Life Expectancy Analysis Using Multiple Linear Regression

Data Synergists

Sai Raga Mounika, Darimireddy

Yogitha Polisetty

Venkata Ashwath Vennela

Anudeep Thummu

Summary

Objective:

The objective of this project is to analyze a dataset containing various health and demographic indicators from different nations and explore their relationships with life expectancy. The main goal is to identify key factors influencing life expectancy and gain insights into population health outcomes.

Methodology:

- **Data Collection:** We obtained the dataset from Kaggle, which includes various socio-economic, health and demographic factors.
- **Data Preprocessing:** We performed data cleaning and preprocessing steps, including handling missing values, renaming variables, and converting categorical variables into factors.
- **Exploratory Data Analysis (EDA):** We conducted EDA to understand the distribution of variables, identify trends, and explore relationships between variables.
- **Statistical Analysis:** We utilized statistical techniques such as correlation analysis and regression analysis to examine associations between health indicators and life expectancy.
- **Visualization:** We created visualizations such as scatterplots, histograms, and boxplots to illustrate key findings and insights.

Key Findings:

- **Factors Influencing Life Expectancy:** We found that several factors, including adult mortality rates, vaccination coverage, GDP, and schooling, are significantly associated with life expectancy.
- **Differences Between Developed and Developing Countries:** There are notable differences in the determinants of life expectancy between developed and developing countries. For example, access to healthcare and socio-economic factors play a more significant role in developed countries.
- **Trends Over Time:** We observed trends in population health outcomes over time, with improvements in life expectancy in many countries, but disparities persist across regions.

1. Introduction

1.1 Background Information:

The dataset used in this analysis comprises various health and demographic indicators collected from diverse nations over multiple years. These indicators include factors such as life expectancy, adult mortality rates, infant deaths, alcohol consumption, vaccination coverage, GDP, and schooling. The dataset aims to provide insights into the factors influencing population health and well-being across different countries and regions.

1.2 Problem Statement:

The primary objective of this analysis is to develop a multiple linear regression model to predict the life expectancy of a country using key economic, social, and health-related factors and explore the relationships between them. Life expectancy is a key measure of overall population health and reflects the average number of years a person can expect to live. By examining the factors associated with variations in life expectancy, we aim to identify potential determinants and drivers of population health outcomes.

1.3 Aim:

In our case study, our aim is to develop a multiple linear regression model to predict the life expectancy of a country using key economic, social, and health-related factors

1.3.1 Objectives:

Specifically, we seek to answer the following questions:

- What are the key factors influencing life expectancy across different countries and regions?
- Are there significant associations between specific health indicators (e.g., vaccination coverage, adult mortality rates) and life expectancy?
- How do socio-economic factors (GDP, Schooling) correlate with life expectancy?
- Are there differences in the determinants of life expectancy between developed and developing countries?
- Can we identify any trends or patterns in population health outcomes over time?

2. Data Description

2.1 Overview:

The dataset used in this analysis contains a comprehensive collection of health and demographic indicators from diverse nations around the world. It includes information on various factors that are known to influence population health outcomes, such as life expectancy, mortality rates, vaccination coverage, socio-economic indicators, and more. The dataset covers multiple years and countries, providing a rich source of information for exploring global health trends.

2.2 Variables:

Country	: Country
Year	: Year
Status	: The country’s development status, whether "Developed" or "Developing.”
Life Expectancy	: Average lifespan in age(Years).
Adult Mortality	: Number of deaths per 1,000 population.
Infant Deaths	: Number of infant deaths per 1,000 live births.
Alcohol	: Average alcohol consumption in liters per capita.
Percentage Expenditure	: Health expenditure as a percentage of a country's GDP.
Hepatitis B	: Measure immunization coverage for Hepatitis B.
Measles	: Number of reported cases per 1,000 population.
BMI	: The average Body Mass Index.
Under-Five Deaths	: Number of deaths under age five per 1,000 live births.
Polio	: Immunization coverage for Polio.
Total Expenditure	: Total health expenditure as a percentage of GDP.
Diphtheria	: Immunization coverage for Diphtheria.
HIV/AIDS	: Prevalence of HIV/AIDS as a percentage of the population.
GDP	: Gross Domestic Product data.

Population	: Witness the ebb and flow of a nation's populace.
Thinness 1-19 Years	: Prevalence of thinness among adolescents aged 1-19.
Thinness 5-9 Years	: Thinness among children aged 5-9.
ICOR	: Composite index reflecting income distribution and resource access.
Schooling	: Average years of schooling.

2.3 Observations:

- The dataset contains 1649 rows and 22 columns(variables) of which 20 are quantitative and 2 are qualitative.
- Each row represents a specific country and year combination, providing a snapshot of health and demographic indicators for that particular time period.
- The dataset spans multiple years, allowing for the analysis of trends over time.

2.4 Source:

The dataset was obtained from Kaggle, which is a popular data repository. The data is publicly available and has been used in numerous research studies and analyses.

2.5 Data Preprocessing:

Before conducting the analysis, we performed several data preprocessing steps, including:

- **Handling missing values:** Examined the dataset for missing values and implemented appropriate strategies for handling them, such as imputation or removal.
- **Renaming variables:** Standardized variable names to ensure consistency and clarity throughout the analysis.
- **Integration of 'Continent' Information:** Incorporated continent information into the dataset by using an external package which performs the country-continent mapping.
- **Reordering the columns:** The columns in the dataset are rearranged for easier glancing.
- **Converting categorical variables:** Converted categorical variables, such as 'Status', 'Country' and ‘Continent’, into factors for easier analysis.

- **Removing variables:**
    - Remove the variables that do not provide additional information to predict life expectancy. The categorical variable country has too many unique levels which results in high cardinality. Each country's economic, social, and health-related data have been included as separate observations in other variables. Therefore, the country variable does not provide additional information beneficial to the study. Thus, remove it.
    - The numerical variable 'Year' is a time series data. Since our study focuses on time-independent economic, social, and health-related predictors of life expectancy,, the Year variable does not provide a benefit to the study. First, by removing the "Country" and "Year" variables from the dataset, unnecessary data is removed, making the dataset more streamlined for analysis.

3. Exploratory data Analysis

The exploratory data analysis (EDA) step is crucial in all data analysis projects. Its purpose is to get a better understanding of the dataset, find patterns, connections, and potential issues that lead us to build ideas for future quantitative study. In the case study on life expectancy, we will utilize EDA to better comprehend the dataset and highlight key findings.

We employ the following components to conduct exploratory data analysis in order to achieve our goal of creating a multiple linear regression model to predict life expectancy.

1. Summary statistics are computed for key variables in the dataset, including measures such as mean, median, standard deviation, minimum, and maximum values for numerical variables. Frequency distributions were also generated for categorical variables.
2. Visualizations such as histograms, box plots, and bar plots were utilized to explore the distributions of variables and identify any outliers or unusual patterns.
3. Scatterplots were employed to visualize relationships between numerical variables, providing insights into potential correlations.

3.1 Summary Statistics:

Summary statistics were computed for key variables in the dataset, including measures such as mean, median, standard deviation, minimum, and maximum values for numerical variables. Frequency distributions were also generated for categorical

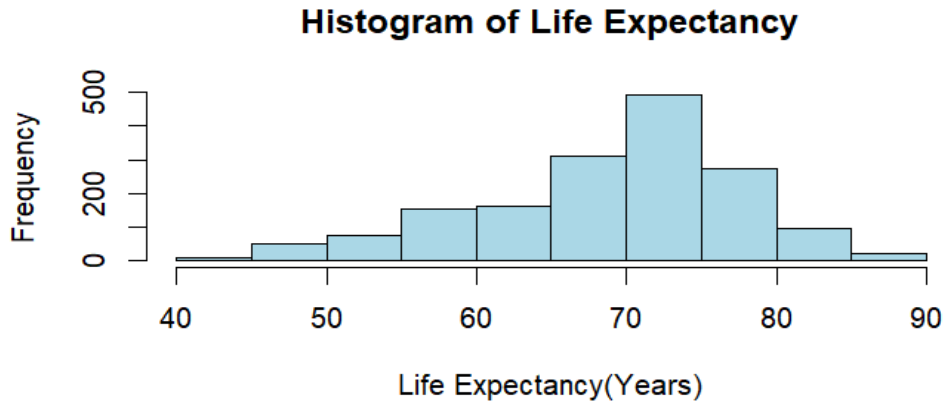
variables.

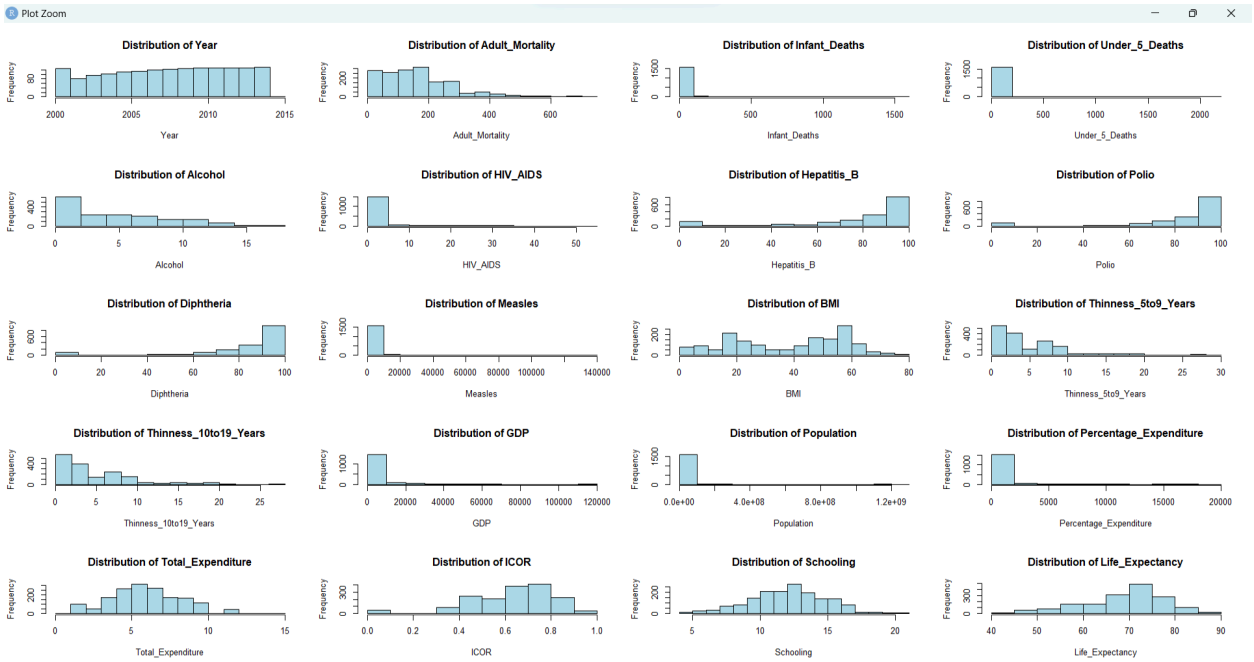
```
> # five-number summary of the given dataset
> summary(Cleaned_Health_Data)
       Country          Year           Status     Life_Expectancy Adult_Mortality Infant_Deaths       Alcohol       Percentage_Expenditure
 Afghanistan:  16   Min.   :2000   Developed : 242   Min.   :44.0    Min.   :  1.0   Min.   :   0.00   Min.   : 0.010   Min.   :    0.00
 Albania    :  16   1st Qu.:2005   Developing:1407   1st Qu.:64.4    1st Qu.: 77.0   1st Qu.:   1.00   1st Qu.: 0.810   1st Qu.:   37.44
 Armenia    :  15   Median :2008                     Median :71.7    Median :148.0   Median :   3.00   Median : 3.790   Median :  145.10
 Austria    :  15   Mean   :2008                     Mean   :69.3    Mean   :168.2   Mean   :  32.55   Mean   : 4.533   Mean   :  698.97
 Belarus    :  15   3rd Qu.:2011                     3rd Qu.:75.0    3rd Qu.:227.0   3rd Qu.:  22.00   3rd Qu.: 7.340   3rd Qu.:  509.39
 Belgium    :  15   Max.   :2015                     Max.   :89.0    Max.   :723.0   Max.   :1600.00   Max.   :17.870   Max.   :18961.35
 (Other)    :1557
  Hepatitis_B       Measles             BMI         Under_5_Deaths        Polio       Total_Expenditure   Diphtheria       HIV_AIDS
 Min.   : 2.00   Min.   :     0   Min.   : 2.00   Min.   :   0.00   Min.   : 3.00   Min.   : 0.740   Min.   : 2.00   Min.   : 0.100
 1st Qu.:74.00   1st Qu.:     0   1st Qu.:19.50   1st Qu.:   1.00   1st Qu.:81.00   1st Qu.: 4.410   1st Qu.:82.00   1st Qu.: 0.100
 Median :89.00   Median :    15   Median :43.70   Median :   4.00   Median :93.00   Median : 5.840   Median :92.00   Median : 0.100
 Mean   :79.22   Mean   :  2224   Mean   :38.13   Mean   :  44.22   Mean   :83.56   Mean   : 5.956   Mean   :84.16   Mean   : 1.984
 3rd Qu.:96.00   3rd Qu.:   373   3rd Qu.:55.80   3rd Qu.:  29.00   3rd Qu.:97.00   3rd Qu.: 7.470   3rd Qu.:97.00   3rd Qu.: 0.700
 Max.   :99.00   Max.   :131441   Max.   :77.10   Max.   :2100.00   Max.   :99.00   Max.   :14.390   Max.   :99.00   Max.   :50.600

      GDP              Population        Thinness_10to19_Years Thinness_5to9_Years      ICOR          Schooling       Continent
 Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.100        Min.   : 0.100      Min.   :0.0000   Min.   : 4.20   Africa  :473
 1st Qu.:   462.15   1st Qu.:1.919e+05   1st Qu.: 1.600        1st Qu.: 1.700      1st Qu.:0.5090   1st Qu.:10.30   Americas:298
 Median :  1592.57   Median :1.420e+06   Median : 3.000        Median : 3.200      Median :0.6730   Median :12.30   Asia    :414
 Mean   :  5566.03   Mean   :1.465e+07   Mean   : 4.851        Mean   : 4.908      Mean   :0.6316   Mean   :12.12   Europe  :345
 3rd Qu.:  4718.51   3rd Qu.:7.659e+06   3rd Qu.: 7.100        3rd Qu.: 7.100      3rd Qu.:0.7510   3rd Qu.:14.00   Oceania :119
 Max.   :119172.74   Max.   :1.294e+09   Max.   :27.200        Max.   :28.200      Max.   :0.9360   Max.   :20.70
```

3.2 Visualization of Key Variables:

Visualizations such as histograms, box plots, and bar plots were utilized to explore the distributions of variables and identify any outliers or unusual patterns. Scatterplots were employed to visualize relationships between numerical variables, providing insights into potential correlations.

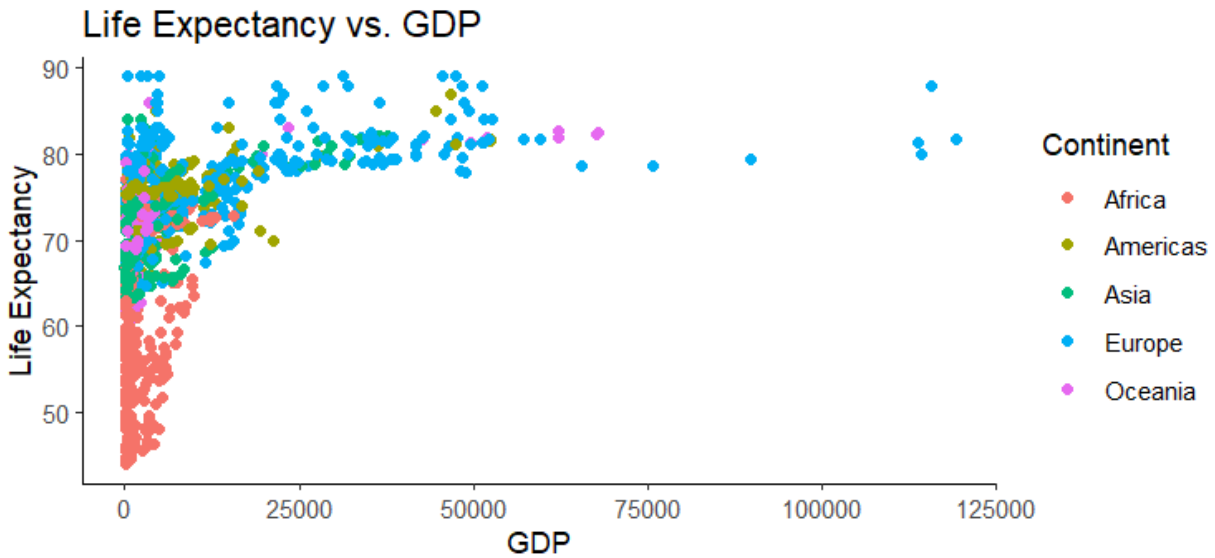
- Histogram:
The histogram of the target variable ‘Life Expectancy’ is below:

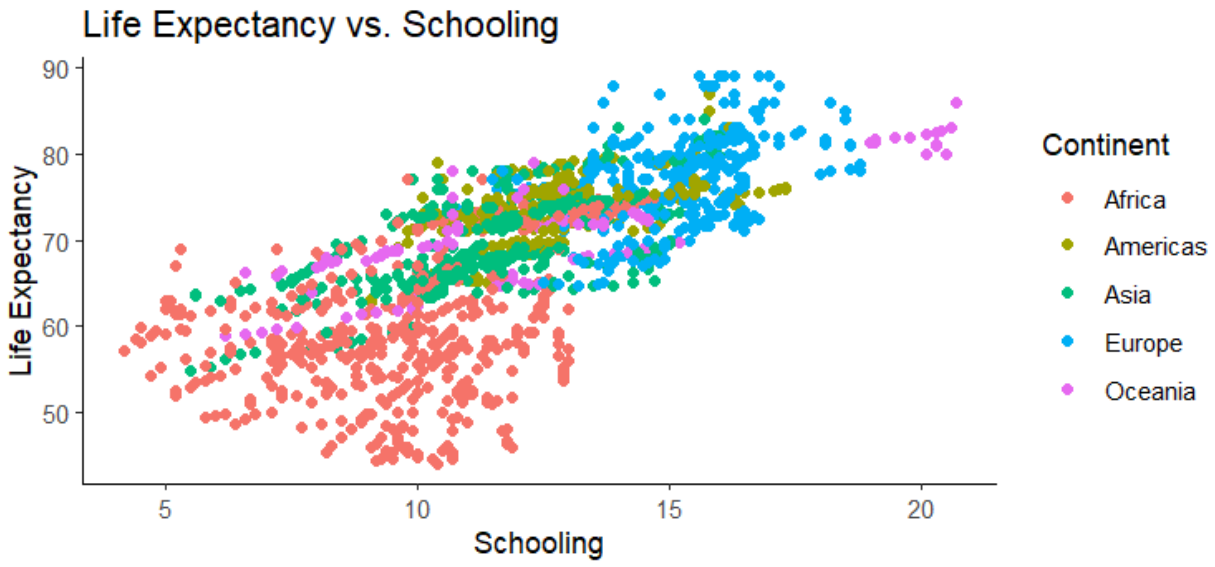


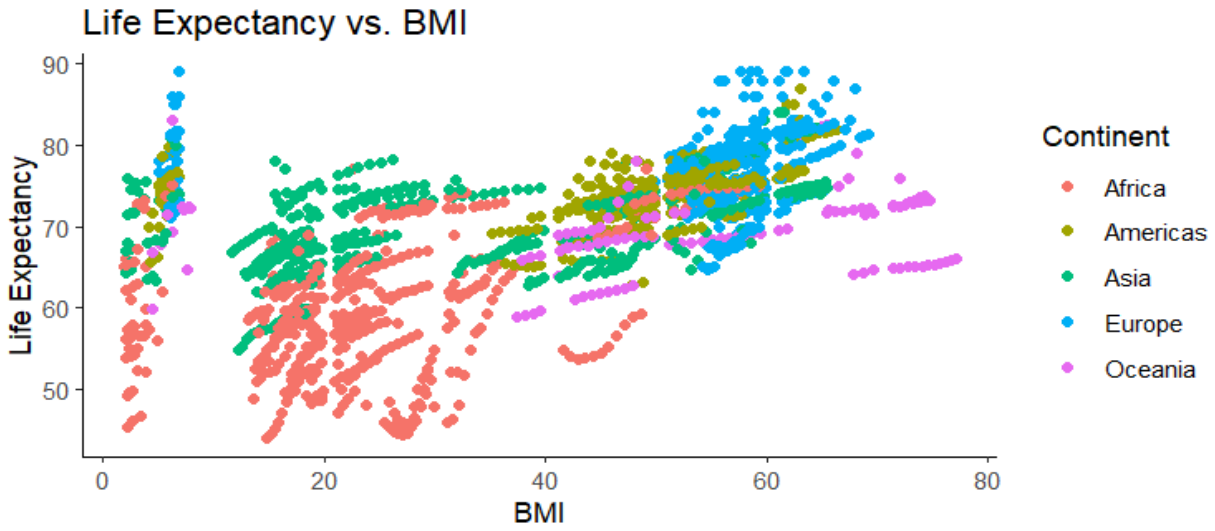


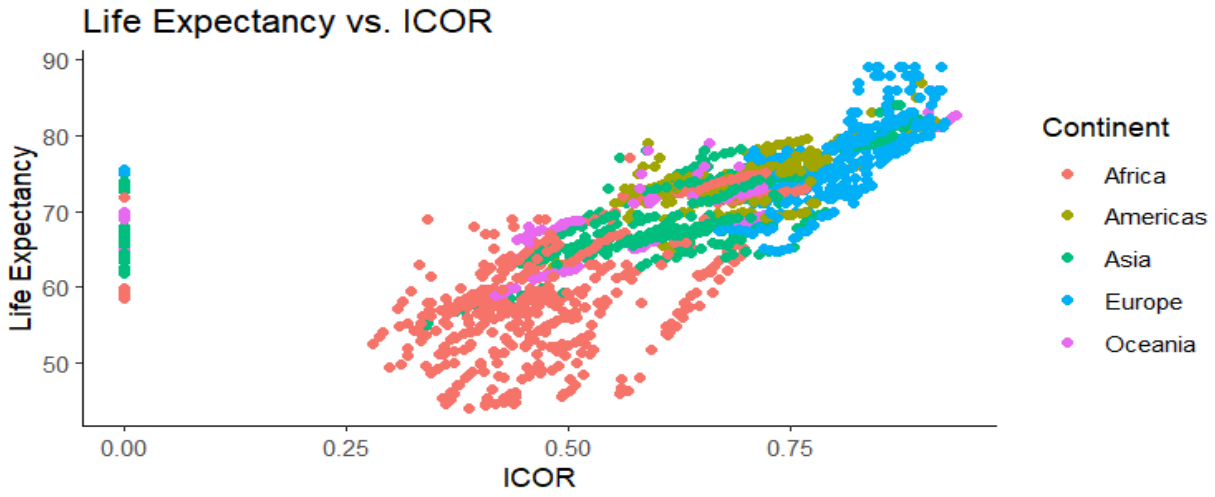
- Scatterplots:

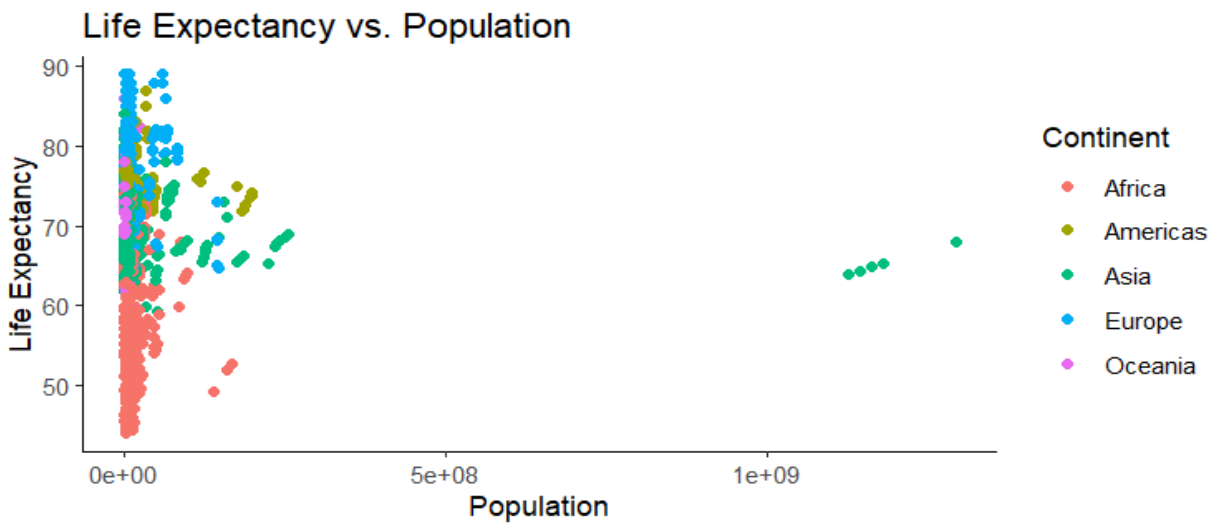
Scatter plots are used to visualize linear relationships. Scatter plots were used to depict the correlations between the dependent variable life expectancy and the other numerical variables. The visualization revealed that several factors might have linear correlations with Life Expectancy, supporting our objective of creating a multiple linear regression model to predict life expectancy

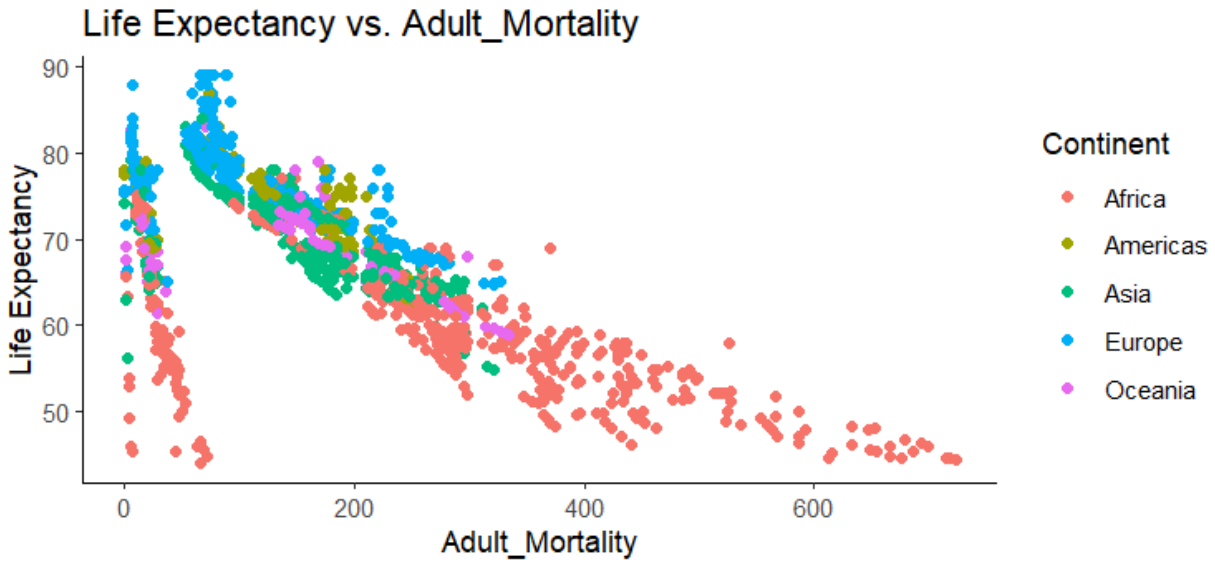


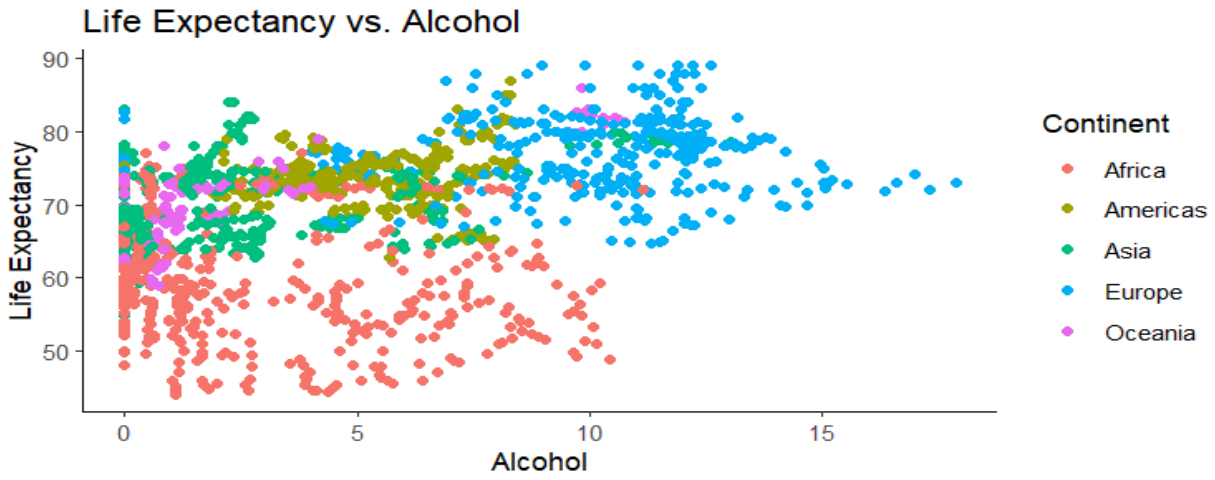

Life Expectancy vs. Schooling
Life Expectancy
90
80
70
60
50
5
10
15
20
Schooling
Continent
Africa
Americas
Asia
Europe
Oceania

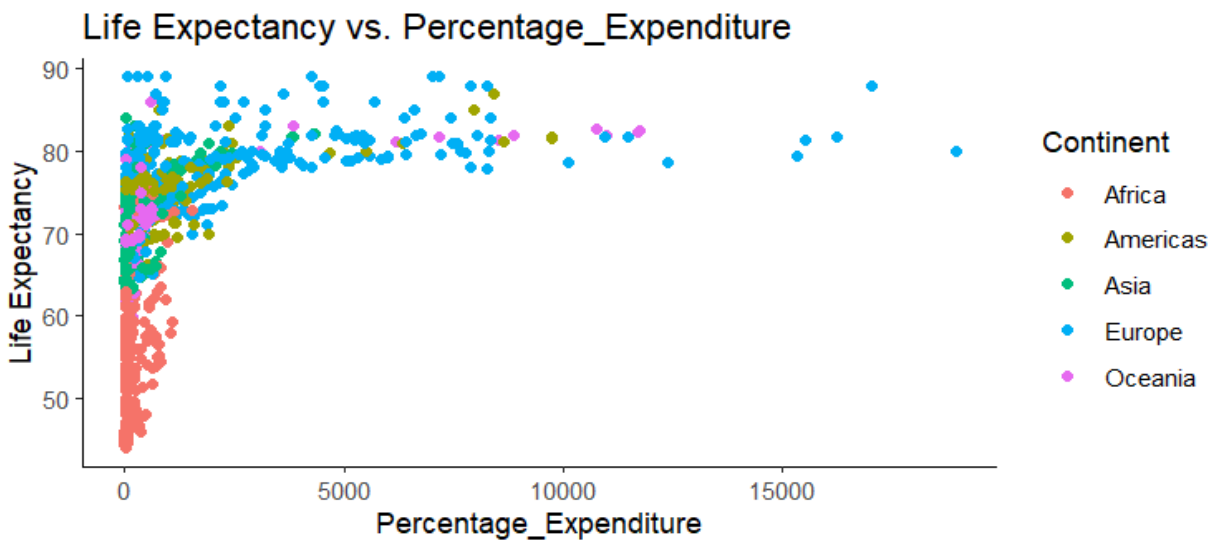

Life Expectancy vs. BMI
Life Expectancy
90
80
70
60
50
0
20
40
60
80
BMI
Continent
Africa
Americas
Asia
Europe
Oceania

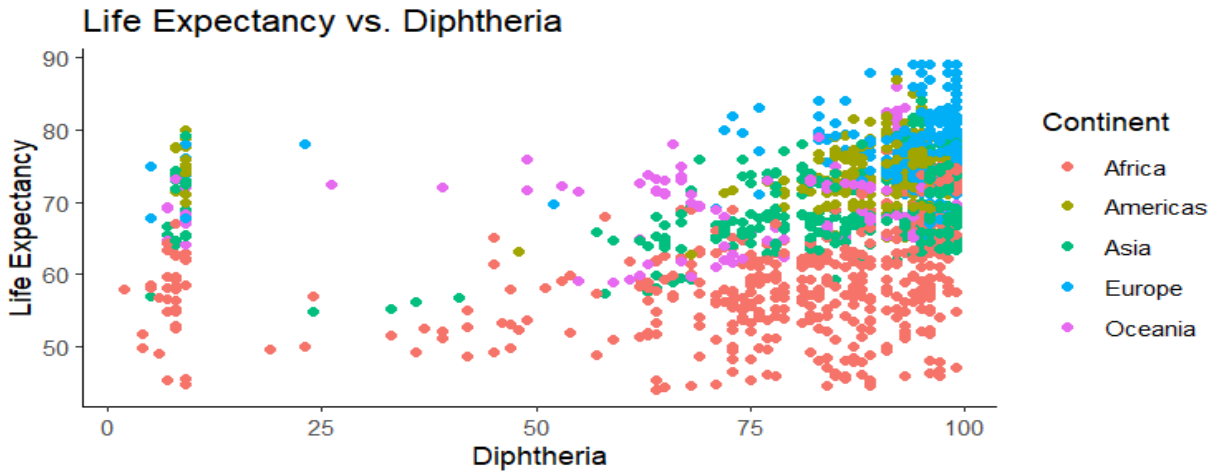

Life Expectancy vs. ICOR
Life Expectancy
90
80
70
60
50
0.00
0.25
0.50
0.75
ICOR
Continent
Africa
Americas
Asia
Europe
Oceania

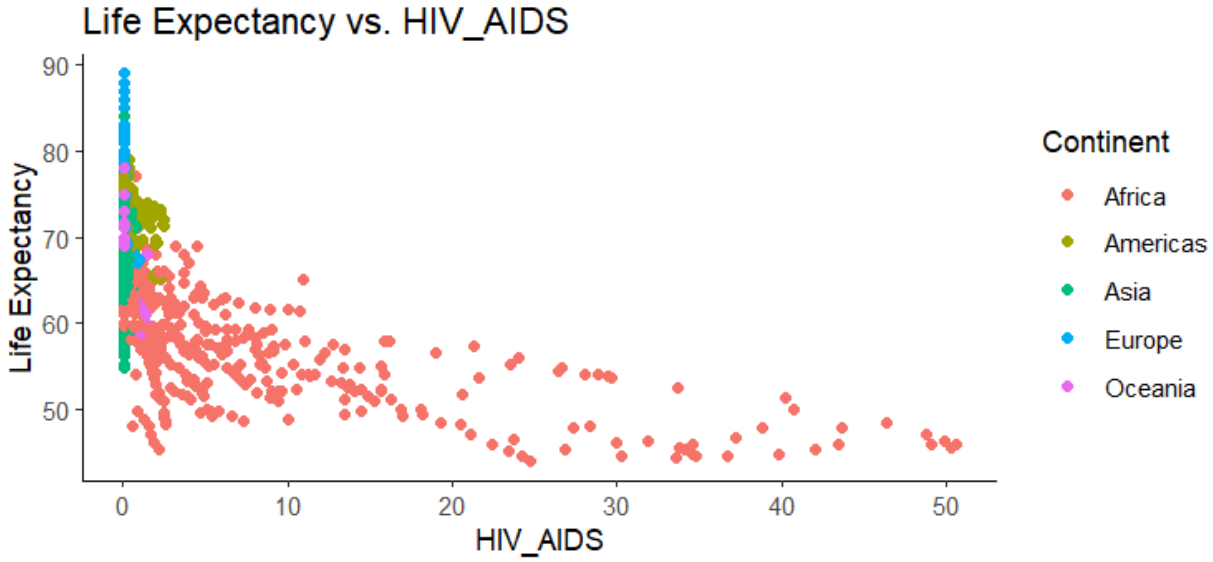

Life Expectancy vs. Population
Life Expectancy
90
80
70
60
50
0e+00
5e+08
1e+09
Population
Continent
Africa
Americas
Asia
Europe
Oceania


Life Expectancy vs. Adult_Mortality
Life Expectancy
90
80
70
60
50
0
200
400
600
Adult_Mortality
Continent
Africa
Americas
Asia
Europe
Oceania


Life Expectancy vs. Alcohol
Life Expectancy
90
80
70
60
50
0
5
10
15
Alcohol
Continent
Africa
Americas
Asia
Europe
Oceania


Life Expectancy vs. Percentage_Expenditure
Life Expectancy
90
80
70
60
50
0
5000
10000
15000
Percentage_Expenditure
Continent
Africa
Americas
Asia
Europe
Oceania


Life Expectancy vs. Diphtheria
Life Expectancy
90
80
70
60
50
0
25
50
75
100
Diphtheria
Continent
Africa
Americas
Asia
Europe
Oceania


Life Expectancy vs. HIV_AIDS
Life Expectancy
90
80
70
60
50
0
10
20
30
40
50
HIV_AIDS
Continent
Africa
Americas
Asia
Europe
Oceania

- Correlation Matrix:

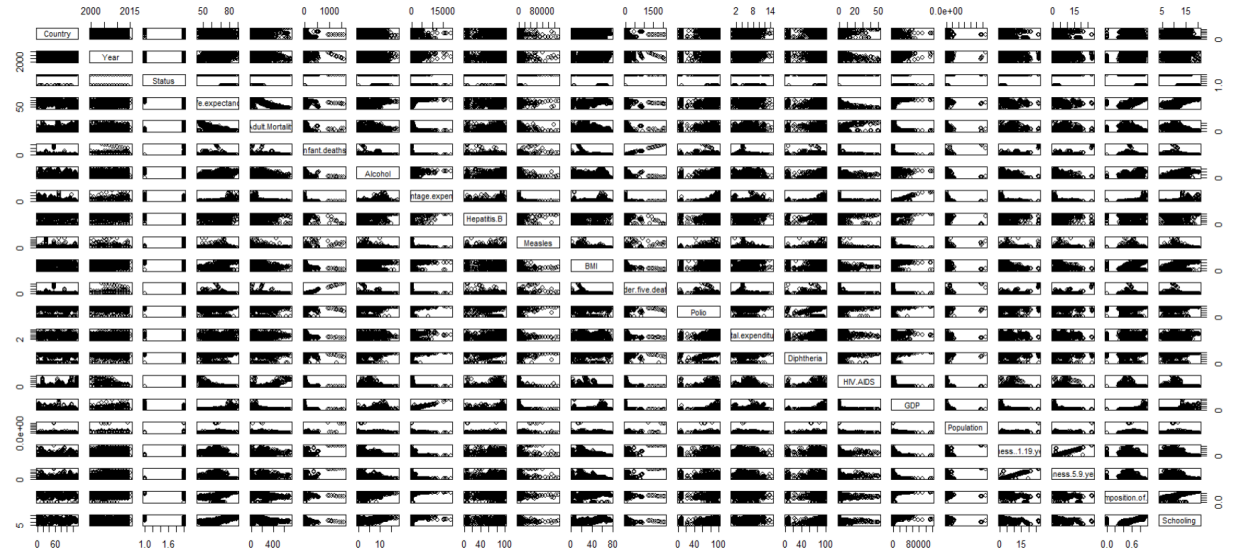
Correlation coefficients were used to quantify the degree and direction of the linear link between each numerical predictor variable and life expectancy. To avoid multicollinearity during the quantitative analysis, strongly correlated numerical predictor variables have been discovered by developing the regression model.

```
> cor(Cleaned_Health_Data[,-c(1,3,4)])
```

	Year	Adult_Mortality	Infant_Deaths	Under_5_Deaths	Alcohol	HIV_AIDS	Hepatitis_B	Polio	Diphtheria	Measles	BMI
Year	1.000000000	-0.037091782	0.008029128	0.01047859	-0.11336476	-0.123404990	0.11489709	-0.01669880	0.02964059	-0.053822046	0.005739061
Adult_Mortality	-0.037091782	1.000000000	0.042450237	0.06036503	-0.17553509	0.550690745	-0.10522544	-0.19985300	-0.19142876	-0.003966685	-0.351542478
Infant_Deaths	0.008029128	0.042450237	1.000000000	0.99690562	-0.10621692	0.007711547	-0.23176894	-0.15692881	-0.16187100	0.532679832	-0.234425154
Under_5_Deaths	0.010478594	0.060365026	0.996905622	1.00000000	-0.10108216	0.019475927	-0.24076603	-0.17116419	-0.17844819	0.517505563	-0.242137398
Alcohol	-0.113364764	-0.175535086	-0.106216917	-0.10108216	1.00000000	-0.027112636	0.10988939	0.24031453	0.24295143	-0.050110235	0.353396205
HIV_AIDS	-0.123404990	0.550690745	0.007711547	0.01947593	-0.02711264	1.000000000	-0.09480197	-0.10788547	-0.11760107	-0.003521854	-0.210896746
Hepatitis_B	0.114897092	-0.105225443	-0.231768937	-0.24076603	0.10988939	-0.094801971	1.00000000	0.46333080	0.58898993	-0.124799993	0.143301786
Polio	-0.016698803	-0.199853000	-0.156928805	-0.17116419	0.24031453	-0.107885468	0.46333080	1.00000000	0.60924547	-0.057850133	0.186267965
Diphtheria	0.029640586	-0.191428759	-0.161871004	-0.17844819	0.24295143	-0.117601074	0.58898993	0.60924547	1.00000000	-0.058605907	0.176294503
Measles	-0.053822046	-0.003966685	0.532679832	0.51750556	-0.05011023	-0.003521854	-0.12479999	-0.05785013	-0.05860591	1.000000000	-0.153245464
BMI	0.005739061	-0.351542478	-0.234425154	-0.24213740	0.35339621	-0.210896746	0.14330179	0.18626797	0.17629450	-0.153245464	1.000000000
Thinness_5to9_Years	0.014122422	0.286722882	0.461907925	0.46228938	-0.38620819	0.183146727	-0.13325099	-0.17448925	-0.18095238	0.174946217	-0.554093981
Thinness_10to19_Years	0.019756611	0.272230044	0.463415256	0.46478470	-0.40375499	0.172591767	-0.12940595	-0.16406959	-0.18724165	0.180641506	-0.547017514
GDP	0.096421485	-0.255034733	-0.098092020	-0.10033126	0.44343279	-0.108080600	0.04184950	0.15680869	0.15843774	-0.064767590	0.266113973
Population	0.012566893	-0.015011838	0.671758310	0.65867969	-0.02888023	-0.027800562	-0.12972265	-0.04538657	-0.03989754	0.321946377	-0.081415982
Percentage_Expenditure	0.069553468	-0.237609890	-0.090764632	-0.09215806	0.41704736	-0.095084991	0.01676017	0.12862605	0.13481324	-0.063070789	0.242738243
Total_Expenditure	0.059492777	-0.085226535	-0.146951117	-0.14580310	0.21488509	0.043100657	0.11332668	0.11976798	0.12991481	-0.113582738	0.189468964
ICOR	0.122891780	-0.442203288	-0.134753863	-0.14809728	0.56107433	-0.248589855	0.18492097	0.31468159	0.34326177	-0.058277256	0.510504831
Schooling	0.088731787	-0.421170523	-0.214371900	-0.22601262	0.61697481	-0.211840201	0.21518159	0.35014660	0.35039793	-0.115660481	0.554843903
Life_Expectancy	0.050771035	-0.702523062	-0.169073804	-0.19226530	0.40271832	-0.592236293	0.19993528	0.32729440	0.34133123	-0.068881222	0.542041588

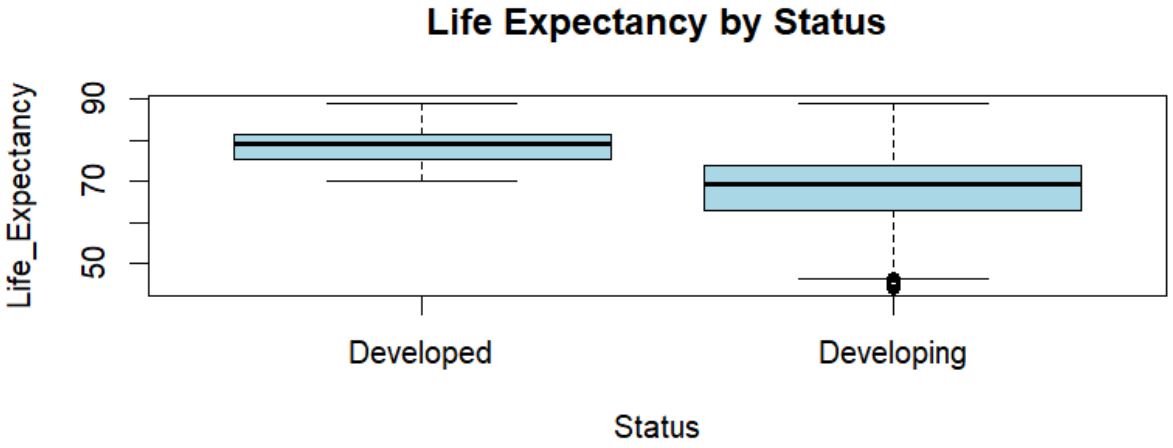
	Thinness_5to9_Years	Thinness_10to19_Years	GDP	Population	Percentage_Expenditure	Total_Expenditure	ICOR	Schooling	Life_Expectancy
Year	0.01412242	0.01975661	0.09642148	0.012566893	0.06955347	0.05949278	0.122891780	0.08873179	0.05077103
Adult_Mortality	0.28672288	0.27223004	-0.25503473	-0.015011838	-0.23760989	-0.08522653	-0.442203288	-0.42117052	-0.70252306
Infant_Deaths	0.46190792	0.46341526	-0.09809202	0.671758310	-0.09076463	-0.14695112	-0.134753863	-0.21437190	-0.16907380
Under_5_Deaths	0.46228938	0.46478470	-0.10033126	0.658679691	-0.09215806	-0.14580310	-0.148097276	-0.22601262	-0.19226530
Alcohol	-0.38620819	-0.40375499	0.44343279	-0.028880232	0.41704736	0.21488509	0.561074332	0.61697481	0.40271832
HIV_AIDS	0.18314673	0.17259177	-0.10808060	-0.027800562	-0.09508499	0.04310066	-0.248589855	-0.21184020	-0.59223629
Hepatitis_B	-0.13325099	-0.12940595	0.04184950	-0.129722655	0.01676017	0.11332668	0.184920970	0.21518159	0.19993528
Polio	-0.17448925	-0.16406959	0.15680869	-0.045386572	0.12862605	0.11976798	0.314681594	0.35014660	0.32729440
Diphtheria	-0.18095238	-0.18724165	0.15843774	-0.039897537	0.13481324	0.12991481	0.343261772	0.35039793	0.34133123
Measles	0.17494622	0.18064151	-0.06476759	0.321946377	-0.06307079	-0.11358274	-0.058277256	-0.11566048	-0.06888122
BMI	-0.55409398	-0.54701751	0.26611397	-0.081415982	0.24273824	0.18946896	0.510504831	0.55484390	0.54204159
Thinness_5to9_Years	1.00000000	0.92791344	-0.27795855	0.277913374	-0.25563544	-0.21786479	-0.438483721	-0.47248203	-0.45750829
Thinness_10to19_Years	0.92791344	1.00000000	-0.27749835	0.282529280	-0.25503460	-0.20987232	-0.453678854	-0.49119921	-0.45783819
GDP	-0.27795855	-0.27749835	1.00000000	-0.020368964	0.95929886	0.18037347	0.446855511	0.46794697	0.44132181
Population	0.27791337	0.28252928	-0.02036896	1.000000000	-0.01679214	-0.07996224	-0.008132466	-0.04031242	-0.02230498
Percentage_Expenditure	-0.25563544	-0.25503460	0.95929886	-0.016792141	1.00000000	0.18387236	0.402169736	0.42208845	0.40963082
Total_Expenditure	-0.21786479	-0.20987232	0.18037347	-0.079962237	0.18387236	1.00000000	0.183653190	0.24378345	0.17471764
ICOR	-0.43848372	-0.45367885	0.44685551	-0.008132466	0.40216974	0.18365319	1.000000000	0.78474058	0.72108259
Schooling	-0.47248203	-0.49119921	0.46794697	-0.040312419	0.42208845	0.24378345	0.784740581	1.00000000	0.72763003
Life_Expectancy	-0.45750829	-0.45783819	0.44132181	-0.022304978	0.40963082	0.17471764	0.721082593	0.72763003	1.00000000

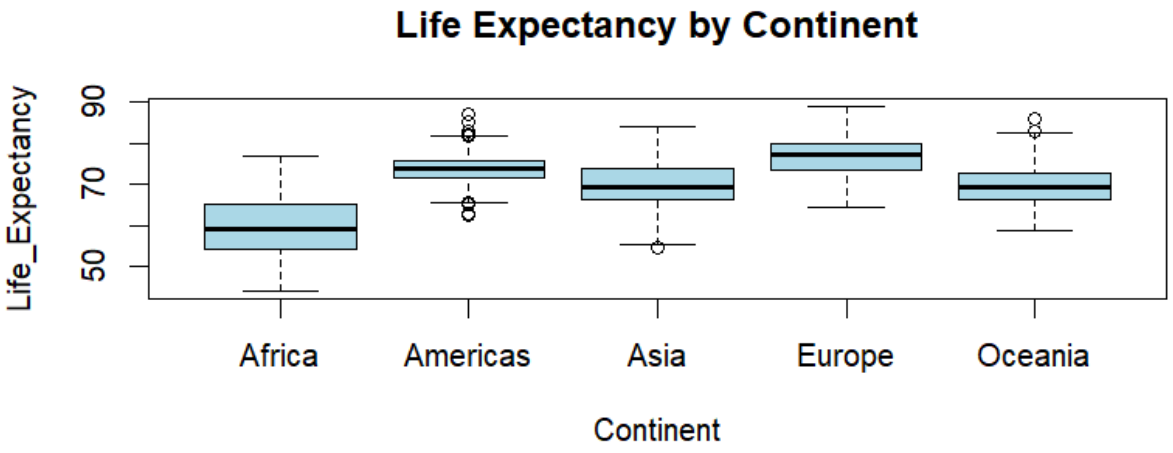
There are few variables which are highly correlated to each other (>0.7) which can be observed from the correlation matrix.
Correlation can also be identified by looking at the scatterplot of independent variables.



- Box Plots:

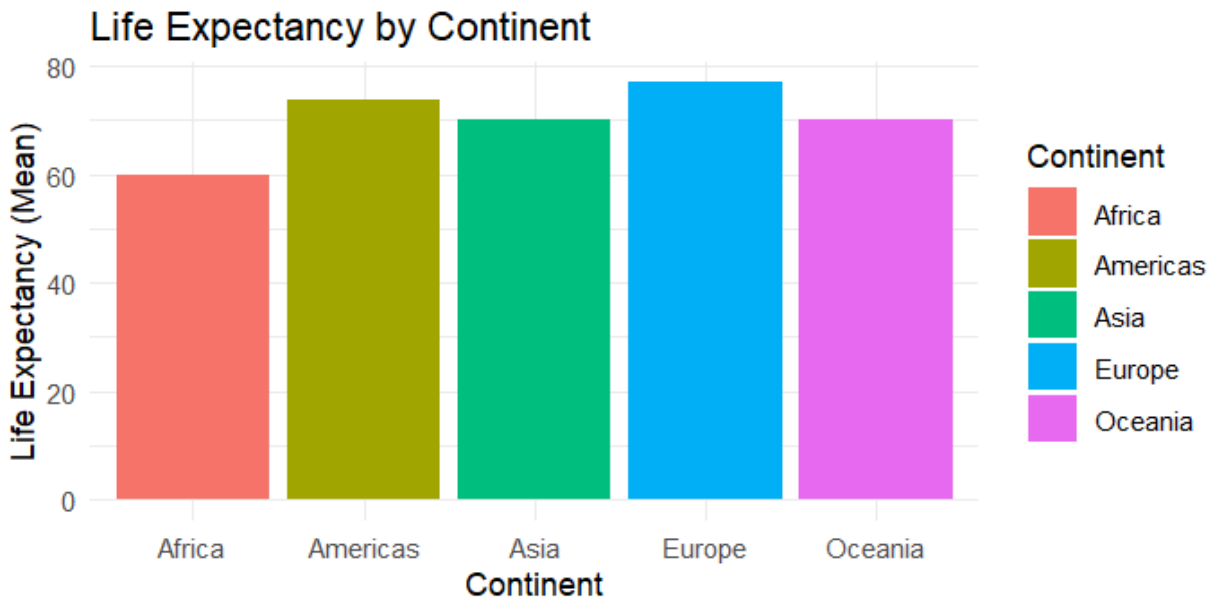
The distribution of life expectancy among different categories of each categorical variable has been investigated to determine whether there is a significant difference in life expectancy values between categories.

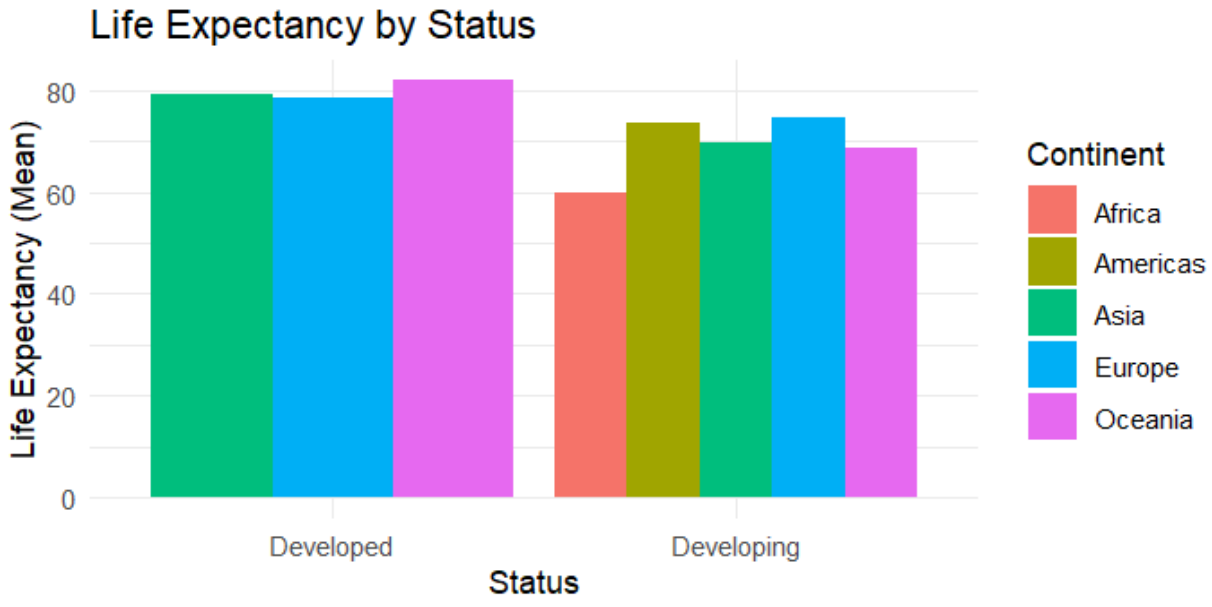




- Life expectancy seems to be higher in developed countries compared to developing ones.
- As per the continent-wise boxplot, the life expectancy is higher in Europe followed by America and Oceania, while Africa has the lowest average life expectancy compared to the other continents.
- Also, there are some outliers observed in the America and Oceania data.

- BarPlots:


Life Expectancy by Continent
Life Expectancy (Mean)
80
60
40
20
0
Africa
Americas
Asia
Europe
Oceania
Continent
Continent
Africa
Americas
Asia
Europe
Oceania


Life Expectancy by Status
Life Expectancy (Mean)
80
60
40
20
0
Developed
Developing
Status
Continent
Africa
Americas
Asia
Europe
Oceania

3.3 Identification of Patterns and Trends:

- A comprehensive analysis of the data was conducted to identify patterns and trends across different variables. This involved examining variations in variables such as GDP, vaccination coverage, and education levels across continents.
- Notable trends such as improvements in life expectancy over time, disparities in health outcomes between developed and developing countries, and associations between socio-economic indicators and health outcomes were observed.

3.4 Insights from EDA:

Several interesting insights emerged from the exploratory data analysis:

- A positive correlation was observed between GDP per capita and life expectancy, suggesting a relationship between economic development and longer life expectancy.
- Vaccination coverage for diseases such as Hepatitis B and Polio exhibited wide variability across continents, indicating disparities in healthcare access and infrastructure.
- Significant variation was noted in income composition of resources and average years of schooling across regions, highlighting differences in socio-economic development and resource allocation.
- Variability in health indicators such as adult mortality rates and prevalence of HIV/AIDS between the continents underscored the importance of addressing healthcare disparities.

3.5 Implications for Further Analysis:

The insights from the exploratory data analysis will inform subsequent analyses, including regression modeling to identify predictors of life expectancy and assess the effectiveness of public health interventions.

3. Quantitative Analysis

The initial multiple linear regression model was developed using the Ordinary Least Squares (OLS) method using all of the available 18 quantitative independent variables and 2 qualitative variables to predict life expectancy.

Outcomes of the model:

1. Life expectancy in developed countries is more likely to be higher than in developing countries.
2. Life Expectancy is negatively related to the variables Adult_Mortality, Under_5_Deaths , Alcohol, Hepatitis_B , HIV_AIDS, Thinness_5to9_Years and Population.
3. Percentage Expenditure, Income Composition of Resources, Schooling, GDP, Immunization Coverage(Polio, Diphtheria), and Total Expenditure are all positively connected to life expectancy.

Adjusted R-Squared Interpretation:

- The model explains 85.1 percent of the variation in the dependent variable Life_Expectancy after accounting for the number of independent variables.
- The model has 12 significant variables out of 20 independent variables based on the 5% significance criterion. To build the optimal regression model, we can use the manual variable elimination, backward elimination or the forward variable selection methods.

Following all the methods, the same regression model is achieved on this dataset.

```
final_model<- lm(Life_Expectancy ~ Status + Adult_Mortality + Infant_Deaths + Under_5_Deaths +
Alcohol + Percentage_Expenditure + BMI +Polio + Diphtheria + HIV_AIDS + Thinness_5to9_Years
+ ICOR + Schooling + Continent,
data = Cleaned_Health_Data)

> summary(final_model)

Call:
lm(formula = Life_Expectancy ~ Status + Adult_Mortality + Infant_Deaths +
    Under_5_Deaths + Alcohol + Percentage_Expenditure + BMI +
    Polio + Diphtheria + HIV_AIDS + Thinness_5to9_Years + ICOR +
    Schooling + Continent, data = Cleaned_Health_Data)

Residuals:
     Min       1Q   Median       3Q      Max
-16.0057  -2.2009   0.1094   2.1074  12.3786
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	54.9298997	0.7755889	70.823	< 2e-16	***
StatusDeveloping	-1.6070430	0.3589032	-4.478	8.07e-06	***
Adult_Mortality	-0.0153076	0.0009020	-16.971	< 2e-16	***
Infant_Deaths	0.0581396	0.0098237	5.918	3.96e-09	***
Under_5_Deaths	-0.0449376	0.0072852	-6.168	8.68e-10	***
Alcohol	-0.1872329	0.0355395	-5.268	1.56e-07	***
Percentage_Expenditure	0.0004767	0.0000561	8.497	< 2e-16	***
BMI	0.0226790	0.0058473	3.879	0.000109	***
Polio	0.0080300	0.0048003	1.673	0.094560	.
Diphtheria	0.0096180	0.0050789	1.894	0.058441	.
HIV_AIDS	-0.3701773	0.0176009	-21.032	< 2e-16	***
Thinness_5to9_Years	-0.0478565	0.0269605	-1.775	0.076074	.
ICOR	8.2859429	0.8039707	10.306	< 2e-16	***
Schooling	0.8316711	0.0568180	14.637	< 2e-16	***
ContinentAmericas	4.2769100	0.3289078	13.003	< 2e-16	***
ContinentAsia	2.7650899	0.2806507	9.852	< 2e-16	***
ContinentEurope	3.2504223	0.4115620	7.898	5.17e-15	***
ContinentOceania	1.2869572	0.4199960	3.064	0.002218	**

Signif. codes: 0 ‘*’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

Residual standard error: 3.393 on 1631 degrees of freedom
Multiple R-squared: 0.8527, Adjusted R-squared: 0.8512
F-statistic: 555.6 on 17 and 1631 DF, p-value: < 2.2e-16

Overall, the model suggests that various factors such as health status, mortality rates, lifestyle factors (like alcohol consumption and BMI), immunization coverage, disease prevalence, and socioeconomic factors (like income composition and schooling) influence life expectancy.

The achieved final model has an R-squared value of 0.8512 which results in explaining about 85% of variance in Life expectancy. Also, the model has all the independent variables as significant when considering alpha at 0.01.

5. Correlation analysis

After building the model, we can check if the variables in the model are correlated or not by checking the Variance Inflation Factor (VIF) using the command vif(model_name) in R. This would list the VIF values for all the independent variables in the model. In general, the variables with VIF >10 are said to be highly correlated to other independent variables and thus we need to remove them from the model.

```
> vif(final_model)
                       GVIF Df GVIF^(1/(2*Df))
Status                2.310048  1       1.519884
Adult_Mortality       1.828522  1       1.352229
Infant_Deaths       201.726086  1      14.203031
Under_5_Deaths      201.579887  1      14.197883
Alcohol               2.934891  1       1.713152
Percentage_Expenditure  1.394210  1       1.180767
BMI                   1.909682  1       1.381913
Polio                 1.662384  1       1.289335
Diphtheria            1.719273  1       1.311210
HIV_AIDS              1.613539  1       1.270252
Thinness_5to9_Years   2.253183  1       1.501061
ICOR                  3.101262  1       1.761040
Schooling             3.610696  1       1.900183
Continent             7.356703  4       1.283321
```

So, the variables Infant_Deaths and Under_5_Deaths are found to have high VIF values. Hence, we need to remove them. However, upon removing the Infant_Deaths variable first, the VIF value of Under_5_Deaths has become normalized. So, it remains in the model.

The variable Thiness_5to9_Years has also been removed from the model since it is not significant at 0.05 level of significance.

The final resulting model now has 12 independent variables which are significant yielding an R squared value of 0.848 which means, the model explains about 85% variance of the predictor variable Life Expectancy.

6. Residual Analysis

Firstly, the error residual values i.e., Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error(MAE) for the final regression below are provided below.

Method	Error Value
MSE	11.38822
RMSE	3.374643
MAE	2.654636

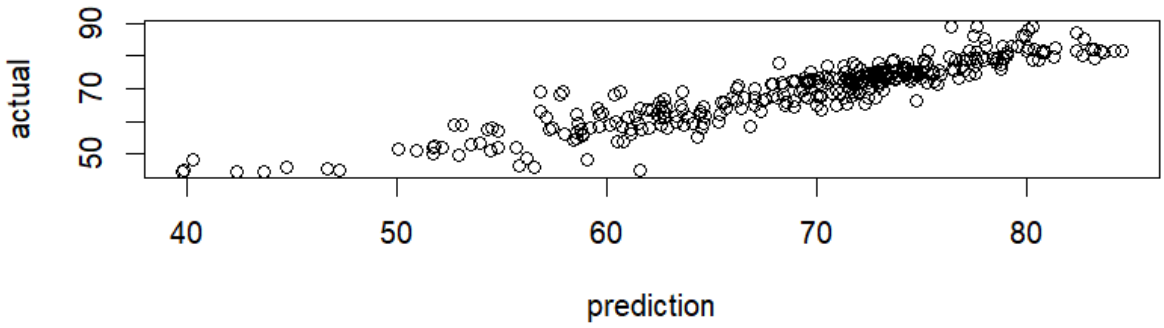
Train and Test Models:

Creating a train and test models on the data, we can identify the predicted Life Expectancy of a created set of data.

```
> prediction <- predict(train_model, test)


> head(prediction)

      4       5       8      11      16      20
64.56593 64.06244 62.87104 61.59536 58.53320 75.17245

> actual = test$Life_Expectancy

> head(actual)

[1] 59.5 59.2 58.1 57.3 54.8 76.9

> cor(prediction, actual)

[1] 0.9215453
```

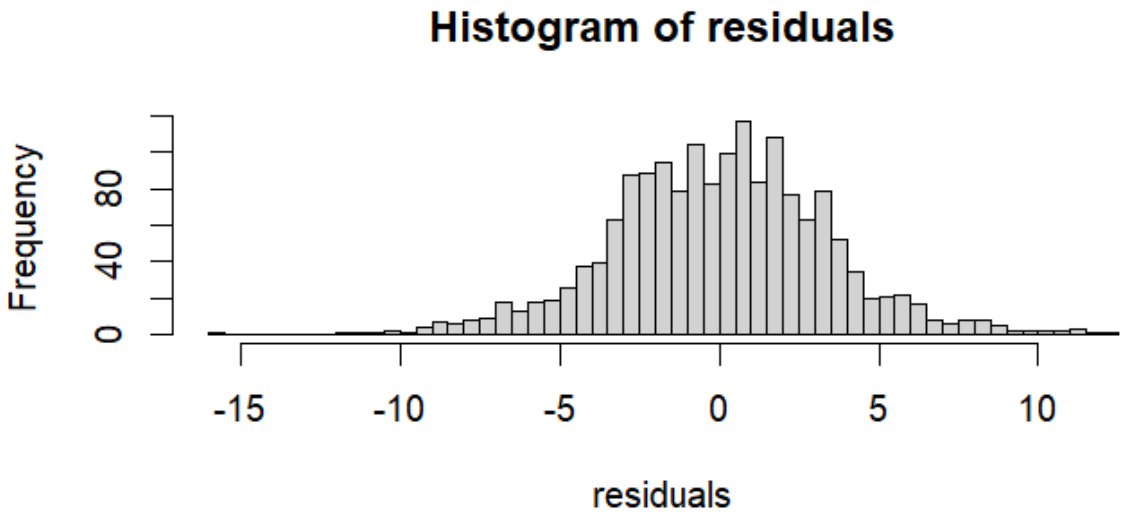
The value 0.92 represents a strong positive linear relationship.

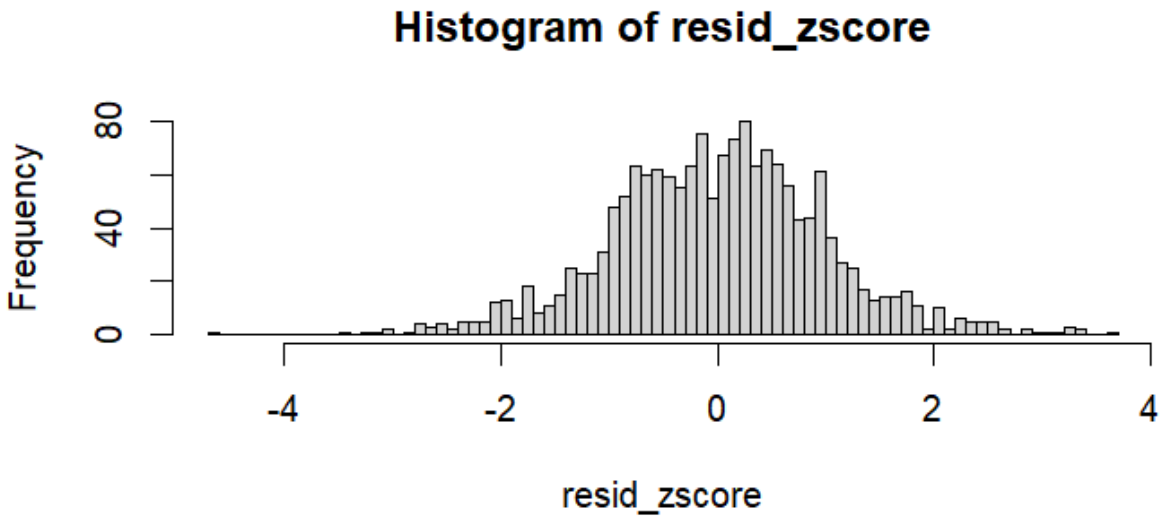
```
> plot(prediction, actual)
```

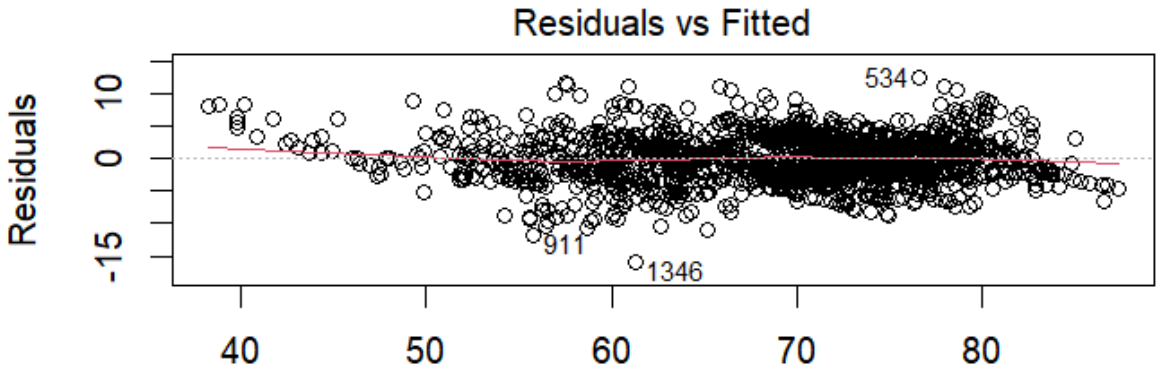


Normality Test:

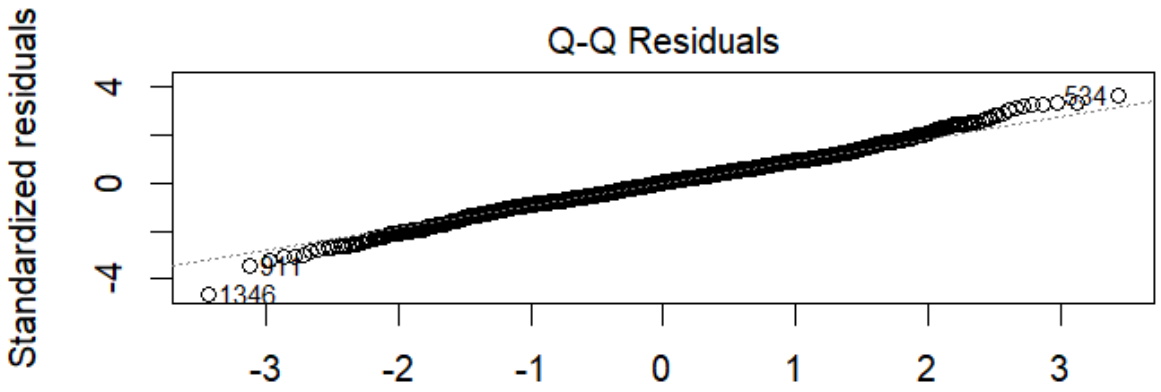
To demonstrate the normality of the residuals, a histogram and a normal q-q plot were utilized.

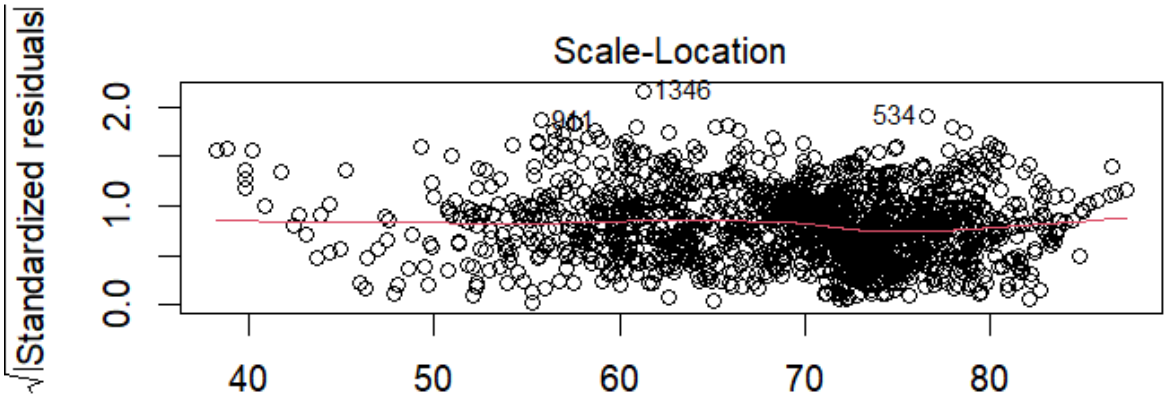


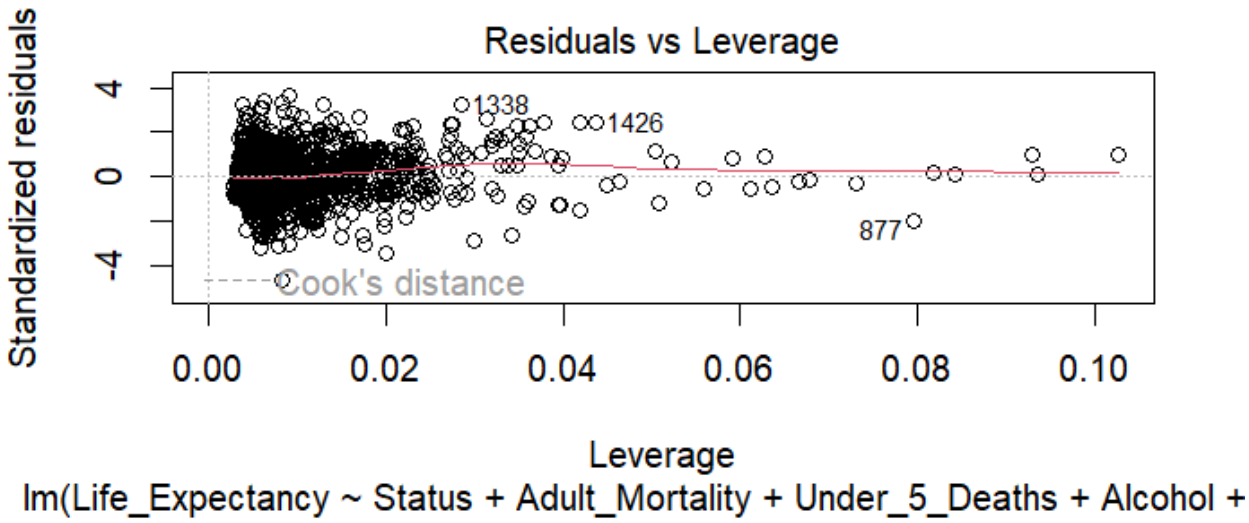




lm(Life_Expectancy ~ Status + Adult_Mortality + Under_5_Deaths + Alcohol +


Q-Q Residuals
Standardized residuals
4
0
-4
534
911
1346
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Life_Expectancy ~ Status + Adult_Mortality + Under_5_Deaths + Alcohol +

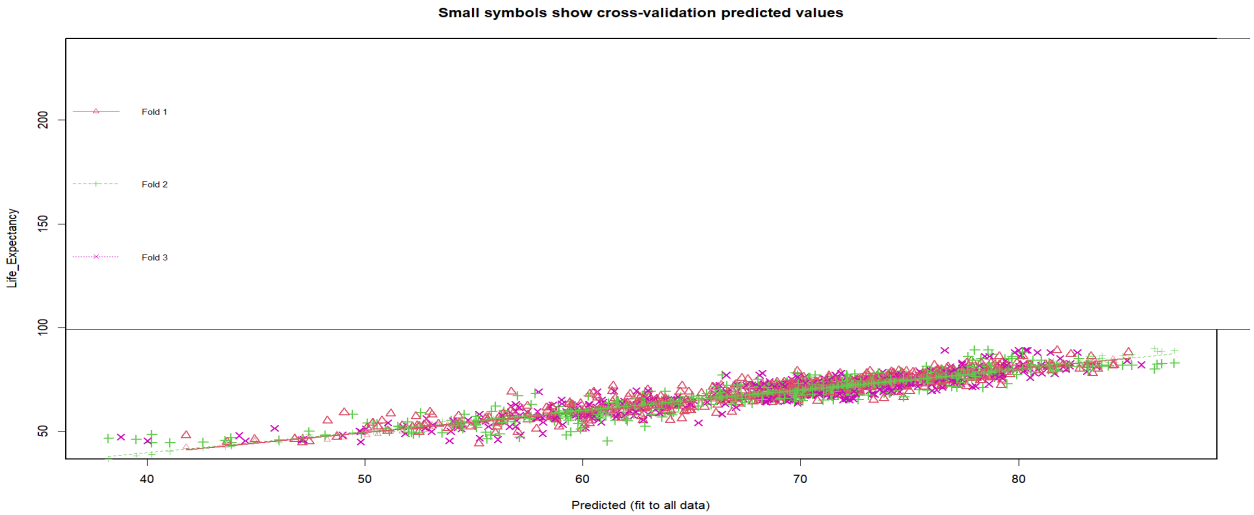

Scale-Location
√|Standardized residuals|
2.0
1.0
0.0
1346
911
534
40
50
60
70
80
Fitted values
lm(Life_Expectancy ~ Status + Adult_Mortality + Under_5_Deaths + Alcohol +



THREE FOLD CROSS VALIDATION

> install.packages("DAAG")
> library(DAAG)

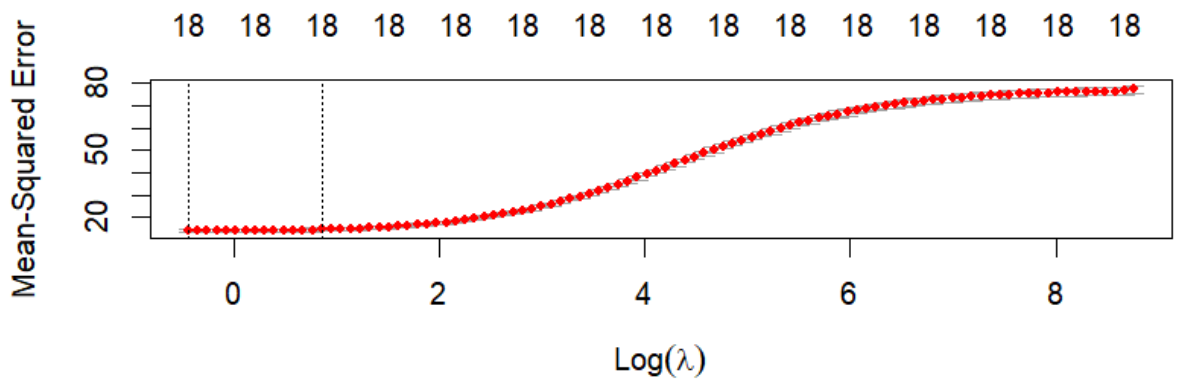
> out <- cv.lm(data = Cleaned_Health_Data, form.lm = formula(Life_Expectancy ~ .), plotit = "Observed", m=3)



5. Advanced Regression Analysis

RIDGE REGRESSION

```
> install.packages("glmnet")
> library(glmnet)

> dim(Cleaned_Health_Data)
[1] 1649   21

> x <- as.matrix(Cleaned_Health_Data[,1:20])

> y <- as.matrix(Cleaned_Health_Data[,21])

> set.seed(123)

> ridge <- cv.glmnet(x, y, family="gaussian", alpha=0)

> plot(ridge)
```

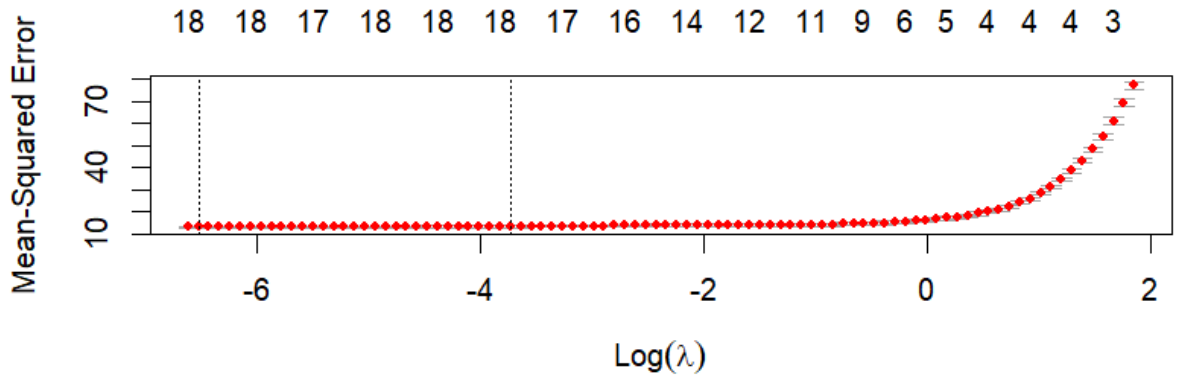



```
> ridge$lambda.min
[1] 0.63989

> coef(ridge, s=ridge$lambda.min)
21 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)          5.312306e+01
Status               .
Continent            .
Adult_Mortality     -1.760321e-02
Infant_Deaths        1.628477e-03
```

```
Under_5_Deaths          -3.588963e-03
Alcohol                 -6.568265e-02
HIV_AIDS                -4.172599e-01
Hepatitis_B             -4.221623e-03
Polio                    1.164462e-02
Diphtheria               1.935137e-02
Measles                  1.222300e-05
BMI                      3.785084e-02
Thinness_5to9_Years     -2.710774e-02
Thinness_10to19_Years   -3.245752e-02
GDP                      3.274683e-05
Population               2.550275e-09
Percentage_Expenditure   2.674838e-04
Total_Expenditure        6.961010e-02
ICOR                     1.008063e+01
Schooling                8.033812e-01
```

LASSO REGRESSION

```
> set.seed(123)
> lasso <- cv.glmnet(x,y,family="gaussian", alpha=1)
There were 21 warnings (use warnings() to see them)
> plot(lasso)
```

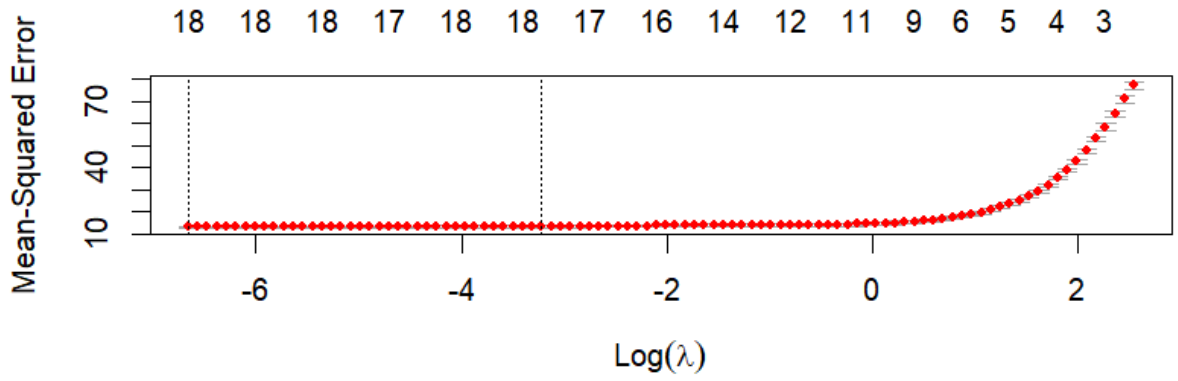



```
> lasso$lambda.min
[1] 0.001478229
> coef(ridge, s=lasso$lambda.min)
21 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)          5.312306e+01
Status                .
Continent             .
Adult_Mortality      -1.760321e-02
Infant_Deaths         1.628477e-03
```

Under_5_Deaths -3.588963e-03
Alcohol -6.568265e-02
HIV_AIDS -4.172599e-01
Hepatitis_B -4.221623e-03
Polio 1.164462e-02
Diphtheria 1.935137e-02
Measles 1.222300e-05
BMI 3.785084e-02
Thinness_5to9_Years -2.710774e-02
Thinness_10to19_Years -3.245752e-02
GDP 3.274683e-05
Population 2.550275e-09
Percentage_Expenditure 2.674838e-04
Total_Expenditure 6.961010e-02
ICOR 1.008063e+01
Schooling 8.033812e-01

ELASTICNET REGRESSION

```
> set.seed(123)
> elasticnet <- cv.glmnet(x,y,family="gaussian", alpha=0.5)
There were 21 warnings (use warnings() to see them)
> plot(elasticnet)
```




```
> elasticnet$lambda.min
[1] 0.00127978
> coef(ridge, s=elasticnet$lambda.min)
21 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)          5.312306e+01
Status               .
Continent            .
Adult_Mortality      -1.760321e-02
Infant_Deaths        1.628477e-03
```

Under_5_Deaths -3.588963e-03
Alcohol -6.568265e-02
HIV_AIDS -4.172599e-01
Hepatitis_B -4.221623e-03
Polio 1.164462e-02
Diphtheria 1.935137e-02
Measles 1.222300e-05
BMI 3.785084e-02
Thinness_5to9_Years -2.710774e-02
Thinness_10to19_Years -3.245752e-02
GDP 3.274683e-05
Population 2.550275e-09
Percentage_Expenditure 2.674838e-04
Total_Expenditure 6.961010e-02
ICOR 1.008063e+01
Schooling 8.033812e-01

6. Conclusion

- The final regression model achieved has an R-squared value of 0.85 which explains about 85% of variance in predicting the Life Expectancy.
- It is observed that the life expectancy increases with an increase in the socio-economic factors like GDP, Percentage_Expenditure, Income Composition of resources, Immunization Coverage(Polio, Hepatitis_B, Diphtheria) and Schooling.
- There is a negative relationship with demographic variables like Adult_Mortality, Infant_Deaths and Alcohol.

7. Future Scope

- The model can further be refined employing various advanced regression methods for a deeper analysis into the data.
- Several other influential factors like pollution, diet, lifestyle and more could be introduced in order to predict the average life expectancy at a better extent.
- For example, the variable ‘alcohol’ is inversely proportional to the response variable. Which means, the increase in the alcohol consumption would lower the life expectancy. But, delving deep into the lifestyle, diet and other factors, there is a potential chance of maintaining the average life expectancy despite the alcohol consumption.