Austin's Clustering

L. Austin Hadamuscin

4/2/2022

Packages

```
library(sampling)
library(tidyverse)
library(rworldmap)
library(mclust)
```

Data

```
dat <- read.csv("C:/Users/hadamul/OneDrive/Graduate school/Semester 4/Data Mining/Data-Mining-Final-Pro
```

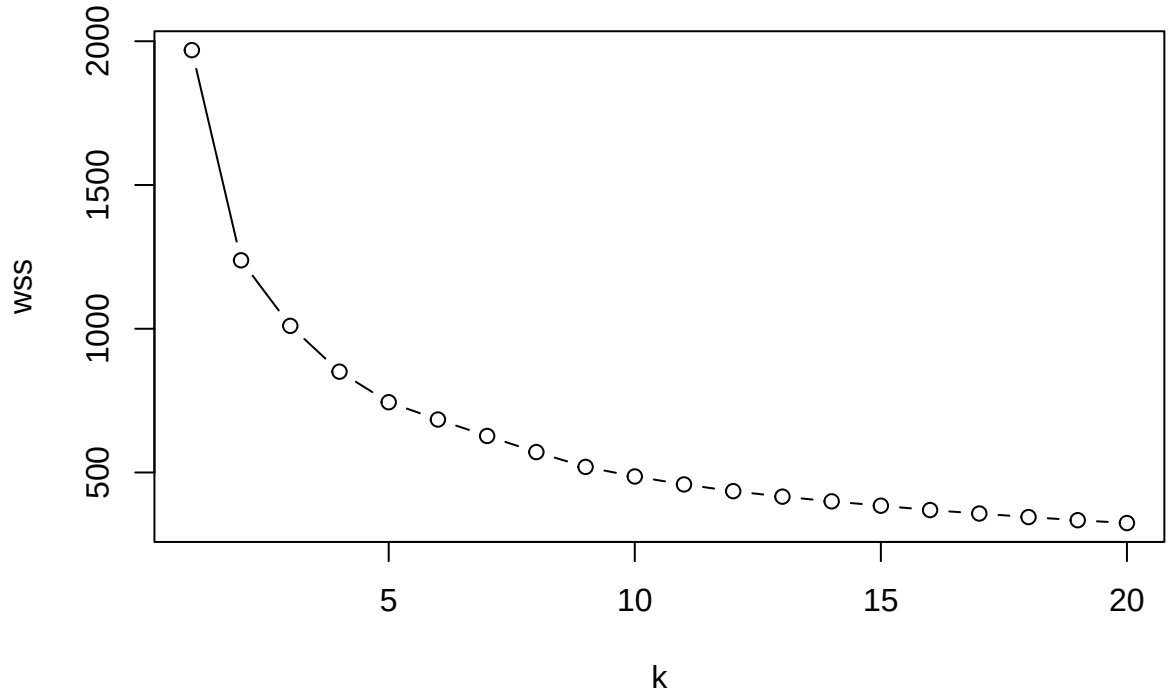
Data Scaling

```
 dat_scaled <- scale(select(dat, -Country, -HALE_Birth))
```

k-means

```
RNGkind (sample.kind = "Rounding")
set.seed(0)

nseeds <- 1000
nk <- 20

seeds <- ceiling(runif(nseeds,0,900000))

# kmean_matrix  <- matrix(NA, nrow = nseeds, ncol = nk)
#
# for (k in 1:nk) {
#   seed_iter = 0
#   for (s in seeds) {
#
#     set.seed(s)
#     seed_iter <- seed_iter + 1
```

```
#
#     kmean_matrix[seed_iter,k] <- kmeans(dat_scaled, centers = k)$tot.withinss
#   }
# }
# save(kmean_matrix, file = "kmean_matrix.RData")

load("kmean_matrix.RData")
```



- 4 or 5, we will see

Finding the seed for the 2- & 6-means

```
kmean_seed2 <- seeds[which(kmean_matrix[,2]==min(kmean_matrix[,2]))[1]]
kmean_seed3 <- seeds[which(kmean_matrix[,3]==min(kmean_matrix[,3]))[1]]
kmean_seed4 <- seeds[which(kmean_matrix[,4]==min(kmean_matrix[,4]))[1]]
```

Doing the 4/5-means

```
set.seed(kmean_seed2)
means2 <- kmeans(dat_scaled, centers = 2)
set.seed(kmean_seed3)
means3 <- kmeans(dat_scaled, centers = 3)
```

```
set.seed(kmean_seed4)
means4 <- kmeans(dat_scaled, centers = 4)
```

Rescaling the cluster centers

```
means2.centers.rescaled <-  means2$centers
for (j in 1:11){
  means2.centers.rescaled[,j] <- attributes(dat_scaled)$`scaled:center`[j] +
    means2$centers[,j]*attributes(dat_scaled)$`scaled:scale`[j]
}

means3.centers.rescaled <-  means3$centers
for (j in 1:11){
  means3.centers.rescaled[,j] <- attributes(dat_scaled)$`scaled:center`[j] +
    means3$centers[,j]*attributes(dat_scaled)$`scaled:scale`[j]
}

means4.centers.rescaled <-  means4$centers
for (j in 1:11){
  means4.centers.rescaled[,j] <- attributes(dat_scaled)$`scaled:center`[j] +
    means4$centers[,j]*attributes(dat_scaled)$`scaled:scale`[j]
}
```

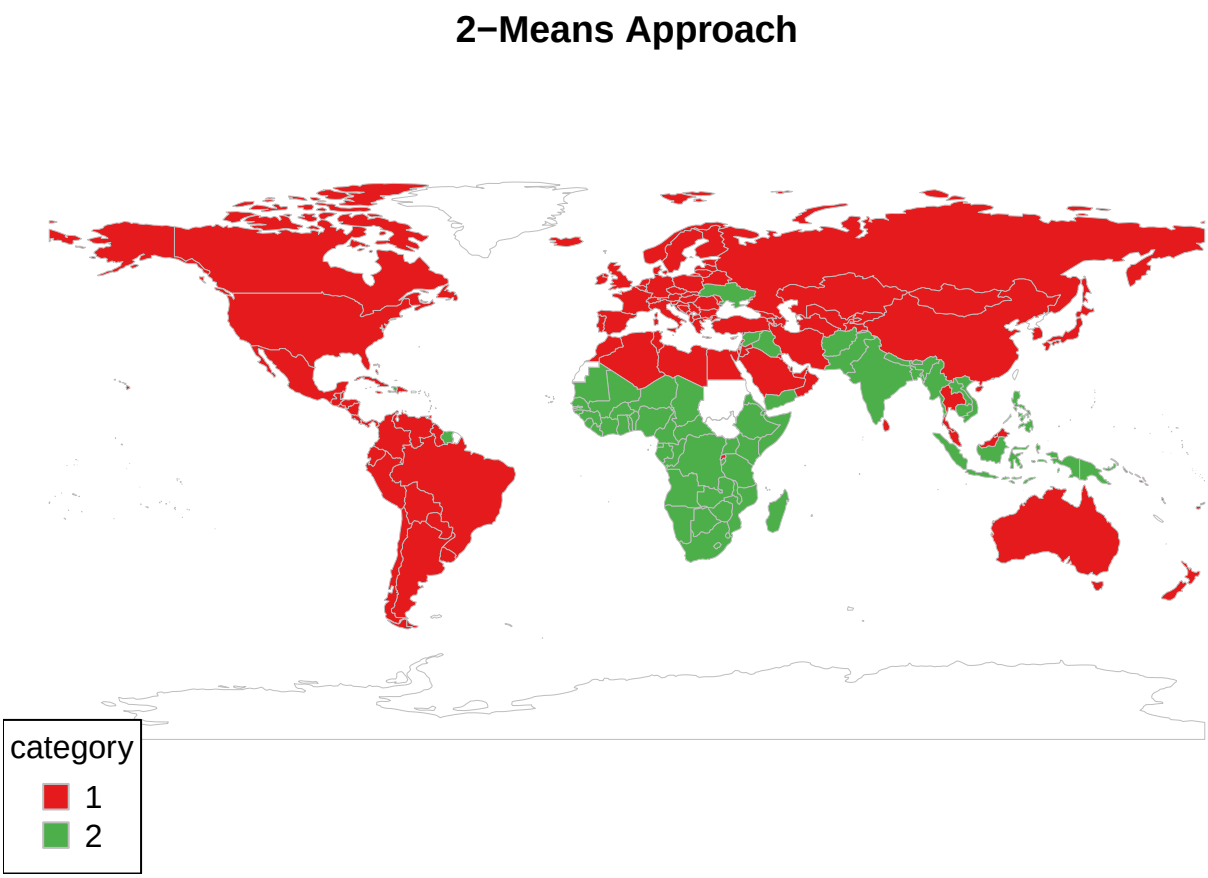
Rearranging cluster groups

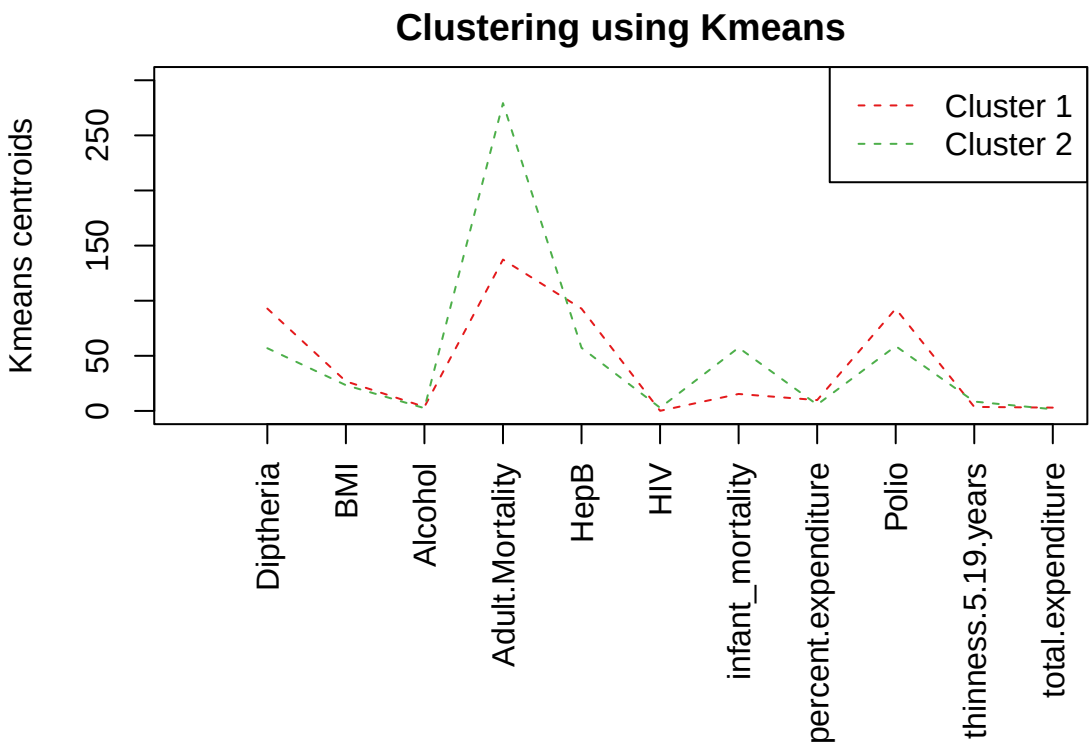
```
km3id <- ifelse(means3$cluster == 1, 3, ifelse(means3$cluster == 2, 2, 1))
km4id <- ifelse(means4$cluster == 1, 2,
                ifelse(means4$cluster == 2, 3,
                       ifelse(means4$cluster == 3, 1, 4)))
```

Grouping data into clusters

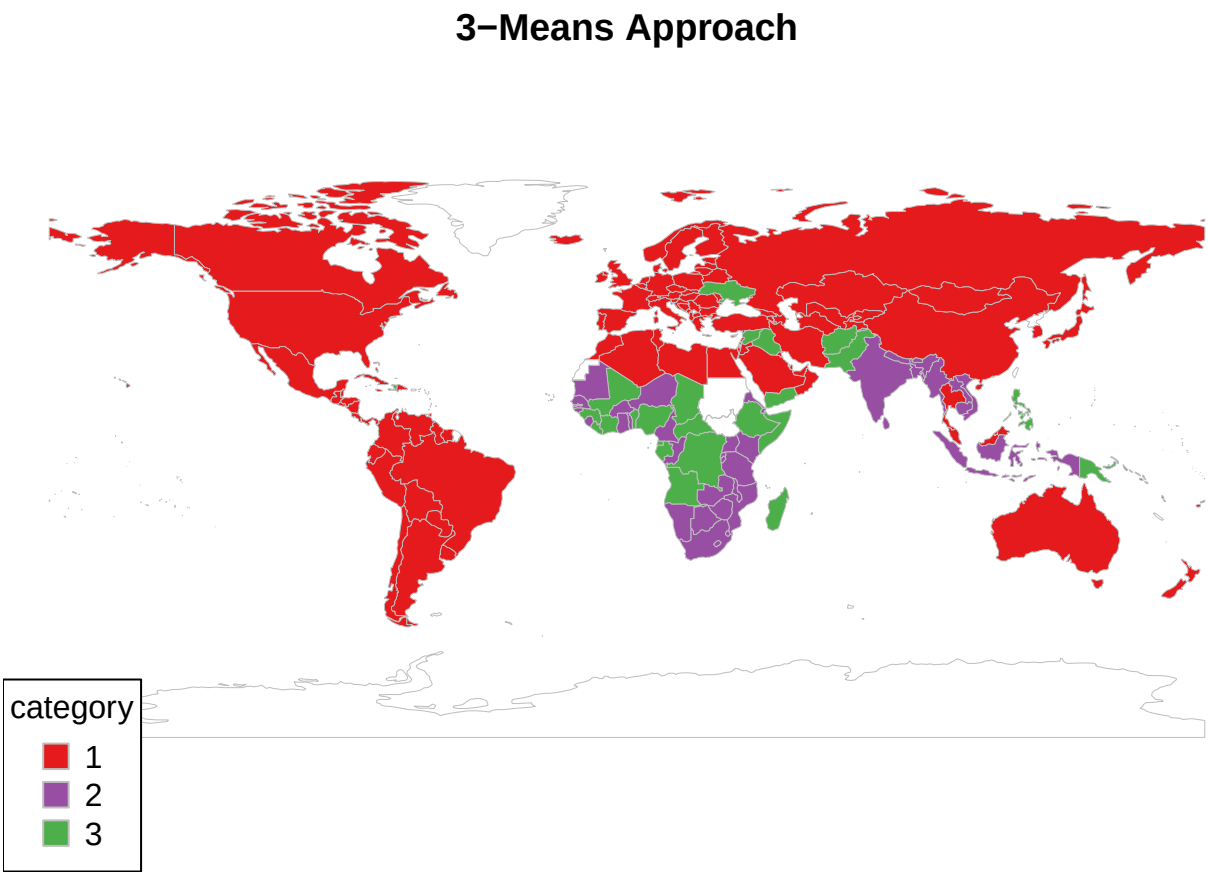
```
clust2 <- cbind(select(dat, Country, HALE_Birth),cluster = means2$cluster,dat)
clust3 <- cbind(select(dat, Country, HALE_Birth),cluster = as.factor(km3id),dat)
clust4 <- cbind(select(dat, Country, HALE_Birth),cluster = as.factor(km4id),dat)
```

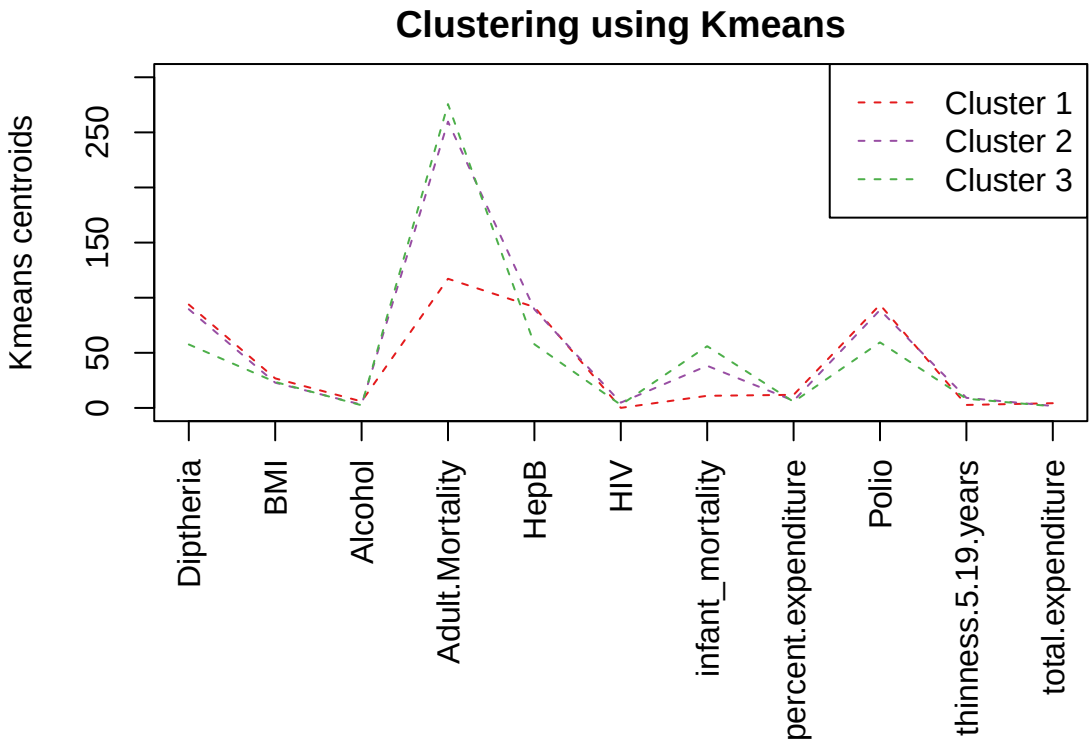
2 Cluster Graph



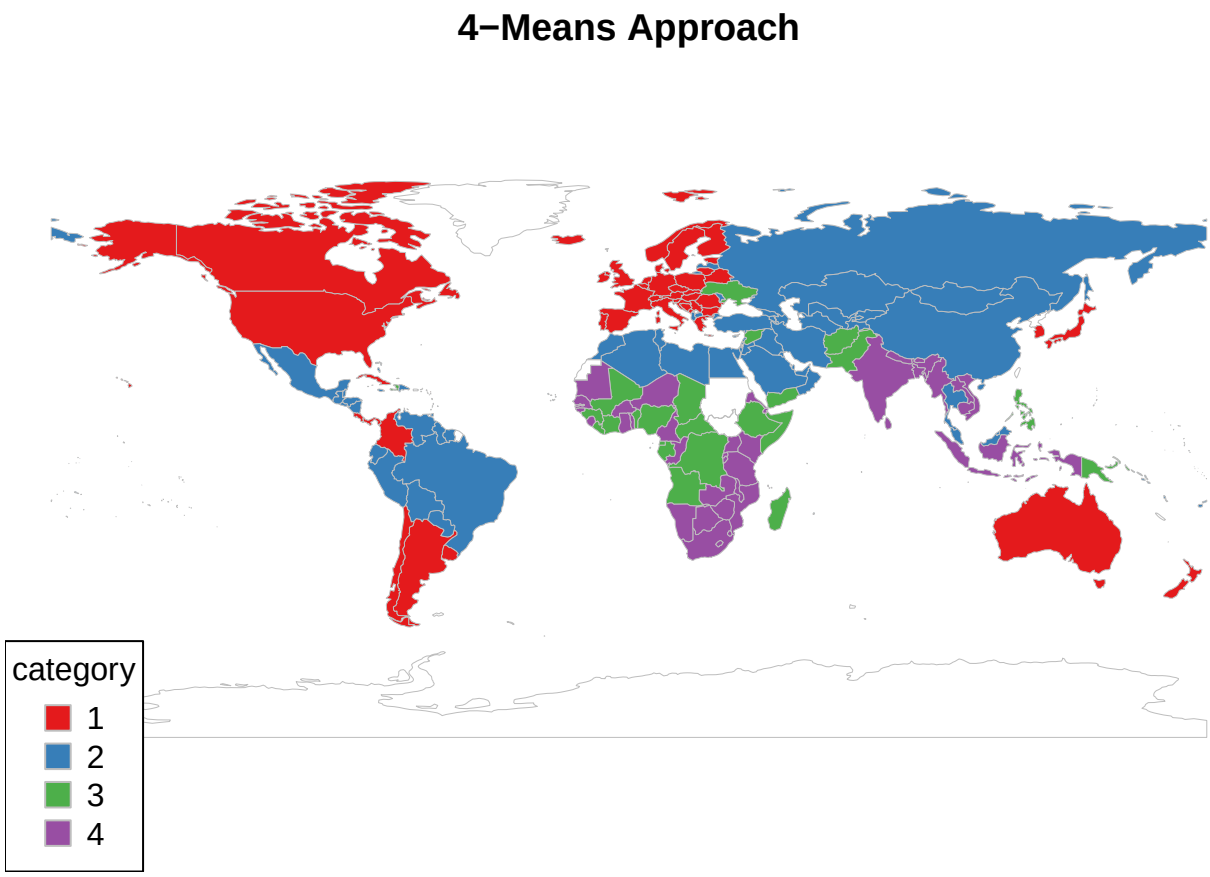

Clustering using Kmeans
Kmeans centroids
0
50
150
250
Cluster 1
Cluster 2
Diptheria
BMI
Alcohol
Adult.Mortality
HepB
HIV
infant_mortality
percent.expenditure
Polio
thinness.5.19.years
total.expenditure

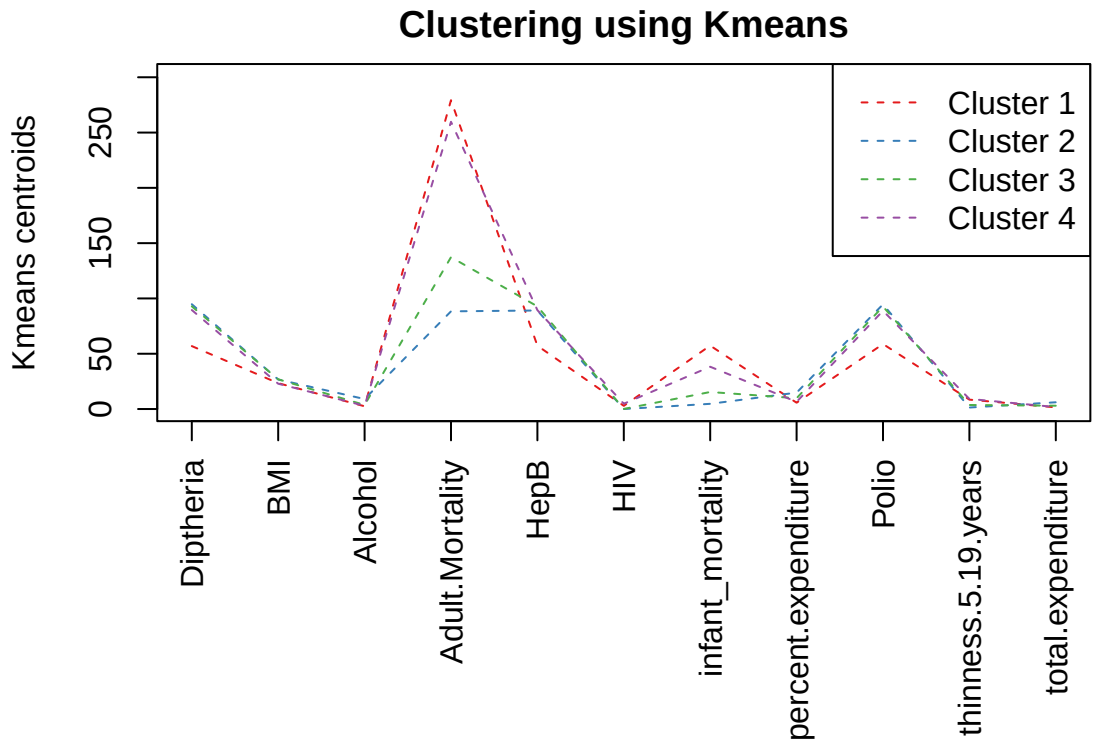
3 Cluster Graph




Clustering using Kmeans
Kmeans centroids
0
50
150
250
Cluster 1
Cluster 2
Cluster 3
Diptheria
BMI
Alcohol
Adult.Mortality
HepB
HIV
infant_mortality
percent.expenditure
Polio
thinness.5.19.years
total.expenditure

4 Cluster Graph

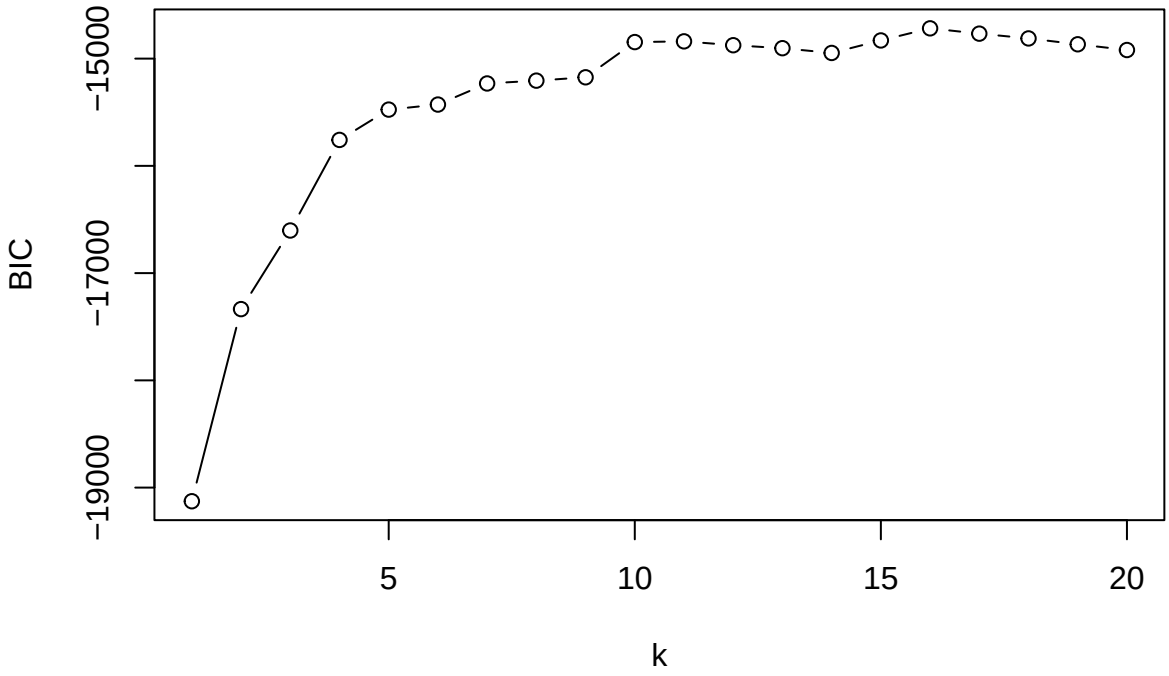




Model Based

```
mb_rep <- 20
mb_matrix <- matrix(NA, 14, mb_rep)

mb_dat <- select(dat, -Country, -HALE_Birth)

for (k in 1:mb_rep) {
  mb_matrix[,k] <- Mclust(mb_dat, k)$BIC
}

# save(mb_matrix,file = "mb_matrix.RData")
#
#
# load("mb_matrix.RData")
```

```
plot(x=1:20,
     y=apply(mb_matrix, 2, function(i) min(i,na.rm = T)),
     type = "b",
     xlab = "k",
     ylab = "BIC")
```



- Either 3 or 4 looks like the best k.

Doing the model clustering with 3 and 4 clusters

```
# mb3 <- Mclust(mb_dat, 3, verbose = F)
# mb4 <- Mclust(mb_dat, 4, verbose = F)
#
# save(mb3, file = "mb3.RData")
# save(mb4, file = "mb4.RData")

load("mb3.RData")
load("mb4.RData")
#
# round(mb3$parameters$mean)
```

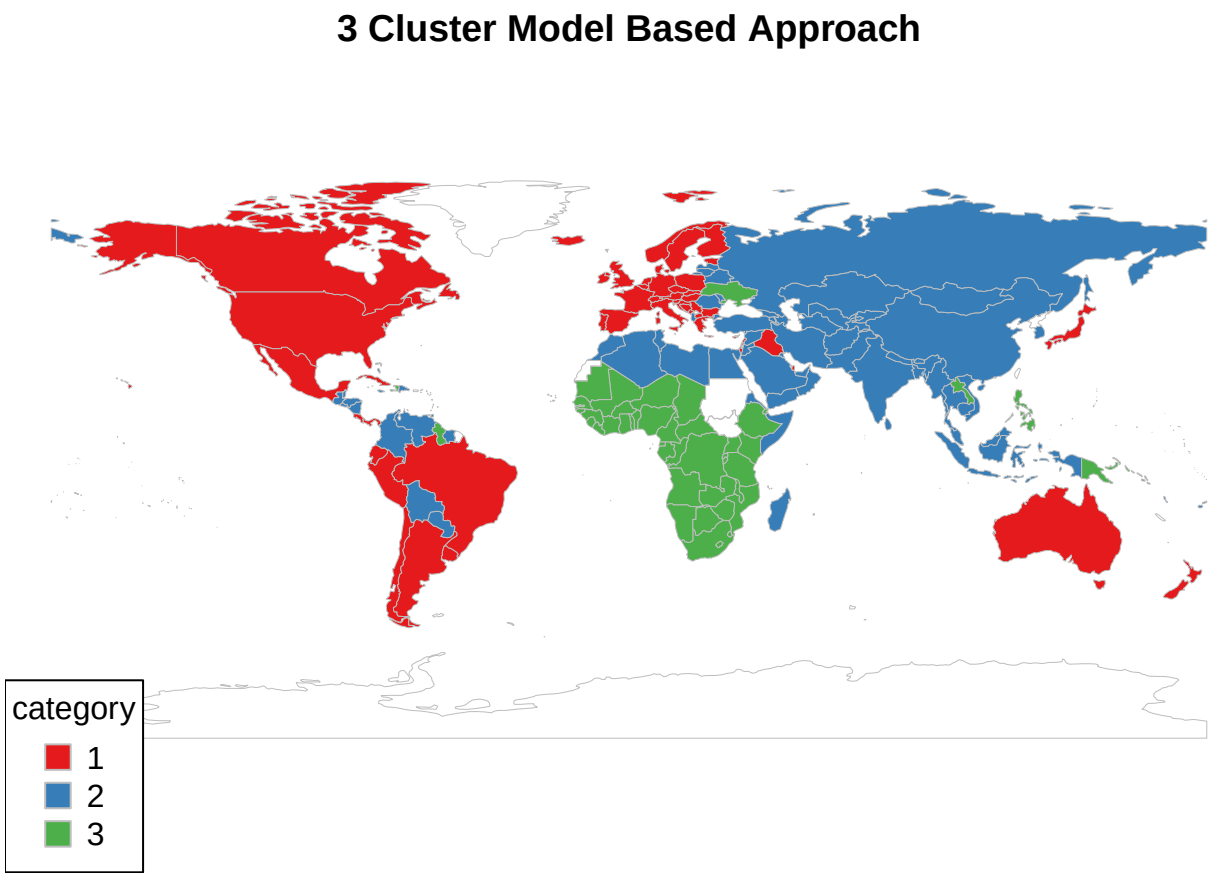
Rearranging the clusters

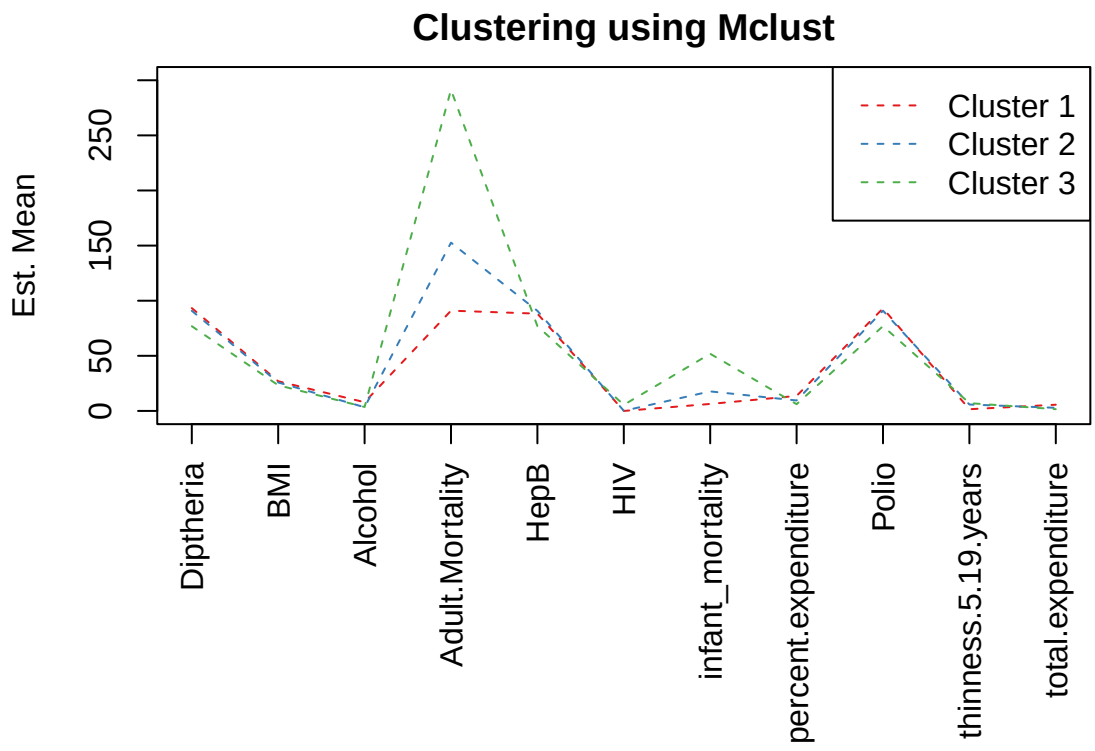
```
mb3id <- ifelse(mb3$classification == 3, 1,
                ifelse(mb3$classification == 2, 2,3))
mb4id <- ifelse(mb4$classification == 3, 1,
                ifelse(mb4$classification == 2, 2,
                       ifelse(mb4$classification == 1, 4, 3)))
```

Grouping data into clusters

```
mbclust3 <- cbind(select(dat, Country, HALE_Birth),cluster = as.factor(mb3id),dat)
mbclust4 <- cbind(select(dat, Country, HALE_Birth),cluster = as.factor(mb4id),dat)
```

Clust 3 Graph

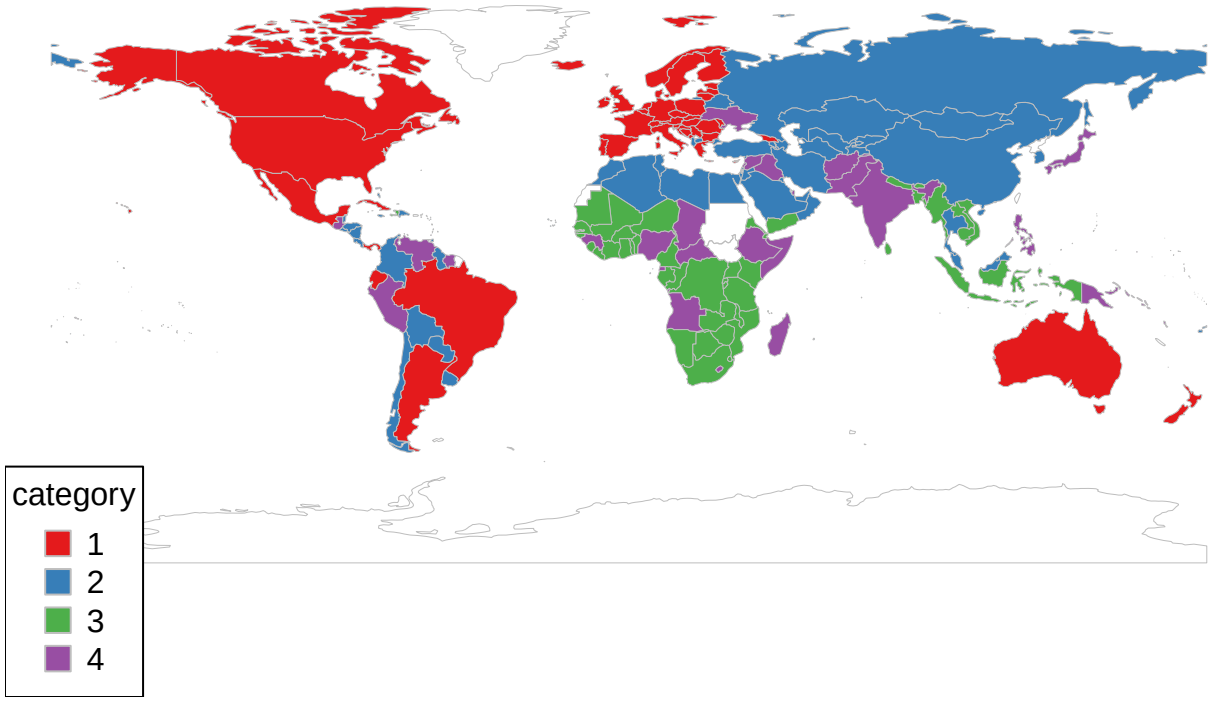


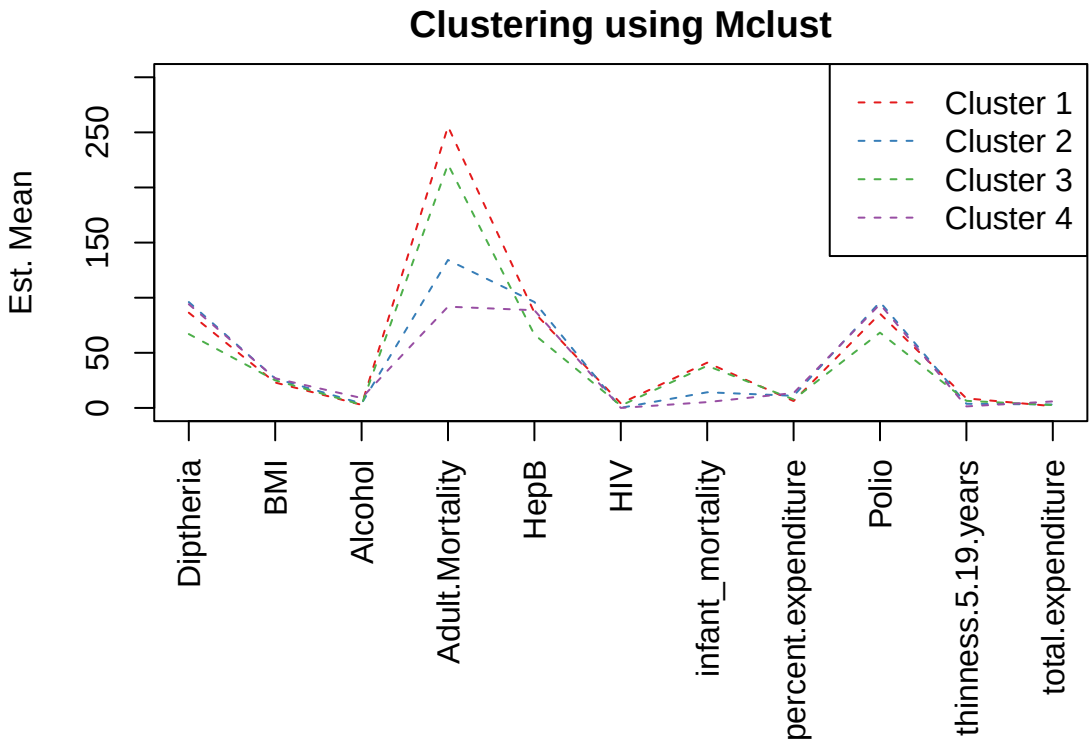


Clust 4 Graph

```
par(mar=c(0,0,1,0))
mapCountryData(mapToPlot = mbclust4_map,
               nameColumnToPlot="cluster",
               catMethod="categorical",
               colourPalette = RColorBrewer::brewer.pal(4,'Set1'),
               mapTitle = "4 Cluster Model Based Approach")
```

4 Cluster Model Based Approach




Clustering using Mclust
Est. Mean
0
50
150
250
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Diptheria
BMI
Alcohol
Adult.Mortality
HepB
HIV
infant_mortality
percent.expenditure
Polio
thinness.5.19.years
total.expenditure