Statistical Analysis of Life Expectancy DataSet

Group 3

24/11/2021

Submitted by :

- Avinandan Patel(215280043)
- Syliva Vincent(215280005)
- Vishakha Verma(215280019)
- Inderesh Singh(215280011)
- Raj Khandagale(215280002)
- Harsh jaiswal(215280022)

Life Expectancy : **Exploratory Data Analysis**

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years , there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years

Goal : **Find a set of features that affect Life Expectancy**.

Contents :

1. **Data Cleaning** :

- Detect and Deal with the Missing values
- Detect the distribution of each factor

2. **Data Exploration and Visualization** :

- What is the Life Expectancy Country-wise
- How different dieseases affect life expectancy in developed and developing countries
- What effect does Schooling and Alcohol have on Life Expectancy

3. **Fitting a multiple regression model**

- Build a Base Model 1
- Plotting The Model1
- Multicollinearity Test

- Build Improved Model 2
- Checking Interaction Terms
- Build Improved Model 3
- Plotting the Model 3
- Fixing Polynomial Terms
- Build Improved Model 4
- Plotting The Model 4
- Conclusion of The Final Model

4. Summary :

Importing the dataset

```
data <- read.csv("E:/CN python/Life Expectancy Data.csv")
```

Section 1 : Data Cleaning

In order to properly clean the data, it is important to understand the variables presented in the data. There are a number of things important to know about each variable:

- What does the variable mean and what type of variable is it (Nominal/Ordinal/Interval/Ratio)?
- Does the variable have missing values? If so, what should be done about them?

Description about the dataset :

This collection is made up of data collected by the World Health Organization from various nations all over the world (WHO for short). The information is a compilation of several indicators for a certain nation and year. In essence, the data is a time series of several metrics divided by nation.

The string values for the columns/variables themselves aren't particularly 'clean,' so here's a little cleaning of the column/variable titles before we go into the variable descriptions.

```
dim(data)
```

```
## [1] 2938   22
```

```
View(data)
```

Number of sample points : = 2938

Number of Variables : = 22

Name of the Variables :

```
colnames(data)
```

```
##  [1] "Country"                        "Year"
##  [3] "Status"                         "Life.expectancy"
##  [5] "Adult.Mortality"                "infant.deaths"
##  [7] "Alcohol"                        "percentage.expenditure"
##  [9] "Hepatitis.B"                    "Measles"
## [11] "BMI"                            "under.five.deaths"
## [13] "Polio"                          "Total.expenditure"
## [15] "Diphtheria"                     "HIV.AIDS"
## [17] "GDP"                            "Population"
## [19] "thinness..1.19.years"           "thinness.5.9.years"
## [21] "Income.composition.of.resources" "Schooling"
```

Variable Descriptions :

Nominal Variable :

- **Country** : The country in which the indicators are from (i.e. United States of America or Congo
- **Status** : Whether a country is considered to be 'Developing' or 'Developed' by WHO standards

Ordinal Variable :

- **Year** : The calendar year the indicators are from (ranging from 2000 to 2015)
- **BMI** Average Body Mass Index (BMI) of a country's total population

Ratio :

- **Life Expectancy** : The life expectancy of people in years for a particular country and year
- **Adult Mortality** : The adult mortality rate per 1000 population (i.e. number of people dying between 15 and 60 years per 1000 population); if the rate is 263 then that means 263 people will die out of 1000 between the ages of 15 and 60; another way to think of this is that the chance an individual will die between 15 and 60 is 26.3%
- **Infant Deaths** : Number of infant deaths per 1000 population; similar to above, but for infants
- **Alcohol** : A country's alcohol consumption rate measured as liters of pure alcohol consumption per capita
- **Percentage Expenditure** : Expenditure on health as a percentage of Gross Domestic Product (gdp)
- **Hepatitis b** : Number of 1 year olds with Hepatitis B immunization over all 1 year olds in population
- **Measles** : Number of reported Measles cases per 1000 population
- **Under five deaths** : Number of people under the age of five deaths per 1000 population
- **Polio** : Number of 1 year olds with Polio immunization over the number of all 1 year olds in population
- **Total Expenditure** : Government expenditure on health as a percentage of total government expenditure
- **Diphtheria** : Diphtheria tetanus toxoid and pertussis (DTP3) immunization rate of 1 year olds
- **Hiv/Aids** : - deaths per 1000 live births caused by HIV/AIDS for people under 5; number of people under 5 who die due to HIV/AIDS per 1000 births
- **Gdp**- Gross Domestic Product per capita
- **population** :- population of a country
- **thinness 1 19 years** rate of thinness among people aged 10-19 (Note: variable should be renamed to thinness_10-19_years to more accurately represent the variable)
- **thinness 5 9 years**rate of thinness among people aged 5-9
- **income composition of resources** Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- **Schooling** average number of years of schooling of a population

Summary of the data

```
summary(data)
```

```
##    Country               Year         Status          Life.expectancy
##  Length:2938        Min.   :2000   Length:2938        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10
##  Mode  :character   Median :2008   Mode  :character   Median :72.10
##                     Mean   :2008                      Mean   :69.22
##                     3rd Qu.:2012                      3rd Qu.:75.70
##                     Max.   :2015                      Max.   :89.00
##                                                       NA's   :10
##  Adult.Mortality infant.deaths       Alcohol        percentage.expenditure
##  Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000
##  1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685
##  Median :144.0   Median :   3.0   Median : 3.7550   Median :   64.913
##  Mean   :164.8   Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251
##  3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534
##  Max.   :723.0   Max.   :1800.0   Max.   :17.8700   Max.   :19479.912
##  NA's   :10                       NA's   :194
##   Hepatitis.B       Measles              BMI        under.five.deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
##  Median :92.00   Median :    17.0   Median :43.50   Median :   4.00
##  Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.04
##  3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##  NA's   :553                        NA's   :34
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.755    Median :93.00   Median : 0.100
##  Mean   :82.55   Mean   : 5.938    Mean   :82.32   Mean   : 1.742
##  3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600
##  NA's   :19      NA's   :226       NA's   :19
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.10
##  1st Qu.:   463.94   1st Qu.:1.958e+05   1st Qu.: 1.60
##  Median :  1766.95   Median :1.387e+06   Median : 3.30
##  Mean   :  7483.16   Mean   :1.275e+07   Mean   : 4.84
##  3rd Qu.:  5910.81   3rd Qu.:7.420e+06   3rd Qu.: 7.20
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.70
##  NA's   :448         NA's   :652         NA's   :34
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.10      Min.   :0.0000                  Min.   : 0.00
##  1st Qu.: 1.50      1st Qu.:0.4930                  1st Qu.:10.10
##  Median : 3.30      Median :0.6770                  Median :12.30
##  Mean   : 4.87      Mean   :0.6276                  Mean   :11.99
##  3rd Qu.: 7.20      3rd Qu.:0.7790                  3rd Qu.:14.30
##  Max.   :28.60      Max.   :0.9480                  Max.   :20.70
##  NA's   :34         NA's   :167                     NA's   :163
```

Missing values

There are few things that must be done concerning missing values:

- Detection of missing values Find nulls Could a null be signified by anything other than null? Zero values perhaps?
- Dealing with missing values Fill nulls? Impute or Interpolate Eliminate nulls?

Missing Values Detection

The simplest and quickest way here is to do a quick is.na() and sum(is.na()) provides total number of missing values.

```
mydata <- data
sum(is.na(mydata))
```

```
## [1] 2563
```

There are 2563 number of Missing values in our model so removing of these values from our dataset means loosing of many information of data from our model. so instead of dropping we do imputation at the place of missing values.

```
df1=data
for(i in 1:ncol(df1)){
  df1[is.na(df1[,i]),i] <- median(df1[,i] , na.rm = T)
}
```

```
## Warning in mean.default(sort(x, partial = half + 0L:1L)[half + 0L:1L]): argument
## is not numeric or logical: returning NA

## Warning in mean.default(sort(x, partial = half + 0L:1L)[half + 0L:1L]): argument
## is not numeric or logical: returning NA
```

```
data=df1
sum(is.na(df1))
```

```
## [1] 0
```

```
mydata=df1
```

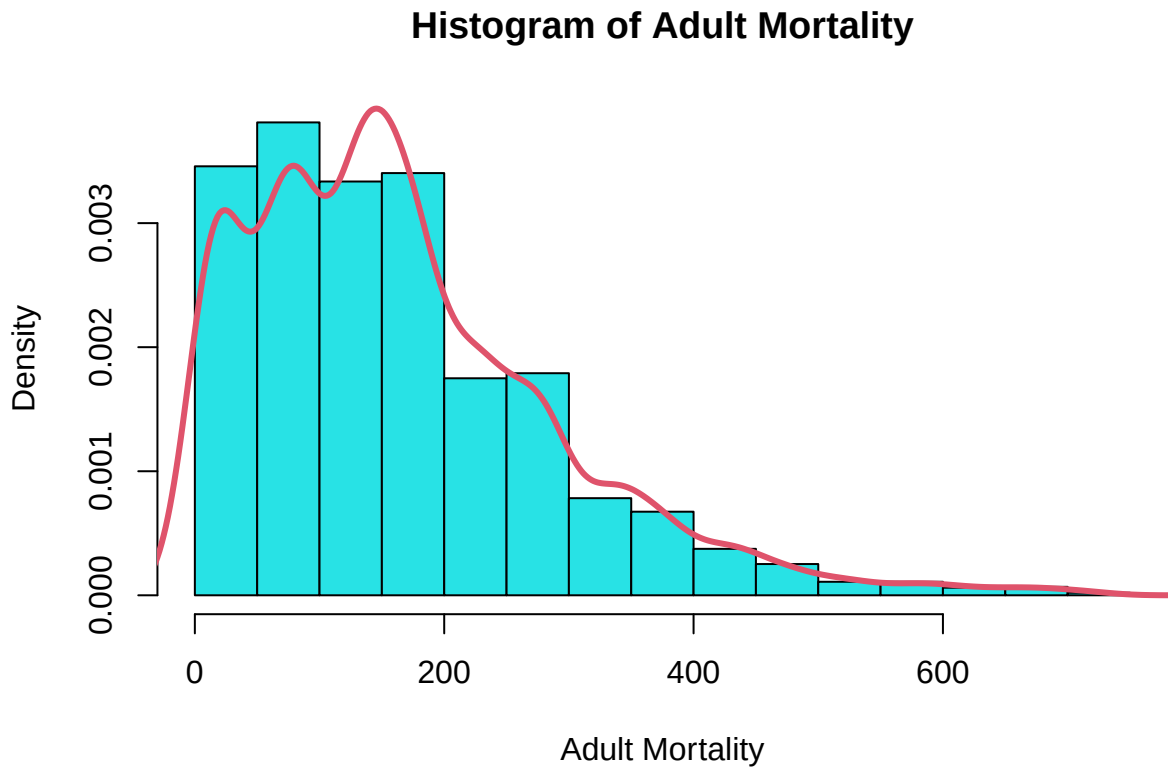
```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.0.5

## Warning: package 'tibble' was built under R version 4.0.5
```

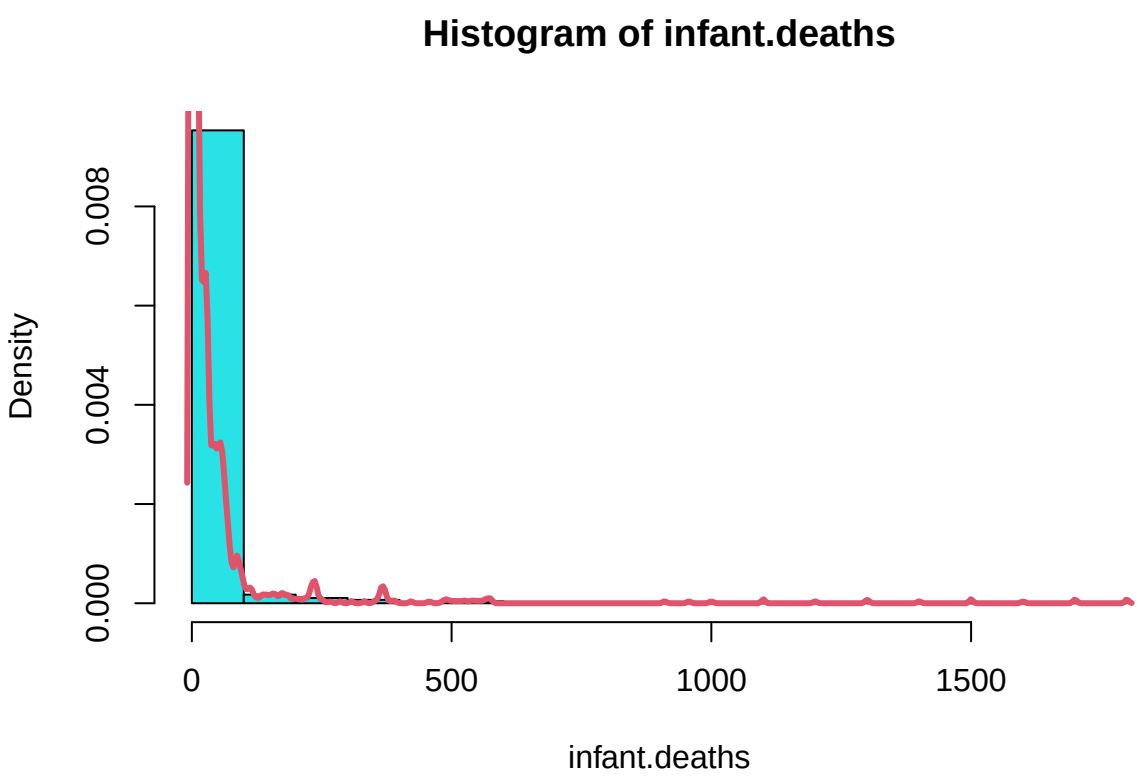
```
## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'purrr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## Warning: package 'stringr' was built under R version 4.0.5

## Warning: package 'forcats' was built under R version 4.0.5

## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data Visualisation :

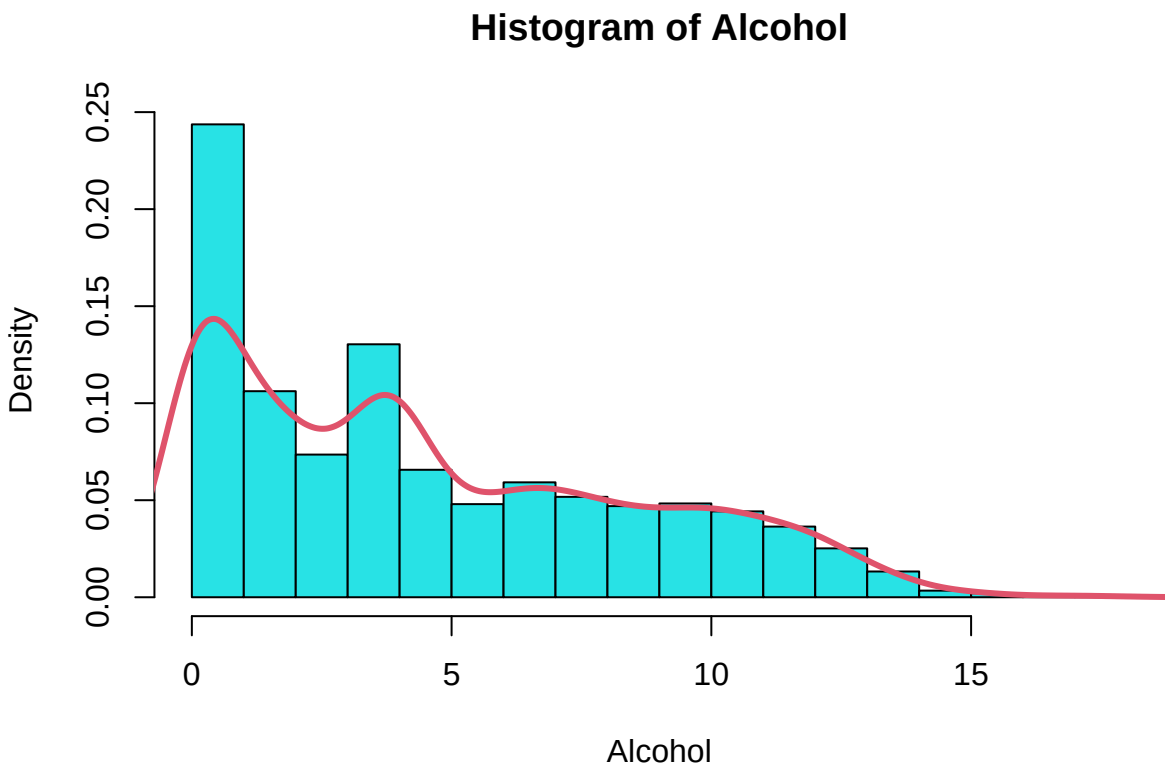
```
hist(mydata$Adult.Mortality,main="Histogram of Adult Mortality",xlab="Adult Mortality",col=5,freq=FALSE)
lines(density(mydata$Adult.Mortality),col=2,lwd=3)
```



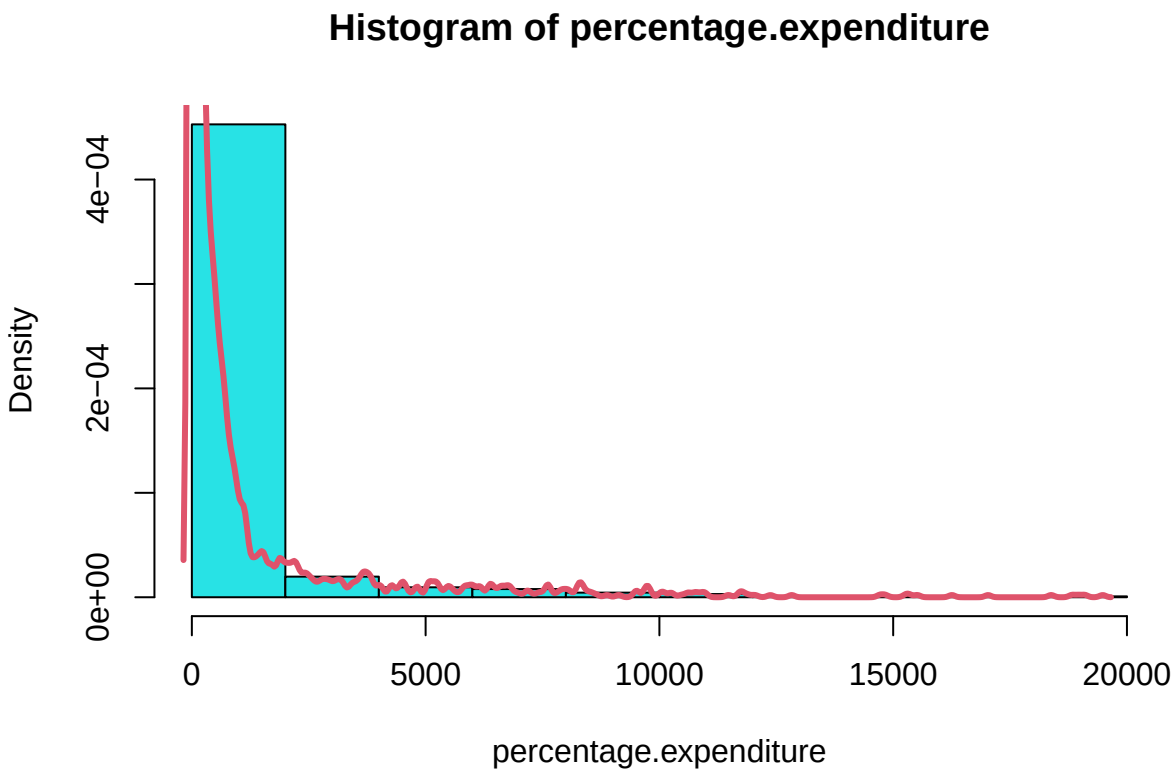

```
hist(mydata$infant.deaths,main="Histogram of infant.deaths",xlab="infant.deaths",col=5,freq =FALSE)#Ske
lines(density(mydata$infant.deaths),col=2,lwd=3)
```



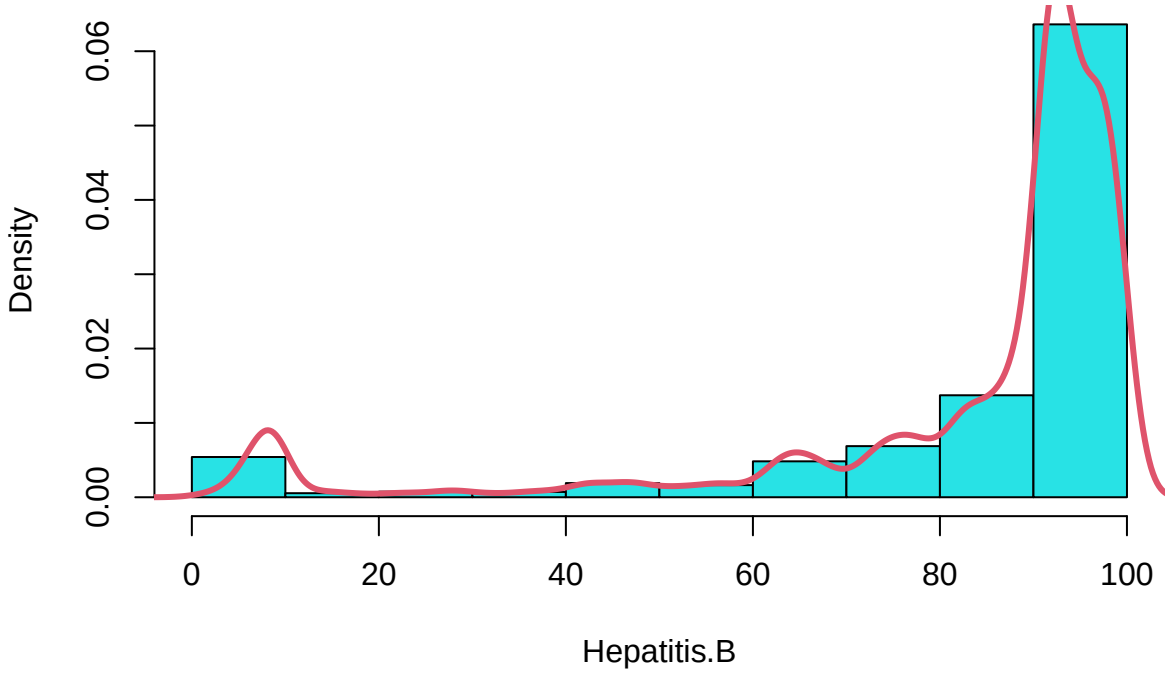

```
hist(mydata$Alcohol,main="Histogram of Alcohol",xlab="Alcohol",col=5,freq=F)#Skewed at right
lines(density(mydata$Alcohol),col=2,lwd=3)
```



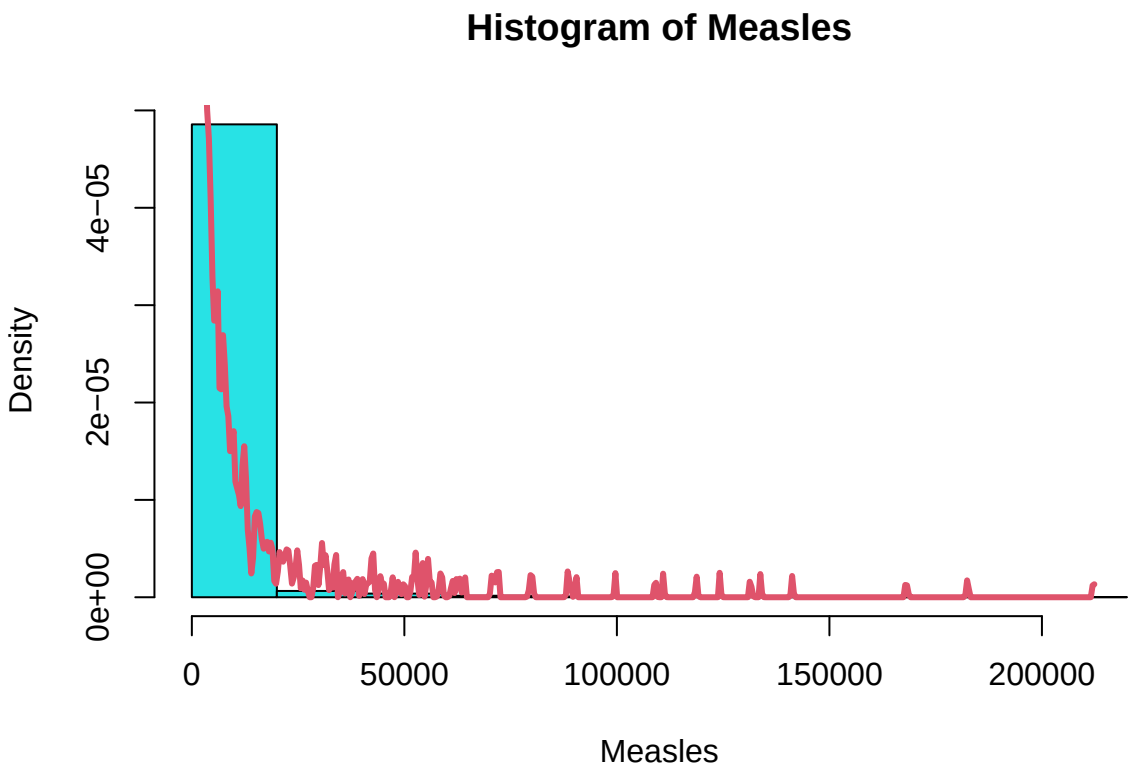

```
hist(mydata$percentage.expenditure,main="Histogram of percentage.expenditure",xlab="percentage.expenditu
lines(density(mydata$percentage.expenditure),col=2,lwd=3)
```



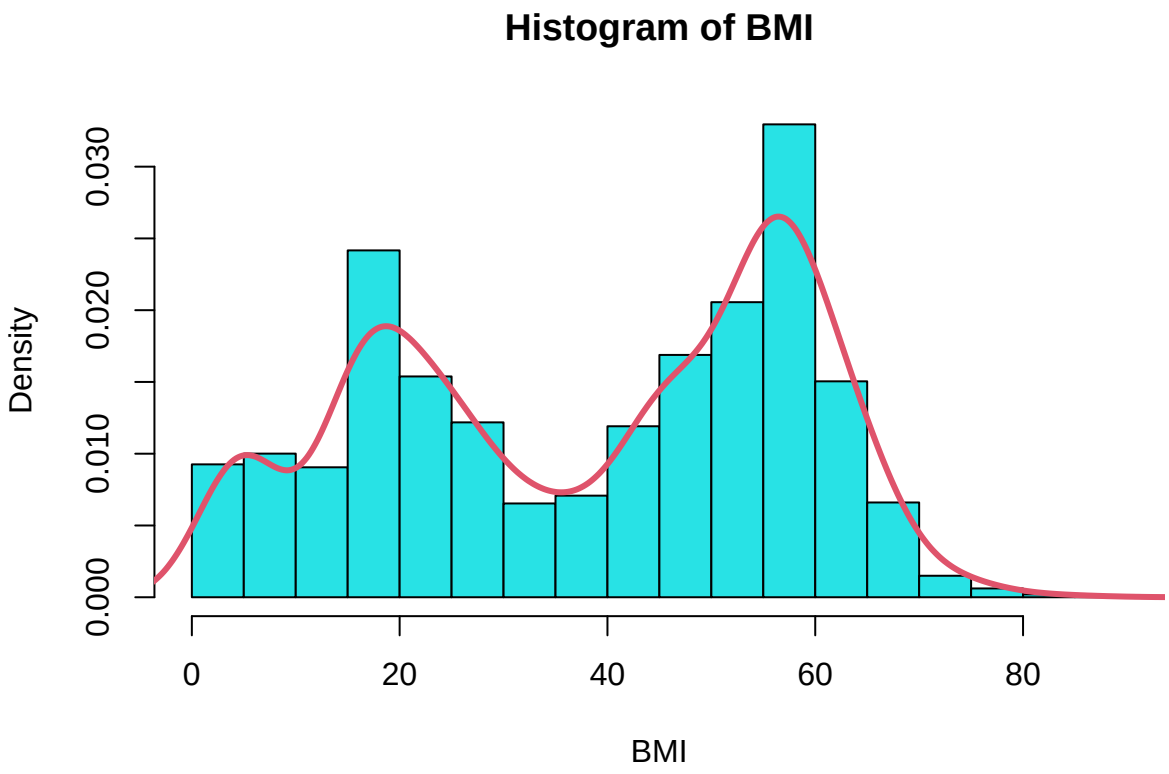

```
hist(mydata$Hepatitis.B,main="Histogram of Hepatitis.B",xlab="Hepatitis.B",col=5,freq=F)#Skewed at left
lines(density(mydata$Hepatitis.B),col=2,lwd=3)
```



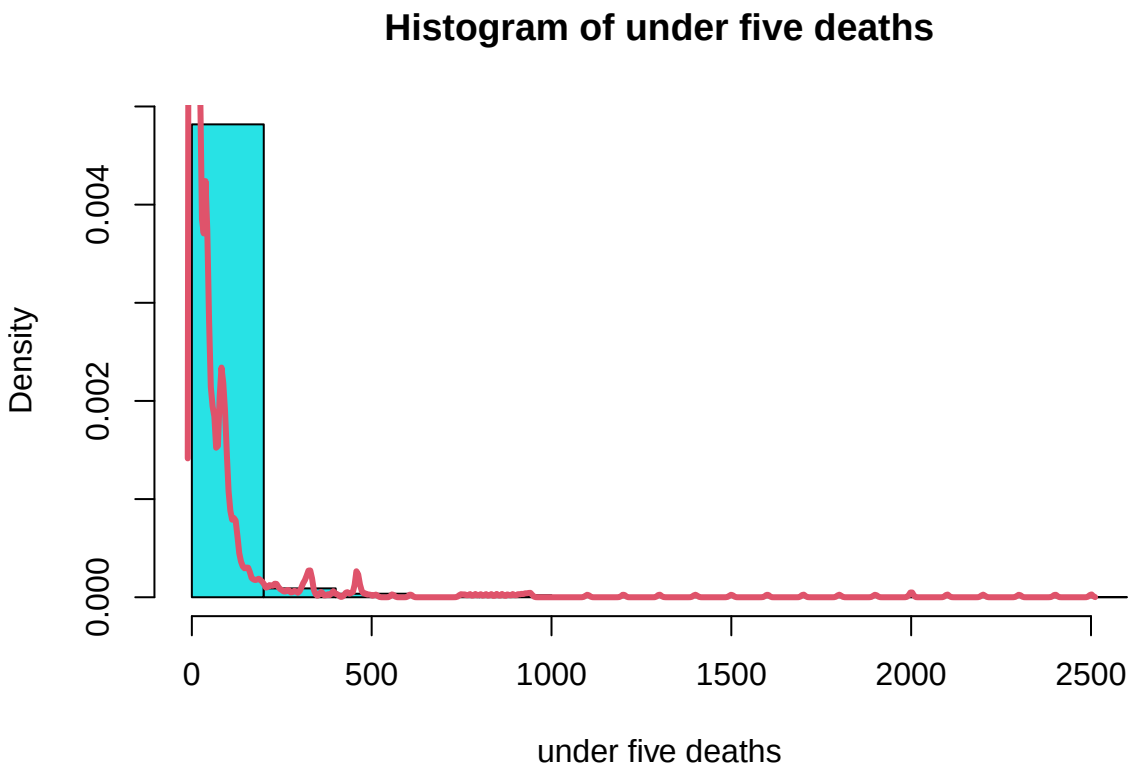

```
hist(mydata$Measles,main="Histogram of Measles",xlab="Measles",col=5,freq=F)#Skewed at right
lines(density(mydata$Measles),col=2,lwd=3)
```



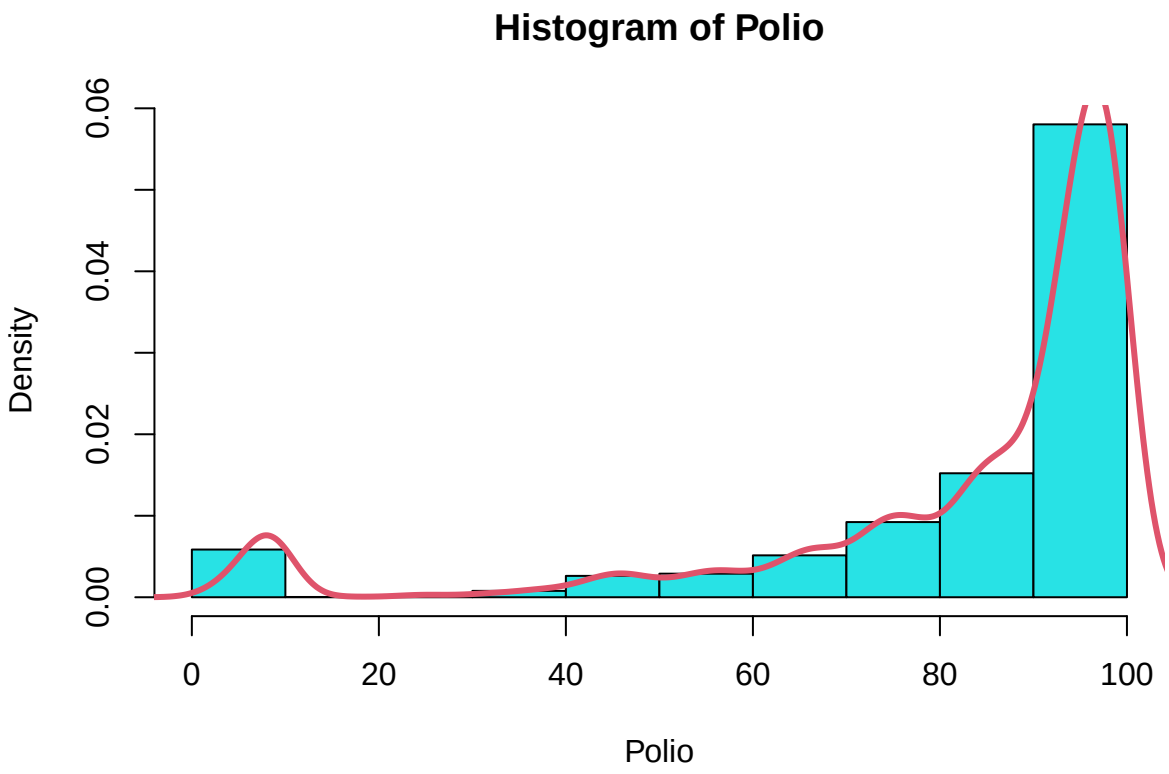

```
hist(mydata$BMI,main="Histogram of BMI",xlab="BMI",col=5,freq=F)#Bimodal
lines(density(mydata$BMI),col=2,lwd=3)
```



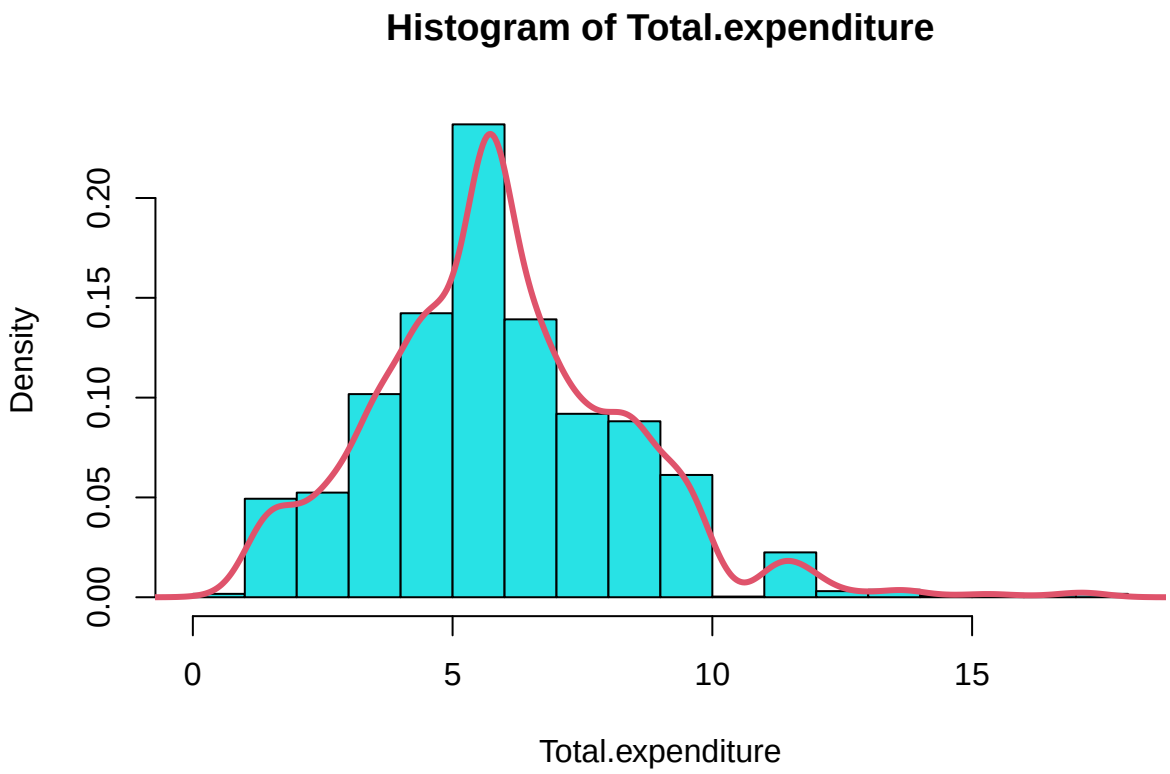

```
hist(mydata$under.five.deaths,main="Histogram of under five deaths",xlab="under five deaths",col=5,freq
lines(density(mydata$under.five.deaths),col=2,lwd=3)
```



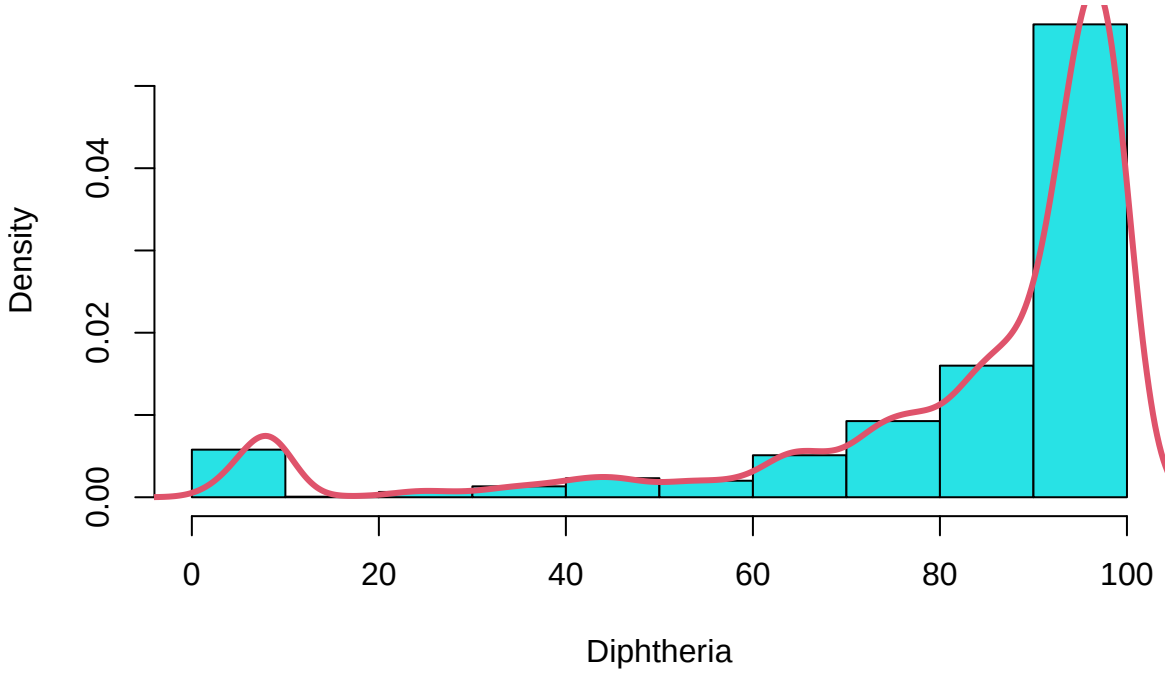

```
hist(mydata$Polio,main="Histogram of Polio",xlab="Polio",col=5,freq=F)#Skewed at left
lines(density(mydata$Polio),col=2,lwd=3)
```



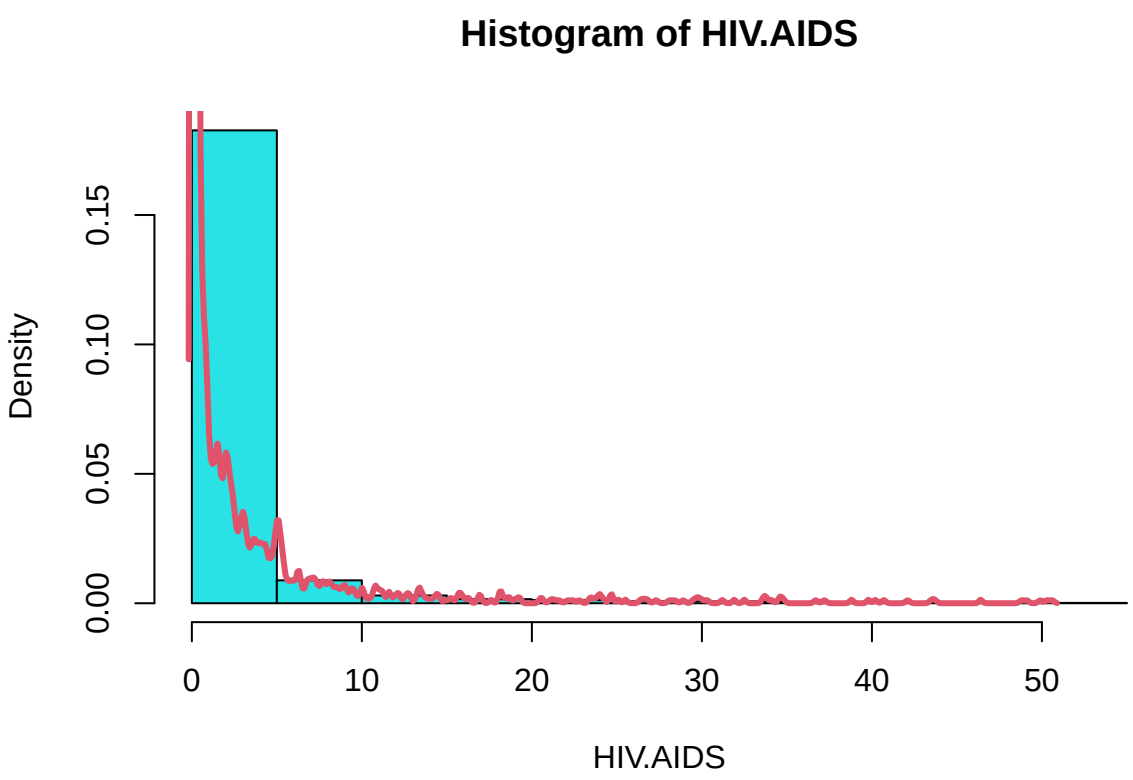

```
hist(mydata$Total.expenditure,main="Histogram of Total.expenditure",xlab="Total.expenditure",col=5,freq
lines(density(mydata$Total.expenditure),col=2,lwd=3)
```



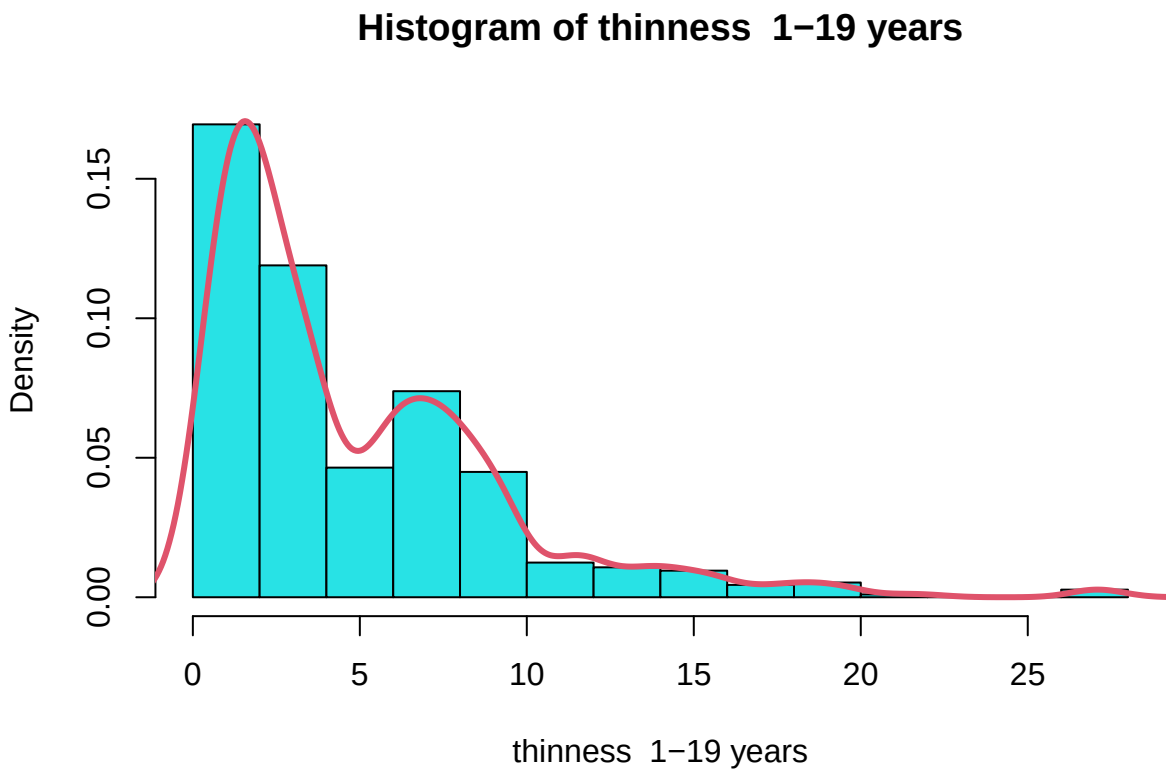

```
hist(mydata$Diphtheria,main="Histogram of Diphtheria",xlab="Diphtheria",col=5,freq=F)#Skewed at left
lines(density(mydata$Diphtheria),col=2,lwd=3)
```



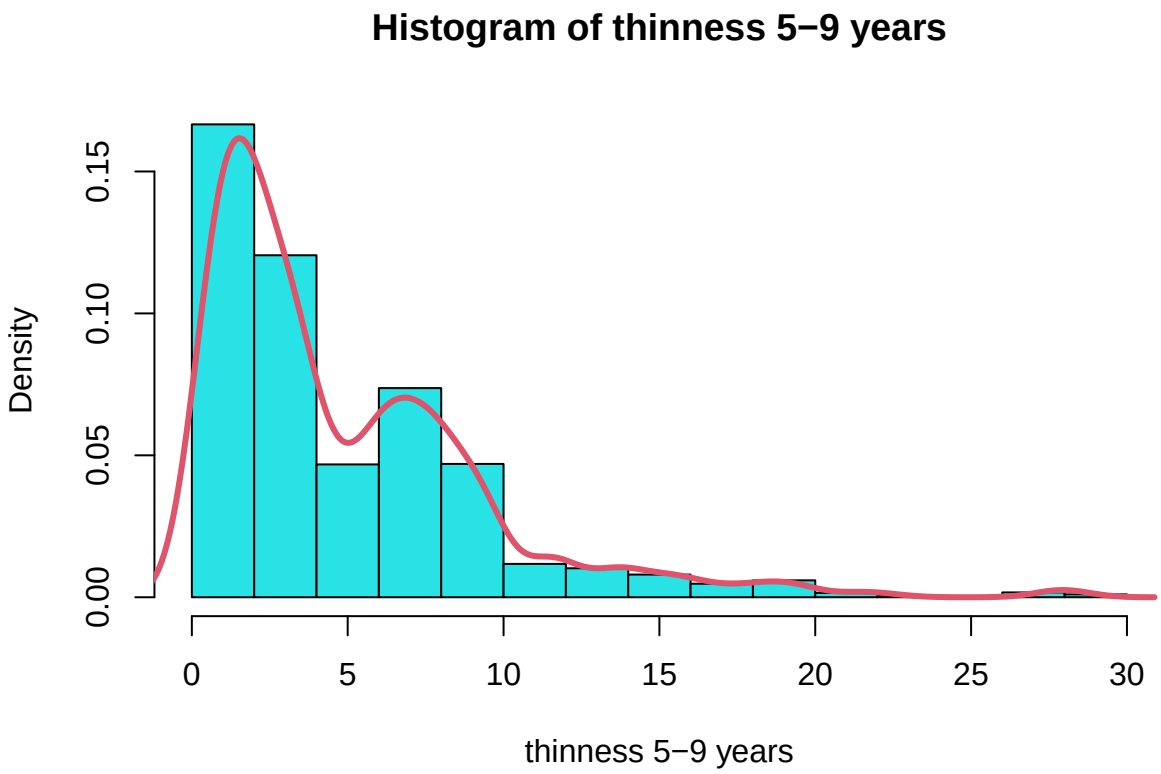

```
hist(mydata$HIV.AIDS,main="Histogram of HIV.AIDS",xlab="HIV.AIDS",col=5,freq=F)#Skewed at right
lines(density(mydata$HIV.AIDS),col=2,lwd=3)
```



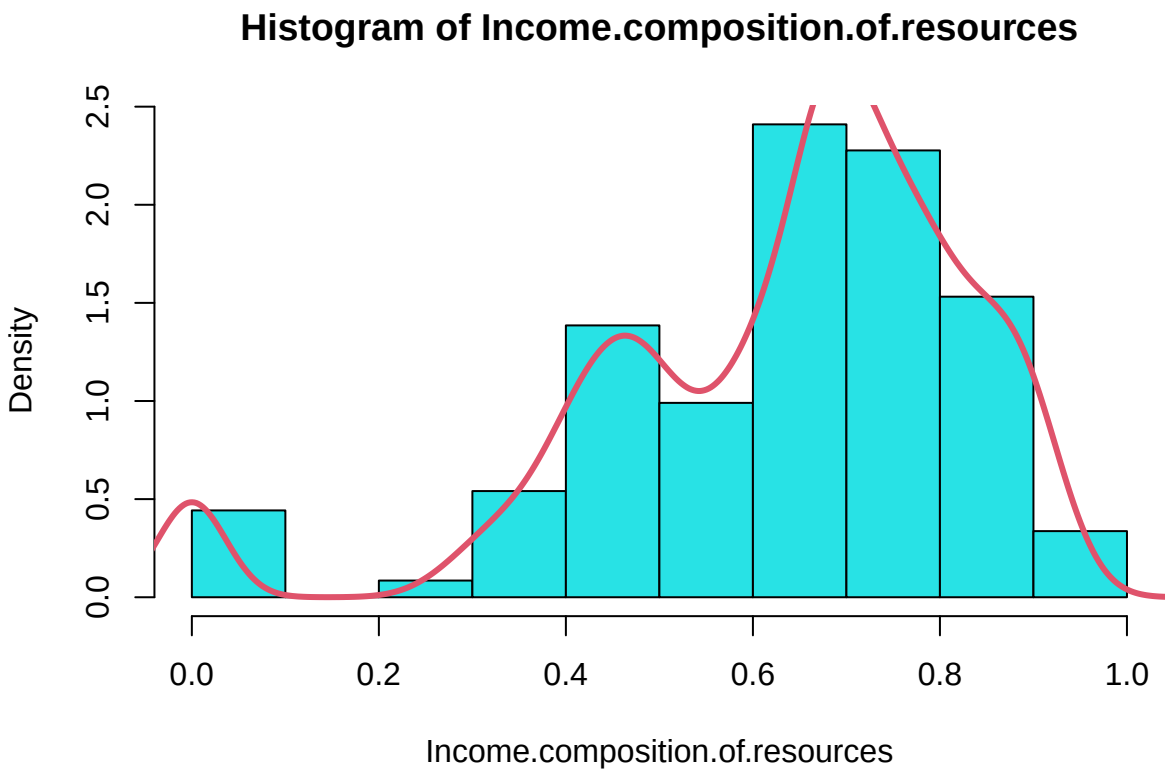

```
hist(mydata$thinness..1.19.years,main="Histogram of thinness  1-19 years",xlab="thinness  1-19 years",co
lines(density(mydata$thinness..1.19.years),col=2,lwd=3)
```



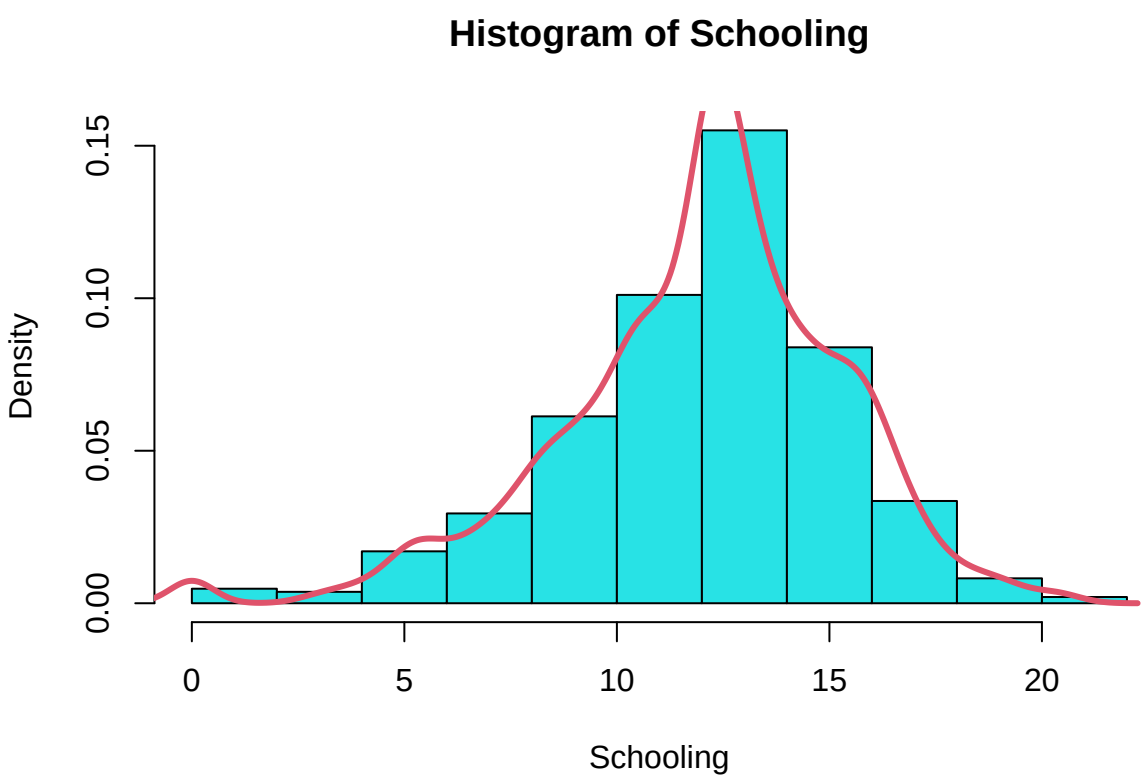

```
hist(mydata$thinness.5.9.years,main="Histogram of thinness 5-9 years",xlab="thinness 5-9 years",col=5,f
lines(density(mydata$thinness.5.9.years),col=2,lwd=3)
```




```
hist(mydata$Income.composition.of.resources,main="Histogram of Income.composition.of.resources",xlab="In
lines(density(mydata$Income.composition.of.resources),col=2,lwd=3)
```




```
hist(mydata$Schooling,main="Histogram of Schooling",xlab="Schooling",col=5,freq=F)#Normal distribution
lines(density(mydata$Schooling),col=2,lwd=3)
```



Dealing with missing values

Here, by looking at the above histogram of all the variables,we concluded that each variables are skewed or normally distributed. Hence Median would be the best imputed value.So we replace na.values with median of respective variables.

```
df1=data
for(i in 1:ncol(df1)){
  df1[is.na(df1[,i]),i] <- median(df1[,i] , na.rm = T)
}
```

```
## Warning in mean.default(sort(x, partial = half + 0L:1L)[half + 0L:1L]): argument
## is not numeric or logical: returning NA

## Warning in mean.default(sort(x, partial = half + 0L:1L)[half + 0L:1L]): argument
## is not numeric or logical: returning NA
```

```
data=df1
sum(is.na(df1))
```

```
## [1] 0
```

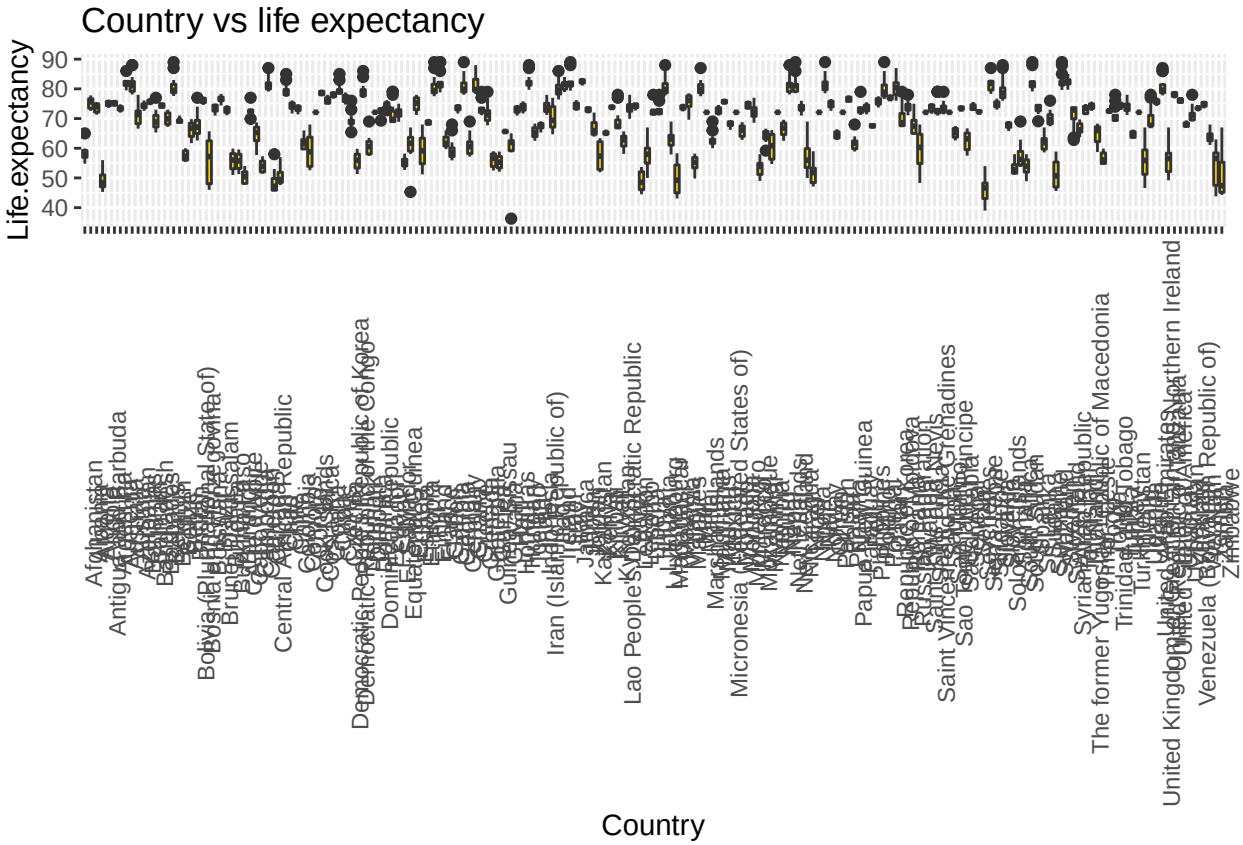
Now, Our data set has Zero na.values, so we proceed to data analysis part.

Exploratory Data Analysis :

```
library(ggplot2)
```

Life Expectancy Country Wise

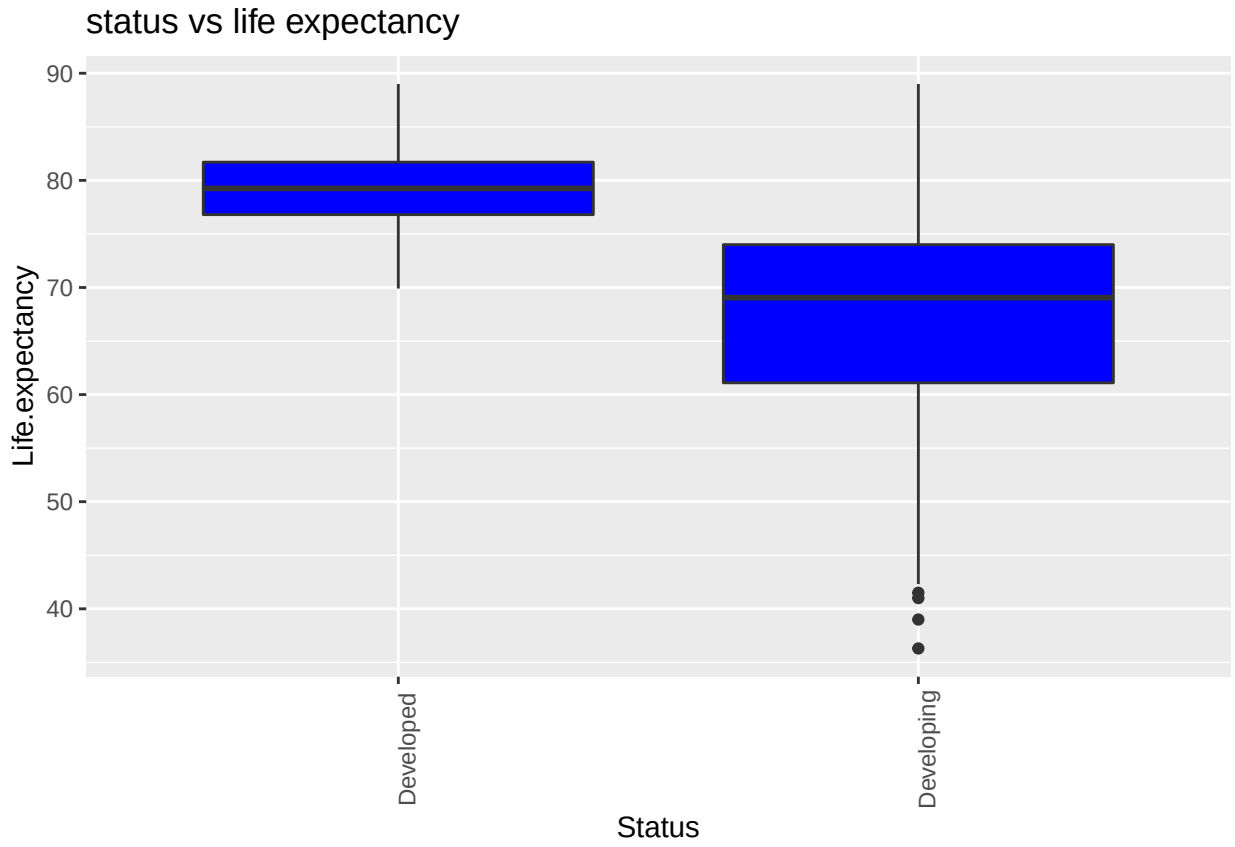
```
#plots
p1<-ggplot(data=df1,aes(x=Country,y=Life.expectancy))+geom_boxplot(fill="gold1")+theme(axis.text.x = el
p1
```



Japan is the country with the highest Life expectancy value followed by Sweden and Sierra Leone has the lowest Life expectancy value

Life Expectancy Comparison in Developed and Developing Countries

```
p2<-ggplot(data=df1,aes(x=Status,y=Life.expectancy))+geom_boxplot(fill="blue")+theme(axis.text.x = eleme
  ggtitle("status vs life expectancy")
p2
```

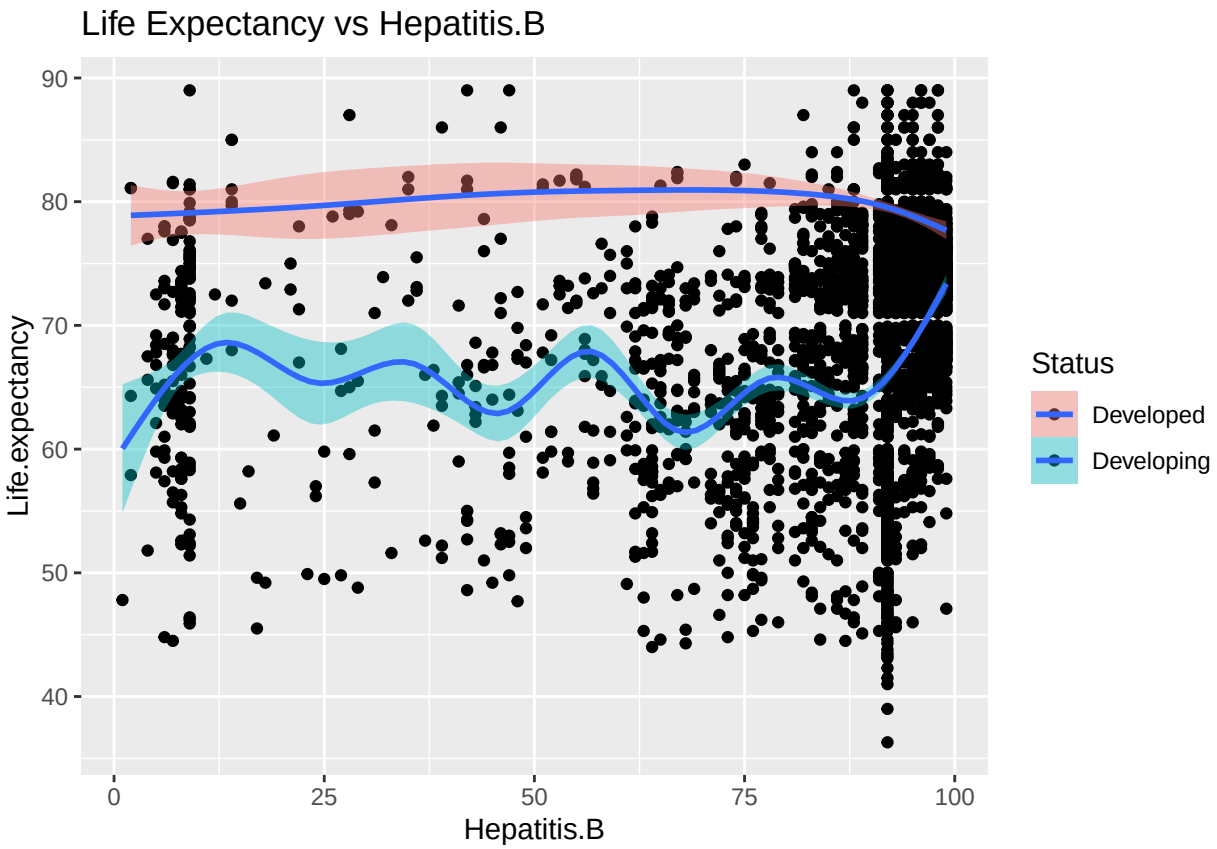


The life expectancy value in case of developing countries is low whereas in case of Developed countries the life expectancy value is comparatively high.

How different diseases affect life expectancy in developed and developing countries

```
#life_expectancy vs heptitis_B
p3 = ggplot(data=df1 , aes(y = Life.expectancy, x = Hepatitis.B,fill= Status ))
f3 = p3 + geom_point() + ggtitle("Life Expectancy vs Hepatitis.B") + stat_smooth()
f3
```

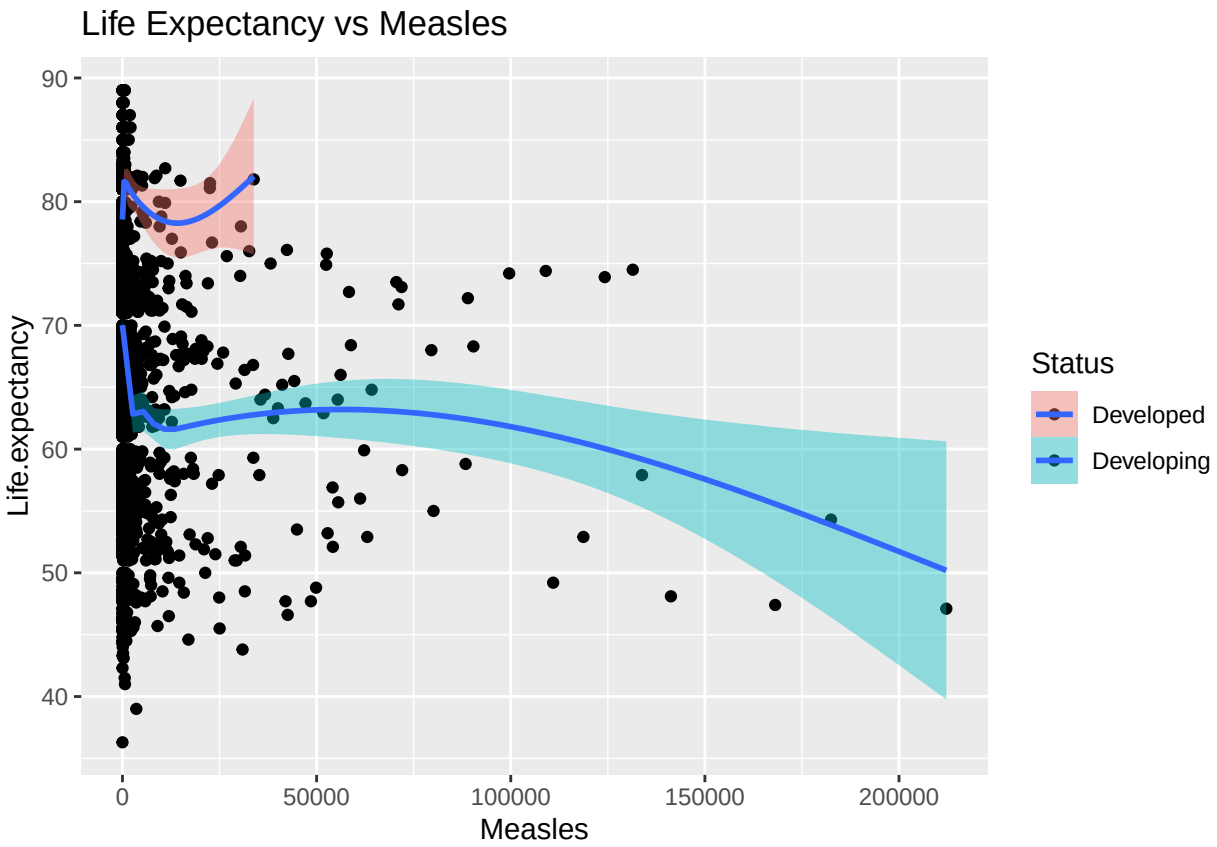
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



There is a slight decrease in the life expectancy value in case of developed countries whereas in case of Developed countries the life expectancy value is gradually rising which means that developing countries are taking measures for setting up vaccine of hepatitis B

```
#life_expectancy vs measles
p4 = ggplot(data=df1 , aes(y = Life.expectancy, x = Measles ,fill=Status))
f4 = p4 + geom_point() + ggtitle("Life Expectancy vs Measles") + stat_smooth()
f4
```

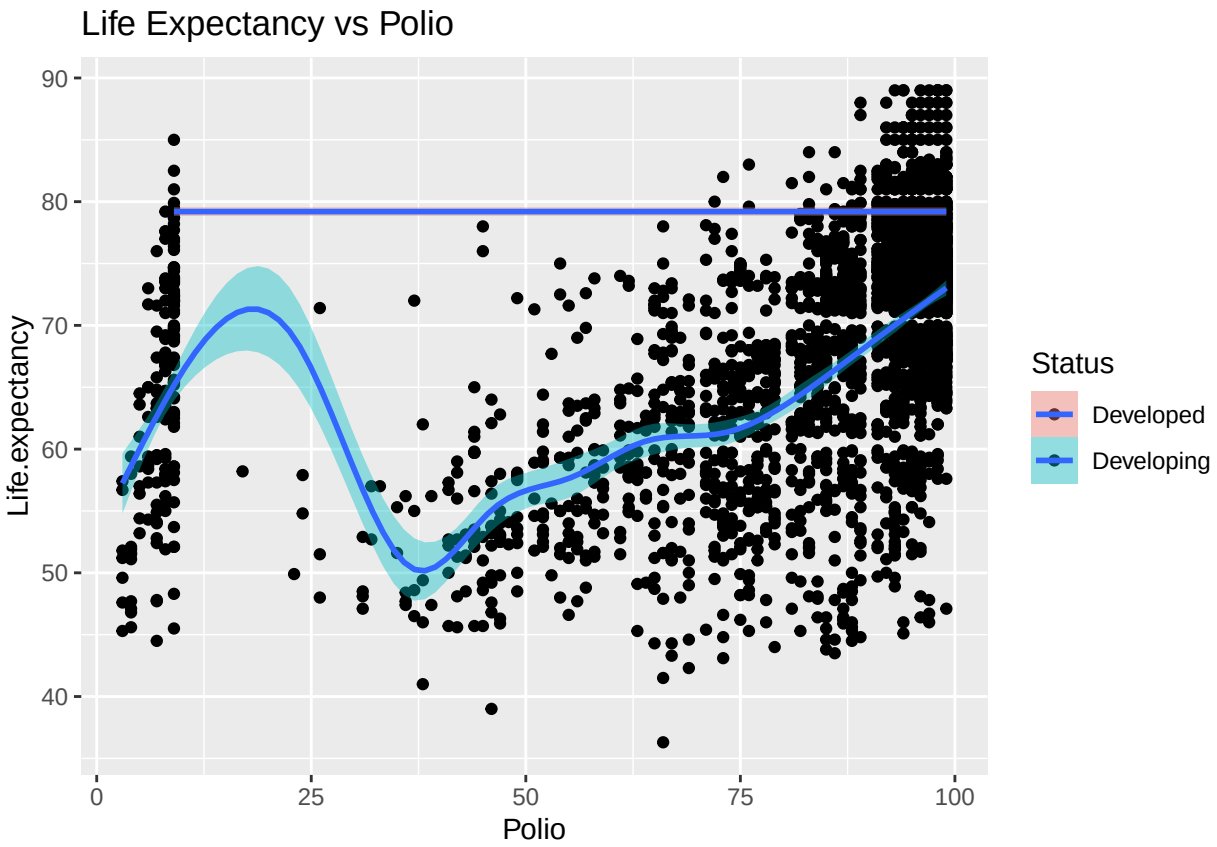
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



In case of Measles, according to the graph the developed countries seems to have vaccines available to tackle measles whereas developing countries life expectancy values is decreasing day by day maybe because of lack of resources to handle measles

```
#life_expectancy vs polio
p6 = ggplot(data=df1 , aes(y = Life.expectancy, x = Polio,fill=Status ))
f6 = p6 + geom_point() + ggtitle("Life Expectancy vs Polio") + stat_smooth()
f6
```

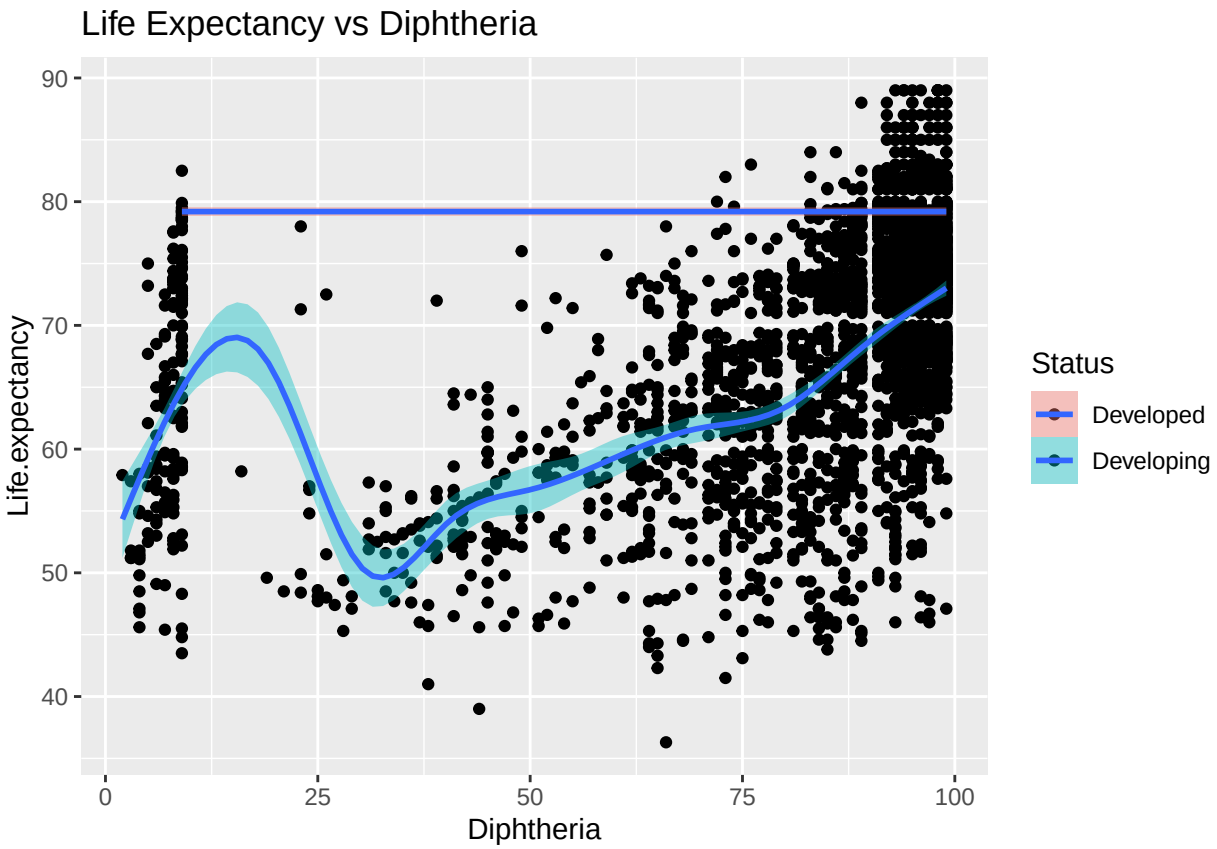
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Developed countries seems to have successfully eradicated polio diesease because of vaccines whereas in developing countries there was low expectancy value initially but now it is gradually increasing maybe because of proper doses being given

```
#life_expectancy vs diphtheria
p8 = ggplot(data=df1 , aes(y = Life.expectancy, x = Diphtheria,fill=Status))
f8 = p8 + geom_point() + ggtitle("Life Expectancy vs Diphtheria") + stat_smooth()
f8
```

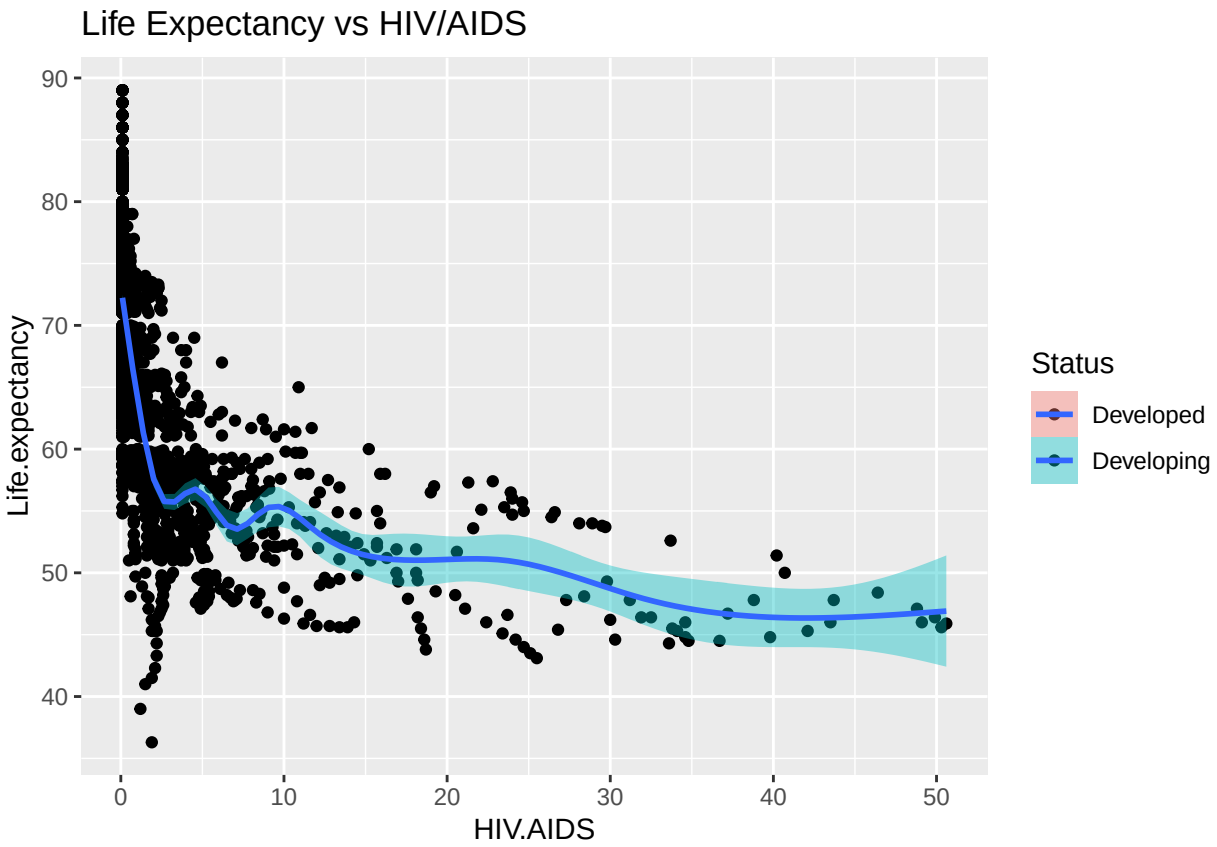
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Developed countries seems to have successfully eradicated diptheria diesease because of vaccines whereas in developing countries there was low expectancy value initially but now it is gradually increasing maybe because of proper doses being given

```
#life_expectancy vs HIV/AIDS
p9 = ggplot(data=df1 , aes(y = Life.expectancy, x = HIV.AIDS,fill=Status ))
f9 = p9 + geom_point() + ggtitle("Life Expectancy vs HIV/AIDS") + stat_smooth()
f9
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

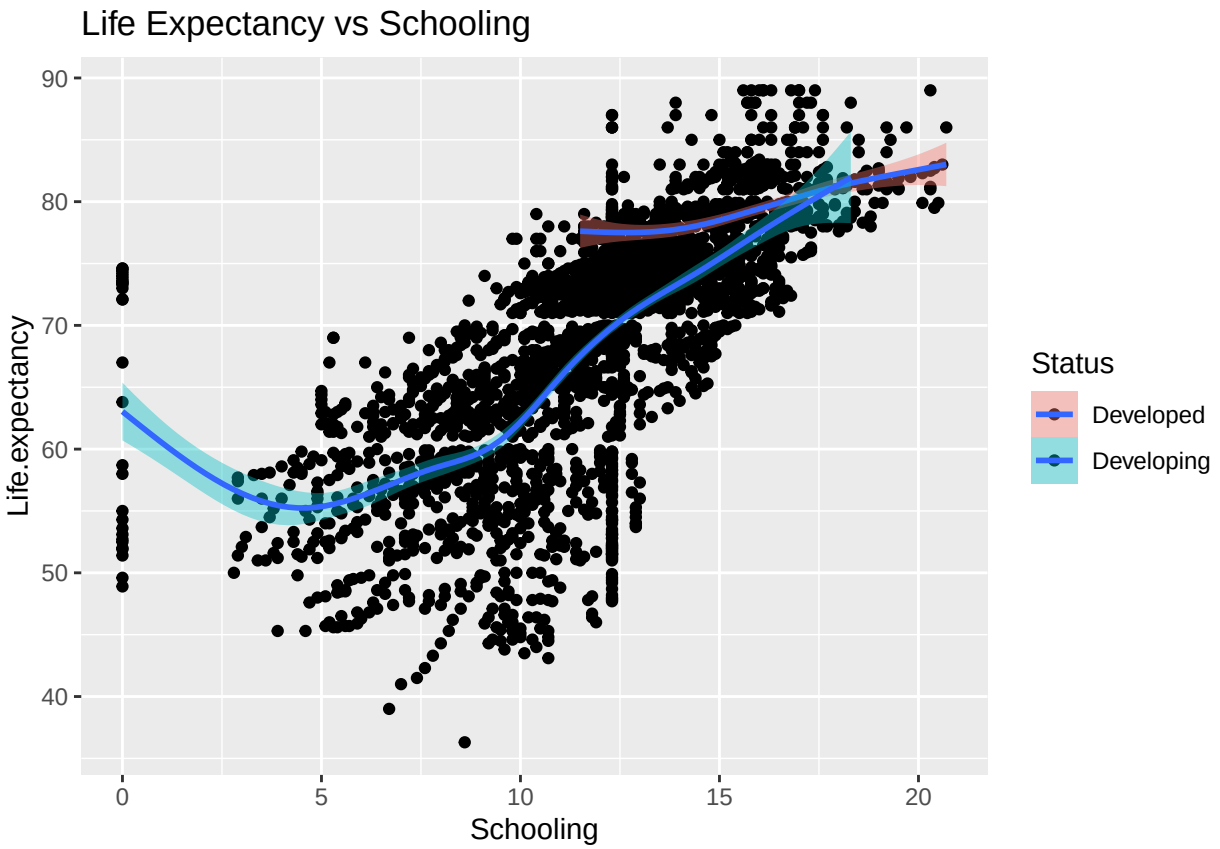


The graph shows that developing countries still have not been able to handle hiv/aids at all as the life expectancy value is decreasing at a rapid range. This can be due to rising population and no education been given

What effect does schooling and alcohol have on life expectancy

```
#life_expectancy vs schooling
p11 = ggplot(data=df1 , aes(y = Life.expectancy, x = Schooling ,fill=Status))
f11 = p11 + geom_point() + ggtitle("Life Expectancy vs Schooling") + stat_smooth()
f11
```

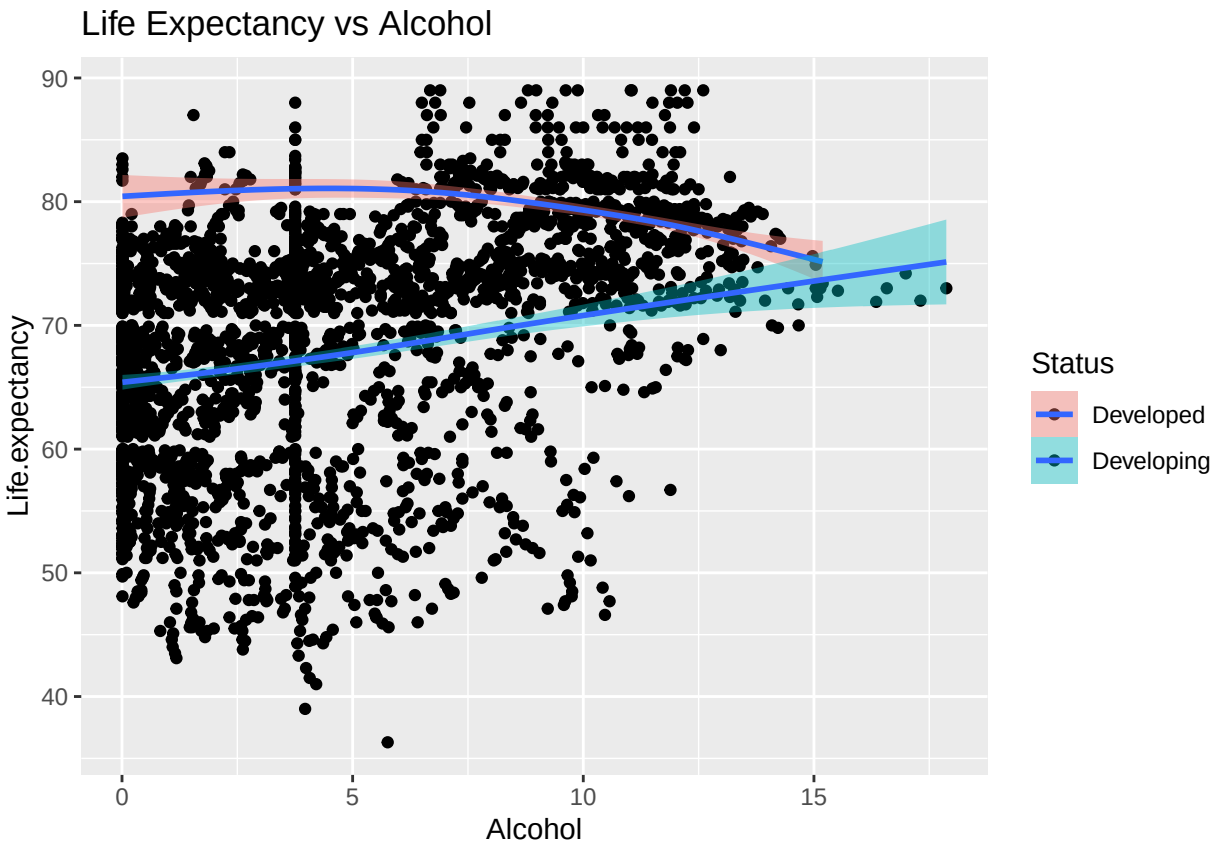
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Schooling can effect life expectancy more in developing countries than developed countries. This may be because education is more established and prevalent in wealthier countries.

```
#life_expectancy vs alcohol
p12 = ggplot(data=df1 , aes(y = Life.expectancy, x = Alcohol ,fill=Status))
f12 = p12 + geom_point() + ggtitle("Life Expectancy vs Alcohol") + stat_smooth()
f12
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



I'm guessing that this is due to the fact that only wealthier countries can afford alcohol or the consumption of alcohol is more prevalent among wealthier populations.

That is why developing countries and alcohol have positive relation and developed countries and alcohol have negative relation.

Fitting a multiple regression model

Baseline Model:

At first, we are making model taking Life Expectancy as the response variable and all others as predictors.

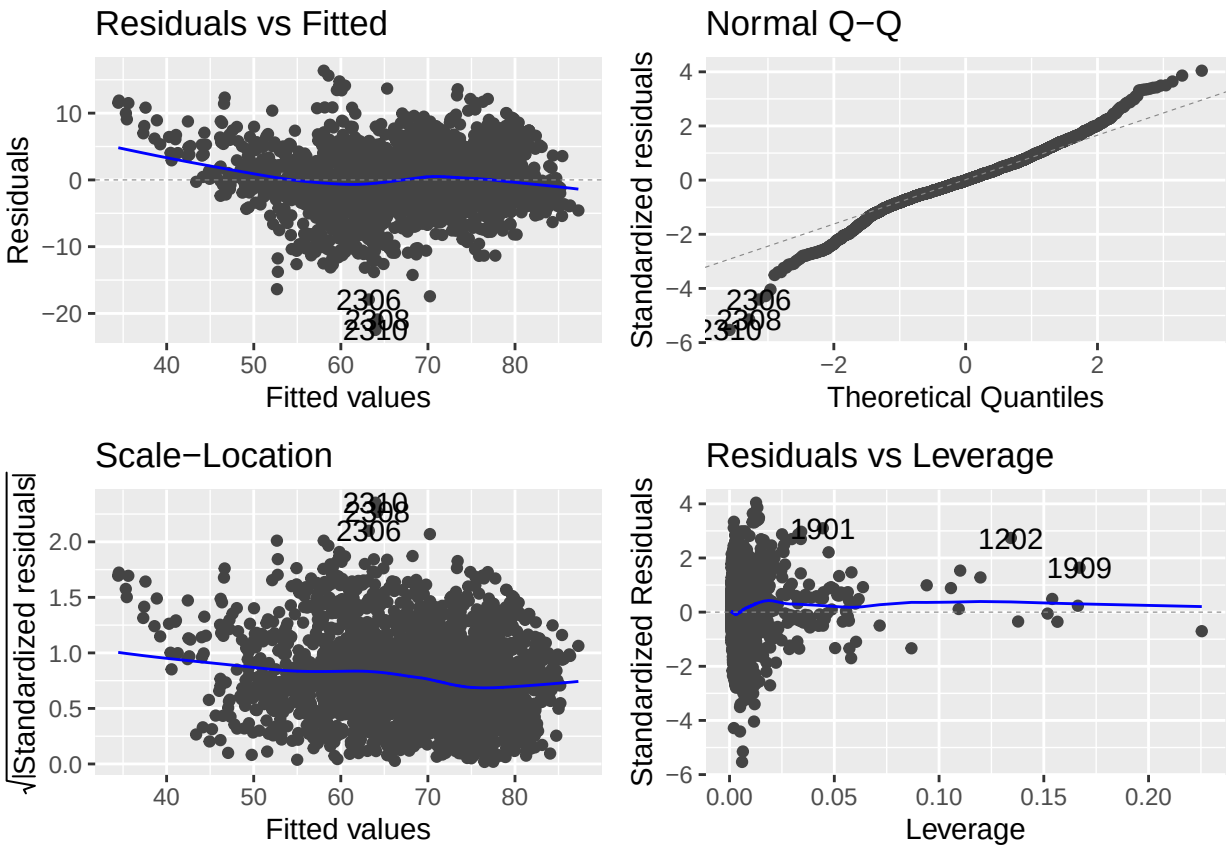
```
data3 = data[,-c(1:3)]
attach(data3)
model1 = lm(Life.expectancy ~ .,data3)
summary(model1)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ ., data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.4719  -2.1696  -0.0549   2.3290  16.3588
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.466e+01  5.793e-01  94.357  < 2e-16 ***
```

```
## Adult.Mortality                 -2.039e-02  7.940e-04 -25.676  < 2e-16 ***
## infant.deaths                    9.946e-02  8.484e-03  11.722  < 2e-16 ***
## Alcohol                          1.190e-01  2.416e-02   4.925 8.92e-07 ***
## percentage.expenditure           7.994e-05  9.057e-05   0.883 0.377483
## Hepatitis.B                     -1.526e-02  3.735e-03  -4.086 4.50e-05 ***
## Measles                         -2.050e-05  7.694e-06  -2.665 0.007749 **
## BMI                              4.394e-02  4.944e-03   8.888  < 2e-16 ***
## under.five.deaths               -7.433e-02  6.217e-03 -11.956  < 2e-16 ***
## Polio                            2.872e-02  4.478e-03   6.415 1.64e-10 ***
## Total.expenditure                9.361e-02  3.410e-02   2.745 0.006081 **
## Diphtheria                       4.025e-02  4.665e-03   8.628  < 2e-16 ***
## HIV.AIDS                        -4.718e-01  1.765e-02 -26.723  < 2e-16 ***
## GDP                              4.681e-05  1.381e-05   3.391 0.000707 ***
## Population                      -7.820e-11  1.698e-09  -0.046 0.963277
## thinness..1.19.years            -8.458e-02  5.063e-02  -1.671 0.094903 .
## thinness.5.9.years               3.579e-03  4.990e-02   0.072 0.942827
## Income.composition.of.resources  5.805e+00  6.331e-01   9.168  < 2e-16 ***
## Schooling                        6.720e-01  4.179e-02  16.079  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.074 on 2919 degrees of freedom
## Multiple R-squared:  0.8176, Adjusted R-squared:  0.8165
## F-statistic:   727 on 18 and 2919 DF,  p-value: < 2.2e-16
```

Plotting the Model :

```
library(ggplot2)
library(ggfortify)
autoplot(model1)
```



Clearly from the above summary of model1 we can conclude that the following parameters such as percentage expenditure, population, thinness(1-19 years), thinness(5-9 years), income composition of resources are insignificant(p-value of t-test is more than 0.05) to estimate our predictor variable.So the above variables are dropped from our model.

MultiCollinearity Test :

VIF Test (Variance Inflation FActor)

```
library(carData)
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
vif(model1)
```

```
##                  Adult.Mortality                      infant.deaths
##                         1.718038                         177.181727
##                          Alcohol           percentage.expenditure
##                         1.589377                           5.737567
##                      Hepatitis.B                            Measles
##                         1.305980                           1.377767
##                              BMI                 under.five.deaths
##                         1.719067                         176.113622
##                            Polio                 Total.expenditure
##                         1.937826                           1.186037
##                       Diphtheria                           HIV.AIDS
##                         2.155516                           1.422386
##                              GDP                         Population
##                         5.965483                           1.489929
##             thinness..1.19.years                thinness.5.9.years
##                         8.773122                           8.869434
## Income.composition.of.resources                          Schooling
##                         2.985808                           3.296186
```

Clearly variables like infant deaths, under five deaths have high vif values(> 100) that means it's highly correlated with other variables.So it violates the independency of explanatory variables(X) so it will be a wise decision to drop these variables.
So we are going to drop 7 variables percentage expenditure, population, thinness(1-19 years), thinness(5-9 years), income composition of resources, infant deaths, under five deaths as these are coming under the low significance case and very high multicollinearity case.

Improved Model 2

```
model2 = lm(Life.expectancy~.-percentage.expenditure-thinness..1.19.years-thinness.5.9.years-Income.comp
summary(model2)
```

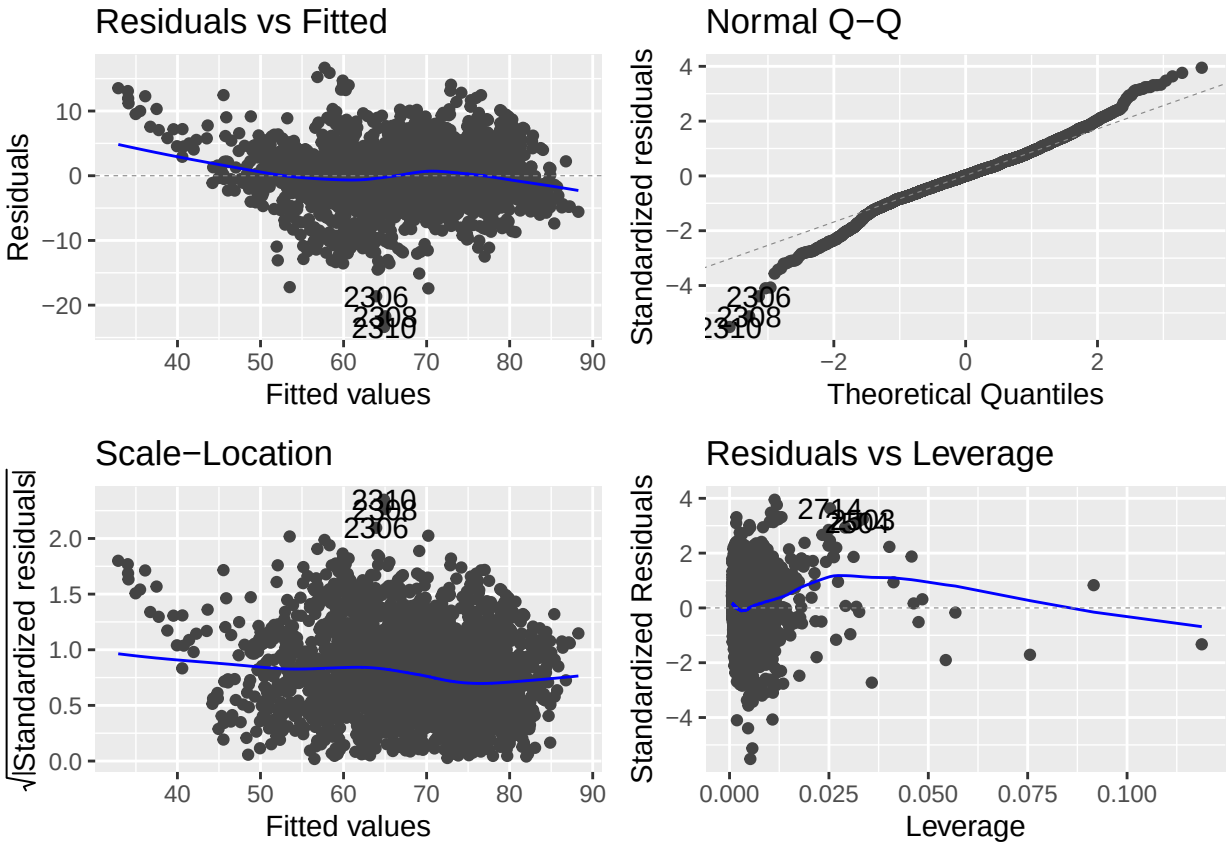
```
##
## Call:
## lm(formula = Life.expectancy ~ . - percentage.expenditure - thinness..1.19.years -
##     thinness.5.9.years - Income.composition.of.resources - Population -
##     infant.deaths - under.five.deaths, data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3723  -2.3559   0.0234   2.5060  16.6845
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.334e+01  5.360e-01  99.514  < 2e-16 ***
## Adult.Mortality   -2.168e-02  8.233e-04 -26.338  < 2e-16 ***
## Alcohol            1.078e-01  2.398e-02   4.495 7.22e-06 ***
## Hepatitis.B       -2.006e-02  3.841e-03  -5.223 1.88e-07 ***
## Measles           -3.657e-05  7.022e-06  -5.209 2.03e-07 ***
## BMI                5.492e-02  4.778e-03  11.496  < 2e-16 ***
## Polio              3.226e-02  4.659e-03   6.924 5.37e-12 ***
## Total.expenditure  9.212e-02  3.482e-02   2.645   0.0082 **
```

```
## Diphtheria          5.101e-02  4.807e-03  10.611  < 2e-16 ***
## HIV.AIDS           -4.942e-01  1.828e-02 -27.036  < 2e-16 ***
## GDP                 6.365e-05  6.652e-06   9.570  < 2e-16 ***
## Schooling           9.697e-01  3.392e-02  28.589  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.251 on 2926 degrees of freedom
## Multiple R-squared:  0.8009, Adjusted R-squared:  0.8002
## F-statistic:  1070 on 11 and 2926 DF,  p-value: < 2.2e-16
```

After removing the 7 insignificant variables model2 is been made and R^2 value of model2 is 80.02 that means it decreases by only 0.01.

Plotting the Model 4 :

```
library(ggplot2)
library(ggfortify)
autoplot(model2)
```

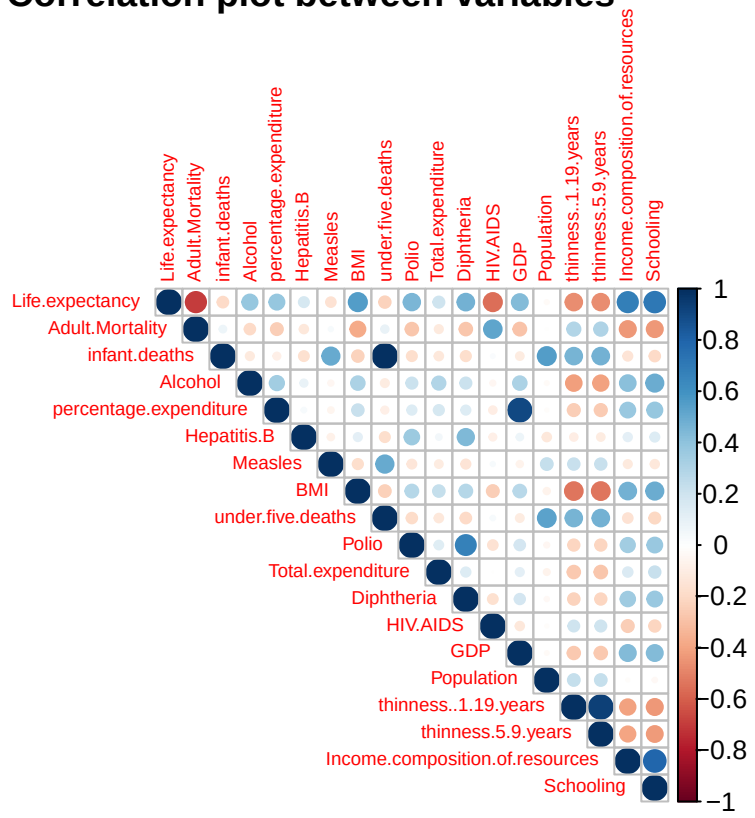


Checkig Interaction terms : Now we can check whether there is correlation between any two variables or not by the help of correlation plot and correlation matrix.

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
corrplot(cor(data3),type="upper",method="circle",title="Correlation plot between variables",
         mar=c(0.7,0.7,0.7,0.7),tl.cex = 0.6)
```




```
cor(Adult.Mortality,HIV.AIDS)
```

```
## [1] 0.5238865
```

```
cor(BMI, thinness..1.19.years)
```

```
## [1] -0.5324869
```

```
cor(BMI, thinness.5.9.years)
```

```
## [1] -0.5393652
```

```
cor(BMI, Schooling)
```

```
## [1] 0.4998062
```

```
cor(Polio, Diphtheria)
```

```
## [1] 0.6739768
```

```
cor(Schooling, Income.composition.of.resources)
```

```
## [1] 0.7953833
```

From the above correlation plot and correlation value of each variables it is found that a decent amount of correlation exsit between Adult-mortality & HIV/AIDS, BMI & thinness(1-19years), BMI & thinness(5-9years), BMI & Schooling, Polio & Diphtheria, Schooling & Income-composition of resources.So their interaction terms is been added in our model.

```
test1 = lm(Life.expectancy~.-percentage.expenditure-thinness..1.19.years-thinness.5.9.years-Income.compo
summary(test1)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ . - percentage.expenditure - thinness..1.19.years -
##     thinness.5.9.years - Income.composition.of.resources - Population -
##     infant.deaths - under.five.deaths + Adult.Mortality:HIV.AIDS +
##     Schooling:Income.composition.of.resources + BMI:Schooling +
##     Polio:Diphtheria + BMI:thinness..1.19.years + BMI:thinness.5.9.years,
##     data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.2899  -2.0040  -0.0193   2.1680  18.1468
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             5.405e+01  8.729e-01  61.918
## Adult.Mortality                        -2.686e-02  8.690e-04 -30.915
## Alcohol                                 8.678e-02  2.315e-02   3.748
## Hepatitis.B                            -1.472e-02  3.583e-03  -4.107
## Measles                                -3.266e-05  6.473e-06  -5.047
## BMI                                     2.278e-01  1.628e-02  13.988
## Polio                                   1.293e-02  8.336e-03   1.551
## Total.expenditure                       4.846e-02  3.259e-02   1.487
## Diphtheria                              2.864e-02  8.290e-03   3.455
## HIV.AIDS                               -8.648e-01  3.061e-02 -28.255
## GDP                                     4.650e-05  6.453e-06   7.206
## Schooling                               8.897e-01  6.786e-02  13.110
## Adult.Mortality:HIV.AIDS                1.040e-03  6.653e-05  15.626
## Income.composition.of.resources:Schooling  5.318e-01  4.755e-02  11.184
## BMI:Schooling                          -1.461e-02  1.248e-03 -11.702
## Polio:Diphtheria                        1.851e-04  1.096e-04   1.688
## BMI:thinness..1.19.years               -2.013e-03  2.653e-03  -0.759
## BMI:thinness.5.9.years                 -1.581e-03  2.621e-03  -0.603
##                                           Pr(>|t|)
## (Intercept)                                < 2e-16 ***
## Adult.Mortality                            < 2e-16 ***
## Alcohol                                   0.000182 ***
## Hepatitis.B                               4.12e-05 ***
## Measles                                   4.77e-07 ***
## BMI                                        < 2e-16 ***
```

```
## Polio                                          0.121051
## Total.expenditure                              0.137164
## Diphtheria                                     0.000558 ***
## HIV.AIDS                                        < 2e-16 ***
## GDP                                            7.33e-13 ***
## Schooling                                       < 2e-16 ***
## Adult.Mortality:HIV.AIDS                        < 2e-16 ***
## Income.composition.of.resources:Schooling       < 2e-16 ***
## BMI:Schooling                                   < 2e-16 ***
## Polio:Diphtheria                               0.091475 .
## BMI:thinness..1.19.years                       0.447948
## BMI:thinness.5.9.years                         0.546402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.899 on 2920 degrees of freedom
## Multiple R-squared:  0.8329, Adjusted R-squared:  0.8319
## F-statistic: 855.8 on 17 and 2920 DF,  p-value: < 2.2e-16
```

From the above testing model we can see that Polio & Diphtheria,BMI & thinness..1.19.years ,BMI & thinness.5.9.years are not significant as their p-values are greater than 0.05.So we can drop this interaction terms.

Improved Model 3 :

```
model3 = lm(Life.expectancy~.-percentage.expenditure-thinness..1.19.years-thinness.5.9.years-Income.comp
summary(model3)
```

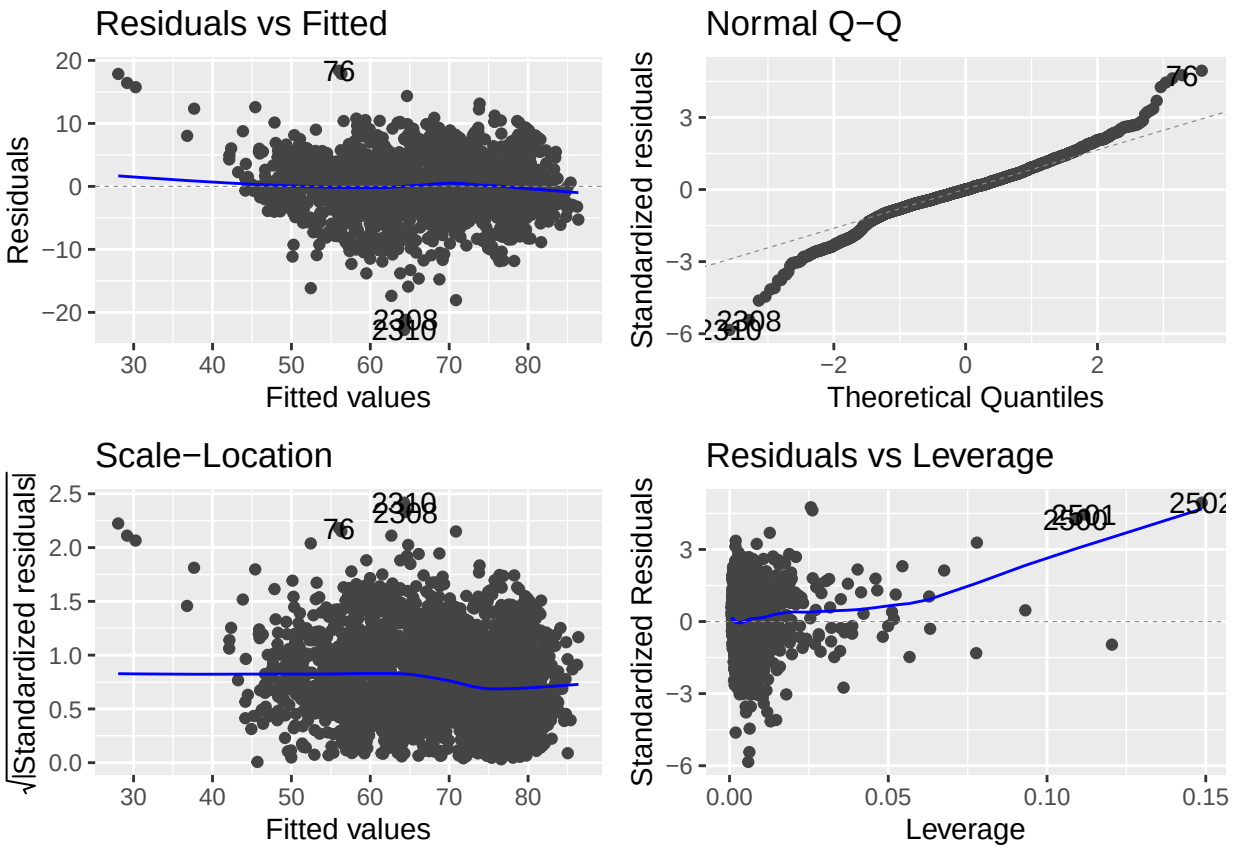
```
##
## Call:
## lm(formula = Life.expectancy ~ . - percentage.expenditure - thinness..1.19.years -
##     thinness.5.9.years - Income.composition.of.resources - Population -
##     infant.deaths - under.five.deaths + Adult.Mortality:HIV.AIDS +
##     Schooling:Income.composition.of.resources + BMI:Schooling,
##     data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7886  -2.0716  -0.0187   2.2203  18.3688
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             5.299e+01  7.432e-01  71.308
## Adult.Mortality                        -2.706e-02  8.690e-04 -31.137
## Alcohol                                 1.145e-01  2.240e-02   5.110
## Hepatitis.B                            -1.431e-02  3.548e-03  -4.034
## Measles                                -3.508e-05  6.469e-06  -5.424
## BMI                                     2.175e-01  1.615e-02  13.467
## Polio                                   2.441e-02  4.302e-03   5.675
## Total.expenditure                       7.560e-02  3.208e-02   2.357
## Diphtheria                              4.022e-02  4.450e-03   9.038
## HIV.AIDS                               -8.766e-01  3.058e-02 -28.668
## GDP                                     4.777e-05  6.464e-06   7.390
## Schooling                               8.706e-01  6.784e-02  12.833
```

```
## Adult.Mortality:HIV.AIDS                    1.042e-03  6.667e-05  15.626
## Income.composition.of.resources:Schooling  5.497e-01  4.755e-02  11.562
## BMI:Schooling                              -1.422e-02  1.247e-03 -11.397
##                                            Pr(>|t|)
## (Intercept)                                 < 2e-16 ***
## Adult.Mortality                             < 2e-16 ***
## Alcohol                                    3.42e-07 ***
## Hepatitis.B                                5.63e-05 ***
## Measles                                    6.32e-08 ***
## BMI                                         < 2e-16 ***
## Polio                                      1.53e-08 ***
## Total.expenditure                            0.0185 *
## Diphtheria                                  < 2e-16 ***
## HIV.AIDS                                    < 2e-16 ***
## GDP                                        1.90e-13 ***
## Schooling                                   < 2e-16 ***
## Adult.Mortality:HIV.AIDS                    < 2e-16 ***
## Income.composition.of.resources:Schooling  < 2e-16 ***
## BMI:Schooling                               < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.912 on 2923 degrees of freedom
## Multiple R-squared:  0.8316, Adjusted R-squared:  0.8308
## F-statistic:  1031 on 14 and 2923 DF,  p-value: < 2.2e-16
```

Clearly by adding the above interaction terms cause 3%increase in our R^2 Value and R^2 value become 83%. Also this cause a decrement of RSE value to 3.912.So It's a better model with respect to model2.

Plotting The Model 3 :

```
library(ggplot2)
library(ggfortify)
autoplot(model3)
```



Fixing the Polynomial Terms

```
test4 = lm(Life.expectancy~.-percentage.expenditure-thinness..1.19.years-thinness.5.9.years-Income.compo
summary(test4)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ . - percentage.expenditure - thinness..1.19.years -
##     thinness.5.9.years - Income.composition.of.resources - Population -
##     infant.deaths - under.five.deaths + Adult.Mortality:HIV.AIDS +
##     Schooling:Income.composition.of.resources + BMI:Schooling -
##     HIV.AIDS + poly(HIV.AIDS, 2) + poly(Polio, 7), data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4817  -2.1095  -0.0122   2.1545  15.1688
##
## Coefficients: (1 not defined because of singularities)
##                                          Estimate Std. Error t value
## (Intercept)                             5.385e+01  7.017e-01  76.743
## Adult.Mortality                        -2.278e-02  8.519e-04 -26.741
## Alcohol                                 1.297e-01  2.112e-02   6.139
## Hepatitis.B                            -8.408e-03  3.457e-03  -2.432
## Measles                                -2.159e-05  6.115e-06  -3.531
## BMI                                     1.635e-01  1.539e-02  10.629
## Polio                                   4.424e-02  4.452e-03   9.936
```

```
## Total.expenditure                             7.150e-02  3.021e-02   2.367
## Diphtheria                                    1.251e-02  4.506e-03   2.777
## GDP                                           5.160e-05  6.051e-06   8.528
## Schooling                                     6.561e-01  6.496e-02  10.101
## poly(HIV.AIDS, 2)1                           -2.311e+02  7.943e+00 -29.095
## poly(HIV.AIDS, 2)2                            5.474e+01  4.142e+00  13.217
## poly(Polio, 7)1                                      NA         NA      NA
## poly(Polio, 7)2                               4.310e+01  4.451e+00   9.682
## poly(Polio, 7)3                              -1.858e+01  3.782e+00  -4.913
## poly(Polio, 7)4                              -1.343e+01  3.698e+00  -3.630
## poly(Polio, 7)5                               2.457e+01  3.809e+00   6.452
## poly(Polio, 7)6                              -2.718e+01  3.756e+00  -7.238
## poly(Polio, 7)7                               1.392e+01  3.688e+00   3.776
## Adult.Mortality:HIV.AIDS                      8.323e-04  6.332e-05  13.145
## Income.composition.of.resources:Schooling     5.421e-01  4.484e-02  12.091
## BMI:Schooling                                -1.057e-02  1.186e-03  -8.917
##                                              Pr(>|t|)
## (Intercept)                                   < 2e-16 ***
## Adult.Mortality                               < 2e-16 ***
## Alcohol                                      9.42e-10 ***
## Hepatitis.B                                  0.015061 *
## Measles                                      0.000421 ***
## BMI                                           < 2e-16 ***
## Polio                                         < 2e-16 ***
## Total.expenditure                            0.018003 *
## Diphtheria                                   0.005514 **
## GDP                                           < 2e-16 ***
## Schooling                                     < 2e-16 ***
## poly(HIV.AIDS, 2)1                            < 2e-16 ***
## poly(HIV.AIDS, 2)2                            < 2e-16 ***
## poly(Polio, 7)1                                   NA
## poly(Polio, 7)2                               < 2e-16 ***
## poly(Polio, 7)3                              9.46e-07 ***
## poly(Polio, 7)4                              0.000288 ***
## poly(Polio, 7)5                              1.28e-10 ***
## poly(Polio, 7)6                              5.80e-13 ***
## poly(Polio, 7)7                              0.000163 ***
## Adult.Mortality:HIV.AIDS                      < 2e-16 ***
## Income.composition.of.resources:Schooling     < 2e-16 ***
## BMI:Schooling                                 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.659 on 2916 degrees of freedom
## Multiple R-squared:  0.853,  Adjusted R-squared:  0.852
## F-statistic: 805.9 on 21 and 2916 DF,  p-value: < 2.2e-16
```

From the above plots and summary of testing model we can conclude that the model can be improved by adding the 2nd degree polynomial of HIV/AIDS and 7th degree polynomial of Polio as the R^2 value increase to 85.3.But on the stake of adding 6 more variables due to the polynomial of Polio the model is improved by only 1. So it will be better if the polynomial terms of Polio are dropped.

Improved Model 4 :

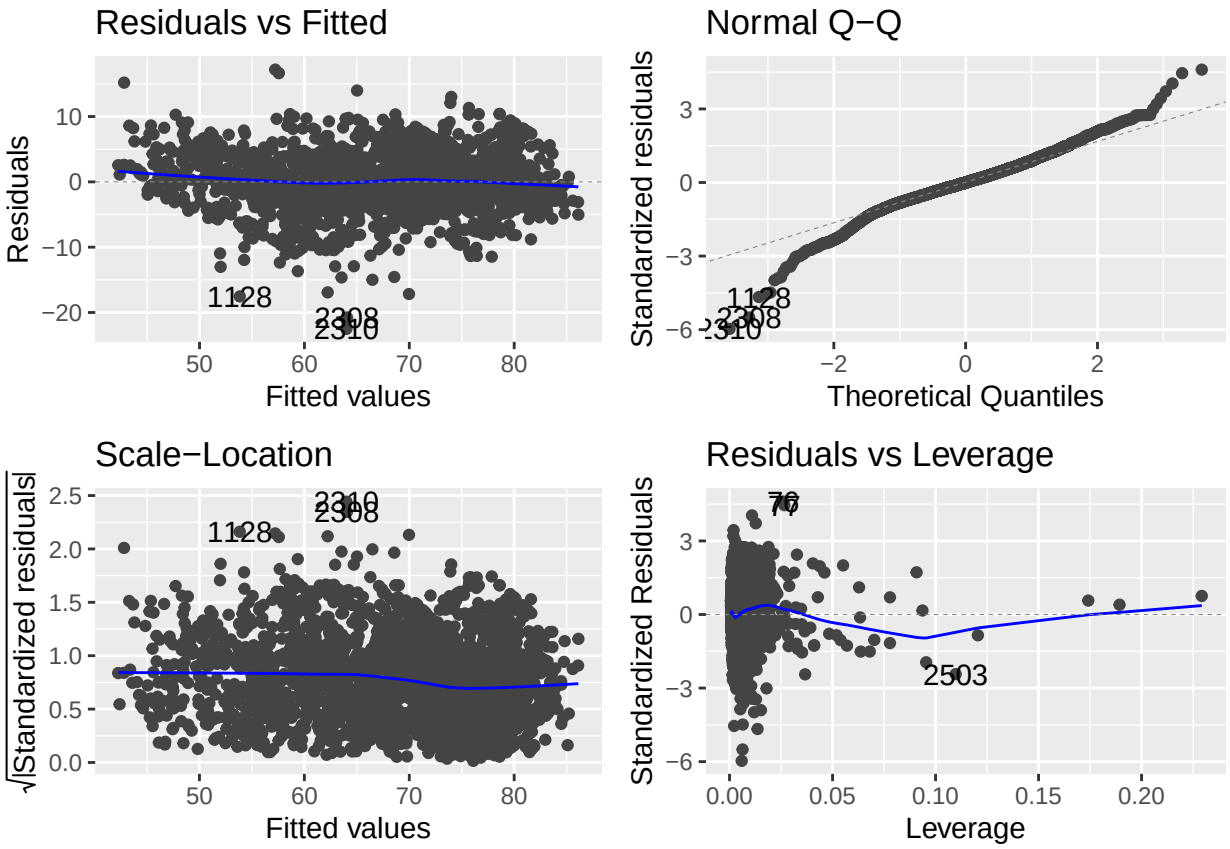
```r
model4 = lm(Life.expectancy~.-percentage.expenditure-thinness..1.19.years-thinness.5.9.years-Income.comp
summary(model4)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ . - percentage.expenditure - thinness..1.19.years -
##     thinness.5.9.years - Income.composition.of.resources - Population -
##     infant.deaths - under.five.deaths + Adult.Mortality:HIV.AIDS +
##     Schooling:Income.composition.of.resources + BMI:Schooling -
##     HIV.AIDS + poly(HIV.AIDS, 2), data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5067  -2.0497  -0.0117   2.1674  17.1798
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             5.218e+01  7.114e-01  73.349
## Adult.Mortality                        -2.359e-02  8.746e-04 -26.968
## Alcohol                                 1.314e-01  2.169e-02   6.057
## Hepatitis.B                            -1.116e-02  3.438e-03  -3.247
## Measles                                -3.221e-05  6.258e-06  -5.147
## BMI                                     1.943e-01  1.570e-02  12.379
## Polio                                   2.118e-02  4.166e-03   5.085
## Total.expenditure                       8.815e-02  3.103e-02   2.841
## Diphtheria                              3.566e-02  4.315e-03   8.264
## GDP                                     5.032e-05  6.252e-06   8.048
## Schooling                               8.188e-01  6.570e-02  12.464
## poly(HIV.AIDS, 2)1                     -2.414e+02  8.136e+00 -29.671
## poly(HIV.AIDS, 2)2                      6.078e+01  4.251e+00  14.297
## Adult.Mortality:HIV.AIDS                9.081e-04  6.515e-05  13.939
## Income.composition.of.resources:Schooling  5.342e-01  4.599e-02  11.616
## BMI:Schooling                          -1.292e-02  1.210e-03 -10.681
##                                         Pr(>|t|)
## (Intercept)                              < 2e-16 ***
## Adult.Mortality                          < 2e-16 ***
## Alcohol                                 1.57e-09 ***
## Hepatitis.B                              0.00118 **
## Measles                                 2.82e-07 ***
## BMI                                      < 2e-16 ***
## Polio                                   3.92e-07 ***
## Total.expenditure                        0.00453 **
## Diphtheria                               < 2e-16 ***
## GDP                                     1.21e-15 ***
## Schooling                                < 2e-16 ***
## poly(HIV.AIDS, 2)1                       < 2e-16 ***
## poly(HIV.AIDS, 2)2                       < 2e-16 ***
## Adult.Mortality:HIV.AIDS                 < 2e-16 ***
## Income.composition.of.resources:Schooling  < 2e-16 ***
## BMI:Schooling                            < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.782 on 2922 degrees of freedom
## Multiple R-squared:  0.8426, Adjusted R-squared:  0.8418
## F-statistic:  1043 on 15 and 2922 DF,  p-value: < 2.2e-16
```

Plotting The Model 4 :

```
library(ggfortify)
autoplot(model4)
```



Conclusion of the final model :

1). R^2 value of our final model is 84.18.
2). From the residual vs fitted graph we can see that the estimated error curve of our final model is almost converge to 0.
3). From the QQ-Plot we can see that the our model shows behave like normal except for the tail parts.
So our Final model is:

Life Expectancy $= (52.18) + (-0.0235) \times$ **Adult.Mortality** $+ (0.1314) \times$ **Alcohol** $+ (-0.01116) \times$ **Hepatitis.B** $+ (-0.00$

Summary :

- The dataset although collected by WHO contained a lot of missing values and we saw that most of the missing values were from the countries with very less population and were data collection is a very tedious task.
- A lot of outliers were detected which could not be removed because doing so we could have losed a lot of informations.

- Japan although being hit badly by world war II came back very strong and is currently the country with the highest life expectancy followed by Sweden which is a big Achievement.
- We largely saw how developing countries have very less life expectancy when we see diseases like HIV/AIDS, polio etc and how Schooling plays a big role in increasing the life expectancy of developing countries as people become much more educated and help improve the welfare and healthcare of the country along with economy.
- Alcoholism is a big issue in the developed country where people have good amount of money to spend and this shows how careless are people in terms of their health when it comes to alcoholism.
- Life Expectancy model is affected by the factors Adult mortality rate,GDP of the country, by the diseases named as Hepatitis B,Polio, Measles, Diphtheria,HIV/AIDS, and by some their factors too such as BMI, Alcoholism, Total expenditure, Schooling.