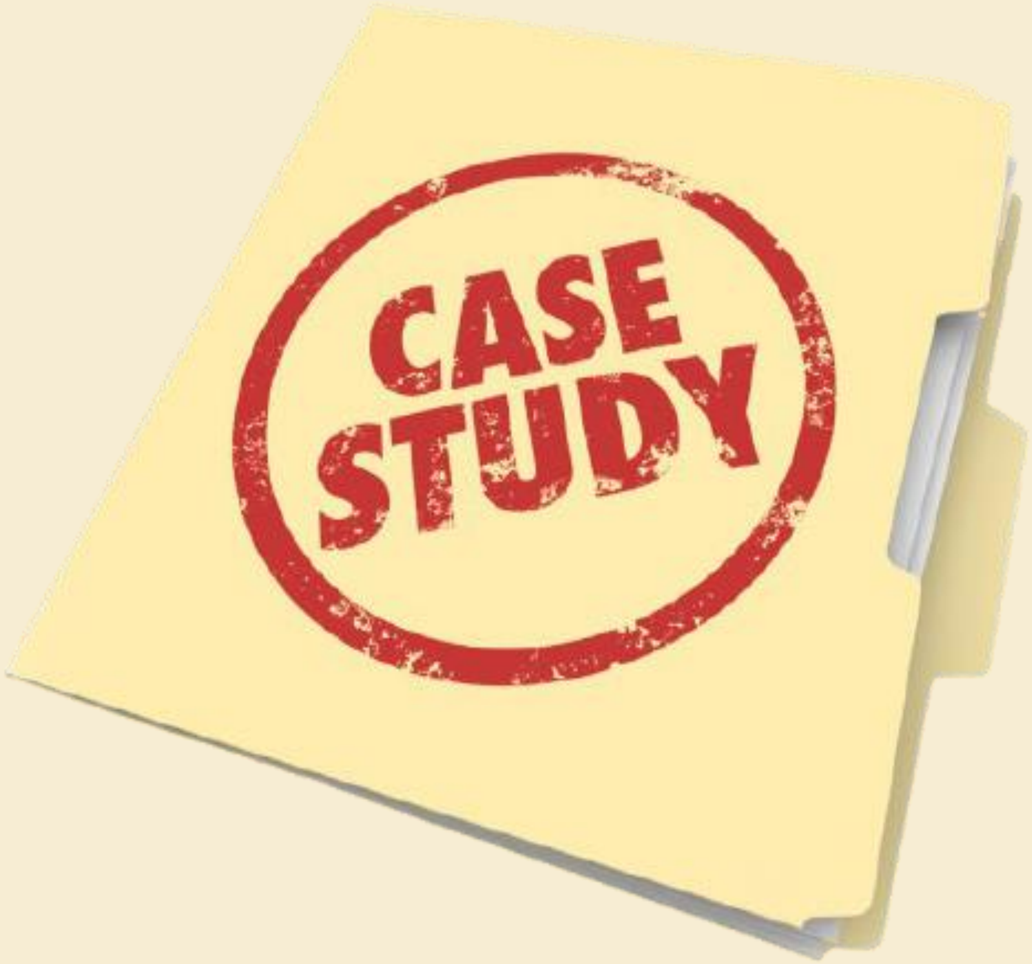
# LIFE EXPECTANCY

→ Research Project by
ALI AHMAD

# CONTENTS

- Case Study
- Objective
- Univariate Analysis
- Bivariate Analysis
- Model Building with Significant Variables
- Observations
- Business Recommendations

# CASE STUDY



- The dataset comes from 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-set related to life expectancy, health factors for 193 countries have been collected from the same WHO data repository website.
- In this case study we have considered data from year 2000-2015 for 193 countries for the analysis.
- It consists of 23 variables with 2938 rows.

# Different Variables

- Country
- Year
- Status
- Life_Expectancy
- Adult_Mortality
- Infant_Deaths
- Alcohol
- Percentage_Expenditure
- Hepatitis_B
- Measles
- BMI
- Under-five_Deaths
- Diphtheria
- Total_Expenditure
- HIV/AIDS
- GDP
- Per_Capita_GDP
- Population
- Thinness_1-19_Years
- Thinness_5-9_Years
- Income_Composition_of_Resources
- Schooling

# OBJECTIVE

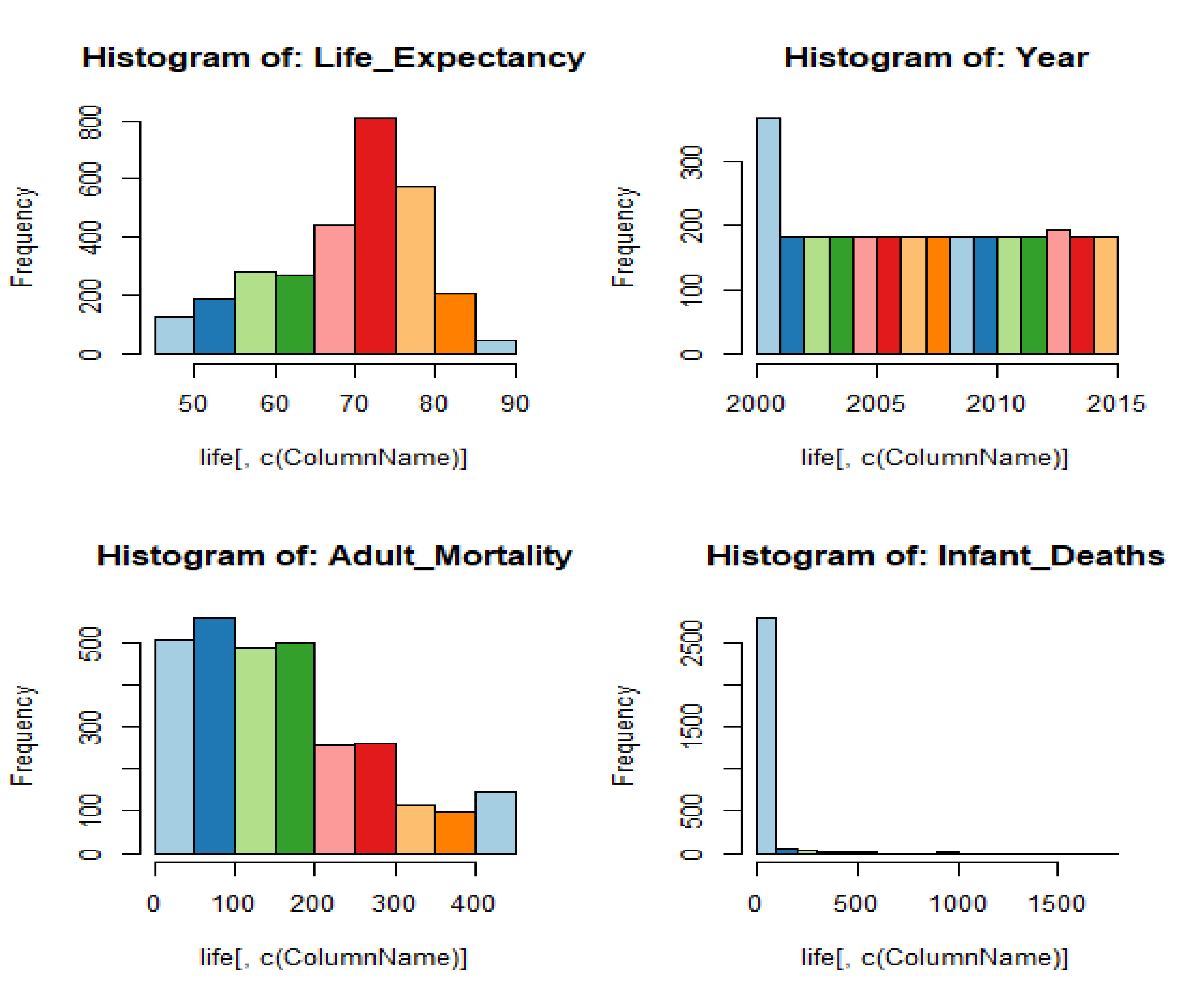
The purpose of this case study is to use Linear regression model to predict the life expectancy of the people for the next year. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well.

In this dataset we are provided average life expectancy of people of 193 Countries from year 2000-2015.

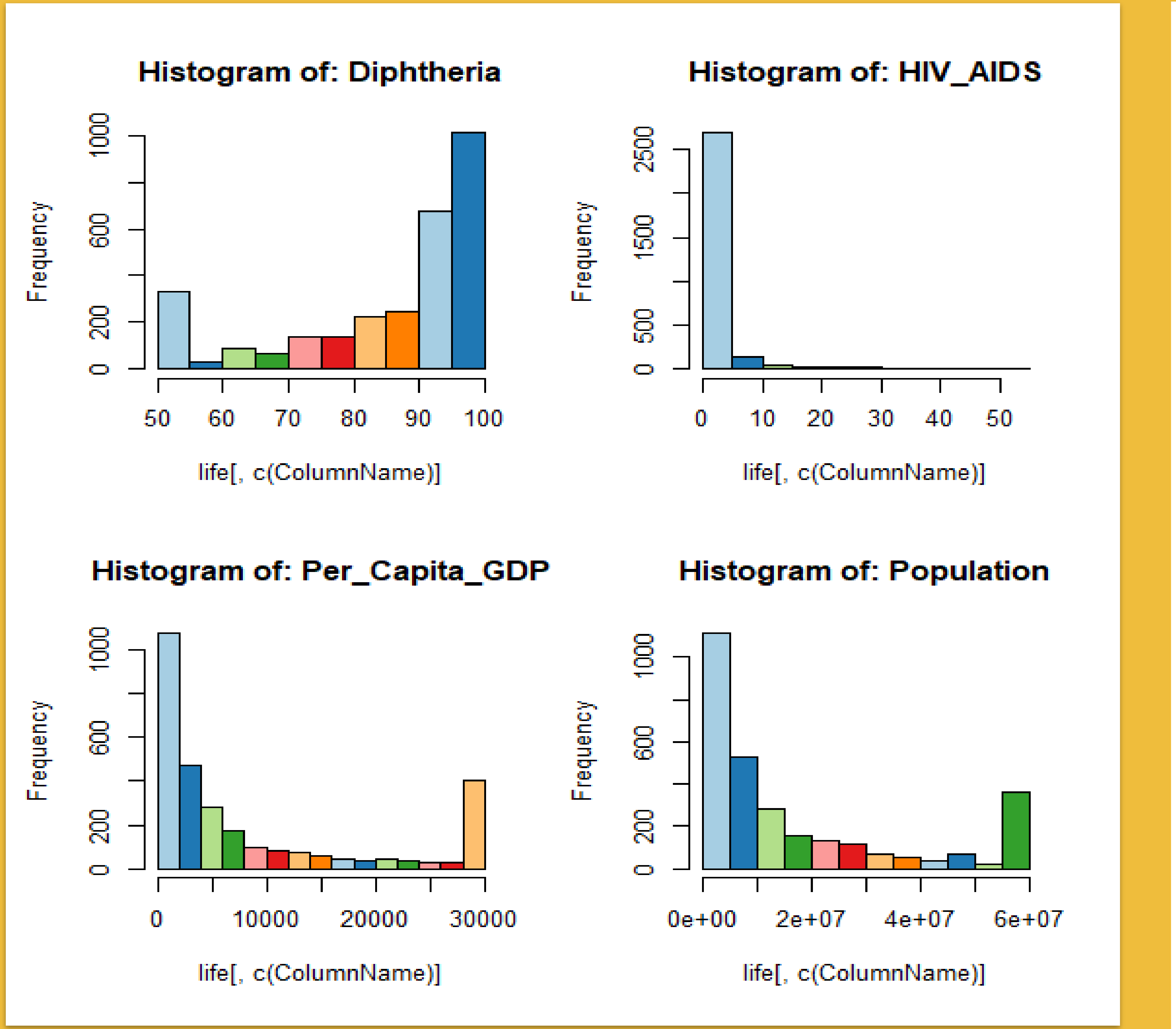
# UNIVARIATE ANALYSIS

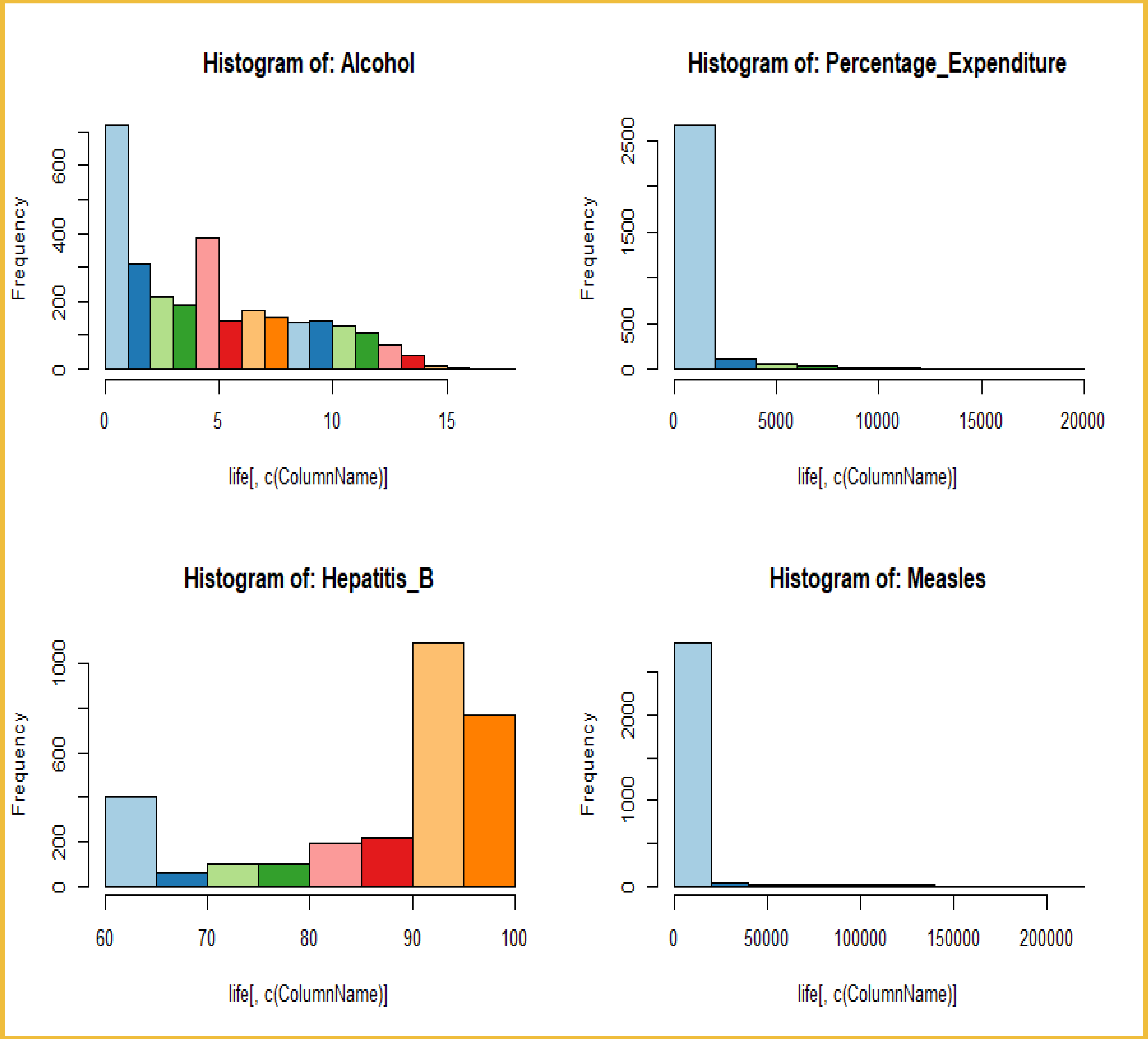
Using Histogram: Continuous column

- The target variable “Life Expectancy” is not distributed perfectly, it is a little left-skewed.
- For variable Infant_Deaths, number of deaths is decreased rapidly.



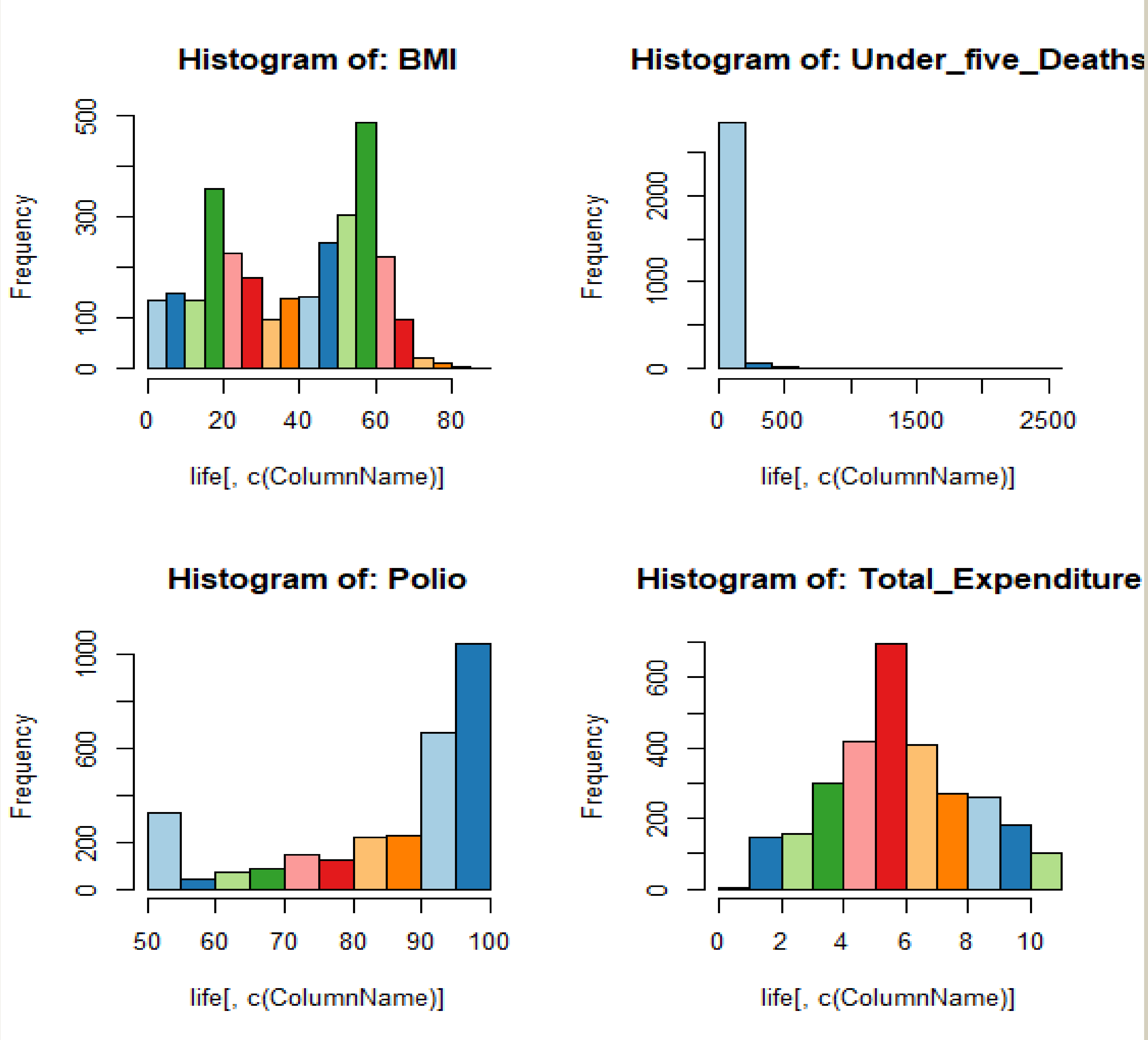
# UNIVARIATE ANALYSIS



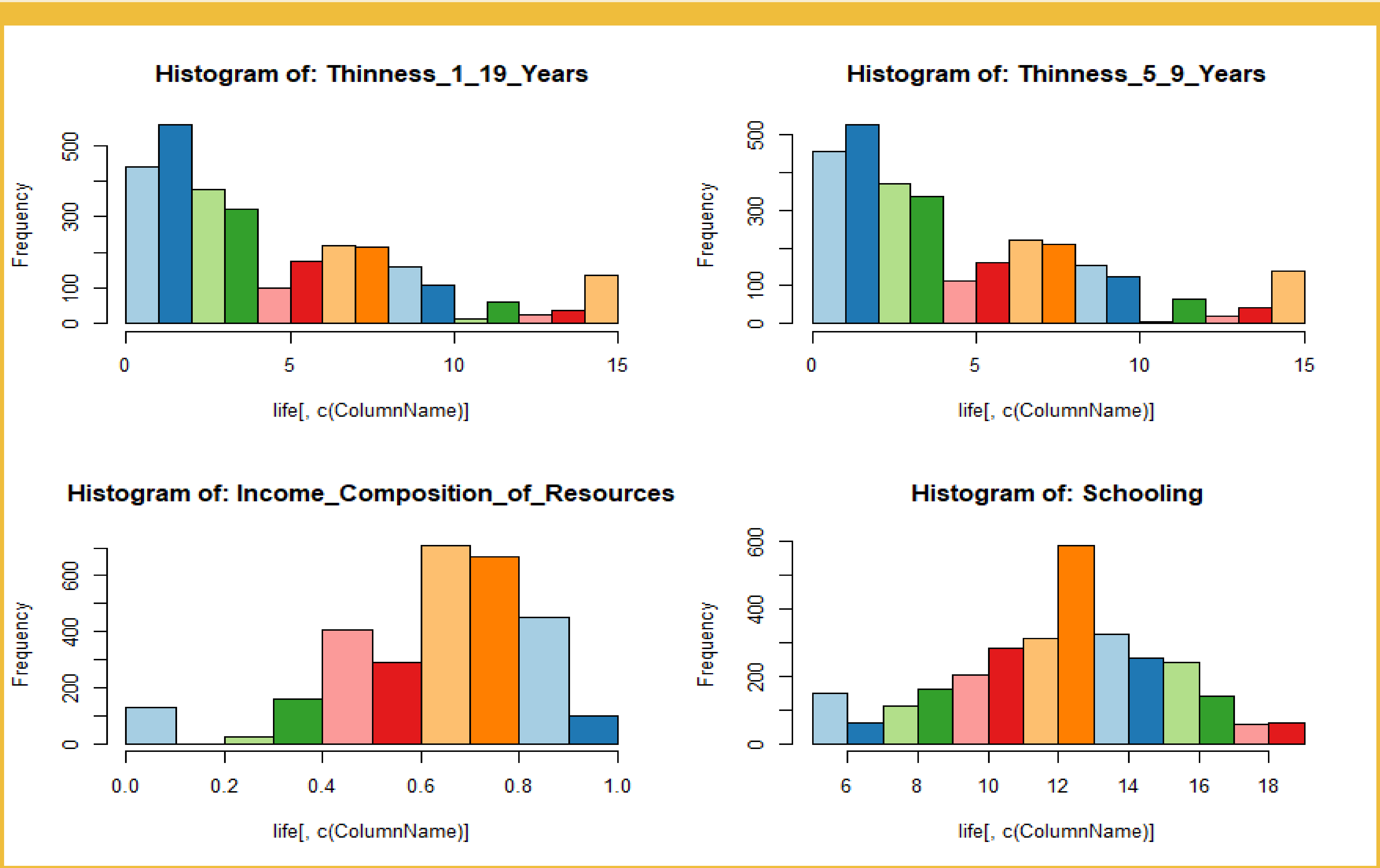


# UNIVARIATE ANALYSIS

- The variable Under_five_deaths is not normally distributed. It is highly right-skewed.
- The variable Polio is also not normally distributed. It is highly left-skewed.



# UNIVARIATE ANALYSIS



Histogram of: Thinness_1_19_Years

Histogram of: Thinness_5_9_Years

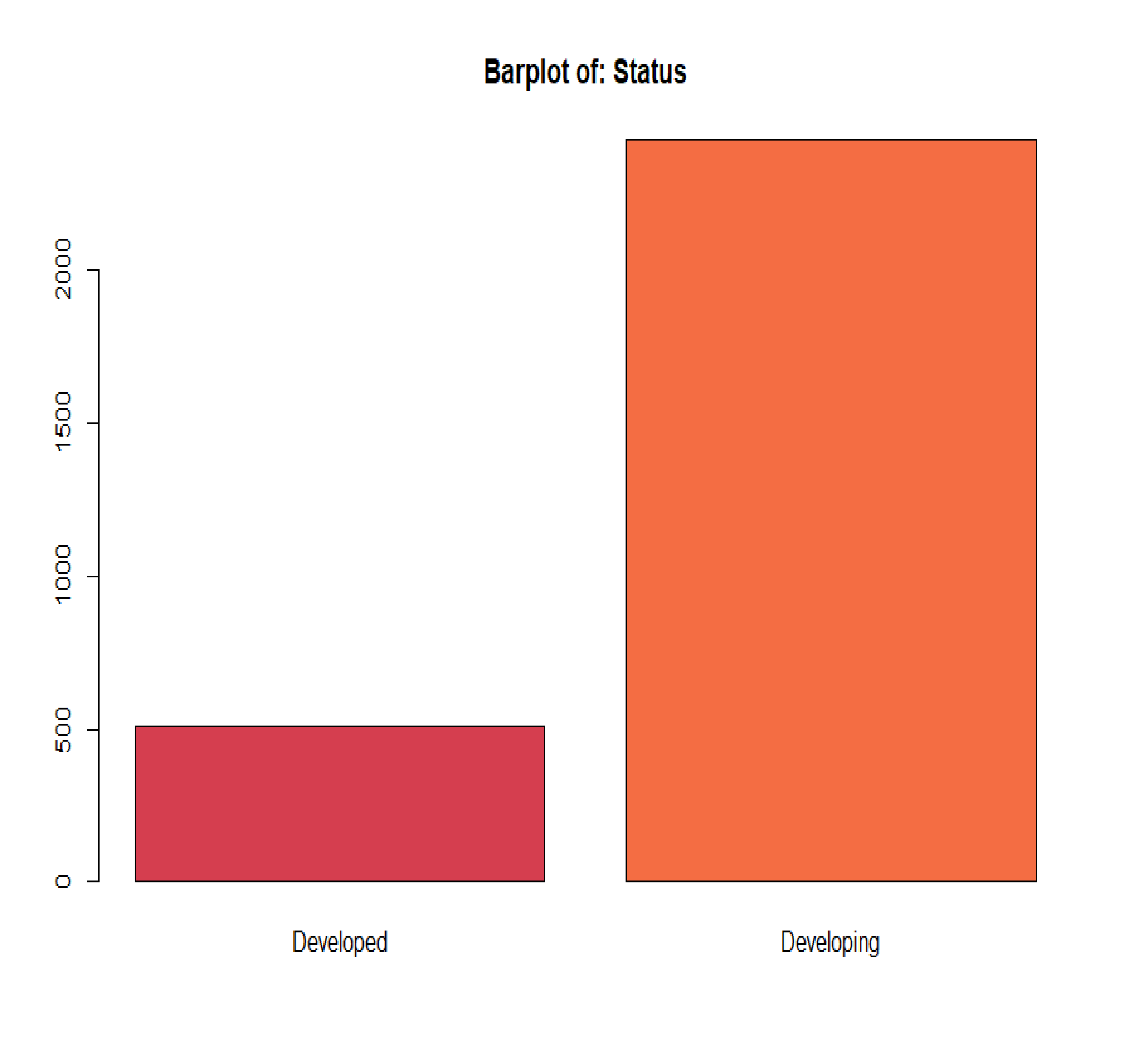
Histogram of: Income_Composition_of_Resources

Histogram of: Schooling

# UNIVARIATE ANALYSIS

Using barplot: Categorical Column

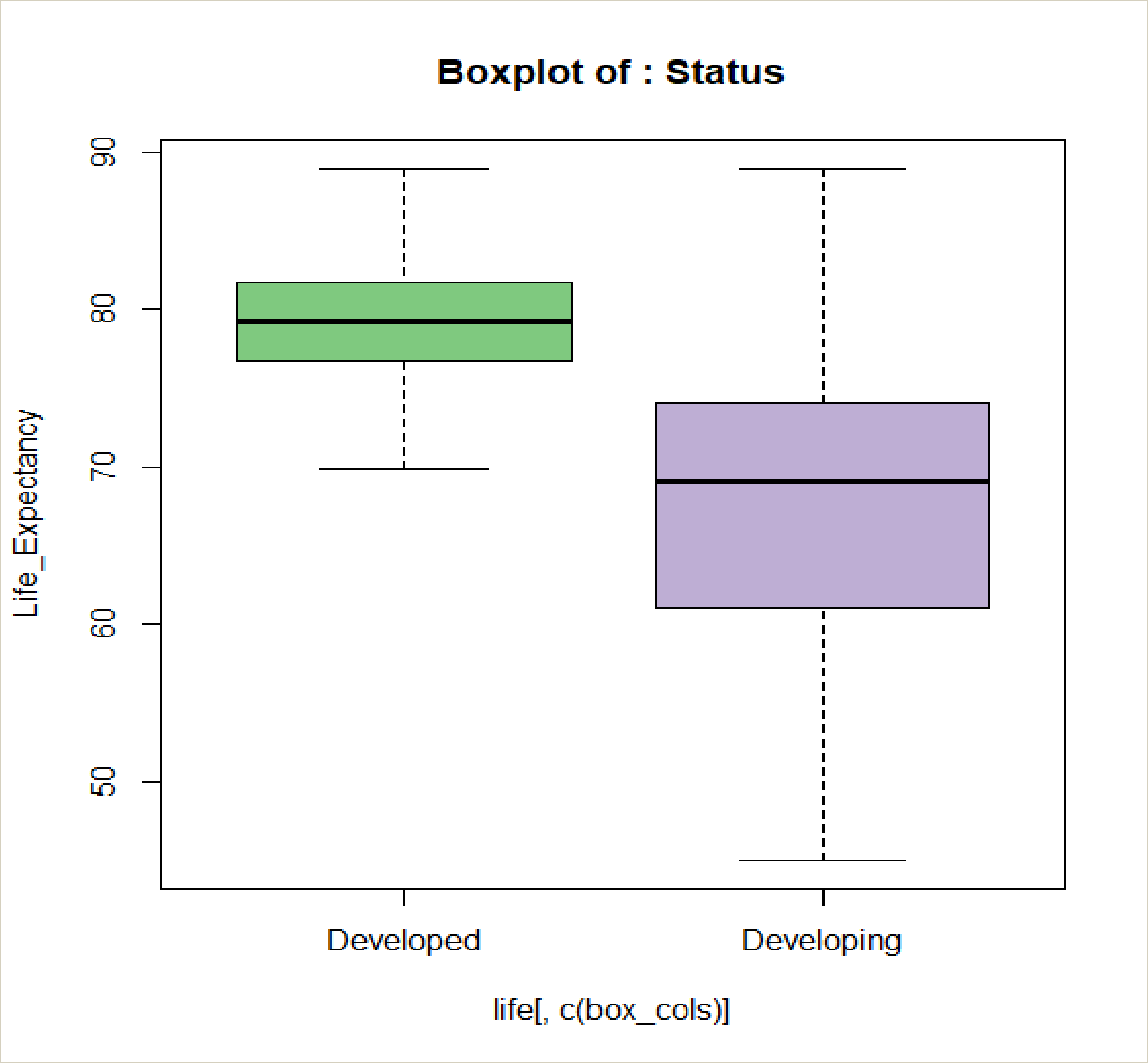
There is a decrease in the life expectancy value in case of developed countries whereas in case of Developing countries, the life expectancy value is gradually rising.



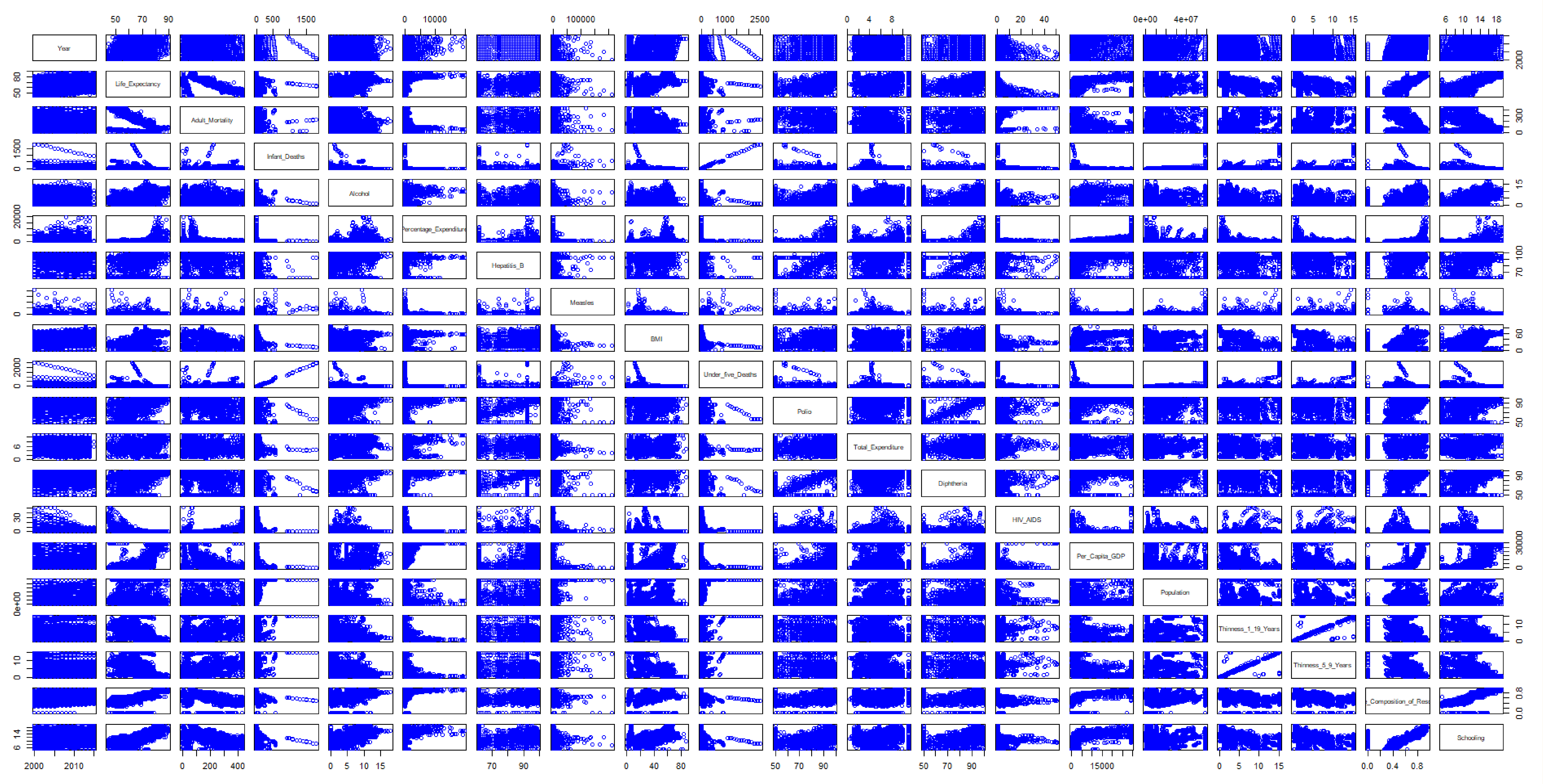
# BIVARIATE ANALYSIS

Using Boxplot: Categorical Column

- The average for developed countries is 79 while for the average for the developing countries is around 70.



# BIVARIATE ANALYSIS : Using Scatterplot



# ANOVA (Analysis of Variance)

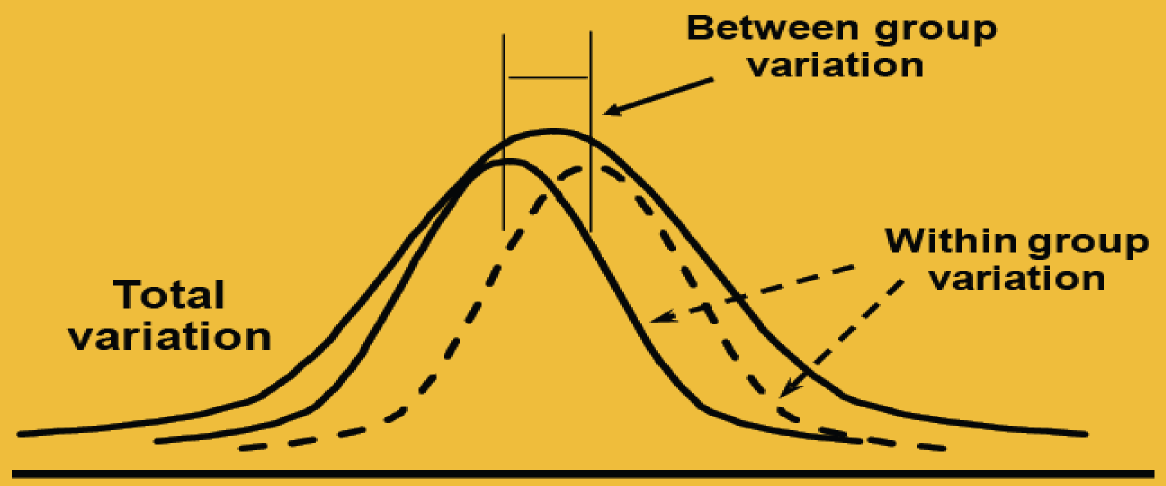
It gives the relationship between continuous and categorical columns whether they are statistically significant or not.

H0- Null hypothesis means the variables are not correlated:

1. If Small P-Value < 5% (0.05)  means the variables are correlated. Null hypothesis H0 is rejected
2. If Large P-Value > 5% (0.05) means the variables are not correlated and the Null hypothesis H0 is accepted

In this dataset, the variables (p-value < 0.05) that are found significant from ANOVA test:

❖ Status



# CORRELATION TEST

Correlation is a statistical technique that predicts whether and how strongly pairs of variables are related.

It gives the relationship between continuous and continuous columns whether they are statistically significant or not.

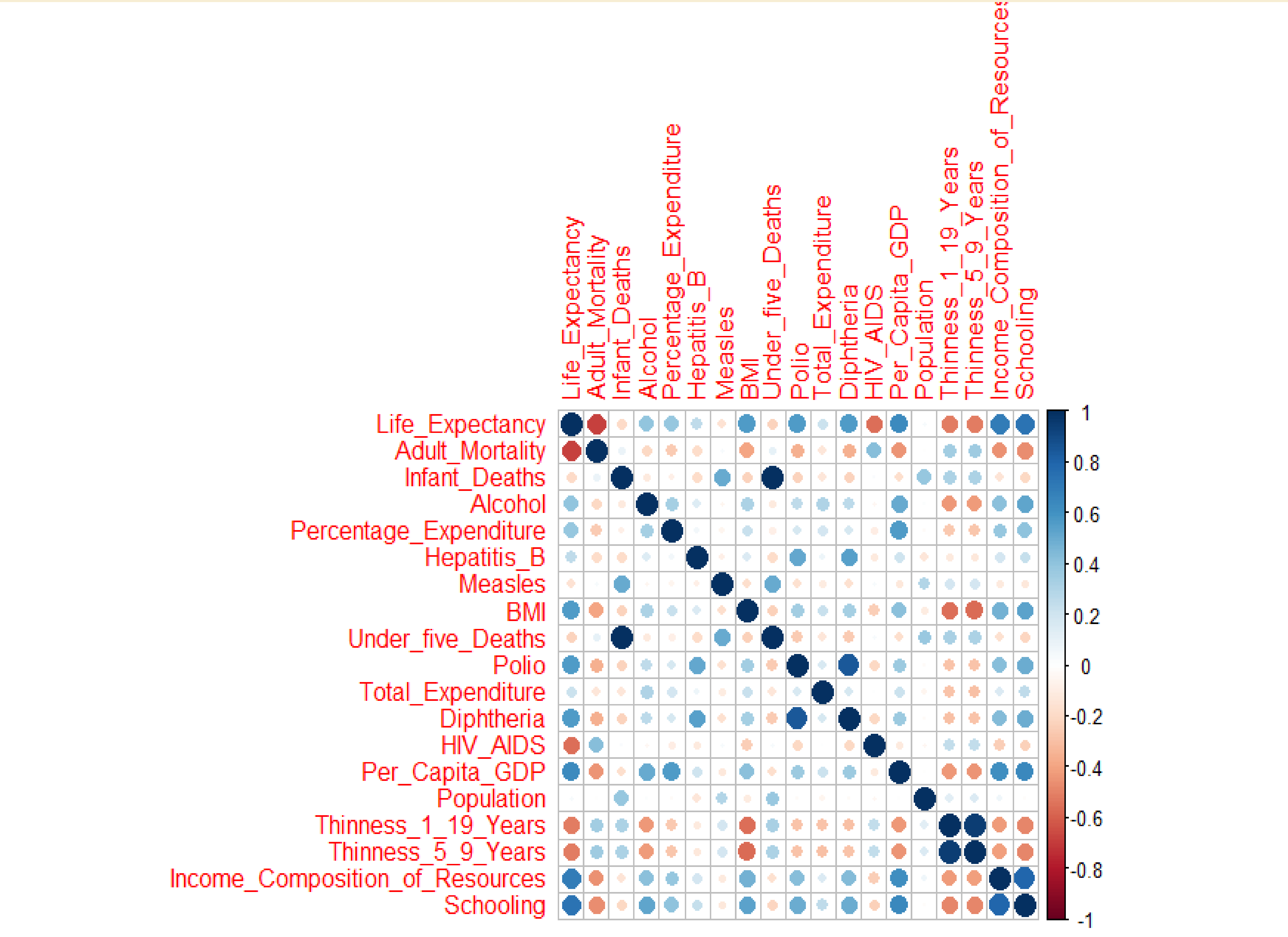
It ranges from -1.0 to +1.0.

In this dataset, the variables that are found significant from Correlation test:

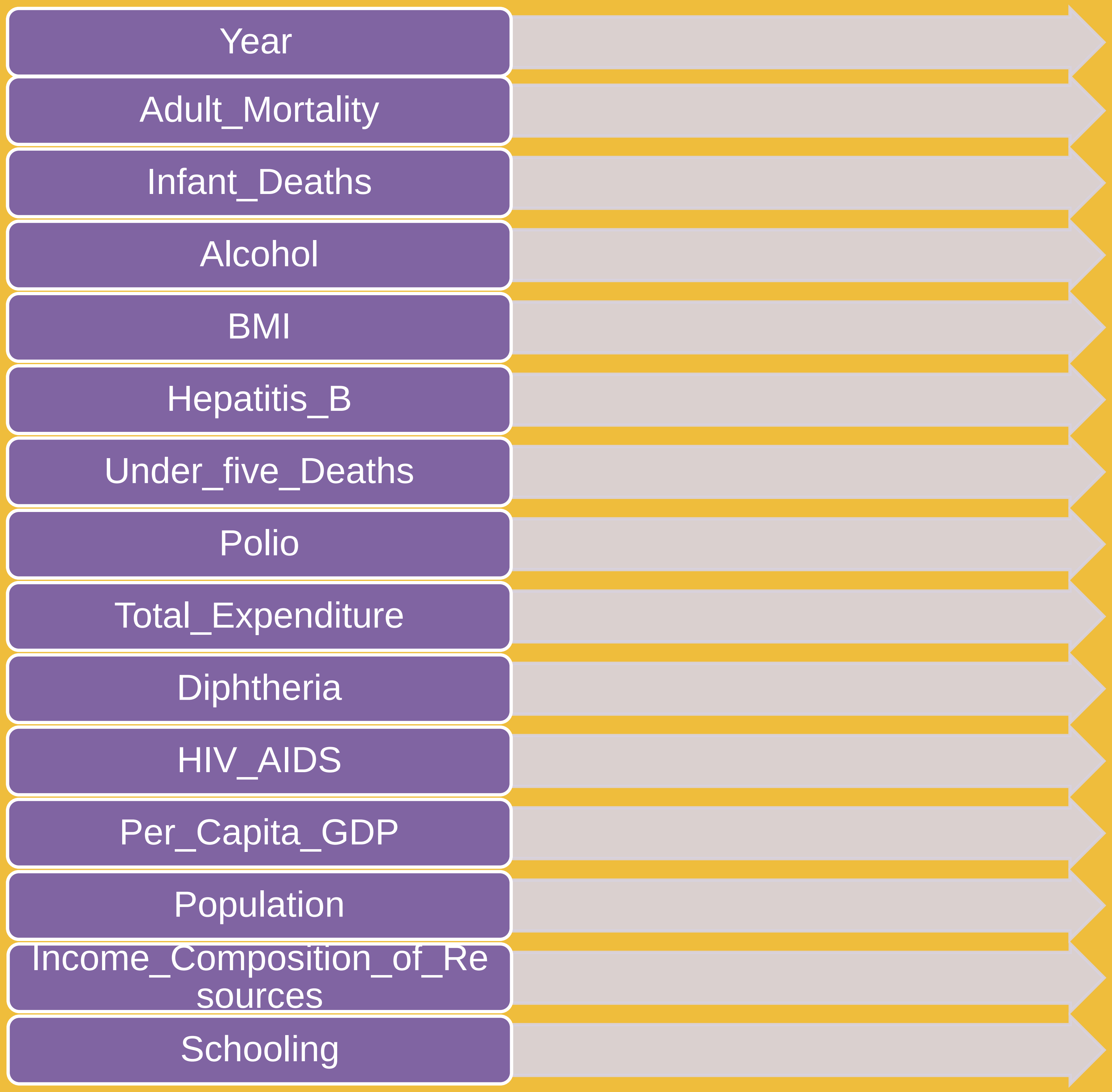
- Adult_Mortality
- Alcohol
- Percentage_Expenditure
- BMI
- Polio
- Diphtheria
- HIV_AIDS
- Per_Capita_GDP
- Thinness_1_19_Years
- Thinness_5_9_Years
- Income_Composition_of_Resources
- Schooling

The variables Infant_Deaths, Hepatitis_B, Measles, Under_five_Deaths, Total_Expenditure, Population were not correlated with the target variable Life Expectancy, so we reject these variables for further analysis.

# HEATMAP OF CORRELATED VARIABLES



# MODEL BUILDING WITH SIGNIFICANT VARIABLES



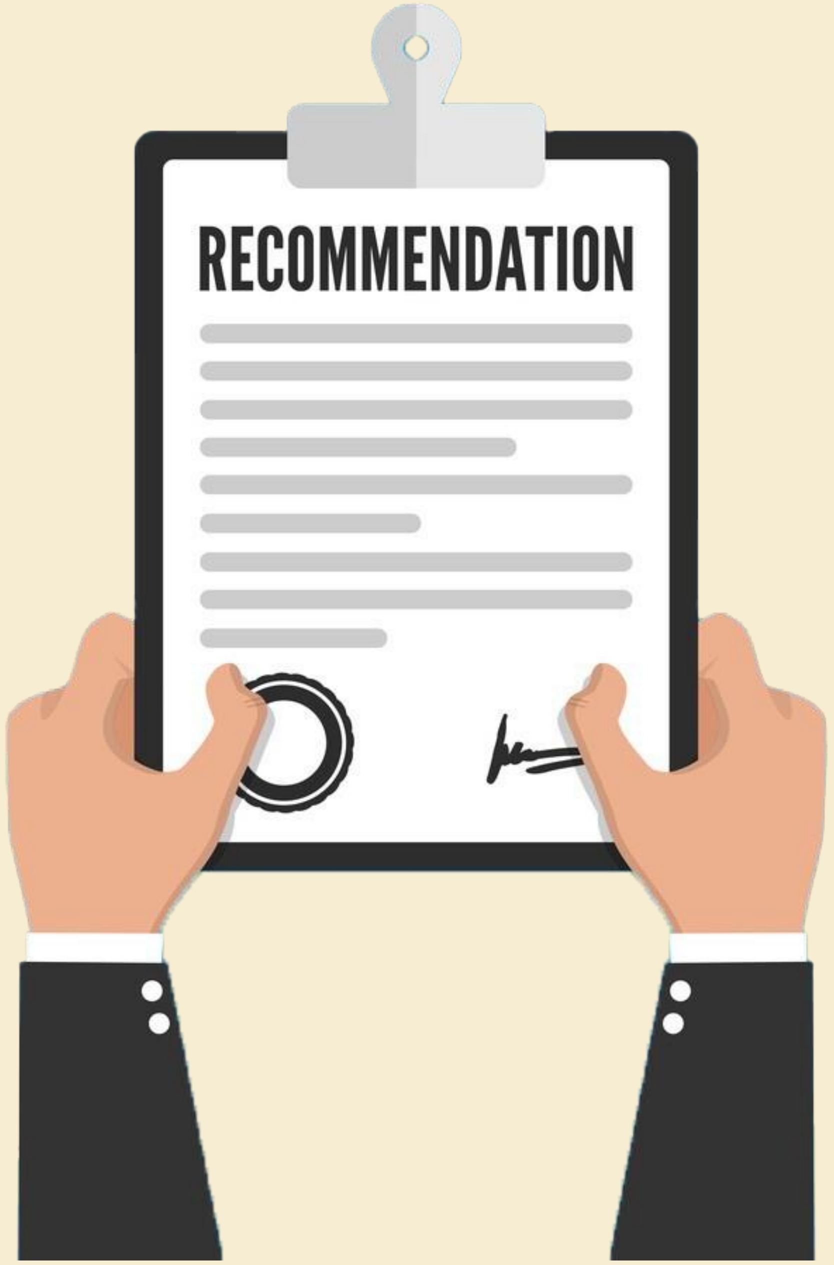
- Year
- Adult_Mortality
- Infant_Deaths
- Alcohol
- BMI
- Hepatitis_B
- Under_five_Deaths
- Polio
- Total_Expenditure
- Diphtheria
- HIV_AIDS
- Per_Capita_GDP
- Population
- Income_Composition_of_Resources
- Schooling

# OBSERVATIONS

- Multiple R-Squared of Linear Regression : 0.843
- Adjusted R-squared of Linear Regression : 0.8421
- Mean Absolute Percentage Error (MAPE) : 4.256897
- Mean Accuracy of Linear Regression : 95.74
- Median Absolute Percentage Error (MDAPE) : 2.935505
- Median Accuracy of Linear Regression : 97.06

# BUSINESS RECOMMENDATIONS

- The Developed countries should help developing countries in eradicating the diseases which are affecting the life of the people by providing vaccinations
- The government of developing countries should launch various schemes to motivate people to send their kids to schools
- Government should organize free healthcare camps to provide free vaccinations for the needy and poor people so that they don't have to spend their money and they also stay healthy to treat their families well.
- The government should increase the subsidy on liquor and increase healthcare and welfare camps to generate awareness among people, how bad drinking is and how it affects your body.
- WHO with the help of developed nations should help the government of developing countries in providing free food, education and organize healthcare camps.



# THANK YOU

ivy™

Professional School

ISO 9001 : 2000 Organisation