NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

MH3511 Data Analysis with Computer
Group Project

Name	Matriculation Number
Nguyen Viet Dung	
Keerthana Jayaraman Karthikeyan	
Teo Wei Yew	
Devlin Nathan Waluja	
Duong Hoang Vu Lam	

Abstract

With the earliest documented work done in the 1660s, Life Expectancy has been a crucial indicator for evaluating a country's well-being. Nevertheless, a considerable discrepancy in life expectancy can be observed worldwide. While major causes for these gaps are the demographic variables of the country including Gross Domestic Product (GDP), development status, health-related factors such as immunization coverage and the Human Development Index indirectly influence life expectancy via mortality rate of young people. Therefore, we want to identify the most significant predictors of life expectancy, helping policymakers pinpoint the key areas for improvement.

Content Page

1. Introduction

With the earliest documented work done in the 1660s, *Life Expectancy* has been a crucial indicator for evaluating a country's well-being. Nevertheless, a considerable discrepancy in life expectancy can be observed worldwide. For example, in 2024, the country with the highest life expectancy recorded was Monaco at 86.49 years old, while the country with the lowest life expectancy was Nigeria at 54.63 years old [1].

In our project, a dataset containing the life expectancy of all countries from the year 2000 to 2015 is used, with other variables such as demographic variables (*GDP, Population, Development Status, etc*) and health-related variables (*HDI, Alcohol Consumption Hepatitis B immunization, Polio immunization, etc*). Based on this dataset, we seek to answer the following popular questions around *Life Expectancy*:

- Does *Life Expectancy* vary over the 16-year period from 2000 to 2015*?*
- How does the level of *Immunization affect Life Expectancy?*
- Are there factors that affect *Life Expectancy* regardless of a country's *Development Status*?

This report will cover the data descriptions and analysis using R language. For each of our research objectives, we performed statistical analysis and drew conclusions in the most appropriate approach, together with explanations and elaborations.

2. Data Description

The dataset, titled "Life Expectancy and Health Factors", is obtained from the Global Health Observatory (GHO) data repository under the World Health Organization (WHO). The original data consists of multiple sources, including health-related data from WHO and economic data from the United Nations and World Bank websites. The dataset covers 179 countries from 2000 to 2015, focusing on immunization, mortality, economic, and social factors. The data was merged into a single dataset with 22 columns.

Before conducting data analysis, we first carried out preliminary data cleaning to ensure that:

- Remove irrelevant columns ;
- Remove all rows having N.A values ;
- Rename columns ;
- Group 2 columns named *Economy_status_Developed* and *Economy_status_Developing* into 1 categorical column named *Status* with 2 options, including *Developed* and *Developing* ;

After all the preparation, 2864 observations with 13 variables are retained for analysis:

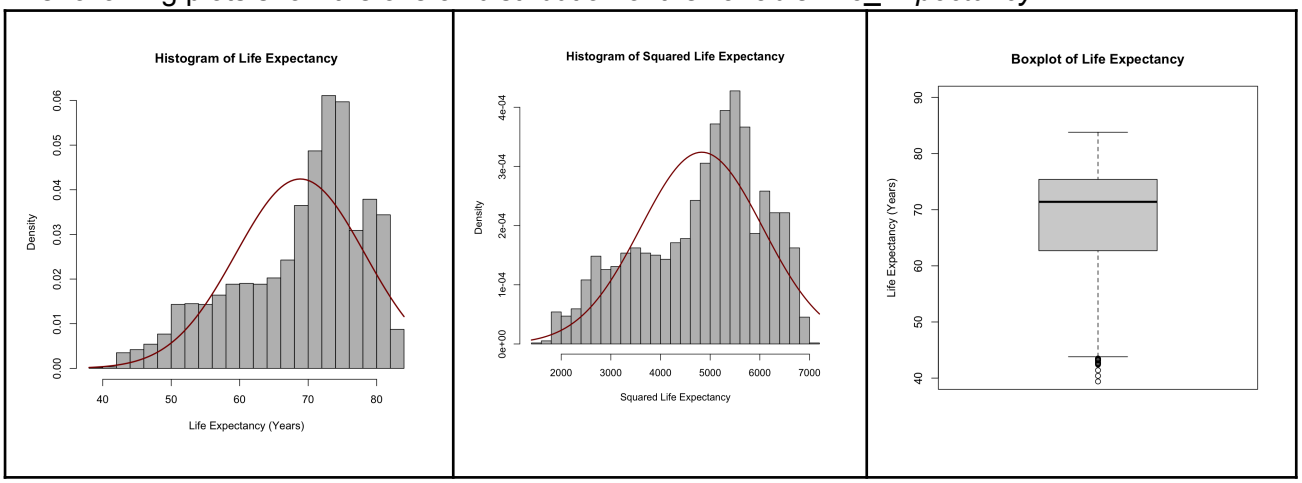
1. *Life_Expectancy* (numeric): Average life expectancy of both genders.
2. *Status* (character): Developed or Developing status.
3. *Country* (character): Name of countries.
4. *Population* (numeric): Population of the country. (in millions)
5. *GDP* (numeric): Gross Domestic Product per capita (in USD).
6. *BMI* (numeric): A measure for indicating nutritional status in adults. It is defined as a person's weight in kilograms divided by the square of the person's height in metres (kg/m^2)

7. *Alcohol* (numeric): Alcohol consumption that is recorded in liters of pure alcohol per capita with 15+ years old.
8. *Hepatitis_B* (numeric): Hepatitis B (HepB) immunization coverage among 1-year-olds (%).
9. *Polio* (numeric): Polio (Pol3) immunization coverage among 1-year-olds (%).
10. *Diphtheria* (numeric): Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%).
11. *Measles* (numeric): Measles containing vaccine first dose (MCV1) immunization coverage among 1-year-olds (%).
12. *Schooling* (numeric): Average years that people aged 25+ spent in formal education.
13. *Year* (numeric): Indicates the year of reporting.

3. Description and Cleaning of Dataset

3.1 Summary statistics for the main variable of interest, *Life_Expectancy*

The following plots show the overall distribution of the variable *Life_Expectancy*:



The life expectancy distribution exhibits a slight left-skewness in its original form:

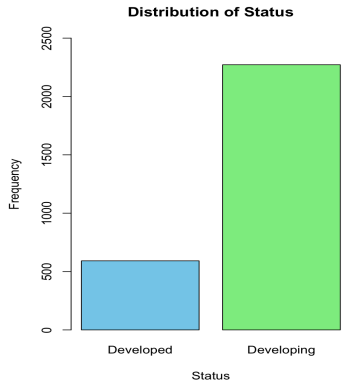
- The tail of the distribution extends further to the left towards lower life expectancies.
- The peak life expectancy is on the right side of the distribution, with a frequency almost double of the other values.

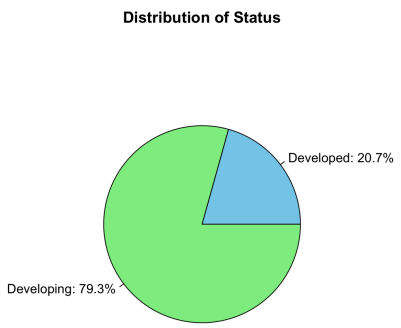
To address this non-normality, we investigated the possibility of a *square transformation* to the data. The transformed distribution reveals a more symmetric pattern, with a pronounced peak around 5000-6000 squared life expectancy units. By squaring the original values, we have mitigated the left-skewed nature of the original distribution, should there be a need for statistical tests that assume normality.

In this case, since the distribution of the original life expectancy distribution appears reasonably symmetric with a clear central tendency at 75-80 years, transformation is not necessary.

3.2 Summary statistics for other variables

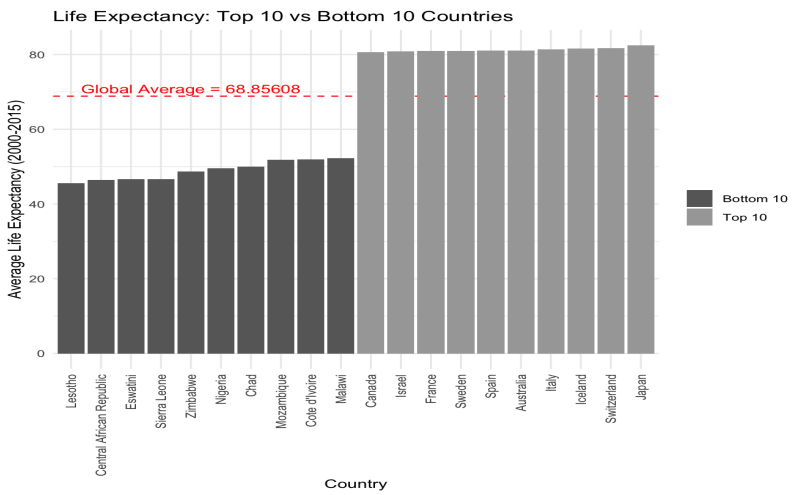
3.2.1 Country's development status (Developed/Developing), *Status*





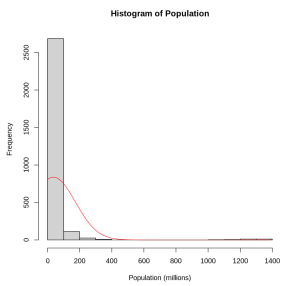
- The dataset spans 16 years (2000–2015), resulting in 2864 observations due to multiple data points recorded annually for each country
- No outlying value of *Status* is removed

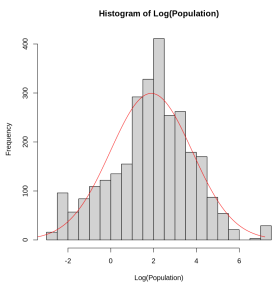
3.2.2 Name of the country, *Country*

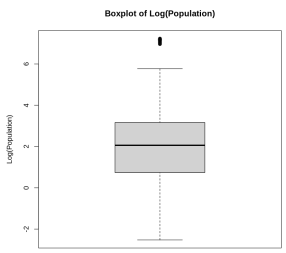


- Average life expectancy of each country was calculated
- Countries were sorted and Top 10 and Bottom 10 were selected to highlight the extremes of global life expectancy
- An evident inequality exists in global health

3.2.3 Total population of the country, *Population*

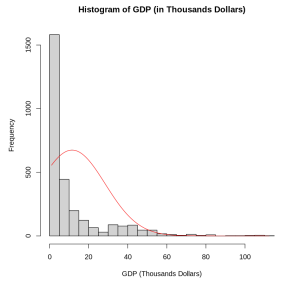


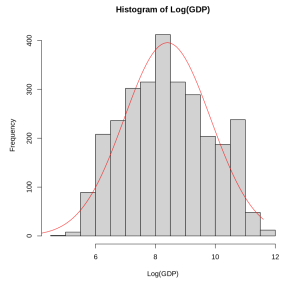


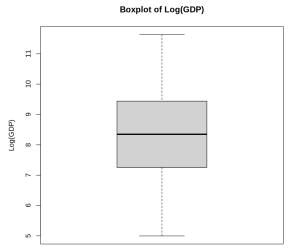


- The log-transformation (base e) is applied

3.2.4 Gross Domestic Product per capita (in USD), *GDP*

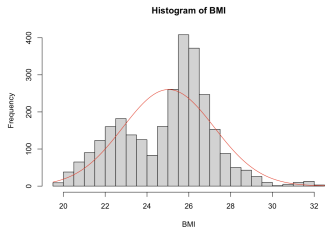


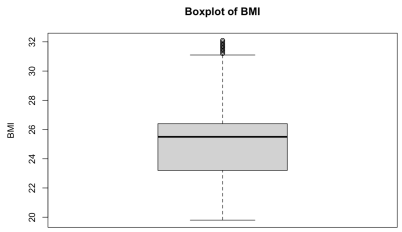




- The log-transformation (base e) is applied

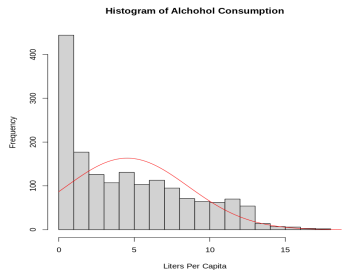
3.2.5 Body Mass Index (in kg/m^2), *BMI*

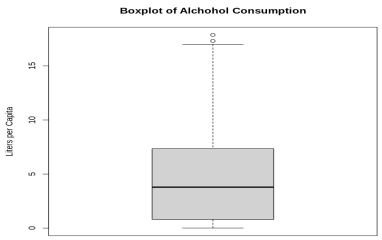




- No transformation is needed

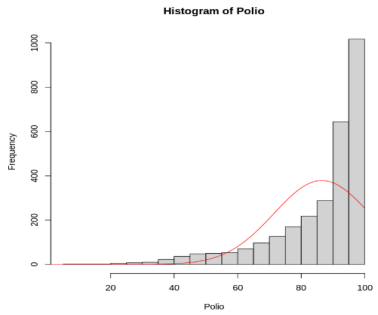
3.2.6 Alcohol consumption per capita (in liters of pure alcohol, age 15+), *Alcohol*

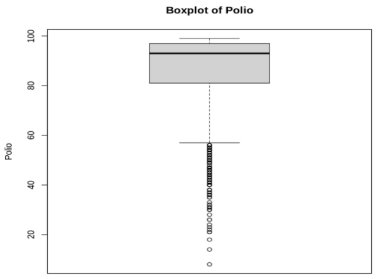




- No further transformation is applied; we wish to preserve the countries with zero alcohol consumption, such as in Muslim-majority countries

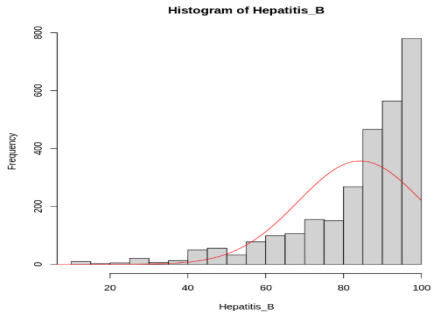
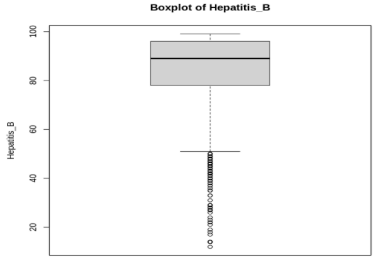
3.2.7 Polio immunization coverage among 1-year-olds (%), *Polio*



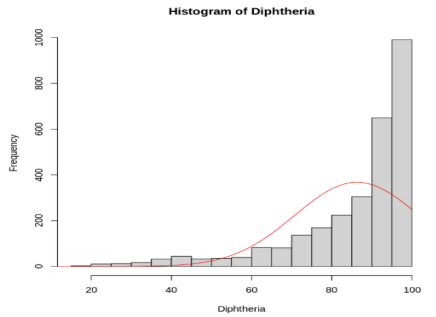
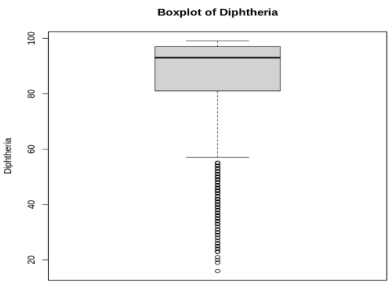


- No further transformation. The continuous values were categorized into two groups (Low, High) for statistical testing.
- No outlying value of *Polio* is removed.

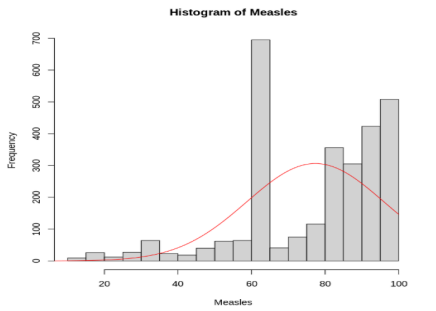
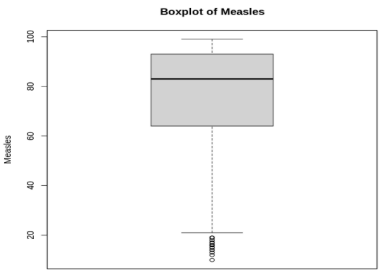
3.2.8 Hepatitis B immunization coverage among 1-year-olds (%), *Hepatitis_B*

		• No further transformation. The continuous values were categorized into two groups (Low, High) for statistical testing. • No outlying value of *Hepatitis_B* is removed.

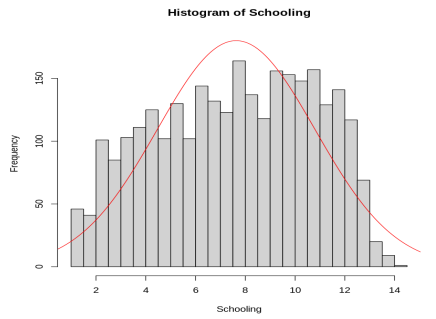
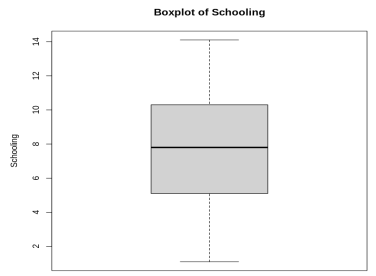
3.2.9 Diphtheria immunization coverage among 1-year-olds (%), *Diphtheria*

		• No further transformation. The continuous values were categorized into two groups (Low, High) for statistical testing. • No outlying value of *Diphtheria* is removed.

3.2.10. Measles immunization coverage among 1-year-olds (%), *Measles*

		• No further transformation. The continuous values were categorized into two groups (Low, High) for statistical testing. • No outlying value of *Measles* is removed.

3.2.11 Average years of schooling, *Schooling*

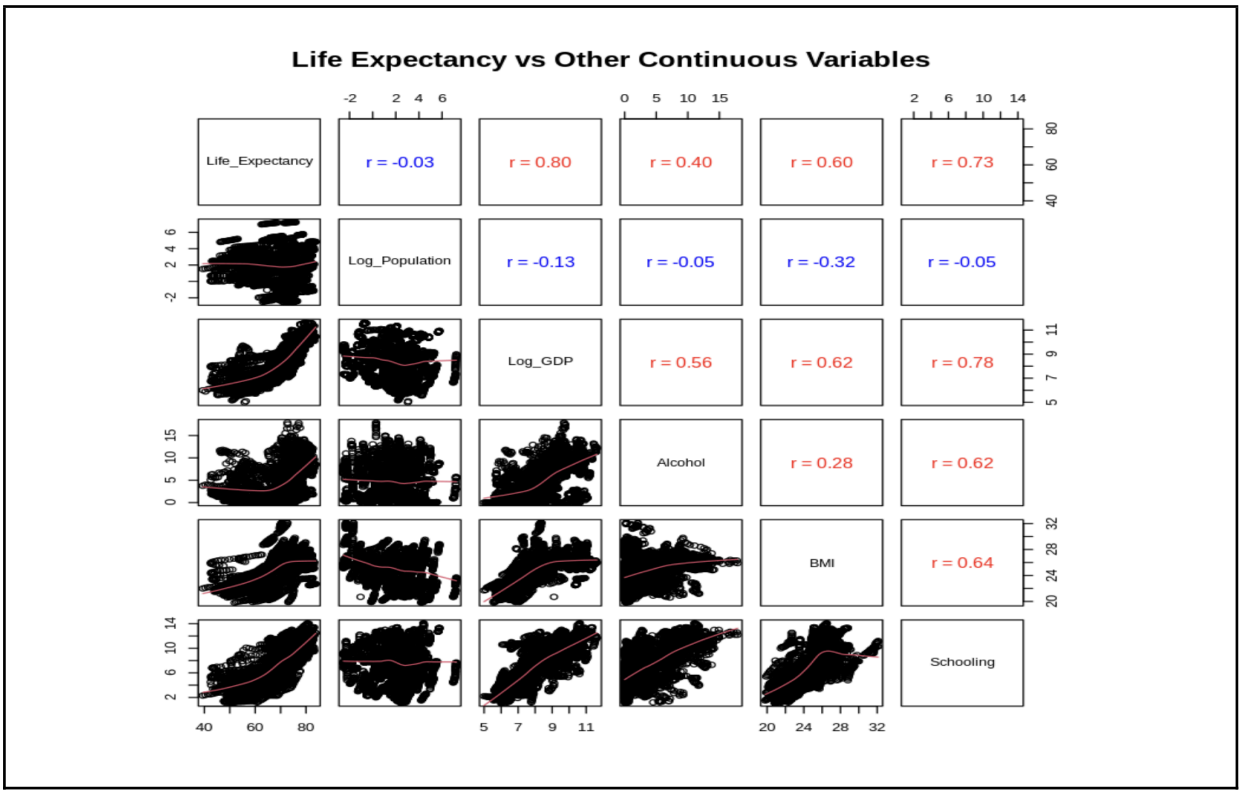
		• No outlying value of *Schooling* is removed.

3.3 Final Dataset for Analysis

Based on the above analysis, the dataset has been refined and transformed as suggested. The final dataset retains important variables across all observations, with transformations applied to improve normality and interpretability. Specifically, log-transformations have been applied to *Population* and *GDP*. After all, the dataset still remains at 2864 observations after applying the suggested transformations.

4. Statistical Analysis

4.1 Correlations between and Other Continuous Variables



Scatter plots and correlation coefficients are useful in studying the possible linear relationships between life expectancy and continuous demographic variables such as *Population, GDP, Alcohol, BMI, Schooling*.

From the plots, it appears that *Life_Expectancy* is positively correlated with all variables except *log(Population)*, in the following order: *log(GDP)* (r = 0.80), *Schooling* (r = 0.73), *BMI* (r = 0.60) and *Alcohol* (r = 0.40).

Among the continuous demographic variables, there are a few interesting observations from this tabulation:

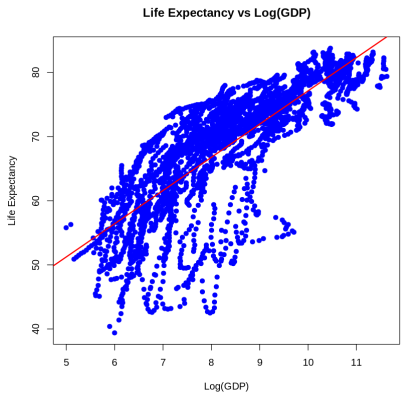
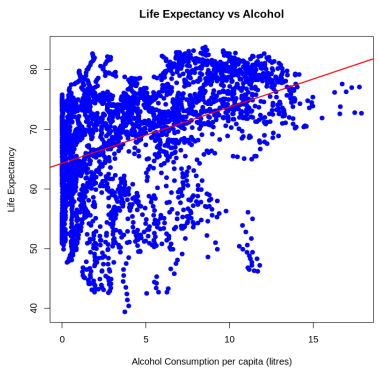
- *log(Population)* is not correlated with any other variables
- *log(GDP)* is quite highly correlated with *Alcohol, BMI* and *Schooling* (r = 0.56, 0.62 and 0.78 respectively)
- *Alcohol* is positively correlated with *BMI* and *Schooling* (r = 0.28 and 0.62 respectively)

- *BMI* and *Schooling* are also positively correlated (r = 0.64)

We shall perform some regression models and statistical tests to confirm some of our observations in the next sections.

4.2 Relationship Models

4.2.1 Linear Regression

Variable (X)	Fitted Model, with Y being Life_Expectancy	p-value	R-squared	Scatter Plot
Schooling	Y = 52.27708 + 2.17227 * X	< 2.2e-16	0.5364	
log(GDP)	Y = 25.33780 + 5.18114 * X	< 2.2e-16	0.6328	
Alcohol	Y = 64.31076 + 0.94284 * X	< 2.2e-16	0.1590	

4.2.2 Polynomial Regression between *Life_Expectancy* and *BMI*

BMI (Body Mass Index) is a ratio of an individual's weight to the square of their height, which is often used as an indicator of their weight classification. A BMI of 18 - 25 is considered normal weight, and any value below that is underweight, any value over that is considered overweight. Hence, we hypothesized that there is a non-linear polynomial relationship between life expectancy and BMI, akin to the law of diminishing returns.


```
Call:
lm(formula = Life_Expectancy ~ poly(BMI, degree = 2), data = data_2015)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.3807  -3.4332   0.1457   4.3047  18.6548 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           71.4637     0.4813 148.485  < 2e-16 ***
poly(BMI, degree = 2)1  55.1107     6.4392   8.559 5.49e-15 ***
poly(BMI, degree = 2)2 -24.1803     6.4392  -3.755 0.000235 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.439 on 176 degrees of freedom
Multiple R-squared:  0.3317,	Adjusted R-squared:  0.3241 
F-statistic: 43.68 on 2 and 176 DF,  p-value: 3.961e-16
```

As can be seen in the graph and statistics, there is a negative quadratic term (-24.1803), signifying an inverted-U shape, meaning that life expectancy increases with BMI, up until BMI ~28, after which it starts to decrease, which is what we hypothesized. The adjusted R-squared value of 0.3241 signifies a 32% variance in life expectancy due to BMI, hinting that despite BMI being a driving factor for life expectancy, there might be more influential factors.

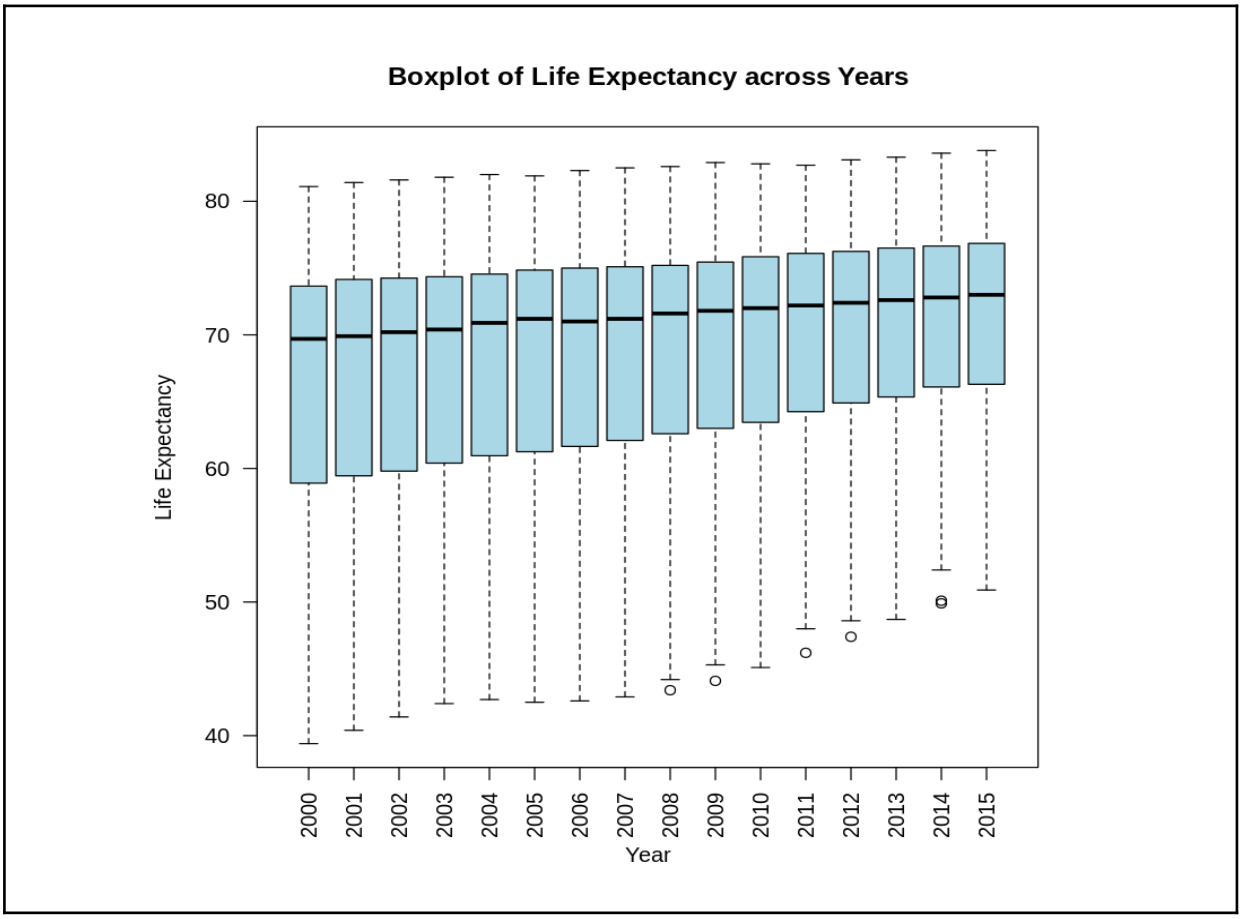
4.3 Statistical Tests

4.3.1 Does *Life_Expectancy* vary across the 16 *Years?*

Life expectancy has changed substantially in recent decades due to healthcare advancements, economic development, and policy interventions. Examining life expectancy trends across specific years (2000–2015) may help us to determine if health outcomes have improved significantly during this period.

Our research questions for this section are the following: ***Has life expectancy significantly changed over the years from 2000 to 2015? Specifically, are there particular years or periods during which life expectancy noticeably improved or declined?***

The following boxplot illustrates the distributions of Life Expectancy across the different years (2000–2015).



Looking at the boxplot, we observe a gradual upward shift in the central tendency (median) of life expectancy over the years, along with a relatively consistent spread in the distribution. This visual trend suggests potential changes in the yearly means, making the ANOVA test appropriate for formally assessing whether these differences are statistically significant. Therefore, we will use the ANOVA test to evaluate if the mean life expectancy is equal across the years. Now, we formally test the following hypotheses:

$$H_0 : \mu_{2000} = \mu_{2001} = \dots = \mu_{2015} \quad \text{against} \quad H_1\text{: not all yearly means are equal}$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Year	15	7719	514.6	5.968	2.02e-12 ***
Residuals	2848	245558	86.2		

< The result of ANOVA test >

This ANOVA test indicates a statistically significant difference among the years ($F(15, 2848) = 5.968$, p-value = 2.02×10^{-12}). The p-value, which is smaller than 0.05, suggests that we reject the null hypothesis, indicating that the observed year-to-year variations in life expectancy are statistically meaningful. Practically, this result implies that global life expectancy experienced notable annual fluctuations during this period rather than remaining stable.

While the ANOVA test confirms that life expectancy differed significantly across the years, it does not specify which years differ from one another. To address this, we continue to conduct pairwise t-tests between all combinations of years (2000–2015) without applying p-value adjustments, to preserve sensitivity in detecting changes over time.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
2001	0.76771	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2002	0.59188	0.80976	-	-	-	-	-	-	-	-	-	-	-	-	-
2003	0.40926	0.59622	0.77249	-	-	-	-	-	-	-	-	-	-	-	-
2004	0.22569	0.35954	0.49934	0.69917	-	-	-	-	-	-	-	-	-	-	-
2005	0.12755	0.21922	0.32319	0.48464	0.75470	-	-	-	-	-	-	-	-	-	-
2006	0.05799	0.10946	0.17383	0.28417	0.49357	0.70974	-	-	-	-	-	-	-	-	-
2007	0.02329	0.04842	0.08307	0.14868	0.29009	0.45595	0.70889	-	-	-	-	-	-	-	-
2008	0.00740	0.01715	0.03211	0.06370	0.14208	0.24777	0.43324	0.68156	-	-	-	-	-	-	-
2009	0.00205	0.00530	0.01084	0.02388	0.06107	0.11857	0.23453	0.41478	0.6853	-	-	-	-	-	-
2010	0.00051	0.00146	0.00326	0.00796	0.02332	0.05046	0.11317	0.22591	0.4232	0.69244	-	-	-	-	-
2011	0.00010	0.00033	0.00081	0.00221	0.00747	0.01813	0.04645	0.10561	0.2270	0.42197	0.68365	-	-	-	-
2012	2.3e-05	8.2e-05	0.00022	0.00065	0.00248	0.00666	0.01918	0.04894	0.1189	0.24847	0.44808	0.72548	-	-	-
2013	4.4e-06	1.7e-05	4.9e-05	0.00016	0.00071	0.00211	0.00686	0.01977	0.0547	0.12955	0.26250	0.47579	0.71738	-	-
2014	7.9e-07	3.4e-06	1.1e-05	3.9e-05	0.00019	0.00063	0.00230	0.00745	0.0234	0.06265	0.14255	0.28958	0.47896	0.72932	-
2015	2.1e-07	9.9e-07	3.3e-06	1.3e-05	6.8e-05	0.00024	0.00097	0.00342	0.0118	0.03461	0.08584	0.19002	0.33732	0.55013	0.80138

< Table: The result of pairwise comparisons using t tests with pooled SD >

Most notably, life expectancy in later years (especially from 2010 to 2015) was significantly higher than in the 2000s. These low p-values suggest a consistent and significant improvement in life expectancy when comparing earlier years (2000–2008) to later years (2012–2015). The statistical significance becomes much progressively stronger as the gap between years widens, indicating a steady improvement over time.

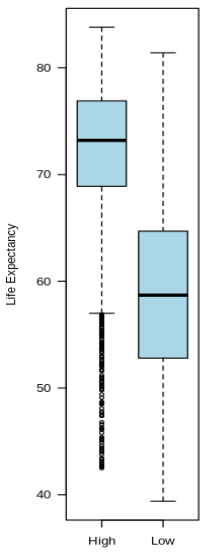
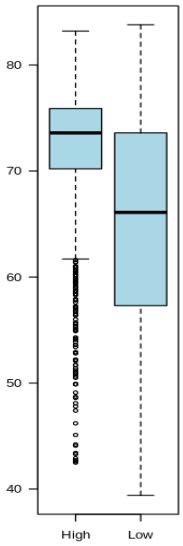
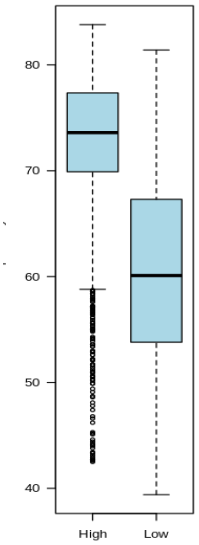
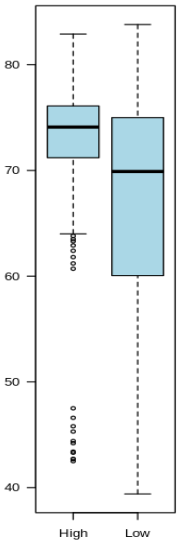
In contrast, consecutive years, such as 2010 and 2011 or 2014 and 2015, were not flagged as significantly different, implying that year-to-year changes were small and gradual. This is consistent with expectations in public health, where improvements accumulate slowly due to long-term interventions.

In conclusion, the combined results of the ANOVA and pairwise t-tests demonstrate that life expectancy has not remained constant from 2000 to 2015. Instead, there is strong statistical evidence of positive shifts, particularly when comparing the early 2000s to the mid-2010s. This reinforces the view that health outcomes are improving globally, though these improvements are incremental and best observed across longer periods.

4.3.2 How does the level of *Immunization affect Life_Expectancy?*

Immunization is a key indicator of public health success. High vaccination coverage for *Polio*, *Hepatitis B*, *Diphtheria*, and *Measles* is expected to reduce the burden of the disease and increase life expectancy. This led us to question: **Does life expectancy differ significantly across countries with low and high immunization coverage levels?**

To answer this, we conducted the analysis separately for each immunization. Coverage was grouped into two categories: High and Low. The threshold for the High group was based on the target recommended by the World Health Organization (WHO) to achieve herd immunity through routine immunization programs for diseases such as *Polio* (80%), *Diphtheria* (85%), and *Measles* (95%) [2] and for *Hepatitis B* (90%) [3].

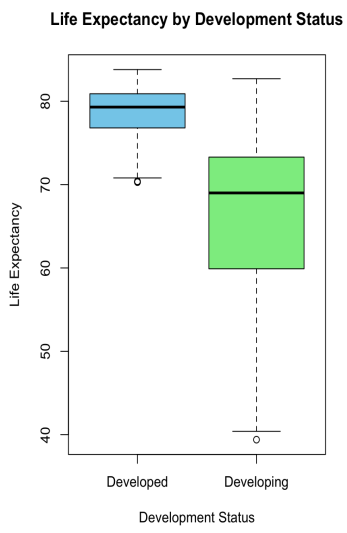
Vaccine	Polio	Hepatitis B	Diphtheria	Measles
Threshold	High >= 80% Low < 80%	High >= 90% Low < 90%	High >= 85% Low < 85%	High >= 95% Low < 95%
Boxplot				
F-test	F = 0.92606, p-value = 0.214	F = 0.40652, p-value <2.2e-16	F = 0.72452, p-value = 1.108e-08	F = 0.3212, p-value <2.2e-16
T-test	t = 37.238, df = 2862, p-value <2.2e-16	t = 20.495, df = 2477.6, p-value <2.2e-16	t = 35.967, df = 1429.9, p-value <2.2e-16	t = 18.32, df = 1800.7, p-value <2.2e-16
Mean Difference Life Expectancy (High - Low)	(71.80-59.04) =12.76	(72.25-65.56) =6.69	(72.46-60.53) =11.93	(73.24-67.63) =5.61

The boxplots clearly show that countries with higher immunization coverage have higher median life expectancy across all four vaccines. The difference in central tendency and the spread of life expectancy is particularly evident for *Polio* and *Diphtheria*, where the gap between the high and low groups is substantial.

Additionally, F-tests were conducted to check variance equality. Since variance differed significantly across groups for *Hepatitis B*, *Diphtheria*, and *Measles* ($p < 0.05$), we applied Welch's t-test, which does not assume equal variances, for these three vaccines. Pooled t-test which assumes equal variance is used for *Polio*. The t-test for each immunization showed very small p-values ($p < 2.2 * 10^{-16}$), indicating that the difference in life expectancy between the high and low immunization is statistically significant and not due to random chance.

Our analysis shows a strong positive relationship between *Immunization* and *Life Expectancy*. Countries with higher levels of immunization for *Polio*, *Hepatitis B*, *Diphtheria*, and *Measles* have significantly higher life expectancies than those with lower immunization rates. This highlights the critical role of public health initiatives and immunization programs in improving overall health and longevity across populations.

4.3.3 Relation between Life_Expectancy and Development Status

	F test to compare two variances data: Life_Expectancy by Status F = 0.12745, num df = 591, denom df = 2271, p-value <2.2e-16 95 percent confidence interval: 0.1123965 0.1452155	Welch Two Sample t-test data: Life_Expectancy by Status t = 53.685, df = 2623.8, p-value <2.2e-16 95 percent confidence interval: 11.71972 12.60832

The boxplot visually suggested a clear gap between the two groups, with Developed countries showing higher and more consistent life expectancy. Since *Status* is a binary categorical variable and *Life Expectancy* is continuous, a two-sample t-test was appropriate. Additionally, a variance test revealed a significant difference in variances ($p < 2.2 * 10^{-16}$), indicating that the assumption of equal variances was violated. Therefore, we applied Welch's t-test.

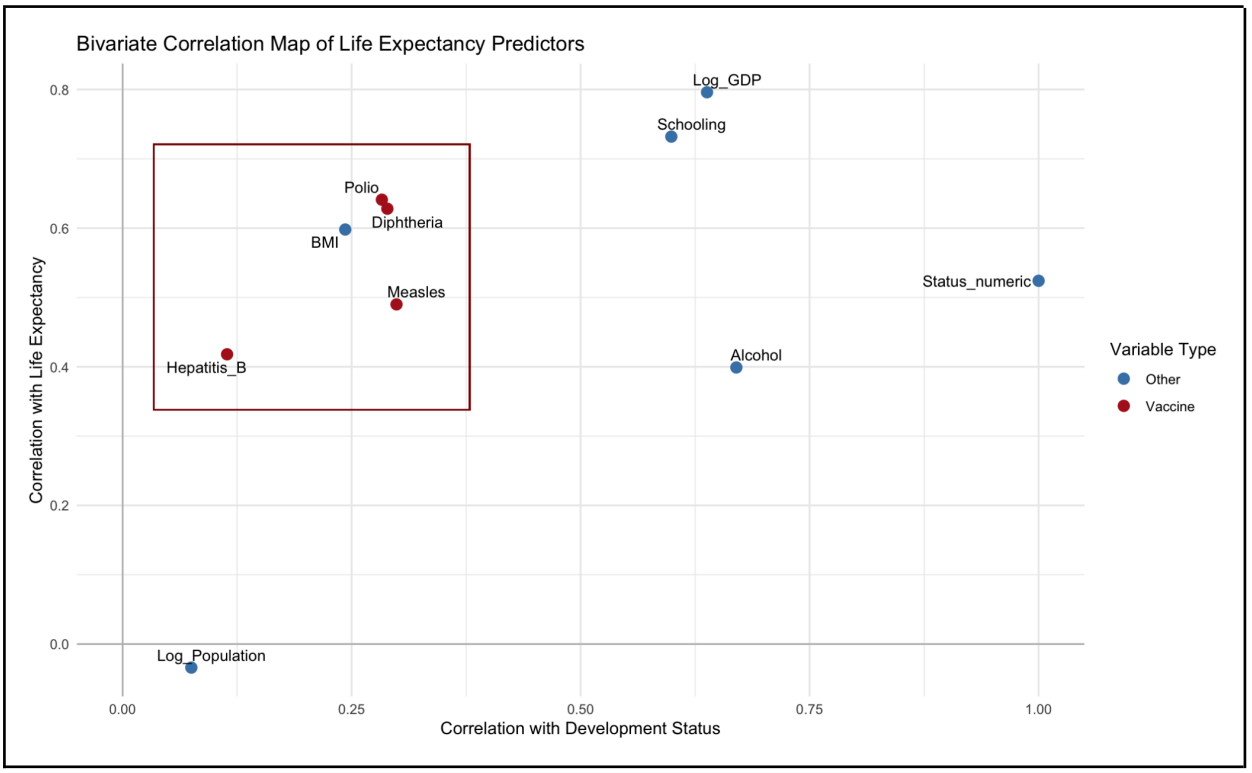
The result was highly significant ($t = 53.685$, $p < 2.2 * 10^{-16}$), and the 95% confidence interval for the difference in means is (11.72, 12.61) excludes zero by a wide margin. This indicates that on average, individuals in developed countries live over a decade longer than those in developing countries. This is a statistically significant and practically meaningful disparity.

4.3.4 Is There A Factor That Transcends Development?

Throughout this study, we examined how factors such as *GDP*, *Schooling*, and *Immunization* relate to *Life Expectancy*. While our analysis shows that these variables are associated with life expectancy, based on contextual knowledge, we tend to differ between *Developed* and *Developing* countries systematically.

This raises an important question: **Are there factors that consistently support longer life expectancy, regardless of a country's development status?** Identifying such predictors could surface interventions that are not just effective in high-income settings, but also scalable across a wide range of national contexts.

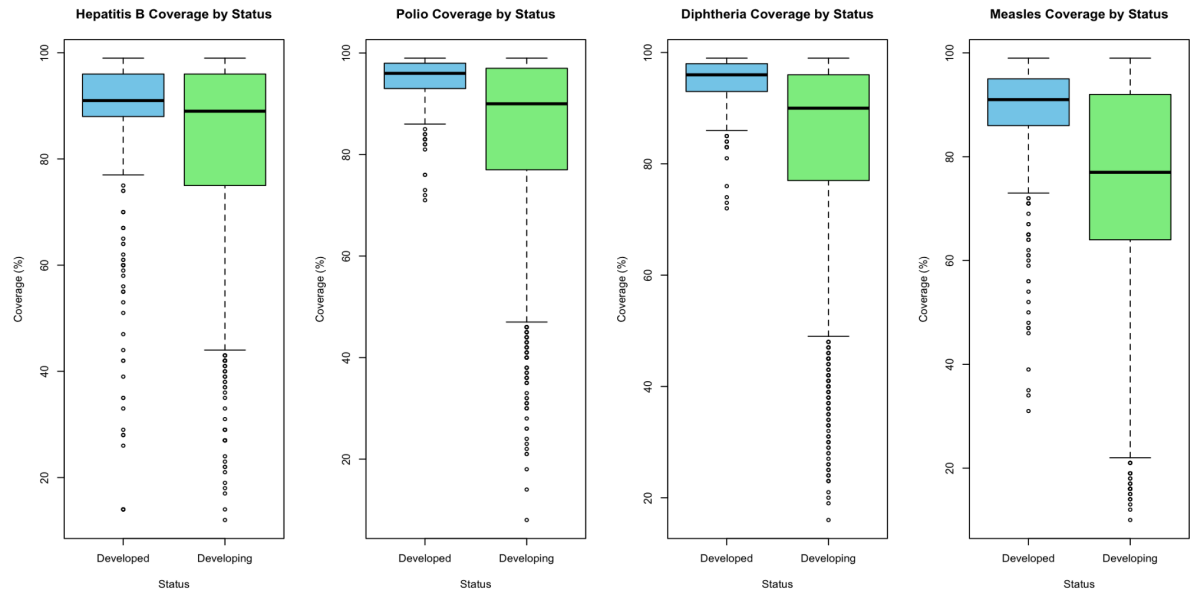
To investigate this, we converted *Status* to a numeric variable (1 = "Developed", 0 = "Developing") to compute Pearson correlations with other numeric predictors. Next, we analyzed the correlation of each numeric variable with *Life_Expectancy* and compared it to its correlation with *Status*.



Not all life expectancy predictors are created equal. While *GDP* and *Schooling* were highly correlated with life expectancy, they also closely tracked with *Development Status* — pointing to structural advantages that are not easily tackled.

In contrast, indicators such as *BMI*, *Polio*, *Diphtheria*, *Measles* and *Hepatitis B* immunization demonstrate strong associations with life expectancy, yet show only weak to moderate ties with development status. These findings highlight immunization as a scalable, impactful intervention, especially in resource-constrained settings.

4.3.4.1 A Closer Look at *Immunization*: Beyond the Developed–Developing Divide



F = 0.80118, p-value = 0.0009441	F = 0.062355 p-value < 2.2e-16	F = 0.053165 p-value < 2.2e-16	F = 0.27024 p-value < 2.2e-16
t = 6.523, df = 1010.1, p-value = 1.088e-10	t = 28.034, df = 2858.9, p-value < 2.2e-16	t = 29.082, df = 2838.2, p-value < 2.2e-16	t = 23.779, df = 1835.9, p-value < 2.2e-16

While our boxplots show notable overlap in vaccine coverage between developed and developing countries, the results of Welch's t-tests reveal that these differences are statistically significant for all four vaccines analyzed ($p < 0.001$). Preliminary F-tests confirmed unequal variances, justifying the use of Welch's method. On average, developed countries maintain higher coverage levels, but some developing countries achieve comparably strong rates, and certain developed nations fall short.

This suggests that vaccine coverage is not solely dictated by a country's economic status, but likely depends on the effectiveness of healthcare infrastructure and public health policies. Therefore, improving access to vaccines represents an impactful health intervention applicable worldwide, offering meaningful potential for improving population health regardless of a country's economic standing.

4.3.4.2 A Closer Look at *BMI*: Beyond the Developed–Developing Divide

As established, BMI's impact on life expectancy transcends development status. This insight provides a clear focus: the sensibilization of residents to the importance of healthy eating and exercise in order to maintain a healthy BMI. Furthermore, education is cost-effective, meaning that educating the public on the importance of maintaining healthy BMI levels provides the greatest price-value proposition.

5. Conclusion and Discussion

Life expectancy is widely regarded as one of the most fundamental indicators of a country's overall development, combined with economic growth, education, healthcare access, and public health strategies. In this project, we explored various demographic, economic and healthcare-related factors to better understand what drives differences in life expectancy across the world. Using data from 2000 to 2015, we conducted statistical analysis, regression modeling, and hypothesis testing to uncover key relationships and patterns.

We found that:

- When a nation's economic strength increases and its people spend more years in school, life expectancy in that country will improve noticeably. We postulate that better education can equip individuals with the knowledge to make healthier choices, while a stronger economy may bring better healthcare, cleaner environments, and improved living conditions. All of these contribute to longer lifespans.

- People with a *BMI* in the moderate-to-high range may experience longer lives. On the other hand, individuals with very low *BMI* may face risks linked to undernourishment, while those with higher *BMI* may encounter health issues related to obesity, such as heart disease or diabetes. This pattern reflects a universal truth: maintaining a healthy body, neither too light nor too heavy, is key to living a longer life.

- *Life Expectancy* around the world had steadily risen from 2000 to 2015. But these changes did not happen in the short term. The most noticeable improvements began to appear in the begin-2010s, suggesting that it may take 7 to 10 years for the full impact of health policies and medical advancements to become visible in population health.

- Countries that achieved high coverage for vaccines such as *Polio*, *Hepatitis B*, *Diphtheria*, and *Measles* consistently had longer-living populations. These results suggest that countries, especially those aiming to improve long-term health outcomes, should prioritize expanding access to routine immunization programs.

- While it's true that developed countries may experience longer and more consistent life expectancies, our findings highlight that some certain factors, such as maintaining a healthy *BMI* and achieving strong immunization coverage, can hold powerful potential regardless of a nation's economic standing.

While the findings of this project are interesting, we must acknowledge its limitations. The analysis is based on observational data from a single 16-year period and may not include other factors such as cultural habits, healthcare quality, or environmental influences. Future research could include more recent data, advanced modeling techniques, and broader indicators to strengthen these findings.

6. Appendix

6.1. Pre-processing Data

```
data = read.csv("Life Expectancy Data Updated.csv")

data$Status = ifelse(
data$Economy_status_Developed == 1 & data$Economy_status_Developing == 0,
"Developed", ifelse(data$Economy_status_Developed == 0 &
data$Economy_status_Developing == 1, "Developing", NA)
)

data =
data[,c('Life_expectancy','Status','Country','Population_mln','GDP_per_ca
pita','Alcohol_consumption','Hepatitis_B','Polio','Diphtheria','BMI','Mea
sles','Schooling', 'Year') ]

colnames(data) =
c('Life_Expectancy','Status','Country','Population','GDP','Alcohol','Hepa
titis_B','Polio','Diphtheria','BMI','Measles','Schooling', 'Year')

cleaned_data = na.omit(data)
```

6.2 Summary statistics for the main variable of interest, Life_Expectancy

```
hist(cleaned_data$Life_Expectancy,
     main = "Histogram of Life Expectancy",
     xlab = "Life Expectancy (Years)",
     breaks = 20,
     freq = FALSE,
     col = "grey")

curve(dnorm(x, mean = mean(cleaned_data$Life_Expectancy, na.rm = TRUE),
            sd = sd(cleaned_data$Life_Expectancy, na.rm = TRUE)),
      add = TRUE,
      col = "darkred",
      lwd = 2)

# Histogram with normal curve for Squared Life Expectancy
cleaned_data$Life_Expectancy_Squared <- cleaned_data$Life_Expectancy^2

hist(cleaned_data$Life_Expectancy_Squared,
     main = "Histogram of Squared Life Expectancy",
```

```
     xlab = "Squared Life Expectancy",
     breaks = 20,
     freq = FALSE,
     col = "grey")

curve(dnorm(x, mean = mean(cleaned_data$Life_Expectancy_Squared, na.rm =
TRUE),
            sd = sd(cleaned_data$Life_Expectancy_Squared, na.rm = TRUE)),
      add = TRUE,
      col = "darkred",
      lwd = 2)
```

6.3 Summary statistics for other variables

6.3.1 Country's development status (Developed/Developing), Status

```
status_counts <- table(cleaned_data$Status)

barplot(status_counts,
        main = "Distribution of Status",
        xlab = "Status",
        ylab = "Frequency",
        ylim = c(0, 2500),
        col = c("skyblue","lightgreen"),
        border = "black")

percent_labels <- paste0(names(status_counts), ": ",
                         round(100 * status_counts / sum(status_counts),
1), "%")

pie(status_counts,
    labels = percent_labels,
    main = "Distribution of Status",
    col = c("skyblue","lightgreen"))
```

6.3.2 Name of the country, Country

```
# average life expectancy per country across all years
country_avg <- aggregate(Life_Expectancy ~ Country, cleaned_data, mean)
sorted_countries <- country_avg[order(country_avg$Life_Expectancy,
decreasing = TRUE),]

# top 10 and bottom 10 countries
```

```
top10 <- sorted_countries[1:10,]
bottom10 <-
sorted_countries[(nrow(sorted_countries)-9):nrow(sorted_countries),]
contrast_countries <- rbind(top10, bottom10)
contrast_countries$Group <- c(rep("Top 10", 10), rep("Bottom 10", 10))

contrast_countries$Country <- factor(contrast_countries$Country,
                                     levels =
contrast_countries$Country[order(contrast_countries$Life_Expectancy)])

ggplot(contrast_countries, aes(x = Country, y = Life_Expectancy, fill =
Group)) +

global_avg <- mean(cleaned_data$Life_Expectancy)
global_avg
ggplot(contrast_countries, aes(x = Country, y = Life_Expectancy, fill =
Group)) +
  geom_bar(stat = "identity") +
  geom_hline(yintercept = global_avg, linetype = "dashed", color = "red")
+
  annotate("text", x = 5, y = global_avg + 2, label = "Global Average =
68.85608", color = "red") +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("Top 10" = "#A9A9A9", "Bottom 10" =
"#696969")) +
  labs(title = "Life Expectancy: Top 10 vs Bottom 10 Countries",
       x = "Country",
       y = "Average Life Expectancy (2000-2015)",
       fill = "") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```

6.3.3 Total population of the country, Population

```
hist(cleaned_data$Population,
    main="Histogram of Population",
    xlab="Population (millions)",
    breaks = 20,
    )
xpt = seq(1,max(cleaned_data$Population),by=0.1)
n_den = dnorm( xpt, mean(cleaned_data$Population),
sd(cleaned_data$Population) )
ypt = n_den*length(cleaned_data$Population)
```

```
lines( xpt , ypt , col = 'red' )

hist(log(cleaned_data$Population),
    main="Histogram of Log(Population)",
    xlab="Log(Population)",
    breaks = 20,
    )
xpt = seq(1,max(log(cleaned_data$Population)),by=0.1)
n_den = dnorm( xpt, mean(log(cleaned_data$Population)),
sd(log(cleaned_data$Population)) )
ypt = n_den*length( log(cleaned_data$Population) )
lines( xpt , ypt , col = 'red' )

boxplot( log(cleaned_data$Population),
       main = "Boxplot of Log(Population)",
       ylab = "Log(Population)")
```

6.3.4 Gross Domestic Product per capita (in USD), GDP

```
hist(cleaned_data$GDP/1000,
    main="Histogram of GDP (in Thousands Dollars)",
    xlab="GDP (Thousands Dollars)",
    breaks = 20,
    )

hist(log(cleaned_data$GDP),
    main="Histogram of Log(GDP)",
    xlab="Log(GDP)",
    breaks = 20,
    )
xpt = seq(1,max(log(cleaned_data$GDP)),by=0.1)
n_den = dnorm( xpt, mean(log(cleaned_data$GDP)),
sd(log(cleaned_data$GDP)) )
ypt = n_den*length( log(cleaned_data$GDP) )
lines( xpt , ypt , col = 'red' )

boxplot( log(cleaned_data$GDP),
       main = "Boxplot of Log(GDP)",
       ylab = "Log(GDP)")
```

6.3.5 Body Mass Index (in kg/m2), BMI

```
hist(cleaned_data$BMI,
```

```
    main="Histogram of BMI",
    xlab="BMI",
    breaks = 20,
)

xpt = seq(0,80,by=0.1)
n_den = dnorm( xpt, mean(cleaned_data$BMI), sd(cleaned_data$BMI) )
ypt = n_den*length( cleaned_data$BMI ) * 1
lines( xpt , ypt , col = 'red' )

boxplot(cleaned_data$BMI,
       main = "Boxplot of BMI",
       ylab = "BMI")
```

6.3.6 Alcohol consumption per capita (in liters of pure alcohol, age 15+), Alcohol

```
hist(cleaned_data$Alcohol,
    main="Histogram of Alchohol Consumption",
    xlab="Liters Per Capita",
    breaks = 20,
)

xpt = seq(0,25,by=0.1)
n_den = dnorm( xpt, mean(cleaned_data$Alcohol), sd(cleaned_data$Alcohol)
)
ypt = n_den*length( cleaned_data$Alcohol ) * 1
lines( xpt , ypt , col = 'red' )

boxplot(cleaned_data$Alcohol,
       main = "Boxplot of Alchohol Consumption",
       ylab = "Liters per Capita")
```

6.3.7 Polio immunization coverage among 1-year-olds (%), Polio

```
hist(cleaned_data$Polio,
     main="Histogram of Polio", xlab="Polio",
     breaks = 20,
     )

xpt = seq(0,100,by=0.1)
n_den = dnorm(xpt, mean(cleaned_data$Polio), sd(cleaned_data$Polio))
ypt = n_den * length(cleaned_data$Polio) * 5
```

```
lines( xpt , ypt , col = 'red')

boxplot(cleaned_data$Polio,
        main = "Boxplot of Polio",
        ylab = "Polio")
```

6.3.8 Hepatitis B immunization coverage among 1-year-olds (%), Hepatitis_B

```
hist(cleaned_data$Hepatitis_B,
     main="Histogram of Hepatitis_B", xlab="Hepatitis_B",
     breaks = 20,
     )

xpt = seq(0,100,by=0.1)
n_den = dnorm(xpt, mean(cleaned_data$Hepatitis_B),
sd(cleaned_data$Hepatitis_B))
ypt = n_den * length(cleaned_data$Hepatitis_B) * 5
lines( xpt , ypt , col = 'red')

boxplot(cleaned_data$Hepatitis_B,
        main = "Boxplot of Hepatitis_B",
        ylab = "Hepatitis_B")
```

6.3.9 Diphtheria immunization coverage among 1-year-olds (%), Diphtheria

```
hist(cleaned_data$Diphtheria,
     main="Histogram of Diphtheria", xlab="Diphtheria",
     breaks = 20,
     )

xpt = seq(0,100,by=0.1)
n_den = dnorm(xpt, mean(cleaned_data$Diphtheria),
sd(cleaned_data$Diphtheria))
ypt = n_den * length(cleaned_data$Diphtheria) * 5
lines( xpt , ypt , col = 'red')

boxplot(cleaned_data$Diphtheria,
        main = "Boxplot of Diphtheria",
        ylab = "Diphtheria")
```

6.3.10. Measles immunization coverage among 1-year-olds (%), Measles

```
hist(cleaned_data$Measles,
```

```
     main="Histogram of Measles", xlab="Measles",
     breaks = 20,
     )

xpt = seq(0,100,by=0.1)
n_den = dnorm(xpt, mean(cleaned_data$Measles), sd(cleaned_data$Measles))
ypt = n_den * length(cleaned_data$Measles) * 5
lines( xpt , ypt , col = 'red')

boxplot(cleaned_data$Measles,
        main = "Boxplot of Measles",
        ylab = "Measles")
```

6.3.11 Average years of schooling, Schooling

```
hist(cleaned_data$Schooling,
    main = "Histogram of Schooling",
    xlab = "Schooling",
    breaks = 20,
    ylim = c(1, 180)
)

xpt = seq(0,25,by=0.1)
n_den = dnorm( xpt, mean(cleaned_data$Schooling),
sd(cleaned_data$Schooling) )
ypt = n_den*length( cleaned_data$Schooling ) * 0.5
lines( xpt , ypt , col = 'red' )

boxplot(cleaned_data$Schooling,
       main = "Boxplot of Schooling",
       ylab = "Schooling")
```

6.4 Final Dataset for Analysis

```
final_data = cleaned_data

final_data$Population = log(final_data$Population)
final_data$GDP = log(final_data$GDP)

colnames(final_data)[colnames(final_data) == "Population"] =
"Log_Population"
colnames(final_data)[colnames(final_data) == "GDP"] = "Log_GDP"
```

6.5. Statistical Analysis

6.5.1 Correlations between and Other Continuous Variables

```
panel.cor <- function(x, y, digits = 2, prefix = "r = ", ...) {
 usr <- par("usr"); on.exit(par(usr))
 par(usr = c(0, 1, 0, 1))
 r <- cor(x, y, use = "complete.obs")
 txt <- paste0(prefix, formatC(r, digits = digits, format = "f"))
 col <- ifelse(r > 0, "red", "blue")
 text(0.5, 0.5, txt, cex = 1.2, col = col)
}

selected_data <- final_data[, c("Life_Expectancy", "Log_Population",
"Log_GDP", "Alcohol", "BMI", "Schooling")]

pairs(selected_data,
     upper.panel = panel.cor,
     lower.panel = panel.smooth,
     main = "Life Expectancy vs Other Continuous Variables")
```

6.5.2 Linear Regression

```
# LE ~ GDP

lm_model <- lm(Life_Expectancy ~ Log_GDP, data = final_data)
summary(lm_model)

plot(final_data$Log_GDP, final_data$Life_Expectancy,
    main = "Life Expectancy vs Log(GDP)",
    xlab = "log(GDP)",
    ylab = "Life Expectancy",
    col = "blue",
    pch = 19)

abline(lm(Life_Expectancy ~ Log_GDP, data = final_data), col = "red", lwd
= 2)

# LE ~ Schooling

lm_model <- lm(Life_Expectancy ~ Schooling, data = final_data)
summary(lm_model)
```

```
plot(final_data$Schooling, final_data$Life_Expectancy,
    main = "Life Expectancy vs Schooling",
    xlab = "Schooling Years",
    ylab = "Life Expectancy",
    col = "blue",
    pch = 19)

abline(lm(Life_Expectancy ~ Schooling, data = final_data), col = "red",
lwd = 2)

# LE ~ Alcohol

lm_model <- lm(Life_Expectancy ~ Alcohol, data = final_data)
summary(lm_model)

plot(final_data$Alcohol, final_data$Life_Expectancy,
    main = "Life Expectancy vs Alcohol",
    xlab = "Alcohol Consumption per capita (litres)",
    ylab = "Life Expectancy",
    col = "blue",
    pch = 19)

abline(lm(Life_Expectancy ~ Alcohol, data = final_data), col = "red", lwd
= 2)
```

6.5.3 Polynomial Regression between Life_Expectancy and BMI

```
# Subset the data for 2015
data_2015 <- subset(final_data, Year == 2015)

model <- lm(Life_Expectancy ~ poly(BMI, degree = 2), data = data_2015)
plot(data_2015$BMI, data_2015$Life_Expectancy,
    xlab = "BMI", ylab = "Life Expectancy",
    main = "Life Expectancy vs BMI (2015)")
bmi_seq <- seq(min(data_2015$BMI, na.rm = TRUE),
               max(data_2015$BMI, na.rm = TRUE),
               length.out = 200)

predicted_life <- predict(model, newdata = data.frame(BMI = bmi_seq))

lines(bmi_seq, predicted_life, col = "red", lwd = 2)

summary(model)
```

6.5.4 Does Life_Expectancy vary across the 16 Years?

```
boxplot(Life_Expectancy ~ Year,
       data = final_data,
       col = "lightblue",
       main = "Boxplot of Life Expectancy across Years",
       xlab = "Year",
       ylab = "Life Expectancy",
       las = 2)

anova_model = aov(Life_Expectancy ~ factor(Year), data = final_data)
summary(anova_model)

pairwise.t.test(final_data$Life_Expectancy, final_data$Year,
p.adjust.method = "none")
```

6.5.5 How does the level of Immunization affect Life_Expectancy?

```
final_data$PolioGroup <- ifelse(final_data$Polio < 80, "Low", "High")
final_data$Hepititis_BGroup <- ifelse(final_data$Hepatitis_B < 90, "Low",
"High")
final_data$DiphtheriaGroup <- ifelse(final_data$Diphtheria < 85, "Low",
"High")
final_data$MeaslesGroup <- ifelse(final_data$Measles < 95, "Low", "High")


par(mfrow = c(1, 4), mar = c(12, 4, 4, 1))

boxplot(Life_Expectancy ~ PolioGroup,
        data = final_data,
        col = "lightblue",
        main = "Polio",
        xlab = "Polio Group",
        ylab = "Life Expectancy",
        las = 1)

boxplot(Life_Expectancy ~ Hepititis_BGroup,
        data = final_data,
        col = "lightblue",
        main = "Hepatitis B",
        xlab = "Hepatitis B Group",
        ylab = "Life Expectancy",
        las = 1)
```

```
boxplot(Life_Expectancy ~ DiphtheriaGroup,
        data = final_data,
        col = "lightblue",
        main = "Diphtheria",
        xlab = "Diphtheria Group",
        ylab = "Life Expectancy",
        las = 1)

boxplot(Life_Expectancy ~ MeaslesGroup,
        data = final_data,
        col = "lightblue",
        main = "Measles",
        xlab = "Measles Group",
        ylab = "Life Expectancy",
        las = 1)

par(mfrow = c(1, 1))

#Polio
var_test <- var.test(Life_Expectancy ~ PolioGroup, data = final_data)
print(var_test)
t_test_polio <- t.test(Life_Expectancy ~ PolioGroup, data = final_data,
var.equal = TRUE)
print(t_test_polio)

#Hepititis B
var_test <- var.test(Life_Expectancy ~ Hepititis_BGroup, data =
final_data)
print(var_test)
t_test_Hepititis_B <- t.test(Life_Expectancy ~ Hepititis_BGroup, data =
final_data)
print(t_test_Hepititis_B)

#Diphtheria
var_test <- var.test(Life_Expectancy ~ DiphtheriaGroup, data =
final_data)
print(var_test)
t_test_Diphtheria <- t.test(Life_Expectancy ~ DiphtheriaGroup, data =
final_data)
print(t_test_Diphtheria)

#Measles
```

```
var_test <- var.test(Life_Expectancy ~ MeaslesGroup, data = final_data)
print(var_test)
t_test_Measles <- t.test(Life_Expectancy ~ MeaslesGroup, data =
final_data)
print(t_test_Measles)
```

6.5.6 Relation between Life_Expectancy and Development Status

```
boxplot(Life_Expectancy ~ Status, data = final_data,
        main = "Life Expectancy by Development Status",
        xlab = "Development Status",
        ylab = "Life Expectancy",
        col = c("skyblue", "lightgreen"))

var.test(Life_Expectancy ~ Status, data = final_data)

# T-test
t_test_result <- t.test(Life_Expectancy ~ Status, data = final_data)
print(t_test_result)
```

6.5.7 Is There A Factor That Transcends Development

```
# convert development status to numeric
final_data$Status_numeric <- ifelse(final_data$Status == "Developed", 1,
0)
numeric_data <- na.omit(final_data[sapply(final_data, is.numeric)])

# compute correlations
cor_life <- cor(numeric_data, use = "pairwise.complete.obs")[,
"Life_Expectancy"]
cor_status <- cor(numeric_data, use = "pairwise.complete.obs")[,
"Status_numeric"]

# results combined into a dataframe
cor_df <- data.frame(
  Variable = names(cor_life),
  Corr_with_Life_Expectancy = cor_life,
  Corr_with_Status = cor_status
)

# filter and clean
cor_df <- cor_df[cor_df$Variable != "Life_Expectancy" & cor_df$Variable
!= "Year", ]
```

```
cor_df$Variable <- as.character(cor_df$Variable)

cor_df$Corr_with_Life_Expectancy <-
round(cor_df$Corr_with_Life_Expectancy, 3)
cor_df$Corr_with_Status <- round(cor_df$Corr_with_Status, 3)
cor_df$Type <- ifelse(cor_df$Variable %in% c("Hepatitis_B", "Polio",
"Diphtheria", "Measles"), "Vaccine", "Other")

# Get the bounding box around the four variables
box_data <- cor_df[cor_df$Variable %in% c("Polio", "BMI", "Measles",
"Hepatitis_B"), ]
x_min <- min(box_data$Corr_with_Status) - 0.08
x_max <- max(box_data$Corr_with_Status) + 0.08
y_min <- min(box_data$Corr_with_Life_Expectancy) - 0.08
y_max <- max(box_data$Corr_with_Life_Expectancy) + 0.08


ggplot(cor_df, aes(x = Corr_with_Status, y = Corr_with_Life_Expectancy,
label = Variable)) +
  geom_point(aes(color = Type), size = 3) +
  geom_text_repel(size = 3.5, max.overlaps = 100) +
  geom_rect(aes(xmin = x_min, xmax = x_max, ymin = y_min, ymax = y_max),
            inherit.aes = FALSE, fill = NA, color = "darkred", linewidth
= 0.5) +
  geom_vline(xintercept = 0, color = "grey") +
  geom_hline(yintercept = 0, color = "grey") +
  scale_color_manual(values = c("Vaccine" = "firebrick", "Other" =
"steelblue")) +
  labs(
    title = "Bivariate Correlation Map of Life Expectancy Predictors",
    x = "Correlation with Development Status",
    y = "Correlation with Life Expectancy",
    color = "Variable Type"
  ) +
  theme_minimal()
```

6.5.7.1 A Closer Look at Immunization: Beyond the Developed–Developing Divide

```
par(mfrow = c(1, 4))

boxplot(Hepatitis_B ~ Status, data = final_data,
        main = "Hepatitis B Coverage by Status",
        ylab = "Coverage (%)",
```

```
        col = c("skyblue", "lightgreen"))

boxplot(Polio ~ Status, data = final_data,
        main = "Polio Coverage by Status",
        ylab = "Coverage (%)",
        col = c("skyblue", "lightgreen"))

boxplot(Diphtheria ~ Status, data = final_data,
        main = "Diphtheria Coverage by Status",
        ylab = "Coverage (%)",
        col = c("skyblue", "lightgreen"))

boxplot(Measles ~ Status, data = final_data,
        main = "Measles Coverage by Status",
        ylab = "Coverage (%)",
        col = c("skyblue", "lightgreen"))

var.test(Hepatitis_B ~ Status, data = final_data)
var.test(Polio ~ Status, data = final_data)
var.test(Diphtheria ~ Status, data = final_data)
var.test(Measles ~ Status, data = final_data)

t.test(Hepatitis_B ~ Status, data = final_data)
t.test(Polio ~ Status, data = final_data)
t.test(Diphtheria ~ Status, data = final_data)
t.test(Measles ~ Status, data = final_data)
```

7. References

[1] Earth Data, *Life expectancy*, 2024. [Online]. Available: https://database.earth/population/life-expectancy/2024

[2] World Health Organization, *Coronavirus disease (COVID-19): Herd immunity, lockdowns and COVID-19*. [Online]. Available: https://www.who.int/news-room/questions-and-answers/item/herd-immunity-lockdowns-and-covid-19

[3] G. H. H. and S. Programmes, *Combating hepatitis B and C to reach elimination by 2030*, May 27, 2016. [Online]. Available: https://www.who.int/publications/i/item/combating-hepatitis-b-and-c-to-reach-elimination-by-2030