Final Project

Shuxin Tan 1007625447

2023-08-10

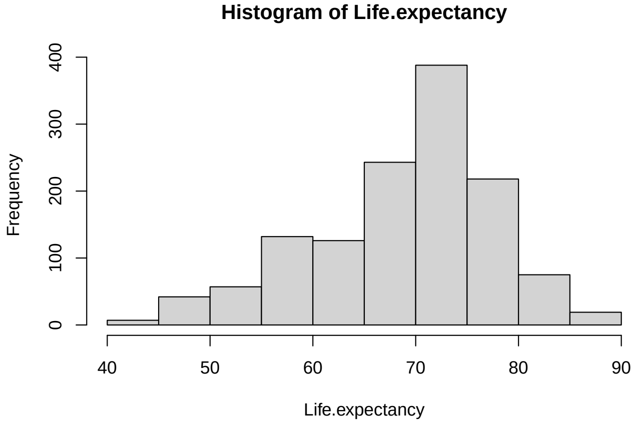
1. Introduction:

This report dissects the intricate influences shaping human life expectancy, accentuated by the current era's rapid technological and economic evolution. As individuals prioritize health and nations strategize to extend citizens' lifespans, a comprehensive exploration of factors impacting longevity becomes crucial. By delving into life expectancy determinants, this study equips nations with insights to implement effective measures for citizens' health and longevity. The dataset, spanning 2000 to 2015 and sourced from WHO and the United Nations, underpins the analysis. A linear regression model is developed, aiming for a generalized understanding beyond specific countries and years. Before embarking on model development, a comprehensive data exploration journey is undertaken. This phase yields invaluable insights into the dataset's characteristics, facilitating the identification of anomalies. Data cleaning and transformations are meticulously executed to ensure data integrity and robustness. Given the richness of the dataset with numerous variables potentially influencing life expectancy, a judicious analysis guides the selection of variables for the model. Rigorous criteria, encompassing statistical significance and theoretical relevance, dictate the inclusion of specific variables. An iterative process ensues, wherein diverse model statistics are compared to ascertain optimal fit. The model's efficacy is subjected to rigorous evaluation through validating the assumptions it rests upon. Subsequently, the model's outcomes are contextualized to real-world scenarios, enriching our comprehension of the intricate relationship between life expectancy and the selected variables. The model's results unveil crucial life expectancy factors, fostering insights for applications and informed decision-making, which would inform critical decisions and shape a healthier future for human beings.

2. Exploratory Data Analysis Section:

The dataset initially comprises 22 variables, spanning Country, Year, Status, Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 10-19 years, thinness 5-9 years, Income composition of resources, Schooling, and Life expectancy. With a focus on unraveling the intricate influences shaping human life expectancy, we prioritize 19 predictor variables: Status, Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 10-19 years, thinness 5-9 years, Income composition of resources, and Schooling. Throughout the analysis, we omit country-specific and year-specific effects, with Life expectancy as the central response variable. We first drop observation with missing values such that the dataset becomes complete. The following is a detailed explanation of each variable.

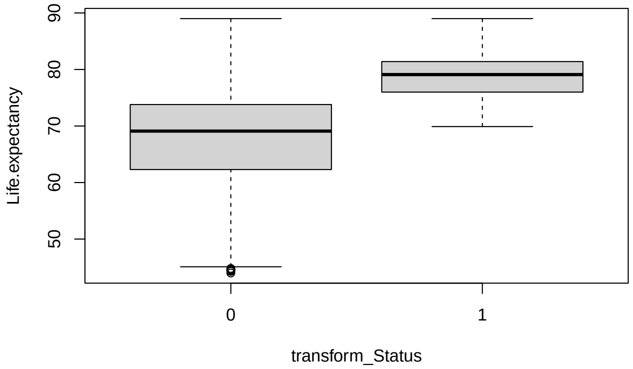
Life expectancy: The country's average Life expectancy in age



According to the above graph, we can see that the distribution of the response variable Life.expectancy is approximately normal distributed and it is continuous, being a little bit left skewed. It has a maximum of 89 and a minimum of 44.

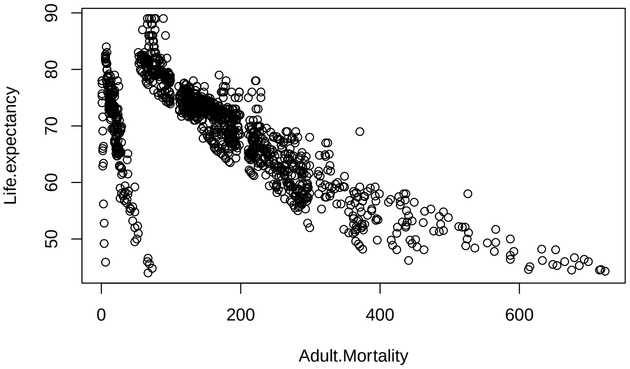
#1 Status: Whether the country is Developed or Developing

Since the character doesn't make sense in statistics, we transform the variable Status into numerical values such that if the Status is Developing, then it's 0, otherwise it's 1 which means if the Status is Developed, it's 1. And use the boxplot to discern its relationship with the response variable.

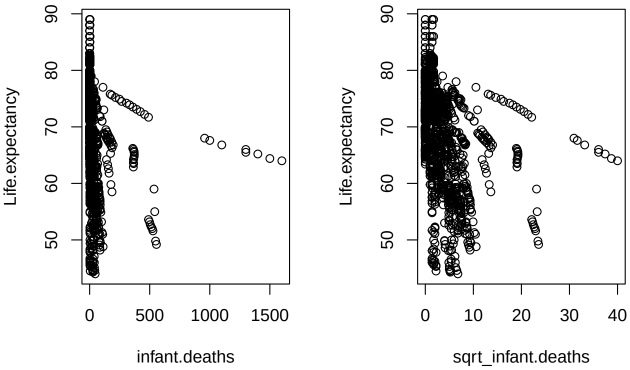


The graphical analysis highlights a discernible linear relationship between the variable in question and the continuous response variable, "Life.expectancy." Specifically, the "Developed" status appears associated with an extended life expectancy. However, a notable observation emerges: within the "Developing" status category, several outliers lie below the minimal value of the corresponding response variable. Furthermore, the median life expectancy for the "Developing" status falls below that of the "Developed" status. Noteworthy is the disparity in box heights—the box representing the "Developing" status surpasses that of the "Developed" status. This discrepancy suggests that fluctuations within the "Developing" status exhibit greater variability compared to the more consistent pattern observed in the "Developed" status.

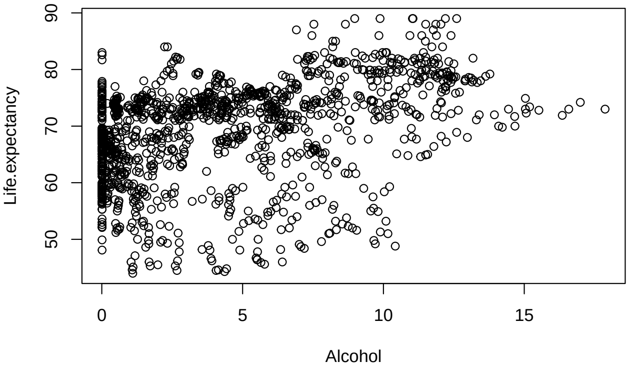
#2 Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)



According to the above scatterplot, the relationship between Adult.Mortality and Life.expectancy is very linear but with several outliers on the left. Also, we can see the variable Adult.Mortality is continuous from the plot. The Adult.Mortality is fairly even distributed with the Minimum of 1 and the Maximum of 699.
#3 infant deaths: Number of Infant Deaths per 1000 population

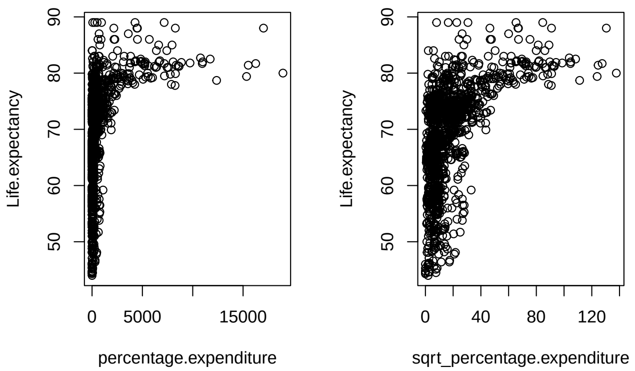


The scatterplot above shows the linearity between infant.deaths and Life.expectancy is very weak as most of the plot concentrating around 0 on the left part of the graph. Since the value of infant.deaths of some observations includes 0, it is better avoid taking logarithm transformation. However, after being imposed a square root transformation on the value of variable infant.deaths, the linearity between the predictor and the response variable Life.expectancy is still weak with most dat centering on the left part of the graph. Thus, it's better to drop this variable for our linear regression model.
#4 Alcohol: Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)



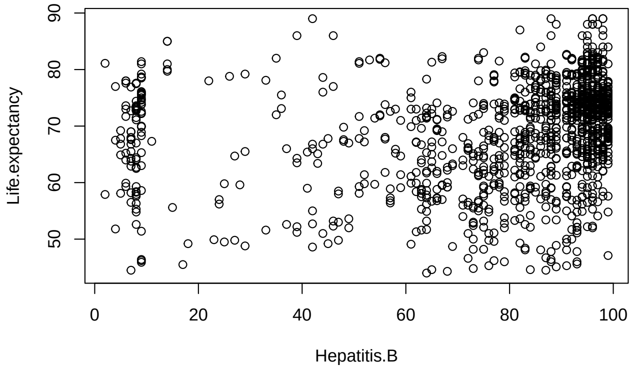
The scatterplot above shows the linearity between Alcohol and Life.expectancy is moderate as a linear trend would be seen in the middle. Also, it shows Alcohol is a continuous variable, distributing fairly evenly between the Maximum of 17.87 and the Minimum of 0.01. The correlation between Alcohol and Life.expectancy is around 0.41 which means the linearity with response variable should be considered relatively strong.

#5 percentage expenditure: Expenditure on health as a percentage of GDP per capita (%)



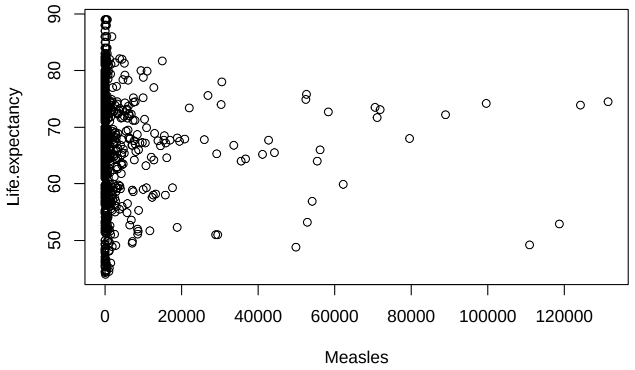
The scatterplot above shows the linearity between percentage.expenditure and Life.expectancy is very weak as it shows a more logarithm relationship. However, since the value of infant.deaths of some observations includes 0, it is better avoid taking logarithm transformation to make our model more meaningful. However, after being imposed a square root transformation on the value of variable infant.deaths, the predictor variable sqrt_percentage.expenditure still shows a logarithm relationship with the response variable Life.expectancy. Thus, it's better to drop this variable for our linear regression model.

#6 Hepatitis B: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)



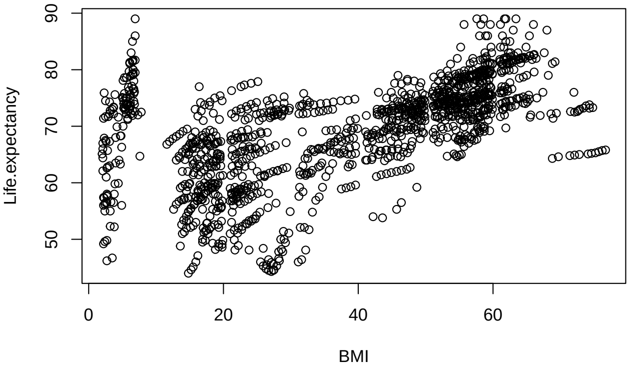
The scatterplot above shows the linearity between Hepatitis.B and Life.expectancy is very weak as most of the plots seems distributed randomly. This holds even after imposing several simple transformation methods on the value of variable Hepatitis.B. Thus, this variable will be dropped for our linear regression model.

#7 Measles: Measles - number of reported cases per 1000 population



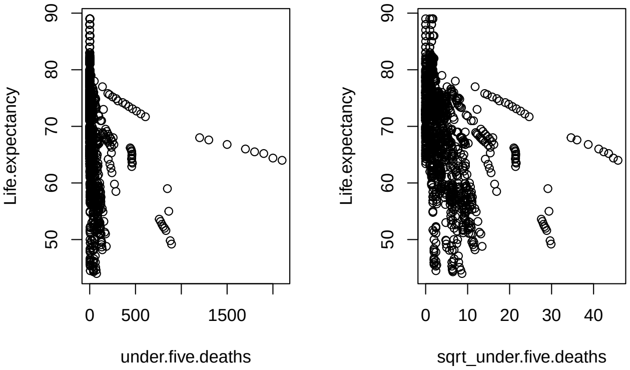
The scatterplot above shows the linearity between Measles and Life.expectancy is very weak as most of the plots are concentrating around 0 on the left part of the graph. This holds even after imposing several simple transformation methods on the value of variable Measles. Thus, this variable will be dropped for our linear regression model.

#8 BMI: Average Body Mass Index of entire population



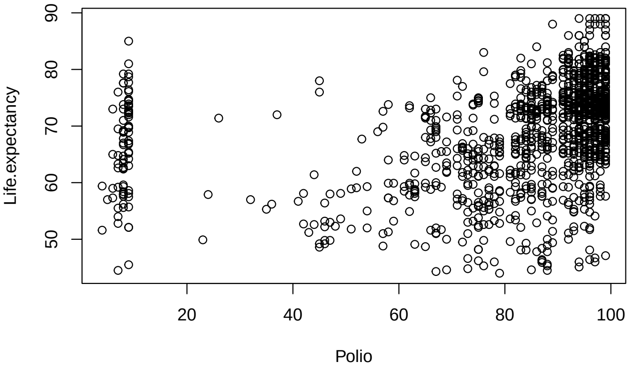
According to the above scatterplot, the relationship between BMI and Life.expectancy is very linear but with several outliers on the left which should be considered as a limitation in the later part. Also, we can see that the variable BMI is continuous. The BMI is fairly even distributed with a Minimum of 2 and a Maximum of 76.7.

#9 under-five deaths: Number of under-five deaths per 1000 population

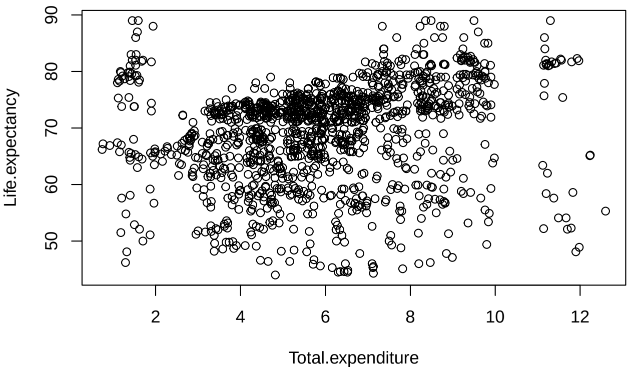


The scatterplot above shows the linearity between under.five.deaths and Life.expectancy is very weak as as most plots are concentrating around 0 on the left part of the graph. However, since the value of infant.deaths of some observations includes 0, it is better avoid taking logarithm transformation to make our model more meaningful. So, after being imposed a square root transformation on the value of variable under.five.deaths, the linearity is still very weak. So, it s better if this variable is dropped for the model.

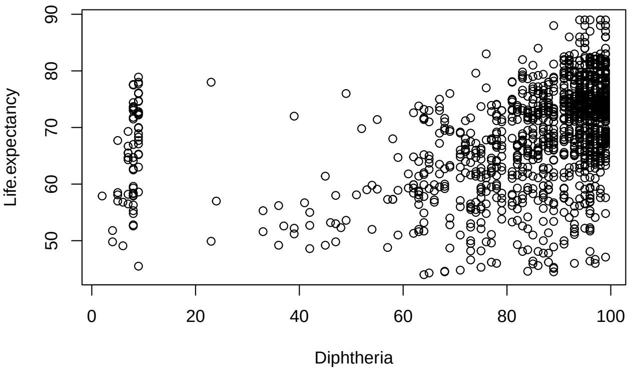
#10 Polio: Polio (Pol3) immunization coverage among 1-year-olds (%)



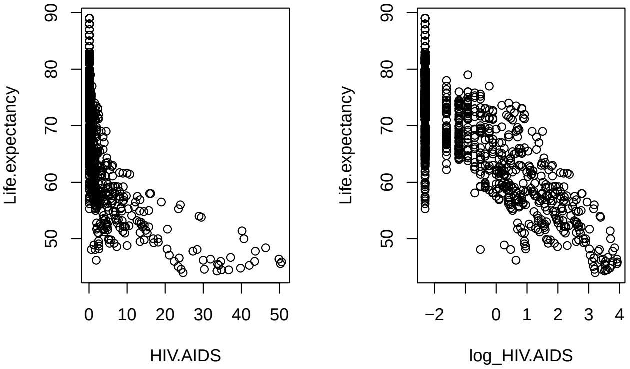
The scatterplot above shows the linearity between Polio and Life.expectancy is very weak as most of the plots concentrating around 0 and around 100 and the correlation between Polio and Life.expectancy is around 0.3. This holds even after imposing several simple transformation methods on the value of variable Measles. Thus, this variable will be dropped for our linear regression model.
#11 Total expenditure: Government expenditure on health as a percent of its total expenditure (%)



The scatterplot above shows the linearity between Total.expenditure and Life.expectancy is very weak as the plots seems to show a flat trend, distributing randomly, and the correlation between Polio and Life.expectancy is around 0.19. This holds even after imposing several simple transformation methods on the value of variable Measles. Thus, this variable will be dropped for our linear regression model.
#12 Diphtheria: Diphtheria tetanus toxoid and pertussis immunization coverage among 1-year-olds (%)

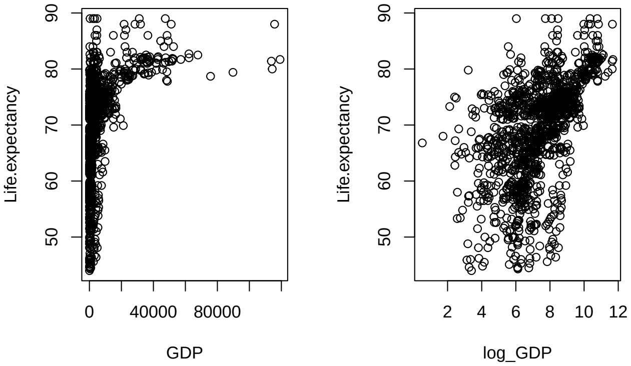


The scatterplot above shows the linearity between Diphtheria and Life.expectancy is very weak as most of the plots concentrating around 0 and around 100 and the correlation between Diphtheria and Life.expectancy is around 0.32. This holds even after imposing several simple transformation methods on the value of variable Measles. Thus, this variable will be dropped for our linear regression model.
#13 HIV/AIDS: Deaths per 1000 live births HIV/AIDS (0-4 years)



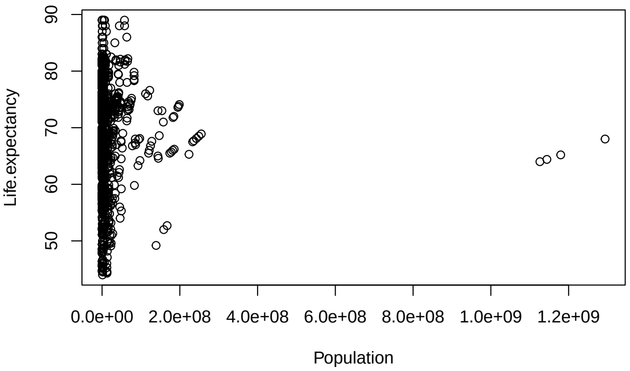
The scatterplot above shows there a more logarithm relationship between HIV.AIDS and Life.expectancy. Since the value of infant.deaths doesn't include 0, we could take a logarithm transformation to make it more linear. After being imposed a logarithm transformation on the value of variable HIV.AIDS, the predictor variable log_HIV.AIDS shows a more linear relationship with the response variable Life.expectancy with a correlation around -0.8, which is also continuous with the Maximum of 3.92 and Minimum of -2.3.

#14 GDP: Gross Domestic Product per capita (in USD)



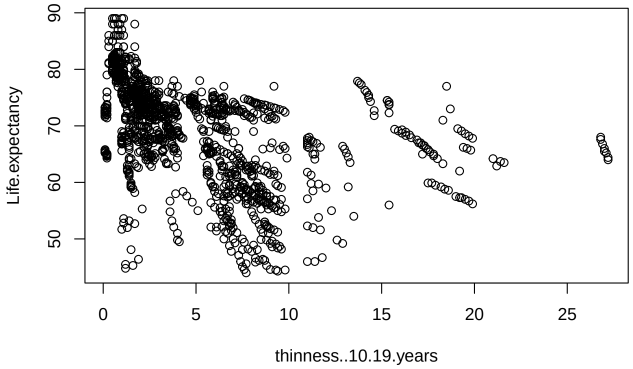
The scatterplot above shows there a more logarithm relationship between GDP and Life.expectancy. Since the value of GDP doesn't include 0, we could take a logarithm transformation to make it more linear. After being imposed a logarithm transformation on the value of variable GDP, the predictor variable log_GDP shows a more linear relationship with the response variable Life.expectancy with a correlation around 0.54, which is also continuous with the Maximum of 11.69 and Minimum of 1.74.

#15 Population: Population of the country

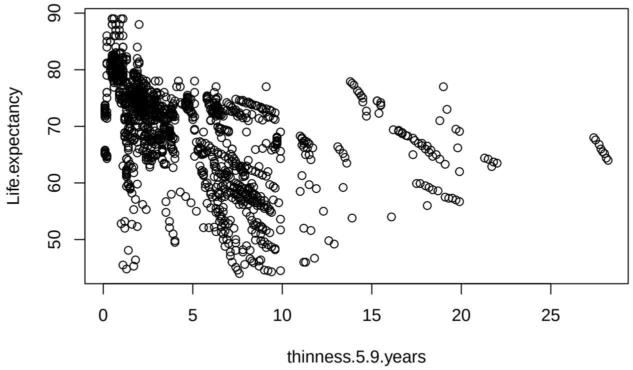


The scatterplot above shows the linearity between Population and Life.expectancy is very weak as most of the plots concentrating around 0 on the left part of the graph. This holds even after imposing several simple transformation methods on the value of variable Measles. Thus, this variable will be dropped for our linear

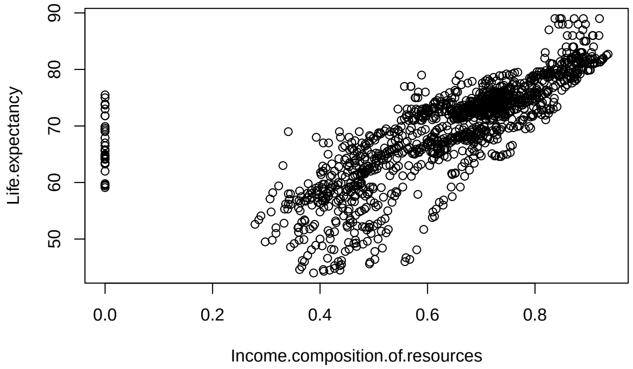
regression model.
#16 thinness 10-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)



According to the above scatterplot, the relationship between thinness..10.19.years and Life.expectancy is very linear but with several outliers which should be considered as a limitation in the later part. Also, we can see that the variable thinness..10.19.years is continuous. And it is fairly even distributed with a Minimum of 0.1 and a Maximum of 27.2.
#17 thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9 (%)

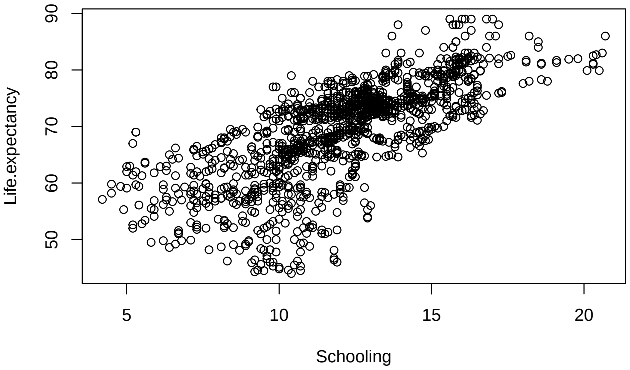


According to the above scatterplot, the relationship between thinness.5.9.years and Life.expectancy is very linear but with several outliers which should be considered as a limitation in the later part. Also, we can see that the variable thinness.5.9.years is continuous. And it is fairly even distributed with a Minimum of 0.1 and a Maximum of 28.2.
#18 Income composition of resources: Income composition of resources



According to the above scatterplot, the relationship between Income.composition.of.resources and Life.expectancy is very linear but with several outliers around 0. Also, we can see that the variable

Income.composition.of.resources is continuous. And it is fairly even distributed with a Minimum of 0 and a Maximum of 0.94.
#19 Schooling: Number of years of Schooling (years)



According to the above scatterplot, the relationship between Schooling and Life.expectancy is very linear. Also, we can see that the variable Income.composition.of.resources is continuous. And it is fairly even distributed with a Minimum of 4.5 and a Maximum of 20.7.

Therefore, after a careful data exploration and necessary data transformations, it's better if we proceed with predictor variables including transform_Status, Adult.Mortality, Alcohol, BMI, log_HIV.AIDS, log_GDP, thinness..10.19.years, thinness.5.9.years, Income.composition.of.resources, and Schooling for the model development section.

3. Model Development Section:
First, a subset of our training dataset is created according to the variables selected in the Exploratory Data Analysis Section and is from the training dataset which is a randomly chosen 80% of the original dataset. The step function can automatically perform stepwise model selection by adding or removing variables by comparing criterion such as AIC as BIC to improve the model's fit. Using the both method which is a combination of backward and forward selection strikes a balance between adding informative variables and removing non-informative ones, being both efficient and avoid overfitting at the same time. The both method starts with an empty model and then alternates between adding and removing predictor variables in each step. Therefore, from the summary of the model using the both stepwise method in R, variables of transform_Status, Adult.Mortality, transform_Status, log_HIV.AIDS, thinness.5.9.years, log_GDP, Income.composition.of.resources, and Schooling, are selected to refine the model, making it become more fitted with the observations and be more statistically significant.

Then, it is necessary to check if the Variance Inflation Factor (VIF) to identify the degree of correlation is high between predictor variables, because if the degree of correlation is high between predictor variables, it can cause multicollinearity problems between predictor variables when fitting and interpreting the regression model.

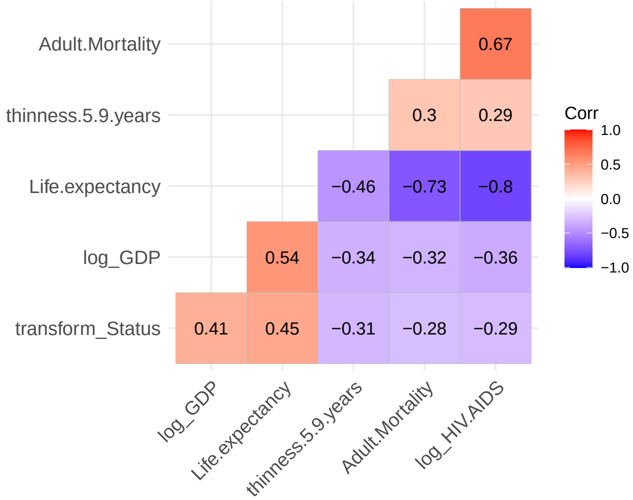
```
## [1] "vif"

##                    log_HIV.AIDS Income.composition.of.resources
##                        2.052980                        2.990713
##                 Adult.Mortality                       Schooling
##                        1.931999                        3.340660
##              thinness.5.9.years                         log_GDP
##                        1.341096                        1.644179
##                transform_Status
##                        1.420320
```

According to the vif table, the vif for log_HIV.AIDS and Income.composition.of.resources are relatively high which are 2.006101 and 2.902301, indicating there is a correlation between log_HIV.AIDS and Income.composition.of.resources. And the vif for Schooling is 3.325976 which is regarded to be high in our initial model, implying there is a moderate correlation between Schooling and other predictor variables. Therefore, it would be better if we drop the predictor variable of Schooling and choose one between log_HIV.AIDS and Income.composition.of.resources to alleviate the problem of multicollinearity, making the predictor variables more independent. After removing the the predictor variable of Schooling, we should decided whether we should include log_HIV.AIDS or Income.composition.of.resources based on various metrics.

```
##                                        R Squared          AIC          BIC
## Drop log_HIV.AIDS                      0.7520516 7604.1831803 7640.4116083
## Drop Income.composition.of.resources   0.7997061 7325.2235999 7361.4520279
```

Then, we use the lm method for both cases to generate models. Calculating R Squared, AIC, and BIC for both models, we could discover that the R Squared when dropping Income.composition.of.resources is higher than the R Squared when dropping log_HIV.AIDS. AIC and BIC are both smaller when dropping Income.composition.of.resources compared with dropping log_HIV.AIDS from the above table. Therefore, as a larger R Squared means a larger proportion of the variance in the response variable is explained by the predictors and a smaller AIC and BIC implies a better balance between model fit and complexity, it is better to drop Income.composition.of.resources and choose log_HIV.AIDS to make the linear regression model more fit. Thus, we proceed with predictors of transform_Status, Adult.Mortality, log_HIV.AIDS, log_GDP, and thinness.5.9.years, with response variable being Life.expectancy.

Then, we further refine our model by checking with the correlation matrix to further assess colinearity between predictors.



From the above visualized correlation matrix, the correlation between Adult.Mortality and log_HIV.AIDS is around 0.65 which is relatively high, suggesting a possibility of dependent predictors. Thus, use carious metrics to decide whether to drop Adult.Mortality or log_HIV.AIDS.

```
##                         R Squared          AIC          BIC
## Drop Adult.Mortality    0.7575853 7572.6830464 7603.7359847
## Drop log_HIV.AIDS       0.6857763 7911.7913690 7942.8443073
```

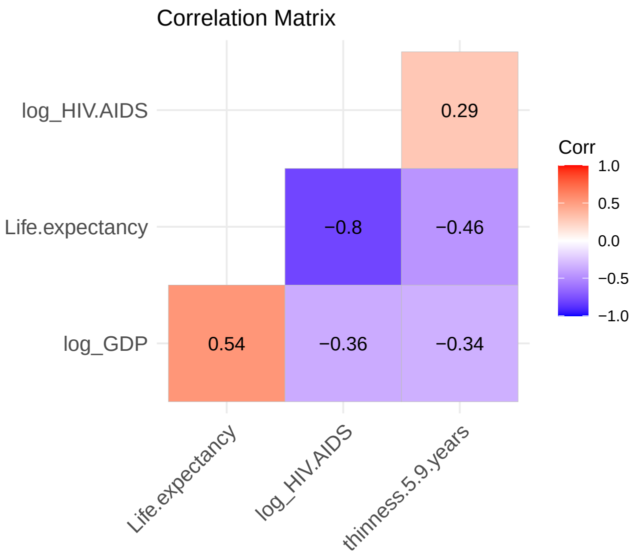
Then, we use the lm method for both cases to generate models. Calculating R Squared, AIC, and BIC for both models, we could discover that the R Squared when dropping Adult.Mortality is higher than the R Squared when dropping log_HIV.AIDS. AIC and BIC are both smaller when dropping Adult.Mortality compared with dropping log_HIV.AIDS from the above table. Therefore, as a larger R Squared means a larger proportion of the variance in the response variable is explained by the predictors and a smaller AIC and BIC implies a better balance between model fit and complexity, it is better to drop Adult.Mortality and choose log_HIV.AIDS to make the linear regression model more fit. Thus, we proceed with predictors of transform_Status, log_HIV.AIDS, log_GDP, and thinness.5.9.years with response variable being Life.expectancy.

Since transform_Status is a categorical variable, it is hard to check the multicolinearity between it and other predictors. However, since the standard criteria for determining whether a country is developed or developing is siginificantly based on the country's GDP, then by the nature of log_GDP and transform_Status, there exists a high correlation between these 2 predictor variables. So, we need to determine whether to drop transform_Status or log_GDP.

```
##                           R Squared          AIC          BIC
## Drop transform_Status     0.7445139 7639.3243254 7665.2017739
## Drop log_GDP              0.6382596 8093.8452077 8119.7226562
```

Then, we use the lm method for both cases to generate models. Calculating R Squared, AIC, and BIC for both models, we could discover that the R Squared when dropping transform_Status is higher than the R Squared when dropping log_GDP. AIC and BIC are both smaller when dropping transform_Status compared with dropping log_GDP from the above table. Therefore, as a larger R Squared means a larger proportion of the variance in the response variable is explained by the predictors and a smaller AIC and BIC implies a better balance between model fit and complexity, it is better to drop transform_Status and choose log_GDP to make the linear regression model more fit. Thus, we proceed with predictors of log_HIV.AIDS, log_GDP, and thinness.5.9.years with response variable being Life.expectancy.

```
## [1] "Correlation Matrix for the Developed Model"

##                     log_HIV.AIDS    log_GDP thinness.5.9.years Life.expectancy
## log_HIV.AIDS          1.0000000 -0.3647440          0.2887156      -0.8007067
## log_GDP              -0.3647440  1.0000000         -0.3391897       0.5408353
## thinness.5.9.years    0.2887156 -0.3391897          1.0000000      -0.4635397
## Life.expectancy      -0.8007067  0.5408353         -0.4635397       1.0000000
```



```
## [1] "vif"
```

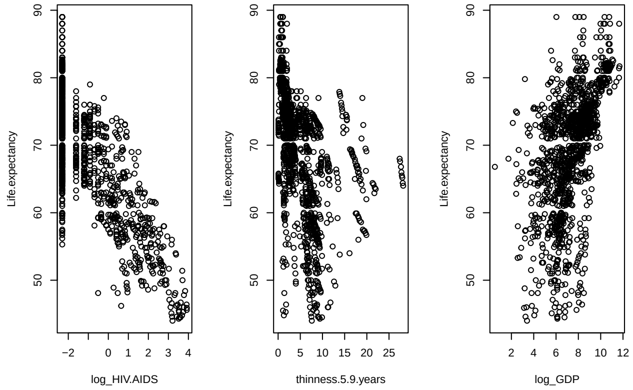
```
##       log_HIV.AIDS            log_GDP thinness.5.9.years
##           1.195889           1.238718           1.171580
```

Checking the visualized correlation matrix for the refined predictors and the vif for the refined model again, we could see that all the correlation between predictors are under 0.4 and all the vif for the model is below 1.5, suggesting the problem of multicolinearity is small enough to neglect. Also, the mean VIF for the model is 1.25191, which is not considerably larger than 1. Thus, it is positive for the model, meaning they are independent to each other. Overall, there is no need for us to drop more predictors.

Hence, the linear regression model has been established based on the provided training dataset.
The final best model: $\hat{Y} = 57.56895 - 3.62363X_1 + 1.23152X_2 - 0.34059X_3$, where
$\hat{Y}$: Life.expectancy
X_1 : log_HIV.AIDS
X_2 : log_GDP
X_3 : thinness.5.9.years
and the actual meaning in the real world will be discussed in detail in the Conclusion Section.

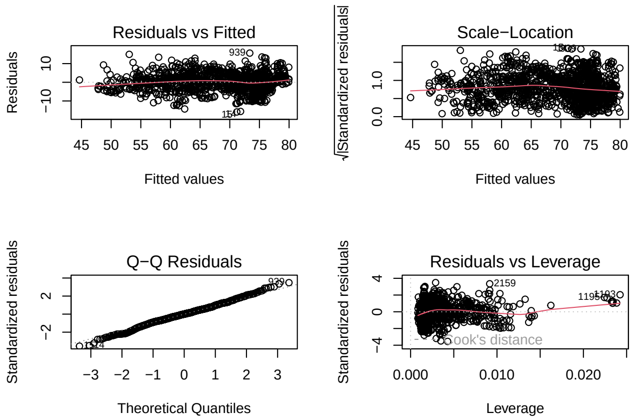
Model Validation:
Subsequently, it becomes imperative to validate this model against the four fundamental assumptions of multiple linear regression. Ensuring the model's reliability involves scrutinizing its goodness by confirming these assumptions: linearity between predictors and the response variable, independence of errors, consistent variance of errors (homoscedasticity), and normality of errors. Evaluating metrics like R-squared and adjusted R-squared, alongside other relevant measures, aids in comprehending the model's fitness. Moreover, gauging the model's effectiveness entails employing an independent test dataset to quantify the disparity between actual and predicted values—a direct approach to assessing the model's practical utility.

First, we start with checking the 4 assumptions of a linear regression model.
(1) Linearity between predictors and the response variable



Examining the scatterplots, a clear observation arises: the four predictor variables—log_HIV.AIDS, thinness.5.9.years, and log_GDP—strongly exhibit linear relationships with the response variable, Life.expectancy. This visual analysis solidifies their inclusion in the linear regression model. However, it's crucial to note that while most scatterplots adhere to the linear pattern, some instances reveal outliers, particularly in thinness.5.9.years. These outliers, situated at the edges of scatterplots, deviate significantly from the linear trend. Acknowledging these outliers is essential as they could influence the model's accuracy and interpretation.

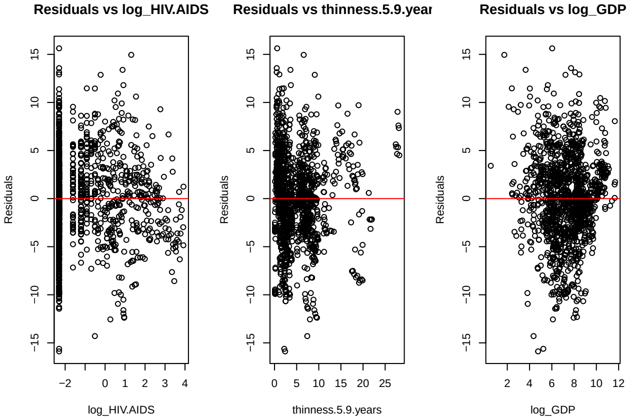
(2) Independence of the error



In the initial scatterplot, residuals exhibit a uniform, random distribution around the central axis (0 line) on the y-axis. This consistency suggests the absence of relationships between residuals and specific variables, confirming error independence. The near-zero correlation coefficient (-1.694763e-16) further supports this notion, validating the assumption of error independence. The second Scale-Location plot portrays a linear trend with scattered points, indicating minimal heteroscedasticity—a consistent spread of residuals across predicted values. This adherence to the pattern implies that the assumption of constant variance (homoscedasticity) is met. The third Normal QQ Residuals plot showcases close alignment with the reference line, indicating a near-normal distribution of residuals. This reinforces the assumption of normality, vital for reliable inference. Lastly, the graph depicting Cook's distance values indicates all observations fall within a reasonable range, suggesting low influence on the model. This implies that individual data points do not disproportionately impact the model's fit and predictions.

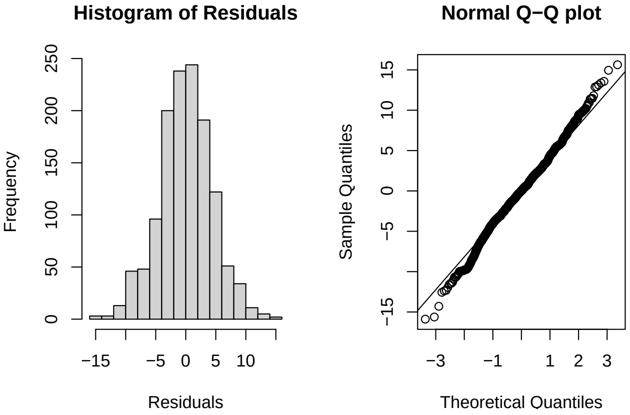
(3) Constant variance of the errors (homoscedasticity)
For the third assumption Constant variance of the errors (homoscedasticity), it uses the same plot as the second assumption does. The scatterlpot between residuals and the fitted values seems relatively more concentrating on the right part of the graph, which is suspicous that there may exist a relationship between residuals and the fitted values for the larger value of the fits. Thus, we should dig deeper into it by plotting

the scatterplot between residuals and each predictor.



To test the assumption of constant error variance (homoscedasticity), we examine scatterplots of residuals versus fitted values. These plots unveil whether the spread of residuals remains consistent across the x-axis range. The scatterplots reveal interesting findings. Residuals plotted against log_HIV.AIDS and log_GDP show a scattered yet stable distribution around the center, supporting homoscedasticity. However, a distinct megaphone-like trend emerges in the thinness.5.9.years scatterplot, indicating potential variance inconsistency. This pattern warrants attention, as it deviates from the assumption. Notably, the variance inconsistency within thinness.5.9.years highlights a limitation affecting model robustness and accuracy. This trend underscores homoscedasticity's importance, emphasizing the need for consistent variance across predictors, a limitation discussed in the Conclusion Section.

(4) Normality of the errors



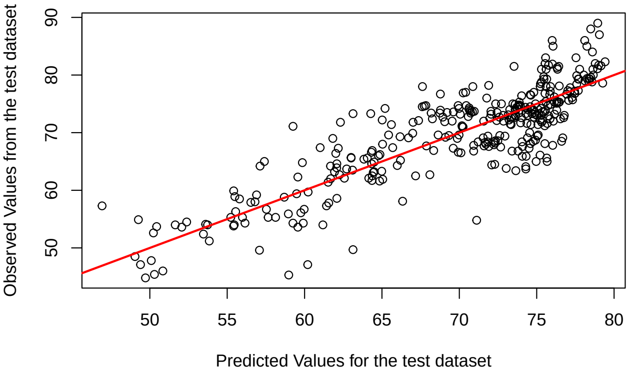
Evaluating normality in error assumptions involves assessing the histogram and normal probability plot of residuals. The histogram approximates a normal curve, while the normal probability plot shows a subtle S-shape deviation from the qqline. This suggests a near-normal distribution with nuanced deviations, especially for lower and higher predictor values which seem to be outliers. While such deviations are expected and can indicate balance, they prompt cautious consideration. These discrepancies are not necessarily model flaws, but acknowledging them is important for a comprehensive analysis. Recognizing these nuances becomes vital as we move toward the latter stages of our assessment. While not implying model inadequacy, these deviations are a limitation to address in our conclusion, highlighting the need for a nuanced perspective beyond assumptions.

Having meticulously scrutinized and validated the four foundational assumptions, the subsequent step involves a comprehensive evaluation of diverse metrics aimed at gauging the model's adeptness in aligning with our dataset.

```
##   R.Squared Adjusted.R.Squared
## 1 0.7445139          0.7439257
```

The R Squared for this model is 0.7414427 and the adjusted R Squared is 0.7408533. An R-squared value of 0.7414427 indicates that approximately 75% of the variability in the response variable (dependent variable) can be explained by the predictor variables (independent variables) included in the model, suggesting a moderate level of fit between the model and the data. An adjusted R Squared of 0.7408533. An adjusted R-squared of approximately 0.7408533 indicates that the model is explaining about 74.08% of the variability in the response variable, considering the number of predictor variables and their respective contributions, which provides a more conservative assessment of the model's fit and helps prevent overfitting by penalizing the inclusion of unnecessary predictor variables. So, the developed model has explained a moderate level of the repsonse variable Life.expectancy. And the summary of the final model shows every predictor is significant with a p-value of between 0 and 0.001, suggesting that the variable is likely to have a significant impact on the response variable Life.expectancy.

Following this, we will employ a test dataset comprising 20% randomly selected instances from the original dataset to assess the model's predictive capability. This comprehensive evaluation ensures the alignment of our model with fundamental assumptions, while also highlighting its practical effectiveness in real-world scenarios. This dual validation reaffirms the model's dependability and its applicability to real-world contexts.



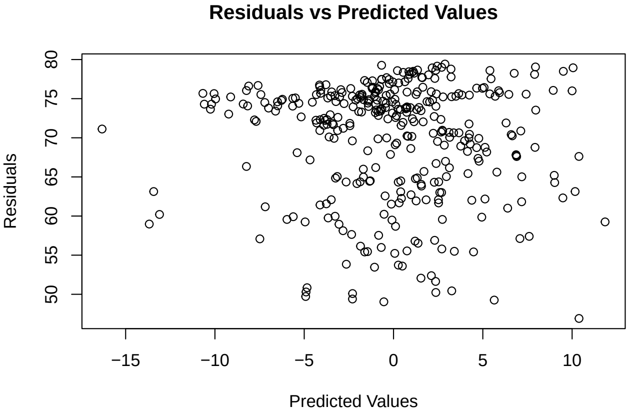
To begin our evaluation process, we initiate the transformation of predictor variables within the test dataset, aligning them with the model's requisites. Subsequently, employing R, we generate a scatterplot that draws a correlation between the actual values of the response variable, Life.expectancy, and the corresponding predicted values of Life.expectancy within the test dataset. Upon examining the resultant scatterplot, a compelling observation emerges: the predictive line closely approximates the actual values of Life.expectancy. This proximity between the predicted and actual values underscores a favorable alignment and a high degree of fitting within the model. This alignment manifests as a testament to the model's accuracy and predictive capacity, signifying a successful and reliable outcome.

Then, we look at some statistic metrics to further quantify the fittness of our model with the test dataset.

```
##   Mean.Absolute.Error.of.Test R.Squared.of.Test
## 1                  0.05040106         0.7231712
##   Mean.Squared.Prediction.Error.of.Test Mean.Squared.Error.of.Train
## 1                              20.09394                    19.93626
```

The Mean Absolute Error for the model is 0.04719227. Since MAE quantifies the average absolute difference between the actual values and the predicted values generated by the model, then a MAE of 0.04719227 indicates a relatively small average prediction error, suggesting that the model is making accurate predictions and exhibiting commendable performance in approximating the true values of the response variable, Life.expectancy. An R.Squared of 0.7388891 suggests approximately 73.89% of the variability in the response variable (Life.expectancy) can be accounted for by the predictor variables included in the model, meaning that the model is capturing a substantial portion of the variance in the response variable using the selected predictors. The Mean Squared Prediction Error (MSPE) for the test dataset is 18.98734, which indicates that, on average, the squared differences between the model's predictions and the actual values amount to approximately 18.98734 units. And the Mean Squared Error for the training dataset is 20.39143, which is very close to the MSPE. Indeed, $\frac{MSPE}{MSE} = 0.917035$, suggesting an appropriate indication of the predictive ability of the model as it is close to 1. Therefore, we have shown that the developed data is reliable and effective for demonstrating the relationship based on the selected predictors.

Moreover, analyzing the correlation between residuals and predicted values provides insight into the quality of the model's fit.



The scatter plot displayed above, meticulously contrasting residuals against their corresponding predicted values, unfurls a reassuring sight. The absence of conspicuous patterns or discernible trends speaks volumes about the fidelity of the model's fit. This intriguing visual insight strongly hints at a harmonious alignment between the model's predictions and the actual values of the response variable Life. expectancy, substantiating the model's effectiveness and reliability.

Therefore, based on the model's performance on the test dataset and the comprehensive analysis of diverse metrics and scatterplots, we can confidently conclude that the developed model effectively and reliably predicts Life.expectancy with a high degree of accuracy.

4. Conclusion Section:
According to the discussion of the goodness of the Model Development shows the developed model has explained most of the data, approximately showing the factors that contributes to the Life.expectancy and there linear relationship based on metrics including validating assumptions, R Squared, MAE, MSE... of the test dataset. Thus, the developed model serves as a crucial tool for unraveling the multifaceted web of factors that contribute to life expectancy. By amalgamating a diverse range of variables, the model offers a comprehensive framework for comprehending the complex dynamics influencing human longevity.

The following is the final model:
$\hat{Y} = 57.56895 - 3.62363X_1 + 1.23152X_2 - 0.34059X_3$, where
$\hat{Y}$: Life.expectancy
X_1 : log_HIV.AIDS

X_2 : log_GDP
X_3 : thinness.5.9.years
The ultimate refined model is expressed as follows: Life expectancy (Y) is predicted based on several key factors. These factors are log_HIV.AIDS, log_GDP, and thinness.5.9.years, each contributing uniquely to the prediction.Let's break it down:

(1) X_1 : log_HIV.AIDS
log_HIV.AIDS: This variable signifies the logarithm of deaths per 1 000 live births HIV/AIDS (0-4 years). The regression coefficient for log_HIV.AIDS is -3.62363. Higher values here are associated with lower life expectancy. So, it means that, on average, if the deaths per 1 000 live births HIV/AIDS (0-4 years) increases by 1%, the average life expectancy will decrease by 3.62363 years, assuming all the predictor variables are held constant. In places with higher HIV/AIDS prevalence, healthcare systems might be strained, and individuals may have compromised immune systems, leading to a shorter average life expectancy.
(2) X_2 : log_GDP
log_GDP (Economic Strength): This variable's logarithm reflects the logarithm of a country's gross domestic product per capita (in USD). The regression coefficient for log_GDP is 1.23152. So, it means that, on average, a 1% increase in a country's gross domestic product per capita (in USD), the average life expectancy will increase by 1.23152 years, assuming all the predictor variables are held constant. A higher log_GDP suggests a stronger economy, which is generally linked to better healthcare access, living standards, and nutrition. These factors contribute to longer life expectancy.
(3) X_3 : thinness.5.9.years
thinness.5.9.years (Thinness Rates among 5 to 9-year-olds): This factor represents the prevalence of thinness in children aged 5 to 9 years with a percentage unit. The regression coefficient for thinness.5.9.years is -0.34059. So, it means on average that, if the thinness Rates among 5 to 9-year-olds increases by 1%, then the averge life expectancy will decrease by 0.34059 year. Higher rates of thinness might indicate issues with nutrition and overall health, which can have a negative impact on life expectancy.
Note that there is no real meaning for the intercept 57.56895 as all the predictors cannot be 0 simultaneously.

In the real world, the model's insights could guide policymakers, healthcare professionals, and governments. For instance, identifying areas with high HIV/AIDS rates can prompt targeted interventions to improve healthcare access and awareness. Similarly, addressing thinness rates among young children can lead to strategies for better nutrition and health education. Overall, this model equips us with a valuable tool to make informed decisions for improving population health and well-being.

While the model showcases notable predictive capabilities, there are certain limitations to consider. Firstly, when examining the histogram of the response variable, Life.expectancy, a slight left-skewed distribution is evident. This characteristic reflects the diverse nature of average life expectancy across different countries. Although the histogram roughly adheres to the shape of a normal distribution, it's important to acknowledge the departure from perfect normality. Secondly, when investigating the scatterplots of each predictor variable against the response variable, an interesting pattern emerges. Outliers, or data points that significantly deviate from the general trend, become apparent. These outliers particularly impact the predictor variable thinness.5.9.years, where the rates of Thinness among 5 to 9-year-olds are notably pronounced. This occurrence suggests a potential breach in the assumption of constant variance, introducing a degree of variability that requires careful consideration. Furthermore, the evaluation of the fourth assumption of our multiple linear regression model involves analyzing the Normal Q-Q plot. Here, a slight S-shape is observed, indicating the presence of outliers in both the lower and higher ranges of predictor variables. These outliers have the potential to influence the construction of confidence and prediction intervals, subsequently affecting the model's precision. While the presence of outliers and departures from perfect normality may pose challenges, the model's overall performance remains reliable and insightful, enriching our understanding of the complex relationship between life expectancy and the selected predictor variables.

In summary, our model underwent a comprehensive validation process involving assumption verification, metric evaluation, and test dataset analysis. Through this process, we not only highlighted the model's limitations but also showcased its commendable performance. Overall, the developed model proves effective and reliable, capable of explaining a substantial portion of the data, rendering it both trustworthy and valuable. This enhanced understanding of the intricate interplay between life expectancy and the selected variables enriches our comprehension. Notably, countries seeking to enhance population well-being and human development can focus on the variables of GDP, deaths per 1,000 live births due to HIV/AIDS (0-4 years), and Thinness Rates among 5 to 9-year-olds as pivotal areas of intervention.

Reference:

WHO and United Nations website. (n.d.). Life expectancy (WHO) (D. Wang & D. Russell, Eds.). Kaggle. Retrieved August 14, 2023, from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Appendix:

```r
knitr::opts_chunk$set(echo = TRUE)
```

```r
setwd("/Users/tanshuxin/Desktop/Second Year s/STA302/final project")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```r
library(ggplot2)
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(ggcorrplot)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
life0=read.csv('Life Expectancy Data.csv', header=TRUE)
life = na.omit(life0)
sample <- sample(c(TRUE, FALSE), nrow(life), replace=TRUE, prob=c(0.8,0.2))
train  <- life[sample, ]
test   <- life[!sample, ]
attach(train)
```

```r
hist(Life.expectancy)
```

```
transform_Status = ifelse(Status=="Developing", 0, 1)
train = train %>% mutate(transform_Status)
options(repr.plot.width = 5, repr.plot.height =2)
boxplot(Life.expectancy~transform_Status, data=train, xlab="transform_Status", ylab="Life.expectancy")
```

```
plot(x=Adult.Mortality, y=Life.expectancy)
```

```
par(mfrow = c(1, 2))
plot(x=infant.deaths, y=Life.expectancy)
plot(x=sqrt(infant.deaths), y=Life.expectancy, xlab='sqrt_infant.deaths')
```

```
sqrt_infant.deaths = sqrt(infant.deaths)
train = train %>% mutate(sqrt_infant.deaths)
```

```
plot(x=Alcohol, y=Life.expectancy)
```

```
par(mfrow = c(1, 2))
plot(x=percentage.expenditure, y=Life.expectancy)
plot(x=sqrt(percentage.expenditure), y=Life.expectancy, xlab='sqrt_percentage.expenditure')
```

```
sqrt_percentage.expenditure = sqrt(percentage.expenditure)
train = train %>% mutate(sqrt_percentage.expenditure)
```

```
plot(x=Hepatitis.B, y=Life.expectancy)
```

```
plot(x=Measles, y=Life.expectancy)
```

```
plot(x=BMI, y=Life.expectancy)
```

```
par(mfrow = c(1, 2))
plot(x=under.five.deaths, y=Life.expectancy)
plot(x=sqrt(under.five.deaths), y=Life.expectancy, xlab='sqrt_under.five.deaths')
```

```
sqrt_under.five.deaths = sqrt(under.five.deaths)
train = train %>% mutate(sqrt_under.five.deaths)
```

```
plot(x=Polio, y=Life.expectancy)
```

```
plot(x=Total.expenditure, y=Life.expectancy)
```

```
plot(x=Diphtheria, y=Life.expectancy)
```

```
par(mfrow = c(1, 2))
plot(x=HIV.AIDS, y=Life.expectancy)
plot(x=log(HIV.AIDS), y=Life.expectancy, xlab='log_HIV.AIDS')
```

```
log_HIV.AIDS = log(HIV.AIDS)
train = train %>% mutate(log_HIV.AIDS)
```

```
par(mfrow = c(1, 2))
plot(x=GDP, y=Life.expectancy)
plot(x=log(GDP), y=Life.expectancy, xlab='log_GDP')
```

```
log_GDP = log(GDP)
train = train %>% mutate(log_GDP)
```

```
plot(x=Population, y=Life.expectancy)
```

```r
plot(x=thinness..10.19.years, y=Life.expectancy)
```

```r
plot(x=thinness.5.9.years, y=Life.expectancy)
```

```r
plot(x=Income.composition.of.resources, y=Life.expectancy)
```

```r
plot(x=Schooling, y=Life.expectancy)
```

```r
train1=subset(train, select = c('transform_Status', 'Adult.Mortality', 'Alcohol', 'BMI', 'log_HIV.AIDS'

#both stepwise method
intercept_only <- lm(Life.expectancy ~ 1, data=train1)
all <- lm(Life.expectancy ~., data=train1)
both <- step(intercept_only, direction='both', scope=formula(all), trace=0)
summary(both)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ log_HIV.AIDS + Income.composition.of.resources +
##     Adult.Mortality + Schooling + thinness.5.9.years + log_GDP +
##     transform_Status, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.575  -1.882  -0.018   1.886  12.617
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     54.434093   0.695071  78.314  < 2e-16 ***
## log_HIV.AIDS                    -2.198342   0.083260 -26.403  < 2e-16 ***
## Income.composition.of.resources 10.000551   0.908075  11.013  < 2e-16 ***
## Adult.Mortality                 -0.015879   0.001036 -15.333  < 2e-16 ***
## Schooling                        0.547725   0.063823   8.582  < 2e-16 ***
## thinness.5.9.years              -0.129163   0.023362  -5.529 3.88e-08 ***
## log_GDP                          0.335977   0.070657   4.755 2.20e-06 ***
## transform_Status                 0.587525   0.321375   1.828   0.0677 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.489 on 1326 degrees of freedom
## Multiple R-squared:  0.8385, Adjusted R-squared:  0.8376
## F-statistic: 983.3 on 7 and 1326 DF,  p-value: < 2.2e-16
```

```r
'vif'
```

```
## [1] "vif"
```

```r
vif(both)
```

```
##                    log_HIV.AIDS Income.composition.of.resources
##                        1.920867                        3.018325
##                 Adult.Mortality                       Schooling
##                        1.791854                        3.474763
##              thinness.5.9.years                         log_GDP
##                        1.333682                        1.618791
##                transform_Status
##                         1.412495
```

```r
#Drop log_HIV.AIDS
both01 <- lm(formula = Life.expectancy ~ Income.composition.of.resources +
    Adult.Mortality  + thinness.5.9.years + log_GDP +
    transform_Status, data = train1)
summary(both01)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Income.composition.of.resources +
##     Adult.Mortality + thinness.5.9.years + log_GDP + transform_Status,
##     data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.7984  -1.9347   0.3023   2.2330  15.5897
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      58.735739   0.807699  72.720  < 2e-16 ***
## Income.composition.of.resources 18.707532   0.930869  20.097  < 2e-16 ***
## Adult.Mortality                 -0.031200   0.001126 -27.697  < 2e-16 ***
## thinness.5.9.years              -0.168835   0.029658  -5.693 1.54e-08 ***
## log_GDP                          0.631096   0.087797   7.188 1.09e-12 ***
## transform_Status                 1.295127   0.403740   3.208  0.00137 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.492 on 1328 degrees of freedom
## Multiple R-squared:  0.7318, Adjusted R-squared:  0.7308
## F-statistic: 724.7 on 5 and 1328 DF,  p-value: < 2.2e-16
```

```r
#Drop Income.composition.of.resources
both02 <- lm(formula = Life.expectancy ~ log_HIV.AIDS +
    Adult.Mortality + thinness.5.9.years + log_GDP +
    transform_Status, data = train1)
summary(both02)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ log_HIV.AIDS + Adult.Mortality +
##     thinness.5.9.years + log_GDP + transform_Status, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.0674  -2.2453   0.1179   2.4414  15.8023
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        63.08608    0.63610  99.177  < 2e-16 ***
## log_HIV.AIDS       -2.64425    0.09325 -28.357  < 2e-16 ***
## Adult.Mortality    -0.01824    0.00119 -15.324  < 2e-16 ***
## thinness.5.9.years -0.27898    0.02584 -10.798  < 2e-16 ***
## log_GDP             0.96382    0.07425  12.980  < 2e-16 ***
## transform_Status    2.57579    0.35465   7.263 6.44e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.049 on 1328 degrees of freedom
## Multiple R-squared:  0.7822, Adjusted R-squared:  0.7813
## F-statistic: 953.6 on 5 and 1328 DF,  p-value: < 2.2e-16
```

```
#Compare the goodness of both models
goodness1 <- matrix(c(summary(both01)$r.squared, summary(both02)$r.squared, AIC(both01), AIC(both02), BI
rownames(goodness1) <- c('Drop log_HIV.AIDS', 'Drop Income.composition.of.resources')
colnames(goodness1) <- c('R Squared', 'AIC', 'BIC')
goodness1 <- as.table(goodness1)
goodness1
```

```
##                                        R Squared          AIC          BIC
## Drop log_HIV.AIDS                      0.7318117 7802.1194392 7838.4909998
## Drop Income.composition.of.resources   0.7821560 7524.7672789 7561.1388394
```

```
train3=subset(train, select = c('transform_Status', 'Adult.Mortality', 'log_HIV.AIDS', 'log_GDP', 'thinn

corr_matrix1=cor(train3)
ggcorrplot(corr_matrix1, hc.order = TRUE, type = "lower", lab = TRUE)
```

```
#Drop Adult.Mortality
both03 <- lm(formula = Life.expectancy ~ transform_Status + log_HIV.AIDS + log_GDP + thinness.5.9.years
summary(both03)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ transform_Status + log_HIV.AIDS +
##     log_GDP + thinness.5.9.years, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0033  -2.7602   0.1031   2.6915  16.2817
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        58.83838    0.62085  94.771  < 2e-16 ***
## transform_Status    3.06988    0.38299   8.015 2.38e-15 ***
## log_HIV.AIDS       -3.46883    0.08258 -42.004  < 2e-16 ***
## log_GDP             1.00893    0.08046  12.540  < 2e-16 ***
## thinness.5.9.years -0.32231    0.02785 -11.573  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.391 on 1329 degrees of freedom
## Multiple R-squared:  0.7436, Adjusted R-squared:  0.7429
## F-statistic: 963.8 on 4 and 1329 DF,  p-value: < 2.2e-16
```

```
#Drop log_HIV.AIDS.
both04 <- lm(formula = Life.expectancy ~ transform_Status + Adult.Mortality + log_GDP + thinness.5.9.yea
summary(both04)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ transform_Status + Adult.Mortality +
##     log_GDP + thinness.5.9.years, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.6535  -2.2047   0.6934   2.8739  17.4549
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        67.294255   0.783458  85.894  < 2e-16 ***
## transform_Status    3.169724   0.448423   7.069 2.52e-12 ***
## Adult.Mortality    -0.037721   0.001231 -30.633  < 2e-16 ***
## log_GDP             1.293674   0.092890  13.927  < 2e-16 ***
## thinness.5.9.years -0.325957   0.032658  -9.981  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.128 on 1329 degrees of freedom
## Multiple R-squared:  0.6502, Adjusted R-squared:  0.6492
## F-statistic: 617.7 on 4 and 1329 DF,  p-value: < 2.2e-16
```

```
goodness2 <- matrix(c(summary(both03)$r.squared, summary(both04)$r.squared, AIC(both03), AIC(both04), BI
rownames(goodness2) <- c('Drop Adult.Mortality', 'Drop log_HIV.AIDS')
colnames(goodness2) <- c('R Squared', 'AIC', 'BIC')
goodness2 <- as.table(goodness2)
goodness2
```

```
##                        R Squared          AIC          BIC
## Drop Adult.Mortality   0.7436363 7739.9665627 7771.1421861
## Drop log_HIV.AIDS      0.6502478 8154.3438776 8185.5195009
```

```
#Drop transform_Status
both05 <- lm(formula = Life.expectancy ~ log_HIV.AIDS + log_GDP + thinness.5.9.years, data = train1)
summary(both05)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ log_HIV.AIDS + log_GDP + thinness.5.9.years,
##     data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1565  -2.7509   0.0666   2.7750  15.8094
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        57.98334    0.62599   92.63   <2e-16 ***
## log_HIV.AIDS       -3.55960    0.08373  -42.52   <2e-16 ***
## log_GDP             1.19825    0.07872   15.22   <2e-16 ***
## thinness.5.9.years -0.36165    0.02806  -12.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.494 on 1330 degrees of freedom
## Multiple R-squared:  0.7312, Adjusted R-squared:  0.7306
## F-statistic:  1206 on 3 and 1330 DF,  p-value: < 2.2e-16
```

```
#Drop log_GDP
both06 <- lm(formula = Life.expectancy ~ transform_Status + Adult.Mortality + thinness.5.9.years, data =
summary(both06)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ transform_Status + Adult.Mortality +
##     thinness.5.9.years, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.1131  -2.3991   0.9452   3.3952  15.4433
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        77.40963    0.31431  246.28   <2e-16 ***
## transform_Status    5.05584    0.45744   11.05   <2e-16 ***
## Adult.Mortality    -0.04049    0.00130  -31.13   <2e-16 ***
## thinness.5.9.years -0.41617    0.03425  -12.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.488 on 1330 degrees of freedom
## Multiple R-squared:  0.5992, Adjusted R-squared:  0.5983
## F-statistic: 662.8 on 3 and 1330 DF,  p-value: < 2.2e-16
```

```
goodness3 <- matrix(c(summary(both05)$r.squared, summary(both06)$r.squared, AIC(both05), AIC(both06), BI
rownames(goodness3) <- c('Drop transform_Status', 'Drop log_GDP')
colnames(goodness3) <- c('R Squared', 'AIC', 'BIC')
goodness3 <- as.table(goodness3)
goodness3
```

```
##                       R Squared          AIC          BIC
## Drop transform_Status    0.7312429 7800.9459213 7826.9256074
## Drop log_GDP             0.5992032 8334.0745267 8360.0542128
```

```
final_train=subset(train, select = c('log_HIV.AIDS', 'log_GDP', 'thinness.5.9.years', 'Life.expectancy'

corr_matrix2=cor(final_train)
'Correlation Matrix for the Developed Model'
```

```
## [1] "Correlation Matrix for the Developed Model"
```

```
corr_matrix2
```

```
##                    log_HIV.AIDS    log_GDP thinness.5.9.years Life.expectancy
## log_HIV.AIDS          1.0000000 -0.3399471          0.2761090      -0.7894263
## log_GDP              -0.3399471  1.0000000         -0.3271145       0.5250147
## thinness.5.9.years    0.2761090 -0.3271145          1.0000000      -0.4557942
## Life.expectancy      -0.7894263  0.5250147         -0.4557942       1.0000000
```

```
ggcorrplot(corr_matrix2, hc.order = TRUE, type = "lower", lab = TRUE, title='Correlation Matrix')
```

```
'vif'
```

```
## [1] "vif"
```

```
vif(both05)
```

```
##       log_HIV.AIDS            log_GDP thinness.5.9.years
##           1.170984           1.211329           1.159759
```

```
res <- summary(both05)$residuals
par(mfrow = c(1, 3))
plot(x = log_HIV.AIDS, y=Life.expectancy)
plot(x = thinness.5.9.years, Life.expectancy)
plot(x = log_GDP, Life.expectancy)
```

```
layout(matrix(c(1,2,3,4),2,2))
plot(both05)
```

```
par(mfrow = c(1, 2))
hist(res, xlab='Residuals', main='Histogram of Residuals')
qqnorm(res,
       main="Normal Q-Q plot")
qqline(res)
```

```
metrics <- data.frame('R Squared'= c(summary(both05)$r.squared),
                      'Adjusted R Squared'= c(summary(both05)$adj.r.squared))

metrics
```

```
##   R.Squared Adjusted.R.Squared
## 1 0.7312429          0.7306367
```

```
#apply the same data transformation on the test dataset
test$log_GDP = log(test$GDP)
test = test %>% mutate(test$log_GDP)

test$log_HIV.AIDS = log(test$HIV.AIDS)
test = test %>% mutate(test$log_HIV.AIDS)

test1 = subset(test, select=c('log_GDP', 'log_HIV.AIDS', 'thinness.5.9.years', 'Life.expectancy'))

plot(predict(both05, newdata=test1), test1$Life.expectancy,
     xlab = "Predicted Values for the test dataset",
     ylab = "Observed Values from the test dataset")
abline(a = 0, b = 1, lwd=2,
       col = "red")
```

```
##Analyze the performance metrics and visualizations to understand the model's performance on the test

y_actual = test1$Life.expectancy
y_predict = predict(both05, newdata=test1)

#1. MEAN ABSOLUTE ERROR (MAE)
MAE = mean(abs((y_actual-y_predict)/y_actual))

#2. R SQUARED error metric -- Coefficient of Determination
R2 = cor(y_actual,y_predict)^2

#3. Mean Squared Error
MSE = sum((both05$residuals^2))/(1320-3-1)
```

```r
#4. Mean Squared Prediction Error
MSPE = mean((test1$Life.expectancy-predict(both05, newdata=test1))^2)

predict_goodness <- data.frame('Mean Absolute Error of Test'= c(MAE), 'R Squared of Test'= c(R2), 'Mean
predict_goodness
```

```
##   Mean.Absolute.Error.of.Test R.Squared.of.Test
## 1                   0.0497341         0.7755739
##   Mean.Squared.Prediction.Error.of.Test Mean.Squared.Error.of.Train
## 1                              19.78617                    20.40944
```

```r
plot(x=test1$Life.expectancy-predict(both05, newdata=test1), y=predict(both05, newdata=test1), ylab='Re
```