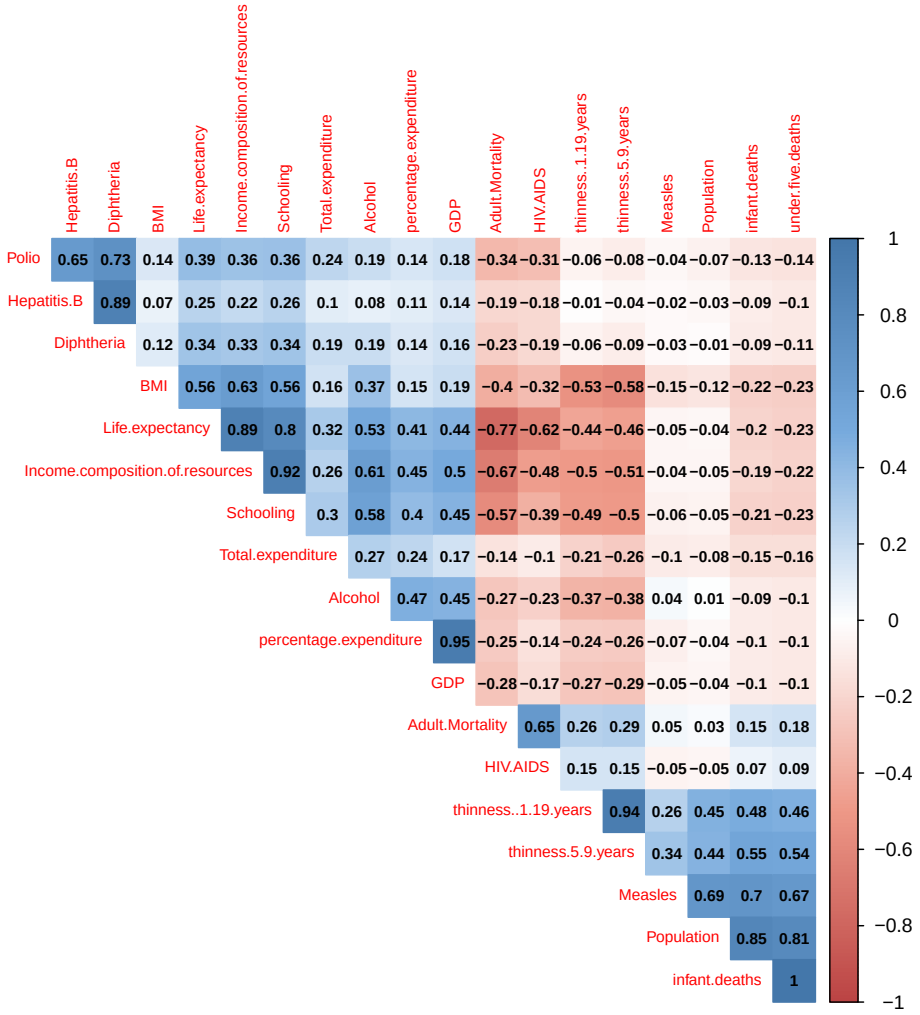
Primary relationship of interest

Team Orange

2022-10-19

Primary relationship of interest

Full correlation map (Maybe in appendix):



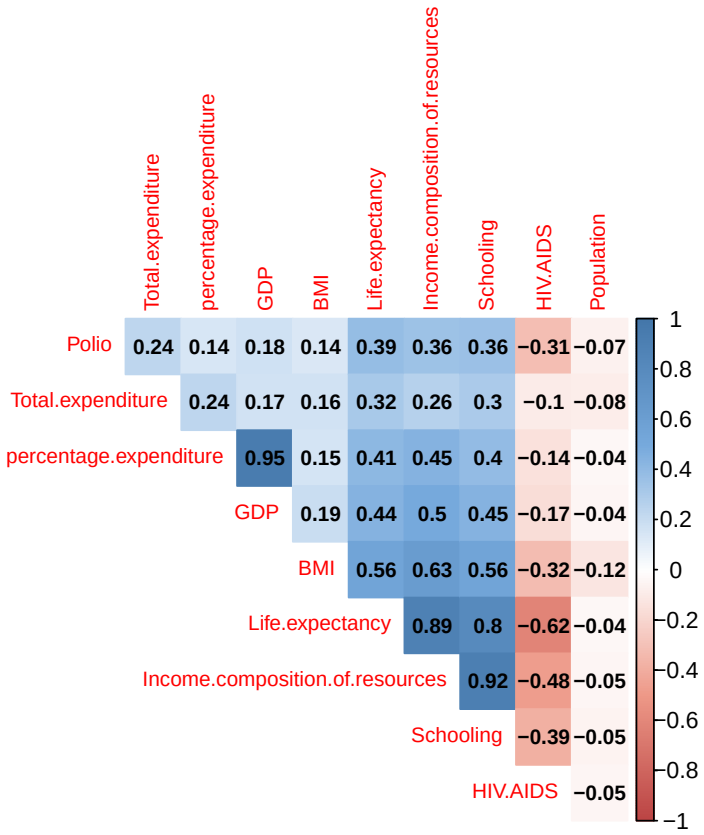
After piror selection:

To represent immunization coverage, it is not wise use all variables when they are highly correlated with each other; among “Hepatitis.B”,“Polio”, “Diphtheria tetanus toxoid and pertussis (DTP3)”, we decide to use “polio”, since it has the highest correlation between Life Expectancy

On the one hand, with domain knowledge, we know “Adult.Mortality”, “infant.deaths” and “under.five.deaths” variables are directly correlated to Life Expectancy, we choose to drop them from the predictor variable list; on the other hand, we are interested in “HIV.AIDS” variable (Deaths per 1 000 live births HIV/AIDS (0-4 years)

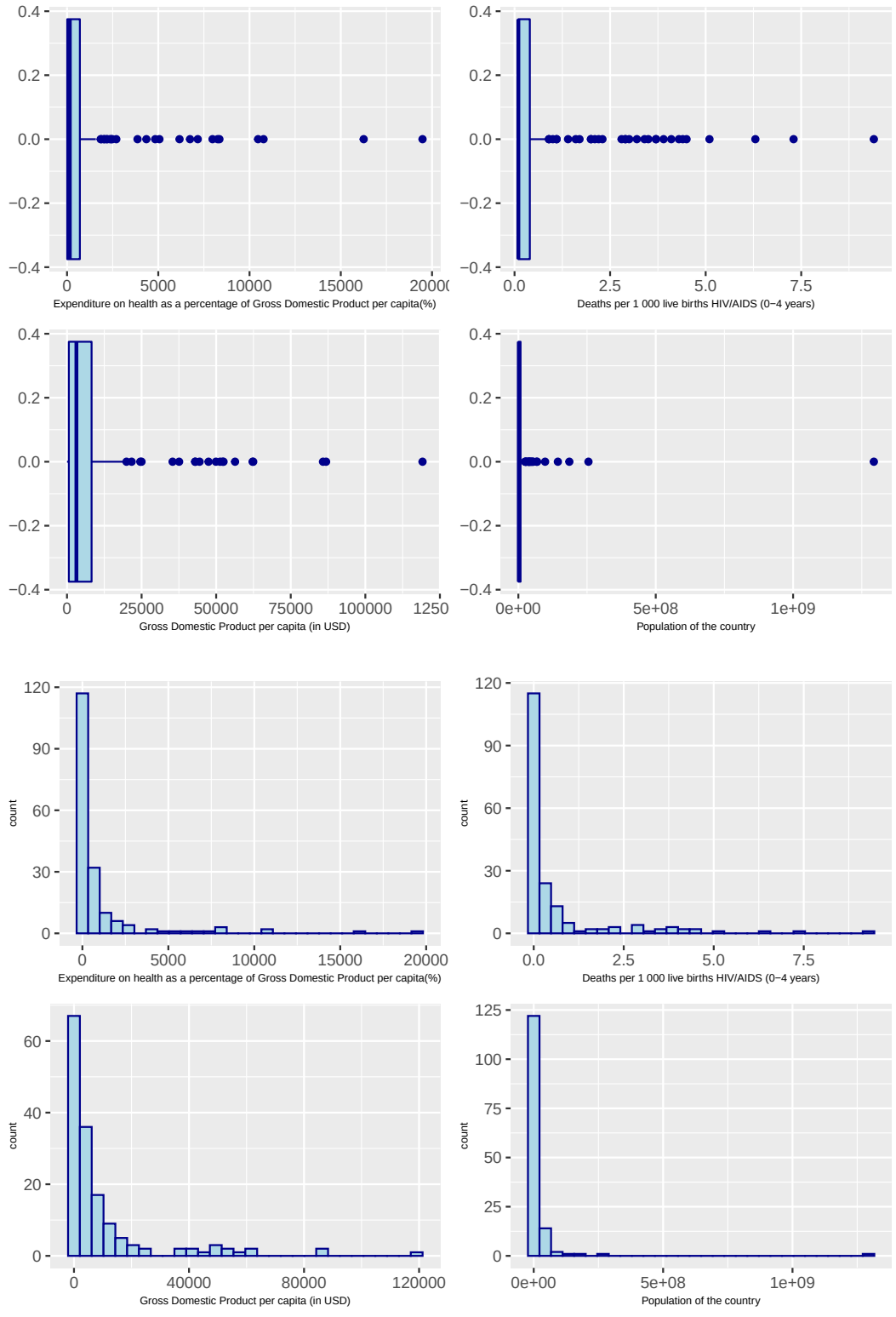
We would like to omit the variables have Low correlation between Life Expectancy: “Measles”; however, we do want to include “population” because of our interest.

As for categorical variable, we would like to keep country status (developing/developed) as one of the predictors



Transformation if needed when modeling

We may want to use log transformation for population and GDP, since the magnitude of gaps are huge. But for other two, we need more investigation, plus the difficulty of interpretation.

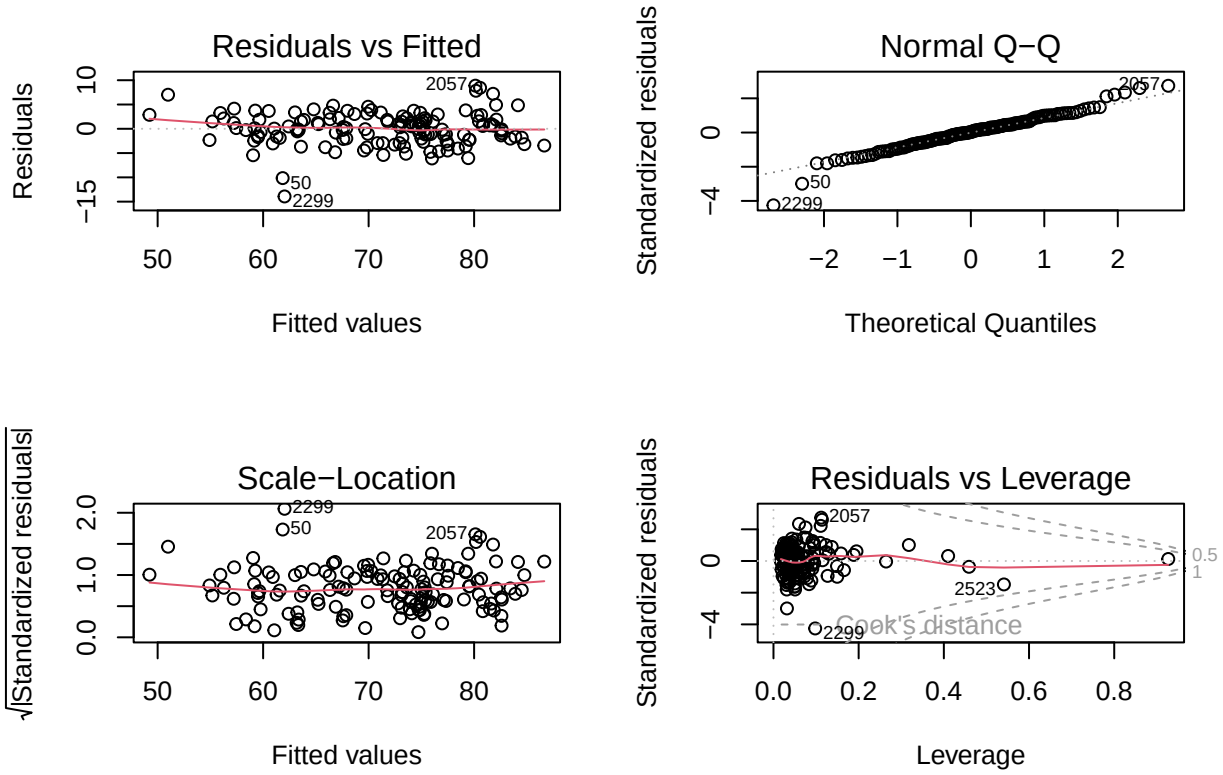


(For our reference)

After selection Model:

Table 1: Regression Summary

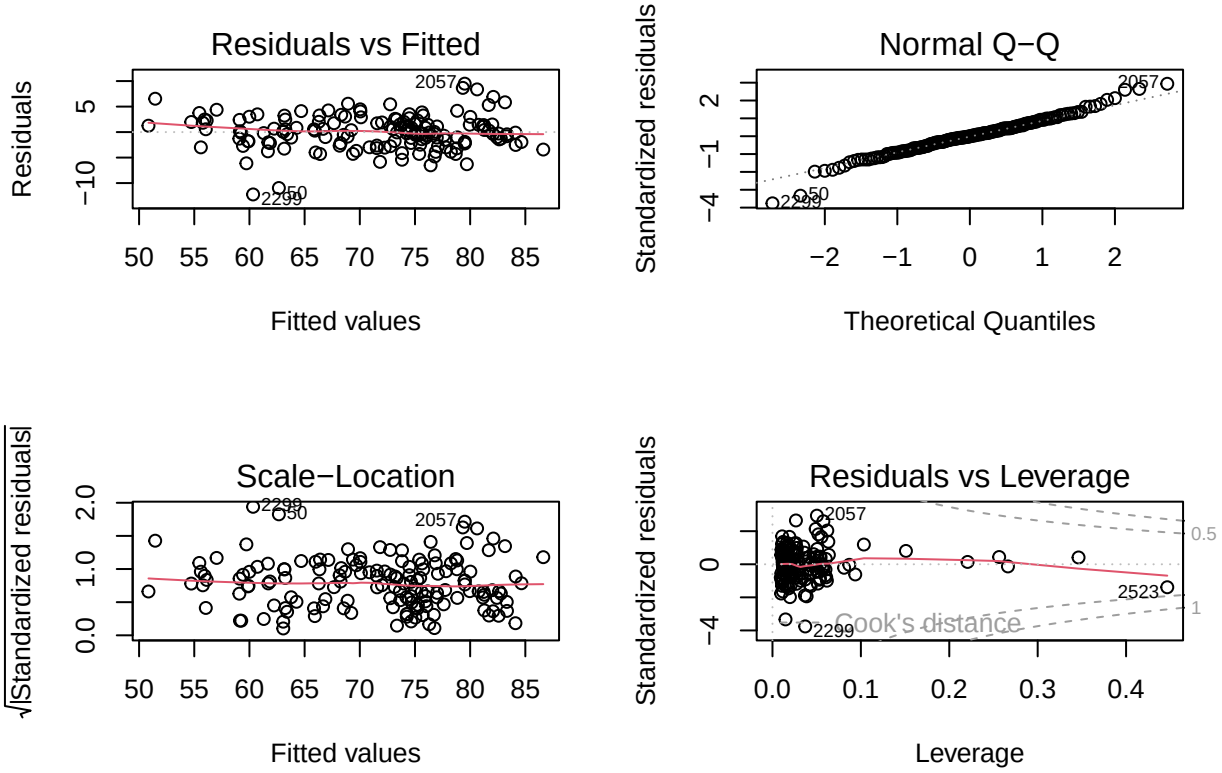
	Dependent variable:
	Life.expectancy
BMI	−0.002 (0.018)
	p = 0.930
GDP	−0.00002 (0.0001)
	p = 0.744
percentage.expenditure	0.0002 (0.0003)
	p = 0.625
Polio	0.010 (0.015)
	p = 0.493
HIV.AIDS	−1.358 (0.228)
	p = 0.00000***
Total.expenditure	0.272 (0.123)
	p = 0.030**
Population	−0.000 (0.000)
	p = 0.900
Income.composition.of.resources	43.748 (5.666)
	p = 0.000***
StatusDeveloping	−0.679 (0.993)
	p = 0.496
Schooling	−0.107 (0.270)
	p = 0.695
Constant	41.852 (2.715)
	p = 0.000***
Observations	139
R^2	0.861
Adjusted R^2	0.851
Residual Std. Error	3.445 (df = 128)
F Statistic	79.607*** (df = 10; 128)
Note:	*p<0.1; **p<0.05; ***p<0.01



Potential Final Model :

Table 2: Regression Summary

	Dependent variable:
	Life.expectancy
GDP	−0.00005 (0.00004)
	p = 0.206
percentage.expenditure	0.0004 (0.0002)
	p = 0.145
HIV.AIDS	−1.392 (0.211)
	p = 0.000***
Income.composition.of.resources	42.536 (2.551)
	p = 0.000***
StatusDeveloping	−1.006 (0.880)
	p = 0.256
Constant	44.053 (2.229)
	p = 0.000***
Observations	154
R^2	0.862
Adjusted R^2	0.857
Residual Std. Error	3.315 (df = 148)
F Statistic	184.448*** (df = 5; 148)
Note:	*p<0.1; **p<0.05; ***p<0.01


Residuals vs Fitted
Residuals
Fitted values
2057
50
2299
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
0.5
1
2523