Linear model to explain the Life Expectancies in 2020

Souranu Maji
Registration Number: 2200470
PRID: MAJIS78506

Abstract:

Life expectancy is defined as a mathematical estimation of how many years a person is expected to survive. This project aims to explain the life expectancy in the world for the year 2020 through a linear model based on the dataset of the World Development Indicators (WDI) derived from a primary World Bank database for development data from officially recognized international sources.

Table of Contents

Word Count: 2047

Introduction:

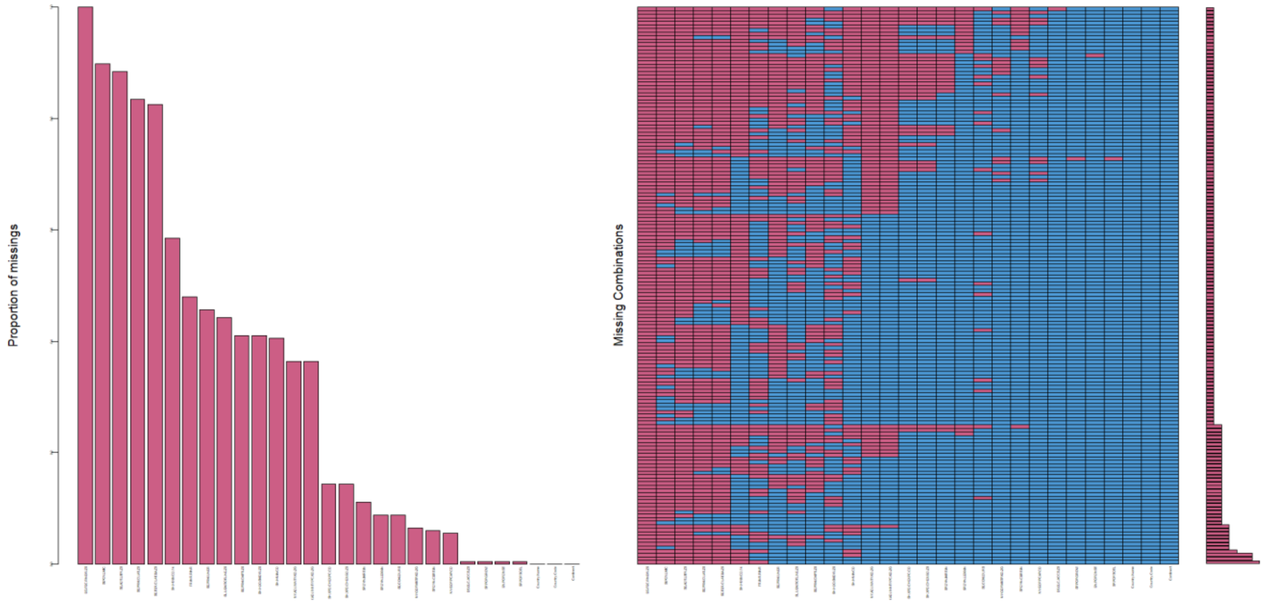
We will analyze the dataset of World Development Indicators (WDI) and to propose a linear model which explains the life expectancies for the year 2020. The WDI includes variables like “Birth Rate”, “Mortality Rate”, “Population Growth”, “Current Health Expenditure per Capita”, etc. In this report we will do a comprehensive analysis of these indicators using regression. But before that we need to do some pre-processing on our data, like removing the NA values and imputing the null values. After that we need to check for collinearity in the data. Then we have to find the best model which defines the life expectancy based on general linear model assumptions.

- Access to electricity
- Adjusted net national income
- Adjusted net national income per capita
- Children newly infected with HIV
- Children out of primary school
- Population of people attaining primary school education
- Total population 25+
- Population of those 25+ attaining a Bachelor’s or equivalent
- Infant mortality rate
- Primary completion rate
- Adult literacy rate (ages 15+)
- Real interest rate
- Annual population growth
- Population density
- Total population
- Current health expenditure per capita
- Current health expenditure
- Total unemployment
- Crude birth rate
- Renewable energy consumption
- Adults (ages 15-49) newly infected with HIV
- Percentage of people using safely managed drinking water services
- Poverty headcount ratio
- Duration of compulsory education

The effect of above variables on the life expectancy will be analysed for the report.

Preliminary Analysis:

- There are 217 rows and 29 columns in our dataset and there are no duplicates.
- The column EG.FEC.RNEW. ZS contains all null values, and we can remove it.
- There are 19 missing values in the response variable SP.DYN.LE00.IN. The mean is 72.93 and the standard deviation is 7.5. The maximum value is 85.08 and the minimum is 53.28, which can be accepted.
- For the rest of the columns, we need to calculate the proportion of null values for every column in the dataset. Then we should remove all the columns where the proportion of null value is more than 80%.



From the figure we can get the proportionality of missing values

Descriptive Analysis:

Dealing with missing values:

While working with the dataset provided, we can observe that many indicators in the dataset contain missing values. There could be distortion in our analysis results by leading to biased estimates. For the values that are missing at random, the deleted values could possibly reduce the bias. But the data removal may also discard plenty of useful information from the data and that's why this may not be the best option in our case, as we don't have enough observations to give a definitive analysis.

There are a grand total of 1981 missing values within the data set. When viewing the missing data, it can be seen that one column contains only missing values. It is best for this column to be removed because the lack of data means that it would be impossible to input relevant values into it. There are six columns, out of the 29, in which over 50% of the data is missing. These columns are removed as keeping them would lead to more imputed values than original data which may skew the results. It also does not significantly affect the data as only 6 of the 25 predictor variables are removed.

Therefore, we will use the Multiple Imputation. Through Multiple Imputation we will avoid the discarding of necessary information in our dataset. At the same time, we can avoid imposing bias in the estimators while reducing the variability of our data.

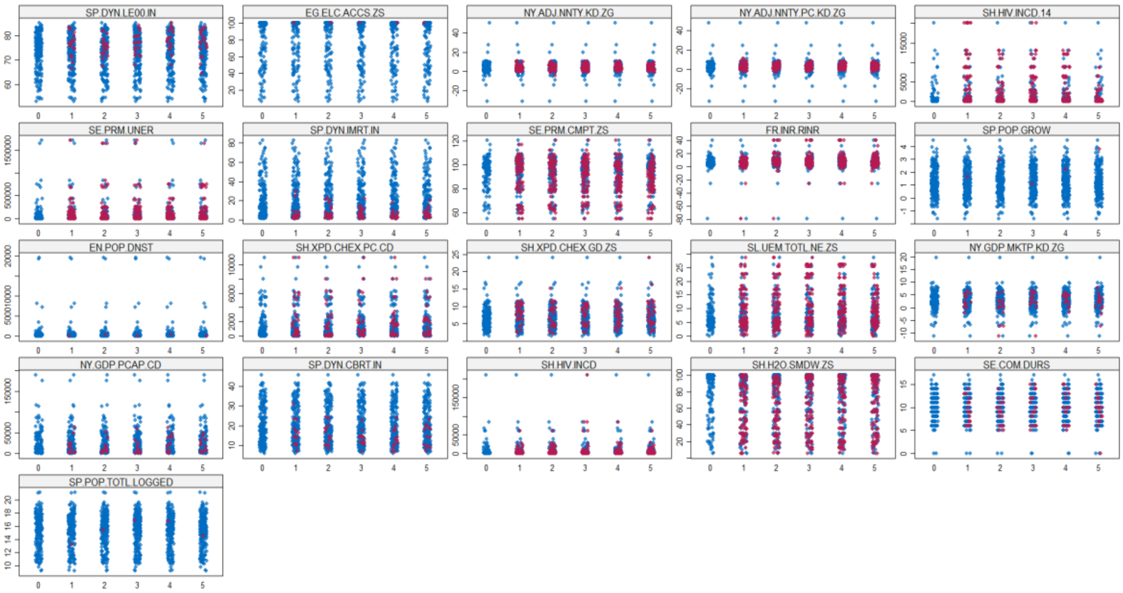
We have decided to calculate the proportion of null values for every column in the dataset. Then we should remove all the columns where the proportion of null value is more than 80%. Subsequently, we have to perform the log transform for the columns "SP.POP.TOTAL" and "NY.GDP.PCAP.CD" for scaling the range of the data.

The MICE (Multivariate Imputation by Chained Equations) package in R creates multiple imputations (replacement values) for multivariate missing data. The imputation method is based on fully conditional specification, where each incomplete variable is imputed by a separate model. We need to implement MICE for imputing the missing values for all the remaining columns of the dataset. MICE have three stages:

1. Imputation Stage
2. Analysis Stage
3. Pooling Stage

We can use the Predictive mean matching (PMM) method with Maximum Iteration (maxit) = 50. After imputing all the null values in every column, we can take help of lattice.stripplot to verify whether the values are imputed correctly. Plotting methods for imputed data using lattice.stripplot produces one-dimensional scatterplots. The function automatically separates the observed and the imputed data.

For our observations the imputed values we will get, is expected to be aligned to the actual values given in the dataset. The imputed values could be checked to ensure that there are no issues, example, negative value(s) when the imputed value(s) should be positive and vice versa.



From Figure we can tell the imputed values is following the same pattern as the observed values, so we can infer that the imputed values are aligned with the observed data and no issues was diagnosed.

From the results of the pooling model we observed that there are some negative values NY.ADJ.NNTY.KD.ZG,SP.DYN.CBRT.IN, NY.GDP.PCAP.CD.LOGGED, SP.DYN.IMRT.IN which can be considered as insignificant. So again we need to check the collinearity among the features.

How to handle the missing life expectancy data for some countries?

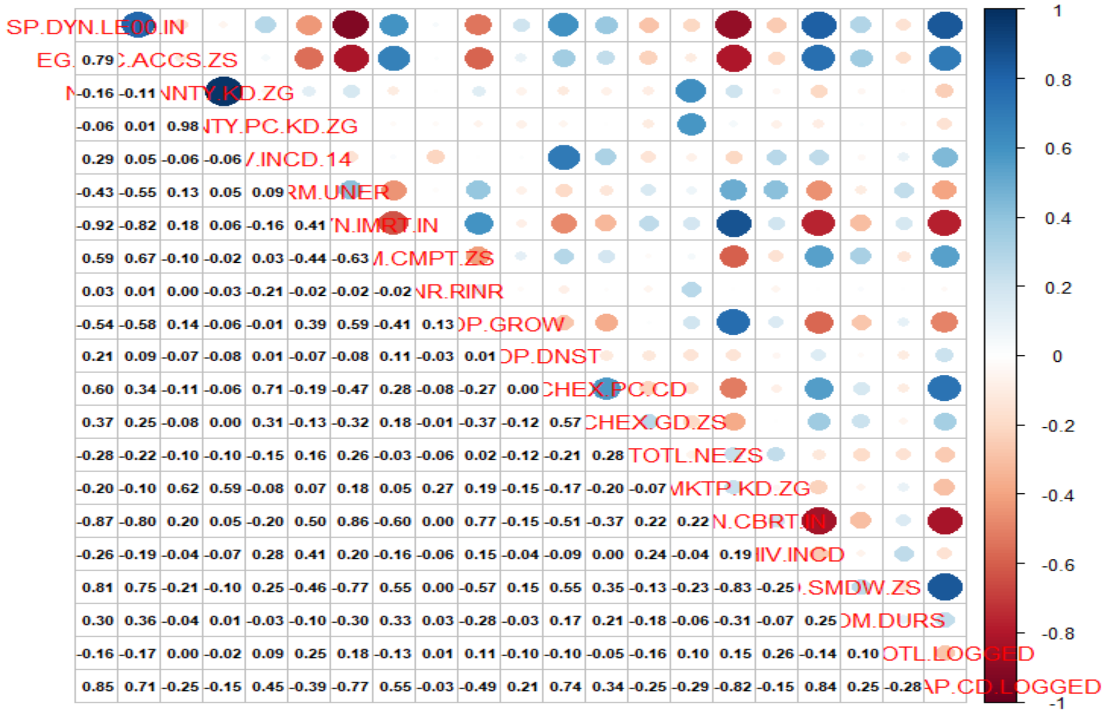
There are 19 rows containing missing life expectancy values. We will handle this scenario by removing the rows from our dataset. This is done in order to create a complete case of response values which means the values will not have to be imputed.

Dealing with collinearity:

Collinearity refers to the linear association between two predictor variables. For our dataset, we plotted the heatmap and found out that two or more predictor variables in our dataset are highly correlated. Multi Collinearity is not desirable as it increases the variance of the coefficient estimates and make the estimation very sensitive to minor variations in the model. We will measure the collinearity using the Variation Inflation Factor (VIF).

We can start our observation by fitting the entire model with imputed dataset. A total of 5 imputed datasets are present for us to analyze and interpret the results.

We can see the presence of multicollinearity. We will start by checking the pair wise correlation between all the predictor variables and then we can use the R Function VIF to calculate the VIF value of each predictor variable. For visualizing the collinearity we can use corrplot library provided by R. The corrplot provides a visual exploratory tool on correlation matrix that supports automatic variable reordering to help detect hidden patterns among variables.



From the figure we can determine the correlation between the different features variables and we have dropped the following columns with VIF value greater than 5 NY.ADJ.NNTY.KD.ZG, SP.DYN.CBRT.IN, NY.GDP.PCAP.CD.LOGGED, SP.DYN.IMRT.IN

Linear Model:

After imputing the values, we get out full model. To eliminate the multicollinearity problem, we have already deducted some features from the given dataset.

We will implement the wrapper method for our model selection, as we are interested in finding out the best subset of features from all the available set of features, so that we can get the best linear model that can predict the life expectancy.

Feature Selection:

First Step: From the ***Forward feature selection method*** we get the below results:

Step: AIC=480.74

SP.DYN.LE00.IN ~ SH.H2O.SMDW.ZS + EG.ELC.ACCS.ZS + SH.HIV.INCD.14 + SH.HIV.INCD + EN.POP.DNST + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS + SE.PRM.CMPT.ZS + SE.PRM.UNER + NY.GDP.MKTP.KD.ZG + FR.INR.RINR

Second Step: From the ***backward feature selection method*** we get the below results:

Step: AIC=480.74

SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SH.HIV.INCD.14 + SE.PRM.UNER + SE.PRM.CMPT.ZS + FR.INR.RINR + EN.POP.DNST + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS + NY.GDP.MKTP.KD.ZG + SH.HIV.INCD + SH.H2O.SMDW.ZS

Comparing the above two results we can observe that the AIC value for both the results are equal and which is 480.74, and the feature variables are also same for above two results, so we can conclude that the following features will contribute towards making the best possible model. The equation for the reduced model after the feature selection is given by

SP.DYN.LE00.IN = 4.896 + 0.1254 * EG.ELC.ACCS.ZS + 0.0002664* SH.HIV.INCD.14 + 0.000002684* SE.PRM.UNER + 0.07525 * SE.PRM.CMPT.ZS + 0.05144 * FR.INR.RINR + 0.000465* EN.POP.DNST + 0.3931 * SH.XPD.CHEX.GD.ZS - 0.1655 * SL.UEM.TOTL.NE.ZS - 0.2107 * NY.GDP.MKTP.KD.ZG - 0.00005089* SH.HIV.INCD + 0.0683* SH.H2O.SMDW.ZS

ANOVA Comparison full Model vs Revised Model

For comparing the reduced and the full model we have used R built-in analysis of variance or anova method. We also have the plots for the residuals vs fitted, Normal Q-Q, scale location and residuals vs leverage plots, for both the models.

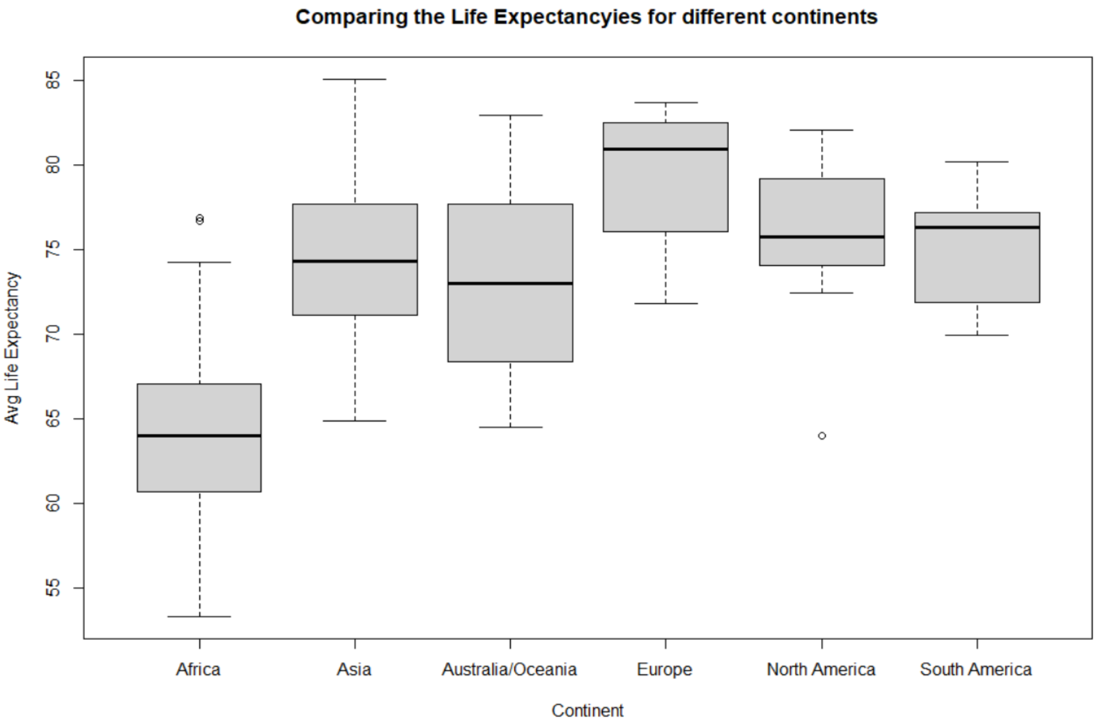
```
Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    186 1988.30
2    177  657.34  9      1331 39.821 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above comparisons we selected the reduced model

Life Expectancies across different Countries:

Before investigating the differences in average life expectancies of six different continents we can evaluate some statistical summary like check the mean for each continent.

In order to compare the life expectancies among six different continents we used the boxplot.

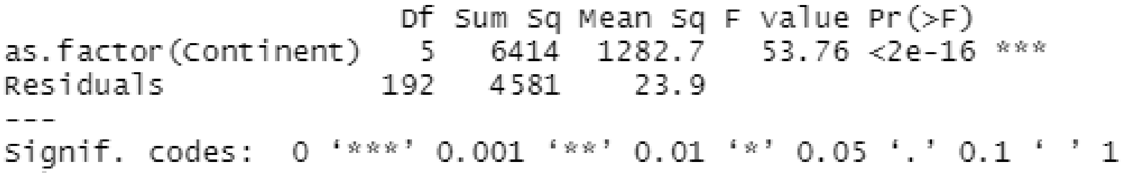


Africa	Asia	Australia/Oceania	Europe	North America	South America
64.11014	74.61739	73.52827	79.28428	76.17408	75.09100

The above values are the mean for life expectancies of the continents respectively.

As the purpose of the study is to understand the average life expectancies across the continents, therefore we employ the one ANOVA to investigate if there is any difference between the mean life expectancies across different continents.

For the above purpose we use the R built in function aov(). This produces the analysis of variance table.



```
                       Df Sum Sq Mean Sq F value Pr(>F)    
as.factor(Continent)    5   6414  1282.7   53.76 <2e-16 ***
Residuals             192   4581    23.9                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above table we can reject the null hypothesis at any significance level.

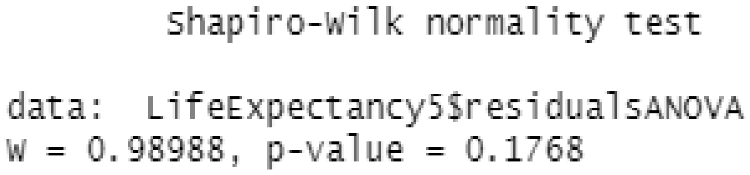
Post hoc test:

To investigate weather all the life expectancy values are different from each other or only one is different from the others we are interested in doing some post hoc tests. These tests are basically t test with adjusted p values for multiple testing. In our case we are using Bonferroni and Tukey’s Honest Significant Differences post -hoc tests and we have observed there is a significant difference between Africa and all other continents, Asia and Europe, Australia and Europe and Europe with North America and South America.

One-way ANOVA assumptions satisfied:

To satisfy the ANOVA assumptions we can also check for normality of residuals and existence of equal variances between different life expectancy values of the continents. The plot for standard residuals and quantiles is include in the appendix section.

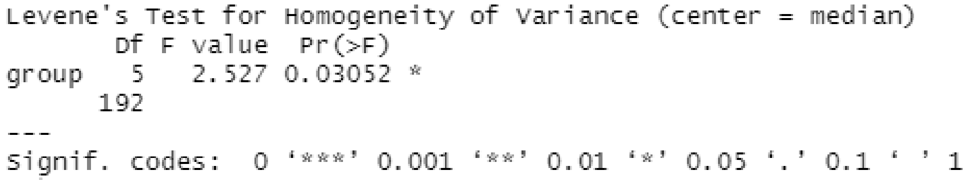
Shapiro test can also be done for checking the same scenario.



```
	Shapiro-Wilk normality test

data:  LifeExpectancy5$residualsANOVA
W = 0.98988, p-value = 0.1768
```

The p value is large we fail to reject the null hypothesis. That means the normality assumption is satisfied for the predicted model.

To check the homogeneity of variance, we conduct the Levene’s test.



```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   5   2.527 0.03052 *
      192
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As the p value is small, we can assume that there is equal variances between the life expectancy of different continents.

Conclusion:

The density plot shows that the pattern of the imputed values is reasonably similar to the observed values. Thus, we can successfully justify the imputed values. By removing the highly correlated variable we reduced multi collinearity in our data.

Thus, the best predicted model for life expectancy after doing the analysis over the given data set depends on the following variables.:

1. Access to electricity
2. Children (ages 0 to 14) newly infected with HIV
3. Children out of school, primary
4. Primary completion rate
5. Real Interest rate
6. Population Density
7. Current health expenditure
8. Unemployment
9. GDP Growth
10. Adults (ages 15-49) newly infected with HIV
11. People using safely managed drinking water services

Appendix:

A1. R code for task one

A2. Figure for task one

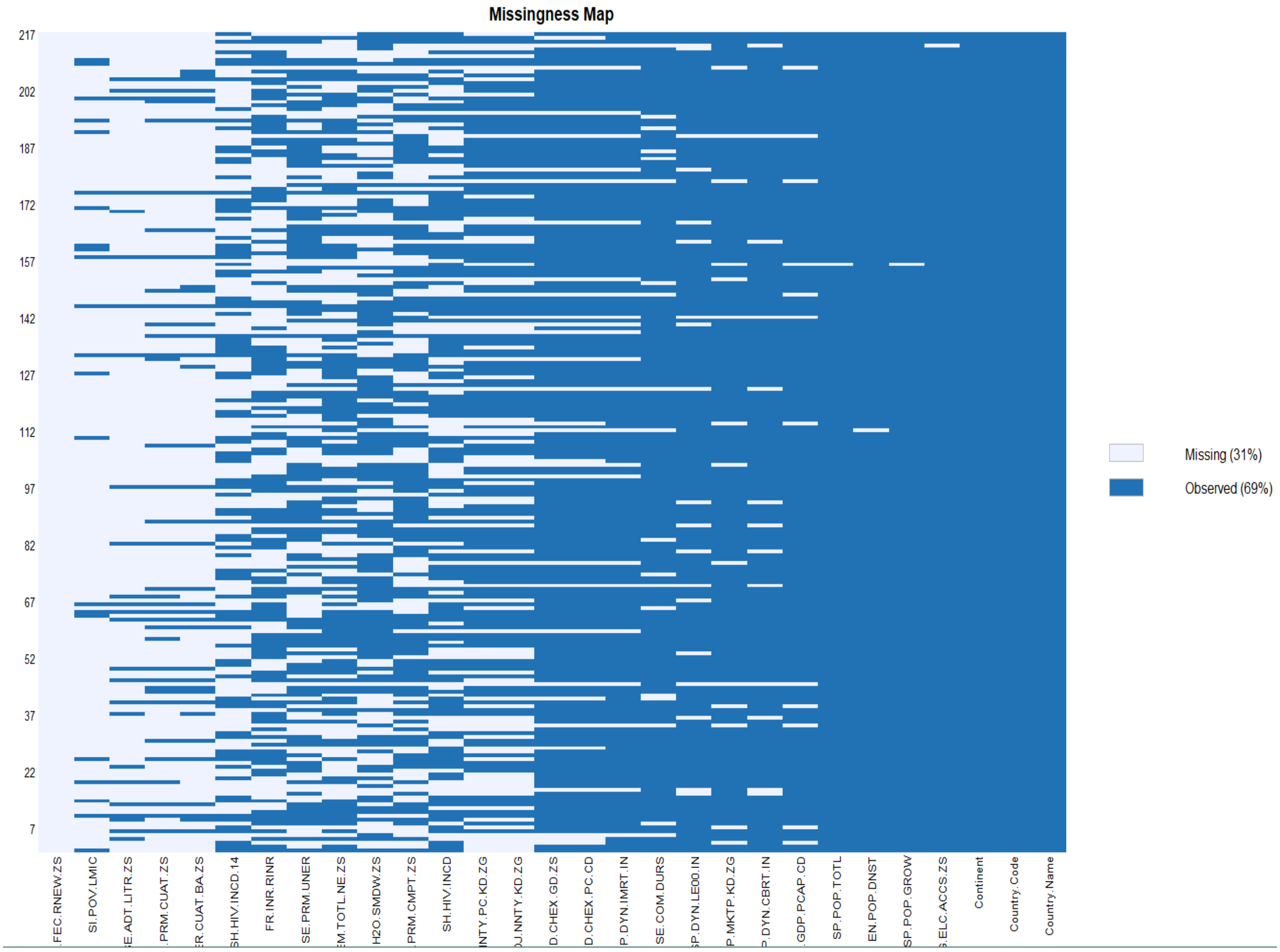
A1. R code for task one

```
library(dplyr)
library(gapminder)
library(tidyverse)
library(corrplot)
library(ggplot2)
library(psych)
library(broom)
library(purrr)
library(tidyr)
library(naniar)
library(Amelia)
library(mice)
library(VIM)
library(GGally)
library(car)
library(reshape2)
library(hrbrthemes)
library(faraway)
library(funModeling)
library(mctest)
library(faraway)
library(corrplot)
library(fun)


####****************************************************Question -
1*********************************************###################
###-----------------------Reading Data------------------------#########
LifeExpectancy <- read.csv("Z:\\Statistics\\Life_Expectancy_Data1.csv")
str(LifeExpectancy)
dim(LifeExpectancy)
df_status(LifeExpectancy)



##################################################------------FILTERING DATA---
-----------------------#######################################################



#-----------------> counting initial Null values per column
initial_na_count <-sapply(LifeExpectancy, function(y) sum(length(which(is.na(y)))))
initial_na_count <- data.frame(initial_na_count)
initial_na_count
missmap(LifeExpectancy)


###---------------- Checking Proportions of Missing Values----#########
aggr(LifeExpectancy, col=mdc(1:2), numbers=TRUE, sortVars=TRUE,
labels=names(LifeExpectancy),
                    cex.axis=.3, gap=3,
                    ylab=c("Proportion of missings","Missing Combinations"))
```

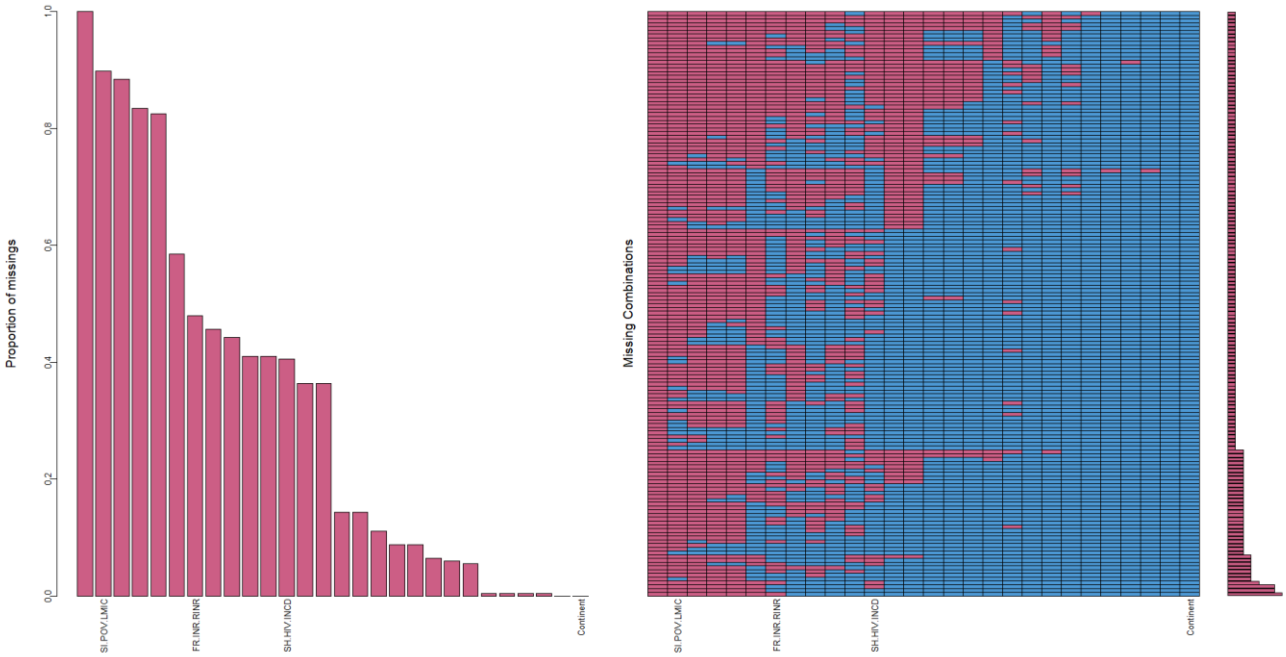
```
#-----------------> eliminating columns containing more than 80% NA
life_expectancy_filtered <- LifeExpectancy %>% select(which(colMeans(is.na(.)) <= 0.8))
names(life_expectancy_filtered)
dim(life_expectancy_filtered)
df_status(life_expectancy_filtered)
summary(life_expectancy_filtered)


###------Rechecking filtered
missmap(life_expectancy_filtered)



####---------------------------Removing Data with missing values for life_expectancy------------
##########


life_expectancy_filtered_with_missing_target <- life_expectancy_filtered
dim(life_expectancy_filtered_with_missing_target)


missingLifeExpectancy <- which(is.na(life_expectancy_filtered$SP.DYN.LE00.IN))
life_expectancy_filtered <- life_expectancy_filtered[-missingLifeExpectancy,]
dim(life_expectancy_filtered)
```

```
###---------------------- Life expectancy VS Continent -------------########
ggplot(life_expectancy_filtered) +
  geom_bar(mapping = aes(x = Continent, fill = Continent))




# x11(width = 20, height = 14)
# layout(matrix(c(1,2), 1, 2, byrow = TRUE))
# hist(life_expectancy_filtered$SP.DYN.LE00.IN)


###-------------Removing other categorical Data for Processing----------------###


LifeExpectancy2 <- life_expectancy_filtered %>% select(-c('Country.Code','Continent',
'Country.Name'))
##LifeExpectancy2 <- LifeExpectancy2 %>% select(-'Country.Name')
names(LifeExpectancy2)



#####---------------------- Analysis of Relationship among Target Variable(Life expectancy) and
the Predictors -------------------###########
eda_Func <- function(x, var_name, dataframe, targetVar){
  x11(width = 20, height = 14)
  #nf <- layout( matrix(c(1,3), ncol=3) )
  par(mfrow = c(1, 3))
  hist(x, main = paste0("Histogram of",var_name))
  boxplot(x,main = paste0("Boxplot of", var_name))
  plot(x, y = dataframe$targetVar, xlab = var_name, ylab = "Life expectancy", main = paste0("Life
exp vs ", var_name))
}




var_names <- names(LifeExpectancy2)
for(i in 1:dim(LifeExpectancy2)[2]){

  print(eda_Func(LifeExpectancy2[,i], var_names[i], LifeExpectancy2, SP.DYN.LE00.IN))
}
```

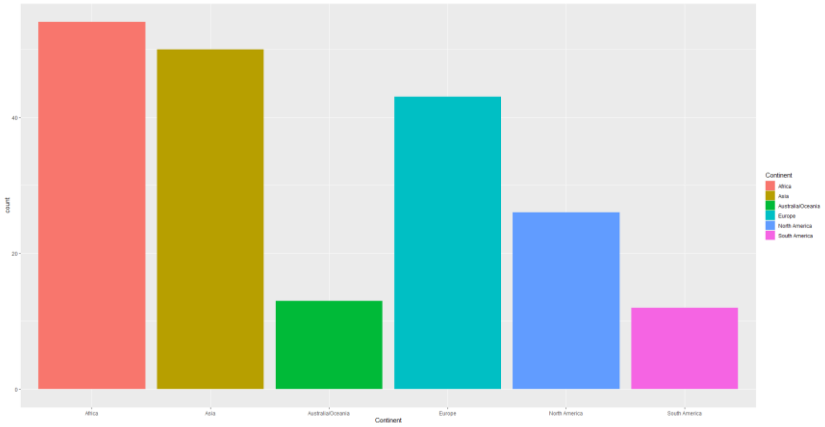
A2. Figure for task One



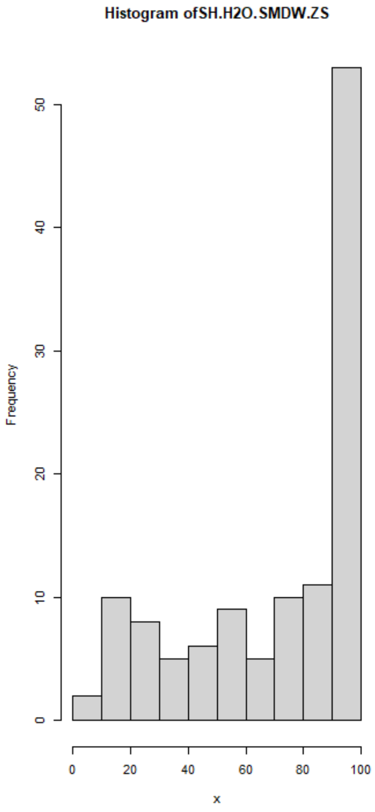
Missing Data Map

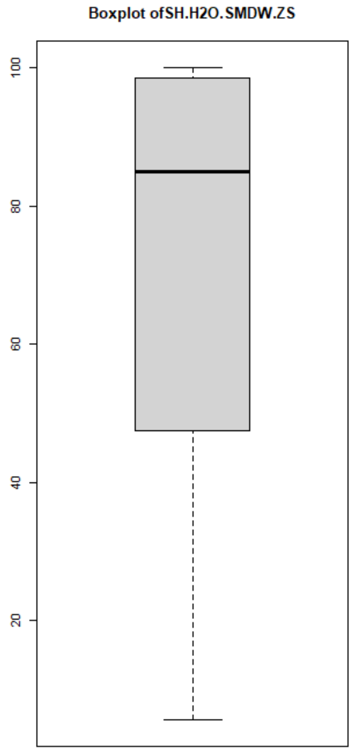


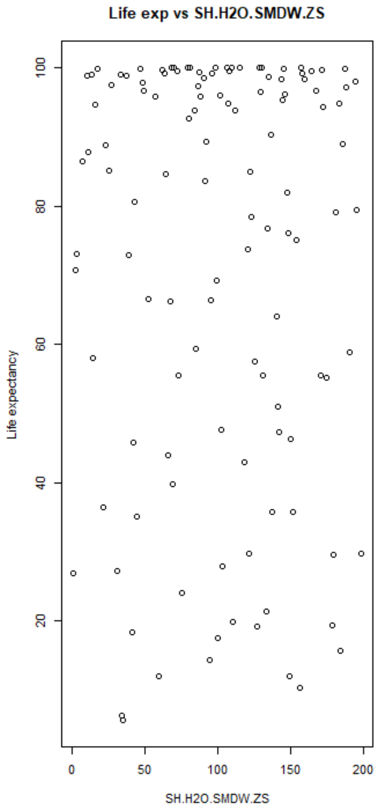
Proportion of Missing Values

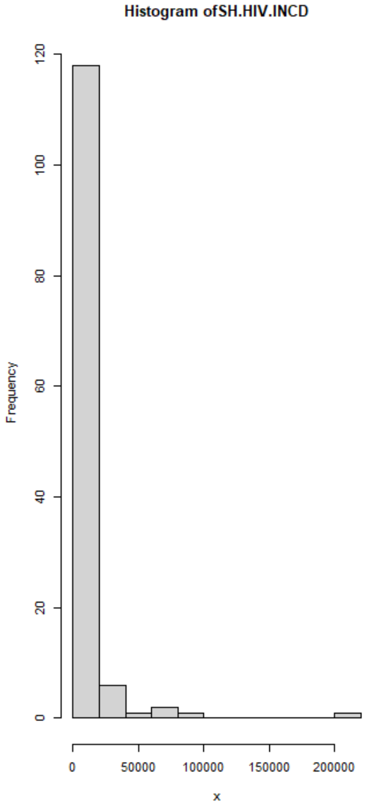


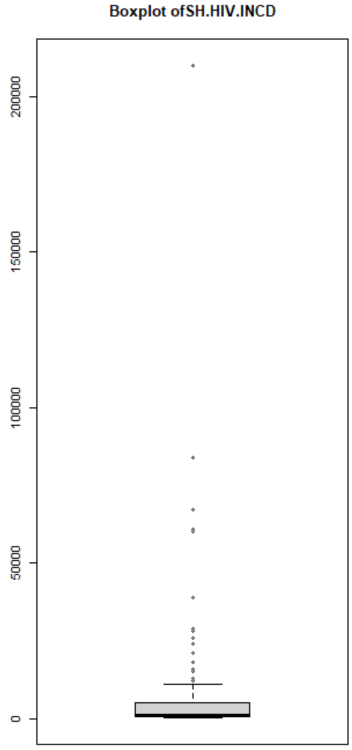
Continent – wise count

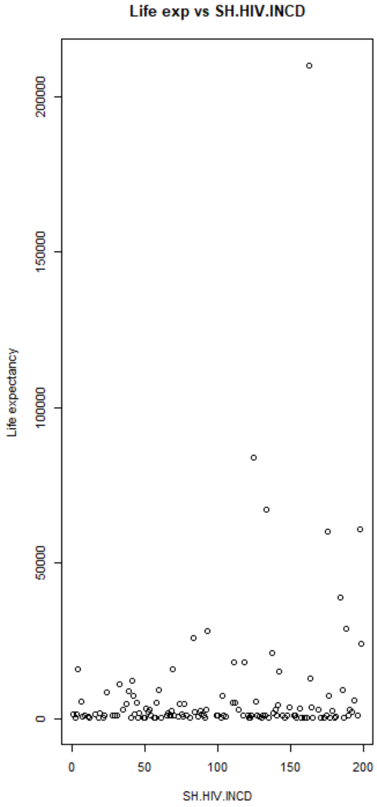

Histogram ofSH.H2O.SMDW.ZS
Frequency
0
10
20
30
40
50
0
20
40
60
80
100
x

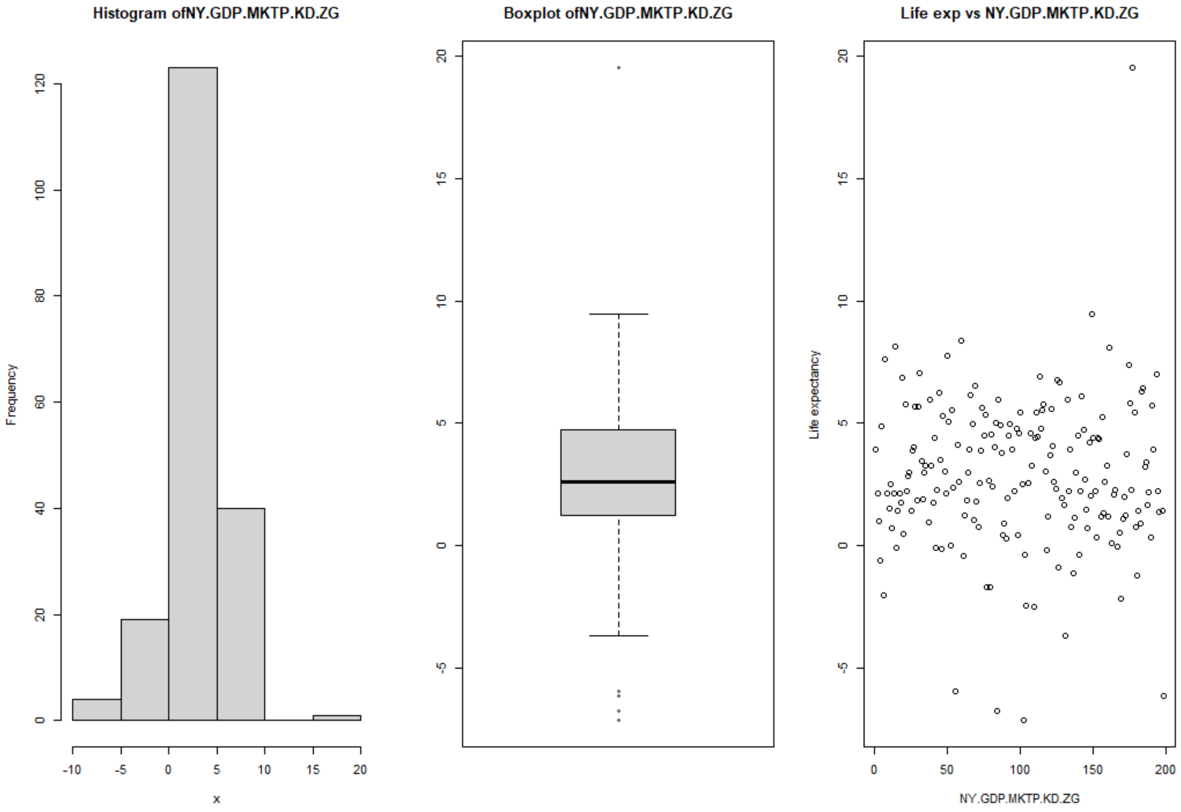

Boxplot ofSH.H2O.SMDW.ZS
20
40
60
80
100

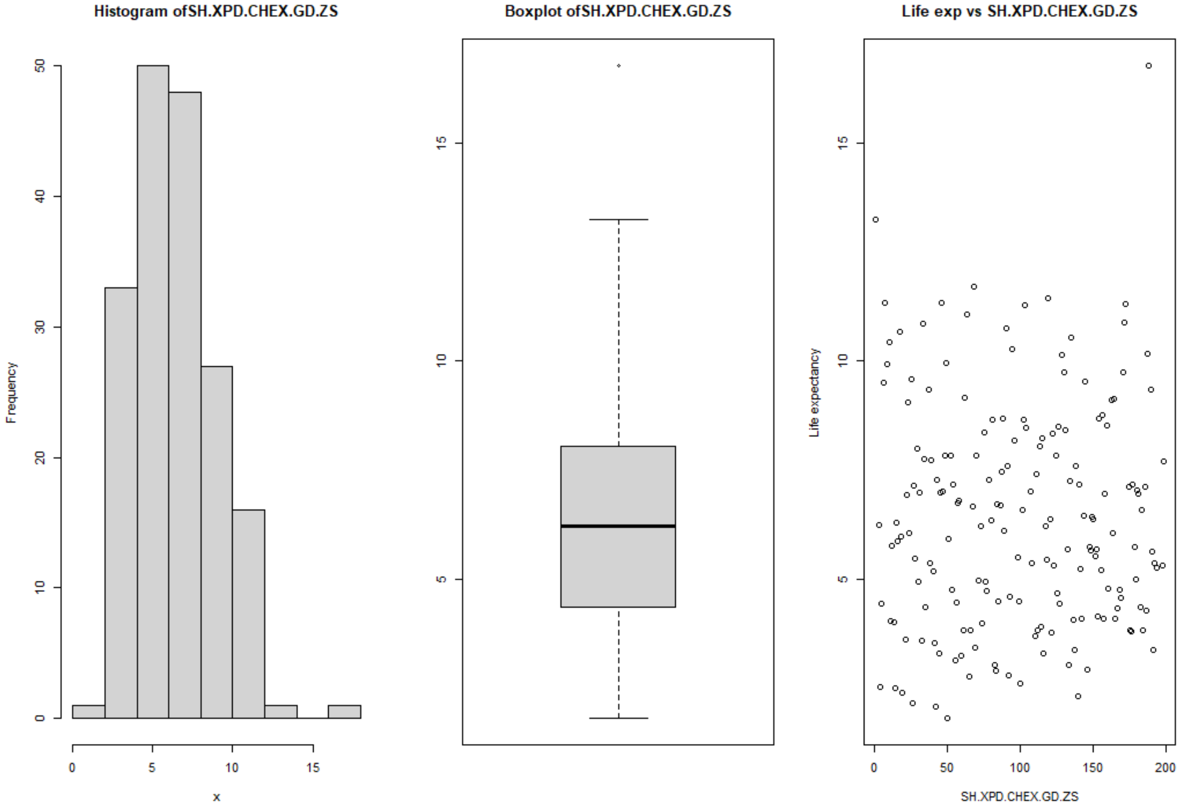

Life exp vs SH.H2O.SMDW.ZS
Life expectancy
20
40
60
80
100
0
50
100
150
200
SH.H2O.SMDW.ZS

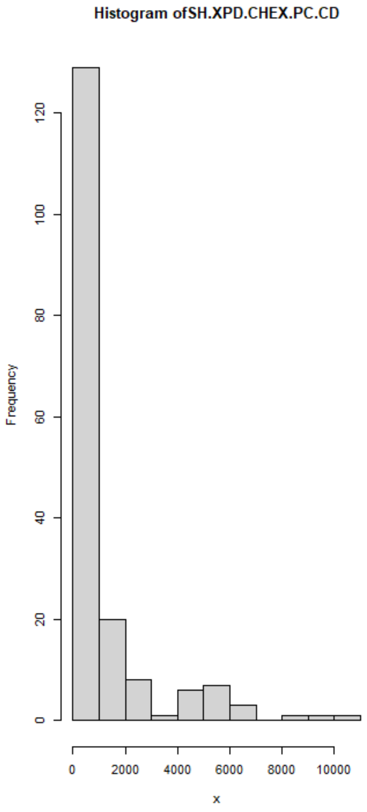

Histogram ofSH.HIV.INCD
Frequency
0
20
40
60
80
100
120
0
50000
100000
150000
200000
x

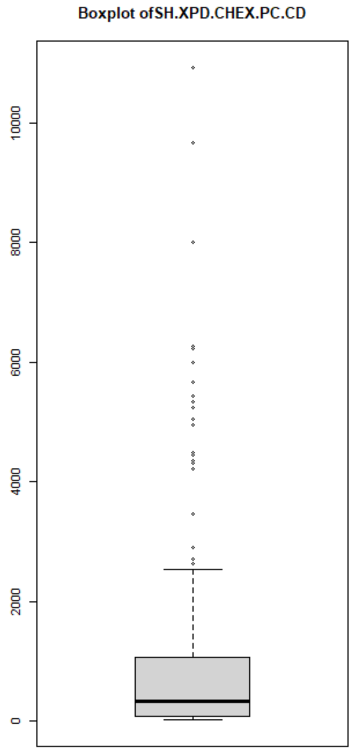

Boxplot ofSH.HIV.INCD
0
50000
100000
150000
200000

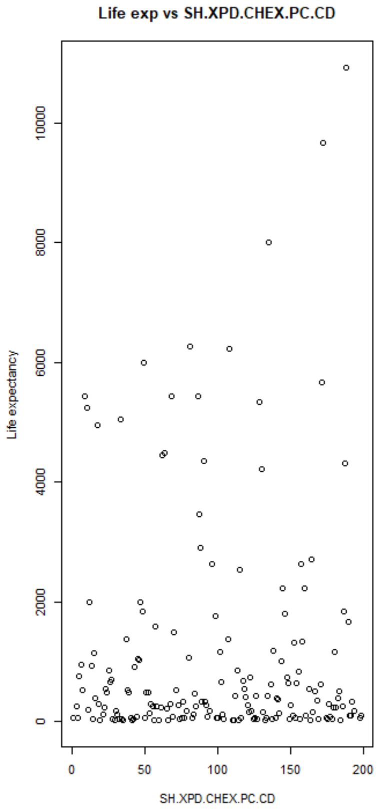

Life exp vs SH.HIV.INCD
Life expectancy
0
50000
100000
150000
200000
0
50
100
150
200
SH.HIV.INCD

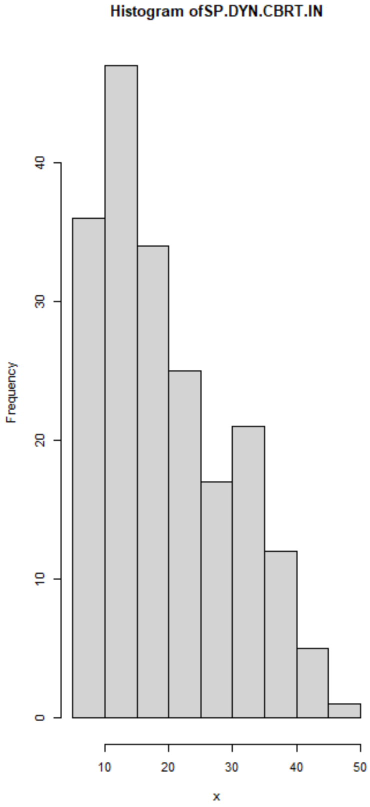

Histogram ofNY.GDP.MKTP.KD.ZG
Frequency
0
20
40
60
80
100
120
-10
-5
0
5
10
15
20
x
Boxplot ofNY.GDP.MKTP.KD.ZG
-5
0
5
10
15
20
Life exp vs NY.GDP.MKTP.KD.ZG
Life expectancy
-5
0
5
10
15
20
0
50
100
150
200
NY.GDP.MKTP.KD.ZG

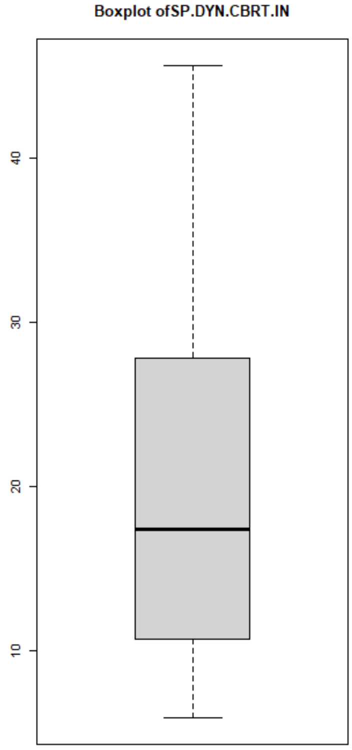

Histogram ofSH.XPD.CHEX.GD.ZS
Frequency
0
10
20
30
40
50
0
5
10
15
x
Boxplot ofSH.XPD.CHEX.GD.ZS
5
10
15
Life exp vs SH.XPD.CHEX.GD.ZS
Life expectancy
5
10
15
0
50
100
150
200
SH.XPD.CHEX.GD.ZS

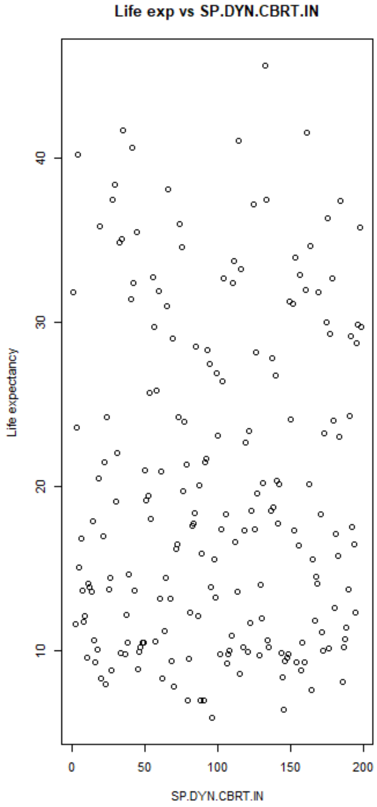

Histogram ofSH.XPD.CHEX.PC.CD
Frequency
0
20
40
60
80
100
120
0
2000
4000
6000
8000
10000
x

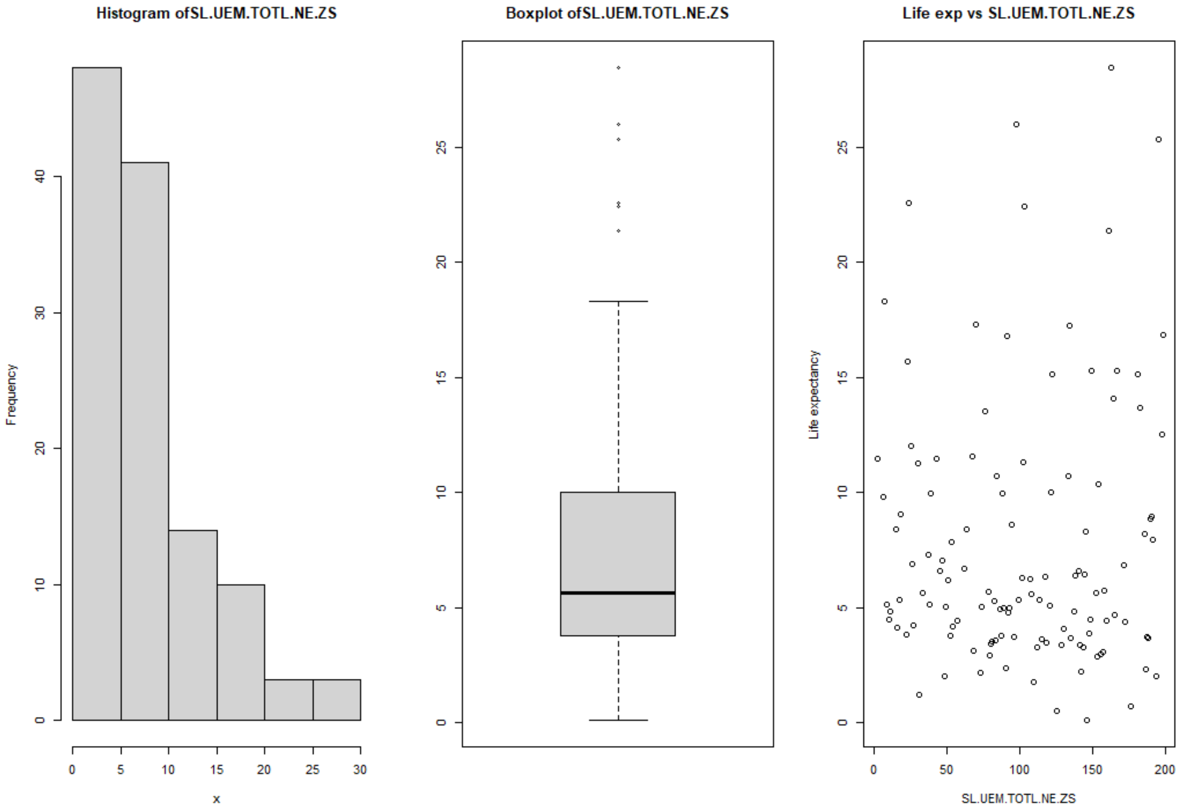

Boxplot ofSH.XPD.CHEX.PC.CD
0
2000
4000
6000
8000
10000

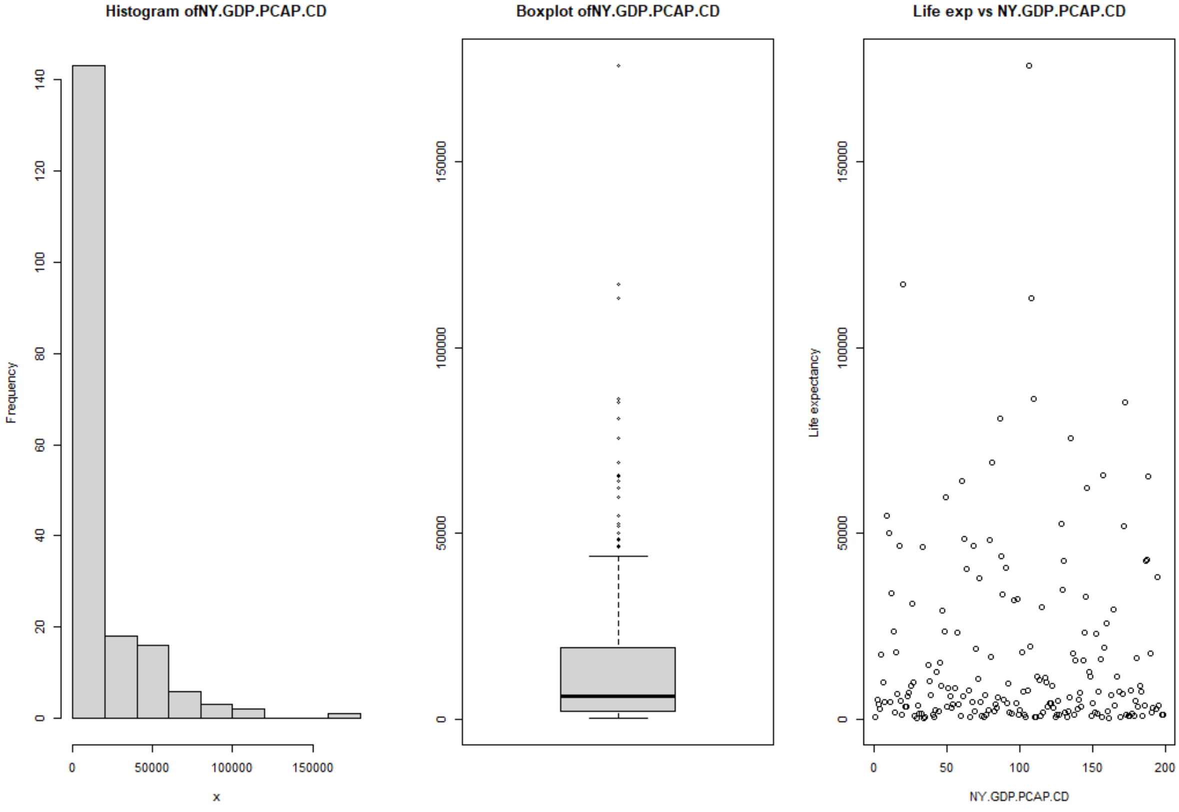

Life exp vs SH.XPD.CHEX.PC.CD
Life expectancy
0
2000
4000
6000
8000
10000
0
50
100
150
200
SH.XPD.CHEX.PC.CD

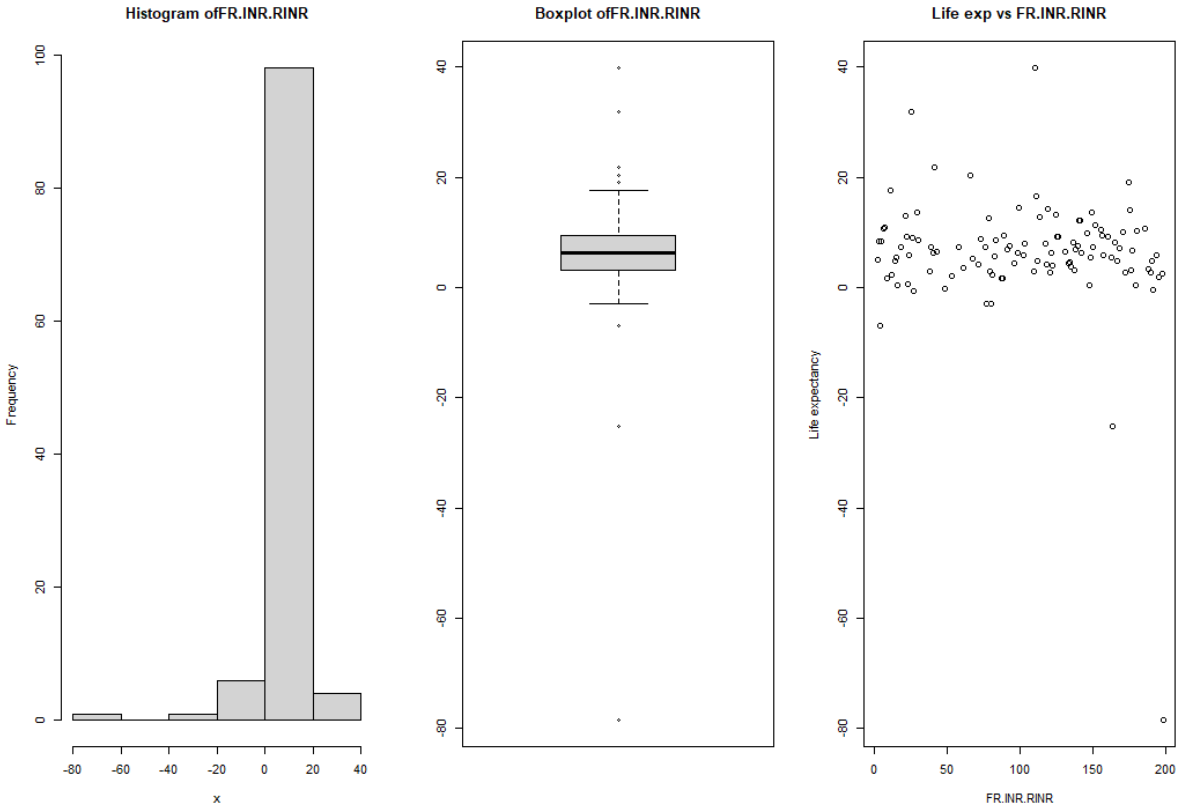

Histogram ofSP.DYN.CBRT.IN
Frequency
0
10
20
30
40
10
20
30
40
50
x

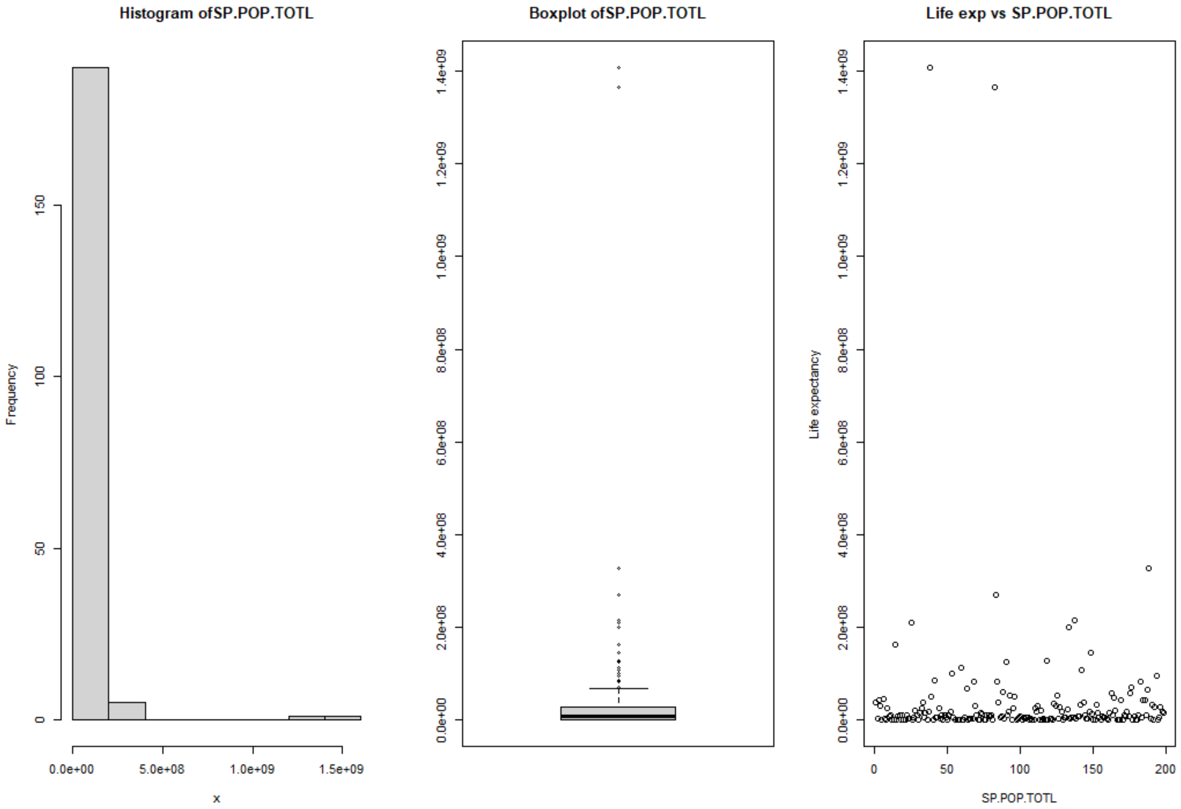

Boxplot ofSP.DYN.CBRT.IN
10
20
30
40

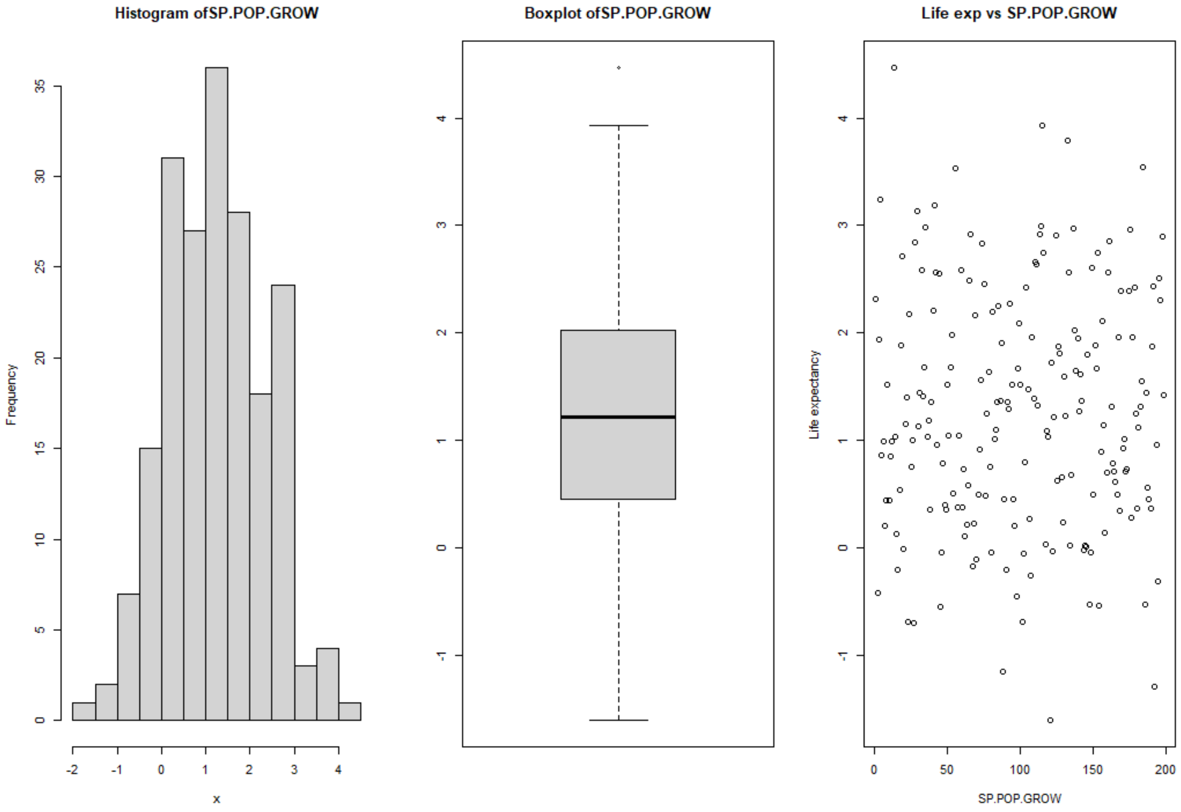

Life exp vs SP.DYN.CBRT.IN
Life expectancy
10
20
30
40
0
50
100
150
200
SP.DYN.CBRT.IN

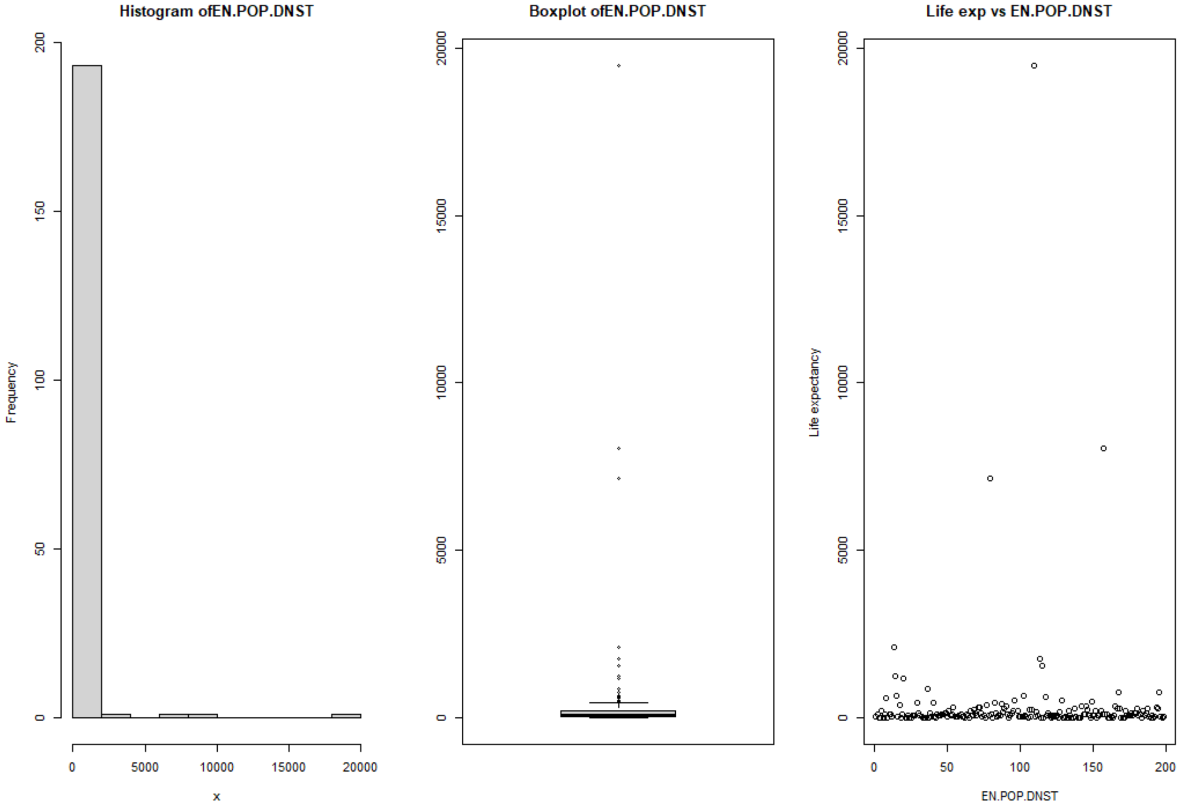

Histogram ofSL.UEM.TOTL.NE.ZS
Frequency
0
10
20
30
40
0
5
10
15
20
25
30
x
Boxplot ofSL.UEM.TOTL.NE.ZS
0
5
10
15
20
25
Life exp vs SL.UEM.TOTL.NE.ZS
Life expectancy
0
5
10
15
20
25
0
50
100
150
200
SL.UEM.TOTL.NE.ZS

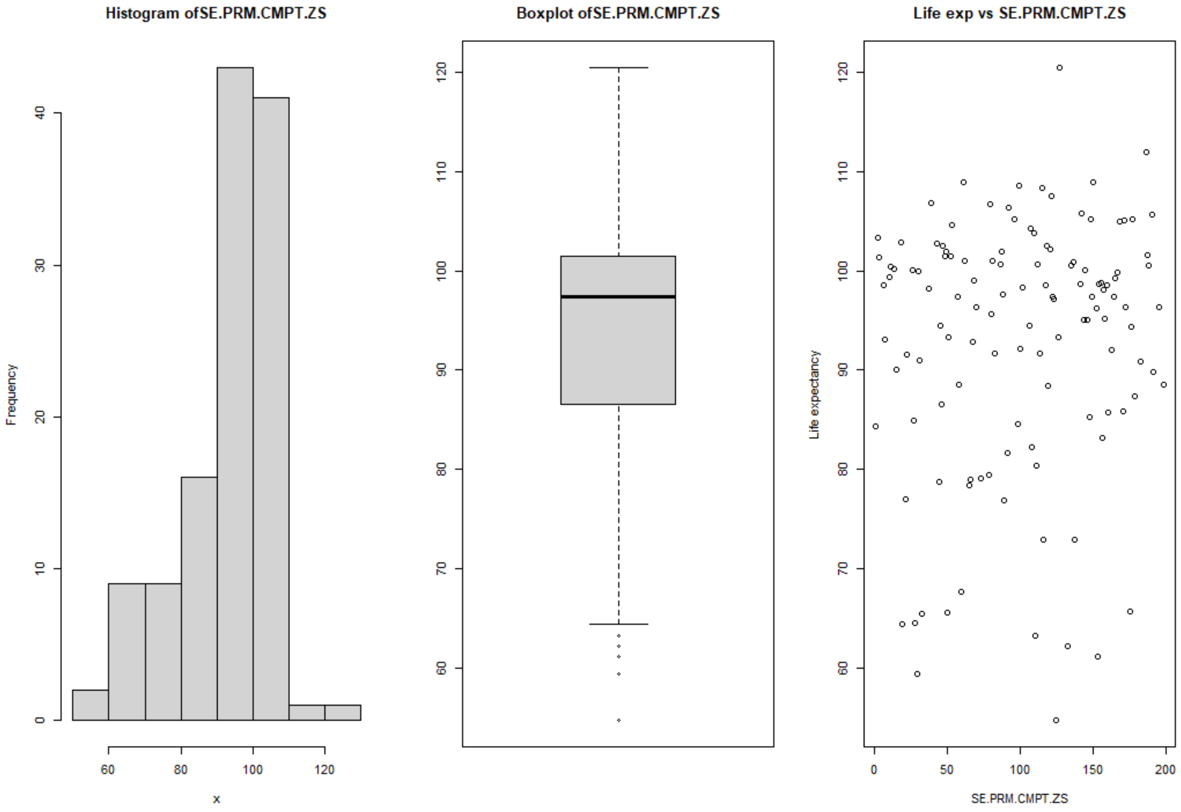

Histogram ofNY.GDP.PCAP.CD
Frequency
0
20
40
60
80
100
120
140
0
50000
100000
150000
x
Boxplot ofNY.GDP.PCAP.CD
0
50000
100000
150000
Life exp vs NY.GDP.PCAP.CD
Life expectancy
0
50000
100000
150000
0
50
100
150
200
NY.GDP.PCAP.CD

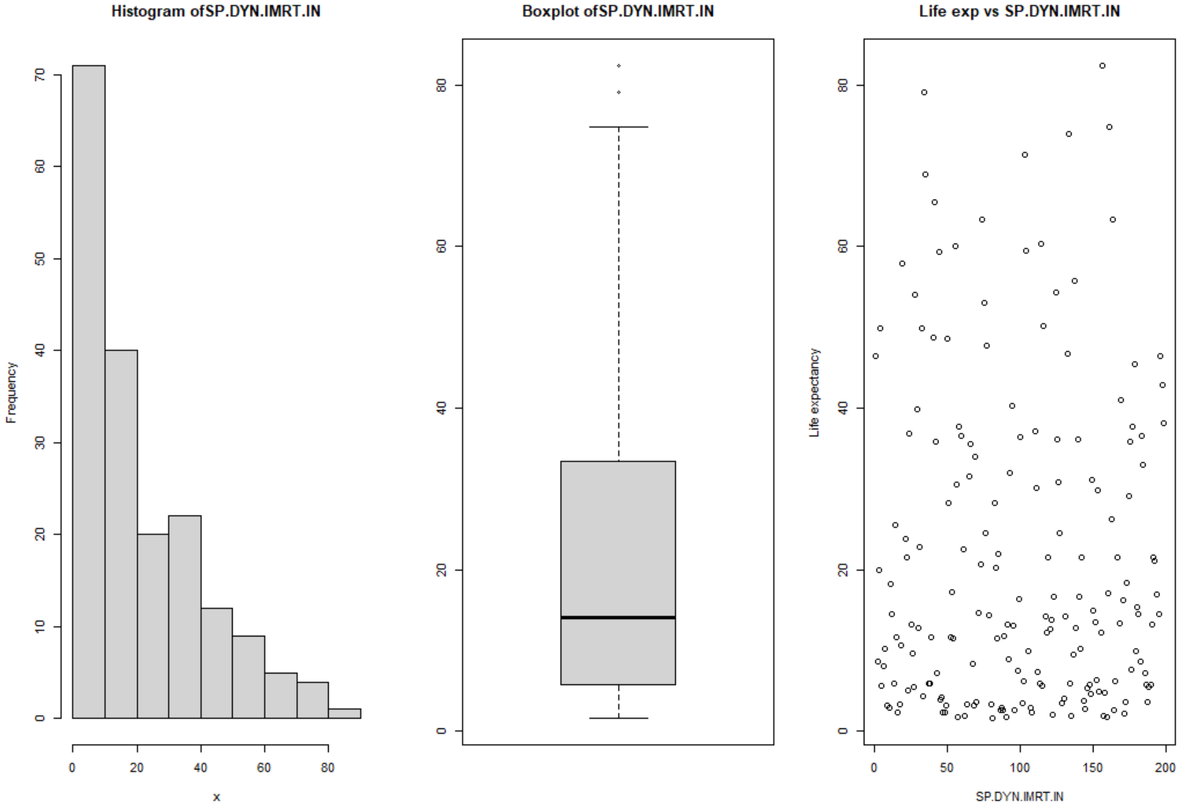

Histogram ofFR.INR.RINR
Frequency
0
20
40
60
80
100
-80
-60
-40
-20
0
20
40
x
Boxplot ofFR.INR.RINR
-80
-60
-40
-20
0
20
40
Life exp vs FR.INR.RINR
Life expectancy
-80
-60
-40
-20
0
20
40
0
50
100
150
200
FR.INR.RINR

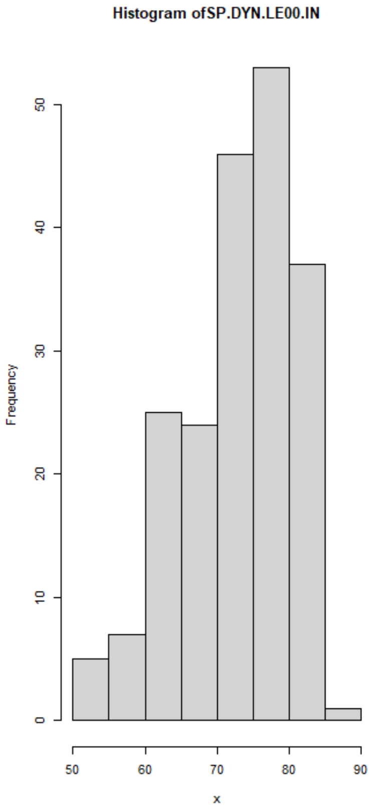

Histogram ofSP.POP.TOTL
Frequency
0
50
100
150
0.0e+00
5.0e+08
1.0e+09
1.5e+09
x
Boxplot ofSP.POP.TOTL
0.0e+00
2.0e+08
4.0e+08
6.0e+08
8.0e+08
1.0e+09
1.2e+09
1.4e+09
Life exp vs SP.POP.TOTL
Life expectancy
0.0e+00
2.0e+08
4.0e+08
6.0e+08
8.0e+08
1.0e+09
1.2e+09
1.4e+09
0
50
100
150
200
SP.POP.TOTL

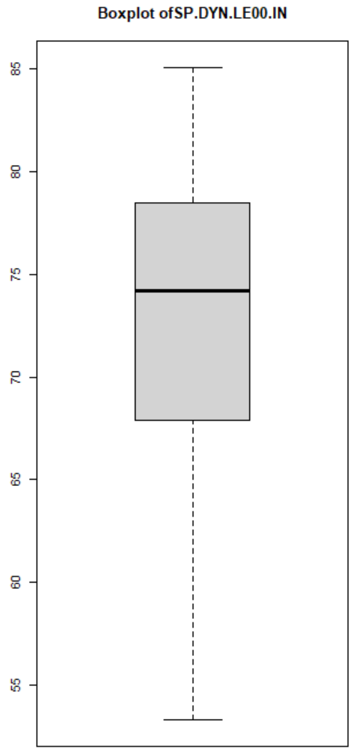

Histogram ofSP.POP.GROW
Frequency
0
5
10
15
20
25
30
35
x
-2
-1
0
1
2
3
4
Boxplot ofSP.POP.GROW
-1
0
1
2
3
4
Life exp vs SP.POP.GROW
Life expectancy
-1
0
1
2
3
4
SP.POP.GROW
0
50
100
150
200

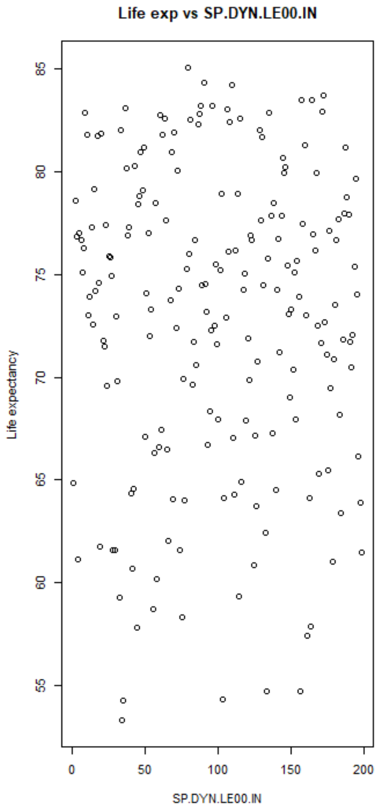

Histogram ofEN.POP.DNST
Frequency
0
50
100
150
200
x
0
5000
10000
15000
20000
Boxplot ofEN.POP.DNST
0
5000
10000
15000
20000
Life exp vs EN.POP.DNST
Life expectancy
0
5000
10000
15000
20000
EN.POP.DNST
0
50
100
150
200

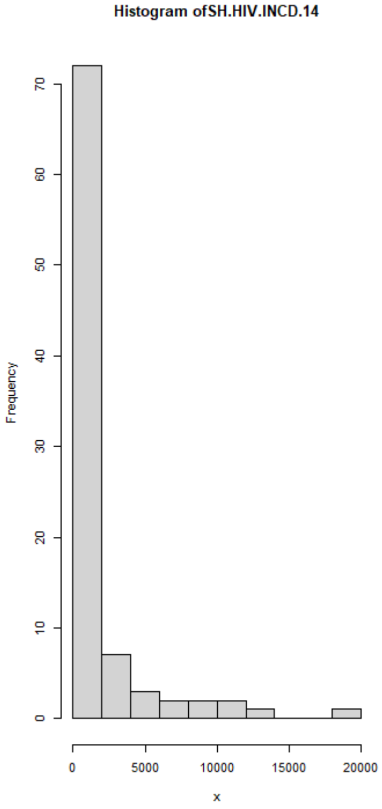

Histogram ofSE.PRM.CMPT.ZS
Frequency
0
10
20
30
40
60
80
100
120
x
Boxplot ofSE.PRM.CMPT.ZS
60
70
80
90
100
110
120
Life exp vs SE.PRM.CMPT.ZS
Life expectancy
60
70
80
90
100
110
120
0
50
100
150
200
SE.PRM.CMPT.ZS

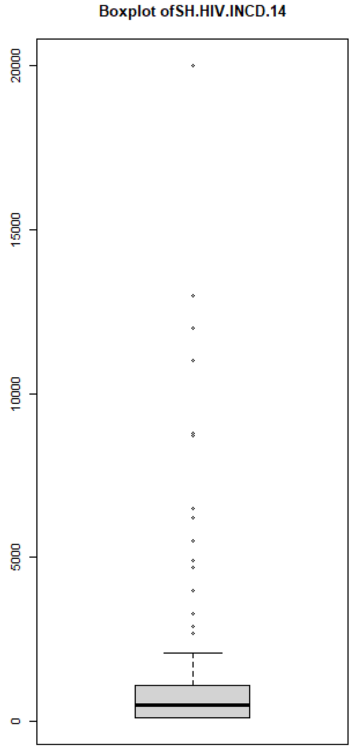

Histogram ofSP.DYN.IMRT.IN
Frequency
0
10
20
30
40
50
60
70
0
20
40
60
80
x
Boxplot ofSP.DYN.IMRT.IN
0
20
40
60
80
Life exp vs SP.DYN.IMRT.IN
Life expectancy
0
20
40
60
80
0
50
100
150
200
SP.DYN.IMRT.IN

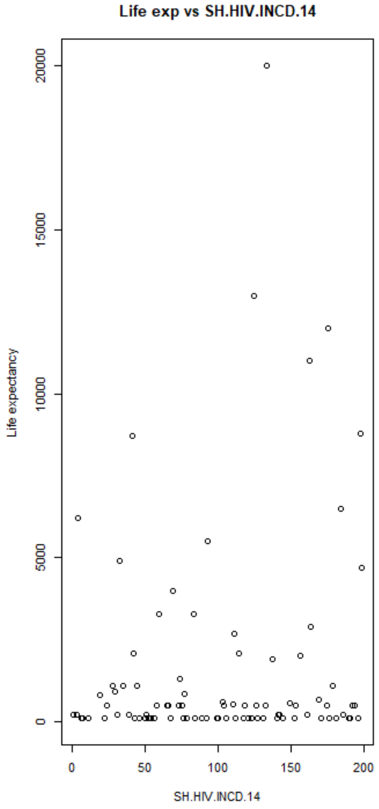

Histogram ofSP.DYN.LE00.IN
Frequency
0
10
20
30
40
50
50
60
70
80
90
x

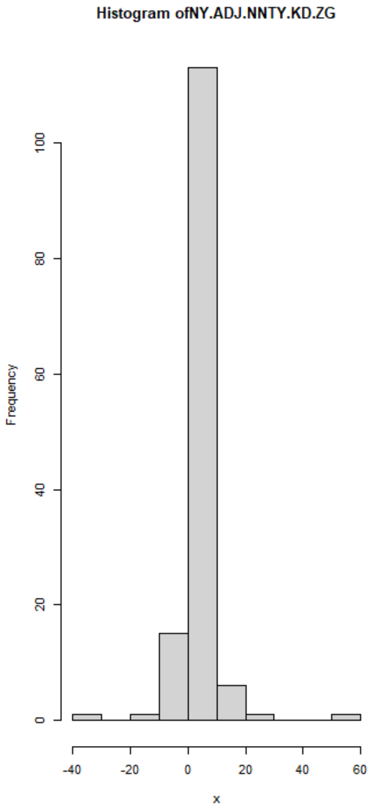

Boxplot ofSP.DYN.LE00.IN
55
60
65
70
75
80
85

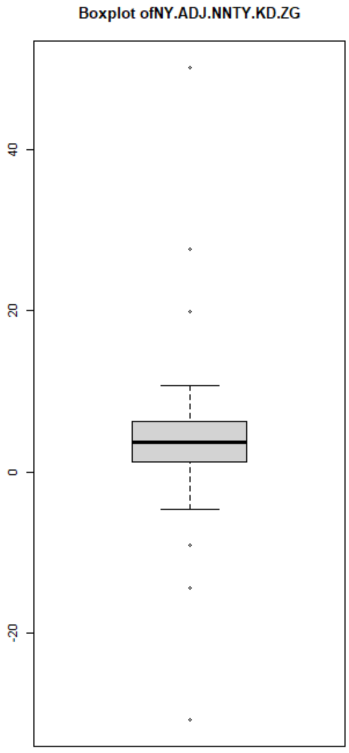

Life exp vs SP.DYN.LE00.IN
Life expectancy
55
60
65
70
75
80
85
0
50
100
150
200
SP.DYN.LE00.IN

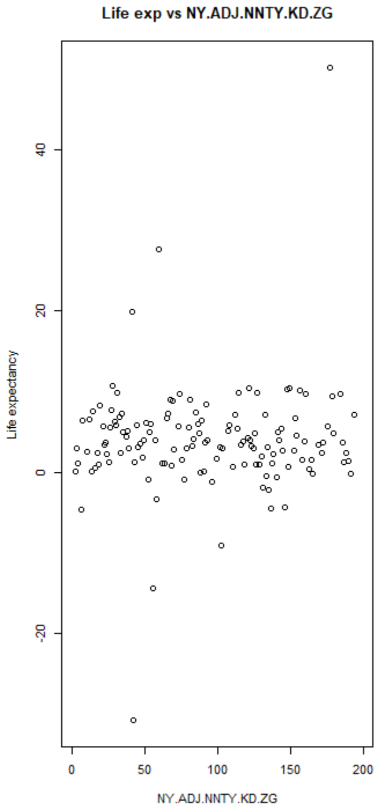

Histogram ofSH.HIV.INCD.14
Frequency
0
10
20
30
40
50
60
70
0
5000
10000
15000
20000
x

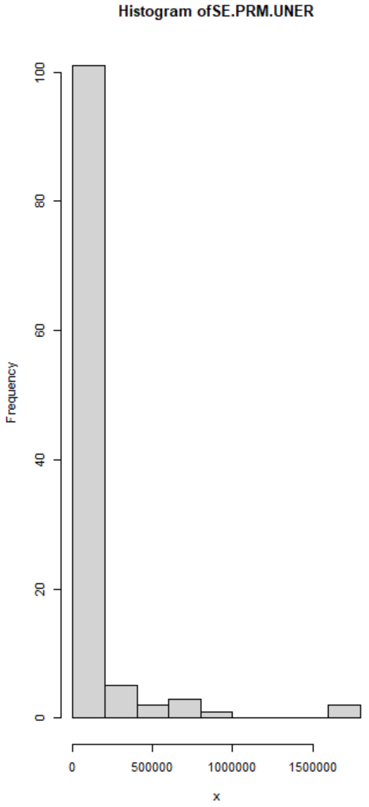

Boxplot ofSH.HIV.INCD.14
0
5000
10000
15000
20000

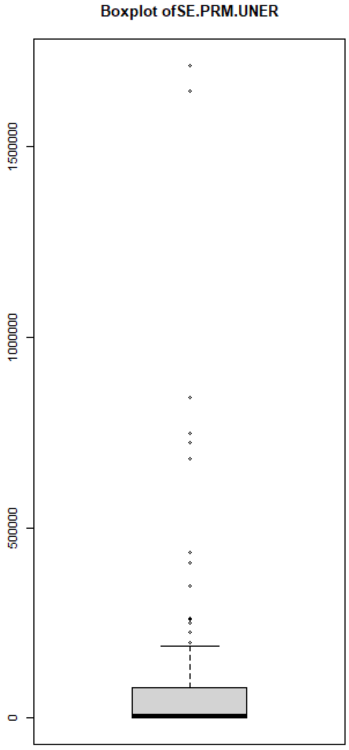

Life exp vs SH.HIV.INCD.14
Life expectancy
0
5000
10000
15000
20000
0
50
100
150
200
SH.HIV.INCD.14

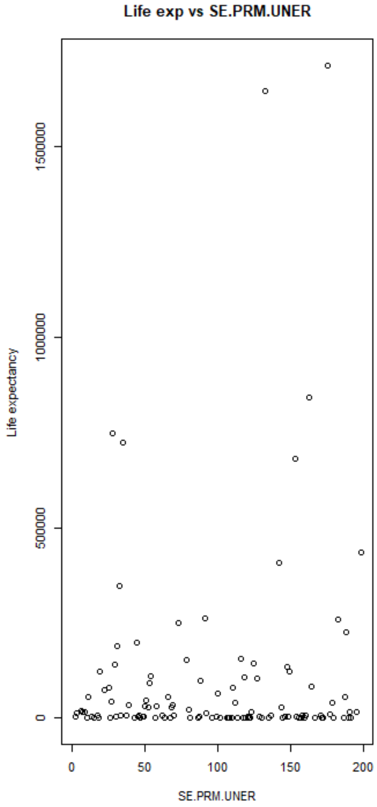

Histogram ofNY.ADJ.NNTY.KD.ZG

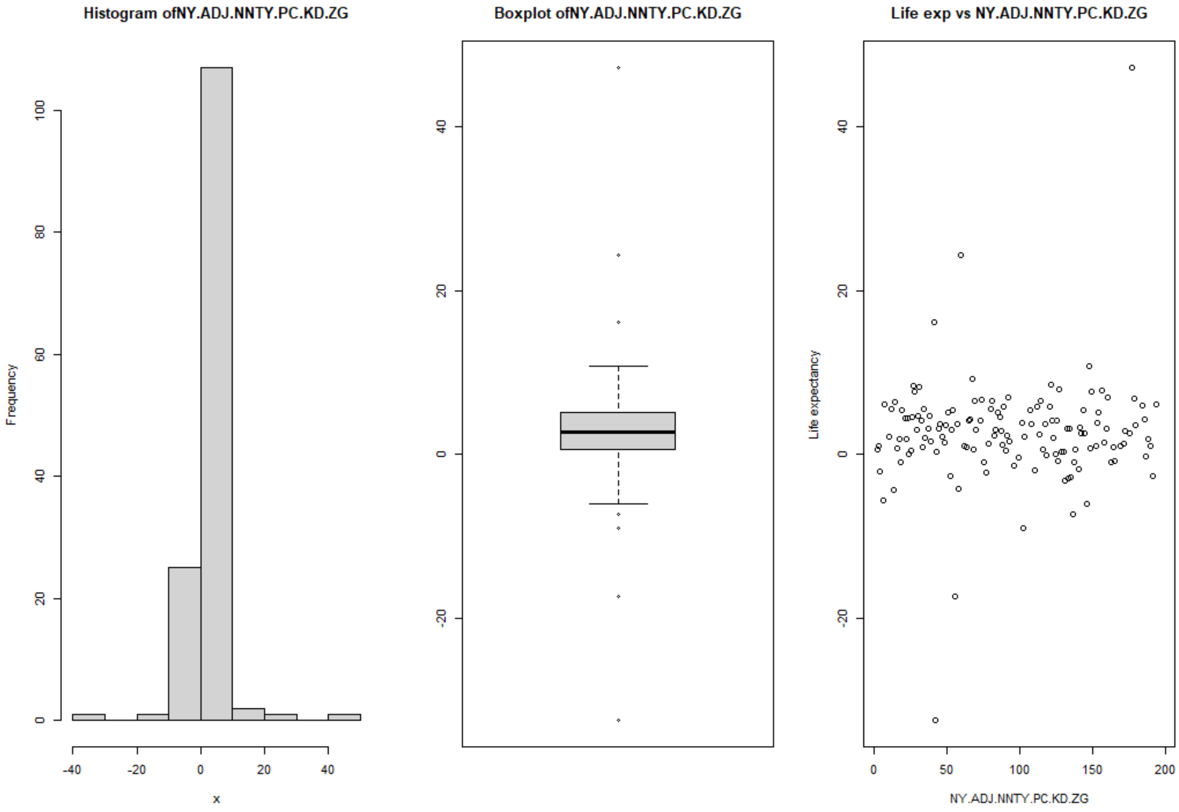

Boxplot ofNY.ADJ.NNTY.KD.ZG

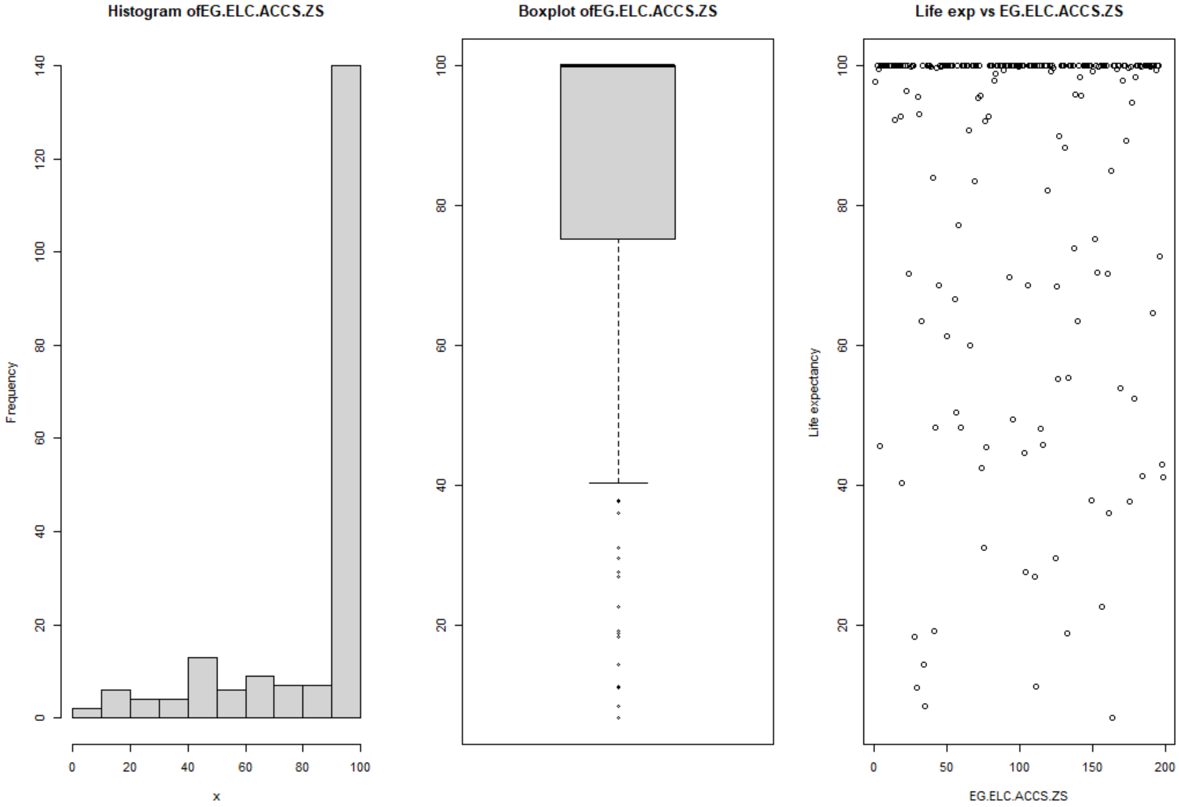

Life exp vs NY.ADJ.NNTY.KD.ZG

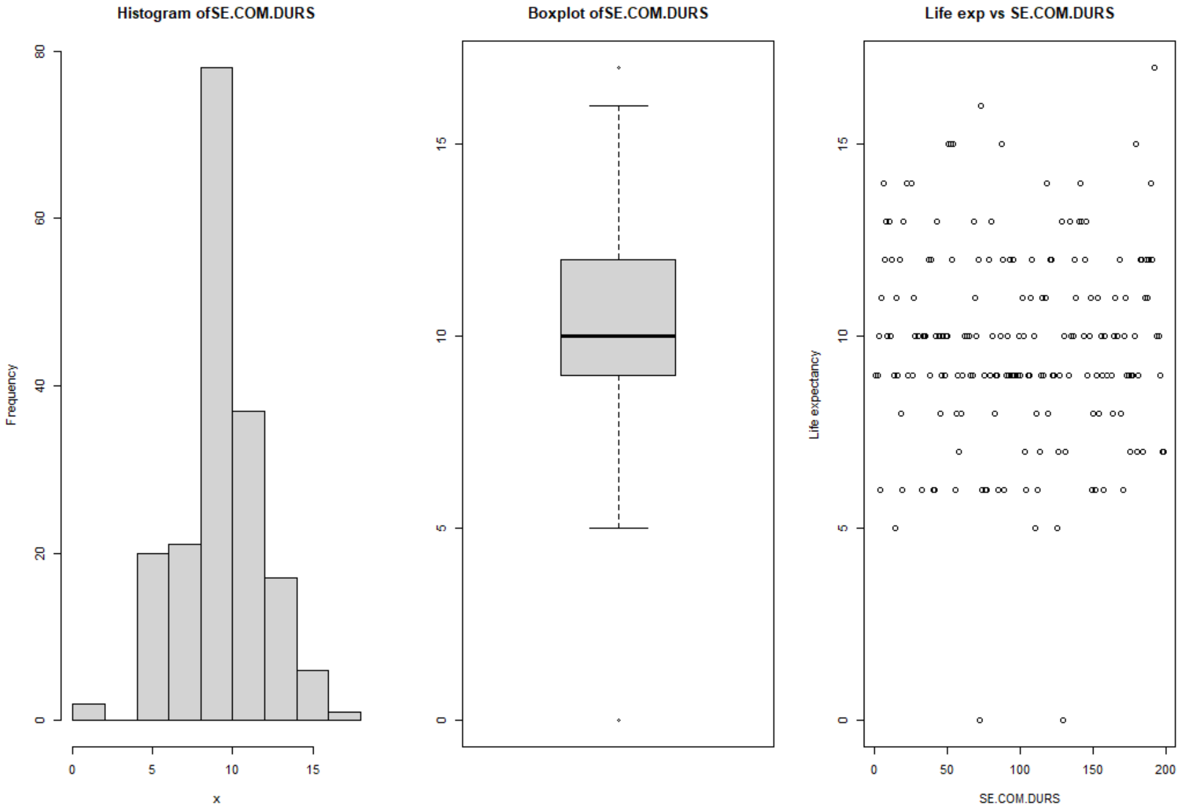

Histogram ofSE.PRM.UNER


Boxplot ofSE.PRM.UNER


Life exp vs SE.PRM.UNER


Histogram ofNY.ADJ.NNTY.PC.KD.ZG
Frequency
0
20
40
60
80
100
-40
-20
0
20
40
x
Boxplot ofNY.ADJ.NNTY.PC.KD.ZG
-20
0
20
40
Life exp vs NY.ADJ.NNTY.PC.KD.ZG
Life expectancy
-20
0
20
40
0
50
100
150
200
NY.ADJ.NNTY.PC.KD.ZG


Histogram ofEG.ELC.ACCS.ZS
Frequency
0
20
40
60
80
100
120
140
0
20
40
60
80
100
x
Boxplot ofEG.ELC.ACCS.ZS
20
40
60
80
100
Life exp vs EG.ELC.ACCS.ZS
Life expectancy
20
40
60
80
100
0
50
100
150
200
EG.ELC.ACCS.ZS



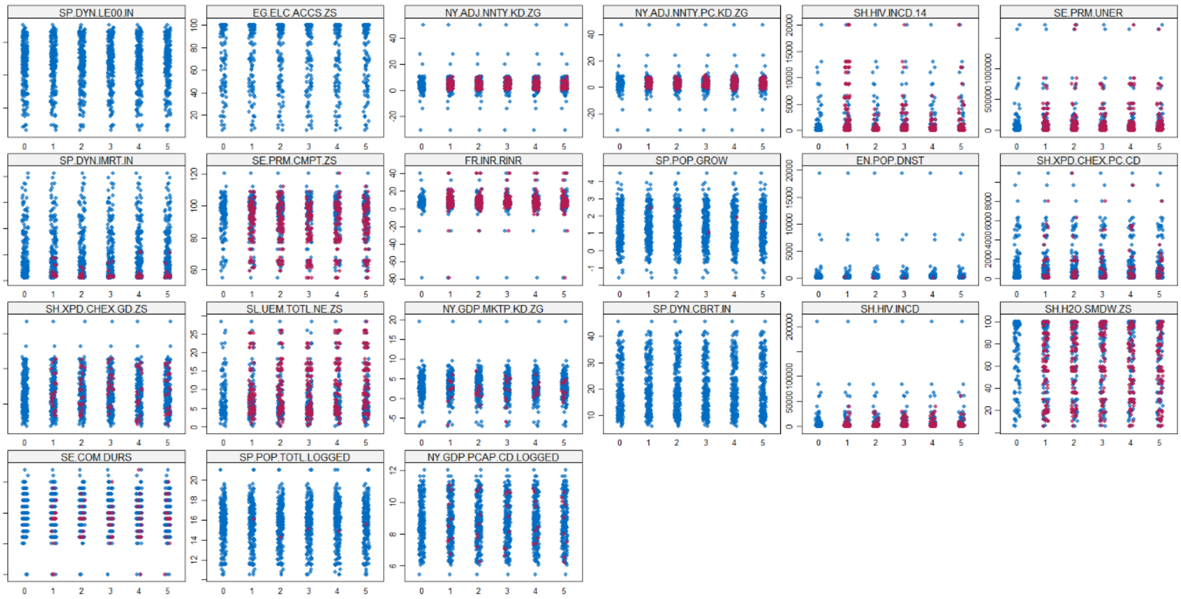
A3. Code for task Two

```
####************************************Question -2*************###################
##################################################------------ MICE
Implementation ----------------
########################################################

######----------------- For scaling the we used Log normalization or SP.POP.TOTL,
NY.GDP.PCAP.CD---------######
LifeExpectancy2$SP.POP.TOTL.LOGGED = log(LifeExpectancy2$SP.POP.TOTL)
LifeExpectancy2 <- LifeExpectancy2[!names(LifeExpectancy2) %in% c('SP.POP.TOTL')]

LifeExpectancy2$NY.GDP.PCAP.CD.LOGGED = log(LifeExpectancy2$NY.GDP.PCAP.CD)
LifeExpectancy2 <- LifeExpectancy2[!names(LifeExpectancy2) %in% c('NY.GDP.PCAP.CD')]
```

```
###----------------Using PMM method for MICE implementation----------------------####
LifeExpectancy2imputated <- mice(LifeExpectancy2,m=5,maxit=50,meth='pmm',seed=500)

LifeExpectancy3 <- complete(LifeExpectancy2imputated)
dim(LifeExpectancy3)


#-------------- Checking on the imputed values to diagnose any issues and how the imputed
values are aligned with the observed data.---------------#########
stripplot(LifeExpectancy2imputated, pch = 20, cex = 1.2)


####----------Checking the imputed values----------------########
LifeExpectancy2imputated$imp


####--------Building Linear Regression Model where SP.DYN.LE00.IN is the target and
considering all the columns of the dataset-------####

modelLifeExpectancy.fit <- with(LifeExpectancy2imputated, lm(SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS
+ NY.ADJ.NNTY.KD.ZG + NY.ADJ.NNTY.PC.KD.ZG
                                        + SH.HIV.INCD.14 + SE.PRM.UNER + SP.DYN.IMRT.IN +
SE.PRM.CMPT.ZS + FR.INR.RINR
                                        + SP.POP.GROW + EN.POP.DNST + SH.XPD.CHEX.PC.CD +
SH.XPD.CHEX.GD.ZS
                                        + SL.UEM.TOTL.NE.ZS + NY.GDP.MKTP.KD.ZG +
SP.DYN.CBRT.IN + SH.HIV.INCD
                                        + SH.H2O.SMDW.ZS + SE.COM.DURS +
SP.POP.TOTL.LOGGED + NY.GDP.PCAP.CD.LOGGED))
summary(modelLifeExpectancy.fit)
```

```
###-------------------------------Analyzing Model------------------------------###
Analyze = capture.output(modelLifeExpectancy.fit)
Analyze


#####--------------------------Pooling ------------------------------------####

LifeExpectancyPooled <- pool(modelLifeExpectancy.fit)
summary(LifeExpectancyPooled)


#####=============== From the result of the pooling model we can observe that there are few high negative values (e.g. - NY.ADJ.NNTY.PC.KD.ZG, SP.POP.GROW etc.)
# which can be considered as insignificant. Again we need to check the collinearity among the features.
```

A4. Figure for task two



Strip Plot for Imputed Values

A5. Code for Task Three

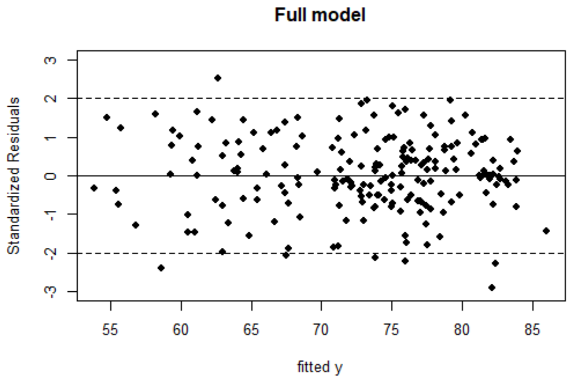
```
####*****************************************************Question - 3**********************************************###################
##################################################-------------VIF for Multi-Collinearity ------------################################################
X <- LifeExpectancy3
LifeExpectancy3.corr<-cor(X)
LifeExpectancy3.corr



corrplot.mixed(LifeExpectancy3.corr, lower.col = "black", number.cex = .7)



VIF_Values <- data.frame(vif(X))
VIF_Values
```

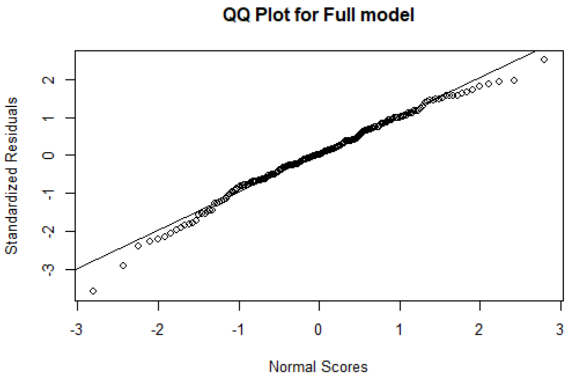
```
###--------------------- NY.ADJ.NNTY.KD.ZG   &&& NY.ADJ.NNTY.PC.KD.ZG ---------####
###--------------------- SP.DYN.IMRT.IN  &&& SP.DYN.CBRT.IN ---------####
LifeExpectancy3 <- LifeExpectancy3 %>% select(-'NY.ADJ.NNTY.KD.ZG')
LifeExpectancy3 <- LifeExpectancy3 %>% select(-'SP.DYN.CBRT.IN')
LifeExpectancy3 <- LifeExpectancy3 %>% select(-'NY.GDP.PCAP.CD.LOGGED')
LifeExpectancy3 <- LifeExpectancy3 %>% select(-'SP.DYN.IMRT.IN')


dim(LifeExpectancy3)



########-----Checking with ANOVA for SH.XPD.CHEX.PC.CD feature importance --------###
testModel1 <- lm(SP.DYN.LE00.IN ~ .-SH.XPD.CHEX.PC.CD, data = LifeExpectancy3)
testModel2 <- lm(SP.DYN.LE00.IN ~ ., data = LifeExpectancy3)


anova(testModel1,testModel2)
#####---------------- reject null hypothesis at 5 % significance level -----------###
LifeExpectancy3 <- LifeExpectancy3 %>% select(-'SH.XPD.CHEX.PC.CD')


##########################################----------- Model Selection with Wrapper
Method -------------------------####################################


names(LifeExpectancy3)


modelForward <-lm(SP.DYN.LE00.IN~1,data=LifeExpectancy3)
```

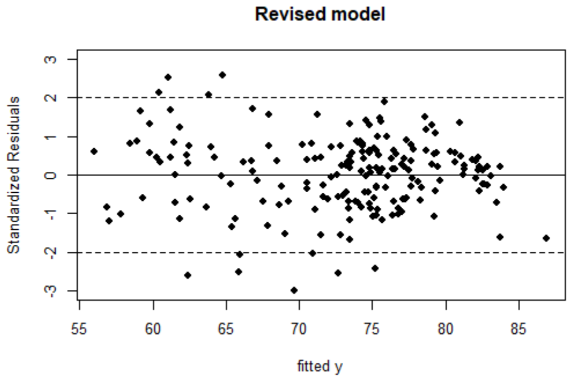
```
#####------------------------------------Forward Seletion --------------------###############
step1 <-step(modelForward,scope =~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.PC.KD.ZG +
SH.HIV.INCD.14 + SE.PRM.UNER + SE.PRM.CMPT.ZS +
           FR.INR.RINR + SP.POP.GROW + EN.POP.DNST + SH.XPD.CHEX.GD.ZS +
           SL.UEM.TOTL.NE.ZS + NY.GDP.MKTP.KD.ZG +
           SH.HIV.INCD + SH.H2O.SMDW.ZS + SE.COM.DURS + SP.POP.TOTL.LOGGED,
         method="forward")
summary(step1)


#####------------------------------------Backard Seletion --------------------###############
modelBackward<-lm(SP.DYN.LE00.IN ~ .,data=LifeExpectancy3)
step2<-step(modelBackward,method="backward")
summary(step2)



######=================== From both Forward and Backward Selection methods we
get the same set of columns ==================#######
#### AIC = 480.74



######-------------------------------------REVISED Model after removing collinearity--------------
-----###############



LifeExpectancy4 <- LifeExpectancy3[names(LifeExpectancy3) %in%
c('SP.DYN.LE00.IN','EG.ELC.ACCS.ZS', 'SH.HIV.INCD.14' , 'SE.PRM.UNER' , 'SE.PRM.CMPT.ZS' ,
                               'FR.INR.RINR' , 'EN.POP.DNST' , 'SH.XPD.CHEX.GD.ZS' ,
'SL.UEM.TOTL.NE.ZS',
                             'NY.GDP.MKTP.KD.ZG' ,'SH.HIV.INCD' ,'SH.H2O.SMDW.ZS')]



revisedModel <- lm(SP.DYN.LE00.IN ~ .,data = LifeExpectancy4)
summary(revisedModel)



####------------------ ANOVA Comparison between Fully-imputed Model and the
RevisedModel(after removing collinearity)--------#######
FullModelWithImputedValues <- lm(SP.DYN.LE00.IN ~ .,data = X)
anova(FullModelWithImputedValues)


anova(revisedModel, FullModelWithImputedValues)



#########----------------plots the standardised residuals against fitted values for FULL model---
---------------###########
x11(width = 20, height = 14)
par(mfrow=c(2,2))


stdres_fullmodel<-rstandard(FullModelWithImputedValues)
plot(FullModelWithImputedValues$fitted.values,stdres_fullmodel,pch=16,
     ylab="Standardized Residuals",xlab="fitted y",ylim=c(-3,3),main="Full model")
abline(h=0)
abline(h=2,lty=2)
abline(h=-2,lty=2)
qqnorm(stdres_fullmodel, ylab="Standardized Residuals",
      xlab="Normal Scores", main="QQ Plot for Full model" )
qqline(stdres_fullmodel)



########----------------plots the standardised residuals against fitted values for REDUCED
model--------------#############
stdres_reducedmodel<-rstandard(revisedModel)
plot(revisedModel$fitted.values,stdres_reducedmodel,pch=16,ylab="Standardized
Residuals",xlab="fitted y", ylim=c(-3,3),main="Revised model")
abline(h=0)
abline(h=2,lty=2)
abline(h=-2,lty=2)
qqnorm(stdres_reducedmodel, ylab="Standardized Residuals",
xlab="Normal Scores", main="QQ Plot for Revised model")
qqline(stdres_reducedmodel)




#########################-----------------Full plot of both models-------------------
############################



###-----------Full Model --------###
x11(width = 20, height = 14)
par(mfrow=c(2,2))
plot(FullModelWithImputedValues)
```

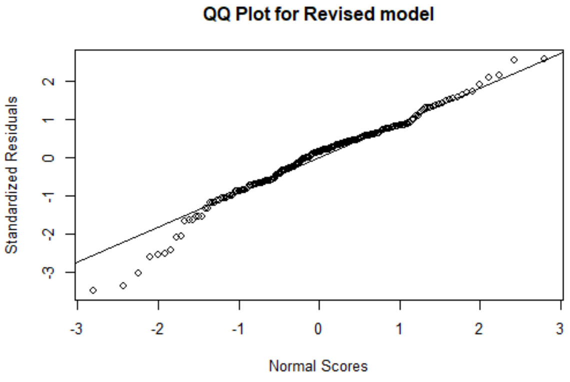
```
###-----------Revised Model --------###
x11(width = 20, height = 14)
par(mfrow=c(2,2))
plot(revisedModel)
```

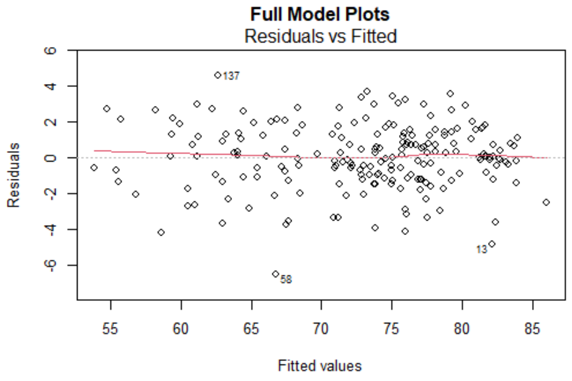
A6. Figure for task three

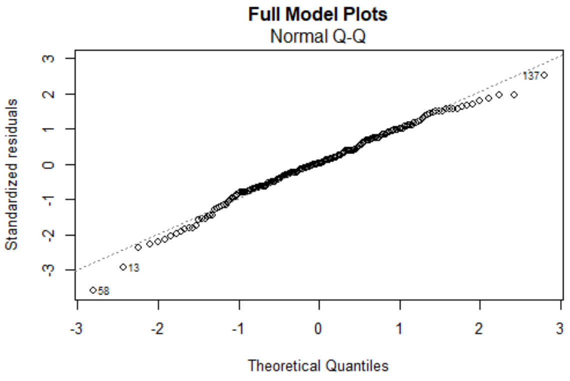

Full model

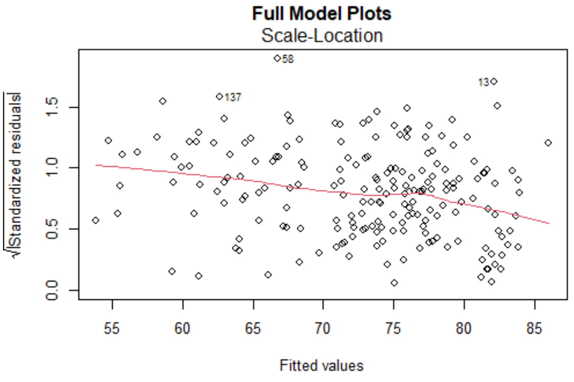

QQ Plot for Full model

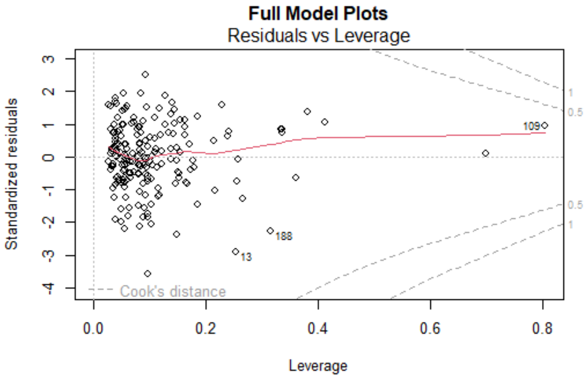

Revised model

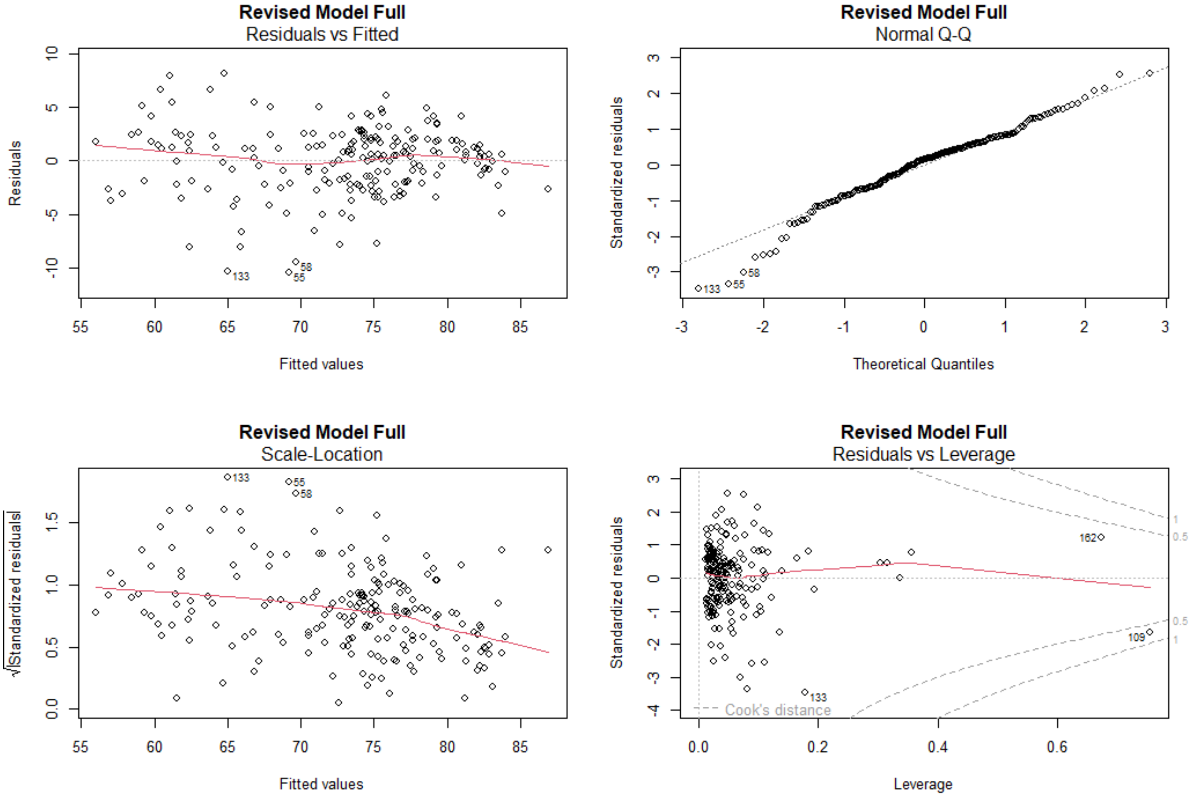

QQ Plot for Revised model


Full Model Plots
Residuals vs Fitted


Full Model Plots
Normal Q-Q


Full Model Plots
Scale-Location


Full Model Plots
Residuals vs Leverage



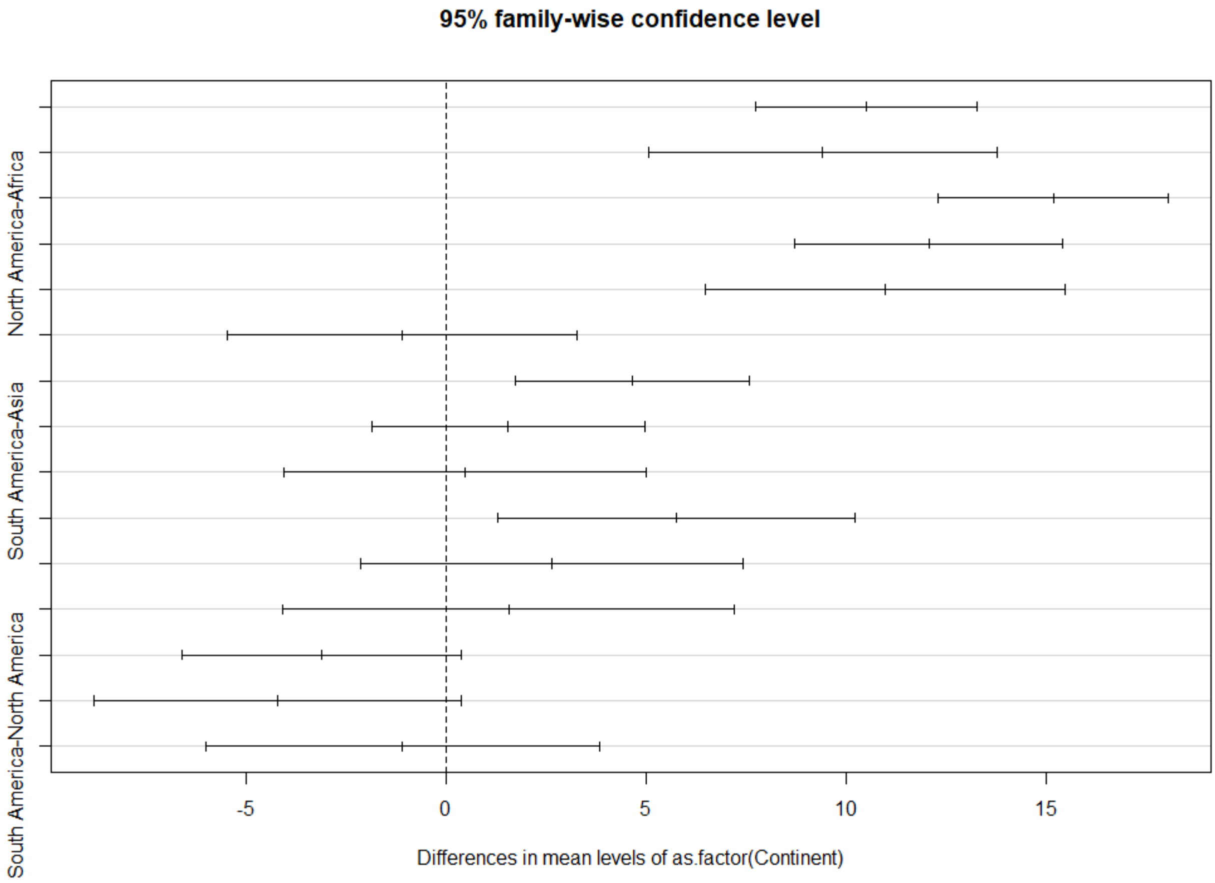
A7. Code for task four

```
####*******************************************************Question -
4************************************************###################
#######--====================================ONE-Way ANOVA for
analyzing Life Expectancy-=============================#################



#dim(LifeExpectancy4)
#names(LifeExpectancy4)



continentNames <- c(life_expectancy_filtered['Continent'])
```

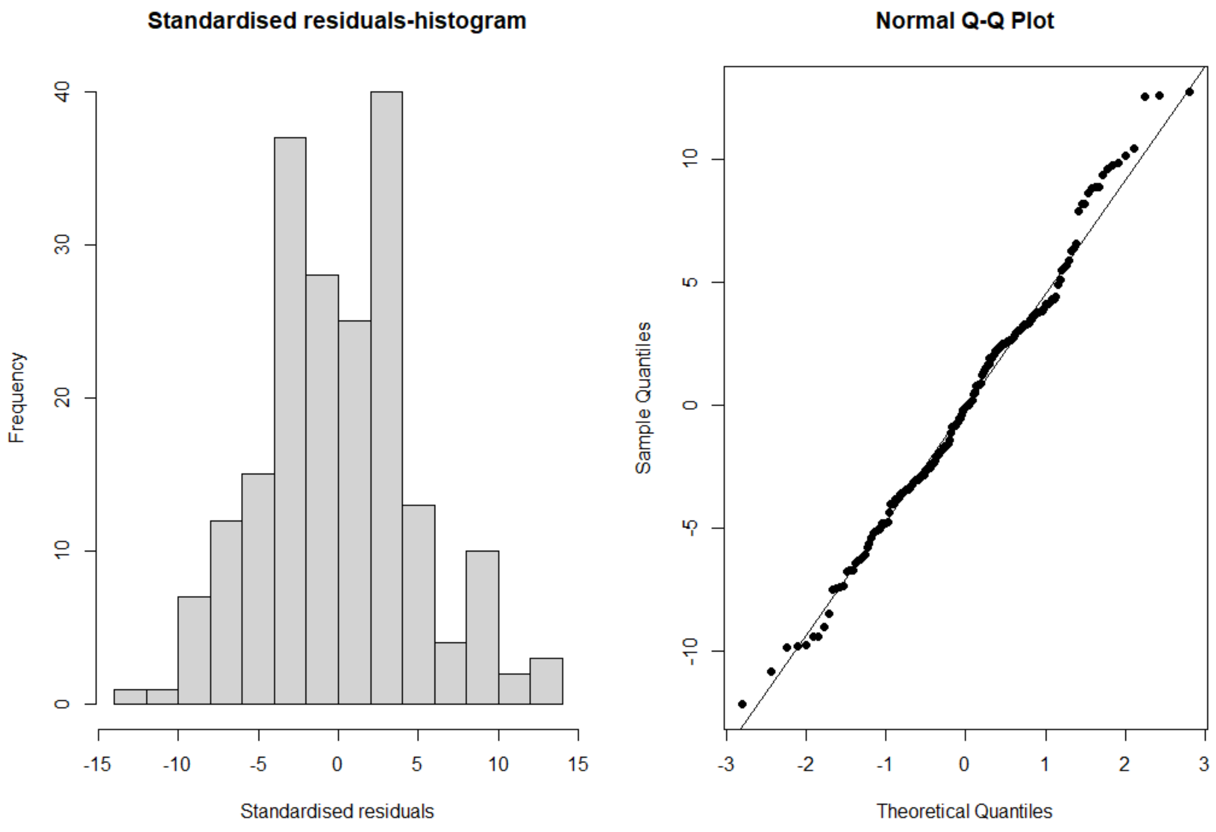
```
###--------------------Adding Continent to the Dataset for Final ANOVA ---------------------------
---##################
LifeExpectancy5 <- cbind(LifeExpectancy4, continentNames)
names(LifeExpectancy5)



###------------------Checking Mean LifeExpectancy for the corresponding Continents-------------
-----------------############################
continent.means <- tapply(LifeExpectancy5$SP.DYN.LE00.IN,LifeExpectancy5$Continent,mean)
continent.means



x11(width = 20, height = 14)
boxplot(LifeExpectancy5$SP.DYN.LE00.IN ~ LifeExpectancy5$Continent,main='Comparing the Life
Expectancyies for different continents',
        xlab='Continent', col="light gray", ylab = "Avg Life Expectancy",)



####------------------- ANOVA Test--------------------------------------------------
#########################
anovaLifeExpectancy<-aov(SP.DYN.LE00.IN ~ as.factor(Continent),data=LifeExpectancy5)
summary(anovaLifeExpectancy)



########-----------------Post HOC Test(For conducting pairwise Comparison)---------------
###################



cat("Conducting Bonferroni post-hoc test for Pairwise Comparison","\n")
```

```
pairwise.t.test(LifeExpectancy5$SP.DYN.LE00.IN, LifeExpectancy5$Continent, p.adj = "bonferroni")


cat("Conducting Tukey post-hoc Test","\n")
tukey.LifeExpectancy<-TukeyHSD(anovaLifeExpectancy)
x11(width = 20, height = 14)
plot(tukey.LifeExpectancy)



######--------------------ANOVA Assumptions checking ---------------------
###########################


LifeExpectancy5$residualsANOVA<-anovaLifeExpectancy$residuals
x11(width = 20, height = 14)
par(mfrow=c(1,2))
hist(LifeExpectancy5$residualsANOVA, main="Standardised residuals-
histogram",xlab="Standardised residuals")
qqnorm(LifeExpectancy5$residualsANOVA,pch=19)
qqline(LifeExpectancy5$residualsANOVA)



shapiro.test(LifeExpectancy5$residualsANOVA)


leveneTest(LifeExpectancy5$SP.DYN.LE00.IN ~ factor(LifeExpectancy5$Continent))
LifeExpectancy.welchtest<-oneway.test(SP.DYN.LE00.IN ~ Continent,data=LifeExpectancy5)
LifeExpectancy.welchtest
```

A8. Figure for task four



Tukey's Post Hoc test



Normality Test for ANOVA Assumption

References:

1. https://www.rdocumentation.org/packages/mice/versions/3.14.0/topics/mice
2. https://search.r-project.org/CRAN/refmans/mice/html/stripplot.mids.html
3. https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
4. Lecture Notes and Slides.