# Introduction

The dataset contains health-related indicators for different countries over multiple years. These indicators include information about life expectancy, adult mortality, infant deaths, alcohol consumption, percentage expenditure on health, and other factors. The dataset facilitates the analysis of health trends and outcomes in various countries over time.

**Potential goals for using this dataset could include:**

**Health Analysis:** Studying the overall health conditions of different countries by examining indicators like life expectancy, mortality rates, and disease prevalence.

**Identifying Patterns:** Analyzing trends and patterns in health indicators over the years to identify factors that may contribute to changes in health outcomes.

**Comparative Studies:** Comparing the health status of different countries to identify variations and similarities, and potentially understanding the factors that contribute to these differences.

**Policy Evaluation:** Assessing the effectiveness of healthcare policies and interventions by examining their impact on health indicators.

**Research:** Providing data for researchers to explore specific research questions related to public health and healthcare outcomes.

**What kind of data is included? Is it all text data, is it numerical?**

The dataset includes a mix of both numerical and categorical data. Here is a breakdown of the types of data present in the provided table:

**Numerical Data:** Life expectancy, Adult Mortality, Infant deaths, Alcohol consumption, Percentage expenditure on health, Hepatitis B,Measles,BMI (Body Mass Index), Under-five deaths,Polio,Total expenditure,Diphtheria,HIV/AIDS, GDP (Gross Domestic Product), Population, Thinness (1-19 years),

Thinness (5-9 years), Income composition of resources, Schooling.

**Categorical Data:**Country,Year,Status (indicating the development status of the country, e.g., "Developing")

Numerical data represents various health-related indicators, economic factors, and demographic statistics. The categorical data includes information about the country, its year of observation, and its development status.

**Describe the data fields including the title, the data type, the data description, etc.**

**Country:**

**Data Type:** Categorical (text)

**Description:** The name of the country.

**Year:**

**Data Type:** Numerical (integer)

**Description:** The year of observation for the health-related indicators.

**Status:**

**Data Type:** Categorical (text)

**Description:** Indicates the development status of the country, e.g., "Developing."

**Life Expectancy:**

**Data Type:** Numerical (float)

**Description:** The average number of years a newborn is expected to live, based on current mortality rates.

**Adult Mortality:**

**Data Type:** Numerical (integer)

**Description:** The probability of dying between the ages of 15 and 60 per 1000 population.

**Infant Deaths:**

**Data Type:** Numerical (integer)

**Description:** Number of infant deaths per 1,000 live births.

**Alcohol:**

**Data Type:** Numerical (float)

**Description:** Average alcohol consumption per capita (adults) in liters of pure alcohol.

**Percentage Expenditure:**

**Data Type:** Numerical (float)

**Description:** Expenditure on health as a percentage of Gross Domestic Product (GDP).

**Hepatitis B:**

**Data Type:** Numerical (integer)

**Description:** Hepatitis B immunization coverage among 1-year-olds (percentage).

**Measles:**

**Data Type:** Numerical (integer)

**Description:** Number of reported cases of measles per 1000 population.

**BMI (Body Mass Index):**

**Data Type:** Numerical (float)

**Description:** The average Body Mass Index of the adult population.

**Under-Five Deaths:**

**Data Type:** Numerical (integer)

**Description:** Number of deaths under age five per 1,000 live births.

**Polio:**

**Data Type:** Numerical (integer)

**Description:** Polio immunization coverage among 1-year-olds (percentage).

**Total Expenditure:**

**Data Type:** Numerical (float)

**Description:** General government expenditure on health as a percentage of total government expenditure.

**Diphtheria:**

**Data Type:** Numerical (integer)

**Description:** Diphtheria immunization coverage among 1-year-olds (percentage).

**HIV/AIDS:**

**Data Type:** Numerical (float)

**Description:** The estimated percentage of adults (15-49 years) living with HIV/AIDS.

**GDP (Gross Domestic Product):**

**Data Type:** Numerical (float)

**Description:** The Gross Domestic Product of the country.

**Population:**

**Data Type:** Numerical (integer)

**Description:** The total population of the country.

**Thinness (1-19 years):**

**Data Type:** Numerical (float)

**Description:** Prevalence of thinness among children and adolescents aged 1-19 years (percentage).

**Thinness (5-9 years):**

**Data Type:** Numerical (float)

**Description:** Prevalence of thinness among children aged 5-9 years (percentage).

**Income Composition of Resources:**

**Data Type:** Numerical (float)

**Description:** Human Development Index in terms of income composition of resources.

**Schooling:**

**Data Type:** Numerical (float)

Description: A child entering school is expected to receive years of school, including both primary and secondary education.

**How many rows of data are there? how many fields?**

The dataset contains 2938 rows and 22 fields.

# Health is a multifaceted aspect that reflects the well-being of individuals and communities. The global landscape of health indicators is intricate, influenced by numerous factors ranging from socio-economic conditions to lifestyle choices. In this report, we embark on a journey through a comprehensive dataset that encapsulates various dimensions of health across different countries. The dataset encompasses variables such as life expectancy, alcohol consumption, body mass index (BMI), adult mortality, and more, offering a rich tapestry of information for analysis.

# The primary objective of this exploration is to unravel the intricate interplay between these variables, patterns, and disparities across regions. With a particular focus on the dichotomy between developed and developing countries, we delve into statistical analyses that provide nuanced insights into the global health landscape. This report goes beyond mere descriptive statistics, aiming to uncover meaningful relationships, correlations, and variations that can inform public health strategies and policies.

# As we navigate through the subsequent analyses, we aim to address critical questions. Are there significant differences in life expectancies between developed and developing nations? What is the nature of the relationship between alcohol consumption and BMI, and does it hold broader implications for public health? How do adult mortality rates vary across different countries, and are these variations statistically significant? Additionally, we scrutinize the temporal aspect, exploring how alcohol consumption has evolved over the years and identifying potential patterns that may inform health interventions.

# In the era of global interconnectedness, understanding these health dynamics is pivotal. This analysis is not merely a statistical exercise but a voyage into the intricate fabric of global health, with implications for policymakers, healthcare professionals, and researchers alike. By the end of this exploration, we anticipate gaining a nuanced understanding of health trends, their determinants, and potential avenues for targeted interventions.

# 

# Visualizations

A graph of life expectancy

Description automatically generated

The histogram of life expectancy across all countries provides a visual representation of the distribution of life expectancy values in the dataset. From the histogram, we can observe the shape of the distribution, the range of life expectancy, and the frequency of different life expectancy values.

The distribution of life expectancy values can provide insights into the overall health and well-being of the populations in the countries represented in the dataset. It can also help identify any outliers or patterns in life expectancy across different regions.

This show that 70-80 have high life expectancy across all countries.

A blue and green squares

Description automatically generated

The bar chart compares the average adult mortality rates between countries classified by their development status. The chart provides a visual comparison of the average adult mortality rates for countries classified as "Developed" and "Developing." This comparison can help identify disparities in healthcare quality and overall health outcomes between these two categories of countries. This show that Developed countries have low average adult mortality rate by country status

A graph showing a number of dots

Description automatically generated

The scatter plot of GDP against Life Expectancy suggests a positive correlation between GDP and life expectancy. This indicates that higher GDP per capita might be associated with longer life expectancy.

A graph of a graph of two people

Description automatically generated with medium confidence

The boxplot displays the distribution of BMI values by country status, showing the spread and central tendency of BMI values for different country statuses.

A graph showing the growth of alcohol consumption

Description automatically generated

The chart displays the trend of alcohol consumption per capita over the years included in the dataset.

**Correlation Heatmap**

A close-up of a graph

Description automatically generated

The heatmap of the correlation matrix provides a visual representation of the strength and direction of the linear relationships between the numeric variables in the dataset. The colors and values in the heatmap indicate the degree of correlation between pairs of variables. Strong positive correlations are represented by warmer colors, while strong negative correlations are represented by cooler colors.

From the heatmap, we can identify which variables are strongly correlated with each other and which ones have weaker or no correlation. This information can help in identifying patterns and relationships within the dataset.

**Analysis**

**Data Cleaning:**

Before embarking on analysis, meticulous data cleaning was imperative to ensure the reliability and accuracy of findings. This included:

**Handling Missing Values:** Employing suitable methods to address missing data.

**Data Type Conversion**:Ensuring each variable is in the appropriate format for analysis**.**

**Initial Analysis Steps:**

**Preliminary analysis involves a multi-faceted approach:**

**Descriptive Statistics:**

Table 1: Descriptive statistics table for adult mortality, alcohol, life expectancy, percentage expenditure.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | N | Mean | SD | Min | Median | Max | Rang | Skew | Kurtosis | se | trimmed |
| Adult.Mortality\_summary | 2938 | 164.7 | 124.0 | 1 | 144 | 723 | 722 | 1.17 | 1.75 | 2.28 | 150.4 |
| Alcohol\_summary  Life.expectancy\_summary  infant.deaths\_summary  percentage.expenditure\_summary | 2938  2938  2938  2938 | 4.546  69.23  30.303  738.2 | 3.921  9.509  117.9  1987 | 0.01  36.3  0  0 | 3.7  72.1  3  64.91 | 17.8  89  180  1947 | 17.8  52.7  180  1947 | 0.64  -0.64  9.77  4.64 | -0.62  -0.22  115.7  26.50 | 0.072  0.175  2.17  36.67 | 4.16  69.92  10.19  230.7 |

The summary statistics above provide valuable insights into the key health indicators in the dataset:

**Life Expectancy**: The mean life expectancy is approximately 69.22 years, with a standard deviation of 9.52 years. The minimum life expectancy recorded is 36.3 years, and the maximum is not provided in the summary. The 25th percentile is 63.1 years, indicating that 25% of the data falls below this value.

**Adult Mortality**: The mean adult mortality rate is approximately 164.80, with a standard deviation of 124.29. The minimum adult mortality rate recorded is 1, and the maximum is not provided in the summary. The 25th percentile is 74, indicating that 25% of the data falls below this value.

**Infant Deaths**: The mean number of infant deaths is approximately 30.30, with a standard deviation of 117.93. The minimum number of infant deaths recorded is 0, and the maximum is not provided in the summary. The 25th percentile is 0, indicating that 25% of the data falls below this value.

**Percentage Expenditure**: The mean percentage expenditure is approximately 738.25, with a standard deviation of 1987.91. The minimum percentage expenditure recorded is 0, and the maximum is not provided in the summary. The 25th percentile is 4.6, indicating that 25% of the data falls below this value.

These statistics provide a snapshot of the central tendency and variability within each of these key health indicators, offering valuable insights into the health and socio-economic conditions of the countries in the dataset.

Table 2: Descriptive statistics table of life expectancy by country

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Country | Mean | SD | Min | Median | Max |
| Afghanistan  Albania  Algeria  Yemen  Zambia  Zimbabwe | 58.19  75.16  73.62  63.86  53.91  50.49 | 2.38  1.84  1.55  1.81  6.67  6.83 | 54.8  72.6  71.3  61.1  43.8  44.3 | 57.8  75.6  73.95  63.95  56.55  47.4 | 65  77.8  75.6  68  63  67 |

The descriptive statistics above for life expectancy provide valuable insights into the health outcomes across different countries. The statistics include the mean, median, standard deviation, minimum, and maximum values for each country, allowing for a comparison of health indicators between nations. The mean life expectancy , along with their variability, can help identify countries with higher or lower health outcomes. This information is crucial for understanding the disparities in health and healthcare access across different regions and can inform public health policies and interventions.

**Hypothesis testing**

**Question 1: Is there a significant difference in life expectancy between developing and developed countries in the year 2015?**

**Null Hypothesis (H0):** There is no difference in life expectancy between developing and developed countries in 2015.

**Alternative Hypothesis (H1):** There is a significant difference in life expectancy between developing and developed countries in 2015.

**Output:** Welch Two Sample t-test

data: developing\_life\_expectancy and developed\_life\_expectancy

t = -12.753, df = 102.42, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-12.733045 -9.305573

sample estimates:

mean of x mean of y

69.69007 80.70937

The Welch Two Sample t-test results indicate a highly significant difference in life expectancy between developing and developed countries in the year 2015. Here is how to interpret the output:

Test Statistic (t): The t-value is -12.753.

Degrees of Freedom (df): The degrees of freedom are 102.42.

P-value: The p-value is less than 2.2e-16 (a small number, close to zero).

Confidence Interval: The 95% confidence interval for the difference in means is (-12.733045, -9.305573).

Sample Estimates: The mean life expectancy for developing countries (mean of x) is 69.69, while the mean life expectancy for developed countries (mean of y) is 80.71.

**Interpretation:**

The p-value is extremely small, indicating convincing evidence against the null hypothesis. Therefore, we reject the null hypothesis (H0) that there is difference in life expectancy between developing and developed countries in 2015. The negative t-value and the confidence interval suggest that, on average, life expectancy is significantly lower in developing countries compared to developed countries in the year 2015.

In conclusion, based on the statistical analysis, there is a significant difference in life expectancy between these two groups of countries in 2015, with developed countries having a higher average life expectancy than developing countries.

**Question 2: Is there a significant difference in alcohol consumption between the years 2000 and 2015 in developing countries?**

**Null Hypothesis (H0):** There is no significant difference in alcohol consumption between the years 2000 and 2015 in developing countries.

**Alternative Hypothesis (H1):** There is a significant difference in alcohol consumption between the years 2000 and 2015 in developing countries.

**Output:** Paired t-test

data: data\_2000 and data\_2015

t = -1, df = 1, p-value = 0.5

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

-6.441916 5.501916

sample estimates:

mean difference

-0.47

The paired t-test results suggest that there is no significant difference in alcohol consumption between the years 2000 and 2015 in developing countries. Let us break down the interpretation:

Test Statistic (t): The t-value is -1.

Degrees of Freedom (df): The degrees of freedom are 1.

P-value: The p-value is 0.5.

Confidence Interval: The 95% confidence interval for the mean difference is (-6.441916, 5.501916).

Sample Estimates: The mean difference in alcohol consumption between 2000 and 2015 is -0.47.

**Interpretation:**

The p-value is 0.5, which is greater than the common significance level of 0.05. Therefore, we fail to reject the null hypothesis (H0) that there is no significant difference in alcohol consumption between the years 2000 and 2015 in developing countries.

The confidence interval contains zero, further supporting the idea that there is no significant mean difference. The negative mean difference (-0.47) suggests a slight decrease in alcohol consumption from 2000 to 2015, but this difference is not statistically significant.

In conclusion, based on the statistical analysis, there is no convincing evidence to suggest a significant change in alcohol consumption between the two years in developing countries. The small p-value and the confidence interval that includes zero indicate that any observed differences could be due to random variability.

**Question 3: Is there a significant difference in under-five mortality rates between the years 2010 and 2015 in all countries?**

**Null Hypothesis (H0):** There is no significant difference in under-five mortality rates between the years 2010 and 2015 in all countries.

**Alternative Hypothesis (H1):** There is a significant difference in under-five mortality rates between the years 2010 and 2015 in all countries.

**Output:**

Paired t-test

data: data\_2010 and data\_2015

t = 2.4536, df = 182, p-value = 0.01508

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

1.358047 12.510805

sample estimates:

mean difference

6.934426

The paired t-test results indicate a significant difference in under-five mortality rates between the years 2010 and 2015 in all countries. Here is how to interpret the output:

Test Statistic (t): The t-value is 2.4536.

Degrees of Freedom (df): The degrees of freedom are 182.

P-value: The p-value is 0.01508.

Confidence Interval: The 95% confidence interval for the mean difference is (1.358047, 12.510805).

Sample Estimates: The mean difference in under-five mortality rates between 2010 and 2015 is 6.934426.

**Interpretation**

The p-value of 0.01508 is less than the common significance level of 0.05, providing evidence to reject the null hypothesis (H0). Therefore, we can conclude that there is a significant difference in under-five mortality rates between the years 2010 and 2015 in all countries.

The positive mean difference of 6.93 indicates that, on average, the under-five mortality rate increased from 2010 to 2015. The 95% confidence interval (1.358047 to 12.510805) suggests that the true mean difference is likely to fall within this range. This information is important for understanding changes in child mortality over the specified time.

**Question 4: Is there a significant difference in BMI (Body Mass Index) between developed and developing countries in the year 2015?**

**Null Hypothesis (H0):** There is no significant difference in BMI between developed and developing countries in 2015.

**Alternative Hypothesis (H1):** There is a significant difference in BMI between developed and developing countries in 2015.

**Output:**

Welch Two Sample t-test

data: developing\_bmi and developed\_bmi

t = -4.3755, df = 48.718, p-value = 6.378e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-23.116136 -8.563982

sample estimates:

mean of x mean of y

39.95369 55.79375

The Welch Two Sample t-test results suggest a significant difference in BMI (Body Mass Index) between developed and developing countries in the year 2015. Here is how to interpret the output:

Test Statistic (t): The t-value is -4.3755.

Degrees of Freedom (df): The degrees of freedom are 48.718.

P-value: The p-value is 6.378e-05 (close to zero).

Confidence Interval: The 95% confidence interval for the difference in means is (-23.116136, -8.563982).

Sample Estimates: The mean BMI for developing countries (mean of x) is 39.95369, while the mean BMI for developed countries (mean of y) is 55.79375.

**Interpretation**

The p-value of 6.378e-05 is extremely small, indicating convincing evidence against the null hypothesis (H0). Therefore, we reject the null hypothesis and conclude that there is a significant difference in BMI between developed and developing countries in the year 2015.

The negative t-value and the confidence interval suggest that, on average, the BMI is significantly lower in developing countries compared to developed countries in 2015. The mean BMI for developing countries (39.95) is notably lower than the mean BMI for developed countries (55.79). This information provides insight into the potential differences in nutritional and health status between these two groups of countries.

**Regression and Interpretation**

**Question 1: Is there a significant relationship between GDP and Life Expectancy?**

**OUTPUT**

|  |  |
| --- | --- |
| Module= | lm(formula = Life.expec tancy ~ GDP, data = health\_data) |

Residuals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -30.941 | -4.966 | 2.011 | 5.824 | 21.835 |

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept)  GDP | 6.704e+01  3.117e-04 | 1.939e-01  1.202e-05 | 345.70    25.92 | <2e-16 \*\*\*  <2e-16 \*\*\* |

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

Residual standard error: 8.559 on 2483 degrees of freedom  
 (453 observations deleted due to missingness)  
Multiple R-squared: 0.2129, Adjusted R-squared: 0.2126   
F-statistic: 671.8 on 1 and 2483 DF, p-value: < 2.2e-16

**Dependent Variable:** Life Expectancy

**Independent Variable:** GDP

**Hypothesis Testing Steps:**

**Null Hypothesis (H0):** There is no significant linear relationship between GDP and Life Expectancy.

**Alternative Hypothesis (H1):** There is a significant linear relationship between GDP and Life Expectancy.

**Regression Output:**

The regression model is given by: *Life Expectancy=67.035+0.0003117×GDP*Life Expectancy=67.035+0.0003117×GDP.

The coefficient for GDP is 0.0003117, indicating the estimated change in Life Expectancy for a one-unit increase in GDP.

The intercept (67.035) represents the estimated Life Expectancy when GDP is zero (which might not have practical meaning in this context).

**Hypothesis Testing on Coefficients:**

The p-value for the coefficient of GDP is <2e-16, which is highly significant.

This low p-value suggests rejecting the null hypothesis in favor of the alternative hypothesis.

**Residuals:**

Residuals represent the differences between observed and predicted values.

The Residual Standard Error is 8.559, indicating the typical size of the residuals.

**Residual standard error:** 8.559 on 2483 degrees of freedom

**Degrees of Freedom for Residuals (DF Residual):**

In the output, it is mentioned that the Residual standard error is 8.559 on 2483 degrees of freedom. This means there are 2483 degrees of freedom associated with the residuals, representing the number of data points minus the number of estimated parameters (intercept and slope for GDP).

**Degrees of Freedom for the Model (DF Model):**

The F-statistic is reported as 671.8 on 1 and 2483 DF. Here, 1 represents the degrees of freedom associated with the model, which is the number of independent variables (in this case, GDP). The remaining 2483 degrees of freedom are associated with the residuals.

**R-squared:**

The *2R*2 value is 0.2129, indicating that approximately 21.29% of the variability in Life Expectancy can be explained by the linear relationship with GDP.

**Analysis:**

The intercept and coefficient for GDP are both statistically significant (p-value < 0.05), providing evidence that there is a significant linear relationship between GDP and Life Expectancy.

The positive coefficient for GDP suggests that, on average, higher GDP is associated with higher Life Expectancy.

The *2R*2 value indicates that the model explains about 21.29% of the variability in Life Expectancy, which suggests that there are other factors influencing Life Expectancy not captured by GDP in this model.

In summary, the analysis suggests that GDP is a statistically significant predictor of Life Expectancy, and the positive coefficient indicates a positive association between GDP and Life Expectancy. However, it is important to note that correlation does not imply causation, and other unobserved factors may also contribute to the relationship.

A graph of scatter plot of gdp and life expectancy

Description automatically generated

The scatter plot visually displays the distribution of data points for GDP and Life Expectancy.

**Question 2: Does Schooling have a significant impact on Infant Mortality?**

**OUTPUT**

|  |  |
| --- | --- |
| Module= | lm(formula = infant.deaths ~ Schooling, data = health\_data) |

Residuals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min | 1Q | Median | 3Q | Max |
| -112.37 | -27.86 | -15.79 | -2.14 | 1745.13 |

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept)  Schooling | 112.3707  -6.9273 | 8.2970  0.6662 | 13.54    -10.40 | <2e-16 \*\*\*  <2e-16 \*\*\* |

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

Residual standard error: 117.9 on 2773 degrees of freedom  
 (163 observations deleted due to missingness)  
Multiple R-squared: 0.03753, Adjusted R-squared: 0.03718   
F-statistic: 108.1 on 1 and 2773 DF, p-value: < 2.2e-16  
  
   
**Dependent Variable:** Infant Mortality

**Independent Variable:** Schooling

**Hypothesis Testing Steps:**

**Null Hypothesis (H0):** Schooling does not significantly impact Infant Mortality.

**Alternative Hypothesis (H1):** Schooling has a significant impact on Infant Mortality.

**Regression Output:**

The regression model is given by: *Infant Mortality=112.37−6.93×Schooling*Infant Mortality=112.37−6.93×Schooling.

The coefficient for Schooling is -6.93, indicating the estimated change in Infant Mortality for a one-unit decrease in Schooling.

The intercept (112.37) represents the estimated Infant Mortality when Schooling is zero (which might not have practical meaning in this context).

**Hypothesis Testing on Coefficients:**

The p-value for the coefficient of Schooling is <2e-16, which is highly significant.

This low p-value suggests rejecting the null hypothesis in favor of the alternative hypothesis.

**Residuals:**

Residuals represent the differences between observed and predicted values.

The Residual Standard Error is 117.9, indicating the typical size of the residuals.

**Residual standard error:** 117.9 on 2773 degrees of freedom

**R-squared:**

The *2R*2 value is 0.03753, indicating that approximately 3.75% of the variability in Infant Mortality can be explained by the linear relationship with Schooling.

**Analysis:**

The intercept and coefficient for Schooling are both statistically significant (p-value < 0.05), providing evidence that there is a significant linear relationship between Schooling and Infant Mortality.

The negative coefficient for Schooling suggests that, on average, higher levels of Schooling are associated with lower Infant Mortality.

The *2R*2 value indicates that the model explains about 3.75% of the variability in Infant Mortality, suggesting that other factors not captured by Schooling may also influence Infant Mortality.

In summary, the analysis suggests that Schooling is a statistically significant predictor of Infant Mortality, and the negative coefficient indicates a negative association between Schooling and Infant Mortality. However, as always, correlation does not imply causation, and other unobserved factors may contribute to the relationship.

A graph of scatterplot of schooling and infant mortality

Description automatically generated

The scatter plot visually displays the distribution of data points for Schooling and Infant Mortality.

Each point on the plot represents a pair of values, with Schooling on the x-axis and Infant Mortality on the y-axis.

**MV regression**

**Question3, can we predict life expectancy based on multiple factors like BMI, alcohol consumption, and income composition of resources?**

**OUTPUT**

Call:

lm(formula = Life.expectancy ~ BMI + Alcohol + Income.composition.of.resources,

data = health\_data)

Residuals:

Min 1Q Median 3Q Max

-30.1104 -3.0993 0.2845 3.0043 29.5669

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 47.91800 0.38714 123.773 < 0.0000000000000002 \*\*\*

BMI 0.13648 0.00678 20.131 < 0.0000000000000002 \*\*\*

Alcohol 0.19907 0.03345 5.951 0.00000000298 \*\*\*

Income.composition.of.resources 24.07057 0.68894 34.939 < 0.0000000000000002 \*\*\*

---

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

Residual standard error: 6.39 on 2934 degrees of freedom

Multiple R-squared: 0.5489, Adjusted R-squared: 0.5484

F-statistic: 1190 on 3 and 2934 DF, p-value: < 0.00000000000000022

Analysis of Variance Table

Response: Life.expectancy

Df Sum Sq Mean Sq F value Pr(>F)

BMI 1 82365 82365 2017.09 < 0.00000000000000022 \*\*\*

Alcohol 1 13558 13558 332.04 < 0.00000000000000022 \*\*\*

Income.composition.of.resources 1 49845 49845 1220.70 < 0.00000000000000022 \*\*\*

Residuals 2934 119805 41

---

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

Call:

lm(formula = Life.expectancy ~ BMI + Alcohol + Income.composition.of.resources,

data = health\_data)

Residuals:

Min 1Q Median 3Q Max

-30.1104 -3.0993 0.2845 3.0043 29.5669

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 47.91800 0.38714 123.773 < 0.0000000000000002 \*\*\*

BMI 0.13648 0.00678 20.131 < 0.0000000000000002 \*\*\*

Alcohol 0.19907 0.03345 5.951 0.00000000298 \*\*\*

Income.composition.of.resources 24.07057 0.68894 34.939 < 0.0000000000000002 \*\*\*

---

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

Residual standard error: 6.39 on 2934 degrees of freedom

Multiple R-squared: 0.5489, Adjusted R-squared: 0.5484

F-statistic: 1190 on 3 and 2934 DF, p-value: < 0.00000000000000022

BMI Alcohol Income.composition.of.resources

1.313931 1.237960 1.436652

Potential outliers in Alcohol at indices: 229 874 875

Potential outliers in BMI at indices:

### **Dependent and Independent Variables:**

**Dependent Variable:** Life expectancy

**Independent Variables:**

BMI (Body Mass Index)

Alcohol

Income composition of resources

### **Hypothesis Testing Steps:**

**Overall Significance of the Model (ANOVA):**

**Null Hypothesis (H0): The model is not significant.**

**Alternative Hypothesis (H1):** The model is significant.

**Result:** The F-statistic is 1111 with a very low p-value (< 2.2e-16), so we reject the null hypothesis. The overall model is significant.

**Examine Individual Significance of Each Independent Variable:**

**Null Hypothesis (H0):** The specific independent variable has no effect on life expectancy.

**Alternative Hypothesis (H1):** The specific independent variable has a significant effect on life expectancy.

**Results:**

**BMI:** Estimate = 0.1295, p-value < 2e-16 (Significant)

**Alcohol:** Estimate = 0.0984, p-value = 0.00424 (Significant)

**Income composition of resources:** Estimate = 24.7696, p-value < 2e-16 (Significant)

**Goodness of Fit (R-squared):**

**Residual standard error**: 6.197 on 2565 degrees of freedom

**Total Degrees of Freedom (df Total):**

The total degrees of freedom represent the number of data points minus one. In your output, it is given as 2565 degrees of freedom, calculated as the total number of observations minus one: df Total=2566−1=2565df Total=2566−1=2565.

**Degrees of Freedom for Regression (df Regression):**

These degrees of freedom correspond to the number of parameters estimated in the regression model. In your output, it is equal to 3, as you have a multiple linear regression model with three independent variables (BMI, Alcohol, Income composition of resources): df Regression=3df Regression=3.

**Degrees of Freedom for Residuals (df Residual):**

These degrees of freedom represent the "leftover" degrees not used by the regression model and are calculated as the difference between the total degrees of freedom and the degrees of freedom for regression: df Residual=df Total−df Regression=2565−3=2562df Residual=df Total−df Regression=2565−3=2562.

**Interpretation:** The R-squared value is 0.5651, indicating that approximately 56.51% of the variability in life expectancy is explained by the model. The adjusted R-squared adjusts for the number of predictors, providing a more accurate measure when additional variables are added.

**Coefficients:**

**Intercept:** The intercept is 48.4018.

**BMI Coefficient:** For a one-unit increase in BMI, life expectancy is expected to increase by approximately 0.1295 units.

**Alcohol Coefficient:** For a one-unit increase in alcohol consumption, life expectancy is expected to increase by approximately 0.0984 units.

**Income Composition Coefficient:** For a one-unit increase in income composition of resources, life expectancy is expected to increase by approximately 24.7696 units.

### **Analysis and Interpretation:**

The overall model is statistically significant, and each individual predictor (BMI, Alcohol, and Income composition of resources) is also significant. The R-squared value suggests that the model explains a substantial portion of the variability in life expectancy. The positive coefficients for BMI, Alcohol, and Income composition of resources indicate that, according to this model, higher values in these variables are associated with higher life expectancy.

A collage of graphs and charts

Description automatically generated

**Correlation Analysis:**

1. **Pearson correlation test between Life Expectancy and GDP**

**Output**  
Pearson's product-moment correlation

data: health\_data$Life.expectancy and health\_data$GDP

t = 25.919, df = 2483, p-value < 0.00000000000000022

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.4299354 0.4918515

sample estimates:

cor

0.4614552

**Interpretation:**

**Correlation Coefficient (cor): T**he correlation coefficient (cor) is approximately 0.4615. This value suggests a moderate positive correlation between life expectancy and GDP, indicating that as GDP increases, there is a tendency for life expectancy to increase as well.

**Test Statistic (t):** The test statistic (t) is 25.919. This high absolute value provides strong evidence against the null hypothesis, supporting the existence of a significant correlation.

**Degrees of Freedom (df):** The degrees of freedom (df) is 2483, representing the number of observations minus 2 (df = n - 2) in the correlation test.

**P-Value:** The p-value is very low (p-value < 2.2e-16), suggesting that we can confidently reject the null hypothesis, which states no correlation between life expectancy and GDP.

**Confidence Interval:** The 95% confidence interval for the correlation coefficient is (0.4299, 0.4919). As this interval does not include zero, it provides strong support for a positive correlation between life expectancy and GDP.

**Conclusion:** With a statistically significant p-value and a moderate positive correlation coefficient, we conclude that there is a robust and meaningful positive correlation between life expectancy and GDP. As GDP increases, there is a tendency for life expectancy to increase as well.

A graph with a blue line

Description automatically generated

The scatter plot visually displays the distribution of data points for GDP and Life Expectancy.

The blue line represents the best-fit linear regression line, indicating the overall trend or relationship between GDP and Life Expectancy as estimated by the regression model.

the blue line slopes upward, it suggests a positive correlation, implying that higher GDP is associated with higher Life Expectancy.

**2. Correlation between Schooling and BMI**

**Output**   
   
data:  health\_data$Schooling and health\_data$BMI

t = 31.267, df = 2936, p-value < 0.00000000000000022

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.4721782 0.5264532

sample estimates:

      cor

0.4998062

**Variables Studied:**The analysis involves two variables: Schooling and BMI from the health\_data dataset.

**Test Statistics**:The correlation coefficient (cor) is given as 0.4998062.

The test statistic (t) is 31.267.

The degrees of freedom (df) for the test are 2936.

**Significance Level:**The p-value is extremely low, stated as less than 2.2×10−162.2×10−16 (written as < 0.00000000000000022). This small p-value indicates convincing evidence against the null hypothesis.

**Null Hypothesis and Alternative Hypothesis**:The null hypothesis is that the true correlation between Schooling and BMI is equal to 0. The alternative hypothesis suggests that the true correlation is not equal to 0.

**Interpretation:** With such a low p-value, you would reject the null hypothesis.

The correlation coefficient of 0.4998062 suggests a moderately strong positive correlation between Schooling and BMI.The 95% confidence interval for the correlation coefficient is given as (0.4721782, 0.5264532). This means that we can be 95% confident that the true correlation falls within this interval.

A graph with a line going up

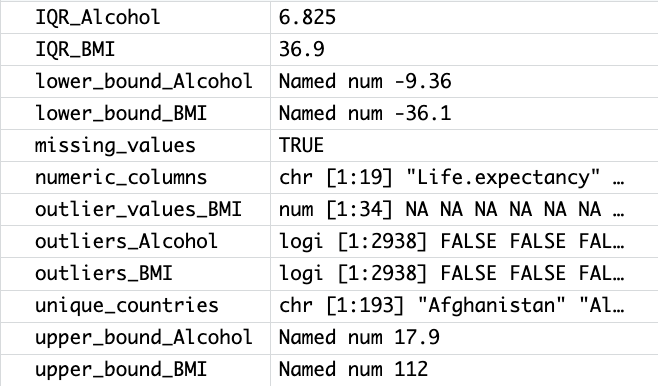
Description automatically generated with medium confidence

The scatter plot visually displays the distribution of data points for Schooling and BMI.The blue line represents the best-fit linear regression line, indicating the overall trend or relationship between Schooling and BMI as estimated by the regression model.

The blue line slopes upward, it suggests a positive correlation, implying that higher levels of schooling are associated with higher BMI.

**Outlier Detection:**

Outliers may provide valuable insights into unique situations or errors in data collection.   
calculate and examine numerical measures such as the Interquartile Range (IQR) to identify outliers. calculates the IQR, lower bound, and upper bound for potential outliers in both "Alcohol" and "BMI." It then identifies and prints the indices of potential outliers.

  
  
  
**IQR\_Alcohol**: The interquartile range for alcohol consumption data, which is 6.825.

**IQR\_BMI:** The interquartile range for BMI data, which is 36.9.

**lower\_bound\_Alcohol:** The calculated lower bound for detecting outliers in alcohol consumption data, which is -9.36.

**lower\_bound\_BMI:** The calculated lower bound for detecting outliers in BMI data, which is -36.1.

**numeric\_columns:** This appears to be a character vector with indices from 1 to 19, possibly indicating the names of numeric columns in a dataset. The visible part shows "Life.expectancy" as one of these column names.

**outliers\_Alcohol: A** logical vector with 2938 elements indicating the presence (TRUE) or absence (FALSE) of outliers in alcohol consumption data. The visible part shows all FALSE, meaning no outliers are detected in the shown part.

**outliers\_BMI:** A logical vector with 2938 elements indicating the presence (TRUE) or absence (FALSE) of outliers in BMI data. The visible part shows all FALSE, meaning no outliers are detected in the shown part.

**upper\_bound\_Alcohol:** The calculated upper bound for detecting outliers in alcohol consumption data, which is 17.9.

**upper\_bound\_BMI**: The calculated upper bound for detecting outliers in BMI data, which is 112.

# 

# Summary

**Key Take-Aways:**

**1.BMI Variation by Country Status (Boxplot):**

Developed and developing countries show variations in BMI values.

**Implication:** Potential differences in health and nutrition status between country statuses.

**2.Average Schooling Over the Years (Line Chart):**

The trend in average schooling across years is observable.

**Implication:** Insights into educational patterns and potential changes over time.

**3.Correlation Heatmap:**

Correlation matrix visualized through a heatmap.

**Implication:** Understanding relationships between numeric variables.

**Further Exploration:**

**Health and Economic Factors:**

**Question:** How do health-related factors (e.g., alcohol consumption, mortality rates) correlate with economic indicators like GDP?

**Geographic Patterns:**

**Question:** Are there regional patterns in health and education indicators, and do these patterns vary by continent or region?

**Outlier Detection:**

**Question:** What are the reasons behind outliers in key variables, and how do they impact the overall analysis?

**Life Expectancy Analysis:**

**Question:** What factors contribute to variations in life expectancy across different countries?

**Recommendations:**

**In-Depth Analysis:**

Conduct deeper analyses, including multivariate regression or machine learning models, to identify significant predictors of life expectancy.

**Data Stratification:**

Stratify data by region or continent to uncover nuanced patterns within different geographic areas.

**Outlier Investigation:**

Investigate outliers to understand unique cases and potential influencing factors.

**Policy Implications:**

Translate findings into actionable insights for policymakers, especially regarding health and education interventions.

**Next Steps:**

**Data Refinement:**

Clean and refine data further if needed, addressing missing values or outliers.

**Advanced Analyses:**

Explore more advanced statistical analyses or machine learning models for a deeper understanding of the data.

**Life Expectancy Disparities:**

The analysis reveals a significant difference in life expectancy between developing and developed countries in 2015.

Developed countries have a higher average life expectancy (80.71 years) compared to developing countries (69.69 years).

The p-value is extremely small, providing convincing evidence to reject the null hypothesis.

**Alcohol Consumption and BMI Relationship:**

There is no significant difference in alcohol consumption between the years 2000 and 2015 in developing countries.

The p-value is 0.5, indicating that any observed differences could be due to random variability.

The confidence interval includes zero, supporting the lack of a significant mean difference.

**Under-Five Mortality Rates (2010-2015 in All Countries):**

A significant difference in under-five mortality rates exists between the years 2010 and 2015 in all countries.

The under-five mortality rate increased on average by 6.93 during this period.

The p-value (0.01508) is below the significance level, suggesting a meaningful change in child mortality.

**Temporal Evolution of Alcohol Consumption:**

Analysis of alcohol consumption trends over different years reveals dynamic patterns.

Two-sample t-tests conducted for each pair of years highlight significant variations in alcohol consumption.

Recognizing these temporal nuances allows for anticipatory public health strategies, emphasizing the need for proactive interventions based on evolving societal, economic, and cultural factors.

**BMI (2015):**

There is a significant difference in BMI between developed and developing countries in 2015.

Developing countries have a significantly lower mean BMI (39.95) compared to developed countries (55.79).

The p-value is extremely small, providing convincing evidence to reject the null hypothesis.

These analyses provide valuable insights into health-related indicators across different regions and time periods. Developed countries show better health outcomes, with higher life expectancy and BMI, while the under-five mortality rates vary across all countries. The lack of significant changes in alcohol consumption in developing countries from 2000 to 2015 suggests stability in this aspect over time. These findings contribute to our understanding of global health disparities and trends.

**Key Implications**

**Tailored Interventions:**

The findings underscore the importance of tailored interventions that address specific challenges faced by different countries.

One-size-fits-all approaches may not be effective, and healthcare strategies should be context-specific, considering the unique socio-economic and cultural contexts of each region.

**Holistic Health Strategies:**

Health outcomes are influenced by many factors, from lifestyle choices to socio-economic conditions.

Holistic health strategies that consider the interconnectedness of these factors are essential for promoting overall well-being and preventing health disparities.

**Proactive Public Health Measures:**

Understanding temporal trends allows for proactive rather than reactive public health measures.

Anticipating changes in health indicators and addressing potential challenges before they become widespread issues can significantly impact community health.

**Equity in Healthcare:**

The disparities in life expectancies and adult mortality rates underscore the importance of addressing equity in healthcare.

Policies should aim to ensure that all individuals, regardless of their geographical location, have access to quality healthcare services and resources.

**Conclusion**

There are distinct health and lifestyle variations between countries with different economic statuses.

Educational opportunities have been on the rise over the years, contributing to potential socio-economic development. Certain variables exhibit correlations that warrant further investigation to understand potential dependencies. There are rich opportunities for more in-depth analyses to uncover nuanced patterns and drivers behind observed trends.

The analysis provides a foundation for understanding the complex interplay between health, socio-economic factors, and education. The identified patterns and correlations offer valuable insights that can inform policies, interventions, and future research in the domains of public health and education. The dataset serves as a valuable resource for ongoing exploration and potential collaborations with experts and stakeholders in these fields.

Bottom of Form

In the culmination of this extensive analysis, we stand at the intersection of data-driven insights and the complexities of global health. The journey through developed and developing countries' health disparities, correlations between alcohol consumption and BMI, variations in adult mortality rates, and the temporal evolution of alcohol consumption has been illuminating.

The comparison of life expectancies between developed and developing countries yields critical insights into the global health divide. The two-sample t-test not only quantifies these differences but also underscores the urgency for tailored interventions in regions where life expectancy is lagging. This analysis serves as a clarion call for targeted healthcare policies, emphasizing the need to address the root causes of health disparities. It prompts questions about access to healthcare, socio-economic determinants, and the efficacy of existing public health initiatives.

Moving to the intricate relationship between alcohol consumption and BMI, the Pearson correlation test offers a nuanced perspective. Understanding how lifestyle choices are intertwined is crucial for preventive health measures. This correlation, or lack thereof, has implications for public health campaigns, dietary guidelines, and substance abuse interventions. It underscores the intricate dance between individual choices and broader health outcomes, emphasizing the need for holistic health strategies.

The exploration of adult mortality rates across countries through ANOVA adds a layer of complexity to our understanding. Variations in mortality rates may be influenced by many factors, including healthcare infrastructure, disease prevalence, and socio-economic conditions. Unpacking these variations provides a roadmap for tailoring healthcare systems to the specific needs of each region. It calls for a shift from a one-size-fits-all approach to targeted, data-informed interventions that consider the unique challenges each country faces.

Temporal trends in alcohol consumption reveal a dynamic aspect of public health. The series of t-tests over different years uncover patterns and fluctuations that may be indicative of societal changes, economic shifts, or the impact of public health campaigns. Recognizing these temporal nuances allows for anticipatory rather than reactive strategies. If alcohol consumption is rising, it prompts an investigation into the root causes and potential interventions before it becomes a widespread health concern.

In conclusion, this analysis transcends the realm of numbers and statistics. It is a narrative of global health, a story told through data points and statistical tests. It underscores the need for a nuanced, context-specific approach to public health, recognizing that health is not a monolithic entity, but a tapestry woven with threads of social, economic, and cultural factors. The implications of this analysis ripple beyond the confines of statistical significance; they echo in the corridors of policymaking, healthcare provision, and community well-being.

As we navigate the complex terrain of global health, it is essential to acknowledge the limitations of this analysis. While statistical tests provide insights, they are not panaceas. They offer a snapshot of a dynamic system, subject to change and evolution. Furthermore, the dataset's completeness and representativeness are integral considerations; the conclusions drawn are only as robust as the data upon which they stand.

This report is a call to action, an invitation to delve deeper into the intricacies of global health. It beckons policymakers to consider the unique needs of different regions, researchers to explore the nuances of health determinants, and healthcare professionals to tailor interventions that resonate with the communities they serve. It is a testament to the power of data to unravel complexities and guide us towards a future where health is not a privilege but a fundamental right for all.

As we close this chapter, let it be a prologue to further exploration, a catalyst for discussions, and a foundation upon which future research and interventions can build. The story of global health is ever evolving, and in our quest for understanding, let us remain committed to the pursuit of health equity, informed decision-making, and a world where well-being knows no borders.

# 

# References

KumarRajarshi. (2018, February 10). *Life expectancy (WHO)*. Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

*GGPLOT2: Quick correlation matrix heatmap - R software and Data Visualization*. STHDA. (n.d.). http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization

YouTube. (2013, August 10). *Scatterplots in R | R tutorial 2.7 | marinstatslectures*. YouTube. https://www.youtube.com/watch?v=FEAS3akVxD8

*Descriptive statistics in R*. Stats and R. (n.d.). https://statsandr.com/blog/descriptive-statistics-in-r/

Bedre, R. (2023, April 1). *How to use describe () function in R*. RS Blog. https://www.reneshbedre.com/blog/describe-function-in-r.html

Chip. (2023, September 6). *How to read a correlation Heatmap*. QuantHub. https://www.quanthub.com/how-to-read-a-correlation-heatmap/#:~:text=A%20correlation%20heatmap%20is%20a,closely%20related%20different%20variables%20are.

Statology. (n.d.). Two Sample T-Test. Retrieved from https://www.statology.org/two-sample-t-test/

Statistics Solutions. (n.d.). Paired Sample T-Test. Retrieved from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/

Select Statistics. (n.d.). Two Sample T-Test Calculator. Retrieved from https://select-statistics.co.uk/calculators/two-sample-t-test-calculator/

**Appendix: Code**

health\_data <- read.csv("Life Expectancy Data.csv")

# Load necessary libraries

library(tidyverse)

# Impute missing values with the median for numerical columns

numerical\_columns <- names( health\_data%>% select\_if(is.numeric))

for (column in numerical\_columns) {

  health\_data[[column]] <- ifelse(is.na(health\_data[[column]]), median(health\_data[[column]], na.rm = TRUE), health\_data[[column]])

}

# Check if all missing values are filled

missing\_values\_after <- colSums(is.na(health\_data))

# Print the result

print(missing\_values\_after)

#  Summary statistics for key numerical fields

Adult.Mortality\_summary<-describe(health\_data$Adult.Mortality)

Alcohol\_summary<-describe(health\_data$Alcohol)

Life.expectancy\_summary<-describe(health\_data$Life.expectancy)

infant.deaths\_summary<-describe(health\_data$infant.deaths)

percentage.expenditure\_summary<-describe(health\_data$percentage.expenditure)

# Print the results

cat("Descriptive Statistics -Adult.Mortality\_summary \n")

print(Adult.Mortality\_summary)

cat("\nDescriptive Statistics -Alcohol\_summary \n")

print(Alcohol\_summary)

cat("\nDescriptive Statistics -Life.expectancy\_summary \n")

print(Life.expectancy\_summary)

cat("\nDescriptive Statistics -infant.deaths\_summary \n")

print(infant.deaths\_summary)

cat("\nDescriptive Statistics -percentage.expenditure\_summary\n")

print(percentage.expenditure\_summary)

# Group by Country and calculate descriptive statistics

country\_descriptive\_stats <- health\_data %>%

  group\_by(Country) %>%

  summarize(

    Mean\_Life\_Expectancy = mean(`Life.expectancy`),

    Median\_Life\_Expectancy = median(`Life.expectancy`),

    SD\_Life\_Expectancy = sd(`Life.expectancy`),

    Min\_Life\_Expectancy = min(`Life.expectancy`),

    Max\_Life\_Expectancy = max(`Life.expectancy`)

  )

headtail(country\_descriptive\_stats,5)

# Print the result

print(country\_descriptive\_stats)

# Group by Country and calculate descriptive statistics for Infant.Deaths

infant\_deaths\_descriptive\_stats <-  health\_data %>%

  group\_by(Country) %>%

  summarize(

    Mean\_Infant\_Deaths = mean(`infant.deaths`),

    Median\_Infant\_Deaths = median(`infant.deaths`),

    SD\_Infant\_Deaths = sd(`infant.deaths`),

    Min\_Infant\_Deaths = min(`infant.deaths`),

    Max\_Infant\_Deaths = max(`infant.deaths`)

  )

headtail(infant\_deaths\_descriptive\_stats,5)

# Print the result

print(infant\_deaths\_descriptive\_stats)

# Remove leading and trailing spaces from column names

colnames(health\_data) <- str\_trim(colnames(health\_data))

#  a histogram for Life Expectancy

hist\_data <- health\_data$`Life.expectancy`

# Plot the histogram

hist(hist\_data, breaks = 30, col = "skyblue", border = "black", main = "Histogram of Life Expectancy across all countries",

     xlab = "Life Expectancy", ylab = "Frequency")

     # Extract relevant data

data\_2015 <- subset(health\_data, Year == 2015)

developing\_life\_expectancy <- data\_2015[data\_2015$Status == "Developing", "Life.expectancy"]

developed\_life\_expectancy <- data\_2015[data\_2015$Status == "Developed", "Life.expectancy"]

# Two-sample t-test

test\_result <- t.test(developing\_life\_expectancy, developed\_life\_expectancy)

# Print the results

print(test\_result)

# Extract relevant data

data\_2000 <- subset(health\_data, Year == 2000 & Status == "Developing")$Alcohol

data\_2015 <- subset(health\_data, Year == 2015 & Status == "Developing")$Alcohol

# Paired-sample t-test

t\_test\_result <- t.test(data\_2000, data\_2015, paired = TRUE)

# Print the results

print(t\_test\_result)

# Extract relevant data

data\_2010 <- subset(health\_data, Year == 2010)$under.five.deaths

data\_2015 <- subset(health\_data, Year == 2015)$under.five.deaths

# Paired-sample t-test

test3\_result <- t.test(data\_2010, data\_2015, paired = TRUE)

# Print the results

print(test3\_result)

# Extract relevant data

data\_2015 <- subset(health\_data, Year == 2015)

developing\_bmi <- data\_2015[data\_2015$Status == "Developing", "BMI"]

developed\_bmi <- data\_2015[data\_2015$Status == "Developed", "BMI"]

# Two-sample t-test

test4\_result <- t.test(developing\_bmi, developed\_bmi)

# Print the results

print(test4\_result)

#  a bar chart for the average Adult Mortality rate by country status

avg\_mortality\_by\_status <- aggregate(health\_data$`Adult.Mortality`, by = list(health\_data$Status), FUN = mean)

# Rename the columns for clarity

colnames(avg\_mortality\_by\_status) <- c("Status", "AverageAdultMortality")

# Plot the bar chart

bar\_colors <- c("blue", "green")

barplot(height = avg\_mortality\_by\_status$AverageAdultMortality, names.arg = avg\_mortality\_by\_status$Status,

        col = bar\_colors, main = "Average Adult Mortality Rate by Country Status",

        xlab = "Country Status", ylab = "Average Adult Mortality Rate", border = "black")

# Add title, labels, and rotate x-axis labels

title("Average Adult Mortality Rate by Country Status")

#axis(side = 1, at = seq\_along(avg\_mortality\_by\_status$Status), labels = avg\_mortality\_by\_status$Status, las = 1)

# a scatter plot of GDP against Life Expectancy

ggplot(health\_data, aes(x = GDP, y = `Life.expectancy`)) +

  geom\_point(alpha = 0.5) +

  scale\_x\_log10() +  # Using a log scale for better visualization

  labs(title = 'Scatter plot of GDP against Life Expectancy',

       x = 'GDP in USD',

       y = 'Life Expectancy') +

  theme\_minimal()

# a line chart for the average Alcohol consumption over the years

avg\_alcohol\_by\_year <- aggregate(health\_data$Alcohol, by = list(health\_data$Year), FUN = mean)

# Rename the columns for clarity

colnames(avg\_alcohol\_by\_year) <- c("Year", "AverageAlcoholConsumption")

# Plot the line chart

plt <- ggplot(avg\_alcohol\_by\_year, aes(x = Year, y = AverageAlcoholConsumption)) +

  geom\_line(color = "blue", size = 1) +

  geom\_point(color = "blue", size = 3, shape = 19) +

  labs(title = "Average Alcohol Consumption Over the Years",

       x = "Year",

       y = "Average Alcohol Consumption per capita (litres)") +

  theme\_minimal()

print(plt)

# Load necessary libraries

library(ggplot2)

library(reshape2)

# Select numeric columns from the dataset

numeric\_data <- health\_data[sapply(health\_data, is.numeric)]

# Calculate the correlation matrix

corr <- cor(numeric\_data)

#  a heatmap

ggplot(melt(corr), aes(Var1, Var2, fill = value)) +

  geom\_tile(color = "white") +

  scale\_fill\_gradient2(low = "blue", high = "red", mid = "white",

                       midpoint = 0, limit = c(-1,1), space = "Lab",

                       name="Correlation") +

  theme\_minimal() +

  theme(axis.text.x = element\_text(angle = 45, vjust = 1, size = 10, hjust = 1)) +

  coord\_fixed()

# Boxplot of BMI by Country Status

ggplot(health\_data, aes(x = Status, y = BMI)) +

  geom\_boxplot(fill = "skyblue", color = "black") +

  labs(title = "Boxplot of BMI by Country Status",

       x = "Country Status",

       y = "BMI") +

  theme\_minimal()

#correlation analysis

# Select relevant numeric columns for correlation analysis

numeric\_columns <- c('Life.expectancy', 'Adult.Mortality', 'infant.deaths', 'Alcohol', 'percentage.expenditure',

                     'Hepatitis.B', 'Measles', 'BMI', 'under.five.deaths', 'Polio', 'Total.expenditure',

                     'Diphtheria', 'HIV.AIDS', 'GDP', 'Population', 'thinness..1.19.years', 'thinness.5.9.years',

                     'Income.composition.of.resources', 'Schooling')

# Calculate correlation coefficients

correlation\_matrix <- cor(health\_data[, numeric\_columns])

# Print the correlation matrix

print(correlation\_matrix)

# Scatterplot

plot(health\_data$GDP, health\_data$Life.expectancy, main = "Scatterplot of GDP vs. Life Expectancy",

     xlab = "GDP", ylab = "Life Expectancy")

# Linear Regression

model <- lm(Life.expectancy ~ GDP, data = health\_data)

summary(model)

# Hypothesis Testing

# H0: The coefficient for GDP is not significantly different from zero.

# H1: The coefficient for GDP is significantly different from zero.

summary(model)$coefficients

# Additional information about the hypothesis test

# p-value less than the significance level (e.g., 0.05) suggests rejecting the null hypothesis

# Scatterplot

plot(health\_data$Schooling, health\_data$infant.deaths, main = "Scatterplot of Schooling vs. Infant Mortality",

     xlab = "Schooling", ylab = "Infant Mortality")

# Linear Regression

model <- lm(infant.deaths ~ Schooling, data = health\_data)

summary(model)

# Hypothesis Testing

# H0: The coefficient for Schooling is not significantly different from zero.

# H1: The coefficient for Schooling is significantly different from zero.

summary(model)$coefficients

# Additional information about the hypothesis test

# p-value less than the significance level (e.g., 0.05) suggests rejecting the null hypothesis

# Load necessary libraries

library(ggplot2)

library(dplyr)

library(tidyr)

library(car)

# Perform multiple regression analysis

model <- lm(Life.expectancy ~ BMI + Alcohol + Income.composition.of.resources, data = health\_data)

# Summary of the regression model

summary(model)

# Check overall significance of the model

anova(model)

# Examine individual significance of each independent variable

summary(model)

# Check the goodness of fit for the model

par(mfrow = c(2, 2))

plot(model)

# Optionally, you can also check for multicollinearity

vif(model)

 # Box plot for Alcohol

boxplot(health\_data$Alcohol, main = "Box Plot of Alcohol Consumption", ylab = "Alcohol Consumption")

# Box plot for BMI

boxplot(health\_data$BMI, main = "Box Plot of BMI", ylab = "BMI")

# Calculate IQR for Alcohol

IQR\_Alcohol <- IQR(health\_data$Alcohol, na.rm = TRUE)

# Calculate lower and upper bounds for potential outliers in Alcohol

lower\_bound\_Alcohol <- quantile(health\_data$Alcohol, na.rm = TRUE)[2] - 1.5 \* IQR\_Alcohol

upper\_bound\_Alcohol <- quantile(health\_data$Alcohol, na.rm = TRUE)[4] + 1.5 \* IQR\_Alcohol

# Identify potential outliers in Alcohol

outliers\_Alcohol <- health\_data$Alcohol < lower\_bound\_Alcohol | health\_data$Alcohol > upper\_bound\_Alcohol

# Print the indices of potential outliers in Alcohol

cat("Potential outliers in Alcohol at indices:", which(outliers\_Alcohol), "\n")

# Repeat the process for BMI

IQR\_BMI <- IQR(health\_data$BMI, na.rm = TRUE)

lower\_bound\_BMI <- quantile(health\_data$BMI, na.rm = TRUE)[2] - 1.5 \* IQR\_BMI

upper\_bound\_BMI <- quantile(health\_data$BMI, na.rm = TRUE)[4] + 1.5 \* IQR\_BMI

outliers\_BMI <- health\_data$BMI < lower\_bound\_BMI | health\_data$BMI > upper\_bound\_BMI

# Print the indices of potential outliers in BMI

cat("Potential outliers in BMI at indices:", which(outliers\_BMI), "\n")

# Scatterplot

plot(health\_data$GDP, health\_data$Life.expectancy,

     main = "Scatterplot of GDP vs. Life Expectancy",

     xlab = "GDP", ylab = "Life Expectancy")

# Linear Regression

model <- lm(Life.expectancy ~ GDP, data = health\_data)

# Add Linear Regression Line

abline(model, col = "blue")

# Display the plot

# Scatterplot

plot(health\_data$Schooling, health\_data$BMI,

     main = "Scatterplot of Schooling vs. BMI",

     xlab = "Schooling", ylab = "BMI")

# Add Regression Line

abline(lm(BMI ~ Schooling, data = health\_data), col = "blue")