# DMPM Assignment 5

Name: Saniya S. Inamdar

SRN: 201900913

Roll no. : 17

CODE:

```
library(dplyr)
library(tidyverse)
library(janitor)
library(Hmisc)
library(caret)
library(reshape2)
library(caTools)
library(ggplot2)
library(scales)
library(Metrics)
lifedf <- read.csv("LifeExpectancyData.csv")
head(lifedf)
summary(lifedf)
glimpse(lifedf)
names(lifedf)

lifedf=clean_names(lifedf)
names(lifedf)

dim(lifedf)

na_count <-sapply(lifedf, function(y) sum(length(which(is.na(y)))))
na_count <- data.frame(na_count)
na_count


lifedf1 = lifedf %>%
  filter(!is.na(life_expectancy),
         !is.na(adult_mortality),
         !is.na(hepatitis_b),
         !is.na(bmi),
         !is.na(polio),
         !is.na(diphtheria),
         !is.na(hiv_aids),
         !is.na(total_expenditure),
         !is.na(thinness_1_19_years),
         !is.na(thinness_5_9_years),
         !is.na(alcohol),
         !is.na(income_composition_of_resources),
```

```
         !is.na(schooling),
         )
summary(lifedf1)
dim(lifedf1)

lifedf1$population = impute(lifedf1$population, fun = median) #
median imputation
lifedf1$gdp = impute(lifedf1$gdp, fun = median) # median imputation


na_count <-sapply(lifedf1, function(y) sum(length(which(is.na(y)))))
na_count <- data.frame(na_count)
na_count

process <- preProcess(as.data.frame(lifedf1), method=c("range"))

lifedf1 = predict(process, as.data.frame(lifedf1))

lifedf1$year

data<-cor(lifedf1[sapply(lifedf1, is.numeric)])
data=melt(data)
correlation = subset(data,data$Var2=="life_expectancy")
correlation


ggplot(data,aes(x = Var1, y = Var2,fill = value))+
  geom_tile()+
  theme(axis.text.x=element_text(angle=90))

#map
library(maps)
mapdata<-map_data("world")
glimpse(mapdata)
mapdata=left_join(lifedf1,mapdata,by="region")
glimpse(mapdata)

map1 = ggplot(mapdata,aes(x=long,y=lat,group=group))+
  geom_polygon(aes(fill=life_expectancy), color="black")
map1

map2 = map1+
scale_fill_gradient(name="LifeExpectancy",low="yellow",high="red",
na.value="grey50")+
  theme(axis.text.x=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks=element_blank(),
        axis.title.x=element_blank(),
        axis.title.y=element_blank(),
        )
```

```
map2

plot(x= lifedf1$schooling, y=lifedf1$life_expectancy,
xlab="Schooling", ylab="LifeExpectancy",
     main="Schooling vs Life Expectancy")
plot(x= lifedf1$hiv_aids, y=lifedf1$life_expectancy,
xlab="hiv_aids", ylab="LifeExpectancy",
     main="hiv_aids vs Life Expectancy")
plot(x= lifedf1$total_expenditure, y=lifedf1$life_expectancy,
xlab="Expenditure", ylab="LifeExpectancy",
     main="total_expenditure vs Life Expectancy")
plot(x= lifedf1$bmi, y=lifedf1$life_expectancy, xlab="BMI",
ylab="LifeExpectancy",
     main="BMI vs Life Expectancy")
plot(x= lifedf1$income_composition_of_resources,
y=lifedf1$life_expectancy, xlab="income_composition_of_resources",
ylab="LifeExpectancy",
     main="income_composition_of_resources vs Life Expectancy")
plot(x= lifedf1$gdp, y=lifedf1$life_expectancy, xlab="GDP",
ylab="LifeExpectancy",
     main="GDP vs Life Expectancy")
plot(x= lifedf1$adult_mortality, y=lifedf1$life_expectancy,
xlab="adult_mortality", ylab="LifeExpectancy",
     main="Adult Mortality vs Life Expectancy")


#split
set.seed(101)
sample = sample.split(lifedf1$life_expectancy, SplitRatio = .70)
train = subset(lifedf1, sample == TRUE)
test  = subset(lifedf1, sample == FALSE)
test_new = within(test, rm(life_expectancy))

#linear regression model
#country+year+infant_deaths+under_five_deaths+hiv_aids+thinness_5_9_
years
model1 = lm(life_expectancy~
region+adult_mortality+schooling+income_composition_of_resources+hiv
_aids , data = train )
summary(model1)

prob = model1 %>% predict(test_new)

test_new$predictedExpectancy = prob

x=cbind(test$life_expectancy,prob)
x=data.matrix(x)
x=rescale(x)
x=as.data.frame(x)
mae=mae(x$V1,x$prob)
mse=mse(x$V1,x$prob)
```

```
rmse=rmse(x$V1,x$prob)
cat("\nMAE:",mae,"\n\nMSE:",mse,"\n\nRMSE:",rmse,"\n\n")

resid = resid(model1)

plot(train$life_expectancy,resid,
     main = "Residual Plot(Schooling and life expectancy)",
     abline(0,0), ylab = "Residuals", xlab
     = "Age(in years)")
```

```
> lifedf <- read.csv("LifeExpectancyData.csv")
> head(lifedf)
       region Year     Status Life.expectancy Adult.Mortality infant.deaths Alcohol
1 Afghanistan 2015 Developing            65.0             263           62    0.01
2 Afghanistan 2014 Developing            59.9             271           64    0.01
3 Afghanistan 2013 Developing            59.9             268           66    0.01
4 Afghanistan 2012 Developing            59.5             272           69    0.01
5 Afghanistan 2011 Developing            59.2             275           71    0.01
6 Afghanistan 2010 Developing            58.8             279           74    0.01
  percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths Polio Total.expenditure
1              71.279624         65    1154 19.1                83     6               8.16
2              73.523582         62     492 18.6                86    58               8.18
3              73.219243         64     430 18.1                89    62               8.13
4              78.184215         67    2787 17.6                93    67               8.52
5               7.097109         68    3013 17.2                97    68               7.87
6              79.679367         66    1989 16.7               102    66               9.20
  Diphtheria HIV.AIDS       GDP Population thinness..1.19.years thinness.5.9.years
1         65      0.1 584.25921   33736494                 17.2                17.3
2         62      0.1 612.69651     327582                 17.5                17.5
3         64      0.1 631.74498   31731688                 17.7                17.7
4         67      0.1 669.95900    3696958                 17.9                18.0
5         68      0.1  63.53723    2978599                 18.2                18.2
6         66      0.1 553.32894    2883167                 18.4                18.4
  Income.composition.of.resources Schooling
1                           0.479      10.1
2                           0.476      10.0
3                           0.470       9.9
4                           0.463       9.8
5                           0.454       9.5
6                           0.448       9.2
```

**There are 22 columns in total.**

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

**There are missing values in the dataset.**

```
> glimpse(lifedf)
Rows: 2,938
Columns: 22
$ region                         <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanist~
$ Year                           <int> 2015, 2014, 2013, 2012, 2011, 2010, 2009, 2008, 2007, 2~
$ Status                         <chr> "Developing", "Developing", "Developing", "Developing",~
$ Life.expectancy                <dbl> 65.0, 59.9, 59.9, 59.5, 59.2, 58.8, 58.6, 58.1, 57.5, 5~
$ Adult.Mortality                <int> 263, 271, 268, 272, 275, 279, 281, 287, 295, 295, 291, ~
$ infant.deaths                  <int> 62, 64, 66, 69, 71, 74, 77, 80, 82, 84, 85, 87, 87, 88,~
$ Alcohol                        <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.03, 0.02, 0~
$ percentage.expenditure         <dbl> 71.279624, 73.523582, 73.219243, 78.184215, 7.097109, 7~
$ Hepatitis.B                    <int> 65, 62, 64, 67, 68, 66, 63, 64, 63, 64, 66, 67, 65, 64,~
$ Measles                        <int> 1154, 492, 430, 2787, 3013, 1989, 2861, 1599, 1141, 199~
$ BMI                            <dbl> 19.1, 18.6, 18.1, 17.6, 17.2, 16.7, 16.2, 15.7, 15.2, 1~
$ under.five.deaths              <int> 83, 86, 89, 93, 97, 102, 106, 110, 113, 116, 118, 120, ~
$ Polio                          <int> 6, 58, 62, 67, 68, 66, 63, 64, 63, 58, 58, 5, 41, 36, 3~
$ Total.expenditure              <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20, 9.42, 8.33, 6.73, 7~
$ Diphtheria                     <int> 65, 62, 64, 67, 68, 66, 63, 64, 63, 58, 58, 5, 41, 36, ~
$ HIV.AIDS                       <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, ~
$ GDP                            <dbl> 584.25921, 612.69651, 631.74498, 669.95900, 63.53723, 5~
$ Population                     <dbl> 33736494, 327582, 31731688, 3696958, 2978599, 2883167, ~
$ thinness..1.19.years           <dbl> 17.2, 17.5, 17.7, 17.9, 18.2, 18.4, 18.6, 18.8, 19.0, 1~
$ thinness.5.9.years             <dbl> 17.3, 17.5, 17.7, 18.0, 18.2, 18.4, 18.7, 18.9, 19.1, 1~
$ Income.composition.of.resources <dbl> 0.479, 0.476, 0.470, 0.463, 0.454, 0.448, 0.434, 0.433,~
$ Schooling                      <dbl> 10.1, 10.0, 9.9, 9.8, 9.5, 9.2, 8.9, 8.7, 8.4, 8.1, 7.9~
```

**There are 3 categorical variables: region, year, status.**

```
> names(lifedf)
 [1] "region"                         "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"
> lifedf=clean_names(lifedf)
> names(lifedf)
 [1] "region"                         "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"
> dim(lifedf)
[1] 2938   22
```

**So I have cleaned the names of the column, made it more, writable and compact.**

**The dataset has 2938 rows and 22 columns.**

```
> na_count
                                 na_count
region                                  0
year                                    0
status                                  0
life_expectancy                        10
adult_mortality                        10
infant_deaths                           0
alcohol                               194
percentage_expenditure                  0
hepatitis_b                           553
measles                                 0
bmi                                    34
under_five_deaths                       0
polio                                  19
total_expenditure                     226
diphtheria                             19
hiv_aids                                0
gdp                                   448
population                            652
thinness_1_19_years                    34
thinness_5_9_years                     34
income_composition_of_resources       167
schooling                             163
```

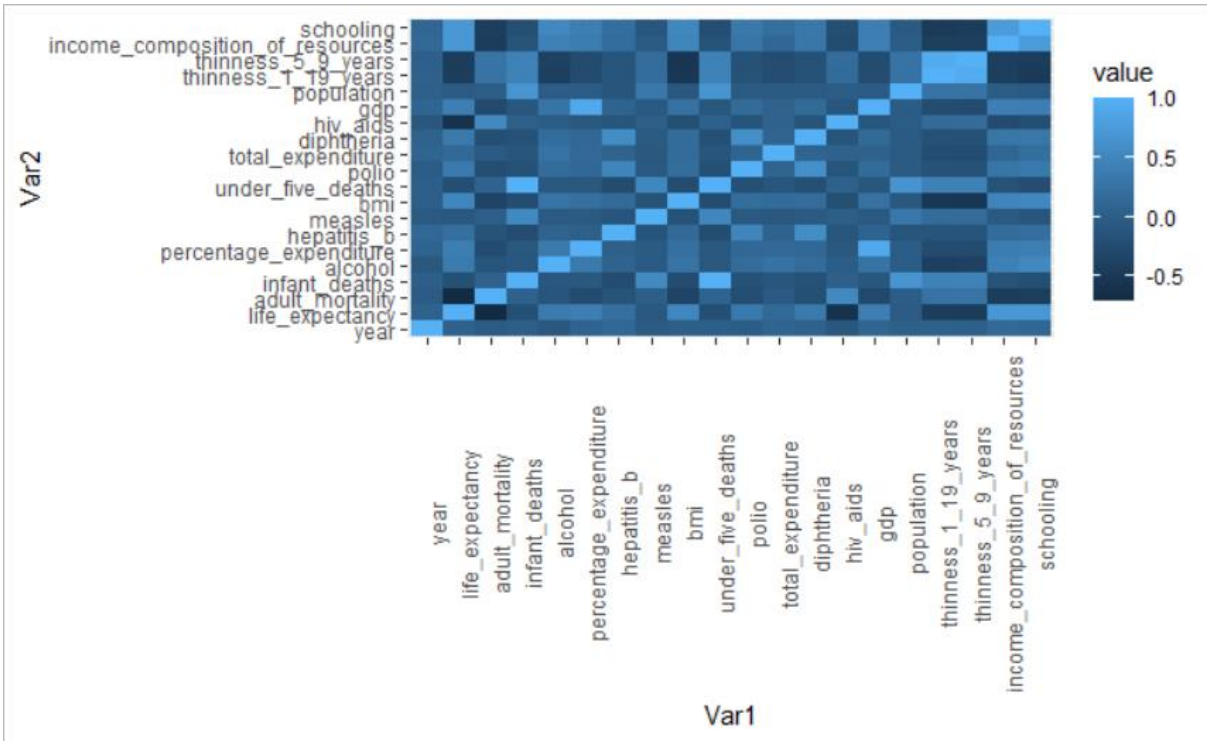
**I have dropped the columns that have na values less than 100 and imputed the others**

```
> na_count
                                 na_count
region                                  0
year                                    0
status                                  0
life_expectancy                         0
adult_mortality                         0
infant_deaths                           0
alcohol                                 0
percentage_expenditure                  0
hepatitis_b                             0
measles                                 0
bmi                                     0
under_five_deaths                       0
polio                                   0
total_expenditure                       0
diphtheria                              0
hiv_aids                                0
gdp                                     0
population                              0
thinness_1_19_years                     0
thinness_5_9_years                      0
income_composition_of_resources         0
schooling                               0
```

**Then I performed MinMAx scaling on the dataset.**

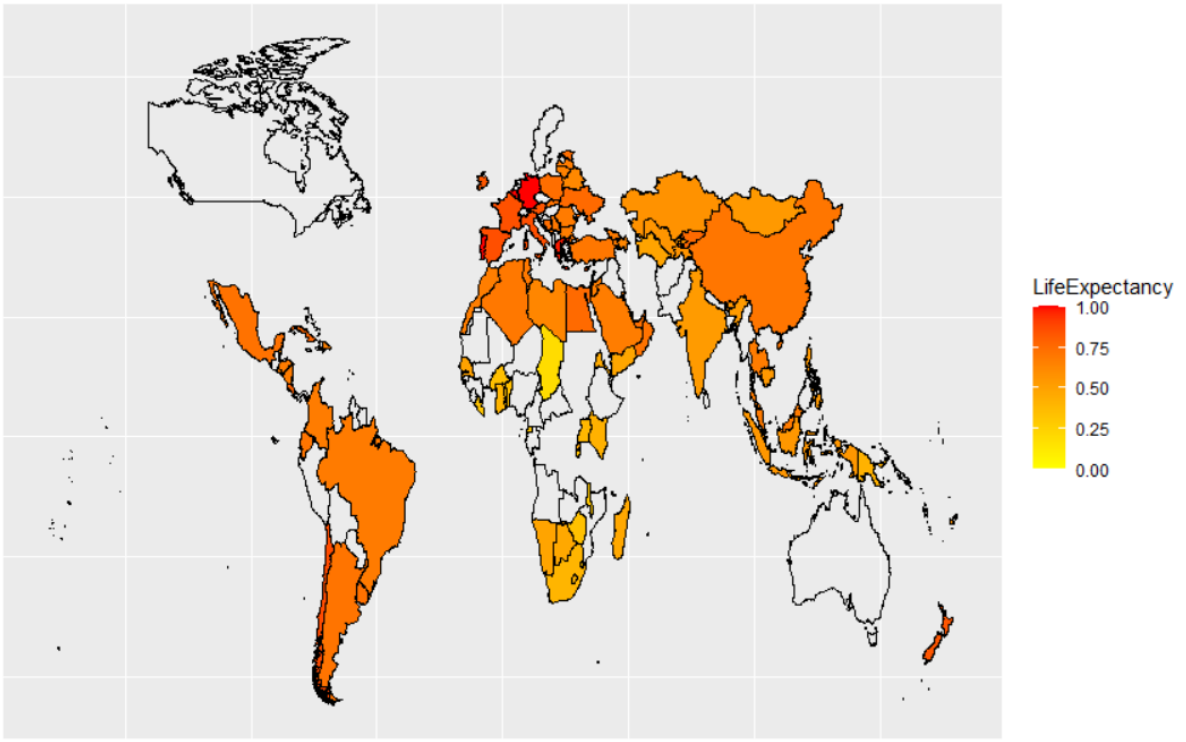
**And calculated the correlations between the features. Below are the results:**

```
> correlation
                              Var1            Var2       value
21                            year life_expectancy  0.05245074
22                 life_expectancy life_expectancy  1.00000000
23                 adult_mortality life_expectancy -0.70359802
24                   infant_deaths life_expectancy -0.17875545
25                         alcohol life_expectancy  0.35300388
26           percentage_expenditure life_expectancy  0.40299865
27                      hepatitis_b life_expectancy  0.22761994
28                          measles life_expectancy -0.08007076
29                              bmi life_expectancy  0.51081647
30               under_five_deaths life_expectancy -0.20102090
31                           polio life_expectancy  0.34839590
32               total_expenditure life_expectancy  0.12946704
33                      diphtheria life_expectancy  0.34399007
34                        hiv_aids life_expectancy -0.58169320
35                             gdp life_expectancy  0.42544890
36                      population life_expectancy -0.03485932
37             thinness_1_19_years life_expectancy -0.43659168
38              thinness_5_9_years life_expectancy -0.43770325
39 income_composition_of_resources life_expectancy  0.69210656
40                        schooling life_expectancy  0.70351805
```

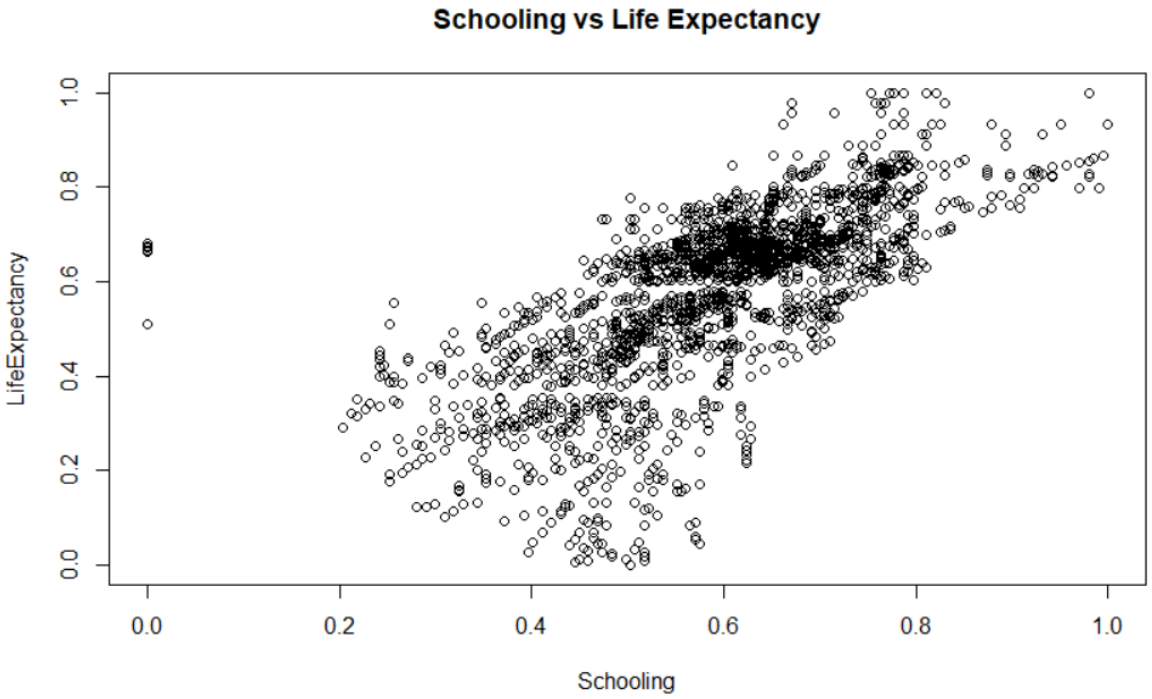


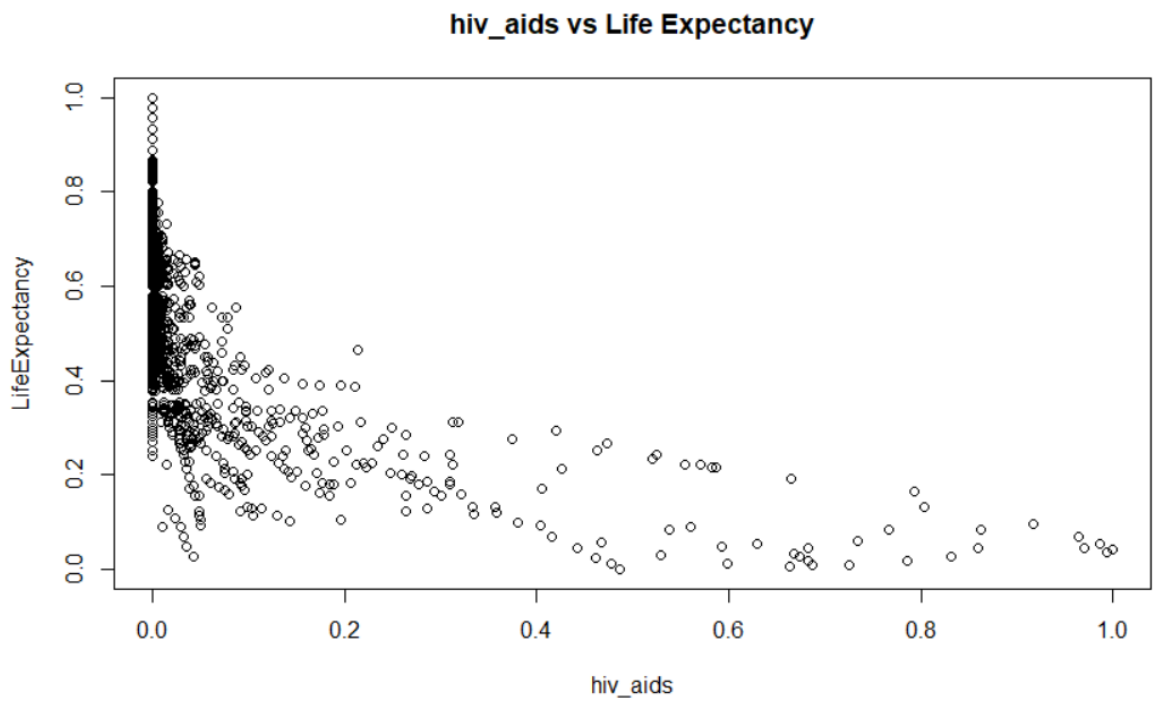

**In order to create geographical plots I imported a wolrd map data, and joined it by region on my original dataset.**

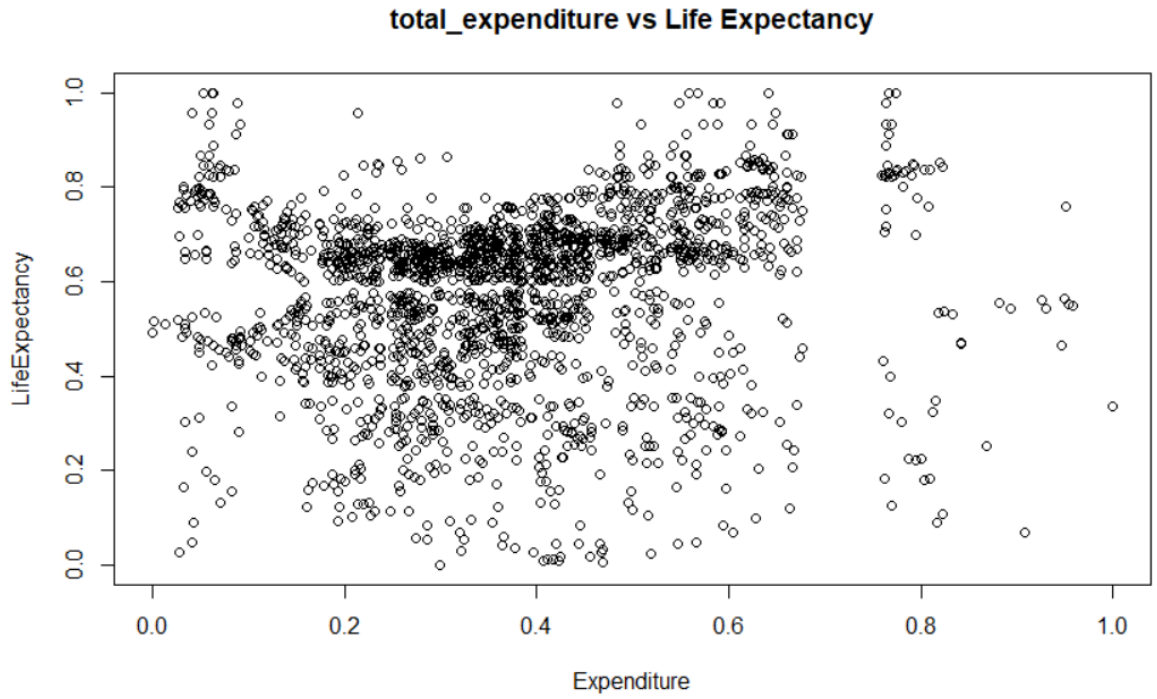
```
> library(maps)
> mapdata<-map_data("world")
> glimpse(mapdata)
Rows: 99,338
Columns: 6
$ long      <dbl> -69.89912, -69.89571, -69.94219, -70.00415, -70.06612, -70.05088, -70.03511, ~
$ lat       <dbl> 12.45200, 12.42300, 12.43853, 12.50049, 12.54697, 12.59707, 12.61411, 12.5676~
$ group     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,~
$ order     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23~
$ region    <chr> "Aruba", "Aruba", "Aruba", "Aruba", "Aruba", "Aruba", "Aruba", "Aruba", "Arub~
$ subregion <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
> mapdata=left_join(lifedf1,mapdata,by="region")
```

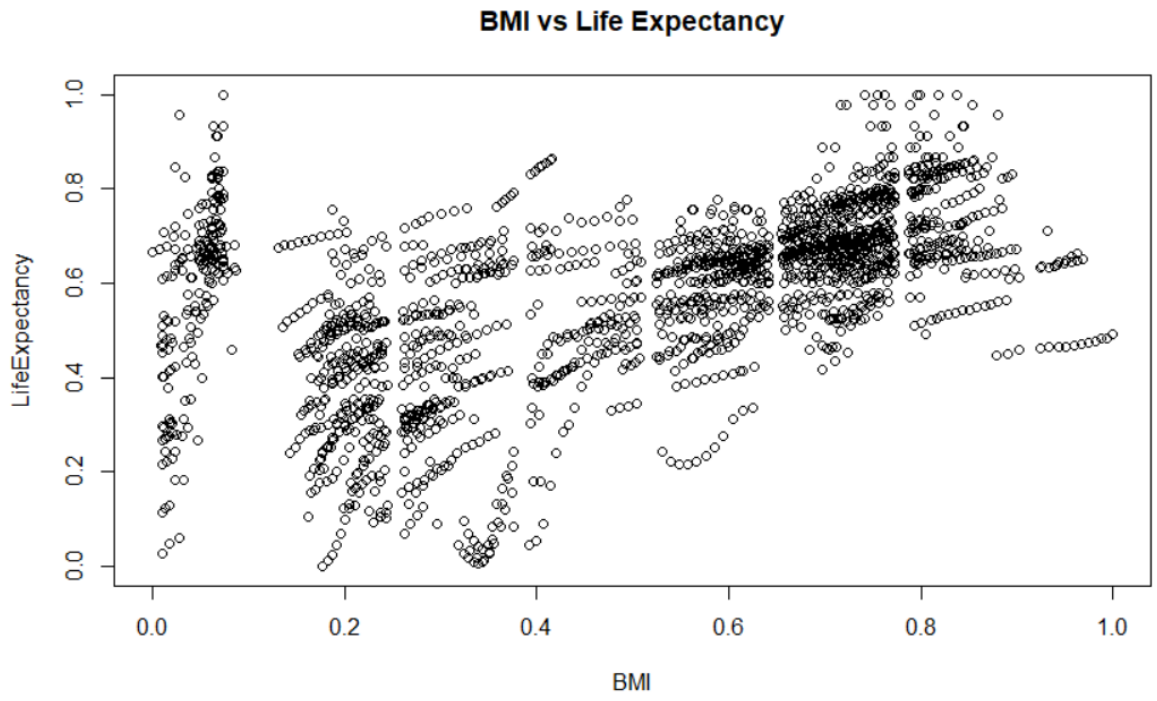


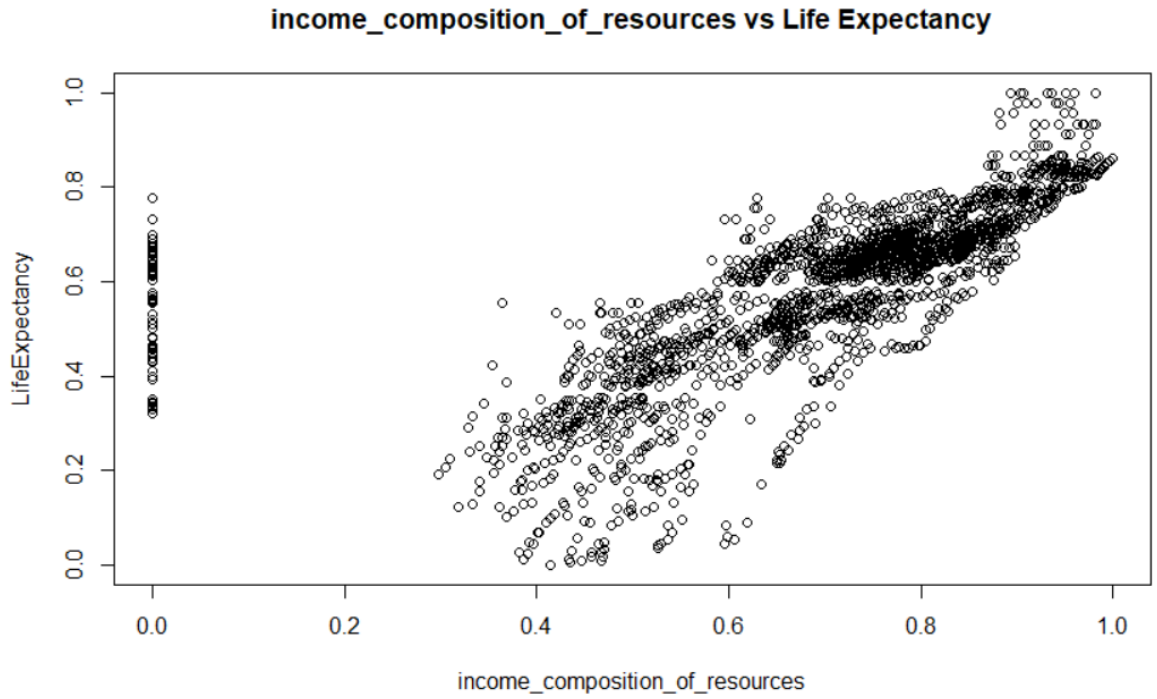

**Then I made some scatter plot to identify the relations between some variables:**

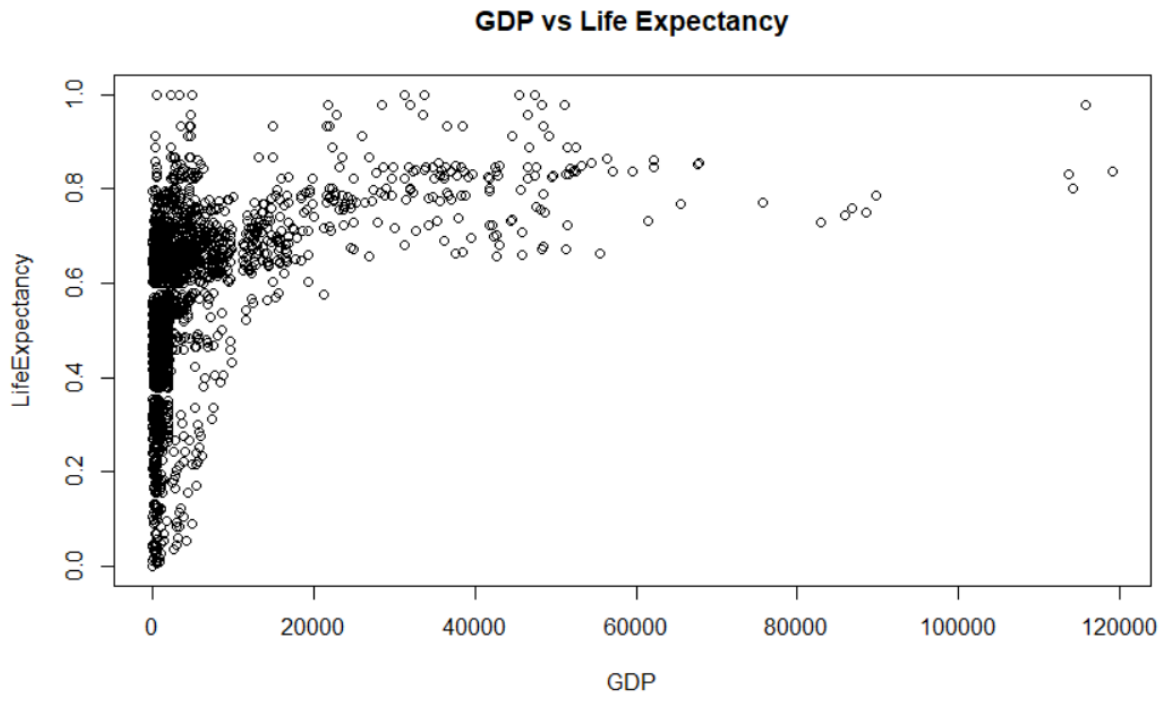



hiv_aids vs Life Expectancy
LifeExpectancy
0.0
0.2
0.4
0.6
0.8
1.0
hiv_aids


total_expenditure vs Life Expectancy
LifeExpectancy
0.0
0.2
0.4
0.6
0.8
1.0
Expenditure


BMI vs Life Expectancy
LifeExpectancy
0.0
0.2
0.4
0.6
0.8
1.0
BMI


income_composition_of_resources vs Life Expectancy
LifeExpectancy
0.0
0.2
0.4
0.6
0.8
1.0
income_composition_of_resources





## Trained the model on significant features:

```
> model1 = lm(life_expectancy~ region+adult_mortality+schooling+income_composition_of_resources+hiv_aids , data = trai
n )
> summary(model1)

Call:
lm(formula = life_expectancy ~ region + adult_mortality + schooling +
    income_composition_of_resources + hiv_aids, data = train)

Residuals:
      Min        1Q    Median        3Q       Max
-0.217048 -0.018669 -0.002413  0.011465  0.203620
```

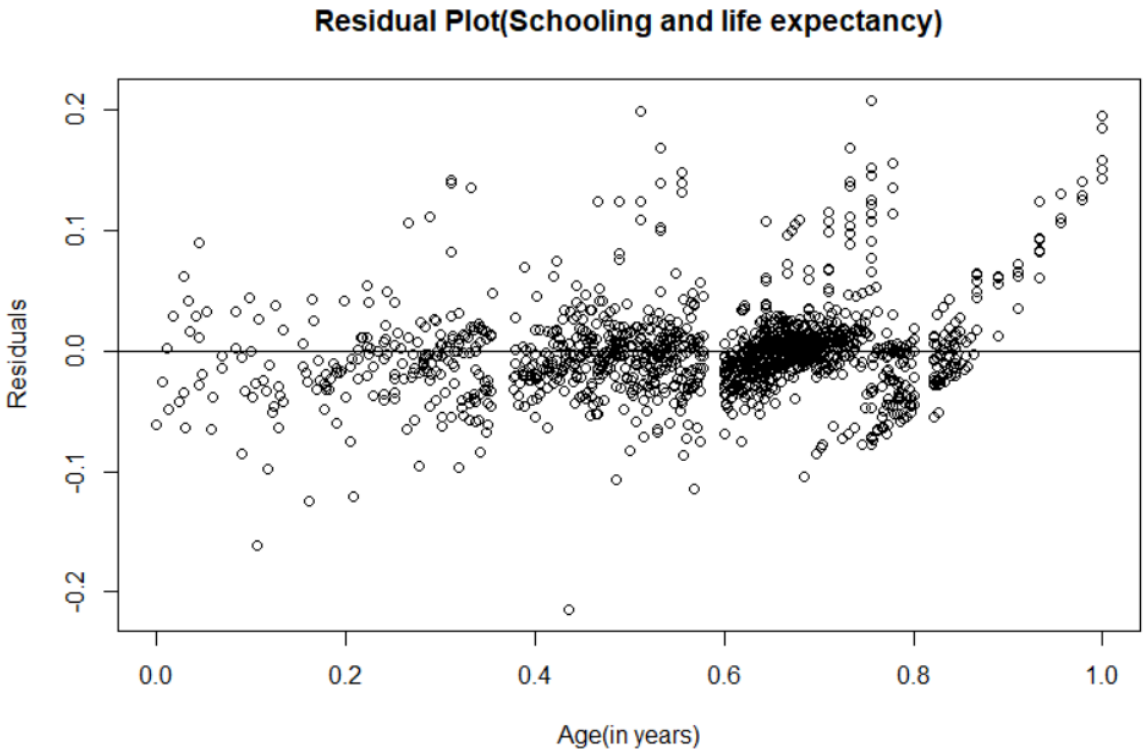
**Results:**

```
MAE: 0.02835596

MSE: 0.001994507

RMSE: 0.04465991
```

**Residual Plot:**




END