Life Expectancy Project

Borruso William Joseph - 902073 , Galli Luca - 905236, Ronchi Davide - 903320

2022-11-01

Contents

1 Prediction of countries' life expectancy

The aim of this study is to build a model capable of predicting the *Life Expectancy* starting from given attributes. The dataset presents statistics about different countries for every year from 2000 to 2015.

Libraries installation and dataset importing:

```
# install.packages("ggcorrplot")
# install.packages("ggplot2")
# install.packages("magrittr")
# install.packages("tidyr")
# install.packages("utils")
# install.packages("nortest")
# install.packages("lmtest")
#install.packages("moments")
library(moments)
library(ggcorrplot)
library(ggplot2)
library(magrittr)
library(tidyr)
library(utils)
library(nortest)
library(lmtest)

directory<- getwd()
setwd(directory)

lifeexp <- read.csv("Life Expectancy Data.csv", header = TRUE, sep = ",", stringsAsFactors = TRUE)
```

Forcing the insertion of NAs in place of missing values, we can then count how many missing values are present for every column.

```
matrice <- data.matrix(lifeexp, rownames.force = NA)

summary(matrice)
```

```
##     Country           Year          Status      Life.expectancy Adult.Mortality
##  Min.   :  1.0   Min.   :2000   Min.   :1.000   Min.   :36.30   Min.   :  1.0
##  1st Qu.: 47.0   1st Qu.:2004   1st Qu.:2.000   1st Qu.:63.10   1st Qu.: 74.0
##  Median : 94.0   Median :2008   Median :2.000   Median :72.10   Median :144.0
##  Mean   : 96.1   Mean   :2008   Mean   :1.826   Mean   :69.22   Mean   :164.8
##  3rd Qu.:146.0   3rd Qu.:2012   3rd Qu.:2.000   3rd Qu.:75.70   3rd Qu.:228.0
##  Max.   :193.0   Max.   :2015   Max.   :2.000   Max.   :89.00   Max.   :723.0
##                                                 NA's   :10      NA's   :10
##  infant.deaths       Alcohol        percentage.expenditure  Hepatitis.B
##  Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000      Min.   : 1.00
##  1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685      1st Qu.:77.00
##  Median :   3.0   Median : 3.7550   Median :   64.913      Median :92.00
##  Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251      Mean   :80.94
##  3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534      3rd Qu.:97.00
##  Max.   :1800.0   Max.   :17.8700   Max.   :19479.912      Max.   :99.00
##                   NA's   :194                              NA's   :553
##     Measles               BMI        under.five.deaths     Polio
##  Min.   :     0.0   Min.   : 1.00   Min.   :   0.00   Min.   : 3.00
##  1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00   1st Qu.:78.00
##  Median :    17.0   Median :43.50   Median :   4.00   Median :93.00
##  Mean   :  2419.6   Mean   :38.32   Mean   :  42.04   Mean   :82.55
##  3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00   3rd Qu.:97.00
##  Max.   :212183.0   Max.   :87.30   Max.   :2500.00   Max.   :99.00
##                     NA's   :34                        NA's   :19
##  Total.expenditure   Diphtheria       HIV.AIDS           GDP
##  Min.   : 0.370    Min.   : 2.00   Min.   : 0.100   Min.   :     1.68
##  1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   463.94
##  Median : 5.755    Median :93.00   Median : 0.100   Median :  1766.95
##  Mean   : 5.938    Mean   :82.32   Mean   : 1.742   Mean   :  7483.16
##  3rd Qu.: 7.492    3rd Qu.:97.00   3rd Qu.: 0.800   3rd Qu.:  5910.81
##  Max.   :17.600    Max.   :99.00   Max.   :50.600   Max.   :119172.74
##  NA's   :226       NA's   :19                       NA's   :448
##    Population        thinness..1.19.years thinness.5.9.years
##  Min.   :3.400e+01   Min.   : 0.10        Min.   : 0.10
##  1st Qu.:1.958e+05   1st Qu.: 1.60        1st Qu.: 1.50
##  Median :1.387e+06   Median : 3.30        Median : 3.30
##  Mean   :1.275e+07   Mean   : 4.84        Mean   : 4.87
##  3rd Qu.:7.420e+06   3rd Qu.: 7.20        3rd Qu.: 7.20
##  Max.   :1.294e+09   Max.   :27.70        Max.   :28.60
##  NA's   :652         NA's   :34           NA's   :34
##  Income.composition.of.resources   Schooling
##  Min.   :0.0000                  Min.   : 0.00
##  1st Qu.:0.4930                  1st Qu.:10.10
##  Median :0.6770                  Median :12.30
##  Mean   :0.6276                  Mean   :11.99
##  3rd Qu.:0.7790                  3rd Qu.:14.30
##  Max.   :0.9480                  Max.   :20.70
```

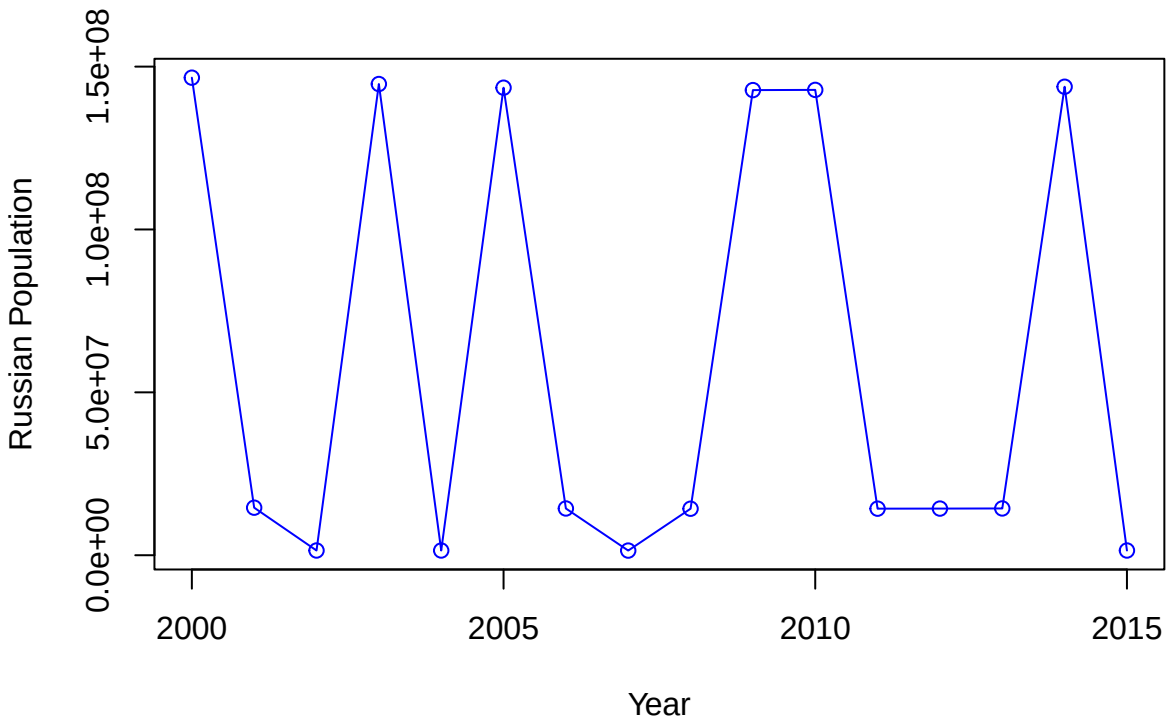
```
##  NA's   :167                     NA's   :163
```

Some columns do have a noticeable amount of missing values.

Furthermore we noticed that many variables present a lot of anomalies probably caused by bad data collection, which has been done using web scraping.

One of these is the *Population* variable: For example, in Russia we observed an abnormally high increase of population equal to 129.462.755 between 2013 and 2014, and then again a decrease of 142.369.979 by 2015. These anomalies were also observed in other years and other countries.

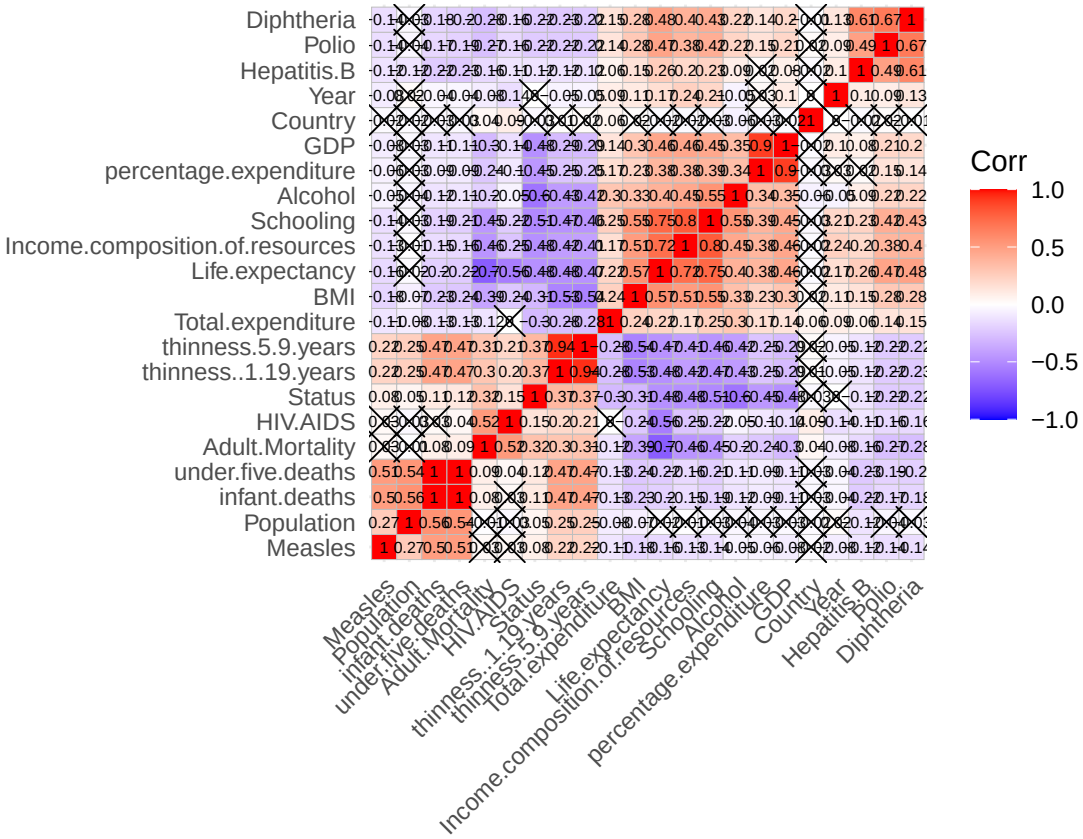
```
data2 <- subset (lifeexp, Country == "Russian Federation")

plot(data2$Year, data2$Population, , xlab = "Year", ylab = "Russian Population", type = "o", col = "blu
```



This led us to reduce the size of the dataset to only the meaningful variables for the analysis.

To do so, let's first see the correlations between all columns based only on complete couples (in other words, excluding couples that present NA values, by using the "Pairwise Complete Correlation") to identify which variables have a high correlation with *Life Expectancy*:

```
cormat_no_na <- round(cor(matrice, use = "pairwise.complete.obs"),2)

p.mat <- cor_pmat(matrice)

ggcorrplot(cormat_no_na, tl.cex = 9, hc.order = TRUE, p.mat = p.mat, lab = TRUE, lab_size = 2)
```



Let's analyse the variables with the highest positive correlation with *Life Expectancy*:

- *Schooling* (75%) - Number of years of Schooling. The high correlation shows how the average number of years spent in school can increase life expectancy;
- *Income composition of resources* (72%) - Human Development Index in terms of income composition of resources (index ranging from 0 to 1), which shows how productive resources are used. It's high correlation shows how life expectancy strongly increases if a country uses its resources productively;
- *BMI* (57%) - Average Body Mass Index of entire population, calculated from a person's weight and height;
- *Diphtheria* (48%) - Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-years-old (%);
- *Polio* (47%) - Polio (Pol3) immunization coverage among 1-years-old (%). This measure is correlated with Diphteria (67%). We assume that this correlation is due to the fact that some countries recognize both of these two vaccines as mandatory, like in Italy;
- *GDP* (46%) - Gross Domestic Product percapita (in USD), highly correlated with *Percentage Expenditure* (90%), which is the expenditure on health as a percentage of Gross Domestic Product percapita (%). We then assume that the more a country is wealthy and developed, the more it invests in healthcare.

Given these considerations we created a new dataset containing only the previously identified attributes:

```
rel_attr <- data.frame(lifeexp$Life.expectancy,
                       lifeexp$Schooling,
                       lifeexp$Income.composition.of.resources,
                       lifeexp$BMI,
                       lifeexp$Diphtheria,
```

```
                     lifeexp$Polio,
                     lifeexp$GDP)
colnames(rel_attr) = c("Life.expectancy",
                     "Schooling",
                     "Income.composition.of.resources",
                     "BMI",
                     "Diphtheria",
                     "Polio",
                     "GDP")
summary(rel_attr)
```

```
##  Life.expectancy   Schooling     Income.composition.of.resources
##  Min.   :36.30   Min.   : 0.00   Min.   :0.0000
##  1st Qu.:63.10   1st Qu.:10.10   1st Qu.:0.4930
##  Median :72.10   Median :12.30   Median :0.6770
##  Mean   :69.22   Mean   :11.99   Mean   :0.6276
##  3rd Qu.:75.70   3rd Qu.:14.30   3rd Qu.:0.7790
##  Max.   :89.00   Max.   :20.70   Max.   :0.9480
##  NA's   :10      NA's   :163     NA's   :167
##       BMI          Diphtheria        Polio            GDP
##  Min.   : 1.00   Min.   : 2.00   Min.   : 3.00   Min.   :     1.68
##  1st Qu.:19.30   1st Qu.:78.00   1st Qu.:78.00   1st Qu.:   463.94
##  Median :43.50   Median :93.00   Median :93.00   Median :  1766.95
##  Mean   :38.32   Mean   :82.32   Mean   :82.55   Mean   :  7483.16
##  3rd Qu.:56.20   3rd Qu.:97.00   3rd Qu.:97.00   3rd Qu.:  5910.81
##  Max.   :87.30   Max.   :99.00   Max.   :99.00   Max.   :119172.74
##  NA's   :34      NA's   :19      NA's   :19      NA's   :448
```

```
attach(rel_attr)
```

2 Observations and Data cleaning

First we started cleaning the dataset by removing NA values:

- We decided to remove the 19 rows that have null values from *Polio* and *Diphtheria*, which happen to be the same rows.

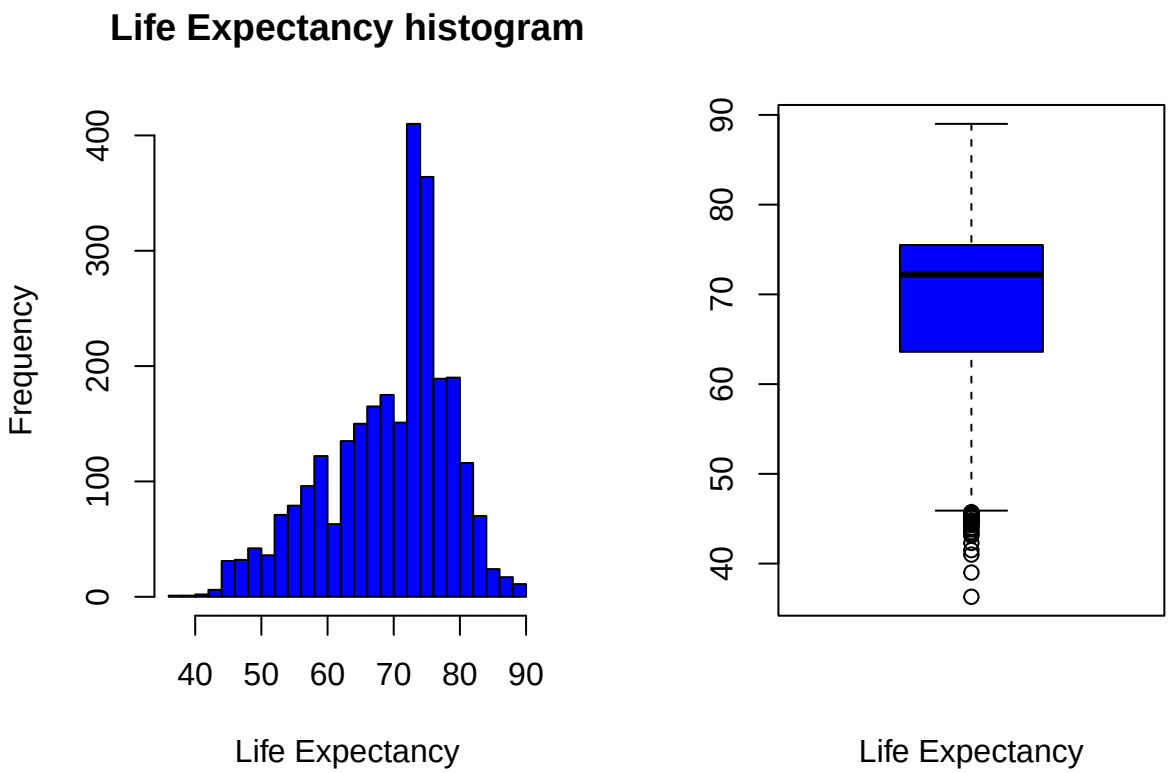
```
rel_attr <- rel_attr %>% drop_na(Diphtheria)
rel_attr <- rel_attr %>% drop_na(Polio)
```

- we dropped the rows that have null values in the *Life Expectancy* column, which will be our dependent variable, the *Schooling* column and the *Income.composition.of.resources* one.

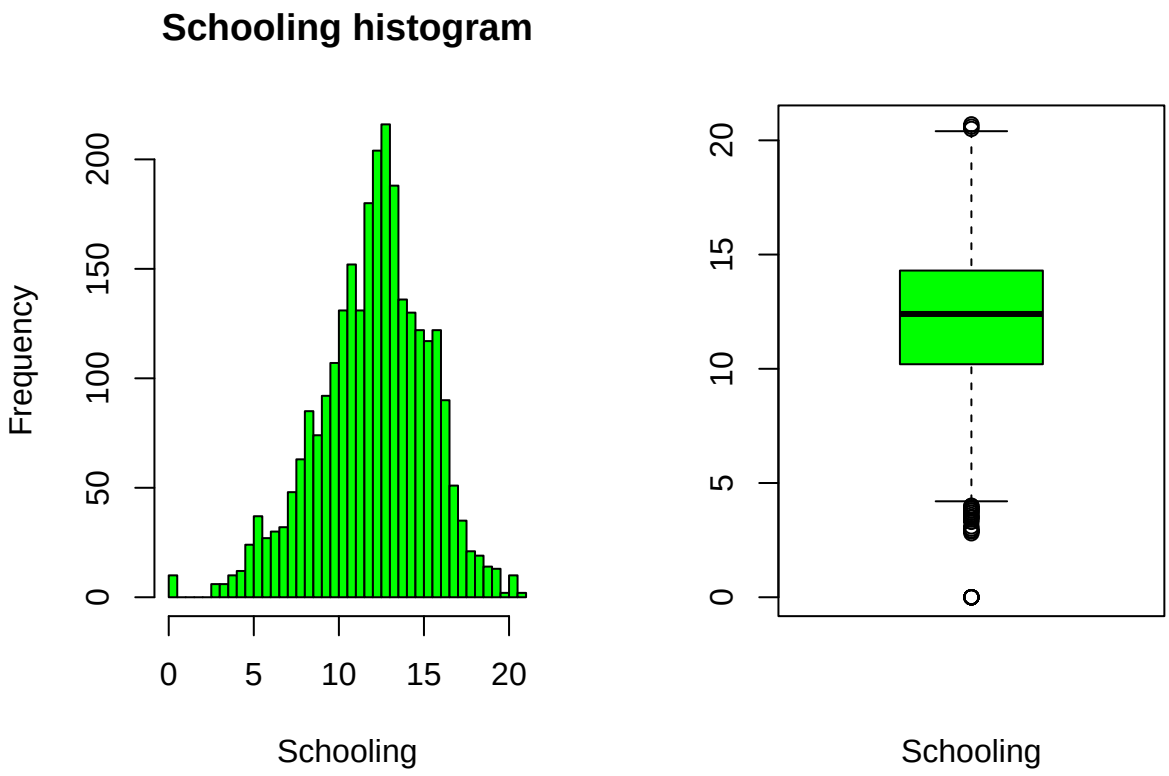
```
rel_attr <- rel_attr %>% drop_na(Life.expectancy)
rel_attr <- rel_attr %>% drop_na(Schooling)
rel_attr <- rel_attr %>% drop_na(Income.composition.of.resources)
```

Let's draw some histograms and box-plots to better visualize the selected variable

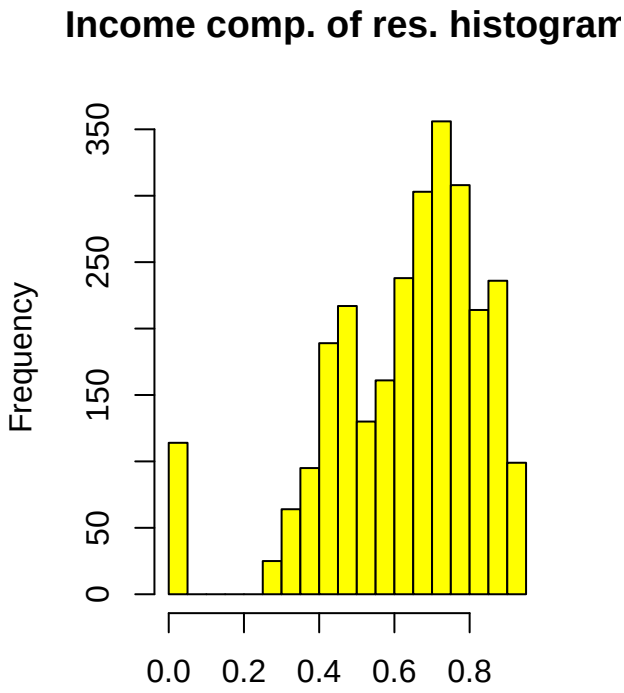
```
par(mfrow = c(1,2))
hist(rel_attr$Life.expectancy, breaks = "FD", main = "Life Expectancy histogram", xlab = "Life Expectanc
boxplot(rel_attr$Life.expectancy, xlab = "Life Expectancy", col="blue")
```

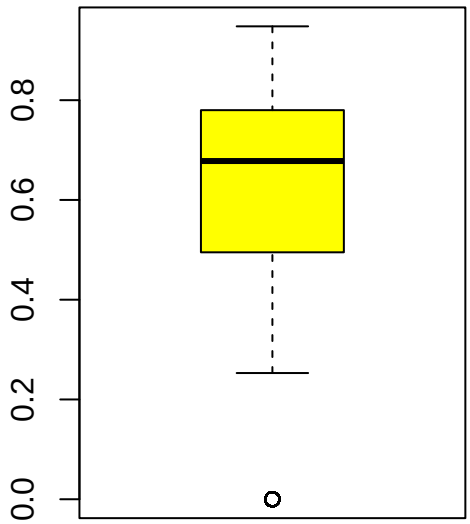



```
hist(rel_attr$Schooling, breaks = "FD", main = "Schooling histogram", xlab = "Schooling", col="green")
boxplot(rel_attr$Schooling, xlab = "Schooling", col="green")
```

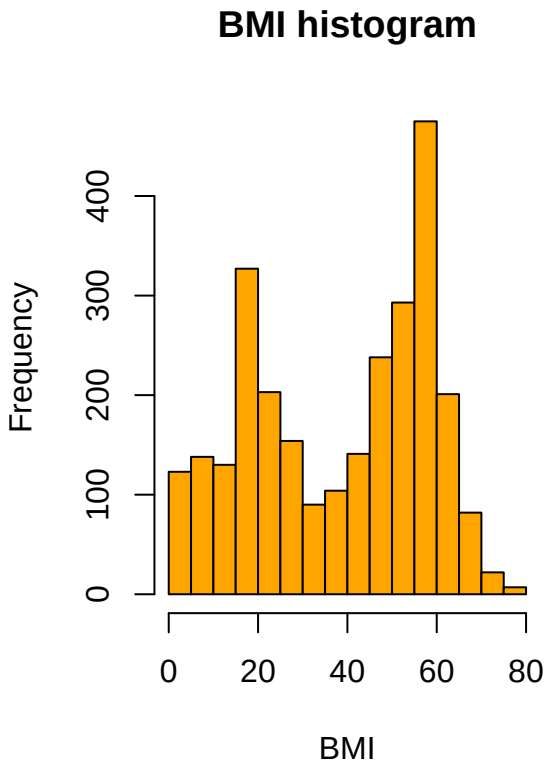


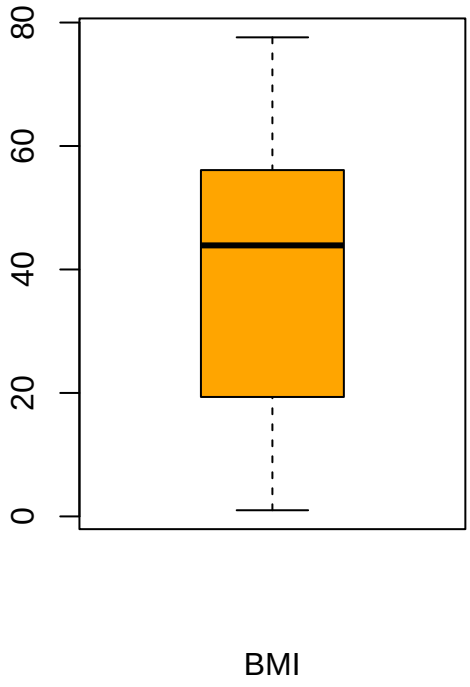

```
hist(rel_attr$Income.composition.of.resources, breaks = "FD", main = "Income comp. of res. histogram", x
boxplot(rel_attr$Income.composition.of.resources,xlab = "Income composition of resources", col="yellow")
```



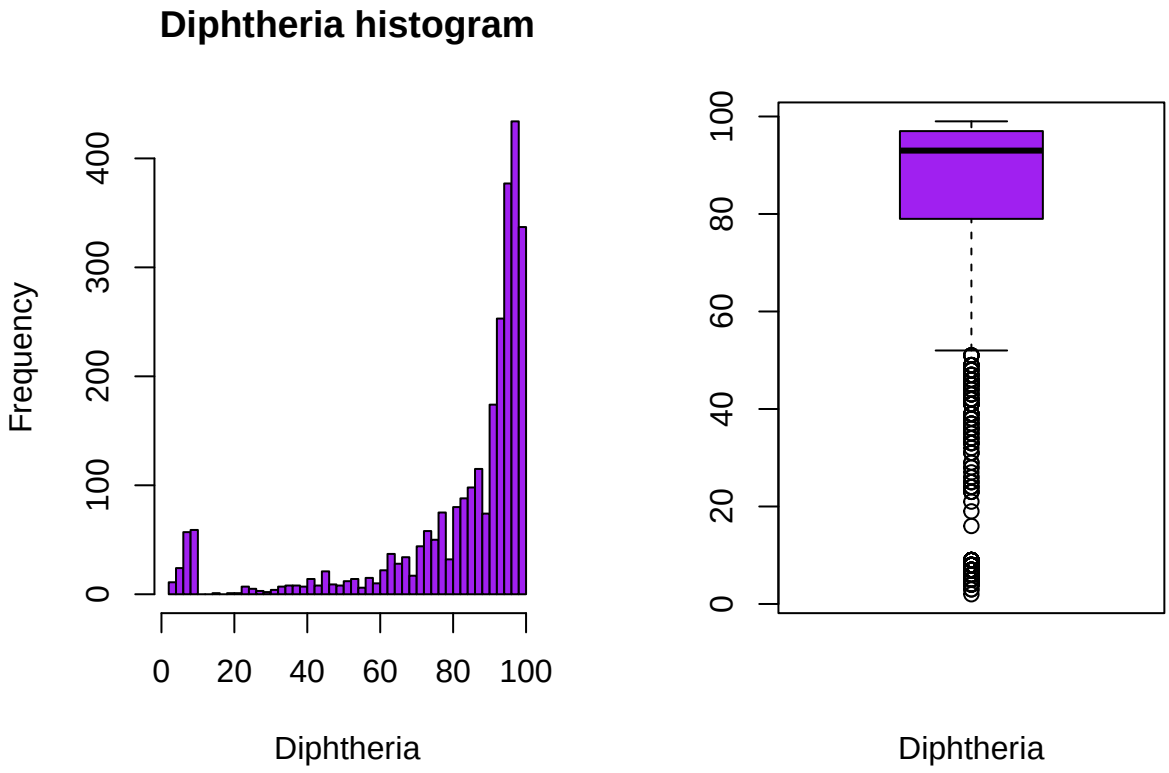



```
hist(rel_attr$BMI, breaks = "FD", main = "BMI histogram", xlab = "BMI", col="orange")
boxplot(rel_attr$BMI,xlab = "BMI", col="orange")
```

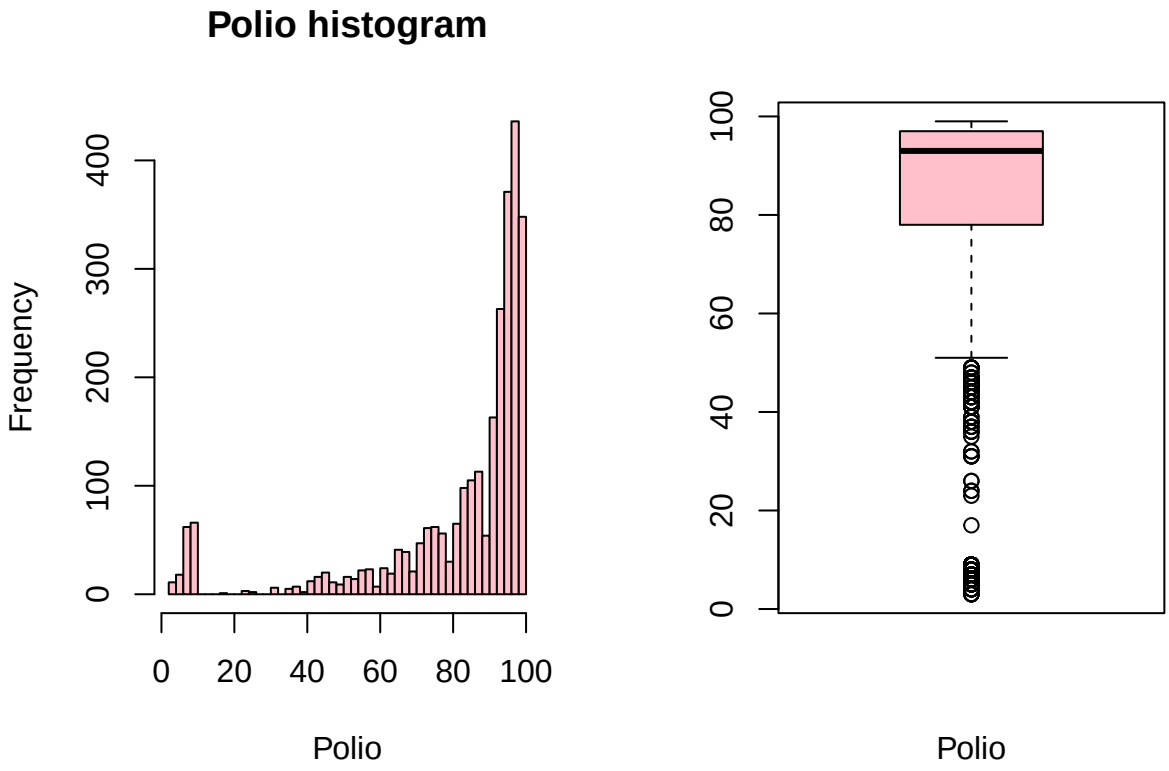




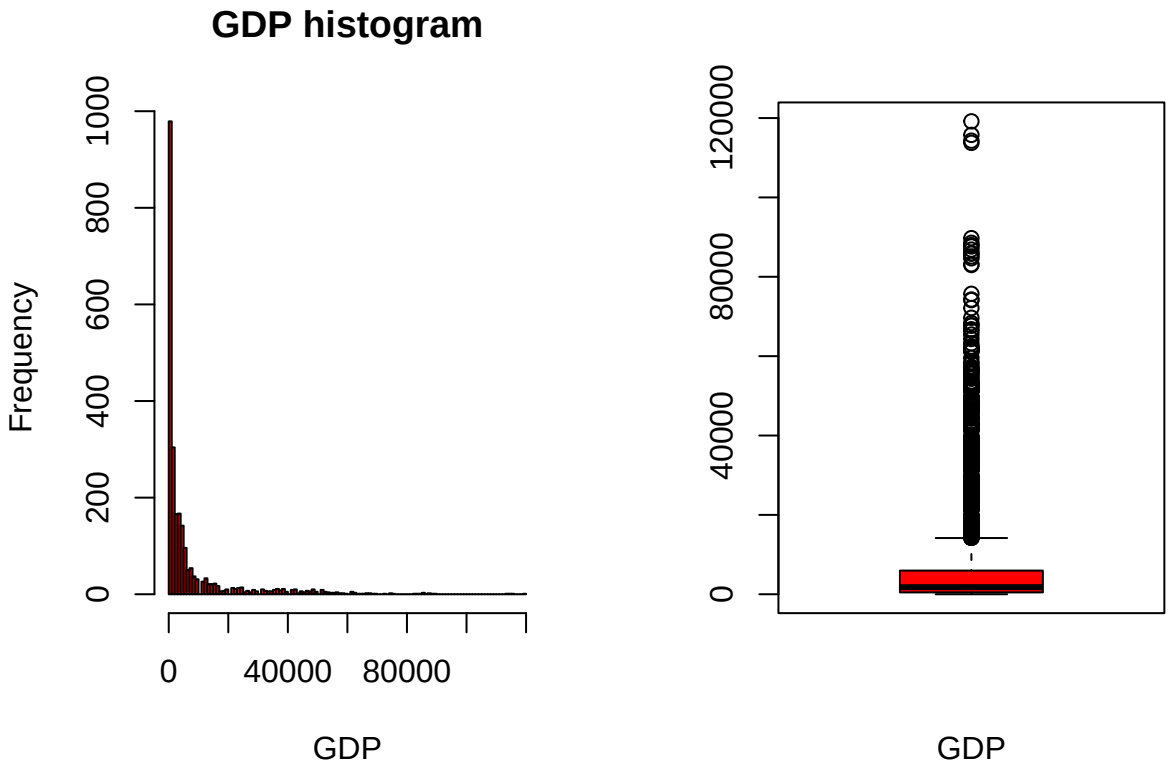

```
hist(rel_attr$Diphtheria, breaks = "FD", main = "Diphtheria histogram", xlab = "Diphtheria", col="purple
boxplot(rel_attr$Diphtheria, xlab = "Diphtheria", col="purple")
```




```
hist(rel_attr$Polio, breaks = "FD", main = "Polio histogram", xlab = "Polio", col="pink")
boxplot(rel_attr$Polio, xlab = "Polio", col="pink")
```




```
hist(rel_attr$GDP, breaks = "FD", main = "GDP histogram", xlab = "GDP", col="red")
boxplot(rel_attr$GDP, xlab = "GDP", col="red")
```



We observed that in the histograms for *Schooling* and *Income composition of resources* there are some 0 values, we assume due to lackluster data collection.

In particular, since it's unlikely that a country has an average of 0 years of schooling, we suspect that the dataset also used the value 0 to refer to null values in addition to NA. In absence of further proofs, we decided to exclude the records that have the *Schooling* value equal to 0 from the dataset.

```
rel_attr <- rel_attr[rel_attr$Schooling != 0, ]
```

Since the *Income composition of resources* is highly correlated with the variable *Schooling* and also contains lots of values equal to 0, probably related to the faulty representation of null values (as already seen in the *Schooling* variable), we decided to avoid using this variable for the regression model.

```
cols.dont.want <- c("Income.composition.of.resources")

rel_attr <- rel_attr[, ! names(rel_attr) %in% cols.dont.want, drop = F]
```

Since we decided to remove the *Population* column due to its anomalies, we also removed the *GDP* column, which is calculated using the *Population* variable and in fact presents a lot of discontinuity.

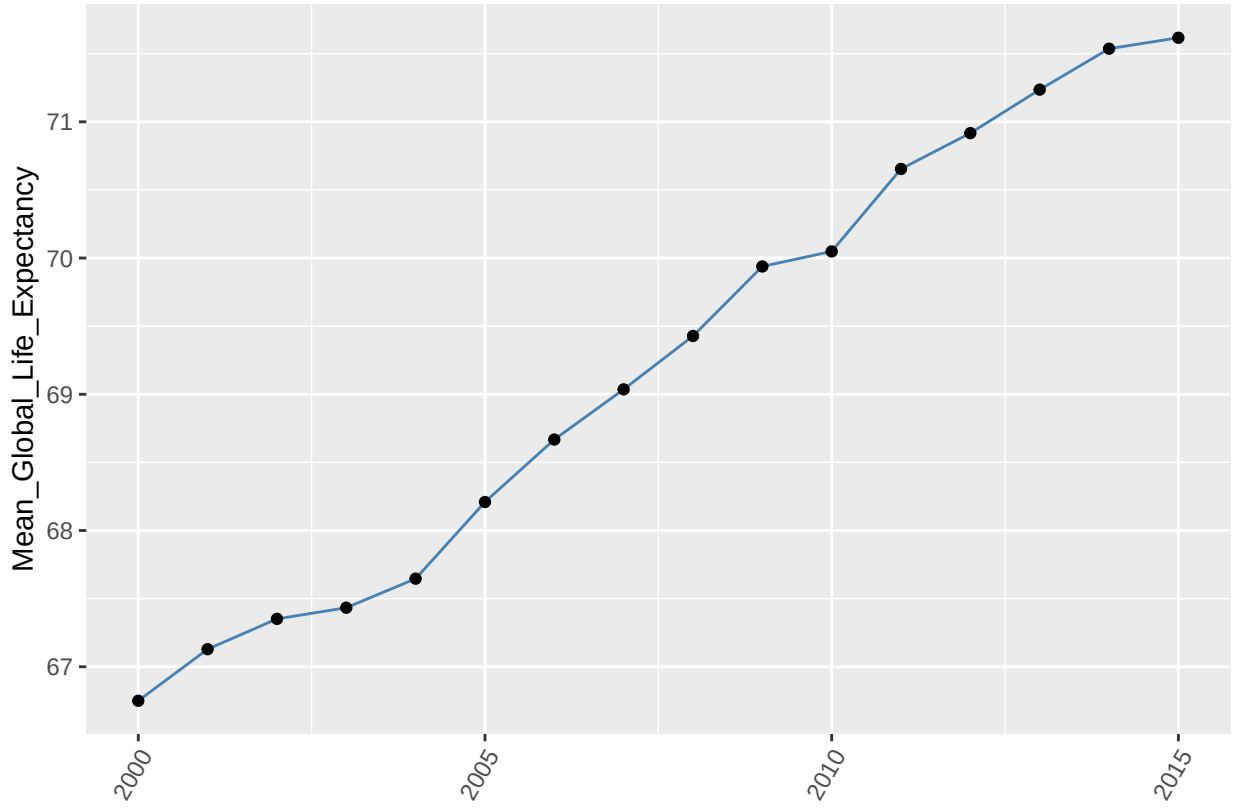
```
cols.dont.want <- c("GDP")

rel_attr <- rel_attr[, ! names(rel_attr) %in% cols.dont.want, drop = F]
```

```
rel_attr2 <- data.frame(lifeexp$Life.expectancy,
                        lifeexp$Schooling,
                        lifeexp$Income.composition.of.resources,
                        lifeexp$BMI,
                        lifeexp$Diphtheria,
                        lifeexp$Polio,
                        lifeexp$GDP,
                        lifeexp$Status,
                        lifeexp$Year)
colnames(rel_attr2) = c("Life.expectancy",
                        "Schooling",
                        "Income.composition.of.resources",
                        "BMI",
                        "Diphtheria",
                        "Polio",
                        "GDP",
                        "Status",
                        "Year")

rel_attr2 <- rel_attr2 %>% drop_na(Life.expectancy)
```

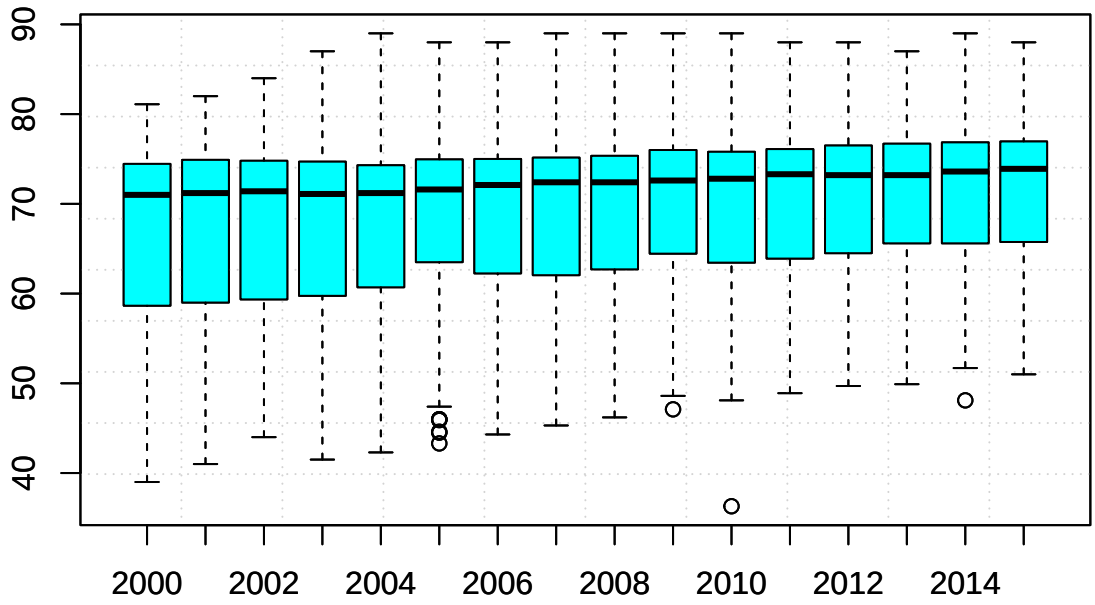
Once cleaned the dataset we analyzed our dependent variable (Life.expectancy).

In particular we plotted a timeline and a box-plot to show its growth over each year.

```
years <- seq(from = 2000, to= 2015)
mean_LE_PerYear <- rep(0,15)
i <- 1
for (anno in years) {
  mean_LE_PerYear[i] = mean(rel_attr2$Life.expectancy[rel_attr2$Year == anno])
  i <- i+1
}
datafr <- data.frame(years, mean_LE_PerYear)
colnames(datafr) <- c('Year',
                      'Mean_Global_Life_Expectancy')

p_LE_ts <- ggplot(datafr, aes(x=Year, y=Mean_Global_Life_Expectancy)) +
  geom_line( color="steelblue") +
  geom_point() +
  xlab("") +
  theme(axis.text.x=element_text(angle=60, hjust=1))
p_LE_ts
```

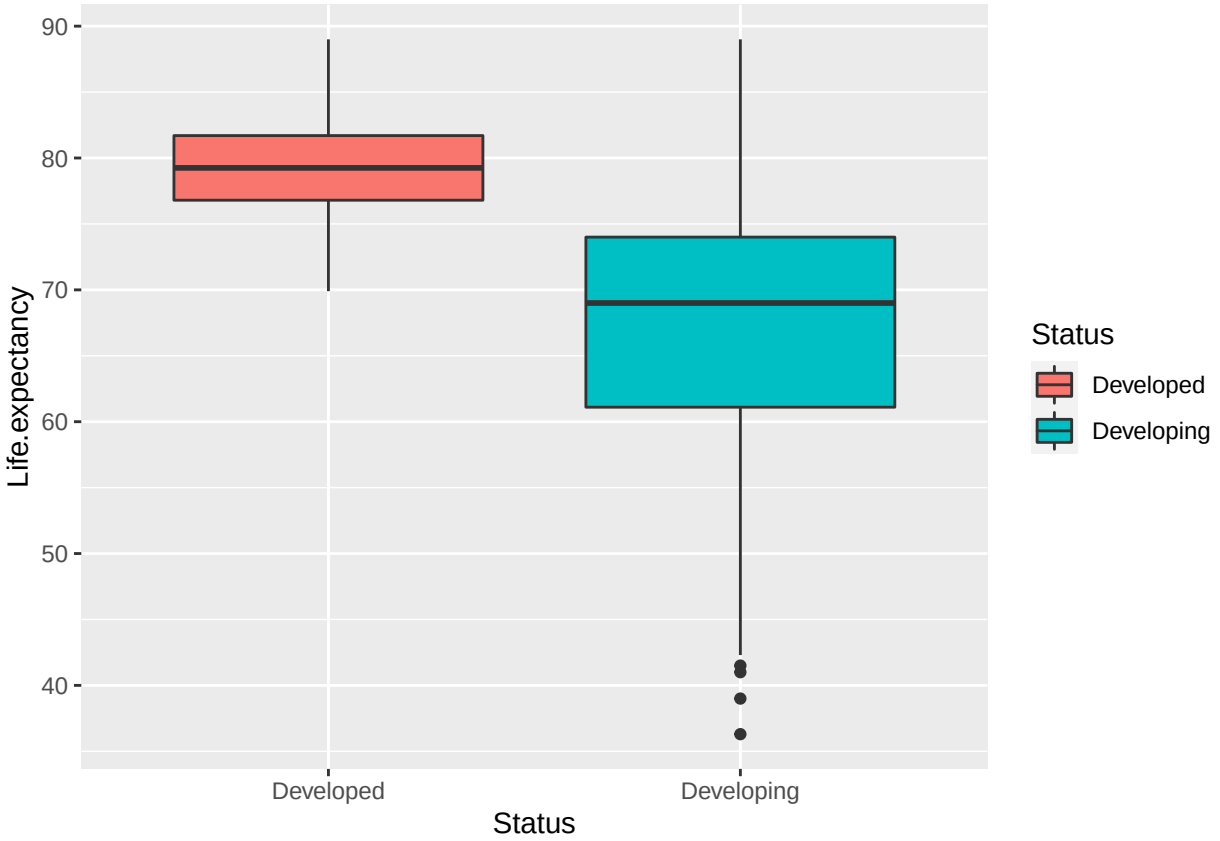



```
years <- seq(from = 2000, to= 2015)

for (anno in years) {
  if (anno == 2000) {
    LE_PerYear <- rel_attr2$Life.expectancy[rel_attr2$Year == anno]
  }
  else {
    LE_PerYear <- cbind(LE_PerYear, rel_attr2$Life.expectancy[rel_attr2$Year == anno])
  }
}

LE_PerYear <- data.frame(LE_PerYear)
colnames(LE_PerYear) <- years

boxplot(LE_PerYear)

grid(nx=10, ny=10)

boxplot(LE_PerYear,
        main = "Boxplot",
        ylab = "Life Expectancy per year",
        xlab = "year",
        col = "cyan",
        border = "black",
        add = T
)
```



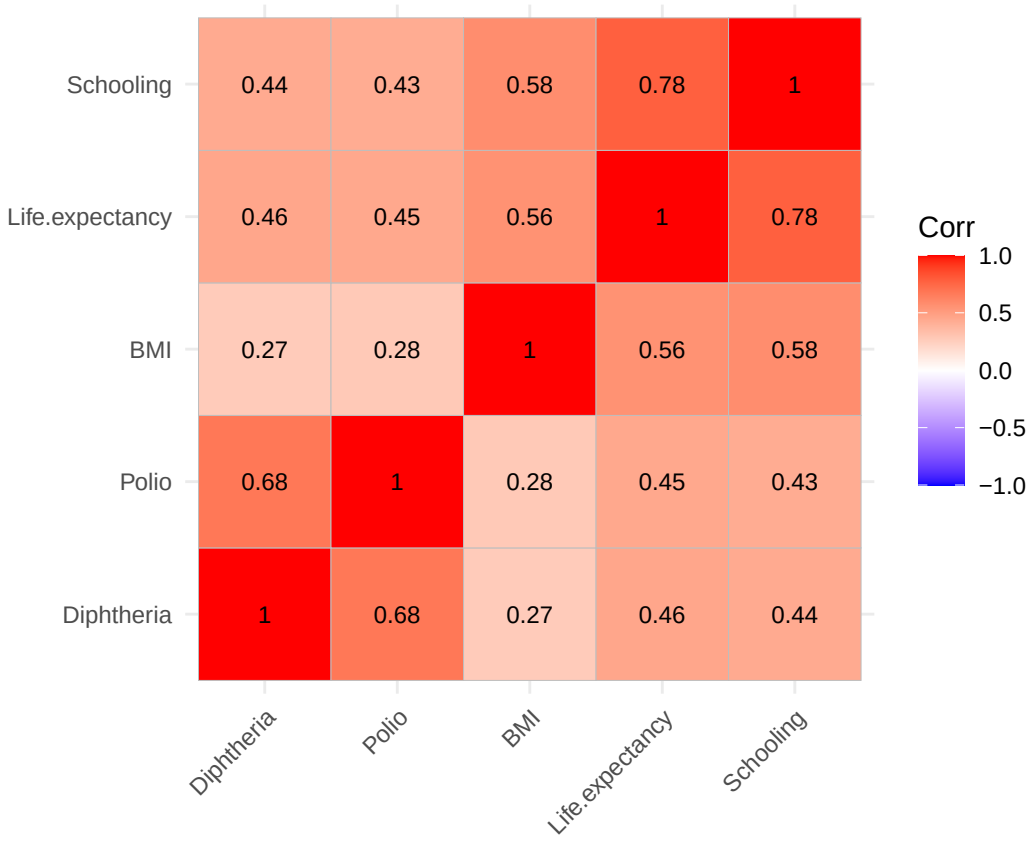
We observe a noticeable growth of the mean worldwide life expectancy through the years (2000-2015). Next we observe the differences between *Developed* and *Developing* countries:

```
ggplot(rel_attr2, aes(x=Status, y=Life.expectancy, fill=Status)) +
geom_boxplot()
```



Finally, let's visualize the correlation matrix of the selected variables after the cleaning of the dataset:

```
matrice_rel <- data.matrix(rel_attr, rownames.force = NA)

cormatRel_no_na <- round(cor(matrice_rel, use = "pairwise.complete.obs"),2)

p.mat <- cor_pmat(matrice_rel)

ggcorrplot(cormatRel_no_na, tl.cex = 9, hc.order = TRUE, p.mat = p.mat, lab = TRUE, lab_size = 3)
```



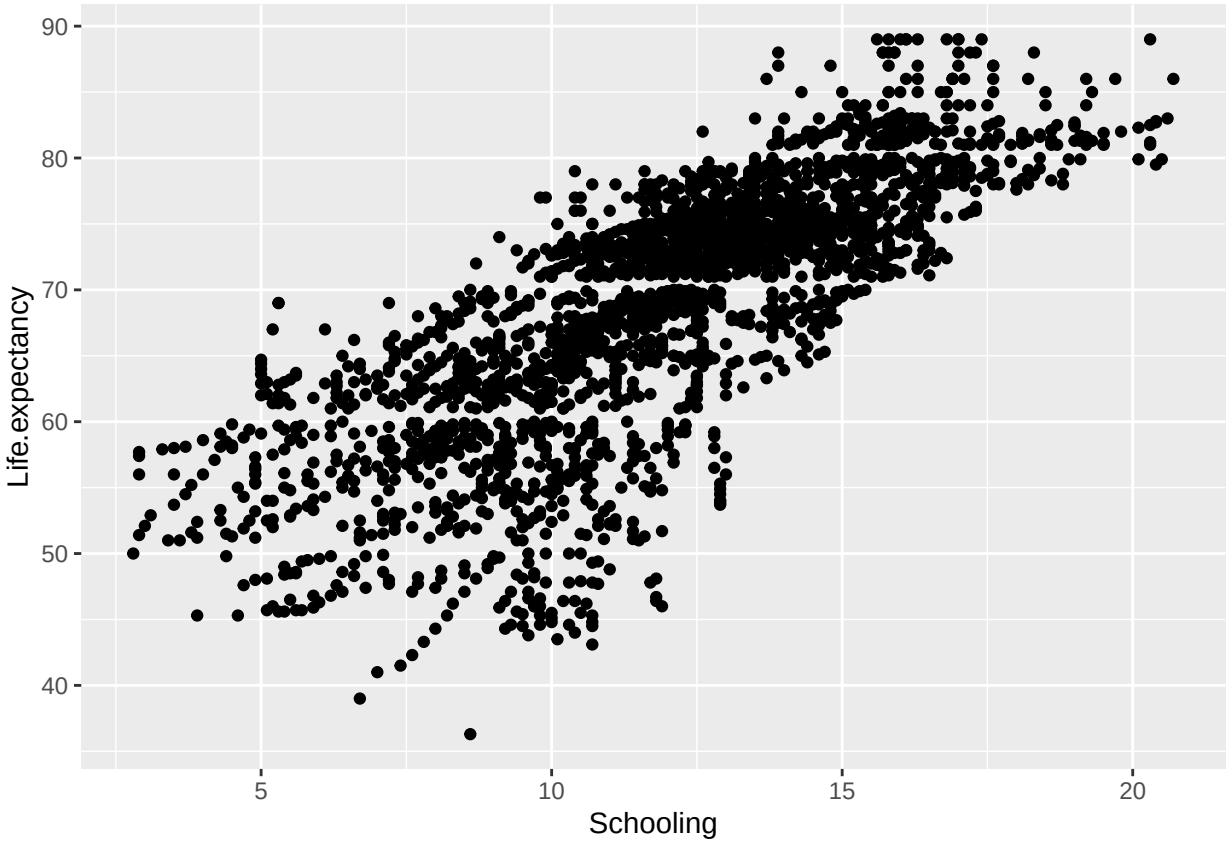
3 Linear Regression

We are now trying to build a regression model which will be capable of predicting the life expectancy of testing samples based on another variable.

As we have previously seen, the variable *Schooling* has the highest correlation with *Life Expectancy*.

Let's visualize the correlation (without NA values):

```
ggplot(data = rel_attr) +
  geom_point(mapping = aes(x = Schooling, y = Life.expectancy))
```




```
cor(rel_attr$Life.expectancy, rel_attr$Schooling, use = "pairwise.complete.obs")
```

```
## [1] 0.7834742
```

As we can see, there's a very noticeable positive linear correlation between the two variables. This brings us to try and build a Linear Regression Model.

Even if we can't verify this assumption with data since the GDP values are badly collected, we can assume that this is a spurious correlation, since *Schooling* is actually highly correlated with *GDP*, which might be a variable that influences the life expectancy more directly.

For the purpose of this study, we will then proceed with a linear regression using the variables:

- y = *Life Expectancy*
- x = *Schooling*

```
instruction.lm <- lm(rel_attr$Life.expectancy ~ rel_attr$Schooling)
summary(instruction.lm)
```

```
##
## Call:
## lm(formula = rel_attr$Life.expectancy ~ rel_attr$Schooling)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

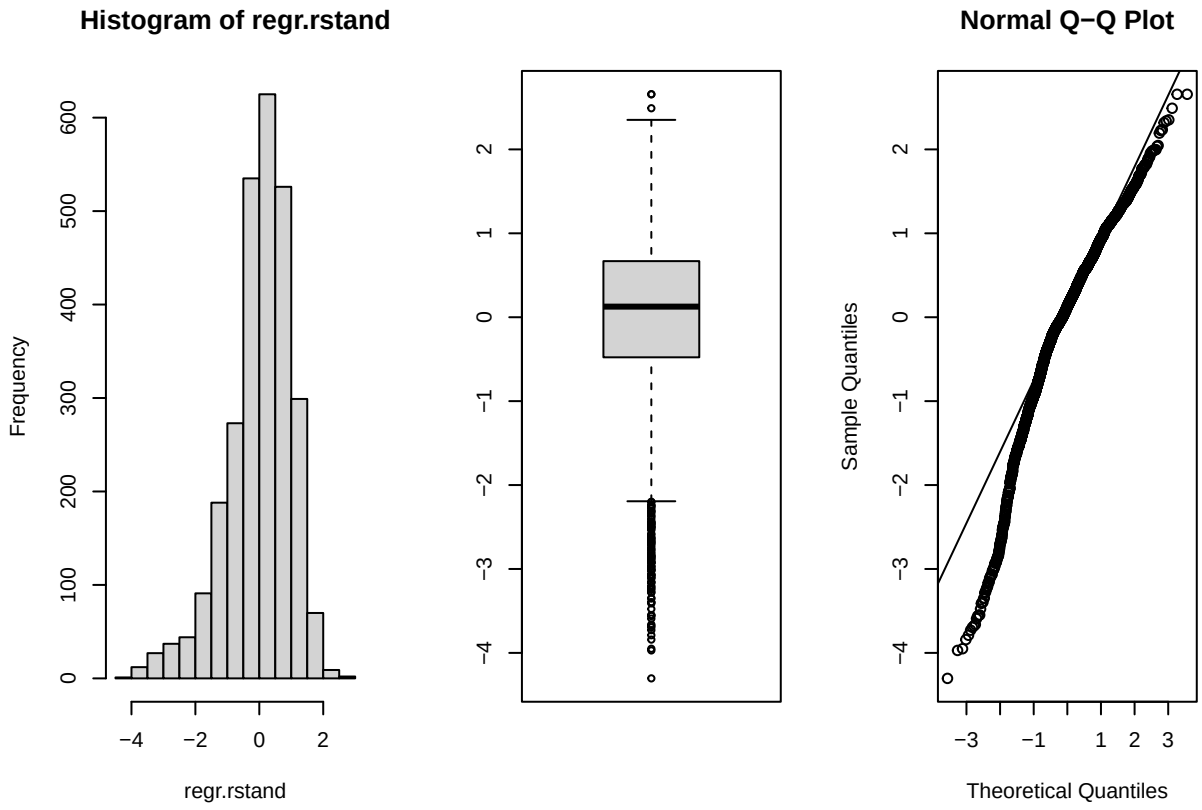
```
## -24.9586  -2.7710   0.7318   3.8762  15.3982
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        41.30452    0.44031   93.81   <2e-16 ***
## rel_attr$Schooling  2.32024    0.03518   65.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.802 on 2737 degrees of freedom
## Multiple R-squared:  0.6138, Adjusted R-squared:  0.6137
## F-statistic:  4351 on 1 and 2737 DF,  p-value: < 2.2e-16
```

As we can see in the summary, the p-values suggest that the variable *Schooling* is statistically significant, also the intercept has around 0 p-values, so it shouldn't be removed.

The R-Squared, which represents a measure of how well a model fits the actual data, has a value around 0.61, which is usually considered to be a moderate correlation.

Let's then analyze the residuals:

```
regr.rstand <- rstandard(instruction.lm)
par(mfrow = c(1, 3))
hist(regr.rstand)
boxplot(regr.rstand)
qqnorm(regr.rstand)
qqline(regr.rstand)
```



The residuals are fairly centered around 0, which could imply the tendency of the residuals to follow a normal distribution. However, as we can see by observing the graphs, there is a negative skewness and a positive kurtosis compared to a normal distribution, and this is why we observe a heavy tail on the left end in the Normal Q-Q Plot, that presents several outliers as showed in the boxplot.

```
matrice_ad <- matrice_rel[, c(1, 2, 3, 4)]

ad.test(matrice_ad[,c(1, 2)])
```

```
##
##  Anderson-Darling normality test
##
## data:  matrice_ad[, c(1, 2)]
## A = 505.91, p-value < 2.2e-16
```

We then conduct the Anderson-Darling normality test, which detects all departures from normality. The result gives a significantly low p-value, which again leads to refuse the normality hypothesis of the residual distribution previously formulated.

This means that the linear model describes the training observations fairly well (R^2 = 61%), but the residuals cannot be identified as normally distributed, implying that they are not a random dataset and that the error is not consistent across all the values of the dependent variable: the model is not fully explaining the behaviour of the training dataset.

4 Multiple Linear Regression

We will now make an attempt to evaluate what variables can be used for a multiple regression, starting from the previous model that we built. We will use a *Forward Selection*: a variable will be chosen if the new regression model has a lower residual sum of squares (RSS) then the previous model. If there are more than one suitable variables, we will choose the one with the lowest RSS.

Let's take the variables *BMI*, *Polio* and *Diphtheria* as possible candidates due to their high correlations with *Life Expectancy*.

We exclude *Income composition of resources* to avoid multicollinearity problems caused by its high correlation with the variable *Schooling*, which has already been used.

We will singularly try to add these terms to our model to see which alternative has lowest RSS.

```
instruction.lm.bmi <- update(instruction.lm, . ~ . + rel_attr$BMI)
instruction.lm.diphtheria <- update(instruction.lm, . ~ . + rel_attr$Diphtheria)
instruction.lm.polio <- update(instruction.lm, . ~ . + rel_attr$Polio)

add1(instruction.lm, . ~ . + rel_attr$BMI + rel_attr$Diphtheria + rel_attr$Polio)
```

We choose the variable *BMI* for our multiple regression, since it has the lowest RSS.

```
summary(instruction.lm.bmi)
```

```
##
## Call:
## lm(formula = rel_attr$Life.expectancy ~ rel_attr$Schooling +
```

```
##     rel_attr$BMI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.2926  -2.7393   0.5778   3.5913  16.6570
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        41.463665   0.440025   94.23   <2e-16 ***
## rel_attr$Schooling  2.058116   0.042848   48.03   <2e-16 ***
## rel_attr$BMI        0.077582   0.006724   11.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.667 on 2715 degrees of freedom
##   (21 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.6316, Adjusted R-squared:  0.6313
## F-statistic:  2328 on 2 and 2715 DF,  p-value: < 2.2e-16
```

Then we repeat the process starting from the new model.

```
add1(instruction.lm.bmi, . ~ . + rel_attr$Diphtheria + rel_attr$Polio)
```

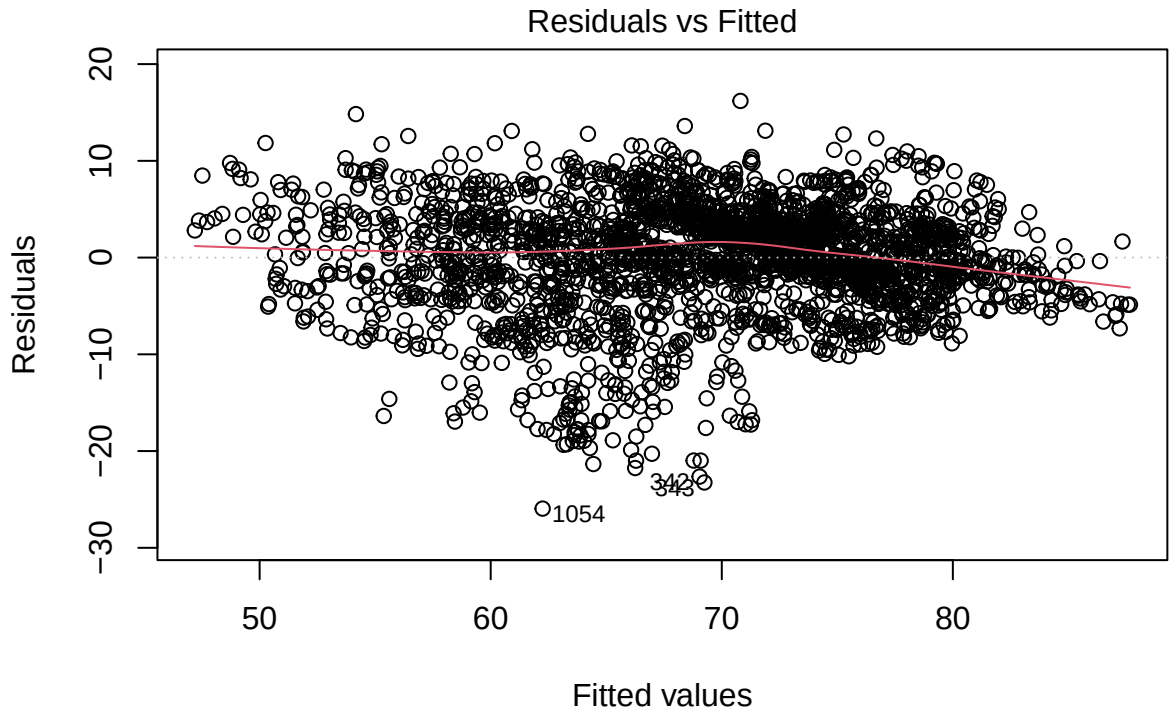
We choose *Diphtheria* as the next variable, due to its lower RSS. To avoid multicollinearity we will avoid using *Polio* for our model since it has an high correlation with the variable *Diphtheria*.

```
instruction.lm.bmi.diph <- update(instruction.lm.bmi, . ~ . + rel_attr$Diphtheria)

summary(instruction.lm.bmi.diph)
```

```
##
## Call:
## lm(formula = rel_attr$Life.expectancy ~ rel_attr$Schooling +
##     rel_attr$BMI + rel_attr$Diphtheria)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9480  -2.6774   0.5779   3.5215  16.1923
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         38.948100   0.484218   80.44   <2e-16 ***
## rel_attr$Schooling   1.878647   0.044787   41.95   <2e-16 ***
## rel_attr$BMI         0.075448   0.006575   11.48   <2e-16 ***
## rel_attr$Diphtheria  0.057708   0.005103   11.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.539 on 2714 degrees of freedom
##   (21 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.6482, Adjusted R-squared:  0.6478
## F-statistic:  1667 on 3 and 2714 DF,  p-value: < 2.2e-16
```

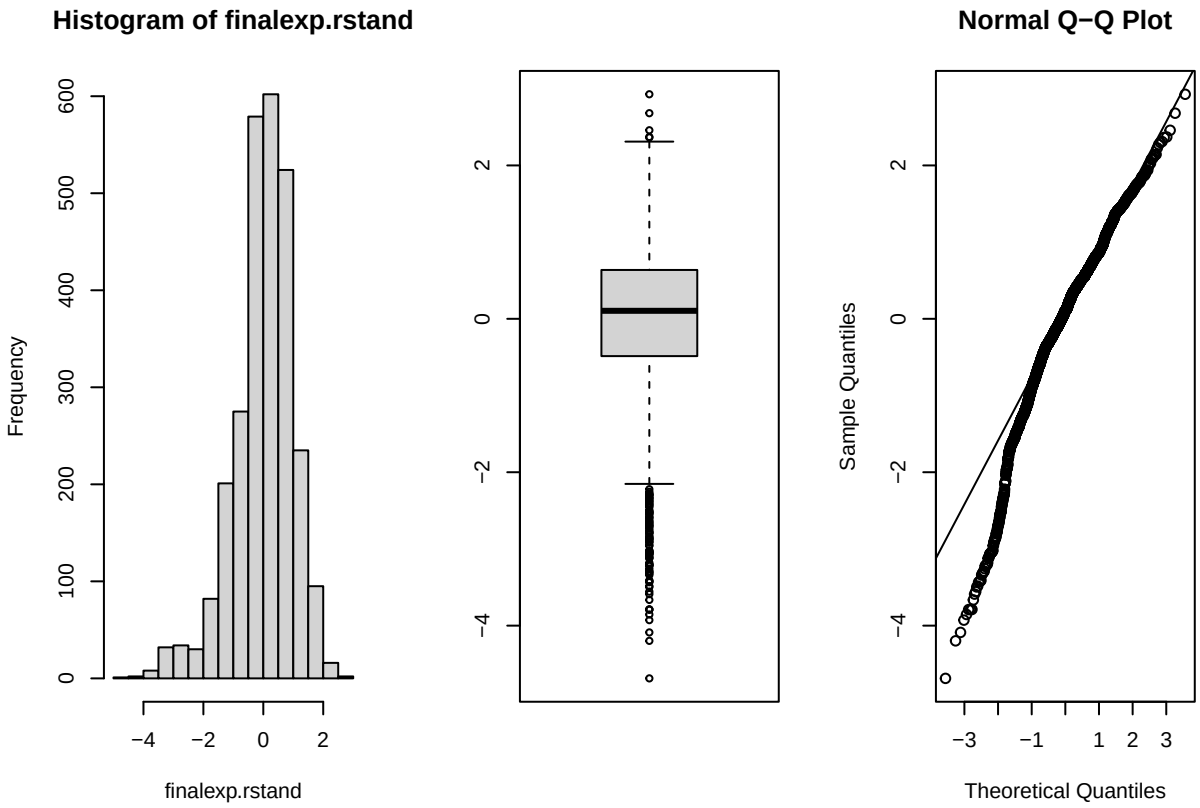
Let's now analyze the residuals of this model:

```
plot(instruction.lm.bmi.diph,
     1,
     sub = "")
```



Analizing the scatter-plot of the residuals, it is possible to observe the absence of a recognizable pattern around 0, and conclude therefore with the verification of the linearity hypothesis of the residuals.

```
finalexp.rstand <- rstandard(instruction.lm.bmi.diph)
par(mfrow = c(1, 3))
hist(finalexp.rstand)
boxplot(finalexp.rstand)
qqnorm(finalexp.rstand)
qqline(finalexp.rstand)
```




```
ad.test(matrice_ad)
```

```
##
##  Anderson-Darling normality test
##
## data:  matrice_ad
## A = 387.86, p-value < 2.2e-16
```

```
dwtest(instruction.lm.bmi.diph)
```

```
##
##  Durbin-Watson test
##
## data:  instruction.lm.bmi.diph
## DW = 0.30326, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

The results show an improvement of the Adjusted R-Squared value using the multiple regression instead of the linear regression.

Despite that, the analysis of the residuals' plots and the AD test don't show a significant improvement in demonstrating the normality of the residuals, since the p-value hasn't increased enough to reach the value 0.05 which is the threshold over which the normality hypothesis is accepted.

We also conducted the Durbin Watson Test which detects the presence of autocorrelation in the residuals. Since this p-value is less than 0.05, we can reject the null hypothesis and conclude that the residuals in

this regression model are autocorrelated. Positive autocorrelation of the residuals implies that they are not independent and identically distributed, and therefore the entire variability of the model was not caught.

Our analysis and construction of the model ends here, but a further improvement would be trying to find other models (using an exponential regression for example, or a mixed model) in order to obtain normally distributed residuals while maintaining a high R^2 value, or to use different variables.