Life Expectancy Prediction Analysis

Group 85

MA317: Modelling Experimental Data

14/12/2022

Submitted by:

Dhanya Pezhumkattil Jayaprakash

Table of contents

Abstract:

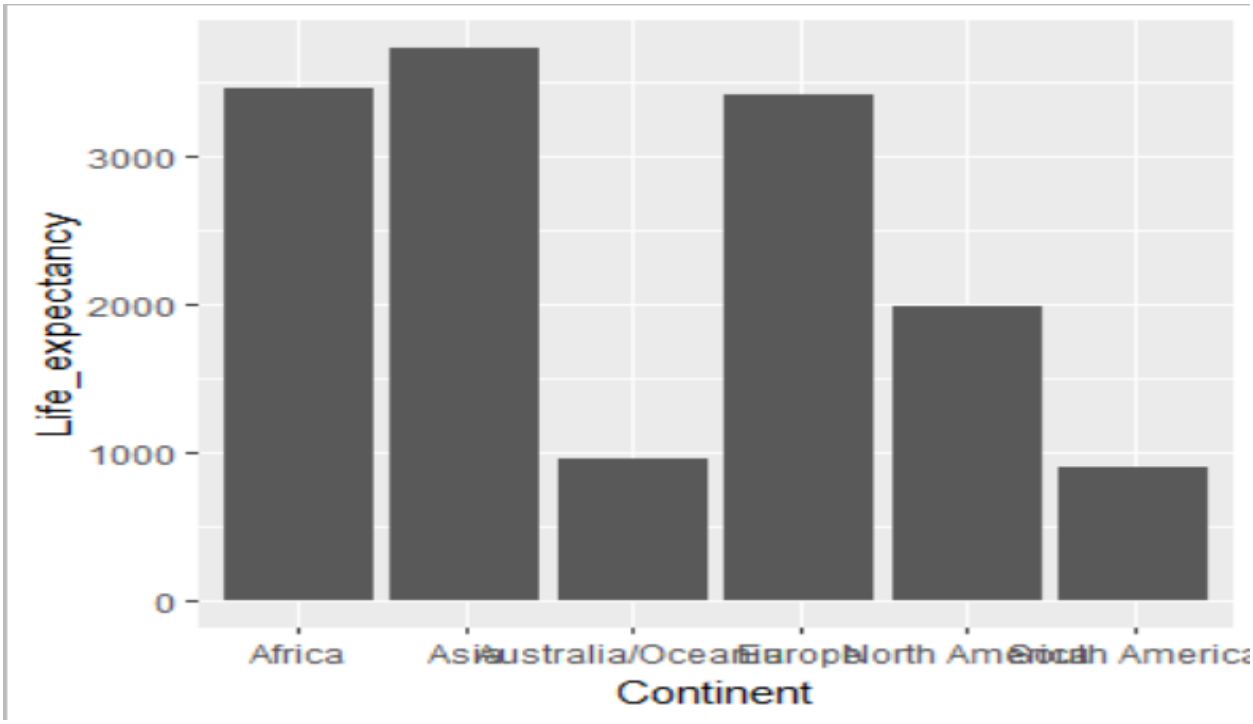
The project has a brief Analysis of the World development indicators (WDI) dataset to predict the Life –Expectancy gave some predictor variables in 2020. The project aims to propose a model which explains life expectancy in the world for 2020 and factors that contribute to predicting life expectancy. Moreover, the project study if there is any significant difference in the average life expectancy across the continents.

Introduction

Life expectancy is the average period of time a particular person is expected to live given some variables based on some demographical factors. It has a crucial role to play in the overall development of the nation. The goal of this coursework is to propose the best model to predict life expectancy in the year 2020.It has given a dataset of the World Development Indicators (WDI), which are derived from a primary World Bank database. This project aims to give the best-fitted linear regression model for forecasting life expectancy for countries.

1. Preliminary Analysis of the life Expectancy dataset

The analysis uses descriptive statistics which is both (graphical and numerical representation) and R language was used for the Life Expectancy data1.cvs dataset. It contains 217 (rows)observations about the features of countries and 29 indicators(columns) such as Access_to_electricity , Adjusted_net_national_income, and so on. The predicted variable(Life Expectancy)has a mean of 72.93 years and the maximum and minimum expected life span is 85.08,53.26 respectively. In this dataset, a certain percentage of data is missing. The column-wise details of missing values are given in the appendix [1].



Graph 1.1: Life_expectancy versus Continent

Using one of the statistical Analyses, the graph above displays the graphical representation of our dataset Life_expectancy when plotted again with one of the predictor variables “Continent”.From this graph, it is clear that South America has the lowest life expectancy compared with Asia has the highest life expectancy rate.

1.1 Dealing with Missing Values

It is very necessary to build a linear model, in which there are no missing values in the dataset. The deletion approach of dealing with the missing value is not the best method however, deleting the predictor variables in some cases is key as using the raw data with missing value as given will result in biased prediction and error in our results. To deal with missing values, the multiple imputation approach is used in this dataset.

This method consists of three stages, in the first stage, all the missing entries are replaced by values from the observed distribution. In this case, there are 5 imputed datasets. In the second stage, each of the imputed datasets was analyzed to find any issues with the imputed data entries. Finally, pool all estimates of the coefficients and of the standard error obtained in stage 2 to derive a single estimate.

In our own case study dataset (Life Expectancy), there are a lot of missing values within in the dataset. For instance, the percentage of the literacy rate (total Adult of people age 15 and above) has a missing value of 192 which represent about 88.5% of the data. Also, some other indicator variable in the given data set for example Poverty Headcount ratio of the predictor variable in the given dataset is also missing 195 values which amount to 89.9% data. While working with the data sets, it was discovered that the missing value for some of the predictor variables are large.

To forecast the best model, removed the variables with more than 80% of missing data. The following table shows the features with a higher ratio of missing data.

Variables	Number of Missing Values
Edu_att_primary	181
Edu_att_Bachelors	179

Literacy_rate	192
Renew-en_consum	217
Poverty_headcount_ratio	195

Table 1: Variables with missing values >80%

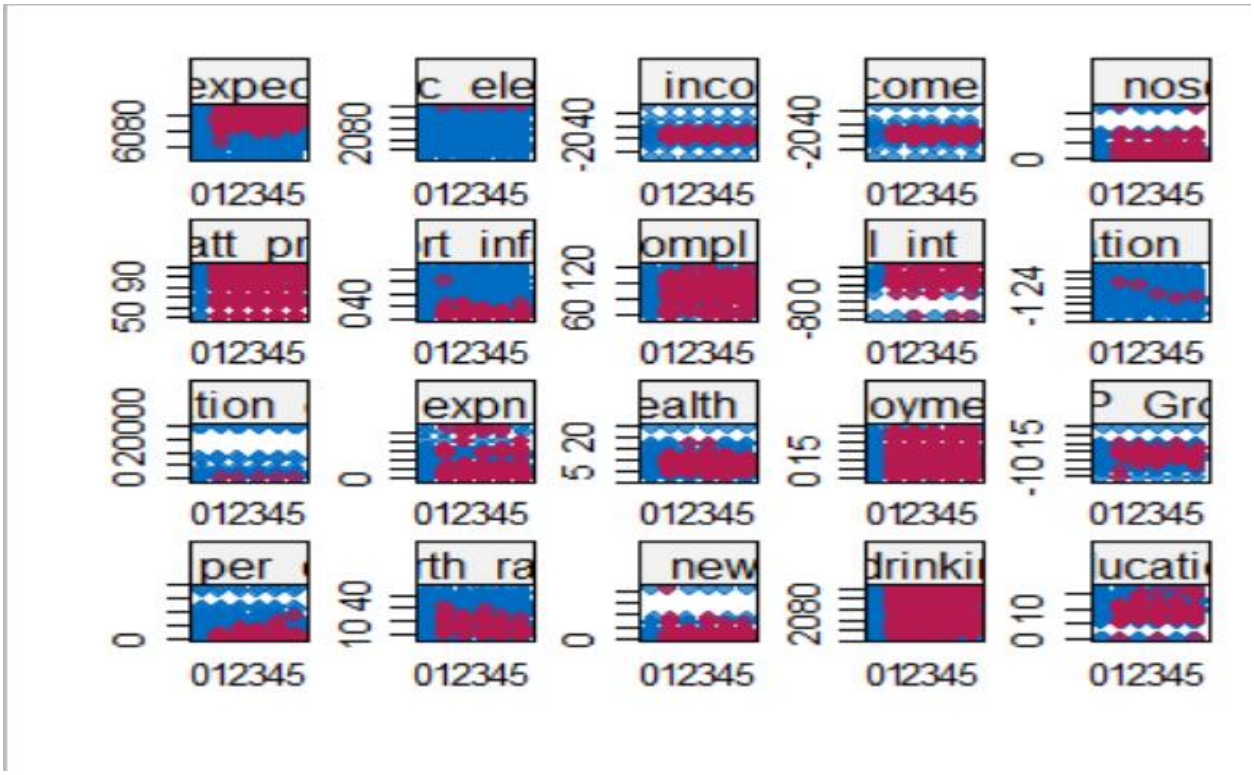


Figure1:Imputed Data Plotted against the rest of the observation

2. Investigating collinearity between the predictor variables

In statistics, collinearity is the correlation between predictor (independent) variables. Collinearity generates a high variance of the estimated coefficients and hence, the coefficient estimates corresponding to those interrelated explanatory variables will not accurately predict the value of the dependent variable. They can become very sensitive to small changes in the model.

To estimate the collinearity between predictor variables, two methods were used.

2.1 The VIF (variance inflation factor):

It measures the multicollinearity among the predictor variables. Based on VIF values remove some of the highly correlated independent variables. Small VIF values(VIF < 3)indicate a low correlation among variables under ideal conditions. The default VIF cutoff value is 5.

2.2 A correlation matrix:

It is a table that displays the correlation coefficients for different predictor variables. Moreover, it is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

In this section, the analysis is carried out on all of the 5 imputed datasets and the following table shows VIF values of variables on 5th iteration of the dataset.

Variables	VIF
Access_to_electricity	5.532769
Adjusted_net_national_income	2182.812164
Adjusted_net_national_income_per_capita	2143.090995
Children_out_of_school_primary	2.102006
Educational_attainment_primary	3.03906
Mortality_rate_infant	7.82368
Primary_completion_rate	3.26538
Real_interest_rate	1.506128
Population_growth	89.699511
Population_density	3.111692
Current_health_expenditure_per_capita	8.840231
Current_health_expenditure	3.373809
Unemployment_total	2.603943
GDP_Growth	2.099569
GDP_per_capita	8.620253
Birth_rate	12.380061
Adults_15_to_49_newly_infected_HIV	1.495663

People_using_safely_managed_drinking_water_services	4.883609
Compulsory_education_duration	1.728147

Table 2.1: The VIF of eah variable for the 5th imputed dataset

- From the table, it is clear that the variables "Adjusted_net_national_income and Adjusted_net_national_income_per_capita" has very high VIF values, even though these variables are removed from the dataset.
- Similarly,the variables" Population_growth, Access_to_electricity , Mortality_rate_infant, Current_health_expenditure_per_capita, GDP_per_capita, Birth_rate " also have VIF value grater than 5.Hence,these variables from the table is dropped.

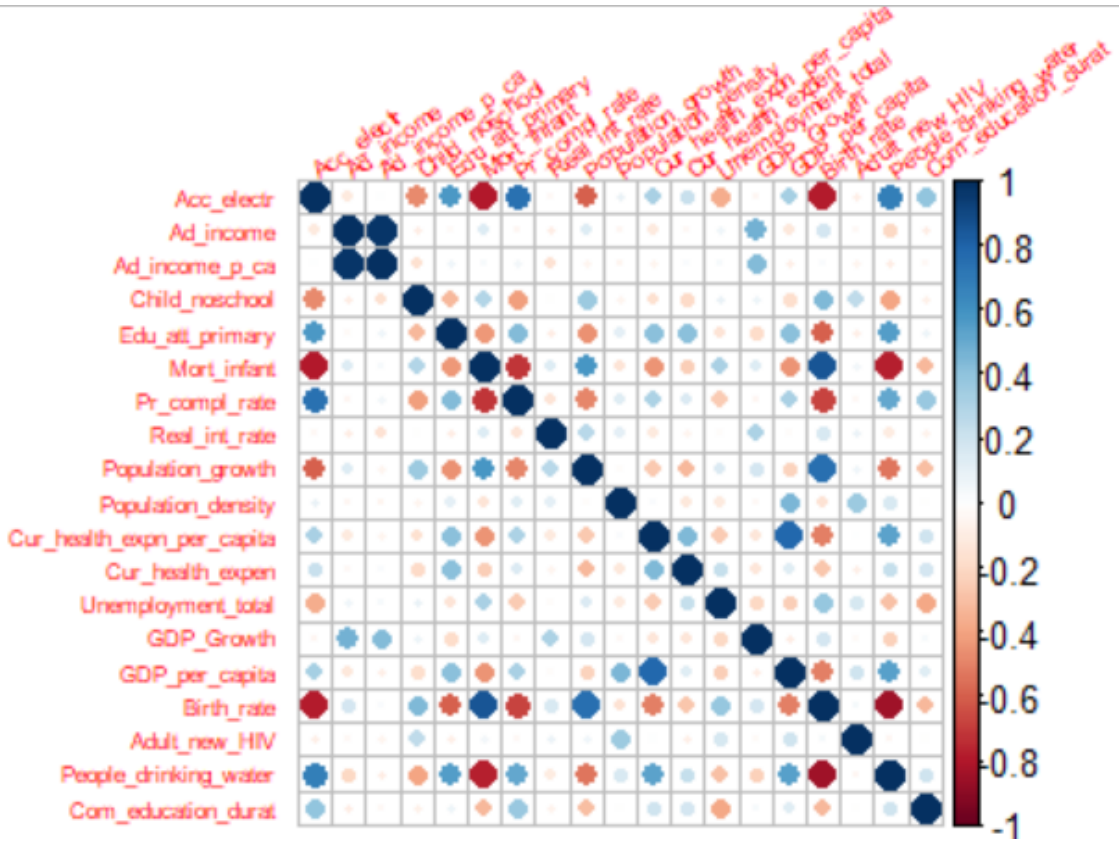
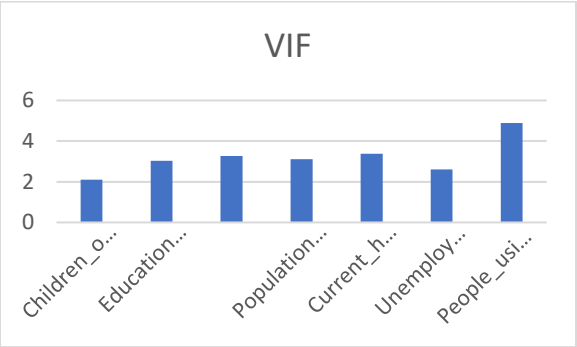


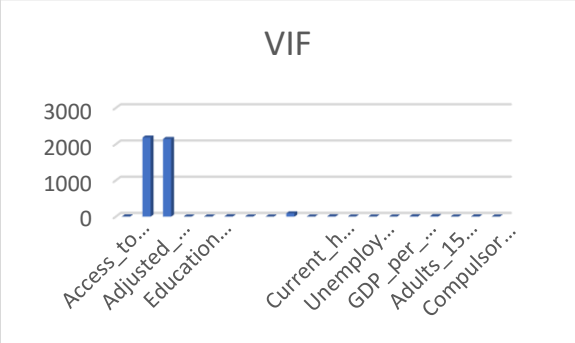
Figure 2.1: Visualization of the Correlation Matrix for the 5th imputed Dataset.

From Figure 2.1, we can see that some features are highly correlated with other independent variables and only eleven features can pass the below 5 thresholds(Appendix1). There are some large correlations so there is evidence for collinearity between some predictors. When taking the multiple linear regression model of these predictor variables, discovered 4 variables specifically, "Real_int_rate, GDP_Growth, Adult_new_HIV, Com_education_durat" have p-value >0.05.

To reduce collinearity, based on the VIF table and Correlation Matrix, the variables with VIF value > 5 were removed from the dataset. Moreover, based on the multiple linear regression analysis the variables with P-values>0.05 were also dropped from the imputed dataset. The tables below compare the VIF values of predictors before and after the removal of variables from the dataset.



Graph 2.1:VIF of each variable after column removal



Graph 2.1:VIF of each variable before column removal

As a result, the multiple linear regression model at this stage is given by,

Life_expectancy= (4.842e+01) + 1.731e-06 Child_noschool - 6.329e-02 Edu_att_primary + 1.925e-01 Pr_compl_rate + 3.149e-04 Population_density + 4.874e-01 Cur_health_expen - 1.623e-01 Unemployment_total + 1.551e-01 People_drinking_water

3. Finding the Best Model

To analyze and forecast better life expectancy in 2020 and to find features that are important for predicting life expectancy, an iterative approach is used. It was deployed on each of the imputed datasets.

3.1 Feature Selection

Stepwise regression is a procedure we can use to build a regression model from a set of predictor variables by entering and removing predictors in a stepwise manner into the model until there is no statistically valid reason to enter or remove any more. Both forward and backward stepwise regression procedure is applied to find the relevant features that contribute to the prediction of life expectancy.

3.1.1Forward Stepwise Regression

- In this method, First, fit the intercept-only model. This model had an AIC of **868.09**.
- Next, fit every possible one-predictor model. The model that produced the lowest AIC and also had a statistically significant reduction in AIC compared to the intercept-only model used the predictor "People_drinking_water". This model had an AIC of **626.58**.
- It turned out that none of these models produced a significant reduction in AIC, and at this point stopped the procedure.
- The final model had an AIC of **525.39 (Appendix2)** and turns out to be:

Life_expectancy = (4.842e+01)+ 1.551e-01 People_drinking_water + 1.925e-01 Pr_compl_rate - 1.623e-01 Unemployment_total + 4.874e-01 Cur_health_expen - 6.329e-02 Edu_att_primary + 3.149e-04 Population_density + 1.731e-06 Child_noschool

3.1.2 Backward Stepwise Regression

- The backward feature selection starts with all variables in the model and then removes the least relevant features one at a time. Initially, the model had an AIC of **525.39 (Appendix3)**.

- In this case, no predictor variables have no significant values as the figure remains the same as the initial AIC of **525.39.** Hence, the predicted model at the current stage is ,

  Life_expectancy =(4.842e+01)+ 1.731e-06 Child_noschool -6.329e-02 Edu_att_primary + 1.925e-01 Pr_compl_rate + 3.149e-04 Population_density +4.874e-01 Cur_health_expen -1.623e-01 Unemployment_total + 1.551e-01 People_drinking_water

3.2. Evaluating the models

To investigate further to find the best model, analyze the forward stepwise regression and backward stepwise regression models separately.

When analyzing the multiple linear regression model of forward feature selection, the p-value of the predictor variable "Child_noschool" is >0.05. So, this variable has no significance in predicting life expectancy. The fitted regression model at this stage is given by(Appendix4),

Life_expectancy = 49.7828579+ 0.1521661 People_drinking_water + 0.1840493 Pr_compl_rate - 0.1639907 Unemployment_total + 0.4775241 Cur_health_expen -0.0650008 Edu_att_primary + 0.0003182 Population_density

Similarly, The multiple linear regression model of feature got in backward feature selection is given by (Appendix5),

Life_expectancy =49.7828579 -0.0650008 Edu_att_primary + 0.1840493 Pr_compl_rate +0.0003182 Population_density + 0.4775241 Cur_health_expen -0.1639907 Unemployment_total + 0.1521661People_drinking_water

3.3Predicting Life Expectancy

From the final analysis, the AIC of both models (forward and backward) is **1144.047**. Since the value of both models is the same, any of these models can be selected as the best model. The most significant features that have an influence on the prediction of life expectancy are

“Edu_att_primary, Pr_compl_rate, Population_density, Cur_health_expen, Unemployment_total, People_drinking_water”.

Conclusion

The analysis concludes that the nation’s development in terms of Life expectancy is primarily focused on education, health expenditure, unemployment, and pure water consumption. Population density is also another feature, it has a less significant role compared to other features.

References

- An Introduction to Statistical Learning: With Applications in R by Gareth James,Trevor Hastie,Robert Tibshirani,Daniela
- MA317 Lab1, MA317 Lab2, MA317 Lab3 ,MA317 Lab4, MA317 Lab5

APPENDIX

1.R code to read the data

```
Dataset1<-read.csv("Life_Expectancy_Data1.csv",header=T)
```

2.#Renaming the variables

```
names(Dataset1)

Z_Last<-Dataset1

names(Z_Last)<-c<- c("Country", "Country Code", "Continent","Life_expectancy",
"Acc_electr", "Ad_income","Ad_income_p_ca", "Child-with_HIV", "Child_noschool",
"Edu_att_primary", "Edu_att_Bachelors", "Mort_infant", "Pr_compl_rate", "Literacy_rate",
"Real_int_rate", "Population_growth", "Population_density", "Population_total",
"Cur_health_expn_per_capita", "Cur_health_expen", "Unemployment_total", "GDP_Growth",
"GDP_per_capita", "Birth_rate", "Renew-en_consum", "Adult_new_HIV",
"People_drinking_water"," Poverty_headcount_ratio", "Com_education_durat")
```

Indicator Name	**Code**
Life expectancy at birth, total (years)	Life_expectancy
Access to electricity (\% of population)	Acc_electr
Adjusted net national income (annual \% growth)	Ad_income
Adjusted net national income per capita (annual \% growth)	Ad_income_p_ca
Children (ages 0-14) newly infected with	Child-with_HIV
Children out of school, primary	Child_noschool
Educational attainment, at least completed primary, population 25+ years, total (\%) (cumulative)	Edu_att_primary
Educational attainment, at least Bachelor's or equivalent, population 25+, total (\%) (cumulative)	Edu_att_Bachelors
Mortality rate, infant (per 1,000 live births	Mort_infant

Literacy rate, adult total (\% of people ages 15 and above) Real interest rate (\%)Population growth (annual \%)	Literacy_rate
Population density (people per sq. km of land area)	Population_density
Population, total	Population_total
Current health expenditure per capita, PPP (current international \$)	Cur_health_expn_per_capita
Unemployment, total (\% of total labor force) (national estimate)	Unemployment_total
GDP per capita, PPP (current international \$)	GDP_per_capita
Birth rate, crude (per 1,000 people)	Birth_rate
Renewable energy consumption (\% of total final energy consumption)	Renew-en_consum
People using safely managed drinking water services (\% of population)	People_drinking_water
Compulsory education, duration (year)	Com_education_durat
Primary completion rate, total (\% of relevant age group)	Pr_compl_rate
Current health expenditure (\% of GDP)	Cur_health_expen
GDP growth (annual \%)	GDP_Growth
Adults (ages 15-49) newly infected with	Adult_new_HIV
Poverty headcount ratio at \$3.20 a day (2011 PPP) (\% of population)	Poverty_headcount_ratio

Appendix[1]

summary(Z_Last) # display the number of NA values in each of the columns displayed

```
> summary(Z_Last)# display the number of NA values in each of the columns displayed
   Country           Country Code         Continent          Life_expectancy   Acc_electr
 Length:217         Length:217         Length:217         Min.   :53.28   Min.   :  6.721
 Class :character   Class :character   Class :character   1st Qu.:67.89   1st Qu.: 84.762
 Mode  :character   Mode  :character   Mode  :character   Median :74.23   Median :100.000
                                                          Mean   :72.93   Mean   : 86.470
                                                          3rd Qu.:78.48   3rd Qu.:100.000
                                                          Max.   :85.08   Max.   :100.000
                                                          NA's   :19      NA's   :1
   Ad_income        Ad_income_p_ca      Child-with_HIV  Child_noschool     Edu_att_primary
 Min.   :-30.792   Min.   :-32.5432   Min.   :  100   Min.   :      0   Min.   : 49.55
 1st Qu.:  1.225   1st Qu.:  0.5222   1st Qu.:  100   1st Qu.:   1262   1st Qu.: 81.77
 Median :  3.660   Median :  2.7583   Median :  500   Median :   7359   Median : 93.69
 Mean   :  4.030   Mean   :  2.6585   Mean   : 1650   Mean   :  98650   Mean   : 87.74
 3rd Qu.:  6.242   3rd Qu.:  5.0702   3rd Qu.: 1100   3rd Qu.:  78956   3rd Qu.: 99.24
 Max.   : 50.172   Max.   : 47.2518   Max.   :20000   Max.   :1712650   Max.   :100.00
 NA's   :79        NA's   :79         NA's   :127     NA's   :99        NA's   :181
 Edu_att_Bachelors  Mort_infant     Pr_compl_rate    Literacy_rate    Real_int_rate
 Min.   : 4.322    Min.   : 1.60   Min.   : 54.73   Min.   : 58.00   Min.   :-78.518
 1st Qu.:11.898    1st Qu.: 5.70   1st Qu.: 85.82   1st Qu.: 89.89   1st Qu.:  3.176
 Median :19.665    Median :14.30   Median : 97.40   Median : 95.74   Median :  6.354
 Mean   :19.864    Mean   :20.97   Mean   : 93.05   Mean   : 92.04   Mean   :  6.220
 3rd Qu.:25.721    3rd Qu.:31.50   3rd Qu.:101.45   3rd Qu.: 97.56   3rd Qu.:  9.214
 Max.   :46.631    Max.   :82.40   Max.   :120.45   Max.   :100.00   Max.   : 39.877
 NA's   :179       NA's   :24      NA's   :89       NA's   :192      NA's   :104
 Population_growth Population_density  Population_total    Cur_health_expn_per_capita
 Min.   :-1.6095   Min.   :    0.137   Min.   :1.076e+04   Min.   :   19.85
 1st Qu.: 0.3882   1st Qu.:   38.177   1st Qu.:7.779e+05   1st Qu.:   85.73
 Median : 1.0946   Median :   92.842   Median :6.661e+06   Median :  392.43
 Mean   : 1.1917   Mean   :  446.043   Mean   :3.545e+07   Mean   : 1143.71
 3rd Qu.: 1.9556   3rd Qu.:  233.011   3rd Qu.:2.544e+07   3rd Qu.: 1160.93
 Max.   : 4.4687   Max.   :19466.444   Max.   :1.408e+09   Max.   :10921.01
```

3. #Plotting Life_expectancy versus Continent in Graph 1.1

```
ggplot(Z_Last, aes(x = Continent, y = Life_expectancy)) +
  geom_bar(stat = "identity")
```

4.Missing Values

```
Z_Last<-Z_Last[,c(-1,-2,-3,-8,-11,-14,-25,-28)]#Removed attribute with high missing values

Z_Last <- Z_Last[-12] # the column population total was removed as it keeps populating erroron the table.
```

```
Removed 19 rows containing missing values (`position_stack()`).
> colSums(is.na(Z_Last))# display on the data set if there is NA values or Not
                    Country                Country Code                   Continent
                          0                           0                           0
            Life_expectancy                  Acc_electr                  Ad_income
                         19                           1                          79
             Ad_income_p_ca             Child-with_HIV             Child_noschool
                         79                         127                          99
            Edu_att_primary          Edu_att_Bachelors                Mort_infant
                        181                         179                          24
              Pr_compl_rate              Literacy_rate              Real_int_rate
                         89                         192                         104
         Population_growth          Population_density           Population_total
                          1                           1                           1
Cur_health_expn_per_capita            Cur_health_expen         Unemployment_total
                         31                          31                          96
                GDP_Growth              GDP_per_capita                 Birth_rate
                         14                          12                          13
            Renew-en_consum                 Adult_new_HIV      People_drinking_water
                        217                          88                          89
   Poverty_headcount_ratio         Com_education_durat
                        195                          19
```

5.Missing value imputation

complete(Z_Last_imp)# these helps to check the imputed values and see if there might have an error at any of the steps while computing the steps

DataC <- complete(Z_Last_imp)# showing computed values of the data set after the imputation function was used.

summary(DataC)

Z_Last_imp$imp # these function helps to analyse and organised the dataset in a logical manner

complete(Z_Last_imp,2) # display the first 2 columns of the data set

stripplot(Z_Last_imp, pch = 20, cex = 1.2) #Draws the Strip plot(One dimentional scatter plots) of the data set

xyplot(Data1_imp, Child_noschool ~ Population_growth | .imp, pch = 20, cex = 1.4)

model.fit <- with (Z_Last_imp, lm(Life_Expectancy_at_birth ~.))

summary(model.fit)

pooled.model.fit<-pool(model.fit)

summary(pooled.model.fit)

```
 DataC <- complete(Z_Last_imp)# showing computed values of the data set after the imputation fur
was used.
 summary(DataC)
Life_expectancy   Acc_electr        Ad_income       Ad_income_p_ca     Child_noschool
Min.   :53.28   Min.   :  6.721   Min.   :-30.792   Min.   :-32.5432   Min.   :      0
1st Qu.:68.19   1st Qu.: 85.000   1st Qu.:  1.365   1st Qu.:  0.6333   1st Qu.:   1948
Median :74.47   Median :100.000   Median :  3.759   Median :  2.9279   Median :   7676
Mean   :73.20   Mean   : 86.533   Mean   :  4.037   Mean   :  2.7793   Mean   :  98939
3rd Qu.:78.57   3rd Qu.:100.000   3rd Qu.:  6.157   3rd Qu.:  5.1267   3rd Qu.:  83083
Max.   :85.08   Max.   :100.000   Max.   : 50.172   Max.   : 47.2518   Max.   :1712650
Edu_att_primary   Mort_infant    Pr_compl_rate    Real_int_rate     Population_growth
Min.   : 49.55   Min.   : 1.60   Min.   : 54.73   Min.   :-78.518   Min.   :-1.6095
1st Qu.: 67.82   1st Qu.: 4.90   1st Qu.: 84.59   1st Qu.:  3.726   1st Qu.: 0.3938
Median : 91.04   Median :12.80   Median : 96.23   Median :  6.354   Median : 1.0979
Mean   : 83.39   Mean   :19.65   Mean   : 91.61   Mean   :  7.122   Mean   : 1.1984
3rd Qu.: 98.53   3rd Qu.:30.10   3rd Qu.:100.98   3rd Qu.: 10.095   3rd Qu.: 1.9611
Max.   :100.00   Max.   :82.40   Max.   :120.45   Max.   : 39.877   Max.   : 4.4687
Population_density  Cur_health_expn_per_capita Cur_health_expen Unemployment_total
Min.   :    0.137   Min.   :   19.85           Min.   : 1.525   Min.   : 0.10
1st Qu.:   38.283   1st Qu.:   92.67           1st Qu.: 4.445   1st Qu.: 3.91
Median :   92.724   Median :  415.20           Median : 6.378   Median : 6.43
Mean   :  444.289   Mean   : 1258.57           Mean   : 6.633   Mean   : 8.99
3rd Qu.:  231.986   3rd Qu.: 1192.82           3rd Qu.: 8.330   3rd Qu.:12.05
Max.   :19466.444   Max.   :10921.01           Max.   :23.962   Max.   :28.47
  GDP_Growth      GDP_per_capita       Birth_rate    Adult_new_HIV    People_drinking_water
Min.   :-11.143   Min.   :   228.2   Min.   : 5.90   Min.   :   100   Min.   :  5.581
1st Qu.:  1.204   1st Qu.:  2276.3   1st Qu.:10.62   1st Qu.:   500   1st Qu.: 36.488
Median :  2.613   Median :  6837.7   Median :17.30   Median :  1400   Median : 73.743
Mean   :  2.796   Mean   : 17995.1   Mean   :19.29   Mean   :  8915   Mean   : 66.670
```

6.VIF of full model

Full_model <- lm(Life_expectancy ~ ., data=DataC)

summary(Full_model)

ols_vif_tol(Full_model)

vif(Full_model)

```
> ols_vif_tol(Full_model)
                    Variables    Tolerance         VIF
1                  Acc_electr 0.1807413379    5.532769
2                   Ad_income 0.0004581246 2182.812164
3               Ad_income_p_ca 0.0004666157 2143.090995
4               Child_noschool 0.4757360396    2.102006
5              Edu_att_primary 0.3290490700    3.039060
6                  Mort_infant 0.1278170883    7.823680
7                Pr_compl_rate 0.3062430330    3.265380
8                Real_int_rate 0.6639540402    1.506128
9            Population_growth 0.0111483328   89.699511
10          Population_density 0.3213685950    3.111692
11 Cur_health_expn_per_capita 0.1131192184    8.840231
12            Cur_health_expen 0.2964008757    3.373809
13         Unemployment_total 0.3840329829    2.603943
14                 GDP_Growth 0.4762883080    2.099569
15              GDP_per_capita 0.1160058718    8.620253
16                  Birth_rate 0.0807750467   12.380061
17               Adult_new_HIV 0.6685999558    1.495663
18      People_drinking_water 0.2047665995    4.883609
19        Com_education_durat 0.5786544397    1.728147
> 
```

7.**Checking for collinearity using R**

install.packages('corrplot')

library(corrplot)

m <- cor(DataC[-1])

m <- cor(DataC[-1])

corrplot(m, tl.pos = 'lt',tl.cex = 0.5,tl.srt=35, method ='circle')

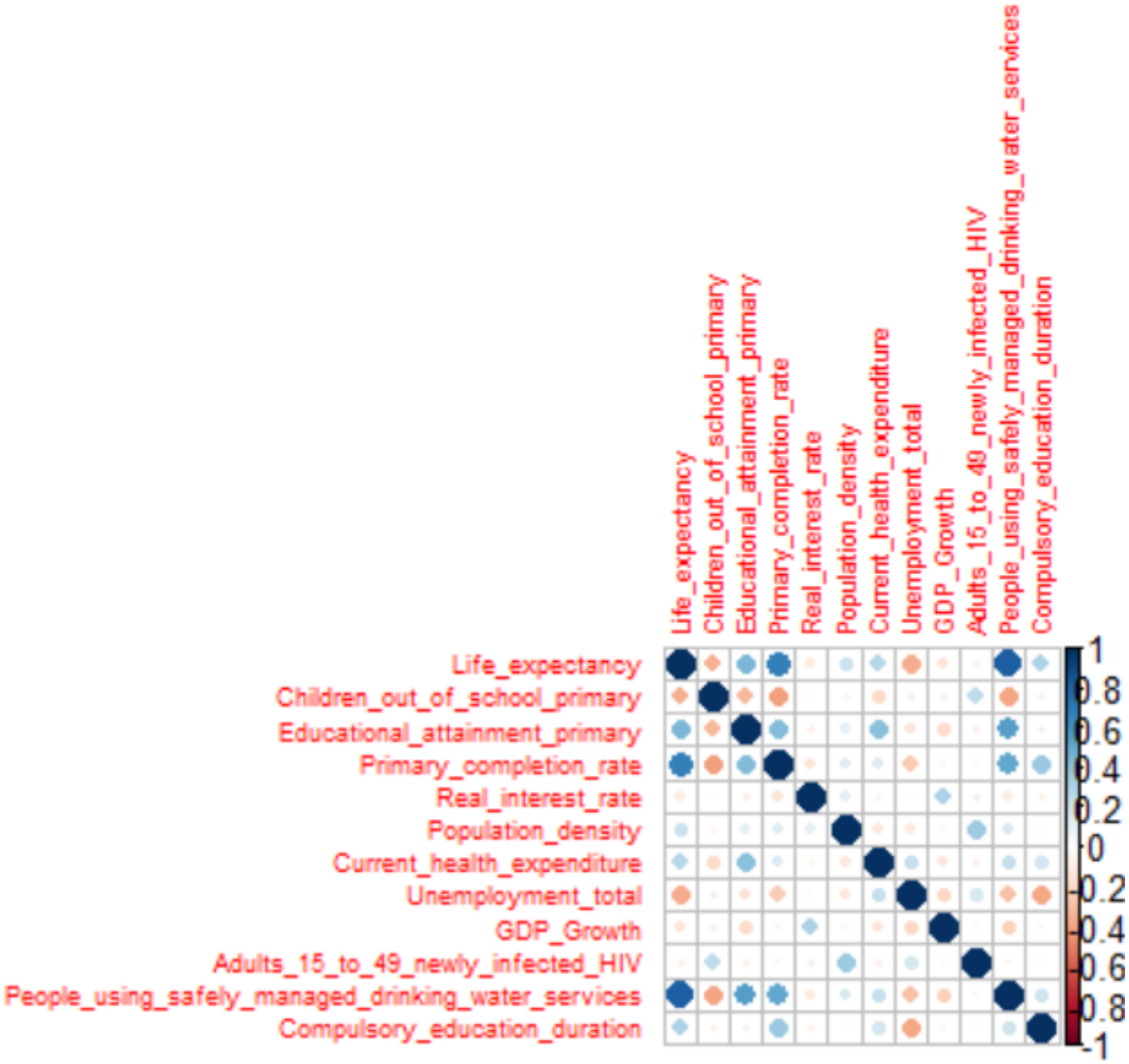
8**. Correlation Matrix After removal of attributes with high VIF values**

Relation2<-DataC[,c(-2,-3,-4,-7,-10,-12,-16,-17)]

FnRelation<-cor(Relation2)

```
corrplot(FnRelation,tl.pos ='lt', tl.cex=0.55,method='circle')

summary(Relation2)
```



```
> summary(Full_modelCor)

Call:
lm(formula = Life_expectancy ~ ., data = Relation2)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6010 -1.9200  0.3045  1.9654  9.7605

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.874e+01  2.174e+00  22.417  < 2e-16 ***
Child_noschool         2.439e-06  1.095e-06   2.228 0.026981 *
Edu_att_primary       -6.920e-02  1.769e-02  -3.912 0.000125 ***
Pr_compl_rate          2.041e-01  2.190e-02   9.319  < 2e-16 ***
Real_int_rate          9.711e-03  2.560e-02   0.379 0.704776
Population_density     4.005e-04  1.284e-04   3.120 0.002067 **
Cur_health_expen       5.128e-01  9.683e-02   5.296 3.04e-07 ***
Unemployment_total    -1.654e-01  4.512e-02  -3.665 0.000314 ***
GDP_Growth            -3.371e-02  7.525e-02  -0.448 0.654660
Adult_new_HIV         -2.025e-05  1.092e-05  -1.854 0.065152 .
People_drinking_water  1.551e-01  1.032e-02  15.031  < 2e-16 ***
Com_education_durat   -9.502e-02  1.041e-01  -0.913 0.362404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.289 on 205 degrees of freedom
Multiple R-squared:  0.8112,    Adjusted R-squared:  0.8011
F-statistic: 80.07 on 11 and 205 DF,  p-value: < 2.2e-16

>
```

Appendix[2]

9. Finding the best model

```
library('faraway')

#Forward feature selection

bestmodel1<-lm(Life_expectancy~1,data=Relation3)

step1<-step(bestmodel1,scope=~ Child_noschool + Edu_att_primary+ Pr_compl_rate + Population_density + Cur_health_expen + Unemployment_total + People_drinking_water,

        method='forward')
```

summary(step1)

AIC(bestmodel1)

BIC(bestmodel1)

```
- Edu_att_primary        1    137.67 2517.8 541.92
- Unemployment_total     1    199.65 2579.8 547.20
- Cur_health_expen       1    253.09 2633.2 551.65
- Pr_compl_rate          1    986.84 3367.0 604.99
- People_drinking_water  1   2707.51 5087.7 694.57

Step:  AIC=526.23
Life_expectancy ~ People_drinking_water + Pr_compl_rate + Unemployment_total +
    Cur_health_expen + Edu_att_primary + Population_density

                        Df Sum of Sq    RSS    AIC
+ Child_noschool         1     29.91 2269.5 525.39
<none>                               2299.4 526.23
- Population_density     1     80.78 2380.2 531.72
- Edu_att_primary        1    155.48 2454.9 538.43
- Unemployment_total     1    200.19 2499.6 542.34
- Cur_health_expen       1    299.26 2598.6 550.78
- Pr_compl_rate          1    957.73 3257.1 599.79
- People_drinking_water  1   2588.92 4888.3 687.89

Step:  AIC=525.39
Life_expectancy ~ People_drinking_water + Pr_compl_rate + Unemployment_total +
    Cur_health_expen + Edu_att_primary + Population_density +
    Child_noschool

                        Df Sum of Sq    RSS    AIC
<none>                               2269.5 525.39
- Child_noschool         1     29.91 2299.4 526.23
- Population_density     1     79.12 2348.6 530.82
- Edu_att_primary        1    146.88 2416.3 537.00
- Unemployment_total     1    195.82 2465.3 541.35
- Cur_health_expen       1    310.46 2579.9 551.21
```

```
- People_drinking_water   1    2606.33 4873.8 689.34
> summary(step1)

Call:
lm(formula = Life_expectancy ~ People_drinking_water + Pr_compl_rate +
    Unemployment_total + Cur_health_expen + Edu_att_primary +
    Population_density + Child_noschool, data = Relation3)

Residuals:
   Min     1Q Median     3Q    Max
-9.447 -1.858  0.229  2.094 10.193

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.842e+01  1.984e+00  24.407  < 2e-16 ***
People_drinking_water 1.551e-01  1.001e-02  15.493  < 2e-16 ***
Pr_compl_rate         1.925e-01  2.024e-02   9.508  < 2e-16 ***
Unemployment_total   -1.623e-01  3.821e-02  -4.247 3.27e-05 ***
Cur_health_expen      4.874e-01  9.116e-02   5.347 2.34e-07 ***
Edu_att_primary      -6.329e-02  1.721e-02  -3.678 0.000299 ***
Population_density    3.149e-04  1.167e-04   2.699 0.007516 **
Child_noschool        1.731e-06  1.043e-06   1.660 0.098493 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.295 on 209 degrees of freedom
Multiple R-squared:  0.8068,    Adjusted R-squared:  0.8003
F-statistic: 124.6 on 7 and 209 DF,  p-value: < 2.2e-16
```

Appendix[3]

#Backward feature selection

bestmodel2<-lm(Life_expectancy~.,data=Relation3)

step2<-step(bestmodel2,method="backward")

summary(step2)

AIC(bestmodel2)

BIC(bestmodel2)

```
 AIC(bestmodel1)
1] 1485.905
 BIC(bestmodel1)
1] 1492.665
 #Backward feature selection
 bestmodel2<-lm(Life_expectancy~.,data=Relation3)
 step2<-step(bestmodel2,method="backward")
tart:  AIC=525.39
ife_expectancy ~ Child_noschool + Edu_att_primary + Pr_compl_rate +
    Population_density + Cur_health_expen + Unemployment_total +
    People_drinking_water

                       Df Sum of Sq    RSS    AIC
none>                                2269.5 525.39
 Child_noschool         1     29.91 2299.4 526.23
 Population_density     1     79.12 2348.6 530.82
 Edu_att_primary        1    146.88 2416.3 537.00
 Unemployment_total     1    195.82 2465.3 541.35
 Cur_health_expen       1    310.46 2579.9 551.21
 Pr_compl_rate          1    981.58 3251.1 601.38
 People_drinking_water  1   2606.35 4875.8 689.34
```

Predicting the best model

modelfit2 <- lm(Life_expectancy ~ Edu_att_primary + Pr_compl_rate + Population_density + Cur_health_expen + Unemployment_total + People_drinking_water, data = Relation3)

summary(modelfit2)

AIC(modelfit2)

```
> summary(modelfit2)

Call:
lm(formula = Life_expectancy ~ Edu_att_primary + Pr_compl_rate +
    Population_density + Cur_health_expen + Unemployment_total +
    People_drinking_water, data = Relation3)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7226 -1.9461  0.2437  1.9864  9.8019

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           49.7828579  1.8146420  27.434  < 2e-16 ***
Edu_att_primary       -0.0650008  0.0172493  -3.768 0.000214 ***
Pr_compl_rate          0.1840493  0.0196793   9.352  < 2e-16 ***
Population_density     0.0003182  0.0001171   2.716 0.007154 **
Cur_health_expen       0.4775241  0.0913408   5.228 4.13e-07 ***
Unemployment_total    -0.1639907  0.0383522  -4.276 2.89e-05 ***
People_drinking_water  0.1521661  0.0098959  15.377  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.309 on 210 degrees of freedom
Multiple R-squared:  0.8042,    Adjusted R-squared:  0.7986
F-statistic: 143.8 on 6 and 210 DF,  p-value: < 2.2e-16
```

modelfit1 <- lm(Life_expectancy ~ People_drinking_water + Pr_compl_rate + Unemployment_total + Cur_health_expen + Edu_att_primary + Population_density , data = Relation3)

summary(modelfit1)

```
> summary(modelfit1)

Call:
lm(formula = Life_expectancy ~ People_drinking_water + Pr_compl_rate +
    Unemployment_total + Cur_health_expen + Edu_att_primary +
    Population_density, data = Relation3)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7226 -1.9461  0.2437  1.9864  9.8019

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           49.7828579  1.8146420  27.434  < 2e-16 ***
People_drinking_water  0.1521661  0.0098959  15.377  < 2e-16 ***
Pr_compl_rate          0.1840493  0.0196793   9.352  < 2e-16 ***
Unemployment_total    -0.1639907  0.0383522  -4.276 2.89e-05 ***
Cur_health_expen       0.4775241  0.0913408   5.228 4.13e-07 ***
Edu_att_primary       -0.0650008  0.0172493  -3.768 0.000214 ***
Population_density     0.0003182  0.0001171   2.716 0.007154 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.309 on 210 degrees of freedom
Multiple R-squared:  0.8042,    Adjusted R-squared:  0.7986
F-statistic: 143.8 on 6 and 210 DF,  p-value: < 2.2e-16
```