WHO Life Expectancy

Final Project: EDA & t-test & regression

ALY6010

Prepared by: Heejae Roh
Presented to: Professor Behzad Ahmadi
Date: Dec 16th, 2022

INTRODUCTION

I am human Humans are finite beings. So, I wonder how much longer I will live. Therefore, I choose WHO's life expectancy data. This data was selected before knowing the concept of collinearity. Therefore, as I continued my analysis, I found that the categorical data in the data itself were deeply related to other variables. I will briefly explain my research process and its contents.

ABSTRACTION

Data itself

My target data is life expectancy. Data Contains other Variables which were determined at the boundary of Country. For example, GDP, disease prevention rates, and mortality rates are factors determined by country.

Data Preparation

I changed column names and removed NA data. Also, among numeric data, data that requires coding has been coded.

EDA

Histogram of Life expectancy is like bell shape. I can see the difference of life expectancy between year range & developed status.

t-test

I have 3 questions about life expectancy. Criteria is developed status, years range, BMI. In BMI t-test my assumption which is ‘life expectancy is lower in BMI high country’ is the completely opposite of the result. Therefore, I do additional t-test to confirm that fact.

Correlation test

I have 3 questions about life expectancy. Criteria is schooling years, the number of people who died because of HIV.AIDS per 1,000 in 0-4 age, and GDP. The assumption of Pearson correlation is normally distributed variables, so I conduct Pearson method only in School. For other two variable which are not normally distributed, I conduct spearman method correlation coefficient.

Regression model

I try to find collinearity between variables, so I check all numeric data and delete VIF high values based on explanation of data. After VIF checking, I make a linear regression model with all numeric and categorical variables. The R2 of this regression model is 0.962. However, the Country data’s VIF was extremely high, and I realize that every other variable is determined by Country in this dataset.

I make the regression model without Country whose R2 is 0.765. With normalization of all variables, I could find the highest impact variables easily. Therefore I only left HIV.AIDS, GDP, Schooling years, because I want to make model simpler. The R2 is 0.742 for simplified model.

I make the equation based on simplified model. ‘Life.Exp=48.25– 0.667*HIV.AIDS + 1.8*School +0.00009*GDP’. I put my numbers into this equation, and I realize that my expected life expectancy is 84.38 according to this equation.

CONCLUSION

I realize that checking data before testing and understanding the logic of dataset itself is very important.

PART 1. Data Cleaning

1. Make Colnames Shorter
2. Delete NA values in Data
3. Coding ‘Year’ dataset to categorical dataset as 2000s and 2010s
4. Coding ‘BMI’ dataset to categorical dataset as high (higher than median) and low

PART 2. SUMMARY of EDA

Headtail of Data 1

	Country	Year	Status	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol
1	Afghanistan	2015	Developing	65	263	62	0.01
2	Afghanistan	2014	Developing	59.9	271	64	0.01
3	Afghanistan	2013	Developing	59.9	268	66	0.01
…				…			
2936	Zimbabwe	2002	Developing	44.8	73	25	4.43
2937	Zimbabwe	2001	Developing	45.3	686	25	1.72
2938	Zimbabwe	2000	Developing	46	665	24	1.68

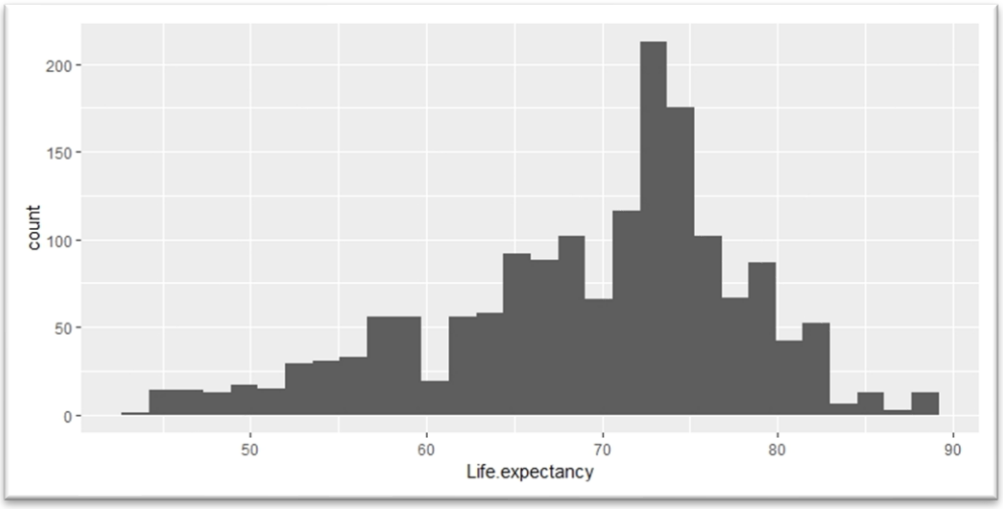
Headtail of Data 2

	%.expenditure	Hepatitis.B	BMI	GDP	Population	thinness.10-19.years	Schooling
1	71.28	65	19.1	584.26	337336494	17.2	10.1
2	73.52	62	18.6	612.7	327582	17.5	10
3	73.22	64	18.1	631.74	31731688	17.7	9.9
…				…			
2936	0	73	26.3	57.35	125525	1.2	10
2937	0	76	25.9	548.59	12366165	1.6	9.8
2938	0	79	25.5	547.36	12222251	11	9.8

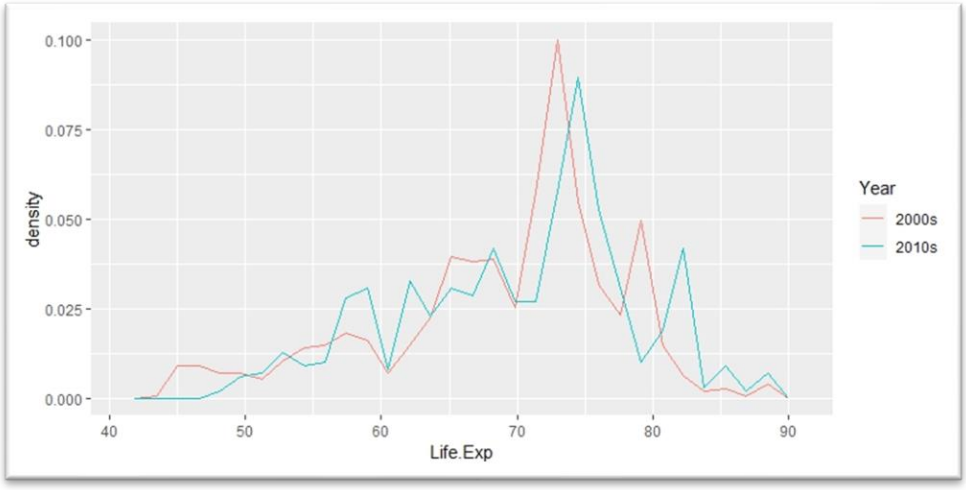
Descriptive Analysis of primary data

	Mean	Sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
Life.Exp	69.30	8.80	71.70	69.91	7.56	44.00	89	45	-0.63	0.03	0.22
Ad.Motal	168.22	125.31	148.00	153.60	109.71	1.00	723	722	1.27	2.38	3.09
BMI	38.13	19.75	43.70	38.89	23.43	2.00	77.1	75.1	-0.23	-1.27	0.49
Polio	83.56	22.45	93.00	88.94	7.41	3.00	99	96	-2.36	5.03	0.55
HIV.AIDS	1.98	6.03	0.10	0.50	0.00	0.10	50.6	50.5	4.97	27.63	0.15
GDP	5566	111476	1593	2779	2078	1.68	119173	119171	4.51	27.89	282.6
Income.Com.R	0.63	0.18	0.67	0.65	0.16	0.00	0.94	0.94	-1.15	2.05	0.00
School	12.12	2.8	12.3	12.18	2.82	4.2	20.7	16.5	-0.13	0.04	0.07

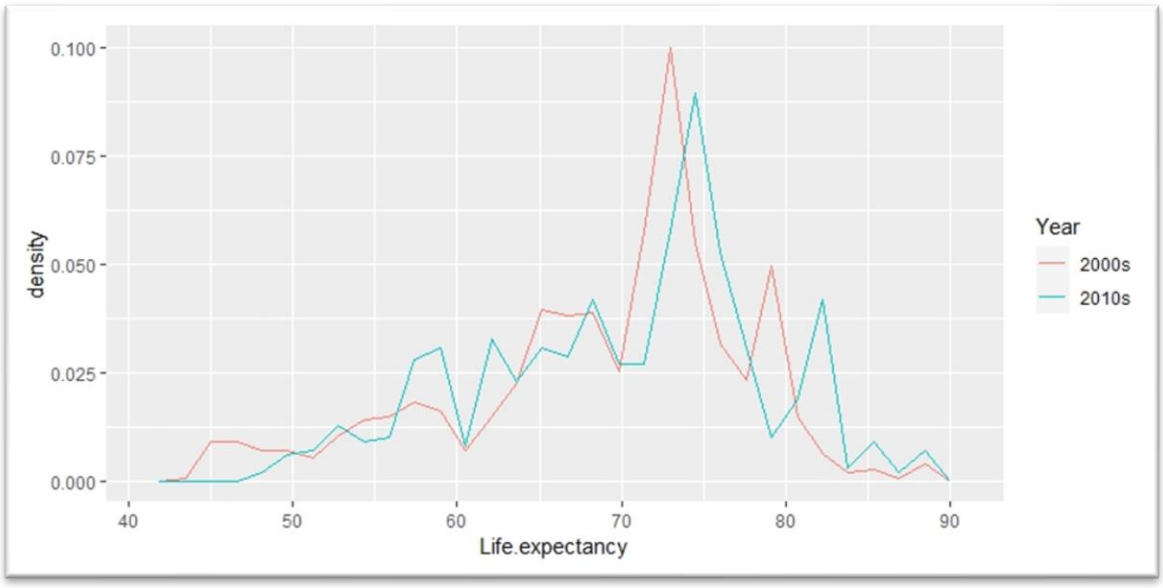
Histogram of Life Expectancy



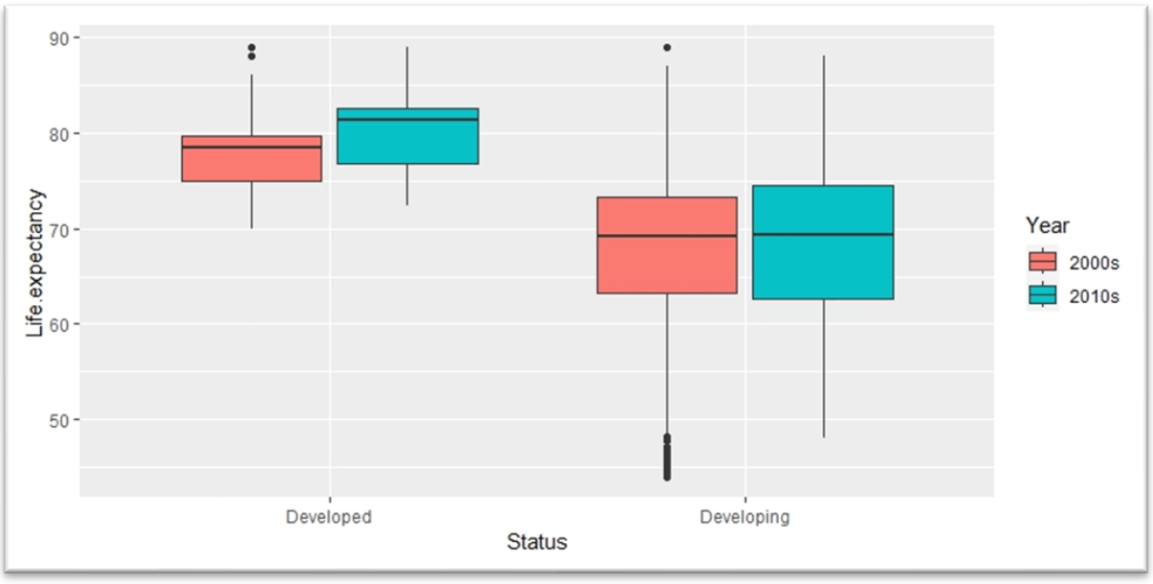
Histogram of Life Expectancy by Status (Developed & Developing)



Histogram of Life Expectancy by Year (2000s & 2010s)



Boxplot of Life Expectancy by Status & Year (2000s & 2010s)



Explanation:

1. 1649 observations, 22 variables after data cleaning
2. Target variable is Life Expectancy. Median of Life Expectancy is 71.7
3. There are two categorical data which is Country & Status
4. Life Expectancy in Deveoped & 2010s is higher than Developing & 2000s, but comparing within in developing the median of life expectancy is almost similar between 2000s and 2010s

PART 3. QUESTIONS

t.test

1. The Life expectancy is higher in Developed comparing with Developing country
2. The Life expectancy is higher in 2010s comparing with 2000s
3. The Life expectancy is lower in BMI high comparing BMI low country.

Correlation

1. There's a correlation between Life expectancy and schooling
2. There's a correlation between Life expectancy and HIV.AIDS
3. There's a correlation between Life expectancy and GDP

PART 4. QUESTIONS to HYPOTHESIS

t.test Null(H0) and alternative(H1) hypothesis/ Claim=(C)

[1] H0: The Life expectancy of Developing and Developed country is equal.
H1: The Life expectancy in Developed country is higher than Developing (C)

[2] H0: The Life expectancy in 2010s and 2000s is equal
H1: The Life expectancy in 2010s is higher than 2000s (C)

[3] H0: The Life expectancy in BMI high and GDP low country is equal
H1: The Life expectancy in BMI high country is higher than BMI low country (C)
BMI high: above BMI median countries, BMI low: below BMI median countries

Correlation Null(H0) and alternative(H1) hypothesis/ Claim=(C)

[1] H0: The correlation coefficient between Life expectancy and schooling is not significantly different from zero.
H1: The correlation coefficient between Life expectancy and schooling is significantly different from zero. (C)

[2] H0: The correlation coefficient between Life expectancy and HIV.AIDS is not significantly different from zero.
H1: The correlation coefficient between Life expectancy and HIV.AIDS is significantly different from zero. (C)

[3] H0: The correlation coefficient between Life expectancy and GDP is not significantly different from zero.
H1: The correlation coefficient between Life expectancy and GDP is significantly different from zero. (C)

PART 5. T-TEST ANALYSIS

T-TEST FOR HYPOTHESIS 1

Step 0 Checking data is normally distributed(APPENDIX).

Step 1 Hypothesis.

The Life expectancy is higher in Developed comparing with Developing country

- H0: μ1 = μ2 and H1: μ1 > μ2 (claim)

Two Sample t-test

data: Developed_Life.Exp		Developing_Life.Exp
t = 20.042	df = 1647	p-value < 2.2e-16
alternative hypothesis:	true difference in means is greater than 0	
95 percent confidence interval:	10.10075	Inf
sample estimates: mean of x		mean of y
78.69174		67.68735

Step 2 Find the critical value.

- α = 0.05, one-tailed CI is 10.10075 Inf

Step 3 Compute the test value.

- t-value is 20.042 with t-test (alternative="greater").

Step 4 Make the decision.

- There is enough evidence to reject H0.

Step 5 Summarize the results.

- The Life expectancy is higher in Developed comparing with Developing country

Interpretation:

1. In this case, at the first t-test, I reject null hypothesis because t-value is in CI
2. The mean of x is 78.69 and y is 67.68. and the p-value is 2.2e-16 which is smaller than α = 0.05
3. The Life expectancy is higher in Developed country, I think, the reason is medical system and food, clothing and shelter.

T-TEST FOR HYPOTHESIS 2

Step 0 Checking data is normally distributed(APPENDIX).

Step 1 Hypothesis.

The Life expectancy is higher in 2010s comparing with 2000s

- H0: μ1 = μ2 and H1: μ1 > μ2 (claim)

Two Sample t-test

data: 2010s_Life.Exp		2000s_Life.Exp
t = 2.7373	df = 1647	p-value = 0.00313
alternative hypothesis:	true difference in means is greater than 0	
95 percent confidence interval:	0.4829348	Inf
sample estimates: mean of x		mean of y
70.03600		68.82492

Step 2 Find the critical value.

- α = 0.05

Step 3 Compute the test value.

- p-value is 0.0031 < 0.05 with t-test (alternative="greater").

Step 4 Make the decision.

- There is enough evidence to reject H0.

Step 5 Summarize the results.

- The Life expectancy is higher in 2010s comparing with 2000s

Interpretation:

1. The p-value is 0.0031 < 0.05, so there is enough evidence to reject H0
2. The Life expectancy is higher in 2010s comparing with 2000s, I think, the technological advancement make life expectancy higher

T-TEST FOR HYPOTHESIS 3 less

Step 0 Checking data is normally distributed(APPENDIX).

Step 1 Hypothesis.

The Life expectancy is lower in BMI high comparing BMI low country

- H0: μ1 = μ2 and H1: μ1 > μ2 (claim)

Two Sample t-test

data: BMI.High_Life.Exp		BMI.Low_Life.Exp
t = 29.628	df = 1647	p-value = 1
alternative hypothesis:	true difference in means is less than 0	
95 percent confidence interval:	-Inf	10.94697
sample estimates: mean of x		mean of y
74.4909		64.1200

Step 2 Find the critical value.

- α = 0.05

Step 3 Compute the test value.

- p-value is 1 > 0.05 with t-test (alternative="less").

Step 4 Make the decision.

- There is not enough evidence to reject H0.

Step 5 Summarize the results.

- The Life expectancy is not lower in BMI high comparing with BMI low

Interpretation:

1. The p-value is 1 > 0.05, so there is not enough evidence to reject H0
2. The mean of x is 74.49 and y is 64.12, so I can think about the life expectancy in BMI higher country is higher than BMI low
3. I will conduct additional t-test about this

T-TEST FOR HYPOTHESIS 3-1 greater

Step 0 Checking data is normally distributed(APPENDIX).

Step 1 Hypothesis.

The Life expectancy is higher in BMI high comparing BMI low country

- H0: μ1 = μ2 and H1: μ1 > μ2 (claim)

Two Sample t-test

data: BMI.High_Life.Exp		BMI.Low_Life.Exp
t = 29.628	df = 1647	p-value < 2.2e-16
alternative hypothesis:	true difference in means is greater than 0	
95 percent confidence interval:	9.794823	Inf
sample estimates: mean of x		mean of y
74.4909		64.1200

Step 2 Find the critical value.

- α = 0.05

Step 3 Compute the test value.

- p-value is 2.2e-16 < 0.05 with t-test (alternative="greater").

Step 4 Make the decision.

- There is enough evidence to reject H0.

Step 5 Summarize the results.

- The Life expectancy is higher in BMI high comparing with BMI low

Interpretation:

1. The p-value is 2.2e-16 < 0.05, so there is enough evidence to reject H0
2. The mean of x is 74.49 and y is 64.12, The Life expectancy is higher in BMI high comparing BMI low country
3. I think BMI is higher in developed countries, so the life expectancy is higher in BMI high countries. I thought that high BMI is not good factor for life expectancy, but extremely low BMI is worse factor for life expectancy.
4. I can conduct more analysis within developing and developed countries

PART 6-1. CORRELATION TEST ANALYSIS

T-TEST FOR HYPOTHESIS 1

Step 0 Test Assumption.

Histogram of Life expectancy and School shows bell-shaped, so we can assump
The variables x and b come from normally distributed populations

Step 1 Hypothesis.

There's a correlation between Life expectancy and schooling

- H0: $\rho = 0$ and H1: $\rho \neq 0$ (claim)

Pearson's product-moment correlation

data: Life.Exp		School
t = 43.048	df = 1647	p-value < 2.2e-16
alternative hypothesis:	true correlation is not equal to 0	
95 percent confidence interval:		0.7040885 0.7495739
sample estimates: cor		**0.72763**

Step 2 Find the critical value.

- $\alpha = 0.05$, p-value is 2.2e-16

Step 3 Compute the test value.

- P-value is 2.2e-16 < 0.05

Step 4 Make the decision.

- There is enough evidence to reject H0.

Step 5 Summarize the results.

- There's a correlation between Life expectancy and schooling

Interpretation:

1. $\alpha = 0.05$, p-value is 2.2e-16 < 0.05
2. There is a correlation between Life expectancy and schooling.
3. R-squared is 0.53, we can predict Life expectancy through schooling of 53%.

PART 6-2. CORRELATION TEST ANALYSIS

T-TEST FOR HYPOTHESIS 1
Step 0 Test Assumption.

Histogram of HIV.AIDS does not normally distributed data, so we cannot assume that two variables are from normally distributed data. Use 'Spearman' method.

Step 1 Find the value.

- Spearman coefficient is '-0.72'

Interpretation:

1. Correlation coefficients whose magnitude are between 0.5 and 0.7 indicate variables which can be considered moderately correlated. Correlation coefficients whose magnitude are between 0.3 and 0.5 indicate variables which have a low correlation (Andrews, n.d.).
2. Spearman's rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function (Aryan, 2022).
3. There is correlation between Life.Exp & GDP, and there are in negative relation

PART 6-3. CORRELATION TEST ANALYSIS

T-TEST FOR HYPOTHESIS 1
Step 0 Test Assumption.

Histogram of GDP does not normally distributed data, so we cannot assume that two variables are from normally distributed data. Use 'Spearman' method.

Step 1 Find the value.

- Spearman coefficient is '0.57'

Interpretation:

1. There is moderate correlation between Life.Exp & GDP, and there are in positive relation

PART 7-1. CHECKING COLLINEARITY

Data preparation for Correlation Analysis:

1. Finding Highly correlated values: Under5.Dth and Inf Dth is 1.0, so delete Inf.Dth from this table. Because Under.5Dth can contain Inf.Dth.
2. Finding Highly correlated values: GDP and Perct.Exp is 0.96, however the

meaning of Perct.Exp is 'Expenditure on health as a percentage of Gross Domestic Product per capita(%)'. I can interpret this that if GDP higher, it could effect positive or negative to Perct.Exp, so I will leave this.

3. Income.Com.R means that 'Human Development Index in terms of income composition of resources (index ranging from 0 to 1)'. I found that this is combination of Life Expectancy Index and Education index and Income index. Therefore, it is highly correlated with others like Life.Exp, GDP, and schooling. I will delete this value from table.
4. Ad.mortal is also highly correlated to Life expectancy, so delete it
5. Thinness of 10-19 and thinness of 5-9 is highly correlated each other, so delte one.

Delete	Leave
Inf.Dth(under5.Dth)	GDP
Income.Com.R(GDP)	Thinness of 10-19
Ad.morta7(Life.Exp)	
Thinness 5-9(Thin 10-19)	
Perct.Exp(GDP)	

Checking in VIF dataset & correlation matrix:

VIF dataset for numeric data

Ad.Motal	Inf.Dth	Alcohol	Perct.Exp	Hep.B	Measles	BMI	Under5.Dth	Polio
1.80	212.18	1.94	12.85	1.65	1.51	1.80	202.00	1.71
Total.Exp	Diphtheria	HIV.AIDS	GDP	Population	`Thin10-19`	`T5-9`	Income.Com.R	School
1.12	2.09	1.48	13.52	1.94	7.6	7.58	2.97	3.51

PART 7-2. MV REGRESSION ANALYSIS

Finding Regression model

Step 0 Checking normality

- In this case, as analyst, I decide to do with all variables in linear regression model. There are some non-normally distributed variables, but I will check the error after regression.
- In fact, linear regression analysis works well, even with non-normal errors. But, the problem is with p-values for hypothesis testing (Bommae, 2015).

Step 1 Normalization with every dataset.

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

Step 2 Make a model with every and find the highest effect values

- 1. HIV.AIDS (-0.72) 2. School (0.51) 3. GDP (0.21) 4. BMI (0.1) 5. Diphtheria (0.06) 6. `Thin10-19` (-0.05) 7. Polio (0.03) 8. Hep.B (-0.02) 9. StatusDeveloping (-0.02) +Country

Model = lm(Life.Exp~ HIV.AIDS + School + GDP + BMI + Diphtheria +`Thin10-19` + Polio + Hep.B + Status + Country, data=norm_scale)

Residuals

Min	1Q	Median	3Q	Max
-0.37243	-0.06140	0.00518	0.06727	0.29395

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5361144	0.0270543	19.816	< 2e-16 ***
HIV.AIDS***	-0.4068593	0.0181286	-22.443	< 2e-16 ***
School***	0.3923937	0.0227213	17.270	< 2e-16 ***
GDP**	0.0447520	0.0158756	2.819	0.004882**
Hep.B**	0.0180854	0.0055276	3.272	0.001093**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

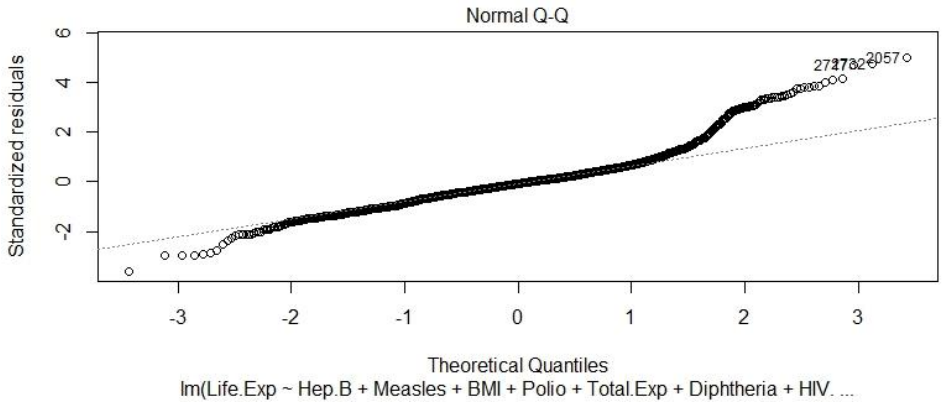
Residual standard error: 0.03984 on 1508 degrees of freedom

Multiple R-squared: **0.962**, Adjusted R-squared: **0.9585**

F-statistic: 272.7 on 140 and 1508 DF p-value: < 2.2e-16

Step 3 Checking vif value, but there is huge collinearity in Country.

- Country GVIF is 3039.05, because all other variables are made from Country. And Normal Q-Q plot with Country model is also not stable.



Step 4 Remove Country and make a model

Model = lm(Life.Exp~ HIV.AIDS + School + GDP + BMI + Diphtheria +`Thin10-19` + Polio + Hep.B + Status, data=norm_scale)

Result

Residuals

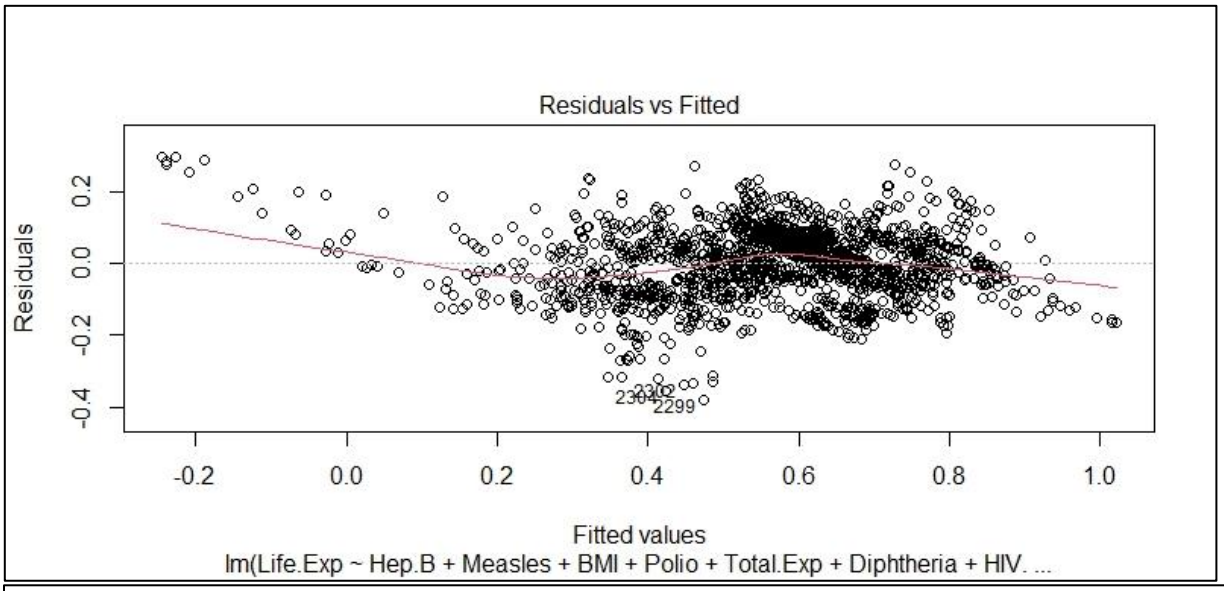
Min	1Q	Median	3Q	Max
-0.37243	-0.06140	0.00518	0.06727	0.29395

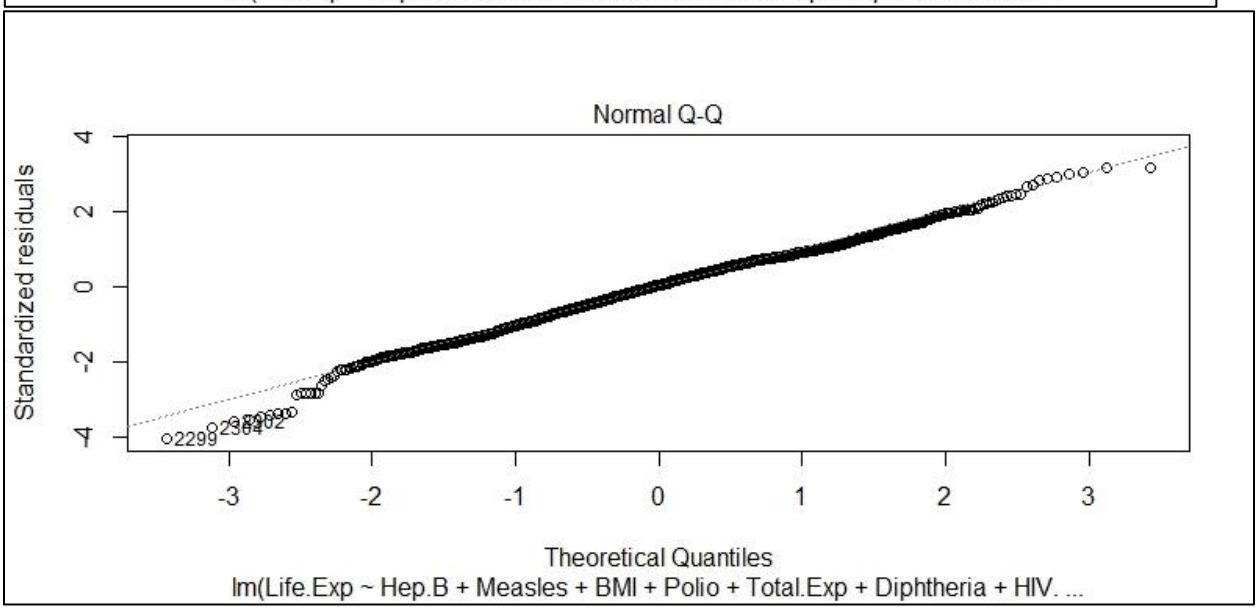
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.09513 on 1639 degrees of freedom
Multiple R-squared: **0.7645**, Adjusted R-squared: **0.7632**
F-statistic: 591.1 on 9 and 1639 DF p-value: < 2.2e-16

Equation1 of Regression model with normalized data

Life.Exp = 0.249–0.718*HIV+0.510*School + 0.217*GDP + 0.095*BMI + 0.06*Diphtheria–0.042*`Thin10-19`+0.033*Polio–0.024*Hep.B-0.021*StatusDeveloping

Plot from last model

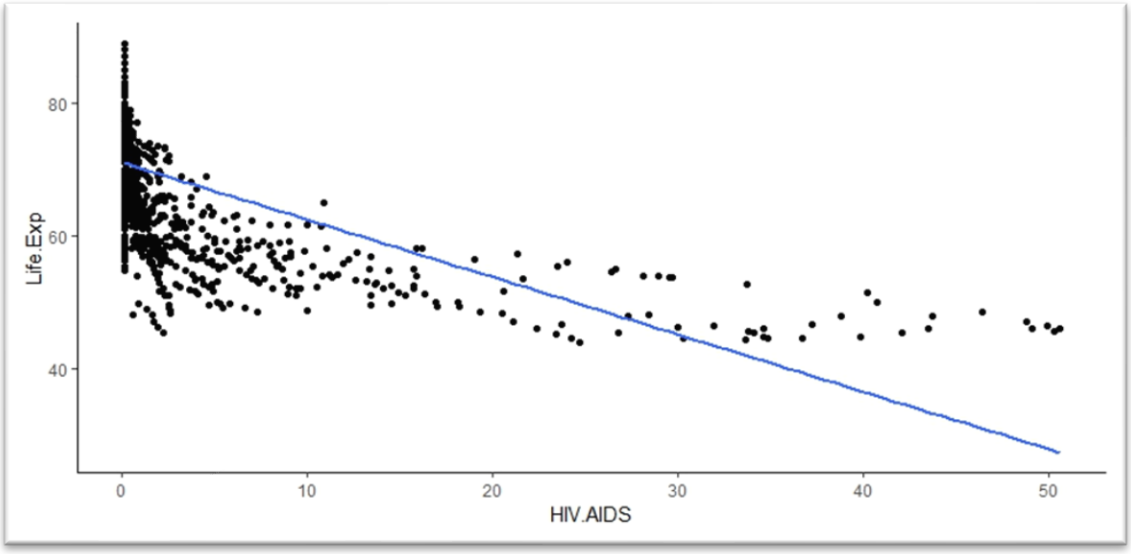


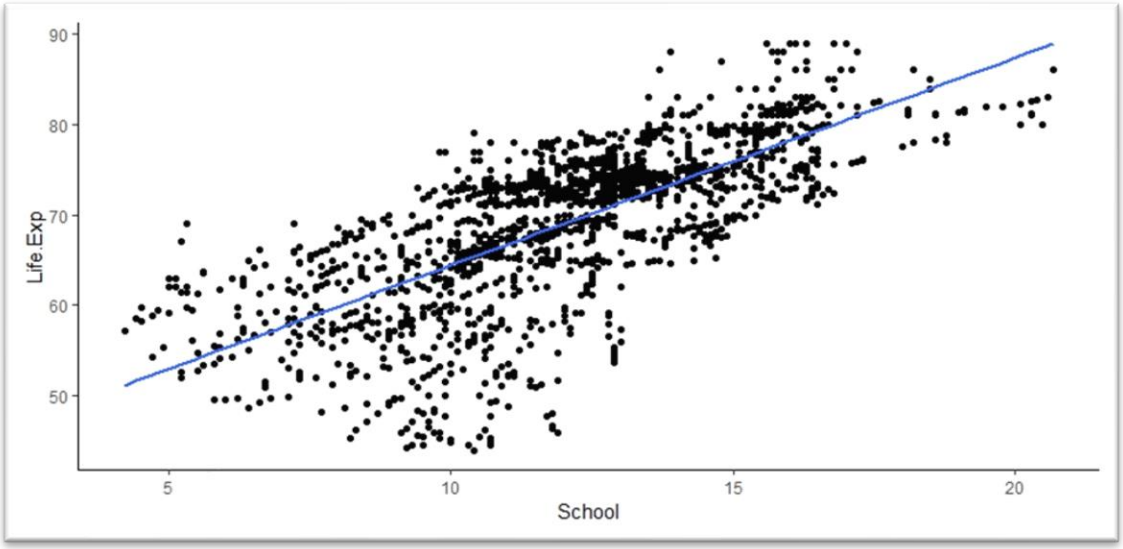


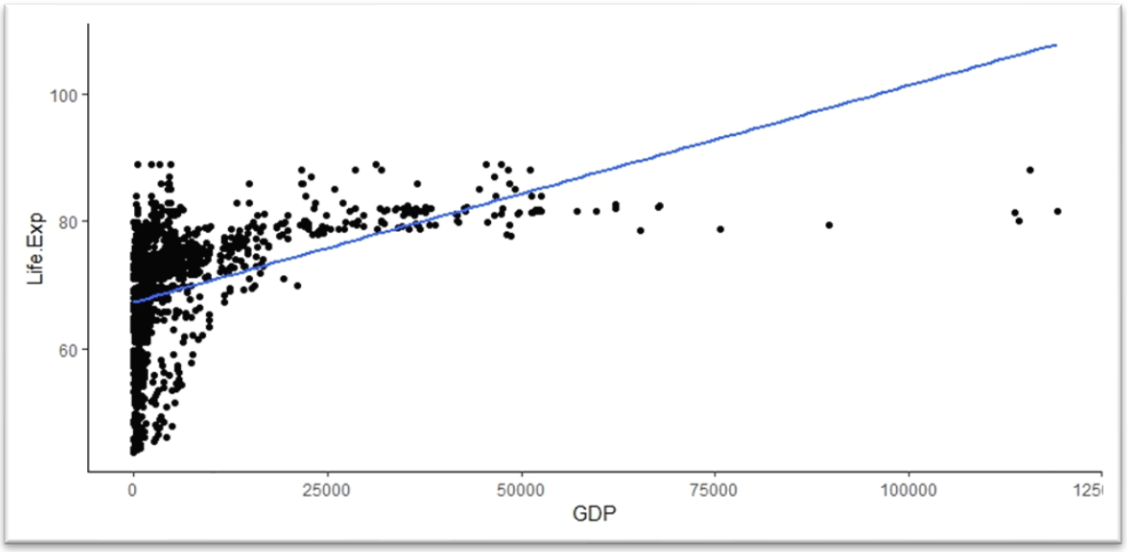
Interpretation:

1. With normalized variables, I could easily find which variable has highest impact on Life expectancy. Graphs about residuals looks stable.
2. When I add Country, the R2 will be highest. However, VIF which calculate collinearity is very high, I think, it's because every other variables is decided based on Country.
3. I want more simplification, so I find the equation 'Life.Exp= 0.261–0.749*HIV.AIDS+0.251*GDP+0.660*School. R2 for this equation is 0.7424 which is not that different from the Equation 1.

Highest correlation coefficient plots:







If x variables of Me: **84.38**(42.25-0.07+32.4+3.8) HIV=0.1, Schooling=18, GDP=42,380

Life.Exp = 48.25– 0.667*HIV.AIDS+1.8*School+0.00009*GDP

RESULT & INTERPRETATION

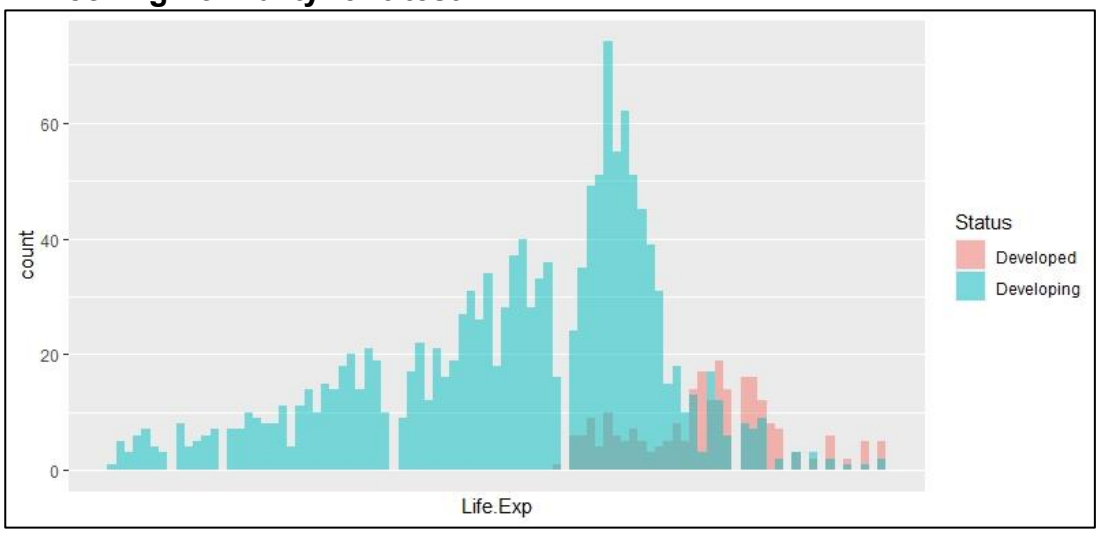
1. According to EDA & t-test, I find ‘life expectancy [Life E] is higher in developed country comparing with developed country’, ‘Life E is higher in 2010s comparing with 2000s’, and ‘Life E is higher in BMI high country’.
2. According to correlation test, I find ‘There is a positive correlation between Life E and Schooling years (by Pearson)’, ‘There is a negative correlation between Life E and HIV.AIDS (by Spearman)’, and ‘There is a moderate positive correlation between Life E and GDP’.
3. Country is very important factor to Life E in this dataset, but the collinearity of two variable is too high. It is because almost all variables are determined by country. I can get R2 0.962 which is very high and means that approximately 96% of variation in Life E can be explained by my model with country, but I cannot make model with country because of collinearity.

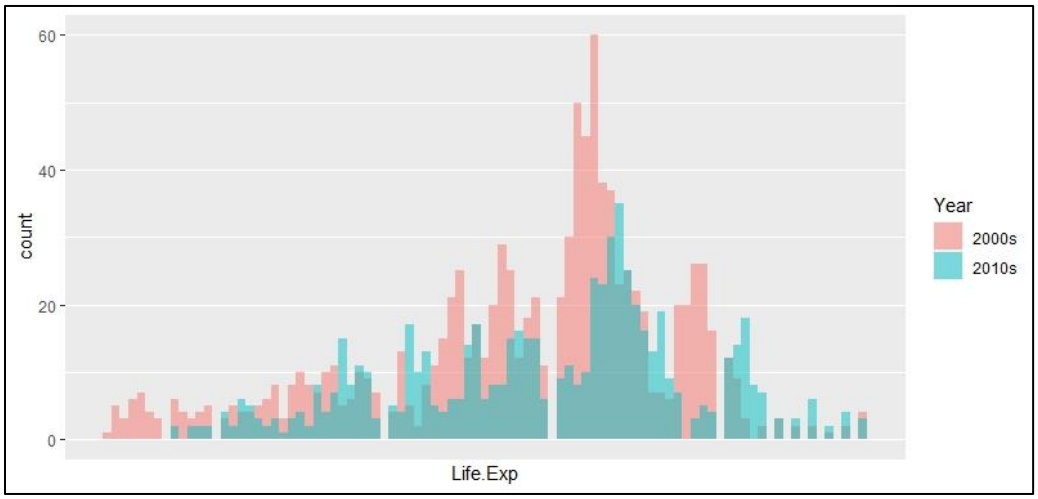
CONCLUSION

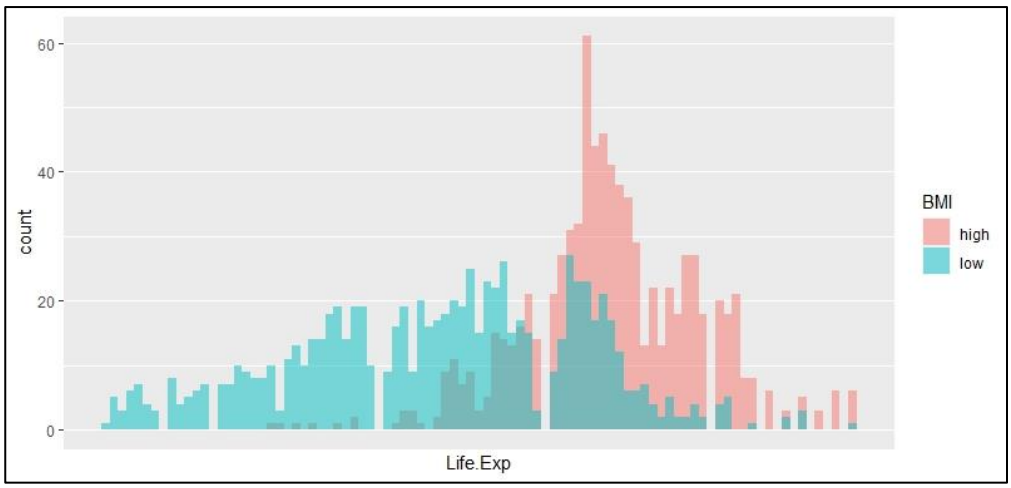
In this last project, I focused specifically on analyzing the assumptions before the test and the results after the test. Among them, it was time to learn more about 'normally distributed' and 'collinearity'. It seems there are more parts that need to be explored more deeply and logically. I realize that checking data before testing and understanding the logic of dataset itself is very important.

APPENDIX

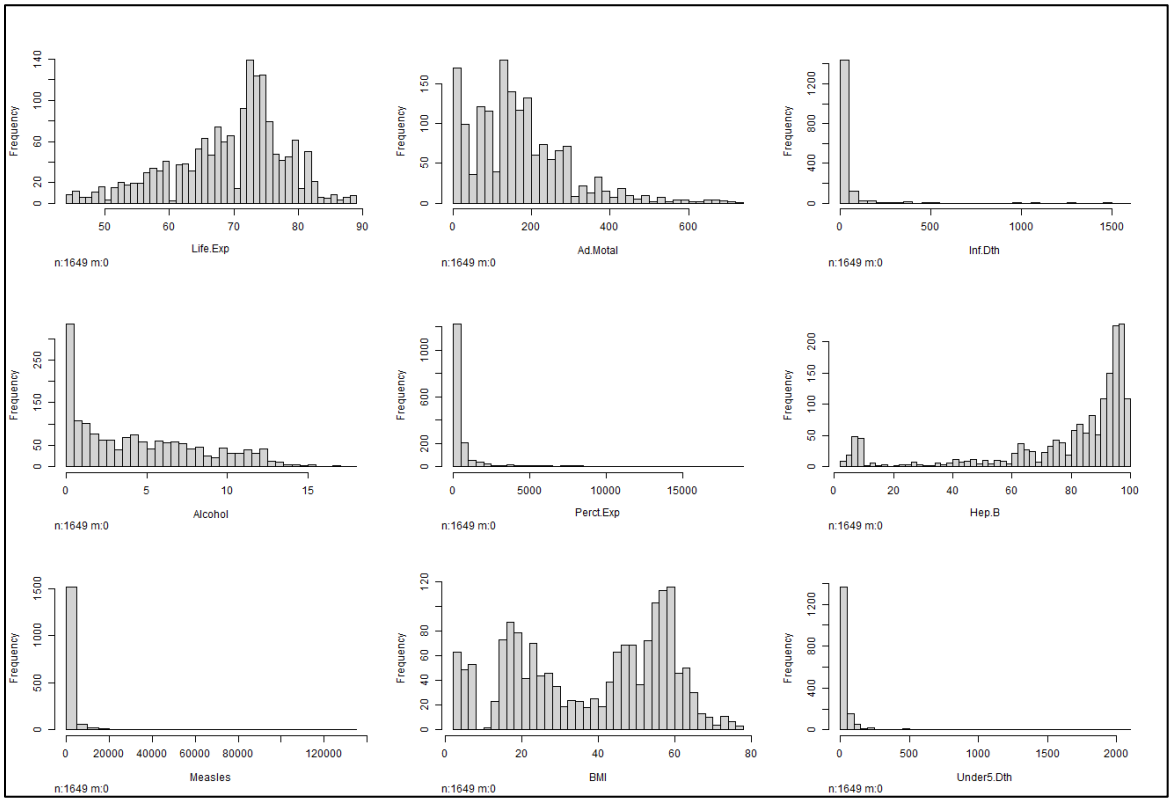
Checking normality for t-test

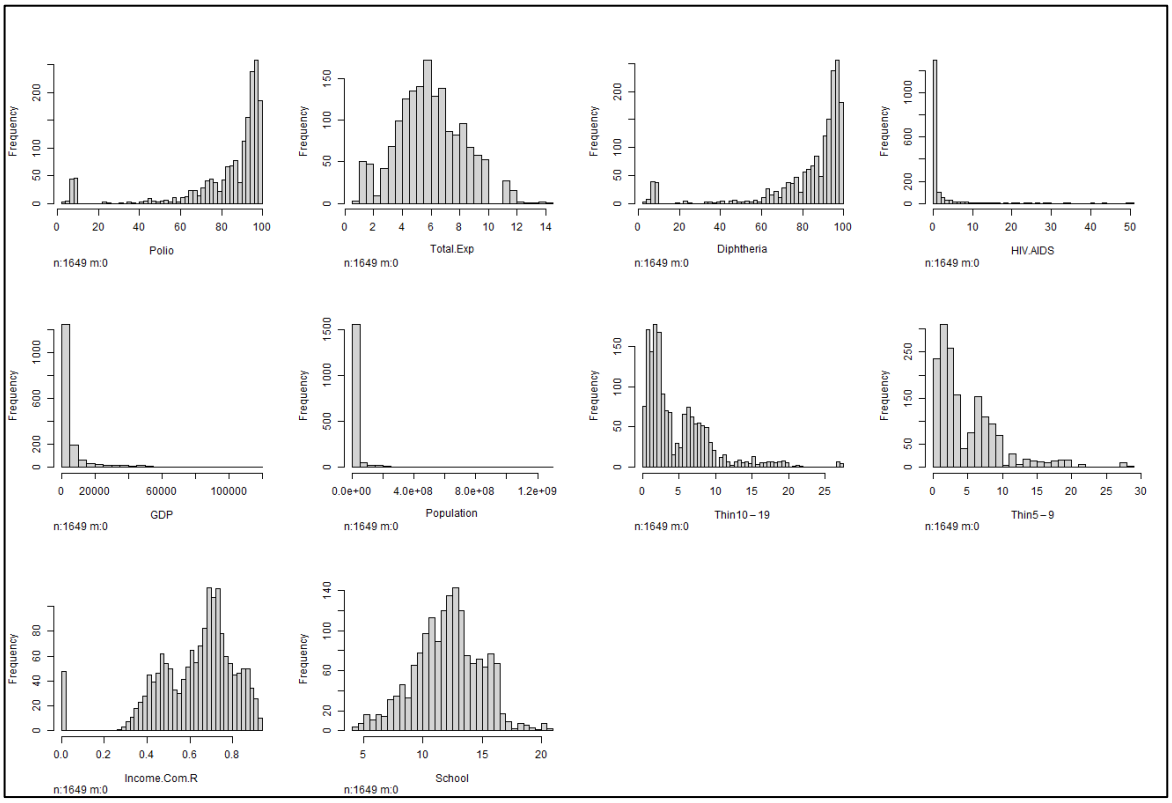




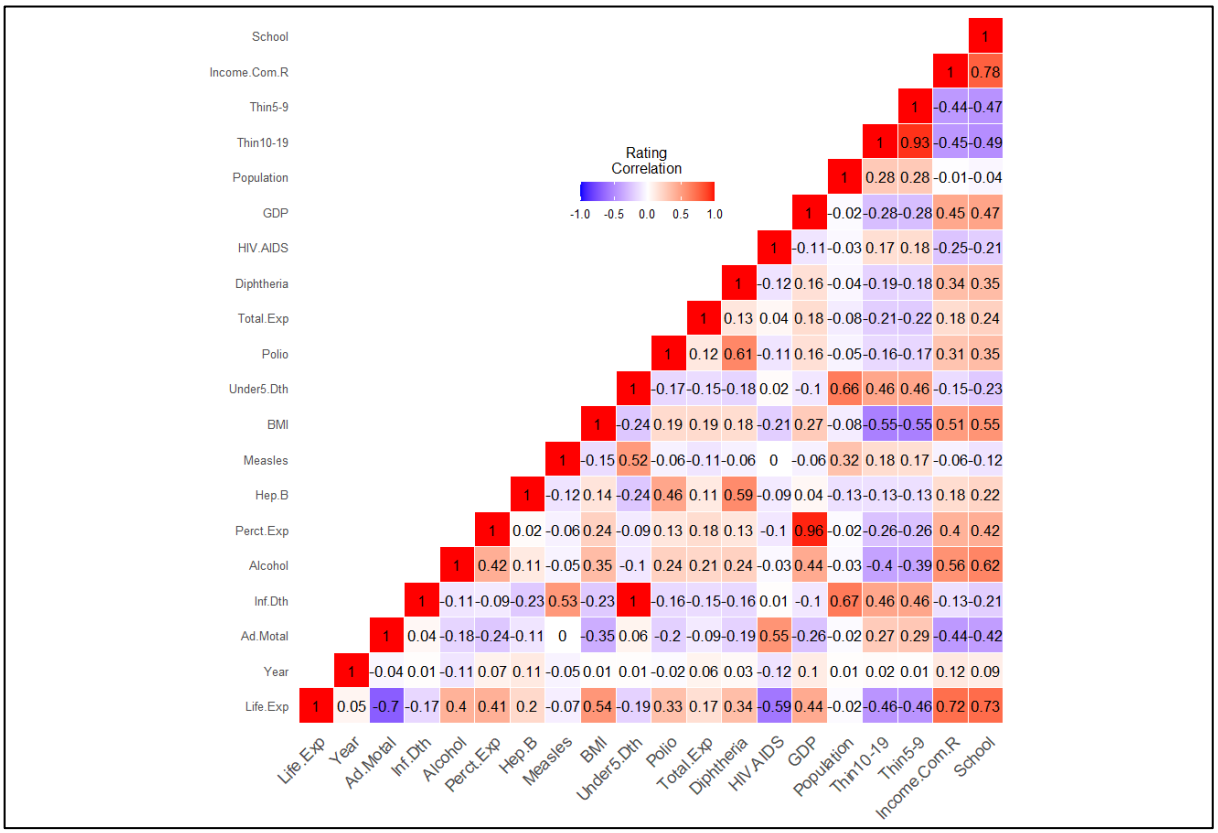


Checking histogram(normality) for correlation test

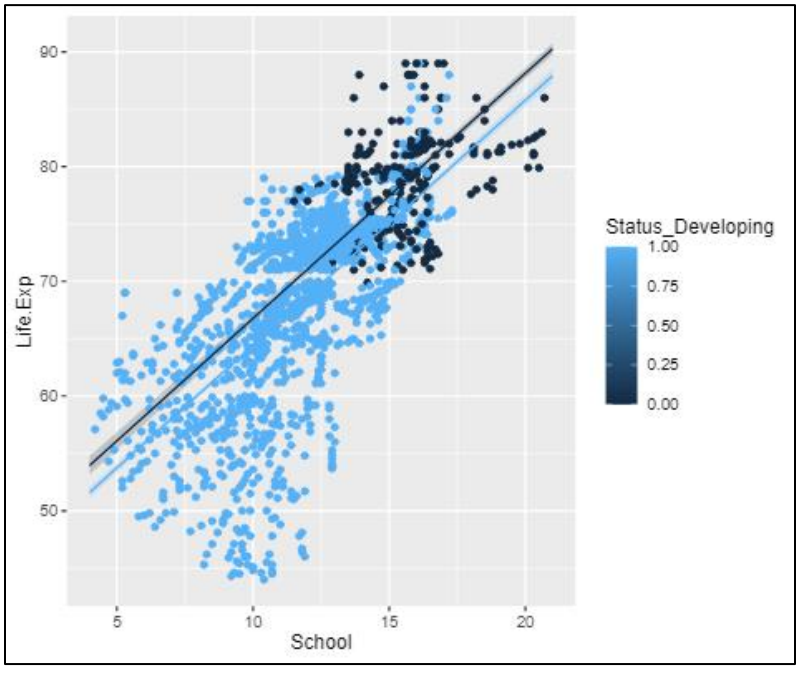




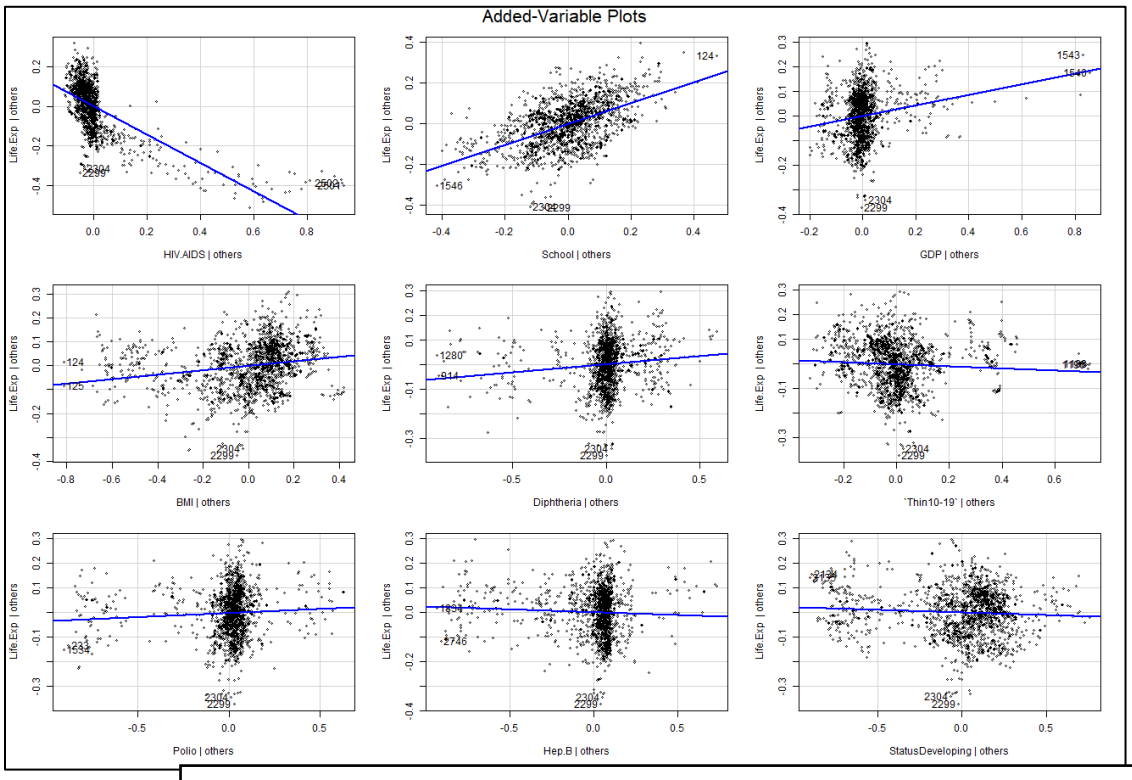
Correlation Matrix



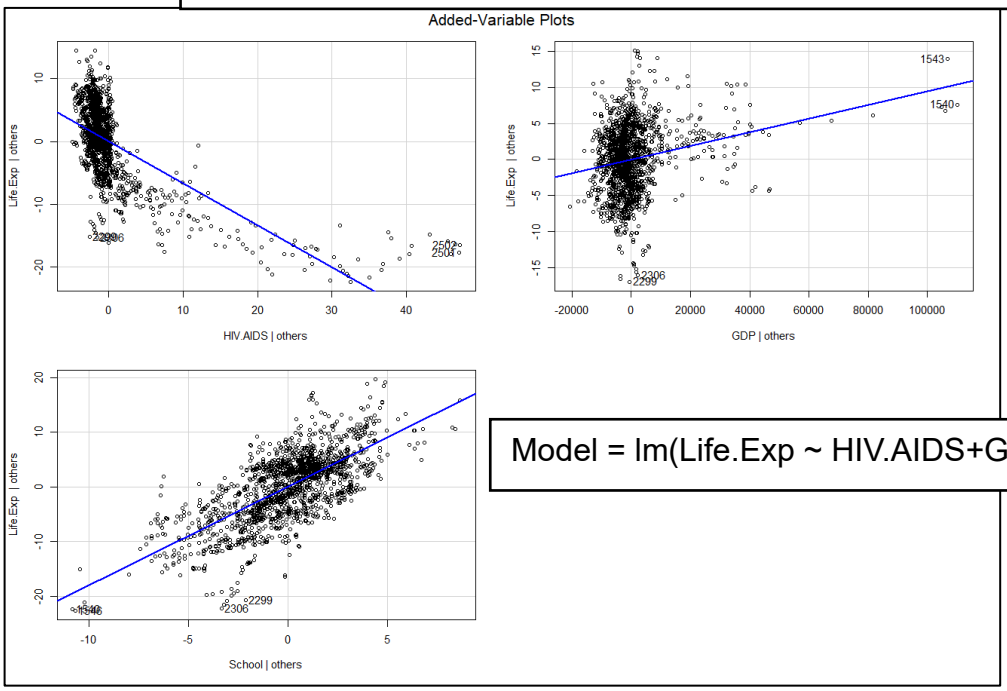
Coding with categorical variable = Developing subset graph (+School)



Avplots with MV regression model



Model = lm(Life.Exp~ HIV.AIDS + School + GDP + BMI + Diphtheria +`Thin10-19` + Polio + Hep.B + Status, data=norm_scale)



Model = lm(Life.Exp ~ HIV.AIDS+GDP+School, data = who)

REFERENCE

Bluman, Allan. (2017). Elementary statistics: a step by step approach 10th edition. McGraw-Hill.

KUMARRAJARSHI. (2018). Life expectancy (WHO). Kaggle. Retrieved from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

ggplot2. (n.d). Histograms and frequency polygons. Retrieved from https://ggplot2.tidyverse.org/reference/geom_histogram.html

NNK. (2022, December 5). How to rename column in R. SparkBy. Retrieved from https://sparkbyexamples.com/r-programming/rename-column-in-r/

STHDA. (n.d.). Correlation test between two variables in R. Retrieved from http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

ZACH. (2021, September 29). How to test for normality in R (4 methods). STATOLOGY. Retrieved from https://www.statology.org/test-for-normality-in-r/

STHDA. (n.d.). Multiple Linear Regression in R. Retrieved from http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/

ECONOMICS. (n.d.). Human development index. Retrieved from https://www.economicshelp.org/blog/glossary/human-development-index/

STHDA. (n.d.). ggplot2 Quick correlation matrix heatmap - R software and data visualization. Retrieved from http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization

Safa, Mulani. (2022, August 3). How to Normalize data in R [3 easy methods]. DigitalOcean. Retrieved from https://www.digitalocean.com/community/tutorials/normalize-data-in-r

Eugenio, Zuccarelli. (2020, July 31). Handling categorical data, the right way. Medium. Retrieved from https://towardsdatascience.com/handling-categorical-data-the-right-way-9d1279956fc6

AmiyaRanjanRout. (2022, June 30). Spearman correlation testing in R programming. geeksforgeeks. Retrieved from https://www.geeksforgeeks.org/spearman-correlation-testing-in-r-programming/

Andrews. (n.d.). Correlation Coefficients. Retrieved from https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm

Aryan Gupta. (2022, November 15). Spearman's rank correlation: the definitive guide to understand. simplilearn. Retrieved from https://www.simplilearn.com/tutorials/statistics-tutorial/spearmans-rank-correlation

Bommae, Kim. (2015, September 14). Should I always transform my variables to make them normal? University of Virginia Library. Retrieved from https://data.library.virginia.edu/normality-assumption/

ZACH. (2021, July 30). How to plot multiple histograms in R (with examples). STATOLOGY. Retrieved from https://www.statology.org/multiple-histograms-r/

ZACH. (2019, May 9). How to calculate variance inflation factor (VIF) in R. STATOLOGY. Retrieved from https://www.statology.org/variance-inflation-factor-r/

ZACH. (2020, December 23). How to plot multiple linear regression results in R. STATOLOGY. Retrieved from https://www.statology.org/plot-multiple-linear-regression-in-r/

R-Codes

```
library("psych")
library("ggplot2")
library("reshape2")
library("dplyr")
library("ggpubr")
library(tidyverse)
library(car)
library(GGally)
library(insight)
library(ggiraphExtra)
library(caret)
library(Hmisc)

#Setwd in file direction
setwd("C:\\Users\\14083\\Desktop")

#Import dataset using read.csv()
who <- read.csv("Life Expectancy Data.csv", stringsAsFactors = T,
                header=T)

#Checking dataset & data structure
headtail(who,5)
str(who)

#NOT YET change colnames
who <- who %>%
  rename('Life.Exp'='Life.expectancy',
         'Ad.Motal'='Adult.Mortality',
         'Inf.Dth'='infant.deaths',
         'Under5.Dth'='under.five.deaths',
         'Perct.Exp'='percentage.expenditure',
         'Total.Exp'='Total.expenditure',
         'Hep.B'='Hepatitis.B',
         'Thin10-19'='thinness..1.19.years',
         'Thin5-9'='thinness.5.9.years',
         'Income.Com.R'='Income.composition.of.resources',
         'School'='Schooling')

str(who)
#check NA value
who <- na.omit(who)
str(who)

#describe again
```

```
describe(who)

#Life expectancy histogram
ggplot(who, aes(Life.Exp)) +
  geom_histogram()

#Life expectancy density histogram by status
ggplot(who, aes(Life.Exp, after_stat(density), colour = Status)) +
  geom_freqpoly()

#year data
who.Y <- who

table(who$Year)
str(who.Y$Year)
who.Y$Year[who.Y$Year < 2010] <- "2000s"
who.Y$Year[who.Y$Year > 2009] <- "2010s"
who.Y$Year <- as.factor(who.Y$Year)


# plot by Year=2010s, Year=2000s
ggplot(who.Y, aes(Life.Exp, after_stat(density), colour = Year)) +
  geom_freqpoly()

str(who)
# Boxplot by developed
ggplot(who.Y, aes(x=Status, y=Life.Exp, fill=Year)) +
  geom_boxplot()

str(who)
#t.test 1 Life expectancy of Developing and Developed country
df.ping <- subset(who, Status == "Developing",
                  select = Status:Life.Exp)
df.ped <- subset(who, Status == "Developed",
                 select = Status:Life.Exp)

#rbind
df.status <- rbind(df.ping, df.ped)

str(df.status)
#Checking normality
ggplot(df.status, aes(x=Life.Exp, fill=Status)) +
  geom_histogram(binwidth=.5, alpha=.5,  position='identity') +
  scale_x_continuous(breaks=0:5)

#describeby
```

```
df.status$Status <- as.factor(df.status$Status)
describeBy(df.status,list(Status=df.status$Status))

df.ping
df.ped

#t-test
t.test(df.ped$Life.Exp, df.ping$Life.Exp, alternative = "greater", var.equal = TRUE)

str(who.Y)
#t.test 2 2010s and 2000s country
df.2010 <- subset(who.Y, Year == "2010s",
                  select = c("Year", "Life.Exp"))
df.2000 <- subset(who.Y, Year == "2000s",
                  select = c("Year", "Life.Exp"))
df.Year <- rbind(df.2010, df.2000)

str(df.Year)
#Checking normality
ggplot(df.Year, aes(x=Life.Exp, fill=Year)) +
  geom_histogram(binwidth=.5, alpha=.5,  position='identity') +
  scale_x_continuous(breaks=0:5)

#t-test
t.test(df.2010$Life.Exp, df.2000$Life.Exp, alternative = "greater", var.equal = TRUE)

# prepering BMI t-test
median(who.Y$BMI)
table(who.Y$BMI)
who.B <- who

str(who.B$BMI)
table(who.B$BMI)
who.B$BMI[who.B$BMI > 43.7] <- "high"
who.B$BMI[who.B$BMI <= 43.7] <- "low"
who.B$BMI[who.B$BMI < 8] <- "low"
who.B$BMI <- as.factor(who.B$BMI)
table(who.B$BMI)

str(who.B)

#t.test checking BMI
df.high <- subset(who.B, BMI == "high",
                  select = c("BMI", "Life.Exp"))
df.low <- subset(who.B, BMI == "low",
                  select = c("BMI", "Life.Exp"))
```

```
df.BMI <- rbind(df.high, df.low)

str(df.BMI)
#Checking normality
ggplot(df.BMI, aes(x=Life.Exp, fill=BMI)) +
  geom_histogram(binwidth=.5, alpha=.5,  position='identity') +
  scale_x_continuous(breaks=0:5)

#t.test 3 2010s and 2000s country
df.B.low <- subset(who.B, BMI == "low",
                   select = c("BMI", "Life.Exp"))
df.B.high <- subset(who.B, BMI == "high",
                    select = c("BMI", "Life.Exp"))

df.B.low
df.B.high

#t-test
t.test(df.B.high$Life.Exp, df.B.low$Life.Exp, alternative = "less", var.equal = TRUE)

#t-test
t.test(df.B.high$Life.Exp, df.B.low$Life.Exp, alternative = "greater", var.equal = TRUE)

str(who)

#extract numeric data only by who data
nu.who <- subset(who, select = c(4:22))
str(nu.who)

#checking histogram
str(nu.who)
nu.who1 <- subset(nu.who, ,select=c(1:9))
nu.who2 <- subset(nu.who, ,select=c(10:19))
nu.who1
hist.data.frame(nu.who1)
hist.data.frame(nu.who2)

#Correlation heatmap
cordata3 <- subset(who, select = c(4, 2, 7, 9:19, 22))
cordata3

aR <- round(cor(cordata3), 2)

cor(cordata3)
aR <- round(cor(cordata3), 2)
```

```
get_upper_tri<-function(aR){
  aR[lower.tri(aR)] <- NA
  return(aR)
}

upper_tri <- get_upper_tri(aR)
upper_tri
melted_cormat <- melt(upper_tri, na.rm = TRUE)
melted_cormat

reorder_cormat <- function(aR){
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  aR <-aR[hc$order, hc$order]
}

aR <- reorder_cormat(aR)
aR
upper_tri <- get_upper_tri(aR)

melted_cormat <- melt(upper_tri, na.rm = TRUE)

#NEW
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Rating\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()

ggheatmap +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
```

```
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))

#scatter plot
ggscatter(who, x = "School", y = "Life.Exp",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Number of years of Schooling", ylab = "Life Expectency")


#histogram of LIfe.Exp
hist(who$Life.Exp, col='steelblue', main='Normal')
hist(who$School, col='steelblue', main='Normal')

#correlation test between Life Expectency and schooling
res <- cor.test(who$Life.Exp, who$School,
                method = "pearson")
cor(who$Life.Exp, who$School)*cor(who$Life.Exp, who$School)

# TEST 3

#scatter plot
ggscatter(who, x = "HIV.AIDS", y = "Life.Exp",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Deaths per 1 000 live births HIV/AIDS (0-4 years)", ylab = "Life Expectency")

#HIV.AIDS subset from who
#histogram of LIfe.Exp
hist(who$Life.Exp, col='steelblue', main='Normal')
hist(who$HIV.AIDS, col='steelblue', xlim=c(0,1), main='Normal')
table(who$HIV.AIDS)

#Spearman test
result <- cor(who$Life.Exp, who$HIV.AIDS, method = "spearman")
print(result)


# TEST 4

#scatter plot
ggscatter(who, x = "HIV.AIDS", y = "Life.Exp",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Deaths per 1 000 live births HIV/AIDS (0-4 years)", ylab = "Life Expectency")
```

```
#HIV.AIDS subset from who
#histogram of LIfe.Exp
hist(who$Life.Exp, col='steelblue', main='Normal')
hist(who$GDP, col='steelblue', main='Normal')
table(who$HIV.AIDS)

#Spearman test
result <- cor(who$Life.Exp, who$GDP, method = "spearman")
print(result)

#histogram of LIfe.Exp
hist(who$Life.Exp, col='steelblue', main='Normal')
hist(who$BMI, col='steelblue', main='Normal')

hist(nu.who$School)

str(nu.who)
#Checking VIF again
reg_mod5 <- lm(Life.Exp ~
Alcohol+Hep.B+Measles+BMI+Polio+Total.Exp+Diphtheria+HIV.AIDS+GDP+Population+`Thin1
0-19`+School, data = nu.who)
summary(reg_mod5)
vif_mod5 <- vif(reg_mod5)

#create horizontal bar chart to display each VIF value
opar <- par(no.readonly = TRUE)
par(fig=c(0.2, 1, 0, 1))
barplot(vif_mod5, main = "VIF Values", horiz = TRUE, las=2, cex.names=0.8, xlim=c(0,5), col =
"steelblue")

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)

# normalization
summary(who)
process <- preProcess(as.data.frame(who), method=c("range"))

norm_scale <- predict(process, as.data.frame(who))
norm_scale

headtail(norm_scale)
str(norm_scale)


# Prediction with normalized data with categorical variables
reg_mod6 <- lm(Life.Exp ~
```

```
Hep.B+Measles+BMI+Polio+Total.Exp+Diphtheria+HIV.AIDS+GDP+Population+`Thin10-
19`+School+Status+Country, data = norm_scale)
summary(reg_mod6)
vif_mod6 <- vif(reg_mod6)
round(vif(reg_mod6),2)
plot(reg_mod6)

# Prediction with normalized data without country
reg_mod60 <- lm(Life.Exp ~
Hep.B+Measles+BMI+Polio+Total.Exp+Diphtheria+HIV.AIDS+GDP+Population+`Thin10-
19`+School+Status, data = norm_scale)
summary(reg_mod60)
vif_mod6 <- vif(reg_mod60)
round(vif(reg_mod60),2)
plot(reg_mod60)

#Scatter plot for highest value variable
ggplot(who,aes(x=HIV.AIDS,y=Life.Exp))+
  geom_point()+
  theme_classic()+
  geom_smooth(method=lm,se=FALSE,fullrange=TRUE)

ggplot(who,aes(x=School,y=Life.Exp))+
  geom_point()+
  theme_classic()+
  geom_smooth(method=lm,se=FALSE,fullrange=TRUE)

ggplot(who,aes(x=GDP,y=Life.Exp))+
  geom_point()+
  theme_classic()+
  geom_smooth(method=lm,se=FALSE,fullrange=TRUE)

# Simpliest model
reg_mod62 <- lm(Life.Exp ~ HIV.AIDS+GDP+School+Status, data = norm_scale)
summary(reg_mod62)
vif_mod6 <- vif(reg_mod62)
round(vif(reg_mod62),2)
plot(reg_mod62)

# not norm
reg_mod61 <- lm(Life.Exp ~ Country +
Hep.B+Measles+BMI+Polio+Total.Exp+Diphtheria+HIV.AIDS+GDP+Population+`Thin10-19`,
data = who)
summary(reg_mod61)
vif_mod61 <- vif(reg_mod61)
round(vif(reg_mod61),2)
```

```
plot(reg_mod61)

# only conuntry.  for checking
reg_mod7 <- lm(Life.Exp ~ Country+Status, data = norm_scale)
summary(reg_mod7)
vif_mod7 <- vif(reg_mod7)
round(vif(reg_mod7),2)

#with Country model
last_mod <- lm(Life.Exp~ HIV.AIDS + School + GDP + BMI +  Diphtheria +`Thin10-19` + Polio +
Hep.B + Country, data=norm_scale)
summary(last_mod)
vif_values <- vif(last_mod)
vif_values

#without Country model
last_mod1 <- lm(Life.Exp~ HIV.AIDS + School + GDP + BMI +  Diphtheria +`Thin10-19` + Polio
+ Hep.B + Status, data=norm_scale)
summary(last_mod1)
vif_values <- vif(last_mod1)
vif_values

#Coding for categorical variables
table(who$Status_Developing)
who$Status_Developing <- ifelse(who$Status == "Developing", 1, 0)
reg_dum<- lm(Life.Exp ~ School+Status_Developing, data=who)
ggPredict(reg_dum, se=TRUE, interactive=TRUE)

plot(reg_mod0)
plot(reg_mod6)

avPlots(reg_mod0)
avPlots(last_mod1)
```