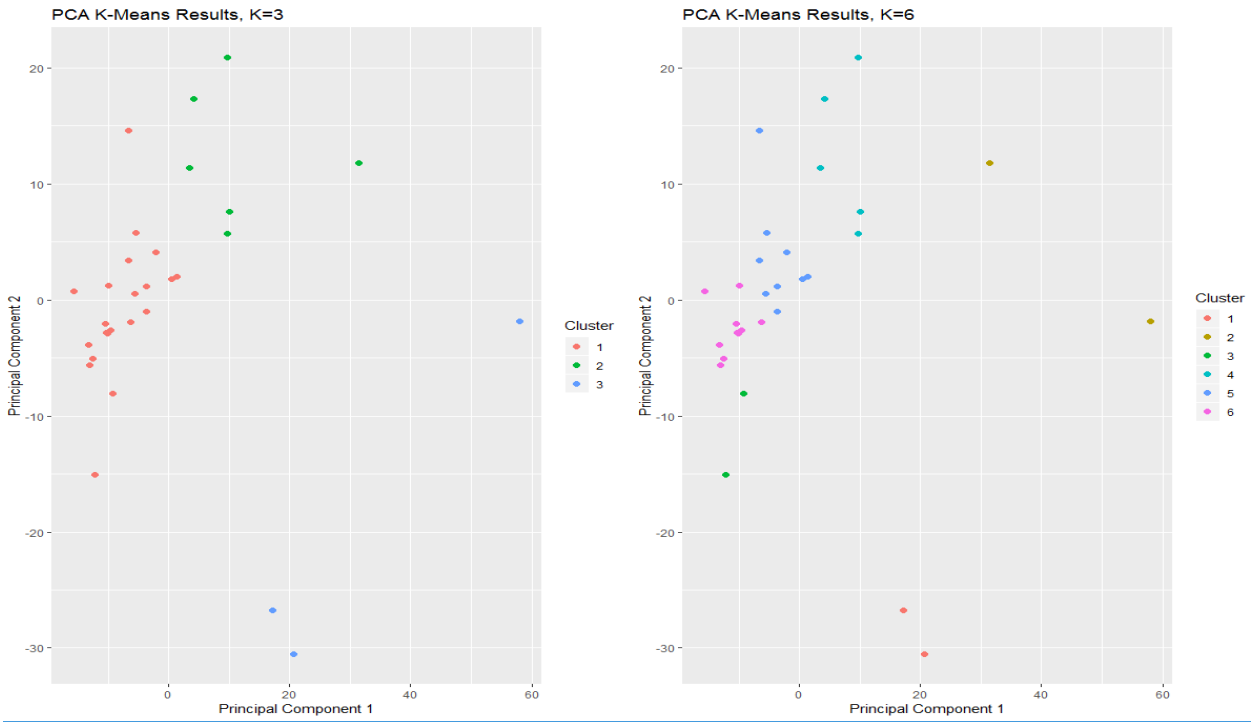
In this step we will use k-means clustering analysis. It is a good idea to use multiple methods for clustering. Hierarchical clustering computes a full cluster tree for k=2 to k=N, it is a computationally expensive clustering technique that cannot be used on larger data sets. Clustering methods that partition the data into k clusters for a specified k are more applicable to larger data sets since they are more computationally efficient. One of, if not THE, most popular clustering technique of the partitioning type is the k-means algorithm.

Let's perform the analogous cluster analysis using k-means for k=3 and k=6. This will allow us to compare the classification accuracy of our different cluster models. 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.

Below are the plots for k=3 and k=6 using original data. Clusters are color coded for each cluster. Clusters are clarity visible when k=3 but there are some overlapping cluster boundaries when k=6 for example cluster 4 and 6. When we increase k values the cluster boundaries starts to overlap.



Below are the K-means cluster plots using 2 principal components PCs. When K=3, there is 1 country which is moved to cluster 1 now. When K=6, with PC data, we can see all 6 clusters clearly and there is no cluster boundary overlapping.



Similar to above step, we will calculate the within-group sums of squares to compare these models accuracy. Below is the matrix of Between Sums of Squares percent for both 9D and 2D data for k=3 and for k=6 for models from both Hierarchical clustering and k-means clustering. 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.

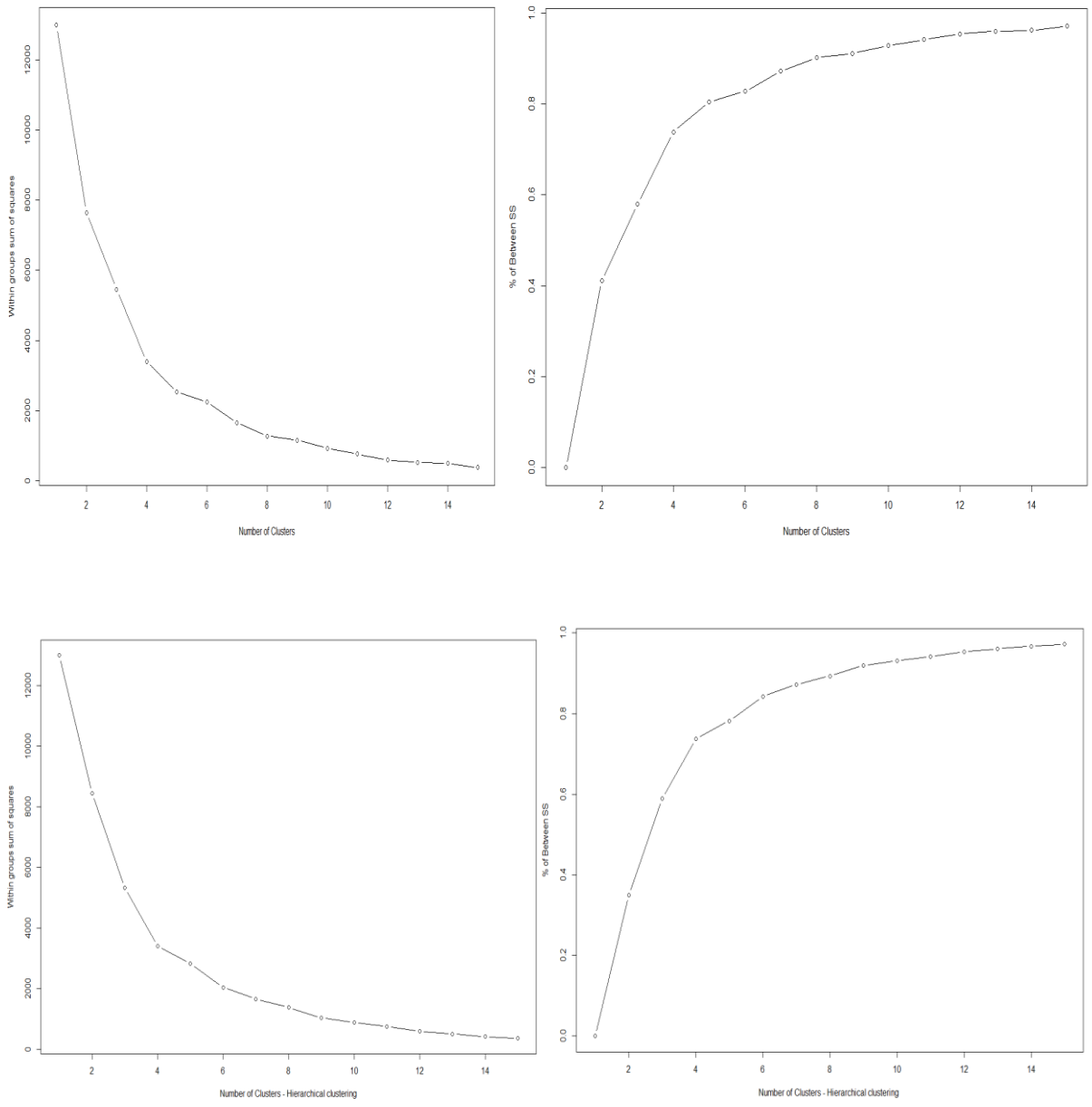
Accuracy Matrix (SS%)	Hierarchical		K-means	
	Original Data	PCA with 2 PCs	Original Data	PCA with 2 PCs
K=3	0.5893374	0.6759739	0.5792964	0.6872469
K=6	0.8421061	0.9291546	0.8341866	0.9017515

From the matrix we can see that Hierarchical clustering using 2 principal component and k=6 gives the most accurate results based on the sums of squares percentage value.

(7) Computing the 'Optimal' Number of Clusters by Brute Force:

Determine the correct number of clusters 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. In our problem we can use the classification accuracy rate of our clusters.

Here we have plotted out the classification accuracy for both the hierarchical and k-means clustering algorithms for k=1 to k=15. Overall we can see that the classification accuracy tends to increase as the number of clusters increase, but the classification accuracy is not strictly monotone.
The plots of the number of clusters for both k-means and hierarchical are very similar. First two plots are k-means and last two are hierarchical. From both the plots, we can see that after 4 or 5 clusters there is no significant gain.



As in any clustering exercise, we must be mindful that better accuracy does not usually mean a better choice in the k number. Overall, the data can often only provide rough guidance regarding the number of clusters we should select; consequently, we should rather revert to practical considerations.

However, first and foremost, we should ensure that our results are interpretable and meaningful. Not only must the number of clusters be small enough to ensure manageability, but each segment should also be large enough to warrant strategic attention.

Reflections & Conclusion:

I found Cluster Analysis very interesting and useful. Whatever approach we decide to perform cluster analysis and choosing correct number of clusters, we should always keep in mind that cluster analysis is primarily an exploratory technique. Thus, practical considerations are of utmost importance when deciding on the number of clusters.