BigData II Case 3 Report

Group F

Introduction

This project works to figure out the correctness of the statement generated by Leo Tolstoy which goes “All happy families are alike; each unhappy family is unhappy in its own way.” And we use the “2017 World Happiness Report” to find whether all happy countries are alike in some ways.

```r
library(tidyverse)
library(ggplot2)
library(tidyr)
library(plyr)
library(factoextra)
library(car)
library(GGally)
```

Data Processing

After having a look at the dataset, we find that there are some missing observations in the variable columns. 8 NAs in the GDPpc and LnGDPpc. 12 NAs in Corruption. 11 NAs in Generosity. 3 NAs in LifeChoice. 1 NA in LifeExp. If we choose to delete all the missing obvervations, we only have 121 countries in the list, but originally we have 141. We think it is a waste of infomation. So we decide to solve the missing problem.

```r
# Read data
happiness<-read.csv("~/Desktop/whr_2017.csv")

# Input some countries' missing GDP values and get their LnGDPpc
happiness[139,9]<-2325.07
happiness[23,9]<-647.8804
happiness[88,9]<-5305.047
happiness[4,9]<-18489.43
happiness[80,9]<-35705.1
happiness$LnGDPpc<-log(happiness$GDPpc,exp(1))
```

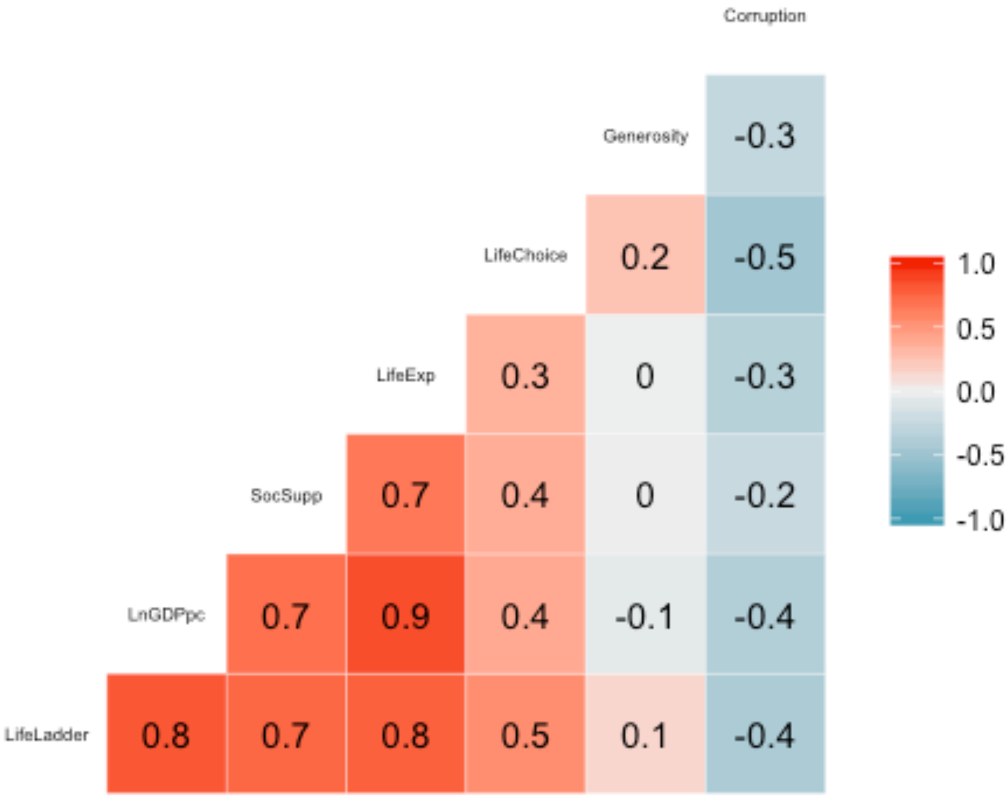
In order to handle the missing value, we first search the dataset of World bank and find the 2016 GDPpc value of 5 missing countries.

```r
# Fill out NAs with Average values
for(i in c(2,4,5,6,7,8)){
  happiness[,i][is.na(happiness[,i])] <- mean(happiness[,i], na.rm = TRUE)
}
```

```
# Take out the rows with missing values
happiness<-na.omit(happiness)
```

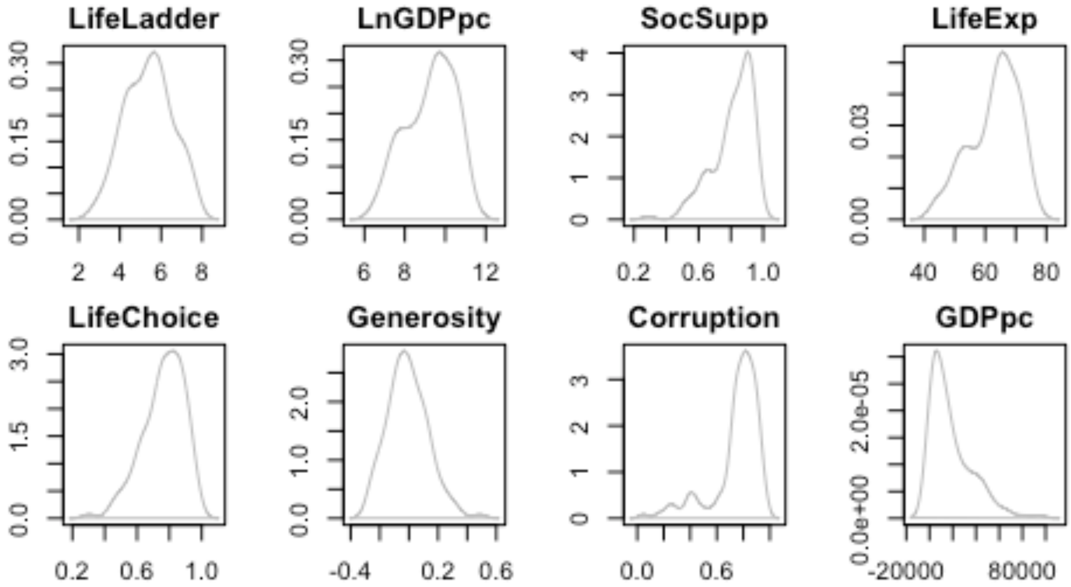
Then, we have a look at variation of Corruption, Generosity, LifeChoice and LifeExp. The result shows that these 4 variables all have low variations, so they are all close to the mean. Meanwhile, we find that the missing varibles of the country is at most 3. If we fill the NAs with mean value of the variables, they still have ability to be clustered. So, we decided to put mean value in the missing observations.

```
# Draw a correlation matrix
ggcorr(happiness[, -c(1,9)], label=TRUE, cex=2)
```



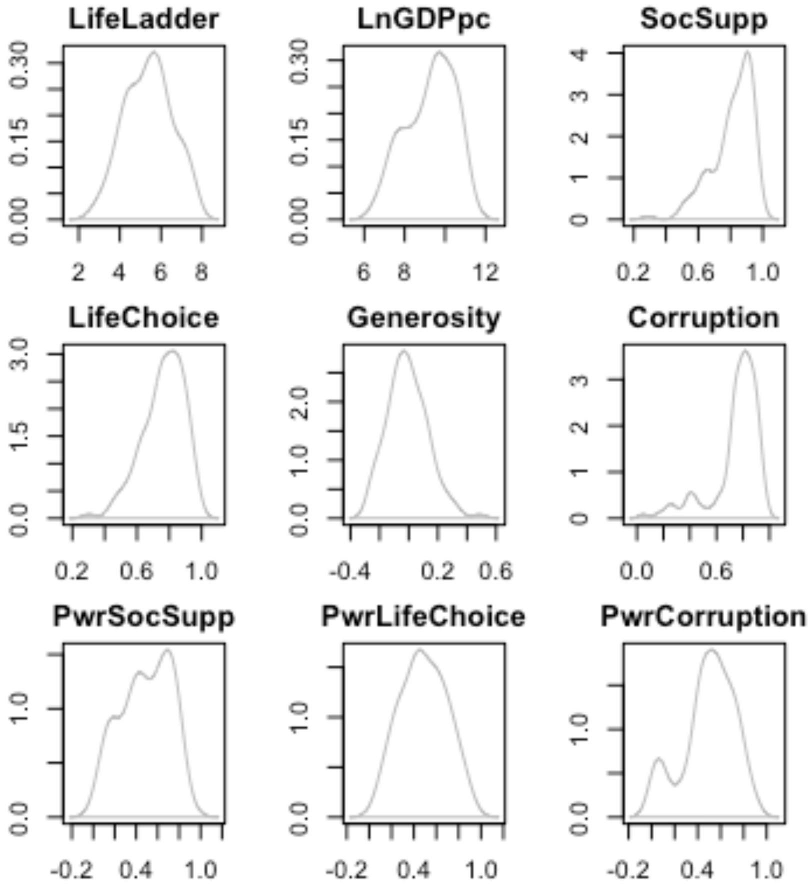
```
# Draw density graphs to see the distrbution of all the variables
opar<-par(no.readonly = TRUE)
par(mfrow=c(3, 4))
par(mar = rep(2, 4))

for (i in 2:9) {
  d <- density(happiness[,i])
  plot(d, type="n", main=colnames(happiness)[i])
  polygon(d, border="gray")}
```

```
par(opar)
```



```
# Do power transformation on variables with skewness and redraw density graphs
powerTransform(happiness[,c(4,6,8)])  ## evaluate best power number

## Estimated transformation parameters
##    SocSupp LifeChoice Corruption
##   4.061827   2.928190   2.836991

happiness$PwrSocSupp<-(happiness$SocSupp)^4.06
happiness$PwrLifeChoice<-(happiness$LifeChoice)^2.93
happiness$PwrCorruption<-(happiness$Corruption)^2.84

opar<-par(no.readonly = TRUE)
par(mfrow=c(3, 4))
par(mar = rep(2, 4))

for (i in 2:12) {
  d <- density(happiness[,i])
  plot(d, type="n", main=colnames(happiness)[i])
  polygon(d, border="gray")}
```

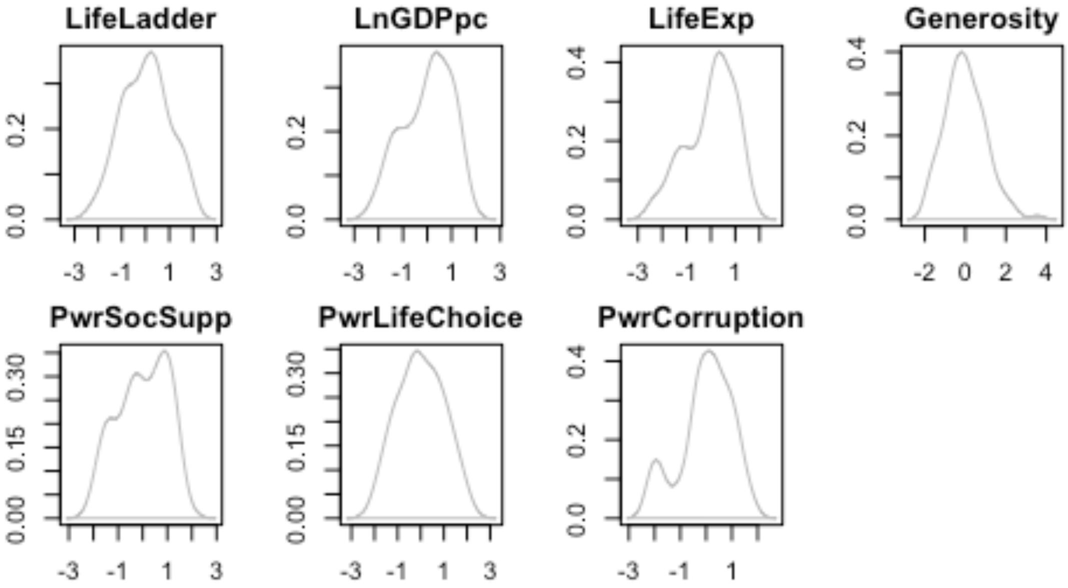
```r
par(opar)
```



Before choosing the variables to put into our model, we are curious about the distribution of these variables. Our clustering methods use Euclidean Distance, which computes the square of the distance between data points, and therefore tails of the distribution have the disproportionate effect on the model. We find that SocSupp, LifeChoice and Corruption have negative skewness. And GDPpc has positive skewness. However, LnGDPpc has already solve the problem of GDPpc. So, we decide to only do power transformation of SocSupp, LifeChoice and Corruption. We use powerTransform fuction to find the best power. The second graph shows the result of power transformation.

```r
# Select variables after analysis
happiness1<-happiness[,c(1,2,3,5,7,10,11,12)]

# Normalize input variables
happiness.norm <- data.frame(sapply(happiness1[,-1], scale))

# Add row names
row.names(happiness.norm) <- happiness1[,1]

# Draw normalized density graphs of selected variables
```

```
opar<-par(no.readonly = TRUE)
par(mfrow=c(3,4))
par(mar = rep(2, 4))

for (i in 1:7) {
  d <- density(na.omit(happiness.norm[,i]))
  plot(d, type="n", main=colnames(happiness.norm)[i])
  polygon(d, border="gray")}

par(opar)
```



Then, we start to choose variables. We use LifeLadder, LnGDPpc, LifeExp, Generosity, powered SocSupp, powered LifeChoice and powered Corruption. Although some variables are highly correlated, for example, LifeExp, SocSupp, LifeLadder and LnGDPpc. It seems like reasonable. Countries with higher GDP may have higher life level because they are able to buy anything they want and get more support from their friends or relatives. Also, people with more money have ability to go to hosiptal and meet better doctors when they are sick. But we think they measures different dimension of countries, GDP can not explain everything. So we still keep them in the inputs. The last thing building the model we have done is normalizing the variable in order to avoid the impact of scale and calculate the distance between countries more accurate.

Model Selection

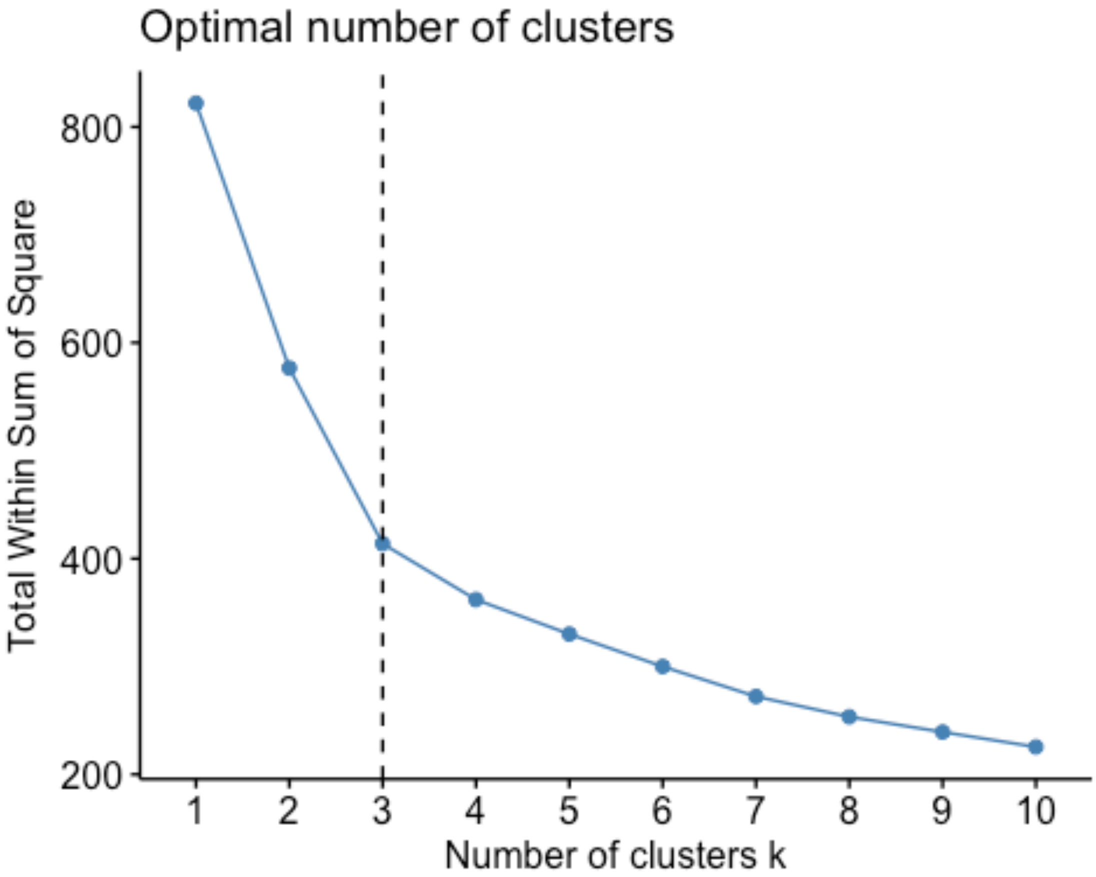
Clustering

Overall, we choose the number of cluster based on the elbow point in WSS diagram for each cluster model. Then comparing the cluster diagram of each model and combining with educational knowledges, we choose the model which generates the most proportional and non-overlapping clusters and provides meaningful and practical significance.

```
# Exclude lifeLadder
happiness.norm2<-happiness.norm[,-1]
```

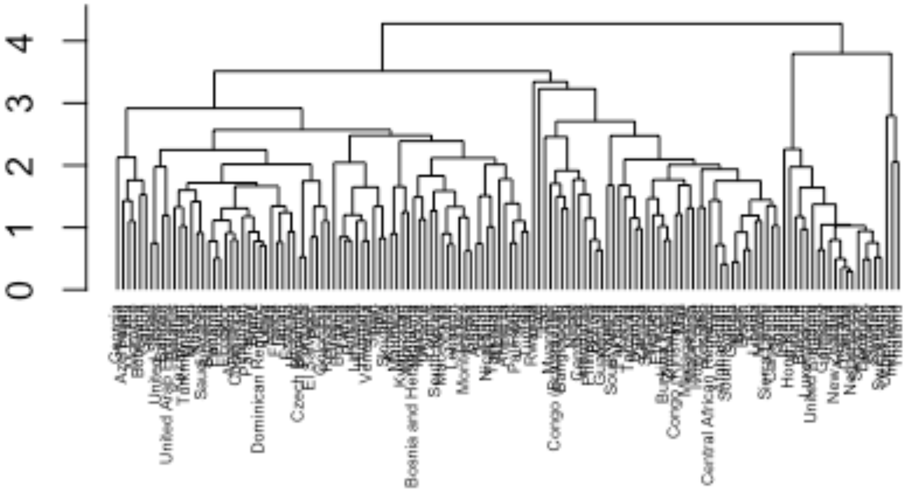
Model 1 Hierarchical model without LifeLadder

```
## Compute Euclidean distance
d.norm <- dist(happiness.norm2, method = "euclidean")

## Choose the optimal k based on WSS
fviz_nbclust(happiness.norm2, hcut, method = "wss") +
  geom_vline(xintercept = 3, linetype = 2)
```



Based on elbow point in WSS diagram, the k equals to 3 in hierarchical cluster method without LifeLadder.

```
## Compare average and complete method
hc1 <- hclust(d.norm, method = "average")
plot(hc1, hang = -1, ann = FALSE,cex=0.45)
```

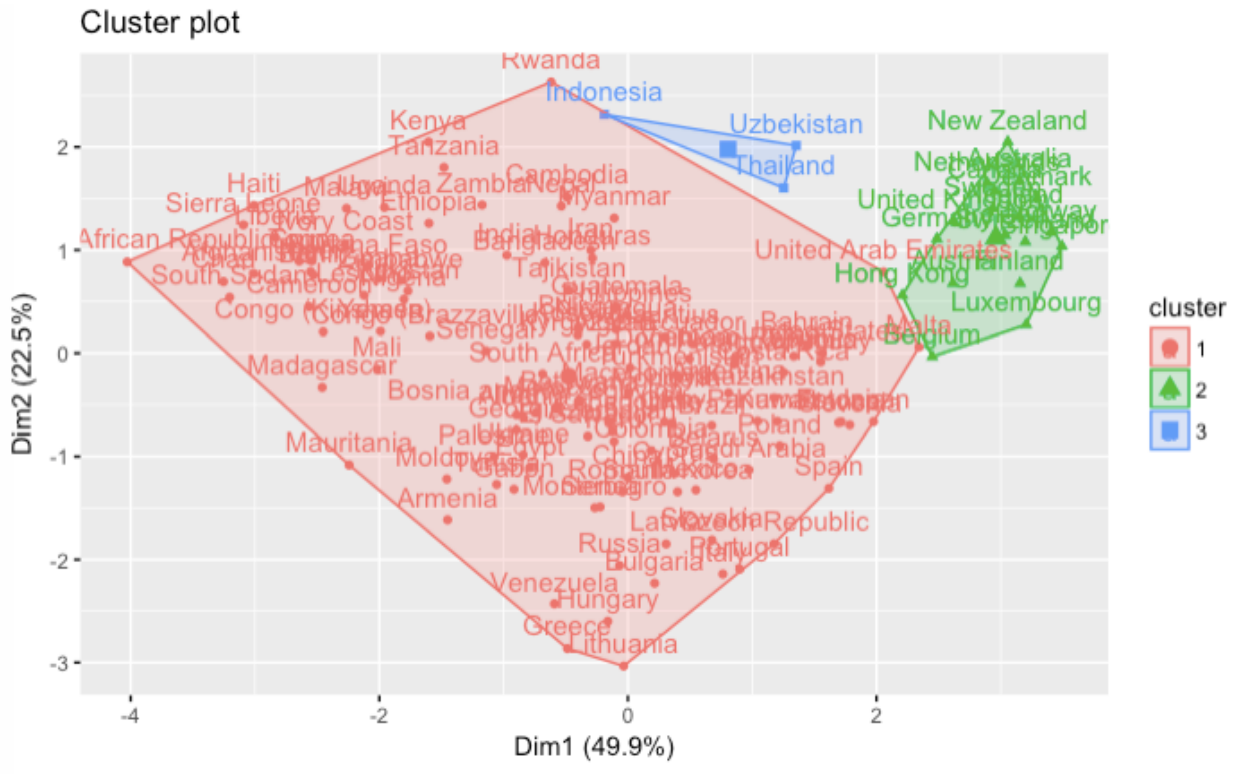


```
memb1 <- cutree(hc1, k = 3)
memb1

##              Afghanistan                  Albania                  Algeria 

##                        1                        1                        1 

##                Argentina                  Armenia                Australia 

##                        1                        1                        2 

##                  Austria               Azerbaijan                  Bahrain 

##                        2                        1                        1 

##               Bangladesh                  Belarus                  Belgium 

##                        1                        1                        2
##                    Benin                  Bolivia   Bosnia and Herzegovina
##                        1                        1                        1
##                 Botswana                   Brazil                 Bulgaria
##                        1                        1                        1
##             Burkina Faso                 Cambodia                 Cameroon
##                        1                        1                        1
##                   Canada Central African Republic                     Chad
##                        2                        1                        1
##                    Chile                    China                 Colombia
##                        1                        1                        1
##     Congo (Brazzaville)         Congo (Kinshasa)               Costa Rica
##                        1                        1                        1
##                   Cyprus           Czech Republic                  Denmark
##                        1                        1                        2
##       Dominican Republic                  Ecuador                    Egypt
##                        1                        1                        1
##              El Salvador                  Estonia                 Ethiopia
##                        1                        1                        1
##                  Finland                   France                    Gabon
##                        2                        1                        1
##                  Georgia                  Germany                    Ghana
##                        1                        2                        1
##                   Greece                Guatemala                   Guinea
##                        1                        1                        1
##                    Haiti                 Honduras                Hong Kong
##                        1                        1                        2
##                  Hungary                  Iceland                    India
##                        1                        2                        1
##                Indonesia                     Iran                     Iraq
##                        3                        1                        1
##                  Ireland                   Israel                    Italy
##                        2                        1                        1
##              Ivory Coast                    Japan                   Jordan
##                        1                        1                        1
##               Kazakhstan                    Kenya                   Kosovo
##                        1                        1                        1
##                   Kuwait               Kyrgyzstan                   Latvia
##                        1                        1                        1
##                  Lebanon                  Lesotho                  Liberia
##                        1                        1                        1
##                    Libya                Lithuania               Luxembourg
##                        1                        1                        2
##                Macedonia               Madagascar                   Malawi
##                        1                        1                        1
##                     Mali                    Malta               Mauritania
##                        1                        1                        1
##                Mauritius                   Mexico                  Moldova
##                        1                        1                        1
##                 Mongolia               Montenegro                  Morocco
##                        1                        1                        1
##                  Myanmar                    Nepal              Netherlands
##                        1                        1                        2
##              New Zealand                Nicaragua                    Niger
##                        2                        1                        1
##                  Nigeria                   Norway                 Pakistan
##                        1                        2                        1
##                Palestine                   Panama                 Paraguay
##                        1                        1                        1
##                     Peru              Philippines                   Poland
##                        1                        1                        1
##                 Portugal                  Romania                   Russia
##                        1                        1                        1
##                   Rwanda             Saudi Arabia                  Senegal
##                        1                        1                        1
##                   Serbia             Sierra Leone                Singapore
##                        1                        1                        2
##                 Slovakia                 Slovenia             South Africa
##                        1                        1                        1
##              South Korea              South Sudan                    Spain
##                        1                        1                        1
##                   Sweden              Switzerland               Tajikistan
##                        2                        2                        1
##                  Tanzania                  Thailand                      Togo

##                         1                         3                         1

##                   Tunisia                    Turkey              Turkmenistan

##                         1                         1                         1

##                    Uganda                   Ukraine      United Arab Emirates

##                         1                         1                         1

##            United Kingdom             United States                   Uruguay

##                         2                         1                         1

##                Uzbekistan                 Venezuela                   Vietnam

##                         3                         1                         1

##                     Yemen                    Zambia                  Zimbabwe

##                         1                         1                         1
```

```
table(memb1)
```

```
## memb1
##   1   2   3 
## 117  18   3
```

```
fviz_cluster(list(data = happiness.norm2, cluster = memb1))
```

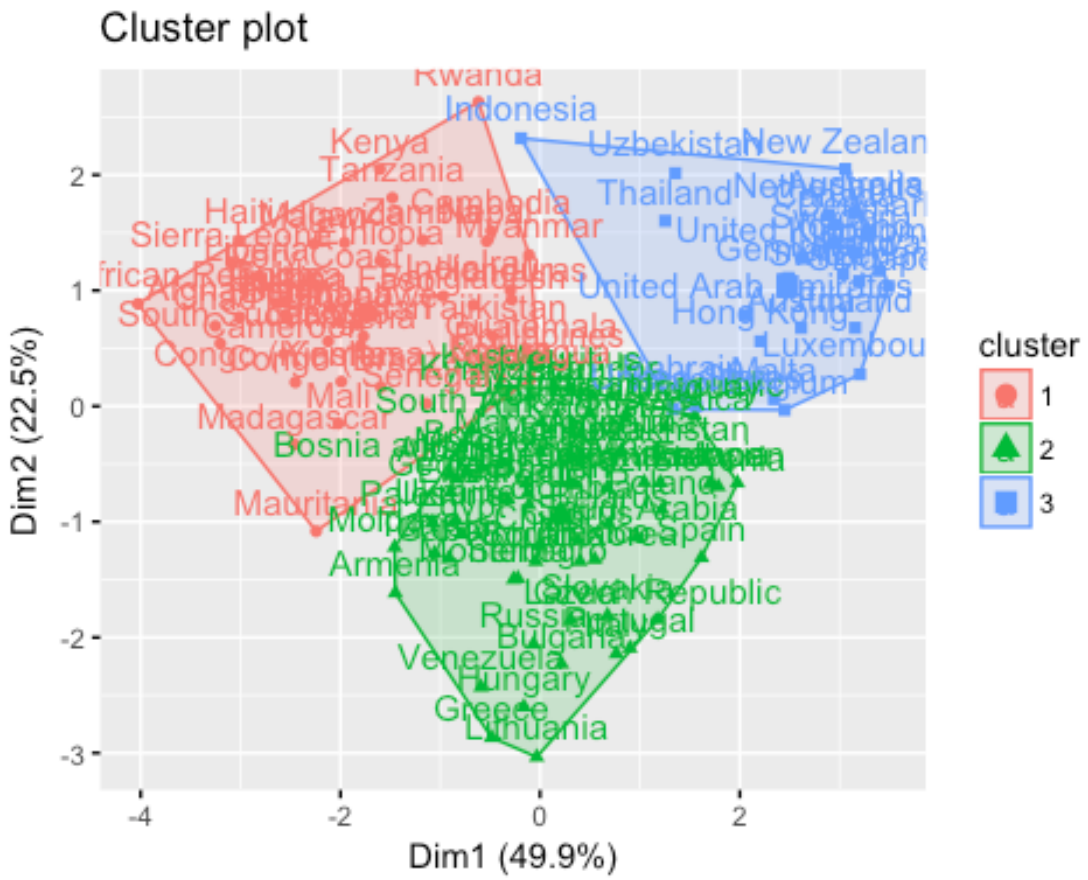


```
hc2 <- hclust(d.norm, method = "complete")
memb2 <- cutree(hc2, k = 3)
memb2

##               Afghanistan                   Albania                   Algeria

##                         1                         2                         2

##                 Argentina                   Armenia                 Australia

##                         2                         2                         3

##                   Austria                Azerbaijan                   Bahrain

##                         3                         2                         3

##                Bangladesh                   Belarus                   Belgium

##                         1                         2                         3

##                     Benin                   Bolivia    Bosnia and Herzegovina

##                         1                         1                         2

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

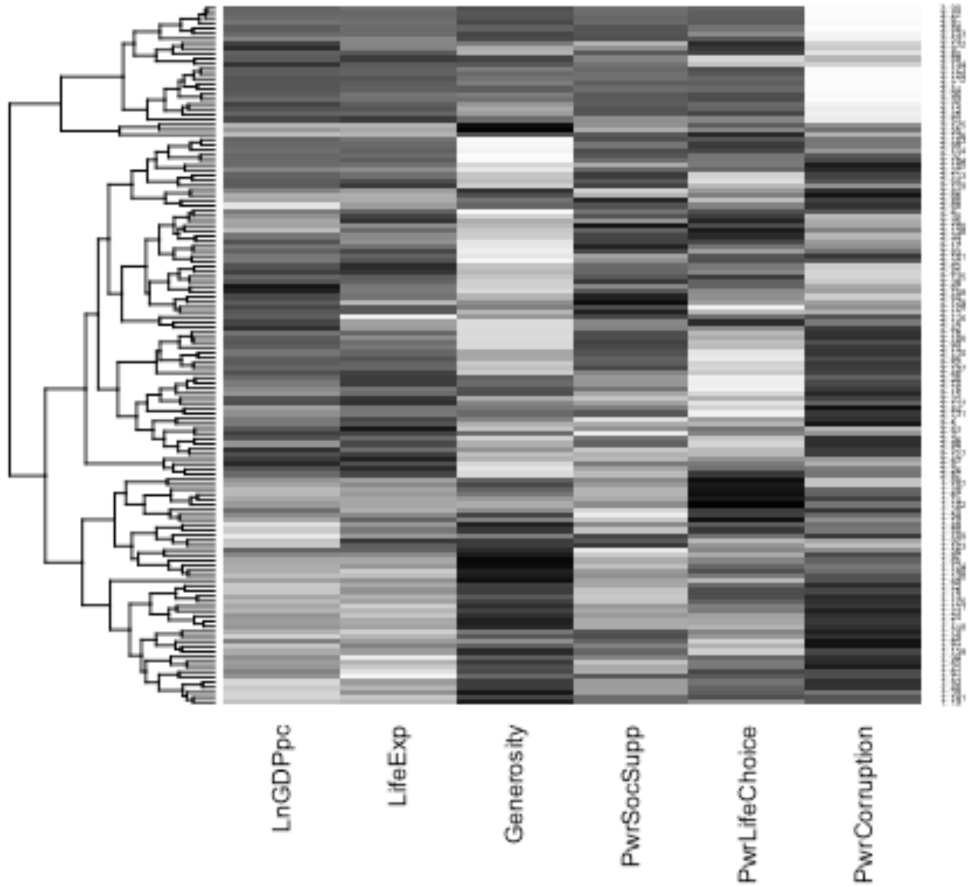
```
table(memb2)
```

```
## memb2
##  1  2  3
## 45 67 26
```

```
fviz_cluster(list(data = happiness.norm2, cluster = memb2))
```



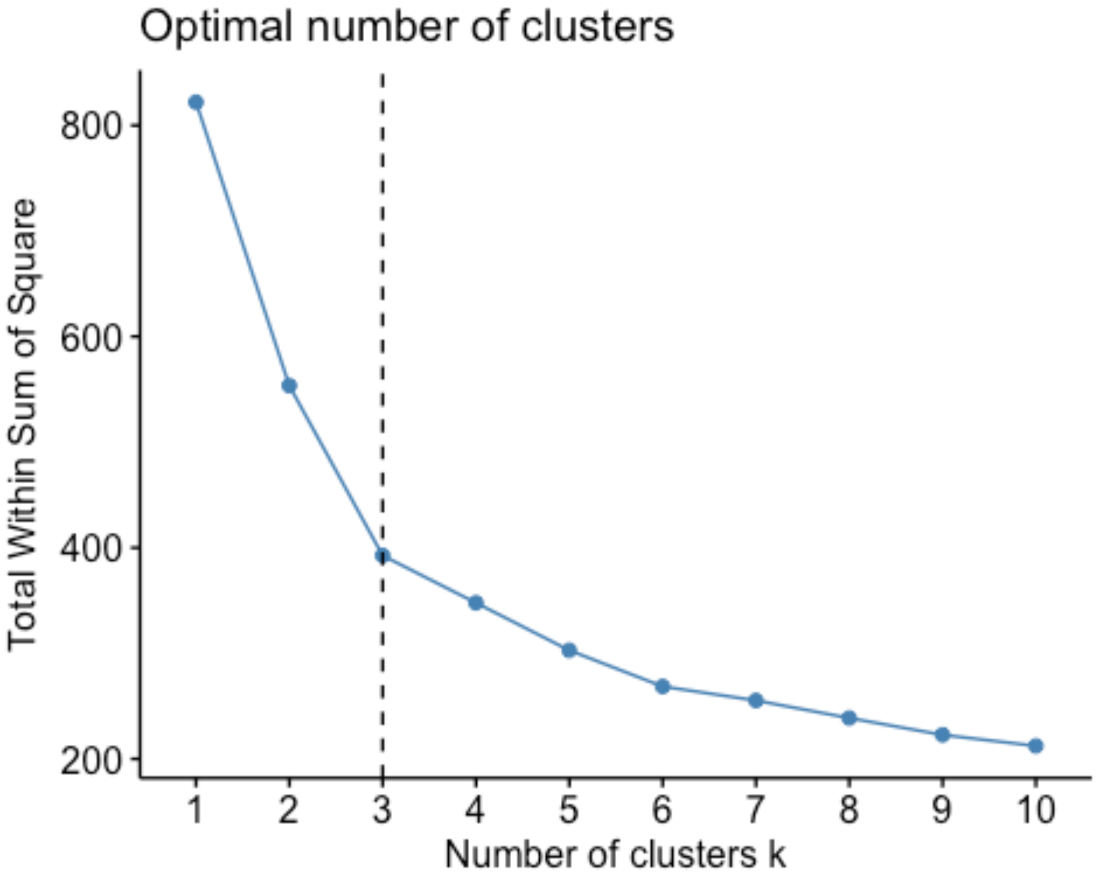
```
## plot heatmap
## rev() reverses the color mapping to large = dark
row.names(happiness.norm2) <- paste(memb2, ": ", row.names(happiness), sep =
"")
heatmap(as.matrix(happiness.norm2), dendrogram="row",Colv = NA, hclustfun = h
clust,key.xlab = "Cm",
        col=rev(paste("gray",1:99,sep="")), cexRow = 0.3,cexCol = 0.8)
```



Comparing average and complete methods, we choose the complete methods which generates more proportional clusters. The number of countries in each cluster respectively is 45, 67 and 26. In Cluster 1, there are countries such as India, Afghanistan, Ethiopia, Liberia and Nepal, where the countries are generally less developed. In Cluster 2, there are countries such as Peru, Portugal, Poland, Vietnam and Japan, where there is a mix of developed and less developed countries. In Cluster 3, there are countries such as Hong Kong, Denmark, United States, United Kingdom and Switzerland, where the countries are generally more developed. The heatmap is showed as below which provides more detailed analysis of each cluster. As the heat diagram shows, the Cluster 3, which is at the top of the heat map, usually has high GDP per capital, life expenditure, generosity, social support and life choice but has low corruption. The Cluster 2 has very similar patternto Cluster 3 in many attributes except generosity and corruption. It usually has a relative low generosity and a high corruption. In contrast, Cluster 1 generally has a relatively high index in corruption and lowest indexes in other four attributes except generosity. The generosity is in supervise higher than the cluster 2's.

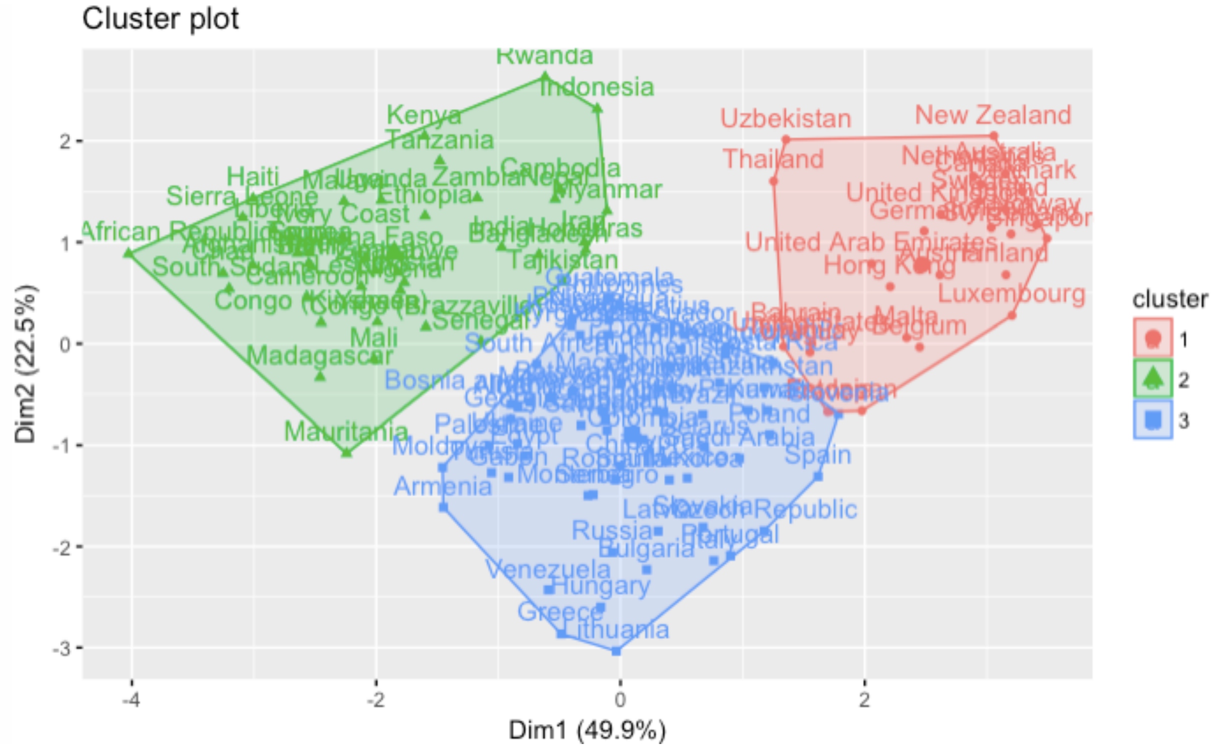
Model 2 K-means model without LifeLadder

```
## Choose the optimal k based on WSS
set.seed(123)
fviz_nbclust(happiness.norm2, kmeans, method = "wss") +
  geom_vline(xintercept = 3, linetype = 2)
```



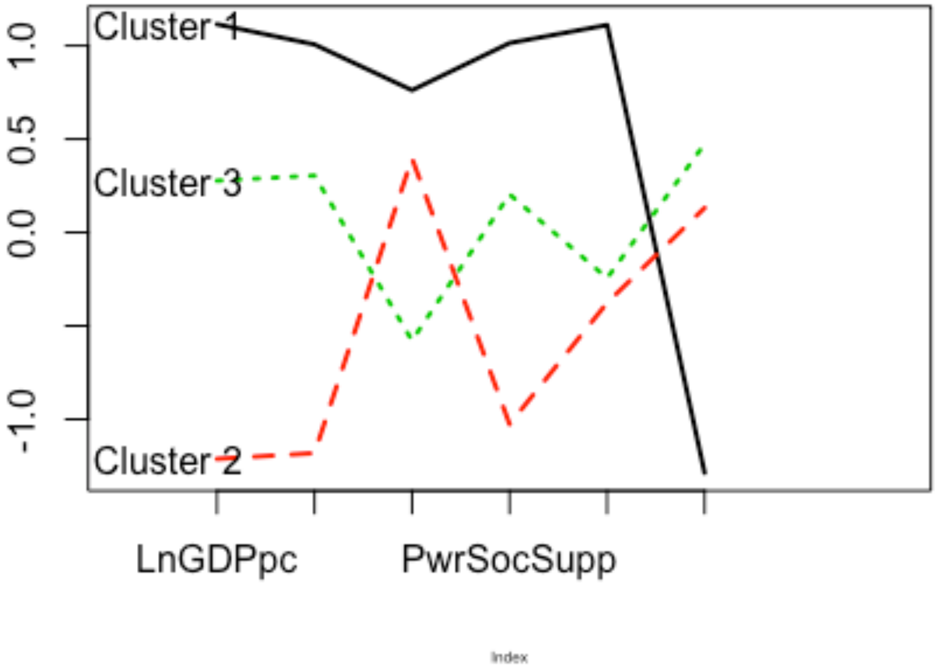
Based on elbow point in WSS, the k equals to 3 without LifeLadder. In later discussion it provides BSS/TSS=52.2%.

```
km <- kmeans(happiness.norm2, 3)
fviz_cluster(km, happiness.norm2)
```



```
km

## K-means clustering with 3 clusters of sizes 29, 42, 67
##
## Cluster means:
##      LnGDPpc    LifeExp Generosity PwrSocSupp PwrLifeChoice PwrCorruption
## 1  1.114994  1.006658  0.7616236  1.0141274     1.1107057    -1.2865436
## 2 -1.212155 -1.181419  0.3966743 -1.0261713    -0.3767489     0.1317047
## 3  0.277249  0.304873 -0.5783195  0.2043209    -0.2445822     0.4743010
##
## Clustering vector:
##   1: 1   2: 2   2: 3   2: 4   2: 5   3: 6   3: 7   2: 8   3: 9  1: 10
##      2      3      3      3      3      1      1      3      1      2
##  2: 11  3: 12  1: 13  1: 14  2: 15  2: 16  2: 17  2: 18  1: 19  1: 20
##      3      1      2      3      3      3      3      3      2      2
##  1: 21  3: 22  1: 23  1: 24  2: 25  2: 26  2: 27  1: 28  1: 29  2: 30
##      2      1      2      2      3      3      3      2      2      3
##  2: 31  2: 32  3: 33  2: 34  2: 35  2: 36  2: 37  2: 38  1: 39  3: 40
##      3      3      1      3      3      3      3      1      2      1
##  2: 41  2: 42  2: 43  3: 44  1: 45  2: 46  1: 47  1: 48  1: 49  1: 50
##      1      3      3      1      2      3      3      2      2      2
##  3: 51  2: 52  3: 53  1: 54  3: 55  1: 56  2: 57  3: 58  3: 59  2: 60
##      1      3      1      2      2      2      3      1      1      3
##  1: 61  2: 62  2: 63  2: 64  1: 65  2: 66  2: 67  2: 68  2: 69  2: 70
##      2      1      3      3      2      3      3      3      3      3
##  1: 71  1: 72  2: 73  2: 74  3: 75  2: 76  1: 77  1: 78  1: 79  3: 80
##      2      2      3      3      1      3      2      2      2      1
##  1: 81  2: 82  2: 83  2: 84  2: 85  2: 86  2: 87  1: 88  1: 89  3: 90
##      2      3      3      3      3      3      3      2      2      1
##  3: 91  1: 92  1: 93  1: 94  3: 96  1: 97  2: 98  2: 99 2: 100 2: 101
##      1      3      2      2      1      2      3      3      3      3
## 1: 102 2: 103 2: 104 2: 105 2: 106 1: 107 2: 108 1: 109 2: 110 1: 111
##      3      3      3      3      3      2      3      2      3      2
## 3: 112 2: 113 2: 114 2: 116 2: 117 1: 118 2: 119 3: 120 3: 121 1: 123
##      1      3      3      3      3      2      3      1      1      2
## 1: 124 3: 125 1: 126 2: 127 2: 128 2: 129 1: 130 2: 131 3: 132 3: 133
##      2      1      2      3      3      3      2      3      1      1
## 3: 134 2: 135 3: 136 2: 137 2: 138 1: 139 1: 140 1: 141
##      1      1      1      3      3      2      2      2
##
## Within cluster sum of squares by cluster:
## [1]  70.30303 129.89983 192.41012
##  (between_SS / total_SS =  52.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

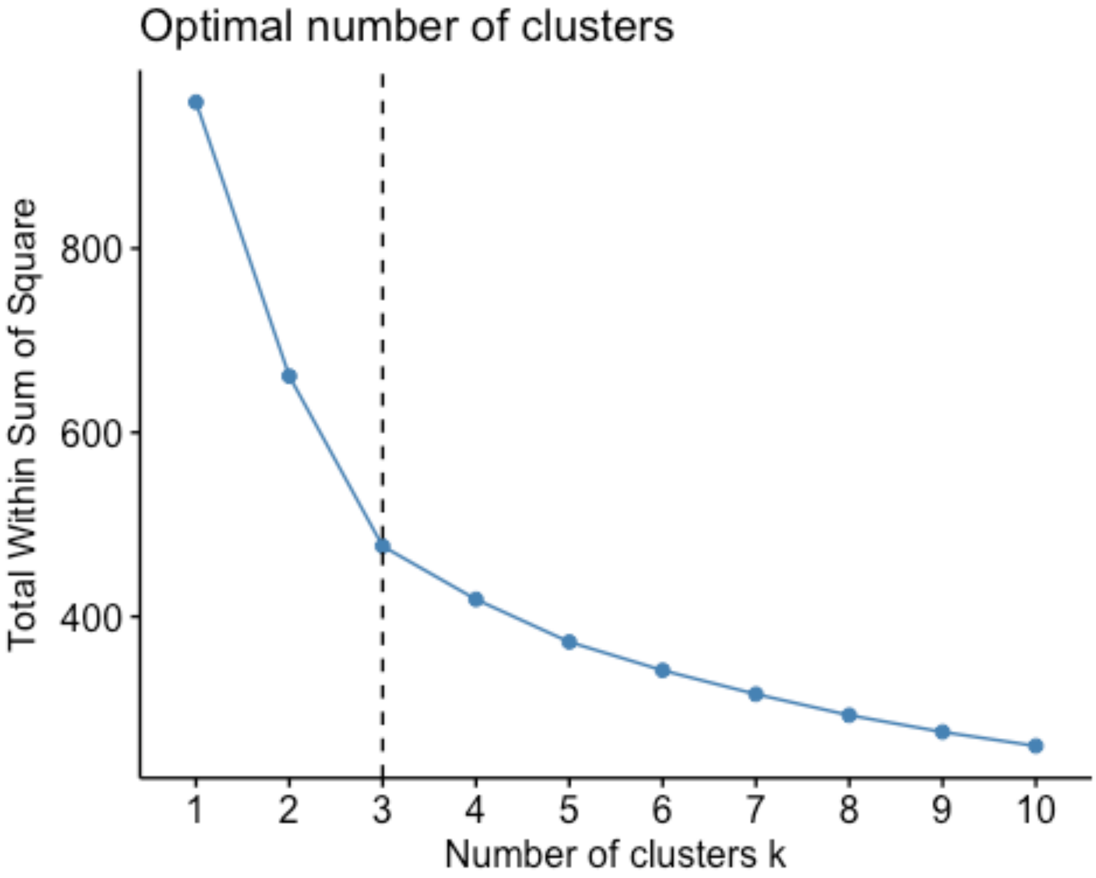
```r
## Plot an empty scatter plot
plot(c(0), xaxt = 'n', ylab = "", type = "l",
     ylim = c(min(km$centers), max(km$centers)), xlim = c(0, 8),cex.lab=0.4)

## Label x-axes
axis(1, at = c(1:6), labels = names(happiness.norm2))

## Plot centroids
for (i in c(1:3))
  lines(km$centers[i,], lty = i, lwd = 2, col = i)

## Name clusters
text(x = 0.5, y = km$centers[, 1], labels = paste("Cluster", c(1:3)))
```



The number of countries in each cluster respectively is 29, 42, 67. In Cluster 1, there are countries such as Austria, Canada, United States, United Kingdom and Singapore, where the countries are generally more developed. In Cluster 2, there are countries such as Afghanistan, India, Nepal, Congo and Liberia, where the countries are less developed. In Cluster 3, there are countries such as Peru, Argentina, Spain, Vietnam and Turkey, where there is a mix of developed and less developed countries. The centroid profile plot is provided as below. The Cluster 1, the same as Cluster 3 in hierarchical clustering, generally has high GDP per capital, life expenditure, generosity, social support and life choice but has a low corruption. The Cluster 2 generally is low in GDP, Life expenditure, social support and life choice, and relatively high in Generosity and Corruption. The Cluster 3 is very similar to Cluster 1 in many attributes except generosity and corruption. It usually has a relatively low generosity and a relatively high corruption.

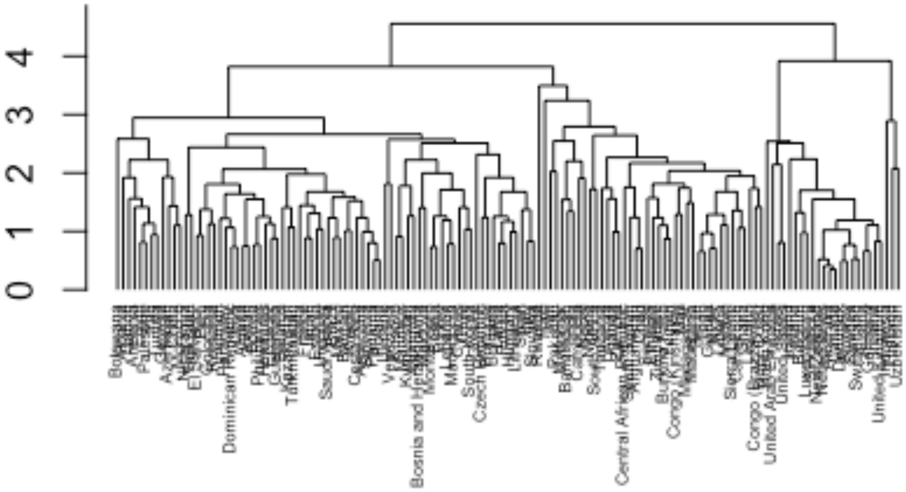
Model 3 Hierarchical model with LifeLadder

```
# Include lifeLadder and compute Euclidean distance
d.norm1 <- dist(happiness.norm, method = "euclidean")

## Choose the optimal k based on WSS
fviz_nbclust(happiness.norm, hcut, method = "wss") +
  geom_vline(xintercept = 3, linetype = 2)
```

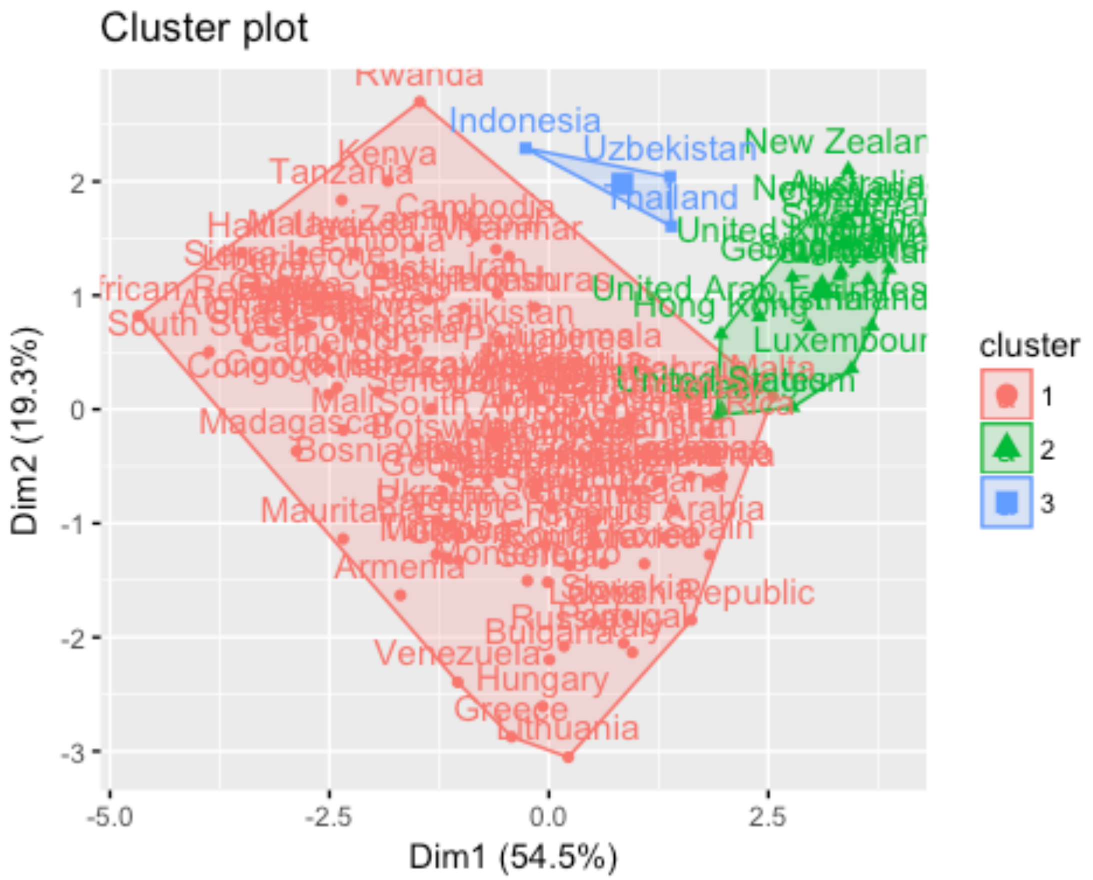


Based on elbow point in WSS diagram, the k equals to 3 in hierarchical cluster method with LifeLadder.

```
## Compare average and complete method
hc3 <- hclust(d.norm1, method = "average")
plot(hc3, hang = -1, ann = FALSE,cex=0.45)
```



```
memb3 <- cutree(hc3, k = 3)
table(memb3)

## memb3
##   1   2   3
## 114  21   3

fviz_cluster(list(data = happiness.norm, cluster = memb3))
```



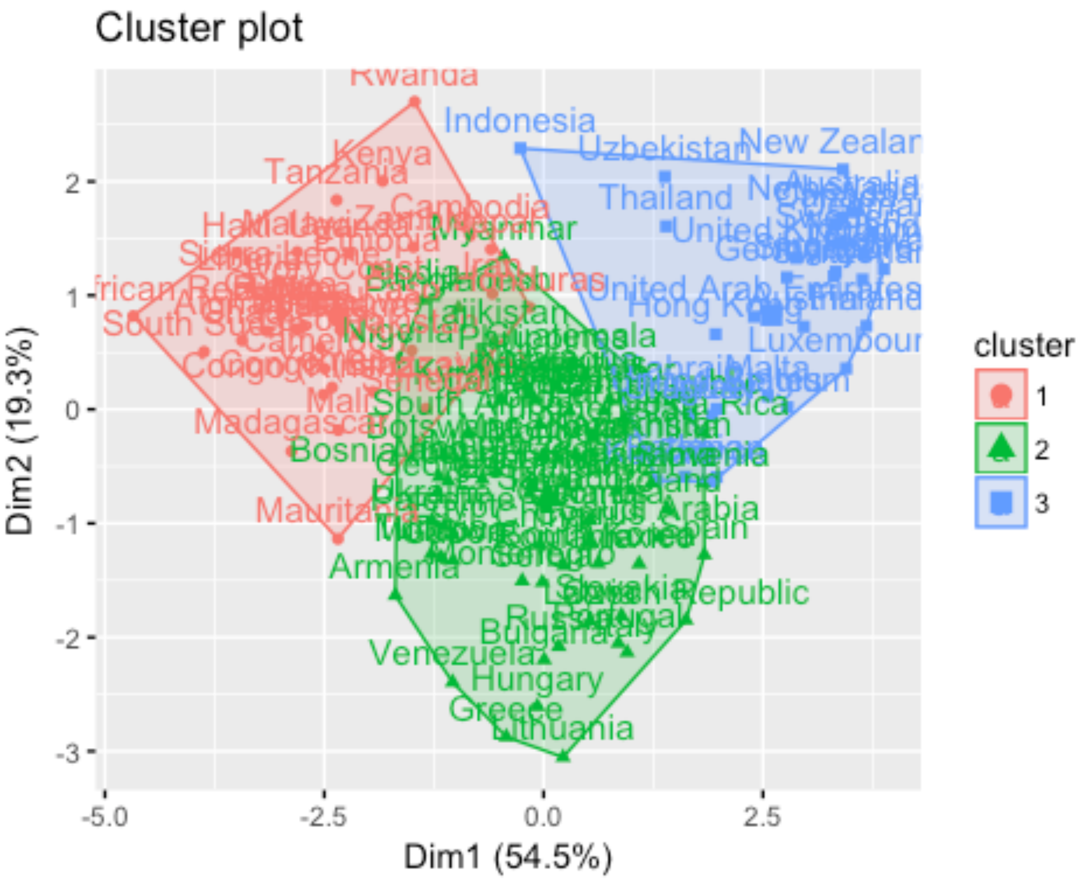
```
hc4 <- hclust(d.norm1, method = "complete")
memb4 <- cutree(hc4, k = 3)
memb4
```

```
##            Afghanistan                Albania                Algeria 

##                      1                      2                      2 

##              Argentina                Armenia              Australia 

##                      2                      2                      3 

##                Austria             Azerbaijan                Bahrain 

##                      3                      2                      3 

##             Bangladesh                Belarus                Belgium 

##                      2                      2                      3 

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

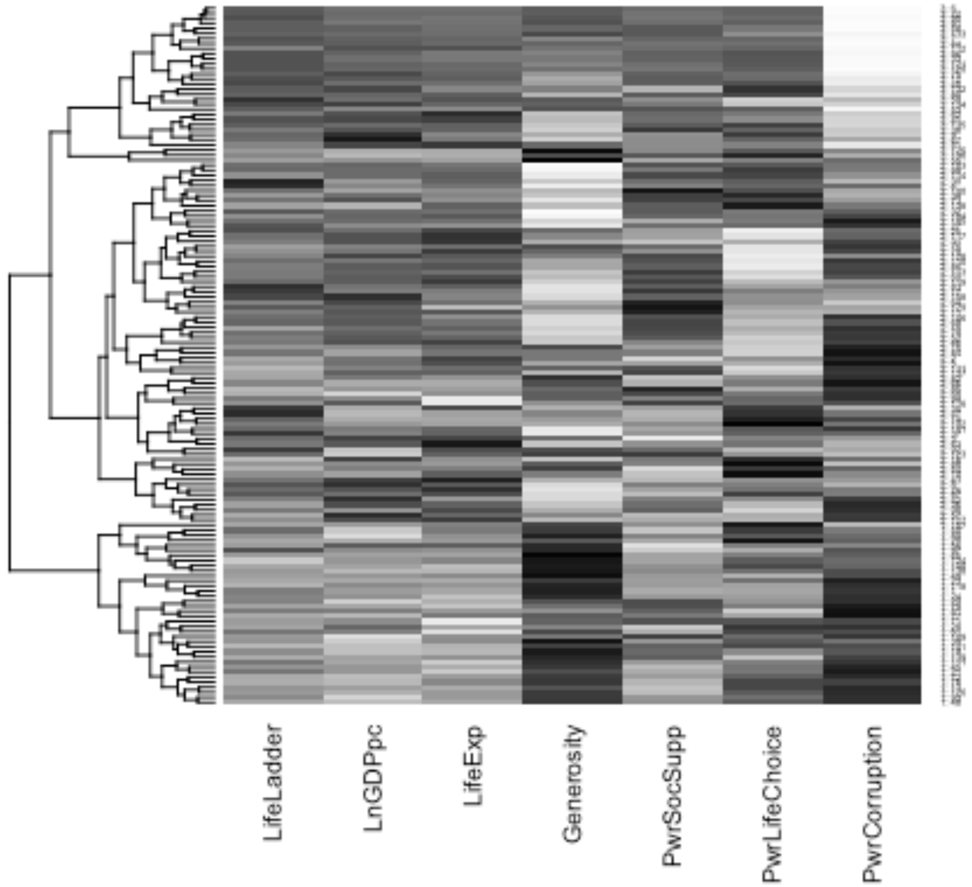
```
table(memb4)
```

```
## memb4
##  1  2  3 
## 36 71 31
```

```
fviz_cluster(list(data = happiness.norm, cluster = memb4))
```



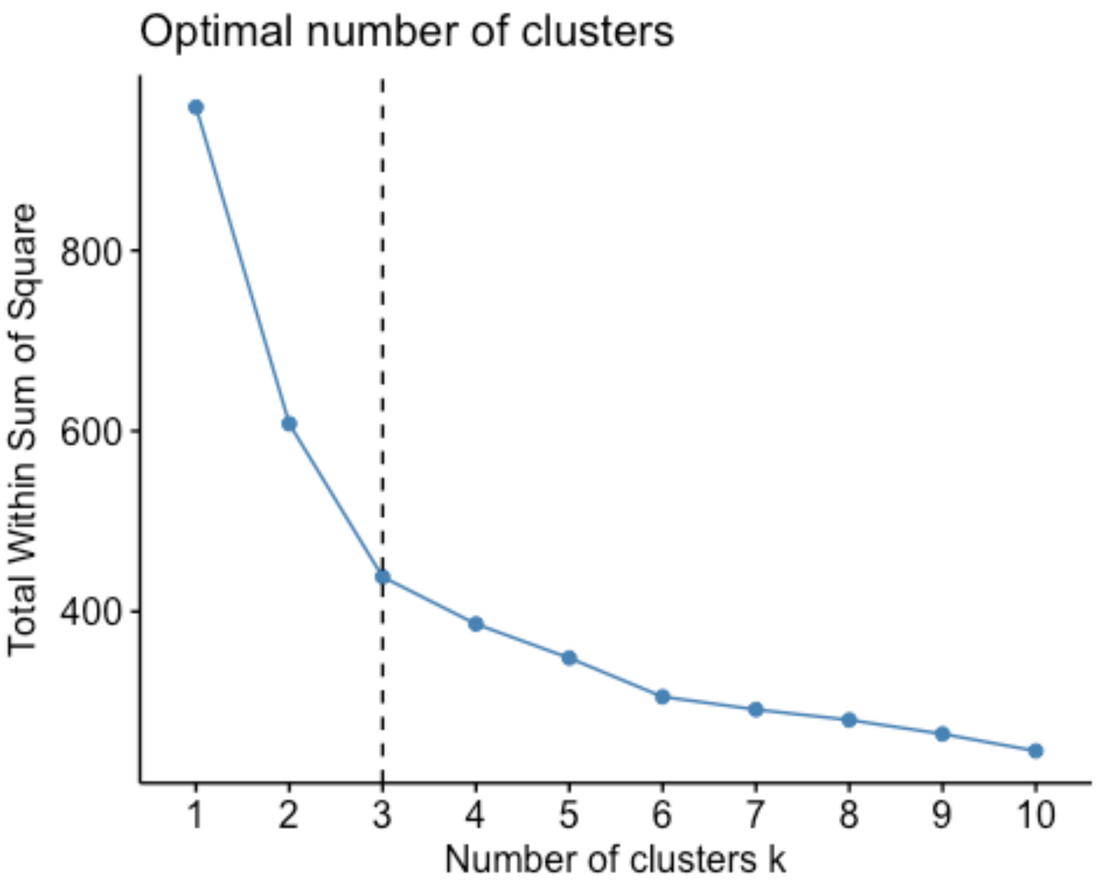
```
## Plot heatmap
## rev() reverses the color mapping to large = dark
row.names(happiness.norm) <- paste(memb4, ": ", row.names(happiness), sep = "
")
heatmap(as.matrix(happiness.norm), dendrogram="row",Colv = NA, hclustfun = hc
lust,key.xlab = "Cm",
        col=rev(paste("gray",1:99,sep="")), cexRow = 0.3,cexCol = 0.8)
```



Comparing average and complete methods, we choose the complete methods which generates more proportional clusters. The number of countries in each cluster respectively is 36, 71 and 31. In Cluster 1, there are countries such as India, Afghanistan, Ethiopia, Liberia and Nepal, where the countries are generally less developed. In Cluster 2, there are countries such as Peru, Portugal, Poland, Vietnam and Labia, where there is a mix of developed and less developed countries. In Cluster 3, there are countries such as Hong Kong, Denmark, United States, United Kingdom and Japan, where the countries are generally more developed. The heat map is showed as below which provides more detailed analysis of each cluster. As the heat diagram shows, the Cluster 3, which is at the top of the heat map, usually has high life ladder, GDP per capital, life expenditure, generosity, social support and life choice but has low corruption. The Cluster 2 has very similar pattern to Cluster 3 in many attributes exceptgenerosity and corruption. It usually has a relative low generosity and a high corruption. In contrast, Cluster 1 generally has a relatively high index in corruption and lowest indexes in other five attributes except generosity. The generosity is in supervise higher than the cluster 2's.

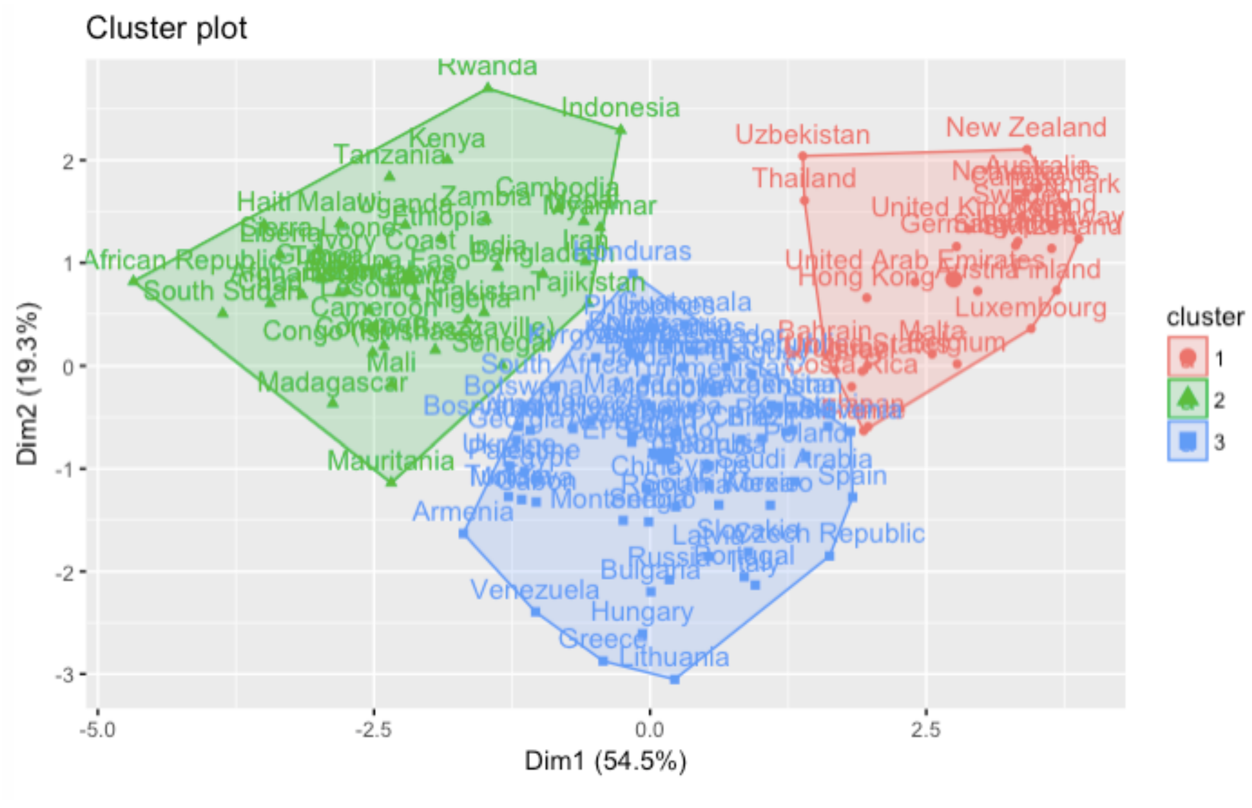
Model 4 K-means model with LifeLadder

```
## Choose the optimal k based on WSS
set.seed(123)
fviz_nbclust(happiness.norm, kmeans, method = "wss") +
  geom_vline(xintercept = 3, linetype = 2)
```



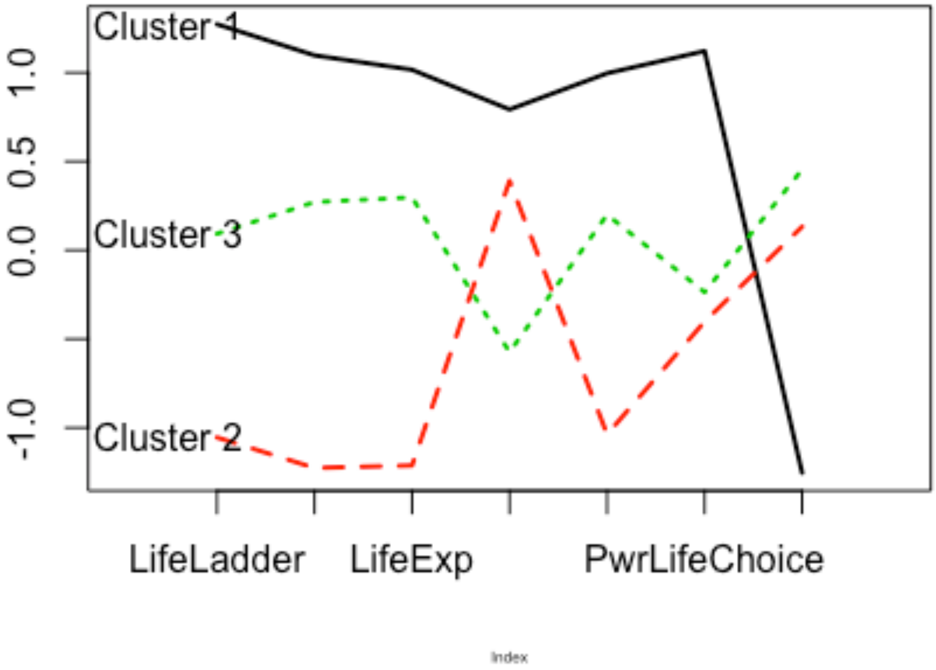
Based on elbow point in WSS, the k equals to 3 with LifeLadder. In later discussion it provides BSS/TSS=54.3%, higher than the k in k-means model without life ladder.

```
km1 <- kmeans(happiness.norm, 3)
fviz_cluster(km, happiness.norm)
```



```
km1

## K-means clustering with 3 clusters of sizes 29, 41, 68
##
## Cluster means:
##    LifeLadder    LnGDPpc     LifeExp Generosity PwrSocSupp PwrLifeChoice
## 1  1.27291777  1.0974951  1.0165406  0.7928399  0.9966661     1.1216921
## 2 -1.05425235 -1.2261693 -1.2122245  0.3908106 -1.0368722    -0.4018333
## 3  0.09279016  0.2712586  0.2973754 -0.5737587  0.2001242    -0.2360869
##   PwrCorruption
## 1    -1.2535017
## 2     0.1327160
## 3     0.4545616
##
## Clustering vector:
##   1: 1   2: 2   2: 3   2: 4   2: 5   3: 6   3: 7   2: 8   3: 9  2: 10
##      2      3      3      3      3      1      1      3      1      2
##  2: 11  3: 12  1: 13  2: 14  2: 15  2: 16  2: 17  2: 18  1: 19  1: 20
##      3      1      2      3      3      3      3      3      2      2
##  1: 21  3: 22  1: 23  1: 24  2: 25  2: 26  2: 27  1: 28  1: 29  2: 30
##      2      1      2      2      3      3      3      2      2      1
##  2: 31  2: 32  3: 33  2: 34  2: 35  2: 36  2: 37  3: 38  1: 39  3: 40
##      3      3      1      3      3      3      3      3      2      1
##  3: 41  2: 42  2: 43  3: 44  1: 45  2: 46  2: 47  1: 48  1: 49  1: 50
##      1      3      3      1      2      3      3      2      2      3
##  3: 51  2: 52  3: 53  2: 54  3: 55  1: 56  2: 57  3: 58  3: 59  2: 60
##       1       3       1       2       2       2       3       1       1       3
##   1: 61   3: 62   2: 63   2: 64   1: 65   2: 66   3: 67   2: 68   2: 69   2: 70
##       2       1       3       3       2       3       3       3       3       3
##   1: 71   1: 72   2: 73   2: 74   3: 75   2: 76   1: 77   1: 78   1: 79   3: 80
##       2       2       3       3       1       3       2       2       2       1
##   1: 81   2: 82   2: 83   2: 84   2: 85   2: 86   2: 87   2: 88   1: 89   3: 90
##       2       3       3       3       3       3       3       2       2       1
##   3: 91   2: 92   1: 93   2: 94   3: 96   1: 97   2: 98   2: 99  2: 100  2: 101
##       1       3       2       2       1       2       3       3       3       3
##  2: 102  2: 103  2: 104  2: 105  2: 106  1: 107  2: 108  1: 109  2: 110  1: 111
##       3       3       3       3       3       2       3       2       3       2
##  3: 112  2: 113  2: 114  2: 116  2: 117  1: 118  2: 119  3: 120  3: 121  2: 123
##       1       3       3       3       3       2       3       1       1       2
##  1: 124  3: 125  1: 126  2: 127  2: 128  2: 129  1: 130  2: 131  3: 132  3: 133
##       2       1       2       3       3       3       2       3       1       1
##  3: 134  3: 135  3: 136  2: 137  2: 138  1: 139  1: 140  1: 141
##       1       1       1       3       3       2       2       2
##
## Within cluster sum of squares by cluster:
## [1]  76.95656 139.38763 221.98106
##  (between_SS / total_SS =  54.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```r
## Plot an empty scatter plot
plot(c(0), xaxt = 'n', ylab = "", type = "l",
     ylim = c(min(km1$centers), max(km1$centers)), xlim = c(0, 8),cex.lab=0.
4)

## Label x-axes
axis(1, at = c(1:7), labels = names(happiness.norm))

## Plot centroids
for (i in c(1:3))
  lines(km1$centers[i,], lty = i, lwd = 2, col = i)

## Name clusters
text(x = 0.5, y = km1$centers[, 1], labels = paste("Cluster", c(1:3)))
```



The number of countries in each cluster respectively is 29, 41, 68, slightly different from kmeans model without life ladder. In Cluster 1, there are countries such as Austria, Canada, United States, United Kingdom and Japan, where the countries aregenerally more developed. In Cluster 2, there are countries such as Afghanistan, India, Nepal, Congo and Liberia, where the countries are less developed. In Cluster 3, there are countries such as Peru, Argentina, Spain, Vietnam and Turkey, where there is a mix of developed and less developed countries. The centroid profile plot is provided as below. The Cluster 1 generally has high life ladder,GDP per capital, life expenditure, generosity, social support and life choice but has a low corruption. The Cluster 2 generally is low in life ladder, GDP per capital, life expenditure, social support and life choice, and relatively high in generosity and corruption. The Cluster 3 is very similar to Cluster 1 in many attributes except generosity and corruption. It usually has a relatively low generosity and a relatively high corruption.

Impact of LifeLadder

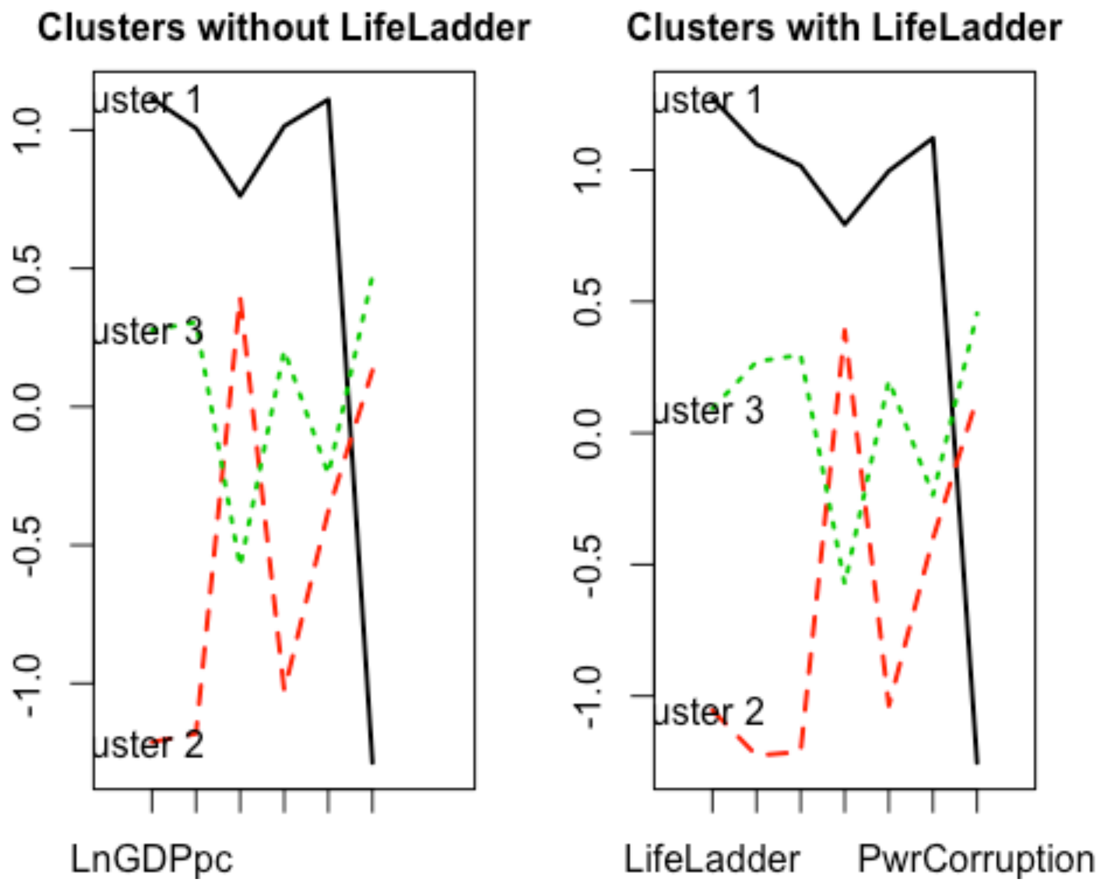
The impact of LifeLadder on the clustering is first measured by the sum of square index. But here with a new variable added, simply using the sum of squares will cause bias in comparison, so we decide to use the Between Sum of Square/Total Sum of Square to adjust the bias caused by different numbers of variables. We'll only look at the impact of LifeLadder on Kmean clustering because it's the better clustering under both circumstances.

```
## Compare two k-means models
km

## K-means clustering with 3 clusters of sizes 29, 42, 67
##
##
## Within cluster sum of squares by cluster:
## [1]  70.30303 129.89983 192.41012
##  (between_SS / total_SS =  52.2 %)
##
##
km1

## K-means clustering with 3 clusters of sizes 29, 41, 68
##
##
## Within cluster sum of squares by cluster:
## [1]  76.95656 139.38763 221.98106
##  (between_SS / total_SS =  54.3 %)
##
```

The BSS/TSS of the kmeans clustering without LifeLadder is 52.2% and The BSS/TSS of the kmeans clustering with LifeLadder is 54.3%. A higher BSS/TSS means a better clustering, so, although to a small extent, LifeLadder did improve the performance of our clustering, which means it contains information the other 6 variables don't.

```
opar<-par(no.readonly = TRUE)
par(mfrow=c(1, 2))
par(mar=rep(2,4))

## Plot an empty scatter plot
plot(c(0), xaxt = 'n', ylab = "", type = "l",
         ylim = c(min(km$centers), max(km$centers)), xlim = c(0, 8),cex.lab=
0.2,main = "Clusters without LifeLadder",cex.main=1)

## Label x-axes
axis(1, at = c(1:6), labels = names(happiness.norm2))

## Plot centroids
for (i in c(1:3))
    lines(km$centers[i,], lty = i, lwd = 2, col = i)

## Name clusters
text(x = 0.5, y = km$centers[, 1], labels = paste("Cluster", c(1:3)))

## Plot an empty scatter plot
plot(c(0), xaxt = 'n', ylab = "", type = "l",
         ylim = c(min(km1$centers), max(km1$centers)), xlim = c(0, 8),cex.lab
=0.2,main = "Clusters with LifeLadder",cex.main=1)
```

```
## Label x-axes
axis(1, at = c(1:7), labels = names(happiness.norm))

## Plot centroids
for (i in c(1:3))
    lines(km1$centers[i,], lty = i, lwd = 2, col = i)

## Name clusters
  text(x = 0.5, y = km1$centers[, 1], labels = paste("Cluster", c(1:3)))
```



```
par(opar)
```

From the above graph, we can easily tell that cluster 1 is “the happiest cluster”, cluster 3 is “the second happiest cluster” and cluster 2 is “the least happy” cluster, and that with or without LifeLadder, the clusterings are almost the same, which further proves that the other variables can already describe most of the levels of happiness. However, there are some little fixings LifeLadder did to make the clustering more precise. With LifeLadder, cluster 1, the obviously “happiest” cluster, has a new member, Costa Rica, and Estonia no longer belongs to cluster 1. From the data set we have, although both countries has a GDPpc higher than average, Costa Rica has significantly lower GDPpc and a much high Generosity level than Estonia. The case where a country with lower GDPpc and higher Generosity is happier is rare because LifeLadder has a very high correlation with GDPpc and very low correlation with Generosity. We can't make a conclusion based on this one case because we can always argue that people in Costa Rica may be easier to be satisfied than people in Estonia, but we can make a guess that although LifeLadder usually agrees with the other 6 variables, it sometimes captures factors related to culture and social environment that can be easily felt but are hard to measure.

Model comparison

We think that the advantages and disadvantages of these two models are: ** Hierarchical Model • Hierarchical model is easy to apply and shows a clear structure on how the choice of k can affect the cluster • Hierarchical model might not show good calculating ability when the dataset is large. • Hierarchical model might be allergic to extreme values. ** K-Means Model • K-means model is easy to apply and has a better clustering ability. • K-means model can provide direct numeric comparison on choices of k and don't have methods confusion. • k-means model is likely to be affected by seeds(although not shown in the case).

Best clustering

As discussed above, we choose k-means model with LifeLadder to be our best clustering model. The reasons are given below. As discussed above, we choose k-means model with LifeLadder to be our best clustering model. First, compared to hierachical model, k-means models provide more reasonable classifications on the dataset. We can see from the comparison of two different types of graphs that the k-means model manages to separate the countries more completely than the hierachical model does, which means that the 3 categories in k-means model are more typical to reflect the differences between countries. And when it comes to the question that whether LifeLadder should be included in the model, we think that LifeLadder is an extra supplementary that help us have a better view understanding the situation of certain countries. Although it may not directly separate some of the countries out of previous cluster, it will add another factor to show the differences between countries. For example, the US does not have high LifeLadder compared to other developed countries in Europe (they all have relative high GDP). So that we can generate a conclusion that there are differences between developed countries and GDP is not an absolute factor used to classify different clusters.

Conclusion

In conclusion, we can see from the "2017 World Happiness Report" that for the most happy countries, they share a few common features and are alike in some ways. However, for countries with not highest happiness index, they are more likely to be different from each other.