TeamD\_Case3

Shine Yao, Tianyu Zhang, Lingyi Kong, Jason Meng

3/20/2018

#load the package  
library(GGally)  
library(factoextra)

## Loading required package: ggplot2

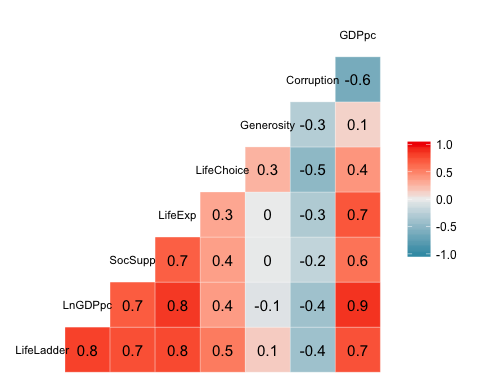
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ

library(ggplot2)

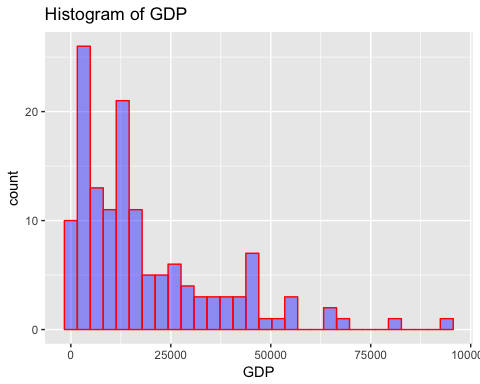
## Part I: Data Pre-processing

## country LifeLadder LnGDPpc SocSupp   
## Afghanistan: 1 Min. :2.693 Min. : 6.633 Min. :0.2902   
## Albania : 1 1st Qu.:4.521 1st Qu.: 8.462 1st Qu.:0.7535   
## Algeria : 1 Median :5.431 Median : 9.486 Median :0.8390   
## Argentina : 1 Mean :5.400 Mean : 9.284 Mean :0.8120   
## Armenia : 1 3rd Qu.:6.140 3rd Qu.:10.220 3rd Qu.:0.9109   
## Australia : 1 Max. :7.660 Max. :11.459 Max. :0.9849   
## (Other) :135 NA's :8   
## LifeExp LifeChoice Generosity Corruption   
## Min. :43.38 Min. :0.3035 Min. :-0.273875 Min. :0.04731   
## 1st Qu.:57.08 1st Qu.:0.6967 1st Qu.:-0.101282 1st Qu.:0.70288   
## Median :64.79 Median :0.7776 Median :-0.020263 Median :0.81052   
## Mean :62.88 Mean :0.7637 Mean :-0.007645 Mean :0.74587   
## 3rd Qu.:68.92 3rd Qu.:0.8650 3rd Qu.: 0.081898 3rd Qu.:0.86592   
## Max. :76.41 Max. :0.9838 Max. : 0.485928 Max. :0.96948   
## NA's :1 NA's :3 NA's :11 NA's :12   
## GDPpc   
## Min. : 759.7   
## 1st Qu.: 4730.6   
## Median :13178.1   
## Mean :18950.2   
## 3rd Qu.:27453.5   
## Max. :94773.8   
## NA's :8

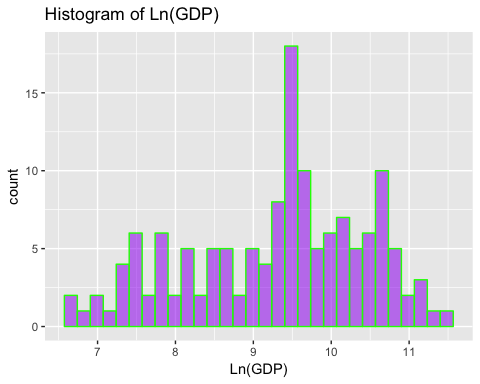
## country LifeLadder LnGDPpc SocSupp LifeExp LifeChoice   
## 0 0 8 0 1 3   
## Generosity Corruption GDPpc   
## 11 12 8



## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

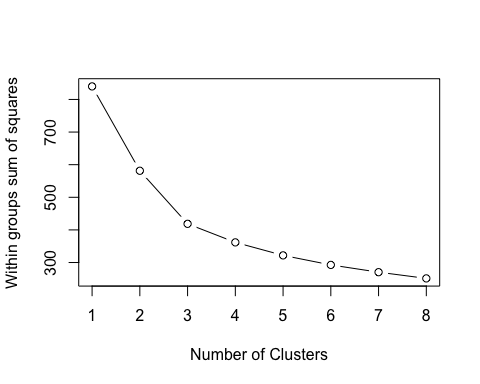


## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 ### Interpretation of Part I: \* After reading the csv file, our group used the summary function to get a basic understanding of the content and structure of the file \* Missing data can be a serious problem if we want to get access to the characteristics of our data and cluster them.Therefore, our group checked the missing value in the dataset by examing the correlation of each variables and replaced them with median value of the dataset. The reason why we used median number is that the median number won’t be affected by extreme values in the dataset. \* Then we visualized the correlation between variables in order to eliminate highly-correlated pairs, if existed. As we can see from the output, GDPpc and LnGDPpc are highly correlated, as they basically represent the same variable. So we need to drop one of them for further analysis. \* In order to decide which one between GDPpc and LnGDPpc should be dropped, our group created the histogram of those two variables to check their distributions. From the output it can be found that Ln(GDP) is less skewwed than GDP data and more conforms with normal distribution. So we keep the Ln(GDP) and drop the GDP column from our dataset.

## Part II: Models without LifeLadder

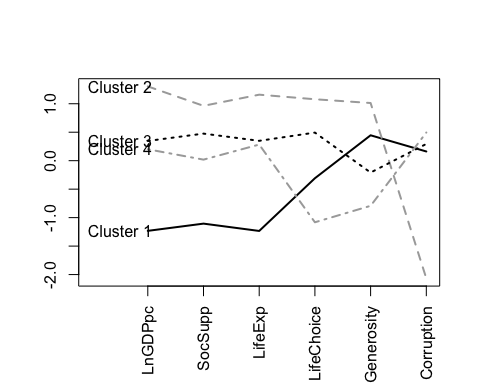
# Normalized the data  
whr\_2017\_DF.scaled<- scale(whr\_2017\_DF[,-1])  
  
#set row names  
row.names(whr\_2017\_DF.scaled) <-whr\_2017\_DF$country  
  
#build k-means model  
par(mfrow = c(1, 1))  
  
# Initialize total within sum of squares error: wss  
wss <- 0  
  
# Look over 3 to 8 possible clusters  
for (i in 1:8) {  
 # Fit the model: km.out  
 km.out <- kmeans(whr\_2017\_DF.scaled, centers = i, nstart = 20, iter.max = 50)  
 # Save the within cluster sum of squares  
 wss[i] <- km.out$tot.withinss  
}  
  
# Produce a scree plot  
plot(1:8, wss, type = "b",   
 xlab = "Number of Clusters",   
 ylab = "Within groups sum of squares")



# Select number of clusters  
k <- 4  
  
# Build model with k clusters: km.out  
km.out <- kmeans(whr\_2017\_DF.scaled, centers = k, nstart = 50, iter.max = 50)  
  
# View the resulting model  
km.out

## K-means clustering with 4 clusters of sizes 39, 18, 53, 31  
##   
## Cluster means:  
## LnGDPpc SocSupp LifeExp LifeChoice Generosity Corruption  
## 1 -1.2322710 -1.10523612 -1.2340219 -0.3085164 0.4459787 0.1616182  
## 2 1.3044912 0.96333139 1.1591966 1.0787345 1.0127625 -2.0673520  
## 3 0.3464848 0.47540003 0.3487467 0.4933891 -0.2084113 0.2940510  
## 4 0.2004527 0.01832393 0.2831528 -1.0817647 -0.7928095 0.4943394  
##   
## Clustering vector:  
## Afghanistan Albania   
## 1 4   
## Algeria Argentina   
## 3 3   
## Armenia Australia   
## 4 2   
## Austria Azerbaijan   
## 2 4   
## Bahrain Bangladesh   
## 3 1   
## Belarus Belgium   
## 4 2   
## Benin Bolivia   
## 1 3   
## Bosnia and Herzegovina Botswana   
## 4 3   
## Brazil Bulgaria   
## 3 4   
## Burkina Faso Cambodia   
## 1 1   
## Cameroon Canada   
## 1 2   
## Central African Republic Chad   
## 1 1   
## Chile China   
## 4 4   
## Colombia Congo (Brazzaville)   
## 3 1   
## Congo (Kinshasa) Costa Rica   
## 1 3   
## Cyprus Czech Republic   
## 3 3   
## Denmark Dominican Republic   
## 2 3   
## Ecuador Egypt   
## 3 4   
## El Salvador Estonia   
## 3 3   
## Ethiopia Finland   
## 1 2   
## France Gabon   
## 3 4   
## Georgia Germany   
## 4 2   
## Ghana Greece   
## 1 4   
## Guatemala Guinea   
## 3 1   
## Haiti Honduras   
## 1 3   
## Hong Kong Hungary   
## 2 4   
## Iceland India   
## 2 1   
## Indonesia Iran   
## 1 1   
## Iraq Ireland   
## 4 2   
## Israel Italy   
## 3 4   
## Ivory Coast Japan   
## 1 3   
## Jordan Kazakhstan   
## 3 3   
## Kenya Kosovo   
## 1 3   
## Kuwait Kyrgyzstan   
## 3 3   
## Latvia Lebanon   
## 4 4   
## Lesotho Liberia   
## 1 1   
## Libya Lithuania   
## 3 4   
## Luxembourg Macedonia   
## 2 3   
## Madagascar Malawi   
## 1 1   
## Mali Malta   
## 1 3   
## Mauritania Mauritius   
## 4 3   
## Mexico Moldova   
## 3 4   
## Mongolia Montenegro   
## 3 4   
## Morocco Myanmar   
## 4 3   
## Nepal Netherlands   
## 1 2   
## New Zealand Nicaragua   
## 2 3   
## Niger Nigeria   
## 1 1   
## Turkish Republic of Northern Cyprus Norway   
## 3 2   
## Pakistan Palestine   
## 1 4   
## Panama Paraguay   
## 3 3   
## Peru Philippines   
## 3 3   
## Poland Portugal   
## 3 3   
## Romania Russia   
## 3 4   
## Rwanda Saudi Arabia   
## 1 3   
## Senegal Serbia   
## 1 4   
## Sierra Leone Singapore   
## 1 2   
## Slovakia Slovenia   
## 3 3   
## Somalia South Africa   
## 1 3   
## South Korea South Sudan   
## 4 1   
## Spain Sweden   
## 3 2   
## Switzerland Taiwan   
## 2 3   
## Tajikistan Tanzania   
## 1 1   
## Thailand Togo   
## 3 1   
## Tunisia Turkey   
## 4 4   
## Turkmenistan Uganda   
## 3 1   
## Ukraine United Arab Emirates   
## 4 3   
## United Kingdom United States   
## 2 3   
## Uruguay Uzbekistan   
## 3 3   
## Venezuela Vietnam   
## 4 3   
## Yemen Zambia   
## 4 1   
## Zimbabwe   
## 1   
##   
## Within cluster sum of squares by cluster:  
## [1] 156.39218 23.99144 103.28379 78.07993  
## (between\_SS / total\_SS = 56.9 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

#plotting profile plot of centroids   
# plot an empty scatter plot  
plot(c(0), xaxt = 'n', ylab = "", xlab ="", type = "l",  
 ylim = c(min(km.out$centers), max(km.out$centers)), xlim = c(0, 6))  
  
# label x-axes  
axis(1, at = c(1:6), labels = colnames(whr\_2017\_DF.scaled), las = 2)  
  
# plot centroids  
for (i in c(1:4))  
lines(km.out$centers[i,], lty = i, lwd = 2, col = ifelse(i %in% c(1, 3, 5),  
 "black", "dark grey"))  
#name clusters  
text(x = 0.5, y = km.out$centers[, 1], labels = paste("Cluster", c(1:4)))



# Create hierarchical clustering model: hclust.out  
hclust.out <- hclust(dist(whr\_2017\_DF.scaled), method="complete")  
  
# Inspect the result  
summary(hclust.out)

## Length Class Mode   
## merge 280 -none- numeric   
## height 140 -none- numeric   
## order 141 -none- numeric   
## labels 141 -none- character  
## method 1 -none- character  
## call 3 -none- call   
## dist.method 1 -none- character

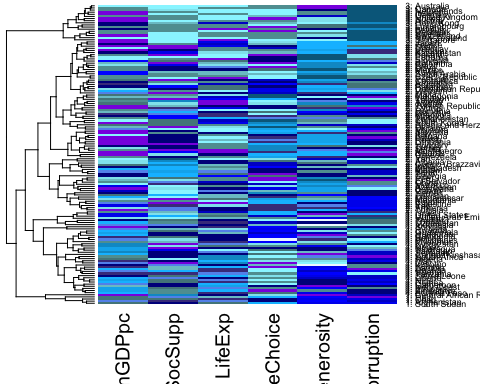
#prune the tree  
hclust.out\_cut <- cutree(hclust.out, k = 4)  
  
#inspect the clusters  
hclust.out\_cut

## Afghanistan Albania   
## 1 2   
## Algeria Argentina   
## 2 2   
## Armenia Australia   
## 2 3   
## Austria Azerbaijan   
## 3 2   
## Bahrain Bangladesh   
## 2 2   
## Belarus Belgium   
## 2 3   
## Benin Bolivia   
## 2 4   
## Bosnia and Herzegovina Botswana   
## 2 2   
## Brazil Bulgaria   
## 2 2   
## Burkina Faso Cambodia   
## 4 4   
## Cameroon Canada   
## 4 3   
## Central African Republic Chad   
## 1 1   
## Chile China   
## 2 2   
## Colombia Congo (Brazzaville)   
## 2 2   
## Congo (Kinshasa) Costa Rica   
## 4 2   
## Cyprus Czech Republic   
## 2 2   
## Denmark Dominican Republic   
## 3 2   
## Ecuador Egypt   
## 2 2   
## El Salvador Estonia   
## 2 2   
## Ethiopia Finland   
## 4 3   
## France Gabon   
## 2 2   
## Georgia Germany   
## 2 3   
## Ghana Greece   
## 4 2   
## Guatemala Guinea   
## 4 4   
## Haiti Honduras   
## 1 4   
## Hong Kong Hungary   
## 3 2   
## Iceland India   
## 4 2   
## Indonesia Iran   
## 4 2   
## Iraq Ireland   
## 2 3   
## Israel Italy   
## 4 2   
## Ivory Coast Japan   
## 4 2   
## Jordan Kazakhstan   
## 2 2   
## Kenya Kosovo   
## 4 2   
## Kuwait Kyrgyzstan   
## 2 4   
## Latvia Lebanon   
## 2 2   
## Lesotho Liberia   
## 4 4   
## Libya Lithuania   
## 2 2   
## Luxembourg Macedonia   
## 3 2   
## Madagascar Malawi   
## 2 2   
## Mali Malta   
## 4 2   
## Mauritania Mauritius   
## 2 2   
## Mexico Moldova   
## 2 2   
## Mongolia Montenegro   
## 2 2   
## Morocco Myanmar   
## 2 4   
## Nepal Netherlands   
## 4 3   
## New Zealand Nicaragua   
## 3 4   
## Niger Nigeria   
## 4 4   
## Turkish Republic of Northern Cyprus Norway   
## 2 3   
## Pakistan Palestine   
## 2 2   
## Panama Paraguay   
## 2 2   
## Peru Philippines   
## 2 4   
## Poland Portugal   
## 2 2   
## Romania Russia   
## 2 2   
## Rwanda Saudi Arabia   
## 4 2   
## Senegal Serbia   
## 4 2   
## Sierra Leone Singapore   
## 4 3   
## Slovakia Slovenia   
## 2 2   
## Somalia South Africa   
## 4 4   
## South Korea South Sudan   
## 2 1   
## Spain Sweden   
## 2 3   
## Switzerland Taiwan   
## 3 2   
## Tajikistan Tanzania   
## 4 4   
## Thailand Togo   
## 4 2   
## Tunisia Turkey   
## 2 2   
## Turkmenistan Uganda   
## 2 4   
## Ukraine United Arab Emirates   
## 2 4   
## United Kingdom United States   
## 3 4   
## Uruguay Uzbekistan   
## 2 4   
## Venezuela Vietnam   
## 2 2   
## Yemen Zambia   
## 2 4   
## Zimbabwe   
## 4

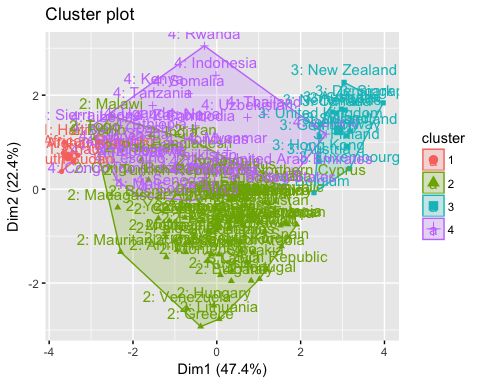
#get size of each cluster  
table(hclust.out\_cut)

## hclust.out\_cut  
## 1 2 3 4   
## 5 80 17 39

#plot the heatmap  
# set labels as cluster membership and utility name  
row.names(whr\_2017\_DF.scaled) <- paste(hclust.out\_cut, ": ", row.names(whr\_2017\_DF.scaled), sep = "")  
  
#set color scheme  
col\_set<- grep("blue",colors())  
h\_col<- colors()[c(col\_set[1:20])]  
  
# rev() reverses the color mapping to large = dark  
heatmap(as.matrix(whr\_2017\_DF.scaled), Colv = NA, hclustfun = hclust,  
 col=rev(h\_col))



#visualize each cluster for hierarchical clustering  
fviz\_cluster(list(data = whr\_2017\_DF.scaled, cluster = hclust.out\_cut))

 ### Interpretation of Part II: \* In this section our group build clustering models without LifeLadder column. First we normalized the data to increase consistency and avoid problems brought by different units. Then we set the row names and built the k-means and hierarchical model. \* The number of clusters has been chosen as 4. Our decision is based on the outcome of screen plot. From the chart it can be seen that increasing number of clusters from 1 to 4 can significantly decrease the variation within each cluster. And including more than 4 clusters do not help decrease the inner variation that much. Actually in this part, we faced with a hard decision about whether choosing 3 or choosing 4 as the optimal k. After trying to plot models using both k=3 and k=4, we found out that choosing k as 4 can gives us a more clear visual understanding of the clusters. To be more specific, in hierarchical model, using k as 4 splits one big, messy, and “unmeaningful” cluster into two smaller clusters, giving us more information and a easier understanding of the clusters. \* We used complete method to build our hierarchical model and generate a heatmap to better interpret the variation within each cluster. In order to be comparable with the k-mean method, the number of cluster of hierarchical model has also been chosed as 4. \* Note that those two methods use different algorithms, so the labelings of the clusters catergorized by the two methods are not one to one correspondence. For example, the cluster 1 of k-mean model is not equal to the cluster 1 of hierarchical model.

### Detailed Interpretation of Model Results:

#### For K-means model, there are four clusters, and the size for each cluster is 31, 39, 53, 18; the five countries for each cluster are:

* Cluster 1: Chile, China, Greece, Italy, Russia
* Cluster 2: Central African Republic, Chad, Congo, India, Kenya
* Cluster 3: France, Japan, Spain, United States, Poland
* Cluster 4: Australia, Canada, Finland, United Kingdom, Germany

#### Similar attributes of the countries within one cluster

From the k-means model, four clusters has been generated and their respective attributes can be summarized as:

#### Cluster that has the highst happiness level: Cluster 4

* Compared to other clusters, Countries in Cluster 4 is obviously the happiest country because they have the highest amount of GDP, the highest level of social support, the highest number of life expenses, the highest level of freedom of life choice, the highest level of generosity and the lowest level of corruption. People live in those countries have the least economic trouble and treat people around themselves very well. Countries in this cluster are highly-developed and live a very happy life.

#### Cluster that has the lowest happiness level: Cluster 2

* Countries in cluster 2 is easy to be differentiated because they have the lowest amount of GDP, the lower level of social support and the lowest number of life expense. Based on those three metrics, it can be concluded that those countries are poor-developed and have unoptimistic economic perspectives. However, they seems to have moderate level of freedom of life choice, the second highest level of generosity and the second lowest level of corruption. In sum our group believes that countries in this cluster represent the least happiness level.

#### Clusters that have the moderate happiness level: Cluster 1 & 3

* Commons in cluster 1 & 3: Countries in cluster 1 & 3 have similar amount of GDP, similar level of social support, similar amount of life expenses and similar level of corruption. But they do have some differences.
* Cluster 1: Countries in cluster 1 is less developed than countries in cluster 3. Those countries have the lowest level of life choice, the lowest level of generosity and the highest level of corruption. It can be inferred that the humanistic environments of those countries are not very good.
* Cluster 3: Countries in cluster 3 have the second highest level of freedom of choice but the second lowest level of generosity. Those countries also have the second high level of corruption (second only to cluster 1).

#### For hierarchical model, when cluster k = 4, the size for each cluster is 5, 80, 17, 39;the five countries for each cluster are:

* Cluster 1: Afghanistan, Haiti, Chad, Central African Republic, South Sudan
* Cluster 2: South Korea, Spain, Russia, Japan, Italy
* Cluster 3: United Kingdom, Switzerland, Germany, Norway, New Zealand
* Cluster 4: Philippines, Nepal, Israel, Indonesia, South Africa

#### Similar attributes of the countries within one cluster:

From the hierarchical model, four clusters has been generated and their respective attributes can be summarized as:

#### Cluster that has the highst happiness level: Cluster 3

* From the heatmap, it can be seem that, compared to other clusters, Countries in Cluster 3 is obviously the happiest country, because they have the highest amount of GDP, the highest level of social support, the highest number of life expenses, the highest level of freedom of life choice, the highest level of generosity and the lowest level of corruption. The typical countries in this cluster are Germany,Norway and New Zealand.People live in those countries have the least economic trouble and treat people around themselves very well. Countries in this cluster are highly-developed and live a very happy life.

#### Cluster that has the lowest happiness level: Cluster 1

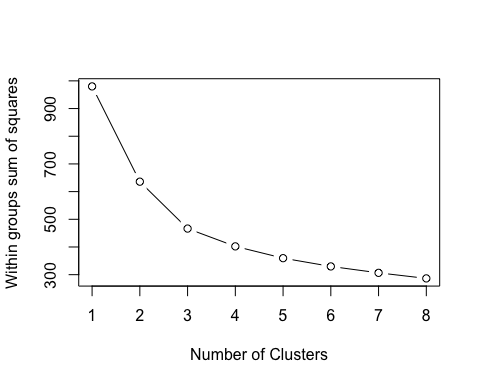
* Countries in cluster 1 is easy to be differentiated because they have the lowest amount of GDP, the lower level of social support and the lowest number of life expense. Based on those three metrics, it can be concluded that those countries are poor-developed and have unoptimistic economic perspectives. However, from the heatmap, it can be seen that they seems to have moderate level of freedom of life choice, the second highest level of generosity and the second lowest level of corruption(only second to some countries in cluster 2). In sum our group believes that countries in this cluster represent the least happiness countries.

#### Clusters that have the moderate happiness level: Cluster 2 & 4

* Commons in cluster 2 & 4: Countries in cluster 2 & 4 have similar amount of GDP, similar level of social suppor and similar amount of life expenses. But they do have some differences.
* Cluster 2: Countries in cluster 2 are more diverse than countries in cluster 4 because cluster 2 contains the most countries in this model.Countries in cluster 2 face with severe corruption problem and show less generosity to other people. The level of freedom of life choice varies in this cluster, but the overall freedom level seems to be pretty.
* Cluster 4: The humanistic environments of cluster 4 are better than that of of cluster 2. Countries in cluster 4 have the second highest level of freedom of choice and the second highest level of generosity. But those countries also face with the second high level of corruption.

## Part III: Models with LifeLadder

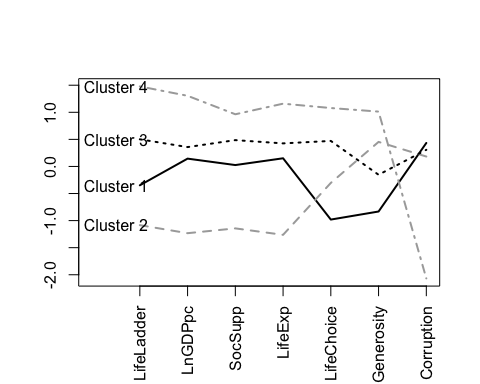
#add LifeLadder column  
whr\_2017\_DF\_LL <- whr\_2017DB[,-9]  
  
# Normalized the data  
whr\_2017\_DF\_LL.scaled<- scale(whr\_2017\_DF\_LL[,-1])  
  
#set row names  
row.names(whr\_2017\_DF\_LL.scaled) <-whr\_2017\_DF$country  
  
#build k-means model  
par(mfrow = c(1, 1))  
  
# Initialize total within sum of squares error: wss  
wss <- 0  
  
# Look over 3 to 8 possible clusters  
for (i in 1:8) {  
 # Fit the model: km.out  
 km.out\_LL <- kmeans(whr\_2017\_DF\_LL.scaled, centers = i, nstart = 20, iter.max = 50)  
 # Save the within cluster sum of squares  
 wss[i] <- km.out\_LL$tot.withinss  
}  
  
# Produce a scree plot  
plot(1:8, wss, type = "b",   
 xlab = "Number of Clusters",   
 ylab = "Within groups sum of squares")



# Select number of clusters  
k <- 4  
  
# Build model with k clusters: km.out  
km.out\_LL <- kmeans(whr\_2017\_DF\_LL.scaled, centers = k, nstart = 50, iter.max = 50)  
  
# View the resulting model  
km.out\_LL

## K-means clustering with 4 clusters of sizes 33, 38, 52, 18  
##   
## Cluster means:  
## LifeLadder LnGDPpc SocSupp LifeExp LifeChoice Generosity  
## 1 -0.3482041 0.1443440 0.0256353 0.1508588 -0.9809241 -0.8321632  
## 2 -1.0838871 -1.2325810 -1.1436733 -1.2627632 -0.3040541 0.4532131  
## 3 0.5013997 0.3575746 0.4860318 0.4257908 0.4712948 -0.1536623  
## 4 1.4780923 1.3044912 0.9633314 1.1591966 1.0787345 1.0127625  
## Corruption  
## 1 0.4332098  
## 2 0.1831711  
## 3 0.3068444  
## 4 -2.0673520  
##   
## Clustering vector:  
## Afghanistan Albania   
## 2 1   
## Algeria Argentina   
## 3 3   
## Armenia Australia   
## 1 4   
## Austria Azerbaijan   
## 4 1   
## Bahrain Bangladesh   
## 3 2   
## Belarus Belgium   
## 1 4   
## Benin Bolivia   
## 2 3   
## Bosnia and Herzegovina Botswana   
## 1 1   
## Brazil Bulgaria   
## 3 1   
## Burkina Faso Cambodia   
## 2 2   
## Cameroon Canada   
## 2 4   
## Central African Republic Chad   
## 2 2   
## Chile China   
## 3 1   
## Colombia Congo (Brazzaville)   
## 3 2   
## Congo (Kinshasa) Costa Rica   
## 2 3   
## Cyprus Czech Republic   
## 3 3   
## Denmark Dominican Republic   
## 4 3   
## Ecuador Egypt   
## 3 1   
## El Salvador Estonia   
## 3 3   
## Ethiopia Finland   
## 2 4   
## France Gabon   
## 3 1   
## Georgia Germany   
## 1 4   
## Ghana Greece   
## 2 1   
## Guatemala Guinea   
## 3 2   
## Haiti Honduras   
## 2 3   
## Hong Kong Hungary   
## 4 1   
## Iceland India   
## 4 2   
## Indonesia Iran   
## 2 2   
## Iraq Ireland   
## 1 4   
## Israel Italy   
## 3 1   
## Ivory Coast Japan   
## 2 3   
## Jordan Kazakhstan   
## 3 3   
## Kenya Kosovo   
## 2 3   
## Kuwait Kyrgyzstan   
## 3 3   
## Latvia Lebanon   
## 1 1   
## Lesotho Liberia   
## 2 2   
## Libya Lithuania   
## 3 1   
## Luxembourg Macedonia   
## 4 3   
## Madagascar Malawi   
## 2 2   
## Mali Malta   
## 2 3   
## Mauritania Mauritius   
## 1 3   
## Mexico Moldova   
## 3 1   
## Mongolia Montenegro   
## 3 1   
## Morocco Myanmar   
## 1 3   
## Nepal Netherlands   
## 2 4   
## New Zealand Nicaragua   
## 4 3   
## Niger Nigeria   
## 2 2   
## Turkish Republic of Northern Cyprus Norway   
## 3 4   
## Pakistan Palestine   
## 2 1   
## Panama Paraguay   
## 3 3   
## Peru Philippines   
## 3 3   
## Poland Portugal   
## 3 3   
## Romania Russia   
## 3 1   
## Rwanda Saudi Arabia   
## 2 3   
## Senegal Serbia   
## 2 1   
## Sierra Leone Singapore   
## 2 4   
## Slovakia Slovenia   
## 3 3   
## Somalia South Africa   
## 2 1   
## South Korea South Sudan   
## 1 2   
## Spain Sweden   
## 3 4   
## Switzerland Taiwan   
## 4 3   
## Tajikistan Tanzania   
## 1 2   
## Thailand Togo   
## 3 2   
## Tunisia Turkey   
## 1 1   
## Turkmenistan Uganda   
## 3 2   
## Ukraine United Arab Emirates   
## 1 3   
## United Kingdom United States   
## 4 3   
## Uruguay Uzbekistan   
## 3 3   
## Venezuela Vietnam   
## 1 3   
## Yemen Zambia   
## 1 2   
## Zimbabwe   
## 2   
##   
## Within cluster sum of squares by cluster:  
## [1] 100.50670 164.33184 109.33138 28.01125  
## (between\_SS / total\_SS = 59.0 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

#plotting profile plot of centroids   
# plot an empty scatter plot  
plot(c(0), xaxt = 'n', ylab = "", xlab = "", type = "l",  
 ylim = c(min(km.out\_LL$centers), max(km.out\_LL$centers)), xlim = c(0, 7))  
  
# label x-axes  
axis(1, at = c(1:7), labels = colnames(whr\_2017\_DF\_LL.scaled), las = 2)  
  
# plot centroids  
for (i in c(1:4))  
lines(km.out\_LL$centers[i,], lty = i, lwd = 2, col = ifelse(i %in% c(1, 3, 5),  
 "black", "dark grey"))  
# name clusters  
text(x = 0.5, y = km.out\_LL$centers[, 1], labels = paste("Cluster", c(1:4)))



# Create hierarchical clustering model: hclust.out  
hclust.out\_LL <- hclust(dist(whr\_2017\_DF\_LL.scaled), method="complete")  
  
# Inspect the result  
summary(hclust.out\_LL)

## Length Class Mode   
## merge 280 -none- numeric   
## height 140 -none- numeric   
## order 141 -none- numeric   
## labels 141 -none- character  
## method 1 -none- character  
## call 3 -none- call   
## dist.method 1 -none- character

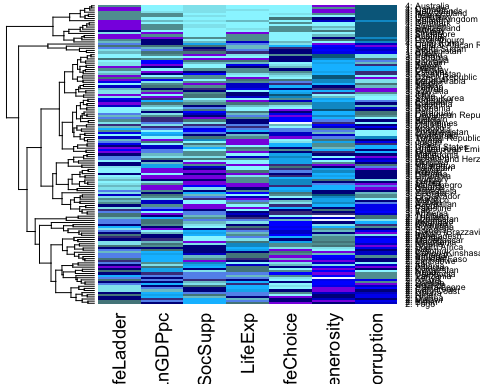
#prune the tree  
hclust.out\_cut\_LL <- cutree(hclust.out\_LL, k = 4)  
  
#inspect the clusters  
hclust.out\_cut\_LL

## Afghanistan Albania   
## 1 2   
## Algeria Argentina   
## 3 3   
## Armenia Australia   
## 3 4   
## Austria Azerbaijan   
## 4 3   
## Bahrain Bangladesh   
## 3 2   
## Belarus Belgium   
## 3 4   
## Benin Bolivia   
## 2 3   
## Bosnia and Herzegovina Botswana   
## 3 2   
## Brazil Bulgaria   
## 3 3   
## Burkina Faso Cambodia   
## 2 2   
## Cameroon Canada   
## 2 4   
## Central African Republic Chad   
## 1 1   
## Chile China   
## 3 3   
## Colombia Congo (Brazzaville)   
## 3 2   
## Congo (Kinshasa) Costa Rica   
## 2 3   
## Cyprus Czech Republic   
## 3 3   
## Denmark Dominican Republic   
## 4 3   
## Ecuador Egypt   
## 3 3   
## El Salvador Estonia   
## 3 3   
## Ethiopia Finland   
## 2 4   
## France Gabon   
## 3 3   
## Georgia Germany   
## 3 4   
## Ghana Greece   
## 2 3   
## Guatemala Guinea   
## 3 2   
## Haiti Honduras   
## 1 3   
## Hong Kong Hungary   
## 4 3   
## Iceland India   
## 4 2   
## Indonesia Iran   
## 2 2   
## Iraq Ireland   
## 3 4   
## Israel Italy   
## 3 3   
## Ivory Coast Japan   
## 2 3   
## Jordan Kazakhstan   
## 3 3   
## Kenya Kosovo   
## 2 3   
## Kuwait Kyrgyzstan   
## 3 2   
## Latvia Lebanon   
## 3 3   
## Lesotho Liberia   
## 2 2   
## Libya Lithuania   
## 3 3   
## Luxembourg Macedonia   
## 4 3   
## Madagascar Malawi   
## 2 2   
## Mali Malta   
## 2 3   
## Mauritania Mauritius   
## 2 3   
## Mexico Moldova   
## 3 3   
## Mongolia Montenegro   
## 3 3   
## Morocco Myanmar   
## 3 2   
## Nepal Netherlands   
## 2 4   
## New Zealand Nicaragua   
## 4 3   
## Niger Nigeria   
## 2 2   
## Turkish Republic of Northern Cyprus Norway   
## 3 4   
## Pakistan Palestine   
## 2 3   
## Panama Paraguay   
## 3 3   
## Peru Philippines   
## 3 3   
## Poland Portugal   
## 3 3   
## Romania Russia   
## 3 3   
## Rwanda Saudi Arabia   
## 2 3   
## Senegal Serbia   
## 2 3   
## Sierra Leone Singapore   
## 2 4   
## Slovakia Slovenia   
## 3 3   
## Somalia South Africa   
## 2 2   
## South Korea South Sudan   
## 3 1   
## Spain Sweden   
## 3 4   
## Switzerland Taiwan   
## 4 3   
## Tajikistan Tanzania   
## 3 2   
## Thailand Togo   
## 2 2   
## Tunisia Turkey   
## 3 3   
## Turkmenistan Uganda   
## 3 2   
## Ukraine United Arab Emirates   
## 3 3   
## United Kingdom United States   
## 4 3   
## Uruguay Uzbekistan   
## 3 2   
## Venezuela Vietnam   
## 3 3   
## Yemen Zambia   
## 2 2   
## Zimbabwe   
## 2

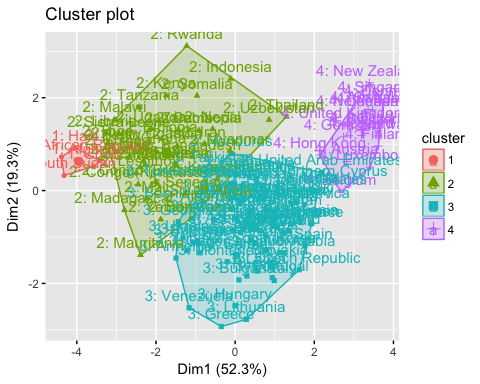
#get size of each cluster  
table(hclust.out\_cut\_LL)

## hclust.out\_cut\_LL  
## 1 2 3 4   
## 5 42 76 18

#plot the heatmap  
# set labels as cluster membership and utility name  
row.names(whr\_2017\_DF\_LL.scaled) <- paste(hclust.out\_cut\_LL, ": ", row.names(whr\_2017\_DF\_LL.scaled), sep = "")  
  
# plot heatmap  
# rev() reverses the color mapping to large = dark  
heatmap(as.matrix(whr\_2017\_DF\_LL.scaled), Colv = NA, hclustfun = hclust,  
 col=rev(h\_col))



#visualize each cluster for hierarchical clustering  
fviz\_cluster(list(data = whr\_2017\_DF\_LL.scaled, cluster = hclust.out\_cut\_LL))

 ### Interpretation of Part III: \* In this section our group added back the LifeLadder column to our dataset. Same as the last part, first we normalized the data to increase consistency and avoid problems brought by different units. Then we set the row names and built the k-means and hierarchical model. \* The number of clusters has been chosen as 4. The main reason for this decision is the same as in the case when we exclude the variable “Life Ladder”. One more point is that after we tried different values for k, ranging from 3 to 6, the results indicate that the choice of 4 clusters is more meaningful, especially in this part where we add the variable “Life Ladeer”. Although the results of our experiments are not all shown here, we do find out that when number of clusters increases from 3 to 6, the most and least happy cluster barely change, only the middle cluster is split into several new clusters. The case of 4 clusters with variable Life Ladder contains more information than 3 clusters and clustering without Life Ladder, which we will explain more detailed later. Further splits add less information, with the side effect that clustering becomes too specific and visualization plots becomes messy. \* We used complete method to build our hierarchical model and generate a heatmap to better interpret the variation within each cluster. In order to be comparable with the k-mean method, the number of cluster of hierarchical model has also been chosed as 4.

### Detailed Interpretation of Model Results:

### For K-means model, there are four clusters, and the size for each cluster is 33,38,52,18; the five countries for each cluster are:

* Cluster 1: Bulgaria, China, Egypt, Greece, Italy
* Cluster 2: Afghanistan, Central African Republic, Haiti, Indonesia, Ethiopia
* Cluster 3: France, Japan, Thailand, Mexico, United States
* Cluster 4: Australia, Belgium, Norway, Netherlands, Luxembourg
* Note that the order of the four clusters is not according to the rank of happiness. We will give the rank after we describe the similarities within each cluster.

### Similar attributes of the countries within one cluster

From the k-means model, three clusters has been generated and their respective attributes can be summarized as:

#### Cluster that has the highest happiness level: cluster 4

* Countries in cluster 4 have highest life ladder score, lnGDP per capita and generosity, longest life expectancy, most social support and life choice, least corruption level way below the level of the other three clusters. Countries in this cluster are therefore of highest happiness level.
* If we look into the countries in this cluster, we will find that they are among the most developed countries, with attractive natural environment, beautiful sceneries, stable political environment, sound social welfare systems, and pleasant life pace.

#### Cluster that has the lowest happiness level: cluster 2

* Countries in this cluster generally have lowest life ladder score and lnGDP per capita, with least social support, shortest life expectancy, life choice more than cluster1 but less than the other two, fair generosity only less than cluster 4, and higher corruption level (very close to cluster1 and cluster3). Countries in this cluster seem to have the lowest happiness score.
* Countries in this cluster is totally opposite to cluster4 in terms of material conditions. As most of them are among the least developed countries, they are struggling with problems such as poverty, illness, natural disasters, or conflicts due to unstable political environments. But the generosity level in those countries is only lower than the happiest cluster. That happens to some countries that are in poverty when the situation of the country unites its people, make them to care about others that share the similar or worse situation with them.

#### Clusters that have the moderate happiness level: cluster 1 & 3

* Things in common: Both cluster 1 and cluster 3 have moderate level of lnGDP per capital, social support, life experience, and life ladder scores higher than cluster 2 but lower than cluster 4. Apart from that, they both have high level of corruption that is very close to the corruption level of cluster 2.
* Cluster 1: Countries in cluster 1 seem to have very little life choice and least generosity even though they are not in poverty. They are mostly developing countries that have good economic environment but have a humanity environment that’s not as good as other clusters. This is a normal stage in the developing process. We find that most of countries, like Egypt, China, Greece, have longer histories and unique cultures, and thus have their traditional thoughts that limit people’s life choice to some extent. It definitely takes them longer time to figure out which part of the tradition to keep and which part to change in order to adapt to the world that is changing fast especially in the recent 100 years.
* Cluster 3: Compared to cluster 1, cluster 3 have higher level of almost all variables except corruption. Life choices are much more and generosity levels are much higher. They are a stage ahead countries in cluster 1 in the developing process.

### For hierarchical model, when cluster k=4, the size for each cluster is 5,42,76,18;the five countries for each cluster are:

* Cluster 1: Afghanistan, Haiti, Chad, South Sudan,Central African Republic
* Cluster 2: India, Ethopia, South Africa, Benin, Kenya
* Cluster 3: China, France, Greece, Italy, United States
* Cluster 4: Australia, Iceland, New Zealand, Norway, United Kingdom
* Note that different from the k-means method, the order of the four clusters here is consistent with the rank of happiness. Therefore the numbers are not the same as thoes in k-means method.
* In the heatmap countries in the same cluster stick together, but the clusters are not in the order of their ranking of happiness level. The order in the heatmap is 4, 1, 3, 2 from top to the bottom.

### Similar attributes of the countries within one cluster:

It is not as easy to compare average variable levels between clusters with hierarchical model as with k-means model, but we still have some obvious findings:

#### Cluster that has the highest happiness level: cluster 4

* We define cluster3 as the most happiest cluster since it is obvious that countries in cluster3 all have the least corruption, and the other variables have higher level than most of other countries.

#### Cluster that has the lowest happiness level: cluster 1

* Among all the countries, countries in this cluster have lowest life ladder score and least life choice, lower level of lnGDP per capital, social support and life expectancy. They also have highest generosity and corruption level. Therefore they seem to be least happy countries.

#### Clusters that have the moderate happiness level: cluster 2 & 3

* Things in common: It’s obvious that Most countries in these two clusters have moderate corruption level and life ladder score. But it’s hard to compare the average level of these two clusters with the levels of the other two clusters, because the hierarchical method doesn’t show the average level with the heatmap and both cluster 2 and cluster 3 have higher variation within cluster than the other two clusters. This is a big weakness of this method.
* Comparing cluster with cluster 3, we can find that countries in cluster 3 generally have higher life ladder score, higher lnGDP per capita, lower level generosity and corruption level. But without the average level shown, we can’t tell the difference in life choice and life expectancy. We can see the in-cluster variation of these variables are significant from the heatmap.

## Part IV:Impact of Happiness Life Ladder

The Life Ladder index shows how happy people think their life is. Comparing to other factors, Life Ladder appears to be a subjective measure and is more or less correlated with GPD and other factors. But after adding this variable to our analysis, our group find out that the models which contain this variable do capture something different and we think this Life Ladder variable is valuable to our analysis. \* After adding the Life Ladder variable, the between ss/total ss ratio increased from 56.9% to 59%, indicating a slight improvement on the accuracy of k-mean model. \* Adding the Life Ladder variable changed the graphic output of hierarchical model. From the cluster plot, it can be seen that the clusters of the model without Life Ladder, especially cluster 2 and 4, seems to be highly overlapped and make it very different for us to derive information from the chart. In contrast, the clusters of the model which contains Life Ladder seems to be less overlapped and each cluster has its own characteristics for us to classify. It seems like the Life Ladder variable further helps us to differentiate the countries.  
\* Sometimes the statistical data might not able to fully represent what people really think in that country. The life ladder variable is the most powerful variable to idenify people’s thoughts. external circumstances will not totally determine how people define happiness. For example, Iceland in the hierarchical model is an example where people still feel happy even though the government declared bankruptcy and corruption is a serious trouble. As a result, Iceland is identified as a country that is moderately happy when we exclude Life Ladder from the factors; it is identified as a country that is highly happy when we include Life Ladder in the factors. In this case, people ‘s attitude matters because it is an important spiritual reference despite the exogenous factors.

## Part V:Most Meaningful Model (chosen by our group)

* Our group thinks that the k-means model with Life Ladder is the most meaningful model.
* The output generated by the hierarchical model is too messy to be interpreted (non-interpretable). Every country has its own unqiue characteristics that makes it difficult for us to get the mean value, or the average performace, of each cluster. In contrast, the centroid profile plot of k-means model delivers a clear, understandable and useful message which is easier for us to get the key characteristics of each cluster.
* K-means plot can also present the inner variation within each cluster by plotting each company’s data, and its visual effect is still better than hierarchical model’s.
* By including Life Ladder, the model takes both objective factors and subjective attitude into consideration. The between ss/total ss ratio increased from 56.9% to 59%, indicating a better model performance.
* In this k-means model, Cluster 1 can be labelled as “moderate economic development/low life choice”, Cluster 2 can be labelled as “poor economic development/high generosity”, Cluster 3 can be labelled as “moderate economic development/high life choice”, Cluster 4 can be labelled as “high economic development/low corruption”.

## Part VI:Is it accurate to generalize Tolstoy’s assertion to countries as well as families?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CLuster # | Models with Life Ladder | Variance |  | Models without Life Ladder | Variance |
| Cluster 1 | 100.50670/33 | 3.05 |  | 78.67993/31 | 2.55 |
| Cluster 2 | 164.33184/38 | 4.32 |  | 156.39218/39 | 4.01 |
| Cluster 3 | 109.33138/52 | 2.10 |  | 103.28379/53 | 1.95 |
| Cluster 4 | 28.01125/18 | 1.56 |  | 23.99144/18 | 1.33 |

Yes, our group believes that it is accurate to generalize Tolstey’s assertion to countries as well as families based on our above analysis. \* The k-means table above shows the variances of different clusters with different input variables. Cluster 4 is the most happiest one and cluster 2 is the least happiest one. It can be seen that, in both cases, cluster 4 which contains the most happiest countries has the least variance, indicating their high similarity. On the other hand, cluster 2 which contains the least happiest countries has the highest variance, which delivers the message that countries in this group have the highest diversity. Those variance numbers are consistent with Tolstoy’s saying. \* We can also derive the conclusion from the graphic output of our models. In the “happy country” cluster of the hierarchical model, we observe roughly similar characteristics of each country. Almost all of them have high Life Ladder, GDP per capita, social support, life expectancy, life choice, generosity and low corruption. However, countries in the “unhappy country” cluster shows more variety in their characteristics. For example, Haiti has relatively high generosity, but its performance in life choice and corruption is extremely poor; Central African Republic has relatively high GDP per capita, but its performance in social supply and corruption is really unsatisfying.

## Part VII: Reflections of the two clustering methods

#### Strength of k-means method:

* Easier to figure out the general characteristic for all variables of each cluster, because the lines shows the average value of each variable.
* Easier to compare between clusters.
* Can show in-cluster variation by calculating Var as SSR/N, where N is the number of countries in that cluster, or by ploting the lines of all countries in that cluster.

#### Weakness of k-means method:

* When we add or change a variable, it’s hard to find out how and why the result is changed from the plot. The only way is to compare two long lists containing country names and cluster numbers.

#### Strength of hierarchical method:

* Shows both the process and results of clustering with a dendrogram.
* With the heatmap, it’s able to view each country’s scores of all variables and compare it with all other countries within or out of its cluster.
* Can show members of three clusters in a scatter plot. Therefore it’s easy to check the impact of adding/changing a variable on the results. For example, we find that Iceland changes cluster after we add the Life Ladder column.
* Purely data driven. There is no need to predetermine a k value. We can decide number of clusters after observing the clustering dendrogram.

#### Weakness of hierarchical method:

* The heatmap doesn’t show the average level of a cluster. So we can summarize a cluster’s characteristic of one variable only when the variation is small. Taking the example of life choice variable, with the significant variation, we can’t compare the average levels between cluster 2 and cluster 3.
* Compared to k-means that reassigns countries till the clusters barely change, hierarchical method only assign one data once thus there’s no chance to reassign if there’s a mistake.
* Depends too much on data. If data changes, such as dropping some rows, the result may be influenced to a large extent.

#### Other comments

* Choice of a better model: The two methods both have strengths and limitations. The hierarchical method seems to be able to provide more information, such as the process of clustering, and each country’s performance within one graph. But choosing a model is really a case-by-case choice. In this case, what we care most is clear differences between clusters and the degree of in-cluster variation. The k-means model just satisfied our demand and the clear and meaningful graph outperforms the heatmap of the hierarchical method.
* Potential improvements: When trying to analyze the reasonability of the clustering results, we find some similarities of countries in the same cluster that definitely has influence in happiness but are not included in the current model, like war and political environment. In the future improvement, we can consider to add more variables like them into the happiness clustering model, if further analysis on the new variables doesn’t show a likelihood of significant side effect.

=======

### Reference

* [1].<https://rpubs.com/mohammadshadan/273129>
* [2].<https://www.kaggle.com/unsdsn/world-happiness/kernels>
* [3]<https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html>
* [4]<http://uc-r.github.io/hc_clustering>
* [5]Data Mining for Business Analytics
* [6]<https://rstudio-pubs-static.s3.amazonaws.com/93706_e3f683a8d77244a5b993b20ad6278f4b.html>