Assignment 8

Jennifer M. Wanat

Introduction

Section 1: Data Check

```
## '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 ...
```

```
##   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
```

Table 1: European Employment Data Set

Country	Group	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
Belgium	EU	2.6	0.2	20.8	0.8	6.3	16.9	8.7	36.9	6.8
Denmark	EU	5.6	0.1	20.4	0.7	6.4	14.5	9.1	36.3	7.0
France	EU	5.1	0.3	20.2	0.9	7.1	16.7	10.2	33.1	6.4
Germany	EU	3.2	0.7	24.8	1.0	9.4	17.2	9.6	28.4	5.6
Greece	EU	22.2	0.5	19.2	1.0	6.8	18.2	5.3	19.8	6.9
Ireland	EU	13.8	0.6	19.8	1.2	7.1	17.8	8.4	25.5	5.8
Italy	EU	8.4	1.1	21.9	0.0	9.1	21.6	4.6	28.0	5.3
Luxembourg	EU	3.3	0.1	19.6	0.7	9.9	21.2	8.7	29.6	6.8
Netherlands	EU	4.2	0.1	19.2	0.7	0.6	18.5	11.5	38.3	6.8
Portugal	EU	11.5	0.5	23.6	0.7	8.2	19.8	6.3	24.6	4.8
Spain	EU	9.9	0.5	21.1	0.6	9.5	20.1	5.9	26.7	5.8
UK	EU	2.2	0.7	21.3	1.2	7.0	20.2	12.4	28.4	6.5
Austria	EFTA	7.4	0.3	26.9	1.2	8.5	19.1	6.7	23.3	6.4
Finland	EFTA	8.5	0.2	19.3	1.2	6.8	14.6	8.6	33.2	7.5
Iceland	EFTA	10.5	0.0	18.7	0.9	10.0	14.5	8.0	30.7	6.7
Norway	EFTA	5.8	1.1	14.6	1.1	6.5	17.6	7.6	37.5	8.1
Sweden	EFTA	3.2	0.3	19.0	0.8	6.4	14.2	9.4	39.5	7.2
Switzerland	EFTA	5.6	0.0	24.7	0.0	9.2	20.5	10.7	23.1	6.2
Albania	Eastern	55.5	19.4	0.0	0.0	3.4	3.3	15.3	0.0	3.0
Bulgaria	Eastern	19.0	0.0	35.0	0.0	6.7	9.4	1.5	20.9	7.5

Country	Group	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
Czech/Slovakia	Eastern	12.8	37.3	0.0	0.0	8.4	10.2	1.6	22.9	6.9
Hungary	Eastern	15.3	28.9	0.0	0.0	6.4	13.3	0.0	27.3	8.8
Poland	Eastern	23.6	3.9	24.1	0.9	6.3	10.3	1.3	24.5	5.2
Romania	Eastern	22.0	2.6	37.9	2.0	5.8	6.9	0.6	15.3	6.8
USSRF	Eastern	18.5	0.0	28.8	0.0	10.2	7.9	0.6	25.6	8.4
YugoslaviaF	Eastern	5.0	2.2	38.7	2.2	8.1	13.8	3.1	19.1	7.8
Cyprus	Other	13.5	0.3	19.0	0.5	9.1	23.7	6.7	21.2	6.0
Gibraltar	Other	0.0	0.0	6.8	2.0	16.9	24.5	10.8	34.0	5.0
Malta	Other	2.6	0.6	27.9	1.5	4.6	10.2	3.9	41.6	7.2
Turkey	Other	44.8	0.9	15.3	0.2	5.2	12.4	2.4	14.5	4.4

```
##  [1] 100.0 100.1 100.0  99.9  99.9 100.0 100.0  99.9  99.9 100.0 100.1
## [12]  99.9  99.8  99.9 100.0  99.9 100.0 100.0  99.9 100.0 100.1 100.0
## [23] 100.1  99.9 100.0 100.0 100.0 100.0 100.1 100.1
```

Table 2: Industries

Abbreviation	Description
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

Section 2: Initial Exploratory Data Analysis

Section 3: Visualizing the Data with Labelled Scatterplots

Section 4: Creating a 2D Projection Using Principal Componenets 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."

"Note that in this application we will not standardize this data to be mean zero with unit variance before we perform the PCA. 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. We are not standardizing the data since large components in some dimensions will require small components in other dimensions in order to sum to 100. This natural constraint creates the natural separations that we have seen in Part 3."

"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. In practice we would run this analysis both ways. We would run the PCA as is, and we would run the PCA on the standardized data,

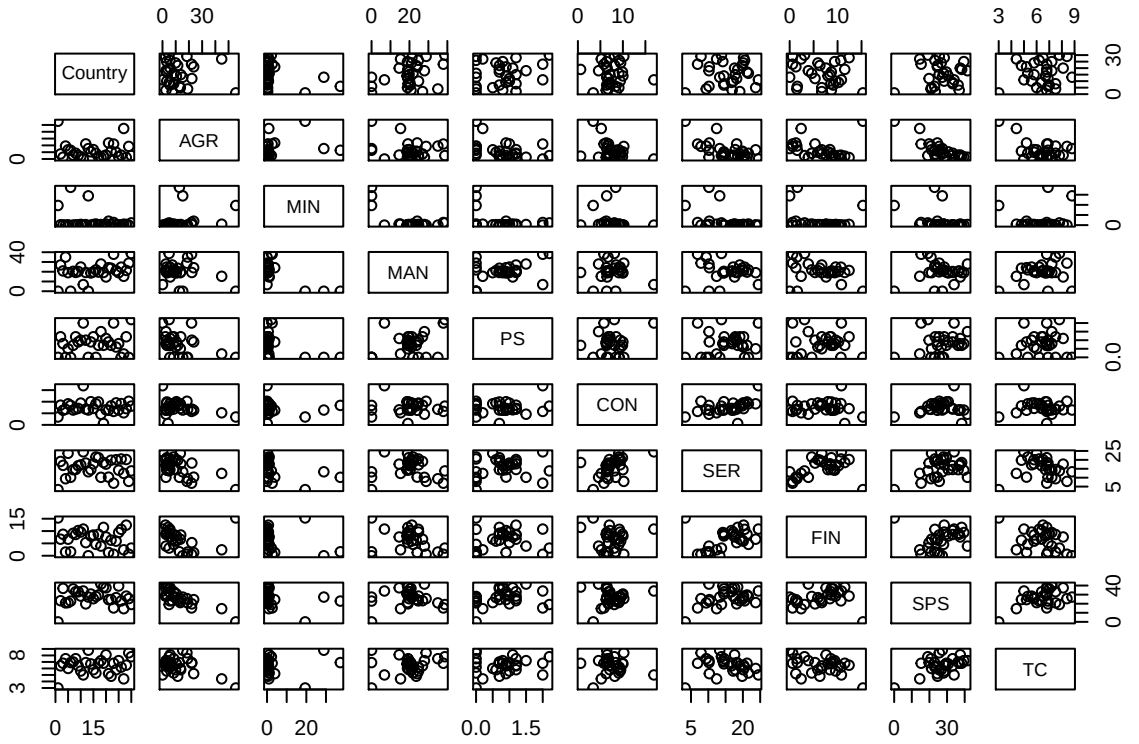


Figure 1: Pairwise Scatterplot

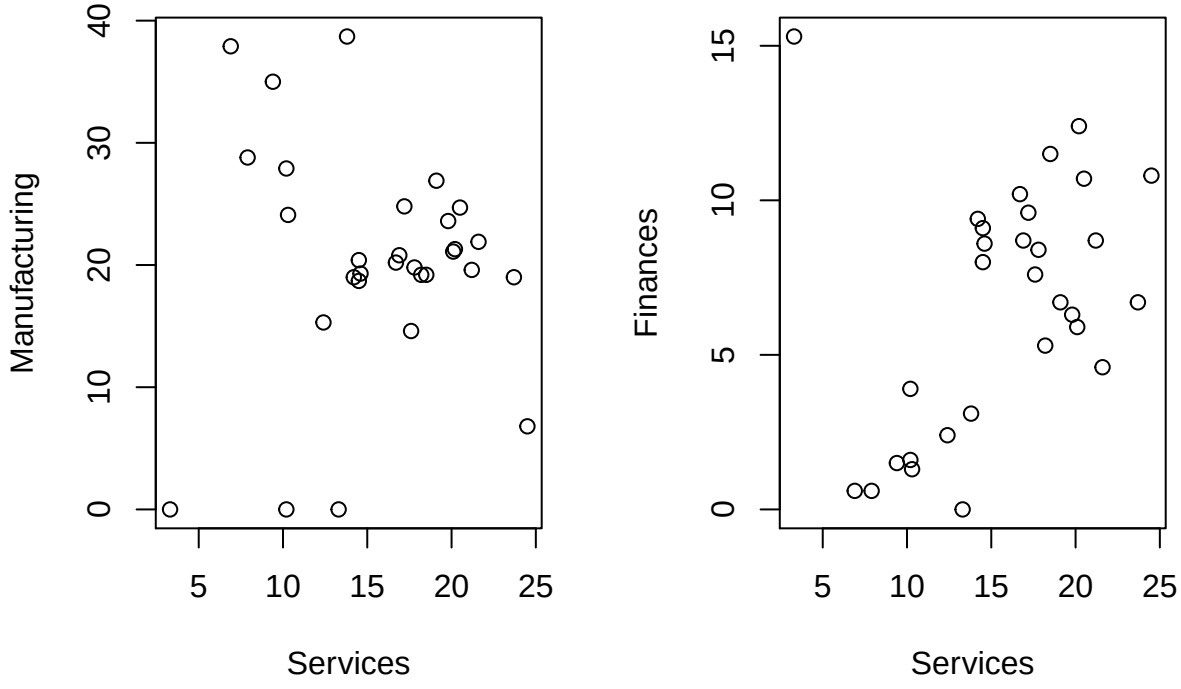


Figure 2: Scatterplots

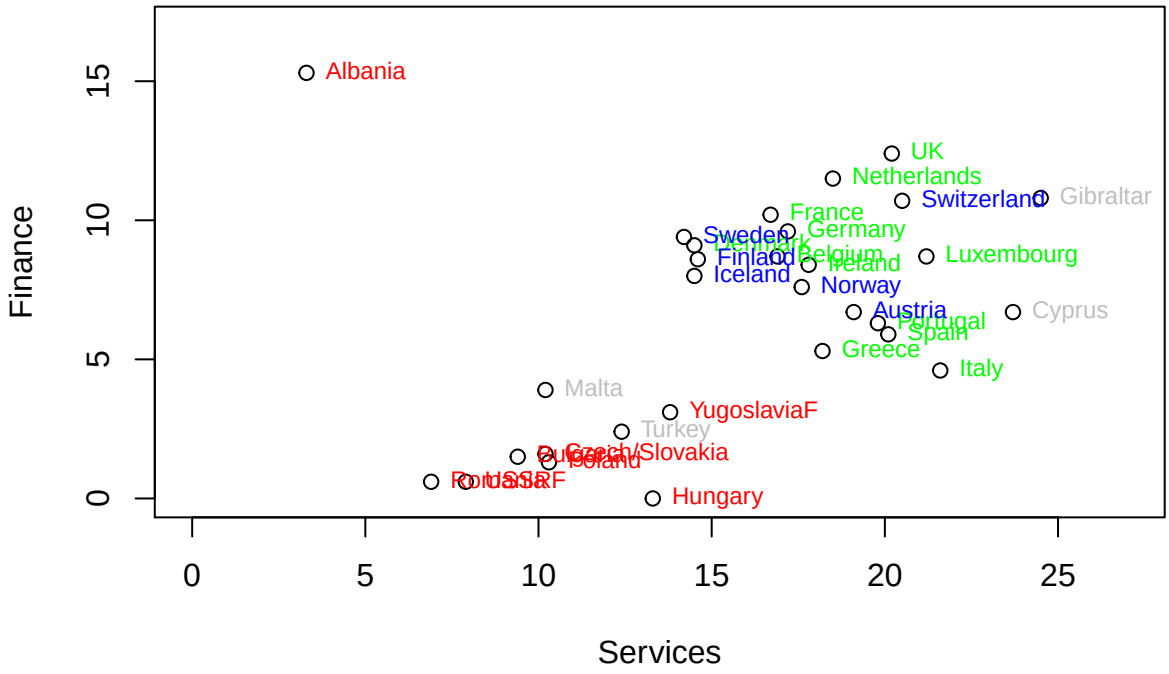


Figure 3: Labelled Scatterplot of Finance vs Services

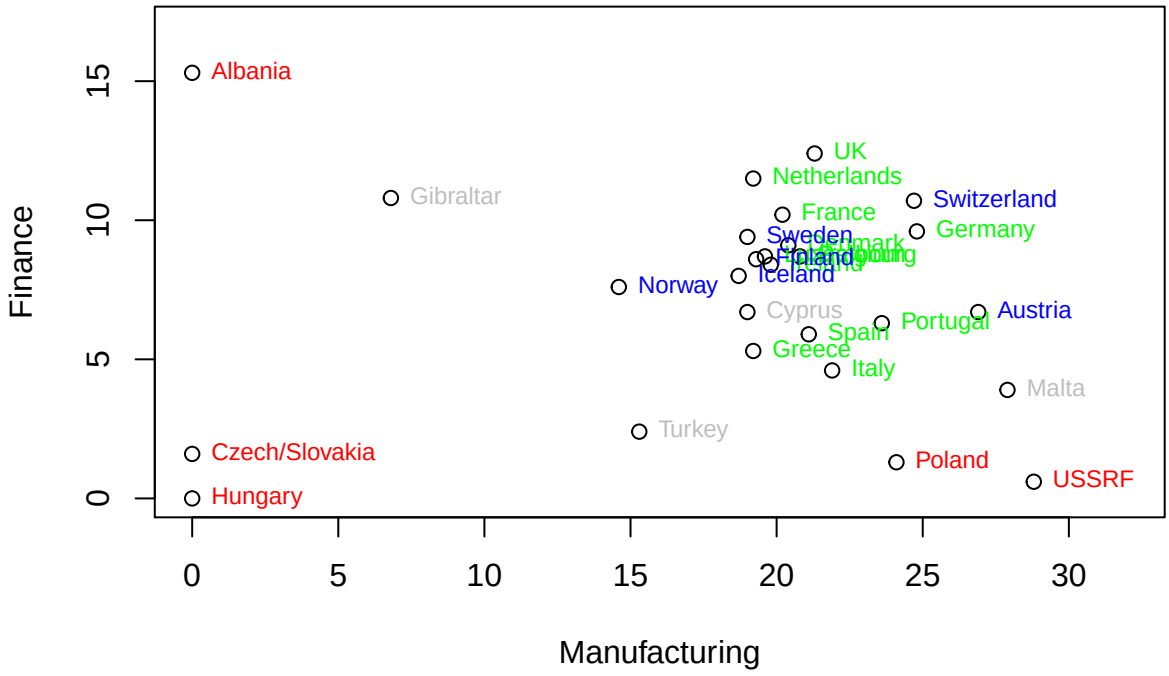


Figure 4: Labelled Scatterplot of Finance vs Manufacturing

and then we would compare the results. In particular we would compare the results in our final application, which in this assignment would be the cluster analysis."

```
##  [1] 100.0 100.1 100.0  99.9  99.9 100.0 100.0  99.9  99.9 100.0 100.1
## [12]  99.9  99.8  99.9 100.0  99.9 100.0 100.0  99.9 100.0 100.1 100.0
## [23] 100.1  99.9 100.0 100.0 100.0 100.0 100.1 100.1
```

```
## [1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"
## [7] "call"
```

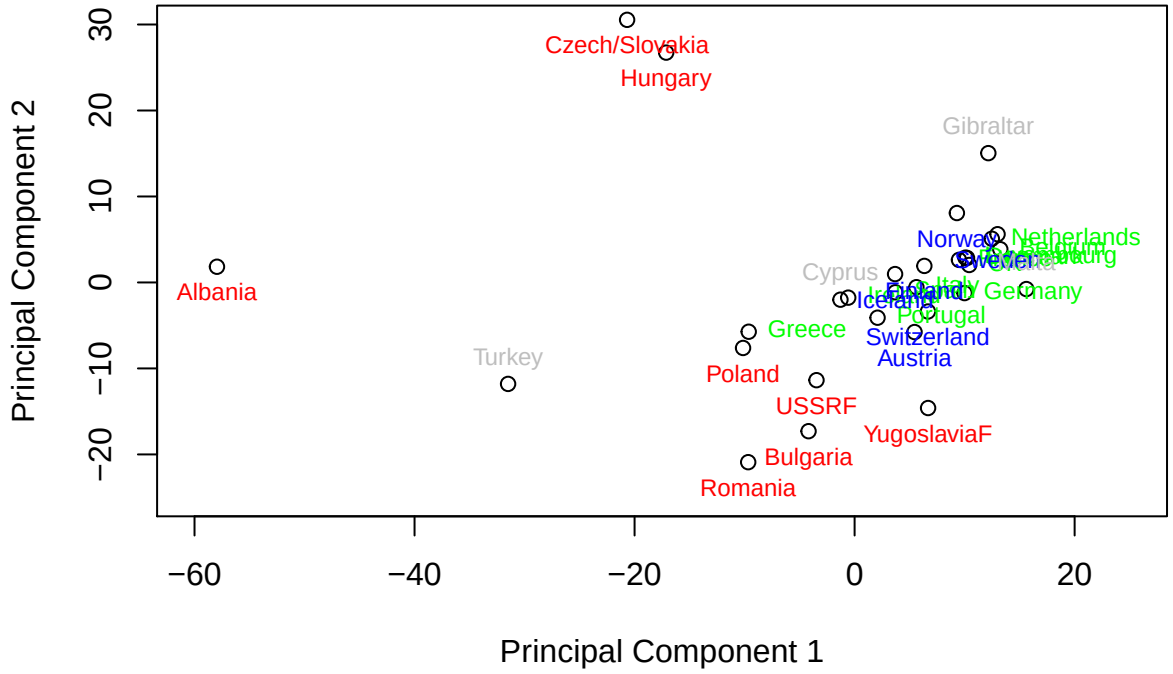


Figure 5: Labeled Scatterplot of PC1 vs PC2

Section 5: Hierarchical Clustering Analysis

"The first cluster analysis that we perform on this data will use hierarchical clustering. In the previous exploratory data analysis of the data we kept the 'Other' category on the data. Since the derived clusters are affect by all included data points, especially outlier data points, we will remove the 'Other' observations from the data so that the clustering algorithms will only use the proper data when creating the clusters."

"We will begin by clustering in the FIN*SER 2D view of the data. Hierarchical clustering is performed in R by using the R function hclust(). 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 dendograms 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."

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

The accuracy of the four hierarchical clusterings was determined by calculating the purity. The purity is a measure of how much each cluster contains a single class. Purity is calculated by first determining the

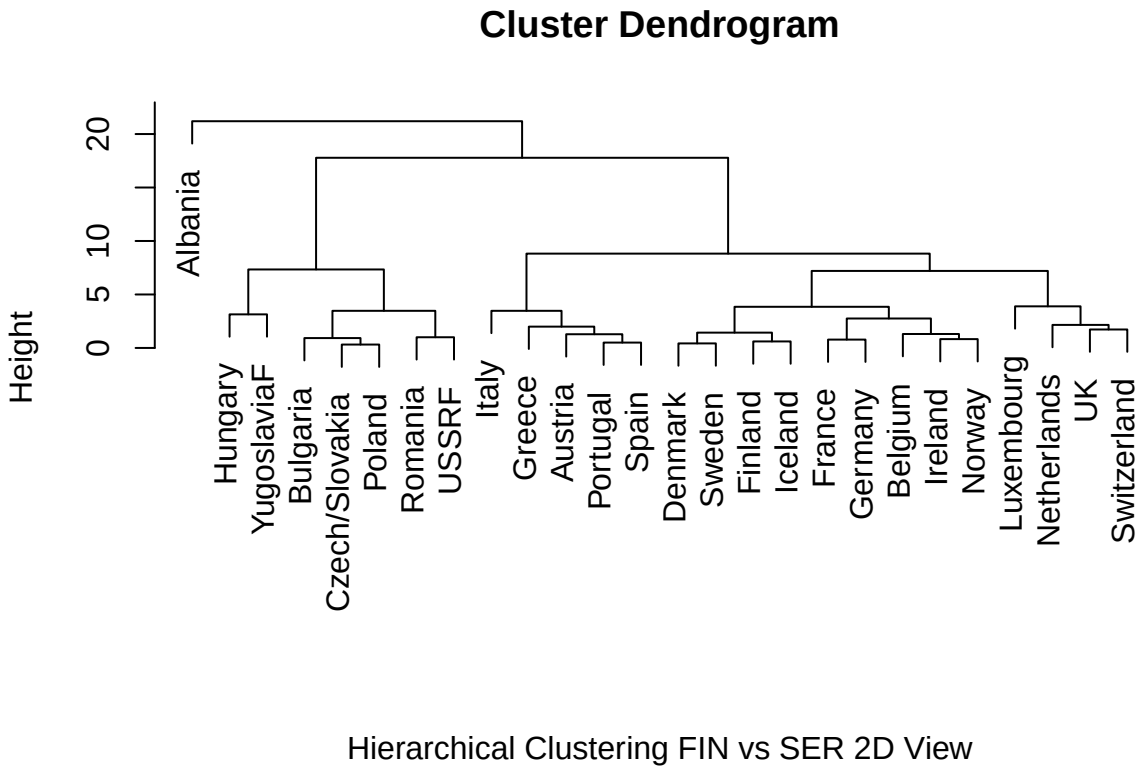


Figure 6: Dendrogram for FIN vs Ser

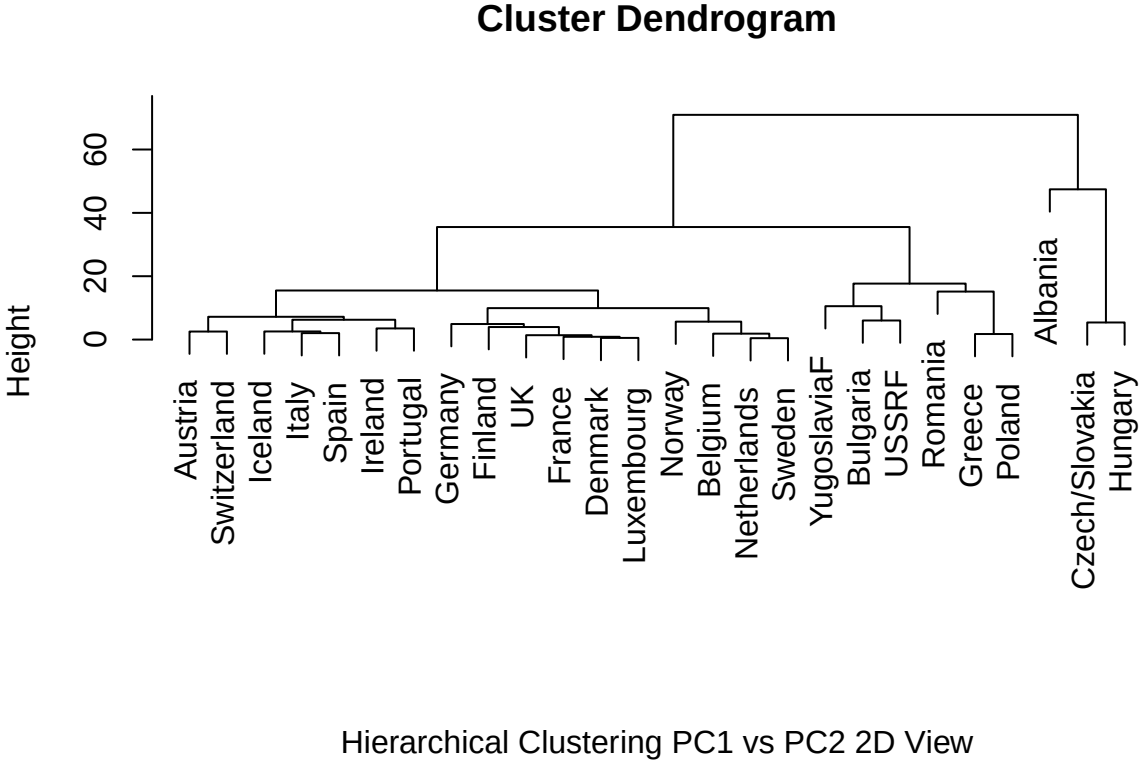


Figure 7: Dendrogram for FIN vs Ser

maximum value of the most common class in each cluster. The sum of the maximum values from each cluster is divided by the total number of countries used in the clustering analysis. The clustering of the countries by group using the finance and services data had the same accuracy of 76.9% using either three or six cluster groups. The second most accurate analysis at 73.0% was the clustering using the first two principal components and six cluster groups. The least accurate analysis was the clustering using the first two principal components and three cluster groups at 57.6%.

Table 3: Accuracy by Cluster Model

	Fin.Ser k=3	Fin.Ser k=6	PCA k=3	PCA k=6
Accuracy	0.7692308	0.7692308	0.5769231	0.7307692

Table 4: Accuracy by Cluster Model

	Accuracy
Fin.Ser k=3	0.7692308
Fin.Ser k=6	0.7692308
PCA k=3	0.5769231
PCA k=6	0.7307692

Section 6: k-Means Clustering Analysis

"Now let's apply k-means clustering to the same data. Do we need to know multiple methods for clustering? Yes. Since 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 clustering was conducted for three clusters with the finance and services data of all countries in the EU, EFTA, and Eastern groups. Three tables were generated from the clustering information. The first table indicates the cluster group for each country by finance and services data. The second table is a tally of the country group by each cluster group. The third table is the accuracy of the clustering analysis determined by the purity. The purity is a measure of how much each cluster contains a single class. Purity is calculated by first determining the maximum value of the most common class in each cluster. The sum of the maximum values from each cluster is divided by the total number of countries used in the clustering analysis.

```
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Table 5: FIN and SER Cluster Groups, k=3

Country	Group	Cluster	FIN	SER
Belgium	EU	2	8.7	16.9
Denmark	EU	1	9.1	14.5
France	EU	2	10.2	16.7
Germany	EU	2	9.6	17.2
Greece	EU	2	5.3	18.2

Country	Group	Cluster	FIN	SER
Ireland	EU	2	8.4	17.8
Italy	EU	2	4.6	21.6
Luxembourg	EU	2	8.7	21.2
Netherlands	EU	2	11.5	18.5
Portugal	EU	2	6.3	19.8
Spain	EU	2	5.9	20.1
UK	EU	2	12.4	20.2
Austria	EFTA	2	6.7	19.1
Finland	EFTA	1	8.6	14.6
Iceland	EFTA	1	8.0	14.5
Norway	EFTA	2	7.6	17.6
Sweden	EFTA	1	9.4	14.2
Switzerland	EFTA	2	10.7	20.5
Albania	Eastern	1	15.3	3.3
Bulgaria	Eastern	3	1.5	9.4
Czech/Slovakia	Eastern	3	1.6	10.2
Hungary	Eastern	3	0.0	13.3
Poland	Eastern	3	1.3	10.3
Romania	Eastern	3	0.6	6.9
USSRF	Eastern	3	0.6	7.9
YugoslaviaF	Eastern	3	3.1	13.8

Table 6: Country Group by Cluster Group, k=3

	1	2	3
Eastern	1	0	7
EFTA	3	3	0
EU	1	11	0
Other	0	0	0

Table 7: Accuracy of k-Means k=3

0.8076923

```
##         FIN      SER
## 1 10.080000 12.22000
## 2  8.328571 18.95714
## 3  1.242857 10.25714
```

"For k-means we can plot the original labels, their assigned clusters, and the cluster centers. Let's take a look at this plot. Here we can see how outliers affect clustering algorithms (they get assigned their own cluster), how our data is split (Eastern Block versus the rest of Europe), and where our four 'Other' countries would belong if we assigned them to a political group based on their economies."

"Now let's perform the k-means analysis with k=6."

```
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

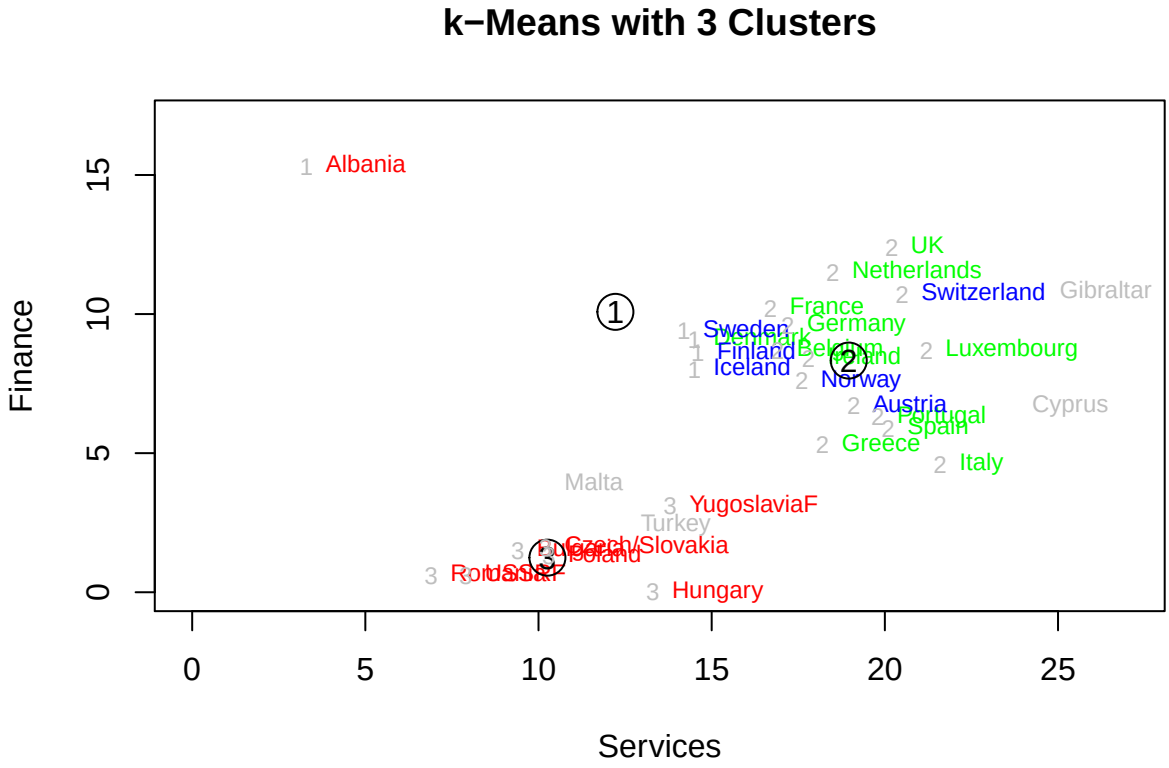


Figure 8: k-Means with k=3 for FIN vs SER

Table 8: FIN and SER Cluster Groups, k=6

Country	Group	Cluster	FIN	SER
Belgium	EU	6	8.7	16.9
Denmark	EU	6	9.1	14.5
France	EU	6	10.2	16.7
Germany	EU	6	9.6	17.2
Greece	EU	2	5.3	18.2
Ireland	EU	6	8.4	17.8
Italy	EU	2	4.6	21.6
Luxembourg	EU	2	8.7	21.2
Netherlands	EU	6	11.5	18.5
Portugal	EU	2	6.3	19.8
Spain	EU	2	5.9	20.1
UK	EU	6	12.4	20.2
Austria	EFTA	2	6.7	19.1
Finland	EFTA	6	8.6	14.6
Iceland	EFTA	6	8.0	14.5
Norway	EFTA	6	7.6	17.6
Sweden	EFTA	6	9.4	14.2
Switzerland	EFTA	6	10.7	20.5
Albania	Eastern	5	15.3	3.3
Bulgaria	Eastern	4	1.5	9.4
Czech/Slovakia	Eastern	4	1.6	10.2
Hungary	Eastern	1	0.0	13.3
Poland	Eastern	4	1.3	10.3

Country	Group	Cluster	FIN	SER
Romania	Eastern	3	0.6	6.9
USSRF	Eastern	3	0.6	7.9
YugoslaviaF	Eastern	1	3.1	13.8

Table 9: Country Group by Cluster Group, k=6

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

Table 10: Accuracy of k-Means k=6

0.7692308

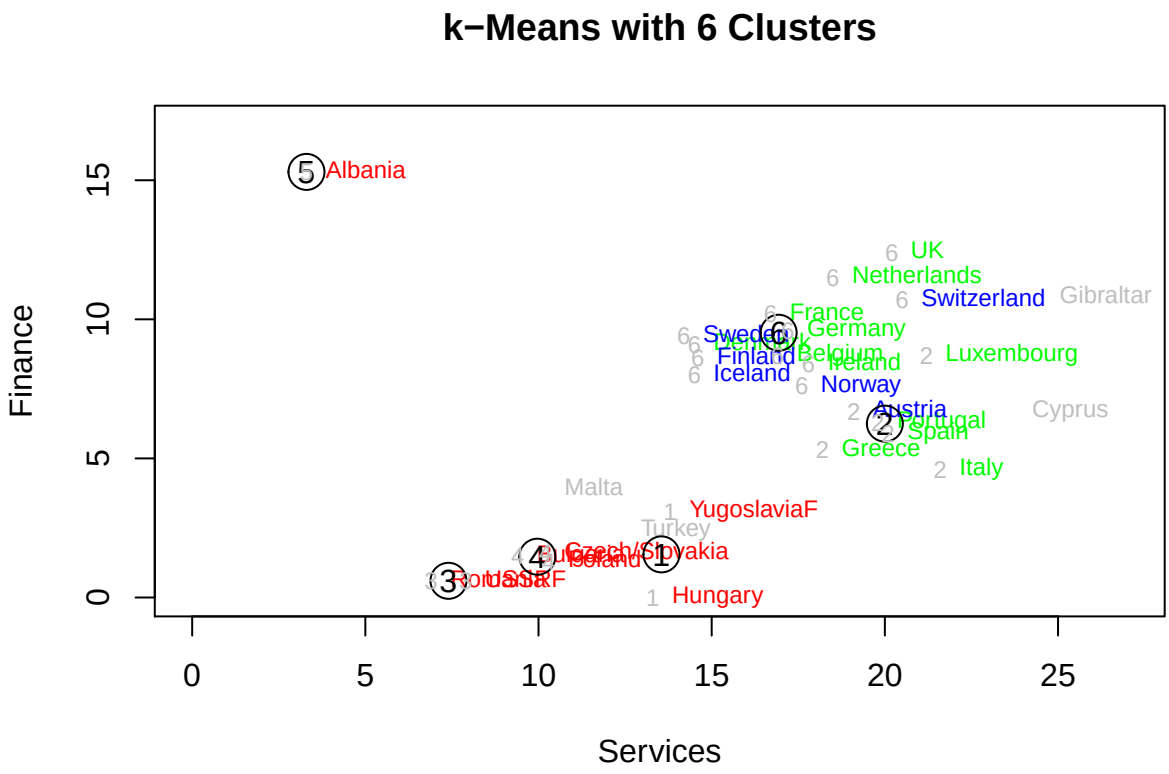


Figure 9: k-Means with k=6 for FIN vs SER

"And now our final set of clusters. Keeping in the mindset of model comparisons let's perform the same analysis in the principal components space using k=3 and k=6."

```
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Table 11: FIN and SER Cluster Groups, PCA k=3

Country	Group	Cluster	pc1	pc2
Belgium	EU	3	12.1400648	6.5255066
Denmark	EU	3	9.3609965	4.9952441
France	EU	3	8.7145586	4.4111197
Germany	EU	3	9.7105887	0.3280569
Greece	EU	2	-8.0714237	-7.3756556
Ireland	EU	3	-0.1417836	-1.8607105
Italy	EU	3	5.6638958	0.4678718
Luxembourg	EU	3	9.3313230	4.4785460
Netherlands	EU	3	11.3474415	7.9147962
Portugal	EU	3	2.8252977	-3.7578151
Spain	EU	3	3.8987954	-0.5443658
UK	EU	3	9.7006165	3.6577158
Austria	EFTA	3	6.2021775	-4.9636265
Finland	EFTA	3	5.7957262	3.2856011
Iceland	EFTA	3	3.3952117	1.6462748
Norway	EFTA	3	7.7162905	10.0525018
Sweden	EFTA	3	11.5508028	8.3008968
Switzerland	EFTA	3	7.0135500	-2.5647462
Albania	Eastern	1	-57.0032636	-9.3478981
Bulgaria	Eastern	2	-1.3443475	-17.6771879
Czech/Slovakia	Eastern	1	-27.1730477	25.6463406
Hungary	Eastern	1	-22.4120312	23.1025479
Poland	Eastern	2	-8.6757695	-9.0223738
Romania	Eastern	2	-6.2958648	-22.4208725
USSRF	Eastern	2	-1.6348877	-11.6619873
YugoslaviaF	Eastern	2	8.3850820	-13.6157808

Table 12: Country Group by Cluster Group, PCA k=3

	1	2	3
Eastern	3	5	0
EFTA	0	0	6
EU	0	1	11
Other	0	0	0

Table 13: Accuracy of PCA k-Means k=3

0.7307692

```
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Table 14: FIN and SER Cluster Groups, PCA k=6

Country	Group	Cluster	pc1	pc2
Belgium	EU	6	12.1400648	6.5255066

Country	Group	Cluster	pc1	pc2
Denmark	EU	2	9.3609965	4.9952441
France	EU	2	8.7145586	4.4111197
Germany	EU	2	9.7105887	0.3280569
Greece	EU	3	-8.0714237	-7.3756556
Ireland	EU	1	-0.1417836	-1.8607105
Italy	EU	2	5.6638958	0.4678718
Luxembourg	EU	2	9.3313230	4.4785460
Netherlands	EU	6	11.3474415	7.9147962
Portugal	EU	1	2.8252977	-3.7578151
Spain	EU	1	3.8987954	-0.5443658
UK	EU	2	9.7006165	3.6577158
Austria	EFTA	1	6.2021775	-4.9636265
Finland	EFTA	2	5.7957262	3.2856011
Iceland	EFTA	2	3.3952117	1.6462748
Norway	EFTA	6	7.7162905	10.0525018
Sweden	EFTA	6	11.5508028	8.3008968
Switzerland	EFTA	1	7.0135500	-2.5647462
Albania	Eastern	5	-57.0032636	-9.3478981
Bulgaria	Eastern	3	-1.3443475	-17.6771879
Czech/Slovakia	Eastern	4	-27.1730477	25.6463406
Hungary	Eastern	4	-22.4120312	23.1025479
Poland	Eastern	3	-8.6757695	-9.0223738
Romania	Eastern	3	-6.2958648	-22.4208725
USSRF	Eastern	3	-1.6348877	-11.6619873
YugoslaviaF	Eastern	1	8.3850820	-13.6157808

Table 15: Country Group by Cluster Group, PCA k=6

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

Table 16: Accuracy of PCA k-Means k=6

0.6923077

The cluster model that was the most accurate was the k-Means clustering using the finance and services data and six cluster groups at 80.7% purity.

Table 17: Accuracy by Cluster Model

	Accuracy
Hierarchical Fin.Ser k=3	0.7692308
Hierarchical Fin.Ser k=6	0.7692308
Hierarchical PCA k=3	0.5769231
Hierarchical PCA k=6	0.7307692

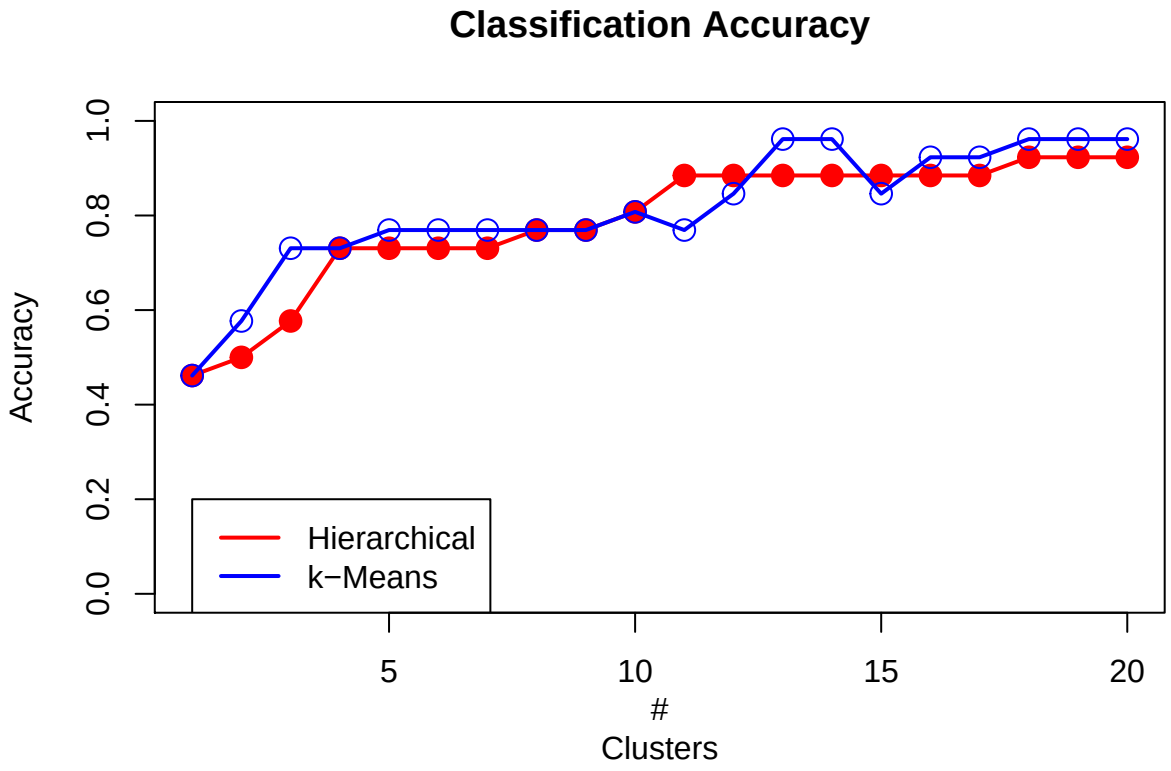
	Accuracy
k-Means Fin.Ser k=3	0.8076923
k-Means Fin.Ser k=6	0.7692308
k-Means PCA k=3	0.7307692
k-Means PCA k=6	0.6923077

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

"After completing our initial cluster analyses we should begin to wonder how many clusters would be the correct number of clusters, and how would we determine the correct number of clusters. 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. 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=20. 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. Maybe the best cluster model is the k-means cluster model with k=14. What do you think?"

My graph does not match the assignment handout. According to the graph below, maybe the best cluster model would be the k-means cluster model with k=16.



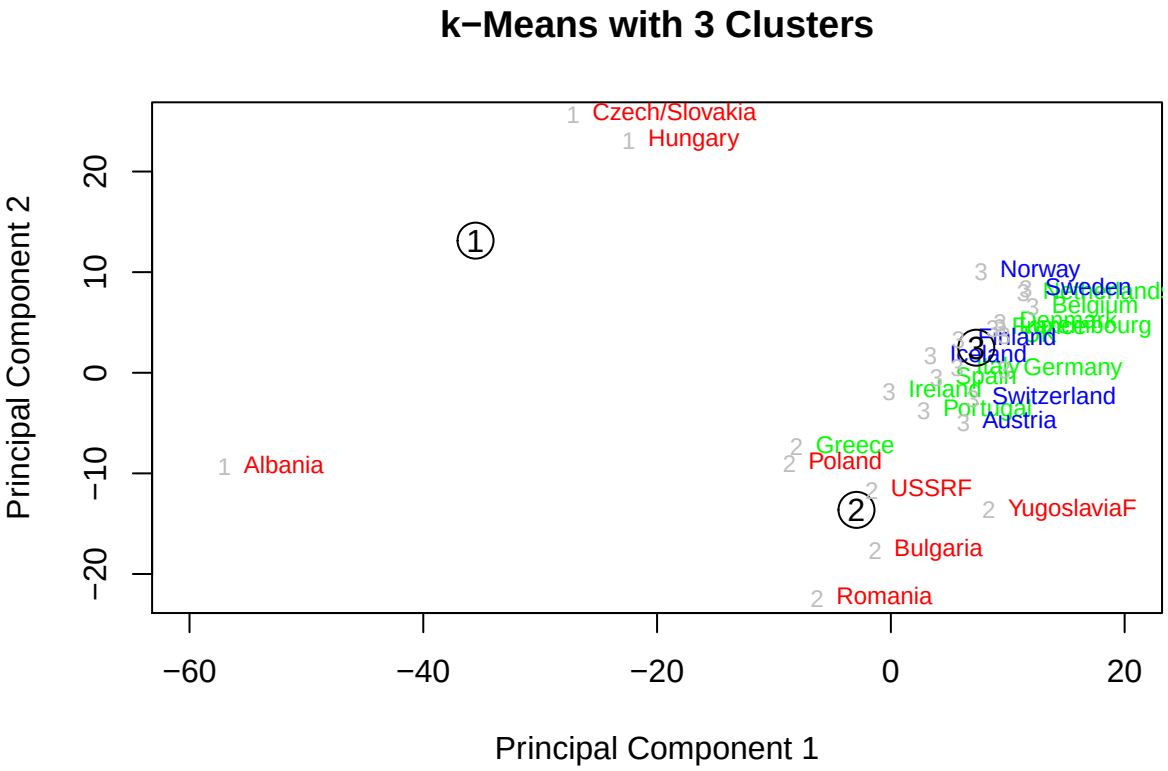


Figure 10: PCA k-Means with k=3

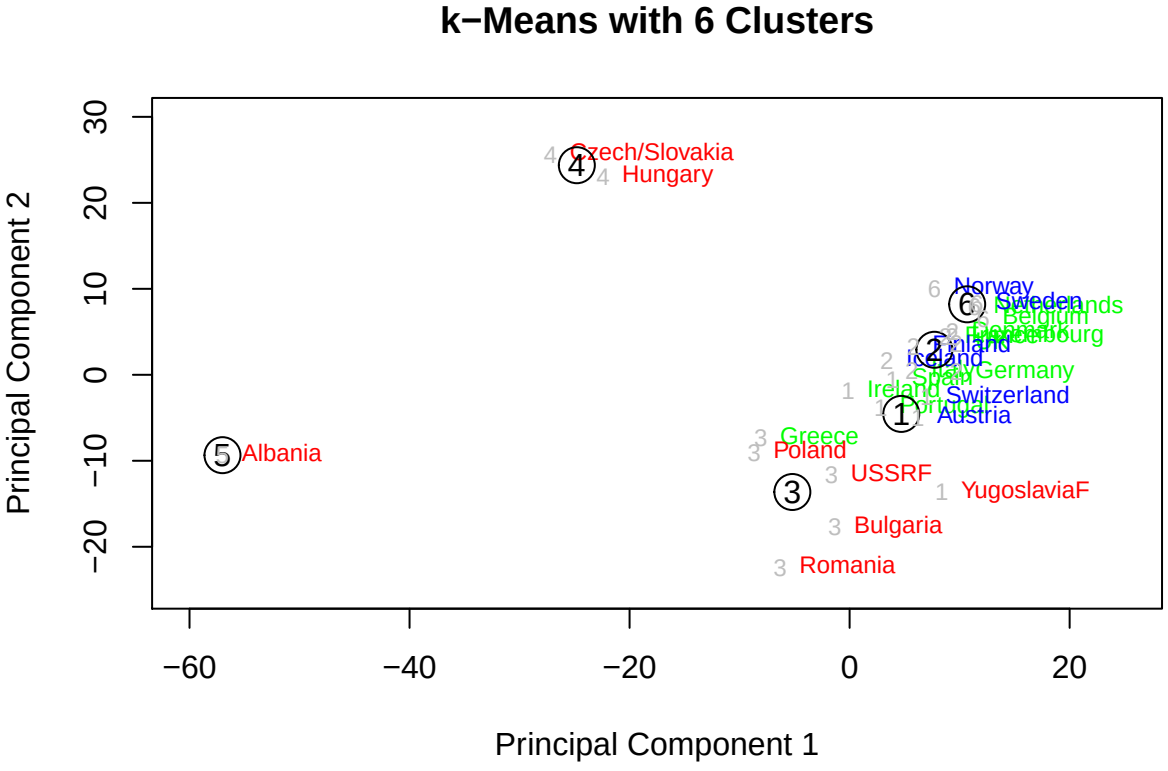


Figure 11: PCA k-Means with k=6