Life Expectancy (WHO)

SKURATIVSKA KATERYNA

CALTRAN LORENZO

ZOLGHADR SHARARE

Introduction

Our motivation in choosing this data-set for further analysis was due to problems that we hope we can answer:

1. Does various predicting factors which has been chosen initially really affect the Life expectancy?
2. What are the predicting variables actually affecting the life expectancy?
3. 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?
4. Does Life Expectancy has positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol etc.
5. What is the impact of schooling on the lifespan of humans?
6. Does Life Expectancy have positive or negative relationship with drinking alcohol?
7. Do densely populated countries tend to have lower life expectancy?
8. What is the impact of Immunization coverage on life Expectancy?
9. Do the sample gives enough evidence to say that Developed countries have more average life expectancy than Developing countries?
10. Do the countries that spend a higher proportion of their resources on human development have a higher life expectancy?
11. What is the most frequent range of life expectancy?

Obtaining data

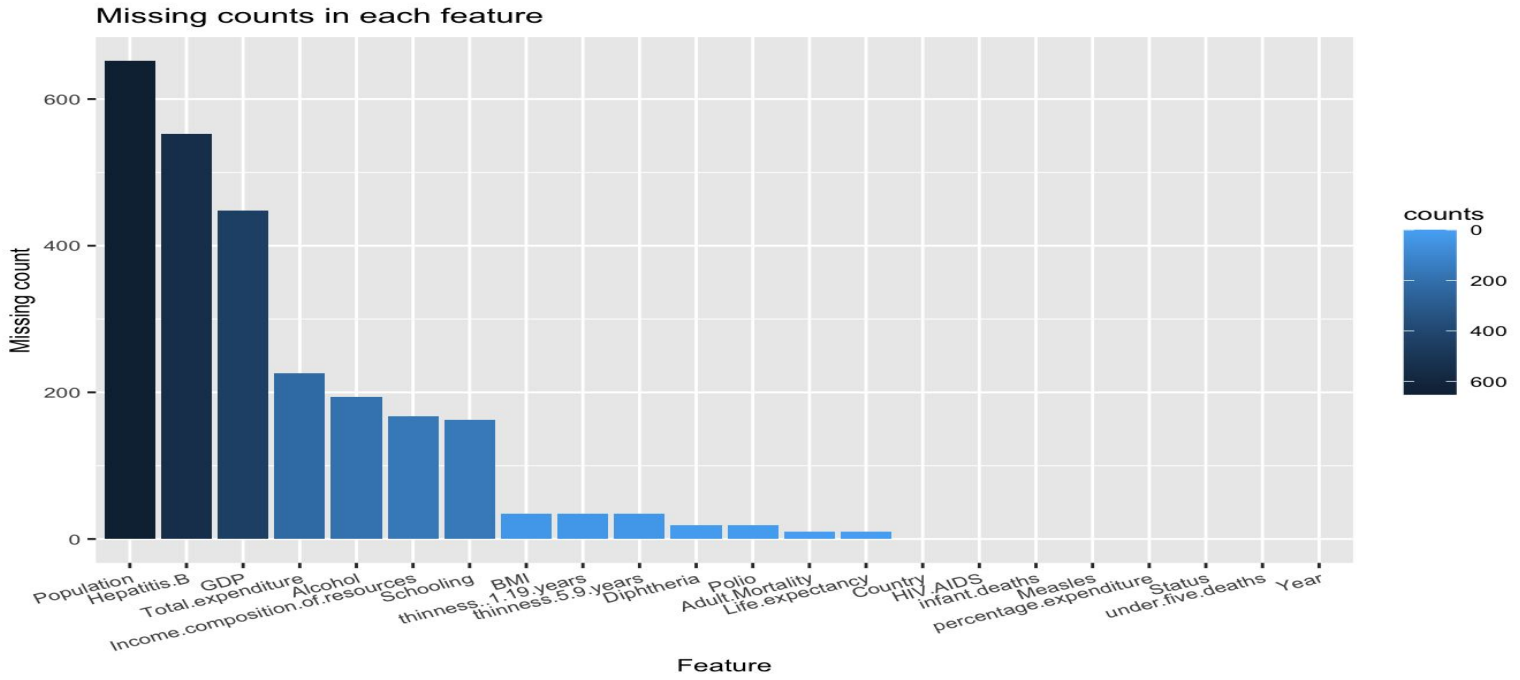
For this project, we obtained the Life Expectancy dataset from Kaggle. The health factors data was collected from the WHO data repository website, and the corresponding economic data was obtained from the United Nations.

The dataset was collected for 193 countries between the years 2000-2015, and it consists of 2938 observations and 22 attributes, of which 20 are meant to be predicting variables.

More specifically, our goal is to use these 20 variables to predict our target feature: Life Expectancy.

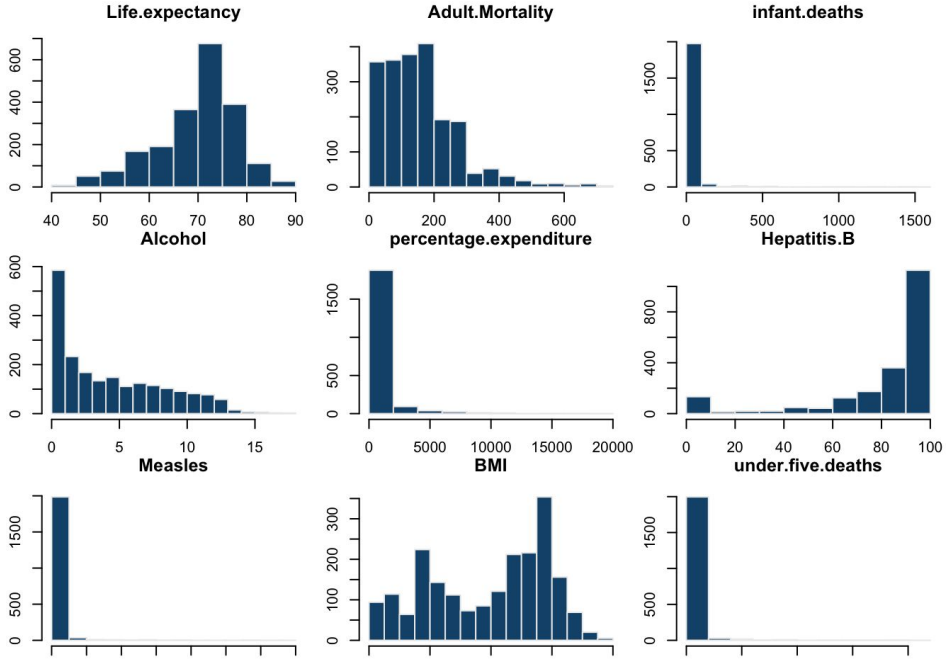
Importing data set and cleaning data

In this step, we examine the presence of missing values in our dataset and perform necessary preprocessing. Approximately 43.87% of the dataset contains missing values, which is nearly half of the data. It is important to analyze how missing values are distributed across different attributes.



We decided to conduct further research and obtained actual values for most of the missing data for Population and GDP from The World Bank website;

To handle the missing data for the other variables, we will utilize mean or median imputation. Mean imputation is suitable for attributes that follow a normal or approximately symmetric distribution without significant outliers. On the other hand, median imputation is more appropriate for attributes with skewed distributions and significant outliers



Handling outliers

There are several methods available for outlier detection, including visual techniques, such as:

- Box-plots
- Histograms

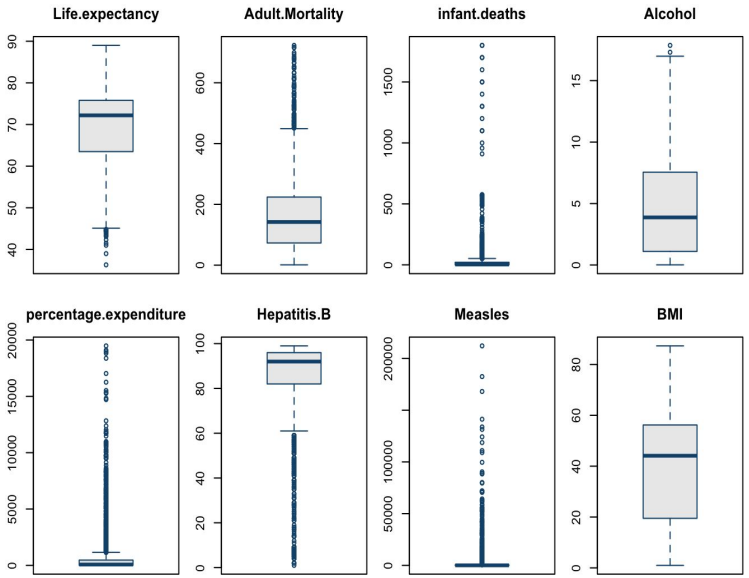
We can also use statistical methods such as:

- Tukey's Method
- Z-score method

Once outliers have been identified using these methods, it is important to preprocess them accordingly.

Several techniques can be employed for outlier preprocessing:

- Dropping outliers
- Limiting/Winsorizing outliers
- Transforming the data



Outliers detection

To see which detection method gives the best result we first detect and remove the outliers detected by using that method, then we build a linear regression model using the dataset with the outliers removed.

Model	Adjusted R-squared
Data_winsorize	0.8722
Data-Z	0.8638
Data-Tukey	0.7602
Data-impute-outlier	0.8159

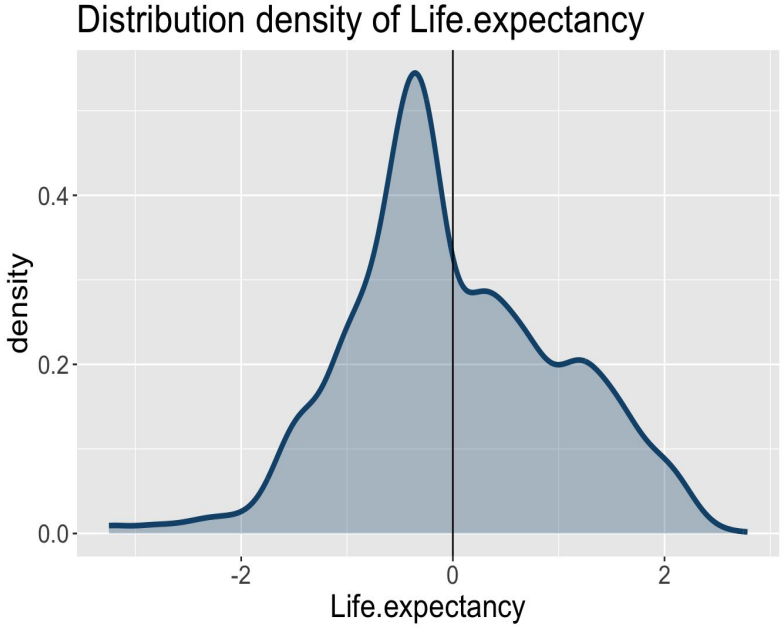
By looking at the results, we observe that the winsorization model gives the highest adjusted R^2, however it is important to note that the winsorization model, may potentially disturb the normality of the distribution of the life expectancy variable.

Taking this into consideration, we will choose the z-score as detection method and we're going to handle outliers by removing them, in fact in this way we obtain a high R^2 score (0.8638), while preserving the normality assumption of the life expectancy distribution.

Data transformation

Since our analysis focuses on predicting Life Expectancy, it is crucial to ensure that this attribute follows a normal distribution. In order to improve the distribution and address any potential skewness, we applied a square root transformation to the Life Expectancy values.

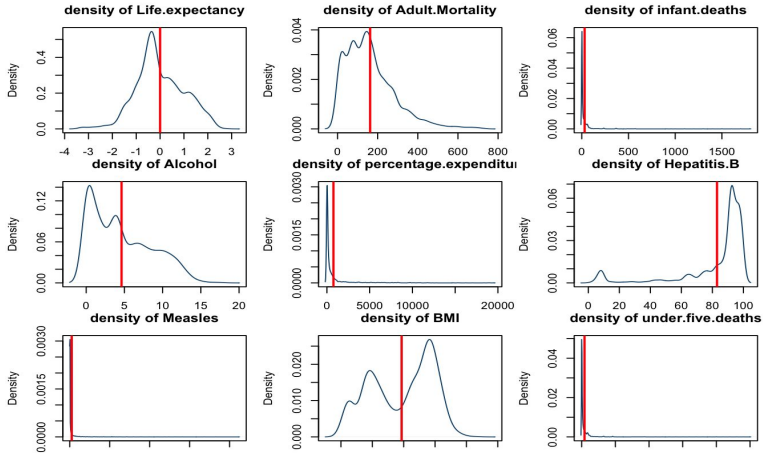
Another important aspect of data preprocessing is scaling the variables. It enables us to compare and analyze variables with different scales and units without any dominance based on their magnitudes. Scaling is particularly beneficial when working with algorithms that are sensitive to variable scales, such as regression models or distance-based algorithms.

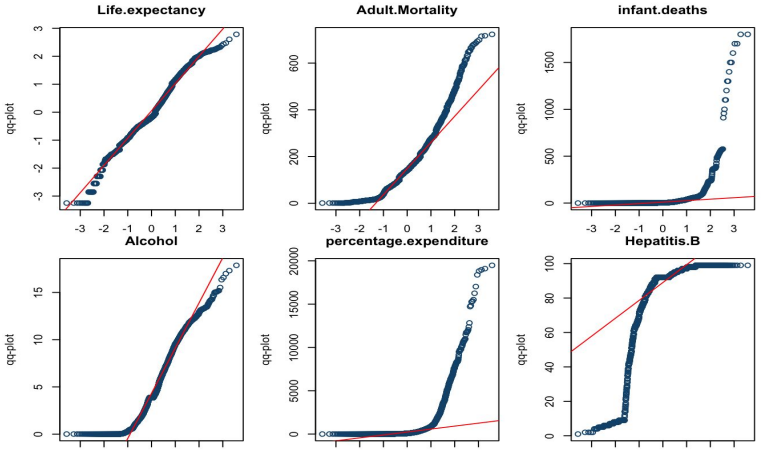


Exploration of the data

In this section we are going to use both Univariate and Bivariate analysis. Our goals are:

- Exploring the relationship between continuous variables and the target variable (Life Expectancy) as well as their interrelationships.
- Investigating the impact of categorical variables on the target variable.
- Examining the relationship between the variables "Country", "Status" and "Year" with continuous variables.

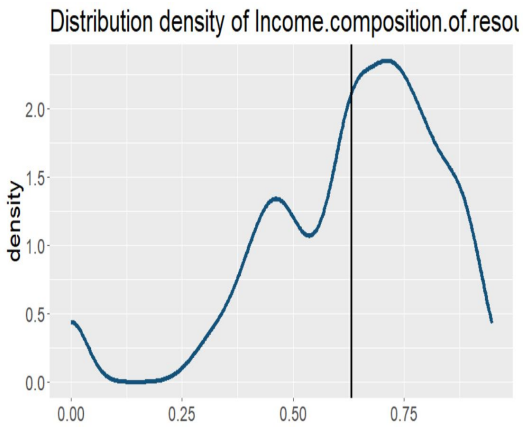


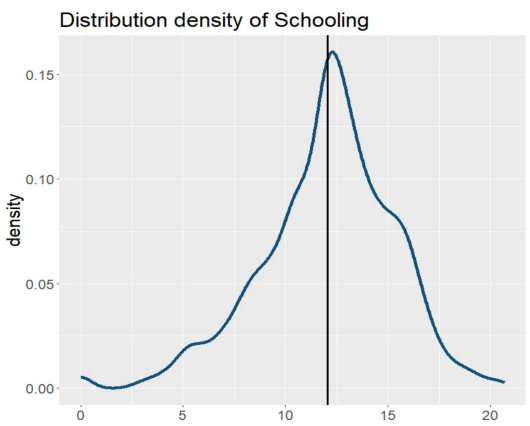


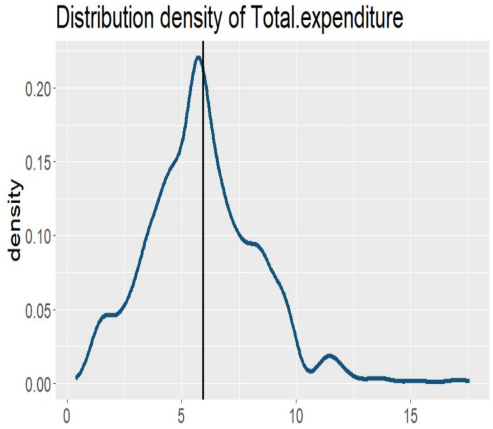
Univariate analysis

Univariate analysis is looking at the data for each variable on its own. This is generally done best by using histograms for continuous data, count/bar-plots for categorical data and of course by getting the descriptive stats.

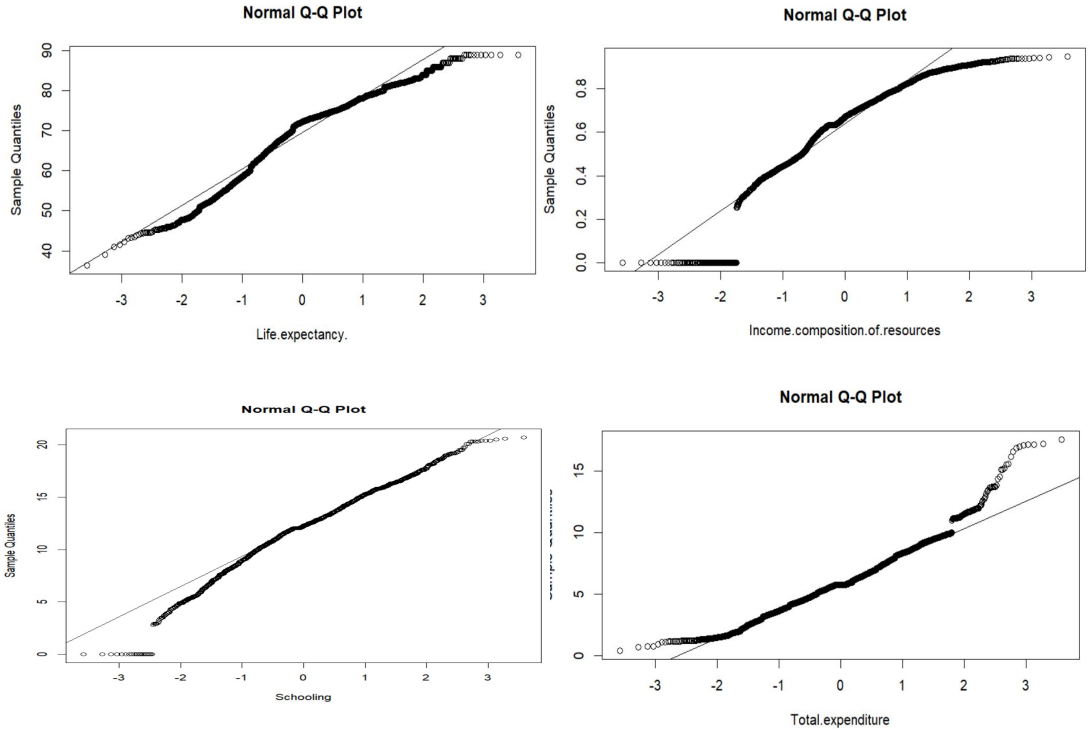
As we can observe, besides Life Expectancy, the variables Total Expenditure, Income Composition of Resources and Schooling look like having normal distributions as well.





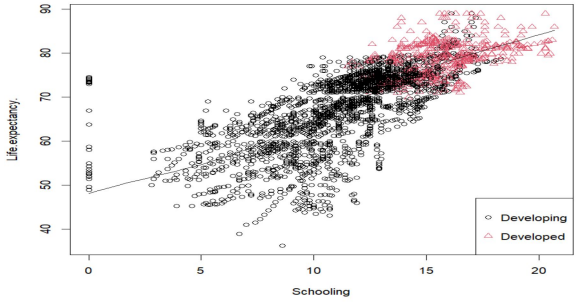


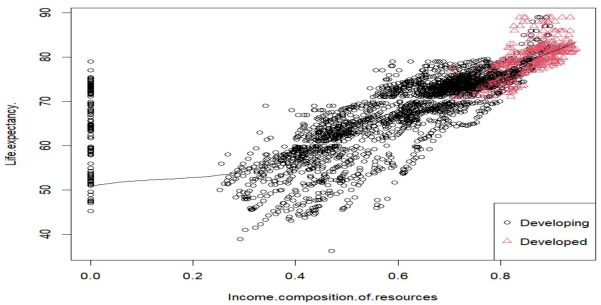
To get a better assessment of the normality of the distributions we can produce the Q-Q Plots

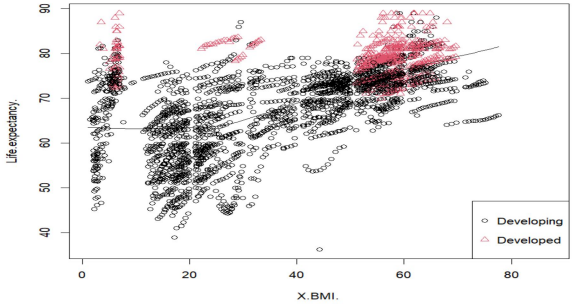


Bivariate analysis

- Continuous variables compared to the life expectancy (target variable) and to one another
- Categorical variables compared to the life expectancy (target variable)
- Comparison of the categorical variables Country, Status and Year to Continuous variables

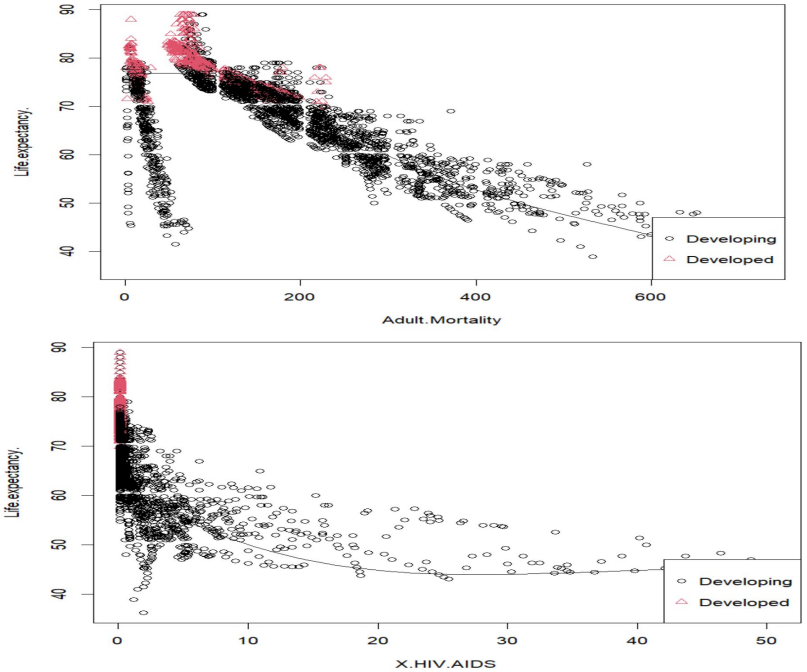




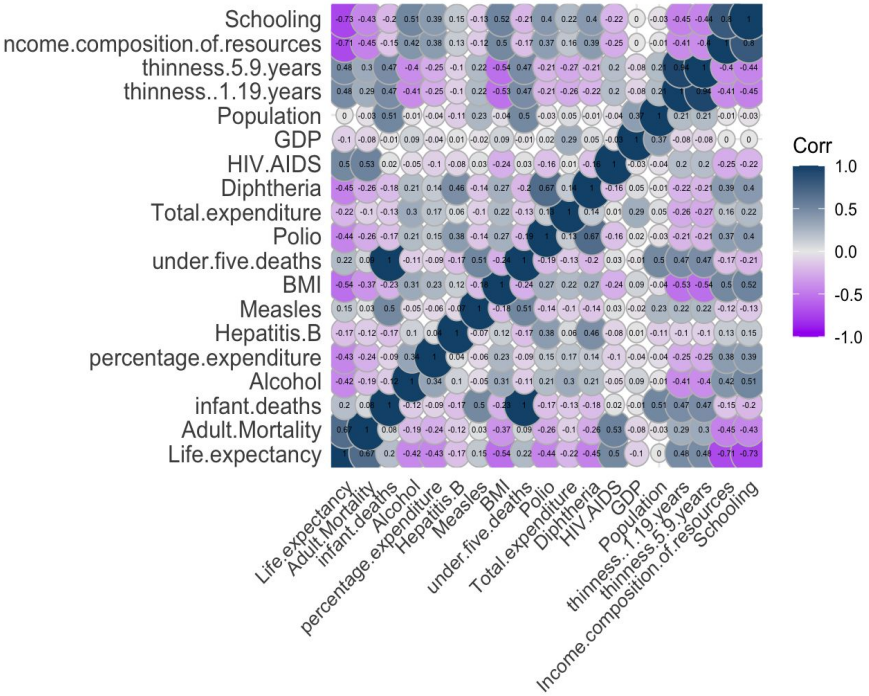


As we can see, these three features (‘Schooling’, ‘Income composition of resources’ and ‘BMI’) have a strong positive correlation with Life Expectancy.

On the other hand **'Adult Mortality'**, **'HIV/AIDS'** have a negative correlation with Life Expectancy.



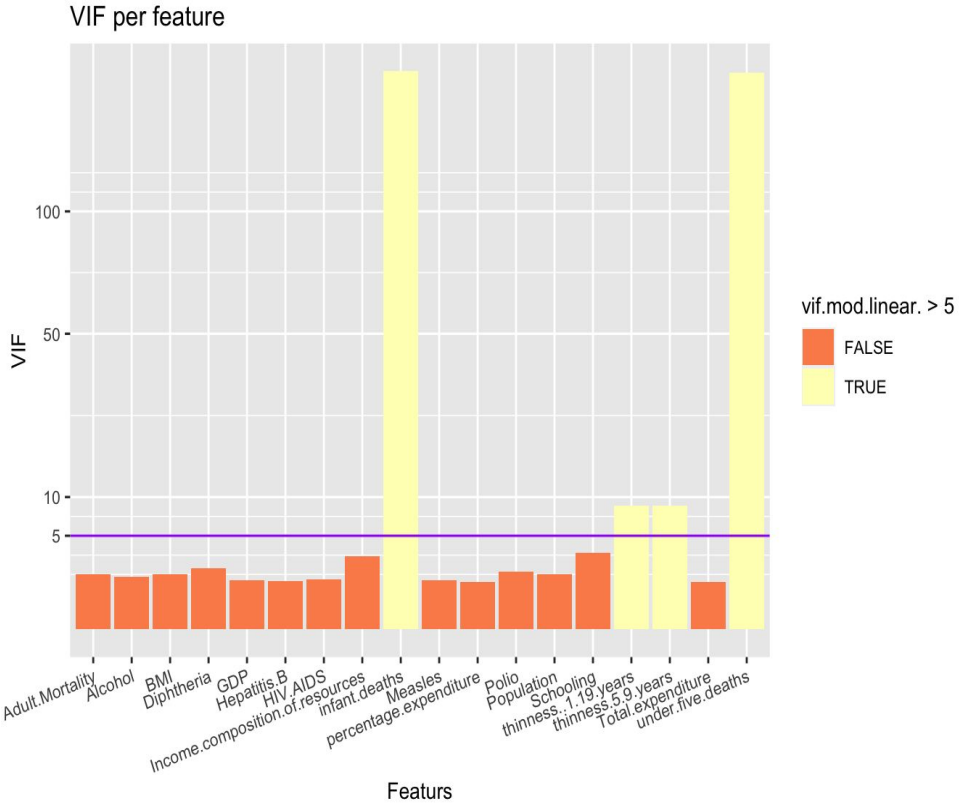
In our analysis, we used the correlation matrix to explore the relationships among the scaled variables. The matrix was visualized as a heatmap, where darker or lighter shades indicated stronger correlations.



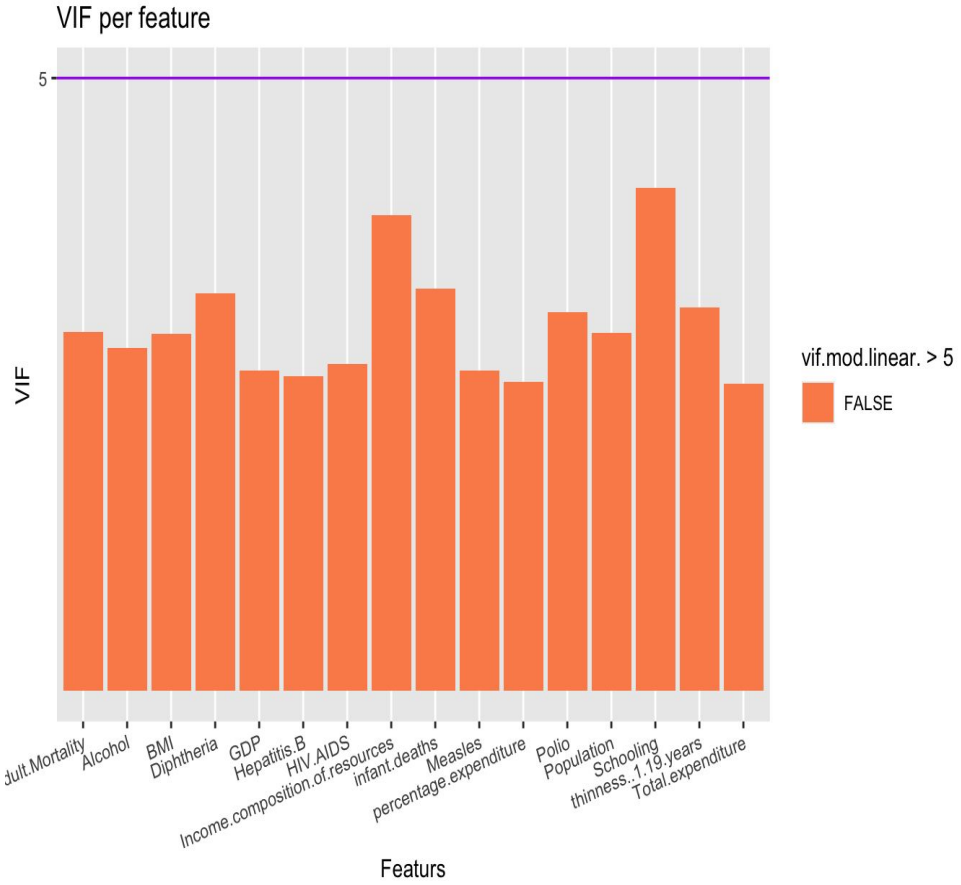
After scaling the variables, we examined the correlation between them and identified several weak correlations:

- Schooling and Income Composition of Resources;
- Thinness 5-9 years and Thinness 1-19 years;
- Under-five deaths and Infant deaths;
- Polio and Diphtheria

To assess the potential collinearity, we recommend calculating the Variance Inflation Factor (VIF) for the variables in a linear model. The VIF helps identify if collinearity is present by quantifying the inflation in the variances of the regression coefficients.



After observing the correlation matrix, it becomes evident that three pairs of variables display high Variance Inflation Factors (VIFs). In order to address this issue of collinearity, we will proceed by omitting one variable from each correlated pair based on their VIF values. Specifically, the variables "infant deaths" and "under five deaths“ exhibit VIFs that significantly exceed the threshold of 5. To resolve this, we will remove the variable “under five deaths" since it possesses a higher VIF value. Similarly, we will eliminate the variable "Income composition of resources" and "thinness.5.9.years”.



Continuous to Life Expectancy comparison

To check if the continuous variables influence Life Expectancy we apply ANOVA Test to each variable.

For each variable we will categorize countries into one of the three categories: 'Low', 'Medium' and 'High' depending on the country's average for that certain feature.

First we group the data by country and find the average life expectancy over the 16 years and we compute the average for the feature we want to test.

We are going to get a new data frame having average life and level of the tested feature (low, medium, or high) as columns and each row corresponding to one among the 193 countries in the dataset.

We then apply the ANOVA Test, where the null hypothesis is H0: mu_low = mu_medium = mu_high and the alternate hypothesis is that not all the means are equal.

Variable	P-value
Adult mortality	<2e-16
Alcohol	9.59e-11
Percentage expenditure	<2e-16
Hepatitis B	0.00144
Measles	8.85e-08
BMI	<2e-16
Total expenditure	0.00153

Variable	P-value
HIV	<2e-16
GDP	<2e-16
Population	0.318
Thinness 1-19 years	<2e-16
Thinness 5-9 years	<2e-16
Income decomposition of Resources	<2e-16
Schooling	<2e-16

As we can see, all the tests return a p-value lower than 0.05, except the one for Population. This means that all the variables, except Population, have effect of Life Expectancy.

In Covid-19 times we all have seen the importance of immunization against the virus to increase life expectancy. Looking at the data, can you show that immunization against Polio and Diphtheria has a significant effect on life expectancy?

We will use a two-way ANOVA test. Here we will divide the countries into two categories for both Polio and Diphtheria. Countries having values of % immunization coverage for one-year-old greater than the median value will get category ‘High’ else ‘Low’.

- Countries with polio (mean) coverage for one-year-old ≤85 will get a label ‘Low’ else ‘High’.
- Countries with Diphtheria (mean) coverage for one-year-old ≤85 will get a label ‘Low’ else ‘High’.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Polio         1   4346    4346   87.34  < 2e-16 ***
Diphtheria    1   1548    1548   31.10 8.64e-08 ***
Residuals   184   9157      50
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value for both Polio and Diphtheria immunization coverage for one-year old is less than 0.05, hence we can say that immunization has a significant impact on the life expectancy.

To assess the relationship between alcohol consumption and adult mortality rate, we can conduct a Pearson correlation test to quantitatively measure the correlation strength between alcohol consumption and adult mortality rate. By doing that, we can gain a deeper understanding of the association between these variables.

```
 Pearson's product-moment correlation

data:  data3$Average_Adult_Mortality and data3$Average_Alcohol
t = -3.8229, df = 185, p-value = 0.00018
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3986006 -0.1322244
sample estimates:
       cor
-0.2705838
```

```
 Pearson's product-moment correlation

data:  data2$Average_Life and data3$Average_Alcohol
t = 6.6054, df = 185, p-value = 4.086e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3129765 0.5461110
sample estimates:
      cor
0.4368508
```

The correlations between alcohol consumption and health indicators yield mixed results. While there is a positive correlation with life expectancy, there is a negative correlation with adult mortality rate. These correlations, however, are not strong enough to draw remarkable conclusions.

Categorical Variables to Life Expectancy Comparison

We divide the dataset into developing countries and developed countries and for each country we compute the mean of the Life Expectancy values obtained through the years.

First we want to check if the variance of the developed countries is the same as the variance of the developing countries, for this we will use a F-test.

Then we are going to see if the developed countries have a higher average life expectancy than Developing countries, for this we are going to use a two sample t-test.

```
        F test to compare two variances

data:  developed$Average_Life and developing$Average_Life
F = 0.13941, num df = 31, denom df = 146, p-value = 2.241e-08
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.08406763 0.25570670
sample estimates:
ratio of variances
         0.1394094
```

```
        Welch Two Sample t-test

data:  developed$Average_Life and developing$Average_Life
t = 13.165, df = 134.02, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 10.49488      Inf
sample estimates:
mean of x mean of y
 79.19785  67.19263
```

Status Variable Compared to other Continuous Variables

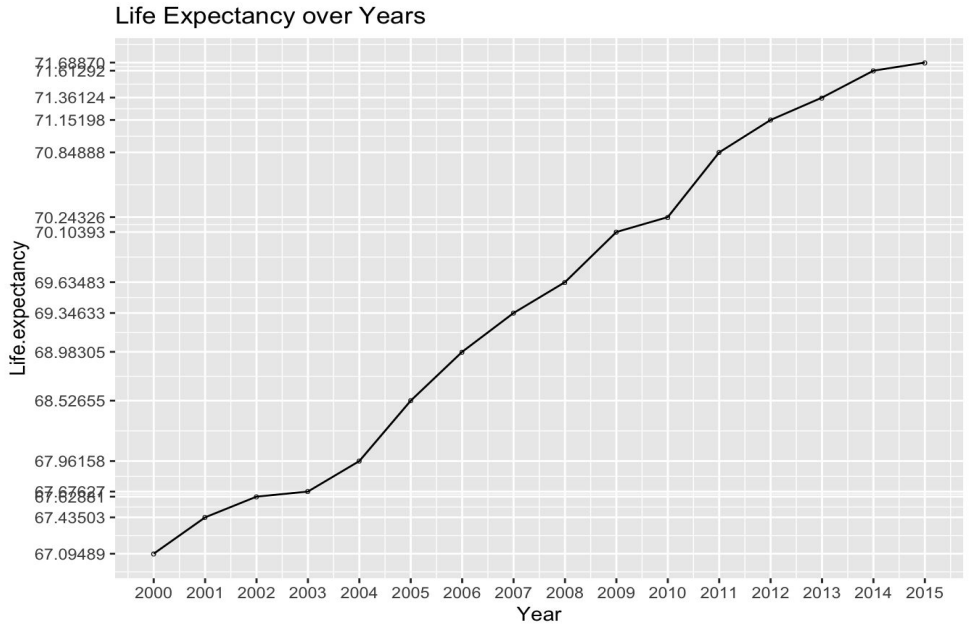
Since the status variable only contains two different values, it is likely best to compare a number of descriptive statistics for those two values with respect to all the other continuous variables.

Variable	P-value
Polio	4.63e-74
Total Expenditure	8.25e-37
Diphteria	7.34e-63
GDP	6.26e-45
Population	0.29
Thinness 1-19 years	1.17e-38
Thinness 5-9 years	6.65e-304
Income decomposition of resources	3.21e-300
Schooling	7.18e-215

Variable	P-value
Life Expectancy	4.45e-312
Adult Mortality	1.53e-166
Infant deaths	5.29e-37
Alcohol	3.70e-198
Percentage Expenditure	8.46e-38
Measles	2.22e-16
Under Five Deaths	1.17e-38
Hepatitis B	6.64e-17
HIV/AIDS	2.78e-60
BMI	1.54e-65

Life Expectancy over the years

In this section we aim to investigate the trend of life expectancy over the years. While there appears to be a positive correlation between life expectancy and the passage of years, it is essential to determine whether the differences observed between each year are statistically significant.



Based on the results of the conducted t-tests, the p-values obtained for all comparisons between consecutive years are greater than 0.05, indicating that there is no significant evidence to support the presence of substantial differences in Life Expectancy between these years.

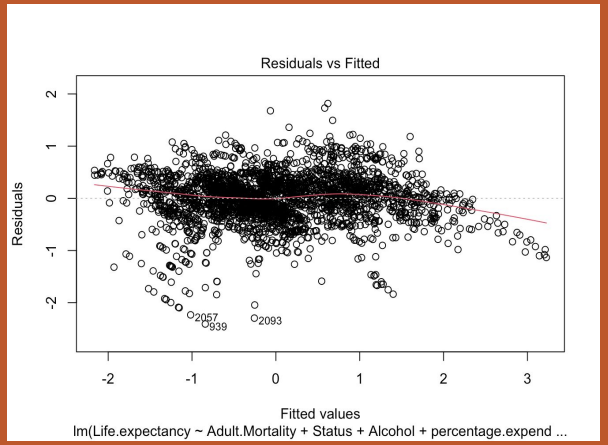
Years	P-value	Years	P-value
2000-2001	0.7413284	2008-2009	0.6288716
2001-2002	0.8458235	2009-2010	0.8541843
2002-2003	0.949058	2010-2011	0.5278882
2003-2004	0.7579069	2011-2012	0.7830066
2004-2005	0.6189956	2012-2013	0.7871843
2005-2006	0.6110444	2013-2014	0.7159781
2006-2007	0.7044814	2014-2015	0.9554911
2007-2008	0.813566		

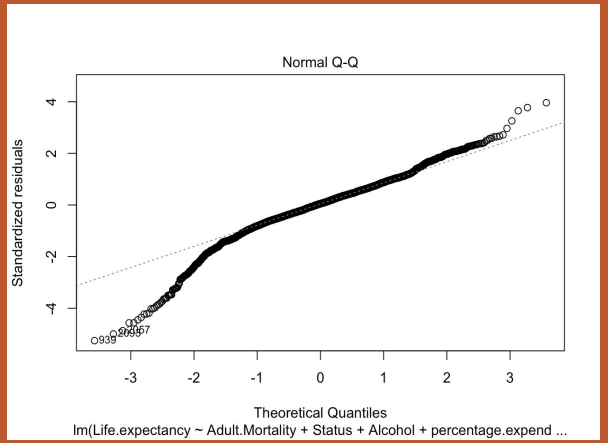
Model selection

To apply linear regression we need to make sure that four conditions are satisfied:

- No multicollinearity: no high correlation between the independent variables
- Linearity: there must be a linear relationship between the target variable (Life Expectancy) and the other variables
- Normality: the residuals must be normally distributed
- Homoscedasticity: the residuals must have a constant variance

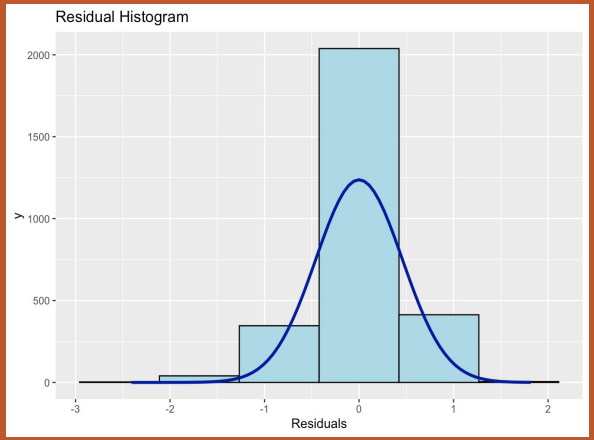
The first condition is already satisfied as we already removed the variables ‘Infant deaths’, ‘Under Five Deaths’, ‘GDP’ and ‘Thinness 1-19 years’, which are the variables that have a higher VIF value, so the ones with strong collinearity.





To check the other three conditions we start by building the linear model. To say that we have linearity the points should be evenly distributed between the two sides of the line and the red line should be approximately horizontal at zero and. This means that we can assume linear relationship between the predictors and the outcome variables.

To check the normality condition we can look at the Q-Q plot, we see that most of the points are located on the diagonal line, except the extremes, which deviate from the line, therefore, this Q-Q plot is inconclusive regarding the normality of the residuals.

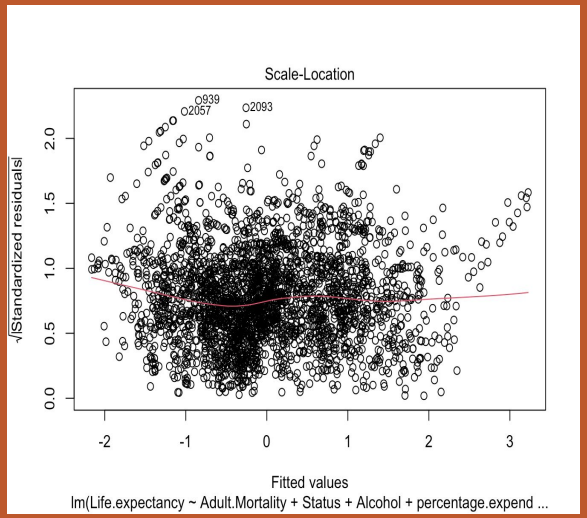



```
studentized Breusch-Pagan test

data:  model
BP = 188.95, df = 15, p-value < 2.2e-16
```

We need to find another way to check if the normality condition is met, let's try by plotting the histogram of the residuals. The histogram shows that most of the residuals fall around zero and the number of observations in the tails (so the extremes) of the histogram is low. We can conclude that residuals of our regression model follow a normal distribution.

To check the homoscedasticity assumption we can use the Breusch-Pagan test. As we can see, the p-value is smaller than 0.05 so there is evidence that the principle of Homoscedasticity is not fulfilled. When the principle of Homoscedasticity is not fulfilled, the estimate of the mean made by the model will continue to be good, but its confidence intervals will not.



This plot shows if residuals are spread equally along the ranges of predictors, so we need to get a line which is close to being horizontal with equally spread points, which is our case. This means that the Homoscedasticity is violated but not to a large extent.

lag	Autocorrelation	D-W Statistic	p-value
1	0.6455986	0.7066183	0

Alternative hypothesis: rho != 0

To check the independence of the residuals we use a Durbin-Watson test. Since, the p-value is less than 0.05, we reject the null hypothesis and conclude that the residuals are autocorrelated. Also, the value of D-W statistic is approximately 0.7, which is close to 0, showing high chances of high positive autocorrelation.

Feature selection

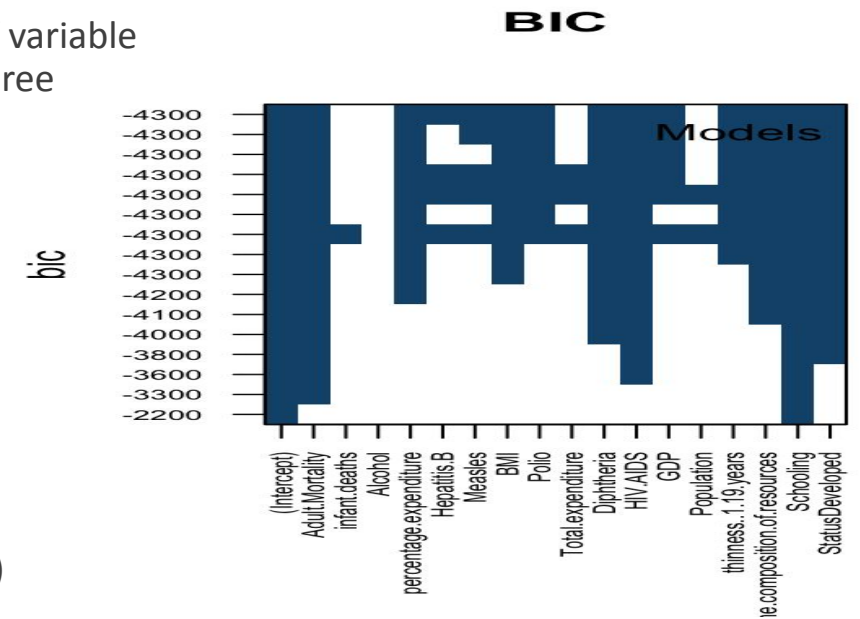
Feature selection is a crucial step in constructing models as it involves identifying a subset of relevant features from a larger dataset.

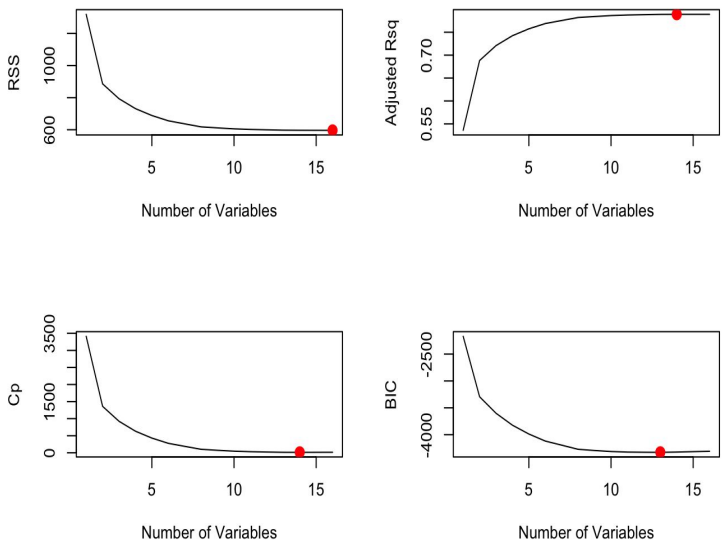
In this report, we explore the process of variable selection on a life expectancy dataset three methods:

- Forward Selection
- Backward Selection
- Mixed Selection

Additionally, we utilize four metrics:

- Residual sum of squares (RSS)
- Adjusted R^2
- Mallow's Cp
- Bayesian information criterion (BIC)



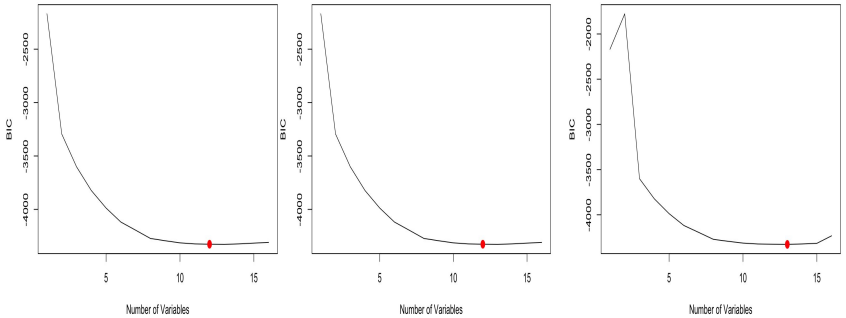


Based on these metrics, we have decided to prioritize the BIC as the criterion for selecting the best subset. This is because the BIC tends to penalize models with a larger number of variables more heavily.

Forward selection

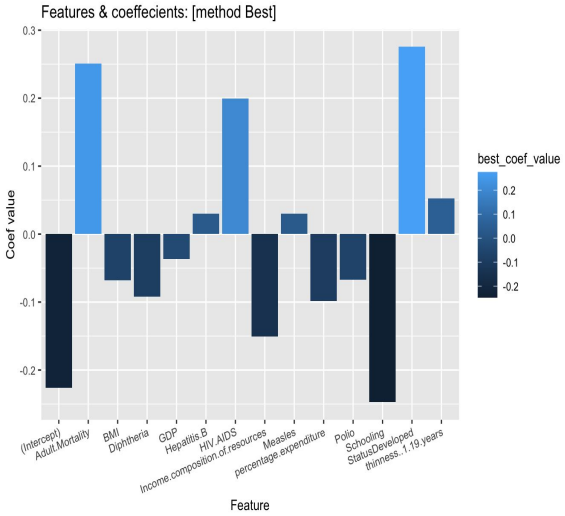
Backward selection

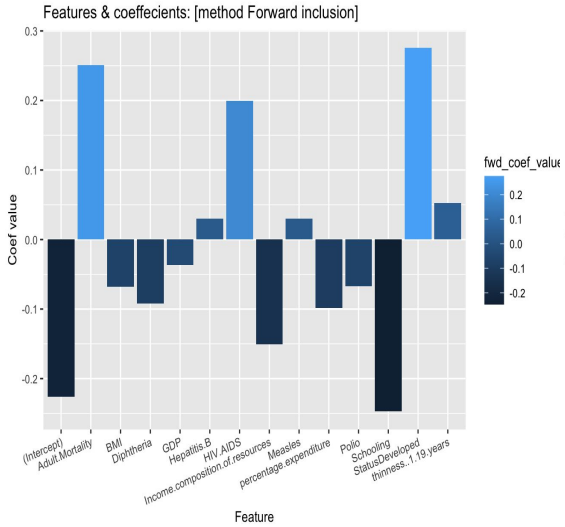
Mixed selection

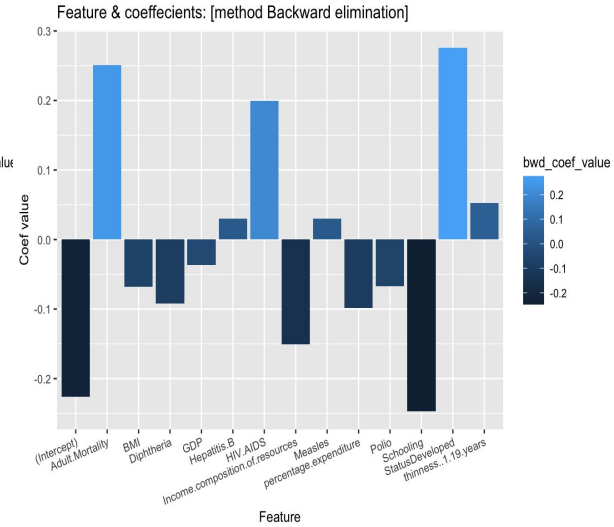


The feature selection analysis, incorporating forward, backward, and mixed techniques, consistently identifies 13 features as the optimal subset for building a predictive model.

By focusing on these 13 features, we can create a streamlined and efficient model that avoids unnecessary complexity and reduces the risk of overfitting. Additionally, we examine the selected variables and their coefficients for each method to gain further insights into the model's behavior.







Overall, when the forward, backward, and mixed methods yield the same coefficients for specific variables, it strengthens the evidence for the importance and reliability of those variables in predicting the target variable.

It provides consistency, stability, and confidence in the selected features, which are crucial for building effective and interpretable regression models.

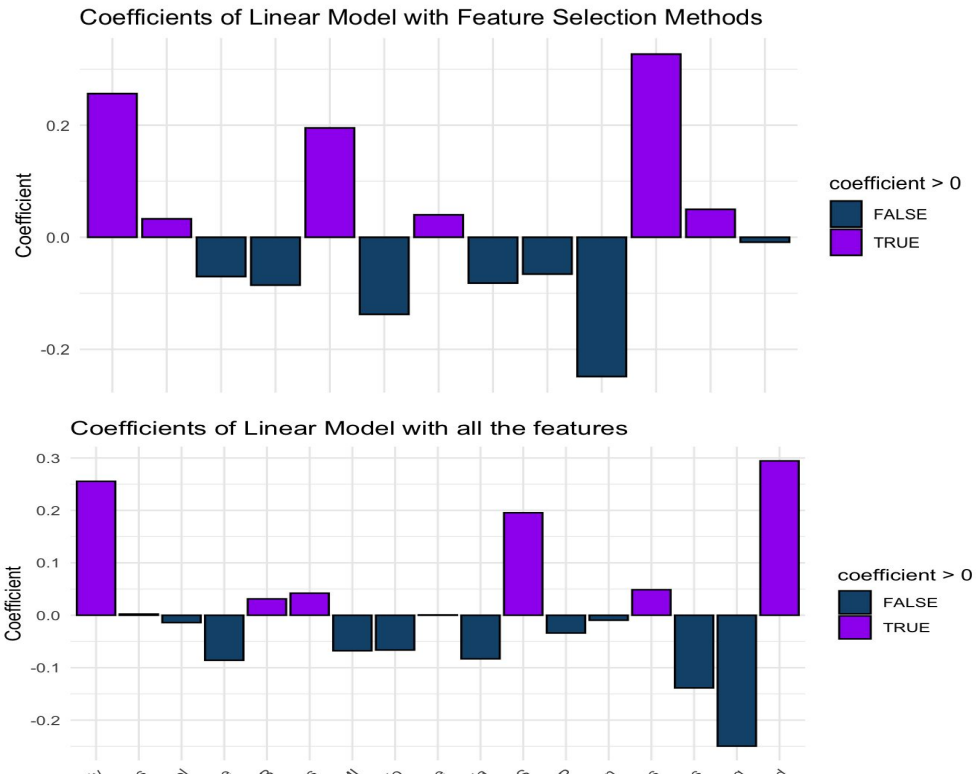
Model data

In this section we're going to apply three models to both the full dataset and the reduced dataset and then we're going to compare the results.

The models that we're going to apply are:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Polynomial Regression

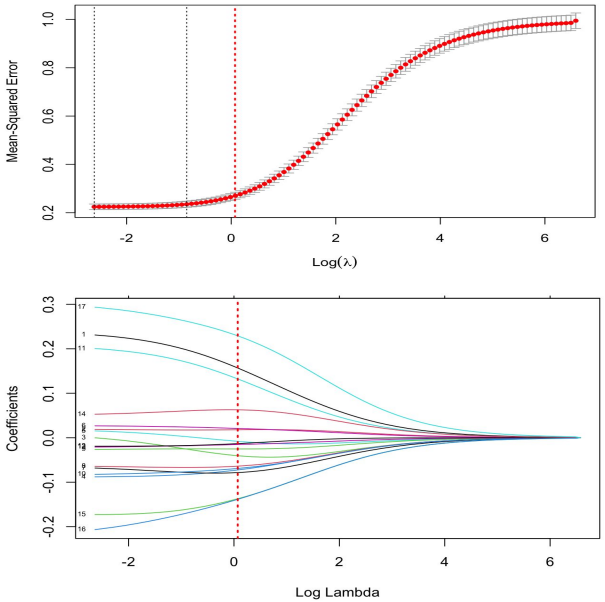
Simple linear model



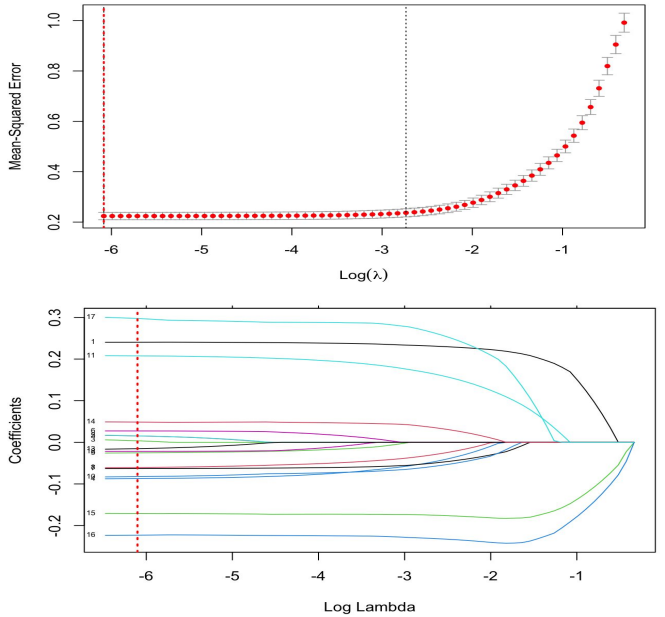
Model Description

For implementing the lasso regression and ridge regression model, we do the cross validation in order to find the best lambda to minimize the Mean Squared Error, also we use the shrinkage parameter in lasso model. here we have the plots of coefficients for different lambda:

Ridge Regression model

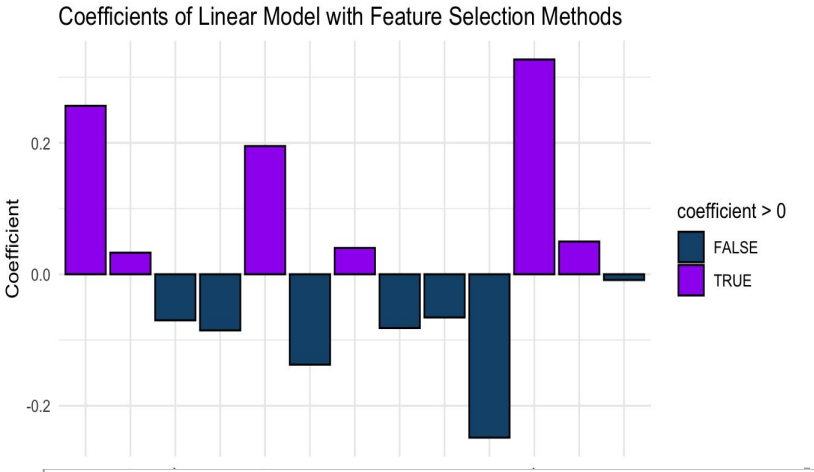


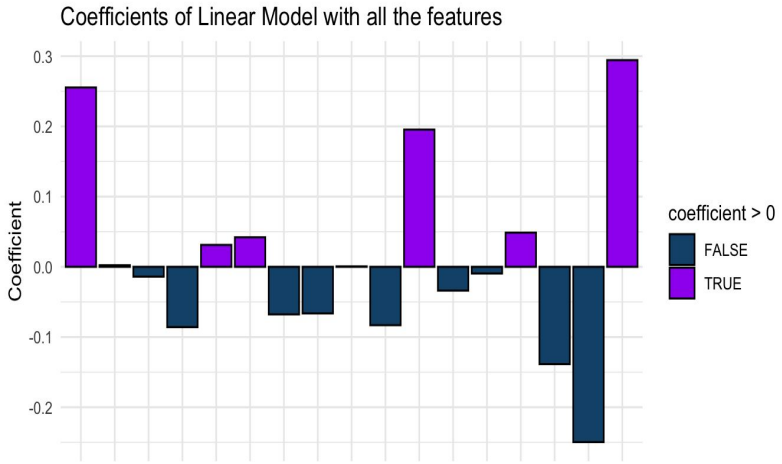
Lasso Regression model



Ridge regression model

The results indicate that there are no significant differences in variances between the two models. Therefore, we conclude that the variances observed are likely due to random variation.

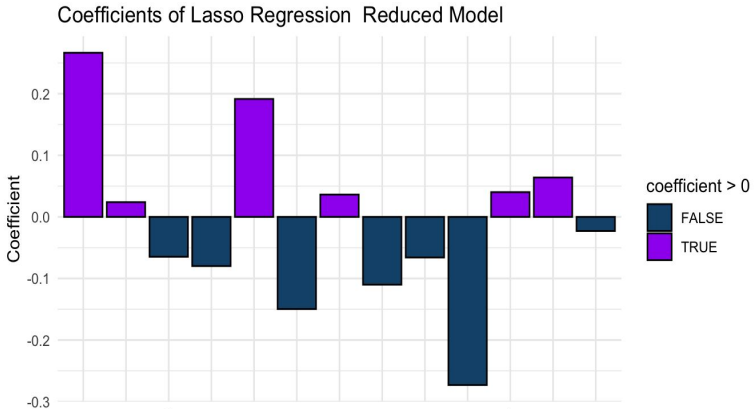


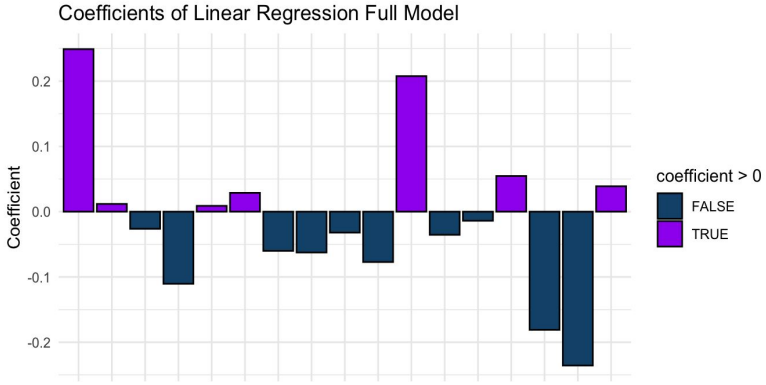


F-test to compare two variances	
P-value	**0.7041**
F-statistic	**1.026519**

Lasso regression model

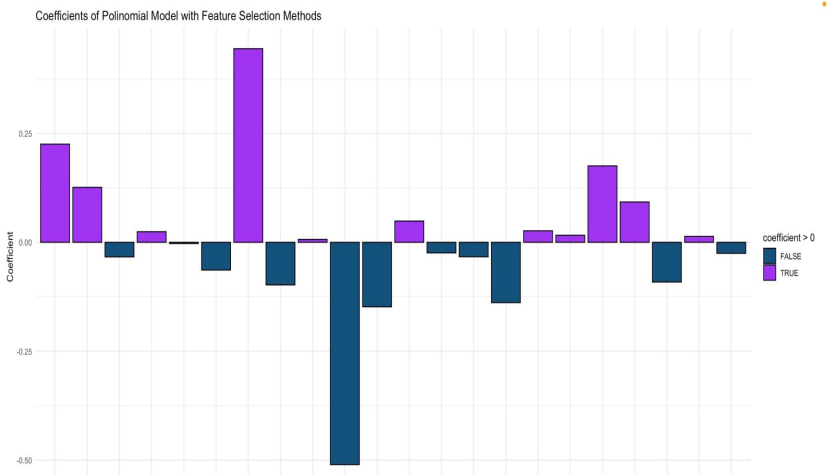
The results indicate that there are no significant differences in variances between the two models. Therefore, we conclude that the variances observed are likely due to random variation.



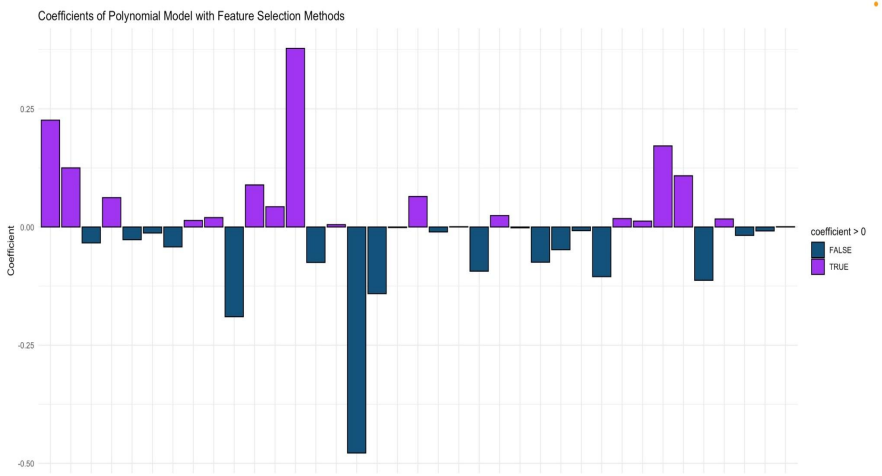


F-test to compare two variances	
P-value	**0.6365**
F-statistic	**1.0331**

Polynomial regression model



Polynomial Regression model on selected features with **polynomial degree 3** for every selected feature



Polynomial Regression model on selected features with **mixed polynomial degree**

Conclusions

Model	Type	Mean Squared Error	R-squared	Adjusted R-squared
Lasso Regression	Reduced Model	0.2139883	0.7881891	0.7848954
Lasso Regression	Full Model	0.2082843	0.7881891	0.7844604
Ridge Regression	Reduced Model	0.214421	0.7881891	0.7844604
Ridge Regression	Full Model	0.2100862	0.7958114	0.7915834
Simple Linear Model	Reduced Model	0.4600969	0.7877	0.7864
Simple Linear Model	Full Model	0.4601041	0.789	0.7864
Polynomial (D=3)	Reduced Model	0.4036827	0.8529	0.8501
Polynomial (D=3)	Full Model	0.4036827	0.8529	0.8501
Polynomial (Mixed)	Reduced Model	0.4112296	0.8471	0.8454
Polynomial (Mixed)	Full Model	0.4112296	0.8471	0.8454

Overall, the regression models (Lasso Regression and Ridge Regression) and the polynomial models demonstrate better performance in terms of MSE and the ability to explain the variance compared to the simple linear models. Both the Lasso and the Ridge models have relatively low mean squared error (MSE) values, indicating good predictive performance.

The models with polynomial terms (Degree 3 and Mixed degrees) generally perform better than the Simple Linear Model in terms of MSE, R-squared, and adjusted R-squared. Among the polynomial models, the Reduced Model has slightly lower complexity but performs similarly to the Full Model, indicating that the additional polynomial terms in the Full Model may not significantly improve the performance. A larger F-statistic with a smaller p-value in Polynomial model compared to Simple linear model indicate that the model is statistically significant and provides a better fit to the data.

Based on the given results, the Lasso Regression model on the whole dataset (Full Model) appears to be the better choice among the three models for predicting the target variable.

It is worth noting that the feature importance varied between the models, has some common features such as:

- Status
- Adult.Mortality
- percentage.expenditure
- Hepatitis.B
- Measles
- BMI
- thinness.5.9.years
- Polio
- Diphtheria
- HIV.AIDS
- Income.composition.of.resources
- Schooling

We can conclude that the Feature selection Model provides good results given that it decreases the number of features.

We also came to the following conclusions:

- Education has a significant impact on life expectancy
- Life Expectancy have negative relationship with drinking alcohol
- Immunization against Hepatitis B and Diphtheria positively impact on life Expectancy
- Countries with higher income composition of resources for human development have a better life expectancy.
- There is no significant difference in proportions of the number of infant deaths and the number of under-five deaths.
- There is no strong correlation between alcohol consumption and life expectancy
- Most frequent range for life expectancy is 65–82 Years and the least frequent range is less than 45 years and more than 85 years.
- Immunization coverage has a significant impact on life expectancy.
- Population doesn’t have big impact on life expectancy