****

**INC 491: Data Science and Intelligent Techniques**

**Mini Project: Life Expectancy (WHO) Analysis**

**Nontawit        Markjan 59070504003  AE**

**To**

**Dr. Issarapong Khuankrue**

**This report is part of the subject**

**INC 491 - Data Science and Intelligent Techniques**

**Developed by Automation Engineering Students**

**King Mongkut's University of Technology Thonburi**

**Semester 1/2019**

**Index**

**Page**

1. Background information 4
2. Data manipulation 6

2.1 Import 6

2.2 Tidy 7

2.3Transform 9

2.4 Visualize 10

2.4.1 Percentage of country status 10

2.4.2 Comparison between developing and developed countries 10

2.4.2.1 Life expectancy 11

2.4.2.2 Adult mortality 11

2.4.2.3 Infant deaths 12

2.4.2.4 Alcohol 12

` 2.4.2.5 Percentage expenditure 13

2.4.2.6 Hepatitis B 13

2.4.2.7 Measles 14

2.4.2.8 BMI 14

2.4.2.9 Under-5 deaths 15

2.4.2.10 Polio 15

2.4.2.11 Total expenditure 16

2.4.2.12 Diphtheria 16

2.4.2.13 HIV/AIDS 17

2.4.2.14 GDP 17

2.4.2.15 Population 18

2.4.2.16 Thinness 1-19 years 18

2.4.2.17 Thinness 5-9 years 19

2.4.2.18 Income composition of resources 19

2.4.2.19 Schooling 20

**Page**

2.4.3 Effect of some factors to life expectancy 20

2.4.3.1 Adult mortality 21

2.4.3.2 Infant Deaths 21

2.4.3.3 Alcohol 22

2.4.3.4 Population 22

2.4.3.5 Schooling 23

2.4.4 Correlation 23

2.5 Modelling 25

2.5.1 Create train and test dataset 25

2.5.2 Multiple linear regression 25

2.5.3 Random forest 26

2.5.4 Gradient boosting 27

2.5.5 Supported vector machine (SVM) 28

2.6 Accuracy 28

2.6.1 Mean-squared error (MSE) 29

2.6.2 Root-mean-squared error (RMSE) 29

2.6.3 Min/max accuracy 30

3. Summary 30

4. Appendix 31

1. Background information

This mini project studies the factors affecting life expectancy by using dataset from the Global Health Observatory (GHO), which is World Health Organization’s (WHO) gateway to health-related statistics repository. It keeps track of the health status as well as many other related factors for all countries. The data-sets are available to public for health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website.

The dataset consists of the factors from year 2000 to 2015. There are some missing data from less known countries such as Vanuatu, Tonga, Togo, Cabo Verde etc. This dataset is the final version that was merged from the different source. It consists of 22 columns and 2,938 rows. The details of each column are shown below.

|  |  |
| --- | --- |
| **Columns** | **Definition** |
| ‘Country’ | Country |
| ‘Year’ | Year |
| ‘Status’ | Developed or Developing status |
| ‘Life expectancy’ | Life Expectancy in age |
| ‘Adult Mortality’ | Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population) |
| ‘infant deaths’ | Number of Infant Deaths per 1000 population |
| ‘Alcohol’ | Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol) |
| ‘percentage expenditure’ | Expenditure on health as a percentage of Gross Domestic Product per capita (%) |
| ‘Hepatitis B’ | Hepatitis B (Hep B) immunization coverage among 1-year-olds (%) |
| ‘Measles’ | Measles, number of reported cases per 1000 population |
| ‘BMI’ | Average Body Mass Index of entire population |
| ‘under-five deaths’ | Number of under-five deaths per 1000 population |
| ‘Polio’ | Polio (Pol3) immunization coverage among 1-year-olds (%) |
| ‘Total expenditure’ | General government expenditure on health as a percentage of total government expenditure (%) |
| ‘Diphtheria’ | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) |
| ‘HIV/AIDS’ | Deaths per 1 000 live births HIV/AIDS (0-4 years) |
| ‘GDP’ | Gross Domestic Product per capita (in USD) |
| ‘Population’ | Population of the country |
| ‘thinness 1-19 years’ | Prevalence of thinness among children and adolescents for Age 10 to 19 (%) |
| ‘thinness 5-9 years’ | Prevalence of thinness among children for Age 5 to 9(%) |
| ‘Income composition of resources’ | Human Development Index in terms of income composition of resources (index ranging from 0 to 1) |
| ‘Schooling’ | Number of years of Schooling (years) |

Table 1: columns definition

1. Data manipulation

Data manipulation in this project is based on the below process where the red part represents exploratory Data Analysis (EDA) - the process for data investigation which discovers the pattern, structure, details, and maximize the insight of the given data. It is the philosophy on how the data should be manipulated. Modelling and analysis are the blue part.

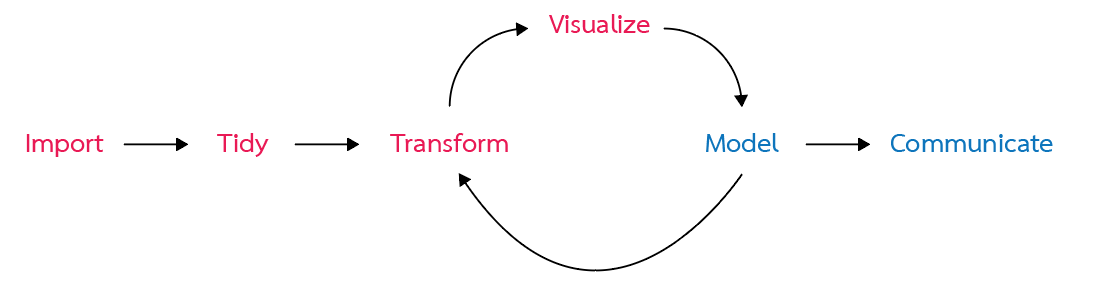


Figure 1: Data exploration process

2.1 Import

Life Expectancy Data.csv is the dataset for this mini project



Figure 2: Dataset

The dataset was imported as data frame in the project with this code by using ‘read\_csv’ library to read comma separated value file. The libraries for data manipulation are also called.

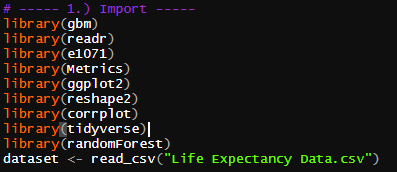


Figure 3: Code for import comma separated value files (.csv) with libraries at the left and data frame as results at the right

2.2 Tidy

The datasets were observed by using these codes.

‘str(dataset)’ – observe the structure of R object, which is ‘dataset’ in this case.

‘summary(dataset)’ – produce the summary details of each attribute in dataset

‘View(dataset)’ – show all entries of dataframe

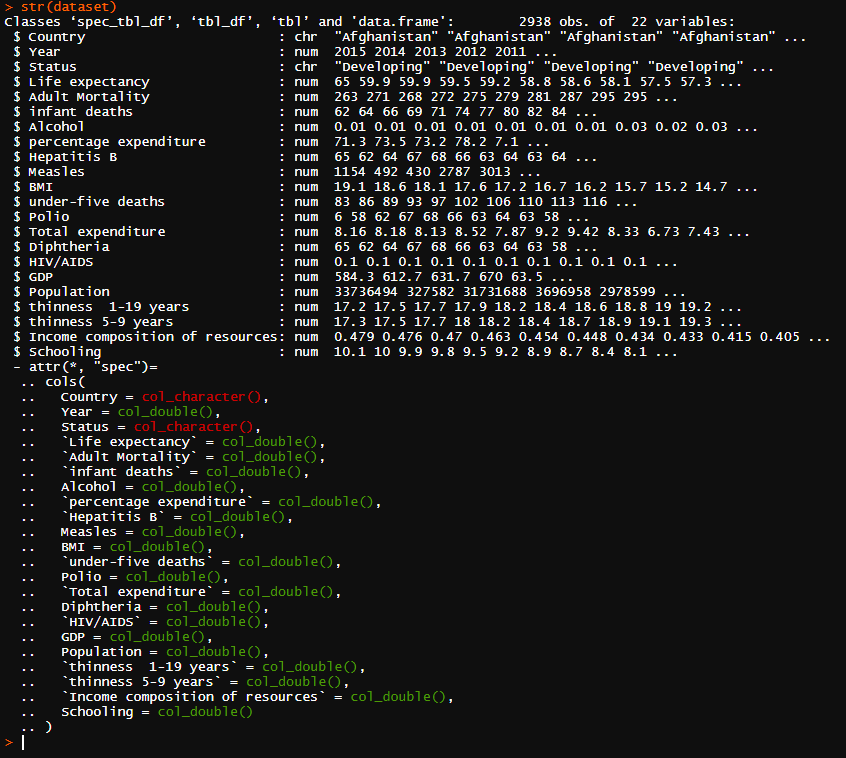


Figure 4: Code for showing ‘dataset’ data frame structure

In addition, we can observe that some variables have inappropriate data type, which is ‘Status’. It should be set as the factor.

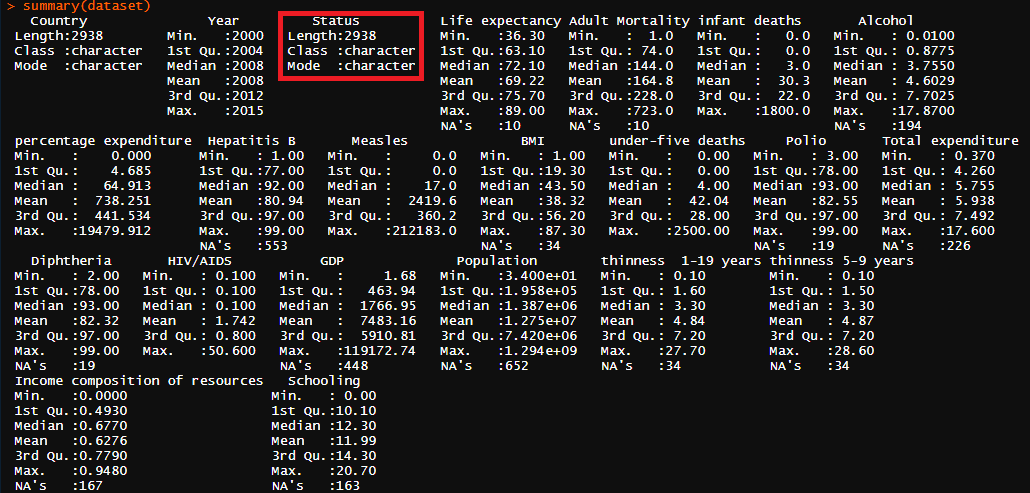


Figure 5: Code for showing ‘dataset’ data frame summary

For ‘view’ function, we can clearly see the data inside the data frame. The structure is the same as we observed by using ‘str’ function but ‘view’ does not tell the data type of each variable.

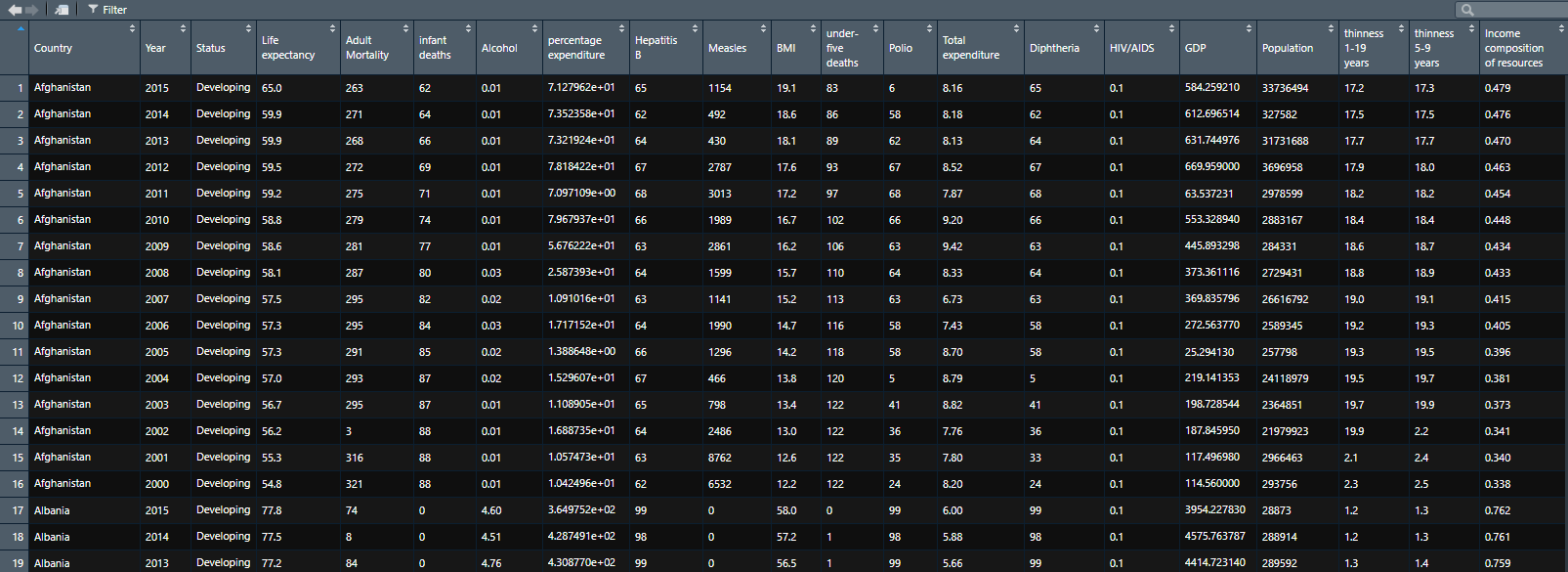


Figure 6: Code for showing ‘dataset’ data frame whole entries

I also observed that there are about a half of dataset that have null value after I omitted the row with the null value. Therefore, I replace the null value with the mean in each column.

Figure 7: Code for omitting the row with null value and create new data frame and the result

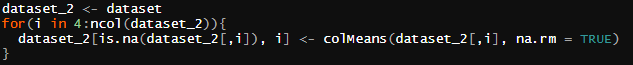
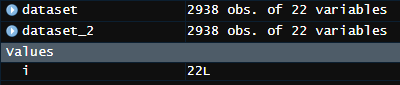
 

Figure 8: Replace null value with the mean in each column and its result

2.3 Transform

From the current data frame structure, I change the name of some columns that has the space in the name to prevent the error from using any library.

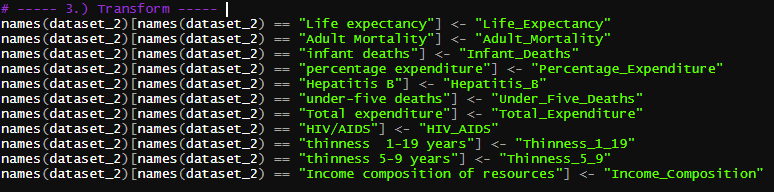


Figure 9: code for rename the columns

The ‘Status’ column is converted as factor, the new column name ‘Status\_type’, which represents the ‘Status’ as numeric, and two new data frames are created as the data for ‘Developing’ countries and ‘Developed’ countries because we are going to see the differences of the factors for both types.

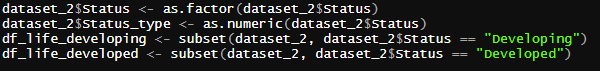


Figure 10: Code for data transformation and the result

2.4 Visualize

From my observation, there are many aspects and relationships between the data as represents below.

2.4.1 Percentage of country status

There are 82.57% as developing countries and 17.42% as developed countries.

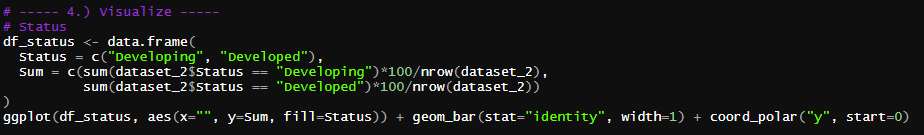


Figure 11: Code for percentage of country status calculation

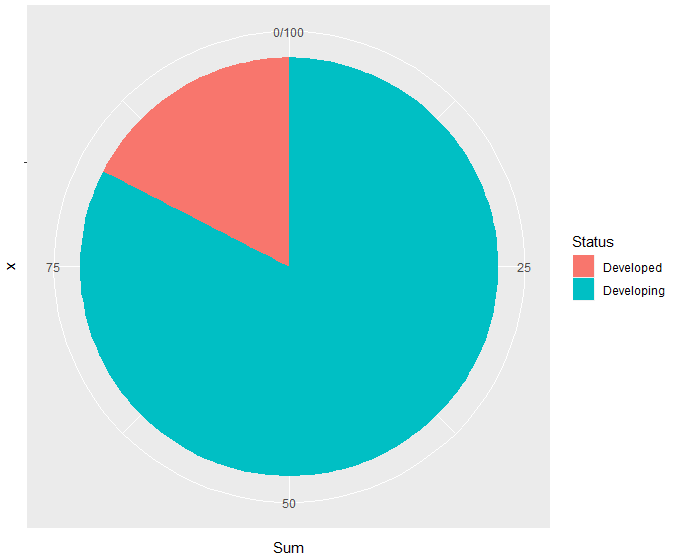


Figure 12: Visualization for percentage of country status calculation

2.4.2 Comparison between developing and developed countries

There are the differences distribution between the developing and developed countries as follows.

2.4.2.1 Life expectancy

People in developed countries rather have a higher life expectancy than developing country

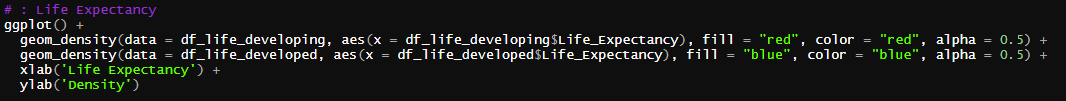


Figure 13: Code for plotting life expectancy between developing and developed countries

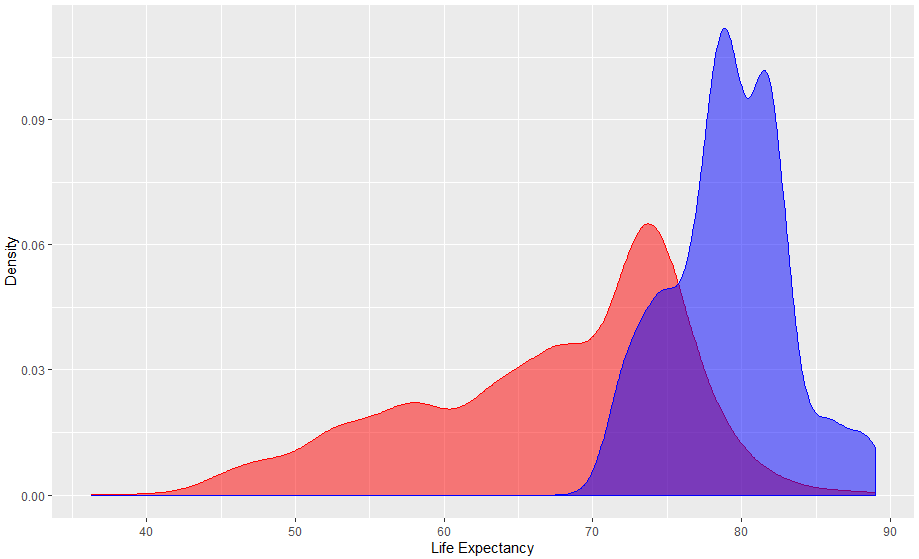


Figure 14: Visualization for plotting life expectancy between developing and developed countries

2.4.2.2 Adult mortality

People in developed countries rather have a lower adult mortality than developing country

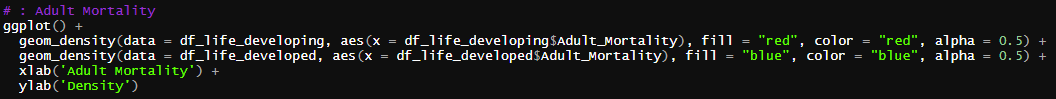


Figure 15: Code for plotting adult mortality between developing and developed countries

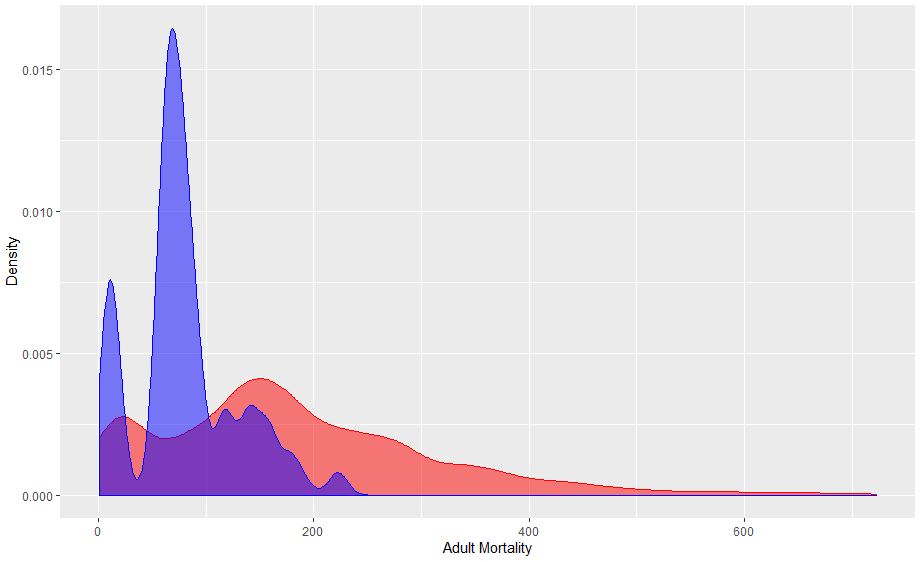


Figure 16: Visualization for plotting adult mortality between developing and developed countries

2.4.2.3 Infant deaths

People in developed countries rather have a lower infant death than developing country. There are some countries that have the high infant deaths which becomes the outlier as shown below.

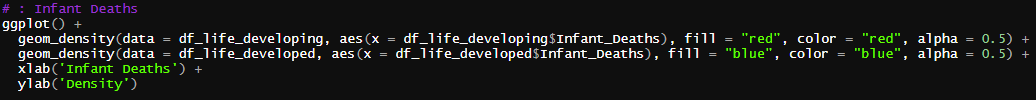


Figure 17: Code for plotting infant deaths between developing and developed countries

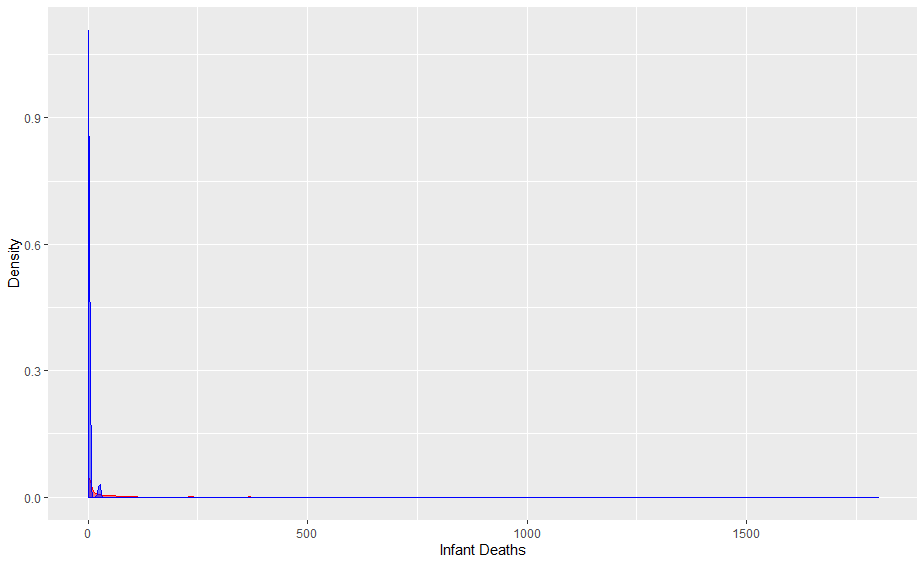


Figure 18: Visualization for plotting infant deaths between developing and developed countries

2.4.2.4 Alcohol

People in developed countries rather have a higher alcohol consumption per capita (liter) than developing country.

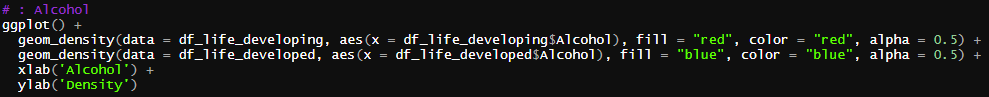


Figure 19: Code for plotting alcohol consumption (liter) between developing and developed countries

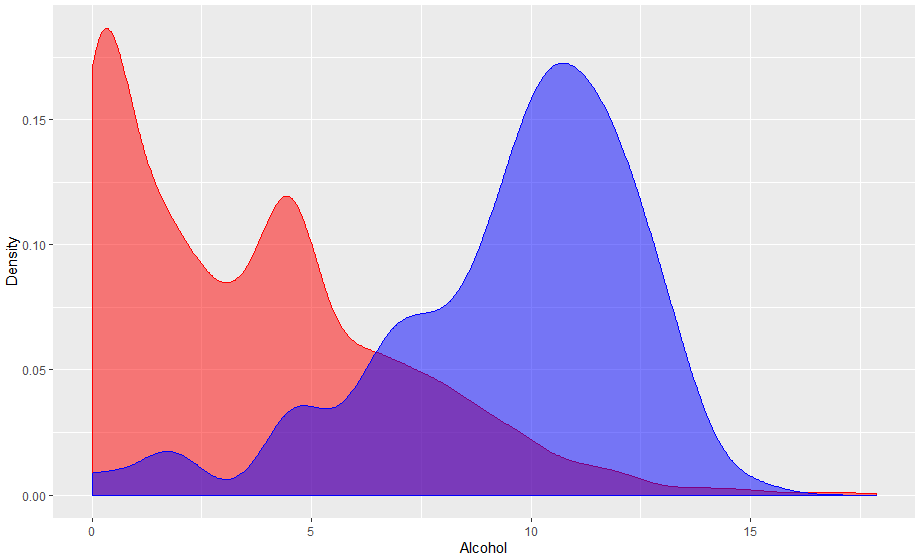


Figure 20: Visualization for plotting alcohol consumption (liter) between developing and developed countries

` 2.4.2.5 Percentage expenditure

People in developed countries rather have a higher percentage expenditure than developing country

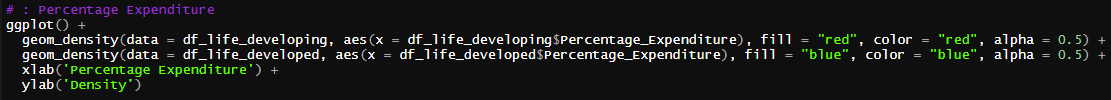


Figure 21: Code for plotting percentage expenditure between developing and developed countries

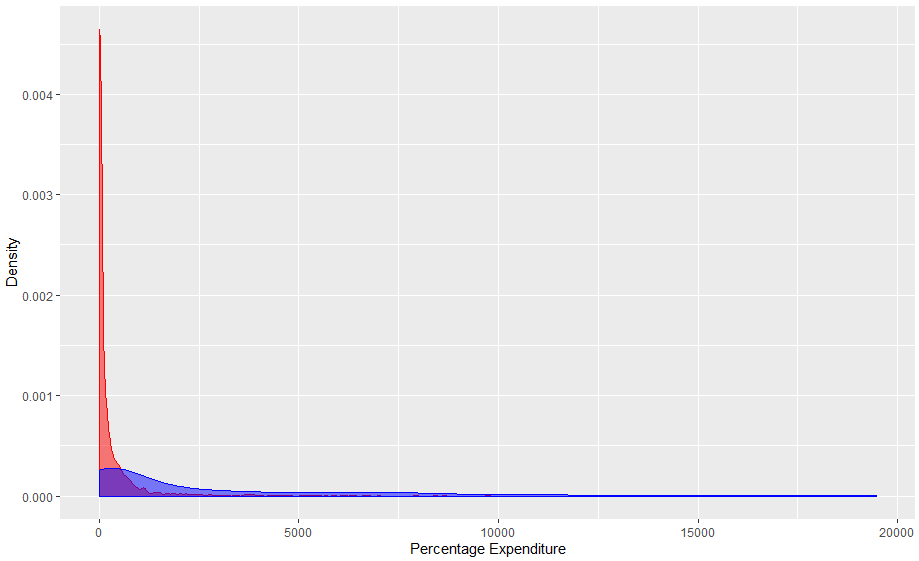


Figure 22: Visualization for plotting percentage expenditure between developing and developed countries

2.4.2.6 Hepatitis B

People in developed countries and developing countries quite have the same distribution of hepatitis B, developed countries tend to have a higher rate.

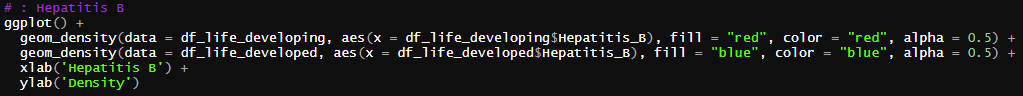


Figure 23: Code for plotting hepatitis B between developing and developed countries

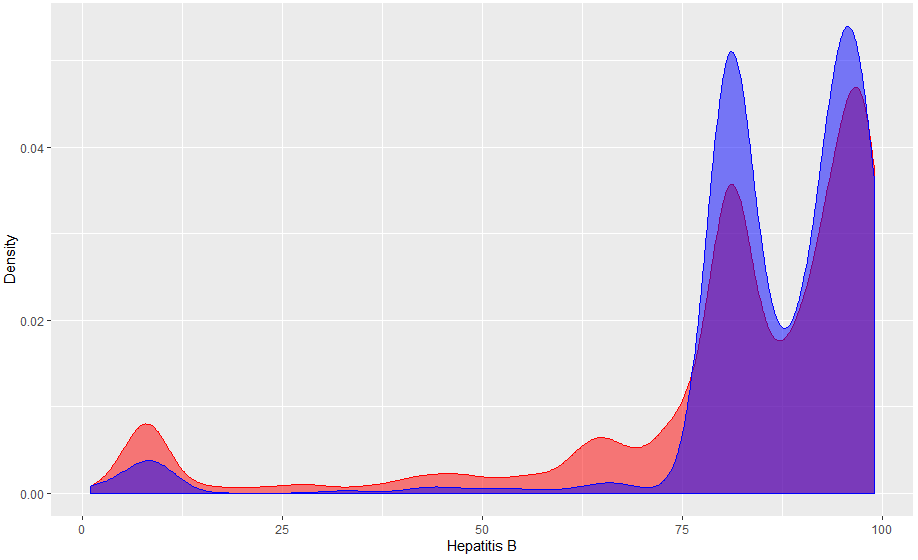


Figure 24: Visualization for plotting hepatitis B between developing and developed countries

2.4.2.7 Measles

People in developed countries and developing country have the same distribution of measles. However, developed countries have higher chance to has less measles due to the higher density compare to developing countries.

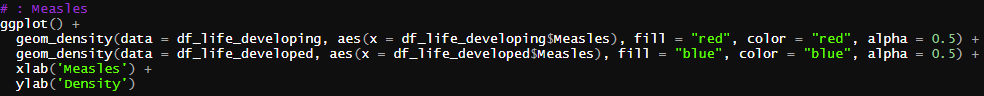


Figure 25: Code for plotting Measles between developing and developed countries

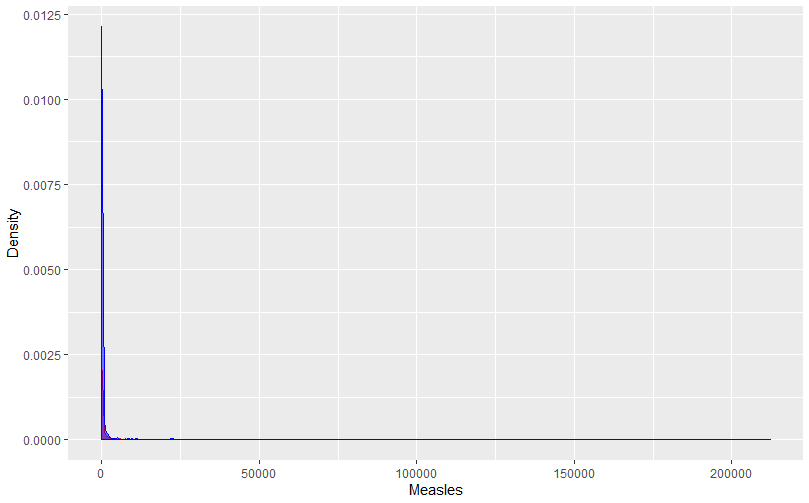


Figure 26: Visualization for plotting measles between developing and developed countries

2.4.2.8 BMI

Developed countries have gap distribution of BMI. Most people have higher BMI than appropriate value (18.5 – 22.9)

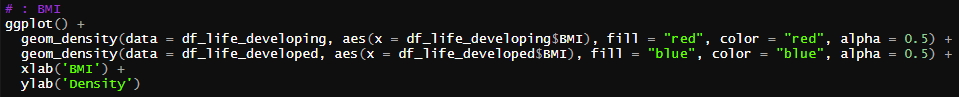


Figure 27: Code for plotting BMI between developing and developed countries

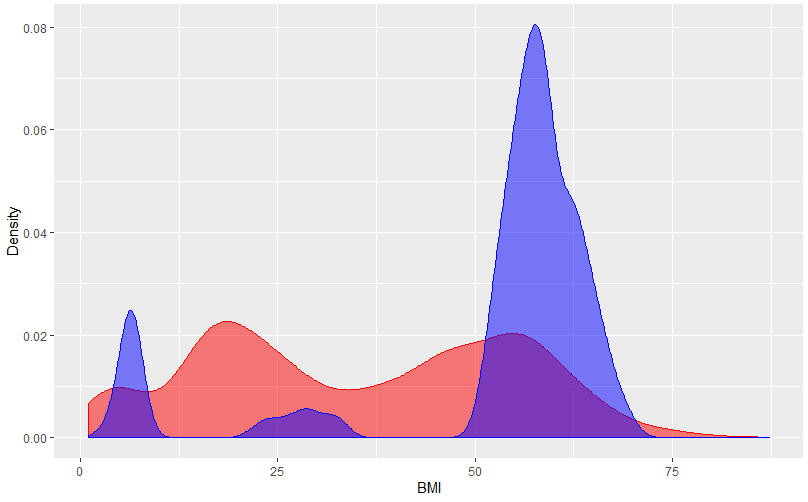


Figure 28: Visualization for plotting BMI between developing and developed countries

2.4.2.9 Under-5 deaths

Developed countries have less under 5 deaths compares to developing countries.

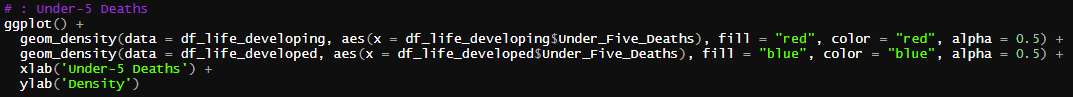


Figure 29: Code for plotting under-5 deaths between developing and developed countries

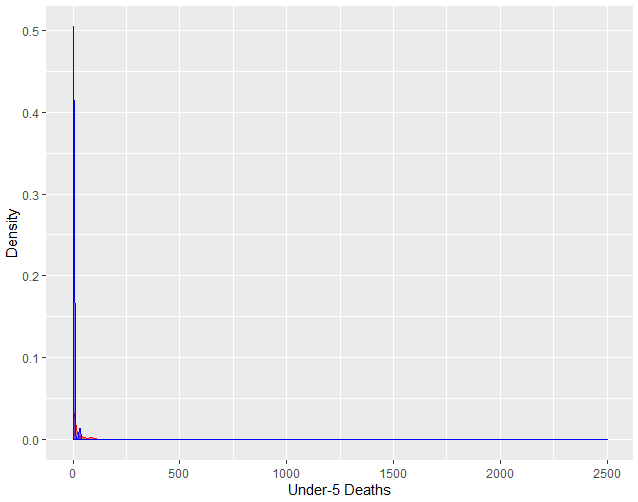


Figure 30: Visualization for plotting under-5 deaths between developing and developed countries

2.4.2.10 Polio

Developed countries have higher polio rates compares to developing countries.

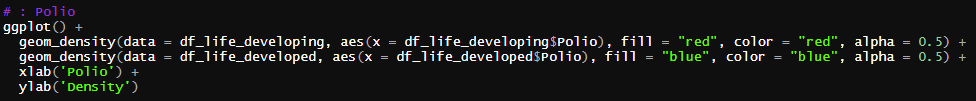


Figure 31: Code for plotting polio between developing and developed countries

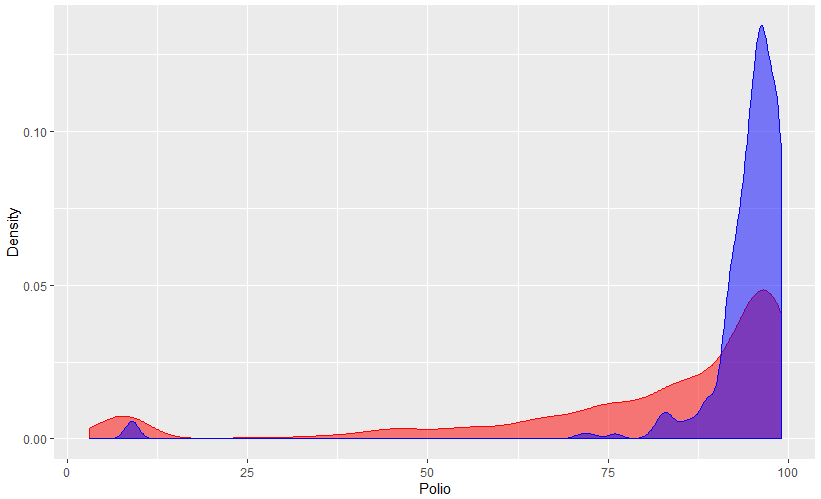


Figure 32: Visualization for plotting polio between developing and developed countries

2.4.2.11 Total expenditure

Government in developed countries expense on health more than developing countries.

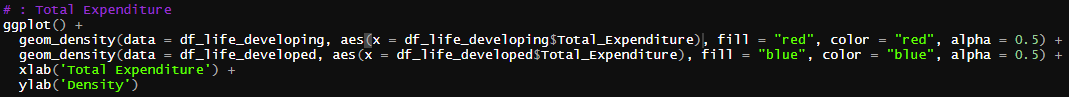


Figure 33: Code for plotting total expenditure between developing and developed countries

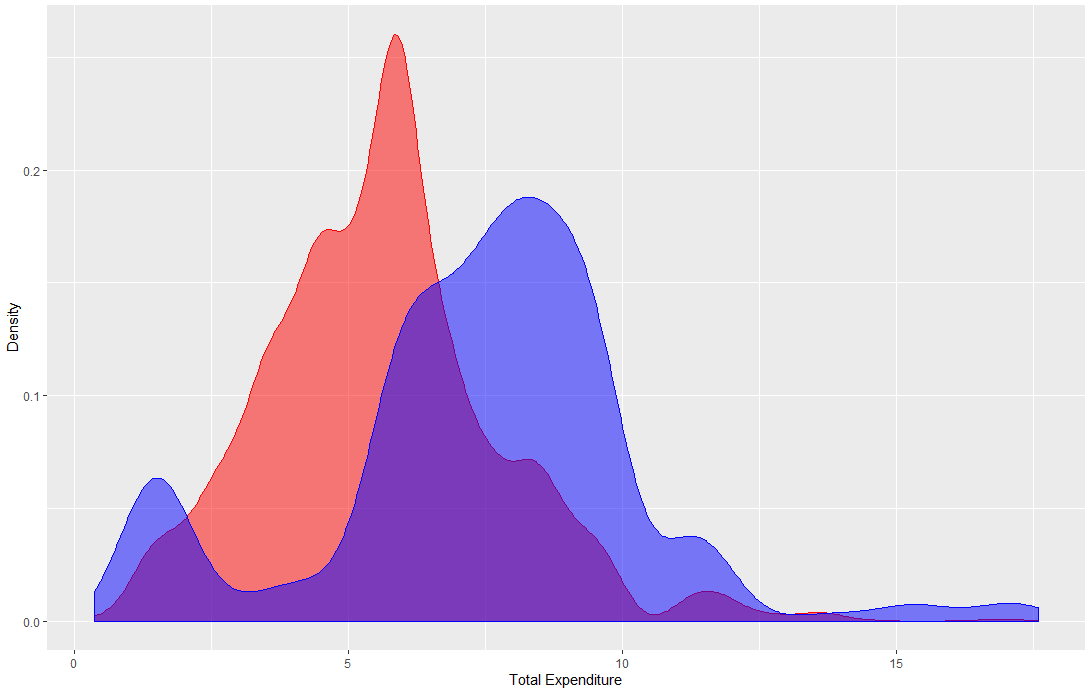


Figure 34: Visualization for plotting total expenditure between developing and developed countries

2.4.2.12 Diphtheria

Developed countries have higher diphtheria rate than developing countries.

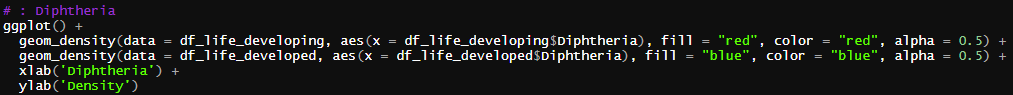


Figure 35: Code for plotting diphtheria between developing and developed countries

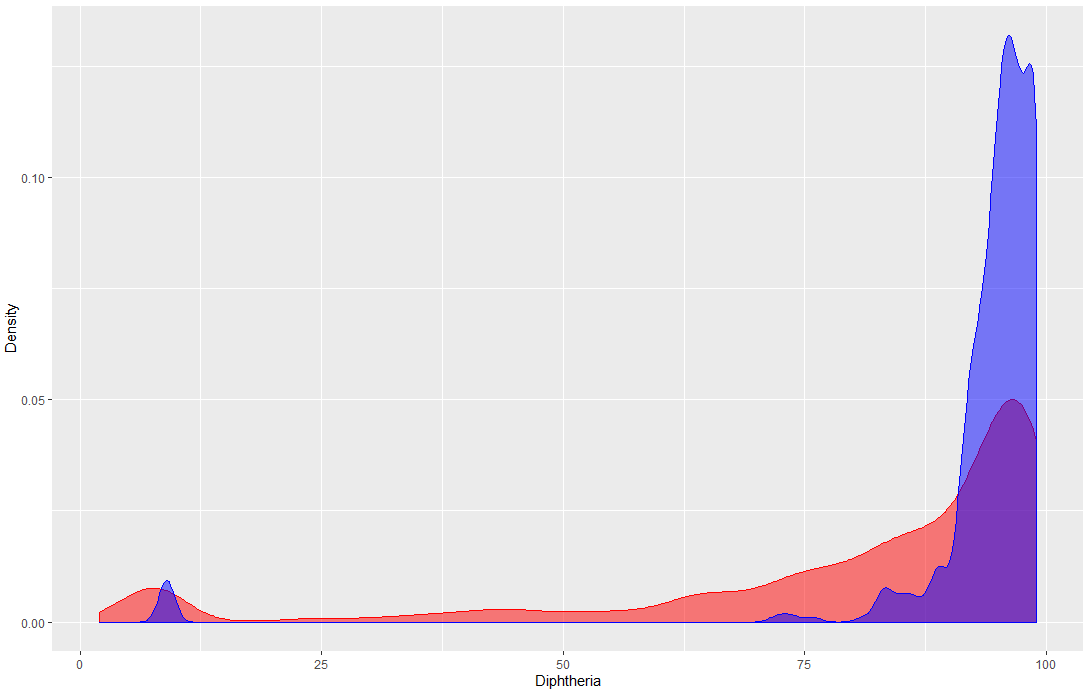


Figure 36: Visualization for plotting polio between developing and developed countries

2.4.2.13 HIV/AIDS

Developed countries have lower death per 1,000 live births HIV/AIDS than developing countries.

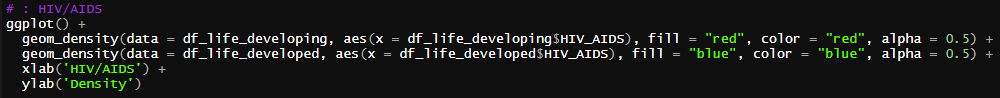


Figure 37: Code for plotting diphtheria between developing and developed countries

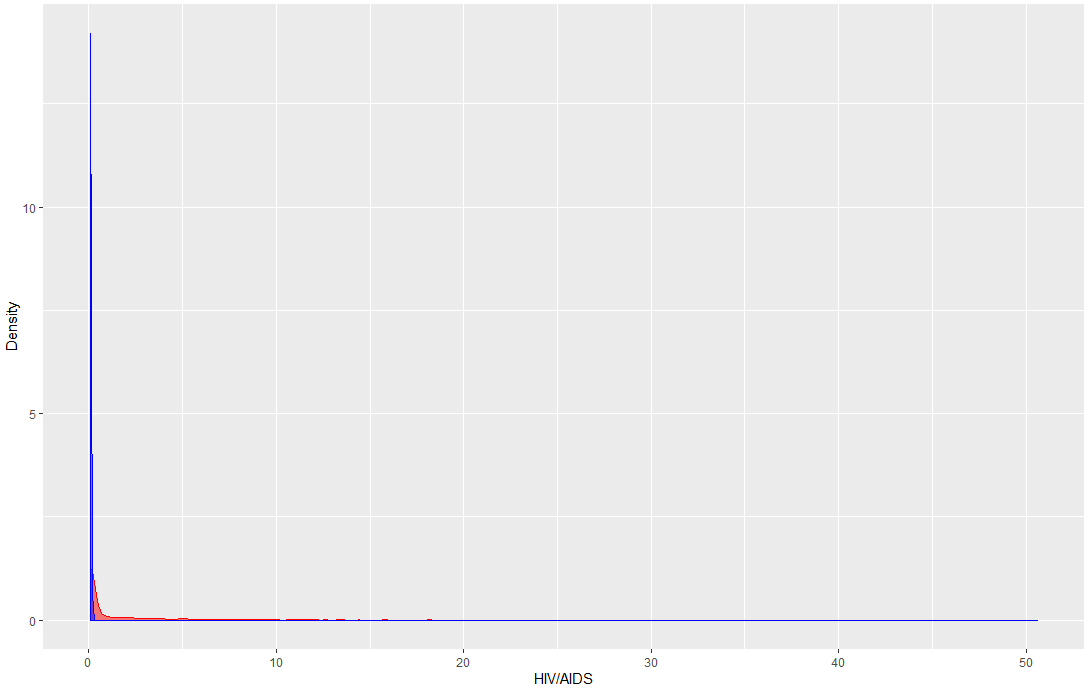


Figure 38: Visualization for plotting HIV/AIDS between developing and developed countries

2.4.2.14 GDP

Developed countries have higher GDP than developing countries.

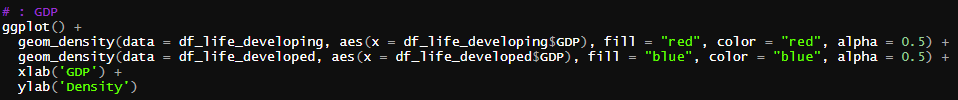


Figure 39: Code for plotting GDP between developing and developed countries

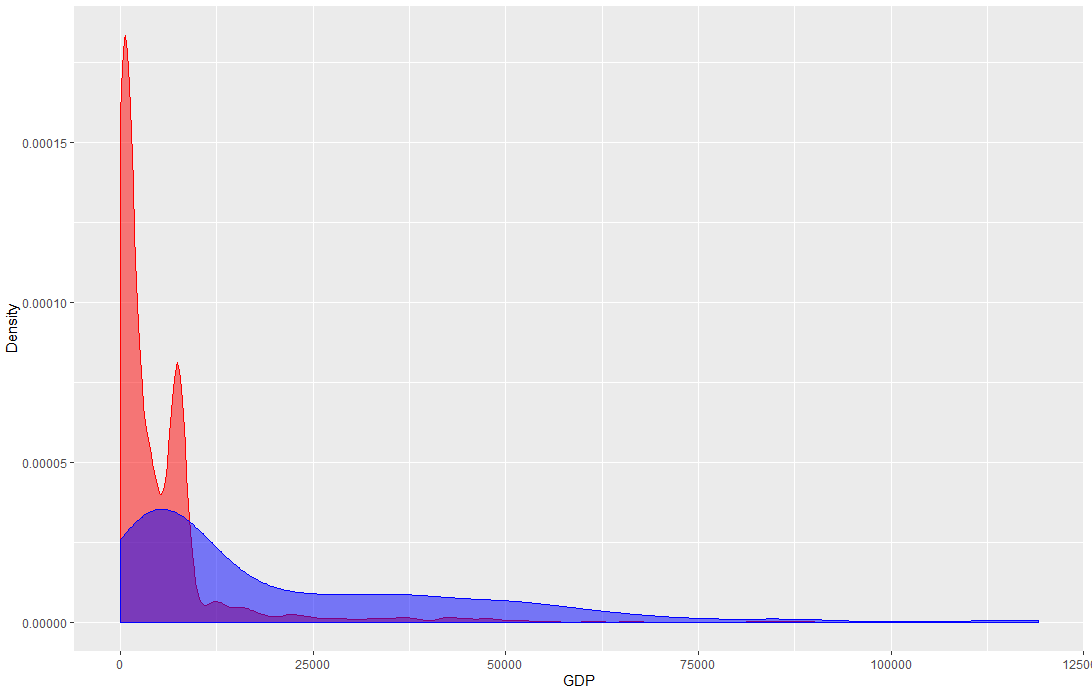


Figure 40: Visualization for plotting GDP between developing and developed countries

2.4.2.15 Population

Developed and developing countries have quite the same distribution of population

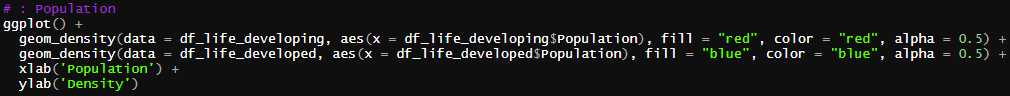


Figure 41: Code for plotting population between developing and developed countries

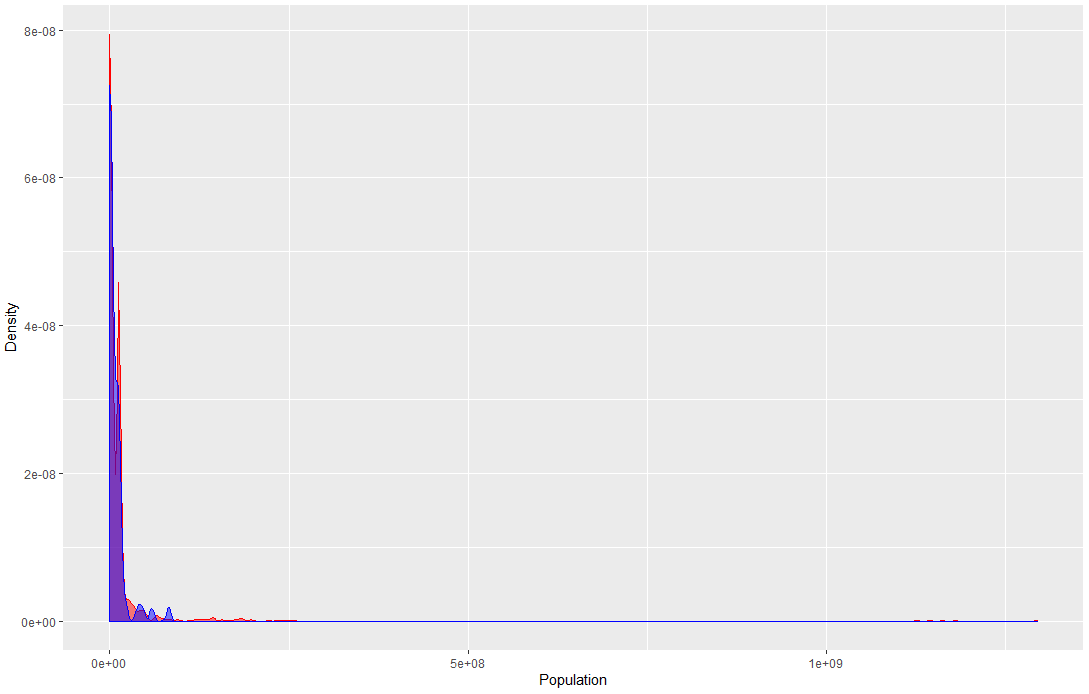


Figure 42: Visualization for plotting population between developing and developed countries

2.4.2.16 Thinness 10-19 years

Developed countries tend to have less thin adolescents between age 10 – 19 than developing countries.

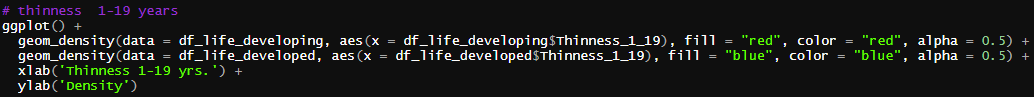


Figure 43: Code for plotting thinness 10-19 years between developing and developed countries

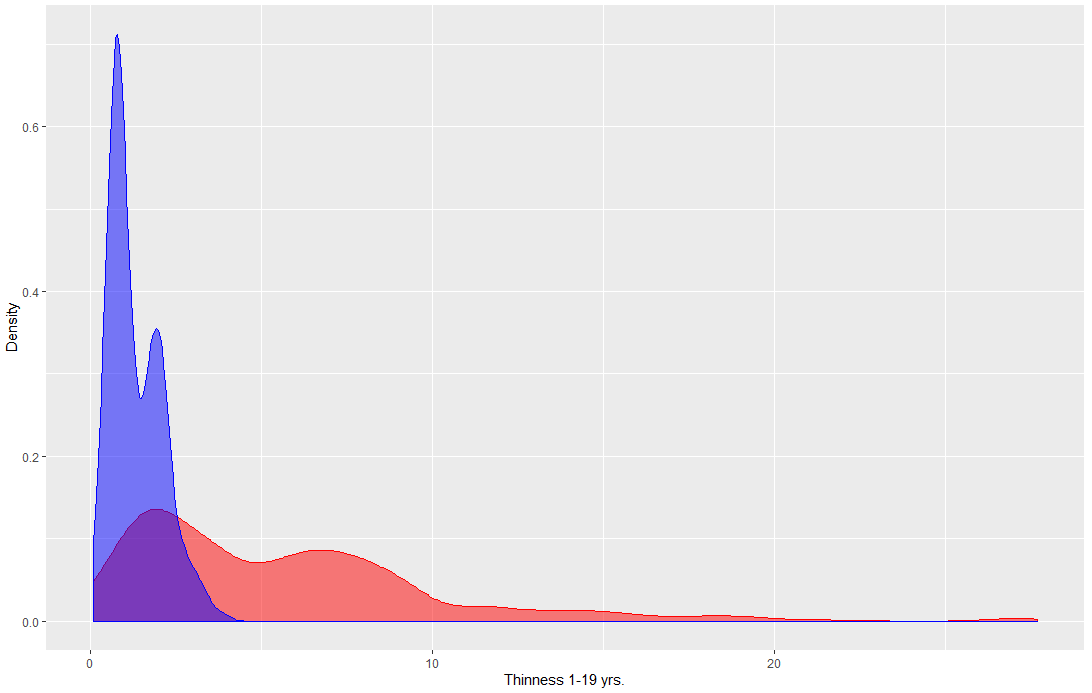


Figure 44: Visualization for plotting Thinness 10-19 years between developing and developed countries

2.4.2.17 Thinness 5-9 years

Developed countries tend to have less thin adolescents between age 5 – 9 than developing countries.

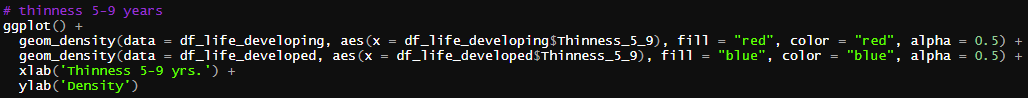


Figure 45: Code for plotting thinness 5-9 years between developing and developed countries

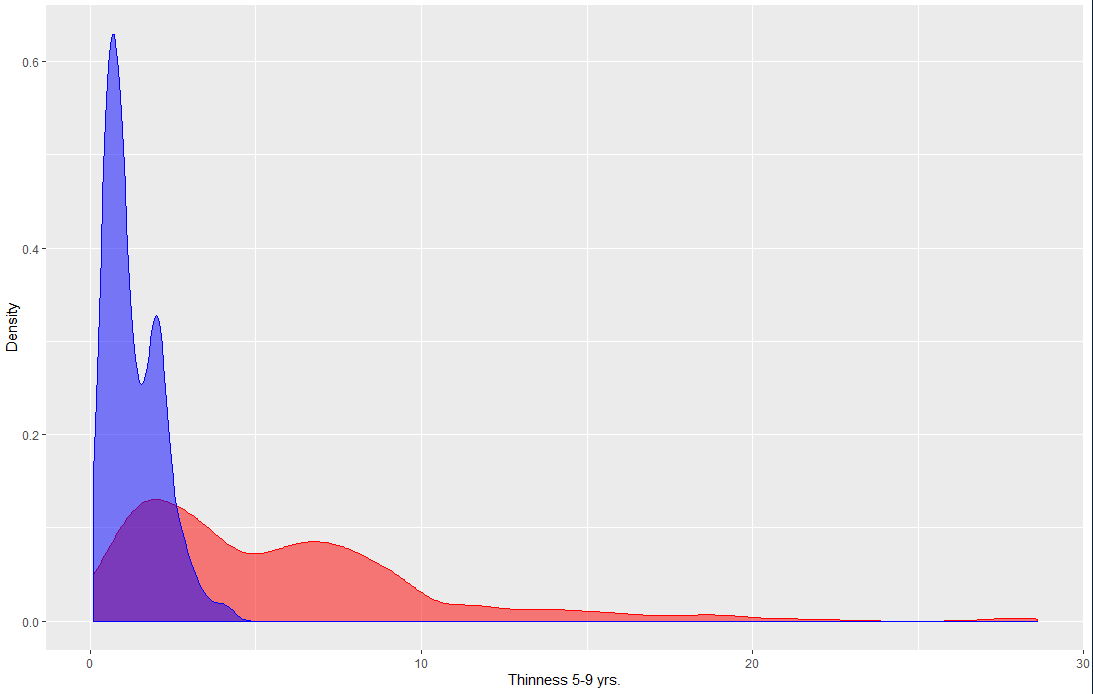


Figure 46: Visualization for plotting thinness 5-9 years between developing and developed countries

2.4.2.18 Income composition of resources

Developed countries tend to have more income composition of resources than developing countries.

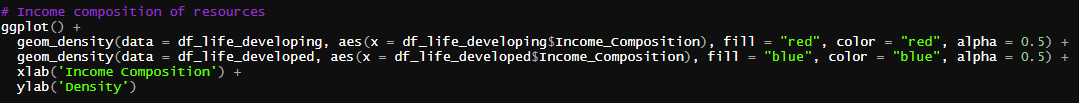


Figure 47: Plotting income composition of resources between developing and developed countries

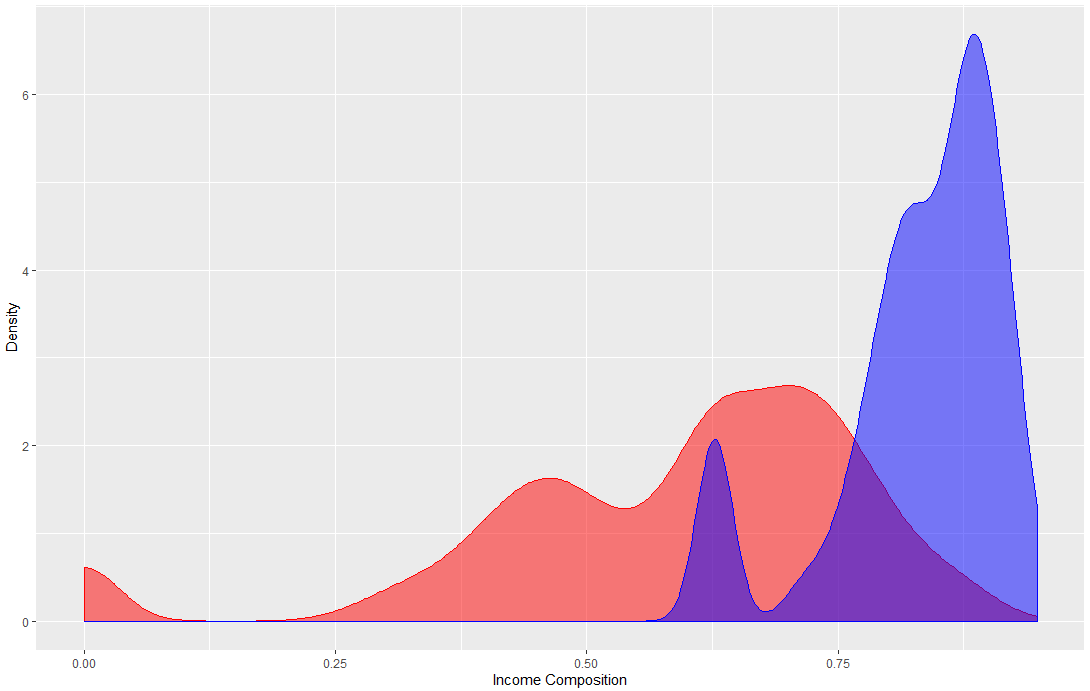


Figure 48: Visualization for plotting income composition of resources between developing and developed countries

2.4.2.19 Schooling

Developed countries tend to have more years of schooling than developing countries.

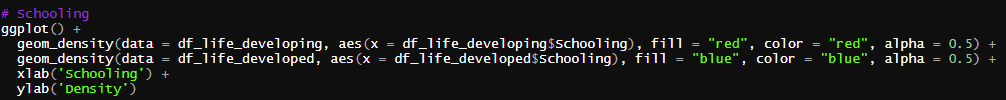


Figure 49: Code for plotting schooling between developing and developed countries

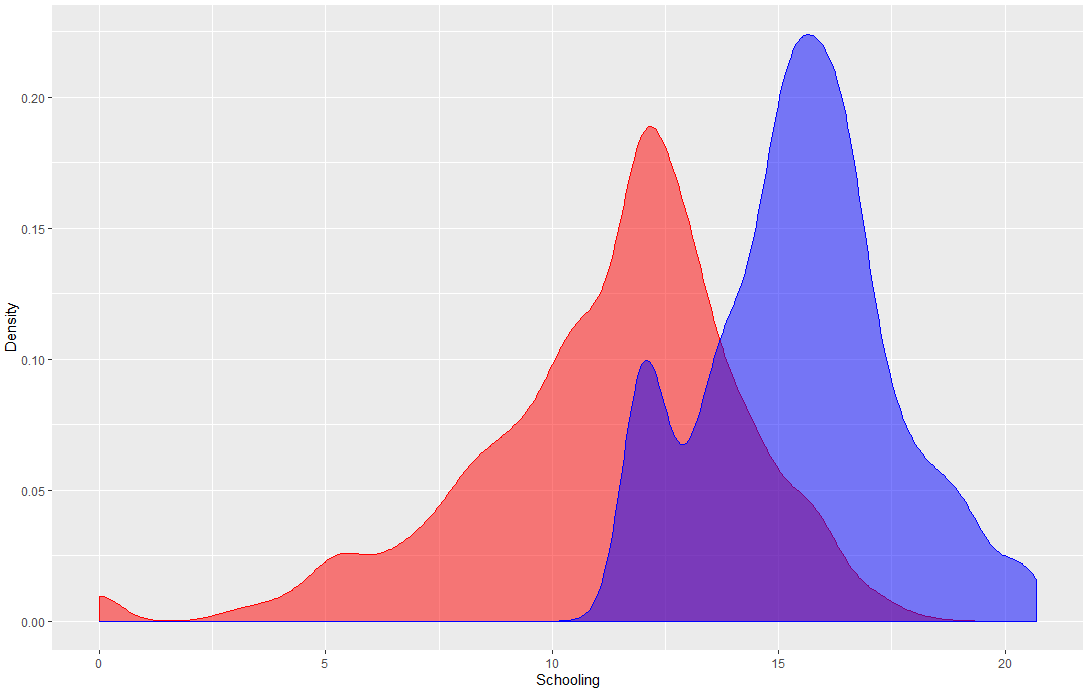


Figure 50: Visualization for plotting schooling between developing and developed countries

2.4.3 Effect of some factors to life expectancy

This section shows the relationship between some factors to life expectancy, consists of adult mortality, infant deaths, alcohol, population, and schooling. The code for implementation is shown below.

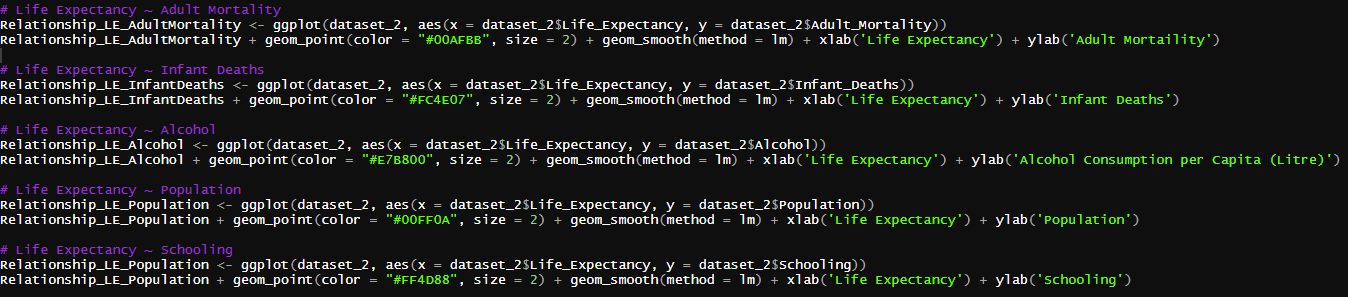


Figure 51: Code for plotting some factors and life expectancy relationship

2.4.3.1 Adult mortality

Adult mortality is inverse proportional to the life expectancy. The higher adult mortality, the less life expectancy that country has.

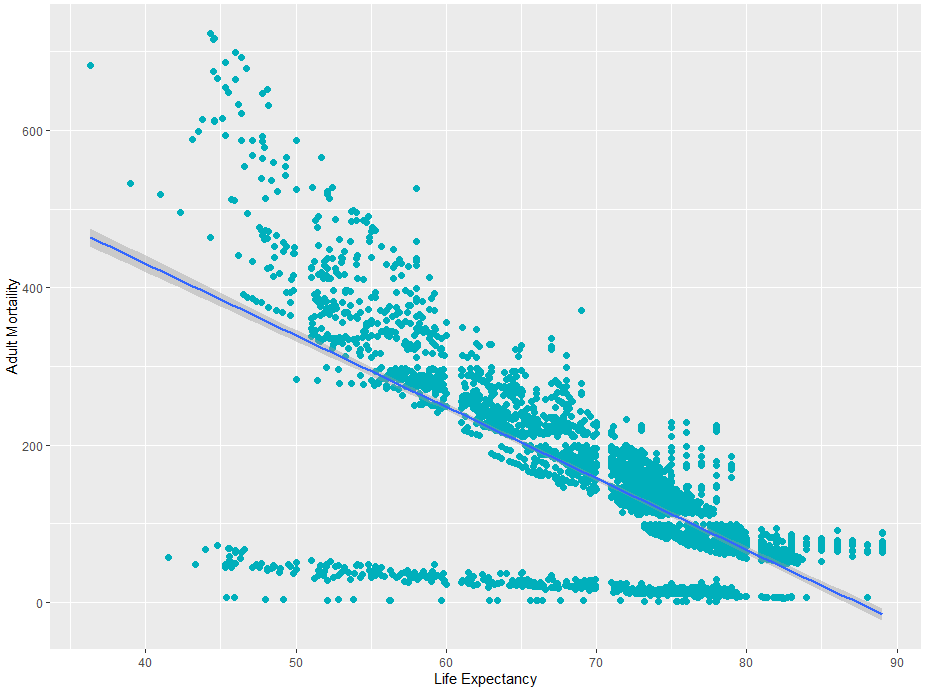


Figure 52: Visualization of relationship between adult mortality and life expectancy

2.4.3.2 Infant deaths

Infant deaths rate is inverse proportional to the life expectancy. The higher Infant deaths rate, the less life expectancy that country has.

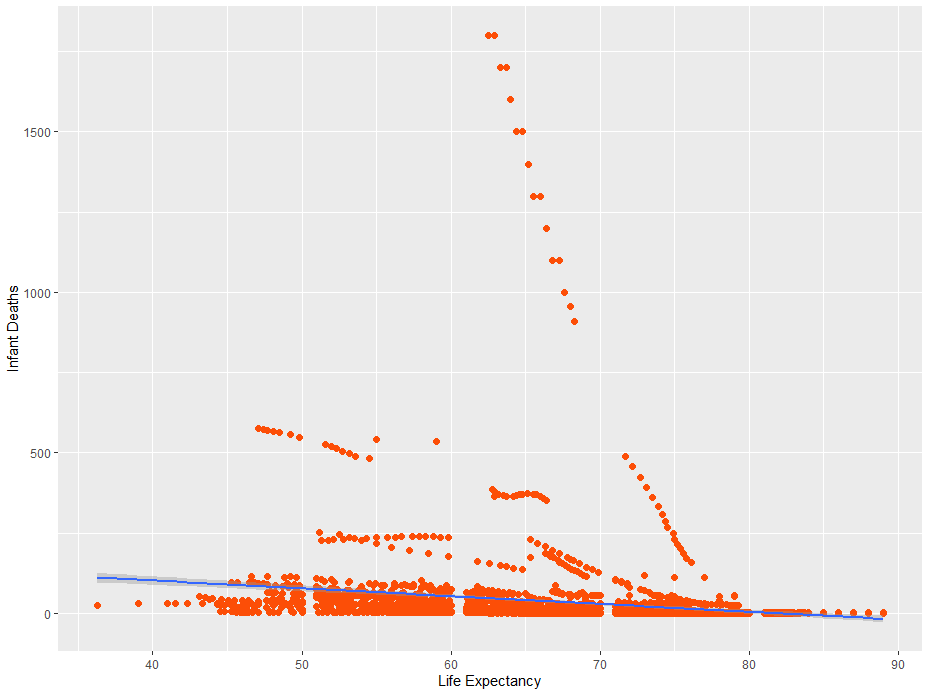


Figure 53: Visualization of relationship between Infant deaths and life expectancy

2.4.3.3 Alcohol

Alcohol consumption per capita (liter) is direct proportional to the life expectancy. The higher alcohol consumption per capita (liter), the higher life expectancy that country has.

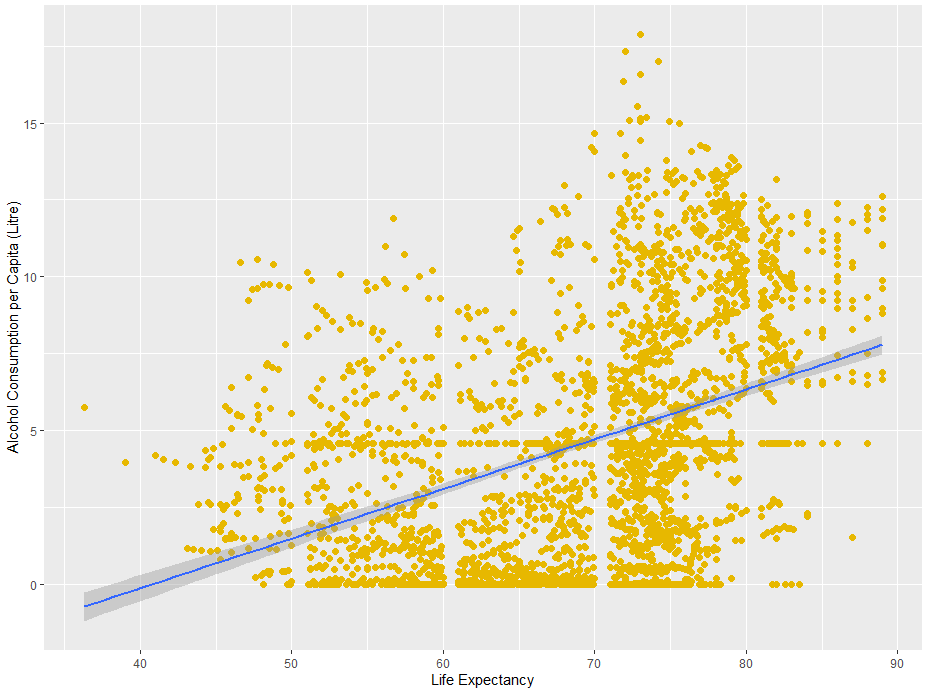


Figure 54: Visualization of relationship between alcohol consumption per capita (liter) and life expectancy

2.4.3.4 Population

Amount of population does not have effect to the life expectancy.

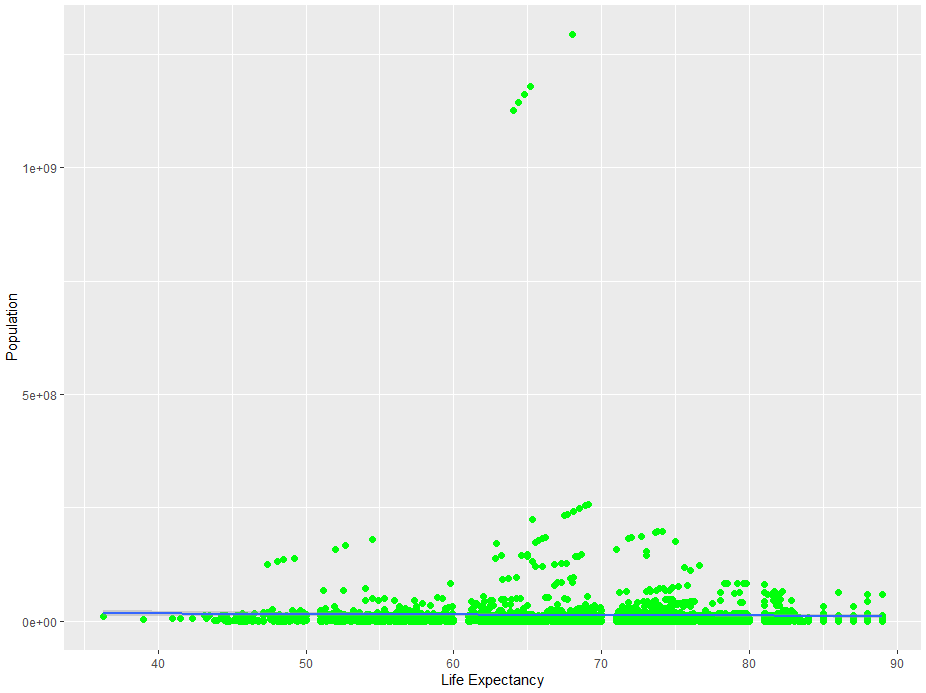


Figure 55: Visualization of relationship between amount of population and life expectancy

2.4.3.5 Schooling

Schooling is direct proportional to the life expectancy. The higher years in schooling, the higher life expectancy that country has.

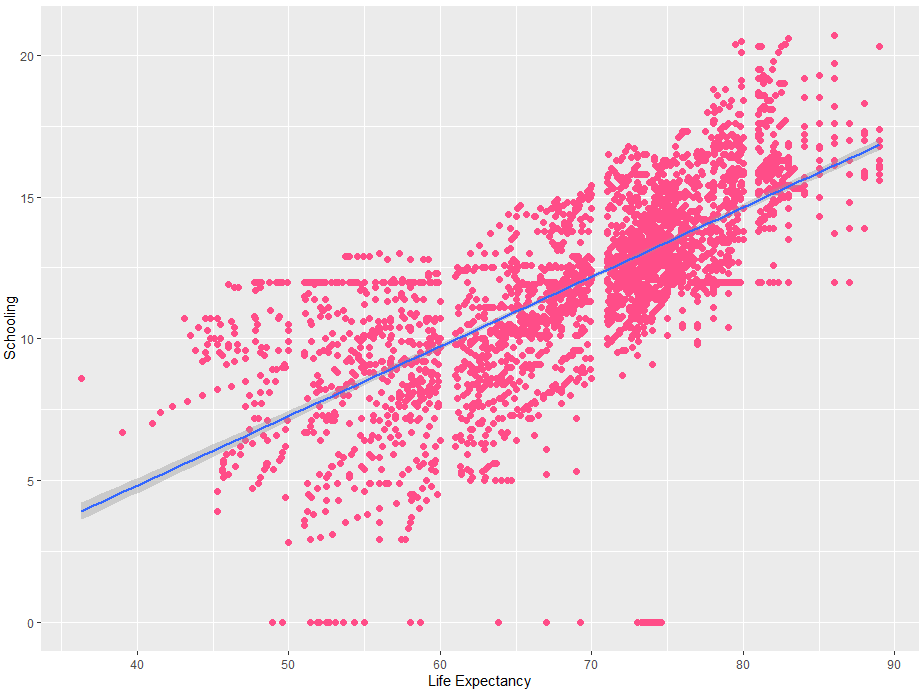


Figure 56: Visualization of relationship between amount of population and life expectancy

2.4.3 Correlation

Heat map is created to represent how strong each factor affects together. However, I create new data frame to represent only each factor affects to life expectancy only for easily interpretation.

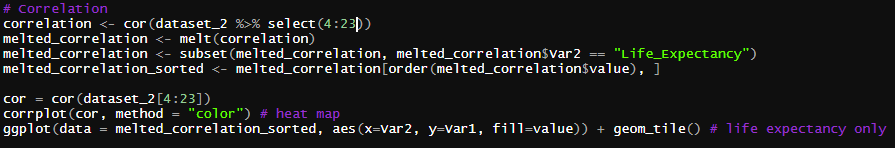


Figure 57: Code for correlation creating, including values and heat map

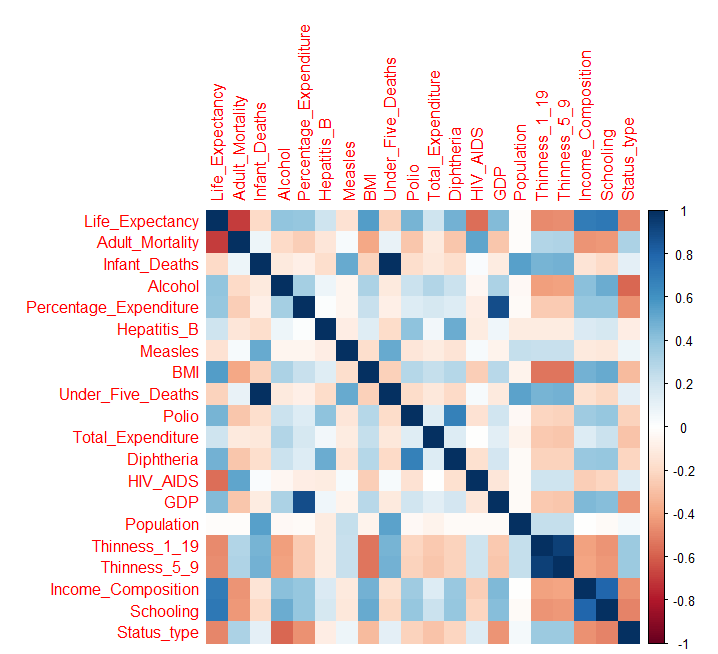


Figure 58: Correlation heat map of each factor

The factor that has the highest directly impact to life expectancy is schooling and the highest inversely impact to life expectancy is adult mortality. The rest has been ordered from lowest to highest as shown below.

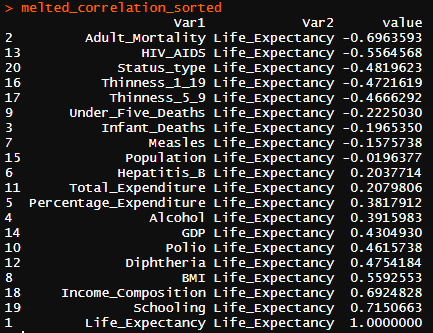
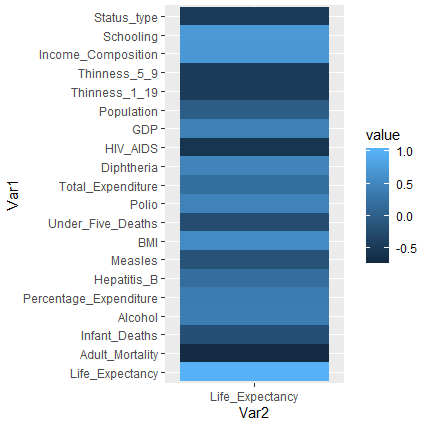


Figure 59: Correlation of each factor to life expectancy as color (left) and value from lowest to highest (right)

2.5 Modelling

2.5.1 Create train and test dataset

The dataset has been selected from fourth column to the last column (4 to 23) and separated to train 70% and test 30% by the code below

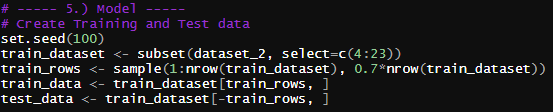


Figure 60: Split the tidied dataset as train 70% and test 30%

2.5.2 Multiple linear regression

For multiple linear regression, function ‘lm()’ has been used. First, the model selects all attributes as the features.



Figure 61: Multiple linear regression model

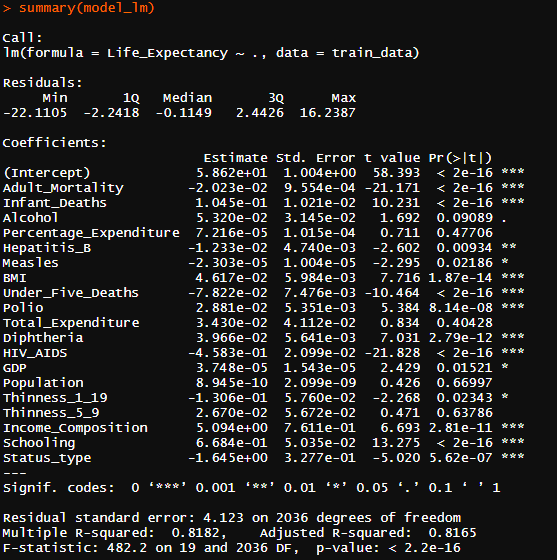


Figure 62: Summary of multiple linear regression model

In addition, I also trained another multiple linear regression model by selected only some features which have significance code as ‘\*\*\*’ from figure 6.2

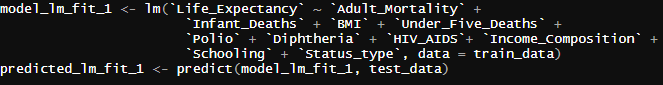


Figure 63: Multiple linear regression model with only high significance features

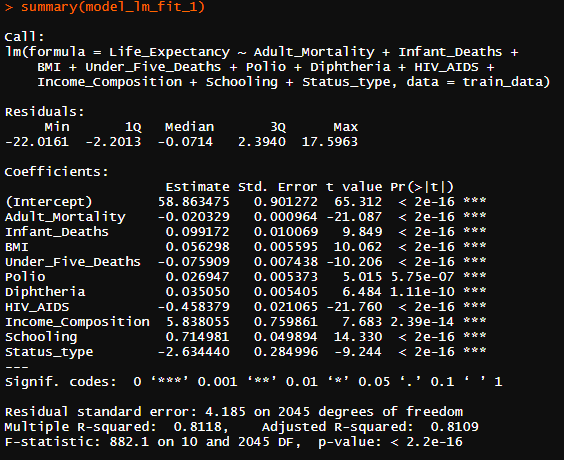


Figure 64: Summary of multiple linear regression model with only high significance features

2.5.3 Random forest

Random forest is another algorithm that was applied to create the model. Function ‘randomForest()’ is used.



Figure 65: Random forest model

2.5.4 Gradient boosting

Gradient boosting is another technique that is used in this project. Some parameters need to be set – distribution type as Gaussian, number of trees as 10,000, learning rate (shrinkage) = 0.01.

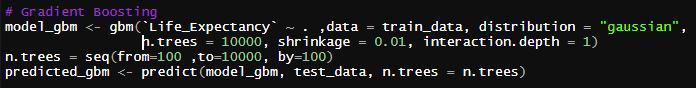


Figure 66: Gradient boosting model

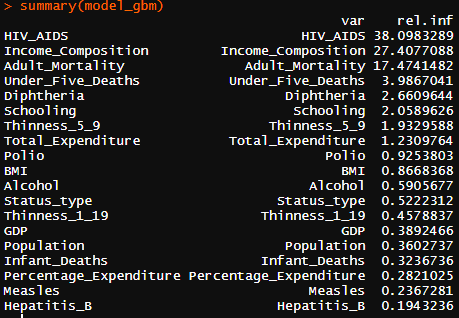


Figure 67: Summary of gradient booster model

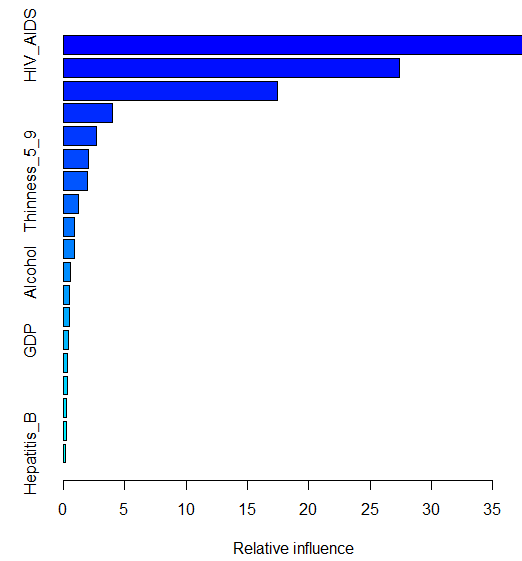


Figure 68: Visualization of gradient booster model summary

2.5.5 Supported vector machine (SVM)

Supported vector machine is mainly used in classification but it can also generate regression model by set the parameter ‘type’ as ‘eps-regression’ and ‘kernel’ as ‘linear’

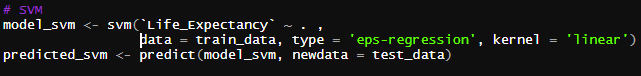


Figure 69: Supported vector machine model

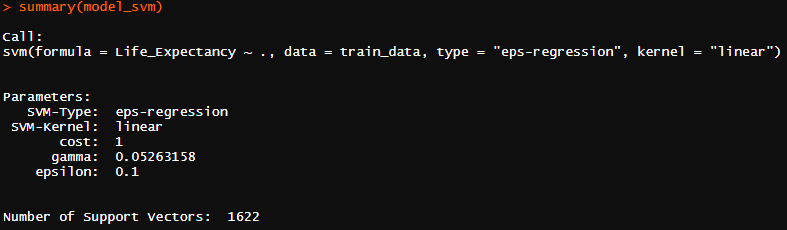


Figure 70: Summary of supported vector machine model

2.6 Accuracy

To evaluate each model efficient. Predictions from each model need to be compared with the test dataset. The criteria that is used consists of mean-squared error, root-mean-squared error, and min/max accuracy.

|  |  |
| --- | --- |
| **Tools** | **Criteria** |
| Mean-squared error (MSE) | Less is good |
| Root-mean-squared error (RMSE) | Less is good |
| Min/max accuracy | More is good |

Table 2: Accuracy tools and criteria

2.6.1 Mean-squared error (MSE)

Five models accuracy were calculated by ‘mse()’ function.

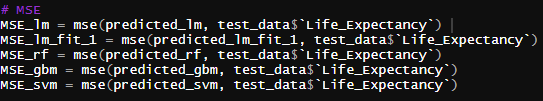


Figure 71: Mean-squared error calculation for five model – multiple linear regression, multiple linear regression with high significance features, random forest, gradient boost, and supported vector machine.

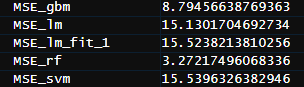


Figure 72: Result of mean-squared error

2.6.2 Root-mean-squared error (RMSE)

Five models accuracy were calculated by root-mean-squared error formula.

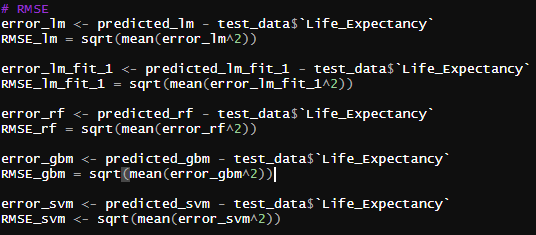


Figure 73: Root-mean-squared error calculation for five model – multiple linear regression, multiple linear regression with high significance features, random forest, gradient boost, and supported vector machine.



Figure 74: Result of root-mean-squared error

2.6.3 Min/max accuracy

Five models accuracy were calculated by min/max accuracy formula.

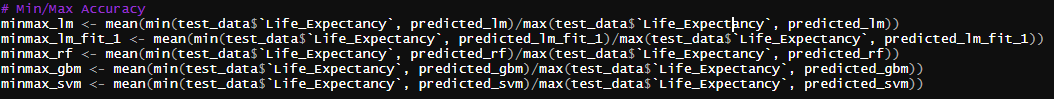


Figure 75: Min/max accuracy calculation for five model – multiple linear regression, multiple linear regression with high significance features, random forest, gradient boost, and supported vector machine.

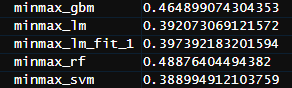


Figure 76: Result of min/max accuracy

1. Summary

All models accuracy are merge as a new data frame and show the result below.

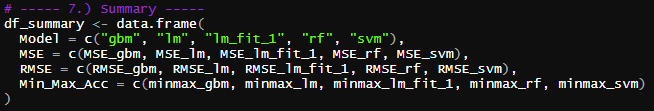


Figure 77: Merge all accuracy as new data frame

From observation, random forest technique gives the best accuracy and least error, followed by gradient boost, multiple linear regression with only high significant features, multiple linear regression, and supported vector machine (SVM) respectively.

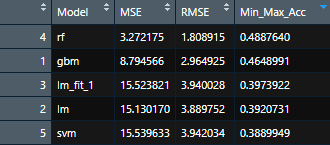


Figure 78: All models accuracy, sorted by min/max accuracy

In the aspect of life expectancy, the factor that has the highest directly impact to life expectancy is schooling and the highest inversely impact to life expectancy is adult mortality. The rest has been ordered from lowest to highest as shown in figure 59.

1. Appendix

#####################################################

#

# INC 491 - Data Science and Intelligent Techniques

# Mini Project - Credit Approval Data Set

#

# Life Expectancy (WHO)

# Statistical Analysis on factors influencing Life

# Expectancy

#

# Developed by: Twitty Manymoon

#

#####################################################

# ----- 1.) Import -----

library**(**gbm**)**

library**(**readr**)**

library**(**e1071**)**

library**(**Metrics**)**

library**(**ggplot2**)**

library**(**reshape2**)**

library**(**corrplot**)**

library**(**tidyverse**)**

library**(**randomForest**)**

dataset **<-** read\_csv**(**"Life Expectancy Data.csv"**)**

# ----- 2.) Tidy up -----

# pattern of data, and find modelling strategies

str**(**dataset**)**

summary**(**dataset**)**

View**(**dataset**)**

# NA removeing

#dataset\_2 <- na.omit(dataset)

dataset\_2 **<-** dataset

**for(**i **in** 4**:**ncol**(**dataset\_2**)){**

dataset\_2**[**is.na**(**dataset\_2**[**,i**])**, i**]** **<-** colMeans**(**dataset\_2**[**,i**]**, na.rm **=** **TRUE)**

**}**

# ----- 3.) Transform -----

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "Life expectancy"**]** **<-** "Life\_Expectancy"

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "Adult Mortality"**]** **<-** "Adult\_Mortality"

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "infant deaths"**]** **<-** "Infant\_Deaths"

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "percentage expenditure"**]** **<-** "Percentage\_Expenditure"

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "Hepatitis B"**]** **<-** "Hepatitis\_B"

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "under-five deaths"**]** **<-** "Under\_Five\_Deaths"

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "Total expenditure"**]** **<-** "Total\_Expenditure"

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "HIV/AIDS"**]** **<-** "HIV\_AIDS"

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "thinness 1-19 years"**]** **<-** "Thinness\_1\_19"

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "thinness 5-9 years"**]** **<-** "Thinness\_5\_9"

names**(**dataset\_2**)[**names**(**dataset\_2**)** **==** "Income composition of resources"**]** **<-** "Income\_Composition"

dataset\_2**$**Status **<-** as.factor**(**dataset\_2**$**Status**)**

dataset\_2**$**Status\_type **<-** as.numeric**(**dataset\_2**$**Status**)**

df\_life\_developing **<-** subset**(**dataset\_2, dataset\_2**$**Status **==** "Developing"**)**

df\_life\_developed **<-** subset**(**dataset\_2, dataset\_2**$**Status **==** "Developed"**)**

# ----- 4.) Visualize -----

# Status

df\_status **<-** data.frame**(**

Status **=** c**(**"Developing", "Developed"**)**,

Sum **=** c**(**sum**(**dataset\_2**$**Status **==** "Developing"**)\***100**/**nrow**(**dataset\_2**)**,

sum**(**dataset\_2**$**Status **==** "Developed"**)\***100**/**nrow**(**dataset\_2**))**

**)**

ggplot**(**df\_status, aes**(**x**=**"", y**=**Sum, fill**=**Status**))** **+** geom\_bar**(**stat**=**"identity", width**=**1**)** **+** coord\_polar**(**"y", start**=**0**)**

# : Life Expectancy

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Life\_Expectancy**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Life\_Expectancy**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Life Expectancy'**)** **+**

ylab**(**'Density'**)**

# : Adult Mortality

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Adult\_Mortality**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Adult\_Mortality**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Adult Mortality'**)** **+**

ylab**(**'Density'**)**

# : Infant Deaths

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Infant\_Deaths**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Infant\_Deaths**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Infant Deaths'**)** **+**

ylab**(**'Density'**)**

# : Alcohol

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Alcohol**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Alcohol**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Alcohol'**)** **+**

ylab**(**'Density'**)**

# : Percentage Expenditure

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Percentage\_Expenditure**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Percentage\_Expenditure**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Percentage Expenditure'**)** **+**

ylab**(**'Density'**)**

# : Hepatitis B

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Hepatitis\_B**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Hepatitis\_B**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Hepatitis B'**)** **+**

ylab**(**'Density'**)**

# : Measles

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Measles**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Measles**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Measles'**)** **+**

ylab**(**'Density'**)**

# : BMI

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**BMI**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**BMI**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'BMI'**)** **+**

ylab**(**'Density'**)**

# : Under-5 Deaths

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Under\_Five\_Deaths**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Under\_Five\_Deaths**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Under-5 Deaths'**)** **+**

ylab**(**'Density'**)**

# : Polio

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Polio**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Polio**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Polio'**)** **+**

ylab**(**'Density'**)**

# : Total Expenditure

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Total\_Expenditure**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Total\_Expenditure**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Total Expenditure'**)** **+**

ylab**(**'Density'**)**

# : Diphtheria

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Diphtheria**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Diphtheria**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Diphtheria'**)** **+**

ylab**(**'Density'**)**

# : HIV/AIDS

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**HIV\_AIDS**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**HIV\_AIDS**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'HIV/AIDS'**)** **+**

ylab**(**'Density'**)**

# : GDP

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**GDP**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**GDP**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'GDP'**)** **+**

ylab**(**'Density'**)**

# : Population

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Population**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Population**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Population'**)** **+**

ylab**(**'Density'**)**

# thinness 10-19 years

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Thinness\_1\_19**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Thinness\_1\_19**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Thinness 1-19 yrs.'**)** **+**

ylab**(**'Density'**)**

# thinness 5-9 years

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Thinness\_5\_9**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Thinness\_5\_9**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Thinness 5-9 yrs.'**)** **+**

ylab**(**'Density'**)**

# Income composition of resources

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Income\_Composition**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Income\_Composition**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Income Composition'**)** **+**

ylab**(**'Density'**)**

# Schooling

ggplot**()** **+**

geom\_density**(**data **=** df\_life\_developing, aes**(**x **=** df\_life\_developing**$**Schooling**)**, fill **=** "red", color **=** "red", alpha **=** 0.5**)** **+**

geom\_density**(**data **=** df\_life\_developed, aes**(**x **=** df\_life\_developed**$**Schooling**)**, fill **=** "blue", color **=** "blue", alpha **=** 0.5**)** **+**

xlab**(**'Schooling'**)** **+**

ylab**(**'Density'**)**

# Life Expectancy ~ Adult Mortality

Relationship\_LE\_AdultMortality **<-** ggplot**(**dataset\_2, aes**(**x **=** dataset\_2**$**Life\_Expectancy, y **=** dataset\_2**$**Adult\_Mortality**))**

Relationship\_LE\_AdultMortality **+** geom\_point**(**color **=** "#00AFBB", size **=** 2**)** **+** geom\_smooth**(**method **=** lm**)** **+** xlab**(**'Life Expectancy'**)** **+** ylab**(**'Adult Mortaility'**)**

# Life Expectancy ~ Infant Deaths

Relationship\_LE\_InfantDeaths **<-** ggplot**(**dataset\_2, aes**(**x **=** dataset\_2**$**Life\_Expectancy, y **=** dataset\_2**$**Infant\_Deaths**))**

Relationship\_LE\_InfantDeaths **+** geom\_point**(**color **=** "#FC4E07", size **=** 2**)** **+** geom\_smooth**(**method **=** lm**)** **+** xlab**(**'Life Expectancy'**)** **+** ylab**(**'Infant Deaths'**)**

# Life Expectancy ~ Alcohol

Relationship\_LE\_Alcohol **<-** ggplot**(**dataset\_2, aes**(**x **=** dataset\_2**$**Life\_Expectancy, y **=** dataset\_2**$**Alcohol**))**

Relationship\_LE\_Alcohol **+** geom\_point**(**color **=** "#E7B800", size **=** 2**)** **+** geom\_smooth**(**method **=** lm**)** **+** xlab**(**'Life Expectancy'**)** **+** ylab**(**'Alcohol Consumption per Capita (Litre)'**)**

# Life Expectancy ~ Population

Relationship\_LE\_Population **<-** ggplot**(**dataset\_2, aes**(**x **=** dataset\_2**$**Life\_Expectancy, y **=** dataset\_2**$**Population**))**

Relationship\_LE\_Population **+** geom\_point**(**color **=** "#00FF0A", size **=** 2**)** **+** geom\_smooth**(**method **=** lm**)** **+** xlab**(**'Life Expectancy'**)** **+** ylab**(**'Population'**)**

# Life Expectancy ~ Schooling

Relationship\_LE\_Population **<-** ggplot**(**dataset\_2, aes**(**x **=** dataset\_2**$**Life\_Expectancy, y **=** dataset\_2**$**Schooling**))**

Relationship\_LE\_Population **+** geom\_point**(**color **=** "#FF4D88", size **=** 2**)** **+** geom\_smooth**(**method **=** lm**)** **+** xlab**(**'Life Expectancy'**)** **+** ylab**(**'Schooling'**)**

# Correlation

correlation **<-** cor**(**dataset\_2 %>% select**(**4**:**23**))**

melted\_correlation **<-** melt**(**correlation**)**

melted\_correlation **<-** subset**(**melted\_correlation, melted\_correlation**$**Var2 **==** "Life\_Expectancy"**)**

melted\_correlation\_sorted **<-** melted\_correlation**[**order**(**melted\_correlation**$**value**)**, **]**

cor **=** cor**(**dataset\_2**[**4**:**23**])**

corrplot**(**cor, method **=** "color"**)** # heat map

ggplot**(**data **=** melted\_correlation\_sorted, aes**(**x**=**Var2, y**=**Var1, fill**=**value**))** **+** geom\_tile**()** # life expectancy only

# ----- 5.) Model -----

# Create Training and Test data

set.seed**(**100**)**

train\_dataset **<-** subset**(**dataset\_2, select**=**c**(**4**:**23**))**

train\_rows **<-** sample**(**1**:**nrow**(**train\_dataset**)**, 0.7**\***nrow**(**train\_dataset**))**

train\_data **<-** train\_dataset**[**train\_rows, **]**

test\_data **<-** train\_dataset**[-**train\_rows, **]**

# Multiple Linear Regression

model\_lm **<-** lm**(**`Life\_Expectancy` **~** ., data **=** train\_data**)**

predicted\_lm **<-** predict**(**model\_lm, test\_data**)**

model\_lm\_fit\_1 **<-** lm**(**`Life\_Expectancy` **~** `Adult\_Mortality` **+**

`Infant\_Deaths` **+** `BMI` **+** `Under\_Five\_Deaths` **+**

`Polio` **+** `Diphtheria` **+** `HIV\_AIDS`**+** `Income\_Composition` **+**

`Schooling` **+** `Status\_type`, data **=** train\_data**)**

predicted\_lm\_fit\_1 **<-** predict**(**model\_lm\_fit\_1, test\_data**)**

# Random Forest

model\_rf **<-** randomForest**(**`Life\_Expectancy` **~** ., data **=** train\_data**)**

predicted\_rf **<-** predict**(**model\_rf, test\_data**)**

# Gradient Boosting

model\_gbm **<-** gbm**(**`Life\_Expectancy` **~** . ,data **=** train\_data, distribution **=** "gaussian",

n.trees **=** 10000, shrinkage **=** 0.01, interaction.depth **=** 1**)**

n.trees **=** seq**(**from**=**100 ,to**=**10000, by**=**100**)**

predicted\_gbm **<-** predict**(**model\_gbm, test\_data, n.trees **=** n.trees**)**

# SVM

model\_svm **<-** svm**(**`Life\_Expectancy` **~** . ,

data **=** train\_data, type **=** 'eps-regression', kernel **=** 'linear'**)**

predicted\_svm **<-** predict**(**model\_svm, newdata **=** test\_data**)**

# ----- 6.) Accuracy -----

# MSE

MSE\_lm **=** mse**(**predicted\_lm, test\_data**$**`Life\_Expectancy`**)**

MSE\_lm\_fit\_1 **=** mse**(**predicted\_lm\_fit\_1, test\_data**$**`Life\_Expectancy`**)**

MSE\_rf **=** mse**(**predicted\_rf, test\_data**$**`Life\_Expectancy`**)**

MSE\_gbm **=** mse**(**predicted\_gbm, test\_data**$**`Life\_Expectancy`**)**

MSE\_svm **=** mse**(**predicted\_svm, test\_data**$**`Life\_Expectancy`**)**

# RMSE

error\_lm **<-** predicted\_lm **-** test\_data**$**`Life\_Expectancy`

RMSE\_lm **=** sqrt**(**mean**(**error\_lm**^**2**))**

error\_lm\_fit\_1 **<-** predicted\_lm\_fit\_1 **-** test\_data**$**`Life\_Expectancy`

RMSE\_lm\_fit\_1 **=** sqrt**(**mean**(**error\_lm\_fit\_1**^**2**))**

error\_rf **<-** predicted\_rf **-** test\_data**$**`Life\_Expectancy`

RMSE\_rf **=** sqrt**(**mean**(**error\_rf**^**2**))**

error\_gbm **<-** predicted\_gbm **-** test\_data**$**`Life\_Expectancy`

RMSE\_gbm **=** sqrt**(**mean**(**error\_gbm**^**2**))**

error\_svm **<-** predicted\_svm **-** test\_data**$**`Life\_Expectancy`

RMSE\_svm **<-** sqrt**(**mean**(**error\_svm**^**2**))**

# Min/Max Accuracy

minmax\_lm **<-** mean**(**min**(**test\_data**$**`Life\_Expectancy`, predicted\_lm**)/**max**(**test\_data**$**`Life\_Expectancy`, predicted\_lm**))**

minmax\_lm\_fit\_1 **<-** mean**(**min**(**test\_data**$**`Life\_Expectancy`, predicted\_lm\_fit\_1**)/**max**(**test\_data**$**`Life\_Expectancy`, predicted\_lm\_fit\_1**))**

minmax\_rf **<-** mean**(**min**(**test\_data**$**`Life\_Expectancy`, predicted\_rf**)/**max**(**test\_data**$**`Life\_Expectancy`, predicted\_rf**))**

minmax\_gbm **<-** mean**(**min**(**test\_data**$**`Life\_Expectancy`, predicted\_gbm**)/**max**(**test\_data**$**`Life\_Expectancy`, predicted\_gbm**))**

minmax\_svm **<-** mean**(**min**(**test\_data**$**`Life\_Expectancy`, predicted\_svm**)/**max**(**test\_data**$**`Life\_Expectancy`, predicted\_svm**))**

# ----- 7.) Summary -----

df\_summary **<-** data.frame**(**

Model **=** c**(**"gbm", "lm", "lm\_fit\_1", "rf", "svm"**)**,

MSE **=** c**(**MSE\_gbm, MSE\_lm, MSE\_lm\_fit\_1, MSE\_rf, MSE\_svm**)**,

RMSE **=** c**(**RMSE\_gbm, RMSE\_lm, RMSE\_lm\_fit\_1, RMSE\_rf, RMSE\_svm**)**,

Min\_Max\_Acc **=** c**(**minmax\_gbm, minmax\_lm, minmax\_lm\_fit\_1, minmax\_rf, minmax\_svm**)**

**)**

summary**(**model\_lm**)**

summary**(**model\_lm\_fit\_1**)**

summary**(**model\_rf**)**

summary**(**model\_gbm**)**

summary**(**model\_svm**)**