DA546: INTRODUCTION TO STATISTICAL LEARNING

REGRESSION ANALYSIS ON FACTORS AFFECTING LIFE EXPECTANCY

SUBMITTED TO

RHYTHM GROVER

ASSISTANT PROFESSOR

MEHTA FAMILY SCHOOL OF DATA SCIENCE AND AI

SUBMITTED BY

RAJENDRA KUJUR (214161008)

M.TECH DATA SCIENCE

MOTIVATION

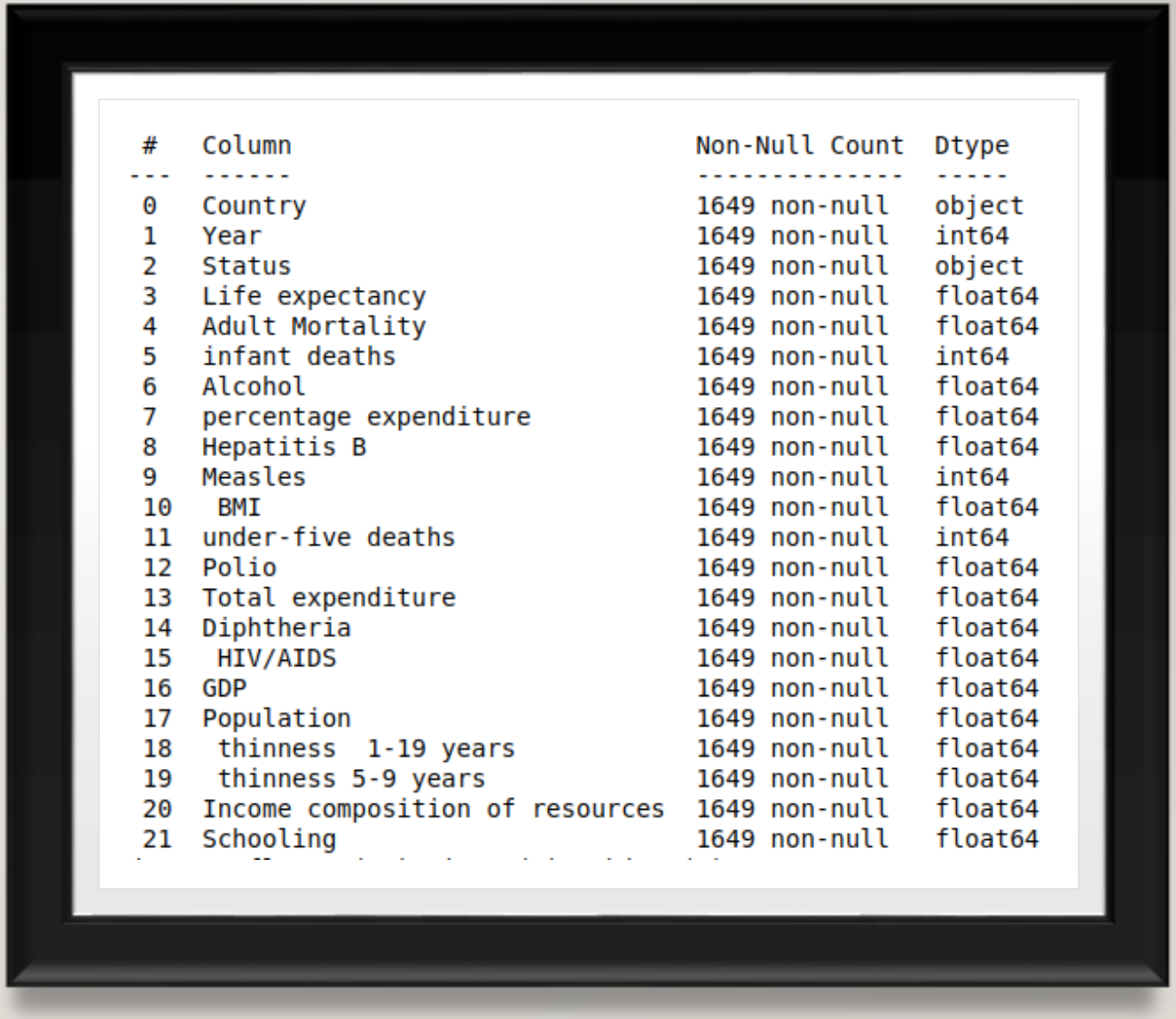
- Should a country having a lower life expectancy value(<65) increase its healthcare expenditure in order to improve its average lifespan?
- How does Infant and Adult mortality rates affect life expectancy?
- Does Life Expectancy has positive or negative correlation with eating habits, lifestyle, exercise.
- Does Life Expectancy have positive or negative relationship with drinking alcohol?
- Do densely populated countries tend to have lower life expectancy?

SOURCE OF DATASET

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The data sets are made available to public for the purpose of health data analysis. The data set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website

ABOUT DATA SET

- We have total 22 features, out of which one feature is Life expectancy (target variable), rest features (independent variables)

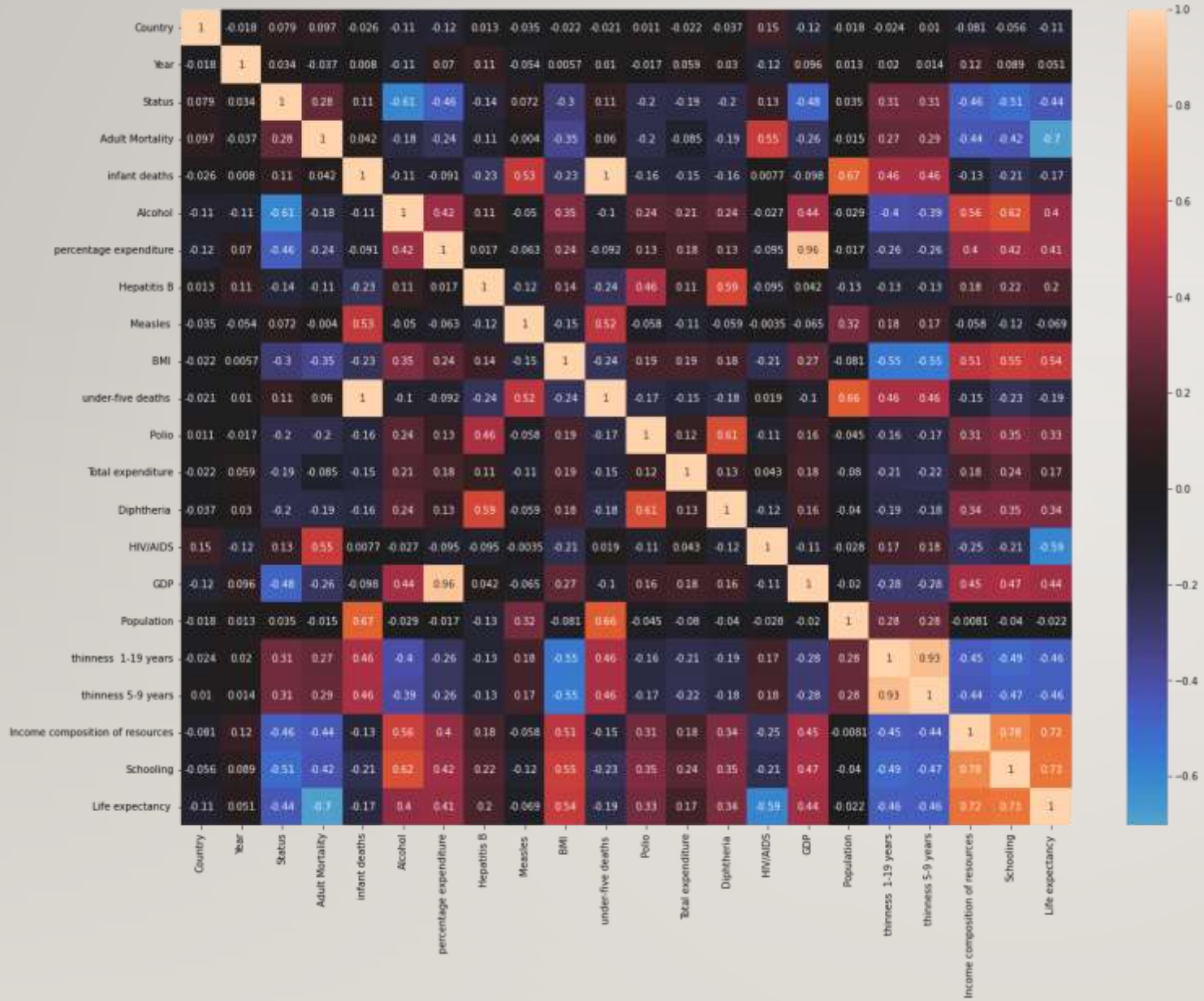


#	Column	Non-Null Count	Dtype
0	Country	1649 non-null	object
1	Year	1649 non-null	int64
2	Status	1649 non-null	object
3	Life expectancy	1649 non-null	float64
4	Adult Mortality	1649 non-null	float64
5	infant deaths	1649 non-null	int64
6	Alcohol	1649 non-null	float64
7	percentage expenditure	1649 non-null	float64
8	Hepatitis B	1649 non-null	float64
9	Measles	1649 non-null	int64
10	BMI	1649 non-null	float64
11	under-five deaths	1649 non-null	int64
12	Polio	1649 non-null	float64
13	Total expenditure	1649 non-null	float64
14	Diphtheria	1649 non-null	float64
15	HIV/AIDS	1649 non-null	float64
16	GDP	1649 non-null	float64
17	Population	1649 non-null	float64
18	thinness 1-19 years	1649 non-null	float64
19	thinness 5-9 years	1649 non-null	float64
20	Income composition of resources	1649 non-null	float64
21	Schooling	1649 non-null	float64

DATA PRE-PROCESSING

Which reduced our dataset from 2938 x 22 to 1649 x 22

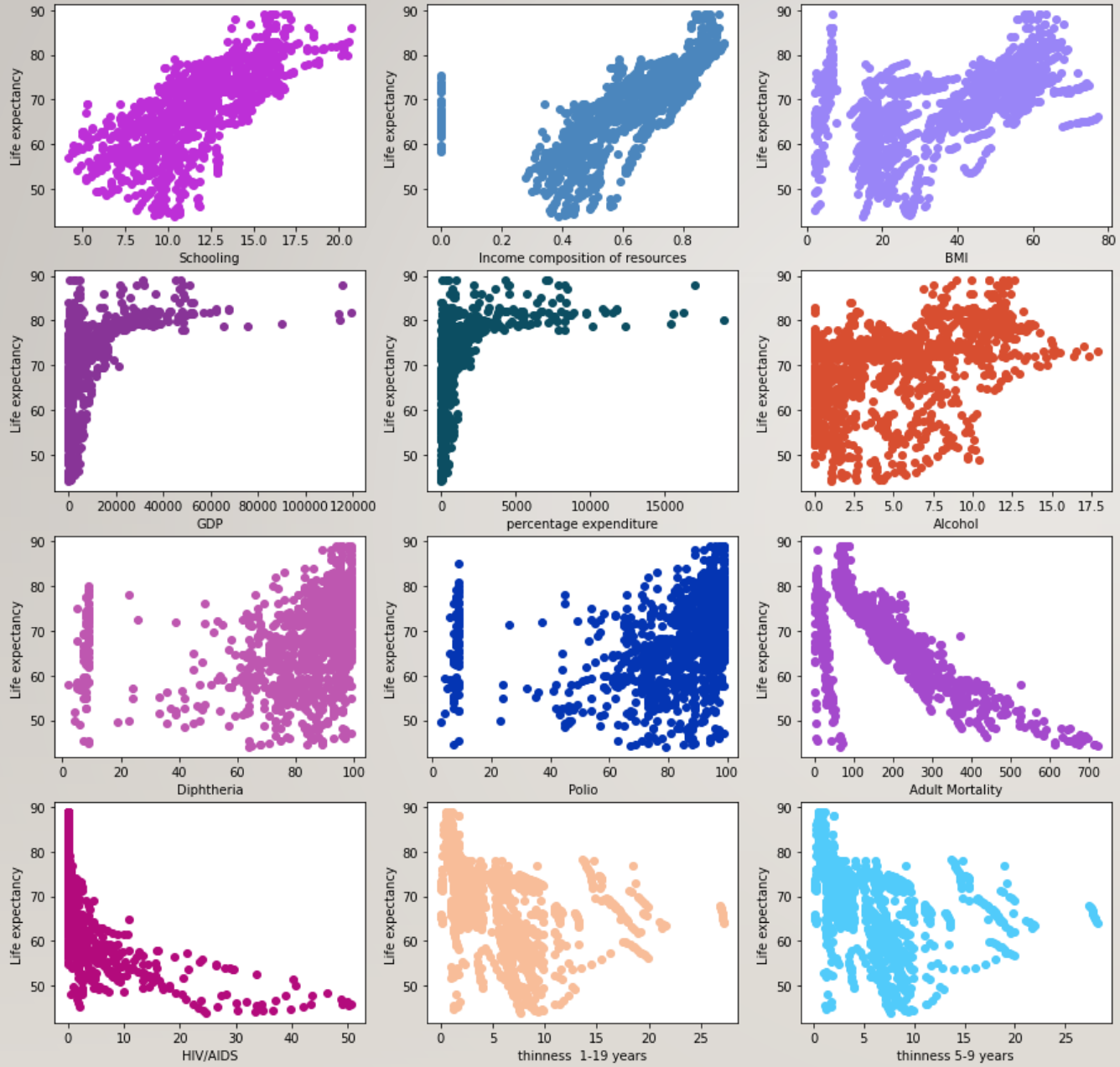
- Removed the missing value rows
- Categorical Values updated to numeric values
- Country Values: 0, 1, 2 …. N
- Status: 0(Developing), 1(Developed)



CORRELATION BETWEEN ALL THE ALL THE FEATURES

Higher the correlation value, better the linear association-ship between two features

- **Schooling**
- **Income composition of resources**
- **Adult Mortality**
- **HIV/AIDS**
- **BMI**



SCATTER PLOT WITH THE HIGHLY CO-RRELATED FEATURES

Features in decreasing order of correlation coefficient

FINAL OUTPUT MODEL

OLS Regression Results

Dep. Variable:	Life expectancy	R-squared:	0.825
Model:	OLS	Adj. R-squared:	0.824
Method:	Least Squares	F-statistic:	928.7
Date:	Fri, 29 Apr 2022	Prob (F-statistic):	0.00
Time:	19:34:52	Log-Likelihood:	-2680.7
No. Observations:	989	AIC:	5373.
Df Residuals:	983	BIC:	5403.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	52.2870	0.665	78.627	0.000	50.982	53.592
Adult Mortality	-0.0175	0.001	-14.457	0.000	-0.020	-0.015
Income composition of resources	12.2441	1.069	11.449	0.000	10.146	14.343
Schooling	0.9775	0.069	14.106	0.000	0.842	1.113
HIV/AIDS	-0.5211	0.025	-20.562	0.000	-0.571	-0.471
BMI	0.0346	0.007	4.877	0.000	0.021	0.049

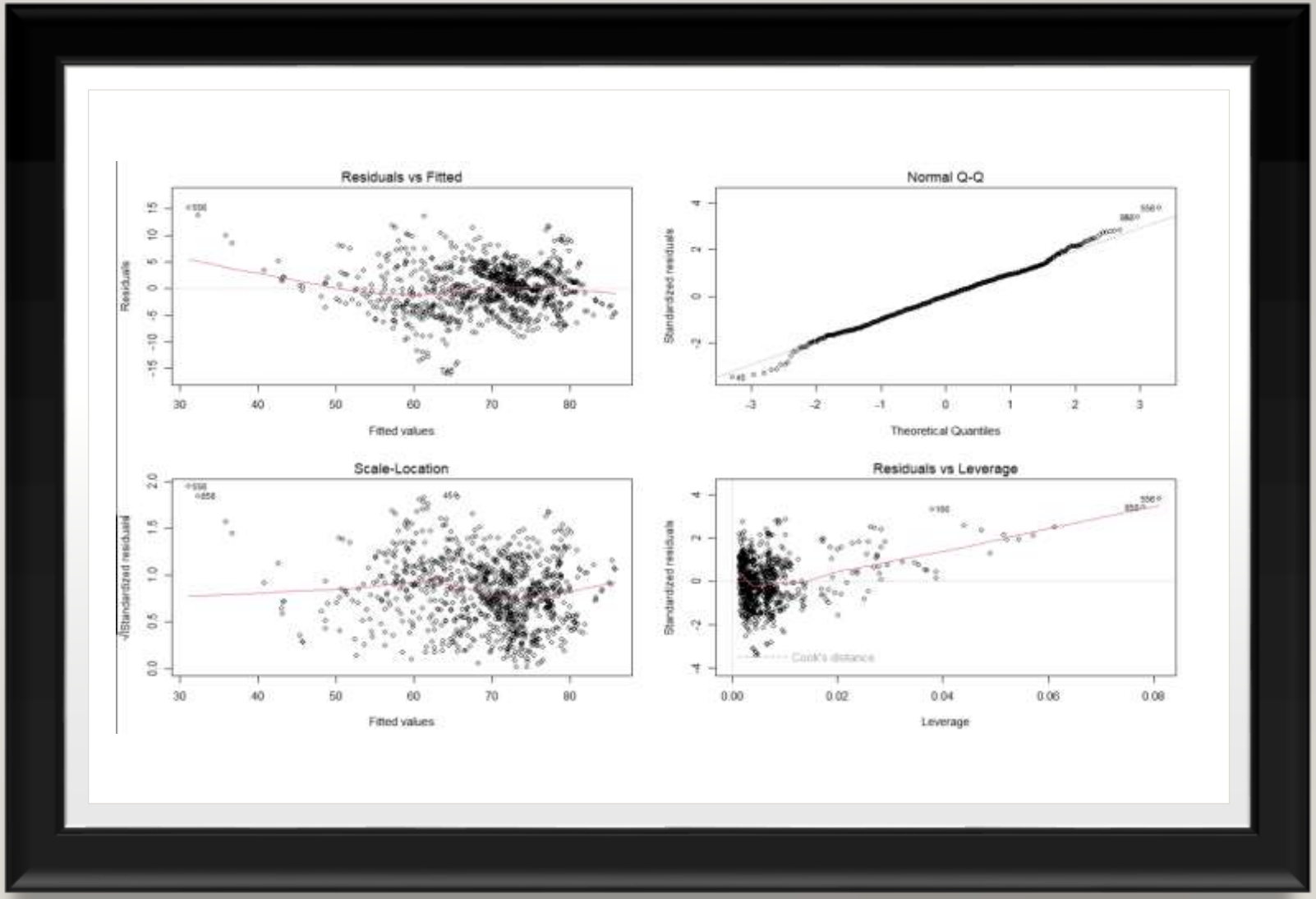
Omnibus:	25.271	Durbin-Watson:	1.851
Prob(Omnibus):	0.000	Jarque-Bera (JB):	43.311
Skew:	-0.187	Prob(JB):	3.94e-10
Kurtosis:	3.955	Cond. No.	1.90e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.9e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Life Expectancy = β0 + β1*(Adult Mortality) + β2*(Income composition of resources) + β3*Schooling + β4*HIV/AIDS + β5*BMI

Life Expectancy = 52.28 - 0.01*Adult Mortality +12.24*Income + 0.97*Schooling - 0.52*HIV/AIDS + 0.03*BMI

RESIDUAL DIAGNOSTICS



CONCLUSION

Adult Morality significantly affects the expected life expectancy.

Income Composition of Resources affects the Life expectancy

Yes, we know alchohol affects the health in a bad manner, but it is not our final predicted model, so we can say on an average its effect tends to go down.

Model doesn't mention about the country origin so it can be concluded that density of population has nothing to do with life expectancy

Better health leads to higher life expectancy.