What to Expect with Life Expectancy?*

Examining How Healthcare Expenditure, Education, and Medical Infrastructure Influence Life Expectancy Across Developed and Developing Regions

Manjun Zhu

December 4, 2024

Life expectancy has seen unprecedented growth in recent decades, driven by advancements in healthcare, improved access to medical services, and reduced child mortality. However, this growth has not been evenly distributed across the globe, with wealthier nations benefiting disproportionately. This research examines the impact of a nation's educational standing and health conditions on life expectancy, exploring potential variables beyond those traditionally studied. By developing a linear model, the analysis identifies key relationships between education and healthcare indicators and their influence on population longevity. The findings aim to inform health policy and education planning, providing actionable insights to address disparities and promote equitable growth in life expectancy worldwide.

Table of contents

1 Introduction

Life expectancy has witnessed remarkable growth over the past century, driven by advancements in healthcare, improved access to medical services, and reduced child mortality rates. However, this progress has not been distributed equitably across the globe. While nations with robust economies and developed healthcare systems have achieved significant gains, others, particularly in underdeveloped regions, continue to face barriers such as inadequate infrastructure, persistent diseases, and economic instability. Despite existing research on life expectancy determinants, gaps remain in understanding the nuanced relationships between economic standing, healthcare access, and other critical variables.

*Code and data are available at: https://github.com/Karrrrmen/Life_Expectancy.

This study aims to examine the factors influencing life expectancy across different nations by developing a linear regression model that incorporates both healthcare and education variables. The response variable in the analysis is life expectancy, while the predictors include national status (Developed vs. Developing), diphtheria immunization coverage, government healthcare expenditure as a percentage of total expenditure, and the number of doctors and nurses per 10,000 people. Additionally, schooling is considered as an important educational factor. By quantifying the effects of these predictors, the study seeks to shed light on disparities in life expectancy between developed and developing nations, and to identify key factors contributing to these differences. Through this analysis, we aim to contribute valuable insights for policy development focused on improving healthcare and economic outcomes.

Previous studies provide valuable context for this research. One study (Garon et al. 2015) examined the long-term effects of diphtheria on survivors, revealing that paralyzed individuals experience greater reductions in life expectancy, which underscores the importance of healthcare infrastructure in mitigating preventable diseases in underdeveloped regions. Another study extended this understanding by exploring the role of the Human Development Index (HDI), emphasizing the influence of public health initiatives, economic conditions, and education on life expectancy in low and medium-HDI nations (Girum et al. 2018). Besides that, developed econometric techniques, including panel cointegration analysis, have demonstrated a positive relationship between health care expenditure per capita and life expectancy, providing a robust foundation for integrating economic variables into this study (He and Li 2020). Additionally, a finding (Hanushek and Woessmann 2016) underscored the importance of combining both educational investments and government support in improving life expectancy, particularly in developing regions. The strong correlation between education and health suggests that higher levels of education improve overall health, reducing mortality rates and extending life expectancy.

The analysis reveals that life expectancy in developed countries is generally higher and more stable, attributed to stronger healthcare systems, consistent access to medical services, and robust economic stability. These factors are typically supported by well-established public health policies. In contrast, life expectancy in developing countries demonstrates a broader range of variation, though with a noticeable upward trend. This improvement is driven by diverse factors like better healthcare access, economic development, and increasing resource availability, but the inconsistency across regions highlights disparities in infrastructure. The faster growth trajectory in these nations indicates progress, yet challenges remain in achieving stable health outcomes. These findings suggest that policymakers in developing countries should prioritize healthcare system strengthening, resource access, and reducing regional inequalities, while in developed countries, maintaining consistent healthcare access and supporting sustained economic growth should remain priorities.

The remainder of this paper is structured as follows: Section Section 2 discusses the data sources and cleaning methods, followed by an analysis of key predictors in Section Section 3. Section Section 4 presents the findings from the linear regression model, and Section Section 5 explores the implications of these findings, concluding with recommendations for policy

interventions and areas for future research.

2 Data

2.1 Overview

This study utilizes two primary datasets related to global health and life expectancy. The first dataset, Life Expectancy and Healthy Life Expectancy (Labs 2021), provides comprehensive global health data, including variables such as GDP, life expectancy, mortality rates, and access to healthcare. The second dataset, Life Expectancy (WHO) (Organization 2024), focuses specifically on life expectancy trends and factors, offering a detailed breakdown by country, year, and socioeconomic variables.

Together, these datasets enable an in-depth exploration of the relationships between life expectancy and various predictors, such as economic indicators, healthcare access, and immunization coverage. Key variables include Life Expectancy, Status (Developed vs. Developing), Diphtheria Immunization Coverage, Government Health Expenditure as percentage of Total Expenditure, Numbers of Doctors and Nurses per 10000, and Schooling.

We performed data cleaning and analysis using R (R Core Team 2023), incorporating a range of packages for efficient data manipulation, visualization, and statistical modeling, leveraging packages such as `tidyverse` for data manipulation and visualization (Wickham et al. 2019), `dplyr` for streamlined data wrangling (Wickham et al. 2023), `ggplot2` for creating high-quality plots (Wickham 2016), `ggridges` for displaying the distribution of variables across different categories with ridge plots (Wilke 2024), `ggbeeswarm` for visualizing the distribution of data points with swarming plots (Clarke, Sherrill-Mix, and Dawson 2023), `here` for managing file paths (Müller 2020), `lubridate` for handling date-related variables (Grolemund and Wickham 2011), `arrow` for data storage and processing (Richardson et al. 2024), `kableExtra` for generating polished tables (Zhu 2024), `modelsummary` for summarizing model results (Arel-Bundock 2022), `broom` for tidying up model outputs into easily readable data frames (Robinson, Hayes, and Couch 2023), and `knitr` for dynamic report generation and the integration of code with results in reproducible workflows (Xie 2023).

2.2 Measurement

The measurement process explains how real-world phenomena—such as healthcare access, economic indicators, or demographic factors—are translated into structured data entries. Each row in the dataset represents a specific country in a given year, capturing various variables related to life expectancy and its determinants. These variables originate from data collection efforts by global organizations like the World Health Organization (WHO), which rely on surveys, national health systems, and standardized reporting frameworks to compile accurate

and comparable statistics across countries. Below is a description of key variables in the dataset:

Life Expectancy (`LifeExpectancy`): This is the dependent variable, representing the average number of years a person is expected to live from birth, assuming current mortality rates. The data is reported by international organizations like the World Health Organization (WHO) and is recorded in numeric format as age in years.

Status (`Status`): This categorical variable identifies whether a country is classified as "Developed" or "Developing." The classification is based on economic and social indicators defined by the World Health Organization (WHO) or related agencies.

Percentage of Health Expenditure (`HealthExpenditure`): This represents the general government expenditure on health as a percentage of total government expenditure. It is reported by national governments or international organizations and recorded as a percentage in the dataset.

Diphtheria Immunization Coverage (`Diphtheria`): This variable reflects the percentage of one-year-olds who have received the DTP3 vaccine (diphtheria, tetanus toxoid, and pertussis). Immunization data is typically gathered through health system reports or household surveys and is recorded as a numeric percentage.

Schooling (`Schooling`): This measures the average number of years of schooling for individuals in the population. Data on schooling is sourced from education ministry reports or international databases like UNESCO, and it is recorded in years as a numeric value.

Medical Doctors (`pct_doctor`): This variable represents the number of medical doctors per 10,000 people in the population. It is typically reported by national health ministries or equivalent institutions. The dataset standardizes this figure as a numeric value to facilitate cross-country comparisons.

Nursing and Midwifery Personnel (`pct_nursing`): This measures the number of nursing and midwifery personnel per 10,000 people in the population. It is collected through similar national reporting systems and is recorded as a numeric value.

2.3 Data Cleaning

The raw Life Expectancy dataset, along with two supplementary datasets for medical doctors and nursing personnel, underwent several cleaning steps to ensure it was accurate, consistent, and ready for analysis. Initially, the columns in the supplementary datasets were renamed for clarity and compatibility with the main dataset. A left join was then performed to merge information on `pct_doctor` and `pct_nursing` into the main dataset. Subsequently, the data was filtered to include only entries from 2004 to 2015, and rows with missing values were removed to maintain data integrity. Column names were standardized, and data types were

adjusted as necessary to ensure compatibility with analytical processes. The cleaned dataset was then saved as a Parquet file for efficient storage and further analysis.

2.4 Outcome Variables

The outcome variable is Life Expectancy. It represents the average number of years a person is expected to live, based on current health conditions and mortality rates. The model aims to understand how various factors, such as healthcare availability, education, and economic status, influence life expectancy across different countries. Figure 1 displays the historical data of life expectancy across 195 countries from 2004 to 2015, comparing developed and developing countries. The data reveals distinct trends in life expectancy, with developed countries generally experiencing higher life expectancy compared to their developing counterparts.

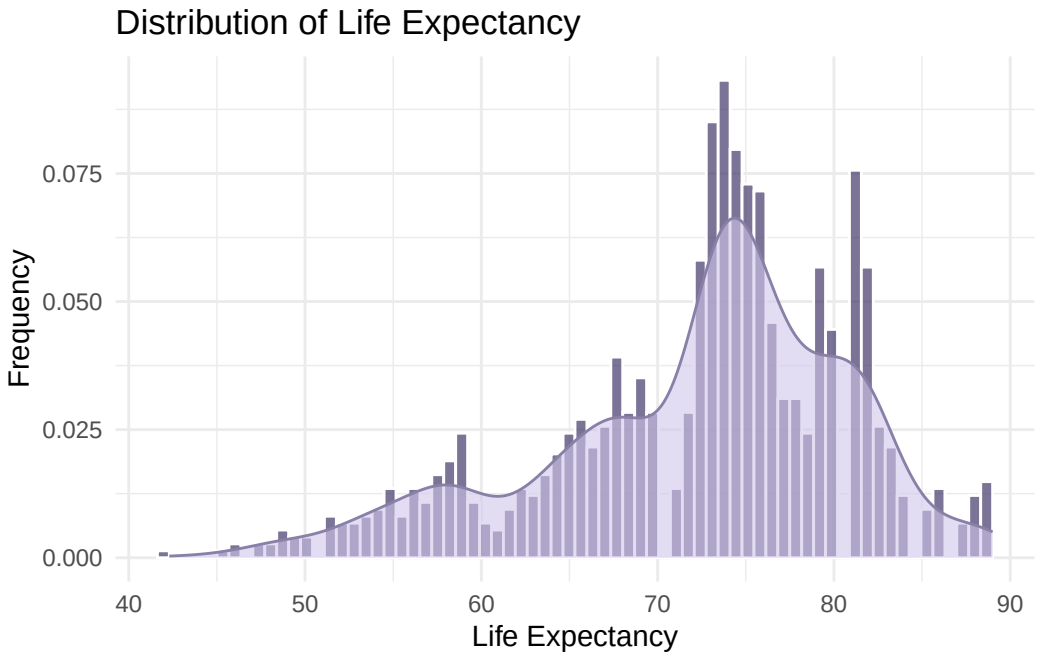

Figure 1: The distribution highlights the frequency of life expectancy values across the dataset. The distribution is slightly right-skewed, with a mean around 73 years, reflecting the overall health trends captured in the data.

2.5 Predictor Variables

The predictor variables (or independent variables) are the factors believed to influence life expectancy:

1. **Status (`Status`)**: This variable categorizes countries into "Developed" and "Developing" groups. The development status of a country plays a significant role in determining life expectancy, with developed countries generally having higher life expectancies due to better healthcare, sanitation, and infrastructure, as depicted in Figure 2.
2. **Diphtheria Immunization Coverage (`Diphtheria`)**: This variable represents the percentage of one-year-olds who have received the Diphtheria, Tetanus, and Pertussis (DTP3) immunization. Immunization coverageFigure 3 is an important health indicator, as it reflects the effectiveness of a country's healthcare system in preventing infectious diseases, which in turn affects life expectancy.
3. **Schooling (`Schooling`)**: The number of years of schooling in a country is used as a proxy for education levels. Higher education levels typically correlate with better health outcomes and, consequently, higher life expectancy, as educated individuals are more likely to engage in healthier behaviors and have access to better healthcare. The Educated condition is shown in Figure 4.
4. **Medical Doctors (`pct_doctor`)**: This variable indicates the number of medical doctors per 10,000 peopleFigure 6. Access to healthcare professionals is a key determinant of health outcomes, and a higher number of doctors generally corresponds with better healthcare services and longer life expectancy.
5. **Nursing and Midwifery Personnel (`pct_nursing`)**: Similar to the number of doctors, this variable represents the number of nursing and midwifery personnel per 10,000 populationFigure 7. Nurses play a significant role in delivering healthcare, and higher availability of nursing staff can contribute to improved health outcomes and increased life expectancy.
6. **Percentage of Health Expenditure (`HealthExpenditure`)**: This variable reflects the general government expenditure on health as a percentage of total government expenditure, as shown in Figure 5. Higher health spending is often associated with better healthcare infrastructure, which can lead to longer life expectancy due to better disease prevention and treatment options.

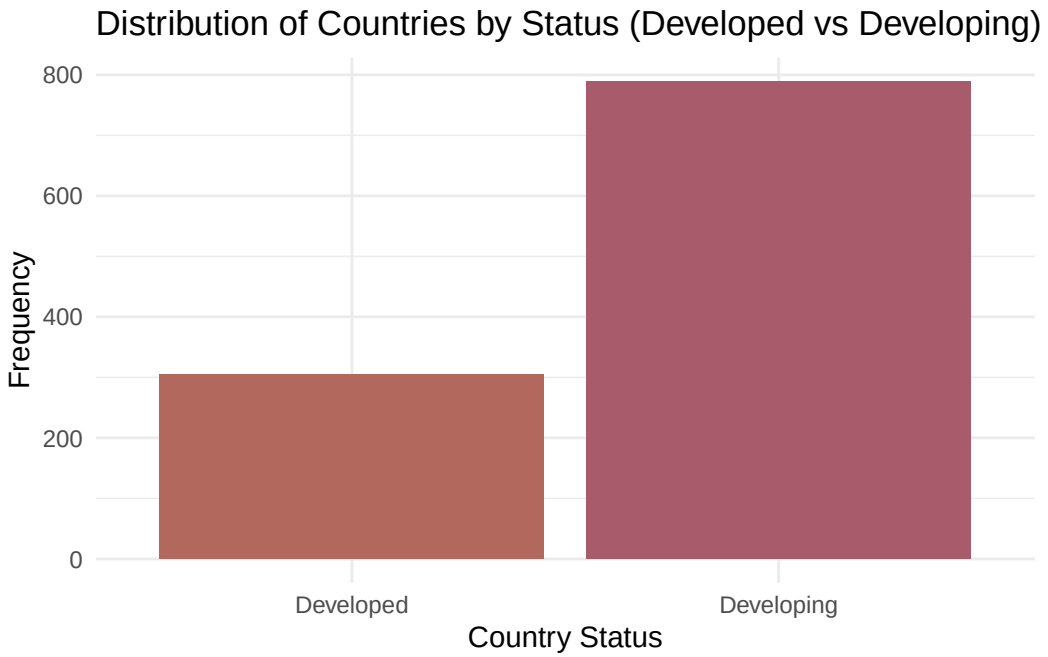

Figure 2: The distribution of Status shows the counts of developed and developing countries in the dataset. There are 305 data points from developed countries and 788 from developing countries.

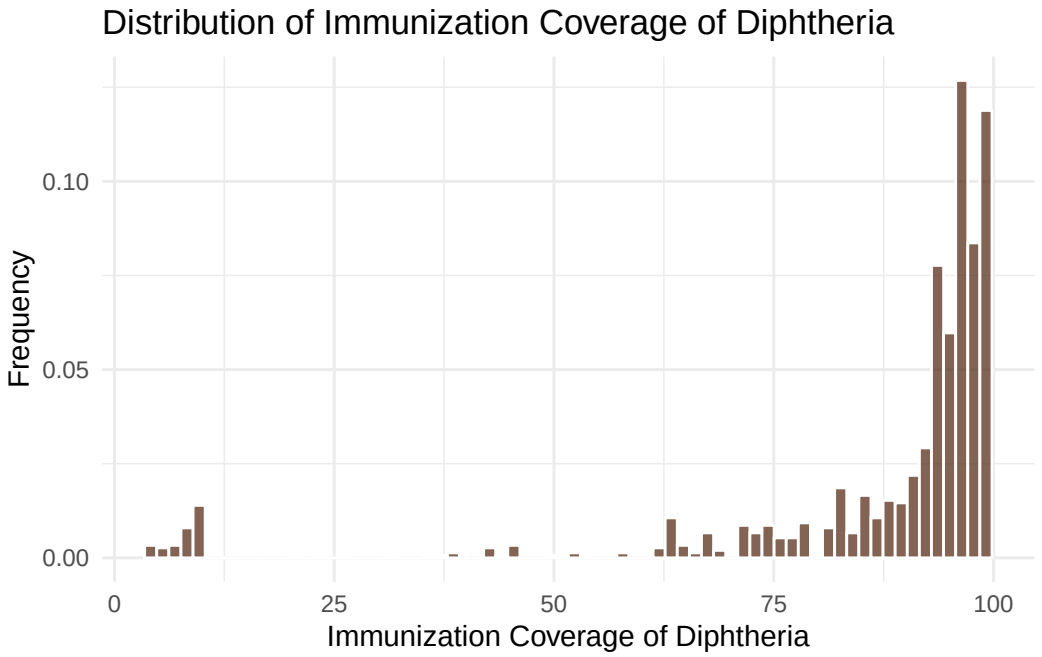

Figure 3: The distribution of immunization coverage for diphtheria reveals that immunization coverage is generally high, with most countries achieving high vaccination rates. However, there is some variance, reflecting disparities in immunization efforts across different regions.

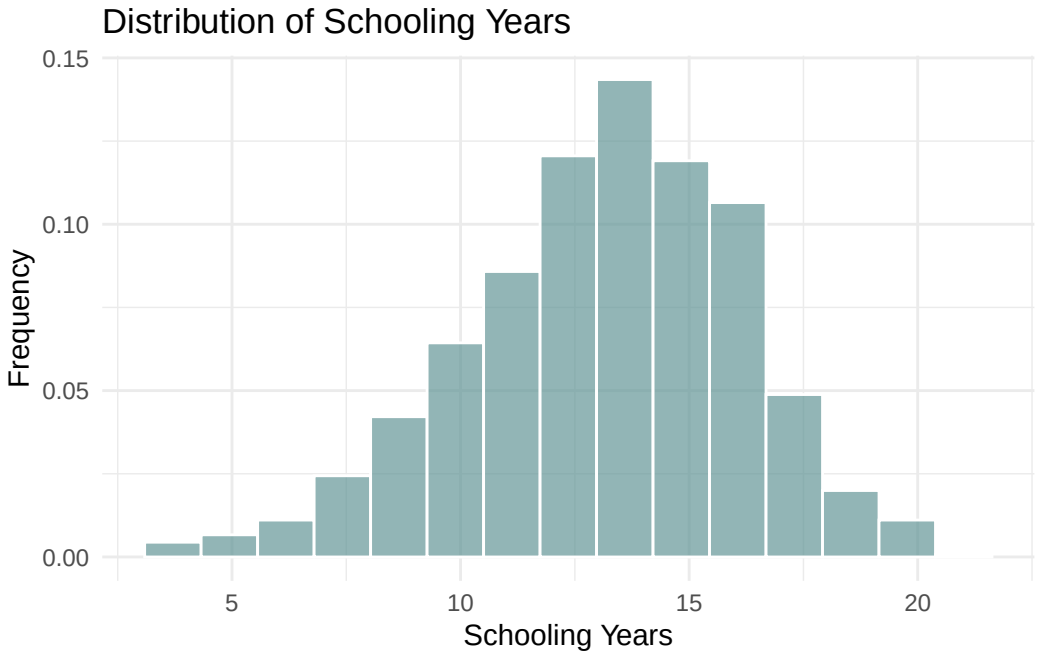

Figure 4: The distribution of Schooling Years is shown as a representation of the Education Factor, the distribution of schooling years across the dataset approaches a normal distribution

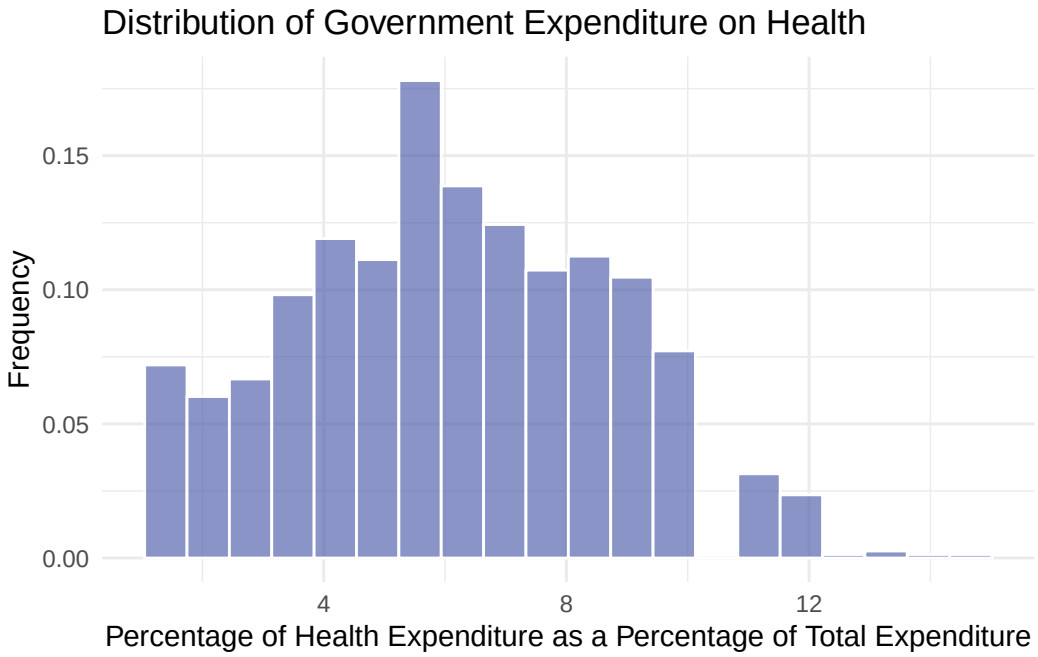

Figure 5: The distribution of government expenditure on health as a percentage of total expenditure reflects the level of medical support provided by governments.

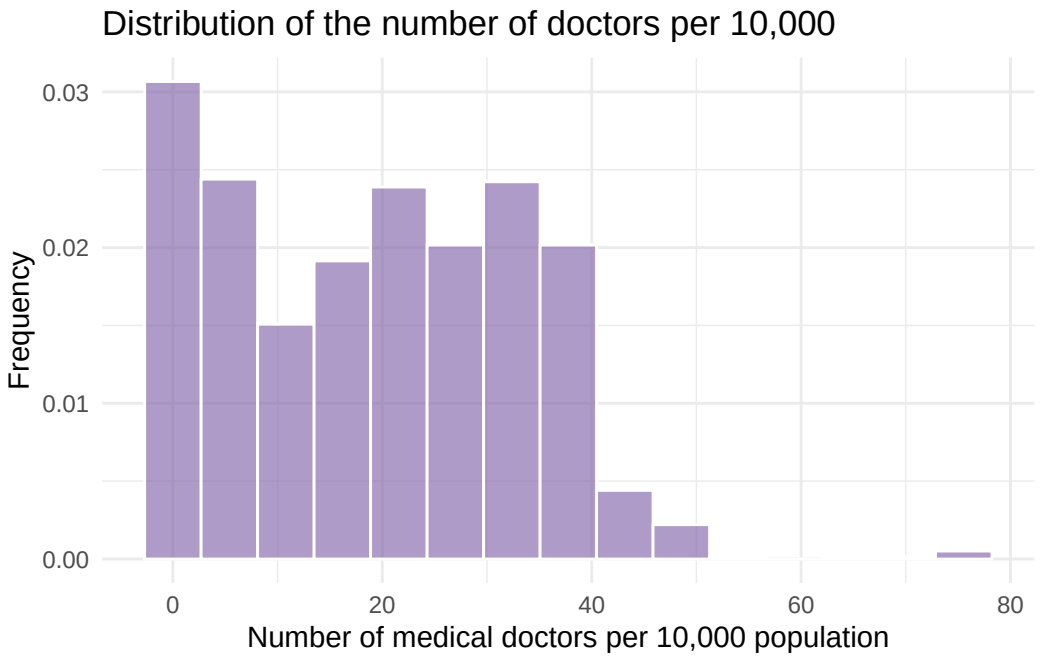

Figure 6: The distribution of doctors per 10,000 population, providing an overview of the distribution of medical doctors globally, reflecting general medical resource availability.

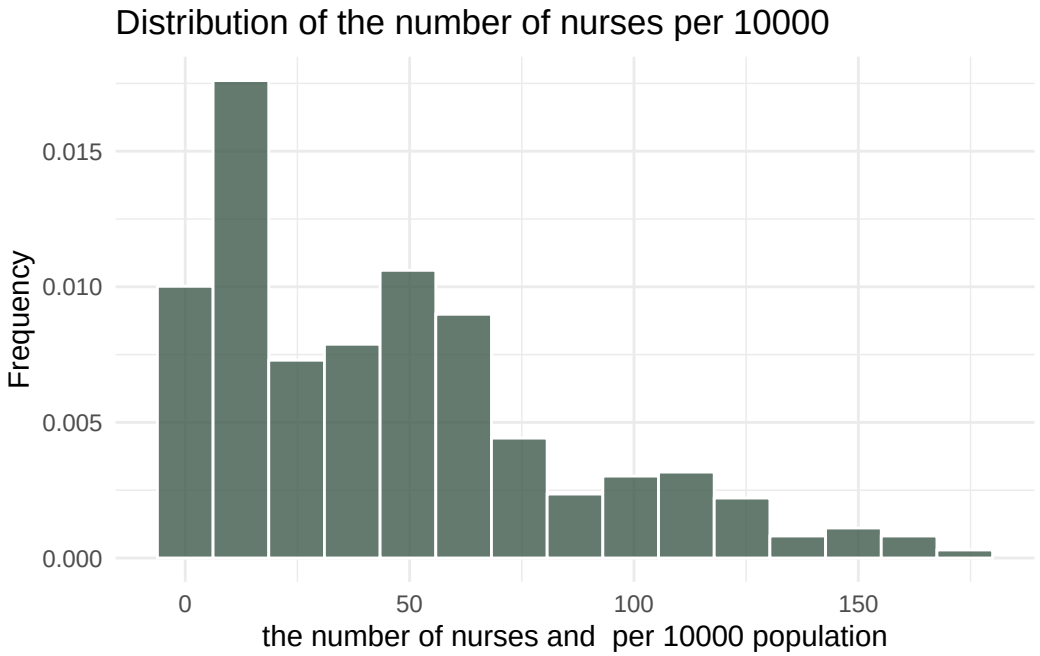

Figure 7: The distribution of nurses per 10,000 population, offering a broad view of nurse density worldwide as an indicator of healthcare support.

3 Model

The goal of this analysis is to construct a linear model to predict life expectancy (`LifeExpectancy`), which represents the average number of years an individual is expected to live. This model incorporates several predictor variables that capture economic and social factors influencing life expectancy.

3.1 Model set-up

The models assume a linear relationship between the predictors and life expectancy, where each predictor contributes independently or interactively to the variation in the outcome variable. Define y_i as the Life Expectancy for country i. The factors x_1, x_2,...represent socioeconomic predictors such as Diphtheria Immunization Coverage, Schooling, Health Expenditure, Number of Doctors and Nurses per 10000, and interaction term.

The linear regression model for developed and developing regions are specified as follows:

$$\begin{aligned}\text{LifeExpectancy}_{\text{developing}} = \beta_0 &+ \beta_1 \text{Diphtheria Immunization}_i + \beta_2 \text{Schooling}_i + \\ &\beta_3 \text{Health Expenditure}_i + \beta_4 \text{Number of Doctors per 10000}_i + \\ &\beta_5 \text{Number of Nurses per 10000}_i + \\ &\beta_6 \text{Health Expenditure}_i \times \text{Doctors per 10000}_i \times \text{Nurses per 10000}_i + \epsilon_i\end{aligned}$$

$$\begin{aligned}\text{LifeExpectancy}_{\text{developing}} = \beta_0 &+ \beta_1 \text{Diphtheria Immunization}_i + \beta_2 \text{Schooling}_i + \\ &\beta_3 \text{Health Expenditure}_i + \beta_4 \text{Number of Doctors per 10000}_i + \\ &\beta_5 \text{Number of Nurses per 10000}_i + \\ &\beta_6 \text{Health Expenditure}_i \times \text{Doctors per 10000}_i \times \text{Nurses per 10000}_i + \epsilon_i\end{aligned}$$

Where:

- `LifeExpectancy`: Represents the average number of years a person is expected to live in a specific country, based on current mortality trends.
- `Diphtheria Immunization`: The percentage of children who received the diphtheria-tetanus-pertussis (DTP3) vaccine, serving as a proxy for the quality and reach of immunization programs.
- `Schooling`: Average years of schooling for individuals aged 25 and above, reflecting the level of education in the population.
- `Health Expenditure`: The percentage of government expenditure allocated to healthcare, indicating the priority given to public health.

- `Number of Doctors per 10,000`: The availability of medical doctors per 10,000 people, capturing the accessibility of professional healthcare services.
- `Number of Nurses per 10,000`: The availability of nursing staff per 10,000 people, further representing healthcare capacity.
- Interaction Term: The combined effect of health expenditure, doctors, and nurses on life expectancy, considering how these variables may amplify or mitigate each other's impact.
- epsilon is the error term, assumed to be normally distributed with a mean of 0 and constant variance $\epsilon_i \sim N(0, \sigma^2)$

The model assumes that the predictors have a linear relationship with the response variable, i.e., the life expectancy is linearly related to the selected features. We are interested in estimating the coefficients β_1, β_2,...,β_4 which represent the effect of each predictor and interaction term in the outcome.

3.2 Model Justification

The linear regression model was chosen for this analysis due to its simplicity and interpretability. The coefficients from this model directly show how each predictor affects life expectancy, assuming other variables remain constant. This makes it ideal for understanding the impact of healthcare and economic factors on life expectancy in both developed and developing countries.

For the developed countries model, predictors like Diphtheria Immunization, Schooling, Health Expenditure, and the number of Doctors per 10,000 were included. These variables are well-established factors influencing life expectancy in high-income nations (Organization 2024; Labs 2021). In the developing countries model, additional predictors such as Schooling, and healthcare access variables like Doctors and Nurses per 10,000 were added to account for differences in healthcare infrastructure.

Linear regression is suitable for our dataset, as it assumes a normally distributed continuous response variable. While more complex models were considered, linear regression was preferred for its transparency and ease of interpretation, aligning with the goals of this analysis. Additionally, previous studies have shown that healthcare access, education, and economic factors play a significant role in determining life expectancy, particularly in developing countries (Hanushek and Woessmann 2016; He and Li 2020). Including these predictors ensures the model is built on sound reasoning and established domain knowledge.

3.3 Model Weaknesses and Limitations

The linear regression model used to estimate life expectancy relies on several key assumptions that must be met for valid results. First, it assumes linearity, meaning there is a straight-line relationship between the predictors (e.g., healthcare condition, education levels) and the outcome variable (`life expectancy`). This assumption ensures that a change in a predictor corresponds to a consistent change in life expectancy. Second, the model assumes that the errors are uncorrelated, meaning the residuals do not exhibit any patterns when plotted. This assumption is significant for making accurate inferences, such as significance tests and confidence intervals. Third, the model assumes homoscedasticity, which means that the variance of the residuals should remain constant across all values of the predictors. This assumption helps avoid bias in estimating the relationship between the predictors and the outcome variable. Fourth, the residuals should follow a normal distribution. This assumption is necessary for performing accurate hypothesis tests and for generating valid confidence intervals, which provide insights into the uncertainty of the model's predictions. Lastly, it is important that the model does not appear multicollinearity, where predictors are highly correlated with each other. These assumptions provide a foundation for the model, violations can lead to misleading results with inaccurate coefficients.

However, several limitations apply to this model. Despite the assumption of linearity, real-world relationships between life expectancy and its predictors may be more complex or non-linear, and such complexity would not be fully captured by this model. For instance, the effect of economic development on life expectancy may vary significantly at different levels of income. Moreover, while the model assumes uncorrelated errors, there may be factors that influence both the predictors and life expectancy, leading to correlated residuals. For example, social determinants like income inequality or access to healthcare might impact both education and life expectancy, potentially violating the assumption of independence. The assumption of constant variance also has limitations, as certain groups (e.g., countries with lower health expenditure) may have more variability in life expectancy than others, potentially leading to not constant variance. Additionally, the assumption of normally distributed errors may not hold in all cases, especially when extreme outliers (such as countries with unusual life expectancy rates) are present in the data.

Other external factors not accounted for in this model may also influence life expectancy. Environmental factors like climate change, sudden public health crises (e.g., pandemics), and political instability could have significant effects on life expectancy, yet these factors may not be included as predictors in the model. These limitations should be considered when interpreting the model's results, particularly when applying the findings to specific populations or making broad predictions.

3.4 Model Validation

The models for both Developed and Developing countries were implemented using R, with applying out-of-sample testing, the data was randomly split into training and test sets, with 80% used for training the model and 20% reserved for testing. This allows us to assess how well the model performs on data that was not used during training. The Root Mean Square Error (RMSE) was calculated for each model to evaluate the performance and generalization to unseen data. A lower RMSE indicates better predictive performance.

In Table 2, the Developed model is more efficient and better fitted primarily due to its higher R^2 (64.5% vs 51.5%), indicating that it explains a greater proportion of variance in life expectancy. Additionally, its RMSE value is lower (3.76 vs 4.94), showing more precise predictions. The Developing model has a larger sample size and variability, which might have diluted the effect of individual predictors. The better fit of the Developed model is partly due to the smaller, more homogeneous sample, allowing it to provide more reliable estimates with fewer outliers.

Extended versions of these tables can be found in Table 3

4 Results

Table 1: The prediction table indicates that the average life expectancy in developed countries is 78.52 years, while in developing countries, it is slightly lower at 71.67 years. This difference of approximately 5.1 years in life expectancy suggests a significant health and infrastructure disparity between developed and developing countries. These results align with broader global trends where developed countries typically benefit from better healthcare systems, higher standards of living, and developed medical technology, contributing to longer life expectancies.

Average Life Expectancy in Developed Countries	Average Life Expectancy in Developing Countries
78.52	71.67

Table 2: Summary of key model estimates for Developed and Developing counties, including coefficients for predictors like Diphtheria, Schooling, and Percentage of HealthExpenditure, pct_doctor, pct_nursing, with standard errors for each estimate. Model performance statistics, such as sample size, R^2, and adjusted R^2, are also displayed.

	Model for Developed	Model for Developing
Diphtheria Immunization	0.023	0.031
	(0.024)	(0.009)
Schooling	0.645	1.694
	(0.161)	(0.106)
Health Expenditure	−1.740	−0.352
	(1.025)	(0.161)
Number of Medical Doctors per 10000	−0.446	0.444
	(0.232)	(0.084)
Number of Nurses per 10000	−0.120	−0.009
	(0.106)	(0.033)
Num.Obs.	305	789
R2	0.191	0.634
R2 Adj.	0.166	0.630
AIC	1695.5	4780.8
BIC	1736.4	4832.1
Log.Lik.	−836.754	−2379.376
F	7.739	150.059
RMSE	3.76	4.94

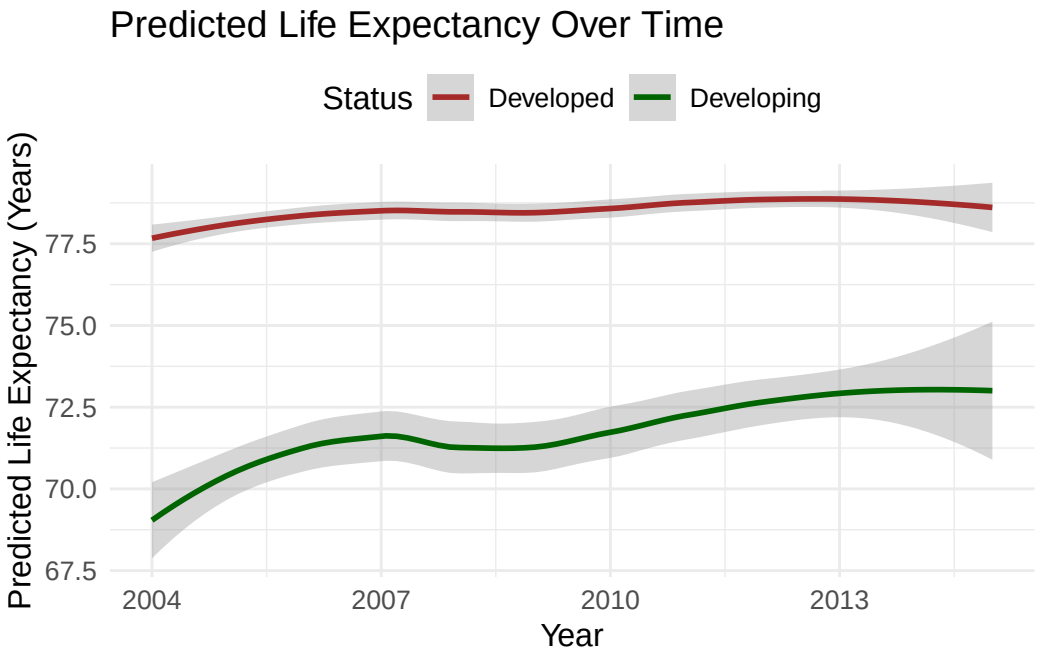

Figure 8: The prediction plots demonstrate that life expectancy in developed countries tends to remain higher with less variance, indicating stability in health outcomes due to better healthcare systems, more consistent access to medical services, and overall economic stability with less variance. In contrast, the life expectancy in developing countries exhibits a broader range of variation, with a more rapid growth trajectory. This suggests that while life expectancy is increasing in these countries, it is influenced by a variety of factors such as improvements in healthcare, economic development, and access to resources, all of which vary more significantly across different regions.

The predicted life expectancy distributions across various predictors reveal important trends. Longer schooling yearsFigure 9 strongly correlate with higher predicted life expectancy, emphasizing the positive impact of education. High diphtheria immunization coverageFigure 10 is also associated with increased life expectancy, though this relationship is less consistent for medium and low coverage levels, indicating variability in its predictive contribution. Similarly, higher government health expenditureFigure 11 aligns with greater life expectancy, while lower expenditure shows limited variability in predicted outcomes. Lastly, the distribution of medical professionalsFigure 12 highlights that developed medical conditions generally contribute to higher life expectancy, though some variability in their influence remains.

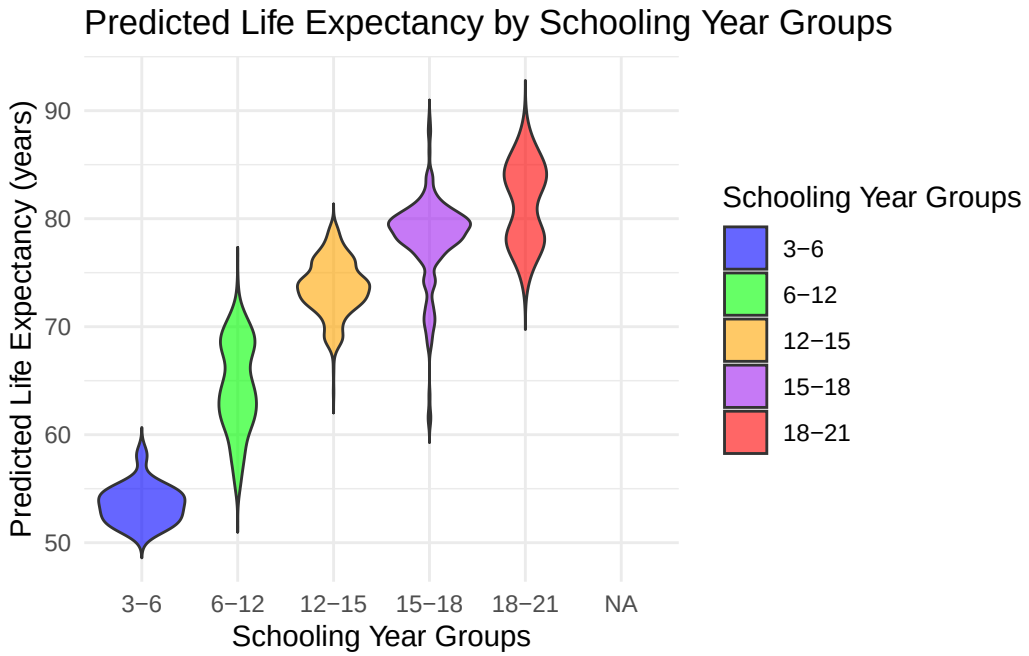

Figure 9: The distribution shows how different levels of education correlate with predicted life expectancy, highlighting group-wise distribution and variability.

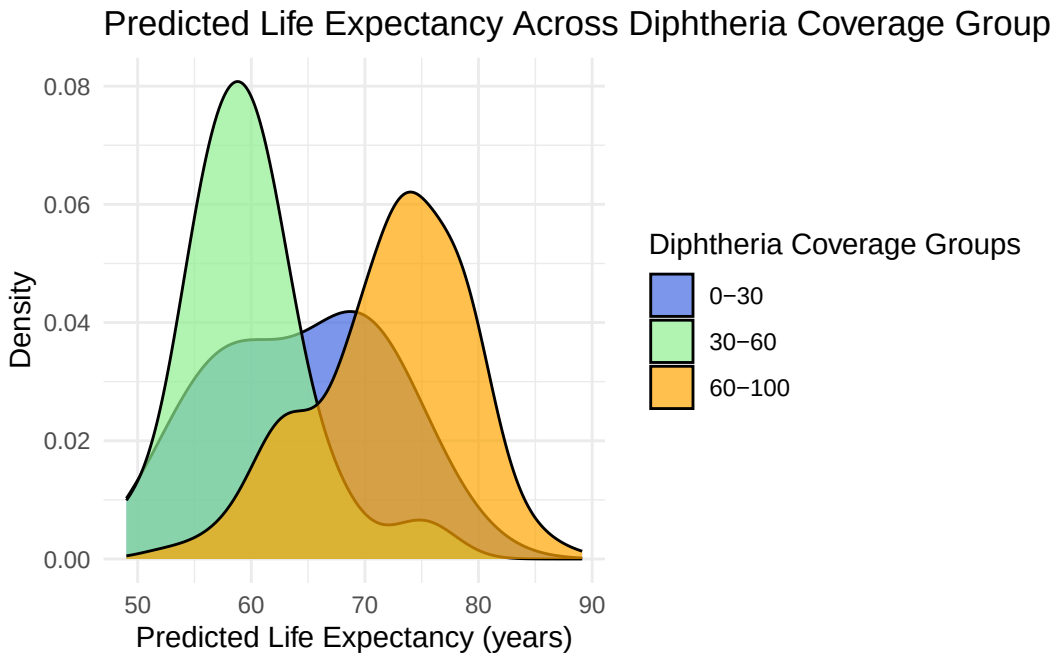

Figure 10: The distribution illustrates how varying immunization coverage levels contribute to differences in predicted life expectancy.

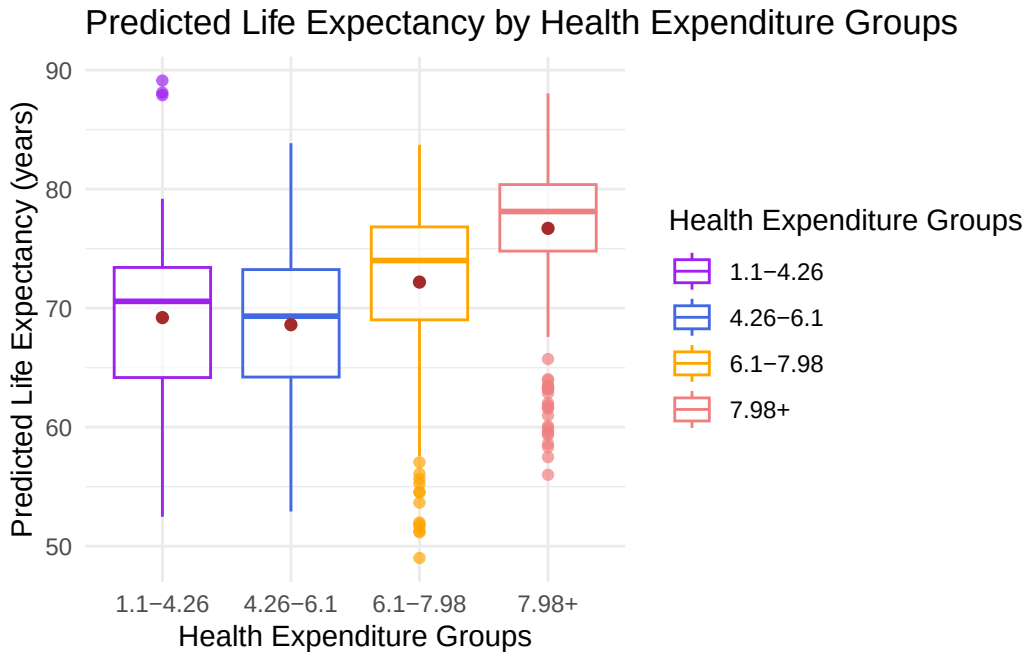

Figure 11: The distribution reveals the distribution of predicted life expectancy across health expenditure levels, showing central tendencies and variability.

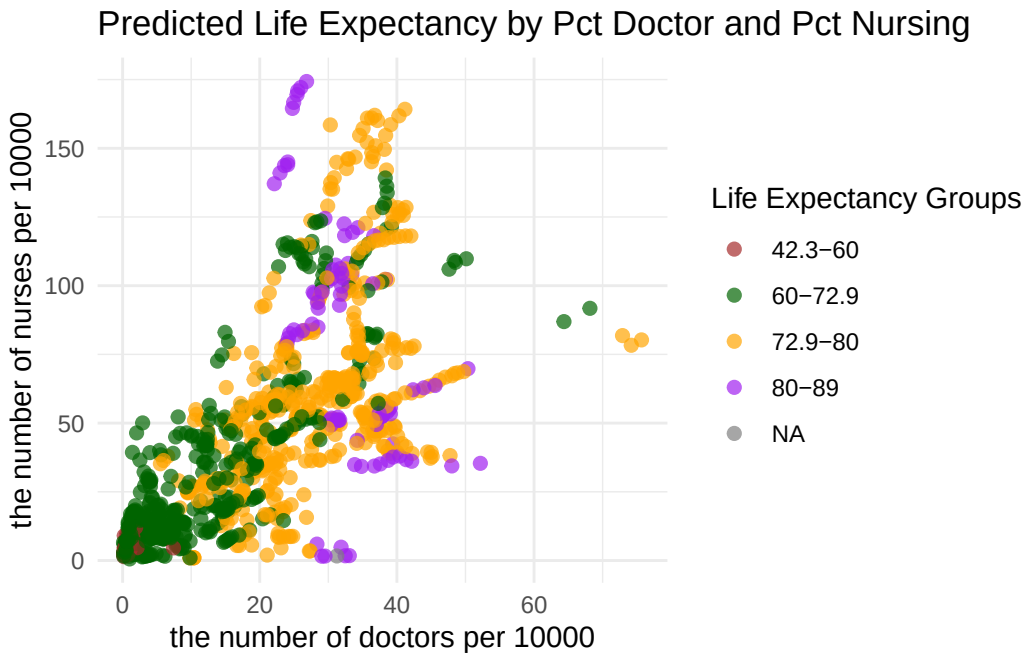

Figure 12: The distribution demonstrates the relationship between medical professional availability (doctors and nurses) and predicted life expectancy, highlighting patterns across groups.

5 Discussion

5.1 Factors Influencing Life Expectancy Across Nations

The study aims to develop a model for predicting life expectancy based on healthcare, economic, and educational factors. By analyzing the relationship between these predictors and actual life expectancy, we gain insights into the factors that influence life expectancy outcomes and evaluate the model's predictive performance.

A comparison of predicted life expectancy distributions reveals several notable trends. Longer schooling years strongly correlate with higher life expectancy, highlighting the critical role of education. High diphtheria immunization coverage generally predicts greater life expectancy, though medium and low coverage levels exhibit variability, suggesting that immunization is not the sole or dominant factor. Similarly, higher percentage of government health expenditure aligns with improved outcomes, while lower spending shows minimal variation. Availability of doctors and nurses also indicates a positive relationship with life expectancy, though some variability suggests other contributing factors are at play.

These findings align with prior research, demonstrating that healthcare and education are pivotal in shaping life expectancy. The model effectively captures these relationships, offering valuable insights for addressing disparities between developed and developing nations.

5.2 Educational and Healthcare Factors as Significant Predictors

A key finding of this study is the distinct roles that educational and healthcare factors play in predicting life expectancy across developed and developing countries. In the model for developed nations, schooling is a significant positive predictor (0.645), though its impact is less pronounced compared to developing nations (1.694). This suggests that while education universally contributes to longer life expectancy, its influence is more significant in countries where educational access is still expanding.

Healthcare variables show contrasting trends between the two groups. Health expenditure negatively correlates with life expectancy in both contexts but is more substantial in developed countries (-1.740) compared to developing countries (-0.352). This reflects diminishing returns on investment in already developed healthcare systems in developed nations, where additional spending may not yield proportional gains in life expectancy. Similarly, the availability of medical doctors demonstrates a divergent impact, showing a negative association in developed nations (-0.446) but a positive one in developing nations (0.444). This suggests that increasing access to medical professionals is a more critical driver of life expectancy in developing regions, where healthcare systems are less established.

These findings highlight the nuanced roles of education and healthcare in shaping life expectancy, emphasizing the need for tailored policy approaches to address disparities across nations.

5.3 Variance in Life Expectancy and the Influence of Social Factors

The analysis underscores significant disparities in life expectancy variance between developed and developing countries, shaped by both systemic and social factors. In developed nations, life expectancy remains higher with less variance, a reflection of robust healthcare infrastructure, widespread access to medical services, and economic stability. These advantages are complemented by stronger social support systems, higher education levels, and improved public health awareness, which collectively enhance the consistency of health outcomes.

In contrast, developing countries exhibit greater variability in life expectancy, driven by a complex interplay of economic growth, healthcare access, and social determinants such as education and community infrastructure. Regions with better access to schooling and healthcare services tend to show faster life expectancy improvements, while others lag behind due to inequities in resource distribution and social support. This dynamic highlights the dual role of economic and social factors in shaping life expectancy trends, underscoring the need for comprehensive policies that prioritize education, equitable healthcare, and community development to reduce disparities and foster sustainable progress.

5.4 Weaknesses and Next Steps

One limitation of the current model is the exclusion of other potential factors that could influence life expectancy. While the model includes several key predictors like schooling, health expenditure, and the number of medical professionals, it does not account for other significant variables such as social determinants of health (e.g., income inequality, access to social services, or political stability), which could have a notable impact, particularly in developing countries. Environmental factors such as air quality, climate change, and water sanitation also play critical roles, especially in low-income regions, yet these were not included in the model. Additionally, there is the potential for bias in the dataset itself. If the dataset used for training the model is not representative of all countries, particularly those with unique socio-economic or healthcare challenges, it could limit the model's generalizability. Bias could also arise if the data predominantly reflects developed nations, where healthcare systems and economic factors are more consistent, leading to skewed predictions for developing countries.

Future research should explore several avenues to improve the model. First, expanding the dataset to include additional variables, such as policies related to social welfare, political stability, and government regulations, would provide a more comprehensive view of the factors influencing life expectancy. Incorporating data on policy changes or government reforms that impact healthcare and social services could also enrich the model's predictive power. Another area for future research could involve exploring the impact of environmental factors, such as pollution levels, access to clean water, and geographical conditions. These variables are especially relevant in developing countries and could reveal new insights into life expectancy

trends. By addressing these limitations and exploring these additional variables, future research can refine models of life expectancy and improve predictions for countries at various stages of economic and healthcare development.

Appendix

5.1 Data cleaning

To prepare the dataset for analysis, I followed these steps:

1. Column Selection and Formatting: Removed irrelevant columns, such as administrative codes, and reformatted the "Year" column into a standardized four-digit format (YYYY) for consistency. Created a "Health_Care_Access" score by aggregating scaled values for "Medical Doctors per 10,000" and "Health Expenditure" to enable focused analysis.
2. Handling Missing Values and Grouping Categories: Addressed missing values in key variables like "Schooling" and "Health Expenditure" by applying group-wise median imputation based on development status (Developed or Developing). For categorical fields like "Status," missing entries were labeled as "Unknown." Countries with over 50% incomplete data were excluded, while less common regions were grouped into an "Other Regions" category to simplify comparisons.
3. Adding New Columns: Created a "Health_Care_Access" score by aggregating scaled values for "Medical Doctors per 10,000" and "Nurses per 10,000" Added a binary "Developed_Status" column, encoding 1 for developed countries and 0 for developing countries, based on established World Bank classifications.
4. Outlier Detection and Export: Verified and handled outliers in "Health Expenditure" and "Life Expectancy" using interquartile range (IQR) thresholds. The cleaned dataset was exported in both CSV and Parquet formats for efficient storage and compatibility with developed tools.

5.2 Idealized Methodology

5.2.1 Population, Frame, and Sample

The population for this study encompasses all individuals worldwide whose life expectancy is influenced by various socioeconomic, healthcare, and environmental factors. This includes individuals living in developed and developing countries, spanning different ages, genders, and social groups.

The frame consists of a comprehensive dataset of individuals with demographic, healthcare, and socioeconomic information relevant to life expectancy. This dataset could be sourced from

global organizations such as the World Health Organization (WHO), World Bank, or national statistical agencies. The frame should ideally include recent data to capture current trends and factors affecting life expectancy.

The sample would be a representative subset of individuals drawn from the frame, stratified by development status (Developed or Developing), age groups, and regions. This stratification ensures diversity and allows for comparisons across different demographics. The sample should include sufficient representation from high-variance populations, such as those in regions undergoing rapid economic or healthcare changes, to enhance the model's predictive accuracy.

5.2.2 Sample Recruitment

- **Recruitment via National and Global Surveys**: Individuals could be recruited through existing national and international health surveys, such as those conducted by the World Health Organization (WHO), World Bank, or national census bureaus. Participants would be invited to contribute additional data on healthcare access, education levels, and socioeconomic factors relevant to life expectancy.
- **Collaboration with Healthcare Providers**: Healthcare facilities, including hospitals and clinics, could assist in recruiting participants by inviting patients to contribute anonymized health data for research purposes. Recruitment could occur during routine health check-ups or through online health portals.
- **Community Outreach Programs**: Recruitment could target communities through local organizations, NGOs, and social programs. These efforts could focus on ensuring representation from underrepresented or high-variance populations, such as rural areas or rapidly developing regions.
- **Incentives**: Participants could be incentivized with benefits such as access to personalized health reports, free health screenings, or community-based health initiatives to improve overall participation rates.
- **Digital Platforms**: Online platforms like health-focused apps, educational websites, and social media groups could provide an avenue for researchers to engage potential participants and collect survey data. These platforms are especially useful for reaching younger or tech-savvy demographics.

5.2.3 Sampling Approach

The study employs a stratified random sampling method, dividing the population into strata based on key factors that influence life expectancy, such as geographic region (e.g., developed vs. developing countries), age group (e.g., 0-20, 21-40, 41-60, 61+), income level (low, middle, high), and access to healthcare (e.g., urban, suburban, rural). This ensures that the sample

represents diverse subgroups affecting life expectancy outcomes. For instance, suppose the study aims to collect data from 1,200 respondents. If stratification assigns 50% of the sample to developed countries and 50% to developing countries, with age distribution set at 20% for 0-20, 30% for 21-40, 30% for 41-60, and 20% for 61+, then 180 participants from developed countries would be aged 21-40. Further stratification could divide these respondents by income level, where 40% fall under the high-income bracket, equating to 72 participants.

- **Representativeness**: Stratified sampling ensures that the sample reflects the variability in life expectancy determinants, enabling more precise insights into how different factors (e.g., healthcare access, socioeconomic status) influence outcomes.
- **Precision**: By ensuring balanced representation from various subgroups, this approach minimizes biases and provides clearer insights into the differences between demographic or regional groups.
- **Complexity**: The need to define and categorize strata, particularly in diverse populations, can make stratified sampling more complex to implement. For example, defining healthcare accessibility uniformly across regions may be challenging.
- **Cost and Logistics**: Stratified sampling may involve higher costs and more administrative effort, particularly in reaching underrepresented or remote populations. Recruitment and data collection across strata could also take longer compared to simpler methods like random or convenience sampling.

5.2.4 Non-response handling

Non-response occurs when individuals or groups do not provide their data, fail to complete surveys, or are otherwise unavailable for participation. Effectively managing non-response is significant to preserving the representativeness and reliability of the findings.

- **Follow-up Invitations**: If participants do not respond initially, reminders can be sent via email, phone calls, or community leaders. Personalized messages highlighting the importance of their input to public health and societal improvements may encourage participation.
- **Weighting for Non-Response**: Statistical weighting can be applied if certain demographics (e.g., rural residents, lower-income groups) have higher non-response rates. Adjusting weights ensures that these groups are proportionally represented in the analysis, mitigating potential biases.
- **Adjustment**: If systematic non-response is identified (e.g., fewer responses from regions with limited internet access), additional data collection efforts in these areas or adjustments during analysis using statistical techniques like imputation can be employed to fill gaps.

- **Engaging Local Organizations**: Partnering with local health agencies or NGOs in areas with traditionally low participation can foster trust and improve response rates.

5.2.5 Standards for the Questionaires

- **More Context**: Participants could be asked to provide detailed information on factors influencing life expectancy, such as diet, physical activity levels, access to healthcare, occupational hazards, and exposure to pollution. Including such variables could enhance the model's ability to capture key influences beyond demographic or geographic data.
- **Flexibility**: The questionnaire could utilize a mix of open-ended and scaled questions to gather both quantitative and qualitative insights. This design would allow participants to elaborate on unique life circumstances or local conditions that may not be captured in structured questions.
- **Self-Reporting Bias**: A significant limitation of questionnaires is the tendency for self-reporting bias. Participants might overestimate healthy behaviors, such as exercise or nutritious eating, or underreport harmful habits like smoking or alcohol consumption. This could distort the findings and reduce model accuracy.

5.3 Idealized Survey

A sample survey can be found at: https://forms.gle/Jzpw6MR5hXgnAfiDA

This survey is conducted as part of a research study to explore how various factors—such as healthcare condition, immunization coverage, and education—impact life expectancy, providing valuable insights into their relationship with overall health outcomes in your community. It includes questions about your healthcare experiences, education, and awareness of health initiatives. Your responses will remain completely anonymous, and no personal data will be shared or used beyond this research. This survey should take 5 to 7 minutes to complete. Thank you for taking the time to share your insights.

Survey Coordinator: KarmenZhu
Email: karmen.zhu@mail.utoronto.ca

Section 1: Personal Information

1. Please select your age group:(Select one)

- Under 20
- 20-29
- 30-39
- 40-49

- 50-59
- 60 or older

2. What is your country of residence?

- (please specify):_____

3. What is your highest level of education completed or currently pursuing? (Select one)

- No formal education
- Primary school
- High school diploma
- Some college/university
- Bachelor's degree
- Master's degree
- Doctoral degree (e.g., Ph.D.)
- Professional degree (e.g., MD, JD, MBA)
- Other (please specify): _____

4. What is your primary occupation? (Select one)

- Full-time employed
- Part-time employed
- Self-employed
- Student
- Retired
- Other (please specify): _____

Section 2: Health and Healthcare Information

5. Have you received a diphtheria immunization in the past 10 years? (Select one)

- Yes
- No
- I don't know

6. What is the approximate distance to the nearest healthcare facility (e.g., clinic, hospital)? (Select one)

- Less than 5 km (3 miles)
- 5-10 km (3-6 miles)
- 11-20 km (7-12 miles)
- More than 20 km (12 miles)

7. How many walking-distance clinics or hospitals are located near your home? (Select one)

- None
- 1-2
- 3-4
- 5 or more

8. How long do you typically wait to receive care at a government healthcare facility? (Select one)

- Less than 1 hour
- 1-3 hours
- 3-6 hours
- More than 6 hours

9. How much of your medical expenses (e.g., doctor visits, hospital stays, medication) does your health insurance cover?

- (please specify):_____

10. How much of your monthly income is spent on healthcare (including insurance, out-of-pocket expenses, etc.)? (Select one)

- Less than 5%
- 5-10%
- 11-20%
- More than 20%

11. Do you have a family doctor? If so, how often do you visit your family doctor? (Select one)

- never
- Rarely (Once a year)
- 2–3 times a year
- More than 3 times a year

Final Section

Thank you for completing this survey! Your responses will help improve public health initiatives.

5.4 Additional Tables & Figures

Table 3: Summary of the Life Expectancy model, which includes key variables such as schooling years and the percentage of government health expenditure. The table presents the model coefficients along with their standard errors.

Term	Estimate	Std. Error	t Value	Pr(>|t|)	Term_Type
Intercept	44.501	1.314	33.865	0.000	Linear
Diphtheria	0.031	0.009	3.582	0.000	Linear
Schooling	1.694	0.106	15.954	0.000	Linear
TotalExpenditure	-0.352	0.161	-2.183	0.029	Linear
pct_doctor	0.444	0.084	5.277	0.000	Linear
pct_nursing	-0.009	0.033	-0.257	0.797	Linear
TotalExpenditurepct_doctor	-0.013	0.012	-1.122	0.262	Linear
TotalExpenditurepct_nursing	0.012	0.005	2.304	0.022	Linear
pct_doctorpct_nursing	-0.006	0.001	-4.508	0.000	Linear
TotalExpenditurepct_doctorpct_nursing	0.000	0.000	0.775	0.438	Linear

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Clarke, Erik, Scott Sherrill-Mix, and Charlotte Dawson. 2023. *Ggbeeswarm: Categorical Scatter (Violin Point) Plots*. https://github.com/eclarke/ggbeeswarm.

Garon, Julie R. et al. 2015. "The Challenge of Global Diphtheria Eradication." *Global Health* 29 (4).

Girum, Tadele et al. 2018. "Determinants of Life Expectancy in Low and Medium Human Development Index Countries." *Medical Studies* 34: 218–25. https://www.researchgate.net/publication/328140185_Determinants_of_life_expectancy_in_low_and_medium_human_development_index_countries.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Hanushek, Eric A., and Ludger Woessmann. 2016. "The Role of Education Quality for Economic Growth." *Economics of Education Review* 55: 13–27. https://hdl.handle.net/10986/7154.

He, L., and N. Li. 2020. "The Linkages Between Life Expectancy and Economic Growth: Some New Evidence." *Empirical Economics* 58: 2381–2402. https://doi.org/10.1007/s00181-018-1612-7.

Labs, John Snow. 2021. "Global Life Expectancy and Healthy Life Expectancy." *John Snow Labs Marketplace*. https://www.johnsnowlabs.com/marketplace/global-life-expectancy-and-healthy-life-expectancy/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Organization, World Health. 2024. "Healthy Life Expectancy at Birth (Years)." *Data.who.int*. https://data.who.int/indicators/i/48D9B0C/C64284D.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://broom.tidymodels.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wilke, Claus O. 2024. *Ggridges: Ridgeline Plots in 'Ggplot2'*. https://wilkelab.org/ggridges/.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/.