Assignment #8

Cluster Analysis

Name: Young, Brent

Predict 410 Section #: 57

Quarter: Summer 2017

Introduction

Context

The data for this assignment is the European employment data set. It consists of various industry segments reported as a percent for 30 European nations. The data consists of 4 groups: Eastern, EFTA, EU, and Other. Additionally, there are 9 quantitative variables that represent percent employment in that particular industry.

Objectives/Purpose

The overall purpose/objective of assignment 8 is to do a cluster analysis based upon the 9 quantitative variables that represent percent employment in that particular industry and group these countries into different number of clusters. This will be accomplished by performing an exploratory data analysis, fitting a hierarchical cluster analysis, fitting a k-means cluster analysis, integrating principal components analysis and cluster analysis, using cluster analysis as a predictive model, and generating R graphics applicable to cluster analysis and multivariate analysis. First, we will prep the data by reading in my.data and taking a quick glance at the 30 countries. Second, we will perform an initial EDA, using a pairwise scatterplot to scan the individual 2-dimensional views of the data. We will then discuss whether we see any interesting 2D views of the data. Third, we will visualize the data with labelled scatterplots (aka: specialized plots), using labels and color so that we can compress more than two dimensions of information into a two dimensional plot. This will be accomplished by plotting FIN versus SER and seeing if there are clusters in the plot, the amount of clusters, and the amount of clusters required to create a segmentation. This will be followed up by plotting MAN versus SER and additional discussion surround this topic will follow (e.g., which two 2D views would be better for supervised clustering). Fourth, we will create a 2D projection using PCA by projecting down the data from 9D to 2D using the first and second principal components. This will allow us to create a new 2D view of the data, and a view of the data that contains information from more than two dimensions. We will then discuss how this 2D projection of the data compares to the two other views of the data that we are considering and the amount of clusters that this 2D projection has. Fifth, we will perform hierarchical clustering on the data by cutting the tree to k=3 and k=6 and compare the classification accuracy of two cluster tree cuts. We will then discuss how we are comparing the cluster accuracy in this R code and address which set of clusters is more accurate. Next, we will perform the same analysis in the principal component space using the first and second principal components. We will then address which four cluster models are the most accurate using a table. Sixth, we will perform a k-Means cluster on the same data using for k=3 and k=6. This will allow us to compare the classification accuracy of our different cluster models. Additionally, we will also analyze the k-means plot. Next, we will perform the same analysis in the principal component space using k=3 and k=6. We will then write down the result of increasing k=3 to k=6 and determine out of these eight cluster models which one is the most accurate using a table. Lastly, we will compute the optimal numbers of clusters using the classification accuracy rate of our clusters. We will then determine whether the cluster model with k=14 is the best cluster model.

Section 1: The Data

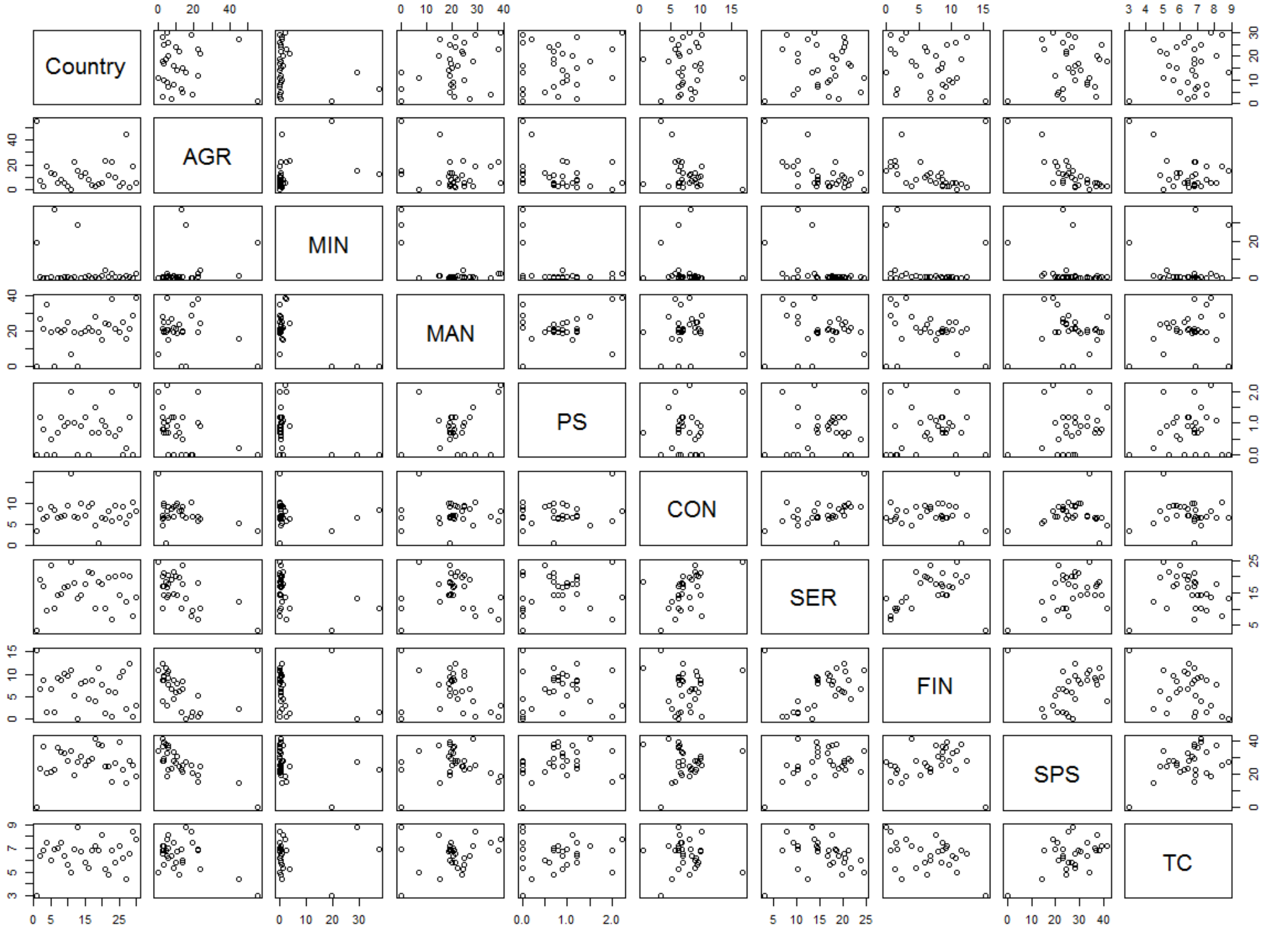
Figure 1: European Employment Data Set

```
     Country   Group  AGR  MIN  MAN   PS  CON  SER  FIN  SPS   TC
1          Belgium      EU  2.6  0.2 20.8 0.8  6.3 16.9  8.7 36.9 6.8
2          Denmark      EU  5.6  0.1 20.4 0.7  6.4 14.5  9.1 36.3 7.0
3           France      EU  5.1  0.3 20.2 0.9  7.1 16.7 10.2 33.1 6.4
4          Germany      EU  3.2  0.7 24.8 1.0  9.4 17.2  9.6 28.4 5.6
5           Greece      EU 22.2  0.5 19.2 1.0  6.8 18.2  5.3 19.8 6.9
6          Ireland      EU 13.8  0.6 19.8 1.2  7.1 17.8  8.4 25.5 5.8
7            Italy      EU  8.4  1.1 21.9 0.0  9.1 21.6  4.6 28.0 5.3
8       Luxembourg      EU  3.3  0.1 19.6 0.7  9.9 21.2  8.7 29.6 6.8
9      Netherlands      EU  4.2  0.1 19.2 0.7  0.6 18.5 11.5 38.3 6.8
10        Portugal      EU 11.5  0.5 23.6 0.7  8.2 19.8  6.3 24.6 4.8
11           Spain      EU  9.9  0.5 21.1 0.6  9.5 20.1  5.9 26.7 5.8
12              UK      EU  2.2  0.7 21.3 1.2  7.0 20.2 12.4 28.4 6.5
13         Austria    EFTA  7.4  0.3 26.9 1.2  8.5 19.1  6.7 23.3 6.4
14         Finland    EFTA  8.5  0.2 19.3 1.2  6.8 14.6  8.6 33.2 7.5
15         Iceland    EFTA 10.5  0.0 18.7 0.9 10.0 14.5  8.0 30.7 6.7
16          Norway    EFTA  5.8  1.1 14.6 1.1  6.5 17.6  7.6 37.5 8.1
17          Sweden    EFTA  3.2  0.3 19.0 0.8  6.4 14.2  9.4 39.5 7.2
18     Switzerland    EFTA  5.6  0.0 24.7 0.0  9.2 20.5 10.7 23.1 6.2
19         Albania Eastern 55.5 19.4  0.0 0.0  3.4  3.3 15.3  0.0 3.0
20        Bulgaria Eastern 19.0  0.0 35.0 0.0  6.7  9.4  1.5 20.9 7.5
21  Czech/Slovakia Eastern 12.8 37.3  0.0 0.0  8.4 10.2  1.6 22.9 6.9
22         Hungary Eastern 15.3 28.9  0.0 0.0  6.4 13.3  0.0 27.3 8.8
23          Poland Eastern 23.6  3.9 24.1 0.9  6.3 10.3  1.3 24.5 5.2
24         Romania Eastern 22.0  2.6 37.9 2.0  5.8  6.9  0.6 15.3 6.8
25           USSRF Eastern 18.5  0.0 28.8 0.0 10.2  7.9  0.6 25.6 8.4
26     YugoslaviaF Eastern  5.0  2.2 38.7 2.2  8.1 13.8  3.1 19.1 7.8
27          Cyprus   Other 13.5  0.3 19.0 0.5  9.1 23.7  6.7 21.2 6.0
28       Gibraltar   Other  0.0  0.0  6.8 2.0 16.9 24.5 10.8 34.0 5.0
29           Malta   Other  2.6  0.6 27.9 1.5  4.6 10.2  3.9 41.6 7.2
30          Turkey   Other 44.8  0.9 15.3 0.2  5.2 12.4  2.4 14.5 4.4
```

Observations (see Appendix for structure, etc.): Figure 1 shows the European employment data set. It consists of various industry segments reported as a percent for 30 European nations. The data consists of 4 groups: Eastern, EFTA, EU, and Other. Additionally, there are 9 quantitative variables that represent percent employment in that particular industry.

Section 2: Initial Exploratory Data Analysis

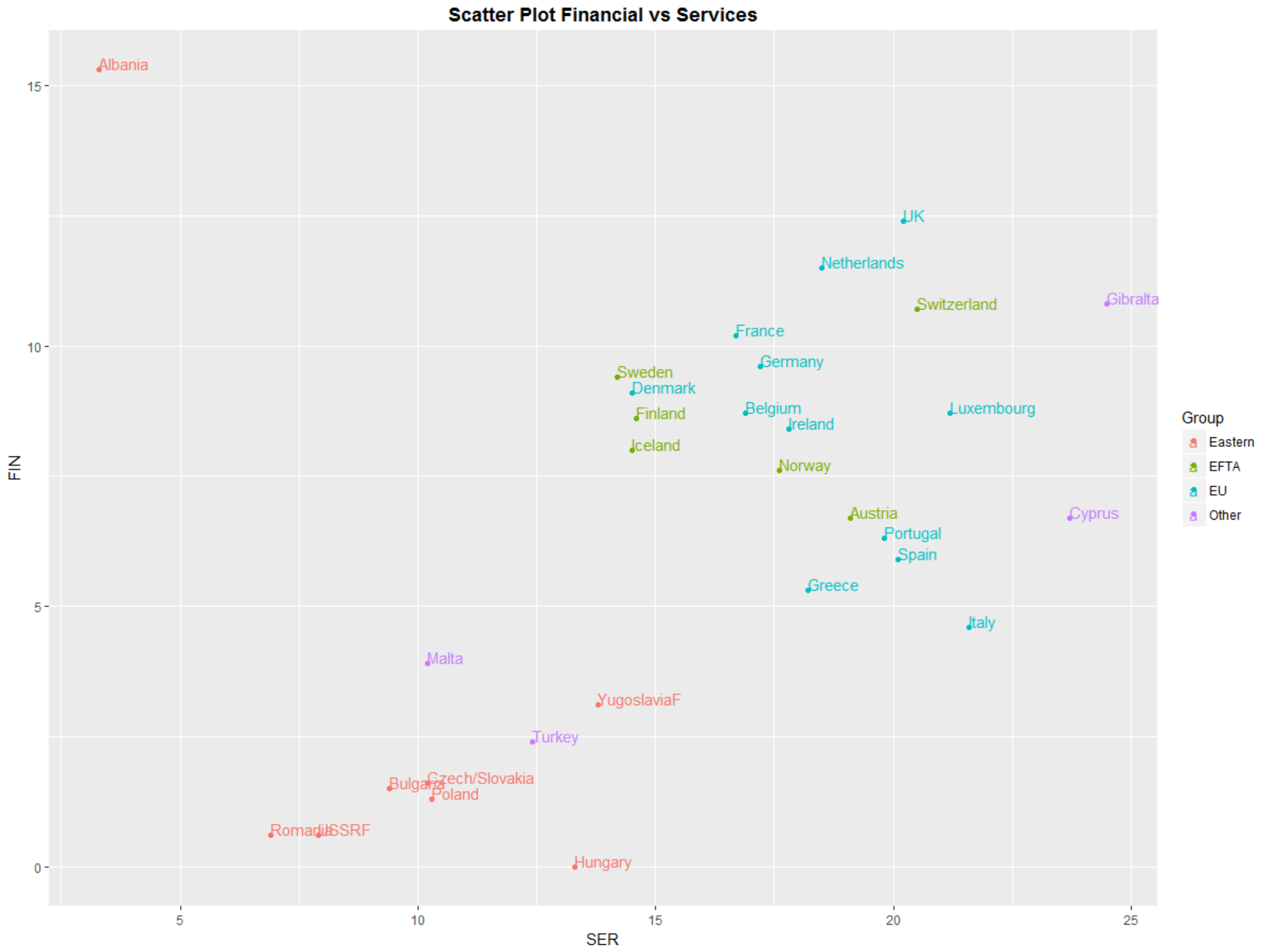
Figure 2: Pairwise Scatterplot



Observations: Figure 2 shows a pairwise scatterplot so that we can scan the individual 2-dimensional views of the data. We consider 'interesting' 2D views as those that are correlated and those that are not correlated. For example, the plot shows that there are some variables that are correlated such as *finance (FIN) vs. services (SER)*, finance (FIN) vs. social and personal services (SPS), etc. However, others are not correlated (e.g., data points are scattered) such as mining (MIN) vs. agriculture (AGR), mining (MIN) versus manufacturing (MAN), and *manufacturing (MAN) vs. services (SER)*.

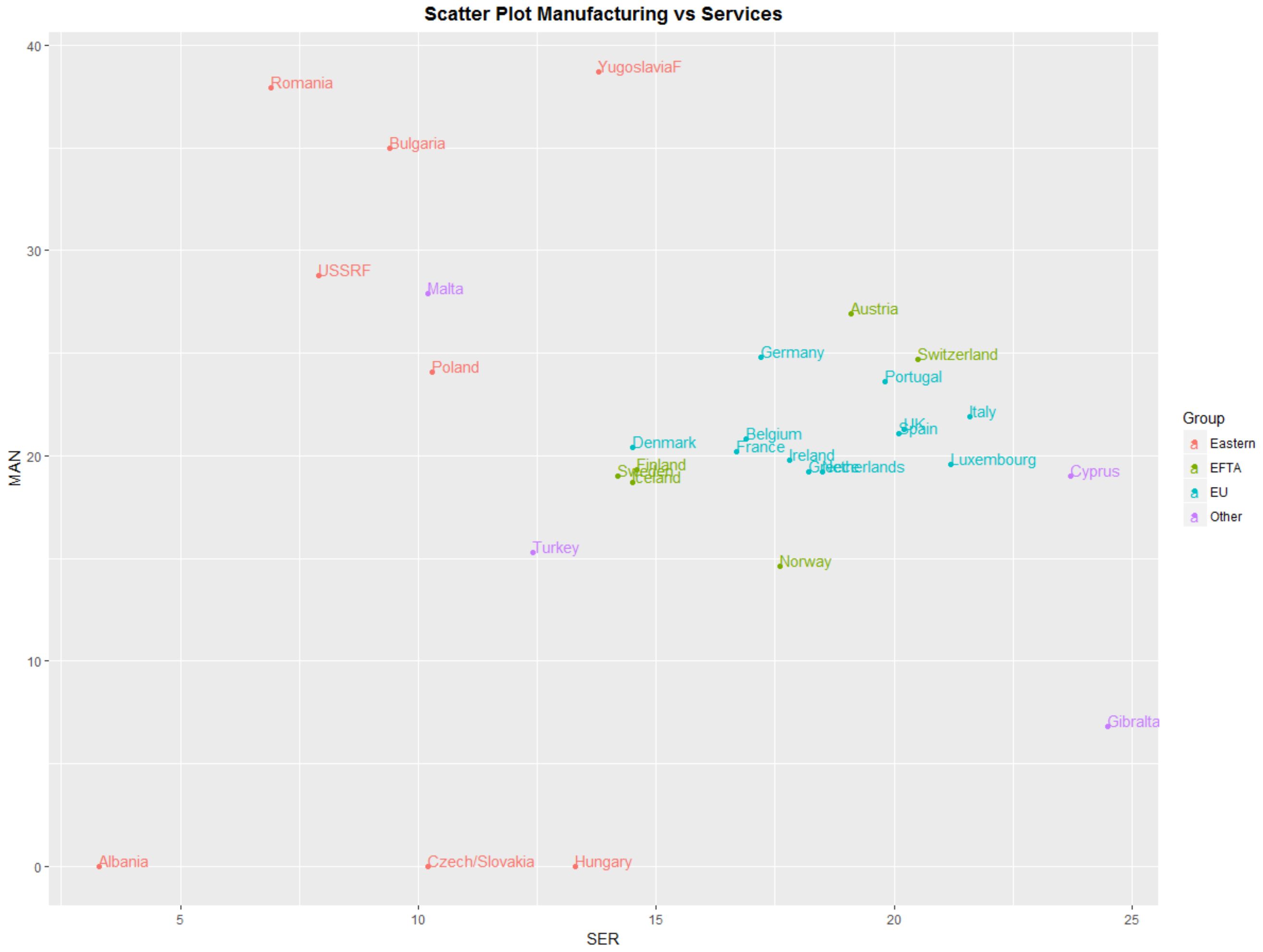
Section 3: Visualizing the Data with Labelled Scatterplots

Figure 3: Scatterplot Financial (FIN) vs. Services (SER)



Observations: Figure 3 shows a scatterplot of FIN vs. SER, which shows data points of all the different countries labeled accordingly and color coordinated to the 4 different groups. The plot shows that there are no clusters in this plot because the colors are not grouped together. For example, we can see red and purple data points below and blue and green points are mixed. However, if we were creating a segmentation we would have 4 clusters since there are 4 groups: Eastern, EFTA, EU, and Other.

Figure 4: Scatterplot Manufacturing (MAN) vs. Services (SER)

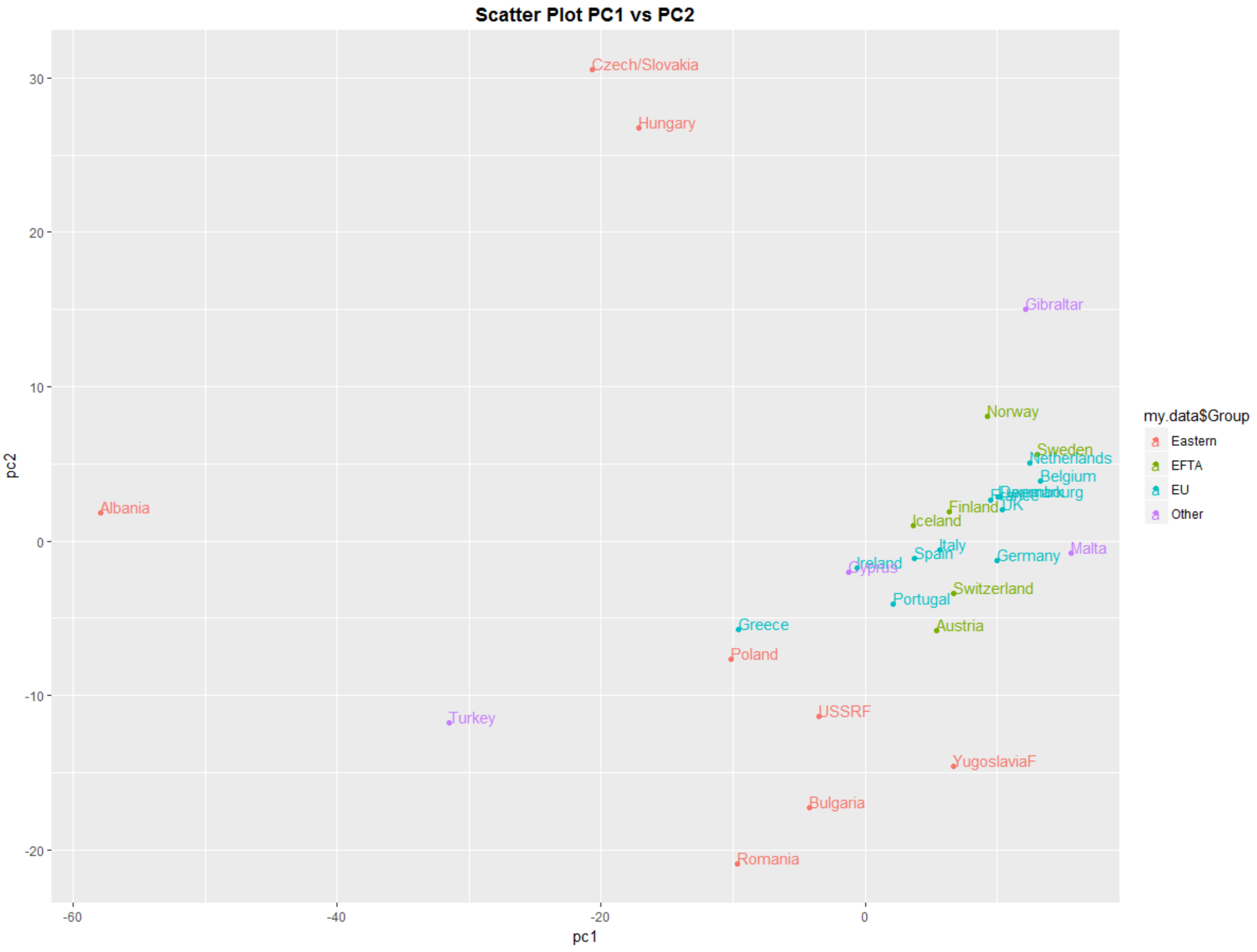


Observations: Figure 4 shows a scatterplot of MAN vs. SER, which shows data points of all the different countries labeled accordingly and color coordinated to the 4 different groups. The plot shows that there are no clusters in this plot because the colors are all over the place and are mixed together. Therefore, these 4 groupings are not clustered based upon employment data. However, if we were creating a segmentation we would have 4 clusters since there are 4 groups: Eastern, EFTA, EU, and Other.

Observations Continued: Of the two 2D views of the data, FIN vs. SER would be the better view for supervised clustering because there is some evidence of some groupings. Additionally, the scatterplot of FIN vs. SER is positively correlated.

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

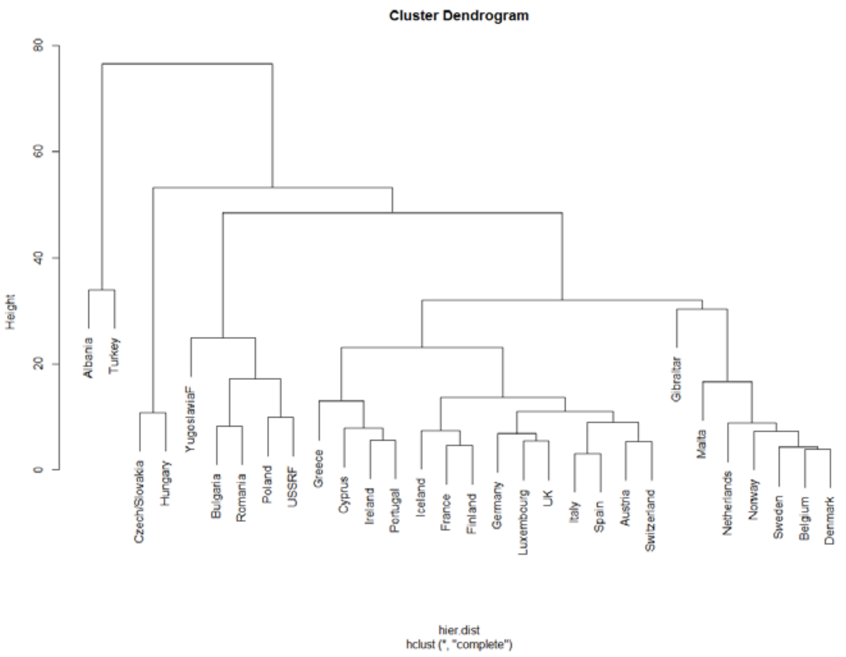
Figure 5: Principal Component 1 vs. Principal Component2



Observations: Figure 5 shows a scatterplot of principle component 1 vs. principle component 2, which shows data points of all the different countries labeled accordingly and color coordinated to the 4 different groups. The results show that the colors are mix matched. However, the plot shows that some countries are close together. For example, countries such as Norway, Austria, and Switzerland are close together. Furthermore, Czech, Slovania, Hungary are also clustered. There are also some countries such as Poland, Bulgaria, and Romania that are close together as well. On the other hand, Albania and Turkey are not clustered. As a result, this 2D projection of the data is different than the other two views because this plot shows that there are some countries that are close together and therefore shows that there are natural clusters in the data. Additionally, this plot uses principle components, while the other two uses variables. Overall, this 2D projection has around 3 natural clusters. In section 5, we will now try to find those clusters.

Section 5: Hierarchical Clustering Analysis

Figure 6: Cluster Dendrogram



Classification Accuracy

Model A: k=3

```
> BetSSPer
[1] 0.5893374
```

Model B: k=3

```
> BetSSPer
[1] 0.8421061
```

Observations: Figure 6 shows a cluster dendrogram along with the between percent sum of squares so that we can compare the classification accuracy for k= 3 and k=6. For k=3, the percent of variation explained by the cluster is 58.9%. On the other hand, for k=6, the percent of variation explained by the cluster is 84.2%. Therefore, since k=6 is closer to 100%, k=6 is the more accurate.

Figure 7: Principal Components, k=3 (Model C)

```
          pc1         pc2 my.data$Country my.data$Group cut.3
1   13.2433600   3.8656680         Belgium            EU     1
2   10.2200362   2.8474626         Denmark            EU     1
3    9.4882760   2.6230785          France            EU     1
4    9.9933150  -1.2373790         Germany            EU     1
5   -9.6250437  -5.7302962          Greece            EU     1
6   -0.5784371  -1.7679693         Ireland            EU     1
7    5.6203757  -0.5533847           Italy            EU     1
8   10.0713613   2.8826637      Luxembourg            EU     1
9   12.4710790   5.0824489     Netherlands            EU     1
10   2.0879800  -4.1015082        Portugal            EU     1
11   3.7027534  -1.1443677           Spain            EU     1
12  10.4214420   2.0305164              UK            EU     1
13   5.4502350  -5.7677262         Austria          EFTA     1
14   6.3403146   1.9227651         Finland          EFTA     1
15   3.6793419   0.9727013         Iceland          EFTA     1
16   9.2986611   8.0722904          Norway          EFTA     1
17  12.9965120   5.6097560          Sweden          EFTA     1
18   6.6650857  -3.3802179     Switzerland          EFTA     1
19 -57.9574320   1.8254605         Albania       Eastern     2
20  -4.1853780 -17.2986029        Bulgaria       Eastern     1
21 -20.6770428  30.5435502  Czech/Slovakia       Eastern     3
22 -17.1253195  26.7320413         Hungary       Eastern     3
23 -10.1389610  -7.6210305          Poland       Eastern     1
24  -9.6705255 -20.9045581         Romania       Eastern     1
25  -3.4626395 -11.3666058           USSRF       Eastern     1
26   6.6803677 -14.6052595     YugoslaviaF       Eastern     1
27  -1.2857888  -2.0018932          Cyprus         Other     1
28  12.1575360  15.0407369       Gibraltar         Other     1
29  15.6147432  -0.7696640           Malta         Other     1
30 -31.4962078 -11.8006764          Turkey         Other     2
```

Figure 8: Clusters vs. Group, k=3

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

Observations: Figure 7 shows the same analysis in the principal component space using the first and second principal components (k=3) along with a crosstab of clusters vs. group (figure 8). The results show that for cluster 1, there are a lot of Eastern, EFTA, and EU. However, in cluster 2, there is only 1 for Eastern and Other. Additionally, for cluster 3, only Eastern shows up. As a result, the fact that there are no matches, shows us that there are two ways to cluster the countries: group classification and cluster.

Figure 9: Principal Components, k=6 (Model D)

```
          pc1         pc2 my.data$Country my.data$Group cut.6
1   13.2433600   3.8656680         Belgium            EU     1
2   10.2200362   2.8474626         Denmark            EU     1
3    9.4882760   2.6230785          France            EU     2
4    9.9933150  -1.2373790         Germany            EU     2
5   -9.6250437  -5.7302962          Greece            EU     2
6   -0.5784371  -1.7679693         Ireland            EU     2
7    5.6203757  -0.5533847           Italy            EU     2
8   10.0713613   2.8826637      Luxembourg            EU     2
9   12.4710790   5.0824489     Netherlands            EU     1
10   2.0879800  -4.1015082        Portugal            EU     2
11   3.7027534  -1.1443677           Spain            EU     2
12  10.4214420   2.0305164              UK            EU     2
13   5.4502350  -5.7677262         Austria          EFTA     2
14   6.3403146   1.9227651         Finland          EFTA     2
15   3.6793419   0.9727013         Iceland          EFTA     2
16   9.2986611   8.0722904          Norway          EFTA     1
17  12.9965120   5.6097560          Sweden          EFTA     1
18   6.6650857  -3.3802179     Switzerland          EFTA     2
19 -57.9574320   1.8254605         Albania       Eastern     3
20  -4.1853780 -17.2986029        Bulgaria       Eastern     4
21 -20.6770428  30.5435502  Czech/Slovakia       Eastern     5
22 -17.1253195  26.7320413         Hungary       Eastern     5
23 -10.1389610  -7.6210305          Poland       Eastern     4
24  -9.6705255 -20.9045581         Romania       Eastern     4
25  -3.4626395 -11.3666058           USSRF       Eastern     4
26   6.6803677 -14.6052595     YugoslaviaF       Eastern     4
27  -1.2857888  -2.0018932          Cyprus         Other     2
28  12.1575360  15.0407369       Gibraltar         Other     1
29  15.6147432  -0.7696640           Malta         Other     1
30 -31.4962078 -11.8006764          Turkey         Other     6
```

Figure 10: Clusters vs. Group, k=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
```

Observations: Figure 9 shows the same analysis in the principal component space using the first and second principal components (k=6) along with a crosstab of clusters vs. group (figure 10). The results show that for cluster 1, there are a lot of EFTA, EU, and Other. Cluster 2 shows that there is also a lot of EFTA, EU, and Other. Cluster 3, shows Eastern only. This is the same for cluster 4 and 5. Lastly, cluster 6 shows Other only. As a result, the fact that there are no matches, shows us that there are two ways to cluster the countries: group classification and cluster.

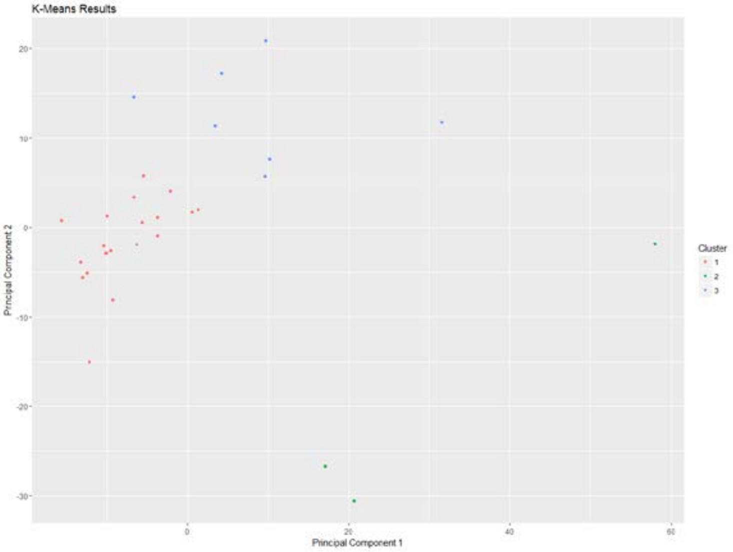
Figure 11: Table – Accuracy of Four 'Cluster Models'

Cluster Model Name	Between Percent Sum of Squares	Crosstab
Model A: k=3	58.9%	
Model B: k=6	84.2%	
Model C: PC, k=3		1 2 3 Eastern 5 1 2 EFTA 6 0 0 EU 12 0 0 Other 3 1 0
Model D, PC, k=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

Observations: Figure 11 shows a table of the four cluster models in regards to accuracy (e.g., between percent sum of squares). The results show that model B is the most accurate since it has the highest between percent sum of squares.

Section 6: k-Means Clustering Analysis

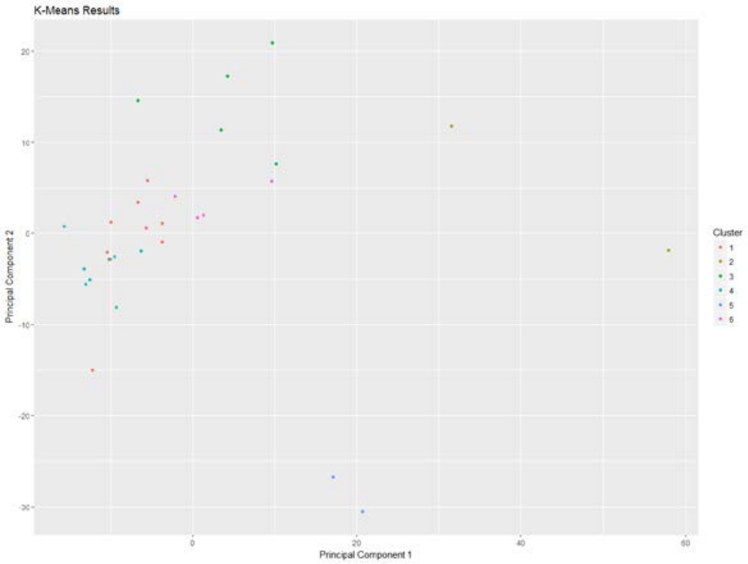
Figure 12: k-means clustering with k=3 clusters (Model E)



```
> BetSSPer
[1] 0.5792964
```

Observations: Figure 12 shows a k-means clustering with k=3 clusters. The results show that the percent of variation explained by the cluster is 57.9%. The plot shows that all the red and blue data points are close together, creating 2 clusters. There is also two green data points on the bottom and 1 green data points on the right side to create a third cluster. Therefore, using the 9 dimensional data, we have generated green, blue, and red natural clusters.

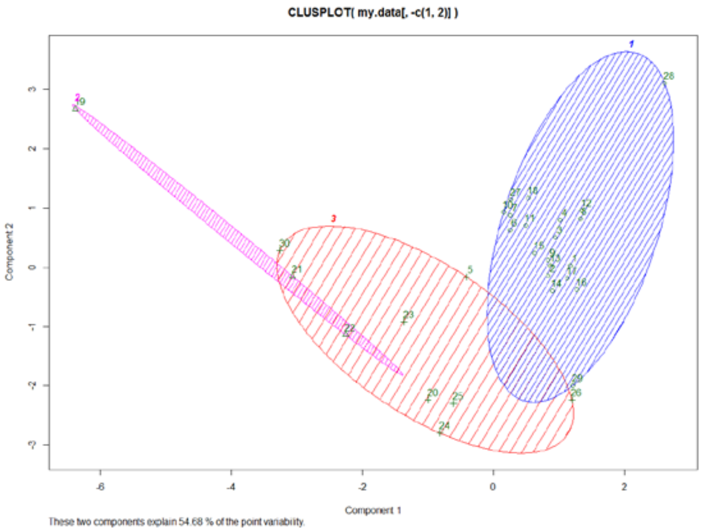
Figure 13: k-means clustering with k=6 clusters (Model F)



```
> BetSSPer
[1] 0.8274197
```

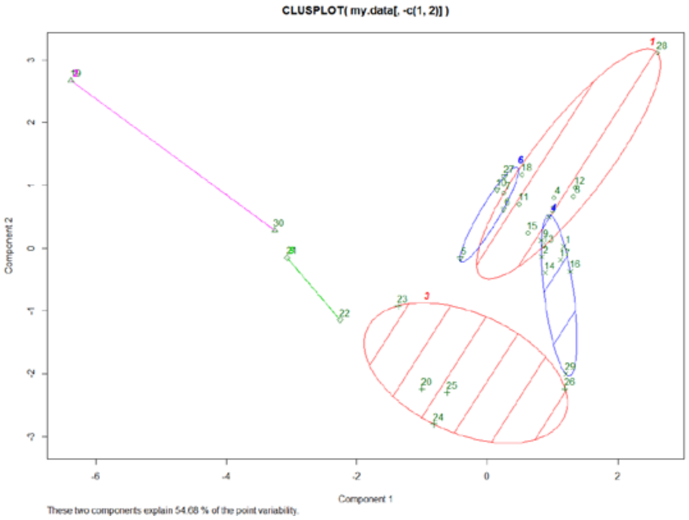
Observations: Figure 13 shows a k-means clustering with k=6 clusters. The results show that the percent of variation explained by the cluster is 82.7%. The plot shows that all the green, magenta, red, and light blue clusters are close together. Additionally, there are two dark data points on the bottom which form another cluster and yellow data points on the right side, which form the sixth cluster. Therefore, using the 9 dimensional data, we have generated 6 natural clusters.

Figure 14: k-means Principal Components, k=3 (Model G)



Observations: Figure 14 shows a k-means clustering in the principal components space using k=3. The results show that the percent of variation explained by the two components is 54.68%. This view shows the countries placed into 3 different clusters with their assigned country number.

Figure 15: k-means Principal Components, k=6 (Model H)



Observations: Figure 15 shows a k-means clustering in the principal components space using k=6. The results show that the percent of variation explained by the two components is 54.68%. This view shows the countries placed into 4 primary clusters with their assigned country number. There are also two other clusters, but they are somewhat outliers. Therefore, as we increase the number of clusters from k=3 to k =6, the clusters tend to be smaller in size.

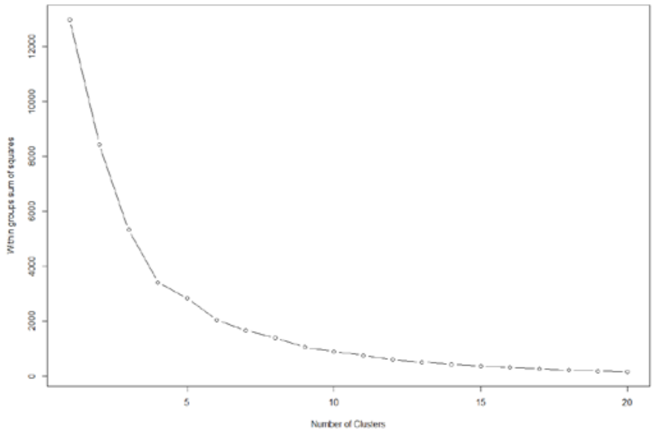
Figure 16: Table – Accuracy of 8 'Cluster Models'

Cluster Model Name	Between Percent Sum of Squares	Crosstab
Hierarchical Cluster Model A: k=3	58.9%	
Hierarchical Cluster Model B: k=6	84.2%	
Hierarchical Cluster Model C: PC, k=3		1 2 3 Eastern 5 1 2 EFTA 6 0 0 EU 12 0 0 Other 3 1 0
Hierarchical Cluster Model D, PC, k=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
k-means Model E: k=3	57.9%	
k-means Model F: k=6	82.7%	
k-means Model G: PC, k=3	54.7%.	
k-means Model H: PC, k=6	54.7%.	

Observations: Figure 16 shows a table of the eight cluster models in regards to accuracy (e.g., between percent sum of squares). The results show that model B is the most accurate since it has the highest between percent sum of squares.

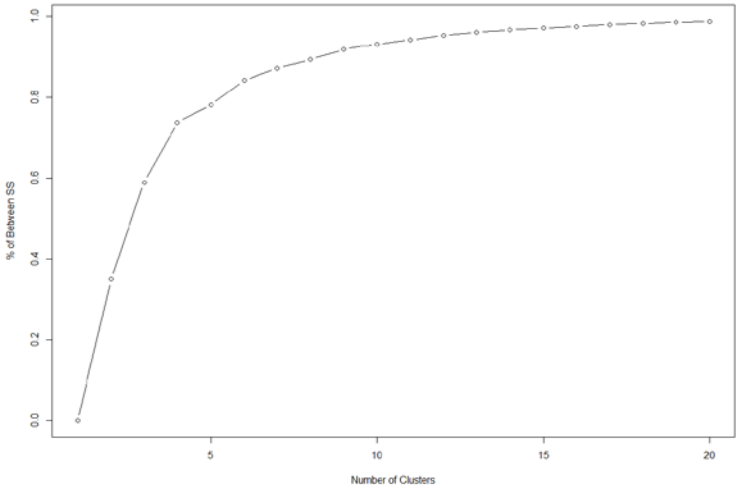
Section7: Optimal Number of Clusters

Figure 17: Hierarchical clustering: 'Within group sums of squares'



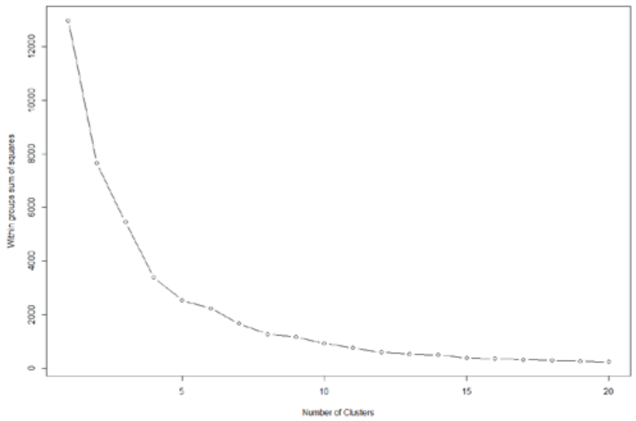
Observations: Figure 17 shows a plot of hierarchical clustering: 'within group sums of squares'. The plot starts off with a high within group sums of squares value, but as the number of clusters increase, the within group sums of squares go down. Based on the plot, 4 or 5 seems to be the optimal number of clusters, after that the benefit goes down very slowly.

Figure 18: Hierarchical clustering: 'Percent of Between SS'



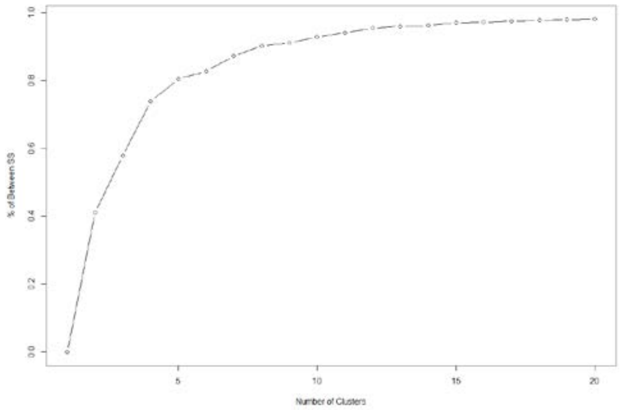
Observations: Figure 18 shows a plot of k-means clustering: 'Percent of Between SS'. When the plot shows 5 clusters, the percent of between SS is 80%. After 5 clusters, it begins to taper off. Therefore, 4 or 5 clusters would be the optimal number of clusters.

Figure 19: k-means clustering: 'Within group sums of squares'



Observations: Figure 19 shows a plot of k-means clustering: 'within group sums of squares'. The plot starts with a high within group sums of squares value, but as the number of clusters increase, the within group sums of squares go down. Based on the plot, 4 or 5 seems to be the optimal number of clusters, after that the benefit goes down very slowly. This is similar to what we saw in figure 17.

Figure 20: k-means clustering: 'Percent of Between SS'



Observations: Figure 20 shows a plot of k-means clustering: 'Percent of Between SS'. When the plot shows 5 clusters, the percent of between SS is 80%. After 5 clusters, it begins to taper off. Therefore, 4 or 5 clusters would be the optimal number of clusters. This is similar to what we saw in figure 18.

In the end, the best cluster model would be a hierarchical cluster model with k=4 or 5, not k=14. Evidence of this is shown in figures 17 to 20.

Conclusion/Summary

In section 1, we prepped the data by reading in my.data and taking a quick glance at the 30 countries. In section 2, we performed an initial EDA, using a pairwise scatterplot to scan the individual 2-dimensional views of the data. The plot showed that there were some variables that were correlated such as *finance (FIN) vs. services (SER)*, finance (FIN) vs. social and personal services (SPS), etc. However, others were not correlated (e.g., data points are scattered) such as mining (MIN) vs. agriculture (AGR), mining (MIN) versus manufacturing (MAN), and *manufacturing (MAN) vs. services (SER)*.

In section 3, we visualized the data with labelled scatterplots (aka: specialized plots), using labels and color so that we could compress more than two dimensions of information into a two dimensional plot. The plot of FIN vs. SER showed that there were no clusters in this plot because the colors were not grouped together. For example, we saw red and purple data points on the bottom and blue and green points that were mixed. However, if we were to create a segmentation we would have 4 clusters since there are 4 groups: Eastern, EFTA, EU, and Other. The plot of MAN versus SER also showed that there are no clusters in this plot because the colors were all over the place and were mixed together. Therefore, these 4 groupings were not clustered based upon employment data. However, if we were to create a segmentation we would have 4 clusters since there w 4 groups: Eastern, EFTA, EU, and Other. However, of the two 2D views of the data, FIN vs. SER would be the better view for supervised clustering because there was some evidence of some groupings. Additionally, the scatterplot of FIN vs. SER was also positively correlated.

In section 4, we created a 2D projection using PCA by projecting down the data from 9D to 2D using the first and second principal components. This allowed us to create a new 2D view of the data, and a view of the data that contains information from more than two dimensions. The results showed that the colors were mix matched. However, the plot showed that some countries were close together. For example, countries such as Norway, Austria, and Switzerland were close together. Furthermore, Czech, Slovania, Hungary were also clustered. There were also some countries such as Poland, Bulgaria, and Romania that were close together as well. On the other hand, Albania and Turkey were not clustered. As a result, this 2D projection of the data was different than the other two views because this plot showed that there were some countries that were close together and therefore showed us that there were natural clusters in the data. Additionally, this plot used principle components, while the other two used variables. Overall, this 2D projection had around 3 natural clusters.

In section 5, we performed hierarchical clustering on the data by cutting the tree to k=3 and k=6 and comparing the classification accuracy of two cluster tree cuts. The results showed that for k=3, the percent of variation explained by the cluster was 58.9%. On the other hand, for k=6, the percent of variation explained by the cluster was 84.2%. Therefore, since k=6 was closer to 100%, k=6 was more accurate. We then performed the same analysis in the principal component space using the first and second principal components. For k=3, the results showed that for cluster 1, there were a lot of Eastern, EFTA, and EU. However, in cluster 2, there was only 1 for Eastern and Other. Additionally, for cluster 3, only Eastern showed up. As a result, the fact that there were no matches, showed us that there were two ways to cluster the countries: group classification and cluster. Furthermore, for k=6, the results showed us that for cluster 1, there are a lot of EFTA, EU, and Other. Cluster 2 showed us that there was also a lot of EFTA, EU, and Other. On the other hand, Cluster 3 showed Eastern only, which was the same for cluster 4 and 5. Lastly, cluster 6 showed Other only. In the end, the results showed that model B (hierarchical clustering, k=6) was the most accurate since it had the highest between percent sum of squares (84.2%).

In section 6, we performed a k-Means cluster on the same data using k=3 and k=6. This allowed us to compare the classification accuracy of our different cluster models. The results showed that for Model E (k=3), the percent of variation explained by the cluster was 57.9%. The plot also showed that green, blue, and red natural clusters were created. Furthermore, for Model F (k=3), the percent of variation explained by the cluster was 82.7%. The plot showed that all the green, magenta, red, and light blue clusters were close together. Additionally, there were two dark data points on the bottom which formed another cluster and yellow data points on the right side, which formed the sixth cluster. We also performed the same analysis in the principal component space using k=3 and k=6. For Model G (k=3), the results showed that the percent of variation explained by the two components was 54.68%. This plot showed the countries placed into 3 different clusters with their assigned country number. Furthermore, for Model H (k=6) the percent of variation explained by the two components was 54.68%. The plot showed the countries placed into 4 primary clusters with their assigned country number. There are also two other clusters, but they are somewhat outliers. Therefore, as we increase the number of clusters from k=3 to k =6, the clusters tended to be smaller in size. In the end, the results showed that model B was still the most accurate since it had the highest between percent sum of squares. Lastly, we computed the optimal numbers of clusters using the classification accuracy rate of our clusters (e.g., ‘Within group sums of squares’ and percent of between sums of squares’ for hierarchical and k-means clustering). The results showed that the best cluster model would be a hierarchical cluster model with k=4 or 5, not k=14. Evidence of this was shown in figures 17 to 20.

Appendix

Section 1

```
> str(my.data)
'data.frame':   30 obs. of  11 variables:
 $ Country: Factor w/ 30 levels "Albania","Austria",..: 3 7 9 10 12 15 16 17 19 22
...
 $ Group  : Factor w/ 4 levels "Eastern","EFTA",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ AGR    : num  2.6 5.6 5.1 3.2 22.2 13.8 8.4 3.3 4.2 11.5 ...
 $ MIN    : num  0.2 0.1 0.3 0.7 0.5 0.6 1.1 0.1 0.1 0.5 ...
 $ MAN    : num  20.8 20.4 20.2 24.8 19.2 19.8 21.9 19.6 19.2 23.6 ...
 $ PS     : num  0.8 0.7 0.9 1 1 1.2 0 0.7 0.7 0.7 ...
 $ CON    : num  6.3 6.4 7.1 9.4 6.8 7.1 9.1 9.9 0.6 8.2 ...
 $ SER    : num  16.9 14.5 16.7 17.2 18.2 17.8 21.6 21.2 18.5 19.8 ...
 $ FIN    : num  8.7 9.1 10.2 9.6 5.3 8.4 4.6 8.7 11.5 6.3 ...
 $ SPS    : num  36.9 36.3 33.1 28.4 19.8 25.5 28 29.6 38.3 24.6 ...
 $ TC     : num  6.8 7 6.4 5.6 6.9 5.8 5.3 6.8 6.8 4.8 ...

> head(my.data)
  Country Group  AGR MIN  MAN  PS CON  SER  FIN  SPS  TC
1 Belgium    EU  2.6 0.2 20.8 0.8 6.3 16.9  8.7 36.9 6.8
2 Denmark    EU  5.6 0.1 20.4 0.7 6.4 14.5  9.1 36.3 7.0
3  France    EU  5.1 0.3 20.2 0.9 7.1 16.7 10.2 33.1 6.4
4 Germany    EU  3.2 0.7 24.8 1.0 9.4 17.2  9.6 28.4 5.6
5  Greece    EU 22.2 0.5 19.2 1.0 6.8 18.2  5.3 19.8 6.9
6 Ireland    EU 13.8 0.6 19.8 1.2 7.1 17.8  8.4 25.5 5.8
```