Assignment #8

Syamala Srinivasan

Introduction:

Data can often be separated into logical categories to assist with analyzing. Employment in Europe can be segregated into multiple groupings based on location. These groups include European Union (EU), European Free Trade Association (EFTA), Eastern Europe (Eastern) and other.

Additionally the types of employment in each of the above groupings can be categorized as well.

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

These data points can be combined into clusters based on similarity and further analyzed. This assignment will explore the creation of these clusters.

Results:

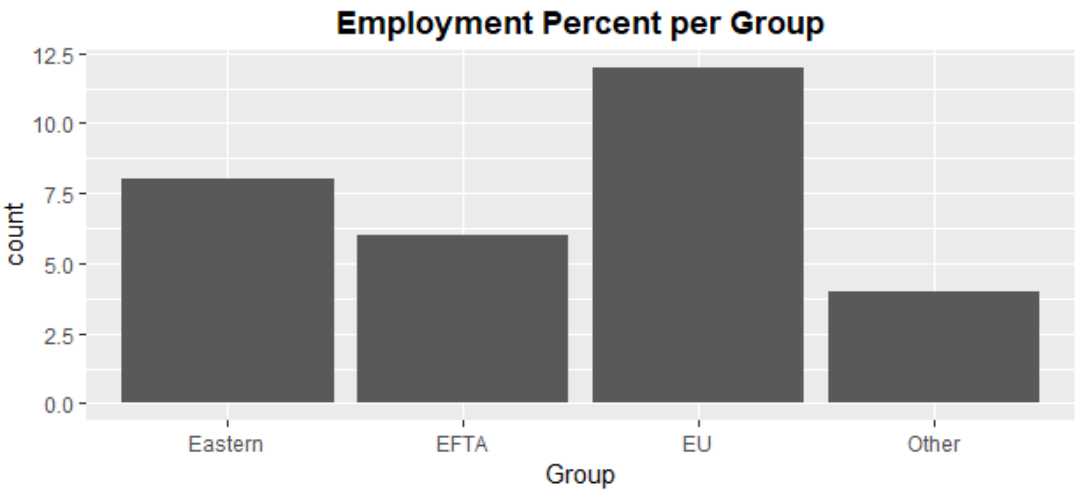
Task 1:

The data-set to be analyzed can be broken down into four categories; EU, EFTA, Eastern, and other. These categories can be utilized further to conduct cluster analysis.

```
          Country       Group        AGR              MIN              MAN              PS              CON              SER
 Albania       : 1   Eastern: 8   Min.   : 0.00   Min.   : 0.000   Min.   : 0.00   Min.   :0.000   Min.   : 0.60   Min.   : 3.30
 Austria       : 1   EFTA   : 6   1st Qu.: 4.40   1st Qu.: 0.125   1st Qu.:19.00   1st Qu.:0.275   1st Qu.: 6.40   1st Qu.:12.62
 Belgium       : 1   EU     :12   Median : 8.45   Median : 0.500   Median :20.30   Median :0.800   Median : 7.05   Median :16.80
 Bulgaria      : 1   Other  : 4   Mean   :12.19   Mean   : 3.447   Mean   :20.29   Mean   :0.800   Mean   : 7.53   Mean   :15.64
 Cyprus        : 1                3rd Qu.:14.93   3rd Qu.: 1.050   3rd Qu.:24.55   3rd Qu.:1.175   3rd Qu.: 9.10   3rd Qu.:19.62
 Czech/Slovakia: 1                Max.   :55.50   Max.   :37.300   Max.   :38.70   Max.   :2.200   Max.   :16.90   Max.   :24.50
 (Other)       :24
      FIN              SPS              TC
 Min.   : 0.000   Min.   : 0.00   Min.   :3.000
 1st Qu.: 3.300   1st Qu.:22.95   1st Qu.:5.800
 Median : 7.150   Median :27.00   Median :6.750
 Mean   : 6.650   Mean   :26.99   Mean   :6.453
 3rd Qu.: 9.325   3rd Qu.:33.17   3rd Qu.:7.150
 Max.   :15.300   Max.   :41.60   Max.   :8.800
```

Before beginning the analysis on this data the data will be reviewed.

Initially the country variable will be reviewed. There are 24 countries, each a part of Europe but separate groupings within Europe. The below graph indicates how many countries fall into each category. The majority of countries are a part of the European union while the least fall into the “other” category.



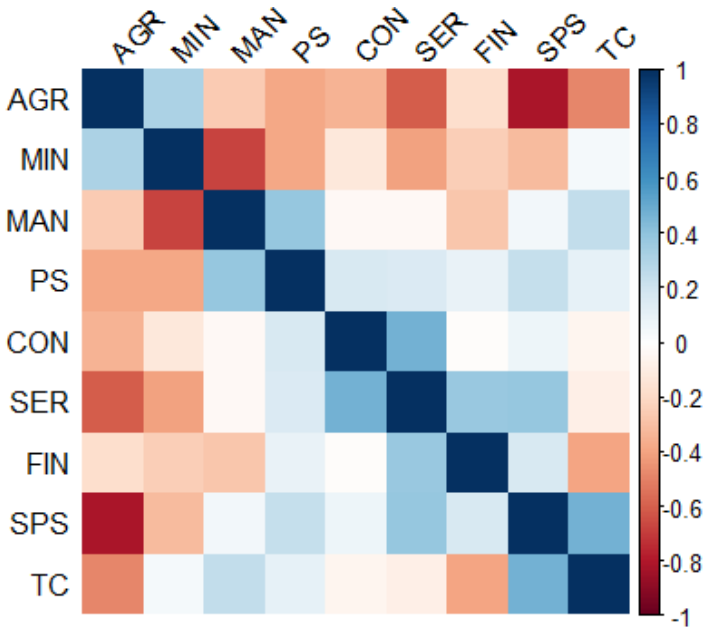
The remaining attributes describe the various sectors of employment among European citizens.

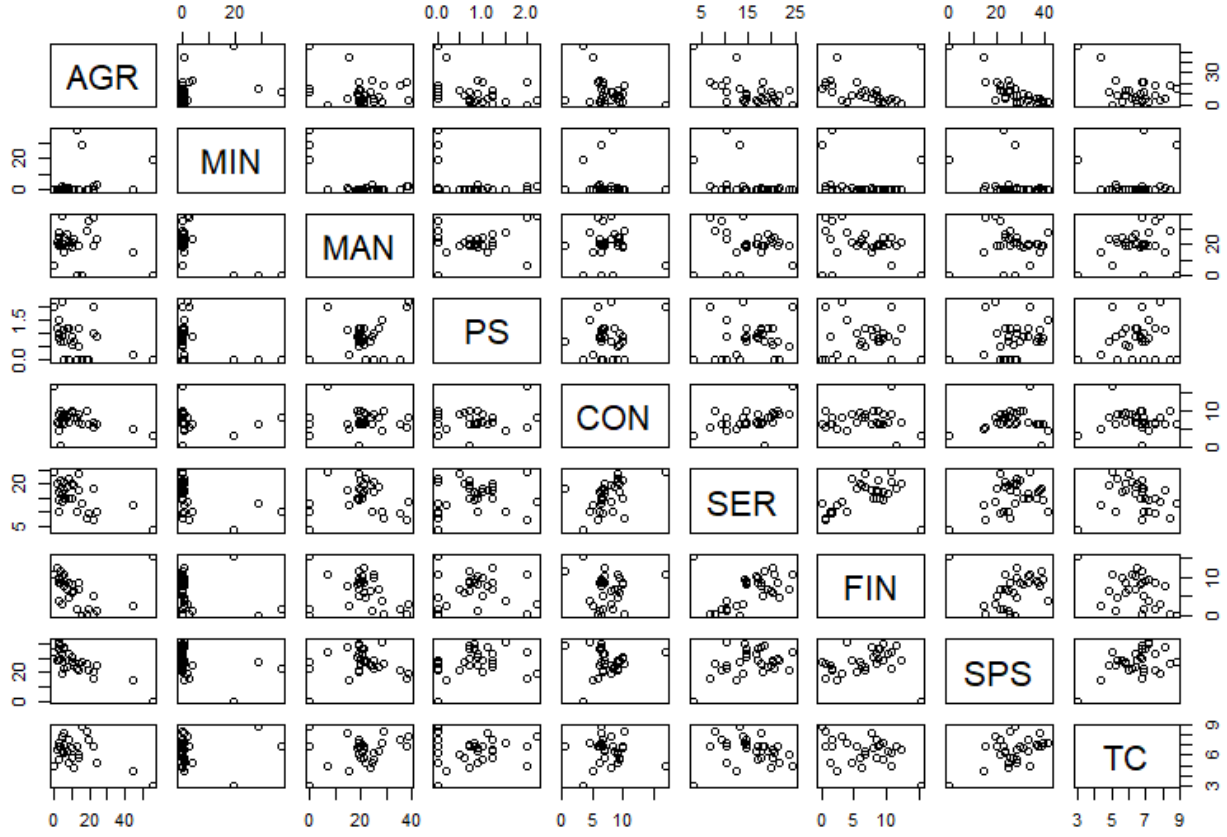
The employment distribution between countries with the largest mean value is social and personal services indicating that this profession has the high percentage of employment among the European groupings.

Task 2:

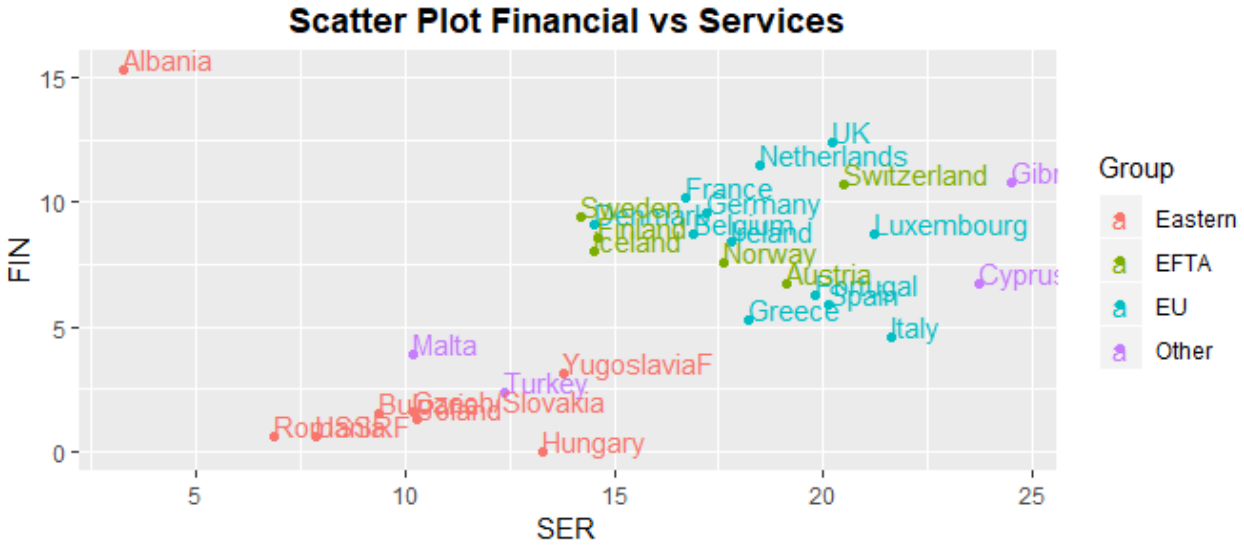
The scatterplot matrix pictured below initially is a lot to taken in due to the large number of variables. One trend that is immediately noticeable is that every combination including mining is skewed to one side. This indicates that mining is likely to be significantly smaller than all other employment opportunities.

The relationships can also be illustrated utilizing a correlation plot. Using the below matrix we can see that there is a very strong negative relationship between Social/Personal services and Agriculture as well as mining and manufacturing. Other than these relationships there are not many others that have significant relationships.

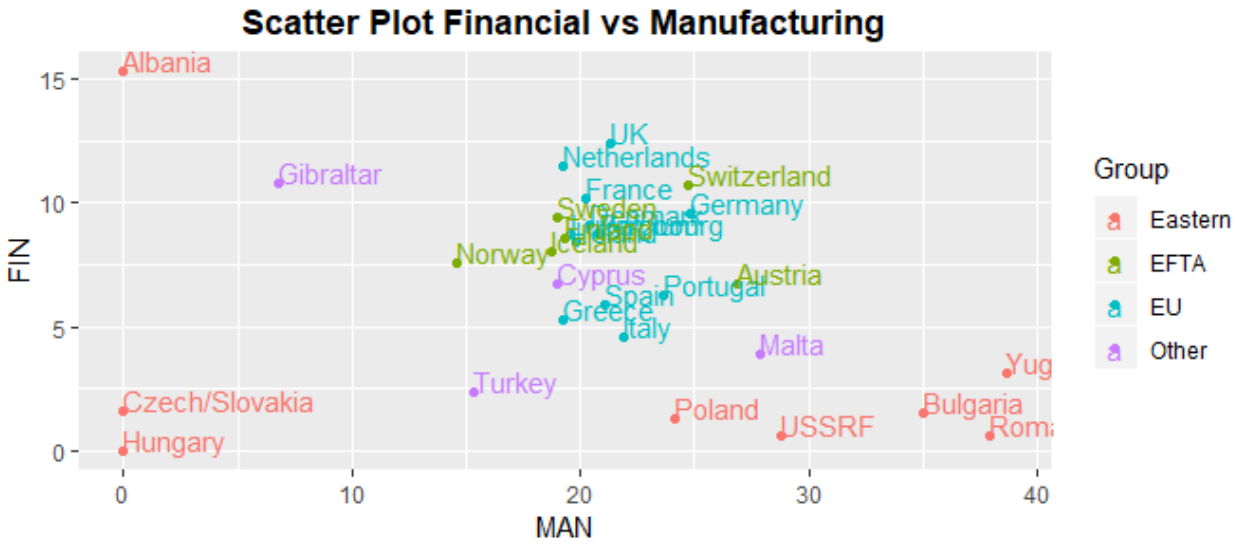

AGR
MIN
MAN
PS
CON
SER
FIN
SPS
TC
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


AGR
MIN
MAN
PS
CON
SER
FIN
SPS
TC

Task 3:



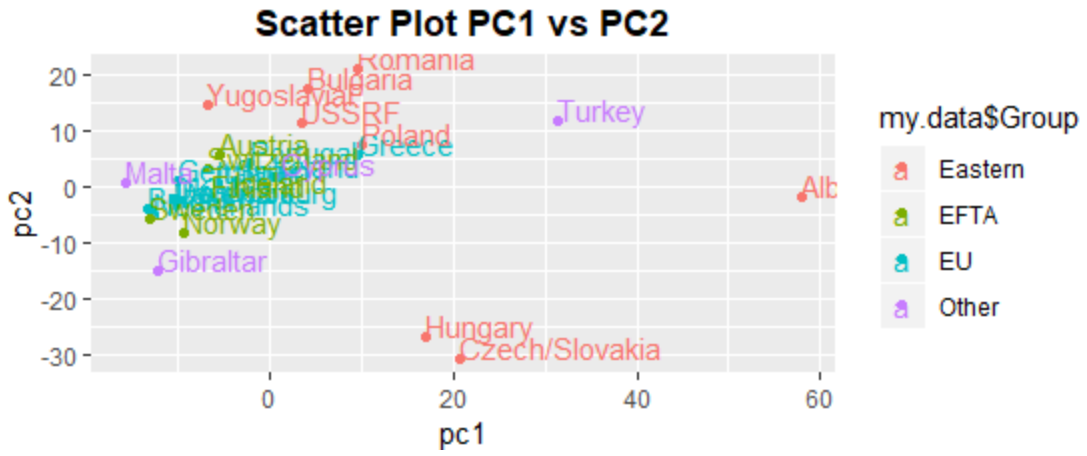
The above pairwise scatterplot illustrates the ratio between service and finance careers. We can see that the Eastern group has a grouping of low numbers of finance careers and a medium number of service careers. EU and EFTA both have large percentages of service careers and a medium percentage of finance careers.



There are two groupings evident in the above graph as well. EU and EFTA both have pretty condense groupings in the low percentages for finance careers and medium percentage for manufacturing careers.

Between the two graphs above the better view for supervised clustering would be Financial careers vs service careers due to the slightly more defied clusters.

Part 4:



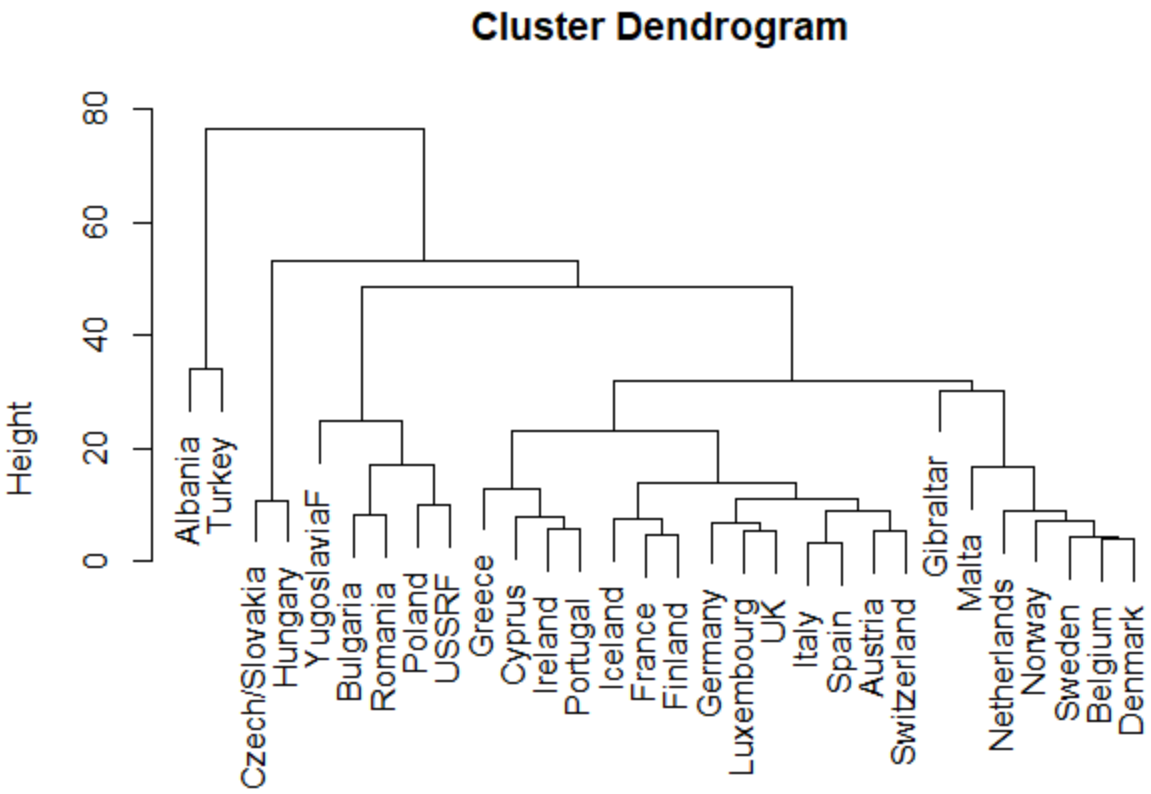
The next step is to utilize principle component analysis to understand what underlying clusters exist in the data. The above graph is an output from creating a new 2D view of the data. In this graph we can see that there are a few clusters of countries that tend to also relate by grouping. For example, Hungry and Czech/Slovakia are together, all of the EFTA and EU countries are grouped together. The only outliers in this graph are the “Other” countries and Albania.

This view of the data presents the most defined clusters compared to the previously created graphs. This graph would appear to have roughly three clusters.

Task 5:

In order to determine the correct number of clusters we will next conduct hierarchical clustering.

The below diagram illustrates the possible created clusters. Moving from left to right we can decide gradually how many clusters we need in the analysis beginning with Albania and Turkey being cluster 1, Czech/Slovakia, Hungry, Yugoslavia, Bulgaria, Romania, Poland and USSR being cluster 2, and so on.



Previously it was estimated that the groupings could be separated into three clusters. If we continue with that approach we achieve the below:

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

We have one large first cluster with all the EU and EFTA countries and a majority of the others, a second cluster with the remaining other and one eastern European country and finally a third cluster that contains the remaining eastern European countries.

The next step is to determine the accuracy of this cluster creation. When computing the accuracy we obtain a value of .5893. We can utilize this number to determine that roughly 59% of the variation is explained by the clusters.

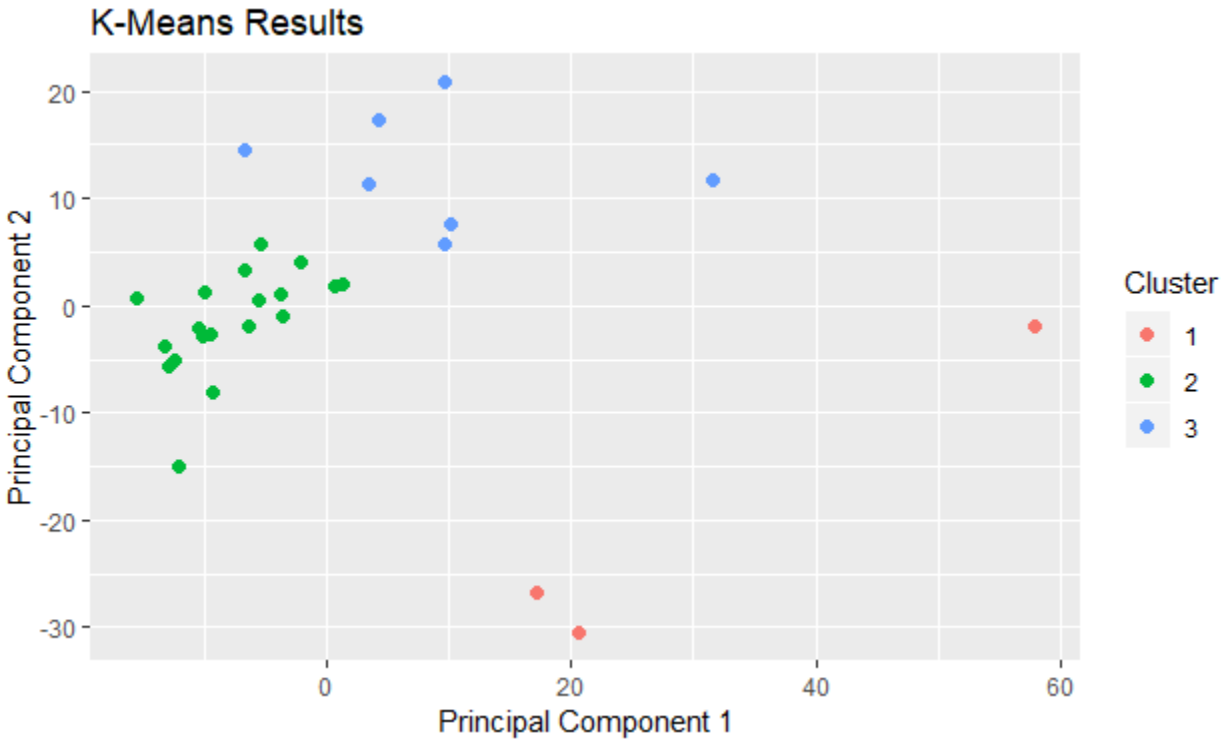
Conversely, if the number of clusters is increased from three to four, there are less countries in the first cluster which are then distributed between the remaining as shown below. Also when the number of

clusters is increased to four the accuracy also increases. The computed value is .7374 or 74% of the variation can be explained utilizing four clusters.

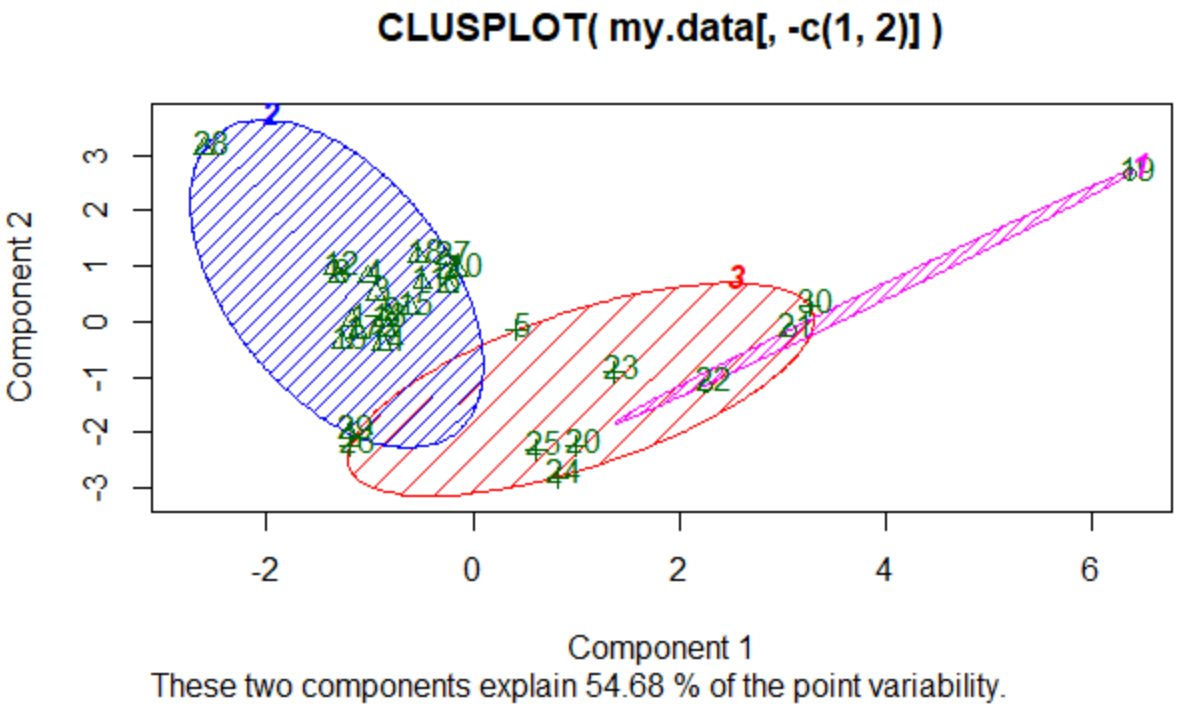
	1	2	3	4
Eastern	0	1	5	2
EFTA	6	0	0	0
EU	12	0	0	0
Other	3	1	0	0

Task 6:

There are other methods of determining the number of clusters in addition to hierarchical. In this section k means will be explored.



To begin the analysis we will be again exploring the analysis utilizing 3 clusters. Viewing the graph above it is easy to see the three separate clusters.

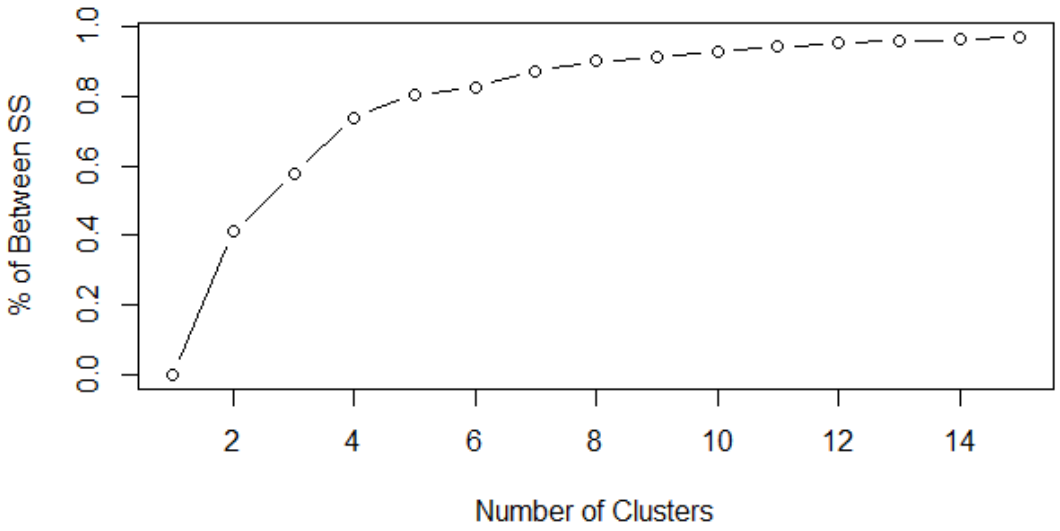


The above graph also illustrates the k means clustering however lists each of the countries individually by their corresponding number. It also shows the area of the graph that each cluster contains.

This analysis produces a slightly less accuracy score of .5793 or 58% of the variation is explained by the three clusters.

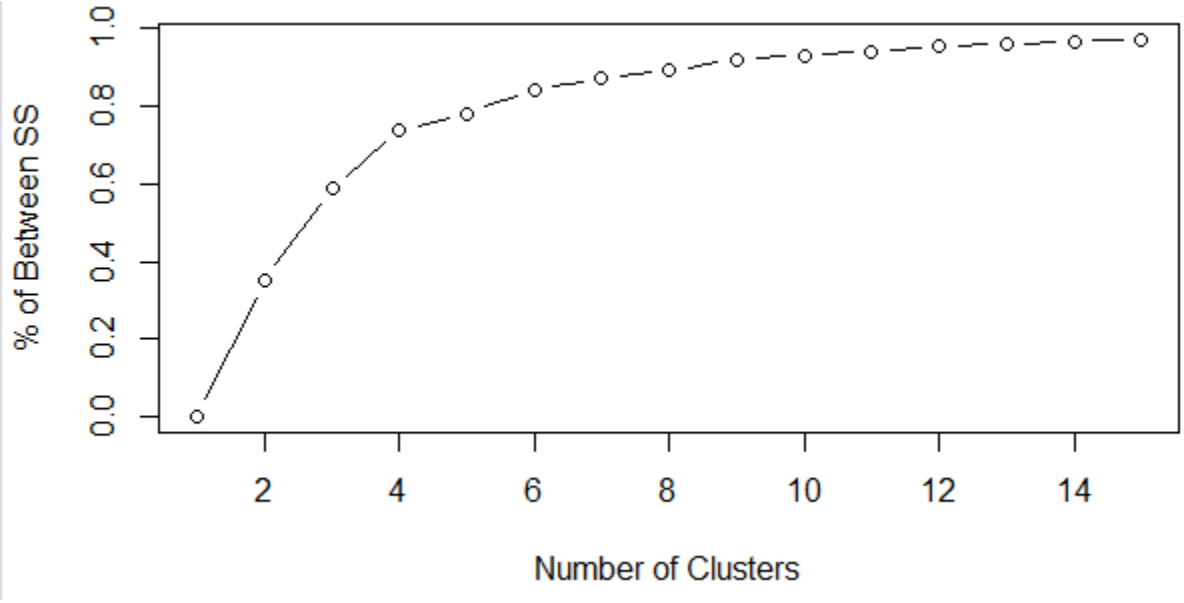
Task 7

Since in cluster analysis you do not typically know the number of clusters ahead of time, it is valuable to see the accuracy of each cluster combination. The below graph illustrates the accuracy for each number of clusters.



Utilizing the above graph we can determine that there is little additional benefit of additional clusters after four clusters as this is where the line begins to become fairly horizontal.

We can also test this by using hierarchical clustering. The below chart is extremely similar to the original k means chart and also indicates that 4 clusters would be the most beneficial.



Conclusions:

Partitioning data into logical clusters allows a simplistic view into large structures of data. These clusters provide a deeper analysis into the existing data. The number of clusters created is highly important to a comprehensive and accurate analysis. There are multiple methods to determining the most accurate number of clusters. Once these clusters are created and verified they can be used to classifying additional information such as diseases, infrastructure breaks and customers.