University of Waterloo
MSE 609 - Fall 2024

Final Project Proposal

Title:
Global Disparities in Life Expectancy: A Cross-Country Comparison of Socioeconomic and Health Determinants

Team Name: CA_Health

Team Members

Hai Ning YIN (h68yin)
Joshua WONG (j89wong)
Sheldon ZHANG (x399zhan)
Vick FENG (z83feng)
Nan WANG (n96wang)

1 Description of Data

1.1 Background

Life expectancy, a key indicator of population health, has become a focal point for policymakers and researchers due to its broad implications for social and economic development. Over the past century, substantial increases in life expectancy have occurred globally, yet vast disparities persist among different regions and countries, reflecting a complex interplay of its determinants. Factors such as income levels, healthcare accessibility, education, environmental conditions, and sociopolitical stability all contribute to variations in life expectancy (Cutler, Deaton & Lleras-Muney, 2006; Oeppen & Vaupel, 2002). While significant progress has been made, understanding the drivers of these differences remains crucial for designing effective policies to improve the overall population health. Our study seeks to identify and quantify the key determinants of life expectancy by employing regression analyses across all countries, focusing on both lifestyle and socioeconomic factors that influence health outcomes. By exploring these determinants, we aim to offer insights that may inform strategies for addressing health disparities and fostering more equitable health outcomes globally.

1.2 Data Sources

We have chosen to utilize data from the World Health Organization's (WHO) Global Health Observatory for our study on life expectancy determinants. The WHO is a globally recognized authority in the field of health, and its data collection methods are designed to ensure reliability and comparability across different countries and regions.

The organization collects data through a combination of surveys, administrative records, and direct reporting from member countries. For example, health surveys are conducted in collaboration with national health authorities to gather information on various health determinants and outcomes. Additionally, vital statistics such as births and deaths are obtained from civil registration systems, which provide accurate and comprehensive data on population dynamics. The WHO also employs standardized procedures for data collection and quality control to minimize errors and biases. This ensures that the data we use is of high quality and can be trusted for our analysis.

The Global Health Observatory offers a wide range of data on factors such as socioeconomic conditions, lifestyle behaviors, health status indicators, and environmental exposures, which are all relevant to our study of life expectancy determinants. The availability of such diverse data allows us to conduct a comprehensive analysis and gain a deeper understanding of the complex relationships between these factors and life expectancy.

1.3 Description of Key Variables

1.3.1 Key response: Life Expectancy

Life expectancy at birth is defined as the average number of years a newborn is expected to live if the current mortality patterns remain constant throughout their life (World Health Organization (WHO), n.d.). It is calculated based on gender and age-specific mortality rates and serves as a comprehensive measure of the health status of a population. This indicator takes into account the mortality risks at all ages, from infancy to old age, and provides an overall picture of the population's health and longevity. According to the United Nations (UN), life expectancy is influenced by a multitude of factors, including access to healthcare, education, nutrition, and environmental conditions (United Nations, n.d.)

1.3.2 Key Predictors

(a) Socioeconomic factors

- **GDP per capita (PCGMP):** This variable represents the GDP per capita in USD, which is the economic output per person in a given area. It serves as a proxy for the overall economic well-being and development of a region. Higher GDP per capita is often associated with better access to healthcare services, improved living conditions, and enhanced resources for public health initiatives (World Bank, n.d.). This can translate into better preventive care, advanced medical treatments, and a more comprehensive healthcare infrastructure, all of which can potentially contribute to increased life expectancy. For example, countries with higher GDP per capita may have more funds to invest in research and development of new medical technologies, as well as in the training and education of healthcare professionals.

- **Average years of schooling (EduAvg):** Education is a crucial determinant of health behaviours and awareness. It emphasizes that individuals with more years of formal education tend to have a better understanding of health-related issues and are more likely to make informed decisions regarding their lifestyle (World Health Organization (WHO), n.d.). They may be more conscious of the importance of a balanced diet, regular exercise, and preventive healthcare measures. Additionally, higher education levels are often associated with better employment opportunities and higher incomes, which can further contribute to improved living standards and access to healthcare, ultimately influencing life expectancy.

(b) Lifestyle factors

- **Alcohol per capita consumption (Alc_pcgmp):** Total APC is defined as the total (sum of three -year average recorded and three-year average unrecorded APC, adjusted for three-year average tourist consumption) amount of alcohol consumed per adult (15+ years) over a calendar year, in litres of pure alcohol. It has been established that excessive alcohol consumption is a significant risk factor for a multitude of health problems. It can lead to liver damage, including cirrhosis, and increase the risk of developing various cancers, such as liver, breast, and esophageal cancer. Moreover, it is associated with cardiovascular diseases, including hypertension and heart failure (World Health Organization (WHO), n.d.). These health conditions can have a severe impact on an individual's

quality of life and can ultimately reduce life expectancy. For instance, chronic alcohol abuse can weaken the immune system, making individuals more susceptible to infections and other diseases.

- **Price of a pack of cigarettes (P_Ciga) and % of Tax, cigarette (Tax_Ciga):** P_Ciga represents the price of a pack 20 cigarettes of most sold brand in USD in the countr; Tax_Ciga represents the percentage of Tax in the price of a pack 20 cigarettes. These variables are directly related to smoking behavior. Higher cigarette prices and taxes can act as economic deterrents, reducing the prevalence of smoking. Smoking is a leading cause of preventable diseases. It is a major risk factor for lung cancer, chronic obstructive pulmonary disease (COPD), and cardiovascular diseases (World Health Organization (WHO), n.d.). The harmful effects of smoking not only impact smokers themselves but also those exposed to secondhand smoke. By influencing smoking rates, these variables can have a significant impact on population health and life expectancy.

- **Prevalence of insufficient physical activity among adults aged 18+ years (Insuf_Phy):** Physical activity is essential for maintaining good health. Regular physical activity is recommended to prevent a range of chronic diseases. Insufficient physical activity is associated with an increased risk of obesity, which in turn is linked to numerous health problems, including type 2 diabetes, heart disease, and certain cancers. Lack of exercise can also lead to muscle weakness, poor cardiovascular function, and reduced flexibility, all of which can contribute to a lower quality of life and potentially shorter life expectancy (World Health Organization (WHO), n.d.).

- **Calorie intake (Cal_Daily):** Adequate calorie intake is necessary for maintaining bodily functions and overall health. Dietary guidelines provided emphasize the importance of a balanced diet in relation to calorie consumption. Consuming too few calories can lead to malnutrition, weakened immune function, and increased susceptibility to diseases. On the other hand, excessive calorie intake, especially from unhealthy sources, can contribute to obesity and its associated health risks (World Health Organization (WHO), n.d.). Therefore, understanding the daily calorie supply per person is crucial in assessing its impact on health and life expectancy.

(c) Health status and medical factors:

- **Prevalence of diabetes (Diab):** Diabetes is a chronic metabolic disorder that affects the body's ability to regulate blood sugar levels. The global burden of diabetes and its associated complications are highlighted. If not properly managed, diabetes can lead to serious health problems, such as cardiovascular disease, kidney damage, nerve damage, and vision loss (World Health Organization (WHO), n.d.). These complications can significantly reduce an individual's quality of life and increase the risk of premature death, thereby affecting life expectancy.

- **Prevalence of hypertension (Hyp_Ten) and Prevalence of raised blood pressure among adults aged 30 - 79 years (R_Bld_P):** High blood pressure is a major risk factor for cardiovascular diseases. It can lead to heart attacks, strokes, and other cardiovascular complications. Managing hypertension through lifestyle modifications and appropriate medical treatment is essential for reducing the risk of these life-threatening events and maintaining good health, which in turn can impact life expectancy (World Health Organization (WHO), n.d.).

- **Immunization coverage variables (DTP3, Hep83, MCV1, Pol3):** Immunization is a cornerstone of public health. These variables indicate the percentage of the population, particularly children, who have been vaccinated against specific diseases. High immunization coverage helps prevent the spread of infectious diseases, protecting individuals from potentially severe and life-threatening illnesses (World Health Organization (WHO), n.d.). By reducing the burden of infectious diseases, immunization can contribute to improved population health and increased life expectancy.

- **Prevalence of HIV (HIV):** HIV is a global health concern, and efforts are actively involved in combating the epidemic. HIV infection weakens the immune system, making individuals more susceptible to opportunistic infections and certain cancers. Without proper treatment and management, HIV can progress to AIDS and significantly reduce life expectancy (UNAIDS, n.d.). Access to antiretroviral therapy and comprehensive HIV prevention programs is crucial in controlling the spread of the virus and improving the health and survival of those affected.

- **Mean HDL cholesterol (HDL_Chol) and Mean non - HDL cholesterol (N_HDL_Chol):** Cholesterol levels play a role in cardiovascular health. HDL cholesterol is often referred to as "good" cholesterol because it helps remove LDL cholesterol (the "bad" cholesterol) from the bloodstream. Higher levels of HDL cholesterol are generally associated with a lower risk of cardiovascular disease, while elevated levels of non - HDL cholesterol are a risk factor for heart disease. Guidelines on healthy cholesterol levels are provided and the importance of maintaining a proper balance to reduce the risk of cardiovascular problems and their impact on life expectancy is emphasized (World Health Organization (WHO), n.d.).

(d) Environmental factors:

- **Percentage, Access to clean water [Water pollution related] (Acs_Water_Service%):** Access to clean and safe drinking water is fundamental for good health. Lack of access to clean water can lead to the spread of waterborne diseases, such as cholera, typhoid, and diarrhea. These diseases can be particularly severe in children and can have a significant impact on the health and well-being of the population (World Health Organization (WHO), n.d.). Improving access to clean water is essential for preventing illness and reducing mortality, especially in areas with limited access to proper sanitation and water treatment facilities.

- **Air pollution (PM2.5):** It has been identified as a major environmental risk to health. Exposure to fine particulate matter (PM2.5) is associated with respiratory and cardiovascular problems, including asthma, bronchitis, lung cancer, and heart disease. Long-term exposure to high levels of air pollution can have a cumulative effect on health, leading to reduced lung function, increased hospital admissions, and premature death (World Health Organization (WHO), n.d.). Mitigating air pollution through policies and interventions is crucial for protecting public health and potentially increasing life expectancy, especially in urban areas where pollution levels are often higher.

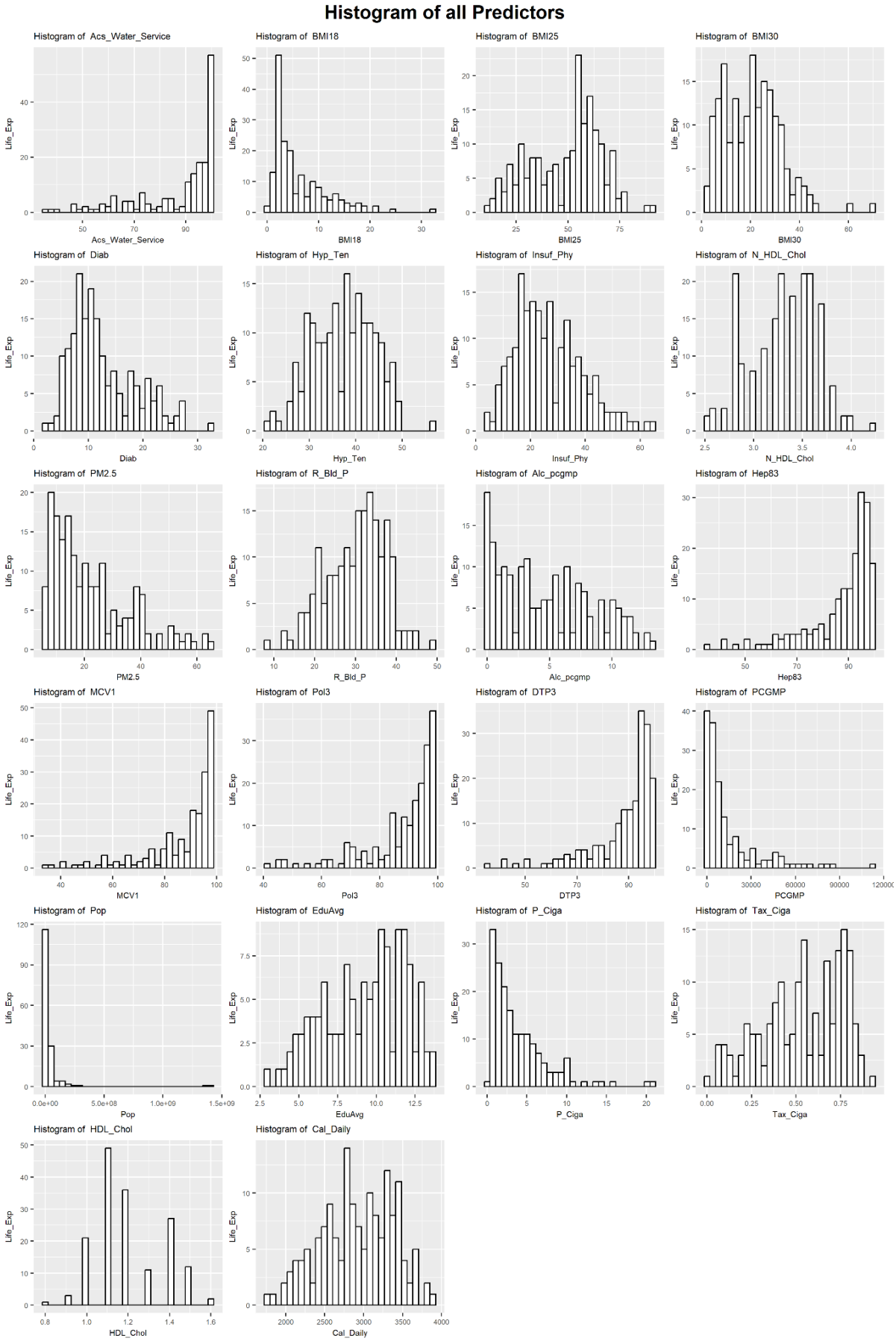
- **Population (Pop):** The size and density of the population can have implications for healthcare delivery, resource allocation, and the spread of diseases. In densely populated areas, there may be greater challenges in providing adequate healthcare services to all individuals. Additionally, population density can influence the transmission of infectious diseases. Understanding the population characteristics is important in assessing the overall health needs of a region and developing appropriate public health strategies to address them (World Health Organization (WHO), n.d.).

- **Prevalence of underweight (BMI18), Prevalence of overweight (BMI25), and Prevalence of obesity (BMI30):** BMI is an important indicator of body fatness and is related to health risks. Classifications of BMI ranges and guidelines on healthy weight are provided. A BMI below 18.5 (BMI18) may indicate undernutrition or underlying health problems, which can increase the risk of infections, osteoporosis, and other health issues. On the other hand, a BMI between 25 and 30 (BMI25) is considered overweight, and a BMI of 30 or above (BMI30) is classified as obese. Overweight and obesity are associated with an increased risk of developing chronic diseases such as diabetes, heart disease, and certain cancers. These conditions can lead to complications and a higher mortality rate, thus affecting life expectancy (World Health Organization (WHO), n.d.).

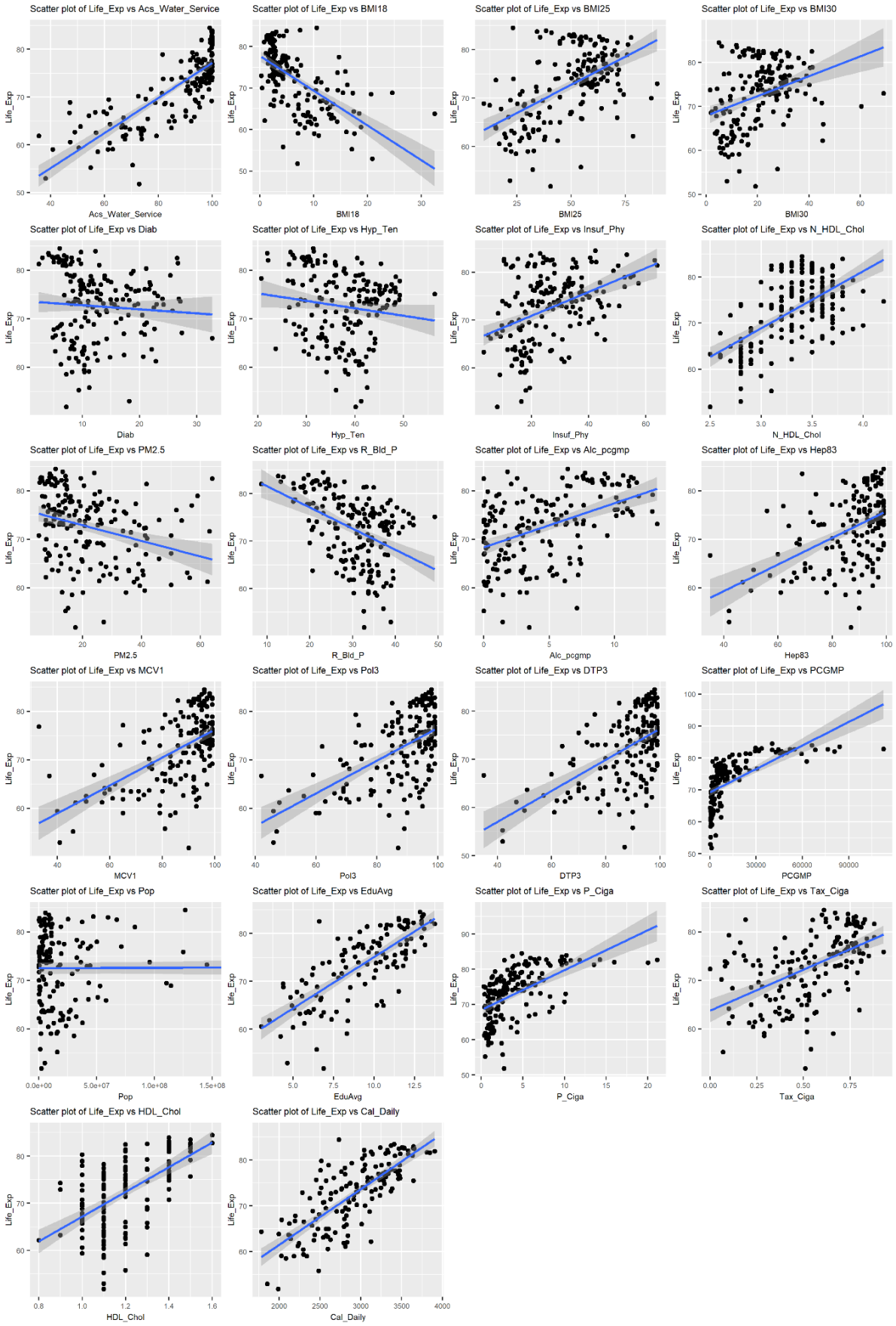
The prevalence of different BMI categories in the data represents the percentage of the population that has BMI within the defined range. Understanding the prevalence of different BMI categories in a population can help identify areas where interventions may be needed to promote healthy weight and improve health outcomes.

2. Descriptive Data Analysis

2.1 Visualization of Data by Scatterplot Diagram & Histogram



Scatterplot of Life Expectancy vs. Predictors with Regression Line



2.2 Discussion of Visualization

The scatterplots reveal relationships between life expectancy and its predictors, providing initial visualizations into the trends observed. We can conclude the scatterplot graphs by the type of trends they show visually above.

Positive Trends

Several variables demonstrate clear positive linear relationships with life expectancy, indicating that improvements in these factors are associated with longer life. For instance, **Access to Water Services (Acs_Water_Service)** suggest that countries with better access to water services tend to have significantly higher life expectancies. Another factor, **Body Mass Index (BMI25 - Overweight),** shows that moderate levels of being overweight appear to be positively correlated with life expectancy. Other variables, such as **Vaccination Coverage (MCV1, DTP3)** indicate that higher vaccination rates for diseases such as measles and diphtheria are strongly linked to longer life expectancy. Furthermore, variables such as Per Capita GDP, Taxation on Tobacco, Educating Index and Good Cholesterol also demonstrate clear positive trends.

Negative Trends

In contrast, several variables exhibit negative relationships with life expectancy, highlighting areas where increased risk factors may lead to poor health outcomes. First, **Diabetes Prevalence (Diab)** demonstrates that higher rates of diabetes are associated with shorter life expectancy. **Hypertension Prevalence (Hyp_Ten)** shows a weak but negative correlation with life expectancy, reflecting cardiovascular risks. **Insufficient Physical Activity (Insuf_Phy)** shows countries with higher levels of physical inactivity tend to have shorter life expectancy. **Air Pollution (PM2.5)** is strongly negatively associated with life expectancy, showcasing the health consequences of environmental pollution.

Unclear Patterns

For certain variables, the scatterplots do not present a clear trend. For example, **BMI30 (Obesity)** appears to be weakly positively related to life expectancy with some variation, making it difficult to draw definitive conclusions. For the **Population (Pop) variable**, the scatterplot for population size and life expectancy shows an almost flat line, with no visible trend. This lack of variation indicates that life expectancy may not be directly related to the size of the population.

Overall, those scatter plots reveal that life expectancy is predominantly influenced by positive linear relationships such as access to water, economic environment and average education. In contrast, negative trends with risk factors, for example, pollution, smoking, and chronic diseases indicate areas with targeted intervention. These consistent linear relations across all variables justify the use of a linear regression model, which quantifies the magnitude and direction of these forces effectively.

Histograms

Variables such as Cal_Daily and PCGMP exhibit relatively normal distributions, suggesting that these predictors have a more even spread across the sampled populations, supporting their strong influence on life expectancy.

A few variables exhibit skewness or clustering at certain ranges, reflecting potential nonlinear relationships with life expectancy. For example, **BMI30** shows right-skewed distributions, suggesting that higher BMI may be concentrated in certain regions or populations.

3 Question of interest

3.1 Research Question

The main research question of our study is whether these variables are associated with life expectancy across different countries. To address this, we reviewed several lines of existing research and selected relevant data for our analysis.

3.2 Related Studies

One source of variation in life expectancy arises from the differing impacts of chronic diseases across different countries. Chronic diseases such as diabetes and cancer pose significant health challenges worldwide and contribute substantially to mortality and morbidity. Their impact on life expectancy has been well-documented, with rising prevalence driven by factors including population aging, lifestyle transitions, and disparities in access to healthcare services (Wild et al., 2004; Ferlay et al., 2015). Diabetes, for example, is a major contributor to cardiovascular disease, kidney disease, and reduced life expectancy, especially in low- and middle-income countries where healthcare access is often limited (Hu, 2011). Similarly, cancer remains a leading cause of death, with cross-country disparities reflecting variations in early detection, prevention, and treatment capabilities (Bray et al., 2018).

Gender differences in various factors also call for a nuanced approach to improve life expectancy globally. Research indicates notable differences between men and women; while women often have longer life expectancies, they may experience higher lifetime morbidity and disability due to chronic conditions (Case & Paxson, 2005). On the other hand, men typically exhibit higher mortality rates from conditions like cancer and cardiovascular disease, possibly due to risk behaviors and delayed healthcare-seeking (Courtenay, 2000).

Our study addresses these gendered disparities by exploring the determinants of chronic disease prevalence and their effects on life expectancy, considering potential heterogeneity between male and female populations across countries. This focus is critical, as gender norms, healthcare-seeking behaviors, and biological differences can influence disease outcomes and health trajectories (WHO, 2019). Furthermore, cross-national comparisons enable the identification of context-specific factors, such as healthcare infrastructure, public health policies, and cultural norms, that mediate disease risk and health outcomes. By examining these determinants comprehensively, our research aims to offer actionable insights into the efforts to improve life expectancy, including mitigating chronic disease burdens, reducing health

disparities, cultivating healthy lifestyles and enhancing socioeconomic status, particularly in diverse international contexts.

3.3 Data relevance

Each predictor has a theoretical and empirical basis for its relationship with life expectancy. Socioeconomic factors influence the availability and quality of healthcare and living conditions. Lifestyle factors directly impact health and the risk of developing chronic diseases. Health status and medical factors are directly related to morbidity and mortality. Environmental factors can have both direct and indirect effects on health and life expectancy. By analyzing these predictors together, we can gain a more comprehensive understanding of the complex determinants of life expectancy and develop more effective public health policies and interventions.

For example, the relationship between BMI and life expectancy is well-established. A higher BMI, especially in the obese range, is associated with an increased risk of developing chronic diseases such as diabetes, heart disease, and certain cancers. These diseases can lead to complications and a higher mortality rate, thus affecting life expectancy. By including BMI-related variables in our analysis, we can assess the impact of body weight distribution in the population on life expectancy and understand how interventions to address overweight and obesity can potentially improve health outcomes and increase life expectancy.

In conclusion, the predictors we have selected are relevant to our study as they capture different aspects of the factors that can influence life expectancy. By understanding these relationships, we can contribute to the development of strategies to improve population health and increase life expectancy.

4. Model Building Approach & Justification

4.1 Model Selection

As reference to our question of interest in section 3, we would like to investigate what factors may be influential to the average life expectancy of the population in different countries. After examining various choices of statistical models available, we chose to perform a Linear Regression Model on this research topic.

Linear Regression Model is an ideal statistical model for evaluating the relationship and statistical significance between a single response (i.e. Y) and multiple predictors (i.e X). In our context, a linear regression model analysis can provide a comprehensive understanding between life expectancy and various health metrics. The formula is expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

Where Y is the response variable (i.e. average Life Expectancy in our context); Xi represents the predictor variables (i.e. health metrics enlisted in section 1), βi are the coefficients to be estimated, and ε is the error term.

Linear Regression Model is chosen over other choices for its fitting properties without question of interest and its simplicity. It is worth noting that there are models that can potentially explain the relationship, for example, the Polynomial Regression Model. However, analyzing the pattern with a simple model such as the Linear Model would be more suitable for our foundational research and can provide a better basis for further studies on the topic in the future.

4.2 Initial Data Processing & Cleaning

The data used in this research is sourced from the online database of The Global Health Observatory (GHO), the official data survey organization of The World Health Organization (WHO). The reference data table, named "FullTable_WHO_DATA_NC_184.csv", was created by full-joining corresponding tables of selected health metrics available from GHO and processed through Excel and R Studio. The combined table consists of data collected from countries and regions of the world by WHO.

Country	Life_Exp		Count of Countries
Afghanista	61.22		183
Albania	77.94		
Algeria	76.55		
Angola	62.49		
Antigua an	75.66		

We built our initial data table based on our major Y response, which is the Life Expectancy Table with 183 initial observations. We matched the other 22 tables of various health metrics by using the "Country" name as the primary key-matching field. We used the full-join method to maintain the total sample size at 183 observations and filled any missing values in any column with NAs initially.

This resulted in an initial data set consisting of 183 total observations on 23 columns. The first column refers to the average life expectancy of the country, which is the major Y response of this research; the rest of the 22 columns consist of health metrics enlisted in section 1, which are the X predictors in the following analysis.

After loading data into R Studio and observing the data structure, there are 77 other observations with incomplete values in various predictors. By observing the missing value pattern, these are mainly small countries. Since small countries may not have sufficient resources and are less eager to perform detailed

annual surveys of collecting national health metrics, we believe this is a missing value not at random (MNAR). Thus, we have to perform imputation on the data.

4.3 Imputation Approach of Missing Values

When choosing the appropriate imputation method, we must consider the nature and missing pattern of our data. There are around 50% of observations that are incomplete with missing values.

Firstly, given that we have only 183 observations available, we cannot directly delete all observations with missing values, as it will lead to a significant reduction in the total sample size and result in bias.

Second, as the values are missing not at random (MNAR), we cannot fill them randomly with on-hand data as it may result in unreasonable results, such as countries with small populations filled with large numbers in other population-related metrics. In addition, we cannot fill them by interpolation either, as this is not a time-series data.

As a result, there are two imputation approaches available for our study, imputing with mean value in each predictor column, and imputation by regression. Also, for comparing model performance, we decided to add the model process with only completed cases. That results in a total of three model imputation and building approaches as follows:

1) Modelling with only complete observations
2) Modelling with Imputation of Mean
3) Modelling with Imputation by Regression

Imputation By Regression & Selection of Imputation Results

There are various regression methods available for imputation. In our analysis, we chose the two common approaches as follows:

1) Classification and Regression Tree (CART)
2) Lasso regression

We would first generate imputation results of both methods for each predictors with missing values. Afterwards, we visualize the distribution of each imputed data on histograms and compare it with the distribution of the original data. The distribution that is closest to the original data would be adopted and used to replace the missing values. The process would be iterated until all missing values are imputed.

4.4 Model Selection by Backward Elimination

In a linear regression model, there can be irrelevant predictors involved in the initial formula, which would reduce the predicting performance of the model. Thus, we decided to adopt the Backward Elimination method in all three imputed approaches mentioned in the previous section to obtain the

optimized model for each imputation approach. We chose Backward Elimination for its capability of providing a simpler, interpretable optimized model in the result.

We performed exactly the same procedures on all three imputation approaches for obtaining a fair result comparison. In each iteration of Backward Elimination, steps are performed as follows:

1) Perform the F-test on all remaining predictors and obtain their p-values using build-in functions of R studio
2) Remove the predictor with the highest p-value (i.e. least statistical relevance) in an F-test.
3) Observe the Akaike information criterion (AIC score) before and after the removal of the predictor. The relative AIC score of the model before and after removal suggests whether the model is improved.
4) Repeat Steps 1-3 if a decrease in AIC score can be observed in the model after removal. Alternatively, if an increase in AIC score is observed. It suggests the model has been optimized and the predictor of the current round would be added back. The Backward Elimination would be halted.

4.5 Comparison of Models

After backward elimination, we shall obtain three optimized models built with three different imputation approaches. The metrics for model comparison we adopted in our analysis are as follows:

1. **Adjusted R squared value:**
   This suggests the "goodness of fit" of the model to our data. A value closer to 1 is preferred.
2. **AIC score of the optimized model**
   The relatively lower score between models suggests which model are likely to be the best on the same data set.

Comparison of Imputed model and Complete case model with fewer observations

In practice, a complete case data set with fewer observations would usually have a lower AIC score as a result of reduced noise and variability in a cleaner data set. But it also sacrifices the potentially informative observations. Therefore, when comparing the two imputed models with the complete case model we would also observe **the Number and patterns of predictors with statistical significance** (i.e. p-value in F-test <0.05) for better judgement of a better model for our topic.

5. Findings & Interpretation of Results

5.1 Approach 1: Modelling with only Complete Cases

5.1.1 Initial Model

```
##
## Call:
## lm(formula = Life_Exp ~ ., data = df_184)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3947 -1.6150 -0.1291  1.9388  5.6356
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.122e+01  6.188e+00   5.045 2.62e-06 ***
## Acs_Water_Service  6.309e-02  4.912e-02   1.284  0.20255
## BMI18              2.389e-01  1.815e-01   1.316  0.19174
## BMI25              2.774e-01  1.202e-01   2.308  0.02349 *
## BMI30             -2.453e-01  1.478e-01  -1.660  0.10072
## Diab              -1.950e-01  8.702e-02  -2.241  0.02770 *
## Hyp_Ten           -2.770e-01  1.548e-01  -1.789  0.07722 .
## Insuf_Phy          9.774e-02  4.156e-02   2.352  0.02104 *
## N_HDL_Chol         3.698e+00  1.280e+00   2.889  0.00493 **
## PM2.5              6.902e-03  3.118e-02   0.221  0.82535
## R_Bld_P            6.323e-02  1.386e-01   0.456  0.64946
## Alc_pcgmp         -2.670e-01  1.684e-01  -1.585  0.11678
## Hep83              6.708e-02  9.246e-02   0.726  0.47018
## MCV1               7.815e-02  8.395e-02   0.931  0.35458
## Pol3               3.545e-02  1.511e-01   0.235  0.81504
## DTP3              -1.188e-01  1.540e-01  -0.771  0.44274
## PCGMP              6.264e-05  3.014e-05   2.078  0.04079 *
## Pop               -4.928e-10  1.860e-09  -0.265  0.79173
## EduAvg             3.907e-01  2.669e-01   1.464  0.14691
## P_Ciga            -3.299e-02  1.417e-01  -0.233  0.81654
## Tax_Ciga           3.721e+00  2.072e+00   1.796  0.07619 .
## HDL_Chol           1.858e+00  3.954e+00   0.470  0.63965
## Cal_Daily          3.080e-03  1.155e-03   2.667  0.00920 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.08 on 83 degrees of freedom
##   (77 observations deleted due to missingness)
## Multiple R-squared:  0.8623, Adjusted R-squared:  0.8257
## F-statistic: 23.62 on 22 and 83 DF,  p-value: < 2.2e-16
```

5.1.2 Backward Elimination & AIC score of each round

R1 - remove highest: PM2.5

```
## AIC test:  559.4082
```

R2 - remove highest: Pol3

```
## AIC test:  557.4869
```

R3 - remove highest: Pop

```
## AIC test:  555.5546
```

R4 - remove highest: P_Ciga

```
## AIC test:  553.6612
```

R5 - remove highest: R_Bld_P

```
## AIC test:  551.8695
```

R6 - remove highest: HDL_Chol

```
## AIC test:  550.1706
```

R7 - remove highest: DTP3

```
## AIC test:  548.9232
```

R8 - remove highest: Hep83

```
## AIC test:  570.3777
```

5.1.3 Final Optimized Model of Approach 1

```
##
## Call:
## lm(formula = Life_Exp ~ Acs_Water_Service + BMI18 + BMI25 + BMI30 +
##     Diab + Hyp_Ten + Insuf_Phy + N_HDL_Chol + Alc_pcgmp + MCV1 +
##     PCGMP + EduAvg + Tax_Ciga + Cal_Daily + Hep83, data = df_184)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5045 -1.5581  0.0904  1.9251  6.0983
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.250e+01  4.872e+00   6.672    2e-09 ***
## Acs_Water_Service  6.431e-02  4.651e-02   1.383  0.17018
## BMI18              1.936e-01  1.657e-01   1.168  0.24571
## BMI25              2.657e-01  1.121e-01   2.371  0.01989 *
## BMI30             -2.566e-01  1.370e-01  -1.873  0.06426 .
## Diab              -1.821e-01  8.029e-02  -2.269  0.02568 *
## Hyp_Ten           -1.865e-01  5.536e-02  -3.368  0.00112 **
## Insuf_Phy          9.480e-02  3.457e-02   2.742  0.00737 **
## N_HDL_Chol         3.650e+00  1.190e+00   3.068  0.00284 **
## Alc_pcgmp         -2.757e-01  1.415e-01  -1.948  0.05451 .
## MCV1               4.066e-02  6.146e-02   0.662  0.50998
## PCGMP              6.613e-05  2.171e-05   3.045  0.00305 **
## EduAvg             3.831e-01  2.307e-01   1.661  0.10029
## Tax_Ciga           3.845e+00  1.944e+00   1.978  0.05102 .
## Cal_Daily          3.089e-03  1.096e-03   2.817  0.00595 **
## Hep83              2.613e-02  5.855e-02   0.446  0.65648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.98 on 90 degrees of freedom
##   (77 observations deleted due to missingness)
## Multiple R-squared:  0.8602, Adjusted R-squared:  0.8369
## F-statistic: 36.92 on 15 and 90 DF,  p-value: < 2.2e-16
```

5.1.4 Model Comparison Metrics of Approach 1

```
significant_C <- tidy(BE_com_lmod) %>%
  filter(p.value < 0.05) %>%
  select(term, p.value)

cat("Model Comparison Result of Models:", "\n\n")

cat("1) Model with Only Complete Case","\n\n")
cat("   Adjusted R squared: ", glance(BE_com_lmod)$adj.r.squared , "\n")
cat("   AIC score:          ",AIC(BE_com_lmod),"\n")
cat("   BIC score:          ",BIC(BE_com_lmod),"\n\n")
cat("   Predictor with p-value< 0.05:","\n")
print(significant_C)
```

```
1) Model with Only Complete Case

   Adjusted R squared:  0.8368908
   AIC score:           548.9232
   BIC score:           594.2017
```

5.2 Approach 2: Modelling by Imputation of Mean

5.2.1 Calculation of Mean of Each Column

Life_Exp	Acs_Water_Service	BMI18	BMI25
7.254951e+01	8.761851e+01	6.175301e+00	4.905760e+01
BMI30	**Diab**	**Hyp_Ten**	**Insuf_Phy**
2.074743e+01	1.262568e+01	3.736940e+01	2.669885e+01
N_HDL_Chol	**PM2.5**	**R_Bld_P**	**Alc_pcgmp**
3.299448e+00	2.215230e+01	3.019820e+01	4.646542e+00
Hep83	**MCV1**	**Pol3**	**DTP3**
8.740449e+01	8.715301e+01	8.806557e+01	8.819126e+01
PCGMP	**Pop**	**EduAvg**	**P_Ciga**
1.411533e+04	4.175932e+07	9.250714e+00	3.888497e+00
Tax_Ciga	**HDL_Chol**	**Cal_Daily**	
5.380925e-01	1.203086e+00	2.913538e+03	

5.2.2 Initial Model After Imputation

```
##
## Call:
## lm(formula = Life_Exp ~ ., data = m_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9171  -1.7910   0.1042   1.8640   9.0493
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.752e+01  4.523e+00   8.296 4.21e-14 ***
## Acs_Water_Service   1.204e-01  2.978e-02   4.042 8.21e-05 ***
## BMI18              -8.234e-02  1.040e-01  -0.792 0.429662
## BMI25               1.652e-01  7.495e-02   2.204 0.028946 *
## BMI30              -2.454e-01  7.927e-02  -3.096 0.002317 **
## Diab               -1.330e-01  6.470e-02  -2.056 0.041440 *
## Hyp_Ten            -1.125e-01  1.144e-01  -0.984 0.326761
## Insuf_Phy           1.117e-01  2.986e-02   3.740 0.000256 ***
## N_HDL_Chol          3.156e+00  9.843e-01   3.206 0.001624 **
## PM2.5               1.374e-02  2.380e-02   0.577 0.564593
## R_Bld_P            -3.852e-03  1.051e-01  -0.037 0.970810
## Alc_pcgmp           3.132e-02  1.007e-01   0.311 0.756293
## Hep83               7.598e-03  7.258e-02   0.105 0.916754
## MCV1                8.119e-03  3.812e-02   0.213 0.831602
## Pol3                1.267e-01  9.850e-02   1.286 0.200201
## DTP3               -6.731e-02  1.139e-01  -0.591 0.555237
## PCGMP               6.119e-05  2.334e-05   2.621 0.009603 **
## Pop                -3.244e-10  1.704e-09  -0.190 0.849258
## EduAvg              1.902e-01  1.712e-01   1.111 0.268266
## P_Ciga              1.080e-01  1.076e-01   1.004 0.317034
## Tax_Ciga            3.639e+00  1.297e+00   2.806 0.005647 **
## HDL_Chol           -1.396e+00  2.766e+00  -0.505 0.614502
## Cal_Daily           1.452e-03  8.960e-04   1.621 0.107001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.081 on 160 degrees of freedom
## Multiple R-squared:  0.8428, Adjusted R-squared:  0.8212
## F-statistic: 38.99 on 22 and 160 DF,  p-value: < 2.2e-16
```

5.2.3 Backward Elimination & AIC score of each round

R1 - remove highest: R_Bld_P

```
## AIC test:  952.5885
```

R2 - remove highest: Hep83

```
## AIC test:  950.6005
```

R3 - remove highest: Pop

```
## AIC test:  948.6471
```

R4 - remove highest: MCV1

```
## AIC test:  946.7159
```

R5 - remove highest: Alc_pcgmp

```
## AIC test:  944.835
```

R6 - remove highest: HDL_Chol

```
## AIC test:  943.076
```

R7 - remove highest: PM2.5

```
## AIC test:  941.2992
```

R8 - remove highest: DTP3

```
## AIC test:  939.6415
```

R9 - remove highest: BMI18

```
## AIC test:  938.275
```

R10 - remove highest: P_Ciga

```
## AIC test:  937.0903
```

R11 - remove highest: EduAvg

```
## AIC test:  936.4524
```

R12 - remove highest: Cal_Daily

```
## AIC test:  938.291
```

AIC increases, so we add back Cal_Daily

5.2.4 Final Optimized Model of Approach 2

```
##
## Call:
## lm(formula = Life_Exp ~ Acs_Water_Service + BMI25 + BMI30 + Diab +
##     Hyp_Ten + Insuf_Phy + N_HDL_Chol + Pol3 + PCGMP + Tax_Ciga +
##     Cal_Daily, data = m_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8964  -1.7093   0.2057   1.7265   8.9824
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.567e+01  3.187e+00  11.191  < 2e-16 ***
## Acs_Water_Service  1.262e-01  2.742e-02   4.605 8.03e-06 ***
## BMI25              2.047e-01  5.261e-02   3.891 0.000143 ***
## BMI30             -2.547e-01  6.879e-02  -3.703 0.000287 ***
## Diab              -1.429e-01  5.301e-02  -2.696 0.007730 **
## Hyp_Ten           -1.270e-01  3.862e-02  -3.288 0.001224 **
## Insuf_Phy          1.064e-01  2.372e-02   4.487 1.32e-05 ***
## N_HDL_Chol         3.119e+00  8.686e-01   3.591 0.000430 ***
## Pol3               7.625e-02  2.260e-02   3.375 0.000915 ***
## PCGMP              7.411e-05  1.643e-05   4.511 1.20e-05 ***
## Tax_Ciga           3.888e+00  1.214e+00   3.202 0.001628 **
## Cal_Daily          1.587e-03  8.337e-04   1.904 0.058604 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.012 on 171 degrees of freedom
## Multiple R-squared:  0.8394, Adjusted R-squared:  0.8291
## F-statistic: 81.28 on 11 and 171 DF,  p-value: < 2.2e-16
```

5.2.5 Model Comparison Metrics of Approach 2

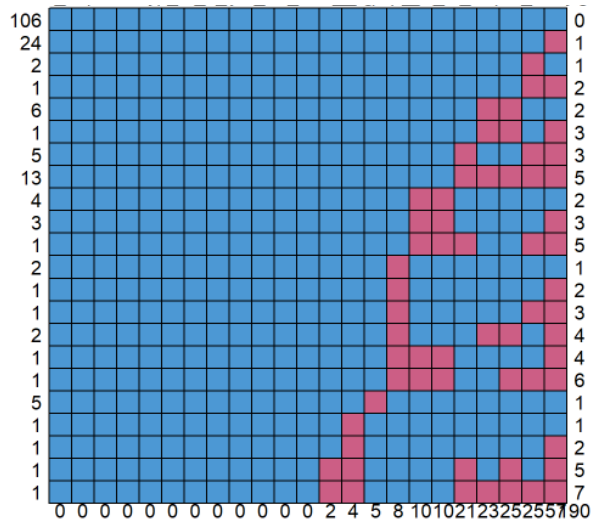
```
significant_M <- tidy(BE_M_lmod) %>%
  filter(p.value < 0.05) %>%
  select(term, p.value)

cat("2) Model imputed with mean","\n\n")
cat("   Adjusted R squared: ", glance(BE_M_lmod)$adj.r.squared , "\n")
cat("   AIC score:          ",AIC(BE_M_lmod),"\n")
cat("   BIC score:          ",BIC(BE_M_lmod),"\n\n")
cat("   Predictor with p-value< 0.05:","\n")
print(significant_M)
```

```
2) Model imputed with mean

   Adjusted R squared:  0.8291135
   AIC score:           936.4524
   BIC score:           978.1757
```

5.3 Approach 3: Modelling by Imputation of Mean

5.3.1 Observation of Missing Value Pattern

Pattern of Missing Values



Detailed Table of Missing Values Distribution in Data Set

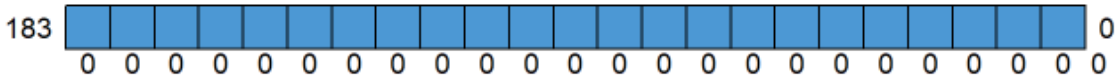
	Life_Exp	BMI18	BMI25	BMI30	Diab	Hyp_Ten	Insuf_Phy	PM2.5	R_Bld_P	MCV1	Pol3
106	1	1	1	1	1	1	1	1	1	1	1
24	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1
	0	0	0	0	0	0	0	0	0	0	0

	DTP3	N_HDL_Chol	Alc_pcgmp	Hep83	Acs_Water_Service	P_Ciga	Tax_Ciga	HDL_Chol	Pop	PCGMP	Cal_Daily	EduAvg	
106	1	1	1	1	1	1	1	1	1	1	1	1	0
24	1	1	1	1	1	1	1	1	1	1	1	0	1
2	1	1	1	1	1	1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1	1	1	1	0	0	2
6	1	1	1	1	1	1	1	1	0	0	1	1	2
1	1	1	1	1	1	1	1	1	0	0	1	0	3
5	1	1	1	1	1	1	1	0	1	1	0	0	3
13	1	1	1	1	1	1	1	0	0	0	0	0	5
4	1	1	1	1	1	0	0	1	1	1	1	1	2
3	1	1	1	1	1	0	0	1	1	1	1	0	3
1	1	1	1	1	1	0	0	0	1	1	0	0	5
2	1	1	1	1	0	1	1	1	1	1	1	1	1
1	1	1	1	1	0	1	1	1	1	1	1	0	2
1	1	1	1	1	0	1	1	1	1	1	0	0	3
2	1	1	1	1	0	1	1	1	0	0	1	0	4
1	1	1	1	1	0	0	0	1	1	1	1	0	4
1	1	1	1	1	0	0	0	1	1	0	0	0	6
5	1	1	1	0	1	1	1	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1	1	1	1	0	2
1	1	0	0	1	1	1	1	0	1	0	1	0	5
1	1	0	0	1	1	1	1	0	0	0	0	0	7
	0	2	4	5	8	10	10	21	23	25	25	57	190

5.3.2 Comparison of Distribution Between Original & 2 Sets of Imputed Data

Please refer to Appendix 1. for detailed results.

5.3.3 Final Imputation Result




```
##     Life_Exp Acs_Water_Service BMI18 BMI25 BMI30 Diab Hyp_Ten Insuf_Phy
## 183        1                 1     1     1     1    1       1         1
##            0                 0     0     0     0    0       0         0
##     N_HDL_Chol PM2.5 R_Bld_P Alc_pcgmp Hep83 MCV1 Pol3 DTP3 PCGMP Pop EduAvg
## 183          1     1       1         1     1    1    1    1     1   1      1
##              0     0       0         0     0    0    0    0     0   0      0
##     P_Ciga Tax_Ciga HDL_Chol Cal_Daily
## 183      1        1        1         1 0
##          0        0        0         0 0
```

5.3.4 Initial Model After Imputation

```
##
## Call:
## lm(formula = Life_Exp ~ ., data = imp_df_184)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2017  -1.7927   0.0147   1.9200   7.7933
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.359e+01  4.289e+00   7.831 6.43e-13 ***
## X                 -1.508e-03  4.490e-03  -0.336 0.737435
## Acs_Water_Service  1.182e-01  3.016e-02   3.920 0.000131 ***
## BMI18              5.112e-02  1.014e-01   0.504 0.615013
## BMI25              1.804e-01  7.168e-02   2.516 0.012855 *
## BMI30             -2.278e-01  7.727e-02  -2.948 0.003681 **
## Diab              -1.076e-01  6.259e-02  -1.719 0.087593 .
## Hyp_Ten           -1.536e-01  1.106e-01  -1.390 0.166594
## Insuf_Phy          1.023e-01  2.893e-02   3.537 0.000531 ***
## N_HDL_Chol         3.375e+00  9.699e-01   3.480 0.000647 ***
## PM2.5              6.158e-03  2.342e-02   0.263 0.792897
## R_Bld_P            1.277e-02  1.016e-01   0.126 0.900165
## Alc_pcgmp          3.506e-02  1.005e-01   0.349 0.727631
## Hep83              1.975e-02  7.345e-02   0.269 0.788393
## MCV1               1.453e-02  3.742e-02   0.388 0.698344
## Pol3               9.764e-02  9.445e-02   1.034 0.302802
## DTP3              -6.288e-02  1.113e-01  -0.565 0.572989
## PCGMP              6.590e-05  2.343e-05   2.813 0.005528 **
## Pop               -9.224e-10  1.655e-09  -0.557 0.578142
## EduAvg            -3.546e-02  1.763e-01  -0.201 0.840790
## P_Ciga             2.640e-02  1.124e-01   0.235 0.814619
## Tax_Ciga           3.916e+00  1.252e+00   3.129 0.002089 **
## HDL_Chol           7.877e-01  2.635e+00   0.299 0.765353
## Cal_Daily          2.425e-03  8.406e-04   2.885 0.004454 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.971 on 159 degrees of freedom
## Multiple R-squared:  0.8547, Adjusted R-squared:  0.8337
## F-statistic: 40.66 on 23 and 159 DF,  p-value: < 2.2e-16
```

5.3.5 Backward Elimination & AIC score of each round

R1 - remove highest: R_Bld_P

```
## AIC test:  940.2198
```

R2 - remove highest:EduAvg

```
## AIC test:  938.2682
```

R3 - remove highest: P_Ciga

```
## AIC test:  936.3316
```

R4 - remove highest: HDL_Chol

```
## AIC test:  934.4028
```

R5 - remove highest: PM2.5

```
## AIC test:  932.4947
```

R6 - remove highest: MCV1

```
## AIC test:  930.5932
```

R7 - remove highest: Alc_pcgmp

```
## AIC test:  928.6987
```

R8 - remove highest: BMI18

```
## AIC test:  926.8999
```

R9 - remove highest: Hep83

```
## AIC test:  925.172
```

R10 - remove highest: DTP3

```
## AIC test:  923.3664
```

R11 - remove highest: Pop

```
## AIC test:  921.6088
```

5.3.6 Final Optimized Model of Approach 3

```
##
## Call:
## lm(formula = Life_Exp ~ Acs_Water_Service + BMI25 + BMI30 + Diab +
##     Hyp_Ten + Insuf_Phy + N_HDL_Chol + Pol3 + PCGMP + Tax_Ciga +
##     Cal_Daily, data = imp_df_184)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2613  -1.7037   0.0616   1.8719   7.9963
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.514e+01  2.967e+00  11.842  < 2e-16 ***
## Acs_Water_Service  1.151e-01  2.647e-02   4.346 2.37e-05 ***
## BMI25              1.607e-01  5.140e-02   3.127 0.002077 **
## BMI30             -2.157e-01  6.650e-02  -3.244 0.001418 **
## Diab              -1.059e-01  5.097e-02  -2.078 0.039238 *
## Hyp_Ten           -1.311e-01  3.722e-02  -3.523 0.000548 ***
## Insuf_Phy          1.014e-01  2.267e-02   4.472 1.41e-05 ***
## N_HDL_Chol         3.366e+00  8.354e-01   4.029 8.41e-05 ***
## Pol3               6.817e-02  2.139e-02   3.187 0.001708 **
## PCGMP              7.224e-05  1.564e-05   4.618 7.60e-06 ***
## Tax_Ciga           3.888e+00  1.163e+00   3.343 0.001020 **
## Cal_Daily          2.482e-03  7.964e-04   3.116 0.002148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.878 on 171 degrees of freedom
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8439
## F-statistic: 90.46 on 11 and 171 DF,  p-value: < 2.2e-16
```

5.3.7 Model Comparison Metrics of Approach 3

```
significant_R <- tidy(BE_imp_lmod) %>%
  filter(p.value < 0.05) %>%
  select(term, p.value)

cat("3) Model imputed with Regression","\n\n")
cat("   Adjusted R squared: ", glance(BE_imp_lmod)$adj.r.squared , "\n")
cat("   AIC score:          ",AIC(BE_imp_lmod),"\n")
cat("   BIC score:          ",BIC(BE_imp_lmod),"\n\n")
cat("   Predictor with p-value< 0.05:","\n")
print(significant_R)
```

```
 3) Model imputed with Regression

    Adjusted R squared:  0.8439117
    AIC score:           919.8766
    BIC score:           961.5999
```

5.4 Model Comparison and Selection of Final Models

Summary of Model Comparison Metrics for All 3 Approaches

Metrics	Original	Imputed Approaches	
	A1: Only Complete Cases	A2: Imputation By Mean	A3: Imputation By Regression
Adopted Observations	106	183	183
Adjusted R squared	0.8369	0.8291135	0.8439117
AIC	548.9232	936.4524	919.8766
BIC	594.2017	978.1757	961.5999
Predictors with p-value < 0.05	1. BMI25 2. Diab 3. Insuf_Phy 4. N_HDL_Chol 5. PCGMP 6. Hyp_Ten 7. Cal_Daily	1. BMI25 2. Diab 3. Insuf_Phy 4. N_HDL_Chol 5. PCGMP 6. Hyp_Ten 7. Acs_Water_Service 8. BMI30 9. Pol3 10. Tax_Ciga	1. BMI25 2. Diab 3. Insuf_Phy 4. N_HDL_Chol 5. PCGMP 6. Hyp_Ten 7. Cal_Daily 8. Acs_Water_Service 9. BMI30 10. Pol3 11. Tax_Ciga

Interpretation of Metrics:

- Adjusted R squared value: A value closer to 1 indicates a better model
- AIC score: A lower value indicates a better model
- BIC score: A lower value indicates a better model

First, by comparing the two imputation approaches, we can observe that A3 - Imputation By Regression, performs better than A2 - Imputation By Mean, with a better value in all three metrics (i.e. Higher Adjusted R squared value, lower AIC & lower BIC score.).

Next, comparing A3 with A1 - the original model, we can observe that A1 has a better AIC & BIC score than A3. However, it is worth noting that a complete case data set with fewer observations would usually have a lower AIC score as a result of reduced noise and variability in a cleaner data set. But it also sacrifices the potentially informative observations and potentially related predictors.

Focusing on the patterns of statistically significant predictors, we can observe that A3 covers all the 7 predictors shown in A1 (i.e. highlighted in yellow) and additionally introduces 4 more predictors (i.e.

highlighted in green) that shall improve the model. Considering the significant increase in observations (i.e. 106 observations in A1 vs. 183 observations in A3), A3 should be more representative on pattern of the data.

Thus, A3 - Imputation By Regression is chosen the be our final model of the analysis and would be used for the rest of this analysis.

5.5 Interpretation of Model Result

5.5.1 Coefficient Interpretation

```
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.514e+01  2.967e+00  11.842  < 2e-16 ***
## Acs_Water_Service   1.151e-01  2.647e-02   4.346 2.37e-05 ***
## BMI25               1.607e-01  5.140e-02   3.127 0.002077 **
## BMI30              -2.157e-01  6.650e-02  -3.244 0.001418 **
## Diab               -1.059e-01  5.097e-02  -2.078 0.039238 *
## Hyp_Ten            -1.311e-01  3.722e-02  -3.523 0.000548 ***
## Insuf_Phy           1.014e-01  2.267e-02   4.472 1.41e-05 ***
## N_HDL_Chol          3.366e+00  8.354e-01   4.029 8.41e-05 ***
## Pol3                6.817e-02  2.139e-02   3.187 0.001708 **
## PCGMP               7.224e-05  1.564e-05   4.618 7.60e-06 ***
## Tax_Ciga            3.888e+00  1.163e+00   3.343 0.001020 **
## Cal_Daily           2.482e-03  7.964e-04   3.116 0.002148 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Discussion on the coefficient can be based on the final result summary of Model A3 - Imputation by Regression. After Backward Elimination, we have obtained 11 predictors that show a statistically significant relationship with major Y response Life Expectancy, in which their p-value of F-Test is lower than 0.05. Within the context of our research, the meaning of these coefficients can be interpreted as follows:

Predictors with Negative Linear Relationship

- **BMI30 (Obesity)**

  With every 1% increase in the prevalence of obesity reduces life expectancy by 0.2547 years.

- **Diabetes (Diab)**

  With every 1% increase in the prevalence of diabetes, life expectancy decreases by 0.1059 years.

- **Hypertension Prevalence (Hyp_Ten)**

With every 1% increase in prevalence of hypertension decreases life expectancy by 0.1270 years, likely due to its cardiovascular impact.

Predictors with Positive Linear Relationship

- **Access to Water Services (Acs_Water_Service)**

  With every 1% increase in the population with access to basic water service corresponds to a 0.1151-year increase in life expectancy.

- **BMI25 (Overweight)**

  With every 1% increase in the overweight population increases life expectancy by 0.2047 years, suggesting better nutrition and healthcare access.

- **Non-HDL Cholesterol (N_HDL_Chol)**

  Every 1 mmol/L increase in non-HDL cholesterol raises life expectancy by 3.119 years.

- **Polio (Pol3) immunization coverage among 1-year-old (%)**

  With every 1% increase in vaccination rate, life expectancy increases by 0.6817 years, suggesting that vaccinations do have health benefits.

- **GDP per capita (PCGMP)**

  With every 1 dollar increase in GDP (in USD), life expectancy increases by 0.7224 years.

- **Insufficient amount of exercise (Insuf_Phy)**

  With every 1% increase in exercise among adults over the age of 18, life expectancy increases by 0.1014 years.

- **Percent of Tax Cigarette (Tax_Ciga)**

  With every 1% increase in the tax on cigarettes, life expectancy increases by 0.3888 years. This may suggest higher taxes on cigarettes reduce people's motivation to purchase cigarettes.

- **Daily Calories Supply per person in kcal (Cal_Daily)**

  With every 1 kcal intake increase in daily supply of calories per person, life expectancy increases by 0.002482 years.

Unexpected Result

Some predictors exhibit unexpected or non-linear relationships with life expectancy.

- **Population (Pop):** The scatterplot, histogram and regression coefficient indicate no significant association, which contrast to the theoretical expectations of population size influencing resource distribution, therefore, life expectancy.
- **Air Pollution (PM2.5):** despite health risks, no significant relationship was observed, indicating that other variables may have an indirect impact on its effectiveness.

Confidence Levels (90% and 95%)

Most variables remained statistically significant at both 90% and 95% confidence levels, including BMI25, HDL Cholesterol, Diabetes Prevalence, Insufficient Physical Activity, and Tobacco Taxation. However, Daily Caloric Intake (Cal_Daily) was solely significant at 90%, demonstrating weaker evidence for its linear relationship with life expectancy.

5.5.2 Confidence Interval of Coefficient

The table below shows the 95% confidence interval of coefficients:

	2.5 %	97.5 %
(Intercept)	29.2785348223	40.9923939051
Acs_Water_Service	0.0627950353	0.1673142140
BMI25	0.0592553284	0.2621684649
BMI30	-0.3470057000	-0.0844593405
Diab	-0.2064884806	-0.0052847459
Hyp_Ten	-0.2045621606	-0.0576352847
Insuf_Phy	0.0566331572	0.1461187479
N_HDL_Chol	1.7168696144	5.0148074487
Pol3	0.0259494114	0.1103857119
PCGMP	0.0000413624	0.0001031265
Tax_Ciga	1.5920075539	6.1843656971
Cal_Daily	0.0009098338	0.0040539710

5.6 Model Validation

There are multiple essential assumption tests for validating the Linear Regression Model. Thus, we decided to perform the following tests/plots on our selected model:

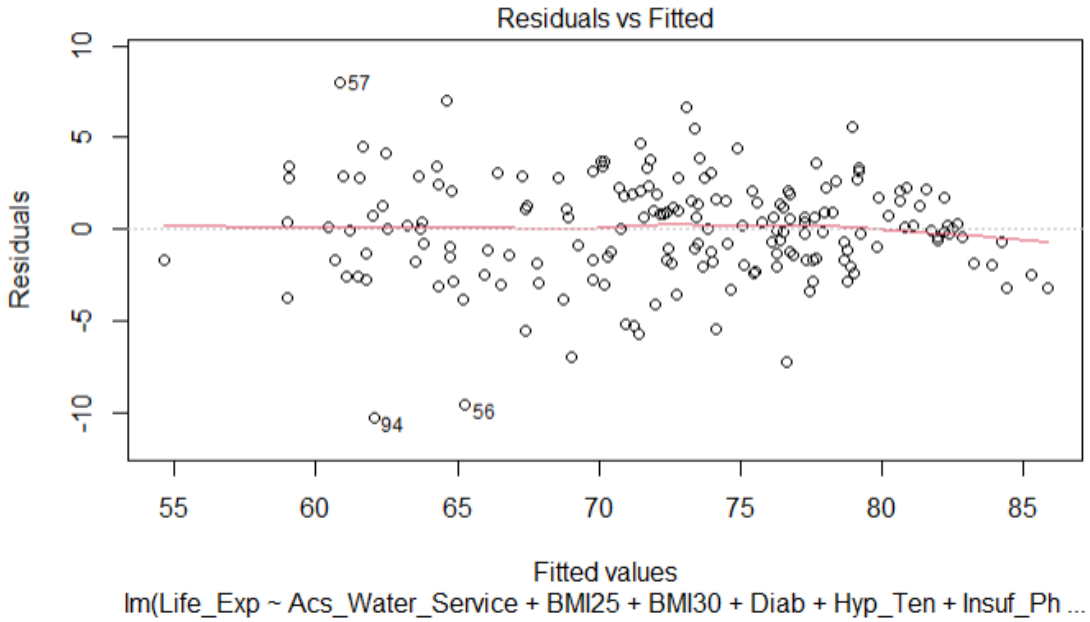
1) Dublin-Watson Test:
   To validate that predictors are independent and observed with negligible error.
2) Residuals Vs Fitted Plot
   To validate that residual errors have a mean of zero
3) Scale-Location Plot
   To validate that residual errors have a constant variance
4) Shapiro-Wilk Test
   To validate that the data is normal distributed.

1) Dublin-Watson Test

```
        Durbin-Watson test

data:  test_mod
DW = 2.0894, p-value = 0.7303
alternative hypothesis: true autocorrelation is greater than 0
```

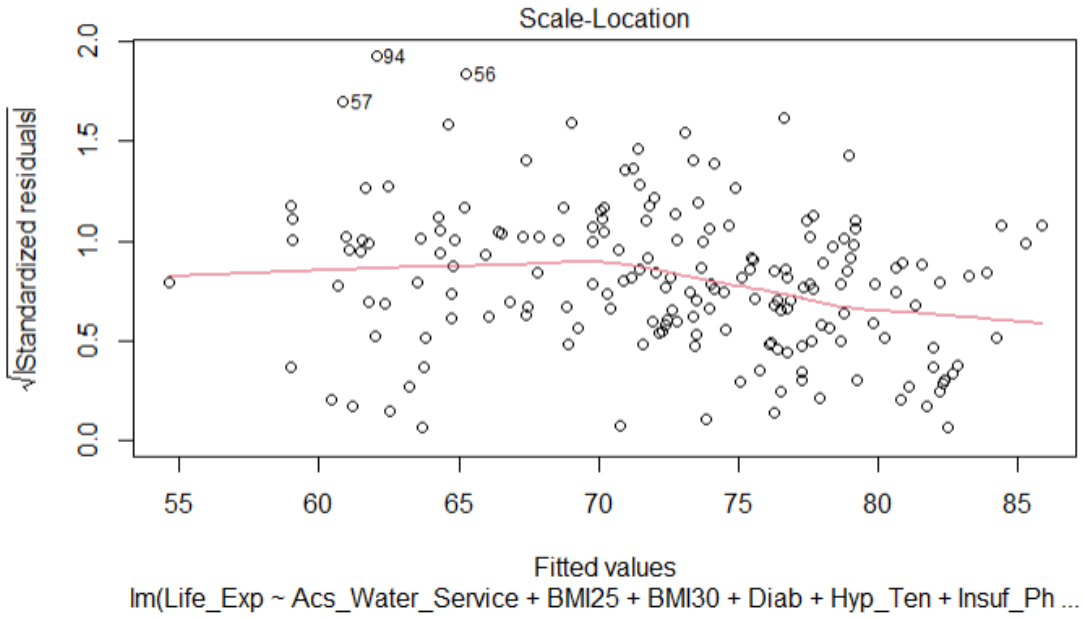
The result shows that our model has a DW statistic of 2.0894 and p-value of 0.7303, which is larger than 0.05. Thus, we can reject the null hypothesis and validate that our model has no significant autocorrelation.

2) Residuals Vs Fitted Plot



From the graph, we can see the lowess smoother (i.e. line in red) is approximately horizontal at zero. Thus, it validates that the residual errors have a mean of zero.

3) Scale-Location Plot



From the graph, we can see the residual points are mostly distributed equally around the line. Transformation on our linear model may improve this result.

4) Shapiro-Wilk Test

```
        Shapiro-Wilk normality test

data:  residuals(test_mod)
W = 0.98235, p-value = 0.02063
```

From the result, the p-value is smaller than 0.05, which suggests that we failed to validate the normal distribution of data. A Transformation on the model may be attempted to improve it.

5.7 Transformation

5.7.1 Observation on predictors

Considering the nature of data set and predictors, there are two predictors that may potentially affect the distribution of data:

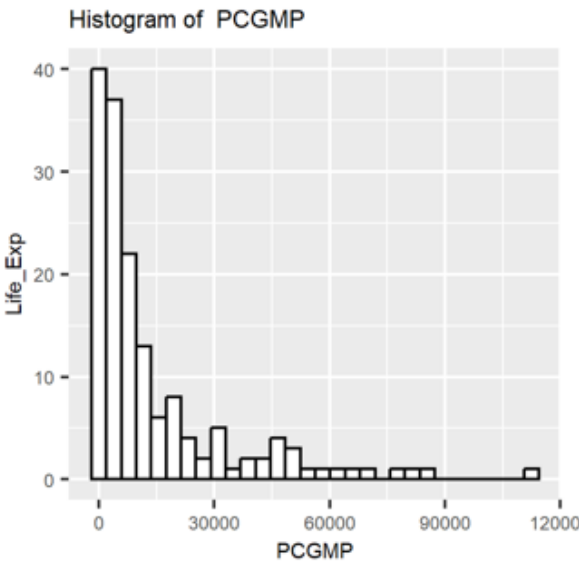
1) PCGMP

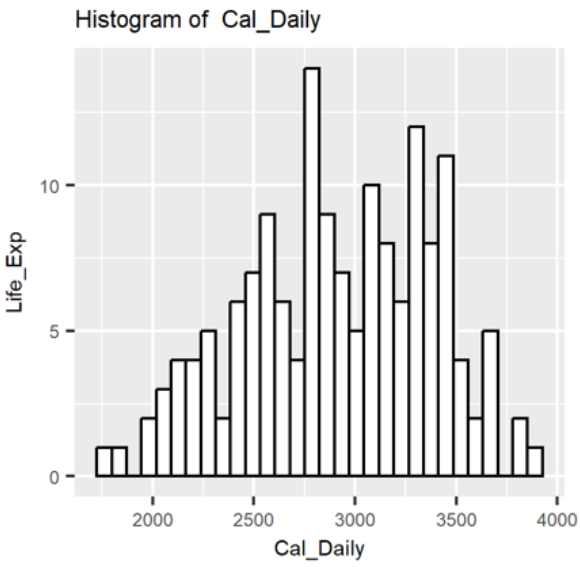
   This predictor represents the GDP per capita (in USD), which can have extreme values in the data which may affect the normal distribution of data,

2) Cal_Daily

   This predictor represents the daily calorie supply per person (in kcal) , which can also have extreme values in the data which may affect the normal distribution of data.

We can validate the assumption from the histogram visualization shown in section 2:





We can observe extreme values in the distribution of PCGMP and slight right-skewness in the distribution of Cal_Daily. Thus, the following transformation can be attempted:

1) Reciprocal transformation on PCGMP
2) Log transformation on Cal_Daily

```
recip_PCGMP <- 1/imp_df_184$PCGMP
sqrt_Cal_Daily <- sqrt(imp_df_184$Cal_Daily)

tran_BE_Imp_lmod <- lm(Life_Exp ~ (Acs_Water_Service) + BMI25 + BMI30 + Diab +
    (Hyp_Ten) + (Insuf_Phy) + (N_HDL_Chol) + (Pol3) + recip_PCGMP + (Tax_Ciga) +
sqrt_Cal_Daily, data = imp_df_184)

test_mod <- tran_BE_Imp_lmod
```

5.7.2 Transformation Result & Diagnostics

```
Call:
lm(formula = Life_Exp ~ (Acs_Water_Service) + BMI25 + BMI30 +
    Diab + (Hyp_Ten) + (Insuf_Phy) + (N_HDL_Chol) + (Pol3) +
    recip_PCGMP + (Tax_Ciga) + sqrt_Cal_Daily, data = imp_df_184)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7710 -1.8647  0.0521  1.9361  7.6427

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.68386    5.18443   4.568 9.38e-06 ***
Acs_Water_Service  0.12496    0.02913   4.290 2.98e-05 ***
BMI25              0.17711    0.05502   3.219 0.001541 **
BMI30             -0.22160    0.07090  -3.126 0.002085 **
Diab              -0.13744    0.05373  -2.558 0.011401 *
Hyp_Ten           -0.19290    0.03723  -5.181 6.16e-07 ***
Insuf_Phy          0.10295    0.02410   4.271 3.22e-05 ***
N_HDL_Chol         2.99500    0.89305   3.354 0.000982 ***
Pol3               0.06453    0.02283   2.827 0.005263 **
recip_PCGMP      134.01810  622.27596   0.215 0.829737
Tax_Ciga           4.76702    1.22382   3.895 0.000141 ***
sqrt_Cal_Daily     0.40515    0.08955   4.524 1.13e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 3.062 on 171 degrees of freedom
Multiple R-squared:  0.8341,    Adjusted R-squared:  0.8234
F-statistic: 78.15 on 11 and 171 DF,  p-value: < 2.2e-16
```

1) Shapiro-Wilk Test

```
        Shapiro-Wilk normality test

data:  residuals(test_mod)
W = 0.98554, p-value = 0.05662
```

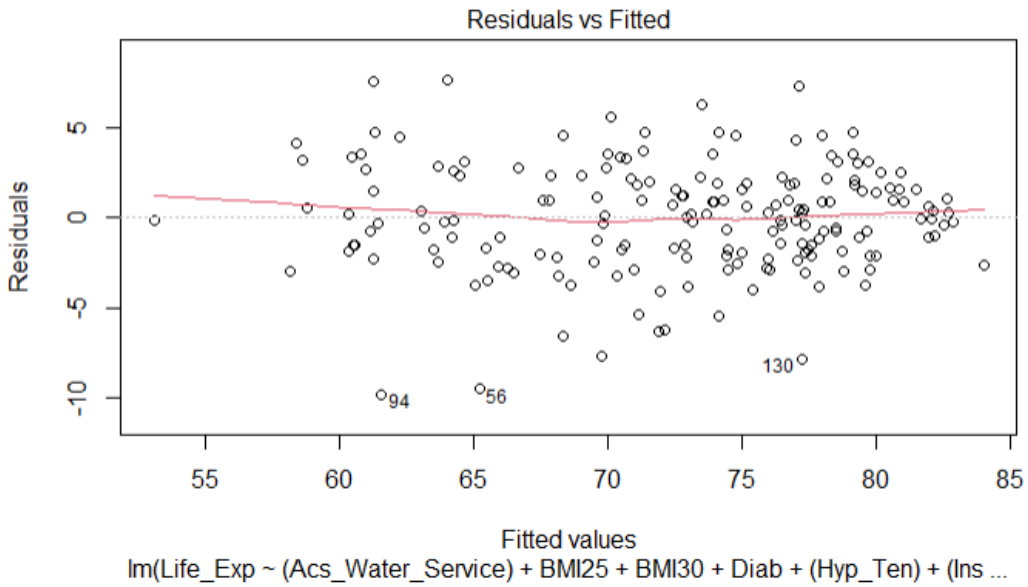
From the result, the p-value is 0.05662, which is now larger than 0.05. Thus, the transformation is successful in fitting the model into the assumption of normal distribution of data.

2) Dublin-Watson Test

```
        Durbin-Watson test

data:  test_mod
DW = 2.0614, p-value = 0.6638
alternative hypothesis: true autocorrelation is greater than 0
```

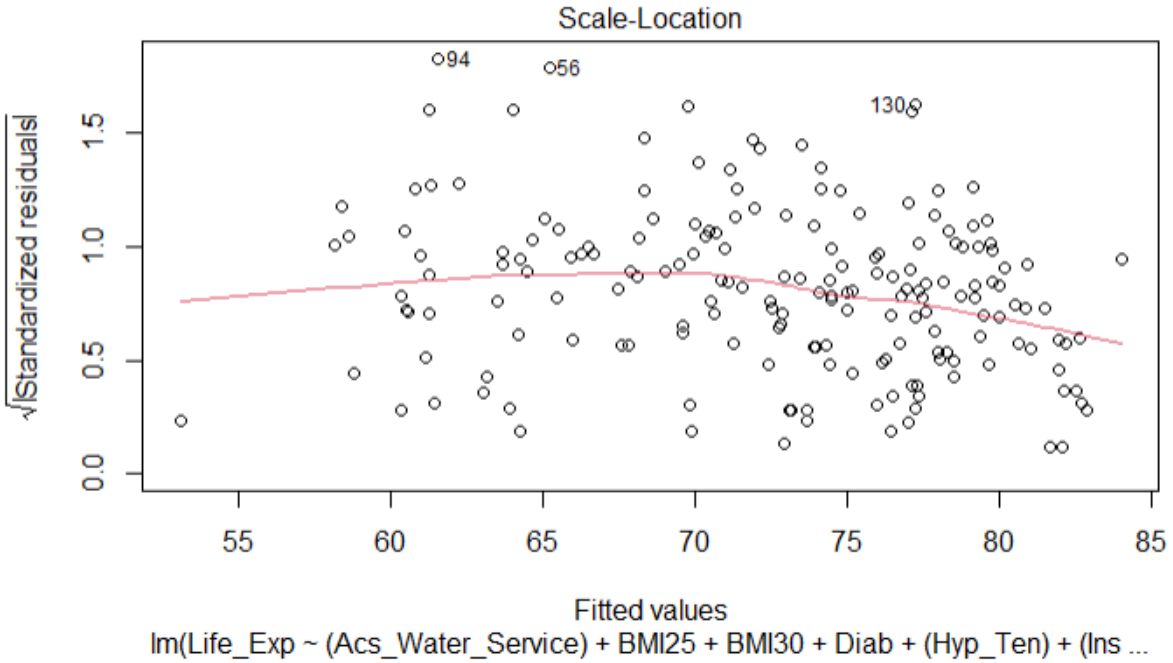
The result shows that our model has a DW statistic of 2.0614 and p-value of 0.6638, which is larger than 0.05. Thus, we can reject the null hypothesis and successfully validate that our model has no significant autocorrelation.

3) Residuals Vs Fitted Plot



From the graph, we can see the lowes smoother (i.e. line in red) is approximately horizontal at zero. Thus, it successfully validates that the residual errors have a mean of zero.

4) Scale-Location Plot



From the graph, we can see the residual points are mostly distributed equally around the line. It validates that our model should have a constant variance.

Further Discussion on predictor results after transformation

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        23.68386    5.18443   4.568 9.38e-06 ***
Acs_Water_Service   0.12496    0.02913   4.290 2.98e-05 ***
BMI25               0.17711    0.05502   3.219 0.001541 **
BMI30              -0.22160    0.07090  -3.126 0.002085 **
Diab               -0.13744    0.05373  -2.558 0.011401 *
Hyp_Ten            -0.19290    0.03723  -5.181 6.16e-07 ***
Insuf_Phy           0.10295    0.02410   4.271 3.22e-05 ***
N_HDL_Chol          2.99500    0.89305   3.354 0.000982 ***
Pol3                0.06453    0.02283   2.827 0.005263 **
recip_PCGMP       134.01810  622.27596   0.215 0.829737
Tax_Ciga            4.76702    1.22382   3.895 0.000141 ***
sqrt_Cal_Daily      0.40515    0.08955   4.524 1.13e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It is worth noting that with the reciprocal transformation on PCGMP, the predictor is no longer statistically significant in improving the linear model of Life expectancy. It may suggest that the transformation improves the performance of the model, as validation in the normality test increases the liability of the coefficient in the linear regression model.

6. Limitations and Further Research

One limitation of this study lies in the scope of our data. We restricted our analysis to a single year due to variable availability. The published statistics on the WHO's website are inconsistent, with certain variables unavailable for specific years. Access to a multi-year, balanced cross-country dataset would allow us to explore the changing trends in factors affecting life expectancy over time. Furthermore, such a dataset would enable us to investigate causal relationships between these factors and life expectancy, rather than merely identifying correlations.

7. Division of Labor

- Sheldon Zhang
  - Initial Statistical Modeling
  - Literature Review
  - Report Writing & Editing
- Vick Feng
  - Initial Statistical Modeling
  - Model Assumption Testing & Transformation
- Joshua Wong
  - Data Collection & Cleaning
  - Major Statistical Modeling & Transformation
  - Visualization of Results on R studio
- Haining Yin
  - Report Writing
- Nan Wang
  - Description of Dataset
  - Data Collection & Cleaning
  - Report Writing & Editing

References

World Health Organization (WHO). (n.d.). *Determinants of health*. World Health Organization (WHO). Retrieved October 9, 2024, from https://www.who.int/news-room/questions-and-answers/item/determinants-of-health

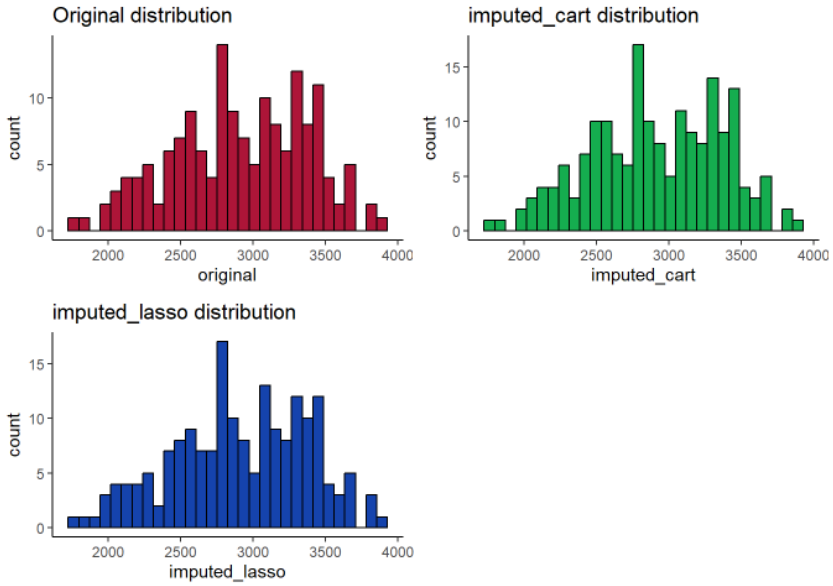
Appendix 1. Comparison Results of Distribution Between Original & 2 Sets of Imputed Data

1) imputation on Cal_Daily

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 25 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_cart would be chosen as imputation.

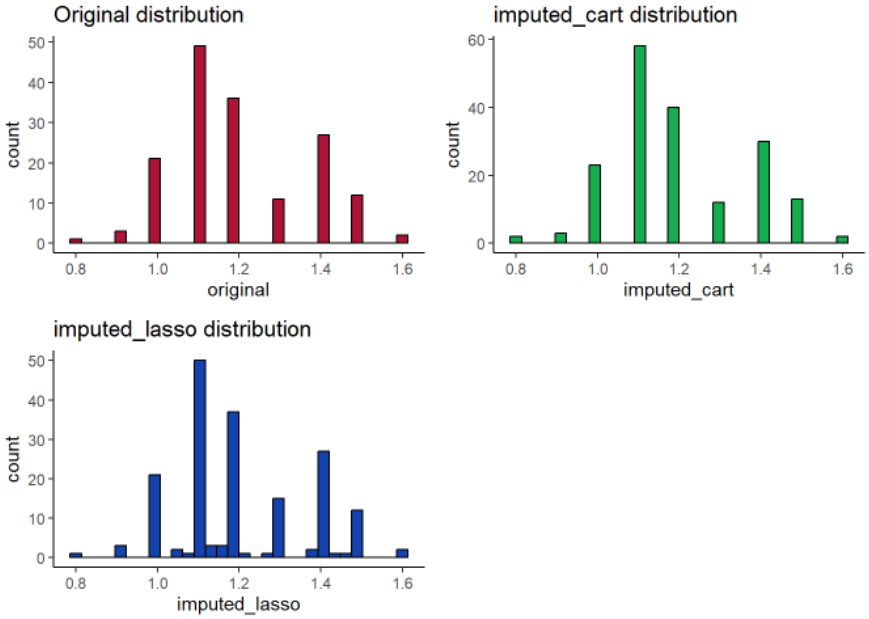
```
#replace with imputed data: Cal_Daily

Re_imp_df184 <- df_184
Re_imp_df184$Cal_Daily <- mice_imputed$imputed_cart
```

2) imputation on HDL_Chol

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 21 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_cart would be chosen as imputation.

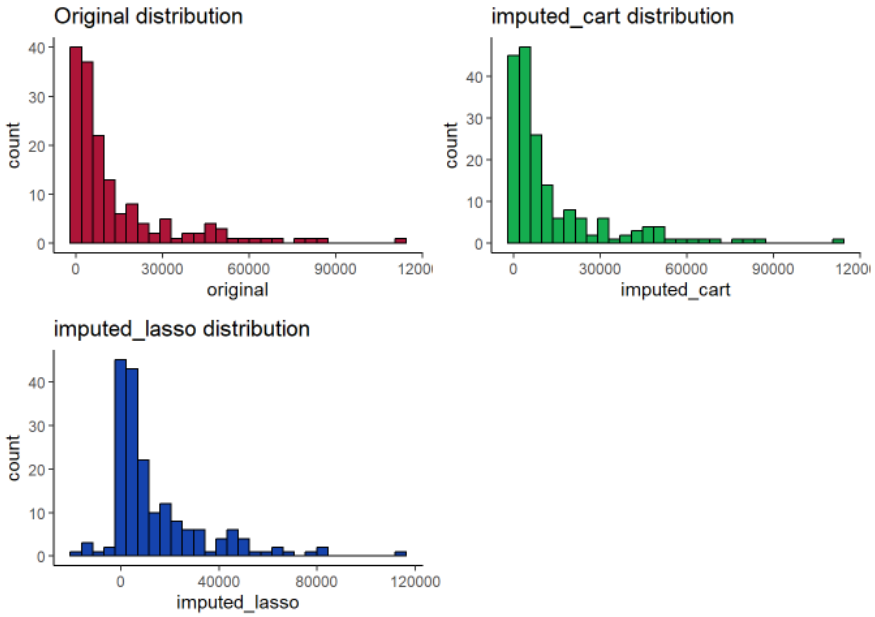
```
#replace with imputed data: HDL_Chol

Re_imp_df184$HDL_Chol <- mice_imputed$imputed_cart
# md.pattern(Re_imp_df184)
```

3) imputation on PCGMP

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 25 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_cart would be chosen as imputation.

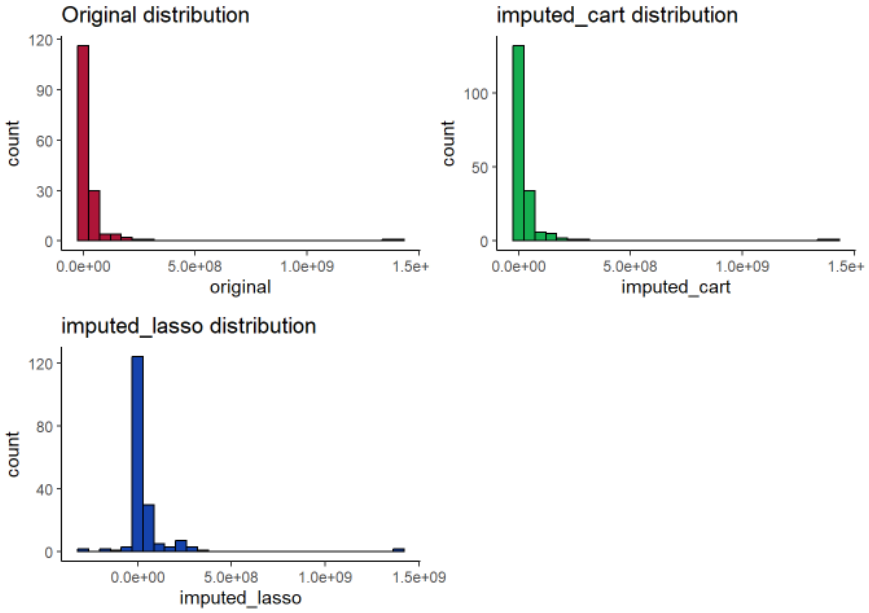
```
#replace with imputed data: PCGMP

Re_imp_df184$PCGMP <- mice_imputed$imputed_cart
md.pattern(Re_imp_df184)
```

4) imputation on Pop

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 23 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_cart would be chosen as imputation.

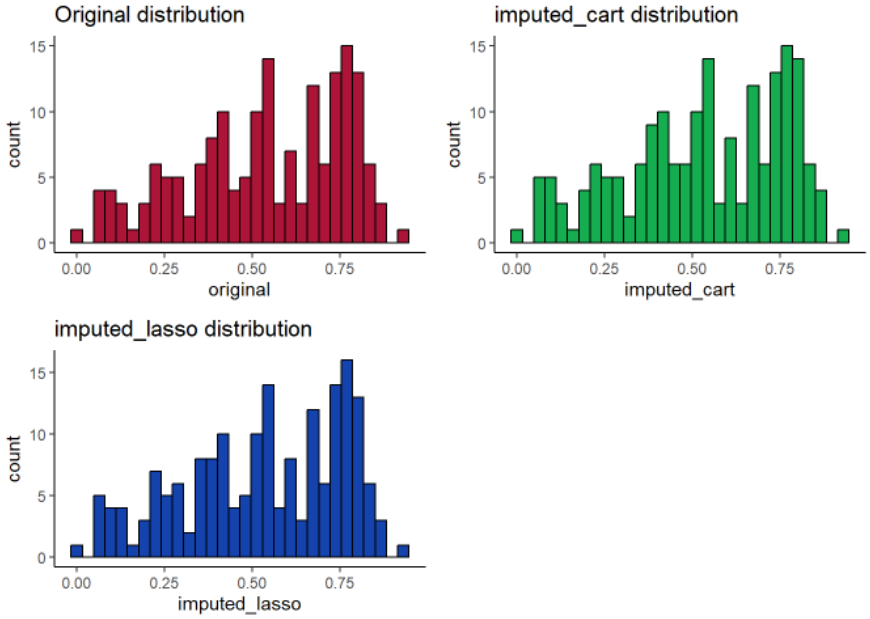
```
#replace with imputed data: Pop

Re_imp_df184$Pop <- mice_imputed$imputed_cart
# md.pattern(Re_imp_df184)
```

5) imputation on Tax_Ciga

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 10 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_lasso would be chosen as imputation.

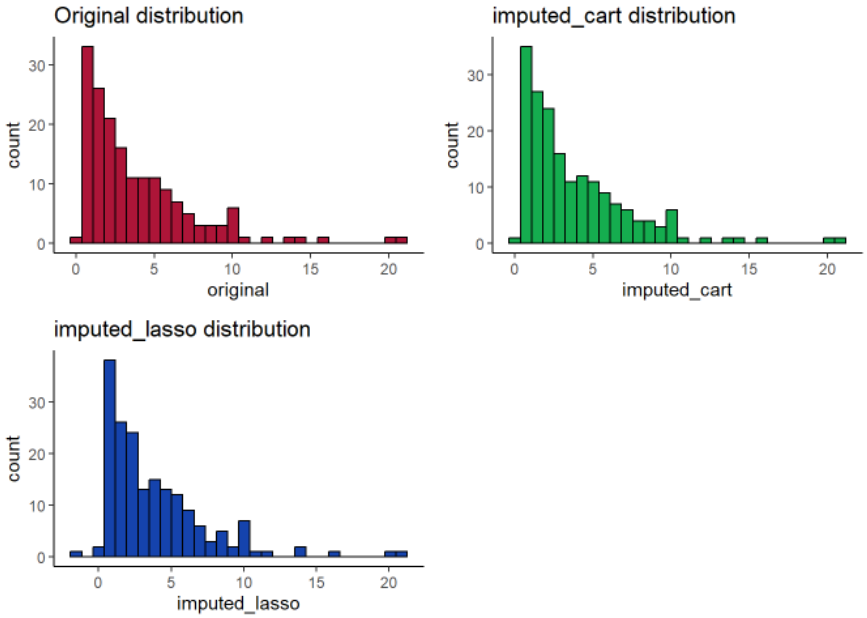
```
#replace with imputed data: Tax_Ciga

Re_imp_df184$Tax_Ciga <- mice_imputed$imputed_lasso
md.pattern(Re_imp_df184)
```

6) imputation on P_Ciga

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 10 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_cart would be chosen as imputation.

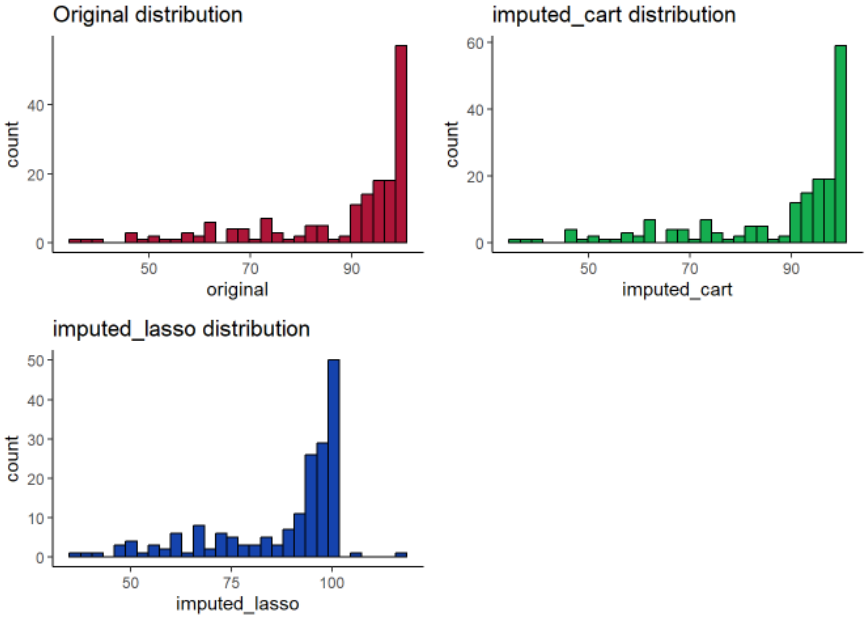
```
#replace with imputed data: P_Ciga

Re_imp_df184$P_Ciga <- mice_imputed$imputed_cart
# md.pattern(Re_imp_df184)
```

7) imputation on Acs_Water_Service

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_cart would be chosen as imputation.

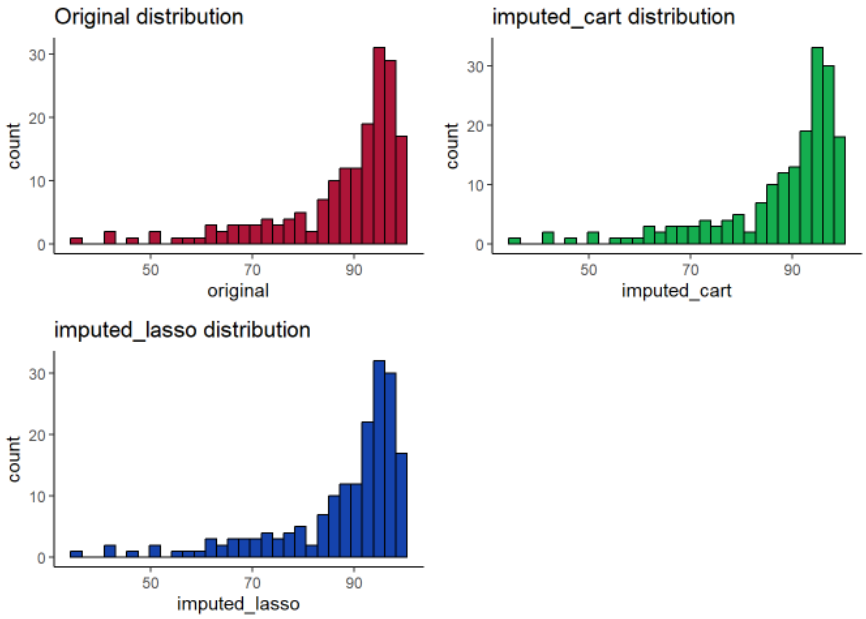
```
#replace with imputed data: Acs_Water_Service

Re_imp_df184$Acs_Water_Service <- mice_imputed$imputed_cart
md.pattern(Re_imp_df184)
```

8) imputation on Hep83

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_cart would be chosen as imputation.

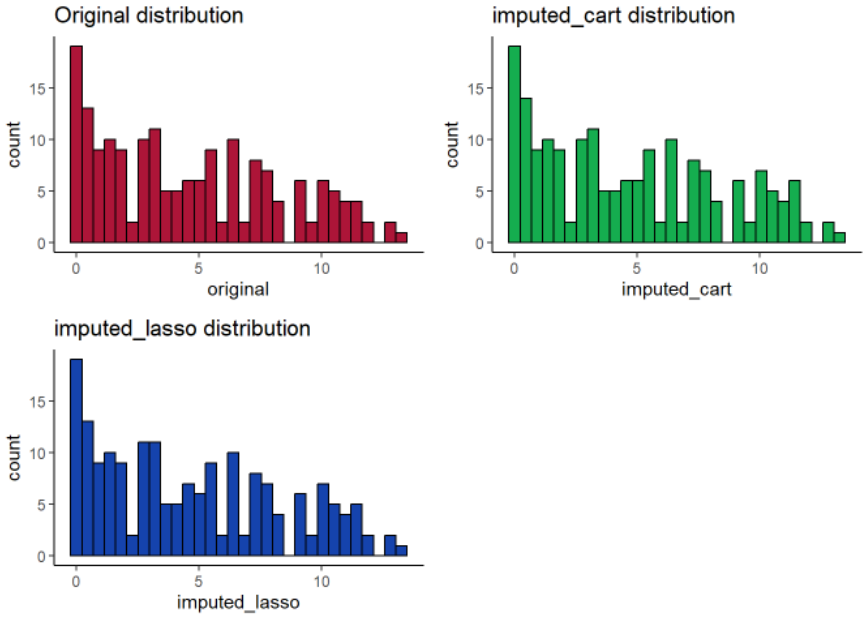
```
#replace with imputed data: Hep83

Re_imp_df184$Hep83 <- mice_imputed$imputed_cart
# md.pattern(Re_imp_df184)
```

9) imputation on Alc_pcgmp

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_lasso would be chosen as imputation.

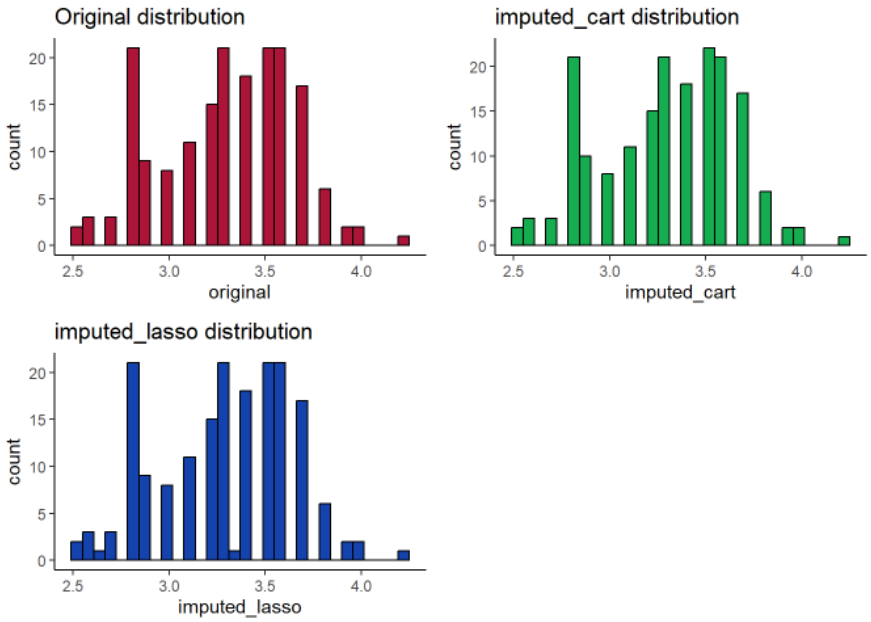
```
#replace with imputed data: Alc_pcgmp

Re_imp_df184$Alc_pcgmp <- mice_imputed$imputed_lasso
md.pattern(Re_imp_df184)
```

10) imputation on N_HDL_Chol

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_cart would be chosen as imputation.

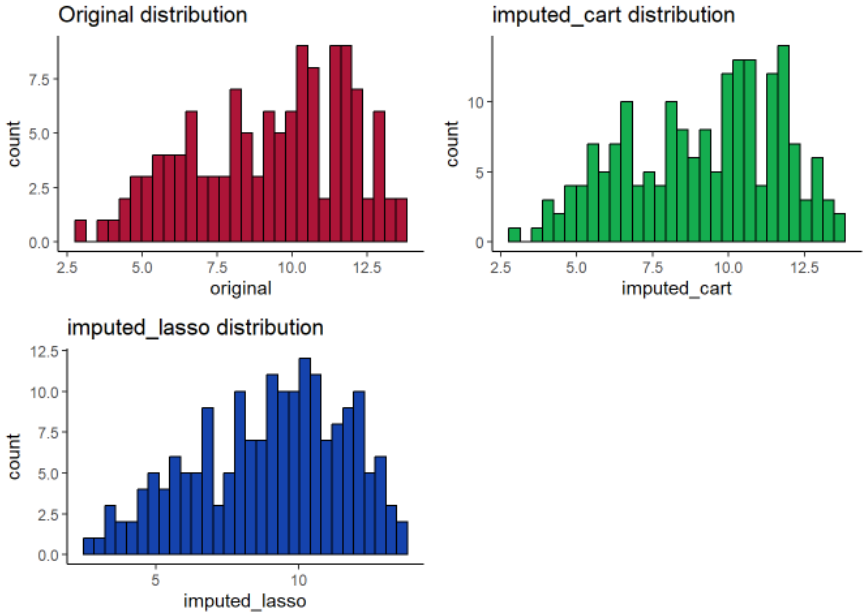
```
#replace with imputed data: N_HDL_Chol

Re_imp_df184$N_HDL_Chol <- mice_imputed$imputed_lasso
md.pattern(Re_imp_df184)
```

11) imputation on EduAvg

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 57 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From above, imputed_cart would be chosen as imputation.

```
#replace with imputed data: EduAvg

Re_imp_df184$EduAvg <- mice_imputed$imputed_cart
md.pattern(Re_imp_df184)
```