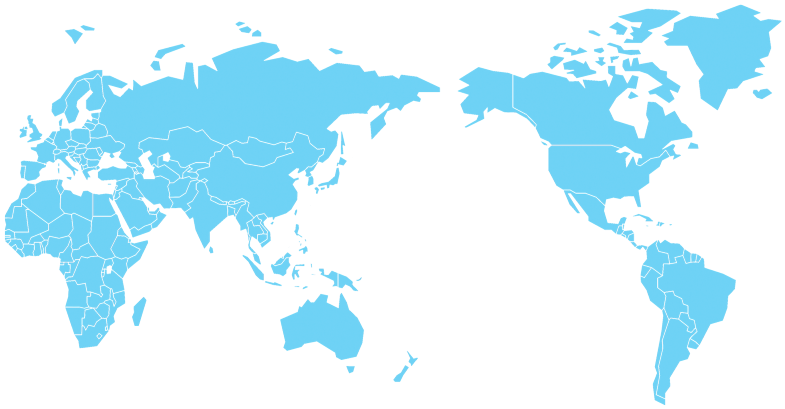


MGSC 661 Individual Project

Identifying Countries Most In Need

Diwei Zhu (260761307)

1 Introduction

As it has been proved, INGOs (International Non-governmental Organizations) can help economic growth and participate in improving public health in developing countries through providing access to capital, supporting entrepreneurship, providing development assistance for health (DAH), and deploying technical cooperation (Murdie & Kakietek, 2012; Laaser & Brand, 2014). Researchers revealed that, in terms of common measures of public health like life expectancy and infant mortality rates, national income level, gross domestic product per capita (GDPP), and per capita spending on healthcare are of significant influential power (Biggs et al., 2010). On the other side, impaired health, and accordingly, impaired life expectancy, exacerbate poverty and undermine development. For example, malaria has lowered the life expectancy of tropical countries, and has been deleterious to economic development by causing shrinks in labour forces and adding burdens to healthcare sectors (Fosu, 2007).

Therefore, an integration of socio-economic and healthcare investments from INGOs acts as a strong power that alleviates public health conditions in developing countries. However, the helping capabilities of INGOs are limited, demanding a wise allocation of resources and a prudent selection of help receivers. The identification of countries that are most in need of help is of high necessity.

Inspired by the above facts, this project set two goals in sequence. The first goal is to re-verify and rank the influential powers of socio-economic and public health factors on life expectancy. With the importance of the factors revealed, we, therefore, will be more knowledgeable about which tool to pick when providing help to developing countries with respect to improving life expectancy levels. The second goal is to classify countries in terms of the influential factors identified above in order to locate the countries that are in most need of supports.

2 Data Description

2.1 Dataset

The dataset this project use contains 5 socio-economic and 4 public health factors of 167 countries.

Socio-economic factors:

gdpp: the GDP per capita. Calculated as the total GDP divided by the total population.

income: net income per person

import: imports of goods and services per capita. Given as % of the gdpp

export: exports of goods and services per capita. Given as % of the gdpp

inflation: the measurement of the annual growth rate of the total GDP

Health factors:

life_exp (target variable): the average number of years a newborn child would live if the current mortality patterns are to remain the same

child_mort: death of children under 5 years of age per 1000 live births

health_spending: total health spending per capita. Given as percentage of GDP per capita

fertility: the number of children that would be born to each woman if the current age-fertility rates remain the same

2.2 Data distribution and correlation

Based on the boxplots (Appendix Fig 1) generated based on each numerical variable (except for *country*). Except for *life_exp* and *fertility*, other variables have low mean val-

ues, have the majority of countries concentrating at lower values, and have outliers of higher values. The outliers, either indicate good socio-economic performances in gdpp, per capita income, and spending on healthcare, or the abnormally high inflation rates, or extremely high infant mortality rate. The target variable, *life_exp*, has 50% of observations concentrated from 65.3 years to 76.8 years (the 25^{th} percentile to the 75^{th} percentile), and three outliers (Haiti, Lesotho, Central African Republic) with very low life expectancy (Appendix Fig 2).

The pairwise correlations between numerical variables were verified and displayed in a correlation matrix (Appendix Fig 3 & 4). *gdpp* and income have a strong positive correlation (0.8956), so as *child_mort* and fertility (0.8485); *life_exp* and *child_mort* have a strong negative correlation (-0.8867). In later feature selection and data analysis, the strong positive and negative correlations between variables will be considered.

3 Model Selection & Methodology

3.1 Feature selection through PCA and Random Forest

With a cleaned dataset where the non-numerical variable country is dropped and the numerical variables are standardized, we conducted the Principal Component Analysis (PCA). According to the 2D PCA plot shown in Appendix Fig 5, the colinearity between *gdpp* and *income*, and between *fertility* and *child_mort*, as shown in the correlation matrix, are proved by PCA because we observe overlaps of the arrows of the two variable pairs. import and export form a 60 degrees angle and a 80 degrees angle with *life_exp* respectively. Although the two arrows representing import and export are not strictly orthogonal to the arrow of *life_exp*, the angles still indicate the two factors' processing weak influential powers on *life_exp*. Pointing to the opposite direction as does the arrow of *life_exp*, *fertility* and *child_mort*, along with inflation, negatively correlate with *life_exp*; pointing to the similar direction as does the arrow of *life_exp*, *health_spending*

and *gdpp*/*income* are mostly correlated to *life_exp*. In this sense, *fertility*/*child_mort*, *inflation*, *health_spending*, and *gdpp*/*income* can be helpful predictors. Appendix Fig 6 showed the PCA plot with only countries that have a *life_exp* value that is higher than the mean value (73.1 years). Except for two outliers, most of the observations concentrate at the top-right of the plot, proving the positive correlations of *life_exp* with *health_spending* and *gdpp*/*income*.

In order to further rank the importance of the features and select the predictors, we built a Random Forest model with:

$$life\_exp = f(exports + imports + health\_spending + inflation + income + gdpp + fertility) \quad (1)$$

child_mort was not included in the model, because it has high colinearity with the target variable. With 10000 "trees" grown in the "forest", the feature importance plot in Appendix Fig 7 shows that *fertility*, *income*, and *gdpp* are of the most importance: if remove fertility from the model, MSE will increase by 116.11% and node impurity would increase by 3843.92; if remove *income* or *gdpp* from the model, MSE will increase by 73.67% and 69.68% respectively, and node impurity would increase by 3203.09 and 2782.65 respectively. Considering the colinearity indicated by the correlation table and the PCA plot, we include *income* as a predictor and discard *gdpp* because the former has a larger influence on the target variable.

Therefore, in later models selection part, we use only *income* and *fertility* as predictors.

$$life\_exp = f(income + fertility) \quad (2)$$

3.2 Tree-based models

Reacting to the first goal of this project, with the predictors selected above, we constructed two tree-based predictive models to verify if the predictors are effective in predicting the *life_exp* values of a country.

The first model is a Decision Tree model. The Decision Tree model splits the tree nodes to minimize the residual sum of squares (RSS), being able to visualize the data division by representing this process as a tree. To minimize the out-of-performance error and avoid overfitting, we picked cp=0.01 as the branching cp threshold, requiring that only when the split reduces the RSS by 1% that the split can be created. Notably, this model in this project is not the main prediction tool, it yet serves as a visualization tool to indicate how countries with different life expectancy values are segmented in terms of income levels and fertility rates.

The second one is a Boosting model that serves as the main predictive model. The boosting model creates large numbers of "simple trees" that grow sequentially while learning from previous ones, and joins all the trees to form up an intelligent final result. We created a boosting forest with 10000 simple trees where each tree has only 4 internal nodes, as the final model that predicts life expectancy of a country with income level and fertility rate inputs.

3.3 Clustering

Responding to the second goal that requires a partitioning of countries and an identification of the countries that are in most need of help, we used the clustering technique. To decide the number of clusters, we applied Elbow and Silhouette method (Appendix Fig 9) and the result plots indicate that both the number of clusters k=2 or k=4 are good choices. Considering the complexities and disparities between countries, it is undoubted that partitioning the countries into 4 clusters is closer to reality. With k=4, we built the

clusters with K-Means Clustering Algorithm.

4 Results

4.1 Tree-based models result

The tree generated by the Decision Tree model (Appendix Fig 8) contains 5 split points among which three split the tree based on income levels and two based on fertility rates. Firstly, whether the fertility rate is higher or lower than 3.3 splits the countries into two branches: one branch contains 31% of the countries with an average *life_exp* at 61 years; the other branch, representing countries with high life expectancy, contains 69% of the observations with an average *life_exp* at 75 years. At the second layer, *income* becomes the classifying predictor at both nodes. As required by the goals, we set our eyes on the subgroups with low *life_exp* values. At the subgroup where the fertility rate is high (on the left), the criterion that whether income level is higher than 3300 divides the countries into two subgroups, one of which contains 22% of the countries with an average *life_exp* at 58 years and the other contains 10% of the countries with an average *life_exp* at 66 years. The third split node of the low life expectancy group is again income level. 11% of the countries where the income level is lower than 1635 and fertility rate is larger or equal to 3.3 have the lowest average *life_exp* of only 56 years. The MSE value at each node varies from 3.94 to 78.61.

The boosting model, as the main predictive model, returned a better MSE value at 0.016. The model also generated a summary of the relative importance of the predictors: fertility has a relative importance of 50.67 and income has a score of 49.33, meaning that the two predictors impose approximately equal influence on the predictive result of the model.

4.2 Clustering result

As for the clustering model, the centroids regarding to each variable of each cluster are shown in the Fig 10 in the Appendix. In clusters 1, 2, and 3, there are 36, 84, and 47 countries respectively. Interpreting from the centroids, cluster 1 contains the countries where *child_mort* (5.00), *inflation* (2.67%) and *fertility* (1.75) are the lowest and *health_spending* (8.81%), *income* (45672.22), *gdpp* (42494.44) and *life_exp* (80.13) are the highest. On the contrary, in cluster 3, the centroids are the highest in *child_mort* (92.96), *inflation* (12.02%) and *fertility* (5.01) and are the lowest in *income* (3942.40), *gdpp* (1922.38) and *life_exp* (59.19). Countries in cluster 3 are countries in between the two extreme clusters.

To select the countries that we recommend INGOs to provide with socio-economic and public health supports, we queried 5 countries with the lowest per capita income levels from cluster 1 (Appendix Fig 11). The countries are: Democratic Republic of Congo, Liberia, Burundi, Niger, and Central African Republic. All selected countries have high children mortality rate (*child_mort* >= 89.3), high fertility rate (*fertility* >= 5.02), and low life expectancy (*life_exp* <= 60.8). Democracy Republic of Congo has the lowest per capita income level that is only 13% of the centroid income level of cluster 2.

5 Classification/prediction and Conclusions

The above rates proved that the socio-economic factor, per capita income level, as well as fertility rate, have influential power on the life expectancy of countries. Higher per capita income level indicates intaking of better food and obtaining higher life standard that contributes to elongation of human life spans. A higher fertility rate is related to the condition in which parents are not able to provide enough care to each child among 6 or even 7 children in one family, thus increasing the children mortality rate and decreasing the expected average number of years a newborn would live.

The clustering model successfully divided the countries into three clusters with clear and differentiated cluster centroids. On top of fertility rate and per capita income level, other features show the pattern that short life expectancy relates to unsatisfactory socio-economic factors (gdpp, import, export, and inflation rate) as well as poor public health conditions (children mortality rate and spendings on healthcare). With the clusters, we identified the 5 countries that suffering from dreadful socio-economic and health conditions the most that could be candidates to receive supports from INGOs.

As for the extensional use of the models constructed in this project, the predictive Boosting model can be used for predicting the life expectancy of countries at different times based on current per capita income levels and fertility rates on hand. Also, the clustering model, based on different criteria, has the potential to identify countries that need helps in terms of different socio-economic/healthcare indicators. Lastly, this project shows the mutual connection between socio-economic factors and public health conditions, highlighting the leverage points–per capita income levels and fertility rates–that international support providers can focus on to improve the life expectancy of countries efficiently.

6 References

1. Biggs, B., King, L., Basu, S., & Stuckler, D. (2010). Is wealthier always healthier? The impact of national income level, inequality, and poverty on public health in Latin America. Social Science & Medicine, 71(2), 266–273. https://doi.org/10.1016/j.socscimed.2010.04.002

2. Fosu, A. K. (2007, October 1). SciELO - Saúde Pública - Poverty and development Poverty and development. SciELO Public Health. https://www.scielosp.org/article/bwho/2007.v85n10/734734/en/

3. Laaser, U., & Brand, H. (2014). Global health in the 21st century. Global Health Action, 7(1), 23694. https://doi.org/10.3402/gha.v7.23694

4. Murdie, A. M., & Kakietek, J. (2012). Do Development INGOs Really Work? The Impact of International Development NGOs on Human Capital and Economic Growth. Journal of Sustainable Society, 01(01). https://doi.org/10.11634/21682585140356

7 Appendix

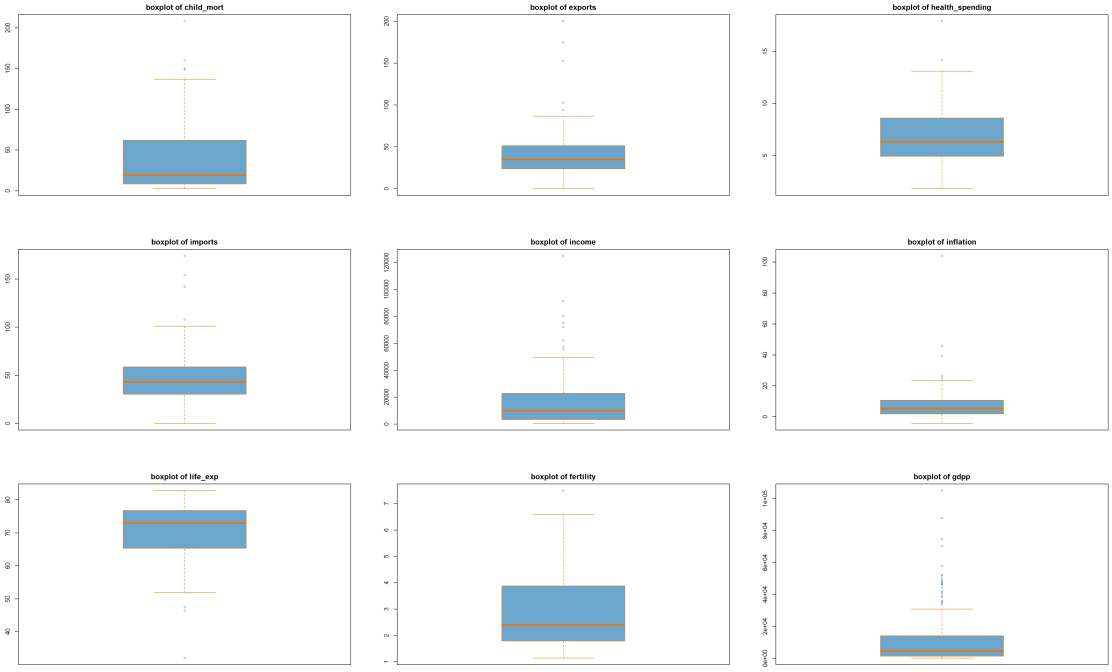


Figure 1: Boxplot of numerical variables

country <chr>	life_exp <dbl>
Haiti	32.1
Lesotho	46.5
Central African Republic	47.5

3 rows

Figure 2: Countries with low life expectancy

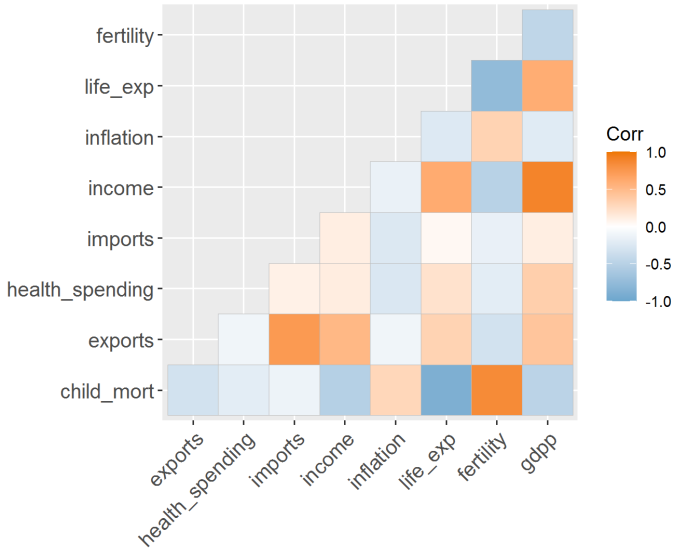


Figure 3: Correlation plot

	child_mort	exports	health_spending	imports	income	inflation	life_exp	fertility	gdpp
child_mort	1.0000	-0.3181	-0.2004	-0.1272	-0.5243	0.2883	-0.8867	0.8485	-0.4830
exports	-0.3181	1.0000	-0.1144	0.7374	0.5168	-0.1073	0.3163	-0.3200	0.4187
health_spending	-0.2004	-0.1144	1.0000	0.0957	0.1296	-0.2554	0.2107	-0.1967	0.3460
imports	-0.1272	0.7374	0.0957	1.0000	0.1224	-0.2470	0.0544	-0.1590	0.1155
income	-0.5243	0.5168	0.1296	0.1224	1.0000	-0.1478	0.6120	-0.5018	0.8956
inflation	0.2883	-0.1073	-0.2554	-0.2470	-0.1478	1.0000	-0.2397	0.3169	-0.2216
life_exp	-0.8867	0.3163	0.2107	0.0544	0.6120	-0.2397	1.0000	-0.7609	0.6001
fertility	0.8485	-0.3200	-0.1967	-0.1590	-0.5018	0.3169	-0.7609	1.0000	-0.4549
gdpp	-0.4830	0.4187	0.3460	0.1155	0.8956	-0.2216	0.6001	-0.4549	1.0000

Figure 4: Correlation table

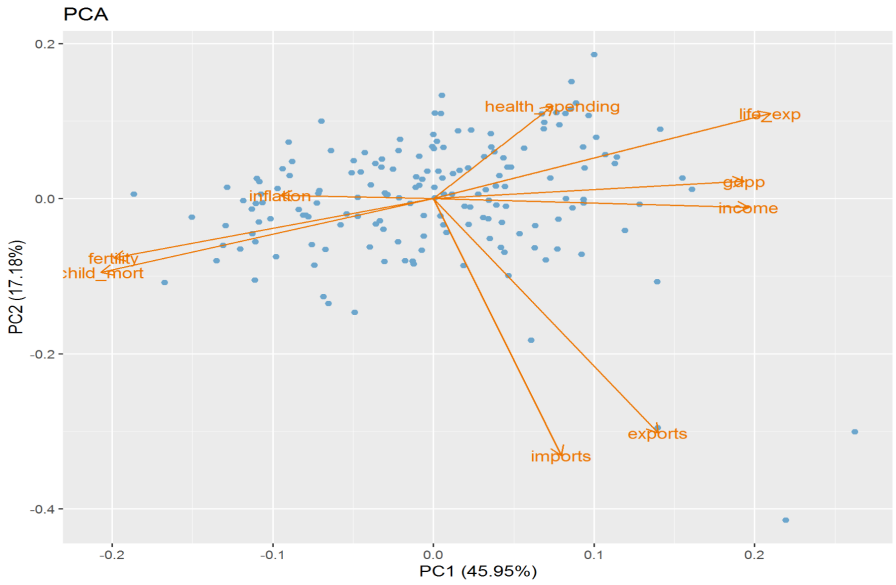


Figure 5: PCA plot

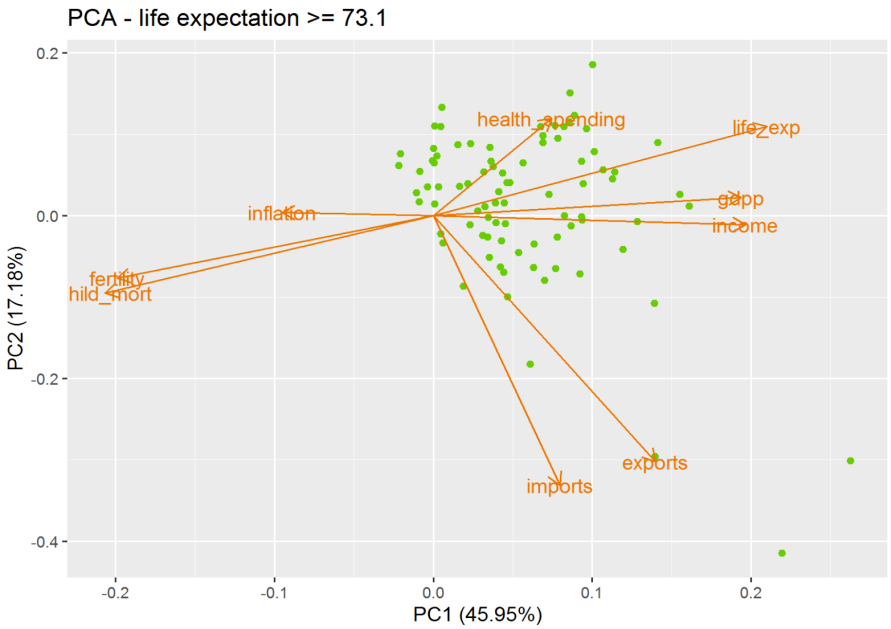


Figure 6: PCA plot showing only observations with life expectancy higher than 73.1 years

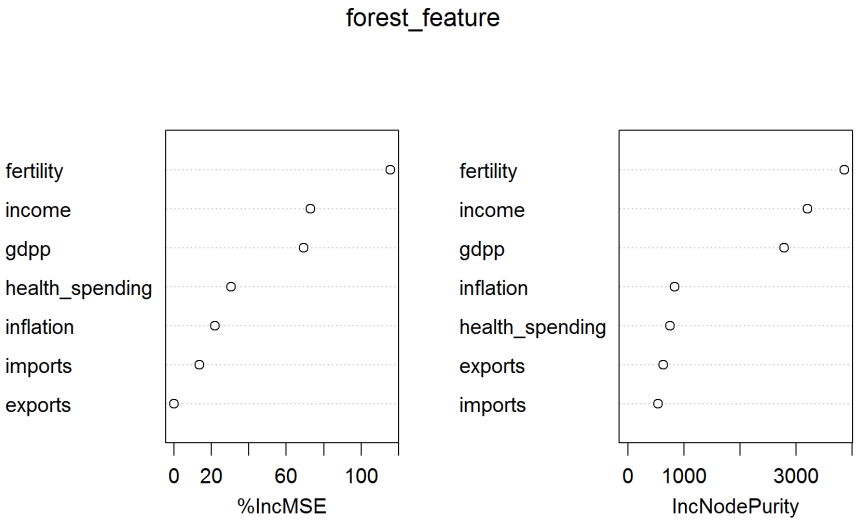


Figure 7: Random Forest feature importance

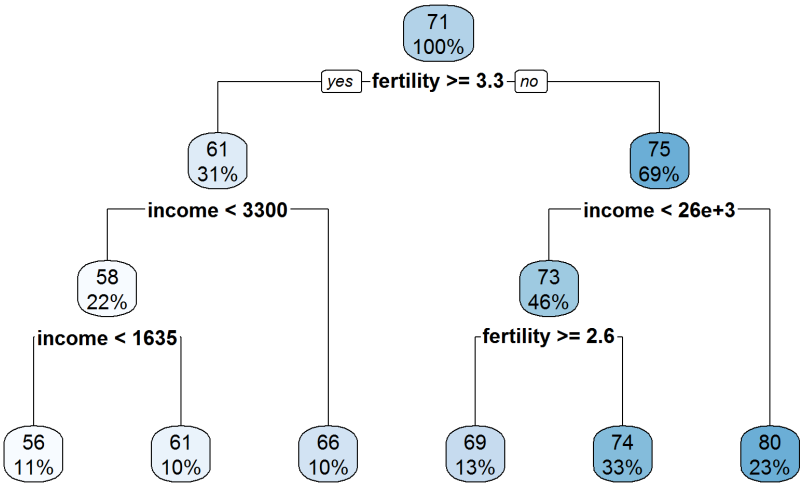


Figure 8: Decision tree visualization

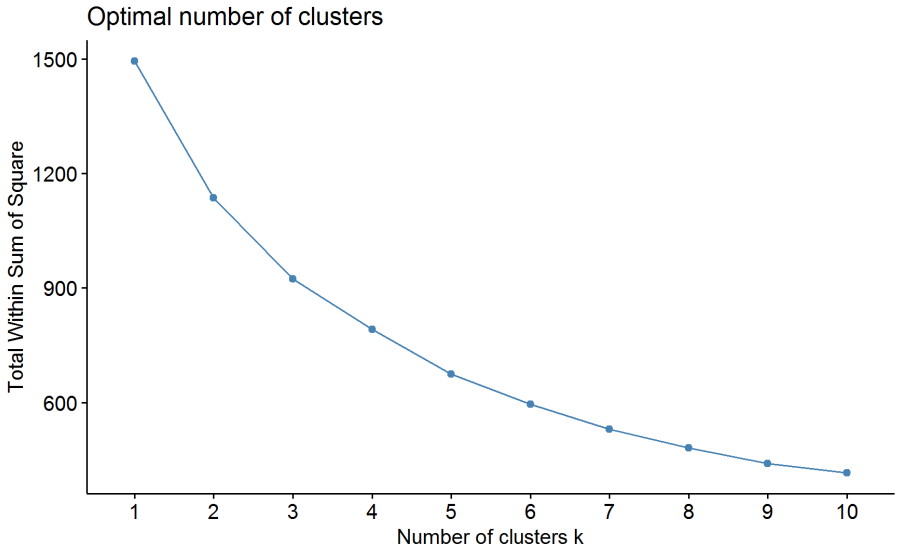


Figure 9: Elbow plot

cluster <int>	n <int>	child_mort <dbl>	exports <dbl>	health_spending <dbl>	imports <dbl>	income <dbl>	inflation <dbl>	life_exp <dbl>	fertility <dbl>	gdpp <dbl>
1	36	5.00000	58.73889	8.807778	51.49167	45672.222	2.671250	80.12778	1.752778	42494.444
2	84	21.92738	40.24392	6.200952	47.47340	12305.595	7.600905	72.81429	2.307500	6486.452
3	47	92.96170	29.15128	6.388511	42.32340	3942.404	12.019681	59.18723	5.008085	1922.383

Figure 10: Centroids of clusters

country <chr>	income <int>	fertility <dbl>	child_mort <dbl>	life_exp <dbl>
Congo, Dem. Rep.	609	6.54	116.0	57.5
Liberia	700	5.02	89.3	60.8
Burundi	764	6.26	93.6	57.7
Niger	814	7.49	123.0	58.8
Central African Republic	888	5.21	149.0	47.5

Figure 11: Selected countries for INGO supports

Final Project – R Coding

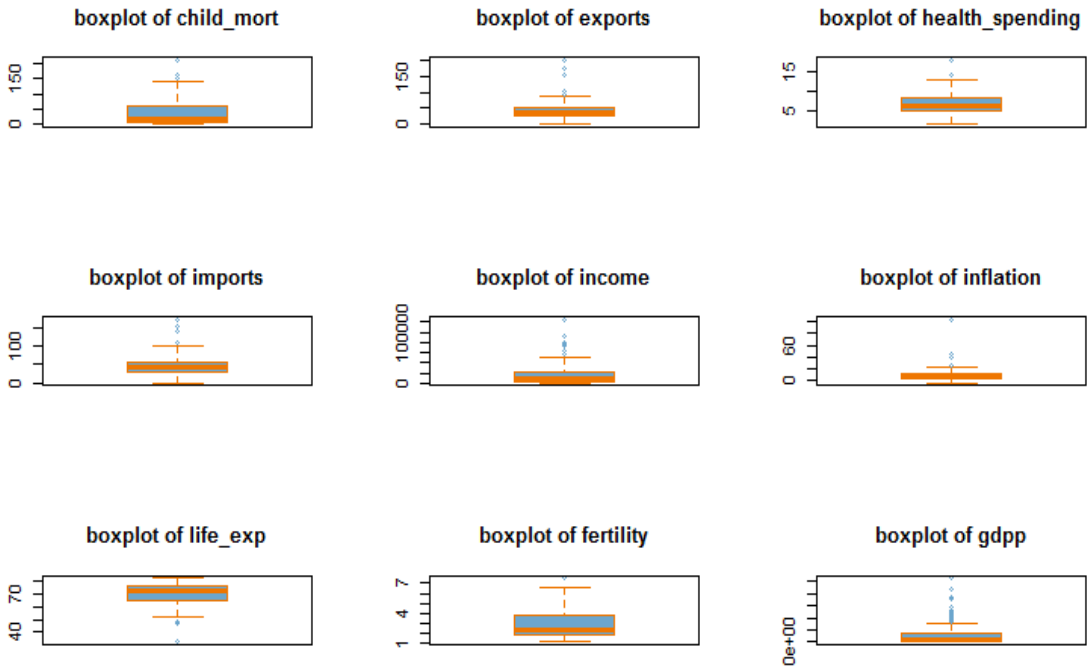
Diwei Zhu

2021/12/10

```
# install.packages("dplyr")
# install.packages("ggcorrplot")
# install.packages("GGally")
# install.packages("ggfortify")
# install.packages("cluster")
# pkgs <- c("factoextra",  "NbClust")
# install.packages(pkgs)
# install.packages("gbm")
# install.packages("sqldf")
library(dplyr)
library(ggcorrplot)
library(ggplot2)
library(GGally)
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

library(ggfortify)
library(cluster)
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at htt
ps://goo.gl/ve3WBa

library(NbClust)
library(tree)
library(rpart)
library(rpart.plot)
library(randomForest)
library(gbm)
library(sqldf)


df <- read.csv("C:/Users/admin/Desktop/Country-data.csv")
colnames(df)

##  [1] "country"         "child_mort"      "exports"         "health_s
pending"
##  [5] "imports"         "income"          "inflation"       "life_exp
"
##  [9] "fertility"       "gdpp"

attach(df)
```

SECTION 1 - Dataset exploration

```
# get list of column names
colname <- colnames(df)
col_list <- c()
for (i in 1:10) {
  x = strsplit(colname,"[[:space:]]")[[i]][1]
  col_list <- c(col_list, x)
}

# getting boxplots
par(mfrow=c(3,3))
boxplot(child_mort, main = "boxplot of child_mort", col="skyblue3", out
col="skyblue3", boxcol="darkorange2", whiskcol="darkorange2", medcol="d
arkorange2", staplecol="darkorange2")

boxplot(exports, main = "boxplot of exports", col="skyblue3", outcol="s
kyblue3", boxcol="darkorange2", whiskcol="darkorange2", medcol="darkora
nge2", staplecol="darkorange2")

boxplot(health_spending, main = "boxplot of health_spending", col="skyb
lue3", outcol="skyblue3", boxcol="darkorange2", whiskcol="darkorange2",
 medcol="darkorange2", staplecol="darkorange2")

boxplot(imports, main = "boxplot of imports", col="skyblue3", outcol="s
kyblue3", boxcol="darkorange2", whiskcol="darkorange2", medcol="darkora
nge2", staplecol="darkorange2")

boxplot(income, main = "boxplot of income", col="skyblue3", outcol="sky
blue3", boxcol="darkorange2", whiskcol="darkorange2", medcol="darkorang
e2", staplecol="darkorange2")

boxplot(inflation, main = "boxplot of inflation", col="skyblue3", outco
l="skyblue3", boxcol="darkorange2", whiskcol="darkorange2", medcol="dar
korange2", staplecol="darkorange2")

boxplot(life_exp, main = "boxplot of life_exp", col="skyblue3", outcol=
"skyblue3", boxcol="darkorange2", whiskcol="darkorange2", medcol="darko
range2", staplecol="darkorange2")

boxplot(fertility, main = "boxplot of fertility", col="skyblue3", outco
l="skyblue3", boxcol="darkorange2", whiskcol="darkorange2", medcol="dar
korange2", staplecol="darkorange2")

boxplot(gdpp, main = "boxplot of gdpp", col="skyblue3", outcol="skyblue
3", boxcol="darkorange2", whiskcol="darkorange2", medcol="darkorange2",
 staplecol="darkorange2")
```




```
# summary of life_exp
summary(life_exp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32.10   65.30   73.10   70.56   76.80   82.80

# countries with low life_exp
sqldf("SELECT country, life_exp FROM df
      ORDER BY life_exp
      LIMIT 3")

##                    country life_exp
## 1                    Haiti     32.1
## 2                  Lesotho     46.5
## 3 Central African Republic     47.5

# get all numerical variables
numerical_var <- select_if(df, is.numeric)

# build correlation matrix and visualize
corr = cor(numerical_var)
corr <- round(corr, 4)
corr

##                child_mort exports health_spending imports  income i
nflation
## child_mort          1.0000 -0.3181         -0.2004 -0.1272 -0.5243
  0.2883
```

```
## exports            -0.3181  1.0000         -0.1144  0.7374  0.5168
 -0.1073
## health_spending    -0.2004 -0.1144          1.0000  0.0957  0.1296
 -0.2554
## imports            -0.1272  0.7374          0.0957  1.0000  0.1224
 -0.2470
## income             -0.5243  0.5168          0.1296  0.1224  1.0000
 -0.1478
## inflation           0.2883 -0.1073         -0.2554 -0.2470 -0.1478
  1.0000
## life_exp           -0.8867  0.3163          0.2107  0.0544  0.6120
 -0.2397
## fertility           0.8485 -0.3200         -0.1967 -0.1590 -0.5018
  0.3169
## gdpp               -0.4830  0.4187          0.3460  0.1155  0.8956
 -0.2216
##                  life_exp fertility    gdpp
## child_mort        -0.8867    0.8485 -0.4830
## exports            0.3163   -0.3200  0.4187
## health_spending    0.2107   -0.1967  0.3460
## imports            0.0544   -0.1590  0.1155
## income             0.6120   -0.5018  0.8956
## inflation         -0.2397    0.3169 -0.2216
## life_exp           1.0000   -0.7609  0.6001
## fertility         -0.7609    1.0000 -0.4549
## gdpp               0.6001   -0.4549  1.0000
```

```
ggcorrplot(corr, type="lower", ggtheme = ggplot2::theme_gray, colors =
c("skyblue3", "white", "darkorange2"))
```

SECTION 2 - Data cleaning

```
# drop the non-numerical column "country"
drops <- "country"
df_num <- df[ , !(names(df) %in% drops)]
head(df_num)
```

```
##   child_mort exports health_spending imports income inflation life_e
xp
## 1       90.2    10.0            7.58    44.9   1610      9.44     56
.2
## 2       16.6    28.0            6.55    48.6   9930      4.49     76
.3
## 3       27.3    38.4            4.17    31.4  12900     16.10     76
.5
## 4      119.0    62.3            2.85    42.9   5900     22.40     60
.1
## 5       10.3    45.5            6.03    58.9  19100      1.44     76
.8
## 6       14.5    18.9            8.10    16.0  18700     20.90     75
.8
##   fertility  gdpp
## 1      5.82   553
## 2      1.65  4090
## 3      2.89  4460
## 4      6.16  3530
## 5      2.13 12200
## 6      2.37 10300
```

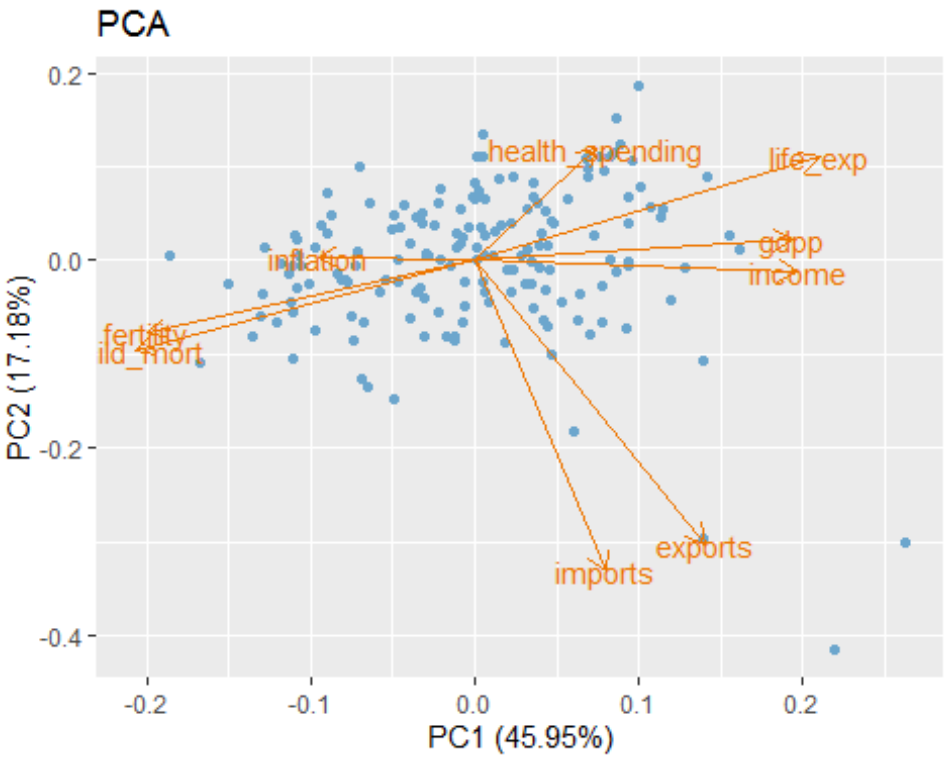
```
# scaling numeric variables
df_std <- scale(df_num)
head(df_std)
```

```
##      child_mort     exports health_spending     imports      income
 inflation
## [1,]  1.2876597 -1.13486665      0.27825140 -0.08220771 -0.80582187
 0.1568645
## [2,] -0.5373329 -0.47822017     -0.09672528  0.07062429 -0.37424335
-0.3114109
## [3,] -0.2720146 -0.09882442     -0.96317624 -0.63983800 -0.22018227
 0.7869076
## [4,]  2.0017872  0.77305618     -1.44372888 -0.16481961 -0.58328920
 1.3828944
## [5,] -0.6935483  0.16018613     -0.28603389  0.49607554  0.10142673
-0.5999442
## [6,] -0.5894047 -0.81019144      0.46756001 -1.27594958  0.08067776
 1.2409928
##        life_exp   fertility        gdpp
## [1,] -1.6142372  1.89717646 -0.67714308
## [2,]  0.6459238 -0.85739418 -0.48416709
## [3,]  0.6684130 -0.03828924 -0.46398018
```

```
## [4,] -1.1756985  2.12176975 -0.51472026
## [5,]  0.7021467 -0.54032130 -0.04169175
## [6,]  0.5897009 -0.38178486 -0.14535428

# Use standardized data for clustering
```

SECTION 3 - PCA

```
pca <- prcomp(df_num, scale=TRUE)
pca

## Standard deviations (1, .., p=9):
## [1] 2.0336314 1.2435217 1.0818425 0.9973889 0.8127847 0.4728437 0.33
68067
## [8] 0.2971790 0.2586020
##
## Rotation (n x k) = (9 x 9):
##                          PC1          PC2         PC3          PC4
     PC5
## child_mort      -0.4195194 -0.192883937  0.02954353 -0.370653262  0.
16896968
## exports          0.2838970 -0.613163494 -0.14476069 -0.003091019 -0.
05761584
## health_spending  0.1508378  0.243086779  0.59663237 -0.461897497 -0.
51800037
## imports          0.1614824 -0.671820644  0.29992674  0.071907461 -0.
25537642
## income           0.3984411 -0.022535530 -0.30154750 -0.392159039  0.
24714960
## inflation       -0.1931729  0.008404473 -0.64251951 -0.150441762 -0.
71486910
## life_exp         0.4258394  0.222706743 -0.11391854  0.203797235 -0.
10821980
## fertility       -0.4037290 -0.155233106 -0.01954925 -0.378303645  0.
13526221
## gdpp             0.3926448  0.046022396 -0.12297749 -0.531994575  0.
18016662
##                         PC6         PC7         PC8         PC9
## child_mort      -0.200628153  0.07948854  0.68274306  0.32754180
## exports          0.059332832  0.70730269  0.01419742 -0.12308207
## health_spending -0.007276456  0.24983051 -0.07249683  0.11308797
## imports          0.030031537 -0.59218953  0.02894642  0.09903717
## income          -0.160346990 -0.09556237 -0.35262369  0.61298247
## inflation       -0.066285372 -0.10463252  0.01153775 -0.02523614
## life_exp         0.601126516 -0.01848639  0.50466425  0.29403981
## fertility        0.750688748 -0.02882643 -0.29335267 -0.02633585
## gdpp            -0.016778761 -0.24299776  0.24969636 -0.62564572

autoplot(pca, data = df_std, loadings = TRUE, loadings.label = TRUE, co
l="skyblue3", loadings.colour= "darkorange2", loadings.label.colour = "
darkorange2", main = "PCA")
```



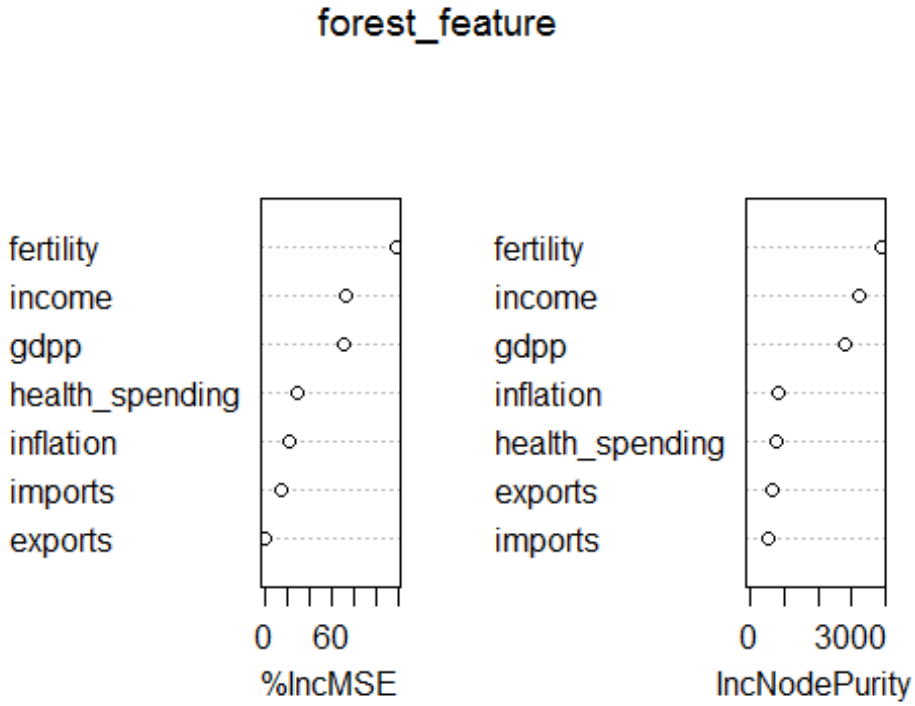

```
# imports/exports may can be not considered (orthogonal)
# positive: gdpp/income, health_spending
# negative: fertility/child_mort, inflation
```

SECTION 4 - Tree-based method to verify the powers of the features in predicting life_exp

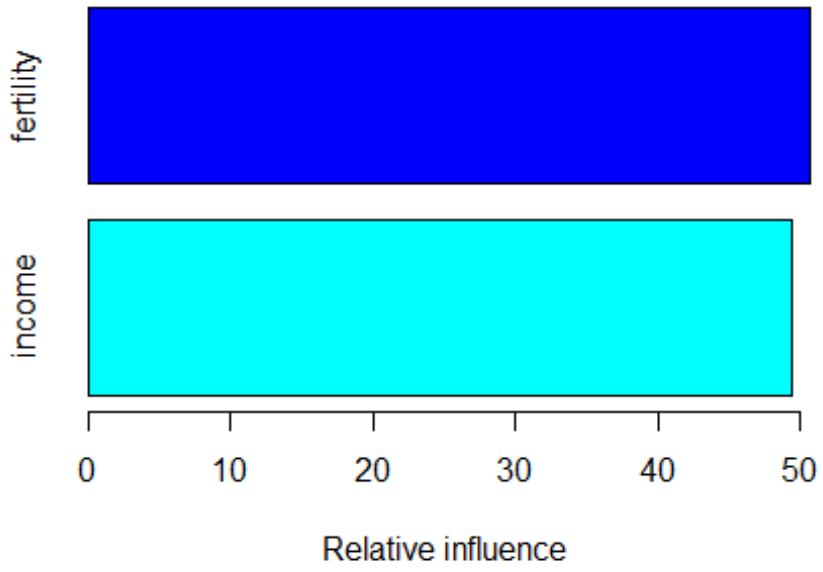
```
# Random Forest
forest_feature = randomForest(life_exp~exports+imports+health_spending+
inflation+income+gdpp+fertility, ntree=10000, data=df_num, importance=T
RUE, na.action=na.omit)

forest_feature
```

```
##
## Call:
##  randomForest(formula = life_exp ~ exports + imports + health_spendi
ng +      inflation + income + gdpp + fertility, data = df_num, ntree =
 10000,      importance = TRUE, na.action = na.omit)
##                Type of random forest: regression
##                      Number of trees: 10000
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 25.52809
##                     % Var explained: 67.53
```

```
# return importance
importance(forest_feature)

##                     %IncMSE IncNodePurity
## exports           1.151482      638.2498
## imports          14.847999      532.6608
## health_spending  30.327718      751.7241
## inflation        23.080455      841.4998
## income           72.443262     3209.7118
## gdpp             70.224387     2772.4378
## fertility       117.374979     3860.2670

varImpPlot(forest_feature)
```




```
# MSE = 36

# Boosting
set.seed(1)
boosted = gbm(life_exp~income+fertility, data=df_num, distribution="gau
ssian",n.trees=10000, interaction.depth=4)
summary(boosted)
```




```
##                   var  rel.inf
## fertility fertility 50.67242
## income       income 49.32758
```

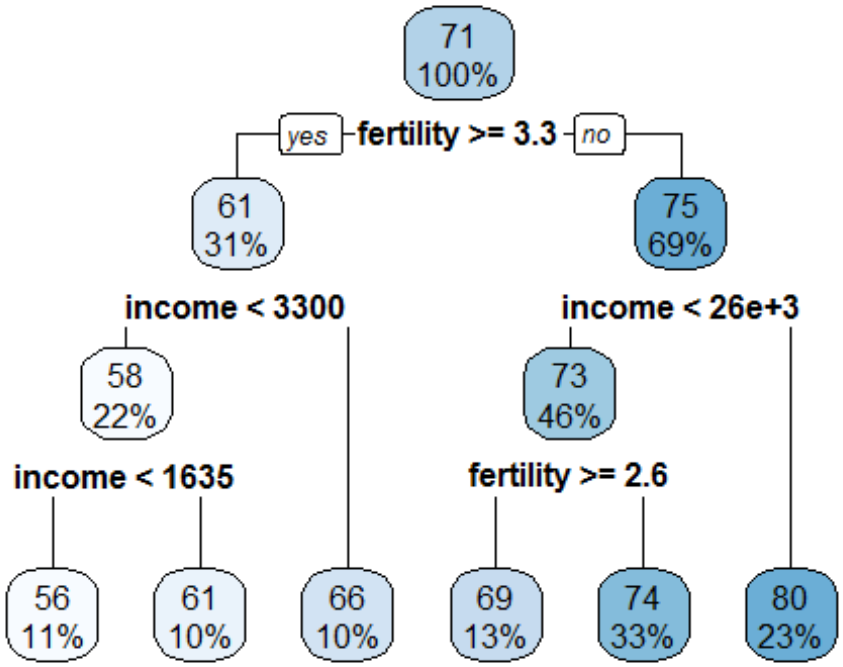
```
# get MSE
predicted_score2=predict(boosted, newdata=df_num, n.trees=10000)
MSE <- mean((predicted_score2 - life_exp)^2)
MSE
```

```
## [1] 0.01594092
```

```
# build decision tree
mytree=rpart(life_exp~income+fertility,control=rpart.control(cp=0.01))

# verify optimal cp value
opt_cp = mytree$cptable[which.min(mytree$cptable[,"xerror"]),"CP"]
print(opt_cp)
```

```
## [1] 0.01
```

```
# print result and draw the tree
rpart.plot(mytree)
```



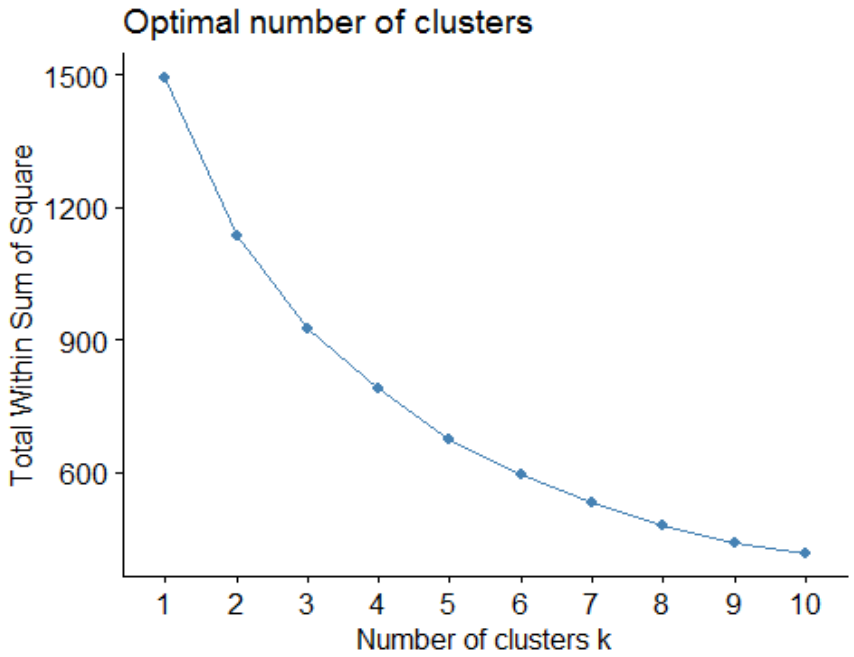

```
summary(mytree)

## Call:
## rpart(formula = life_exp ~ income + fertility, control = rpart.contr
ol(cp = 0.01))
##   n= 167
##
##           CP nsplit rel error    xerror       xstd
## 1 0.56218849      0 1.0000000 1.0070849 0.13743745
## 2 0.10291111      1 0.4378115 0.5106222 0.08855553
## 3 0.04521877      2 0.3349004 0.4056803 0.08090816
## 4 0.03526809      3 0.2896816 0.3893436 0.07307674
## 5 0.01144472      4 0.2544135 0.3484388 0.06872534
## 6 0.01000000      5 0.2429688 0.3435662 0.07421586
##
## Variable importance
## fertility    income
##        53        47
##
## Node number 1: 167 observations,    complexity param=0.5621885
##   mean=70.55569, MSE=78.61492
##   left son=2 (52 obs) right son=3 (115 obs)
##   Primary splits:
##       fertility < 3.25  to the right, improve=0.5621885, (0 missing)
##       income    < 6160  to the left,  improve=0.5253823, (0 missing)
##   Surrogate splits:
##       income < 3545  to the left,  agree=0.898, adj=0.673, (0 split)
```

```
##
## Node number 2: 52 observations,    complexity param=0.04521877
##   mean=60.66923, MSE=50.05944
##   left son=4 (36 obs) right son=5 (16 obs)
##   Primary splits:
##       income    < 3300  to the left,  improve=0.2280609, (0 missing)
##       fertility < 5.035 to the right, improve=0.1186970, (0 missing)
##   Surrogate splits:
##       fertility < 4.16  to the right, agree=0.731, adj=0.125, (0 spl
it)
##
## Node number 3: 115 observations,    complexity param=0.1029111
##   mean=75.02609, MSE=27.3461
##   left son=6 (77 obs) right son=7 (38 obs)
##   Primary splits:
##       income    < 26200 to the left,  improve=0.4296259, (0 missing)
##       fertility < 2.235 to the right, improve=0.3348740, (0 missing)
##   Surrogate splits:
##       fertility < 1.395 to the right, agree=0.696, adj=0.079, (0 spl
it)
##
## Node number 4: 36 observations,    complexity param=0.01144472
##   mean=58.41667, MSE=45.68528
##   left son=8 (19 obs) right son=9 (17 obs)
##   Primary splits:
##       income    < 1635  to the left,  improve=0.09135824, (0 missing
)
##       fertility < 4.32  to the left,  improve=0.01855131, (0 missing
)
##   Surrogate splits:
##       fertility < 4.44  to the right, agree=0.694, adj=0.353, (0 spl
it)
##
## Node number 5: 16 observations
##   mean=65.7375, MSE=22.79734
##
## Node number 6: 77 observations,    complexity param=0.03526809
##   mean=72.61818, MSE=21.34928
##   left son=12 (22 obs) right son=13 (55 obs)
##   Primary splits:
##       fertility < 2.585 to the right, improve=0.2816628, (0 missing)
##       income    < 7940  to the left,  improve=0.1770320, (0 missing)
##   Surrogate splits:
##       income < 6130  to the left,  agree=0.766, adj=0.182, (0 split)
##
## Node number 7: 38 observations
##   mean=79.90526, MSE=3.942604
##
## Node number 8: 19 observations
##   mean=56.48421, MSE=48.94343
```

```
##
## Node number 9: 17 observations
##   mean=60.57647, MSE=33.20533
##
## Node number 12: 22 observations
##   mean=68.74091, MSE=32.40787
##
## Node number 13: 55 observations
##   mean=74.16909, MSE=8.507226
```

SECTION 5 - Clustering

```
# find the number of clusters we'll need to use

# Elbow
fviz_nbclust(df_std, FUN = hcut, method = "wss")
```




```
# K-Means with k=3
km3 <- kmeans(df_std, 3)
clusters_info <- km3$cluster

# Adding cluster classification to dataset as a column
df_clustered <- mutate(df, cluster = clusters_info)
head(df_clustered)
```

```
##               country child_mort exports health_spending imports inc
ome
## 1         Afghanistan       90.2    10.0            7.58    44.9   1
610
## 2             Albania       16.6    28.0            6.55    48.6   9
930
```

```
## 3             Algeria       27.3    38.4           4.17    31.4  12
900
## 4              Angola      119.0    62.3           2.85    42.9   5
900
## 5 Antigua and Barbuda       10.3    45.5           6.03    58.9  19
100
## 6           Argentina       14.5    18.9           8.10    16.0  18
700
##   inflation life_exp fertility  gdpp cluster
## 1      9.44     56.2      5.82   553       3
## 2      4.49     76.3      1.65  4090       2
## 3     16.10     76.5      2.89  4460       2
## 4     22.40     60.1      6.16  3530       3
## 5      1.44     76.8      2.13 12200       2
## 6     20.90     75.8      2.37 10300       2
```

```
# printing out the cluster centroids
centroids <- df_clustered %>%
    group_by(cluster) %>% # Grouping by the respective country group
    summarise(n = n(), # Listing all the variables and calculating thei
r means
        child_mort = mean(child_mort),
        exports = mean(exports),
        health_spending = mean(health_spending),
        imports = mean(imports),
        income = mean(income),
        inflation = mean(inflation),
        life_exp = mean(life_exp),
        fertility = mean(fertility),
        gdpp = mean(gdpp))

centroids
```

```
## Registered S3 method overwritten by 'cli':
##   method     from
##   print.tree tree
```

```
## # A tibble: 3 x 11
##   cluster     n child_mort exports health_spending imports income in
flation
##     <int> <int>      <dbl>   <dbl>           <dbl>   <dbl>  <dbl>
  <dbl>
## 1       1    36        5      58.7            8.81    51.5 45672.
   2.67
## 2       2    84       21.9    40.2            6.20    47.5 12306.
   7.60
## 3       3    47       93.0    29.2            6.39    42.3  3942.
  12.0
## # ... with 3 more variables: life_exp <dbl>, fertility <dbl>, gdpp <
dbl>
```

```r
# find countries most in need for help based on the order of per capita
 income
sqldf("SELECT country, income, fertility, child_mort, life_exp
      FROM df_clustered
      WHERE cluster == 3
      ORDER BY income
      LIMIT 5")
```

```
##                     country income fertility child_mort life_exp
## 1         Congo, Dem. Rep.    609      6.54      116.0     57.5
## 2                   Liberia    700      5.02       89.3     60.8
## 3                   Burundi    764      6.26       93.6     57.7
## 4                     Niger    814      7.49      123.0     58.8
## 5 Central African Republic    888      5.21      149.0     47.5
```