Assignment8_Seshadri

Sri Seshadri

8/12/2017

Assignment

In this assignment we will learn how to 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.

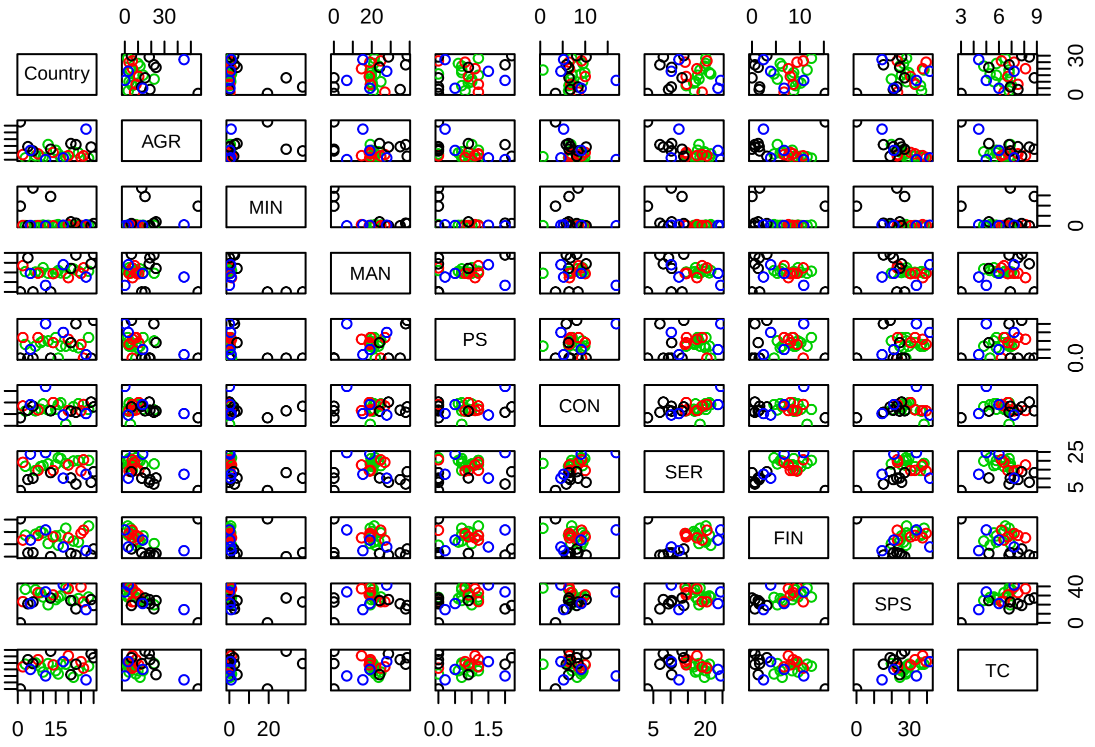
Part 1- The data

Let's begin by reading in the data.

```
my.data <- read.csv("European_Employment.csv", header = T)
levels(my.data$Group)
```

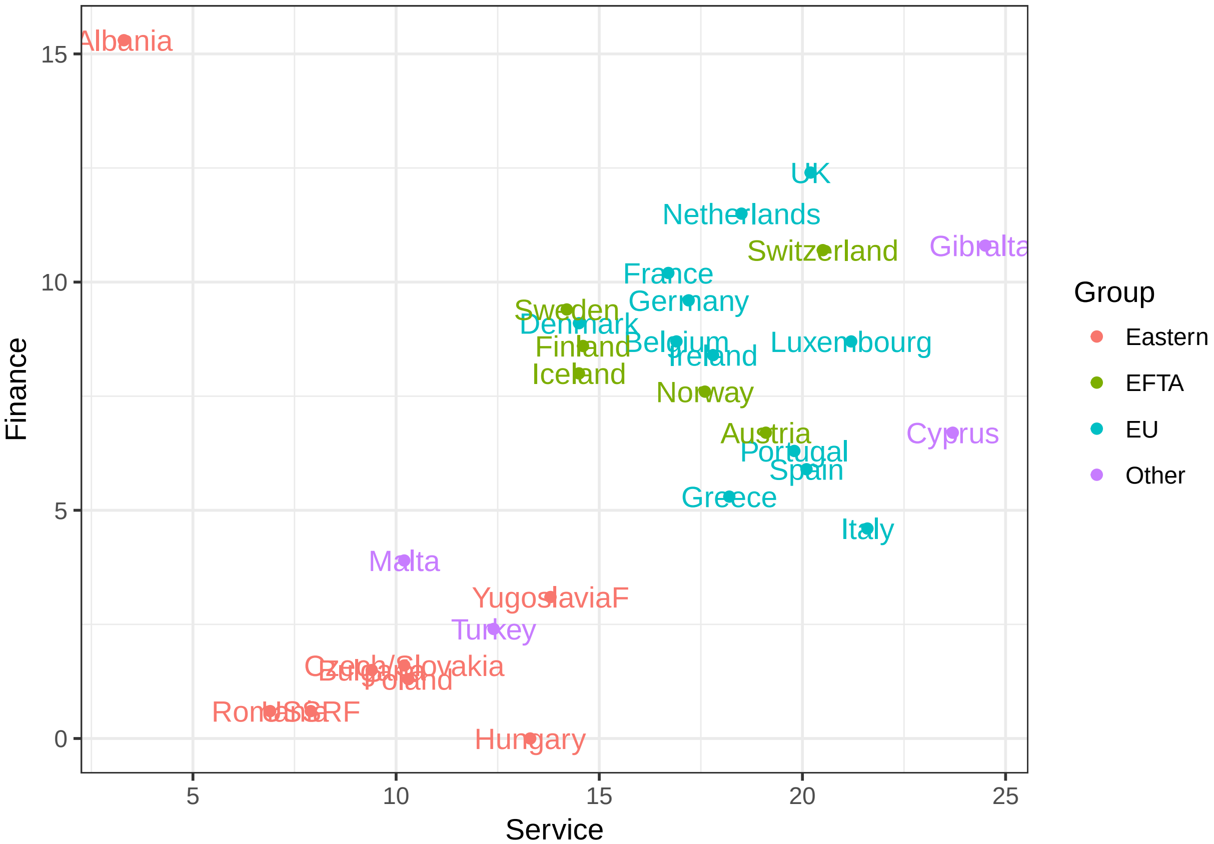
```
## [1] "Eastern" "EFTA"    "EU"      "Other"
```

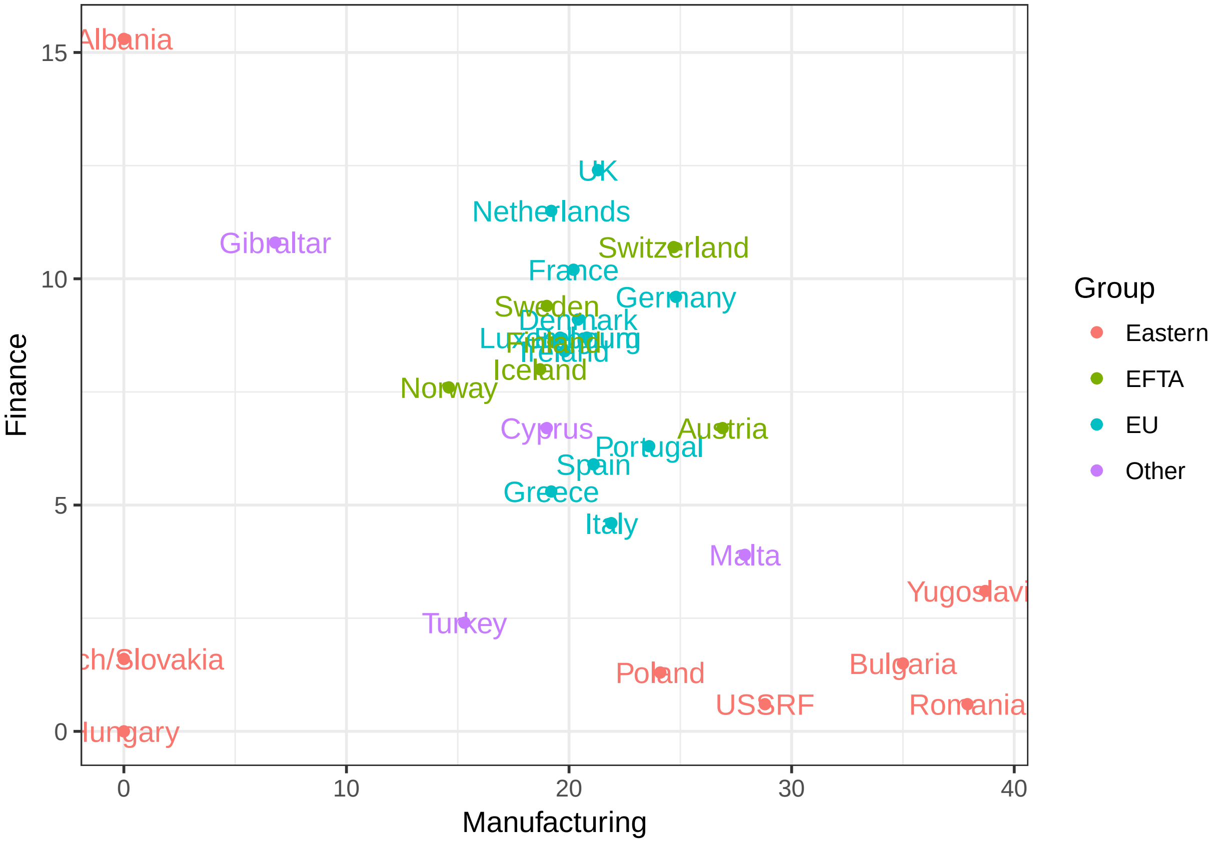
```
pairs(my.data[, -2], col = my.data$Group)
```



Appears that FIN vs SER and MAN vs FIN are interesting

2D plots


Finance
Service
15
10
5
0
5
10
15
20
25
Albania
UK
Netherlands
Switzerland
Gibralta
France
Germany
Sweden
Denmark
Finland
Belgium
Ireland
Luxembourg
Iceland
Norway
Austria
Cyprus
Portugal
Spain
Greece
Italy
Malta
YugoslaviaF
Turkey
Czech/Slovakia
Bulgaria
Poland
Romania
USSRF
Hungary
Group
● Eastern
● EFTA
● EU
● Other


Finance
Manufacturing
15
10
5
0
0
10
20
30
40
Albania
UK
Netherlands
Gibraltar
Switzerland
France
Sweden
Germany
Denmark
Luxembourg
Belgium
Finland
Ireland
Iceland
Norway
Cyprus
Austria
Portugal
Spain
Greece
Italy
Malta
Yugoslavi
Turkey
ch/Slovakia
Bulgaria
Poland
USSRF
Romania
lungary
Group
● Eastern
● EFTA
● EU
● Other

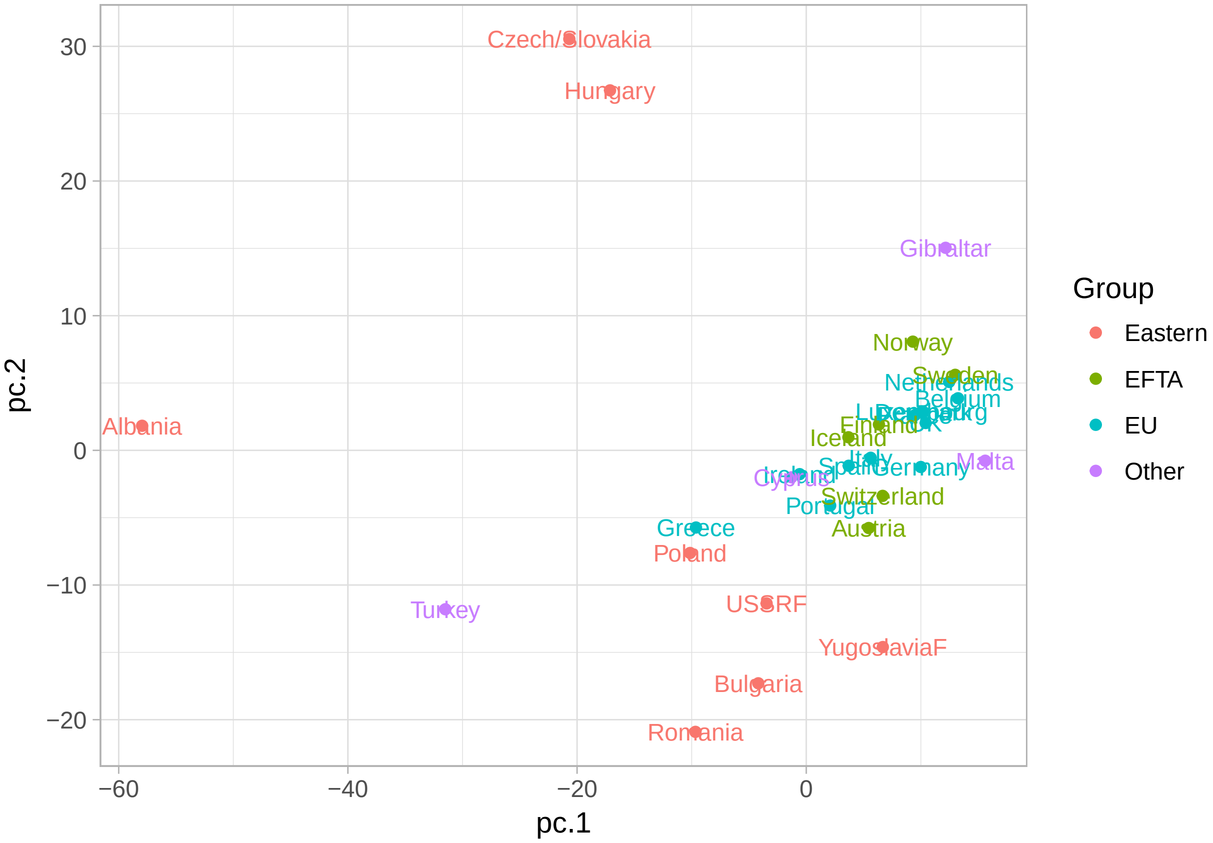
PCA

Each row is desogned to add to 100%. So if the data is standardized and scaled, we may lose the property. Hence PCA will be done without standardization.

```
pca <- princomp(my.data[, c(-1, -2)], cor = F)
summary(pca)
```

```
## Importance of components:
##                           Comp.1     Comp.2     Comp.3     Comp.4
## Standard deviation     15.380500 10.6893286 6.37294665 5.05126799
## Proportion of Variance  0.546685  0.2640566 0.09385918 0.05896537
## Cumulative Proportion   0.546685  0.8107416 0.90460079 0.96356616
##                            Comp.5     Comp.6      Comp.7       Comp.8
## Standard deviation     3.17639096 2.16103131 0.841192653 0.5459686803
## Proportion of Variance 0.02331654 0.01079241 0.001635261 0.0006888611
## Cumulative Proportion  0.98688270 0.99767510 0.999310365 0.9999992256
##                              Comp.9
## Standard deviation     1.830545e-02
## Proportion of Variance 7.743848e-07
## Cumulative Proportion  1.000000e+00
```

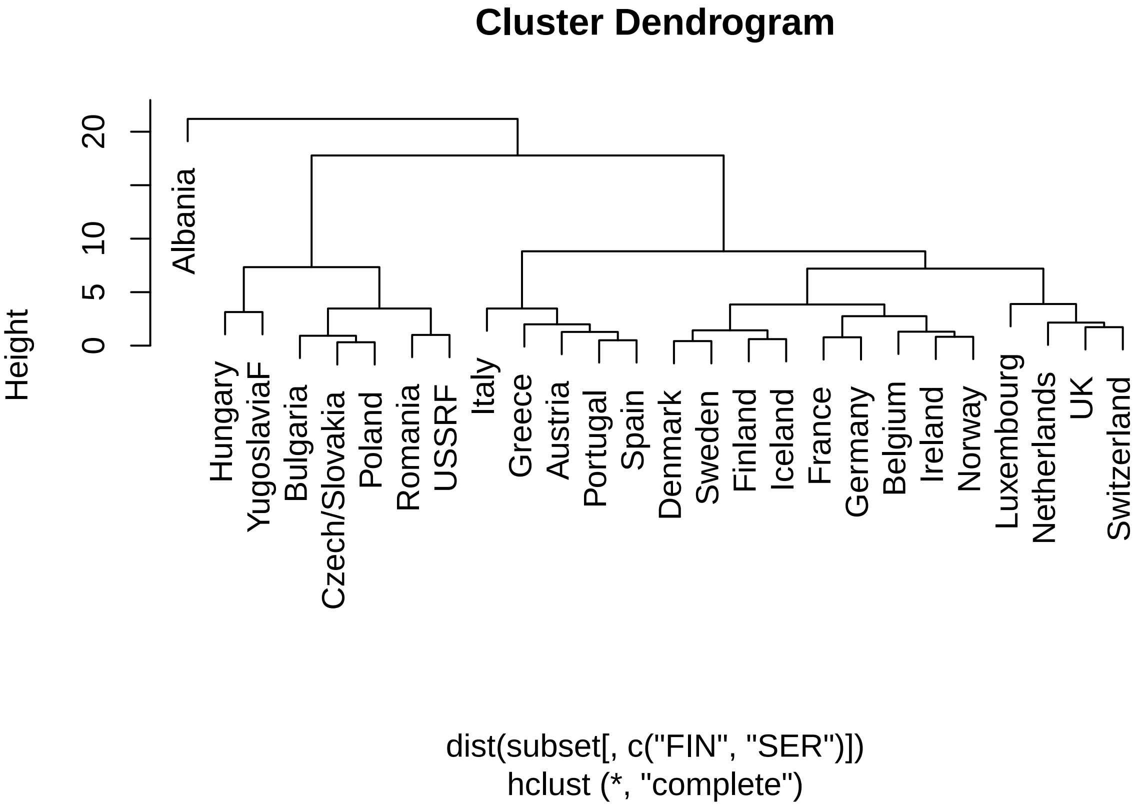
```
my.data$pc.1 <- pca$score[, 1]
my.data$pc.2 <- pca$score[, 2]
ggplot(data = my.data, mapping = aes(x = pc.1, y = pc.2, color = Group,
    label = Country)) + geom_point() + geom_text(size = 3, show.legend = F) +
    theme_light()
```



Hirearchical Clustering analysis

Let the drop the group - “other”. And what does the cluster look like when we attempt to reproduce the FIN vs Ser

```
my.data <- dplyr::as_tibble(my.data)
subset <- my.data %>% dplyr::filter(Group != "Other")
fin.ser <- hclust(dist(subset[, c("FIN", "SER")]), method = "complete")
plot(fin.ser, labels = subset$Country)
```



```
class <- cutree(fin.ser, k = 3)
class2 <- cutree(fin.ser, h = 12)  # class 1 & 2 are the same
subset$clust3 <- class
tab <- table(subset$Group, subset$clust3)
tab <- tab[1:3, ]
# t(tab) * (1/apply(tab,FUN = sum, MARGIN = 2)) tab

round(t(tab)/colSums(tab), 2)
```

```
##
##     Eastern EFTA   EU
##   1    0.00 0.33 0.67
##   2    1.00 0.00 0.00
##   3    1.00 0.00 0.00
```

```
class6 <- cutree(fin.ser, k = 6)
subset$clust6 <- class6
tab2 <- table(subset$Group, subset$clust6)
tab2 <- tab2[1:3, ]
```

```
# t(tab) * (1/apply(tab,FUN = sum, MARGIN = 2)) tab
round(t(tab2)/colSums(tab2), 2)
```

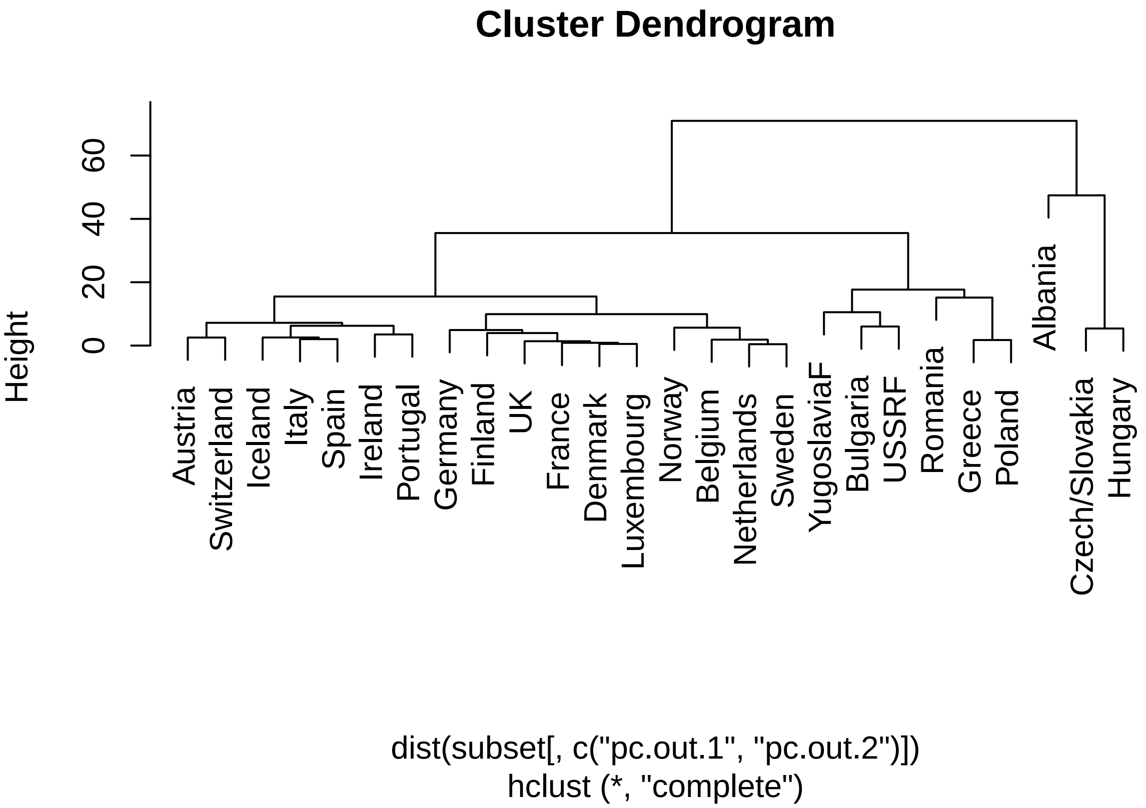
```
##
##     Eastern EFTA   EU
##   1    0.00 0.44 0.56
##   2    0.00 0.20 0.80
##   3    0.00 0.25 0.75
##   4    1.00 0.00 0.00
##   5    1.00 0.00 0.00
##   6    1.00 0.00 0.00
```

Lets try a PCA again

```
pca.out <- princomp(subset[, 3:11], cor = F)
summary(pca.out)
```

```
## Importance of components:
##                            Comp.1     Comp.2     Comp.3     Comp.4
## Standard deviation     14.9885834 10.6827988 6.06712838 4.36644154
## Proportion of Variance  0.5531982  0.2810151 0.09064127 0.04694776
## Cumulative Proportion   0.5531982  0.8342133 0.92485460 0.97180236
##                            Comp.5      Comp.6      Comp.7       Comp.8
## Standard deviation     2.72333343 1.756197858 0.848524441 0.4797367447
## Proportion of Variance 0.01826254 0.007594627 0.001772917 0.0005667161
## Cumulative Proportion  0.99006490 0.997659529 0.999432446 0.9999991616
##                              Comp.9
## Standard deviation     1.845179e-02
## Proportion of Variance 8.383717e-07
## Cumulative Proportion  1.000000e+00
```

```
subset$pc.out.1 <- pca.out$scores[, 1]
subset$pc.out.2 <- pca.out$scores[, 2]
pca.hclust <- hclust(dist(subset[, c("pc.out.1", "pc.out.2")]), method = "complete")
plot(pca.hclust, labels = subset$Country)
```



```
tab3 <- table(subset$Group, cutree(pca.hclust, k = 3))
round(t(tab3)/colSums(tab3), 2)
```

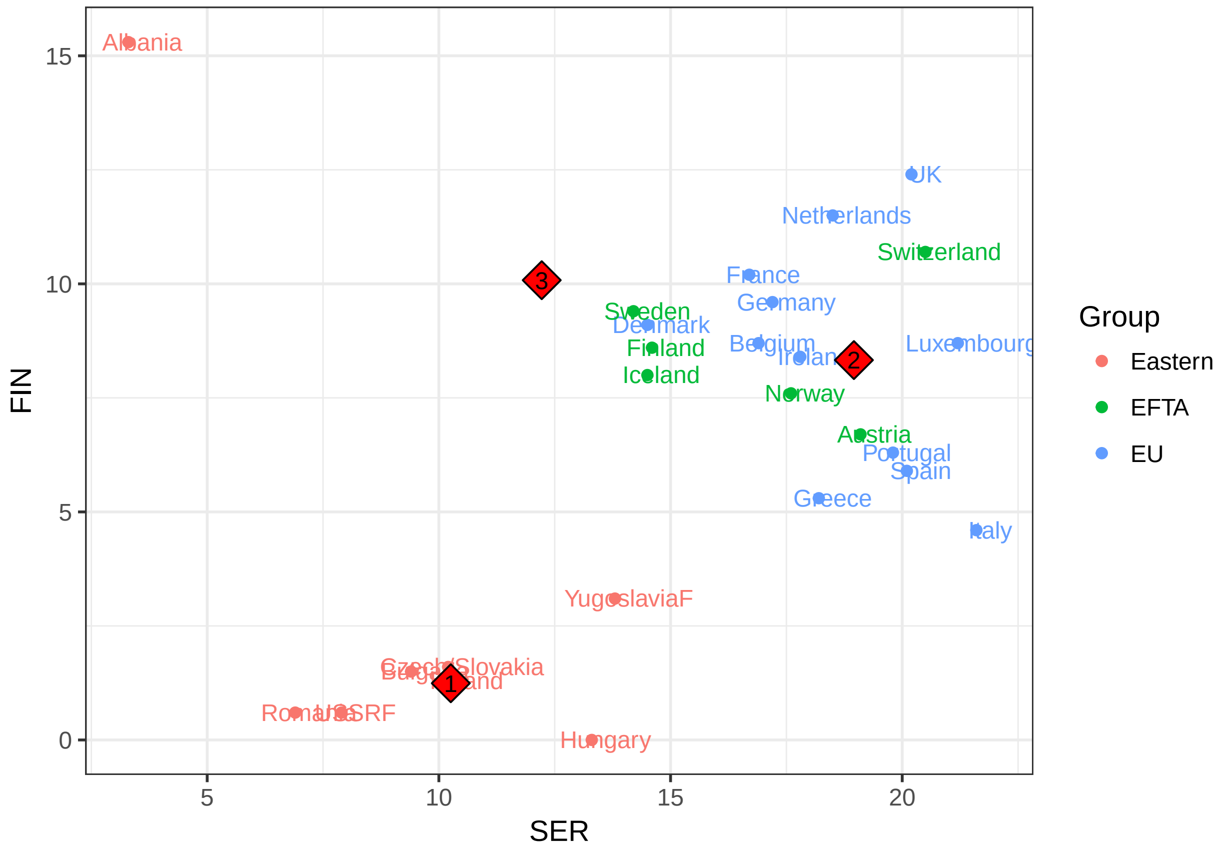
```
##
##     Eastern EFTA   EU Other
##   1    0.22 0.26 0.52  0.00
##   2    1.00 0.00 0.00  0.00
##   3    1.00 0.00 0.00  0.00
```

```
tab4 <- table(subset$Group, cutree(pca.hclust, k = 6))
round(t(tab4)/colSums(tab4), 2)
```

```
##
##     Eastern EFTA   EU Other
##   1    0.00 0.30 0.70  0.00
##   2    0.67 0.00 0.33  0.00
##   3    0.00 0.43 0.57  0.00
##   4    1.00 0.00 0.00  0.00
##   5    1.00 0.00 0.00  0.00
##   6    1.00 0.00 0.00  0.00
```

K means clustering

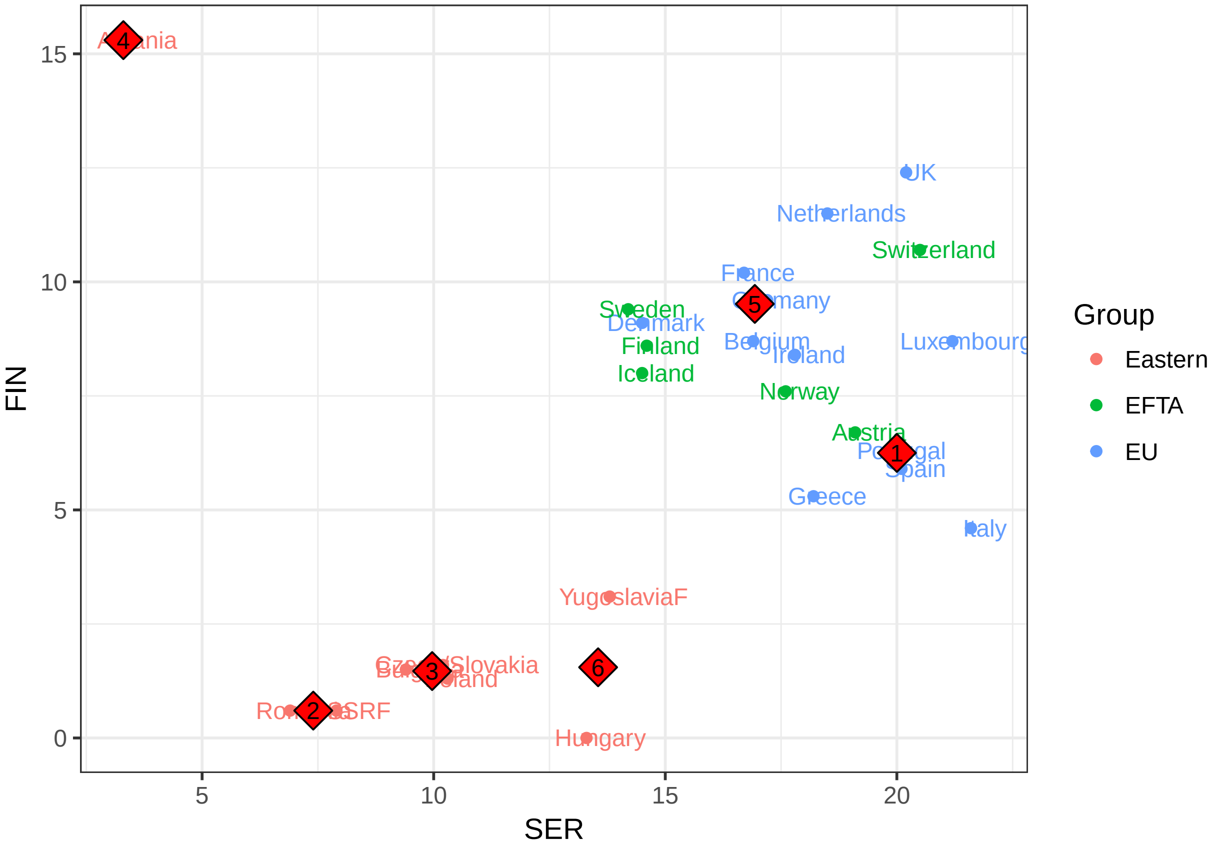
```
finser <- kmeans(x = subset[, c("FIN", "SER")], centers = 3)
subset$kmeans3 <- finser$cluster
centers <- as.data.frame(finser$centers)
ggplot() + geom_point(data = subset, mapping = aes(y = FIN, x = SER, color = Group)) +
    geom_text(data = subset, mapping = aes(y = FIN, x = SER, color = Group,
        label = Country), show.legend = F, size = 3, nudge_x = 0.3, nudge_y = 0.01) +
    geom_point(data = centers, mapping = aes(y = FIN, x = SER), shape = 23,
        fill = "red", size = 5) + geom_text(data = centers, mapping = aes(y = FIN,
    x = SER, label = rownames(centers)), size = 3) + theme_bw()
```



```
tab5 <- table(subset$Group, subset$kmeans3)
t(tab5)/colSums(tab5)
```

```
## 
##       Eastern      EFTA        EU     Other
##   1 1.0000000 0.0000000 0.0000000 0.0000000
##   2 0.0000000 0.2142857 0.7857143 0.0000000
##   3 0.2000000 0.6000000 0.2000000 0.0000000
```

```
finser <- kmeans(x = subset[, c("FIN", "SER")], centers = 6)
subset$kmeans6 <- finser$cluster
centers <- as.data.frame(finser$centers)
ggplot() + geom_point(data = subset, mapping = aes(y = FIN, x = SER, color = Group)) +
    geom_text(data = subset, mapping = aes(y = FIN, x = SER, color = Group,
        label = Country), show.legend = F, size = 3, nudge_x = 0.3, nudge_y = 0.01) +
    geom_point(data = centers, mapping = aes(y = FIN, x = SER), shape = 23,
        fill = "red", size = 5) + geom_text(data = centers, mapping = aes(y = FIN,
    x = SER, label = rownames(centers)), size = 3) + theme_bw()
```

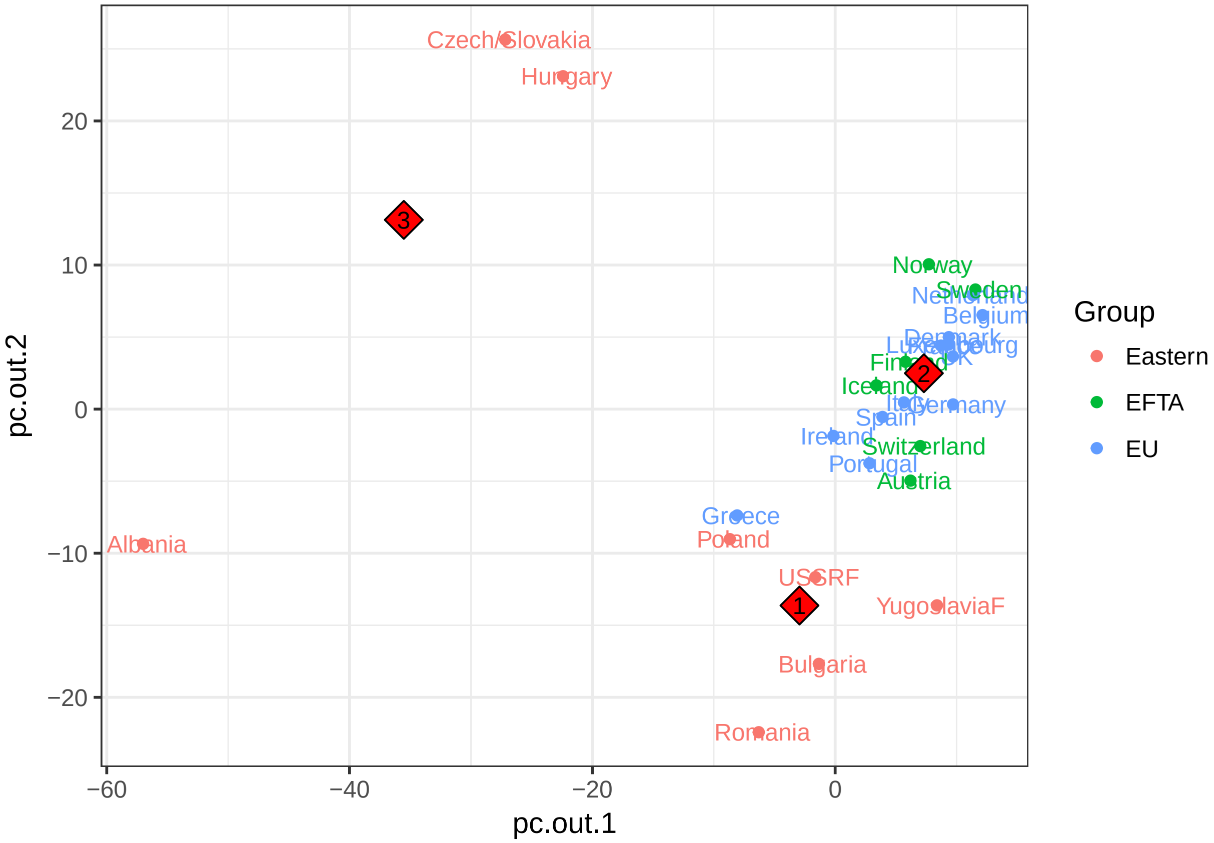


```
tab6 <- table(subset$Group, subset$kmeans6)
t(tab6)/colSums(tab6)
```

```
##
##       Eastern      EFTA        EU     Other
##   1 0.0000000 0.1666667 0.8333333 0.0000000
##   2 1.0000000 0.0000000 0.0000000 0.0000000
##   3 1.0000000 0.0000000 0.0000000 0.0000000
##   4 1.0000000 0.0000000 0.0000000 0.0000000
##   5 0.0000000 0.4166667 0.5833333 0.0000000
##   6 1.0000000 0.0000000 0.0000000 0.0000000
```

Hmmmm. . . . Let’s do this on the principal component scores

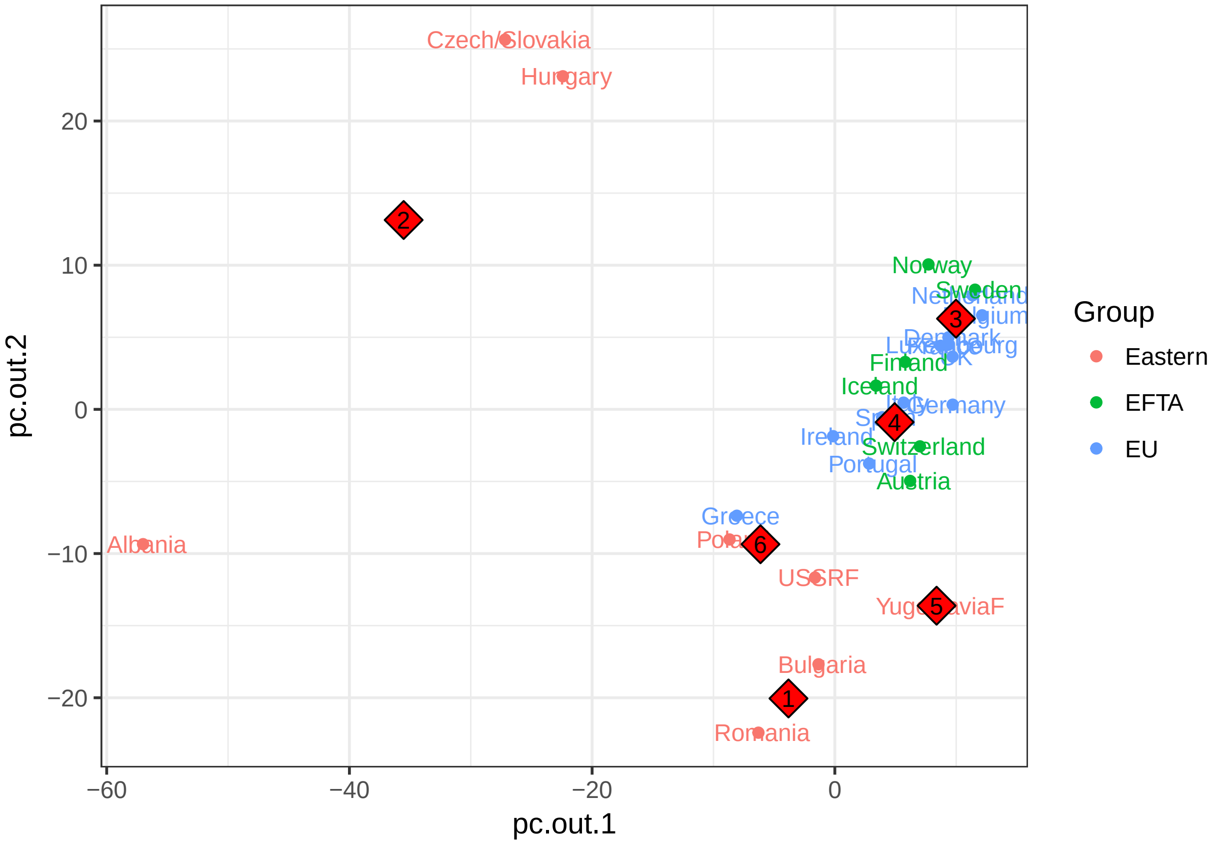
```
finser <- kmeans(x = subset[, c("pc.out.1", "pc.out.2")], centers = 3)
subset$kmeans3 <- finser$cluster
centers <- as.data.frame(finser$centers)
ggplot() + geom_point(data = subset, mapping = aes(y = pc.out.2, x = pc.out.1,
    color = Group)) + geom_text(data = subset, mapping = aes(y = pc.out.2,
    x = pc.out.1, color = Group, label = Country), show.legend = F, size = 3,
    nudge_x = 0.3, nudge_y = 0.01) + geom_point(data = centers, mapping = aes(y = pc.out.2,
    x = pc.out.1), shape = 23, fill = "red", size = 5) + geom_text(data = centers,
    mapping = aes(y = pc.out.2, x = pc.out.1, label = rownames(centers)),
    size = 3) + theme_bw()
```



```
tab5 <- table(subset$Group, subset$kmeans3)
t(tab5)/colSums(tab5)
```

```
##
##       Eastern      EFTA        EU     Other
##   1 0.8333333 0.0000000 0.1666667 0.0000000
##   2 0.0000000 0.3529412 0.6470588 0.0000000
##   3 1.0000000 0.0000000 0.0000000 0.0000000
```

```
finser <- kmeans(x = subset[, c("pc.out.1", "pc.out.2")], centers = 6)
subset$kmeans3 <- finser$cluster
centers <- as.data.frame(finser$centers)
ggplot() + geom_point(data = subset, mapping = aes(y = pc.out.2, x = pc.out.1,
    color = Group)) + geom_text(data = subset, mapping = aes(y = pc.out.2,
    x = pc.out.1, color = Group, label = Country), show.legend = F, size = 3,
    nudge_x = 0.3, nudge_y = 0.01) + geom_point(data = centers, mapping = aes(y = pc.out.2,
    x = pc.out.1), shape = 23, fill = "red", size = 5) + geom_text(data = centers,
    mapping = aes(y = pc.out.2, x = pc.out.1, label = rownames(centers)),
    size = 3) + theme_bw()
```



```
tab5 <- table(subset$Group, subset$kmeans3)
t(tab5)/colSums(tab5)
```

```
##
##      Eastern      EFTA        EU     Other
##   1 1.0000000 0.0000000 0.0000000 0.0000000
##   2 1.0000000 0.0000000 0.0000000 0.0000000
##   3 0.0000000 0.2500000 0.7500000 0.0000000
##   4 0.0000000 0.4444444 0.5555556 0.0000000
##   5 1.0000000 0.0000000 0.0000000 0.0000000
##   6 0.6666667 0.0000000 0.3333333 0.0000000
```

```
sum(tab5[1:3, ])
```

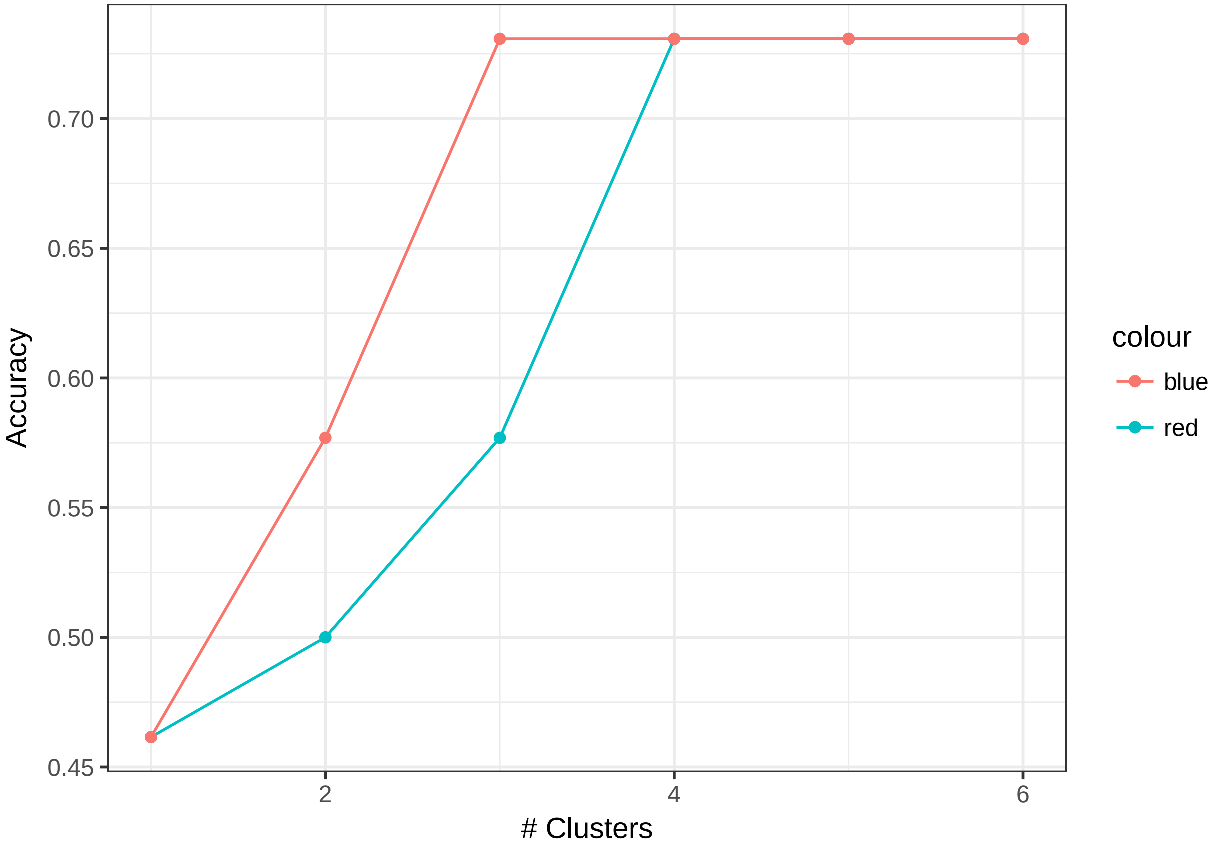
```
## [1] 26
```

```
Accuracy <- function(cluster, data = subset) {
    # lets get the hierarchical clustering accuracy
    hclust <- hclust(dist(data[, c("AGR", "MIN", "MAN", "PS", "CON", "SER",
        "FIN", "SPS", "TC")]), method = "complete")
    data$hiercluster <- cutree(hclust, k = cluster)
    tab <- table(data$Group, data$hiercluster)
    hier.Accuracy <- ifelse(cluster == 1, max(tab[1:3, ])/sum(tab[1:3,
        ]), sum(apply(tab[1:3, ], MARGIN = 2, FUN = max))/sum(colSums(tab[1:3,
        ])))

    # k means
    kclust <- kmeans(data[, c("AGR", "MIN", "MAN", "PS", "CON", "SER",
        "FIN", "SPS", "TC")], centers = cluster)
    data$kclust <- kclust$cluster
    tab2 <- table(data$Group, data$kclust)
    kmean.Accuracy <- ifelse(cluster == 1, max(tab[1:3, ])/sum(tab[1:3,
        ]), sum(apply(tab2[1:3, ], MARGIN = 2, FUN = max))/sum(colSums(tab2[1:3,
        ])))

    return(data.frame(cluster = cluster, hier.Accuracy = hier.Accuracy,
        kmean.Accuracy = kmean.Accuracy))
}

AccuracyResults <- purrr::map_df(.x = 1:6, .f = Accuracy)
ggplot(data = AccuracyResults, mapping = aes(x = cluster)) + geom_line(aes(y = hier.Accuracy,
    color = "red")) + geom_point(aes(y = hier.Accuracy, color = "red")) +
    geom_line(aes(y = kmean.Accuracy, color = "blue")) + geom_point(aes(y = kmean.Accuracy,
    color = "blue")) + theme_bw() + xlab("# Clusters") + ylab("Accuracy")
```


Accuracy
# Clusters
colour
blue
red