Project 1: Exploration and Modeling of Life Expectancy

MSDS-6372

Authors: Taylor Bonar, Alexandre Jasserme, & Halle Purdom

When examining life expectancy, traditionally studies have used demographic variables, income composition, and mortality rates, but what about other variables? Could we use current records of immunization records, mortality, economic, social, and other, more recent health related factors to identify trends and behaviors on life expectancy between various countries? In this project, we will explore some potential new variables from the Global Health Observatory (GHO) from the World Health Organization (WHO) in combination with United Nations economic data.

Looking at the Data

The original dataset is derived from a mixture of the Global Health Observatory (GHO) data repository managed by the World Health Organization for health statuses and other related factors for countries while economic data (GDP, percent expenditure, etc.) was collected from the United Nations website. Although the dataset is not a full, comprehensive list from both sources, critical factors remained that are believed to be more representative of influence from developments in the health industry. In a preliminary analysis of the dataset, it was found that there is a significant amount of missing data (~3% total) where population, Hepatitis B and GDP make up more than half of the missing data (~2% missing for population/GDP/HepB).

However, we determined that we could fill in missing data for Population and GDP from the World Bank as a supplement to the United Nations data gathered. This decision made sense as the World Bank is a financial institution that collects economic data and provides loans and grants to variety of countries, while being a separate part of the UN system. Instead, each institution is owned by its member governments.

Additionally, we found that many missing fields were from 2013 and 2015. To limit the amount of data we needed to gather, we decided to drop 2013 onward to prevent potential outliers. We also chose to remove countries that did not have the full range of data we were looking at from 2000 to 2013. These 10 countries were mostly small islands and small countries. Lastly, we chose to drop South Sudan from our analysis, as there were many missing or questionable datapoints from the country. South Sudan gained independence from the Republic of the Sudan in 2011, and Sudan is already in our dataset. By restricting the year range, excluding the 11 select countries, and filling in the missing population and GDP observations, we were able to successfully reduce the missing data from 3% to 1.5%.

Lastly, we reduced some variables from our models by examining a Pearson's correlation matrix (Appendix 2.C.i) between all the variables. We found that the following pairs had a high correlation:

- GDP and Percent Expenditure (.9) – we decided to keep GDP as we were able to supplement the missing data and there were potential errors of min/max values for Percent Expenditure.
- Thinness of 5 to 9 and Thinness of 1-19 (.9) -- chosen to keep thinness of 1-19 as it encompassed a more comprehensive age range of youths.
- Under-5 Deaths and Infant Deaths (1) -- These were 1:1 correlated, either variable can be chosen.

Comprehensive Variable List from Original Dataset and Examination of Pearson's Correlation:

- Country: this variable is only an identifier and does not offer predictive ability so it will not be considered for the model; rows related to countries where we did not have data for all the years have been removed
- Year: year 2013 to 2015 will not be considered in the model because of high percentage of missing values for 2013 and 2015 and to have only consecutive years
- Status: this will be a candidate for the model after adjusting some incorrect data
- Life.expectancy: models will be built to predict this value; for interpretability and valid comparison between models, we will not log transform this value.
- Adult.Mortality: this will be a candidate for the model.
- Infant.deaths: this variable has an extremely high correlation (0.997) with "Under.five.deaths". These variables have no missing values and we have decided to keep "infant.deaths" as candidate.
- Alcohol: this will be a candidate for the model.
- Percentage.expenditure: this variable has a high correlation (0.9) with "GDP", and it will not be considered as candidate for the model.
- Hepatitis.B: this will be a candidate for the model.
- Measles: this will be a candidate for the model.
- BMI: this will be a candidate for the model.
- Under.five.deaths: this variable has an extremely high correlation (0.997) with "infant.deaths". These variables have no missing values and we have decided to keep "infant.deaths" as candidate.
- Polio: this will be a candidate for the model.
- Total.expenditure: this will be a candidate for the model.
- Diphtheria: this will be a candidate for the model.
- HIV.AIDS: this will be a candidate for the model; to help with constant variance and linearity assumptions, we will use the log of this variable for the parametric models.
- GDP: this variable has a high correlation (0.9) with "percentage.expenditure" and it will be used it as a candidate variable; to help with constant variance and linearity assumptions, we will use the log of this variable for the parametric models.
- Population: this will be a candidate for the model.
- Thinness..1.19.years: this variable has a high correlation (0.9) with "Thinness.5.9.years". They have the same number of missing values (34) and we have decided to keep this variable as candidate.
- Thinness.5.9.years: this variable has a high correlation (0.9) with "Thinness..1.19.years". They have the same number of missing values (34) and we have decided to keep "thinness..1.19.years" as candidate.
- Income.composition.of.resources: this will be a candidate for the model.
- Schooling: this will be a candidate for the model.

In objectives one and two, we will be producing five models using a variety of parametric and non-parametric approaches using the variables specified above. To achieve this, we are adapting code from RLab4.R that was provided in Unit 6 of DS6372.

Objective 1

Parametric Model: Multiple Linear Regression with Stepwise Variable Selection

With the information gathered from our exploratory data analysis, we can build a first model to interpret any patterns or potential key relationships. In our first model, we have chosen to use the stepwise variable selection technique as our automatic variable selector to reduce statistically insignificant variables. This decision, in turn, will allow us to analyze a smaller set of variables to create a more interpretable model. Stepwise variable selection chooses which variables to add to the multiple linear regression model. First, variables are added one by one to the model and tested for significance. Then at each step after a variable is added, the model's current variables are retested to make sure no variable's significance dropped below a certain level.

Checking the Assumptions

Linearity

We are assuming the relationship between each explanatory variable and the response is a linear relationship. See Appendix 2.C.i and 2.C.ii for scatterplots of each variable's relationship with life expectancy. From looking at the graphs, we decided to log transform GDP and HIV/AIDS to get a more linear relationship with life expectancy.

Normality

MLR also assumes a normal distribution for the residuals as well as for each response variable for a fixed explanatory value. Based on the residual plots and QQ-plot seen in figure Appendix 4.B the residuals mostly appear to be randomly distributed with both ends of the QQ-plot only slightly deviating from the normality line. By observing the scatterplots in figure Appendix 2.C.i and 2.C.ii, the distribution of the response variables for a fixed explanatory is normal, there is not much obvious skewness. However, a lot of the plots show odd distributions of points that are not a perfectly normal distribution.

Constant Variance

The variance of the response variable for a fixed explanatory variable is assumed to be constant, as well as the variance of the residuals for a fixed explanatory variable. In the scatterplots in figure Appendix 2.C.i and 2.C.ii we can see that diphtheria and polio specifically have nonconstant variance, however the log transformation of the response variable did not drastically affect that variance and log transformations of the variables themselves supplied groupings of observations in the plots while the increasing variance remained. The residuals appear to have a mostly even variance throughout as seen in the residual plots in figure Appendix 4.B.

Independence of Observations

It is assumed all observations in the study are independent for multiple linear regression. In this case, there are violations of independence, and we will continue with caution as we're resampling the same country over a period of years. Specifically, we are looking at multiple observations from the same country, over a range of years. This would ideally be modeled with time series analysis to take this into account, that is however not the goal of this specific work. In addition, to complete time series analysis we would need a wider range of data.

Influential Point Analysis

In looking at the Cook's d plots in Appendix 4.b, all points are around or below 0.03, no point has a particularly high Cook's d value comparatively. In looking at the leverage plot in Appendix 4.b, no points showed a significantly high leverage.

Final Model

To produce the final model, we used the following 17 predictors for the stepwise analysis to run through and find the optimal variables for the model. For the importance of each variable from the stepwise analysis, see Appendix 4.C.

Developed, Year, AdultMortality, InfantDeaths, Alcohol, Hepatitis B, Measles, BMI, Polio, Total Expenditure, Diphtheria, log(HIV/AIDS, log(GDP), Population, Thinness 1-19 years, Income Composition of Resources, Schooling

Life expectancy = 54.504992 - 2.232 log(HIV/AIDS) + 0.489 Schooling - 0.016 AdultMortality + 6.076 IncCompOfResources + 0.442 log(GDP) + 5.56Developed - 0.003 InfantDeaths + 0.017 Diphtheria - 0.073 Thinness1-19years + 0.072 TotalExpenditure

Interpretation

For untransformed parameters: Schooling, AdultMortality, IncCompOfResources, InfantDeaths, Diphtheria, Thinness1-19years, TotalExpenditure

Holding all other variables constant, a one unit increase or decrease (depending on sign) in the specific response variable is associated with a β increase in life expectancy. See below table for specific values of β for a parameter.

Schooling	Adult Mortality	IncCompOf Resources	Infant Deaths	Diphtheria	Thinness 1-19 years	Total Expenditure
0.489	-0.016	6.076	-0.003	0.017	-0.073	0.072

For logged parameters: GDP and HIV/AIDs

Holding all other response variables constant, a doubling of a specific response variable is associated with a βlog(2) unit change in the mean of life expectancy, where β_i represents the parameter estimates from the model. See below table for specific β and βlog(2) values for a specific parameter.

Variable	GDP	HIV/AIDs
β	0.442	-2.232
βlog(2)	0.306	-1.547

Explanatory Parameter: Status: Developed/Developing

In the model, we adapted the status variable to be called “Developed” where a country was assigned 0 or 1 based on its status of “developed” or “developing”. This essentially creates a different model for developed and developing countries, differing on the intercept by a value of 54.50. Interaction terms between status and other continuous variables were tested and found to produce a less accurate model.

Objective 2

The goal in this analysis is to use the data compiled from the WHO and United Nations to decide which factors predict a country’s life expectancy. In our first model, we used the parametric technique multiple linear regression with the stepwise variable selection technique. For the next model, we will be creating another multiple linear regression with added complexity through the elastic net variable selection technique. And lastly, we will be creating a couple models using nonparametric techniques KNN regression and regression trees.

Parametric Model: Multiple Linear Regression with Elastic Net

For our more complex multiple linear regression model, we will be using the elastic net variable selection technique to produce the model for our data. All the earlier work done to address the assumptions of multiple linear regression also applies for these two models.

The elastic net technique is a hybrid regression model that allows the flexibility of both LASSO and Ridge for dealing with multicollinearity. The technique introduces penalties for handling variance by introducing bias while training. It does this by incorporating the penalties of LASSO and Ridge regression. As a quick refresher, Ridge regression works best when most of the variables are useful and significant, as its penalty does not eliminate coefficients of the variables but rather, shrink their coefficients towards zero, while LASSO regression penalty allows the removal of coefficients by allowing them to be exactly zero. This combination of penalties allows for elastic net to perform variable selection and regularization simultaneously by grouping and shrinking the parameters associated with the correlated variables and will leave them in the equation or remove them all at once.

Final Models

Life expectancy = Year + StatusDeveloping + AdultMortality + InfantDeaths + Alchohol + PercentageExpenditure + HepatitisB + Measles + BMI + UnderFiveDeaths + Polio + TotalExpenditure + Diphtheria + log(HIV/AIDS) + Thinness1-19years + Thinness5-8years + IncCompOfResources + Schooling + Developed + log(GDP) + Population

When comparing with the stepwise model, the elastic net technique did include other variables that were found previously to have multicollinearity with other parameters and subsequently chosen one or the other. Such variables observed are thinness for 1-19 with thinness for 5-9, income composition of resources with schooling, infant deaths with under five deaths, etc.

Nonparametric Models: An Overview

Parametric approaches to modeling data rely on the assumption that the data is distributed and behaves in a certain way. For example, in linear models we assume that the explanatory and response variables are linearly related. This can be particularly useful, however in situations where this assumption is far from correct the model will perform poorly, specifically in making predictions. Nonparametric approaches are helpful in that they do not assume the data is in any specific form. This method of regression is more flexible and can perform better than parametric methods in some situations depending on the data.

In this case, we will be looking at KNN regression and regression trees as a nonparametric approach to modeling the life expectancy data and comparing its performance to our earlier parametric models.

Nonparametric Model: KNN Regression

K-nearest neighbors (KNN) regression can overfit data depending on the value of k chosen for the model. KNN regression models data by averaging the closest k datapoints to a point to make a prediction. The smaller the k value is, the more flexible the model will be in fitting the data. And a higher k value will yield a smoother fit. The best fitting k for a dataset depends on the bias variance trade-off.

Data overfitting can occur in KNN regression, specifically when the chosen k is exceedingly small. In this analysis we tested the data to find the best value for k in the KNN regression model.

Response Variable: *Life Expectancy*
Explanatory Variables: *Developed + Year + AdultMortality + InfantDeaths + Alcohol + HepatitisB + Measles + BMI + Polio + TotalExpenditure + Diphtheria + HIV/AIDS + GDP + Population + Thinness1-19years + IncCompOfResources + Schooling*

The KNN model was developed from the above 17 variables. In testing many values, the optimal k for the model was found to be k=2, as seen in Appendix 9.B.

Nonparametric Model: Regression Trees

Regression trees create models for prediction by segmenting the training observations into regions called predictor spaces. The training data is split into binary partitions until a certain sized predictor space, or terminal node, is reached. The model can be visualized as a tree, and each split shows the variable and value the separation was made on. Predictions are then made by then taking the mean or mode of the relevant predictor space. Trees that are too complex will lead to overfitting, so a method called pruning is used where you produce a large regression tree, then prune it to create a smaller subtree. Choosing the best subtree for your analysis will result in a model with less variance than the full first regression tree. In this analysis, we tested different terminal node sizes to find the best model for the data.

Different complexity parameters (cp) were tested and the one that provided the lowest RMSE, which was for a cp of 0.0008, was selected. In Appendix 7.A and 7.B, you can see the final regression tree and the rating of the predictor value's importance.

Comparing the Models

Model	RMSE Train	MAE Train	RMSE Validate	MAE Validate
Stepwise	3.4688	2.581109	11.64075	9.15677
Elastic Net w/ Country	1.825657	1.186341	11.93222	9.28125
Elastic Net w/o Country	3.435034	2.558124	11.677036	9.196641
K-Nearest-Neighbor	2.747412	1.728554	11.584792	9.110282
Regression Trees	2.507126	1.793151	11.630447	9.109415

Looking at the results on the training set (85% of the original dataset), the best fit was obtained using the elastic net model (with the country as a factor) using RMSE (Root Mean Squared Error) and Mean Absolute Error (MAE) as selection criteria.

When the models were applied to the validation set (15% of the original dataset), the best performance was obtained with K-Nearest-Neighbor (using RMSE as the selection criterion) and Regression Trees (using MAE as the selection criterion).

An interesting observation is that the elastic net model (with the country) is the best fitting model on the training set but the worst fitting model on the validation set. This demonstrates that adding the country as a factor lead to overfitting of the model on the training set.

The 2 models that performed best on the validation set only have average performance on the training set, so they do not have the same tendency to overfit as elastic net.

In Conclusion

In this analysis and modeling, we attempted to find the factors that were associated with a country's life expectancy and build a predictive model based on those findings. We developed five models using different variable selection techniques with multiple linear regression and the nonparametric approaches KNN regression and regression tree modeling.

From our models, we found that the regression tree model or KNN regression model to be the most accurate and predictive models when examining the validation set's RMSE and MAE values. We found that our elastic net with or without countries were close but tended to overfit on the training dataset while attempting to reduce variance.

Scope of Inference

This study we analyzed was an observational study, so no causal inferences between the variables considered in the models and a country's life expectancy. This dataset was also not a random sample, meaning the results of this study can only be applied to the specific countries in the study from the years 2000 to 2012. Ten countries without the full range of years were removed, most of which were exceedingly small islands and with little to no records. In addition, South Sudan was removed from the dataset. These models therefore cannot be applied to these 11 countries. See *Looking at the Data* for more information and explanation on what data was used in this study.

Many errors were also found in the data in comparing it with World Bank data. For example, over 7 countries across many years had incorrect data for infant deaths and under five deaths, and the smallest population values were also found to be incorrect. The dataset is so large that many errors remained undiscovered. The models were built off this given data, so when any predictions or analysis is done, it must be considered that the erroneous data could lead to models that are misleading.

Further Work

In the given time, we were able to successfully replace the missing values from the population and GDP columns of data (~22% missing for population and ~15% missing for GDP) with data from the World Bank. With more time, first we would have replaced all missing data in the entire dataset, as many columns had missing values, particularly Hepatitis B as it is missing ~23% of its data. Then we would have checked all this data with a second source, given how many errors we found in this preliminary analysis. In the dataset, most countries had data for the range of years 2000 to 2015, so any countries that did not have the full range of years would have supplemented from other sources. All this would leave us with the ability to build a model that could be applied to a wider range in years and that we were more confident in since it would be derived from a dataset that was more complete and less erroneous.

As referenced previously in the independence assumption of multiple linear regression, this dataset spans range of multiple years for each country. Time series could be another, potentially better, possibilities for analysis of the data.

These variables in the dataset are not the only factors that could be correlated with life expectancy. For example, other variables we could collect data on could include country's obesity, motor vehicle accidents, suicide rate, and potential political situations (e.g., war status). We would have liked to conduct complete analysis on this data to completely determine if they have any relationship with life expectancy.

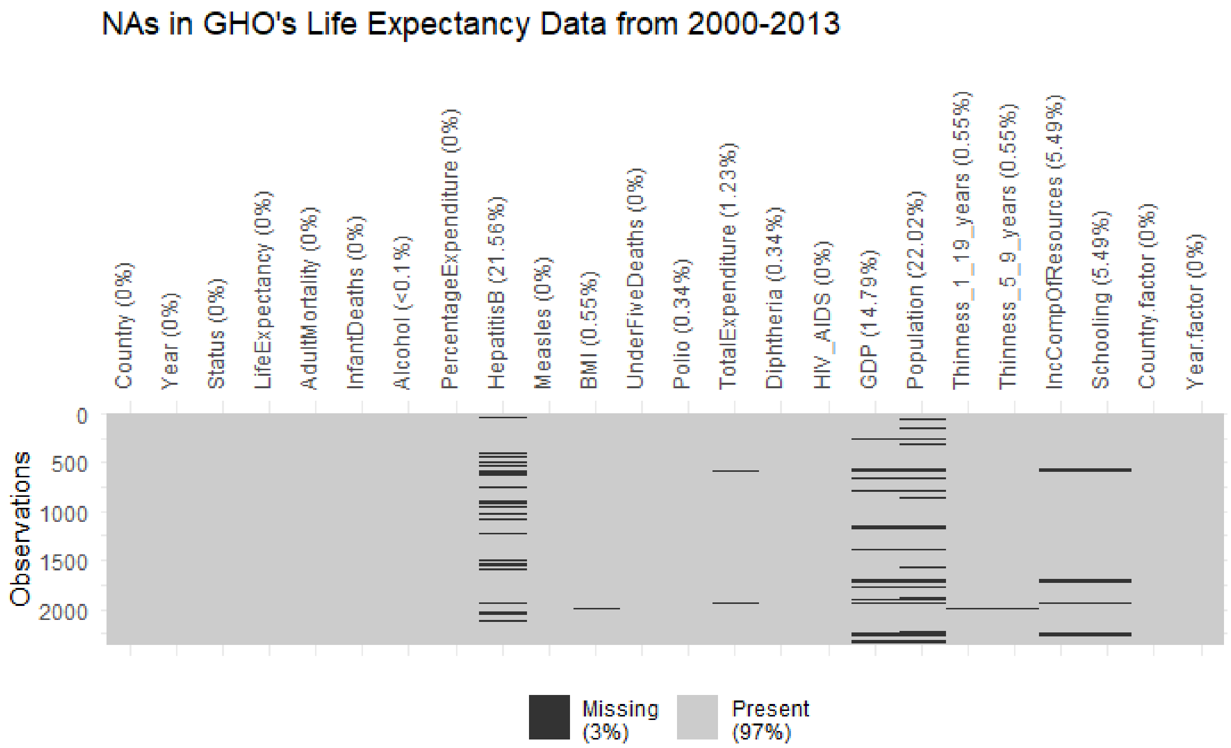
Appendix

1) Github Repository

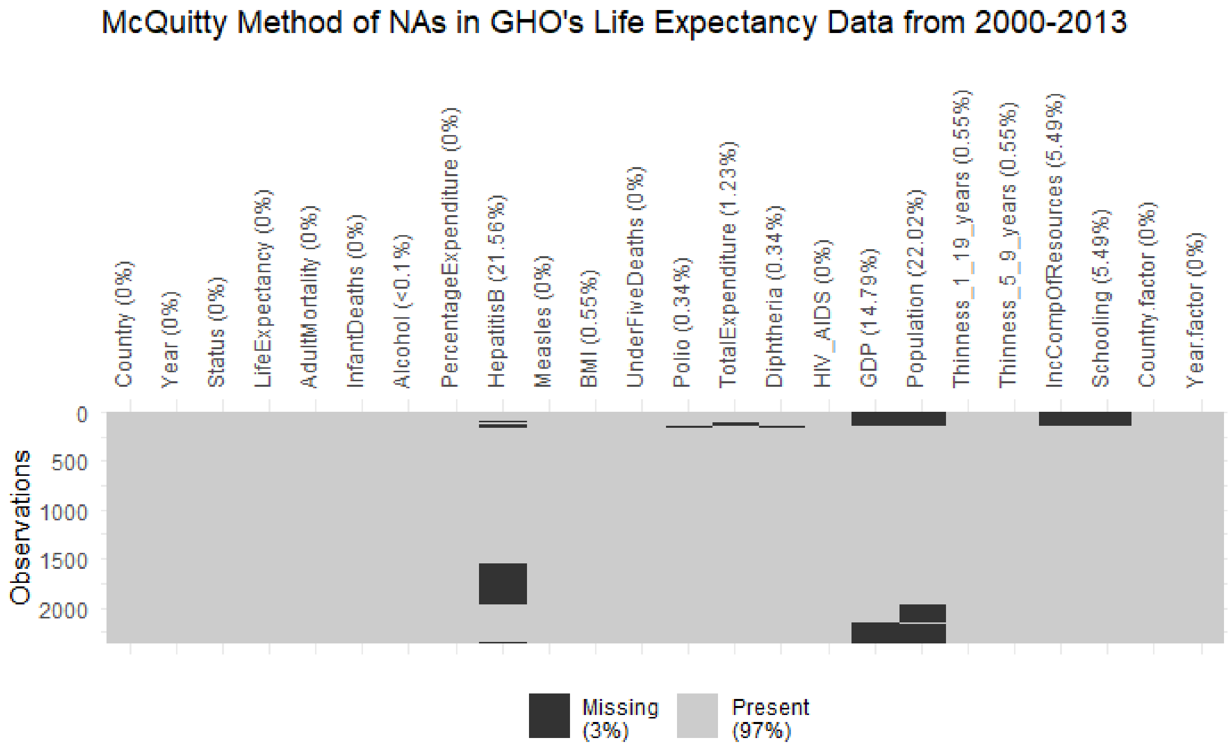
https://github.com/tbonar/MSDS-6372-KaggleLifeExpectancies

2) Initial Dataset Metrics

A) Spread of Missing Data

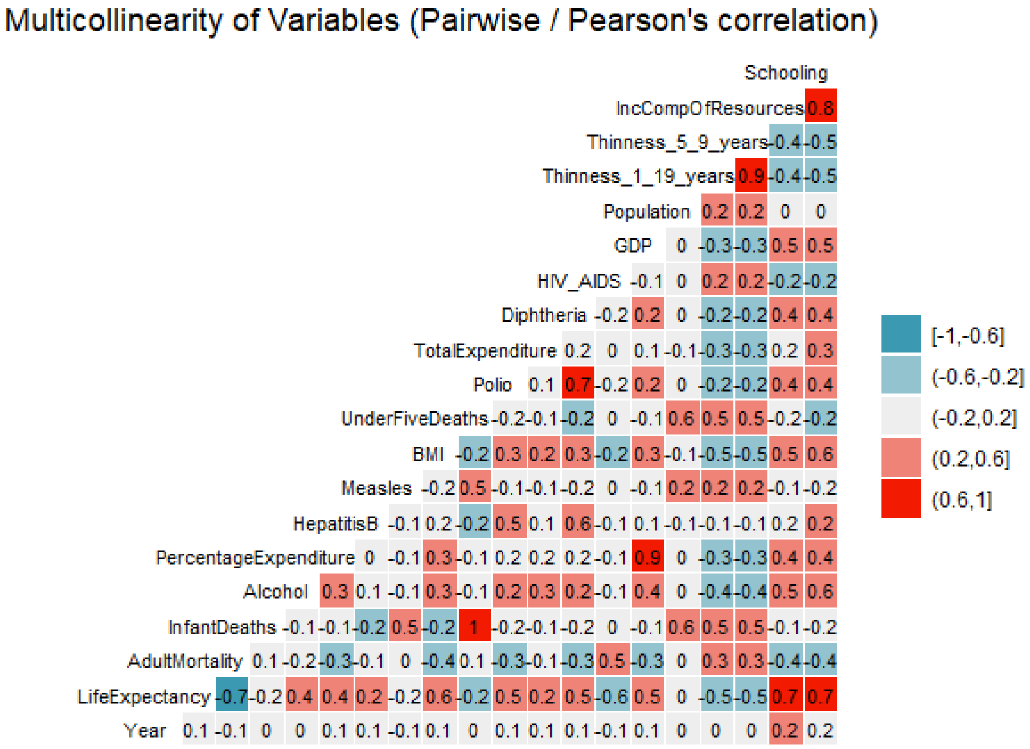


B) Clustered Missing Data

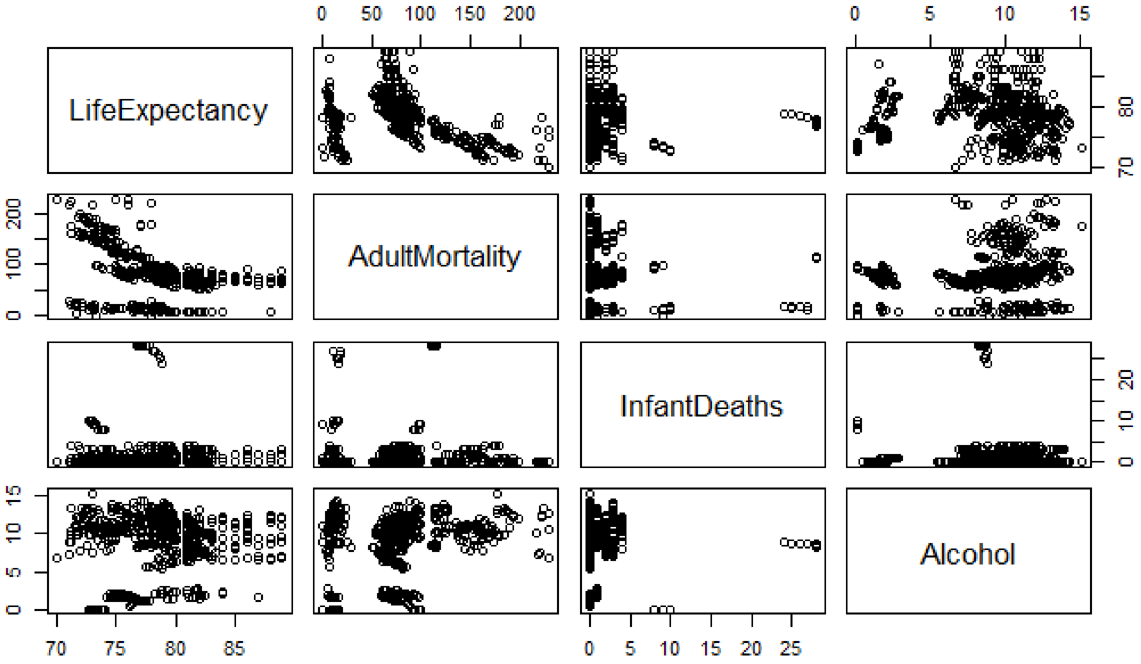


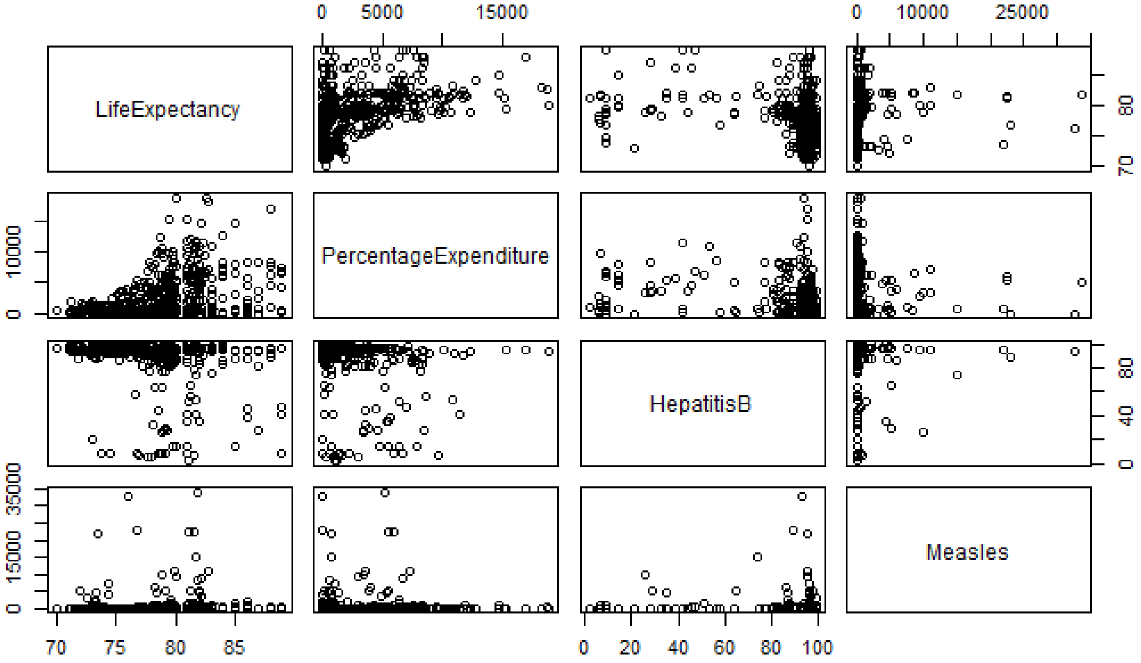
C) Pearson's Correlation

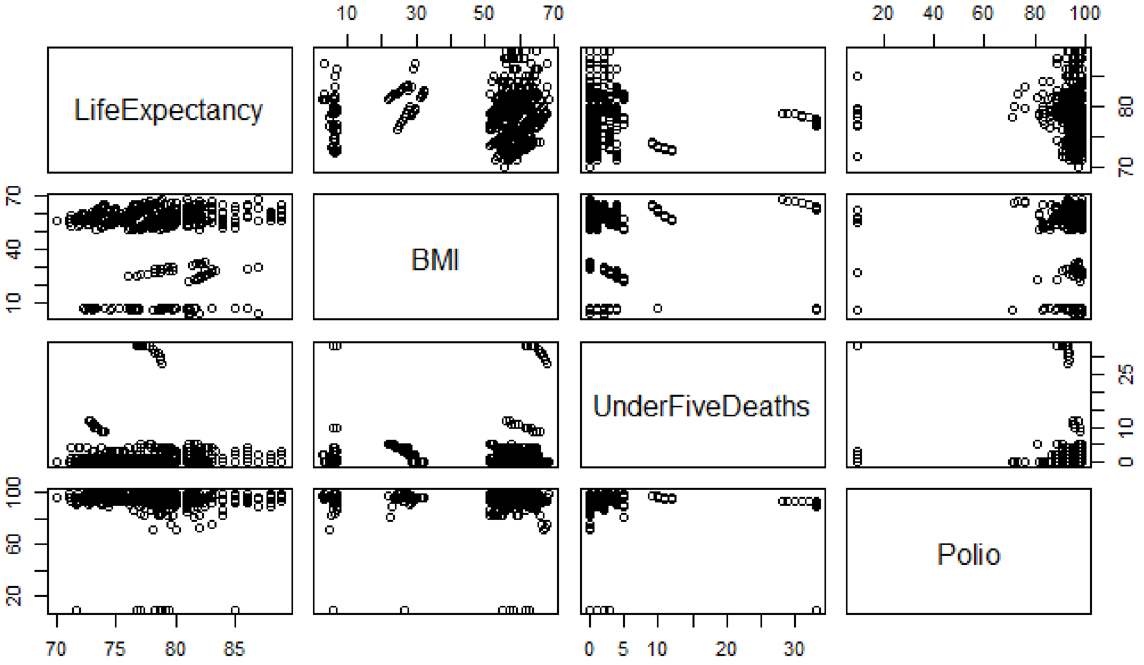
i) Matrix of Variables

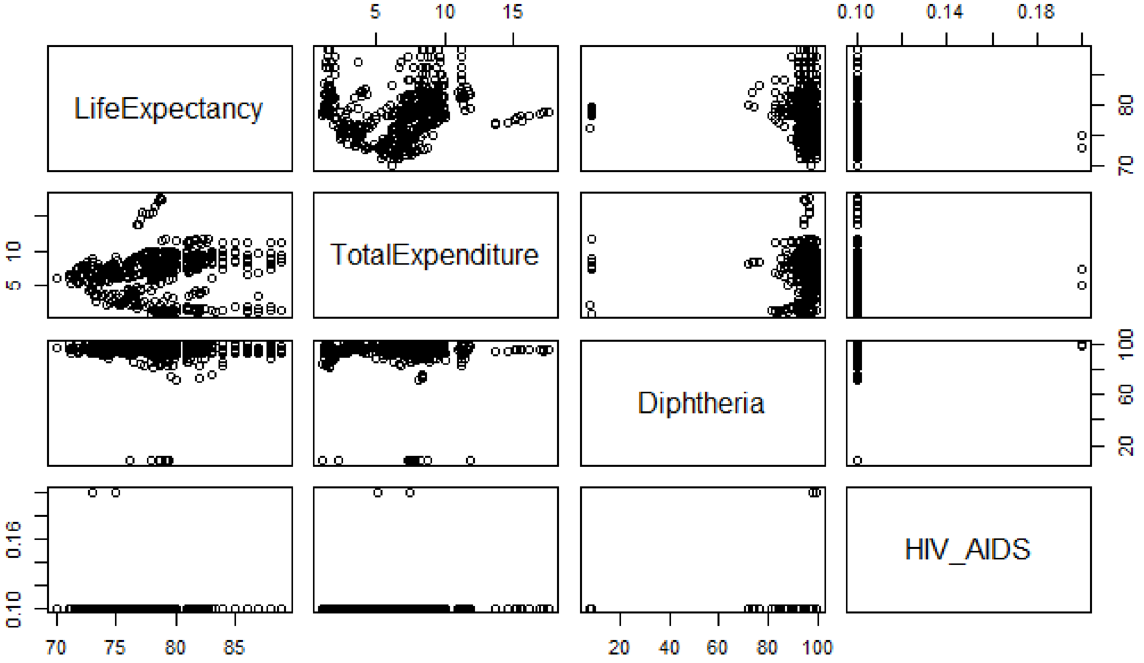


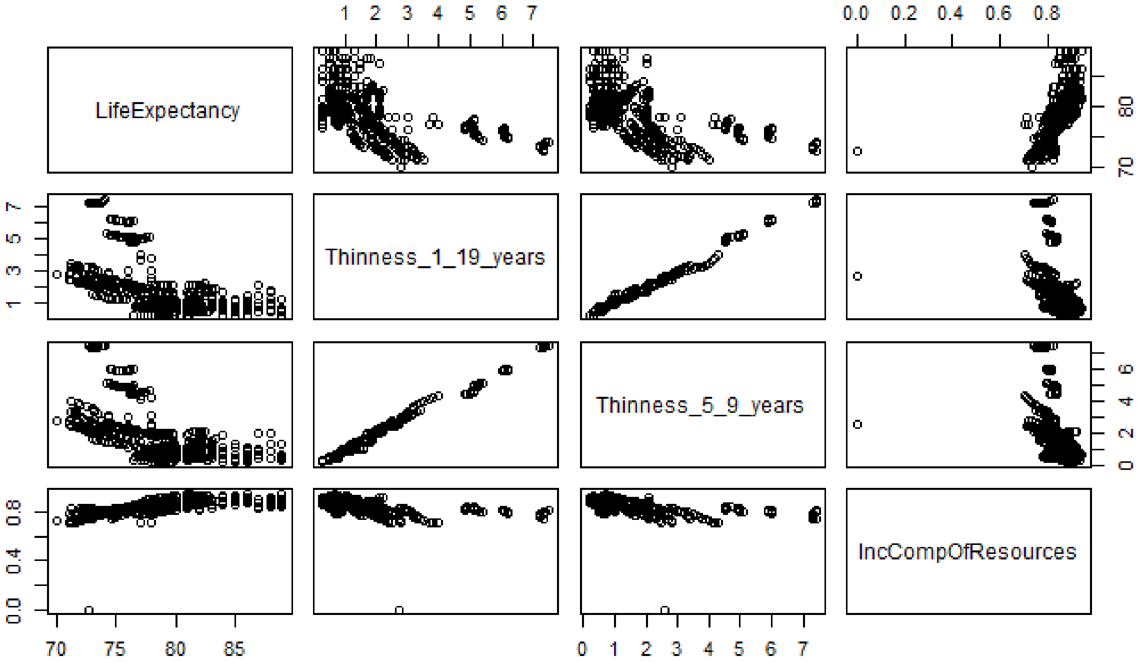
ii) Variables Correlation for Developed Countries

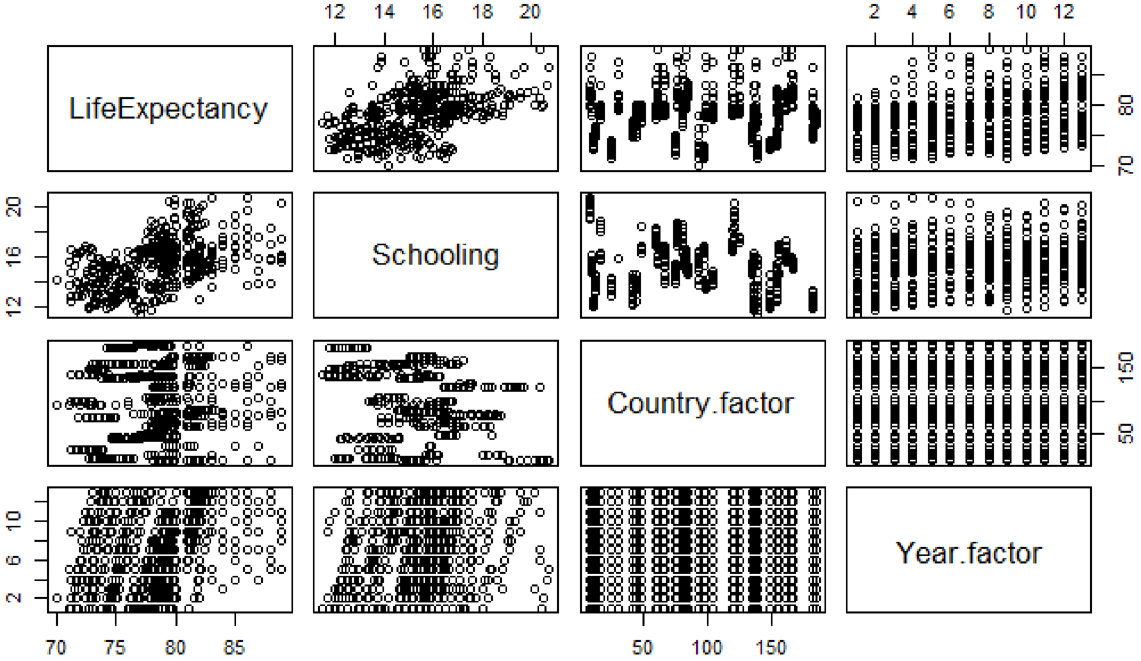


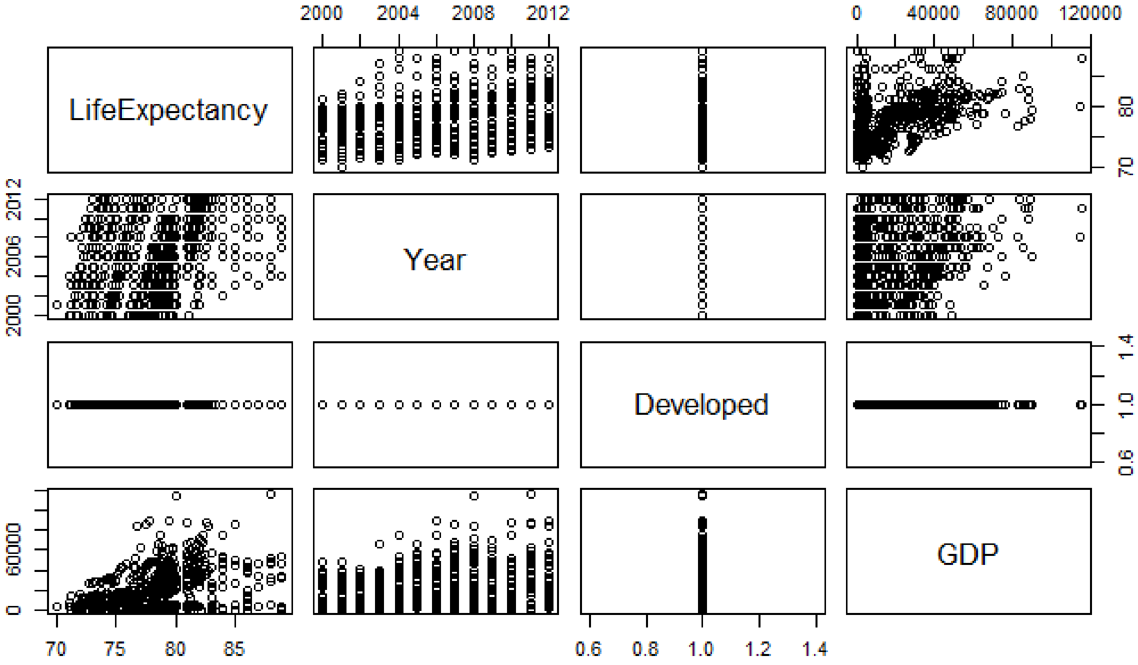



LifeExpectancy
BMI
UnderFiveDeaths
Polio
10
30
50
70
20
40
60
80
100
80
70
70
40
10
25
10
0
100
60
20
70
75
80
85
0
5
10
20
30

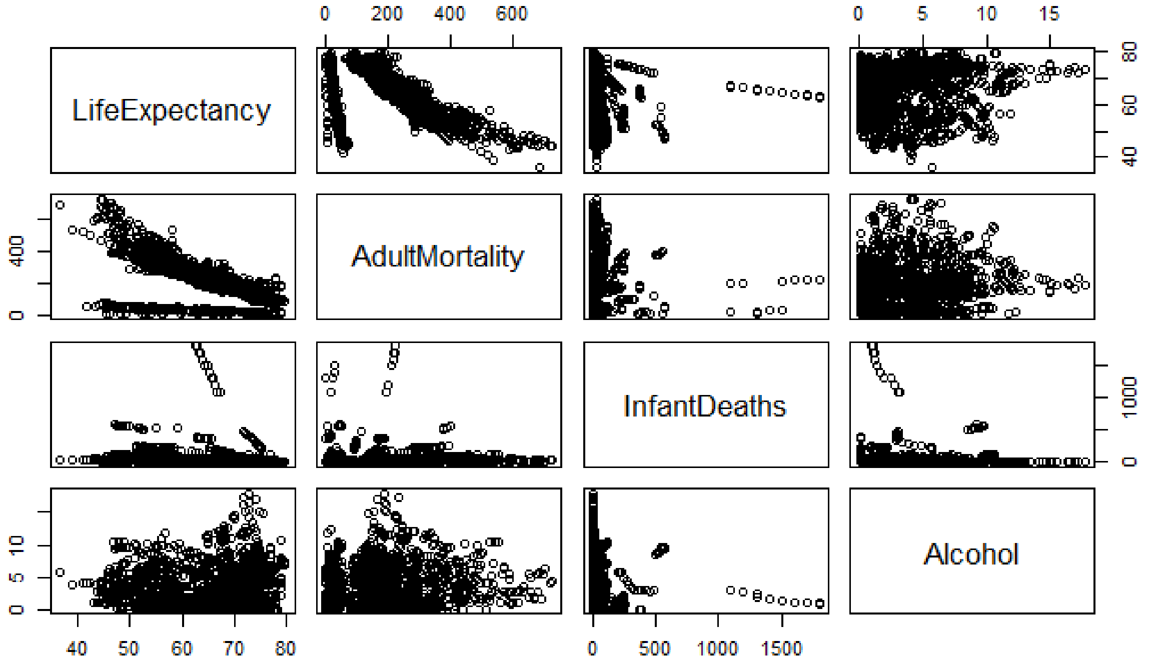

LifeExpectancy
TotalExpenditure
Diphtheria
HIV_AIDS
5
10
15
0.10
0.14
0.18
80
70
10
5
100
60
20
0.16
0.10
70
75
80
85
20
40
60
80
100

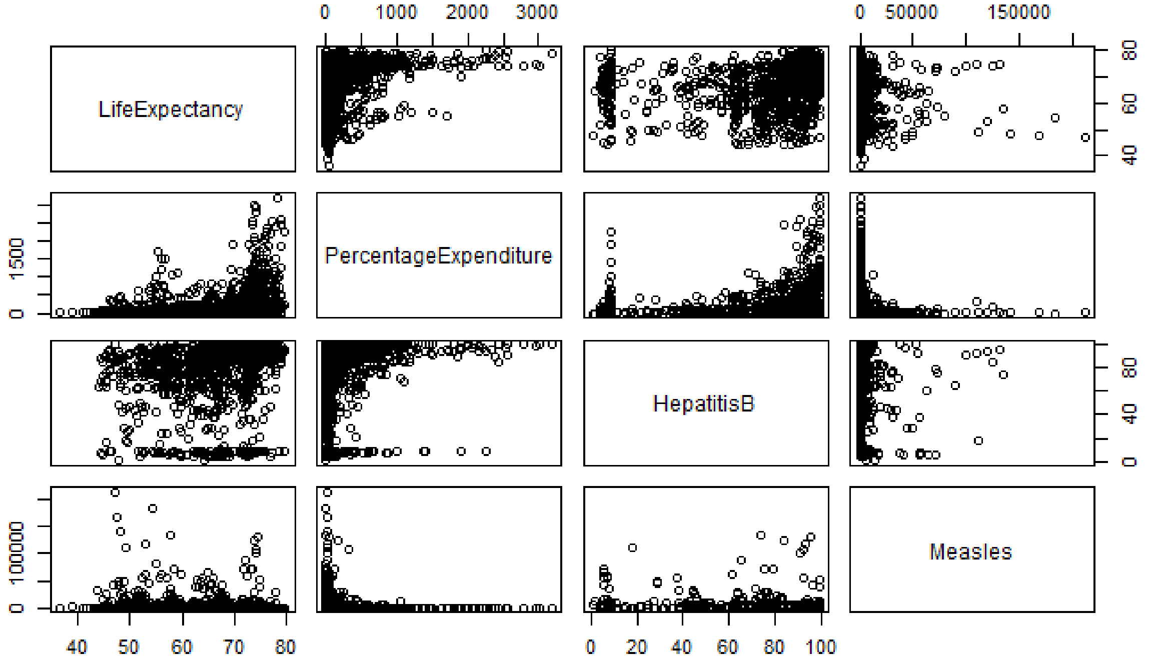

LifeExpectancy
Thinness_1_19_years
Thinness_5_9_years
IncCompOfResources
1 2 3 4 5 6 7
0.0 0.2 0.4 0.6 0.8
70 80
1 3 5 7
0 2 4 6
0.0 0.4 0.8
70 75 80 85
0 1 2 3 4 5 6 7

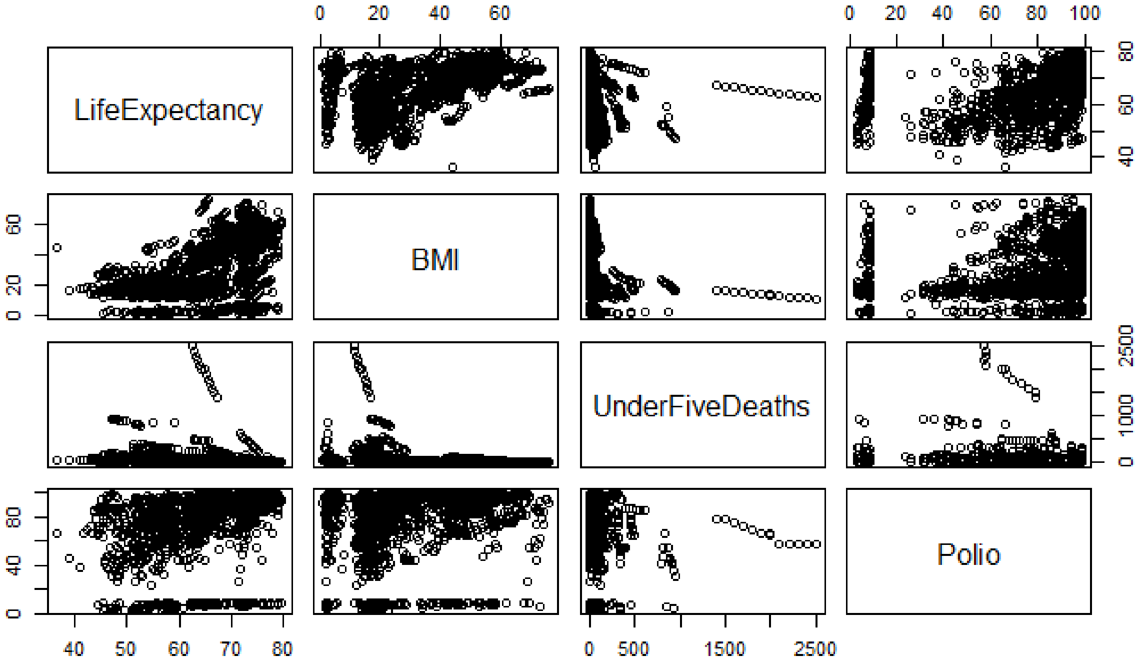

LifeExpectancy
Schooling
Country.factor
Year.factor
12 14 16 18 20
2 4 6 8 10 12
70 80
12 16 20
50 150
2 6 10
70 75 80 85
50 100 150

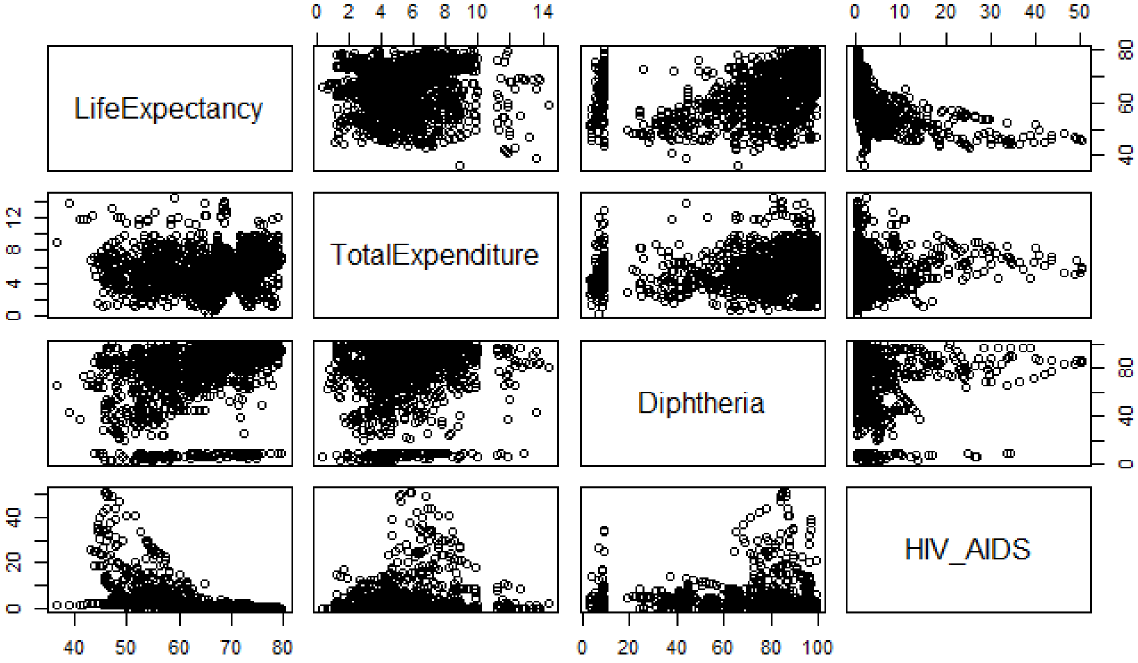

LifeExpectancy
Year
Developed
GDP
2000
2004
2008
2012
0
40000
80000
120000
70
80
2000
2006
2012
0.6
1.0
1.4
0
60000
70
75
80
85
0.6
0.8
1.0
1.2
1.4

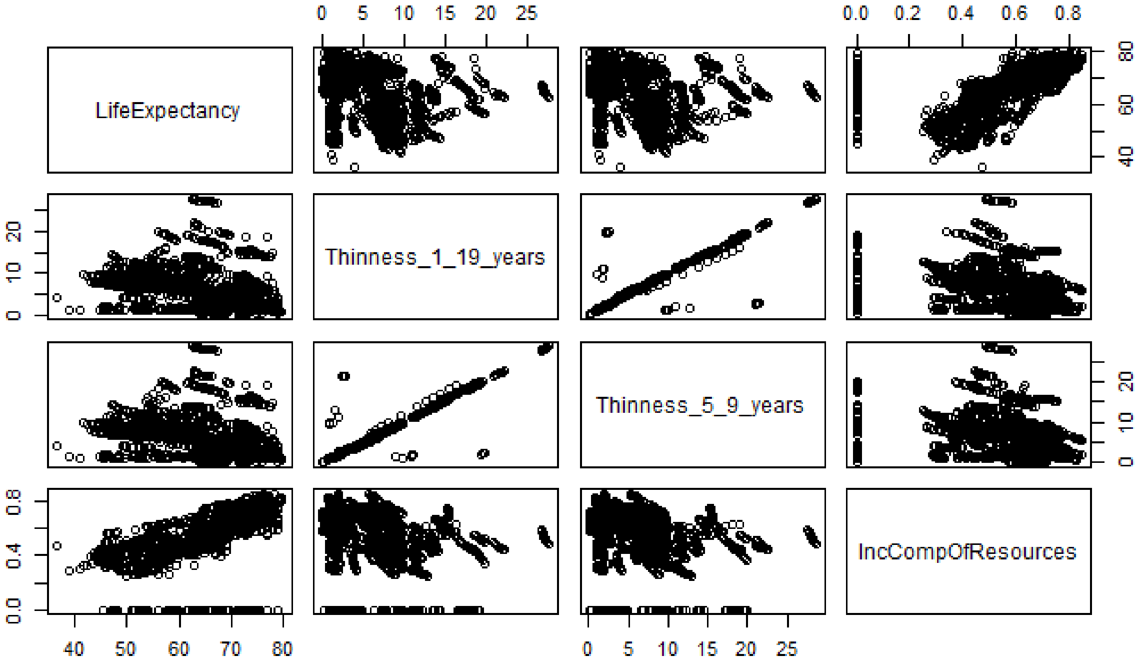
iii) Variables Correlation for Developing Countries

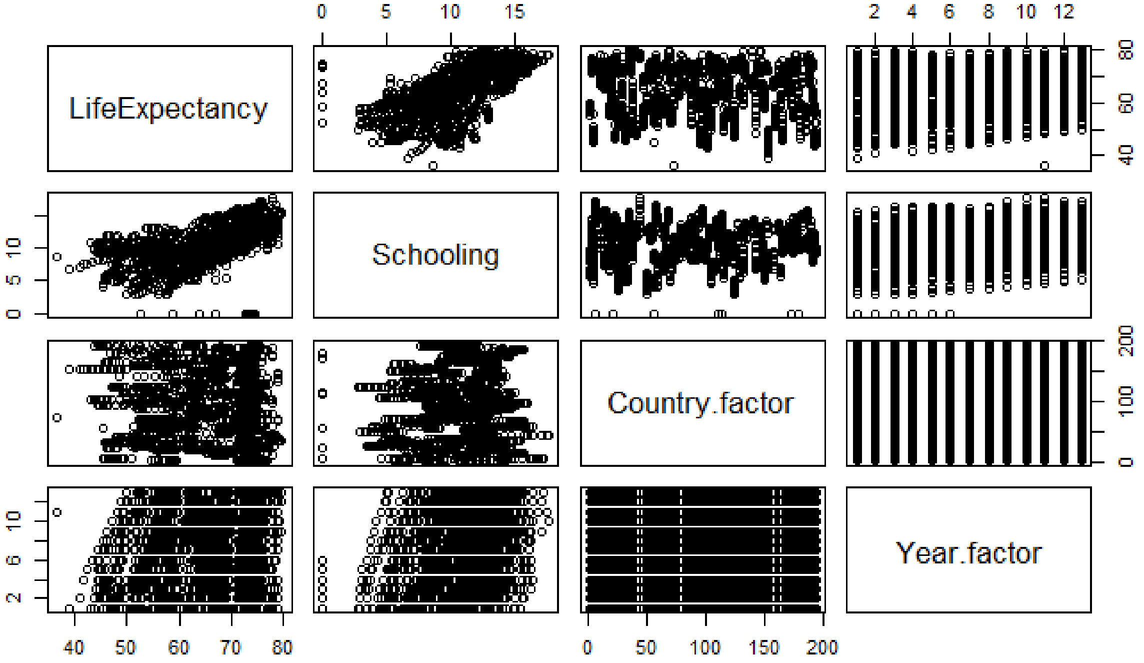


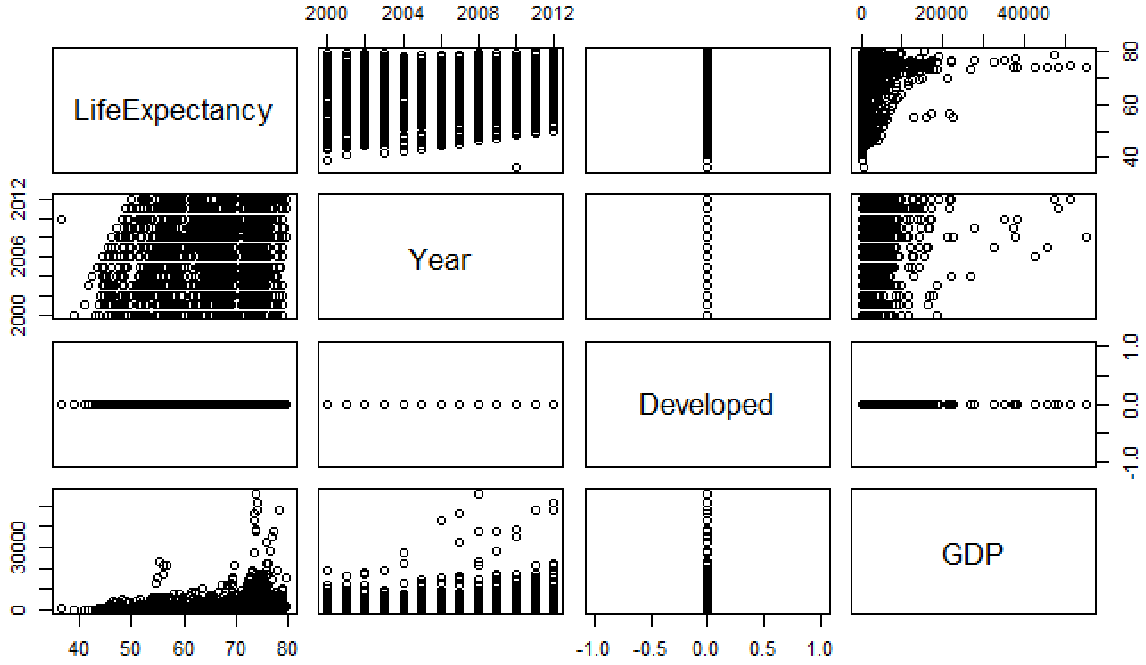



LifeExpectancy
BMI
UnderFiveDeaths
Polio
0 20 40 60
0 20 40 60 80 100
40 60 80
60 20 0
2500 1000 0
80 40 0
40 50 60 70 80
0 500 1500 2500

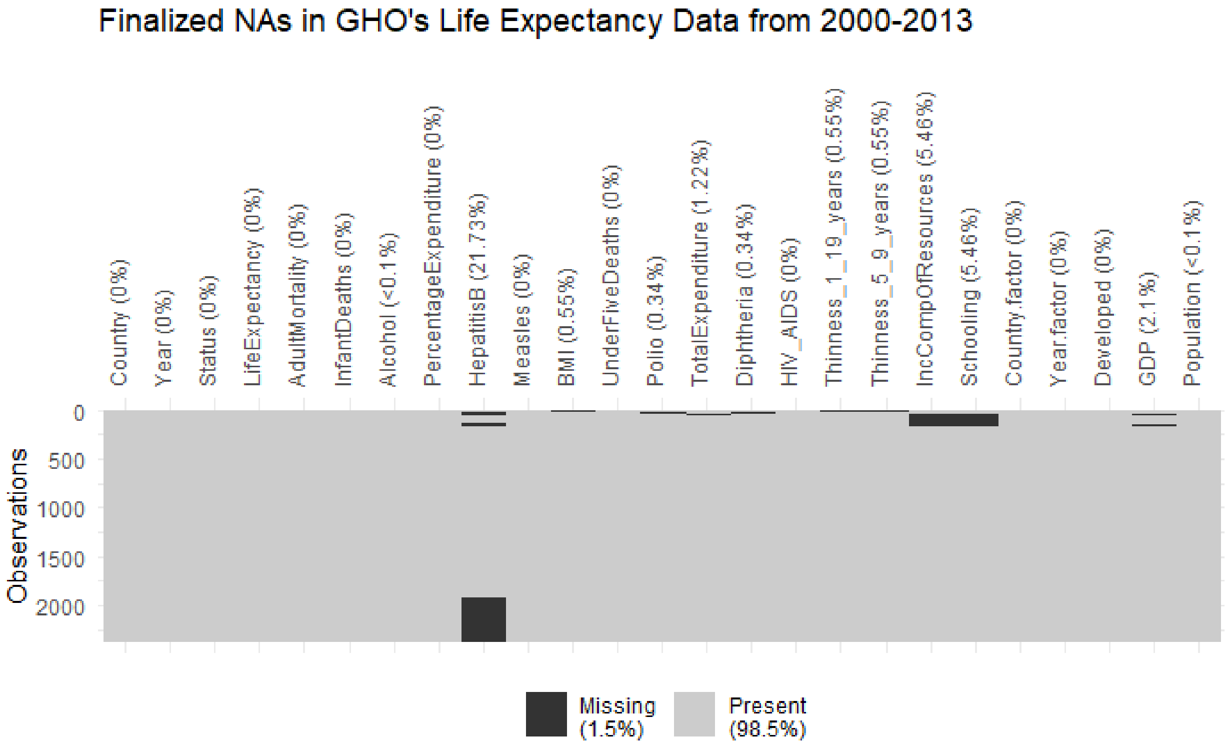

LifeExpectancy
TotalExpenditure
Diphtheria
HIV_AIDS
0 2 4 6 8 10 14
0 10 20 30 40 50
80 60 40
12 8 4 0
80 40 0
40 20 0
40 50 60 70 80
0 20 40 60 80 100


LifeExpectancy
Thinness_1_19_years
Thinness_5_9_years
IncCompOfResources
0 5 10 15 20 25
0.0 0.2 0.4 0.6 0.8
40 60 80
0 10 20
0.0 0.4 0.8
40 50 60 70 80
0 5 10 15 20 25


LifeExpectancy
Schooling
Country.factor
Year.factor
0 5 10 15
2 4 6 8 10 12
40 60 80
0 5 10
0 100 200
2 6 10
40 50 60 70 80
0 50 100 150 200



D) Finalized Cleaned Dataset for Missing NAs



3) Assumptions of Models

Multi-Linear Regression

1. Relationship between explanatory and response variables is linear.

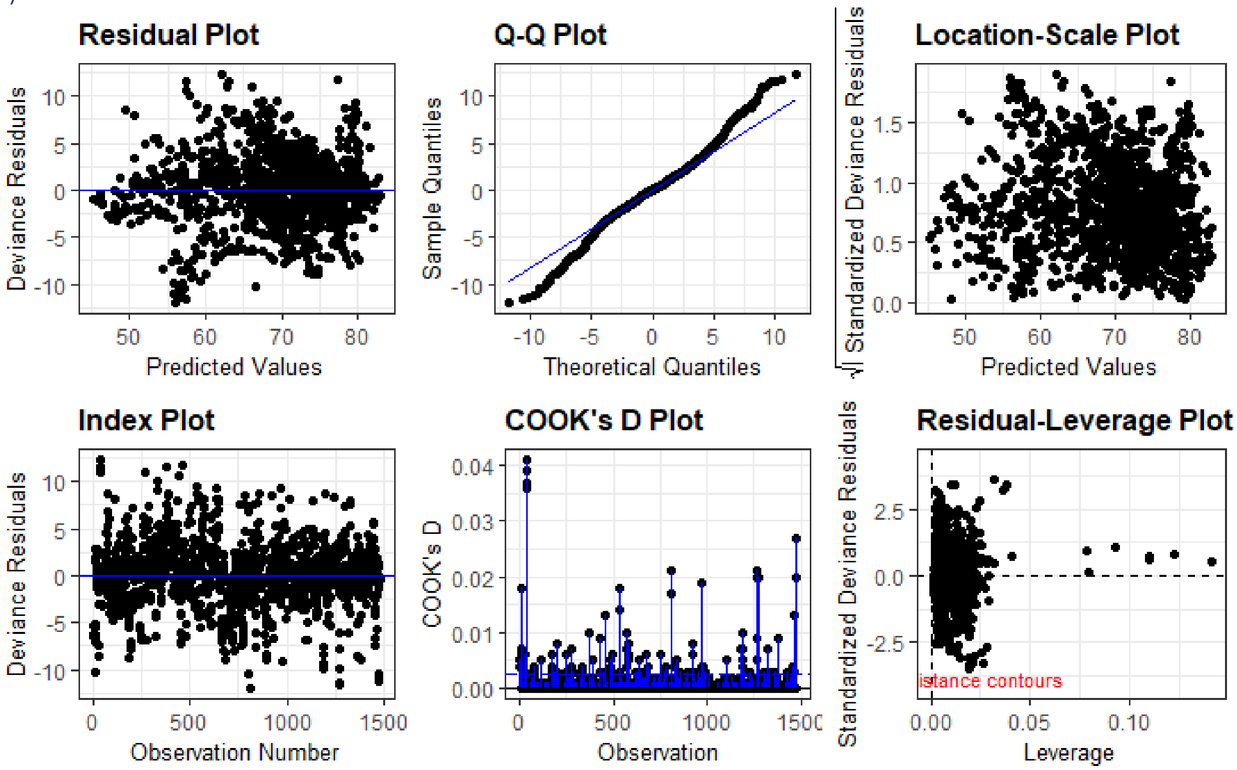
2. Distribution of response variable for a fixed value of explanatory variables is normal.
3. Distribution of residuals is normal.
4. Normal distributions of the response should have constant variance for fixed values of explanatory variables.
5. The normal distributions of residuals should have constant variance for fixed values of explanatory variables.
6. Observations are independent.

4) Step-Wise Model

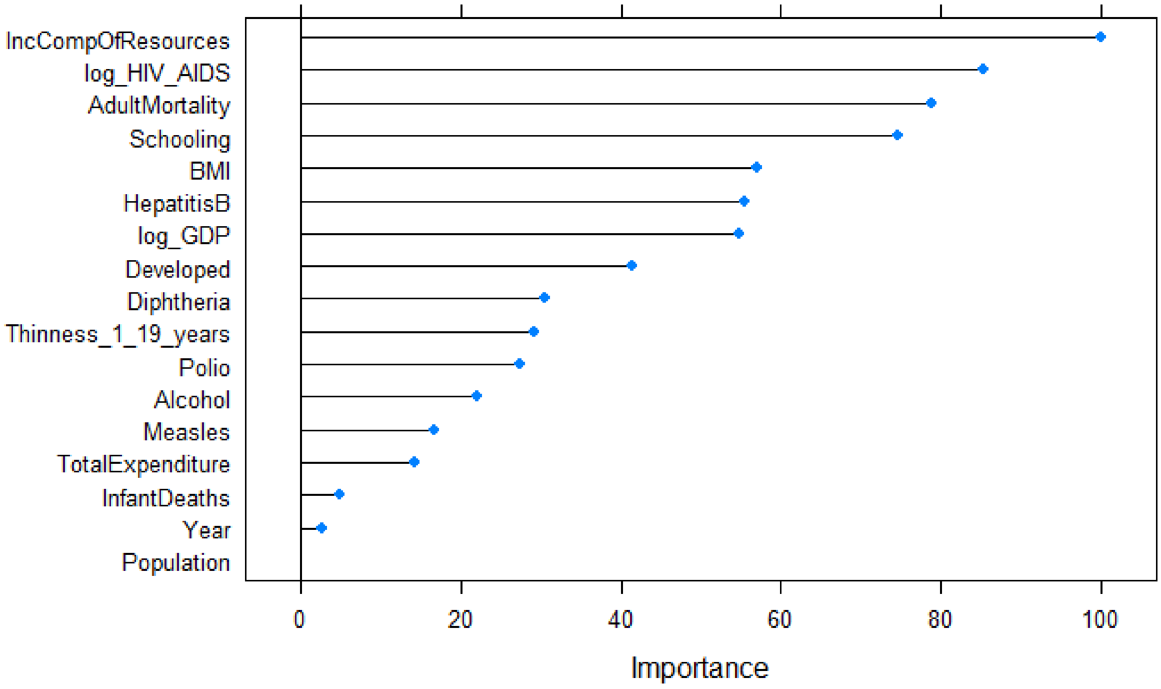
A) Summary Statistics

```
Generalized Linear Model with Stepwise Feature Selection

2024 samples
  17 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 1329, 1330, 1327, 1328, 1328, 1329, ...
Resampling results:

  RMSE    Rsquared   MAE
  3.4688  0.8279992  2.581109


Call:  NULL

Coefficients:
        (Intercept)         log_HIV_AIDS            Schooling       AdultMortality
IncCompOfResources
          54.504992            -2.232302             0.488641            -0.016498
6.075941
            log_GDP            Developed         InfantDeaths           Diphtheria
Thinness_1_19_years
           0.442267             1.555180            -0.002748             0.017352
-0.073413
   TotalExpenditure
           0.072360

Degrees of Freedom: 1475 Total (i.e. Null);  1465 Residual
Null Deviance:            103300
Residual Deviance: 17410        AIC: 7855
loess r-squared variable importance

```

B) Residuals



C) Predictors of Importance



Predictor	Importance
IncCompOfResources	100.000000
log_HIV_AIDS	90.777657
AdultMortality	82.597901
Schooling	73.549850
HepatitisB	58.968162
BMI	57.573693
log_GDP	53.008802
Developed	37.301209
Diphtheria	31.411300
Polio	28.116982
Thinness_1_19_years	27.996204
Alcohol	20.054762
Measles	16.403955
TotalExpenditure	12.630369
InfantDeaths	4.838753
Year	2.938862
Population	0.000000

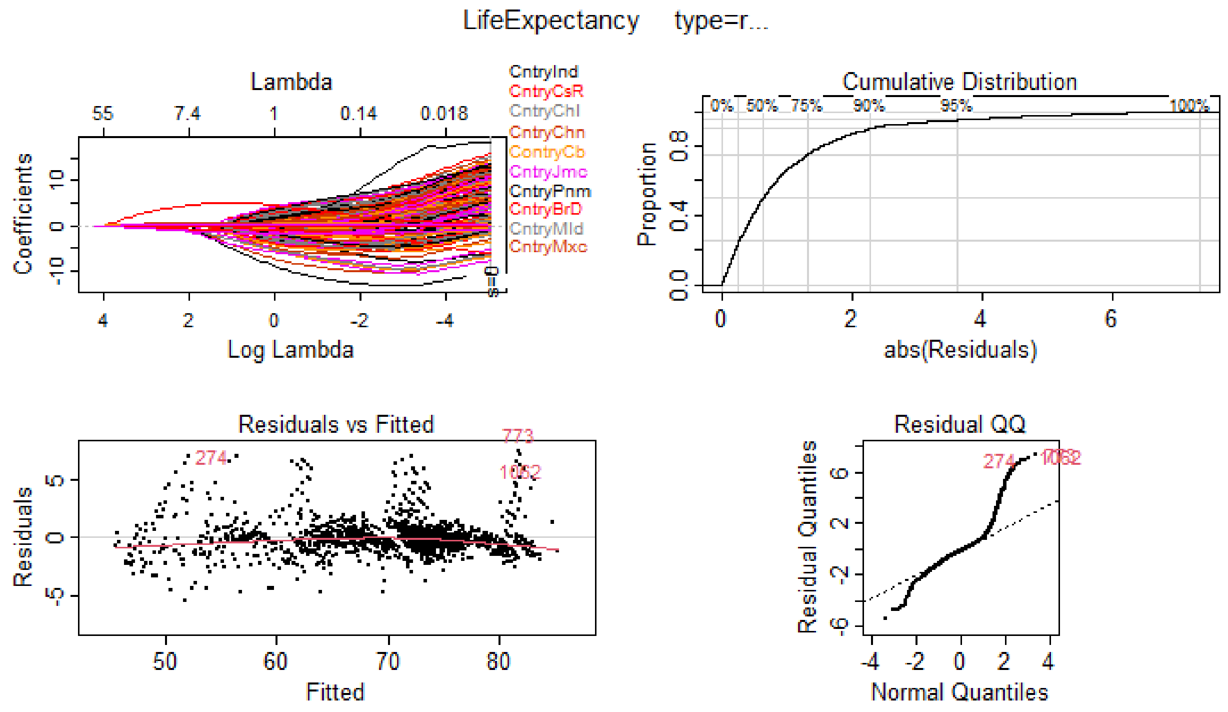
6) Elastic-Net/GLMNET Model

A) With Country

i) Summary Statistics

```
glmnet

2024 samples
  29 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 1327, 1329, 1329, 1329, 1329, 1328, ...
Resampling results across tuning parameters:

  alpha  lambda      RMSE      Rsquared   MAE
  0.10   0.01324987  1.825657  0.9526130  1.186341
  0.10   0.13249874  1.970056  0.9447715  1.312798
  0.10   1.32498744  2.680475  0.9041246  2.007959
  0.55   0.01324987  1.902339  0.9484653  1.249919
  0.55   0.13249874  2.263105  0.9278776  1.597811
  0.55   1.32498744  3.664193  0.8259611  2.741576
  1.00   0.01324987  1.958708  0.9453267  1.303467
  1.00   0.13249874  2.554527  0.9087463  1.889315
  1.00   1.32498744  3.949436  0.8158161  3.000260

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0.1 and lambda = 0.01324987.

```

ii) Residual Plots



iii) Prediction's RSME

11.93222

iv) Predictor Ranking

Predictors	Overall
CountryIndia	100.00000
CountrySierra Leone	91.62896
CountryCosta Rica	71.05132
CountryChina	70.59412
CountryAngola	66.07597
CountryChile	64.51760
CountryCuba	60.93097
CountryPanamá	58.55487
CountryViet Nam	58.23433
CountryMaldives	58.23073
CountryBosnia and Herzegovina	58.19121
CountryAntigua and Barbuda	57.77920
CountryBrunei Darussalam	56.04374
CountryMexico	55.90785
CountrySyrian Arab Republic	54.78828
CountryJamaica	54.65850
CountryLebanon	53.56119
CountrySamoa	52.64764

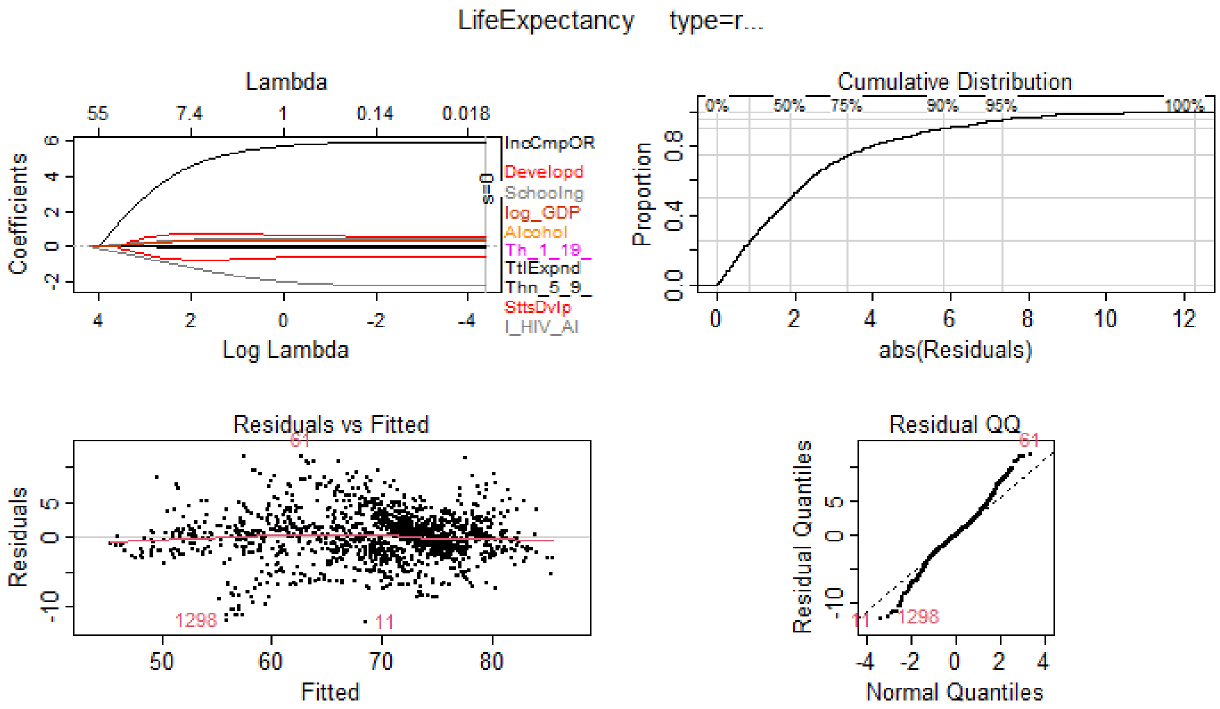
CountryEcuador	52.22268
CountryBangladesh	51.71391

B) Without Country

i) Summary Statistics

```
glmnet

2024 samples
  29 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 1329, 1330, 1327, 1328, 1328, 1329, ...
Resampling results across tuning parameters:

  alpha  lambda      RMSE      Rsquared   MAE
  0.10   0.01324987  3.435034  0.8313779  2.558124
  0.10   0.13249874  3.439274  0.8309549  2.565673
  0.10   1.32498744  3.481892  0.8302658  2.604403
  0.55   0.01324987  3.441956  0.8306787  2.566089
  0.55   0.13249874  3.436744  0.8312934  2.562925
  0.55   1.32498744  3.660162  0.8259800  2.741490
  1.00   0.01324987  3.441309  0.8307329  2.567860
  1.00   0.13249874  3.440673  0.8311172  2.563571
  1.00   1.32498744  3.945265  0.8160767  2.998612

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0.1 and lambda = 0.01324987.

```

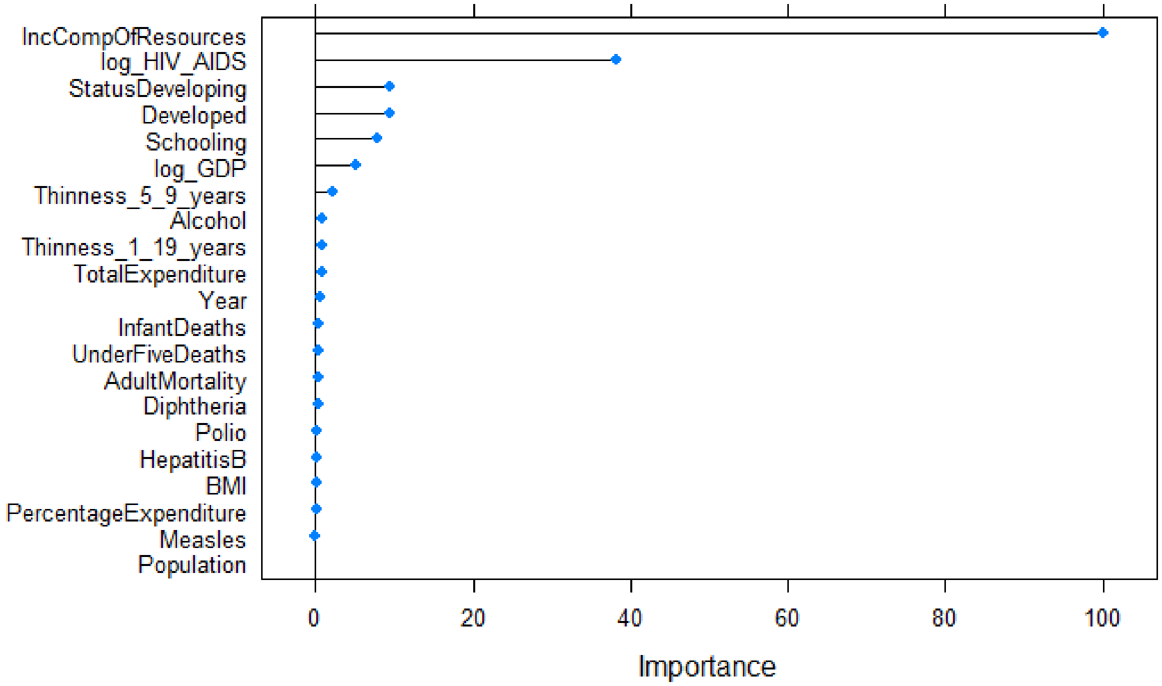
ii) Residual Plots



iii) Prediction's RSME

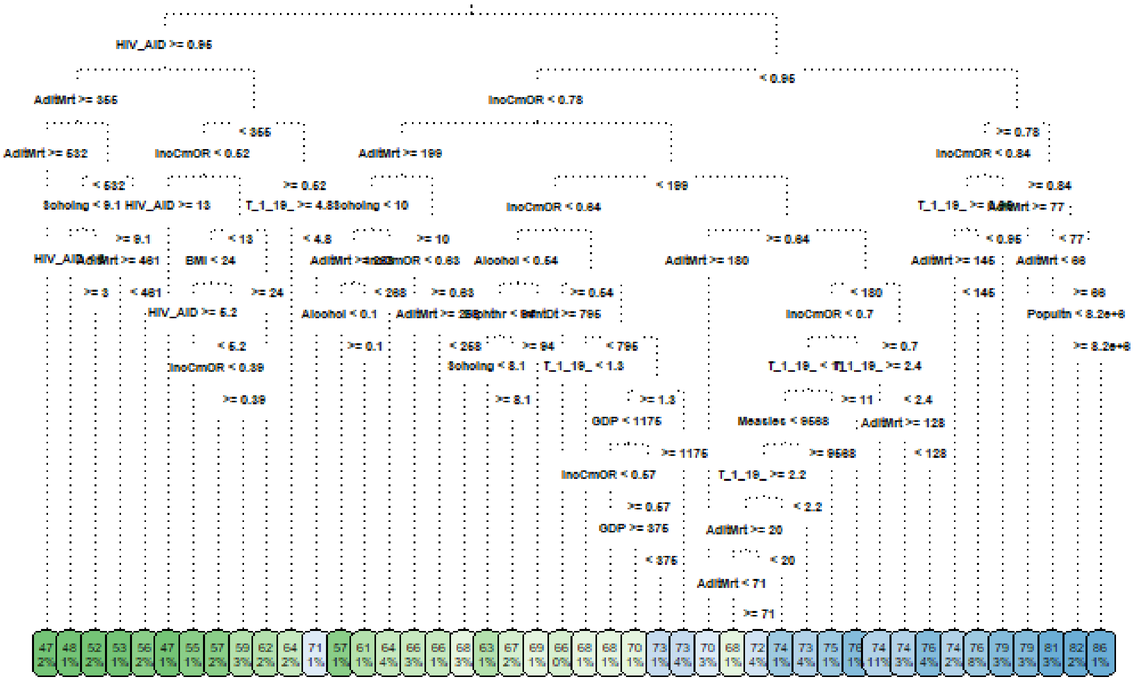
11.67704

iv) Predictor Ranking

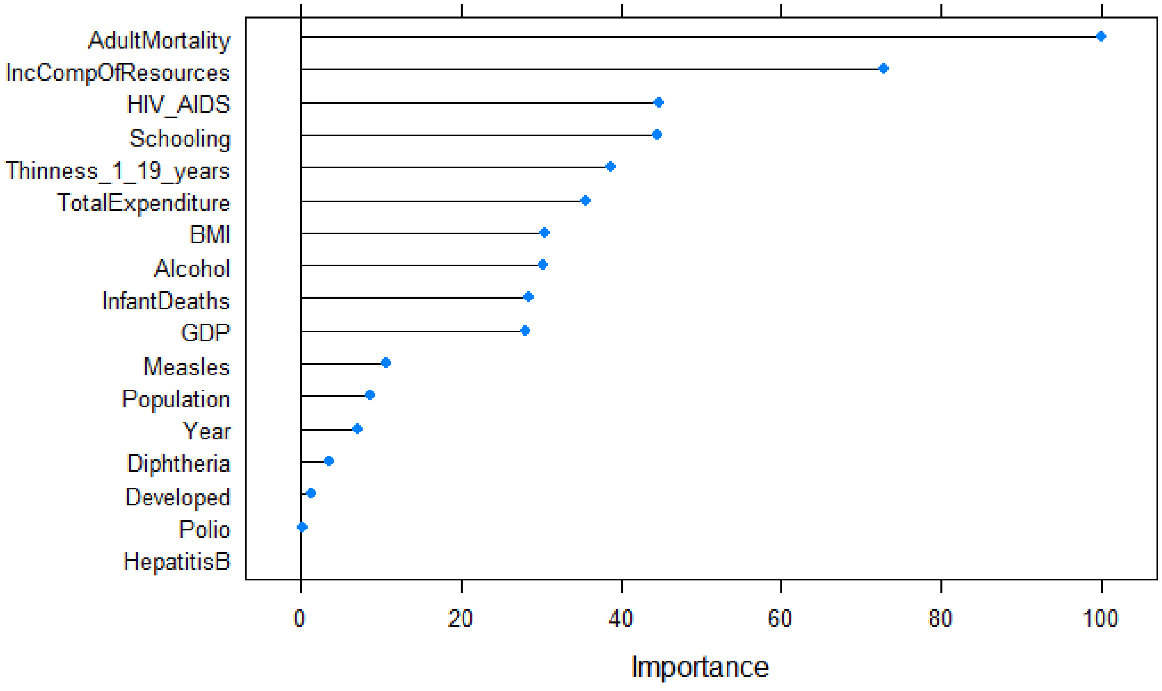


7) Regression Tree Model

A) RPart Plot



B) Predictor Ranking

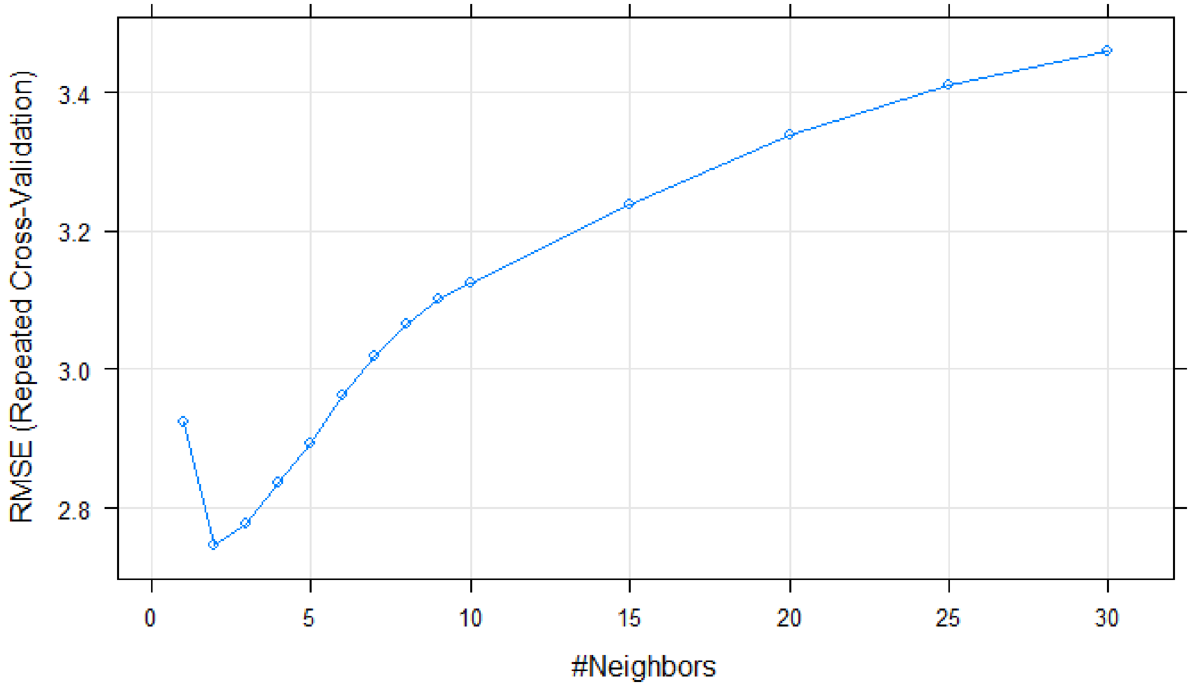


9) Knn Model

A) Summary Statistics

```
k-Nearest Neighbors

2024 samples
  17 predictor

Pre-processing: centered (17), scaled (17)
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 1329, 1330, 1327, 1328, 1328, 1329, ...
Resampling results across tuning parameters:

  k   RMSE      Rsquared   MAE
   1  2.923140  0.8801498  1.708424
   2  2.747412  0.8919239  1.728554
   3  2.776255  0.8894473  1.813962
   4  2.836714  0.8849493  1.921303
   5  2.892727  0.8806092  2.006667
   6  2.962947  0.8747369  2.084330
   7  3.019031  0.8703141  2.147218
   8  3.066195  0.8666054  2.204201
   9  3.102457  0.8638602  2.247936
  10  3.124294  0.8623254  2.275139
  15  3.239134  0.8536511  2.406599
  20  3.337851  0.8460022  2.515153
  25  3.410219  0.8404458  2.583871
  30  3.458544  0.8371208  2.621366
RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 2.
```

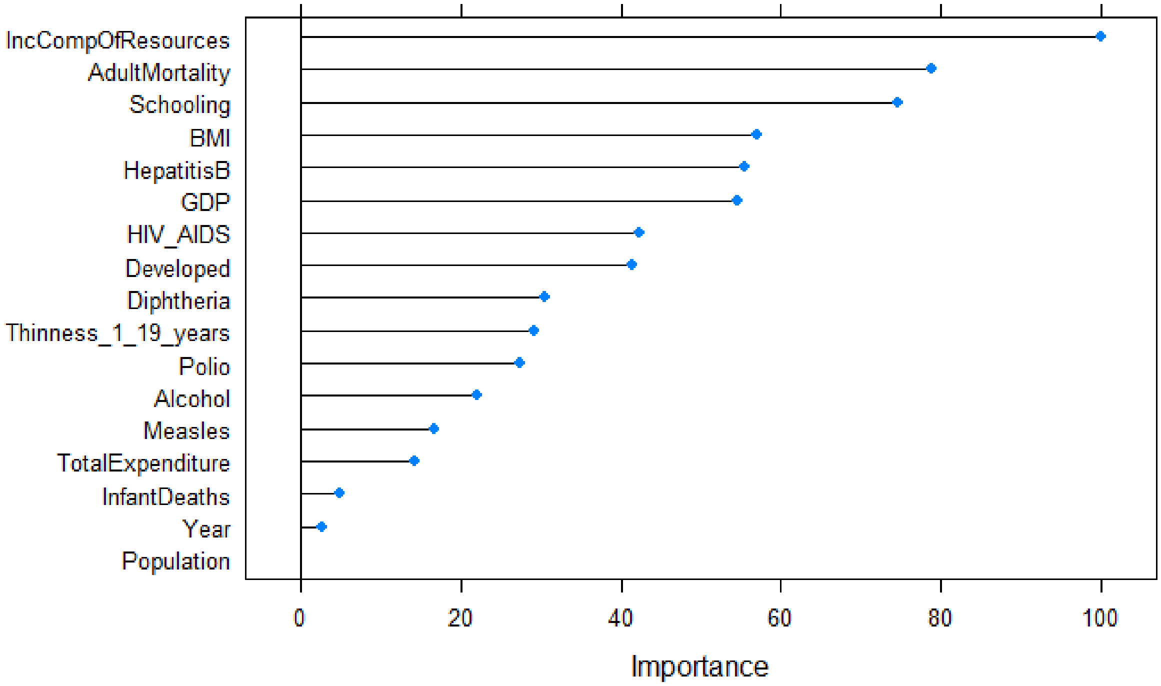
B) K Fit



C) Validation Fit

RMSE	Rsquared	MAE
11.584792	NA	9.110282

D) Predictor Ranking



Predictor	Overall Importance
IncCompOfResources	100.000000
AdultMortality	78.804190
Schooling	74.667472
BMI	56.985648
HepatitisB	55.533967
GDP	54.596349
HIV_AIDS	42.340491
Developed	41.410502
Diphtheria	30.426621
Thinness_1_19_years	29.261760
Polio	27.369131
Alcohol	21.946861
Measles	16.703981
TotalExpenditure	14.318948
InfantDeaths	4.936347
Year	2.768503
Population	0.000000