Clustering

Problem 1 With real data, true cluster labels are unknown. Use the Europe Employment data. Dataset shows the percentage of people employed in nine industry sectors in Europe for the years 1989 to 1995. The variables are the percentages employed in • AGR: Agriculture, forestry, and fishing • MIN: Mining and quarrying • MAN: Manufacturing • PS: Power and water supplies • CON: Construction • SER: Services • FIN: Finance • SPS: Social and personal services • TC: Transport and communications

```
euro <- read.csv("https://bit.ly/3ktLWfr", header=TRUE, row.names=1)
euro.c <- euro[-c(19,28), ]
head(euro)
```

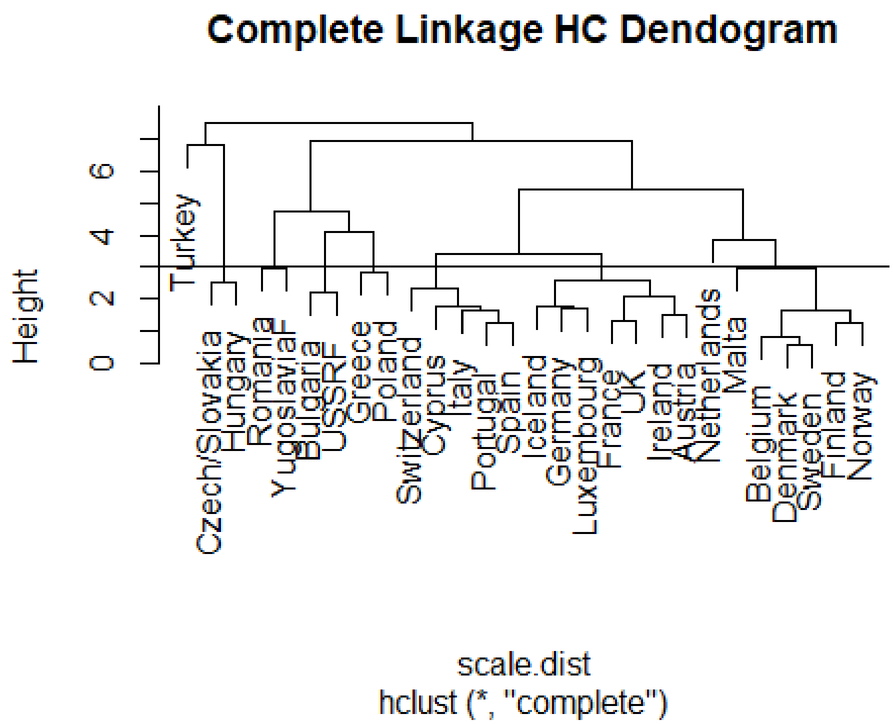
```
##         Group  AGR MIN  MAN  PS CON  SER  FIN  SPS  TC
## Belgium    EU  2.6 0.2 20.8 0.8 6.3 16.9  8.7 36.9 6.8
## Denmark    EU  5.6 0.1 20.4 0.7 6.4 14.5  9.1 36.3 7.0
## France     EU  5.1 0.3 20.2 0.9 7.1 16.7 10.2 33.1 6.4
## Germany    EU  3.2 0.7 24.8 1.0 9.4 17.2  9.6 28.4 5.6
## Greece     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
```

```
mydata = euro.c[, -1]
head(mydata)
```

```
##          AGR MIN  MAN  PS CON  SER  FIN  SPS  TC
## Belgium  2.6 0.2 20.8 0.8 6.3 16.9  8.7 36.9 6.8
## Denmark  5.6 0.1 20.4 0.7 6.4 14.5  9.1 36.3 7.0
## France   5.1 0.3 20.2 0.9 7.1 16.7 10.2 33.1 6.4
## Germany  3.2 0.7 24.8 1.0 9.4 17.2  9.6 28.4 5.6
## Greece  22.2 0.5 19.2 1.0 6.8 18.2  5.3 19.8 6.9
## Ireland 13.8 0.6 19.8 1.2 7.1 17.8  8.4 25.5 5.8
```

a) Create a hierarchical clustering dendrogram based on complete linkage (default). Don't forget to extract your distance matrix from the scaled data.

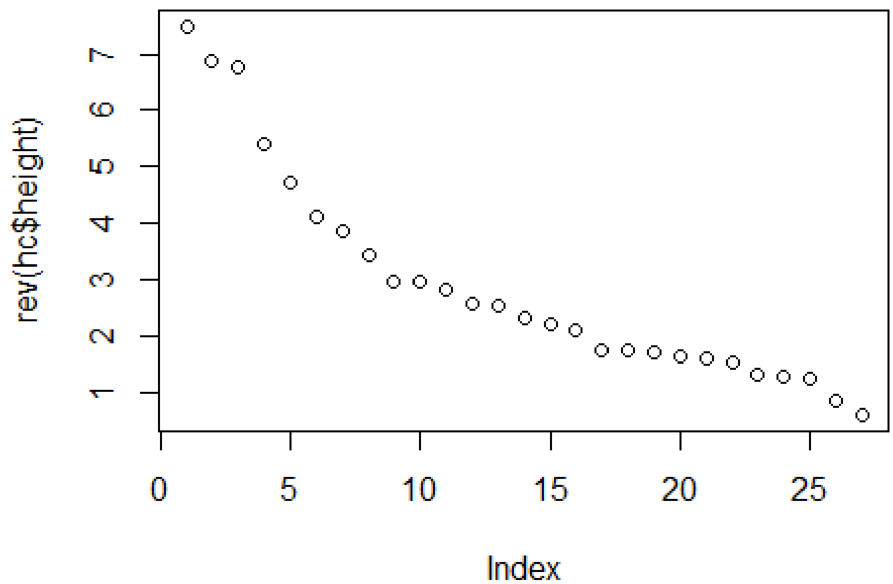
```
scale.dist = dist(scale(mydata))
hc <- hclust(scale.dist, "complete")
plot(hc, main = "Complete Linkage HC Dendogram")
abline(h=3)
```



b) Identify the appropriate number of clusters in Hierarchical clustering using a scree plot (hint: plot the reverse of hc$height).

The Number of clusters is: 3 based on the elbow test

```
round(hc$height,2)

##  [1] 0.61 0.85 1.26 1.29 1.34 1.54 1.62 1.66 1.71 1.74 1.75 2.11 2.20 2.33
2.52
## [16] 2.56 2.83 2.96 2.97 3.43 3.87 4.11 4.73 5.41 6.77 6.89 7.49

plot(rev(hc$height))
```



c) Based on your decision in part b, determine what countries are in which group?

The following countries are within cluster 1,2, and 3 below.

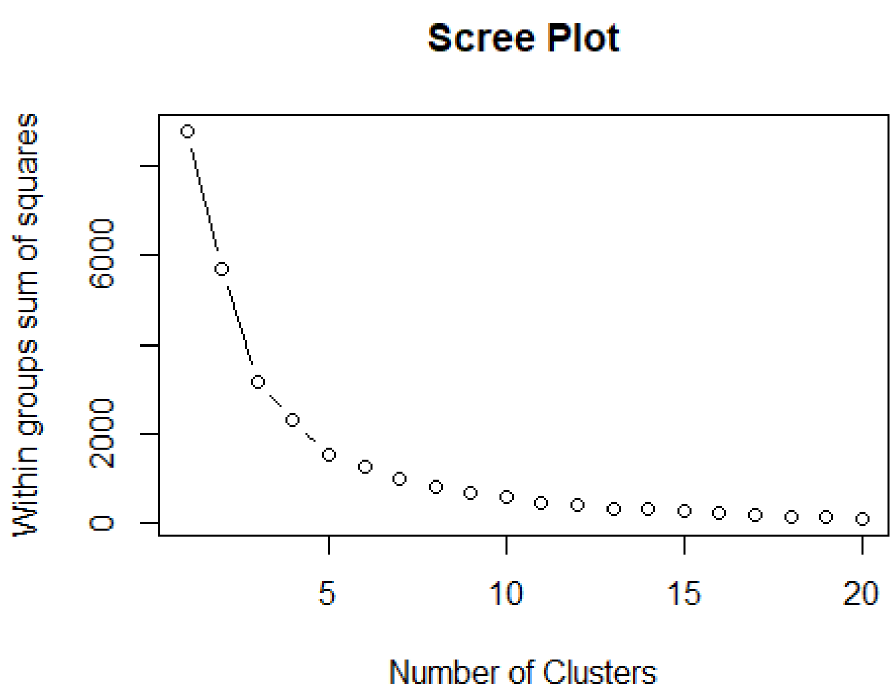
```
ct <- cutree(hc,3)
cat("\nCountries in cluster 1 \n")

##
## Countries in cluster 1

names(ct[ct == 1])

##  [1] "Belgium"     "Denmark"     "France"      "Germany"     "Ireland"
##  [6] "Italy"       "Luxembourg"  "Netherlands" "Portugal"    "Spain"
## [11] "UK"          "Austria"     "Finland"     "Iceland"     "Norway"
## [16] "Sweden"      "Switzerland" "Cyprus"      "Malta"

cat("\nCountries in cluster 2 \n")

##
## Countries in cluster 2

names(ct[ct == 2])

## [1] "Greece"      "Bulgaria"    "Poland"      "Romania"     "USSRF"
## [6] "YugoslaviaF"

cat("\nCountries in cluster 3 \n")
```

```
##
## Countries in cluster 3
names(ct[ct == 3])
## [1] "Czech/Slovakia" "Hungary"        "Turkey"
```

d) Identify the appropriate number of clusters in kmeans clustering based on the WGSS scree plot function.

From the scree plot we can deduce there are 3 clusters

```
plot.wgss = function(mydata, maxc) {
  wss = numeric(maxc)
  for (i in 1:maxc)
    wss[i] = kmeans(mydata,centers=i, nstart = 10)$tot.withinss
  plot(1:maxc, wss, type="b", xlab="Number of Clusters",
  ylab="Within groups sum of squares", main="Scree Plot")
}

plot.wgss(mydata, 20)
```



e) Based on your decision in part d, perform k-means clustering and determine what countries are in which group?

There are 2 countries in cluster 1, 6 countries in cluster 2, and 20 countries in cluster 3.They are listed below.

```
km <- kmeans(mydata, center = 3, nstart = 10)
table(km$cluster)

##
##  1  2  3
##  6 20  2

cat("\nCountries in cluster 1 \n")

##
## Countries in cluster 1

names(km$cluster[km$cluster == 1])

## [1] "Greece"   "Bulgaria" "Poland"   "Romania"  "USSRF"    "Turkey"

cat("\nCountries in cluster 2 \n")

##
## Countries in cluster 2

names(km$cluster[km$cluster == 2])

##  [1] "Belgium"     "Denmark"     "France"      "Germany"     "Ireland"
##  [6] "Italy"       "Luxembourg"  "Netherlands" "Portugal"    "Spain"
## [11] "UK"          "Austria"     "Finland"     "Iceland"     "Norway"
## [16] "Sweden"      "Switzerland" "YugoslaviaF" "Cyprus"      "Malta"

cat("\nCountries in cluster 3 \n")

##
## Countries in cluster 3

names(km$cluster[km$cluster == 3])

## [1] "Czech/Slovakia" "Hungary"
```

f) Attempt to identify the meanings of the clusters you found in part e by finding and interpreting the cluster centroids. • AGR: Agriculture, forestry, and fishing • MIN: Mining and quarrying • MAN: Manufacturing • PS: Power and water supplies • CON: Construction • SER: Services • FIN: Finance • SPS: Social and personal services • TC: Transport and communications

Looking at the clusters we see certain variables which stand out across clusters, such as AGR (Agriculture), MIN (Mining), MAN(Manufacturing) and PS (Power and water supplies), FIN. Apart from these three SPS and SER also show differences.

The very first cluster has most of its people working tertiary sector such as FIN , SER, SPS or secondary sector manufacturing. Very small amount of its population is into agriculture and mininig.

The second cluster is primarily engaged in mining and agriculture (47 %). There are no jobs in manufacturing.

The third cluster shows AGR and MAN and some services jobs.

```
round(km$centers,2)

##     AGR   MIN   MAN   PS  CON   SER  FIN   SPS   TC
## 1 25.02  1.32 26.72 0.68 6.83 10.85 1.95 20.10 6.53
## 2  6.60  0.50 22.07 0.89 7.49 17.63 8.00 30.25 6.53
## 3 14.05 33.10  0.00 0.00 7.40 11.75 0.80 25.10 7.85
```

g) Attempt to identify the meanings of the clusters you found in part f by plotting different pairs of principal component scores, the (PC1,PC2), (PC1,PC3), and (PC2,PC3) scatterplots, with points labeled (or colored) according to the assigned cluster. You can look at the loading of the first three PCs to find a meaning for each PC.

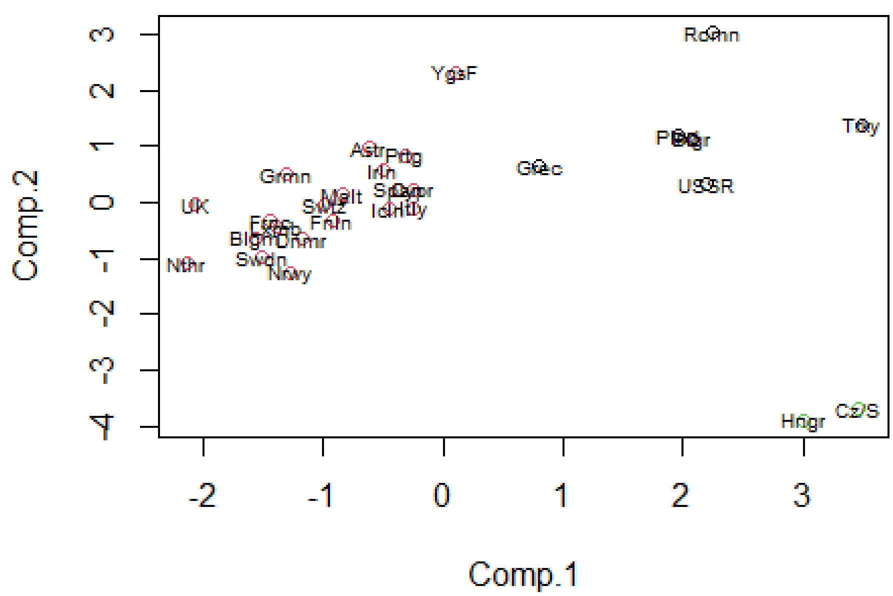
The first plot between components 1 and 2 we can see that cluster 2 show very high degree of separation from clusters 1 and 3 , and comp2 has a high correlation with mining and manufacturing both of which separate countries in cluster 2 from other countries. Similar separation is also seen in 3rd plot which also has component 2.

In the second plot we see that cluster 2 and 3 are close on the component 1 which has high correlation with agriculture.

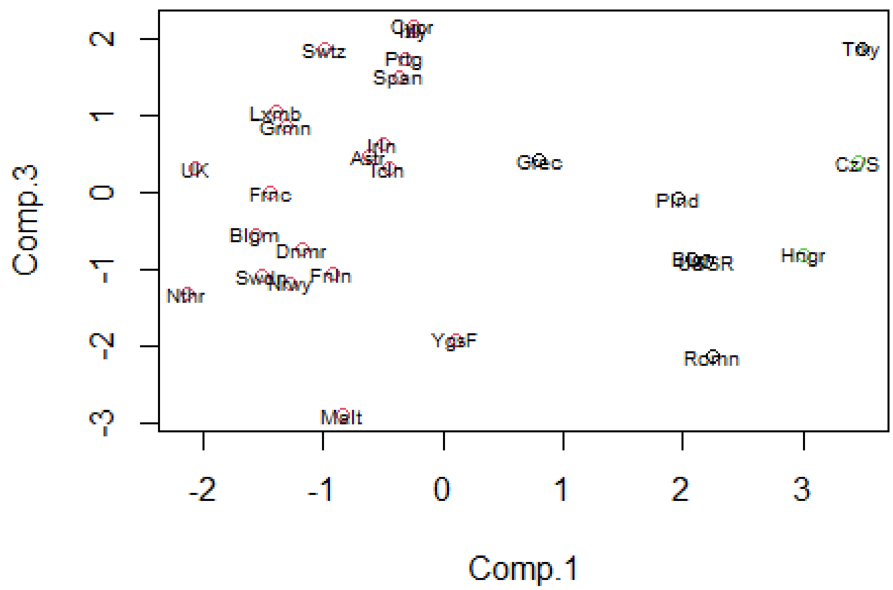
```
pca <- princomp(mydata, cor = T)
pca$loadings[,1:3]
```

```
##           Comp.1      Comp.2      Comp.3
## AGR  0.48364840  0.21179179  0.15321979
## MIN  0.34260148 -0.49066649 -0.03803186
## MAN -0.09502752  0.61846227 -0.20394526
## PS  -0.20066474  0.36413697 -0.38254381
## CON  0.02671019  0.04530866  0.39543849
## SER -0.38420771 -0.06921631  0.48893685
## FIN -0.53994266 -0.08089840  0.15245056
## SPS -0.39848393 -0.33498839 -0.29358673
## TC   0.02748388 -0.27146598 -0.53129903
```

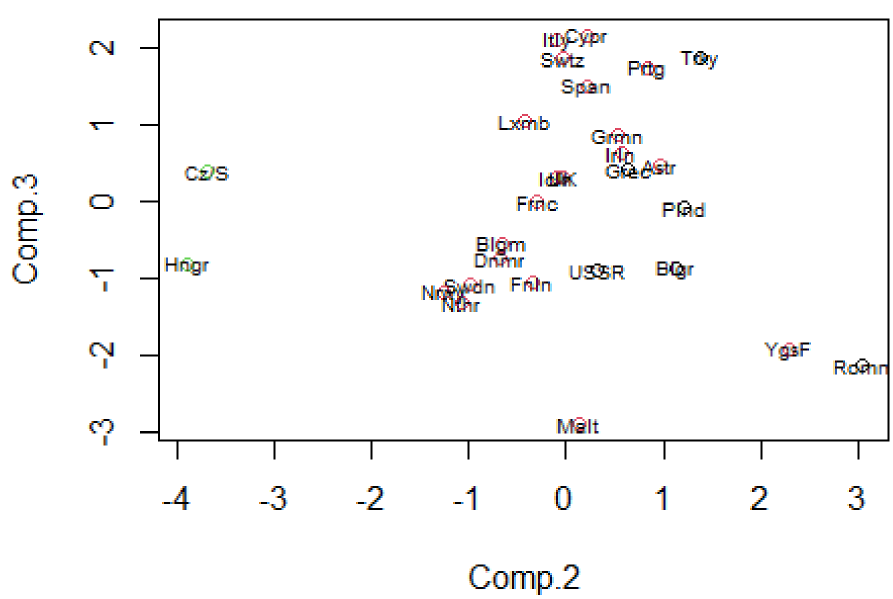
```r
plot(pca$scores[, 1:2], col = km$cluster)
text(pca$scores[,c(1,2)], labels = abbreviate(rownames(mydata)), cex = 0.6, c
ol = km$classification)
```




```r
plot(pca$scores[, c(1,3)], col = km$cluster)
text(pca$scores[,c(1,3)], labels = abbreviate(rownames(mydata)), cex = 0.6, c
ol = km$classification)
```



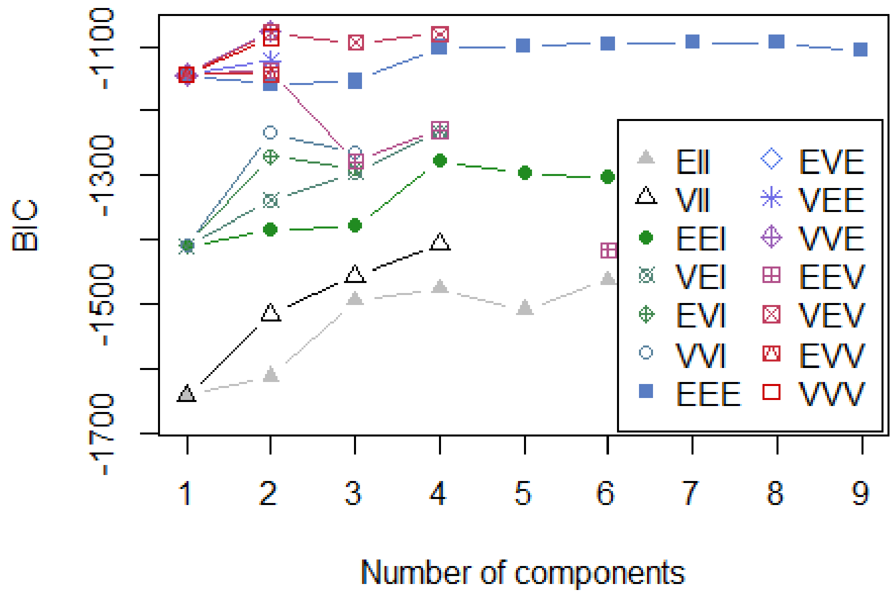

```r
plot(pca$scores[, 2:3], col = km$cluster)
text(pca$scores[,c(2,3)], labels = abbreviate(rownames(mydata)), cex = 0.6, c
ol = km$classification)
```



h) Perform model-based clustering without identifying the number of clusters. Plot the result of classification. How many groups are identified in your data? Determine what countries are in which group?

There are 2 clusters. listed below are the countries in each cluster.

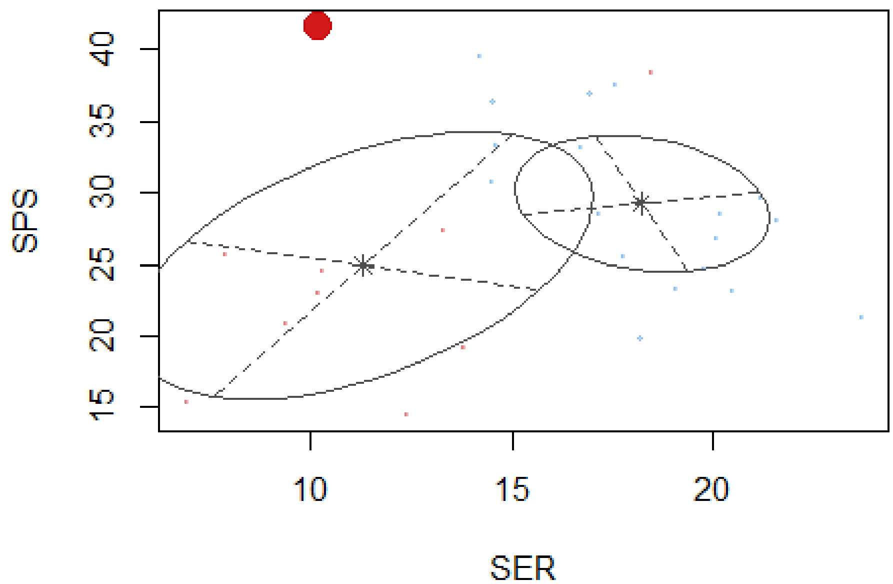
```
library("mclust")

## Warning: package 'mclust' was built under R version 4.0.3

## Package 'mclust' version 5.4.6
## Type 'citation("mclust")' for citing this R package in publications.

mc <- Mclust(mydata)
mc.clust <- mc$classification
table(mc.clust)

## mc.clust
##  1  2
## 18 10

cat("\nCountries in cluster 1 \n")

##
## Countries in cluster 1

names(mc.clust[mc.clust == 1])

##  [1] "Belgium"     "Denmark"     "France"      "Germany"     "Greece"
##  [6] "Ireland"     "Italy"       "Luxembourg"  "Portugal"    "Spain"
## [11] "UK"          "Austria"     "Finland"     "Iceland"     "Norway"
## [16] "Sweden"      "Switzerland" "Cyprus"

cat("\nCountries in cluster 2 \n")

##
## Countries in cluster 2

names(mc.clust[mc.clust == 2])

##  [1] "Netherlands"    "Bulgaria"       "Czech/Slovakia" "Hungary"
##  [5] "Poland"         "Romania"        "USSRF"          "YugoslaviaF"
##  [9] "Malta"          "Turkey"

plot(mc, "BIC")
```



i) Use “plot” on your fitted mclust object, and report the “uncertainty” plot for variables (SER, SPS), which is dimens = c(6,8). Explain the grouping of what country is more uncertain with what probability of uncertainty.

By cross referencing the values we see that the grouping of country Malta is uncertain.

```
plot(mc,  what = "uncertainty", dimens = c(6,8))
```



j) Construct the appropriate contingency table between the given grouping in the original cleaned data (euro.*cGroup*)*andthegroupswefoundinthemodel* − *basedclustering*. *Interpretthetable*, *andexplainhowwelldothemodel* − *basedclusterscorrespond*? *Codeforcontingencytableislike*: *table*(*euro*. *c*Group, mc$classification).

We found 2 clusters from mclust where as there were 4 groups in the cleaned data. Eastern and others were part of the 2nd cluster with one country from EU group belonging to the 2nd cluster which is Netherlands. The model based clustering works well but it was not able to distinguish between EFTA and EU countries. but neither were other models. Moreover if we look at the data we see that the averages for almost all variables are very close for EFTA and EU countries.

```
library("dplyr")

## Warning: package 'dplyr' was built under R version 4.0.3

##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:reshape':
##
##     rename

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
table(euro.c$Group, mc$classification)
```

```
##
##             1  2
##   Eastern  0  7
##   EFTA     6  0
##   EU      11  1
##   Other    1  2
```

```
euro %>% group_by(Group) %>% summarise(across(everything(), mean))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 4 x 10
##   Group     AGR   MIN   MAN    PS   CON   SER   FIN   SPS    TC
##   <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Eastern 21.5  11.8   20.6 0.638  6.91  9.39  3     19.4  6.8
## 2 EFTA     6.83  0.317 20.5 0.867  7.9  16.8   8.5   31.2  7.02
## 3 EU       7.67  0.45  21.0 0.792  7.28 18.6   8.39  29.6  6.21
## 4 Other   15.2   0.45  17.2 1.05   8.95 17.7   5.95  27.8  5.65
```