Homework 2 - Correspondence and Cluster Analyses

Matija Jakovac, Dunja Petrović, Léo Serra

2025-05-23

1. First do the exploratory data analysis.

```
olympic_df <- read.csv("olympic2000.csv", stringsAsFactors = FALSE)
str(olympic_df)
```

```
## 'data.frame':    66 obs. of  13 variables:
##  $ Country       : chr  "United States" "Russia" "China" "Australia" ...
##  $ Gold2000      : int  39 32 28 16 14 13 14 11 11 8 ...
##  $ Silver2000    : int  25 28 16 25 17 14 8 11 10 9 ...
##  $ Bronze2000    : int  33 28 15 17 26 11 13 7 7 11 ...
##  $ Total2000     : int  97 88 59 58 57 38 35 29 28 28 ...
##  $ Population    : int  274028 147434 1255698 18520 82133 58683 57369 11116 58649 23348 ...
##  $ Athletes      : int  664 468 295 655 497 359 413 245 355 292 ...
##  $ GDP           : int  7567100 356030 906079 367802 2364632 1533619 1140484 18600 1152136 483130 ..
##  $ Total1996     : int  101 63 50 41 65 37 35 25 15 27 ...
##  $ Log1996       : num  4.62 4.14 3.91 3.71 4.17 ...
##  $ Log.population: num  12.52 11.9 14.04 9.83 11.32 ...
##  $ Log.GDP       : num  15.8 12.8 13.7 12.8 14.7 ...
##  $ Log.athletes  : num  6.5 6.15 5.69 6.48 6.21 ...
```

a) Import the data set correctly to R and assign type of each variable correctly and assign the country names as labels to the rows of the data frame

```
rownames(olympic_df) <- olympic_df$Country
olympic_df$Country <- NULL
head(olympic_df, n=3)
```

```
##               Gold2000 Silver2000 Bronze2000 Total2000 Population Athletes
## United States       39         25         33        97     274028      664
## Russia              32         28         28        88     147434      468
## China               28         16         15        59    1255698      295
##                   GDP Total1996  Log1996 Log.population  Log.GDP Log.athletes
## United States 7567100       101 4.615121       12.52099 15.83932     6.498282
## Russia         356030        63 4.143135       11.90114 12.78277     6.148468
## China          906079        50 3.912023       14.04320 13.71688     5.686975
```

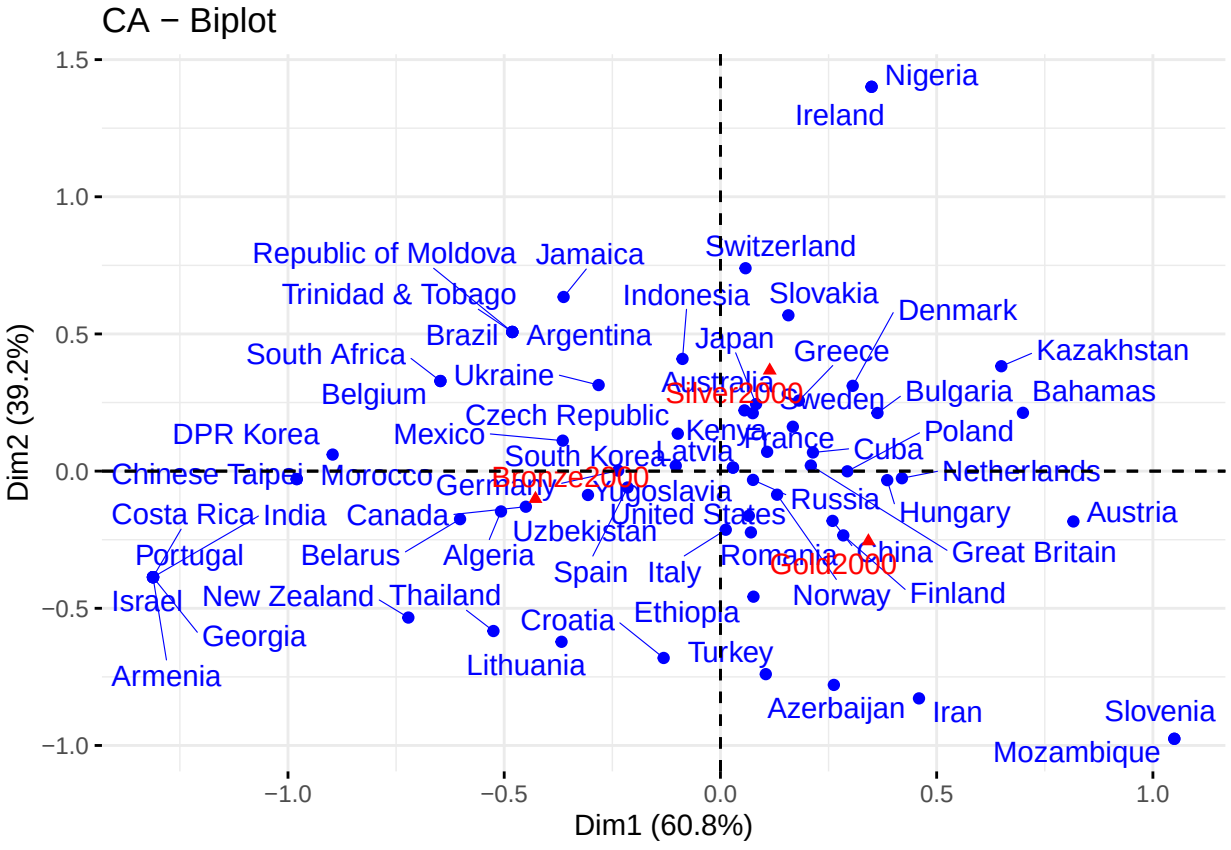
b) Create a data frame only consisting of the variables Gold, Silver, Bronze number of medals and the logarithm of the variables population, GDP and athletes.

```
eda_df <- olympic_df[, c("Gold2000", "Silver2000", "Bronze2000",
                         "Log.population", "Log.GDP", "Log.athletes")]
head(eda_df, n=3)
```

```
##               Gold2000 Silver2000 Bronze2000 Log.population  Log.GDP
## United States       39         25         33       12.52099 15.83932
## Russia              32         28         28       11.90114 12.78277
## China               28         16         15       14.04320 13.71688
##               Log.athletes
## United States     6.498282
## Russia            6.148468
## China             5.686975
```

2. Application of Correspondence Analysis (CA)

```
# Creation of a contingency table with only the medal counts
medals_df <- olympic_df[, c("Gold2000", "Silver2000", "Bronze2000")]
# Run Correspondence Analysis
ca_result <- CA(medals_df, graph = FALSE)
# Better visualization of the CA plot
fviz_ca_biplot(ca_result, repel = TRUE)
```



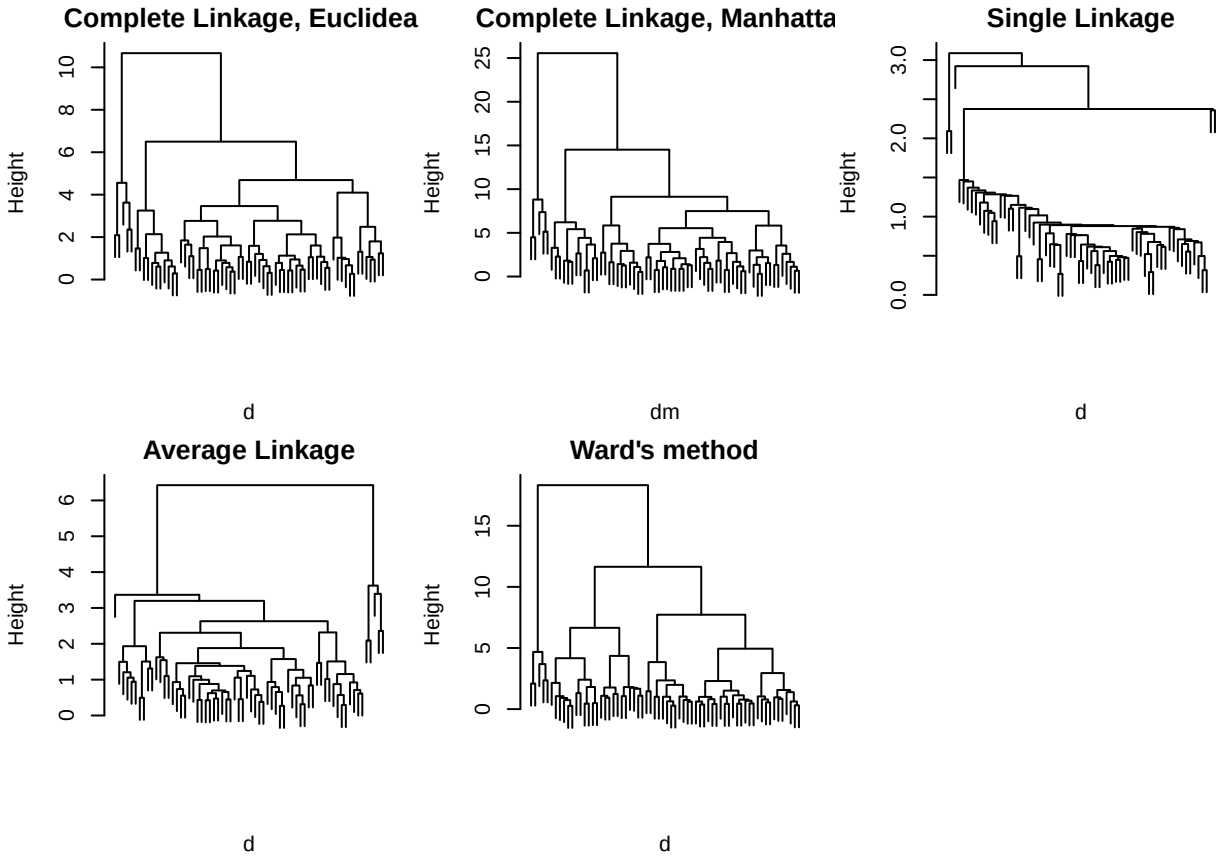
In the CA biplot, the **first dimension** clearly separates countries based on medal quality. Countries positioned on the right, near **Gold2000** and **Silver2000** (such as **United States, Russia, China, France**, etc...), are more strongly associated with higher-value medals. In contrast, countries on the left side, closer to **Bronze2000** (like **Spain, Canada, Morocco, Belarus**, etc...), are more aligned with bronze medals or generally lower medal performance. This suggests that **Dimension 1** captures the gradient from top-performing to lower-performing countries in terms of medal type.

Dimension 2 appears to capture the **dominant or exclusive type of medal** a country earned. Countries at the top of this axis, like **Ireland** and **Nigeria**, are characterized by **silver-only** performances. In contrast, those at the bottom, such as **Slovenia** and **Mozambique**, are distinguished by **gold-only** outcomes. Countries near the center, including many larger delegations, tend to have a **more balanced medal profile**, winning across multiple categories. Thus, **Dimension 2** doesn't just reflect diversity vs. concentration — it also reveals a **vertical spectrum of medal type dominance**, with silver-specialized nations at the top and gold-specialized ones at the bottom.

3. Hierarchical and Non-Hiearchical Clustering

d) Apply hierarchical clustering considering all the variables in the data frame constructed in section (b). Show the dendrogram that is constructed by the most interpretable method.Decide how many clusters to use.

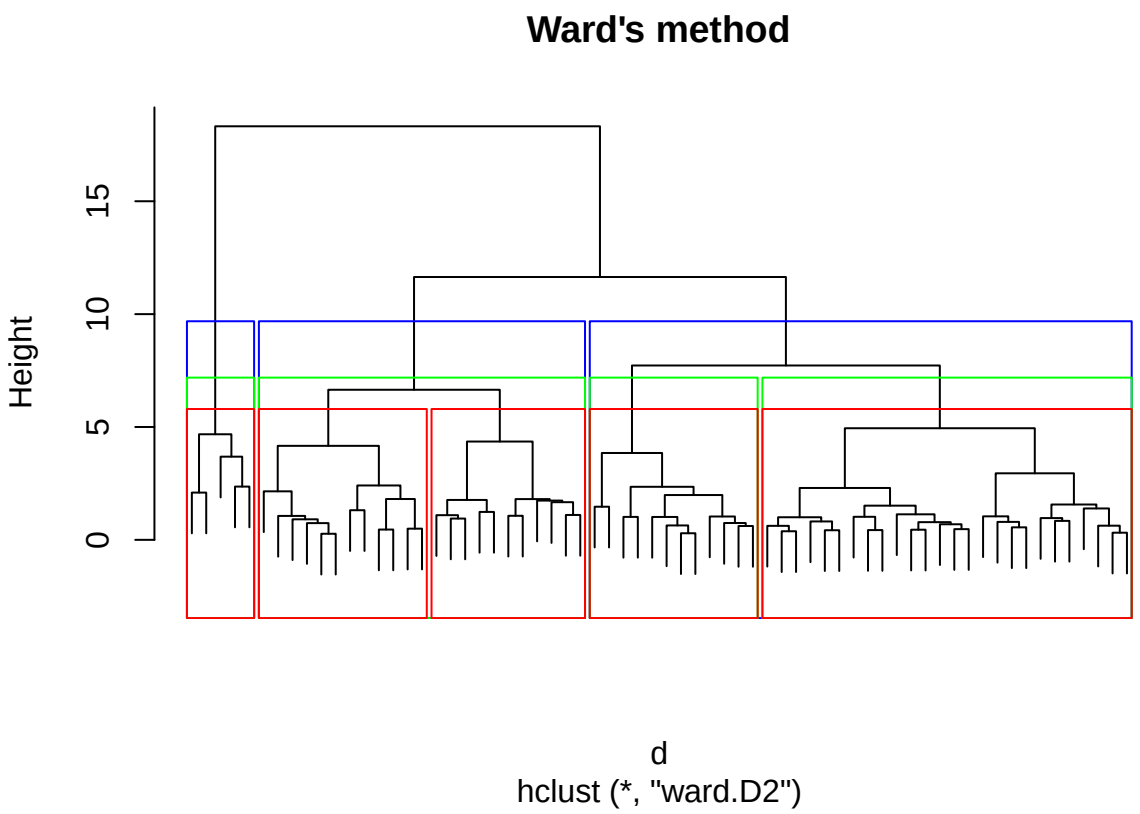
```
eda_df_s<-scale(eda_df)
d <- dist(eda_df_s, method = "euclidean") # Euclidean distance matrix
dm <- dist(eda_df_s, method = "manhattan")
```



The most interpretable method seems to be the Ward's method because all six variables are positively

correlated and Ward's criterion matches exactly that geometry (most clusters would be among one main axis - PC1) whereas criteria that are distance based are more sensitive to pairwise distances and because of that they separate the clusters more (single linkage).

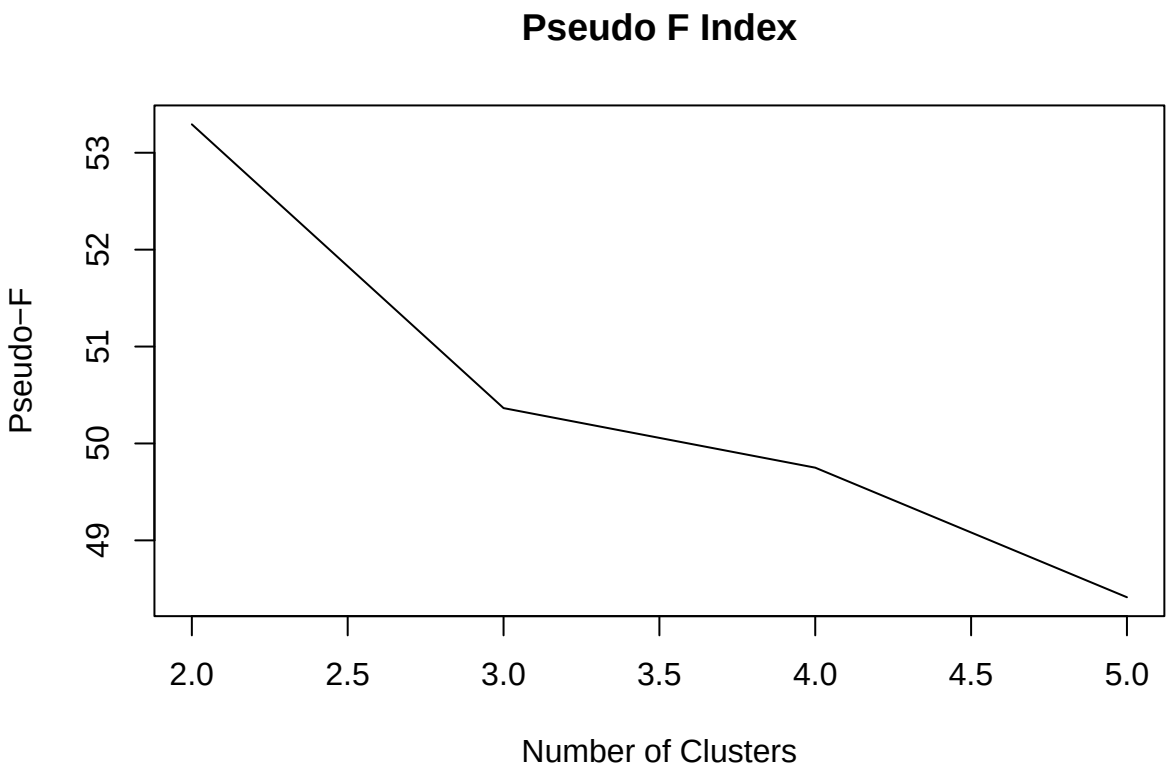
```
fit4 <- hclust(d, method="ward.D2")
```



Deciding on how many clusters we use is observing the dendrogram and looking for the largest "vertical jump" between increasing number of clusters. In other words, we look for biggest gap between two consecutive heights. We can examine that the largest just, besides k=2, is when k=3 dividing into 3 blue clusters. We would not choose k=2 as that is the too basic approach and commonly clusters are not separated into 2. So the largest jump besides that is when k=3 as we can see largest gap in that part so we can the dendrogram at k=3. By doing that we can get clusters with more interpretable structure than just dividing into two groups.

e) Use Pseudo-F index to decide number of clusters.

```
aux<-c()
for (i in 2:5){
  k<-kmeans(eda_df_s,centers=i,nstart=25)
  aux[i-1]<-((k$betweenss)*(nrow(eda_df)-i))/((k$tot.withinss)*(i-1))
}
plot(2:5,aux, xlab="Number of Clusters", ylab="Pseudo-F", type="l", main="Pseudo F Index")
```



The chosen number of clusters is 3. This also complies with cutting the dendrogram at k=3 height. the highest pseudo-F value is at k=2 but the drop in value is that large with k=3 and with three clusters we would get richer and interpretable structure as well.

f) Apply k-means clustering by taking k as the selected number of clusters.

```
k=3
fit <- kmeans(eda_df_s, k)
eda_df_kmeans <- data.frame(eda_df, fit$cluster)
```

```
##   Group.1  Gold2000 Silver2000 Bronze2000 Log.population   Log.GDP Log.athletes
## 1       1  1.105263   1.210526   1.631579       8.576638  9.013486     3.619149
## 2       2  3.050000   3.500000   3.600000      10.060652 11.890379     4.829433
## 3       3 22.285714  19.000000  20.428571      11.649313 14.002930     6.133389
##   fit.cluster
## 1           1
## 2           2
## 3           3
```

g) Interpret constructed clusters.

When interpreting these results, it has to be taken into account that some of these values are logarithms of actual values, so even if the numbers don't seem as different cluster-to-cluster, the actual differences are great.

The countries such as USA, Russia, China, Australia and Germany are countries with extremely high medal counts (around 20 in each category), large economies (logGDP $>$ 13.5), big populations and big athletic delegations. The exact countries in this cluster are the ones we usually see winning most medals at every

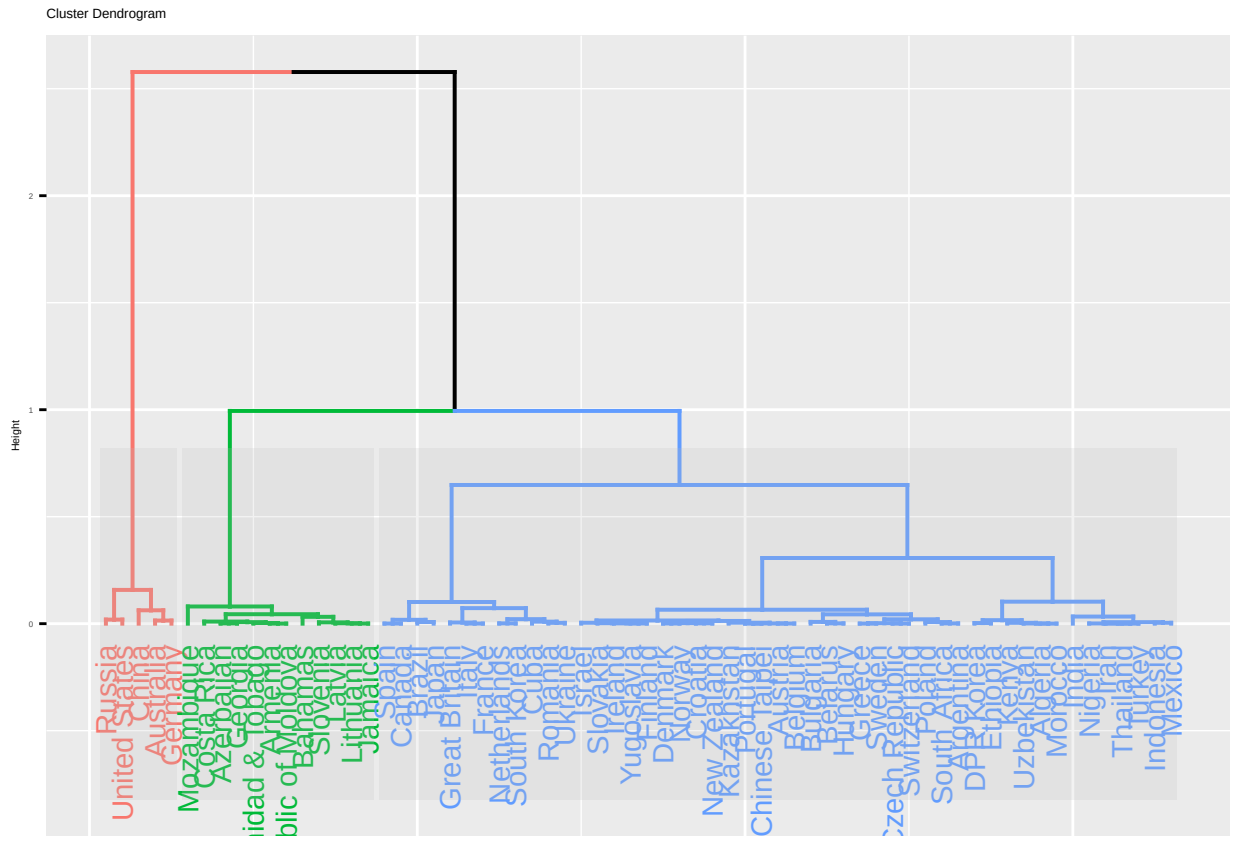
Olympics.

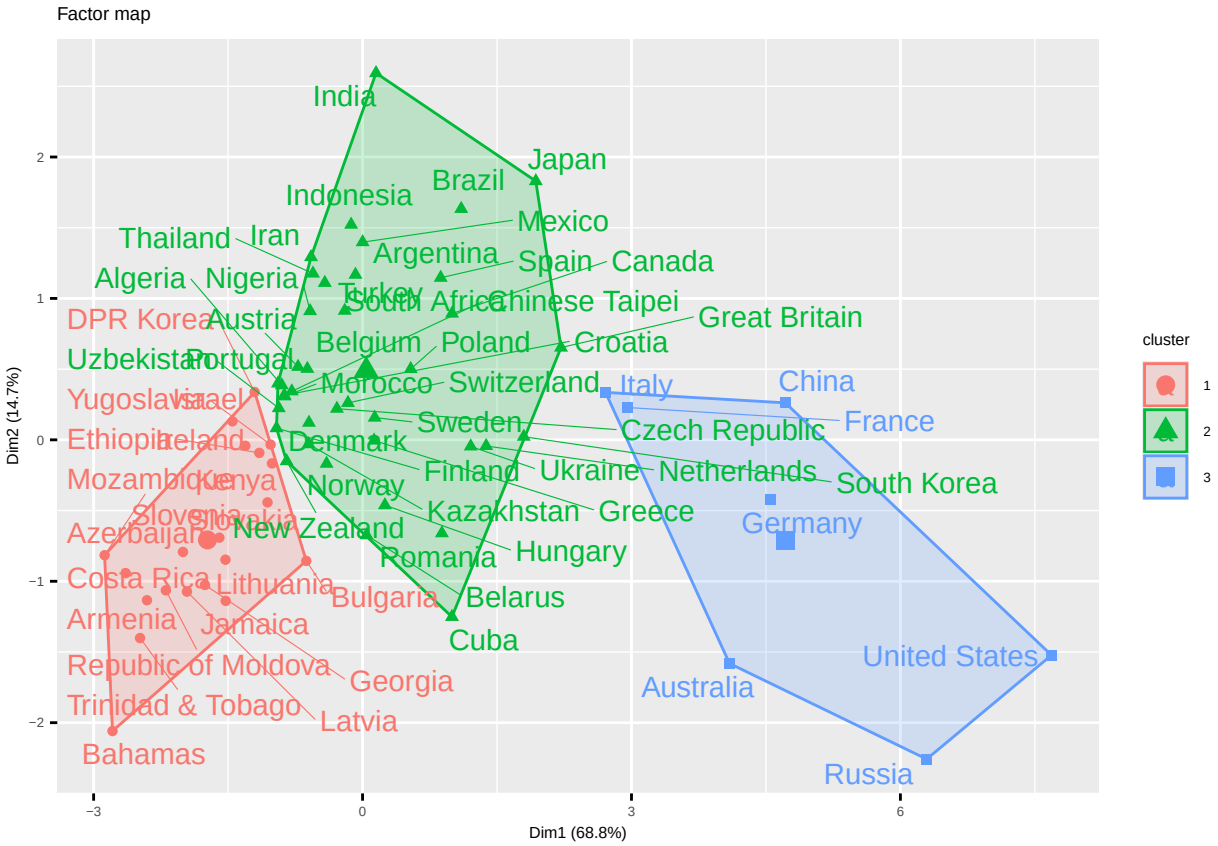
The countries such as Spain, Italy, Great Britain etc. are the countries with mid medal counts (around 5 in each category), as well as the populations and GDP (logGDP around 12). They also have smaller athletic delegations than countries in the first cluster. They still have good results, just not as good as the countries mentioned in the first cluster.

The countries in the last remaining cluster have lower gold, silver and bronze medal counts (around 1.5 in each category), lower average populations and lower average GDP (log-value around 10). Also, the number of athletes representing them at the Olympics is smaller than in other clusters. These countries are likely developing countries or those that don't invest as much in Olympic sports.

h) Apply hierarchical clustering on PCA scores by using HCPC() function in R. Interpret your findings.

```
pca_res <- PCA(eda_df, ncp = 3, graph = FALSE)
hcpc_res <- HCPC(pca_res, graph = FALSE)
```





Factor map interpretation

We can see PC1 component accumulates for nearly 70% of inertia and PC2 for around 15% which means calculating with three components is enough as they accumulate for nearly 80% of inertia.

On positive side of PC1 there are countries with high all medal counts with also high population, GDP and athletic delegations. Such countries are United States, Russia, China etc. Some countries on the opposite side of PC1 are Mozambique, Ireland, Israel as they won only one medal on Olympics and do not have big number of athletes representing them. This axis represent overall size and success.

High on the positive side of PC2 we can see India as they high population but they got small number of medals compared to their size. Japan and Brazil would also go into that category as they have big population and GDP compared to numbers of medals won. On the negative side of PC2 we can find over-achievers such as Jamaica and Cuba as they won a lot of medals compared to their relatively small population and GDP.

Cluster interpretation

Those PCA representations/scores translate also to the hierarchical clustering because it is based on PCA scores. After performing HCPC() with Ward's method dendrogram and and PC plot both show separation into 3 clusters.

The red, number 1, cluster is positioned mostly on the left side of factor map, specifically on the bottom left part where there are negative values of both PC1 and values of PC2 vary. Example of countries in that cluster are Kenya, Ethiopia, Latvia, Bulgaria, Costa Rica,... Those countries have small number of medals won as well as lower GDP and small population than the rest of countries. Most of them over-achieve as some of them won a lot of medals despite their small population and low GDP making "medal-efficiency" high.

The green, number 2, cluster is the biggest cluster. The center of the cluster is positioned very close to the

origin of factor map but most of the countries lay in first quadrant having positive PC1 and PC2 values. Some of the countries in the cluster are Great Britain, Spain, Brazil, Switzerland,... They have higher GDP and larger population than the ones in the red cluster with also higher medal count. It could be said those countries are emerging powers having solid number of medals won but also moderate GDP and population.

The blue, number 3, cluster clearly represent global powerhouses as they position themselves on the extreme right of factor map having high PC1 value and moderate PC2 value. Those countries have a lot of medals won especially USA with over 100 total medals won. Other countries that are in cluster are Russia, Australia, France, China, Italy,... When there is talk about sports including those countries all of them invest a lot into sports. One of the reasons they can do that is larger GDP and as they have larger population they statistically have more chance of having great athletes that can compete at the Olympics and win a medal.

i) Which of the above hierarchical clustering methods would you choose? Why?

Among the methods displayed in the analysis the most appropriate choice would probably be Ward's method with Euclidean distance dendrogram. It produces compact clusters as it minimises increase in variance within cluster. That improves the separation of clusters as countries that are in the same clusters have as much similar features as they can. Single linkage and complete/average linkage have flaws with either chaining all countries into one branch or having many singletons. We can also see that Ward's method is consistent with two other criteria, pseudo F-index and HCPC using PCA scores. They also show same cluster separation but are not that reliable as Ward's but they prove the results Ward's method gives.