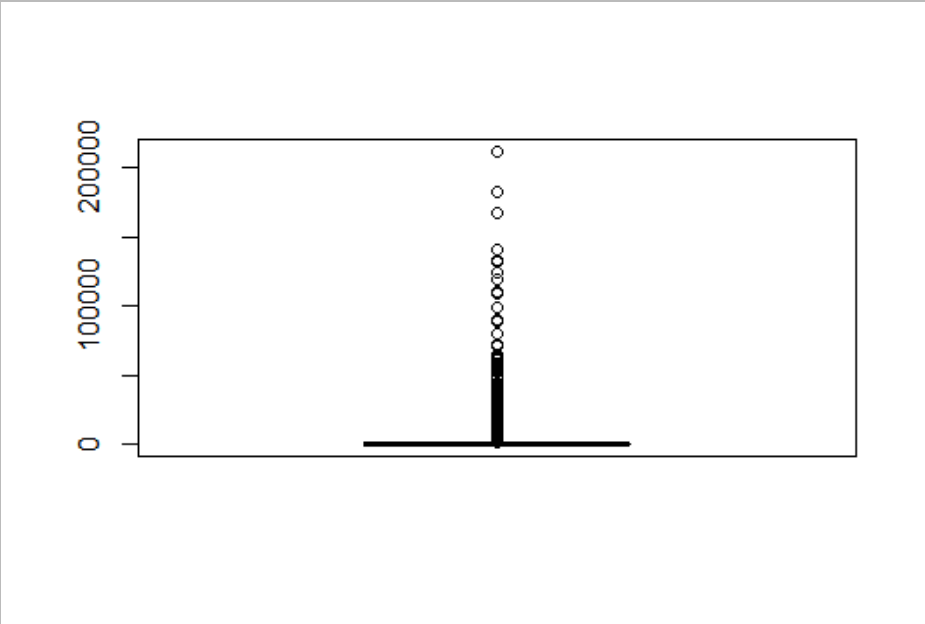
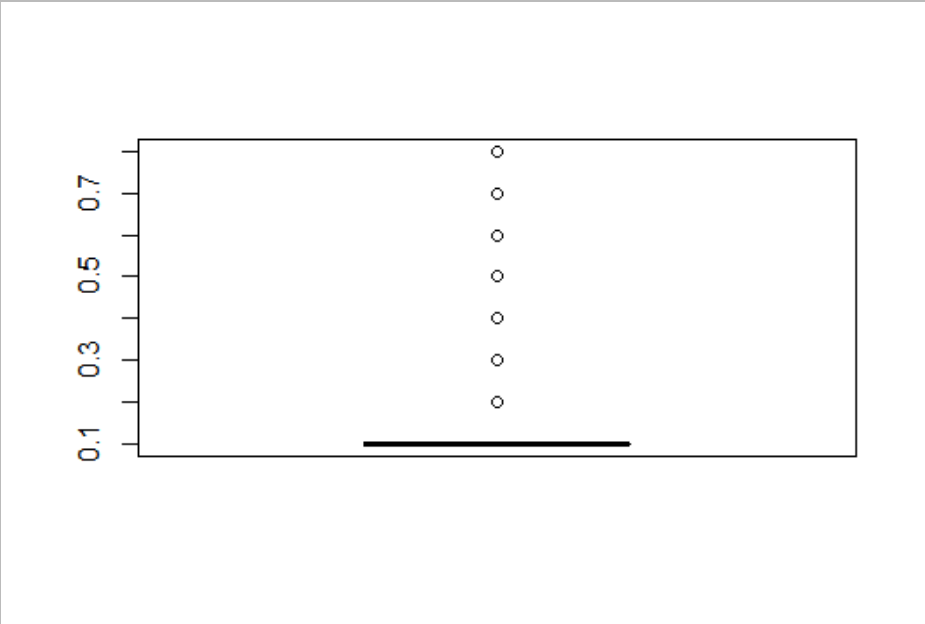
## **Problem description:**

In this case study we’re provided with average life expectancy of people of 193 Countries. We have to predict next year value using linear regression.

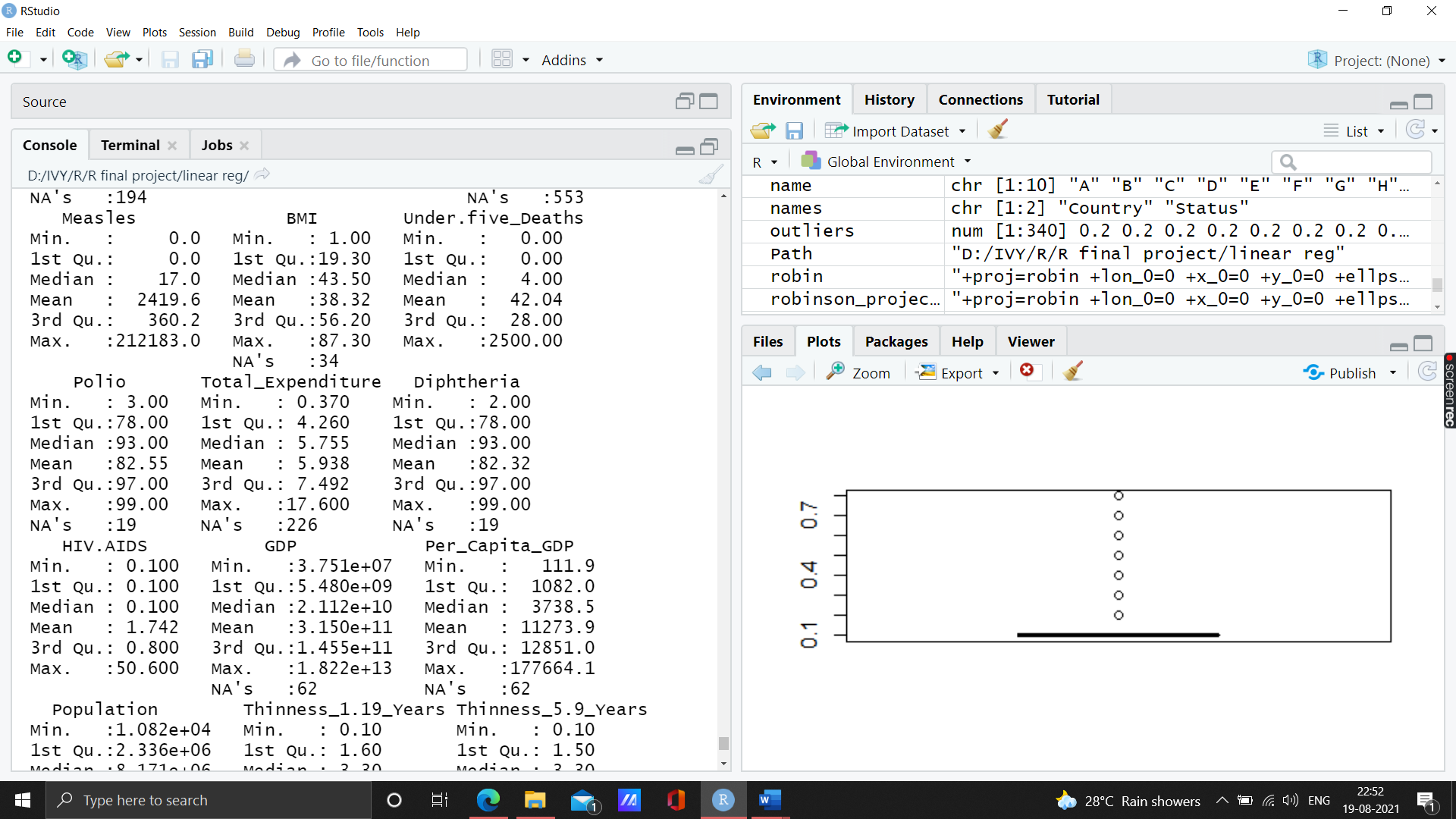
**Data Wrangling**

1. Life Expectancy dataset consist of 2938 rows and 23 columns where Life\_Expectancy is the dependent variable.
2. **Country** and **Status (Developed and Developing)** have been converted into categorical variables.
3. No outlier present in **Life\_Expectancy** but **outliers** are found in **Measles** and **HIV.AIDS** variables as seen in the boxplot.

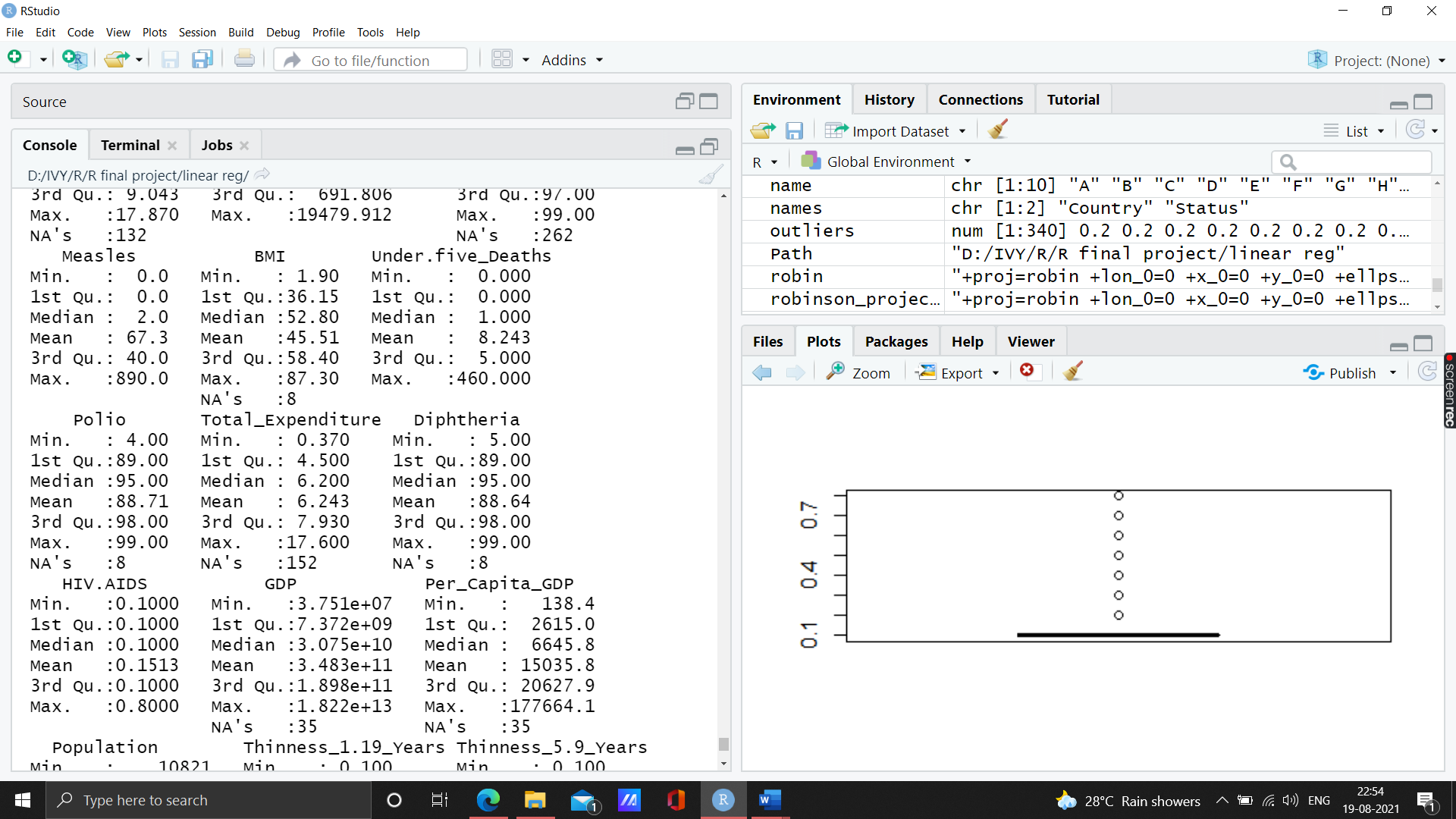


**Measles outliers**

**HIV.AIDS outliers**



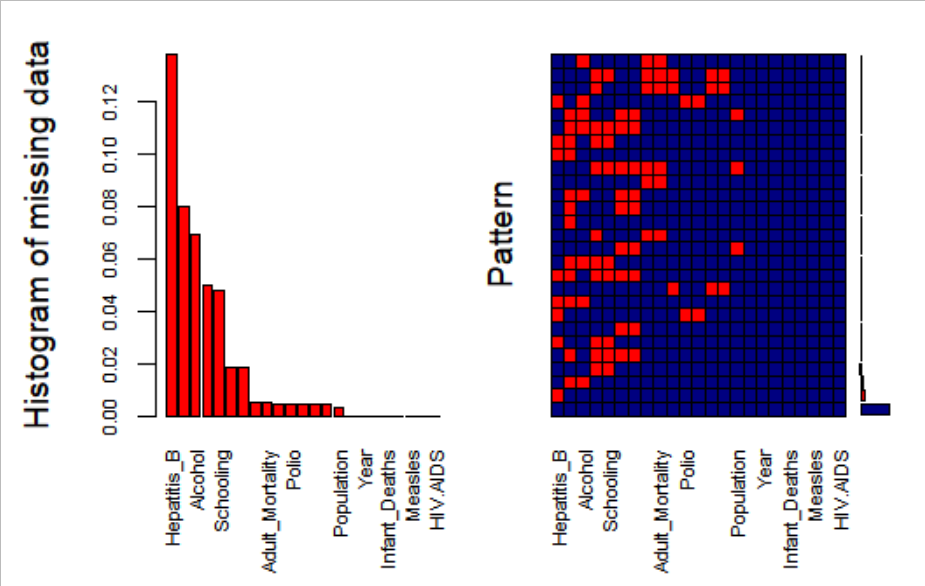
**After outlier treatment, in the summary of the dataset we have 1896 rows and 23 columns:**



1. The dataset consist of **missing values** as follows:

|  |  |
| --- | --- |
| **Column names** | **Missing values** |
| Country | 0 |
| Year | 0 |
| Status | 0 |
| Life\_Expectancy | 10 |
| Adult\_Mortality | 10 |
| Infant\_Deaths | 0 |
| **Alcohol** | **132** |
| Percentage\_Expenditure | 0 |
| **Hepatitis\_B** | **262** |
| Measles | 0 |
| BMI | 8 |
| Under.five\_Deaths | 0 |
| Polio | 8 |
| **Total\_Expenditure** | **152** |
| Diphtheria | 8 |
| HIV.AIDS | 0 |
| GDP | 35 |
| Per\_Capita\_GDP | 35 |
| Population | 6 |
| Thinness\_1.19\_Years | 8 |
| Thinness\_5.9\_Years | 8 |
| Income\_Composition\_of\_Resources | 95 |
| Schooling | 91 |

The missing values has been treated with **Mice** Package:

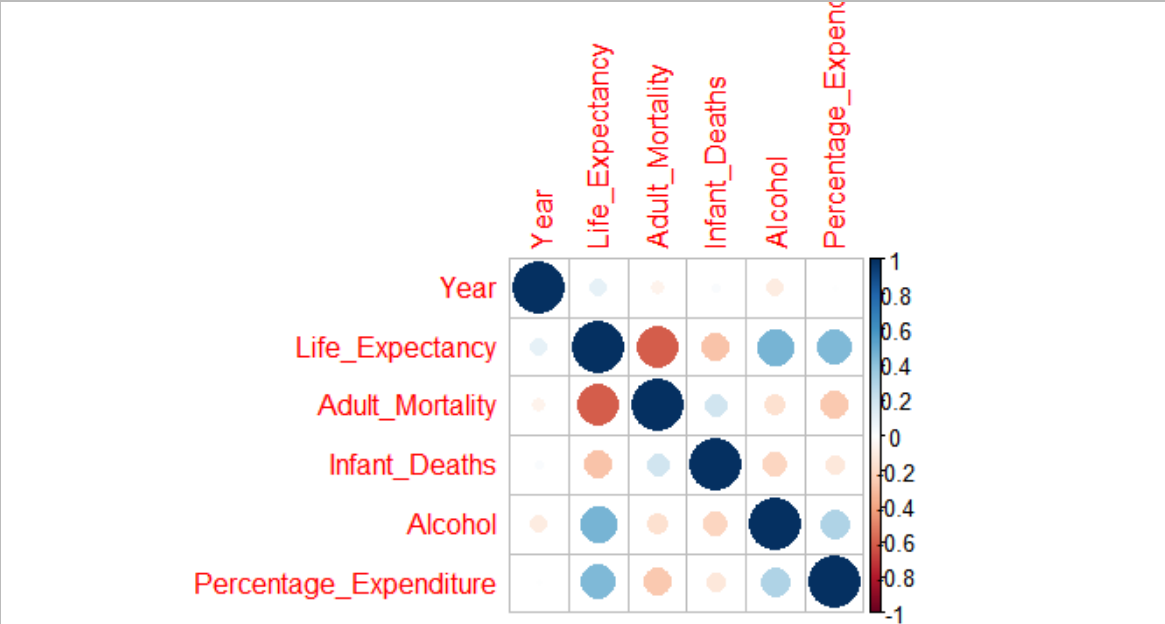


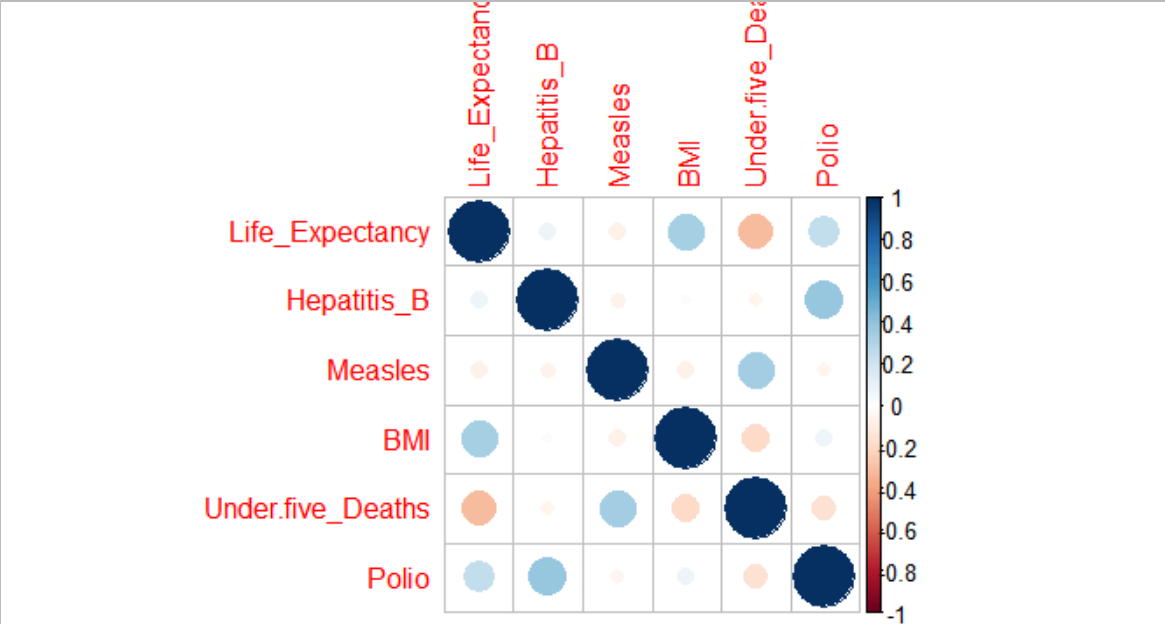
**Hepaptitis\_B** has the highest missing value followed by **Alcohol** and **Schooling**. There has been a maximum of 5 imputations , 20 iterations. Out of which 3rd imputation set has been taken.

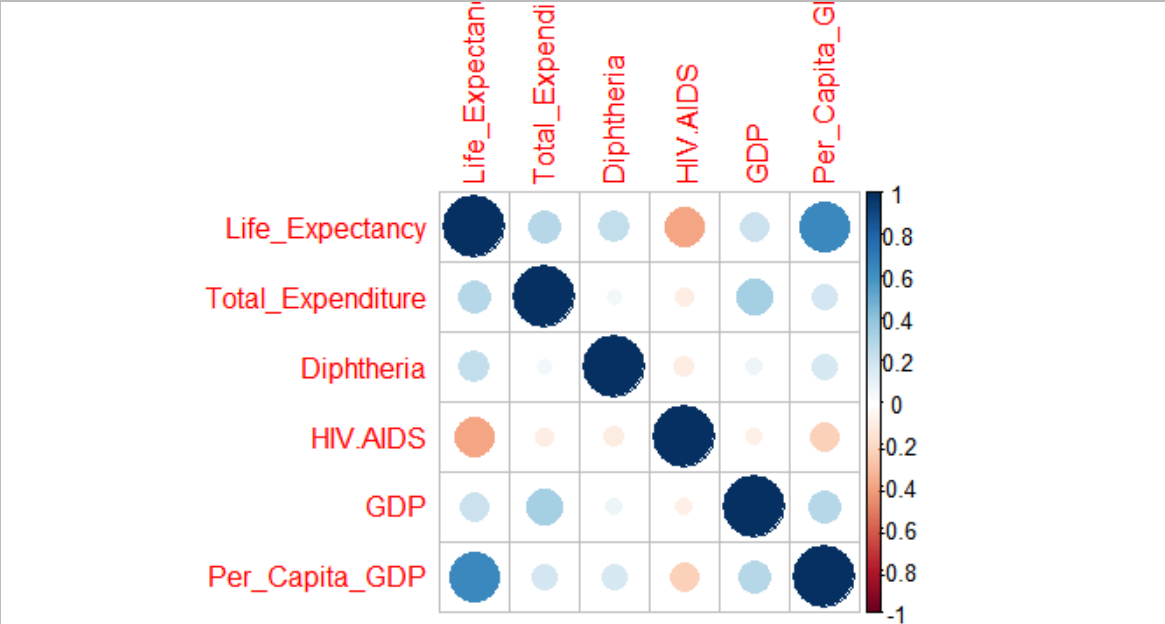
Final\_data1.csv consist of the treated data.

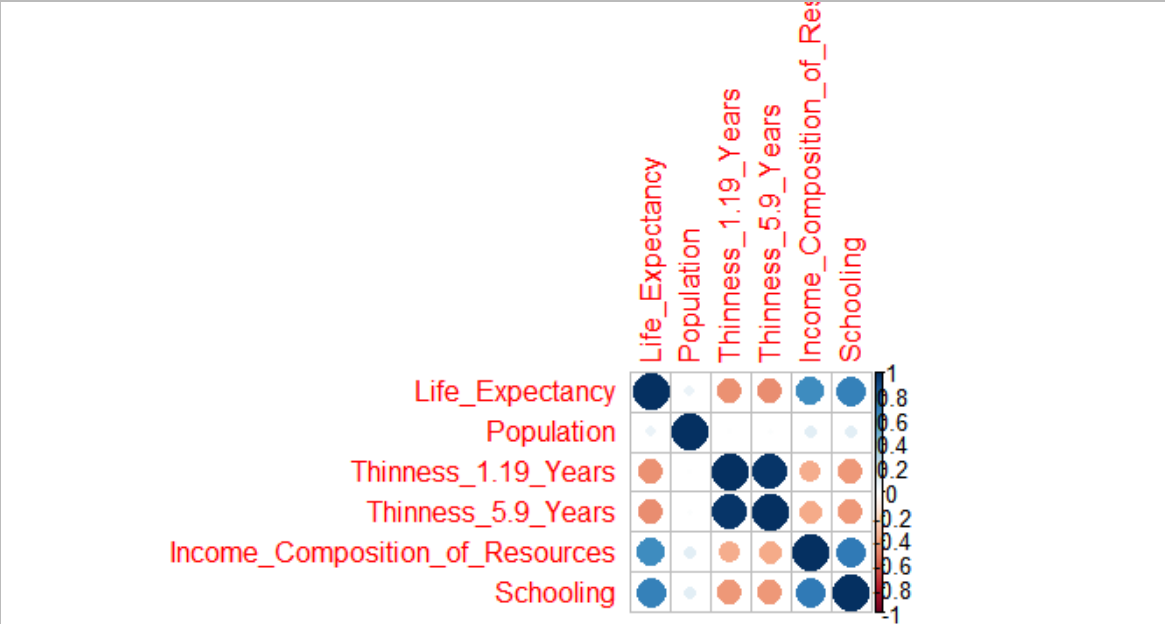
**Basic Exploration of the data**

1. **Correlations** of target variable with other variables.



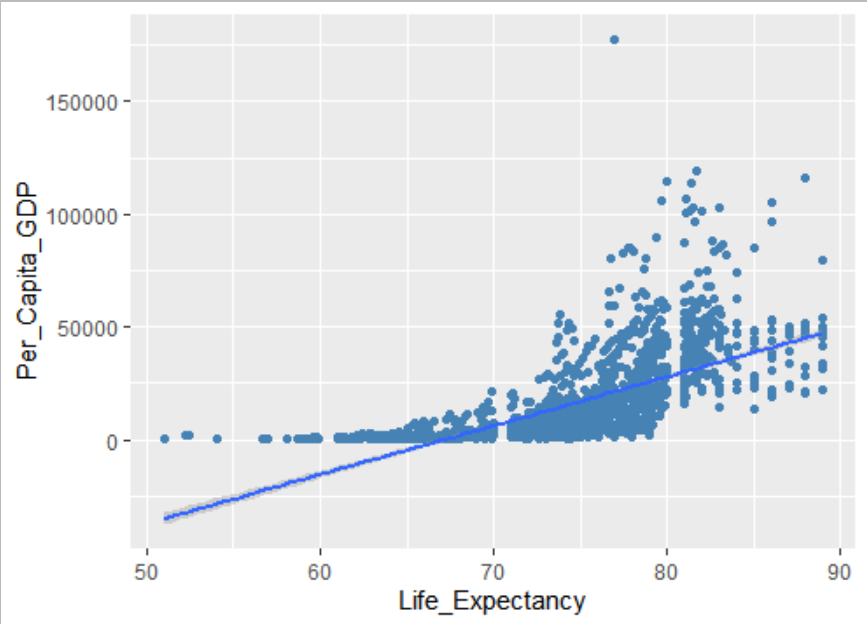


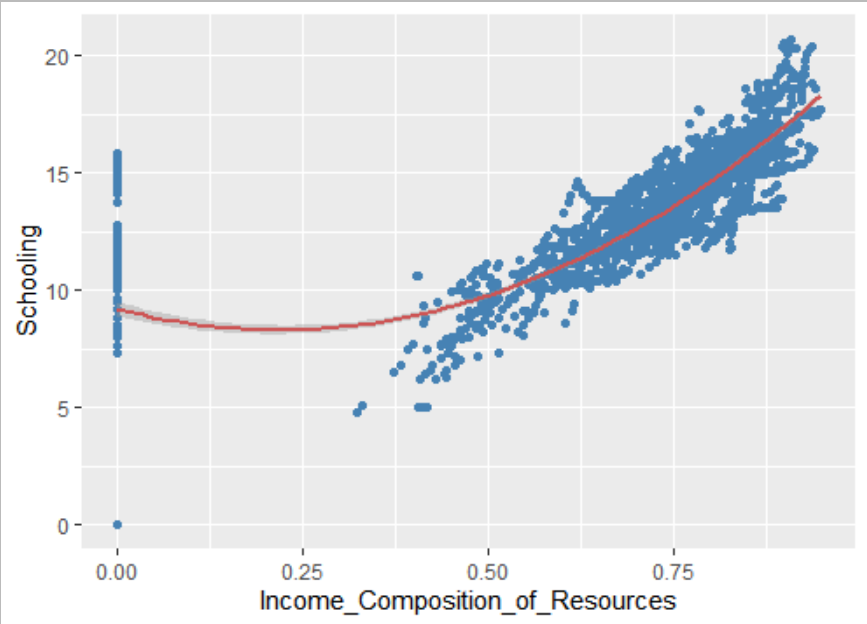




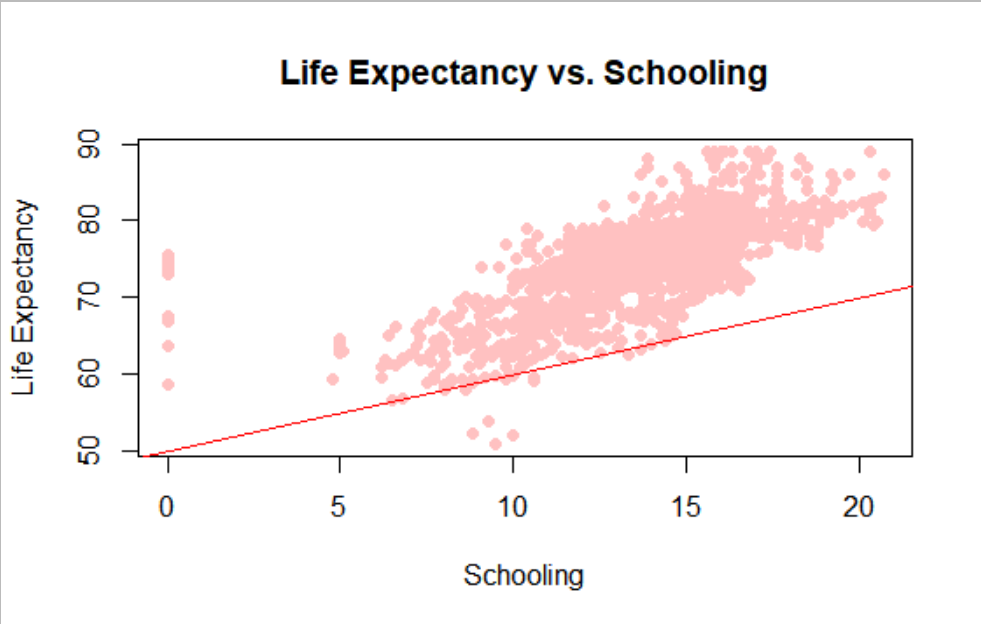
From these correlations plot we can see that Income\_Composition\_of \_Resources and Schooling , Thinness\_1.19\_Years and Thinness\_5.9\_Years, Under.five\_Deaths and Infant\_Deaths are highly correlated.

Per\_Capita\_GDP, Income\_Composition\_of \_Resources and Schooling are positively correlated with Life\_Expectancy while Adult\_Mortality is negatively correlated.





Income\_Composition\_of\_Resources and Schooling are strongly positively correlated

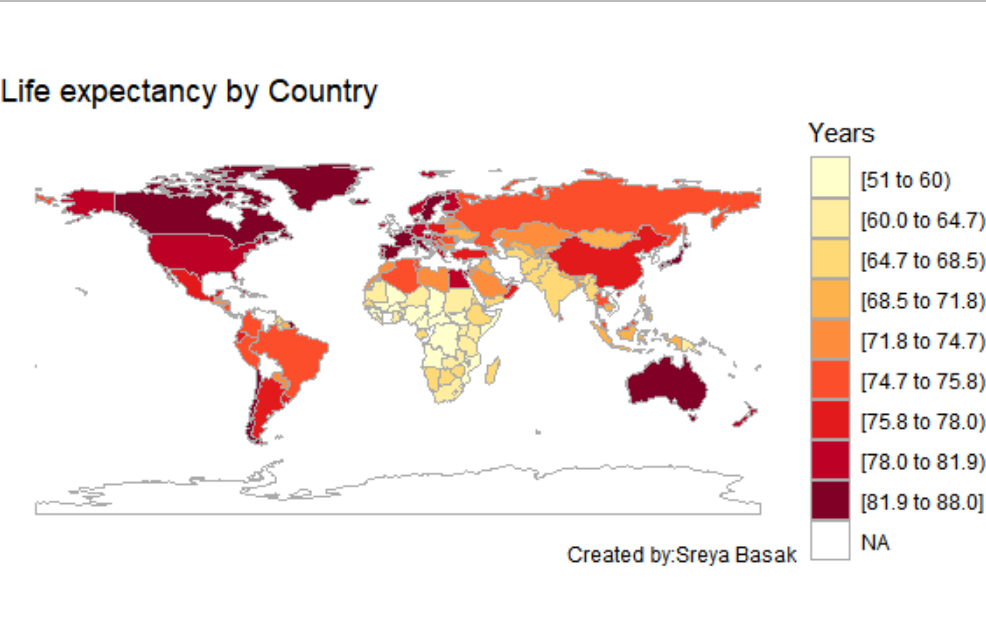


Life Expectancy and Schooling are positively correlated.



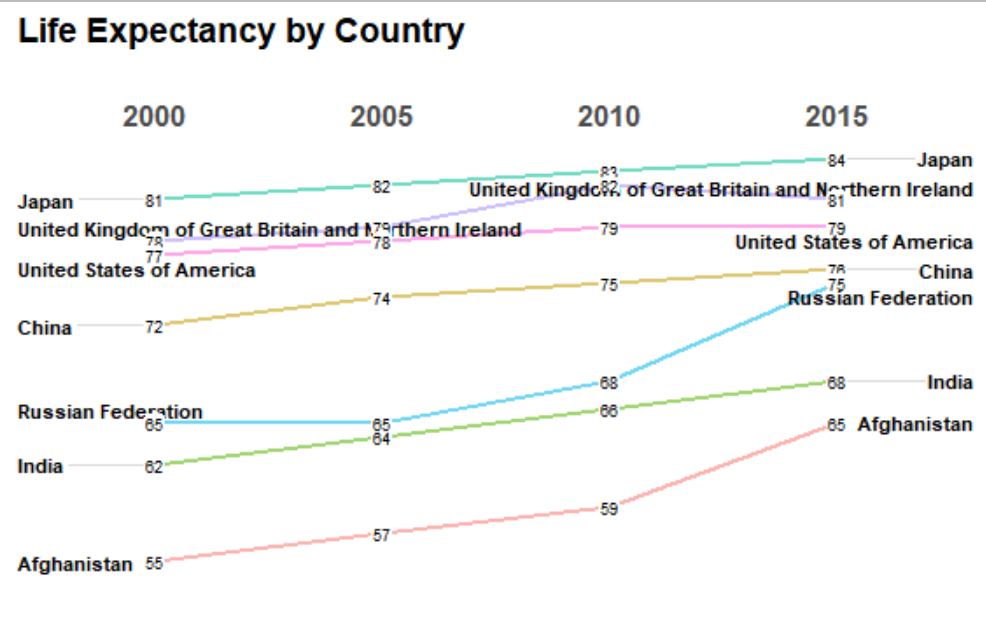
Life\_Expectancy and Adult\_Mortality are negatively correlated.

1. **Life Expectancy all over the world**

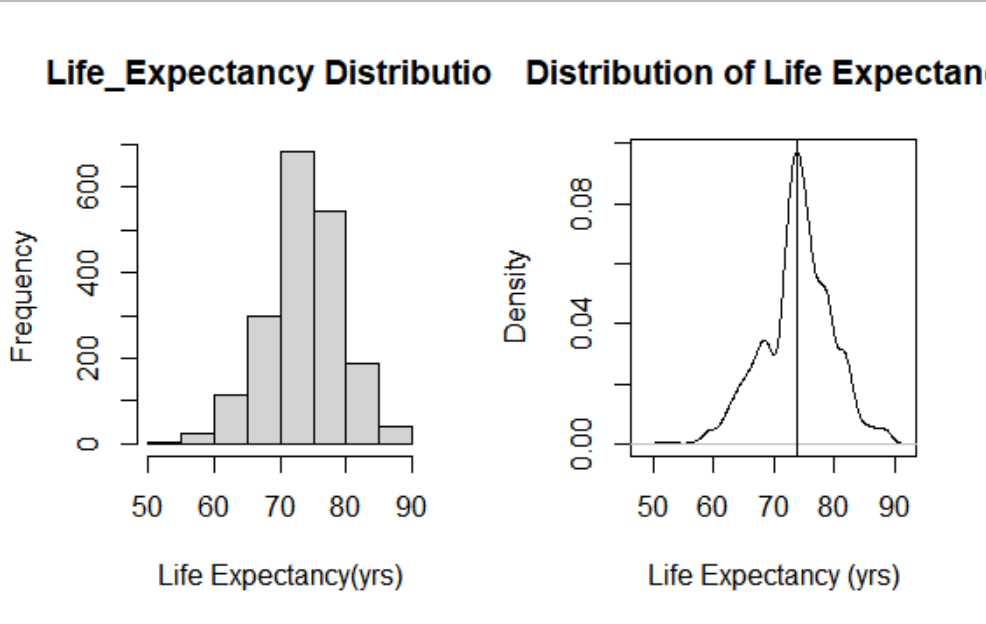


In 2015, Slovenia had the highest life expectancy 88.0 years while Sierra Leone had the lowest life expectancy 51 years.

In 2000, Japan had the highest life expectancy 81.1, Sierra Leone had the lowest life expectancy 39 years.



In India, Life expectancy has increased from 62 in the year 2000 to 68 in the year 2015.



**Observation findings :**

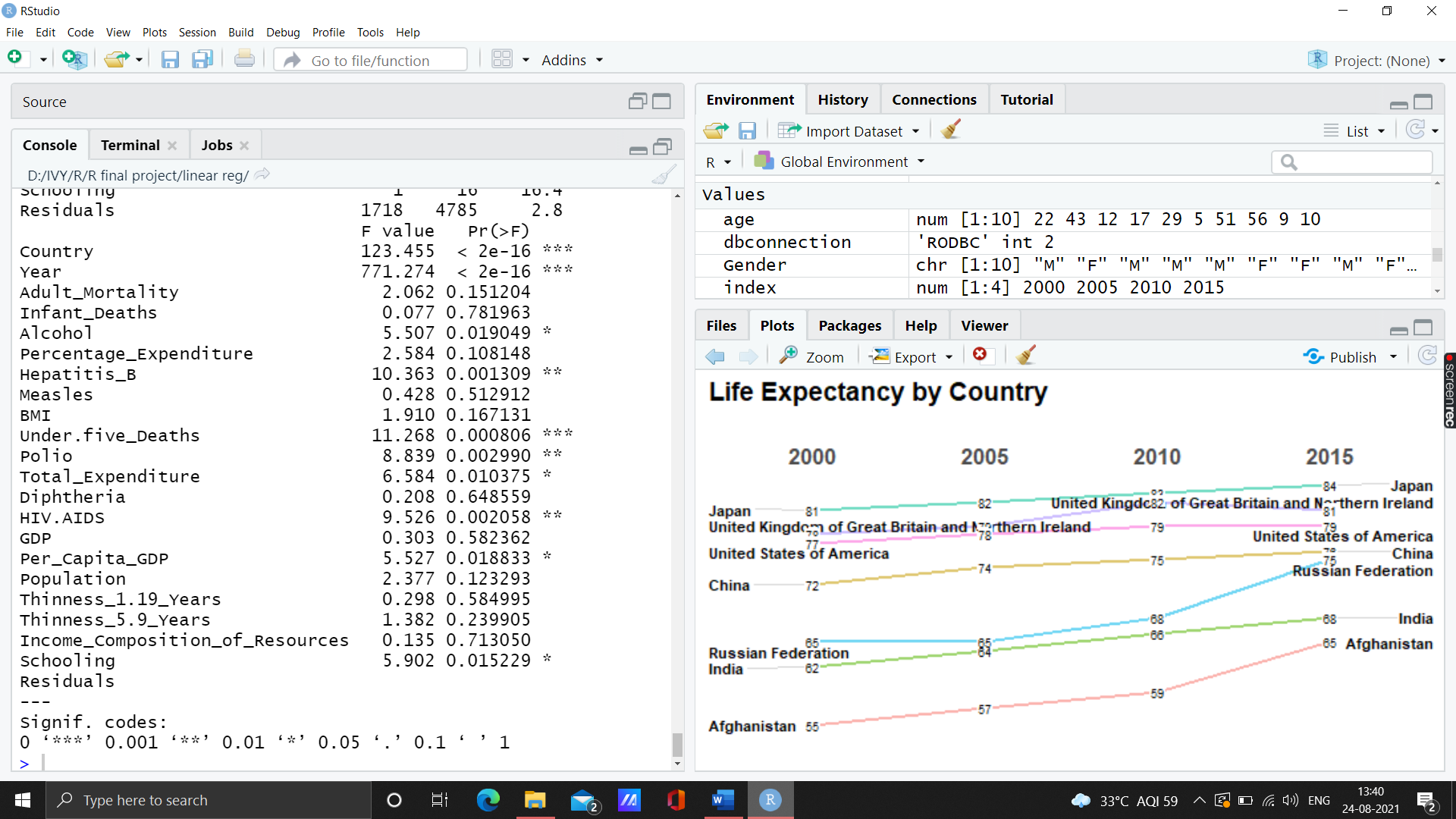
1. The target variable Life\_Expectancy is not distributed perfectly normal, it is a little left-skewed.

2. The unit of Life Expectancy is number of years.

3. Maximum Distribution of Life Expectancy is between 70-80 years.

1. **Anova test**

Variables: Country, Year, Alcohol,Hepaptits\_B,Under.five\_Deaths, Polio, Total\_Expenditure, HIV.AIDS, Per\_Capita\_GDP, Schooling are significant variables with 95% confidence level.



1. First I removed the column Country, as it has too many levels, Infant\_Deaths, Thinness\_5.9\_Years as they are correlated with Under.five\_Deaths and Thinness\_1.19\_Years respectively.

**Modelling and Predictions**

1. **Null or the baseline model**

With only intercept,

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 73.9922 0.1304 567.3 <2e-16 \*\*\*

**Observation findings :**

1. The nullModel does not have any predictor.
2. If there is no predictor, then the nullModel will predict the future value using the mean value of the target variable or the intercept value.
3. **EDA based model – model with predictors that strongly correlated to target variable.** The predictors are **Schooling, Adult.Mortality,** and **Income\_Composition\_of\_Resources.**

t value Pr(>|t|)

(Intercept) 120.81 <2e-16 \*\*\*

Schooling 17.89 <2e-16 \*\*\*

Adult\_Mortality -22.46 <2e-16 \*\*\*

Income\_Composition\_of\_Resources 10.08 <2e-16 \*\*\*

Multiple R squared: 0.6063, Adjusted R-Squared: 0.6056

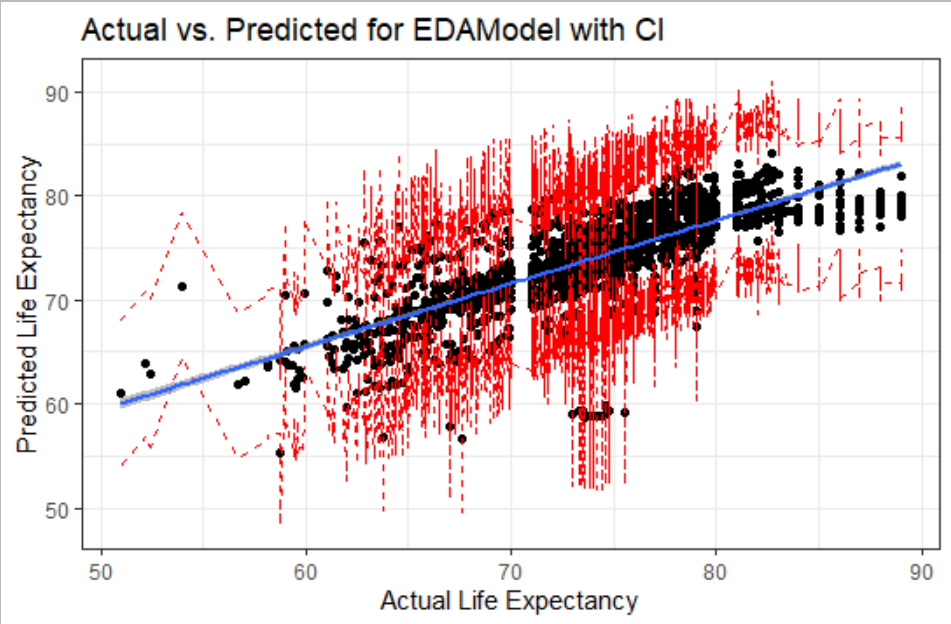
Observation findings :

* 1. The EDAModel has the largest parameter estimate that is Income\_Composition\_of\_Resources which is 5.99.

2.The Income\_Composition\_of\_Resources will affect the Life.expectancy the most in a positive direction.

3.The p-value of all predictors are much lower than 0.05, thus indicating they are very significant predictors for Life.expectancy.

4.The EDAModel has R-squared value 0.6063, which indicates the EDAModel can describe its predictors condition by 60.6%.



Here is the plot of the Actual vs Predicted of EDAModel with Confidence Interval.

The blue line is the regression line, surrounding which in grey shade is the prediction interval. The confidence interval for the prediction is indicated by the dotted red line both above and below the regression line. The plot shows that almost all the data points lie well within the confidence interval of 95%.

1. **Backward Step Model**

**The data set is divided into training and test dataset in the ratio of 7:3.**

**In the original dataset the Backwardstep model is first used.**

The predictors of the BackwardStepModel are Status, Adult.Mortality, Under.five\_Deaths, Per\_Capita\_GDP, Population, Total\_Expenditure, HIV.AIDS, GDP, Thinness\_1.19\_Years, Income\_Composition\_of\_Resources, and Schooling.

Multiple R-squared: 0.7061, Adjusted R-squared: 0.7037

F-statistic: 287.9 on 11 and 1318 DF, p-value: < 2.2e-16

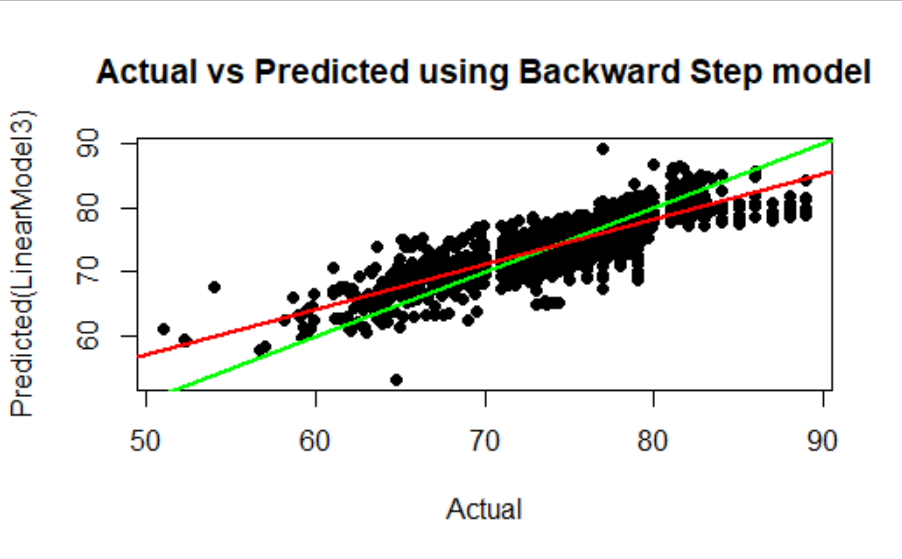
**Observation findings :**

1.The BackwardStepModel has the largest parameter estimate that is Income\_Composition\_of\_Resources which is 3.93, followed by HIV.AIDS which is -3.69.

2.The Income\_Composition\_of\_Resources will affect the Life\_Expectancy the most in a positive direction.

3.On the other hand, the HIV.AIDS will affect the Life\_Expectancy the most in a negative direction.

4.The BackwardStepModel has R-squared value 0.706, which indicates the BackwardStepModel can describe its predictors condition by 70.6%.



Here is the Actual vs Predicted Plot of BackwardStepModel. The green line represents a perfect prediction, while the red line represents the regression line.

1. **Vif reduced Model**

I(Status == "Developing") Adult\_Mortality

**1.647627 1.369585**

Under.five\_Deaths HIV.AIDS

**1.138897 1.154306**

Per\_Capita\_GDP Thinness\_1.19\_Years

**1.747159 1.376806**

Total\_Expenditure Income\_Composition\_of\_Resources

**1.167366 1.585855**

Multiple R-squared: 0.6795, Adjusted R-squared: 0.6776

All the values are less than 2, hence no multicollinearity exists.

# **Final Conclusion**

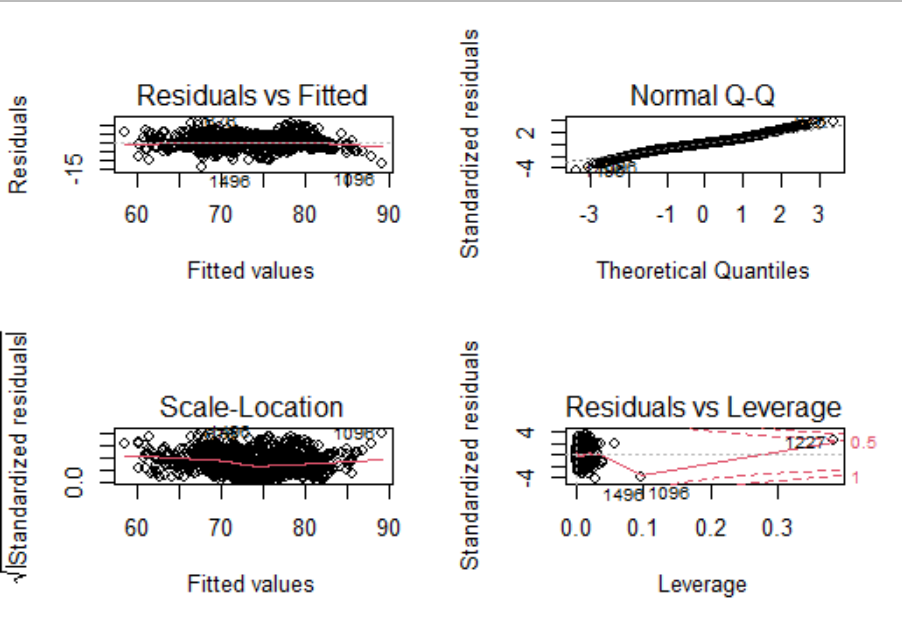
Here are the criteria to find the best model to predict Life Expectancy:

1. The model has the highest value of Adjusted R-Squared
2. The model has the lowest value of RMSE
3. The model has the least predictors.

| **Model** | **Adjusted.R.squared** | **RMSE** | **Number.of.Predictors** |
| --- | --- | --- | --- |
| EDAModel | 0.6063 | 3.562733 | 3 |
| BackwardStepModel | 0.706 | 3.092301 | 11 |
| LinearModel4 | 0.6795 | 3.229481 | 8 |
|  |  |  |  |

So, the best model that fits the criteria goes to

**The LinearModel4—vif reduced model.**



Linear regression makes several assumptions about the data, such as linearity of the data, normality of the residuals (error), homogeneity of residuals variance (homoscedasticity), and independece of residuals error terms (Non-Multicolinearity).

To check the assumptions, we use the diagnostic plots. The diagnostic plots show residuals (error) in four different ways:

**Residuals vs Fitted**. Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, what is good.

**Normal Q-Q**. Used to examine whether the residuals are normally distributed. It’s good if residuals points follow the straight dashed line.

**Scale-Location (or Spread-Location)**. Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity.

**Residuals vs Leverage**. Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis.

Observation findings :

**The Residuals vs Fitted plot** is to check linearity assumption.

There is no pattern in the residual plot. This suggests that we can assume linear relationship between the predictors and the outcome variables.

**The Scale-Location plot** is to check homoscedasticity assumption.

The residuals are spread adequately equal along with the ranges of predictors.

**The Normal Q-Q plot** is to check normality of the residuals (error) assumption.

All the points fall adequately along the reference line, so it’s sufficient to assume that the data has the normality of the residuals.

**The Residuals vs Leverage** is to check linearity assumption.

The plot highlights the top 3 most extreme points (#1496, #1096, and #1227), with a standardized residuals below -4. However, there are outliers that exceed 3 standard deviations.

1. **The mape of the vif reduced model is : 0.03316524**
2. **Autocorrelation test :**

lag Autocorrelation D-W Statistic p-value

1 0.632204 0.7337626 0

Alternative hypothesis: rho != 0

As p-value is <0.05, Autocorrelation exists.

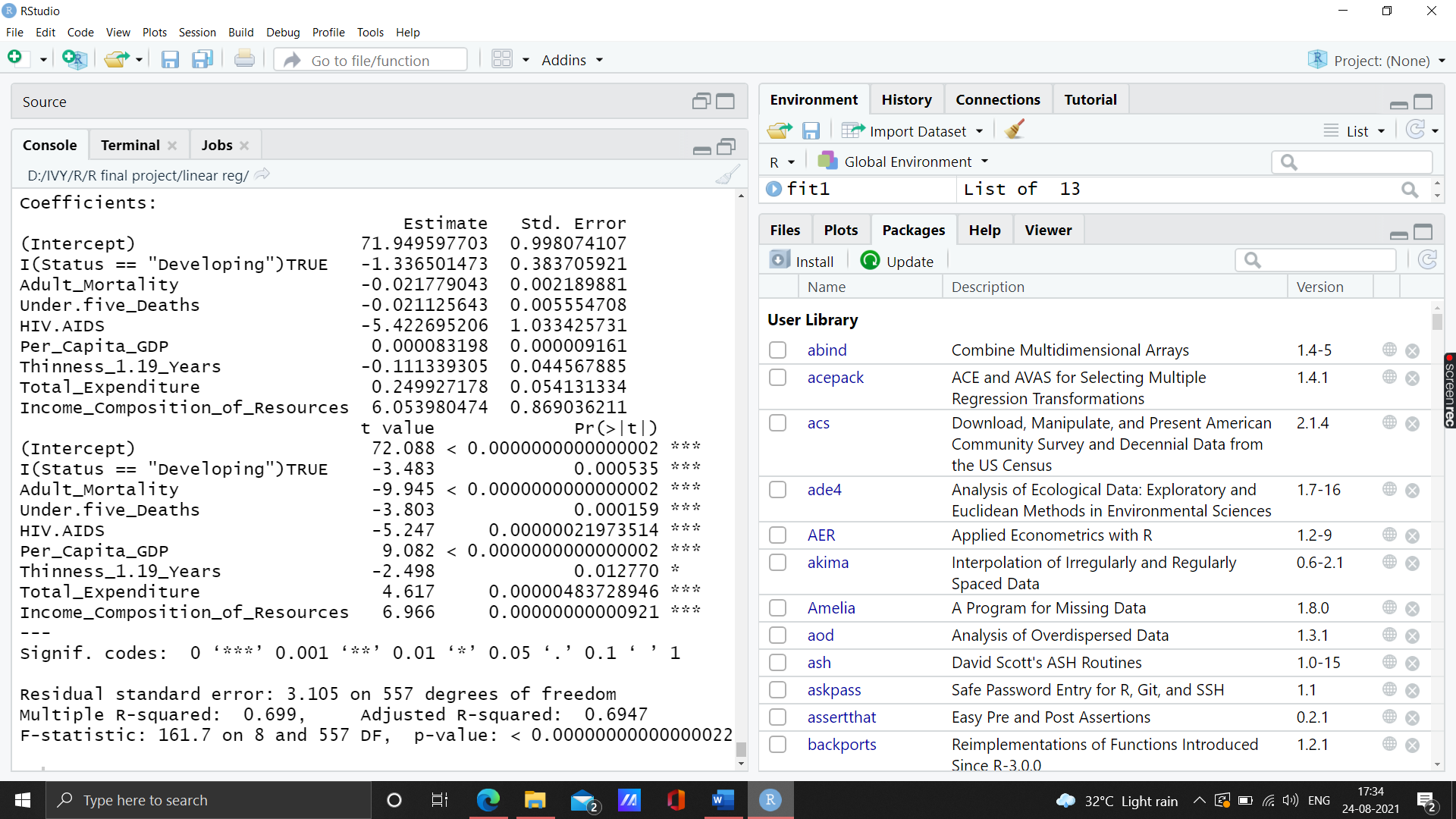
1. In vif test, it is found that all the values are less than , hence no multicollinearity.
2. Breush Pagan test

studentized Breusch-Pagan test

data: FinalModel

BP = 115.69, df = 8, p-value < 0.00000000000000022

1. In the test dataset, we use the vif reduced model and the following is the summary of the model where all the variables are significant at 95% confidence interval.



All the variable have Variance Inflation Factor less than <2 and p-value <0.0001, which shows that there is no significant correlation exists between the variables. Hence, **no multicollinearity exists**.

**I(Status == "Developing") Adult\_Mortality**

1.639379 1.397512

**Under.five\_Deaths HIV.AIDS**

1.187477 1.171004

**Per\_Capita\_GDP Thinness\_1.19\_Years**

1.753472 1.525998

**Total\_Expenditure Income\_Composition\_of\_Resources**

1.226553 1.661689

18. The mape of the model is: **0.03124635**

19. RMSE of the model is: **3.080402**

20. **Durbin Watson statistic in test dataset:**

* The null hypothesis (H0) is that there is no correlation among residuals, i.e., they are independent.
* The alternative hypothesis (Ha) is that residuals are autocorrelated.

lag Autocorrelation D-W Statistic p-value

1 0.3618349 1.269555 0

Alternative hypothesis: rho != 0

**Alternate hypothesis is accepted**

**21.** studentized Breusch-Pagan test

data: fit1

BP = 53.894, df = 8, p-value = 0.000000007239

Alternate hypothesis is accepted

22. The Regression Equation:

*Life\_Expectancy = 71.949597703 - 1.336501473StatusDeveloping*

***(< 0.00000) (0.000535)***

*-0.021779043 Adult Mortality - 0.021125643Under.five\_Deaths*

***(<0.00000000) (0.000159)***

*- 5.422695206HIV.AIDS + 0.000083198Per\_Capita\_GDP*

***(0.00000021973514) (< 0.00000)***

*- 0.111339305Thinness\_1.19\_Years + 0.249927178Total\_Expenditure*

***(0.012770) (0.00000483728946)***

*+ 6.053980474Income\_Composition\_of\_Resources*

***(0.00000000000921)***

**Interpretation of the variables, its statistical significance**:

**StatusDeveloping** -> If Status = “Developing” compared to Status = “Developed”, then **Life\_Expectancy** will decrease by **-1.336501473**

**Adult Mortality** -> If Adult\_Mortality increases by unit then **Life\_Expectancy** will decrease by **-0.021779043**

**Under.five\_Deaths** ->If Under.five\_Deaths per 1000 population increases by unit then **Life\_Expectancy** will decrease by **-0.021125643**

**HIV.AIDS** -> If HIV.AIDS(Deaths per 1 000 live births HIV/AIDS) increases by unit then **Life\_Expectancy** will decrease by **-5.422695206**

**Per\_Capita\_GDP** -> If Per\_Capita\_GDP increases by unit then **Life\_Expectancy** will increase by **0.000083198**

**Thinness\_1.19\_Years** -> If Thinness\_1.19\_Years increases by unit then **Life\_Expectancy** will decrease by -**0.111339305**

**Total\_Expenditure** -> If Total\_Expenditure(General government expenditure on health as a percentage of total government expenditure (%)) increases by unit then **Life\_Expectancy** will increase by **0.249927178**

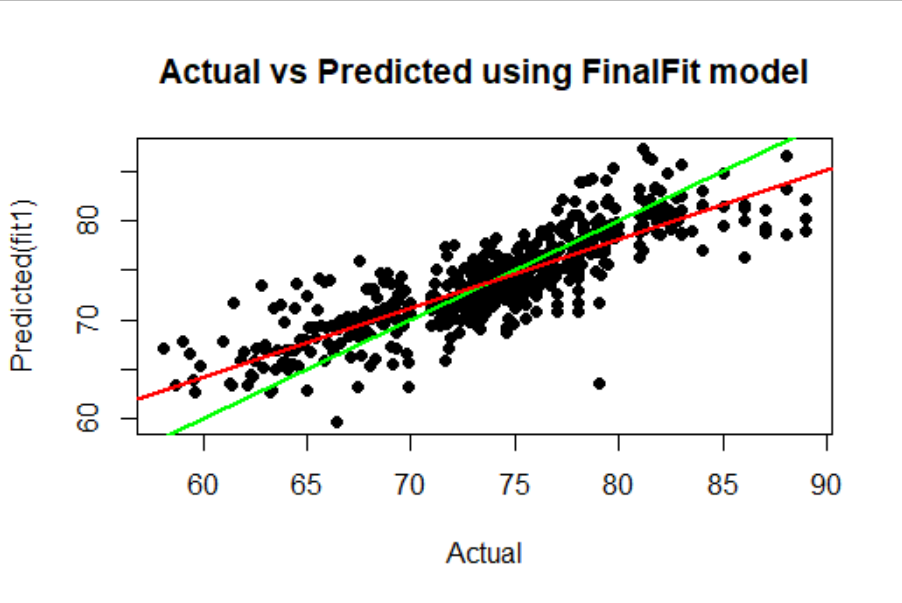
**Income\_Composition\_of\_Resources** -> If Human Development Index in terms of Income\_Composition\_of\_Resources increases by unit then **Life\_Expectancy** will increase by **6.053980474**

**Impact of the explanatory variables on the target variable:**

|  |  |
| --- | --- |
| ***Negative Impact*** | ***Positive Impact*** |
| *StatusDeveloping* | *Per\_Capita\_GDP* |
| *Adult Mortality* | *Total\_Expenditure* |
| *Under.five\_Deaths* | *Income\_Composition\_of\_Resources* |
| *HIV.AIDS* |  |
| *Thinness\_1.19\_Years* |  |

|  |  |
| --- | --- |
| **Multiple R-squared: 0.699** | **Adjusted R-squared: 0.6947** |
| **F-statistic: 161.7 on 8 and 557 DF** | **p-value:**  **< 0.00000000000000022** |
| **MAPE : 0.03124635** | **RMSE: 3.080402** |

**Interpretation:**



The 8 explanatory variables (**StatusDeveloping, Adult\_ Mortality,Under.five\_Deaths, HIV.AIDS, Thinness\_1.19\_Years, Total\_Expenditure, Income\_Composition\_of\_Resources**) are jointly able to explain 69.9 % of the variation in Life\_Expectancy, as represented by the R Square. Also, the small difference between R square and adjusted R-square the represents that we are not overfittingthe model.

MAPE: **0.03124635** which shows the on an average shows the forecast

are off by 0.03%